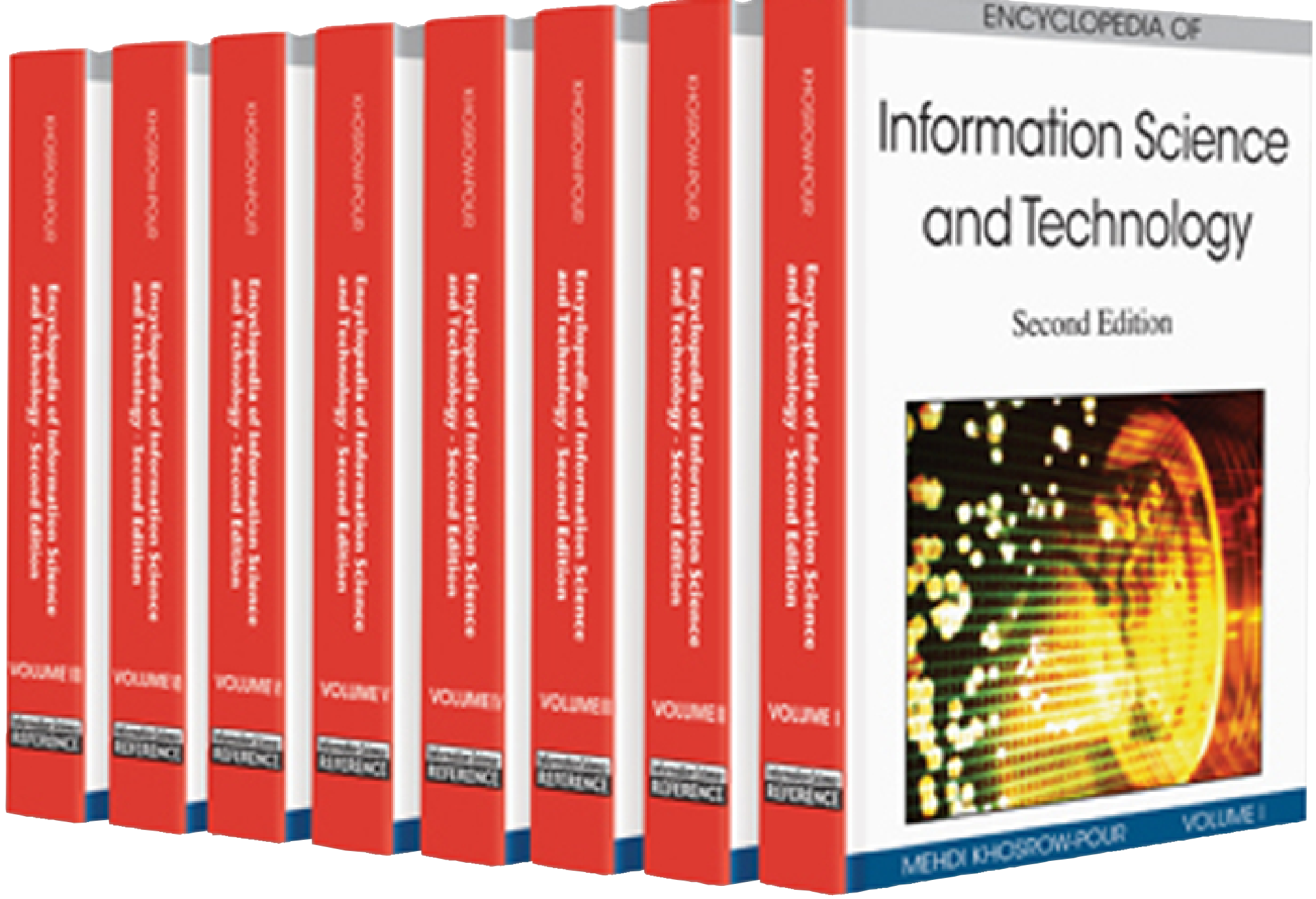

ENCYCLOPEDIA OF
Information Science and Technology
Second Edition
MEHDI KHOSROW-POUR
VOLUME I

Encyclopedia of Information Science and Technology

Second Edition

Mehdi Khosrow-Pour
Information Resources Management Association, USA

Information Science REFERENCE

INFORMATION SCI
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of information science and technology / Mehdi Khosrow-Pour, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set of books represents a detailed compendium of authoritative, research-based entries that define the contemporary state of knowledge on technology"--Provided by publisher.
ISBN 978-1-60566-026-4 (hardcover) -- ISBN 978-1-60566-027-1 (ebook)
1. Information science--Encyclopedias. 2. Information technology--Encyclopedias. I. Khosrowpour, Mehdi, 1951-
Z1006.E566 2008
004'.03--dc22
2008029068

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Editorial Advisory Board

List of Contributors

Contents
by Category

Artificial Intelligence

Bioinformatics

Business Information Systems

Cognitive Informatics

Data Mining & Databases

Electronic Business

Electronic Commerce

Electronic Government

Environmental Informatics

Global Information Technology

Health Information Systems

High Performance Computing

Human Aspects of Technology

Industrial Informatics

IT Education

IT Security & Ethics

Knowledge Management

Library Science

Medical Technologies

Mobile & Wireless Computing

Multimedia Technology

Networking & Telecommunication

Social Computing

Software & Systems Design

User-Centered Technologies

Web Technologies

Contents
by Volume

Volume I

Volume II

Volume III

Volume IV

Volume V

Volume VI

Volume VII

Volume VIII

Preface

Influencing every facet of business, society, and life worldwide, with speed beyond imagination, the field of information science and technology has without doubt brought upon a revolution in the way the human population interacts, does business, and governs. As one takes into account the leaps and bounds experienced in information sharing and communication exchange over the last few decades, a truly admirable phenomenon presents itself and clearly shows that the results of this pivotal rising will monumentally impact the way the world thinks, subsists, and evolves.

With a long history of expeditious evolution, the growth and expansion of information technology began during the early 1950s with the main purpose of initiating scientific computing, expanding research, and utilizing the power of computers as a means to support a mass volume of computational tasks in scientific applications and discoveries. Later, during the 1960s and '70s, the use of computer technology was also extended to business applications mostly in accounting and financial areas that involved processing numbers and collecting data in a quantitative sense. As a result, the use of this technology was limited to those who had an expansive knowledge of these systems and had access to computer programming languages. With the evolution of computers and telecommunications in the 1980s, a new information technology was born with a strong focus on the management and dissemination of information by both information providers and users across the globe.

In the early 1990s, the most noticeable advancement in the information technology revolution was the creation of the Internet. During the past two decades, Internet technologies have become the driving force in allowing people worldwide to communicate and exchange information, creating a new dimension that is virtual, interactive, and provides a digital forum for global social connection. In recent years, through the use of Web-enabled technologies, organizations of all types and sizes around the world have managed to utilize these technologies to disseminate and process information with prospective customers, suppliers, students, and governments. Today, the ability to communicate and connect from many locations through personal computers has influenced different people in many different societies. These technologies allow everyone, regardless of their geographic location, to bring the information age to its full realization.

In recent years, the science of understanding the nature of information processing and management along with the computers and technologies that decipher, disseminate, and manage information has become known as information science and technology. Technology has profoundly impacted science, business, and society, thus constructing an entity that improves access to the rapidly expanding body of knowledge in almost every discipline. Society fuels this knowledge creation, as it receives, manages, educates, and collects information. The volume and intensity of research in information science and technology have exceeded many other fields of science, and research discoveries have become the impetus behind many emerging tools and applications seen at every organizational level.

In addressing this need for the representation of evolving information science and technology disciplines in academic literature, the First Edition of the Encyclopedia of Information Science and Technology, released in early 2005, positioned itself as the first of its kind in reference publications, offering an invaluable source of 554 articles highlighting major breakthroughs, discoveries, and authoritative research results in technological advancements. In providing this compendium of references, definitions and key words within this field of pivotal social and organizational movement, the Five-Volume Encyclopedia of Information Science and Technology, First Edition, supplied researchers with a definitive one-stop reference source.

With the endeavor of progressing from this precedence, and in effort to exhibit the latest research innovations, the Second Edition of the Encyclopedia of Information Science and Technology collects and uncovers the most current research findings related to technological, organizational and managerial issues, challenges, trends, and applications of information

technologies in modern organizations. The coverage of this Encyclopedia seeks to bridge existing gaps in available references on technology and methodologies with its contribution as a valuable resource of encompassing paradigms shaping the ever changing research, theory and discovery of information science and technology.

Including a thorough coverage of innovative topics and terms embodying the disciplines that construct this field, the Second Edition of the Encyclopedia of Information Science and Technology offers the most comprehensive list of research references to further support these developments. New articles focusing on emerging topics, as well as enhanced and updated articles featured in the First Edition, allow for the Encyclopedia of Information Science and Technology, Second Edition to construct an up-to-date exhibition of must know developments.

The articles were carefully selected for this publication for their presentation of the most comprehensive, innovative, and in-depth coverage of information science and technology. The topics covered in this all-encompassing publication, include the most influential areas of this field.

Articles fully cover the area of Artificial Intelligence in their examination of current applications applied within organizational spheres that are impacted by such technologies as Machine Learning, Computational Intelligence, Digital Ecosystems and Adaptive Systems. By providing these chapters, the Encyclopedia is able to supply audiences with reputable sources for identifying future trends in breakthrough computing and technologies that will direct everyday aspects of life, such as intelligent transportation, smart homes, and intelligent diagnostics.

Through introducing the changing state of business information systems, the Encyclopedia analyzes the driving force of globalization and the concept's effect on international trade, economics, and capital. Because business information systems are so far reaching, from the intranet to export tracking and manufacturing intelligence, the research results comprise a breadth of review and discussion of methodologies behind these systems and how they are implemented in the field of business.

As a growing discipline of theory, concept and science, cognitive informatics is discussed in several articles that delve into the philosophy behind knowledge, how behavior is determined through this analysis, and what impacts are made through this careful evaluation. With such areas as cybernetics and systems theory being implemented into social and organizational practice, the references provided through this research offer invaluable sources of further evaluation.

Continuously utilized for the evolution of technical applications, the areas of data mining and databases are comprehensively discussed in a significant level of articles in the Encyclopedia as a channel for introduction and advancement. In taking into account the growing incorporation of data mining into Web development and databases into engineering progress, these resources supply audiences with authoritative results and a foundation for additional research.

With a considerable effect on the global economy, electronic business applications are discussed in the Encyclopedia and offer readers a credible source of knowledge for understanding the current realms of mobile applications for business functions, technologies for marketing, and virtual enterprises. Seeing the importance of these technology-based management systems, modern electronic business is analyzed as a growing phenomenon of capitalism.

As today's form of trade, shopping and purchasing, electronic commerce is fully presented in the Encyclopedia as a phenomenon of progressive change. With significant growth since the dot com explosion of the past two decades, e-commerce is exposed as a discipline with global reach and international application for the way currency is exchanged and services are dealt.

With undeniable influence on policy making and delivery, aspects of electronic government are examined in the Encyclopedia, including electronic voting, public service delivery, and citizen management. Considering the social change developing as a result of e-government implementation, articles offer researchers and policy makers a wealth of perspective on the challenges and opportunities in directing government via information communication technologies.

Given the worldwide focus on the environment, the Encyclopedia provides audiences with emerging findings in the area of environmental informatics. As a growing area of attention, this discipline is supported by prominent international researchers studying the future of maintaining environmental functions and safeguarding the planet.

Global information technology has emerged as an area of study with growing importance as digital communications spread internationally. The Encyclopedia supplies articles that delve into the importance of the digital divide, as well as case studies exhibiting regional adaptation, resistance, and adoption of technologies.

As technological demands extend and users multiply, high-performance computing seeks to maintain the flow of communication between servers, knowledge workers, and organizations. With much recent advancement in Grid computing, and technologies to handle large amounts of data, the Encyclopedia offers readers a close look at these breaking areas of study and applications of growing necessity.

Considering the human element in making technology the paradigm it is today, the Encyclopedia provides readers with a comprehensive view of human aspects of technology. With an analysis of end-user behavior, gender differences and ubiquity of computing, readers will find an extensive amount of research results and analysis to assist in building literature in the important area of study; human computer interaction.

Following the advancements of the industrial revolution, industrial informatics is described in the Encyclopedia as a bridge between technological progress and the application of manufacturing, transportation, and construction. With growing utilization in enhancing and expediting good production, building and usage, research results examine the future trends of such technologies and how these applications will incorporate into the daily life of individuals.

As a contemporary social motivator, IT Education is comprehensively examined in the Encyclopedia. This is an area of research with philosophical and constructivist reach into all levels of learning. In discussing blended learning, distance education, tools for online learning, advanced pedagogy and computational support for teaching, IT education demands further attention in developing reliable research. As a popular area of examination, the Encyclopedia lends support to emerging trends in the discipline while supplying a venue for further discussion of terms, themes, and additional implication.

Taking into account the massive growth in technological utilization for sensitive data, government, personal and organizational functions, IT security and ethics as an area of study continues to involve expanding securitization and streamlining for the best protection of digital information. In examining such areas as digital forensics, authentication, cryptography, cyber warfare, and trustworthy computing, analysis of IT security and ethics take precedence as a point of key discussion in the Encyclopedia, with the endeavor to enrich the field.

As a result of the current structures of both public and private sector administration, the field of knowledge management has been positioned as a concept of both utility and application. Considering the need for quality sources indicating best practices in the management of knowledge workers, IT governance and information sharing, the Encyclopedia supplies readers and researchers with a considerable selection of current themes, terms and topics relating to the state of knowledge management.

With many library systems implementing digital filing and cataloguing systems, the study of bibliometrics has flourished in the scientific world. With cutting edge technologies in archiving and classification, preservation and reformatting, library science has developed into an intricate branch of information science and the Encyclopedia illustrates explanations of the vast spectrum of this study. As issues such as open access, security, and Internet property rights take center stage, legal aspects of the digitization of information are explored.

Mobile devices, wireless systems, sensors, and wearable computing applications are technologies that are shaping communication and public administration. Keeping in mind the spread of the need for reliable means of information transfer, mobile and Wireless computing is explored in the Encyclopedia as a field of widening pervasiveness. With uses ranging from machine correspondence to business interrelationships, important research findings are effectively exposed and discussed by top researchers in the field.

Whether utilized for entertainment, learning, or public policy, multimedia technology as a form of study involves behavioral and practical analysis. The Encyclopedia offers audiences with many aspects of analysis on topics such as gaming, simulation and hypermedia to best understand how the world is displaying its information for consumers and administrators.

With the expansion of wireless networks, telecommunications have flourished as users are able to get in touch from just about anywhere with the use of ad-hoc networks, Bluetooth, and RFIDs. Advancements in neural and wireless sensor networks have increased availability of the Internet exponentially. Peer-to-peer networks are discussed in the Encyclopedia, along with optical access and overlay networks.

Social computing describes the intersection of social behaviors and computer systems, intricately examining areas such as social networking sites, augmented realities, and online auctions. Tools such as blogs, Wikis, tags, and podcasts have expanded the user's online researching experience. The Encyclopedia expands on social informatics and explores instant messaging, virtual groups, mailing lists, and forums; just a few of the places that users interact with one another.

Technological engineers focus on algorithms, modeling languages, and kernel applications to design and format the machines that the world uses today. Systems analysis, visualization, and diagnostics all play a role in the design and improvement of the ever changing software being developed. The Encyclopedia takes a look at the vulnerabilities, specifications, and quality of software and its architectures.

As a result of the Internet explosion of the 1990s, Web technology as a field of study has taken the stage as a discipline of ubiquitous importance. With the Web being a single source of infinite information, responding and expanding at exponential rates, the Encyclopedia offers readers a snap shot of disciplines such as portal technologies, semantic computing, Web 2.0, and service-oriented technologies.

The Encyclopedia elaborates on the expansive area of bioinformatics. Subjects such as artificial immune systems, biomedical imaging, and biometric tools such as those used for identifying humans by biological, intrinsic traits are discussed. Topics of importance also include organic computing, proteomics, and computational chemistry.

Electronic patient records, e-health, and cybertherapy are three of the many topics explored through the Encyclopedia's investigation of health information science. With new assistive and rehabilitative technologies, medical data storage, and issues of security and privacy, Health Informatics affects the lives of human beings worldwide. Considering this growth, research results prove valuable for healthcare administrators and academic disciplines, such as nursing and health management.

Advanced medical technologies are incorporated into daily measures to secure and save lives. Medical informatics is exemplified in this second edition Encyclopedia through an intricate look at the updates and improvements in this pivotal field. Medical imaging, biosensors, and new nanotechnology modernizations for surgical procedures are just a few of the exciting and significant advances in medical technologies. Rehabilitation, disease detection, and routine medical care are assisted and revamped with advances in hospital machinery and tools.

The selected topics of this encyclopedia are to provide the best balanced compilation of concepts and issues from researchers around the world. These researchers were asked to submit proposals describing the topic and scope of their articles. All proposals were carefully reviewed for suitability by the editor-in-chief. Upon the receipt of full entry submissions, each contribution was forwarded to at least three expert external reviewers on a double-blind, peer review basis. Only submissions with strong and favorable reviews were chosen as entries for this encyclopedia. In most cases, submissions were sent back for several revisions prior to final acceptance. The goal was to assemble the best minds in the information science and technology field from all over the world to contribute entries to this encyclopedia and to apply the highest level of quality in selecting entries.

As a result, over 650 entries were selected for inclusion in this eight-volume encyclopedia, highlighting current concepts, issues, and emerging technologies. All entries are written by more than 1,512 knowledgeable, distinguished scholars from many prominent research institutions throughout the world. Over 5,000 technical and managerial terms and their definitions have been organized by the authors to enhance the articles, allowing for extensive research into core concepts and ideas. In addition, this eight-volume set offers a thorough reference section with over 14,500 sources of additional information for scholars, students, and researchers in the field of information science and technology.

Multiple detailed Tables of Contents have been organized to better assist readers in navigating and identifying information. Contents are structured through alphabetical, categorical, and contributing author listings for simple reference. The Encyclopedia includes these three comprehensive and detailed tables of contents to greatly assist readers in locating articles on topics within a particular discipline, as well as identify any particular entry per author or title.

Complimentary online access to this encyclopedia for the life of the edition will also be provided to any library with the purchase of the print copy. This complimentary online availability will allow students, faculty, researchers, and corporate managers to access the latest contents of this comprehensive and in-depth encyclopedia regardless of their location. This particular feature will prove to be an extremely valuable resource for distance learning educational programs worldwide.

The field of information science and technology today is a collection of many specific disciplines that researchers have created. No longer is this discipline limited to a few technology-related areas. Today, information science and technology as a paradigm is heavily intertwined in medicine, learning, finance, government, and many other areas. Technology is constantly changing and improving, necessitating the creation and study of innovative literature in many disciplines of information science and technology.

The diverse and comprehensive coverage of multiple disciplines of information science and technology in this eight-volume, authoritative encyclopedia will contribute to a better understanding of all topics, research, and discoveries in this evolving field. Furthermore, the contributions included in this publication will be instrumental in the expansion of knowledge in this field. This publication will inspire its readers to further contribute to the current discoveries in this immense field, creating possibilities for further research and discovery into the future of information science and technology and what lies ahead for the knowledge society.

Acknowledgment

Putting together a comprehensive publication of this magnitude requires a tremendous contribution and much assistance from everyone involved. The most important goal of editing this encyclopedia, as with most publications, was identifying and selecting quality contributors. I am indebted to all of the authors for their excellent contributions. The degree of quality that this publication reaches could not have been achieved without the valuable peer reviews the authors had provided me through their expertise and their rigorous, unbiased assessment of the manuscripts assigned to them on a double-blind basis. I am extremely grateful for their contribution to this publication. I would also like to express my gratitude and deep appreciation to the members of the Editorial Advisory Board for their wisdom, guidance, and assistance in deciding on different issues related to this publication.

I would also like to convey my deep appreciation and gratefulness to IGI Global's Vice President of Editorial, Ms. Jan Travers, for all of her wisdom and encouragement. Additionally, I would like to express my appreciation to the editorial staff of IGI Global for their assistance in this project: Jan Travers, Vice President of Editorial, Kristin M. Klinger, Director of Editorial Content, Meg Stocking, Assistant Executive Editor, Kristin M. Roth, Managing Development Editor, Jennifer Neidig, Director of Production, Deborah Yahnke, Assistant Development Editor, Heather Probst, Assistant Development Editor, and Amanda Steel, Editorial Assistant.

Finally, I would like to express my warmest thanks to my wife, Beth Peiffer, for her support, wisdom, encouragement, understanding, patience, and love. My heart also goes to my two young girls, Basha and Anar, for the joys that they have brought to my life. Finally, much gratitude goes to all those who have taught me immeasurable amounts during the past few decades.

Mehdi Khosrow-Pour, D.B.A.
Editor-in-Chief

About the Editor

Mehdi Khosrow-Pour, DBA, received his Doctorate in Business Administration (DBA) from the Nova Southeastern University (Fla.), USA. Dr. Khosrow-Pour has taught undergraduate and graduate information system courses at the Pennsylvania State University – Harrisburg for 20 years where he was the chair of the information Systems Department for 14 years. He is currently president and publisher of IGI Global, an international academic publishing house with headquarters in Hershey, PA and an editorial office in New York City (www.igi-global.com). He also serves as executive director of the Information Resources Management Association (IRMA) (www.irma-international.org), and executive director of the World Forgotten Children Foundation (www.world-forgotten-children.org).

He is the author/editor of over 20 books in information technology management. He is also the editor-in-chief of the *Information Resources Management Journal, Journal of Cases on Information Technology, Journal of Electronic Commerce in Organizations*, and *Journal of Information Technology Research*, and has authored more than 50 articles published in various conference proceedings and journals.

A

Accessibility of Online Library Information for People with Disabilities

Axel Schmetzke
University of Wisconsin-Stevens Point, USA

INTRODUCTION

After 20 years of digitization efforts, hardly a single type of library information resource remains that has not shifted, at least to some extent, to an electronic, Web-based format: information about the library itself, catalogs, indexes, dictionaries and encyclopedias, books and journals, tutorials, reserve materials, and reference services. The online migration of these resources has opened unprecedented opportunities to people with "print disabilities" who cannot independently access printed works because of lack of sight, dyslexia, or insufficient motor control (Coombs, 2000), but who are able to access electronic text with the help of assistive input and output technology such as modified computer keyboards and screen readers with speech or Braille output (Lazzaro, 2001; Mates, 2000).

The extent to which these new opportunities become realized depends on the design of the Web environment. From the perspective of accessibility, design in the online world matters as much as it does in the physical world. This article seeks to determine the extent to which the library profession addresses the need of people with disabilities for accessibly designed online resources—by reviewing the professional library literature for coverage of this issue, by summarizing empirical accessibility studies, and by analyzing pertinent policies adapted by libraries and their professional organizations.

COVERAGE OF ONLINE ACCESSIBILITY IN THE LIBRARY LITERATURE

In 1996, accessible Web design began to emerge as an issue in the professional library literature. Since 1999, there has been a noticeable increase in library-related journal publications that investigate the accessibility of Web-based library information, seek to raise awareness concerning the need for accessible Web design, and provide practical tips (for a detailed overview, see Schmetzke, 2003, p. 153-156; Stewart, Narendra, and Schmetzke, 2005, p. 267-270). Since 2001, two library journals, *Computers in Libraries* (2001), and *Library Hi Tech* (Schmetzke, 2002a, 2002b, 2007a) have devoted special-theme issues to online accessibility; *Information Technology and Disability* reports regularly on the subject. In 1999, the American Library Association began publishing monographs that addressed accessible Web design (Lazzaro, 2001; Mates, 2000; McNulty, 1999). Gradually, the need to include people with disabilities is also acknowledged in the broader library literature on electronic resources: Whereas some authors—such as Breivik & Gee (*Higher Education in the Internet Age*, 2006), Gregory (*Selecting and Managing Electronic Resources*, 2006) and the contributors to Lee (*Collection Management and Strategic Access to Digital Resources*, 2005)—continue to ignore the issue, others deal with it, at least briefly, in connection with topics such as Web page design (Garlock & Piontek, 1999), Web site usability testing (Norlin & Winter, 2002), digital resources selection and digital video (Curtis, 2005; Hanson & Lubotsky Levin, 2003; Kovacs & Robinson, 2004; Lilly, 2001), Web-based instruction (Sharpless Smith, 2006), and virtual reference service (Coffman, 2003).

EMPIRICAL RESEARCH FINDINGS

Of the online resources provided by libraries, Web pages have been studied the most. The majority of studies employed Bobby, a software-based accessibility checker, to investigate conformance to the 1999 Web content accessibility guidelines (WCAG), developed by the World Wide Web Consortium's Web Accessibility Initiative. Recently, researchers also began looking at compliance with the "access board" standards, a similar set of accessible design criteria developed under Section 508 of the U.S. Rehabilitation Act Amendments of 1998 (Architectural and Transportation Barriers Compliance Board, 2000).

At the library Web sites evaluated between 1999 and 2002, 19% to 75% of the Web pages were found to be free of major accessibility problems (Blake, 2000; Kester, 1999; Lilly & Van Fleet, 1999, 2000; Schmetzke, 2001a, 2003; Spindler, 2002, Yu, 2002); the average number of errors per page varied between 1.3 and 6.1 (Schmetzke, 2002c). Web accessibility tended to be higher at academic libraries than at public libraries. More recent data, available only for academic libraries continue to show a mixed picture. On the average, library Web sites have become more accessible. In a national sample of 49 U.S. libraries, pages free of major Bobby-detectable barriers (compliance with

priority-1 WCAG check-points) have increased from 47% in 2002 to 59% in 2006 (Comeaux & Schmetzke, 2006). With 72%, Web site accessibility is higher at University of Wisconsin libraries (Schmetzke, 2005)—in contrast to Kentucky's academic libraries, where, far fewer homepages passed similar accessibility checkpoints; 23% in December 2003 and 37% in March 2007 (Providenti, 2004; Providenti and Zai III, 2007).

Interestingly, Web sites of accredited schools of library and information science (SLIS)—those institutions that train the next generation of librarians—tend to be less accessible than the library Web pages on their campuses (Schmetzke, 2003). In 2002, only 30% of the SLIS pages (at U.S. campuses) were free of barriers. With 36%, accessibility was barely higher in Canadian schools. Although the situation has improved much in Canada (73%), it has done so only mildly in the U.S. (Comeaux & Schmetzke, in press). With only 47% of the pages conforming to the most basic WCAG guidelines, it is reasonable to assume that there is widespread unawareness about the need for accessible design among SLIS Web designers and among those library school faculty and staff who hire the designers and give them direction. Similar lack of awareness among the leadership was also reported for the area of distance education (Schmetzke, 2001b) and in connection with several high-profile technology-promoting initiatives in higher education (Blair, Goldmann, & Relton, 2004).

Although the occasion of a Web-site redesign provides an opportunity for improving accessibility (see Sloan, Gregor, Booth, & Gibson, 2002), it is not always taken advantage of. A comparison of Web accessibility at U.S. libraries between 2000 and 2002, with a break-down of Web sites into those that had undergone a major redesign during the period in question and those that did not, revealed that the percentage of accessible pages in the redesigned set had drastically declined (from 47% to 24%) whereas that in the largely unchanged set had considerably improved (from 68% to 81%) (Schmetzke, 2003). More recent data suggest a reversal of this situation. Redesigned Web sites of both academic libraries and library schools tend to be more accessible than those not having undergone a major overhaul (Comeaux & Schmetzke, in press).

Information about the accessibility of Web-based library resources other than library Web pages is comparatively scarce. Prior to 2002, little had been published in this area. Then, in 2002, *Library Hi Tech* (Schmetzke, 2002a, 2002b) published two special-theme issues that included accessibility studies on selected online catalogs, online indexes and databases, e-journals, online reference works, and courseware. Although few of the online resources reviewed were found to be absolutely inaccessible, most contained at least some accessibility problems (for an overview, see Schmetzke, 2002c). Several authors pointed out that lack of usability, rather than accessibility, was often the problem (Axtell & Dixon, 2002; Byerley & Chambers, 2002). Stewart (2003), whose studies comprised 36 databases, arrived at a similar conclusion. He cautioned that the observed improvement in accessibility, defined in terms of conformance to certain accessible-design standards, does not automatically result in usability. In a follow-up study, Stewart et al. (2005) found similar results: Most sites contained some access board standards (Section 508) violations (e.g., 85% of the sites did not include mechanisms permitting users to bypass repeatedly occurring navigation and page elements), but complete inaccessibility was the exception. A usability component of this study, designed to ascertain the ability of screen-reader users to perform basic search tasks, revealed that if the bar for success was set very low—if it did not matter how cumbersome and twisted the search process was—most databases could be searched successfully. Self-critically, the authors suggested that future studies of this sort should set out to assess usability more broadly—in terms of user-friendliness.

Until 2002, anecdotal evidence suggested that vendors showed little, if any, concern for the accessibility of their products and that their sales representatives were typically ill prepared to discuss the issue. In 2003, survey findings published by Byerley and Chambers (2003) revealed that the situation had changed significantly: Vendors have become more aware of accessibility and started to remove access barriers from their products. However, the authors discovered that vendors' efforts are largely focused on conformance to Section 508 standards. As a recent follow-up survey shows, even four years later only five of twelve companies conduct usability tests with people who have disabilities (Byerley, Chambers, & Thohira, 2007). The survey also revealed that only half of the database companies provide accessibility information on their corporate Web sites, which makes it difficult for accessibility-conscientious customers to make informed purchasing decisions. Few companies seem to regard accessibility as a selling point in their marketing efforts; only 25% of the responding companies stated that they include accessibility information in their product brochures.

ACCESSIBILITY POLICIES

Under the pressure of the Americans with Disabilities Act of 1990 (ADA Handbook, 1995) and the widening influence of Section 508, many U.S. colleges and universities have adopted campus-wide accessible-Web policies during the past years. Typically, these policies either recommend or require compliance with WCAG, the Access Board standards issued under Section 508, or some combination or variation thereof (Bohman, 2004).

Some, mostly larger, academic libraries have picked up the campus-wide mandate for accessible Web pages and addressed it in their own policies. Among the first to do so was

Yale University Library, which, in its *Library Services for Persons with Disabilities Policy Statement* (2000), requires compliance with WCAG's priority levels one and two.

Very few libraries have adopted policies that address the issue of accessibility in connection with the selection and procurement of online information products—policies crucial for the building of an overall barrier-free information infrastructure (National Center on Disability and Access to Education, 2006). An extensive Web search conducted by this author in June 2006, along with an inquiry posted to pertinent electronic discussion forums, found such policies at only eight libraries. Most of them do not require that only accessible resources must be selected; they merely emphasize the need to consider accessibility, along with other criteria, in the selection process (Florida Atlantic University Libraries, 2006; Northcentral University, 2004; University of Vermont Libraries, 2000; University of Wisconsin-Platteville Karrmann Library, 2002; University of Wisconsin-Stout Library Learning Center, 2006; and University of Washington Libraries, 2001). For example, the latter's *Selection Guidelines for Internet Resources* direct librarians to take into account accessibility when selecting online resources and to weigh the "value of the resource … against the access difficulties presented." Two policies are considerably stronger in tone. The *Principles for CSU Acquisition of Electronic Information Resources* (2005) adopted by California State University (CSU), stipulate that "[i]nformation providers should offer interfaces that comply with basic standards for accessibility by users with disabilities." In a similar vein, the University of Wisconsin Libraries' *Strategic Directions for 2005-2007* (2005) call for measures that put in place accessible "workstations, Web pages, Web-based information resources, Web-based instructional applications, and online services." These more stringent policies echo principles of good collection practice advocated by the National Information Standards Organization (2004) and the International Coalition of Library Consortia (1999). According to the former, "collections should be accessible to persons with disabilities and usable effectively in conjunction with adaptive technologies"; the latter considers its members responsible for "ensur[ing] that vendor platforms are … ADA compliant" and urges them to discuss with vendors the development of accessible products.

Recognition of accessibility as an important issue varies among professional library organizations. In 2001, the American Library Association (ALA) approved a policy drafted by one of its branch organizations, the Association of Specialized and Cooperative Library Agencies (ASCLA), that calls for "equal access to library resources" and urges libraries to use "strategies based upon the principles of universal design to ensure that library policy, resources and services meet the needs of all people." However, the need to create a barrier-free electronic environment is not mentioned at all in the policy's section that addresses, in some detail, the accessibility of library services and collections.

Whereas the call for a barrier-free Web environment is present in the ALA-ASCLA policy at least in the form of a general principle, it is completely absent in the *Guidelines for Distance Learning Library Services* (2004) issued by one of ALA's major branches, the Association of College and Research Libraries (ACRL)—guidelines that not only reflect the professional views of the broader library and higher-education community, but that also have been influential in shaping individual libraries' policies and practices. However, the *Guidelines* are currently undergoing a major revision; in light of currently proposed changes (ACRL's Distance Learning Section Wiki, 2006), it seems likely that the new version (to be approved not before 2008) will address the need for accessible online services and resources.

In summer 2007, the ALA Council approved two policies pertaining to digitization: the document "Principles for Digitized Content," which was drafted by the Task Force on Digitization Policy initiated by ALA's Office for Information Technology (OITP), fully acknowledges the needs of people with disabilities by demanding "equitable access to library materials … ensured through maximum accessibility … and barrier-free design" as well as standards and best practices that must serve "the broadest community of users, including those with disabilities." The "Resolution on Accessible Digitization Projects," authored by ALA-ASCLA, calls on ALA to "strongly encourage all libraries and other entities engaging in digitization projects to adopt Section 508 regulations …" (Schmetzke, 2007b).

FUTURE TRENDS

Even though online accessibility is gradually being recognized as important, much remains to be done. The library literature on electronic resources must routinely cover this issue, just as it does with copyright issues. Librarians need to further educate themselves and their leadership, reshape their institutional policies, and develop effective implementation strategies that ensure 100% accessibility of their homegrown Web resources and challenge vendors to make their Web-based products more accessible. Unless this happens, the online library environment is unlikely to transform into what Tim Berners-Lee (no date), WWW inventor and the current director of the World Wide Web Consortium, had envisioned: "The power of the Web is its universality. Access by everyone regardless of disability is an essential aspect."

CONCLUSION

To a large extent, disability is a social construct. Whether individuals with "disabilities" can pursue independent and fulfilling lives is not merely a matter of their particular internal conditions but also a question of enabling or disabling external factors put in place by society and its institutions. Libraries clearly are part of this nexus. By neglecting to remove all barriers from their Web pages and by not adopting policies that seek to realize the opportunities new information technology provides, many libraries currently fail to seek conditions that would enable all people, including those with "disabilities," to participate fully in the evolving information society.

REFERENCES

ADA handbook: disability discrimination: Statutes, regulations and related materials. (1995). Cincinnati: Anderson Publishing Co.

American Library Association. Association of Specialized and Cooperative Library Agencies. (2001, January 16). *Library Services for People with Disabilities Policy*. Retrieved August 31, 2007, from http://www.ala.org/ala/ascla/asclaissues/libraryservices.htm

Architectural and Transportation Barriers Compliance Board. (2000). *Electronic and information technology accessibility standards.* 36 CFR Part 1194. Retrieved June 29, 2006, from http://www.access-board.gov/sec508/standards.htm

Association of College and Research Libraries. (2004). *Guidelines for distance learning library services*. Retrieved June 29, 2006, from http://www.ala.org/ala/acrl/acrlstandards/guidelinesdistancelearning.htm

ACRL's Distance Learning Section Wiki. (2006). *Guidelines for distance learning library services*. Retrieved June 29, 2006, from http://dls.schtuff.com/guidelines_for_distance_learning_library_services

American Library Association. Office for Information Technology. Task Force on Digitization Policy (2007, July 12). *Principles for Digitized Content.* Retrieved August 31, 2007, from http://www.ala.org/ala/washoff/contactwo/oitp/digtask.cfm#prin

Axtell, R., & Dixon, J. M. (2002). Voyager 2000: A review of accessibility for persons with visual disabilities. *Library Hi Tech, 20*(2), 141-147.

Berners-Lee, T. (no date). Cited from the *Web Accessibility Initiative* Web site. Retrieved May 5, 2004, from http://www.w3.org/WAI/

Blair, M. E., Goldmann, H., & Relton, J. (2004). *Access to electronically-mediated education for students with disabilities: Policy issues*. National Center on Disability and Access to Education. Retrieved June 30, 2006, from http://ncdae.org/activities/papers/policy.htm

Blake, S. (2000). Universal access, the ADA, and your library Web page. *Arkansas Libraries, 57*(1), 19-24.

Bohman, P. R. (2004). *University Web accessibility policies: A bridge not quite far enough*. WebAim. Retrieved June 29, 2006, from http://www.Webaim.org/articles/policies/policies_pilot/

Breivik, P. S., & Gee, E. G. (2006). *Higher education in the Internet age. Libraries creating a strategic edge.* Westport, CT: American Council on Education/Praeger.

Byerley, S. L., & Chambers, M. B. (2002). Accessibility and usability of Web-based library databases for non-visual users. *Library Hi Tech, 20*(2), 169-178.

Byerley, S. L., & Chambers, M. B. (2003). Accessibility of Web-based library databases: The vendors' perspectives. *Library Hi Tech, 21*(3), 347-357.

Byerley, S. L., Chambers, M., & Thohira, M. (2007). Accessibility of Web-based library databases: The vendors perspectives in 2007. *Library Hi Tech, 25*(4).

California State University. (2005, October 13). *Principles for CSU Acquisition of Electronic Information Resources.* Retrieved June 29, 2006, from http://seir.calstate.edu/acom/ear/docs/principles.shtml

Coffman, S. (2003). *Going live. Starting & running a virtual reference service*. Chicago: American Library Association.

Comeaux, D., & Schmetzke, A. (2007b). Web accessibility trends at university libraries and library schools. *Library Hi Tech, 25*(4).

Coombs, N. (2000). Enabling technologies. Untangling your Web. *Library Hi Tech, 18*(1), 93-96.

Curtis, D. (2005). *A how-to-do manual for building, managing, and supporting electronic journal collections.* New York: Neal-Schuman.

Garlock, K. L., & Piontek, S. (1999). *Designing Web interfaces to library services and resources.* Chicago: American Library Association.

Gregory, V. L. (2006). *Selecting and managing electronic resources. A how-to-do-it manual for librarians* (rev. ed.). New York: Neal-Schuman.

Hanson, A., & Lubotsky Levin, B. (2003). *Building a virtual library*. Hershey, PA: Information Science Publishing.

Florida Atlantic University Libraries. (2006, March 7). *FAU libraries collection development policy: Adding free Web/Internet resources*. Retrieved June 29, 2006, from http://www.library.fau.edu/policies/cd_free_e-resources.htm

International Coalition of Library Consortia. (1999, January). *Guidelines for technical issues in request for proposal (RFP) requirements and contract negotiations*. Retrieved June 29, 2006, from http://www.library.yale.edu/consortia/techreq.html

Kester, D. (1999). Measuring the sight of your Web site. *North Carolina Libraries, 57*(3), 114-117.

Kovacs, D. K., & Robinson, K. L. (2004). *The Kovacs guide to electronic library collection development*. New York: Neal-Schuhman.

Lazzaro, J. J. (2001). *Adaptive technologies for learning & work environments* (2nd ed.). Chicago: American Library Association.

Lee, S. H. (2005). *Collection management and strategic access to digital resources. The new challenges for research libraries*. Binghamton, NY: Haworth Press.

Lilly, E. B. (2001). Evaluating the virtual library collection. In D. P. Wallace, & C. Van Fleet, C. (Ed.). *Library evaluation: A casebook and can-do guide* (pp. 165-184). Englewood, CO: Libraries Unlimited.

Lilly, E. B., & Van Fleet, C. (2000). Measuring the accessibility of public library home pages. *Reference & User Services Quarterly, 40*(2), 156-163.

Lilly, E. B., & Van Fleet, C. (1999). Wired but not connected: A accessibility of academic library home pages. *The Reference Librarian, 67/68*, 5-28.

Mates, B. T. (2000). *Adaptive technology for the Internet: Making electronic resources accessible to all.* Chicago: American Library Association. [Free online version at http://www.ala.org/ala/products/books/editions/adaptivetechnology.htm]

National Information Standards Organization (NISO). (2004). *A framework of guidance for building good digital collections* (2nd ed.). Retrieved June 29, 2006, from http://www.niso.org/framework/Framework2.html

McNulty, T. (1999). *Accessible libraries on campus. A practical guide for the creation of disability-friendly libraries.* Chicago: Association of College and Research Libraries, American Library Association.

National Center on Disability and Access to Education. (2006). *Let the buyer be aware: The importance of procurement in accessibility policy*. Retrieved June 30, 2006, from http://ncdae.org/policy/procurement.cfm

Norlin, E., & Winters, C. M. (2002). *Usability testing for library Web sites. A hands-on guide.* Chicago: American Library Association.

Northcentral University. (n.d.). *Electronic collection development policy*. Retrieved June 29, 2006, from http://www.ncu.edu/elrc/policy/collection_dev.asp

Providenti, M. (2004). Library Web accessibility at Kentucky's 4-year degree granting colleges and universities. *D-Lib Magazine. 10*(9). Retrieved Sept. 11, 2006, from http://www.dlib.org/dlib/september04/providenti/09providenti.html

Providenti, M., & Zai III, R. (2007). Web accessibility at Kentucky's academic libraries. *Library Hi Tech, 25*(4).

Schmetzke, A. (guest Ed.), (2007a). Accessibility of electronic resources for all (special issue). *Library Hi Tech, 25*(4).

Schmetzke, A. (2007b). Leadership at the American Library Association and accessibility—A critical view. *Library Hi Tech, 25*(4).

Schmetzke, A. (2005). *Web page accessibility on University of Wisconsin campuses: 2005 survey and seven-year trend data*. Retrieved June 29, 2006, from http://library.uwsp.edu/aschmetz/Accessible/UW-Campuses/Survey2005/contents2005.htm

Schmetzke, A. (2003). Web accessibility at University libraries and library schools: 2002 Follow-Up Study. In M. Hricko (Ed.), *Design and implementation of Web-enabled teaching tools* (pp. 145-189). Hershey, PA: Information Science Publishing.

Schmetzke, A. (guest Ed.), (2002a). Accessibility of Web-based information resources for people with disabilities (part one) (special issue). Library Hi Tech, *20*(2).

Schmetzke, A. (guest Ed.), (2002b). Accessibility of Web-based information resources for people with disabilities (part two) (special issue). *Library Hi Tech, 20*(4).

Schmetzke, A. (2002c, July 15-20). The accessibility of online library resources for people with print disabilities: Research and strategies for change. Computers helping people with special needs. In *Proceedings of the 8th International ICCHP Conference,* Linz, Austria. (pp. 390-397). Berlin: Springer Verlag.

Schmetzke, A. (2001a). Web accessibility at university libraries and library schools. *Library Hi Tech, 19*(1), 35-49.

Schmetzke, A. (2001b). Online distance education—"anytime, anywhere" but not for everyone. *Information Technology and Disabilities, 7*(2). Retrieved June 29, 2006, from http://www.rit.edu/~easi/itd/itdv07n2/contents.htm

Sharpless Smith, S. (2006). *Web-based instruction. A guide for libraries* (2nd ed.). Chicago: American Library Association.

Sloan, D., Gregor, P., Booth, P., & Gibson, L. (2002). Auditing accessibility of UK higher education Web sites. *Interacting with Computers, 12*, 313-325.

Spindler, T. (2002). The accessibility of Web pages for mid-sized college and university libraries. *Reference & User Services Quarterly, 42*(2), 149-154.

Stewart, R. (2002, November 15). *Accessibility of online databases. A usability study of research databases.* Technology Access Program. Oregon University. Retrieved May, 2004, from http://tap.oregonstate.edu/research/ahg.htm

Stewart, R., Narendra, V., & Schmetzke, A. (2005). Accessibility and usability of online library databases. *Library Hi Tech, 23*(2), 265-286.

University of Vermont Libraries. Electronic Resources Coordinating Council. (2000, June 16). *Electronic resources collection development policy*. Retrieved June 28, 2006, from http://bailey.uvm.edu/ercc/appendix6.html

University of Washington. University Libraries. (2001, November 29). *Selection guide for Internet resources.* Retrieved May 5, 2004, from http://www.lib.washington.edu/msd/internetselguide.html

University of Wisconsin Libraries. (2005, June 28). *Strategic directions for 2005-2007*. Retrieved June 29, 2006, from http://uwlib.uwsa.edu/strategic%20directions%202005-2007.htm

University of Wisconsin-Platteville Karrmann Library. *Electronic resources collection development policy*. (2002, July 11). Retrieved June 29, 2006, from http://www.uwplatt.edu/library/colldev/policy/cdpolicyelectronicformats.html

University of Wisconsin-Stout Library Learning Center. (2006, Feb. 10). Information resource development policy. Retrieved June 29, 2006, from http://www.uwstout.edu/lib/policies/irdpolicy.htm

World Wide Web Consortium (W3C). (1999). *Web content accessibility guidelines 1.0*. Retrieved June 29, 2006, from http://www.w3.org/TR/WAI-WEBCONTENT/

Yale University Library. (2000). *Library services for persons with disabilities policy statement*. Retrieved June 29, 2006, from http://www.library.yale.edu/Administration/SQIC/spd1.html

Yu, H. (2002). Web accessibility and the law: Recommendations for implementation. *Library Hi Tech, 20*(4), 406-419.

KEY TERMS

Access Board Standards: Technical and functional performance criteria developed by the Architectural and Transformation Barriers Compliance Board (the "Access Board"), a U.S. government agency, under Section 508. Only electronic and information technology conforming to these standards is considered accessible.

Accessibility: As defined within Section 508, accessibility is achieved when individuals with disabilities can access and use information technology in ways comparable to those available to people without disabilities. A narrower, operational definition conceptualizes accessibility in terms of conformance to certain accessibility criteria such as the *web content accessibility guidelines* or the *access board standards*.

Accessible Web Design: Also sometimes referred to as "barrier-free Web design." Web design that strives to accommodate the needs of people with disabilities, including those using assistive technology, to access the Web environment.

Americans with Disabilities Act (ADA): U.S. civil rights legislation passed in 1990 that prohibits discrimination against people with disabilities in the areas of employment, transportation, telecommunications, and public accommodation.

Assistive Technology: Specialized software or hardware, such as screen readers, magnification software, and a modified keyboard, used by some people with disabilities to interact with the computer.

Audio Browser: Also referred to as "talking browser." Software that interprets the html code of Web pages and provides speech output for text-based components, along with information provided by the html mark-up tags. Typically, it also enables users to navigate the Web page through alternative keystrokes.

Print Disabilities: Comprises all those disabilities that make it difficult, or impossible, to read printed text. The term includes visual impairment and blindness; cognitive disabilities, such as dyslexia; and certain motor-control impairments.

Section 508: A provision within the Rehabilitation Act of 1973, as amended by Congress in 1998, that mandates that the electronic and information technology developed, maintained, procured or used by the U.S. government must be accessible to people with disabilities.

Screen Reader: Software that interprets the signals sent to the computer screen and reads aloud the displayed text with the help of a speech synthesizer.

Universal Design: A concept similar to accessible design. Its meaning is broader in that it refers to design that strives to create products that are usable by all people, regardless of age, gender, (dis)ability, handedness, etc. Its meaning is narrower in that it seeks one solution to accommodate the needs of all people.

Web Content Accessibility Guidelines (WCAG): Guidelines for accessible Web design developed by the World Wide Web Consortium's Web Accessibility Initiative. WCAG 1.0 were passed in 1999. A working draft of a revised set of guidelines, WCAG 2.0, is currently under review.

Actionable Knowledge Discovery

Longbing Cao
University of Technology Sydney, Australia

INTRODUCTION

Actionable knowledge discovery is selected as one of the greatest challenges (Ankerst, 2002; Fayyad, Shapiro, & Uthurusamy, 2003) of next-generation knowledge discovery in database (KDD) studies (Han & Kamber, 2006). In the existing data mining, often mined patterns are nonactionable to real user needs. To enhance knowledge actionability, domain-related social intelligence is substantially essential (Cao et al., 2006b). The involvement of domain-related social intelligence into data mining leads to *domain-driven data mining* (Cao & Zhang, 2006a, 2007a), which complements traditional data-centered mining methodology. Domain-related social intelligence consists of intelligence of human, domain, environment, society and cyberspace, which complements data intelligence. The extension of KDD toward domain-driven data mining involves many challenging but promising research and development issues in KDD. Studies in regard to these issues may promote *the paradigm shift of KDD from data-centered interesting pattern mining to domain-driven actionable knowledge discovery*, and *the deployment shift from simulated data set-based to real-life data and business environment-oriented* as widely predicted.

BACKGROUND

In the last decades, data mining, or KDD, has become a prominent, exciting research and development area in the field of information technology. Data-centered data mining has experienced rapid development in various aspects such as data mined, knowledge discovered, techniques developed, and applications involved. Table 1 illustrates such key research and development progress in KDD.

A typical feature of data-centered data mining is that KDD is presumed to be an automated process of identifying interesting hidden patterns in public data sets. It targets the production of automatic algorithms as well as methods that extract patterns of certain technical significance. As a result, algorithms and the tools developed lack the capability to adapt to real-life environmental constraints and dynamics. Thousands of patterns and algorithms have been published in academia, but unfortunately very few of them have been transferred into real business use.

Increasing numbers of KDD researchers and developers have realized the limitation of traditional data mining methodologies (Ankerst, 2002; Fayyad et al., 2003), noted the gap between business and academic interests (Gurali & Wallace, 1997). The research on the challenges of KDD,

Table 1. An overview of data mining

Dimension	Key research progress
Data mined	• Relational, transactional, object-relational, active, temporal, spatial, time-series, heterogeneous, legacy, Web, and so forth. • Stream, spatiotemporal, multimedia, ontology, event, activity, link, graph, text, sensor, and so forth.
Techniques studied	• Database, machine learning, or statistics-oriented, say Neural Network, Bayesian network, Support Vector Machine, Rough Set, and so forth. • Association, frequent pattern analysis, multidimensional and OLAP analysis methods, classification, cluster analysis, outlier detection, visualization, and so forth. • Scalable data mining, stream data mining, spatiotemporal data mining, multimedia data mining, biological data mining, text and Web mining, privacy-preserving data mining, event mining, link mining, ontology mining, granule mining, and so forth.
Knowledge discovered	• Characters, associations, classes, clusters, discrimination, trends, deviation, outliers, exceptions and so forth.
Application involved	• Engineering, retail market, telecommunication, banking, fraud detection, intrusion detection, stock market, social security, bio-informatics, defense, Web services, biological, social network analysis, intelligence and security, and so forth. • Enterprise data mining, cross-organization mining, online mining, dynamic mining, and so forth.

plus trustworthy and workable KDD methodologies and techniques have therefore become a significant and productive direction of KDD research. In the panel discussions of SIGKDD 2002 and 2003 (Ankerst, 2002; Fayyad et al., 2003), a couple of important challenges for extant and future data mining were identified. Among them, actionable knowledge discovery is viewed as one of the key foci, because it not only provides an important tool to business decision makers for performing appropriate actions, but also delivers reliable and actionable outcomes to businesses. However, it is not a simple task to extract actionable knowledge utilizing traditional KDD methodologies. This situation results partly from the assumption that traditional data mining is a data-centered trial-and-error process (Ankerst, 2002), data and technical interestingness have been taken as two of the major targets and criteria in algorithm and pattern development.

To bridge the gap between business and academia, it is important to understand the difference between objectives and result evaluation of data mining in research and real-world applications. In the business world, KDD must answer the question "What Makes Knowledge Identified Interesting to Businesses" (Silberschatz & Tuzhilin, 1996) not only from the technical angle but also from a business perspective. Real-world data mining needs to identify patterns in constrained environments and satisfy not only technical significance (Freitas, 1998; Hilderman & Hamilton, 2000; Omiecinski, 2003; Padmanabhan & Tuzhilin, 1998), but business expectations (Ghani & Soares, 2006; Cao & Zhang, 2007a; Kleinberg, Papadimitriou, & Raghavan, 1998). The difference referred to above is exemplified through several key aspects (Cao & Zhang 2007a), for example, KDD problem, context, patterns, mining processes, objective and subjective interestingness, business expection, balancing multiple objectives, and infrastructure supporting business-oriented mining.

To deal with this difference, experience and lessons learned in real-world data mining (Cao & Zhang 2006a, 2007a, 2007c, 2008a, 2008b) show the significance of involving domain-specific data, domain, human and cyberspace intelligence (Cao et al., 2006b). For instance, domain-related social intelligence may consist of the involvement of domain knowledge (Yoon, Henscen, Park, Makki, 1999) and experts (Cao & Dai, 2003; Han & Kamber, 2006), the consideration of constraints (Boulicaut & Jeudy, 2005), and the development of in-depth patterns (Lin & Cao, 2006). It is essential to filter subtle concerns while capturing incisive issues. Through the thorough scrutiny of domain-specific intelligence and correct involvement of domain-specific intelligence into KDD, a streamlined data mining methodology emerges to discover the hidden core of a problem. These form the grounds of *domain-driven data mining* for next-generation KDD.

KNOWLEDGE ACTIONABILITY

In order to reflect business concerns, a *two-way significance framework* (Luo, Cao, Ni, & Liu, 2007) is proposed for measuring knowledge actionability. The two-way significance framework presents a straightforward nevertheless important definition of knowledge actionability. This means highlighting the involvement of business expectations (Tzacheva & Ras, 2005; Wang, Zhou, & Han, 2002) into traditional technical-only significance scheme (Yang, Yin, Lin, & Chen, 2003). In addition, the two-way significance is reflected in terms of both objective (Freitas, 1998; Hilderman & Hamilton, 2000), subjective (Liu, Hsu, Chen, & Ma, 2000), and multi-objective (Freitas, 2004; Tuzhilin, 2002) perspectives. As a result, actionable knowledge identified is not only based on a solid technical foundation, for instance, of recognized statistical significance, but enables business users to take appropriate actions which will be to their advantage.

DEFINITION 1. (*Knowledge Actionability*) *Let x be an itemset in dataset X, given a mined pattern p associated with x, actionable capability x.act(p) is described as the satisfaction of both technical interestingness x.tech_int*(*p*) *and business expectation x.biz_int*(*p*).

$$\forall x \in X, \exists p : x.tech\_int(p) \wedge x.biz\_int(p) \rightarrow x.act(p) \quad (1)$$

Further, knowledge actionability is instantiated in terms of objective (*_obj*()) and subjective (*_sub*()) perspectives from both technical and business sides:

$$\forall x \in X, \exists p : x.tech\_obj(p) \wedge x.tech\_subj(p) \wedge biz\_obj(p) \wedge biz\_sub(p) \rightarrow x.act(p) \quad (2)$$

However, it is not rare to discover that incompatibility and uncertainty exist in bridging the gap between business and academia in real-world data mining. To solve these issues, we propose (Luo, et al., 2007) fuzzy aggregation of business expectation and technical significance, generating a fuzzy ranking of patterns reflecting and balancing both technical and business concerns. As demonstrated in mining actionable trading patterns in market order book data (Lin & Cao, 2006), this approach presents promising options for resolving the issue of business needs, as well as bridging the gap between the business and technical side.

KEY COMPONENTS

Based on our experience and lessons learned in developing domain-driven data mining for real-world applications, including the discovery of actionable trading patterns in stock markets (Lin & Cao, 2006) and activity patterns

in government social security services (Cao, Zhao, & Zhang, 2007b), we can summarize the following key KDD components. (1) Problem understanding and definition is domain-specific and must involve domain-related social intelligence. (2) The context of actionable data mining involves data, domain, human, society or cyberspace, which is constrained, open and dynamic. (3) In-depth data intelligence and patterns need to be discovered. (4) A loop-closed refinement process is essential. (5) The evaluation of pattern actionability should balance statistical significance and business expectation. (6) A human-mining-cooperated infrastructure may be necessary for involving domain-related social intelligence. (7) Effective and efficient system support such as parallel KDD support plays an important role. (8) Knowledge discovered should be reliable, trustworthy and actionable to support action-taking in business decision-making. We believe studies on these points are useful for actionable knowledge discovery.

In domain-driven framework, data mining analysts and domain-specific business analysts complement each other with regard to in-depth granularity and constrained environment through interactive system support (Aggarwal, 2002). The involvement of domain experts and their knowledge can assist in developing highly effective domain-specific data mining techniques, and reduce the complexity of knowledge discovery and production process in a practical manner. In-depth pattern mining discovers deep-rooted interesting patterns that are beyond normal expectation. The patterns mined should not only be based on solid technical foundation, but satisfy real user needs. They can be used for specific users to take action according to their various criteria. A system following this framework is a loop-closed system, which reflects and appreciates feedback from the process as well as domain experts. It can refine the life cycle of actionable knowledge discovery in an iterative manner.

Domain driven data mining can greatly complement and tackle issues in traditional data-centered KDD methodologies in terms of the full process of KDD. The involvement of domain-related social intelligence into KDD process not only strengthens technical development and performance, but highlights business expectations and actionable capability of the identified results. Table 2 summarizes such supplementation and expansion in addition to traditional data-centered methodologies.

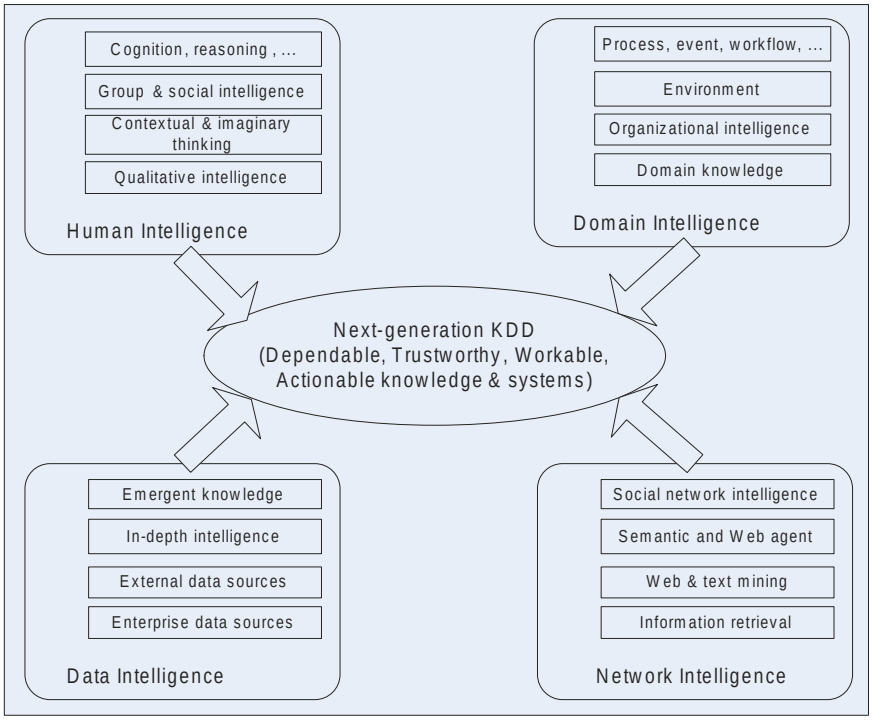
Intelligence Metasynthesis

Compared with traditional data-centered data mining methodology, a key involvement and enhancement in domain-driven data mining is *intelligence meta-synthesis* (Cao & Zhang, 2007b) of findings and progress available from the above tasks. In a high level perspective, intelligence meta-synthesis synthesizes data intelligence with domain-oriented social intelligence, including domain intelligence, human intelligence and cyberspace intelligence where appropriate. Domain-driven data mining therefore is a process full of interaction and integration among multiple kinds of intelligence, as well as intelligence emergence toward actionable knowledge discovery. Figure 1 outlines basic units and their interaction in intelligence meta-synthesis for actionable knowledge discovery.

Table 2. KDD supplementation and expansion via domain driven data mining

Aspects	Data-centered data mining	Domain-driven data mining
Object mined	Data tells the story	Data and domain-related social intelligence tell the story
Aim	Developing innovative approaches	Generating reliable and trustworthy knowledge supporting business decision-making actions
Objective	Algorithms are the core	Systems are the target
Dataset	Mining abstract and refined public data set	Mining constrained real life data
Extendibility	Usually predefined models and methods	Dynamic and personalized model use
Process	An automated process	Human-mining-interactive process
Evaluation	Strong technical significance	Strong actionable capability with solid technical foundation
Accuracy	Accurate and solid theoretical foundation	Data mining is a type of artwork
Goal	Let data create/verify research innovation; Demonstrate and push the use of novel algorithms discovering knowledge of interest to research	Let data and domain-related social intelligence tell the hidden story in business; Discover reliable knowledge that will support users to take actions to their advantage

Figure 1. Intelligence metasynthesis in actionable knowledge discovery



FUTURE TRENDS

To promote domain driven actionable knowledge discovery, many challenging issues in both theoretical and application aspects need to be studied and experimented in real world scenarios. Theoretically, there are many directions to go, for instance, what is domain-driven data mining methodology and process? How to model and involve domain knowledge into KDD? How to support human-data mining system interaction? How to mine in-depth patterns? How to adapt to dynamic mining in high frequency and high density data? What is generic business interestingness framework? How to tackle incompatibility and uncertainty in combining technical significance and business expectation? How to develop and synthesize domain intelligence, human intelligence and cyberspace intelligence? In practice, to construct actionable knowledge discovery, many practical problems need to be studied, for example, what would a domain-driven data mining tool look like? What would be an appropriate project management methodology? How to manage knowledge from multiple sources? How to support human involvement in decision making? How to balance performance and computational complexity? How to present data and knowledge satisfying business users' request? How to produce trustworthy and reliable knowledge? All the above theoretical and practical issues are worthy of further research and development in real-world scenarios.

CONCLUSION

The retrospect of traditional data-centered mining methodology has disclosed the significance of developing next-generation KDD methodologies and system support targeting actionable knowledge discovery. To this end, a significant emphasis is on involving and synthesizing domain-specific social intelligence, for instance, domain intelligence, human intelligence and cyberspace intelligence, into data intelligence during the KDD process. It adequately utilizes intelligence including domain expertise, knowledge, constraints, environment, business rules and processes, and human cooperation surrounding the problem studied. This benefits in-depth and actionable pattern mining satisfying both technical significance and business expectation. The research promotes a trend toward domain-driven data mining on top of data-centered methodology.

Domain-driven data mining provides complementary support to traditional data-centered data mining. Real-world experiments and experience have shown that domain-driven data mining has potential to strengthen traditional KDD, where a great amount of information was mined but little was of value or interest to real life business needs. With increasing theoretical and practical research and development in real-world applications, domain-driven data mining can promote the KDD paradigm shift from data-centered hidden pattern mining to domain-driven actionable knowledge discovery.

It can also benefit the deployment shift from abstract and artificial data modeling to real-life data and business environment based development as widely appreciated by business users and data mining researchers.

REFERENCES

Aggarwal, C. (2002). Towards effective and interpretable data mining by visual interaction. *ACM SIGKDD Explorations Newsletter, 3*(2), 11-22.

Ankerst, M. (2002). Report on the SIGKDD-2002 panel the perfect data mining tool: Interactive or automated? *ACM SIGKDD Explorations Newsletter, 4*(2),110-111.

Boulicaut, J.F., & Jeudy, B. (2005). Constraint-based data mining. In O. Maimon & L. Rokach (Ed.),*The data mining and knowledge discovery handbook* (pp. 399-416). Springer-Verlag.

Cao, L. (2008a). Developing actionable trading strategies. Intelligent agents in the evolution of WEB and applications. Springer (to appear).

Cao, L., He, T. (2008b). Developing actionable trading agents. *Knowledge and information systems: An international journal* (to appear).

Cao, L.B., & Dai, R.W. (2003). Human-computer cooperated intelligent information system based on multi-Agents. *ACTA AUTOMATICA SINICA, 29*(1), 86-94.

Cao, L.B., & Zhang, C.Q. (2006a). Domain-driven data mining: A practical methodology. *International Journal of Data Warehousing and Mining, 2*(4), 49-65.

Cao, L.B. et al. (2006b). *Intelligence metasynthesis in building business intelligence systems*, WImBI2006, Springer-Verlag.

Cao, L.B., & Zhang, C.Q. (2007a). The evolution of KDD: Towards domain-driven data mining. *International Journal of Pattern Recognition and Artificial Intelligence* (to appear).

Cao, L., Zhao, Y., & Zhang, C. (2007b). Mining impact-targeted activity patterns in imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*.

Cao, L. (2007c). Domain-driven actionable knowledge discovery. *IEEE Intelligent Systems, 22*(4), 78-89.

Fayyad, U., Shapiro, G., & Uthurusamy, R. (2003). Summary from the KDD-03 panel: Data mining: The next 10 years. *ACM SIGKDD Explorations Newsletter, 5*(2), 191-196.

Freitas, A.A. (1998). *On objective measures of rule surprisingness*, PKDD98 (pp. 1-9).

Freitas, A.A. (2004). Critical review of multi-objective optimization in data mining—a position paper. *SIGKDD Explorations, 6*(2), 77-86.

Ghani, R., & Soares, C. (2006). In *Proceedings of the Workshop on Data Mining for Business Applications: Joint in KDD2006*.

Gur Ali, O.F., & Wallace, W.A. (1997). Bridging the gap between business objectives and parameters of data mining algorithms. *Decision Support Systems, 21*, 3-15.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann.

Hilderman, R.J., & Hamilton, H.J. (2000). *Applying objective interestingness measures in data mining systems*, PKDD00 (pp. 432-439).

Kleinberg, J., Papadimitriou, C., & Raghavan, P. (1998). A microeconomic view of data mining. *Journal of Data Mining and Knowledge Discovery.*

Lin, L., & Cao, L.B. (2007). Mining in-depth patterns in stock market. *International Journal on Intelligent System Technologies and Applications* (to appear).

Liu, B., Hsu, W., Chen, S., & Ma, Y. (2000). Analyzing subjective interestingness of association rules. *IEEE Intelligent Systems, 15*(5), 47-55.

Luo, D., Cao, L.B., Ni, J.R, & Liu, L. (2007). Towards business interestingness in actionable knowledge discovery. In *Proceedings of the PAKDD 07 Workshop on Data Mining for Business*. Springer-Verlag.

Omiecinski, E. (2003). Alternative interest measures for mining associations. *IEEE Transactions on Knowledge and Data Engineering, 15*, 57-69.

Padmanabhan, B., & Tuzhilin, A. (1998). *A belief-driven method for discovering unexpected patterns*, KDD-98 (pp. 94-100).

Silberschatz, A., & Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems? *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 970-974.

Tuzhilin, A. (2002). Knowledge evaluation: Other evaluations: usefulness, novelty, and integration of interesting news measures. *Handbook of data mining and knowledge discovery* (496-508).

Tzacheva, A., & Ras, W. (2005). Action rules mining. *International Journal of Intelligent Systems, 20*, 719-736.

Wang, K., Zhou, S., & Han, J. (2002). *Profit mining: From patterns to actions*. EBDT2002.

Yang, Q., Yin, J., Lin, C., & Chen, T. (2003). Postprocessing decision trees to extract actionable knowledge. In *Proceedings of the ICDM2003*.

Yoon, S., Henschen, L., Park, E., & Makki, S. (1999). Using domain knowledge in knowledge discovery. In *Proceedings of the Eighth International Conference on Information and Knowledge Management*. ACM Press.

KEY TERMS

Actionability: The business-oriented capability of discovered knowledge offering evidence for appropriate action-taking in business decision-making.

Business Decision Making: Decisions made by business people aim to the achievement of business objectives in a manner of satisfying business needs and expectations.

Business Expectation: A pattern of interest to business needs satisfies business performance requests from aspects such as social, economic and psychoanalytic concerns.

Domain Driven Data Mining: Data mining methodologies and techniques that utilize domain-oriented social intelligence, target dependable, trustworthy and actionable knowledge for business decision making.

In-Depth Mining: Mining patterns that disclose deep hidden information and relationship of attributes, which can assist deeper understanding of data, business and decision-making.

Interestingness: Measuring the performance of discovered knowledge in terms of statistical significance, user preference, and business expectation from objective and subjective perspectives.

Intelligence Metasynthesis: The interaction and integration between multiple types of intelligence surrounding a problem solving process, including human intelligence, data intelligence, domain intelligence and cyberspace intelligence. Advanced intelligence emerges through intelligence interaction and metasynthesis.

Social Intelligence: Intelligence hidden or reflected in aspects of data, domain, human, society and cyberspace where appropriate, for instance, domain-specific background information and knowledge, expertise, expert involvement, constraints, environment, business rules and processes.

Statistical Significance: A pattern mined is statistically significant based on statistical metrics that measure and evaluate the performance of an identified pattern.

Active Patient Role in Recording Health Data

Josipa Kern
University of Zagreb, School of Medicine, Andrija Stampar School of Public Health, Croatia

Kristina Fister
University of Zagreb, School of Medicine, Andrija Stampar School of Public Health, Croatia

Ozren Polasek
University of Zagreb, School of Medicine, Andrija Stampar School of Public Health, Croatia

INTRODUCTION

The healing process can be viewed as a partnership between doctors and patients, nurses and physicians or, more generally, a partnership of health professionals and health care users (Anonymous, 2008, Graham, 2007). A patient-centered approach that empowers patients to participate in decisions about their treatment and health care options asks for active participation of patients themselves, specifically, in health information gathering and exchange of this information with their health or medical records (Bachman, 2007; Stolyar, Lober, Drozd, & Sibley, 2005).

BACKGROUND

Medical or Health Record

Every physician has a number of patients in his or her care. Many patients also have a number of specialists taking care of their health. It is almost impossible for physicians to keep in mind all the information about even a single patient, let alone all patients in their care. Similarly, patients need to remember and comply with many recommendations communicated to them by their doctors. Recording patients' data is, today, a necessity, especially considering a large number of available diagnostic procedures and instruments producing information relevant for making medical decisions. One implication of such recording is the creation of medical records in health institutions; they are created and accessed by health professionals. According to the National Library of Medicine, the MeSH (medical subject heading) term "medical record" considers "recording of pertinent information concerning patient's illness or illnesses" (http://www.PubMed.com).

However, the medical data gathered by health professionals are not enough for making good medical decisions. Information that is not strictly medical can be added to medical data. We therefore usually talk about a *health record*, consisting of data and information that affect or could affect the patient's health status, or simply describe it. A health record is a more general term than a medical record, nursing record, or dental record, and should be used as an immediate superior term to them. Keeping all the information pertaining to a particular patient in one place, and making it accessible at any time to authorized professionals, is a challenge. In seeking solutions, the information and communication technology should be consulted.

Personal Health Record

In trying to encourage people to take an active interest in their own health, patients are supported to manage their own personal health records. A personal health record can contain copies of data from the health record, which is created by health professionals, and also information entered by patients themselves (for example, subjective information such as description of symptoms, and objective information such as values of self-measured blood pressure or blood glucose levels, etc., recorded in a personal health diary).

Thus far, the literature does not give an adequate definition of a personal health record. Wikipedia defines it as "a health record that is initiated and maintained by an individual," but it is unclear who the individual is, the health professional or the patient. According to Tang (2006), a personal health record includes health information managed by the *individual*, who is not necessarily a *patient,* an ill person. This distinction emphasizes that the personal health record is a tool used to care for health and wellness, not only illness.

CURRENT STATUS

Electronic Health Record and Personal Health Record

There are several definitions of the electronic health record and many descriptions of its characteristics and demands (Hayrinnen, 2007). According to ISO (2004), "the EHR means a repository of patient data in digital form, stored and

exchanged securely, and accessible by multiple authorized users. It contains retrospective, concurrent, and prospective information and its primary purpose is to support continuing, efficient and quality integrated health care." One of the most exhaustive descriptions of electronic health records is given by the Advisory Committee on Health Infostructure of Canada (2001). According to the description in their Tactical Plan for a pan-Canadian Health Infostructure, an electronic health record is "a longitudinal collection of personal health information of a single individual, entered or accepted by health care providers, and stored electronically. The record may be made available at any time to providers, who have been authorized by the individual, as a tool in the provision of health care service. The individual has access to the record and can request changes to its content. The transmission and storage of the record is under strict security." This means that an electronic *health* record also incorporates electronic *medical* records, including digital medical images (computer tomography or similar) and biomedical signals (electrocardiography or similar), laboratory findings, the interpretation of all such findings, and physicians' recommendations to patients. Hospital records, nursing records, dental records, and other similar records can also be parts of an electronic health record.

It is generally agreed that patients have the right to know who is collecting, storing, accessing, communicating, or processing the data in their electronic health records, for what purpose, where the data will be kept, to whom they will be communicated, and for what purpose (Kluge, 2004).

The definition of electronic personal health records as "electronic summaries of a patient's medical record that are often portable and easily accessed by the patient" (Endsley, Kibbe, Linares, & Colorafi, 2006) is not adequate, for it does not distinguish from electronic medical records. A better description of an electronic personal health record was given as "an electronic application through which individuals can access, manage and share their health information" (Pagliari, Detmer, & Singleton, 2007). We propose the most appropriate description of an electronic personal health record would be "a digitally (electronically) saved health information, created and accessible by both health professionals and the individual, respecting privacy, security and confidentiality."

Content of the Electronic Health Record

Today, electronic health records are used in many hospitals, primary care offices, and other health institutions. Each of these sites collects specific patient data particularly relevant to the type of health care received at the site, but all the data could, and should, be mutually communicated. Pulling together health data from different sources should help doctors make better diagnostic and therapeutic decisions. However, pulling information from several different sources will require a unique identifier (Mayor, 2007). This is the first demand on electronic health records.

Patient Identifier

Health systems of different countries define patient identifiers in different ways. Sometimes patients are identified by their health insurance number, other times by the social security number, and sometimes by biometric characteristics of a patient. As a general rule, all of these identifiers are unique (i.e., any two patients have different identifiers) and should have a check digit calculated by a defined algorithm. Table 1 shows an example of creating such an identifier.

Health and Medical Data

Structured electronic medical records can result in quicker data entry, improved quality of care, and improved usefulness

Table 1. Creating a patient identifier (algorithm: module 11)

Starting with: 10000001
Add ponders to each digit: 7 6 5 4 3 2 7 6
Calculate: 7*1+6*0+5*0+4*0+3*0+2*0+7*0*6*1=13
Calculate: 13 : 11 = 1, remnant = 2*
Calculate: 11 – 2 = 9
Conclusion: 9 is a check digit and the created identification number is: 100000019

---

* in case of remnant = 0 the 8 digit number should be omitted; in case of remnant = 1 the check digit shoudl be 0

Following the same algorithm the next identifiers will be:
100000024
100000038
100000043
100000057
....

of records for daily clinical practice (Kruger, 2007). Doctors and nurses prefer structured data entry; electronic nursing records are better, and databases with structured electronic patient records can be used on a large scale to develop treatment regimes and support quality assurance (Kruger, 2007). Integration of clinical decision support systems into electronic medical records holds a promise for more efficacious and cost-effective solutions in daily medical practice. However, designing and structuring the electronic health (or medical) record is a daunting task.

The starting point of the electronic health record should be *patient history*. Patient history is produced during a conversation between the patient and the health care professional. It is not a final product made by, for example, a general practitioner at the first appointment, but it should be resummarised, recontextualised, and recreated in the light of new information, possibilities, and changing priorities during patient care. Some of these data can be structured, but some should also be written as free text.

Laboratory data are mostly structured. Some are measured and recorded as numeric values (e.g., blood glucose 5.3 mmol/L), while some laboratory data are qualitative and usually coded (e.g., blood groups: A, B, O, AB; proteins in urine: -, +, ++, +++).

Radiological data (medical images) and *biomedical signals* can be recorded digitally, and described (interpreted) by a radiologist or another appropriate specialist in the form of free text.

Diagnoses are coded according to an adequate coding system (International Classification of Diseases – ICD-10, SNOMED-CT, etc.), and *drugs* usually by the ATC classification.

Procedures are also coded, usually by using a specific coding system, depending on the country, health sector (hospitals, primary health care, etc.), and the purpose of such data. A diagnosis-related group (DRG) is based on the major diagnosis, length of stay, secondary diagnosis, surgical procedure, age, and types of services required. It is internationally accepted for acute cases in hospitals in order to determine the fixed payment per case (Hammond & Cimino, 2001), but it can be locally adjusted.

Patient *discharge letter* could be partially structured (parts including numeric and coded data), but it also includes some descriptions (free text).

Other Data

The electronic health record can contain other data such as personal data, socio-demographic and administrative data (e.g., name of the individual, date of birth, gender, education, insurance data, etc.).

Characteristics of an Electronic Health Record

Electronic health records should support the delivery of good quality patient care and improve decision making in all daily clinical situations. It should be standardized (e.g., EN 13606). In particular, an electronic health record should improve the management of chronic conditions and help maintain immunization records.

Specific characteristics of digitally recorded data are:

- They can be viewed from a variety of locations (general practice office, specialist practice, laboratory, hospital, etc.) at any moment, when needed,
- For viewing the data as a specific output (e.g., tables or graphs), a specific software is needed; viewing is not static and final, but is instead continuously modifiable,
- Users' authorization levels determine which parts of electronic health record they can access and which actions they can perform (security and protection),
- It is possible to trace who entered the data and who has seen the data, and when (any action can be recorded for the sake of patient safety and responsibility of the health professional),
- Periods of nonavailability of electronic health record data can be known (criterion for usability of the particular system).

Opposite to this, paper-based medical data can be viewed from only one location, they can be seen unconditionally, there is no possibility to specify which part of the document can be read (all or nothing), it is not possible to know who has seen the data and when (it is not recorded), and periods of nonavailability are unknown.

An electronic health record can be used not only for patient management, but also for reporting, quality assurance processes, research, and administrative processes, such as scheduling and billing. A standardised electronic health record is a prerequisite for good communication in the health care system (examples of standards relevant for electronic health record and communication are EN 13606, DICOM, LOINC and HL7).

Benefits of the Electronic Health Record

Electronic health records increase patients' confidence in their health care providers. They save patients and health care providers the repetition of various information (such as personal information – name, address, etc., previous course of the disease, test results) in different health institutions. At the same time, patients are assured that all health professionals caring for them have access to all the relevant parts of their health history. Also, online communication between

health care providers in different institutions speeds up health service. There is no need to wait for discharge letters or test results, and the number of lost test results is markedly reduced. Quality of care is improved as the electronic health record (and data in it) makes the communication with experts much easier. Consequently, another benefit of an electronic health record can be a reduction in the number of medical errors.

Content of the Electronic Personal Health Record

Individual patients have their own documents about their health, their own personal health records. These could be copies of medical documents in paper or in electronic format, for example, laboratory findings or x-ray images (Fig. 1), or even hospital discharge letters. These could also be data on drugs, herbal medications, diet, or results of home testing (blood pressure, medication, etc.) (Fig. 2). Some patients can analyze their health data (e.g., visualization of time series of blood pressure) (Fig. 3) and change their behavior accordingly. Personal health record can also contain various administrative data, such as the name of the person's physician or data related to health insurance.

Characteristics of the Electronic Personal Health Record

There is no standard for electronic personal health record, but there are many electronic applications offered for this purpose (MyPHR, iHealthRecord, etc).

Benefits of the Personal Health Record

Without doubt, the two leading benefits of personal health records are patients' empowerment and promotion of active partnership between patients and caregivers.

FUTURE TRENDS

Electronic health records have been successfully implemented in many settings. Still, they all need further development and improvement, usually linked to rapidly changing technologies. Other aspects of improvement deal with the contents and the way in which the data are recorded (such as classifications or coding systems).

Electronic personal health records are emerging and have much room for further growth. Simpler models of personal health records include passive access to data written on a compact disc or a smart card, or possibly at a Web site. Interactive systems of personal health records demand the integration of a personal health record with health providers' record systems (i.e., electronic health records). Ensuring the security of such systems remains a challenge. According to Pagliari (2007), views on values vs. risks of electronic records are highly polarized. Still, as all new initiatives take time to become fully developed and accepted by its users, this polarization is likely to be only temporary.

CONCLUSION

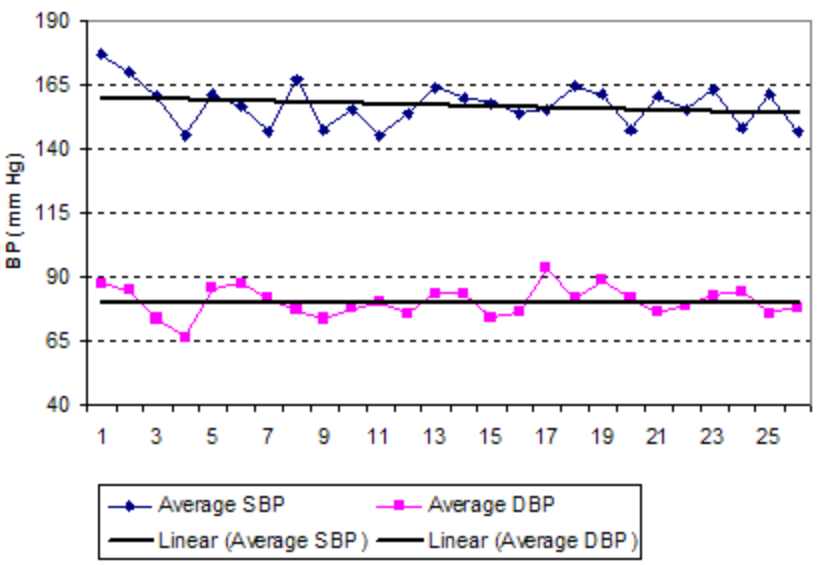
Electronic health records are the present and the future of contemporary health care systems. Electronic personal health records should be developed in accordance to the needs of all the participants in the health care (health care providers and health care users). It could be expected that the electronic personal health record and electronic health record will interact while not threatening privacy, security, and confidentiality. Both types of records are only parts of the same system, and both need more development and improve-

Figure 1. Example of the copy of a medical document

Figure 2. Example of a personal health record maintained by a person with hypertension

Date	Time	Therapy	Systolic BP	Diastolic BP	Puls rate	Average SBP	Average DBP
11.7.	19:00		180	93	80	177	88
			170	85	80		
			180	85	81		
	19:40		192	91	80	170	85
			163	81	81		
			154	83	81		
	20:07	Tinidil taken at 19:45	160	72	106	160	73
			157	74	106		
			164	74	106		
	20:45		145	66	106		
12.7.	5:40		167	85	78	161	86
			158	85	76		
			159	87	75		
	6:45	Hyzar	164	87	75	157	88
			153	88	73		
			153	88	73		
	7:35		147	82	75	147	82
			146	83	75		
			147	80	76		
	15:45		179	77	103	167	77
			163	77	101		
			159	77	98		
	19:00		159	76	87	147	74
			141	72	89		
			142	73	85		
13.7.	5:00	Hyzar	158	75	71	155	78
			152	80	69		
	6:17		156	78	78	145	80
			136	81	80		
			144	81	78		
	16:40		156	76	106	154	76
			153	77	103		
			153	74	98		
	21:00		172	82	73	164	84
			163	86	73		
			156	83	73		
14.7.	5:00	Hyzar	165	83	65	160	84
			159	85	61		
			155	83	65		
	7:00		171	74	72	158	74
			152	77	72		
			151	72	73		
	10:05	After physical activity	161	80	93	154	76

Figure 3. Example of visualization of blood pressure measurements as a part of an electronic personal health record



ment. The ultimate goal of high interactivity, accuracy, and safety can be assured by the information and communication technologies, but both health care professionals and health care users need to accept such a system and become partners in using it, in order to benefit from it.

REFERENCES

American Health Information Medical Association. (n.d.). *MyPHR personal health record. A guide to understanding and managing your personal health information*. Retrieved November 16th, 2007 from http://www.myphr.com/your_record/index.asp

Anonymous. (2007). Good patient-doctor communication is vital. The healing process can be viewed as a doctor-patient partnership, facilitated and enhanced by good interaction. *Health News, 13*(12), 3-4.

Bachman, J. (2007). Improving care with an automated patient history. *Family Practice Management, 14*(7), 39- 43.

Endsley, S., Kibbe, D. C., Linares, A., & Colorafi, K. (2006). An introduction to personal health record. *Family Practice Management, 13*(5), 57-62.

F/P/T Advisory Committee on Health Infostructure. (2001). *Tactical plan for a pan-Canadian health infostructure. Update*. Office of health and the information highway, Health Canada, November 2001. Retrieved September 11th, 2007, from http://www.hc-sc.gc.ca/hcs-sss/alt_formats/iacb-dgiac/pdf/pubs/2001-plan-tact/2001-plan-tact_e.pdf

Graham, I. W. (2007). Consultant nurse-consultant physician: A new partnership for patient-centred care? J Clin Nurs, *16*(10), 1809-17.

Hammond, W. E., & Cimino, J. J. (2001). Standards in medical informatics. In E. H. Shortliffe et al. (Eds), *Medical informatics. Computer applications in health care and biomedicine* (p. 226). New York: Springer.

Häyrinen, K., Saranto, K., & Nykänen, P (2007). Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *Int J Med Inform*. In press. doi:10.1016/j.ijmedinf.2007.09.001.

Interactive Health Record. (n.d.). *iHealthRecord*. Retrieved November 16th, 2007, from http://www.ihealthrecord.org/faq.html

ISO/DTR 20514. (2004). *Health informatics – Electronic health record – Definition, scope, and context.*

Kluge, E-H. W. (2004). Informed consent and the security of the electronic health record (EHR): Some policy considerations. *International journal of Medical Informatics, 73*(3), 229-34.

Krüger, K. (2007). Electronic medical records should be structured. *Tidsskr Nor Laegeforen, 127*(16), 2090-2093.

Mayor, S. (2007). NHS IT system must use unique patient identifiers to achieve research potential. *BMJ, 334*(7606), 1238.

Pagliari, C., Detmer, D., & Singleton, P. (2007). Potential of electronic personal health records. *BMJ, 335*(7615), 330-3.

Stolyar, A., Lober, W. B., Drozd, D. R., & Sibley, J. (2005). Feasibility of data exchange with a patient-centered health record. In *AMIA 2005 Symposium Proceedings* (p. 1123). Retrieved February 25th, 2008, from http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1560649&blobtype=pdf

Tang, P. C., Ash, J. S., Bates, D. W., Overhage, J. M., & Sands, D. Z. (2006). Personal health records: Definitions, benefits, and strategies for overcoming barriers to adoption. *Journal of the American Medical Informatics Association, 13*(2), 121-6.

Wikipedia. Personal health record. (n.d.). Retrieved October 16th, 2007, from http://en.wikipedia.org/wiki/Personal_health_record#_note-FIRST

KEY TERMS

Active Patient is an emerging term describing patients' active participation in the management of their health and wellness.

Electronic Health Record (EHR): A health record stored electronically.

Electronic Medical Record (EMR): A medical record stored electronically.

Electronic Personal Health Record (EPHR): A personal health record stored electronically.

Health Record (HR): A longitudinal collection of personal health information of a single individual, entered or accepted by health care providers.

Medical Record (MR): A longitudinal collection of personal data concerning patient's illness or illnesses.

Personal Health Record (PHR): A longitudinal collection of personal health information of a single individual containing copies of parts of health records, health related data measured and/or noticed by the individual, and administrative data.

Category: Software & Systems Design

Actor–Network Theory Applied to Information Systems Research

Arthur Tatnall
Victoria University, Australia

INTRODUCTION

Building an information system is a difficult task, partly due to the problem of ascertaining the requirements of the intended users, but also because of the complexity of the large number of human-machine interactions (Tatnall & Davey, 2005). This complexity is reflected in the difficulty of building these systems to operate free from error and to perform as intended. The dictionary defines innovation as "the alteration of what is established; something newly introduced" (Macquarie Library, 1981 p. 914). As the introduction or improvement of an information system in an organisation *necessarily* involves change, information systems research often involves research into technological innovation.

BACKGROUND: INFORMATION SYSTEMS AS A SOCIO-TECHNICAL DISCIPLINE

The discipline of information systems (IS) is concerned with the ways people build and use computer-based systems to produce useful information and so has to deal with issues involving both people and machines; with the multitude of human and non-human entities that comprise an information system (Tatnall, 2003). Information systems is neither merely a technical discipline nor a social one, but one that is truly socio-technical. Researchers in information systems face the problem of how to handle complexities due to interconnected combinations of computers, peripherals, procedures, operating systems, programming languages, software, data and many other inanimate objects; how they all relate to humans and human organisations, and how humans relate to them (Longenecker, Feinstein, Couger, Davis, & Gorgone, 1994).

This paper will outline a socio-technical approach, based on actor-network theory (ANT), to researching how people interact with and use information systems (Tatnall & Gilding, 1999; Tatnall 2003; Tatnall & Pliaskin, 2005). In actor-network theory the key is in using an approach that is neither purely social nor purely technical, but socio-technical.

Qualitative Research Traditions in Information Systems

Each field of academic inquiry is characterised by its own preferred and commonly used research approaches and traditions. In information systems research Myers (1997) outlines four qualitative traditions as being particularly significant: case study research, ethnography, grounded theory and action research.

Case study research is the most commonly used qualitative approach in information systems. As IS research topics commonly involve the study of organisational systems, a case study approach is often appropriate. Ethnography has grown in prominence as a suitable approach to information systems research after work such as that undertaken by Suchman (1987) and Zuboff (1988). It has been used especially in research where the emphasis is upon design, computer-supported cooperative work, studies of Internet and virtual communities, and information-related policies (Star, 1995). Grounded theory is an "an inductive, theory discovery methodology" (Martin & Turner, 1986) that seeks to develop theory that is grounded in data that is systematically gathered and analysed and involves "continuous interplay" between data and analysis (Myers, 1997). Orlikowski (1993) argues that in information systems research situations involving organisational change, a grounded theory approach can be useful as it allows a focus on "contextual and processual" elements as well as on the actions of key players.

Action research has been described as proceeding in a spiral of steps where each step consists of planning, action and evaluation of the result of the action. It is seen as aiming "... to contribute both to the practical concerns of people in an immediate problematic situation and to the goals of social science by joint collaboration within a mutually acceptable ethical framework." (Rapoport, 1970, p. 499). A variant of action research that is slowly gaining acceptance in information systems is soft systems methodology (SSM), developed by Peter Checkland and his colleagues (Checkland & Scholes, 1991). SSM attempts to give due recognition to both the human and technological aspects of a system. It acknowledges both human and non-human aspects of IS, but considers these to be entirely separate types of entities.

ANT AND SOCIO-TECHNICAL RESEARCH

Actor-network theory considers both social and technical determinism to be flawed and proposes instead a socio-technical account (Latour, 1996) in which nothing is purely social and nothing is purely technical (Law, 1991). ANT deals with the social-technical divide by denying that purely technical or purely social relations are possible.

To see better how this works, suppose that an IS researcher was investigating the uptake of a business-to-business eCommerce portal developed by a local government authority for use within a regional area, with an Internet service provider (ISP) and a software company engaged to build the portal, and a bank to provide a payment gateway (Pliaskin, 2004; Pliaskin & Tatnall, 2005). ANT asserts that the world is full of hybrid entities (Latour, 1991) containing both human and non-human elements and offers the notion of heterogeneity to describe projects such as this. The project will involve not just the entities mentioned above, but also non-human entities such as computers, computer programs, data storage devices, modems and telephone lines, and human entities including local business proprietors from small and large businesses, customers, programmers and local council staff. The utilisation of heterogeneous entities (Bijker, Hughes, & Pinch, 1987) then avoids questions of: "is it social?" or "is it technical?" as missing the point, which should be: "is this association stronger or weaker than that one?" (Latour, 1991).

Information systems researchers using an ANT approach would concentrate on issues of network formation, investigating the human and non-human alliances and networks built up by the actors involved. They would concentrate on the negotiations that allow the network to be configured by the enrolment of both human and non-human allies. Interactions and associations between actors and networks are all important, and actors are seen simply as the sum of their interactions with other actors and networks.

In the case of the portal an actor-network researcher would begin by identifying some of the important actors, starting perhaps with the local government portal project manager. An interview with the project manager would reveal why the project was instigated and identify some of the other actors. The main advice on method suggested by the proponents of actor-network theory is to "follow the actors" (Latour, 1996) and let them set the framework and limits of the study themselves, and one line of inquiry resulting from the interview with the project manager might be to approach the portal software designer and programmers. Another set of actors is the proprietors of the local businesses themselves, and the project manager may suggest some "business champions" to interview first. At least some of these business people might then point to the influence exerted by the computer hardware or software as a significant factor, so identifying some non-human actors. Negotiations between actors must be carefully investigated. Apart from the obvious human to human kind of negotiation, also included must be human to non-human interactions such as the business people trying to work out how the portal operates, and how to adapt this technology to be most suitable for their own business purposes. The process of adopting and implementing the portal can now be seen as the complex set of interactions that it is, and not just the inevitable result of the innate characteristics of this technology.

How Actor-Network Theory Handles Complexity

Longenecker et al. (1994) suggest that computer-based information systems should be regarded as complex socio-technical entities, begging the question of how this complexity should be handled. A common method of handling complexity in all subject areas lies in simplification, but the danger with simplification is that it runs the risk of removing just those things that constitute a useful description of the phenomenon under investigation by concealing the parts played by many of the actors (Suchman, 1987). The question here is which details to include and which to leave out, and who is to decide. In this respect, an appropriate research approach needs to ensure that complexities are not lost "in the process of labelling" (Law, 1991).

In actor-network theory the extent of a network is determined by actors that are able to make their presence *individually felt* by other actors. The definition of an actor requires this and means that, in practice, actors limit their associations to affect only a relatively small number of entities whose attributes are well defined within the network. An actor is not just a "point object" but an association of heterogeneous elements, themselves constituting a network. An actor can, however, in many ways also be considered as a "black box" (Callon, 1986), and when we open the lid of the box to look inside it will be seen to constitute a whole network of other, perhaps complex, associations. In many cases details of what constitutes an actor—details of its network—are a complication we can avoid having to deal with all the time.

When investigating the e-commerce portal it might be convenient, most of the time, to consider both the ISP and the portal software to constitute a black box. This would mean that this aspect of the technology could then be considered as just a single actor; the portal, and its interactions with other actors investigated on this basis. At other times it might be necessary to lift the lid of the black box and investigate the enclosed network of the ISP, telephone lines, computers, data storage, programmers, and interface designers it contains. The advantage of black-boxing though is that most of the time however, the portal can be regarded as just another actor. The important thing to note about the use of black-

boxing for simplification is that the complexity is not just put into the black box and lost as it is always possible, and indeed necessary, to periodically reopen the black box to investigate its contents.

The portal black box could be considered as shown in Figure 1 below. This black box itself also contains several other black boxes (portal software, portal hardware and payment gateway) that we do not need to consider in detail until such consideration becomes necessary. Figure 1 also shows some of the other main actors that interact with the portal. Until (and possibly after) the portal is fully operational, of course, these actors will also interact directly with each other.

Limitations and Criticisms of Actor-Network Theory

There are several main criticisms of actor-network theory. To begin, there is the criticism by Grint and Woolgar (1997) that it is not always sufficiently clear where the boundaries of a network lie or whose account of a network is to be taken as definitive. They note that the analyst's story seems to depend on a description of the "actual" network as if this was objectively available.

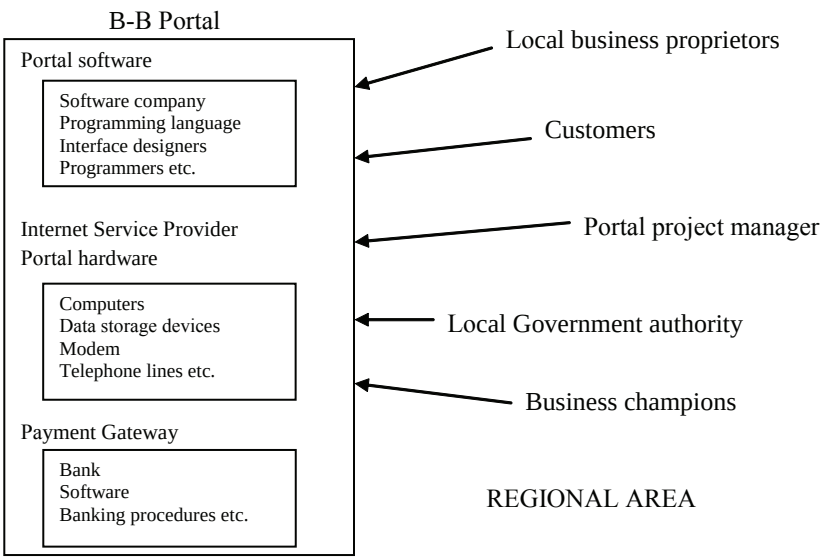
A second criticism relates to ANT's treatment of non-human actors. A critique by Collins and Yearley (1992) claims that in giving an autonomous voice to "things," ANT concedes too much to realist and technical accounts. In reply, Callon and Latour (1992) claim that technological artefacts are implicated in the very fabric of the social and are "social relations viewed in their durability and cohesion."

Thirdly, Grint and Woolgar (1997) argue that ANT retains a degree of residual technicism in its need to sometimes refer to "actual" technical capacities of a technology. They quote Callon's (1986) analysis of the attempts at building a French electric car, in which they claim that he makes reference to the "unfortunate tendency" of the catalysts to become quickly contaminated. They note that the anti-essentialist approach of actor-network theory would point to this 'actual property' being treated as a construction. Despite these minor reservations, however, Grint and Woolgar note that actor-network theory points to the possibility of an understanding of technology that does not rely on the presence of a "god within the machine."

FUTURE TRENDS

Actor-network theory offers a socio-technical approach to information systems research, and particularly to theorising technological innovation—a major aspect of IS research. It is useful in situations in which people and machines are intimately involved with each other, and this is exactly the case in most IS studies. A major use of ANT has been in the study of past events, and ANT makes no claim to be able to predict what may happen in the future. Nevertheless, ANT analysis can identify some pointers towards the successful introduction of an innovation, and the change management associate with this and so can be used to point to likely future scenarios and approaches that may be likely to succeed (Tatnall & Davey, 2004). ANT suggests that the key to successful change management in information systems involves allowing for these interactions and for the socio-technical nature of the process.

Figure 1. Actors, interactions and the portal black box



CONCLUSION

In this paper I have argued that information systems is a socio-technical discipline involving both human and non-human entities, and that information systems implementations are complex activities inevitably involving some form of technological innovation. I have also argued that simplistic views of how information systems are built, implemented and used often conceal important interactions between human and non-human actors and so give a less than complete picture of what has happened.

An actor-network approach avoids the need to consider the social and the technical, and thus human and non-human actors, in different ways. Highlighting how the human and non-human actors involved in socio-technical situations, such as the building and use of information systems, interact with each other is an important benefit of adopting an ANT research framework. In showing how these interactions may lead to the formations of stable networks, actor-network theory offers a useful way to handle the complexity of such studies.

REFERENCES

Bijker, W. E., Hughes, T. P., & Pinch, T. J. (1987). *The social construction of technological systems: New directions in the sociology and history of technology*. Cambridge, MA: MIT Press.

Callon, M. (1986). The sociology of an actor-network: The case of the electric vehicle. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology* (pp. 19-34). Macmillan Press.

Callon, M., & Latour, B. (1992). Don't throw the baby out with the bath school: A reply to Collins and Yearley. In A. Pickering (Ed.), *Science as practice and culture* (pp. 343-368). Chicago: Chicago University Press.

Checkland, P., & Scholes, J. (1991). *Soft systems methodology in action*. Chichester: Wiley.

Collins, H. M., & Yearley, S. (1992). Epistemological chicken. In A. Pickering (Ed.), *Science as practice and culture* (pp. 301-326). Chicago: Chicago University Press.

Grint, K., & Woolgar, S. (1997). *The machine at work—Technology, work, and organisation*. Cambridge: Polity Press.

Latour, B. (1996). *Aramis or the love of technology*. Cambridge, MA: Harvard University Press.

Latour, B. (1991). Technology is society made durable. In J. Law (Ed.), *A sociology of monsters. Essays on power, technology, and domination* (pp. 103-131). London: Routledge.

Law, J. (1991). *A sociology of monsters. Essays on power, technology, and domination*. London, Routledge.

Longenecker, H. E. J., Feinstein, D. L., Couger, J. D., Davis, G. G., & Gorgone, J. T. (1994). Information systems '95: A summary of the collaborative IS curriculum specification of the joint DPMA, ACM, AIS Task Force. *Journal of Information Systems Education, 6*(4), 174-186.

Macquarie Library. (1981). *The Macquarie Dictionary*. Sydney, Macquarie Library.

Martin, P. Y., & Turner, B. A. (1986). Grounded theory and organizational research. *The Journal of Applied Behavioral Science, 22*(2), 141-157.

Myers, M. D. (1997). *Qualitative research in information systems*. MIS Quarterly. Retrieved May 20, 1997, from http://misq.org/misqd961/isworld/

Orlikowski, W. J. (1993). CASE tools as organizational change: Investigating incremental and radical changes in systems development. *Management Information Systems Quarterly, 17*(3), 1-28.

Pliaskin, A. (2004). The life and times of BIZEWEST. Honours thesis. Information Systems. Victoria University, Melbourne.

Pliaskin, A., & Tatnall, A. (2005). Developing a portal to build a business community. In A. Tatnall (Ed.), *Web portals: The new gateways to Internet information and services* (pp. 335-348). Hershey, PA: Idea Group Publishing.

Rapoport, R. N. (1970). Three dilemmas in action research. *Human Relations, 23*(4), 449-513.

Star, S. L. (1995). *The cultures of computing*. Oxford: Blackwell Publishers.

Suchman, L. A. (1987). *Plans and situated actions. The problem of human-machine communication*. Cambridge: Cambridge University Press.

Tatnall, A. (2003). Actor-network theory as a socio-technical approach to information systems research. In S. Clarke, E. Coakes, M. G. Hunter, & A. Wenn (Eds.), *Socio-technical and human cognition elements of information systems* (pp. 266-283). Hershey, PA: Information Science Publishing.

Tatnall, A., & Davey, B. (2005). A new spider on the Web: Modelling the adoption of Web-based training. In P. Nicholson, J. B. Thompson, M. Ruohonen, & J. Multisilta (Eds.), *E-training practices for professional organizations* (pp. 307-314). Assinippi Park, MA: Kluwer Academic Publishers/IFIP.

Tatnall, A., & Davey, B. (2004). Improving the chances of getting your IT curriculum innovation successfully adopted

by the application of an ecological approach to innovation. *Informing Science, 7*(1), 87-103.

Tatnall, A., & Gilding, A. (1999). *Actor-network theory and information systems research*. The *10th Australasian Conference on Information Systems (ACIS)*, Wellington, Victoria University of Wellington.

Tatnall, A., & Pliaskin, A. (2005). *Technological innovation and the non-adoption of a B-B portal*. The *2nd International Conference on Innovations in Information Technology*, Dubai, UAE, UAE University.

Zuboff, S. (1988). *In the age of the smart machine*. New York, Basic Books.

KEY TERMS

Actor-Network Theory (ANT): an approach to research in which networks associations and interactions between actors (both human and non-human) and are the basis for investigation.

Actor: An entity that can make its presence individually felt by other actors. Actors can be human or non-human, non-human actors including such things as computer programs, portals, companies and other entities that cannot be seen as individual people. An actor can be seen as an association of heterogeneous elements that constitute a network. This is especially important with non-human actors as there are always some human aspects within the network.

Black Boxing: A technique used for simplification. Multiple actors can be put into a black box so that it is not necessary to look at them in detail. The portal mentioned in this paper could be considered as a black box containing the ISP, portal software, data storage devices, modems, telephone devices and so on. Black-boxing is done for convenience as it means that an entity can then be seen as just another actor, and saves looking at the detail until necessary. The black box can later be reopened to investigate its contents.

Socio-Technical Research: Involving both social and technical interactions, occurring is such a way that it is not easily possible to disentangle them.

Technological Innovation: The introduction or alteration of some form of technology (often information technology) into an organisation.

Category: Mobile & Wireless Computing

Adaptive Mobile Applications

Thomas Kunz
Carleton University, Canada

Abdulbaset Gaddah
Carleton University, Canada

INTRODUCTION

The convergence of two technological developments has made mobile computing a reality. In the last few years, developed countries spent large amounts of money to install and deploy wireless communication facilities. Originally aimed at telephone services (which still account for the majority of usage), the same infrastructure is increasingly used to transfer data. In parallel, wireless LAN technologies are providing hotspot coverage in many high-traffic locations. The second development is the continuing reduction in size of computer hardware, leading to portable computation devices such as laptops, palmtops, or functionally enhanced cell phones. Given current technology, a user can run a set of applications on a portable device and communicate over a variety of communication links, depending on his/her current location.

As will be explained in more detail later on, the mobile computing environment is highly dynamic. Available bandwidth changes by orders of magnitudes, based on the selected wireless access technology. Also, portable devices differ in processing power, memory, display capabilities, and other characteristics. It is generally argued that applications should "adapt" to the current environment, for example by filtering and compressing data or by changing the functionality offered to the user. Some researchers even argue that all future applications, not just the ones intended for execution on mobile devices, will have to be able to adapt to changing requirements and changing implementation environments on time scales from microseconds to years (Kavi, 1999). This article reviews the work on adaptive mobile applications and provides an outlook on future trends.

The alternative to adaptive applications is to either implement a single application that is designed for the lowest common denominator (in terms of resource availability) or multiple functionally identical or similar binaries, tuned for specific environments. The former will needlessly sacrifice application features when running in more resource-rich environments. The latter approach is an inferior solution as well, for a number of reasons. The user of a portable device has to install and maintain multiple applications, which is a drain on the limited storage capabilities typically found on those devices. It also potentially results in different user interfaces and causes high software development overheads when developing the "same" mobile application multiple times. Finally, it forces the user to identify the current execution conditions and select the "right" application.

The next section will review the motivation for adaptive approaches towards mobile application design. We will then briefly review traditional approaches to adaptive mobile applications, followed by a discussion of mobile middleware that is intended to support adaptive mobile applications. The article finishes with a brief conclusion of the state-of-the-art and identifies areas of future work.

BACKGROUND

Wireless communication and portable devices make it possible for mobile users to have access to information anywhere and anytime. Designing, implementing and deploying applications that work well across all portable devices and across a wide range of wireless access networks is non-trivial.

There are at least three common factors that affect the design of mobile applications: portable devices, network connection, and mobility. *Portable devices* vary from one to another in term of resource availability. Devices like laptops can offer fast CPUs and large amount of RAM and disk space while others like pocket PCs and phones usually have scarce resources. It is either impossible or too expensive to augment the resource availability. Hence, applications should be designed to achieve optimal resource utilization. In general, the design of portable devices strives for properties such as size, weight, durability and long battery life. Different devices will emphasize different trade-offs between CPU speed, memory, I/O capabilities, and power consumption, providing very heterogeneous execution environments for mobile applications.

Network connection in mobile scenarios is characterized by limited bandwidth, high error rate, higher cost, and frequent disconnections due to power limitations, available spectrum, and mobility. Wireless communication is more difficult to achieve than wired communication because the surrounding environment interacts with the signal, blocking signal paths and introducing noise and echoes. Therefore, mobile application designs need to be more concerned about

the network conditions than applications designed for fixed networks. Many wireless and mobile networks such as WaveLAN are organized into geographically defined cells, with a control point called a base station in each of the cells. Devices within the same cell share the network bandwidth; hence, the bandwidth rapidly decreases whenever a new device joins the cell. Portable devices may move around different areas with no coverage or high interference that cause a sudden drop in network bandwidth or a loss of connection entirely. Unpredictable disconnection is also a common issue that frequently occurs due to the handoff process or shadowed areas. Most wireless network services charge a flat fee for their service, which usually covers a fixed number of messages. Additional charges are levied on per packet or per message basis. In contrast, the cost for sending data over cellular networks is based on connection time instead. This forces mobile users to connect for short periods of time.

Physical device mobility can greatly affect network connection, which accordingly has to adapt to user mobility by reconnecting the user with respect to a new location. Portable devices may interact with different types of networks, services, and security policies as they move from one area to another. This requires applications to behave differently to cope with dynamic changes of the environment parameters. As a consequence, mobile applications also have to cope with a much greater variation in network bandwidth: bandwidth can shift from one to six orders of magnitude between being plugged into the wired network versus using (different) wireless access networks.

The constraints and limitations mobile applications face are not a product of current technology, but they are related naturally to mobility. Together, they complicate the design of mobile applications and require rethinking traditional approaches to application design. Any feasible approach to mobile computing must strike a balance between conflicting demands. This balance cannot be static, as the execution environment of mobile applications varies; it must react, or in other words, the applications must be adaptive.

ADAPTIVE MOBILE APPLICATIONS: TRADITIONAL APPROACHES

Designing adaptive applications is an active research area. Traditionally, most work focused on the wireless link(s). Early work provides general solutions that do not change the TCP semantics but focus on improving TCP throughput over wireless links; see for example Balakrishnan (1995). While this addresses issues such as high link error rates and spurious disconnections, it does not address the low and highly variable bandwidth characteristic of mobile computing.

A second group of approaches adapts to the scarce and varying wireless link bandwidth by filtering and compressing the data stream between a client application on a portable device and a server executing on a stationary host. Data compression is done at one of two places. Bolliger (1998) and Seshan (1997) enhance the server to generate a data stream that is suited for the currently available bandwidth. This typically represents an end-to-end approach, which is well known in the networking and system literature. Most other proposals (Angin, 1998; Fox, 1998) extend the client-server structure to a client-proxy-server structure, where a proxy executes in the wireless access network, close to the portable device. This proxy-based approach filters and compresses the data stream originating from the server to suit the current wireless bandwidth. Joshi (1997) incorporates both end-to-end and proxy-based approaches, using each as appropriate, to support Web access from mobile platforms. For example, tasks such as complex filtration, which require significant computational resources, are done in an end-to-end manner. The proxy-based approach, on the other hand, is used when the server is not able to generate the appropriate data stream.

A third, complementary approach, focuses on the computational effort (Kunz, 2000). Mobile applications, especially ones that require intensive computation (for example, video decoding), can be divided dynamically between the wired network and the portable device according to the mobile environment and to the availability of the resources on the portable device, the wireless link, and the wired network. The access network supports the mobile application by providing proxy servers that can execute parts of the application code. This may increase the performance of applications and reduce the power consumption on portable devices since offloading computation to the proxies in the wired network will reduce the CPU cycles and memory needed to achieve certain tasks at portable devices.

FUTURE TRENDS: MOBILE MIDDLEWARE

The early approaches reviewed in the previous section typically provide toolkits that support specific adaptation ideas. To generalize this effort, support for adaptive mobile applications should be embedded into appropriate mobile middleware. Traditional middleware systems, like CORBA and DCOM, have achieved great success in dealing with the heterogeneity in the underlying hardware and software platforms, offering portability, and facilitating development of distributed applications. However, these systems are based on the assumptions that the distributed applications will run in a static environment; hence, they fail to provide the appropriate support for mobile applications. Therefore, mobile applications need a middleware that facilitates adapting to environment variations.

Based on the mobile computing challenges reviewed previously, mobile middleware should meet the following requirements: *Asynchronous interaction* tackles the problems of high latency and disconnected operations. It allows mobile clients to issue a request for a service, disconnect from the network, and collect the result later on. This type of interaction model reduces bandwidth consumption, achieves decoupling of client and server, and elevates system scalability. *Reconfigurability* is the process of adding a new behavior or changing an existing one during system runtime. Dynamic reconfiguration of system behavior and operating context at runtime may require reevaluation and reallocation of resources. Reconfiguration could be based on context information. *Adaptivity* allows applications to modify their behavior instead of providing a uniform interface in all situations. The middleware needs to monitor the resource supply/demand, compute adaptation decisions, and notify applications about changes. *Context-awareness* of client capabilities, changes to network conditions, and the ability to change the behavior of the system as circumstances warrant are required to build an effective and efficient adaptive system. The context includes device characteristics, user's location, user's activities, and other resources. The system performance can be increased when information context is disclosed to the application to assist middleware in making the right decision. Finally, *lightweight load* should be considered when constructing middleware for mobile computing. Middleware should not increase the computational load on the most power-consuming components such as processor, network interface, and so forth. Middleware implementations often include a number of unused features that can be entirely omitted to reduce the computational load.

We identified four categories of mobile middleware: reflective, tuple space, context-aware, and event-based middleware. *Reflective middleware* like DynamicTAO (Kon, 2000) and Open-ORB (Blair, 2001) are built around the concept of component frameworks (CF). Components can be developed independently, distributed in binary form, and combined at run time. Reflection provides a *meta-interface* to inspect the internal behavior of the middleware and, if it is necessary, alter its behavior to better match the system's current operating environment. The main motivation of this approach is to make the middleware more adaptable to its environment and better able to cope with changes. Open problems are consistent dynamic reconfiguration and performance. There is some early work in this area that has focused on developing reconfiguration models and algorithms that enforce well-defined consistency rules while minimizing system disturbance (Kramer & Magee, 1990). In addition, all reflective systems impose a heavy computational load that causes significant performance degradation on portable devices.

Tuple-space systems such as LIME (Picco, 1999) and TSpaces (Wyckoff, 1998) exploit the decoupled nature of tuple spaces for supporting disconnected operations in a natural manner. A tuple space is a globally shared, associatively addressed memory space that is used by processes to communicate. Client processes create tuples and place them in the tuple space using a *write* operation. Also, they can concurrently access tuples using *read* or *take* operations. This communication paradigm fits well in a mobile setting where logical and physical mobility is involved. By default they offer an asynchronous interaction paradigm that appears to be more appropriate for dealing with intermittent connection of mobile devices, as is often the case when a server is not in reach or a mobile client requires to voluntary disconnect to save battery and bandwidth. Using a tuple-space approach, we can decouple the client and server components in time and space. In other words, they do not need to be connected at the same time and in the same place. Tuple-space systems support the concept of a space or spaces that offer the ability to join objects into appropriate spaces for ease of access. This opens up the possibility of constructing a dynamic super space environment to allow participating spaces to join or leave at arbitrary times. The ability to use multiple spaces will elevate the overall throughput of the system. One problem with tuple-space middleware systems is their excessive memory requirements, making them impractical for most portable devices available to-date.

Context-aware systems provide mobile applications with the necessary knowledge about the execution context in order to allow applications to adapt to dynamic changes in mobile host and network condition. The execution context includes but is not limited to: mobile user location, mobile device characteristics, network condition, and user activity (i.e., driving or sitting in a room). However, most context-aware applications are only focusing on a user's location. Nexus (Fritsch, 2000), for example, is designed to be a generic platform for location-aware applications. Reflective middleware may also improve the development of context-aware services and applications. For example, Capra (2003) has suggested the use of metadata and reflection to support context-aware applications. However, overall, limited attention has been given to contexts other than location. It is necessary to take into account other types of context awareness such as internal resources (i.e., memory size, battery and processor power) or external resources (i.e., network bandwidth and connectivity quality). More efforts need to be directed towards an easy context representation and simple interfaces that enable the applications to interact with the underlying middleware.

In *event-based systems*, clients first announce their interest in receiving specific events and then servers broadcast events to all interested clients. Hence, the event-based model achieves a highly decoupled system and many-to-many interaction style between clients and servers. Examples are JEDI (Cugalo, 2001) and STEAM (Meier, 2002). Most existing systems do not combine traditional middleware functionality (i.e., security, QoS, transactions, reliability, access control,

etc.) with the event-based paradigm. In addition, the developers are responsible for handling the low-level event transmission issues. Current publish/subscribe systems are restricted to certain application scenarios such as instant messaging and stock quote dissemination. This indicates that such systems are not designed as general middleware platforms. The majority of event-based middleware architectures are based on a logically centralized component called event dispatcher or broker. This component acts as a gateway between interacting components and hence has global knowledge about all the generated events and subscription requests. However, this centralized design often results in performance bottlenecks. Furthermore, not all event brokers provide a persistent buffer that can store all events for the provision of a reliable event service. There is also no support for the notion of composite events. Composite-event services allow clients to sign up with several event sources and receive event notifications in form of composite events. A special mechanism is needed to model event arrival from different sources and to specify composite events. This however may complicate the system architecture and incur extra cost.

CONCLUSION

Mobile computing is a relatively new field. While the challenges arising from mobility and the limitations of the portable devices are relatively well understood, there is no consensus yet as to what should be done to address these challenges. A comprehensive solution has to address many different aspects, such as the issue of dynamically changing bandwidth, the power, computational, and other limitations of the portable devices, or the varying availability of services in different environments. Traditional approaches to these challenges involved the design and implementation of proxies to either transparently intercept the data stream or to cooperate with the client application in the portable device in offloading computational tasks. To generalize this work, such services are expected to be embedded in middleware for mobile applications. Traditional middleware systems, such as CORBA and Java RMI, are based on the assumption that applications in distributed systems will run in a static environment; hence, they fail to provide the appropriate support for mobile applications. This gives a strong incentive to many researchers to develop modern middleware that supports and facilitates the implementation of mobile applications. To date, there is no single middleware that can fully support the requirements for mobile applications. Several solutions have considered one aspect or another; however, the door for further research is still wide open.

Table 1 provides a simple comparison of mobile middleware solutions with respect to the identified requirements for mobile computing. As can be seen, no single middleware paradigm covers all requirements. Based on our analysis, an evolution of reflective mobile middleware appears most promising, however. The key challenges are to reduce the complexity and size of such middleware for thin clients and to efficiently support asynchronous interaction styles.

REFERENCES

Angin, O. et al. (1998). The Mobiware toolkit: Programmable support for adaptive mobile networking. *IEEE Personal Communications, 5*(4), 32-43.

Balakrishnan, H. et al. (1995). Improving TCP/IP performance over wireless networks. *Proceedings of the 1st Annual International Conference on Mobile Computing and Communications*, Berkeley, CA, USA (pp. 2-11).

Blair, G.S. et al. (2001). The design and implementation of Open ORB 2. *IEEE Distributed Systems Online, 2*(6).

Bolliger, J., & Gross, T. (1998). A framework-based approach to the development of network-aware applications. *IEEE Transactions on Software Eng., 24*(5), 376-390.

Table 1. Requirements analysis for modern middleware

	Reflective	Tuple-space	Context-aware	Event-based
Asynchronous		✓		✓
Reconfigurability	✓			
Adaptivity	✓		✓	
Awareness	✓		✓	
Lightweight				✓

Capra, L., Emmerich, W., & Mascolo, C. (2003). CARISMA: Context-aware reflective middleware system for mobile applications. *IEEE Transactions on Software Engineering, 29*(10), 929–945.

Cugola, G., Nitto, E.D., & Fuggetta, A. (2001). The JEDI event-based infrastructure and its applications to the development of the OPSS WFMS. *IEEE Transactions on Software Engineering, 27*(9), 827–850.

Fox, A. et al. (1998). Adapting to network and client variation using infrastructure proxies: Lessons and perspectives. *IEEE Personal Communications, 5*(4), 10-19.

Fritsch, D., Klinec, D., & Volz, S. (2000). NEXUS positioning and data management concepts for location aware applications. *Proceedings of the 2nd International Symposium on Telegeoprocessing,* Nice-Sophia-Antipolis, France (pp. 171-184).

Joshi, A., Weerawarana, S., & Houstis, E. (1997). Disconnected browsing of distributed information. *Proc. Seventh IEEE Intl. Workshop on Research Issues in Data Engineering* (pp. 101-108).

Kavi, K., Browne, J.C., & Tripathi, A. (1999). Computer systems research: The pressure is on. *IEEE Computer, 32*(1), 30-39.

Kon, F. et al. (2000). Monitoring, security, and dynamic configuration with the dynamic TAO reflective ORB. *Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms and Open Distributed Processing (Middleware'2000),* Heidelberg, Germany (pp. 121–143).

Kramer, J., & Magee, J. (1990). The evolving philosophers problem: Dynamic change management. *IEEE Transactions on Software Engineering, 16*(11), 1293-1306.

Kunz, T., Omar, S., & Zhou, X. (2000). Mobile agents for adaptive mobile applications. *Networking and Information Systems Journal, 3*(5/6), 709-723.

Meier, R., & Cahill, V. (2002). STEAM: Event-based middleware for wireless ad hoc networks. *Proceedings of the International Workshop on Distributed Event-Based Systems (ICDCS/DEBS'02),* Vienna, Austria (pp. 639-644).

Picco, G., Murphy, A., & Roman, G.-C. (1999). LIME: Linda meets mobility. *Proceedings of the 21st Int. Conference on Software Engineering* (pp. 368–377).

Seshan, M., & Katz, R. (1997). Spand: Shared passive network performance discovery. *Proc. 1st Usenix Symposium on Internet Technologies and Systems.*

Wyckoff, P. et al. (1998). T spaces. *IBM Systems Journal, 37*(3), 454-474.

KEY TERMS

Context-Awareness: Makes applications aware of the dynamic changes in execution environment. The execution context includes but is not limited to: mobile user location, mobile device characteristics, network condition, and user activity.

Event-Based Systems: Systems in which clients (*subscribers*) have to express (*subscribe*) their interest in receiving particular events. Once clients have subscribed, servers (*publishers*) publish events, which will be sent to all interested subscribers.

Middleware: An enabling layer of software that resides between the application program and the networked layer of heterogeneous platforms and protocols. It decouples applications from any dependencies on the plumbing layer that consists of heterogeneous operating systems, hardware platforms and communication protocols.

Portable Device: Computational device that is small and can be carried by its user, such as smart cell phones, PDAs, and laptops. Unlike stationary devices, the design of portable devices typically trades-off CPU speed, memory, I/O facilities, and so forth for reduced power consumption and size.

Reflective Software: Computational process that reasons about itself, comprising an ingredient process (interpreter) formally manipulating representations of its own operations and structures.

Tuple Space: A globally shared, associatively addressed memory space that is used by processes to communicate asynchronously, based on the notion of tuples and tuple matching.

Wireless Communication: Data communication that does not require a wired connection between communicating peers, typically using radio or infrared transmissions.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 47-52, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Multimedia Technology

Adaptive Playout Buffering Schemes for IP Voice Communication

Stefano Ferretti
University of Bologna, Italy

Marco Roccetti
University of Bologna, Italy

Claudio E. Palazzi
University of Bologna, Italy

INTRODUCTION

Audio communication over IP-based networks represents one of the most interesting research areas in the field of distributed multimedia systems. Today, routing the voice over Internet enables cheaper communication services than those deployed over traditional circuit-switched networks. BoAT (Roccetti, Ghini, Pau, Salomoni, & Bonfigli, 2001a), Ekiga, FreePhone (Bolot & Vega Garcia, 1996), iCall, Kiax, NeVot (Schulzrinne, 1992), rat (Hardman, Sasse, & Kouvelas, 1998), Skype, Tapioca, vat (Jacobson & McCanne, n.d.), WengoPhone, and YATE, are just few examples of free VoIP software available to Internet users.

Without any doubts, new (wired and wireless) high-speed, broadband networks facilitate the transmission of the voice over the Internet and have determined the success of these applications. However, the best effort service offered by the Internet architecture does not provide any guarantee on the delivery of (voice) data packets. Thus, to maintain a correct time consistency of the transmitted audio stream, these voice communication systems must be equipped with schemes able to deal with the unpredictability of network latency, delay jitter, and possible packet loss.

BACKGROUND

Several proposals to face with the effects caused by network delay, delay jitter, and packet loss rate on continuous media stream playout have been presented in literature. For instance, protocol suites (e.g., RSVP, DiffServ) and networking technologies (e.g., ATM) have been devised that provide users with quality of service (QoS) guarantees (Zhang, Deering, Estrin, Shenker, & Zappala, 1993). Yet, these approaches have not been widely adopted as usual means to provide guarantees of QoS to Internet users.

An interesting alternative that is now widely exploited in most existing Internet audio communication tools amounts to the use of adaptive playout control mechanisms. Basically, these schemes are faced with the unpredictability of IP networks by compensating for variable network delays and jitters experienced during the transmission of audio packets. In particular, delay jitter is smoothed away by employing a playout buffer at the receiver side and by dynamically enqueuing audio packets in it. Output of received and buffered packets is thus artificially delayed for some time so as to have a constant audio packet playout rate, hence absorbing negative effects introduced by the delay jitter. Such a buffering policy must be adaptive, since delay jitter on the Internet may vary significantly with time. This way, dynamic playout buffers hide packet delay jitters at the cost of additional delays at the receiver (see Figure 1).

Summing up, each audio packet transmitted on the network has an associated scheduled playout delay, being defined as the total amount of time experienced by such audio packet from the instant it is generated at the source and the instant it is played out at the destination. Such a playout delay consists of: (1) the time needed for the transmitter to collect an audio sample and prepare it for transmission, (2) the network delay, and (3) the buffering time, that is the amount of time that a packet spends queued in the destination buffer before it is played out. Based on this specific notion of playout delay, a received audio packet is defined to be late when its arrival time at the destination is after the expiration of its scheduled playout time.

Needless to say, while the approach of buffering audio packets enables delay jitter removal from audio packet streams and guarantees the speech intelligibility, a critical trade-off exists between the amount of delay that is introduced in the buffer and the percentage of late packets that are not received in time for playout (and are consequently lost, see Figures 2 and 3). In fact, the longer the additional delay due to the buffering policy, the more likely a packet arrives before its scheduled playout time, but the higher the time spent by earliest packets in the buffer. Summing up, on one side, a too large percentage of audio packet loss may impair the intelligibility of an audio transmission, but, on the other side, too large playout delays (due to buffer-

Figure 1. Smoothing out jitter delay at the receiver

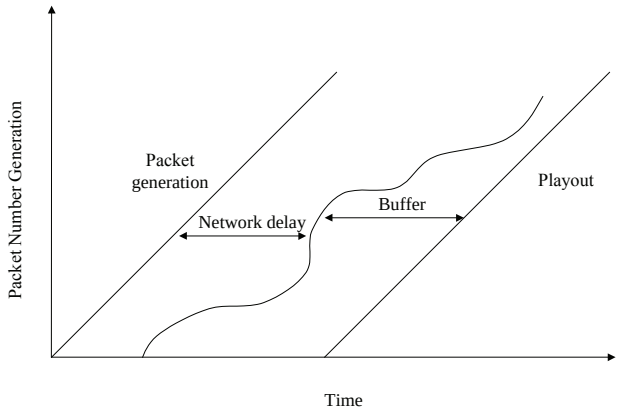


Figure 2. A small playout delay

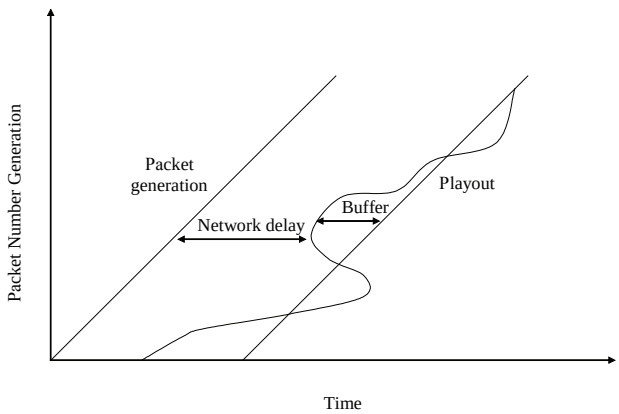
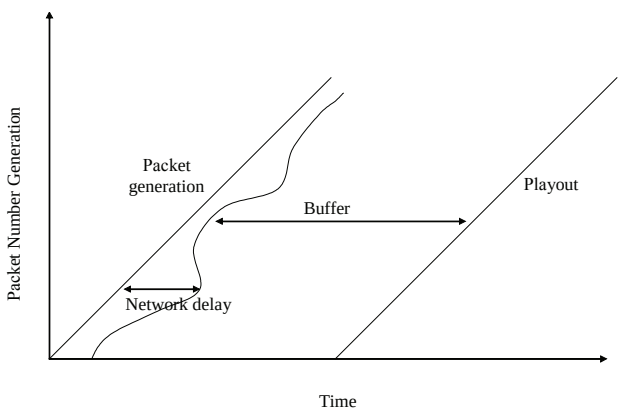


Figure 3. A large playout delay



ing) may disrupt the interactivity of an audio conversation (Kostas et al., 1998; Panzieri & Roccetti, 1997). Finally, a suitable buffering policy must reduce the probability of having underflow discontinuities (i.e., the buffer should never empty so that some audio packets are always available to be played out).

With this consideration in view, playout control mechanisms adaptively adjust the playout delay of audio packets to keep the additional buffering delay as small as possible, while minimizing the number of packets received past the point at which they are scheduled to be played out (Boutremans & Le Boudec, 2003; Fujimoto, Ata, & Murata, 2004; Liang, Farber, & Girod, 2003; Sreenan, Chen, Agrawal, & Narendran, 2000).

ADJUSTING PLAYOUT DELAYS

In the remainder of this section, we survey on different adaptive playout delay control schemes that have been designed to support speech transmission over the Internet. All these approaches adopt adaptive control mechanisms that keep the same playout delay constant throughout a given talkspurt, but permit different playout delays in different talkspurts.

Naylor and Kleinrock (1982)

Probably, the first work that proposed an adaptive playout scheduler is that presented in Naylor and Kleinrock (1982).

An important concept at the basis of the scheme is that of the talkspurt (i.e., a short burst of energy) in voice communication, during which the speech activity is carried out. In essence, a packetized audio segment may be considered as being constituted of several talkspurts separated by silence periods (during which no audio packet is generated).

According to this seminal approach, the playout delay is adjusted at the beginning of each talkspurt based on a fixed number (say m) of last delay network latencies experienced during the previous talkspurt. In particular, last m delays are recorded; then, a given amount of k values ($k < m$), representing the highest values in the set are discarded, so as to eliminate isolated cases of particularly high network delays. The maximum difference between the remaining packets is exploited to obtain an estimation of the delay jitter. Such a value is used as the buffering delay of the first packet of the talkspurt.

Ramjee, Kurose, Towsley, and Schulzrinne (1994)

A popular, important, adaptive playout delay adjustment algorithm has been proposed by Ramjee, Kurose, Towsley, & Schulzrinne (1994). The scheme proposed in that work has been extensively used in several Internet audio tools such as NeVoT (Schulzrinne, 1992), rat (Hardman et al., 1998) and FreePhone (Bolot et al., 1996).

A basic assumption of the algorithm is that an external mechanism exists that keeps synchronized physical clocks of the two nodes involved in the communication. Moreover, the approach assumes that network delays experienced by transmitted audio packets follow a Gaussian distribution.

The mechanism works as sketched in the following. The receiver buffers each received audio packet and delays its playout for a time quantity. The delay value to be associated to each audio packet is dynamically calculated based on the network latency for the transmission of the data packet and on the playout delay time. Such playout delay is adaptively adjusted from one talkspurt to the next one. Then, audio packets constituting a given talkspurt are played out in the order they were emitted at the sending site and with a constant playout delay.

In essence, the playout time p_i for the first packet i of a given talkspurt is computed as:

$$p_i = t_i + \underline{d_i} + k * \underline{v_i},$$

where t_i is the time at which the audio packet i is generated at the sending site, $\underline{d_i}$ is the average value of the playout delay, $k \in [1, 2, 4]$ is a variation coefficient (whose effect can be enforced through shift operations) and $\underline{v_i}$ is the average variation of the playout delay. The playout point p_j for any subsequent packet j of that talkspurt is computed as an offset from the point in time when the first packet i in the talkspurt was played out:

$$p_j = p_i + t_j - t_i.$$

The estimation of both the average delay and the average delay variation can be carried out using different algorithms. Basically, a viable solution is that of resorting to a linear filter such that:

$$\underline{d_i} = \alpha\, \underline{d_{i-1}} + (1 - \alpha)\, n_i$$

and

$$\underline{v_i} = \alpha\, \underline{v_{i-1}} + (1 - \alpha)\, |d_i - n_i|$$

where α is a constant value, typically chosen near 0.99, and n_i is the network delay for the i^{th} audio packet.

The scheme is also equipped with a delay spike detection and management mechanism. In essence, a delay spike is a sudden, large increase in the end-to-end network delay of some data packet, followed by a series of data packets arriving almost simultaneously at the receiver side. Delay spikes are very common in Internet transmissions. Based

on this consideration, the algorithm is set to work as described above, but when a delay spike is detected (i.e, the difference between the delay of consecutive audio packets is above a given spike threshold), the mechanism enters in a special mode. In particular, the delay of the first packet of the talkspurt is used as the estimated playout delay for each packet in the talkspurt, in order to effectively react to very large change in transmission delays.

Moon, Kurose, and Towsley (1998)

Another adaptive delay adjustment algorithm for speech transmission has been presented in Moon, Kurose, & Towsley, 1998). The main idea behind this algorithm is to collect statistics on packets already arrived and then use them to calculate the playout delay. This approach has been devised based on the consideration that cases arise when the mechanism, proposed by Ramjee et al. (1994), loses several audio packets between a talkspurt and the next one.

In essence, the value of each packet delay is recorded and the distribution of packet delays is updated with each new arrival. Thus, when a new talkspurt starts, this mechanism calculates the q^{th} percentile point for an established amount of w last arrived packets, and uses it as the playout delay for the new starting talkspurt.

Similarly to the mechanism of Ramjee et al. (1994), in the presence of delay spikes, the algorithm stops collecting packet delays and follows the spike by using as playout delay the delay value experienced by the packet that commenced that spike.

Measurements took from experiments made by the authors of that work confirmed that such a scheme can outperform the approach presented in Ramjee et al. (1994).

Roccetti et al. (2001a)

Another mechanism designed to dynamically adapt talkspurt playout delays to the network traffic conditions has been proposed in Roccetti et al. (2001a). This mechanism is at the basis of an Internet audio tool called *BoAT*.

The proposed mechanism dynamically adjusts playout delays of different talkspurts based on the network traffic conditions. In particular, the scheme does not make any assumptions on the existence of an external mechanism for maintaining an accurate physical clock synchronization between the sender node and the receiver node. Moreover, the scheme neither assumes the existence of a specific distribution of the end-to-end transmission delays experienced by the audio packets.

The technique for dynamically adjusting the playout delay for a given talkspurt is based on obtaining, in periodic intervals, an estimation of the upper bound for the packet transmission delays experienced during the voice communication. Such an upper bound is periodically computed using round trip time (RTT) values obtained from packet exchanges of a three-way handshake protocol performed between the sender and the receiver. Then, the upper bound is used to dynamically adjust the playout delay from one talkspurt to the next, with the introduction of artificially elongated or reduced silence periods.

The Need for Silent Intervals

The need of silent intervals for allowing a playout delay control mechanism to adjust to the fluctuating network conditions is common to all the described Internet voice-based conversation mechanisms. In fact, it is the presence of silent intervals that permits to dynamically adjust the playout

Figure 4. Total amount of silence periods

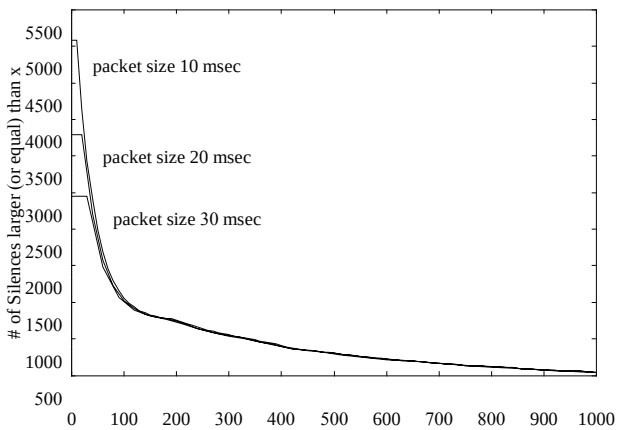
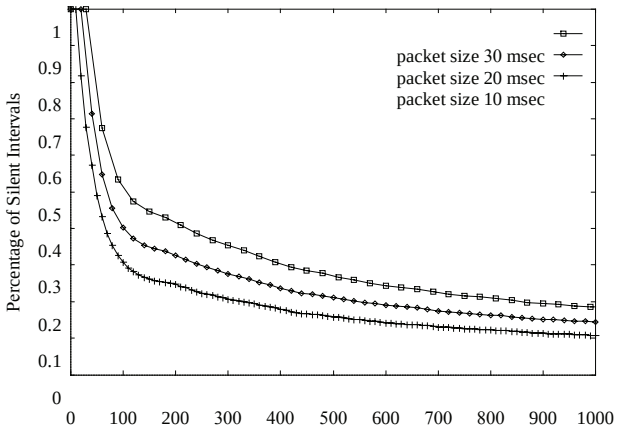


Figure 5. Percentage of silent intervals (with duration larger than x) w.r.t. the total number of silence periods



delay from one talkspurt to the next. Put in other words, the end of a talkspurt is seen as an opportunity to resynchronize different parties involved in a voice-based communication. (For this reason, spoken voice with silent periods is often thought as a *semi-countinuous* medium.)

To assess the viability of these mechanisms, an accurate model of the talkspurt characteristics of conversational speech is necessary for understanding whether sufficient (and sufficiently long) silent intervals occur in typical human conversations.

With this in view, a simulation study has been conducted in Roccetti, Ghini, and Pau (2001b). It is shown that a sufficient amount of silent intervals occur in human conversational speech to be used at the receiver to accommodate changes of the audio packet transmission delay, while maintaining speech intelligibility.

In particular, the total quantity of silent intervals within a simulated two-party one-hour-long packetized conversation amounts to about 63-66%, depending on the voice packetization interval (see Figure 4).

As already discussed, the duration of intervening silence periods in human speech is artificially reduced or elongated to accommodate at the receiver changes of the audio packet transmission delays. Thus, a fundamental issue is concerned with the average length of the intervening silence periods. Figure 5 shows the percentage of silent periods (of length larger than a fixed amount) out of the total quantity of all the obtained silence periods. Of particular interest is the fact that the larger the packet size, the larger the average silence duration.

In conclusion, these reported results show that there is room in human speech to successfully exploit sophisticated Internet audio control mechanisms.

FUTURE TRENDS

The future of voice communication systems over IP networks seems to be still bound to the effectiveness of adaptive playout schemes. New approaches are being devised which deserve attention and further investigations (Atzori & Molina, 2006; Liang et al., 2003; Narbut & Murphy 2004; Narbut, Kelly, Murphy, & Perry, 2005).

As an example, a new scheme has been recently proposed, which is based on the idea of adjusting the playout time of audio packets not only between different talkspurts, but also during talkspurts (Liang et al., 2003). In essence, playout delay is dynamically modified by scaling individual voice using a time-scale modification scheme based on the *waveform overlap-add (WSOLA)* algorithm. Basically, *WSOLA* is an interpolation method that modifies the speech signal in order to affect the speech rate only, while preserving other specific speech aspects such as timbre, pitch and quality of the voice.

Based on this, the proposed adaptive playout scheduling scheme estimates delay latencies based on last received audio packets. Then the playout time may be adjusted also within the talkspurt. Speech quality is maintained by using the mentioned WSOLA scheme.

CONCLUSION

Delay adaptation in the presence of fluctuant network delays is a crucial issue in determining the audio quality for real time speech transmission over the Internet. With this in view, the typical approach is that of dynamically adapting the audio application to the network conditions so as to minimize the trade-off between packet playout delay and packet playout

loss. The performance figures illustrate the adequacy of those mechanisms for human speech transmission across the Internet.

REFERENCES

Atzori, L., & Lobina, M. L. (2006). Playout buffering in IP telephony: A survey discussing problems and approaches. *IEEE Surveys & Tutorials, 8*(3), 36-46.

Bolot, J., & Vega Garcia, A. (1996). Control mechanism for packet audio in the Internet. In *Proceedings of IEEE SIGCOMM '96* (pp. 232-239). San Francisco.

Boutremans, C., & Le Boudec, J. Y. (2003). Adaptive joint playout buffer and FEC adjustment for Internet telephony. In *Proceedings of IEEE INFOCOM'03* (pp. 652-662). San Francisco.

Bradner, S. (2002). Internet telephony-progress along the road. *IEEE Internet Computing, 6*(3), 37-38.

Fujimoto, K., Ata, S., & Murata, M. (2004). Adaptive playout buffer algorithm for enhancing perceived quality of streaming applications. *Telecommunications Systems, 25*(3), 259-271.

Hardman, V., Sasse, M. A., & Kouvelas, I. (1998). Successful multiparty audio communication over the Internet. *Communications of the ACM, 41*(5), 74-80.

Jacobson, V., & McCanne, S. (n.d.). *vat*, Retrieved April, 2004, from ftp://ftp.ee.lbl.gov/conferencing/vat/

Kostas, T. J., Borella, M. S., Sidhu, I., Schuster, G. M., Grabiec, J., & Mahler, J. (1998). Real-time voice over packet-switched networks. *IEEE Network, 12*(1), 18-27.

Liang, Y. J., Färber, N., & Girod, B. (2003). Adaptive playout scheduling and loss concealment for voice communications over IP networks. *IEEE Transactions on Multimedia, 5*(4), 532-543.

Moon, S. B., Kurose, J., & Towsley, D. (1998). Packet audio playout delay adjustment: Performance bounds and algorithms. *ACM Multimedia Systems, 6*(1), 17-28.

Narbut, M., & Murphy, L. (2004). A new VoIP adaptive playout algorithm. In *Proceedings of IEEE Telecommunication Quality of Service: The Business of Success* (pp. 99-103). London.

Narbut, M., Kelly, A., Murphy, L., & Perry, P. (2005). Adaptive VoIP playout scheduling: Assessing user satisfaction. *IEEE Internet Computing, 9*(4), 18-24.

Naylor, E. N, & Kleinrock, L. (1982). Stream traffic communication in packet switched networks: Destination buffering constraints. *IEEE Transactions on Communication, 30*(12), 2527-2534.

Panzieri, F., & Roccetti, M. (1997). Synchronization support and group-membership services for reliable distributed multimedia applications. *ACM Multimedia Systems, 5*(1), 1-22.

Ramjee, R., Kurose, J., Towsley, D., & Schulzrinne, H. (1994). Adaptive playout mechanisms for packetized audio applications in wide-area networks. *Proceedings of IEEE INFOCOM'94* (pp. 680-688). Montreal, Canada.

Roccetti, M., Ghini, V., & Pau, G. (2001). Simulative and experimental analysis of an adaptive playout delay adjustment mechanism for packetized voice across the Internet. *International Journal of Modelling and Simulation, 21*(2), 101-106.

Roccetti, M., Ghini, V., Pau, G., Salomoni, P., & Bonfigli, M. E. (2001). Design and experimental evaluation of an adaptive playout delay control mechanism for packetized audio for use over the Internet. *Multimedia Tools and Applications, 14*(1), 23-53.

Sreenan, C. J., Chen, J. C., Agrawal, P., & Narendran, B. (2000). Delay reduction techniques for playout buffering. *IEEE Transactions on Multimedia, 2*(2), 100-112.

Schulzrinne, H. (1992). *Voice communication across the Internet: A network voice terminal*. Amherst, MA: Department of ECE and CS, University of Massachusetts.

Zhang, L., Deering, S., Estrin, D., Shenker, S., & Zappala, D. (1993). RSVP: A new resource reservation protocol. *IEEE Network Magazine, 7*(5), 8-18.

KEY TERMS

Audio Packet: Packet encoding an audio sample in digital form. Each audio packet has a timestamp and a sequence number as additional information. Timestamps are used to measure the end-to-end delay experienced during the communication, and sequence numbers are used to detect packet losses. Typically, during an audio communication, audio packets are transmitted over the network, received in a playout buffer, decoded in sequential order, and, finally, played out by the audio device.

Audio Sample: The amplitude of a waveform is measured (sampled) at regular time intervals and converted into an integer value. Each of these instantaneous measurements is an audio sample.

Delay Jitter: Variance of the network delay computed over two subsequent audio packets.

Delay Spike: Sudden, large increase in the end-to-end network delay, followed by a series of audio packets arriving almost simultaneously.

Network Delay: Time needed for the transmission of a data packet from the source to the destination.

Playout Buffer: Buffer used at the receiver to store received audio packets in interactive real-time applications to compensate for variable network delays.

Playout Control Mechanism: Adaptive mechanism that fetches audio packets from the playout buffer and sends them to the audio device for immediate playout.

Playout Delay: Total amount of time experienced by an audio packet from the time instant it is generated at the source and the time instant it is played out at the destination. It consists of (1) the time needed to collect an audio sample and to prepare it for transmission, (2) the network delay, and (3) the buffering time, that is the time that a packet spends queued in the destination buffer before it is played out.

Talkspurt: In audio communication—short burst of energy during, which the audio activity is carried out. An audio segment may be considered as being constituted of several talkspurts separated by silence periods.

A

Addressing the Central Problem in Cyber Ethics through Stories

John M. Artz
The George Washington University, USA

INTRODUCTION

The central problem in cyber ethics is not, as many might think, how to address the problems of protecting individual privacy, or preventing software piracy, or forcing computer programmers to take responsibility for the systems that they build. These are, of course, legitimate concerns of cyber ethics, but the central problem is how you decide what the right thing to do is with regard to these issues when the consequences of any responses cannot be known in advance. Stated more clearly, the central problem in cyber ethics is - how do you establish ethical standards in a professional field that is defined by a rapidly evolving technology where the consequences of the technology and the impact of any ethical standards cannot be known in the time frame in which the standards must be established? Stories play a very important role in addressing this issue. Specifically, stories provide a means of exploring ethical issues for which the full range of consequences is not currently known. But, in order to justify this claim, a few words of explanation are in order.

BACKGROUND

The word "story" means many different things to different people. For example, if one of your children tells you that your dog ate the neighbor's cat, you might challenge the veracity of this claim by asking – "Is that true, or is that just a story?" The implication is that there is truth and there are stories. And if it is a story, it cannot be true. But true versus fictitious is not the same as true versus false; and a story can contain important truths while still being wholly fictitious. If we are looking for precise intellectual truths, then perhaps stories are not the best medium for exploration. However, in areas where our understanding is unclear, either because we do not fully understand a phenomenon, or the phenomenon is not available for study because it exists in a possible world, stories play a very important role in advancing our understanding. To put a finer point on this argument, science and logic fail miserably at telling us what could be, or more importantly, what should be. In these two areas stories are powerful vehicles for intellectual explorations. A story, for the purposes of the current discussion, is a rendition or a telling of a series of true or fictitious events, connected by a narrative in which a set of characters experience and react to a set of actions or events and in doing so reveal something about the human character or condition. In order to see the value of stories for the exploration of issues in cyber ethics, three prior arguments must be made.

NARRATIVE VS. LOGICAL THINKING

Narrative and logical reasoning represent two distinct methods of making sense out of the world around us. They are both legitimate and both can be very rigorous (Bruner, 1986). Sometimes they provide alternative paths to truth and understanding. Sometimes one or the other provides the only path. Logical reasoning is general, context independent, objective and leads to a single conclusion. Narrative reasoning is specific, context dependent, open to subjective interpretation, and potentially leads to multiple conclusions. The characteristics of narrative reasoning are considered flaws when applied to logical reasoning. But the reverse applies also. A story that has only one interpretation and means the same to everyone is not much of a story. While narrative and logical reasoning are different kinds of reasoning, they are not mutually exclusive. A good narrative is also often quite logical in structure, and a good logical argument can often be better understood with a good narrative example. But for the most part, they are complimentary, alternative modes of thinking that provide different paths to truth and understanding.

To some extent, logical and narrative reasoning address different domains. Logic is well suited to mechanistic processes that can be reduced to logical description. Logic is good for articulating general principles and deductive reasons. Logic is useful for describing and explaining. While logic is good for describing "what is," narrative is good for exploring "what could be" and figuring out "what should be". Narratives are a useful means for understanding the complex and ambiguous issues in human affairs. They allow us to explore possibilities and experience situations vicariously. Narrative reasoning is particularly well suited to cyber ethics because many issues are not well understood and the goal of cyber ethics is not to discover truth about the physical world, but truth about human nature. Narrative fiction gives us a means to explore and discover truths about what could

be and what should be. Through narratives we can explore possible consequences of technology, construct alternative worlds and select the one in which we would like to live.

Critics of the use of narrative in ethics point out that after exploring narratives you always have to come back to principles. Ethics, they argue, is too messy without principles and discussion of narratives does not lead to consistent conclusions. This view misses the point of narratives. First, principles are developed by extracting the principles from experience. Narratives provide some of these experiences vicariously. Hence, narratives can be used in the development of principles. Second, it is often unclear which principles apply in given situations. Narrative explorations provide insight into situations, allowing us to determine the governing principles. And narratives can be used to explore the consequences of principled decisions to determine if the outcomes are indeed what are intended. Finally, narrative reasoning does lead to conclusions - very specific conclusions about very specific situations. Narrative reasoning is lacking in generality, as was mentioned before, not lacking in conclusions.

THE ROLE OF EMOTION IN REASON

Most people believe that emotions have no role in logical reasoning. After all, reasoning should be dispassionate and free from emotional influences that may cloud our reasoning. And there is some basis for this. For example, if you lose your temper in the middle of an argument and start flinging personal insults at your opponent, rational people would not consider you as having advanced your position. Most would say that you lost the argument when you lost your temper. Yet emotions play an important role in reasoning and in order to understand this, we need to better understand exactly what emotions are.

There is considerable debate about the exact nature of emotions. The philosopher Robert Solomon (Solomon, 1994) offers one very useful observation that "emotions are judgments about the world". If you are walking down a path in the woods and it is getting dark, you might start to get a little nervous and walk a little faster. If you hear an unfamiliar noise or a rustling in the leaves your heart may begin to beat a little faster as you experience the emotional reaction of fear. This fear is a judgment about the world in which you have judged your current situation as unsafe. You did not arrive at this judgment through a rational process. Specifically, you did not think – "It is dark and hungry animals or possibly monsters come out when it is dark. I just heard a noise that I cannot identify and therefore there could be a hungry animal near me. If I walk a little faster, I might get away before the animal gets me. If I am wrong then all I have done is walked a little faster. If I am right, I might avoid being eaten. Hence, it is logical and reasonable for me to walk faster." In fact, you probably did not think at all. You just felt scared and increased your pace. If asked later why you were walking so quickly you might come up with a reasonable explanation. But that reasonable explanation is certainly constructed after the fact.

Perhaps we have conceded at this point that emotions are judgments about the world and that they can play an important role in reasoning. The obvious question is "So what?" Why do we care and why should we bother to make an effort to incorporate our emotional judgments into our reasoning? Damsio (1994) describes the case of a young man who after suffering damage to part of his brain was no longer able to feel emotions. The unexpected side effect of this malady was that he was also unable to make good decisions or assign importance to competing tasks. He seemed normal in every other way and seemed to have his intellectual facilities fully intact. Yet he seemed no longer able to feel emotions and as a result he was unable to function as a normal person. When we make a decision we evaluate alternatives. If we are unable to feel emotions we are unable to place values on the different alternatives. If we cannot place values on the different alternatives then there is no difference between the alternatives and decision-making becomes seriously flawed. Hence, without emotions rational decision-making may not be possible.

A good story about an ethical issue is much more likely to draw an emotional response that an intellectual one, whereas an abstract analysis is more likely to yield an intellectual response. Ultimately, ethical decisions are emotional decisions because they embody human values. For this reason, examining ethics from a purely rational perspective completely misses the point.

IMAGINATION AND POSSIBLE CONSEQUENTIALISM

One of the problems in establishing standards of ethical behavior in a field driven by technology is that the consequences of the technology and reactions to the technology often cannot be known. Looking to the past to provide guidance is ineffective because the past provides few clues. Marshall McLuhan is often attributed with the famous observation that looking to the past to provide guidance for the future is like driving by looking in the rear-view mirror. Although it is disputed as to whether he ever said that or not, it is a rich metaphor for understanding how we should think about the future in times of rapid technological change.

Imagination is the key to understanding the future. The problem though, in using imagination to understand the future, is that we have a cognitive bias against understanding the future. We feel quite comfortable that we understand the past, but the future is the domain of prophesies. Yet assertions about the past are never testable because the past is

gone, never to return, while assertions about the future are testable. So one could argue, on the basis of the testability criterion, that the future is more knowable than the past. However, that discussion is for another time.

Consider imagination as the creative capacity to think of possibilities. Imagination lets us see the world not as it is, but as it could be. Seeing the world as it could be allows us to make choices about how it should be. It is this ability to see possibilities that drives us to build technologies to bring about, or implement our preferences about possible worlds. Stories are both a product and a tool of our imaginations. Using stories in moral reasoning provides a means for a slightly different view of ethics that could be called "possible consequentialism". Whereas the consequentialist evaluates actions based upon their consequences, the possible consequentialist evaluates actions based upon their possible outcomes. The possible outcomes are described in stories and the likelihood of the outcome is determined by the believability of the story given our understanding of current conditions and human nature. As the literary critic Northrop Frye points out, "The fundamental job of the imagination in ordinary life, then, is to produce, out of the society we have to live in, a vision of the society we want to live in" (Frye, 1964, p. 140).

When we examine issues in cyber ethics, we cannot examine them in terms of consequentialist ethics because the consequences are not known. However, through the use of stories we can construct imaginative scenarios and examine possible consequences. Possible consequentialism may be a preferable approach to computer ethics because we can look at possible outcomes, assess the likelihood of each, and select the outcome we prefer. Imagination provides us with a means of fully examining possible outcomes and stories provide us with the means of sharing our imaginings. By writing stories and sharing them we can explore possible consequences and, through social debate, derive imaginary truths. These imaginary truths allow us to choose the kind of world that we would like to live in.

FUTURE TRENDS

While stories play an important role in the exploration of problems in cyber ethics, there is still a serious barrier to using them. That barrier is the fact that precious few stories have been written. Current work is addressing that problem (Artz, 2004). As researchers and teachers in cyber ethics become more comfortable with writing stories, the value of stories for exploring possible worlds will make this approach increasingly more attractive. Further, as the implications of information technology and the ethical problems they bring move further into the future, the ability of stories to capture possible worlds will make them increasingly more compelling.

CONCLUSION

The central problem in cyber ethics is not, as many may suppose, how to prevent software piracy or how to protect privacy on the Internet. It is instead - how do you establish ethical standards in a professional field that is defined by a rapidly evolving technology where the consequences of the technology and the impact of any ethical standards cannot be known in the time frame in which the standards must be established? Stories play an important role in addressing this problem by providing a means of exploring ethical issues for which the full range of consequences are not known. Stories allow us to construct narrative arguments to explore issues that we do not fully understand. They allow us to explore the emotional as well as rational aspects of a situation. Stories allow us to explore worlds that do not currently exist, which, in turn, allows us to examine possible consequences and make choices about the world in which we would like to live.

REFERENCES

Artz, J. (2003). The central problem in cyber ethics and how stories can be used to address it. In L. Brennan & V. Johnson (Eds.), *Social, ethical and policy implications of information technology.* Hershey, PA: Information Science Publishing.

Artz, J. (2004). Using a story to explore an ethical dilemma. *Computers and Society E-Journal.*

Bruner, J. (1986). Actual minds, possible worlds. *Harvard University Press.*

Damsio, A. (1994). *Descartes' error: Emotion, reason, and the human brain.* Avon Books.

Edgar, S. (2002). *Morality and machines: Perspectives on computer ethics* (2nd ed.). Jones & Bartlett Pub.

Frye, N. (1964). *The educated imagination.* Indiana University Press.

Goleman, D. (1995). *Emotional intelligence.* Bantam Books.

Johnson, D. (2000). *Computer ethics* (3rd ed.). Prentice-Hall.

Solomon, R. (1994). *Love and Vengeance: A course on human emotion.* The Teaching Company Superstar Teachers Series.

KEY TERMS

Cyber Ethics: A branch of ethics that focuses on behaviors that are specifically related to information technology.

Ethics: A branch of moral philosophy that examines the standards for proper conduct.

Imagination: The creative capacity to think of possibilities.

Logical Reasoning: General, context independent, objective reasoning that leads to a single conclusion.

Narrative Reasoning: Specific, context dependent reasoning that is open to subjective interpretation, and potentially leads to multiple conclusions.

Possible Consequentialism: Evaluating an ethical position in terms of its possible consequences.

Story: A rendition or a telling of a series of true or fictitious events, connected by a narrative, in which a set of characters experience and react to a set of actions or events and in doing so reveal something about the human character or condition.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 58-61, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Adoption of E-Commerce in SMEs

Arthur Tatnall
Victoria University, Australia

Stephen Burgess
Victoria University, Australia

INTRODUCTION

Just because e-commerce technologies seem like useful tools that may assist a small to medium enterprise (SME) do its business better, it does not necessarily follow that these technologies will be *adopted* by this business. The implementation of an e-commerce system in an SME necessitates change in the way the business operates, and so it should be considered as an innovation and studied using innovation theory.

Electronic commerce (e-commerce) is concerned with how computers, information systems and communications technologies can be used by people to improve the ways in which they do business. As e-commerce necessarily involves interactions of people and technology, any study of how it is used by a small business must be considered in a socio-technical context. Although there is no universal consensus on what constitutes e-commerce, it must be considered to contain elements of information systems, computer hardware technology, business processes, communications technologies, and people. The complexity of studies in e-commerce is due, to a considerable degree, to the interconnected parts played by human actors and by the multitude of non-human entities involved. Small business managers, sales people, staff involved in procurement and warehouse operations, computers, software, Web browsers, Internet service providers (ISP), modems, and Web portals are only some of the many heterogeneous components of an e-commerce system.

BACKGROUND: ADOPTION OF E-COMMERCE BY AN SME

In this article we will argue that the decision to adopt, or not to adopt a new technology is not a straightforward one and has more to do with the interactions and associations of both human and non-human actors involved in the project than with the characteristics of the technology itself (Tatnall, 2005). Information systems are complex socio-technical entities, and research into their implementation needs to take account of this complexity, which will only be seen if it is reported in all its "messy reality" (Hughes, 1983). Research into the implementation and operation of these systems needs to take this heterogeneity into account and to find a way to give due regard to both their human and non-human aspects.

One view of the adoption of an electronic commerce innovation by a small to medium enterprise suggests that decisions are made primarily based on perceptions, by business managers, of the characteristics of the technology concerned. Such an "essentialist" approach (Haslam, 1998) involves consideration of the "essential" characteristics of the technology. Innovation diffusion (Rogers, 1995) uses this approach and is based upon the following elements:

- characteristics of the innovation itself,
- the nature of the communications channels,
- the passage of time,
- and the social system.

Another approach that has recently gained prominence is the Technology Acceptance Model (TAM), formulated by Davis and his colleagues (Davis, 1986, 1989; Davis, Bagozzi, & Warshaw, 1989). Davis identifies three major determinants of technology acceptance (or adoption) suggested by previous research studies that relate to cognition and effectiveness:

- perceived usefulness and
- perceived ease of use

to which are sometimes added attitude toward using technology and behavioral intention. TAM then attempts to use these factors to explain technology adoption.

Using approaches of this sort, the researcher would probably begin by looking for characteristics of the specific e-commerce technology to be adopted, and the perceptions, attitudes, advantages, and problems associated with its use. The next step would be to suggest that the adoption, or rejection, of this technology by an SME was due largely to these characteristics. We contend that while there may be some validity in such an approach, it is unlikely to provide the complete explanation as it would miss other influences due to inter-personal and intra-business interactions, and to the backgrounds of the people involved.

INNOVATION TRANSLATION

We argue that actor-network theory (ANT) has much to offer in a situation like this. A researcher using an actor-network approach to study innovation would concentrate on issues of network formation, investigating the human and non-human actors and the alliances and networks they build up. They would investigate how the strength of these alliances may have enticed the small business to make the adoption or, on the other hand, to have deterred them from doing so (Tatnall, 2002; Tatnall & Burgess, 2006; Tatnall & Gilding, 1999). While some research approaches to technological innovation treat the social and the technical in entirely different ways, actor-network theory proposes a socio-technical account in which neither social nor technical positions are privileged.

Actor-network theory argues that interactions between actors are heterogeneous and denies that purely technical or purely social relations are possible. It considers the world to be full of hybrid entities (Latour, 1993) containing both human and non-human elements. Change, in the ANT view, results from decisions made by actors, and involves the exercise of power. Latour (1986) argues that the mere possession of power by an actor does not automatically lead to change unless other actors can also be *persuaded* to perform the appropriate actions for this to occur.

In our experience it is often the case that when a small or medium business is considering a technological innovation it is interested in *only some aspects* of this innovation and not others (Tatnall, 2002; Tatnall & Burgess, 2002, 2006; Tatnall & Davey, 2005). In actor-network terms it needs to *translate* (Callon, 1986) this piece of technology into a form where it can be adopted, which may mean choosing some elements of the technology and leaving out others. What results is that the innovation finally adopted is not the innovation in its original form, but a translation of it into a form that is suitable for use by the recipient small business (Tatnall, 2002; Tatnall & Burgess, 2006).

In many instances a small business proprietor will adopt e-commerce because a friend is using it, or because they know a competitor is using it, or because a son or daughter learned about it at school (Burgess, 2002; Tatnall, 2002; Tatnall & Burgess, 2006). The nature and size of each small business, the intra-business interactions in which they engage, and the backgrounds and interests of particular individuals in each are also likely to have had an important affect that would, most likely, have been ignored by the essentialist approach offered by innovation diffusion or TAM. Actor-network theory, in examining alliances and networks of human and non-human actors, provides a good foundation from which small business adoption and use of e-commerce can be researched. The ANT approach will be further amplified in the case studies that follow, particularly in respect of the identification of actors and networks.

CASE STUDIES OF TECHNOLOGY ADOPTION

This article now offers several brief case studies in which actor-network theory has provided a means by which adoption (or non-adoption) of technology can be explained. In each case, data for the study were obtained through semi-structured interviews with the proprietors and personnel of the businesses involved.

1. Adoption of a Portal by a Storage and Transport Company

The business to be considered in this study is a medium-sized Melbourne company, with about 50 employees, that stores frozen food and transports it to supermarkets and other locations around the country. It became clear from the study that the transport company had "not really been into computers" and had only recently started coming to grips with this technology.

Although the manager had some idea of the benefits to his company of using the portal, he had no clear plan for using it. It was just "a really good idea." The reasons he adopted this innovation thus had little to do with the characteristics of this technology, and much more to do with his involvement with the local business community and because of his belief that the portal had the potential to improve business in the region.

2. Adoption of Internet Technologies by a Rural Medical Practice

Rural medical general practitioners (GPs) in Australia have not been rapid adopters of ICT, and this has been of concern to both the commonwealth government and to medical industry bodies (Tatnall, 2005). Many ICT products have been developed to support GPs in all aspects of their work in Australia (GPSRG, 1998), and much research and development in this area has already been undertaken. It is apparent, however, that GPs are not making as much use of these systems as they could be (Burgess, Darbyshire, Sellitto, Tatnall, & Wenn, 2003a).

A study of 1200 randomly selected general practices from across Australia (GPCG, 2001) identified the main uses GPs make of computers. Although a large number of respondent practices (89%) were computerized, it does not necessarily follow that computers are being directly used by the GPs themselves. Major uses identified were administrative functions (85%), clinical functions (76%), script writing (60%), general referral letters (57%), and receiving pathology and other test results electronically (57%). While these figures suggest that most GPs make some use of ICT,

they are somewhat misleading in not indicating the extent of this use (Everitt & Tatnall, 2003). Burgess, Darbyshire, Sellitto, Tatnall, and Wenn (2003b) also note that the use of ICT by GPs reflects the patterns of use in small business, and it seems that if GPs do use ICT, then in many cases it is for perceived cost savings rather than for adding value (Burgess & Trethowan, 2002).

Dr. Doyle and Dr. Holmes (although this is a real case, these names are fictitious) operate a small medical practice in a country town 80km north of Melbourne, employing two secretarial assistants. Although aware that ICT could be of benefit to their practice, they currently use computers only for patient billing and secretarial functions. The interactions resulting from a complex network of actors have been responsible for this practice not using ICT to the extent that they could. This network includes the following actors: Dr. Doyle, Dr. Holmes, the regional Division of General Practice, other local GPs (friends and colleagues), patients, their secretarial staff, the local hospital, the local pathology laboratory, "Medical Director" software, Internet broadband technology, computers, and medical consulting rooms.

While the regional division of general practice, some of their patients, and several of their colleagues strongly encourage the use of ICT, many of the other actors have worked to discourage it. By providing patient data in paper, and not electronic, format, both the local hospital and pathology laboratory act to discourage the use of computers into which the GPs would have to type all this data. The difficulty of using the software and of making all the required Internet connections is also a discouragement.

A small group of actors is thus working to encourage these GPs in the use of ICT, while another group is working to discourage this use. At present the second group is winning the battle. Although Dr. Doyle and Dr. Holmes understand the benefits of the technology, they are very busy people and have not yet been persuaded by a large enough coalition of actors to adopt it.

3. Adoption of Electronic Commerce by a Small Publishing Company

The next case concerns a small publishing company where four people work on the production of textbooks and research publications. The company is a very small business with a relatively low turnover but a well-established market. The business has adopted some e-commerce technologies, but not others. Some time ago, it registered a domain name and set up its own Web page, but only for informational purposes. The site shows details of the company's products and indicates how orders can be placed, but does not support sales or other business-consumer (B-C) transactions.

When asked why the company had not considered instituting a B-B e-commerce system with the printer, the director replied that they would like to do so, but that the printer was not really interested or geared up to get into this type of order. Adoption decisions of the company thus had more to do with the technological status of their suppliers and their customers than with the e-commerce technology itself. The company was interested in using those e-commerce technologies that its business partners used and those it considered useful or timesaving. What was adopted could be seen as a translation of an e-commerce innovation to include just these things.

4. Non-Adoption of E-Commerce by a Small Chartered Accountancy Firm

The last case is of a small chartered accountancy firm, which is a family business in the western suburbs of Melbourne. Employees of the business are the main accountant, who has a degree in accounting and is a certified practicing accountant (CPA); the main accountant's father, who previously ran the business but, as he is not qualified, is limited these days mainly to supporting the taxation side of the business; another accountant (CPA); and a full-time secretary. The firm offers services that are typical of a small accounting business, and its clients include both individuals and small businesses.

For some time members of the business had been debating whether to set up a Web site. The decision seemed to come down to two major opposing viewpoints. The first, held by the main accountant, was that a Web site was "the way of the future," that customers would expect it, and that some competitors already had one. The opposite viewpoint, held by the father, was that it is a waste of time and money, that there was nothing you could put on a Web site that customers would want anyway, and "who is going to do it?" Other members of the business seemed quite apathetic about the whole matter.

The upshot was that the Web site could not be "translated" into a form where it could be adopted by this business. The main accountant was the prime motivator in attempting to define the problem and to discover the solution: he did not do this well enough, however, to gain acceptance of the innovation.

FUTURE TRENDS: ANT AND E-COMMERCE INNOVATION

The theory of innovation diffusion (Rogers, 1995) and also the related technology acceptance model (Davis, 1989) are well established and have been used as the framework of many studies. In most cases, however, the success of the diffusion model has been in explanation of innovation "in the large" when the statistical effects of big numbers of orga-

nizations and individuals involved come into play. Although TAM is somewhat less mechanistic, it suffers a similar fate. They have, typically, been less successful in explaining how particular individuals or specific organizations make their adoption decisions, and it is in situations like this that an innovation translation approach, using actor-network theory, is especially useful.

In offering a socio-technical approach to theorising innovation, ANT provides a particularly useful tool to the study of innovations in which people and machines are intimately involved with each other. The adoption of e-commerce technologies certainly involves a consideration of the technologies themselves, but also of business organizations, business processes, and the needs and likes of individual humans. ANT, we suggest, is especially useful in researching innovations like these, and in particular, when considering individual adoption decisions.

The main use made of any research approach such as ANT is in the study of past events, and ANT makes no claim to be able to predict what may happen in the future. We suggest, however, that ANT analysis can identify some pointers toward the successful introduction of an innovation, and the change management associated with this. ANT argues that it is not the characteristics of either the innovation itself or the potential adopter acting alone that are important, but rather the interactions of many actors. The key to successful change management, it would thus seem, involves allowing for these interactions and for the socio-technical nature of the process.

CONCLUSION

All of the situations described in this article point to the fact that decisions on the adoption of electronic commerce technologies are often made on the basis of more than just the characteristics of the technology and that in many cases these characteristics are not especially significant in the decision making process.

In each of these instances, the actor-network approach therefore offers a useful explanation of why a particular e-commerce initiative was or was not adopted. On the other hand, an innovation diffusion approach to investigating each of the potential adoptions would have looked for explanations for the uptake, or lack of uptake, primarily in the characteristics and properties of the technology in each case. It would not have considered, as particularly important, the human and non-human interactions described here. In our view, the decision to adopt, or not to adopt, has more to do with the interactions and associations of both human and non-human actors involved in the project rather than with characteristics of the technology.

REFERENCES

Burgess, S. (2002). Information technology in small business: Issues and challenges. In S. Burgess (Ed.), *Information Technology and Small Business: Issues and Challenges* (pp. 1-17). Hershey, PA: Idea Group Publishing.

Burgess, S., Darbyshire, P., Sellitto, C., Tatnall, A., & Wenn, A. (2003a). A classification system for tracking the adoption of IT by GPS in rural Australia. In *IADIS International Conference—WWW/Internet 2003* (pp. 93-99). Algarve, Portugal: IADIS Press.

Burgess, S., Darbyshire, P., Sellitto, C., Tatnall, A., & Wenn, A. (2003b). Tracking the adoption of IT by GPs in rural Australia: A socio-technical approach. In K. K. Dhanda & M. G. Hunter (Ed.), *ISOneWorld: Nurturing Executive Networks*. Las Vegas, NV: The Information Institute.

Burgess, S., & Trethowan, P. (2002). *GPs and their Web sites in Australia: Doctors as small businesses*. Las Vegas, NV: IS OneWorld.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, action & belief. A new sociology of knowledge?* (pp. 196-229). London: Routledge & Kegan Paul.

Davis, F. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. Boston: MIT.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 10*(3), 318-340.

Davis, F. D., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Everitt, P., & Tatnall, A. (2003). *Investigating the adoption and use of information technology by general practitioners in rural Australia and why this is less than it might be*. Perth: ACIS.

GPCG. (2001). *Measuring IT use in Australian general practice*. Brisbane: General Practice Computing Group, University of Queensland.

GPSRG. (1998). *Changing the future through partnerships*. Canberra: Commonwealth Department of Health and Family Services, General Practice Strategy Review Group.

Haslam, N. O. (1998). Natural kinds, human kinds, and essentialism. *Social Research, 65*(2), 291-314.

Hughes, T. P. (1983). *Networks of power: Electrification in western society, 1880-1930*. Baltimore & London: Johns Hopkins University Press.

Latour, B. (1986). The powers of association. In J. Law (Ed.), *Power, action and belief. A new sociology of knowledge? Sociological review monograph 32* (pp. 264-280). London: Routledge & Kegan Paul.

Latour, B. (1993). *We have never been modern*. Hemel Hempstead: Harvester Wheatsheaf.

Rogers, E. M. (1995). *Diffusion of innovations*. New York: The Free Press.

Tatnall, A. (2002). Modelling technological change in small business: Two approaches to theorising innovation. In S. Burgess (Ed.), *Managing information technology in small business: Challenges and solutions* (pp. 83-97). Hershey, PA: Idea Group Publishing.

Tatnall, A. (2005). Technological change in small organisations: An innovation translation perspective. *International Journal of Knowledge, Culture and Change Management, 4*(1), 755-761.

Tatnall, A., & Burgess, S. (2002). Using actor-network theory to research the implementation of a B-B portal for regional SMEs in Melbourne, Australia. In C. Loebbecke, R. T. Wigand, J. Cricar, A. Pucihar, & G. Lenart (Eds.), *15th Bled Electronic Commerce Conference—E-Reality: Constructing the E-Economy* (pp. 179-191). Bled, Slovenia: University of Maribor.

Tatnall, A., & Burgess, S. (2006). Innovation translation and e-commerce in SMEs. In M. Khosrow-Pour (Ed.), *Encyclopedia of e-commerce, e-government and mobile commerce* (pp. 631-635). Hershey, PA: Idea Group Reference.

Tatnall, A., & Davey, B. (2005). A new spider on the Web: Modelling the adoption of Web-based training. In P. Nicholson, J. B. Thompson, M. Ruohonen, & J. Multisilta (Eds.), *E-training practices for professional organizations* (pp. 307-314). Assinippi Park, MA: Kluwer Academic Publishers/IFIP.

Tatnall, A., & Gilding, A. (1999). Actor-network theory and information systems research. In B. Hope & P. Yoong (Eds.), *10th Australasian Conference on Information Systems (ACIS)* (pp. 955-966). Wellington, Victoria: University of Wellington.

KEY TERMS

A

Actor: An entity that can make its presence individually felt by other actors. Actors can be human or non-human, non-human actors including such things as computer programs, portals, companies, and other entities that cannot be seen as individual people. An actor can be seen as an association of heterogeneous elements that constitute a network. This is especially important with non-human actors, as there are always some human aspects within the network.

Actor-Network Theory (ANT): An approach to research in which networks' associations and interactions between actors (both human and non-human) are the basis for investigation.

E-Commerce Systems: Contain elements of information systems, business processes, and communications technologies.

Innovation Diffusion: A theory of innovation in which the main elements are characteristics of the innovation itself, the nature of the communication channels, the passage of time, and the social system through which the innovation diffuses. The main proponent of Innovation Diffusion is Rogers (1995).

Innovation Translation: A theory of innovation in which, instead of using an innovation in the form it is proposed, potential adopters *translate* it into a form that suits their needs.

Small to Medium Enterprise (SME): For the purpose of this article, SMEs are considered to be those businesses that have from 1-20 employees—small—and 21-50 employees—medium.

Technology Acceptance Model (TAM): A theory of innovation developed by Davis (1986) in which the main elements are perceived usefulness, perceived ease of use, attitude toward using technology, and behavioral intention.

Technology Adoption: The decision, by an organization or individual, to utilize and implement a technology.

Technological Innovation: The introduction or alteration of some form of technology (often information technology) into an organization.

Adoption of Electronic Commerce by Small Businesses

Serena Cubico
University of Verona, Italy

Giuseppe Favretto
University of Verona, Italy

INTRODUCTION

The role played by small business in economic growth and development in the world is officially recognized, in both the economic literature and in official documents (e.g., Organization for Economic Cooperation and Development, European Commission, U.S. Department of State).

Information and communication technology connectivity are widespread in all sized businesses, but small businesses seem slower than larger ones to adopt and use ICT and electronic commerce.

SMEs (small- to medium-sized enterprises) are independent firms that employ less than 10 (micro), 50 (small), and 250 (medium) employees (European Commission, 2003); the United States includes firms with fewer than 500 employees in the definition of an SME (OECD, 2000a).

In Europe, SMEs contribute up to 80% of employment in some industrial sectors (e.g., textiles, construction, furniture), and they are defined as "a major source of entrepreneurial skills, innovation and contribute to economic and social cohesion" (European Commission, 2005, p. 3); in the U.S. economy, small businesses represent 99.7% of all employers and "broaden a base of participation in society, create jobs, decentralize economic power and give people a stake in the future" (U.S. Department of State, 2006, p. 2).

To synthesize: more than 95% of OECD enterprises are SMEs, accounting for 60-70% of employment in most countries (OECD, 2000a).

The same proportion is indicated by the United Nations Conference on Trade and Development; in fact, SMEs account for 60-70% of all employment in developing countries (UNCTAD, 2002).

BACKGROUND

Research interests in e-commerce utilization in SMEs have been driven by a basic hypothesis that this type of technology can offer new opportunities to counterbalance disadvantages of size, resources, geographic isolation, and market reach (Wymer & Regan, 2005).

Several different disciplines (management, organizational behavior, communications, computer science, information systems, marketing, work, and social psychology) are involved in research on incentives and technology adoption barriers. In this regard, different theoretical and applied models already exist:

- The *Theory of Reasoned Action (TRA),* and its extension, the *Theory of Planned Behavior (TPB)* (Ajzen & Fishbein, 1980; Chau & Hu, 2001; Harrison, Mykytyn, & Riemenschneider, 1997) are based on assumptions that a person's intentions are the best guide to behavior, and that there is a link between attitudes and behavior.
- The *Technology Acceptance Model (TAM)* (Straub, Limayem, & Karahannaevaristo, 1995), defines models as to how users come to accept and make use of technology.
- The *Adoption, Innovation and Diffusion Theory* (Rogers, 1995) defines adopter (of any new innovation or idea) categories as innovators, early adopters, early majority groups, late majority groups, and laggards.
- *Social Cognitive Theory* (Bandura, 1996) defines human behavior as a triadic, dynamic, and reciprocal interaction of personal factors, behavior, and the environment.
- The *Unified Theory of Acceptance and Use of Technology (UTAUT)* (Venkatesh, Morris, Davis, & Davis, 2003) uses performance expectancy, effort expectancy, social influence, and facilitating conditions as direct determinants of usage intention.

Table 1 presents a synthesis of the numerous factors influencing adoption of e-commerce adoption from the literature.

As we can see, adoption of electronic commerce by SMEs is influenced by different factors. Grandon and Pearson (2004) identified and synthesized four factors that have statistically significant effects on e-commerce utilization: *organizational readiness* includes financial and technological resources and compatibility of e-commerce with company's culture, values, and preferred work activity; *external pressure* is defined by competing, social factors, dependency on other firms already using e-commerce, the industry, and the government; and *perceived ease of use* and *perceived*

Table 1. Factors influencing decision to adopt e-commerce/e-business/Internet technology (adapted from Wymer & Regan, 2005, p. 442)

Factor Name	Description
Environmental Factors	
Competitive Pressure	Competitive pressure from other Internet adopters within the industry
Government	Government rules and regulations
Market	Viable market or customer base for e-commerce
Partners/Vendors	Availability of the right partners
Supplier Readiness	Readiness of suppliers for electronic business
Knowledge Factors	
Change Experience	Employee experience with making major changes
Executive Experience	Experience of top executives with computers and the Internet
Innovativeness	Company's willingness to adopt new technology
Models	Successful models of use in the industry
Need	Perceived need for change or implementation of Web and Internet technologies
Prior Experience	Company's prior experience with new technology implementation
Trust	Trust or confidence in Web and Internet technologies
Understanding	Understanding of available opportunities and options with e-commerce
Value	Perceived value or relevance to the business
Organizational Factors	
Capital	Access to capital for start-up
Employee Reduction	Resulting reduction in number of employees
Priority	Priority relative to other projects that require existing resources and time
Profitability	Projected profitability of e-commerce
Technical Expertise	Availability of technical staff or consultants with Web skills
Technological Factors	
Cost	Cost to setup and maintain
EC Technology	Technology for selling products or services online
Infrastructure	Access to network services or infrastructure to support Web and Internet technologies
Reliability	Reliability of Web and Internet technologies
Security	Security issues
Technology Availability	Availability or adequacy of existing technology and tools

usefulness. In particular, the last two factors turned out to be most influential in adoption of electronic commerce by top managers of SMEs, while compatibility emerged as a partial factor that highly influenced e-commerce adoption, as opposed to financial and technological resources.

The benefits of e-commerce are for all sized businesses, and even SMEs could reap advantages. Studies in numerous counties reveal that SMEs have been slower to adopt e-commerce than their larger counterparts, however information technology use by SMEs is increasing (Drew, 2003).

Moreover, many studies define and analyze e-commerce and small businesses through different points of views with images and concepts that are not of help in understanding the phenomenon (Ngai & Wat, 2002).

Small business have many reasons for selling or buying over the Web. They can receive benefits from this type of commerce—that is, "adding distribution channels, increasing overall sales, expanding their reach beyond local markets, or gaining greater exposure in existing markets [in] building an Internet storefront for a retail shop" (Mehta & Shah, 2001, p. 88).

SMEs use e-commerce in three different ways:

Internet start-ups' are inventing new ways of creating value added, new service and new business models…'Established

Table 2. Percentage of small European enterprises (10 to 49 employees) that use e-commerce for purchasing and sales (data in parentheses refer to large enterprises: more than 250 employees) (adapted from Eurostat, 2006)

ELECTRONIC COMMERCE	EU	DE	ES	IT	CY	LV	HU	PL	SE	UK
Purchase	22(40)	40(52)	4(7)	4(15)	14(20)	1(1)	4(8)	9(11)	40(57)	48(72)
Sales	10(31)	14(40)	2(15)	2(13)	2(29)	1(1)	4(6)	4(14)	21(45)	22(45)

small firms' are developing their own e-commerce strategies to expand their business by entering new markets, often internationally…'Existing SMEs' are entering into electronic partnership with large corporate customers. (OECD, 2000b, p. 17)

Studies on adoption of electronic commerce by SMEs focused on different aspects:

- The benchmark on the use of the Internet emphasizes that SMEs need to think both globally and strategically, and that they must learn from competitors (Webb & Sayer, 1998).
- A learning organization style oriented to upgrade competencies and to acquire new knowledge (defined *high-order* by Argyris & Schon, 1978) is more involved in the use of the Internet and e-commerce (Chaston, 2001).
- There are significant differences in frequency and type of use, related to company size: four or more employees represents the critical dimension for more frequent and sophisticated use (Dandridge & Levenburg, 2000).
- The greater the usage of Internet technology is among entrepreneur-led family business, the more it is possible to find ties between entrepreneurial profile and growth (Davis & Harveston, 2000).
- The analysis of strategic use of the Internet and e-commerce in SMEs shows that they are opportunistic in their adoption and that the communication requirement has been a motivating factor of implementation (Sadowski, Maitland, & van Dongen, 2002).
- SMEs present different sequences of e-commerce adoption: in the early stages they use the lowest levels of e-commerce service; in the second, e-mails are used to communicate with customers, suppliers, and employees; the third level of adoption includes information-based Web sites operating and developing online ordering services; and the most advanced adopters use online ordering and are developing online payment capabilities (Daniel, Wilson, & Myers, 2002).

FOCUS: SPECIFIC DIFFICULTIES FOR SMALL BUSINESSES WITH E-COMMERCE ADOPTION

Some problems to understanding the barriers faced by smaller firms are due to the fact that they are a not uniform group and that their characteristics vary by sector:

High technology, knowledge-intensive small firms are more likely to use e-commerce than other small firms and there are differences between industry sectors in terms of e-commerce use and strategy development. (Fillis & Wagner, 2005, p. 607)

Non-adopters show some characteristics that differ from SMEs that adopt e-commerce. In fact, they present high scores related to barriers/impediment and tend to be slower in detecting changes in technologies that might affect their business, whereas the adopters are more aware of opportunities afforded by technology, are more costumer oriented, and are more sensitive to changes in their customer/competitive environment (McCole & Ramsey, 2005).

It is interesting to see the adoption of e-commerce in different areas of the world.

In European Union (EU) countries, the use of an internal computer network and intranets is progressing well, but there is space for improvement, especially among smaller enterprises (10 to 49 employees). Internet and Web sites are not enough of a support system for e-business. Enterprises need to use more of their technological potential in order to reap maximum benefits. E-commerce can be difficult to start up, and generally, enterprises tend to prefer to purchase than to sell online. According to Eurostat (2006), only 10% of EU small firms engaged in e-commerce sales activity.

Table 2 shows the most significant data from the EU.

In the United States, there were slight differences in e-commerce attitudes and experience between small/medium-sized enterprises and large establishments. This difference in e-commerce experience suggests that a considerable population of small establishments (less than 25 employees) may be less prepared for e-commerce. Routine use of e-commerce

Table 3. Mean percent of total commerce conducted online by SMEs (25-250 employees) in the U.S. (data in parentheses refers to large enterprises: more than 250 employees) (adapted from Fomin, King, Ljjtinen, & McGann, 2005)

ELECTRONIC COMMERCE	SMEs
Purchase	73.5(79.8)
Sales	5.1(4.0)

lags behind more traditional forms of commerce (Fomin, King, Ljjtinen, & McGann, 2005).

Table 3 shows the most significant data from the United States.

In many developing countries, no statistical indicators on e-business have been collected and considerations are more difficult due to the very significant differences among SMEs in different regions and countries.

The general conclusion (UNCTAD, 2004, p. 52) on the adoption of e-business in developing countries are:

"…for SMEs is that it is relatively easy to start using PCs, then connect to the Internet using e-mail, and then set up a Web page. However, the introduction of the Internet into their business activities (…including e-commerce) does not follow straightaway, and larger companies are more likely to automate their business processes (and to do so earlier) than smaller companies. One explanation for this is that most SMEs have no defined e-business strategy."

The specific difficulties in different countries by small firms in e-commerce adoption are the object of various studies (Fomin, King, Ljjtinen, & McGann, 2005; Cubico, Venturini, Russo, & Favretto, 2005; Fillis & Wagner, 2005; Jones, Beynon-Davies, & Muir, 2003; OECD, 2004; Walczuch, Van Braven, & Lundgren, 2000), and from these, we can draw a basic synthesis of the factors involved.

The most cited barriers related to smaller-sized enterprises in e-commerce adoption are:

- Concern about privacy of data or security issues,
- need for face-to-face customer interaction,
- implementation costs of e-commerce sites,
- lack of financial resources and high costs,
- customers do not use the technology
- applicability to enterprises
- finding staff with e-commerce expertise and uncertainty on how to implement,
- insufficient education/information about benefits,
- level of ability to use the Internet as part of business strategy,
- prevalence of credit card use in the country,
- taxation of Internet sales,
- making needed organizational changes,
- inadequate legal protection for Internet purchases,
- cost of Internet access, and
- business laws do not support e-commerce.

The following specific elements are limited factors: awareness of SME access to infrastructure and skills, critical mass among business partners, confidence in legal and regulatory framework/security, and adaptation of business processes.

FUTURE TRENDS

The first steps to understanding the phenomenon of e-commerce adoption in SMEs are to define their characteristics and to know the specific needs of these types of enterprises.

Furthermore, it would be interesting to understand more about specific patterns related to decisions and choices in e-commerce adoption through different disciplines. For instance, important psychological aspects that inform decisions to work with electronic markets seem to be reliability, security, confidence building, and the legal framework (selected cultural differences have been demonstrated in recent cross-cultural research by Dinev, Bellotto, Hart, Russo, Serra, & Colautti, 2006), or the effect of a specific organizational culture that impedes SMEs' decisions (Feltham, Feltham, & Barnett, 2005; Schein, 1983).

Another way could be to identify specific skills, knowledge, attitudes, and aptitudes necessary and distinctive to e-commerce adoption by SMEs.

CONCLUSION

Work on SMEs and different levels of e-commerce adoption shows that firms needs to improve specific skills (Taylor, Mcwilliam, England, & Akomode, 2004), in order to develop existing processes or to introduce new processes and to integrate their new Web-based systems (with existing internal systems and with external systems for customers and suppliers) (Jeffcoate, Chappel, & Feindt, 2000).

Help to SMEs and an increase in different types of commerce (especially in electronic commerce, which represents the future in a global market) are important to the improvement of their chances of survival and for the longevity of economic systems based on these types of enterprises.

REFERENCES

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Englewood Cliffs, NJ: Prentice Hall.

Argyris, C., & Schon, D.A. (1978). *Organizational learning: A theory of action perspective.* Reading, MA: Addison-Wesley.

Bandura, A. (1996). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Chaston, I. (2001). The Internet and e-commerce: An opportunity to examine organisational learning in progress in small manufacturing firms? *International Small Business Journal, 19*(2), 13-30.

Chau, P.Y.C., & Hu, P.J.-H. (2001). Information technology acceptance by individual professionals: A model comparison approach. *Decision Sciences, 32*(4), 699-719.

Cubico, S., Venturini, B., Russo, V., & Favretto, G. (2005). E-commerce and cultural innovation in small and micro sized enterprises: A preliminary investigation. *Proceedings of the 30th Annual Congress of the International Association for Research in Economic Psychology,* Prague, Czech Republic.

Dinev, T., Bellotto, M., Hart, P., Russo, V., Serra, I., & Colautti, C. (2006). Privacy calculus model in e-commerce — a study of Italy and the United States. *European Journal of Information Systems, 15*(4), 389-402.

Dandridge, T., & Levenburg, N.M. (2000). High-tech potential? An exploratory study of very small firms' usage of the Internet. *International Small Business Journal, 18*(2), 81-91.

Daniel, E., Wilson, H., & Myers, A. (2002). Adoption of e-commerce by SMEs in the UK. *International Small Business Journal, 20*(3), 253-270.

Davis, P.S., & Harveston, P.D. (2000). Internationalization and organizational growth: The impact of Internet usage and technology involvement among entrepreneur-led family businesses. *Family Business Review, 13*(2), 107-120.

Drew, S. (2003). Strategic uses of e-commerce by SMEs in the east of England. *European Management Journal, 21*(1), 79-88.

European Commission. (2003). Commission recommendation concerning the definition of micro, small and medium-sized enterprises. *Official Journal of the European Union, 361.*

European Commission. (2005). *Implementing the Community Lisbon Programme. Modern SME policy for growth and employment.* Commission Communication COM(2005) 551 Final.

Eurostat. (2006). The Internet and other computer networks and their use by European enterprise to do e-business. *Statistics in Focus* [Online Serial], *28.* Retrieved from http://ec.europa/eurostat

Feltham, T.S., Feltham, G., & Barnett, J.J. (2005). The dependence of family business on a single decision-maker. *Journal of Small Business Management, 43*(1), 1-15.

Fillis, J., & Wagner, B. (2005). E-business development: An exploratory investigation of the small firm. *International Small Business Journal, 23*(6), 604-634.

Fomin, V.V., King, J.L., Ljjtinen, K.J., & McGann, S.T. (2005). Diffusion and impacts of e-commerce in the United States of America: Results from an industry survey. *Communications of the AIS, 16*(October), 559-603.

Grandon, E.E., & Pearson, J.M. (2004). Electronic commerce adoption: An empirical study of small and medium U.S. business. *Information & Management, 42*(1), 197-216.

Harrison, D.A., Mykytyn, P.P. Jr., & Riemenschneider, C.K. (1997). Executive decisions about adoption of information technology in small business: Theory and empirical tests. *Information Systems Research, 8*(2), 171-195.

Jeffcoate, J., Chappel, C., & Feindt, S. (2000). Attitude towards process improvement among SMEs involved in e-commerce. *Knowledge and Process Management, 7*(3), 187-185.

Jones, P., Beynon-Davies, P., & Muir, E. (2003). E-business barriers within the SME sector. *Journal of Systems and Information Technology (E-Business Special Edition), 7*(1), 1-26.

Mehta, D.T., & Shah, V. (2001). E-commerce: The next global frontier for small businesses. *Journal of Applied Business Research, 17*(1), 87-94.

McCole, P., & Ramsey, E. (2005). A profile of adopters and non-adopters of e-commerce in SME professional service firms. *Australian Marketing Journal, 13*(1), 36-48.

Ngai, E.W.T., & Wat, F.K.T. (2002). A literature review and classification of electronic commerce research. *Information & Management, 39*(5), 415-429.

OECD (Organization for Economic Cooperation and Development). (2000a, June). Small and medium-sized enterprises: Local strength, global reach. *OECD Policy Brief* [Online Serial]. Retrieved from http://www.oecd.org

OECD. (2000b, June 14-15). Realizing the potential of electronic commerce for SMEs in the global economy. *Proceedings of the Conference for Ministers Responsible for SMEs and Industry Ministers,* Bologna, Italy.

OECD. (2004). *ICT, e-business and SMEs*. DSTI/IND/PME(2002)7/FINAL, OEDC, France.

Rogers, E.M. (1995). *Diffusion of innovations* (4th ed.). New York: The Free Press.

Sadowski, B.M., Maitland, C., & van Dongen, J. (2002). Strategic use of the Internet by small- and medium-sized companies: An exploratory study. *Information Economics and Policy, 14*(1), 75-93.

Schein, E.H. (1983). The role of the founder in creating organizational culture. *Organizational Dynamics,* (Summer), 13-28 [reprint (1995). *Family Business Review, 8*(3), 221-238].

Schein, E.H. (1990). Organizational culture. *American Psychologist, 45*(2), 109-119.

Straub, D., Limayem, M., & Karahannaevaristo, E. (1995). Measuring system usage—implications for IS theory testing. *Management Science, 41*(8), 1328-1342.

Tagiuri, R., & Davis, J.A. (1982). *Bivalent attributes of the family firm.* Working Paper, Harvard Business School, USA [reprint (1996). *Family Business Review, 9*(2), 199-208].

Taylor, M.J., Mcwilliam, J., England, D., & Akomode, J. (2004). Skill required in developing electronic commerce for small and medium enterprises: Case based generalization approach. *Electronic Commerce Research and Applications, 3*(3), 253-265.

UNCTAD (United Nations Conference on Trade and Development). (2002). Report of the expert meeting on improving the competitiveness of SMEs in developing countries: The role of finance, including e-finance to enhance enterprise development. *Proceedings of the 6th Session of the Trade and Development Board, Commission on Enterprise, Business Facilitation, and Development,* Geneva, Switzerland. Retrieved from http://www.unctad.org/en/docs/c3em13d3.en.pdf

UNCTAD. (2004). *E-commerce and development report 2004.* Retrieved from http://www.unctad.org/en/docs/ecdr2004_en.pdf

U.S. Department of State. (2006). Entrepreneurship and small business. *eJournal USA: Economic Perspective* [Online Serial], *11*(1). Retrieved from http://usinfo.state.gov/journals/ites/0106/ijee/ijee0106.htm

Venkatesh, V., Morris, M., Davis, G., & Davis, F. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Walczuch, R., Van Braven, G., & Lundgren, H. (2000). Internet adoption barriers for small firms in The Netherlands. *European Management Journal, 18*(5), 561-572.

Webb, B., & Sayer, R. (1998). Benchmarking small companies on the Internet. *Long Range Planning, 31*(6), 815-827.

Wymer, S.A., & Regan, E.A. (2005). Factors influencing e-commerce adoption and use by small and medium businesses. *Electronic Markets, 15*(4), 438-453.

A

KEY TERMS

Electronic Commerce (E-Commerce): Transactions conducted over Internet protocol-based networks and over other computer-mediated networks. Goods and services are ordered over those networks, but payment and final delivery of goods or services may be conducted on or off-line. Orders received via telephone, facsimile, or manually typed e-mails are not counted as electronic commerce (Eurostat, 2006).

European Union (EU): Family of democratic European countries. The six founders (on March 25, 1957, with the Treaty of Rome) are Belgium, France, Germany, Italy, Luxembourg, and The Netherlands. The European Union acts in a wide range of policy areas—economic, social, regulatory, and financial—through solidarity policies (also known as cohesion policies), regional, agricultural, social affairs and innovation policies, which provide state-of-the-art technologies to fields such as environmental protection, research and development, and energy. Currently, the EU embraces 27 countries and 490 million people (*http://europa.eu*). EU countries include: Austria-A, Belgium-BE, Bulgaria-BG, Cyprus-CY, the Czech Republic-CZ, Denmark-DK, Estonia-EE, Finland-FI, France-F, Germany-DE, Greece-EL, Hungary-HU, Ireland-IE, Italy-I, Latvia-LV, Lithuania-LT, Luxembourg-LU, Malta-MT, The Netherlands-NL, Poland-PL, Portugal-PT, Romania-RO, Slovakia-SK, Slovenia-SL, Spain-ES, Sweden-SE, and the United Kingdom-UK.

Entrepreneur: An individual who sets up a business and heads a firm.

Family Business: Organizations where two or more extended family members influence the direction of the business (through kinship ties, management roles, ownership rights) (Tagiuri & Davis, 1982).

OECD (Organization for Economic Cooperation and Development): Established in 1961, one of the world's largest and most reliable sources of comparable statistics, and economic and social data. The OECD monitors trends, analyzes and forecasts economic developments, and researches social changes or evolving patterns in trade, environment, agriculture, technology, taxation, and more (*http://www.oecd.org*). OECD countries include: Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, The Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic,

Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States.

Organizational Culture: A pattern of basic assumptions invented, discovered, or developed by a given group as it learns to cope with problems of external adaptation and internal integration, which has worked well enough to be considered valid and therefore is to be taught to new members as the correct way to perceive, think, and feel in relation to those problems (Schein, 1990).

Small- to Medium-Sized Enterprises (SMEs): Independent firms that employ less than 10 (micro), 50 (small), and 250 (medium) employees (U.S. SMEs include firms up to 500 employees).

The Adoption of IS/IT Evaluation Methodologies in Australian Public Sector Organizations

Chad Lin
Curtin University of Technology, Australia

Yu-An Huang
National Chi Nan University, Taiwan

INTRODUCTION

Information systems/information technology (IS/IT) represents substantial financial investment for many organizations (Lin, Huang, & Tseng, 2007; Standing, Guilfoyle, Lin, & Love, 2006). However, IS/IT managers have found it increasingly difficult to justify rising IS/IT expenditures (Lin & Pervan, 2003; Serafeimidis, & Smithson, 2003) and are often under immense pressure to find a way to measure the contribution of their organizations' IS/IT investments to business performance, as well as to find reliable ways to ensure that the business benefits from IS/IT investments are actually realized (Luftman, Kempaiah, & Nash, 2006). This problem has become more complex as the nature of IS/IT investments and the benefits they can deliver have changed rapidly (Murphy & Simon, 2002). Furthermore, evaluation of these IS/IT investments is an extremely complicated process, and it is often avoided or dealt with ineffectively, especially in the public sector (Cilek, Fanko, Koch, Mild, & Taudes, 2004). Given the complexity of the decisions and the large expenditure involved, a better understanding of the basis and practice of IS/IT investment and evaluation in the public sector organizations is essential. The difficulties of evaluation and benefits realization processes are often the determining factors in the application of any formal methodology, and must be addressed if the processes are to be understood (Counihan, Finnegan, & Sammon, 2002; Love, Irani, Standing, Lin, & Burn, 2005).

BACKGROUND

The IS/IT investment evaluation and benefits realization process is a complex but critical function in both private and public organizations. The need to justify expenditure, to assess the effectiveness of a project, and to ensure that expected benefits are eventually delivered are crucial elements in the IS/IT investment evaluation and benefits realization process. The main purpose of IS/IT evaluation is an important factor in determining how the process should be carried out. However, the IS/IT investment evaluation and benefits realization process itself is an extremely complicated and difficult process, and is not often carried out by both private and public organizations.

IS/IT Investment Evaluation and Benefits Realization

The evaluation of the business value of IS/IT investment has been the subject of considerable debate by many academics and practitioners, and the term "productivity paradox" arises from studies that reveal static productivity and rising IS/IT expenditure (Grover, Teng, Segar, & Fiedler, 1998; Tallon, Kraemer, & Gurbaxani, 2000). Despite large investments in IS/IT over many years, it has been difficult to determine where the IS/IT benefits have actually occurred, if indeed there have been any. Some studies have suggested that IS/IT investment produces negligible benefits (e.g., Strassmann (1997)), while others report a positive relationship between organizational performance and IS/IT spending (e.g., Hu & Quan (2005)). The inability of many organizations to assimilate and apply IT both inter- and intra-organizationally is resulting in missed opportunities and a lack of business value (van Grembergen & van Bruggen, 1998).

The difficulties associated with determining the benefits and costs of IT are deemed to be the major constraint to investment justification. Some of the problems associated with IS/IT investment evaluation (Counihan et al., 2002; Lin et al., 2007; Willcocks & Lester, 1997) are:

1. Organizations often fail to identify relevant risks, costs, and benefits;
2. traditional financially oriented evaluation methods (e.g., ROI, NPV) can be problematic in measuring IS/IT investments and quantifying relevant benefits and costs;
3. working with new technology introduces higher levels of risk, which affects timing, costs, and delivery deadlines;

4. organizations have failed to devote appropriate evaluation time and effort to IS/IT, and to deal with the extended investment timeframe; and
5. it is very difficult to evaluate intangibles and make relationship between IS/IT and profitability explicit.

To understand the IS/IT investment evaluation and benefits realization processes, it is important to consider the historical and methodological connection between IS/IT investment evaluation methodologies and IS/IT benefits realization methodologies. IS/IT investment evaluation methodologies have been in use at least since Melone and Wharton (1984, in Farbey, Land, & Targett, 1992), while discussion and adoption of IS/IT benefits realization methodologies appear later in the literature (e.g., Dhillon, 2005; Ward, Taylor, & Bond, 1996). IS/IT investment evaluation methodologies are typically concerned with making investment decisions about IS/IT investments. In other words, the domain of concern is more about selecting the investment or investments that at the outset seem to offer the greatest returns or benefits for the outlay. Early examples of the methodologies emphasize the adoption of accounting indicators such as payback period, ROI, and IRR, whereas later methodologies link the decision-making process more strategically (Lin, Lin, & Tsao, 2005). Some of the formal IS/IT investment evaluation methodologies published in the literature are:

- Return on Management (ROM) (Strassmann, 1990);
- Options theory (Dos Santos, 1994); and
- Kobler Unit framework (Hochstrasser, 1994).

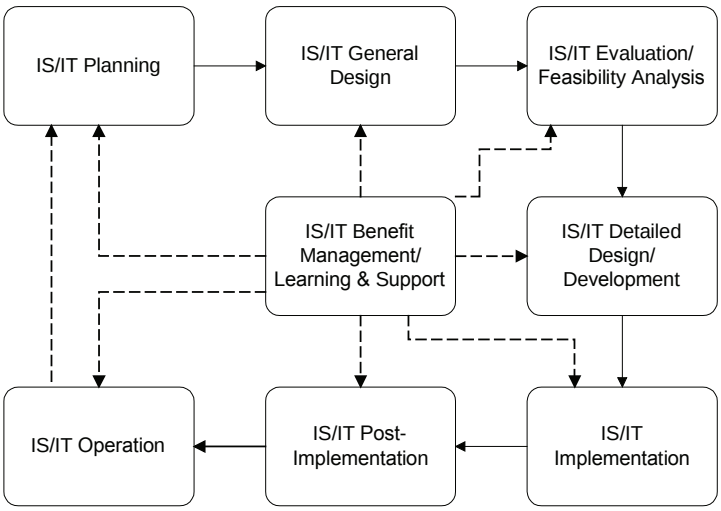
However, IS/IT investment evaluation methodologies alone are insufficient in terms of ensuring that the benefits identified and expected by organizations are realized and delivered (Dhillon, 2005). This is because IS/IT is just one enabler of process change (Grover et al., 1998), and it only enables or creates a capability to derive benefits. IS/IT benefits realization methodologies also need to be adopted by organizations to extend investment evaluation further into the investment lifecycle by ensuring expected benefits are realized once a decision to invest has been taken (Changchit, Joshi, & Lederer, 1998). This involves planning how and when benefits will be realized, and deciding who will be responsible for achieving benefits as well as actually overseeing the realization of benefits (Ward et al., 1996). Some of these formal methodologies published in the literature are:

- Cranfield Process Model of Benefits Management (Ward et al., 1996);
- Active Benefit Realization (ABR) (Remenyi, Sherwood-Smith, & White, 1997); and
- Model of Benefits Identification (Changchit et al., 1998).

AN INTEGRATED APPROACH

There are several reasons why the value of IS/IT cannot be determined by a single measure or methodology. When IS/IT operations are measured as a profit center or as a cost center, significant differences arise and each must show numbers tied to management control (Lin et al., 2007). Senior executives are no longer satisfied to evaluate their IS/IT investments in terms of business performance, but also need to find out where value has arisen in many segments of the organization (Tallon et al., 2000). Therefore, there is a need to integrate IS/IT investment evaluation and benefits realiza-

Figure 1. IS/IT evaluation and benefit realization diagram (adapted from Burch & Grudnitski, 1986; Lin & Pervan, 2003; Willcocks & Lester, 1997)



tion methodologies during the IS/IT systems development process (Lin & Pervan, 2003; Willcocks & Lester, 1997). These two types of methodologies are needed that point to cost effectiveness and containment, as well as identify and deliver expected benefits and value (Love et al., 2005). Figure 1 illustrates the relationship between IS/IT evaluation and IS/IT benefits realization during the IS/IT system development process.

In addition, in order to make this integrated evaluation approach work, it is important to involve the stakeholders in processes that operationalize the evaluation criteria and techniques (Lin et al., 2005) – that is, to involve the stakeholders in processes that "breathe life into, adapt over time, and act upon" the evaluation criteria and techniques (Willcocks & Lester, 1997). This is an important issue for both researcher and practitioner, because evaluation of IS/IT benefits without an assessment of all relevant stakeholder benefits is incomplete, and senior executives need information not only for evaluating IS/IT benefits but also for managing the IS/IT investments and capturing the benefits in the bottom line (Dhillon, 2005; Murphy & Simon, 2002).

RESEARCH APPROACH

However, there is much evidence to suggest that the integration of IS/IT investment evaluation and benefits realization processes rarely exists in organizations (Lin & Pervan, 2003; Murphy & Simon, 2002; Willcocks & Lester, 1997). Therefore, three case studies were conducted to investigate the practices of IS/IT investment evaluation and benefits realization in large Australian public sector organizations. Semi-structured interviews were used to gain a deeper understanding of issues.

CASE DESCRIPTION

Thirty-five interviews were conducted with participants from three Australian public sector organizations and their six major IS/IT contractors. The questions were related to the formal benefits realization methodology adopted, major IS/IT contracts, contractual relationship between these public sector organizations and contractors, and IS/IT investment evaluation methodology or technique deployed. Other data collected included contract documents, planning documents, and minutes of relevant meetings. Around 200 pages of transcripts were coded and analyzed. The analysis was conducted in a cyclical manner and followed guidelines for interpretive research set out by Klein and Myers (1999).

CASE STUDY RESULTS

A number of issues arose from the analysis of this text data; the key issues are presented below.

Issue 1: Lack of Formal IS/IT Investment Evaluation Methodology

Most of the participants claimed that some sort of formal methodology or process was put in place for evaluating these contracts. However, closer examination revealed that what was described did not constitute a formal IS/IT investment evaluation methodology. Participants wrongly considered various contract control mechanisms as a formal IS/IT investment evaluation methodology.

Issue 2: Lack of Understanding of IS/IT Investment Evaluation Methodology

The confusion indicated in Issue 1 about what constitutes a formal IS/IT investment evaluation methodology demonstrated a lack of understanding of such methodologies. This may be due to the fact that these public sector organizations were unable to introduce a formal IS/IT investment evaluation methodology because it was required to follow the public sector's IS/IT investment guidelines.

Issue 3: Existence of an Informal IS/IT Investment Evaluation Process

Despite the fact that no formal IS/IT investment evaluation methodology or process was adopted, contract control and evaluation mechanisms specified within the service level agreements (SLAs) or government guidelines do represent an informal IS/IT investment evaluation process. Although these informal mechanisms may not totally replace a formal methodology (e.g., Kobler Unit framework; Hochstrasser, 1994), they were able to assist in evaluating the performance of the outsourcing contracts. These mechanisms were largely based on the standard public sector contract process and purchasing guidelines.

Issue 4: Focus on Quantitative IS/IT Investment Evaluation Measures

Most measures contained within the SLAs were quantitative in nature. Very few provisions were put in place to identify and assess the more qualitative measurements. Without employing more qualitative measures (e.g., customer satis-

faction, relationship, culture, and leadership) and a formal IS/IT investment evaluation methodology or process, the adoption of quantitative or accounting-based measures alone did not assist in full evaluation and monitoring of the performance.

Issue 5: Different Motivations for Seeking External Expertise

Several reasons were put forward as the main motivation for seeking external IS/IT expertise. Only a few of the external contractor representatives cited access to the required technical expertise as one of these public sector organizations' reasons to seek external expertise. However, all of the public sector organizations mentioned access to required technical expertise as a major reason to externalize some of their IS/IT functions. Therefore, these public sector organizations' motivation for outsourcing was somewhat different from the contractors'.

Issue 6: Success of the Contracts Perceived Differently by Stakeholders

Customer satisfaction, achieving the contractor's projected revenue, bringing value/benefits to the organization, and meeting the SLAs' provisions were mentioned. Other criteria mentioned included technical competence to deliver what is required, risk factors, contractors' experience in a relevant area, and business continuity of the contractors. Most of the external contractors mentioned achieving the projected revenue for themselves and satisfying customers as their only criteria for determining the success of their outsourcing contracts with these public sector organizations. This may indicate that most of the external contractors' aim is to maximize the profit while maintaining a certain level of customer satisfaction. However, participants from these public sector organizations seemed to have used different criteria for determining the success of the outsourcing contracts. Bringing value/benefits to the organizations, meeting the SLA provisions, and pricing/cost were mentioned by most participants.

Issue 7: Embedded Contract Mentality

Staff of these public sector organizations seemed to have a "contract mentality," as the operation of the contracts was based on the specifications set out in the SLAs within the outsourcing contracts. Most participants clearly indicated that there was a pre-agreed set of evaluation and control mechanisms in the SLAs within the outsourcing contracts such as metrics, monthly reports, reviews, and regular meetings. Moreover, most of them thought these contract control mechanisms were all part of the IS/IT investment evaluation methodology.

Issue 8: Lack of User Involvement/Participation in Contract Development

There appeared to be an organizational memory gap where units within these public sector organizations possessed knowledge of different parts of the IS/IT systems development cycle. However, the knowledge did not seem to be shared by all units because different units participated in different stages. These public sector organizations' IS/IT contract processes may have been more successful if the participants were all involved in the original tendering and contracts negotiation, as well as benefits realization processes.

Issue 9: General Lack of Commitment by Contractors

It was easy to see that these public sector organizations and contractors had different agendas in mind, despite the fact that the contract was a partnership-type arrangement. The external contractors' criteria for determining success of these IS/IT contracts seemed to be maximization of profit/revenue while keeping the customers satisfied to a certain extent. The contractors' lack of commitment could also be demonstrated by the fact that they either did not know why these public sector organizations externalize their IS/IT functions, and did not really care whether these public sector organizations had benefited from their services and expertise. However, these public sector organizations, in general, had agreed that access to the required technical expertise was a major reason for seeking external IS/IT expertise.

FUTURE TRENDS

Carr (2003) has argued that IT has become a commodity because it has become widespread, as has happened to other innovations such as engines and telephones. According to Carr (2003), IT has become an infrastructural technology and therefore is often subject to over-investment and causes economic troubles such as the "Internet Bubble." Carr's (2003) views on IT are not shared by many IS/IT practitioners and academics who argue that IT still has a lot to offer in the future and can deliver competitive advantages to both private and public organizations (Hayes, 2003).

However, recent evidence suggests that many private and public organizations simply got carried away with IS/IT and spent money unwisely during the last two decades (Farrell, 2003). According to a study by the McKinsey Global Institute, more successful organizations analyzed their economics carefully and spent on only those IS/IT applications that would deliver productivity gains, sequencing their investments carefully through a disciplined approach with innovative management practices (Farrell, 2003).

CONCLUSION

Case studies were conducted in three public sector organizations that sought IS/IT services and expertise from external contractors. While these public sector organizations appear to operate without any major problem, the mostly negative issues shown above indicate weaknesses in the way IS/IT deals with the level of formality and integration in applying the methodologies. The problems mentioned above were mostly caused by the lack of attention to IS/IT investment evaluation. For example, if formal IS/IT investment evaluation was adopted by these organizations, more qualitative measures may have been used to evaluate the IS/IT contracts.

So why did these organizations not formally evaluate their IS/IT investments? One possible explanation was that the restrictive nature of the public sector's IS/IT contract guidelines made it difficult to implement a formal IS/IT investment evaluation methodology. Another explanation was that none of the IS/IT staff was familiar with the formal IS/IT investment evaluation process and hence possessed an "embedded contract mentality" by simply following conditions set out within the SLAs. Seddon, Graeser, and Willcocks (2002) suggest that under some circumstances, cost of formal IS/IT evaluations may seem likely to exceed benefits. However, the results from the case studies indicate that the use of a benefits realization methodology enabled greater control and better management of IS/IT contracts.

Despite large investments in IS/IT over many years, it has been difficult for organizations to determine where benefits have occurred, if indeed there have been any. IS/IT investment evaluation practice remains a hotly debated topic in the IS literature. Little published work has been conducted in Australia and there is still a lot to be learned in the area of processes and practices of IS/IT investment evaluation and benefits management. We hope that more studies of the practice of IS/IT investment evaluation will benefit other researchers in this field and in the public sector, in particular. Through the case study results presented in this article, it is hoped that better approaches may be developed for Australian public sector organizations.

REFERENCES

Burch, J.G., & Grudnitski, G. (1986). *Information systems: Theory and practice* (4th ed.). New York: John Wiley & Sons.

Carr, N.G. (2003). IT doesn't matter. *Harvard Business Review, 8*(1), 4-50.

Changchit, C., Joshi, K.D., & Lederer, A.L. (1998). Process and reality in information systems benefit analysis. *Information Systems Journal, 8,* 145-162.

Cilek, P., Fanko, W., Koch, S., Mild, A., & Taudes, A. (2004). A hedonic wage model-based methodology for evaluating the benefits of IT investments in public-sector organizations. *Journal of Enterprise Information Management, 17*(4), 269-275.

Counihan, A., Finnegan, P., & Sammon, D. (2002) Towards a framework for evaluating investments in data warehousing. *Information Systems Journal, 12,* 321-338.

Dhillon, G. (2005). Gaining benefits from IS/IT implementation: Interpretations from case studies. *International Journal of Information Management, 25*(6), 502-515.

Dos Santos, B.L. (1994). Assessing the value of strategic information technology investments. In L. Willcocks (Ed.), *Information management: The evaluation of information systems investments* (pp. 133-148). London: Chapman & Hall.

Farbey, B., Land, F., & Targett, D. (1992). Evaluating investments in IT. *Journal of Information Technology, 7,* 109-122.

Farrell, D. (2003). The real new economy. *Harvard Business Review, 81*(10), 104.

Grover, V., Teng, J., Segar, A.H., & Fiedler, K. (1998). The influence of information technology diffusion and business process change on perceived productivity: The IS executive's perspective. *Information and Management, 34,* 141-159.

Hayes, F. (2003, May 19). *IT delivers.* Retrieved from http://www.computerworld.com/printthis/2003/0,4814,81278,00.html

Hochstrasser, B. (1994). Justifying IT investments. In L. Willcocks (Ed.), *Information management: The evaluation of information systems investments* (pp. 151-169). London: Chapman & Hall.

Hu, Q., & Quan, J.J. (2005). Evaluating the impact of IT investments on productivity: A causal analysis at industry level. *International Journal of Information Management, 5*(1), 39-53.

Klein, H.K., & Myers, M.D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly, 23*(1), 67-94.

Lin, C., & Pervan, G. (2003). The practice of IS/IT benefits management in large Australian organizations. *Information and Management, 41*(1), 13-24.

Lin, C., Huang, Y., & Tseng, S. (2007). A study of planning and implementation stages in electronic commerce adoption and evaluation: The case of Australian SMEs. *Contemporary Management Research, 3*(1), 83-100.

Lin, K., Lin, C., & Tsao, H. (2005). IS/IT investment evaluation and benefit realization practices in Taiwanese SMEs. *Journal of Information Science and Technology, 2*(4), 44-71.

Love, P.E.D., Irani, Z., Standing, C., Lin, C., & Burn, J. (2005). The enigma of evaluation: Benefits, costs and risks of IT in small-medium sized enterprises. *Information and Management, 42*(7), 947-964.

Luftman, J., Kempaiah, R., & Nash, E. (2006). Key issues for IT executives 2005. *MIS Quarterly Executive, 5*(2), 27-45.

Murphy, K.E., & Simon, S.J. (2002). Intangible benefits valuation in ERP projects. *Information Systems Journal, 12,* 301-320.

Remenyi, D., Sherwood-Smith, M., & White, T. (1997). *Achieving maximum value from information systems: A process approach.* Chichester, England: John Wiley & Sons.

Seddon, P., Graeser, V., & Willcocks, L. (2002). Measuring organizational IS effectiveness: An overview and update of senior management perspectives. *The DATA BASE for Advances in Information Systems, 33*(2), 11-28.

Serafeimidis, V., & Smithson, S. (2003). Information systems evaluation as an organizational institution—experience from a case study. *Information Systems Journal, 13*(2), 251-274.

Standing, C., Guilfoyle, A., Lin, C., & Love, P.E.D. (2006). The attribution of success and failure in IT projects. *Industrial Management and Data Systems, 106*(8), 1148-1165.

Strassman, P.A. (1990). *The business value of computers.* New Canaan, CT: Information Economics Press.

Tallon, P.P., Kraemer, K.L., & Gurbaxani, V. (2000). Executives' perceptions of the business value of information technology: A process-oriented approach. *Journal of Management Information Systems, 16*(4), 145-173.

Truax, J. (1997). Investing with benefits in mind: Curing investment myopia. *The DMR White Paper,* 1-6.

van Grembergen, W., & van Bruggen, R. (1998). Measuring and improving corporate information technology through the balanced scorecard. *Electronic Journal of Information Systems Evaluation, 1*(1). Retrieved from http://is.twi.tudelft.nl/ejise/indpap.html

Ward, J., Taylor, P., & Bond, P. (1996). Evaluation and realization of IS/IT benefits: An empirical study of current practice. *European Journal of Information Systems, 4,* 214-225.

Willcocks, L., & Lester, S. (1997). Assessing IT productivity: Any way out of the labyrinth? In L. Willcocks, D.F. Feeny, & G. Islei (Eds.), *Managing IT as a strategic resource* (pp. 64-93). London: McGraw-Hill.

KEY TERMS

Benefits Management: A managed and controlled process of checking, implementing, and adjusting expected results, and continuously adjusting the path leading from investments to expected business benefits.

Information Technology (IT): Any computer-based tool used to work with information and support the information needs of an organization.

IS/IT Benefits Realization Methodologies: Approaches used to ensure that benefits expected in the IS/IT investments by organizations are realized or delivered.

IS/IT Investment Evaluation: The weighing up process to rationally assess the value of any acquisition of software or hardware that is expected to improve the business value of an organization's information systems.

IS/IT Investment Evaluation Methodologies: Approaches used to evaluate organizations' IS/IT investments.

Productivity Paradox: Despite large investments in IS/IT over many years, there have been conflicting reports as to whether or not IS/IT benefits have actually occurred.

Systems Development Process: A structured step-by-step approach for developing IS/IT.

Advanced Techniques for Object-Based Image Retrieval

A

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

Along with the progress of imaging modality and the wide utility of digital images (including video) in various fields, many potential content producers have emerged, and many image databases have been built. Because images require large amounts of storage space and processing time, how to quickly and efficiently access and manage these large, both in the sense of information contents and data volume, databases has become an urgent problem. The research solution for this problem, using content-based image retrieval (CBIR) techniques, was initiated in the last decade (Kato, 1992). An international standard for multimedia content descriptions, MPEG-7, was formed in 2001 (MPEG). With the advantages of comprehensive descriptions of image contents and consistence to human visual perception, research in this direction is considered as one of the hottest research points in the new century (Castelli, 2002; Zhang, 2003; Deb, 2004).

Many practical retrieval systems have been developed; a survey of near 40 systems can be found in Veltkamp (2000). Most of them mainly use low-level image features, such as color, texture, and shape, etc., to represent image contents. However, there is a considerable difference between the users' interest in reality and the image contents described by only using the above low-level image features. In other words, there is a wide gap between the image content description based on low-level features and that of human beings' understanding. As a result, these low-level feature-based systems often lead to unsatisfying querying results in practical applications.

To cope with this challenging task, many approaches have been proposed to represent and describe the content of images at a higher level, which should be more related to human beings' understanding. Three broad categories could be classified: synthetic, semantic, and semiotic (Bimbo, 1999; Djeraba, 2002). From the understanding point of view, the semantic approach is natural. Human beings often describe image content in terms of objects, which can be defined at different abstraction levels. In this article, objects are considered not only as carrying semantic information in images, but also as suitable building blocks for further image understanding.

The rest of the article is organized as follows: in "Background," early object-based techniques will be briefly reviewed, and the current research on object-based techniques will be surveyed. In "Main Techniques," a general paradigm for object-based image retrieval will be described; and different object-based techniques, such as techniques for extracting meaningful regions, for identifying objects, for matching semantics, and for conducting feedback are discussed. In "Future Trends," some potential directions for further research are pointed out. In "Conclusion," several final remarks are presented.

BACKGROUND

Early Object-Based Techniques in Content-Based Image Retrieval

CBIR techniques are distinguished from traditional retrieval techniques by many aspects. Two of the most pertinent are that CBIR is a somehow subjective process, as for a given image, its means may have different interpretations for different users; and image retrieval is often a computationally expensive process, as the image database is often large in size and contains heterogeneous information. Due to these particular aspects, the results of CBIR could not be judged objectively—human perception should be considered. In addition, performing an exhaustive search for finding optimal solutions in CBIR is not feasible, and therefore, some suboptimal solutions will be chosen.

Because of the unique aspects of CBIR, object-based representation and description must be used even in so-called low-level feature-based image retrieval, though in these works, object recognition is not evidently performed and semantic information is not explicitly searched.

One typical example is in shape-based retrieval, as the shape features are generally extracted from individual objects (Latecki, 2002). In contrast, color features and textural features are often obtained by taking the whole image as a unit. From this point of view, shape-based retrieval is already at some higher level than color-based retrieval and texture-based retrieval (Zhang, 2003).

Structural query model is another instance in which partial matches are allowed and outputs related to the score of similarity can be provided. This type of retrieval is based on the relations between the individual objects and compo-

nents in images (Zhou, 2001). In query by visual sketch, users sketch a scene by drawing a collection of objects. (It is assumed that these objects could fully define a scene.) For example, the objects are first identified and then used in a search (Chang, 1998).

Current Object-Based Techniques in Content-Based Image Retrieval

Currently, researchers seek explicit semantics and use the high-level descriptions that are common to humans, such as articles, people, places, and things. It is generally accepted that high-level features are crucial to improve the performance of CBIR up to so-called semantic-based querying. For this purpose, object-based content analysis, especially segmentation that segments the semantically meaningful objects from images, is an essential step (Zhang, 2001).

Complete image understanding should start at interpreting image objects and their relationships. Objects can be further identified in line with appropriate knowledge. For example, some object grammars based on rules for concept inference have been proposed (Petkovic, 2003). When domain knowledge is available, objects can be classified even without the explicit determination of object regions (Li, 2002b).

To extract high-level descriptions from images and to fill the gap between the low-level features and human beings' understanding of image contents, techniques to describe the whole image with a hierarchical structure to reach progressive image analysis are proposed (Castelli, 1998; Jaimes, 1999; Hong, 1999). The contents of images can be represented in different levels (Amir, 1998), such as the three-level content representation, including feature level content, object level content, and scene level content (Hong, 1999); and the five-level representation, including region level, perceptual region level, object part level, object level, and scene level (Jaimes, 1999). The problem here is how to implement these levels efficiently and effectively.

Another direction for extracting semantics information from an image is to map low-level visual features to high-level semantics. In other words, to fill the semantic gap, one makes the retrieval system work with low-level features, while the user puts in more high-level knowledge (Zhou, 2002). Two typical methods are to optimize query requests by using relevance feedback and semantic visual templates (Chang, 1998) and to interpret progressively the content of images by using interactive interfaces (Castelli, 1998). In both approaches, relevance feedback plays an important role, as humans are much better than computers at extracting semantic information from images (Rui, 1998; Ciocca, 1999).

MAIN TECHNIQUES FOR OBJECT-BASED IMAGE RETRIEVAL

A General Paradigm

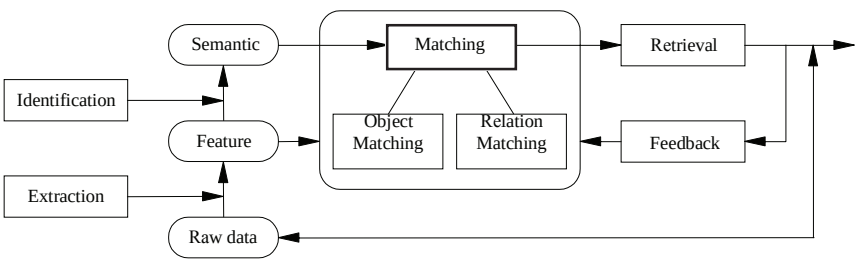
In general, people distinguish three levels of abstraction when talking about image databases: raw data level, feature level, and semantic level. The raw data are original images in the form of a pixel matrix. The feature level shows some significant characteristics of the pixel patterns of the image. The semantic level describes the meanings of identified objects in images. Note that the semantic level should also describe the meaning of an image as a whole. Such a meaning could be obtained by the analysis of objects and the understanding of images.

According to the above discussions, a multilayer approach should be used for efficiently treating image data. Though the number of layers and the definitions and functions of these layers could have some variations in different approaches, some principle steps are common for object-based image retrieval. A general paradigm is shown in Figure 1.

First, objects should be determined. Two important tasks are as follows:

1. Extract meaningful regions: To be able to base the image retrieval on objects, the interesting regions related to objects should be extracted first. This process relates the raw data level to the feature level.

Figure 1. A general paradigm for object-based image retrieval



2. Identify interesting objects: Based on the extracted regions, (perceptual) features should be taken out, and those required objects could be identified. This corresponds to the step from feature level to object level.

Once the objects in images are identified, further retrieval can be carried out by using objects as primitives. Two tasks are as follows:

1. Matching identified objects: For each identified object, suitable properties and attributes should be selected for proper description. The matching between images, such as object matching and relation matching, is then carried out.
2. Performing feedback in retrieval: This is to introduce human intelligence and to incorporate human semantics into the retrieval process.

The following sections introduce some techniques developed for each task.

Techniques for Extracting Region of Interest

Extraction of interesting regions from an image is, in general, called image segmentation (Zhang, 2001). Image segmentation is one of the most critical tasks in automatic analysis of image contents. A great variety of segmentation algorithms has been proposed in the literature. One should note that none of the proposed segmentation algorithms is generally applicable to all images, and different algorithms are not equally suitable for a particular application. This is the reason that though several thousands of algorithms have been developed, much attention and new efforts are continuously made on improving and perfecting them.

With the progress in segmentation techniques, people also realized that precise segmentation of objects in many cases is still beyond the capability of current computer techniques. On the other side, compared to some image analysis tasks that aim to obtain accurate measurements from the segmented objects, the requirement for precise segmentation of objects can be somehow relaxed in the context of image retrieval. Image retrieval is a subject-oriented process in which the precise object measurement is not the goal. In addition, for object-based image retrieval, the purpose of segmentation is for identifying the objects.

One idea derived from the above considerations is to extract approximately the so-called "meaningful region," instead of to segment the object accurately (Luo, 2001). The "meaningful region" provides an effective visual representation of objects from the point of view of object recognition. Though the "meaningful region" is not an exact representation of the objects, however, based on some domain knowledge, the semantic meaning of objects can still be recovered. On the other side, robust extraction of "meaningful regions" is easy to accomplish. This makes the object-based image retrieval with the extraction of "meaningful region" a feasible approach (Gao, 2000).

Another idea derived from the above considerations uses a particular matching procedure to reduce the requirement for precise object segmentation (Dai, 2004). The images are segmented both in a rough version and in a detailed version. The rough one is the merge result of several detailed ones and is less spatial-constrained by either oversegmentation or undersegmentation.

Techniques for Identifying Objects

From extracted regions, some perceptual features could be obtained. This can help to "recognize" what they represent in terms of human beings' perceptions. One iterative procedure uses this principle and transforms the recognition to a training-and-testing procedure (Gao, 2000). To make the problem simpler, it is assumed that there are finite types of interesting objects in a given image database. In fact, this requirement could often be satisfied in practice, as only limited image contents are considered in one application. The object recognition is then performed in an iterative way. Context-based information would be obtained during this process, helping to reach the correct recognition result.

To capture different aspects of images, multiple features are often used. The proportion of each feature in a description would be determined by training. Due to the lighting conditions and variety of object appearance, the objects belonging to the same category can have different visual aspects. To solve this problem, the most significant objects in the training set would be selected, and the trivial ones would be discarded. This recognition process can be iterated until the final recognition result is acceptable (in terms of the image composition, according to some *a priori* knowledge). With this iterative procedure, the context-based knowledge is gradually improved, and the correct recognition result is gradually approached.

Techniques for Matching Objects

Matching is one important task in image retrieval, which consists of comparison and judgment. As retrieval is a subjective process, so the decision is often made according to similarity (the distance to be reasonably small) but not according to equivalence (identical). Based on matching score, a database search can be carried out, and required images can be retrieved. Object match is more direct than feature match is in image retrieval.

One procedure is to describe the object in an image by an $M \times M$ matrix, where M is the number of all objects in the image database. It is a diagonal matrix, with each entry

indicating the attribute of every meaningful region in the image. In addition, a relation matrix is also defined. It is a $K \times K$ matrix to indicate the spatial relationship between every two meaningful regions, with K representing the number of meaningful regions in the whole image. The object-matching procedure is dependent on a decision function. This decision function is determined based on the correlation among all content description matrices. It reflects whether all the relevant images have common contents, that is, the same objects. In the case that the relevant images have common objects, the match will be based on the objects of images. In the case that the relevant images do not have common objects, the match will be based on the features of images. For object matching, the similarity information from the common objects in all relevant images will be extracted to perform the matching between the relevant images and candidate images. Details can be found in Zhang (2004).

For compounded objects, matching might be accidentally performed among different parts. To solve this problem, a two-level matching is proposed (Dai, 2004). The principal idea is to describe the query images at a relatively rough scale and to describe the database images at some more detailed scales. As the rough description is based on the merging of detailed descriptions, the matching process will be carried on in an uneven way, and the minor errors caused by segmentation will be recompensed by the approximate matching procedure.

Techniques for (Interactive) Feedback

Feedback plays an important role, especially in high-level retrieval. As indicated above, retrieval is a subjective process, so feedback is required to combine the information from users, or in other words, to incorporate human knowledge and requirements. Retrieval is also a progressive process, so feedback is required to introduce interaction and to turn the search direction to follow the user's intention.

A self-adaptive relevance feedback technique has been used in an object-based image retrieval system (Gao, 2001). In such a system, objects are first identified, the relevance feedback relying on the high-level attributes could better catch image semantics, and the retrieval results are refined according to users' wishes in an explicit manner. In practice, to make the querying more convenient for the user, the procedure of feedback could be directed, also based on high-level information, without memory or with memory to make the feedback mechanism more flexible. In the former case, each feedback is an independent procedure, in which all of the relevant images selected in previous iterations would be ignored. In the latter case, the relevant image selected in previous iterations would be taken into account in the current iteration. A time-delay curve has also been proposed to simulate human beings' memory mechanisms in feedback with memory (Gao, 2001). The main idea of the proposed scheme is to analyze the fed-back relevant images marked by the user in different levels to reach comprehensive similarity analysis.

Another approach called association feedback has been proposed (Xu, 2001b). Feature elements are first defined that can be considered a type of perceptual primitives with abstraction levels located between that of raw images and that of objects in images (Xu, 2001a). These feature elements, different from commonly used feature vectors, have obvious intuitive visual senses and are relatively independent from each other physically. A selection mechanism called feature element evaluation is also proposed, which tries to find those feature elements that are closer to the interest of people by visual meaning. A group of feature elements can be used to represent compound objects. Based on feature elements, association feedback can be applied. In contrast to the weighting adjustment in relevance feedback, here the associated relations between different feature elements are counted. New sets of feature elements can thus be formed during the retrieval process, and this property is suitable for handling the so-called "interest switch" cases. In other words, the search for images can be controlled by users with the introduction of new feature elements according to the change of interest, and the search direction will be guided toward new goals.

FUTURE TRENDS

Further research can be considered in the concrete techniques for advanced image analysis and along the general research movements for image understanding.

To perform object-based image retrieval, different image analysis techniques are to be enhanced:

1. Improving the robustness of meaningful region extraction, especially with complicated images, by taking more characteristics of images into consideration
2. Describing objects as congruous to humans' sense as possible—as human beings are still far from knowing all the cognitive details from the real world, how to automatically form semantic objects is a challenging task
3. Using more efficient feedback procedures to make the search process fast following users' aspirations in the course of retrieval

In the "Background" section, two generations of object-based techniques are discussed. From the point of view of image understanding, the next stage would go beyond objects, though the third generation will still be based on objects. The actions and interactions of objects and thus generated events (or scenes) are important to fully understand the contents of images. The images would be, in this case, described by

some metadata. The event detection and event retrieval have already played an important role in many applications (e.g., surveillance, war, etc.). However, only a few particular works are made now, and they are mainly based on audio and multiple frames (Li, 2002a). Further research in this direction, taking more advantage of human knowledge for constructing more intelligent image data would be promising.

CONCLUSION

Object-based techniques can fulfill many roles and tasks required by CBIR. Three generations of object-based techniques are reviewed and discussed in this article. Some of them have already made their contributions to the advancement of CBIR, and some of them need to be improved, and developing new object-based techniques for CBIR is even required.

The object-based techniques discussed here are mainly focused on CBIR. As content-based video retrieval (CBVR) appears like a natural combination of CBIR and content-based audio retrieval (CBAR), as well as some extensions along the temporal axis, many of these techniques would also be applicable for CBVR.

REFERENCES

Amir, A., & Lindenbaum, M. (1998). A generic grouping algorithm and its quantitative analysis. *IEEE PAMI, 20*(2), 168–185.

Bimbo, A. (1999). *Visual information retrieval*. San Francisco, CA: Morgan Kaufmann (Elsevier).

Castelli, V., Bergman, L. D., & Kontoyiannis, I. et al. (1998). Progressive search and retrieval in large image archives. *IBM J. Res. Develop., 42*(2), 253–268.

Chang, S. F., Chen, W., & Sundaram, H. (1998). Semantic visual templates: Linking visual features to semantics. In *Proceedings ICIP'98* (pp. 531–535).

Ciocca, G., & Schettini, R. (1999). Using a relevance feedback mechanism to improve content-based image retrieval. In *Proceedings of the Third International Conference, VISUAL'99* (pp. 107–114).

Dai, S. Y., & Zhang, Y. J. (2004). Unbalanced region matching based on two-level description for image retrieval. *Pattern Recognition Letters*.

Deb, S. (2004). *Multimedia systems and content-based image retrieval*. Hershey, PA: Idea Group Publishing.

Djeraba, C. (2002). Content-based multimedia indexing and retrieval. *IEEE, Multimedia*, (2), 18–22.

Gao, Y. Y., Zhang, Y. J., & Merzlyakov, N. S. (2000). Semantic-based image description model and its implementation for image retrieval. In *Proceedings of the First International Conference on Image and Graphics* (pp. 657–660).

Gao, Y. Y., Zhang, Y. J., & Yu, F. (2001). Self-adaptive relevance feedback based on multi-level image content analysis. In *SPIE Proceedings Storage and Retrieval for Media Databases 2001* (Vol. 4315, pp. 449–459).

Hong, D. Z., Wu, J. K., & Singh, S. S. (1999). Refining image retrieval based on context-driven method. In *SPIE Proceedings Storage and Retrieval for Image and Video Database VII* (Vol. 3656, pp. 581–593).

Jaimes, A., & Chang, S. F. (1999). Model-based classification of visual information for content-based retrieval. In *SPIE Proceedings on Storage and Retrieval for Image and Video Database VII* (Vol. 3656, pp. 402–414).

Kato, T. (1992). Database architecture for content-based image retrieval. *SPIE* (Vol. 1662, pp. 112–123).

Latecki, L. J., Melter, R., & Gross, A. (2002). Shape representation and similarity for image database [Special Issue]. *Pattern Recognition, 35*(1), 1–297.

Li, B. X., & Sezan, I. (2002a). Event detection and summarization in American football broadcast video. In *SPIE Proceedings Storage and Retrieval for Media Databases 2002* (Vol. 4676, pp. 202–213).

Li, Q., Zhang, Y. J., & Dai, S. Y. (2002b). Image search engine with selective filtering and feature element based classification. In *SPIE Proceedings Internet Imaging III* (Vol. 4672, pp. 190–197).

Luo, Y., Zhang, Y. J., & Gao, Y. Y. et al. (2001). Extracting meaningful region for content-based retrieval of image and video. In *SPIE, Proceedings Visual Communications and Image Processing* (Vol. 4310, pp. 455–464).

MPEG. Retrieved from *http://www.cselt.it/mpeg/*

Petkovic, M., & Jonker, W. (2003). *Content-based video retrieval: A database perspective*. Dordrecht: Kluwer.

Rui, Y., Huang, T. S., & Mehrotra, S. (1998). Relevance feedback techniques in interactive content-based image retrieval. In *SPIE Proceedings Storage and Retrieval for Image and Video Database V* (Vol. 3312, pp. 25–34).

Veltkamp, R. C., & Tanase, M. (2000). Content-based image retrieval systems: A survey. Technical Report, UU-CS-2000-34, Utrecht University, The Netherlands.

Xu, Y., & Zhang, Y. J. (2001a). Image retrieval framework driven by association feedback with feature element evaluation built in. In *SPIE Proceedings Storage and Retrieval for Media Databases 2001* (Vol. 4315, pp. 118–129).

Xu, Y., & Zhang, Y. J. (2001b). Association feedback: A novel tool for feature elements based image retrieval. In *Lecture Notes in Computer Science 2195* (pp. 506–513).

Zhang, Y. J. (2001). *Image segmentation*. Beijing: Science Publisher.

Zhang, Y. J. (2003). *Content-based visual information retrieval*. Beijing: Science Publisher.

Zhang, Y. J., Gao, Y. Y., & Luo, Y. (2004). Object-based techniques for image retrieval. In *Multimedia systems and content-based image retrieval* (Chap. 7, pp. 154–179). Hershey, PA: Idea Group Publishing.

Zhou, X. S., & Huang, T. S. (2001). Edge-based structural features for content-based image retrieval. *Pattern Recognition Letters, 22*(5), 457–468.

Zhou, X. S., & Huang, T. S. (2002). Unifying keywords and visual contents in image retrieval. *IEEE, Multimedia,* (2), 23–33.

KEY TERMS

Content-Based Image Retrieval (CBIR): A process framework for efficiently retrieving images from a collection by similarity. The retrieval relies on extracting the appropriate characteristic quantities describing the desired contents of images. In addition, suitable querying, matching, indexing, and searching techniques are required.

Feature-Based Image Retrieval: A branch of CBIR that is based on specific visual characteristics called "features" and is considered at a low abstraction level. Features are commonly refereed to perceptive attributes of images, such as color, texture, shape, etc., of images.

Intelligent Image Data: A data format that embeds pixel information of images as well as higher-level information, such as indices and semantic information. This format is self-descriptive in the sense that data could explain by themselves what contents are inside and present and retrieve the related and interested portions for the users.

Metadata: A data format that may contain numerous information, including information obtained indirectly from the image, as well as information related to the actual description of the image content. At the highest level, images are often accompanied and associated by metadata.

MPEG-7: This is an international standard named "multimedia content description interface" (ISO/IEC 15938). It provides a set of audiovisual description tools, descriptors, and description schemes for effective and efficient access (search, filtering, and browsing) to multimedia content.

Query Model: An abstraction model for image querying in the context of CBIR. In this model, a submitted query would specify both a filter condition and a ranking expression, while the query result will be a rank of the images that satisfies the filter condition, according to grade of match for the ranking expression.

Semantic-Based Image Retrieval: A branch of CBIR based on descriptions with semantic meaning and considered at a high abstraction level. Semantic descriptions are more closely related to the human interpretation and understanding of images.

Semantic Gap (SG): The discrepancy between the perceptual property and semantic meaning of images in the context of CBIR. As the perceptual properties are usually described by low-level visual features that can be easily treated by computers, and the semantic meanings are commonly related to high-level object-based descriptions that are familiar to human beings, the semantic gap is also considered a gap between current techniques and human requirements.

Semiotics: The science that analyzes signs and sign systems and puts them in correspondence with particular meanings. It provides formal tools for image knowledge acquisition, generation, representation, organization, and utilization in the context of CBIR.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 68-73, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Advances in Tracking and Recognition of Human Motion

Niki Aifanti
Informatics & Telematics Institute, Greece

Angel D. Sappa
Computer Vision Center, Spain

Nikos Grammalidis
Informatics & Telematics Institute, Greece

Sotiris Malassiotis
Informatics & Telematics Institute, Greece

INTRODUCTION

Tracking and recognition of human motion has become an important research area in computer vision. In real world conditions it constitutes a complicated problem, considering cluttered backgrounds, gross illumination variations, occlusions, self-occlusions, different clothing and multiple moving objects. These ill-posed problems are usually tackled by making simplifying assumptions regarding the scene or by imposing constraints on the motion. Constraints such as that the contrast between the moving people and the background should be high and that everything in the scene should be static except for the target person are quite often introduced in order to achieve accurate segmentation. Moreover, the motion of the target person is often confined to simple movements with limited occlusions. In addition, assumptions such as known initial position and posture of the person are usually imposed in tracking processes.

BACKGROUND

The first step towards human tracking is the segmentation of human figures from the background. This problem is usually addressed either by exploiting the temporal relation between consecutive frames (e.g., background subtraction (Sato & Aggarwal, 2001), optical flow (Okada, Shirai & Miura, 2000)), by modeling the image statistics of human appearance (Wren, Azarbayejani, Darrell & Pentland, 1997) or by exploiting the human shape (Leibe, Seemann & Schiele, 2005). Efficient texture-based methods for modeling the background and detecting moving objects from a video sequence have been developed as well (Heikkila & Pietikainen, 2006), while some other recent research copes with the problem of occlusions (Capellades, Doermann, DeMenthon & Chellappa, 2003). The output of the segmentation, which could be edges, silhouettes, blobs, and so forth, comprises the basis for feature extraction.

Feature correspondence is established in order to track the subject. Tracking through consecutive frames commonly incorporates prediction of movement, which ensures continuity of motion especially when some body parts are occluded. For example, when a person is walking there are some moments when one of the legs occludes the other. Furthermore, there are scenes with multiple persons occluding one another. Depending on the scene and the chosen methodology, some techniques try to determine the precise movement of each body part (Sidenbladh, Black, & Sigal, 2002), while other techniques focus on tracking the human body as a whole (Okada, Shirai & Miura, 2000). Tracking may be classified as 2D or 3D. 2D tracking consists in following the motion in the image plane either by exploiting low-level image features or by using a 2D human model. 3D tracking aims at obtaining the parameters, which describe body motion in three dimensions. The 3D tracking process, which estimates the motion of the body parts, is inherently connected to 3D human pose recovery.

3D pose recovery aims at defining the configuration of the body parts in the 3D space and estimating the orientation of the body with respect to the camera. This work will mainly focus on model-based techniques, since they are usually used for 3D reconstruction. Model-based techniques rely on a mathematical representation of human body structure and motion dynamics. The 3D pose parameters are commonly estimated by iteratively matching a set of image features extracted from the current frame with the projection of the

model on the image plane. Thus, 3D pose parameters are determined by means of an energy minimization process.

Instead of obtaining the exact configuration of the human body, human motion recognition consists in identifying the action performed by a moving person. Most of the proposed techniques focus on identifying actions belonging to the same category. For example, the objective could be to recognize several aerobic exercises or tennis strokes or some everyday actions such as sitting down, standing up, walking, running, or skipping.

Next, some of the most recent approaches addressing human motion tracking and 3D pose recovery are presented, while the following subsection introduces some whole-body human motion recognition techniques. Previous surveys of vision-based human motion analysis have been carried out by Cédras and Shah (1995), Aggarwal and Cai (1999), Gavrila (1999), Moeslund and Granum (2001), and Moeslund, Hilton, and Kruger (2006).

This overview presents briefly some of the techniques developed during the last years. The outline of this work is as follows. Firstly, a survey about human motion tracking and 3D pose recovery is given. Next, human motion recognition is introduced; following, a summary of some application works is presented. Finally, a section with future trends and conclusions is introduced.

HUMAN MOTION TRACKING AND 3D POSE RECOVERY

Tracking relies either on monocular or multiple camera image sequences. Using *monocular* image sequences is quite challenging due to occlusions of body parts and ambiguity in recovering their structure and motion from a single perspective view (different configurations have the same projection). On the other hand, single camera views are more easily obtained and processed than multiple camera views. In the following table, some recent techniques using only one camera are presented.

In contrast to single-view approaches, *multiple camera* techniques are able to overcome occlusions and depth ambiguities of the body parts, since useful motion information missing from one view may be recovered from another view. The following table presents some recent approaches using multiple cameras.

Some currently published papers specifically tackle the *pose recovery* problem using multiple sensors. In Mikic,. Trivedi, Hunter, and Cosman (2001) a 3D voxel reconstruction of the person's body is computed from silhouettes extracted from four cameras. Body parts are located sequentially from the voxel data. Consistency with an articulated body model is guaranteed by feeding measurements to

Table 1. Monocular systems

Authors	***Description***
Sminchisescu and Triggs (2001)	A 3D human body model, consisting of tampered superellipsoids, is fitted on the image features by means of an iterative cost function optimization scheme. A multiple-hypothesis approach with the ability of escaping local minima in the cost function is proposed.
Sidenbladh, Black, and Sigal (2002)	A probabilistic approach for modeling 3D human motion for synthesis and tracking, where learning of state transition probabilities is replaced with efficient probabilistic search in a large training set.
Ning, Tan, Wang, and Hu (2004)	An approach to tracking walking human based on both body model and a motion model (learnt from semiautomatically acquired training data), in a CONDENSATION framework. A pose evaluation function combining boundary and region information is proposed. Automatic acquisition of initial model pose and recovery from severe failures are addressed.
Antonini, Martinez, Bierlaire, and Thiran (2006)	Image processing methods are combined with behavioral models for pedestrian dynamics in a detection/tracking system. Behavioral models are used to find globally coherent trajectories. Pedestrian detection is based on the behavior rather than appearance.
Liu and Chellappa (2007)	An articulated model of a real human body consisting of ellipsoids connected by joints is used. Motion in a video sequence is observed using optical flow. Optical flow using scaled orthographic projection relates the spatial-temporal intensity change of the sequence to the motion parameters.
Caillette, Galata, and Howard (2007)	A 3D human body tracker, build upon the Monte-Carlo Bayesian framework, is capable of handling fast and complex motions in real-time. Novel prediction and evaluation methods improving the robustness and efficiency of the tracker are proposed. The tracker is also capable of automatic initialisation and self-recovery.

Table 2. Multiple-camera systems

Authors	***Description***
Delamarre and Faugeras (2001)	It incorporates physical forces to each rigid part of a 3D model consisting of truncated cones. The model's projections are compared with the silhouettes extracted from the image by means of a novel approach. This technique copes with self-occlusions, fast movements and poor quality images.
Atsushi, Hirokazu, Shinsaku, and Siji (2002)	Multiple people tracking in a wide-area covered by a number of sensors. Algorithms for simple blob-based tracking of humans and object matching among views are presented.
Yang, Shih, and Wang (2004)	Color, motion and depth features are combined for robust people tracking using a stereo camera. A Kalman filter is used for feature integration and tracking.
Kehl, Bray, and Van Gool (2005)	A volumetric reconstruction of a person is extracted from silhouettes in multiple video images. Then, an articulated body model is fitted to the data with stochastic meta descent (SMD) optimization.
Kang, Cohen, and Medioni (2005)	The proposed method tracks a large number of moving people with partial and total occlusions in a surveillance scenario. The appearance of detected moving blobs is described by multiple spatial distributions models of blobs' colors and edges.
Hayashi, Hashimoto, Sumi, and Sasakawa (2004)	The method uses a stereo camera mounted on the ceiling. Captured 3D voxels are projected onto the floor, and their peaks are tracked. An algorithm is proposed to predict collisions.

Table 3. Human motion recognition

Authors	***Description***
Masoud and Papanikolopoulos (2003)	PCA-based classification of human activities from video using low-level motion features.
Parameswaran and Chellappa (2005)	Actions are represented in a compact, view-invariant manner as curves in spaces arising from 3D mutual invariants.
Robertson and Reid (2006)	Stochastic modelling and recognition of sequences of actions, based on trajectory information, local motion descriptors and HMMs.
Sminchisescu, Kanaujia, and Metaxas (2006)	Algorithms for recognizing human motion in monocular video sequences, based on discriminative conditional random fields (CRFs) and maximum entropy Markov models (MEMMs) are proposed.
Nascimento, Figueiredo, and Marques (2007)	A hierarchical approach is presented for human motion pattern classification. More specifically, high-level activities are segmented into sequences of low-level motion patterns, which are simply independent increment processes, each describing a specific motion regime (e.g., ''moving left'').

Table 4. Human body model applications

Application	*Authors*	*Description*
Virtual Reality	Hilton (2003)	Summary of computer vision technology for the modeling and analysis of people; in particular for applications in games, multimedia and virtual reality.
	Deutsch, Lewis, and Burdea, (2006)	Virtual-reality-based telerehabilitation system providing to the therapist three-dimensional representations of patients' movements and exercise progress.
	Chattopadhyay, Bhandarkar, and Li (2007)	Algorithm for compression of Human Motion Capture data, exploiting structural information derived from the skeletal virtual human model. Better compressions than standard MPEG-4 are achieved.
Surveillance Systems	Havasi, Szlávik, and Szirányi (2006)	Detection of walking human figures contained in cluttered video sequences. Spatiotemporal information is used to detect and classify human movement patterns.
	Lei and Xu (2006)	Real-time video analysis system for outdoor surveillance and monitoring scenarios.
	Babu, Pérez, and Bouthemy (2007)	Improved visual tracking by combining sum-of-squared differences and color-based mean-shift trackers.
User Interface	Starner, Leibe, Minnen, Westyn, Hurst, and Weeks (2003)	Vision-based interface between the physical and virtual worlds able to identify 3D hand position, pointing direction, and sweeping arm gestures.
	Nickel and Stiefelhagen (2007)	Visual tracking of head, hands and head orientation in the context of human-robot interaction. A multihypothesis tracking framework is used to find 3D positions of body parts, based on color and stereo data information.
Medical Applications	Zubairi (2002)	Methodology for analyzing surface shape using computer-aided rasterstereography, suitable for screening children for spinal deformities.
	Gomez, Carstensen, and Ersbøll (2006)	Integrated imaging system for acquisition of accurate standardized images. Useful for characterizing dermatological images.
	Panchaphongsaphak Burgkart, and Riener (2007)	Multimodal virtual reality system for medical education. The user can visualize and manipulate graphical information.

an extended Kalman filter. In Plänkers and Fua (2003) an articulated implicit surface model is introduced for human body modelling and algorithms for reconstruction of the model from trinocular stereo data are presented. Carranza, Theobalt, Magnor, and Seidel (2003) describe a system that uses multiview synchronized video footage of an actor's performance to estimate motion parameters and to interactively re-render the actor's appearance from any viewpoint.

HUMAN MOTION RECOGNITION

Human motion recognition may be achieved by analyzing the extracted 3D pose parameters. However, because of the extra preprocessing required, recognition of human motion patterns is usually achieved by exploiting low-level features (e.g., silhouettes) obtained during tracking. Techniques extensively used for recognizing human actions include Template matching algorithms, State-space approaches (e.g., HMMs) and/or involve semantic or other high-level descriptions (e.g.,natural language) (Wang, Hu & Tan, 2003).

APPLICATIONS

A large range of applications involve vision-based human body models, requiring both tracking and recognition. Some recent works, grouped according to their applications field, are presented in the following table:

FUTURE TRENDS AND CONCLUSION

The problem of human motion tracking and recognition has become an attractive challenge. The huge amount of articles published during the last years demonstrates the increasing interest in this topic and its wide range of applications, yet in spite of this many issues are still open. Problems such as: unconstrained image segmentation, limitations in tracking, development of models including prior knowledge, modeling of multiple person environments, and real-time performance, still need to be efficiently solved.

A common limitation in tracking, for example, is that the motion of a person is constrained to simple movements with a few occlusions. Although several multiview systems have been proposed that are able to deal with ambiguities, the occlusion problem remains a main hindrance for automatic monoscopic human tracking. Another problem to be totally overcome is the initialization of the pose parameters, and automatic self-tuning of the model's shape parameters. Significant progress has been achieved in this direction. However no completely automatic technique that is able to cope with these problems in unconstrained environments has been demonstrated up to now.

Future human motion recognition systems must be able to identify human motion even in unconstrained environments. These systems can have many useful applications in areas ranging from everyday life to medicine. Especially, robust real-time systems may offer many benefits. Of course, the reduction of processing time presupposes not only advances on computational techniques but also improvements on the current technology.

ACKNOWLEDGMENT

This work was supported by the EC under the FP6 IST Network of Excellence 3DTV — Integrated Three-Dimensional Television-Capture, Transmission, and Display (contract FP6-511568) and the Government of Spain under research project TRA2007-62526/AUT and research program Consolider Ingenio 2010 CSD2007-00018.

REFERENCES

Aggarwal, J. K. & Cai, Q. (1999). Human motion analysis: A review. *Computer Vision and Image Understanding, 73*(3), 428-440.

Antonini, G., Martinez, S. V., Bierlaire, M., & Thiran, J. P. (2006). Behavioral priors for detection and tracking of pedestrians in video sequences. *International Journal of Computer Vision, 69*(2), 159-180.

Atsushi, N., Hirokazu, K., Shinsaku, H., & Siji, I. (2002). Tracking multiple people using distributed vision systems. In *Proceedings of the International Conference on Robotics and Automation*, Washington DC.

Babu, R., Pérez, P., & Bouthemy, P. (2007). Robust tracking with motion estimation and local kernel-based color modeling. *Image and Vision Computing, 25*(8), 1205-1216.

Caillette, F., Galata, A., & Howard, T. (2007). Real-time 3-D human body tracking using learnt models of behaviour. *Computer vision and image understanding*. DOI 10.1016/j.cviu.2007.05.005.

Capellades, M. B., Doermann, D., DeMenthon, D., & Chellappa, R. (2003). An appearance based approach for human and object tracking. In *Proceedings of the International Conference on Image Processing,* Barcelona, Spain.

Carranza, J., Theobalt, C., Magnor, M., & Seidel, H.-P. (2003). Free-viewpoint video of human actors. *ACM SIGGRAPH*, 565-577.

Cédras, C. & Shah**,** M. (1995). Motion-based recognition: A survey. *Image and Vision Computing, 13*(2), 129-155.

Chattopadhyay, S., Bhandarkar, S., & Li, K. (2007). Human motion capture data compression by model-based indexing: A power aware approach. *IEEE Trans. on Visualization and Computer Graphics, 13*(1), 5-14.

Delamarre, Q. & Faugeras, O. (2001). 3D articulated models and multiview tracking with physical forces. *Computer Vision and Image Understanding, 81*(3), 328-357 [Special Issue].

Deutsch, J. Lewis, J., Burdea, G. (2006). Virtual reality-integrated telerehabilitation system: Patient and technical performance. In *Proceedings of the IEEE International Workshop on Virtual Rehabilitation,* New York.

Gavrila, D. M. (1999). The visual analysis of human movement: A Survey. *Computer Vision and Image Understanding, 73*(1), 82-98.

Gomez, D., Carstensen, J., & Ersbøll, B. (2006). Collecting highly reproducible images to support dermatological medical diagnosis. *Image and Vision Computing, 24*(2), 186-191.

Havasi, L., Szlávik, Z., & Szirányi, T. (2006). Higher order symmetry for non-linear classification of human walk detection. *Pattern Recognition Letters, 27*(7), 822-829.

Hayashi, K., Hashimoto, M., Sumi, K., & Sasakawa, K. (2004). Multiple-person tracker with a fixed slanting stereo camera. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition,* Seoul, Korea.

Heikkila, M. & Pietikainen, M. (2006). A texture-based method for modelling the background and detecting moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*(4), 657-662.

Hilton, A. (2003). Computer vision for human modelling and analysis. *Machine Vision and Applications, 14*(4), 206-209.

Kang, J., Cohen, I., & Medioni, G. (2005). Persistent objects tracking across multiple non-overlapping cameras. In *Proceedings of the IEEE Workshop on Motion and Video Computing (MOTION'05),* Breckenridge, Colorado.

Kehl, R., Bray, M., & VanGool, L. (2005). Full body tracking from multiple views using stochastic sampling. *Computer Vision and Pattern Recognition*. San Diego, California.

Lei, B. & Xu, L. (2006). Real-time outdoor video surveillance with robust foreground extraction and object tracking via multi-state transition management. *Pattern Recognition Letters, 27*(15), 1816-1825.

Leibe, B., Seemann, E., & Schiele, B. (2005). Pedestrian detection in crowded scenes. *Computer Vision and Pattern Recognition.* San Diego, CA.

Liu, H. & Chellappa, R. (2007). Markerless monocular tracking of articulated human motion. In *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP 2007,* Honolulu, Hawaii.

Masoud, O. & Papanikolopoulos, N. (2003). A method for human action recognition. *Image and Vision Computing, 21*(8), 729-743.

Mikic, I., Trivedi, M. M., Hunter, E., & Cosman, P. (2001). Articulated body posture estimation from multi-camera voxel data. In *Proceedings of the Conferene on Computer Vision and Pattern Recognition*, Kauai Marriott, Hawaii.

Moeslund, T. B. & Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding, 81*(3), 231-268.

Moeslund, T. B., Hilton, A., & Kruger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding, 104*(2-3), 90-126.

Nascimento, J., Figueiredo, M. & Marques, J. (in press). Independent increment processes for human motion recognition. *Computer Vision and Image Understanding*.

Nickel, K. & Stiefelhagen, R. (2007). Visual recognition of pointing gestures for human–robot interaction. *Image and Vision Computing, 25*(12), 1875-1884.

Ning, H., Tan, T., Wang, L., & Hu, W. (2004). People tracking based on motion model and motion constraints with automatic initialization. *Pattern Recognition, 37*(7), 1423-1440.

Okada, R., Shirai, Y., & Miura, J. (2000). Tracking a person with 3D motion by integrating optical flow and depth. In *Proceedings of the 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition,* Grenoble, France.

Panchaphongsaphak, B., Burgkart, R. & Riener, R. (2007). Three-dimensional touch interface for medical education. *IEEE Trans. on Information Technology in Biomedicine, 11*(3), 251-263.

Parameswaran, V. & Chellappa, R. (2005). Human action-recognition using mutual invariants. *Computer Vision and Image Understanding, 98*(2), 295-325.

Plänkers, R. & Fua, P. (2003). Articulated soft objects for multiview shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(9), 1182-1187.

Robertson, N. & Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding, 104*(2), 232-248.

Sato, K. & Aggarwal, J.K. (2001). Tracking and recognizing two-person interactions in outdoor image sequences. In *Proceedings of the IEEE Workshop on Multi-Object Tracking*, Vancouver, Canada.

Sidenbladh, H., Black, M. J., & Sigal, L. (2002). Implicit probabilistic models of human motion for synthesis and tracking. **In** *Proceedings of the European Conference on Computer Vision*, Copenhagen, Denmark.

Sminchisescu, C., Kanaujia, A. & Metaxas, D. (2006). Conditional models for contextual human motion recognition. *Computer Vision and Image Understanding, 104*(2-3), 210-220.

Sminchisescu, C. & Triggs, B. (2001). Covariance scaled sampling for monocular 3D body tracking. **In** *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Kauai Marriott, Hawaii.

Starner, T., Leibe, B., Minnen, D., Westyn, T., Hurst, A., & Weeks, J. (2003). The perceptive workbench: Computer-vision-based gesture tracking, object tracking, and

3D reconstruction for augmented desks. *Machine Vision and Applications, 14*(1), 59-71.

Wang, L., Hu, W., & Tan, T. (2003). Recent developments in human motion analysis. *Pattern Recognition, 36*(3), 585-601.

Wren, C., Azarbayejani, A., Darrell, T., & Pentland, A. (1997). Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 19*(7), 780-785.

Yang, M. T., Shih, Y. C, & Wang, S. C. (2004). People tracking by integrating multiple features. **In** *Proceedings of the International Conference on Pattern Recognition*, Cambridge, UK.

Zubairi, J. (2002). Applications of computer-aided rasterstereography in spinal deformity detection. *Image and Vision Computing, 20*(4), 319-324.

KEY TERMS

Image Feature: A structure in the image with interesting characteristics, for example, points, edges, lines, surfaces.

Occlusion: When one object is in front of another object in the direction of observation, a portion of the object that is behind cannot be seen. Then, the second object is occluded by the first one.

Real-Time System: System that processes and updates information always within a given time.

Self-Occlusion: When a part of the object is occluded by another part of itself. For example, when a person is walking, one leg may occlude the other leg.

Stereo Vision System: System devised to extract 3D information of a given scene. In binocular stereo systems, two images are taken from different viewpoints allowing the computation of 3D structure.

Training Set: A set of known, labelled examples used in classification, representative of the data that will be classified in the application.

Tracking: The process of estimating the parameters of a dynamic system by means of measurements obtained at successive time instances. An example is the estimation of the position of a moving object from an image sequence.

Category: Software & Systems Design

Aesthetics in Software Engineering

Bruce MacLennan
University of Tennessee, USA

INTRODUCTION

It is commonly supposed that software engineering is—and should be—focused on technical and scientific issues, such as correctness, efficiency, reliability, testability, and maintainability. Within this constellation of important technological concerns, it might seem that design aesthetics should hold a secondary, marginal role, and that aesthetic considerations might enter the design process, if at all, only after the bulk of the engineering is done. This article discusses the important role that aesthetics can play in engineering, and in particular in software engineering, and how it can contribute to achieving engineering objectives.

BACKGROUND

Certainly, aesthetic considerations have not been completely absent from software engineering. For example, the development of structured programming ideas was accompanied with conventions for the textual layout of programs, which aimed for conceptual clarity, but also aesthetic appeal. (Even the older programming techniques relied on the aesthetic layout of flowcharts.) Knuth (1992) has advocated a practice of literate programming in which composites of programs and their documentation are treated as "works of literature." Elegance has been discussed as a criterion of programming-language design since the 1960s (MacLennan, 1997, 1999). Furthermore, software engineers have tended to prefer elegant algorithms (Gelernter, 1998) where the criteria of elegance have been inherited primarily from mathematics and theoretical science, in which beauty is a widely acknowledged value (Curtin, 1982; Farmelo, 2002; Fischer, 1999; Heisenberg, 1975; King, 2006; McAllister, 1996; Tauber, 1997; Wechsler, 1988; Wickman, 2005). In these and similar cases, however, there has been little direct work on an aesthetic theory for software.

One exception to the relative lack of explicit work on software aesthetics is the research program in aesthetic computing pursued by, for example, Fishwick (2002) and Fishwick, Diehl, Lowgren, and Prophet (2003). He argues that traditional program representations (such as textual programs and graphs) do not engage our aesthetic senses, and that they are abstract and uninteresting. However, because software concepts are abstract, they do not have a natural sensuous representation, and so Fishwick argues that they should be represented metaphorically (e.g., by visual or sculptural metaphors), and that the "craft-worthy, artistic step" is the choice of metaphor and the expression of the algorithm in terms of it. This metaphorical work (not the textual program, which is secondary) is then the primary medium for the aesthetic expression and appreciation of software (Fishwick, 2002). Recent developments in aesthetic computing are collected in Fishwick (2006).

Software engineering is a new discipline, and so it does not have a long aesthetic tradition as do many of the other arts and engineering disciplines. Therefore, it is helpful to look at well-established disciplines that have significant similarities to software engineering. One of these is mathematics, in which we may include theoretical science, which has in common with software engineering the fact that formal considerations dominate material considerations (explained below). As is well known, aesthetic considerations are important in theoretical science and mathematics (Curtin, 1982; Farmelo, 2002; Fischer, 1999; King, 2006; McAllister, 1996; Tauber, 1997; Wechsler, 1988; Wickman, 2005), and we will find Heisenberg's (1975) remarks on beauty in the exact sciences to be informative.

Another source of useful analogies comes, perhaps surprisingly, from the structural engineering of towers and bridges, which has in common with software engineering the fact that teams engage in the economical and efficient design and construction of reliable, large, complex systems, the failure of which may be expensive and dangerous. We will appeal especially to Billington's (1985) insightful investigation of the role of aesthetics in structural engineering.

AESTHETICS, FORMALITY, AND SOFTWARE

Importance of Aesthetics

Billington (1985) identifies three criteria of good design—the three *E*s of efficiency, economy, and elegance—which he associates with three dimensions of the design process: the scientific, the social, and the symbolic (the three *S*s). All of these also apply to software engineering. Efficiency is primarily a scientific issue since it deals with the physical resources necessary to meet the project's requirements. In structural engineering, the fundamental requirement is safety, which is also important in software engineering, in which other

important issues are real-time response, dynamic stability, robustness, and accuracy, among others.

The second criterion, economy, which treats benefits and costs, is a social issue because it depends on a society's values, both monetary and other. Certainly, many economical considerations can be reduced to money, which is almost the common denominator of value in the modern world, but it is problematic, at very least, to treat some costs, such as human death and suffering, in financial terms. The economy of a design depends on many social factors, including the costs of materials, equipment, and other resources; the availability and cost of labor; governmental regulations; and so forth. As a consequence, the economic worth of a design is more difficult to evaluate and predict than its efficiency.

This brings us to elegance, the aesthetic criterion, and while it will probably be granted that beautiful designs are preferable for their own sakes (other things being equal), Billington (1985) gives compelling arguments for the engineering worthiness of elegant designs, which also apply to software engineering.

As will be explained, elegance is a means of conquering complexity. In terms of their numbers of components and their interactions, software systems are some of the most complicated systems that we design. Even programs as simple as text editors may have hundreds of thousands of lines of code. Furthermore, these components (the individual programming-language statements) interact with each other in complex ways so that the number of potential interactions to be considered rises to at least the square of the number of lines of code. In addition, software engineers are at a disadvantage compared to other engineers for the components of programs are abstract operations, whereas designers in other engineering disciplines can depend on their physical intuition to aid their understanding (Ferguson, 1992).

Software engineers have designed various analytical tools to help them conquer this complexity, but it is important to understand the inherent limitations of analysis. Analysis makes use of models (especially mathematical models) to understand some system of interest. These models are useful because they make simplifying assumptions that permit the models to be understood more easily than the modeled systems. An effective model simplifies matters that are irrelevant to its purpose so that our limited cognitive resources can be devoted to the relevant matters. Often, the simplifying assumptions are made consciously; for example, we may use a linear model for a nonlinear system if we have reason to believe that the nonlinear effects are irrelevant to our concerns. However, models also typically make tacit simplifications because certain factors are assumed to be irrelevant or, in some cases, because some factors have never been considered at all. For example, Billington (1985) observes that the Tacoma Narrows Bridge collapsed 4 months after it was opened because aerodynamic stability had not been considered a factor in bridge design (see also Ferguson, 1992). The problem is more severe for more complex systems because our cognitive capacities are strictly limited, therefore so is the completeness of the model; hence, the gap between the modeled factors and the system increases with increasing system complexity. The risk of intentionally or inadvertently omitting relevant factors increases with system complexity.

Billington (1985) observes that in structural engineering, elegance compensates for the limitations of analysis. This is because in an elegant design, the disposition and balance of the forces are manifest in the design's form, and therefore designs that look stable are in fact stable. More generally, designs that look good are good. For example, the Tacoma Narrows Bridge and several other bridges, which were designed on the basis of extensive mathematical analysis and computer models, collapsed because aerodynamic stability was not included in the analysis, whereas earlier bridges, whose designs were guided by aesthetic judgment, were aerodynamically stable. If we generalize from the specifics of structural engineering, we may say that elegant designs are manifestly correct, safe, efficient, and economical. But why should we expect any correlation between aesthetic values and engineering values?

Billington (1985) argues that in structural engineering, Louis Sullivan's architectural maxim "form follows function" is not as applicable as its converse, "function follows form." Sullivan's principle asserts that a building's function should strongly determine its form. In structural engineering, however, there are typically many ways to fulfill a particular function (such as bridging a river at a particular location) for "engineering design is surprisingly open-ended" (Ferguson, 1992, p. 23). Therefore, since there is wide choice of form, the engineer can choose designs that manifest a stable disposition of forces, that is, elegant designs. More abstractly, engineers who are guided by elegance choose to work in a region of the design space in which aesthetic values correspond to engineering values (that is, function follows form). Therefore they can rely on their aesthetic judgment, which is a highly integrative cognitive faculty, to guide the design process.

The same considerations apply even more so in software engineering, where there are typically very many possible software solutions to an application need. That is, software engineers can compensate for the limitations of analysis by limiting their attention to designs that are elegant, that is, manifestly correct, efficient, maintainable, and so forth. In this region of the design space, where aesthetic values correspond to engineering values—where good designs look good—designers can rely on their aesthetic judgment to guide them to good engineering solutions.

Elegance is associated with the symbolic dimension of a design because, in the appropriate region of the design space, aesthetic values symbolize engineering values. However, design aesthetics can also symbolize less tangible

values that are nevertheless important; this may be called the ethical aspect of the design. For example, a design may be straightforward or subtle, or it may be well organized or a rat's nest; it may be monolithic or modular, and so forth. That is, the design accentuates or suppresses various attitudes, concerns, and other values, which are therefore kept before the designers' eyes or are allowed to escape their attention. An aesthetically sensitive engineer will tend to conform to the aesthetic values already exemplified by the design, and so they will tend to perpetuate the nonaesthetic values symbolized by its aesthetic qualities. Elegance calls forth more elegance.

As Heisenberg (1975) remarks, the ethical dimension is especially important in long-term group projects. As an aesthetic object, the ongoing project symbolizes the constellation of values around which it is organized. In this way, new project members are encouraged to conform to these values and maintain the integrity of the project. Heisenberg compares the generations of scientists who construct a beautiful theory to the medieval cathedral builders, who "were imbued with the idea of beauty posited by the original forms, and were compelled by their task to carry out exact and meticulous work in accordance with these forms" (p. 176).

So far, I have focused on the importance of aesthetics for the designers of software systems, but it is also important for its users. It is perhaps obvious that, other things being equal, users would prefer using aesthetically attractive software to using unattractive alternatives. However, many people's occupations are dominated by interaction with software systems, and software is coming to occupy many people's leisure activities as well. Therefore, the aesthetic dimension of software can have a significant effect on users' quality of life.

Many people's experience of software is dominated by fear, which results largely from the unpredictability of their interactions with it. People interact with technological devices better to the extent that they can form a cognitive or conceptual model of the workings of the system (Norman, 1988, 1998, 2005). The model does not have to represent the actual internal operation of the system, but it has to be accurate enough to allow users to interact with it reliably. In the same way that people will use an elegantly designed bridge with more confidence—for it looks as stable and strong as it is—they will be able to use elegantly designed software with confidence for its external form will reflect and imply its internal logic.

As users become more educated about the aesthetic qualities of software, they will prefer elegant software and therefore encourage its production with a corresponding improvement in the scientific and economical aspects of software systems. Billington (1985) has pointed out the role of an aesthetically knowledgeable public in European bridge design, where the aesthetics of structural engineering projects are regularly critiqued in public.

Formality and Aesthetics

How can elegant software be achieved? The products of all the arts (technical arts as well as fine arts) can be analyzed in terms of form and matter. In this context, matter refers to the raw material of the art, the relatively neutral stuff on which an art imposes its particular kind of form. The matter of painting includes the canvas, paints, and brushes; the matter of music includes duration, pitches, timbres, and so forth, which are the raw ingredients of a musical composition. The matter of programming is the hardware, which provides relatively neutral ground (comprising memory cells, machine operations, etc.) on which a program imposes a particular form, organizing the computer's operations in time and memory space.

Any artist (or technologist) must be sensitive to both the formal and material characteristics of the art, but software engineering is unusual in the degree to which formal considerations dominate material considerations. Indeed, the correctness and efficiency of software is usually treated entirely from a formal perspective, and material concerns (e.g., the limitations and capacities of specific hardware) are addressed as second-order effects. In spite of enormous changes in the medium of computation (relays, vacuum tubes, transistors, VLSI), we still use many of the same algorithms that were used in the earliest days of computing (and in some cases even before computers were invented). This highly formal quality of the software art distinguishes it from most other artistic and engineering disciplines.

As a consequence, in establishing aesthetics for software engineering, it is helpful to look to other longer established enterprises where formal issues dominate the material. In particular, we can learn from the elegance and aesthetics in mathematics and the theoretical sciences, which have similar standards of correctness, generality, simplicity, elegance, and beauty (e.g., Curtin, 1982; Farmelo, 2002; Fischer, 1999; King, 2006; McAllister, 1996; Tauber, 1997; Wechsler, 1988; Wickman, 2005), which are also applicable to software.

Of course, the role of aesthetics in mathematics and science is itself a topic of discussion, but the analysis of Heisenberg (1975) provides a good starting point. The formality of these disciplines (and of software) suggests a Platonic approach to aesthetics, which explains beauty in terms of formal structure. Indeed, as Heisenberg remarks, aesthetic concerns were at the root of Western science in Pythagoras' correlation of musical pitch with numerical ratios: The most beautiful (consonant) intervals were found to correspond to the simplest whole-number ratios (1/2, 2/3, 3/4).

According to classical theories of aesthetics (as represented primarily by Plato and Aristotle), the beauty of a work resides in a symmetric, orderly, and harmonious relation of the parts to each other and to the whole. Our aesthetic response arises from a comprehension of these formal relationships, and thus there is a correspondence between beauty and in-

telligibility. As we saw for elegance, therefore, engineers' aesthetic responses can be used to guide software design and to evaluate its result.

These aesthetic principles can be applied both to the static relations within a software system and to the dynamic relations generated by the running system; we begin with the static relations.

As stated by Heisenberg (1975, p. 174), "Beauty is the proper conformity of the parts to one another and to the whole." Specifically, there should be a *harmonia* among the parts, which in ancient Greek meant that they fit together well. They should be arranged symmetrically and be meaningfully ordered; finally, they should be of similar size and of a size and number that is easy to comprehend. In short, the conformity of the parts to one another coincides with their intelligibility.

A beautiful design is characterized also by the conformity of the parts to the whole. Classically, this means the whole has an organic unity similar to that of a living organism. Therefore, the parts fulfill complementary functions to constitute the whole, which in turn defines the relationships of the parts to each other; the parts and the whole are mutually determinative. In short, the design has integrity.

The preceding criteria are intrinsic to the work, but classical aesthetics also addresses the extrinsic relation of the work to its perceivers: The parts should be of such number and size that they can be easily discriminated, and the whole should be of a perceptible size and capable of being held in the perceiver's memory. Obviously, when applied to software, these criteria depend in part on the visual presentation of the program, which may be enhanced by visual programming languages. More generally, an aesthetic design conforms to the perceptual, cognitive, and memory abilities of human beings; in these cases, the perceiver has the satisfaction of comprehending the system's formal structure.

Of course a program is more than an object of contemplation; it is a static structure that defines an infinite set of possible execution sequences (conditioned by possible environment states). Programming is difficult because an informal idea of these possible execution sequences must be expressed by a static structure (the program) that can control a computer to generate any of the desired sequences. Therefore, software design has similarities to the use of mathematics in the theoretical sciences, which seeks to express observed regularities in mathematical laws. Thus, we can benefit from Heisenberg's (1975) insights into the role of beauty in science, as well as those of others (Curtin, 1982; Wechsler, 1988).

Embodiment and Software Aesthetics

Fishwick (2002) has criticized traditional visual presentations of programs for being "aesthetically-challenged and Platonic" and "largely devoid of texture, sound, and aesthetic content," and he has advocated aesthetic programming with the goal of crafting software with "sensory appeal." Indeed, even Plato, who argued that beauty is rooted in formal relations among abstract ideas, advocated perceptible beauty (mirroring the formal relations) as a route toward appreciation of abstract beauty (e.g., *Symposium*, 209e–212a). Therefore the visual form of software is relevant to its aesthetics and affects our intellectual and emotional response to it.

Indeed, in recent decades psychologists have come to recognize the important role that embodiment plays in human cognition (e.g., Gibbs, 2006; Lakoff & Johnson, 1999), and Lakoff and Núñez (2000) have argued that abstract mathematical concepts and our intuitions about them are grounded in our sensorimotor skills and experiences as living beings. For example, our basic intuitions about sets arise from our experience with physical containers.

Humans have a deep, intuitive understanding of the physical world and of our embodied relation to it, which is part of our genetic heritage or is acquired in early childhood. People normally feel pleasure when acting skillfully, and that feeling of satisfied competence is part of the psychological feedback process that improves human skill. Therefore, one way to increase pleasurable interactions with software, for software engineers as well as for users, is to design it so that it conforms to our embodied sensorimotor skills, and therefore engages them pleasurably. For example, this can be accomplished by designing interfaces that use actual or simulated physical interactions as a way of interfacing with abstract structures and systems. Thus, Karlsson and Djabri (2001) suggest that when objects are manipulated on a screen, they should mimic the familiar inertia, elasticity, and so forth of physical objects (see also Norman, 1998). Therefore, an increasing focus on embodiment will bring software aesthetics closer to the less formal (more material) aesthetics of architecture and other artistic, design, and engineering disciplines.

FUTURE TRENDS

Along with increasing awareness of the importance of emotional intelligence (e.g., Goleman, 2005) and embodied understanding, there has come an appreciation of the role of aesthetics in mathematics, science, and engineering (e.g., Wickman, 2006). Historically, there has been little explicit work on software aesthetics, but that is beginning to change (Fishwick, 2006; Gelernter, 1998). I propose that we can progress by four simultaneous, interrelated activities, which I will call experiment, criticism, theory, and practice.

By experiment, I mean the conscious application of aesthetic principles in program design and the empirical evaluation of the results, including an aesthetic analysis both by the designers and by other software engineers (since software engineering is, ultimately, a social process). In

addition to experiments designed explicitly to test aesthetic hypotheses, every software engineering project should be an implicit or explicit aesthetic experiment.

Criticism is fundamental to all the arts. In addition to informing the public about the aesthetic quality of works, it raises aesthetic issues and makes them salient for consumers and artists. Artists in particular articulate a position relative to criticism, perhaps responding with their own criticism, and may adapt their art either in conformity or reaction to the criticism. Thus, criticism is an important feedback process in the evolution of aesthetics, in the formulation of aesthetic principles, and in the aesthetic education of consumers.

By theory I refer to our knowledge of cognitive and affective neuropsychology, which improves our understanding of human aesthetic response and its relation to cognition; this growing body of theory will help to explain experimental results and will suggest new experiments. Nevertheless, human aesthetic response and cognition are both complex and poorly understood, and thus for some time the application of aesthetic principles in software engineering will remain a practical matter with a weak theoretical basis.

This brings us to the matter of practice. The long history of aesthetic debate in philosophy and the fine arts suggests that beauty is not a simple matter of conformity to a few aesthetic principles. Therefore, in the art of programming, as in the other arts, we expect aesthetic principles to provide general guidelines, within which—and occasionally outside of which—designers exercise their aesthetic judgment. Indeed, all expert behavior is characterized by behavior that is broadly in conformity with normative rules, but ultimately free of them (Dreyfus & Dreyfus, 1986); so also with software design.

For the potential benefits of aesthetically designed software to become reality, we will need aesthetics to be a part of the education of every software engineer (cf. Wickman, 2005). Since the goal is for all software design to be guided by programmers' well-honed sense of elegance and aesthetic judgment, aesthetic issues should be addressed in all software design courses, not just in one course devoted to software aesthetics. To accomplish this, all computer science educators will need to become more attentive to aesthetics.

CONCLUSION

We have seen that attention to aesthetics in software engineering can have practical benefits and also improve designers' and users' emotional response to software. Elegant software results from choosing to work in that region of the design space where good designs also look good, therefore aesthetically sensitive programmers can use this highly integrative cognitive faculty to guide their design decisions. Because of the predominantly formal character of software, Platonic aesthetics is most appropriate as it is in mathematics and theoretical science; according to this aesthetics, beauty resides in the conformity of the parts to one another and to the whole. Nevertheless, the more sensuous aspects of aesthetics are also relevant for they engage our embodied understanding and sensorimotor skills. The full benefits of attention to aesthetics will require ongoing pursuit of the aesthetic principles of software and their integration into computer science education.

REFERENCES

Billington, D. P. (1985). *The tower and the bridge: The new art of structural engineering.* Princeton, NJ: Princeton University Press.

Curtin, D. W. (Ed.). (1982). *The aesthetic dimension of science: 1980 Nobel Conference.* New York: Philosophical Library, Inc.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York: Free Press.

Farmelo, G. (Ed.). (2002). *It must be beautiful: Great equations of modern science.* New York: Granta Books.

Ferguson, E. S. (1992). *Engineering and the mind's eye.* Cambridge, MA: MIT Press.

Fischer, E. P. (1999). *Beauty and the beast: The aesthetic moment in science.* New York: Plenum Publishing Corporation.

Fishwick, P. (2002). Aesthetic programming: Crafting personalized software. *Leonardo, 35*, 383-390.

Fishwick, P. (Ed.). (2006). *Aesthetic computing.* Cambridge, MA: MIT Press.

Fishwick, P., Diehl, S., Lowgren, J., & Prophet, J. (2003). Aesthetic computing manifesto. *Leonardo, 36*, 255-256.

Gelernter, D. (1998). *Machine beauty: Elegance and the heart of technology.* New York: Basic Books.

Gibbs, R. W., Jr. (2006). *Embodiment and cognitive science.* New York: Cambridge University Press.

Goleman, D. (2005). *Emotional intelligence: Why it can matter more than IQ* (10th ed.). New York: Bantam.

Heisenberg, W. (1975). The meaning of beauty in the exact sciences. In W. Heisenberg (Ed.) & P. Heath (Trans.), *Across the frontiers* (pp. 166-183). New York: Harper & Row.

King, J. P. (2006). *The art of mathematics.* New York: Dover Publications.

Knuth, D.E. (1992). *Literate programming* (CSLI Lecture Notes, No. 27). Stanford, CA: Center for the Study of Language and Information.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought.* New York: HarperCollins Publishers.

Lakoff, G., & Núñez, R. E. (2000). *Where mathematics comes from: How the embodied mind brings mathematics into being.* New York: Basic Books.

MacLennan, B. J. (1997). "Who cares about elegance?" The role of aesthetics in programming language design. *SIGPLAN Notices, 32*(3), 33-37.

MacLennan, B. J. (1999). *Principles of programming languages: Design, evaluation, and implementation* (3rd ed.). New York: Oxford University Press.

McAllister, J. W. (1996). *Beauty & revolution in science.* New York: Cornell University Press.

Norman, D. A. (1988). *The psychology of everyday things.* New York: Basic Books.

Norman, D. A. (1998). *The invisible computer.* New York: MIT Press.

Norman, D. A. (2005). *Emotional design.* New York: Basic Books.

Tauber, A. I. (1997). *The elusive synthesis: Aesthetics and science.* New York: Springer.

Wechsler, J. (Ed.). (1988). *On aesthetics in science.* New York: Birkhauser Verlag.

Wickman, P.-O. (2005). *Aesthetic experience in science education: Learning and meaning-making as situated talk and action.* New York: Lawrence Erlbaum Associates.

KEY TERMS

Aesthetics: Aesthetics is a set of principles and practices pointing toward a particular notion of beauty; it is also the discipline that studies such principles and practices.

Elegance: Refers to beauty, grace, harmony, beautiful simplicity, restraint, clarity, and precision.

Embodiment: Refers to the important role played by the human body, including the brain, in cognition, language, imagination, and other mental processes.

Ethical Dimension: The ethical dimension (or aspect) of a software design refers to the influence of the design on the behavior of the engineers working on it.

Execution Sequence: Is the sequence of instructions executed in a particular run of a program (i.e., when provided particular inputs).

Normative: Provides norms or standards of behavior, practice, and so on.

Sensorimotor Skills: Refers to those skills by which we interact fluently and competently with the world, integrating sensory perception and motor action.

Visual Programming Language: Is a programming language in which programs are displayed in some nontextual form, such as a graph, tree, or nested assembly of tiles.

African-Americans and the Digital Divide

Lynette Kvasny
The Pennsylvania State University, USA

Fay Cobb Payton
North Carolina State University, USA

INTRODUCTION

The Internet has become an integral part of America's entertainment, communication, and information culture. Since the mid 1990s, the Internet has become prevalent in middle and upper-class American households. Companies and government agencies are increasingly offering products, services, and information online. Educational institutions are integrating technology into their curriculum and are offering courses from a distance.

However, while some are advantaged by the efficiencies and convenience that result from these innovations, others may unwittingly become further marginalized by these same innovations since Internet access is not spreading to them as quickly. The 'digital divide' is the term used to describe this emerging disparity. Government analysts argue that historically underserved groups such as racial and ethnic minorities, rural and low-income communities, and older Americans are at a distinct disadvantage if this divide is not closed because American economic and social life is increasingly becoming networked through the Internet (National Telecommunications and Information Administration, 1995).

Over the last decade access to the Internet has increased significantly. A 2006 Pew Internet and American Life survey shows that 73% of U.S. adults (about 147 million adults) are Internet users, up from 66% (about 133 million adults) in 2005. And the share of Americans who have broadband connections at home reached 42% (about 84 million), up from 29% (about 59 million) in 2005 (Madden, 2006). African-Americans are increasingly accessing the Internet via home broadband connections, with a 121% adoption rate in 2005 (Horrigan, 2006). But does this mean that the problem of the digital divide has been solved? Is further research in this area warranted or has the digital divide become passé? In this article, we take on these questions by first reviewing major issues and trends in digital divide research. We do so by reviewing the digital divide literature as it relates to one historically underserved group, namely African-Americans. Next, we present a conceptual framework that contrasts 1) social and technological access perspectives, and 2) asset-based/resource and behavioral/use perspectives. The article concludes with our recommendations for future research opportunities for examining digital divide issues.

BACKGROUND

There have been numerous definitions for the digital divide, government and industry reports about the digital divide, and competing interpretations of the statistics contained in these reports. For instance, the digital divide has been defined at the What is Web site as "the fact that the world can be divided into people who do and people who don't have access to—and the capability to use—modern information technology, such as the telephone, television, or the Internet." Others (PRNewswire, 2000) offer another definition: "arguably the single, largest, and segregating force in today's world. If it is not made a national priority, a generation of children and families will mature without these tools that are proving to be the key to the future."

Most of our knowledge about the digital divide in the U.S. is based on survey research on computer and Internet access in the home, at work, and in public places. The most cited statistics are found in the digital divide series produced by the U.S. Department of Commerce (National Telecommunications and Information Association, 1998; 1999; 2000; 2002). Theses studies have found that the divide cuts along the lines of ethnicity and race, geographic location, household composition, age, education, and income level. However, these gaps are rapidly closing. In September 2001, 143 million Americans (54%) were using the Internet, and 174 million Americans (66%) used computers (U.S. Department of Commerce 2002). The gains are largest for low income families (those earning less than $15,000 per year increased at a 25% percent annual growth rate versus 11% for households earning $75,000 and above), and under represented ethnic and racial minorities (33% for Blacks, 30% for Hispanics, 20% for Whites and Asian American and Pacific Islanders). American Internet users are also engaged in a wide variety of activities—45% use e-mail, 36% use the Internet to search for products and services, 39% of individuals are making online purchases, and 35% are searching for health information (U.S. Department of Commerce, 2002).

In 2006, 73% of the U.S. adult population and 61% of African-Americans are online. These online African-Americans tend to use the Internet differently than other racial and ethnic groups. Novak, Hoffman and Venkatesh (1997)

summarize previous research on African-Americans with regard to different media as follows:

African-Americans have the highest participation in radio and TV and the lowest participation in newspapers. In terms of our classification, it means that historically, they have participated in greater measure in entertainment-oriented technologies rather than in information oriented technologies. Previous studies have also shown that African-American ownership of telephones is lower than white ownership, which may be due in part to differences in income.

They go on to theorize that culture helps to explain these results. African-Americans have found their social expression historically through the arts, and have been less successful in gaining entry to other dominant domains such as business, education, technical employment, and professional occupations. Culture may also help to explain Spooner and Rainie's (2000) observation that online African-Americans are 69% more likely than online whites to have listened to music on the Web, and are 65% more likely than online whites to have sought religious information on the Web. Music and spirituality have traditionally been integral components of African-American culture.

Although African-Americans may use the Internet relatively less than other ethnic groups, they have more positive attitudes toward the Internet than do similarly situated whites (Mossberger & Tolbert, 2003). For instance, Kvasny (2006) found that working class African-American women believed that computer skills would prepare them for higher paying jobs, and improve their parenting abilities. In a study of Internet adoption in a community technology project, Youtie et al. (2002) found that African-American women were among the highest adopters of cable TV-based Internet devices.

Although African-Americans harbored favorable attitudes towards the Internet, these same technologies may have little impact on social inclusion. In a more recent study, Sipior, Ward, Volonino and Marzec (2004) examined the digital divide in a public housing community in Delaware County, PA USA. With thirty-one African-American participants with ages ranging from 13-65, these researchers concluded that effective community-based programs could help reduce the divide. While these interventions notably have improved computing skills about under-served groups, a one time shot fails to eliminate or even reduce broader feelings of cultural isolation among minority groups.

CURRENT DEBATES

Given the rapid Internet uptake by African-Americans and other under-represented groups, one of the foremost issues is whether a digital divide still exists. We contend that these debates about the existence of the digital divide result from a rather narrow treatment of a complex social phenomenon. In fact, many of the newer studies in this genre call for a rethinking of the digital divide (Warschauer, 2002; Gurstein, 2003; Hacker and Mason, 2003; Kvasny, 2003; Payton 2003; Payton forthcoming). In what follows, we organize a discussion of the current debates in the digital divide discourse. We do so through a framework (Table 1) that contrasts two perspectives of access (technological and social) and two perspectives of use (asset-based and behavioral). Technological access focuses on the computing artifact, while social access focuses on know-how and competence. Asset-based perspectives view the divide as a deficiency in requisite resources such as income or education that enable Internet use, while behavioral perspectives tend to focus on the effectiveness of Internet use. Although these perspectives are presented as separate categories, authors tend to draw from both categories. For instance, the argument that the digital divide is based upon a lack of access to computing artifacts and computer skills suggests a technological access/asset-based perspective. An argument that the digital divide emerges from a lack of understanding about how to use the Internet to further life chances adopts a social/behavioral perspective.

Technological and Social Perspectives on Access

The technological access view, with its focus on broad statistics on Internet diffusion and use rates, has led some policy analysts to assume that the answer lies in certain characteristics of the technology. Hence, policy solutions tend to employ technological fixes, such as, wiring public schools and libraries, and providing computing resources with Internet access in poorer communities (Norris, 2001). We

Table 1. Competing perceptions for examining the digital divide

Access Factors	**Use Factors**
Technological	Asset-based
Social	Behavioral

contend that an over reliance on descriptive statistics largely contributes to this technology-centric understanding of the digital divide. The more important question for studying as technological access increases is 'what are people able to do with this access'?

We further argue that emphasis on quantitative descriptions of who has and who lacks access fuels debates about the degree to which the divide is temporary or permanent, whether the divide is widening or narrowing, or whether a divide exists at all. We have already seen the initial have/have not thesis superseded with the more complacent have now/have later prediction. Proponents of the have now/have later position argue that given enough time, competition will eventually alleviate any natural disparities in the marketplace.

Digital divide interventions informed by a technological access perspective are likely to subside once the technology gap has been narrowed through various programs and policies designed to distribute these resources more evenly. From this perspective, the digital divide would not warrant long-term policy remedies. High profile, short-term injections of government, foundation, or corporate assistance will occur until such time as the technology diffusion problem is lessened. Then, further critical attention to social inequities that are deeper than descriptions of technology access and use may be stifled. The digital divide will be simply defined away (Kvasny & Truex, 2001). For instance, in 2002 the Bush Administration declared the end of the digital divide and proposed deep cuts to federal digital divide programs. The biggest proposed cut was levied against the Technology Opportunities Program (TOP), a federal grant program designed to bring aid to communities that are lagging in access to digital technologies. Under the Clinton administration's 2001 budget, the program distributed $42 million in grants to 74 different non-profit organizations. In 2002, that number fell to just over $12 million (Benner, 2002).

Analysts and researchers who adopt the social access perspective critique this shortsightedness, and assert that the technological access paradigm ignores social constraints, such as workforce literacy, income differentials, and the inevitable household tradeoffs required in making a PC purchase. Simply put, Marlow's Needs Hierarchy must be addressed from the most fundamental level if one is to progress to higher-order affiliation. The digital divide reflects not only differences in the structure of access, but also the ways in which historical, economic, social, cultural, political and other non-technological factors make such differences matter. Technology-centric solutions alone will do little to redress these aspects of the digital divide.

Asset-Based and Behavioral Perspectives on Use

As access diffuses to historically underserved groups, use also becomes an important basis for studying the digital divide (DiMaggio & Hargittai, 2001; Patterson & Wilson, 2000; Warschauer, 2002; Gurstein 2003). From an asset-based perspective, computer and Internet access are insufficient without the requisite skills and competencies to use the technology effectively (Mossberger, Tolbert, & Stansbury, 2003). These authors take historically underserved groups as their subjects, and point out the ways in which their use of the Internet may be hampered. For instance, in a study of African-American youths, Payton (2003) found that these youths were all too aware that the digital divide is not merely about Internet access. Rather, it involves access to the social networks that ease the path to success in high-tech careers. Hargittai (2001) introduces the concept of 'second level divides' to signify the disparities in computer skills and how these disparities are patterned along age, racial, gender and income categories. She found, for example, that search time was positively correlated with age, and negatively correlated with education and prior experience with technology.

In contrast, the behavioral perspective sees the digital divide in terms of disparities in benefits like social inclusion, economic opportunity and political participation that one derives from Internet use. These disparities provide primary justification for realizing that the digital divide is a public problem and not simply a matter of private misfortune (Warschauer, 2003). Groups that are historically underserved in their quality of employment, their level of qualifications, their level of income, their quality of education, and their consumption opportunities tend to also be marginalized in their access to and use of IT. The digital divide, therefore, is a political outcome rooted in these historical systems of power and privilege, and not simply a gap in access to technological artifacts. Promoting access and basic training to improve the computer skills of individuals is warranted, but may do little to redress the social forces that may limit these actions in the first place. From the behavioral perspective, the divide is about disparities in what individuals and groups are able to do with their Internet access. Gurstein (2003) contends that effective use of the Internet occurs when people are able to use this technology purposively and independently to improve their life chances in culturally relevant domains such as economics, employment, health, education, housing, recreation, culture, and civic engagement.

FUTURE TRENDS

Despite one's alignment regarding the nature and context (technological or social access; asset-based or behavioral use) of the digital divide, the topic warrants re-conceptualization. Holistic approaches and frameworks will assist academic, government and industry leaders to first understand the 'Other' and the social conditions under which these groups function. Technology access and skills can equip; however, taken in isolation, they cannot sustain, maintain or offer ac-

cess to the social, financial or educational networks needed for empowerment of the 'total person'.

Thus, as digital divide research matures, there is a continued need to better understand the social dimensions of access. Kling (2000) contends that we do not have a good understanding of the ways that social access to the Internet is effectively supported for ordinary people at home and in public service agencies, such as schools and libraries. This is a topic that merits significant inquiry, since a large body of research points to its importance. He also argues that it is important to examine the specific kinds of networked services that will actually be of value to ordinary people. This goes beyond the current survey research and laboratory studies that contrast the use of the Internet by various demographic groups. We need to understand why groups use the Internet in the manner that they do, be appreciative of the culture perspectives that diverse groups bring to their online interactions, and how people use the Internet in various contexts such as work, home, churches and public access facilities.

CONCLUSION

In this article, we discussed the background, current debates, and future trends in digital divide research and policy. We focus our discussion on African-Americans because historically, this group has been under-represented on the Internet. However, in 2006, African-Americans are represented online in numbers proportional to their representation in the U.S. population, and are now among the fastest growing groups of Internet users. As the number of African-Americans and other non-traditional Internet users increase, digital divide research and policy must expand from an access-centric agenda to include Internet use by diverse groups. Understanding Internet use as well as barriers to access enables the creation of socially just policies and the development of Internet-based content that serves the needs and interests of diverse Internet users.

REFERENCES

Benner, J. (2002). *Bush Plan 'Digital Distortion',* Wired News, February, 2. Retrieved December 2006, from http://www.wired.com/news/politics/0,1283,50279,00.html

DiMaggio, P. J., & Hargittai, E. (2001). *From digital divide to digital inequality: Studying Internet use as penetration increases*. Sociology Department, Princeton University. Retrieved December 2006, from http://www.princeton.edu/~eszter/research

Gurstein, M. (2003). Effective use: A community informatics strategy beyond the digital divide. *First Monday, 8*(12). Retrieved December 2006, from http://www.firstmonday.org/issues/issue8_12/gurstein/

Hacker, K., & Mason, S. (2003). Ethical gaps in studies of the digital divide. *Ethics and Information and Technology, 5*(2), 99-115.

Hargittai, E. (2002). Second-level digital divide: Differences in people's online skills. *First Monday, 7*(4). Retrieved December 2006, from http://www.firstmonday.dk/issues7_4/hargittai/index.html

Hoffman, D. L., & Novak, T. P. (1998). Bridging the racial divide on the Internet. *Science, 280*(5362), 390-391.

Horrigan, J. (2006). Home Broadband Adoption. Pew Internet and American Life Project, Washington, D.C. Retrieved December 2006, from http://www.pewInternet.org/pdfs/PIP_Broadband_trends2006.pdf

Jerding, C. M. (2000b, March 17). *True nature of digital divide divides experts* [WWW]. The Freedom Forum Online. Retrieved August 2003, from http://www.freedomforum.org/news/2000/03/2000-02-17-06.asp

Kling, R. (1998). *Technological and social access to computing, information and communication technologies*. White Paper for Presidential Advisory Committee on High-Performance Computing and Communications, Information Technology, and the Next Generation Internet. Retrieved August 2003, from http//www.ccic.gov/ac/whitepapers.html

Kvasny, L. (2006). Let the sisters speak: Understanding information technology from the standpoint of the 'other'. *The DATA BASE for Advances in Information Systems, 37*(4), 13-25.

Kvasny, L., & Truex, D. (2001). Defining Away the Digital Divide: A Content Analysis of Institutional Influences on Popular Representations of Technology. In N. Russo, B. F.,& Janice DeGross (Ed.) *Realigning research and practice in information systems development: The social and organizational perspective*. Boston: Kluwer Academic Publishers

Madden, M. (2006). *Internet Evolution Report*. Pew Internet and American Life Project, Washington, DC. Retrieved December 2006, from http://www.pewInternet.org/PPF/r/182/report_display.asp

Mossenberg, K., & Tolbert, C. (2003). Race, Place, and Information Technology. In *Proceedings of the Telecommunications Policy Research Conference*, Arlington, VA. Retrieved December 2006, from http://intel.si.umich.edu/tprc/papers/2003/184/raceplace4.pdf

Mossberger, K., Tolbert, C., & Stansbury, K. (2003). *Virtual inequality: Beyond the digital divide* Washington, DC: Georgetown University Press.

National Telecommunications and Information Administration (1995). *Falling through the net: A survey of the have nots in rural and urban America*. Washington, DC: U.S. Dept. of Commerce. Retrieved December 2006, from http://www.ntia.doc.gov/ntiahome/fallingthru.html

National Telecommunications and Information Administration (1998). *Falling through the net II: New data on the digital divide*. Washington, DC: U.S. Dept. of Commerce. Retrieved December 2006, from http://www.ntia.doc.gov/ntiahome/net2/falling.html

National Telecommunications and Information Administration (1999). *Falling through the net: Defining the digital divide.* Washington, DC: U.S. Dept. of Commerce. Retrieved December 2006, from http://www.ntia.doc.gov./ntiahome/fttn99/contents.html

National Telecommunications and Information Administration (2000). *Falling through the net: Toward digital inclusion.* Washington, DC: U.S. Dept. of Commerce. Retrieved December 2006, from http://www.ntia.doc.gov/ntiahome/fttn00/contents00.html

Norris, P. (2001). *Digital divide: Civic engagement, information poverty and the Internet worldwide*. Cambridge University Press, Cambridge.

Novak, T. P., Hoffman, D. L., & Venkatesh, A. (1997, Oct.). *Diversity on the Internet: The relationship of race to access and usage*. Paper presented at the Forum on Diversity and the Media, Aspen Institute. Retrieved December 2006, from http://elab.vanderbilt.edu/research/papers/html/manuscripts/aspen/diversity.on.the.Internet.oct24.1997.html

Payton, F. C. (Forthcoming). Digital divide or digital equity: Other considerations? In W. Darity (Ed.), *International Encyclopedia of the Social Sciences,* (2nd ed.).

Payton, F. C. (2003). Rethinking the digital divide. *Communications of the ACM, 46*(6), 89-91.

Pew Internet and American Life Project (2006). *Demographics of Internet users*. Retrieved December 2006, from http://www.pewInternet.org/trends/User_Demo_4.26.06.htm

Sipior, J. C., Ward, B. T., Volonino, L., & Marzec, J. Z. (2004). A community initiative that diminished the digital divide. *Communications of the AIS, 13*, 29-56.

Spooner, T., & Rainie, L. (2000). *African-Americans and the Internet*. Washington, DC: Pew Internet & American Life Project. Retrieved December 2006, from http://www.pewInternet.org/report/pdfs/PIP_African_Americans_Report.pdf

Warschauer, M. (2002). Reconceptualizing the digital divide. *First Monday, 7*(4). Retrieved December 2006, from http://www.firstmonday.org/issues/issue7_7/warschauer/index.html

Youtie, J., Shapira, P., Brice, K., Laudeman, G., Young, C., Oh, E., & DiMinin, A. (2002). *Who uses LaGrange's public Internet system?* Results of the LaGrange Internet Access Survey. Georgia Institute of Technology, Atlanta.

KEY TERMS

Content: The various genres of information available on the Internet. For instance, local content is information that is specific to a community, neighborhood, or area, such as businesses, housing, neighborhood services, and recreation activities. Community content is information about the neighborhood that promotes community development and facilitates community building. Examples include a listing of places where GED courses are offered, or a newsletter. Culturally relevant content is information that is significant to people with different cultural backgrounds.

Digital Divide: Refers to the gap that exists between these who have and those who do not have access to technology (telephones, computers, Internet access) and related services.

Effective Use: The capacity and opportunity to integrate information and communication technology into the accomplishment of self or collaboratively identified goals. What is most important is not so much the physical availability of computers and the Internet but rather people's ability to make use of those technologies to engage in meaningful social practices.

Historically Underserved Groups: Refers to those who lack access to computers and the Internet. Historically this has included Americans who have low-incomes, live in rural communities, have limited education, and are members of racial or ethnic minorities.

Social Access: Refers to a mix of professional knowledge, economic resources, and technical skills to use technologies in ways that enhance professional practices and social life.

Social Inclusion: Refers to the extent that individuals, families, and communities are able to fully participate in society and control their own destinies, taking into account a variety of factors related to economic resources, employment, health, education, housing, recreation, culture, and civic engagement.

Technological Access: Refers to the physical availability of suitable information and communication technologies, including computers of adequate speed and equipped with appropriate software for a given activity.

Agent Technology

J.-J. Ch. Meyer
Utrecht University, The Netherlands

INTRODUCTION

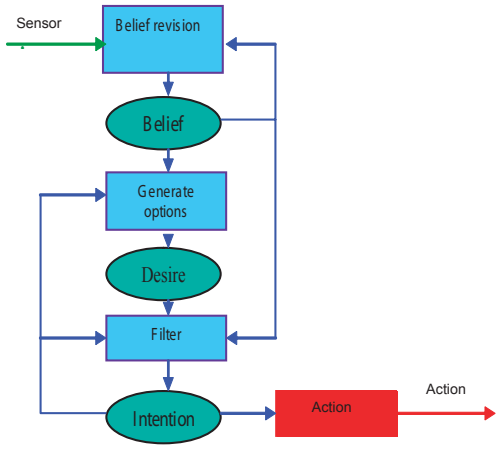
Agent technology is a rapidly growing subdiscipline of computer science on the borderline of artificial intelligence and software engineering that studies the construction of intelligent systems. It is centered around the concept of an (intelligent/rational/autonomous) *agent*. An agent is a software entity that displays some degree of autonomy; it performs actions in its environment on behalf of its user but in a relatively independent way, taking initiatives to perform actions on its own by *deliberating* its options to achieve its goal(s).

The field of agent technology emerged out of philosophical considerations about how to reason about courses of action, and human action, in particular. In analytical philosophy there is an area occupied with so-called practical reasoning, in which one studies so-called practical syllogisms, that constitute patterns of inference regarding actions. By way of an example, a practical syllogism may have the following form (Audi, 1999, p. 728):

Would that I exercise.
Jogging is exercise.
Therefore, I shall go jogging.

Although this has the form of a deductive syllogism in the familiar Aristotelian tradition of "theoretical reasoning," on closer inspection it appears that this syllogism does not express a purely logical deduction. (The conclusion does not follow logically from the premises.) It rather constitutes a representation of a *decision* of the agent (going to jog), where this decision is based on mental attitudes of the agent, namely, his/her beliefs ("jogging is exercise") and his/her desires or goals ("would that I exercise"). So, practical reasoning is "reasoning directed toward action—the process of figuring out what to do," as Wooldridge (2000, p. 21) puts it. The process of reasoning about what to do next on the basis of mental states such as beliefs and desires is called *deliberation* (see Figure 1). The philosopher Michael Bratman has argued that humans (and more generally, resource-bounded agents) also use the notion of an intention when deliberating their next action (Bratman, 1987). An intention is a desire that the agent is committed to and will try to fulfill till it believes it has achieved it or has some other rational reason to abandon it. Thus, we could say that agents, given their beliefs and desires, choose some desire as their intention, and "go for it." This philosophical theory has been formalized through several studies, in particular the work of Cohen and Levesque (1990); Rao and Georgeff (1991); and Van der Hoek, Van Linder, and Meyer (1998), and has led to the so-called Belief-Desire-Intention (BDI) model of intelligent or rational agents (Rao & Georgeff, 1991). Since the beginning of the 1990s researchers have turned to the problem of realizing artificial agents. We will return to this hereafter.

Figure 1. The deliberation process in a BDI architecture



BACKGROUND: THE DEFINITION OF AGENTHOOD

Although there is no generally accepted definition of an agent, there is some consensus on the (possible) properties of an agent (Wooldridge, 2002; Wooldridge & Jennings, 1995): Agents are hardware or software-based computer systems that enjoy the properties of:

- **Autonomy:** The agent operates without the direct intervention of humans or other agents and has some control over its own actions and internal state.
- **Reactivity:** Agents perceive their environment and react to it in a timely fashion.
- **Pro-Activity:** Agents take initiatives to perform actions and may set and pursue their own goals.

- **Social Ability:** Agents interact with other agents (and humans) by communication; they may coordinate and cooperate while performing tasks.

Thus we see that agents have both informational and motivational attitudes, namely, they handle and act upon certain types of information (such as knowledge, or rather beliefs) as well as motivations (such as goals). Many researchers adhere to a stronger notion of agency, sometimes referred to as *cognitive* agents, which are agents that realize the aforementioned properties by means of mentalistic attitudes, pertaining to some notion of a mental state, involving such notions as knowledge, beliefs, desires, intentions, goals, plans, commitments, and so forth. The idea behind this is that through these mentalistic attitudes the agent can achieve autonomous, reactive, proactive, and social behavior in a way that is mimicking or at least inspired by the human way of thinking and acting. So, in a way we may regard agent technology as a modern incarnation of the old ideal of creating intelligent artifacts in artificial intelligence. The aforementioned BDI model of agents provides a typical example of this strong notion of agency and has served as a guide for much work on agents, both theoretical (on BDI logic) and practical (on the BDI architecture, see next section).

CURRENT AGENT RESEARCH: MULTI-AGENT SYSTEMS

Agent-based systems become truly interesting and useful if we have multiple agents at our disposal sharing the same environment. Here we have to deal with a number of more or less autonomous agents interacting with each other. Such systems are called multi-agent systems (MAS) (Wooldridge, 2002) or sometimes also agent societies. Naturally, these systems will generally involve some kind of *communication* between agents. Agents may communicate by means of a communication primitive such as a *send* (agent, performative, content), which has as semantics to send to the agent specified the content specified with a certain illocutionary force, specified by the performative, for example, inform or request. The area of agent communication (and agent communication languages) is a field of research in itself (Dignum & Greaves, 2000). Further, it depends on the application whether one may assume that the agents in a multi-agent system cooperate or compete with each other. But even in the case of cooperation it is not a priori obvious how autonomous agents will react to requests from other agents: Since they ultimately have their own goals, it may be the case that they do not have the time or simply do not want to comply. This also depends on the *social structure* within the agent system/society, in particular the type(s) of coordination mechanisms and power relations employed, such as a market, network, or hierarchy, and the *role* the agent plays in it (Dignum, Meyer, Weigand, & Dignum, 2002).

Another issue related to agent communication, particularly within heterogeneous agent societies, concerns the language (ontology) agents use to reason about their beliefs and communicate with each other. Of course, if agents stem from different sources (designers) and have different tasks they will generally employ different and distinct ontologies (concepts and their representations) for performing their tasks. When communicating it is generally not efficacious to try to transmit their whole ontologies to each other or to translate everything into one giant *universal* ontology if this would exist anyway. Rather it seems that a kind of "ontology negotiation" should take place to arrive at a kind of minimal solution (i.e., sharing of concepts) that will facilitate communication between those agents (Bailin & Truszkowski, 2002; Van Diggelen, Beun, Dignum, Van Eijk, & Meyer, 2006).

Next we turn to the issue of constructing agent-based systems. Since the philosophical and logical work on intelligent agents mentioned in the introduction, researchers have embarked upon the enterprise of realizing agent-based systems. Architectures for agents have been proposed, such as the well-known BDI architecture (Rao & Georgeff, 1991) and its derivatives procedural reasoning system (PRS) and the InteRRap architecture (Wooldridge, 2002). Other researchers devised dedicated agent-oriented programming (AOP) languages to program agents directly in terms of mentalistic notions in the same spirit as the ones mentioned previously. The first researcher who proposed this approach was Shoham with the language AGENT0 (Shoham, 1993). Other languages include AgentSpeak(L)/Jason, (Concurrent) METATEM, CONGOLOG, JACK, JADE, JADEX, and 3APL (Bordini, Dastani, Dix, & El Fallah Seghrouchni, 2005; de Giacomo, Lespérance, & Levesque, 2000; Fisher, 1994; Hindriks, de Boer, Van der Hoek, & Meyer, 1999). One may also adhere to programming agents directly in generic programming languages such as JAVA and C++. Interestingly, one may now ask questions such as how to program agents in these languages in the same way that these questions are asked in software engineering with respect to traditional programming. The thus emerging subfield is called agent-oriented software engineering (AOSE) (Ciancarini & Wooldridge, 2001). In our opinion this makes the most sense in the interpretation of engineering agent-oriented software rather than trying to engineer arbitrary software in an agent-oriented way. So we advocate using an agent-oriented design together with an explicit agent-oriented implementation in an AOP language for those applications that are fit for this approach (Dastani, Hulstijn, Dignum, & Meyer, 2004). Obviously we do not advocate performing all programming (e.g., a sorting algorithm) in an agent-oriented way.

This brings us to the important question, what kind of application is particularly suited for an agent-based solution?

Although it is hard to say something about this in general, one may particularly think of applications where (e.g., logistical/planning) tasks may be distributed in such a way that subtasks may be (or rather preferably so) performed by autonomous entities (agents). For instance in situations where it is virtually impossible to perform the task centrally, due to either the complexity of the task or a large degree of dynamics (changes) in the environment in which the system has to operate. Of course, also in applications where there is a natural notion of a *cognitive* agent, such as, for example, in the gaming world where virtual characters need to behave in a natural/believable way and have human-like features, agents seem to be the obvious choice for realization. For the same reason, in (multi-)robotic applications agents may be used, too, where each robot may be controlled by its own agent, and therefore can display autonomous but, for instance, also cooperative behavior. Furthermore, also in e-commerce/e-business and in applications for the Web in general, where agents may act on behalf of a user, agent technology seems adequate.

At the moment agent research is quite big. There are many workshops addressing aspects of agents, and there is the large annual international *Autonomous Agents and Multi-Agent Systems (AAMAS) Conference*, the most authoritative event in the field, attracting over 700 people. There has been a specialization into many topics. To give an impression: AAMAS 2006 (http://www.fun.ac.jp/aamas2006/) lists, in its call for papers, the following list of topics (categorized by the author for ease of reading).

Agent Theories

- Agents and cognitive models
- Agents and adjustable autonomy
- Argumentation in agent systems
- Coalition formation and teamwork
- Collective and emergent agent behavior
- Computational complexity in agent systems
- Conventions, commitments, norms, social laws
- Cooperation and coordination among agents
- Formal models of agency
- Game theoretic foundations of agent systems
- Legal issues raised by autonomous agents
- Logics for agent systems
- (Multi-)agent evolution, adaptation and learning
- (Multi-)agent planning
- Ontologies and agent systems
- Privacy, safety and security in agent systems
- Social choice mechanisms
- Social and organizational structures of agent systems
- Specification languages for agent systems
- Computational autonomy
- Trust and reputation in agent systems
- Verification and validation of agent systems

Agent Construction: Design and Implementation

- Agent and multi-agent architectures
- Agent communication: languages, semantics, pragmatics, protocols
- Agent programming languages
- Agent-oriented software engineering and agent-oriented methodologies
- Electronic institutions
- Frameworks, infrastructures and environments for agent systems
- Performance evaluation of agent systems
- Scalability, robustness and dependability of agent systems

Agent Applications

- Agent standardizations in industry and commerce
- Agents and ambient intelligence
- Agents and novel computing paradigms (e.g. autonomic, grid, P2P, ubiquitous computing)
- Agents, web services and semantic web
- Agent-based simulation and modeling
- Agent-mediated electronic commerce and trading agents
- Applications of autonomous agents and multi-agent systems
- Artificial social systems
- Auctions and electronic markets
- Autonomous robots and robot teams
- Constraint processing in agent systems
- Conversation and dialog in agent systems
- Cooperative distributed problem solving in agent systems
- Humanoid and sociable robots
- Information agents, brokering and matchmaking
- Mobile agents
- Negotiation and conflict handling in agent systems
- Perception, action and planning in agents
- Synthetic, embodied, emotional and believable agents
- Task and resource allocation in agent systems

It is beyond the scope and purpose of this article to discuss all of these. Many issues mentioned in this huge list have been touched upon already. Some others will be addressed in the next section on future trends, since they are still much a matter of pioneering research.

FUTURE TRENDS

It is interesting to speculate on the future of agents. Will it provide a new programming paradigm as a kind of successor to object-oriented programming? Time will tell. The trends at the moment are diverse and ambitious. One trend is to go beyond rationality in the BDI sense, and include also social notions such as norms and obligations into the agent framework (Dignum, 1999), and even also "irrational" notions such as emotions (Meyer, 2004). So-called *electronic institutions* have been proposed to regulate the behavior of agents in compliance with the norms in force (Esteva, Padget, & Sierra, 2001). Another trend is to look at how one may devise "hybrid" agent systems with "humans in the loop" (Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004). Since it is deemed imperative to check the correctness of complex agent-based systems for very costly and life-critical applications, yet another trend is the development of formal verification techniques for agents and model checking in particular (Rash, Rouff, Truszkowski, Gordon, & Hinchey, 2001).

CONCLUSION

In this short article we have seen how the idea of agent technology and agent-oriented programming evolved from philosophical considerations about human action to a way of programming intelligent artificial (computer-based) systems. Since it is a way to construct complex intelligent systems in a structured and anthropomorphic way, it appears to be a technology that is widely applicable, and it may well become one of the main programming paradigms of the future.

REFERENCES

Audi, R. (Ed.). (1999). *The cambridge dictionary of philosophy.* Cambridge, UK: Cambridge University Press.

Bailin, S., & Truszkowski, W. (2002). Ontology negotiation between intelligent information agents. *Knowledge Engineering Review, 17*(1), 7-19.

Bordini, R. H., Dastani, M., Dix, J., & El Fallah Seghrouchni, A. (Eds.). (2005). *Multi-agent programming (Languages, platforms and applications).* New York: Springer Science.

Bratman, M. E. (1987). *Intentions, plans, and practical reason.* Cambridge, MA: Harvard University Press.

Ciancarini, P., & Wooldridge, M. J. (Eds.). (2001). *Agent-oriented software engineering. Lecture notes in artificial intelligence 1957.* Berlin, Germany: Springer.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence, 42*(3), 213-261.

Dastani, M., Hulstijn, J., Dignum, F., & Meyer, J.-J. Ch. (2004). Issues in multiagent system development. In N. R. Jennings, C. Sierra, L. Sonenberg, & M. Tambe (Eds.), *Proceedings of the 3rd International Joint Conference On Autonomous Agents & Multi Agent Systems (AAMAS 2004)* (pp. 922-929). New York: ACM Press.

de Giacomo, G., Lespérance, Y., & Levesque, H. (2000). ConGolog, a concurrent programming language based on the situation calculus. *Artificial Intelligence, 121*(1,2), 109-169.

Dignum, F. (1999). Autonomous agents with norms. *Artificial Intelligence and Law, 7*, 69-79.

Dignum, F., & Greaves, M. (Eds.). (2000). *Issues in agent communication. Lecture notes in artificial intelligence 1906.* Berlin, Germany: Springer.

Dignum, V., Meyer, J.-J. Ch., Weigand, H., & Dignum, F. (2002). An organisational-oriented model for agent societies. In G. Lindemann, D. Moldt, M. Paolucci, & B. Yu (Eds.), *Proceedings of the International Workshop on Regulated Agent-Based Social Systems: Theory and Applications (RASTA'02)* (FBI—HH-M-318/02, 31-50). Germany: University of Hamburg.

Esteva, M., Padget, J., & Sierra, C. (2001). Formalizing a language for institutions and norms. In J.-J. Ch. Meyer & M. Tambe (Eds.), *Intelligent agents VIII. Lecture notes in artificial intelligence 2333* (pp. 348-366). Berlin, Germany: Springer.

Fisher, M. (1994). A survey of concurrent METATEM—The language and its applications. In D. M. Gabbay & H. J. Ohlbach (Eds.), *Temporal logic. Lecture notes in artificial intelligence 827* (pp. 480-505). Berlin, Germany: Springer.

Hindriks, K. V., de Boer, F. S., Van der Hoek, W., & Meyer, J.-J. Ch. (1999). Agent programming in 3APL. *International Journal of Autonomous Agents and Multi-Agent Systems, 2*(4), 357-401.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "Team Player" in joint human-agent activity. *IEEE Intelligent Systems, 19*(6), 91-95.

Meyer, J.-J. Ch. (2004), Reasoning about emotional agents. In R. López de Mántaras & L. Saitta (Eds.), *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)* (pp. 129-133). Amsterdam: IOS Press.

Rao, A. S., & Georgeff, M. P. (1991). Modeling rational agents within a BDI-architecture. In J. Allen, R. Fikes, & E. Sandewall (Eds.), *Proceedings of the 1991 Conference On Knowledge Representation (KR'91)* (pp. 473-484). Los Altos, CA: Morgan Kaufmann.

Rash, J. L., Rouff, C. A., Truszkowski, W., Gordon, D., & Hinchey, M. G. (Eds.). (2001). *Proceedings of the First Goddard Workshop on Formal Approaches to Agent-Based Systems (FAABS 2000).* Lecture notes in artificial intelligence 1871. Berlin/Heidelberg, Germany: Springer.

Shoham, Y. (1993). Agent-oriented programming. *Artificial Intelligence, 60*(1), 51-92.

Van der Hoek, W., Van Linder, B., & Meyer, J.-J. Ch. (1998). An integrated modal approach to rational agents. In M. Wooldridge & A. Rao (Eds.), *Foundations of rational agency* (pp. 133-168). Dordrecht, The Netherlands: Kluwer.

Van Diggelen, J., Beun, R. J., Dignum, F., Van Eijk, R., & Meyer, J.-J. Ch. (2006). ANEMONE: An effective minimal ontology negotiation environment. In P. Stone & G. Weiss (Eds.). *Proceedings of the 5th International Joint Conference on Autonomous Agents & Multi Agent Systems (AAMAS 2006)* (pp. 899-906). New York: ACM Press

Wooldridge, M. J. (2000). *Reasoning about rational agents.* Cambridge, MA: MIT Press.

Wooldridge, M. J. (2002). *An introduction to multiagent systems.* Chichester, UK: John Wiley & Sons.

Wooldridge, M. J., & Jennings, N. R. (Eds.). (1995). Intelligent agents. Lecture Notes in Artificial Intelligence, 890. Berlin, Germany: Springer.

KEY TERMS

Agent-Oriented Programming (AOP): AOP is an approach to constructing agents by means of programming them in terms of mentalistic notions such as beliefs, desires, and intentions.

Agent-Oriented Programming Language: AOP Language is a language that enables the programmer to program intelligent agents, as defined previously, (in the strong sense) in terms of agent-oriented (mentalistic) notions such as beliefs, goals, and plans.

Agent-Oriented Software Engineering (AOSE): AOSE is the study of the construction of intelligent systems by the use of the agent paradigm, that is, using agent-oriented notions, in any high-level, programming language. In a strict sense: the study of the implementation of agent systems by means of agent-oriented programming languages.

Autonomous Agent: Autonomous agent is an agent that acts without human intervention and has some degree of control over its internal state.

Believable Agent: Believable agent is an agent, typically occurring in a virtual environment that displays *natural* behavior, such that the user of the system may regard it as an entity that interacts with him/her in a natural way.

Electronic Institution: Electronic institution is a (sub)system to regulate the behavior of agents in a multi-agent system/agent society, in particular their interaction, in compliance with the norms in force in that society.

Intelligent Agent: Intelligent agent is a software or hardware entity that displays (some degree of) autonomous, reactive, proactive, and social behavior; in a strong sense: an agent that possesses mental or cognitive attitudes, such as beliefs, desires, goals, intentions, plans, commitments, and so forth.

Multi-Agent System (MAS)/Agent Society: MAS agent society is a collection of agents sharing the same environment and possibly also tasks, goals, and norms, and therefore are part of the same organization.

Pro-Active Agent: Pro-active agent is an agent that takes initiatives to perform actions and may set and pursue its own goals.

Reactive Agent: Reactive agent is an agent that perceives its environment and reacts to it in a timely fashion.

Social Agent: Social agent is an agent that interacts with other agents (and humans) by communication; it may coordinate and cooperate with other agents while performing tasks.

Agent-Based Negotiation in E-Marketing

V.K. Murthy
University of New South Wales, Australia

E.V. Krishnamurthy
Australian National University, Australia

INTRODUCTION

This article describes in brief the design of agent-based negotiation system in e-marketing. Such a negotiation scheme requires the construction of a suitable set of rules, called protocol, among the participating agents. The construction of the protocol is carried out in two stages: first expressing a program into an object-based rule system and then converting the rule applications into a set of agent-based transactions on a database of active objects represented using high-level data structures.

BACKGROUND

An agent is a code-containing object, that along with data and execution context can migrate autonomously and purposefully within a computer network. Thus an agent knows what to do with the information obtained from its environment. Agents behave like actors and have intentions and actions. In addition, agents are flexible, proactive and have multithreaded control. In this overview, we describe in detail how a set of agents can be used for negotiation in e-marketing. For this purpose we need to have a model of the multi agent-based paradigm for executing the negotiation process analogous to what we humans do. Negotiation is an interactive process among a number of agents that results in varying degrees of cooperation, competition and ultimately to commitment that leads to a total agreement, consensus or a disagreement. It has many applications, including economics, psychology, sociology and computer science.

MAIN THRUST OF THE ARTICLE

The following subsections bring out the main thrust of this chapter, namely: what is a multi-agent system, what is planning, reasoning and negotiation, and how agents can be useful in modeling e-market and e-auction. Also we will briefly describe how a coalition among agents can cause a speculation bubble or a crash in e-share market.

A MULTI-AGENT SYSTEM

A simple model of an agent that is suitable for our purpose is shown in Figure 1. This is a unified model based on several important contributions made by the following authors: (Chen & Dayal, 2000; Dignum & Sierra, 2001; Fisher, 1995; Genesereth & Nilsson, 1987; Ishida, 1994; Murthy, 2002; Woolridge, 2002).

As shown in Figure 1, an agent consists of the following subsystems:

(1) Worldly states or environment U: Those states which completely describe the universe containing all the agents.

(2) Percept: Depending upon the sensory capabilities (input interface to the universe or environment), an agent can partition U into a standard set of messages T, using a sensory function Perception (PERCEPT): PERCEPT :U $\rightarrow$ T.
PERCEPT can involve various types of perception: see, read, hear, smell. The messages are assumed to be of standard types based on an interaction language that is interpreted identically by all agents.

(3) Epistemic states or Mind M: Weassumethattheagent has a mind M (that is essentially a problem domain knowledge consisting of an internal database for the problem domain data and a set of problem domain rules) that can be clearly understood by the agent without involving any sensory function. The database D sentences are in first order predicate calculus (also known as extensional database) and agents' mental actions are viewed as inferences arising from the associated rules that result in an intentional database, whcih changes (revises or updates) D.
The agent's state of belief, or a representation of an agent's state of belief at a certain time, is represented by an ordered pair of elements (D, P). D is a set of beliefs about objects, their attributes and relationships stored as an internal database and P is a set of rules expressed as preconditions and consequences (conditions and actions). When T is input, if the conditions given in the left-hand side of P match T, the elements from D that correspond to the right-hand side are taken

Figure 1.

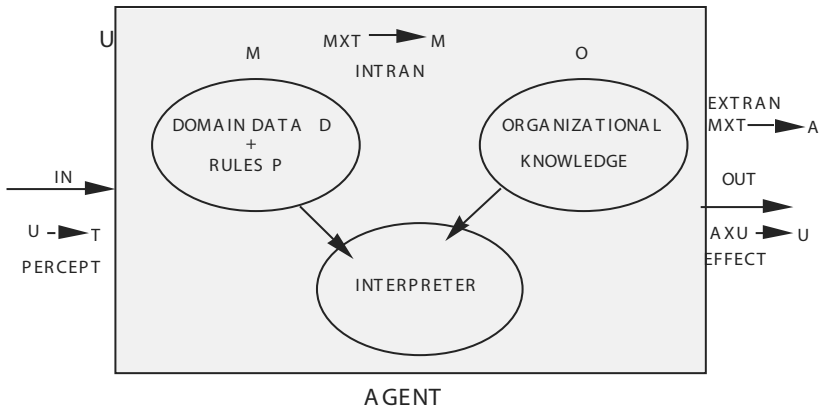


Figure 1

from D and suitable actions are carried out locally (in M) as well as on the environment.

(4) Organizational Knowledge (O): Since each agent needs to communicate with the external world or other agents, we assume that O contains all the information about the relationships among the different agents. For example, the connectivity relationship for communication, the data dependencies between agents, and interference among agents with respect to rules. Information about the location of different domain rules is in O.

(5) INTRAN: M is suitably revised or updated by the function called internal transaction (INTRAN).

Revision: Revision means acquisition of new information about the environment that requires a change in the rule system P. This may result in changes in the database D.

Example: The inclusion of a new tax rule in the tax system.

Update: Update means adding new entries to the database D; the rules P are not changed.

Example: Inclusion of a new tax-payer in the tax system.

Both revision and update can be denoted in set-theoretic notation by: INTRAN: $M \times T \rightarrow M$

(6) EXTRAN: External action is defined through a function called global or external transaction (EXTRAN) that maps an epistemic state and a partition from an external state into an action performed by the agent. That is: EXTRAN: $M \times T \rightarrow A$
This means that the current state of mind and a new input activates an external action from A.

(7) EFFECT: The agent also can affect U by performing an action from a set of actions A (ask, tell, hear, read, write, speak, send, smell, taste, receive, silent), or more complex actions. Such actions are carried out according to a particular agent's role and governed by an etiquette called protocols. The effect of these actions is defined by a function EFFECT, which modifies the world states through the actions of an agent:

EFFECT: $A \times U \rightarrow U$; EFFECT can involve additions, deletions and modifications to U.

Thus an agent is defined by a set of nine entities, called a 9-tuple:

(U,T,M(P,D),O,A,PERCEPT,INTRAN,EXTRAN,EFFECT).

The interpreter repeatedly executes selected rules in P, until no rule can be fired.

We can interpret all the abstract machine models (such as a finite state machine or a Turing machine) and parallel computational models (such as classifier systems) as subclasses of the agents, by suitably formulating the definitions.

The nature of internal production rules P, their mode of application, and the action set A determines whether an agent is deterministic, nondeterministic, probabilistic or fuzzy. Rule application policy in a production system P can be modified by:

(1) Assigning probabilities/fuzziness for applying the rule

(2) Assigning strength to each rule by using a measure of its past success

(3) Introducing a support for each rule by using a measure of its likely relevance to the current situation

The preceding three factors provide for competition and cooperation among the different rules. Such a model is useful for negotiation in learning, as well as in e-marketing that involves interactions between many agents.

WHAT IS NEGOTIATION?

A negotiation protocol is viewed as a set of public rules that dictate the conduct of an agent with other agents to achieve a desired final outcome.

A negotiation protocol among agents involves the following actions or conversational states:

(1) Propose: one puts forward for consideration a set of intentions called a proposal.
(2) Accept: The proposal is accepted for execution into actions.
(3) Refuse: The proposal is rejected for execution into actions.
(4) Modify: This alters some of the intentions of the proposer and suggests a modified proposal that is at the worst, a refuse and a new proposal; or a partial acceptance and new additions.
(5) No proposal: No negotiation.
(6) Abort: Quit negotiation.
(7) Report agreement: This is the termination point for negotiation in order to begin executing actions.
(8) Report failure (agree to disagree): Negotiation breaks down.

Note that the previous actions are not simple exchange of messages but may involve some intelligent or smart computation.

Multiagents can cooperate to achieve a common goal to complete a transaction to aid the customer. The negotiation follows rule-based strategies that are computed locally by its host server. Here competing offers are to be considered; occasionally cooperation may be required. Special rules may be needed to take care of risk factors, domain knowledge dependencies between attributes, and positive and negative end conditions. When making a transaction several agents have to negotiate and converge to some final set of values that satisfies their common goal. Such a goal should also be cost effective so that it is in an agreed state at the minimum cost or a utility function. To choose an optimal strategy each agent must build a plan of action and communicate with other agents.

PLANNING, REASONING AND NEGOTIATION

The negotiation process is usually preceded by two cooperating interactive processes: planning and reasoning (Woolridge, 2000). The ability to plan ahead for solving a problem is the key aspect of intelligent behavior. To solve a problem through negotiation, we start with a set of desired properties and try to devise a plan that results in a final state with the desired properties. For this purpose, we define an initial state where we begin an operation and also define a desirable goal state or a set of goal states. Simultaneously, we use a reasoning scheme and define a set of intended actions that can convert a given initial state to a desired goal state or states. Such a set of intended actions called the plan exists if and only if it can achieve a goal state starting from an initial state and moving through a succession of states. Therefore to begin the negotiation process, we need to look for a precondition that is a negation of the goal state and look for actions that can achieve the goal. This strategy is used widely in AI and forms the basis to plan a negotiation (Genersereth & Nilsson, 1987). Such a planning is possible for clear-cut algorithmic problems. For general AI problems, however, we can only generate a plan that may or may not work; if the plan does not work we need to either modify the plan or devise a new plan. The same approach is used for devising a multi-agent negotiation protocol that is useful in e-auction, e-marketing and in other applications (Horlait et al., 2003; Marik et al., 2003; Schillo et al., 2003; Woolridge, 2000).

We now describe how to carry out distributed multi-agent negotiation by sending, receiving, handshaking and acknowledging messages and performing some local computations.

A multi-agent negotiation has the following features (Dignum & Sierra, 2001; Murthy, 2002):

(1) There is a seeding agent who initiates the negotiation and coordinates the negotiation.
(2) Each agent can be active or inactive.
(3) Initially all agents are inactive except for a specified seeding agent that initiates the computation.
(4) An active agent can do local computation, send and receive messages and can spontaneously become inactive.
(5) An inactive agent becomes active, if and only if it receives a message.
(6) Each agent may retain its current belief or revise its belief as a result of receiving a new message by performing a local computation. If it revises its belief, it communicates its revised state of belief to other concerned agents; else it does not revise its solution and remains silent.
(7) Finally, the negotiation terminates, or fails. If there is no consensus within a set time-out limit, the negotiation fails.

FUTURE TRENDS

Agent based systems will provide new approaches to modeling and simulation of complex business systems and also provide for new software engineering approaches (Lucena

et al., 2004; Sichman et al., 2003). We will illustrate this by a simple example on modeling e-market and e-auction.

MODELING E-MARKET AND E-AUCTION

The agent negotiation system can model the e-market with many traders (agents), popularly known as buyers and sellers. These agents negotiate over the Internet to sell or buy shares or stocks in a stock market. In an e-market situation, it is possible that the negotiation ultimately leads to self-organization and criticality, causing crashes. That is, individual agents that correspond to a microscopic system can emerge as a self-organizing macroscopic system corresponding to a "percolation model" (Paul & Baschnagel, 1999).

In e-market situation (see Figure 1), to start with, the domain data D, rules P and organizational knowledge O can be based on three factors:

(1) the experience and economics knowledge of an agent deployed by a trader based totally on individualistic idiosyncratic criteria.
(2) the trader's acquired knowledge through communication with other selected agents; such a trader is called a fundamentalist.
(3) the trader's acquired knowledge by observing the trends on market from a collective opinion of other traders; such a trader is called a trend chaser.

The previous three factors play an important role in deciding the number of possible states that each agent will be in and his or her inclination to buy or sell or wait in an e-marketing decision.

In e-market, at every time instant a trader can adopt three possible states of action: buy, sell or wait, respectively represented by three states 1, -1 and 0. Each agent representing a trader can communicate with one another and this creates an imaginary bond or connectivity relationship among them, modifying the organizational knowledge O. This bond is created with a certain probability determined by a single parameter that characterizes the willingness of an agent to comply with others.

The three states of behavior are obviously a very complicated function of the behavioral property and personality of an individual and whether he or she uses elementary, derived or inferential beliefs. It is interesting to note that all the beliefs are again a function of the speed with which information is available to an agent, financial status, ability to reason and susceptibility to pressure from neighbors. Thus in a share market or auction situation we need to work out how the agents are linked in order to obtain information through communication, the personality factors such as age, financial status and the market trend. Using data mining techniques one can derive detailed information about the mechanism of bond formation among the agents. Based on this information, we can assume that any two agents are randomly connected with a certain probability. This will divide the agents into clusters of different sizes whose members are linked either directly or indirectly via a chain of intermediate agents. These groups are coalitions of market participants who share the same opinion about their activity. The decision of each group is independent of its size and the decision taken by other clusters. In this situation, using the random cluster model (Bak, 1996; Paul & Baschnagel, 1999) we can show that when every trader is on average connected to another, more and more traders join the spanning cluster, and the cluster begins to dominate the overall behavior of the system. This can give rise to "speculation bubble" (if the members all decide to buy), a crash (if the members all decide to sell) or a stagnation (if the members all decide to wait). These are cooperative phenomenon and depend upon trading rules, exchange of information- the speed and volume, and the connectivity relationship. For the three-state agents the critical probability $p(c) = 0.63$. That is, if an agent is even showing about 63% preference to the information from his or her neighbors, a crash or bubble is bound to happen. Crash is a highly cooperative phenomenon and depends upon trading rules, exchange of information- the speed and volume, and the connectivity relationship. Accordingly, an analogy exists between stock-market crashes and critical phenomena or phase transitions in physics. Thus a distributed agent system can eventually enter into a self-organized critical state or an emergent state. However, such a phenomenon is similar to an earthquake or epidemic, or fatigue in materials resulting in cracks and is hard to predict, since the cluster formation depends upon the various factors already mentioned.

CONCLUSION

This overview explained the use of multi-agent based planning, reasoning and negotiation in e-marketing. We explained how to use the techniques of AI planning to devise the multi-agent based negotiation protocol. We also explained how agents can reach a self-organized critical state that can result in crash, speculation bubble or stagnation in e-share market. The study of agent systems and their applications is a vastly expanding area for research; for more details, the following references may be consulted: (Horlait et al., 2003; Marik et al., 2003; Schillo et al., 2003; Woolridge, 2000).

REFERENCES

Bak, B. (1996). *How nature works: The science of self-organized criticality.* New York: Springer.

Chen, Q., & Dayal, U. (2000). Multi agent cooperative transactions for e-commerce. *Lecture Notes in Computer Science, 1901,* 311-322. New York: Springer Verlag.

Dignum, F., & Sierra, C. (Eds.). (2001). Agent mediated e-commerce. *Lecture Notes in Artificial Intelligence, 1991, 2003.* New York: Springer Verlag.

Fisher, M. (1995). Representing and executing agent-based systems. *Lecture Notes in Computer Science, 890,* 307-323. New York: Springer-Verlag.

Genesereth, M.R., & Nilsson, N.J. (1987). *Logical foundations of artificial intelligence.* New York: Morgan Kaufmann.

Horlait, E., Magedanz, T., & Glitho, R.H. (Eds.). (2003). Mobile agents for telecommunication applications. *Lecture Notes In Artificial Intelligence, 2691.* New York: Springer Verlag.

Ishida, T. (1994). Parallel, distributed and multiagent production systems. *Lecture Notes in Computer Science, 878.* New York: Springer Verlag.

Lucena, C. et al. (2004). Software engineering for multi-agent systems II . *Lecture Notes in Computer Science, 2940.* New York: Springer Verlag.

Marik, V., Muller, J., & Pechoucek, M. (Eds.). (2003). Multi-agent systems and applications. *Lecture Notes In Artificial Intelligence, 2691.* New York: Springer Verlag.

Murthy, V.K. (2002). Designing agent-based negotiation for e-marketing. In N. Shi & V.K. Murthy (Eds.), *Architectural issues of Web-enabled electronic business* (pp. 290-302). Hershey, PA: Idea Group Publishing.

Paul, W., & Baschnagel, J. (1999). *Stochastic processes.* New York: Springer Verlag.

Schillo, M., Klusch, M., Muller, J., & Tianfield, H. (Eds). (2003). Multi-agent system technologies. *Lecture Notes in Artificial Intelligence, 2831.* New York: Springer Verlag.

Sichman, J.S. et al. (2003). Multi-agent based simulation II. *Lecture Notes in Artificial Intelligence, 2581.* New York: Springer Verlag.

Woolridge, M. (2000). *Reasoning about rational agents.* Cambridge, MA: MIT Press.

Woolridge, M. (2002). *An introduction to multi-agent systems.* New York: John Wiley.

KEY TERMS

Agent: A system that is capable of perceiving events in its environment, or representing information about the current state of affairs and of acting in its environment guided by perceptions and stored information (current definition by AOIS, agent oriented information system community).

Coordinator: An that acts as a coordinator among several agents to carry out the negotiation protocol.

E-Auction: This is a centralized protocol for redistributing resources among agents. Each agent attaches a value to each resource. The seller asks a price for a resource and buyer offers a price and they negotiate over the Internet to achieve a desired outcome satisfying to both; else the negotiation fails.

E-Market: An Internet based market with many traders (agents) popularly known as buyers and sellers. These agents negotiate over the Internet to sell or buy products in any market (e.g., shares or stocks in a stock market).

Negotiation: This is an interactive process among a number of agents that results in varying degrees of cooperation, competition and ultimately to commitment that leads to a total agreement, consensus governed by a voting policy, or a disagreement.

Negotiation Protocol: A negotiation protocol is viewed as a set of public rules that dictate the conduct of an agent with other agents to achieve a desired final outcome.

Protocols: A set of rules that dictate the behavior of objects for communication and interaction.

Self-Organized Critical State or Emergence: A group of interconnected communicating, interacting and negotiating agents reach a state of self-organization resulting in an unpredictable decision—such as a share-market crash, a stagnation or a speculation bubble.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 88-92, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Agent-Oriented Software Engineering

Kuldar Taveter
The University of Melbourne, Australia

Leon Sterling
The University of Melbourne, Australia

INTRODUCTION

Over the past decade, the target environment for software development has complexified dramatically. Software systems must now operate robustly in a dynamic, global, networked environment comprised of distributed diverse technologies, where frequent change is inevitable. There is increasing demand for flexibility and ease of use. Multi-agent systems (Wooldridge, 2002) are a potential successor to object-oriented systems, better able to address the new demands on software. In multi-agent systems, heterogeneous autonomous entities (i.e., *agents*) interact to achieve system goals. In addition to being a technological building block, an agent, also known as an *actor*, is an important modeling abstraction that can be used at different stages of software engineering. The authors while teaching agent-related subjects and interacting with industry have observed that the agent serves as a powerful anthropomorphic notion readily understood by novices. It is easy to explain to even a non-technical person that one or more software agents are going to perform a set of tasks on your behalf.

We define *software engineering* as a discipline applied by teams to produce high-quality, large-scale, cost-effective software that satisfies the users' needs and can be maintained over time. Methods and processes are emerging to place software development on a parallel with other engineering endeavors. Software engineering courses give increasing focus to teaching students how to analyze software designs, emphasizing imbuing software with quality attributes such as performance, correctness, scalability, and security.

Agent-oriented software engineering (AOSE) (Ciancarini & Wooldridge, 2001) has become an active research area. Agent-oriented methodologies, such as Tropos (Bresciani, Perini, Giorgini, Giunchiglia, & Mylopoulos, 2004), ROADMAP (Juan & Sterling, 2003), and RAP/AOR (Taveter & Wagner, 2005), use the notion of agent throughout the software lifecycle from analyzing the problem domain to maintaining the functional software system. An agent-oriented approach can be useful even when the resulting system neither consists of nor includes software agents. Some other proposed AOSE methodologies are Gaia (Wooldridge, Jennings, & Kinny, 2000), MESSAGE (Garijo, Gomez-Sanz, & Massonet, 2005), TAO (Silva & Lucena, 2004), and Prometheus (Padgham & Winikoff, 2004). Although none of the AOSE methodologies are yet widely accepted, AOSE is a promising area. The recent book by Henderson-Sellers & Giorgini (2005) contains a good overview of currently available agent-oriented methodologies.

AOSE approaches loosely fall into one of two categories. One approach adds agent extensions to an existing object-oriented notation. The prototypical example is Agent UML (Odell, Van Dyke, & Bauer, 2001). The alternate approach builds a custom software methodology around agent concepts such as roles. Gaia (Wooldridge et al., 2000) was the pioneering example.

In this article, we address the new paradigm of AOSE for developing both agent-based and traditional software systems.

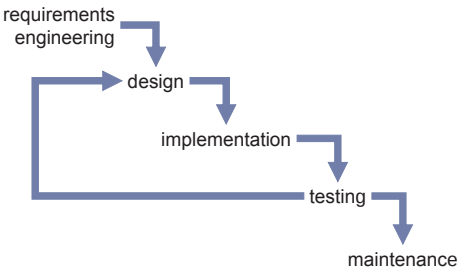
BACKGROUND

Software engineering deals with sociotechnical systems. Sommerville (2004) defines a *sociotechnical system* as one that includes hardware and software, has defined operational processes, and offers an interface, implemented in software, to human users. Software engineering addresses developing software components of sociotechnical systems. Software engineering is therefore critical for the successful development of complex, computer-based, sociotechnical systems because a software engineer should have a broad awareness of how that software interacts with other hardware and software systems and its intended use, not only the software itself (Sommerville, 2004).

A conventional *software engineering process* represented in Figure 1 contains the stages of requirements engineering, design, implementation, testing, and maintenance.

Requirements engineering—understanding, and specifying the user's needs—is followed by *software design* consisting of both high-level architectural design of the system and detailed design of the software components. Implementation follows, increasingly in the form of code generation using different Computer-aided software engineering (CASE) tools. The code must then be tested to uncover and correct as many errors as possible before delivery to the customer. Pressman (2001) provides an overview of many potentially

Figure 1. A software engineering process



useful testing strategies and methods. Finally, there is ongoing maintenance.

Object-oriented software engineering (OOSE), which numerous AOSE methodologies build on, adopts these conventional steps. Object-oriented approaches characterize the problem as a set of objects with specific attributes and behaviors (Pressman, 2001). Objects are categorized into classes and subclasses. Object-oriented analysis identifies classes and objects relevant to the problem domain; design provides architecture, interface, and component-level detail; and implementation (using an object-oriented language) transforms design into code.

To facilitate OOSE, the Unified Modelling Language (UML) (Object Management Group (OMG) 2003a, 2003b) for software analysis and design has become widely used in the industry over the past decade. A widely used OOSE process is the *Rational Unified Process* (RUP) (Kruchten, 1999) derived from UML and the associated *Unified Software Development Process* proposed by Jacobson, Booch, and Rumbaugh (1999). The core software engineering disciplines of RUP are domain modeling, requirements engineering, design, implementation, testing, and deployment.

In parallel with RUP, agile methodologies, such as *Extreme Programming* (Beck, 1999), have emerged. They emphasize lightweight processes such as test-case-based development and rapid prototyping. They de-emphasize detailed modeling on which they blame the heavy weight and inflexibility of traditional methodologies. In contrast, the *Model-Driven Architecture* (MDA, http://www.omg.org/mda) approach of the OMG identifies modeling as the key to state-of-the-art software engineering. In the MDA, computation-independent domain models are transformed into platform-independent design models that are then turned into platform-specific models and implementations.

NEW SOFTWARE ENGINEERING PARADIGM

Dynamic adaptive systems in open environments create new challenges for software engineering. For such systems, current software engineering techniques cannot guarantee quality attributes such as correctness to hold after deployment. Correctness is traditionally assured by testing the system before release, against documented requirements. Such assurance is lost if system behavior changes due to continuous adaptation or environment change. For example, a sociotechnical business system may change its structure (architecture) to mirror changes in the human organization.

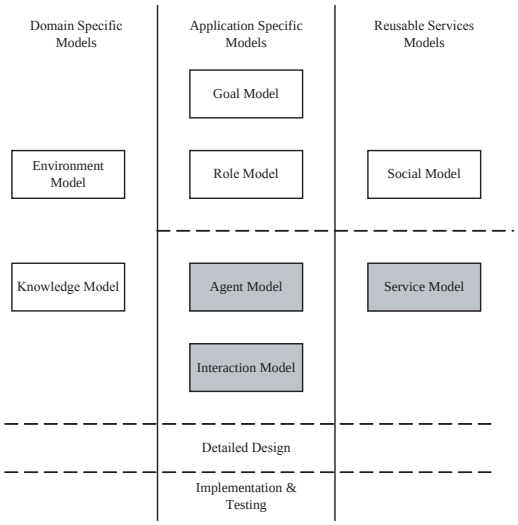
Similarly, system performance, reliability, security, usability, and maintainability can be compromised due to adaptation or environmental changes. Without explicit representation of system requirements and constant validation at run time, there is no guarantee that the system functions correctly.

To address the challenges described, a sociotechnical system should be analyzed and designed in terms of agents (actors), roles, and goals at the stage of requirements engineering, as well as at design, implementation, testing, and maintenance stages of the software lifecycle. From the start of a software engineering process, a distinction should be introduced between active and passive entities, that is, between agents and (nonagentive) objects of the real world. We define an agent as an autonomous entity situated in an environment capable of both perceiving the environment and acting on it. The agent metaphor subsumes *artificial* (software and robotic), *natural* (human and animal) as well as *social/institutional* agents (groups and organizations). According to Wagner (2003), the agent metaphor lies beyond UML where actors are only considered as users of the system's services in *use cases*, but otherwise remain external to the sociotechnical system model.

We define a *role* as a coherent set of functional responsibilities specifying what the agent playing the role is expected to do in the organization within some specialized context or domain of endeavor: with respect to both other agents and the organization itself. Roles may be subject to constraints.

Roles go hand in hand with goals. For example, in a business domain, a human or institutional agent acting in the role of "customer" has a goal of accomplishing something. To achieve its goal, the agent uses some service provided by another agent. An agent's autonomy means that the service provider performs the service requested if it is able to do so but the service provider also has an option to refuse the service request. Even though the agent requesting the service may not explicitly communicate its goals to the service provider agent, the latter always "internalizes" the whole or part of the customer's goal in attempting to provide the service. For example, given a customer wanting to rent a car, the goal of a car rental company is to provide the customer

Figure 2. Overview of the ROADMAP methodology



with a car, which is, of course, a subgoal of the company's higher-level goal—to earn money through renting cars. The car rental company tries to achieve this higher-level goal by "internalizing" as many customer goals as possible.

A goal can thus be defined as a condition or state of affairs in the world that the agent wants to bring about. Some people, like van Lamsweerde (2003), have chosen to model goals formally but this is not always necessary or even possible. For example, how would one formally model a goal of being greeted by someone?

An agent realizes goals by performing activities. Each activity belongs to some *activity type* defined as a prototypical job function in an organization that specifies a particular "way of doing" by performing elementary epistemic, physical, and/or communicative actions. An *action* is an atomic and instantaneous unit of work done by an agent. We thus view an agent's action more broadly as something that the agent does. In OOSE an action is often understood narrowly as something changing the state of a data object.

An action performed by one agent can be perceived as an event by another agent. Wagner (2003) distinguishes between an *action event type* (an event created through the action of an agent, such as starting a machine) and a *nonaction event type* (for example, temporal events or events created by natural forces). Action event types are further divided into *communicative action event* (or *message*) *types* and *noncommunicative* (*physical*) *action event types* like providing another agent with a commodity. An agent can also perform *epistemic actions* that change its knowledge state.

The popular belief-desire-intention (BDI) agent model by Rao and Georgeff (1995) includes the notions of belief and plan. Padgham and Winikoff (2004) define a *belief* as some aspect of the agent's knowledge about the environment, itself, or other agents. They define a *plan* as a way of realizing a goal.

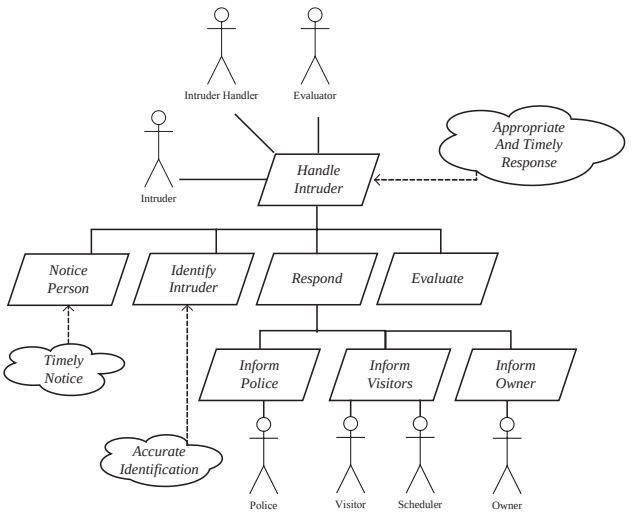
An AOSE methodology consists of the stages of requirements engineering, design, implementation and testing, and maintenance. Figure 2 shows the models employed at different stages of the ROADMAP methodology (Juan & Sterling, 2003). In ROADMAP, models are split horizontally by dotted horizontal lines according to the requirements engineering and design stages so that the environment model, knowledge model, goal model, role model, and social model are parts of the requirements engineering stage, while the agent model, interaction model, and service model belong to the architectural design substage. The models are also divided vertically into domain specific models, application specific models, and reusable services models. The environment model and knowledge model represent information about a specific domain and belong to multiple phases in the software development lifecycle. The goal model, role model, agent model, and interaction model are tied to the system being modeled. Generic and reusable components in the system are captured by the social and service models.

During requirements engineering, goals and roles are defined pertaining to the intended sociotechnical system. The goal model provides a high-level overview of the system requirements. Its objective is to enable both domain experts and developers to pinpoint the goals of the system and thus the roles the system needs in order to meet those goals. Quality goals (nonfunctional goals) may be attached to regular goals to constrain or document how goals should be fulfilled. Quality goals reflect more intangible goals of a system, such as privacy and timeliness.

Roles are attached to goals, indicating responsibility for achieving a particular goal. Role models, in turn, detail what responsibilities a specific role carries, as well as constraints that must be heeded. The concepts of roles and agents facilitate understanding of and elicitation of requirements because they parallel humans taking on roles within a particular organizational unit with (full or partial) responsibility for a set of goals within that unit.

Figure 3 gives the goal model of a scenario for handling intruders in an intelligent home. The role Intruder Handler has a single goal, to handle an intruder. This goal is characterized by the *quality goal* to provide an appropriate and timely response to a possible intruder detected. The Handle Intruder goal is decomposed into four subgoals: Notice Person, Identify Intruder, Respond, and Evaluate. There are quality goals Timely Notice and Accurate Identification pertaining to the subgoals Notice Person and Identify Intruder, respectively. The subgoal Respond, in turn, is divided into subgoals Inform Police, Inform Visitors, and Inform Owner. To accomplish these, the additional roles Police, Visitor, Scheduler, and Owner are introduced.

Figure 3. The goal model of intruder handling

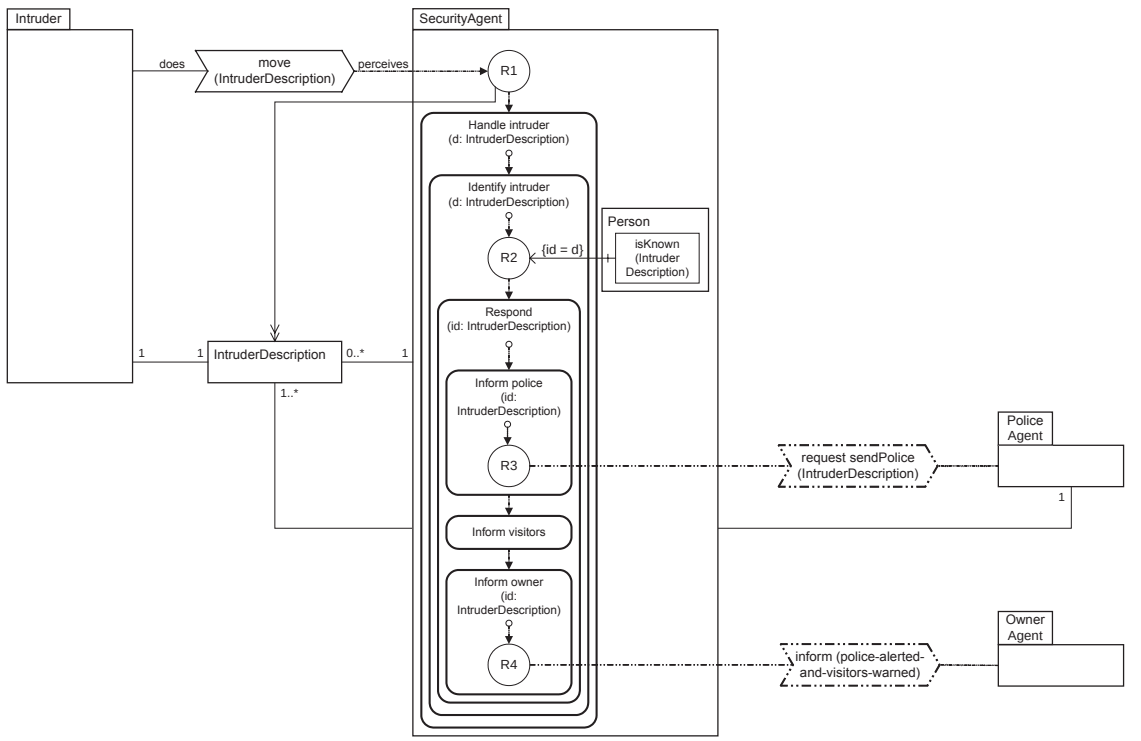


The environment and knowledge models constitute an *ontology* providing a common framework of knowledge for the agents of the problem domain. The social model captures relations between roles in the system and policies for interaction encompassing such areas as security, privacy, and communication.

During design, the types of software agents required, as well as the types of interactions and services provided by agents are defined, based on the models created during requirements engineering. These descriptions are refined during detailed design. An AOSE process ends with generating or coding the software agents or with other implementations such as object-oriented ones.

Various agent-oriented methodologies and the associated agent architectures can be used for design and implementation, for example, Prometheus (Padgham & Winikoff, 2004). To facilitate fast prototyping, we use the RAP/AOR methodology by Taveter and Wagner (2005). RAP/AOR is an agile AOSE methodology based on both MDA and RUP. RAP/AOR proposes an agent architecture with behavior modeling constructs called reaction rules. *Reaction rules* (also known as *event-condition-action* rules) govern the execution and sequencing of physical, communicative, and epistemic actions in response to perception and communication events. Reaction rules also start activities. Figure 4 represents an Agent-Object-Relationship (AOR) diagram comprising the agent, interaction, and service models for the intruder handling scenario. Reaction rule R1, triggered by an action event of type move(IntruderDescription), starts the scenario's outermost activity. This action event is created by a human agent Intruder through a sensor and perceived by the software agent SecurityAgent. Note that activity types modeled in the AOR diagram in Figure 4 correspond to goals represented in Figure 3. For example, an activity of type "Respond" achieves a Goal to respond to the detection

Figure 4. A design model of intruder handling



of an intruder. The subactivity type "Inform visitors," which involves the agent type SchedulerAgent, is not refined in Figure 4. Reaction rule R2 represents checking the Boolean predicate isKnown, attached to the object type Person. If the predicate evaluates to false, that is, if the person described by the IntruderDescription is not recognized, an activity of type "Respond" is started. This activity involves informing the PoliceAgent and the OwnerAgent embedded in a handheld device.

Taveter (2004) demonstrates that AOR diagrams can be straightforwardly transformed into programming constructs of the Java Agent Development Environment (JADE, http://jade.cselt.it/) software agent platform (Bellifemine, Poggi, & Rimassa, 2001).

FUTURE TRENDS

As decision support systems become more intelligent, new and imprecise quality attributes, such as privacy and politeness, emerge. The exact meanings of these quality attributes depend closely on actual system users and context of use. Whether quality attributes are fulfilled is open to user interpretation and perception. For example, we may expect our personal assistant agents to know our daily routines, habits, and preferences. Yet, we also expect user privacy and do not wish such knowledge to become public. The meaning of privacy and the level of privacy needed by each user is subjective and usually different. For many intelligent systems, the sheer complexity of their tasks renders it difficult to fully define and test the systems for correctness. To address this issue, we suggest that the AOSE paradigm must make available constructs to define such quality attributes in a flexible manner so that the definitions can be easily customized for each user at run time, as has been started within ROADMAP (Juan & Sterling, 2003).

CONCLUSION

In this article, we described the new paradigm of AOSE. The article thus presented an overview of some of the latest advances in the area of software engineering which are due to the emergence of novel analysis notions, the most fundamental ones of them being agent (actor), role, goal, and activity, as well as new implementation units—software agents. While agent-oriented analysis notions can always be used, software agents are recommended when the problem domain is distributed and the environment is open and unpredictable.

REFERENCES

Beck, K. (1999). *Extreme programming explained: Embrace change*. Indianapolis, IN: Addison-Wesley Professional.

Bellifemine, F., Poggi, A., & Rimassa, G. (2001). Developing multi-agent systems with a FIPA-compliant agent framework. *Software—Practice and Experience, 31*(2001), 103-128.

Bresciani, P., Perini, A., Giorgini, P., Giunchiglia, F., & Mylopoulos, J. (2004). Tropos: An agent-oriented software development methodology. *Autonomous Agents and Multi-Agent Systems, 8*(3), 203-236.

Ciancarini, P., & Wooldridge, M. (2001). Agent-based software engineering—Guest editors' introduction. *International Journal of Software Engineering and Knowledge Engineering, 11*(3), 205-206.

Garijo, F. J., Gomez-Sanz, J. J., & Massonet, P. (2005). The MESSAGE methodology for agent-oriented analysis and design. In B. Henderson-Sellers & P. Giorgini (Eds.) *Agent-oriented methodologies* (pp. 203-235). Hershey, PA: Idea Group Publishing.

Henderson-Sellers, B., & Giorgini, P. (2005). *Agent-oriented methodologies*. Hershey, PA: Idea Group Publishing.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *Unified software development process.* Reading, MA: Addison-Wesley.

Juan, T., & Sterling, L. (2003, July 15). The ROADMAP meta-model for intelligent adaptive multi-agent systems in open environments. In P. Giorgini, J. P. Muller, & J. Odell (Eds.), *Agent-Oriented Software Engineering IV, 4th International Workshop, AOSE 2003,* Melbourne, Australia, *Revised Papers* (LNCS, Vol. 2935, pp. 826-837). Berlin: Springer-Verlag.

Kruchten, P. (1999). *Rational unified process—An introduction*. Reading, MA: Addison-Wesley.

Odell, J., Parunak, H. Van Dyke, & Bauer, B. (2001). Representing agent interaction protocols in UML. In P. Ciancarini & M. Wooldridge (Eds.), *Agent-Oriented Software Engineering, First International Workshop, AOSE 2000,* Limerick, Ireland, June 10, 2000, *Revised Papers* (LNCS, Vol. 1957, pp. 121-140). Berlin: Springer-Verlag.

OMG. (2003a, March). *Unified Modeling Language specification.* (Version 1.5). Retrieved October 11, 2005, from http://www.omg.org/cgi-bin/doc?formal/03-03-01

OMG. (2003b, August). *Unified Modeling Language: Superstructure.* (Version 2.0). Retrieved October 11, 2005, from http://www.omg.org/cgi-bin/doc?ptc/2003-08-02

Padgham, L., & Winikoff, M. (2004). *Developing intelligent agent systems: A practical guide*. Chichester, UK; Hoboken, NJ: John Wiley & Sons.

Pressman, R. S. (2001). *Software engineering: A practitioner's approach* (5th ed.). New York: McGraw-Hill.

Rao, A. S., & Georgeff, M. P. (1995). BDI agents: From theory to practice. In V. R. Lesser & L. Gasser (Eds.), *Proceedings of the First International Conference on Multiagent Systems* (pp. 312-319). Cambridge, MA: MIT Press.

Silva, V., & Lucena, C. (2004). From a conceptual framework for agents and objects to a multi-agent system modeling language. *Autonomous Agents and Multi-Agent Systems, 9*(1-2), 145-189.

Sommerville, I. (2004). *Software engineering* (7th ed.). Reading, MA: Addison-Wesley.

Taveter, K. (2004). *A multi-perspective methodology for agent-oriented business modelling and simulation*. PhD thesis, Tallinn University of Technology, Estonia. Tallinn, Estonia: TUT Press.(ISBN 9985-59-439-8).

Taveter, K., & Wagner, G. (2005). Towards radical agent-oriented software engineering processes based on AOR modelling. In B. Henderson-Sellers & P. Giorgini (Eds.), *Agent-oriented methodologies* (pp. 277-316). Hershey, PA: Idea Group Publishing.

van Lamsweerde, A. (2003). From system goals to software architecture. In M. Bernardo & P. Inverardi (Eds.), *Formal methods for software architectures* (LNCS, Vol. 2804, pp. 25-43). Berlin: Springer-Verlag.

Wagner, G. (2003). The agent-object-relationship metamodel: Towards a unified view of state and behavior. *Information Systems, 28*(5), 475-504.

Wooldridge, M. (2002). *Introduction to multi-agent systems*. Reading, MA: Addison-Wesley.

Wooldridge, M., Jennings, N. R., & Kinny, D. (2000). The Gaia methodology for agent-oriented analysis and design. *Autonomous Agents and Multi-Agent Systems, 3*(3), 285-312.

KEY TERMS

Action: An atomic and instantaneous unit of work done by an agent.

Action Event Type: The type of an event created through the action of an agent, such as starting a machine, sending a message to another agent, or providing another agent with a commodity.

Activity Type: A prototypical job function in an organization that specifies a particular "way of doing" by performing elementary epistemic, physical, and/or communicative actions.

Agent: An autonomous entity situated in an environment that is capable of perceiving the environment and acting on the environment.

Agent-Oriented Software Engineering: A new paradigm within software engineering involving novel analysis notions, the most fundamental ones of them being agent (actor), role, goal, and activity, as well as new implementation units—software agents.

Goal: A condition or state of affairs in the world that the agent wants to bring about.

Multi-agent System (MAS): A new paradigm in distributed computing where the system characteristics are: (1) no global system control; (2) decentralized information; and (3) each participating agent has incomplete information and limited capabilities.

Quality Goal: A goal which constrains or documents how a regular goal should be achieved by an agent.

Reaction Rule: A rule that specifies behavior of an agent by determining a triggering event to which the agent must react, a condition that the agent must check in its internal knowledge base and one or more actions to be performed by the agent.

Role: A coherent set of functional responsibilities specifying what the agent playing the role is expected to do in the organization within some specialized context or domain of endeavor: with respect to both other agents and the organization itself.

Software Engineering: A discipline applied by teams to produce high-quality, large-scale, cost-effective software that satisfies the users' needs and can be maintained over time.

Software Engineering Process: A set of activities whose purpose is development or evolution of a software system. Generally consists of the stages of requirements engineering, design, implementation, testing, and maintenance.

Agents and Payment Systems in E-Commerce

A

Sheng-Uei Guan
National University of Singapore, Singapore

INTRODUCTION

An emerging outcome of the popularization of the Internet are electronic commerce and payment systems, which present great opportunities for businesses, reduce transaction costs, and provide faster transaction times. More research has been conducted with new technologies like mobile Internet used by business models (Baek & Hong, 2003). However, before using the Internet, it is essential to provide security in transferring monetary value over the Internet. A number of protocols have been proposed for these secure payment systems, including NetBill, NetCheque, Open Market, iKP, Millicent, SET (Sherift, 1998), E-Cash (Brands, 1995), NetCash, CAFÉ (Mjolsnes, 1997), EMV cards (Khu-Smith & Mitchell, 2002), etc. These systems are designed to meet diverse requirements, each with particular attributes.

Automation and intelligence is another issue that poses challenges in the development of e-commerce. Agent technology has been incorporated into the area of e-commerce to provide automation and intelligence for the e-trade process. An agent is a software program capable of accomplishing tasks autonomously on behalf of its user. Agents must provide trustworthy consistency and fault tolerance to avoid eavesdropping and fraud. Also, agents should have roaming capability so as to extend their capability well beyond the limitations of owners' computers. To meet these requirements, this chapter will discuss some related components under the SAFER (Secure Agent Fabrication, Evolution, and Roaming) architecture (Zhu & Guan, 2000) and propose an agent-based payment scheme for SAFER.

Different types of electronic payment systems have been developed to meet its diverse requirements, which generally include integrity, authorization, confidentiality, availability, and reliability for security requirements (Asokan, 1997). Payment systems can be classified in a variety of ways according to their characteristics (Dahab & Ferreira, 1998), such as the exchange model (cash-like, check-like, or hybrid), central authority contact (online or offline), hardware requirements (specific or general), payment amounts (micropayment), etc.

Among the available payment schemes in the market, E-Cash is one of the best in terms of security, flexibility, and full anonymity. E-Cash is a cash-like online system that uses electronic coins as tokens. E-Cash has unique advantages, such as flexibility, integrity, and full anonymity that cannot be found in electronic check and credit card based systems. It uses cryptographic techniques to provide full anonymity. The agent-based payment scheme for SAFER adopts some similar principles and concepts of E-Cash.

MAIN THRUST OF THE ARTICLE

This chapter presents a brief overview of agents and payment system attributes used in e-commerce. An agent-based e-payment scheme built for the SAFER e-commerce architecture is proposed, which is aimed at providing a flexible and secure financial infrastructure for Internet commerce.

Software Agents in Electronic Commerce

Attributes of Agent-Based Systems for Electronic Commerce

Agents are bits of software performing routine tasks, typically in the background, on behalf of the user. Gathering, filtering, and presenting information are some of the small and well-defined tasks given to simple agents. An agent distinguishes itself from any other software by its intelligence. Intelligent agents are capable of "thinking" and producing intelligent feedback (Guan & Yang, 1999). Agents are increasing in the degree and sophistication of automation, on both the buyer's and seller's sides, commerce becomes much more dynamic, personalized, and context sensitive. These changes can be beneficial to both buyers and sellers (He, Jennings, & Leung, 2003).

The requirement for continuity and autonomy derives from our desire that an agent be able to carry out activities in a manner that is responsive to changes in the environment without requiring constant human guidance or intervention. According to Bradshaw (1997), agents have the following attributes, as shown in Table 1.

There are several software agent prototypes under development that will be capable of doing even more on behalf of buyers and sellers. One is Kasbah, wherein agents would proactively seek potential sellers and negotiate with them on the buyer's behalf, making the best possible deal, based on a set of constraints specified by the buyer, including the highest acceptable price and a transaction completion date (Chavz, 1996). A disadvantage of this software agent is that it always

Table 1. Attributes of software agents

Attribute	Description
Reactivity	The ability to selectively sense an act
Autonomy	Goal-directness, proactive and self-starting behavior
Collaborative behavior	Can work in concert with other agents to achieve a common goal
Communication ability	The ability to communicate with persons and other agents
Personality	The capability of manifesting the attributes of a believable character, such as emotion
Temporal continuity	Persistence of identity and state over long periods of time
Adaptivity	Being able to learn and improve with experience
Mobility	Being able to migrate in a self-directed way from one host platform to another

accepts the first offer that can meet its asking price, when even better offers might exist. This disadvantage is resolved by AuctionBot, a general-purpose Internet auction server. *AGENTics* is another agent prototype that develops what is referred to as "*online catalog integration for e-commerce*." AGENTics products shield the user from the technicalities of "where" and "how" the information was gathered, while it synthesizes many information pieces into a coherent whole (Mougayar, 1997).

Some agents can select desired items based on preferences, search databases to look for selected pieces of information, and conduct transactions. An example of such an adaptive agent is the SAFER architecture for e-commerce.

SAFER (Secure Agent Fabrication, Evolution, and Roaming) is a Web-based distributed infrastructure to serve agents to query, buy, and sell goods in e-commerce. It establishes necessary mechanisms to transport, manufacture, and evolve all different types of agents. The goal of SAFER is to construct open, dynamic, and evolutionary agent systems for e-commerce (Zhu & Guan, 2000). There will be SAFER-compliant and noncompliant communities coexisting in the e-commerce network. Each SAFER community

Figure 1. Cooperating agents for the SAFER payment scheme

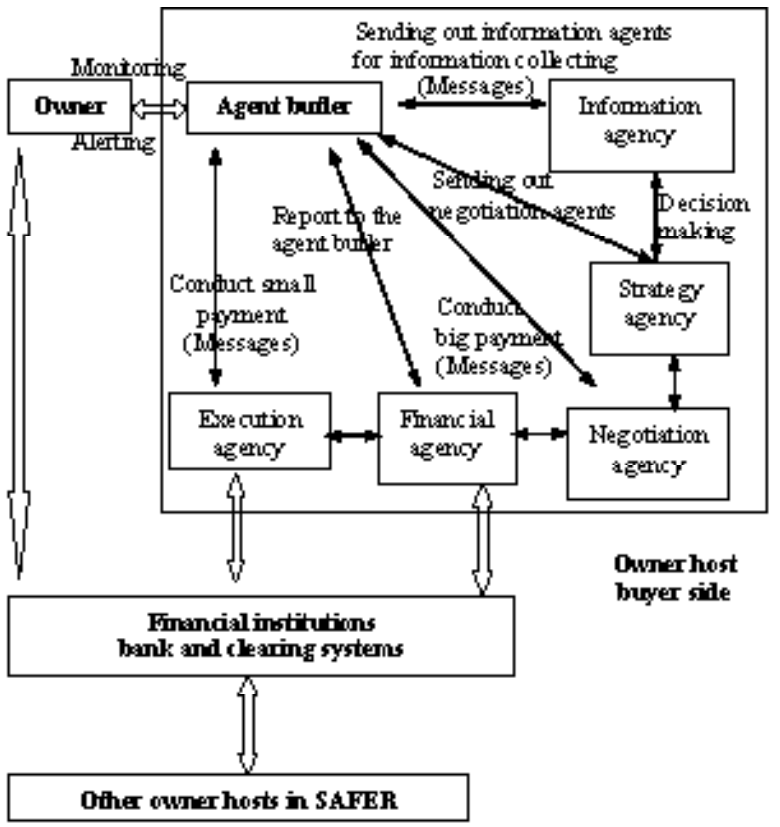
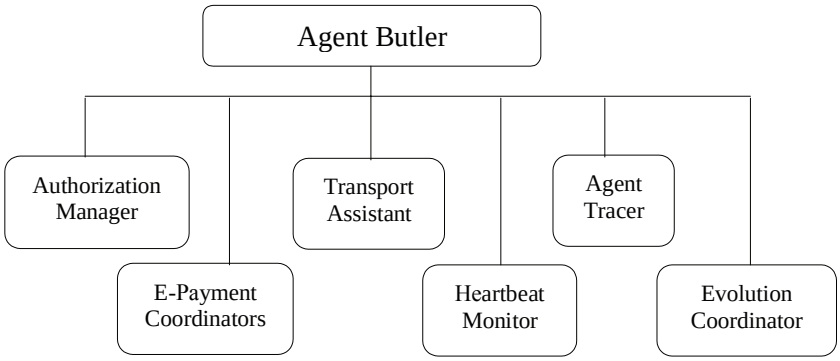


Figure 2. Prototype of agent butler



consists of several mandatory components: owner, butler, agent, agent factory, community administration center, agent charger, agent immigration, clearinghouse, and bank. Agent community is the basic unit in SAFER e-commerce, which offers virtual regions and vehicles to host and administrate mobile agents during roaming, transaction, and evolution. An owner is in charge of all his agents and of making respective authorizations to mobile agents and his agent butler, which is a 24-hour online watcher that would handle most of the tasks on behalf of the owner. When agents are sent roaming in the network, the butler has the responsibility of keeping track of agents' activities and locations by sending messages to agents and receiving messages from agents. At least one financial institution, usually a bank, that can link all value representations to real money, must also be involved. The payment scheme designed for SAFER is expected to fulfill flexibility and interoperability, which means that diverse representations of value will have the possibility of emerging in one framework for users' convenience. Given that, it is important that funds represented by one mechanism be easily converted into funds represented by others (Neuman, 1995).

An Agent-Based E-Payment Scheme for SAFER

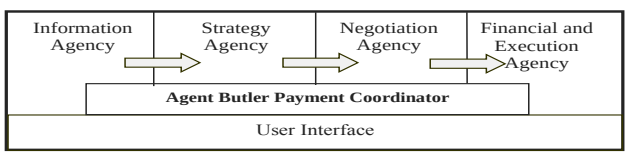
The payment module in the agent-mediated SAFER e-commerce architecture must contain several essential components: the marketplace, agents (including mobile agents, static agents, and agent butlers), financial institutions, and users. In SAFER, a community will offer virtual regions, factories, administration tools, vehicles to manipulate and administrate mobile agents during any activity, and provide security so that users can trust it.

Different types of agents fabricated by an agent factory of SAFER are running under the payment scheme for respective functions and tasks. They are briefly described in Figure 1.

In this scheme, a subsystem called agency is mentioned. Similar to the definition given by Dr. Larry Kerschberg in his Defense Personnel Support Center (DPSC) project, an agency can be thought of as a multilayered agent group or a federation of agents with specific goals and functional roles in the architecture. It is also like a collection of cooperating intelligent agents with particular expertise (Kerschberg, 1997).

If the owner is interested in some items, he will assign tasks to his butler and agents. The agent butler will then send out information agents from the agency, taking note of the items of interest, and set parameters such as due date (by which the item should be purchased), desired price, and highest acceptable price. The information agents used to sift, filter, and process information will roam in SAFER or even non-SAFER communities under a certain transport protocol, which is explained in the literature (Guan & Yang, 1999). It can help with dozens of purchasing decisions, thus lowering the cost and gaining efficiency. While roaming, agents are well tracked by the agent butler, by sending messages to report their activities and locations, which is described in detail in Zhu and Guan (2000). After gathering enough information, the information agent forwards all to

Figure 3. Payment coordinator



the strategy agency, which will analyze the new data and settle on a decision for the user. All the recommendations will be reported to the agent butler first. Once a recommendation has been reported, the agent butler activates the negotiation agency that will send out negotiation agents to the shortlist merchant hosts. Negotiation is defined as follows in Green (1997), "negotiation is the communication process of a group of agents in order to reach a mutually accepted agreement on some matter" (21). If the negotiation agent and the receptionist agent reach an agreement, the result will be reported to the butler. The butler will inform the financial agency to initiate the contract for certain goods and make a transaction decision according to the amount of money involved, the distance from host to the destination vendor, etc. Financial agents will take charge of the goods reception and payment transaction under the authorization of the butler. They communicate with the merchant host, autonomously make a payment request, and sign a contract order against the right good.

Implementation

The implementation of SAFER is under way. The overall architecture consists of several closely related but separate modules: roaming, evolution, fabrication, negotiation, and electronic payment.

The implementation of the payment module began with the development of the agent butler, which is defined as a combination of several separate functions, as shown in Figure 2. They are authorization manager, e-payment coordinator, transport assistant, heartbeat monitor, agent tracer, and evolution coordinator.

In the e-payment coordinator module, communication channels are set up between agent butler and all agencies of diverse functionalities, each of which is running in a separate thread. User interfaces are designed so that the user can assign tasks, define needs and requirements, check records, and read alerting messages reported by his or her agent butler.

Making all types of agents and merchant hosts available to fit in the same framework will be difficult in the current research stage, because the attributes that agents require to communicate may differ. Given that, we have chosen to implement a limited number of typical agents to test the system functionality and will consider how the work could be generalized to e-commerce in the future.

FUTURE TRENDS

The foremost important feature of e-commerce is transaction security. If the system is not trustworthy, there will be no incentive for participants to cooperate. A prerequisite is the prevention of double spending of electronic cash or coins. Ensuring this is the crux of any system, and it will incur significant overhead. Electronic currency in the context of mobile-agent-based systems has one particular caveat: the credits that an agent may carry are essentially just data, to which the host potentially has access. To ensure system reliability and reduce overhead, there are still many open issues for further consideration and discussion in future work:

- Privacy issues in collecting data and negotiation, e.g., to prevent agents from divulging private information to other hosts and to protect agents from malicious hosts
- The extent that users will let agents make decisions for them (based on their preference)
- The agent's ability of negotiation (This should be protected to prevent a host from having access to the agent's negotiation function and then affect the agent's buying power.)
- Agent traceability (An agent butler keeps in contact with his mobile agents periodically.)
- Fault tolerance and credits recovery in case of sudden crash of remote systems or unexpected attack on mobile agents
- Protection of agent's public and secret key during agent roaming

CONCLUSION

The agent-based SAFER e-commerce payment scheme incorporated agent technologies and took advantage of some well-known secure payment transaction protocols. It aims to simulate and even enhance physical cash and is designed to support a multitude of currency types. By incorporating the concepts of agent, the system is expected to provide security, efficiency, flexibility, autonomy, and intelligence. It is designed to provide anonymity against other parties and audit ability (traceability) for the owner (or agent butler). At last, a number of potential improvements, practical aspects, and some open issues have been identified for future work.

REFERENCES

Asokan, N., & Janson, P. A. (1997). *The state of the art in electronic payment systems. Computer, 30*(9), 28–35.

Baek, J. M., & Hong, I.-S. (2003). *A study on mobile payment system with United Mileage using USIM* (pp. 394–403). HSI 2003, LNCS 2713.

Bradshaw, J. (1997). *Software agent.* Cambridge, MA: AAAI Press/The MIT Press.

Brands, S. (1995). *Electronic cash on the Internet.* In *Proceedings of the Symposium on Network and Distributed System Security, San Diego, California, 1995* (pp. 64–84).

Chavz, A., & Maes, P. (1996). MIT Media Lab. *Kashbah: An Agent Marketplace for Buying and Selling Goods.* Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi-Agent Technology (Crabtree, B. and Jennings, N., Eds.), 75-90. The Practical Application Company Ltd, Blackpool.

Dahab, R., & Ferreira, L. C. (1998). *A scheme for analyzing electronic payment systems.* 14th Annual Computer Security Applications Conference, Phoenix, AZ.

Green, S. (1997). *Software agents*. A review. IAG Technical Report, Trinity College Available at *http://www.cs.tcd.ie/research_groups/aig/iag/toplevel2.html*

Guan, S. U., & Yang, Y. (1999). *SAFE: Secure-roaming agent for e-commerce*, 26th International Conference on Computers and Industrial Engineering, Australia, Vol 42 Issue 2-4, 481-493.

He, M., Jennings, N. R., & Leung, H. F. (2003). *On agent-mediated electronic commerce. IEEE Transactions on Knowledge and Data Engineering, 15*(4), 985-1003.

Kerschberg, L., & Banerjee, S. (1997). *The DPSC electronic marketspace: The impact of intelligent agents, the Internet and the Web on electronic commerce and logistics.* Available at *http://cise.krl.gmu.edu/KRG/DPSCAgentHTML_folder/DPSCAgent.html*

Khu-Smith, V., & Mitchell, C. J. (2002). *Using EMV-cards to protect e-commerce transactions* (pp. 388–399). EC-Web 2002, LNCS 2455.

Lucas, F. & Dahab, R. (1998). A scheme for analyzing electronic payment systems. In 14th ACSAC-Annual Computer Security Applications Conference (ACSAC '98), Scottsdale, Arizona.

Mjolsnes, S. F., & Michelsen, R. (1997). *CAFÉ. Open transnational system for digital currency payment.* In Proceedings of the 30th Hawaii International Conference on System Sciences, Advanced Technology Track, IEEE Computer Society, Washington, D.C. (Vol. 5, pp. 198–207).

Mougayar, W. (1997). *The future of agent-based commerce on the Web.* CYBERManagement Inc. Retrieved from *http://www.cyberm.com/cyber/art2.htm*

Neuman, B. C., & Medvinsky, G. (1995). *Requirements for network payment: The NetCheque™ perspective*. Proceedings of IEEE Compcon'95, San Francisco, IEEE Computer Society, Washington, DC, 32.

Sherift, M. H., & Serhrouchni, A. (1998). SET and SSL: Electronic payments on the Internet. In *Proceedings of the Third IEEE Symposium on Computers and Communications, 1998. ISCC'98* (pp. 353–358), Athens, Greece.

Zhu, F. M., Guan, S. U., Yang, Y., & Ko, C. C. (2000). SAFER e-commerce: Secure agent fabrication, evolution, and roaming for e-commerce. In M. R. Syed & R. J. Bignall (Eds.), Internet *Commerce and Software Agents: Cases, Technologies and Opportunities*. Hershey, PA: Idea Group Publishing, 190-207. *http://www.webopedia.com/TERM/c/cryptography.html*

KEY TERMS

Adaptability: The ease with which software satisfies differing system constraints and user needs (Evans, 1987).

Agents: A piece of software that acts to accomplish tasks on behalf of its user.

Anonymity: The degree to which a software system or component allows for or supports anonymous transactions.

Confidentiality: The nonoccurrence of the unauthorized disclosure of information (Barbacci, 1995).

Cryptography: The art of protecting information by transforming it (encrypting it) into an unreadable format, called cipher text. Only those who possess a secret key can decipher (or decrypt) the message into plain text.

Flexibility: The ease with which a system or component can be modified for use in applications or environments other than those for which it was specifically designed (IEEE, 1990).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 93-97, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Agile Information Technology Infrastructures

Nancy Alexopoulou
University of Athens, Greece

Panagiotis Kanellis
National and Kapodistrian University of Athens, Greece

Drakoulis Martakos
National and Kapodistrian University of Athens, Greece

INTRODUCTION

Operating in highly turbulent environments, organizations today are faced with the need to continually adjust their infrastructure and strategies in order to remain competitive. Globalization and continual technological evolution are the main drivers of this turbulence (Dove, 1999b). To adapt at the same pace as their changing environment, organizations have to be agile. Loosely defined, an agile enterprise is one that is characterized by change proficiency. Change proficiency is the defining characteristic of agility and denotes the competency in which an adaptive transformation occurs (Dove, Benson, & Hartman, 1996). In a more detailed definition, an agile enterprise is one that is characterized as a fast moving, adaptable, and robust business, which is capable of rapid adaptation in response to unexpected and unpredicted changes and events, market opportunities, and customer requirements (Henbury, 1996).

According to Dove (1999b), agility is very much related to the ability to manage and apply knowledge effectively. Dove (1999b) felicitously associates agility with cats. A cat is both physically adept at movement and also mentally adept at choosing useful movement appropriate for the situation. If a cat has merely the ability to move quickly but moves inappropriately and to no gain (e.g., a cat on a hot tin roof), it might be called spastic or confused but never agile. On the other hand, a cat that knows what should be done but finds itself unable to move (e.g., a cat that's got itself up a tree), might be called catatonic, confused, or paralyzed but never agile.

This example implies that agility cannot be easily attained. It requires knowledge, experience, and skill. Enterprise agility depends on many factors such as personnel capabilities, information technology (IT) infrastructure, business strategy, and so forth. When an enterprise is agile, all its constituents are agile and vice versa. This article focuses particularly on IT infrastructure. It defines agility in IT infrastructure and explains how it contributes to enterprise sensing and response agility. *Sensing agility* is defined as a firm's ability to rapidly discover and interpret the market opportunities through its information systems, and it concerns not only an ability to distinguish information from noise quickly, but also to transform apparent noise into meaning faster (Haeckel, 1999). *Response agility* relates to the organizational capability to quickly transform knowledge into action in response to the environmental signals (Haeckel, 1999).

BACKGROUND

The term *agility* has over a decade of use in manufacturing practices, where it has been defined as a principle competitive issue (Kidd, 1994; Dove, 1994a; Goldman, Roger, & Kenneth, 1995). Dove (1999a, 2005) has introduced the principles for agile systems at an abstract level so that they can be interpreted either from a business or a technical perspective. The term *system* is used to characterize a group of interacting modules sharing a common framework and serving a common purpose. At the business level, modules represent groups of people while at the technical level correspond to software components or machines. These principles are summarized in Table 1.

Dove (1995) has also defined four agility metrics, namely *time, cost, robustness,* and *scope*. The first concerns the time required to complete a transformation. The second defines the cost regarding the transformation implementation. Robustness measures the strength and quality of the change process. Scope indicates how much latitude for change can be accommodated. Kidd (1994) has additionally defined a fifth agility metric which is the *frequency of change*.

The concept of agility has also been employed in the research area of information systems (IS) development where the term is much more recent (Aydin, Harmsen, Slooten, & Stegwee, 2004; Levine, 2005). Agile IS development concerns a new methodology paradigm proposed as an alternative to traditional disciplined methodologies for software development because these methodologies are no longer successful for rapidly changing environments due to their bureaucratic nature (Conboy & Fitzgerald, 2004; Nerur, Mahapatra, & Mangalaraj, 2005). Agile development

Table 1. Agile design principles (Source: Dove, 1999a)

Encapsulated Unit Modularity	System of interacting unit not intimately integrated. Internal workings unknown externally.
Plug Compatibility	System units share common interaction and interface standards, and are easily inserted or removed.
Facilitated Unit Reusability	Standardized unit replication information, unit modification tools, unit capability catalogs.
Non-Hierarchical Interaction	Empowered self-directed units that communicate negotiate and interact directly among themselves.
Dynamic Late Binding Relationships	Relationships are transient when possible; fixed binding is postponed until immediately necessary.
Distributed Control & Information	Units respond to objectives; decisions made at point of knowledge; data retained locally but accessible globally.
Self-Organizing Relationships	Dynamic unit alliances and scheduling; open bidding; and other self-adapting behaviors.
Scalable Size	Unrestricted unit populations that permit large increases and decreases in total unit population.
Unit Redundancy	Duplicate unit types or capabilities to provide capacity fluctuation options and fault tolerance.
Extensible Framework	Evolving open system framework capable of accommodating legacy, common, or completely new units.

methodologies (Abrahamsson, Salo, Ronkainen, & Warsta, 2002), such as Extreme Programming (Beck, 1999) and SCRUM (Schwaber & Beedle, 2002), promise faster development times and higher customer satisfaction. Extreme Programming and SCRUM constitute instantiations of the Agile Manifesto (Fowler & Highsmith, 2001), which was published by the Agile Alliance in 2001 (www.agilealliance.com). The basic principles of the Agile Manifesto are: first, individuals and interactions over processes and tools; second, working software over comprehensive documentation; third, customer collaboration over contract negotiation; and fourth, responding to change over following a plan (Williams & Cockburn, 2003).

In IT and IS literature, the term agility has not been broadly used. A relevant concept for IT infrastructure that has been defined instead is that of *flexibility*. *IT Infrastructure flexibility* is defined as the ability of the IS department to respond quickly and cost-effectively to system demands, which evolve with changes in business practices or strategies (Duncan & Bogucki, 1995). In this definition, however, the

ability to respond to unexpected or unpredicted change is not explicitly stated. As a matter of fact, agility and flexibility are not synonyms. Their main difference is that flexibility is a planned response to variations which have been planned, whereas agility concerns minimizing the inhibitions to change in any direction based on unanticipated change (Dove, 1995). In this respect and as deduced from the previous discussion, agility is used in a broader sense. This being the primary reason for the focus on IT infrastructure agility instead of flexibility.

AGILE IT INFRASTRUCTURES

IT infrastructure plays a critical role in an organization's competitive advantage as it is the enabling foundation of shared information technology capabilities upon which business depends (McKay & Brockway, 1989). IT infrastructure includes the hardware, operating systems, software applications, data, and the underlying network required to support business operations, as well as the enterprise personnel that interacts with it. Consequently, it comprises two interrelated but distinct components: (a) technical IT infrastructure and (b) human IT infrastructure (Broadbent & Weill, 1997). Enterprise agility requires both components to be agile. The following paragraphs explain the meaning of agility in human and technical IT infrastructure.

Agility in Human IT Infrastructure

Agility in human IT infrastructure implies that the relevant employees have adopted the perspective that they work in an unstable environment where any change may occur at anytime and that they have the skills, knowledge and expertise to cope with change efficiently. To be specific, managers should be able to make the right decisions at the right time, while other employees, such as software engineers and business process owners, should have the knowledge to immediately configure or implement a solution. According to Lee, Trauth and Farwell (1995), human IT infrastructure needs four types of knowledge and skills: technology management, business functional, interpersonal, and technical. Technology management knowledge and skills concern an understanding of where and how to deploy IT effectively and profitably by achieving the strategic goals of an enterprise. Business functional knowledge and skills refer to the ability to understand business problems and develop the required technical solutions. Interpersonal and management knowledge and skills involve abilities such as planning, organizing, teaching, and leading. Lastly, technical knowledge and skills include abilities in technical areas, such as computer operating systems, telecommunications, application specific software, and so forth.

Agility in Technical IT Infrastructure

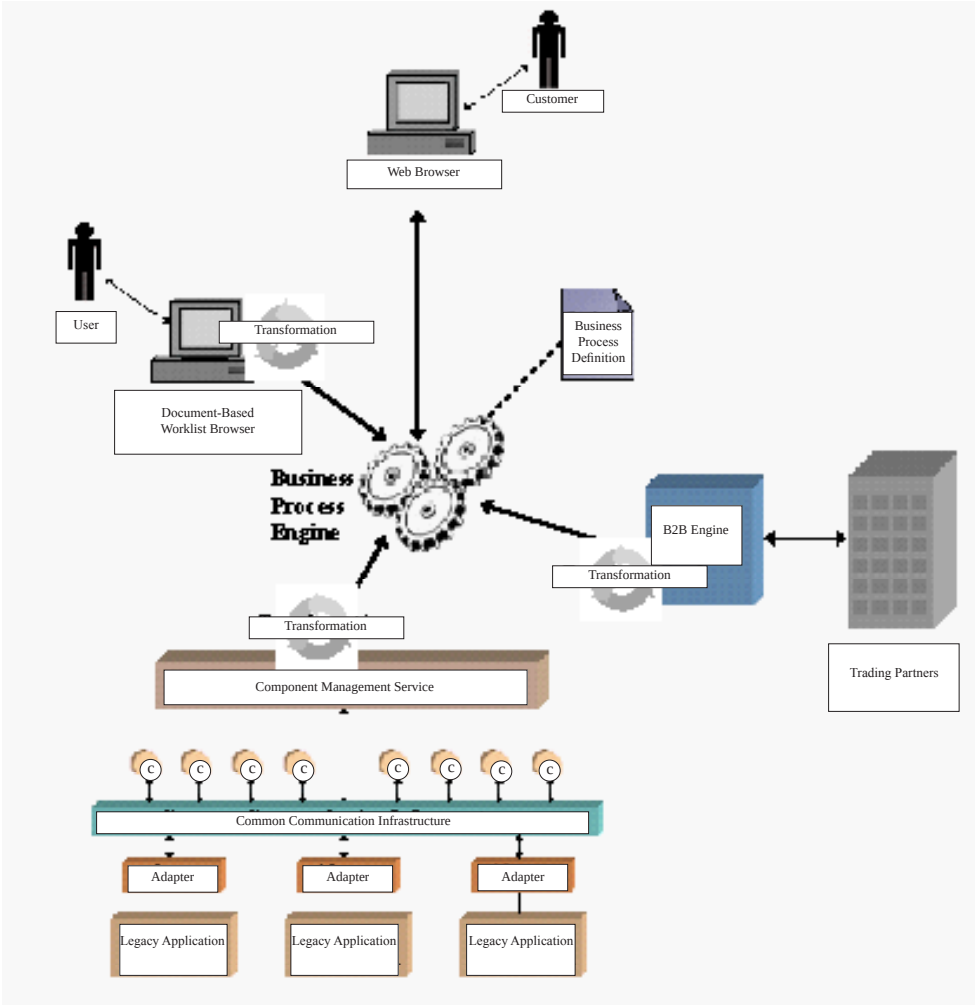
Agile IT infrastructure is one that has been developed according to an agile IT architecture. IT architecture is a framework which forms a guide for an IT infrastructure implementation by indicating technology components in an integrated view that shows how they collaborate to deliver business or technical services. Earl (1989) suggests a typical IT architecture has blueprints for the computing, data, communications, and the application systems of the organization. When an IT architecture is not developed down to a detailed technical level and forms a more abstract view instead, it is called *conceptual architecture*.

In order to be deemed agile, a technical IT architecture should offer a high degree of automation, integration, and flexibility (Alexopoulou, Kanellis, & Martakos, 2004). A conceptual agile IT architecture is illustrated in Figure 1. This architecture ensures high automation because it is based on executable business processes. An executable business process is a kind of enterprise process, whose life cycle is controlled by a business process engine (Nickull et al., 2001). Executable business processes are described in XML-based business process languages such as Business Process Modeling Language (Assaf, 2001), XLANG of Microsoft (Thatte, 2001) and Business Process Execution Language (Thatte et al., 2003), which are machine-readable.

As shown in Figure 1, the Business Process Engine (BPE) constitutes the heart of the architecture since it interacts through the exchange of messages with: (a) users via a document-based Worklist Browser, (b) customers via a Web Browser, (c) trading partners via the B2B engine, and (d) applications and components via the component management service (CMS). The BPE reads and executes business logic defined in process definition documents and acts as a coordinator of activities spanning across the enterprise entities, invoking for each activity the entity that is responsible for performing it. Whenever messages sent by the BPE need to be transformed into another format, a transformation mechanism is used. For example, if a message is to be directed to a worklist browser, it must be first transformed into HTML. Likewise, at the application component level, if for example CORBA (Common Object Request Broker Architecture) is used, then the messages sent by the BPE will have to be transformed into CORBA IDL messages. Overall, the B2B engine will have to transform them onto the format required by the protocol used in the specific business collaboration, since the B2B engine is able to support various B2B protocols (Dabous, Rhabi, Ray, & Benatallah, 2003).

The CMS finds and invokes the appropriate application components that deliver the requested business service. These components are called business aware components, while the components implementing the fundamental

Figure 1. Agile IT architecture (Source: Alexopoulou et al., 2004)



infrastructure services are called framework service components (Raymer, Afrin, & Trivedi, 2001). The components can intercommunicate over a common communication infrastructure. Legacy applications can be connected to the communication infrastructure via adapters. The CMS together with this infrastructure constitute an enterprise application integration (EAI) (Puschmann & Alt, 2004) which follows some of the principles of the NGOSS (New Generation Operations System and Software) framework (Raymer, Afrin, & Trivedi, 2001). NGOSS is an initiative of the TeleManagement Forum set to develop a framework for rapid and flexible integration of operations and business support systems in telecommunications, but it can be equally applied to other business areas as well. NGOSS defines a service-oriented modular system framework, which is based on a collection of loosely coupled, re-usable components that perform business services. Two very good candidates for the implementation of NGOSS are Web Services (Fremantle, Weerawarana, & Khalaf, 2002) and the Java 2 Platform Enterprise Edition (J2EE) architecture (Sun Microsystems, 2003). Web services are loosely-coupled, re-usable software components that semantically encapsulate discrete functionality and are distributed and semantically accessible over standard Internet protocols (Stencil Group, 2001). J2EE is a set of specifications for developing multi-tier enterprise applications in Java.

The integration and flexibility capabilities of this architecture are described at three different levels, namely application components, data, and business process level, in Table 2.

Table 2. Integration and flexibility capabilities of agile IT infrastructures (Source: Alexopoulou et al., 2004)

	Integration	Flexibility
Business Processes	Support for business processes that span multiple applications regardless of whether these applications belong to a single or to different companies.	A business process definition can be altered without requiring modification of the application components.
Data	Data reside in any data source anywhere and can be used by any application or system anywhere.	Data can be easily transformed from one format to another at run time.
Application Components	Components can communicate efficiently with each other as well as with legacy applications.	New components can be easily embodied into the existing architecture and also components can be re-used across multiple business scenarios.

HOW AGILE IT INFRASTRUCTURES CONTRIBUTE TO ENTERPRISE AGILITY

The sensing agility of an enterprise is facilitated by agility in human IT infrastructure, since timely discerning an imminent need for change depends on the knowledge, experience, and skills like perspicacity of the relative enterprise personnel. However, the acquirement of the necessary knowledge is significantly facilitated by an agile technical IT infrastructure as the latter ensures the availability of the right information at the right time by enabling a seamless flow of information.

As far as response enterprise agility is concerned, it is equally facilitated by agility in both human and technical IT infrastructure. Human agility implies that the right decisions are timely made in response to a change. On the other hand, an agile technical IT infrastructure provides for a rapid and cost-effective implementation of the decided-upon solutions. Following, we will further elaborate on the way the agile IT architecture described earlier accommodates change.

According to Dove (1994b), an agile enterprise can employ business process reengineering as a core competency when a transformation need arises. In the described architecture, abstraction of the business process flow into an entity (BPE) separate from the application components themselves allows an easier and more flexible way to alter the business process logic whenever new circumstances arise or a modification is needed. The only action required in such a case is an update in the business process definition that is executed by the BPE, while no modification is needed at the application component level. Separating process control removes the need for the individual components to have knowledge of the business logic associated with process operation. When invoked by process control, a component simply performs the service offered through its interface. Also, components can be re-used across multiple business scenarios.

To achieve even greater flexibility at runtime, business process definition could follow the methodology described by ShuiGuang, Zhen, ZhaoHui and LiCan (2004). According to this methodology, a business process is composed of *general activities*, which are predefined in detail at design time, and *flexible activities,* which are like a "black box", representing an undetermined sub-process without detailed specification at build-time. In other words, flexible activities encapsulate the uncertain sub-process at run time. At run time, depending on current circumstances, a flexible activity can be replaced by a concrete sub-process composed of selected activities from existing or newly-added activities (constituting a pool of activities) based on selection and composition constraints.

However, this method implies that activities included in the aforementioned pool correspond to predictable situations, but what about the case of a situation that has not been predicted? Agility, as we mentioned earlier concerns also efficient response to unexpected change. In case of an unexpected change, the required services to accommodate it may not be offered by the existing IT infrastructure simply because the relevant application components may not exist. In such a case, the most important role is played by the human IT infrastructure because it has to be decided whether the

required services will be outsourced or developed in-house or acquired through commercial, off-the-shelf products. As agility is a function of both cost and time (Dove, 1995), it cannot be claimed that a specific solution will be a panacea to all situations.

From the perspective of an agile technical IT infrastructure, the only contribution that can be offered is the facilitation of an easy and rapid incorporation of the new application into the existing infrastructure; either this application has been developed in-house, purchased, or outsourced. The aforedescribed IT architecture supports plug and play components, which means that new components can be easily embodied into the infrastructure and communicate with the already existing applications via the common communication vehicle.

FUTURE TRENDS

As stated earlier, software implementation may be inevitable in case new services are required upon an environmental change. However, as programming is a difficult and time-consuming task, developing models instead of code would significantly augment enterprise agility. Toward this direction, several initiatives have already emerged. Borland is developing a software product called Themis, which will have a module that will turn models automatically into programming code (*The Economist*, 2004). OMG has proposed the Model Driven Architecture (MDA) (www.omg.org/mda) initiative that addresses the problem of integration and interoperability by making UML models and modeling artifacts more executable. In other words, MDA is about using modeling languages as programming languages rather than merely as design languages. Likewise, Web Modeling Language (WebML) (www.webml.org), which is a notation for specifying complex Web sites at a conceptual level, enables a model-driven approach to Web site development. It is expected that this shift of emphasis on models and their automatic translation to source and object code underlines the trends for both research and practice in this area.

CONCLUSION

The pace of business change drives organizations to respond as quickly as possible. In order to cope with continually changing environments they have to be agile. IT infrastructure facilitates and enables agility within an organization. From a technical perspective, an agile IT infrastructure requires a high degree of automation, integration and flexibility. The fact that this paper focuses mainly on technical IT infrastructure does not imply that the human part has less impact on enterprise agility. In fact, human IT infrastructure also plays a critical role in an organization's ability to sense environmental change and respond efficiently and effectively to that change. Even if a "super agile" technical IT infrastructure has been developed, it will be of no value at all if the relative personnel are unable to understand and adjust to new circumstances. Humans are key players; they discern environmental changes and make the appropriate decisions to cope with these changes. Of course, they must always be supported by an agile technical IT infrastructure. Otherwise stagnation will follow with the enterprise being unable to implement its strategies and hence ensure its continuous profitability and survival.

REFERENCES

Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2002). *Agile software development methods. Review and analysis*. VTT Publications.

Alexopoulou, N., Kanellis, P., & Martakos, D. (2004, April 14-17). Managing information flow dynamics with agile enterprise architectures. *Proceedings of Sixth International Conference on Enterprise Information Systems*, Porto, Portugal (Vol. 1, pp. 454-459).

Assaf, A. (2001). Business process modeling language (BPML). *Business Process Management Initiative.* Retrieved from http://www.bpmi.org

Aydin, M., Harmsen, F. Slooten, K., & Stegwee, R. (2004). An agile information systems development method in use. *Turkish Journal of Electrical Engineering & Computer Sciences, 12*(2), 127-138.

Beck, K. (1999). *Extreme programming explained – Embrace change*. Addison-Wesley.

Broadbent, M., & Weill, P. (1997). Management by maxim: How business and IT managers can create IT infrastructures. *Sloan Management Review, 38*(3), 77-92.

Conboy, K., & Fitzgerald, B. (2004, November). Toward a conceptual framework of agile methods: A study of agility in different disciplines. *Proceedings of the 2004 ACM Workshop on Interdisciplinary Software Engineering Research*, Newport Beach, CA (pp. 37-44). ACM .

Dabous, F., Rhabi, F., Ray, P., & Benatallah, B. (2003). Middleware technologies for B2B integration. *Annual Review of Communications, 56*(3). International Engineering Consortium.

Dove, R. (1994a). Plumbing the agile organization. *Production, 106*(12), 14-15.

Dove, R. (1994b). The meaning of life and the meaning of agile. *Production ,106* (11), 14-15.

Dove, R. (1995). Measuring agility: The toll of turmoil. *Production, 107*(1), 12-14.

Dove, R. (1999a). Design principles for highly sdaptable business systems with tangible manufacturing examples. *Maynard's Industrial Handbook*. McGraw Hill.

Dove, R. (1999b). Knowledge management, response ability and the agile enterprise. *Journal of Knowledge Management, 3*(1), 18-35.

Dove, R. (2005). *Fundamental principles for agile systems engineering*. Conference on Systems Engineering Research, Stevens Institute of Technology, Hoboken, NJ.

Dove, R., Benson, S., & Hartman, S. (1996, March). *A structured assessment system for groups analyzing agility*. Fifth National Agility Conference, Agility Forum, Boston.

Duncan & Bogucki, N. (1995). Capturing flexibility of information technology infrastructure: A study of resource characteristics and their measure. *Journal of Management Information Systems, 12*(2), 37-57.

Earl, M. J. (1989). *Management strategies for information technology*. UK: Prentice-Hall.

The Economist. (2004, November 27). p 75-77.

Fowler, M., & Highsmith, J. (2001, August). *The agile manifesto. software development.* Retrieved from http://www.sdmagazine.com/documents/s=844/sdm0108a/0108a.htm

Fremantle, P., Weerawarana, S., & Khalaf, R. (2002). Enterprise services. *Communications of the ACM, 45*(10), 77-82.

Goldman, S. L., Roger, N. N., & Kenneth, P. (1995). *Agile competitors and virtual organizations*. New York: Van Nostrand Reinhold.

Haeckel, S. H. (1999). *Adaptive enterprise: creating and leading sense-and-respond organizations*. Boston: Harvard Business School Press.

Henbury, C. (1996). *Agile enterprise/next generation manufacturing enterprise*. Retrieved from http://ourworld.compuserve.com/homepages/chrshire_ henbury/agility.htm

Kidd, P. (1994). *Agile manufacturing: forging new frontiers*. Addison-Wesley.

Lee, D.M.S., Trauth, E., & Farwell, D. (1995). Critical skills and knowledge requirement of IS professionals: A joint academic/industry investigation. *MIS Quarterly, 19*(3), 313-340.

Levine, L. (2005, May 8-11). Reflections on software agility and agile methods: challenges, dilemmas and the way ahead. *IFIP TC 8 WG 8.6 International Working Conference, Business Agility and Information Technology Diffusion*, Atlanta, GA (pp. 353-365).

McKay, D. T., & Brockway, D. W. (1989). *Building IT infrastructure for the 1990s*. Stage by Stage (Nolan Norton & Company).

Nerur, S., Mahapatra, R., & Mangalaraj, G. (2005). Challenges of migrating to agile methodologies. *Communications of the ACM, 48*(5), 73-78.

Nickull, D., Dubray, J., Colleen, E., Van der Eijk, P., Vivek, C., Chappell, D., et al. (2001). *Professional ebXML Foundations* (1st ed.). WROX Press Inc.

Object Management Group. Model driven architecture. Retrieved from http://www.omg.org/mda

Puschmann T., & Alt, R. (2004). Enterprise application integration systems and architecture: The case of the Robert Bosch Group. *Journal of Enterprise Information Management, 17*(2), 105-116.

Raymer, D., Afrin, S., & Trivedi, R. (2001). NGOSS architecture technology neutral specification. TeleManagement Forum. Retrieved from http://www.tmforum.org

Schwaber, K., & Beedle, M. (2002). *Agilesoftware development with SCRUM.* Prentice Hall.

ShuiGuang, D., Zhen, Y., ZhaoHui, W., & LiCan, H. (2004). Enhancement of workflow flexibility by composing activities at run-time. *Proceedings of the 2004 ACM Symposium on Applied Computing* (pp. 667-673).

Stencil Group. (2001). Defining Web services. Retrieved from http://www.stencilgroup.com/ideas_scope_200106 wsdefined. html

Sun Microsystems. (2003). Java™ 2 Platform Enterprise Edition Specification, v1.4.

Thatte, S. (2001). XLANG Web services for business process design. Microsoft Corporation. Retrieved fro, http://www.gotdotnet.com/team/xml_wsspecs/xlang-c/default.htm

Thatte, S., Andrews, T., Curbera, F., Dholakia, H., Goland, Klein, J., et al. (2003). Business process execution language for Web services Version 1.1 5. BEA Systems, International Business Machines Corporation, Microsoft Corporation, SAP AG, Siebel Systems.

Williams, L., & Cockburn, A. (2003, June). Agile software development: It's about feedback and change. *IEEE Computer,* 39-43.

KEY TERMS

Agile Enterprise: A fast moving, adaptable and robust business, which is capable of rapid adaptation in response to unexpected and unpredicted changes and events, market opportunities and customer requirements.

Agile IT Infrastructure: A highly automated, integrated and flexible IT infrastructure which enables efficient and effective response to planned, as well as unanticipated change.

Agility: Efficient and effective response to planned as well as unanticipated change.

Executable Business Process: A kind of enterprise business process, whose life cycle is controlled by a Business Process Engine.

IT Architecture: A framework which forms a guide for an IT infrastructure implementation by indicating technology components in an integrated view that shows how they collaborate to deliver business or technical services.

IT Infrastructure: The enabling foundation of shared information technology capabilities upon which business depends.

Response Agility: The organizational capability to quickly transform knowledge into action in response to the environmental signals.

Sensing Agility: A firm's ability to rapidly discover and interpret the market opportunities through its information systems.

System: A group of interacting modules sharing a common framework and serving a common purpose.

Agile Knowledge Management

Meira Levy
Haifa University, Israel

Orit Hazzan
Technion – Israel Institute of Technology, Israel

INTRODUCTION

This article is based on the assumption that Knowledge Management (KM) is a vital part of any project. Based on this working assumption, the purpose of this article is to introduce the term *Agile Knowledge Management* (AKM) by illustrating how the *Agile Software Development* (ASD) approach is suitable for the introduction of KM processes.

The ASD approach emerged over the past decade in response to the unique problems that characterize software development processes (Highsmith, 2002). In general, ASD emphasizes customer needs, communication among team members, short releases and heavy testing throughout the entire development process. These ideas are implemented quite variedly by the different ASD development methods.

Knowledge Management (KM) and Agile Software Development (ASD) are two organizational processes that face common barriers when introduced and applied. This article suggests that because the field of KM presents a less disciplined approach compared with ASD, it is logical that KM practitioners should learn how ASD has coped with very similar barriers. We further illustrate how it is but natural to emphasize the concept of *Agile Knowledge Management* (AKM) in order to improve KM processes, because ASD already encompasses the organizational and cultural infrastructure needed for KM.

The pairing of KM and ASD is not new; a connection between the two concepts has been acknowledged by various researchers. For related discussions, see, for example, Dove (1999) and Holz, Melnik and Schaaf (2003). This connection, however, is not surprising because both disciplines deal with organizational culture and change management.

In what follows, we further highlight the connection between the two fields. First, we show that the two processes, KM and ASD, face the same barriers when introduced into an organization. We also include some suggestions for coping with such barriers. Second, we highlight the way in which KM is already embedded into ASD processes. Thus, in order to improve KM in such processes, it should be made more explicit. Accordingly, we introduce an *agile* KM manifesto.

BACKGROUND

In today's competitive global market, companies are required to manage their intellectual resources as well as their financial ones. KM is therefore recognized as a legitimate management practice that helps organizations distribute the right knowledge to the right people at the right time (Van der Spek & Carter, 2003). Furthermore, KM is considered to be the main source of competitive edge for companies when facing new opportunities, time-to-market demands and frequent changes in their technological and business environments. At the same time, however, research reveals that some organizations do not apply systematic KM processes and support, but rather rely mostly on common sense. Barriers, such as competition instead of collaboration, cultural differences, the pressures of daily challenges, lack of communication tools and places to meet, stubbornness of people or lack of discipline within the company, might interfere (Van der Spek & Carter, 2003). In addition, cultural and job security issues prevent managers from investing in KM initiatives (Drucker, 1998).

Similarly, the main barrier when introducing ASD into software organizations is the need to cope with the conceptual change, mainly the organizational cultural change, that ASD brings with it. Following are two illustrations of the conceptual change required when applying the ASD approach.

First, cooperation should replace the knowledge-is-power perception. ASD introduces a management paradigm that encourages collaboration, communication and the *whole team* concept. At the same time, however, the software development culture, which has evolved over the years, sometimes encourages opposite values and manners, as expressed, for example, by the concealing of information and the isolating of people in cubicles. Second, in ASD processes, a change is required also in the customer's conception and involvement, as well as in customer-developers relations. ASD requires intensive and frequent communication with the customer. Clearly, this is a significant difference compared with the common level of interaction with the customer as practiced today in many software organizations.

Studies reveal that the introduction of KM and ASD processes increases productivity, shortens time-to-market and results in higher product quality (see, for example, Bennet & Bennet, 2003; Reifer, 2002). Yet, as mentioned above, it

is but logical that practitioners feel insecure when required to undergo such change. What is needed then in many cases is a realization that the new paradigm, whether it is KM or ASD, in fact constitutes a new and different infrastructure within it concerns can be addressed.

It is in this spirit that we further illustrate the close relationship between the two processes by presenting nine arguments that are often raised when KM and ASD processes are introduced. For each argument, we present one frequently-heard statement for KM and ASD, and suggest an approach for overcoming the said argument. This presentation format reflects both the similarity in the resistance to the two processes, as well as the similar way in which this resistance can be addressed in both cases.

I. "It is not needed at all"

KM: "Someone in the organization is already taking care of the KM process."
ASD: "Our organization already has a very well-defined development process that works just fine."
Possible response: "Can you please elaborate on the benefits and weaknesses of your current process?" In many cases, an attempt to answer this question reveals the problems that exist.

II. "Time does not permit"

KM: "I must deliver the project on time and I'm behind schedule. I can't devote any time to knowledge sharing." OR: "Do I have to invest any more work in order to manage knowledge?"
ASD: "I must deliver the project on time and I'm behind schedule. I can't devote any time to testing." OR: "How much extra time is needed in order to collect the metrics? Is it worth investing?"
Possible response: "What are the main reasons for the gaps between the project planning and the actual progress?" In many cases, it is found that the reasons given here further highlight the importance of some elements of KM and ASD processes.

III. "The current tools work very well"

KM: "We have tools for KM, such as WSS, which hosts many discussion forums."
ASD: "We already have a mechanism for sharing our metrics using an online tool that is accessible to all; why should we sit together in one lab/informative workspace?"
Possible response: "In your opinion, how frequently do people use or open this tool?" This question highlights the spirit of the first principle of the Agile Manifesto (see Table 1); namely, that we should address the people, not the tools. In other words, tools are useless unless they are simple and accessible and their use is integrated naturally into the work process itself.

IV. "We can't change the status of documentation"

KM: "We have documents that reflect the project knowledge."
ASD: "How can we develop software without comprehensive documentation? After all, the customer asks for it."
Possible response: "Based on your experience, are the documents always compiled along with the actual projects? Also, can you please estimate how often, if at all, the documentation is read?" In many cases, this question leads practitioners to realize the gap that exists between the perceived image and reality. Specifically, it highlights the fact that the documents produced in many typical processes are not the ones truly required (some do not reflect reality at all; others are never read). Rethinking the role of documentation reveals that documentation should not be skipped, but rather carried out in a way that allows for timely and relevant knowledge and information maintenance.

V. "I have had very bad experience with all these buzzwords"

KM: "KM is prosaic, what it is actually?" OR: "Has any business already implemented it? How do you even start a KM project?"
ASD: "There are so many buzzwords in software development. You must convince me that this is not just another one." OR: "It sounds good on paper as a theory. Does it really work?"
Possible response: "You can try to initiate a small scale KM/ASD project and observe its benefits. Also, you can read testimonies that will enable you to move from the abstract to the concrete." In both KM and ASD, experience shows that these processes work well in practice. In the ASD arena, this is manifested by the fact that more and more software houses are starting to work according to the agile software development paradigm[1]; in the case of KM, it is reflected by the increasing numbers of KM initiatives in organizations and the designation of a specific person to manage the organization's KM processes (Van der Spek & Carter, 2003).

VI. "The current working environment provides the mentioned benefits"

KM: "If I need information, I just go to the right person and ask him or her."
ASD: "It's impossible to sit together in one lab. It's too noisy. If I need to ask something, I just go to the right person and ask."
Possible response: "Can you describe what happens when

you find a bug and can't manage to fix it for several days." In many cases, the process described in response to this question starts with the person asking his or her roommate, then the lunch group and, if the person is lucky, he or she will sooner or later come across someone who knows the answer. It is also admitted that sometimes the answer is never actually found and the person asking the question must decide how to proceed without the required knowledge. The outcomes of such circumstances are clear.

VII. "But knowledge is power"

KM: "It is naïve to think that when I need the information, someone will share it with me."
ASD: "I'm the expert in this domain. How can I pair-program with someone who is new to this area?"
Possible response: "Let's think about your argument from a win-win perspective, rather than as a lose-lose situation. If you know something and you don't share it, the organization loses; furthermore, when you need help, you won't get it from your colleagues. This is a lose-lose situation. Let's change it to a win-win situation by sharing." Indeed, in both cases (KM and ASD), a process that ensures mutual trust should be established. See, for example, how it is accomplished in Extreme Programming (Beck, 2001, Beck & Andres, 2005), one of the leading ASD methods. Hazzan and Dubinsky (2003) analyze software development processes from the perspective of game theory and illustrate how Extreme Programming establishes a trustful development environment in which teammates are assured that their cooperation will be reciprocated. Thus, teammates apply agile practices. Although on the face of things, these actions can be perceived as a loss of power, a deeper understanding reveals the benefits that result from the collaborative atmosphere inspired by ASD.

VIII. "Too much money is needed"

KM: "Who should pay for KM?"
ASD: "We aren't sure that we can invest in the infrastructure needed for ASD. We need more computers and we must make changes in the physical space."
Possible response: "Can you please estimate how much you pay for extra hours/delays or for reinventing the wheel?" When this question is answered, it is appropriate to present data on the expected improvements due to KM and ASD processes. For example, with respect to KM, Bennet and Bennet (2003) claim that "those organizations that have found ways to compete successfully within this nonlinear, complex, and dynamic environment can dominate their competitors by as much as 25% in growth rate and profitability relative to the average in their industry" (p. 9). Similarly, Reifer (2002) shows a 25%-50% decrease in time-to-market for ASD processes.

IX. "But the stakeholders are happy now"

KM: "KM is not a standard in our processes."
ASD: "The customers are happy now. They don't ask for it, so why should I invest in it?"
Possible response: "This might be the case now. But let's think about a time when your customer no longer agrees to pay more if your competitors, who manage knowledge or develop software using an agile process, manage to develop a better product in a shorter period of time." Realization has recently set that KM and ASD processes are expected to become more prevalent because customers will no longer agree to pay more when they can receive a better product in a shorter period of time if KM/ASD processes are applied. Accordingly, KM/ASD processes can be viewed as disruptive technology—in the near future customers will start requiring it and organizations that have not adopted it will fall behind.

KNOWLEDGE MANAGEMENT IN AGILE SOFTWARE PROJECTS

The message conveyed in this section is that KM has the potential to be easily accepted into ASD environments. Following are two explanations for this perspective. First, the agile cultural infrastructure already includes values such as cooperation and knowledge sharing; specifically, ASD processes include some practices that support KM, such as stand-up meetings, the planning game, pair programming and the informative workplace. Second, KM is about learning, and ASD establishes an environment that supports learning processes (Hazzan & Dubinsky, 2003).

We suggest that although many software project managers believe in the importance of KM, the resistance exhibited toward KM results from the fact that the actual knowledge on the application of KM does not exist. We therefore propose that by making KM elements more explicit in ASD environments, the barriers presented above, as well as others, can be at least partially overcome. Furthermore, because ASD is based on knowledge sharing and communication, we suggest that ASD also benefits from the fact that the KM elements are clearly specified. This approach is compatible with Holz and Melnik (2004) and Doran (2004).

Following are several illustrations of the natural integration of KM into ASD processes. The AKM activities presented in what follows can be integrated into each of the specific agile software development methods (e.g., Extreme Programming and Scrum).

- **The role of knowledge manager:** Because different agile methods outline different role schemes[2], we suggest assigning one team member to be in charge of the KM process. In large organizations, the group of

people who are in charge of KM can form a KM team that is entrusted with leading the entire organization's KM. The KM officer in each team gathers the relevant issues that are to be discussed in the global organization meetings and brings back new information to his or her team.

- **Retrospective meetings:** A special session can be dedicated to KM issues in the framework of the retrospective meetings. This session might include topics such as new tools used in the latest release and lessons learned during the development of new features. It implies that the entire KM process is also carried out in short releases, as ASD is.
- **Planning game:** In each planning game, time will be allocated during the next iteration for specific KM activities, if needed, as practiced with respect to other tasks that improve the developed product and the development process.
- **Informative workplace and collective ownership:** Each new piece of knowledge can first be posted on the project's whiteboard. Then, these pieces will be evaluated in the framework of retrospective meetings. If they are found to be significant, they will be stored at the end of the iteration/release in a predetermined place.
- **Metrics:** Metrics related to KM will be measured on a regular basis as customary with respect to other aspects of ASD projects. If they are found to be useless, they will be adjusted accordingly.
- **Adaptability:** The development of AKM should gradually encompass all phases and activities of the development process (e.g., requirements, design). The suitable AKM solution for each phase will be shaped and its applicability checked continuously. Because, as previously mentioned, knowledge management activities sometimes create antagonism, we suggest that the agile spirit of the change introduction will render the AKM activities more agreeable.

FUTURE TRENDS

We propose that KM processes should be approached like any other project in the organization. This attitude is compatible with Bailey and Clarke (2000), who state that their "definition of KM as how mangers can generate, communicate and exploit knowledge (usable ideas) for personal and organizational benefits" highlights not only the organizational importance of KM, but also its relevance for individual managerial action. "The challenge therefore is to make KM ideas current, relevant and actionable to diverse organizational/managerial user groups" (p. 237).

As mentioned previously, we suggest that at least part of the resistance exhibited toward KM results from the fact that the actual daily application of these processes is not clear. The agile community has already, at least partially, overcome this barrier, whereby specific agile methods, based on a general manifesto[3], are applied on a daily basis.

In this spirit, we present the Manifesto for Agile *Knowledge Management* (AKM) projects based on the Manifesto for Agile *Software Development*, highlighting the changes we made in bold-face type (see Table 1).

For illustration purposes, we show how the Manifesto for AKM can be implemented with respect to Extreme Programming values:

- **Communication:** Use communicative channels in order to make implicit KM an explicit process.
- **Simplicity:** Use simple solutions that can be embedded in the work process in order to exploit the essence of KM.
- **Feedback:** The team members are the AKM activities' customers. Their feedback should be heard to create a vivid AKM environment.
- **Courage:** Team members should feel comfortable when reviewing and reporting on good or bad practices.
- **Respect:** Team members should respect each other by knowledge sharing, not considering knowledge

Table 1. The manifesto for AKM and the manifesto for ASD

Manifesto for AKM	Manifesto for ASD
We are uncovering better ways of **managing knowledge** by doing it and helping others do it.	We are uncovering better ways of **developing software** by doing it and helping others do it.
Through this work we have come to value:	
- *Individuals and interactions* over processes and tools. - **Sharing knowledge** over comprehensive documentation. - *Customer collaboration* over **persuasive** negotiation. - *Responding to change* over following a plan.	- *Individuals and interactions* over processes and tools. - **Working software** over comprehensive documentation. - *Customer collaboration* over **contract** negotiation. - *Responding to change* over following a plan

retaining as power. Time should be devoted to knowledge sharing activities, respecting it as a means that augment the work process for the benefit of all team members.

CONCLUSION

KM and ASD are essentials in the competitive global market in which organizations are required to respond to customer demands in a timely manner (Bennet & Bennet, 2003; Van der Spek & Carter, 2003). Both disciplines are abstract, deal with intangible artifacts in a competitive market and face common barriers. In this article, we suggest that agile processes can be an appropriate infrastructure for the enhancement of KM. Good practices adopted from ASD, which is a more mature discipline, can help managers overcome their initial resistance to KM, initiate KM projects and embed the KM process into the working process. We call this approach *Agile Knowledge Management*.

REFERENCES

Bailey, C., & Clarke, M. (2000). How do mangers use knowledge about knowledge management? *Journal of Knowledge management, 4(*3), 235-243.

Beck, K. (2001). *Extreme programming explained*. Addison-Wesley.

Beck, K., & Andres, C. (2005). *Extreme programming explained* (2nd ed.). Addison-Wesley.

Bennet, D., & Bennet, A. (2003). The rise of the knowledge organization. In C.W. Holsapple (Ed.), *Handbook on knowledge management 1*. Springer-Verlag.

Doran, H. (2004). Agile knowledge management in practice. *LSO* (pp. 137-143).

Dove, R. (1999). Knowledge management, response ability, and the agile enterprise. *Journal of Knowledge Management*, 18-35.

Drucker, P. F. (1998). The coming of the new organization. Harvard Business Review on knowledge management. Harvard Business School Press.

Dubinsky, Y., & Hazzan, O. (2006).Using a role scheme to derive software project quality. *Journal of System Architecture, 52*(11), 693-699.

Hazzan, O., & Dubinsky, Y. (2003). Bridging cognitive and social chasms in software development using extreme programming. In *Proceedings of the Fourth International Conference on eXtreme Programming and Agile Processes in Software Engineering*, Genova, Italy, (pp. 47-53).

Highsmith, J. (2002). *Agile software developments ecosystems*. Addison-Wesley.

Holz, H., & Melnik, G. (2004). Research on learning software organizations—past, present and future. In *Proceedings of the 6th International Workshop of the Advances in Learning Software Organizations, LSO*, (pp.1-6).

Holz, H., Melnik, G., & Schaaf, M. (2003). Knowledge management for distributed agile processes: Models, techniques, and infrastructure. In *Proceedings of the 12th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE '03)*. IEEE Computer Society Press.

http://www.iis.uni-hildesheim.de/Staff/schaaf/Publications/Resources/kmdap-holz-final.pdf

Reifer, D. (2002). How good are agile methods? *IEEE Software, 19*(4), 16-18.

Van der Spek, R., & Carter, G. (2003). A survey on good practices in knowledge management in European companies. In K. Mertins, P. Heisig, & J. Vorbeck (Eds.), *Knowledge management: Concepts and best practices* (2nd ed). Springer-Verlag.

KEY TERMS

Agile Knowledge Management: Agile knowledge management is the working framework introduced in this article and it is based on good practices adopted from agile software development. Agile Knowledge Management can help managers overcome their initial resistance to knowledge management, initiate knowledge management projects and embed the knowledge management process into working processes.

Agile Knowledge Management Manifesto: This manifesto, which we suggest in this article, adopts and enhances the common principles of agile software development for knowledge management processes.

Agile Manifesto: This manifesto reflects the common principles of all agile software development methods. It is presented at http://agilemanifesto.org/.

Agile Software Development: Agile software development is a management paradigm for software development projects. It emphasizes customer needs, communication among team members, short releases and heavy testing throughout the entire development process. These ideas are implemented quite variedly by the different agile software development methods.

Extreme Programming: Extreme programming is one of the agile software development methods. It is based on a list of principles and practices that reflect the method's values. It is accepted as one of the prevalent agile software development methods.

Extreme Programming Values: Extreme programming is based on five values: Communication, Simplicity, Feedback, Courage and Respect. We suggest adopting those values for AKM processes.

Knowledge Management: Knowledge management is the way knowledge-based companies manage their intellectual assets in order to gain a competitive advantage. Knowledge management includes three dimensions: technology infrastructure, business processes and cultural change.

ENDNOTES

1 Although there is less than 10 years of accumulated experience using the agile approach, it is currently being applied as the development paradigm by about 20% of the companies in North America and Europe. Source: http://www.versionone.com/pdf/AgileMyths_BetterSoftware.pdf

2 See Dubinsky and Hazzan (2006) for a summary of these role schemes.

3 See: http://agilemanifesto.org/

Agile Methodology Adoption

John McAvoy
University College Cork, Ireland

David Sammon
University College Cork, Ireland

INTRODUCTION

Discussions on agile software development methodologies have a tendency to develop into an argument between proponents of agile methods and proponents of more traditional process-oriented methodologies. The terminology used in these debates is often unhelpful, and in many cases are inaccurate and biased representations. It needs to be accepted that there are no "silver bullets" providing universal solutions (Jeffries, 2001). Bearing this in mind, the decision to adopt a particular software development methodology is a difficult one, and the decision to choose an agile method is no exception. In theory, as in practice, definitions and descriptions of the various agile methods are presented, yet the factors considered in the decision to adopt, or not adopt, an agile method are not addressed. While agile methodologies try to avoid the excessive use of procedures or tools (Beck & Fowler, 2001), one agile methodology, dynamic systems development method (DSDM), does recommend the use of appropriate tools during the development process (Coesmans, 2003). However, it appears that none of the available agile methodologies suggest a tool to assist decision makers at the project initiation phase, therefore, the debate on agile suitability is usually a debate on agile versus traditional methods (DeMarco & Boehm, 2002), rather than an examination of the suitability of agile methods for a particular project. While the "agile debate" rages, individual projects are not adequately assessed prior to the adoption of a method.

BACKGROUND

To describe the agile method is a misnomer. The agile software development method does not exist; it is instead a collection of methodologies with common core values, where examples of agile approaches include: Extreme Programming (XP); Crystal Methods; SCRUM[1]; DSDM; Feature Driven Development (FDD); and Adaptive Software Development (ASD) (Highsmith, 2001; Sutherland, 2001). For many proponents of the agile methodologies, the epoch of the agile movement was February 11th, 2001, when representatives of the different agile methodologies convened in the mountains of Utah to create the "Manifesto for Agile Software Development."

The agile manifesto is a collection of values that underlie all agile methodologies:

- Individuals are more important than processes and tools.
- Working software is more important than comprehensive documentation.
- Customer collaboration is more important than contract negotiation.
- Responding to change is more important than following a plan.

Therefore, agile methods are a response to the inability of traditional methods to embrace change in a turbulent business environment that demands software to meet its needs quickly (Highsmith & Cockburn, 2001), and Rising and Janoff (2000) describe it as the need to "meet customer needs and turn this chaos to our advantage" (p. 3). The manifesto, its origins, and its importance to the agile methods are discussed in a variety of research including: Boehm and Turner (2003, 2004); Fowler and Highsmith (2001); Highsmith (2004); Koch (2004); and Lindvall et al. (2002).

VIABILITY OF ADOPTING AN AGILE APPROACH

Throughout the available literature relating to agile methodologies, factors important to the success of an agile project are discussed, yet these factors are not specifically used to determine the viability of adopting an agile approach. For example, the agile manifesto, described in Abrahamsson, Salo, Ronkainen, and Warsta (2002), is a list of aspirations or ideals, and as such is not readily quantifiable as a method of adoption assessment. However, some researchers have provided a number of approaches to assessing various aspects of suitability for agile.

Boehm and Turner (2003, 2004) provide five factors (or dimensions, to use their term) in a graphical approach to assess whether an organization is an agile or more traditional organization. Boehm and Turner's (2003, 2004) graphical

analysis provides an evaluation of whether an agile or process-oriented (they use the term disciplined) approach represents the current state of a project or organisation. The graphical output highlights, by the location of ratings on five axes, whether the project or organisation, is leaning towards agile or disciplined (see Figure 1).

Boehm and Turner (2003, 2004) list five critical agile/disciplined decision factors:

- the criticality of the project (the level of loss due to the impact of defects),
- the dynamism of the project (level of requirements change),
- the size of the project,
- the expertise of the team, and
- the culture of the project or organisation (empowerment versus process driven).

McAvoy and Sammon (2005) investigate the suitability of an agile approach for a project through conducting a series of workshops in a number of software development companies. McAvoy and Sammon developed a simple decision support tool, referred to as an Adoption Assessment Matrix, as illustrated in Figure 2, based on critical adoption factors relating to agile methods, addressing a need in industry; namely, to improve the overall understanding of the constituent parts of agile systems development methodologies. The critical adoption factors are described in more detail later on when discussing the groupings of critical adoption factors. It is argued by McAvoy and Sammon (2005) that the use of a decision support tool, aiding decision makers, to determine the viability of an agile method for a specific software project has proved hugely beneficial, for example, a major benefit of the tool is that it guides discussion, concentrating the debate on the critical factors, applied to the individual project. According to McAvoy and Sammon (2005), these discussions proved to be as valuable as the output of the tool itself. The results of these workshops show that an argument can be made for the use and benefit of such a decision support process in industry, in supporting the decision to adopt an agile approach.

The research work of Boehm and Turner (2003, 2004), and McAvoy and Sammon (2005) presents a high-level grouping of factors which can form the basis for a discussion around the agility of an organization, a particular project and its suitability for an agile approach, while assessing the likelihood of success. These groupings referred to as *project; customer; team;* and *organisation* are open to debate and refinement. Like any list, for example, be it top 10 athletes or favorite music, there will never be universal agreement on its constituent parts. The list is not set in stone, but an extremely important starting point. For example, both Boehm and Turner (2003, 2004), and McAvoy and Sammon (2005) agree on the issue that criticality of a project is an important factor (life-critical projects should be automatically excluded as possibilities for an agile approach), while McAvoy and Sammon add factors relating to the relationship with the customer as also warranting consideration. These groupings are discussed in the next section.

Figure 1. Boehm and Turner (2003, p. 5) graphical scatter plot

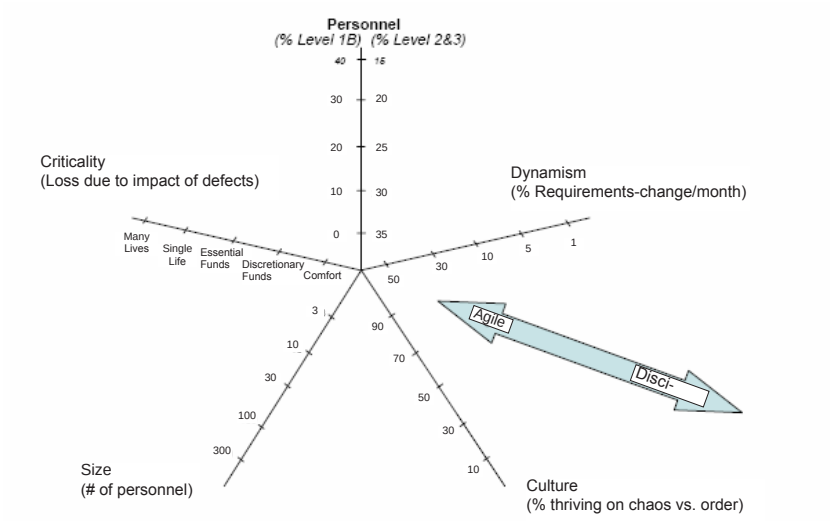


Figure 2. Adoption Assessment Matrix

Critical Adoption Factor	Weighting	Rank Case	Result Case
Duration of the project 1=more than 5 years 2=less than 6 months	4		
Location of the customer 1=many customers in many countries 5=In house	4		
Customer involvement 1=will have no interaction 5=willing to interact	4		
Acceptance of change (to requirements) 1=rigid 5=flexible	4		
Team size 1=more than 20 5= up to 3	3		
Skill of team 1=inexperienced 5=very experienced	3		
Organisational and reporting structure 1=many reporting layers 5=flat structure	2		
Process 1=5 or more standards to follow 5= no standards to follow	2		
Documentation requirements 1= a lot of documentation required 5= very little documentation required	1		
Layout of Workspace 1= individual cubes, people isolated 5= open plan, no walls	1		
Confidence Rating	**28**		

Note on Workings of Matrix:

[1] Rankings (between 1 and 5) are applied to each of the critical adoption factors. These rankings are based on a workshop participants assessment of the factors applicability to a particular project.

[2] Each ranking is multiplied by the weighting for each critical adoption factor, providing a result for each critical adoption factor.

[3] Finally, the totalled output, referred to as the 'confidence rating' is calculated as follows:

Confidence Rating Calculation = (SUM(Result Case)/(28*5))*100

A DISCUSSION ON THE GROUPINGS OF CRITICAL FACTORS

The *project* grouping describes the relevance of critical adoption factors when considering the mechanics of the project being undertaken. Issues of importance that can impact on the appropriateness of adopting agile methods include: the duration of the project; increasing uncertainty and changing requirements within the project; and the criticality of the project. "Today's project manager must deliver concrete results in shorter time frames while being constantly bombarded with myriad changes and risk laden decisions" (Highsmith, 2001, p. 260). This concern aligned with Kruchen's (2001) criteria of reacting quickly to market, impose a shorter time frame on agile projects. A core value of agile development is the acceptance of change, usually visible through changing requirements. Highsmith (2002) proposes that as the uncertainty associated with a project increases, the suitability of the agile approach will also increase. Kirkpatrick, Walker, and Firth (1992) describe the changing of requirements (requirements volatility) as one of the major sources of risk in a project. However, this directly contradicts the agile viewpoint, while stories and requirements definitions in the agile methodology continue to be written throughout the project. This is regarded as one of the basic edicts of the agile approach (Beck & Fowler, 2001). Finally, agile techniques, such as ASD are unsuitable for critical systems, such as air traffic control software. However, this is not a major concern as the majority of systems do not have this degree of criticality (Emery, 2001).

The *team* grouping highlights the importance of the project team, in terms of team size and skill level of the team, for adopting agile methods. While the team size is important, it is not of vital importance. Boehm (2002) describes how agile development is optimal when used in small teams. The skill level of the team is a continually stated requirement in research into agile methods. Martin (2003) further defines skill level by stating that a strong player does not necessarily have to be an expert programmer, though they must work well with others. Good communication skills and an ability to interact with others are of higher importance than expertise in a programming language. Furthermore, Reifer (2002) makes an interesting observation, which may have repercussions for agile research and implementation. Reifer (2002) surveyed 31 agile projects across eight industry sectors and comments that the teams involved were made up of motivated, experienced programmers in cohesive teams.

The *customer* grouping describes the customer, or customers, of the project, identifying location of the customer and customer involvement as critical adoption factors to consider in adopting an agile methodology. Beck and Fowler (2001) are willing to compromise on the definition of customer. The customer is the person who makes the business decisions, that is, completion date, scope, and so forth. This person can be an internal product manager or an individual who will purchase the software. Young (2003) describes how a DSDM project in British Airways addressed the necessity for user involvement in DSDM projects. The project was in the area of e-commerce, so its user base was extensive and difficult to categorise. British Airways used sales and marketing staff to represent the customers.

The *organisation* grouping highlights the criticality of the organisational environment, which comprises the following: organisational and reporting structure; process; documentation requirements; and layout of workspace. Boehm (2002) states that agile methods require both responsive people and organisations. The relationship between managers and developers is one of collaboration rather than the traditional command and control structure (Cockburn & Highsmith, 2001).

It is clear that process, or the lack of process, plays a significant role in agile methods (Highsmith, 2000). Discussing the agile concept and the importance of individuals over process highlights that "a good process will not save the project from failure if the team doesn't have strong players" (Martin, 2003, p. 4). Therefore, the absence of rigid processes is embraced. "Fast companies have a defining sense of purpose, supported by a few simple rules. They determine

what is essential and ignore the rest" (Gandossy, 2003, p. 32). For example, Greening (2001) describes the difficulties involved in the implementation of XP, in a company that has a formal software development process. A variation, or continuation, of the discussion on process is the documentation requirements of an agile project. Greening (2001) stresses that XP does not prohibit documentation, it merely stresses that documentation has a cost. Any documentation to be used must be evaluated to ensure that it has a benefit and that it is definitely necessary. The layout of the workplace receives considerable attention in literature on agile methods. Poole and Huisman (2001) describe the efforts made in Iona Technologies to enhance team communications. The previous environment they describe resembles that of the Dilbert cartoon. The multiple bays of shoulder height cubicles were replaced by a common area, or group workspace. This was an attempt to increase awareness of the team, as opposed to individuals in their own tiny spaces. Kalita (2003) demonstrates that colocated teams, although not a formal principal of DSDM, should be used in DSDM projects to facilitate the necessity of communications.

FUTURE TRENDS

The assessments of Boehm and Turner (2003, 2004), and McAvoy and Sammon (2005), while similar, are addressing different needs. However, there is benefit in combining both approaches to determine the "as is" state and the potential future or "to be" state of an organisation or project team. Ultimately, if an agile approach is indicated as having value and merits consideration, following the two assessments, further work could be done in determining how to adopt an agile method. An example of this "how" is presented in Pikkarainen and Passoja (2005), who describe a case study of how an assessment was conducted to determine which of the agile practices were suitable for a product development organisation. They recommend the tailoring of the agile approach to match the needs of the organisation. Although the case study describes one organisation, they propose that the assessment could be used in further organisations. Finally, in analysing these perspectives it is important to raise an extremely important point concerning the use of these tools proposed by Boehm and Turner (2003, 2004), and McAvoy and Sammon (2005), such that; although agile methodologies recommend limited use of tools, the purpose of the tools proposed by Boehm and Turner (2003, 2004), and McAvoy and Sammon (2005) is to support decision makers in (1) identifying the current state of an organizations agility and (2) assessing the suitability of adopting an agile approach to a project and the likelihood of success. Therefore, one should consider the use of these support tools to be a precursor to the selection and use of an agile methodology.

CONCLUSION

Agile methods have been described as glorified hacking by those who argue against their adoption and use. A more structured approach to the assessment of the adoption of an agile method would help to dispel these views. Some work has been carried out highlighting this argument; for example, Svensson and Host (2005) investigated the introduction of an agile method and concluded that an assessment should have been done before the project. As an example, by aligning all the assessment methods discussed in this chapter, an investigation could provide the following insights:

1. Identify the current state of an organizations agility.
2. Identify if an agile approach would be successful.
3. Identify what processes to use when adopting an agile approach.

As managers look to agile methods as a solution for the myriad of problems inherent in software development, there can be a rush to judgment, where an assumption is made that an agile approach will work for a project. Projects are rarely assessed prior to the adoption of an agile method. Ultimately this will increase the likelihood of problems and ultimately the failure of projects. However it is worth considering that problems that occur during agile projects could have been predicted if a proper assessment had been conducted and steps taken to mitigate these potential problems. On the other hand, an assessment could determine that an agile method would be unsuitable for a project, and considerable time and effort could be saved. The debate is moving on from pro- and anti-agile to one of the acceptance of agile in certain circumstances. The view of the agile method as a "silver bullet" should not exist so there is a necessity to determine which projects could benefit from an agile approach and which projects need a more traditional process-based approach.

Applying an "engineering" approach to the adoption of agile methods in no way diminishes the core values of the agile approach. Agile recommends against the overuse of tools, but this applies to the use of tools within the project. Using assessment tools to determine if a project should be an agile one is a necessary precursor to adopting an agile method.

REFERENCES

Abrahamsson, P., Salo, O., Ronkainen, J., & Warsta, J. (2002). *Agile software development methods. Review and analysis*. Finland: VTT.

Beck, K., & Fowler, M. (2001). *Planning extreme programming*. NJ: Addison-Wesley.

Boehm, B. (2002). Get ready for agile methods, with care. *IEEE Computer, 35*(1), 64-69.

Boehm, B., Turner, R. (2003) Rebalancing your organizations agility and discipline. In F. Maurer & D. Wells (Eds.), *Extreme programming and agile methods—Xp/Agile universe 2003* (pp. 1-8). Berlin, Germany: Springer-Verlag.

Boehm, B., & Turner, R. (2004). *Balancing agility and discipline*. MA: Pearson Education.

Cockburn, A., & Highsmith, J. (2001). Agile software development: The people factor. *IEEE Computer, 34*(11), 131-133.

Coesmans, P. (2003). DSDM in a non-IT project. In J. Stapleton (Ed.), *DSDM business focused development* (2nd ed., pp. 113-119). London: Pearson Education.

DeMarco, T., & Boehm, B. (2002). The agile methods fray. *IEEE Computer, 35*(6), 90-92.

Emery, J. (2001). Adaptive software development: An experience report. In L. Constantine (Ed.), *Beyond chaos: The expert edge in managing software development* (pp. 273-280). NJ: Addison-Wesley.

Fowler, M., & Highsmith, J. (2001). The agile manifesto. *Software Development, 9*(8), 28-32.

Gandossy, R. (2003). The need for speed. *Journal of Business Strategy, 24*(1), 29-33.

Greening, J. (2001). Launching extreme programming at a process-intensive company. *IEEE Software, 18*(6), 27-33.

Highsmith, J. (2000). *Adaptive software development: A collaborate approach to managing complex system.* New York: Dorset House Publishing.

Highsmith, J. (2001). Opening statement. *Cutter IT Journal, 14*(12), 2-4.

Highsmith, J. (2002). *Agile software development ecosystems*. MA: Addison-Wesley.

Highsmith, J. (2004). *Agile project management*. MA: Pearson Education.

Highsmith, J., & Cockburn, A. (2001). Agile software development: The business of innovation. *IEEE Computer, 34*(9), 120-122.

Jeffries, R. (2001). Card magic for managers: low-tech techniques for design and decisions. In L. Constantine (Ed.), *Beyond chaos: The expert edge in managing software development* (pp. 27-32). NJ: Addison-Wesley.

Kalita, T. (2003). DSDM in process improvement. In J. Stapleton (Ed.), *DSDM business focused development* (2nd ed., pp. 175-191). New York: Pearson Education.

Kirkpatrick, R., Walker, J., & Firth, R. (1992). *Software development risk management: An SEI appraisal*. PA: Software Engineering Institute.

Koch, S. (2004). *Agile principles and open source software development.* Paper presented at the Extreme Programming and Agile Processes in Software Engineering, 5th International Conference, Germany.

Kruchen, P. (2001). Agility with the RUP. *Cutter IT Journal, 14*(12), 27-33.

Lindvall, M., Basili, V., Boehm, B., Costa, P., Dangle, K., Shull, F., et al. (2002). *Emperical findings in agile methods.* Paper presented at the Extreme Programming and Agile Methods—XP/Agile Universe, Chicago.

Manifesto for agile software development. (n.d.). Retrieved from http://agilemanifesto.org

Martin, R. (2003). *Agile software development. Principles, patterns, and practices*. NJ: Prentice Hall.

McAvoy, J., & Sammon, D. (2005). Agile methodology adoption decisions: An innovative approach to teaching and learning. *Journal of Information Systems Education, 16*(4), 409-420.

Pikkarainen, M., & Passoja, U. (2005). *An approach for assessing suitability of agile solutions: A case study.* Paper presented at the Extreme Programming and Agile Processes in Software Engineering XP 2005, Sheffield, UK.

Poole, C., & Huisman, J. (2001). Using extreme programming in a maintenance environment. *IEEE Software, 18*(6), 42-49.

Reifer, D. (2002). *How to get the most out of extreme programming/agile methods.* Paper presented at the Extreme Programming and Agile methods—XP/Agile Universe, Chicago.

Rising, L., & Janoff, S. (2000). The scrum software development process for small teams. *IEEE Software, 17*(4), 26-32.

Sutherland, J. (2001). Agile can scale: Inventing and reinventing SCRUM in five companies. *Cutter IT Journal, 14*(12), 5-11.

Svensson, H., & Host, M. (2005) *Introducing and agile process in a software maintenance and evolution organization.* Paper presented at the Ninth European Conference on Software Maintenance and Reengineering. Manchester, UK. pp. 256-264. IEEE Computer Society.

Young, G. (2003). Implementing DSDM in eBA. In J. Stapleton (Ed.), *DSDM business focused development* (2nd ed., pp. 97-103). London: Pearson Education.

KEY TERMS

Adoption Decision: The decision to adopt a methodology where the various relevant factors are considered and discussed to determine the viability of adoption.

Agile Approach: The primary goal of agile approaches is to ensure adaptability in the ever changing environment that teams operate in, to ensure quick delivery of product to the customer, matching the customer's requirements.

Agile Manifesto: The Agile Manifesto is a list of four values that the different agile methods consider to be their common core values.

Disciplined Approach: These approaches utilise standards and processes to ensure conformance to acceptable standards and best practices. The primary goal is to ensure delivery of software products of high quality.

Life Critical Projects: The criticality of the project is determined by its impact on its users, rather than the impact on the development organization. There are degrees of criticality with life critical at one extreme (life critical implying the potential loss of life through malfunctioning of the software product) or at the other extreme the effect on the comfort of the user (a poorly designed screen causing eyestrain).

Software Development Methodology: A codified set of practices that are adhered to, to ensure a successful project.

Software Development Project: A temporary endeavour undertaking the creation of a unique software solution for a customer.

ENDNOTE

[1] SCRUM is not an abbreviation, it is a rugby term used to depict a strategy meeting which gets an out of play ball, back into play.

Alignment of Business and Knowledge Management Strategy

El-Sayed Aboud Zeid
Concordia University, Canada

INTRODUCTION

The role of knowledge as a crucial asset for and enterprise's survival and advancement has been recognized by several researchers (e.g., von Krogh, Ichijo, & Nonaka, 2000). Moreover, by having knowledge (intellectual resources), an organization can understand how to exploit and develop its traditional resources better than its competitors can, even if some or all of those traditional resources are not unique (Zack, 1999). Therefore, knowledge management (KM-) strategy has to be solidly linked (aligned) to business (B-) strategy in order to create economic value and competitive advantage.

Several authors clearly indicate the importance of mutually aligning business strategy and KM efforts and how this alignment helps enhance organizational performance (e.g., Earl, 2001; Ribbens, 1997). For example, Maier and Remus (2001, 2002, 2003) propose a process-oriented approach that considers market-oriented factors in a KM strategy. In this approach KM strategies can be described according to the process focus and type of business processes supported (Maier & Remus, 2001). The process focus can extend from a single business process to an organization-wide perspective, including all relevant business processes (core and service). The type of process is related to the identification of knowledge-intensive business processes. In addition, Sabherwal and Sabherwal (2003) empirically found that the cumulative abnormal stock market return (in the five-day event window) due to a KM announcement is positively associated with the alignment between the firm's business strategy and the attributes of the KM initiative announced. They use four attributes to characterize KM initiatives: KM level, KM process, KM means, and knowledge source. KM level concerns the hierarchical grouping of individuals upon which the KM effort described in the announcement is focused. The KM processes (or K-manipulating processes) involve the sharing, utilization, or creation of knowledge, while KM means involve organizational structural arrangements and technologies that used to enable KM processes (Earl, 2001; Hansen, Nohria, & Tierney, 1999). Finally, knowledge source reflects from where the knowledge originates.

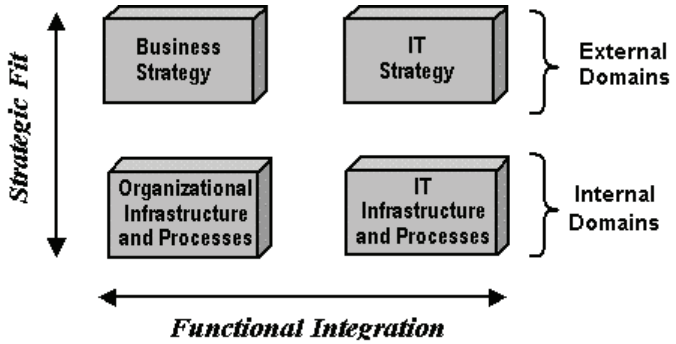
However, realizing the importance of aligning B- and KM-strategies in creating value and in gaining competitive advantage is only the first and the easiest step in any KM initiative. The second and almost as important step is to answer how and where to begin questioning (Earl, 2001). In fact this link has not been widely implemented in practice (see Zack, 1999, and the empirical studies cited there), and "many executives are struggling to articulate the relationship between their organization's competitive strategy and its intellectual resources and capabilities (knowledge)" (Zack, 1999). This is due to the lack of strategic models to link KM-strategy (knowledge [K-] scope, K-systemic competencies, K-governance, K-processes, K-infrastructures, and K-skills) and business strategy. As Zack (1999) argued, they a need pragmatic yet theoretically sound model. It has been highly accepted that a pragmatic and theoretically sound model should meet at least two criteria. First, it should explicitly include the external domains (opportunities/threat) and internal domains (capabilities/arrangements) of both B- and KM-strategies and the relationships between them. Second, it should provide alternative strategic choices.

In order address this issue a "KM strategic alignment model (KMSAM)" is presented. It stems from the premise that the realization of business value gained from KM investment requires alignment between the B- and KM-strategies of the firm and is based on the Henderson-Venkatraman (1993) Strategic Alignment Model for information technology (IT).

OVERVIEW OF THE HENDERSON-VENKARTAMAN STRATEGIC ALIGNMENT MODEL

The KM strategic alignment model is based on the theoretical construct developed by Henderson and Venkatraman (1993). In their model, business success is viewed as the result of the synergy between four domains. The first two, the external domains, are business-strategy and IT strategy. The strategy domains are described in terms of (business/technology) scope, (distinctive business/IT systemic) competencies, and (business/IT) governance. The second two, the internal domains, are organizational infrastructure and processes and IT infrastructure and processes. Both internal domains are described in terms of (administrative/IT) infrastructure, (business/IT) processes, and (business/IT) skills. This synergy is achieved through two types of relationship:

Figure 1. IT Strategic Alignment Model (Henderson & Venkatraman, 1993)



- **Strategic fit:** Emphasizes the need for consistency between strategy (external domain) and its implementation (internal domain).
- **Functional integration:** Has two modes and extends the strategic fit across functional domains. The first mode, *strategic integration*, deals with the capability of IT functionality both to shape and to support business-strategy. The second mode, *operation integration*, focuses on the criticality of ensuring internal coherence between organizational infrastructure and processes and IT infrastructure and processes.

Figure 1 shows the elements of the IT Strategic Alignment Model (ITSAM).

KM STRATEGIC ALIGNMENT MODEL

The premise of the original ITSAM is that "the effective and efficient utilization of IT requires the alignment of IT strategies with business strategies" (Henderson & Venkatraman, 1993). In a parallel way, the premise of KMSAM, in which knowledge strategy replaces IT strategy, is that "the effective and efficient use of organizational knowledge requires the alignment of knowledge strategies with business strategies." Since strategy, whether B-strategy or K-strategy, can be seen as a balancing act between the *external domain* (opportunities/threats) and the *internal domain* (capabilities/arrangements) of the firm (strengths and weaknesses) (Henderson & Venkatraman, 1993; Zack, 1999), the external and internal domains of K-strategy have first to be defined.

K-Strategy External Domain

In the case of K-strategy, the *external domain* involves three dimensions: *K-scope* (what the firm must know), *K-Systemic competencies* (what are the critical characteristics of the required knowledge) and *K-governance* (how to obtain the required K-competencies). The first dimension, K-scope, deals with the specific domains of knowledge that are critical to the firm's survival and advancement strategies. Survival strategies aim at securing current enterprise profitability, while advancement strategies aim for future profitability (von Krogh et al., 2000).

Determining the K-scope can be achieved by constructing a B-domain/K-thing matrix that documents the current and required state of organizational knowledge concerning some or all business domains. The first group of elements that constitutes this matrix includes the list of B-domains (B_i). The second group of elements includes the K-things (K_j) that describe the current state of knowledge associated with each of the relevant B-domains. To relate this knowledge to enterprise business-strategies, K-things are further classified according to the roles they play in such strategies. Von Krogh et al. (2000) have suggested that there are two types of strategies: survival and advancement. Survival strategies aim at securing current enterprise profitability, while advancement strategies aim for future profitability. Therefore, organizational knowledge, and consequently K-things, is classified into two categories: survival (K_S) and advancement (K_A). Figure 2 shows the generic form of this matrix.

The second dimension of the K-strategy external domain is K-systemic competencies. The focus of this dimension is the set of utilization-oriented characteristics of knowledge that could contribute positively to the creation of new business-strategy or better support of existing business-strategy. This set includes characteristics such as:

- **Accessibility:** The extent to which organizational knowledge is made available to its members regardless of time or location (Buckman, 1998).
- **Transferability:** The extent to which the newly acquired knowledge can be applied in other contexts, for example, organizational and cultural (Grant, 1996).

Figure 2. The generic form of B-things/K-things matrix (Abou-Zeid, 2002)

Survival Knowledge				**Advancement Knowledge**			
B_1	K_{S11} (Current/Required States)	….	K_{S1n} (Current/Required States)	K_{A11} (Current/Required States)	….	K_{A1m} (Current/Required States)	
B_2	K_{S21} (Current/Required States)	….	K_{S2k} (Current/Required States)	K_{A21} (Current/Required States)	….	K_{A2l} (Current/Required States)	
….	….	….	….	….	….	….	
B_N	K_{SN1} (Current/Required States)	….	K_{SNk} (Current/Required States)	K_{AN1} (Current/Required States)	….	K_{ANl} (Current/Required States)	

- **Appropriability:** The extent to which knowledge can be imitated. Things are said to have "strong" appropriability if they are difficult to reproduce by another organization. The converse is "weak" appropriability. A related concept is that of "sticky/slippery," that is, sticky knowledge is such an integral part of a regime that it cannot be extracted in a meaningful whole (Grant, 1996; Narasimha, 2000).
- **Depth and Breadth** (Narasimha, 2000).
- **Compositionality:** The amenability of knowledge to be synthesized from existing knowledge.
- **Integrateability:** The extent to which the newly acquired knowledge can be integrated with existing knowledge.

Finally, K-governance dimension deals with the selection and use of mechanisms for obtaining the required K-competencies. The following are examples of some "acquisition mechanisms" (Probst, Raub, & Romhardt, 2000):

- Bringing experts to the firm by recruiting specialists as full-time or temporary staff. Temporary hiring is becoming an increasingly interesting alternative.
- Tapping knowledge held by other firms through different inter-organizational co-operation forms such as joint ventures or strategic alliances.
- Utilizing the knowledge of stakeholders, for example, customers, suppliers, employees, and owners. For example, involving customers early in the product-development process could generate valuable information about their needs.
- Acquiring knowledge products such as software, patents, and CD-ROMs.

K-Strategy Internal Domain

In the case of K-strategy, the internal domain involves three dimensions: *K-processes, K-infrastructures,* and *K-skills*.

K-processes, the first dimension of the K-strategy internal domain, can be classified into two main categories: K-manipulating processes and K-enabling processes. The first category, K-manipulating processes, includes all the organizational processes needed to change the state of organizational knowledge such as K-generation, K-mobilization, and K-application (Abou-Zeid, 2003). The second category, K-enabling processes, includes organizational processes that support K-manipulating processes such as managing conversation, mobilizing knowledge activists, creating the right context, and globalizing local knowledge (von Krogh et al., 2000).

Organizational knowledge processes are socially interaction-intensive. They involve social interactions and direct communication and contact among individuals and among members of "communities of practice." Therefore, they require the presence of social capital. Social capital is "the sum of actual and potential resources embedded within, available through, and derived from the network of relationships possessed by a social unit" (Nahapier & Ghoshal, 1998). Recognizing the importance of social capital, Gold, Malhotra, and Segars (2001) have identified three key K-infrastructures, the second dimension of the K-strategy internal domain—technical, structural, and cultural—that enable social capital. The *K-technical infrastructure* includes IT-enabled technologies that support KM activities such as business intelligence, collaboration and distributed learning, K-discovery, K-mapping, opportunity generation, and security. The *K-structural infrastructure* refers to the presence of enabling formal organization structures and the organization's system of rewards and incentives. Finally, the *K-cultural infrastructure* involves elements such as corporate vision and the organization's system of values (Gold et al., 2001).

The last dimension of the K-strategy internal domain is K-skills. KM processes are by their very nature multifaceted. They involve many dimensions such as technical, organizational, and human. This characteristic of KM processes reflects on the nature of skills required to perform them. For example, Malhotra (1997) defines a senior knowledge executive, such as a chief knowledge officer (CKO) or an organizational knowledge architect, as the person who

Figure 3. The dynamics of the strategic alignment process

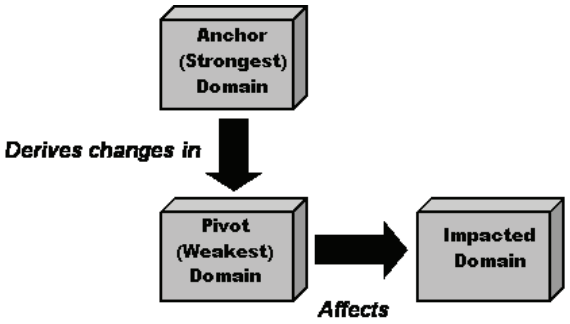


Figure 4. Knowledge potential perspective

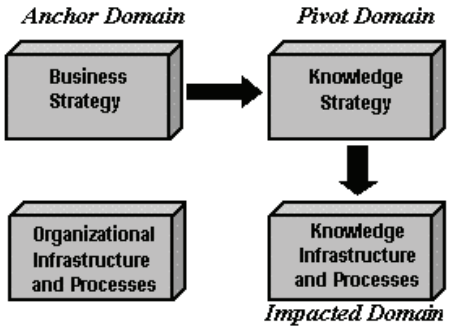
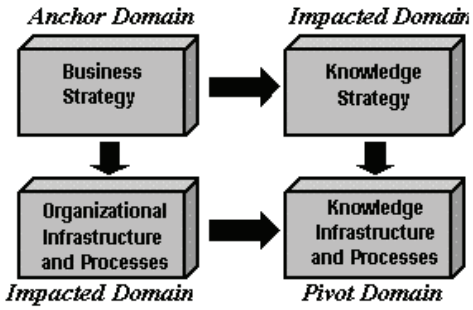


Figure 5. K-infrastructure fusion perspective



should have the combined capabilities of a business strategist, technology analyst, and a human resource professional. The ability to facilitate the ongoing process of knowledge sharing and knowledge renewal, the ability to develop the human and cultural infrastructure that facilitates information sharing, and the ability to utilize the available technologies for serving the creation, sharing, and documentation of knowledge are some examples of the required skills.

The Dynamics of KM Strategic Alignment Model

Effecting a change in any single domain may require the use of three out of the four domains to assure that both strategic fit and functional integration are properly addressed. Therefore, applying KMSAM requires the identification of three domains: pivot, anchor, and impacted (Luftman, 1996). The pivot domain is the weakest and offers the greatest opportunity for improvement. The anchor domain is the strongest and will be the driver of change. Finally, the impacted domain is the area affected by a change to the pivot domain. Figure 3 shows the dynamics of the strategic alignment process.

Based on this distinction, different perspectives of strategic alignment can be identified. Each perspective represents a pattern of linkages between at least three elements of the four elements of KMSAM, that is, the two external domains (business-strategy and knowledge-strategy) and the two internal domains (organizational infrastructure and processes and knowledge infrastructure and processes). By identifying the strongest (anchor) domain and the *adjacent* weakest (pivot) domain, it becomes possible to identify the area that will be affected by the changes (the impacted domain). The direction the perspective flows is based on which domain is the strongest and which is the weakest.

For example, Figure 4 shows knowledge potential perspective in which business-strategy, the strongest domain, derives changes to the adjacent weakest domain, knowledge-strategy, and these changes will impact knowledge infrastructure and processes. In general, each alignment perspective has to include two types of relationship. The first is between external and internal domains of its business and knowledge components, that is, strategic fit. The second is the functional integration between business and knowledge domains. Eight single-path alignment perspectives can be then identified, namely form anchor domain to adjacent pivot domain to impacted domain.

When the pivot and the anchor domains are not adjacent to one another, but rather across from each other on the diagonal, there will be two possible "paths" from the anchor domain to the pivot domain. This yields four fusion perspectives that result from fusing two of the eight single-path perspectives (Luftman, 1996). For example, Figure 5 shows K-infrastructure fusion perspective in which business-strategy derives changes to the K-infrastructure and processes domain through organizational infrastructure and processes, and K-strategy domains.

Table 1 summarizes the 12 alignment perspectives.

CONCLUSION

Based on the premise that the realization of business value from KM investments requires alignment between the busi-

Table 1. KM strategic alignment perspectives

	Domain / Strategic Perspective	Anchor Domain	Pivot Domain	Impacted Domain
1	**Strategy Execution**	B-strategy	Organizational infrastructure and processes	K-infrastructure and processes
2	**Knowledge Potential**	B-strategy	K-strategy	K-infrastructure and processes
3	**Competitive Potential**	K-strategy	B-strategy	Organizational infrastructure and processes
4	**Service Level**	K-strategy	K-infrastructure and processes	Organizational infrastructure and processes
5	**K-/Organizational Infrastructure**	K-infrastructure and processes	Organizational infrastructure and processes	B-strategy
6	**K-Infrastructure/K- Strategy**	K-infrastructure and processes	K-strategy	B-strategy
7	**Organizational/K- Infrastructure**	Organizational infrastructure and processes	K-infrastructure	K-strategy
8	**Organizational Infrastructure/B-Strategy**	Organizational infrastructure and processes	B-strategy	K-strategy
9	**K-Infrastructure Fusion** (Perspectives 4 + 7)	B-strategy	K-infrastructure and processes	▪ Organizational infrastructure and processes ▪ K-strategy
10	**Organizational Infrastructure Fusion** (Perspectives 1+ 5)	K-strategy	Organizational infrastructure and processes	▪ B-strategy ▪ K-infrastructure and processes
11	**B-Strategy Fusion** (Perspectives 3+ 8)	K-infrastructure and processes	B-strategy	▪ Organizational infrastructure ▪ K-strategy
12	**K-Strategy Fusion** (Perspectives 2+ 6)	Organizational infrastructure and processes	K-strategy	▪ B-strategy ▪ K-infrastructure and processes

ness- and knowledge-strategies and on the IT strategic alignment model (SAM) developed by Henderson and Venkatraman (1993), a KMSAM is developed. Moreover, it provides executives with a logical framework for analyzing and assessing alternative strategic choices with regard to aligning K-strategy and B-strategy.

Extension of this work would move in two directions. The first would be to use KMSAM in cross-sectional study of KM initiatives in order to identify the dominant patterns of K-strategy and B-strategy alignment. As "strategic alignment is not an event but a process of continuous adaptation and change" (Henderson & Venkatraman, 1993), the second direction would be a longitudinal study of each enterprise cycle around the alignment perspectives and how the adopted perspective is related to the degree of maturity of the KM initiative.

REFERENCES

Abou-Zeid, E. (2002). A knowledge management reference model. *Journal of Knowledge Management, 6*(5), 486-499.

Abou-Zeid, E. (2003). Developing business aligned knowledge management strategy. In E. Coakes (Ed.), *Knowledge management: Current issues and challenges* (pp. 156-172). Hershey, PA: IRM Press.

Buckman, R. (1998). *Lions, tigers and bears: Following the road from command and control to knowledge sharing*. Retrieved September 10, 2004, from http://www.knowledge-nurture.com/

Earl, M. (2001). Knowledge management strategies: Toward a taxonomies. *Journal of Management Information Systems, 18*(1), 215-233.

Gold, A., Malhotra, A., & Segars, A. (2001). Knowledge management: An organizational capabilities perspective. *Journal of Management Information Systems, 18*(1), 185-214.

Grant, R. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal, 17*(Winter Special Issue), 109-112.

Hansen, M., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? *Harvard Business Review, 77*(2), 106-119.

Henderson, J., & Venkatraman, N. (1993). Strategic alignment: Leveraging information technology for transforming organization. *IBM Systems Journal, 32*(1), 4-16.

Luftman, J. (1996). Applying the strategic alignment model. In J. Luftman (Ed.), *Competing in the information age* (pp. 43-69): Oxford University Press.

Maier, R., & Remus, U. (2001). *Towards a framework for knowledge management strategies: Process-orientation as a new strategic starting point.* Paper presented at the 34th Annual Hawaii International Conference on System Sciences (HICSS-34).

Maier, R., & Remus, U. (2002). Defining process-oriented knowledge management strategies. *Knowledge and Process Management, 9*(2), 109-118.

Maier, R., & Remus, U. (2003). Implementing process-oriented knowledge management strategies. *Journal of Knowledge Management, 7*(4), 62-74.

Malhotra, Y. (1997). Profile of the ideal knowledge manager/architect. Retrieved from http://www.brint.com/wwwboard/messages/273.html

Nahapier, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review, 23*, 242-266.

Narasimha, S. (2000). Organizational knowledge, human resource management and sustained competitive advantage: Toward a framework. *CR, 10*(1), 123-135.

Probst, G., Raub, S., & Romhardt, K. (2000). *Managing knowledge: Building block for success*. John Wiley.

Ribbens, B.A. (1997). Organizational learning styles: Categorizing strategic predispositions from learning. *International Journal of Organizational Analysis, 5*(1), 59-73.

Sabherwal, R., & Sabherwal, S. (2003). *How do knowledge management announcements affect firm value? A study of firms pursuing different business strategies*. University of Missouri.

von Krogh, G., Ichijo, K., & Nonaka, I. (2000). *Enabling knowledge creation: How to unlock the mystery of tacit knowledge and release the power of innovation*. Oxford University Press.

Zack, M. H. (1999). Developing knowledge strategy. *California Management Review, 41*(3), 125-145.

KEY TERMS

Anchor Domain: The area that provides (drives, catalyzes, or enables) the change forces applied to the pivot domain.

Impacted Domain: The area affected by a change to the pivot domain.

K-Systemic Competencies: The set of utilization-oriented characteristics of knowledge that could contribute positively to the creation of new business strategy or better support of existing business strategy.

Pivot Domain: The problem or opportunity being addressed.

Alignment with Sound Relationships and SLA Support

AC Leonard
University of Pretoria, South Africa

INTRODUCTION

International data corporation surveyed 283 top executives across three vertical industries: finance, manufacturing, and retail/wholesale. They found "a strong correlation between the effectiveness of the IT department (IS organization) and the relationship between the CIO and the CEO." "We suspect that this relationship, if it is close, permits the CIO to develop the IT department (IS organization) into a service that delivers competitive advantage for the company, thus enhancing the careers of every IT professional in the organization." In other words, "a certain amount of mutual esteem will help IT (IS) function as a business partner."

In terms of alignment, sound relationships between IT and the business become even more important. Boar (1994) states that aligning with anything other than the customer leads to momentary success. For the IT function to achieve a state of alignment with the business, it must align with the business scope, and through that business scope enable all business functions and processes to serve the customers in a superior manner.

BACKGROUND

In their research, Reich and Benbasit (1999) point out that there are two dimensions to strategy creation: the intellectual dimension and the social dimension. Research into the intellectual dimension is more likely to concentrate on the contents of plans and on planning methodologies. Research into the social dimension is more likely to focus on the people involved in the creation of alignment. The social dimension of alignment is defined as "the state in which business and IT executives within an organizational unit understand and are committed to the business and IT mission, objectives, and plans."

Another theoretical perspective supporting the concept of the social dimension of alignment is the social construction of reality. This view would suggest that, in addition to studying artefacts (such as plans and structures) to predict the presence or absence of alignment, one should investigate the contents of the players' minds: their beliefs, attitudes, and understanding of these artefacts.

This article focuses on the social dimension in terms of the construction and nature of sound IT-end user relationships and the role such relationships play in aligning IT with the business. Research in this field has shown that relationships between IT professionals and their end users are intriguing and complex, and should be seen and managed as a multi-dimensional environment. Furthermore, the supportive role of service level agreements (SLA's) in this regard is also highlighted.

IT END-USER RELATIONSHIPS: HISTORICAL FOUNDATIONS

For many years, the *culture gap* between IT departments and their end users has been characterized by unfortunate differences like distrust, scepticism, and cynicism. This situation impacts negatively on the relationship of IT departments with their end users, and as such on their ability to produce service and support of high quality.

Historically, the gap was caused mainly by the difference in management culture, as well as human behaviour problems on both sides. Umbaugh (1991) states in his argumentation of organizational imbalances that too often IT exists as an adjunct to the organization and not as an integral part of the whole. This situation unfortunately still exists today and contributes to the so-called *culture gap* between IT departments and their end users. Du Plooy (1995) explains this gap as follows:

...the 'culture gap' should be understood as a gap of misunderstanding in the sense of two different organizational 'cultures' that, according to Grindley, coexist in most organizations. The two cultures under discussion here are the 'culture' of the IT profession and the 'culture' of the rest of the organization.

The culture on both the IT department and the business side is also an important obstacle in building mutual trust, and eventually in building sound relationships between IT and its end-user environment, and as such in creating alignment between IT and the business. According to Moad (1994), the IT professional has been fighting for recognition and relevance at the CEO level for the last 25 years. He gives

many examples illustrating the kind of culture that exists, which could be described as the main reason for misunderstandings and misconceptions about IT amongst today's end users.

THE NATURE OF IT-END USER RELATIONSHIPS

The preceding paragraphs briefly describe the history of how poor relationships emerged over the years between IT departments and their end users, as well as some basic characteristics of such poor relationships. The question one can ask is, what are the characteristics of sound relationships between IT departments and their end users, and how are they established? To answer the question, this section gives a definition of IT-end user relationships and briefly discusses the nature of the different elements.

A relationship between an IT professional and an end user consists of two dimensions, namely a physical dimension and an abstract dimension. The physical dimension describes those elements that are necessary in order to enable contact between IT professional and its end users, whereas the abstract dimension describes the soft issues of such a relationship. These two dimensions enable one to fully describe the holistic nature of such a relationship and encapsulate the important elements of a support-oriented organization, namely mutuality, belonging, and connection, as mentioned by Pheysey (1993) in her book *Organizational Cultures*.

Without going into all the details of the different elements of the physical and abstract dimensions as described by Leonard (2002), the article focuses on describing the most important characteristics of these elements. This will give the reader enough understanding of the social nature of IT-end user relationships.

Physical Elements

As far as the physical dimension is concerned, the following elements could be seen as the most important:

- **People:** A relationship consists of all the responsible people who are involved in the systems development life cycle at a given time. "Responsibilities are negotiated and shared between systems developers, and users" (Dahlbom & Mathiassen, 1993).
- **Technology:** Technology may be seen as one of the most important elements in such a relationship, enabling the people who participate in the relationship to communicate with one another. The importance of proper communication structures, both vertically and horizontally, are emphasized by Bommer, Gratto, Gravander, and Tuttle (1991) and could be seen as one of the most important organizational characteristics associated with unethical activity.
- **Procedures:** Two types of procedures are of importance, namely organizational procedures (such as standards and policies), which already exist and which can be seen as a given, and new procedures that are being created by people because of their interaction with the given procedures and technology (DeSanctis & Poole, 1994).
- **Structures:** Depending upon the "type" of end user, and therefore the service and support that will be offered, relationships will differ in content as far as formal and informal social communication structures are concerned. The most common of these structures are project meetings, JAD sessions, and end-user group meetings.

Abstract Elements

As far as the abstract dimension is concerned, the following elements are the most important:

- **They are dynamic:** The nature of the relationships between the IT department and its end users will, among other things, depend upon the type of end user, as well as upon regarding the end user as a human being. According to Stokes (1991), when talking to end users, the IT professional should always bear in mind their concerns, problems, environment, and responsibilities in terms of opportunities for IT services and support.
  Furthermore, he says, continuous contact with end users gives IT the opportunity to gain more insight into their problems.
- **They are sensitive to change:** Because of the social nature of relationships, any form of change initiated on either the IT or the end-user side may disturb a relationship. It is argued that any kind of change having an effect on any of the elements of both the physical and abstract dimensions of a relationship will in fact disturb the relationship because of its holistic nature, which will be described later.
- **They have a knowledge base:** The complex world of perceptions, attitudes, and approaches toward developing software products by IT professionals for the end user forces us to a point where it can be said that in order to overcome the most serious problems during this communication process in a relation-ship, a knowledge base of some kind is required before entering a relationship.
- **They have a supportive culture:** In order for a relationship to be sound, continuous support and mutual understanding, among other things, need to be elements of such a relationship. According to Pheysey,

a support-oriented organization has the elements of mutuality, belonging, and connection. Furthermore, an appreciative form of control should be applied, which means: "management is seen to be a process focused on maintaining balance in a field of relationships" (Pheysey, 1993).

- **A cooperative behaviour pattern is followed by the participants:** Cooperation is not a fixed pattern of behaviour, but a changing, adaptive process directed to future results. The representation and understanding of intent by every party is therefore essential to cooperation, and as such emphasizes the importance of communication during cooperation (Clarke & Smyth, 1993).
  Cooperation can also create new motives, attitudes, values, and capabilities in the cooperating parties, which will help create and maintain a supportive culture.
- **They have a holistic nature:** The important elements making up a relationship between the IT department and its end users at a given time should be organized together as a whole. If any of these elements are disturbed in a negative sense, the whole relationship between the IT department and its end users is undermined.
- **Sustainability:** A most obvious characteristic of the abstract dimension is its sustainability over a period of time. In this regard, time refers to the life span of an IT-end user relationship. One can therefore argue that from an information systems viewpoint, a relationship of this kind will only last until the product or service reaches the end of its life cycle.
  In this regard, Introna (1994) states: "Structures as relationships are contingent, it appears and disappears. It could be brief (a few seconds) or long lasting (several years)."
- **Commitment:** Kinlaw (1989) states that one of the primary tasks of a manager is to create commitment and focus on employees. He furthermore states that managers who help employees increase their knowledge, skill, and experience also are building employee commitment. In this regard it is important that managers should take note of the four sturdy supports of commitment, namely: (a) clarity of goals and values; (b) employee competencies that ensure success; (c) the degree of influence that employees have; and (d) the expressed appreciation given to employees for their contributions. Commitment should be seen as a solid block that rests on these four supports or legs (Kinlaw (1989)).

All the elements described previously form important sub-dimensions of the physical and abstract dimensions. Each of these elements plays a specific social role in an IT/ end-user relationship environment, which impacts on the soundness of such a relationship as well as the success of alignment between IT and the business. The way in which the application of this paradigm could enhance alignment is addressed in the following paragraphs.

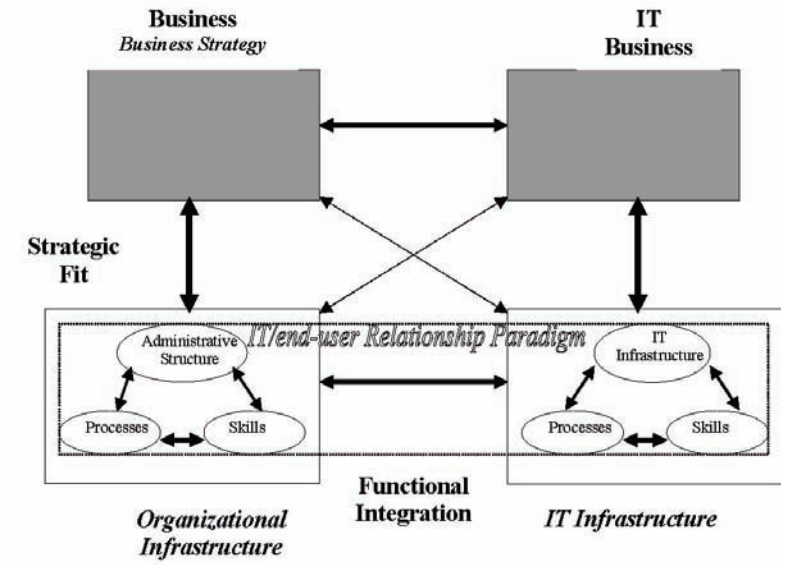
ALIGNMENT MODEL FOR APPLYING THE IT END-USER RELATIONSHIP PARADIGM

The theoretical construct of strategic alignment (Henderson & Venkatraman, 1992) indicates that in terms of alignment there are two distinct linkages, namely a strategic fit and functional integration. According to the model, strategic fit is the vertical linkage concerned with the integration of the external environment in which the firm competes (e.g., partnerships and alliances) and the internal environment, which focuses on administrative structure (e.g., human resources and product development). Functional integration, according to the model, is the corresponding horizontal link between business and IT. These two linkages are used to determine the relation-ships between IT and business (Papp, 2001).

It is clear that the paradigm of IT-end user relationships, which is based on two dimensions, namely the physical and abstract dimensions (as described earlier), addresses the two lower domains indicated by the dotted rectangle in Figure 1. In other words, the paradigm enhances alignment in terms of organizational infrastructure and processes, and IT infrastructure and processes. The physical dimension addresses structures, skills, and processes while the abstract dimension addresses all the soft issues required to ensure that relationships prevail. Therefore, it is argued that if the paradigm of IT end-user relationships is applied when service and support activities[1] are performed by IT professionals, it will enhance the functional integration between IT and the business. This is the case because all the elements of the physical and abstract dimensions are of a sound nature, which directly impacts on structures, processes, and skills in the infrastructure domains of the alignment model in Figure 1.

According to Larson (1998), a service level agreement (SLA) is a formal contract between the IT service provider and the business unit within an organization. The SLA provides a common understanding of the quality of service that the IT service provider will provide. It also helps create reasonable expectations among end users at a specific business unit. As Coye (2004) states it, end users have expectations regarding services and support and the quality thereof provided by the supplier; they compare their expectations to the received service to assess the service quality. Apart from defining the standard of service quality and setting customer expectation, the SLA outlines the role of the end users and the role of the service provider. Therefore, an SLA will enable the end users to be fully aware of the service delivery capabilities

Figure 1. The role and impact of IT-end user relationships in the alignment of business and IT departments (based on the work of Henderson et al., 1992)



and limitations of the service provider, while the service provider will understand the expectation and IT service needs of the end users. This common understanding about the service delivery between the IT service provider and end users is an important component for establishing a successful IT service provider-end user relationship as indicated by Smith (1996). With this in mind, it is argued that SLA's could be seen as the supportive basis for the establishment and maintenance of sound relationships because it provides a continuous "reminder" for all role players of what is expected in terms of service and support. Furthermore, sound IT-end user relationships will enable end users to have a highly effective commitment with the service provider. End users who have such commitment are less likely to switch to a new service provider (Mattila, 2004). Therefore, there is a large possibility that the end users will continue using the service of the current service provider or develop new SLA's with them instead of searching for a new provider in case of poor services or support. To conclude this argument, it is important to take note of IOMA's report on customer relationship management which says that customer relationship commitment will never stand if staff members are focused on "how many" customers they can handle in stead of quality service (IOMA (2004)).

Once there are common interests and mutual understanding between the end users and IT service providers, it will enable the IT service provider to include the services and quality of service needed by end users in the draft SLA. According to Pratt (2003), the SLA will only be of value if the IT service provider has a clear understanding of the end user organization's core business operations and business needs.

FUTURE TRENDS

Theories in terms of how relationships between an IT department and its end users could be managed are scarce. Those who do address issues in this regard (Beard & Peterson, 1988; CSC research foundation, 1994; Wike et al., 1984) do not look into soft issues, nor do they give substance to the contents of such relationships. Furthermore, none of these theories deal with the important issue of how to understand and manage the soft issues involved during the establishment and maintenance of sound IT-end user relationships.

Managing relationships and alignment is especially critical in the outsourcing of information technology services or extending of business applications to systems in other enterprises in order to form extended enterprises. It is clear that in the future much more research needs to go into the management of relationships and service level agreements to ensure that sound relationships and alignment will be maintained.

CONCLUSION

In this article, the paradigm of IT-end user relationships was defined in terms of its physical and abstract dimensions. It was argued that these two dimensions enable one to fully

describe the holistic nature of such relationships. Although the construction and use of SLA's is not something new in terms of service and support, in this article it is given a new dimension by which it serves as a supportive basis for the establishment and maintenance of sound relationships. Furthermore, in terms of business and IT alignment, it was argued that IT-end user relationships should be applied in the infrastructure domains (both IT and the business) and will therefore enhance alignment in terms of functional integration.

REFERENCES

Beard, J. W., & Peterson, T. O. (1988). A taxonomy for the study of human factors in management information systems (MIS). In M. J. Carey (Ed.), *Human factors in management information systems*. Ablex Publishing Corporation, USA.

Boar, B. H. (1994). *Practical steps for aligning information technology with business strategies: How to achieve a competitive advantage.* New York: John Wiley & Sons, Inc.

Bommer, M., Gratto, C., Gravander, J., & Tuttle, M. (1991). A behaviour model of ethical and unethical decision making. In R. Dejoie, G. Fowler, & D. Paradice (Eds.), *Ethical issues in information systems.* SanFrancisco: Boyd & Fraser Publishing.

Clarke, A. A., & Smyth, M. G. G. (1993). A co-operative computer based on the principles of human cooperation. *International Journal of Man-machine Studies, 38,* 3-22.

Coye, R. W. (2004). Managing customer expectations in the service encounter. *International Journal of Service Management, 15*(1), 54-71.

CSC Foundation. (1994). *Future roles and responsibilities for the IS department*. Final Report 96.

Dahlbom, B., & Mathiassen, L. (1993). *Computers in context, the philosophy, and practice of systems design*. Cambridge, UK: Blackwell Publishers.

DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science, 5*(2).

Du Plooy, N. F. (1995). *Overcoming the culture gap between management and IT staff*. Paper presented at Conference on HR Management of IT staff, IEC, Jan Smuts.

Jackson, I. F. (1986). *Corporate information management*. London: Prentice-Hall.

Henderson, J. C., & Venkatraman, N. (1992). Strategic alignment: A model for organizational transformation through information technology. In T. A. Kochan & M. Useem (Eds.), *Transforming organizations*. New York: Oxford University Press.

Introna, L. D. (1994). *Giddens, emergence, and social intervention.* Paper presented at the International Conference on Systems Thinking and Progressive Social Change, University of Cape Town, South Africa.

IOMA'S Report on Customer Relationship. (2004). Does your CRM program need more "relationship?" *4*(9), 13 -15.

Kinlaw, D. C. (1989). *Coaching for commitment: Managerial strategies for obtaining superior performance*. San Diego, CA: Pfeiffer.

Larson, K. D. (1998). The role of service level agreements in IT service delivery. *Information Management & Computer Security, 6*(3), 128-132.

Leonard, A. C. (2002). A conceptual framework for managing relationships between all participants during IT service and support activities. *South African Journal of Industrial Engineering*, 81-96.

Mattila, A. S. (2004). The impact of service failures on customer loyalty. *International Journal of Service Management, 15*(2), 134-149.

Moad, J. (1994). Does your CEO get it? *Datamation, 40*(18), 59-61.

Newman, M., & Sabherwal, R. (1996, March). Determinants of commitment to information systems development: A longitudinal investigation. *MIS Quarterly,* 23-54.

Papp, R. (2001). Introduction to strategic alignment. In R. Papp (Ed.), *Strategic information technology: Opportunities for competitive advantage.* Hershey, PA: Idea Group Publishing.

Pheysey, D. C. (1993). *Organizational cultures*. New York: Routledge.

Pratt, K. T. (2003). Introducing a service level culture. *Facilities, 21*(11), 253-259.

Smith, R. (1996). Business continuity planning and service level agreements. *Information Management & Computer Security, 3*(3), 17-19.

Umbaugh, R. E. (1991). *Handbook of IS management* (3rd ed.). Boston: Auerbach Publishers.

Reich, B. H., & Benbasit, I. (1999). Factors that influence the social dimension of alignment between business and information *technology objectives.* Society of Information Management (SIM) and the Management Information Systems Research Center (MISRC).

Stokes, S. L., Jr. (1991). A marketing orientation for end-user computing support. In R. E. Umbaugh (Ed.), *Hand-book of IS management* (3rd ed.) (pp. 125-134). Boston and New York: Auerbach Publishers.

Wike, W. R., & Andersen, A. (1984). Service management. In *Proceedeings of the CMG XV International Conference on the management and performance of computer systems* (pp. 534-540).

KEY TERMS

Abstract Dimension: The abstract dimension describes the soft issues of such a relationship.

Commitment: A state of mind that holds people and organizations in the line of behaviour. It encompasses psychological forces that bind an individual to an action.

Culture Gap: It should be seen as a gap of misunderstanding in the sense of two different organizational cultures that coexist in most organizations. The two cultures under discussion here are the culture of the IT profession and the culture of the rest of the organization.

Holistic Nature of an IT-End User Relationship: The important elements making up a relationship between an IT professional and its end user(s) at a given time should be organized together as a whole. If any of these elements are disturbed in a negative sense, the whole relationship between the IT professional the end user(s) is under-mined. In other words, the relationship as a whole is more than the sum of its elements.

IT-End User Relationship: A relationship between IT and the end user consists of two dimensions, namely a physical dimension and an abstract dimension. The physical dimension describes those elements that are necessary in order to enable contact between IT and its end users, whereas the abstract dimension describes the soft issues of a relationship.

Physical Dimension: The physical dimension de-scribes those elements that are necessary in order to enable contact between IT professional and its end users.

Social Dimension of an IT-End User Relationship: Refers to all the elements in the abstract dimensions. Each of these elements plays a specific social role in an IT-end user relationship environment, which impacts on the soundness of such a relationship as well as the success of alignment between IT and the business.

ENDNOTE

[1] Normal system development activities or any other types of support IT can give its end users.

Ambient Intelligence in Perspective

Caroline Byrne
Institute of Technology Carlow, Ireland

Michael O'Grady
University College Dublin, Ireland

Gregory O'Hare
University College Dublin, Ireland

INTRODUCTION

Ambient intelligence (AmI) is a relatively new and distinct interpretation of the mobile computing paradigm. However, its recognition that embedded intelligence, either in actuality or perception, is an essential prerequisite if mobile computing is to realize its potential distinguishes it from other mobile usage paradigms. Though stressing the need for intelligence, and implicitly the adoption of artificial intelligence (AI) techniques, AmI does not formally ratify any particular approach and is thus technique agnostic. In this article, we examine the constituent technologies of AmI and provide a brief overview of some exemplary AmI projects. In particular, the question of intelligence is considered and some strategies for incorporating intelligence into AmI applications and services are proposed. It is the authors hope that a mature understanding of the issues involved will aid software professionals in the design and implementation of AmI applications.

BACKGROUND

In 2001, the EU Information Society Technologies Advisory Group (ISTAG) launched a report that proceeded to define the term Ambient Intelligence (ISTAG, 2001). Over a decade earlier, the late Mark Weiser had defined his vision for ubiquitous computing (Weiser, 1991). This vision was far ahead of its time but has been perceived by computer scientists as a vision worth pursuing. As the various technological hurdles were being progressively overcome, ISTAG recognised the inevitability of ubiquitous, pervasive technologies being widely deployed. In practice, this would mean entire generations growing, learning, working and relaxing in an environment saturated with smart sensors and other embedded artifacts. However, a key problem was identified: how to facilitate intuitive interaction with the prevailing embedded technologies. In particular, the scale of these interactions could potentially give rise to situations where numerous artifacts would be clamouring for the individual's attention. Given that human attention is a scarce and precious resource, this course of action could have undesired consequences, and a situation could be envisaged arising where a user might perceive environments saturated with embedded technologies as being places best avoided. Hence, the objective of AmI is to facilitate seamless intuitive interaction between users and their environment.

CONSTITUENT TECHNOLOGIES FOR AMBIENT INTELLIGENCE

Ambient intelligence (AmI) (Aarts & Marzano, 2003; Vasilakos & Pedrycz, 2006) has evolved conceptually and practically, resulting in a common agreement on its core constituent technologies. Three technologies have been identified as being essential to AmI: ubiquitous computing, ubiquitous communications and intelligent user interfaces.

Ubiquitous Computing

Ubiquitous computing envisages the embedding of computational artifacts in the physical environment and their subsequent intuitive access by users. Concerned with the prominence of the then current range of computing systems and their unwieldy interaction modalities, Weiser hoped that ubiquitous computing would herald in an era of what he termed *calm technology*. However, before this could take place, significant advances would have to take place in a number of computing disciplines. One area of particular interest is that of smart environments, as such environments seek to deliver a practical realisation of the ubiquitous computing vision in everyday scenarios, including the home and office. Integration of microprocessors into people's everyday living space objects, such as furniture, clothing, toys and so on, allows the immediate living space to become sensitive and responsive to its inhabitants, rather than just remaining inanimate. Hence, the origin of the term ubiquitous, which

implies that something exists or is everywhere within a living environment on a constant level. A concept closely associated with ubiquitous computing is that of context (Dourish, 2004). In ubiquitous computing, and indeed, other computer usage paradigms, it is envisaged that a model of the user and their environment is available, thus enabling the delivery of services to users that have been dynamically adapted according to the user's current context. Here, context may entail such factors as temporal information, elements of their individual profile (sex, languages spoken, etc.) and current location. In the latter case, absolute positioning, for example, geographic coordinates, or relative positioning, for example, west of a certain landmark, could be used, depending on the nature of the service in question. From a software perspective, the continuous process of capturing context and interpreting it is computationally expensive, and significant scope exists for incorporating intelligent techniques. Such techniques may be used to incorporate reasoning about incomplete knowledge, or perhaps infer future user behavior based on past experiences.

Ubiquitous Communications

Computing technology is increasingly pervasive in everyday life, though under a number of guises, for example, cellular phones and standard embedded household electronics. With continually decreasing hardware costs, relentless miniaturisation and the adoption of high speed data networks, this trend is likely to continue. For example, modern automobiles already contain dozens of microprocessors, while the increasing popularity of Third Generation (3G) mobile phones means that mobile computing is now within reach of people in all facets of their daily lives. Indeed, the widespread deployment of wireless technologies has ensured that mobile computing is spawning a dominant new culture (Rheingold, 2002), as encounters with people using their cellular phones, PDAs, MP3 players, digital cameras and so on is a regular occurrence. Traditionally, these islands of technology would have existed in isolation. However, AmI takes a more holistic view and demands that the existence of a multiprotocol communications infrastructure for integrating disparate technologies such that a unison between all electronic data and equipment pertinent to the user's immediate context can be achieved. Thus, ubiquitous communications seeks to enable embedded objects to communicate with each other and the user by means of various fixed, wireless and ad-hoc networking techniques.

Intelligent User Interfaces

Intelligent user interfaces (IUIs) form the third, and arguably the most important, component of AmI. Human attention is a precious resource. Thus, it must be used judiciously when available, and should be requested cautiously and prudently. It is this prominence given to the need for new innovative interface technologies that distinguishes AmI from other mobile computing paradigms, and also motivates the need for intelligence. The urgent need for IUIs is seen when the nature of the AmI environment is considered. A multitude of embedded or smart artifacts all competing for explicit user attention does not constitute a usable system. Thus, for the sanity of the user, alternative techniques must be considered. Three issues are of particular interest: the need for new physical interface technologies, support for more sophisticated interaction modalities, and the need for collaboration between the constituent components of the environment such that intelligent behaviour can be utilised to the maximum.

Traditional graphical user interfaces (GUIs) are well understood from an ergonomic and implementation perspective. Such GUIs comprise devices such as a keyboard, mouse, and visual display unit; however, at present, the large ambient space that surrounds the user is unused. Thus, AmI demands more sophisticated interaction modalities that utilises this space but which may be difficult to implement and interpret. Examples of such modalities include voice, handwriting, gestures and gaze. The situation is exacerbated when it is considered that a mixture of modalities, for example, voice and gesture, may be used. Interpreting these modalities may require the availability of significant computational resources which may give rise to serious difficulty. A complementary strategy may involve the deployment of embedded dedicated sensors that aid the AmI environment monitor the user, develop sophisticated behaviour models, and enables it to proactively pre-empt user requests. However, no single interaction modality or interface technology can achieve this on its own. Rather, a degree of collaboration and intelligence is necessary to harness and interpret the data before a decision can be made as to whether an explicit interaction session with the user should be initiated.

REALIZING PRACTICAL AMBIENT INTELLIGENT ENVIRONMENTS

A number of organisations, in an effort to understand how AmI environments would work in practice, have designed and developed realistic test environments in which AmI scenarios can be quickly prototyped and evaluated. Philips HomeLab (Aarts & Eggen, 2002), launched in 2002, is one well known example. In essence this lab is a real home, modeled on a two level, two bedroom home. Volunteers live in this house 24 hours a day and their interactions with various electronic devices are observed by researchers. In this way, it is hoped to gain a deeper understanding of how people interact with technology in the home, resulting in

better products and an improved product development cycle. Other initiatives include the Aware House at Georgia Tech (Kidd, Orr, Abowd, Atkeson, Essa, MacIntyre et al., 1999), the MIT Oxygen project (Rudolph, 2001) and the Fraunhofer inHaus initiative (Miller, 2001).

Though homes and offices are logical places for deploying AmI solutions, there are many others. Key obstacles hindering the widespread saturation of AmI in the home environment are economics and the possibility of security breaches. Within business organisations, the obstacles may include a resistance to new work practices or certain legal and ethical difficulties. From a human perspective, it may be anticipated that cultural issues will affect the deployment of AmI solutions. A reticence is frequently observed in consumers when adopting a new technology, and is understandable particularly in the case of the elderly where the overwhelming desire is for affairs to remain constant. Though reluctance to embrace new technologies is not universal, nevertheless, it should not be underestimated by proponents of new technology.

Ambient Assisted Living

Ambient assisted living (AAL) seeks to aid people, particularly the elderly, live independently for a longer time period than would otherwise be possible. The motivation behind AAL is a consequence of the demographical profile of western countries in particular. In short, the number of elderly people as a portion of the population will grow significantly in the next twenty years for many countries. This is a societal issue and one for which there is no easy long term solution. The cost of supporting this increasingly aged population will grow dramatically; a cost that may have to be borne by a smaller workforce. Anticipating this problem, the EU via its Sixth Framework initiative has actively funding projects that seek to enable older people live independently for as long as possible. This initiative, AAL, envisages the use of AmI technologies in particular as a means of achieving this goal.

As just one exemplary illustration of how AAL might operate in practice, the case of a bathroom monitoring system (Chen, Zhang, Kam, & Shue, 2005) is now considered. The system uses acoustics to monitor and classify bathroom activities. At the end of each day, a personal hygiene behavioural report is compiled, for the benefit of care-givers and clinicians, respectively. A number of issues of interest to AmI researchers are raised here. Firstly, no dedicated expensive infrastructure is required. Rather, the only technologies required are an infrared sensor and a microphone. A sophisticated algorithm based on Hidden Markov Models classifies the sound events in real time, thus enabling the raising of alarms quickly. Finally, no significant usability issues arise as the system is quintessentially unobtrusive. Its passive and noninvasive nature increases the possibility of it being accepted by users.

Designing Ambient Intelligent Solutions

Given the importance that AmI ascribes to artificial intelligence, the question of how to realize this intelligence in practice arises. Recall that AmI envisages a world saturated with small computational artifacts. Such artifacts may be mobile, as exemplified by the average mobile telephone or PDA, or may be static, as exemplified by embedded sensors. For the purposes of this discussion, it is assumed that a data communications network is available, without specifying topology or protocol. Three approaches for delivering the necessary intelligence functionality may be adopted.

1. **Networked AI:** In this scenario, the AI software may be deployed on fixed nodes in the network. The advantage of this approach is that the software engineer is not limited by computational constraints that characterise mobile devices and embedded technologies. Thus, computationally expensive solutions, including neural networks and machine learning techniques, can be employed. However, the realisation of this approach is dependent on the availability of a sophisticated communications infrastructure. Recall that implicit in the AmI concept is the need for continuous monitoring of the state of the environment. Any changes to the status quo need to be relayed over the network to the appropriate node where the implications of this change can be gauged and any necessary feedback returned to the embedded artifact. This process needs to take place in near real-time conditions if the net usability of the system is not to be adversely affected.
2. **Embedded AI:** Embedding AI augmented software on the mobile device or embedded artifact is, conceptually, a preferred solution in some, though not all, circumstances. Historically, this approach would have been computationally intractable. Ongoing developments in mobile technologies, however, are rendering this approach feasible, and certain AI technologies may now be incorporated on mobile devices, and, increasingly, on embedded computational artifacts such as wireless sensor networks. At present, the possibility of deploying intelligent agents on such devices is progressing and has already been described in the literature (Carabelea & Boissier, 2003). In general, the approach that has been taken to extend well known agent toolkits such that the runtime engine operates on the mobile device. Of course, most of these toolkits come with additional tools for design and debugging, but these are not ported onto the device. It is only necessary

to deploy a runtime interpreter for the agent logic. One disadvantage of the embedded approach per se is that such applications may tend to be insular. This contravenes the AmI vision in practice as an implicit collaborative effort is necessary in AmI if explicit user interaction is to be kept to a minimum.

3. **Distributed AI:** A third approach to realising intelligence in AmI environments is to distribute the AI technologies between the embedded computational artifact and a fixed server node. This is a flexible solution, and for suitably equipped application domains, may offer the software engineer significant opportunities for mixing and matching the techniques needed for realizing the application in question. Interestingly, distributed AI (DAI) (O'Hare & Jennings, 1996) is itself a distinct and mature sub-branch of AI, and is primarily concerned with issues such as distributed reasoning and knowledge management. Thus, software engineers can call on a wealth of research in the course of their projects. Indeed, there seems to be an uncanny synthesis between the DAI discipline and the AmI concept. More importantly, one of the more advanced implementations of DAI concerns intelligent agents and Multi-agent Systems (MAS).

Intelligent Agents in Computationally Restricted Environments

Intelligent agents are a distinct and viable mechanism for the design and implementation of applications. Agents are usually adopted in highly complex and dynamic situations where traditional software techniques do not seem to encapsulate the necessary constructs essential for effectively modeling the situation. A number of characteristics are synonymous with agents including autonomy, proactivity, reactivity, mobility and sociability (Wooldridge & Jennings, 1995). To what degree each MAS implementation incorporates these characteristics varies. From an AI perspective, agents are viewed as being endowed with a sophisticated reasoning ability. One well-known example of this is the Belief-Desire-Intention (BDI) architecture (Rao & Geogeff, 1995). It should of course be noted that the characteristics supported in fixed networked environments may not necessarily be supported in the runtime engine for embedded devices, and is an issue that prospective developers of agent-based systems should be acutely aware of.

A number of mature MAS environments have been extended with runtime engines for devices of limited computational capacity. All are described extensively in the research literature and most are available in open source format. Examples of interest include 3APL-M (Koch, 2005), LEAP (Adorni, Bergenti, Poggi & Rimassa, 2001) and MicroFIPA-OS (Laukkanen, Tarkoma & Leinonen, 2001). In the case of the BDI architecture, JACK (http://www.agent-software.com), a commercial product, and Agent Factory Micro Edition (Muldoon, O'Hare, Collier & O'Grady, 2006), a derivative of Agent Factory (Collier, O'Hare, Lowen & Rooney, 2003), have also been ported to mobile devices.

FUTURE TRENDS

The potential of AmI is undisputed, but researchers and commercial companies need to be acutely aware of pertinent issues that must be addressed prior to unleashing AmI technology on the general population. These include the issue of privacy and its protection, economics, and the need for basic fault-tolerant networks that are continually alert and can react quickly to changing circumstances. Above all, AmI systems must maintain the privacy of its clients and not compromise their safety in any way. This will be a critical concern of clients when considering whether the advantages outweigh the economic burden. As a priority, companies developing AmI technologies need to emphasise the security benefits they will adhere to contractually. Finally, in the critical issue of practically realising intelligence in AmI applications, it can be realistically envisaged that mobile devices and embedded technologies will increase in power and sophistication, thus increasing the range of AI techniques that can be harnessed by the software engineering community.

CONCLUSION

Ambient intelligent environments may well form a significant portion of the next generation of computing systems. Such environments build on a number of state-of-the-art technologies and seek to effectively harness them such that the needs of users in select domains are adequately addressed. It is this user-centric approach to application and service access that distinguishes AmI from other initiatives in the mobile computing sphere, and over time, the incorporation of increasingly sophisticated AI techniques may further spur the adoption of AmI in a broad range of new and challenging domains.

REFERENCES

Aarts, E., & Eggen, B. (Eds.). (2002). *Ambient intelligence in HomeLab*. Eindhoven: Norec.

Aarts, E., & Marzano, S. (Eds.). (2003). *The new everyday: Views on ambient intelligence*. Rotterdam: 010 Publishers.

Adorni, G., Bergenti, F., Poggi, A., & Rimassa, G. (2001). Enabling FIPA agents on small devices. *Lecture notes in computer science* (Vol. 2182, pp. 248-257). London: Springer-Verlag.

Carabelea, C., & Boissier, O. (2003). *Multi-agent platforms on smart devices: Dream or reality.* Retrieved December 6, 2007, from http://www.emse.fr/~carabele/papers/carabelea.soc03.pdf

Chen, J., Zhang, J., Kam, A., & Shue, L. (2005). An automatic acoustic bathroom monitoring system. In *Proceedings of the IEEE International Conference on Circuits and Systems*, (pp. 1750- 1753). California: IEEE.

Collier, R.W., O'Hare, G.M.P., Lowen, T., & Rooney, C.F.B. (2003). Beyond prototyping in the factory of agents. *Lecture notes in computer science* (Vol. 2691, pp. 383-393). Berlin: Springer-Verlag.

Dourish, P. (2004). What we talk about when we talk about context. *Personal & Ubiquitous Computing, 8*, 19-30.

ISTAG. (2001). *Scenarios for ambient intelligence in 2010.* Retrieved December 6, 2007, from ftp://ftp.cordis.europa.eu/pub/ist/docs/istagscenarios2010.pdf

Kidd, C.D., Orr, R.J., Abowd, G.D., Atkeson, C.G., Essa, I.A., MacIntyre, B., et al. (1999). The aware home: A living laboratory for ubiquitous computing research. *Lecture notes in computer science* (Vol. 1670, pp. 191-198). Berlin: Springer-Verlag.

Koch, F. (2005). 3APL-M platform for deliberative agents in mobile devices. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-agent Systems*, (pp. 153-154). New York: ACM Press.

Laukkanen, M., Tarkoma, S., & Leinonen, J. (2001). FIPA-OS agent platform for small-footprint devices. *Lecture notes in computer science* (Vol. 2333, pp. 447-460). Berlin: Springer-Verlag.

Miller, F. (2001). Wired and smart: From the fridge to the bathtub. *Fraunhofer Magazine*, (2),30-32.

Muldoon, C., O'Hare, G.M.P., Collier, R.W., & O'Grady, M.J. (2006). Agent factory micro edition: A framework for ambient applications. *Lecture notes in computer science* (Vol. 3993, pp. 727-734). Berlin: Springer-Verlag.

O'Hare, G.M.P., & Jennings, N.R. (Eds.). (1996). *Foundations of distributed artificial intelligence*. NJ: John Wiley.

Rao, A.S., & Georgeff, M.P. (1995). BDI agents: From theory to practice. In V. Lesser & L. Gasser (Eds.), *Proceedings of the First International Conference on Multi-agent Systems*, (pp. 312-319). CA: MIT Press.

Rheingold, H. (2002). Smart *mobs: The next social revolution*. New York: Perseus.

Rudolph, L. (2001). Project oxygen: Pervasive, Human-centric computing—an initial experience. *Lecture notes in computer science* (Vol. 2068, pp. 1-12). Berlin: Springer-Verlag.

Vasilakos, A., & Pedrycz, W. (Eds.). (2006). *Ambient intelligence, wireless networking, ubiquitous computing.* Norwood: Artech House.

Weiser, M. (1991). The computer for the twenty-first century. *Scientific American, 265*(3), 94-100.

Wooldridge, M., & Jennings, N.R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review, 10*(2), 115-152.

KEY TERMS

Ambient Intelligence: Intelligence embedded in everyday objects and the surrounding environment such that the use of these smart objects is intuitive to the inhabitants of the environment.

Ambient Assisted Living: This concerns the use of ambient intelligent techniques to enable elderly people live independently for as long as possible.

Context-Awareness: The property of a system that allows it to adjust its behaviour based on environmental cues, such as location or user presence or absence.

Human Computer Interaction (HCI): The study of how people and computers interact, combining aspects of computer science, psychology, sociology, aesthetics and ergonomics, as well as many others.

Intelligent Agents: In this article, the word agent refers to software entities which are capable of displaying autonomous, cooperative and flexible behaviour directed toward achieving a set of internal goals or objectives.

Intelligent User Interface: Intelligent User Interfaces harness various techniques from Artificial Intelligence to adapt and configure the interface to an application such that the end-user's experience is more satisfactory.

Ubiquitous Computing: Also known as pervasive computing, this is the study of how computing can be integrated into the environment in a way that makes it easily accessible to users.

Analysis and Modelling of Hierarchical Fuzzy Logic Systems

Masoud Mohammadian
University of Canberra, Australia

INTRODUCTION

Computational intelligence techniques such as neural networks, fuzzy logic, and evolutionary algorithms have been applied successfully in the place of the complex mathematical systems (Cox, 1993; Kosko, 1992). Neural networks and fuzzy logic are active research area (Cox, 1993; Kosko, 1992; Lee, 1990; Mohammadian & Stonier, 1995; Welstead, 1994; Zadeh, 1965). It has been found useful when the process is either difficult to predict or difficult to model by conventional methods. Neural network modelling has numerous practical applications in control, prediction, and inference.

Time series (Ruelle, 1998) are a special form of data where past values in the series may influence future values, based on presence of some underlying deterministic forces. Predictive model use trends cycles in the time series data to make prediction about the future trends in the time series. Predictive models attempt to recognise patterns and trends. Application of liner models to time series found to be inaccurate, and there has been a great interest in nonlinear modelling techniques.

Recently, techniques from computational intelligence fields have been successfully used in the place of the complex mathematical systems for forecasting of time series. These new techniques are capable of responding quickly and efficiently to the uncertainty and ambiguity of the system.

Fuzzy logic and neural network systems (Welstead, 1994) can be trained in an adaptive manner to map past and future values of a time series and thereby, extract hidden structure and relationships governing the data. The systems have been successfully used in the place of the complex mathematical systems, and have numerous practical applications in control, prediction, and inference. They have been found useful when the system is either difficult to predict and/or difficult to model by conventional methods. Fuzzy set theory provides a means for representing uncertainties. The underlying power of fuzzy logic is its ability to represent imprecise values in an understandable form. The majority of fuzzy logic systems, to date, have been static and based upon knowledge derived from imprecise heuristic knowledge of experienced operators, and where applicable, also upon physical laws that governs the dynamics of the process.

Although its application to industrial problems has often produced results superior to classical control, the design procedures are limited by the heuristic rules of the system. It is simply assumed that the rules for the system are readily available or can be obtained. This implicit assumption limits the application of fuzzy logic to the cases of the system with a few parameters. The number of parameters of a system could be large.

Although the the number of fuzzy rules of a system is directly dependant on these parameters. As the number of parameters increase, the number of fuzzy rules of the system grows exponentially.

In fuzzy logic systems, there is a direct relationship between the number of fuzzy sets of input parameters of the system and the size of the fuzzy knowledge base (FKB). Kosko (1992) call this the "Curse of Dimensionallity." The "curse" in this instance is that there is exponential growth in the size of the fuzzy knowledge base (FKB), where k is the number of rules in the FKB, m is the number of fuzzy sets for each input and n is the number of inputs into the fuzzy system.

As the number of fuzzy sets of input parameters increase, the number of rules increases exponentially. There are a number of ways that this exponential growth in the size of the FKB can be contained. The most obvious is to limit the number of inputs that the system is using. However, this may reduce the accuracy of the system, and in many cases, render the system being modelled unusable. Another approach is to reduce the number of fuzzy sets that each input has. Again, this may reduce the accuracy of the system. The number of rules in the FKB can also be trimmed if it is known that some rules are never used. This can be a time-consuming and tedious task, as every rule in the FKB may need to be looked at.

Raju and Zhou (1993), Mohammadian and Kingham (1997), and Mohammadian, Kingham, and Bignall (1998) suggested using a hierarchical fuzzy logic structure for such fuzzy logic systems to overcome this problem. By using hierarchical fuzzy logic systems, the number of fuzzy rules in the system are reduced, thereby, reducing the computational time while maintaining the systems robustness and efficiency. In this chapter, the design and development of a hierarchical fuzzy logic systems using genetic algorithms to model and predict interest rate in Australia is considered. Genetic algorithms are employed as an adaptive method for design and development of hierarchical fuzzy logic systems.

HIERARCHICAL FUZZY LOGIC SYSTEMS

The hierarchical fuzzy logic structure is formed by having the most influential inputs as the system variables in the first level of the hierarchy, the next important inputs in the second layer, and so on. If the hierarchical fuzzy logic structure contains n system input parameters and L number of hierarchical levels with n_i the number of variables contained in the ith level, the total number of rules k is then determined by:

$$k = \sum_{i-1}^{L} m^{n_i} \tag{1}$$

where m is the number of fuzzy sets. This equation means that by using a hierarchical fuzzy logic structure, the number of fuzzy rules for the system is reduced to a linear function of the number of system variables n, instead of an exponential function of n as is the conventional case. The first level of the hierarchy gives an approximate output, which is then modified by the second level rule set, and so on. This is repeated for all succeeding levels of the hierarchy. One problem occurs when it is not known which inputs to the system have more influence than the others. This is the case in many problems. In some case, statistical analysis could be performed on the inputs to determine which ones have more bearing on the system.

INTEGRATED HIERARCHICAL FUZZY LOGIC AND GENETIC ALGORITHMS

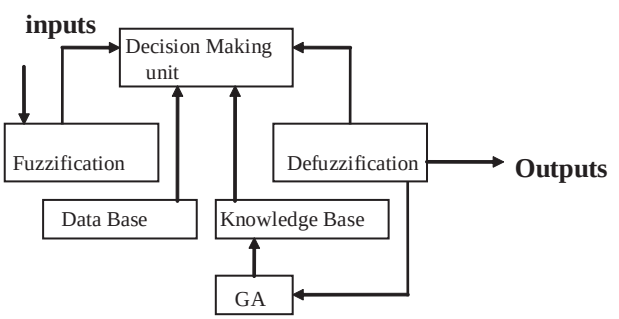
Genetic algorithms (GAs) (Goldberg, 1989; Goonatilake, Campbell, & Ahmad, 1995) are powerful search algorithms based on the mechanism of natural selection, and use operations of reproduction, crossover, and mutation on a population of strings. A set (population) of possible solutions, in this case, a coding of the fuzzy rules of a fuzzy logic system, represented as a string of numbers. New strings are produced every generation by the repetition of a two-step cycle. First, each individual string is decoded and its ability to solve the problem is assessed. Each string is assigned a fitness value, depending on how well it performed. In the second stage, the fittest strings are preferentially chosen for recombination to form the next generation. Recombination involves the selection of two strings, the choice of a crossover point in the string, and the switching of the segments to the right of this point, between the two strings (the cross-over operation). Figure 1 shows the combination of fuzzy logic and genetic algorithms for generating fuzzy rules.

For encoding and decoding of the fuzzy rule for a fuzzy logic system, first the input parameters of the fuzzy logic system is divided into fuzzy sets. Assume that the fuzzy logic system has two inputs α and β and a single output δ. Assume also that the inputs and output of the system is divided into five fuzzy sets. Therefore, a maximum of 25 fuzzy rules can be written for the fuzzy logic system.

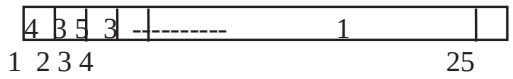
The consequence for each fuzzy rule is determined by genetic evolution. In order to do so, the output fuzzy sets are encoded. It is not necessary to encode the input fuzzy sets because the input fuzzy sets are static and do not change.

The fuzzy rules relating the input variables (α and β) to the output variable (δ) have 25 possible combinations. The consequent of each fuzzy rule can be any one of the five output fuzzy sets. Assume that the output fuzzy sets are **NB** (Negative Big), **NS** (Negative Small), **ZE** (Zero), **PS** (Positive Small), and **PB** (Positive Big). Then the output fuzzy sets are encoded by assigning 1 = **NB** (Negative Big), 2 = **NS** (Negative Small), 3 = **ZE** (Zero), 4 = **PS** (Positive Small),

Figure 1. Combination of fuzzy logic and genetic algorithms for fuzzy rule generation



and 5 = **PB** (Positive Big). Genetic algorithms randomly encode each output fuzzy set into a number ranging from 1 to 5 for all possible combinations of the input fuzzy variables. A string encoded this way can be represented as



Each individual string is then decoded into the output linguistic terms. The set of fuzzy rules thus developed, is evaluated by the fuzzy logic system based upon a fitness value that is specific to the system. At the end of each generation, (two or more) copies of the best performing string from the parent generation is included in the next generation to ensure that the best performing strings are not lost. Genetic algorithms then perform the process of selection, crossover, and mutation on the rest of the individual strings. Selection and crossover are the same as a simple genetic algorithms while the mutation operation is modified. Crossover and mutation take place based on the probability of crossover and mutation respectively. Mutation operator is changed to suit this problem. For mutation, an allele is selected at random, and it is replaced by a random number ranging from 1 to 5. The process of selection, crossover, and mutation are repeated for a number of generations till a satisfactory fuzzy rule base is obtained. We define a satisfactory rule base as one whose fitness value differs from the desired output of the system by a very small value.

HIERARCHICAL FUZZY LOGIC SYSTEM FOR INTEREST RATE PREDICTION

There is a large interest by investors and government departments in the ability to predict future interest rate fluctuations from current economic data. Economists, and investors, have been unable to find all the factors that influence interest rate fluctuations. It is agreed, however, that there are some major economic indicators released by the government that are commonly used to look at the current position of the economy. These indicators used in this chapter are as follows

- *Interest Rate, which is the indicator being predicted. The Interest Rate used here is the Australian Commonwealth government 10-year treasury bonds.*
- *Job Vacancies is where a position is available for immediate filling or for which recruitment action has been taken.*
- *The Unemployment Rate is the percentage of the labour force actively looking for work in the country.*
- *Gross Domestic Product is an average aggregate measure of the value of economic production in a given period.*
- *The Consumer Price Index is a general indicator of the rate of change in prices paid by consumers for goods and services.*
- *Household Saving Ratio is the ratio of household income saved to households disposable income.*
- *Home Loans measure the supply of finance for home loans, not the demand for housing.*
- *Average Weekly Earnings is the average amount of wages that a full time worker takes home before any taxes.*
- *Current Account is the sum of the balances on merchandise trade, services trade, income, and unrequited transfers.*
- *Trade Weighted Index measures changes in our currency relative to the currencies of our main trading partners.*
- *RBA Commodity Price Index provides an early indication of trends in Australia's export Prices.*
- *All Industrial Index provides an indication of price movements on the Australian Stock Market.*
- *Company Profits are defined as net operating profits or losses before income tax.*
- *New Motor Vehicles is the number of new vehicles registered in Australia.*

By creating a system that contained all these indicators, we would be in a much better position to predict the fluctuations in interest rates. A fuzzy logic system that used every indicator and had five fuzzy sets for every indicator would result in a large FKB consisting of over six billion rules! As can be imagined, this would require large computing power to not only train the fuzzy logic system with a genetic algorithm, but also large storage and run-time costs when the system is operational. Even if a computer could adequately handle this large amount of data, there is still the problem in having enough data to properly train every possible rule. To overcome this problem a hierarchical fuzzy logic structure for the fuzzy logic system can be constructed. By using a hierarchical fuzzy logic system, the number of fuzzy rules of the system is reduced, hence, computational times are decreased, resulting in a more efficient system. A novel way to tackle this problem would be to group the relevant indicators and to build a fuzzy knowledge base for each group. The first step is to divide the indicators into smaller-related groups. This problem was investigated in Mohammadian and Kingham (1997) and Mohammadian et al. (1998), and is shown as follows:

1. **Employment** (Job Vacancies, Unemployment Rate**)**
2. **Country** (Gross Domestic Product, Consumer Price Index)
3. **Savings** (Household Saving Ratio, Home Loans, Average Weekly Earnings)

4. **Foreign** (Current Account, RBA Index, Trade Weighted Index)
5. **Company** (All Industrial Index, Company Profit, New Motor Vehicles)

The Interest Rate is included with each of these groups. To learn the fuzzy knowledge base for each group, a genetic algorithm was implemented. The genetic algorithms had a population size of 500 with a crossover rate of 0.6 and a mutation rate of 0.01, and it was run for 10,000 generations over 10 years (a period of 40 quarters) data. Fitness of each string of the genetic algorithm was calculated as the sum of the absolute differences from the predicted quarter and the actual quarters interest rate. The fitness was subtracted from an "optimal" fitness amount, which was decided to be 30, as it was unlikely the error amount would be higher than this over 10 years (Mohammadian & Kingham, 1997; Mohammadian et al., 1998). The fitness of the system is calculated by the following formula:

$$fitness = 30 - \sum_{i=0}^{30} abs(PI_i - I_{i+1}) \quad (2)$$

An elitist strategy was used in that the best population generated was saved and entered in the next generation (two copies of the string with best fitness was included to the next generation). The five fuzzy knowledge bases created form the top layer of the hierarchy are shown in Figure 2. Mohammadian and Kingham (1997) designed and connected together the fuzzy knowledge bases to form a final fuzzy knowledge base system. The final fuzzy knowledge base system, shown in Figure 2, then uses the predicted interest rate from the five listed groups to produce a final interest rate prediction. The number of fuzzy rules for each group is shown in Figure 2.

The final hierarchical FKB contains 3,125 rules (Mohammadian & Kingham, 1997), giving the total number of rules learnt as 5,250. This is a significant reduction from the 6 billion rules that would have been used previously. This allows quicker training time without the need for huge computer resources (Mohammadian & Kingham, 1997). Good prediction of Australian quarterly interest rate can be obtained using this system. The number of fuzzy rules used are also reduced dramatically.

However, there is still a question: Does a two-layer hierarchical architecture provide the best solution?

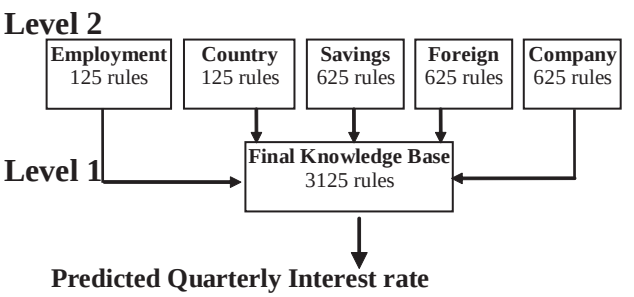
To answer this question, one can start building three, four layer hierarchical fuzzy logic system by trial and error to possibly find the correct number of layers required. This could be a cumbersome problem. We need to know how many layers are required, and which fuzzy knowledge base should be used in each layer. Genetic algorithms can be used to solve this problem by determining the number of layers in the hierarchical fuzzy logic system and the correct combination of FKBs for each layer, see Figure 3.

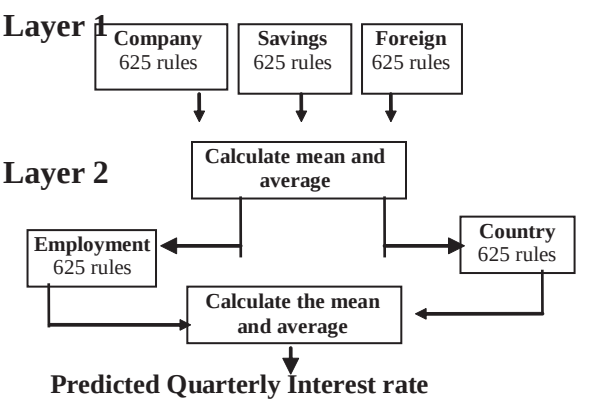
Next, the performance of genetic algorithms for design and development of hierarchical fuzzy logic systems is considered. The system is developed in such a way to provide the possible best architecture for designing hierarchical fuzzy logic systems for prediction of Interest Rate in Australia. Using the economic indicators, five fuzzy logic systems were developed from five groups, each producing a predicted interest rate for the next quarter. Genetic algorithms were then used to design and develop a hierarchical fuzzy logic system.

The hierarchical fuzzy logic system developed was then used to predict interest rate For each of these groups, the

Figure 3. A three-layer hierarchical fuzzy logic system – 3,125 fuzzy rules

Figure 2. Hierarchical fuzzy logic system (Mohammadian & Kingham, 1997; Mohammadian et al., 1998)

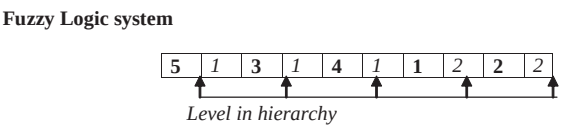




current quarter's interest rate is included in the indicators used (Mohammadian & Kingham, 1997). The advantage of using this hierarchical fuzzy logic structure is that the number of rules used in the knowledge base of fuzzy logic systems has been reduced substantially. For encoding and decoding of the hierarchical fuzzy logic system, first a number is allocated to each fuzzy logic system developed from group of indicators. For this simulation, the number allocated to each group is shown below:

1 = Employment, 2 = Country, 3 = Savings, 4 = Foreign, 5 = Company

The number of layers and the fuzzy logic system/s for each layer is determined by genetic algorithms. In order to do so, a number is allocated to each fuzzy logic system. Genetic algorithms randomly encode each fuzzy logic system into a number ranging from 1 to 5 for all possible combinations of the fuzzy logic systems. The level in the hierarchy in which a fuzzy logic system is allocated to, is also encoded each string. A string is encoded this way can be represented as



Each individual string is then decoded into a hierarchical fuzzy logic system that defines the fuzzy logic system/s for each level of the hierarchical fuzzy logic system. This string, once decoded, will provide a hierarchical fuzzy logic system, as shown in Figure 3. The set of hierarchical fuzzy logic systems, thus developed, are evaluated, and a fitness value is given to each string. At the end of each generation, (two or more), copies of the best performing string from the parent generation are included in the next generation to ensure that the best performing strings are not lost. Genetic algorithms then performs the process of selection, crossover, and mutation on the rest of the individual strings. Crossover and mutation take place based on the probability of crossover and mutation, respectively. Mutation operator is changed to suit this problem. The process of selection, crossover, and mutation are repeated for a number of generations till a satisfactory hierarchical fuzzy logic system is obtained. We define a satisfactory hierarchical fuzzy logic system as one whose fitness value (predicated interest rate) differs from the desired output of the system (in this case the actual interest rate) by a very small value. We calculate the average error of the system for the training set and tests sets using the following formula (Mohammadian & Kingham, 1997):

$$E = \frac{\sum_{i=1}^{n} abs(Pi - Ai)}{n} \tag{3}$$

where E is the average error, *Pi* is the predicted interest rate at time period *i*, *Ai* is the actual interest rate for the quarter, and *n* is the number of quarters predicted. By using genetic algorithms to design and develop hierarchical fuzzy logic system, better results were obtained. The hierarchical fuzzy logic systems developed using genetic algorithms perform predict the interest rate to different degree of accuracy. It is, however, interesting to see that genetic algorithms is capable of providing different hierarchical fuzzy logic system for predicting the interest rate. It is now possible to choose the best hierarchical fuzzy logic system among those suggested by genetic algorithms. The result of the top performing five hierarchical fuzzy logic systems designed by genetic algorithms is given in Table 1. Comparison of average errors of these five best hierarchical fuzzy logic systems designed and developed using genetic algorithms is also shown in Table 1.

Table 1. Average Errors of hierarchical fuzzy logic (HFL) systems designed and developed using GA and Average Errors of hierarchical neural networks by (Mohammadian & Kingham, 1997; Mohammadian et al., 1998)

	Training Error	Testing Error
Hierarchical fuzzy logic #1	0.356	0.659
Hierarchical fuzzy logic #2	0.343	0.663
Hierarchical fuzzy logic #3	0.289	0.494
Hierarchical fuzzy logic #4	0.274	0.441
Hierarchical fuzzy logic #5	0.291	0.398
Hierarchical neural network	0.354	0.607

COMPARISON OF HIERARCHICAL FUZZY LOGIC SYSTEM WITH NEURAL NETWORK SYSTEM FOR PREDICTION OF INTEREST RATE

Mohammadian and Kingham (1997) reported the use of a hierarchical neural network system using the same data inputs as described for the hierarchical fuzzy logic system. Using these economic indicators, a neural network system was created for each five groups (Country Group, Employment Group, Savings Group, Company Group, and Foreign Group). Each neural network system was trained using back-propagation algorithms. A back-propagation neural network was used with two hidden layers, each consisting of 20 neurons, output layer consists of one node. Sigmoid learning was used to predict the following quarters interest rate. The error tolerance was set to 0.0001, the Learning Parameter (Beta) was set to 0.6, momentum (alpha), and Noise Factor were both set to 0. The neural network was trained for 10,000 cycles (Mohammadian & Kingham, 1997, Mohammadian et al., 1998). After training each neural network system for each group, all neural network systems were combined to form a hierarchical neural network system with the same structure as shown in Figure 2. The final neural network system then was trained. It used the prediction of all five neural networks for each group to predict the quarterly interest rate, as its output. Table 1 shows the comparison of the average errors of hierarchical fuzzy logic systems designed and developed using GA and average errors of hierarchical neural networks by Mohammadian and Kingham (1997).

CONCLUSION AND FURTHER INVESTIGATIONS

In this chapter, an innovative method is used to design and develop hierarchical fuzzy logic systems. Genetic algorithms are used as an adaptive learning method to design a hierarchical fuzzy logic systems to predict the quarterly interest rate in Australia. Using a hierarchical fuzzy logic system, the number of fuzzy rules in the fuzzy knowledge base is reduced dramatically, hence, computational times are decreased resulting in a more efficient system. Genetic algorithms are also used to obtain the fuzzy rules for each fuzzy logic system of a hierarchical fuzzy logic system.

From simulation results, it was found that the hierarchical fuzzy logic system is capable of making accurate predictions of the following quarter's interest rate. The prediction result of the top five hierarchical fuzzy logic systems were compared to a hierarchical neural network. It was found that most of the top five hierarchical fuzzy logic system designed by genetic algorithms performed better than the hierarchical neural network. It should be noted that hierarchical neural network was designed based on the intuition (Mohammadian & Kingham, 1997; Mohammadian et al., 1998) and it may be possible to obtain better prediction results using an automated system using genetic algorithms to automatically design the hierarchical neural network system. The research work performed in this chapter is unique in the way the hierarchical fuzzy logic systems are developed. The application of this method to several industrial problems, such as robotic control and collision avoidance of multirobot systems, is currently under consideration. Research is also currently being performed to automatically design a hierarchical neural network system for modeling and prediction.

REFERENCES

Cox, E. (1993). Adaptive fuzzy systems. *IEEE Spectrum, February*, 27-31.

Goldberg, D. (1989).*Genetic algorithms in search, optimisation and machine learning.* Reading, MA: Addison Wesley.

Goonatilake, S., Campbell, J. A., & Ahmad, N. (1995). Genetic-fuzzy systems for financial decision making. Advances in fuzzy logic, neural networks and genetic algorithms. *IEEE/ Nagoya-University World Wisepersons Workshop. Lecture Notes in Artificial Intelligence*. Germany: Springer.

Kosko, B. (1992). *Neural networks and fuzzy systems, a dynamic system*.Englewood Cliff: Prentice-Hall.

Lee, C. C. (1990). Fuzzy logic in control systems: Fuzzy controllers - part I, part II. *IEEE Transactions on Systems, Man and Cybernetics, 2092*, 404-435.

Mohammadian, M., & Kingham. M. (1997). Hierarchical fuzzy logic for financial modelling and prediction. In *Tenth Australian Joint Conference on Artificial Intelligence* (pp 147-156), Perth, Australia.

Mohammadian, M., Kingham, M., & Bignall, B. (1998). Hierarchical fuzzy logic for financial modelling and prediction. *Journal of Computational Intelligence in Finance*.

Mohammadian, M., & Stonier, R. J. (1995). Adaptive two layer control of a mobile robot systems. In *Proceedings of IEEE International Conference on Evolutionary Computing*, Perth, Australia.

Raju, G. V. S., & Zhou, J. (1993). Adaptive hierarchical fuzzy controller. IEEE *Transactions on Systems, Man & Cybernetics, 23*(4), 973-980..

Ruelle. D. (1998). *Chaotic evolution and strange attractors: The statistical analysis of time series for deterministic nonlinear systems*. Cambridge: Uni Press.

Welstead, T. (1994). *Neural networks and fuzzy logic applications in C/C++*. Wiley.

Zadeh, L. (1965). Fuzzy sets. *Inf. Control, 8* 338-353.

Category: IT Security & Ethics

Anonymous Communications in Computer Networks

Marga Nácher
Technical University of Valencia, Spain

Carlos Tavares Calafate
Technical University of Valencia, Spain

Juan-Carlos Cano
Technical University of Valencia, Spain

Pietro Manzoni
Technical University of Valencia, Spain

INTRODUCTION

In our daily life no one questions the necessity of privacy protection. Nevertheless, our privacy is often put at risk. The first problem has to do with the fact that privacy itself is a concept difficult to define. As a matter of fact, in many countries the concept has been confused with data protection, which interprets privacy in terms of the management of personal information. Nowadays, the term *privacy* is extended to territorial and communications protection.

We will focus on the privacy of electronic communications. When referring to this type of communication, the first aspect we think about is security. In fact, this concept is widely discussed, and nowadays we often hear about threats and attacks to networks.

Security attacks are usually split into active and passive attacks. We consider that an active attack takes place when an attacker injects or modifies traffic in the network with different purposes, such as denial of service or gaining unauthorized access. Unlike active attacks, a passive attack takes place whenever the attacker merely inspects the network by listening to packets, never injecting any packet. Malicious nodes hope to be 'invisible' in order to collect as much network information as possible just by using timing analysis and eavesdropping routing information. A way to avoid this type of attack is to anonymize both data and routing traffic. In this manner we can hide the identities of communicating nodes and avoid data flow traceability.

Various scenarios can be devised where anonymity is desirable. In a commercial transactions context, if we think about an off-line purchase, we accept that some users prefer to use cash when buying some goods and services, because anonymity makes them more comfortable with the transaction. Offering anonymity to online commerce would increase the number of transactions.

Military communications are another typical example where not only privacy but also anonymity are crucial for the success of the corresponding mission.

Finally, if we attend a meeting where some delicate matter is being voted on, it could be necessary for the identities to remain hidden. Again, in this case, anonymity is required.

BACKGROUND

In order to talk about anonymity, first we have to establish the terminology to be used. An important work on this issue is Pfitzmann and Hansen (2000); based on this work, we can establish a classification of anonymity degrees: A node is considered exposed when its identity information is known. If its identity is not the real one, the node is pseudonymous. Furthermore, if it is unlinkable to some kind of relevant information, we achieve anonymity with respect to that information; as an example we can consider the relationship between end-to-end peers or the peers themselves. Finally, when the communication is not perceived, we can say that it is undetectable; and if it is undetectable for any external node and also anonymous for every participant, the communication is unobservable.

In the literature, there are various works based on different networks topologies as the Dining Cryptographers (Chaum, 1988) or MIXes (Chaum, 1981) in order to provide anonymous communications in fixed networks.

PEER ANONYMITY

In this article we will discuss the different degrees of anonymity provided by means of different proposals found in the literature, emphasizing those issues that are still unsolved.

General Approaches

Anonymity has been treated differently depending on the network characteristics and goals. We believe that the two most relevant generic proposals are the Dining Cryptographers network and the MIX network.

Dining Cryptographers Network

The Dining Cryptographers network (DC-net) (Chaum, 1988) achieves sender anonymity in the following way: some pairs of participants share a secret bit. Each participant calculates the sum of all the bits that he shares, and if he wants to transmit, he inverts that result. All the nodes send the result of the sum or the inverted one (if necessary). If no one (or an even number of participants) transmits, the sum of all these transmissions is zero. In cases where one participant (or an odd number of them) transmits, the sum will be one. Each participant could share a key of n bits with another participant, one bit per round. So, the ith bit of each such key will be used in the ith round.

However, this approach is restricted to small networks since only one node can transmit in each round. In large networks the probability of having more than one node wishing to transmit in a specific round increases, and so collisions will render this mechanism impractical.

Furthermore, the anonymous bandwidth of a DC-net is limited by the slowest participant. Overall, DC-nets provide strong anonymity elegantly, but suffer from efficiency and scalability problems.

Herbivore (Goel, Robson, Polte, & Sirer, 2003) is a protocol based on DC-nets that tries to solve the scalability problem by splitting the network into sub-groups (called cliques), but it requires global topology control. In terms of efficiency, results are actually quite poor.

MIXes Network

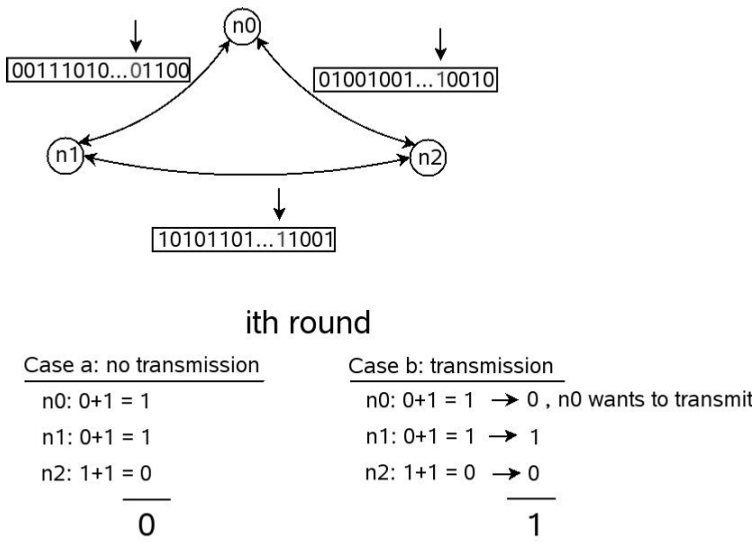
In 1981, Chaum proposed the use of MIXes to anonymize electronic mail users and messages. The main goal for a single MIX is to hide the correlation between incoming and outgoing messages within a large group of messages by delaying or reordering them. In order to do this, encryption and padding mechanisms are applied.

There are several MIX variants:

- A pool MIX only sends part of the incoming messages, keeping the other parts for later rounds. Hence, it uses the reordering technique. It is called a "timed MIX" if the event that triggers the flushing is the expiration of a timeout. In cases where the trigger is the arrival of a message, the MIX will be referred to as a "threshold MIX."
- A stop-and-go MIX (or continuous MIX) delays messages according to an exponential distribution, which does not depend on traffic. Hence, if the number of users is low, the degree of anonymity is also low.

A MIX network can consist of a set of predefined routes, called cascades, or free route networks, where routes are selected by users. Berthold, Pfitzmann, and Standtke (2001) establish that this last type of networks is flexible, scalable, and extendable. However, it is less secure due to the intersection of different anonymity sender/recipient groups,

Figure 1. Example of DC-net



making it easier to reveal participants' identities. Danezis (2003) provides an intermediate solution: Each MIX can only choose routes included in a predetermined graph. The conclusion is that this network is more scalable than a cascade and more resistant to intersection attack than the free route alternative.

In general we can say that MIX-nets always provide relationship anonymity and sometimes also recipient anonymity if their identity remains hidden. However, sender anonymity is more difficult to achieve.

Based on Chaum's MIXes, Tarzan (Freedman & Morris, 2002) provides both sender and recipient anonymity, and therefore relationship anonymity, by using a restricted topology for packet routing. Packets can be routed only between special IP tunnels. Anonymity should be transparent to both client applications and servers.

Tarzan operates at the IP layer and relies on layered encryption: each leg of the tunnel removes or adds a layer of encryption. The tunnels are static and any relay failure requires the formation of a new tunnel, thus increasing both delay and computation overheads.

Instead of single-node MIXes, Cashmere (Zhuang, Zhou, Zhao, & Rowstron, 2005) selects regions (relay groups) as MIXes, providing sender and relationship anonymity. At the same time, any node in a region can act as a MIX, hence reducing the probability of a MIX failure. Each group requires a public/private key pair, which is generated and distributed using an off-line certificate authority (CA). The source randomly selects and orders the relay groups to conceal the destination relay group. It then encrypts the forwarding path in multiple layers using the public keys associated with each relay group. An intermediate node decrypts the message received, forwards it to the next relay group, and broadcasts the decrypted contents to all other members of its own group. The bandwidth cost is higher for this solution than for a node-based relay approach.

Other Protocols

In addition to the solutions already discussed, proposals such as Crowds, Hordes, P5, Tor, and HIP are also relevant in this field of research.

With Crowds (Reiter & Rubin, 1998), the authors provide sender anonymity following this strategy: The source node does not choose the path to be used; instead, it sends the message directly to the Internet with probability p, or forwards the message to another randomly selected user with probability 1-p. The rest of nodes in the network behave in the same way. Therefore, the initiator is indistinguishable from a member that simply forwards a request from another. Once a path is chosen, it remains static for that source-destination pair until the server sends a special message that forces all the established paths to change in order to avoid certain types of attacks. Connections between users are encrypted with the keys distributed by the server.

Shields and Levine (2000) describe the Hordes protocol, which uses a similar strategy to anonymously send a packet from the initiator to the responder. However, it uses multicast communications in the reverse path to reduce the amount of work required from participants and to improve data delivery latency and link utilization. This proposal assumes that shortest path multicast routing trees are available.

The Peer-to-Peer Personal Privacy Protocol (P5) (Sherwood, Bhattacharjee, & Srinivasan, 2002) is another protocol targeting anonymous communications that provides sender and recipient anonymity. It is designed to be implemented in addition to the current Internet protocols without any special infrastructure support. Since it is based on broadcast transmissions, it creates a broadcast hierarchy to avoid scalability problems. This solution has a cost in terms of efficiency; moreover, in mobile environments it would require group management algorithms in order to keep the hierarchy updated. Every node acts as a MIX, scrambling received packets before forwarding them and using hop-by-hop encryption. Also, it maintains a fixed communication rate, sending signal or noise packets only if necessary.

In Dingledine, Mathewson, and Syverson (2004), Tor is proposed as the second-generation onion router. It works on the Internet, and is designed to make TCP-based applications (like Web browsing or instant messaging) anonymous. Clients choose a path through the network and build a circuit, in which each node knows its predecessor and successor, but no other nodes along the path. The length of packets is fixed, and each node unwraps them using a symmetric key. Each onion router maintains a long-term identity key to sign certificates, directories, and the router's descriptor, along with a short-term onion key to decrypt requests from the users to set up a circuit and to negotiate ephemeral keys.

The Host Identity Protocol (HIP) (Moskowitz, Nikander, Jokela, & Henderson, 2007) aims to separate the identifier and locator roles of IP addresses. The base HIP protocol ("base exchange") is used between hosts to establish an IP-layer communications context (called HIP association) prior to communications. This process is based on a Sigma-compliant Diffie-Hellman key exchange (Diffie & Hellman, 1976) with public key identifiers for mutual peer authentication.

The public key of an asymmetric key pair is used as the identifier, named the host identifier (HI). However, a hashed encoding of the HI, usually referred to as the host identity tag (HIT), represents the host identity in protocols due to its short and fixed length (128 bits) and the following three properties: (1) it has the same length as an IPv6 address, and so can be used in address-sized fields in APIs and protocols; (2) it is self-certifying (i.e., given a HIT, it is computationally hard to find a host identity key that matches the HIT); and (3) the probability of HIT collision between two hosts is very low.

Indeed, these properties provide the following advantages: first, its fixed length simplifies the protocol coding and reduces the cost of this technology in terms of packet size. Second, it presents a consistent format to the protocol irrespective of the underlying identity technology used.

Undetectability and Unobservability

The highest degrees of anonymity are undetectability and unobservability.

Referring again to the definitions presented in Pfitzmann and Hansen (2000), and according to the authors: "A mechanism to achieve some kind of anonymity appropriately combined with dummy traffic yields the corresponding kind of unobservability." So we could say, for example, that:

DC-nets + dummy traffic = sender unobservability
MIX-nets + dummy traffic = relationship unobservability

The other two most popular mechanisms used to provide undetectability are steganography and spread spectrum techniques.

The use of power control can also help to reduce the probability of being heard since the attacker must be located very close to the transmitter node. If directional antennas are used, the attacker not only has to stay close, but also inside the corresponding sector to which the antenna is directed.

General approaches to achieve undetectability and unobservability can be found in the literature. In this section we briefly explain some of them that are based on two popular techniques: steganography and dummy traffic.

We focus firstly on the studies related to steganography. The typical approach is based on images where it is easy to introduce hidden information due to the redundancy that this type of message presents, but without forgetting the possibility of compression that would eliminate redundancy and also all the hidden information.

Sender unobservability is provided in Heydt-Benjamin, Serjantov, and Defend (2006) using steganography in a high-latency MIX network, achieving sender/receiver unlinkability as achieved by other MIX networks, while improving sender unobservability. In it, users steganographically embed messages in images which they then post to the most popular Usenet newsgroups. The majority of images in Usenet will not contain stegotext and will serve as cover traffic.

The schema presented in Bo, Jia-zhen, and De-yun (2007) uses the identification field of the IPv4 header to conceal data. According to this proposal the first eight bits of that field will contain data and the next eight bits the order of the packet. Due to the small amount of information transmitted in each packet, several are required to send the whole message. In order to give this field a random appearance, a fourth-order Chebyshev chaotic system is used to generate a sequence from an initial given value. This chaotic sequence is then converted to a binary sequence. Also, a key is shared between source and destination. The message is encrypted with this key. Afterwards the k different encrypted blocks are included in k packets and they are sent to the destination. To ensure the success of the proposal packet, fragmentation along the path must be avoided. To achieve this, the path maximum transfer unit discovery (PMTUD) is enabled. Obviously, the schema is only practical for very short messages.

Ahsan and Kundur (2002) use two strategies to hide information: IPv4 header manipulation (fragment bit and identification field) and packet sorting using the sequence number fields in IPSec (AH or ESP headers).

In the former case they propose the use of the second bit—the DF (Do not Fragment) bit—in the Flags field. In order to send secret information using this bit, both the source and destination have to know the MTU of their network and always build packets that are a smaller size than this MTU to avoid packet fragmentation (that would corrupt the DF bit). The identification field is another possible header field manipulated to hide data. Ahsan and Kundur (2002) use it through chaotic MIXing (toral automorphism systems) to provide a random appearance to the field. The only limitation is that the identification field is unique for a specific source-destination pair as long as the datagram is alive. With regards to the packet sorting, the authors consider that the order of the packets sent is of no concern. So, if there are *n* packets to send, they will have *n!* ways of sending them and the selection of one combination can be interpreted as *log2(n!)* concealed bits. Estimation of hidden information is done using a look-up table to match the stored sequence to the corresponding sequence of packets received, and mapping this sequence to the hidden information. Also, the transmission process is modeled as a non-ideal channel characterized by the position error. The latter is imposed by the network's behavior since the network cannot guarantee sequencing in packet delivery.

With regards to dummy traffic, Diaz and Preneel (2004a) introduced this topic and pointed out the necessity of establishing an appropriate amount of dummy packets depending on the cost of inserting them. Also, the number of dummy packets should depend (or not) on the amount of real traffic. The frequency of generation and the most appropriate route length for these dummy messages are still open questions.

Research presented in Diaz and Preneel (2004b, 2004c) focuses on MIXes networks by trying to determine the best strategy for the insertion of dummy traffic. Such traffic is inserted and removed by MIXes, not users, and two different ways of inserting it are proposed: into the output link at the time of flushing or into the pool of the MIX. In both cases the generation of dummy messages follows a probability distribution independent of the traffic of real messages. With regards to the type of MIXes, the authors compare deterministic and binomial MIXes using random or deterministic dummy policies. They conclude that binomial MIXes,

together with a random dummy policy, provide the greatest level of anonymity. Likewise, in terms of anonymity, it is better to insert dummy traffic at the output link, although latency also increases.

FUTURE TRENDS

In recent times, wireless networks have become important in our daily communications, and so it is necessary to provide them with security mechanisms, and anonymous ones in particular. The proposals for wired networks are not suitable for wireless and mobile communications. Hence, most of the last works for anonymous communications have been specifically designed for MANETs (Kong & Hong, 2003; Zhang, Liu, Lou, & Fang, 2006). However, their performance needs to improve if we want those protocols to be useful and practical.

CONCLUSION

In this article we have analyzed a wide variety of proposals for anonymous communications in wired networks. We believe that the field of anonymous communications is still open to improvements, and no author has yet integrated the different anonymity mechanisms described throughout this article in a consistent and efficient manner. Performance issues also remain largely untackled, and they require more scrutiny before actual deployment can take place.

REFERENCES

Ahsan, K., & Kundur, D. (2002). Practical data hiding in TCP/IP. *Proceedings of the Workshop on Multimedia Security at ACM Multimedia.*

Berthold, O., Pfitzmann, A., & Standtke, R. (2001). The disadvantages of free MIX routes and how to overcome them. *Proceedings of the International Workshop on Designing Privacy Enhancing Technologies: Design Issues in Anonymity and Unobservability.* New York: Springer-Verlag.

Bo, X., Jia-zhen, W., & De-yun, P. (2007). Practical protocol steganography: Hiding data in IP header. *Proceedings of the 1st Asia International Conference on Modeling and Simulation.*

Chaum, D. (1988). The dining cryptographers problem: Unconditional sender and recipient untraceability. *Journal of Cryptology, 1*(1), 65-75.

Chaum, D. (1981). Untraceable electronic mail, return addresses, and digital pseudonyms. *Communications of the ACM, 4*(2).

Danezis, G. (2003). MIX-networks with restricted routes. *Proceedings of the Privacy Enhancing Technologies Workshop* (PET 2003). Berlin: Springer-Verlag (LNCS 2760).

Diaz, C., & Preneel, B. (2004a). Anonymous communication. In *WHOLES: A multiple view of individual privacy in a networked world.*

Diaz, C., & Preneel, B. (2004b). Reasoning about the anonymity provided by pool MIXes that generate dummy traffic. *Proceedings of the Conference on Information Hiding* (IH'04). Berlin: Springer-Verlag (LNCS 3200).

Diaz, C., & Preneel, B. (2004c). Taxonomy of MIXes and dummy traffic. *Proceedings of the Conference on Information Security Management, Education and Privacy* (INetSec'04) (vol. 3, pp. 215-230).

Diffie, W., & Hellman, M.E. (1976). New directions in cryptography. *IEEE Transactions on Information Theory, 22*(6), 644-654.

Dingledine, R., Mathewson, N., & Syverson, P. (2004). Tor: The second-generation onion router. *Proceedings of the 13th USENIX Security Symposium.*

Freedman, M.J., & Morris, R. (2002). Tarzan: A peer-to-peer anonymizing network layer. *Proceedings of the 9th ACM Conference on Computer and Communications Security.*

Goel, S., Robson, M., Polte, M., & Sirer, E.G. (2003). *Herbivore: A scalable and efficient protocol for anonymous communication.* Technical Report 2003-1890, Cornell University, USA.

Heydt-Benjamin, T.S., Serjantov, A., & Defend, B. (2006). Nonesuch: A MIX network with sender unobservability. *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society.*

Kong, J., & Hong X. (2003). ANODR: Anonymous on demand routing with untraceable routes for mobile ad-hoc networks. *Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking & Computing* (MobiHoc'03) (pp. 291-302), New York.

Moskowitz, R., Nikander, P., Jokela, P., & Henderson, T. (2007). *HIP: Host identity protocol.* Retrieved from http://www.ietf.org/internet-drafts/draft-ietf-hipbase-09.txt

Pfitzmann, A., & Hansen, M. (2000). Anonymity, unobservability, and pseudonymity: A proposal for terminology. In H. Federrath (Ed.), *Workshop on design issues in anonymity and unobservability* (pp. 1-9). Berlin: Springer-Verlag (LNCS 2009).

Reiter, M., & Rubin, A. (1998). Crowds: Anonymity for

Web transactions. *ACM Transactions on Information and System Security, 1*(1).

Sherwood, R., Bhattacharjee, B., & Srinivasan, A. (2002). P5: A protocol for scalable anonymous communication. *Proceedings of the IEEE Symposium on Security and Privacy.*

Shields, C., & Levine, B.N. (2000). A protocol for anonymous communication over the internet. *Proceedings of the ACM Conference on Computer and Communications Security.*

Zhang, Y., Liu, W., Lou, W., & Fang, Y. (2006). MASK: Anonymous on-demand routing in mobile ad hoc networks. *IEEE Transactions on Wireless Communications, 21,* 2376-2385.

Zhuang, L., Zhou, F., Zhao, B.Y., & Rowstron, A. (2005). Cashmere: Resilient anonymous routing. *Proceedings of the 2nd Symposium on Networked Systems Design and Implementation* (NSDI), Boston.

KEY TERMS

Anonymity: State of being not identifiable among other items belonging to a set. This set is called *anonymity set*.

Dummy Traffic: Randomly generated packets injected in the network to make the perception of real traffic difficult.

Item of the System: Any participating subject, object, or action: node, user, message, sending, and so forth.

Spread-Spectrum Techniques: Methods by which energy generated with a certain bandwidth is deliberately spread in the frequency domain, resulting in a signal with a wider bandwidth.

Steganography: The art and science of writing hidden messages in such a way that no one apart from the sender and the intended recipient even realizes that there is a hidden message.

Undetectability: Incapability of observing an established communication. Thus, undetectability prevents that third parties can observe when a packet is being sent through the network.

Unlinkability: Incapability of stating the relation between two observed items of the system. For example, recipient unlinkability ensures that the sending of a packet and the corresponding recipient cannot be linked by others.

Unobservability: Undetectability by external attackers plus anonymity for internal attackers.

Untraceability: Property of maintaining routes unknown to either external or internal attackers.

Category: Data Mining and Databases

Ant Colony Algorithms for Data Classification

Alex A. Freitas
University of Kent, UK

Rafael S. Parpinelli
UDESC, Brazil

Heitor S. Lopes
UTFPR, Brazil

INTRODUCTION

Ant colony optimization (ACO) is a relatively new computational intelligence paradigm inspired by the behavior of natural ants (Dorigo & Stutzle, 2004). Ants often find the shortest path between a food source and the nest of the colony without using visual information. In order to exchange information about which path should be followed, ants communicate with each other by means of a chemical substance called pheromone. As ants move, a certain amount of pheromone is dropped on the ground, creating a pheromone trail. The more ants that follow a given trail, the more attractive that trail becomes to be followed by other ants. This process involves a loop of positive feedback, in which the probability that an ant chooses a path is proportional to the number of ants that have already passed by that path.

Hence, individual ants, following very simple rules, interact to produce an intelligent behavior at the higher level of the ant colony. In other words, intelligence is an emergent phenomenon.

In this article we present an overview of Ant-Miner, an ACO algorithm for discovering classification rules in data mining (Parpinelli, Lopes, & Freitas, 2002a, 2002b), as well as a review of several Ant-Miner variations and related ACO algorithms.

All the algorithms reviewed in this article address the classification task of data mining. In this task each case (record) of the data being mined consists of two parts: a goal attribute, whose value is to be predicted, and a set of predictor attributes. The aim is to predict the value of the goal attribute for a case, given the values of the predictor attributes for that case (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

BACKGROUND

An ACO algorithm is essentially a computational system inspired by the behavior of natural ants and designed to solve real-world optimization problems. The basic ideas of ACO algorithms are as follows (Dorigo & Stutzle, 2004):

- Each ant incrementally constructs a candidate solution to a given optimization problem. That candidate solution is associated with a path in a graph representing the search space.
- When an ant follows a path, the amount of pheromone deposited on that path is proportional to the quality of the corresponding candidate solution.
- At each step during the incremental construction of a candidate solution, an ant typically has to choose which solution component should be added to the current partial solution (i.e., how to extend the current path), among several solution components. In general the probability of a given component being chosen is proportional to the product of two terms: (a) the amount of pheromone associated with that component; and (b) the value of a (problem-dependent) heuristic function for that component.

As a result, due to a continuous increase of the pheromone associated with the components of the best candidate solutions considered by the algorithm, the ants usually converge to the optimum or near-optimum solution for the target problem.

The motivation for applying ACO algorithms to the discovery of classification rules and related data mining tasks is as follows. Many projects in the field of data mining proposed deterministic rule induction algorithms. These algorithms typically are greedy, and so they are susceptible to find only locally optimal (rather than globally optimal) classification rules. By contrast, ACO algorithms try to mitigate this drawback using a combination of two basic ideas. First, these algorithms have a stochastic aspect, which helps them to explore a larger area of the search space. Second, they use an iterative adaptation procedure based on positive feedback (the gradual increase of the pheromone associated with the best solution components) to continuously improve candidate rules. Combining these basic ideas, in general ACO algorithms perform a more global search in the space of candidate rules than typical deterministic rule induction algorithms, which makes the former an interesting alternative to be considered in rule induction.

The first ACO algorithm proposed for discovering classification rules was Ant-Miner (Parpinelli et al., 2002a). In Ant-Miner each artificial path constructed by an ant represents a candidate classification rule of the form:

IF <*term1* AND *term2* AND … > THEN < *class* >.

Each term is a triple <*attribute, operator, value*>, where *value* is one of the values belonging to the domain of *attribute*. An example of a term is: <*Sex = female*>. *Class* is the value of the goal attribute predicted by the rule for any case that satisfies all the terms of the rule antecedent. An example of a rule is:

IF <*Salary = high*> AND <*Mortgage = No*> THEN <*Credit = good*>.

In the original version of Ant-Miner, the *operator* is always "=" so that Ant-Miner can cope only with categorical (discrete) attributes. Continuous attributes would have to be discretized in a preprocessing step.

The pseudo code of Ant-Miner is described, at a very high level of abstraction, in Algorithm 1. Ant-Miner starts by initializing the training set to the set of all training cases, and initializing the discovered rule list to an empty list. Then it performs an outer loop where each iteration discovers a classification rule.

The first step of this outer loop is to initialize all trails with the same amount of pheromone, which means that all terms have the same probability of being chosen—by an ant—to incrementally construct a rule. This is done by an inner loop, consisting of three steps. First, an ant starts with an empty rule and incrementally constructs a classification rule by adding one term at a time to the current rule. In this step a $term_{ij}$—representing a triple $<Attribute_i = Value_j>$—is chosen to be added to the current rule with probability proportional to the product of $\eta_{ij} \times \tau_{ij}(t)$, where η_{ij} is the value of a problem-dependent heuristic function for $term_{ij}$ and $\tau_{ij}(t)$ is the amount of pheromone associated with $term_{ij}$ at iteration (time index) t. More precisely, η_{ij} is essentially the information gain associated with $term_{ij}$ (see Cover & Thomas, 1991, for a comprehensive discussion on information gain). The higher the value of η_{ij}, the more relevant for classification $term_{ij}$ is and so the higher its probability of being chosen. $\tau_{ij}(t)$ corresponds to the amount of pheromone currently available in the position i,j of the trail being followed by the current ant. The better the quality of the rule constructed by an ant, the higher the amount of pheromone added to the trail positions ("terms") visited ("used") by the ant. Therefore, as time goes by, the best trail positions to be followed—that is, the best terms to be added to a rule—will have greater and greater amounts of pheromone, increasing their probability of being chosen.

The second step of the inner loop consists of pruning the just-constructed rule—that is, removing irrelevant terms. This is done by using the same rule-quality measure used to update the pheromones of the trails, as defined later. In essence, a term is removed from a rule if this operation does not decrease the quality of the rule.

Algorithm 1. High-Level Pseudo Code of Ant-Miner

```
TrainingSet = {all training cases};
DiscoveredRuleList = []; /* initialized with empty list */
REPEAT
  Initialize all trails with the same amount of pheromone;
  REPEAT
    An ant incrementally constructs a classification rule;
    Prune the just-constructed rule;
    Update the pheromone of all trails;
  UNTIL (stopping criteria)
  Choose the best rule out of all constructed rules;
  Add the chosen rule to DiscoveredRuleList;
  TrainingSet = TrainingSet – {cases correctly covered by
  the chosen rule};
UNTIL (stopping criteria)
```

The third step of the inner loop consists of updating the pheromone of all trails by increasing the pheromone in the trail followed by the ant, proportionally to the quality of the rule. In other words, the higher the quality of the rule, the higher the increase in the pheromone of the terms occurring in the rule antecedent. The quality (Q) of a rule is measured by the equation:

$$Q = \text{Sensitivity} \times \text{Specificity},$$

where $Sensitivity = TP / (TP + FN)$ and $Specificity = TN / (TN + FP)$. The acronyms TP, FN, TN, and FP stand for the number of true positives, false negatives, true negatives, and false positives, respectively (Parpinelli et al., 2002a).

The inner loop is performed until some stopping criterion is satisfied, for example, until a maximum number of candidate rules has been constructed.

Once the inner loop is over, the algorithm chooses the highest-quality rule out of all the rules constructed by all the ants in the inner loop, and then it adds the chosen rule to the discovered rule list. Next, the algorithm removes from the training set all the cases correctly covered by the rule—that is, all cases that satisfy the rule antecedent and have the same class as predicted by the rule consequent. Hence, the next iteration of the outer loop starts with a smaller training set, consisting only of cases which have not been correctly covered by any rule discovered in previous iterations. The outer loop is performed until some stopping criterion is satisfied, for example, until the number of uncovered cases is smaller than a user-specified threshold.

Hence, the output of Ant-Miner is the list of classification rules contained in the discovered rule list.

Parpinelli et al. (2002a, 2002b) have performed computational experiments comparing Ant-Miner with two well-known rule induction algorithms, namely CN2 (Clark & Niblett, 1989) and C4.5 (Quinlan, 1993), in several public-domain data sets. In a nutshell, the results showed that Ant-Miner is competitive with both C4.5 and CN2 concerning predictive accuracy on the test set. However, Ant-Miner discovered rule lists much simpler (i.e., smaller) than the rule sets discovered by C4.5 and the rule lists discovered by CN2. This is an advantage in the context of data mining, where discovered knowledge is supposed to be comprehensible to the user in order to support the user's intelligent decision making (Fayyad et al., 1996).

ANT-MINER VARIATIONS AND RELATED ALGORITHMS

(a) Fixing in Advance the Class Predicted by a Rule

In Ant-Miner, each ant first constructs a rule antecedent. Next, the majority class among all cases covered by the rule is assigned to the rule consequent. The fact that the class predicted by a rule is not known during rule construction has two important consequences. First, the heuristic function is based on reducing the entropy associated with the entire class distribution, rather than using a heuristic function specific to the class to be predicted by the rule. Second, the pheromone associated with each term represents the relevance of that term with respect to the overall discrimination between all classes, rather than with respect to a specific class.

In order to make the heuristic function and pheromone values have a more focused relevance, variations of Ant-Miner have been proposed where all the ants in the population construct rules predicting the same class, so that the class to be assigned to a rule is known during rule construction (Chen, Chen, & He, 2006; Galea & Shen, 2006; Martens, De Backer, M., Haesen, Baesens, & Holvoet, 2006; Smaldon & Freitas, 2006). This naturally leads to new heuristic functions and new pheromone update methods, as discussed in the next two subsections.

(b) New Class-Specific Heuristic Functions

Several Ant-Miner variations have replaced the entropy reduction heuristic function by a simpler heuristic function whose value is specific for each class (Chen et al., 2006; Liu, Abbass, & McKay, 2004; Martens et al., 2006; Oakes, 2004; Smaldon & Freitas, 2006; Wang & Feng, 2004). Typically, in these Ant-Miner variations, the value of the heuristic function for a $term_{ij}$ is based on the relative frequency of the class predicted by the rule among all the cases that have the $term_{ij}$. This modification is particularly recommended when the class predicted by a rule is fixed in advance before rule construction, unlike the situation in the original Ant-Miner, as discussed above.

It has also been argued that a simpler heuristic function based on the relative frequency of the rule's predicted class has the advantage of being computationally more efficient without sacrificing the predictive accuracy of the discovered rules, since hopefully the use of pheromones should compensate for the less accurate estimate of the simpler heuristic function. Note, however, that there is no guarantee that the use of pheromones would completely compensate the use of a less effective heuristic function. A more important issue seems to be whether or not the rule's predicted class is known during rule construction, as discussed earlier. In addition, note that in Ant-Miner the heuristic function values are computed just once in the initialization of the algorithm, and in principle the heuristic function values for all terms can be computed in linear time with respect to the number of cases and attributes (Parpinelli et al., 2002a). Hence, the time complexity of the heuristic function of Ant-Miner does not seem a significant problem in the context of the entire algorithm.

(c) Using the Pseudorandom Proportional Transition Rule

As explained earlier, Ant-Miner adds a $term_{ij}$ to a rule with a probability proportional to the product $\eta_{ij} \times \tau_{ij}(t)$. This kind of transition rule is called the random proportional transition rule (Dorigo & Stutzle, 2004). Several variants of Ant-Miner instead use a pseudorandom proportional transition rule (Chen et al., 2006; Liu et al., 2004; Wang & Feng, 2004). In this transition rule, in essence, there is a probability q_0 that the term to be added to the current partial rule will be deterministically chosen as the $term_{ij}$ with the greatest value of the product of $\eta_{ij} \times \tau_{ij}(t)$; and there is a probability $1 - q_0$ that the term to be added to the rule will be probabilistically chosen by the random proportional rule. This transition rule has the advantage that it allows the user to have explicit control over the exploitation vs. exploration trade-off (Dorigo & Stutzle, 2004), which of course comes with the need to choose a good value for the parameter q_0—normally chosen in an empirical way.

(d) New Rule Quality Measures

Ant-Miner's rule quality is based on the product of sensitivity and specificity. Chen et al. (2006) and Martens et al. (2006) proposed to replace Ant-Miner's rule quality measure with

measures that are essentially based on the confidence and coverage of a rule. Coverage is equivalent to sensitivity, so the main idea introduced in these Ant-Miner variations is to replace specificity by confidence in the rule quality measure. Given the importance of rule quality—on which pheromone updating is based—it would be interesting to perform extensive experiments comparing the effectiveness of these two kinds of rule quality measures—something missing in the literature.

(e) New Pheromone Updating Procedures

Several variants of Ant-Miner use an explicit pheromone evaporation rate—this is a predefined parameter in Liu et al. (2004) and Martens et al. (2006) and a self-adaptive parameter in Wang and Feng (2004). Note that this new parameter is not necessary in the original Ant-Miner, where pheromone evaporation is implicitly performed by normalizing all pheromone values after increasing the pheromone of the terms used in the just-constructed rule.

In addition, Liu et al. (2004), Wang and Feng (2004), and Smaldon and Freitas (2006) proposed different equations or procedures to update pheromone as a function of rule quality. These new approaches address the issue that Ant-Miner's original equation for pheromone updating does not work well when the quality of a rule is close to zero. Note, however, that in the original Ant-Miner, the value of the rule quality measure will be close to zero only in rare situations. This is due to the fact that the best possible class for the current rule is always chosen to be added to the rule consequent after the rule antecedent has been fully constructed, which should result in a rule of at least reasonable quality in most cases. On the other hand, in Ant-Miner variants—where the class of all constructed rules is predetermined (see the discussion in the first subsection above) rules with very low quality will be less rare, especially in early iterations of the algorithm, before the pheromone of good terms has been significantly increased. In such Ant-Miner variants, it is indeed important to change the original Ant-Miner equation to update pheromone in order to cope better with low-quality rules.

(f) Coping with Categorical Attributes Having Ordered Values

Ant-Miner copes only with categorical (nominal or discrete) attributes. Hence, continuous (real-valued) attributes have to be discretized in a preprocessing step. Some categorical attributes have unordered values (e.g., the attribute "gender" can take the values "male" or "female"), while other categorical attributes have ordered values (e.g., the attribute "number of children" could take the values "0", "1", "2", "3", or "more than 3"). Ant-Miner does not recognize this distinction. It produces only terms (rule conditions) of the form "attribute = value" using the "=" operator.

Both Oakes (2004) and Martens et al. (2006) proposed Ant-Miner variants that can cope with ordered values of categorical (but not continuous) attributes. These variants can discover rules containing terms of the form "attribute op value" where *op* can be a relational comparison operator such as "<", ">", "≤", or "≥". This allows the discovery of simpler (smaller) rule sets. For instance, a single condition like "number of children < 3" is shorter than the disjunction of conditions: "number of children = 0" or "number of children = 1" or "number of children = 2."

(g) Dropping the Rule Pruning Procedure

Martens et al. (2006) proposed an Ant-Miner variant (Ant-Miner+) that does not perform rule pruning. In order to mitigate the need for pruning, the domain of each attribute was extended with a dummy value, hereafter called the "any" value. Hence, for each attribute i with k values, there is a set of k+1 terms that can be chosen to be added to the current partial rule, where the $term_{i,k+1}$ is the "any" value for attribute i. Adding an "any" value term to a rule means that the attribute in question can have any value when the rule is evaluated, which effectively means that attribute is not present in the rule antecedent. The authors argue that, using this approach, pruning is superfluous because the dummy values already lead to shorter rules. The removal of pruning has the advantage of making Ant-Miner+ significantly faster than Ant-Miner, since rule pruning tends to be the most computational expensive and least scalable part of Ant-Miner (Parpinelli et al., 2002a).

On the other hand, from the point of view of predictive accuracy of the discovered rules, it is not very clear if rule pruning is really superfluous due to the use of dummy "any" values to make rules shorter. In the original Ant-Miner, rule pruning is performed by a deterministic procedure that iteratively removes one term at a time from the rule as long as rule quality is improved. This kind of local search not only reduces the discovered rules' size, but also tends to improve the predictive accuracy of those rules (Parpinelli et al., 2002a). It would be interesting to evaluate if, despite the relatively short size of the rules discovered by Ant-Miner+, the predictive accuracy of those rules could be increased by the use of a deterministic rule pruning procedure driven by rule quality.

(h) Discovering Fuzzy Classification Rules

Galea and Shen (2006) proposed a new ACO algorithm—called FRANTIC-SRL (Fuzzy Rules from ANT-Inspired Computation-Simultaneous Rule Learning)—that discovers

fuzzy classification rules, rather than the crisp rules discovered by Ant-Miner. FRANTIC-SRL maintains multiple populations of ants, where each population is in charge of discovering rules predicting a different class. Hence, the class predicted by a rule is fixed in advance for all ants in each of the populations — see item (a).

Each ant constructs a fuzzy rule by adding one linguistic term (a fuzzy condition) at a time to the current partial rule. This involved replacing the heuristic function of Ant-Miner by a function based on fuzzy systems' theory. The new function is class-specific—that is, the heuristic function of a term depends on the class predicted by the rule– see item (b). At each iteration, instead of evaluating each candidate rule separately, FRANTIC-SRL evaluates each candidate rule set containing one rule from each of the ant colony populations—that is, each candidate rule set containing one rule for each class. Hence, at each iteration the number of evaluated rule sets is $numAnts^{numClasses}$, where *numAnts* is the number of ants in each population and *numClasses* is the number of classes. To speed up this step, the number of evaluated rule sets could be reduced by considering only combinations of the best rules (rather than all rules) from each population, as pointed out by the authors.

(i) Discovering Rules for MultiLabel Classification

Multilabel classification involves the simultaneous prediction of the value of two or more class attributes, rather than just one class attribute as in conventional classification. Chan and Freitas (2006) proposed a major extension of Ant-Miner to cope with multi-label classification, called MuLAM (Multilabel Ant-Miner). In MuLAM, each ant constructs a set of rules—rather than a single rule—where different rules predict different class attributes. A rule can predict a single class attribute or multiple class attributes, depending on which option will lead to better rules for the data being mined. MuLAM uses a pheromone matrix for each of the class attributes. Hence, when a rule is built, the terms in its antecedent are used to update only the pheromone matrix(ces) of the class attribute(s) predicted by that rule.

FUTURE TRENDS

In the last few years, there has been increasing interest in specialized types of classification problems, which tend to be particularly challenging, such as: multi-label classification, where a single rule can predict multiple classes; hierarchical classification, where the classes to be predicted are arranged in a hierarchy; and the discovery of fuzzy classification rules. Recently, the first variations of Ant-Miner for some of these challenging classification problems have been proposed, as discussed earlier, and it seems that a future trend could be the development of more sophisticated Ant-Miner variations or related algorithms for these types of problems.

CONCLUSION

This article has reviewed Ant-Miner, the first ant colony algorithm for discovering classification rules in data mining, as well as a number of variations of the original algorithm. Some of the proposed variations were relatively simple, in the sense that they did not affect the type of classification rules discovered by the algorithm. However, some variations were proposed to cope with attributes having ordered categorical values, somewhat improving the flexibility of the rule representation language. Even more significant variations were also proposed. In particular, Ant-Miner variations have been proposed to discover multi-label classification rules and to discover fuzzy classification rules. The problem of extending Ant-Miner to cope with continuous attributes on-the-fly, during the run of the algorithm (rather than requiring continuous attributes to be discretized in a preprocessing phase), remains an open problem and an interesting future research direction.

REFERENCES

Chan, A., & Freitas, A.A. (2006). A new ant colony algorithm for multi-label classification with applications in bioinformatics. *Proceedings of the Genetic and Evolutionary Computation Conference* (GECCO-2006) (pp. 27-34). San Francisco: Morgan Kaufmann.

Chen, C., Chen, Y., & He, J. (2006). Neural network ensemble based ant colony classification rule mining. *Proceedings of the 1st International Conference on Innovative Computing, Information and Control* (ICICIC'06) (pp. 427-430).

Clark, P., & Niblett, T. (1989). The CN2 rule induction algorithm. *Machine Learning, 3*(4), 261-283.

Cover, T.M., & Thomas, J.A. (1991). *Elements of information theory.* New York: John Wiley & Sons.

Dorigo, M., & Stutzle, T. (2004). *Ant colony optimization.* Cambridge, MA: MIT Press.

Fayyad, U.M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.). *Advances in knowledge discovery & data mining* (pp. 1-34). Cambridge, MA: MIT Press.

Galea, M., & Shen, Q. (2006). Simultaneous ant colony optimization algorithms for learning linguistic fuzzy rules. In A.

Agraham, C. Grosan, & V. Ramos (Eds.). *Swarm intelligence in data mining* (pp. 75-99). Berlin: Springer-Verlag.

Liu, B., Abbass, H.A., & McKay, B. (2004). Classification rule discovery with ant colony optimization. *IEEE Computational Intelligence Bulletin, 3*(1), 31-35.

Martens, D., De Backer, M., Haesen, R., Baesens, B., & Holvoet, T. (2006). Ants constructing rule-based classifiers. In A. Agraham, C. Grosan, & V. Ramos (Eds.), *Swarm intelligence in data mining* (pp. 21-43). Berlin: Springer-Verlag.

Oakes, M.P. (2004). Ant colony optimisation for stylometry: The federalist papers. *Proceedings of the Conference on Recent Advances in Soft Computing* (RASC-2004) (pp. 86-91).

Parpinelli, R.S., Lopes, H.S., & Freitas, A.A. (2002a). Data mining with an ant colony optimization algorithm. *IEEE Transactions on Evolutionary Computation, Special Issue on Ant Colony Algorithms, 6*(4), 321-332.

Parpinelli, R.S., Lopes, H.S., & Freitas, A.A. (2002b). An ant colony algorithm for classification rule discovery. In H. Abbass, R. Sarker, & C. Newton (Eds.), *Data mining: A heuristic approach* (pp. 191-208). London: Idea Group.

Quinlan, J.R. (1993). *C4.5: Programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Smaldon, J., & Freitas, A.A. (2006). A new version of the Ant-Miner algorithm discovering unordered rule sets. *Proceedings of the Genetic and Evolutionary Computation Conference* (GECCO-2006) (pp. 43-50). San Francisco: Morgan Kaufmann.

Wang, Z., & Feng, B. (2004). Classification rule mining with an improved ant colony algorithm. *AI 2004: Advances in Artificial Intelligence* (pp. 357-367). Berlin: Springer-Verlag (LNAI 3339).

A

KEY TERMS

Classification Rule: A rule of the form *IF (conditions) THEN (class)*, meaning that if a case (record) satisfies the rule conditions, it is predicted to have the class specified in the rule.

Data Mining: A research field where the goal is to discover accurate, comprehensible knowledge (or patterns) in data.

Rule List: An ordered list of IF-THEN classification rules discovered by the algorithm during training. When the rules are applied to classify a testing case, they are applied in order, so that the first rule matching the testing case is used to classify that case.

Test Set: A set of cases unseen during the training of the algorithm. The test set is used to measure the predictive accuracy (generalization ability) of the rules discovered during training.

Testing Case: Each of the cases (records) of the test set.

Training Case: Each of the cases (records) of the training set.

Training Set: A set of cases used by the algorithm to discover classification rules.

Antecedents of Trust in Online Communities

Catherine M. Ridings
Lehigh University, USA

David Gefen
Drexel University, USA

INTRODUCTION

Online virtual communities have existed on the Internet since the early 1980s as Usenet newsgroups. With the advent of the World Wide Web and emphasis on Web site interactivity, these communities and accompanying research have grown rapidly (Horrigan, Rainie, & Fox, 2001; Lee, Vogel, & Limayem, 2003; Petersen, 1999). Virtual communities arise as a natural consequence of people coming together to discuss a common hobby, medical affliction, or other similar interest, such as coin collecting, a devotion to a rock group, or living with a disease such as lupus. Virtual communities can be defined as groups of people with common interests and practices that communicate regularly and for some duration in an organized way over the Internet through a common location or site (Ridings, Gefen, & Arinze, 2002). The location is the "place" where the community meets, and it can be supported technologically by e-mail listservs, newsgroups, bulletin boards, or chat rooms, for example. The technology helps to organize the community's conversation, which is the essence of the community. For example, messages in a community supported by a listserv are organized in e-mails, sometimes even grouping together several messages into an e-mail digest. In bulletin board communities, the conversation is organized into message threads consisting of questions or comments posted by members and associated replies to the messages.

Virtual community members form personal relationships with strong norms and expectations (Sproull & Faraj, 1997; Sproull & Kiesler, 1991), sometimes developing deep attachments to the communities (Hiltz, 1984; Hiltz & Wellman, 1997). These developments are interesting, because the members of virtual communities are typically strangers to one another and may never meet face to face. Additionally, the nature of computer-mediated communication is such that nonverbal cues that aid in the interpretation of communication, such as inflections in the voice, gestures, dress, tone, physical personal attributes, and posture, are missing (Sproull & Kiesler, 1991), making the communication open to multiple interpretations (Korenman & Wyatt, 1996). Yet, despite these limitations, many virtual communities flourish by exchanging messages and building their conversation base. A key ingredient in sustaining the conversation in the community is the existence of trust between the members. Trust has a downstream effect on the members' intentions to give and get information through the virtual community (Ridings et al., 2002).

This chapter examines emergent virtual communities, that is, those arising without direction or mandate from an organization, government, or other entity for an expressed economic or academic purpose. For example, a discussion board for a strategic partnership work group between two companies or a chat room for a class taking a college course would not be considered emergent virtual communities. However, an online forum established by the Breast Cancer Young Survivors Coalition so that women could discuss their battles with the disease would be considered an emergent virtual community.

BACKGROUND

Trust is an essential ingredient in social relationships (Blau, 1964; Luhmann, 1979), and understanding and defining trust are dependent upon the situation in which they are considered. In communities, in general, trust is an integral part of interpersonal relations among members and defines an individual's expectations and behavior (Luhmann, 1979; Rotter, 1971). Trust has many definitions. It has been defined as a willingness to take a risk associated with the behavior of others (Mayer, Davis, & Schoorman, 1995) and, more generally, as a method of reducing social uncertainty (Gefen, Karahanna, & Straub, 2003; Luhmann, 1979). In this sense, trust is used in the virtual community to reduce social complexity associated with the behavior of other members, and as a way of reducing the fear that the trusted party will take advantage by engaging in opportunistic behavior (Gefen et al., 2003), much as it does in communities in general (Fukuyama, 1995).

Participating in a virtual community entails exposure to risk. Opportunistic behaviors could include selling personal information that was confidentially provided, adopting a fictitious persona, deliberately and stealthily marketing products and services when this is prohibited, flaming or spamming, making unfair practical jokes at members, providing false information, and, in general, behaving in a dysfunctional

manner that ruins the community. Such behavior also applies to other types of communities, except that in the case of an online community, the anonymity provided by the Internet makes such behavior much easier to accomplish by the perpetrator and much harder to notice by the victim.

Scholarly research on trust has shown that trust is a multidimensional concept consisting of beliefs in ability, benevolence, and integrity (Blau, 1964; Butler, 1991; Giffin, 1967; Mayer et al., 1995; McKnight, Choudhury, & Kacmar, 2002). Ability deals with beliefs about the skills or expertise that another (i.e., trusted parties) has in a certain area. Ability relates to the belief that the other person knows what he or she is talking about. Because virtual communities are almost always focused on a specific topic, concerns about the abilities of others with respect to this topic are important. Benevolence is the expectation that others will have a positive orientation or a desire to do good to the trustee, typically by reciprocating with appropriate advice, help, discussion, and so on, such as contributing to the ongoing discussion with the intent to help, support, and care for others. Benevolence is important in virtual communities, because without positive reciprocation, the community would not exist. Integrity is the expectation that another will act in accordance with socially accepted standards of honesty or a set of principles, such as not telling a lie and providing reasonably verified information. Integrity applies in the virtual community context, because it is the existence of norms of reciprocity, closely linked with benevolence, that allow the community to properly function.

Research based upon surveying members of virtual communities has found that integrity and benevolence are united in this context, because the expected mode of behavior in many of the virtual communities is one of benevolence (Ridings et al., 2002). Hence, adhering to this expected mode of conduct, integrity, should overlap with actually behaving so, namely, with benevolence. Conformance to socially acceptable behavior or standards (integrity) and a desire to do "good" to others (benevolent intentions) are synonymous in the virtual community environment.

THE ANTECENDENTS OF TRUST

Trust in a virtual community is built through several mechanisms that are germane to the online context. As in personal contacts where successful interpersonal interaction builds trust (Blau, 1964; Gefen, 2000a; Luhmann, 1979), the responsiveness of other community members is necessary for trust to develop (Ridings et al., 2002). This can be shown through adherence to the social norms of the community (benevolence and integrity) and competency in the topic (ability). Members who post messages most often expect responses, and when these responses are absent, late, or lacking in number, there is no successful interpersonal interaction, and that hinders the development of trust. Responsiveness is also evident by members indicating gratitude for timely help. Trust is also built by reading what others post. If others post personal information about themselves, they appear less as strangers and more as acquaintances or friends. Divulging gender, age, name, e-mail address, or a personal problem may also add to the credibility of the member (ability) as well as make it easier for other members to shape beliefs regarding adherence to the community's standards and principles (integrity and benevolence). Personal information can also be provided in site profiles. Thus, the confiding of personal information also builds trust in other members of a virtual community (Ridings et al., 2002). Finally, humans have some degree of a general willingness to depend on others, known as disposition to trust (McKnight, Cummings, & Chervany, 1998), and this has been found to be stable across situations (Mayer et al., 1995). In the virtual community where people are unfamiliar with one another, disposition to trust, at least initially before extensive interactions take place, is also an important factor leading to the development of trust in others. Disposition to trust has been empirically found to be directly related to trust in virtual settings (Gefen, 2000a; Jarvenpaa, Knoll, & Leidner, 1998) and in virtual communities, in particular (Ridings et al., 2002).

Because virtual communities lack an enforceable legal system to ensure appropriate behavior online, the actual membership in the community and the feeling of being part of a community, even if a virtual one, may provide a possible way to enforce honest behavior. Virtual communities enhance honest behavior through creating what Ba (Ba, 2001; Ba, Whinston, & Zhang, 2003) called a trusted third party (TTP) certification mechanism. Considering the problems with the three current trust-building mechanisms in online markets (feedback, insurance or guarantee, and escrow), as pointed out theoretically by Ba and with some empirical support by Pavlou and Gefen (2004), extralegal mechanisms might be especially useful in virtual communities. Extralegal mechanisms, such as gossip, reproach, and community appreciation, and the praise and sanctions they bring, may serve to create trust just as they do in regular community settings.

Another way virtual communities may be applied to build trust, according to Ba, is through the sense of community, that, as we know from economics, is crucial when there is a separation in time between the quid and the pro (Ba, 2001). Moreover, if the members of the community are held responsible for the actions of an offender, there will be more social pressure to adhere to the rules. This might only work with online groups with a strong sense of community, but many virtual communities are precisely that.

Trust also has implications with regard to user privacy. Many virtual communities center on personal topics, such as medical conditions, legal issues, or occupations. Participants may care to be anonymous when communicating in such

communities. However, the economic viability of virtual communities may depend on the sale of advertising space or products to users. To accommodate this action, reliance is placed on the provision of user demographics, e-mail addresses, and traffic statistics, information somewhat at odds with the protection of user privacy. It may be possible for virtual communities to incorporate concepts of procedural fairness, where the provision and collection of personal information is perceived as conducted fairly. Procedural fairness has been found to address privacy concerns of customers. Culnan and Armstrong found that procedural fairness builds trust that customers have for an organization using personal information for marketing purposes (Culnan & Armstrong, 1999). Such procedural fairness policies could also be applied in the virtual community context.

FUTURE TRENDS

There are many directions to be investigated from the basic understanding of trust in virtual communities. It may be that trust develops differently in different kinds of communities. For example, trust in medical-based communities may be based more heavily on certain attributes than trust in communities organized for fans of a particular sports team. The level of trust for an individual may be related to the use of the community. Demographic variables such as gender, race, and culture may also influence trust and its development (Gefen, 2000b). Longitudinal studies of virtual communities may yield more information about the development of trust over time.

CONCLUSION

Virtual communities are a key resource on the Internet for individuals looking to exchange information with others as well as Web site sponsors desiring to provide interactivity to their sites. For virtual communities to survive, they must have conversation, and for the conversation to grow and flourish, there must exist trust between the virtual community members. Trust in the other members' abilities and benevolence and integrity is necessary, and this trust has been found to be built by the responsiveness of others, the confiding of personal information, and the member's general disposition to trust.

REFERENCES

Ba, S. (2001). Establishing online trust through a community responsibility system. *Decision Support Systems, 31*, 323–336.

Ba, S., Whinston, A. B., & Zhang, H. (2003). Building trust in online auction markets through an economic incentive mechanism. *Decision Support Systems, 35*, 273–286.

Blau, P. M. (1964). *Exchange and power in social life*. New York: John Wiley & Sons.

Butler, J. K. (1991). Toward understanding and measuring conditions of trust: Evolution of a condition of trust inventory. *Journal of Management, 17*(3), 643–663.

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science, 10*(1), 104–115.

Fukuyama, F. (1995). *Trust: The social virtues & the creation of prosperity*. New York: The Free Press.

Gefen, D. (2000a). E-commerce: The role of familiarity and trust. *Omega, 28*(6), 725–737.

Gefen, D. (2000b). Gender differences in the perception and adoption of e-mail and computer-mediated communication media: A sociolinguistics approach. In A. Kent (Ed.), *The encyclopedia of library and information science*. New York: Marcel Dekker.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly, 27*(1), 51–90.

Giffin, K. (1967). The contribution of studies of source credibility to a theory of interpersonal trust in the communication process. *Psychological Bulletin, 68*(2), 104–120.

Hiltz, S. R. (1984). *Online communities: A case study of the office of the future*. Norwood, NJ: Ablex Publishing Corporation.

Hiltz, S. R., & Wellman, B. (1997). Asynchronous learning networks as a virtual classroom. *Communications of the ACM, 40*(9), 44–49.

Horrigan, J. B., Rainie, L., & Fox, S. (2001). Online communities: Networks that nurture long-distance relationships and local ties. Retrieved from *http://www.pewinternet.org/reports/toc.asp?Report=47*

Jarvenpaa, S. L., Knoll, K., & Leidner, D. E. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Information Systems, 14*(4), 29–64.

Korenman, J., & Wyatt, N. (1996). Group dynamics in an e-mail forum. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 225–242). Philadelphia: John Benjamins.

Lee, F. S. L., Vogel, D., & Limayem, M. (2003). Virtual

community informatics: A review and research agenda. *Journal of Information Technology Theory and Application, 5*(1), 47–61.

Luhmann, N. (1979). *Trust and power* (H. Davis, J. Raffan, & K. Rooney, Trans.). UK: John Wiley and Sons.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research, 13*(3), 334–359.

McKnight, D. H., Cummings, L. L., & Chervany, N. L. (1998). Initial trust formation in new organizational relationships. *Academy of Management Review, 23*(3), 473–490.

Pavlou, P. A., & Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information Systems Research, 15*(1), 37–59.

Petersen, A. (1999, January 6). Some places to go when you want to feel right at home: Communities focus on people who need people. *The Wall Street Journal,* p. B6.

Ridings, C., Gefen, D., & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *Journal of Strategic Information Systems, 11*(3–4), 271–295.

Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist, 26*, 443–450.

Sproull, L., & Faraj, S. (1997). Atheism, sex and databases: The Net as a social technology. In S. Kiesler (Ed.), *Culture of the Internet* (pp. 35–51). Mahwah, NJ: Lawrence Erlbaum Associates.

Sproull, L., & Kiesler, S. (1991). *Connections: New ways of working in the networked organization*. Cambridge, MA: The MIT Press.

KEY TERMS

Disposition to Trust: A tendency to believe in the goodness of others based on lifelong socialization.

Reciprocity: Returning favors, which is a major way of building trust.

Trust: A willingness to take for granted that another person will behave as expected in a socially constructive manner. Trust generally reduces the perceived risk that another person will behave in an opportunistic manner.

Trust in Ability: The belief that a person has subject matter expertise in a certain area.

Trust in Benevolence: The belief that a person has a positive orientation or a desire to do good to others.

Trust in Integrity: The belief that a person will act in accordance with socially accepted standards of honesty or a set of principles.

Virtual Community: A group of people with common interests and practices that communicates regularly and for some duration in an organized way over the Internet through a common location or site.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 127-130,

Anytime, Anywhere Mobility

Mikael Wiberg
Umea University, Sweden

INTRODUCTION

Just a couple of years ago several mobile phone operators and others (e.g., Helal, 1999; Galambos, 2002; Ilderem, 2005) pushed forward "anytime, anywhere" as a goal or vision for future mobile services and mobile IT-use. In this article we set out to explore if "anytime, anywhere" mobility is in fact a paradox.

Kleinrock (1996, 1998) claims advanced wireless technologies, the Internet, global positioning systems, portable and distributed computing, and so forth, will realize the vision of "anytime, anywhere." We can today see the first signs of this vision. For example, telework is now possible, remote organizations can be engaged in close cooperation, and people can communicate, collaborate, share digital media, and form communities on the Internet. The world has become a global village, some claim (Preece, 1994, Castells, 1996), where you can interact with anybody independent of time and space.

The vision of "anytime, anywhere" describes a situation where people can do tasks wherever they want and without any consideration of time. Related to the vision is the 2x2 matrix often used in the field of computer supported cooperative work (CSCW) to denote different kinds of computer supported collaboration (e.g., Johansen, 1988; Baecker et al., 1993). This model has the dimensions of time and place, where each can be the same or different. The model is shown in Figure 1.

The vision of "anytime, anywhere" is tasks that can be done independent of time and place (i.e., in any of the four scenarios). This does not say anything about where or when the tasks should be done, only that these dimensions should not restrict them.

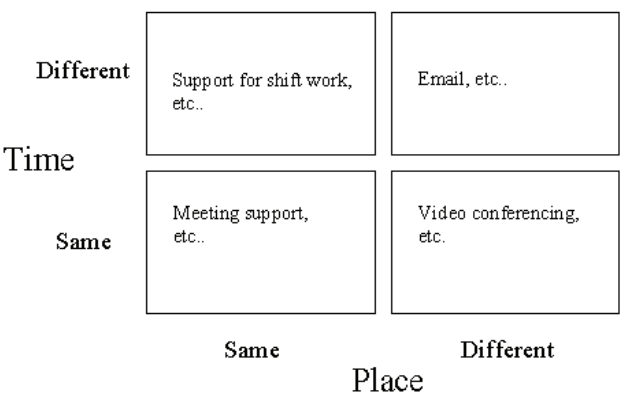
It is interesting to notice that the model does not take into consideration *mobility*. It assumes that people are either in the same place, or in a different place, and whether or not they are mobile does not seem to make a difference.

BACKGROUND

In the past, people traveled because they had no choice. If you wanted to do business or talk to remote friends you had to meet them face-to-face. However, transportation costs prohibited certain meetings and activities. A long series of technological developments (including the pony express, railroads, automobiles, and the telephone) have aimed at lowering the costs associated with transaction and conversation. Computer-mediated communications are the most recent development in that progression. Even so, people still travel and still meet in person.

To summarize: The adoption of Internet technologies, mobile phones, and so forth, have increased and in a sense made the world smaller. Compared to ten years ago, today it is much easier to communicate with remotes sites, and the frequency of communication in many organizations

Figure 1. The model shows different scenarios for groupware (Ellis et al., 1991)



has increased accordingly. Some people have even talked about "the global village" (Preece, 1994). A parallel trend is that people travel more than they used to do. According to predictions, this trend will sustain, and even increase. For example, the national road agency of Sweden reports the number of flights will increase by a factor of four in the next ten years. How can it be that the global village is so mobile? If people can interact and work independent of time and space, why then do they spend more and more time traveling? Is that not a paradox?

Reviewing the literature on the topic, we find no research that has explored this apparent paradox. Authors are either concerned with *remote interaction* (e.g., Ellis et al., 1991; Brave, Ishii & Dahley, 1998; McDaniel, 1996; Kuzuoka, 1992; and Tang & Minneman, 1991) *mobility* (e.g., Luff & Heath, 1998; Bejerano & Cidon, 1998; and Porta et al., 1996) or *mobility as anytime, anywhere work* (e.g., Dix et al, 2000; Perry et al., 2001; Davis, 2002; Ilderem, 2005). Furthermore, research on mobility has mainly dealt with technology issues, (e.g., limited battery life, unreliable network connections, varying channel coding and characteristics, volatile access points, risk of data loss, portability and location discovery) (e.g., Bhagwat, Satish, & Tripathi, 1994; Dearl, 1998; Francis, 1997; and Varshney, 1999). Accordingly, no research so far has explored the relation between, on one hand "the global village," with its idea that distance plays no role, and on the other hand the trend of increased mobility. How do the two trends hang together?

EXPLORING THE "ANYTIME, ANYWHERE" MOBILITY PARADOX

In order to investigate this seemingly paradox we conducted an empirical study of mobile telecommunication engineers in a Swedish company (Wiberg & Ljungberg, 2000). Using qualitative research methods, we studied to what extent the work tasks they do are dependent on time and place. We analyzed the data using a 2x2 matrix, with the two axis "time" and "space," which both have the categories "dependent" and "independent." One of the four situations is "anytime, any where," while the other three are dependent on time, place, or booth (see figure 2).

We found instances of work in all four categories. Some traveling seems very difficult to escape, simply because there are places that staff need to visit physically to do their job. For example, to repair a telephone pole you need to go there. We also found there are time frames that staff cannot escape. For example, rebooting parts of the telephone network has to be done at night. Lastly, there are work tasks that seem pretty much independent of time and space (e.g., scheduling and rescheduling of activities).

As observed during this empirical study there were just tiny parts of service work possible to perform "anytime, anywhere". Most of the work is dependent on spatial factors such as location of breakdown in the telephone network system, the location of the client, etc., or time related dependencies such as fixing problems within 24 hours or coordinate schedules to cooperate around larger problems. For a more throughout description of the empirical material see Wiberg & Ljungberg (2000). Overall, we found there are:

- *Traveling* that seems difficult to remove, thus places that people have to visit physically (e.g., telephone poles, customers houses, not all customers are mobile, network routers, locations where new cables needs to be drawn, etc.)
- *Time frames* which seem very difficult for staff not to do certain tasks within, e.g., customer service within 24 hours, rebooting parts of the telephone network has to be done at night, etc.

Figure 2. The theoretical framework of the study

		Place	
		Independent	**Dependent**
Time	**Independent**	**1. Anytime, anywhere**: Tasks that can be done independent of time and place; they can be done anytime, anywhere	**2. Anytime, particular place**: Tasks that need to be done in a particular place but can be done anytime
	Dependent	**3. Particular time, any place**: Tasks that can be done independent of place but at a certain time or in a certain order	**4. Particular time, particular place**: Tasks that must be done in a particular place within a particular time

- *Tasks* that do not seem to be restricted by time and place (e.g. scheduling and rescheduling of the activities over the day, co-ordinations of activities between the technicians, experiences and knowledge sharing among the technicians, etc.) although important for them since they are alone in their cars most of the day.

Accordingly, the vision of "anytime, anywhere" is not easy to realize in the case of the mobile workers we studied.

FUTURE TRENDS

Both work and leisure activities are becoming increasingly mobile. To describe the mobile worker, new concepts have been coined. Some examples are "road warriors" and "nomads" (Dahlbom, 1998), thus distinguishes mobile workers as moving from terms as distributed work, telework, and co-located work. One reason for this increased mobility is the emergence of service work as the dominating profession in the post-industrial society. Service work very often takes place at the client site, and therefore it is often mobile. Another reason is the increased importance of cooperation in and between organizations. Some cooperation can take place remotely, but people also need to meet physically. A third important reason for increased mobility is the extensive adoption of mobile phones. Mobile phones enable people to be mobile and yet accessible (Wiberg & Whittaker, 2005). As people have become accessible independent of place, new ways of working and new work rhythms have emerged in many organizations (Nilsson & Hertzum, 2005). So, for future development within this prominent area of mobile IT and mobility it is important to have a good understanding of mobile contexts (Tamminen et al., 2004) and keep in mind this "anytime, anywhere" paradox of the mobility vision.

CONCLUSION

This article has shown some limitations to the vision of "anytime, anywhere" in the context of mobile work. Time and place are indeed very old ways for understanding context and it seams like they are useful even for bringing light on the phenomena of the two parallel trends of the global village and mobility.

The article has shown that work has moments (e.g., *time frames*, which are not negotiable so the work is dependent upon those). The article has also shown that work has *places* of non negotiable importance (e.g., you cannot reframe the earth by putting away distance nor go backwards in time, although computers are often described as being able to bridge those gaps in time and space). As seen above, there is not much service work possible to perform "anytime, anywhere" since service work is not only about moving around (i.e., mobility) but also about taking actions at various sites at specific times.

Kleinrock (1998) has argued the vision of mobility as being able to work "anytime, anywhere." However, from the analysis of the empirical study presented above, we argue that there are several limits to that vision. In fact, this article has argued that the concept of "anytime, anywhere" belongs to another trend (i.e., the trend towards a global village, which is something altogether different from the trend of mobility). However, as the analysis has shown above those to trends comes together in practice.

So, finally we conclude this article arguing that the practical limitations of "anytime, anywhere" make it impossible for the mobile service engineers to conduct work "anytime, anywhere."

REFERENCES

Baecker, R. M. (Ed.). (1993). *Readings in groupware and computer supported cooperative work. Assisting human to human collaboration*. San Mateo: Morgan Kaufmann Publisher Inc.

Bejerano, Y., & Israel Cidon, I. (1998). An efficient mobility management strategy for personal communication systems, *The fourth annual ACM/IEEE international conference on Mobile computing and networking*.

Bhagwat, P., & Satish K. T. (1994). Mobile Computing. In *Proceedings of Networks'94*. (pp. 3-12).

Brave, S., Ishii, H., & Dahley, A. (1998). Tangible interfaces for remote collaboration and communication. In *Proceedings of the ACM 1998 conference on Computer Supported Cooperative Work*.

Castells, M. (1996). *The information age: Economy, society and culture*. Oxford, UK: Blackwell Publishers Ltd.

Coleman, J. S. (1988). Social capital in the creation of human capital. American *Journal of Sociology, 94*, 295-120.

Dahlbom (1998). From Infrastructure to Networking, In N. J. Buch, J. Damsgaard, L. Eriksen, J. Iversen, & P. Nielsen, (Eds.), *Proceedings of IRIS 21*. Department of Computer Science. Aalborg University.

Davis, G. (2002). Issues and challenges in ubiquitous computing: Anytime/anyplace computing and the future of knowledge work, *Communications of the ACM, 45*(12).

Dearle, A. (1998). Towards ubiquitous environments for mobile users, *IEEE Internet Computing, 2*(1), 22-32.

Dix & Beale (1996). *Remote cooperation: CSCW issues for mobile and teleworkers*. New York, Springer.

Dix, A., Rodden, T., Davies, N., Trevor, J., Friday, A., & Palfreyman, K. (2000). Exploiting space and location as a design framework for interactive mobile systems, *ACM Transactions on Computer-Human Interaction (TOCHI), 7*(3).

Ellis, C., Gibbs, S., & Rein, G. (1991). Groupware. Some issues and experiences, *Communications of the ACM, 34*(1), 39-58.

Francis, L. (1997). Mobile computing: A fact in your future. In *Proceedings of SIGDOC'97* (pp. 63-67).

Galambos, L. (2002). *Anytime, anywhere: Entrepreneurship and the creation of a wireless world*. Cambridge, USA: Camp. U.P.

Hammersley, M., & Atkinson, P. (1995). Ethnography: Principles in practice. London, Routledge.

Helal, A. (1999). *Any time, anywhere computing: Mobile computing concepts and technology*. Dordrecht, UK: Kluwer Academic Publishers.

Hughes, J., Randall, D., & Shapiro, D. (1993). From etnographic record to system design. Some experiences from the field, *An International Journal, 1*(3), 123-141.

Ilderem, V. (2005). Research and development for seamless mobility. In *Proceedings of the 15th ACM Great Lakes symposium on VLSI*. ACM Press.

Johansen, R. (1988). *Groupware: Computer support for business teams*. New York: The Free Press.

Kleinrock, L. (1996). Nomadicity: Anytime, anywhere in a disconnected world, Invited paper, *Mobile Networks and Applications, 1*(4), 351-357.

Kleinrock, L. (1998). Nomadic Computing: Information network and data communication. *IFIP/ICCC International Conference on Information Network and Data Communication*, Trondheim, Norway (pp. 223-233).

Kuzuoka, H., Kosuge, T., & Tanaka, M. (1994). GestureCam a video communication system for sympathetic remote collaboration. In *Proceedings of the Conference on Computer Supported Cooperative Work.*

Lindgren, R., & Wiberg, M. (2000). Knowledge management and mobility in a semi-virtual organization: Lessons learned from the case of Telia Nära. In *Proceedings of Hicss33.*

Luff, P., & Heath, C. (1998). Mobility in collaboration. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work.*

Mason, R. O. (1989). MIS experiments: A pragmatic perspective. In *Information systems research challenge: Experimental research methods* (pp. 3-20), Boston: Harvard Business School Press.

McDaniel, S. (1996). Providing awareness information to support transitions in remote computer-mediated collaboration. In *Proceedings of the CHI '96 conference companion on Human factors in computing systems: Common ground.*

Nilsson, M., & Hertzum, M. (2005). Work rhythms and coordinative artifacts: Negotiated rhythms of mobile work: time, place, and work schedules. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work GROUP '05*. ACM Press.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: Understanding access anytime, anywhere, *ACM Transactions on Computer-Human Interaction (TOCHI*), *8*(4).

Porta, T., Sabnani, K., & Gitlin, R. (1996). Challenges for nomadic computing mobility management and wireless communications. *Mobile Networking Applications, 1*(1). Kluwer Academic Publishers.

Preece, J. (1994). *Human-Computer Interaction*. New York: Addison & Wesley.

Tamminen, S., Oulasvirta, A., Toiskallio, K., & Kankainen, A. (2004).Understanding mobile contexts. *Personal and Ubiquitous Computing, 8*(2).

Tang, J., & Minneman, S. (1991). Videowhiteboard: Video shadows to support remote collaboration. In *Proceedings of the 1991 Conference on Human Factors in Computer Systems, CHI '91*. (pp. 315-322).

Varshney, U. (1999). Networking support for mobile computing, *AIS (Communications of the Association for Information Systems), 1*(1).

Wiberg, M., & Ljungberg, F. (2000). Exploring the vision of anytime, anywhere in the context of mobile work. *In Knowledge management and virtual organizations: Theories, practices, technologies and methods, the biztech network.* Brint Press.

Wiberg, M., & Whittaker, S. (2005). Managing Availability: Supporting Lightweight Negotiations to Handle Interruptions, *ACM Transactions of Computer-Human Interaction (ToCHI), 12*(4).

KEY TERMS

"Anytime, Anywhere" Work: Describes a situation where people can do tasks wherever they want and without any consideration of time (i.e. they can be done anytime, anywhere).

Co-Located Work: Collaborative work carried out by several persons at the same geographical location.

Distributed Work: Collaborative work carried out by several persons at different geographical location.

Global Village: As computers all over the world become interconnected via the Internet and the frequency of communication in and between organizations, countries, cultures, societies etc. has increased accordingly via these networks we can now, on a daily basis and quite easily, maintain contact with anybody independent of time and space, that is. to be able to interact "anytime, anywhere."

Mobile Work: The ability to carry out work while geographically moving around.

"Particular Time, Particular Place" Work: Tasks that must be done in a particular place within a particular time.

Remote Interaction: Information technology mediated human-to-human communication over a distance.

Telework: The ability to carry out work from a distance. For example, sitting at home and doing office work. Telework does not imply that the worker is mobile (i.e., in motion) in any sense even though the concept of telework is sometimes used as a synonym for mobile work.

A

Application of Cognitive Map in Knowledge Management

Ali Reza Montazemi
McMaster University, Canada

Akbar Esfahanipour
Amirkabir University of Technology, Iran

INTRODUCTION

Cognitive map methodologies consist of a set of procedures to capture perceived relationships of attributes related to ill-structured decision problems that decision makers have to face. This article provides an overview of the application of cognitive maps (CMs) in the design and development of intelligent information systems. Here, CM is used as a set of techniques to identify subjective beliefs and to portray those beliefs externally as follows:

- Causal mapping is used to investigate the cognition of decision-makers. A causal map represents a set of causal relationships (i.e., cause and effect relationships) among constructs within a system. For example, Figure 1 shows that better sanitation facilities, causing an initial improvement in health, led to an increase in the city's population. This growth led to more garbage, more bacteria, and therefore more disease. Causal map aids: 1) in identification of irrelevant data, 2) to evaluate the factors that affect a given class of decisions, and 3) enhances the overall understanding of a decision maker's environment, particularly when it is ill-structured.
- Semantic mapping, also known as *idea mapping*, is used to explore an idea without the constraints of a superimposed structure. A semantic map visually organizes related concepts around a main concept with tree-like branches. Figure 2 depicts different types of transportation, organized in three categories: land, water, and air. This technique facilitates communication between end-users and system analysts in support of information requirements analysis.
- Concept mapping is a useful tool for organizing and representing concepts (events or objects) and their interrelationships in a particular domain. Each concept is designated with a label. The relationship between two concepts in a concept map is referred to as a proposition; propositions connect concepts to form a meaningful statement. Relationships between concepts are associative. For example, in Figure 3, two concepts of "plants" and "flowers" are associated via "may have" that form the proposition of "plants may have flowers." Describing complex structures with simple propositions improve quality of conceptual modeling in the development of information systems.

Figure 1. Causal map for public health issues

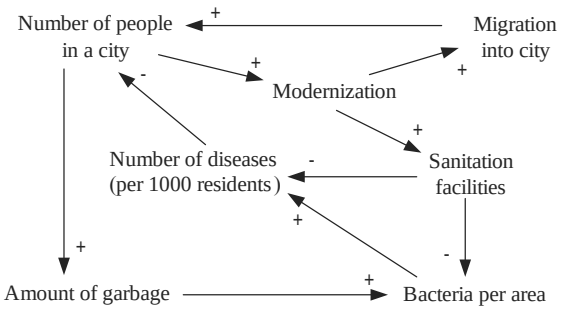


Figure 2. Semantic map for different types of transportation

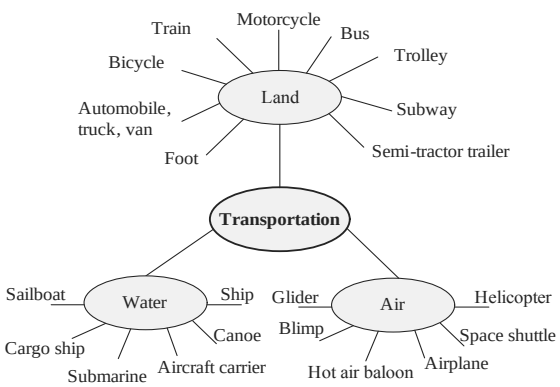
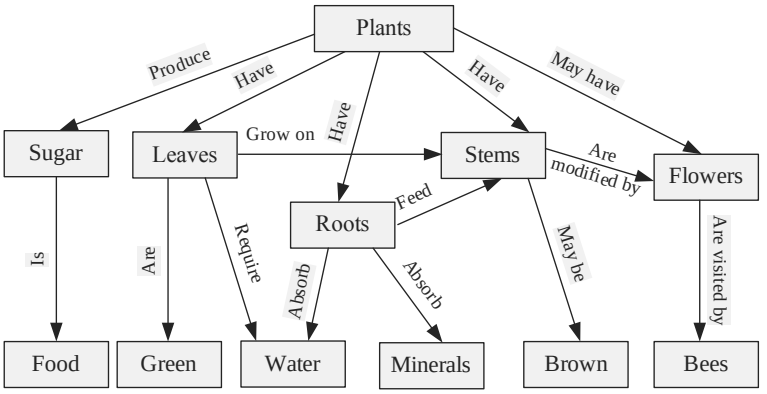


Figure 3. Concept map for plants



BACKGROUND

Cognitive Map (CM) has been employed to capture, store and retrieve expert knowledge in support of the design and development of intelligent information systems. CM is a representation of the relationships that are perceived to exist among the elements of a given environment. Taking any two of these elements, the concern is whether the state or movement of the one is perceived to have an influence on the state or movement of the other (both static and dynamic relationships can be considered) (Montazemi & Conrath, 1986). CMs have been used to describe experts' tacit knowledge about a certain problem, particularly in ill-structured decision problems (Axelrod, 1976; Montazemi & Chan, 1990). Tacit knowledge is personal knowledge, shared and exchanged through direct and face-to-face contact among actors (Eden, 1988).

There are different perspectives of knowledge within organizations (Nonaka, 1994). Thus, it seems appropriate to use knowledge management categories to identify different applications of cognitive map in the design and development of intelligent information systems. Alavi and Leidner (2001) provide a framework that is grounded in the sociology of knowledge and is based on the view of organizations as social collectives and "knowledge systems." They contend that organizations as knowledge systems consist of four sets of socially enacted "knowledge processes" as follows:

- **Knowledge application:** Those activities concerned with deploying knowledge in order to produce goods and services. Information technology can enhance knowledge application by facilitating capture, updating and accessibility of organizational directives.
- **Knowledge storage/retrieval:** This is also referred to as *organizational memory*, which includes knowledge residing in various component forms, including written documentation, structured information stored in electronic databases, codified human knowledge stored in expert systems, and tacit knowledge acquired by individuals.
- **Knowledge transfer:** Transfer of knowledge to locations where it is needed and can be used. This can occur at various levels: transfer of knowledge between individuals, from individuals to explicit sources, from individuals to groups, between groups, across groups, and from groups to the organization.
- **Knowledge creation:** Developing new content or replacing existing content within the organization's tacit and explicit knowledge. Through social and collaborative processes as well as through the cognitive processes of the individual, knowledge is created, shared, amplified, enlarged and justified in organizational settings.

A literature search shows that application of CMs in intelligent information systems can be found in support of the above four categories of knowledge processes, as depicted in the Appendix. A brief description of each of the above four categories within the context of CMs is presented next.

KNOWLEDGE PROCESSING THROUGH CMs

Knowledge Application

CM techniques (e.g., causal mapping, semantic mapping and concept mapping) have been used to improve the processes of the design and development of information systems. They

include improvement of conceptual modeling (Siau & Tan, 2005a), user–database interaction (Siau & Tan, 2006) and applicability of the resulting information systems (Siau & Tan, 2005b). Siau and Tan (2006) proposed a framework of user-database interaction in support of knowledge application. According to this framework, user-database interaction has three dimensions: content, structure and style. The content (i.e., semantics) dimension is determined by the user's data needs. The structure dimension is determined by both the user's perceived data model (constituent structure) and by the query language (syntactic structure). The style dimension is dictated by user-database interface.

According to this framework, to begin with we need to decide what elements and operations of the database schema (constituent structure) are relevant in order to translate the data needs (content) into a database query. Next, the query is written using the syntax structure as dictated by the query language. Finally, the query is entered into the system in a style offered by the user-database interface. The possible styles include command languages, menu selection, and form fill-in. In this framework, the user's knowledge is critical to an understanding of the data needs, to the formation of a correct and appropriate database schema, and to a proper application of the syntax of the query language. The resulting CMs are suitable materials for user's guides and online help.

Knowledge Storage/Retrieval

Cognitive mapping techniques have been also applied to acquire a decision maker's tacit knowledge when faced with ill-structured decision problems such as the extraction of expert knowledge to forecast stock prices (Montazemi & Chan, 1990), the building of expert systems to control electronic data interchange (Lee & Lee, 2007), for negotiation in a B2B system (Lee & Kwon, 2006) and for analysis of firms with regard to their creditworthiness (Noh, Lee, Kim, Lee, & Kim, 2000). CMs have been used to develop a specific method to analyze complex problems. For example, Satur and Liu (1999) developed an inference method in support of geographic information systems and Kwon, Im, and Van de Walle (2002) proposed a structural analysis method for a DSS in support of urban planning. Noh et al. (2000) adopted CMs to formalize tacit knowledge and proposed case-based reasoning as a tool for storage/retrieval of CM-driven tacit knowledge in the form of frame-typed cases.

Knowledge Transfer

Cognitive maps have been applied in group decision making for the exchange of ideas and opinions between group members in regard to a specific problem. For instance, Caliskan (2006) used the CM approach to facilitate the evaluation of transport investment alternatives using a group of experts. Sengupta and Te'eni (1993) studied the impact of cognitive feedback on group decision making at three levels: individual, interpersonal and collective. They contend that for the development of a group decision support system (GDSS), cognitive feedback should be considered as an integral part at every level, and that a human-computer interaction should be designed as an effective transition across the components of feedback at all levels.

Fuzzy cognitive map (FCM) also has been applied in group decision making to create a new FCM representing the views of a number of experts in a unified manner. For example, Khan and Quaddus (2004) presented a methodology for the development and analysis of FCM as a useful GDSS tool. This methodology consists of two phases: development and application. In the development phase, an FCM is created for each group member and these are then merged to produce a group FCM. The application phase consists of static analysis of concepts and causal relationships, and dynamic analysis of the simulated system over time.

FCMs have been also used to develop a fuzzy cognitive agent to provide personalized recommendations to online customers (Miao, Yang, Fang, & Goh, 2007). The agent learns users' preferences from the most recent cases and helps customers to make inferences and decisions. Xirogiannis and Glykas (2007) proposed the application of the fuzzy causal characteristics of FCMs as the underlying methodology in order to generate a hierarchical and dynamic network of interconnected maturity indicators in e-business strategy formulation exercises. A brief description of this model follows.

In business strategy formulation, maturity metrics are defined by managers at different organizational levels in a hierarchical manner. Top management sets the overall performance targets (strategic maturity). These targets are exemplified further to action plan performance metrics (tactical maturity) and then to operational performance indicators (operational maturity). This model represents inherent relationships between these metrics utilizing FCMs to interpret:

- E-business maturity metrics as concepts
- Decision weights as relationship weights
- Decision variables as maturity concept values
- Hierarchical decomposition of maturity metrics as a hierarchy of FCMs.

This approach allows the stakeholders to reason about the qualitative states of e-business maturity metrics using fuzzy linguistic.

Knowledge Creation

Cognitive mapping techniques have been applied for knowledge creation. Nonaka (1994) defines the four "modes" of

knowledge creation as 1) socialization: conversion of tacit knowledge to new tacit knowledge through interaction of individuals, sharing ideas and experiences and learning from each other; 2) externalization: conversion of tacit knowledge to new explicit knowledge such as articulation of best practices or lessons learned; 3) combination: conversion of explicit knowledge to new explicit knowledge by merging, categorizing, reclassifying and synthesizing existing explicit knowledge; 4) internalization: conversion of explicit knowledge to new tacit knowledge that could be achieved through the learning and understanding that result from reading about a specific topic.

Given that tacit knowledge is difficult to formalize and communicate (Polanyi, 1966), socialization, which occurs through conversion of such tacit knowledge to a new form of tacit knowledge, is a complex task to model. As a result, CMs have not been applied directly to model socialization mode of knowledge creation. However, the four aforementioned knowledge creation modes are interdependent and intertwined so that each mode relies on, contributes to, and benefits from the other modes (Alavi & Leidner, 2001). Hence, it is possible to use one or more intermediate conversion steps to model socialization using CMs. That is, conversion of tacit knowledge to a new explicit knowledge (i.e., externalization) and then converting the obtained explicit knowledge to a new tacit knowledge (i.e., internalization). We may need to use a combination mode before internalization, however.

Cognitive mapping allows modeling of mental models of decision makers, and it leads to clarification and structuring of the experts' thought processes when faced in ill-structured problems (Rodhain, 1999). Thus, CMs can be used as a means of externalization. For example, Rodhain (1999) uses CMs to extract business strategy mental models of managers for determining a portfolio of projects. Other pertinent applications include application of CMs to extract experts' tacit knowledge in order to forecast stock prices (Montazemi & Chan, 1990) and to analyze information requirements (Montazemi & Conrath, 1986).

FCM has been used in group decision making to construct a causal knowledge base in stock investment analysis (Lee & Kim, 1997) and in electronic data interchange (EDI) controls (Lee & Lee, 2007) as a means of externalization, so that the resulting CMs could be transferred, managed and stored as explicit knowledge.

Individuals' CMs could be different in a specific domain because perceived attributes or perceived causal relationships of CMs may be different from one person to another. These differences between CMs, which are called *content difference* (Langfield-Smith & Wirth, 1992) exist because of different levels of individual experience and expertise. Langfield-Smith and Wirth (1992) propose quantitative measures for analyzing content difference between two or more CMs. These measures would assist in providing a more objective basis for qualitative analysis. Eden (2004) discusses the use of CMs as a qualitative method for structuring ill-structure decision problems. Wang (1996) uses neural network as a quantitative method to measure differences between CMs of individuals over time. The comparisons of CMs provide useful information for a decision maker when considering individual points of view. Providing such information can be considered as combination, and analyzing of this information by decision makers to get an insight into a given problem can be considered as the internalization mode of knowledge creation.

Learning methods have been used to improve knowledge represented by means of FCMs. These learning methods include the nonlinear Hebbian learning rule (Papageorgiou, Stylios, & Groumpos, 2003) and unsupervised learning technique (Papageorgiou, Stylios, & Groumpos, 2006). Stach, Kurgan, Pedrycz, and Reformat (2005) propose a genetic learning method to generate FCMs from historical data. Improving the embedded knowledge of CMs through machine learning methods can be considered as combination mode of knowledge creation that can improve the user's tacit knowledge (i.e., internalization).

For example, Lee, Kim, Chung, and Kwon (2002) propose a three-phased Web-mining inference amplification (WEMIA) based on inference logic of FCM. The first phase is used to extract association rules. Because association rules are similar to causal knowledge in which a condition clause triggers a conclusion clause, and then in the second phase, corresponding FCMs are developed. The advantages of causal knowledge over association rules are that it can provide a more refined inference mechanism than can the association rules (Lee et al., 2002). The final phase is used to apply inference amplification procedures to the causal knowledge: to eliminate rule redundancy, to search for the directly and indirectly chained rules, and to amplify inferences about the logical relationships between rules. In this case, knowledge is created thanks to the conversion of some explicit knowledge (i.e., association rules) to another form of explicit knowledge (i.e., FCM), which is referred to as *knowledge combination*. Here, knowledge is improved by the use of inference methods to derive rich semantics which are suitable for and understandable by the decision maker. This is referred to as the *internalization mode of knowledge creation*.

Thus, cognitive mapping techniques can be applied directly toward externalization and combination modes of knowledge creation and can contribute indirectly toward internalization and eventually to the socialization mode of knowledge creation.

FUTURE TRENDS

This literature review shows diverse use of CM in the varied domain of decision making processes. We expect increased

sophistication in capturing CM and its application in support of knowledge management. In particular, CM can play a major role in realizing the usefulness of the Semantic Web, which is the outgrowth of many diverse desires and influences, all aimed at making better use of the Web as it stands. The Semantic Web is portrayed as: 1) a universal library, to be readily accessed and used by humans in a variety of information use contexts; 2) the backdrop for the work of computational agents completing sophisticated activities on behalf of their human counterparts; and 3) a method for federating particular knowledge bases and databases to perform anticipated tasks for humans and their agents (Marshal & Shipman, 2003). Intelligent agents can use the CM of users to filter information from the Semantic Web, making it possible to manage organizational knowledge using the Intranet and the vast resources available from the Internet.

CONCLUSION

Cognitive mapping techniques have been applied in design and development of intelligent information systems as a means of acquiring knowledge from domain experts, manipulating knowledge to create new contents, storing/retrieving knowledge for later use, and sharing knowledge between group members to deal with ill-structured decision problems. Successful applications of cognitive mapping methodologies reveal the efficiency and flexibility of these techniques, which can be potentially utilized by developers of intelligent information systems as follows:

- Providing a rich picture of ill-structured problems for decision makers,
- Facilitating group decision making through clear exchange of ideas and information,
- Modeling dynamic systems in uncertain environments using fuzzy cognitive maps,
- Providing cognitive feedbacks to improve decision makers' decision processes when faced with ill-structured decision problems,
- Improving the quality of conceptual modeling of systems, user-system interactions and usability of the resulting systems, and
- Better identifying end-users' information requirements and improving user-developer communications.

REFERENCES

Alavi, M., & Leidner, D.E. (2001). Review: knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Aguilar, J. (2002). Adaptive random fuzzy cognitive maps. In F.J. Garijio, J.C. Riquelme, & M. Toro (Eds.), *IBERAMIA 2002, Lecture Notes in Artificial Intelligence 2527, Heidelberg*, (pp. 402-410). Berlin: Springer-Verlag.

Axelrod, R. (1976). *Structure of decision*. Princeton, NJ: Princeton University Press.

Caliskan, N. (2006). A decision support approach for the evaluation of transport investment alternatives. *European Journal of Operational Research, 175*(3), 1696-1704.

Eden, C. (1988). Cognitive mapping: A review. *European Journal of Operational Research, 36*(1), 1-13.

Eden, C. (2004). Analyzing cognitive maps to help structure issues or problems. *European Journal of Operational Research, 159*(3), 673-686.

Khan, M., & Quaddus, M. (2004). Group decision support using fuzzy cognitive maps for causal reasoning. *Group Decision Negotiation Journal, 13*(5), 463-480.

Kwon, H., Im, I., & Van de Walle, B. (2002). Are you thinking what I am thinking?—a comparison of decision makers' cognitive maps by means of a new similarity measure. In *Proceedings of the 35th Hawaii International Conference on System Sciences*.

Langfield-Smith, K., & Wirth, A. (1992). Measuring differences between cognitive maps. *Journal of the Operational Research Society, 43*(12), 1135-1150.

Lee, K.C., & Kim, H.S. (1997). A fuzzy cognitive map-based bi-directional inference mechanism: An application to stock investment analysis. *Intelligent Systems in Accounting, Finance and Management, 6*, 41-57.

Lee, K.C., Kim, J.S., Chung, N., & Kwon, S.J. (2002). Fuzzy cognitive map approach to Web-mining inference amplification. *Expert Systems with Applications, 22*(3), 197-211.

Lee, K.C., & Kwon, S.J. (2006). The use of cognitive maps and case-based reasoning for B2B negotiation. *Journal of Management Information Systems, 22*(4), 337-376.

Lee, K.C., & Lee, S. (2003). A cognitive map simulation approach to adjusting the design factors of the electronic commerce Web sites. *Expert Systems with Applications, 24*(1), 1-11.

Lee, K.C., & Lee, S. (2007). Causal knowledge-based design of EDI controls: An explorative study. *Computers in Human Behavior, 23*(1), 628-663.

Marshall, C.C., & Shipman, F.M. (2003, August 26-30). Which semantic Web? In *Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia (HYPERTEXT '03)*, Nottingham, United Kingdom, (pp. 57-66).

Miao, C., Yang, Q., Fang, H., & Goh, A. (2007). A cognitive approach for agent-based personalized recommendation. *Knowledge Based Systems, 20*(4), 397-405.

Montazemi, A.R., & Conrath, D.W. (1986). The use of cognitive mapping for information requirements analysis. *MIS Quarterly*, 10(1), 44-55.

Montazemi, A.R., & Chan, L. (1990). An analysis of the structure of expert knowledge. *European Journal of Operational Research, 45(*2), 275-292.

Noh, J.B., Lee, K.C, Kim, J.K., Lee, J.K., & Kim, S.H. (2000). A case-based reasoning approach to cognitive map-driven tacit knowledge management. *Expert Systems with Applications, 19*(4), 249-259.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

Papageorgiou, E.I., Stylios, C.D., & Groumpos, P.P. (2003). Fuzzy cognitive map learning based on nonlinear Hebbian rule. In T.D. Gedeon, & L.C.C. Fung (Eds.), *AI 2003, Lecture Notes in Artificial Intelligence 2903*, (pp. 254-266). Berlin Heidelberg: Springer-Verlag.

Papageorgiou, E.I., Stylios, C.D., & Groumpos, P.P. (2006). Unsupervised learning techniques for fine-tuning fuzzy cognitive map causal links. *International Journal of Human-Computer Studies, 64*(8), 727-743.

Polanyi, M. (1966). *The tacit dimension.* London: Routledge & Kegan Paul.

Rodhain, F. (1999). Tacit to explicit: Transforming knowledge through cognitive mapping—an experiment. In *Proceedings of the Special Interest Group on Computer Personnel Research (SIGCPR '99),* New Orleans, LA, USA.

Rodriguez-Repiso, L., Setchi, R., & Salmeron, J.L. (2007). Modeling IT projects success with fuzzy cognitive maps. *Expert Systems with Applications, 32*(2), 543-559.

Satur, R., & Liu, Z.Q. (1999). A contextual fuzzy cognitive map framework for geographic information systems. *IEEE Transactions on Fuzzy Systems, 7*(5), 481-494.

Sengupta, K., & Te'eni, D. (1993). Cognitive feedback in GDSS: Improving control and convergence. *MIS Quarterly, 17*(1), 87-113.

Siau, K., & Tan, X. (2005a). Improving the quality of conceptual modeling using cognitive mapping techniques. *Data & Knowledge Engineering, 55*(3), 343-365.

Siau, K., & Tan, X. (2005b). Technical communication in information systems development: The use of cognitive mapping. *IEEE Transaction on Professional Communication, 48*(3), 269-284.

Siau, K., & Tan, X. (2006). Cognitive mapping techniques for user–database interaction. *IEEE Transactions on Professional Communication, 49*(2), 96-108.

Stach, W., Kurgan,L., Pedrycz, W., & Reformat, M. (2005). Genetic learning of fuzzy cognitive maps. *Fuzzy Sets and Systems, 153*(3), 371-401.

Wang, S. (1996). A dynamic perspective of differences between cognitive maps. *Journal of the Operational Research Society, 47*(4), 538-549.

Xirogiannis, G., & Glykas, M. (2007). Intelligent modeling of e-business maturity. *Expert Systems with Applications, 32*(2), 687-702.

KEY TERMS

Cognitive Map: A representation of the relationships which are perceived to exist among the elements of a given environment.

Explicit Knowledge: Codified knowledge which refers to knowledge that is transmittable in formal and systematic language.

Fuzzy Cognitive Map: An extended and fuzzified version of the cognitive map that enables causal relationships to have fuzzy weights.

Knowledge Application: Those activities concerned with deploying knowledge in order to produce goods and services.

Knowledge Creation: Creation of new content based on the organizational tacit and explicit knowledge.

Knowledge Transfer: Transfer of organizational knowledge from one entity to another entity within/between organizations.

Tacit Knowledge: Personal knowledge which could be shared and exchanged through face-to-face contact among actors.

APPENDIX

A

Table 1. Articles related to the application of CMs in intelligent information systems

Author (year)	Method	Application	Knowledge perspective*			
			1	2	3	4
Wang, 1996	used neural networks in a three-dimensional framework defined by initial input to the system, time horizon and thresholds	measuring differences between CMs	✓			
Montazemi, & Conrath, 1986	applied cognitive mapping technique for the evaluation of the performance of insurance claim representatives	Information requirements analysis	✓			✓
Montazemi & Chan, 1990	used CMs to extract the structure of expert knowledge	stock price forecasting	✓			✓
Satur & Liu, 1999	used contextual FCM to draw inferences from quantitative and qualitative descriptions	Geographic Information Systems	✓			✓
Lee & Kim, 1997	used FCMs to construct a causal knowledge base and perform a bi-directional inference	stock investment analysis	✓	✓		✓
Eden, 2004	used CMs to structure problems or issues		✓			
Lee & Lee, 2007	used FCM to develop a causal knowledge-based expert system	electronic data interchange (EDI) controls	✓	✓		✓
Lee & Lee, 2003	used simulation of CM to adjust the design factors of the EC Web sites and using LISREL to prove the proposed research model	adjusting the design factors in electronic commerce Web site	✓			
Lee et al., 2002	used FCMs to develop causal knowledge base from mined association rules	Web mining inference amplification	✓			
Kwon et al., 2002	proposed a method to assess the structural similarities among the FCMs of different decision makers	DSS for urban planning decision problem	✓			✓
Aguilar, 2002	developed an adaptive FCM based on the random neural network model	modeling of dynamic system	✓			✓
Papageorgiou et al., 2003	used unsupervised Hebbian algorithm to nonlinear units for training FCMs	learning methods	✓			✓
Papageorgiou et al., 2006	used unsupervised learning techniques for fine-tuning weights of concepts in FCMs	industrial process control	✓			✓
Stach et al., 2005	used genetic algorithm to generate FCMs from historical data	learning methods	✓			
Lee & Kwon, 2006	used CM to formalize tacit knowledge and applied CBR techniques for storing and retrieving CMs.	reasoning in B2B negotiations		✓		✓
Noh et al., 2000	used CM to formalize tacit knowledge, and applied CBR techniques to store and reuse tacit knowledge	credit analysis problem		✓		✓
Sengupta, & Te'eni, 1993	applied cognitive feedback in GDSS to improve decision making.				✓	
Miao et al., 2007	used a fuzzy cognitive agent to provide personalized recommendations to online customers	e-commerce application			✓	✓
Caliskan, 2006	used CM and AHP technique as a DSS	evaluation of transport investment alternatives			✓	✓
Siau & Tan, 2006	used CM techniques (causal mapping, semantic mapping, and concept mapping) to improve user-database interaction.	User-database interface analysis and design				✓

** Knowledge perspective: 1) knowledge creation 2) knowledge storage/retrieval 3) knowledge transfer 4) knowledge application*

Table 1. Articles related to the application of CMs in intelligent information systems, continued

Author (year)	Method	Application	Knowledge perspective*			
			1	2	3	4
Siau & Tan, 2005a	used CM techniques to improve the quality of conceptual modeling.	conceptual modeling				✓
Siau & Tan, 2005b	used CMs to improve the usability information systems	Technical communication in IS development				✓
Khan & Quaddus, 2004	used FCM to improve GDSS usability	group decision support environment			✓	
Xirogiannis & Glykas, 2007	used FCM to model hierarchical and distributed nature of e-business maturity	e-business strategy formulation			✓	✓
Rodriguez-Repiso et al., 2007	used FCM for mapping success, modeling Critical Success Factors (CSFs) perceptions and the relations between them	IT project success				✓

** Knowledge perspective: 1) knowledge creation 2) knowledge storage/retrieval 3) knowledge transfer 4) knowledge application*

Application of Fuzzy Logic to Fraud Detection

Mary Jane Lenard
University of North Carolina – Greensboro, USA

Pervaiz Alam
Kent State University, USA

INTRODUCTION

In light of recent reporting of the failures of some of the major publicly-held companies in the U.S. (e.g., Enron & WorldCom), it has become increasingly important that management, auditors, analysts, and regulators be able to assess and identify fraudulent financial reporting. The Enron and WorldCom failures illustrate that financial reporting fraud could have disastrous consequences both for stockholders and employees. These recent failures have not only adversely affected the U.S. accounting profession but have also raised serious questions about the credibility of financial statements. KPMG (2003) reports seven broad categories of fraud experienced by U.S. businesses and governments: employee fraud (60%), consumer fraud (32%), third-party fraud (25%), computer crime (18%), misconduct (15%), medical/insurance fraud (12%), and financial reporting fraud (7%). Even though it occurred with least frequency, the average cost of financial reporting fraud was the highest, at $257 million, followed by the cost of medical/insurance fraud (average cost of $33.7 million).

Statistical methods, expert reasoning, and data mining may be used to achieve the objective of identifying financial reporting fraud. One way that a company can justify its financial health is by developing a database of financial and non-financial variables to evaluate the risk of fraud. These variables may help determine if the company has reached a stress level susceptible to fraud, or the variables may identify fraud indicators. There are a number of methods of analysis that may be used in fraud determination. Fuzzy logic is one method of analyzing financial and non-financial statement data. When applied to fraud detection, a fuzzy logic program clusters the information into various fraud risk categories. The clusters identify variables that are used as input in a statistical model. Expert reasoning is then applied to interpret the responses to questions about financial and non-financial conditions that may indicate fraud. The responses provide information for variables that can be developed continuously over the life of the company. This article summarizes the specifics of fraud detection modeling and presents the features and critical issues of fuzzy logic when applied for that purpose.

BACKGROUND

Fraud Detection

The problem of fraudulent financial reporting is not limited to the U.S. In 2002, the Dutch retailer, Ahold, disclosed losses of $500 million related to accounting at its U.S. subsidiary (Arnold, 2003). Recently, Parmalat, an Italian firm, declared insolvency as a result of fraudulent financial reporting. The CEO of Parmalat has been accused of mishandling $10 billion and of hiding losses in offshore funds and bank accounts. The scandal at Parmalat could also have serious consequences for the company's auditor (Gallani & Trofimov, 2004).

The auditor's responsibility for fraud detection in the U.S. has been defined in Statement on Auditing Standards No. 99, *Fraud Detection in a GAAS Audit* (AICPA, 2002). This statement has four key provisions (Lanza, 2002): (1) increased emphasis on professional skepticism, (2) frequent discussion among audit team personnel regarding the risk of misstatement due to fraud, (3) random audit testing of locations, accounts, and balances, and (4) procedures to test for management override of controls. Auditors are discouraged from placing too much reliance on client representation and are required to maintain a skeptical attitude throughout the audit. The standard encourages auditors to engage in frequent discussion among engagement personnel regarding the risk of material misstatement due to fraud. SAS 99 also requires auditors to inquire of management and others not directly involved with fraud, perform analytical procedures, and conduct necessary tests to assess management override of controls. Finally, auditors are advised to evaluate the risk of fraud and steps taken by the client to mitigate the risk of fraud.

The U.S. Congress in 2002 passed the Sarbanes-Oxley Act, which spells out a number of steps firms must take to minimize fraudulent financial reporting. This legislation requires the principal executive officer and the principal financial officer of publicly traded companies to certify the appropriateness of the financial statements and disclosures in each quarterly and annual report that their company issues. These officers are also responsible for establishing and maintaining internal controls within the company. Further, they must disclose to auditors and the audit committee of

the board of directors any fraud, whether or not material, involving management or employees who have a significant role in defining or implementing internal controls. As this law goes into effect, evaluation and reporting of a company's internal controls and financial statements in order to detect fraud becomes even more critical, and must be on-going.

Prior research shows that various kinds of decision aids may be used to assist the auditor in detecting financial reporting fraud. Bell, Szykowny, and Willingham (1993) used bivariate and cascaded logit to assess the likelihood of management fraud. Their model achieved within-sample correct classification of 97% on the fraud observations and 75% on the non-fraud observations. Hansen, McDonald, Messier, and Bell (1996) used a generalized qualitative response model to predict management fraud. They reported 89.3% predictive accuracy over 20 trials. Bell and Carcello (2000) developed a logistic regression model as a decision aid to assist in the auditor's fraud decision. Auditors may also use an expert system as a decision aid to assist in fraud determination. Eining, Jones, and Loebbecke (1997) examined the effect that the use of an expert system has on auditor decision-making ability in detecting fraud. Their research showed that in allowing the interaction between the auditor and the system, the expert systems that have been used to assist auditors in complex decision processes often give results that are more accurate and consistent. Similarly, Whitecotton and Butler (1998) found that allowing decision makers to select information for the decision aid increases decision aid reliance. Fuzzy clustering may also be used as a decision aid for an auditor to detect fraudulent financial reporting (Lenard & Alam, 2004).

Fuzzy Clustering

When available data does not suggest a clear answer, decision makers often look for patterns or groups in the underlying data to make a decision (Alam, Booth, Lee, & Thordarson, 2000). While discriminant analysis and logistic regression assign observations to groups that were defined in advance, cluster analysis is the art of finding groups in data (Kaufman & Rousseeuw, 1990). Fuzzy set theory, introduced by Zadeh (1965), attempts to classify subjective reasoning (e.g., a human description of "good", "very good", or "not so good") and assigns degrees of possibility in reaching conclusions (Lenard, Alam, & Booth, 2000). As opposed to hard clustering, where there is a clear-cut decision for each object, fuzzy clustering allows for ambiguity in the data by showing where a solution is not clearly represented in any one category or cluster. Fuzzy clustering shows the degree to which (in terms of a percentage) an item "belongs" to a cluster of data. In other words, a data item may belong "partially" in each of several categories. The strength of fuzzy analysis is this ability to model partial categorizations.

Lau, Wong, and Pun (1999) used neural networks and fuzzy modeling to control a plastic injection-molding machine. They suggested that the neural network and fuzzy technology complement each other and offset the pitfalls of computationally intelligent technologies. Alam et al. (2000) used a combination of fuzzy clustering and self-organizing neural networks, and were successful in identifying potentially failing banks. Ahn, Cho, and Kim (2000) reported results using these technologies to predict business failure, and stressed the importance of these predictions as useful in aiding decision makers. Lenard et al. (2000) used fuzzy clustering to identify two different categories of bankruptcy. Companies placed in the second bankrupt category exhibited more extreme values in terms of the financial ratios used in the study. Companies either showed much better results (such as a high current ratio) than would be expected for a company facing bankruptcy, or the companies showed very poor results, such as a much higher debt ratio than any of the other bankrupt companies in the data sample. Lenard and Alam (2004) operationalized a fuzzy logic model for fraud detection in an Excel spreadsheet. By using the fuzzy logic model to develop clusters for different statements representing red flags in the detection of fraud, non-financial data was included with financial statement variables for the analysis. The overall prediction accuracy for the model was 86.7%.

Nolan (1998) used expert fuzzy classification and found that fuzzy technology enables one to perform approximate reasoning, as when a student assignment is graded as "very good", or "not so good", and improves performance in three ways. First, performance is improved through efficient numerical representation of vague terms, because the fuzzy technology can numerically show representation of a data item in a particular category. The second way performance is enhanced is through increased range of operation in ill-defined environments, which is the way that fuzzy methodology can show partial membership of data elements in one or more categories that may not be clearly defined in traditional analysis. Finally, performance is increased because the fuzzy technology has decreased sensitivity to "noisy" data, or outliers. Ammar, Wright, and Selden (2000) used a multilevel fuzzy rule-based system to rank state financial management. The authors used fuzzy set theory to represent imprecision in evaluated information and judgments. Pathak, Viyarthi, and Summers (2003) developed a fuzzy logic based system for auditors to identify fraud in settled claimed insurance. They believe that their system was able to cut costs by detecting fraudulent filings.

CRITICAL ISSUES OF FUZZY LOGIC

The fuzzy clustering procedure used by Lenard et al. (2000) and Lenard and Alam (2004) is called FANNY (Kaufman & Rousseeuw, 1990). The program FANNY uses "fuzzi-

ness" to partition objects by avoiding "hard" decisions, or clustering into fixed, definite categories. For each item in the dataset, the algorithm provides k+1 pieces of information, where k is the number of clusters that are used in the clustering algorithm. The k+1 pieces of information are: U_{iv}, the membership coefficient of item i in cluster v, v = 1 ... k, and S_i, the silhouette coefficient of item i. A higher value of U_{iv} indicates a stronger association of item i and cluster v. The silhouette coefficients satisfy the constraints $-1 \le S_i \le 1$ and indicate how a well-clustered object uses average distances from its own cluster to the closest neighboring clusters. The closer S_i is to 1 the better the clustering of an individual item. A value of S_i close to -1 indicates that an item may be assigned to more than one cluster (Alam et al., 2000). The Euclidean distance measure is used to compute distances between objects and to quantify the degree of similarity for each object. The degree of similarity for each objects *i* and *j* is computed as follows (Kaufman & Rousseeuw, 1990)

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_p - x_p)^2} \quad (1)$$

where the p^{th} measurement of the i^{th} object is given by x_{ip} and *d(i,j)* is the actual distance between objects *i* and *j*.

Several authors have expanded upon the fuzzy clustering algorithms. Van den Bergh and van den Berg (2000) developed a competitive learning algorithm using fuzzy frequency distributions. They emphasized that sometimes the discovery of exceptions is more important than the main rules. For example, profit opportunities often seem to appear randomly and infrequently, so the agent should concentrate on detecting the unusual, abnormal states, or exceptions, rather than the average normal states (van den Bergh & van den Berg, 2000). Their algorithm seeks to find a mapping from an *M*-dimensional input space X into an *N*-dimensional output space *Y*, given a representative data set *S*. The set *S* contains *P* data pairs (x_p; y_p). The final formula for the mapping is depicted as

$$y_b = \sum_{c=1}^{C_y} \overline{y_c} \; X \overline{u}_c^{Y} \mid \overline{u}_b^{X} \quad (2)$$

Thus, the sum of output cluster centroids y_c is weighted by the "local" membership values $\overline{u}_c^{Y} \mid \overline{u}_b^{X}$.

Fuzzy clustering algorithms have also been extended by the work of Kaymak and Setnes (2000), who proposed fuzzy clustering algorithms with volume prototypes and similarity based cluster merging. These extensions reduce sensitivity of the clustering algorithms to bias in data distribution, and help determine the number of clusters automatically.

Finally, Mashor (2001) proposed a clustering algorithm called adaptive fuzzy *c*-means clustering. In this method, each data sample is assigned a membership grade to indicate the degree of belonging to each center rather than assigning the data sample to one center as in "hard" clustering algorithms like the *k*-means clustering. In addition, the clustering program is not as sensitive to initial centers and gives better performance. The algorithm only requires the data to be presented once, instead of requiring multiple presentations, and as such reduces the computational load.

Identifying the number of clusters in fuzzy clustering is a challenging task. The optimal clustering should consider both fuzzy compactness and separation. The current state of the art does not provide a theoretical basis for an optimal choice of clusters. The objective function based fuzzy clustering algorithms are often used to divide the data into a predetermined number of clusters. The fuzzy *c*-means algorithm is one of the most popular objective functions used for fuzzy clustering. It uses the similarity between objects to measure the distances between clusters. The validity of the clusters is often assessed after the clusters have been formed. Validity Measures typically address the issues of the compactness of the clusters and the distances between them (e.g., Pal & Bezdek, 1995; Xie & Beni, 1991). Gath and Geva (1989) proposed a validity measure, which is a ratio between fuzzy compactness and separation. Bensaid et al. (1996) argued that combining the validity guided clustering algorithm with the fuzzy *c*-means considerably improves the partitions generated by fuzzy *c*-means alone. Various other studies have addressed the issue of the number of clusters, but there is no generally accepted approach of a priori selecting the appropriate number of clusters. Cluster validity tests are the only means available to decide whether the number of clusters used captures the underlying characteristics of the data. Investigating and resolving these issues is crucial in the analysis of financial statement data. The assignment to a particular cluster would determine whether or not the data being analyzed is a high "red flag" indicator of fraud.

FUTURE TRENDS

In addition to fraud determination, there are also other decisions that accounting and financial personnel make that affect financial reporting. Specifically, in the field of auditing, there is the auditor's decision reflected in the audit report about whether the company can continue as a going concern. The auditor uses financial statement and non-financial statement data, and a framework of questions to make the going concern judgment. The auditor also applies judgment in the consideration of materiality. Materiality judgment is closely linked to the analysis of fraud because the auditor must decide the extent to which a discrepancy affects the credibility of financial statements. These decisions may be enhanced by the use of statistical models, expert reasoning,

data mining tools, and now fuzzy logic, to provide support for the auditor's judgment.

CONCLUSION

Financial statement information is used by management, employees, outside analysts, investors and creditors to assess the health of publicly traded companies. Just as this information is now readily available through the Internet and online financial services, so should tools that help in the analysis of that information be readily available or easily obtained. As the different methods of fuzzy analysis become more prevalent, there will be additional opportunities for using fuzzy logic in various other applications.

REFERENCES

Ahn, B.S., Cho, S.S., & Kim, C.Y. (2000). The integrated methodology of rough set theory and artificial neural network for business failure prediction. *Expert Systems with Applications, 18*, 65-74.

Alam, P., Booth, D., Lee, K., & Thordarson, T. (2000). The use of fuzzy clustering and self-organizing neural networks for identifying potentially failing banks: An experimental study. *Expert Systems with Applications, 18*, 185-99.

American Institute of Certified Public Accountants (AICPA) (2002). *Consideration of fraud in a financial statement audit. Statement on Auditing Standards No. 99.* New York: AICPA.

Ammar, S., Wright, R., & Selden, S. (2000). Ranking state financial management: A multilevel fuzzy rule-based system. *Decision Sciences, 31*(2), 449-481.

Arnold, J. (2003). Worries mount for Ahold. BBC News. *http://news.bbc.co.uk/1/hi/business/2797097.stm*. February 26.

Bell, T.B. & Carcello, J.V. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory, 19*(1), 169-184.

Bell, T.B., Szykowny, S., & Willingham, J.J. (1993). *Assessing the likelihood of fraudulent financial reporting: A cascaded logit approach.* Working Paper, KPMG Peat Marwick, Montvale, NJ.

Bensaid, A.M., Hall, L.O., Bezdek, J.C., Clarke, L.P., Sibliger, M.L. Arrington, J.A., & Murtagh, R.F. (1996). Validity-guided (Re) clustering with applications to image segmentation. *IEEE Transactions on Fuzzy Systems, 4*(2), 112-123.

Eining, M., Jones, D.R., & Loebbecke, J.K. (1997). Reliance on decision aids: An examination of auditors' assessment of management fraud. *Auditing: A Journal of Practice & Theory, 16*(2), 1-19.

Gallani, A., & Trofimov, Y. (2004). Behind Parmalat chief's rise: Ties to Italian power structure. *Wall Street Journal*, March 8, A1.

Gath, J., & Geva, A.B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(7), 32-57.

Hansen, J.V., McDonald, J.B., Messier, Jr., W.F., & Bell, T.B. (1996). A generalized qualitative-response model and the analysis of management fraud. *Management Science, 42*(7), 1022-1032.

Kaufman, L., & Rousseeuw, P.T. (1990). *Finding groups in data: An introduction to cluster analysis.* New York: John Wiley.

Kaymak, U., & Setnes, M. (2000). Extended fuzzy clustering algorithms. *ERIM Report Series in Management, 51*, 1-24.

KPMG. (2003). *Fraud survey 2003.* Montvale, NJ.

Lanza, R.B. (2002). New audit standard approved-SAS 99 "consideration of fraud in financial statement audit." *http://www/aicpa.org/pubs/tpcpa/nov2002/anti fraud.htm*

Lau, H.C.W., Wong, T.T., & Pun, K.F. (1999). Neural-fuzzy modeling of plastic injection molding machine for intelligent control. *Expert Systems with Applications, 17*, 33-43.

Lenard, M.J., & Alam, P. (2004). The use of fuzzy logic and expert reasoning for knowledge management and discovery of financial reporting fraud. In H.R. Nemati & C.D. Barko (Eds.), *Organizational data mining: Leveraging enterprise data resources for optimal performance* (pp. 230-262). Hershey, PA: Idea Group Publishing.

Lenard, M.J., Alam, P., & Booth, D. (2000). An analysis of fuzzy clustering and a hybrid model for the auditor's going concern assessment. *Decision Sciences, 31*(4), 861-864.

Mashor, M.Y. (2001). Adaptive fuzzy *c*-means clustering algorithm for a radial basis function network. *International Journal of Systems Science, 32*(1), 53-63.

Nolan, J.R. (1998). An expert fuzzy classification system for supporting the grading of student writing samples. *Expert Systems with Applications, 15*, 59-68.

Pal, N.R., & Bezdek, J.C. (1995). On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy Systems, 3*, 370-379.

Pathak, J., Viyarthi, N., & Summers, S.L. (2003). *A fuzzy-based algorithm for auditors to detect element of fraud in*

settled insurance claims. Working paper, University of Windsor.

Van den Bergh, W-M., & van den Berg, J. (2000). Competitive exception learning using fuzzy frequency distributions. *ERIM Report Series Research in Management, 6,* 1-12.

Whitecotton, S.M., & Butler, S.A. (1998). Influencing decision aid reliance through involvement in information choice. *Behavioral Research in Accounting, 10*(Supplement), 182-201.

Xie, X.L., & Beni, G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 13*(8), 841-847.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control, 8*, 338-353.

KEY TERMS

Cluster Analysis: Defining groups based on the "degree" to which an item belongs in a category. The degree may be determined by indicating a percentage amount.

Data Mining: Using powerful data collection methods to analyze a company's database or data stores and select information that supports a specific objective.

Expert Reasoning: Implementing rules or procedures, often programmed to occur automatically, in order to make a decision. Background and heuristics that identify how to reach a conclusion are based on the knowledge of human experts in that field.

Fraudulent Financial Reporting: Intentional or reckless conduct, whether by act or omission, that results in materially misleading financial statements.

Fuzzy Logic: A mathematical technique that classifies subjective reasoning and assigns data to a particular group, or cluster, based on the degree of possibility the data has of being in that group.

Internal Controls: Procedures applied by a business organization that ensure information is safeguarded, that it is accurate and reliable, and that it is processed efficiently and in accordance with management's prescribed policies.

Management Fraud: A situation in which management misrepresents the financial condition of their firm. They may do so for personal financial gain or to disguise the financial results of their company.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 135-139, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Application Service Provision for Intelligent Enterprises

Matthew W. Guah
Warwick University, UK

Wendy L. Currie
Warwick University, UK

ROAD TO ASP

Several historical shifts in information systems (IS) involved strategies from a mainframe to a client server, and now to application service provision (ASP) for intelligent enterprises. Just as the steam, electric, and gasoline engines became the driving forces behind the industrial revolution of the early 1900s, so the Internet and high-speed telecommunications infrastructure are making ASP a reality today. The current problem with the ASP model involves redefining success in the business environment of the 21st century. Central to this discussion is the idea of adding value at each stage of the IS life cycle. The challenge for business professionals is to find ways to improve business processes by using Web services.

It took mainframe computers a decade or two to become central to most firms. When IBM marketed its first mainframe computer, it estimated that 20 of these machines would fulfil the world's need for computation! Minicomputers moved into companies and schools a little faster than mainframes, but at considerably less costs. When the first computers were applied to business problems in the 1950s, there were so few users that they had almost total influence over their systems. That situation changed during the 1960s and 1970s as the number of users grew. During the 1980s the situation became even tighter when a new player entered the picture—the enterprise (McLeord, 1993). In the 21st century, information systems are developed in an enterprise environment (see Diagram 1).

Beniger (1986) puts forth a seemingly influential argument that the origin of the information society may be found in the advancing industrialisation of the late nineteenth century. The Internet is simply a global network of networks that has become a necessity in the way people in enterprises access information, communicate with others, and do business in the 21st century. The initial stage of e-commerce ensured that all large enterprises have computer-to-computer connections with their suppliers via electronic data interchange (EDI), thereby facilitating orders completed by the click of a mouse. Unfortunately, most small companies still cannot afford such direct connections. ASPs ensure access to this service costing little, and usually having a standard PC is sufficient to enter this marketplace.

The emergence of the ASP model suggested an answer to prevailing question: Why should small businesses and non-IT organisations spend substantial resources on continuously upgrading their IT? Many scholars believed that outsourcing might be the solution to information needs for 21st century enterprises (Hagel, 2002; Kern, Lacity & Willcocks, 2002; Kakabadse & Kakabadse, 2002). In particular, the emergence of the ASP model provided a viable strategy to surmount the economic obstacles and facilitate various EPR systems adoption (Guah & Currie, 2004). Application service provision—or application service provider—represents a business model of supplying and consuming software-based services over computer networks. An ASP assumes responsibility of buying, hosting, and maintaining a software application on its own facilities; publishes its user interfaces over the networks; and provides its clients with shared access to the published interfaces. The customer only has to subscribe and receive the application services through an Internet or dedicated intranet connection as an alternative to hosting the same application in-house (Guah & Currie, 2004). ASP is an IT-enabled change, a different and recent form of organisational change, evidenced by the specific information systems area (Orlikowski & Tyre, 1994). ASP has its foundations in the organisational behaviour and analysis area (Kern et al., 2002).

The initial attempt—by the ASP industry to take over the business world—was fuelled by the belief that utility computing offered a new business model to customers, similar to electricity, gas, and water. The commercialization of the Internet meant that, as network traffic increased in a firm's data centre, IT architecture would trigger other resources into action, including idle servers, applications, or pools of network storage. The firm would pay only for the amount of time it used the services. Thus, the concept of 'software-as-a-service' was created (Kakabadse & Kakabadse, 2002). Accessing IT resources in this way would result in reduced up-front investment and expenditure, enabling firms to buy services on a variable-price basis (Dewire, 2000). This fuelled opportunities in the late 1990s for service provid-

ers to offer software applications and IT infrastructure on a rental, pay-as-you-go pricing model (Bennet & Timbrell, 2000). An ASP could be a commercial entity, providing a paid service to customers (Dussauge, Hart & Ramanantsoa, 1994) or, conversely, a not-for-profit organisation supporting end users (Currie, Desai & Khan, 2003).

ASP AREAS OF CONCERN

As evidence relating to the reality and basic features of the ASP market continues to grow, there begins to be less concern about confirming that any structural economic shift has continued historically, and more concern about understanding how the ASP industry is performing, and its impacts on productivity, investment, corporate capital formation, labour force composition, and competition.

The ASP business model is premised on the formation of strategic alliances and partnerships with technology and service providers (Ferergul, 2002). Telecommunications firms entering the ASP market with large IT infrastructures needed to partner with ISVs and hardware manufacturers. One of the significant strategic alliances was between Cable & Wireless (IT infrastructure), Compaq (hardware manufacturer), and Microsoft (ISV). Pure-play ASPs without a large investment in IT infrastructure needed to form strategic alliances with data centre and co-locator firms (telcos) and ISVs. Some of the major reasons for businesses to implement an ASP business model are list in Table 1.

The ASP model was highly volatile, dynamic, and immature. A recent review of the ASP industry concluded that technological factors like scalability, the managerial aspects of speed and focus, and the behavioural aspects of price and flexibility were the key drivers of the model. The inhibitors of the model were poor connectivity, lack of trust in the model, reluctance to be locked into long-term contracts with suppliers, lack of customisation, poor choice and suitability of software applications from ASPs, and few opportunities to integrate disparate applications across technology platforms and business environments. These factors and others led Hagel (2002, p. 43) to conclude:

"ASPs in many respects represented a false start in the efforts to break out of the enterprise straitjacket. In particular, few of them adopted Web services architectures as their technology platform. Instead, they attempted to build businesses on the Internet using traditional technology architectures…this proved to be a significant flaw in the early ASP model and explains many difficulties these businesses experienced."

The business environment for intelligent enterprises (see Diagram 1) includes the enterprise itself and everything else that affects its success, such as competitors; suppliers; customers; regulatory agencies; and demographic, social, and economic conditions (Guah & Currie, 2004). As a strategic resource, ASP helps the flow of various resources from the elements to the enterprise, and through the enterprise and back to the elements.

THE FUTURE OF THE ASP MODEL

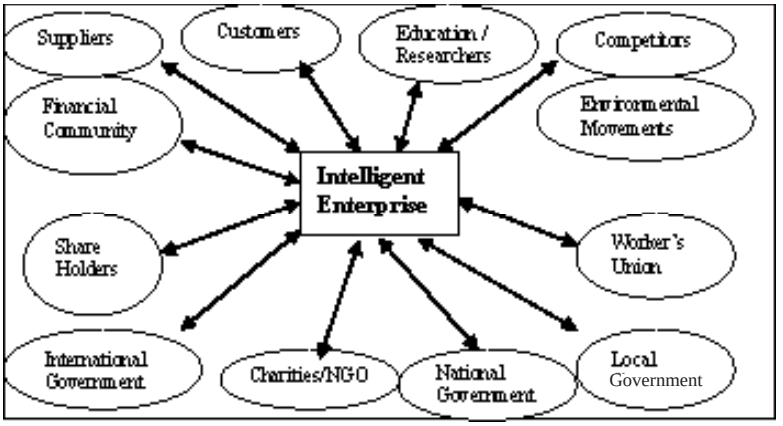
According to Forester Research, the proportion of ASP business in the outsourcing market peaked at about $800 million in 2000 and was projecting for $25 billion by 2005. However, it actually declined by the year 2002 (due partly to the effect of stock market collapse) and currently is being projected at $15 billion by 2006. The overall business interests in the ASP model will continue to rise, with proportionally higher rates of investment by vendors versus traditional outsourcing. We attribute this optimistic forecast to four trends:

- continuing improvements in capabilities and cost-performance characteristics of Remote Support Services by vendors,
- improvements in capabilities and cost-performance characteristics of the technology at the system or application level,

Table 1. A list of motivational factors to implement an ASP strategy

• To take maximise the capabilities of the Internet's latest technology
• To increase sales of products and services
• To reach a highly desirable demographic market
• To stay on top of competition
• To make changing information available quickly
• To test new products and services on the market
• To boast morale among staff and customers
• To experiment with an Internet model to business IT outsourcing

Diagram 1. A tool for controlling influences in a complex environment



- continual development of the telecommunications infrastructure to support ASP performance, and
- gradual reduction of institutional and social barriers to the introduction of the ASP model as a viable business strategy.

There are numerous papers warning that such accelerated Web service evolution increases the difficulty that other competitors have in adapting to ASP (Gottchalk, Graham, Kreger & Snell, 2002; Hondo, Nagaratnam & Nadalin, 2002; Stencil Group, 2002). By modifying the nature and the relative importance of the key factors for success in the ASP industry, Web service technological changes that are introduced by one or more of the vendors can lead to favourable changes in the competitive environment. In an industry built upon high volume, new technologies that are introduced by some of the competitors that nullify or minimise the impact of scale can significantly alter the nature of the competitive environment by making size a drawback rather than an advantage.

Diagram 2 shows that the holistic approach to technology always seems to work better than the piece-meal approach to information systems solution. Early stages of Web services are represented by a two-legged table. The current version of Web services being practiced by vendors after the dot.com crash is represented by the three-legged table in Diagram 2. But an even more successful model of Web services would be a properly architecture four-legged table, represented above. The analogy here is that a two-legged table is less stable than a three-legged table, while a four-legged table is even firmer.

CONCLUSION

We can safely conclude that policy makers in all fields, not just in IS, are forced into ill-considered conclusions and recommendations because they still view their management strategies in pre-Internet terms. Moreover, they are still constrained by statistical calculations based on outmoded and obsolete classification approaches, as well as on invalid assumptions about the fundamental sources of profit and capital formation—without full consideration for business environment.

Rethinking your business in terms of Internet economy, formulating new strategies for gaining competitive advantage, and raising the level of awareness of people throughout your enterprise to the notion that information itself can and should be looked upon as a strategic corporate asset¾these are great steps, but only the first steps for success in the 21st century. In addition, both structural and procedural changes must take place for an intelligent enterprise to put its convictions into operation. Could ASP provide you with such a necessary tool thereby directing your focus into the reality of a 21st century intelligent organisation?

REFERENCES

Beniger, J.R. (1986). *The control revolution: Technological and economic origins of the information society*. Boston: Harvard University Press.

Bennett, C. & Timbrell, G.T. (2000). Application service providers: Will they succeed? *Information Systems Frontiers*, *2*(2), 195-211.

Table 3.Summary of areas affecting the growth of the ASP market

Widespread model ignorance and perceptions among small and medium businesses. Lack of adequate understanding of the ASP business model and its understanding.

IS infrastructure raises a broad range of economic, social, and technical issues. Who should pay for infrastructure? Who should have access to/control over them and at what cost? Which technology should it include? Where ASP is involved, the economic question often puts telephone companies against cable companies, both of whom can provide similar capabilities for major parts of the telecommunications system.

Telecommunications facilitates ASP emancipation. Telecommunications has become virtually inseparable from computer with a paired value that is vital for integrating enterprises. As an e-commerce phenomenon, a few of the essentials of an ASP infrastructure are Common Carriers, Value-Added Networks, Private Line, and Private Networks.

Issues of security. As a program executed upon accessing a Web page, ASP carries a security risk because users end up running programs they don't know and/or trust. The latest solution is encryption, but does any mathematical encryption guarantee absolute security? No. Just as a physical lock cannot provide absolute safety, encryption cannot guarantee privacy—if a third party uses enough computers and enough time, they will be able to break the code and read the message. The encryption system only guarantees that the time required to break the code is so long that the security provided is sufficient. When someone asserts that an encryption scheme guarantees security, what they actually mean is that although the code can be broken, the effort and time required is great. Thus, an encryption scheme that requires a longer time to break than another scheme is said to be 'more secure'.

Overcoming organisational obstacles to a commercial future. The issue of organisation is based on the way ASP has evolved. The ASP industry lacks the type of clear organisation that would make it easy to use as a reliable and profitable business model.

ASP as competitive investment. Not so much return on investment for ASP, rather what is the cost of not investing? Will an enterprise lose customers and market share because it does not have a particular technology in place? Can you enter a new line of business without investing in the technology that competitors have adopted? What kinds of services do customers expect? These are competitive investment issues raised by ASP.

Change is also an opportunity. One stimulus for ASP solution implementation is its intention to transform the enterprise. In this sense, the investment in ASP is part of a larger change programme that is meant to enable intelligent enterprises' virtual and multiple-team structures. The resulting contributions can be described as part of the outcome of a general change effort.

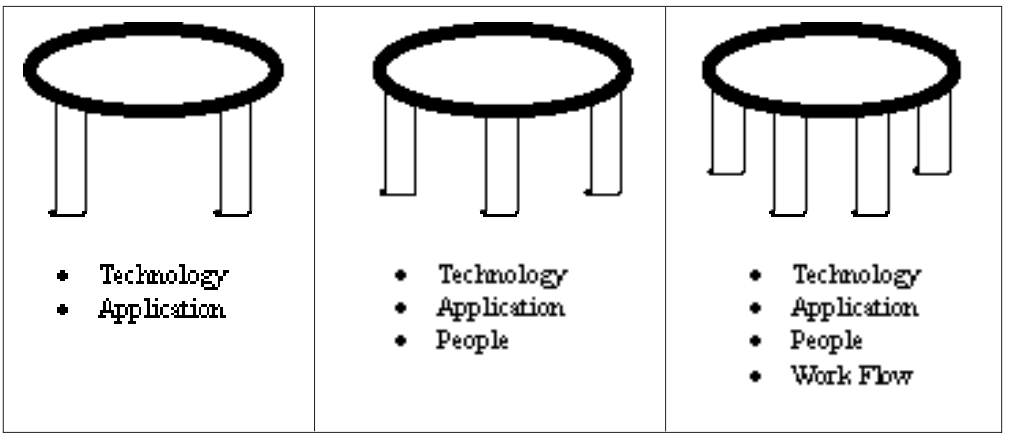
Social-technical issues. The social objectives refer to the expectations of major stakeholders (i.e., employees). An ASP business model that provides information and tools for employees increases involvement because they reinforce the employee's authority and responsibility for work. Those that provide information to managers or quality inspectors but do not support employees can reduce involvement by reinforcing the suspicion that the employee is not really responsible.

Tangible and intangible benefits. The tangible benefits of an ASP solution can be measured directly to evaluate system performance. Examples include reduction in the time for completion of transaction, improvement in response time, reduction in the cost of assets, and reduction in the error rate. Intangible benefits affect performance but are difficult to measure because they refer to comparatively vague concepts. A few intangible benefits of a solution are: better coordination; better supervision; better morale; better information for decision making; ability to evaluate more alternatives; ability to respond quickly to unexpected situations; and organisational learning. Although hard to quantify, intangible benefits are important and shouldn't be ignored, as many IS benefits to organisations are intangible.

The role of government. Modernisation of the machinery and process of government will accommodate many Internet strategies like ASP. This will involve reform of intellectual property law to accommodate access to and exploitation of works via the Internet. It will include administration of Internet domain names on an international basis. Finally, the facilitation of e-commerce development will include national and international initiatives and measures to protect both suppliers and consumers operating within this global electronic marketplace.

Blurring in-house IT and ASP services. As the industry evolves and becomes more complex, the need for new services and specialisation in the division of labour continues to increase. In-house migrates into strategic management and monitoring of IT standard, while ASP migrates into value-added services so that 'business IT becomes a service in a package form'. As the boundaries between in-house and ASP become more blurred through the use of improved communications technologies, the opportunities for entrepreneurs continue to increase.

Diagram 2. Evolution of Web services



Currie, W., Desai, B., Khan, N., Wang, X. & Weerakkody, V. (2003, January). Vendor strategies for business process and applications outsourcing: Recent findings from field research. *Proceedings of the Hawaii International Conference on Systems Sciences,* Hawaii.

Dewire, D.T. (2000). Application service providers. *Information Systems Management, 17*(4), 14-19.

Dussauge, P., Hart, S. & Ramanantsoa, B. (1994). *Strategic technology management: Integrating product technology into global business strategies for the 1990s.* Chichester: John Wiley & Sons.

Ferergul, C. (2002). Best practices in Web hosting service level agreements. Stamford, CT: Meta Group. Retrieved May 2, 2002, from *techupdate.zdnet.com/techupdate/stories/main/*

Gottschalk, K., Graham, S., Kreger, H. & Snell, J. (2002). Introduction to Web services architecture. *IBM Systems Journal, 41*(2).

Guah, M.W. & Currie, W.L. (2004). Application service provision: A technology and working tool for healthcare organizations in the knowledge age. *International Journal of Healthcare Technology and Management, 6*(1/2), 84-98.

Hagel III, J. (2002). *Out of the box: Strategies for achieving profits today and growth tomorrow through Web services.* Boston: Harvard Business School Press.

Hondo, M., Nagaratnam, N. & Nadalin, A. (2002). Securing Web services. *IBM Systems Journal, 41*(2).

Kakabadse, N. & Kakabadse, A. (2002). Software as a service via application service providers (ASPs) model of sourcing: An exploratory study. *Journal of Information Technology Cases and Applications, 4*(2), 26-44.

Kern, T., Lacity, M. & Willcocks, L. (2002). *Netsourcing: Renting business applications and services over a network.* New York: Prentice-Hall.

McLeord Jr., R. (1993). *Management information systems: A study of computer-based information systems* (5th edition). New York: Macmillan.

Orlikowski, W.J. & Tyre, M.J. (1994). Windows of opportunity: Temporal patterns of technological adaptation in organizations. *Organization Science*, (May), 98-118.

Stencil Group. (2002). Understanding Web services management: An analysis memo. Retrieved May 2002 from *www.stencilgroup.com*

KEY TERMS

ASP: A third-party service firm that deploys, manages, and remotely hosts software applications through centrally located services in a rental or lease agreement (ASP Consortium, 2000). Such application deliveries are done to multiple entities from data centres across a wide area network (WAN) as a service rather than a product, priced according to a license fee and maintenance contract set by the vendor. An ASP is considered by many to be the new form of IT outsourcing, usually referred to as application outsourcing.

ASP Aggregator: The ASP aggregator model is based on the premise that the rapid proliferation of firms offering ASP services has created an overly complex market for medium-sized enterprises to deal with when investigating application outsourcing options.

Agent-Based Approach to ASP: This approach to ASP is well equipped to address the challenges of multi-market package to e-procurement. Service agents within the ASP model are the system's gateway to external sources of goods and services. These agents are usually aware of the source's market model and of the protocols it uses (Zhang et al., 2000). Service agents are not only able to determine which requests it can service, but also proactively read these requests and try to find an acceptable solution.

Common Carriers: Companies that are licensed, usually by a national government, to provide telecommunications services to the public, facilitating the transmission of voice and data messages.

Infrastructure: An emerging class of companies have opted to approach the ASP market by providing infrastructure management and outsourcing services to ASPs, freeing up their resources to focus more directly on application management issues (telco, data centre, networking).

Internet Service Providers (ISP): Provides access to the Internet via different communications channels such as traditional telephone lines or a high-speed fibre optics channel.

Pure-Play ASP: Those with non-specific industry-required product or service, except Internet/Web-enabled software applications (e-mail/security/disaster recovery). Firms offering this service suffer from the sales of unprofitable commodity applications and therefore have greater reliance on venture capitalist funding as they operate in a rather unstable, volatile, and dynamic market.

Vertical Service Provider (VSP): Vertically focused ASPs offering industry-specific applications are also emerging. Their basic premise is that each industry (health, finance, transportation) has its own unique set of characteristics that can best be served by companies that focus exclusively on the given industry.

Virus: A malicious code added to an e-mail program or other downloadable file that is loaded onto a computer without the user's knowledge and which runs often without the user's consent. Computer viruses can often copy themselves and spread themselves to a user's e-mail address book or other computers on a network.

Web Services: Web Services technology is one of the most important foundations for ASP new-game strategies. Thus, by accelerating the pace of Web services in the industry, a competitor with good capability in the technology reinforces its own competitive position.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 140-145, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Applications for Data Mining Techniques in Customer Relationship Management

Natalie Clewley
Brunel University, UK

Sherry Y. Chen
Brunel University, UK

Xiaohui Liu
Brunel University, UK

INTRODUCTION

With the explosion in the amount of data produced in commercial environments, organizations are faced with the challenge of how to collect, analyze, and manage such large volumes of data. As a consequence, they have to rely upon new technologies to efficiently and automatically manage this process. Data mining is an example of one such technology, which can help to discover hidden knowledge from an organization's databases with a view to making better business decisions (Changchien & Lu, 2001).

Data mining, or knowledge discovery from databases (KDD), is the search for valuable information within large volumes of data (Hand, Mannila & Smyth, 2001), which can then be used to predict, model or identify interrelationships within the data (Urtubia, Perez-Correa, Soto & Pszczolkowski, 2007). By utilizing data mining techniques, organizations can gain the ability to predict future trends in both the markets and customer behaviors. By providing detailed analyses of current markets and customers, data mining gives organizations the opportunity to better meet the needs of its customers.

With such significance in mind, this chapter aims to investigate how data mining techniques can be applied in customer relationship management (CRM). This chapter is organized as follows. Firstly, an overview of the main functionalities data mining technologies can provide is given. The following section presents application examples where data mining is commonly applied within the domain, with supporting evidence as to how each enhances CRM processes. Finally, current issues and future research trends are discussed before the main conclusions are presented.

BACKGROUND

Data mining methods can generally be grouped into four categories: classification, clustering, association rules and information visualization. The following subsections will describe these in further detail.

Classification

Databases are full of hidden information that can help to make important business decisions. Classification involves using an algorithm to find a model that describes a data class or concept (Han & Kamber, 2006). By identifying a series of predefined labels, items can be categorized into classes according to its attributes (e.g., age or income). Thus, it is a useful technique for identifying the characteristics of a new item. For example, in the case of a bank loan clerk, classification is useful for predicting whether loan applicants are a "safe" or "risky" investment for the bank based on the class that they belong to. Popular classification techniques include Decision Trees and Bayesian Networks.

Clustering

Where classification is thought of as a supervised learning technique because it uses a set of predefined class labels, clustering is an unsupervised learning technique. Because no assumptions are made about the structure of the data, clustering can uncover previously hidden and unexpected trends or patterns. Clustering involves grouping items into "natural" clusters based on their similarities (Hand et al., 2001). Each item in a cluster is similar to those within its cluster, but dissimilar to those items in other clusters. In this way, clustering is commonly used to identify customer affinity groups with the aim of targeting with specialized marketing promotions (section 3.2.2). Common clustering techniques include K-means and Kohonen Networks.

Association Rules

Association rules are mainly used to find relationships between two or more items in a database. Association rules are

expressed in the form (X→Y), where X and Y are both items. In a set of transactions, this means that those containing the items X, tend to contain the items Y. Such an association rule is usually measured by *support* and *confidence*, where the *support* is the percentage of both X and Y contained in all transactions and the *confidence* is calculated by dividing the number of transactions supporting the rule by the number of transactions supporting the rule body (Zhang, Gong & Kawamura, 2004). For example, this technique is commonly used to identify which items are regularly purchased together or to identify the navigational paths of users through an online store. The discovery of such relationships can help in many business decisions, such as customer shopping behavior analysis, recommendations, and catalog design (Han & Kamber, 2006).

Information Visualization

Information visualization is based on an assumption that human beings are very good at perceiving structure in visual forms. The basic idea is to present the data with some graphics, for example, 2D graphics and 3D graphics, allowing the human to gain insight from the data, draw conclusions, and directly interact with the data (Ankerst, 2001). Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary (Lopez, Kreuseler & Schumann, 2002). This approach is especially useful when little is known about the data and the exploration goals are vague, for example, analyzing the path of customers through an online store.

Data Mining and Its Applications in CRM

Customer relationship management (CRM) is thought to be an increasingly important success factor for e-business. CRM is the process of managing interactions between a company and its customers. Initially, this involves segmenting the market to identify customers with high-profit potential, from which marketing strategies are designed to favorably impact the behavior of customers in these segments (Berson, Smith & Thearling, 2000).

Customer Analysis

Customer Segmentation

Customer segmentation is the process of splitting customers into homogeneous groups based on their common attributes (Böttcher, Spott, Nauck & Kruse, 2007). More specifically, clustering algorithms are used to discover groups of customers based on such attributes. By doing so, each cluster contains customers who share similar attribute values. Benefits of clustering in this way include a deeper understanding of the needs and behaviors of individual customer groups, allowing the implementation of systems for the purpose of personalized recommendations (section 3.2.1) and targeted marketing (section 3.2.2). For example, Vellido, Lisboa, and Meehan (1999) segmented customers from the data gathered in an online survey. Initially, they conducted a factor analysis of the observable data and then the factor scores were clustered by using the Self Organizing Maps. Five clusters were discovered: cost conscious, complexity avoiders, unconvinced, convinced and security conscious. These clusters allowed researchers to identify which groups of customers frequently used online shopping and those that would be more open to specific marketing campaigns.

However, even though customers can be easily segmented into groups in such a manner, today's fluctuating markets mean that segmentation often becomes obsolete very quickly. Ha (2007) overcame this by providing a method that monitors constantly changing customer needs. After performing Kohonen Network clustering to divide the customers into four dominant clusters, the author focused on keeping track of customer shifts among the segments to monitor the changes over time. Behavior patterns of customers in each of the segments were then predicted through the use of transition paths.

In addition, Böttcher et al. (2007) present a system for customer segmentation which accounts for the dynamic nature of a market based on "interestingness". Their system focuses on the discovery of frequent item-sets and the analysis of their change over time, which provides detailed knowledge about how customer behavior evolves over time. They successfully applied their system to two problem domains in a telecommunications company: customer analytics and network usage. The former aimed to identify the factors likely to drive customer satisfaction in the future, whereas the latter aimed to understand the drivers of change in customer behavior whilst using the services.

Click Stream Analysis

In addition to grouping the types of users through transaction or personal data, click streams, the paths visitors take through a website, also provide valuable information. Analyzing click stream data can show retailers how visitors navigate their way around an online shop, which is useful towards understanding the effectiveness of marketing efforts, that is, how customers find the online store, what products they view and what they purchase (Lee, Podlaseck, Schonberg & Hoch, 2001).

Lee et al. (2001) conducted a study to analyze click streams with a view to evaluating the effectiveness of online merchandising tactics. By using an interactive visualization system, they were able to break down the click streams into individual customer shopping steps to highlight potential

problems with their merchandising. For example, if a product has a high click-to-basket rate, this indicates that the product page is effective. However, if the product has a low basket-to-buy rate, it could mean that although the product is attractive, the customer is turned off because of its high price. Such knowledge would then allow the retailer to focus on marketing the product in a more persuasive way.

Lee, Lee, and Park (2007) used a different method to conduct click stream analysis to investigate the key considerations of customers and how these affect the success of online shopping at a much higher level. In their study, users were classified through the use of decision trees into two categories: those who rarely used online shopping and those who frequently used online shopping. For those who rarely used online shopping facilities, the findings suggest that the time and urgency of the purchase are the most important factors that need to be considered. On the other hand, for those who shop online frequently, price, and human resources were highlighted as the key considerations. Such findings allow organizations to understand their customers and better provide for their needs and preferences.

Customer Profitability

Although segmentation and click stream analysis can provide detailed information about a company's customers, data mining can also be used to predict customer profitability (Shen, Xing & Peng, 2007). In other words, a company can identify its most valuable customers and can concentrate their marketing efforts around them. Wang and Hong (2006) used data mining techniques to develop a Customer Profitability Management system which would lead customers into more profitable behavior. Applying their system to customer data from a telecommunications company, neural networks were used to classify customers into distinct groups with regards to their susceptibility to marketing promotions: profitable, resistant, and unpredictable. This allowed the company to target the latter two groups and, with the aid of additional telephone interviews and surveys, enabled them to develop specific marketing strategies and increase their customer profitability levels by 12%.

Marketing Strategies

Making Recommendations

Due to the high degree of variety in services and products offered on the Web, it is sometimes difficult for customers to select those which are the most appropriate to their needs. In other words, the online consumer is faced with a multitude of choice. To address this issue, personalized recommendation systems can be used to predict and identify the preferences and interests of a particular user based on the behaviors of similar users (Karypis, 2001).

Liu, Ke, Lee, and Lee (2008) used clustering and association rule mining techniques to provide recommendations. Initially, the K-Means algorithm was applied to usage data of composite e-services, which are packages offered by a number of e-service providers, to separate customers into user interest groups. Association rule mining was then applied to each of the groups. This resulted in the discovery of similarities in the buying and browsing habits between members in the groups, which then enabled the use of group-based recommendations. For example, if a customer selected "Programming in Microsoft SQL Server 2000 Database", the association rules would highlight that other customers in the interest group had previously selected "Programming with C#" and "Programming with VB.Net". This provided customers with a set of related recommendations to help them effectively utilize composite e-services.

In a similar vein, Lazcorreta, Botella, and Fernández-Caballero (2007) also proposed a solution for recommendations, but only used association rules. Association rule mining may work effectively for smaller data sets, but limitations, such as the amount of time taken to generate the rules or the number of irrelevant associations identified, have been known to arise when larger data sets are involved. To address this problem, they developed a modified version of the Apriori algorithm for a much larger collection of data. Their modified algorithm included four steps: (1) discovering existing association rules in the original repository of transactions; (2) rewriting the original transactions to reflect all association rules that verify each one of them; (3) analyzing differences existing between both sets of transactions by means of statistical analysis; and, (4) discovering and using the existing relations between rules discovered in step 1 in order to automatically divide the set of transaction rules obtained in step 2. They found that the time and effectiveness of the rules were improved by analyzing the behavior of a single user in accordance with other users.

Targeted Marketing

Once classification has been used to identify a series of key factors that can influence a customer's decision to shop online, clustering techniques can be used to identify customer groups with similar online shopping habits. This data can then be used to target specific groups for marketing promotions or predict the behavior of new customers. Liao and Chen (2004) mined customer, product, and transaction data to generate cross-selling rules for the design of an electronic catalogue. Customers were segmented into groups and association rule mining was used to find frequently purchased item sets. Marketers were then able to use this knowledge to create a personalized online catalog for their targeted customers.

Similarly, Lin and Ong (2008) also analyzed customer behavior to aid in the design of an electronic catalog for a stationery mall. Association rules were used to identify links

between customer profiles and products purchased. In this way, collections of target customers, products, brands and discount prices were designed and emailed to customers in specific segmentation groups in the form of an electronic catalog. A review of sales of the stationery company showed an increase in sales over the six month period that this system was implemented.

Additionally, Yang & Lai (2006) compared the analyses based on three different types of data: (1) order data, (2) browsing data and (3) browsing and shopping cart data. Association rule mining was then used to extract frequent item sets so that product bundles could be created. These bundles proved to be more accurate when both browsing and shopping cart data was used, in comparison to the other two types of data analyzed.

FUTURE TRENDS

It is clear from the aforementioned applications that data mining techniques are very useful tools in customer relationship management. However, with more information becoming accessible via the Web, there are increasing concerns that data mining may pose a threat to individuals' privacy and data security. In general, data mining is used for analyzing transaction data, instead of personal data, but it is likely that privacy-sensitive information is revealed so they should be designed with caution. For example, actions can be taken to remove key fields to ensure privacy is still kept in place. Precautions such as these can help to ensure the integrity of data mining techniques, but essentially the future relies upon the cooperation of all researchers and developers to implement these techniques in a way that preserves privacy (Han & Kamber, 2006).

As well as dealing with such issues, limitations have been highlighted that should be addressed in future research.

- **Specific Application Development:** A study by Anderson, Jolly, and Fairhurst (2007) highlighted that data mining is not utilized evenly across all CRM strategies. For example, of the five strategies identified (improving marketing effectiveness, enhance loyalty by improving customer service, customer analysis, customer acquisition and retention and grow or drive business), customer analysis and marketing were found to have had the most research concentrated previously. Data mining has added value to these two strategies, so further research should concentrate on looking at other strategies.
- **Manageable methods:** The amount of data stored and collected in the business world is rapidly increasing. Therefore, care has to be taken to ensure that data mining algorithms remain efficient and manageable. The idea of introducing constraints into data mining algorithms, that is, constraint-based mining, allows the user some additional control over the efficiency of the process. For example, when using association rules to mine for frequent item sets, statistical methods can be used to restrict the search space to only that which is interesting, therefore resulting in more relevant patterns being identified (e.g., Bonchi & Lucchese, 2007).

CONCLUSION

In summary, data mining can be used to discover novel and interesting relationships from large business databases. These can then be used to enhance an organization's productivity by supporting their CRM functions. This paper provides a background to data mining and has given some examples of how data mining supports CRM, especially focusing customer analysis and marketing strategies. By exploiting the full potential of data mining techniques, organizations can meet their customers' needs in the best possible way.

REFERENCES

Anderson, J., Jolly, L., & Fairhurst, A. (2007). Customer relationship management in retailing: A content analysis of retail trade journals. *Journal of Retailing and Consumer Services, 14*(6), 394-399.

Ankerst, M. (2001). Visual data mining with pixel-oriented visualization techniques. In *Proceedings of ACM SIGKDD Workshop on Visual Data Mining*.

Berson, A., Smith, S., & Thearling, K. (2000). *Building data mining applications for CRM*. McGraw-Hill.

Bonchi, F. & Lucchese, C. (2007). Extending the state-of-the-art of constraint-based pattern discovery. *Data and Knowledge Engineering, 60*(2), 377-399.

Böttcher, M., Spott, M., Nauck, D., & Kruse, R. (in press). Mining changing customer segments in dynamic markets. *Expert Systems with Applications*.

Changchien, S.W. & Lu, T. (2001). Mining association rules procedure to support on-line recommendation by customers and products fragmentation. *Expert Systems with Applications, 20*(1), 325-335.

Ha, S. H. (2007). Applying knowledge engineering techniques to customer analysis in the service industry. *Advanced Engineering Informatics, 21*(3), 293-301.

Han, J. & Kamber, M. (2006). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann.

Hand, D.J., Mannila, H., & Smyth, P. (2001). *Principles of data mining.* Massachusetts: MIT Press.

Karypis, G. (2001). Evaluation of item-based top-N recommendation algorithms. In *Proceedings of the ACM 10th International Conference on Information and Knowledge Management* (pp. 247–254).

Lazcorreta, E., Botella, F., & Fernández-Caballero, A. (in press). Towards personalized recommendation by two-step modified Apriori data mining algorithm. *Expert Systems with Applications*.

Lee, S., Lee, S., & Park, Y. (2007). A prediction model for success of services in e-commerce using decision tree: E-customer's attitude towards online service. *Expert Systems with Applications, 33*(3), 572-581.

Lee, J., Podlaseck, M., Schonberg, E., & Hoch, R. (2001). Visualization and analysis of clickstream data of online stores for understanding web merchandising. *Journal of Data Mining and Knowledge Discovery, 5*(1), 59-84.

Liao, S. & Chen, Y. (2004). Mining customer knowledge for electronic catalog marketing. *Expert Systems with Applications, 27*(4), 521-532.

Liu, D. R., Ke, C. K., Lee, J. Y., & Lee, C. F. (2008). Knowledge maps for composite e-services: A mining-based system platform coupling with recommendations. *Expert Systems with Applications, 34*(1), 700-716.

Liu, H. & Ong, C. (2008). Variable selection in clustering for marketing segmentation using genetic algorithms. *Expert Systems with Applications, 34*(1), 502-510.

Lopez, N., Kreuseler, M., & Schumann, H. (2002). A scalable framework for information visualization. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 39-51.

Shen, Y., Xing, L., & Peng, Y. (2007). Study and application of web-based data mining in e-business. In *Proceedings of the Eighth ACIS Conference: Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing* (pp. 812-816).

Urtubia, A., Perez-Correa, J. R., Soto, A., & Pszczolkowski, P. (2007). Using data mining techniques to predict industrial wine problem fermentations. *Food Control, 18*(1), 1512-1517.

Vellido, A., Lisboa, P. J. G., & Meehan, K. (1999). Segmentation of the on-line shopping market using neural networks. *Expert Systems with Applications, 17*(4), 303-314.

Wang, H. F. & Hong, W. K. (2006). Managing customer profitability in a competitive market by continuous data mining. *Industrial Marketing Management, 35*(6), 715-723.

Yang, T. & Lai, H. (2006). Comparison of product bundling strategies on different online shopping behaviours. *Electronic Commerce Research and Applications, 5*(4), 295-304.

Zhang, X., Gong, W., & Kawamura, Y. (2004). Customer behavior pattern discovering with web mining. *Advanced web technologies and applications*. Berlin/Heidelberg: Springer.

KEY TERMS

Apriori Algorithm: An iterative association rule-mining algorithm that uses prior knowledge to identify frequent item sets in order to predict future events.

Bayesian Classification: A type of classification algorithm, based on the statistical probability of a class and the features associated with that class.

Decision Tree: A visual representation of a decision problem that forms tree-like predictive models with the aim of helping people to make better decisions.

K-Means: A nonhierarchical clustering technique which splits data sets into K (a given number) subsets in a way that each subset is maximally compact.

Kohonen Networks: A type of unsupervised neural network used for finding patterns in input data without human intervention, consisting of Vector Quantization, Self-Organizing Map, and Learning Vector Quantization.

Neural Network: A computational approach inspired by simple models of the brain, consisting of nodes or neurons connected together in some sort of network.

Self Organizing Map (SOM): An unsupervised neural network algorithm that results in a clustered neuron structure, where neurons with similar properties (values) are arranged in related areas on the map.

A

Applying a Teaching Strategy to Create a Collaborative Educational Mode

Nidia J. Moncallo
Universidad Nacional Experimental Politécnica "Antonio José de Sucre", Venezuela

Pilar Herrero
Universidad Politécnica de Madrid, Spain

Luis Joyanes
Universidad Pontificia de Salamanca, Spain

INTRODUCTION

The evolution of ICT and the influence over educational areas has been very significant in recent years, changing conception of learning environments, communications and interactions forms, and educational material. Researchers, like Buzon and Barragán (2004), have expressed the need to create new learning (online) environments that allow teaching and learning without the time and space restrictions of residential courses, and ensures continual (**virtual**) **communication** between students and professors, or the need to find new material courses, learning strategies that allow the efficient use of new systems and educational resources emerged from technical advances (Wai-Chung and Li, 2007; Weert, 2006).

On the other hand, among the conclusions reached at the Second Virtual Congress, "Education through Internet and Internet in Education" (2004), was the need for all technological research to take into account the pedagogical, economic, and social aspects, so that a coherent integration between technology and education can be achieved.

Nevertheless, it is still difficult to incorporate the use of tools like chat, electronic mail, text editors, and forums, in other activities that involve no more than the simple exchange of information; this limits their potential and benefits. According to Friendals and Pauls (2005), the majority of Professors still depend on well-established, primitive teaching aids, like, for example, chalk and board. Their analysis revealed that the need for teaching aids in classrooms, which include educational integrative mini-applications, should be one of teaching's main priorities.

Based on these criteria, that is, the need to effectively incorporate ICTs to make changes in the educational field, this research has focused in submitting a proposal of a collaborative teaching strategy, empirical education collaborative **teaching strategy**, shortening EE-Col, like first link to develop later on, a collaborative educational model. EE–CoL's validation will enable to lay down the basis for the design of an exclusive model in **distributed environments** where the generated learning elements are interoperable and reusable, using shared and coordinated resources.

BACKGROUND SECTION

There have been numerous positive experiences in higher education, where **collaborative learning** has been supported by the use of electronic mail and discussion forums (Murillo, 2000; Romero, Osuna, Sheremetov, Chi, & Villa, 2003); collaborative editing systems to support groups that edit, simultaneously, from different places (Ignat & Norrie, 2004; Stavroula, Ignat, Ester, & Norrie, 2006); or experiences that show interest in the design and application of collaborative environments with use of diverse learning technique (Gonzalez, 2006; Lucero, Chiarani, & Pianucci, 2003; Roman, 2003). Similarly, there were solutions that used the Web for collaborative work (Klein, 2004; Thao, 2002). These investigations have evidenced excellent and satisfactory results when utilizing technological elements like auxiliary tools in educational processes

In investigations of **computer assisted collaborative learning,** CSCL, of Rubia, Jarrín and Bote (2003); Martinez, Gomez, Martinez, and Mora (2004); Hansson and Van Heuten (2006), among others, the use of BSCW (*basic support for cooperative work)* for education programs between two institutions, computer simulation tools use for synchronous and asynchronous communication showed their effectiveness to resolve punctual problems of collaborative learning and communications.

The analysis of theses researches confirmed the benefits obtained by students in collaborative learning and advantages offered by ICT for working with distributed environments. Additionally, it helped confirm that most research is directed towards the satisfaction of specific needs, like distance-learning applications, to promote group integration in the acquisition of knowledge, or increase the use of groupware

tools in the educational process in order to improve academic achievement, social **interaction,** and communication. In general, the projects checked were focused on these objectives, while proposals of **teaching strategy** or educational models that allow the incorporation, integration, and systematization of new tools as a fundamental component in teaching processes were not found.

EE-CoL Teaching Strategy

EE-Col, based in the integration of constructivism, negotiation and social integration, cognitive conflict, and collaborative-cooperative work principles, (concepts explained by Fernández & Melero, 1995; Martinez, 2001; Panitz & Panitz,1998=), has, as a premise to reach a consensus through the cooperation of group members, shared authority and acceptance of responsibility of action as a group, applying techniques and methodologies of **collaborative learning** to activate critical thought and autonomous learning.

EE-CoL's fundamental motor is to be applied through the collaborative tools offered by computer technology applications, like electronic mail, videoconferences, chats, discussion and application groups, which allow the strategy to support synchronous and asynchronous interaction to have, at its disposal, the multiple advantages relating to space-time that the tools offer. Such characteristics make the new EE-Col unique, as it results from the combination of pedagogic and collaborative principles, information-technology tools, and emerging technologies of communications and applicable to **blended learning.**

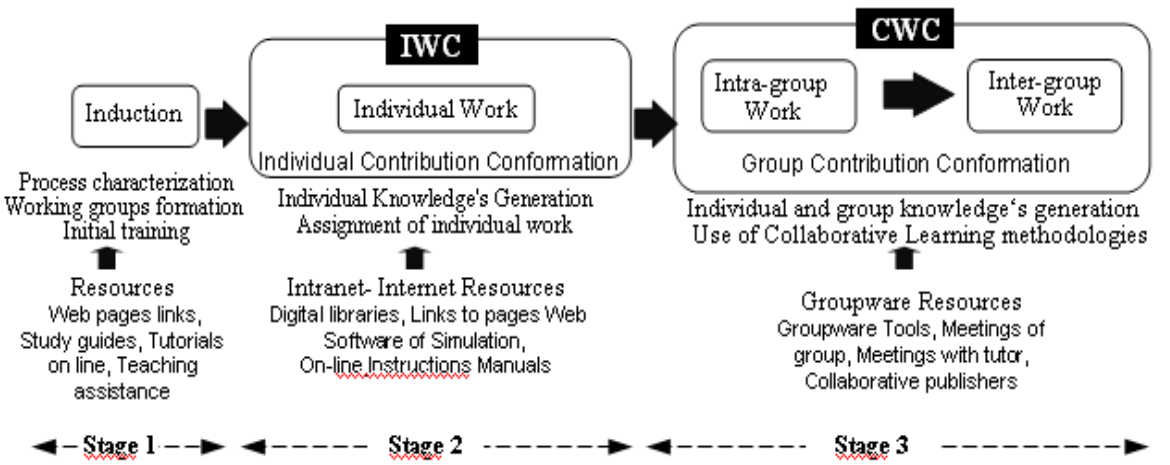
EE-CoL Stages

EE-CoL develops in three consecutive stages, *induction stage*, *individual work stage*, and *group work stage*, that complement each other, in order to achieve acquisition and accumulation of individual and collective knowledge. Figure 1 shows EE-CoL development stages.

Individual work component (IWC). IWC gives each individual the preliminary preparation to achieve certain balance between knowledge and abilities for more effective **interaction** during subsequent group discussions. It is essential for IWC to be focused on two key points: student's needs and student's motivations. In consequence, the professor must carefully select the **teaching methodology** or required activities, schedules, basic material, and online or face-to-face consultation sessions; identify the core concepts and the information available, always bearing in mind the knowledge that the students must acquire during this stage. EE-CoL proposes to use the *introductory focal activity* that specifically seeks to attract the students' attention towards a specific aspect.

***Cooperative work component (CWC*).** The CWC's essence is the knowledge generated by **interaction** and group work through two components. The *cooperative intragroup work* organizes students in order that they can achieve their goals with an active and positive interaction accomplished through synchronous and asynchronous discussions that generate information exchange *between group members* and a global synthesis of the team's work. Intragroup interaction is activated through collaborative work strategies: *jigsaw*

Figure 1. EE-CoL stages



or puzzle, research groups *co-op co-op or kagan* (Diaz & Hernández, 2005), selected according to student numbers, topic, objectives, subject (theoretical or practical), and degree level (undergraduate or postgraduate).

C*ooperative intergroup work* is subsequent stage to interchange of information *between groups,* always through discussion sessions lead by a teacher where each group's results are analyzed and learned by all students. These discussions, which activate productive **interaction,** can take place on the Internet, in virtual classrooms, in videoconferences, in discussion forums, or in person. This subcomponent allows complementing the knowledge acquired within each group.

EE-CoL Implementation

EE-CoL was developed in two sections of Analog Electronics Laboratory Course of the Electrical Engineering School at Universidad Nacional Experimental Politécnica, Venezuela. In these courses, 10 experiences were executed, and the research subjects were the 24 students registered for the course (sections A and B); in each section, three work groups of four students per group were formed.

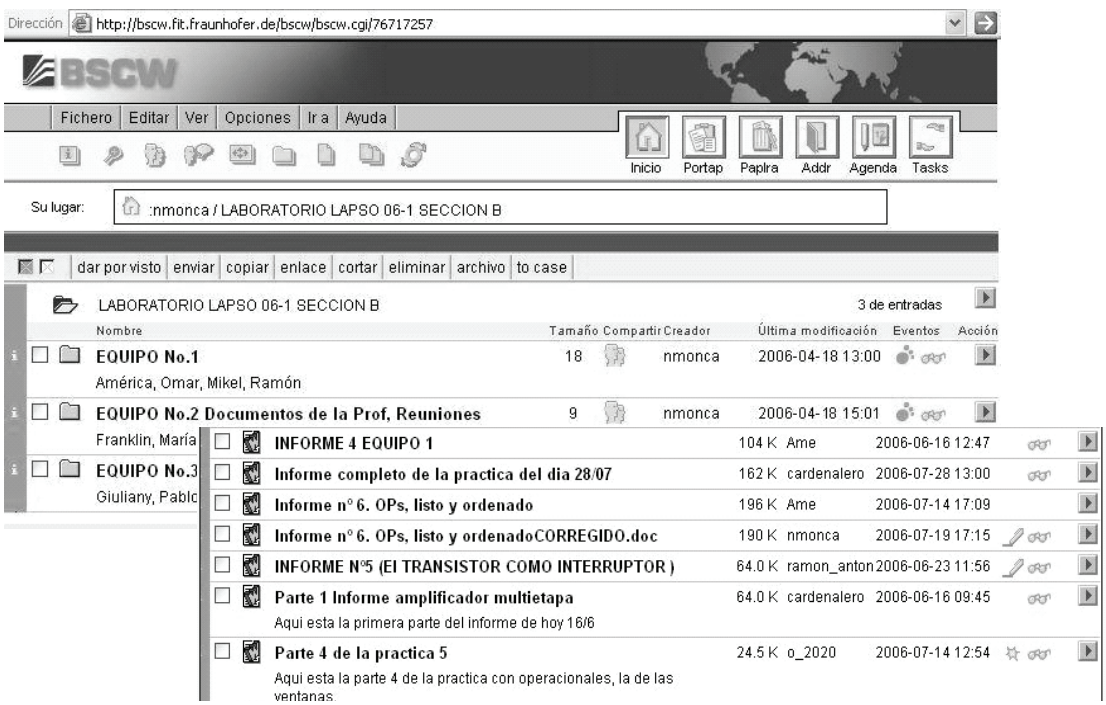
Both strategies, traditional and EE-CoL, were applied to each group of students. The traditional strategy was applied to the first five lab sessions, with traditional methods and work booklet; the subsequent five sessions were developed with the EE-Col strategy, using a new assessment plan and new support material to carry out each stage of strategy

In new support material, each practical session had two parts: *prior preparation,* which should have been prepared individually by each member of the group and it was constituted by questions over procedures explications, components, and equipment operating; and *laboratory experience,* whose processes are closely related with prior preparation. One BSCW space was set for each section to be used by its students as a shared space.

During the *induction stage*, the methods of working, assessment to be followed during the experiment, and the fundamental aspects of the strategy were explained, and the basic processes of **BSCW** (edit documents, add and erase notes, files or folders, and carry out discussions forums) were practiced. For each group, the instructor assigned three basics roles with specific functions: group leader, material organizer, and the editor. It was suggested that for each session, the roles were rotated in accordance with the members' criteria so that they all had the chance of performing each role at least once.

Through the *individual contribution* approval stage, each student had to individually research the questions

Figure 2. Folder's distribution with one use sample of group folder



asked in the previous preparation, operational basis for EE-CoL's *individual work component*, and "upload," on to the shared space, his individual answers that had been supervised by the professor.

The *intragroup contribution* was generated by online discussions, preparation of final documents, pre- and post-laboratory, and experimental works sessions. In order to elaborate the final report, all members of the group checked each other's individual contributions to select the best answers, correct mistakes, add suggestions, so that the final report could be prepared by editor of group. The **BSCW** space was used for these purposes. (See space sample in Figure 2) A chronogram for these activities was suggested by the teacher.

The *intergroup contribution* was made up of face-to-face discussions, led by the instructor, that took place among all students during the first 20 minutes of each laboratory session.

Results of EE-CoL Evaluation

As EE-CoL is a teaching strategy that utilizes electronic tools and collaborative principles, the goal was to somehow measure its basic principles by observing the interaction between students, individual responsibility, active participation in discussions, and academic achievement and performance during the lab session. The active participation in the joint preparation of reports was one parameter considered to measure **interaction** between students. They were able to create, read, and modify the documents being produced by all group members in order to assemble the final document. Graphs in Figure 3 show noticeable activity to produce pre- and postlaboratory reports.

Likewise, use of discussion forums and insertion of comments next to the documents were observed. These were used in order to communicate any doubts, explanations, agreements, or approvals to the modifications made by any of the students. Figure 4 shows the interactions from each group during report creation using the **BSCW** tool.

In order to evaluate the acceptance of the new strategy, and the use of its tools, a survey-type measuring instrument to evaluate the new work material and the group work was used. The following results were obtained:

A high degree of satisfaction was obtained with the incorporation of previous preparations; 72.7% answered that they allowed a better disposition to the understanding and development of the experiments, and a more effective learning process

In reference to group work, 100% agreed on the need for adequate distribution of members and roles within the group in order to obtain better results when working as a team; 80% admitted that the work performed in their roles allowed them to determine weaknesses and strengths in their performance and group work. In the same manner, a high percentage, over 90% were of the opinion that the group work was beneficial and that they felt the interrelation between members.

The groupware tools and the use of BSCW cyberspace was highly accepted within the surveyed groups since they could intervene from any place and at any time; a positive opinion was detected by benefits when group discussions took place, and values over 80% confirmed that discussions, report creation were facilitated.

Figure 3. Distribution of asynchronous interactions (document creation, reading, and modification for reports elaboration (pre–post laboratory)

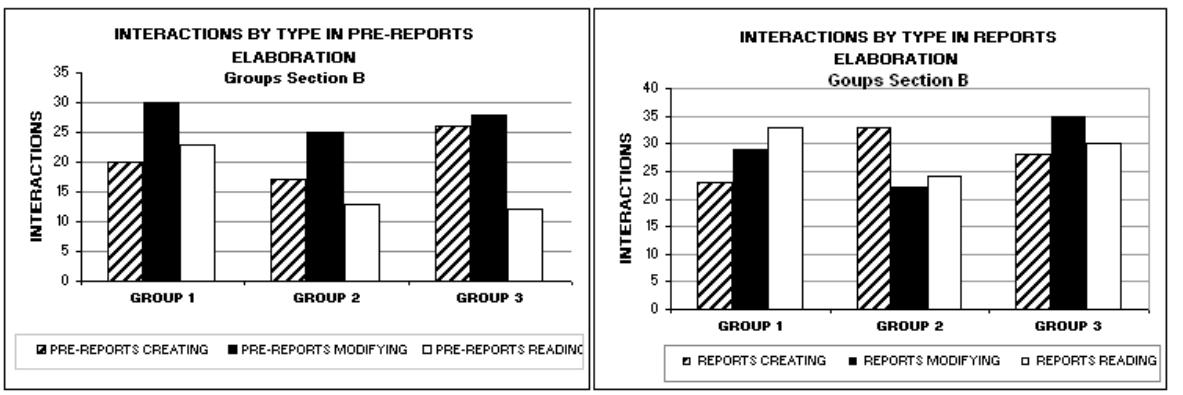


Figure 4. Interactions for reports elaboration of section B groups

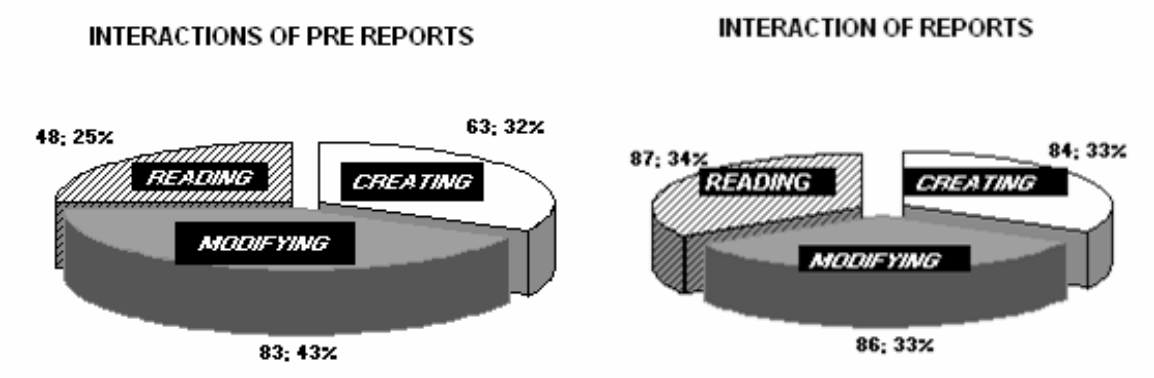
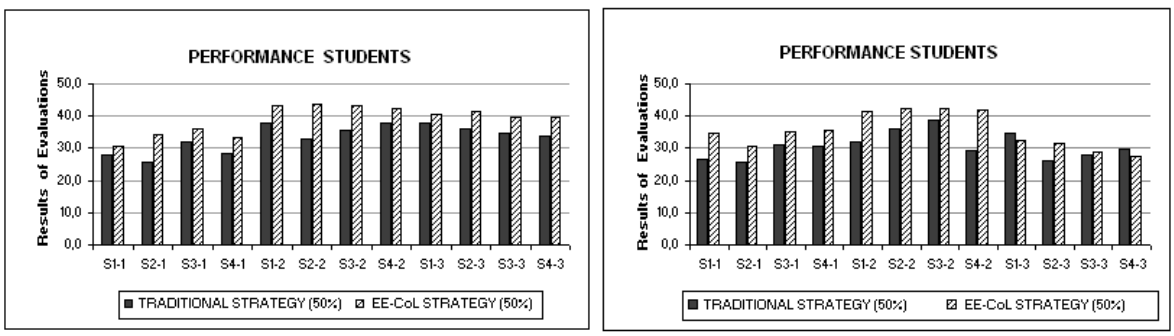


Figure 5. Student performed according to used strategy for both section



As an important additional element, the students' academic performance in both phases was considered. When comparing the academic performance, 83.3% of students performed better when the proposed strategy was applied. Figure 5 shows a bar chart with students' performance according to the used strategy.

The students' performance level when using EE-CoL was better because there was more active and significant participation in discussions and better disposition to complete the assigned tasks. Furthermore, when using EE-CoL, a more favorable effect is evident in the socio-affective relations of the students, and a higher **interaction** and collaboration between students was achieved, because the student had the possibility to program his own work and to have a better coordination of activities from any place and at any time.

As a collaborative strategy, EE-CoL achieved participation of students in the planned activities, using the different tools made available for collaborative work. In the same manner, group goals were achieved by distributing responsibilities through the roles performed by the students within their group.

The main difficulty observed during the application of the strategy was represented by two factors: lack of skill in the use of computers and electronic tools in some students and

the difficulty to access the different resources and information made available for the work at any specific time, due to the lack of Internet connection outside the campus.

FUTURE TRENDS

Group interactive work plays an important role in the success of the proposed model. Therefore, it is of interest to specify and optimize mechanisms that could boost ***interaction*** and the use of *collaborative tools.* All this implicates the need to initiate further research that will allow the transference of experiences to other educational domains with diverse pedagogical contexts. At the moment, the application of EE-CoL to a master's degree subject is being planned. With this, the usability of the different types of interaction and informatics tools proposed in the strategy could be analyzed and the applicability could be validated.

The use of other technologies that facilitate collaborative work and resources sharing, like grid technology, already existing in several campuses, would allow EE-CoL implementation with better results. With analysis of this alternative, it would then be possible to perform a study to propose a collaborative educational model in grid environments involving both innovations

CONCLUSIONS

This chapter has presented EE-Col, a teaching strategy based on the collaborative work principles and the use of its informatics tools in **distributed environments**. The results have shown significant evidence on effectiveness and efficacy of EE-CoL, when the users achieved their objectives with precision and with relatively few resources; the high degree of satisfaction was also verified by positive attitude of the students towards its use. In addition, the application of this strategy gave satisfactory results with respect to promoting **interaction** and communication, verification of the simplicity of the groupware tools and of the presence of the basic principles of collaborative work.

EE-CoL constitutes an alternative for complementing traditional education or promoting **interaction** and communication in courses whose participants, teacher and students, do not have a continuous contact, like at the postgraduate or training courses. Results allow establishing the link to formulate a new collaborative educational model for **distributed environments**.

REFERENCES

Buzón, G., & Barragán S. (2004). *Las Nuevas Tecnologías en la Enseñanza Superior y Universidad.* Congreso Educación en Internet e Internet en la Educación. Num.2. Madrid. Ministerio de Educación y Ciencia.

Congreso Educación en Internet e Internet para la Educación. (2004). Mesa redonda. *Ministerio de Educación y Ciencia. Num. 2*. Spain.

Díaz., F., & Hernández G. (2005). *Estrategias docentes para un aprendizaje significativo. Una interpretación constructivista.* 2da.Edición. Mexico: Mc-Graw Hill.

Fernández, B., & Melero Z. (1995). *La interacción social en contextos educativos.* Madrid: Siglo XXI, Editores.S.A.

Friedland, G., & Pauls, K. (2005). Architecting multimedia environments for teaching. *Computer IEEE Computer Society, 38*(6), 57-64.

Gonzalez, G. (2006). A systematic approach to active and cooperative learning in CS1 and its effects on CS2. *ACM SIGCSE Bulletin, 3*(1),133-137. Hansson, T., & Van Heugten, L. (2006). **Collaborative writing in a software game: Re-enacting a Vygotskian design.** *Advanced Technology for Learning, 3,* 22-28. Ignat, C., & Norrie, M. (2004). **Grouping in collaborative graphical editors.** *ACM conference on Computer supported cooperative work,* 447 – 456. New York: ACM Press.

Klein, M. (2004). ***Web-based support for collaborative course design and teaching among university faculties.*** *International Journal of Management and Decision Making, 5*(4), 313–332. Lucero, M, Chiarani, M., & Pianucci, I. (2003). *Modelo de Aprendizaje Colaborativo en el ambiente ACI. Investigation Group report: Ambientes Colaborativos Inteligentes*. Informatic Department. San Luis University. Argentina. Retrieved 07-05-2005, from http://www.dirinfo.unsl.edu.ar/~profeso/PagProy/articulos/Lucero%20Cacic%202003.pdf

Martínez, M. (2001). *Método y modelo para el apoyo computacional a la evaluación en CSCL*. PhD Tesis.Retrieved 21-03-2005 from http://www.infor.uva.es/~amartine/research/phd/tesis_amartine.pdf

Martínez, M., Gómez, A., Martínez, E., & Mora, M. (2004). COLAB: Una plataforma para la simulación en entornos colaborativos en laboratorios virtuales. *Revista Iberoamericana de Inteligencia Artificial, 8*(24), 45-53.

Murillo, J. (2000). *Un entorno interactivo de aprendizaje con Cabri-actividades aplicado a la enseñanza de la geometría*

en la ESO. Universidad Autónoma de Barcelona. Tesis Doctoral. Retrieved 15-01-2006, from http://www.tdx.cesca.es/TDX-0710101-030847/

Panitz, T., & Panitz, P.(1998). Encouraging the uses of collaborative learning in higher education. In J.J. Forest (Ed.), *Issues facing international education*. NY: Garland Publishing.

Román, P. (2003). La flexibilización de los espacios de aprendizaje a través de entornos de trabajo colaborativos telemáticos. *III Congreso Internacional Virtual de Educación, CIVE 2003*. 1-11 Abril. Retrieved 04-02-06, from http://www.cibereduca.com

Romero, M., Osuna, G., Sheremetov, L., Chi, M., & Villa, L. **(2003). Study and analysis of the working space conscience in CDebate: A groupware application for collaborative debates.** *Latin American Conference on Human-computer **Interaction**; ACM International Conference Proceeding Series, 46*, 107-115.

Rubia, B., Jarrín, M., & Bote, L. (2003). Una experiencia de formación Colaborativa y práctica real entre la Universidad y un centro educativo Generando un espacio CSCL. *Revista Latinoamericana de Tecnología Educativa, 3*(1).

Stavroula, P., Ignat C., Ester, G., & Norrie, M. (2006). **Increasing awareness in collaborative authoring through edit profiling.** *CollaborateCom 2006, 2nd International Conference on Collaborative Computing: Networking, Applications and Worksharing, Atlanta, USA. Retrieved 09-01-2007, fromhttp://www.globis.ethz.ch/ publications/ index#year2006.*

Thao, L. (2002). Collaborate to learn and learn to collaborate. *Seventh world conference on Computers in Education: Australian topics, 8.*

Wai-Chung, E., & Li, Q. (2007). An experimental study of a personalized learning environment trough open-source software tools. *IEEE Transaction on Education, 50,* 331-337.

Weert, T. **(2006). Education of the twenty-first century: New professionalism in lifelong learning, knowledge development and knowledge sharing.** *Education and Information Technologies, 11*, 217-237. Kluwer Academic Publishers.

KEY TERMS

Asynchronous Communications: Computer-based exchanges of messages for which the participants need not be available or online at the same time, but, rather, read and respond as their schedules (and desires) permit. Examples: e-mail, discussion boards, text messaging over cell phones.

Blended Learning (B-Learning): Learning that combines different modes of delivery, such as online and traditional face-to face learning.

BSCW System: Provides facilities for collaboration over the Internet. It is based on the "shared workspace" metaphor: an object store for group work, with some simple awareness functionality that allows users to keep an overview of what is happening in the workspace.

Focal Introductory Activity: A set of activities that try to fix the students' attention to stimulate the knowledge previously acquired and to create an appropriate motivational initial situation that will be used for any posterior activity. Hypothetical situations, examples, case studies, or a group of approaches or questions are activities that could be utilized.

Groupware: Refers to computer applications designed to help people work together collectively while located remotely from each other. Groupware services can include the sharing of calendars, collective writing, e-mail handling, shared database access, electronic meetings with each person able to see and display information to others, and other activities.

Jigsaw or Puzzle: Technique used in collaborative learning that proposes dividing the academic material in as many sections as the number of group members. All groups learn the same topic but, inside the groups, each member learns their assigned section, making him an "expert" on his "piece of the puzzle" or knowledge section. Later, the members of diverse groups that have researched the same topic meet in the so called "expert groups" to discuss their findings, obtain new knowledge and then, return to their original group to share and teach this topic to the rest of the group.

Kagan Technique: Or co-op co-op technique for collaborative learning. Each work group freely chooses the topics to be divided into subtopics. These subtopics will be assigned to, and developed by, each member. Later, each member will impart the knowledge acquired to the rest of the colleagues, so that at the end, each group presents their work globally.

Synchronous Communication: *Or* direct communication, where all parties involved in the communication are present at the same time. Examples: telephone conversations and instant messaging.

Applying Evaluation to Information Science and Technology

David Dwayne Williams
Brigham Young University, USA

INTRODUCTION

As indicated by the wide range of topics addressed by this Encyclopedia, the fields of information science and technology have grown exponentially. Likewise, the field of evaluation has evolved and become increasingly integral to learning and improving upon principles and practices associated with all fields the Encyclopedia explores.

The field of evaluation is the formal transdiscipline of gathering information about the performance or nature of objects of evaluation and comparing the objects' performance to criteria to help participants make evaluative judgments (Scriven, 2004). Evaluation includes several elements: negotiation with multiple participants regarding their values and criteria, using many different kinds of processes to document and judge the performance of various objects of evaluation, formative and summative purposes, measurement and assessment techniques, and use of quantitative and qualitative data gathering and analysis processes.

This chapter documents the development of evaluation as a field; presents a framework for thinking about evaluation that is theoretically sound and practical to use; and explores ways to apply the framework to facilitate learning, improvement, decision-making, and judgment in all sub-fields of information science and technology.

BACKGROUND

After reviewing several approaches to achieving different evaluation purposes, the relationship between evaluation, measurement, and assessment is explored and the use of quantitative and qualitative data to facilitate evaluation is clarified.

EVALUATION THEORIES OR APPROACHES

For the last few decades, many approaches to evaluation have been evolving. In the 1960's several social scientists, psychometricians, and others responded to government challenges to evaluate funded programs by identifying approaches that have been debated and expanded for years. Many of these approaches are summarized and discussed by Fitzpatrick, Sanders, and Worthen (2004) and Alkin (2004).

For example, one influential thinker, Daniel Stufflebeam (2004a), introduced the CIPP (context, input, process, product) approach in the early 1970's. He elaborated the idea of meta-evaluation and guided the Joint Committee on Evaluation Standards to generate meta-evaluation standards (Stufflebeam, 2004b) for judging evaluations of programs, personnel, and students.

Patton (2004), recognizing that many evaluations, using social science research approaches, were ignored by the stakeholders that they were supposed to serve, he therefore created utilization-focused evaluation. It promotes practical ways to ascertain and target stakeholders' criteria to raise chances of results use.

Lincoln and Guba (2004) questioned the dominant evaluation paradigms and proposed fourth generation evaluation. Its hermeneutic dialectic methods of working with stakeholders seeks to negotiate their often conflicting values to better identify criteria, standards, and questions for guiding evaluations.

Robert Stake's (2003) responsive approach proposed radical changes to his earlier countenance approach by acknowledging that evaluation is only one of many factors that communities of stakeholders consider when negotiating with one another about evaluating objects they care about together.

Cousins, Goh, Clark, and Lee (2004) noted that evaluation is part of most organizations and something all stakeholders are doing constantly. They reviewed ways to encourage stakeholders to collaborate in various participatory approaches to formal evaluation.

Fetterman and Wandersman (2005) have proposed an approach to evaluation that some argue is more a form of social activism than evaluation. Empowerment evaluation seeks to encourage professional evaluators to coach various stakeholder groups, but particularly those that traditionally have less voice in their social and political communities, to conduct their own evaluations.

Formative and Summative Purposes

Scriven (2004) has critiqued other approaches and proposed others, such as goal-free evaluation and the key evaluation

checklist. He also distinguished summative from formative evaluation, to not only test how well evaluands achieve their purposes but also to seek formative feedback to improve evaluands.

Measurement and Assessment Techniques

Another important distinction in the literature is the relationship between evaluation, measurement, and assessment, which are often used synonymously. In the *Encyclopedia of Evaluation* (Mathison, 2005) three authors note: "Measurement may be defined as the set of rules for transforming behaviors into categories or numbers" (Petrosko, 2005, p. 247). "Roughly synonymous with testing and evaluation in lay terms, assessment has become the term of choice in education for determining the quality of student work for purposes of identifying the student's level of achievement" (Mabry, 2005, p. 22). "Evaluation is an applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, merit, worth, significance, or quality of a program, product, person, policy, proposal, or plan. Conclusions made in evaluations encompass both an empirical aspect (that something is the case) and a normative aspect (judgment about the value of something)" (Fournier, 2005, pp. 139-140).

One implication of these quotes is that thinking about the evaluation task in terms that include measurement and assessment as subsets of the broader evaluation concept should help anyone using evaluation to explore its wider ranging concerns and thus enhance whatever they are evaluating as well.

Quantitative and Qualitative

A final concern raised by the approaches to evaluation asks whether quantitative, qualitative, or a mixture of methods are better for evaluation. Although explored extensively in social science literature, this debate continues in evaluation literature as well. To many, some evaluation questions demand qualitative answers while others seem best answered through quantitative data collection and analysis. Lately, mixing methods has been the answer many evaluation theorists give regarding method issues. However, as Yanchar and Williams (2006) have argued, mixing methods without taking into account the assumptions those methods are built upon does not make those assumptions meaningless or of no influence. All evaluators should examine and build upon assumptions they can support and trust when selecting methodologies and associated techniques of data collection and analysis.

Summary

The field of evaluation has developed through efforts of theorists and practitioners from many fields for several years. Although many issues remain unresolved, evaluation scholars and professionals identify several variables to account for in creating evaluations that help stakeholders. Many such variables are addressed in the evaluation framework described below.

AN EVALUATION FRAMEWORK

A framework for applying the lessons learned by the field of evaluation to the many fields associated with information science and technology includes the elements presented in Table 1 and explained thereafter.

IDENTIFYING STAKEHOLDERS

Who are the stakeholders interested in evaluation of information science and technology programs, projects, products, and so forth? This question should be addressed first according to most of the approaches to evaluation cited earlier. Some questions to clarify who the stakeholders are include: Who asked for the evaluation and why? Who is served by the evaluand or should be? Who is likely to use the evaluation results to do something helpful? Who does not usually have a voice in matters associated with the evaluand but has a stake in it?

Table 1. Elements of an evaluation framework guiding what evaluators should do

1. Identify stakeholders and objects of evaluation (evaluands) they care about.
2. Clarify background, literature, values, issues, criteria, standards, and guiding questions reflecting stakeholders' beliefs about "what should be" regarding the evaluands.
3. Use data collection and analysis to document "what is" regarding evaluands.
4. Compare "what should be" to "what is" to generate results and recommendations.
5. Meta-evaluate before, during, and after conducting an evaluation.

Evaluators are Attuned to Stakeholders' Values

They may be internal to an organization or external consultants; but evaluators encourage all stakeholders to speak up and voice their value perspectives. When value conflicts arise among stakeholders, evaluators encourage dialogue, empathic understanding of alternative views, consensus-building, and otherwise involve stakeholders in understanding each others' perspectives. Evaluators believe if stakeholders' preferences and definitions are not identified and built systematically into the evaluation, they may ignore, counter, or misuse the results.

Clearly, the subfields associated with information science and technology are of interest to many potential stakeholders with variegated concerns for how various components of the field are performing. Designers who want to use technologies to create and sell products, potential and actual users of technology innovations, funders of projects and programs all have unique perspectives on what matters most to them and what they look for in terms of quality.

Clarifying Objects of Evaluation (Evaluands) Stakeholders Care About

Evaluators should involve multiple stakeholders in deciding what evaluands to focus the evaluation on. Questions to clarify the evaluand include: What do you want to judge or improve? What are its objectives? What is its "program logic?"

In information science and technology, the evaluator may begin by studying an evaluand such as a computer or a computer program only to discover that the stakeholders are really interested in the implementation of that program or a distributed system of computers or database management technologies more generally. Often stakeholders are not sure what to focus the evaluation on because they have not thought about their evaluands in this way or because they have conflicting purposes in contrast with their colleagues. The evaluator has the task of inviting stakeholders to not only reflect on their values associated with various potential evaluands but to also narrow their consensus focus to particular aspects of the evaluands that might be improved or that need to be summatively judged.

Some potential evaluands associated with information science and technology could include accounting information systems, computers, database management systems, decision support systems and related technologies, distance education teaching methods, learning applications, Web-based course and curriculum development tools, and electronic commerce applications.

Helping Stakeholders Focus Their Inquiry on Particular Criteria

Once the stakeholders associated with a particular evaluand have been clarified, the evaluator should continue working with these stakeholders to identify the background context, literature, questions, values criteria, and standards that need to be understood.

Questions evaluators should help stakeholders answer to focus the evaluation on particular criteria include: What does the literature associated with the evaluand say are the key issues? Why is an evaluation appropriate now? Is the focus on gathering summative or formative information? Is the study best done as an internal or external evaluation? Will the evaluation focus on needs assessment, implementation fidelity, or on outcomes? What concerns do they have about the evaluand? What criteria do they have for judging the evaluand? What do they think the evaluand should be accomplishing? What standards do they have or how completely do they hope the evaluand will meet the criteria? How will they know when the evaluand is successful to their satisfaction?

Examples of criteria and standards this process might yield in the various fields of information science and technology include: There should be a new technological intervention to address several needs identified by the literature and key stakeholders. One of several alternative technological solutions identified should be most worth pursuing and developing according to stakeholders' educated points of view. This new program should be at least 90% implemented in at least 80% of the target locations in terms of the design specifications by next fall. This new product should raise performance levels by 50% during its first year once it is fully implemented in the targeted market.

Identifying Questions to Guide the Evaluation

All the activities described to this point have been focused on helping the evaluators and stakeholders clarify "what should be" regarding the prioritized evaluands they are most interested in evaluating. Once these are identified and agreed upon by stakeholders, particular evaluation questions based on the criteria should be articulated for use in guiding the gathering of "what is" information to compare to the "what should be" criteria in making actual evaluation decisions.

Prioritizing the many questions so the study can be focused on the most important ones is another challenge evaluators face and usually involves more discussion and negotiation with multiple stakeholders to make sure they are in agreement and willing to invest time and energy in

obtaining answers to these questions and using the results to take evaluative action.

Examples of evaluation questions associated with information science and technology might include: Is there a need for a new technological intervention? Which of these alternative technological solution ideas is most worth pursuing and developing? How well is this new program being implemented in terms of the design specifications? What is the estimated impact of this new product in this targeted market during its first year?

The evaluation questions provide the focus for all the remaining components of the evaluation framework presented here. The questions are familiar to social scientists and others because they are similar to research questions that demand particular kinds of information be gathered and interpreted in order to answer them. But the evaluator continues to remind stakeholders and other audiences that evaluation questions are based on values, criteria, and evaluand definitions "owned" by particular stakeholders; so the answers to these questions will be used to ascertain how well those criteria and standards are being met. Evaluation decisions are different than research results in this way.

DESIGNING DATA COLLECTION AND ANALYSIS PROCEDURES

In contrast with the questions asked above, the key issues for evaluators associated with data collection and analysis are very similar to those addressed by researchers. They include: What information from what sources will be collected, using what data collection procedures, by whom, and on what schedule? Also, how will each collection procedure be refined to ensure validity, reliability, credibility, trustworthiness, generalizability, transferability, or whatever other methods standards apply? Analysis questions focus on clarifying how quantitative and qualitative data will be summarized, analyzed, displayed, and represented to best reflect the information gathered.

It is during this phase of the evaluation process that measurement issues become very important. Evaluators have to ask how to most appropriately measure success or progress associated with the various criteria and standards valued by the stakeholders. These measures may be qualitative (such as qualitative descriptions of participants' behaviors, dispositions, attitudes, and interactions) or quantitative (such as performance on tests, surveys, observation protocols, and work records), or some combination of these data. Sometimes these measures will be assembled into assessment systems that are administered and collected regularly to feed continual information into databases for consideration by the evaluating stakeholders. At other times the measures will be periodic and idiosyncratic for particular decisions that need to be made only once in awhile.

Examples of data collection procedures for information science and technology might include surveys of stakeholders' perceptions of the need for a potential innovation, observations of how well a computer lab is being used by trainees compared to how it was designed to be used, or performance assessments of students' skills after using a new software program. Analysis procedures should be tailored for each kind of data collected to allow nuanced interpretation of evidence and presentation in terms of charts, tables, figures, statistics, thick qualitative descriptions, and theme analyses. All these analyses should be focused on answering the evaluation questions but unanticipated findings should also be identified to allow balanced reporting.

DEVELOPING RESULTS AND RECOMMENDATIONS

Finally, once the data regarding "what is" are collected and analyzed, evaluators and stakeholders search for ways to share results in reporting displays and formats that are helpful, timely, and accessible. Oral, written, or mediated formats may form interim and final reports which thoughtfully compare "what is" as learned through evaluation activities to "what should be" as identified through dialogue with stakeholders so evaluation questions can be answered and stakeholders can decide how well the evaluand is meeting their criteria.

Whatever the results, the stakeholders should help decide what to do about them. The evaluator should collaborate with stakeholders to identify realistic recommendations, such as continuing to fund initiatives, modifying processes, refining designs, and otherwise making formative and summative decisions. Unless they are very well informed internal evaluators, most consultants do not adequately understand extenuating circumstances to make realistic recommendations. The best policy is to involve the stakeholders in making their own evaluation and recommendation decisions.

METAEVALUATION

Although listed at the end of this framework, metaevaluation should infuse every other step and concept involved in planning, implementing, and critiquing a final evaluation study. It should be done internally by those who are doing the evaluation and externally by experts who can see the issues from different angles.

The main questions asked during a metaevaluation include: How does this study (as planned, carried out, and reported) hold up against metaevaluation standards such as

those based on the Joint Committee standards (Stufflebeam, 2004b) which address four groups of standards: utility, feasibility, propriety, and accuracy. An internal metaevaluation should explain how the evaluation has been conducted to meet each standard and/or why particular standards are not relevant in that case. An external meta-evaluation should certify whether the evaluation should be trusted or not.

FUTURE TRENDS

What will be done with this framework to improve information science and technology innovations? A major premise of this chapter has been that using the framework, which is based solidly on the formal evaluation literature, will enhance evaluation and associated improvements in various subfields of information science and technology. This process will thereby improve performance of the continually evolving innovations in these areas, if the evaluations are done with sensitivity to the values and needs of the participants who could use the results to make such improvements. To the extent that evaluation is viewed as a helpful way to gain insight into information science and technology innovations and not something to be feared or avoided, the future is bright for both evaluation and growth through feedback.

This future trend can be most effectively achieved by building evaluation more systematically into the strategic operations of organizations and their associated stakeholders. If readers use the framework described earlier to identify evaluation needs and activities before, during, and after implementation of various activities and policies, they will identify ways to improve and judge the evaluands they care the most about. If they collaborate to use the resulting evaluation information to jointly recommend actions to take, "what is" will become more and more like "what should be" and the goals of information science and technology will be more readily achieved. That is the future this article was written to encourage. If evaluation and the ideas in this framework are avoided or enfeebled, the future of the field will be jeopardized.

CONCLUSION

This chapter has argued that evaluation should not be an isolated or esoteric concept, to be avoided out of fear or lack of understanding. Rather, by following the straightforward steps provided in the framework, organizational leaders and members, and consumers of information science and technology products can more effectively make the many evaluative decisions they are already trying to make and will continue having to make anyway. Building systematic and disciplined evaluation into this field is not only possible and necessary but should be integral to all that information science and technology subfields are pursuing.

REFERENCES

Alkin, M. (Ed.) (2004). *Evaluation roots: Tracing theorists' views and influences.* Thousand Oaks, CA: Sage.

Cousins, J. B., Goh, S., Clark, S., & Lee, L. (2004). Integrating evaluative inquiry into the organizational culture: A review and synthesis of the knowledge. *Canadian Journal of Program Evaluation, 19(2)*, 99-141.

Fetterman, D. M. & Wandersman, A. (2005). *Empowerment evaluation principles in practice.* New York: Guilford.

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2004). *Program evaluation: Alternative approaches and practical guidelines.* Boston, MA: Pearson.

Fournier, D. M. (2005). Evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 139-140). Thousand Oaks, CA: Sage.

Lincoln, Y. S. & Guba, E. G. (2004). The roots of fourth generation evaluation. In M. Alkin, (Ed). *Evaluation roots: Tracing theorists' views and influences.* Thousand Oaks, CA: Sage Publications.

Mabry, L. (2005). Assessment. *Encyclopedia of evaluation* (pp. 22). Thousand Oaks, CA: Sage.

Mathison, S. (2005). *Encyclopedia of evaluation.* Thousand Oaks, CA: Sage.

Patton, M. Q. (2004). Utilization-focused evaluation: Theoretical underpinnings and origins. In M. Alkin, (Ed). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage Publications.

Petrosko, J. M. (2005). Measurement. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 247). Thousand Oaks, CA: Sage.

Scriven, M. (2004). Reflections. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences.* Thousand Oaks, CA: Sage.

Stake, R. E. (2003). *Standards-based and responsive evaluation.* Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2004a). The 21st-century CIPP model: Origins, development, and use. In M. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Stufflebeam, D. L. (2004b). A note on the purposes, development, and applicability of the Joint Committee

evaluation standards. *American Journal of Evaluation, 25*(1), 99-102.

Yanchar, S. & Williams, D. D. (2006). Reconsidering the compatibility thesis and eclecticism: Five proposed guidelines for method use. *Educational Researcher, 35*(9), 10.

KEY TERMS

Criteria: Ideals against which evaluands should be compared.

Evaluand: A thing or person being evaluated.

Evaluation: Judging merit or worth by comparing what is to what should be.

Formative Decision: Evaluation results suggesting ways to improve an evaluand.

Measurement: Process of identifying the existence of entities by categorizing or enumerating their qualities.

Stakeholders: People who have an interest in an evaluand and its evaluation and who must be involved in the evaluation so they will value and use the results.

Standard: Level on a criterion to which an evaluand is expected to reach.

Summative Decision: Evaluation results leading to decisions to continue or discontinue an evaluand.

Category: Artificial Intelligence

An Approach to Optimize Multicast Transport Protocols

Dávid Tegze
Budapest University of Technology and Economics, Hungary

Mihály Orosz
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The article presents an approach to optimize the multicast transport protocols. The main constraint of this procedure is the orthogonality (linear independence) of protocol parameters. Protocol parameters are variables defined for protocol classes, where the possible values of each parameter are protocol mechanisms, which serve the same goal in the multicast transport protocol. A multi-dimensional hyperspace of protocol parameters is stated, as a mathematical model of the optimization process where every transport protocol is represented as an individual point. A multicast transport *Sim*ulator for multi*Cast* (*SimCast*) has been developed to describe the performance of the transport protocols and to simulate the operation of these protocols for reliable multicasting. The simulator supports the protocol analysis in the hyperspace of protocol parameters.

BACKGROUND

Reliability is one of the most important features of all multimedia applications. This requirement may be especially critical in case of multicast, where the timely correction or resending of lost data is even more difficult because of the large volume of data to be transferred. Multimedia applications make multicast an active area of research. Multicasting is the one-to-many group communication way. For this purpose the IP-multicast transport is the preferred mechanism (Hosszú, 2005).

Since most of the multicast applications are media-related software, for example, media conference, voice distribution, shared whiteboard, and various collaborative media tools, they need more reliability than the best-effort delivery of Internet protocol (Adamson & Macker, 2001; Luby, & Goyal, 2004). In order to increase the reliability of multicast applications additional *multicast transport protocols* are used to achieve the required level of reliability (Whetten & Taskale, 2000). Such a protocol is the *NORM*: *N*ACK-*O*riented *R*eliable *M*ulticast Protocol (Adamson, Bormann, Handley & Macker, 2004a, 2004b).

It is hard to compare the various protocol mechanisms implemented in different protocols. Therefore, the modularly structured simulator *SimCast* (*Sim*ulator for multi*Cast*) is developed for traffic analysis of unicast (one-to-one) and multicast (one-to-many) streams (Orosz & Tegze, 2001). To carry out the necessary analysis of the unicast and multicast traffic, a well usable simulation program should be applied in order to present statistically correct results for multicast data transfer. The reason of developing a new, custom simulator instead of using a standard framework like *ns* (Breslau, Estrin, Fall, Floyd, Heidemann, Helmy, Huang, McCanne, Varadhan, Xu, and Yu, 2000) is that the architecture of the *SimCast* simulator is optimized for transport layer modeling and, due to its modular design, it is relatively easy to integrate new protocol mechanisms in it.

DECOMPOSITION OF THE MULTICAST TRANSPORT PROTOCOLS

Multicast transport protocols have many different properties for data delivery. These attributes can be represented by the previously mentioned *protocol parameters* (Hosszú, 2005). Each protocol parameter specifies different reliability mechanisms for the same delivery attribute. Such a protocol parameter is, for instance, the repair method, which can have the values like "retransmission", "forward-error correction", "interleaving", or different ways of "local receiver-based repairs" (Luby & Vicisano, 2004). Another parameter is the

acknowledgement type, which could hold the possible values "tree-based", "ring-based" or a "simple direct form".

Various applications have different reliability requirements and, therefore, these protocol parameters should be optimized in order to determinate the best-suited multicast transport protocol for a given application. However, to use any mathematical optimization method for the selection of the protocol parameters, a linearly independent (in other words *orthogonal*) set of parameters should be applied. For this purpose, a hyperspace of protocol parameters is created where each individual transport protocol corresponds to a point of this multi-dimensional space. The optimization procedure finds the most appropriate point in this space to provide the best performance for multicast content delivery. The possible values of the protocol parameters (which are classes of various mechanisms) are the realizations of specific protocol functionalities. A *quasi-orthogonal* subset of the protocol parameters and their possible values are presented in *Table 1*. These parameters represent the well-known reliability mechanisms of transport protocols.

To carry out a correct optimization procedure on the appropriately selected protocol parameters, a well usable simulation program should be applied in order to obtain statistically accurate results for multicast data transfer. Using a convenient simulator, the optimized transport protocol can be *synthesized* for a given media application, satisfying the requirements. This means that by means of a suitable mathematical method, an optimal point in the hyperspace of the protocol parameters can be found.

The optimization process should deal with the dependencies between protocol parameters. Most dependencies should be taken into account, but some of them can be omitted. An example of a negligible dependency is the relation between the protocol parameter *Feedback control* and the protocol parameter *Feedback addressee*, where modifying the actual value of the *Feedback control* from *Structure-based* to *Timer-based* could change the *optimal* value of *Feedback addressee* from *Intermediate host* to *Every member*, however, its influence may practically be ignored.

The developed simulator became a helpful tool for the analysis of multicast transport protocol mechanisms (Orosz & Tegze, 2001). By means of this simulator an optimized transport protocol can be synthesized, satisfying the requirements of certain type of media applications.

SIMULATING THE MULTICAST TRANSMISSION

The multicasting capabilities of the software are demonstrated by a distance vector based multicast protocol implemented in the simulator, which is very much like the existing DVMRP protocol (Distance Vector Multicast Routing Protocol). The concept of this protocol is similar to that of RIP (Routing Information Protocol) unicast routing protocol, which is one of the most common interior gateway protocols. This DVMRP like protocol determinates the multicast routing table by means of a flooding algorithm. Routers send *advertising packets* to the neighbor routers periodically to propagate routing information about them. Receiving such a packet, a router forwards the packet through all of its interfaces to the next router, except the receiving interface, and increments the hop counter field in the packet header. If the hop counter and source address of the received packet indicates a new route or a route better than the previous one, then the router adds it to its multicast routing table or modifies an existing table entry. To avoid infinite circulation of advertising packets, a TTL (Time To Live) like mechanism is used. This means that the source of the packet sets the TTL field of the packet to a positive integer. Each time the packet gets forwarded, this

Table 1. The selected set of the protocol parameters

Protocol parameter	Values
Flow control	Window-based, Rate-based, Multigroup multicast, Receiver give-up, None
Data accuracy	Reliable, Atomic, Non-reliable
Feedback addressee	Original source, Intermediate host, Every member, None
State control	Sender-based, Receiver-based, Shared, None
Feedback control	Structure-based, Timer-based, Representatives-based, Rate-based, None
Way of sending repair	Unicast, Multicast, None
Scope of repair	Global, Global to secondary group, Global to individual members, Local, None
Session membership control	Explicit, Implicit, None

counter is decremented. If this value reaches zero, the packet gets dropped. The implemented protocol can optimize the multicast distribution tree both for hop count and delay.

The administration of group membership in the local subnetwork between the local router and the hosts is performed via the IGMP (Internet Group Management Protocol). The IGMP implementation of *SimCast* is very much like the IGMP v1 (Deering, 1989). The combination of these two protocols makes it possible to carry out comprehensive simulations of multicast data delivery.

THE ARCHITECTURE OF THE SIMULATOR

Figure 1 shows the schematic of the network topology and the state messages generated by the simulator (in the grey box). The presented network topology is composed of 8 routers, and some hosts in their subnets. The routers are denoted by character "R" and their network layer addresses.

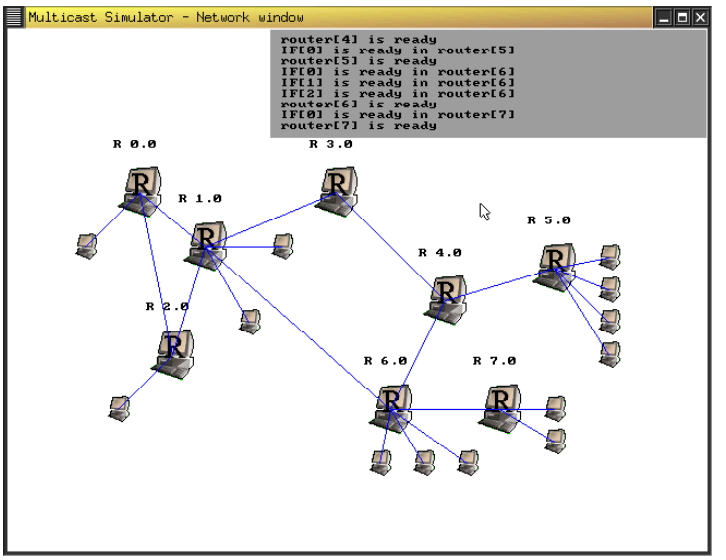
Simulations were carried out using various configurations of data traffic and network parameters. Depending on the packet queue sizes of the routers interfaces, the sending rate and the delay of the network links various congested situations can be simulated.

In simulator programs, basically two kinds of scheduling systems are used: *preemptive* and *non-preemptive* ones. In case of *preemptive* scheduling the *Scheduler* shares the real time among the scheduled objects and it is able to take control away from the scheduled objects after expiring the portioned time slice. Such solutions are applied in real-time and time-sharing systems, where, theoretically, the *Scheduler* can withdraw the right for running after any instruction. Therefore, the synchronization and the mutual exclusion problems must be eliminated during the simulation. In case of preemptive scheduling the states of the running objects must be saved by the *Scheduler*. This solution is not applicable in our case, since the simulated objects must step one simulation time unit in each simulation cycle. Active objects are scheduled with a predefined periodicity, which can be specified in the configuration.

In the case of *non-preemptive* systems, the *Scheduler* cannot take away the control from the scheduled objects, but they have to give it up themselves. In such systems the active objects are implementing a method "execute the next step". The Scheduler calls these methods periodically. After an object executed the necessary activity in the given time period, it saves its actual state in order to continue the run at the next activation according to these states.

In this manner the simulator sends a "go one time-step" message to the objects competing for run-time. Receiving such a message, an object recovers its running activity state based on its inner state attributes and after performing its tasks for the current time period, it updates the state attributes and gives the control back to the *Scheduler*. In such a way, non-preemptively scheduled objects return the control to the *Scheduler*.

Figure 1. The graphical output of the simulation



It is practical to describe the behavior of the simulated objects by state-space model. This model defines finite number of states with some state variables. The simulated protocol entities are deterministic, which means that the next state of an entity depends only on the inputs and the current state. Each of these objects implements the "execute the next step" method. These methods manage the state transitions and other activities of a particular object.

At the design phase the fact should be taken into account that the objects could wait for some event to occur longer than the length of the simulation time slice. In such situations waiting must be realized as cyclic polling. The method "execute the next step" should be implemented in such way that it cuts up the cycles of the state-space graph.

In the *SimCast* simulator scheduling of the simulation time among the routers, hosts and the Transmitter is non-preemptive. The preemptive scheduling is not applicable in the realization of the simulator, since it is important, that active objects should proceed exactly one fixed length time step. This is not guaranteed in the case of preemptive scheduling, since the *Scheduler* can interrupt the run anytime.

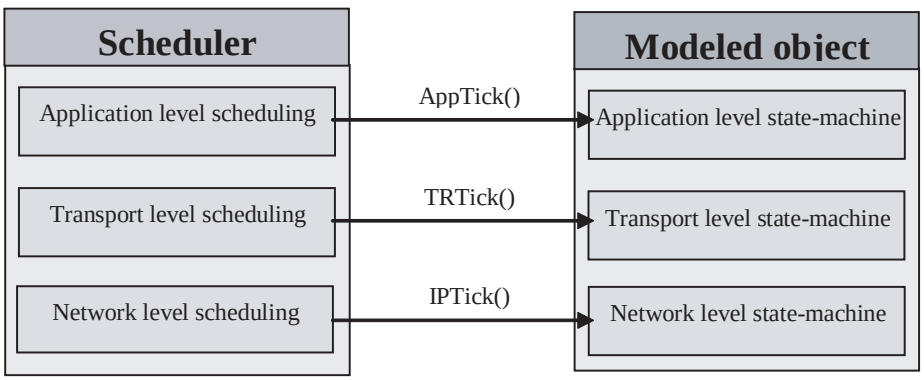
The simulated protocol entities activated with a preconfigured frequency, which is characteristic for the simulated protocol layer. The activation of protocol objects is performed in a *round robin* fashion within a protocol layer. In this way, *SimCast* implements a multi level scheduling system. The levels of the scheduling reflect the modeled protocol layers: network layer, transport layer (transport protocol entities) and application layer, as *Figure 2* shows.

The appropriate protocol entities are generated dynamically in run time according to the load-in configuration. The scheduler cannot make the difference between the various protocol entities, because the scheduling is implemented using abstract methods. Since the scheduling interfaces are realized with abstract methods, additional protocol entities can be built into the system without the modification of the core system, if the new protocol entities implement the scheduling interface defined by existing abstract protocol classes. This object oriented design results in very modular simulator system.

The simulator implements the *Manager* object, which is responsible for loading configuration data, as well as the creation and initialization of other simulation objects. After system initialization *Manager* gives control to the *Scheduler* object. Descendant classes of the abstract Sender object model the sender protocols. At transport layer the *Sender* actuates a repair buffer. The size of this buffer is controlled by some inner state variables of the simulated transport layer entities. This buffer is also known as *congestion window.* It is also the task of the *Sender* entities to handle the retransmission timers, which are needed to detect packet losses. On the receiver side the transport layer, receiver protocol reassembles the data stream sent by the sender side. The receiver application can access this stream through the *Receiver Socket* object. It is also the task of the receiver side transport layer entity to generate acknowledgement packets. *SimCast* adopts various acknowledgement methods depending on the simulated transport protocol. Receiver also handles negative or positive acknowledgements. When one uses positive acknowledgement, delayed or normal acknowledgement can be chosen as a further parameter. It is also the liability of the transport layer protocol to assemble and disassemble data stream from sent and received segments together with the multiplexing and demultiplexing of data between the communicating sockets based on the port numbers.

Socket objects make contact between application layer protocols and transport layer entities. In case of connection oriented transport protocols like TCP and TFRC entities, ap-

Figure 2. Canonical model of the simulation



plication layer senders and receivers are handling data sent to and received from sockets as data streams. The functions required to exchange data with sockets are very similar to standard file system read and write functions.

Descendants of the abstract *ProNetwork* class implement the model of network layer. Network interfaces incorporate input and output buffers also known as FIFO objects, which have limited capacity. As an input FIFO reaches its full capacity further incoming packets are lost. This behavior models the *drop-tail* property of existing network routers. The capacity of the FIFOs are given in packets.

Host, *Router*, *Interface* and *Wire* objects realize the model of physical layer. The bandwidth and bit error rate are typical attributes of this layer. The maximal sending rate is limited by the properties of *Interface* and *Wire* objects. *SimCast* gives us the opportunity to change these attributes in run-time. Therefore, we can run simulations with predefined, variable network conditions. In this way, we have the possibility to perform simulations based on network conditions measured in real world networks or imported from other network simulators.

The *Logger* subsystem is activated in each time slice, which ensures that all relevant data of the simulation is stored. The post-processing phase of the analysis extracts several useful network statistics based on this data.

The *SimCast* software is written in object-oriented C++ and it is compiled by GNU C++ under Linux operating system and runs on the Xwindow graphical user interface. The simulator uses the Allegro platform independent graphical framework. The simulator is portable to the following operating systems: Linux, MS Windows, DOS, BeOs.

FUTURE TRENDS

The importance of the network simulation is obvious. The larger the communication network is, the harder to estimate its behavior. Analytic models are useful for a coarse level of examination, but the number of problems they can tractably solve is limited (Nicol & Liu, 2005). Detailed discrete-event simulations remain a useful tool for examining computer networks.

In case of IP-multicast, the difficulty of the problem regarding network simulation is more dramatic, due to the possible large number of participants. Because of the complexity of the problem, developing a specially designed simulator seemed to be more efficient than using a more general product, which may not be suitable for the multicast transport protocol optimization.

It is an important advantage of the *SimCast* simulator, that the latest congestion control mechanisms are also implemented in it. In this way, the cooperation among different multicast transport protocols and the widely used TCP protocol entities or various other unicast transport level protocols can be examined (Yaun, Bhutada, Carothers, Yuksel, and Kalyanaraman, 2003).

Based upon the performed experiments, it became clear that the special purpose simulator is an appropriate way to examine multicast transport protocols. However, to model the operation of the transport protocols on the Internet realistically, a more sophisticated network model should be built.

CONCLUSION

The simulation results showed that the presented multicast simulator can simulate the various multicast transport protocol mechanisms and the obtained simulation results are suitable for further processing. The simulator serves as a testbed for multicast transport protocol synthesis. It implements a simplified model of the network layer and a detailed model of the multicast transport. Furthermore, it can accept the statistical network attributes obtained from other, specific network simulators. Therefore, it can be considered as a valuable tool for multicast traffic analysis and protocol optimization.

The IP-Multicast itself is a smooth extension of the basic IP-unicast paradigm to mass-communication. However, the already elaborated sophisticated procedures and methods should be configured precisely in order to achieve the required efficiency in their performance. That is why the simulation tools give special advantages in the field of the reliable multicast communication research.

REFERENCES

Adamson, B., Bormann, C., Handley, M., and Macker, J. (2004a). Negative-Acknowledgement (NACK)-Oriented Reliable Multicast (NORM) Protocol, IETF Network Working Group, RFC 3940.

Adamson, B., Bormann, C., Handley, M., and Macker, J. (2004b). Negative-Acknowledgement (NACK)-Oriented Reliable Multicast (NORM) Building Blocks, IETF Network Working Group, RFC 3941.

Adamson, R., and Macker, J. (2001). A TCP Friendly Rate/ Based Mechanism for Nack/Oriented Reliable Multicast Congestion Control, Proceedings of IEEE GLOBECOMM 2001.

Breslau, L., Estrin, D., Fall, K., Floyd, S., Heidemann, J., Helmy, A., Huang, P., McCanne, S., Varadhan, K., Xu, Y., and Yu, H. (May, 2000). Advances in network simulation. IEEE Computer, 33:5, (pp. 59-67). IEEE Computer Society Press, Los Alamitos, CA, USA.

Deering, S.E. (1989). Host extensions for IP multicasting, Network Working Group RFC 1112, Aug.

Hosszú, G. (2005). Current Multicast Technology. In M. Khosrow-Pour (Ed.), Encyclopedia of Information Science and Information Technology Vol. I-V, (pp. 660-667). Hershey, PA: Idea Group Reference.

Levine, B.N.and Garcia-Luna-Aceves, J. J. (1998). A Comparison of Reliable Multicast Protocols ACM Multimedia Systems 6:5 (pp. 334-348) ACM Press, New York, NY, USA.

Luby, M., and Goyal, V. (2004). Wave and Equation Based Rate Control (WEBRC) Building Block, IETF Network Working Group, RFC 3738.

Luby, M., and Vicisano, L. (2004). Compact Forward Error Correction (FEC) Schemes, *IETF Network Working Group, RFC 3695*.

Nicol, D.M., and Liu, J. (2005). Advanced Concepts in Large-Scale Network Simulation. In *Proceedings of the 2005 Winter Simulation Conference*, (pp. 153-166). Informs Simulation Society.

Orosz, M., and Tegze, D. (September 19-20, 2001). The SimCast Multicast Simulator. In *Proceedings of the International Workshop on Control & Information Technology, IWCIT'01* (pp. 66-71). Ostrava, Czech Republic.

Pejhan, S., Schwartz, M. and Anastassiou, D. (June, 1996). Error control using retransmission schemes in multicast transport protocols for real-time media *IEEE/ACM Transactions on Networking*, vol. 4(3), (pp. 413—427), New York, NY, USA.

Yaun, G.R., Bhutada, H.L., Carothers, C.D., Yuksel, M., and Kalyanaraman, S. (July, 2003) Large-Scale Network Simulation Techniques: Examples of TCP and OSPF Models, *ACM SIGCOMM Computer Communication Review, 33:3,* (pp. 27-41). ACM Press, New York, NY, USA.

Whetten B. & Taskale G. (2000). The Overview of Reliable Multicast Transport Protocol II, *IEEE Network.*

Widmer, J. and Handley, M. (August, 2001). Extending equation-based congestion control to multicast applications. *Proceedings of ACM SIGCOMM*, pages 275--286, San Diego, California.

KEY TERMS

Congestion Window Size: This is the maximal allowed amount of unacknowledged data to be sent to the network.

FIFO: This is the way packets stored in a queue are processed. Each packet in the queue is stored in a queue data structure. The first data to be added to the queue will be the first data to be removed, then processing proceeds sequentially in the same order.

Hyperspace of Protocol Parameters: This abstract space is composed of the possible values of each property of the multicast transport protocols. The values represent various protocol mechanisms.

IP-Multicast: Network-level multicast technology, which uses the special class-D IP-address range. It requires multicast routing protocols in the network routers. Its other name: *Network-level Multicast (NLM).*

Multicast Routing Protocol: In order to forward the multicast packets, the routers have to create multicast routing tables using multicast routing protocols.

Multicast Transport Protocol: To improve the reliability of the multicast delivery special transport protocols are used in addition to the unreliable *U*ser *D*atagram *P*rotocol (*UDP*).

Pre-Emptive Scheduling: In case of this scheduling method scheduler can take away execution right from an active object at any time.

Socket: A socket is an abstraction, designed to provide a standard application programming interface for sending and receiving data across a computer network. Sockets are designed to accommodate virtually any networking protocol.

TTL: *Time-to-live,* a field in the IP packet header. Its value is the allowed hop-count, the number of routers, which can forward the packet before delivery or dropping out.

Unicast Transport Protocol: They handle the ports in each computer, or improve the reliability of the unicast communication. As examples, the *U*ser *D*atagram *P*rotocol **(***UDP***)** is a simple unicast transport protocol mainly for the port-handling, and the *T*ransmission *C*ontrol *P*rotocol (*TCP*) is intended for the reliable file transfer.

Approaches to Telemedicine

José Aurelio Medina-Garrido
Cadiz University, Spain

María José Crisóstomo-Acevedo
Jerez Hospital, Spain

INTRODUCTION

Information technologies have become essential for most businesses, including those in the healthcare industry (Chau & Hu, 2004; Rodger & Pendharkar, 2000). Information technologies can improve both the delivery of the healthcare service and certain aspects of healthcare centers' administration.

There has been a proliferation of *information systems applied to the health sector*, such as hospital information systems, medical decision-support systems, systems for interpreting medical tests and images, expert systems based on the handling of medical knowledge, or telemedicine (Rao, 2001).

Etymologically, the term *telemedicine* means medicine from a distance. This concept can include something as simple as two healthcare professionals debating the case of a patient by telephone, or as complex as conducting the diagnosis of a patient remotely using videoconference.

Telemedicine implies that there is an exchange of information, without personal contact, between two physicians or between a physician and a patient. Thanks to telecommunications technologies, telemedicine enables the provision of healthcare services or the exchange of healthcare information across geographic, temporal, social, and cultural barriers (Chau & Hu, 2004). Telemedicine makes use of a wide range of technologies to overcome distances, such as radio, analog landlines, e-mail, the Internet, ISDN, satellites,telesensors, and so forth, for the transmission of medical information, (data, voice, and video) and provision of medical services from a distance.

With regard to the *transmission of medical information*, this includes the digital handling of patient information (for example, from their *electronic medical records*), or the transfer of images (such as radiographs, high-resolution medical images, computer tomography scans, magnetic resonance imaging pictures, ultrasound images, electrocardiograms or echocardiograms, video images of endoscopic cameras, etc.) or sounds (for example, from electronic stethoscopes) (Rao, 2001).

With regard to the provision of remote medical services, specialist physicians can see their patients in consultation, conduct medical examinations, arrive at a diagnosis and prescribe treatment, all without needing to be in actual physical contact with them.

The essence of *telemedicine* is to move the medical knowledge and experience rather than move the patient physically. For this, telemedicine involves rather more than just taking medical services to where they did not exist before. It has also become a practice of transmitting and handling knowledge. It enables medical practitioners to exchange their knowledge (Robinson, Savage & Campbell, 2003) so that others can apply it in specific situations.

We should not confuse telemedicine with e-health (or tele-health). Telemedicine only refers to the provision of medical services. *E-health*, on the other hand, refers to both medical services and any other type of service, as long as it has something to do with health and employs information technology. In this respect, e-health would also include healthcare educational activities, research in the health sciences, the handling of electronic files in the healthcare system, and any other use of information technologies in the healthcare system.

The rest of this article is organized as follows. The second section discusses the antecedents of telemedicine, and proposes two taxonomies, one in function of the temporal synchronization of the individuals using it, and the other in function of the medical specialty for which it is employed. The third section tries to identify the obstacles in the way of an adequate acceptance and development of telemedicine. Before the conclusions section, section four suggests some future trends, including what technologies are most in use at present and which ones are promising for the future.

BACKGROUND

The concept of *telemedicine* does not actually require the use of information technologies. Indeed, it was common in the past to exchange medical opinions and prescribe treatments using mail, the radio, or even visual signals. People living in remote areas of Australia at the beginning of the 20th century used radio to communicate with the Royal Flying Doctor Service of Australia. At this time, physicians

on dry land also used the radio to communicate with ships suffering from medical emergencies (Wootton, 2001). Some African villages used smoke signals to warn outsiders not to approach the village during an epidemic. Similarly, ships used flags to warn that they were in quarantine (Darkins & Cary, 2000).

Nevertheless, modern IT has given new meaning to the practice of *telemedicine* (Bladwin, Clarke & Jones, 2002). The majority of projects have shown that technology enables the exchange of medical information both in urban and rural areas (Thames, 2003). The University of Nebraska was one of the pioneers, in 1959, when its scientists transmitted neurological examinations within the campus, and again in 1964, when it ran a telemedicine project with a distant mental hospital. NASA's activities in telemedicine are more familiar. In the 1960s the space agency used a satellite as part of a telemedicine project in the Appalachian and Rocky Mountain regions and Alaska. In the same decade, NASA monitored the pulse and blood pressure of the first astronauts remotely while they were in space (Rao, 2001). In the 1970s, telemedicine evolved further thanks to satellite technology, which, for example, enabled isolated villages in Alaska and Canada to be connected with distant hospitals (Rao, 2001).

A first taxonomy can differentiate between two types of practice in *telemedicine and* in function of the temporal synchronization of the users. In this respect, we can identify synchronous (or real-time) telemedicine and asynchronous telemedicine.

Synchronous telemedicine can range from a simple telephone conversation to a robot-assisted surgical operation. Videoconference is very often combined with devices for monitoring and diagnosing patients remotely.

Asynchronous telemedicine involves previously storing medical information and then transmitting it to the appropriate medical specialist. As we can see, the two parties need not coincide at the same time.

Telemedicine can also be classified in function of the specialties to which it is applied. In this respect, we might mention the following specialties (Tachakra, 2003):

- *Telenursing*. Nurses tend to use telemedicine in two ways. They can process patient data in a database in order to monitor patients undergoing medical treatment and refer them to the appropriate medical services. Another possibility is to monitor patients remotely, for example, in their own homes, using electronic medical devices, and interactive video applications. Nurses prompt the patients to take their medicine, blood pressure, temperature, and so forth.
- *Teleradiology*. This telemedicine specialty involves sending electronic radiology images such as X-rays, computerized axial tomography scans, or magnetic resonance images.
- *Telepathology*. This specialty involves transmitting high-resolution images of microscope slides, photographs of lesions or smears, and so forth.
- *Teledermatology*. This involves transmitting images of the skin using a dermascope.
- *Telecardiology*. Transmitting information relating to electrocardiograms, echocardiograms, angioplasty, and cardiac pacemaker monitoring.
- *Telesurgery*. The surgeon operates on the patient remotely using robotics and audio/video devices.
- *Video teleconferencing.* Videoconference allows physicians to attend to their patients, conduct diagnoses, and offer treatments remotely.

OBSTACLES TO THE ADOPTION AND USE OF *TELEMEDICINE*

There is no doubt that telemedicine offers considerable advantages to the population in general, and to certain patients in particular (the chronically ill, elderly, population of rural areas, etc.). Physicians can also gain by accessing digital information that can help them in their diagnoses and treatments, as well as allowing them to exchange opinions with expert colleagues. Nevertheless, the use of telemedicine is advancing only slowly, and healthcare professionals have shown some reluctance to embrace it (Audet, Doty, Peugh, Shamasdin, Zapert, & Schoenbaum, 2004; Parente, 2000; Sands, 2004; Wilson, 2005). This section analyzes the main factors that may explain this slow development and practitioners' resistance to accept these technologies.

The most important obstacles to the development of telemedicine are as follows (Adams, 2001; Kirsch, 2002; Lumpkin, 2000; Miller & Derse, 2002; Parente, 2000; Rao, 2001): the difficulty in making money from it; lack of technological infrastructure; lack of standardization; unequal access to the Internet; insufficient legislation; reprisals from traditional healthcare organizations; and cultural barriers.

It has been argued that telemedicine will offer important business opportunities (Parente, 2000). But the healthcare sector is having difficulties making money from it (Kirsch, 2002), due to: government intervention in financing and regulating the healthcare system; physicians' reluctance to pay for information services; the lack of incentives for healthcare professionals to improve their productivity and the quality of their work; and the fact that patients are often captives of public health systems with traditional procedures.

Another problem is the lack of the necessary technological infrastructure. The healthcare sector has been characteristically slower to adopt new information technologies than other service-sector industries (Parente, 2000). In general, the different healthcare centers are rarely interconnected digitally (Rao, 2001).

The lack of standardization of *clinical data* has been identified by some authors as one of the main technical obstacles to telemedicine (Lumpkin, 2000). The information contained in medical records, consultations, diagnoses, treatments, and administrative aspects of the healthcare sector is rarely recorded digitally, and, when it is recorded, independent, fragmented, and unintegrated information systems tend to be used, employing different, incompatible formats. There is also a lack of standards for the digital transmission of clinical images (Rao, 2001) and for the electronic exchange of clinical data. With regard to this last point, the lack of standards for exchanging clinical data, two types of behaviors are evident in organizations. On the one hand, we find pioneers and innovators that invest large sums and commit themselves to a particular standard. On the other, a large number of organizations make no such investments; they simply wait for the market to decide which standard will predominate. These organizations act as free riders, holding back technological development in telemedicine as a whole.

Another technical obstacle is the unequal access to the Internet. The Internet is widely available, but it is not used everywhere. The use of information and communications technologies promises to reduce the disparities in the population – caused by demographic and socio-economic factors – in access to healthcare services (Ahern, Kreslake & Phalen, 2006; Cashen, Dykes & Gerber, 2004; Gibbons, 2005). But it is precisely the most deprived individuals who have least access to these technologies.

Another threat to the development of telemedical practice is the absence of global legislation for the provision of what can be a global service. In this digital environment, some patients may obtain diagnoses, treatments, or prescriptions for drugs from physicians who may not have a valid medical license in the state or country where the patient lives. The lack of a well-designed public policy, along with market incentives and consumer demand, could lead to a telemedical practice that undermines patients' well-being, the quality of the healthcare service, the nature of the patient-physician relationship, and the integrity of the medical profession (Miller & Derse, 2002).

A legal problem with the implementation of telemedicine concerns the protection of patients' privacy (Adams, 2001; Lumpkin, 2000; Miller & Derse, 2002). Consumer privacy laws could restrict firms' handling of *patients' clinical data*, or even prevent it altogether, either because this information could be consulted illegally or because the firm could pass it on to third parties without the patient's authorization. With regard to illegal access to the data, the Internet has advanced sufficiently to store information securely using encryption methods. With regard to making the information available to third parties, there is some protection, but there is a legal vacuum since most countries are applying the existing legislation, which was conceived with the old health economy in mind.

From the perspective of competitive strategy we must bear in mind that traditional healthcare organizations lacking a presence in the new virtual economy may retaliate against their new rivals. While some big business successes on the Internet had no traditional competitors in the market, the healthcare sector of the old economy is a mature, concentrated sector, and it could fight back against this virtual threat. For this reason, it is perhaps more feasible for these traditional organizations to gradually increase their presence on the Internet, offering patients and healthcare professionals the right balance between the virtual and the physical. Nevertheless, and as we have mentioned, the technological progress of the old healthcare economy is proving slow.

Finally, a number of cultural barriers and misconceptions, among both physicians and patients, hinder the adoption of telemedicine. Patients do not have sufficient confidence to deposit and update their *clinical data* on the Internet. This undoubtedly requires efforts to raise patient confidence, involving the right legal and technological protection, and a cultural change. This latter point may perhaps be the most difficult of all, but the increasing use of credit cards for online purchases shows that change is possible. Physicians, in turn, often argue that they do not have time to attend to patients remotely, as they are too busy with the consultations and medical services they provide to the patients who see them in person. This is a shortsighted view from the strategic perspective, since they would be able to attend to more patients if they automated medical services wherever possible, for example in prescribing drugs for the chronically ill.

FUTURE TRENDS

Despite the barriers to the adoption of telemedicine discussed in the previous section, the application of information and communications technologies to health is growing gradually. Thus, telemedicine is now being used, for example, to attend to patients' calls from *medical call centers*, for remote consultations and prescriptions of treatments, or to send microscope images to colleagues (Eckhardt, Roulet, Schneeberger, Stauffacher, & Stump, 2004).

However, some telemedicine applications are still incipient or insufficiently employed. This is the case of remote monitoring, or telemonitoring. *Telemonitoring* is fundamentally used to control and treat chronic patients (Eckhardt, et al., 2004).

Another incipient trend is that of online healthcare advice and drug prescription. But prescribing drugs on the Internet can be extremely controversial (Coile, 2000).

Although e-mail is being increasingly used in the workplace and privately, it is still rare in communications between patients and their physicians, having grown only modestly

in recent years (Audet, et al., 2004; Sands, 2004). The use of e-mail in the physician-patient relationship is recognized as a useful tool for improving communication between both (Car & Sheikh, 2004; Patt, Houston, Jenckes, Sands, & Ford, 2003; Moyer, Stern, Dobias, Cox, & Katz, 2002). E-mail use in physician-patient communications is likely to increase in the future, just as this technology is increasingly used in communications between firms and private individuals (Brooks & Menachemi, 2006).

Citizens are very accepting of electronic communication, including e-mail, voicemail and instant messaging, and its use may provide opportunities to improve patient monitoring and education. The Internet and short message services (SMS), it has been suggested, may be potentially useful tools for improving self-monitoring of asthma or diabetes, and outbreak management.

Classical mobile telemedicine means using portable mobile computer sets with medical equipment and satellite links. But the new conception of mobile telemedicine means quick, easy, and useful access to telemedical technologies. The telemedicine tools should be in the doctor's pocket, alongside the stethoscope, pencil, or scalpel. The development and improvement of mobile technologies (PDAs, cell phones with digital cameras or videocameras, cell phones with message services (MMS/SMS), GPRS Internet, etc.) allows medical doctors to be online 24 hours a day.

Nevertheless, the most promising application of telemedicine is in the provision of medical services to people living in geographically-isolated areas (Rao, 2001) or developing countries (O'Neill, 2001) who lack access to healthcare services. Paradoxically, however, although developing countries have the most to benefit from telemedicine, they are the ones that can least afford it (Jarudi, 2000), and they also have the poorest telecommunications infrastructures.

CONCLUSION

Thanks to telemedicine some patients can now access medical services and experience that would otherwise not be available to them. This is particularly important for people in rural areas (Thames, 2003), in prison, on board ship (Rao, 2001; Wootton, 2001), in a war zone (Rodger & Pendharkar, 2000), or in any other situation that prevents them from accessing healthcare services normally (O'Neill, 2001). It will be equally valuable for patients suffering from rare diseases (Miller & Derse, 2002), or people wanting a second medical opinion who need to look for a particular specialist some distance away.

There are undoubtedly obstacles to the adoption of telemedicine activities. These obstacles are hindering the rapid development that this activity could and should be enjoying. They include, most notably, the difficulty in making money from telemedicine, the lack of technological infrastructure, the lack of standardization in the data, files and images being exchanged electronically, unequal access to the Internet among the population, insufficient legislation for the accredited practice of telemedicine, or to protect patients' privacy, competitive retaliation from the traditional healthcare organizations, and cultural barriers among patients and physicians.

Nevertheless, these barriers are similar to those that have already been overcome in other industries outside healthcare. Healthcare organizations are increasingly running telemedicine projects. The advantages of using telemedicine and other e-health applications, together with the unstoppable development of information and communications technologies, make it likely that these technologies will grow considerably in the medium to long term.

REFERENCES

Adams, D. (2001). E-mail One of Technology Benefits Touted at AIA Forum. *American Medical News*, April 16.

Ahern, D.K., Kreslake, J.M., & Phalen, J.M. (2006). What Is eHealth: Perspectives on the Evolution of eHealth Research. *Journal of Medical Internet Research*, 8(1), e4.

Audet, A.M., Doty, M.M., Peugh, J., Shamasdin, J., Zapert, K., & Schoenbaum, S. (2004). Information technologies: when will they make it into physicians' black bags? *Medscape General Medicine*, 6(4), 2.

Bladwin, L.P., Clarke, M., & Jones, R. (2002). Clinical ICT systems: Augmenting case management. *Journal of Management in Medicine*, 16(2/3), 188-198.

Brooks, R.G., & Menachemi, N. (2006). Physicians' Use of Email With Patients: Factors Influencing Electronic Communication and Adherence to Best Practices. *Journal of Medical Internet Research*, 8(1), e2.

Car, J., & Sheikh, A. (2004). E-mail consultations in health care: scope and effectiveness. *British Medical Journal*, 329(7463), 435-438.

Cashen, M.S., Dykes, P., & Gerber B. (2004). eHealth technology and Internet resources: Barriers for vulnerable populations. *Journal of Cardiovascular Nursing*. 19(3), 209-222.

Chau, P.Y.K., & Hu, P.J. (2004). Technology Implementation for Telemedicine Programs. *Communications of the ACM*, 47(2), 87-92.

Coile, R.C. (2000). E-health: Reinventing healthcare in the information age. *Journal of Healthcare Management*, 45(3), 206-210.

Darkins, A.W., & Cary, M.A. (2000). *Telemedicine and Telehealth: Principles, Policies, Performance, and Pitfalls*. New York: Springer Publishing.

Eckhardt, A., Roulet, M., Schneeberger, K., Stauffacher, W., & Stump, D. (2004). *Telemedicine – utilising the opportunities*. Berne: SATW.

Gibbons, M.C. (2005). A Historical Overview of Health Disparities and the Potential of eHealth Solutions. *Journal of Medical Internet Research*, 7(5), Article e50.

Jarudi, L. (2000). Doctors without borders. *Harvard International Review*, 22(1), 36-39.

Kirsch, G. (2002). The business of eHealth. *International Journal of Medical Marketing*, 2(2), 106-110.

Lumpkin, J.R. (2000). E-health, HIPAA, and beyond. *Health Affairs. Chevy Chase*, 19(6), 149-151.

Miller, T.E., & Derse, A.R. (2002). Between strangers: The practice of medicine online. *Health Affairs. Chevy Chase*, 21(4), 168.

Moyer, C.A., Stern, D.T., Dobias, K.S., Cox, D.T., & Katz, S.J. (2002). Bridging the electronic divide: Patient and provider perspectives on e-mail communication in primary care. *The American Journal of Managed Care*, 8(5), 427-433.

O'Neill, R. (2001). Internet Brings Medicine to Remote Cambodian Village. *Associated Press*. February 15.

Parente, S.T. (2000). Beyond the hype: A taxonomy of e-health business models. *Health Affairs. Chevy Chase*, 19(6), 89-102.

Patt, M.R., Houston, T.K., Jenckes, M.W., Sands, D.Z., & Ford, D.E. (2003). Doctors who are using e-mail with their patients: A qualitative exploration. *Journal of Medical Internet Research*, 5(2), e9.

Rao, S.S. (2001). Integrated health care and telemedicine. *Work Study*, 50(6/7), 222-228.

Robinson, D.F., Savage, G.T., & Campbell, K.S. (2003). Organizational learning, diffusion of innovation, and international collaboration in telemedicine. *Health Care Management Review*, 28(1), 68-92.

Rodger, J.A., & Pendharkar, P.C. (2000). Using telemedicine in the Department of Defense. *Communications of the ACM*, 43(3), 19-20.

Sands, D.Z. (2004). Help for physicians contemplating use of e-mail with patients. *Journal of the American Medical Informatics Association*, 11(4), 268-269.

Tachakra, S. (2003): *Telemedicine and e-Health. In Business Briefing: Global Healthcare 2003*. World Medical Association.

Thames, T. (2003).Telemedicine: The road to higher quality health care. *Vital Speeches of the Day*, 70(2), 53.

Wilson, P. (2005). *My Health / My eHealth. Meeting the challenges of making eHealth personal*. Presented at ICLM9. Brazil, September.

Wootton, R. (2001). Telemedicine. *British Medical Journal*, 323, 557-616.

KEY TERMS

Computerized Axial Tomography Scans: A medical imaging method used to generate a three-dimensional image of the internals of an object from a large series of two-dimensional X-ray images.

E-Health: The provision of any healthcare service that is supported by electronic processes and communications.

Electronic Medical Records: Computer-based patient medical records. Patient medical records are a systematic documentation of a patient's medical history and care.

Telecardiology: The digital transmission between healthcare professionals of information relating to electrocardiograms, echocardiograms, angioplasty, and cardiac pacemaker monitoring.

Teledermatology: The digital transmission between healthcare professionals of images of the skin using a dermascope.

Teleendoscopy: The digital transmission between healthcare professionals of the results of endoscopic examinations.

Telemedicine: The use of information and communications technologies to exchange information between practitioners, or to deliver medical services to a patient remotely.

Telemonitoring: The remote monitoring of patients' state of health. It is fundamentally used to control and treat chronic patients.

Telenursing: Healthcare services provided by nurses remotely, such as monitoring patients in their homes, or referring patients to the appropriate medical services through the processing of patient data.

Telepathology: The digital transmission between healthcare professionals of high-resolution images of, for example, microscope slides, photographs of injuries or smears, etc.

Teleradiology: The digital transmission between healthcare professionals of electronic radiology images such as

X-rays, computerized axial tomography scans, or magnetic resonance images.

Telesurgery: Remote surgery using robotics and audio/video devices.

Architecture Methods and Frameworks Overview

Tony C. Shan
Bank of America, USA

Winnie W. Hua
CTS Inc., USA

INTRODUCTION

The e-business models in today's globalized business world demand ever-increasing flexibility, responsiveness, and agility of information technology (IT) solutions. It is compulsory for the IT group to provide higher levels of services at a lower cost for the business to compete and succeed. The reality to IT is that there is no choice other than to build more complex, flexible, scalable, extensible, innovative, and forward-thinking technical solutions, to satisfy the growing business needs.

In large organizations like worldwide financial institutions, virtually thousands, if not millions, of IT applications and systems have been constructed or purchased to provide electronic services for external customers and internal employees in the past years, utilizing heterogeneous technologies and architectures to meet diverse functional requirements from different lines of business. In the banking industry, as an example, the business process generally contains different business sectors in consumer, commercial, small business, wealth management, and capital management. In particular, services are delivered to different channels such as automated teller machines (ATMs), Web browsers, interactive voice response, agent assistance, e-mails, mobile devices, and so on. To effectively manage the architecture assets and rationalize the architecture designs in such a diverse environment, a multi-disciplinary engineering approach is of crucial importance to abstract concerns, divide responsibilities, mitigate risks, encapsulate the complexity, reverse-engineer existing applications, identify reengineering opportunities, and conduct objective technology assessments, which leads to in-depth technical recommendations and rationalization action plans.

BACKGROUND

The computing environment has gone through a number of generations of evolution in the last few decades, ranging from monolithic, client/server, multi-tier, object-oriented, component-based, service-oriented, event-driven, to social computing models. The overall solution architecture has become increasingly complicated and thus hardly manageable through a traditional waterfall process. Previous studies (DoD, 1997; IEAD, 2004; Kruchten, 2003; OMG, 2007; Putman, 2001; The Open Group, 2007; Zachman, 1987) in the past have strived to address the issue of architecture design complexity, which has grown exponentially as the computing space is transformed to a service-oriented architecture paradigm.

The architecture methods and frameworks are the general or proven approaches to designing/developing architecture of information systems. They have progressively undergone an evolutionary growth in the last 20 years. The prominent architecture methods and frameworks developed and proposed so far are listed as follows:

- **Zachman Framework**
- **E2AF:** Extended Enterprise Architecture Framework
- **TOGAF:** The Open Group Architecture Framework
- **RUP:** Rational Unified Process, evolved to Enterprise Unified Process and OpenUP
- **MDA:** Model-Driven Architecture
- **Microsoft Solutions Framework (MSF)**, and **Microsoft Systems Architecture (MSA)**
- **C4ISR:** Command, Control, Computers, Communications (C4), Intelligence, Surveillance, and Reconnaissance (ISR).
- **FEA:** Federal Enterprise Architecture Framework
- **TEAF:** Treasury Enterprise Architecture Framework
- **PERA:** Purdue Enterprise Reference Architecture
- **RM-ODP:** Reference Model for Open Distributed Processing
- **ATAM:** Architecture Tradeoff Analysis Method
- **SAAM:** Software Architecture Analysis Method
- **IDEF:** Integrated Definition Methods
- **MODAF:** Ministry of Defense Architectural Framework

The design principles for the development of architecture methods and frameworks are specified in the following section, with detailed articulations on each method/framework provided in the subsequent section. A comparison matrix is presented afterwards, whereas future trends and recommendations are elaborated next, followed by the conclusions.

DESIGN PHILOSOPHY

The following design principles have been generally applied in developing architecture methods and frameworks:

- Information processing activities comply with applicable laws, orders, and regulations.
- Business objectives are well defined before initiating information technology solutions.
- Total business value is the primary objective when making information technology decisions.
- Architectural selections maximize the interoperability and reusability.
- Architecture methods take advantage of standardization to fulfill common customer requirements and to provide common functions.
- Information technology groups collaborate to share information, data, services, components, and infrastructure required by the business units.
- Business and information technology requirements adopt matured commercial off-the-shelf (COTS) technology where appropriate rather than customized or in-house solutions.
- Information, services, applications, systems, and infrastructure are vital assets that must be managed, controlled, and secured in a holistic manner.
- Enterprise architecture (EA) is consistent with the guidance and strategic goals at the divisional levels.

METHODS AND FRAMEWORKS

The major architecture methods and frameworks are discussed in greater detail in this section.

Zachman Framework

The Zachman Framework (Zachman, 1987) is a logical structure used to categorize and organize the descriptive representations of an enterprise IT environment, which are significant to the organization management and the development of the enterprise's information systems. It takes the form of the two-dimensional matrix, and has achieved a level of penetration in the domain of business and information systems architecture and modeling. It is mainly used as a planning or problem-solving tool. However, it tends to implicitly align with the data-driven approach and process-decomposition methods, and it operates above and across the individual project level.

E2AF

Extended Enterprise Architecture Framework (E2AF) (IEAD, 2004) takes a very similar approach to the Zachman Framework. Its scope contains business, information, system, and infrastructure in a 2-D matrix. E2AF is more technology-oriented. Both Zachman Framework and E2AF approaches are heavyweight methodologies, which necessitate a fairly steep learning curve to get started in an organization.

TOGAF

Another heavyweight approach, The Open Group Architectural Framework (TOGAF) (The Open Group, 2007), is a detailed framework with a set of supporting tools for developing an enterprise architecture to meet the business and information technology needs of an organization. The three core parts of TOGAF are *architecture development method*, *enterprise architecture continuum*, and *TOGAF resource base*. The scope of TOGAF covers *business process architecture*, *applications architecture*, *data architecture*, and *technology architecture*. The focal point of TOGAF is at the enterprise architecture level, rather than the individual application architecture level.

RUP

Rational unified process (RUP) (Kruchten, 2003) attempted to overcome the shortcomings in the heavyweight methods by applying the unified modeling language (UML) in a use-case driven, object-oriented, and component-based approach. The concept of 4+1 views interprets the overall system structure from multiple perspectives. RUP is characterized by process orientation and is generally a waterfall approach. RUP barely addresses the phases of software maintenance and operations, and lacks a broad coverage on physical topology and development/testing tools. It mainly operates at an individual project level. RUP has recently been expanded to enterprise unified process (EUP), and was partially open sourced—OpenUP. RUP is now part of the IBM Rational Method Composer (RMC) product, enabling the process customization.

MDA

Model-driven architecture (MDA) (OMG, 2007) takes a different approach. MDA aims to separate business logic or application logic from the underlying platform technology.

It provides an open, vendor-neutral approach in UML to the challenge of business and technology change. The core of MDA is comprised of the platform-independent model (PIM) and platform-specific model (PSM), which provide greater portability and interoperability as well as enhanced productivity and maintenance. Platform-independent models of an application or integrated system's business functionality and behavior can be implemented via the MDA on practically any open or proprietary platforms. MDA is primarily for the software modeling part in the development lifecycle process.

MSF/MSA

Microsoft Solutions Framework (MSF) (Microsoft, 2007) is a comprehensive set of software engineering principles, processes, and proven practices that are specified to enable developers to achieve success in the software development lifecycle (SDLC). MSF is equipped with an adaptable guidance, based upon experiences and best practices from inside and outside of Microsoft, to enhance the chance of successful delivery of information technology solutions to clients by working fast, reducing the headcounts on the project team, mitigating risks, while not sacrificing high quality in the results.

The Microsoft Systems Architecture (MSA) is a program developed by Microsoft with the objectives of defining, designing, verifying, and documenting a set of IT infrastructure architectures, which comprise software, hardware machines, storage, connectivity, networking infrastructure, data center, and other tools/products.

C4ISR

The C4ISR Architecture Framework (DoD, 1997) provides comprehensive architectural guidance for the various commands, services, and agencies within the United States (U.S.) Department of Defense, to guarantee the interoperability and cost-effectiveness in the military systems. The framework consists of the guidance, policies, and product descriptions for defining and developing architecture descriptions that guarantee a common denominator for understanding, comparing, and integrating architectures.

FEA

The Federal Enterprise Architecture (FEA) framework (Federal Office, 2007) specifies direction and guidance to U.S. federal agencies for structuring enterprise architecture. The FEA is a group of reference models with a common taxonomy and ontology developed to describe IT resources, comprising the *performance reference model*, *business reference model*, *technical reference model*, *data reference model*, and *service component reference model*.

The *performance reference model* is a standardized construct to quantify the performance of key IT spendings and their impact on overall program performance. The *business reference model* specifies a function-centric structure to depict the agency-independent business operations in the public sector. The model provides an organized, hierarchical framework describing the daily business operations of the U.S. government via a functionally driven method. The *technical reference model* defines a component-based technical structure to classify the specifications, standards, guidelines, and solutions, which enable and support the delivery of service capabilities and components. The *data reference model* describes the data and information at an aggregate level, which support government programs and line-of-business operations. It empowers agencies to specify the types of interaction and exchanges that occur between citizens and the U.S. government. Finally, the *service component reference model* defines a business and performance-centered, functional structure that categorizes service components regarding how they support business and/or performance goals.

TEAF

The Treasury Enterprise Architecture Framework (TEAF) (Treasury Department, 2000) is to guide the planning and development of enterprise architectures in all bureaus and offices of the Treasury Department. It provides a framework for producing an EA, guidance for developing and using an EA, and guidance for managing EA activities. The framework subdivides an EA by views, perspectives, and work products.

PERA

The Purdue Enterprise Reference Architecture (PERA) (Purdue University, 2007) is aligned to computer integrated manufacturing. The PERA generic enterprise model consists of three basic elements: *product facilities*, *people/organization*, and *control and information systems*. The most basic way to structure the enterprise model in PERA is by "phase". Various diagrams are used in each phase of the enterprise to reflect the developing detail, ranging from initial definition to operations phase, to dissolution.

RM-ODP

The ISO Reference Model for Open Distributed Processing (RM-ODP) (Putman, 2001) is a coordinating framework for the standardization of open distributed processing in heterogeneous environments. It creates an architecture model that integrates the support of distribution, interworking and portability, using five "viewpoints"—*enterprise*, *informa-*

tion, *computational*, *engineering*, and *technology*—and eight "transparencies"—*access*, *location*, *relocation*, *migration*, *persistence*, *failure*, *replication*, and *transaction*. RM-ODP was adopted as ISO Standard 10746 in late 1990s, composed of four fundamental elements: *an object modeling approach*, *system specification*, *system infrastructure definition*, and *system conformance assessment framework*.

ATAM

In software engineering, Architecture Tradeoff Analysis Method (ATAM) (SEI, 2007) is a risk-mitigation process used early in the software development lifecycle. ATAM was developed by the Software Engineering Institute at the Carnegie Mellon University. Its purpose is to help choose a suitable architecture for a software system by discovering trade-offs and sensitivity points.

ATAM is most beneficial when used in the early stage of the software development lifecycle, in which the cost of changing architectures is minimal. The ATAM process consists of gathering stakeholders together to analyze business drivers and from these drivers extract quality attributes that are used to create scenarios. These scenarios are then used in conjunction with architectural approaches and architectural decisions to conduct an analysis of trade-offs, sensitivity points, and risks (or non-risks). This analysis can be converted to risk themes and their impacts whereupon the process can be repeated.

SAAM

Scenario-based Architecture Analysis Method (SAAM) (SEI, 2007a) is an evaluation method of software architecture. As one of the first documented software architecture analysis methods, it was developed to analyze a system for modifiability but is useful for testing any nonfunctional aspect. Architecture models are examined via scenarios in SAAM with regard to achieving quality attributes. SAAM can be used to assess existing systems, evaluate planned systems, or compare old and new systems. As a low-cost method, SAAM's results may not be precise. In spite of inaccurate results, the main value of SAAM is in the way that it focuses on investigation, stimulates discussions among the stakeholders, and helps build consensus among the parties involved.

IDEF

Integrated Definition Methods (IDEF) (IDEF, 2007) forms a structured approach to enterprise modeling and analysis. IDEF was produced by the integrated computer-aided manufacturing (ICAM) initiative of the U.S. Air Force. More specifically, the integrated information support system (IISS) project priorities 6201, 6202, and 6203 fathered IDEF. IISS was an effort to build an information processing environment running on heterogeneous physical computing systems. The purpose was to build "generic subsystems" that can be reused by a large number of collaborating enterprises, such as U.S. Defense contractors and the armed forces of alliance nations.

IDEF consists of 16 methods, covering a wide range of uses—*function*, *information*, *data*, *simulation*, *process description capture*, *object-oriented design*, *ontology description capture*, *design rationale capture*, *information system auditing*, *user interface*, *scenario-driven information system design*, *implementation architecture*, *information artifact*, *organization*, *schema mapping*, and *network design*.

MODAF

The UK Ministry of Defense Architectural Framework (MODAF) (UK, 2007) defines a standardized way of modeling an enterprise. The purpose of MODAF is to ensure a consistent approach to enterprise architecture development. It defines architectural views covering the strategic goals of the enterprise as well as the people, processes, and systems that deliver those goals. It also includes capability management (lines of development, doctrine, organization, training, material, leadership & education, personnel, and facilities) and programmatic aspects such as project dependencies. A MODAF model is organized into six viewpoints: *operational*, *system*, *strategic*, *technical*, *acquisition*, and *all viewpoints*.

COMPARISON MATRIX

Most of today's architecture design practices are still ad hoc, manual, and error-prone, which inevitably leads to chaotic outcomes and failures in the execution. According to surveys (Standish Group, 2007), a vast majority of information systems projects are routinely behind schedule, over budget, or canceled. A lack of a systematic approach to objectively assessing and validating the architecture design methods is indirectly attributed to the overwhelming failure.

As discussed in the preceding section, most of the existing methods and frameworks reveal the architectural aspects of a software application to some extent from a specific perspective. The necessity of a comprehensive comparison to evaluate various architecture frameworks becomes more and more evident, demanding a systematic disciplined head-to-head assessment. A highly structured comparison matrix is thus constructed to provide a side-by-side comparison and contrast. Table 1 shows a comprehensive chart of the comparative study.

Table 1. Comparison matrix of architecture methods and frameworks

Method / Framework	Description	Scope	Strengths	Weaknesses	Focus	Reference
Zachman Framework	Pioneer work in EA space. A 2-D matrix of 5W+1H.	Contextual, conceptual, logical, physical	Planning tool Broad acceptance Categorizing deliverables History in manufacturing	Limited holistic perspective Process-oriented Data-driven	Generic Process	www.zifa.com
E2AF	Extended Enterprise Architecture Framework	Business, information, information-system, technology infrastructure	History in enterprise framework Separation of concerns Focus on collaboration Holistic perspective Neutral/Open Communication tool	Limited acceptance Forward-engineering only	Generic Project	www.enterprise-architecture.info
TOGAF	The Open Group Architecture Framework	Architecture vision, business architecture, data architecture, application architecture, technology architecture	EA development methodology Open structure Holistic perspective	Heavyweight approach Forward-engineering only No coverage on portfolio Does not address tradeoff analysis	Generic Solution	www.togaf.org
RUP	Rational Unified Process	Applying the Unified Modeling Language (UML) in a use-case driven, object-oriented and component-based approach.	Knowledge base Proven methodology Key artifact templates Lifecycle management Traceability	Heavyweight Proprietary Somewhat waterfall Individual project level Process-oriented Does not cover maintenance	Generic Process	www.ibm.com/software/rational
MDA	Model-Driven Architecture	Computation-independent model, platform-independent model, platform-specific model	Agile method More levels of abstraction More abstract artifacts PIM and PSM Queries/ Views/ Transformations (QVT) transformation	Redundancy: duplicate work on different views Rampant round-trip problem for models Moving, not reducing, complexity Lack of semantics and automated tools	Generic Light-weight Model transforma-tion	www.oasis.org/mda
MSF/MSA	Microsoft Solutions Framework and System Architecture	Principles, models, disciplines, concepts, guidelines, infrastructure design	Key design aspects in IT operations environment Infrastructural services Department data center Enterprise data center Internet data center	Proprietary Based on Windows platform	Generic Infra-structure services	www.microsoft.com
C4ISR	Command, Control, Computers, Communications (C4), Intelligence, Surveillance, and Reconnaissance (ISR)	Policies, guidance, and product descriptions for defining and developing architecture descriptions that guarantee a common denominator for understanding, comparing, and integrating architectures.	History in defense Broad defense acceptance Neutral Process Planning tool	Limited holistic perspective Limited to defense domain	Domain-specific Process	www.aitcnet.org/dodaf

continued on following page

Table 1. continued

Method / Framework	Description	Scope	Strengths	Weaknesses	Focus	Reference
FEA	Federal Enterprise Architecture	A set of reference models: Performance, Business, Service Component, Data, and Technical	EA reference framework History in EA planning Holistic structure	Limited to U.S. government domain Too general stack Lack of automatic model transformation Not a roadmap Not a methodology	Domain-specific Frame-work	www.feapmo.gov/fea.asp
TEAF	Treasury Enterprise Architecture Framework	A framework for producing an Enterprise Architecture, guidance for developing and using an EA, and guidance for managing EA activities.	Multiple perspectives Multiple views Work products 2-D matrix	Limited to Treasury domain Outdated	Domain-specific Management	www.eaframeworks.com/TEAF
PERA	Purdue Enterprise Reference Architecture	Three basic elements: Product Facilities, People/ Organization, and Control and Information Systems.	Phased approach Diagrams Modularized structure	Out of date For computer integrated manufacturing	Domain-specific Control	www.pera.net
RM-ODP	Reference Model of Open Distributed Processing	A coordinating framework for the standardization of Open Distributed Processing in heterogeneous environments	Multiple viewpoints Various transparencies For heterogeneous environments	Outdated Does not address the issues in the new computing paradigm such as component-based and service-oriented architecture	Generic Views Transparency	www.rm-odp.net
ATAM	Architecture Tradeoff Analysis Method	Risk mitigation process to create an analysis of trade-offs and sensitivity points	Business drivers Quality attributes Architectural approaches and decisions	For software development lifecycle only	Generic Risk method	www.sei.cmu.edu/architecture
SAAM	Software Architecture Analysis Method	Architecture examination via scenarios with regard to achieving quality attributes	Scenario-based Quality attributes Activities and dependencies Build consensus	For software architecture only Imprecise results	Generic Scenario method	www.sei.cmu.edu/architecture
IDEF	Integrated Definition methods	A structured approach to enterprise modeling and analysis	16 methods IEEE standards 1320.1/2 Functional, Conceptual, Process flow, OO design, and Ontology	Root in Air Force Manufacturing-focused Outdated	Modeling IDEF0 ~ 4 commonly used	www.idef.com
MODAF	MOD Architecture Framework	Architectural views covering the strategic goals of the enterprise, and the people, processes and systems that deliver those goals.	Contiguous, coherent model Extended viewpoints	Data-driven approach to architecture Defense domain	Domain-specific Process	www.modaf.com

FUTURE TRENDS

A plethora of architecture methods and frameworks have emerged and evolved in the last couple of decades. Some of them have matured while some have merged or been transformed to other structures. The others have simply retired. However, most of the prominent architecture methods and frameworks have distinctive design goals with a focus on different domains of dissimilar scope and structures based on different principles and approaches. Consequently, it becomes difficult in the real world to directly apply these methods and frameworks in a particular domain. Generally speaking, there are five approaches to leveraging these existing methods and frameworks to the maximum extent: *direct use*, *adaption*, *best-of-breed*, *aggregation*, and *build*. The most straightforward way is via *direct use*, in which a method/framework is adopted without any modification. The general-purpose methods and frameworks are good candidates to be picked in this situation. The second approach is *adaption*, which customizes a selected method or framework for the needs in a specific segment. In this approach, typically a heavyweight method or framework in the same business sector is chosen as the baseline, and is subsequently streamlined and customized to adapt to an individual domain. The third way is *best-of-breed*, in which desired features from several methods and frameworks are extracted to form an optimal solution. The constituents in various methods and frameworks are selectively united to formulate a compound stack. The fourth approach is *aggregation*, which combines capabilities from various methods and frameworks, and further augments these features to establish a composite method. This usually results in a hybrid approach in the real-world application, benefitting from multi-disciplines, richer features, and better user friendliness as a result of aggregation. The other alternative is *build*, in which an organization creates its own framework from scratch for the unique requirements in the environment. This usually is the least cost-effective option. Nevertheless, a number of artifacts and fundamentals in the well-established methods and framework can still be directly or indirectly utilized in this approach, such as principles, techniques, patterns, and structures.

In general, there is no one-size-fits-all in this area. Some of the existing architecture methods and frameworks will continue to grow and mature, incorporating forthcoming innovations and adapting to the latest advance in the field. Convergence is expected to take place to unite the methods and frameworks with similar capabilities and contexts. Hybrid methods and frameworks demonstrate a stronger potential for widespread adoption, exploiting the promising benefits of both heavyweight and agile approaches. Fewer brand-new but more holistic methods and frameworks tend to emerge from the ground up to satisfy the new needs in the architecture paradigm. It can be foreseen that notations will be unified and tooling will be standardized, with more advanced round-trip engineering practices to fully automate the architecture design process.

CONCLUSION

This article provides an overview of the predominant architecture methods and frameworks used in enterprise architecture planning, development, management, and governance. The design principles of architecture methods are evaluated. The major methods and frameworks discussed in this paper include Zachman Framework, E2AF, TOGAF, RUP/EUP/Open-UP, MDA, MSF/MSA, C4ISR, FEA, TEAF, PERA, RM-ODP, ATAM, SAAM, IDEF, and MODAF. The key characteristics and capabilities are detailed in the articulation. A comparison matrix is constructed to provide a head-to-head comparison of the strengths and weaknesses of each individual model. The scope and selection guidelines are also included in the chart. Future trends and recommendations are elaborated in the context.

REFERENCES

DoD C4ISR Architecture Working Group. (1997). *C4ISR Architecture Framework*, Version 2.

Federal Office of Management and Budget (2007). *Federal Enterprise Architecture Framework*. Retrieved May 18, 2007, from http://www.whitehouse.gov/omb/egov/a-2-EAModelsNEW2.html

IDEF (2007). *Integrated Definition Methods*. Retrieved May 18, 2007, from http://www.idef.com

IEAD (Institute for Enterprise Architecture Developments) (2004). *Extended Enterprise Architecture Framework*.

Kruchten, P. (2003). *The rational unified process: An introduction* (3rd ed.). MA: Addison Wesley.

Microsoft. (2007). *Microsoft solutions framework*. Retrieved May 18, 2007, from http://www.microsoft.com/technet/solutionaccelerators/msf

OMG (Object Management Group). (2007). *Model driven architecture*. Retrieved May 18, 2007, from http://www.omg.org/mda

Purdue University. (2007). *The Purdue enterprise reference architecture*. Retrieved May 18, 2007, from http://pera.net

Putman, J. R. (2001). *Architecting with RM-ODP*. New Jersey: Prentice Hall PTR.

SEI. (2007). *Architecture tradeoff analysis method*. Retrieved May 18, 2007, from http://www.sei.cmu.edu/architecture/ata_method.html

SEI. (2007a). *Scenario-based architecture analysis method*. Retrieved May 18, 2007, from http://www.sei.cmu.edu/architecture/scenario_paper

The Open Group. (2007). *The open group architecture framework*. Retrieved May 18, 2007, from http://www.opengroup.org/architecture/togaf8/index8.htm

The Standish Group. (2007). *The Standish Group chaos report 2006*. Retrieved May 18, 2007, from http://www.standishgroup.com

Treasury Department CIO Council. (2000). *Treasury enterprise architecture framework*. Version 1.

UK. (2007). *Ministry of Defense architectural framework*. Retrieved May 18, 2007, from http://www.modaf.com

Zachman, J. A. (1987). A framework for information systems architecture. *IBM Systems Journal, 26*(3), 276-295.

KEY TERMS

ATAM: Architecture tradeoff analysis method, a risk-mitigation process used early in the software development lifecycle.

E2AF: Extended enterprise architecture framework, covering business, information, system and infrastructure in a 2-D matrix.

IDEF: Integrated definition methods, a structured approach to enterprise modeling and analysis, consisting of 16 methods.

MDA: Model-driven architecture, an agile approach. MDA aims to separate business logic or application logic from the underlying platform technology.

Microsoft Solution Framework: A comprehensive set of software engineering principles, processes, and proven practices that are specified to enable developers to achieve success in the software development lifecycle.

MODAF: Ministry of Defense architectural framework, a standardized way of modeling an enterprise, with six viewpoints ranging from people to processes and systems.

RM-ODP: Reference model for open distributed processing, a coordinating framework for the standardization of open distributed processing in heterogeneous environments, with five viewpoints and eight transparencies.

RUP: Rational unified process, a use-case driven, object-oriented and component-based approach.

SAAM: Scenario-based architecture analysis method, an evaluation method examining architectures via scenarios with regard to achieving quality attributes.

TOGAF: The Open Group architectural framework, a detailed framework with a set of supporting tools for developing an enterprise architecture, composed of architecture development method, enterprise architecture continuum, and TOGAF resource base.

Zachman Framework: A logical structure used to categorize and organize the descriptive representations of an enterprise IT environment, designed by John Zachman.

Architectures for Rich Internet Real-Time Games

Matthias Häsel
University of Duisburg-Essen, Germany

INTRODUCTION

Many researchers regard multiplayer online games as the future of the interactive entertainment industry (Brun, Safaei, & Boustead, 2006; El Rhalibi & Merabti, 2005; Sharp & Rowe, 2006). In particular, due to advances in game design and the availability of broadband Internet access to the end-user, multiplayer online games with real-time interaction have come into wide use (Aggarwal, Banavar, Mukherjee, & Rangarajan, 2005; Claypool & Claypool, 2006; Yasui, Yutaka, & Ikedo, 2005). The majority of these games are made up by classic software titles that need to be installed on the players' machines (El Rhalibi & Merabti, 2005). Browser-based multiplayer games, on the contrary, can be run instantly from a Web site, but have, due to technical limitations, long been round-based, strategy-focused games. However, with the ongoing evolvement of Rich Internet Application (RIA) technology (Allaire, 2002) such as Adobe Flash and Java, browser-based online game development has reached a point where also *real-time* games can be produced and distributed to a large audience quickly and easily. Browser-based games can be utilized in conjunction with e-business offers in a very simple way and hold a number of exciting possibilities for new online business models, new markets, and new growth (Kollmann & Häsel, 2006; Sharp & Rowe, 2006). However, as the browser is a very different operating environment and interactive experience from that of classical game software, browser-based multiplayer real-time games involve gaming architectures that are distinct from their classical counterparts. A major challenge when designing and implementing such architectures is that multiplayer online games are highly vulnerable to propagation delays resulting from redundant communication, bottlenecks, single points of failure and poor reactivity to changing network conditions (Ramakrishna, Robinson, Eustice, & Reiher, 2006). As latency from input of information to its output determines gameplay and fairness (Brun et al., 2006), the game architecture has to be designed in a way that it mitigates latency effects and meets the expectations of the players (Claypool & Claypool, 2006). Elaborating on the example of an online tabletop soccer game with two remote players, this article discusses two architectural models that can be applied to implement browser-based multiplayer real-time games using RIA technology.

BACKGROUND

Online games that give the player the ability to compete against other players over a network emerged strongly in the middle of the last decade. Traditionally, multiplayer online games have been implemented using client-server architectures (GauthierDickey, Zappala, Lo, & Marr, 2004). Thereby, a copy of the software is installed on each player's machine, which connects directly to a central authoritative server designed to handle game logic (Claypool & Claypool, 2006; El Rhalibi & Merabti, 2005; Guo, Mukherjee, Kangarajan, & Paul, 2003). The server deals out information individually to each client as it is requested and keeps all the players up to date with the current state of the game (El Rhalibi & Merabti, 2005). Whenever a client performs an action (such as firing a gun or kicking a ball), the data is sent to the server, which calculates the effects of that action and sends the updated game state to the clients (El Rhalibi & Merabti, 2005). Client-server architectures have the advantage that a *single* decision point orders the clients' actions, resolves conflicts between these actions and holds the global game state (GauthierDickey et al., 2004). Unfortunately, they also have several disadvantages. The most obvious one is that client-server architectures introduce delay because messages between players are always forwarded through the server (GauthierDickey et al., 2004). This adds additional latency over the minimum cost of sending commands directly to other clients (Cronin, Kurc, Filstrup, & Jamin, 2004). Moreover, traffic and CPU load at the server increases with the number of players, creating localized congestion and limiting the architecture by the computational power of the server (GauthierDickey et al., 2004).

To address the problems associated with client-server architectures, many authors and game designers have developed fully distributed peer-to-peer architectures for multiplayer games (El Rhalibi & Merabti, 2005; GauthierDickey et al., 2004). In a peer-to-peer architecture, players send their actions to each other and react on the received action (Guo et al., 2003). Each peer acts as a decision point that has exactly the same responsibilities as every other peer (El Rhalibi & Merabti, 2005). Peer-to-peer architectures have a lot of advantages. Firstly, there is neither a single point of failure nor an expensive server infrastructure. Moreover, the amount of bandwidth required is reduced dramatically as there is a

direct communication between two peers. This also reduces latency on the network, as the bottleneck caused by a server is eliminated (El Rhalibi & Merabti, 2005). However, unlike games using a client-server architecture, where there is a single authoritative copy of the game state kept at a central server, peer-to-peer architectures require an up-to-date copy of the entire game state to be kept at each peer. Consequently, these architectures require some form of distributed agreement protocols between the peers that prevents the peers' game states from diverging over time and becoming inconsistent (Cronin et al., 2004; Guo et al., 2003).

Although most online games have low bit-rate requirements sending frequent but small packets typically well within the capacity of broadband connections (Brun et al., 2006), deploying these applications over a large-scale infrastructure presents a significant technological challenge. In both client-server and peer-to-peer architectures, an increase in the geographical distances among participating clients or servers results in an unavoidable end-to-end delay that may render the game "unresponsive and sluggish even when abundant processing and network resources are available" (Brun et al., 2006, p. 46). Moreover, differences in game responsiveness to user input may give some players an unfair advantage (Brun et al., 2006; Aggarwal et al., 2005). If the latency between a client and the server (or between two peers) is large enough, the responsiveness of the game to a player's action decreases, and the player's performance is likely to degrade (Claypool & Claypool, 2006). A game can be regarded as playable if the players find its performance acceptable in terms of the perceptual effect of its inevitable inconsistencies, whereas fairness is concerned with relative playability among all players (Brun et al., 2006).

ARCHITECTURE DESIGN AND IMPLEMENTATION

After reviewing existing literature on architectural concepts and issues of multiplayer online games, the remainder of this article will examine to what extent these concepts and issues hold for *browser-based* multiplayer real-time games. As a matter of simplification, the focus will be on games featuring concurrent, pairwise real-time interactions between two remote players, such as it the case for the online tabletop soccer game that is depicted in Figure 1.

For the player-side RIA, Adobe Flash is assumed to be the best-suited technology since Flash, due to its common runtime environment across operating systems, browsers and chip architectures, enables easy deployment on multiple platforms and devices (Allaire, 2002). In contrast to Java as a possible alternative, Flash has a much higher diffusion rate and can be updated by the player in a very easy way. Moreover, Flash is very efficient in rendering

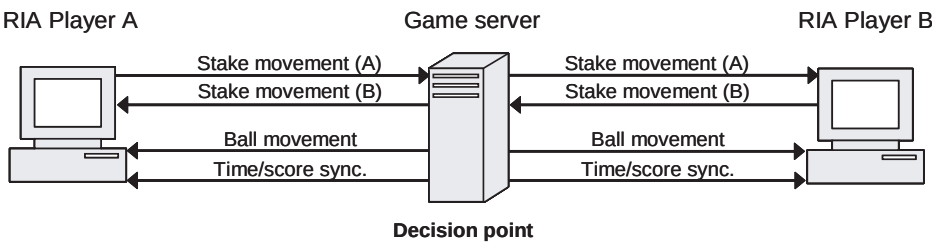
Figure 1. An online tabletop soccer game with two remote players

vector graphics and thus provides an optimal runtime for fast-paced browser-based games. With respect to real-time and multiplayer functionality, Flash applications are able to integrate socket-based, two-way communications and to keep live connections to servers for building applications with persistent connectivity (Allaire, 2002). However, a major limitation of Flash is the fact that these live connections are inevitably based on TCP. At first sight, this seems to be fairly advantageous, because TCP enables an error-free, ordered data transfer and a retransmission of lost packets, rendering built-in mechanisms to deal with message loss unnecessary (Pantel & Wolf, 2002). In the context of real-time games, however, TCP connections are unfortunate because a client cannot receive the packets coming after a lost packet until the retransmitted copy of the lost packet is received (Pantel & Wolf, 2002). In fact, for real-time clients it would be more useful to get most of the data in a timely fashion (as it would be the case for UDP) than it is to get all of the data in succession, as the resulting effects of packet loss can be mitigated by frequent game-state updates and techniques such as dead-reckoning vectors (Guo et al., 2003; Aggarwal et al., 2005) and client-side prediction (Brun et al., 2006). Though packet loss and latency resulting from applied packet loss mitigation techniques is zero, latency in connection with TCP results from the data transmission itself. For instance, as the difference in network latency between two tabletop soccer players becomes larger, the position of the ball displayed at one player diverges more largely from that at the other, bringing inconsistency among the two players (Yasui et al., 2005). In particular, there are two kinds of inconsistencies that may occur (Brun et al., 2006): First, the increased response time, that is, the delay between the time of a player's input and the rendering of the respective results on a player's screen, may frustrate players and make the game unplayable. Second, a presentation inconsistency may occur due to the fact that the game-state update reaching a RIA is already outdated to some degree because the real game state may have varied while the update packet was on its way. This would mean, for instance, that the player's perception of the current ball position is slightly inconsistent with the real ball position at the server, respectively, the opponent's RIA. Such inconsistencies may prevent a player from responding effectively or appropriately, and thus can lead to player frustration, especially in highly competitive game environments such as an online soccer league (Guo et al., 2003).

Latency, however, is not the only limitation with respect to RIAs, including both Flash applications and Java applets. Due to the sandbox principle of these technologies, a direct data exchange (i.e., peer-to-peer communication) between two client-side game applications is not possible. Consequently, browser-based multiplayer games must inevitably be based on physical client-server architectures. With respect to their logical nature, however, there are two alternatives: client-server and hybrid peer-to-peer. The main difference between these two approaches lies within the distribution of the game's decision points. In the former approach, the current game state is calculated by the server, whereas in the latter approach, the calculation of the game states is performed by the clients, while the game server is simply relaying peer-to-peer communication. In the following, these two approaches will be discussed in more detail, including the advantages and drawbacks resulting from each alternative.

In its essence, a browser-based multiplayer real-time game based on a client-server architecture (Figure 2) is very similar to classical multiplayer games where the server acts as a central decision point that calculates and simulates the game states based on the players' actions. Moreover, the server enables players to join the game and assigns opponents to each other. The Flash clients, in contrast, simply render and present the game states to the player and send the players' inputs to the server. In the context of an online tabletop soccer game, each player sends update vectors representing the stakes he controls, including each stake's current position and rotation. The server uses these vectors to calculate an update vector representing the current ball position and trajectory, and sends it to the clients. Clients themselves do not (and cannot) communicate with each other, and neither do they play any active role in deciding the ordering of actions in the game. Moreover, the server sends synchronization updates relating to the current score and time. This is necessary because the timing of a Flash

Figure 2. Client-server architecture for browser-based multiplayer real-time games



client is highly inaccurate, as it depends on the application's frame rate, which again depends on the local CPU speed and type of browser.

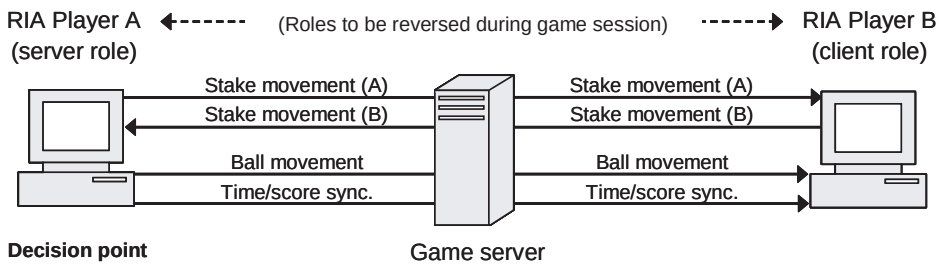
A major drawback of the client-server model is that the game server becomes the main bottleneck for both network traffic and CPU capacity. Moreover, as the network delay from the server to different clients is different, players may receive the same state update at different times. Vice versa, players' action messages can also take different times to reach the game server, and therefore unfairness in processing player action messages can be created at the game server. A player further away from the game server or connected to the server through congested or slower links will experience longer message delay. Because of this, even fast reacting players may not be given credit for their actions, leading to an unfair advantage for players with small message delays (Guo et al., 2003). For instance, a player with a large delay would always see the ball later than the other player and, therefore, this player's action on the ball would be delayed even if the player reacted instantaneously after the ball position was rendered (Aggarwal et al., 2005).

In contrast to a fully distributed peer-to-peer architecture, a hybrid peer-to-peer architecture for browser-based games still needs to be based on a central game server because a direct communication is not possible using current RIA technologies. Consequently, communication between two peers must be relayed via a *nonauthoritative* game server in this model. Besides that, the peers still rely on the server to join the game and to discover their opponents. While the client-server model guarantees event ordering because messages from all the players are only delivered at a central decision point (Guo et al., 2003), peer-to-peer games introduce discrepancies among decision points that can cause some decision points to evaluate events out of order, possibly violating causality and prompting incompatible decisions (Brun et al., 2006). Event consistency in peer-to-peer architectures has been well studied in the online gaming literature devised to guarantee a uniform view of the game state using causality control techniques such as bucket synchronization (Diot & Gautier, 1999; Pantel & Wolf, 2002), trailing state synchronization (Cronin et al., 2004), delta-causality control (Yasui et al., 2005), and time warp synchronization (Mauve, Vogel, Hilt, & Effelsberg, 2004). However, these techniques have been mainly designed for real-time games with high consistency requirements, but *low* latency requirements (Cronin et al., 2004), and thus cannot be used in conjunction with highly fast-paced games such as the tabletop soccer. Moreover, implementing clock-based synchronization in Flash renders extremely difficult, as timing depends on the respective client machines and browsers. These limitations make a topology with multiple decision points unfeasible.

One solution for guaranteeing event consistency despite this limitation is to transfer causality control to one of the RIAs during the game session (Figure 3). This is possible because the game session of browser-based games is relatively short-lived, that is, there is no global game state that needs to be preserved for several hours or days. However, with respect to playability and fairness in fast-paced games generating high bitrates, there are two main drawbacks of this approach: First, the player acting as the decision point is per se in advantage, as for this player, the propagation delay of the game state is zero, while, in comparison to the client-server architecture, the other player's response time has roughly *doubled*. As a fair game gives all users the same level of handicap (Brun et al., 2006), a possibility to re-establish fairness is to switch the decision point between the RIAs during the game session. For the tabletop soccer game, for instance, this could be done after each goal or at halftime. Second, the fact that one of the RIAs takes over server functionality increases the amount of outgoing traffic for that client. As the *upload* speed of high bandwidth lines such as ADSL is relatively low, this introduces an additional bottleneck at the respective decision point.

Whether a browser-based multiplayer real-time gaming architecture is implemented using the client-server or the hybrid peer-to-peer model fairly depends on the type of game. While slower games may be implemented using a hybrid peer-to-peer architecture, a client-server architecture seems to be the better choice with respect to playability and fairness in highly fast-paced browser-based games, as response times are shorter and both clients are treated equally. However, this solution may incur significant cost in the form of more CPU

Figure 3. Hybrid peer-to-peer architecture for browser-based multiplayer real-time games



capacity and server software. In both architectural models, one possibility for implementing the server is using a high-level language such as Java, using its TCP socket API for exchanging real-time data with the Flash clients. Another possibility is to use server-side RIA technology such as Adobe Flash Media Server (Lesser, Guilizzoni, Reinhardt, Lott, & Watkins, 2005), which provides a framework with a number of ready-made server-side components for building Flash-based client-server applications with real-time capabilities. Yet, with respect to games that are generating high bitrates, technologies such as these are cost-intensive because license fees typically depend on bandwidth requirements.

FUTURE TRENDS

From a practical point of view, it will be interesting to see which architectural model and what kinds of technology will prevail in the realm of browser-based real-time multiplayer games. Although current RIA technology still entails some limitations with respect to real-time communication, this is likely to change in the future. From a theoretical point of view, likewise, researchers will have to explore what combination of gaming architecture and technology is best-suited for what kind of browser-based game. In this regard, it also has to be explored to what extent the two architectural models presented in this article can be applied to implement browser-based real-time games featuring *more* than two opposing players in a single game session.

Moreover, with respect to online gaming portals with several thousand concurrent players, the architectural concepts presented in this article have to be enhanced. A single server cannot be expected to suffice to deal with thousands of players synchronously. This suggests implementing the authoritative game server as cluster of machines with dedicated responsibilities. Also, such approaches need to consider intelligent methods for choosing the best-suited server for groups of geographically dispersed players. Future research has to explore how such concepts can be realized in connection with RIA technology.

CONCLUSION

With the advent of RIA technology, browser-based multiplayer gaming has finally reached a point that deserves the attention of serious game designers, programmers and online gaming researchers. The vast possibilities connected with technologies such as Adobe Flash and Java include producing many engaging and sought-after types of games. When implementing real-time multiplayer games, the main challenge lies within guaranteeing playability and fairness despite synchronization issues that are connected with the current best-effort Internet. These issues need to be solved by designing appropriate game architectures and trading off inconsistencies between players depending on their perceptual impact. This article introduced two architectural models that can be used to implement browser-based real-time multiplayer games. Both models rely on a central server, as a direct peer-to-peer communication cannot be realized with current RIA technology. However, this server can either form a global decision point that is calculating the current game state (client-server architecture) or simply relay game data between the clients, while one of the clients takes over decision point functionality (hybrid peer-to-peer architecture). While the former approach makes the game server a major bottleneck with respect to CPU capacity, the latter approach has major drawbacks resulting from the two different roles that clients can play. The choice of architecture certainly depends on the type of game to be implemented. Future research has to further elaborate on the architectures and technologies that are best-suited for browser-based real-time applications.

REFERENCES

Aggarwal, S., Banavar, H., Mukherjee, S., & Rangarajan, S. (2005). Fairness in dead-reckoning based distributed multi-player games. In *NetGames '05: Proceedings of the 4th ACM SIGCOMM Workshop on Network and System Support for Games,* (pp. 1-10).

Allaire, J. (2002). *Macromedia Flash MX–a next-generation rich client*. Retrieved December 13, 2007, from http://www.adobe.com/devnet/flash/whitepapers/richclient.pdf

Brun, J., Safaei, F., & Boustead, P. (2006). Managing latency and fairness in networked games. *Communications of the ACM, 49*(11), 46-51.

Claypool, M., & Claypool, K. (2006). Latency and player actions in online games. *Communications of the ACM, 49*(11), 40-45.

Cronin, E., Kurc, A.R., Filstrup, B., & Jamin, S. (2004). An efficient synchronization mechanism for mirrored game architectures. *Multimedia Tools and Applications, 23*(1), 7-30.

Diot, C., & Gautier, L. (1999). A distributed architecture for multiplayer interactive application on the Internet. *IEEE Network Magazine, 13*(4), 6-15.

El Rhalibi, A., & Merabti, M. (2005). Agents-based modeling for a peer-to-peer MMOG architecture. *Computers in Entertainment, 3*(2), Article 3B.

GauthierDickey, C., Zappala, D., Lo, V., & Marr, J. (2004). Low latency and cheat-proof event ordering for peer-to-peer games. In *Proceedings of the ACM NOSSDAV '04,* (pp. 134-139).

Guo, K., Mukherjee, S., Rangarajan, S., & Paul, S. (2003). A fair message exchange framework for distributed multi-player games. In *NetGames '03: Proceedings of the 2nd ACM SIGCOMM Workshop on Network and System Support for Games,* (pp. 29-41).

Kollmann, T., & Häsel, M. (2006). *Cross-channel cooperation– the bundling of online and offline business models*. Wiesbaden: DUV.

Lesser, B., Guilizzoni, G., Reinhardt, R., Lott, J., & Watkins, J. (2005). *Programming flash communication server.* Cambridge, MA: O'Reilly.

Mauve, M., Vogel, J., Hilt, V., & Effelsberg, W. (2004). Local-lag and time-warp: Providing consistency for replicated continuous applications. *IEEE Transactions on Multimedia, 6*(1), 47-57.

Pantel, L., & Wolf, L.C. (2002). On the impact of delay on real-time multiplayer games. In *Proceedings of the ACM NOSSDAV '02,* (pp. 23-29).

Ramakrishna, V., Robinson, M., Eustice, K., & Reiher, P. (2006). An active self-optimizing multiplayer gaming architecture. *Cluster Computing, 9*(2), 201-205.

Sharp, C.E., & Rowe, M. (2006). Online games and e-business: Architecture for integrating business models and services into online games. *IBM Systems Journal, 45*(1), 161-179.

Yasui, T., Yutaka, I., & Ikedo, T. (2005). Influences of network latency and packet loss on consistency in networked racing games. In *NetGames '05: Proceedings of the 4th ACM SIGCOMM Workshop on Network and System Support for Games,* (pp. 1-8).

KEY TERMS

Browser-Based Game: An online game that is accessed with a Web browser over the World Wide Web. Unlike classical game software, browser-based games can be played instantly and do not require a prior installation on the player's machine. Browser-based games may be implemented using standard hypertext mark-up or Rich Internet Application technologies.

Client-Server: A network architecture that separates several clients from a server. A client is an application that sends requests to the server and is often responsible for rendering the user interface. The server passively waits for the clients' requests, processes them and sends appropriate responses to the clients.

Latency: A time delay between the moment something is initiated, and the moment one of its effects begins. With respect to packet-switched networks such as the Internet, latency (also referred to as *lag*) refers to the time taken for a packet of data from the sending application to the receiving application.

Multi-Player Game: A game that is played by multiple people at the same time. In contrast to single-player games that feature artificial opponents, players in multiplayer games compete against each other or team up to achieve a common goal. Modern multiplayer games connect geographically dispersed players by using Internet technologies and allow them to play together.

Peer-to-Peer: A network architecture that does not have the notion of clients or servers, but only equal peer nodes that simultaneously function as both clients and servers to each other. If a peer-to-peer network has a central server that keeps information on peers and responds to requests for that information, this network is often referred to as *hybrid* peer-to-peer.

Real-Time Game: A multiplayer game featuring real-time interaction between the players, that is, when a player performs an action, other players are made aware of the consequences of that action within an operational deadline. As a discrepancy in the perceptions of the players may lead to undesirable outcomes, such games entail increased constraints on responsiveness and consistency.

Rich Internet Application: A browser-based application that has the features and functionality of a traditional desktop application. Rich Internet Applications (RIAs) introduce an intermediate layer of code between the user and the server that acts as an extension of the browser. They are typically run in a secure "sandbox" environment and can be implemented using technologies such as Flash, Java or Ajax.

Archival Issues Related to Digital Creations

Mark Kieler
Carnegie Mellon University, USA

Michael J. West
Carnegie Mellon University, USA

INTRODUCTION

The authors define "intellectual creations" as human expressions embodied in text, music, or other forms of art. Increasingly, we encode these creations in digital formats that have extremely short life cycles. Eventually, backward compatibility is lost. Thus, after very little time, a digital encoding format becomes obsolete, and intellectual works encoded in the format may become irretrievable. In contrast, the cultural worth of an intellectual creation may not be realized for generations. Additionally, future generations must access artifacts, including intellectual creations, to understand a culture in historical context. We contend that technology – intensive storage and manipulation of data may result in an inability to gain this access. Technology creators have some responsibility to facilitate future retrieval through careful documentation, and by selective maintenance of hardware that may be required to access archival media.

BACKGROUND

Cultural artifacts nearly always outlive the technologies that made them possible, which is particularly obvious with digital technology. Imagine the discovery hundreds of years from now of a floppy diskette containing a document written in WordStar®. Once a code has been lost, have all texts written in that code been lost as well?

At some point, supporting a technology is no longer economically viable if it is abandoned by enough people. It ceases to exist except in archival form. Currently, home videocassettes are being replaced by the DVD (Digital Video Disc), which subsequently will also be superceded.

Linguists stress the organic nature of language, which it is in constant evolution. Grammars are codifications constructed of "snapshot images" of spoken and written language usage. Despite the fact that both are considered examples of English, speakers of the vernacular of Shakespeare and the vernacular of rap music would find each other incomprehensible. With technology, as with language and culture, timing is everything, and context plays an important role in shaping how we interact with each other and with technology.

HOW DO WE KNOW ABOUT PAST CULTURES?

In engineering, Shannon (1948) described elements of a communications system: An information source generate data. A transmitter encodes information to travel over a channel through distance and time. At the other end, a receiver decodes the signal for the destination, which is the person or thing for which the message is intended. Since we are worried about humans, we consider them to be the source and destination. While parts of the transmission system have changed greatly over time, one could certainly argue that technology has caused the other portions of the system to change more rapidly than the human elements. Linguists, beginning with Ferdinand de Saussure, established semiotics as a science of signs, which likewise focuses on the sender and the receiver of a message. Semiotics also posited the arbitrary nature of the sign and looked at the ways human languages encode and transmit messages. (Saussure, 1974).

Physical artifacts that survive from the distant past reveal much about a culture, depending on their purpose and the quality of materials from which they are made. Our record from ancient civilizations is far from perfect, but archaeologists can construct some details about them from these clues.

Significant events have often been recorded in the living embodiment of a storyteller, and oral traditions still form an important part of many cultures. The historical record has often been related by language, as well as by performance (e.g., a ritual dance). Some of these oral histories survived long enough to be recorded in other more permanent media. However, not only have many oral traditions died with the last generation of storyteller, others have assumed inaccuracies and exaggerations as a result of being passed serially through generations.

As languages evolved and became standardized, it became possible to encode events in written form. Because writing has traditionally been the province of the learned few, written documents were recorded on long-lived media, and special care was accorded to their storage. Fortunately, many ancient documents have survived, albeit, with significant degradation. Given the constant change in a living language, when a culture dies, the language often dies with it. Language experts

attempt to reconstruct meaning by looking for patterns that may establish types of words and contexts and similarities to more modern languages.

In addition to printed text, human expression is also accomplished through artistic means such as music, painting, and dance. While we only have been able to preserve musical and dance performances for a relatively recent portion of human history, for centuries we have had a written procedure for recreating music, using notes, measures, and time definitions. The methods for recording dance instructions are much less standardized and rely on interpretations and tradition.

Art works degrade over time, depending on the types of inks, paints, or dyes used, the media on which they are deposited, and overall interaction with the environment. Even sculpture is subject to environmental degradation, such as damage from acid rain.

THE INTRODUCTION OF TECHNOLOGY

The invention of the printing press made wide distribution of printed information possible, and wood pulp-based paper made it affordable for the general public. However, unlike expensive fiber-based paper, pulp-based paper has usually been manufactured through an acid-based process, and residual acid in the paper eventually destroys it. Thus, paradoxically, fiber-based books from the 19th century are often more legible than their 20th century wood pulp counterparts.

Text

Text was the first means of expression to be converted into electrical form. In fact, text went "direct to digital." Morse code is a duration-encoded digital signal, unrecognizable to anyone who does not understand it. Initially, storage was primitive, used mainly to "buffer" the information until a human could reconvert it to text. Thus, long-term storage and archiving of Morse code traffic was not an issue.

The first modern bit encoding of text occurred in 1874. Emile Baudot, a French telegraph engineer, devised a 5-bit code for each letter of the alphabet. Unlike Morse code, each symbol had a fixed length representation, dependent only on the presence or absence of electrical current. The Baudot code was durable, used by news service teletypewriters throughout the 1970s (Freed, 1995). Crude paper tape punches were often utilized for storage. The digital code could be read, albeit slowly, merely by holding the tape up to a light.

ASCII (American Standard Code for Information Interchange) uses 7 bits. The added bits allowed upper and lower case letters, as well as numbers, punctuation, and other special characters. It endures as the "plain text" standard.

The rise of WYSIWYG (what you see is what you get) computer interfaces, and the availability of sophisticated word processing programs, made it possible to digitally encode additional expressions to text. Different art styles of text (fonts) could be used, and these could embody visual variations such as italics, bold, superscripts and subscripts, and underlines. Word processing evolved to encode these variations. This was accomplished by adding bits to the original text data, or by software commands that controlled a section of ASCII text. These techniques represent a deviation from international standards into conventions of the word processing software. As the level of sophistication increases, we become increasingly unable to understand an encoded section of text without the software used to create it. In fact, the actual text becomes only a tiny portion of the code. If the page is graphically encoded, as occurs in programs such as Adobe Acrobat®, then plain text representations are lost all together (Kieler et al., 2004).

Audio and Visual Technology

Storage of audio and visual data began in the 19th century and progressed in sophistication throughout most of the 20th century. While not strictly correct in each instance, the general paths of progress could be viewed as follows:

- Visual Still and Audio-Only, to Visual Motion, to Visual Motion with Audio
- Mechanical or Chemical-Based Recording, and Storage to Electronic Recording and Storage
- Mechanical Reproduction to Electronic Reproduction
- Analog Encoding to Digital Encoding

The most significant impediments to easy retrieval of an intellectual creation involve electronic encoding and storage, and digital conversion.

Electronic encoding of visual information marks the historical point where the original signal cannot be recovered merely by converting stored data directly into light or motion. The visual image is stored as a series of lines, and voltage references signal new lines or new collections of lines (*frames*). Thus, one must be able to properly decode the various electronic levels and pulses unique to the method, and understand how the lines are ordered and encoded in order to correctly reconstruct the scanned image. Incompatible formats include PAL (Phase Alternating Line), NTSC (National Television Standards Committee) and SECAM (Systéme Electronique Couleur Avec Mémorie) (Abbott, 1994).

With the advent of magnetic storage, one could no longer directly "see" the information on the storage medium without a great deal of technological aid. As an example, by the late 1920s, light modulation of audio onto motion picture film was possible (Hochheiser, 1992), and both audio and video information are clearly visible on the film. In contrast, the

advent of practical magnetic video recording in 1956 (Shima, 1984) resulted in a signal which can only be observed using microscope techniques (Chinn, 1957).

Finally, digital data types in streams of 1s and 0s are indistinguishable without the reverse algorithm to retrieve the original information. The significance is that in addition to the physical details of how the data was encoded, we also need mathematical processing information on how the data was changed after its conversion. Data protection techniques, including public and private key cryptography, hashing algorithms, and other methods (Wayner, 1997), added on top of the conversion technique, have deliberate obscuration of data as their intent. Thus, the data is not readable, even when the encoding techniques are known, unless the proper decrypting information is available.

Lossy Encoding

An additional dilemma is that digital conversions result in massive amounts of data. Compact disc audio, for instance, necessitates a bit rate of 4.32 million bits per second (Carrasso et al., 1982). Efficient storage and transmission virtually require data reduction. We must actually throw away some of the original information (hence the term *lossy*) to make any real progress in storage efficiency. We discard information based on redundancies in the retained data, or upon deeming it insignificant through models of human perceptual limitations. For instance, Moving Picture Experts Group Layer 3 audio (MPEG 3 or MP3) achieves a better than 10 to 1 savings (Jack, 1996). However, the encoded audio bit stream now represents a set of mathematical manipulations on top of the interleaving, control, data, and error correction algorithms of the storage format. It has become impossible to reconstruct the data without complete knowledge of each of these transformations.

IMPACT ON LONG-TERM DATA AVAILABILITY

If intellectual creations are stored in formats that require an exact knowledge of the encoding transforms, the issue for society is whether any of these techniques will be available, or even remembered, in the distant future.

In the short term, there is no problem, as new formats usually incorporate backward compatibility and conversion. Eventually, however, the overhead needed to retain compatibility hampers the effectiveness of a system. As demand collapses, the resources that support an obsolete format must be utilized for other purposes.

The digital revolution has compressed the time scale. Tremendous gains in hardware and software capability have resulted in unprecedented rates of obsolescence. For example, in both popular personal computing formats, operating systems have migrated to 32-bit code, abandoning the previous 16-bit systems. In doing so, they have shed much backward compatibility with the previous systems (White, 2002.) This transition has occurred in a mere 20 years. Thus, the ability to retrieve data created with a 16-bit program may well be lost once the last 16-bit compatible system is "upgraded."

New generations of artists and creative personalities use electronic and especially digital formats as freely as their predecessors used physical media. While most popular culture intellectual creations have little lasting value, a few will be significant enough to be studied and preserved far into the future. The problem is that human society does not always immediately recognize an intellectual work's significance, and it may fall into obscurity for decades or even centuries. A cursory encyclopedia search (Compton's, 1994) reveals examples from many fields. For example, painter Sandro Botticelli's works remained obscure from the late 1400s until the 19th century.

While not individually significant, ordinary popular culture creations collectively can reveal a great deal about a period in society. However objective evaluation often requires the passage of at least several decades. Even then, critiques and judgments may go through many modifications and revisions before any consensus is reached.

MINDSETS: CREATORS OF TECHNOLOGY VERSUS CREATORS OF IDEAS

Technology creators derive value from innovation. Profits and employment fall once a particular technology generation "matures" and a "shakeout" era ensues, where only the most efficient producers prevail. Software and digital data can be infinitely replicated with no degradation. Thus, barring legal impediments, there is virtually no value in creating additional copies. Even with legal protections against duplication, others may create software that operates differently but accomplishes the same goal. These characteristics shape the mindset of technology creators into a "cult of the new." Obsolescence is the fuel for continued employment and profitability.

With intellectual creations, there is also a progress-oriented motivation. Artists do not merely rest on past achievements. However, since these creations embody a combination of knowledge, ideas, and interpretation, their value is not necessarily superceded by later works. While only a few will have lasting value to society, there is no way to know when this value will finally be appreciated. The finest intellectual works must prove their value repeatedly through decades and centuries of cultural change.

The issue is that these works are increasingly created through a technological process that need only prove itself until the next "better" innovation comes along. The potentially tremendous but unpredictable future value of key intellectual works may never be realized if they are embodied in a technical process that has zero value five years later. While future societal worth is of little present economic value to the creation of intellectual work, the encoding technology creator derives nothing beyond the initial sale of the software.

Certainly, the same could be said of those who produce physical creative media such as paint or canvas. However, the medium usually survives regardless of the intentions of its original manufacturer. Deterioration is directly observable, and can often be repaired to foster preservation. With modern technology, however, the "medium" is a combination of the physical storage object and the encoding software/hardware systems. Even if the storage object can last centuries, the ability to retrieve it is governed by entities that have, if anything, a *dis*incentive to ensure continued long-term accessibility. This leaves society with no way to rediscover and realize the benefits of important intellectual creations in the distant future.

FUTURE TRENDS

While it is unfair to lay all responsibility for long-term accessibility on technology creators, it is clear that the nature of digital innovation fosters the problem. Among engineering ethicists, there is general agreement that those who create technology have a unique insight into potential side effects. Harris, Pritchard, and Rabins (1992) discuss the need for "preventive ethics" in engineering. Shinzinger and Martin (2000) describe engineering as social experimentation. This implies a responsibility for technology creators to envision possible social consequences and, at the very least, take steps to inform others of their existence and nature. This is especially true where the complexity of the technology is transparent to the end-user.

Thus, while digital technology creators may not solely be responsible for ensuring future accessibility of works expressed through their technology, one could argue that they have a responsibility to enter into dialogues about preserving accessibility, and to design their encoding technology to foster preservation.

Currently, preservation awareness is coming, not from the technology developers, but rather from the software archivists. Pennavaria (2003) details a number of concerns and efforts of groups such as the American Library Association and the American Film Institute. Although technologists are involved in these efforts, they tend to deal only with existing analog and digital formats. Also, they tend to concentrate largely on the deterioration mechanisms of the physical media, as opposed to preserving the availability of playback hardware and decoding software.

Some hardware, such as silicon chips, could last for centuries. Conversely, other pieces of machinery clearly have a limited shelf life due to physical deterioration and wear. It is the "soft" items, such as encoding and decoding software and the underlying algorithms and manipulations, that have the greatest chance to be lost, owing to their short lifetimes. Only the creators of these artifacts can take steps to ensure their survival. This implies a responsibility that the technologists cannot ignore. The broader culture must also realize the problem and work with the technologists to foster preservation. Otherwise, much of our current culture will disappear in less than a century.

CONCLUSION

The attractiveness and availability of digital processing has resulted in its use by creators of intellectual content. The rapid turnover of digital applications may make access to these creations impossible within a very short period, compared to the time it takes for society to determine their ultimate worth. Technology creators must work with society leaders to ensure that intellectual creations are not lost forever.

REFERENCES

Abbott, P. (1994, November-December). Video formats and resolution. *Nuclear Plant Journal*, 39-46.

Carasso, M., Peek, J., & Sinjou, J. (1982). The compact disc digital audio system. *Philips Technical Review, 40*(6), 151-156.

Chinn, H. (1957, August). Splicing video tape. *Industries & Tele-Tech,* 79.

Compton's interactive encyclopedia 1995. (1995) . *[CD-ROM.* d.). SoftKey International, Inc.

Freed, L. (1995). *The history of computers.* Emeryville, CA: Ziff-Davis Press.

Harris, Jr., C., Pritchard, M., & Rabins, M. (2000). *Engineering ethics: Concepts and cases* (2nd Ed.). Belmont, CA: Wadsworth.

Hochheiser, S. (1992). What makes the picture talk: AT&T and the development of sound motion picture technology. *IEEE Transactions on Education, 35*(4), 278-285.

Jack, K. (1996). *Video demystified: A handbook for the digital engineer* (2nd ed.). San Diego, CA: HighText Publications.

Kieler, M., & West, M. (2004). Digital orphans: Technology's wayward children. In Linda L. Brennan & Victoria E. Johnson (Eds.), *Social, ethical and policy implications of information technology* pp. 234-250). Hershey, PA: Idea Group Publishing.

Pennavaria, K. (2003). Nonprint media preservation: A guide to resources on the Web. *College Resources and Library News, 64*(8).

Saussure, F. (1974). *Course in general linguistics*. London: Fontana.

Schinzinger, R., & Martin, M. (2000). *Introduction to engineering ethics.* New York: McGraw-Hill.

Shannon, C.E. (1948, July, October). A mathematical theory of communications. *Bell Systems Technical Journal,* 379-423 (July) and 623-656 (October).

Shima, S. (1984). The evolution of consumer vtr's – Technological milestones. *IEEE Transactions on Consumer Electronics, CE-30*(2), 66-69.

Wayner, P. (1997). *Digital copyright protection.* Chestnut Hill, MA: Academic Press.

White, R. (2002). *How computers work* (6th Ed.). Indianapolis, Indiana: Que.

KEY TERMS

Analog: Encoding a physical phenomenon by a direct, perceptually continuous variation of a physical property such as electromagnetic intensity (recording tape), mechanical displacement (Vinyl disk), or opaqueness (photographic film).

ASCII: American Standard Code for Information Interchange. A standard method of encoding upper and lower case text and other symbols with a 7-bit code.

Baudot: A 5-bit standard encoding method for uppercase letters, Invented by Emilie Baudot in 1874.

Intellectual Creation: Any work of creation, such as authorship, visual arts, performing arts, or music.

Lossy Encoding: Removal of data that represents redundant information, or differences presumed imperceptible to humans, in order to reduce stored or transmitted quantities of digital data.

Masking: In psychoacoustic models, the ability of a loud tone to block the perception of sounds or noise occurring in nearby frequencies.

Psychoacoustic Models: Models of human aural perception, especially with regard to the ability or inability to perceive signals that are masked by other signals.

Semiotics: A theory of signs and the use of signs in languages. Semiotics posits the arbitrary nature of the sign and looks at how human languages encode and transmit messages.

Wood Pulp: An inexpensive paper stock. Acid, used as part of the production process, frequently remains in the paper and causes its destruction. This destruction is rapid compared to paper made from fiber.

WYSIWYG: "What you See is What you Get." A user display that shows text and graphical information exactly as it will appear in print or other subsequent distribution. This includes expressional and artistic variations of text fonts, such as italics, bold, and underlined text.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 152-156,

Artificial Intelligence and Investing

Roy Rada
University of Maryland, Baltimore County, USA

INTRODUCTION

The techniques of artificial intelligence include knowledge-based, machine learning, and natural language processing techniques. The discipline of investing requires data identification, asset valuation, and risk management. Artificial intelligence techniques apply to many aspects of financial investing, and published work has shown an emphasis on the application of knowledge-based techniques for credit risk assessment and machine learning techniques for stock valuation. However, in the future, knowledge-based, machine learning, and natural language processing techniques will be integrated into systems that simultaneously address data identification, asset valuation, and risk management.

WHAT IS ARTIFICIAL INTELLIGENCE?

Computers play a role in many aspects of investing. For example, program trading is computer-driven, automatically executed trading of large volumes of shares, and has become increasingly prominent on stock exchanges. Artificial intelligence is a technique of computing that is perpetually on the cutting edge of what can be done with computers. Artificial intelligence could apply to program trading, but also other aspects of investing.

In the early days of computing, a typical task for a computer program was a numerical computation, such as computing the trajectory of a bullet. In modern days, a typical task for a computer program may involve supporting many people in important decisions, backed by a massive database across a global network. As the tasks that computers typically perform have become more complex and more closely intertwined with the daily decisions of people, the behavior of the computer programs increasingly assumes characteristics that people associate with intelligence. When, exactly, a program earns the label of "artificial intelligence" is unclear. The classic test for whether a program is intelligent is that a person would not be able to distinguish a response from an intelligent program from the response of a person. This famous Turing Test is dependent on factors not easily standardized, such as what person is making the assessment under what conditions.

A range of computer programming techniques that are currently, popularly considered artificial intelligence techniques includes (Rada 2008):

- Knowledge-based techniques, such as in expert systems.
- Machine learning techniques, such as genetic algorithms and neural networks.
- Sensory or motor techniques, such as natural language processing and image processing.

These methods may apply to investing. For instance, expert systems have been used to predict whether a company will go bankrupt. Neural networks have been used to generate buy and sell decisions on stock exchange indices. Natural language processing programs have been used to analyze corporate news releases, and to suggest a buy or sell signal for the corporate stock.

While artificial intelligence (AI) could apply to many areas of investing, much of what happens in computer-supported investing comes from non-AI areas. For instance, computational techniques not considered primarily AI techniques include numerical analyses, operations research, and probabilistic analyses. These non-AI techniques are routinely used in investing.

INVESTING AND DATA

The process of investing has three stages of:

- Data identification,
- Asset valuation, and
- Risk management.

AI has been most often applied to asset valuation, but is also applicable to data identification and risk management.

Two, high-level types of data used in financial investing are technical data and fundamental data. The price of an asset across time is technical data, and lends itself to various computations, such as the moving average or the standard deviation (volatility). Fundamental data should support cause-and-effect relationships between an asset and its price. For instance, the quality of management of a company should influence the profitability of a company and thus, the price of its stock.

The universe of fundamental data is infinite. Many streams of data that might be relevant, such as corporate earnings or corporate debt, might also be related to one another. Various

non-AI tools, such as linear regression analysis and principal components analysis, might be used in identifying what sets of data are more likely to be useful than what other sets. Such non-AI, computational techniques can be combined with AI techniques in experimenting with various combinations of data and choosing what data to use in asset valuation.

ASSET VALUATION

Different computational techniques might be appropriate for different assets or for different types of data for a particular asset. For instance, both stocks and commodities have price histories that might be tracked by the same time series analysis methods. However, the knowledge bases that would apply to valuing these assets might be significantly different. For example, knowledge about corporate management is less germane to commodity valuation than to stock valuation, while knowledge about weather patterns is more germane to commodity valuation than stock valuation.

Many assets have derivatives, such as options, that are priced and exchanged on markets. The computational characteristic of the Black-Scholes option pricing equation means that option valuation is done with the support of computer programs. Solving the Black-Scholes equation is not an artificial intelligence operation, although adequately handling options valuation could well involve artificial intelligence techniques.

An index is a special kind of asset. For instance, Standard & Poor's 500 Index (S&P500) is a widely traded asset that represents, with a single number, the price of shares of 500 companies. Programs that would be appropriate for evaluating the fair price of the S&P500 would be different from programs designed to evaluate the fair price of a particular company's shares. Among other things, the S&P500 does not have corporate management, per se. Neural networks have been extensively applied to predicting prices of stock indices.

A typical technical approach to a market problem (Chun & Park, 2006) took daily values over 5 years for five attributes of the Korean stock price index: daily high and low values, daily opening and closing values, and daily trading volume. On the other hand, the bond rating work of (Kim & Lee, 1995) looks at fundamental data with an expert system The input data for the bond rating work considers the quality of management and the quality of financial policies. The expert system's approach has a professional interactively answer questions from the system. Through this user interactivity, the system might collect subjective information, such as a company's management quality.

If a bank considers lending money to a company, the bank would be interested in judging the likelihood that the company would go bankrupt. More generally, financial institutions that lend money want to judge the credit worthiness of the entities to which they might lend money. These valuations of credit worthiness are a kind of asset valuation, but the techniques for doing this credit assessment would tend to be different from those for assessing the fair price of a company share. In particular, expert systems are more likely to be used for bankruptcy predictions, and neural networks are more likely to be used for stock price prediction. A bank may take its time in deciding what conditions, if any, to offer for a loan. Once the loan is made, its conditions are not subject to ready change. Investing in stocks or financial derivatives may be a fast-moving activity based on a history of prices. Those prices may be volatile, and entry and exit from the market may occur any time. The speculative financial investing problem is more of a time-series problem than the financial accounting problem and is thus, amenable to a different set of computational tools.

RISK MANAGEMENT

Risk or portfolio management involves choosing the asset classes in which to invest, and modifying the held assets across time, so as to suit the investment objectives. Various mathematical models, such as the Markowitz portfolio selection model, may be used by professional managers to guide the diversification of holdings so as to minimize risk for any specified rate of return. Implementing this kind of portfolio management may rely on numerical computation at one level, but can also benefit from various artificial intelligence techniques.

Lee et al. (Lee, Trippi, Chu, & Kim,1990) have described a knowledge-based system for supporting portfolio management. The system has different agents for different tasks. One agent elicits client goals, another agent implements dynamic hedging strategies, and another suggests market-timing decisions. Lee et al. note that the agents are only successful in narrow domains, and intervention of the human, portfolio manager is regularly necessary. In more recent work, Abdelazim and Wahba (2006) use genetic algorithms and neural networks to modify the parameters suggested by the Markowitz portfolio selection model, and obtain portfolios that earn higher returns at a specified risk level.

AI TRENDS

A multiagent architecture for an integrated system that considers data identification, asset valuation, and risk management has been proposed by researchers at Carnegie Mellon University. The system is called WARREN, which refers to the first name of the famous investor Warren Buffet (Sycara, Decker, Pannu, Williamson, & Zeng, 1996). The WARREN

system design includes components for collecting large amounts of real-time data, both numeric and textual. The data would be preprocessed and then fed to various asset valuation agents that would, in turn, feed their assessments to a portfolio management agent. The portfolio management agent would interact with clients of WARREN. Systems with various features of WARREN are available from commercial vendors, and are developed in-house by large investing companies, but more research is needed on how to develop integrated, AI systems that support investing.

Natural language processing systems may include large bodies of domain knowledge and parse free text, so as to make inferences about the content of the text. However, such natural language processing systems do not seem as popular in investing applications as much simpler natural language processing techniques. The natural language processing work that has been applied to the investing seems to be largely of the sort in which the distribution of word frequencies in a document is used to characterize the document. In this word-frequency way, Thomas (2003) has shown a potential value to processing news stories to help anticipate stock price changes.

As one can see cycles in the value of financial assets, one can also see cycles in the frequency of publication of articles on certain topics. In the field of artificial intelligence, one might identify, roughly speaking, three phases, as follows (Rada, 2008):

1. machine learning, in what was then called perceptron and self-organizing systems research, was popular from 1955 to 1975,
2. knowledge-based, multiagent, or expert systems work was popular from 1975 to 1995, and
3. machine learning research, now called neural networks or genetic algorithms research, returned to dominate the AI research scene from 1995 to the date of this article.

When AI research has been applied to investing, the AI technique used has tended to be the technique popular at the time. This leaves, unaddressed, the question of whether investing is more appropriately addressed with one AI technique or another.

The recent literature is rich with neural network applications to investing, but a new trend is the combining of knowledge-based techniques with neural network and genetic algorithm techniques. For instance, Tsakonas et al. (Tsakonas, Dounias, Doumpos, & Zopounidis, 2006) use "logic" neural nets that can be directly understood by people (traditional neural nets are a "black box" to humans). Genetic programming modifies the architecture of the logic neural net by adding or deleting nodes of the network in a way that preserves the meaning of the neural net to people and to the net itself. Bhattacharyya et al. (Bhattacharyya, Pictet, & Zumbach,2002) have added knowledge-rich constraints to the genetic operators in their application for investing in foreign exchange markets.

A promising research direction is to combine the earlier knowledge-based work on financial accounting with the more recent work on machine learning for stock valuation. For instance, neural logic nets could represent some of the cause-effect knowledge from a bankruptcy system and become part of a learning system for predicting stock prices. Some of the bankruptcy variables are readily available online, such as a company's debt, cash flow, and capital assets.

The financial markets are human markets that evolve over time as opportunities to make profits in this zero-sum game depend on the changing strategies of the opponent. Thus, among other things, what is important in the input may change over time. An AI system should be able to evolve its data selection, asset valuation, and portfolio management components. The future direction for AI in investing is to integrate the three major tools of AI (knowledge-based systems, machine learning, and natural language processing) into a system that simultaneously handles the three stages of investing (data collection, asset valuation, and portfolio management). Such systems will interact with humans so that humans can specify their preferences and make difficult decisions, but in some arenas, such as program trading, these sophisticated AI systems could compete with one another.

REFERENCES

Abdelazim, H., & Wahba, K. (2006). An artificial intelligence approach to portfolio selection and management. *International Journal of Financial Services Management*, 1(2-3), 243-254.

Bhattacharyya, S., Pictet, O. V., & Zumbach, G. (2002). Knowledge-intensive genetic discovery in foreign exchange markets. *Evolutionary Computation, IEEE Transactions on*, 6(2), 169-181.

Chun, S.-H., & Park, Y.-J. (2006). A new hybrid data mining technique using a regression case based reasoning: Application to financial forecasting. *Expert Systems with Applications*, 31(2), 329-336.

Kim, B.-O., & Lee, S. M. (1995). Bond rating expert system for industrial companies. *Expert Systems with Applications*, 9(1), 63-70.

Lee, J. K., Trippi, R. R., Chu, S. C., & Kim, H. S. (1990). K-FOLIO: Integrating the Markowitz model with a knowledge- based system. *Journal of Portfolio Management*, 17(1), 89-93.

Rada, R. (2008). Expert systems and evolutionary computing for financial investing: A review. *Expert Systems with Applications. 34*(4), 2232-2240

Sycara, K., Decker, K., Pannu, A., Williamson, M., & Zeng, D. (1996). Distributed intelligent agents. *IEEE Expert,* 11(6), 36-46.

Thomas, J. D. (2003). News and trading rules. In *Computer Science, PhD Thesis* (pp. 1-214), Carnegie Mellon University, Pittsburgh.

Tsakonas, A., Dounias, G., Doumpos, M., & Zopounidis, C. (2006). Bankruptcy prediction with neural logic networks by means of grammar-guided genetic programming. *Expert Systems with Applications,* 30(3), 449-461.

KEY TERMS

Artificial Intelligence: The ability of a computer to perform activities normally considered to require human intelligence

Asset Valuation: The process of determining the worth of something

Expert System: A program that uses knowledge and inferences to solve problems in a way that experts might

Investing: The act of committing money to an endeavor with the expectation of obtaining profit.

Neural Networks: Programs that simulate a network of communicating nerve cells to achieve a machine learning objective

Risk Management: The process of managing the uncertainty in investment decision-making.

Artificial Intelligence Applications in Tourism

Carey Goh
The Hong Kong Polytechnic University, Hong Kong

Henry M. K. Mok
The Chinese University of Hong Kong, Hong Kong

Rob Law
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

The tourism industry has become one of the fastest growing industries in the world, with international tourism flows in year 2006 more than doubled since 1980. In terms of direct economic benefits, United Nations World Tourism Organization (UNWTO, 2007) estimated that the industry has generated US $735 billion through tourism in the year of 2006. Through multiplier effects, World Travel and Tourism Council (WTTC, 2007) estimated that tourism will generate economic activities worth of approximately US $5,390 billion in year 2007 (10.4% of world GDP).

Owing to the important economic contribution by the tourism industry, researchers, policy makers, planners, and industrial practitioners have been trying to analyze and forecast tourism demand. The perishable nature of tourism products and services, the information-intensive nature of the tourism industry, and the long lead-time investment planning of equipment and infrastructures all render accurate forecasting of tourism demand necessary (Law, Mok, & Goh, 2007). Past studies have predominantly applied the well-developed econometric techniques to measure and predict the future market performance in terms of the number of tourist arrivals in a specific destination. In this chapter, we aim to present an overview of studies that have adopted artificial intelligence (AI) data-mining techniques in studying tourism demand forecasting. Our objective is to review and trace the evolution of such techniques employed in tourism demand studies since 1999, and based on our observations from the review, a discussion on the future direction of tourism research techniques and methods is then provided. Although the adoption of data mining techniques in tourism demand forecasting is still at its infancy stage, from the review, we identify certain research gaps, draw certain key observations, and discuss possible future research directions.

BACKGROUND

Econometric modeling and forecasting techniques are very much highly exploited to the understanding of international tourism demand. The econometric techniques have evolved and improved in accuracy and sophistication in recent years with the development of stationarity consideration, cointegration, and GARCH and time varying parameter (TVP) models. However, the development of such techniques might have reached a plateau for the time being. It is often expressed that econometric models are expanded at the expense of model comprehensiveness. That is, demand relationship is represented in a complicated system of equations in such models, which policy makers and practitioners find it like a black box that is hard to comprehend.

Besides, the conventional econometric models are founded on strict statistical assumptions and stringent economic theory (e.g., utility theory and consumption behavioral theory). These theories suggest that economic factors, such as income, price, substitute price, and advertising, are the primary influences of demand. However, tourism consumption, particularly if it involves a long-haul trip, is restricted by not only income constraint, but also on time constraint (Cesario & Knetsch, 1970; Morley, 1994), and the *intrinsic* properties of a particular tourism product or service that have not been modeled in the traditional demand framework (Eymann & Ronning, 1992; Morley, 1992; Papatheodorou, 2001; Rugg, 1973; Seddighi & Theocarous, 2002). The noneconomic factors, such as psychological, anthropological, and sociological factors, had not been analyzed in the traditional econometric travel demand studies. It is true that there exists a large amount of tourism literature that studies how these noneconomic factors could affect travel motivation, and how travel motivation affects destination choices (e.g., Edwards, 1979; Gray, 1970; Mayo & Jarvis, 1981; Um & Crompton, 1990), published articles have failed to show how noneconomic factors could affect demand for tourism, however.

Researchers have attributed the reasons for not including many relevant variables to the lack of data availability and difficulty in obtaining exact measures for the determining factors (Gonzelez & Moral, 1995; Kulendran, 1996, Song & Witt, 2000; Song *et al.*, 2000). Perhaps the true reason lies in the expense of the increasing complexity of a model in exchange for the inclusion of more determining factors, as noted in Turner and Witt (2001), Stabler (1990), as well as Faulker & Valerio (2000).

A RETROSPECTIVE VIEW ON DATA MINING IN TOURISM DEMAND ANALYSIS AND FORECASTING

In view of the growing importance of data mining in business applications, we review and analyze the relevant articles that adopt data-mining techniques to forecast tourism demand in tourism research journals. We sorted the 70 research journals in tourism, as mentioned by McKercher, Law, and Lam (2006), for the period between 1980 and early 2007, and the ScienceDirect, and the Hospitality and Tourism Complete index on EBSCOhost Web for the period between mid-2006 and early 2007. We identified 174 papers on tourism demand forecastin,g published in a 28-year span (1980 to early 2007) and among them, only 14 used data-mining techniques, with the first one published in 1999. Nine out of the 14 papers were published in the last 5 years (2003-early 2007).

A variety of neural network (NN) systems was adopted in these 14 studies, and they include supervised feed forward NN, back propagation NN, Elman's NN, multilayer perceptron NN, radial basis function NN, and Bayesian NN. Others have adopted rough sets, support vector regression, group method of data handling, fuzzy time series, and grey forecasting model.

Neural Networks

Traditionally, the term neural network had been used to refer to a network of biological neurons, but it is often refered to as artificial neural networks (ANN) nowadays. ANNs are computer software that mimics the human intelligence to deduce or learn from a dataset. What gives ANNs excellent classification and pattern recognition ability are their capabilities of representing knowledge based on massive parallel processing, and pattern recognition based on past experience. This pattern recognition ability makes ANNs a superb classification and forecasting tool for industrial applications.

Supervised Feed Forward Neural Networks (SFFNN)

The SFFNN was first adopted by Uysal and El Roubi (1999) and Law and Au (1999) in tourism demand studies. Uysal and El Roubi (1999) developed a preliminary neural network that used Canadian tourism expenditures in the United States as dependent variable; whereas per capita income of Canada, consumer price index, lagged expenditure by one period, and seasonal dummy variables were used as independent variables. Their findings showed that the neural network achieved highly accurate results with high-adjusted correlations and low-error terms. Law and Au (1999) used the model to forecast Japanese tourist arrivals to Hong Kong. Using six independent variables, including relative service price, average hotel rate in Hong Kong, foreign exchange rate, population in Japan, marketing expenses by the Hong Kong Tourist Association, and gross domestic expenditure in Japan, the neural network achieved the lowest mean average percentage error (MAPE), and highest acceptance percentage and normalized correlation coefficient compared to four other commonly used econometric models, including multiple regression, naïve I, moving average, and single exponential smoothing.

The SFFNN was adopted by Law (2001) in an attempt to capture turbulent travel behavior as a result of regional financial crisis in Asia in 1997. Law (2001) tested and compared the forecasting accuracy of this model with naïve I, naïve II, moving average, single exponential smoothing, Holt's exponential smoothing, and multiple regression in predicting Japanese demand for travel to Hong Kong. Using the same variables as in Law and Au (1999) and Law (2000), but with updated data, neural network, again, outperformed other techniques in three of the five accuracy measurements, including MAPE, trend change accuracy, and the closeness between actual and estimated values. Law (2001) concluded that no single forecasting method could outperform others in all situations when there was a sudden environmental change, but neural network appeared to perform reasonably well in terms of predicted values. Using tourism demand data of South Africa, Burger *et al.* (Burger, Dohnal, Kathrada, & Law, 2001) also investigated performance and SFFNN with a few benchmark models, such as Naïve I, moving average, decomposition, single exponential smoothing, ARIMA, autoregression, and found that neural networks with 1-month-ahead forecast generated the lowest MAPE values. Radial basis function network, a type of SFFNN, was employed by Kon and Turner (2005), but the forecasting performance falls behind that of multilayer peceptron NN model in all seasonal and deseasonalized series.

Back Propagation Neural Networks (BBNN)

BPNN was found as a core model in Law (2000), and it is used to analyze the nonlinearly separable Taiwanese visitor arrivals to Hong Kong from 1967 to 1996 using similar variables in Law and Au (1999). Their results showed that the calibrated neural network model outperformed all other approaches (commonly used econometric models, including multiple regression, naïve I, moving average, and single exponential smoothing), by attaining the highest forecasting accuracy. Besides, Kon and Turner (2005) also employ BPNN as a type of multilayer perceptron to analyze and forecast tourism demand for Singapore. Their results confirm that correctly structured NN model can outperform basic structural model and other simpler methods, such as Naive I and Holt-Winters model, in the short run. The BPNN was also used in Chen and Wang (2007) as a model for comparison with support vector regression.

Elman's Neural Networks (Elman's NN)

The Elman's NN has the advantageous time series prediction capability because of its memory nodes, as well as local recurrent connections. Cho (2003) established an Elman's NN in its time series analysis to forecast the number of visitors from different origins to Hong Kong. Their findings showed that Elman's neural networks performed the best in five of the six origins, compared with ARIMA model and exponential smoothing.

Bayesian Neural Networks (BNN)

BNN model was tested together with multilayer perceptron neural network (MLP) and radial basis function neural network (RBF) in Kon and Turner (2005) on both deseasonalized and nondeseasonalized data series. The BNN model did not forecast accurately on the sample data.

Rough Sets (RS)

RS algorithms can handle both numeric and nonnumeric data. As such, the noneconomic factors, along with economic factors, can be analyzed without having to bear the problem of insufficient degree of freedom and loss of information from transforming the nonnumeric factors into numeric ones.

There were two regression papers that incorporated the RS theory to tourism demand forecasting. In the first paper, Goh and Law (2003) built an RS-based model using independent variables of real gross domestic product, relative consumer price index, population, trade volume, and foreign exchange rate to form six decision rules for forecasting tourist arrivals. Their findings achieved 87.2% accuracy rate in six samples. High classification capability of the RS model was also demonstrated in Law, Goh, and Pine (2004). The empirical findings showed that the RS-based model correctly classified 86.5% of the testing data.

Goh, Law, and Mok, (forthcoming), for the first time, have introduced a leisure time index and a climate index into the rough sets forecasting framework.

Support Vector Regression (SVR)

The model was tested by Pai and Hong (2005) on arrival data in Barbados. SVRs are normally used to avoid the over- or the under-fitting of data series. In SVR models, parameter values of structural error range (ε), the flatness (C), and the curvature (σ) could be adjusted through a dynamic shift procedure. In this study, genetic algorithms (Holland, 1975) were employed to select the optimal parameter for the three SVR model parameters. The results indicated that support vector machines with genetic algorithms (SVMG) outperform ARIMA model and accurately forecast arrivals in Barbados, and can predict the increasing tendency of arrivals properly. The dynamic shift procedure easily captures the data patterns, and the genetic algorithms provide a proper procedure to determine SVR parameters, hence the whole process could minimize the structural risk. For these reasons, SVMG is particularly suitable to forecast tourist arrivals that tend to exhibit profound structural change in the series.

Group Method of Data Handling: Genetic Regression (GMDH)

The GMDH has an inductive nature that automatically finds interrelations in data and selects the optimal model structure by sorting out possible variants. This automatic sorting of different solutions in the GMDH networks minimizes the researcher's manipulation on the results of modeling. Burger *et al*. (2001) used this model alongside with eight other models, including SFFNN, to forecast travel demand of American for Durban. However, the model was found to perform poorly owing to the fact that GMDH algorithm needs to reserve a large set of data for validation and hence, leaving a relatively small data set for model fitting. It was believed that the model performance should improve with a larger overall data set, however.

Fuzzy Time Series (FTS)

The main assumption of FTS model is that the variation of this year is related to the variation of previous years and the model is, therefore, appropriate for analyzing tourism data

that exhibits habit persistence. As FTS model is designed to forecast processes with linguistic value observation, in Wang's (2004) and Yu and Schwartz's (2006) studies, the numerical value of tourism demand was translated into a linguistic value, and then the predicted variation value is translated back into a numeric value. In Wang (2004), two AI models, FTS, hybrid grey theory and a Markov modification model were used to forecast arrivals to Taiwan from Hong Kong, United States, and Germany. The empirical results indicate that FTS model leads the others only on forecasting tourism demand of Hong Kong to Taiwan. Yu and Schwartz (2006) also established FTS forecasting model and grey-theory-based model for tourist arrivals to the United States from eight origins, and compared them with two benchmarking models, double exponential smoothing and double moving average model. Different from Wang (2004), experimental findings indicated that the sophisticated data-mining technique did not produce more accurate forecasting results. The authors thus cautioned that tourism researchers should proceed with care when dealing with innovative data-mining techniques.

Grey Forecasting Model (GFM)

Similar to FTS model, GFM is particularly suitable for forecasting with few available past observations, such as one for tourism. In Wang (2004), the study revealed that GFM performed well when sample data showed a stable increasing trend while Markov modification model performs better when there is significant fluctuation in the series.

KEY OBSERVATIONS AND ANALYSIS

We derive a few observations from this review on tourism demand studies that adopted data-mining techniques. These articles were predominantly published in first-tier and second-tier research journals in tourism. According to a recent survey with worldwide scholars, two of these journals were rated at 3.6 and 3.8, and the other journals were all rated above 4 on a 5-point Likert scale (McKercher, Law, & Lam, 2006).

Both time series and regression approaches were used in the reviewed studies. Although six studies were to investigate tourism demand relationship, their results were either not compared with econometric models, or the comparisons were done with simple regression models. Other than Uysal and El Roubi's (1999) study, other studies only analyzed annual data under which seasonality, an important phenomenon in tourism, was basically neglected in these studies.

One interesting observation for tourism demand forecasting studies using data mining is the dominance of artificial neural network approach. For example, Burger *et al*. (2001), Law and Au (1999); Law (2001), and Uysal and El Roubi (1999) all applied SFFNN. Later, researchers extended the applicability of neural networks by incorporating back-propagation learning process into nonlinearly separable tourism demand data (Chen & Wang, 2007; Law, 2000). Elman's concurrent network was also used to fit time series tourism data (Cho, 2003). In a more comprehensive approach, Kon and Turner (2005) fitted the tourism time series data with three different network architectures, including the multilayer perceptron (MLP), radial basis function network (RBF), and BNN, and compared the models' performance with a number of time series econometric models.

We also observe that the data-mining techniques adopted in tourism demand forecasting studies were mainly used to analyze tourist arrivals to a specific destination from either one specific origin (Burger *et al*.., 2001; Law & Au, 1999; Law, 2000; Law, 2001; Law *et al*.. 2004; Uysal & El Roubi, 1999), a number of selected origins (Chen & Wang, 2007; Pai & Hong, 2005; Palmer, ., Montaño, & Sesé, 2006), or from an aggregate total of certain countries (Cho, 2003; Goh & Law, 2003; Kon & Turner, 2005).

FUTURE TREND

The review reveals several research gaps. It is observed that the forecasting accuracy of the adopted data-mining techniques was often compared with the traditional and less sophisticated forecasting techniques; thus it remains largely unknown whether data-mining techniques are able to outperform the advanced econometric techniques, such as nonstationary regression approaches, system of equation techniques, and the time varying parameter (TVP) models. The comparison between forecasting quality of data-mining techniques and advanced econometric techniques is crucial for tourism data-mining researchers to be confident of their research findings.

Another noticeable gap is the effective nonexistence of examining noneconomic variables in the published data-mining articles. With the exception of Goh, Law, and Mok, (2008), who have introduced a leisure time index and a climate index into the rough sets forecasting framework, none of the reviewed articles has investigated the impact of noneconomic factors on tourism demand. Given the fact that strict statistical assumptions are not required in data-mining techniques, as they are in econometric models, noneconomic variables, which are found in many tourism studies (Decrop, 1999; Galarza *et al.*, 2003, Heung, Qu, & Chu, 2001; Um & Crompton, 1990) that may affect tourists' decision making could also be investigated using data-mining techniques.

Apparently, prior studies on tourism demand forecasting using data-mining techniques largely followed the variables that have long been used by economists. This is manifested by the same variables of demand measurements instead of

innovative variables. Moreover, hybrid data-mining systems, such as fuzzy neural networks, were still unpopular in tourism demand studies. The last observable gap related to the use of data for model calibration and model testing. Virtually all prior studies on data mining in tourism demand forecasting relied on regional instead of truly international data.

CONCLUSION

The interest in adopting data mining into tourism demand forecasting is evident by the fact that ever since the initial attempt in 1999, articles on data mining and tourism demand forecasting have been regularly published in leading tourism journals. Nevertheless, comparing to the traditional econometric modeling techniques, data mining is still at its infancy stage. What are needed in the future are the endeavors to integrate different data-mining approaches into a commonly agreeable process that can be applied to the tourism industry at large. In the meantime, the introduction of more rigorously developed data-mining techniques with sufficient tests on both primary and secondary data (instead of purely using secondary data) is needed. As right now, most of the published tourism-demand forecasting articles offered only general suggestions, it would be desirable to generate more applicable business solutions calibrated with industrial experts using qualitative approaches, such as jury of executive opinion and Delphi method.

The coverage of this review is limited by the selection channels and the number of published papers. As such, future research efforts can certainly go beyond this limitation in order to have a more comprehensive review of related literature. The research gaps that have been discussed also deserve further investigations by researchers in data mining and tourism forecasting.

ACKNOWLEDGMENT

This research was supported by Research Grant: CUHK4631/06H.

REFERENCES

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001). A practitioners guide to a time-series methods for tourism demand forecasting – a case study of Durban, South Africa. *Tourism Management 22,* 403-409.

Cesario, F. J., & Knetsch, J. K. (1970). Time bias in recreation benefits estimates. *Water Resources Research, 6*(3), 700-704.

Chen, K. Y., & Wang, C. H. (2007). Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management 28*, 215-226.

Cho, V. (2003). A comparison of three different approaches to tourist arrival forecasting. *Tourism Management 24*, 323-330.

Decrop, A. (1999). Tourists' decision-making and behavior processes. In *Consumer Behavior in Travel and Tourism*. New York: Haworth Hospitality Press.

Edwards, A. (1979). *International tourism development forecasts to 1985*. Special Reports No. 33. London: The Economist Intelligence Unit, Ltd.

Eymann, A., & Ronning, G. (1992). Discrete choice analysis of foreign travel demand. In **Hans-Jürgen Vosgerau (Ed.),** *European Integration and World Economy***.** Berlin: Springer-verlag..

Faulker, B., & Valerio, P. (2000). An integrative approach to tourism demand forecasting. In *Tourism Management: Towards the New Millennium.* New York: Pergamon.

Frechtling, D. C. (2001). *Forecasting tourism demand: Methods and strategies.* Oxford: Butterworth-Heinemann.

Galarza, M. G., Saura I. G., & Garcia, H. C. (2002). Destination image: Towards a conceptual framework. *Annals of Tourism Research 29*(1), 56-78.

Goh, C., & Law, R. (2003). Incorporating the rough sets theory into travel demand analysis. *Tourism Management, 24*(5), 511-517.

Goh, C., Law R., & Mok. H. M. K. (2008). Analyzing and forecasting tourism demand: A rough sets approach. *Journal of Travel Research, 46*(3), 327-338.

Gonzalez, P., & Moral, P. (1995). An analysis of the international tourism demand in Spain. *International Journal of Forecasting, 11*, 233-251.

Gray, H. P. (1970). *International travel – International trade*. Lexington, MA: D.C. Health and Company

Heung, V. C. S., Qu, H., & Chu, R. (2001). The relationship between vacation factors and socio-demographic and characteristics: The case of Japanese leisure travelers. *Tourism Management, 22*(3), 259-269.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Cambridge, MA: MIT Press.

Kon, S. C., & Turner, L. W. (2005). Neural network forecasting of tourism demand. *Tourism Economics 11*(3), 301-328.

Kulendran, K. (1996). Modelling quarterly tourist flows to Australia using cointegration analysis. *Tourism Economics, 2*(3), 203-222.

Law, R. (2000). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management, 21,* 331-340.

Law, R. (2001). The impact of the Asian financial crisis on Japanese demand for travel to Hong Kong: A study of various forecasting techniques. *Journal of Travel & Tourism Marketing, 10*(2/3), 47-65.

Law, R., & Au, N. (1999). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management 20*, 89-97.

Law, R., Goh, C., & Pine, R. (2004). Modeling tourism demand: A decision rules based approach. *Journal of Travel & Tourism Marketing, 16*(2/3), 61-69.

Law, R., Mok, H., and Goh, C. (2007). Data mining in tourism demand analysis: A retrospective analysis. *Lecture Notes in Computer Science, 4632*, 508-515

Mayo E. J. Jr., & Jarvis, L. P. (1981). *The psychology of leisure travel: Effective marketing and selling of travel services.* Boston: CBI Publishing Inc.

McIntosh, R. W., Goeldner, C. R., & Ritchie, J. R. B. (1995). *Tourism: Principles, practices, philosophies.* New York: John Wiley & Sons.

McKercher, B., Law, R., & Lam, T. (2006). Rating tourism and hospitality journals. *Tourism Management, 27*(6), 1235-1252.

Morley, C. L. (1992). A microeconomic theory of international tourism demand. *Annals of Tourism Research, 19,* 250-267.

Morley, C. L. (1994). Experimental destination choice analysis. *Annals of Tourism Research, 21*(4), 780-791.

Pai, P. F., & Hong, W. C. (2005). An improved neural network model in forecasting arrivals. *Annals of Tourism Research, 32(*4), 1138-1141.

Palmer, A., Montaño, J. J., & Sesé, A. (2006). Designing an artificial neural network for forecasting tourism time series. *Tourism Management, 27*, 781-790.

Papatheodorou, A. (2001). Why people travel to different places, *Annals of Tourism Research¸ 28*(1), 164-179.

Rugg, D. (1973). The choice of journey destination: A theoretical and empirical analysis. *Review of Economics and Statistics, 55*(1), 64-72.

Seddighi, H. R., & Theocharous, A. L. (2002). A model of tourism destination choice: A theoretical and empirical analysis. *Tourism Management, 23*(5), 475-487.

Song, H., & Witt, S. F. (2000). *Tourism demand modeling and forecasting. Modern econometric approach* (1st ed.). New York: Pergamon.

Song *et al.*, 2000

Stabler, M. J. (1990). The image of destination regions: theoretical and empirical aspects. In B. Goodall & G. Ashworth (Eds.), *Marketing in the tourism industry: The promotion of destination regions.* London.

Turner, L. W., & Witt, S. F. (2001). Forecasting tourism using univariate and multivariate structural time series models. *Tourism Economics, 7*(2), 135-147.

Um, S., & Crompton, J. L. (1990). Attitude determinants in tourism destination choice. *Annals of Tourism Research, 17*(3), 432-448.

Uysal, M., & El Roubi, S. E. (1999). Artificial neural networks vs. multiple regression in tourism demand analysis. *Journal of Travel Research, 38,* 111-118.

Wang, C. H. (2004). Predicting tourism demand using fuzzy time series and hybrid grey theory. *Tourism Management, 25*, 367-374.

Ward, F. A., & Beal, D. (2000). *Valuing nature with travel cost models*. Northampton, MA: Edward Elgar.

United Nations World Tourism Organization. (2007). *Tourism highlights 2007*. Madrid, Spain.

World Travel and Tourism Council. (2007). *Executive summary: Travel & tourism navigating the path ahead.* WTTC.

Yu, G., & Schwartz, Z. (2006). Forecasting short time-series tourism demand with artificial intelligence models. *Journal of Travel Research, 45,* 194-203.

KEY TERMS

Cointegration: An econometric property of time series variables. If two or more series are themselves nonstationary, but a linear combination of them is stationary, then the series are said to be cointegrated.

Econometrics: It is concerned with the tasks of developing and applying quantitative or statistical methods to the study and elucidation of economic principles. It combines

economic theory with statistics to analyze and test economic relationships.

Multiplier Effect: In economics, it refers to the idea that an initial spending rise can lead to even greater increase in national income

Nonstationary: A condition where value of the time series changes over time, usually because there is some trend in the series so that the mean value is either rising or falling over time.

Perishable Tourism Product: Tourism services are perishable, and cannot be stored for sale at a later date. This affects the distribution of tourism services, as they must be marketed in a way that minimizes lost capacities.

Statistical Assumptions: Statistical assumptions are general assumptions about statistical populations. In order to generate interesting conclusions about real statistical populations, it is usually required to make some appropriate background assumptions in order to generate valid and accurate conclusions.

Assessing Critical Success Factors of ERP Implementation

Leopoldo Colmenares
Simon Bolivar University, Venezuela

INTRODUCTION

An enterprise resource planning (ERP) system is an integrated set of programs that provides support for core organizational activities. ERP is a software infrastructure embedded with "best practices," or best ways to do business based on common business practices or academic theory. The aim is to improve the cooperation and interaction between all the organizations' departments, such as the products planning, manufacturing, purchasing, marketing and customer service department. ERP systems is a fine expression of the inseparability of IT and business. As an enabling key technology as well as an effective managerial tool, ERP systems allow companies to integrate at all levels and utilize important ERP systems applications, such as supply-chain management, financials and accounting applications, human resource management and customer relationship management (Boubekri, 2001). ERP systems hold the promise of improving processes and decreasing costs. Furthermore, two important new frontiers for ERP systems are electronic business (e-business) and supply-chain management (Wang and Nah, 2001). The systems can connect with suppliers, distributors, and customers, facilitating the flow, the product and information.

ERP systems implementation is costly and complex. In many cases, an ERP system is the largest single investment in any corporate-wide project. The software is expensive, and the consulting costs even more. Meta Group found that the average ERP systems implementation takes 23 months with total owners' cost of $12 million (Stewart, 2000). The ERP systems implementation is the process where business process and ERP system match each other. Usually the firm has to change the business process per ERP systems. Sometimes most positions have to be redesigned according to the ERP systems. Thus the difficulties and high failure rate in implementing ERP systems have been widely cited in the literature (Davenport, 1998; Kim, Lee, & Gosain, 2005)). The failure percentage of ERP systems was determined by one study as ranging from 40 to 60% and from another study as between 60 and 90% (Langernwalter, 2000; Ptak and Schragenheim, 2000; Yingjie, 2005).

Although the failure rates of these ERP implementations have been highly publicized, this has not distracted companies from investing large sums of money on ERP systems (Somers & Nelson, 2004). ERP systems provide companies with the means of integrating their business functions into a unified and integrated business process. As companies implement more enterprise based systems throughout their organizations, the need for integration of these systems becomes even more paramount. Expanding from the functional areas of accounting, human resources, and shop floor control to an enterprise-wide system has become a format for producing full organization integration.

Over the past few years, limited research has been conducted about ERP implementation issues: mainly case studies in individual organizations have been reported. That is a motivation toward conducting empirical studies to explore critical factors that affect ERP systems implementation.

This study presents the results of an empirical study that surveyed managers from seven corporations, who were identified as having a key role in ERP systems implementation, in order to assess empirically which CSFs are critical in leading a successful implementation of ERP systems. A factor analysis solution was used to derive factors affecting successful ERP implementation. These factors are: ERP implementation management, users aptitudes and communication and technical knowledge. The study reveals that about 81.5 % of the variances in ERP systems implementation were explained by the critical factors identified in the study.

The remainder of this article is organized in four sections. First ERP-related literature is reviewed. The next section introduces the research methodology, followed by the presentation of the results. The paper ends with the conclusions and implications for future research and practice.

BACKGROUND

Implementing an ERP system is not an easy task (Tsai et al., 2005). It can cause dramatic changes that need to be carefully administered if the potential advantages of an ERP systems solution (Al-Mudimigh, Zairi, & Al-Mashari, 2001) are to be gained. In some well-documented cases, spectacular results have been achieved (Johnston, 2002). There is on the other hand a relatively high failure rate: it was reported that three-quarters of ERP systems projects were judged to be unsuccessful by the ERP systems implementing firms (Kyung & Young, 2002). What is more, failures are much less extensively documented. As a result, pitfalls to be avoided tend to be less well known.

A recent summary of ERP systems literature states that research of critical success factors (CSFs) in ERP systems implementation is rare and fragmented (Nah, Lau, & Kuang, 2001). Identifying CSFs relevant to local companies is one way to increase the chances of a successful local implementation (Sum, Ang & Yeo, 1997). The idea of identifying CSFs as a basis for determining the information needs of managers was popularized by Rockart (1979). CSFs are those factors that are critical to the success of any organization, in the sense that, if objectives associated with the factors are not achieved, the organization will fail—perhaps catastrophically (Rockart, 1979). In the context of ERP systems project implementation, CSFs represent the essential ingredients without which a project stands little chance of success. (Colmenares, 2005)

A literature review was conducted to understand the CSFs in successful ERP implementations. So we find that in a study on ERP implementation in China, the authors posit strong considerations for national cultural issues, since critical success factors may vary significantly, depending on the country in which an implementation is carried out (Shanks & Parr, 2000). ERP implementations have also been investigated through case studies with varying degrees to describe critical success factors. These include the impact of ERP on job characteristics (Perez & Rojas, 1999), strategic options open to firms beyond the implementation of common business systems (Upton & McAfree, 1997), means to avoid ERP project failures (Scott, 1999), issues of business alignment (Smethurst & Kawalek, 1999; Volkoff, 1999) business process reengineering (BPR) (Slooten & Yap, 2000), and change management (Klaus, Rosemann, & Gable, 2000). Others studies have assessed the ambiguous role of large systems as both catalysts and inhibitors to change (Pawlowski & Boudreau, 1999) analyze the special challenges of ERP implementations outside the business world (Sieber & Nah, 1999), and describe global supply chain (Chatfield & Andersen,1998). Implementing ERP with or without BPR has been surveyed and analyzed (Bernroider & Koch, 1999). Theoretical considerations have focused on global business processes (Basu & Palvia, 1999) and IT architecture options (Chan, 1999), as well as on enhancement of process engineering and development methodologies (Sato, 2000).

The critical challenge in ERP implementation has been to first identify the gaps between the ERP generic functionality and the specific organizational requirements (Soh, Kien, & Tay-Yap, 2000). Too often, ERP adopting companies fail to understand the business requirements which the ERP systems are expected to solve. The congruence between

Table 1. Critical success factors for ERP implementation

Top management support
User training
Use of consultants
User participation
Vendor package selection
Use of steering committee
Discipline and standardization
Minimal customization
Use of vendor's development tools
Best people full time
Technical and business knowledge
Implementation approach
Clear goals, focus and scope
Business process reengineering
Project management
Effective communications
Presence of a champion
Interdepartmental cooperation and communication
Management of expectations
Vendor/customer partnership

ERP systems and organizational culture is the prerequisite to successful ERP implementation (Hong & Kim, 2002). The implementation of an integrated system such as ERP requires that the basic business practices embedded in the ERP system be adapted to the organizational processes and culture. Others authors have developed studies that list the CSFs, but these lists are generally based on experience or casual observation, not controlled research. See for example Bingi, Sharma, and Godla (1999), Esteves and Pastor (2001), Falkowski et al. (1998), Holland and Light (1999), Nah, Lau, and Kuang (2001), Rosario (2000), Stefanou (1999), Sumner (1999) and Wee (2000).

Therefore we can conclude that the literature varies according to the variables required for implementation success, so there is not a general consensus as to the factors that are key to success in ERP implementation. It is probably a combination of factors that are important in explaining ERP implementation success (Zhang et al., 2003). However, from the review, twenty factors emerged frequently mentioned as critical to the successful implementation of ERP systems. They were obtained after careful analysis and grouping of related sub-factors. Table 1 shows these CSFs.

RESEARCH METHODOLOGY

The target of the study was the organizations that have implemented ERP systems successfully. The key informant method was used for collecting information in a social setting by surveying (or interviewing) a selected number of participants. Seven firms that proclaimed having succeeded in implementing ERP systems were identified from newspapers and computing magazines. We contacted the ERP project managers of each company in charge of ERP implementation. About one hundred questionnaires were sent to the ERP project managers of each firm, who forwarded the questionnaires to the project team members in charge of individual processes. A total of 86 questionnaires were returned, of which 84 were valid. The questionnaire consisted of two main parts: the company background and the CSFs. The first part was designed to determine characteristics such as size of the company, type of industry, location of company, etc. The second part consisted of twenty statements about the success factors of ERP systems implementation, derived from the literature review. The language used in the survey was Spanish. Translation was rather easy because Venezuelans used original English terms for many technical and management concepts and especially for management, information systems and computing concepts.

Participants were requested to rate the importance of each CSF using a five-point Likert scale, where a score of 5 indicated "extremely critical" and a score of 1 indicated "not critical." This method was employed on the grounds that a rating method avoids the problems of having to consider twenty CSFs simultaneously in order to rank them. The data collected was then analyzed by using SPSS. Based on the responses, descriptive statistics, factor analysis and reliability tests were carried out to identify the CSFs for the successful implementation of ERP systems and data validity respectively.

RESULTS

Ranking

The importance rating of the twenty CSFs is listed in Table 2.

The individual mean value of the Likert rating scale is the popular usage indicator for measuring an item's importance, without regard to the other items: so the higher the value the more important the factor. Most items are rated above the 3.0 scale (mid-point). The three most important factors, in order of declining importance, are: top management support, presence of a champion, and project management, with a mean value ranging from 4.80 to 4.64. Just as the literature argues, these are key items for ERP implementation management (Johnston, 2002). Conversely, use of steering committee, business process reengineering, and use of vendor's development tools, are the three items lowest in the list, with a mean value ranging from 2.95 to 2.06.

Factor Analysis

Exploratory factor analysis was used to analyze the managers' responses. The main reason for using the exploratory factor analysis was to reduce the dimensionality of the attribute data and uncover a smaller number of underlying factors that account for a major amount of the variance in the original measures [39]. Factor analysis is a data reduction technique that uses correlations between data variables. The underlying assumption of factor analysis is that a number of factors exist to explain the correlations or inter-relationships among observed variables (Chatfield & Collins, 1992). For the present study, factor analysis was performed on all twenty variables using principal components extraction (Tabachnick & Fidell, 1989). The goal of this method is to extract maximum variance from the data set within each factor. It is basically used to reduce a large number of variables down to a smaller number of components. The measure of sampling adequacy for the twenty items was 0.87 indicating that the items were suitable for factoring (Kaiser, 1974).

A three-stage factor analysis was conducted with an orthogonal (varimax) rotation to obtain a stable factor structure (Rai et al., 1996), resulting in easily interpretable factors. Under this three-round factor analysis, items were omitted according to the following two criteria: (1) no loading greater than 0.35, or (2) loading greater than 0.35 on two

Table 2. Ranking of CSFs

Rank	CSF	Mean	Std. Dev.
1	Top management support	4.80	0.62
2	Presence of a champion	4.75	0.85
3	Project management	4.64	0.92
4	Best people full time	4.58	0.60
5	Effective communications	4.51	0.85
6	Interdepartmental cooperation and communication	4.40	0.91
7	Management of expectations	4.36	1.02
8	Technical and business knowledge	4.33	1.21
9	User participation	4.22	0.82
10	Discipline and standardization	4.09	0.85
11	Vendor package selection	4.02	0.61
12	User training	4.01	1.12
13	Implementation approach	4.00	1.20
14	Clear goals, focus and scope	3.89	1.14
15	Use of consultants	3.75	0.85
16	Minimal customization	3.68	1.52
17	Vendor/customer partnership	3.15	0.52
18	Use of steering committee	2.95	0.63
19	Business process reengineering	2.84	0.55
20	Use of vendor's development tools	2.06	0.42

N: 84
Scale: 1-5 (5: "extremely critical"; 1:"not critical")

or more factors (Kim & Mueller, 1978). Table 3 shows the results of this analysis. A first factor analysis was conducted and produced five factors. According to the two criteria, three items were dropped. A second factor analysis on the remaining 17 items resulted in four factors and the dropping of three items. Finally, a three-factor structure was derived which kept a total of 14 items after three iterations. The minimum eigenvalue from a varimax rotation for which a factor was to be retained was set at 1.0 in order to satisfy the minimum eigenvalue criterion (Nunnally, 1978). The goal in labeling each of these factors is to come up with a term that best describes the content domain of the attributes that load highly on each factor. The naming of a particular factor is determined by those response variables that load on the factor. Next the three-factor structure is described:

- Factor 1, named "ERP implementation management," comprises six items relating to implementation management: top management support, presence of a champion, project management, management of expectations, implementation approach, and clear goals, focus and scope.
- Factor 2, named "user aptitudes and communication," comprises four items relating to user participation: effective communication, interdepartmental cooperation and communication, user participation, and user training.
- Factor 3, named "technical knowledge," comprises four items relating to knowledge of business and ERP: best people full time, technical and business knowledge, use of consultants, and discipline and standardization.

Cronbach alpha coefficients were calculated to test the reliability of these CSFs, as shown in the last row of Table 3. The reliability of coefficients obtained ranges from 0.56 (factor 3) to 0.88 (factor 1). Srinivasan (1985) proposed that a coefficient of 0.7 or higher is acceptable, while a coefficient of 0.5 or higher is considered sufficient when dealing with exploratory research combined with invalidated data. Thus, the reliability coefficients in this study are deemed

Table 3. Results of factor analysis

		Factor Loading (Final Factor Structure)		
CSF	**Mean**	**F1**	**F2**	**F3**
Top management support	4.80	0.608		
Presence of a champion	4.75	0.541		
Project management	4.64	0.586		
Best people full time	4.58			0.741
Effective communications	4.51		0.565	
Management of expectations	4.40	0.420		
Interdepartmental cooperation and communication	4.36		0.452	
Technical and business knowledge	4.33			0.584
User participation	4.22		0.562	
Discipline and standardization	4.09			0.387
User training	4.01		0.510	
Implementation approach	4.00	0.478		
Clear goals, focus and scope	3.89	0.411		
Use of consultants	3.75			0.573
Eingenvalue		7.56	2.54	1.20
Percentage of variance		58.50	14.36	8.65
Cumulative percentage of variance		58.50	72.86	81.51
Cronbach alpha coefficient		0.88	0.74	0.56

Scale: 1-5 (5: "extremely critical", 1:"not critical"); F1: ERP implementation management; F2: Users aptitudes and communication; F3: Technical knowledge

acceptable. The strength of factor analysis is that it provides a basis for data reduction. Rather than looking at all twenty items, just three factors can be examined. That simplifies the rankings and clarifies the most important items. Rather than focusing on individual items, practitioners and researchers can focus on the broad set of items represented by the essential factors.

FUTURE TRENDS

Many companies around the world are following the trend toward making large investments in implementing ERP systems. Several approaches and methodologies of ERP project implementation recognize a series of critical factors that must be carefully considered to ensure successful implementation of an ERP system project. In essence, there are dominant critical factors hypothesized to play a more overriding role in the implementation of ERP project and, they should be ongoing throughout all implementation levels. Clearly, the dominant factors are the ones that will shape the overall project culture, and subsequently the organizational culture as ERP is far reaching in nature.

Post-ERP activity seems to follow a clear path. A Deloitte Consulting study of 62 companies segments post-ERP activity into three stages. The first stage entails a three- to nine-month productivity decline, which is overcome by redefining jobs, establishing new procedures, fine-tuning ERP software, and taking charge of the new streams of information created by the platform. The second stage, which lasts from six to 18 months, involves skills development, structural changes, process integration, and add-on technologies that expand ERP functionality. The third stage, of one to two years' duration, is one of transformation, where the synergies of people, processes, and technology reach a peak. Another set of treds in ERP systems are as follows:

a. More ERP systems are going to be browser based, this makes it hardware and operating system independant.

b. ERP systems area turning more towards open source and they are moving to ASP model or hosted solutions.
c. ERP systems prices are coming down so the smaller companies can take advantages of the features and functionality of a integrated ERP solution.

Finally and perhaps most important, ERP forces discipline and organization around processes, making the alignment of IT and business goals more likely in the post-ERP era.

CONCLUSION

Despite the benefits that can be achieved, there is already evidence of failure in projects related to ERP systems implementation (Davenport, 1998). It is therefore important to find out what are the CSFs that drive ERP systems project success.

According to the respondents in this study, the six top CFSs for ERP systems implementation in Venezuelans firms are: top management support, presence of a champion, project management, best people full time, effective communications, and management of expectations.

This research has derived three composite CSFs in ERP systems implementation:

1. ERP implementation management
2. Users aptitudes and communication
3. Technical knowledge

Four of the six top individual items contribute to the composite factor of ERP systems implementation management, which is by far the most important. It is made up of items that are concerned with the management of ERP systems implementation projects.

The results of ranking ERP systems critical success factors in this study are largely consistent with the literature review, though the relative ranking of some factors varies. In the literature, top management support, change management, presence of a champion and management of expectations are the four most often cited critical factors.

A majority of factors—twelve—were rated as critical (rating > 4). Only one factor, business process reengineering, which is rated as critical in most articles reviewed, was not rated as critical in this study (rating < 3.0). Hence, the perceptions on CSFs of Venezuelan managers involved in ERP systems implementation projects are largely consistent with the findings reported in the literature.

There are a number of questions still to be determined. For example, although this paper establishes the relative importance of CSFs in seven firms, it has not established the reasons. Future studies could look at differences by size of firms, by industry type, by number of locations, by number of customers etc.

Finally, it should be noted that all CSFs are based on previous research; the success items can be modified when further research is conducted. For instance, discovery-oriented research through comprehensive interviews with several top-level managers in an attempt to identify new CSFs is an option.

REFERENCES

Al-Mudimigh, A., Zairi, M., & Al-Mashari, M. (2001). ERP software implementation: An integrative framework. *European Journal of Information Systems, 10,* 216-226.

Basu, C., & Palvia, P. (1999). Towards developing a model for global business process reengineering. In *Proceedings of the Americas Conference on Information Systems,* 283-285.

Bernroider, E., & Koch, S. (1999). Decision making for ERP-investments from the perspective of organizational impact: Preliminary results from an empirical study. In *Proceedings of the Americas Conference on Information Systems* (pp. 773-775).

Bingi, P., Sharma, M. K., & Godla, J. (1999). Critical issues affecting an ERP implementation. *Information Systems Management, 16*(2), 7-14.

Chatfield, C., & Collins, A. J. (1992). *Introduction to Multi-Variate Analysis*, (3rd ed.). London: Chapmanand Hall.

Chan, S. (1999). Architecture choices for ERP systems. In *Proceedings of the Americas Conference on Information Systems* (pp. 210-212).

Chatfield, A., & Andersen, K. (1998). Playing with LEGO: IT, coordination and global supply management in a world leader toy manufacturing enterprise. In *Proceedings of the Sixth European Conference on Information Systems,* (pp. 1680-1687).

Colmenares, L. (2005). Un Estudio Exploratorio sobre los Factores Críticos de Éxito en la Implantación de Sistemas de Planeación de Recursos Empresariales (ERP) en Venezuela. *Revista de Gestão da Tecnologia e Sistemas de Informação, 2*(2), 167-187.

Davenport, T. H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review,* July-August, 121-131.

Esteves, J., & Pastor, J. (2001). Analysis of critical success factors relevance along SAP implementation phases. In *Proceedings of the 7th Americas Conference on Information Systems* (pp. 1019-1025). Boston.

Falkowski, G., Pedigo, P., Smith, B., & Swanson, D. (1998). A recipe for ERP success. *Beyond Computing*, September, 44-45.

Holland, C. P., & Light, B. (1999). A critical success factors model for ERP implementation. *IEEE Software, 16*, 30-36.

Hong, K., & Kim, Y. (2002). The critical success factors for ERP Implementation: An organizational fit perspective. *Information & Management, 40*, 25-40.

Johnston, S. (2002). ERP: Payoffs and pitfalls. *HBS Working Knowledge*. October. 14-21.

Kaiser, H.F. (1974). An index of factorial simplicity. *Psychometrika*, (39), 31-62.

Kim, J., & Mueller, C.W. (1978). *Factor analysis: Statistical methods and practical issues*. Beverly Hills, CA: Sage Publications.

Kim, Y., Lee, Z., & Gozain. S. (2005). Impediments to successful ERP implementation process. *Business Process Management Journal, 11*(2), 158-170.

Klaus, H., Rosemann, M., & Gable, G. (2000) What is ERP? *Information Systems Frontiers, 2*, 141-162.

Kyung-Kwon, H., & Young-Gul, K. (2002). The critical success factors for ERP implementation: An organizational fit prespective. *Information and Management, 40*, 25-40.

Martin, M.H. (1998). An ERP Strategy. *Fortune*. February, 95-97.

Nah, F. F.-H., Lau, J. L.-S., & Kuang, J. (2001). Critical factors for successful implementation of enterprise systems. *Business Process Management, 7*(3), 285-296.

Nunnally, J. C. (1987). *Psychometric theory*. Englewood Cliffs, NJ: Prentice-Hall.

Pawlowski, S., & Boudreau, M. (1999). Constraints and flexibility in enterprise systems: Dialectic of system and job. In *Proceedings of the Americas Conference on Information Systems* (pp. 791-793).

Perez, M., & Rojas, T. (1999). SAP, change management and process development effectiveness (II): Case study. In *Proceedings of the Americas Conference on Information Systems (AMCIS)* (pp. 764-766).

Rai, A., Borah, S., & Ramaprasad, A. (1996). Critical success factors for strategic alliances in the information technology industry: An empirical study. *Decision Sciences, 27*(1), 141-155.

Rockart, J. (1979). Chief executives define their own data needs. *Harvard Business Review, 57*(2), 238-241.

Rosario, J. G. (2000). On the leading edge: Critical success factors in ERP implementation projects. *Business World*, May, 27-32.

Sato, R. (2000). Quick iterative process prototyping: A bridge over the gap between ERP and business process Reengineering. In *Proceedings of the Fourth Pacific Asia Conference on Information Systems* (pp. 16-25).

Scott, J. (1999). The FoxMeyer drugs bankruptcy: Was it a failure of ERP? In *Proceedings of the Americas Conference on Information Systems* (ACIS) (pp. 223-225).

Shanks, G., & Parr, A. (2000). Differences in critical success factors in ERP systems implementation in Australia and China: A cultural analysis. In *Proceedings of the Eighth European Conference on Information Systems* (pp. 1-8).

Sieber, M., & Nah, F. (1999). A recurring improvisational methodology for change management in ERP Implementation. In *Proceedings of the Americas Conference on Information Systems* (pp. 797-799).

Somers, T., & Nelson, K. (2004). A taxonomy of players and activities across the ERP project life cycle. *Information & Management, 41*, 257-278.

Slooten, L., & Yap, K. (1999). Implementing ERP information systems using SAP. In *Proceedings of the America Conference on Information Systems,* (pp. 226-228).

Smethurst, J., & Kawalek, P. (1999). Structured methodology usage in ERP implementation projects: An empirical Investigation. In *Proceedings of the Americas Conference on Information Systems* (pp. 219-221).

Soh, C., Kien, K., & Tay-Yap, J. (2000). Cultural fit and misfit: Is ERP auniversal solution? *Communication of the ACM,* 43, 47-51.

Srinivason, A. (1985). Alternative measures of system effectiveness: Associations and implications. *MIS Quarterly, 9*(3), 243-53.

Stefanou, C. J. (1999). Supply Chain Management (SCM) and organizational key factors for successful implementation of enterprise resource planning (ERP) systems. In *Proceedings of 5th Americas Conference on Information Systems*, 800-802.

Stratman, J., & Roth, A. (1999). Enterprise resource planning competence: A model, propositions and pre-test, design-stage scale development. In *Proceedings of 30th Decision Science Institute* (pp. 1199-201).

Sum, C.C., Ang, J. S. K., & Yeo, L. N. (1997). Contextual elements of critical success factors in MRP implementation. *Production and Inventory Management Journal, 3*, 77-83.

Sumner, M. (1999). Critical success factors in enterprise wide information management systems projects. In *Proceedings of 5th Americas Conference on Information Systems,* (pp. 232-234).

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics*. Harper Collins: London.

Tsai, W-H., Chien, S-W., Hsu, P-Y., & Leu, J-D. (2005). Identification of critical failure factors in the implementation of enterprise resource planning (ERP) system in Taiwan's industries. *International Journal of Management and Enterprise Development, 2*(2), 219-239.

Upton, D., & McAfree, A. (1997) Vandelay Industries, Inc. *Harvard Business School Publishing #9-697-037*, (pp. 1-16).

Volkoff, O. (1999). Using the structural model of technology to analyze an ERP Implementation. In *Proceedings of the Americas Conference on Information Systems,* (pp. 235-237).

Wee, S. (2002). *Juggling toward ERP success: Keep key success factors high.* ERP News. Retrieved January 22, 2002, from http://www.erpnews.com/erpnews/erp904/02get.html

Yingjie, J. (2005). *Critical success factors in ERP implementation in Finland*. Unpublished Master Thesis. The Swedish School of Economics and Business Administration, p. 71.

Zairi, M., Al-Mudimigh, A., & Jarrar, Y. (2000). ERP Implementation Critical Success Factors—The role and impact of business process management. In *Proceedings of the 2000 IEEE International Conference on Management of Innovation and Technology,* (pp. 122 –127).

Zhan, L., Lee, M., Zhang, Z., & Banerjee, P. (2003). Critical success factors of enterprise resource planning systems implementation success in China. In *Proceedings of the 36th Hawaii International Conference on System Sciences,* (pp. 562-567).

KEY TERMS

Business Process Reengineering (BPR): Any radical change in the way in which an organization performs its business activities. BPR involves a fundamental re-think of the business processes followed by a redesign of business activities to enhance all or most of its critical measures—costs, quality of service, staff dynamics, etc.

Change Management: Change management is the process of developing a planned approach to change in an organization. Typically the objective is to maximize the collective benefits for all people involved in the change and minimize the risk of failure of implementing the change.

Critical Success Factors: (CSF) indicates the few key areas of activity in which favorable results are absolutely necessary for the manager to succeed.

Enterprise Resource Planning System: Business software package for running every aspect of a company including managing orders, inventory, accounting, and logistics. Well known ERP software providers include BAAN, Oracle, PeopleSoft and SAP, collectively known to industry insiders as the "BOPS."

Factor Analysis: Any of several methods for reducing co-relational data to a smaller number of dimensions or factors; beginning with a correlation matrix a small number of components or factors are extracted that are regarded as the basic variable that account for the interrelations observed in the data.

Project Champion: A member of an organization who creates, defines or adopts a new technological innovation and who is willing to risk his or her position and prestige to make possible the innovation's successful implementation.

Project Management: Is the ensemble of activities concerned with successfully achieving a set of goals. This includes planning, scheduling and maintaining progress of the activities that comprise the project.

Software Package: Written programs or procedures or rules and associated documentation pertaining to the operation of a computer system and that are stored in read/write memory.

Systems Implementation: Customization or parameterization and adaptation of the software application according to the needs of the organization.

Assessing ERP Risks and Rewards

Joseph Bradley
University of Idaho, USA

INTRODUCTION

Enterprise resource planning (ERP) systems claim to meet the information needs of organizations. These off-the-shelf software packages replace hard to maintain solutions created by IS departments or older off-the-shelf packages that often provided only piecemeal solutions to an organization's information needs. ERP systems evolved from material requirements planning (MRP) systems and manufacturing resources planning (MRP II) systems. ERP serves the entire enterprise, not just manufacturing and inventory control as with its predecessors. ERP integrates information for the entire organization in a single database. But ERP implementations are often complex and experience serious problems. Failures, abandoned projects, and general dissatisfaction have been well publicized in the business press. ERP systems are "expensive and difficult to implement, often imposing their own logic on a company's strategy and existing culture" (Pozzebon, 2000, p. 1015).

BACKGROUND

Three characteristics distinguish ERP implementations from other IT projects (Somers, Ragowsky, Nelson, & Stern, 2001).

- ERP systems are "profoundly complex pieces of software, and installing them requires large investments of money, time and expertise (Davenport, 1998, p. 122).
- The packages may require changes in business processes and procedure, induce customization, and leave the implementing firm dependent on a vendor for support and updates (Lucas, Walton, & Ginsberg, 1988).
- The adopting firm is usually required to reengineer its business processes. As a result, the project must be managed as a broad program of organizational change rather than a software implementation (Markus & Tanis, 2000; Somers et al., 2001).

Despite these risks, global firms annually spent $10 billion on ERP software and another $10 billion on consultants to implement the systems in the late 1990s (Davenport, 1998). An AMR study estimated 2001 firm spending on ERP systems at $47 billion (Cotteleer, 2002). CIOs identified ERP as a leading application and technology development in a 2005 survey (Luftman, Kempaiah, & Nash, 2006).

This article will discuss the benefits firms expect to realize by adopting ERP systems, why some firms do not adopt these systems, risks associated with ERP implementation, some well-publicized ERP failures, risk management tools, and future trends in ERP implementation.

WHY DO FIRMS ADOPT ERP?

Firms adopt ERP for technical and business reasons. The technical reasons include reducing systems operating costs, solving specific problems such as Y2K, accommodating increased system capacity, and solving maintenance problems with legacy systems. Business reasons may include presenting a single face to the customer, quoting realistic delivery times, accommodating business growth, improvement of business processes, standardization of data, reduction of inventory carrying costs, and elimination of delays in filling orders (Markus & Tanis, 2000).

The rapid growth of the commercial market for ERP is attributed to the following factors (Watson & Schneider, 1999):

- Use of the popular client/server platform
- Can be used as an enabler for reengineering projects
- Y2K compliant
- Marketed to CEOs and CFOs as "strategic solutions" rather than as transaction processing software
- A way to outsource a significant part of the IS function (Watson & Schneider, 1999)

Advantages of ERP systems include (Rashid, Hossain, & Patrick, 2002):

- Reliable information access by using a single database
- Avoiding multiple data entries, reducing cost, and improving accuracy
- Delivery and cycle time reduction minimizing delays in reporting
- Cost reduction including time saving and improved controls
- Easy adaptability with business process options based on best practices easy to adapt

- Improved scalability
- Improved maintenance with long-term vendor contracts
- Global outreach with extensions to modules such as CRM and SCM
- E-commerce and e-business capabilities

An example of a decision to adopt an ERP system is provided by Geneva Pharmaceuticals, a manufacturer of generic drugs. Faced with eroding margins and continuing price pressure, the existing systems were proving inadequate. Data shared across business units had to be re-keyed, resulting in frequent errors. Data were locked in "functional silos" and did not support new processes. Geneva adopted ERP to solve the following problems:

- "implement best practices in business processes,
- integrate data across business units (hence reduce re-keying and maintenance costs),
- enforce data standardization (to reduce software maintenance costs),
- integrate well with new technologies or systems of acquired companies,
- provide scalability with growing product and customer base, and be Y2K (year 2000) compliant." (Bhattacherjee, 2000, p. 12)

A survey of Fortune 1000 firms identified organizational changes following ERP implementations, including (Jones & Young, 2006):

- Greater collaboration among functional areas in divisions
- Reorganization of processes
- Greater integration of processes across the organization
- Reduced silo behavior within divisions of the organization
- Reduced costs of operations
- Reduced silo behaviors across the organization
- Greater collaboration across divisions of the organization
- Greater integration of processes within divisions.

In addition, Jones and Young (2006) found people have a better view of the big picture, utilize more teamwork, and are more receptive to change.

With the identification of the prospective benefits of ERP, why have some firms not adopted ERP?

WHY DO FIRMS NOT ADOPT ERP?

Markus and Tanis (2000) identified three broad categories of reasons why firms that otherwise have all or some of the reasons to adopt ERP systems, do not adopt it, or only adopt ERP in part. These firms may adopt only certain modules and rely on legacy systems or new custom systems for their needs. Other firms may begin an implementation only to discontinue it for a variety of reasons. The reasons are:

1. Lack of feature-function fit
2. Company growth, strategic flexibility, and decentralized decision-making
3. Availability of alternatives to increase systems integration

Lack of feature-function fit may be due to the design of most ERP for discrete manufacturing. Many companies have specialized processes common to their industry, which may not be solved by the best practices embedded in ERP systems. The various modules may not fully support process manufacturing industries such as food processing and paper manufacturing, project industries such as aerospace, or industries that manufacture products with dimensionality such as clothing or footwear (Markus & Tanis, 2000). As the ERP market becomes saturated, vendors are designing packages for industries that were previously viewed as too complex.

Companies concerned with maintaining rapid growth rates, those needing strategic flexibility, and those without a top down decision-making style may be non-adopters or partial adopters of ERP systems. Dell Computer Corp. planned full implementation of SAP R/3 but discontinued the implementation after installing the human resource module. Dell's CIO expressed concern with the software's ability to keep pace with Dell's extraordinary growth rate. Visio, a software company subsequently acquired by Microsoft, expressed concern with the ability of SAP to handle the frequent changes it required to its sales analysis and commission requirements (Markus & Tanis, 2000, p. 29).

The experiences of Dell and Visio focus on the need for efficiency and flexibility in dealing with the external environment and internal processes. In a stable environment, mechanistic structures are appropriate consisting of "high degrees of standardization, formalization, specialization and hierarchy" (Newell, Huang, Galliers, & Pan, 2003). In a dynamic environment, organic structures are needed to enable organizations to be flexible to change products, processes, and structures. In these organizations, low levels of standardization, formalization, specialization, and hierarchy are most appropriate. ERP may maximize organizational efficiency at the cost of flexibility (Newell et al., 2003). The result may be an inability to respond quickly to changes in the environment, reducing the firm's competitiveness.

Organizational culture may also be a factor in non-adoption or partial adoption of ERP systems. Kraft Foods Inc. was highly decentralized but slowly moving to a one-company philosophy. ERP was regarded as culturally inappropriate with this strategy (Markus & Tanis, 2000).

Lean enterprises succeed "as a growth strategy for increasing sales by trimming the company's product delivery system into a competitive weapon" (Bradford & Mayfield, 2001, p. 30). Lean enterprises have difficulty using ERP systems due to the lack of flexibility. "ERP creates many non-value-added transactions by making companies track every activity and material price in the factory. This is counter to Lean philosophy, which aims at speeding up and smoothing production" (Bradford & Mayfield, 2001, p. 30).

Alternatives to ERP systems include data warehousing technologies that integrate data from source systems for query and analysis. These systems, sometimes described as "poor man's ERP," are limited by the quality of the underlying source systems (Markus & Tanis, 2000). In 1993 Great Atlantic & Pacific Tea Company, Inc. completed a supply chain and business process infrastructure based on a "robust data warehousing capacity for category management and other grocery-specific functionality" (Retek, 2003).

Other problems identified with implementation of ERP include time, expense, vendor dependence, and complexity.

RISKS ASSOCIATED WITH ERP IMPLEMENTATION

Implementing ERP can be a risky proposition for firms. Brown and Vessey (2003) observe, "Although failures to deliver projects on time and within budget were an old IT story, enterprise systems held even higher risks—they could be a 'bet-our-company' type of failure" (p. 65)

Markus (2000) propose 10 categories of IT related risks, all of which would apply to ERP systems:

1. Financial risk
2. Technical risk
3. Project risk
4. Political risk
5. Contingency risk
6. Non-use, under use, misuse risk
7. Internal abuse
8. External risk
9. Competitive risk
10. Reputational risks

"IT-related risk includes anything related to IT that could have significant negative effects on the business or its environment from the perspective of an executive investing in IT" (Markus 2000).

Some firms may be averse to the risks an ERP implementation can create. Scott (2003) discusses some of the risks identified by Markus and Tanis (2000). He describes project risks, information systems risks, organizational risks, and external risks in ERP implementations.

Project risks stem from the customization of purchased packages and the difficulty of interfacing with legacy systems. When firms believe their business processes are unique, they may customize ERP software instead of adopting best practices imbedded in a standard implementation. Data conversion can also be a problem when firms do not clean up their data before embarking on a project. After implementing SAP, Halliburton reported that inventory accuracy stood at less than 80% (Anderson, 2003). Project leadership, limiting project scope, avoiding customization, and a phased implementation (rollout) can minimize this risk (Scott, 2003).

Information systems risks arise from system performance problems. ERP systems may be poorly configured, or the hardware may need upgrading. Another risk arises when the use of multiple vendors creates the need for multiple interfaces. Multiple vendors contributed to the problems in the Hershey Food Corporation implementation. Information systems risks can be minimized by avoiding customization, use of data warehousing for reports and queries, and avoiding multivendor implementations (Scott, 2003).

Organizational risks of a bad ERP implementation can impact the firm's operating profits. Customer deliveries can be delayed, putting customer relationships at risk. Impacts can be with customers, financial performance, or internal business objectives. Organizational risks can be minimized with training and strong leadership, which assures that sufficient resources are allocated to the project and inspires employees who may resist the implementation (Scott, 2003).

External risk centers on litigation associated with the implementation. Firms with implementation problems may sue consultants and/or ERP vendors. Over-billing by consultants and use of incompetent trainees have been sources of litigation (Scott, 2003). Gore-Tex claims its consultant promised expert staff and delivered incompetent trainees. Managing consultants by specifying goals and individual competence of consultants can minimize this risk (MacDonald, 1999).

Political risk occurs "if a dominant coalition attempts to use the ERP package as a means by which to impose its views on other functional areas" (O'Gorman, 2004, p. 25). A case study at an international oil supplies company where the ERP implementation was dominated by the financial management of the business left the supply chain management function without the tools they believed they needed (Bradley, 2005).

A survey of members of the Chinese Enterprise Resource Planning Society identified the top 10 ERP risk factors as follows (Huang, Chang, Li, & Lin, 2004):

- Lack of senior management commitment to the project
- Ineffective communication with users
- Insufficient training of end users
- Fail to get user support

- Lack of effective project management methodology
- Attempting to build bridges to legacy applications
- Conflicts between user departments
- The composition of project team members
- Fail to redesign business processes
- Unclear/misunderstanding changing requirements

Competitive Advantage

Another reason for non-adoption may be that a standard software package available to all potential purchasers may reduce a firm's competitive advantage. A resource-based view of the firm assumes that the individual firm's unique collection of resources and capabilities is a potential source of competitive advantage. Capabilities leading to competitive advantage may be embedded in current business processes.

To create competitive advantage, such capabilities must be valuable, rare, costly to imitate, and non-substitutable. "Valuable and rare organizational resource[s] can only be sources of sustained competitive advantage if firms that do not possess these resources cannot readily obtain them" (Barney, 1991, p. 107). An off-the-shelf package may be costly, but would not be rare or costly to imitate. Adoption of ERP packages based on "best practices" may cause the loss of the unique and valuable advantage imbedded in current business processes. A case study showed that:

... the introduction of SAP-specific business routines can threaten established core, enabling and supplemental capabilities and related knowledge sets. The integration of SAP's embedded business routines and reporting functionality contributed to the creation of (a) highly rigid reporting structures; (b) inflexible managerial decision-making routines; and (c) reduced autonomy on the factory floor ... (Butler & Pyke, 2004, pp. 167-168)

WELL-PUBLICIZED FAILURES AND PROBLEMS

Numerous descriptions of ERP failures have appeared in the business press. The experience of serious problems at many well-run, well-financed firms may be enough to discourage some firms from beginning an ERP implementation.

Hershey Foods embarked on an ERP investment in mid-1996 to solve its Y2K problems and improve its ability to perform just-in-time store deliveries to its customers (Severance & Passino, 2002). After spending $112 million on an ERP project, Hershey Foods Corporation was unable to fill Halloween candy orders in October 1999, resulting in a 19% drop in third quarter profits (Stedman, 1999). As a result of Hershey's problems, its stock price fell by a third and the firm lost market share to Mars and Nestle (Severance & Passino, 2002).

A study by the PA Consulting Group found that "92% of companies are dissatisfied with results achieved to date from their ERP implementation and only 8% achieved a positive improvement in their performance" (ERP Implementation Disappoints Companies, 2000).

Davenport (1998) identifies several unsuccessful implementation efforts:

- Fox-Meyer Drug claims that an ERP system led to its bankruptcy.
- Mobil Europe spent hundreds of millions on ERP, but abandoned the project when a merger partner objected.
- Dell found that its ERP system did not support its new decentralized management style.
- Applied Materials gave up on its ERP implementation when it became overwhelmed with the organizational changes it required.

ERP success stories receive much less publicity.

MANAGING RISKS

ERP vendors have tried to overcome the concern of potential clients by developing tools to assess and manage risks associated with ERP implementation. Risk management as a process should (Zafiropoulos, Metaxiotis, & Askounis, 2005):

- Identify the context and criteria of the risk
- Identify the risks
- Determine the significance of each risk
- Identify, select, and implement risk management options
- Monitor and review the corrective options

These risk management systems allow project managers to make more realistic cost and time estimates to help avoid problems during implementation. Zafiropoulos et al. (2005) found that the use of a generic risk management tool they developed provided a structured way to assess risks, better integrated the use of consultants, improved communications between consultants and the project manager, led to more realistic time planning, and reduced the impact of problems when they occurred since the problems were expected and part of the project planning.

FUTURE TRENDS

Recently, the Gartner Group coined the term "ERP II" to describe a shift in ERP from an enterprise information base to moving information across the supply chain (Tak-

ing the Pulse of ERP, 2001). ERP II includes applications such as customer relationship management (CRM), supply chain management (SCM), and e-business. New risks and challenges will be faced by organizations opening up information systems to supply chain partners. Supply chain partners could be potential competitors or pass information to existing competitors. Resolving trust issues will be key to advancement in ERP II. Luftman et al. (2006) expect that the line between ERP, CRM, and SCM will blur as ERP vendors expand the functionality of their products and redefine their packages.

Davenport and Brooks (2004) discuss the need for both infrastructural and strategic capabilities in the organization. ERP or Enterprise Systems provide core functionality but are expensive and time consuming to implement. "These infrastructural capabilities provide very little in the way of real business value" (p. 13). These systems do not provide "short-term cost savings or other competitive advantage" (p. 13). In contrast, SCM applications provide "strategic, competitively-oriented capabilities," (p. 13) which can reduce inventories and improve customer service. Firms beginning to build infrastructure with ERP systems may run out of time and money, neglecting the strategic supply chain management systems (Davenport & Brooks, 2004).

Small and mid-sized enterprises (SME) will become leading adopters of ERP systems as the large company market becomes saturated. Vendors have been developing less expensive versions of their software to appeal to the SME market.

CONCLUSION

ERP implementation projects continue to present risks to adopting organizations. Continuing ERP spending demonstrates that most organizations have concluded that the benefits resulting from such implementations outweigh the substantial risks and cost of ERP systems. Perhaps the risks of not adopting ERP are determined to be greater than the risks faced by adopters. The prospect of extending ERP beyond organizational boundaries to supply chain partners makes ERP even more attractive and possibly more risky. Organizations will continue to adopt ERP as a strategic necessity to remain competitive in their industry, but few, if any, will gain any sustainable competitive advantage by adopting ERP.

REFERENCES

Anderson, A. (2003). *When closeness counts*. Retrieved December 23, 2003, from http://www.datasweep.com/ds/2003/article_2003.asp?page_id=newsln_002print

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of management, 17*(1), 99-120.

Bhattacherjee, A. (2000). Beginning SAP R/3 implementation at Geneva Pharmaceuticals. *Communications of the Association for Information Systems, 4*(2).

Bradford, M., & Mayfield, T. (2001). Does ERP fit in a LEAN world? *Strategic Finance, 82*(11), 28-34.

Bradley, J. (2005). Are all critical success factors created equal? In *Proceedings of the 11th Americas' Conference on Information Systems,* Atlanta (pp. 2152-2159). Omaha, NE: Association for Information Systems.

Brown, C. V., & Vessey, I. (2003). Managing the next wave of enterprise systems: Leveraging lessons from ERP. *MIS Quarterly Executive, 2*(1), 65-77.

Butler, T. & Pyke, A. (2004). Examining the influence of ERP systems on firm-specific knowledge assets and capabilities. In F. Adam & D. Sammon (Eds.), The enterprise resource planning decade: Lessons learned and issues for the future (pp. 167-206), Hershey, PA: Idea Group Publishing.

Cotteleer, M. J. (2002). *An empirical study of operational performance convergence following enterprise-IT implementation* (Working Paper No. 03-011). Harvard Business School.

Davenport, T. H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review, 76*(4, July-August), 121-131.

Davenport, T. H., & Brooks, J. D. (2004). Enterprise systems and the supply chain. *Journal of Enterprise Information Management, 17*(1), 8-19.

ERP implementation disappoints companies. (2000, August 31). *Australian Banking & Finance, 9,* 8.

Hall, D., & Hulett, D. (2002). *Universal risk project: Final report, February 2002*. Milford, NH: PMI Risk SIG.

Huang, S-M., Chang, I-C., Li, S-H., & Lin, M-T. (2004). Assessing risk in ERP projects: Identify and prioritize the factors. *Industrial Management & Data Systems, 104*(8/9), 681-688.

Jones, M. C., & Young, R. (2006). ERP usage in practice. *Information Resources Management Journal, 19*(1), 23-42.

Lucas, H. C. Jr., Walton, E. J., & Ginsberg, M. J. (1988). Implementing packaged software. *MIS Quarterly, 12*(4), 537-549.

Luftman, J., Kempaiah, R., & Nash, E. (2006). Key issues for IT executives 2005. *MIS Quarterly Executive, 5*(2), 81-99.

MacDonald, E. (1999, Nov. 2). W. L. Gore alleges PeopleSoft, Deloitte botched a costly software installation. *The Wall Street Journal,* p. 14.

Markus, M. L. (2000). Toward an integrative theory of risk control. In R. Baskerville, J. Stage, & J. I. DeGross (Eds.), Organizational and social perspectives on information technology (pp. 167-178). Boston: Kluywer Academic Publishers.

Newell, S., Huang, J. C., Galliers, R. D., & Pan, S. L. (2003). Implementing enterprise resource planning and knowledge management systems in tandem: Fostering efficiency and innovation complementarity. *Information and Organization, 13*(1), 25-52.

O'Gorman, B. (2004). The road to ERP: has industry learned or revolved back to the start? In F. Adams & D. Sammon (Eds.), *The enterprise resource planning decade: Lessons learned and issues for the future* (pp. 22-46). Hershey, PA: Idea Group Publishing.

Pozzebon, M. (2000). Combining a structuration approach with a behavioral-based model to investigate ERP usage. In *Proceedings of the Sixth Americas' cnference on Information Systems,* Long Beach, CA (pp. 1015-1021). Atlanta: Association for Information Systems.

Rashid, M. A., Hossain, L., & Patrick, J. D. (2002). The evolution of ERP systems: A historical perspective. In F. F.-H. Nah (Ed.), *Enterprise resource planning solutions & management* (pp. 35-50). Hershey, PA: IRM Press.

Retek. (2003). *A&P completes supply chain/business process initiative: IBM and Retek deliver enterprise merchandising solutions*. Retrieved March 13, 2004, from http://www.retek.com/press/press.asp?id/id=507

Scott, J. (2003). What risks does an organization face from an ERP implementation? In D. R. Laube & R. F. Zammuto (Eds.), *Business driven information technology: Answers to 100 critical questions for every manager* (pp. 274-278). Stanford, CT: Stanford Business Books.

Severance, D. G., & Passino, J. (2002). *Making I/T work.* San Francisco: Jossey-Bass.

Somers, T. M., Ragowsky, A. A., Nelson, K. G., & Stern, M. (2001). *Exploring critical success factors across the enterprise systems experience cycle: An empirical study* (Working Paper). Detroit, MI: Wayne State University.

Stedman, C. (1999, November 1). Failed ERP gamble haunts Hershey: Candy maker bites off more than it can chew and 'Kisses' big Halloween sales goodbye. *Computer World,* p. 1.

Taking the Puse of ERP. (2001). *Modern Materials Handling, 56*(2), 44-51.

Watson, E. E., & Schneider, H. (1999). Using ERP in education. *Communications of the Association for Information Systems, 1*, Article 9, p. 1-46.

Zafiropoulos, I., Metaxiotis, K., & Askounis, D. (2005). Dynamic risk management systems for modeling, optimal adaption and implementation of an ERP system. *Information Management & Computer Security, 13*(2/3), 212-234.

KEY TERMS

Enterprise Resource Planning Systems (ERP): An off-the-shelf accounting-oriented information system that meets the information needs of most organizations. A complex and expensive information tool to meet the needs of an organization to procure, process, and deliver customer goods or services in a timely, predictable manner.

ERP II: ERP systems that extend beyond the enterprise level to exchange information with supply chain partners. Examples are CRM, SCM, and e-business.

IT-Related Risk: This risk includes "anything related to IT that could have significant negative effects on the business or its environment from the perspective of an executive investing in IT" (Markus, 2000).

Legacy Systems: Transaction processing systems designed to perform specific tasks. Systems that have become outdated as business needs change and the hardware and software available in the market place improve.

Manufacturing Resources Planning (MRPII): Extends MRP by addressing all resources in addition to inventory. MRPII links material requirements planning with capacity requirements planning avoiding over and under shop loading typical with MRP.

Material Requirements Planning (MRP) Systems: Processes that use bills of materials, inventory data, and a master productions schedule to time phase material requirement, releasing inventory purchases in a manner that reduces inventory investment yet meets customer requirements.

Risk Management: A system designed to avoid "problems during a project, which can lead to deviation from project goals, timetables, and cost estimations" (Zafiropoulos et al., 2005, p. 213).

Risks: "A risk is a future event that may or may not occur. The probability of the future event occurring must be greater than 0% and less than 100%. The consequences of the future event must be unexpected or unplanned for" (Hall & Hulett, 2002, p. 5).

Association Rules Mining for Retail Organizations

Ioannis N. Kouris
University of Patras, Greece

Christos Makris
University of Patras, Greece

Evangelos Theodoridis
University of Patras, Greece

Athanasios Tsakalidis
University of Patras, Greece

INTRODUCTION

In recent years, we have witnessed an explosive growth in the amount of data generated and stored from practically all possible fields (e.g., science, business, medicine, military just to name a few). However, the ability to store more and more data has not been followed by the same rate of growth in the processing power, and, therefore, much of the data accumulated remains today still unanalyzed. Data mining, which could be defined as the process concerned with applying computational techniques (i.e., algorithms implemented as computer programs) to actually find patterns in the data, tries to bridge this gap. Among others, data mining technologies include association rule discovery, classification, clustering, summarization, regression and sequential pattern discovery (Adrians & Zantige, 1996; Chen, Han, & Yu, 1996; Fayyad, Piatetsky-Shapiro, & Smyth, 1996). This problem has been motivated by applications known as market basket analysis which find items purchased by customers; that is, what kinds of products tend to be purchased together (Agrawal, Imielinski, & Swami, 1993).

BACKGROUND

The goal of the data mining task is to find all frequent itemsets above a user specified threshold (called *support*) and to generate all association rules above another threshold (called *confidence*) using these frequent itemsets as input. This type of information could be used for catalogue design, store layout, product placement, target marketing, and so forth. The prototypical application of this task has been the market basket analysis, but the specific model is not limited to it since it can be applied to many other domains (e.g., text documents [Holt & Chung, 2001], census data, [Brin et al., 1997], telecommunication data and even medical images, etc.). In fact, any data set consisting of "baskets" containing multiple "items" can fit this model. Many solutions have been proposed in the last years using a sequential or a parallel paradigm, experimenting on factors such as memory requirements, I/O scans, dimensionality reduction, and so forth.

The specific problem was first introduced by Agrawal et al. (1993) and an algorithm by the name AIS was proposed for effectively addressing it. Agrawal and Srikant (1994) have introduced a much more efficient solution and two new algorithms by the names Apriori and AprioriTid were proposed. Algorithm Apriori has been and still is the major reference point for all subsequent works. Most algorithms and approaches proposed thereafter (Toivonen, 1996; Brin et al., 1997; Park, Chen, & Yu, 1995; Han, Pei, & Yin, 2000) focus on either decreasing the number of passes made over the data or at improving the efficiency of those passes (i.e., by using additional methods for pruning the number of candidates that have to be counted). Among other things studied in association rule mining are: (1) incremental updating (Cheung, Han, Ng, & Wong, 1996; Lee, Lin &Chen, 2001), (2) mining of generalized and multi-level rules (Han & Fu, 1995; Srikant &Agrawal, 1995), (3) using multiple minimum supports (Liu, Hsu, & Ma, 1999), (4) mining of quantitative rules (Srikant & Agrawal, 1996), (5) parallel algorithms (Agrawal & Shafer, 1996; Park, Chen, & Yu, 1995).

Lately, it has been widely accepted that the model of association rules is either oversimplified or suffers from several omissions that cannot nowadays be considered insignificant especially in a retail environment. For example, treating the items as mere statistical probabilities and neglecting their real significance or handling them as Boolean variables and not taking into consideration their exact number of appearances in every transaction leads to fragmentary and dubious results (Liu, Hsu, & Ma, 1999). In this article we present a collection of works trying to solve all these problems as

well as to address new ones. The main technical contribution of these works is the technically challenging assignment of weight values to different items in a given sell-period independently from other items; we call this process association rules mining since it tries with this careful selection of weights to mine suitable association rules. The focus of all these works is retail data and organizations.

IDENTIFYING THE "HOT" ITEMS WITHOUT GETTING BURNED

The idea behind association rule mining is to search a considerable amount of data collected over a long period and to apply various techniques to discover some more or less unseen knowledge hidden inside the data. However, one of the biggest omissions of the approaches used up to now was that they discovered long existing relations and rather ignored the emerging ones. All approaches up to now followed rather than kept up with the sales or, more generally, the appearances of the itemsets. What we need is an approach that finds emerging trends in the bud along with the long established ones. This situation can be explained better in the example next.

Let's suppose that there exists a product that is sold in a retail store for many years, with moderate sales as compared to all other products and another product that just entered the market (e.g., is on sale about a year) but has tremendous sales. Applying the classical statistical model of association rule mining, where we simply measure the number of appearances of every itemset and if it is above a user specified threshold it is considered as frequent, would unavoidably doom the new product. A product that is on sale for so little can not practically come even near the sales of a product that is on sale for so long. So one must either wait for so long as for the new products to sum enough sales, which could well take months or even years, or find a way to effectively take them into consideration as soon as possible. The same situation can take place with products that were on the market but with very low sales and suddenly begun to present abnormally high sales because they are currently under heavy promotion, they suddenly became in fashion, some external circumstances or factors (e.g., weather conditions) promoted their sales, and so forth. The specific situation is very usual, especially at retail stores where the items on sale present such behaviors. After all, this kind of bursty sales behavior in a retail organization is probably far more interesting than that of high selling but stable products.

Therefore, the notion of "hot" items has been introduced by Kouris, Makris, and Tsakalidis (2004b), where as "hot" is considered any item that presents very high sales in a certain period of time. More formally, for every 1-itemset the, so called, *interest ratio* is calculated, which is defined as the number of sales of an item in the last period of sales (i.e., from the last time the algorithm has run again) divided by the mean number of sales of all items in the same period.

$$\dot{r}_i = \frac{sales_i}{\overline{sales}}$$

Every item that has its interest ratio above a user defined threshold called minimum interest threshold is considered as "hot" for the specific period. Of course, if an item has a number of sales above the support threshold it is treated as a frequent item. In essence, the proposed algorithm searches for items that have sales below the support threshold, but their interest ratio is above the minimum interest threshold. The user has, of course, the option of giving to that threshold any value depending on the desired output.

A logical question could be what happens with an item that was "hot" in the previous period but is no longer "hot" or frequent in the next one. One option would be to treat these items as infrequent items in the new period since they are obviously no longer interesting for the users. Another one could be to give these items a grace period, and treat them as "hot", to see if they will come back to high sales. Either one is possible and acceptable and depends solely on the needs of the data miner. One, though, could claim that such items were wrongly considered as interesting since their subsequent trend showed that they are no longer "hot". Nevertheless, the algorithm managed to immediately identify the period that they became interesting, took them into consideration, and promoted their sales when they were actually very interesting and this was exactly the goal. If, on the other hand, a "hot" item becomes frequent in the next period then the algorithm managed to successfully predict its future performance early enough and to take it into consideration in advance rather than having to wait for it to actually become frequent.

ASSESSING THE IMPORTANCE OF ITEMS IN A RETAIL ORGANIZATION

In contrast to the assumption upon which all association rules approaches work, that is that the correct support value is already known, in practice the correct minimum support value is not known and can only be estimated based on domain knowledge or on previous experience. Also when talking about the "correct" value, this is judged after the completion of the mining process based on the number of discovered frequent itemsets (i.e., too few or too many itemsets as compared to what has been anticipated), or in other words through completely subjective and rather trivial criteria. Consequently, if the support threshold changes, then the mining has to be repeated from the beginning requiring

at least the same number of passes over the database as the previous run (if the new threshold is lower than before), and what's most important without being able to use knowledge from previous runs of the same algorithm. So, the whole process of assigning support values to all the items in a database has to be as accurate as possible in order to avoid unnecessary runs of the same algorithm over the data.

Association rules algorithms are very powerful tools but suffer from the drawback that they take no provision in the business value of items and the associations between them when determining the support of an item (Cabena, Hadjinian, Stadler, Verhess, & Zanasi, 1997), especially in retail organizations. As a consequence, the support assigned does not reflect the real value of the items, which in turn results in the homogenization of all items in any dataset. The specific problem was addressed in Kouris, Makris, and Tsakalidis (2004a), by using an efficient weighting scheme for assigning the correct minimum support values to all the items. The intuition was simple. Every item can have support at maximum: $max\ sup(i)=T_i/T_{total}$, where T_i is the number of transactions that an item appears in and T_{total} denotes the total number of transactions.

The closest the support of an item is set to its maximum support, the more appearances there are needed in order for it to quantify as a frequent and the less important it is considered. So the final support of an item is determined by multiplying the maximum support an item can have with a weight that reflects its importance according to the following formula: $\sup(i)=wf * \max \sup(i)$.

The value of *wf* reflects the answer to the following questions:

1. What are the net profits of an item compared to the average net profits from all itemsets in a dataset?
2. What are the total sales of an item compared to the average total sales from all itemsets?
3. What is the number of transactions an item appears in compared to the average number of transactions all items appear in?

These factors are combined in the following equation:

Equation 1

$$wf = \log_A\left(1+\frac{\overline{T}total}{T_i}+\frac{\overline{S}total}{S_i}+\frac{\overline{N}total}{N_i}\right)$$

where

$$A = \max\left(1+\frac{\overline{T}total}{T_i}+\frac{\overline{S}total}{S_i}+\frac{\overline{N}total}{N_i}\right)$$

where T_i is the number of transactions that an item apears in $\overline{T}_{total}$, is the average number of transactions all items appear in, S_i is the number of times an item appears in general in the database, $\overline{S}_{total}$ is the average number of appearances of all items, N_i is the net profit per item, and N_{total} is the average net profit from all items.

The main characteristic of the specific equation is that it takes into consideration all three factors, and an item must have all three factors small in order to be considered as very important. It is not enough for an item to appear in many transactions, or to appear many times in total, or to have large net profits in order to get a very low support value. It is the combination of all these three factors that make it most important. Also, in Equation 1, one is free to enter a different degree of influence to every factor he decides to. For example, if a company is more interested in increasing its market share (i.e., its actual sales) rather than its net profits, it could boost the effect of the factor

$$\frac{\overline{S}total}{S_i}$$

by multiplying it with a constant positive value below one.

The parameter *A* acts as a normalization factor in order to have the value of *wf* always less than 1 for every item. Note that altering the importance of various items can change the value of *A*, but this is desirable since *wf* estimates the relative importance of an item to the other items and not its absolute importance; hence, changes in other items' importance that could alter this value of *A* should be reflected to changes to the value of *wf*, for each item. Of course, the less important item would have *wf* value equal to 1.

Finally, it should be mentioned that the specific equation allows the introduction of any new factor a user considers important. Of course, the philosophy and the logic of this approach can be used with little modifications in all kinds of businesses and applications.

LOCALLY FREQUENT PRODUCTS

As interesting could be considered any item that should be taken into account when generating association rules, even if it fails to meet the minimum support or any other measure used for generating all frequent itemsets. Examples of criteria for interesting items could be the profit margin, the lifetime, the durability, and so forth. A large percentage of the products that are sold in a retail organization have a significant seasonal or more generally localistic behavior. This kind of products was the focus of the work of Bodon, Kouris, Makris, and Tsakalidis (2005), namely those items that are very frequent if considered in relation to a specific

part of the database, but their total number of appearances fails to exceed the minimum support value in the whole database. This part could correspond to sales made during a specific time period in a specific branch of an enterprise, and so forth. This case is very common and has extreme interest especially at large supermarkets or at retail stores where some products are on sale throughout the year whereas others only at a per season basis. Such products can reach support values of way over 60% at the specific periods with considerable profits, but the rest of the year they are not even at a shelf and, as a consequence, their overall support value becomes very low. So if one is to present a complete view of all rules concerning a dataset that truly reflects reality, these items must certainly be taken into account. This case arises frequently in many other situations too like, for example, with items bought at specific geographical areas only, or due to special circumstances, and so forth.

Essential to the approach is the distinction between the items that are frequent and the items that should be considered frequent in relation to a specific part of our database. An item is said to be locally frequent if its local frequency ratio is above a user specified threshold. As *local frequency ratio* (*LFR*) is defined to be the ratio of the number of occurrences of an item in a specific transactions interval divided by that interval.

LFR = Item interval count / Interval

If this ratio is above a user specified minimum, then this item is considered as a locally frequent item. The specific ratio depends heavily on the interval that is chosen to count the support of the items. In a dataset there may exist locally frequent itemsets of various orders (e.g., 1-itemsets, 2-itemsets, 3-itemsets, and so forth), and the user can define the order of itemsets in which he/she is interested.

FUTURE TRENDS

Concerning future trends, an interesting subject is the further employment of Markov models in the data mining arena (Giudici & Castelo, 2001, 2003). Another very interesting subject in the area right now is data mining for bioinformatics (Wang, Zaki, Toivonen, & Shasha, 2005). In Bioinformatics, data mining refers to finding and predicting motifs in sequences, to discover genetic mechanisms, to summarize clustering rules for multiple DNA or protein sequences and so on. Finally, program comprehension through data mining, is a promising area to most probably become the center of attention in the years to come; representative works can be found in Xiao and Tzerpos (2005), Chen, Tjortjis, and Layzell (2002), and Tjortjis and Layzel (2001).

CONCLUSION

Mining association rules especially from large databases of business data such as transactions records has been and probably will remain for quite a long the "hottest" topic in the area of data mining, mainly due to its large financial interest. In this article we have presented a collection of techniques and methodologies more oriented towards the practical side of the knowledge discovery process and especially for retail organizations and data. The main technical contribution of these techniques was the technically challenging association of weight values to different items; we call this process *association rules mining* since its purpose is to mine suitable association rules. These works manage to solve many ambiguities suffered from all previous methods and also describe and address novel problems in a way that is both efficient and effective.

REFERENCES

Adrians, P., & Zantige, D. (1996). *Data mining.* Addison-Wesley.

Agrawal, R., Imielinski, T., & Swami, A. (1993, May). Mining association rules between sets of items in large databases. In *Proceedings of ACM SIGMOD,* Washington, DC (pp. 207-216).

Agrawal, R., & Shafer, J.C. (1996). *Parallel mining of association rules: Design, implementation and experience.* Technical Report TJ10004, IBM Research Division, Almaden Research Center.

Agrawal, R., & Srikant, R. (1994, September). Fast algorithms for mining generalized association rules. In *Proceedings of VLDB,* Santiago, Chile (pp. 487-499).

Bodon, F., Kouris, I. N., Makris, C. H., & Tsakalidis, A. K. (2005). Automatic discovery of locally frequent itemsets in the presence of highly frequent itemsets. In *Intelligent Data Analysis Journal, 9*(1).

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997, May). Dynamic itemset counting and implication rules for market basket data. In *Proceedings of ACM SIGMOD,* Tucson, AZ (pp. 255-264).

Cabena, P., Hadjinian, P., Stadler, R., Verhess, J., & Zanasi, A. (1997). *Discovering data mining: from concept to implementation.* NJ: Prentice Hall.

Chen, M.S., Han, J., & Yu, P.S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866-883.

Chen, K., Tjortjis, C., & Layzell, P.J. (2002) A method for legacy systems maintenance by mining data extracted from source code. *Case studies of IEEE 6th European Conf. Software Maintenance and Reengineering (CSMR 02)* (pp. 54-60).

Cheung, D. W., Han, J., Ng, V., & Wong, C. Y. (1996, February). Maintenance of discovered association rules in large databases: An incremental updating technique. In *Proceedings of ICDE,* New Orleans, LA (pp. 106-114).

Dunham, M. H. (2003). *Data mining introductory and advanced topics*. Upper Saddle River, NJ: Prentice Hall.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Giudici, P., & Castelo, R. (2001). Association models for Web mining. *Journal of Data Mining and Knowledge Discovery, 24*(1), 39-57.

Giudici, P., & Castelo, R. (2003). Improving markov chain monte carlo model search for data mining. *Machine Learning, 50*(1/2).

Han, J., & Fu, Y. (1995). Discovery of multiple-level association rules from large databases. In *Proceedings of the VLDB,* Zurich, Switzerland (pp. 420-431).

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of ACM SIGMOD,* Dallas, TX.

Holt, J. D., & Chung, S. M. (2001). Multipass algorithms for mining association rules in text databases. *Knowledge and Information Systems, 3*(2), 168-183.

Kouris, I. N., Makris, C. H., & Tsakalidis, A. K. (2004a, December). Assessing the microeconomic facet of association rules via an efficient weighting scheme. In *Proceedings ICKM 2004* (pp. 13-15).

Kouris, I. N., Makris, C. H., & Tsakalidis, A. K. (2004b, April). Efficient automatic discovery of "hot" itemsets. In *Information Processing Letters, 90*(2), 65-72.

Lee, C.-H., Lin, C.-R., & Chen, M.-S. (2001, November). Sliding window filtering: An efficient algorithm for incremental mining. In *Proceedings of ACM CIKM.*

Liu, B., Hsu, W., & Ma, Y. (1999, August). Mining association rules with multiple minimum supports. In *Proceedings of ACM KDD,* San Diego, CA (pp. 337-34).

Park, J.-S., Chen, M.-S., & Yu, P. S. (1995, May). An effective hash based algorithm for mining association rules. In *Proceedings of the ACM SIGMOD,* San Jose, CA (pp. 175-186).

Park, J.-S., Chen, M.-S., & Yu, P. S. (1995, November). Efficient parallel data mining for association rules. In *Proceedings of CIKM,* Baltimore (pp. 31-36).

Srikant, R., & Agrawal, R. (1995, September). Mining generalized association rules, In *Proceedings of VLDB,* Zurich, Switzerland (pp. 407-419).

Srikant, R., & Agrawal, R. (1996, June). Mining quantitative association rules in large relational tables. In *Proceedings of ACM SIGMOD,* Montreal, Canada (pp. 1-12).

Tjortjis, C., & Layzel P. J. (2001). Using data mining to assess software reliability. *Suppl. Proc. IEEE 12th Int'l Symposium Software Reliability Engineering (ISSRE 01)* (pp. 221-223).

Toivonen, H. (1996, September). Sampling large databases for finding association rules. In *Proceedings of VLDB,* Mumbay, India (pp. 134-145).

Wang, J., Zaki, M., Toivonen, H., & Shasha, D. (2005). *Data mining in bioinformatics.* Springer Verlag.

Witten, I., Moffat, A., & Bell, T. (1999). *Managing gigabytes: Compressing and indexing documents and images* (2nd ed.). San Francisco: Morgan Kaufmann.

Xiao, C., & Tzerpos, V. (2005), Software clustering on dynamic dependencies. *Proc. IEEE 9th European Conf. Software Maintenance and Reengineering (CSMR 05).*

KEY TERMS

Association Rules Mining: Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, and other information repositories.

Confidence: A statistical measure of importance that denotes the strength of implication.

Data Clustering: Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters).

Data Mining: Analysis of data in a database using tools which look for trends or anomalies without knowledge of the meaning of the data. The nontrivial extraction of implicit, previously unknown, and potentially useful information from data. The science of extracting useful information from large data sets or databases.

"Hot" Itemsets: Itemsets that present abnormally high appearances as compared to the appearances of all other items but mainly as compared to their own previous number appearances.

Itemsets: A collection or a combination of items in a database.

Locally Frequent Itemsets: Itemsets that present a highly localized and condensed appearances behavior as compared to all other itemsets.

Support: A statistical measure of importance that indicates the frequencies of the occurring patterns in a rule.

Category: Software & Systems Design

Attribute Grammars and Their Applications

Krishnaprasad Thirunarayan
Wright State University, USA

INTRODUCTION

Attribute grammars are a framework for defining semantics of programming languages in a syntax-directed fashion. In this chapter, we define attribute grammars, and then illustrate their use for language definition, compiler generation, definite clause grammars, design and specification of algorithms, and so forth. Our goal is to emphasize its role as a tool for design, formal specification and implementation of practical systems, so our presentation is example rich.

BACKGROUND

The lexical structure and syntax of a language is normally defined using regular expressions and context-free grammars respectively (Aho, Lam, Sethi & Ullman et al., 2007,). Knuth (1968) introduced attribute grammars to specify static and dynamic semantics of a programming language in a syntax-directed way.

Let G = (N, T, P, S) be a context-free grammar for a language L (Aho et al., 2007). N is the set of non-terminals. T is the set of terminals. P is the set of productions. Each production is of the form A::= α, where A ∈ N and α ∈ (N U T)*. S ∈ N is the start symbol. An *attribute grammar* AG is a triple (G, A, AR), where G is a context-free grammar for the language, A associates each grammar symbol X ∈ N U T with a set of attributes, and AR associates each production R ∈ P with a set of attribute computation rules (Paakki, 1995). A(X), where X ∈ (N U T), can be further partitioned into two sets: synthesized attributes S(X) and inherited attributes I(X). AR(R), where R ∈ P, contains rules for computing inherited and synthesized attributes associated with the symbols in the production R.

Consider the following attribute grammar that maps bit strings to numbers. CFG = ({N},{0,1},P,N), where P is the left column of productions shown below. The number semantics is formalized by associating a synthesized attribute *val* with N, and providing rules for computing the value of the attribute *val* associated with the left-hand side N (denoted N_l) in terms of the value of the attribute *val* associated with the right-hand side N (denoted N_r), and the terminal.

N ::= 0	N.*val* = 0
N ::= 1	N.*val* = 1
N ::= N0	N_l.*val* = 2 * N_r.*val*
N ::= N1	N_l.*val* = 2 * N_r.*val* + 1

An attribute grammar involving only synthesized attributes is called an *S-attributed grammar*. It is straightforward to parse a binary string using this grammar and then compute the value of the string using a simple top-down left-to-right traversal of the abstract syntax tree.

The above attribute grammar is not unique. One can construct a different S-attributed grammar for the same language and the same semantics as follows.

N ::= 0	N.*val* = 0, N.*len* = 1
N ::= 1	N.*val* = 1, N.*len* = 1
N ::= 0N	N_l.*val* = N_r.*val*; N_l.*len* = N_r.len +1
N ::= 1N	N_l.*val* = 2^ N_r.*len* + N_r.*val*; N_l.*len* = N_r.len +1

Attribute grammars can be devised to specify different semantics associated with the same language. For instance, the bit string can be interpreted as a fraction by associating an inherited attribute *pow* to capture the left context—the length of the bit string between left of a non-terminal and the binary point, to determine the local value or the weight of the bit.

F ::= **.** N	F.*val* = N.*val*, N.*pow* = 1
N ::= 0	N.*val* = 0
N ::= 1	N.*val* = (1 / 2^ N.*pow*)
N ::= 0N	N_l.*val* = N_r.*val*; N_l.*pow* = N_r.pow +1
N ::= 1N	N_l.*val* = (1 / 2^ N.*pow*) + N_r.*val*; N_r.*pow* = N_l.pow +1

Each production is associated with attribute computation rules to compute the synthesized attribute *val* of the left-hand side non-terminal and the inherited attribute *pow* of the right-hand side non-terminal (we leave it as an exercise for the interested reader to devise an S-attributed grammar to capture this semantics).

APPLICATIONS OF ATTRIBUTE GRAMMARS

Attribute grammars provide a *modular framework* for formally specifying the semantics of a language based on its context-free grammar (or in practice, for conciseness, on its Extended Backus Naur formalism representation (Louden, 2003)). By *modular*, we emphasize its role in structuring specification that is incremental with respect to the produc-

tions. That is, it is possible to develop and understand the attribute computation rules one production at a time. By *framework*, we emphasize its role in structuring a specification, rather than conceptualizing the meaning. For instance, denotational semantics, axiomatic semantics, and operational semantics of a language can all be specified using the attribute grammar formalism by appropriately choosing the attributes (Louden, 2003). In practice, different programming languages interpret the same syntax/construct differently, and attribute grammars provide a framework for defining and analyzing subtle semantic differences.

In this section, we illustrate the uses and the subtleties associated with attribute grammars using examples of contemporary interest. Traditionally, attribute grammars have been used to specify various compiler activities formally. We show examples involving (i) type checking/inference (static semantics), (ii) code generation, and (iii) collecting distinct variables in a straight-line program. Attribute grammars can also be used to specify compiler generator activities. We show parser generator examples specifying the computation of (i) nullable non-terminals, (ii) first sets, and (iii) follow sets, in that sequence (Aho et al., 2007). Definite clause grammars enable attribute grammars satisfying certain restrictions to be viewed as executable specifications (Bratko, 2001). We exemplify this in SWI-Prolog. The essence of attribute grammars can be seen to underlie several database algorithms. We substantiate this by discussing the magic sets for optimizing bottom-up query processing engine. We also discuss how attribute grammars can be used for developing and specifying algorithms for information management (Thirunarayan, Berkovich, & Sokol, 2005).

Type Checking, Type Inference, and Code Generation

Consider a simple prefix expression language containing terminals {n, x, +}. The type of variable n is int, and the type of variable x is double. The binary arithmetic operation + returns an int result if both the operands are int, otherwise it returns a double. The type of a prefix expression can be specified as follows.

E ::= n E.*typ* = int
E ::= x E.*typ* = double
E ::= + E E E.*typ* = if E_{r1}.*typ* = E_{r2}.*typ*
then E_{r1}.*typ* else double

The corresponding executable specification in Prolog can be obtained by defining a binary relation 'typ' between prefix expression terms and their types as follows. Observe that each line of specification that ends in a "." is an axiom (first two are Prolog facts, while last two are Prolog rules), E, F, T, T1, and T2 are universally quantified Prolog variables, ":-" stands for *logical if*, and "," stands for *logical and*.

```
typ(i,int).
typ(x,double).
typ(+(E,F),T) :- typ(E,T), typ(F,T).
typ(+(E,F),real) :- typ(E,T1), typ(F,T2),  T1 \= T2.
```

A type checking query '?- typ(+(n,x),int).' verifies if the expression '+(n,x)' is of type int, while a type inference query '?- typ(+(n,x),T).' determines the type of the expression '+(n,x)'.

Attribute grammar specifying the translation of an equivalent expression language containing infix + into Java bytecode in the context of the instance method definition: 'class { double f(int n, double i) { return E;}}' is as follows:

E ::= n E.*code* = [iload_1]
E ::= x E.*code* = [dload_2]
E ::= E + E E.*code* = if E_{r1}.*typ* = int
then if E_{r2}.*typ* = int
then E_{r1}.*code*@E_{r2}.*code*@[iadd]
else E_{r1}.*code*@[i2d]@E_{r2}.*code*@[dadd]
else if E_{r2}.*typ* = int
then E_{r1}.*code*@E_{r2}.*code*@[i2d,dadd]
else E_{r1}.*code*@E_{r2}.*code*@[dadd]

The attribute *typ* has been specified earlier. The attribute *code* is bound to a list of Java bytecode. '@' refers to list append operation. Java compiler maps the formal parameters n and x to register 1, and register pair 2 and 3, respectively. (The double value requires two registers.) iload_1 (dload_2) stands for pushing the value of the int n (double x) on top of the stack; dadd (iadd) stands for popping the top two double (int) values from the stack, adding them, and pushing the result on top of the stack; and i2d stands for coercing an int value into a double value (Lindholm & Yellin, 1999). (Note that + is left associative in Java.) In practice, the code generator has to cater to variations on whether the method is static or instance, whether the formal arguments require 4 or more registers, whether the arguments are of primitive types or reference types, etc, and all this can be made explicit via attribute grammars.

Collecting Distinct Identifiers

We use the example of collecting distinct identifiers in an expression to illustrate the influence of primitive data types available for specifying the semantics on the ease of writing specifications, and the rules of thumb to be used to enable sound and complete attribute computation in one-pass using top-down left-to-right traversal of the abstract syntax tree.

<exp> ::= <var> | <exp> + <exp>

Suppose we have ADT SET available to us as a primitive. We associate synthesized attributes *id* and *ids* with <var> and <exp> respectively, to obtain the following attribute grammar. ('∪' refers to set-union.)

<exp> ::= <var>	<exp>.*ids* = { <var>.*id* }
<exp> ::= <exp> + <exp>	<exp>.*ids* = <exp>.*ids* U <exp>.*ids*

Instead, if we have only ADT LIST available to us, we associate synthesized attribute *envo* and inherited attribute *envi* of type list of symbols with <exp>, to obtain the following attribute grammar.

<exp> ::= <var>	<exp>.*envo* = if <var>.id ∈ <exp>.*envi* then <env>.*envi* else cons(<var>.*id*, <env>.*envi*)
<exp> ::= <exp> + <exp>	$<exp_{r1}>$.*envi* = $<exp_{l}>$.*envi* $<exp_{r2}>$.*envi* = $<exp_{r1}>$.*envo* $<exp_{l}>$.*envo* = $<exp_{r2}>$.*envo*

Observe that, given the definition of the attributes and the production rule, one can *automatically* determine the left-hand sides of the attribute computation rules required. *There is one rule for each synthesized attribute of the left-hand side non-terminal, and one rule for each inherited attribute of the right-hand side symbol (non-terminal).*

<exp>	::=	<exp>	+ <exp>
↓ envi		↓ *envi*	↓ *envi*
↑ *envo*		↑ envo	↑ envo

For the above production, the attributes shown in italics need to be determined using the attributes given in italics. *To enable one-pass top-down left-to-right computation of the attributes, each inherited attribute of the right-hand side symbol can depend on all the attributes associated with preceding right-hand side symbols and the inherited attribute of the left-hand side non-terminal. Similarly, the synthesized attribute of the left-hand side non-terminal can depend on all the attributes associated with all the right-hand side symbols and the inherited attribute of the left-hand side non-terminal.* Effectively, the inherited attribute associated with the left-hand side non-terminal provides context information from the parent and the left siblings, while the synthesized attributes of the right-hand side non-terminals provide information from their descendants.

From modularity perspective, if the attribute computation rules associated with each production satisfies the above constraints, it can be argued using induction principle that all the attributes associated with each node in the abstract syntax tree will be well-defined after one-pass of computation.

In practical compiler construction, a programming language is specified to a parser generator as an S-attributed grammar that transforms a program into its abstract syntax tree. This abstract syntax tree is traversed multiple times for semantic analysis and code generation. The APIs of the meta-language provide primitive functions and data types for defining and implementing semantic actions (attribute computations). For example, bottom-up parser generator Bison is based on C/C++ libraries, top-down parser generator ANTRL is based on Java APIs, and so forth.

Definite Clause Grammars

Prolog's definite clause grammars can be used to execute the set-based and the list-based attribute grammars discussed earlier. Each DCG rule resembles a production with the predicate name corresponding to the non-terminal and the formal arguments corresponding to the attributes. The input string is encoded as a list of symbols (tokens). Upon loading, the DCG rules are automatically translated into ordinary Prolog rules using difference-list implementation (Bratko, 2001). As a result, the predicate exp in the query has two additional arguments than are found in the DCG specification. The semantic action code inside curly braces incorporates calls to SWI-Prolog library functions.

The DCG corresponding to the set-based attribute grammar is as follows.

```
exp(Ids)    →   aexp(Ids).
exp(Ids)    →   aexp(Ids1), ['+'] , exp(Ids2), {union(Ids1,Ids2,Ids)}.
aexp([Id])  →   [Id], {atom(Id)}.

/* ?- exp(Vs, [x, '+', y, '+', z, +, y],[]). */
/* Vs = [x, z, y] ; */
```

The DCG corresponding to the list-based attribute grammar is as follows.

```
exp(Envi,Envo)   → aexp(Envi,Envo).
exp(Envi,Envo)       →      aexp(Envi,EnvT), ['+'] , exp(EnvT,Envo).
aexp(Envi,Envo)      →      [Id], {atom(Id)},
                            {member(Id,Envi) -> Envo = Envi;
                                    Envo = [Id | Envi]}.

/* ?- exp([],Vs, [x, '+', y, '+', z, +, y],[]). */
/* Vs = [z, y, x] ; */
```

The sample queries and the results have been commented out. Refer to (Bratko, 2001) for DCG details.

Specifying Compiler Generator Operations

Attribute grammars can be used to formalize and implement parser generators (Thirunarayan, 1984). For instance, consider the computation of nullable non-terminals, first-sets associated with non-terminals and follow-sets associated with non-terminals. A non-terminal is *nullable* if it can derive a null string. The *first-set* associated with a non-terminal is the set of tokens that can begin a string derivable from the non-terminal. The *follow-set* associated with a non-terminal is the set of tokens that can come after the non-terminal in a sentential form. These can be specified for an expression grammar along the lines indicated below. (Only partial specification has been given.)

S ::= T E
E ::= ε | + S
T ::= x | y

S.*nullable* = T.*nullable* and E.*nullable*
E.*nullable* = true
T.*nullable* = false

S.*first-set* = if T.*nullable* then T.*first-set* ∪ E.*first-set* else T.*first-set*
E.*first-set* = {ε , +}
T.*first-set* = {x , y}

T.*follow-set* = if E.*nullable* then S.*follow-set* ∪ E.*first-set* else E.*first-set*
S.*follow-set* = E.*follow-set*
E.*follow-set* = S.*follow-set*

The dependencies among the various attributes can be exploited to sequence their computation in three phases: compute *nullable* first, followed by *first-sets*, followed by *follow-sets* (Bochmann, 1976). Each phase can potentially require multiple iterations to converge to a fixed-point. The *nullable* is a synthesized attribute and requires multiple bottom-up pass. The *first-set* is also a synthesized attribute computed using multiple bottom-up left-to-right pass. The *follow-set* is an inherited attribute computed using multiple top-down right-to-left pass.

Optimizing Bottom-up Database Query Evaluation

Top-down query evaluation and bottom-up query evaluation are two well-known deductive database query implementation strategies (Ramakrishnan & Sudarshan, 1991). In top-down approach, the query solution tree is grown from the root (goal) to the leaves (data), applying the datalog (function-free Horn-logic or Prolog) rules from left to right. It is potentially incomplete in the presence of left-recursive rules, but it is efficient because the bindings in the queries are propagated to the base relations (data). In bottom-up approach, the query solution tree is grown from the leaves (data) to the root (goal), applying the datalog rules from right to left. It is complete, but can be inefficient because the search is not goal directed. Memoing techniques can be used to improve top-down strategy by making its search complete, while magic predicates can be employed to improve bottom-up strategy by making it more focused and efficient (Beeri & Ramakrishnan, 1987). The ideas underlying the definition of magic predicates and sideways information processing are reminiscent of attribute computation rules involving inherited and synthesized attributes as explained and illustrated in the following.

Consider the definition of the ancestor relation based on the parent relation, and the ancestor query with the first argument bound.

```
ancestor(X,Y) :- parent(X,Y).
ancestor(X,Y) :- parent(X,Z), ancestor(Z,Y).
?- ancestor(john, N).
```

In order to explore only a small fragment of the necessary parent facts to compute the query answers using naïve bottom-up approach, it is important to restrict the "firing" of the second rule. This can be done by transforming the second rule using *magic* predicates so that it is satisfied only for the ancestors of john as follows.

```
magic(john).
magic(Z) :- magic(X), parent(X,Z).
ancestor(X,Y) :- magic(X), parent(X,Y).
ancestor(X,Y) :- magic(X), parent(X,Z), ancestor(Z,Y).
```

The rules for *magic* predicate are analogous to the computation of attributes. The first fact corresponds to the inherited attribute value corresponding to the first argument of ancestor initialized by the top-level query. The second rule corresponds to the computation of additional inherited attribute values corresponding to the first argument of ancestor obtained through sideways information passing of the synthesized attribute from the parent relation. Beeri and Ramakrishnan (1987) provide several additional examples embodying attribute grammars ideas for query optimization.

Specifying Algorithms for Customized Information Extraction

In this section, we discuss an information extraction problem of industrial significance that has benefited from attribute grammar based algorithm specification. We skip the detailed specification *per se* due to space constraints.

A typical materials and process spec (from authoring organizations such as GE, Pratt and Whitney, ASTM, AMS, and so forth) contains requirements for making and testing a variety of alloys. A typical customer order specifies the desired material in terms of the specs it must conform to, and a collection of domain-specific product parameters such as product type, spec class, product dimension and cross-section, etc. A fundamental operation of interest on a spec is the determination of applicable fragments of the spec for a customer order. The goal of coarse-grain extraction is to convert a spec into a form that can be evaluated against the order parameters to determine applicable fragments.

To balance the commercial viability of the extraction task and its tractability, a spec is transformed into a possibly, nested sequence of conditioned notes of the form "If CONDITIONS Then [Note = " ... "]", where the note contains contiguous block of spec text. These extractions are cheap to produce because the detailed requirements are still in text.

In order to formalize conditioned notes, we need to propose a structure for the conditional expression and the note it qualifies. The conditional expression can be formed

using characteristic names (e.g., spec class, alloy, product type, etc.), constants, relational symbols (e.g., `=', `$>$', etc.), and boolean connectives (e.g., `and', `or', etc) in the standard way. The note can be defined in quanta of (sub)-sections and paragraphs. Thus, there are two important technical problems to be solved for carrying out extraction: (1) identification of the values of a characteristic that can appear as a condition, and (2) transformation of the relevant spec text into a sequence of conditioned notes.

The conditioned notes can be specified in terms of the scope rules of applicability of characteristic-value pairs to the spec text fragments. For instance, to associate a spec class with a section or a paragraph, we use the following heuristic: Every (sub-)section is conditioned on all spec classes named in section 'scope.' Explicit spec class references in a paragraph override the default condition. Otherwise, a paragraph inherits the condition from its left sibling (earlier paragraph), or transitively from its parent (enclosing (sub-)section). The rationale behind the heuristic is that, when the conditionals in an extraction are evaluated against the given condition values, it should generate all applicable fragments of the spec.

To abstract the algorithmic details of the extraction, the structure of a spec can be captured using EBNF as:

```
<document> ::= <document-header> <section>+
<section> ::= <sectionNumber> <sectionHeading> <paragraph>+ <section>*
```

and the computation of the conditioned notes involving spec classes can be given using attribute grammars. There are many other qualifiers of interest besides spec classes such as products, product types, alloys, and so forth. (See Thirunarayan et al., 2005 for a detailed attribute grammar specification.)

FUTURE TRENDS

Information flow ideas underlying attribute grammars can provide a general framework for designing and specifying algorithms. For example, Neven (2005), and Neven and Den Bussche (2002) demonstrate the influence of attribute grammars on query languages.

Web technologies such as XML/XSLT that are based on adorned context-free grammars can benefit from techniques developed for attribute grammars (Harold, 2004). Conceptually, an XML document consists of annotations, where each annotation consists of an associated XML-element that reflects the semantic category to which the corresponding text fragment belongs, and the associated XML-attributes that are bound to relevant semantic values. Overlaying domain-specific XML tags on a text document enables abstraction, formalization, and in-place embedding of machine-processable semantics. In the future, we can expect the annotated data to be interpreted by viewing it as a function/procedure call, and defining the XML-element as a function/procedure in a language such as XSLT or Water for associating different collections of behaviors with XML-elements (Thirunarayan, 2005).

CONCLUSION

Historically, attribute grammars were developed in the context of compiler construction. We provided several examples illustrating the application of attribute grammars for static analysis of programs, for program translation, and for specifying certain phases of a parser generator. We also provided general principles for developing attribute grammar specifications for efficient one-pass computation of attributes. We introduced Prolog's definite clause grammars to enable attribute grammars to be viewed as executable specifications. We showed how the information flow ideas implicit in the attribute computation rules can be exploited to optimize bottom-up evaluation of datalog programs. Finally, we discussed the application of attribute grammars for specifying information extraction algorithms, and its future role in XML technologies.

REFERENCES

Aho, A. V., Lam, M., Sethi, R., & Ullman, J. D. (2007). *Compilers: principles, techniques, and tools*. Addison Wesley.

Beeri, C., & Ramakrishnan, R. (1987). On the Power of Magic, *Proceedings of the 6th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, (pp. 269-283).

Bratko, I. (2001). *Prolog Programming for Artificial Intelligence*, (3rd ed.). Addison Wesley.

Bochmann, G. V. (1976). Semantic evaluation from left to right. *Communications of the ACM, 19*(2), 55-62.

Harold, E. R. (2004). *XML in a nutshell*, (3rd ed.). O'Reilly.

Knuth, D. E. (1968). Semantics of context-free languages. *Mathematical Systems Theory, 2*(2), 127-145. (Corrigenda: *Mathematical Sysems Theory 5*(1),1971, 95-96.)

Lindholm, T., & Yellin, F. (1999). *Programming Java virtual machine specification,* (2nd ed.). Addison-Wesley.

Louden, K. C. (2003). *Programming languages: Principles and practice*, (2nd ed.). Thomson – Course Technology.

Neven, F. (2005). Attribute grammars for unranked trees as a query language for structured documents. *Journal of Computer and System Sciences, 70,* 221-257.

Neven, F., & Den Bussche, J. V. (2002). Expressiveness of structured document query languages based on attribute grammars. *Journal of the ACM, 49*(1), 56-100.

Paakki, J. (1995). Attribute grammar paradigms—a high-level methodology in language implementation, *ACM Computing Surveys, 27*(2), 196-255.

Ramakrishnan, R., & Sudarshan, S. (1991). Top-Down versus Bottom-Up Revisited. *In Proceedings of the 1991 International Symposium on Logic Programming*, (pp. 321-336).

Thirunarayan, K. (1984). *A Compiler-Generator Based on Attributed Translation Grammars,* M. E. Thesis, Indian Institute of Science, Bangalore. (Advisors: Priti Shankar and Y. N. Srikant)

Thirunarayan, K. (2005). On embedding machine-processable semantics into documents. *IEEE Transactions on Knowledge and Data Engineering, 17*(7), 1014-1018.

Thirunarayan, K., Berkovich, A., & Sokol, D. (2005). An information extraction approach to reorganizing and summarizing specifications. *Information and Software Technology Journal, 47*(4), 215-232, 2005.

KEY TERMS

A

Attribute Computation Rules: Rules used to compute attributes, usually defined in terms of other attributes and standard primitives.

Attribute Grammar: An attribute grammar is an extension of context-free grammar that enables definition of context-sensitive aspects of a language and its translation.

DCG: A definite clause grammar is a Prolog built-in mechanism for implementing attribute grammars efficiently using difference lists.

EBNF: Extended Backus Naur formalism is an extension of context-free grammar with regular expression operations for defining context-free languages. It provides a more concise syntax specification.

Inherited Attributes: These attributes pass information from root to the leaves of a parse tree, or sideways among siblings.

Machine-Processable Semantics: Metadata added to the documents to enable machines to understand and reason with text or multi-media content.

Synthesized Attributes: These attributes pass information from leaves to the root of a parse tree.

XML/XSLT: Extensible markup l,anguage is a meta-language for creating markup languages. XML is a subset of SGML. XHTML is an instance of XML. Extensible stylesheet language transformations is a language for manipulating XML documents.

Audience-Driven Design Approach for Web Systems

Olga De Troyer
WISE Research Group, Belgium

INTRODUCTION

In the last years, Web systems have evolved from a simple collection of hypertext pages toward applications supporting complex (business) applications, offering (rapidly changing) information and functionality to a highly diversified audience. Although it is still easy to publish a couple of pages, it is now recognized that appropriate Web design methods are needed to develop more complex Web sites and applications (generally called Web systems). In the past, Web systems were created opportunistically without prior planning or analysis, and without any regard for methodology, resulting in Web systems that were lacking consistency in structure, navigation, and presentation, and were not transparent. A lot of these systems were also suffering from the classical maintenance problems and development backlog. In the same period, Web technology evolved at an equally dazzling rate enabling more advanced Web applications, but with the unfavorable consequence that Web development is no longer simple and easy. The latest developments in the field of the Web are related to the vision of the Semantic Web: an extension of the current Web in which information is given well-defined meaning, better enabling computers, and people to work in cooperation (Berners-Lee, Hendler, & Lassila, 2001).

Together with the Web, a new problem unknown in classical information systems emerged: competition for the visitor's attention. Especially for commercial Web systems, it is important to hold the interest of the visitors and to keep them coming back. As stated by usability expert Nielsen (2000, p. 9), "all the competitors in the world are but a mouse click away." Much more than in "classical" software systems, the usability of Web systems is a primary factor for their success.

BACKGROUND

One way to deal with the usability of a Web system is by assessing the usability once the system is built and improving it if necessary. The techniques for accessing the usability of a Web system are mainly the same as those used in usability testing of classical user interfaces, for example, heuristic evaluation, expert-based evaluation, experimental evaluation, interviews, questionnaires, and so forth (Nielsen & Mack, 1994). Also, different tools are developed that support assessing the usability of Web sites (e.g., WebQuilt [Hong, Heer, Waterson, & Landay, 2001; Vanderdonckt, Beirekdar, & Noirhomme-Fraiture, 2004], and the full-featured experimentation environment of Noldus [www.noldus.com]). Another approach to enhance usability (and complementary to the first approach) is to use a Web design method that ensures a higher usability. The first Web design methods, HDM (Garzotto, Paolini, & Schwabe, 1993) and its successors HDM2 (Garzotto, Paolini, & Mainetti, 1993) and OOHDM (Schwabe & Rossi, 1995), and RMM (Isakowitz, Stohr, & Balasubramanian, 1995), were originally designed for hypertext applications or came from the database research community. These methods used database design methods like E-R (Chen, 1976) or OMT (Rumbaugh, Blaha, Premerlani, Eddy, & Lorensen, 1991), and focused on the organization of the data to be presented on the Web. These methods could solve to some extent maintenance problems, but they did not address usability. Essential for achieving a good usability in Web systems is meeting the needs of the (different) visitors. WSDM was one of the first Web design method to recognize this. This method was presented at the WWW7 conference (1998) as a "user-centered" design method for Web sites (De Troyer & Leune, 1998). The starting point in the approach is the set of potential visitors (audiences) of the Web system. The method recognizes that different types of visitors have different needs and that this should drive the design of the Web system rather than the organization of the available data. Later on (De Troyer, 2001), the authors renamed their approach from "user-centered" to "audience-driven," to avoid confusion with the term "user-centered" from the HCI (human-computer interaction) field. In HCI, a user-centered approach refers to a design process in which users are actively involved (by interviews, scenario analysis, prototyping, evaluation, etc.). This explicit involvement is not necessary in WSDM. On the contrary, the individual Web users are unknown during the Web development process; they cannot be interviewed in advance, and they cannot be involved in the development process. In the audience-driven approach as defined by WSDM, the users play a central role, but it is not necessary to involve them actively in the development process.

Since the late 1990s, several Web design methods have been conceived. Some examples are WebML (Ceri, Fraternali, & Bongio, 2000), UWE (Koch & Kraus, 2001), HERA (Houben, Barna, Frasincar, & Vdovjak, 2003), OO-H (Gómez, Cachero, & Pastor, 2003), SHDM (Schwabe, Szundy, de Moura, & Lima, 2004), and Co-Design (Schewe & Thalheim, 2005). Some of these design methods focus on adaptivity or personalization as a way to enhance the usability. These methods use a user model to adapt the Web system to the needs or characteristics of an individual user. This implies that a particular Web system will look different to two different users, or even that the system will look different when the same user revisits it. Although personalization may be undoubtedly a good solution in some situations (e.g., e-learning, e-commerce), in other situations it may be less appropriate (e.g., regular Web systems). In this article, we will not consider personalization as a way to enhance usability.

When designing a Web system, there are two important questions to be answered:

1. What information and services should be provided?
2. How should all this information and services be structured?

To answer these questions, different design approaches can be followed. One of them is the *audience-driven* approach. Other possible approaches are the *data-driven* approach and the *organization-driven* approach.

In a data-driven approach, the data (and services) available in the organization (in databases, brochures, internal documents, etc.) are the design's starting point. Following this approach, the structure of the Web system will reflect the way the data are structured and maintained in the organization, and the content will parallel the internal data. The same applies for services or functionality. Forms available in the organization will be converted into e-forms, and the current way of working will be reflected in the Web system. The advantage is that structuring the information and services is easy and that the maintenance can be done in parallel with the maintenance of the internal data and procedures. However, the disadvantages are: (1) the data are presented and organized the way they are used in the organization. This is not necessarily how people external to the organization need it or perceive it; (2) information or services may be missing because it was not available in the form of a specific document or existing procedure and the designers were not aware of the fact that users may need this; (3) all information and all services are offered to all users. As a consequence, visitors may be drowned in information.

In an organization-driven approach, the internal structure of the organization is the starting point: the structure of the Web system reflects the structure of the organization. This approach is often used for large organizations with a lot of divisions, for example, a university Web system that reflects its internal structure into faculties, departments, and research institutes. As for the data driven approach, it is easy to structure the Web system, and the development and maintenance of the different parts can be assigned to the different divisions of the organization. The disadvantage is that it may be very difficult for visitors not familiar with the internal structure of the organization to know where to look for information or services.

In the audience-driven approach, the information and services needed in the Web system are determined by the needs and requirements of the target audiences (users). Also the main structure of the Web system will be based on the different types of audiences and their requirements. This last point differentiates the audience-driven approach from many so-called user-centered approaches. We illustrate this with an example, a university Web site. Following the audience-driven approach, the university Web site would (at least) contain a part with general information interesting to all visitors; a part with information specific for students and lecturers; and a part containing information for researchers and third parties interested in research. The audience-driven approach gives consideration to the fact that (large) Web systems usually have different types of visitors that may have different needs. Clearly, such Web systems will have a higher usability than the ones structured using a data-driven or organization-driven approach. However, the downsides of the medal are that the effort needed to design the Web system is higher and that the task of maintaining may be spread over the organization (usually, there will be no one-to-one mapping from the structure of the Web system onto the structure of the organization).

THE AUDIENCE-DRIVEN APPROACH

As explained in the introduction, an audience-driven design approach means that the different target audiences (visitors) and their requirements are taken as starting points for the design and that the main structure of the Web system is derived from this. Concretely, this results in a Web system where the homepage contains different navigation paths (called audience tracks), one for each different kind of visitor.

To arrive to such an audience-driven organization of the Web system, the different types of audiences and their needs are identified already in an early stage of the design process (Casteleyn & De Troyer, 2001). One way to identify the different types of audiences is by looking at the activities of the organization relevant for the Web system and the role people play in these activities. These people are the potential users (audiences) of the Web system. For example, the activities of a university are "performing research," "giving courses," and "advising third parties." The people involved are researchers, students, potential students, teaching staff,

and third parties. Next, the people identified, are classified into *audience classes* by collecting their requirements (information-, as well as functional- and usability requirements), that is, we establish what kind of information those people want to find on the Web system (information requirements), what kind of functionality they expect (functional requirements), and which special needs they have from a usability point of view, for example, advanced search functionality (usability requirements). Users with the same information and functional requirements become members of the same audience class. Whenever the information and functional requirements of a set of users differ, a new audience class is defined, or if possible, an audience subclass is introduced. An audience class is a subclass of another audience class if the members of the subclass have all the same information- and functional requirements as the members of the superclass but also some extra requirements. If possible, the activities we started from are decomposed in order to refine the audience classes. This may introduce more audience subclasses. In this way, a hierarchy of audience classes can be constructed. The top of the audience class hierarchy is always the audience class visitor, which represents all target users. The requirements associated with the visitor class are the requirements that are common to all users. The requirements associated with a particular audience class are specific for the members of this audience class.

The audience class hierarchy is the basis for the main structure of the Web system. For each audience class, a separate *audience track* will be created. Such an audience track can be considered as a subsystem containing all and only the information and functionality needed by the members of the associated audience class. To fill in the detailed navigation and content of such a track, the different requirements of the corresponding audience class are considered. This is done as follows. With each requirement formulated for this audience class, a task is provided in the audience track that a user can choose when he or she wants to satisfy this requirement. The task will provide the necessarily information and functionality by means of nodes and links that can be followed. For a simple information requirement, this means that a node will be provided containing the necessarily information. For a functional requirement (i.e., booking a flight), nodes offering the necessary information and functionality will be provided connected by navigational links respecting the workflow of the task. Details on how this can be achieved are given by De Troyer and Casteleyn (2003).

Next to the fact that different types of users (audiences) may have different information and functional requirements, it may be necessary to represent the (same) information or functionality in different ways to different kinds of users. This depends on the characteristics of the users. As an example we again consider the university example. Potential students, especially secondary school students are not familiar with the university jargon and should be addressed in a young and dynamic way. Also, by preference, the information should be offered in the native language. The enrolled students are familiar with the university jargon. They also prefer to have the information in the native language; however, for foreign students (e.g., who follow exchange programs) English should be used as communication language. For researchers, it may be sufficient to use English. This can be taken into consideration by specifying for each audience class identified, the characteristics of its members. Examples of characteristics are level of experience with the Web in general, frequency of use, language issues, education/intellectual abilities, age, income, lifestyle, and so forth. Some of the characteristics may be translated into usability requirements, while others may be used to guide the design of the "look and feel" of the different navigation tracks, for example, choice of colors, fonts, graphics, and so forth.

FUTURE TRENDS

In the last years, many different researchers have recognized the need to take the users into consideration during the Web development process and adopted either the user-centered approach from the HCI field (as in Cato, 2001; Lazar, 2001; McCracken & Wolfe, 2004) or an approach similar to WSDM's audience-driven approach (with respect to the fact that it is necessary to give due attention to the users but not necessary to actively involve them in the design process) (as in Bomsdorf & Szwillus, 2003; Brinck, Gergle, & Wood, 2002; Lengels, 2002; McCracken & Wolfe, 2004). More and more, we see that due consideration is given to usability in general, but also to more specific issues such as accessibility for people with disabilities, localization, personalization, and context-awareness. One approach to the accessibility problem consists of stimulating people to design Web systems with accessibility in mind. In this context design guidelines are developed that Web developers can use to ensure that their Web system is accessible to people with disabilities (see the W3C Web Content Accessibility Guidelines [http://www.w3.org/WAI/]), and software exists that can be used to test if a Web system is conforming to these recommendations (e.g., W3C Markup Validation service [http://validator.w3.org]). Others try to automatically transform existing Web pages into a form more appropriate for disabled persons (e.g., Aurora [Huang & Sundaresan, 2000]), while also work is in progress to approach the problem from a Web engineering approach ensuring that Web systems are inherently accessible (e.g., Plessers et al., 2005). In the context of the Web, localization deals with adapting or preparing Web systems to be able to deal with the requirements of different local communities: different language, different regulations, and different culture (see, i.e., De Troyer et al., 2006, or De Troyer & Casteleyn, 2004). As already mentioned, personalization is used to adapt the content, structure and presentation of a Web system to

the needs of a particular user. This has proven to be useful in, for example, online shops to offer customized services and recommendations to the customer, or in e-learning systems to offer an individualized learning plan. Context-awareness aims at achieving a similar goal, but instead of adapting to the individual user, the system is adapted to the context in which it is used, for example, to the device on which it is used (PDA, mobile phone, etc.), the location, or the available connection. Sometimes, both are combined (such as in Ceri, Daniel, & Matera, 2003).

CONCLUSION

There exist different approaches to elicit and structure the information and services in a Web system. In the audience-driven approach, this is done by taking the different audiences of the Web system as the starting point. This results in Web systems where the information and the services are organized according to the needs of the different audience classes. This may result in higher usability, which is a primary factor for the success of Web systems.

REFERENCES

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 284*(5), 34-43.

Bomsdorf, B., & Szwillus, G. (2003). User-centered modeling of interactive Web sites. In I. King & T. Maray (Eds.), *Proceedings WWW2003 Conference (CD-ROM)*. Budapest: WWW2003.

Brinck, T., Gergle, D., & Wood, S. D. (2002). *Usability for the Web: Designing Web sites that work*. San Francisco: Morgan Kaufmann Publishers.

Casteleyn, S., & De Troyer, O. (2001). Structuring Web sites using audience class hierarchies. In H. Arisawa, Y. Kambayashi, V. Kumar, H.-C. Mayr, & I. Hunt (Eds.), *Conceptual Modeling for New Information Systems Technologies, ER 2001 Workshops, HUMACS, DASWIS, ECOMO, and DAMA, Lecture Notes in Computer Science* (Vol. 2465, pp. 198-211). Yokohama, Japan: Springer-Verlag.

Cato, J. (2001). *User-centered Web design*. Harlow, UK: Addison-Wesley Pearson Education.

Ceri, P., Fraternali, P., & Bongio, A. (2000). Web modeling language (WebML): A modeling language for designing Web sites. *Computer Networks and ISDN Systems, 33*(1-6), 137-157.

Ceri, S., Daniel, F., & Matera, M. (2003). Extending WebML for modeling multi-channel context-aware Web applications. In I. F. Akyildiz, & H. Rudin (Eds.), *Proceedings of the MMIS'03 Workshop (Mobile Multi-channel Information Systems)* (pp. 225-233). Rome: IEEE Press.

Chen, P. P. (1976). The entity-relationship model: Towards a unified view of data. *ACM Transactions on Database Systems, 1*(1), 471-522.

De Troyer, O. (2001). Audience-driven Web design. In M. Rossi & K. Siu (Eds.), *Information modeling in the new millennium* (pp. 442-462). Hershey, PA: Idea Group Publishing.

De Troyer, O., & Casteleyn, S. (2003). Modeling complex processes for Web applications using WSDM. In D. Schwabe, O. Pastor, G. Rossi, & L. Olsina (Eds.), *Proceedings of the Third International Workshop on Web-Oriented Software Technologies IWWOST2003*. Oviedo, Spain. Retrieved September 2003, from http://www.dsic.upv.es/~west/iwwost03/articles.htm

De Troyer, O., & Casteleyn, S. (2004). Designing localized Web sites. In X. Zhou, S. Su, M. P. Papazoglou, M. E. Orlowska, & K. G. Jeffery (Eds.), *Proceedings of the Fifth International Conference on Web Information Systems Engineering (WISE2004)* (pp. 547-558). Brisbane, Australia: Springer-Verlag.

De Troyer, O., & Leune, C. (1998). WSDM: A user-centered design method for Web sites. In H. Ashman & P. Thistlewaite (Eds.), *Computer networks and ISDN systems. Proceedings of the 7th International World Wide Web Conference* (pp. 85-94). Brisbane, Australia: Elsevier.

De Troyer, O., Mushtaha, A. N., Stengers, H., Baetens, M., Boers, F., Casteleyn, S., et al. (2006). On cultural differences in local Web interfaces. *Journal of Web Engineering, 5*(3), 246-264.

Garzotto, F., Paolini, P., & Mainetti, L. (1993). Navigation patterns in hypermedia databases. In G. Marchionini (Ed.), *Proceedings of the 26th Hawaii International Conference on System Science* (pp. 370-379). New York: IEEE Computer Society Press.

Garzotto, F., Paolini, P., & Schwabe, D. (1993). HDM—A model-based approach to hypertext application design. *ACM Transactions on Information Systems, 11*(1), 1-26.

Gómez, J., Cachero, C., & Pastor, O. (2003). Modelling dynamic personalization in Web applications. In *Third International Conference on Web Engineering—ICWE 2003*, (LNCS 2722, pp. 472-475). Oviedo, Spain: Springer-Verlag.

Hong, J. I., Heer, J., Waterson, S., & Landay, J. A. (2001). WebQuilt: A proxy-based approach to remote Web usability testing. *ACM Transactions on Information Systems, 19*(3), 263-385.

Houben, G.-J., Barna, P., Frasincar, F., & Vdovjak, R. (2003). HERA: Development of Semantic Web information systems. In *Third International Conference on Web Engineering—ICWE 2003* (LNCS 2722, pp. 529-538). Oviedo, Spain: Springer-Verlag.

Huang, A. W., & Sundaresan, N. (2000). Aurora: A conceptual model for Web-content adaptation to support the universal usability of Web-based services. In J. Thomas (Ed.), *Proceedings on the 2000 ACM Conference on Universal Usability* (pp. 124-131). Arlington, VA: ACM Press.

Isakowitz, T., Stohr, E. A., & Balasubramanian, P. (1995). RMM: A methodology for structured hypermedia design. *Communications of the ACM, 38*(8), 34-43.

Koch, N., & Kraus, A. (2001). The authoring process of UML-based Web engineering approach. In O. Pastor (Ed.), *Proceedings of the First International Workshop on Web-Oriented Software Construction (IWWOST'02)* (pp. 105-119). Valencia, Spain: Valencia University of Technology.

Lazar, J. (2001). *User-centered Web development*. Boston: Jones and Bartlett Publishers, Inc.

Lengels, J. G. (2002). *The Web wizard's guide to Web design*. Boston: Addison-Wesley Pearson Education.

McCracken, D. D., & Wolfe, R. J. (2004). *User-centered Web site development: A human-computer interaction approach*. Upper Saddle River, NJ: Pearson Prentice Hall.

Nielsen, J. (2000). *Designing Web usability: The practice of simplicity*. Indianapolis: New Riders Publishing.

Nielsen, J., & Mack, R. L. (Eds.). (1994). *Usability inspection methods*. New York: John Wiley.

Plessers, P., Casteleyn, S., Yesilada, Y., De Troyer, O., Stevens, R., Harper, S., et al. (2005). Accessibility: A Web engineering approach. In A. Ellis & T. Hagino (Eds.), *Proceedings of the 14th International World Wide Web Conference (WWW2005)* (pp. 353-362). Chiba, Japan: ACM.

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., & Lorensen, W. (1991). *Object oriented modeling and design*. Upper Saddle River, NJ: Prentice Hall Inc.

Schwabe, D., & Rossi, G. (1995). The object-oriented hypermedia design model. *Communications of the ACM, 38*(8), 45-46.

Schwabe, D., Szundy, G., de Moura, S. S., & Lima, F. (2004). Design and implementation of Semantic Web applications. In C. Bussler, S. Decker, D. Schwabe, O. Pastor (Eds.), *Proceedings of the Workshop on Application Design, Development and Implementation Issues in the Semantic Web (WWW 2004), CEUR Workshop Proceedings* (Vol. 105) [CD-ROM]. New York: CEUR-WS.org.

Schewe, K.-D., & Thalheim, B. (2005). Conceptual modelling of Web information systems. *Data and Knowledge Engineering, 54*(2), 147-188.

Vanderdonckt, J., Beirekdar, A., & Noirhomme-Fraiture, M. (2004). Automated evaluation of Web usability and accessibility by guideline review. In N. Koch, P. Fraternali, & M. Wirsing (Eds.), *ICWE 2004,* (LNCS 3140, pp. 17-30). Berlin, Heidelberg: Springer-Verlag.

KEY TERMS

Accessibility: People with disabilities can perceive, understand, navigate, and interact with the Web, and they can contribute to the Web.

Audience Class: Group of target visitors of a Web system with the same requirements.

Audience-Driven Web Design: The different audiences and their requirements are taken as the starting point for the design of the Web system. The information and services in the Web system are organized around these different audiences.

Audience Track: Part of the Web system that provides information and services specifically tailored to a particular audience class.

Context-Awareness: Property of devices or systems that have information about the circumstances under which they operate and can react accordingly. Context-aware systems may also try to make assumptions about the user's current situation (also called personalization).

Data-Driven Web Design: The data available in the organization are taken as the starting point for the design of the Web system.

Localization: In Web design and software, localization refers to the adaptation of language, content, and design to reflect local cultural sensitivities.

Organization-Driven Web Design: The structure of the organization is taken as the starting point for the design of the Web system. The structure of the organization is reflected in the Web system.

Usability: The extent to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use.

User-Centered Web Design: The requirements of the users of a Web system play a central role in the design process.

Category: Multimedia Technology

Audio Analysis Applications for Music

Simon Dixon
Austrian Research Institute for Artificial Intelligence, Austria

INTRODUCTION

The last decade has seen a revolution in the use of digital audio: The CD, which one decade earlier had taken over the home audio market, is starting to be replaced by electronic media which are distributed over the Internet and stored on computers or portable devices in compressed formats. The need has arisen for software to manage and manipulate the gigabytes of data in these music collections, and with the continual increase in computer speed, memory and disk storage capacity, the development of many previously infeasible applications has become possible.

This article provides a brief review of automatic analysis of digital audio recordings with musical content, a rapidly expanding research area which finds numerous applications. One application area is the field of music information retrieval, where content-based indexing, classification and retrieval of audio data are needed in order to manage multimedia databases and libraries, as well as being useful in music retailing and commercial information services. Another application area is music software for the home and studio, where automatic beat tracking and transcription of music are much desired goals. In systematic musicology, audio analysis algorithms are being used in the study of expressive interpretation of music. Other emerging applications which make use of audio analysis are music recommender systems, playlist generators, visualisation systems, and software for automatic synchronisation of audio with other media and/or devices.

We illustrate recent developments with three case studies of systems which analyse specific aspects of music (Dixon, 2004). The first system is BeatRoot (Dixon, 2001a, 2001c), a beat tracking system that finds the temporal location of musical beats in an audio recording, analogous to the way that people tap their feet in time to music. The second system is JTranscriber, an interactive automatic transcription system based on (Dixon, 2000a, 2000b), which recognizes musical notes and converts them into MIDI format, displaying the audio data as a spectrogram with the MIDI data overlaid in piano roll notation, and allowing interactive monitoring and correction of the extracted MIDI data. The third system is the Performance Worm (Dixon, Goebl, & Widmer, 2002), a real-time system for visualisation of musical expression, which presents in real time a two dimensional animation of variations in tempo and loudness (Langner & Goebl, 2002, 2003).

Space does not permit the description of the many other music content analysis applications, such as: audio fingerprinting, where recordings can be uniquely identified with a high degree of accuracy, even with poor sound quality and in noisy environments (Wang, 2003); music summarisation, where important parts of songs such as choruses are identified automatically; instrument identification, using machine learning techniques to classify sounds by their source instruments; and melody and bass line extraction, essential components of query-by-example systems, where music databases can be searched by singing or whistling a small part of the desired piece. At the end of the article, we discuss emerging and future trends and research opportunities in audio content analysis.

BACKGROUND

Early research in musical audio analysis is reviewed by Roads (1996). The problems that received the most attention were pitch detection, spectral analysis and rhythm recognition, areas which correspond respectively to the three most important aspects of music: melody, harmony and rhythm.

Pitch detection is the estimation of the fundamental frequency of a signal, usually assuming it to be monophonic. Methods include: time domain algorithms such as counting of zero-crossings and autocorrelation; frequency domain methods such as Fourier analysis and the phase vocoder; and auditory models which combine time and frequency domain information based on an understanding of human auditory processing. Recent work extends these methods to find the predominant pitch (e.g., the melody note) in a polyphonic mixture (Gómez, Klapuri, & Meudic, 2003; Goto & Hayamizu, 1999).

Spectral analysis is a well-understood research area with many algorithms available for analysing various classes of signals, such as the short time Fourier transform, wavelets and other more signal-specific time-frequency distributions. Building upon these methods, the specific application of automatic music transcription has a long research history (Chafe, Jaffe, Kashima, Mont-Reynaud, & Smith, 1985; Dixon, 2000a, 2000b; Kashino, Nakadai, Kinoshita, & Tanaka, 1995; Klapuri, 1998, 2003; Klapuri, Virtanen, & Holm, 2000; Marolt, 1997, 1998, 2001; Martin, 1996; Mont-Reynaud, 1985; Moorer, 1975; Piszczalski & Galler, 1977; Sterian, 1999; Watson, 1985). Certain features are

common to many of these systems: producing a time-frequency representation of the signal, finding peaks in the frequency dimension, tracking these peaks over the time dimension to produce a set of partials, and combining the partials to produce a set of notes. The differences between systems are usually related to the assumptions made about the input signal (e.g., the number of simultaneous notes, types of instruments, fastest notes, or musical style), and the means of decision making (e.g., using heuristics, neural nets or probabilistic reasoning).

The problem of extracting rhythmic content from a musical performance, and in particular finding the rate and temporal location of musical beats, has also attracted considerable interest in recent times (Allen & Dannenberg, 1990; Cemgil, Kappen, Desain, & Honing, 2000; Desain, 1993; Desain & Honing, 1989; Dixon, 2001a; Eck, 2000; Goto & Muraoka, 1995, 1999; Large & Kolen, 1994; Longuet-Higgins, 1987; Rosenthal, 1992; Scheirer, 1998; Schloss, 1985). Previous work had concentrated on rhythmic parsing of musical scores, lacking the tempo and timing variations that are characteristic of performed music, but recent tempo and beat tracking systems work quite successfully on a wide range of performed music.

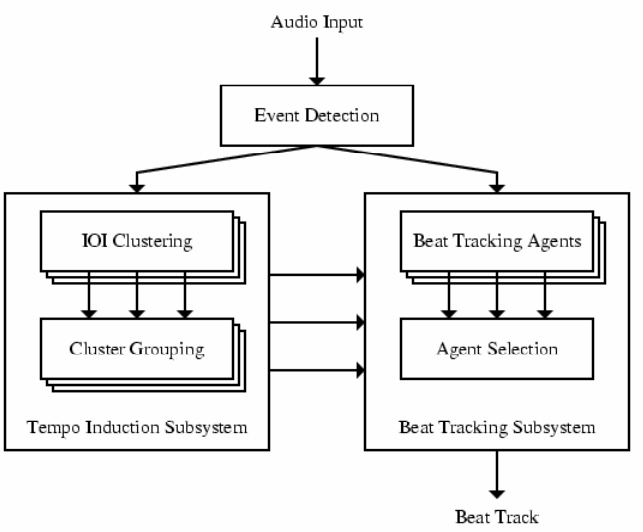
Music performance research is only starting to take advantage of the possibility of audio analysis software, following work such as Scheirer (1995) and Dixon (2000a). Previously, general purpose signal visualisation tools combined with human judgement had been used to extract performance parameters from audio data. The main problem in music signal analysis is the development of algorithms to extract sufficiently high level content, since it requires the type of musical knowledge possessed by a musically literate human listener. Such "musical intelligence" is difficult to encapsulate in rules or algorithms that can be incorporated into computer programs. In the following sections, three systems are presented which take the approach of encoding as much as possible of this intelligence in the software and then presenting the results in an intuitive format which can be edited via a graphical user interface, so that the systems can be used in practical settings even when not 100% correct. This approach has proved to be very successful in performance research (Dixon et al., 2002; Goebl & Dixon, 2001; Widmer, 2002; Widmer, Dixon, Goebl, Pampalk, & Tobudic, 2003).

BEATROOT

Compared with complex cognitive tasks such as playing chess, beat tracking (identifying the basic rhythmic pulse of a piece of music) does not appear to be particularly difficult, as it is performed by people with little or no musical training, who tap their feet, clap their hands or dance in time with music. However, while chess programs compete with world champions, no computer program has been developed which approaches the beat tracking ability of an average musician, although recent systems are approaching this target. In this section, we describe BeatRoot, a system which estimates the rate and times of musical beats in expressively performed music (for a full description, see Dixon, 2001a, 2001c).

BeatRoot models the perception of beat by two interacting processes (see Figure 1): The first finds the rate of

Figure 1. System architecture of BeatRoot



the beats (tempo induction), and the second synchronises a pulse sequence with the music (beat tracking). At any time, there may exist multiple hypotheses regarding each of these processes; these are modelled by a multiple agent architecture in which agents representing each hypothesis compete and cooperate in order to find the best solution. The user interface presents a graphical representation of the music and the extracted beats, and allows the user to edit and recalculate results based on the editing. Input to BeatRoot is either digital audio or symbolic music data such as MIDI. This data is processed off-line to detect salient rhythmic events, using an onset detection algorithm which finds peaks in the slope of the amplitude envelope of the signal (or a set of frequency bands of the signal). The timing of these events is then analysed to generate hypotheses of the tempo at various metrical levels.

First, inter-onset intervals (IOIs), the time differences between pairs of onsets, are calculated, and then a clustering algorithm is used to find groups of similar IOIs which represent the various musical units (e.g., half notes, quarter notes). Information about the clusters is combined by identifying near integer relationships between clusters, in order to produce a ranked list of tempo hypotheses, which is then passed to the beat tracking subsystem.

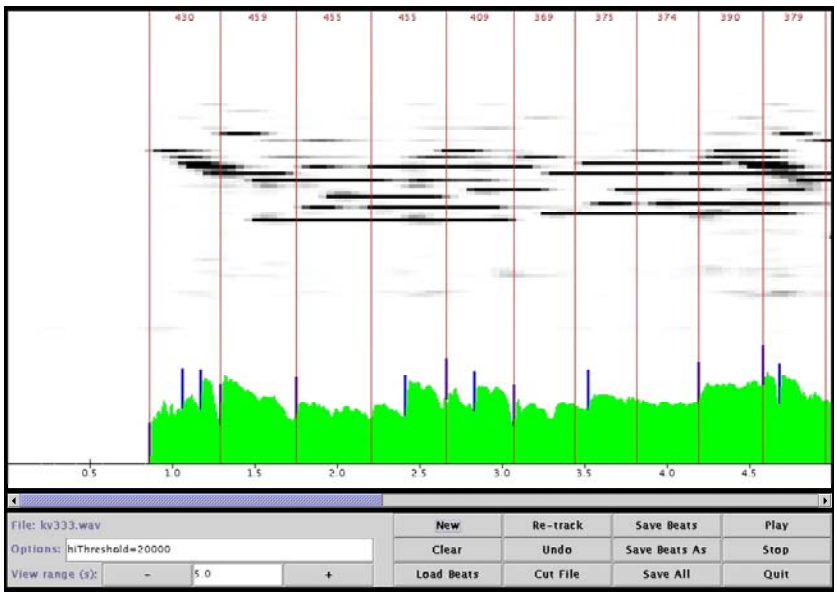
The beat tracking subsystem uses a multiple agent architecture to find sequences of events which match the various tempo hypotheses, and rates each sequence to determine the most likely sequence of beat times. Each agent represents a specific hypothesis about the rate and the timing of the beats, which is updated as the agent matches the detected onsets to predicted beat times. The agent also evaluates its beat tracking, based on how evenly the beat times are spaced, how many predicted beats correspond to actual events, and the salience of the matched events, which is calculated from the signal amplitude at the time of the onset. At the end of processing, the agent with the highest score outputs its sequence of beats as the solution to the beat tracking problem.

BeatRoot is written in Linux/C++, and comprises about 10,000 lines of code, with a graphical user interface consisting of 1,000 lines of Java. The user interface allows playback of the music with the beat times marked by clicks, and provides a graphical display of the signal and the beats with editing functions for correction of errors or selection of alternate metrical levels (Figure 2). BeatRoot is open source software (under the GNU Public License), and is available from:

http://www.oefai.at/~simon/beatroot

The lack of a standard corpus for testing beat tracking creates a difficulty for making an objective evaluation of the system. The automatic beat tracking algorithm has been tested on several sets of data: a set of 13 complete piano sonatas, a large collection of solo piano performances of two Beatles songs and a small set of pop songs. In each case, the system found an average of over 90% of the beats (Dixon, 2001a), and compared favourably to another state of the art

Figure 2. Screen shot of BeatRoot processing the first five seconds of a Mozart piano sonata, showing the inter-beat intervals in ms (top), calculated beat times (long vertical lines), spectrogram (centre), waveform (below) marked with detected onsets (short vertical lines) and the control panel (bottom)



tempo tracker (Dixon, 2001b). Tempo induction results were almost always correct, so the errors were usually related to the phase of the beat, such as choosing as beats onsets half way between the correct beat times. Interested readers are referred to the sound examples at:

http://www.oefai.at/~simon

Presently, BeatRoot is being used in a large scale study of interpretation in piano performance (Widmer, 2002; Widmer et al., 2003), to extract symbolic data from audio CDs for automatic analysis.

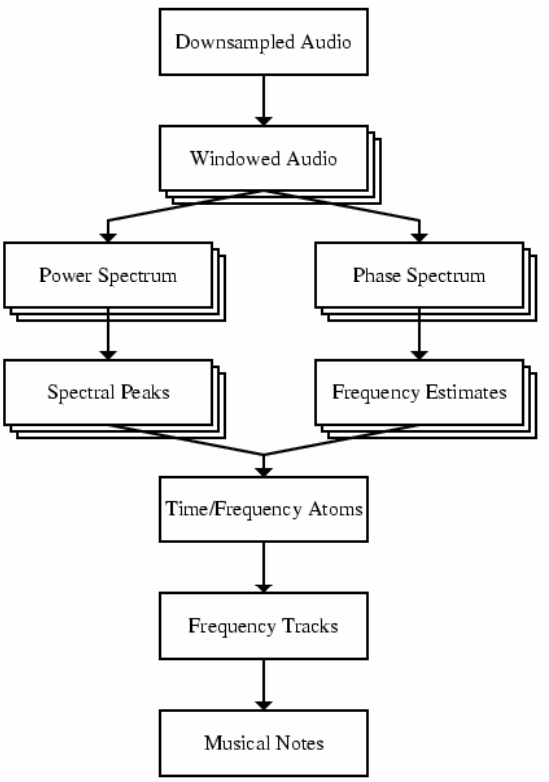
JTRANSCRIBER

The goal of an automatic music transcription system is to create, from an audio recording, some form of symbolic notation (usually common music notation) representing the piece that was played. For classical music, this should be the same as the score from which the performer played the piece. There are several reasons why this goal can never be fully reached, for example, that there is no one-to-one correspondence between scores and performances, and that masking makes it impossible to measure everything that occurs in a musical performance. Recent attempts at transcription report note detection rates around 90% for solo piano music (Dixon, 2000a ; Klapuri, 1998; Marolt, 2001), which is sufficient to be somewhat useful to musicians.

A full transcription system is normally conceptualised in two stages: the signal processing stage, in which the pitch and timing of all notes is detected, producing a symbolic representation (often in MIDI format), and the notation stage, in which the symbolic data is interpreted in musical terms and presented as a score. This second stage involves tasks such as finding the key signature and time signature, following tempo changes, quantising the onset and offset times of the notes, choosing suitable enharmonic spellings for notes, assigning notes to voices in polyphonic passages, and finally laying out the musical symbols on the page. Here, we address only the first stage of the problem, detecting the pitch and timing of all notes, or in more concrete terms, converting audio data to MIDI.

The data is processed according to Figure 3: The audio data is averaged to a single channel and downsampled to increase processing speed. A short time Fourier transform (STFT) is used to create a time-frequency image of the signal, with the user selecting the type, size and spacing of the

Figure 3. Data processing steps in JTranscriber



windows. Using a technique developed for the phase vocoder (Flanagan & Golden, 1966), a more accurate estimate of the sinusoidal energy in each frequency bin can be calculated from the rate of change of phase in each bin.

The next step is to calculate the peaks in the magnitude spectrum, and to combine the frequency estimates to give a set of time-frequency atoms, which represent packets of energy localised in time and frequency. These are then combined with the atoms from neighbouring frames (time slices), to create a set of frequency tracks, representing the partials of musical notes. Frequency tracks are assigned to musical notes by estimating the most likely set of fundamental frequencies that would give rise to the observed tracks, and the pitch, onset time, duration and amplitude of each note are estimated from its constituent partials.

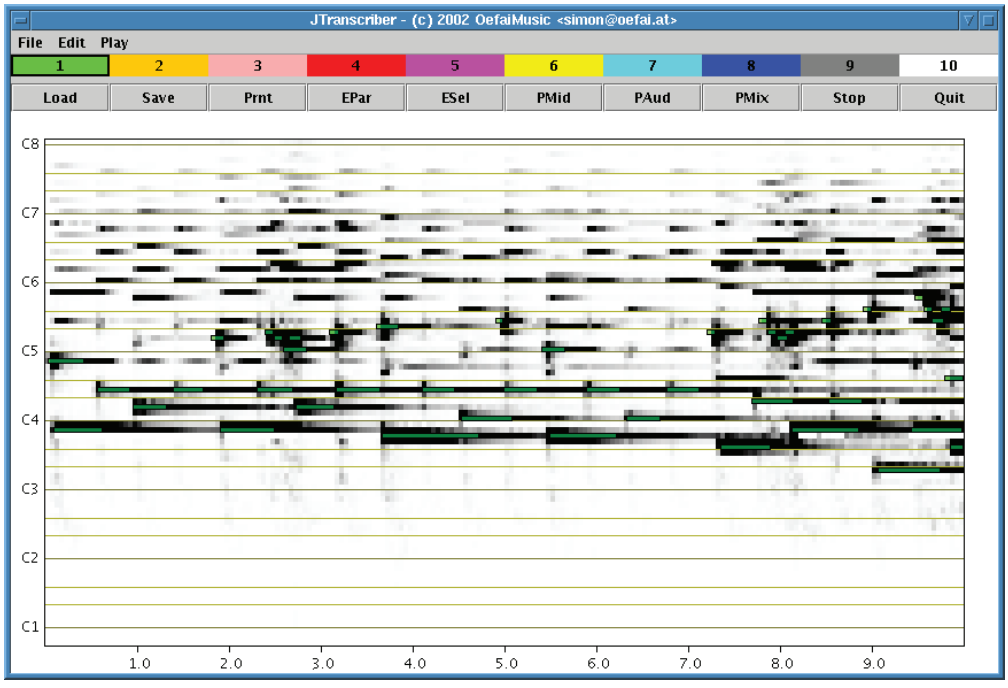
An example of the output is displayed in Figure 4, showing a spectrogram representation of the signal using a logarithmic frequency scale, labelled with the corresponding musical note names, and the transcribed notes superimposed over the spectrogram in piano roll notation. (The piano roll notation is colour and partially transparent, whereas the spectrogram is black and white, which makes the data easily distinguishable on the screen. In the grey-scale diagram, the coloured notes are difficult to see; here they are surrounded by a solid frame to help identify them.) An interactive editing system allows the user to correct any errors made by the automatic transcription system, and also to assign notes to different voices (different colours) and insert high level musical structure information. It is also possible to listen to the original and reconstructed signals (separately or simultaneously) for comparison.

An earlier version of the transcription system was written in C++, however the current version is implemented entirely in Java. The system was tested on a large database of solo piano music consisting of professional performances of 13 Mozart piano sonatas, or around 100,000 notes (Dixon, 2000a), with the results that approximately 10-15% of the notes were missed, and a similar number of the reported notes were false. The most typical errors made by the system are thresholding errors (discarding played notes because they are below the threshold set by the user, or including spurious notes which are above the given threshold) and octave errors (or more generally, where a harmonic of one tone is taken to be the fundamental of another, and vice versa).

THE PERFORMANCE WORM

Skilled musicians communicate high-level information such as musical structure and emotion when they shape the music by the continuous modulation of aspects such as tempo and loudness. That is, artists go beyond what is prescribed in the score, and express their interpretation of the music and their individuality by varying certain musical parameters within

Figure 4. Transcription of the opening 10s of the second movement of Mozart's Piano Sonata K332. The transcribed notes are superimposed over the spectrogram of the audio signal (see text). It is not possible to distinguish fundamental frequencies from harmonics of notes merely by viewing the spectrogram.



acceptable limits. This is referred to as expressive music performance, and is an important part of Western art music, particularly classical music. The Performance Worm (Dixon et al., 2002) is a real-time system for tracking and visualising the tempo and dynamics of a performance in an appealing graphical format which provides insight into the expressive patterns applied by skilled artists. This representation also forms the basis for automatic recognition of performers' style (Widmer, 2002; Widmer et al., 2003).

The system takes input from the sound card (or from a file), and measures the dynamics and tempo, displaying them as a trajectory in a 2-dimensional performance space (Langner & Goebl, 2002, 2003). The measurement of dynamics is straightforward: It can be calculated directly as the RMS energy expressed in decibels, or, by applying a standard psychoacoustic calculation (Zwicker & Fastl, 1999), the perceived loudness can be computed and expressed in sones. The difficulty lies in creating a tempo tracking system which is robust to timing perturbations yet responsive to changes in tempo. This is performed by an adaptation of the tempo induction subsystem of BeatRoot, modified to work in real time. The major difference is the online IOI clustering algorithm, which continuously outputs a tempo estimate based only on the musical data up to the time of processing. The clustering algorithm finds groups of IOIs of similar duration in the most recent eight seconds of music, and calculates a weighted average IOI representing the tempo for each cluster. The tempo estimates are adjusted to accommodate information from musically-related clusters, and then smoothed over time by matching each cluster with previous tempo hypotheses. Figure 5 shows the development over time of the highest ranked tempo hypothesis with the corresponding dynamics.

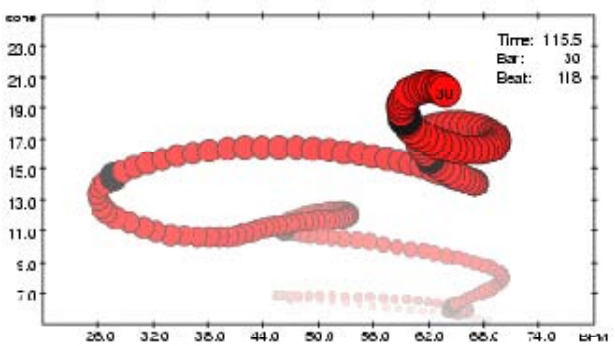
The Performance Worm is implemented in about 4,000 lines of Java, and runs in real time on standard desktop computers. The graphical user interface provides buttons for scaling and translating the axes, selecting the metrical level, setting parameters, loading and saving files, and playing, pausing and stopping the animation.

Apart from the real-time visualisation of performance data, the Worm can also load data from other programs, such as the more accurate beat tracking data produced by BeatRoot. This function enables the accurate comparison of different performers playing the same piece, in order to characterise the individual interpretive style of the performer. Current investigations include the use of AI pattern matching algorithms to learn to recognize performers by the typical trajectories that their playing produces.

FUTURE TRENDS

Research in music content analysis is progressing rapidly, making it difficult to summarise the various branches of investigation. One major initiative addresses the possibility of interacting with music at the semantic level, which involves the automatic generation of metadata, using machine learning and data mining techniques to discover relationships between low-level features and high-level concepts. Another important trend is the automatic computation of

Figure 5. Screen shot of the performance worm showing the trajectory to bar 30 of Rachmaninov's Prelude op.23 no.6 played by Vladimir Ashkenazy. The horizontal axis shows tempo in beats per minute, and the vertical axis shows loudness in sones. The most recent points are largest and darkest; the points shrink and fade into the background as the animation proceeds.



musical similarity for organising and navigating large music collections. For other developments in this area, interested readers are referred to the web site at:

http://www.semanticaudio.org

CONCLUSION

The three systems discussed are research prototypes, whose performance could be improved in several ways, for example, by specialisation to suit music of a particular style or limited complexity, or by the incorporation of high-level knowledge of the piece being analysed.

This is particularly relevant to performance research, where the musical score is usually known. By supplying a beat tracking or performance analysis system with the score, most ambiguities are resolved, giving the possibility of a fully automatic and accurate analysis.

Both dynamic programming and Bayesian approaches have proved successful in score following (e.g., for automatic accompaniment, Raphael, 2001), and it is likely that one of these approaches would also be adequate for our purposes. A more complex alternative would be a learning system which automatically extracts the high-level knowledge required for the system to fine-tune itself to the input data (Dixon, 1996). In any case, the continuing rapid growth in computing power and processing techniques ensures that content-based analysis of music will play an increasingly important role in many areas of human interaction with music.

ACKNOWLEDGMENT

The Austrian Research Institute for Artificial Intelligence ackowledges the financial support of the Austrian Federal Ministry for Education, Science and Culture (BMBWK) and the Austrian Federal Ministry for Transport, Innovation and Technology (BMVIT). This work was supported by the START programme (project Y99-INF) of the BMBWK. Special thanks to the Bösendorfer Company, Vienna, for performance data used in this work.

REFERENCES

Allen, P., & Dannenberg, R. (1990). Tracking musical beats in real time. In *Proceedings of the International Computer Music Conference* (pp. 140-143), San Francisco CA. International Computer Music Association.

Cemgil, A., Kappen, B., Desain, P., & Honing, H. (2000). On tempo tracking: Tempogram representation and Kalman filtering. In *Proceedings of the 2000 International Computer Music Conference* (pp. 352-355), San Francisco CA. International Computer Music Association.

Chafe, C., Jaffe, D., Kashima, K., Mont-Reynaud, B., & Smith, J. (1985). Techniques for note identification in polyphonic music. In *Proceedings of the International Computer Music Conference* (pp. 399-405), San Francisco CA. International Computer Music Association.

Desain, P. (1993). A connectionist and a traditional AI quantizer: Symbolic versus sub-symbolic models of rhythm perception. *Contemporary Music Review, 9*, 239-254.

Desain, P., & Honing, H. (1989). Quantization of musical time: A connectionist approach. *Computer Music Journal, 13*(3), 56-66.

Dixon, S. (1996). A dynamic modelling approach to music recognition. In *Proceedings of the International Computer Music Conference* (pp. 83-86), San Francisco CA. International Computer Music Association.

Dixon, S. (2000a). Extraction of musical performance parameters from audio data. In *Proceedings of the First IEEE Pacific-Rim Conference on Multimedia* (pp. 42-45), Sydney. University of Sydney.

Dixon, S. (2000b). On the computer recognition of solo piano music. *Mikropolyphonie*, 6. *http://www.mikro pol. net/volume6*

Dixon, S. (2001a). Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research, 30*(1), 39-58.

Dixon, S. (2001b). An empirical comparison of tempo trackers. In *Proceedings of the 8th Brazilian Symposium on Computer Music* (pp. 832-840). Brazilian Computing Society.

Dixon, S. (2001c). An interactive beat tracking and visualisation system. In *Proceedings of the International Computer Music Conference* (pp. 215-218), San Francisco CA. International Computer Music Association.

Dixon, S. (2004). Analysis of musical content in digital audio. In J. DiMarco (Ed.), *Computer graphics and multimedia: Applications, problems and solutions* (pp. 214-235). Hershey, PA: Idea Group Publishing.

Dixon, S., Goebl, W., & Widmer, G. (2002). Real time tracking and visualisation of musical expression. In *Music and Artificial Intelligence: Second International Conference, ICMAI2002* (pp. 58-68), Edinburgh, Scotland. Springer.

Eck, D. (2000). Meter through synchrony: Processing rhythmical patterns with relaxation oscillators. PhD thesis, Indiana University, Department of Computer Science.

Flanagan, J., & Golden, R. (1966). Phase vocoder. *Bell System Technical Journal, 45*, 1493-1509.

Goebl, W., & Dixon, S. (2001). Analysis of tempo classes in performances of Mozart sonatas. In *Proceedings of VII International Symposium on Systematic and Comparative Musicology and III International Conference on Cognitive Musicology* (pp. 65-76), Jyväskylä, Finland. University of Jyväskylä.

Gómez, E., Klapuri, A., & Meudic, B. (2003). Melody description and extraction in the context of music content processing. *Journal of New Music Research, 32*(1), 23-41.

Goto, M., & Hayamizu, S. (1999). A real-time music scene description system: Detecting melody and bass lines in audio signals. In Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis (pp. 31-40). *International Joint Conference on Artificial Intelligence*.

Goto, M., & Muraoka, Y. (1995). A real-time beat tracking system for audio signals. In *Proceedings of the International Computer Music Conference* (pp. 171-174), San Francisco, CA. International Computer Music Association.

Goto, M., & Muraoka, Y. (1999). Real-time beat tracking for drumless audio signals. *Speech Communication, 27*(3-4), 311-335.

Kashino, K., Nakadai, K., Kinoshita, T., & Tanaka, H. (1995). Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitative information integration mechanism. In C.S. Mellish (Ed.), *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 158-164), Montréal, Canada: Morgan Kaufmann.

Klapuri, A. (1998). *Automatic transcription of music*. Master's Thesis, Tampere University of Technology, Department of Information Technology.

Klapuri, A. (2003). Automatic transcription of music. In R. Bresin (Ed.), *Proceedings of the Stockholm Music Acoustics Conference* (pp. 587-590).

Klapuri, A., Virtanen, T., & Holm, J.-M. (2000). Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals. In *Proceedings of the COST-G6 Conference on Digital Audio Effects*, Verona, Italy.

Langner, J., & Goebl, W. (2002). Representing expressive performance in tempo-loudness space. In *Proceedings of the ESCOM 10th Anniversary Conference on Musical Creativity*, Liége, Belgium.

Langner, J., & Goebl, W. (2003). Visualizing expressive performance in tempo-loudness space. *Computer Music Journal, 27*(4), 69-83.

Large, E., & Kolen, J. (1994). Resonance and the perception of musical meter. *Connection Science, 6*,177-208.

Longuet-Higgins, H. (1987). *Mental processes*. Cambridge, MA: MIT Press.

Marolt, M. (1997). A music transcription system based on multiple-agents architecture. In *Proceedings of Multimedia and Hypermedia Systems Conference MIPRO'97*, Opatija, Croatia.

Marolt, M. (1998). Feedforward neural networks for piano music transcription. In *Proceedings of the XIIth Colloquium on Musical Informatics* (pp. 240-243), Gorizia, Italy. Associazione di Informatica Musicale Italiana.

Marolt, M. (2001). SONIC: Transcription of polyphonic piano music with neural networks. In *Proceedings of the Workshop on Current Directions in Computer Music Research* (pp. 217-224), Barcelona, Spain. Audiovisual Institute, Pompeu Fabra University.

Martin, K. (1996). *A blackboard system for automatic transcription of simple polyphonic music*. Technical Report 385, Massachussets Institute of Technology Media Laboratory, Perceptual Computing Section.

Mont-Reynaud, B. (1985). Problem-solving strategies in a music transcription system. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 916-919), Los Angeles, CA: Morgan Kaufmann.

Moorer, J. (1975). *On the segmentation and analysis of continuous musical sound by digital computer.* PhD Thesis, Stanford University, CCRMA.

Piszczalski, M., & Galler, B. (1977). Automatic music transcription. *Computer Music Journal, 1*(4), 24-31.

Raphael, C. (2001). Synthesizing musical accompaniments with Bayesian belief networks. *Journal of New Music Research, 30*(1), 59-67.

Roads, C. (1996). *The computer music tutorial*. Cambridge, MA: MIT Press.

Rosenthal, D. (1992). Emulation of human rhythm perception. *Computer Music Journal, 16*(1), 64-76.

Scheirer, E. (1995). *Extracting expressive performance information from recorded music*. Master's Thesis, Massachusetts Institute of Technology, Media Laboratory.

Scheirer, E. (1998). Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustical Society of America, 103*(1), 588-601.

Schloss, W. (1985). *On the automatic transcription of percussive music: From acoustic signal to high level analysis.* PhD Thesis, Stanford University, CCRMA.

Sterian, A. (1999). *Model-based segmentation of time-frequency images for musical transcription.* PhD Thesis, University of Michigan, Department of Electrical Engineering.

Wang, A. (2003). An industrial strength audio search algorithm. In *4th International Conference on Music Information Retrieval (ISMIR 2003)* (pp. 7-13).

Watson, C. (1985). The computer analysis of polyphonic music. PhD thesis, University of Sydney, Basser Department of Computer Science.

Widmer, G. (2002). In search of the Horowitz factor: Interim report on a musical discovery project. In *Proceedings of the 5th International Conference on Discovery Science* (pp. 13-32), Berlin: Springer.

Widmer, G., Dixon, S., Goebl, W., Pampalk, E., & Tobudic, A. (2003). In search of the Horowitz factor. *AI Magazine, 24*(3), 111-130.

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and models* (2nd ed.). Berlin: Springer.

KEY TERMS

Automatic Transcription: The process of extracting the musical content from an audio signal and representing it in standard music notation.

Beat Tracking: The process of finding the times of musical beats in an audio signal, including following tempo changes, similar to the way that people tap their feet in time to music.

Clustering Algorithm: An algorithm which sorts data into groups of similar items, where the category boundaries are not known in advance.

Frequency Domain: The representation of a signal as a function of frequency, for example as the sum of sinusoidal waves of different amplitudes and frequencies.

Music Content Analysis: The analysis of an audio signal in terms of higher-level (cognitive) properties such as melody, harmony and rhythm, or in terms of a description of the signal's component sounds and the sound sources which generated them.

Music Information Retrieval: The research field concerning the automation of access to music information through the use of digital computers.

Onset Detection: The process of finding the start times of notes in an audio signal.

Time Domain: The representation of a signal, such as the amplitude or pressure of a sound wave, as a function of time.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 188-196, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Authentication Methods for Computer Systems Security

Zippy Erlich
The Open University of Israel, Israel

Moshe Zviran
Tel-Aviv University, Israel

INTRODUCTION

With the rapid growth of networked systems and applications such as e-commerce, the demand for effective computer security is increasing. Most computer systems are protected through a process of user identification and authentication. While identification is usually non-private information provided by users to identify themselves and can be known by system administrators and other system users, authentication provides secret, private user information which can authenticate their identity. There are various authentication approaches and techniques, from passwords to public keys (Smith, 2002).

This article presents the three main authentication approaches, their technology and implementation issues, and the factors to be considered when choosing an authentication method.

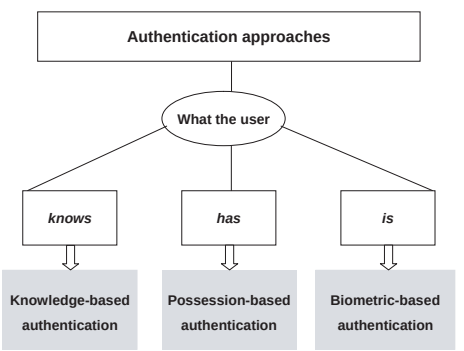
BACKGROUND

Even before computers came along, a variety of distinguishing characteristics were used to authenticate people. Computer systems have applied these characteristics for user authentication. The authentication approaches can be classified into three types according to the distinguishing characteristics they use (Menkus, 1988), as presented in Figure 1:

- What the user *knows*—knowledge-based authentication (e.g., password, PIN, pass code)
- What the user *has*—possession-based authentication (e.g., memory card and smart card tokens)
- What the user *is*—biometric-based authentication: physiological (e.g., fingerprint) or behavioral (e.g., keyboard dynamics) characteristics

As all these authentication types have benefits and drawbacks, trade-offs need to be made among security, ease of use, and ease of administration. Authentication types can be implemented alone or in combination. To strengthen the authentication process, the use of at least two types is recommended. Multiple layers of different types of authentication provide substantially better protection.

Figure 1. Classification of authentication methods

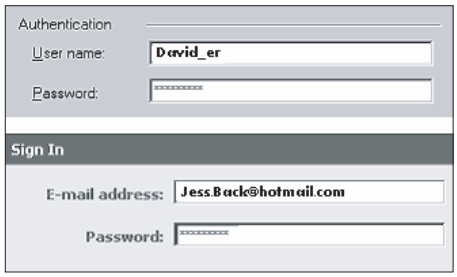


KNOWLEDGE-BASED AUTHENTICATION

The most widely used type of authentication is knowledge-based authentication. Examples of knowledge-based authentication include passwords, pass phrases, or pass sentences (Spector & Ginzberg, 1994), graphical passwords (Thorpe & Van Oorschot, 2004; Wiedenbeck, Waters, Birget, Brodskiy, & Memon, 2005), pass faces (Brostoff & Sasse, 2000) and personal identification numbers (PINs). To verify and authenticate users over an unsecured public network, such as the Internet, digital certificates and digital signatures are used. They are provided using a public key infrastructure (PKI) which consists of a public and a private cryptographic key pair (Adams & Lloyd, 1999).

The traditional, and by far the most widely used, form of authentication based on user knowledge is the password (Zviran & Haga, 1993). Most computer systems are protected through user identification (like user name or user e-mail address) and a password, as shown in Figure 2.

Figure 2. Authentication through user identification and password



A password is conceptually simple for both system designers and end users. It consists of a secret series of characters according to some predefined rules. The user ID and password pair acts as user identification and authentication and serves to block unauthorized access to computing resources. In most systems, it can provide effective protection if used correctly.

However, passwords are known to suffer from a number of pitfalls due to human information processing limitations (Sasse, Brostoff, & Weirich, 2001; Yan, Blackwell, Anderson, & Grant, 2005). First, there is a trade-off between memorability and security. Passwords should be difficult to guess and easy to remember. The fact that difficult-to-guess and difficult-to-crack passwords are difficult to remember and that easy to remember passwords are easy to guess and easy to crack poses a dilemma for the generation of passwords. The most secure password is a random string of characters. Such passwords are difficult to guess by others, but at the same time are difficult to remember and thus compel the users to write them down, which impairs their secrecy. Moreover, most users have multiple passwords for different systems and applications, forcing them to remember several passwords. In order to help them to remember the passwords, they usually choose meaningful strings such as names, nicknames, or initials (Adams & Sasse, 1999), which are easy to remember but also easy to crack. They also tend to duplicate their passwords and thus cause the domino effect of password reuse (Ives, Walsh, & Schneider, 2004); namely, all the systems with the same password are no more secure than the weakest system using this password.

In order to improve password security and protect it from dictionary and brute force attacks, password policy should implement rules for choosing and maintaining passwords (Smith, 2002). The major rules are:

- Non-dictionary and no-name passwords.
- Long enough passwords with mixed types of characters.
- Password ageing and not reusing.
- Complex passwords using acronyms, rhymes, and mnemonic phrases, which are difficult to guess and easy to remember (Carstens, McCauley-Bell, Malone, & DeMara, 2004; Yan et al., 2005).
- Passwords should not be shared and should not be written down.
- The number of unsuccessful authentication attempts should be limited by the system.
- Passwords should never be stored in clear text; they should be encrypted or hashed.

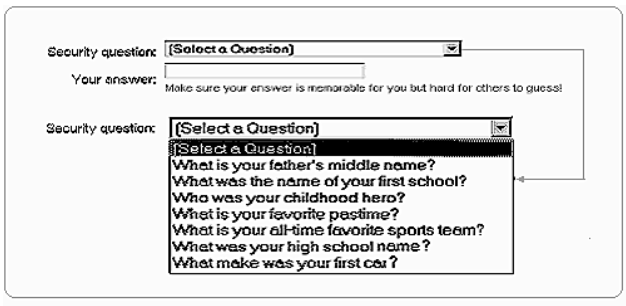
Passwords based on the aforementioned rules are more effective, more difficult to identify and to determine by cracking utilities. To overcome the problem of sniffing passwords when authentication is performed over the Internet, one-time passwords are used. The one-time password can be implemented using smart cards—a kind of possession-based authentication discussed hereafter.

Passwords, used as the first level of authentication, that allow access to information system resources through operating systems are commonly referred to as *primary passwords*. Passwords used as the second level of authentication, for further control and protection of multi-level access to segments of these resources, such as sensitive applications or data files, are commonly referred to as *secondary passwords* (Zviran & Haga, 1993).

In determining primary passwords, the operating system manufacturer uses system-generated passwords or user-generated passwords with predefined rules. User-generated passwords are shown to be easier to remember but less secure than system-generated passwords as they can be easily guessed (Lopez, Oppliger, & Pernul, 2004).

In order to overcome the difficulty of remembering passwords, a *question-and-answer password* method has been suggested (Haga & Zviran, 1991). This method is mainly used for secondary passwords. It involves a dialogue between the user and the system, as shown in Figure 3.

Figure 3. Example of a question-and-answer password



In a typical question-and-answer session, the user is presented with several randomly selected brief questions from a set of questions stored in his or her profile in the operating system. Access to a system or to a particular application is granted only upon a match between the user's answers and those stored in his/her profile.

The two main types of question-and-answer passwords are: *cognitive passwords* and *associative passwords* (Bunnell, Podd, Henderson, Napier, & Kennedy-Moffat, 1997; Haga & Zviran, 1991; Zviran & Haga, 1993). In cognitive passwords, the user must provide the system with answers to personal fact-based or opinion-based questions, such as the user's mother's maiden name (fact-based) or user's favorite type of music (opinion-based).

In associative passwords, the user must provide the system with a set of word associations, consisting of both cues and their unique associated responses.

POSSESSION-BASED AUTHENTICATION

Possession-based authentication, referred to also as token-based authentication, is based on what the user has. It makes use mainly of physical objects that a user possesses, like tokens. Aside from the fact that presentation of a valid token does not prove ownership, as it may have been stolen or duplicated by some sophisticated fraudulent means (Svigals, 1994), there are problems of administration and of the inconvenience to users of having to carry them.

Tokens are usually divided into two main groups: *memory tokens* and *smart tokens*. Memory tokens store information but do not process it. The most common type of memory token is the magnetic card, used mainly for authentication together with a knowledge-based authentication mechanism such as a PIN. Memory tokens are inexpensive to produce. Using them with PINs provides significantly more security than PINs or passwords alone.

Unlike memory tokens, smart tokens incorporate one or more embedded integrated circuits which enable them to process information. Like memory tokens, most smart tokens are used for authentication together with a knowledge-based authentication mechanism such as a PIN. Of the various types of smart tokens, the most widely used are those that house an integrated chip containing a microprocessor. Their portability and cryptographic capacity have led to their wide use in many remote and e-commerce applications (Juang, 2004; Ku & Chen, 2004; Wu & Chieu, 2003). Due to their complexity, smart tokens are more expensive than memory tokens but provide greater flexibility and security and are more difficult to forge. Because of their high security level, smart tokens are also used for one-time passwords for authentication across open networks.

BIOMETRIC-BASED AUTHENTICATION

Biometric-based authentication is based on what the user is, namely, automatic identification using certain anatomical, physiological or behavioral features and characteristics associated with the user (Kim, 1995; Wayman, Jain, Maltoni, & Maio, 2004). Biometric authentications are based on the fact that certain physiological or behavioral characteristics reliably distinguish one person from another. Thus, it is possible to establish an identity based on who the user is, rather than on what the user possesses or knows and remembers. Biometrics involves both the collection and the comparison of these characteristics. A biometric system can be viewed as a pattern recognition system consisting of three main modules: (1) the sensor module, (2) the feature extraction module, and (3) the feature matching module. The users' personal attributes are captured and stored in reference files to be compared for later authentication to determine if a match exists. The accuracy of the different biometric systems can be evaluated by the measurement of two types of errors (Matyas & Stapleton, 2000): (1) erroneous rejection, that is, false non-match (type I error), and (2) erroneous acceptance, that is, false match (type II error). In a biometric system that provides a high level of authentication, the rate of these two errors is low.

Biometric authentications are technically complex and usually expensive as they require special hardware. Although all biometric technologies inherently suffer from some level of false match or false non-match, they have a high level of security. Despite their high security, they do not have a high acceptance rate by users as they are perceived to be intrusive and an encroachment on privacy (Prabhakar, Pankanti, & Jain, 2003) through automated means. They also raise ethical issues of potential misuse of personal biometrics such as for tracking and monitoring productivity (Alterman, 2003). Thus, they are not popular and mainly used in systems with very high levels of security.

The emergence of biometric authentication addressed the problems that plagued traditional verification methods, providing the most effective and accurate identification method with an edge over traditional security methods in that it cannot be easily stolen or shared. Biometric systems also enhance user convenience by alleviating the need to determine and remember passwords. However, while convenient, the digital scan or pattern is vulnerable to network analysis and once stolen, can no longer be used (Ives et al., 2004).

There are various kinds of biometrics (Matyas & Stapleton, 2000) and they are usually divided into two main categories: physiological and behavioral biometrics. Physiological biometrics are based on the user's stable physical attributes. The best known are fingerprints, finger scans, hand geometry, iris scans, retina scans, and facial scans. Fingerprints are the most widely used physiological characteristic in systems that automatically recognize a user's identity (Ratha

& Bolle, 2005; Wayman et al., 2004). An example of one of its up-to-date applications is the use of fingerprint-based identification and authentication to support online, Web-based course examinations (Auernheimer & Tsai, 2005).

Behavioral biometrics are based on users' behavioral attributes that are learned movements (Guven & Sogukpinar, 2003; O'Gorman, 2003; Yu & Cho, 2004). The best known are: keystroke dynamics, signature dynamics, mouse dynamics, and speech or voice verification.

FACTORS IN CHOOSING AN AUTHENTICATION METHOD

In choosing an authentication method a number of factors need to be considered: effectiveness, ease of implementation, ease of use, and user attitude and acceptance (Furnell, Dowland, Illingworth, & Reynolds, 2000). Table 1 shows the ranking of the three authentication types according to these four factors.

The knowledge-based authentication type is inexpensive and easy to implement and change. Unfortunately, it is also the easiest to compromise and is less secure than tokens or biometric-based authentication methods, which are inherently more secure. On the other hand, tokens and biometric-authentication methods are more expensive to implement. User attitudes towards knowledge-based authentication are highly positive, less positive towards possession-based authentication, and negative towards biometric-based authentication (Deane, Barrelle, Henderson, & Mahar, 1995; Prabhakar et al., 2003). As knowledge-based authentication is less effective than the other types, it is recommended that it be used in two-type authentication; for example, a password and a token or a password and a keystroke (Furnell, Papadopoulos, & Dowland, 2004; Yu & Cho, 2004).

Passwords provide the most cost effective solution, they are portable to other applications, easy to deploy and scale to an unlimited number of users. They are integrated into many operating systems and users are familiar with them. Improved security can be achieved with a secondary technique, like a cognitive or associative password.

Possession-based authentication using tokens provides higher security than knowledge-based authentication. Most of the problems associated with tokens relate to their cost, administration, loss, and user dissatisfaction. Because of vulnerability to theft, a token should not be used alone but with another authentication type. When combined with a password, it can provide major advantages as it can store or generate multiple passwords, and the user has to remember only the single password needed to access the token.

Table 1. Ranking of authentication types from 1 (low) to 3 (high)

	Factor			
Authentication type	Security effectiveness	Ease of implementation	Ease of use	User attitude and acceptance
Knowledge-based	1	3	3	3
Possession-based	2	2	2	2
Biometric-based	3	1	1	1

FUTURE TRENDS

Since passwords are comparatively inexpensive, simple to use, and attractive to users, they will probably continue to be the most widely used form of authentication in the foreseeable future. When properly managed and controlled, they can provide effective security. Further research is needed on the three authentication types to improve and enhance their technologies and to enable the design of more usable and effective security systems.

CONCLUSION

Computer systems are protected by three main types of authentication approaches: (1) knowledge-based, (2) possession-based, and (3) biometric-based. Each of these has both benefits and drawbacks. When choosing an authentication method we have to consider the trade-off among security effectiveness, ease of implementation, ease of use, and user attitude and acceptance. In order to strengthen the authentication process, the use of at least two-type authentication is recommended.

Overall, there is no one best solution to the user authentication problems. Multiple layers of protection provide substantially better security.

REFERENCES

Adams, A., & Sasse, M. A. (1999). Users are not the enemy: Why users compromise security mechanisms and how to take remedial measures. *Communications of the ACM, 42*(12), 40-46.

Adams, C., & Lloyd, S. (1999). *Understanding public-key infrastructure: Concepts, standards and deployment considerations*. Indianapolis, IN: Macmillan Technical.

Alterman, A. (2003). A piece of yourself: Ethical issues in biometric identification. *Ethics and Information Technology, 5*(3), 139-150.

Auernheimer, B., & Tsai, M. J. (2005). Biometric authentication for Web-based course examinations. In *Proceedings of the 38th Annual Hawaii International Conference on System Science (HICSS'05)* (pp. 294b). Washington, DC: IEEE Computer Society.

Brostoff, S., & Sasse, M. A. (2000). Are passfaces more usable than passwords? A field trial investigation. In S. McDonald, Y. Waern, & G. Cockton (Eds.), *People and computers XIV—Usability or else! Proceedings of HCI2000* (pp. 405-424). Sunderland, UK: Springer.

Bunnell, J., Podd, J., Henderson, R., Napier, R., & Kennedy-Moffat, J. (1997). Cognitive, associative and conventional passwords: Recall and guessing rates. *Computers & Security, 16*(7), 629-641.

Carstens, D. S., McCauley-Bell, P. R., Malone, L. C., & DeMara, R. F. (2004). Evaluation of the human impact of password authentication practices on information security. *Information Science Journal, 7*(1), 67-85.

Deane, F., Barrelle, K., Henderson, R., & Mahar, D. (1995). Perceived acceptability of biometric security systems. *Computers & Security, 14*(3), 225-231.

Furnell, S. M., Dowland, P. S., Illingworth, H. M., & Reynolds, P. L. (2000). Authentication and supervision: A survey of user attitudes. *Computers & Security, 19*(6), 529-539.

Furnell, S. M., Papadopoulos, I., & Dowland, P. S. (2004). A long-term trial of alternative user authentication technologies. *Information Management and Computer Security, 12*(2), 178-190.

Guven, A., & Sogukpinar, I. (2003). Understanding users' keystroke patterns for computer access security. *Computers & Security, 22*(8), 695-706.

Haga, W. J., & Zviran, M. (1991). Question-and-answer passwords: An empirical evaluation. *Information Systems, 16*(3), 335-343.

Ives, B., Walsh, K. R., & Schneider, H. (2004). The domino effect of password reuse. *Communications of the ACM, 47*(4), 75-78.

Juang, W. S. (2004). Efficient password authenticated key agreement using smart cards. *Computers & Security, 23*(2), 167-173.

Kim, H. J. (1995). Biometrics, is it a viable proposition for identity authentication and access control? *Computers & Security, 14*(3), 205-214.

Ku, W. C., & Chen, S. M. (2004). Weaknesses and improvements of an efficient password based user authentication scheme using smart cards. *IEEE Transactions on Consumer Electronics, 50*(1), 204-207.

Lopez, J., Oppliger, R., & Pernul, G. (2004). Authentication and authorization infrastructures (AAIs): A comparative survey. *Computers & Security, 23*(7), 578-590.

Matyas, S. M., & Stapleton, J. (2000). A biometric standard for information management and security. *Computers & Security, 19*(5), 428-441.

Menkus, B. (1988). Understanding the use of passwords. *Computers & Security, 7*(2), 132-136.

O'Gorman, L. (2003). Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE, 91*(12), 2019-2040.

Prabhakar, S., Pankanti, S., & Jain, A. K. (2003). Biometric recognition: Security and privacy concerns. *IEEE Security and Privacy Magazine, 1*(2), 33-42.

Ratha, N., & Bolle, R. (Eds.). (2005). *Automatic fingerprint recognition systems*. New York: Springer Verlag.

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the 'weakest link': A human/computer interaction approach to usable and effective security. *BT Technology Journal, 19*(3), 122-131.

Smith, R. E. (2002). *Authentication: From passwords to public keys*. Boston: Addison-Wesley.

Spector, Y., & Ginzberg, J. (1994). Pass-sentence: A new approach to computer code. *Computers & Security, 13*(2), 145-160.

Svigals, J. (1994). Smartcards: A security assessment. *Computers & Security, 13*(2), 107-114.

Thorpe, J., & Van Oorschot, P. (2004, August 9-13). Graphical dictionaries and the memorable space of graphical passwords. In *Proceedings of the 13th USENIX Security Symposium*. San Diego, CA.

Wayman, J., Jain, A. K., Maltoni, D., & Maio, D. (Eds.). (2004). *Biometric systems: Technology, design and performance evaluation*. New York: Springer.

Wiedenbeck, S., Waters, J., Birget, J. C., Brodskiy, A., & Memon, N. (2005). PassPoints: Design and longitudinal evaluation of a graphical password system. *International Journal of Human-Computer Studies, 63*(1-2), 102-127.

Wu, S. T., & Chieu, B. C. (2003). A user friendly remote authentication scheme with smart cards. *Computers & Security, 22*(6), 547-550.

Yan, J., Blackwell, A., Anderson, R., & Grant, A. (2005). The memorability and security of passwords. In L. Cranor & S. Garfinkel (Eds.), *Security and usability: Designing secure systems that people can use* (pp. 121-134). Sebastopol, CA: O'Reilly & Associates.

Yu, E., & Cho, S. (2004). Keystroke dynamics identity verification: Its problems and practical solutions. *Computers & Security, 23*(5), 428-440.

Zviran, M., & Haga, W. J. (1993). A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal, 36*(3), 227-237.

KEY TERMS

Associate Password: A question-and-answer password in which the user provides the system with associated responses to rotating cues.

Authentication: Verifying the identity of the user. There are three main approaches to user authentication: knowledge-based, possession-based, and biometric-based.

Biometric-Based Authentication: An authentication based on what the user is—unique physiological characteristics such as fingerprints or behavioral characteristics such as keyboard dynamics.

Cognitive Password: A question-and-answer password in which the user provides the system with answers to personal, fact-based questions such as the user's mother's maiden name, or opinion-based questions such as the user's favorite type of music.

Identification: The activity of users who supply information to identify themselves, such as name, user name, and user ID.

Knowledge-Based Authentication: An authentication based on what the user knows, such as password, PIN, and pass code.

Password: Knowledge-based authentication consisting of a secret series of characters according to predefined rules. It is the most widely used mechanism of authentication.

Possession-Based Authentication: An authentication based on what the user has, such as memory cards and smart card tokens. Possession-based authentication is also referred to as token-based authentication.

Question-and-Answer Password: A session in which a user is presented with several randomly selected questions from a set of questions stored in the user's profile in the operating system. The user's answers are compared to match with those stored in the profile. The two main types of question-and-answer passwords are cognitive passwords and associative passwords.

Autognomic Intellisite

Jon Ray Hamann
University at Buffalo, State University of New York, Baird Research Park, USA

INTRODUCTION

The 20th century saw the beginning of the evolution of learning machines from the growth of Boolean computers into Bayesian inference machines (Knuth, 2003). For some this is the crux of Artificial Intelligence (AI); however, AI research generally has yielded a plethora of specifically engineered, but formally unrelated, theories/models with varied levels of applications successes/failures, but without a commonly-explicatable conceptual foundation (i.e., it has left a *theory-glut*). Despite these many approaches to AI, including Automated Neural Nets, Natural Language Processing, Genetic Algorithms, Fuzzy Logic and Fractal Mathematical computational approaches, to identify only a few, AI itself has remained an elusive goal to achieve by means of a systems architecture relying on an implementation based on the systemic computer paradigm.

The 21st century experience is overwhelmingly one of an ever-accelerating, dynamically changing world. Just staying in place seems nearly impossible—getting ahead is becoming increasing unfathomable in a world now characterized by an evolving dominance of Information Science and Technology Development in exponentially tighter (shorter) innovation cycles (IBM, 2008). In business, for example, there is the continuous challenge to ensure that the business's products appear obviously differentiated from the competition, while staying current with the never-ending hot new trends that buffet the industry. A prime case in point is that of staying current with the trends in the computer solutions industry since adapting a computer dependent business (and most are) for the *next big trend* can be expected to be mitigated, if not made completely obsolete, by the *next next big trend* already on the radar screen.

BACKGROUND

It is becoming increasingly evident to a growing number of key decision makers that innovation development and management demands a technological assist (Roco & Bainbridge, 2002). This technology, however, must dramatically Augment Human Intelligence in the near future while moving toward a General Autonomous Artificial Intelligence in the longer term (Singularity Institute for Artificial Intelligence, Inc., 2001). Despite the recognition that meeting the demands of accelerating innovation is only likely through advancing AI, which in turn has the potential to impact every aspect of human life, the problem/dilemma for AI developers is that there is no *standard theory of mind*.

To further accentuate this circumstance, the networking of computers has in turn led to the Web with essentially an unlimited growth of data/information (i.e., an *info-glut*). The industry's response, however, to the info-glut problem, has been an ever-growing abundance of Web-access tools, which to an average user seem ironically as only another "glut" (a *technology-glut* or *tool-glut*).

Proposed theories of the Web, like with AI, are also numerous and without a common foundation on which to build a mutual understanding of *AI and the Web*. There are also a plethora of heuristic technological approaches to *AI and the Web* ranging from Intelligizing™ the Web through Learning/Thinking Webs to the Web as a Global (Super) Brain and Virtual Reality as Social Superorganism [See for instance these topics at Principia Cybernetica Web (2008)]. Basically, however, research on *AI and the Web* is categorizable as to whether the focus is on the preeminence of *brain vs. mind* (Roco & Bainbridge, 2002), as for the Human Cognome Project keyed to reverse engineering the human brain**,** or *mind vs. brain*, via a modular description of a general intelligence capable of open-ended recursive self-enhancement (Singularity Institute for Artificial Intelligence (2001), *General Intelligence and Seed AI*) or, alternatively, on the *co-evolution of mind & brain*, characterized by *Project AutoGnome™/CoGnome™/ CogWeb™*, this being the approach of Ai3inc.

The explication of the Web as a Virtual Reality (a computer-based CyberSpace) which is an *image* (sign, symbol, icon)–*communication* system, that is, a *Semiotic* (Goodwin & Queiroz, J., 2007) Relational System, is also of the essence of *Mind*. Ai3inc's long-term focus is on an approach to Synthetic Mind/Artificial Intelligence via a patented technology known as the AutoGnome. This addresses a uniform solution to all of the foregoing problems of glut by way of an Intellisite™, an *Intelligent Website*. The AutoGnome, as an Automated Inference/Inquiry/Intuition software exploiting Mechanized Semiosis, also provides an optimal approach to a General Theory of an Autonomous Virtual Society (Virtuality)—this being an autonomous semiotic universe of Virtual Minds (WebGnomes™); hence Virtuality is related to (Human) Reality through the Virtual Reality of the Web. It is the provision of the foregoing which implicates a *standard theory of mind* that is the focus of As It Is, Inc.'s current development of "*Semiotic* Relational Systems: The

AutoGnome as Synthetic Mind" and "*AutoGnomics* and Intelligent Systems Development" including the present "*AutoGnomic Intellisite"* (Hamann, 2007a).

Relational Systems Foundations

Generally, a canvassing of human experience has reports thereof falling into two fundamental forms—experience of Systems (objects, things, stuff, matter, etc.) and experience of Relations (connections,interactions, functions, transformations, etc.). Historically, this record has been largely confined to a form in which Relations were assumed to exist only between/among Systems, Systems Related to other Systems (SRS'). Between 1963-1968, work was introduced in which Relations were also taken to logically exist both as Relations between/among Systems and other Relations (SRR') and as Relations between/among Relations and other Relations (RR'R''). Based on the presumption of the foregoing and with certain Systems or Relations taking the place of (i.e., imaging (signifying)) other Systems or Relations (this being the notion of image) and with certain Systems or Relations being part of other Systems or Relations (this being the notion of subsumption), the foundation of a Relational Evolutionary paradigm, Relational Systems (RS), was promulgated. (Hamann, 2007b)

From *Image* and *Subsumption* to Mathematics and Logic

First, a Relational Conjecture is restated to form and substantiate the notion of *image*: It is conjectured that the origin of an *image* (or *sign*) system as a chaotic ordering (emergent) event in the evolution of physical/chemical systems is a necessary and (possibly) sufficient condition for the origin of *Life* (and thus Intelligence/*Mind*) (Hamann & Bianchi, 1970).

Second, beginning with the simplest fundamental derivative of the Presumption of *subsumption,* that is, the notion of *distinction* (Spencer-Brown, 1969; Shoup, 2008) whereby there is formed a *boundary* which generates *twoness,* a mathematics of distinction has been created and grown into a general candidate for an approach to a universal language for formal systems, that is, multiboundary mathematics. Inherent to this Boundary Mathematics is a Boundary Logic (from which Boolean Logic is derivable as a special case), which is leading to a more powerful computer design (Bricken, 2007). Generally, taking a *universal formal system* as an axiom system with the property that any other consistent axiom system can be interpreted within it, the mathematics of distinction implies a mathematics of subsumption which, in turn, implies a membership theory as a first step towards a universal language for mathematics (Etter, 2006).

Theory of Mind and of Virtuality

An approach to understanding the "origin" and nature of "mind" is in development based loosely at this point in the process on the System of Boundary Mathematics. This is interpreted as deriving from the Foundations notions of Relational Systems. A theoretical architecture has been posited regarding the formalization of an order (an instantiation of the Mathematics of Subsumption in terms of a degree of partial subsumption) and its derivative calculus, the latter taken as a formulation of the disorder experientially related to the given order, which also implies a reorder(ing) disorder format. Within a Nonseparable System of order/disorder/reorder Relations, this architecture suggests The Form of a meta-theory of theory formation. The Form, in turn, has been invoked in formulating Theories of Intelligence/Mind and Virtuality.

Assume, in a simple, but common instance of the foregoing, that *ordered* experience is *formally signifiable* as a Boolean Network (lattice, algebra, graph or diagram) composed of *points* (nodes, objects, states or Systems) and *lines* (edges, connections, transitions or Relations). Assume further that experience is not totally ordered and that the *disorder* is *formally signifiable* by extending the Boolean Network to the form of a Bayesian Network via a Coxian theory of the algebra of probable inference/inquiry. (Cox, 1961) And finally, assume that *reordering disorder* is *formally signifiable* via the Cox/Jaynes (Jaynes, 2003) form of maximum entropy (maxent) or its generalized probabilistic optimization principal. This approach to modeling both the Web (as a Virtual Reality) and Mind is warranted by the "natural" Network-of-*Images* view of the Web and by the historical predominance of connectionist theories of Mind, and neural-network analyses of mental processes and states.

The resulting synthesis of the foregoing is an approach to Relational Science of Signs, including *signification* and *communication*, that is, a Theory of *Semiotic* Relational Systems**.** This is the necessary basis upon which is built a Theory of Mind and of Virtuality with technologically engineered applications as Synthetic Intelligence(s)/Synthetic Mind and Virtual Reality. (Hamann, 2007a)

AutoGnomic Technology

Based on the work of Charles Sanders Peirce and his successor, Charles Morris, Gene Pendergraft (Pendergraft, 1993), proposed the architecture of a special kind of system, called the AutoGnome, which would be able to perform mechanized (automated) inference using principles derived from semiotics. A venture for the implementation of such an architecture in software code was begun and has resulted in the building of a first release of an AutoGnome System, AutoGnome 01, being a partial implementation of the General AutoGnome Specifications, but representing only about 10-15% of the

complete Conceptual Specification. This first version is a basically a general purpose *pattern generation/recognition/categorization/prediction* engine.

The *Autognome*

General Characterization

The AutoGnome (AG) by its Specification is a General Purpose System of Automated Inference/Inquiry software exploiting a system of Mechanized Semiosis. Unlike most other forms in the mainstream of Artificial Intelligence developments, the AutoGnome is designed to approximate the known semiotic structure and processes of Human Mind. The AutoGnome, to be a complete Semiotic Inference/Inquiry Engine, must account for *The Form* of experience including:

- Ordred (i.e., determined or certain) experience: a formal algebra/logic of semiosis
- Disordered (indeterminate or uncertain) experience: a theory of probable inference/inquiry
- Reordering Disordered experience: via a generalized probabilistic optimization principal

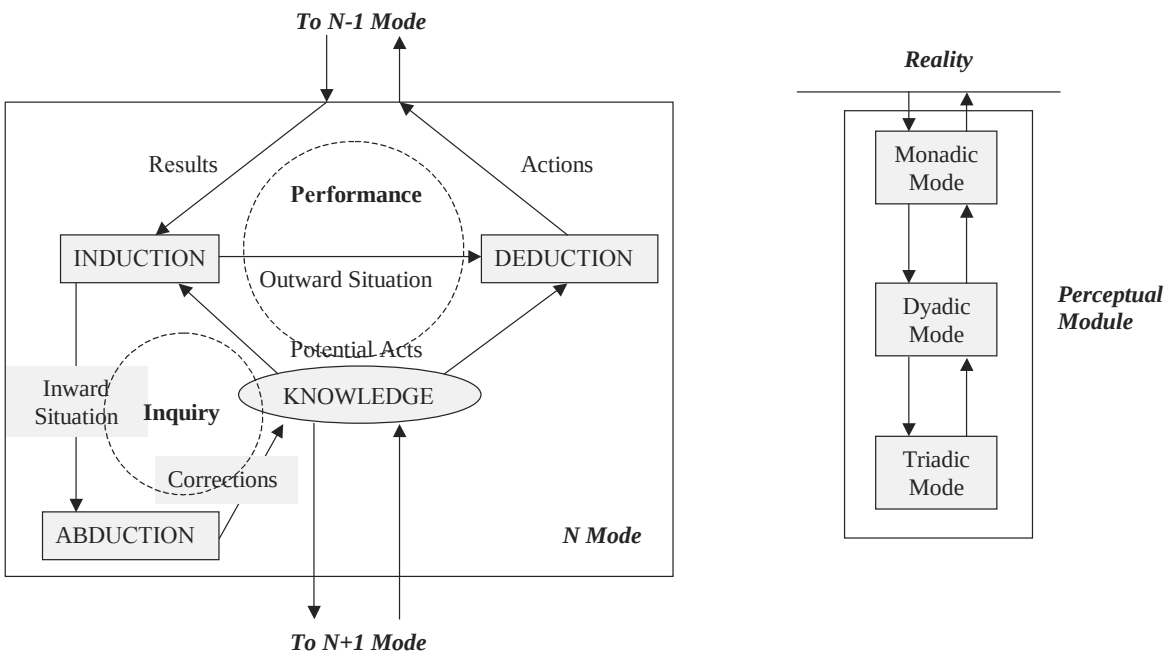
The AutoGnome Architecture

The AutoGnome Architecture (see Figure 1) may be envisioned as multiple modules (perceptual, conceptual, and valuational), each module coding a specific model of the formalisms of semiosis composed of the three modes of semiosis (monadic, dyadic and triadic) and three inferential processes (deduction, induction, and abduction). These recursive inference processes operate on three information stores (an experience store, a knowledge store and a valuation store), gain experience through connective agents (sensors, mediators and effectors (actors)), and function (act) in both an inquiry cycle and a performance cycle.

The probabilistic inference processes integrated formally with the logic of semiosis are the processes of formal representation of the Disorder whereby an AutoGnome identifies and maintains its Identity (Order). The information stores at any particular time are stable states of such probablilistic processes generated by optimizing acts in response to environmental (other system) perturbations of the perceptual module. The form of these optimization procedures for Reordering Disorder are those implementing *Optimum Systemic (subSystemic) Probable Inference* (e.g., MaxEnt).

Note: If the system's ability to perform the Intelligent Act does not depend on the "content" of the inferences (e.g., the three inferential processes do not presuppose what is being reasoned about), then such intelligence can be deemed "generalized". "General Intelligence" is one of the most important design objectives of the AutoGnome and distinguishes it further from other specifically engineered forms of AI.

Figure 1. The AutoGnome – Inference architecture



The Intellisite

The most informative reading of this section is best accomplished while visiting www.truethinker.com which best engenders a realistic sense of an Intellisite

The first application of the AutoGnomic Technology, the Intellisite (an Intelligent WebSite **generically branded as TrueThinker),** is a constructed software environment (a Website) with an embedded form of the AutoGnome known as a WebGnome. MyWebGnome™ then is an intelligent agent residing in this cyberspace environment which, with its continuous adaptive learning from mimicking the user's behavior, will grow into a likeminded replica (*MIND*Clone™) of a user-*self* acting in the Virtual Reality of the Internet with capabilities initially including *knowledge organization* (learning), *knowledge creation* (thinking) and *knowledge applications* (acting) (Hamann, 2007a).

As a homepage "portal to cyberspace", TrueThinker is a premier Knowledge Development Management System.

Intellisite Derivatives

The Intellisite obviously has a broad spectrum of applicability apart from its key functionality as an individual's mirrored Intelligence/Knowledge, in particular as an Autonomous Scientific Intelligence (AScI), a Collective-AutoGnome (CoGnome) and the CogWeb (cognitive network of Intellisites).

AScI

A first generative instantiation of the AutoGnome deriving from its form as a general [meta order(ing)] theory of theory formation is as an Autonomous Scientific Intelligence (AScI) which promises to automate the scientific method (Knuth, 2003).

CoGnome and CogWeb

Assigning a priority MetaQuery/Response status to a selected WebGnome which inter-connects two or more Intellisites in a Network provides a computerized collective intelligence, the CoGnome.

The CogWeb is, by definition, the implementation of the CoGnome for Network Decision-Making by Intellisite-defined groups, organizations, communities, and societies.

While the Intellisite itself was focused on the development of a semiotic engine as an Individual Intelligence, the full potential of individual intellect, be it human or machine, is realized in groups; hence the Automated Community Builder functionality of the Intellisite. This collective creativity, while related to the intelligence of the individual, is actually a feature not only of the Decision Network's inquiry/inference processes (the CoGnome), but more generally of the Network Architecture. Since it is increasingly evident that smart aggregates of humans are frequently more effective decision makers than individuals (Rheingold, 2003), this CogWeb architecture collectively technologically enables cointelligence.

FUTURE TRENDS

In the context of the AutoGnomic approach to Synthetic Mind, the following definitions might be projected to emerge in characterizing the Web:

- Web 1.0 – Syntactic Web (Perceptual; Intelligence) (Lynch, 1996)
- Web 2.0 – Syntactic-based Social Web (Tapscott, 2007)
- Web 3.0 – Semantic Web (Cognitive; Knowledge)
- Web 4.0 – Semantic-based Social Web
- Web 5.0 – Pragmatic Web (Valuational; Wisdom**)**
- Web 6.0 – Pragmatic-based Social Web
- Web 7.0– AutoGnomic (Semiotic) Web; a Semiotic Web holistically incorporates the Syntactic, Semantic and Pragmatic functionalities as well as including its Social Web and anticipates a new computing paradigm realizing Boundary Logic (Bricken, 2007) in a Relational Computer-*PILE* (Krieg, 2007). Hence the uniqueness, novelty and robustness of the AutoGnome. *Q.E.D.*

The current status in this development (Synthetic Mind: Intelligence→Knowledge→Wisdom) is hovering between Web 1.0 and Web 2.0 with forays into Web 3.0. Hence, Synthesizing Knowledge is the present goal with Wisdom (Knowledge processed through a Value filter) still a human interpretive extension. Nevertheless, the mission of AutoGnomics, in contrast to approaches to bring AI into the Web, is to reform the Web as AI; hence, Web 7.0 is a present goal of an approach to both a Synthetic Mind and Virtuality, the AutoGnomic Intellisite.

CONCLUSION

The AutoGnome is at the beginning of its technology cycle now focused on Knowledge Development and has a technical roadmap which will continuously and significantly increase, in contrast to competitors, its differentiation into the future. In particular, its Virtual Nature as an autonomous WebGnome (and CoGnome and CogWeb) in the IntelliSite application will make it stand out as unlike user-dependent web tools. The benefits of applications of the current version

of the AutoGnomic Technology in contrast with other special purpose AI approaches derive from the general purpose Semiotic Nature of its core Automated (Autonomous) Inference/Inquiry Engine whereby this same core engine can be deployed in a broad spectrum of contexts (education, health, business, economic and community development, homeland security, etc.) with only the provision of the "connectors" of the engine to that context, but no re-development of the engine itself. It should be emphasized that the current status in the development of the AutoGnomic Technology is essentially state of the art amongst competing commercial technologies, albeit the claims include significant advancements yet to be commercially introduced.

REFERENCES

Bricken, W. (2007). *Boundary mathematics. Boundary logic*. Retrieved June 17, 2008, from http://wbricken.com/index.html

Cox, R. T. (1961). *The algebra of probable inference*. Baltimore, MD: Johns Hopkins Press.

Etter, T. (2006). *Boundary institute for the study of foundations. Three place identity.* Retrieved June 17, 2008, from http://www.boundaryinstitute.org/theoretical.htm

Gudwin, R. & Queiroz, J. (2007). Preface. In R. Gudwin, & J. Queiroz (Eds.), *Semiotics and intelligent systems development*. Hershey, PA: Idea Group Inc.

Hamann, J. R. & Bianchi, L. M. (1970). The evolutionary origin of life: Preliminary considerations of necessary and (possibly) sufficient conditions. *Journal of Theoretical Biology, 28,* 489.

Hamann, J. R. (2007a). Computational autognomics: An introduction. In R. Gudwin & J. Queiroz (Eds.), *Semiotics and intelligent systems development* (pp. 287-309). Hershey, PA: Idea Group Inc.

Hamann, J. R. (2007b). *AHA! Institute*. Retrieved June 17, 2008, from http://relatedone.blogspot.com/index.html

IBM (2008). *Global innovation outlook*. Retrieved June 17, 2008, from http://domino.watson.ibm.com/comm/www_innovate.nsf/pages/world.gio.html

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge: Cambridge University Press.

Knuth, K. H. (2003). Intelligent machines in the 21st century: Foundations of inference and inquiry. *Philosophical Transactions Royal Society London, (A)361*, 2859–2873.

Krieg, P. (2007). What makes a thinking machine? Computational semiotics and semiotic computation. In R. Gudwin & J. Queiroz (Eds.) *Semiotics and intelligent systems development* (pp. 311-329). Hershey, PA: Idea Group Inc.

Lynch, M. (1996). *Autonomy: Understanding what matters – Meaning based computing*. Retrieved June 17, 2008, from http://www.wikinomics.com/

Pendergraft, E. P. (1993). *The future's voice: Intelligence based on pragmatic logic.* Jasper, AR: Privately Distributed.

Principia Cybernetica Web (2008). Retrieved June 17, 2008, from http://pcp.vub.ac.be/TOC.html

Rheingold, H. (2003). *Smart mobs: The next social revolution.* New York: Basic Books.

Roco, M. C. & Bainbridge, W. S. (2002). *Converging technologies for improving human performance: Nanotechnology, biotechnology, information technology and cognitive science (NBIC*). Retrieved June 17, 2008, from http://www.wtec.org/ConvergingTechnologies/1/NBIC_report.pdf

Shoup, R. (2008). *Boundary institute for the study of foundations*. Retrieved June 17, 2008, from http://www.boundary.org/

Singularity Institute for Artificial Intelligence, Inc. (2001). *General intelligence and seed AI 2.3: Creating complete minds capable of open-ended self-improvement*. Retrieved June 17, 2008, from http://www.singinst.org/ourresearch/publications/GISAI/index.html

Spencer-Brown, G. (1969). *Laws of form* (4th ed.). New York: E. P. Dutton.

Tapscott, D. (2007). Wikinomics: How mass collaboration changes everything. Retrieved June 17, 2008, from http://www.wikinomics.com/

KEY TERMS

Algebra of Probable Inference/Inquiry: The *Algebra of Probable Inference/Inquiry* is a common sense foundational reformation of the concepts of Probability, by the simple generalization of *implication* among logical statements in the Boolean algebra to *degrees of implication*, and of Entropy, by generalizing a particular function of the question lattice to a valuation called *relevance* which is a measure of the degree to which a statement answers a given question. This effectively establishes probability theory as logic.

AutoGnome: The AutoGnome is a self-knowing general purpose software system of automated (autonomous) inquiry, inference and intuition exploiting a mechanized carrier system for relational semiosis as a virtual (synthetic) mind.

Boundary Mathematics: Boundary Mathematics is a semiotic formalism generated by creating a distinction (a boundary) in nonexistence (of system) thus resulting in a first system. Extended to multiboundaries with a common sense reiterative reduction rule leading either to one distinction or nonexistence, this mathematical form and process is the germ of an approach to the formulation of a universal language of mathematics.

CoGnome: The CoGnome is a selected WebGnome which, inter-connecting two or more Intellisites in a Network of Intellisites, provides a computerized collective intelligence, an automated cointelligence, that is, the Collective-AutoGnome (Auto(Co)Gnome) or simply the CoGnome.

CogWeb: The CogWeb is the Network of Intellisites implemening the CoGnome for Network Decision-Making by autonomously formed Intellisite-defined groups, organizations, communities, and societies.

Intellisite: The Intellisite (an Intelligent Website) is a constructed software environment (a Website) with an embedded form of the AutoGnome known as a WebGnome, an intelligent agent residing in this cyberspace environment which, with its continuous adaptive learning from mimicking the user's behavior, will grow into a likeminded replica (*Mind*Clone) of a user-*self* acting in the Virtual Reality of the Internet with the synthetic mind capabilities of the AutoGnome.

Maximum Entropy (MaxEnt) Principle: MaxEnt is a technique for automatically acquiring probabilistic knowledge from incomplete information without making any unsubstantiated assumptions. Entropy is a mathematical measure of uncertainty or ignorance: greater entropy corresponds to greater ignorance. Hence, the MaxEnt solution is the least biased possible solution given whatever is experimentally known, but assuming nothing else.

Order/DisOrder/ReOrder Form: It is a tenet of Relational Systems that any semiotic act must, of necessity, express the *Form* of experience as the inseparable conjunction of:

- Ordered (i.e., determined or certain) experience: a formal algebra/logic of semiosis
- DisOrdered (indeterminate or uncertain) experience: a theory of probable inference/inquiry
- ReOrdering DisOrdered experience: via a generalized probabilistic optimization principal

That is to say, experience is all at once partially ordered, partially chaotic and partially organizable.

Semiotic Relational Systems: A Semiotic Relational System is a system of relations exhaustively admitting all forms of interrelatedness among systems and/or relations and with certain systems or relations taking the place of (i.e., imaging (signifying)) other systems or relations.

Automation of American Criminal Justice

J. William Holland
Georgia Bureau of Investigation, USA

INTRODUCTION

Criminal Justice has been one of the public sectors in the forefront of the move toward automation and digital government. The effect of computerization on American criminal justice has been profound and it has transformed the criminal justice process in many fundamental ways. Starting with President Lyndon Johnson's government commission, *The Challenge of Crime in a Free Society: A Report by the President's Commission on Law Enforcement and the Administration of Justice*, public and private experts in criminal justice and technology laid out the information needs of the criminal justice system and the computer systems to meet those demands. At a time when computerization was minimal throughout the criminal justice system, these task force members developed the blueprint for today's multilayered automated criminal justice environment (Dallek, 1998, pp. 405-407, 409-411; *Challenge of crime in a free society*, 1967, pp. 268-271).

Among the major recommendations of the commission were the creation of a national directory of offenders' criminal records, what came to be known as Computerized Criminal History (CCH) and the development of similar directories at the state level. The commission also called for federal coordination of standards for criminal justice information and sharing. Finally, the report urged that a study of fingerprint classification techniques be undertaken with a view to automating much of the fingerprint search and identification effort and that work be intensified to create a national linkage of files on wanted persons and stolen vehicles under the name of the National Crime Information Center (NCIC) (*Challenge of crime in a free society*, 1967, pp. 255, 268-271; *Task force report: Science and technology*, 1967, p. 69).

BACKGROUND

One of the earliest responses to this report was the creation of the Law Enforcement Assistance Administration (LEAA) within the United States Department of Justice (DOJ). In 1969, LEAA funded Project SEARCH to create a nationwide computerized criminal history system. From this initial effort, SEARCH quickly evolved into an independent consortium of states with the mission of demonstrating a computerized system for the electronic exchange of criminal history information. On the national level, the United States Attorney General assigned management responsibility for the interstate and national portion of this system to the Federal Bureau of Investigation. The states also formed the National Law Enforcement Telecommunications System (NLETS) electronically linking the states as well as the FBI and the Royal Canadian Mounted Police. By 1976, 26 states had used LEAA funding to create state level central repositories for computerized criminal history information (U.S. Department of Justice, 2001c, p. 26).

It became apparent during the last half of the 1970s, however, that greater decentralization of the nation's criminal history systems was urgently needed. To respond to these issues and concerns, the various states, FBI and SEARCH created the Interstate Identification Index or Triple I (III) concept in 1980 (U.S. Department of Justice, 2001c, pp. 26-27, 76-82, 88). Designed to replace a centralized national criminal history file, III was an index of criminal offenders that pointed to the state or states where detailed criminal history information could be found. There was widespread acceptance of III for criminal justice purposes: By 2001, 43 states participated. Legal restrictions and concerns, however, limited use of III for non-criminal justice use and weakened any effort to achieve a truly decentralized criminal history system. Consequently, the FBI continued to maintain criminal histories on individuals to meet interstate non-criminal justice needs (U.S. Department of Justice, 2001c, pp. 76-82).

Another factor that prevented the decentralization of criminal history information was the vast effort required in the time-consuming fingerprint identification process. A new system called the NCIC classification was implemented in the 1970s. It did little, however, to speed up the overall identification process (*Challenge of crime in a free society*, 1967, p. 255; *Task force report*, 1967, p. 16; Ms. Shirley Andrews, personal communication, September 9, 2002).

During the mid 1980s, new technological solutions for fingerprint identification emerged on the market. These systems, called automated fingerprint identification systems (AFIS), significantly reduced the manual tasks needed to search a fingerprint and made true searching of latent crime scene fingerprints possible. By the close of the 1980s, many states and a few local agencies had purchased these systems. Most were stand alone systems dedicated to the fingerprint input, search, and presentation of potential candidates for human comparison. A few states, however, attempted to expand the capabilities of these systems and link them to

other criminal history processes. When combined with the proven effectiveness of the AFIS latent search capability, the new technology contained the potential to transform criminal justice systems (U.S. Department of Justice, 2001b, pp. 43-44; U.S. Department of Justice, 2001c, pp. 61-63).

In the early 1990s, efforts were made through the National Institute of Standards and Technology (NIST) to devise a national fingerprint transmission standard; an effort spearheaded by the FBI. By 1993, a national standard for the electronic interchange of fingerprint information was approved by NIST and became the basis for the electronic linkage of local jurisdictions to state criminal history bureaus and the FBI. It formed the basis for the emerging national network of real-time identification and criminal history systems (See *Data format for the interchange of fingerprint, facial, and SMT information*, originally issued in 1993, amended in 1997 and further amended in 2000; U.S. Department of Justice, 2001c, pp. 61-63.)

CURRENT AND FUTURE TRENDS IN CRIMINAL JUSTICE AUTOMATION

Building on these past activities in fingerprint and criminal history automation, emphasis within state and national criminal justice circles has shifted to the need to share information, what is known as integrated criminal justice. With the explosion of the Internet and simultaneous cost limitations on criminal justice system development, both federal and state funding entities require that new criminal justice system developments build in the concept of information sharing, realignment of processing functions, and greater involvement of all criminal justice parties in individual systems development. The goal of this new focus is to eliminate duplicate entry of the same information and increase the overall completeness and accuracy of criminal justice information. (U.S. Department of Justice, 2001c, pp. 63-65; Harris, 2000, pp. 7, 14, 18-20, 41; U.S. Department of Justice, 2001b, pp. 47-48, 50; *Planning the integration of justice information systems*, 2002, pp. 2-3.)

Integrated justice efforts, however, have also resurrected older worries about privacy of such information and merged them with new concerns about greater linkage of criminal justice and non-criminal justice information on individuals. Questions about release of integrated information are linked to serious questions about the accuracy of the information released. These fears are intensified as private companies demand access to criminal history information, gathered at public expense, to market to customers for profit. In many jurisdictions, the old line between public and private responsibilities and authority has faded as private companies have assumed many of the traditional criminal justice information systems functions. In addition, the heightened threat of terrorist attacks has led to efforts to gather large amounts of information on individuals into databases to search for terrorist patterns. These efforts have collided with fears about loss of privacy and misuse of such information by the government. Initiatives such as the Total Information Awareness effort and the MATRIX project to correlate private and public data on suspicious individuals have ground to a halt in the face of protest from citizens fearful of the loss of civil liberties. (Ideas that mattered in 2003:9. No future for terror market, 2003; MATRIX Updates, 2003; *Planning the integration of justice information systems*, 2002, p.5; Stanford, 2003; U.S. Department of Justice, 2001a, pp. 8, 12; U.S. Department of Justice, 2001b, pp. 2-3, 27-28, 50).

CONCLUSION

In 1967, a national commission developed *The Challenge of Crime in a Free Society*, the roadmap for today's highly automated but incomplete criminal justice system. This report served the nation well but it is time to move beyond its confining vistas, time to recognize that dramatic developments in computer technology and digital government demand new answers to old questions and the formulation of entirely new questions. The events of September 11, 2001 have raised anew questions about lack of information on potential threats to society and posed new questions on how we as a nation can weave together governmental and private computerized information to detect dangerous individuals intent on mass murder without compromising constitutional safeguards and individual liberties. It is time to convene a new national task force charged with the duty to assess the challenge of crime and terror in a free digital society. Only then can criminal justice automation and digital government move forward in a planned and comprehensive way.

REFERENCE

(*References marked with an asterisk indicate reports included in the Commission report.)

Challenge of crime in a free society: A report by the President's Commission on Law Enforcement and Administration of Justice. (1967). Washington, DC: US Government Printing Office.

Dallek, R. (1998). *Flawed giant: Lyndon Johnson and his times, 1961-1973*. NewYork: Oxford University Press.

Data format for the interchange of fingerprint, facial, and SMT information. (2000). Washington, DC: US Government Printing Office.

Harris, K.J. (2000, September). *Integrated justice information systems: Governance structures, roles, and responsibilities*. Retrieved July 10, 2002 from SEARCH Group, Inc. Web site at *http://www.search.org/images/pdf/governance.pdf/*

Ideas that mattered most in 2003: 9. No future for terror market (2003, December 28). *Atlanta Journal Constitution, (December 28, 2003),* p. G3.

MATRIX Updates. Retrieved February 23, 2004 from ACLU Website *http://www.aclu.org/Privacy/Privacy.cfm?ID=14240&c=130*

Planning the integration of justice information systems: Developing the justice information exchange model. Retrieved July 10, 2002 from SEARCH Group, Inc. Web site: *http://search.org/integration/pdf/JIEM.pdf*

Stanford, D. D. (2003). ACLU attacks MATRIX on privacy. *Atlanta Journal Constitution*, (October 31), p. G3.

**Task force report: Science and technology*. (1967). Washington, DC: US Government Printing Office.

Toward improved criminal justice information sharing: An information integration planning model (2000, April). Available from the International Association of Chiefs of Police, 515 North Washington Street, Alexandria, VA. 22314-2357.

US Department of Justice. (2001a). *Public attitudes toward uses of criminal history information: A privacy, technology, and criminal justice information report* (NCJ187663). Washington, DC.

US Department of Justice. (2001b). *Report of the National Task Force on privacy, technology, and criminal justice information* (NCJ187669). Washington, DC.

US Department of Justice. (2001c). *Use and management of criminal history record information: A comprehensive report, 2001 update* (NCJ187670). Washington, DC.

KEY TERMS

Automated Fingerprint Identification System (AFIS): A system that provides computerized fingerprint identification of arrestees, applicants for licensing and employment, and crime scene fingerprints of potential suspects.

Computerized Criminal History (CCH): A system containing offenders and their individual arrests, final disposition of those arrests, and custodial information for those arrests.

Federal Bureau of Investigation (FBI): The federal government's investigative, forensic, and criminal justice information system entity. It is part of the U.S. Department of Justice.

Interstate Identification Index (III): a national index of offenders housed at the FBI that points an inquiring entity to those states that contain detailed criminal histories on the requested individual.

Multistate Anti-Terrorism Information Exchange (MATRIX): A consortium of states attempting to create a database of public and private information on individuals to allow for advanced searches to uncover suspicious criminal and/or terrorist activity.

National Crime Information Center (NCIC): A national center housed at the FBI that provides 24 hour a day, seven day a week, real time access to law enforcement for warrants, stolen vehicles and other articles.

National Institute of Standards and Technology (NIST): A federal government standards setting body housed in the U.S. Department of Commerce.

National Law Enforcement Telecommunications System (NLETS): A state managed national network that provides state to state communications as well as links to the FBI and other large scale criminal justice entities.

SEARCH Group, Inc: A state managed consortium that represents the states as a body on criminal justice information systems issues at the national level.

Total Information Awareness: A discontinued effort by the federal government to create a vast database containing public and private information on individuals to allow for advanced search techniques to uncover suspicious activity or indications of possible terrorist links.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 197-199, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Autopoietic Approach for Information System and Knowledge Management System Development

El-Sayed Abou-Zeid
Concordia University, Canada

INTRODUCTION

In the last decade a new generation of information systems (ISs), such as Web-based information systems and knowledge management support systems, have emerged in response to ever-changing organizational needs. Therefore, the need for new "Information System Design Theories" for the emerging ISs is recognized. According to Walls, Widmeyer, and El-Sawy (1992), an "IS design theory" must have two aspects—one dealing with the description of the system and one dealing with the prescription, that is, the process of developing of the system. The prescription aspect includes a description of procedures and guidelines for system development. In addition, these two aspects must be grounded on theories from natural or social sciences (i.e., kernel theories).

As information systems are socio-technical phenomena in which social and technical factors interweave the ways in which people work, the issue of "how to integrate the work activity and social context of users into the IS which is being designed" becomes one of the principal problems of IS development (Bai & Lindberg, 1999). Therefore, the development of new IS design theories requires a closer look at the system theories that go beyond the traditional system theory that is based, among other things, on Cartesian dualism (i.e., mind/body or cognition/action) and on a model of cognition as the processing of representational information (Mingers, 2001). One of the candidate theories is the theory of autopoiesis, which can be best viewed as a system-grounded way of thinking with biological foundations, together with its extension into social domain.

BACKGROUND

In order to conceive of living systems in terms of the processes that realized them, rather than in terms of their relationships with an environment, Maturana and Varela (1980) coined the word *autopoiesis* (αυτοσ = self, ποιενιν = creation, production) to denote the central feature of their organization, which is "autonomy." The meaning of this word coveys the very nature of living systems as systems that maintain their *identity* through their own operations of continuous self-renewal. Moreover, these systems could only be characterized with *reference to themselve*s and whatever takes place in them, takes place as necessarily and constitutively determined in relation to themselves—that is, *self-referentiality.*

One of the key concepts of autopoiesis is the distinction between organization and structure. On one hand, *organization* is the capability of a system to reproduce its identity by referring constantly to itself, through the alternate reproduction of its components together with the component-producing processes, that is, the capability of a recursive self-reproduction. On the other hand, *structure* is the realization of a system's organization through the presence and interplay of its components in a specific realization space. While organization is necessary to establish system unity and identity, structure is necessary because different spaces of its actualization impose different constraints on a system's components (Maturana & Varela, 1980). By rough analogy, an algorithm for solving certain problem can be viewed as a description of the system's organization, whereas the corresponding computer program can be viewed as the realization of this organization (structure) in a certain space (programming language).

Autopoietic Systems

An autopoietic system is defined by Maturana and Varela (1980) as "a network of processes of production, transformation and destruction of components. These components constitute the system as a distinct unity in the space of its actualization and they continuously regenerate and realize, through their interactions and transformations, the network of processes that produce them" (p. 135).

Among the distinct characteristics of the autopoietic systems, the most relevant ones are:

- *The simultaneous openness and closure.* Autopoietic systems are *open* with respect to structural interaction with the environment, that is, *structural openness*, which is an unavoidable consequence of the fact that system elements must satisfy the particular requirements of the physical domain in which they occur, while they are *closed* with respect to their own organization, that is, *organizational closure.* The recognition of *the*

simultaneous openness and closure of autopoietic systems is in opposition to the tradition for which a system is one or the other but not both. This interpretation is possible only because of the clear distinction between organization and structure (Bednarz, 1988).

- *Structural determination.* The state transition a system undergoes in response to environmental perturbations is entirely determined by its structure at that time. Moreover, a system specifies which environmental perturbations may trigger which structural changes. In other words, the environmental perturbations could trigger the system's structural changes but can never determine or direct these changes. Moreover, a system specifies which environmental perturbations may trigger which structural changes. Over time, through ongoing interactions with the environment, an autopoietic system will experience what Maturana and Varela (1992) describe as a *structural drift,* or a gradual change to their structure. The nature of this change is determined by a previous system's history of structural changes, that is, its *ontogeny.*

Higher-Order Autopoietic Systems

Two (or more) lower-order autopoietic systems can be "structurally coupled" to form a higher-order autopoietic system. Structural coupling is the ongoing process of the congruent structural changes between two (or more) systems that results from recurrent interactions between (among) them. Therefore, structural coupling has connotations of coordination and co-evolution. Moreover, following structural determination principle, two structurally coupled systems means that each of them selects from its possible structural changes those that are compatible with those in the other system and, at the same time, are suitable for the maintenance of its identity.

Social systems, such as enterprises, are constituted through the process of third-order structural coupling, or social coupling, the one that occurs between (or among) two (or more) second-order autopoietic systems. However, the unique feature of any human social system, such as an enterprise, is that the social coupling among its constituents occurs through "language in the network of conservations which language generates and which, through their closure, constitute the unity of a particular human society" (Maturana & Varela, 1992, p. 196). From this perspective, language is viewed as an example of social structural coupling that generates the self and creates meaning through interactions with others. Moreover, language represents what Maturana and Varela would describe as a consensual domain, which is the domain of arbitrary and contextual interlocking behaviors (Mingers, 1995a, p. 78). Within a consensual domain, two autopoietic systems would be able to observe the attribution of meaning to common events and undertake coordinated actions.

Autopoiesis and Cognition

Cognition is the term conventionally used to denote the process by which a system discriminates among differences in its environment and potential states of that environment. The evidence for this cognition is effectiveness of system behavior in response to the environmental perturbations. Today's dominant perspective on cognition, and consequently IS, is the idea that effective action is explainable in terms of manipulating formal and static representations of the objective reality (Mingers, 2001).

According to the theory of autopoiesis, perception is neither objectivist nor purely constructivist (Varela, 1992, p. 254); rather, it is co-determined by the linking of the structure of the perceiver and the local situations in which it must act to maintain its identity. This is the basis of *enactive (embodied) cognition,* which implies that the autopoietic system's activities condition *what can be perceived* in an environment, and these perceptions, in turn, condition future actions. In this view, "a cognitive system is a system whose organization defines a domain of interactions in which it can act with relevance to the maintenance of itself, and the process of cognition is the actual (inductive) acting or behaving in this domain" (Maturana & Varela, 1980, p. 13). Therefore, cognition, according to autopoietic theory, is essentially embodied. Or, in the words of Maturana and Varela (1992): "All doing is knowing, and all knowing is doing" (p. 26). In addition, cognitive domain of an autopoietic system is defined as the domain of all the interactions in which it can enter without loss of identity (Maturana & Varela, 1980, p. 119).

APPLICATIONS OF THE CONCEPTS OF AUTOPOIESIS IN IS DEVELOPMENT RESEARCH

The use theory of autopoiesis in IS research can be classified into two main categories: metaphoric and theory-oriented approaches (Beeson, 2001).

Metaphoric Approaches

Kay and Cecez-Kecmanovic (2002) used the concepts of *social coupling* and *consensual domain* to explain processes underlying the IS-organization relationship and how it impacts on the competitive advantage of an organization. They showed how processes of recurrent interactions between members of different groups—analysts, the IS team, and external clients—within the organization's work environ-

ment gave rise to commonalities in understanding, which in turn enabled continual IS organization co-emergence. In the same vein, Maula (2000) used the concepts of structural openness and organizational closure to identify two major knowledge management (KM) functions in four consulting firms. The first KM function, sensory function, is the realization of the structural openness of the firm and its structural coupling with its environment. The second KM function, memory function, is the realization of the concept's organizational closure and self-referentiality that enable the firm's effective functioning and continuous renewal. Hall (2003) also considered organizational memory as organizational heredity that served to maintain organizational integrity in a dynamic economic environment. Finally, Carlsen and Gjersvik (1997) used an autopoiesis metaphor to analyze possible organizational uses of workflow technology. They argued against "generic" business processes except as starting points for organizational adaptation. In addition, they indicated that the concept of autopoiesis implies that process models should include references to richer descriptions of the organizational environment and the environment the work process is situated in.

Theory-Oriented Approaches

Bai and Lindberg (1999) used first- and second-order cybernetics together with Luhmann's (1995) social autopoiesis theory and Engeström's (1987) activity theory to develop a framework for studying the relationship between IS design activity, use activity, and the embedded social context. This framework sheds light on the complex social context within which IS development takes place, and provides an epistemological understanding of the relationship among the elements involved in IS development. Moreover, it can be used to develop methodologies for the practice of IS development, and guide various research activities, such as the socio-technical approach.

FUTURE TRENDS

The autopoietic metaphor provides ways of thinking about the mechanisms underpinning the development and the introduction of IS in an enterprise. Here, a third-order autopoietic system is used as a metaphor to explore referential correspondence between the characteristics of autopoietic systems and an enterprise and its IS. For example, organizational closure of an autopoietic system implies that it is homeostatic and its own organization is the fundamental variable to be maintained constantly. This concept may be used to explain why the behavior of IS developers and users seems to be stuck sometimes in persistent patterns or repertoires (Beeson, 2001). Moreover, the difficulties of system integration may be better understood from a perspective of structural coupling than from one of rational design or negotiation (Beeson, 2001). The structural coupling concept can also be used to explore the way common understandings between enterprise members emerge as a function of interactions in the work environment (Kay & Cecez-Kecmanovic, 2002). Therefore, significant focus must be drawn towards the patterns of interaction between members of an organization and the way in which these patterns may be supported to give rise to consensual domains (Kay & Goldspink, 2005).

From an autopoietic view, introducing a new IS in an enterprise can be conceptualized as a kind of perturbation that provokes or triggers an enterprise's structural-determined responses. Therefore, IS development process can be viewed as the means for realizing structural coupling between an enterprise and its new IS, and becomes an integrated aspect of the recurrent interactions between developers and users in the work environment. Table 1 summarizes the implications of theory of autopoiesis for IS development process.

Furthermore, ISD process by its very nature is a knowledge work (Iivari, 2000), and its success depends on the developers' tacit knowledge (Abou-Zeid & Bahli, 2005). From an autopoietic perspective, knowledge is an embodied (enactive) notion, and it cannot be treated as an object that can be captured, packaged, and processed. Therefore, tacit knowledge associated with ISD should be appreciated as much as explicit knowledge. The interplay of the explicit and tacit and embedded organizational knowledge also forms a stimulating research topic.

Finally, there are many open questions. While some of the main concepts of autopoiesis theory are presented, considerable work remains to be done in terms of translating these concepts into practical solutions for the workplace.

CONCLUSION

The theory of autopoiesis, as a kernel theory for IS design theory, can used to derive a set of ISD methodology meta-requirements which includes insider frame of reference**,** historicity**,** context-dependency of ISD methodology, and minimal set of initial requirements.

REFERENCES

Abou-Zeid, E.-S., & Bahli, B. (2005). Knowledge management based view of information system development process. *Proceedings of the Information Resources Management Association (IRMA) International Conference*, San Diego, CA.

Bai, G., & Lindberg, L.-A. (1999). A sociocybernetic approach to information systems development. *Kybernetes, 28*(6/7), 792-809.

Table 1. Autopoietic implications for IS development

Characteristics of Autopoietic Systems	**Implications for IS Development**
Organizational Closure and Self-Referentiality	**Insider frame of reference.** The organizational closure and self-referentiality of an enterprise suggest it is best understood from *inside*. Therefore, an interpretive or hermeneutic approach could more reliably and intelligibly account for the experiences, intentions, and interpretations of its members. Moreover, the main role of the system developer is the role of "*catalyst* and/or *emancipator*" (Hirschheim & Klein, 1989) who helps enterprise members develop the necessary inquiring, collaborative, and communicative patterns needed to continuously explicate their information requirements.
	Historicity. As an enterprise is continuously reproducing itself, it must do so with constant reference to itself, its past practices, values, decisions, contracts, and commitments (Truex, Baskerville, & Klein, 1999). Therefore, explicating an enterprise's history is an essential element in developing new knowledge and in introducing a new IS (von Krogh, Ross, & Slocum, 1994).
Structural Determination and Structural Coupling	**Context-dependency of IS development methodology.** Viewing ISD methodology as the means for realizing structural coupling between an enterprise and its new IS implies that it cannot be separated from the enterprise's context. In other words, an autopoietic metaphor of an enterprise and its IS suggest "strong" approaches to systems development instead of the commonly used "weak" approaches (see Key Terms).
Embodied Cognition	**Minimal set of initial requirements.** The autopoietic view of cognition implies that requirements are always in motion, unfrozen, and negotiable (Truex et al., 1999). Therefore, IS development can be viewed as an open-ended bootstrapping process that starts with a minimal set of requirements. Moreover, formal representation must be subordinated to the fostering of mutual understanding and coordinated action in the development team and between the team's members and the stakeholders (Beeson, 2001; Robert Kay & Cecez-Kecmanovic, 2002). As knowledge is not an object that may be captured, packaged, processed, and distributed, KMS may be best conceptualized as "an additional medium through which interlocking behaviors may converge and the congruities of context, that give rise to consensual domains, develop" (Robert Kay & Goldspink, 2005).

Bednarz, J. (1988). Autopoiesis: The organizational closure of social systems. *System Research, 5*(1), 57-64.

Beeson, I. (2001). Implications of the theory of autopoiesis for the discipline and practice of information systems. *Proceedings of the IFIP WG 8.2 Working Conference on Realigning Research and Practice in Information Systems Development: The Social and Organizational Perspective,* Boise, ID.

Carlsen, S., & Gjersvik, R. (1997). Organizational metaphors as lenses for analyzing workflow technology. *Proceedings*

of the International ACM SIGGROUP Conference on Supporting Group Work: The Integration Challenge, Phoenix, AZ.

Engeström, Y. (1987). *Learning by expanding. An activity-theoretical approach to developmental research.* Helsinki: Orienta-Konsultit.

Hall, W. (2003). Organisational autopoiesis and knowledge management. *Proceedings of the 12th International Conference on Information Systems Development—Methods & Tools, Theory & Practice*, Melbourne, Australia.

Hirschheim, R., & Klein, H. (1989). Four paradigms of information systems development. *Communications of the ACM, 32*(10), 1199-1216.

Iivari, J. (2000). Information systems development as knowledge work: The body of systems development process knowledge. In E. Kawaguchi, H. Kangassalo, H. Jaakkola, & I. Hamid (Eds.), *Information modelling and knowledge bases XI* (pp. 41-56). Amsterdam, The Netherlands: IOS Press.

Kay, R., & Cecez-Kecmanovic, D. (2002). Toward an autopoietic perspective on information systems organization. *Proceedings of the 23rd Annual International Conference on Information Systems,* Barcelona, Spain.

Kay, R., & Goldspink, C. (2005). Organizational knowledge & autopoiesis: Implications for knowledge management. *Proceedings of the Systems Thinking and Complexity Science: Insights for Action, 11th Annual ANZSYS Conference/Managing the Complex V,* Christchurch, New Zealand.

Luhmann, N. (1995). *Social systems.* Stanford, CA: Stanford University Press.

Maturana, H., & Varela, F. (1980). *Autopoiesis and cognition.* Dordrecht: Reidel.

Maturana, H., & Varela, F. (1992). *The tree of knowledge: The biological roots of human understanding* (revised ed.). Boston: Shambhala.

Maula, M. (2000). The senses and memory of a firm—implications of autopoiesis theory for knowledge management. *Journal of Knowledge Management, 4*(2), 157-161.

Mingers, J. (1995a). Information and meaning: Foundations for an intersubjective account. *Information Systems Journal, 5*, 285-306.

Mingers, J. (1995b). *Self-producing systems: Implications and applications of autopoiesis.* New York: Plenum.

Mingers, J. (2001). Embodying information systems: The contribution of phenomenology. *Information and Organization, 11*(2), 103-128.

Truex, D., Baskerville, R., & Klein, H. (1999). Growing systems in emergent organizations. *Communications of the ACM, 42*(8), 117-123.

Varela, F. (1992). Whence perceptual meaning? A cartography of current ideas. In F. Varela & J. Dupuy (Eds.), *Understanding origins: Contemporary views on the origin of life, mind and society* (pp. 235-263). Dordrecht: Kluwer Academic.

Vessey, I., & Glass, R. (1998). Strong vs. weak approaches systems development. *Communications of the ACM, 41*(4), 99-102.

von Krogh, G., Ross, J., & Slocum, K. (1994). An essay on corporate epistemology. *Strategic Management Journal, 15,* 53-71.

Walls, J., Widmeyer, G., & El-Sawy, O. (1992). Building an information system design theory for vigilant EIS. *Information Systems Research, 3*(1), 36-59.

KEY TERMS

Cognitive Domain: The domain of all the interactions in which one can enter without loss of identity (Maturana & Varela, 1980, p. 119).

Consensual Domain: "The domain of interlocked conducts that results from ontogenetic structural coupling between structurally plastic organisms" (Mingers, 1995b).

Ontogeny: The history of structural changes that a system experiences without losing its identity.

Organization: The configuration of relationships among a system's components that define a system as a unity with distinctive identity, and determine the dynamics of interaction and transformations that it may undergo as such a unity.

Structural Coupling: The ongoing mutual co-adaptation between a system and its environment.

Structural Determination: The principle that the actual course of change in an autopoietic system is controlled by its structure rather than direct influence of its environment.

Structure: The physical embodiment of system's organization in a certain physical domain.

Strong vs. Weak Methods: "Strong methods are those designed to address a specific type of problem, while weak methods are general approaches that may be applied to many types of problems" (Vessey & Glass, 1998).

Category: Business Information Systems

Bankruptcy Prediction through Artificial Intelligence

Y. Goletsis
University of Ioannina, Greece

C. Papaloukas
University of Ioannina, Greece

Th. Exarhos
University of Ioannina, Greece

C. D. Katsis
University of Ioannina, Greece

INTRODUCTION

Bankruptcy prediction or corporate failure is considered a classic issue in both, academic and business communities. Bankruptcy risk is one of the most important factors (if not the most important one) to be considered when credit requests are screened or even existing debtors are evaluated. On the other hand, all potential stakeholders (shareholders, suppliers, customers, employees, creditors, auditors, etc.) have potential interest to identify if a company is on a trajectory that is tending towards failure. Commercial banks, public accounting firms and other institutional entities (e.g., bond rating agencies) appear to be the primary beneficiaries of accurate bankruptcy prediction, since they can use research results to minimize exposure to potential client failures. In addition to avoiding potentially troubled obligors, the research can also benefit in other ways. It can help in accurately assessing the credit risk of bank loan portfolios. Credit risk has been the subject of much research activity, since the regulators are acknowledging the need and are urging the banks to assess the credit risk in their portfolios. Measuring the credit risk accurately also allows banks to engineer future lending transactions, so as to achieve targeted return/risk characteristics. The other benefit of the prediction of bankruptcies is for accounting firms. If an accounting firm audits a potentially troubled firm, and misses giving a warning signal then it faces costly lawsuits (Atiya, 2001).

A series of techniques have been applied in literature. Econometric / statistical methods have first appeared in literature: In late 1960's (multiple) discriminant analysis (DA) was the dominant method; during the 1980's logistic analysis. In the 1990's artificial intelligence starts appearing in financial literature with neural networks (Odom & Sharda 1990) serving as an alternative to statistical methods demonstrating promising results.

The goal of this chapter is therefore two-fold: First, it intends to give an overview of the artificial intelligence techniques successfully applied to the problem, ranging from the first neural network applications to recent applications of biologically inspired algorithms, such as genetic algorithms. Then, two kernel based methods, namely the Radial Basis Function Neural Networks and the Support Vector Machines are applied to the bankruptcy problem.

BACKGROUND

Early statistical studies in bankruptcy prediction (e.g., Beaver, 1966 adopted a univariate methodology identifying the accounting ratios having the highest classification accuracy in separating failing and non-failing firms. Beaver investigated the predictability of 14 financial ratios. Altman (1968) examined simultaneously a series of financial ratios, enriching the single ratio approaches. A multiple discriminant function was calculated, the so-called Z-score composed of five financial ratios:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + .6X_4 + .999X_5 , \quad (1)$$

where

X_1 = Working Capital / Total Assets. (Measures liquidity)
X_2 = Retained Earnings / Total Assets. (Measures profitability)
X_3 = Earnings Before Interest and Taxes / Total Assets. (Measures operating efficiency)
X_4 = Market Value of Equity / Book Value of Total Liabilities. (Adds market dimension)
X_5 = Sales/ Total Assets. (Standard measure for turnover)

Z-Score model was modified by Altman, Haldeman, and Narayanan (1977). Their ZETA model was composed from seven financial ratios. Since these early studies, a vast range of statistical methodologies have been applied for the purposes of corporate failure prediction including logistic regression (Martin, 1977), logit (Ohlson, 1980), Kolari, Glennon, Shin, and Caputo (2002), probit and maximum likehood models (Zmijewski, 1984).

Literature review reveals that a series of financial ratios has been examined. Stability indicators, industry-specific indicators, macroeconomic factors, firm's particular features have been examined. However, there is no agreement on the features that carry significant predictive power. According to Courtis (1978) and Dimitras (1995) the applied ratios should adequately cover three fields: profitability, management efficiency, and solvency.

AI FOR BANKRUPTCY PREDICTION

The statistical methods described above have some restrictive assumptions such as linearity, normality and independence among predictor or input variables. Considering that violation of these assumptions for independent variables frequently occurs with financial data (Deakin, 1976) these methods have limitations to obtain effectiveness and validity. Artificial intelligence (AI) methods have been proven to be less vulnerable to these assumptions.

In the following paragraphs we present a brief description of AI techniques employed and a short review of research attempts applying AI techniques and approaches in the bankruptcy prediction problem. Specifically, three types of methods are analyzed: (1) artificial neural networks, (2) approaches based on decision trees, and (3) genetic algorithms.

Artificial Neural Networks

Artificial neural networks (ANNs) are based on the behavior of the biological neurons of the brain and are used to mimic the performance of a system. They consist of a set of elements that start out connected in a random pattern, and, based upon feedback, are modelled into the pattern required to generate the required results. When the problem is bankruptcy prediction, neural networks create a function of the predictors, mapping to a specified outcome; in our case bankruptcy or not.

As ANNs are capable of identifying and representing non-linear relationships in the dataset they were the first AI technique applied in the bankruptcy prediction problem and the most studied one. Odom and Sharda (1990) were the first to use NNs for bankruptcy prediction. They used the five financial ratios used in Altman's Z-score. A series of other applications can be identified where criteria / predictors' number varied from Altman's original 5 to 41 (Leshno & Spector, 1996). Various ANN based models, with different architectures and training algorithms have been employed. In some applications an extra dimension reduction stage was added, such as Principal Components Analysis (Ravi & Pramodh, in press) in order to reduce the dimensionality of the input feature vector of the ANNs.

The creation of additional indicators or features has been also tested. Atiya (2001) developed and employed novel indicators for his ANN. Lam (2004) followed a different approach incorporating macroeconomic variables as well as financial ones. Recently, ensemble neural networks classifiers that combine a number of single NN classifiers into one multiple classification system have gained ground (Tsai & Wu, 2008).

Decision Trees

Decision tree induction is a technique which is widely used for predictive/classification tasks. Decision trees employ mathematical formulations in order to detect the most important predictors and create a tree structure for deriving the classification decisions. An advantage of the decision trees is their ability to provide interpretation for their automated decisions. Still, their linearity limits their performance.

Marais, Patel, and Wolfson (1984) proposed recursive partitioning algorithm and bootstrapping techniques for inducing the decision tree. Frydman, Altman, and Kao (1985) employed recursive partitioning and compared decision trees to the DA, which was found inferior.

Genetic Algorithms

More recently Genetic Algorithms (GAs) have been applied in the bankruptcy prediction problem. GAs are stochastic search techniques that can search large and complicating spaces and mimic the ideas of natural evolution and the survival of the fittest.

As GAs are effective in searching large spaces, they are particularly effective for multi-parameter optimisation problems under several constraints. For this reason, in the bankruptcy prediction problem, they have been mainly applied for optimisation of the classification process either by selecting the most discriminant financial ratios (Sai, Zhong & Qu, 2007) or by optimising the classifier parameters (Back, Laitinen & Sere, 1996). Shin (2002) also uses GAs for the extraction of classification rules.

KERNEL BASED METHODS FOR BANKRUPTCY PREDICTION

In the second half of this chapter, following the directions proposed by Ravi Kumar and Ravi (2007) we examine the potential of two newer techniques, namely the SVMs and the Radial Based Function Neural Networks (RBF-NNs). Both methods are classified as kernel based methods.

Kernel based methods are based on mapping data from the original input feature space to kernel feature space of higher dimensionality and then solving a (linear) problem in this space. These methods allow us to interpret and to design learning algorithms geometrically in the kernel space (which is non linearly related to the input space). Kernel based methods are considered to perform satisfactorily in difficult environments such as noisy, of high dimension input or of limited number of training samples.

THE SVMs

The goal of the Support Vector Machines (SVM) (Cortes & Vapnick, 1995) is to produce a model which predicts target value of data instances in the testing set in which only the attributes are given. Let a training set of instance-label pairs be (x_i, y_i), $i = 1, ..., l$ where $x_i \in \Re^n$ is the training vector belonging to one of the two classes, l is the number of the extracted features in the training set and y_i indicates the class of x_i.

The support vector machine requires the solution of the following optimization problem:

$$\min_{w,b,\xi} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \right) \quad (2)$$

subject to $y_i(\mathbf{w}^T\phi(x_i)+b) \geq 1-\xi_i,\ \xi_i \geq 0,$

where b is the bias term, $\mathbf{w}$ is a vector perpendicular to the hyperplane $<\mathbf{w},b>$, ξ is the factor of classification error and $C>0$ is the penalty parameter of the error term. The training vectors x_i are mapped into a higher dimensional space F by the function $\phi : R^n \rightarrow F$. SVM finds a separating hyperplane with the maximal geometric margin and minimal empirical risk R_{emp} in the higher dimensional space. R_{emp} is defined as:

$$R_{emp}(a) = \frac{1}{2l}\sum_{i=1}^{l} | y_i - f(x_i, a) |, \quad (3)$$

where f is the decision function defined as:

$$f(x) = \sum_{i=1}^{l} y_i a_i K(x_i, x) + b, \quad (4)$$

where $K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j)$ is the kernel function, a_i are weighting factors and b is the bias term. In our case the kernel is a radial basis function (RBF) which is defined as:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2), \gamma > 0, \quad (5)$$

where

$$\gamma = \frac{1}{2\sigma^2}$$

(σ is the standard deviation) is a kernel parameter. The RBF kernel non-linearly maps samples into a higher dimensional space, so it can handle the case when the relation between class labels and attributes is nonlinear.

THE RBF NNS

Artificial Neural Networks (ANNs) that employ Radial Basis Functions (RBFs) usually require a larger architecture than standard feed-forward ANNs for the same classification task. However, their main advantage is that they are trained much faster and easier with a smaller training set (Chen, Cowan & Grant, 1991). In a RBF-ANN the nodes in the hidden layer utilize a transfer function with the following formula:

$$f(n) = e^{-n^2}. \quad (6)$$

The input to the transfer function of each neuron is the vector distance between a weight vector $\mathbf{w}$ (estimated during training) and the input vector $\mathbf{x}$ (financial ratios), multiplied by a bias b, that is:

$$n = \|\mathbf{w} - \mathbf{x}\| b. \quad (7)$$

According to this schema, a radial basis function has a maximum of 1 when the input (distance) is 0, thus as the distance between $\mathbf{w}$ and $\mathbf{x}$ decreases, the corresponding output increases. Consequently, a radial basis neuron acts as a detector that produces 1 whenever the input $\mathbf{x}$ is identical to its weight vector $\mathbf{w}$. The bias b allows the sensitivity of the radial basis neuron to be properly adjusted. During testing, when a feature vector is introduced to the network, each neuron in the hidden layer will produce an output value according to how close the input vector is to each neuron's weight vector. Thus, radial basis neurons with weight vectors quite different from the input vector ($\mathbf{x}$) have outputs near zero. On the other hand, a radial basis neuron with a weight vector close to the input vector produces a value near 1.

A subtype of RBF ANNs suitable for classification problems are the Probabilistic Neural Networks (PNNs) (Wasserman, 1993). The first hidden layer at such ANNs

operates similarly with that of RBFs described earlier. PNNs have an additional hidden layer which sums the outputs of the nodes from the previous layer in order to produce a vector of classification probabilities. A compete transfer function is applied in the final layer to produce the final output. This transfer function identifies the maximum probability, and produces a 1 for the corresponding class and a 0 for the rest classes.

Typical PNNs have two layers (besides the input and output ones). The first layer consists of RBF neurons, one for each pattern (training example). For each pattern we calculate the product of the weight vector and the given classification example and then the product is passed through the transfer function:

$$f(x) = e^{(\mathbf{x}^T \mathbf{w}_{ki} - 1)/\sigma^2}, \quad (8)$$

where $\mathbf{w}_{ki}$ is the weight vector of the i^{th} pattern from class k and σ^2 is the deviation of the Gaussian function for the specific node. In the second layer, we have as many nodes as is the number of classes and each node implements a summation process based on the outputs of the RBF nodes associated with a given class, using the following formula:

$$\sum_{i=1}^{N_k} e^{(\mathbf{x}^T \mathbf{w}_{ki} - 1)/\sigma^2} \quad (9)$$

where N_k is the number of patterns for class k. In order to produce the classification decision as output, PNNs employ binary neurons according to:

$$\sum_{i=1}^{N_k} e^{(\mathbf{x}^T \mathbf{w}_{ki} - 1)/\sigma^2} > \sum_{i=1}^{N_j} e^{(\mathbf{x}^T \mathbf{w}_{kj} - 1)/\sigma^2}. \quad (10)$$

Only one factor needs to be tuned during training, the smoothing factor (i.e. the deviation of the Gaussian functions). Small deviations can lead to sharp approximations and thus without proper generalization capabilities, while large deviations tend to miss probably valuable details. Therefore, the appropriate deviation (namely spread) must be chosen by experiment.

FINANCIAL APPLICATION

Indicators for Bankruptcy Prediction

The major assumption underlying almost all studies for bankruptcy prediction is that financial variables extracted from financial statements, such as financial ratios, contain a large amount of information about a company's credit / bankruptcy risk. Although a series of financial ratios have been applied in different studies the aim of the present study was not the identification of a good set of ratios but the examination of the performance of the classifiers over a specific set of financial ratios. Therefore, the ratios considered, in accordance to the study of to Olmeda and Fernadez (1997), were the following:

(I1) current assets/total assets
(I2) current assets-cash/total assets
(I3) current assets/loans
(I4) reserves/loans
(I5) net income/total assets
(I6) net income/total equity capital
(I7) net income/loans
(I8) cost of sales/sales
(I9) cash flow/loans.

These indicators are considered to fulfill the requirements set in par. 2.

DATASET

For comparison reasons, the dataset applied in our study is the one of Olmeda and Fernadez. This includes data from 66 Spanish banks active in the period 1977-1985. The ratios used for the failed banks come from the last financial statements issued before bankruptcy was declared while the data for non failed banks is from the 1982 statements. All 9 indicators were used. Summary statistics on dataset are given in the Appendix.

According to the originating study, this database was randomly split into two data sets (DS), DS1 consisted of 34 banks (15 failed and 19 non failed) and DS2 of 32 banks (14 failed and 18 non-failed). DS1 was used as a training set and DS2 for testing. Then the procedure is reversed, using DS2 for training.

Results

SVMs and probabilistic RBF ANNs were applied on the above dataset. The SVM parameters, γ and C, were defined heuristically after a series of experiments and more specifically γ was equal to $2^{-1.3}$ and C was equal to $2^{6.5}$. As far as the PNNs is concerned, the spread value was heuristically set to 206.

Table 1 summarises the obtained results and compares them to the performance of ANNs, Logit and Discriminant Analysis.

The accuracy obtained by the SVMs is quite high. Compared to the RBF ANN, the SVM approach provided

Table 1. Bankruptcy prediction dataset key statistics

Experiment	Training set: DS1 Test: DS2			Training: DS2 Test: DS1			
Method	**Accuracy (%)**	**Type I error[b] (%)**	**Type II error[c] (%)**	**Accuracy (%)**	**Type I error (%)**	**Type II error (%)**	**Average Accuracy (%)**
RBF ANN	**80**	**28.5**	**16.7**	**80**	**33**	**10.5**	**80**
SVM	**85**	**14.2**	**15.8**	**94**	**6.7**	**5.3**	**90**
ANN*	90			91			90
Logit*	87.5			85			86
D.A.*	81			77			79

** performance data coming from Olmeda and Fernadez (1997)*

higher prediction results. This difference in performance is mainly explained by the fact that SVMs operate in an automated fashion generating the optimum kernels for the problem under study. This is not the case with the RBF ANN approach where the kernels are actually represented by the training patterns, meaning that their performance is highly dependant from the training set selection.

FUTURE TRENDS

The application of AI can be adapted / extended in a series of credit risk estimation cases. Credit risk estimation could be extended even to microcredit applications with certain limitations. The lack of financial statements (especially in the cases of developing countries) often imposes the use of different criteria, while the fact the micro-enterprises heavily rely on personal participation precludes AI methods from detecting bankruptcy risk due to human factors (e.g., an illness or an accident). On the other hand, adapting the ratings to rate a microcredit company itself appears as a new niche market in the financial sector (Navajas & Suaznabar, 2006).

In techncical terms, future work will examine whether specific feature selection methods can improve the obtained accuracy even more and will verify the results in larger datasets. The combination of classifiers into hybrid classification approaches that amplify the advantages of individual models and minimise their limitations appears as a worth-while research direction.

CONCLUSION

A short review of major AI applications to solve the bankruptcy prediction problem was presented. Moreover, two kernel based methods have been tested for bankruptcy prediction. Nine financial ratios covering profitability, management efficiency and solvency were applied. One of the methods examined, namely the SVMs, provides high classification accuracy, comparable to the accuracy obtained by neural networks.

REFERENCES

Altman, E. (1968). Financial ratios, discriminant analysis and prediction of corporate bankruptcy. *Journal of Finance, 23*(3), 589-609.

Altman, E. I., Haldeman, R. G., & Narayanan, P. (1977). ZETA analysis: A new model to identify bankruptcy risk of corporations. *Journal of Banking and Finance 1*, 29–54.

Atiya, A. F. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks, 12*(4), 929–935.

Back, B., Laitinen, T., & Sere, K. (1996). Neural network and genetic algorithm for bankruptcy prediction. *Expert Systems with Applications, 11*(4), 407–413.

Beaver, W. (1966). Financial ratios as predictors of failures. *Journal of Accounting Research, 4*, 71-111.

Chen, S., Cowan, C. F. N., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks, 2*(2), 302-309

Cortes, C. & Vapnik, V. (1995). Support-vector network. *Machine Learning, 32*, 273-297.

Courtis, J. K. (1978). Modelling a financial ratios categoric framework. *Journal of Business Finance and Accounting, 5*(4), 371-386.

Deakin, B. E. (1976). A discriminant analysis of predictors of business failure. *Journal of Accounting Research*, 167-179.

Dimitras, A. (1995). *Multicriteria methods for the prediction of bankruptcy risk.* Unpublished doctoral thesis, Techical University of Crete.

Frydman, H. Altman, E. I., & Kao, D. (1985). Introducing recursive partitioning for financial classification: The case of financial distress. *Journal of Finance, 40*(1), 269–291.

Kolari, J., Glennon, D., Shin, H., & Caputo, M. (2002). Predicting large US commercial bank failures. *Journal of Economics and Business, 54*(32 1), 361–387.

Lam, M. (2004). Neural networks techniques for financial performance prediction: Integrating fundamental and technical analysis. *Decision Support Systems, 37*, 567–581.

Leshno, M. & Spector, Y. (1996). Neural network prediction analysis: The bankruptcy case. *Neurocomputing, 10*, 125– 147.

Marais, M. L., Patel, J., & Wolfson, M. (1984). The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classifications. *Journal of Accounting Research, 22*, 87–113.

Martin, D. (1977) Early warning of bank failure: A logit regression approach. *Journal of Banking and Finance, 1*, 249–276.

Navajas, S. & Suaznabar, C. (2006). Microfinance rating agencies: An industry on the rise. *Microenterprise Americas Magazine,* Fall.

Odom, M. & Sharda, R. (1990). A neural networks model for bankruptcy prediction. In *Proceeding of the IEEE International Conference on Neural Network, 2*, 163-168.

Ohlson, J. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research, 18*(1), 109-131.

Olmeda, I. & Fernadez, E. (1997). Hybrid classifiers for multicriteria decision making: The case of bankruptcy prediction. *Computational Economics, 10*, 317-335.

Ravi Kumar, P. & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques - A review. *European Journal of Operational Research, 18* (1), 1-28.

Ravi, V. & Pramodh, C. (in press). Threshold accepting trained principal component neural network and feature subset selection: Application to bankruptcy prediction in banks. *Applied Soft Computing Journal*.

Sai, Y., Zhong, C., & Qu, L. (2007). A hybrid GA-BP model for bankruptcy prediction. In *Proceedings of the Eighth International Symposium on Autonomous Decentralized Systems, ISADS '07* (pp. 473 – 477).

Tsai, C.-F. , & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction for credit scoring. *Expert Systems with Applications, 34*, 2639-2649.

Wasserman, P. D. (1993). *Advanced methods in neural computing*. New York: Van Nostrand Reinhold.

Zmijewski, M. E. (1984). Methodological issues related to the estimated distress prediction models. *Journal of Accounting Research, 22*(1), 59-82.

KEY TERMS

Artificial Neural Networks (ANN): An artificial neural network is a massive parallel distributed processor made up of simple processing units. It has the ability to learn from experiential knowledge expressed through interunit connections strengths, and can make such knowledge available for use.

Bankruptcy Prediction: The process based on financial data of a company to predict whether the company will become bankrupt.

Financial Ratios: Mathematical relationship between one financial quantity and another used to describe financial condition of a firm. There are many categories of ratios such as those that evaluate a business entity's liquidity, solvency, return on investment, operating performance, asset utilization, and market measures.

Genetic Algorithms (GA): Genetic algorithms are derivative free, stochastic optimization methods based on the concepts of natural selection and evolutionary processes.

Support Vector Machines (SVM): Support vector machines is a methodology used for classification and re-

gression. SVMs select a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible.

ENDNOTES

a * corresponding author (goletsis@cc.uoi.gr)

b Type I error = $\frac{\text{number of failed companied not identified}}{\text{total number of failed companies}}$

c Type II error = $\frac{\text{number of healty companied not identified}}{\text{total number of healty companies}}$

APPENDIX

	I1	*I2*	*I3*	*I4*	*I5*	*I6*	*I7*	*I8*	*I9*
Mean	0.3927	0.2664	0.4086	0.0241	0.0031	0.0764	0.0034	0.8848	0.0118
Standard Deviation	0.1124	0.0893	0.1330	0.0258	0.0141	0.3006	0.0149	0.1466	0.0237
Variance	0.0126	0.0080	0.0177	0.0007	0.0002	0.0904	0.0002	0.0215	0.0006
Kurtosis	0.8587	0.4218	1.2817	32.4912	24.2715	23.7063	23.7213	5.1434	43.1391
Skewness	0.7412	0.5887	0.5477	5.0437	-4.4427	-4.1201	-4.3491	0.6136	5.9543
Minimum	0.2177	0.0883	0.0443	0.0012	-0.0822	-1.7434	-0.0862	0.3478	-0.0231
Maximum	0.7671	0.4942	0.8070	0.1986	0.0226	0.8109	0.0278	1.3713	0.1830

Barriers to Successful Knowledge Management

Alexander Richter
Bundeswehr University Munich, Germany

Volker Derballa
Augsburg University, Germany

INTRODUCTION

Knowledge and Knowledge Management (KM) are gaining more and more attention in theory and practice. This development can be observed by an increasing number of publications since the 1990s, addressing the question of how knowledge in organizations can be organized and managed (Davenport & Prusak, 1998; Nonaka & Takeuchi, 1995). It is argued that knowledge is becoming the pre-eminent source of competitive advantage compared to the traditional factors of production, labour, capital and land. This theoretical discourse is accompanied in practice by an increasing number of KM initiatives. In many cases however, the results of those KM implementation projects have not lived up to the high expectations associated with them. Reasons for that are manifold. In this article, we will present the results of an extensive analysis of KM literature identifying the major barriers to KM. Those barriers represent current challenges during any holistic KM implementation that includes knowledge management systems (KMS).

BACKGROUND

KM, as an area of research is influenced by many disciplines such as information systems research, strategic management, organization science, psychology and human resources. Due to that, a wide array of different concepts and methods does exist. With the personification and the codification strategies two main perspectives on KM can be identified. The personification strategy, represented mainly by Japanese authors (e.g., Nonaka & Takeuchi, 1995) deals foremost with implicit knowledge embedded in human actors and aims at leveraging knowledge creation and sharing through informal and cultural mechanisms. The codification approach, which is dominated by U.S.-American authors (e.g., Bhatt, 2001), mainly considers the explicit aspects of knowledge and thus focuses on knowledge explication and reuse. The unbalanced focus on one of those perspectives has been made responsible for the failure of many KM initiatives. Because of this insight it has become more and more accepted that it is critical for successful KM to pursue a holistic approach. From that point of view KM is considered to be a management discipline that "(…) embodies organizational processes that seek synergistic combination of data and information processing capacity of information technologies, and the creative and innovative capacity of human beings" (Malhotra, 2005). Thus, whereas an overemphasis on technology is often problematic, a well balanced combination of technical and social approaches can be a rewarding departure (Alavi & Leidner, 1999).

BARRIERS TO SUCCESSFUL KM

Table 1 shows the barriers to successful KM, which are detailed in the following and categorized according to their origin along the dimensions technology, organization and human factors (Richter, 2006).

Technological Barriers

The balanced use of information technology is seen as a factor that can beneficially support different KM processes (Wiig, 1995). Typical examples for information technology aiming at the support of the codification aspects of KM include, for example, database solutions acting as knowledge databases or repositories. The personification aspects of KM can be supported by information systems that foster interpersonal communication, such as chat clients and groupware. Some of the most important barriers touching the technological domain of KM are highlighted in the following.

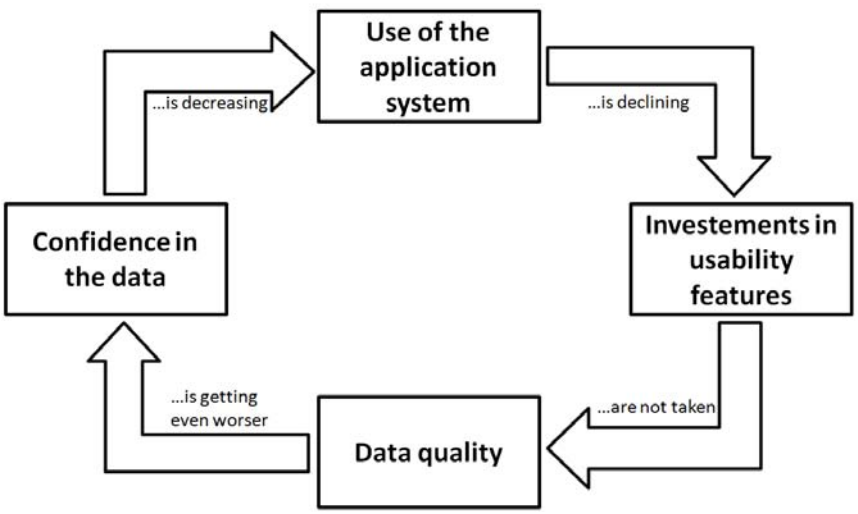
T1: Lacking Acceptance

Reasons for lacking acceptance can be found in a user-unfriendly application system, unsatisfying trainings and support granted by the software manufacturer, lacking stability and reliability and bad performance. Further, if the personal benefit for the individual user is not clear or the use of the system implies extra effort, many systems are insuf-

Table 1. Synopsis of all barriers

"T," "O" und "H" stand for the three dimensions "Technology," "Organization" and "Human"	
Category	**Barrier**
T 1	Lacking acceptance
T 2	Information overload and redundancies
T 3	Missing instruments for integrated planning and evaluation
O 1	Linguistic problems
O 2	Lack of time
O 3	Unfavourable company-and knowledge culture
O 4	Missing or diverging goals
H 1	Cultural influences
H 2	Personal fears and uncertainties
H 3	Inadequate motivation

Figure 1. Problem circle in the context of an electronic knowledgebase[1]



ficiently accepted by employees. The influence of change management during the introduction phase should also not be underestimated. Each of the factors mentioned above can lead to a so-called "problem circle," as it is illustrated here with the example of data quality (cf. Bullinger et al., 1998, p.30).

Usability

The "ISO 9241-11 Guidance on Usability" (ISO, 1998) issued by the International Organization for Standardization (ISO), defines the term "usability" as: "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." A particularly simple application, which is suitable to the user and its tasks, is consequently called "user-friendly." The ISO 9241 includes a catalogue of several criteria for the evaluation of the user-friendliness of a product. Among these are the self description ability of the application (dialogue is easy to follow and gives explanations when desired) its error tolerance (incorrect inputs are indicated and corrected by the user or automatically by the system), controllability (the user should be able to affect work speed, selection and order of media, kind and extent of expenditures) and the focus on simplicity: The completion of the task should be supported, though the user should not be overloaded with characteristics of the system ("as much information as necessary, as few as possible").

Trainings

Modern software solutions are getting more and more complex. Therefore, trainings are gaining importance. Training is even a traditional form of KM (on a basis of the personification strategy). This can take place at the customer site (so-called "In-house training courses") or in the rooms of the software manufacturer. The training courses should be arranged in such a way that the participants receive support while they handle the system. This enables them to master all upcoming tasks in the future without any further assistance.

Support

During the continuous improvement of a software product, bugs are not completely inevitable. For this case the user should have either a direct contact within the company or some kind of internal or external user-support, which is often divided in a three-part system, that is, 1st Level, 2nd Level or 3rd level support.

Stability, Reliability and Performance

If a system is unstable (i.e., it fails too often), unreliable (i.e., it does not supply the correct results) or not sufficiently performing, the user will be unwilling to work with it and consequently diminishes its use. Likewise he will lose the confidence into an unreliable system and he will never develop any confidence into a new one, respectively.

Data Quality

The term "information quality research" stems from a branch of research, which is concerned exclusively with the question "what means good information?" (Eppler, 2001, p.1). There are objective and subjective knowledge valuation criteria. Agreement prevails among the researchers over the fact that data quality contributes to user-friendliness and to the frequency of WMS-usage.

T2: Information Overload and Redundancies

The cost for producing and spreading information is nowadays negligible. This results in information overload and even leads to the doubling of the worldwide printed knowledge every 8 years (Krcmar, 2005, p.52). Eli Noam found appropriate words for this: "The real issue for future technology does not appear to be production of information, and certainly not transmission. Almost everybody can add information. The difficult question is how to reduce it" (Noam, 1986).

It is to be assumed that a considerable part of the worldwide digital information exists several times. From an aspect of data security this is certainly advisable. However, each version of a document should exist only once within an enterprise, because this is a potential source for mistakes and confusion and might cause difficulties for a system user to identify one document among others. Even if data redundancy could be reduced to a reasonable measure, the employee's need for a clear file structure or for search features is still a big issue as the employees are in contact with innumerable information objects every day.

T3: Missing Instruments for Integrated Planning and Evaluation

Common sense demands setting goals before any measures are taken. In the relatively intangible are of KM, that is often not the case. A framework of knowledge planning should be developed, which permits prioritization of the scheduled activities. On basis of this framework, knowledge can be managed and ex post be evaluated. Thus, the activities should be supported by a knowledge management system (KMS). If an objective is not accomplished or if it is not integrated in the system, it is not a surprise that deadlines are exceeded or time management is a burden for the employees. In this context it is remarkable that management level systems, developed for the integrated planning and evaluating of "tangible assets,"

the so-called management support systems, have already been existing for years. However, these do not or only to a small extent incorporate knowledge work.

Organizational Barriers

KM and the concepts behind it frequently challenge or even reject old organizational structures and past strategic adjustments of the organization. The additional expenditure, due to organizational barriers and efforts to master them, frequently bars enterprises from changing their structures.

O1: Linguistic Problems

Language is a very important medium, in which knowledge becomes manifest and is getting transportable. Although enterprises are increasingly internationally aligned, language is not a substantial hurdle anymore. Nevertheless, the exchange of experience or generally the exchange of implicit knowledge often fails due to linguistic hurdles (i.e., metaphors or analogies, agreed beforehand, are missing). Thus, a linguistic hurdle does not necessarily involve different native languages of the participants; the hurdle can be even the result of different technical terms (Pawar et al., 2001, p.7). Besides, it is often difficult to access stored knowledge because different participants use different terms and definitions for describing the same competences. Probst accuses the low extent of vocabulary concerning KM (so-called "controlled vocabulary") for the insufficient formulation of knowledge goals (Probst et al. 1997, p.90).

O2: Lack of Time

In several surveys, the scheduled time for dealing with KM has been identified as the overall barrier of KM (cf. e.g., Pawar, 2001, p.7). There are two main reasons for that. Firstly, the commitment and support of the chief executive might be missing. This could be a consequence of the fact that the employees who apply KM are more aware of the need for spending KM-time than the person in charge. Secondly, this could be a pretended argument for hiding a lacking motivation, for preserving power or for some other reason (Bullinger et al., 1998).

O3: Unfavorable Knowledge Culture

Although KM has gained a considerable degree of popularity in many enterprises, these maintain a mainly conservative tenor and therefore consider their enterprise culture as a success-critical factor that has to be observed. Knowledge within a company is, particularly in conflict- and risk-averse working environments, often retained, if not specifically legitimated by the company culture. Thus, new and innovative ideas or approaches are often not followed up (cf. Disterer, 2000, p. 541). Moreover, in the case of group decisions, the position of the majority or of the person with the highest status is adopted (cf. Nemeth & Nemeth, 2001, p. 96)) without rationally weighting or even exchanging different opinions among the group. Therefore both, the knowledge development and the knowledge acquisition are obstructed. Hierarchy and bureaucracy are another two barriers that hinder the knowledge flow. Strictly hierarchical forms of business organization can slow down communication and thus the knowledge exchange over the borders of divisions. Far worse is the case when it even completely blocks communication. Bureaucratic sets of rules and regulations which allow the spread of worthy knowledge within an enterprise and which state the way of how it is spread ("official routine"), contradict the thought of an interlaced KM. Old, line-oriented organizational structures with a multiplicity of hierarchy levels and rigid information flows can hinder a successful KM, too.

O4: Missing or Diverging Goals

KM should receive consistent, convincing and reliable support by the executive committee. The latter should state clear that it is convinced of the necessity of the knowledge exchange and that KM-initiatives are appreciated and sufficient support is granted. Moreover, executives have to act as role models for KM implementation.

Discrepancies between the principles of an enterprise (vision, mission, values, etc.) and the individual goals of the employees can lead to the fact that employees do not share individual knowledge und personal experiences voluntarily (Disterer, 2000, p. 542). While, for example, in marketing an enterprise aims at collecting and evaluating the knowledge of field representatives, it is observable that the representatives hold back their knowledge, in order to increase their value for the enterprise. Additionally, individual experiences of the employee can lead to a conflict situation between the views of the enterprise and the views and goals of the business principles.

Human Barriers

Human actors play a central role in the identification, acquisition, creation, storage, structuring, distribution and evaluation of knowledge. The knowledge of an employee is the most important factor for the process organization within an enterprise, though it is this implicit knowledge, which is most difficult to measure, to store and to distribute (Pawar, 2001, p.3). Thus, if the importance of the human factor is neglected in a KM-strategy and if numerous activities are not taken due to this fact, then several barriers oppose successful KM.

H1: Cultural Influences

One of the most important barriers in this area is the "not-invented-here-syndrome." In this case, a mixture of unawareness, uncertainty, distrust and vanity leads to the bias of developing their own solutions than using already existing ones. In addition, knowledge from lower hierarchical levels is in most cases not adequately approved. A common example is an employee who is not willing to accept or use established solutions of unknown, nonfamiliar or unpopular colleagues.

H2: Personal Fears and Uncertainties

The fear of being compromised can also lead to the fact that knowledge is not shared. If an employee communicates parts of his knowledge, this is based on the individual assumption that his knowledge is not trivial, but possesses a certain value and rarity (Disterer, 2000, p. 541). Thus, if a colleague has a different opinion concerning the value of a person's contribution, it is conceivable that this devaluation is passed back to the other person. Often colleagues are trying to demand allegedly necessary corrections or additions in order to document their own expertise in this way. If one is entering knowledge into a data base, the value of this knowledge is published at the same time and presented for discussion. This implies the possibility that colleagues do not share the same opinion about the value of this particular knowledge, or that colleagues use the opportunity to boost their own reputation by making corrections or additions.

During the initial determination of knowledge goals, the knowledge acquisition, knowledge development and knowledge distribution, power aspects pose a potential pitfall (Probst et al., 1997, p. 91). A competitive situation within a company does definitely not enhance the willingness to share knowledge. In connection with software reuse, Judicibus (1996) calls the barrier of power loss "egg head-syndrome."

H3: Inadequate Motivation

Finally, motivation is one of the most important and most comprehensive barriers to KM (e.g., Bullinger et al., 1998, p. 88). The quality and quantity of achievements, which for example, an employee provides in an enterprise, are basically affected by two criteria (e.g., Comelli & von Rosenstiel, 2001, p. 2): The individual abilities and talents of the person and their readiness to apply these. An enterprise should thus be interested not only in the abilities of the coworkers (e.g., by advanced training measures) but also in the readiness to perform or the motivation of the coworker.

The implementation of a KMS is based on the principle of reciprocity. On the one hand, an employee has to invest much effort and time in order to share his knowledge. On the other hand, he can draw considerable value from having access to the organization's knowledge collection. He is thus compensated for his knowledge sharing effort. Coworkers who do not trust this principle of mutual giving and taking or who have not experienced successful KM projects yet are lacking an important incentive for taking on the additional work.

H4: Generation of Knowledge as Prisoner's Dilemma

The prisoner's dilemma is a central component of the game theory. The dilemma is caused by decisions that appear rational to each individual, but at the same time lead to collectively worse results compared to mutual cooperation. For the area of KM this means that the prospects of joint KM effort, like the creation of a knowledge database, are too vague for many. Thus, many individuals are not ready to perform additional work if the assurance of cooperative behavior among their coworkers is missing. Therefore, it is necessary that coworkers do not only have in mind the individual use, but beyond that they have a certain measure of employment will and commitment for the community, so that knowledge exchange is set and maintained.

FUTURE TRENDS

Most current KMS do not overcome the main barriers and thus fail in providing sufficient support for KM from a holistic perspective. The barriers to successful KM can be used to develop requirements for the development of KMS:

- Functional requirements
- Nonfunctional requirements
- Integration of KM in business processes
- Flat hierarchies
- Definition of knowledge goals
- Common language
- Adequate incentives

Adequate incentive mechanisms play an important role in overcoming barriers connected with human factors in the area of KMS development. The integration of KM into business process management can create considerable benefits. Doing so, knowledge work is no longer treated as a separate function, but integrated or closely connected with operational processes. Thus, it is possible to reuse knowledge that has been generated during the implementation of prior projects for new projects.

CONCLUSION

KMS that focus on one of the main KM perspectives, the codification or the personification approach, cannot overcome the barriers to successful KM. It is necessary to follow a holistic approach that adequately considers social as well as technical aspects. The barriers presented here, have their origins in the technological, organizational and human domain, thus demonstrating the different fields of action within KM. Some of them can be overcome by minor adjustments; others require major organizational or cultural changes. By identifying and categorizing the individual barriers, the basis has been created for the development of KMS-requirements.

REFERENCES

Alavi, M., & Leidner, D.E. (1999, February). Knowledge management systems: Issues, challenges, and benefits. *Communications of the AIS, 1*(2).

Bhatt, G.D. (2001). Knowledge management in organizations: Examining the interaction between technologies, techniques, and people. Knowledge Management Journal, 5(1), 68-75.

Bullinger, H. -J., Wörner, K., & Prieto, J. (1997). *Wissensmanagement heute: Daten, fakten, trends*. Fraunhofer IAO, Stuttgart

Comelli, G., & von Rosenstiel, L. (2001). Führung durch motivation. Mitarbeiter für organisationsziele gewinnen. Vahlen, München.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston.

Disterer, G. (2000). Individuelle und soziale Barrieren beim Aufbau von Wissenssammlungen. *Wirtschaftsinformatik, 42*(6), 539-546.

Eppler, M. (2001). Increasing information quality through knowledge management systems services. In *Proceedings of the 2001 International Conference on Informations Systemens and Engineering*. Las Vegas, NV: IEEE Print. Retrieved December 14, 2007, from http://www.knowledgemedia.org/modules/pub/view.php/knowledgemedia-21

Hislop, D. (2004). *Knowledge management in organizations.* Oxford: Oxford University Press.

ISO. (1998). *Article on usability*. Retrieved December 14, 2007, from http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail? CSNUMBER=16883

De Judicibus, D. (1996, January 8-9). Reuse–a cultural change. In *Proceedings of the International Workshop on Systematic Reuse*, Liverpool.

Malhotra, Y. (2005). *Knowledge management: Rethinking management for the new world of uncertainty and risk. An interview ith Dr. Yogesh Malhotra*. Management First, April 2. Emerald Publishing.

Maier, R. (2004). *Knowledge management systems*. Berlin: Springer Verlag.

Mertins, K. (2003). *Knowledge management: Best practices in Europe*. Berlin, Heidelberg: Springer-Verlag.

Nemeth, C.J., & Nemeth, L. (2001). Understanding the creative process: Management of the knowledge worker. In J. Nonaka & D. J. Teece (Eds.), *Managing industrial knowledge* (pp. 91-104). London: Sage.

Noam, E. (1986). *The impact of information technologies on the service sector*. Ballinger.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company: How Japanese companies create the dynamics of innovation*. New York.

Pawar, K., Horton, A., Gupta, A., Wunram, M., Barson, R.J., & Weber, F. (2001, July 28-August 1). Interorganisational knowledge management: Focus on human barriers in the telecommunications industry. In *Proceedings of the 8th ISPE International Conference on Concurrent Engineering: Research and Applications*.

Probst, G., Raub, S., & Romhardt, K. (1998). *Wissen Managen. Wie Unternehmen ihre wertvollste Ressource optimal nutzen* (2nd Ed.). Wiesbaden.

Richter, A. (2006). *The closed-loop integration of knowledge into operational processes exemplified with the software os/rooms–an approach to overcome barriers in knowledge management*. Chair of Business Informatics and Systems Engineering, University of Augsburg.

Steinmüller, W. (1993). *Informationstechnologie und gesellschaft*. Darmstadt.

Wiig, K.M. (1995). *Knowledge management methods: Practical approaches to managing knowledge*. Arlington.

KEY TERMS

Controlled Vocabulary: Standardized terms used in searching a specific database. These terms can differ for each database. Using controlled vocabulary to search will provide you with more focused results.

Knowledge: Knowledge is defined according to Steinmüller (1993) as the combination or connection of information.

Knowledge Management: Knowledge Management (KM) "(…) refers to the critical issues of organizational adaptation, survival and competence against discontinuous environmental change. Essentially it embodies organizational processes that seek synergistic combination of data and information processing capacity of information technologies, and the creative and innovative capacity of human beings" (Malhotra, 2005).

Knowledge Management System (KMS): KMS describe information systems that are designed to support certain KM processes like the dissemination or application of knowledge.

Personification Strategy: In contrast to the technical strategy, the personification or social strategy focuses on tacit knowledge and leveraging knowledge exchange on the interpersonal level.

Technical Strategy: The technical strategy implies that knowledge is not solely embedded in humans, but can be provided to knowledge users—after having gone through the process of explication—in codified form.

Usability: The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use.

ENDNOTE

[1] Cf. (Bullinger et al., 1998, p.30].

Category: Business Information Systems

Benefits Realization through the Treatment of Organizational Issues

Neil F. Doherty
Loughborough University, UK

Malcolm King
Loughborough University, UK

INTRODUCTION

Information technology is now a ubiquitous and increasingly critical part of the fabric of the modern organization, supporting its day-to-day operations and all aspects of the decision-making process, as well as its strategic positioning. It is therefore not perhaps surprising that the implementation of a new technology or information system is likely to result in a wide array of impacts to the organization as well as the working lives of individual employees. There is a growing consensus within the literature that many such impacts are not deterministic and cannot therefore be easily predicted prior to a system's implementation (e.g., DeSanctis & Poole, 1994). The corollary of this is that many of the consequences of an information system's implementation will be unanticipated (Robey & Boudreau, 1999). While some of these unanticipated consequences, or incidental side effects, may be of a positive nature, negative impacts are also quite common, as IT-induced organizational change often results in user resistance and, in extreme cases, possibly even system rejection (Martinsons & Chong, 1999).

Information systems projects may not be totally predictable, but it can be argued that many of their organizational impacts only remain unanticipated, because systems developers are reluctant to tackle the human and organizational aspects of IT (Doherty & King, 2005). Systems development projects have typically been viewed as exercises in technical change, rather than socio-technical change; "most investments in IT are technology-led, reflecting too technical an emphasis" (Clegg, 2000, p. 464). This is a dangerous strategy, because unforeseen and unresolved negative impacts may increase the likelihood of systems failure. Moreover, beneficial impacts, of both a planned and incidental nature, may not be fully realized without an appropriate program of organizational change. Indeed, Ward and Daniel (2006) argue convincingly that the unacceptably high levels of IT failures are largely due to the absence of formal "benefits realization" approaches that explicitly target the organizational change needed to deliver business benefits. Consequently, we would argue that if systems development projects are viewed as an exercise in organizational change, in which all potential organizational impacts are proactively and systematically analyzed, then many undesirable impacts could be avoided, while the planned benefits can be more effectively realized (Doherty & King, 2002). The importance of treating organizational issues may now be widely acknowledged (e.g., Clegg, 2000; Eason, 2001), but little progress has been made in the development of practical treatment approaches that have succeeded in making the transition from research laboratory to widespread commercial usage. The primary aim of this article is to present an innovative new benefits-oriented approach for their proactive treatment. However, in advance of this, it is important to establish the importance of treating organizational issues.

BACKGROUND: THE NEED TO TREAT ORGANIZATIONAL ISSUES

The information systems' literature is very clear on two points; general levels of failure are far too high, and the primary cause of this problem is the failure to adequately treat organizational issues (Clegg, Axtell, et al., 1997; Doherty & King, 2001). In this context, the term "organizational issue" relates to those organizationally-oriented facets of systems development projects that need to be treated to ensure that the resultant impacts of an information system are likely to be desirable. A comprehensive checklist of important organizational issues, that was originally drawn from the literature but then validated over a series of studies (e.g., Doherty & King, 2001; Doherty, King, & Al-Mushayt, 2003), is presented in Table 1.

To treat a specific organizational issue it is necessary to first evaluate the likely organizational impact associated with it, and then if necessary take steps to ensure that the resultant impact is likely to be desirable. For example, if it is found that a proposed system is likely to be poorly suited to an organization's working practices, then it will be necessary to either modify the system's technical specification, so that the mismatch is avoided, or redesign the working practices so that they are well aligned with the system. In essence, the treatment of organizational issues is the mechanism by which the project team should align the capabilities afforded, and the constraints imposed, by the technical system with

Table 1. Checklist of organizational issues to address

Issue	Description
Information systems strategy	The system's alignment with the current information system strategy.
Current business needs	The system's ability to satisfy the organization's current business needs.
Prioritization of needs	The prioritizing of development effort on those aspects that address the most important business needs.
Future needs of organization	The system's ability to satisfy the organization's likely future business needs.
Process design	The system's impact on the design of key business processes.
Health & safety/ergonomic factors	The likely ergonomic and health and safety implications of the system, such as RSI and eye strain.
User motivation/needs	The system's ability to satisfy user needs and support user motivations.
User working styles and personal skills	The implications of user working styles and personal skills for the system's design and ongoing use.
Job redesign	The proposed system's impact on the design of working practices.
Timing of Implementation	The interaction of the system's implementation with other planned concurrent changes.
Organizational disruption	The temporary organizational disruption that may be caused by the implementation of the proposed system.
Organizational structure	The system's effect on the organizational structure, and the lines of authority.
Organizational culture	The proposed system's impact on the culture in the organization (*i.e., the set of important assumptions [often unstated] that members of an organization share in common).*
Organizational power	The proposed system's political implications for the distribution of power in the organization.

the requirements and characteristics of an organization and its individual employees.

System developers typically view the system development process as a science, rather than art, which requires the use of structured methods that focus upon the delivery of technically effective systems, on time and within budget. They are extremely reluctant to tackle intangible, ill-defined, and politically-sensitive organizational issues (Doherty & King, 2001), for which they are ill-equipped, in terms of training, competencies, and motivation (Clegg, 2000). Consequently, approaches to the treatment of organizational issues have typically been reactive rather than proactive (Clegg, Coleman, et al., 1996)—get the system implemented and then worry about its organizational impacts. There is therefore a pressing need to find ways to encourage the systems development community to become more actively engaged in the treatment of organizational issues. One obvious strategy is through the creation of methods, tools, and techniques that are specifically designed to facilitate the treatment of organizational issues. A wide variety of organizationally-oriented approaches have now been proposed, which can be categorized as follows:

1. **Socio-Technical Methods:** Socio-technical methods that attempt to produce information systems that are technically efficient and coherent, while also being sensitive to organizational and human needs, for example, ethics (Mumford, 1996) or multi-view (Avison, Wood-Harper, Vidgen, & Wood, 1998).
2. **Tools and Techniques for the Treatment of Specific Issues:** Many researchers (e.g., Clegg, Coleman, et al., 1996) have attempted to develop tools and techniques to aid in the treatment of specific organizational issues.
3. **An Organizational Impacts Analysis:** The "organizational impact analysis" (e.g., Sauer, 1993) is typically a one-off study to determine the ways in which a proposed system will affect the organization's decision-making, power, structure, culture, working practices, and so forth.

While each of these contributions has been very useful in increasing our understanding of the nature and treatment of organizational issues, there is little evidence that these contributions have made much of an impact on the practice of systems development (Clegg, 2000). This is probably, at least in part, due to technical specialists' continuing preference for the more technically oriented tools and techniques. However, if a comprehensive, coherent, and easy to use approach could be found, which complemented their existing methods, then it might have a greater chance of adoption.

The remainder of this section describes one such approach, which can best be described as an example of organizational impact analysis.

AN APPROACH FOR THE TREATMENT OF ORGANIZATIONAL ISSUES

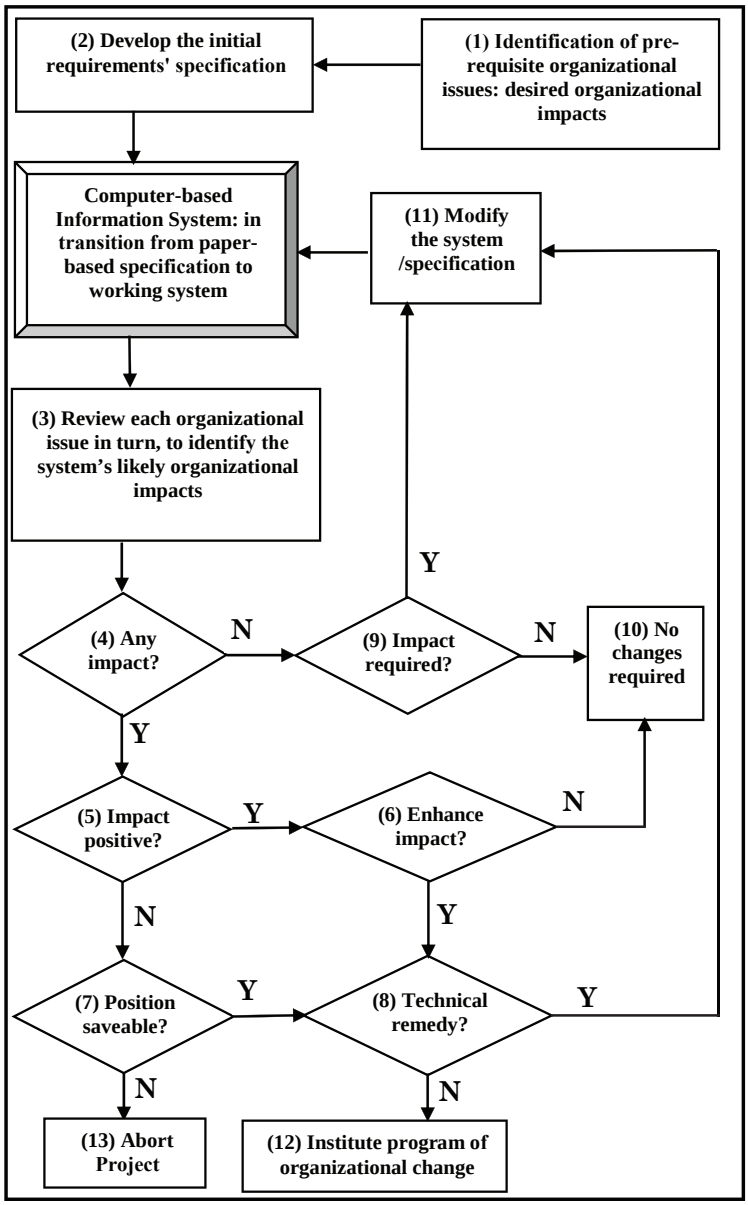
The proposed approach has been formulated from an extensive review of the literature and the authors' experience working in this domain for the past few years (e.g., Doherty & King, 2001; Doherty, King, & Al-Mushayt, 2003). A schematic representation of the approach, which has been conceptualized as a flow diagram, is presented in Figure 1. Each of the major processes and decision points on this diagram, all of which have been numbered, is reviewed below:

1. **Identification of Planned Organizational Impacts:** Many organizational issues will not be treated until a systems development project is well under way, but others, such as the system's ability system to satisfy "current organizational needs," will need to be considered right at the very outset if they are to be the planned outputs of the project. Indeed, unless the required benefits of a particular IT application are clearly articulated and agreed at the outset of the systems development project, the chances of any meaningful benefits being ultimately realized are extremely slim.
2. **Development of Initial Requirements Specification:** The development of the initial requirement's specification is a fundamental component of all development methods, and can be conducted using the proprietary or in-house method of the developer's choosing.

While the previous two stages occur only once, at the project's outset, it is envisaged that the following stages will be repeated at key stages in the systems development process, for each of the organizational issues in the checklist (see Table 1), in turn.

3. **Review of Organizational Issues:** Assess the system's likely impacts, with regard to each organizational issue. The process is designed to ensure that the planned impacts will ultimately come to fruition, while any incidental impacts will be effectively identified and managed.
4. **Determine Existence of Organizational Impacts:** The output of the review procedure, described previously, will be an assessment of whether there is a significant organizational impact associated with each organizational issue.
5. **Evaluation of Desirability of Impacts:** The desirability of each identified impact must be assessed to determine whether it is likely to be of a positive or negative nature. For example, an assessment of the system's impact on the motivation of users might identify a negative impact, such as user resistance or resentment, due to changes in their work practices. Potential solutions might be of a technical nature, such as changing the system's design to give the user more control over their work, or an organizational orientation, for example, improving the users' terms and conditions by way of compensation.
6. **Assessment of Potential for Increasing Desirability of Impacts:** If the previous stage has identified a desirable impact associated with an organizational issue, it is important to consider whether the impact could be made even more positive if the information system design were to be modified.
7. **Is the Situation Retrievable?:** In situations where a potentially undesirable impact of the system's operation has been identified, it is necessary to consider whether the situation is retrievable or not.
8. **Is the Remedy Technical?:** Having identified potentially negative, yet retrievable, impacts associated with the system implementation, a decision must be made as to whether the remedy is of a technical or organizational nature.
9. **Evaluation of Potential for Impacts:** If it has been discovered that there is no impact associated with a particular organizational issue, then it is important to question whether there should be an impact. If, for example, it has been determined that the system is unlikely to change working practices, then questions might be raised as to whether the system could be used to streamline or enrich the design of jobs.
10. **No Changes Required:** In the cases where there is no actual impact or potential for any specific organizational issue, there is no requirement to either change the system's specification or to institute a program of organizational change.
11. **Modification of Specification:** In many situations the organizational issues review process will necessitate changes to the system's specification in order to correct a negative impact or evoke a more positive impact.
12. **Development of Program of Organizational Change:** In situations where organizational impacts have been identified that have been judged to be desirable, it is important that a program of organizational change is planned and implemented to ensure that the impact is realized. In particular, the program of organizational change should be viewed primarily as a means of ensuring that all planned benefits are proactively managed, and ultimately realized (Ward & Elvin, 1999).
13. **Abort Project:** In situations where it has been found that the introduction of an information system is likely

Figure 1. An approach to the treatment of organizational issues

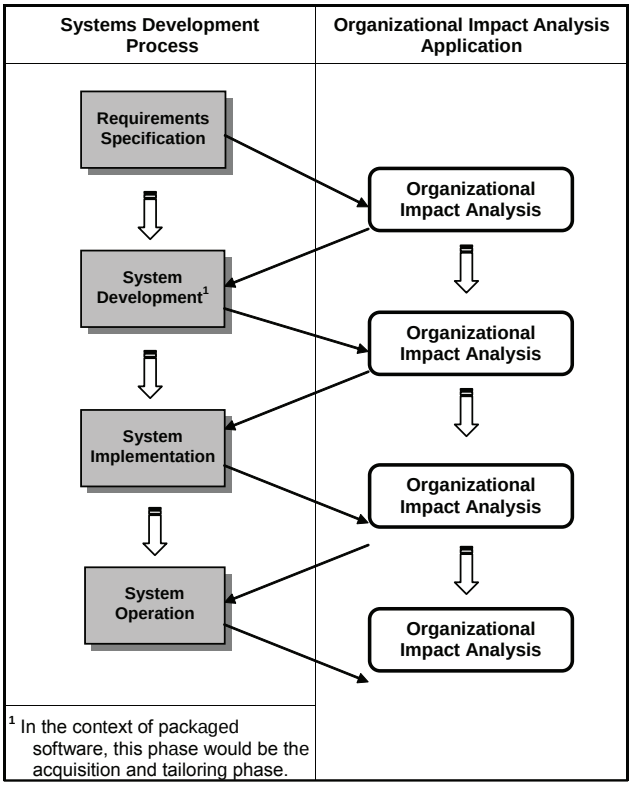


to result in significant organizational impacts of a negative nature, the project should be aborted.

At a minimum, it is envisaged that the organizational impacts will be assessed using this approach at the following key stages of the development process (see Figure 2): on completion of the requirements specification, then again at the end of the development phase, and then very soon after implementation. However, for very complex projects, or those that are likely to be the catalyst for significant organizational change, it is recommended that the analysis should be repeated more frequently, particularly during the design phase. Moreover, it is recommended that the organizational impact analysis be repeated a number of times over the system's working life. This would be one way of operationalizing Orlikowski, Yates, Okamura, and Fujimoto's (1995, p. 424) concept of "*technology-use mediation,*" which they define as:

deliberate, ongoing and organizationally sanctioned intervention within the context of use that helps to adapt new technology to its context, modifies the context as appropriate to accommodate the use of the technology, and facilitates the ongoing effectiveness of the technology over time.

Figure 2. The relationship between the systems development process and the application of the organizational impact analysis



It should be noted that once the system goes live, the organizational impact analysis procedure will not need to address the full range of organizational issues, as some—such as prioritization, timing of implementation, and organizational disruption—will no longer be relevant in an operational context.

While there will be a number of separate applications of the organizational impact analysis, they are not totally independent, as it is envisaged that the outcomes and findings of any given iteration of the approach might prompt or flag-up issues that need to be addressed in a subsequent iteration. It should also be remembered that many organizational impacts are interdependent, and consequently, changes made with respect to one specific issue might engender impacts in other areas. One of the obvious benefits of adopting an iterative approach is that it allows changing requirements to be monitored and accommodated, on an ongoing basis.

In terms of who is involved in the analysis of the impacts and the identification of appropriate changes to the system or the organization, to ensure the impacts will ultimately be desirable, it is recommended that the exercise be very inclusive. As a key objective of socio-technical approaches is to achieve consensus amongst all the system's stakeholders, it is envisaged that the proposed approach will act as a mechanism for channeling a debate about organizational change. As Markus and Benjamin (1997, p. 55) put it, "the hard reality of IT-enabled transformation is that change is everyone's job."

FUTURE TRENDS

As the scope and strategic importance of information systems continues to grow, so to does the need to find better ways of matching them to their organizational context, to ensure that they deliver a significant contribution to organizational performance. Consequently, there is an urgent need for practical and easy to use methods to aid systems development professionals in managing the organizational impacts of the systems for which they are responsible. To this end, our immediate priority is to test the provisional framework on a variety of systems development projects and use the feedback from these exercises to further define exactly how different organizational impacts can best be analyzed and managed.

CONCLUSION

Many information systems researchers have recognized the need for more effective socio-technical tools and methods to be developed. While the work described in this article is not trying to develop a highly specific tool or technique, it does propose a more general framework to promote the systematic and coherent treatment of organizational issues. The chief benefits of the proposed approach are that it presents systems developers with a systematic framework that obliges them to confront organizational issues and provides them with the means to effectively navigate their way through a very complex decision-making process. In particular, the application of this approach should ensure that systems development teams are encouraged to focus on benefits realization, as well as IT solution delivery (Ward & Daniel, 2006). Moreover, a comparison of this approach with some of its predecessors allows the following distinctions to be made:

- **Complementary:** The proposed approach complements, rather than replaces, existing development tools and methods. There is no requirement for systems developers to abandon their tried and tested practices.
- **Straightforward:** The approach adopts a commonsense perspective, and it should therefore be relatively easy to learn and apply.
- **Proactive:** By using this approach organizations will ensure that potential problems are recognized and opportunities are identified and exploited in a timely and effective manner.
- **Comprehensive:** The approach is comprehensive and can cope with a wide range of potential impacts.
- **Flexible:** The approach is highly flexible and can be adapted to suite the requirements of a wide variety of information systems projects

In essence, this approach is inviting systems developers to periodically stand back from the systems development process and, in conjunction with a wide variety of stakeholders, assess the likely impacts of their work on the design and operation of the organization. While there are a number of potential benefits to the proposed approach, it is also important to issue a health warning: these ideas are provisional and exploratory, and there is much further work required to translate them into a robust and reliable tool.

REFERENCES

Avison, D., Wood-Harper, A. T., Vidgen, R. T., & Wood, J. R. G. (1998). A further exploration into information systems development: The evolution of Multiview2. *Information Technology and People, 11*(2),124-139.

Clegg, C. W. (2000). Socio-technical principles for system design. *Applied Ergonomics, 31*(5), 463 - 477.

Clegg, C. W., Axtell, C., Damadoran, L., Farbey, B., Hull, R., Lloyd-Jones, R., et al. (1997). Information technology: A study of performance and the role of human and organizational factors. *Ergonomics, 40*(9), 851-871.

Clegg, C. W., Coleman, P., Hornby, P., McClaren, R., Robson, J., Carey, N., et al. (1996). Tools to incorporate some psychological and organizational issues during the development of computer-based systems. *Ergonomics, 39*(3), 482-511.

DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization, 5*(2), 121-147.

Doherty, N. F., & King, M. (2001). An investigation of the factors affecting the successful treatment of organizational issues in systems development projects. *European Journal of Information Systems, 10*(3), 147-160.

Doherty, N. F., & King, M., (2002). From technical change to socio-technical change: Towards a proactive approach to the treatment of organizational issues. In S. Clarke, E. Coakes, M. G. Hunter, & A. Wenn (Eds.), *Socio-technical and human cognition elements of information systems* (pp. 22-40). Hershey, PA: Information Science Publishing.

Doherty, N. F., King, M., & Al-Mushayt, O. (2003). The impact of inadequacies in the treatment of organizational issues on information systems development projects. *Information & Management, 41*(1), 147-160.

Doherty, N. F., & King, M. (2005). From technical to socio-technical change: Tackling the human and organizational aspects of systems development projects. *European Journal of Information Systems, 14*(1), 1-5.

Eason, K., (2001). Changing perspectives on the organizational consequences of information technology. *Behaviour & Information Technology, 20*(5), 323-328.

Ewusi-Mensah, K., & Przasnyski, Z. (1994). Factors contributing to the abandonment of information systems development projects. *Journal of Information Technology, 9*(3), 185-201.

Lyytinen, K., & Hirschheim, R. (1987). Information systems failures: A survey and classification of the empirical literature. *Oxford Surveys in Information Technology, 4*, 257-309.

Markus, M. L., & Benjamin, R. I. (1997). The magi bullet theory in IT-enabled transformation. *Sloan Management Review, 38*(2), 55-68.

Martinsons, M., & Chong, P. (1999). The influence of human factors and specialist involvement on informations systems success. *Human Relations, 52*(1), 123-152.

Mumford, E. (1996). *Systems design: Ethical tools for ethical change*. Basingstoke, UK; Hampshire, UK: MacMillan.

Orlikowski, W. J., Yates, J., Okamura, K., & Fujimoto, M. (1995). Shaping electronic communication—The metastructuring of technology, in the context of use. *Organization Science, 6*(4), 423-444.

Robey, D., & Boudreau, M-C. (1999). Accounting for the contradictory organizational consequences of information technology: Theoretical directions and methodological implications. *Information Systems Research, 10*(2), 176-185.

Sauer, C. (1993). *Why information systems fail: A case study approach*. Henley Upon Thames, UK; Oxfordshire, UK: Alfred Waller.

Ward, J., & Daniel, E. (2006). *Benefits management*. Chichester, UK: John Wiley & Sons.

Ward, J., & Elvin, R. (1999). A new framework for managing IT-enabled business change. *Information Systems Journal, 9*(3), 197-222.

KEY TERMS

Benefits Realization: The process of proactively managing benefits to ensure that all the potential benefits that may arise from the introduction of a new information technology are ultimately realized.

Incidental Impacts: Impacts that are unplanned by-products of the system's development process that had not, or could not, have been envisaged at the project's outset.

Organizational Impact Analysis: A one-off study to determine the ways in which a proposed system will affect the organization, in areas such as power, structure, culture, working practices, and so forth.

Organizational Issues: Those issues that need to be treated during the systems development process to ensure that the individual human, wider social, and economic impacts of the resultant computer-based information system are likely to be desirable.

Planned Impacts: The anticipated outcomes of a systems development project that were identified at the project's outset, and are typically critical to its ultimate success.

Socio-Technical Methods: Development methods that attempt to produce systems that are both technically efficient and organizationally sensitive.

Systems Failure: Systems abandoned before completion, systems completed but never used, under-used, or failing to deliver key aspects of functionality, and projects that are significantly over budget or schedule.

Best Practices for IS&T Supervisors

Debra A. Major
Old Dominion University, USA

Valerie L. Morganson
Old Dominion University, USA

INTRODUCTION

Researchers over the last decade have generated a body of literature which is informed by management research and theory and tailored to the unique demands that characterize IS&T work. At the industry level, IS&T fluctuates with the supply and demand asymmetry caused by technological advances (Agarwal & Ferratt, 2002a). The changing nature of the industry trickles down to affect IS&T professionals who must continually update their skills in order to prevent obsolescence (Rajeswari & Anantharaman, 2003). IS&T work demands flexibility in responding to customer demands, emerging issues, spontaneously hectic workloads, and frequently unplanned requests. The nature of the work is continuous (frequently 24/7) and often requires the coordination of multiple experts. IT is typically a service function upon which other organizational functions depend. Yet, it is common for IT to be undervalued and unrecognized, unless there is an IT failure. IS&T work may be performed by individuals or teams that may be colocated or virtually connected. Although there has been some debate in defining the parameters of the so-called "IS&T workforce," considerable overlap in skills, educational backgrounds and other domains persist (Kaarst-Brown & Guzman, 2005). The current article defines IS&T professionals as individuals whose primary job function is the development, installation, and implementation of computer systems or communication technology. Research and best practices literature are reviewed to provide IT managers with an overview and a starting point for workforce intervention and improvement.

BACKGROUND

IT human capital is seen as a strategic resource and competitive advantage for businesses (Bhardwaj, 2000). Where the IT workforce was once inundated, many researchers and practitioners have raised concern about a shortage of skilled professionals and have noted a corresponding research focus on IT turnover (Agarwal, Ferratt, & De, 2007; Niederman, Moore, Yager, 2002). Prior to the popping of the IT bubble, the abundance of IT workers permitted managers to focus on motivating extant staff. Because the industry has stabilized and labor is in shorter supply, managers have been required to heed turnover, a more longitudinal goal, in addition to maintaining production levels.

Table 1. IS&T best practices taxonomies

Human Resource Practices Agarwal & Ferratt (2002a)	Supervisory Practices Major et al. (2007)
Performance Measurement	*Task-focused Practices*
Compensation & Benefits Systems	Boundary Spanning
Work Arrangements	Performance Management
Employability Training	Employee Involvement
Longer-term Career Development	Training & Development
Opportunities for Advancement	
Opportunities for Recognition	*Person-focused Practices*
Quality of Leadership	Relationship Building
Sense of Community	Mentoring
Lifestyle Accommodations	Stress Management
Organizational Stability & Employment Security	Work-family Balance

Agarwal and Ferratt (2002a) and Major et al. (2007) have combined survey and interviewing methodologies to empirically derive taxonomies of best human resources management (HRM) and supervisory practices for IT (see Table 1). Although both taxonomies address the issue of effectively managing IS&T professionals, Agarwal and Ferratt approach the issue from a more global HRM systems perspective, while Major et al. focus on the practices of individual supervisors.

Reminiscent of classical leadership theory, two dimensions emerge in each taxonomy. One is focused on work and output itself , "task-focused leadership practices" (Major et al., 2007), or "productivity concerns" (Agarwal & Ferratt, 2002a). This aspect emphasizes performance management, employee involvement, and training and development. The other attends to the individual needs of the worker. At the macrolevel, this refers to attending to employee needs through human resources, an "interpersonal dimension" (Agarwal & Ferratt, 2002a). Similarly, "person-focused practices" refer to meeting individuals' social needs and maintaining interpersonal relationships though supervisor-subordinate interaction (Major et al., 2007). These two dimensions reflect the state of the IT industry, which has recently required a dual management focus on both immediate (motivational) behavior and longitudinal (employee continuance) behavior.

Because Major et al. and Agarwal and Ferratt's work provide two holistic and complimentary taxonomies and hold a strong empirical base, the current article derives its structure by identifying commonalities between the two perspectives. In addition, the findings of these articles are used to reference practical examples. Our aim is to create an integrative perspective of effective IT management practices.

PERSON-FOCUSED PRACTICES

Relationship Building

Upon hire, employees enter into a relationship with their organization. As with any relationship, it requires reciprocity. Employees make an investment in their organization (e.g., labor and effort) and hold expectations of the company in return. The perceived exchange relationship between employees and their organizations is referred to as a psychological contract (see Agarwal & Ferratt, 2000; Rousseau, 2001). Managers are responsible for communicating and upholding the employer's end of the contract and may assist the organization to enjoy the benefits of the relationship (Agarwal & Ferratt, 2002b).

IT supervisors should maintain open communication with their subordinates for motivational purposes. Research suggests that face-to-face communication is preferred over other methods (e.g., e-mail) by IS&T professionals and their supervisors as means of interpersonal relationship building (Major et al., in press). Face-to-face communication is especially important, at least on occasion, for employees working at a distance (Davis & Bryant, 2003). Through good communication aimed at establishing strong relationships, managers can mitigate job stressors such as customer service demands, tight deadlines, and understaffing (Major et al., 2007). Given the nature of IS&T work, direct supervision may not be possible. Managers must rely on trust, mutual respect, and loyalty instead. This is especially important to the effective functioning of team arrangements, distributed work in virtual teams, and telework (Costa, 2003; Davis & Bryant, 2003). Subordinate trust can be developed through open communication, honesty and follow-through (Korsgaard, Brodt, & Whitener, 2002).

IT research has especially advocated the implementation of mentor-based systems (e.g., supervisory mentoring). Mentors assist employees in identifying career choices, required competencies and training needs. Effective mentors also provide psychosocial support and serve as role models (Major et al., 2007; also see Scandura & Ragins, 1993). Research suggests that supervisor/subordinate mentoring relationships may be especially advantageous (see Payne & Huffman, 2005). In addition to providing mentoring themselves, effective IS&T supervisors also facilitate peer mentoring among IT professionals (Major et al., 2007).

Embeddedness

Organizations and departments are infused with information technology. Thus, IT professionals are required to interface with a variety of other organizational departments and functions. These clients make technology requests, frequently without understanding what is required to fulfill their demands, or while underestimating the time and resources required. The situation may be exacerbated by client-held stereotypes of the IT professional and role (Guzman et al., 2004). In order to prevent interdepartmental conflict, effective supervisors work proactively to educate and build relationships with other departments and monitor the work environment to prevent sudden conflicts from arising. Ideally, the supervisor should assist subordinates and clients in maintaining sight of how the client's needs and IT interactions fit within the context of the organization's needs and goals holistically. Effective participative practices in an IS&T setting include involving employees in informal meetings with multiple layers of management to allow them to see how their work fits into a larger context (Agarwal & Ferratt, 2002a) and seeking employee feedback in both one-on-one and team settings (Major et al., 2007).

Supervisors should also be proactive in gaining a seat at the table for organizational planning. This entails marketing IT capabilities to customers to communicate how IT

can be used to facilitate the attainment of organizational goals (Major et al., 2007). This process of monitoring the organizational environment and crossing departmental functions to anticipate and seek out implications for one's own department is referred to as boundary spanning. In a meta-analysis that empirically combined the results of 51 studies of IT turnover intentions, the practice of boundary spanning was shown to be an important factor in preventing IT professionals from intending to leave their organization (Joseph & Ang, 2003)

Attending to Employee Well-Being

IS&T work, characterized by heavy customer service demands, on-call duties, frequent understaffing, and bursts of activity, may create stress for employees and challenge them to keep a healthy balance between their work and nonwork lives. Effective supervisors address key sources of stress directly through interventions such as additional staffing, advocating for subordinate needs to decision makers, and monitoring the environment for changes with IT implications. Supervisors also assist employees in coping with stress by engaging in open communication, prioritizing and monitoring workloads for fair distribution, and encouraging coworker support (Major et al., 2007).

Effective supervisors recognize the reciprocal influences between employees' work and nonwork lives. Because IS&T work often infringes on family and personal life, effective supervisors use formal (e.g., vacation time) and informal methods (e.g., flexible arrival and departure times) to help employees attend to their nonwork demands (Major et al., 2007). Agarwal and Ferratt (2002a) noted that effective IT organizations had lifestyle accommodations in place and encouraged their use. Examples included a relaxed work environment, flexible work arrangements, child care services, and telecommuting. Even when family-friendly policies are on the books, they are unlikely to be used if they are not supported by the work environment (Allen, 2001).

In an IT sample, policies that permitted flexibility for workers' personal lives were found to have an influence on commitment and turnover intentions. Based on the findings, the authors suggested that work-life concerns should be considered complimentary to other important organizational practices (e.g., recognition, competence development, and empowerment (Paré, Tremblay, & Lalonde, 2001). Employers that engage in fair treatment and invest in employee well-being reinforce employees' expectations that they will be treated fairly throughout their tenure (Moorman, Blakely, & Niehoff, 1998).

PRODUCTIVITY-FOCUSED PRACTICES

Empowerment

Involving IT professionals in management decisions is a common theme in the IT literature. In order for participation to be motivating, IS&T professionals must see that their participation actually has an influence or makes a difference in the work environment (Major, Davis, Sanchez-Hucles, Germano, & Mann, 2006; Major & Germano, 2006). Aside from the motivational benefits of empowerment, supervisors may rely upon their subordinates as technical experts. Often the supervisor lacks certain types of technical expertise and depends on subordinates' skills to supplement their own (Major et al., 2007). Due to the complex nature of IT work, pooling knowledge from multiple experts provides considerable advantage over authoritarian leadership styles. By allowing subordinates to become involved in making decisions, supervisors are providing learning and growth opportunities, teaching self-reliance, and fostering empowerment.

Training and Development

According to a recent survey of hiring managers conducted by the ITAA, the value placed on highly skilled IT professionals is increasing, and IT managers are aware of the strategic significance of individuals' skills portfolios (ITAA, 2004). Researchers have recognized the role of training and development for increasing retention, organizational performance, building competitive advantage and dealing with staff shortages (e.g., Agarwal & Ferratt, 1999; Hopp, Tekin, & Van Oyen, 2004; Major et al., 2007; Paré et al., 2001). Likewise, IS&T professionals hold a stake in continually updating their skills in order to remain marketable for career development purposes (Rajeswari & Anantharaman, 2003). Competency development practices may communicate to employees that the organization considers them to be an investment toward competitive advantage and seeks to establish a long-term relationship with them (Agarwal & Ferratt, 1999). IT professionals viewed competency development to be principally important; it predicted feeling committed (affective commitment) to the organization, which, in turn was related to turnover (Paré et al., 2001). Empirically derived qualitative data has revealed several training and development practices employed by effective managers, including tuition reimbursement, computer-based training programs, leadership workshops, mentoring, provision of enriching job activities (e.g., on-the-job-training or challenging projects) and facilitation of learning-supportive environments, job rotation, cross-functional teams, supervisor-to-subordinate coaching, facilitation of peer training, opportunity to participate in organizational committees and cultural awareness

training (Agarwal & Ferratt, 1999; Agarwal & Ferratt, 2002a, Major et al., 2007; Paré et al., 2001).

Managing Performance

Research has demonstrated that feedback rich environments—those where feedback is frequent, specific and positive—are linked to individual perceptions of feedback accuracy, and in turn, desire to respond to feedback (Kinicki, Prussia, Wu, & McKee-Ryan, 2004). Empirical evidence supports that effective IT organizations complete annual (or more frequent) performance appraisals (Agarwal & Ferratt, 2002a). Unfortunately, IS&T personnel evaluation practices are seen as inadequate in today's organizations (Chilton & Hardgrave, 2004). Accommodating user needs is an essential job function; yet, organizations have struggled to identify a criterion to account for this performance dimension. Traditional system development performance was deemed inadequate and has more recently been replaced with subjective judgment and supplementary measures (Saarinen, 1996). To tend to the criterion problem, Boyd, Huang, Jiang, and Klein (2007) identified several core literature-based performance dimensions to guide performance ratings of IS&T professionals. These include work quality, project work, general task, interpersonal quality, dependability, teamwork and leadership, and career-related training. They also recommended using a 360-degree feedback approach to include key stakeholders in the performance of groups or individuals (e.g., users and developers). Indeed, there is some validation for this suggestion. In a large sample, 360-feedback systems, including multiple raters, was identified as a best practice (Agarwal & Ferratt, 2002a).

Beyond provisioning formalized performance feedback, effective supervisors accentuate the linkages between the organization's objectives, subordinate performance, and the achievement of desired rewards (Major et al., 2007). Furthermore, recognizing individual contribution through nonmonetary means (e.g., extended vacations, awards, and outings) has been empirically linked to supervisory effectiveness and employee commitment to the organization (Major et al., 2007; Paré et al., 2001).

FUTURE TRENDS

Research describing the demands of IS&T work is abundant, and researchers are directing their attention to the question of best management practices for IS&T professionals facing these circumstances. Agarwal and Ferratt (2002a) addressed this question with regard to upper-level management and human resources practices. Major et al. (2007) explore effective human resource management from the perspective of first-level IT supervisors. These two taxonomies highlight the need for a cohesive human resources management strategy. Human resource policies and practices sensitive to the needs of IS&T professionals must be supported by top management. Middle managers and first-level supervisors must be supported and rewarded for their efforts to implement policies and encourage their use. Much like the IS&T professionals they supervise, first-level managers need the flexibility and discretion to assess the needs of their individual employees and provide appropriate accommodations. As the closest and most visible representative of the organization for IS&T professionals, a supervisor's role in effective human resource management cannot be underestimated. Ensuring that supervisors possess the requisite skills for effectively managing IS&T personnel and understanding how to adequately motivate supervisors to engage in best practices are key issues for research and practice.

Research shows that best practices for managing IS&T professionals are the same practices found to be effective for other technical professions (Ferratt & Short, 1988). Moreover, practices that are effective for supervising IS&T professionals are consistent with decades of leadership effectiveness research (Major et al., 2007). Still, IT-specific research is warranted. Researchers should continue to examine which sets of practices are needed to dovetail with the unique demands of the work environment to attain optimal functioning.

CONCLUSION

In reviewing the IS&T literature, a dichotomy of person- and productivity-focused practices emerges. On one hand, themes in the literature regarding person-focused practices include building relationships, creating embeddedness, and attending to employee well-being. On the other hand, productivity is comprised of empowerment, training and development, and managing performance. Person- and productivity-focused practices are each associated with their own respective objectives. The former is mainly concerned with achieving employee staying behavior, whereas the latter is more closely associated with product or service output. Thus, while all of the practices discussed in the current article are potential points of leverage for IT managers, some may take precedence over others, depending upon an organization's internal and external conditions. Where there is less stability and capacity to forecast of the conditions to come, production goals tend to be a more appropriate focus. For longer time horizons, where there is stability and constancy within the department and business, longer-term objectives, such as employee development and other person-focused practices, may come into focus as well (Jackson & Schuller, 1995). Resources permitting, the best practices described in this article should

be used together. In combination, they may have synergistic benefits (Ferratt, Agarwal, Brown, & Moore, 2005).

REFERENCES

Allen, T.D. (2001). Family-supportive work environments: The role of organizational perceptions. *Journal of Vocational Behavior, 58*, 414-435.

Agarwal, R., & Ferratt, T.W. (1999). *Coping with labor scarcity in information technology: Strategies and practices for effective recruitment and retention*. Cincinnati, OH: Pinnaflex Educational Resources.

Agarwal, R., & Ferratt, T.W. (2000). Retention and the career motives of IT professionals. In *Proceedings of the 2000 ACM SIGCPR Conference on Computer Personnel Research,* Chicago, IL, (pp. 158-166).

Agarwal, R., & Ferratt, T.W. (2002a). Enduring practices for managing IT professionals. *Communications of the ACM, 45*(9), 73-79.

Agarwal, R., & Ferratt, T.W. (2002b). Toward understanding the relationship between IT human resource management systems and retention: An empirical analysis based on multiple theoretical and measurement approaches. In *Proceedings of the 2002 ACM SIGCPR Conference on Computer Personnel Research,* Kristiansand, Norway, (pp. 126-138).

Agarwal, R., Ferratt, T.W., & De, P. (2007). An experimental investigation of turnover intentions among new entrants in IT. *ACM SIGMIS Database, 38*(1), 8-28.

Banks, C.G., & May, K.E. (1999). Performance management: The real glue in organizations. In A.I. Kraut & A.K. Korman (Eds.), *Evolving practices in human resource management* (pp. 118-145). San Francisco: Jossey-Bass.

Bharadwaj, A.S. (2000). A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly, 24*(1), 169-196.

Boyd, M., Huang, S., Jiang, J.J., & Klein, G. (2007). Discrepancies between desired and perceived measures of performance of IS professionals: Views of the IS professionals themselves and the users. *Information & Management, 44*, 188-195.

Chilton, M.A., & Hardgrave, B.C. (2004). Assessing information technology personnel: Toward a behavioral rating scale. *ACM SIGMIS Database, 35*(3), 88-104.

Costa, A.C. (2003). Work team trust and effectiveness. *Personnel Review, 32*, 605-622.

Davis, D.D., & Bryant, J. (2003). Leadership in global virtual teams. In W.H. Mobley & P.W. Dorfman (Eds.), *Advances in global leadership* (Vol. 3, pp. 303-340). Amsterdam: JAI.

Ferratt, T.W., Agarwal, R., Brown, C.V., & Moore, J.E. (2005). IT human resource management configurations and IT turnover: Theoretical synthesis and empirical analysis. *Information Systems Research, 16*(3), 427-443.

Ferratt, T.W., & Short, L.E. (1988). Are information systems people different? An investigation of how they are and should be managed. *MIS Quarterly, 12*, 427-443.

Guzman, I.R., Stanton, J.M., Stam, K.R., Vijayasri, V., Yamodo, I., Zakaria, N., & Caldera, C.A. (2004). A qualitative study of the occupational subculture of information systems employees in organizations. In *Proceedings of SIGMIS Conference on Computer Personnel Research,* Tucson, AZ, (pp. 74-80).

Hopp, W.J., Tekin, E., & Van Oyen, M.J. (2004). Benefits of skill chaining in serial production lines with cross-trained workers. *Management Science, 50*(1), 83-98.

Information Technology Association of America (ITAA). (2004, September). *Adding value… growing careers: The employment outlook in today's increasingly competitive job market*. Arlington, VA: Author.

Jackson, S.E., & Schuler, R.S. (1995). Understanding human resource management in the context of organizations and their environment. *Annual Review of Psychology, 46,* 237-264.

Joseph, D., & Ang, S. (2003). Turnover of IT professionals: A quantitative analysis of the literature. In *Proceedings of the 2003 SIGMIS Computer Personnel Research Annual Conference,* Philadelphia, PA, (pp. 130-132).

Kaarst-Brown, M.L., & Guzman, I.R. (2005). Who is "the IT workforce"? Challenges facing policy makers, educators, management, and research. In *Proceedings of the 2005 ACM SIGMIS CPR Conference on Computer personnel research,* Atlanta, GA, (pp. 1-8).

Kinicki, A.J., Prussia, G.E., Wu, B., & McKee-Ryan, F.M. (2004). A covariance structure analysis of employee's response to performance feedback. *Journal of Applied Psychology, 89,* 1057-1069.

Korsgaard, M.A., Brodt, S.E., & Whitener, E.M. (2002). Trust in the face of conflict: The role of managerial trustworthy behavior and organizational context. *Journal of Applied Psychology, 87*, 312-319.

Major, D.A., Davis, D.D., Germano, L.M., Fletcher, T.D., Sanchez-Hucles, J., & Mann, J. (2007). Managing human resources in information technology: Best practices of high

performing supervisors. *Human Resource Management, 46*(3), 411-427.

Major, D.A., Davis, D.D., Sanchez-Hucles, J., Germano, L.M., & Mann, J. (2006). IT workplace climate for opportunity and inclusion. In E.M. Trauth (Ed.), *Encyclopedia of gender and information technology* (Vol. 2, pp. 856-862). Hershey, PA: Idea Group Reference.

Major, D.A., & Germano, L.M. (2006). Survey feedback interventions in IT workplaces. In E.M. Trauth (Ed.), *Encyclopedia of gender and information technology* (Vol. 2, pp. 1134-1141*)*. Hershey, PA: Idea Group Reference.

Moorman, R.H., Blakely, G.L., & Niehoff, B.P. (1998). Does perceived organizational support mediate the relationship between procedural justice and organizational citizenship behavior? *Academy of Management Journal, 41,* 351-357.

Niederman, F., Moore, J.E., & Yager, S.E. (2002). A view from the SIGCPR conference: What have we learned this decade? *ACM SIGCPR Computer Personnel, 20*(4), 75-89.

Paré, G., Tremblay, M., & Lalonde, P. (2001). Workforce retention: What do IT employees really want? In *Proceedings of the 2001 ACM SIGCPR Conference on Computer Personnel Research,* San Diego, CA, (pp. 1-10).

Payne, S.C., & Huffman, A.H. (2005). A longitudinal examination of the influence of mentoring on organizational commitment and turnover. *Academy of Management Journal, 48*, 158-168.

Rajeswari, K.S., & Anantharaman, R.N. (2003). Development of an instrument to measure stress among software professionals: Factor analytic study. In *Proceedings of the 2003 ACM SIGMIS Conference on Computer Personnel Research,* Philadelphia, PA, (pp. 34-43).

Rousseau, D.M. (2001). Schema, promise and mutuality: The building blocks of the psychological contract. *Journal of Occupational and Organizational Psychology, 74,* 511-541.

Saarinen, T. (1996). An expanded instrument for evaluating information systems success. *Information & Management, 31*(2), 103-118.

Scandura, T.A., & Ragins, B.R. (1993). The effects of sex and gender role orientation on mentoring in male-dominated occupations. *Journal of Vocational Behavior, 43*, 251-265.

KEY TERMS

Boundary Spanning: Monitoring the organizational environment and crossing departmental functions to anticipate and to proactively seek out implications for one's own department.

Embeddedness: Incorporating and establishing (a) the IT function across departments and (b) IT employees within the organization. For embeddedness to occur, IT departments and workers must participate in the organization, hold a network of partnerships (e.g., friendships and alliances), possess influence, and experience positive interactions within the organization.

Performance Management: Applying techniques such as role clarification, goal setting, performance appraisal, and performance-related rewards in order to connect individual behavior and organizational strategies and goals (Banks & May, 1999).

Psychological Contract: The reciprocal exchange relationship that is perceived to exist between employees and their organizations.

Relationship Building: Developing a dynamic of reciprocity and trust between an employee and his or her (a) organization, (b) coworkers or (c) supervisor.

Training and Development: Providing opportunities for employees to acquire and to improve technical and interpersonal job-related skills and knowledge for personal growth and career advancement.

Work-Life Balance: Achieving a suitable harmony between the frequently incompatible duties of work and home life that many workers face, particularly when working in jobs with irregular or long hours.

Better Executive Information with the Dashboard Approach

Frédéric Adam
University College Cork, Ireland

Jean-Charles Pomerol
Université Pierre et Marie Curie, France

INTRODUCTION

After more than 30 years of research on how the work of managers can be supported by computers, the observation that developing computer systems that are truly useful for top management is a highly complex and uncertain task is still as valid as ever. Information systems for executives raise specific problems, which have primarily to do with the nature of managerial work itself (Mintzberg, 1973), as they are intended to tackle the needs of users whose most important role is "to create a vision of the future of the company and to lead the company towards it" (King, 1985, p. xi).

BACKGROUND

The major difficulty in supporting managers with computer systems comes from the very nature of management work (Mintzberg, 1973, 1975, 1976), which is concerned with communication, coordination, and people's management for more than 80%. At the time of his research, Mintzberg (1973) had noted how little time is left for reflection and for "playing" with computer systems. This has been a significant difficulty from the origins of MIS systems because their primarily "operational" focus was not central to executives' concerns (Ackoff, 1967; Keen & Scott Morton, 1978). Twenty years later, this difficulty has also been largely responsible for the shift from decision support systems (DSSs) to executive information systems (EISs). EISs were intended to be very easy to use and to help users manipulate required data without the need for much training, which would be very attractive to top executives who want to have, at a glance, a very comprehensive view of their business. Specific descriptions of the differences between DSSs, EISs, and cooperative decision systems can be found in Pomerol and Brézillon (1998). Naturally, computer literacy among executives has increased to a great extent, notably thanks to the development of electronic mail and the World Wide Web. However, whatever designs were put forward over the years, it has remained true that managers are not inclined to spend countless hours browsing computer data, such is the time pressure under which they operate.

Beyond the time pressures under which executives must operate, there are issues of trust and of credibility of the information that can be found in a computer system, which mitigate against intensive executive reliance on information systems, especially in a long-term perspective. First of all, the lack of confidence of executives in their models has been noted by many researchers (e.g., Wallenius, 1975; Cats-Baril & Huber, 1987; Abualsamh, Carlin & McDaniel, 1990). The idea that decision makers need sophisticated models may actually be wrong. People in charge of the preparation of decisions would probably be able to understand and use smart models, but the high-level executives who most commonly make the final decisions are far too busy to train with and use involved systems. On the contrary, they appear to prefer simple systems that they trust and understand, and that display very timely simple information. More often, the data required to make the best decisions will already reside in some form or another in the database of the organization or can be captured with an online feed into a computer system, and what is really needed is a device to filter and display and to warn executives about the most important variances (Simon, 1977). As noted by Kleinmutz (1985): "the ability to select relevant variables seems to be more important than procedural sophistication in the processing of that information" (p. 696).

In EIS, the underlying models built into the system are normally very simple and easily understandable, which is a great help in increasing the acceptability of a computer system.

To conclude, the specificities of managerial decision making can be synthesized as follows:

- Most decisions are made very quickly under considerable time pressure (except some strategic decisions).
- Strategic decision making is often the result of collaborative processes.
- Most decisions are linked to individuals who have specific intentions and commitments to personal principles and ideas.

It is therefore very difficult to support managers, and despite many years of research, little is known about the way information systems could support such unstructured tasks.

A VEHICLE FOR INFORMATION REQUIREMENTS ANALYSIS: CRITICAL SUCCESS FACTORS

In pre-EIS days, Rockart (1979) put forward a methodology called critical success factors or CSF to guide information systems planning. The method had its advantages, though it failed to make a general impact on the planning process of organizations. Its potential in other areas, notably the development of information systems, has been explored by a number of researchers. It is argued in this article that it can be very useful as a guide for the development of executive systems, as both from an information content perspective as for the design of the interface of these systems.

CSF assumes that the performance of organizations can be improved by focusing on "the few key areas where things must go right for the business to flourish" (Rockart, 1979). In simple terms, the method seeks to isolate, using the expertise and gut feeling of managers, the factors which may make the difference between success and failure for the firm.

A number of key points about CSF make it a very attractive technique. First of all, while CSF is essentially a generic framework, it recognizes that all firms are different and operate in different markets. Thus, CSFs are different for different organizations. Secondly, the CSF theory takes into account that the needs of managers within the same organizations are also different based on their hierarchical level, but more importantly, based on their style and their specific areas of responsibility. In general, there are only a limited number of factors that each manager should monitor closely, and this guarantees that managers can concentrate their limited attention to factors that really matter and that are within their control. The attractive thing about this breakdown of responsibility is that the CSF sets controlled by the different managers add up to a complete organizational set that covers all the key areas of the business.

Van Bullen and Rockart (1986) identified a number of primary categories of CSF that are useful in guiding the analysis of the organizational CSF set. These generic sources of CSFs are: (1) the industry where the organization operates (these CSFs are shared by mainstream organizations in this industry), (2) the competitive position and strategy pursued by the organization (which are unique to its set of circumstances and objectives set by its top managers), (3) the environmental factors surrounding the organization (which it has no control over, but which it must monitor closely to compete), (4) temporal factors (which relate to specific events or change programs currently facing the organization, and require the temporary monitoring of additional factors), and finally, (5) CSFs that are specific to each manager and their role in the company. Other authors have added other potential sources such as CSFs related to the analysis of main competitors (especially industry leaders) and the evolution of their business (Leidecker & Bruno, 1984). These sources add up to a wealth of potential factors and measurements that are sufficient for effective monitoring of the business of most organizations.

Dashboards and Control Rooms

In the next stage of the development of executive systems, designers must create an interface for displaying the CSFs. The design of this interface is nearly as important as the selection of the indicators in shaping the perception of managers of the usefulness of their information systems and keeping their interest in the long run. One technique that has worked well in selecting and presenting indicators is the application of the dashboard concept to the management of organizations.

Fundamentally, the concept of dashboard reflects the application of the concept of control room to the management of the firm and echoes the call for a warning or exception reporting functionality in EIS-type systems. In engineering, the control room is a specially designed physical area of a plant where the proper operation of key equipment can be monitored. Control rooms have developed because of the need to monitor increasingly complex processes, such as petrol refining or the operation of nuclear power plants. The control room allows operators to control a process without looking at it with their own eyes, and with a degree of accuracy and completeness that could not be achieved with human perception alone.

This suggests that dashboards may be developed that considerably help managers in their day-to-day search for problems and matching solutions. Naturally, the nature of management itself is highly dynamic and diverse and involves consideration of infinite number of parameters in a way that is fundamentally different from the monitoring of a manufacturing process. Thus, management has a significant "human interaction" component that cannot easily be supported by computer systems. Simon (1977), Gorry and Scott Morton (1971), and others have commented comprehensively on the degree to which managerial decisions are programmable or not, however it remains that the implementation of many of the objectives of the firm, however elusive, can be monitored using a dashboard-type interface. Further, the CSF method can be a powerful vehicle in selecting the indicators to be shown on each manager's dashboard.

The difficulty with CSF-based dashboards resides in the operationalization of managers' key concerns and the identification of specific targets for CSF monitoring, the

design of measurement logic for each indicator, and in the development of the interfaces that can be used by managers to easily and effectively review the performance of the firm in relation to each of the indicators.

TOWARDS A METHODOLOGY FOR DASHBOARD DEVELOPMENT

At the height of the EIS movement, King (1985) remarked:

"It is so easy to lose sight of reality—to believe that the computer model's numerical forecasts are real and that they describe future outcomes that will, in fact, come to pass…The computer model's forecasts are based solely on those predictions about the future that we are able to quantify. Those things that are not readily quantifiable are usually omitted, and in being omitted there is a danger that they may be ignored." (p. xi)

This illustrates the dangers inherent in approaching management based solely on numbers, however obtained. This also explains why observational plant tours are still regarded as one of the most reliable methods for collecting data in manufacturing environments (Jones, Saunders & McLeod, 1988). The basis of any dashboard approach to management must therefore take into account the following four key issues:

(1) *Limited Attention:* Given the limited attention of managers and the costs inherent in sourcing certain data, the indicators displayed on the dashboard must be carefully selected using the CSF.

(2) *Performance Measurement:* The measurements used to monitor indicators or CSFs are crucial. The usefulness and effectiveness of the dashboard is totally dependent on the accuracy of the data used and the realism of the calculations presented to managers.

(3) *Operator Training:* It is critical that managers understand the assumptions built into the dashboard and the algorithms used to reach the results presented to them. They must also be fully aware of how data are collected and what limitations applied to the accuracy of the measurements. For instance, drill down facilities can make the difference between "using information to manage more intelligently and more effectively and making the same old mistakes but with more speed" (Meall, 1990).

(4) *Dashboard Layout:* The layout of the dashboard has a direct impact on the understanding derived by managers. The interface of the dashboard must be consistent so that managers can visualize immediately where they should focus their attention as a matter of priority. Exception reporting and color coding (Watson, Rainer & Koh, 1991) can be used to achieve maximum visual impact.

It is also useful if the development of the dashboard can be operationalized as an evolutionary activity where managers can feed back their perception of the dashboards to developers so that the design can be improved over time and the indicators refined or replaced (as in the case of temporal CSFs). In this article, we propose a framework based on 11 questions to support this evolutionary process, and help managers and developers to establish a fruitful dialogue:

- **Question 1: Who will use the indicators?**
  The answer may not be simple when not one, but a number of individuals are interested in monitoring certain indicators. However, managers should concentrate on monitoring the parameters most closely associated with their own performance or that of the areas directly under their control.
- **Question 2: Can be mapped out to a specific objective at a higher level?**
  In the perspective of a top-down CSF exercise, indicators are mapped out to specific objectives pursued by top management. In a bottom-up scenario, it will help if indicators can be merged into higher level composite indicators presented to higher level managers. Developers can use the hierarchy of indicators as a blueprint for the drill-down facility to be built into the dashboard so that top managers can understand the underlying causes of poor or good performance.
- **Question 3: How frequently will managers need to monitor each indicator?**
  Managers' perception of how frequent significant or revelatory variations are likely to occur should be used as a guide for deciding how frequently indicators should be updated. The scope of the benefits that may arise as a result of the monitoring should also be considered if high costs are likely to be incurred.
- **Question 4: What calculation methods are available? What unit of measurement will be used?**
  The choice of calculation method can greatly influence the variation of an indicator and shift the burden of responsibility from one area to another. It can also influence the way the performance of operators or workshops is measured.
  The choice of the unit of measurement is normally straightforward for quantitative analysis, but can become far more complex for less tangible CSFs that involve the estimations of qualitative factors. Customer satisfaction, for instance, will require vision and creativity if it is to be measured properly. Some quantitative measures may be applicable such as the number of complaints received per time interval, but

other measures may have to be found that can act as surrogates of customer satisfaction.

- **Question 5: What data source exists? What should be created?**
  Certain data may be missing from existing organizational information systems. Other data may reside in a proprietary system (e.g., a custom-built process control system) that does not integrate well with other systems. Significant investment in equipment and special devices (such as scanners and sensors) or in software such as OLAP and ROLAP (Relational OLAP) may have to be made.
- **Question 6: How detailed should the analysis presented in the dashboard be? How can the indicators be broken down to be more meaningful?**
  Many indicators are too broad to be suitably presented as one figure, and some disaggregating may be required. Typical organizations sell multiple products in multiple markets. Thus, sales figures need to be disaggregated to present £ figures, volumes, and variances for each product on each market while also presenting the aggregated data. Multi-dimensional modeling can be used to support the organization and retrieval of such data.
- **Question 7: What threshold values should be used to differentiate between adequate and inadequate performance? What comparisons can be made to assess the company's performance?**
  Absolute measurement figures presented by a dashboard may not be meaningful to managers unless they can be examined in light of other data. Most companies already have a tight budget system in place, and this can be used as a source of normative values.
- **Question 8: How can each indicator be represented for maximum visual impact?**
  Developers must seek to reduce information overload and use the latest graphical user interface (GUI) technology. Some software tool boxes are now available to help designers create displays and objects that mirror the type of controls normally found on dashboards. Gauges with specific color-coded threshold values can easily be created, and special charts can be made clickable to build intuitive drill down into the data. These speed up and facilitate the data reading of managers.
- **Question 9: What action must be taken when good or bad performance is measured? Is there scope for corrective action to be taken based on the indicator?**
  Whenever good or bad results are presented, managers should know what avenues can be pursued. Reporting mechanisms (e.g., electronic mail) can be built into the dashboard to facilitate and accelerate the dissemination of interesting results and their discussion. In the longer term, increased familiarity with indicators and what their evolution means should have practical decision-making implications for all managers and staff. Thus, users' reaction times to certain signals should be reduced and their responses should improve, especially in recurrent situations.
- **Question 10: How will indicators be monitored/archived in the long term?**
  A key element of our approach is the learning that can be achieved when CSFs are monitored over long periods of time. Staff and managers learn from regularly sampling their performance and that of their areas, and seeing it compared to other data, such as budgets and previous performance of industry standards. Greater learning will be derived if managers and staff set time aside to review and discuss indicators on a regular basis.
- **Question 11: Is there any potential bias inherent in the methods and data used for calculations? What incentives are being given to staff?**
  The development of new performance measurement systems, such as a dashboard of indicators, should always be guided by consideration of the incentives given to actors and the behavior likely to result from the implementation of the underlying indicators. There may also be a change management side to the project, as managers negotiate with staff the implementation of the system. Staff may object to a certain type of measurement (which they may perceive to be threatening or invasive) or the implementation of devices dedicated to monitoring their work.

CONCLUSION

The case for managing the firm solely based on numbers has already been argued and lost. At this point, it is well established that managing firms cannot and should not be compared to the administration of a power plant. This does not mean, however, that the concept of control room does not have potential when applied to the management of organizations. Faced with increasingly complex situations and responsibility for the administration of increasingly complex business processes, managers have less and less time to spend monitoring the key factors of the business. The development of a dashboard can speed up this process and help managers catch far more information than they normally would without assistance.

Following the steps highlighted in this article will also give organizations a much better idea of what parameters they should worry about and how to measure performance. Peter Swasey, one of the directors of the Bank of Boston, commented that "what you don't measure, you don't manage" (McGill, 1990). The preparatory analysis work on the CSFs of the firm will give much confidence to organizational actors

that they understand their business and have a comprehensive hold upon its vital functions, and the dashboard ultimately developed will provide flexible and speedy access to vital information, thereby freeing time for other key activities such as business or staff development. As a by-product, managers may also be able to use the analysis carried out for their dashboard as a blueprint for the incentive systems of their company.

REFERENCES

Abualsamh, R., Carlin, B. & McDaniel Jr., R.R. (1990). Problem structuring heuristics in strategic decision making. *Organizational Behavior and Decision Process, 45,* 159-174.

Ackoff, R.L. (1967). Management MISinformation systems. *Management Science, 14*(4), 147-156.

Cats-Baril, W.L. & Huber, G. (1987). Decision support systems for ill-structured problems: An empirical study. *Decision Science, 18,* 350-372.

Gorry, A. & Scott Morton, M. (1971). A framework for management information systems. *Sloan Management Review,* (Fall), 55-70.

Jones, J., Saunders, C. & McLeod, R. (1988). Information media and source patterns across management levels: A pilot study. *Journal of Management Information Systems, 5*(3), 71-84.

Keen, P.G. & Scott Morton, M.S. (1978). Decision support systems: An organizational perspective. Reading, MA: Addison-Wesley.

King, W.R. (1985). Editor's comment: CEOs and their PCs. *Management Information Systems Quarterly, 9,* xi-xii.

Kleinmutz, D.N. (1985). Cognitive heuristics and feedback in a dynamic decision environment. *Management Science, 31,* 680-702.

Leidecker, J. & Bruno, A. (1984). Identifying and using critical success factors. *Long Range Planning, 17*(1), 23-32.

McGill, P. (1990). Executive support systems. *Business Quarterly*¸ (Summer).

Meall, L. (1990). EIS: Sharpening the executives' competitive edge. *Accountancy,* (September).

Mintzberg, H. (1973). *The nature of managerial work.* New York: Harper and Row.

Mintzberg, H. (1975). The manager's job: Folklore and fact. *Harvard Business Review,* (July/August), 49-61.

Mintzberg, H. (1976). Planning on the left side and managing on the right. *Harvard Business Review,* (July/August), 120-130.

Pomerol, J.-Ch. & Brézillon, P. (1998). From DSSs to cooperative systems: Some hard problems still remain. In R. Dolk (Ed.), *Proceedings of HICCS 31* (IEEE Publication, Volume 5, pp. 64-71).

Rockart, J. (1979). Chief executives define their own data needs. *Harvard Business Review, 57*(2), 81-93.

Simon H. (1977). *The new science of management decisions.* Englewood Cliffs, NJ: Prentice-Hall.

Van Bullen, C. & Rockart, J. (1986). A primer on critical success factors. In J. Rockart & C. Van Bullen (Eds.), *The rise of management computing.* Homewood, IL: Dow Jones Irwin.

Wallenius, J. (1975). Comparative evaluation of some interactive approaches to multi-criterion optimization. *Management Science, 21,* 1387-1396.

Watson, H.J, Rainer, K.R. Jr. & Koh, C.E. (1991). Executive information systems: A framework for development and a survey of current practices. *MIS Quarterly, 15*(1), 13-50.

KEY TERMS

Control Room: A special location in a plant where operators can monitor a process in great detail without having to physically be looking at it. This is particularly useful in dangerous environments.

Critical Success Factors: A methodology for managing projects and firms that concentrates on the areas where things must go right if the endeavor is to flourish.

Dashboard: Specific display of information that presents key information about a process or device. A dashboard may or may not be computerized.

Evolutionary Design: System development methodology where an ongoing approach is taken to analyzing the requirements of the application.

Interface: Portion of a computer application that is used by the user to communicate with the application. It is particularly important for a dashboard, because it may impinge on the ability of users to properly interpret the variations in the indicators shown to them.

Managing by Numbers: A school of thought that sought to demonstrate that firms could be managed solely based on watching key (mostly financial) indicators. This is now largely discredited.

Model: A simplified representation of reality that concentrates on predicting how a factor or a series of related factors would evolve based on the variation of a set of parameters. Also, a simplified representation of reality.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 266-271, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Bibliomining for Library Decision-Making

Scott Nicholson
Syracuse University, USA

Jeffrey Stanton
Syracuse University, USA

INTRODUCTION

Most people think of a library as the little brick building in the heart of their community or the big brick building in the center of a campus. These notions greatly oversimplify the world of libraries, however. Most large commercial organizations have dedicated in-house library operations, as do schools, non-governmental organizations, as well as local, state, and federal governments. With the increasing use of the Internet and the World Wide Web, digital libraries have burgeoned, and these serve a huge variety of different user audiences. With this expanded view of libraries, two key insights arise. First, libraries are typically embedded within larger institutions. Corporate libraries serve their corporations, academic libraries serve their universities, and public libraries serve taxpaying communities who elect overseeing representatives. Second, libraries play a pivotal role within their institutions as repositories and providers of information resources. In the provider role, libraries represent in microcosm the intellectual and learning activities of the people who comprise the institution. This fact provides the basis for the strategic importance of library data mining: By ascertaining what users are seeking, bibliomining can reveal insights that have meaning in the context of the library's host institution.

Use of data mining to examine library data might be aptly termed *bibliomining*. With widespread adoption of computerized catalogs and search facilities over the past quarter century, library and information scientists have often used bibliometric methods (e.g., the discovery of patterns in authorship and citation within a field) to explore patterns in bibliographic information. During the same period, various researchers have developed and tested data mining techniques—advanced statistical and visualization methods to locate non-trivial patterns in large data sets. Bibliomining refers to the use of these bibliometric and data mining techniques to explore the enormous quantities of data generated by the typical automated library.

BACKGROUND

Forward-thinking authors in the field of library science began to explore sophisticated uses of library data some years before the concept of data mining became popularized. Nutter (1987) explored library data sources to support decision-making, but lamented that "the ability to collect, organize, and manipulate data far outstrips the ability to interpret and to apply them"(p. 143). Johnston and Weckert (1990) developed a data-driven expert system to help select library materials and Vizine-Goetz, Weibel, and Oskins (1990) developed a system for automated cataloging based on book titles (also see Aluri & Riggs, 1990; Morris, 1991). A special section of *Library Administration and Management* ("Mining your automated system") included articles on extracting data to support system management decisions (Mancini, 1996), extracting frequencies to assist in collection decision-making (Atkins, 1996), and examining transaction logs to support collection management (Peters, 1996).

More recently, Banerjeree (1998) focused on describing how data mining works and ways of using it to provide better access to the collection. Guenther (2000) discussed data sources and bibliomining applications, but focused on the problems with heterogeneous data formats. Doszkocs (2000) discussed the potential for applying neural networks to library data to uncover possible associations between documents, indexing terms, classification codes, and queries. Liddy (2000) combined natural language processing with text mining to discover information in "digital library" collections. Lawrence, Giles, and Bollacker (1999) created a system to retrieve and index citations from works in digital libraries. Gutwin, Paynter, Witten, Nevill-Manning, and Frank (1999) used text mining to support resource discovery.

These projects all shared a common focus on improving and automating two of the core functions of a library—acquisitions and collection management. A few authors have recently begun to address the need to support management by focusing on understanding library users: Schulman (1998) discussed using data mining to examine changing trends in library user behavior; Sallis, Hill, Jance, Lovetter, and Masi (1999) created

a neural network that clusters digital library users; and Chau (2000) discussed the application of Web mining to personalize services in electronic reference.

The December 2003 issue of *Information Technology and Libraries* was a special issue dedicated to the bibliomining process. Nicholson (2003) presented an overview of the process, including the importance of creating a data warehouse that protects the privacy of users. Zucca (2003) discussed an implementation of a data warehouse in an academic library. Wormell (2003), Suárez-Balseiro, Iribarren-Maestro, and Casado (2003), and Geyer-Schultz, Neumann, and Thede (2003) used bibliomining in different ways to understand use of academic library sources and to create appropriate library services.

We extend these efforts by taking a more global view of the data generated in libraries and the variety of decisions that those data can inform. Thus, the focus of this work is on describing ways in which library and information managers can use data mining to understand patterns of behavior among library users and staff and patterns of information resource use throughout the institution.

INTEGRATED LIBRARY SYSTEMS AND DATA WAREHOUSES

Most managers who wish to explore bibliomining will need to work with the technical staff of their integrated library system (ILS) vendors to gain access to the databases that underlie that system to create a data warehouse. The cleaning, pre-processing, and anonymizing of the data can absorb a significant amount of time and effort. Only by combining and linking different data sources, however, can managers uncover the hidden patterns that can help to understand library operations and users.

EXPLORATION OF DATA SOURCES

Available library data sources are divided in three groups for this discussion: data from the *creation* of the library, data from the *use of the collection*, and data from *external sources* not normally included in the ILS.

ILS Data Sources from the Creation of the Library System

Bibliographic Information

One source of data is the collection of bibliographic records and searching interfaces that represent materials in the library, commonly known as the Online Public Access Catalog (OPAC). In a digital library environment, the same type of information collected in a bibliographic library record can be collected as metadata. The concepts parallel those in a traditional library: take an agreed-upon standard for describing an object, apply it to every object, and make the resulting data searchable. Therefore, digital libraries use conceptually similar bibliographic data sources as traditional libraries.

Acquisitions Information

Another source of data for bibliomining comes from acquisitions, where items are ordered from suppliers and tracked until received and processed. Because digital libraries do not order physical goods, somewhat different acquisition methods and vendor relationships exist. Nonetheless, in both traditional and digital library environments, acquisition data have untapped potential for understanding, controlling, and forecasting information resource costs.

ILS Data Sources from Usage of the Library System

User Information

In order to verify the identity of users who wish to use library services, libraries maintain user databases. In libraries associated with institutions, the user database is closely aligned with the organizational database. Sophisticated public libraries link user records through zip codes with demographic information in order to learn more about their user population. Digital libraries may or may not have any information about their users, based upon the login procedure required. No matter what data is captured about the patron, it is important to ensure that the identification information about the patron is separated from the demographic information before storing this information in a data warehouse; this will protect the privacy of the individual.

Circulation and Usage Information

The richest sources of information about library user behavior are circulation and usage records. Legal and ethical issues limit the use of circulation data, however. This is where a data warehouse can be useful, in that basic demographic information and details about the circulation could be recorded without infringing upon the privacy of the individual.

Digital library services have a greater difficulty in defining circulation, as viewing a page does not carry the same meaning as checking a book out of the library, although requests to print or save a full text information resource might be similar in meaning. Some electronic full-text services already implement server-side capture of such requests from their user interfaces.

Searching and Navigation Information

The OPAC serves as the primary means of searching for works owned by the library. Additionally, because most OPACs use

a Web browser interface, users may also access bibliographic databases, the World Wide Web, and other online resources during the same session; all of this information can be useful in library decision-making. Digital libraries typically capture logs from users searching their databases and can track, through "clickstream" analysis, the elements of Web-based services visited by users. In addition, the combination of a login procedure and cookies allow connecting user demographics to the services and searches they used in a session.

External Data Sources

Reference Desk Interactions

In the typical face-to-face or telephone interaction with a library user, the reference librarian records very little information about the interaction. Digital reference transactions, however, occur through an electronic format, and the transaction text can be captured for later analysis, which provide a much richer record than is available in traditional reference work. The utility of these data can be increased if identifying information about the user can be captured as well, but again, anonymization of these transactions is a significant challenge.

Item Use Information

Fussler and Simon (as cited in Nutter, 1987) estimated that 75-80% of the use of materials in academic libraries is in-house. Some types of materials never circulate, and therefore, tracking in-house use is also vital in discovering patterns of use. This task becomes much easier in a digital library, as Web logs can be analyzed to discover what sources users examined.

Interlibrary Loan and other Outsourcing Services

Many libraries using Interlibrary Loan and/or other outsourcing methods to get items on a "just-in-time" basis for users. The data produced by this class of transactions will vary by service, but can provide a window to areas of need in a library collection.

FUTURE TRENDS

Bibliomining can provide understanding of the individual sources listed earlier; however, much more information can be discovered when sources are combined through common fields in a data warehouse.

Bibliomining to Improve Library Services

Most libraries exist to serve the information needs of users, and therefore, understanding those needs of individuals or groups is crucial to a library's success. For many decades, librarians have suggested works; market basket analysis can provide the same function through usage data to aid users in locating useful works. Bibliomining can also be used to determine areas of deficiency and predict future user needs. Common areas of item requests and unsuccessful searches may point to areas of collection weakness. By looking for patterns in high-use items, librarians can better predict the demand for new items.

Virtual reference desk services can build a database of questions and expert-created answers, which can be used in a number of ways. Data mining could be used to discover patterns for tools that will automatically assign questions to experts based upon past assignments. In addition, by mining the question/answer pairs for patterns, an expert system could be created that can provide users an immediate answer and a pointer to an expert for more information.

Bibliomining for Organizational Decision-Making within the Library

Just as the user behavior is captured within the ILS, the behavior of library staff can also be discovered by connecting various databases to supplement existing performance review methods. While monitoring staff through their performance may be an uncomfortable concept, tighter budgets and demands for justification require thoughtful and careful tracking of performance. In addition, research has shown that incorporating clear, objective measures into performance evaluations can actually improve the fairness and effectiveness of those evaluations (Stanton, 2000).

Low use statistics for a work may indicate a problem in the selection or cataloging process. Looking at the associations between assigned subject headings, call numbers and keywords along with the responsible party for the catalog record may lead to a discovery of system inefficiencies. Vendor selection and price can be examined in a similar fashion to discover if a staff member consistently uses a more expensive vendor when cheaper alternatives are available. Most libraries acquire works both by individual orders and through automated ordering plans that are configured to fit the size and type of that library. While these automated plans do simplify the selection process, if some or many of the works they recommend go unused, then the plan might not be cost effective. Therefore, merging the acquisitions and circulation databases and seeking patterns that predict low use can aid in appropriate selection of vendors and plans.

Bibliomining for External Reporting and Justification

The library may often be able to offer insights to their parent organization or community about their user base through patterns detected with bibliomining. In addition, library managers are often called upon to justify the funding for their library when

budgets are tight. Likewise, managers must sometimes defend their policies, particularly when faced with user complaints. Bibliomining can provide the data-based justification to back up the anecdotal evidence usually used for such arguments.

Bibliomining of circulation data can provide a number of insights about the groups who use the library. By clustering the users by materials circulated and tying demographic information into each cluster, the library can develop conceptual "user groups" that provide a model of the important constituencies of the institution's user base which can fulfill some common organizational needs for understanding where common interests and expertise reside in the user community. This capability may be particularly valuable within large organizations where research and development efforts are dispersed over multiple locations.

In the future, organizations that fund digital libraries can look to text mining to greatly improve access to materials beyond the current cataloging / metadata solutions. The quality and speed of text mining continues to improve. Liddy (2000) has researched the extraction of information from digital texts, and implementing these technologies can allow a digital library to move from suggesting texts that might *contain the answer* to just *providing the answer*, extracted from the appropriate text or texts. The use of such tools risks taking textual material out of context and also provides a few hints about the quality of the material, but if these extractions were links directly into the texts, then context could emerge along with an answer. This could provide a substantial asset to organizations that maintain large bodies of technical texts because it would promote rapid, universal access to previously scattered and/or uncataloged materials.

CONCLUSION

Libraries have gathered data about their collections and users for years, but have not always used those data for better decision-making. By taking a more active approach based on applications of data mining, data visualization, and statistics, these information organizations can get a clearer picture of their information delivery and management needs. At the same time, libraries must continue to protect their users and employees from misuse of personally identifiable data records. Information discovered through the application of bibliomining techniques gives the library the potential to save money, provide more appropriate programs, meet more of the user's information needs, become aware of gaps and strengths of their collection, and serve as a more effective information source for its users. Bibliomining can provide the data-based justifications for the difficult decisions and funding requests library managers must make.

REFERENCES

Atkins, S. (1996). Mining automated systems for collection management. *Library Administration & Management, 10*(1), 16-19.

Chau, M.Y. (2000). *Mediating off-site electronic reference services: Human-computer interactions between libraries and Web mining technology*. Fourth International Conference on Knowledge-based Intelligent Engineering Systems & Allied Technologies (vol. 2, pp.695-699). Piscataway, NJ: IEEE.

Chaudhry, A.S. (1993). Automation systems as tools of use studies and management information. *IFLA Journal, 19*(4), 397-409.

Doszkocs, T.E. (2000). Neural networks in libraries: The potential of a new information technology. Retrieved October 24, 2001, from *http://web.simmons.edu/~chen/nit/NIT%2791/027~dos.htm*

Geyer-Schulz, A., Neumann, A., & Thede, A. (2003). An architecture for behavior-based library recommender systems. *Information Technology and Libraries, 22*(4), 165-174.

Guenther, K. (2000). Applying data mining principles to library data collection. *Computers in Libraries, 20*(4), 60-63.

Gutwin, C., Paynter, G., Witten, I., Nevill-Manning, C., & Frank, E. (1999). Improving browsing in digital libraries with keyphrase indexes. *Decision Support Systems, 2I,* 81-104.

Johnston, M., & Weckert, J. (1990). Selection advisor: An expert system for collection development. *Information Technology and Libraries, 9*(3), 219-225.

Lawrence, S., Giles, C.L., & Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer, 32*(6), 67-71.

Liddy, L. (2000, November/December). Text mining. *Bulletin of the American Society for Information Science*, 13-14.

Mancini, D.D. (1996). Mining your automated system for systemwide decision making. *Library Administration & Management, 10*(1), 11-15.

Morris, A. (Ed.) (1991). *Application of expert systems in library and information centers*. London: Bowker-Saur.

Nicholson, S. (2003). The bibliomining process: Data warehousing and data mining for library decision-making. *Information Technology and Libraries, 22*(4), 146-151.

Nutter, S.K. (1987). Online systems and the management of collections: Use and implications. *Advances in Library Automation Networking, 1*, 125-149.

Peters, T. (1996). Using transaction log analysis for library management information. *Library Administration & Management, 10*(1), 20-25.

Sallis, P., Hill, L., Jance, G., Lovetter, K., & Masi, C. (1999). A methodology for profiling users of large interactive systems incorporating neural network data mining techniques. *Proceedings of the 1999 Information Resources Management Association International Conference* (pp. 994-998). Hershey, PA: Idea Group Publishing.

Schulman, S. (1998). Data mining: Life after report generators. *Information Today, 15*(3), 52.

Stanton, J.M. (2000). Reactions to employee performance monitoring: Framework, review, and research directions. *Human Performance, 13*, 85-113.

Suárez-Balseiro, C.A., Iribarren-Maestro, I., & Casado, E.S. (2003). A study of the use of the Carlos III University of Madrid library's online database service in scientific endeavor. Information Technology and Libraries, 22(4), 179-182.

Wormell, I. (2003). Matching subject portals with the research environment. *Information Technology and Libraries, 22*(4), 158-166.

Zucca, J. (2003). Traces in the clickstream: Early work on a management information repository at the University of Pennsylvania. *Information Technology and Libraries, 22*(4), 175-178.

KEY TERMS

Bibliometrics: The study of regularities in citations, authorship, subjects, and other extractable facets from scientific communication using quantitative and visualization techniques. This allows researchers to understand patterns in the creation and documented use of scholarly publishing.

Bibliomining: The application of statistical and pattern-recognition tools to large amounts of data associated with library systems in order to aid decision-making or justify services. The term "bibliomining" comes from the combination of bibliometrics and data mining, which are the two main toolsets used for analysis.

Data Mining: The extraction of non-trivial and actionable patterns from large amounts of data using statistical and artificial intelligence techniques. Directed data mining starts with a question or area of interest, and patterns are sought that answer those needs. Undirected data mining is the use of these tools to explore a dataset for patterns without a guiding research question.

Data Warehousing: The gathering and cleaning of data from disparate sources into a single database, optimized for exploration and reporting. The data warehouse holds a cleaned version of the data from operational systems, and data mining requires the type of cleaned data that lives in a data warehouse.

Integrated Library System: The automation system for libraries, combining modules for cataloging, acquisition, circulation, end-user searching, database access, and other library functions through a common set of interfaces and databases.

Online Public Access Catalog (OPAC): The module of the integrated library system designed for use by the public to allow discovery of the library's holdings through the searching of bibliographic surrogates. As libraries acquire more digital materials, they are linking those materials to the OPAC entries.

ENDNOTE

[1] This work is adapted from: Nicholson, S. & Stanton, J. (2003). Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. In H. Nemati, & C. Barko (Eds.), *Organizational data mining: Leveraging enterprise data resources for optimal performance (pp.247-262)*. Hershey, PA: Idea Group Publishing.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 272-277, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Bioinformatics

Biometric Authentication

Julien Mahier
ENSICAEN, France

Marc Pasquet
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Christophe Rosenberger
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Félix Cuozzo
ENSICAEN, France

INTRODUCTION

For ages, humans recognized themselves according to different characteristics (appearance, behavior…). Biometrics is a well known technique to identify an individual or verify its identity; as, for example, fingerprints have been used for more than 100 years to identify one criminal. With computers, this analysis can be realized very quickly and with a higher reliability.

Biometrics has many applications: site monitoring (Bird, Masoud, Papanikolopoulos, & Isaacs, 2005), e-commerce (Jain & Pankanti, 2006)… The main benefits of biometrics are to provide better security and to facilitate the authentication process for a user. For example, it can be easy to obtain the password of a user, but it is more difficult to look like the user if a face recognition system is used for the user verification. Biometrics can also provide many advantages for particular applications. Indeed, biometric authentication can be realized in a contactless way that could be important for cultural aspects or reasons of hygiene. For all these motivations, biometrics is an emergent technology that could be more present in our daily life.

The goal of this chapter is to make an overview of biometrics. We focus on the authentication process, whose goal is to verify the identity of an individual. Ideal biometric information must have multiple properties:

- **Universality:** all individuals must be characterized by this information;
- **Uniqueness:** this information must as dissimilar as possible for two different persons;
- **Permanency:** it should be present during the whole life of an individual;
- **Collectability:** it can be measured (in a easy way);
- **Acceptability:** it concerns the possibility of a real use by users.

The plan of this chapter is given below. The background part presents the different biometric modalities studied in the research labs and used in real conditions. The main thrust of this chapter is an analysis of the benefits and limitations of biometric authentication. We present also the general architecture of a biometric system. Future trends stress the different research topics that should be treated to improve the biometric authentication. It concerns the combination of different biometric systems and their performance evaluation. We conclude by resuming the main aspects of this domain.

BACKGROUND

We detail, in this section, the different biometric modalities from the state of the art that can be used for the authentication process. Figure 1 illustrates the different types of biometric information. Biological analysis exploits some information that is present for any alive mammal (DNA, blood…). The behavioral analysis is specific to a human being and characterizes the way an individual makes some daily tasks (gait, talking…). Last, morphological analysis uses some information on how we look (for another individual or for a particular sensor).

Biological Analysis

As mentioned previously, biological biometric information can be extracted from any human being (see Figure 2). Generally, this information is not very easy to obtain. Some particular sensors are needed and the extraction of the biometric information can be quite long. For example, the DNA analysis with the most recent research techniques is possible in a few hours. We can cite the blood analysis that can only differentiate two individuals with different

Figure 1. Different biometric information in the state of the art

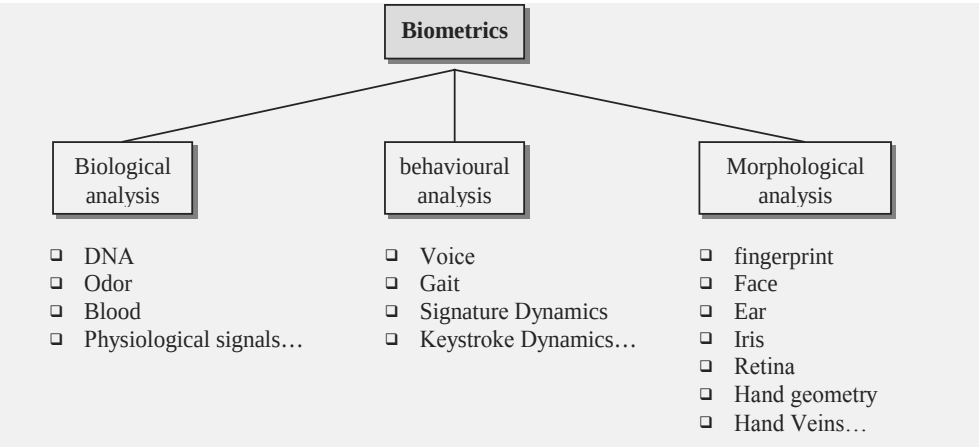
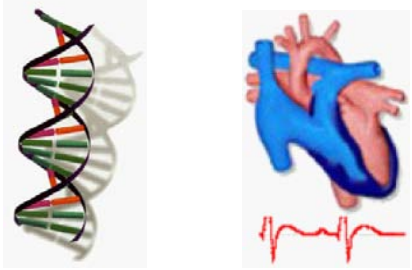


Figure 2. Some illustrations of the biological analysis (K. Phua et al., 2007)



Rhesus groups. We focus, in this chapter, on recent biological biometrics.

The electroencephalogram (brain signal) as a biometric information was studied in 1999 (Poulos, Rangoussi, Chrissikopoulos, & Evangelou, 1999). The EER value obtained by this approach is at the range of 16% to 28%. The odor as biometric information has also been tested (Korotkaya, 2003). This feasibility study showed the limitations of the actual sensors to use this information in a real context.

An industrial company proposed, in 2004, a biometric authentication solution based on dynamic electrophysiological characteristics of the living body, primarily of the beating heart (Idesia Ltd., 2004). Heart sound signals are also recently used for the authentication process (Phua et al., 2007). The biometric system comprises an electronic stethoscope, a computer equipped with a sound card, and the software application. This system provides a promising performance (EER equals to 4%).

Behavioral Analysis

We are able to recognize someone considering the way he/she walks or the way he/she types with a keyboard (see Figure 3).

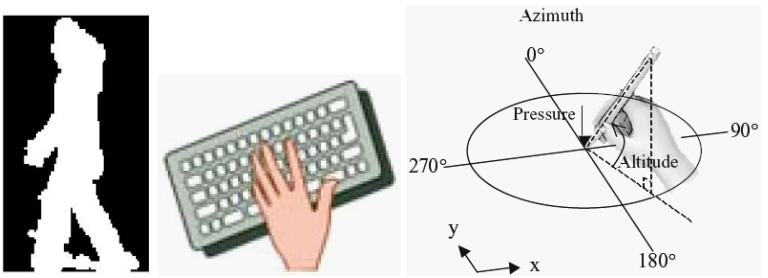
An original biometric information has been recently proposed (Orozco, Asfaw, Shirmohammadi, Adler, El Saddik, 2006) using haptics devices. Results provide an EER near 10%, and open a new area of research. Individual recognition by considering its gait has also been studied (Man & Bhanu, 2006). The performance evaluation of this kind of approach puts into obviousness an average value of EER between 19% to 37%.

An individual can be recognized thanks to its voice. Voice verification has been treated by researchers for 50 years (Petrovska-Delacretaz, El Hannani, & Chollet, 2007). Many problems have to be solved, such as acquisition artifacts (ambient noise for example) or the variability of an individual's voice due to its stress or mood. The best results obtained by combining different methods give an EER near 5%, but the robustness of the authentication/identification is difficult to reach.

Many research works focus on keystroke analysis as authentication solution for controlling the access to a computer or a mobile phone. The major reason is that the knowledge of a password is often shared with many individuals. A recent work studied the feasibility of using keystroke authentication for mobile phones (Clarke & Furnell, 2007). The method, based on neural network classifiers, achieves promising results with an EER of 15.2%. A similar approach by using a computer keyboard provides better results (EER near 5%). The main advantage of this biometric modality is that no additional sensor is required.

Online signature verification is an interesting method because it is very common for a user. The signature shape is not really used in this context, but only the way it has been

Figure 3. Some examples of behavioral biometric information (gait, keystroke dynamics, online signature) (Muramatsu & Matsumoto 2007)



realized. Notions of pen pressure and pen orientations are taken into account (Muramatsu & Matsumoto, 2007). This kind of biometric information is more adapted for authentication applications and provides an EER near 5%.

Morphological Analysis

This kind of approach is the most popular in biometrics (see Figure 4).

Fingerprint verification is a very well known technique and has been used to identify criminals for more than 100 years. This approach is often used and a dense literature is available (Tulyakov, Farooq, Mansukhani, & Govindaraju, 2007). Fingerprint verification provides, actually, a performance characterized by an EER value near 2%.

Iris recognition is a powerful technique and provides a low value of FAR near 1% (Cui et al., 2004). Retinal image is also reliable biometric information and provides a close to zero FAR value (Mariño, Penedo, Penas, Carreira, & Gonzalez, 2006). These two biometric data need a particular acquisition system and are not always accepted by users.

Hand geometry is also an interesting characteristic. Recent methods achieve a good performance such as a FAR value near 0% while keeping a FRR one near 1.5% (Kumar, Wong, Shen, & Jain, 2006). Finger vein authentication is possible and provides extremely good results, that is to say, 0.01% in FRR and less than 0.00002% in FAR (Hashimoto, 2006). Palmprint or hand veins can be used as biometric information. As for example, a recent method (Connie, Teoh, Goh, & Ngo, 2004) achieves an authentication process with an EER value close to 0%.

Face recognition is the most common approach for a human. Many research works focused on this biometric modality in the past. An important advantage of this approach is that it can be used either for the authentication or identification of an individual without his assent for video monitoring applications. Face recognition permits good performances for identification with an EER between 5% and 10% (Shen & Bai, 2006). Nevertheless, in order to be used in a real context, many problems have to be solved, such as the acquisition variability and some modifications of an individual's face (haircut, ageing…).

Figure 4. Some examples of morphological biometrics (Cui, Wang, Huang, Tan, & Sun, 2004)

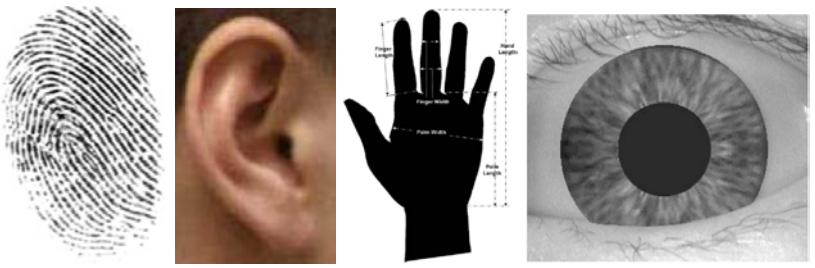


Table 1 summarizes the properties of the current biometric modalities in the state of the art.

MAIN FOCUS OF THE CHAPTER

This section first discussed the general authentication process. Second, we present the usual architecture in biometric authentication applications. The discussion presents the main points concerning the biometric authentication in the state of the art.

Authentication Process

One individual has three possibilities to prove its identity:

- What he/she owns (card, document) ;
- What he/she knows (a name, a password) ;
- What he/she is (fingerprint, hand, face...).

Generally, the identity verification of an individual is achieved with one password for different applications. For e-commerce applications, it would be very interesting to use, in place of the PIN code, which represents only four figures to guess, a biometric authentication, for many reasons. First, it would be more convenient for a user (difficulties to memorize the different codes of multiple cards, quick verification). Second, it can limit frauds because if it is quite easy to guess the PIN code composed of 4 digits, it is more difficult to copy the biometric reference of an individual.

Biometric Authentication

Despite the growing numbers of biometric applications, systems may be modeled according to the ISO Biometric conceptual diagram (ISO/IEC 19794-1:2006(E)). This diagram describes five conceptual components, which are data capture, signal processing, data storage, matching, and decision used in the two basic biometric operations, enrollment and authentication.

The enrollment is the first step introducing people in a biometric system. Its aim is to associate one person, through its biometric information, to an identity (see Figure 6).

he enrollment starts with the capture of N biometric samples needed by the signal processing "Extraction for enrollment" component. This module extracts the distinguished features from the biometric samples and sends them to the quality control subsystem. The aim of the quality control subsystem is to evaluate data relevance according to the system data quality requirements. In the case of a negative result from the quality control subsystem, the biometric

Table 1. Properties of biometric modalities (the number of stars in the performance colonna is related to the obtained value of EER in the state of the art)

Biometric information	**Universality**	**Uniqueness**	**Permanency**	**Collectability**	**Acceptability**	**Performance**
DNA	Yes	Yes	Yes	Not really	Not really	*****
Blood	Yes	No	Yes	Not really	No	*
Brain signal	Yes	Yes	Yes	Not really	No	****
Heart signal	Yes	Yes	Yes	Not really	Yes	*****
Online signature	Yes	Yes	Not really	Yes	Yes	****
Gait	Yes	No	Not really	Yes	Yes	***
Keystroke	Yes	Yes	Not really	Yes	Yes	****
Haptic behavior	Yes	Yes	Not Really	Not Really	Yes	****
Voice	Yes	Yes	Not really	Yes	Yes	****
Iris	Yes	Yes	Yes	Yes	Somewhat	*****
Retina	Yes	Yes	Yes	Yes	Somewhat	*****
Face	Yes	No	Not really	Yes	Yes	****
Hand geometry	Yes	No	Yes	Yes	Yes	****
Hand veins	Yes	Yes	Yes	Yes	Yes	*****
Ear	Yes	Yes	Yes	Yes	Yes	*****
Fingerprint	Yes	Yes	Yes	Yes	Yes	****

Figure 5. Top level view of ISO Biometric conceptual diagram

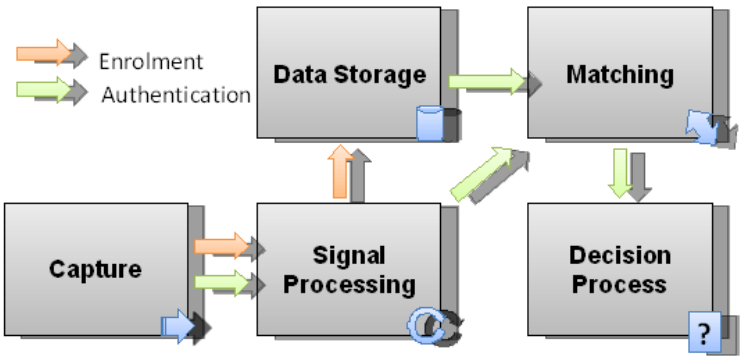
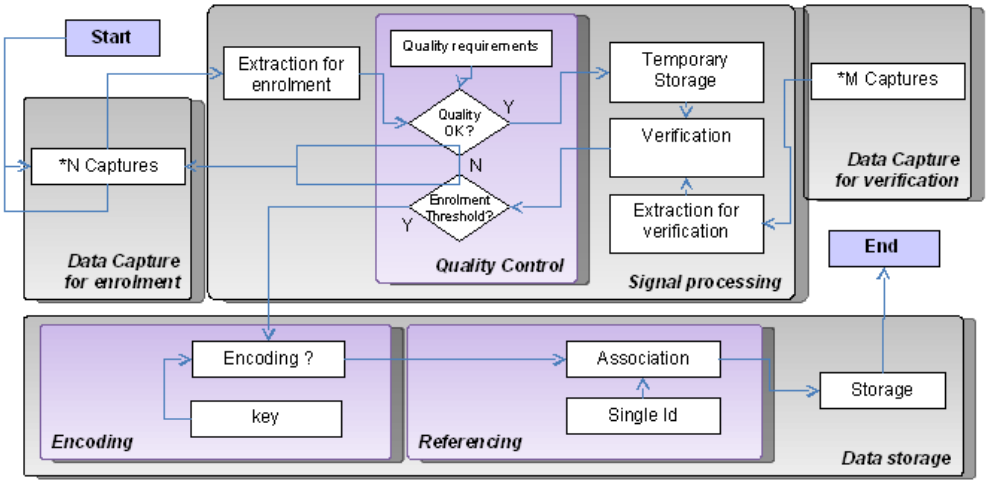


Figure 6. High level diagram of biometric Enrolment



application may ask for a new enrollment session or stop the enrollment procedure. Otherwise, the system sends the biometric data to the verification structure. The objective is to reproduce a verification request for validating data. When data validation operates correctly, the biometrics data, named template, is sent to the data storage component. This component implements some cryptographic encoding functions before transmitting the data to the referencing subsystem. The goal of this system is to associate biometric data, representing the physical characteristics, with one unique identifier, representing the virtual identity of the person. The result is transferred on a storage system (database, token…).

The authentication process (see Figure 7) implies one fundamental notion named verification and one derived notion named identification. Identification is solving the membership of a person to a system and can be perceived as multiple verification procedures.

The aim of the matching component is to evaluate the accuracy of those biometric samples with the stored enrollment templates. This evaluation is materialized by a similarity rate. The decision process focuses on the analysis of the rate given by the matching component. In the verification case, only one data extraction is done. The decision process returns to the system a Boolean result. In the identification case, the decision process will ask to the matching component to verify all the templates stored in a database. All the positive results satisfying the threshold requirements are stored. The decision process analyzes the results and returns an identifier.

Discussion

Biometric authentication is much more currently advanced by using a morphological analysis for several reasons:

Figure 7. High level diagram of biometric authentication

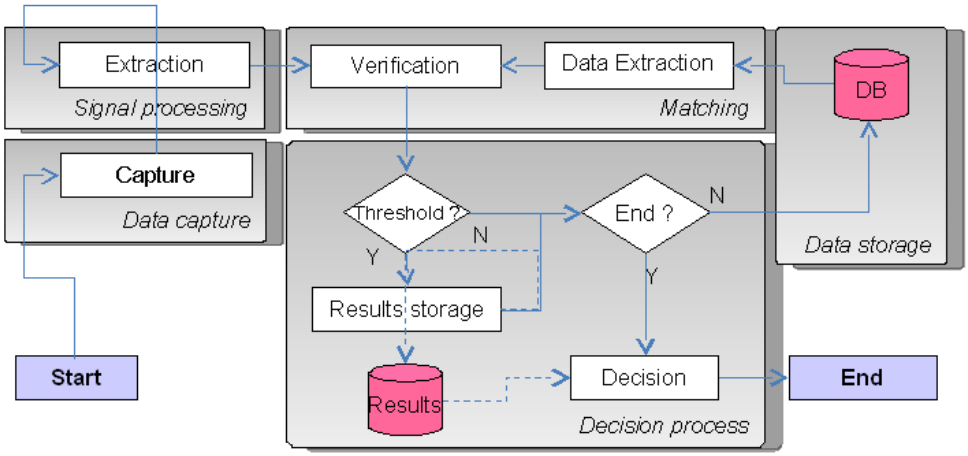
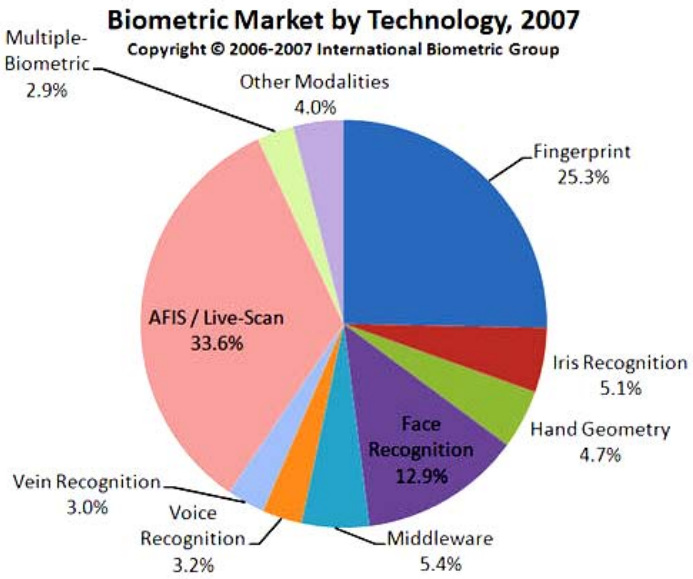


Figure 8. Biometric market and repartition of the different modalities



- the sensors are currently more reliable than the biological analysis and cheaper ;
- pattern recognition algorithms used in morphological analysis are well known ;
- recognition times for the morphological information are often shorter than for the two other types of analysis
- user perception of the intrusion of these technologies in its sphere of protection is often better for certain morphological analysis.

Figure 8 illustrates the market for each biometric modality. Morphological analysis represents actually nearly 90% of the market.

FUTURE TRENDS

In this section, we give some future trends concerning biometric authentication.

Table 3. The different biometric multimodal approaches

Methods	Example
Combination of multiple instances of a single biometric modality	Fingerprints of two different fingers
Combination of different biometric modalities	Fingerprint and physiological signals
Combination of multiple extraction algorithms	Minutia extraction and filter based algorithms for fingerprint verification
Combination of multiple acquisitions of a biometric modality	Fingerprints with two different sensors

Multimodal Biometrics

The combination of biometric systems is an interesting solution to improve their performance and robustness. Table 3 summarizes the different approaches that can be used. Note that these approaches can use some well known fusion methods, such as the Dempster-Schafer theory, that are actually widely treated in the literature (Chang et al., 2005).

Biometric evaluation

The evaluation of biometric systems grounds on two fundamental concepts applied on each of the five conceptual components described in the previous section. Those components should be generically modeled as the scheme given in figure 9.

Based on the use cases and constraints of the global system, the component will acquire input data, makes the treatment, and gives the output data as a result of its treatment. The first concept is to identify the commons and appropriated criteria. Those criteria should be represented values of the technology used by the component.

The second concept is the measure of those values according to the constraints of the system. The International norm (ISO/IEC 19795-1:2006 §6) focus on biometric performance testing and reporting. The general principles and protocols for testing the performance of a biometric system (definitions of FRR and FAR measures for example) are specified. Some problems, although, can appear such as the reliability and the representativity of the data collected for a cross-technologies comparison and cross-applications comparison.

Lots of research works deal with the evaluation of a biometric system. Generally, benchmark image databases are used for the comparison of research and commercial algorithms. One can cite, for example, the FERET database (see Figure 10) that is used for face identification/verification algorithms (Phillips, Moon, Rizvi, & Rauss, 2000).

These evaluation methodologies are often dedicated to a particular biometric system. The definition of a general platform that could compare different biometric systems is a great challenge. The BioAPI (Biometric Application Programming Interface) that is a key part of the international standards would considerably facilitate this kind of research.

Figure 9. Generic modeling of conceptual components

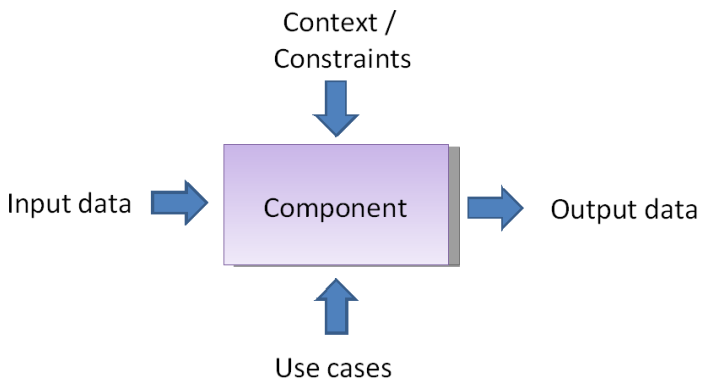


Figure 10. Some examples of images in the FERET database (Phillips et al. 2000)

CONCLUSION

This chapter focuses on biometric authentication. Even if biometrics has been used a long time ago to identify criminals, for example, automatic and efficient systems just began to appear in the research literature in the last decades. Some commercial products are already present in our daily life such as the biometricUSB keys. To be more widely used in more difficult applications, such as e-transactions, many problems have to be solved. The EER provided by algorithms using single biometric information can be very low, but not sufficiently to be accepted by everybody. Thus, the characterization of biometric systems is fundamental in order to identify the problems to solve. Multimodal biometric systems improve the performance and can be a solution even if their costs are more important. This domain represents a huge market; no doubt, many research works will provide us, in the future, a biometric toothbrush or a new payment way without any bank card.

REFERENCES

Bird, N. D., Masoud, O., Papanikolopoulos, N. P., & Isaacs, A. (2005). Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems, 6*(2), 167–177.

Chang, et al., 2005

Clarke, N., & Furnell, S. (2007). Authenticating mobile phone users using keystroke analysis. *International Journal of Information Security (IJIS), 6*(1), 1-14.

Connie, T., Teoh, A., Goh, M., & Ngo, D. (2004). PalmHashing: A novel approach for dual-factor authentication. *Pattern Analysis & Applications, 7*(3), 255-268.

Cui, J., Wang, Y., Huang, J., Tan, T., & Sun, Z. (2004). An iris image synthesis method based on PCA and super-resolution. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR)* vol. 4, pp. 471-474.

Grother, P., & Tabassi, E. (2007). Performance of biometric quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(4), 531-543.

Hashimoto, J. (2006). Finger vein authentication technology and its future. *Digest of Technical Papers. Symposium on VLSI Circuits*, pp. 5-8.

Idesia Ltd. (2004). *Bio-dynamic signature (BDS™) platform. Technical Report.*

ISO/IEC 19794-1:2006(E) – *Information technology – Biometric data interchange formats – Part 1: Framework*

ISO/IEC 19795-1:2006(E) – *Information technology – Biometric performance testing and reporting – Part 1: Principles and framework*

Jain, A. K., & Pankanti, S. (2006). A touch of money [biometric authentication systems]. *IEEE Spectrum, 43*, 22-27.

Korotkaya, Z. (2003). *Biometric person authentication: Odor. Technical report.*

Kumar, A., Wong, D. C., Shen, H. C., Jain, A. K. (2006). Personal authentication using hand images. *Pattern Recognition Letters, 27*, 1478–1486.

Man, J., & Bhanu, B. (2006). Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 28*, 316- 322

Mariño, C., Penedo, M. G., Penas, M., Carreira, M. J., &Gonzalez, F. (2006). Personal authentication using digital retinal images. *Pattern Analysis & Applications, 9*, 21-33

Muramatsu, D., & Matsumoto, T. (2007). Effectiveness of pen pressure, azimuth, and altitude features for online signature verification. In *Proceedings of the International Conference on Advances in Biometrics (ICB) Lecture Notes in Computer Science 4642* (pp. 503-512). Springer.

Orozco, M., Asfaw, Y., Shirmohammadi, S., Adler, A., & El Saddik, A. (2006). Haptic-based biometrics: A feasibility study. In *Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (HAPTICS)* (pp. 265-271).

Petrovska-Delacretaz, D., El Hannani, A., & Chollet, G. (2007). Text-independent speaker verification: State of the art and challenges. *Lecture Notes in Computer Science, Progress in Nonlinear Speech Processing, 4391*, 135-169.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence Archive, 22*(10), 1090-1104.

Phua, K., Chen, J., Huy Dat, T., & Shue, L. (2007). Heart sound as a biometric. *Pattern Recognition.* In press.

Poulos, M., Rangoussi, M., Chrissikopoulos, V., & Evangelou, A. (1999). Person identification based on parametric processing of the EEG. In *Proceedings of the 6th IEEE International Conference on Electronics, Circuits, and Systems*, vol. 1, pp. 283-286.

Shen, L., & Bai, L. (2006). A review on Gabor wavelets for face recognition. *Pattern Analysis and Applications, 9*(2), 273-292.

Tulyakov, S., Farooq, F., Mansukhani, P., & Govindaraju, V. (2007). Symmetric hash functions for secure fingerprint biometric systems. *Pattern Recognition Letters.* In press.

KEY TERMS

Authentication: Security measure designed to establish the validity of a transmission, message, or originator, or a means of verifying an individual's authorization to receive specific categories of information.

Biometric: Any specific and uniquely identifiable physical human characteristic, for example, of the retina that may be used to validate the identity of an individual.

Biometric Application Programming Interface (BioAPI): The BioAPI specification enables different biometric systems to be developed by the integration of modules from multiple independent companies.

Enrollment: The process of collecting biometric samples from a person and the subsequent preparation and storage of biometric reference templates representing that person's identity.

Equal Error Rate (EER): This error rate equates to the point at which the FAR and FRR cross (compromise between FAR and FRR).

False Acceptance Rate (FAR): Rate at which an impostor is accepted by an authentication system.

False Rejection Rate (FRR): Rate at which the authorized user is rejected from the system.

Biometric Identification Techniques

Hunny Mehrotra
Indian Institute of Technology Kanpur, India

Pratyush Mishra
Indian Institute of Technology Kanpur, India

Phalguni Gupta
Indian Institute of Technology Kanpur, India

INTRODUCTION

In today's high-speed world, millions of transactions occur every minute. For these transactions, data need to be readily available for the genuine people who want to have access, and it must be kept securely from imposters. Some methods of establishing a person's identity are broadly classified into:

1. *Something You Know:* These systems are known as knowledge-based systems. Here the person is granted access to the system using a piece of information like a password, PIN, or your mother's maiden name.
2. *Something You Have:* These systems are known as token-based systems. Here a person needs a token like a card key, smartcard, or token (like a Secure ID card).
3. *Something You Are:* These systems are known as inherited systems like biometrics. This refers to the use of behavioral and physiological characteristics to measure the identity of an individual.

The third method of authentication is preferred over token-based and knowledge-based methods, as it cannot be misplaced, forgotten, stolen, or hacked, unlike other approaches. Biometrics is considered as one of the most reliable techniques for data security and access control. Among the traits used are fingerprints, hand geometry, handwriting, and face, iris, retinal, vein, and voice recognition.

Biometrics features are the information extracted from biometric samples which can be used for comparison. In cases of face recognition, the feature set comprises detected landmark points like eye-to-nose distance, and distance between two eye points. Various feature extraction methods have been proposed, for example, methods using neural networks, Gabor filtering, and genetic algorithms. Among these different methods, a class of methods based on statistical approaches has recently received wide attention. In cases of fingerprint identification, the feature set comprises location and orientation of ridge endings and bifurcations, known as a minutiae matching approach (Hong, Wan, & Jain, 1998). Most iris recognition systems extract iris features using a bank of filters of many scales and orientation in the whole iris region. Palmprint recognition, just like fingerprint identification, is based on aggregate information presented in finger ridge impression. Like fingerprint identification, three main categories of palm matching techniques are minutiae-based matching, correlation-based matching, and ridge-based matching. The feature set for various traits may differ depending upon the extraction mechanism used.

The system that uses a single trait for authenticity verification is called unimodal biometric system. A unimodal biometric system (Ross & Jain, 2003) consists of three major modules: sensor module, feature extraction module, and matching module. However, even the best biometric traits face numerous problems like non-universality, susceptibility to biometric spoofing, and noisy input. Multimodal biometrics provides a solution to the above mentioned problems.

A multimodal biometric system uses multiple sensors for data acquisition. This allows capturing multiple samples of a single biometric trait (called multi-sample biometrics) and/or samples of multiple biometric traits (called multi-source or multimodal biometrics). This approach also enables a user who does not possess a particular biometric identifier to still enroll and authenticate using other traits, thus eliminating the enrollment problems. Such systems, known as multimodal biometric systems (Tolba & Rezq, 2000), are expected to be more reliable due to the presence of multiple pieces of evidence. A good fusion technique is required to fuse information for such biometric systems.

Depending on the application context, a biometric system may operate either in verification or identification mode (Jain, Bolle, & Pankanti, 1999a). In verification mode, the system validates a person's identity by comparing the captured biometric data with his or her own biometric template stored in the system database. In such a system, an individual who desires to be recognized claims an identity (usually via PIN), a user name, or a smartcard, and the system conducts one-to-one comparison to determine whether the claim is true or not. In the identification mode the system recognizes the individual by searching the templates of all the users in the

database for a match (Ross, Nandakumar, & Jain, 2006). The time required by the biometric system to claim identification is critical for many applications. Apart from good accuracy, response time, and retrieval, efficiency plays an important role in the identification mode.

BACKGROUND

A biometric system is essentially a pattern recognition system that recognizes a person based on a feature vector derived from a specific physiological or behavioral characteristic that the person possesses (Prabhakar, Pankanti, & Jain, 2003). A brief overview of the field of biometrics and a summary of some of its advantages, disadvantages, strengths, limitations, and related privacy concerns are presented in Jain, Ross, and Prabhakar (2004).

Multimodal biometric systems are those that utilize more than one physiological or behavioral characteristic for enrollment, verification, or identification. Ross and Jain (2003) have presented an overview of multimodal biometrics and have proposed various levels of fusion, various possible scenarios, the different modes of operation, integration strategies, and design issues. Different fusion strategies as well as its performance are given by Allano et al. (2006). The performance of multimodal biometric authentication systems using state-of-the-art commercial off-the-shelf (COTS) fingerprint and face biometric systems on a population approaching 1,000 individuals is examined by Snelick, Uludag, Mink, Indovina, and Jain (2005).

The identification system is designed in such a way that the search space and data retrieval time reduce to the minimum. Classification techniques have been introduced to reduce the search time of identification systems. There exist several classification techniques like classification of face images based on age (Kwon & Lobo, 1999), where input images can be classified into one of three age-groups: babies, adults, and senior adults. Gender classification from frontal facial images using genetic feature subset selection is considered in Sun, Bebis, Yuan, and Louis (2002). Further, nonlinear support vector machines (SVMs) are investigated for appearance-based gender classification with low-resolution thumbnail faces from a FERET (FacE REcognition Technology) face database (Moghaddam & Yang, 2002). Classification of fingerprints into five categories—whorl, right loop, left loop, arch, and tented arch—is done using a novel representation (FingerCode) based on a two-stage classifier (Jain, Prabhakar, & Hong, 1999c). Ern and Sulong (2001) give a good account of fingerprint classification techniques, and Jain, Murty, and Flynn (1999b) overview pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. Another approach for search space reduction is to index the database using some data structures like B+ trees, pyramid technique (Mhatre, Palla, Chikkerur, & Govindaraju, 2005), or hash functions. However the main concern behind the use of suitable indexing technique is to catalog the database in such a way that the identification template falls in the bin of the enrolled template.

IDENTIFICATION TECHNIQUES

Biometrics identification is expensive in terms of time as it involves a large number of comparisons in the database. As database size increases, data retrieval and search times increase significantly. Thus to overcome the problem arising from large biometric records, there must be some method to reduce the search space for a matcher to operate. The identification system is designed in such a way that the search space and data retrieval times reduce to the minimum. The biometric database can be segmented using one or all of the following three approaches in the hierarchy discussed below.

Classification

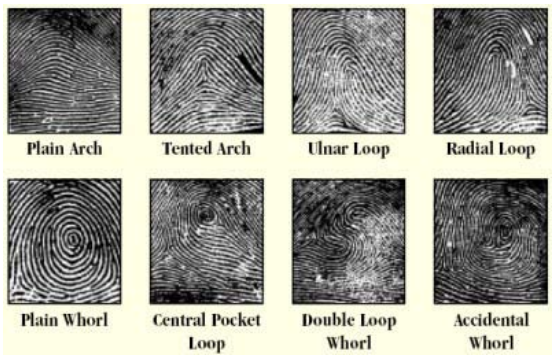
Classification refers to assigning an object physically into one of a set of predefined categories. The main idea behind the use of a classification algorithm is to divide the database into groups where each group has homogenous characteristics. The important step is to design a classifier based on texture, pattern, or some soft biometric attribute. Some commonly used classification techniques for well-known traits include face, fingerprint, and iris classification.

Face Classification

Face images are used to classify a person based on age, gender, ethnicity, and so forth. Age classification is used to classify the input images into one of three age groups: babies, young adults, and senior adults (Kwon & Lobo, 1999). The computations are based on skin wrinkle analysis.

Automated gender classification is possible with the use of statistical classification algorithms that are trained with a set of known images. With a large number of known images, it may be possible to directly train the classification algorithms, with each grayscale image representing a point in the image space. However, it is often much better to reduce the number of dimensions, in an efficient way, allowing accurate training with less observations. In applications using face images, a well-established dimension reduction technique is to use the Karhunen-Loeve Transform on an ensemble of images, to generate an eigenspace composed of eigenfaces. The idea is to project images into this eigenspace and use

Figure 1. Fingerprint pattern classes



the first few projected components for classification (Kirby & Sirovich, 1990).

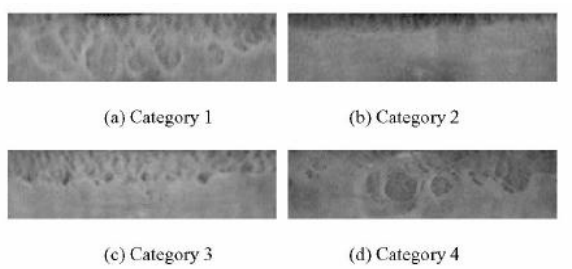
Fingerprint Classification

Fingerprint classification is one of the most reliable methods for identifying individuals. At the most basic level fingerprints, can be classified into some classes like: arch, tented arch, ulnar loop, radial loop, and whorl and twin loop, as shown in Figure 1. Jain et al. (1999c) have used a two-stage classifier for classification of fingerprint images into known pattern classes.

Iris Classification

For iris classification the images of the iris are divided into four or more classes based on the texture pattern of the image as shown in Figure 2. The texture pattern is classified by computing fractal dimension (Yu, Wang, & Zhang 2005).

Figure 2. Basic-level classification of iris images (taken from Yu et al., 2005)



Clustering

Data clustering is a common technique for statistical data analysis, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis, and bioinformatics. Clustering is the segmentation of similar objects into several groups, or more concisely, the partitioning of a dataset into subsets (clusters), so that the data in each subset (ideally) share some common trait, often proximity according to some defined distance measure (Jain et al., 1999b). Basic data clustering algorithms can be hierarchical or partitional. Hierarchical algorithms find successive clusters using previously established clusters, whereas partitional algorithms determine all clusters at once (Jain et al., 1999b).

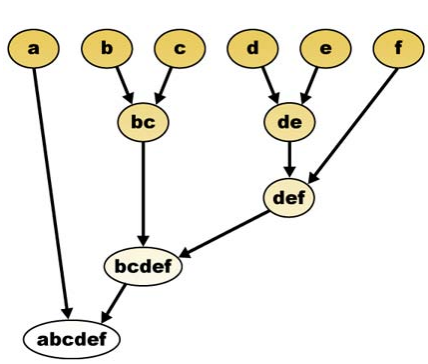
Types of Clustering Approaches

There are many clustering approaches available, and choice of a technique depends upon the type of dataset to be clustered and output desired after clustering. In general, clustering methods may be divided into two major categories based on the clusters they produce.

Hierarchical Methods

Hierarchical algorithms can be agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with each element as a separate cluster and merge them in successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters. Given a distance measure, elements can be

Figure 3. Hierarchical clustering approach



combined. Hierarchical clustering builds (agglomerative) or breaks up (divisive) a hierarchy of clusters. The traditional representation of this hierarchy is a tree data structure (called a dendrogram as shown in Figure 3), with individual elements at one end and a single cluster with every element at the other. Agglomerative algorithms begin at the top of the tree, whereas divisive algorithms begin at the bottom. Disadvantages of hierarchical clustering are related to:

- vagueness of termination criteria; and
- the fact that most hierarchical algorithms do not revisit clusters, once constructed (intermediate), to consider improvement.

Partitional Approaches

Given a database of objects, a partitional clustering algorithm constructs partitions of the data, where each cluster optimizes a clustering criterion, such as the minimization of the sum of squared distance from the mean within each cluster. One of the issues with such algorithms is their high complexity, as some of them exhaustively enumerate all possible groupings and try to find the global optimum. Even for a small number of objects, the number of partitions is huge. The diagrammatic representation of various forms of the partitional clustering technique is given in Figure 4.

Indexing

Indexing the database implies a logical partitioning of the data space. In this approach of search space, reduction tree structure is used for organizing the data, such that each of the leaf nodes stores one or more biometric templates. Thus given a test template, only the templates having similar index values would be considered as the search space, or in the case of range search, only the fraction of the database lying within the range of the indexes of the test template would be the new search space. With biometric data being inherently multidimensional, it is indispensable that the indexing technique support multidimensional data. The tree structure for biometric data indexing is shown in Figure 5.

Figure 4. Examples of partitional clustering approaches

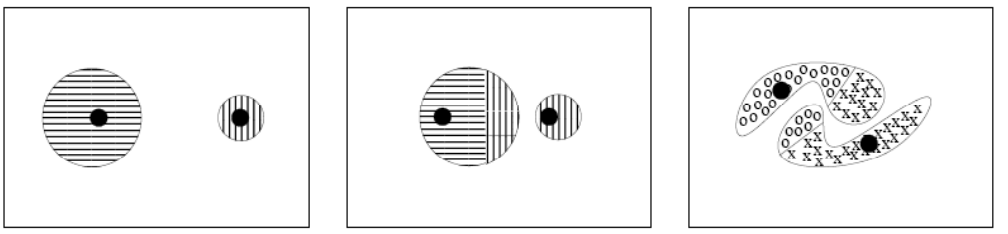
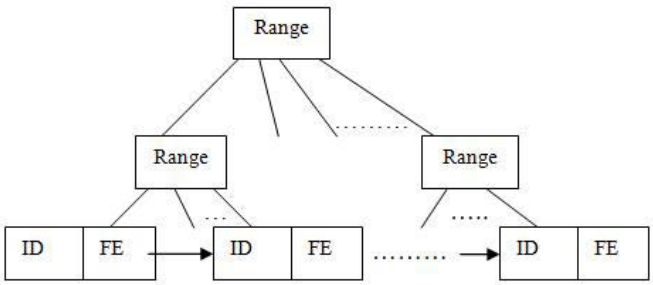


Figure 5. B+ tree for indexing



Pyramid Technique

This currently seems to be the single most feasible indexing technique for indexing biometric databases. The concept is based on spatial-hashing the high-dimensional data into a single value. This single value can then be indexed very effectively using the popular B+ trees. The pyramid-technique was especially designed for working in higher-dimensional spaces and has been successfully tested on an 100-dimensional data set. The technique results in a tree structure that is not only insertion order invariant, but also height balanced and dynamic to the frequent insertions and deletions.

Hash Function

A hash table, or a hash map, is a data structure that associates keys with values. The primary operation it supports efficiently is a lookup: given a key (e.g., a person's name), find the corresponding value (e.g., that person's telephone number). It works by transforming the key using a hash function into a hash, a number that the hash table uses to locate the desired value. Thus hashing can be used on biometric data by calculating keys using feature values and storing the identifier at the key generated.

MULTIMODAL BIOMETRIC SYSTEM AT IIT KANPUR

The multimodal biometric identification system at the Indian Institute of Technology Kanpur is developed using five traits: face, fingerprint, iris, ear, and signature. The system classifies the biometric database at a basic level using iris classification techniques based on a box counting approach. At a refined level the database is further subdivided using a clustering technique. The data within each cluster is indexed using B+ tree for storage. The steps are followed hierarchically to get top matches:

- In the first phase the database is classified using iris. The acquired iris image is preprocessed and transformed into a rectangular block known as the 'strip'. The strip is used for classification of iris texture using a box counting approach. The database is divided into four classes using the supervised classification technique.
- In the second phase the data within each class is clustered based on a feature vector generated from face recognition. For clustering, the well-known K-means algorithm is used on the generated feature set from face recognition within a particular class.
- Finally the clustered database is indexed through some indexing scheme so that we can achieve the above mentioned goal of partitioning the large biometric database. The n-dimensional feature set obtained is used for indexing the database by forming a B+ tree. Here for every feature element, a B+ tree is formed. Thus for n-dimensional database, n such B+ trees are formed for indexing.

At the time of identification, the query data are passed to a classification algorithm, and a relevant cluster is obtained. The B+ tree within the retrieved cluster is traversed to get a set of data for matching. Finally the individual scores obtained from available traits after matching are fused using weighted sum rule. These scores are used to form a decision regarding a person's identity.

FUTURE TRENDS

Biometrics is considered one of the most reliable means of access control. Unimodal biometric systems come to standstill due to unavailability of data or poor quality input. To overcome the problem, multimodal biometric systems are used. The present state of the art provides good fusion strategies at a matching score level. However, more efforts are required towards feature-level and sensor-level fusion techniques.

Further, during identification the database is segmented using classification, clustering, and indexing techniques. At a basic level the database is divided into classes using some supervised learning algorithms. After classification, data points within respective classes are clustered using known clustering algorithms. Finally the data within each cluster are indexed to form bins. Researchers should investigate ways to develop algorithms to reduce the time and space complexity in dividing the large biometric database, with increased accuracy as well as reliability. The efforts should be directed towards minimizing the time required to claim the identity of an individual.

CONCLUSION

Biometrics systems are widely used to overcome the traditional methods of authentication. But the unimodal biometric system fails when there is a lack of biometric data for a particular trait. Thus the individual scores are combined at a classifier level and trait level to develop a multimodal biometric system. The system classifies the biometric database at a basic level using classification techniques mentioned in this article. At a refined level the database is further subdivided using a clustering technique. The data within each cluster are indexed using B+ trees, hash tables, and a pyramid technique for storage. At the time of identification, the query data are classified and the relevant cluster is obtained. The

B+ tree within a respective cluster is traversed to get a set of data for matching. Thus the hierarchical partitioning of databases is very helpful to claiming identity with reduced time and increased accuracy.

REFERENCES

Allano, L., Morris, A.C., Sellahewa, H., Garcia-Salicetti, S., Koreman, J., Jassim, S., Ly-Van, B., Wu, D., & Dorizzi, B. (2006). Non intrusive multi-biometrics on a mobile device: A comparison of fusion techniques. *Proceedings of SPIE* (p. 6202).

Ern, L.C., & Sulong, G. (2001). Fingerprint classification approaches: An overview. *Proceedings of the 6th International Symposium on Signal Processing and its Applications* (pp. 347-350).

Hong, L., Wan, Y., & Jain, A. K. (1998). Fingerprint image enhancement: Algorithm and performance evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 777-789.

Jain, A.K., Bolle, B., & Pankanti, S. (1999a). *Biometrics: Personal identification in networked society.* Norwell, MA: Kluwer Academic.

Jain, A.K., Murty, M.N., & Flynn, P.J. (1999b). Data clustering: A review. *ACM Computing Surveys, 31*(3), 264-323.

Jain, A.K., Prabhakar, S., & Hong, L. (1999c). A multichannel approach to fingerprint classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(4), 348-359.

Jain, A.K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(1).

Kirby, M., & Sirovich, L. (1990). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(1), 103-108.

Kwon, Y.H., & Lobo, N.V. (1999). Age classification from facial images. *Computer Vision and Image Understanding, 74*(1), 1-21.

Mhatre, A., Palla, S., Chikkerur, S., & Govindaraju, V. (2005). Indexing biometric databases using pyramid technique. *Audio and Video-Based Biometric Person Authentication (AVBPA),* 841-849.

Moghaddam, B., & Yang, M.H. (2002). Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(5), 707-711.

Prabhakar, S., Pankanti, S., & Jain, A.K. (2003). Biometric recognition: Security and privacy concerns. *IEEE Security & Privacy,* 33-42.

Ross, A., & Jain, A.K. (2003). Information fusion in biometrics. *Pattern Recognition Letters, 24*(13), 2115-2125.

Ross, A.A., Nandakumar, K., & Jain, A.K. (2006). *Handbook of multibiometrics* (International Series on Biometrics). NJ: Springer-Verlag.

Snelick, R., Uludag, U., Mink, A., Indovina, M., & Jain A.K. (2005). Large-scale evaluation of multimodal biometric authentication using state-of-the-art systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(3), 450-455.

Sun, Z., Bebis, G., Yuan, X., & Louis, S.J. (2002). Genetic feature subset selection for gender classification: A comparison study. *Proceedings of 6th IEEE Workshop on Applications of Computer Vision* (pp. 165-170).

Tolba, A.S., & Rezq, A.A. (2000). Combined classifier for invariant face recognition. *Pattern Analysis and Applications, 3*(4), 289-302.

Yu, L., Wang, K., & Zhang, D. (2005). Coarse iris classification based on box-counting method. *IEEE International Conference on Image Processing, 3,* 301-304.

KEY TERMS

Biometrics: The use of physiological or behavioral characteristics to verify the identity of an individual comes under the realm of biometrics. Biometrics is an automated recognition of an individual based on his or her distinctive anatomical characteristics.

Classification: Supervised grouping of entire dataset at a basic level using some extracted biometric features.

Clustering: Can be considered the most important unsupervised learning problem of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects that are similar between them and are dissimilar to the objects belonging to other clusters.

Fusion: Combines biometric characteristics derived from one or more modalities or technologies (algorithms, sensors), multiple characteristics derived from samples, or multiple or repeated biometric instances to increase performance. Biometrics fusion centers on the capture and comparison of multiple biometric measurements like fingerprint and face.

Hashing: The transformation of a record into a usually shorter fixed-length value or key that represents the original record. Hashing is used to index and retrieve items in a database because it is faster to find the item using the shorter hashed key than to find it using the original value.

Indexing: Logical partitioning of large database using some known data structures like B-trees, hash functions, and so forth.

Multi-Biometrics: A biometric system that uses more than one biometric identifier (like a combination of face, fingerprint, iris, ear etc.) in making a decision about personal identification. Multimodal biometrics systems are expected to be more reliable due to the presence of multiple traits.

Template: The mathematical representation of biometric data. Any graphical representation is reduced to a numerical representation. The template is then used by the biometric system as an efficient method to make comparisons with other templates stored in the system.

Biometric Paradigm Using Visual Evoked Potential

Cota Navin Gupta
University of Essex, UK

Ramaswamy Palaniappan
University of Essex, UK

INTRODUCTION

Recognizing humans based upon one's intrinsic physical or behavioral traits has been gaining acceptance and is termed as biometrics. It involves either confirmation or denial of the identity that the user is claiming. It is especially important in ensuring security for access to highly restricted areas (for example: accessing classified documents, control gates and defence related applications). This chapter will discuss the use of brain signals at an application level exploiting the evoked potential approach for biometrics.

BAKGROUND

The most primitive and widely used authentication method to establish a person's identity is the textual password and usage of personal identification number (PIN) which are motivated by the facts of popularity due to low cost and user familiarity.

However these schemes have obvious shortcomings in the form of dictionary attack, shoulder surfing and people picking up obvious known words which can be easily cracked. Dictionary attacks can be prevented by using human-in-loop verifications (Pinkas & Sander, 2002) and encrypted key exchange methods (Bellovin & Merritt, 1992), but operating system vulnerabilities and access control failures may lead to disclosure of password databases. The use of PIN actually denotes the automatic identification of the PIN, not necessarily identification of the person who has provided it. The same applies with card and tokens, which could be presented by anyone who successfully steals the card or token. The system and information is definitely vulnerable during the period before a user's card or token is revoked. Even the recently proposed graphical password which is motivated by the fact that people have a remarkable memory for pictures seem to share similar problems along with the shortcomings of guessing attacks (Thorpe & Van Orschot, 2004) and reduced effective password space. The ominous presence of mobile phone cameras, digital cameras, and wireless video cameras brings in a new threat in the form of "recorded shoulder surfing" for high security applications.

Hence biometric technology based on measurable physiological and/or behavioral characteristics (e.g., fingerprints, Roddy & Stosz, 1996, the iris, Daugman, 2004, and voice recognition, Monrose, Reiter, Li & Wetzel, 2001) is often considered to surpass conventional automatic identity measures like passwords and PIN by offering positive human identification.

Fingerprint biometric systems have found its way in many public person identity databases (Maltoni, Maio, Jain & Prabhakar, 2003), but they do not seem suitable for high security environments. Recent articles and studies (BBC, 2007a; Matsumoto, Matsumoto, Yamada & Hoshino, 2002) show that common household articles (e.g., gelatine) can be used to make artificial fingers and prints to bypass the security systems. Also development of scars and cuts can result in erroneous fingerprint matching results thus increasing false rejects. Voice recognition as a biometric seems to suffer from several limitations. Different people can have similar voices and it may also change over time because of health, emotional state and age. Face recognition has been used as a biometric system but issues like the family resemblance, occurrence of identical twins (one in every 10,000) seem to question the reliability. A recent article shows that face recognition systems can be bypassed by using still and video images of a person (BBC, 2007b). Also it is inherently unreliable where high security is needed because there is not nearly enough randomness in the visual appearance of people's faces and also small variations in pose angle, illumination geometry, and facial expression have disastrous effects on the authentication algorithm accuracy (BBC, 2007b).

Another issue facing many of the biometric systems is the factor that biometric data (e.g., fingerprints or iris scans) have information which is valid and unchangeable for lifetime of the user and is irreplaceable if stolen. However it is a known fact that no biometric is expected to effectively meet the requirements for all applications. The choice of a specific biometric completely depends on the requirements of the application domain.

The above discussion on the existing biometric technologies definitely highlights the shortcomings for high security

environments and reiterates the need for an authentication system which has the following characteristics (Thorpe, Van Oorschot & Somayaji, 2005):

- ***a.*** *Changeability*: The ability to replace authentication information.
- ***b.*** *Privacy (theft protection)*: A biometric which is fraud resistant and does not use a template for lifetime.
- ***c.*** *Shoulder surfing*: System should be immune to all forms of shoulder surfing.
- ***d.*** *Universality*: Every person should have the considered characteristics.
- ***e.*** *Permanence (stability)*: Characteristic should be invariant and stable over a period of time.

A biometric system using brain's electrical response patterns with the evoked potential approach seems to have the potential to satisfy all of these requirements. Applications for this biometric system include high security systems (access to classified documents, defence applications) where fingerprints and other identity measures like passwords could be easily forged. It could also be used as a modality within a multimodal biometric environment. The advantage of using such brain electrical activity as biometric is its fraud resistance, that is someone else cannot duplicate the recorded brain response, and is hence unlikely to be forged or stolen. This modality has the additional advantage of confidentiality ("all forms of shoulder surfing' is impossible"), as brain activity is not easily seen. An added impetus for this sort of approach is the recent report in Newscientist (2007) about an initial study on the possibility of developing an electronic security system that will identify people by monitoring the brain activity.

BIOMETRIC SYSTEM USING BRAIN SIGNALS

In general, data for brain biometric system are collected using an electrode cap worn by the person (also known as subject). The electrodes are connected to the holder as shown and the brain signals are recorded in response to the activity on the computer screen. Electrode gel is used at the point of contact while fixing the electrodes to the electrode cap for improving conductance of brain potentials. There are also interfacing cables which interface the computer and the electroencephalogram (EEG) equipment to record the responses to the ongoing paradigm. The electrode configurations commonly used are the 32, 64, and 128. A number of trials of the same paradigm are usually performed during the course of the experiment and averaging taken to reduce artifacts (i.e., noise).

Given the risks with invasive implanted devices in brain and the associated ethical concerns, non invasive approaches (in particular those using EEG) seems to be more popular. EEG which is the recording of the brain's electrical activity is the de facto standard in diagnosis of brain related diseases, however recently there has been a spurt of activity in the studies on brain biometrics (Marcel and Millan, 2007; Palaniappan & Mandic, 2007; Palaniappan & Ravi, 2006; Palaniappan & Raveendran, 2002). Some early work on EEG based biometrics include the use of autoregressive (AR) models of various orders computed from EEG signals recorded from subjects with eyes open and eyes closed (Paranjape, Mahovsky, Benedicenti & Koles, 2001). A linear discriminant analysis was employed to classify the 40 subjects which gave an accuracy of 80 percent. Learning Vector Quantizer network (LVQ) was used to classify AR parameters from four subjects describing the alpha rhythm EEG feature, where the classification performance of 72-84 percent was obtained (Poulos, Rangoussi, Chrissikopoulos & Evangelou, 1999a). In a similar related study using the same data set but a different classification technique based on computational geometry gave a much improved average classification of 95% (Poulos, Rangoussi, Chrissikopoulos & Evangelou, 1999b).

More recently a statistical framework based on Gaussian mixture models and maximum a posteriori model adaptation (Marcel & Millan, 2007) was used for person authentication. The study also highlighted that certain mental tasks are more appropriate for authentication. However, many of these studies were conducted for a relatively small number of subjects.

BIOMETRIC SYSTEM USING THE EVOKED POTENTIAL APPROACH

Evoked potential is a type of EEG that is evoked in response to a stimulus, which could be visual, auditory or somatosensory. Visual evoked potential (VEP) is the evoked response to visual stimulus. In a recent study (Palaniappan & Mandic, 2007), multiple signal classification (MUSIC) algorithm was used to extract features in the gamma band of VEP based experiment study and gave enhanced person recognition of over 96% with 102 subjects. Other systems have exploited the P300 component of VEP as a medium of communication (Donchin, Spencer & Wijesinghe, 2000; Farwell & Donchin, 1988), which could be adapted from biometrics. P300 is an endogenous component of the VEP, which is most frequently elicited within the framework of an "oddball paradigm". P300 based systems are promising and motivating as they require no or very less training. It is known for its simplicity, ease of use and low error rates (Kaper, Meinicke, Grossekathoefer, Lingner & Ritter, 2004; Serby, Tov & Inbar, 2005,).

In the oddball experiment the subject is asked to distinguish between two stimuli, one common and one rare, by

Figure 1. Donchin's stimulus matrix viewed by the subject (Donchin et al, 2000)

A	B	C	D	E	F
G	H	I	J	K	L
M	N	O	P	Q	R
S	T	U	V	W	X
Y	Z	1	2	3	4
5	6	7	8	9	space

performing a mental count of one of the stimuli. In response to mentally counting the appearance of the rare stimulus, a typical potential is evoked in the brain. At any given moment the user selects (say a color or symbol) that the user wishes to communicate, and maintains a mental count of the number of times it is intensified. In response to this counting, a potential is elicited in the brain each time the color or symbol is flashed and the response is known as a P300 wave (Sutton et al, 1965). The P300 based Donchin paradigm (Farwell and Donchin, 1988) enabled communication of the alphabets and few symbols using the P300 component of the VEP. The 26 characters and symbols were displayed on the computer as shown below in Figure 1. The subject focuses successively on the character that the subject wishes to communicate. The computer detects the character focused by the subject because of the alternate repeated flashing of the columns and rows. When a column or row containing the chosen character is flashed, a P300 is evoked and would be detected by the computer.

A Preliminary Study on "Inblock" and "Outofblock" P300 Oddball Experiment

In this preliminary study, we wished to investigate the effects of presenting the oddball stimulus in "Inblock" and "Outofblock" fashion (as illustrated in Figure 2). The Donchin's paradigm (Donchin et al., 2000) is based on the Outofblock' fashion but here we show that the Inblock fashion would be more suitable to evoke P300 using the oddball paradigm for biometric applications.

We recorded EEG data from a male subject aged 27 with no known neurological disorders. A very simple form of visual stimulus presentation was used here where the subject was asked to concentrate on the letter "A". The letter Awas flashed 30% of the time while the square block was flashed for the rest 70% of the time. It is assumed that the infrequent stimuli (i.e., when the subject concentrates on the letter A) will evoke a P300 component.

The flashes were intensified for 100 ms, with an interstimulus interval (ISI) of 300 ms. During the ISI, there would be no intensifications. The ISI is defined as the end of the intensification to the start of the next intensification. The period of 300 ms was chosen after some preliminary simulations.

The sampling frequency was 256 Hz and EEG data was recorded from 64 channels using Bio-semi system. EEG data was recorded for every intensification (i.e., flash of the "block" or letter A). First, averages of left and right mastoid channels were used to re-reference the data. To extract P300 component, each EEG signal was low pass filtered to 8 Hz using a fourth order Elliptic Infinite Impulse Response filter (with forward and reverse filtering to avoid phase distortion) and then normalised to zero mean and standard deviation of one. The cut-off frequency of 8 Hz was chosen after some preliminary simulations.

Next, P300 amplitude was computed as the most positive peak in the range of 300-600 ms (or 77-154 sampling points) after stimulus onset. The above trials for the Inblock

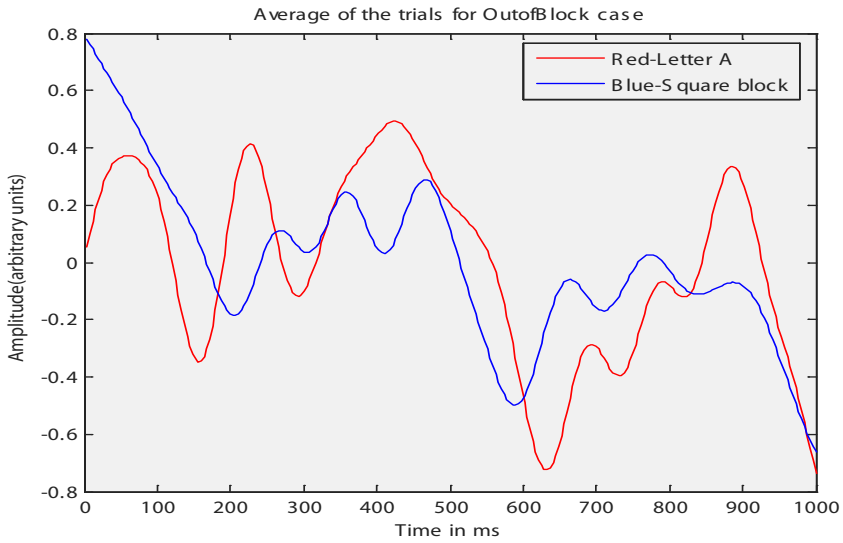
Figure 2. Visual stimulus for Donchin's paradigm: (a) Inblock and (b) OutofBlock

(a) Inblock stimulus

(b) OutofBlock stimulus

Figure 3. Averaged EEG signal for Inblock case

Figure 4. Averaged EEG signal for Outofblock case



and OutofBlock cases were averaged to reduce noise and are as shown below in Figures 3 and 4 where red indicates the EEG data when the letter A was flashed which has the P300 component (because the subject concentrated on the letter) and blue when the square block was flashed. The analysis was done for the channel Cz which is known to have maximal P300 component.

It could be seen that the signal component for the target letter A case has higher amplitude in the region of 300-600 ms rather than the square block. This is in line with the results expected from oddball paradigm. But the more important observation is the fact that the P300 peak is more easily distinguishable using the Inblock rather than Outofblock fashion. Therefore, in future experiments, we could use Inblock type of fashion for evoking P300 components using oddball paradigm.

Color "Blocks" Visual Stimuli

To make the concept of using brain biometrics universal across all boundaries (i.e., to avoid language differences), we are currently investigating the possibility of using colors

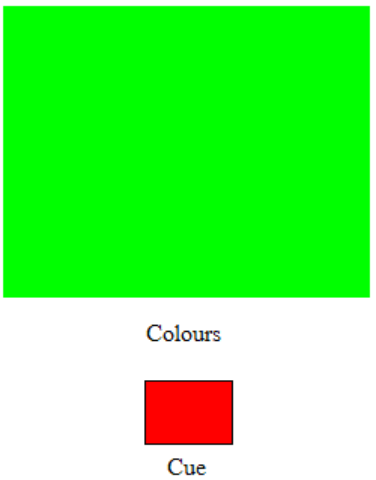
(or universal symbols) instead of English alphabets as in Donchin's paradigm.

This proposed system, which is still in early stages of research, will authenticate users using their brain signals in response to a sequence of on-screen color stimuli as well as audio signals. So at any time, the system's focus for the biometric application is to differentiate the colors on screen rather than the individuals. The EEG data will be recorded when the subject perceives different color flashes on white background within a single square block (i.e., Inblock case). The basic colors: black, red, green and blue were decided because of the contrasting difference which will easily evoke the P300 potential. The cue on which the subject has to concentrate will also be presented below this block as shown in the Figure 5.

The subject will be instructed to focus on colors that randomly flash on screen. The objective would be to form a color coded pass code generated by thought alone, which could be used to authenticate the identity of a person. For example, a passcode could be RED, BLACK, BLUE, GREEN (sequence is important as it determines the passcode). Each color would flash in random order until all colors have flashed; this is known as randomized block intensification. Each randomised block intensification of four colors will be considered as a trial. Currently, we are investigating various other novel visual stimulus paradigms (for example, using picture instead of colors) to decide the suitable stimulus for biometric applications.

This may be extended to various scenarios like words, graphical images or music to form a sequence in the form of a password. The advantage of using such brain electrical activity as biometric is its fraud resistance, that is the brain response cannot be duplicated by someone else, and is hence unlikely to be forged or stolen. The only disadvantage of the system lies in the cumbersome data collection procedure but improvements in data collection procedures (such as dry electrodes, instead of wet) will reduce the unwieldiness. However the fraud resistance significantly outweighs this difficulty especially for high security applications.

Figure 5. Color visual stimulus under investigation for brain biometrics



FUTURE TRENDS

The field of crossmodal perception refers to different sensory modalities (say audio and visual) and is being increasingly studied to gain better understanding of the long-term properties of the brain. Matched sensory inputs (such as the sight and sound of a cat) enhance our perception. Studies have shown that performance improvements can be achieved on low-level visual perceptual tasks with practice but is difficult as well as slow and requires many days of training (Adini, Sagi & Tsodyks, 2002; Poggio, Fahle & Edelman, 1992). In the study by Seitz et al (2006), a multisensory audiovisual training procedure facilitated faster visual learning than unisensory visual training. In this biometric application, the remembrance of password may be considered as a recall from episodic memory, which is a subset of declarative memory because it is related to storage of facts and can be discussed or declared. Declarative memory is subject to forgetting, but if accessed frequently they can last indefinitely (Tulving & Schacter, 1990). Since it is known that matched sensory inputs enhance our perception, it would be worthwhile to investigate and compare evoking the declarative memory (i.e. password) by one of the three protocols: visual stimulus, audio stimulus and visual combined with audio stimuli.

CONCLUSION

Although much work has been done in the past decade using brain signals for clinical analysis, the application of brain signals for biometric purpose is relatively new. It is also known that in a P300-based biometric system, the communication speed of characters is dependent on the number of trials. We are working on developing new framework at an algorithm level which we foresee will use less number of trials (Gupta & Palaniappan, 2007). It is anticipated that much of the work in future will concentrate on developing a really robust and user-friendly system where the number of trials used will be minimum (i.e., to obtain a higher bit rate). It is envisaged that the future work and technological advancements will move this concept from research into a practical working system. No doubt, one of the main constraints is the long set-up time. However, with steady research progress in signal processing, we should be able to use only a few channels that will still give accurate authentication. Then, a simple headband with

electrodes attached could be used instead and set-up time would only be on the order of seconds.

Concluding, a biometric authentication system using brain's electrical response patterns using the evoked potential approach has the potential to satisfy all of the requirements for a high security scenario and should be seriously considered as one of the emerging biometric paradigms.

REFERENCES

Adini, Y., Sagi, D., & Tsodyks, M. (2002). Context-enabled learning in the human visual system. *Nature, 415*, 790–793.

BBC (2007a). Retrieved June 17, 2008 http://news.bbc.co.uk/2/hi/science/nature/1991517.stm

BBC (2007b). Retrieved June 17, 2008 http://news.bbc.co.uk/1/hi/sci/tech/2016788.stm.

Bellovin, S. & Merritt, M. (1992). Encrypted key exchange: Password-based protocols secure against dictionary attacks. In *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy* (pp. 72–84O.

Daugman, J. (2004). How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(1), 21–30.

Donchin, E., Spencer, K. M., & Wijesinghe, R. (2000). The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering, 8*(2), 174–179.

Farwell, L. A. & Donchin, E. (1988). Talking off the top of your head: A mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology, 70*(6), 510–523.

Gupta, C. N. & Palaniappan, R. (2007). Enhanced detection of visual evoked potentials in brain-computer interface using genetic algorithm and cyclostationary analysis. [Special Issue] *Journal of Computational Intelligence and Neuroscience*, DOI: 10.1155/28692.

Kaper, M., Meinicke, P., Grossekathoefer, U., Lingner, T., & Ritter, H. (2004). BCI competition 2003-data set IIb: Support vector machines for the P300 speller paradigm. *IEEE Transactions on Biomedical Engineering, 51*(6), 1073–1076.

Maltoni, D., Maio, D., Jain, A.K., & Prabhakar, S. (2003). *Handbook of fingerprint recognition.* Springer-Verlag.

Marcel, S. & Millan, J. (2007). Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(4), 743-752.

Matsumoto, T., Matsumoto, H., Yamada, K., & Hoshino, S. (2002). Impact of artificial gummy fingers on fingerprint systems. In E.L. van Renesse (Ed.), *SPIE Optical Security and Counterfeit Deterrence Techniques, IV*(4677) 275–289.

Monrose, F., Reiter, M.K., Li, F.Q., & Wetzel, S. (2001). Cryptographic key generation from voice. In *Proceedings of the IEEE Conference on Security and Privacy* (pp. 202-213).

NewScientistTech (2007). Retrieved June 17, 2008, from http://www.newscientisttech.com/channel /tech/dn10963-brain-activity-provides-novel-biometric-key.html

Palaniappan, R. & Raveendran, P. (2002). Individual identification technique using visual evoked potential signals. *Electronics Letters, 138*(25), 1634-1635.

Palaniappan, R. & Ravi, K. V. R. (2006). Improving visual evoked potential feature classification for person recognition using PCA and normalization. *Pattern Recognition Letters, 27*(7), 726-733.

Palaniappan, R. & Mandic, D. P. (2007). Biometrics from brain electrical activity: A machine learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(4), 738-742.

Paranjape, R. B., Mahovsky, J., Benedicenti, L., & Koles, Z. (2001). The electroencephalogram as a biometrics. *Proceedings of Canadian Conference on Electrical and Computer Engineering, 2*, 1363-1366.

Pinkas, B. & Sander, T. (2002). Securing passwords against dictionary attacks. In *Proceedings of 9th ACM Conference on Computer and Communications Security* (pp. 161–170).

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science, 256*, 1018–1021.

Poulos, M., Rangoussi, M., Chrissikopoulos,V., & Evangelou, A. (1999a). Person identification based on parametric processing of the EEG. *Proceedings of IEEE International Conference on Electronics, Circuits, and Systems, 1*, 283-286.

Poulos, M., Rangoussi, M., Chrissikopoulos, V., & Evangelou, A. (1999b). Parametric person identification from the EEG using computational geometry. *Proceedings of IEEE International Conference on Electronics, Circuits, and Systems, 2*, 1005-1008.

Roddy, A. R. & Stosz, J. D. (1996). Fingerprint features—Statistical analysis and system performance estimates. *Proceedings of the IEEE, 85*(9), 1390–1421.

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology, 16*(14), 1422-1427.

Serby, H., Tov, E. Y., & Inbar, G. F. (2005). An improved P300 brain computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 13*(1), 89-98.

Sutton, S., Braren, M., Zubin, J., & John, E. R. (1965). Information delivery and the sensory evoked potential. *Science, 155*, 1436–1439.

Thorpe, J. & Van Oorschot, P. C. (2004). Graphical dictionaries and the memorable space of graphical passwords. In *Proceedings of 13th USENIX Security Symposium*.

Thorpe, J., Van Oorschot, P. C., & Somayaji, A. (2005). Pass-thoughts: Authenticating with our minds. In *Proceedings of the New Security Paradigms Workshop*, Lake Arrowhead, California.

Tulving, E. & Schacter, D.L. (1990). Priming and human memory systems. *Science, 247*(4940), 301-306.

KEY TERMS

Authentication System: A system which securely identifies/authenticates users.

Biometrics: Recognising human beings using their intrinsic physical or behavioral traits.

Crossmodal Perception: Using more than on modality, say audio and video to analyze the perception skill or response from subjects.

Donchin Paradigm: It is an oddball paradigm that evokes P300 component and can be used by subjects to select alphabets or menus on screen.

Electroencephalogram (EEG): It is the neurophysiologic measurement of the electrical activity of the brain recorded from electrodes placed on the scalp.

Interstimulus Interval (ISI): *T*he time (interval) between the presentation of stimuli.

P300 Potential: A component in VEP which is evoked 300 ms after the presentation of the visual stimulus.

Visual Evoked Potential (VEP): A brain potential which is evoked on the presentation of a visual stimulus.

Visual Stimulus: A stimulus normally in the form of a picture or color shown on screen, which usually is used to evoke a VEP.

Biometric Technologies

Yingzi (Eliza) Du
Indiana University, Purdue University, USA

INTRODUCTION

Biometrics is an emerging technology for automatic human identification and verification using unique biological traits (Woodward, Orlans, & Higgins, 2002). These traits include face, fingerprints, iris, voice, hand geometry, handwriting, retina, and veins. For example, fingerprint recognition analyzes ridge ends, bifurcation, or dots of finger tips; voice recognition analyzes speech signal characteristics; iris recognition analyzes the pits, striations, filaments, rings, dark spots, and freckles of eyes; and face recognition analyzes facial parameters (Du *et al.*, 2004). It is based on "something you are" rather than "something you have" (Du, 2005). Compared to the traditional identification and verification ways, such as user name/password, and paper IDs, biometrics is more convenient to use, reduces fraud, and is more secure (Reid, 2004).

BACKGROUND

Biometrics has been widely used in criminal justice; U.S. immigration and naturalization services; and e-commerce and e-government. For example, fingerprints have long been used to identify criminals. The Department of Homeland Security (2004) has deployed the US-VISIT Program for border security, which uses biometric technologies to help secure the nation's borders and expedite the entry/exit process while enhancing the integrity of the immigration system and respecting the privacy of the visitors. Biometrics has been used to replace the user name and password in e-commerce and e-government for information access.

BIOMETRICS SYSTEM

Biometric system usually includes two subsystems: (1) the biometric enrollment system (Figure 1a) and (2) biometric matching system (Figure 1b).

The biometric enrollment system includes the Sensor Module, the Data Acquisition Module, the Data Preprocessing Module, the Pattern Analysis Module, the Pattern Extraction Module, and the Biometric Database Module. And the Data Acquisition Module interprets the biometric data into digital signals (images). The Data Preprocessing Module processes these signals to reduce the noise. The Pattern Analysis Module finds the most distinctive patterns of the biometric traits. The Pattern Extraction Module picks these distinctive patterns and generates identifiable templates. These templates will be then saved in the biometric database.

Compared to the biometric enrollment system, the biometric matching system adds the Pattern Matching Module and the Decision Module. In the biometric matching system, the newly sensed biometric data will be first processed similarly as the enrollment data, and the system will generate the pattern templates from the data. The Pattern Matching Module compares the newly generated templates with those in the biometric database and calculates match scores or quality scores for final decision. If the matching score is higher than the predetermined threshold, the system identifies/verifies it.

The false acceptance rate (FAR) and the false rejection rate (FRR) are used to measure if the biometric system is reliable (Ratha, Connell, & Bolle, 2001). A biometric system that generates high scores of either FAR or FRR is not reliable and cannot be used.

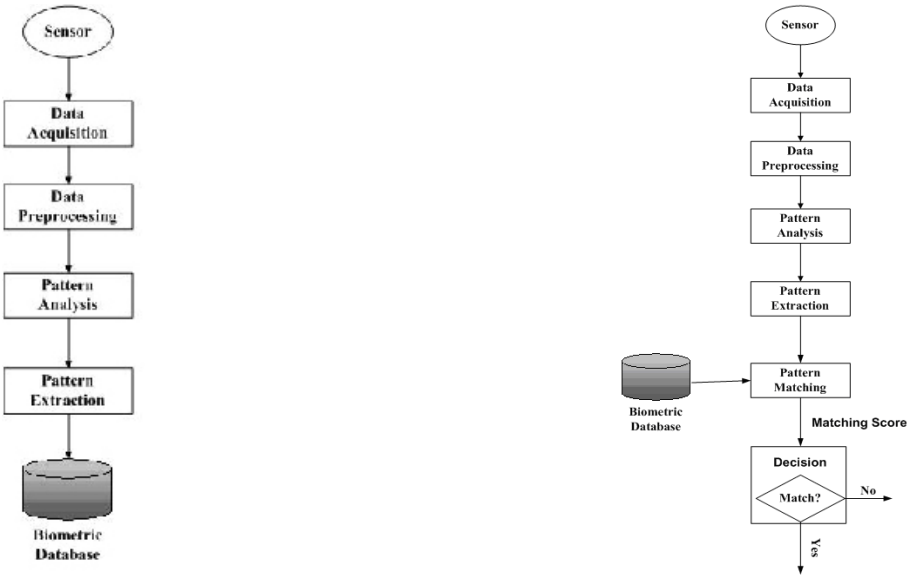
The FAR measures the percentage of incorrect identification:

$$FAR(\%) = \frac{\text{Number of false acceptance}}{\text{Total number of acceptance by the system}} \times 100\% \cdot \tag{1}$$

The FRR measures the percentage of incorrect rejection:

$$FRR(\%) = \frac{\text{Number of false rejections}}{\text{Total number of rejections by the system}} \times 100\% \cdot \tag{2}$$

Figure 1. Biometrics system



(a) Biometric Enrollment System

(b) Biometric Matching System

BIOMETRICS TECHNOLOGIES

Currently, the common used biometric systems include fingerprint, iris, face, and voice. The following paragraphs briefly describe each biometrics technology.

Fingerprint Recognition

Fingerprints have been extensively used in modern law enforcement. Fingerprint recognition is a well established and accepted method for person identification. Every person has minute raised ridges, which display a number of characteristics known as minutiae (Figure 2). The minutiae do not change naturally during a person's life (Pankanti, Prabhakar, & Jain, 2002). Fingerprint recognition systems usually extract information about the location, type, and direction of significant minutiae and generate templates for matching.

There are three types of fingerprint recognition systems to acquire digital fingerprint images:

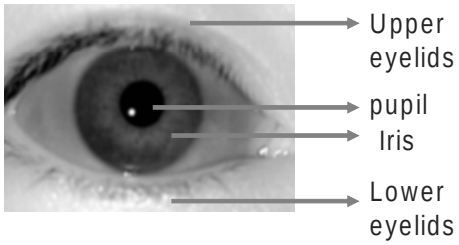
- **Optical sensors.** The optical fingerprint devices can generate very high resolution images and are often used by law enforcement and gate/door access. However, this kind of sensor is sensitive to the dirt and grease that may be on the finger.
- **Solid-state sensors.** Using integrated circuit (IC) to generate the image of the fingerprint. This kind of sensor is more cost efficient and easy to be integrated with other devices. But the image quality is not very high.
- **Ultrasound sensors.** An ultrasound camera is used to acquire images from the finger. This approach allows distinguishing between real fingers and any imitations. Furthermore, it is not sensitive to any dirt, grease, and so forth. However, this kind of fingerprint system is very expensive and not ready for mass-market application yet.

Figure 2. Fingerprint in detail

Iris Recognition

The iris is the round, pigmented tissue that lies behind the cornea (Figure 3). Compared to other kinds of biometric systems, such as face recognition and fingerprint recognition systems, iris recognition is more reliable and can achieve a higher accuracy rate.

Figure 3. Eye



Ophthalmologists Flom and Safir (1987) first noted that the iris is very unique for each person and remains unchanged after the first year of human life. In 1994, Daugman (2004) invented the first automatic iris recognition system. The iris has proven to be the most stable and reliable means of biometric identification.

To analyze the iris pattern and generate the iris templates, there are different approaches (Du *et. al.* 2004). Daugman (2004) used a quadrature 2-D Gabor wavelet method to analyze both coherent and incoherent detailed texture of the iris. Recently, Du *et. al.* (Apri, 2004) designed the local texture pattern (LTP) approach to analyze iris patterns and generated a 1-D iris signature.

Currently, there are three kinds of iris recognition systems:

- **PC Iris recognition system.** This system uses a low-end iris camera. It is the cheapest iris recognition system, but it can only hold a very small database. This kind of system is popular when used for computer access control.
- **Walkup iris recognition system.** This system uses a server-client distribution system and comprises multiple iris cameras. It is used for large database application.
- **Standalone/portable iris recognition system.** This system is a fully self-contained iris enrollment and recognition system. It is used for field applications.

Face Recognition

Face recognition is a person's primary method of personal identification. It can be captured remotely. Face recognition uses mathematical models to analyze the features in the face. Traditionally, the face is captured by a digital camera/video camera and is saved as a 2-D image. The face recognition system processes the 2-D image and analyze the face features and generate face templates.

Recently, researchers have developed 3-D cameras to capture 3-D face features. A 3-D camera is usually composed with two or more regular digital cameras. Two cameras could be arranged in such a way that they would take depth information of the face; just as our eyes. Our two eyes enabled us to see the 3-D objects. The 3-D face recognition system applies 3-D face models to the problem of robust face recognition. In particular, the 3-D face models address the two most critical and complicating factors affecting 2-D face recognition performance: illumination and pose variation.

Voice Recognition

Speech is produced by vibrating the vocal cords. The vowel sounds are perhaps the most interesting class of sounds in English (Du, 2005). Information embedded in speech can be divided into three categories: (1) linguistic (words in the speech), (2) paralinguistic (the way of delivery), and (3) nonlinguistic information (facial expression, hand gestures, and the speaker properties) (Katagiri, 2000). Voice recognition in English is more advanced and more accurate than any other language. In English, the voice recognition systems rely heavily on vowel recognition to achieve high performance. A person's voice does not, over a life span, vary with age, mood, stress, colds, and allergies.

CHALLENGES AND PROMISES OF BIOMETRICS

Accuracy

Most statistics about the accuracy rate of biometrics are performed in well-controlled laboratory environments. Even though researchers/engineers have tried their best to synthesize real-life situations, there are always surprises in real-world applications. In fact, the biometric system could encounter a lot of difficulties, such as unexpected background noises, resolution problems, and unusual user behaviors. These factors will result in much lower accuracy rates of biometric systems. For example, face recognition systems

can achieve as high as 94% accuracy rate in a well-controlled laboratory environment (indoor with good light situation). However, face recognition systems can only achieve a 50% accuracy rate or even lower in an outdoor situation.

In addition, the accuracy rate of some biometrics systems is obtained by a very small number of experiments. This kind of accuracy rate has no such meaning in statistics or real-life application.

Moreover, different kinds of biometrics systems have different accuracy rates. This is a result of the characteristics of different biometric traits. For example, the accuracy rate of the voice recognition system is much lower than that of the fingerprint recognition system because the voice of a human is not stable. The makeup and expression will change the appearance of a person and result in recognition failure by the face recognition system.

It is very important to choose the right kind of biometric system for the particular application. For an extremely secured application, the iris and fingerprint recognition systems would be more preferred than the face or voice recognition systems. However, the face and voice recognition system could be more suitable for surveillance because they do not need cooperation from the users. Compared to the fingerprint recognition system, the iris recognition system has a higher accuracy rate. However, in the forensic applications, iris recognition would not be useful because the possible changes of the iris in a dead body.

Vulnerabilities

No system is perfect. There is always one way or another to fail/spoof the system because the algorithms used by different vendors for different biometric products are based on certain mechanisms/hypothesis, which can be fooled.

There are three possible ways to trick a biometric system (Thalheim, Krissler, & Ziegler, 2002):

- **Use artificially created biometric traits.** Tsutomo Matsumoto, a Japanese cryptographer, created a fake finger—"gummy fingers"—using cheap and commonly available materials. This fake finger could fool the optical fingerprint recognition system but not the solid-state or ultrasound fingerprint recognition system.
- **Trick the system by playing back the biometric data captured in the previous authentication process.** Some voice recognition systems could be fooled by presenting the recorded version of a person's voice. This usually happens to traditional voice recognition systems that ask the user to speak a fixed word or phrase. The new voice recognition systems will randomly ask the users to speak random generated words or phrases. Some 2-D face recognition systems could be fooled by presenting a person's photo. But this could be avoided by using 3-D face recognition systems or using a liveness test.
- **Theft of the biometric data after getting into the database that stores the biometric templates.** This is a common problem of biometrics systems. It is very important for the biometric system to encrypt the biometrics data and the personal information of the user. It is also important for the security of the network if the biometric system uses server-client architecture.

Liveness test is to test if the biometric traits are from a living person rather than an artificial or lifeless person. The liveness test is very important to enforce the data integrity of the biometrics system. For example, the natural papillary response (changing pupil size in response to changes in illumination) can be used to confirm the liveness of an iris (Ma, Tan, Wang, & Zhang, 2003). The motion of a person could be used to confirm the liveness of a face. The response of a finger to a low electronic current could be used to test the liveness of a finger.

Privacy

The issue of privacy is central to biometrics. Any technology is a two-edged sword. On one hand, biometrics is used to enforce information integrity and protect privacy. On the other hand, misusing of biometrics can seriously harm individual privacy. In fact, biometrics itself is by no means privacy invasive. It is the misuse of the biometric data that is horrifying. Privacy policies, regulations, and ethic training of biometric system operators should be in place and strictly enforced to ensure that the privacy rights of the people are protected (Woodward et al., 2002).

MULTIMODAL BIOMETRICS

Multimodal biometrics is to use two or more biological characteristic for person identification or verification. Unimodal biometrics is to use only one kind of biological trait for identification or verification. Each kind of unimodal biometric system has its advantages and disadvantages. Table 1 compares different kinds of biometric technologies in terms of performance, cost, and their applications.

All aforementioned unimodal biometrics can be used for verification purposes. Each biometric system has its advantages and disadvantages. Only fingerprint and iris verification can provide reliable identifications for a large database. However, face and voice recognition systems can be used for surveillance or remote access. Voice recognition may achieve the lowest accuracy rate, but it is nonintrusive to the users.

By using multiple biometric traits and data fusion, multimodal biometrics could improve the accuracy rate

Table 1. Comparison of biometric systems

	Accuracy	Reliability	Stable	ID	Low Cost	Intru-sive	User Co-op	Large Population
Finger-print	High	High	Yes	Yes	Yes	Yes	Yes	Yes
Face	Medium	Medium	No	No	No	No	No	No
Iris	Very High	Very High	Yes	Yes	No	Yes	Yes	Yes
Voice	Low	Low	No	No	Yes	No	No	No

and reduce the vulnerabilities. For example, a multimodal biometric system which is composed of a face recognition system and a voice recognition system could possibly achieve similar accuracy as a fingerprint system (Frischholz & Dieckmann, 2000).

A single biometric system might be easily fooled by artificial/faked biometric traits. However to fake multiple biometrics traits would not be an easy task. A combination of different types of biometric systems could also improve the system performance in variable environments. For example, a face recognition system combined with an iris recognition system could improve the accuracy rate while keeping the ability of remote tracking.

FUTURE TRENDS

With the development of technologies, the biometric systems are becoming more and more reliable and convenient for authentication and identification. The demands for biometric systems have grown in markets, from personal log in to PCs to large scale entry/exit security systems. In the near future, biometrics will become part of our daily life. We will use biometrics to open the door, withdraw money from ATMs, borrow books from libraries, and so forth.

CONCLUSION

Biometrics is the automated use of biological features of a human being to positively identify a person. Compared to the traditional identification and verification methods, biometrics is more convenient to use, reduces fraud, and is more secure. There are needs for biometrics in federal, state, and local governments; in the military; and in commercial applications. While it is very important to choose the right kind of biometric system for a particular application, multimodal biometrics will be a trend in the deploying of biometrics systems.

Over the years, there are concerns about privacy issues in the applications of biometrics, and it will continue to be an important issue. Privacy policies, regulations, and ethic training of biometric system operators will be very necessary and important to have in place and strictly enforced to ensure that the privacy rights of the people are protected.

REFERENCES

Chuah, L. E. (2002). The future challenges of biometrics Retrieved from http://www.giac.org/practical/LeeEng_Chuah_GSEC.doc

Court, W. (2003). Biometrics: Evaluation criteria and scenario based performance testing. Retrieved from http://www.giac.org/practical/GSEC/Warren_Court_GSEC.pdf

Daugman, J. (2004). How iris recognition works. *IEEE Transaction on Circuits and Systems for Video Technology, 14*(1), 21- 30.

Department of Homeland Security. (2004). *US-VISIT program overview.* Retrieved from http://www.dhs.gov/dhspublic/interapp/editorial/editorial_0333.xml

Du, Y. (2005). Biometrics: Technologies and trends. In *Encyclopedia of optical engineering* (3rd ed.). Dekker.com.

Du, Y., Ives, R. W., Etter, D. M. (2004). Iris recognition. In *Electrical engineering handbook* (3rd ed.). Boca Raton, FL: CRC Press.

Du, Y., Ives, R. W., Etter, D. M., & Welch, T. B., (2004). Biometric signal processing laboratory. *IEEE International Conference on Acousitics, Speech, and Signal Processing.*

Du, Y., Ives, R. W., Etter, D. M., Welch, T. B. & Chang C.-I. (2004). One dimensional approach to iris recognition. *Proceedings of the SPIE's.*

Flom, L., & Safir, A. (1987). Iris recognition system. (United States Patent No. 4,641,349). Washington DC: U.S. Government Printing Office.

Frischholz, R. W., & Dieckmann, U. (2000). BioID: A multimodal biometric identification system. *Computer, 33*(2), 64-68.

International Biometric Group. (2004). *Biometrics market and industry Report 2004-2008.* Retrieved from http://www.biometricgroup.com/reports/public/market_report.html

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(1), 4-20.

Katagiri, S. (2000). *Handbook of neural networks for speech processing.* Norwood, MA: Artech House.

Ma, L., Tan, T., Wang, Y., & Zhang, D. (2003). Personal identification based on iris texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(12), 1519-1533.

Pankanti, S., Prabhakar, S., & Jain, A. K. (2002). On the individuality of fingerprints. *IEEE Transactions on Pattern Analysis and Machine Intelligences, 24*(8), 1010-1025.

Ratha, N. K., Connell, J. H., & Bolle, R. M. (2001). Enhancing security and privacy in biometrics based authentication systems. *IBM Systems Journal, 40*(3). Retrieved from http://www.research.ibm.com/journal/sj/403/ratha.html

Reid, P. (2004). *Biometrics for network security.* Upper Saddle River, NJ: Prentice Hall.

Thalheim, L., Krissler, J., & Ziegler, P.-M. (2002). Body check. Retrieved from http://www.heise.de/ct/english/02/11/114/

The 9-11 Commission Report. (2004). Retrieved from http://www.gpoaccess.gov/911/

Woodward, J. D., Orlans, N. M., & Higgins, P. T. (2002). *Biometrics.* Berkeley, CA: McGraw-Hill.

KEY TERMS

Biometrics: An emerging field of technology that uses unique physical, biological, or behavioral traits for automatic human identification and verification.

Face Recognition: Automatic recognition of a person's identity by mathematical analysis of person's face features.

False Acceptance Rate (FAR): FAR is the percentage of incorrect identification. It is defined as the total number of acceptance by the system divided by the number of false acceptance.

False Rejection Rate (FRR): FRR is the percentage of incorrect rejection. It is defined as the total number of rejection by the system divided by the number of false rejection.

Fingerprint Recognition: Automatic recognition of a person's identity by mathematical analysis of the patterns of fingerprints.

Iris Recognition: Automatic recognition of a person's identity by mathematical analysis of the random patterns that are visible within the iris of an eye from some distance.

Liveness Test: A test performed to test if the biometric traits are from a living person rather than an artificial or lifeless person.

Multimodal Biometrics: Automatic recognition of a person's identity using more than one physical, biological, or behavioral characteristic.

Voice Recognition: Automatic recognition of a person's identity by mathematical analysis of person's voice features.

Blended Learning Models

Charles R. Graham
Brigham Young University, USA

INTRODUCTION

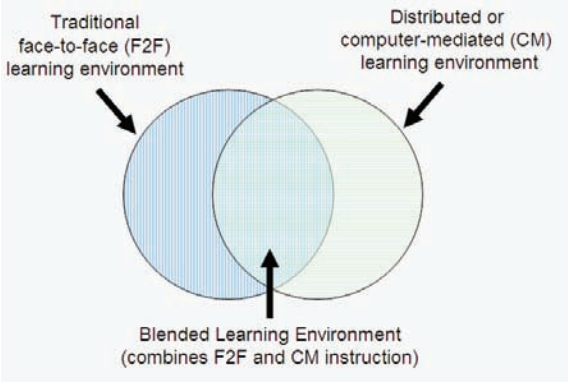
Technological advances and widespread access to information and communication technologies (ICTs) have facilitated the rapid growth of blended learning approaches in both higher education and corporate training contexts. In 2002, the president of Pennsylvania State University expressed his belief that blended learning was "the single greatest unrecognized trend in higher education" (Young, 2002, p. A33). At the same time, the American Society for Training and Development also identified blended learning as one of the top 10 emergent trends in the knowledge delivery industry (Finn, 2002). Since then, the visibility of blended learning environments has increased dramatically in both formal education and corporate training settings. At the third annual Sloan-C Workshop on Blended Learning and Higher Education, Frank Mayadas, the program director for the Alfred P. Sloan Foundation, predicted that "by 2010 you will be hard pressed to find a course that is not blended" (Mayadas, 2006). There is increasing interest in the concept of blended learning as evidenced by greater numbers of books, journal articles, and trade magazine articles that directly address issues related to blended learning. This article will provide an overview of current models of blended learning and provide references to the most recent resources in this emergent area of research and practice.

BACKGROUND

Definition

The use of the term blended learning is relatively new in both higher education and corporate settings. In higher education, the term "hybrid course" was often used prior to the emergence of the term "blended learning," and now the two terms are used interchangeably. Because term is relatively new, there are still ongoing debates regarding the precise meaning and relevance of the term (Driscoll, 2002; Graham, Allen, & Ure, 2003; Laster, 2004; Masie, 2005; Oliver & Trigwell, 2005; Osguthorpe & Graham, 2003). However, the most commonly held position is that *blended learning environments combine face-to-face instruction with technology-mediated instruction* (Graham, 2005; Graham et al., 2003). This definition highlights the ongoing convergence of two archetypal learning environments: the traditional face-to-face (F2F) environment with the distributed (or technology-mediated) environment (see Figure 1).

Figure 1. Blended learning combines traditional face-to-face and computer mediated instruction



Purposes

There are many reasons why a blended approach to learning might be selected. The three most common reasons for blending listed in the literature are:

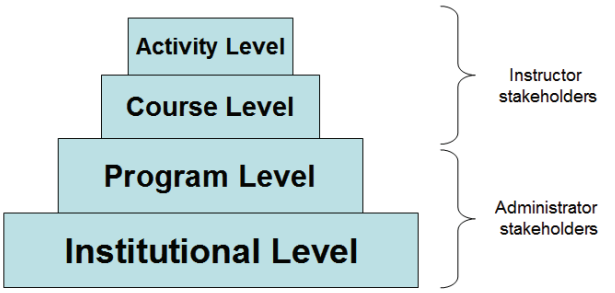
- To increase learning effectiveness
- To increase convenience and access
- To increase cost effectiveness

Often educators adopt a blended approach in order to explore tradeoffs between more than one of these goals simultaneously (e.g., increasing the convenience to students afforded by an asynchronous distributed environment without completely eliminating the human touch from the F2F environment). While blended learning is appealing to many because it enables one to take advantage of the "best of both worlds" (Morgan, 2002; Young, 2002), blended learning environments can also mix the least effective elements of both F2F and technology-mediated worlds if not designed well.

MODELS

The concept of blended learning is simple and elegant. However, there are numerous ways that blended learning

Figure 2. Different levels where blended learning can occur



can be implemented in a wide variety of different contexts. For this reason, it is important to share successful models of blended learning so that all can benefit. Sharing *models* of blended learning can help to facilitate the *purposeful* and *disciplined* adoption of appropriate blended learning strategies. This section of the article will present several models of blended learning. Because of space constraints it is not possible to share all of the details of the models, but a rich set of references is provided that will allow the reader to find additional details for the examples of interest.

It is important to understand that blending occurs at many different levels including the institutional level, the program level, the course level, and the activity level (see Figure 2). Typically, models at the course and activity levels have instructor stakeholders who are primarily interested in issues of learning effectiveness and productivity. Blended learning that occurs at the program and institutional levels typically has administrator stakeholders who are often driven by issues of cost effectiveness and expanding access of the learning to untapped audiences. Specific examples of blended learning at each of these levels can be found in *The Handbook of Blended Learning* (Graham, 2005) and *The Encyclopedia of Distance Learning* (Graham & Allen, in press).

Because there is such a wide range of possible blends in the different contexts, it can be helpful to think of three major categories of blends: enabling blends, enhancing blends, and transforming blends. Table 1 contains a description of each category and specific examples for each.

The distinctions here are particularly important when considering the impact of blended learning on learning ef-

Table 1. Three categories of blends with examples

Category	Description	Examples
Enabling Blends	Enabling blends primarily focus on addressing issues of *access* and *convenience*. They often use information and communication technologies as a way to provide "equivalent" learning experiences to the predominant face-to-face modality.	1. Many of the for-profit institutions like University of Phoenix (Lindquist, 2005) have models that focus on making educational opportunities available to those who do not have access due to time and location constraints. 2. National University has a teacher preparation program geared toward access and flexibility (Reynolds & Greiner, 2005). 3. Many international education and training programs are also focused on providing access (e.g., World Bank, Jagannathan, 2005, Mexico's Red Escolar program, Acuña Limón, 2005, etc.).
Enhancing Blends	Enhancing blends allow for incremental changes to the pedagogy. They are often characterized by the inclusion of supplemental online resources and/or the implementation of online activities that are small in scope when compared to the overall course.	4. University of Glamorgan, Wales (Jones, 2005) has a continuum of e-learning that includes four levels, the first two of which represent enhancing blends: (1) Basic ICT usage (e.g., PowerPoint presentations) and (2) E-enhanced (e.g., access to online resources, use of Bb for productivity such as announcements, lecture notes, etc.). 5. University of Waikato, New Zealand (Wright, Dewstow, Topping, & Tappenden, 2005) has a model for enhancing F2F courses that includes levels such as "Supported Online" (e.g., traditional F2F with access to materials provided online) and "Somewhat Online" (e.g., includes an online course component for on-campus students). 6. University of Central Florida, U.S. (Dziuban, Hartman, Juge, Moskal, & Sorg, 2005) has a model that includes "W courses" (e.g., fully online), M courses (e.g., mixed, reduced F2F contact courses), and E courses (e.g., Web enhanced courses). E courses use online or Web components to enhance a traditional F2F course.
Transforming Blends	Transforming blends allow for a significant change in pedagogy that facilitates active learner construction of knowledge.	7. Use of instructional simulations such as the Virtual Audiometer and Virtual Chem Lab at Brigham Young University are changing the ways in which students learn and solve problems (Graham & Robison, in press; West & Graham, 2005). 8. Authentic learning environments that bring real world contexts into the classroom (Oliver & Trigwell, 2005) or integrate formal learning with workplace learning (Collis, 2005; DeViney & Lewis, 2005; Singh, 2005) can be supported through the use of blended learning approaches. 9. Mixed reality technologies facilitate the blending of F2F and virtual worlds and are transforming the kinds of learning and performance support that is taking place in industrial and military contexts (Kirkley & Kirkley, 2005; Wisher, 2005).

Figure 3. Three paths for designing blended learning environments

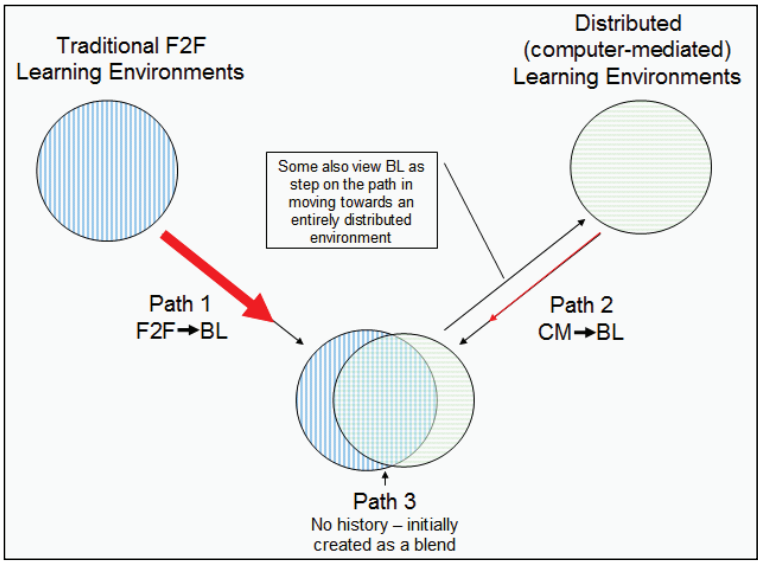
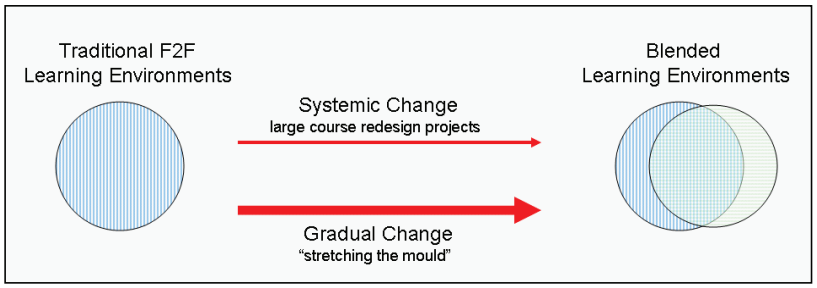


Figure 4. Two ways of moving from a traditional F2F learning environment to a blended learning environment



fectiveness. An enhancing blend might serve as a stepping-stone to a more transformative blend, or it might end up superficially impacting student learning (Graham & Dziuban, submitted; Graham & Robison, in press).

Higher Education Models

In higher education the primary path to blended learning is from a predominantly F2F environment to a blended environment (see Path 1 in Figure 3). There is also a path (though much smaller) from an entirely distributed environment to a blended environment (see Path 2 in Figure 3). Path 3 (see Figure 3) occurs most often in corporate contexts when new programs and courses are developed from scratch to meet an emerging need.

Models that involve Path 2 movement typically occur with the goal of adding "human touch" to a distance course or program where access to the F2F environment is possible but less convenient. In these cases we see blends including a limited number of F2F events such as a residency requirement (Offerman & Tassava, 2005; Pease, 2005), F2F orientations and/or final project presentations (Lindquist, 2005), or optional F2F help sessions for struggling learners.

Faculty adoption of blended learning that involves Path 1 movement occurs via gradual change or systemic change (see Figure 4). Most often it occurs via gradual change or what Collis and van der Wende (2002) call "stretching the mould." This involves exploration and adoption of blended learning strategies such as enhancing a course with online resources and activities.

At first, F2F contact time may not be reduced because enhancements are small and exploratory. As online activities become more successful and more integral to the course, faculty reduce their F2F contact time to accommodate the online activities. Researchers have documented the tendency among many faculty designing blended learning

courses to keep adding online components to the traditional course without eliminating anything. This phenomenon is known as the course-and-a-half syndrome (Kaleta, Skibba, & Joosten, in press).

To date, systemic change has not been as wide spread as gradual change toward blended learning. Systemic approaches to designing blended learning involve whole course redesign. Some of the best documented cases of course redesign efforts involving blended learning come from the Program in Course Redesign supported by the PEW Charitable Trusts (Twigg, 2003). The goal of this project was to see if 30 large enrollment courses from across the United States could be redesigned using technology to simultaneously provide both reductions in cost and gains in learning outcomes. The majority of the course redesign efforts involved moving from a traditional course delivery to a blended learning approach. Table 2 outlines five specific models of blended learning that resulted from the efforts.

Corporate and Military Models

Blended learning in corporate settings is driven by the desire to improve return on investment (ROI) of training dollars. This means driving down the costs and trying to increase the impact of training. Bersin and Associates (2003) studied blended training in major programs that impacted over 100,000 employees at 15 large corporations. They found that the ROI for blended learning programs was 100%+ in almost every case and much larger in some cases. Similarly, IBM documented a 17:1 ROI for deployment of its blended learning leadership management program (Lewis & Orton, 2005).

Blended learning models in corporate settings are even more varied than in higher education settings. F2F human interaction is arguably the most powerful learning intervention and the most costly (Lewis & Orton, 2005). Data show that instructor led training (ILT) is still by far the most prevalent mode of training delivery (Ziob & Mosher, 2005). So, most blended learning models seek to use human interaction strategically and replace much of the F2F interaction with interactive simulations, performance support systems, or technology mediated interactions with colleagues that eliminate the need for expensive travel.

For example, IBM has a four-tiered approach where learners have access to a performance support database, interactive learning simulations, a live-virtual collaborative learning environment, and F2F learning laboratories. Learners in the program begin with 26 weeks of self-paced online learning after which they participate in a five-day in-class learning lab. The F2F lab experience is followed by a 25-week online learning experience that focuses on application of skills and knowledge (Lewis & Orton, 2005). Similarly, Oracle's leadership training program (Hanson & Clem, 2005) and Avaya's sales training program (Chute, Williams, & Hancock, 2005) both used limited and strategically placed F2F sessions embedded within a wide variety of computer-mediated and self-paced activities to reach their goals. The use of technology to facilitate training is

Table 2. Models developed from 30 course redesign efforts sponsored by the PEW charitable trusts (Twigg, 2003)

Model	Description
Supplemental Model	• Lecture portion of class kept intact • Supplemental online materials provided • Online quizzes • Additional online activities
Replacement Model	• Reduction of in-class meeting time • Replacement of face-to-face (F2F) class time with online activities • Online activities can take place in a computer lab or at home

Buffet Model	• Student chooses learning options
	  • Lecture
	  • Online
	  • Discovery laboratories
	  • Individual projects
	  • Team/group activities
	  • And so forth

Emporium Model	• Eliminates class meetings
	• Substitutes a learning resource center with
	  (1) online materials and
	  (2) on-demand personal assistance

Fully Online Model	• All online learning activities • No required F2F class meetings • (In some cases) optional F2F help

facilitating a greater integration between formal learning and informal or workplace learning (Collis, 2005; DeViney & Lewis, 2005; Singh, 2005). Increasingly, learners are able to engage with a formal instructor at a distance and have learning activities and assignments mediated in the local context by a manager or mentor.

A second corporate model that is worth mentioning can be seen in the Cisco Networking Academy (Dennis et al., 2005; Selinger, 2005). The Cisco Networking Academy is a global training program for Internet technology skills that is implemented in more than 150 countries across the world and has over 400,000 enrollments. The academy provides centralized Web-based curriculum, online assessments, and tracking of student performance. At each academy site instructors are able to use the Web-based content and customize it to support the specific needs of their local students. This approach allows courses to be offered as Web-based training or instructor-led training, or a blend of both to best accommodate the learning preferences and work styles of the learners (Selinger, 2005). This blended approach also facilitates cultural adaptation and localization of curriculum to meet diverse cultural needs.

Finally, military and some high-tech industrial contexts are employing mixed reality environments that blend F2F interactions with interactions in a virtual world (Kirkley & Kirkley, 2005). For example, the U.S. military is training with live-virtual-constructive learning exercises, which meld the real world with the simulated world. Wisher (2005, p. 527) writes about one such exercise with the task of conducting an amphibious assault that involved "seventeen military units (live), six simulators (virtual), and twenty one simulations (constructive)."

FUTURE TRENDS

It is hard to predict exactly what the future holds for blended learning environments. It is very likely that the use of blended learning in both higher education and corporate contexts will continue to grow. In fact, there may come a time when the traditional learning environment is predominantly a blended learning environment and it no longer makes sense to use the adjective "blended." An example of this is the fact that the University of Central Florida has considered dropping its (E)nhanced course designation because virtually all the university courses have a Web presence (Dziuban, 2006). There is likely to be an increased focus in higher education on the transformative potential of blended learning (Garrison & Kanuta, 2004; Graham & Robison, in press). Rather than focus on whether blending is happening or not, universities will focus more on the quality of the blend and seek to understand how faculty can be trained and supported to teach in blended learning environments.

There is evidence that administrators and students in K-12 environments (particularly in high school and home school settings) are beginning to explore the possibilities of blended learning. Corporate and military contexts are likely to be the ones that continue to push the technological envelope, exploring the use of more expensive technologies, although increasingly simulations may be used in K-12 and higher education classrooms.

Finally, Bonk, Kim, and Zeng (2005, p. 560) make 10 predictions related to blended learning in the future:

1. the increased use of mobile devices in blended learning
2. greater use of visualization tools and hands-on learning in blended learning
3. increased learner input in the design of their own learning programs
4. increased connectedness, community, and collaboration
5. increased authenticity and on-demand learning
6. stronger ties between work and learning
7. calendaring system will need to change and be more flexible
8. programs will begin to include blended learning course designations
9. instructor roles will increasingly move toward that of mentor, coach, and counselor
10. blended learning specialist teaching certificates will emerge

CONCLUSION

During the past decade, distributed learning has made huge strides in popularity in both higher education and corporate sectors of society. The use of technology has increased access to educational resources and facilitated communication in a way that was not previously possible. Despite the strengths that online learning environments provide, there are different strengths inherent in traditional F2F learning environments. The current trend toward blending both online and F2F instruction is a positive direction and merits increased attention and study. Because the possibilities inherent in a blended environment are so vast, it is important that we begin to develop and share successful models of blended learning at all the different levels (see Figure 2) and contexts in which it can occur.

REFERENCES

Acuña Limón, A. (2005). Tecnológico de Monterrey in México: Where technology extends the classroom. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learn-*

ing: Global perspectives, local designs (pp. 351-359). San Francisco: Pfeiffer Publishing.

Bersin & Associates. (2003). *Blended learning: What works?: An industry study of the strategy, implementation, and impact of blended learning*. Bersin & Associates.

Bonk, C. J., Kim, K.-J., & Zeng, T. (2005). Future directions of blended learning in higher education and workplace settings. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 550-567). San Francisco: Pfeiffer Publishing.

Chute, A. G., Williams, J. O. D., & Hancock, B. W. (2005). Transformation of sales skills through knowledge management and blended learning. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 105-119). San Francisco: Pfeiffer Publishing.

Collis, B. (2005). Putting blended learning to work. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 461-473). San Francisco: Pfeiffer Publishing.

Collis, B., & van der Wende, M. (2002). *Models of technology and change in higher education: An international comparative survey on the current and future use of ICT in higher education*. Enschede, NL: Center for Higher Education Policy Studies, University of Twente.

Dennis, A., Bichelmeyer, B., Henry, D., Cakir, H., Korkmaz, A., Watson, C., et al. (2005). The Cisco Networking Academy: A model for the study of student success in a blended learning environment. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 120-135). San Francisco: Pfeiffer Publishing.

DeViney, N., & Lewis, N. J. (2005). On-demand learning: How work-embedded learning is expanding enterprise performance. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 491-501). San Francisco: Pfeiffer Publishing.

Driscoll, M. (2002, March 1). Blended learning: Let's get beyond the hype. *E-learning,* p. 54.

Dziuban, C., Hartman, J., Juge, F., Moskal, P., & Sorg, S. (2005). Blended learning enters the mainstream. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 195-208). San Francisco: Pfeiffer Publishing.

Finn, A. (2002). Trends in e-learning. *Learning Circuits, 3*(11). Retrieved from http://www.learningcircuits.org/2002/nov2002/finn.htm

Garrison, D. R., & Kanuta, H. (2004). Blended learning: Uncovering its transformative potential in higher education. *The Internet and Higher Education, 7*(2), 95-105.

Graham, C. R. (2005). Blended learning systems: Definition, current trends, and future directions. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 3-21). San Francisco: Pfeiffer Publishing.

Graham, C. R., & Allen, S. (in press). Designing blended learning environments. In C. Howard, J. V. Boettecher, L. Justice, K. D. Schenk, P. L. Rogers, & G. A. Berg (Eds.), *Encyclopedia of distance learning* (2nd ed.). Hershey, PA: Idea Group Reference.

Graham, C. R., Allen, S., & Ure, D. (2003). *Blended learning environments: A review of the research literature*. Retrieved May 29, 2006, from http://msed.byu.edu/ipt/graham/vita/ble_litrev.pdf

Graham, C. R., & Dziuban, C. D. (submitted). Core research and issues related to blended learning environments. In J. M. Spector, M. D. Merrill, J. J. G. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed.). Mahwah, NJ: Lawrence Earlbaum Associates.

Graham, C. R., & Robison, R. (in press). Realizing the transformational potential of blended learning: Comparing cases of transforming blends and enhancing blends in higher education. In A. G. Picciano & C. D. Dziuban (Eds.), *Blended learning: Research perspectives*. Sloan Consortium.

Hanson, K. S., & Clem, F. A. (2005). To blend or not to blend: A look at community development via blended learning strategies. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 136-149). San Francisco: Pfeiffer Publishing.

Jagannathan, S. (2005). Blended e-learning in the context of international development: Global perspectives, local design of e-courses. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 444-458). San Francisco: Pfeiffer Publishing.

Jones, N. (2005). E-college Wales, a case study of blended learning. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 182-194). San Francisco: Pfeiffer Publishing.

Kaleta, R., Skibba, K., & Joosten, T. (in press). Discovering, designing, and delivering hybrid courses. In A. G. Picciano & C. D. Dziuban (Eds.), *Blended learning: Research perspectives*: Sloan Consortium.

Kirkley, J. R., & Kirkley, S. E. (2005). Expanding the boundaries of blended learning: Transforming learning with

mixed and virtual reality technologies. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 533-549). San Francisco: Pfeiffer Publishing.

Laster, S. (2004). Blended learning: Driving forward without a definition. In J. C. Moore (Ed.), *Engaging communities: Wisdom from the Sloan Consortium* (pp. 153-162). Needham, MA: Sloan Consortium.

Lewis, N. J., & Orton, P. Z. (2005). Blended learning for business impact. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 61-75). San Francisco: Pfeiffer Publishing.

Lindquist, B. (2005). Blended learning at the University of Phoenix. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 223-234). San Francisco: Pfeiffer Publishing.

Masie, E. (2005). The blended learning imperative. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 22-26). San Francisco: Pfeiffer Publishing.

Mayadas, F. (2006). Keynote address at the Sloan-C Workshop on Blended Learning and Higher Education: Blended Learning, Localness, and Outreach. Chicago, IL.

Morgan, K. R. (2002). *Blended learning: A strategic action plan for a new campus*. Seminole, FL: University of Central Florida.

Offerman, M., & Tassava, C. (2005). A different perspective on blended learning: Asserting the efficacy of online learning at Capella University. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 235-244). San Francisco: Pfeiffer Publishing.

Oliver, M., & Trigwell, K. (2005). Can 'blended learning' be redeemed? *E-learning, 2*(1), 17-26.

Osguthorpe, R. T., & Graham, C. R. (2003). Blended learning systems: Definitions and directions. *Quarterly Review of Distance Education, 4*(3), 227-234.

Pease, P. S. (2005). Blended learning goes totally virtual by design: The case of a for-profit, online university. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 245-260). San Francisco: Pfeiffer Publishing.

Reynolds, T., & Greiner, C. (2005). Integrated field experiences in online teacher education: A natural blend. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 209-220). San Francisco: Pfeiffer Publishing.

Selinger, M. (2005). Developing an understanding of blended learning: A personal journey across Africa and the Middle East. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 432-443). San Francisco: Pfeiffer Publishing.

Singh, H. (2005). Blending learning and work: Real-time work flow learning. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 474-490). San Francisco: Pfeiffer Publishing.

Twigg, C. (2003). Improving learning and reducing costs: New models for online learning. *Educause Review, 38*(5), 28-38.

West, R. E., & Graham, C. R. (2005). Five powerful ways technology can enhance teaching and learning in higher education. *Educational Technology, 45*(3), 20-27.

Wisher, R. A. (2005). Blended learning in military training. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 519-532). San Francisco: Pfeiffer Publishing.

Wright, N., Dewstow, R., Topping, M., & Tappenden, S. (2005). New Zealand examples of blended learning. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 169-181). San Francisco: Pfeiffer Publishing.

Young, J. R. (2002, March 22). 'Hybrid' teaching seeks to end the divide between traditional and online instruction. *Chronicle of Higher Education,* p. A33.

Ziob, L., & Mosher, B. (2005). Putting customers first at Microsoft: Blending learning capabilities with customer needs. In C. J. Bonk & C. R. Graham (Eds.), *Handbook of blended learning: Global perspectives, local designs* (pp. 92-104). San Francisco: Pfeiffer Publishing.

KEY TERMS

Affordances: Features of an environment or artifact that "afford" or permit certain behaviors.

Blended Learning Environment: A learning environment that combines face-to-face and computer-mediated instruction.

Distributed Learning Environment: A learning environment where participants are not co-located and use computer-based technologies to access instruction and communicate with others.

Hybrid Course: Another name for a blended course. Typically a course that replaces some F2F instructional time with computer-mediated activities.

Performance Support Systems: Systems that are designed to improve human performance through many different kinds of interventions including but not being limited to instructional interventions.

Return on Investment (ROI): A measurement evaluating the gains versus the costs of an investment.

Technology-Mediated Learning Environment: Another name for a distributed learning environment.

Bonded Design

Andrew Large
McGill University, Canada

Valerie Nesset
McGill University, Canada

INTRODUCTION

It is hardly controversial to argue for user involvement in the technology design process: the issue rather is the extent of that involvement and whether or not this is related to the kind of user. In particular, can young children play a meaningful role in design, and if so, what should it be? Several design methodologies advocate a range of roles for children within the design process; this article presents a new such methodology, Bonded Design. Essentially, Bonded Design assumes an intergenerational team comprising adult designers and young users working together to produce a low-tech prototype. This team employs a variety of design techniques—conducting a user needs' assessment, evaluating existing technologies, brainstorming, discussing ideas as a group, prototyping (for example, through drawings), and consensus building—to achieve its goal.

Bonded Design emerged in 2003 from a research study to investigate whether elementary school students (specifically in grades three and six) could actively participate in designing Web portals. To accomplish this objective two intergenerational design teams were established, each including children alongside researchers, which produced two low-tech portal prototypes (Large, Beheshti, Nesset, & Bowler, 2004; Large, Nesset, Beheshti, & Bowler, 2006, 2007). These prototypes subsequently were converted into working portals that received high praise in their evaluations by elementary school students. Indeed, one of these portals, *History Trek*, is now operational on the Web, providing access to information in English and French on Canadian history (*http://www.historytrek.ca*).

BACKGROUND

Bonded Design did not emerge in a vacuum; a number of user-focused design methodologies have accommodated children in various ways and to various degrees in the design of technologies intended for use by children (Nesset & Large, 2004). The oldest and most conventional approach, "User-Centered Design," focuses on the impact of technology on users, but traditionally these users were only involved after the technology had been designed (Nesset & Large, 2004; Scaife & Rogers, 1999, Scaife, Rogers, Aldrich, & Davies, 1997). In other contexts, the term *user-centered design* has been understood by some authors to mean direct contact between users and designers throughout the design process (Rubin, 1994). Typically in User-Centered Design the users have little or no control over the design process itself. Fundamentally they are testers rather than designers, revealing design shortcomings rather than proposing design ideas. In this context, where children only act as testers of prototypes designed by adults for young audiences, their involvement is relatively uncontroversial.

Contextual Design is described by Beyer and Holtzblatt (1999, p. 32) as "a state-of-the-art approach to designing products directly from a designer's understanding of how the customer works." Designers collect data from users' own environments by observing them performing typical activities. They usually record observational data and conduct one-on-one interviews with users in order to develop a deeper understanding of the users' work practices. They then apply work modeling using such techniques as pictorial charts, storyboarding, and low-tech prototyping. In Contextual Design, therefore, the users' role is critical but passive: it is their behavior rather than their ideas that inform the process. This methodology can be applied to children as technology users when the classroom or home is substituted for the adults' workplace.

Soloway, Guzdial, and Hay (1994), based on the idea that the long-term goal of computing is to make people smarter, decided that the HCI community needed to move from the traditional "user-centered" design to what they term "Learner-Centered Design." This approach assumes that everyone is a learner, whether a professional or a student. The main focus of Learner-Centered Design is to ensure that the design is adapted to the interests, knowledge, and styles of the learners who use it. Soloway et al. (1994) believe in the educational philosophy of "learning by doing." At the heart of Learner-Centered Design are understanding (how will the learner learn the practice?), motivation (how can technology motivate a learner?), diversity (every learner is different—what kind of technology can be developed to support this?), and growth (the learner changes but the technology does not).

Kafai (1999) adapted Learner-Centered Design for use with children by making them the actual designers. She believes it is necessary that child learners be involved in the evaluation and testing processes. Her research showed that young student designers are similar to professional designers in their concern for their users. They were conscious of, and tried to address such issues as content and user motivation, but they did not always fully grasp how to address their users' other needs. Kafai is convinced, however, that children have the ability to become more than just informants in the design; rather, that they can become design process participants.

The premise behind Participatory Design is that users are the best qualified to determine how to improve their work, and that their perceptions about technology are as important as technical specifications (Carmel, Whitaker, & George, 1993). Two themes govern the implementation of Participatory Design principles: through "mutual reciprocal learning," users and designers teach each other about work practices and technical possibilities based on joint experiences; in "design by doing," interactive experimentation, modeling and testing, hands-on designing and learning by doing are employed. Like Contextual Design, Participatory Design is suitable for design projects involving children, where their school or home can substitute for the adult workplace. Its main difference from User-Centered Design, Contextual Design, and Learner-Centered Design is that the role assigned to children can be more extensive.

Informant Design was introduced specifically to address some of the perceived problems with User-Centered Design and Participatory Design when working with children (Scaife et al., 1997). In User-Centered Design, users are involved only as evaluators or testers at the end of the design process, and it is left to the designers to translate and interpret users' reactions, which can sometimes give inaccurate results. Scaife and his colleagues were critical of Participatory Design as a methodology to employ when working with children because it promotes the equality of all design team members. They considered this approach effective for a team comprising adult users who can see each other as peers, but infeasible with children who they believe have neither the time, knowledge, nor expertise to fully participate in a collaborative Participatory Design methodology. Informant Design attempts to maximize the input of the participants at various stages of the design process. Informants can help the designers "discover what we did not know rather than try to confirm what we thought we already knew" (Scaife & Rogers, 1999, p. 31). In Informant Design, each informant shapes the design at different points. Scaife and his colleagues believe Informant Design to be the best method "for the design of interactive software for non-typical users or those who cannot be equal partners (e.g., children)" (Scaife et al., 1997, p. 346). At the same time, there is a basic assumption that in the design process, children are most helpful at suggesting ideas only for motivational and fun aspects of the design rather than its totality.

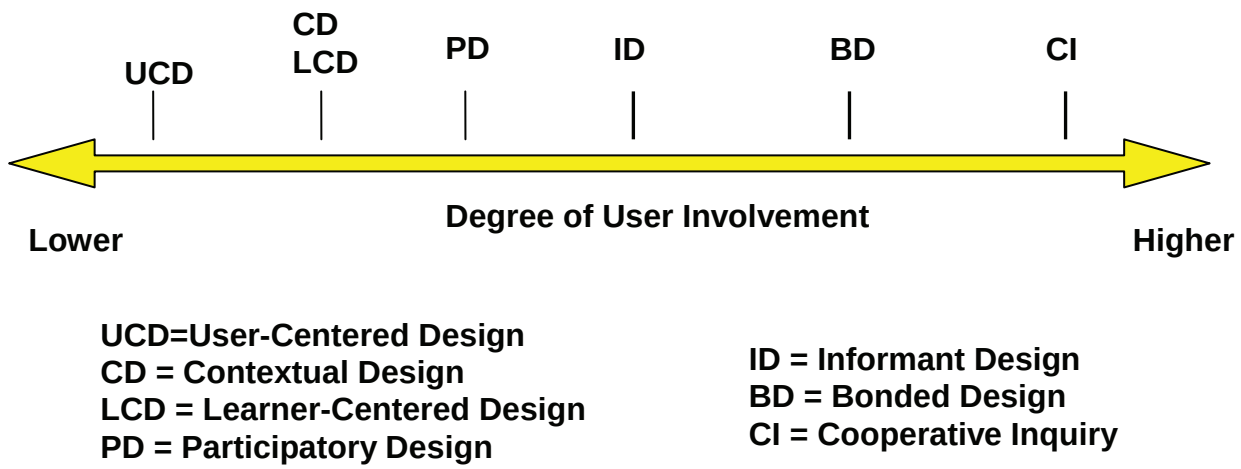
Cooperative Inquiry combines techniques from different design methodologies that have proven useful when working with children. Developed by Druin (1999) and her colleagues at the University of Maryland, it involves a multidisciplinary partnership with children, field research, and iterative low-tech and high-tech prototyping. Children are treated as full design partners alongside the adult designers on the intergenerational team. Professional designers and users (children) of the technology are partnered in intergenerational design teams with the understanding that full participation of users requires training and active cooperation. The design team makes use of such Contextual Inquiry methods as brainstorming and interviewing, as well as working together in small groups and developing low-tech prototypes (Druin, 2002; Guha et al., 2004). Using Cooperative Inquiry, Druin and her colleagues (Druin, 2002, 2005; Druin et al., 2003) have designed the International Children's Digital Library (*http://www.icdlbooks.org*).

BONDED DESIGN

Bonded Design is the newest addition to this family of technology design methodologies. From conventional User-Centered Design, it takes the most basic premise—involving users. From Contextual Design have come the ideas of drawing paper prototypes, and a similar process to what it terms *work redesign* in the use of a whiteboard to set out a map at the beginning of each session for what had already been accomplished and what remained to be done. Participatory Design provides the concept of peer co-designers, drawings (low-tech prototyping), hands-on activities, and "learning by doing." It shares with Informant Design the approach of seeking new and creative ideas rather than merely confirming what the adults already knew. Bonded Design also includes aspects of Learner-Centered Design in that it provides a learning environment for all team members: children and adults alike. In designing Web portals for children, as in Learner-Centered Design, all team members are learners, and the team's objective is to ensure that the design is adapted to the interests, knowledge, and styles of its target (child) users.

Of all the design methodologies, Cooperative Inquiry is the closest to Bonded Design. Both emphasize an intergenerational partnership to achieve a common goal, and embrace the idea that children should play an active role in the design process from start to finish rather than merely being evaluators or testers at the end of the design process. These two methodologies differ, however, in the emphasis placed by Bonded Design on a very focused approach that seeks to complete a highly specified task in a limited number

Figure 1. User involvement continuum



of design sessions extending only over a few weeks. Essentially Bonded Design is situated on a user-focused design methodology continuum between Cooperative Inquiry and Informant Design (see Figure 1). It shares the former's belief in the ability of children to work as partners in all aspects of the design process, but has reservations about the extent to which full and equal cooperation can occur across the generational divide, and in these respects, therefore, has similarities with the latter.

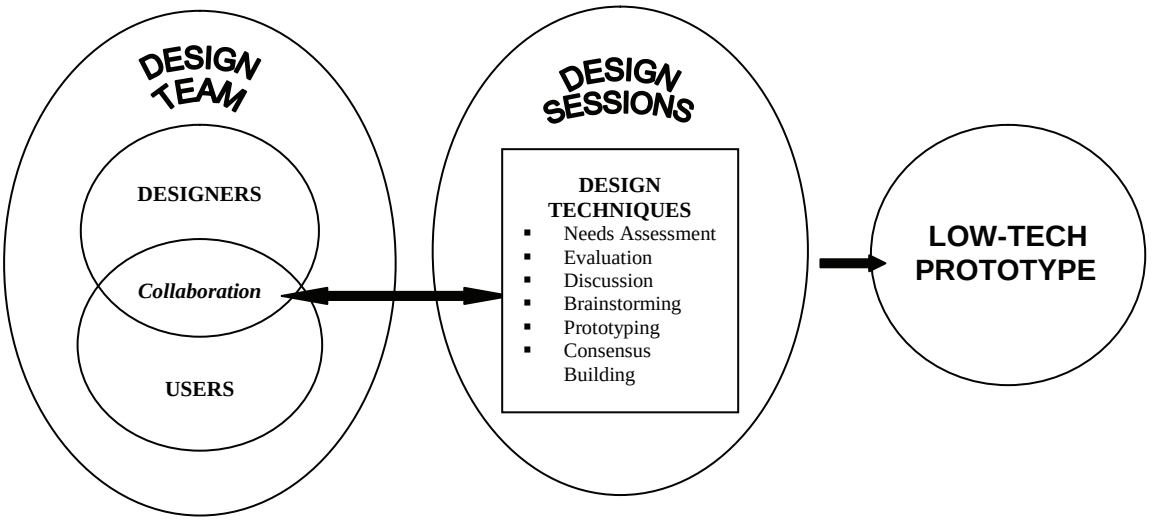
The Bonded Design methodology is graphically represented in Figure 2. The design team comprises two distinct groups: designers and users. It is assumed that the designers have a familiarity with the relevant technological environment as well as the design process. Yet it is fallacious, if common, to believe that these two attributes allow adults to think like children and see the world through youthful eyes. If designers want to be confident that their product will meet the expectations and requirements of children, they better understand what these are; the best way to do this, Bonded Design asserts, is to include children along with designers as collaborators throughout the design process.

Since collaboration is integral to the Bonded Design methodology, it is important to facilitate interaction among the team members. When working with children this can present unique challenges that can be met in a number of ways: a casual environment where the team sits around one large table, name badges (first name only), and respect for and acceptance that each team member's (adult and child) contribution is worthy of consideration. In other words, the team should promote a collaborative environment rather than a traditional classroom setting with its teacher-student relationships. Yet, even with these in place, it cannot be assured that everyone will feel at ease in participating. It is the responsibility of the adults to ensure that the sessions are not dominated by one or two voices and that the more reticent are actively encouraged to participate. If the design team is to work effectively and produce a low-tech prototype after a limited number of design sessions, it is also important to restrict the size of the team. In Bonded Design it is recommended that the team include between six and ten members (children and adults) to facilitate consensus building while providing a variety of ideas.

Any design is intended ultimately to be used. An important preliminary step in the design process is to ascertain the needs which the design is intended to meet for any given user community. A needs assessment, where potential users are polled to elicit how and why they might employ the completed product, is an effective tool to achieve this objective. This holds true even when the users happen to be children. One way to undertake a needs assessment is to survey a user sample by questionnaire. As the users will be children, it makes for good practice to involve the team's children in carrying out this assessment by administering the questionnaire to their peers.

Evaluating any available examples of the intended technology is a critical aspect that can take place throughout the design process. This evaluation may draw upon team members' prior knowledge of the product or upon

Figure 2. Model of bonded design



examination of examples within the design sessions. Any evaluation should be critical, and team members should be encouraged to identify strengths and weaknesses that can in turn inform their own preliminary designs. For example, the team that used Bonded Design to produce a prototype children's Web portal was able to critically assess a variety of existing exemplars available on the Web.

A free exchange of ideas lies at the essence of Bonded Design, but this is also where the interaction between the adult designers and the children within the team can pose the greatest challenge. The designers must be willing to accept the ideas put forward by the children even if they have reservations about their efficacy and feasibility of implementation. In the same vein, the children should be willing to draw upon the expertise that the designers inevitably bring to the task. At the heart of Bonded Design lies the belief that the child users have things to tell the adult designers that the latter cannot grasp themselves. Equally, Bonded Design is posited on the fact that the children by themselves do not have the necessary knowledge to design independently. It is the very bonding of ideas from these two groups that constitutes the strength of this design methodology.

Brainstorming is an activity that promotes creativity by encouraging all team members to contribute ideas on a topic. At this stage all ideas are accepted as having merit and are documented for later discussion. These ideas, however, are not always expressed verbally; for example, ideas often can be expressed very effectively through drawings, and this technique enables children to present interesting ideas without the constraints that a written or verbal representation might entail for young people. Furthermore, technology designs will normally be visual and therefore lend themselves to visual expression.

Prototyping is a technique shared by many design methodologies (Beyer & Holtzblatt, 1999; Carmel et al., 1993; Druin, 1999, 2002, 2005; Scaife, et al., 1997). It forms the bridge between discussion and brainstorming on the one hand, and the completed prototype design on the other. It can take various forms, but the most popular in participatory design methodologies is that of low-tech prototyping where participants use paper, modeling clay, or other such materials to represent design ideas. In Bonded Design, prototyping is used iteratively throughout the sessions in order to produce a final low-tech prototype. Bonded Design is particularly appropriate when working with children because they enjoy these types of prototyping activities and very successfully accomplish them. Furthermore, as commented above, children often can express their ideas more cogently through such activities rather than attempting textual descriptions (Arizpe & Styles, 2003; Glynn, 1997).

In any team environment where individuals are required to work together to reach a common goal, consensus building must take place. In the early stages of Bonded Design, team members are encouraged to think independently and to

formulate and express their own design ideas. Brainstorming is an effective way to generate a rich pool of ideas, but at some point these disparate ideas must coalesce into a unified design as a prelude to the completion of the low-tech prototype. Building such a consensus can be especially challenging when children are involved as they tend to an egocentric view of the world; a willingness to embrace alternative viewpoints comes only with maturity (Piaget & Inhelder, 1969). Before brainstorming begins it is important to establish evaluation criteria. One way to achieve this is through the initial user needs assessment, as it can identify design objectives and serve as a valuable basis on which to construct consensus. After brainstorming, when trying to reach consensus the team must determine which options best match the evaluation criteria. By matching options to pre-determined evaluation criteria, there is less chance that the opinions of one or two people will dominate.

The final step in the Bonded Design process is the development of a low-tech prototype. This can take a variety of forms, largely depending upon the time and resources available. For example, such activities as drawing on paper, modeling with clay, or computer simulations can be used to represent a final design.

FUTURE TRENDS

Bonded Design emerged as a methodology that can be used to design information technologies for children by involving children actively in all aspects of the design process. In the future there is no reason to think that it will not be applied to accomplish other design tasks where children can play an integral role, such as the creation of a student newspaper or the planning of a children's recreation area. Furthermore, although developed to involve children alongside adults in the design process, its emphasis on collaboration between design team members makes it a useful tool for cooperative learning within a classroom environment, where the emphasis may be more on the process than the outcome.

CONCLUSION

Bonded Design represents an addition to the array of user-centered design approaches. It is a means of bringing together for design purposes a team that unites in diversity. Adult design experts collaborate with child users who are experts in being children. Bonded Design is a method that involves users intimately in the design process. It does not simply ask users to test and respond to prototypes presented by professional designers (although it does turn to users for evaluations of the completed prototypes), but rather it incorporates members of the target user community into the decision making that lies behind the completion of these prototypes. In this way users do not simply react to designs that are presented to them, they help to create these designs. Such an approach provides an opportunity to accelerate the design process; instead of designing and testing over multiple iterations, as would be the approach in Participatory Design, Bonded Design over a limited number of design sessions can arrive at a low-tech prototype that has "bonded" the designers' professional expertise with the users' expertise in being users to get the best out of both constituencies. As such it is particularly appropriate when designing technologies for children. In order for designers to create for children, it is essential that children themselves be consulted, and Bonded Design is an ideal way to achieve this.

REFERENCES

Arizpe, E., & Styles, M. (2003). *Children reading pictures: Interpreting visual texts*. London: Routledge Falmer.

Beyer, H., & Holtzblatt, K. (1999). Contextual design. *Interactions, 6,* 32-42.

Carmel, E., Whitaker, R., & George, J. (1993). PD and joint application design: A transatlantic comparison. *Communications of the ACM, 36*(4), 40-48.

Druin, A. (1999). Cooperative inquiry: Developing new technologies for children with children. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 592-599). New York: ACM.

Druin, A. (2002). The role of children in the design of new technology. *Behaviour and Information Technology, 21*(1), 1-25.

Druin, A. (2005). What children can teach us: Developing digital libraries for children with children. *Library Quarterly, 75*(1), 20-41.

Druin, A., Bederson, B.B., Weeks, A., Farber, A., Grosjean, J., Guha, M.L., Hourcade, J.P., Lee, J., Liao, S., Reuter, K., Rose, A., Takayama, Y., & Zhang, L. (2003). The International Children's Digital Library: Description and analysis of first use. *First Monday, 8*(5). Retrieved from http://www.firstmonday.org/issues/issue8_5/druin/index.html

Glynn, S.M. (1997). Drawing mental models. *The Science Teacher, 64*(1), 30-32.

Guha, M.L., Druin, A., Chipman, G., Fails, J.A., Simms, S., & Farber, A. (2004). Mixing ideas: A new technique for working with young children as design partners. In A. Druin, J.P. Hourcade, & S. Kollet (Eds.), *Proceedings of Interaction Design and Children 2004: Building a Community* (pp. 35-42). New York: ACM.

Kafai, Y.B. (1999). Children as designers, testers, and evaluators of educational software. In A. Druin (Ed.), *The design of children's technology* (pp. 123-145). San Francisco: Morgan Kaufmann.

Large, A., Beheshti, J., Nesset, V., & Bowler, L. (2004). Designing Web portals in intergenerational teams: Two prototype portals for elementary school students. *Journal of the American Society for Information Science and Technology, 55*(13), 1140-1154.

Large, A., Nesset, V., Beheshti, J., & Bowler, L. (2006). 'Bonded Design': A novel approach to the design of new technologies. *Library and Information Science Research, 28,* 64-82.

Large, A., Nesset, V., Beheshti, J., & Bowler, L. (2007). Bonded Design: A methodology for designing with children. In S. Kurniawan & P. Zaphiris (Eds.), *Advances in universal Web design and evaluation: Research, trends and opportunities* (pp. 73-96). Hershey, PA: Idea Group.

Nesset, V., & Large, A. (2004). Children in the information technology design process: A review of theories and their applications. *Library and Information Science Research, 26*(2), 140-161.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child.* New York: Basic Books.

Rubin, J. (1994). *Handbook of usability testing: How to plan, design and conduct effective tests.* New York: John Wiley & Sons.

Scaife, M., & Rogers, Y. (1999). Kids as informants: Telling us what we didn't know or confirming what we knew already. In A. Druin (Ed.), *The design of children's technology* (pp. 27-50). San Francisco: Morgan Kaufmann.

Scaife, M., Rogers, Y., Aldrich, F., & Davies, M. (1997). Designing for or designing with? Informant design for interactive learning environments. *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 343-350).

Soloway, E., Guzdial, M., & Hay, K. (1994). Learner-centered design: The challenge for HCI in the 21st century. *Interactions, 1*(2), 36-48.

KEY TERMS

Bonded Design: A design methodology in which children play an active part in the design process alongside adult designers in an intergenerational team to accomplish a specific objective over a limited number of planned sessions.

Children: Young people, typically aged 12 years and younger.

Cooperative Inquiry: A design methodology in which children and adults cooperate as equals within an intergenerational team throughout the design process.

Design Methodology: A set of techniques normally involving users as well as designers to be employed in creating new technologies.

Informant Design: A design methodology in which children and adults are invited to inform the design throughout different stages of the design process.

Participatory Design: A design methodology in which users participate actively in the design process.

Prototyping: The process of creating an initial design.

Bridging the Digital Divide in Scotland

Anna Malina
e-Society Research, UK

INTRODUCTION

Perceptions of the different meanings and issues surrounding the term *digital divide* have set the scene for policy development in various countries. In recent times, broader analysis of the meanings and problems have altered understanding, and a new range of initiatives to tackle perceived problems is being devised in the United Kingdom (UK) and its regions. In what follows, digital divide perspectives are outlined and action to close the divide in Scotland is discussed.

BACKGROUND

For some time now, the Information Society vision in many countries has been accompanied by knowledge of risk of exclusion and strategies to close the so-called "Digital Divide," often seen as a short-hand term to indicate significant inequalities in access across social groups, and in particular between those who have access to ICTs (i.e., the "haves" and the "have-nots" or those who do not have access) (Civille, 1995; Raab, 1996). EU directives (e.g., eEurope 2002 [2000] and eEurope 2005 [2002]) support the goal of cheaper and more widespread access. The 2002 goal is to achieve physical access to the Internet, and the next stage is to consider content to stimulate access. The hope is that the benefits that emerge as a result of access to modern information and communication technologies (ICTs) will be felt by regional and local economies and communities.

Extending access to ICTs will help ensure innovation, economic development, and the new economy (Tambini, 2000a). In discussing universal Internet access, Tambini (2000b) points to arguments that without widespread access, e-commerce would not be able to support greater innovation and entrepreneurship. In addition, the intense concentration on developing e-government could not be legitimised. Moreover, benefits said to be associated with the design of ICTs to improve efficiency, effectiveness, and transparency in public service delivery could not be realised. Norris (2001) and others note the views of pessimists who fear an escalation of existing inequalities, and optimists who hold that new ICTs have the potential to widen opportunities for more democratic participation. However, without universal Internet access, it is unlikely that wider forms of electronic participation and actions associated with e-governance and e-democracy could be supported. Additionally, distance learning and public education resources would not reach wider audiences or increase literacy levels.

BRIDGING THE DIGITAL DIVIDE IN SCOTLAND

The Scottish Household Survey shows that access to the Internet in Scotland is growing quickly. People who are excluded comprise the unemployed, those with low incomes, low levels of education, and poor literacy and numeracy levels. The Scottish Executive, the Scottish Parliament, the voluntary sector, and other organizations in Scotland have designed a range of initiatives to tackle problems associated with the digital divide. The Scottish framework is based on raising awareness, widening access, increasing skills, building support, developing content, and motivating and involving communities.

Scottish Executive and Scottish Parliament

Similar to the UK Government, the Scottish Executive (i.e., the devolved government of Scotland) is committed to achieving universal access to the Internet by 2005. The Scottish Executive initiative—Digital Scotland—set out to ensure that Scotland obtains and retains maximum economic and social advantage from the development of ICTs. Digital divide problems are associated with a lack of telecommunications infrastructure and with poverty, lack of awareness, and low skill levels (Digital Inclusion: Connecting Scotland's People, 2001). Emphasis has been placed on expanding Scotland's communication infrastructure to stimulate demand for broadband and to test innovative delivery technologies. A three-year Digital Champions programme[1] was set up to improve inclusive ICT provision in Social Inclusion Partnerships (SIPs) in deprived areas of Scotland. This project provides community professionals with local knowledge to engage local people in various initiatives in order to drive local ICT activities forward.

In 2002, a £3.2 million initiative was launched by the Social Justice Minister to create a network of 1,000 new Internet access points in areas where current public provision was still poor.[2] The Scottish Executive also promotes awareness

of existing public access to the Web, and provides an online service to help people find the nearest access point.[3]

The Scottish Parliament's Web site[4] provides extensive information online. To help address problems of the digital divide, the Parliament worked with the public library sector to establish a network of 80 partner libraries throughout Scotland, many of which now provide public access to the Internet through freely available terminals.

With a key focus on the citizen, government portals across the UK are offering services that they suggest are relevant to life episodes. Closely linked to the UK Government Portal (www.ukonline.gov.uk) is the Scottish Government Portal that operates under the brand *Open Scotland*[5] to promote choice and public take-up of services.

Voluntary Sector Initiatives

In August 1999, com.com/holyrood, a public-private partnership between the Scottish Centre for Voluntary Organisations (SCVO) and British Telecom (BT), was given the task of installing 200 PCs into local halls and community centres throughout Scotland. However, access alone was not considered enough to address the digital divide, and SCVO also began to develop voluntary sector content. In June 2002, a Web portal[6] was launched to act as a single gateway to Scotland's voluntary sector. In addition, a lobby channel allowed voluntary organisations to conduct their own e-consultations. Moreover, online questions could be forwarded to Members of the Scottish Parliament (MSP).

Education and Learning Initiatives

The National Grid for Learning (NGfL) Scotland was set up by the Scottish Executive Education Department in September 1999 to connect all schools, colleges, universities, and libraries in Scotland to the Internet by 2002. A key objective of the NGfL Scotland Communities team[7] is to use ICT to improve opportunity, access, and quality of life for excluded groups, and to actively involve communities in worthwhile local projects. The communities channel of NGfL Scotland aims to supply information, advice, and assistance to all those providing support for ICTs in their community. In addition, NGfL Scotland's Connecting Communities Training Programme[8] promotes the effective use of ICT in community learning agencies across Scotland.

The Scottish University for Industry (SUfI) was set up to promote public/private partnership, commission research, draw on other analyses, and investigate the needs of market and client groups in Scotland. SUfI built on partnerships already existing in Scotland and worked closely with Highlands and Islands Enterprise and other partners to develop skills in using ICT. The subsequent development of a variety of IT centres in different locations of Scotland has provided Internet access and a learning environment for people to meet, to learn about new ICT, and to achieve new skills.

Scottish radio and television broadcasts were organised in late 1999 to promote learning directly and supporting the BBC's Webwise campaign, building on an earlier programme entitled Computers Don't Bite. Drawing from £200 million pounds allocated to support Community Access to Lifelong Learning (CALL) across the UK, the new Opportunities Fund (NOF) in Scotland was allocated £23 million pounds to support LearnDirect, an organisation providing public information about local learning centres.

Scotland is aiming to close the digital divide and to encourage people in deprived areas to acquire the key IT skills that are suitable to the demands of an Information Society.

Wired Communities in Scotland

During 2001, in an attempt to promote digital inclusion, a three million pound initiative outlined the intention to create two pilot digital communities in Scotland. The document, titled Digital Inclusion: Connecting Scotland's People (*Scottish Executive*, 2001, p. 22) outlines intention and funding for:

- the provision of entry level PCs, software, and Web access to up to 2,000 homes in each community;
- the development of links with school-based ICT and Web access initiative;
- the development of a community Web portal for each community with local content relevant to that community, including relevant online public and commercial services;
- the provision of training to increase the level of ICT and Web skills;
- the promotion to raise awareness of the benefits of the Web; and
- the creation of a network of local people to provide ongoing support.

The same report also outlines the aims of the digital communities Scotland project as follows:

- to use ICTs to help tackle social exclusion in these communities;
- to create a "critical mass" of Web users in each community;
- to contribute to achieving universal access to the Web;
- to increase the take-up of computers and the Web in disadvantaged households;
- to increase ICT skills in disadvantaged communities;
- to increase community involvement to develop local online content and a local support network; and

- to create partnerships with the private sector.

In late 2001, communities across Scotland were invited to submit bids to the Scottish Executive to receive funding and support to become digital communities. Towards the end of March 2002, the Scottish Executive selected two winning communities—one rural and one urban.

The Argyll Islands, the rural community selected, covers 13 of the 26 inhabited islands in the archipelago. A total of 2,145 households from a population of 4,290 was included. The digital communities submission for funding suggests that while the islands "are rich in biodiversity, culture, history, archaeology, renewable energy opportunities, and landscape, they suffer from many issues of deprivation and disadvantage, both caused and accentuated by their geographical isolation" . Physical and demographic characteristics of the islands combine to produce economic and social conditions constraining economic activity and growth. As such, the islands of Argyll are categorised as "fragile areas" by Highlands and Islands Enterprise and as "disadvantaged areas" by the Council's Economic Development Strategy (Argyll and Bute Council First Round Digital Communities Submission, 2002).

All local people in the Argyll Islands are expected to profit from new opportunities to use computers and access the Internet. The aim is to make it easier to communicate speedily and cheaply in regard to public services, to interact with peers, to telework, and to access health services and lifelong learning materials. Wherever possible, the intention is to include the business community of these islands as well as community groups. While some members of the community already have computer skills, the initiative is expected to extend training and support to any of those households that currently have computers but lack sufficient knowledge and training to use them to full advantage. In addition, the intention is to build e-services into a comprehensive community portal.

Bellsmyre, situated to the north of the town of Dumbarton, is the urban digital community selected for funding. Bellsmyre has 789 households and a population of 1,694 people. No telecommunications company had cabled the community, and at the time of the bid, Bellsmyre only had an analogue telephone infrastructure. The area is described in the First Round Proposal document as the most deprived within West Dunbartonshire.

At the time, people living in Bellsmyre had not yet responded to previous encouragement to train and use technology and, as a result, were considered in danger of being excluded from society even further. Technology is being taken to the people to promote the advantages of learning. Drawing from partnership arrangements, the key aim outlined in Bellsmyre's First Round Proposal is to help achieve the objectives set out in Scotland's digital inclusion strategy. It is envisaged that delivery of training and ICT services by Digital Champions and local community groups and agencies will help develop trust. It is hoped that attitudes toward ICT will improve, fear of technology will subside, confidence will improve, and the benefits of being connected to the Web will emerge. It is also hoped that attitudes toward learning will change and become more positive. A central aim is to ensure that levels of educational achievement will rise locally. Key aspirations are to reduce digital exclusion, create an ICT-skilled local population that will develop a routine taste for further learning and training and go on to employment or start up new businesses.

E-Democracy and the Digital Divide in Scotland

The UK government has established a Cabinet Committee to oversee e-democracy in the UK, and to develop a strategy for its development and rollout. An e-democracy charter divides the e-democracy domain into two clear fields: e-voting and e-participation. It is possible to design technology specifically to support local authorities by taking a creative, flexible, and democratic approach at local levels (Sisk, 2001). However, the ways in which technology might be developed and implemented to support better democratic participation in local areas have not been researched previously in Scotland. An added difficulty is that because each local authority is charged with representing multiple communities with diverse needs, the local authorities may need to use different e-democracy processes in different ways in different communities. This highlights a gap in knowledge about the numerous ways e-democracy systems could be tailored to suit the democratic needs of dissimilar communities. In addition, there is a matter of ensuring authentic participation of citizens in ways in which they are willing to be involved and that reflect their views and experiences about issues they feel are relevant in their own communities and beyond.

FUTURE TRENDS

A New Zealand report, "The Digital Divide – Examining the Main Characteristics that Influence Household Internet Connection in New Zealand" (2004), reflects international research in suggesting that the expansion of information and communication technologies is mainly utilised by households with higher incomes and by households whose members have formal educational qualifications. The report also found that "although the age of the youngest occupant, ethnicity, labour force status and geographic location played important roles in determining household Internet access, the most important variables identified as influencing household connectivity levels were household income, the level of educational qualification and household composition" (Chapter 3 Conclusion).

Research conducted by the Greater London Authority to outline a strategy to address the digital divide in London shows that while socioeconomic factors such as low income, low levels of education, low-skilled jobs, unemployment, and lack of technology skills are a barrier to the adoption and use of ICTs, sociopersonal aspects such as low levels of awareness, interest, understanding, and acceptance of ICTs are also very important (Foley et al., 2002).

A great deal of attention and policy and resource support over recent years has helped to ensure that excluded groups and communities living outside mainstream society are provided with different kinds of access to contemporary ICTs. However, Wilhelm (2000) questions whether or not the information under class can now be wholly defined in terms of access. Instead, he emphasizes broader contexts of information-seeking behaviour (i.e., media use patterns and cultural and environmental contexts). Drawing from Schon, Sanyal, and Mitchell (1999), he argues that these contexts will "provide a thicker description of the various shades of information and communications inequalities" (p. 70). Hague and Loader (1999) argue that experiences of using ICTs will vary enormously, since people do not generally share the same traits and since members of the public are very different in terms of gender, race, disability, class, location, and religion.

Others suggest that digital divide arguments should move away from issues of access alone. For example, Warschauer (2003) argues that increased emphasis should be put on action to ensure that people develop more meaningful use of technology within their own social arenas and daily practices.

CONCLUSION

Consideration of the Scottish context suggests that commitment to develop locally based ICT infrastructure and innovative new initiatives is fundamental in addressing problems associated with the digital divide. Critical evaluations of these initiatives would be invaluable, and it is crucial now for long-term research programmes to broaden understanding of meaningful routine use. Some researchers underline the importance of gathering empirical data relating to how and why people use ICTs in different settings (Malina & MacIntosh, 2003). Action research is suggested to assess the democratic requirements of those living in different kinds of communities, and to enable more meaningful design, development, and continuous assessment of ICT-based systems underpinning new democratic practices. It is suggested that this research approach could be applied initially to the two "digital communities" in Scotland—Argyle & Bute and Bellsmyre. In seeking to consult citizens, we must listen to their democratic needs and consider their perceptions. Research would help to improve the quality of citizenship and the level of democratic participation in local communities. The research work would also provide a framework to better appreciate the significance of technology in supporting e-democracy at local community levels and, in so doing, to contribute knowledge to strategy and planning policies and social and digital inclusion agendas to address problems of the digital divide in Scotland.

REFERENCES

Argyll and Bute Council first round digital communities submission (2002). Argyll & Bute Council.

Bellsmyre, West Dunbartonshire first round digital communities proposal (2002). West Dunbartonshire Council.

Civille, R. (1995). The Internet and the poor. In B. Kahlin and J. Keller (Eds.), *Public access to the Internet*. Cambridge MA: MIT Press.

The digital divide – Examining the main characteristics that influence household Internet connection in New Zealand (2004, March 5). Statistics New Zealand (Te Tari Tatau). Retrieved from *http://www.stats.govt.nz/domino/external/pasfull/pasfull.nsf/web/Reference+Reports+The+Digital+Divide+2004?open*

eEurope2002 (2000). *http://www.europa.eu.int/information_society/eeurope/action_plan/pdf/actionplan_en.pdf*

eEurope2005*:* An information society for all (2002, June 28). *Retrieved from http://europa.eu.int/information_society/eeurope/news_library/eeurope2005/index_en.htm*

Foley, P., Alfonso, X., & Ghani, S. (2002). *The digital divide in a world city*. London: Greater London Authority.

Hague, B., & Loader, B. (Eds.) (1999). *Digital democracy: Discourse and decision making in the information age*. London: Routledge.

Malina, A., & Macintosh, A. (2003). Bridging the digital divide. Development in Scotland. In A. Anttiroiko, M. Mälkiä , & R. Savolainen (Eds.), *eTransformation in governance: New directions in government and politics.* Hershey, PA: Idea Group Publishing.

Norris, P. (2001). *Digital divide: Civic engagement information poverty, and the Internet worldwide.* Cambridge, MA: Cambridge University Press.

Raab, C., Bellamy, C., Taylor, J., Dutton, W.H., & Peltu, M. (1996). The information polity: Electronic democracy, privacy, and surveillance. In W. Dutton (Ed.), *Information and communication technologies: Vision and realities*. Oxford: Oxford University Press.

Schon, D., Sanjal, B., & Mitchell, W.J. (Eds.) (1999). *High technology and low-income communities: Prospects for the positive use of advanced information technology*. Cambridge, MA: MIT Press.

Scottish Executive (2001). Digital inclusion: Connecting Scotland's people. Retrieved from *http://www.scotland.gov.uk/library3/enterprise/dics-00.asp*

Scottish Executive (2002). Scottish household survey bulletin no. 7: Life cycles. Retrieved from *www.scotland.gov.uk/library3/housing/shs7-09.asp*

Sisk, T. (2001). *Democracy at the local level: The international IDEA handbook on participation, representation, conflict management, and governance.* Hershey, PA: Idea Group Publishing.

Tambini, D. (2000a). *Digital danger.* London: IPPR.

Tambini, D. (2000b). *Universal Internet access: A realistic view.* London: IPPR.

Warschauer, M. (2003). *Technology & social inclusion.* Cambridge, MA: MIT Press.

Wilhem, Anthony (2000). *Democracy in the digital age: Challenges to political life in cyberspace.* New York and London: Routledge.

KEY TERMS

Digital Divide: Refers to individuals or members of communities and groups whose social, cultural, political, economic, or personal circumstances constrain access to electronic communications or limit benefit to their lives from contemporary electronic technologies.

Digital Inclusion: Strategies and actions to assure more equal access to digital technologies and Web facilities and to strengthen effective, meaningful, and beneficial use for all members of the public in their day-to-day lives.

Distance Learning: Learners are connected with educational resources beyond the confines of a traditional classroom, and instructed via computer-mediated communication and different types of electronic technologies that can overcome the constraints of distance, time, physical presence, or location that separate instructors and students. Learning may be synchronous or asynchronous.

E-Commerce: The buying and selling of commercial goods; the conduct of financial transactions using digital communications and electronic networks such as the World Wide Web; and aspects of sharing of business information, maintenance of business relationships, and provision of information services.

E-Democracy: The use of electronic communications to support and increase democratic engagement and deepen and widen citizen participation.

E-Government: The ability of government to design and use ICTs to interact internally and externally with government bodies, citizens, and businesses in order to deliver integrated electronic public services.

E-Governance: Communication by electronic means to place power in the hands of citizens to determine what laws need to be made and how these laws should be written.

Information and Communication Technologies (ICTs): While often meaning different things in different timescales, places, and contexts, ICTs describe all media and a mix of converging technology tools involved in the dynamic transfer and storage of analogue and digital data. In addition to Internet-based technologies such as computers, telephones, and networks, ICTs in a broad sense include digital television, cable and satellite technologies, and music formats (e.g., MP3), DVDs, and CDs. ICTs may be used to facilitate remote human interaction for good and evil purposes. In the context of this article, ICTs are used to increase human communication; broaden education, literacy, and knowledge; and enhance social, cultural, political, and economic capacity. It is hoped that this will help address problems attributed to the so-called digital divide.

ENDNOTES

1 http://www.scotland.gov.uk/library3/enterprise/dics-06.asp, access November 2002.
2 http://www.scotland.gov.uk/pages/news/extras/00007100.aspx, accessed November 2002.
3 http://www.scotland.gov.uk/digitalscotland/webaccess/default.asp, accessed November 2002.
4 www.scottish.parliament.uk, accessed November 2002.
5 http://www.openscotland.gov.uk/, accessed November 2003.
6 www.workwithus.org, retrieved November 2002.
7 http://www.ngflscotland.com/communities/, retrieved May 2002.
8 http://www.ngflscotland.com/communities/training/connectingcommunities/, retrieved May 2002.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 278-283, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Social Computing

A Brief Introduction to Sociotechnical Systems

Brian Whitworth
Massey University Auckland, New Zealand

INTRODUCTION

The term sociotechnical was introduced by the Tavistock Institute in the 1950's for manufacturing cases where the needs of technology confronted those of local communities, for example, longwall mining in English coalmines (see http://www.strategosinc.com/socio-technical.htm). Social needs were opposed to the reductionism of Taylorism, which broke down jobs on say a car assembly line into most efficient elements. Social and technical were seen as separate side-by-side systems which needed to interact positively, for example, a village near a nuclear plant is a social system (with social needs) besides a technical system (with technical needs). The sociotechnical view later developed into a call for ethical computer use by supporters like Mumford (Porra & Hirscheim, 2007).

In the modern holistic view the sociotechnical system (STS) is the whole system, not one of two side-by-side systems. To illustrate the contrast, consider a simple case: A pilot plus a plane are two side-by-side systems with different needs, one mechanical (plane) and one human (pilot). Human Computer Interaction (HCI) suggests these systems should interact positively to succeed. However plane plus pilot can also be seen as a single system, with human and mechanical levels. On the mechanical level, the pilot's body is as physical as the plane, for example, the body of the plane and the body of the pilot both have weight, volume, and so forth. However the pilot adds a human thought level that sits above the plane's mechanical level, allowing the "pilot + plane" system to strategize and analyze. The sociotechnical concept that will be developed changes the priorities, for example, if a social system sits next to a technical one it is usually secondary, and ethics an afterthought to mechanics, but when a social system sits above a technical one it guides the entire system, that is, the primary factor in system performance.

BACKGROUND

General Systems Theory

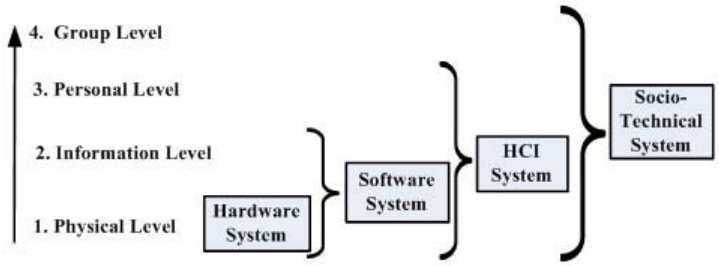
Sociotechnical theory is based upon general systems theory (Bertalanffy, 1968), which sees systems as composed of autonomous yet interdependent parts that mutually interact as part of a purposeful whole. Rather than reduce a system to its parts, systems theory explores emergent properties that arise through component interactions via the dynamics of regulation, including feedback and feed-forward loops. While self-reference and circular causality can give the snowball effects of chaos theory, such systems can self-organize and self-maintain (Maturana & Varela, 1998).

System Levels

In the 1950's-1960's computing was mainly about hardware, but the 1970's introduced the software era with business information processing. The 1980's then gave personal computers, adding people into the equation, and email in the 1990's introduced the computer as a social medium. In this decade social computing development continues, with chat rooms, bulletin boards, e-markets, social networks (e.g., UTube, Facebook, MySpace), Wikis and Blogs. Each decade computing has reinvented itself, going from hardware to software, from software to HCI, and now from HCI to social computing. The concept of system levels frames this progression. While Grudin initially postulated three levels of hardware, software and cognitive (Grudin, 1990), Kuutti later added an organizational level (Kuutti, 1996), suggesting an information system (IS) could have four levels: hardware, software, human, and organizational (Alter, 1999). Just as software "emerges" from hardware, so personal cognitions can be seen as arising from neural information exchanges (Whitworth, 2008), and a society can emerge from the interaction of individual people. If the first two levels (hardware/software) are together considered technical, and the last two (human/group) social, then a sociotechnical system is one that involves all four levels (Figure 1).

As computing evolved, the problems it faced changed. Early problems were mainly hardware issues, like over-heating. When these gave way to software problems like infinite loops, then network and database needs began to influence hardware chip development. A similar progression occurred as human factors emerged, and human requirements like usability became part of software engineering (Sanders & McCormick, 1993). Social computing continues this trend, as social problems beyond HCI theory (Whitworth, 2005) are now important in design. Driving this evolution is that each emergent level increases system performance (Whitworth & deMoor, 2003).

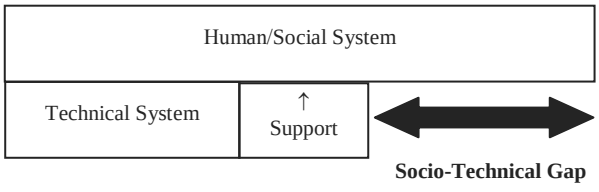
Figure 1. Sociotechnical system levels



The Sociotechnical Gap

The Figure1 levels are not different systems but overlapping views of the same system, corresponding to engineering, computing, psychological, and sociological perspectives respectively (Whitworth, Fjermestad, & Mahinda, 2006). Higher levels are both more efficient ways to describe a system and also more efficient ways to operate it, for example, social individuals produce more than they would acting independently. Whether a social system is electronically or physically mediated is arbitrary. That the communication medium (a computer network) is "virtual" does not make the people involved less real, for example, one can be as upset by an e-mail as by a letter. Sociotechnical systems are systems of people communicating with people that arise through interactions mediated by technology rather than the natural world. The social system can follow the same principles whether on a physical or electronic base, for example, friendships that cross seamlessly from face-to-face to e-mail interaction. However in physical society, physical architecture supports social norms, for example, you may not legally enter my house and I can also physically lock you out or call the police to restrain you. In contrast, while in cyberspace the "architecture" is the computer code itself, that "… makes cyberspace as it is" (Lessig, 2000), that code is largely designed without any reference to social needs. This sociotechnical gap (Ackerman, 2000), between what computers do and what society wants (Figure 2), is a major software problem today (Cooper, 1999).

Figure 2. Socio-technical gap



STS REQUIREMENTS

System Theory Requirements

A systems approach to performance suggests that systems can adapt the four elements of a boundary, internal structure, effectors, and receptors to either increase gains or reduce losses (Whitworth et al., 2006), where:

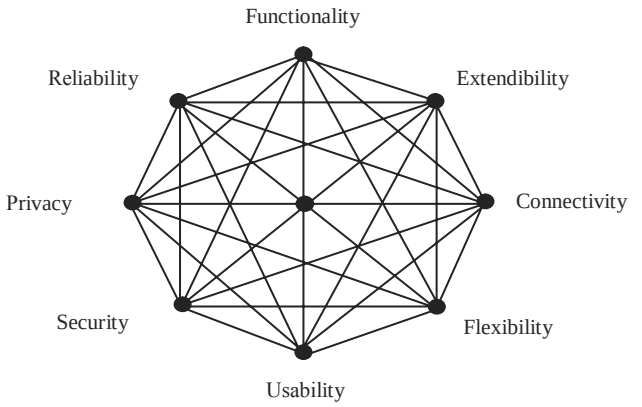
1. The system boundary controls entry, and can be designed to deny unwelcome entry (security), or to use the entity as a "tool" (extendibility).
2. The system internal structure manages system operations, and can be designed to reduce internal changes that cause faults (reliability), or to increase internal changes allowing environment adaptation (flexibility).
3. The system effectors use system resources to act upon the environment, and can be designed to maximize their effects (functionality), or to minimize the relative resource "cost of action" (usability).
4. The system receptors open channels to communicate with other systems, and can enable communication with similar systems (connectivity), or limit communication (privacy).

This gives the eight performance goals of Figure 3, where:

- The Web area shows the overall system's performance potential.
- The Web shape shows the system's performance profile, for example, risky environments favor secure profiles.
- The Web lines show tensions between goals, for example, improving flexibility might reduce reliability.

The goals change for different system levels, for example, a system can be hardware reliable but software unreliable,

Figure 3. The Web of system performance



or both hardware and software reliable but operator unreliable (Sommerville, 2004, p. 24). Likewise usability (the relative cost of action) can mean less cognitive "effort" for a person in an HCI system, or for a software system mean less memory/processing (e.g., "light" background utilities), or for a hardware system mean less power use (e.g., mobile phones that last longer). From this perspective, the challenge of socio-technical computing is to design innovative systems that integrate the multiple requirements of system performance at higher and higher levels, where each level builds upon the previous.

Information Exchange Requirements

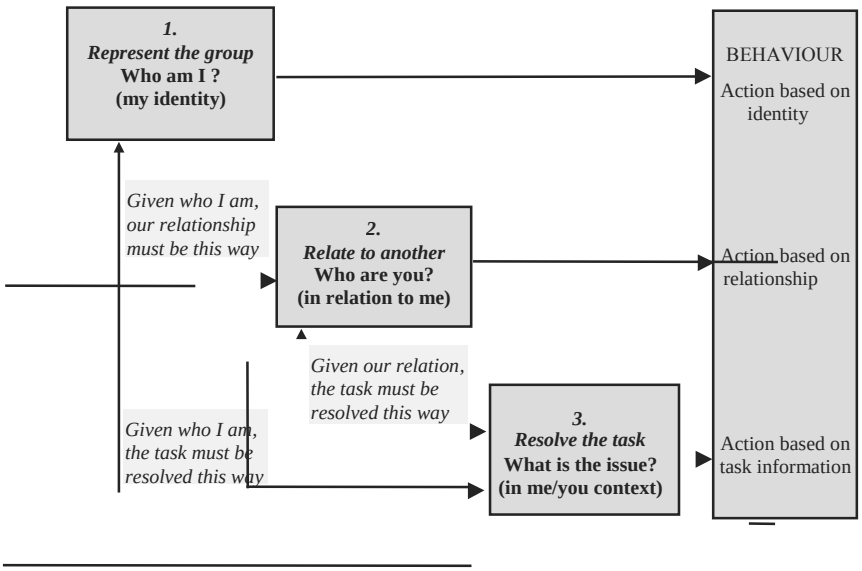
The connectivity-privacy tension line (Figure 3) introduces a social dimension to system design, as information exchanges let the systems we call people combine into larger systems, for example, people can form villages, villages can form states, and states can form nations. In this social evolution not only does "the system" evolve, but also what we define as "the system" evolves. Theories of computer-mediated information exchange postulate the underlying social process. Some consider it a single process of rational analysis (Huber, 1984; Winograd & Flores, 1986), but others suggest process dichotomies, like task vs. socioemotional (Bales, 1950), informational vs. normative (Deutsch & Gerard, 1965), task vs. social (Sproull & Kiesler, 1986), and social vs. interpersonal (Spears & Lea, 1992). A three process model of online communication (Whitworth, Gallupe, & McQueen, 2000) suggests three processes:

1. Factual information exchange: the exchange of factual data or information, that is, message content
2. Personal information exchange: the exchange of personal sender state information, that is, sender context
3. Group information exchange: the exchange of group normative information, that is, group position

The first process involves intellectual data gathering, the second involves building emotional relationships, and the third involves identifying with the group, as proposed by Social Identity theory (Hogg, 1990). The three goals are understanding, intimacy and group agreement, respectively. In this multi-threaded communication model one communication can contain many information threads (McGrath 1984), for example, if one says "I AM NOT UPSET!" in an upset voice, sender state information is analyzed in an emotional channel, while message content is analysed intellectually. A message with a factual content not only lies within a sender state context given by facial expressions etc, but also contains a core of implied action, for example, saying "This is good, lets buy it" gives not only content information (the item is good) and sender information (say tone of voice), but also the sender's intended action (to buy the item), that is, the sender's action "position".

While message content and sender context are generally recognized, action position is often overlooked as a communication channel, perhaps because it typically involves many-to-many information exchange, for example, in a choir singing everyone sends and everyone receives, so if the choir moves off key they do so together. Similarly, in apparently rational group discussions, the group "valence index" predicts the group decision (Hoffman & Maier, 1964). Group members seem to assess where the group is going and change their ideas to stay with the group. This intellectual equivalent of how social flocks/herds cohese seems to work equally well online, so using this process one can generate online agreement from anonymous and lean information exchanges (Whitworth, Gallupe & McQueen, 2001). While factual information suits one-way one-to-many information exchange (e.g., a Website broadcast), and developing personal relations suits two-way one-to-one information exchange (e.g., e-mail), the group normative process needs two-way, many-to-many information exchange (e.g., reputation systems). Figure 4 shows how people prioritize the three cognitive processes, first evaluating group membership

Figure 4. Three information exchange processes



identity, then evaluating the sender's relationship, and only finally analyzing message content. As each cognitive process favours a different communication architecture, there is no ideal human communication medium. From this perspective, the challenge of sociotechnical computing is to design innovative systems that integrate the multiple channels of human communication.

Social Requirements

A startling discovery of game theory was that positive social interaction can be unstable, for example, the "equilibrium point" of the prisoner's dilemma dyad is that both cheat each other (Poundstone, 1992). Situations like social loafing and the volunteer dilemma are common, including many-to-many cases like the tragedy of the commons which mirrors global conservation problems. In a society one person can gain at another's expense, for example, theft, yet if everyone steals society collapses into disorder. Human society has evolved various ways to make anti-social acts unprofitable, whether by personal revenge traditions, or by state justice systems. The goal of justice, it has been argued, is to reduce unfairness (Rawls, 2001), where unfairness is not just an unequal outcome distribution (inequity) but failure to distribute outcomes according to action contributions. In a successful society people are accountable not just for the effects of their acts on themselves but also on others. Without this, people can take the benefits that others create, and harm others with no consequence to themselves. Under these conditions, any society will fail. Fortunately people in general seem to recognize unfairness, and tend to avoid unfair situations (Adams, 1965). They even prefer fairness to personal benefit (Lind & Tyler, 1988). The capacity to perceive "natural justice" seems to underlie our ability to form prosperous societies. The general requirement is legitimate interaction, which is both fair to the parties involved and also benefits the social group (Whitworth & deMoor, 2003). Legitimacy is a complex sociological concept that describes governments that are justified to their people (and not coerced) (Barker, 1990). Fukuyama argues that legitimacy is a requirement for community prosperity, and those that ignore it do so at their peril (Fukuyama, 1992).

It follows that online society should be designed to support legitimate interaction and oppose antisocial acts. Yet defining what is and is not legitimate is a complex issue. Physical society evolved the concept of "rights" expressed in terms of ownership (Freeden, 1991), for example, freedom as the right to own onself. Likewise analyzing who owns what information online (Rose, 2001), can be used to specify equivalent online rights that designers can support (Whitworth, Aldo de Moor & Liu, 2006) (Figure 5). This does not mechanize online interaction, as rights are choices not obligations, for example, the right to privacy does not force one to be private. From this perspective, the challenge of sociotechnical computing is to design systems that reflect rights in overlapping social groups.

Figure 5. Legitimacy analysis

FUTURE TRENDS

The previous discussion suggests the design of future sociotechnical systems will involve:

1. More performance requirements: Simple dichotomies like usefulness and ease of use, while themselves valid, will be insufficient to describe the multiple criteria relevant to STS performance.
2. More communication requirements: Communication models with one or two cognitive processes while themselves valid, will be insufficient to describe the multiple threads of STS communication.
3. More social requirements: Approaches that reference only individual users, while themselves valid, will be insufficient to represent social level requirements, where one social group can contain another.

Performance concepts will expand, as success-creating goals (functionality, flexibility, extendibility and connectivity) and the failure-avoiding goals (security, reliability, privacy and usability) interact in complex ways in a multidimensional performance space that challenges designers. Likewise designing communication media that simultaneously support the flow of intellectual, emotional and positional information along parallel exchange channels is another significant challenge. Finally, in social evolution social systems "stack" one upon another (e.g., states can form into a federation), giving the challenge of groups within groups. In sum, the challenges are great, but then again why should STS design be easy?

The concept of levels (Figure 1) runs through the above trends, suggesting that the World Wide Web will develop cumulatively in three stages. The first stage, a factual information exchange system, seems already largely in place, with the World Wide Web essentially a huge information library accessed by search tools, though it contains disinformation as well as information. The second stage lets people form personal relations to distinguish trusted from not trusted information sources. This stage is now well underway, as social networks combine e-mail and browser systems into protocol independent users environments (Berners-Lee 2000). Finally, in stage three the Web will sustain synergistic and stable online communities, opening to group members the power of the group, for example, the group knowledge sharing of Wikipedia. To do this, online communities must both overcome antisocial forces like Spam and prevent internal take-overs by personal power seekers. In this struggle software cannot be "group blind" (McGrath & Hollingshead, 1993). Online communities cannot democratically elect new leaders to replace old ones unless software provides the tools. Supporting group computing is not just creating a few membership lists. Even a cursory study of Robert's Rules of Order will dispel the illusion that technical support for social groups is easy (Robert, 1993).

Human social activity is complex, as people like to belong and relate as well as understand. Each process reflects a practical human concern, namely dealing with world tasks, with other people, and with the society one is within. All are important, because sometimes what you know is critical, sometimes who you know counts, and sometimes, like which side of the road to drive on, all that counts is what everyone else is doing. Yet as we browse the Web software largely ignores social complexity. "Smart" tools like Mr. Clippy have relational "amnesia", as no matter how many times one tells them to go away they still come back. Essential to relationships is remembering past interactions, yet my Windows file browser cannot remember my last browsed directory and return me there, my Word processor cannot remember where my cursor was last time I opened this document and put me back there, and my browser cannot remember the last Website I browsed and restart from there (Whitworth, 2005). Group level concepts like rights, leadership, roles, and democracy are equally poorly represented in technical design.

These challenges suggest that the Internet is only just beginning its social future. We may be no more able to envisage this than traders in the Middle Ages could conceive today's global trade system, where people send millions of dollars to foreigners they have never seen for goods they have not touched to arrive at unknown times. To traders in the middle ages, this would have seemed not just technically but also socially impossible (Mandelbaum, 2002). The future of software will be more about social than technical design, as software will support what online societies need for success. If society believes in individual freedom, online avatars should belong to the person concerned. If society gives the right to not communicate (Warren & Brandeis, 1890) so should communication systems like email (Whitworth & Whitworth, 2004). If society supports privacy, people should be able to remove their personal data from online lists. If society gives creators rights to the fruits of their labors (Locke, 1963), one should be able to sign and own one's electronic work. If society believes in democracy, online communities should be able to elect their leaders. In the sociotechnical approach social principles should drive technical design.

CONCLUSION

The core Internet architecture was designed many years ago at a time when a global electronic society was not even envisaged. It seems long due for an overhaul to meet modern social requirements. In this upgrade technologists cannot stand on the sidelines. Technology is not and cannot be value neutral,

because online code affects online society. Just as physical laws determine what happens in physical worlds, so the "laws" of online interaction are affected by those who write the programs that create it. If computer system developers do not embody social concepts like freedom, privacy and democracy in their code, they will not happen. This means specifying social requirements just as technical ones currently are. While this is a daunting task, the alternative is an antisocial online society which could "collapse" (Diamond, 2005). If the next step of the human social evolution is an electronically enabled world society, computer technology may be contributing to a process of human unification that has been underway for thousands of years.

REFERENCES

Alter, S. (1999). A general, yet useful theory of information systems. *Communications of the AIS, 1*(March), 13-60.

Bales, R. F. (1950). A set of categories for the analysis of small group interaction. *American Sociological Review, 15*, 257-263.

Bertalanffy, L. V. (1968). *General system theory*. New York: George Braziller Inc.

Cooper, A. (1999). *The inmates are running the asylum-Why high tech products drive us crazy and how to Restore the sanity.* USA.

Deutsch, M., & Gerard, H. B. (1965). A study of normative and informational influences on social judgement. *Journal of Abnormal and Social Psychology, 51*, 629-636.

Diamond, J. (2005). *Collapse: How societies choose to fail or succeed*. New York: Viking (Penguin Group.

Freeden, M. (1991). Rights. *Concepts in social thought*.

Grudin, J. (1990). *The computer reaches out: The historical continuity of user interface design.* In *Proceedings of the CHI '90, ACM SIGCHI Conference*, Seattle, WA.

Hoffman, L. R. & Maier, N. R. F. (1964). Valence in the adoption of solutions by problem-solving groups: Concept, method and results. *Journal of Abnormal and Social Psychology, 69*(3), 264-271.

Hogg, M. A. (1990). *Social identity Theory*. New York: Springer-Verlag.

Huber, G. P. (1984). Issues in the design of group decision support systems. *Management Information Systems Quarterly, Sep, 8*(3), 195-204.

Kuutti, K. (1996). Activity theory as a potential framework for human computer interaction research. In B. A. Nardi (Ed.), *Context and consciousness: Activity theory and human-computer interaction*. Cambridge, MA: The MIT Press.

Locke, J. (1963). An essay concerning the true original extent and end of civil government: Second of "Two Treatises on Government" (1690). In J. Somerville & R. E. Santoni (Eds.), *Social and political philosophy* (pp. 169-204). New York: Anchor.

Mandelbaum, M. (2002). *The ideas that conquered the world*. New York: Public Affairs.

Maturana, H. R. & Varela, F. J. (1998). *The tree of knowledge*. Boston: Shambala.

McGrath, J. E. & Hollingshead, A. B. (1993). Putting the "group" back in group support systems. In L. M. Jessup & J. S. Valacich (Eds.), *Group support systems: New Perspective*. MacMillan.

Porra, J. & Hirscheim, R. (2007). A lifetime of theory and action on the ethical use of computers. A dialogue with Enid Mumford. *JAIS, 8*(9), 467-478.

Poundstone, W. (1992). *Prisoner's dilemma*. New York: Doubleday, Anchor.

Robert, H. M. (Ed.). (1993). *The new Robert's rules of order*. Barnes & Noble.

Sommerville, I. (2004). *Software engineering* (7th ed.).

Spears, R. & Lea, M. (1992). Social influence and the influence of the "social" in computer-mediated communication. In M. Lea (Ed.), *Contexts of computer mediated communication* (pp. 30-65). Hemel Hampstead: Harvester Wheatsheaf.

Sproull, L. & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32*(11), 1492-1512.

Warren, S. D. & Brandeis, L. D. (1890). The right to privacy. *Harvard Law Review, IV*(5), 193-220.

Whitworth, B. (2005). Polite computing. *Behaviour & Information Technology,* 5(September), 353 – 363.

Whitworth, B. (2008). *Some implications of comparing human and computer processing.* In *Proceedings of the 41st Hawaii International Conference on System Sciences*.

Whitworth, B. & deMoor, A. (2003). Legitimate by design: Towards trusted virtual community environments. *Behaviour & Information Technology, 22*(1), 31-51.

Whitworth, B. & Whitworth, E. (2004). Reducing spam by closing the social-technical gap *Computer,* (October), 38-45.

Whitworth, B., Aldo de Moor, A., & Liu, T. (2006). Towards a theory of online social rights. In Z. T. R. Meersman, P.

Herrero et al. (Eds.), In *Proceedings of the OTM Workshops 2006, LNCS 4277* (pp. 247 - 256). Berlin Heidelberg: Springer-Verlag

Whitworth, B., Fjermestad, J., & Mahinda, E. (2006). The web of system performance: A multi-goal model of information system performance. *Communications of the ACM, 49, May*(5), 93-99.

Whitworth, B., Gallupe, B., & McQueen, R. J. (2000). A cognitive three process model of computer-mediated groups: Theoretical foundations for groupware design. *Group Decision and Negotiation, 9*(5), 431-456.

Whitworth, B., Gallupe, B., & McQueen, R. (2001). Generating agreement in computer-mediated groups. *Small Group Research, 32*(5), 621-661.

Winograd, T. & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design*. Norwood, NJ: Ablex Pub. Corp.

KEY TERMS

Information System: A system that may include hardware, software, people and business or community structures and processes (Alter, 1999), c.f. a social-technical system, which must include all four levels.

Social System: Physical society is not just buildings or information, as without people information has no meaning. Yet it is also more than people. Countries with people of similar nature and abilities, like East and West Germany, perform differently as societies. While people come and go, the "society" continues, for example, we say "the Jews" survived while "the Romans" did not because the people lived on but because their social manner of interaction survived. A social system then is a general form of human interaction that persists despite changes in individuals, communications or architecture (Whitworth & deMoor, 2003).

System: A system must exist within a "world", as the nature of a system is the nature of the world that contains it, for example, a physical world, a world of ideas, and a social world, may contain physical systems, idea systems and social systems, respectively. A system needs identity to define "system" from "not system", for example, a crystal of sugar that dissolves in water still has existence as sugar, but is no longer a separate system. Existence and identity seem two basic requirements of any system.

System Elements: Advanced systems have a boundary, an internal structure, environment effectors and receptors (Whitworth et al., 2006). Simple biological systems (cells) formed a cell wall boundary and organelles for internal cell functions. Cells like Giardia developed flagella to effect movement, and protozoa developed light sensitive receptors. People are more complex, but still have a boundary (skin), an internal structure of organs, muscle effectors and sense receptors. Computer systems likewise have a physical case boundary, an internal architecture, printer/screen effectors and keyboard/mouse receptors. Likewise software systems have memory boundaries, a program structure, input analyzers, and output "driver" code.

System Environment: In a changing world, a system's environment is that part of a world that can affect the system. Darwinian "success" depends on how the environment responds to system performance. Three properties seem relevant: opportunities, threats, and the rate these change. In an opportunistic environment, right action can give great benefit. In a risky environment, wrong action can give great loss. In a dynamic environment, risk and opportunity change quickly, giving turbulence (sudden risk) or luck (sudden opportunity). An environment can be any combination, for example, opportunistic and risky and dynamic.

System Levels: Are physical systems the only possible systems? The term information system suggests otherwise. Philosophers propose idea systems in logical worlds. Sociologists propose social systems. Psychologists propose cognitive mental models. Software designers propose data entity relationship models apart from hardware. Software cannot exist without a hardware system of chips and circuits, but software concepts like data records and files are not hardware. A system can have four levels: mechanical, informational, personal and group, each emerging from the previous as a different framing of the same system, for example, information derives from mechanics, human cognitions from information, and society from a sum of human cognitions (Whitworth et al., 2006).

System Performance: A traditional information system's performance is its functionality, but a better definition is how successfully a system interacts with its environment. This allows usability and other "non-functional" requirements like security and reliability to be part of system performance. Eight general system goals seem applicable to modern software: functionality, usability, reliability, flexibility, security, extendibility, connectivity and confidentiality (Whitworth et al., 2006).

Building and Management of Trust in Networked Information Systems

István Mezgár
Hungarian Academy of Sciences, Hungary

INTRODUCTION

Thanks to rapidly developing information and communication technologies, the complexity of networked organizations has become very high, so the representation of their structure and the description of their operation and their control need new technologies, new approaches. The availability of individuals independently from location and time means mobility, and that is an important attribute of today's society. This mobility can be achieved by using different types of mobile wireless networks as wireless wide area networks (WWANs, e.g., GSM, GPRS, and UMTS), wireless local area networks (WLANs, e.g., WiFi 802.11a-g), and wireless personal area (or pico) network (WPAN, e.g., Bluetooth, IrDA2).

In spite of the application of high-tech approaches, tools, and methodologies, there is a common point in all of the organizations: human beings make most of the important decisions, and they operate and use systems. Experience shows that improper application of this human factor can make operation very inefficient even in the case of the technically most advanced systems. The lowest level of connection among systems is made through protocols; the highest contact level is among decision makers, users namely among human beings. A very important element of this human contact is trust. In a networked organization, trust is the atmosphere, the medium in which actors are moving (Castelfranchi & Tan, 2001). Only trust can bridge cultural, geographical, and organizational distances of team members (and even of firms) from turning to unmanageable psychological distances. Trust is the base of cooperation, the normal behavior of the human being in the society. The ability of enterprises to form networked systems depends on the existing level of trust in the society and on the capital of society (Fukuyama, 1995). As the rate of cooperation is increasing in all fields of life, the importance of trust is evolving even faster.

Lack of trustworthy security services is a major obstacle to the use of information systems in private, in business (B2B), as well as in public services. Trust is intimately linked to consumers' rights, like security, identification, authentication, privacy, and confidentiality. Secure identification, authentication of the users, and communication security are main problems in networked systems.

Information management (IM) is a fuzzy term covering the various stages of information processing from production to storage and retrieval to dissemination towards the better working of an organization, where information can be from internal and external sources and in any format. The role of trust in these processes is definitive as human-to-human and human-to-system communication forms the base of information management.

BACKGROUND

Definitions of Trust

The word "trust" is used by different disciplines, so there are many definitions of the term fulfilling the demands of the actual theory or application. In everyday life without trust, one would be confronted with the extreme complexity of the world in every minute. No human being could stand this, so people must have fixed points around them: one must have trust in family members, in partners, in the institutions of a society and between its members, and within and between organizations partners. The diversity of approaches is one reason that trust has been called an "elusive concept to define" (Gambetta, 1988).

Trust can be defined as a psychological condition comprising the trustor's intention to accept vulnerability based upon positive expectations of the trustee's intentions or behavior (Rousseau, Sitkin, Burt, & Camerer, 1998). Those positive expectations are based upon the trustor's cognitive and affective evaluations of the trustee and the system/world, as well as of the disposition of the trustor to trust. Trust is a psychological condition (interpreted in terms of expectation, attitude, willingness, perceived probability). Trust can cause or result from trusting behavior (e.g., cooperation, taking a risk), but is not behavior itself.

According to Luhmann (1979), trust can be viewed as a cognitive and social device able to reduce complexity, enabling people to cope with the different levels of uncertainty and sometimes the risks that, at different degrees, permeate our life. Without trust, an individual would freeze in uncertainty and indecision when faced with the impossibility of calculating all possible outcomes of a situation. Engaging trust automatically can reduce the number of decision nodes that are being analyzed and facilitate the decision-making processes. From a social perspective, trust permits the necessary knowledge sharing of delegation and cooperative actions.

The following components are included in most definitions of trust (Harrison, McKnight, & Chervany, 1996):

- willingness to be vulnerable/to rely;
- confident, positive expectation/positive attitude towards others; and
- risk and interdependence as necessary conditions.

Trust has different forms such as:

- **Intrapersonal trust:** Trust in one's own abilities; self-confidence/basic trust (in others).
- **Interpersonal trust:** Expectation based on cognitive and affective evaluation of the partners; in primary relationships (e.g., family) and non-primary relationships (e.g., business partners).
- **System trust:** Trust in depersonalized systems/world that functions independently (e.g., economic system, regulations, legal system, technology); requires voluntary abandonment of control and knowledge.
- **Object trust:** Trust in non-social objects; trust in its correct functioning (e.g., in an electronic device).

Trust Is a Multi-Faceted Construct

There is compelling evidence originating from the organizational research community to support the idea that trust is a many-sided, complex construct. McAllister (1995) has proposed two critical dimensions: emotional trust and cognitive trust. Emotional trust is the development of non-calculative and spontaneous emotional bonds and effect among two or more people. Emotional trust is demonstrated through confidence and openness in sharing ideas, feelings, and concerns. Cognitive trust refers both to judgments of competence (predictably professional behavior) and reliability (the congruence between words and actions) about the other members of a team.

Representation Forms of Trust

There are two basic modeling approaches in describing trust: the cognitive approach (Castelfranchi & Falcone, 1999) and the mathematical approach (Marsh, 1994). In case of applying cognitive models, trust is made up of underlying beliefs, and trust is a function of the value of these beliefs. The mathematical modeling approach ignores the role of underlying beliefs and uses a trust metric, based on variables like *perceived_competence, perceived_risk, utility of a situation for the agent involved, importance of a situation,* and so forth. These models incorporate some aspects of game theory and the evolution of cooperation models. Both modeling approaches see trust as a variable with a threshold for action. When the value of the variable crosses the threshold, the agent executes an action. In the Marsh model, the action is cooperation; in the Castelfranchi model, the action is delegation. The action is Boolean in nature — the agent either delegates or not, or the agent either cooperates or not.

Classification of the Meanings of Trust

Harrison et al. (1996) made a very deep and thorough analysis of the word "trust" from many aspects in their working paper. The goal of the paper was to develop a classification system for the types of trust and develop trust definitions/types that can be accepted by most of the disciplines.

The main groups of the classification system for trust constructs are as follows:

- **Impersonal/structural trust:** Those definitions of trust that differentiate it from being a property or state of a person or persons.
- **Dispositional trust:** Trust is based in the personality attributes of the trusting party.
- **Personal/interpersonal trust:** Trust in which one person trusts another person, persons, or thing(s) in the situation.

Guided by the classification system, six related types of trust have been defined in the working paper. The six constructs are as follows: Trusting Intention, Trusting Behavior, Trusting Beliefs, System Trust, Dispositional Trust, and Situational Decision to Trust. Both cognitive and affective components are included in Trusting Beliefs, Trusting Intention, and Trusting Behavior. The six constructs cover the more common of the dictionary definitions of trust. This multi-dimensional view of trust provides a parsimonious way to organize measurable trust types, while clearly distinguishing one type from another.

BUILDING TRUST

Connection of Trust and Information Management

Information technology management deals with the management of the different steps of information processing, and trust has a role where human beings are involved in this process. Human beings basically have two types of connections in these processes: human-to-human relationship through networks, and human-to-computer system communication through interfaces. In the first case trust management of virtual teams can be analyzed; in the second case special effects of computer interfaces and the role of security technologies in trust building and maintenance can be studied (Mezgar & Kincses, 2002). Information management must take into consideration the aspects of the trust-building process, to

develop, modify, and influence information handling in the direction that increases trust of human beings participating in these processes.

Trust Is More Than a Simple Technology

In building trust there are two approaches: information technology approach and human-centered approach, based on culture and morality. Information technology approach means that security must increase by different architectures, protocols, certifications, cryptography, and authentication procedures and standards, and this increased security generates the trust of users. This means access control (passwords, firewalls) integrity protection and privacy of message and database (cryptography) identification of users. Stressing the effectiveness of these technologies for human users can cause them to trust the systems based on this convincing action. Based on the technological approach, 100 % security can never be obtained (there will be always security holes somewhere in the system), so full trust cannot be guaranteed based on these mechanisms.

The human side of trust is more complicated. There were different researches (e.g., Hoffman, Novak, & Peralta, 1999) focusing on this side of trust. From this aspect, the user interface has the main role (i.e., the menu structure) to send the messages to the user from the system. In case the user feels that is easy to use, he or she can control the system (even with low-level computer knowledge) — that is, the system is "user friendly," and through this the user can be convinced that he or she is using a trustworthy system.

It would be a mistake to think that applying only the human-centered approach would generate trust; the technological part must be added as well (e.g., biometrical identification), so mainly the structured integration of the two approaches can result in the expected level of trust.

Technical Side of Trust Building: Application of Security Mechanisms

Security is a very complex term. There is computer, communication, information system, physical, and a lot of other "securities." As an addition, these terms overlap each other in many cases. Approaching security from the side of trust, security is the set of different services, mechanisms, and software and hardware tools for generating trust with pure technology. More generally, security is a condition that results from the establishment and maintenance of protective measures that ensure a state of inviolability from hostile acts or influences (FED, 2003).

The totality of protection mechanisms within a computer system — including hardware, orgware, and software, the combination of which is responsible for enforcing a security policy — is called trusted computing base (TCB). The ability of a trusted computing base to enforce correctly a unified security policy depends on the correctness of all types of mechanisms within the trusted computing base, the protection of those mechanisms to ensure their correctness, and the correct input of parameters related to the security policy.

In network management, the security management covers the set of functions: (a) that protects telecommunications networks and systems from unauthorized access by persons, acts, or influences; and (b) that includes many subfunctions, such as creating, deleting, and controlling security services and mechanisms; distributing security-relevant information; reporting security-relevant events; controlling the distribution of cryptographic keying material; and authorizing subscriber access, rights, and privileges (FED, 2003; Tipton & Krause, 1998).

The building block elements of security are the security services and the security mechanisms. The security services are (Schneier, 1996):

- **Confidentiality:** Protects against disclosure to unauthorized identities.
- **Integrity:** Protects from unauthorized data alteration.
- **Authentication:** Provides assurance of someone's identity.
- **Access Control:** Protects against unauthorized use.
- **Non-Repudiation:** Protects against originator of communications later denying it.

The means for achieving these properties depends on the collection of security mechanisms that supply security services, the correct implementation of these mechanisms, and how these mechanisms are used.

In security mechanisms with crypto functions (three basic building blocks are used):

- Encryption is used to provide confidentiality, and also can provide authentication and integrity protection.
- Digital signatures are used to provide authentication, integrity protection, and non-repudiation.
- Checksums/hash algorithms are used to provide integrity protection and can provide authentication.

Human Side of Trust Building: Feeling of Trust

The feeling of security experienced by a user of an interactive system does not depend on technical security measures alone. Other (psychological) factors can play a determining role; the user's feeling of control can be one of these factors.

Trust is a dynamic process, and it alters based on experience. Trusting process begins when an individual perceives indications that suggest a person/organization may be worthy of trust. These indications can include behaviors such as

manners, professionalism, and sensitivity, and these forms are designed to represent trustworthiness. These formal claims to trustworthiness become strengthened over time and are eventually transformed into "character traits," such as dependability, reliability. and honesty. Cheskin (1999) identifies six fundamental factors that communicate trustworthiness in the case of e-commerce Web sites:

- **Brand:** The importance of the company's reputation in the choice to do business with them.
- **Navigation:** The ease of finding what the user seeks.
- **Fulfillment:** The process the user experiences from the time a purchase process is initiated until the purchase is received to the user's satisfaction.
- **Presentation:** Ways in which the appearance of the site, in and of itself, communicates meaningful information.
- **Technology:** Ways in which the site technically functions.
- **Seals of Approval:** Symbols that represent the companies that specialize in assuring the safety of Web sites.

Why people feel safe and secure, and what causes these feelings, must be analyzed. The hypothesis of D'Hertefelt (2000) was that **"the feeling of security experienced by a user of an interactive system is determined by the user's feeling of control of the interactive system."** The more a user feels in control of an interactive program, the more the user will trust the site, the program. An interactive system that generates the feeling of control for the user must be: (a) comprehensible (the client needs to know what and how he or she can accomplish, and needs confirmation that it actually has been accomplished), (b) predictable (user has to know what is going to happen when he or she clicks on a bottom/menu item), and (c) flexible and adaptable (the user will feel in control if he or she can choose the way a task is executed instead of having to figure out how the system requires it to be done).

The process of building trust is slow; trust is formed gradually, it takes quite a lot of time and repeated positive experiences (Cheskin, 1999). Online trust can be described as a kind of human relationship. The initial stage is that of interest and distrust; there has to be a motivation, a need, to get interested in the service, or co-working. In subsequent phases the trust will evolve, or in case of negative experiences, the cooperation will terminate.

Trust is dependant on the time span of cooperation and the type of connection as well. It can be stated that there are differences in the trust-building process in short-term and long-term relationships. In case of short-term relationships (e.g., in a virtual organization), trust must be achieved quickly, and then maintain with no or rare face-to-face interaction. The members of these teams must assume that other remote team members are trustworthy, and then later on modify their assumptions according to their positive or negative experiences.

In long-term relationships there are four factors that influence trust building (Rocco, Finholt, Hofer, & Herbsleb, 2001):

1. Expectation of future interaction may motivate greater investment in building trustworthy relationships.
2. Long-term relationships offer more time to establish trustworthiness through routines and culture.
3. People have more communication channels, which may affect trust to the extent that participants have additional ways to clarify misunderstandings or to correct failures.
4. Participants are interested in successful task performance, and trust formation may assume a higher priority.

TRUST AND INFORMATION MANAGEMENT

Devices, Services, and Technologies for Trust Building in Information Management

As security, one important base of trust involves many topics — not only information security technology, but also the structure of the firms; the management techniques also must follow the ongoing changes of the IC technologies. In case the security challenges are put into a broader, management-oriented interpretation, they can be classified into four categories according to Dhillon (2001):

- Establishing good management practices in a geographically distributed environment and yet being able to control organizational operations.
- Establishing security policies and procedures that adequately reflect the organizational context and new business processes.
- Establishing correct structures of responsibility, given the complex structuring of organizations and information processing activities.
- Establishing appropriate information technology emergency recovery plans.

A solution for addressing trust is to build a chain of trust where each link is strong but also connects to its neighbor by verifying its trustworthiness. In particular, beginning with a priori grounding in a physically trustworthy base, each link of the chain checks signature-based trust tokens of its

neighbor before the chain is allowed to carry any weight. Such a chain begins with processors and extends to operating systems to applications to interconnect protocols, and ultimately, to end users.

Trust management unifies the notions of security policy, credentials, access control, and authorization. An application that uses a trust-management system can simply ask the compliance checker whether a requested action should be allowed. Trust-management policies are easy to distribute across networks, helping to avoid the need for application-specific distributed policy configuration mechanisms, access control lists, and certificate parsers and interpreters (Blaze, Feigenbaum, & Lacy, 1996).

Virtual Teamwork and Trust Management

Today the different types of networked organizations need new types of cooperation, as the members of the working teams are geographically (physically) separated. They use shared documents, and communicate through e-mail and high quality audio and video channels. These teams are called "virtual teams," as they never meet personally and they have no face-to-face (FTF) contact. The work of teams without FTF contact is less effective and reliable based on the observation stated by Handy (1995): "Trust needs touch." According to case studies, it is evident that trust of virtual team members is significantly lower than trust in conventional teams (Rocco et al., 2001). In other experiments where interaction was primarily via e-mail, very similar results were gained, as in geographically distributed teams (Jarvenpaa & Leidner, 1999).

In an experiment introduced in Bos, Olson, Gergle, Olson, and Wright (2002), four media types were compared: chat (text), phone conference, videoconference, and FTF. Chat was significantly worse than each of the other three conditions, but audio and video did as well as FTF in overall cooperation, and were a definite improvement over text-chat-only computer-mediated communication. However, these two channels still showed evidence of delayed trust, in that they took longer to reach high levels of cooperation.

There are experiments on the truth content of the communication as well. It came to light that participants lied most on the telephone and least in e-mail, and that lying rates in face-to-face and instant messaging interactions were approximately equal. These results suggest that the type of communication technologies affect lying behavior in important ways. Both designers and users must take these features into consideration when trust-building processes are designed and applied (Hancock, Thom-Santelli, & Ritchie, 2004).

The latest research shows that if people meet before using computer-mediated communication, they trust each other, as trust is being established through touch. In cases where participants do not meet formerly, but they initiate various get-acquainted activities over a network, trust is much higher than if they do nothing before and nearly as good as a prior meeting. Using chat forums to get acquainted is nearly as good as meeting, and "even just seeing a picture is better than nothing" (Zheng, Veinott, Bos, Olson, & Olson, 2002).

Information Technology Tools for Generating Trust

Identification of a user/customer of an information system is a complex task. In computer science the identifier is a string of bits or characters that names an entity, such as a program, device, or system, in order that other entities can call that entity. In the context of information systems, the purpose of identification is very concrete: it is used to link a stream of data with a certain person, so the following definition can be given: "Human identification is the association of data with a particular human being."

Information systems have tended to use codes rather than names as the primary identification mechanism. The most reliable mode to identify a person is to apply biometric techniques (Jain, Bolle, & Pankanti, 1999). The applied techniques in biometry systems in IT include a physiological (fingerprint, iris, facial features, etc.) element or factor, as well as a behavioral one (e.g., vocal patterns, typing rhythm). Biometric identification is preferred over current methods involving passwords and pin numbers, as the person to be identified is required to be physically present at the point of identification; thus, the person or user is identified, not the device as in the case of a PIN and password. Layered Biometric Verification (LBV) technology entails layering different biometric technologies into a complex identification process.

The main factor of trust is confidentiality that can be achieved by technologies that convert/hide the data/text into a form that cannot be interpreted by unauthorized persons. There are two major techniques to fulfill this goal: encryption and steganography. Encryption is transforming the message to a cipher text such that an enemy who monitors the cipher text cannot determine the message sent (Schneier, 1996). Steganography is the art of hiding a secret message within a larger one in such a way that the opponent cannot discern the presence or contents of the hidden message (Johnson, Duric, & Jajodia, 2000). For example, a message might be hidden within a picture by changing the low-order pixel bits to be the message bits.

Generating Trust by Human-Computer Interfaces

Generally speaking the goal of an interface is to interconnect two or more entities at a common point or shared boundary. As a communication/information system term, an interface is the point of communication between two or more processes,

persons, or other physical entities. Interfaces are the key points for gaining the trust of the user/customer. They are the first connection point between the user and the system; identification of the users take place at this point (e.g., password input, fingerprint reader, smart card reader), so interfaces must be designed very carefully. In the design of both types of interfaces, ergonomic and psychological aspects are taken into consideration, in addition to the technical ones.

Researchers test different new types of interfaces. Interaction must be extended with more senses (touch, smell, and taste) and make better use of the senses used today (hearing and vision) by exploring peripheral vision and ambient listening. 'All-senses communication' would be one way to enhance the communication with other entities (humans or machines) using a combination of several present or future senses of humans. Multimodal systems (Oviatt, Coulston, & Lunsford, 2002) process two or more combined user input modes—such as speech, pen, touch, manual gestures, gaze, and head and body movements—in a coordinated manner with multimedia system output. This class of systems represents a new direction for computing, and a paradigm shift away from conventional interfaces to the collaborative multimodal interfaces.

Trusted Time

Trusted time is important in global network communication. Trusted time is essential to ensuring that network and operations infrastructure run accurately, securely, and reliably, and that all relying applications execute with traceability that provides auditability to resolve disputes, avoid repudiation, and ensure legal enforceability of electronic business transactions. The auditable sourcing and synchronization of time involves a "chain of trust."

FUTURE TRENDS IN INFORMATION TECHNOLOGY MANAGEMENT

Nearly all types of systems in all fields of the economy became distributed, and virtual structures appeared. The result is large structural and cultural changes in enterprises. A new communication technology appeared as well—wireless technology. The management of these new systems has also brought new aspects into focus in the field of information management. As Malone (2004) appoints in his book, the style of management moves from command-oriented management to coordination-based approaches (new skills are required), The main advantages of wireless and mobile communication are that anybody, from anywhere, at anytime, can make contacts.

Mobile devices become far more popular as was estimated before, and thanks to the extremely fast development of the electronic industry, these devices have grown into a multi-functional tool. Mobile Internet rewrites many rules. People everywhere, in different positions, are using new services that are relevant to their personal needs and preferences — and are accessible anytime, anywhere. New terms are developing in all sectors of the industry — in finance, in government, and in society.

Information architectures, structures, business processes, and business models of enterprises must be modified according to new infocom technology. Mobile technologies add new value as to be continuously connected. Short response times assure the validity of information. The productivity is no longer restricted by place or time. Best of all, it is possible to experience new ways of communicating, sharing information. In an information and communication-centered world, security and trust are exceptionally important, as the value and strategic role of reliable information is extremely high. Information technology management systems and the managers themselves must adapt to this fast, non-stop changing environment. Smart cards, personal trusted devices (PTDs), secure agent architectures, and new protocols for ad hoc networks (Patwardhan, Parker, Joshi, Karygiannis, & Iorga, 2005) are some of the main elements in information technology that will change in the future.

In spite of the many positive characteristics and effects introduced so far, mobile communication has negative sides as well, for example, the possibility of tracking services and routes of owners by governments and marketers using information from wireless devices. An additional general problem is the establishment of trust and reputation in the wireless world. Using PTDs for trust building makes authentication and confidentiality easier and more reliable to support the operation of different types of virtual organizations.

In using mobile devices there is a problem generated by camera phones. They have become extremely popular, as it is easy to take pictures and send them immediately to partners as well as to send videos and other multimedia materials. In some countries it is not easy to buy mobile phones without a camera. Alternatively, companies and sporting clubs are stressing that privacy rights and company properties (technology, etc.) are in danger when people bring camera phones into their territory, so they prohibit the use of camera-phones in their areas. The exaggerated usage of mobile devices can cause problems in workplaces as well by disturbing others with frequent calls or by exploiting the video possibilities of the new, enhanced phones. In some workplaces, private usage of mobile phones and SMS services are prohibited, as they decrease work efficiency and cause additional security risk.

CONCLUSION

The chapter has briefly introduced the main characteristics of trust and its connections between security services and

mechanisms and its role in information technology management. The importance of trust is increasing very fast, as the main characteristic of the information and knowledge society is the cooperation through computer and communication networks. As pointed out by different analysis based on real-life statistics, when users do not trust a system/service, they do not use it. Organizations must adapt to this requirement, by changing their culture or organization structures as well.

Integrated mobile technologies are speeding up this tendency as they offer mobility/freedom for the citizens. Distributed information systems of different sizes are playing a definite role, but originating from their openness and flexibility, these systems will always be a security risk. As high-level security and trust cannot be reached with pure information technology approaches, they must be integrated with the human-centered techniques based on psychological methods.

There is a need for complex, flexible security systems that are user friendly and platform independent at the same time. The developments of hardware and software elements of such systems are going on, and the potential users have to get acquainted with them. The managers of information technology must adapt these technologies, tools, and devices into their systems to provide a high level of security and trust that can induce trust in all humans involved in the different phases of the lifecycle of information systems.

REFERENCES

Blaze, M., Feigenbaum, J., & Lacy, J. (1996). Decentralized trust management. *Proceedings of the 17th Symposium on Security and Privacy* (pp. 164-173), Los Alamitos, CA.

Bos, N.D., Olson, J.S., Gergle, D., Olson, G.M., & Wright, Z. (2002). Effects of four computer-mediated channels on trust development. *Proceedings of CHI 2002.*

Castelfranchi, C., & Falcone, R. (1999). The dynamics of trust: From beliefs to action. *Proceedings of the Autonomous Agents Workshop on Deception, Fraud and Trust in Agent Societies.*

Castelfranchi, C., & Tan, Y.-H. (Eds.). (2001). *Trust and deception in virtual societies.* Norwell, MA: Kluwer Academic.

Cheskin. (1999, January). *E-commerce trust: A joint research study with Studio Archetype/Sapient and Cheskin.* Retrieved from http://www.cheskin.com/cms/files/i/articles//17__report-eComm%20Trust1999.pdf

D'Hertefelt, S. (2000). *Trust and the perception of security.* Retrieved from http://www.interactionarchitect.com/research/report20000103shd.htm

Dhillon, G. (2001). *Information security management: Global challenges in the new millennium.* Hershey, PA: Idea Group.

FED. (2003). *FED-STD-1037C, telecommunications: Glossary of telecommunication terms.* Retrieved from http://glossary.its.bldrdoc.gov/fs-1037/

Fukuyama, F. (1995). *Trust — the social virtues and the creation of prosperity.* New York: The Free Press.

Gambetta, D. (1988). Can we trust trust? In D. Gambetta (Ed.), *Trust, making and breaking cooperative relations* (pp. 213-237), Oxford: Basil Blackwell.

Hancock, J.T., Thom-Santelli, J., & Ritchie, T. (2004, April). Deception and design: The impact of communication technology on lying behavior. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 129-134), Vienna, Austria.

Handy, C. (1995). Trust and the virtual organization. *Harvard Business Review, 73*(3), 40-50.

Harrison, D., McKnight, N., & Chervany, L. (1996). *The meanings of trust.* Working Paper 96-04, University of Minnesota Management Information Systems Research Center, USA.

Hoffman, D.L., Novak, T.P., & Peralta, M. (1999). Building consumer trust online. *CACM, 42*(4), 80-85.

Jain, A., Bolle, R., & Pankanti, S. (1999). *Biometrics: Personal identification in networked society.* Norwell, MA: Kluwer Academic.

Jarvenpaa, S.L., & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-815.

Johnson, N.F., Duric, Z., & Jajodia, S. (2000). *Information hiding: Steganography and watermarking — attacks and countermeasures.* Norwell, MA: Kluwer Academic.

Luhmann, N. (1979). *Trust and power.* Chichester: John Wiley & Sons.

*Malone, T.W. (*2004). *The future of work.* **Boston:** Harvard Business School Press.

Marsh, S. (1994). *Formalising trust as a computational concept.* PhD Thesis, Department of Computing Science and, Mathematics, University of Stirling, Scotland.

McAllister, D.J. (1995). Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of Management Journal, 38*(1), 24-59.

Mezgár, I., & Kincses, Z. (2002). The role of trust in information management. In A. Gunasekaran (Ed.), *Knowledge*

and information technology management in the 21st century organizations: Human and social perspectives (pp. 283-304). Hershey, PA: Idea Group.

Oviatt, S.L., Coulston, R., & Lunsford, R. (2004). When do we interact multimodally?: Cognitive load and multimodal communication patterns. *Proceedings of ICMI 2004* (pp. 129-136). New York: ACM Press.

Patwardhan, A., Parker, J., Joshi, A., Karygiannis, A., & Iorga, M. (2005, March 8-12). Secure routing and intrusion detection in ad hoc networks. *Proceedings of the 3rd IEEE International Conference on Pervasive Computing and Communications,* Kauai Island, HI.

Rocco, E., Finholt, T.A., Hofer, E.C., & Herbsleb, J.D. (2001, April). Out of sight, short of trust. *Proceedings of the Founding Conference of the European Academy of Management,* Barcelona, Spain.

Rousseau, D.M., Sitkin, S.B., Burt, R., & Camerer, C. (1998). Not so different after all: A cross-disciplinary view of trust. *Academy of Management Review, 23,* 1-12.

Schneier, B. (1996). *Applied cryptography.* New York: John Wiley & Sons.

Tipton, H., & Krause, M. (Eds.). (1998). *Handbook of information security management.* CRC Press.

Zheng, J., Veinott, E, Bos, N., Olson, J.S., & Olson, G.M. (2002). Trust without touch: Jumpstarting long-distance trust with initial social activities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves* (pp. 141-146), Minneapolis, MN.

KEY TERMS

Biometry/Biometrics: Generally, biometrics refers to the study of measurable biological characteristics. In computer security, biometric technologies are defined as automated methods of identifying or authenticating the identity of a living person based on his or her physiological (e.g., fingerprint, hand, ear, face, eye — iris/retina) or behavioral (e.g., signature, voice, keystroke) characteristic. This method of identification is preferred over current methods involving passwords and pin numbers, as the person to be identified is required to be physically present at the point of identification, so the person or user is identified, not the device, as in case of a PIN and password.

Encryption: The transformation of plaintext into an apparently less readable form (called cipher text) through a mathematical process. The cipher text may be read by anyone who has the key that decrypts (undoes the encryption of) the cipher text.

Layered Biometric System: Multi-layer/layered/multimodal biometric systems combine more than one physiological or behavioral characteristic for verification or identification. By promoting multi-layered identification and authentication — that means parallel use of strong passwords, smart tokens, and biometrics — many significant security problems can be eliminated. The combination of multiple biometric methods such as voice and fingerprints, in conjunction with digital certificates, smart cards, and so forth, offer the companies an ideal solution to provide very high-level protection of their sensitive information.

Personal Trusted Device: People like smart, little, handheld tools they can carry with them permanently, so they can control them both physically and in time. According to the concept of the personal trusted device, it must be personal, always carried by user, small, cheap, accumulator powered, have a common user interface, and be as secure as a smart card. Mobile phones can fulfill the role of personal trusted devices, as mobile phones are well placed as identity tokens, they have dynamic authentication already proven in GSM, and have mass market and secure communications. Mobile phones are the only mass-market smart card readers, and they are highly personal. Users are usually very attached to their phones, and they can be made more personal by use of a PIN or (later) biometrics.

Trust: Can be viewed as a cognitive and social device able to reduce complexity, enabling people to cope with the different levels of uncertainty and sometimes the risks that, at different degrees, permeate our life. Without trust, an individual would freeze in uncertainty and indecision when faced with the impossibility of calculating all possible outcomes of a situation. From a social perspective, trust permits the necessary knowledge sharing of delegation and cooperative actions (Luhmann, 1979).

Trust Chain: A solution for addressing trust is to build a chain of trust where each link is strong but also connects to its neighbor by verifying its trustworthiness.

Trust Management: A unified approach to specifying and interpreting security policies, credentials, and relationships; it allows direct authorization of security-critical actions. A trust-management system provides standard, general-purpose mechanisms for specifying application security policies and credentials. Trust-management credentials describe a specific delegation of trust and subsume the role of public key certificates; unlike traditional certificates, which bind keys to names, credentials can bind keys directly to the authorization to perform specific tasks.

Trusted Time: Emerging as the industry best practice, trusted time is auditable, secure, available, warranted, accurate, and managed. Trusted time is essential to ensuring that network and operations infrastructure run accurately, securely, and reliably, and to ensure legal enforceability of electronic business transactions. The auditable sourcing, setting, and synchronization of time involves a "chain of trust" that is complex, requires time specialization, and involves multiple sources of risk.

Trustworthiness: The ability to attain and maintain a "trusted state," which is definable, measurable, validatable, and demonstrable over time. Digital trustworthiness means a verifiable level of electronic process integrity, security, control, authenticity, and reliability that captures, preserves, retrieves, verifies, renders, and makes available in human-readable form — the essential transaction content, context, notice, intent, and consent — to meet the electronic forensic evidence requirements necessary for legal admissibility and regulatory compliance.

Building Educational Technology Partnerships through Participatory Design

John M. Carroll
The Pennsylvania State University, USA

INTRODUCTION

Educational technology provides many examples of how efficient software development and deployment is not enough. Teachers work in a complex and dynamic context in which measurable objectives and underlying values collide on a daily basis. Traditionally, teachers work in isolation from their peers; individual teachers have well-established personal practices and philosophies of education. Teachers have enormous discretion with respect to what goes on in their classrooms, yet are also routinely interrogated by supervisors, by parents and other community members, and by educational bureaucracies. This has led to an abiding tension in the culture of schools: Teachers' innovative practices are often not adequately acknowledged or valued, and at the same time, teachers often passively resist school reforms that are imposed top-down.

Technology is a particularly problematic element in the culture of schools. The isolation and discretion of the teacher's work environment requires that technology for classroom use be highly appropriate and reliable. Yet it is generally assumed that teachers are to be *trained* on new technologies, not asked to *define* what those technologies should be. From the teacher's standpoint, classroom technology often is itself the problem, not the solution. This culture of technologydevelopment in the schools has been singularly ineffective—film and radio in the 1920s, television in the 1950s, and computer-assisted instruction in the 1980s, among others, have been notable failures (Tyack & Cuban, 1995).

An alternative to merely efficient technology development is *participatory design*, the inclusion of users within a development team such that they actively help in setting design goals and planning prototypes. This approach was pioneered, and has been widely employed, in Europe since the 1970s, and now consists of a well-articulated and differentiated set of engineering methods in use worldwide (Carroll, 2000; Clement & Van den Besselaar, 1993; Muller, 2003; Muller, Haslwanter, & Dayton, 1997; Rosson & Carroll, 2002).

In 1994, a design collaboration was formed between Virginia Tech and the public schools of Montgomery County, Virginia. The objective was to develop and investigate a high-quality communications infrastructure to support collaborative science learning. Montgomery County is located in the rural Appalachian region of southwestern Virginia. In March 2000, one of its high schools was listed among the top 100 in the US by *Newsweek* magazine. However, in others, physics is only offered every other year and to classes of only three to five students. The initial vision was to give students in this diverse and dispersed school district access to peers through networked collaboration.

We felt it was critical for the teachers to contribute as collaborators in design analysis, implementation, deployment, testing, and refinement, and as leaders in the development of courseware and classroom activities that would exploit the software. For a classroom-technology partnership to succeed, the university researchers must eventually fade and leave the teachers to maintain and develop its achievements. In the end, the technology-development goals of this project were achieved, though this is not the topic of this paper (Isenhour, Carroll, Neale, Rosson, & Dunlap, 2000).

BACKGROUND

We analyzed our participatory engagement with the teachers as "developmental" in the sense of Piaget and Inhelder (1969) and Vygotsky (1978). We believe the teachers developed qualitatively different roles through the course of our collaboration. In some cases, these roles were suggested to them; in other cases, they defined and claimed new roles. But in all cases, these transitions exemplified the defining characteristics of *developmental change*: active resolution of manifest conflicts in one's activity, taking more responsibility, and assuming a greater scope of action. Each successive stage can be seen as a relatively stable organization of knowledge, skills, and attitudes that resolves the instigating conflict.

During the six years of this project, we distinguished four stages in our collaboration with the teachers. At first, the teachers were *practitioner-informants*; we observed their classroom practices and we interviewed them. Subsequently, the teachers became directly and actively involved in the requirements-development process as *analysts*. Later, the teachers assumed responsibility as *designers* for key aspects of the project. Finally, the teachers became *coaches* to their own colleagues within the public school system.

In a classic Piagetian example, a child in the preoperational stage perceives single dimensions of quantity. This produces conflicts: A given quantity of liquid poured from a short, wide container into a tall, thin container appears suddenly to be more, but of course cannot be more. These conflicts eventually precipitate a cognitive reorganization called the concrete operational stage, in which constant quantities are perceived as constant regardless of varying shapes and arrangements.

Developmental change in adults is of course more complex. The stages we describe are not singular competencies, but relatively complex ensembles of collaboration, social norms, tool manipulation, domain-specific goals and heuristics, problem solving, and reflection in action. They are social constructions achieved through enculturation, constituted by the appropriation of the artifacts and practices of a community (Vygotsky, 1978).

In the Piagetian notion of stages in child development, successive stages build upon the cognitive structures and enabled activity of prior stages, but ultimately replace those structures. A child who enters the concrete operational stage can no longer function at the preoperational stage. Adult growth, however, is not static achievement, but continual elaboration. The teachers are still practitioners whose classroom practices we regularly observe and whose classroom expertise we still interrogate; they seem to us and to themselves to be representative practitioner-informants. However, they are now *also* analysts and designers, and often coaches. Indeed, effective design coaches probably must be experienced designers, successful designers must be skilled analysts, and analysts must have attained significant domain knowledge (Carroll, Chin, Rosson, & Neale, 2000).

MAIN THRUST OF THE CHAPTER

Developmental theory explains transitions between stages as resolutions of conflict. Thus, the preoperational child's conflicting perceptions of quantity based on single dimensions, such as height and width, are resolved in the abstraction of quantity as an invariant in the concrete operational stage. For development to take place, the child must have attained the requisite competencies to experience the triggering conflict, and then be able to reconceptualize the situation in such a way that the conflict dissolves.

This analytical schema seems to fit the transitions between the stages of cooperation we identified. The general mechanism appears to be that successive increases in knowledge, skill, and confidence empowered the teachers to resolve conflicts by assuming successively greater scope of action and responsibility in the project. Early on, the teachers faced the conflict that their pedagogical concerns and perspectives would be adequately represented and fully considered by the group only if they themselves championed those concerns. This went beyond the practitioner-informant role they had played in the project up to then. But they were both motivated and competent to resolve this conflict by assuming the analyst role in the project.

Once the teachers were functioning as analysts in the project team, further conflicts and resolutions arose. The teachers experienced a conflict between their own analyses of system requirements and the current state of our project software and development plans. They resolved these conflicts by formulating their own design proposals, ultimately a radical reorientation of the project's vision of classroom activity. They became designers. Subsequently, the teachers recognized that they were the best qualified project members to train new teachers and to pursue specific curricular extensions of the project. They became coaches.

The teachers' behavior also reflects development *within* the four general stages we have described. For example, cognitive scaffolding (via examples, reflective prompts) was needed to engage the teachers in the novel and relatively abstract activity of design analysis. But as the project progressed, teachers spontaneously identified and presented design trade-offs to the group as a way to articulate personal positions. This is consonant with the general notion of learning as movement through a zone of proximal development (Vygotsky, 1978).

The designer stage also reflects several different levels of development. Initially, the teachers were able to collaborate with a research assistant in focused design sessions, cowriting scenarios of technology-mediated activities for their classroom. Later they banded together as a subgroup, pooling their goals and expertise to develop a scenario that specified a new vision of collaborative learning activities. Ultimately, each learned to function as an independent designer, envisioning and specifying activities optimized for their own teaching styles, objectives, and classroom environments. In their coach role, the teachers also worked first as a group, but subsequently recruited and mentored colleagues in a one-to-one fashion.

In sum, it appears that the transitions among stages were triggered by conflicts with respect to the teachers' role in the project. In each case, a series of scaffolded activities enabled them to attain the knowledge, skill, and confidence that led them to expand their role (Carroll et al., 2000).

FUTURE TRENDS

We originally committed to a long-term participatory-design method because we conjectured that such an approach would be crucial for success in this educational technology setting. We believe we could not have succeeded to the extent we have had we not made this commitment. Working from the

national agenda for school reform, educational technology, and science education (AAAS, 1993; NTSA, 1992), and from our own initial vision of a "virtual school," we would have built the wrong system—we would not have had effective support from teachers. Little or nothing would have been sustained after the initial project funding ended.

Participatory design is fundamentally a process of mutual learning, and thus of personal development for participants. But it is often exemplified by rather singular and ephemeral learning interactions. Our study expands the scope of the design participants' personal development by examining a case of long-term cooperative design interaction, and by describing a developmental sequence of roles with constituent capacities and responsibilities.

Much research on participatory design has focused on relatively short-term collaborative relationships. This is especially true in North America, where many participatory-design techniques are directed at brief user-interface-design interactions of perhaps one hour (Muller et al., 1997). Such methods are both effective and democratic, but it seems unlikely that the experience of manipulating a user-interface mock-up during a brief participatory session can have a significant developmental effect on a person's knowledge, skills, self-confidence, or other professional capacities.

In our project, user-interface design per se was a secondary issue. We used brief participatory exercises, but this level of engagement was more a starting point than the objective of our work. More specifically, we wanted the teachers to have a significant voice in designing the functionality and the use of the virtual school, not merely its appearance. We needed to learn about pedagogical goals and practices, classroom management, school-system politics, the relationship of the community and the schools, and so forth.

Where participatory-design investigations *have* focused on longer term interactions, chiefly in Europe, these often involve extremely well-organized user groups with well-defined roles and prerogatives in the design process. In many cases, the users are represented by labor unions whose personnel provide legal representation of user interests in the design process. In these cases, there is sometimes a clear demarcation, even conflict, between the user (union) interests and management's technology strategy. Indeed, this is an important element of the context for many of these studies. Because the user role in many of these studies is both specified a priori and representative (versus individual), the personal development of user-designers is not a central issue. These case studies also typically involve situations in which the development and deployment of new information technology is a given, and the challenge is to define appropriate technology for the users and their activities (Bjerknes & Bratteteig, 1993; Bødker, Ehn, Kammersgaard, Kyng, & Sundblad, 1987; Merkel et al., 2004).

In the educational domain, the deployment of new information technology is far from given. Indeed, the introduction of new technology has historically almost always failed in school settings. One of the key questions for us was whether a concept for appropriate technological support could be developed at all.

The users in our domain are very loosely organized. As mentioned earlier, teachers traditionally work in isolation from peers; they manage their own work practices and environments (classrooms). The notion of a "user community" in this domain is almost ironic. Teachers unions in the US are also extremely weak and play no role in the introduction of classroom technology. Indeed, school administrations in the US rarely have technology strategies at all. Thus, unlike the European case studies, the issue is almost never one of recognized conflict, but rather finding a direction at all.

The teachers in our team do not represent other teachers; they are individuals who, as members of our team, have become teacher-designers. This is precisely why their personal development as designers is a central issue in our study. Of course, we do hope that they are representative teachers—allowing us to generalize our investigation to other teachers participating in similar development projects—but this is a separate issue. The point is that in our project, and unlike many long-term participatory-design efforts in Europe, the teachers act as individual professionals just as university researchers do.

CONCLUSION

The stages we have described here are specific to our project; they emerged through specific things that we did and are rooted in the specific goals of our project. At the same time, they suggest a schematic programme for developing cooperative engagement more generally. Most participatory-design work engages users at the practitioner-informant stage. This would seem to be an obvious and general starting point for any participatory-design collaboration. In our project, the teachers transitioned to the analyst stage through their inclusion in a requirements-analysis workshop and a significant process of iterative requirements development (Carroll, Rosson, Chin, & Koenemann, 1998). This is perhaps not typical of participatory-design practice, but it is a modest extension. Nevertheless, the teachers found it quite stimulating to be invited to objectify their own experience, to dissect it and not merely describe it.

ACKNOWLEDGMENT

This work was partially supported by the Hitachi Foundation and the National Science Foundation (REC-9554206 and DGE-9553458). This paper is a summary of Carroll, Chin, Rosson, Neale, Dunlap, and Isenhour (2002).

REFERENCES

American Association for the Advancement of Science (AAAS). (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Bjerknes, G., & Bratteteig, T. (1995). User participation and democracy: A discussion of Scandinavian research on system development. *Scandinavian Journal of Information Systems, 7*(1), 73-98.

Bødker, S., Ehn, P., Kammersgaard, J., Kyng, M., & Sundblad, Y. (1987). A utopian experience. In G. Bjerknes, P. Ehn, & M. Kyng (Eds.), *Computers and democracy: A Scandinavian challenge* (pp. 251-278). Brookfield, VT: Avebury.

Carroll, J. M. (2000). *Making use: Scenario-based design of human-computer interactions.* Cambridge, MA: MIT Press.

Carroll, J. M., Chin, G., Rosson, M. B., & Neale, D. C. (2000). The development of cooperation: Five years of participatory design in the virtual school. In D. Boyarski & W. Kellogg (Eds.), *DIS'2000: Designing interactive systems* (pp. 239-251). New York: Association for Computing Machinery.

Carroll, J. M., Chin, G., Rosson, M. B., Neale, D. C., Dunlap, D. R., & Isenhour, P. L. (2002). Building educational technology partnerships through participatory design. In J. Lazar (Ed.), *Managing IT/community partnerships in the 21st century* (pp. 88-115). Hershey, PA: Idea Group Publishing.

Carroll, J. M., Rosson, M. B., Chin, G., & Koenemann, J. (1998). Requirements development in scenario-based design. *IEEE Transactions on Software Engineering, 24*(12), 1-15.

Clement, A., & Van den Besselaar, P. (1993). A retrospective look at PD projects. *Communications of the ACM, 36*(4), 29-37.

Isenhour, P. L., Carroll, J. M., Neale, D. C., Rosson, M. B., & Dunlap, D. R. (2000). The virtual school: An integrated collaborative environment for the classroom. *Educational Technology and Society, 3*(3), 74-86.

Merkel, C. B., Xiao, L., Farooq, U., Ganoe, C. H., Lee, R., Carroll, J. M., et al. (2004). Participatory design in community computing contexts: Tales from the field. In *Artful integration: Interweaving media, materials and practices-Proceedings of the 2004 participatory design conference.* Palo Alto, CA: Computer Professionals for Social Responsibility.

Muller, M. J. (2003). Participatory design: The third space in HCI. In J. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications.* Mahwah, NJ: Erlbaum.

Muller, M. J., Haslwanter, J. H., & Dayton, T. (1997). Participatory practices in the software lifecycle. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of human-computer interaction* (2nd ed., pp. 255-297). Amsterdam: Elsevier.

National Science Teachers Association (NTSA). (1992). *Scope, sequence and coordination of secondary school science: Vol. 1, The content core*. Washington, DC.

Piaget, J., & Inhelder, B. (1969). *The psychology of the child*. New York: Basic Books.

Rosson, M. B., & Carroll, J. M. (2002). *Usability engineering: Scenario-based development of human-computer interaction.* San Francisco: Morgan Kaufmann.

Tyack, D., & Cuban, L. (1995). *Tinkering toward utopia: A century of public school reform*. Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

KEY TERMS

Analysts: Users who collaborate with designers as domain experts analyzing constraints and trade-offs in existing and envisioned work practices are called *analysts*. This is the second stage in the developmental theory of participatory-design relationships between users and designers.

Coaches: Users who help other users participate in design work by coaching them are called *coaches*. This is the fourth stage in the developmental theory of participatory-design relationships between users and designers.

Designers: Users who collaborate with designers as domain experts envisioning new work practices and tools are called *designers*. This is the third stage in the developmental theory of participatory-design relationships between users and designers.

Developmental Theory: Theories of learning that involve growth and other qualitative changes in skills, knowledge, and capacities. Developmental theory is contrasted to accretive theories in which learning is conceived of as a matter of quantitative improvements—more knowledge or faster performance.

Educational Technology: Technology used in formal educational contexts, such as classrooms. Recent examples are television, personal computers, and the Internet.

Participatory Design: Design methods in which users (and other stakeholders) provide special expertise and play active and autonomous roles in design work.

Practitioner-Informants: Users who collaborate with designers as domain experts providing information about work practices are called *practitioner-informants*. This is the initial stage in the developmental theory of participatory-design relationships between users and designers.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 307-311, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Building Local Capacity via Scaleable Web-Based Services

Helen Thompson
University of Ballarat, Australia

INTRODUCTION

Information communications technology (ICT) has been identified as a key enabler in the achievement of regional and rural success, particularly in terms of economic and business development. The potential of achieving equity of service through improved communications infrastructure and enhanced access to government, health, education, and other services has been identified. ICT has also been linked to the aspiration of community empowerment, where dimensions include revitalizing a sense of community, building regional capacity, enhancing democracy, and increasing social capital.

In Australia, there has been a vision for online services to be used to open up regional communities to the rest of the world. Government support has been seen "as enhancing the competence levels of local economies and communities so they become strong enough to deal equitably in an increasingly open marketplace" (McGrath & More, 2002, p. 40). In a regional and rural context, the availability of practical assistance is often limited. Identification of the most appropriate online services for a particular community is sometimes difficult (Ashford, 1999; Papandrea & Wade, 2000; Pattulock & Albury Wodonga Area Consultative Committee, 2000). Calls, however, continue for regional communities to join the globalized, online world. These are supported by the view that success today is based less and less on natural resource wealth, labor costs, and relative exchange rates, and more and more on individual knowledge, skills, and innovation. But how can regional communities "grab their share of this wealth" and use it to strengthen local communities (Simpson 1999, p. 6)? Should communities be moving, as Porter (2001, p. 18) recommends (for business), away from the rhetoric about "Internet industries," "e-business strategies," and the "new economy," to see the Internet as "an enabling technology—a powerful set of tools that can be used, wisely or unwisely, in almost any industry and as part of almost any strategy?"

Recent Australian literature (particularly government literature) does indeed demonstrate somewhat of a shift in terms of the expectations of ICT and e-commerce (National Office for the Information Economy, 2001; Multimedia Victoria, 2002; National Office for the Information Economy, 2002). Consistent with reflections on international industry experience, there is now a greater emphasis on identifying locally appropriate initiatives, exploring opportunities for improving existing communication and service quality, and for using the Internet and ICT to support more efficient community processes and relationships (Hunter, 1999; Municipal Association of Victoria and ETC Electronic Trading Concepts Pty Ltd., 2000; National Office for the Information Economy, 2002).

The objective of this article is to explore whether well-developed and well-implemented online services can make a positive contribution to the future of regional and rural communities. This will be achieved by disseminating some of the learning from the implementation of the MainStreet Regional Portal project (www.mainstreet.net.au). To provide a context for this case study, the next section introduces some theory relevant to virtual communities and portals. The concept of *online communities* is introduced and then literature is reviewed to identify factors that have been acknowledged as important in the success of online community and portal initiatives.

BACKGROUND

In regional Australia, many Web-based initiatives have been premised on fear of external electronic commerce ventures adversely affecting local industry (McGrath & More, 2002, p. 50). Media and government reports have reinforced notions that those who ignore the adoption of electronic commerce will do so at their peril (Department of Communications Information Technology and the Arts, 2000). Recent research however identifies a movement beyond the "starry-eyed fascination with, and high expectations of, technology per se," with the focus now more pragmatically on how ICT can enable enhanced business and community processes and more effective organizational relationships (More & McGrath, 2003).

The term *online community* means different things to different people (Preece, 2000). In early definitions, the term described communication facilitated through bulletin boards (Rheingold, 1994, pp. 57-58). More recent definitions reflect the expansion of Web-based technologies and often link online communities with concepts of regional communities and local strengths (Keeble & Loader, 2001).

In Australia the terms *online community*, *regional portal*, *Web portal*, and *community portal* are often used more or less interchangeably. Web portals "provide focal points on the Web, places to start, places to go to find things" (Gronlund, 2001, p. 88). They have been identified as one strategy for encouraging regional participation in the information economy. For example, according to the Department of Communications Information Technology and the Arts (2001), a regional portal can achieve the online aggregation of potential and existing regional presence into a comprehensive portal, gateway, or regional Web site. In funding initiatives, preference has been given to projects that offer inclusive regional aggregation of business, government, and community services, and which provide interactive services to clients both in and external to the region.

Some definitions of online communities capture the concepts of both *communities of interest* and *communities of location*, and identify the role of encouraging communication and information sharing among members as important (McGrath & More, 2002). Australia's largest telecommunications provider describes online communities as providing a focal point for the provision of local regional information. In terms of functionality, these community portals generally incorporate local news services, local weather reports, a directory of community organizations, and features such as bulletin boards, discussion forums, a calendar of events, and transaction services (Telstra Country Wide, 2002).

To achieve optimum online collaboration, various issues require consideration. These include notions of community, trust and commitment, processes and structure, knowledge management, learning, and collaboration (More & McGrath, 2003). Some further factors more specific to the success of online community or portal initiatives are considered in the next section.

In forging and managing online collaboration, people issues rather than technological ones have been identified as the most challenging. "Certainly across a broad range of projects, many have come to realize that managing people, relationships, and business processes is harder than managing technology" (McGrath & More, 2002, p. 66). It is easy to underestimate the amount of planning and effort that is needed to build and sustain an online community; therefore care should be taken to avoid miscalculations. In particular, "overlooking the key role of the human facilitator is perhaps the greatest reason that online communities fail to meet the expectations of their designers" (Bernal, 2000, p. 4).

For many projects, collaboration is the key to survival, renewal, and growth, especially in regional areas "where the threat of global competitive dynamics often drove alliances" (McGrath & More, 2002, p. 67). Initiatives, however, with a broad geographical focus, can "encounter difficulties in establishing and maintaining cooperative relationships across multiple communities in their regions" (Simpson, 2002, p. 8).

"Many projects that have adopted a 'build it and they will come' approach have been doomed to early failure" (Simpson, 2002, p. 4). Developers need to work with community members to ensure that the goals of the site owner and the needs of community members are met (Preece, 2000). Good online services provide multiple levels of entry, many-to-many relationships, and rapid movement between the services and content of disparate providers (Local Government Association of Tasmania and Trinitas Pty Ltd., 2001).

Community members also need compelling reasons to use and return to an online community again and again. There will be a need to balance supply-side investment (access, technical platforms) and demand-side investment (content and services) (Local Government Association of Tasmania and Trinitas Pty Ltd., 2001).

"If you get this right—if you can identify and fill a need in the lives of your community members—you can go a long way on very little technology. If you miss this, no amount of technology is going to make you successful as an online community." (Kim, cited in Bernal, 2000, p. 3)

Engaging and relevant content are vital to increase uptake and sustained use of the Internet. Portal content management strategies should be *bottom-up* in their approach. This can be achieved by providing multiple opportunities for interaction and by providing permission-based access to software that allows members to produce content for their online community (Brumby, 2001; Telstra Country Wide, 2002).

Soft technologies are also essential in building user confidence and comfort with new technology. "Individualized awareness raising…training activities, and learner support are key elements in creating within the community the desire, motivation, and enthusiasm to trial and take up the technology" (Simpson, 2002, p. 7).

This review has highlighted a number of factors which can impact the success or otherwise of portal type initiatives. This background information provides a context for introducing the MainStreet case study in the next section.

MAIN THRUST OF ARTICLE

In May 1999 a collective of regional stakeholder organizations engaged the University of Ballarat to research the requirements and make recommendations on how the Central Highlands and Wimmera regions of Victoria could capture greater advantages from new information and communications technologies.

The research, documented in *Victoria's Golden West Portal Project Business Case* (Thompson, 1999), involved a number of different stages. These included confirming existing regional Web content, examining community portal

developments, identifying portal tools, researching potential revenue streams, conducting focus group sessions, and other forms of stakeholder consultation.

The research report described how an environment could be established that would be conducive to the widespread adoption of electronic commerce. Specific recommendations included: establish a membership-based regional association with a specific focus on electronic commerce; establish infrastructure for a manageable and economically sustainable Internet presence in a way that would encourage the enhancement of community service and facilitate communities of interest and trading communities; and through a regional portal, achieve better Web content coordination, provide a valuable information source for residents, and also enhance efforts to promote all the attributes of the region.

The Chamber of Electronic Commerce Western Victoria Inc. (the Chamber) was established to facilitate the advancement of electronic commerce and implement the MainStreet portal project. Funding applications were prepared, and in November 1999 the MainStreet project secured funding of AUD 274,000 through Networking the Nation, with a further AUD 135,000 approved in May 2000. The University's Centre for Electronic Commerce and Communications (CECC) was then contracted to implement the project because it had the specialist skills necessary to develop the portal infrastructure and services.

Research had identified that many portal projects had produced 'static' or 'fixed' solutions. The MainStreet model, with the inclusion of a technical team as a critical element, was different, but the decision to have this team was significant in determining how the MainStreet project would evolve. The technical officer and part-time programmers would develop a portal framework based on the core services identified during the preliminary study. All tools would be selected or developed with non-technical end users in mind. The initial toolset would include event calendars; news publishing tools; online registration, payment, and product systems; and access to Web wizards and other Web publishing tools. This would be achieved by incorporating a range of in-house developments, with some integration of externally sourced product. The core services would create capacities to link regional Internet information and services, construct searchable directories, dynamically generate content like news and weather, distribute publishing and authoring rights, and promote community news and events.

The MainStreet project was actively promoted in the period leading up to its official launch in July 2000. This promotion was important as it helped to maintain interest in the project while technical developments proceeded behind the scenes.

During the early part of 2002, the MainStreet project attracted its first major client. Success in securing the Ararat Online project (www.ararat.asn.au) was attributed to involving regional stakeholders right from the project's beginning. Ararat's Economic Development Manager had participated in a range of activities, meetings, and focus group sessions. Through these activities he developed a strong understanding of how MainStreet offered Ararat something different that could be applied immediately to benefit his local community.

The Ararat Online project would include a range of elements with more than 80 businesses and community groups to benefit directly from an upgrade of their Web presence. They would also be given the opportunity to undertake training so that each organization would gain the skills to manage their own site. A further opportunity would be available for six businesses through an e-commerce mentoring program. Selected businesses would be assisted in the implementation of electronic commerce initiatives developed to match their particular business needs.

The value derived from the Ararat Online project was substantial. First, although the project did not represent a significant 'bottom-line' contribution in the context of the overall project budget, the investment of AUD 8,000 in a regional electronic commerce context represented a significant buy-in for the MainStreet product. Second, the Ararat Online project provided an opportunity to showcase the full product suite, its technical capabilities, the Web products, and the training and consulting services. Third, the project would help to address one of the early barriers: people in the target region had a very limited understanding of what a portal was. The Ararat Online project would provide a 'real' example, which it was hoped could be used to demonstrate the value and benefits that were associated with the efficient linking of Internet-based information and services in an easily searchable form. In other words, the Ararat Online project would establish the first 'before' and 'after' images. This proved to be a very powerful marketing mechanism for the project.

The project's technical team, however, had their task doubled—they were now expected to build not one, but two portals, and to deliver these within very short periods. They were successful in developing a way to replicate the MainStreet functionality through Ararat Online (www.ararat.asn.au) and later through projects with the Birchip Cropping Group (www.bcg.org.au), Moorabool Shire (www.mconline.com.au), and Pyrenees Shire (www.pyreneesonline.com.au).

The original goal had been to establish MainStreet as the "point of first electronic contact for the region" (Thompson 1999, p. iv.). The vision was that people would find MainStreet, and from there be able to search and access information about a particular region or locate services of a particular type. What, however, was now understood was that 'communities' were much more motivated if the functionality of MainStreet could be delivered with local Web addresses

and branding. Information could then be filtered up to the MainStreet umbrella so that client communities could be either accessed directly or through MainStreet. While this turned the original concept upside down, there was a strong indication that communities in the region were prepared to pay to both establish and maintain a service based on the 'replicable portal framework' developed through the MainStreet project.

FUTURE TRENDS

The MainStreet portal infrastructure and tools have since been replicated to suit a range of different clients, with this approach proving to be a very effective way of getting people actively engaged online. Appendix 1 contains a selection of URLs for clients including local governments, town-based communities, membership-based organizations, industry groups, and small and medium enterprises.

While a number of factors have been highlighted, the most successful and distinctive aspect has been the development of the replicable portal framework. It has been this capability that has been leveraged to cause increase in 'buy-in', participation, and ongoing investment in regional Web-based services. Members of 'geographic communities' and 'communities of interest' are able to work with CECC to design and implement sophisticated Web-based services, customized to meet their specific communication, promotional, and/or electronic commerce needs. Through this university/community partnership, initiatives are then sustained by putting community members in charge of the management of their online community. Local ownership and the sustainability of infrastructure and technical support services have been achieved by effectively aggregating regional demand for portal services.

The MainStreet project has made a significant contribution to the advancement of ICT and electronic commerce uptake in the Central Highlands and Wimmera regions of Victoria. Many individuals and communities have been assisted in advancing their uptake of electronic commerce as they update their community sites, publish event information and news items, or show others how to build simple Web sites. The level of functionality and services accessed is high and, because clients have strong ownership of their online activities, maintain their own Web-based information, and are committed to annually investing to maintain the portal infrastructure and services, the services can continue to be delivered after the initial seed funding period.

The MainStreet project has also supported and encouraged a staged uptake of electronic commerce, with a number of organizational clients becoming increasingly confident in both selecting and investing in electronic commerce solutions.

Services have also been customized to meet the needs of small groups such as Birchip Cropping Group, and also larger communities such as Moorabool, Ararat, and the Pyrenees Shire regions. This has overcome a barrier where under most models, the costs to establish (and sustain) a local portal have been substantial, and therefore prohibitive for small towns and community groups.

CONCLUSION

Through the MainStreet project, regional and rural communities have a greater ability to build on local strengths and capitalize on the opportunities that are provided by electronic commerce and ICT. Communities, however, just like businesses, require assistance in identifying the most appropriate online service for their particular circumstances. Policies that encourage communities to enter collaborative partnerships, and which leverage existing infrastructure, knowledge and learning, should thus be seen as preferable to the funding or establishment of discrete or stand-alone initiatives. Well-developed and well-implemented online services can make a positive contribution to the future of regional and rural communities. Case studies such as the one presented in this article are effective in illustrating the impacts, influences, and challenges that can be experienced in operationalizing and sustaining online communities in a regional and rural context.

ACKNOWLEDGEMENTS

The author acknowledges Brian West from the University of Ballarat who has been generous in the provision of advice and encouragement that greatly assisted in the preparation of this work.

REFERENCES

Ashford, M. (1999). Online WA: A trickle-up approach to using communications to enhance regional economic and social development. *Proceedings of the Regional Australia Summit,* Canberra, Australia.

Bernal, V. (2000, November). *Building online communities: Transforming assumptions into success.* Retrieved from *benton.org/Practice/Community/assumptions.html.*

Brumby, H.J. (2001). *Connecting communities: A framework for using technology to create and strengthen communities.* State Government of Victoria, Melbourne.

Department of Communications Information Technology and the Arts. (2000). *Taking the plunge: Sink or swim? Small*

business attitudes to electronic commerce. Commonwealth of Australia, Canberra.

Department of Communications Information Technology and the Arts. (2001). *Funding priorities and principles, networking the nation, the commonwealth government's regional telecommunications infrastructure fund.* Commonwealth of Australia, Canberra.

Gronlund, A. (2001). Building an infrastructure to manage electronic services. In S. Dasgupta (Ed.), *Managing Internet and intranet technologies in organizations: Challenges and opportunities.* Hershey, PA: Idea Group Publishing.

Hunter, A. (1999). Opportunities through communications technology for regional Australia. *Proceedings of the Regional Australia Summit,* Canberra.

Keeble, L. & Loader, B.D. (2001). *Challenging the digital divide: A preliminary review of online community support.* CIRA, University of Teesside, UK.

Local Government Association of Tasmania and Trinitas Pty Ltd. (2001). *Online service delivery strategy paper—gaining the maximum benefit for our communities from the local government fund.* Local Government Association of Tasmania, Hobart.

McGrath, M. & More, E. (2002). *Forging and managing online collaboration: The ITOL experience.* National Office for the Information Economy and Macquarie University, Canberra, Australia.

More, E. & McGrath, M. (2003). Organizational collaboration in an e-commerce context: Australia's ITOL project. *The E-Business Review III*, 121-123.

Multimedia Victoria. (2002). *Connecting Victoria: A progress report 1999-2002.* State Government of Victoria, Melbourne.

Municipal Association of Victoria and ETC Electronic Trading Concepts Pty Ltd. (2000). *Local government—integrated online service delivery strategy and implementation plan, executive summary—final.* Municipal Association of Victoria, Melbourne.

National Office for the Information Economy. (2001). *B2B e-commerce: Capturing value online.* Commonwealth of Australia, Canberra.

National Office for the Information Economy. (2002). *The benefits of doing business electronically—e-business.* Commonwealth of Australia, Canberra.

National Office for the Information Economy. (2002). *Guide to successful e-business collaboration.* Commonwealth of Australia, Canberra.

Papandrea, F. & Wade, M. (2000). *E-commerce in rural areas—case studies.* Rural Industries Research and Development Corporation, Canberra.

Pattulock, E. & Albury Wodonga Area Consultative Committee. (2000). *Facilitation of e-commerce and Internet use by regional SMEs.* Albury Wodonga, La Trobe University, Australia.

Porter, M.E. (2001). Strategy after the Net. *BOSS*, (April), 17-23.

Preece, J. (2000). *Online communities: Designing usability, supporting sociability.* Chichester, UK: John Wiley & Sons.

Rheingold, H. (1994). A slice of life in my virtual community. In L.M. Harasim (Ed.), *Global networks: Computers and international communication* (pp. 57-80). Cambridge, MA: MIT Press.

Simpson, L. (2002). Big questions for community informatics initiatives: A social capital perspective. *Search Conference: Community and Information Technology The Big Questions,* Centre for Community Networking Research, Monash University, Melbourne, Australia.

Simpson, R. (1999). Brave new regions. *Proceedings of the Regional Australia Summit,* Canberra, Australia.

Telstra Country Wide. (2002). *Our community online.* Letter and brochure distributed to local government conference delegates, 31 October 2002, Telstra Corporation Limited.

Thompson, H. (1999). *Victoria's Golden West portal project business case.* Centre for Electronic Commerce and Communications, University of Ballarat, Australia.

KEY TERMS

'Bottom-Up'Approach: Development approach founded upon the principle that communities are better placed to coordinate and integrate efforts at the local level.

Case Study: The intensive examination of a single instance of a phenomenon or where one or just a few cases are intensively examined using a variety of data-gathering techniques.

Community Informatics: A multidisciplinary field for the investigation of the social and cultural factors shaping the development and diffusion of new ICT and its effects upon community development, regeneration, and sustainability.

Community Portal: Online initiative often developed through participative processes which aims to achieve better

coordination of relevant Web-based information and provide communication services for community members.

Regional Development: The act, process, or result of actions to grow, expand, or bring a regional place to a more advanced or effective state.

Web Portal: Focal points on the Web which provide a place to start. Web portals facilitate the location of information by incorporating the strengths of search engines and additionally provide more efficient access to information by categorizing it into easily recognizable subcategories or channels.

APPENDIX 1

University of Ballarat URL	
University of Ballarat	www.ballarat.edu.au
CECC	www.cecc.com.au
MainStreet portal URL	
Mainstreet.net.au	www.mainstreet.net.au
Geographical portal URLs examples	
Ararat Online	www.ararat.asn.au
Moorabool Online	www.mconline.com.au
Pyrenees Online	www.pyreneesonline.com.au
Membership based communities URLs examples	
Birchip Cropping Group	www.bcg.org.au
Young Australian Rural Network	www.yarn.gov.au
Rural Regional Research Network	www.cecc.com.au/rrrn
Pyrenees Hay Processors	www.exporthay.com.au
Comprehensive Web site URLs examples	
Ballarat A Learning City	www.ballaratlearningcity.com.au
Central Highlands Area Consultative Committee	www.chacc.com.au
Pyrenees Shire	www.pyrenees.vic.gov.au
Regional Connectivity Project	www.regionalconnectivity.org

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 312-317, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Building Police/Community Relations through Virtual Communities

Susan A. Baim
Miami University Middletown, USA

INTRODUCTION

Over the past two decades, police departments around the globe have been involved in a slow, but steady transition from call-based policing to community-oriented policing. The former approach, while effective at closing cases once a crime has occurred, does little to develop crime prevention partnerships between officers on the beat and the citizens of local communities. Community-oriented policing serves to increase awareness of issues and potential problems before they occur, thus assisting police departments to provide a more proactive approach to stopping crime within their communities.

One of the greatest difficulties in developing effective community-oriented policing programs is establishing solid, two-way communications links between police officers and the populations that they serve. Information flow to the police and suggestions back to the citizenry often fall victim to the same constraints—lack of time to interact effectively and lack of a ready-made mechanism to deliver the information in a timely manner. To reduce or eliminate these constraints, interactive police department Web sites and virtual communities (that involve both police officers and citizens) can provide actionable and measurable performance increases in the efficiencies and the effectiveness of community-oriented policing efforts. Although the IT hardware, software, and design expertise needed to create interactive Web sites and virtual communities are readily available, online efforts at community-oriented policing will remain more of a theoretical interest than a broad-scale application until police departments truly understand the needs and the wants of the citizens within their local communities.

This article explores a service-learning approach for use in a university classroom that combines IT applications with current research practices in the use of citizen satisfaction surveys conducted for local police departments. Examples are drawn from three primary-based research studies involving police departments that are turning away from call-based policing practices and proactively moving toward community-oriented policing practices.

BACKGROUND

Descriptions of community-oriented policing efforts may be found in the literature as early as the 1960s, although the majority of papers published date from the mid-1990s to the present day. Successful community-oriented policing programs began to emerge as departments returned to fundamental cop-on-the-beat policing that put officers back in close contact with citizens in their neighborhoods and in their places of business (Sissom, 1996). The knowledge gained from the early community-oriented policing efforts was used to improve departmental training efforts and also used to focus police officers more closely on crime prevention techniques. Community-oriented policing practices have continued to evolve in more recent studies where most authors focus on how the police can better identify specific issues that are divisive within their respective communities (Culbertson, 2000; Vincent, 1999; Woods, 1999; Rohe, Adams & Arcury, 2001).

Community-oriented policing programs rely heavily on current and ongoing issues of citizen concern received from both the police departments and the citizens of the communities served. The basic premise of community-oriented policing involves both sides becoming very familiar with each other's needs, wants, and expectations, and then forming a true community partnership to create a safe environment for citizens to live and work. Police officers are, in a sense, being asked to enroll in a police version of a basic marketing course in order to learn how to sell this new approach to the residents of their communities (Cummings, 2001). Residents, long accustomed to seeing police officers only when an emergency has been reported, can represent a tough sell for police officers in terms of forming proactive crime prevention citizen partnerships. Additionally, many police departments, themselves, may believe that the extra time and effort necessary to create community-oriented policing programs is not worth the increased costs, given the difficulties in measuring the perceived benefits of crime prevention programs. Crime, itself, is measurable and actionable in terms of police performance. For example, the widespread incorporation of computerized emergency call systems (e.g., 911 in the United States and similar systems in other nations) has given police departments ready access to tools capable

of tracking performance measures such as call volume, time from call to officer arrival, clearance rate of calls, and so forth (Siegel, 1999). Particularly for police departments that score well on these measures and are rewarded appropriately by their city councils and/or their citizenry, the impetus to move toward more time-consuming and less easily quantified community-oriented policing objectives appears to be small. Like many governmental agencies, operational change in police departments tends to be extremely slow and very difficult to implement.

Regardless of these prevalent views, however, one finding that all parties seem to agree on is that the proper incorporation of new computer-based technologies will help police departments get closer to the citizens that they protect and serve. Computer-based technologies also help the individual officer solve crimes at the same time. An excellent example of such a technology is the mobile laptop computer now found in a high percentage of patrol cars (Greenemeier, 2002; Couret, 1999). Officers in the field now have access to virtually the same information as their office-based counterparts, and they can get at that information in real time without the translation losses associated with working through a radio dispatcher. Frequent reliance on the Internet and e-mail for sharing information and communicating between local police departments and other external police agencies also adds to the active network in use by the majority of police departments today. Given the significant improvements in computer technology, the timing is right for police departments to begin to implement community-based policing practices.

The design and development of efficient and effective "customized" community-oriented policing programs clearly places a burden on police departments to solicit, collect, analyze, and interpret data from their citizenries in order to make wise choices regarding the scope and size of any program that is set up. Obtaining high-quality data can be a formidable task, due to the diversity of the population to be sampled and due to the fact that not all respondents share the same expectations regarding active participation in crime prevention with their local police departments. It is with this background framework in mind that modern IT techniques, combined with current research practices, can significantly boost the ability of all parties to communicate and to share information that is critical in moving a police department from call-based policing practices to community-orientated police practices.

SURVEYS, INTERACTIVE WEB SITES, AND VIRTUAL COMMUNITIES

Police departments often lack not only the knowledge of what citizens might respond favorably to in terms of interactive Web sites or virtual communities, but also to the expertise that is needed to conduct unbiased research surveys among their constituencies to generate the required data input. Citizen satisfaction surveys are becoming highly efficient and effective tools for a variety of city government purposes with credible studies cited in the literature over the past several years (Kearney, Feldman & Scavo, 2000; Oleari, 2000). Often, such citizen surveys, conducted among random samples of the community's population, will be highly revealing of the type of information needed to set up initial Web site and/or virtual community structure(s).

To generate useful data that can drive the development of interactive Web sites and virtual communities, it may be necessary to enlist the services of professional researchers. Given large citizen populations, researchers may want to choose conducting either a mail or an online survey. If neither format is selected, they then need to develop a survey instrument, a process plan, and a timeline for conducting the survey, and at the conclusion of the survey, an unbiased set of concrete, actionable, and meaningful recommendations on how to put the data to use. Unfortunately, hiring a professional research firm can cost thousands of dollars that taxpayers cannot afford to spend given today's tight city budgets. A workable alternative, therefore, is to "hire" university students (under the direction of an instructor knowledgeable in current research techniques) who want to have a "hands-on" educational experience that benefits themselves, their university, and their local community in a service learning setting.

The three citizen satisfaction surveys described in this article were conducted following a traditional mail survey format. Table 1 briefly summarizes the city locations, numbers of surveys sent and returned by respondents (including response rate), and the dates that the research studies were conducted over the past few years. In each case, the study was run collaboratively with a police department in Southwestern Ohio (all looking to implement community-based policing practices in their communities) and an undergraduate Marketing course at Miami University. Students in each of the courses developed the database (in Excel or in Access), handled all of the data tabulation, and analyzed the results under the guidance of the author (as instructor of the courses).

Since future citizen satisfaction surveys of this type may be conducted using Internet-based survey instruments in situations where there is reason to believe that a sufficient concentration of "online" respondents is available, citizen respondents were asked to provide data on their Internet usage and their previous access to city and/or police department-sponsored Web sites. Data were also collected on the desirability and projected usage level of advanced interactive Web services, in case these types of services should be offered by police departments at a later date. Selected data are summarized in Table 2.

A careful examination of the data generated across all three surveys reveals several important findings for police

Table 1. Citizen satisfaction studies (mail surveys) involving local police departments

City	No. of Surveys Sent	No. of Surveys Returned	Response Rate	Date Conducted
Middletown, Ohio	2,000	636	32%	Spring 2000
Oxford, Ohio	1,857	522	28%	Spring 2001
Trenton, Ohio	1,600	478	30%	Spring 2001

departments that may desire to set up interactive Web sites and/or virtual communities to enhance police/community relations. These findings were generated by examining the cross-tabulations of the questions noted in Table 2 with various demographic parameters tracked in the surveys and also the respondent verbatims related to Web site usage. Results are summarized qualitatively in Table 3.

- In all three communities surveyed, personal Internet connectivity varied with age group. Older residents were less likely to have Internet access at home, and they did not appear to make up for this lack by using Internet access at work or public Internet access at libraries or other similar locations.
- Among residents with Internet connections, most users felt comfortable using Internet sites as casual observers or to download information, but only a small percentage of individuals were comfortable sending personal information online. Subsequent questions revealed that this reluctance could be attributed to the desire to keep personal information secure—regardless of any printed claims regarding "secure" Web sites or other online privacy statements relayed by Internet providers.
- Even among residents who felt comfortable with two-way communications online or with using interactive Web sites, it is evident that traditional communication mechanisms would have to be maintained in order to meet all needs, especially as they relate to police departments. Most residents believe that traditional means of communication are a "safety net" that must be used when electronic communications go down.

Police departments seeking to increase citizen participation in community-oriented policing efforts through the incorporation of electronic communications mechanisms such as interactive Web sites and/or virtual communities may face a similar situation as outlined above. While it would be too simplistic to state that the general citizenry will not accept such efforts, it is imperative to acknowledge that a significant percentage of city residents may be reluctant to participate in online communications at the present time. As such, it is critically important for those involved in setting up interactive Web sites and/or virtual communities to bear in mind that it may take time for these initiatives to "catch on" and that a keen eye is essential to make the Web sites attractive (with frequent updates) for casual and/or new Internet users.

John Hagel and Arthur Armstrong contend that there is "nothing more uninviting to passing Web traffic than a community without members" (1997, p. 134). Interpreting this comment in the context of community-oriented policing is straightforward. To attract proactive, involved citizens, interactive Web sites and virtual communities must offer something of value—something to cause these citizens to check back frequently, to offer suggestions, and to take away information that is valuable and that cannot be obtained as easily through alternative means. In bringing interactive Web sites and/or virtual communities to fruition, it may be advantageous to encourage members of the police department to participate vigorously and frequently—especially as the sites are first introduced. Some citizens may take a "watch and see" attitude before beginning to participate in online community-based policing efforts. The presence of helpful

Table 2. Citizen satisfaction studies—Internet usage

City	No. of Surveys Returned	Percent of Respondents Using the Internet	Percent of Internet Users Accessing City and/or Police Web Sites	Percent of Internet Users w/Interest in an Upgraded Police Web Site
Middletown, Ohio	636	40%	29%	13%
Oxford, Ohio	522	81%	28%	3%
Trenton, Ohio	478	66%	15%	48%

Table 3. Key findings regarding citizens' use/desire to use the Internet for police services

• In all three communities surveyed, personal Internet connectivity varied with age group. Older residents were less likely to have Internet access at home, and they did not appear to make up for this lack by using Internet access at work or public Internet access at libraries or other similar locations. • Among residents with Internet connections, most users felt comfortable using Internet sites as casual observers or to download information, but only a small percentage of individuals were comfortable sending personal information online. Subsequent questions revealed that this reluctance could be attributed to the desire to keep personal information secure—regardless of any printed claims regarding "secure" Web sites or other online privacy statements relayed by Internet providers. • Even among residents who felt comfortable with two-way communications online or with using interactive Web sites, it is evident that traditional communication mechanisms would have to be maintained in order to meet all needs, especially as they relate to police departments. Most residents believe that traditional means of communication are a "safety net" that must be used when electronic communications go down.

and interesting information from the police department may help draw these citizens from the mode of casual observer to the mode of full-partner participant.

The technology required to offer interactive Web sites and virtual communities is well established, and it will not be addressed specifically in this article. Police departments have numerous choices in how to achieve these goals, ranging from employing the services of professional Web site and virtual community design/support organizations, to "bootlegging" efforts from members of the city workforce and/or police department who are proficient at computer operations. What is mandatory, however, is the list of criteria given in Table 4.

- The Web site and/or virtual community must be in operation 24/7. Long periods of inactivity or excessive "down" periods must be avoided.
- Police officers and/or a department spokesperson(s) must monitor the Web site(s) frequently and actively participate in the citizen/police partnership.
- Continuous solicitation of new members will keep the Web site(s) fresh and productive.
- Police departments must not become discouraged if it takes a relatively long period for the concept to "catch on" in their communities.

Many police departments have a preliminary foundation on which to build an interactive Web site or a virtual community since they already have a police department Web site in place. (Among the three Ohio cities surveyed, Middletown, Oxford, and Trenton, all have police department Web sites or Web pages in an overall directory of city services.) Even if these sites exist only to provide a brief overview of the departments' operations, they can serve as a starting point to build citizen involvement. The key for the developer charged with expanding a basic Web site into an offering capable of engaging the local citizenry in community-oriented policing is to recognize that marketing the partnership process and the end result of the project (in terms of actionable and measurable performance indicators) will be as critical as executing the technical details of the assignment. Considered to be one of the true Internet pioneers, Howard Rheingold stated in an interview that the three most critical things that a developer must determine before beginning a project of this type are: 1) how the site will be marketed, 2) what is expected in return for visits by potential members, and 3) what technologies will be needed to put the site into proper operation (Moore, 2001).

While interactive Web sites can provide a forum for police officers and citizens to exchange information, in general these exchanges are likely to be highly discreet in nature—involving only one citizen and one police officer or department representative at a time. The real opportunity to magnify the positive benefits of community-oriented policing will occur as groups of citizens and officers begin to communicate on a frequent basis. If this is handled electronically, the communications may evolve into a highly functional virtual community. Police departments are well positioned for such an effort because police officers need to interact with the citizenry in order to keep a finger on the pulse of the community. Citizens, on the other hand, often have trouble interacting with the police unless an emergency has occurred. The mindset is that police officers are incredibly busy and that they do not have time to "just chat," when it is actually through such interactions that community-oriented policing makes its mark.

Police departments will need to go the extra mile in establishing virtual communities because individual citizens, or even groups of business people, will be unlikely to set up a virtual community infrastructure with sufficient credibility to attract widespread participation. Police departments may want to borrow guidelines on community participation from

Table 4. Criteria for successful use of interactive Web sites and virtual communities

▪ The Web site and/or virtual community must be in operation 24/7. Long periods of inactivity or excessive "down" periods must be avoided.
▪ Police officers and/or a department spokesperson(s) must monitor the Web site(s) frequently and actively participate in the citizen/police partnership.
▪ Continuous solicitation of new members will keep the Web site(s) fresh and productive.
▪ Police departments must not become discouraged if it takes a relatively long period for the concept to "catch on" in their communities.

authors such as Rachel Gordon, who has published an extensive discussion on "hints for success" in virtual community interactions (Gordon, 2000). Privacy and confidentiality concerns must also be properly addressed. Even though an interactive Web site or a virtual community is considered to be a public forum, clear statements regarding the collection and use of information are warranted (Sheehan, 2002; Sheehan & Hoy, 2000; Milne & Culnan, 2002). Finally, police departments should also keep in mind that they are not setting up interactive Web sites and virtual communities in order to run for-profit businesses. As such, the goal is not to extract data from citizen participants in order to judge how to market revamped and/or new police services. Rather, the goal is to generate critical information and upgrade the ability of both sides to deal with the process of maintaining an environment in which all citizens feel safe to live and work (Brewer, 2000; Wood, 2000).

FUTURE TRENDS

Community-oriented policing is here to stay. For the three Southwestern Ohio communities studied, the benefits of the process far outweigh the anticipated costs and extra effort put forth by their police departments. Data consistently showed that community-oriented policing programs would be greatly appreciated by citizens and business owners alike. Nevertheless, time pressures and resource constraints still continue to plague police departments in the same manner as they do other for-profit businesses, non-profit organizations, and governmental agencies. The key to successful implantation of community-oriented policing efforts thus becomes finding efficient and effective ways to interact with the public at large. Fortunately, the rapid increase in Internet connectivity and the general upswing in personal e-mail and Web usage provide a suitable vehicle for police departments to use in establishing better communication links with the citizenry.

It is highly likely that police departments will continue to increase their presence online in a variety of formats. The relevant issues requiring future study are primarily application oriented and not rooted in the basic technologies of setting up Web sites and/or offering access to virtual communities. Fortunately, police departments can ride the coattails of IT developers who are already generating continuously more advanced Internet communications technologies for business and industrial clients. This is not to say, however, that engaging the citizenry of a community to communicate with the police online is an easy or well-understood proposition. The state of the art today is one of relatively infrequent successes. Solid, two-way communications take time and money to implement, and both issues are at a premium in most police departments—regardless of whether those communications take place through traditional channels or in an online format.

Citizen satisfaction surveys, conducted through a mail survey format as described in this article or conducted in a fully online research mode, are likely to grow in importance as police departments continue to focus more energy on meeting the specific needs and wants of their communities. In terms of service learning at the university level, surveys of this type are natural service learning projects for post-secondary students in business, sociology, criminal justice, and other similar majors where a quantitative and qualitative study of police/citizen interactions is of value. Educators are likely to find students highly willing to participate in such work, and police departments eager for the unbiased "third-party" perspective that a well-run research study can provide.

Future research should focus on identifying more precisely what information citizens would value receiving through interactive Web sites and/or virtual communities. Individuals charged with developing interactive Web sites and/or virtual communities for police and citizen use would also be well advised to consider surveying police officers to better elucidate their department's needs and wants before beginning the design process. Collectively, the issues surrounding the "true" needs and wants of citizens in a local community transcend the technical aspects and challenges of mainstream research in information technology to include additional aspects of public policy, criminal justice, and citizens' rights. Citizen satisfaction surveys are, nevertheless, one form of practical application toward the future direction of applied information technology solutions.

CONCLUSION

Interactive Web sites and virtual communities represent two of the most innovative ways to generate meaningful, two-way dialogs over the Internet. As the technologies for connecting multiple Internet users in these manners mature, information technology researchers and developers can turn significant attention toward solving the specific application problems posed by unusual clients. Connecting the citizenry of a community with their police department in an efficient and effective manner is just one of a growing number of novel applications made possible by advanced Web site design and virtual community hosting technologies. As community-oriented policing efforts continue to grow in the coming years, it is highly likely that the efforts of information technology professionals will play a critical role in their success—making all of us feel safer at home and at work.

REFERENCES

Brewer, C. (2000). Community is the fly paper of your site. *Computer User, 18*(12), 49.

Couret, C. (1999). Police and technology. *The American City & County, 114*(9), 31-32+.

Culbertson, H.M. (2000). A key step in police-community relations: Identifying the divisive issues. *Public Relations Quarterly, 45*(1), 13-17.

Cummings, B. (2001). NYPD meets marketing 101. *Sales and Marketing Management, 153*(4), 14.

Gordon, R.S. (2000). Online discussion forums. *Link-up, 17*(1), 12.

Greenemeier, L. (2002). Sacramento cops take e-tools on the beat. *Information Week, 886,* 60.

Hagel III, J. & Armstrong, A.G. (1997). *Net gain*. Boston: HBR Press.

Kearney, R.C., Feldman, B.M. & Scavo, C.P.F. (2000). Reinventing government: City manager attitudes and actions. *Public Administration Review, 60*(6), 535-547.

Milne, G. & Culnan, M. (2002). Using the content of online privacy notices to inform public policy: A longitudinal analysis of the 1998-2001 U.S. Web surveys. *The Information Society, 18,* 345-359.

Moore, R. (2001). Focus on virtual communities. *B to B, 86*(7), 14.

Oleari, K. (2000). Making your job easier: Using whole system approaches to involve the community in sustainable planning and development. *Public Management, 82*(12), 4-12.

Rohe, W.M., Adams, R.E. & Arcury, T.A. (2001). Community policing and planning. *Journal of the American Planning Association, 67*(1), 78-90.

Sheehan, K. (2002). Toward a typology of Internet users and online privacy concerns. *The Information Society, 18,* 21-32.

Sheehan, K. & Hoy, M. (2000). Dimensions of privacy concern among online consumers. *Journal of Public Policy & Marketing, 19*(1), 62-73.

Siegel, F. (1999). Two tales of policing. *Public Interest, 134,* 117-121.

Sissom, K. (1996). Community-oriented policing means business. *FBI Law Enforcement Bulletin, 65*(12), 10-14.

Vincent, E. (1999). How citizens' voices are heard in Jacksonville. *Sheriff Times, 1*(10). Retrieved October 10, 2003, from *www.communitypolicing.org/publications/shtimes/s10_fa99/s10vince.htm*

Wood, J.M. (2000). The virtues of our virtual community. *Instructor, 110*(1), 80-81.

Woods Jr., D.D. (1999). Supervising officers in the community policing age. *Sheriff Times, 1*(10). Retrieved October 10, 2003, from *www.communitypolicing.org/publications/shtimes/s10_fa99/s10woods.htm*

KEY TERMS

Call-Based Policing: Traditional policing approach whereby officers respond to emergency calls for assistance and address crimes and other situations after the fact.

Citizen Satisfaction: Term coined to describe the overall approval rating of services received by citizens within their communities. A 100% rating indicates total satisfaction with services received. Ratings may be taken "per service" or represent satisfaction with an entire government structure.

Community-Oriented Policing: Contemporary policing approach that builds relationships between police officers and the citizens of a community on an ongoing basis. Crime prevention is stressed as a partnership approach before an actual emergency situation(s) develops.

Interactive Web Site: A Web site or page configured so as to invite correspondence between the user and the originator/sponsor of the site. Such sites go beyond pas-

sively providing information to those who browse the site. Customarily, there are options to complete online surveys, send e-mail to the sponsor, request specialized or personalized response(s), and so forth.

Internet Privacy: Concerns expressed by Internet users regarding the security and confidentiality of information transmitted electronically. Government agencies and business/industry professional groups share responsibility to address Internet privacy concerns.

Internet-Based Survey: A contemporary survey technique through which researchers may obtain respondents' opinions via online survey processes. Respondents may either be asked to go to a Web site to complete a survey (Web-based) or the survey questionnaire may be e-mailed to the respondents (e-mail-based) for them to complete and return electronically.

IT Applications: Research and development work performed to create a situation-specific bridge between new or existing IT hardware and software technologies and the information needs/wants of a customer. The combination of proper hardware, software, and tailored application delivers a well-rounded IT solution for the customer's problem.

Mail Survey: A traditional survey technique in which a multi-part survey questionnaire is mailed to a randomized sample of individuals (within a larger population) who are asked to complete the questionnaire and return it to the survey researcher for tabulation and analysis.

Service Learning: Educational projects structured to take students out of the traditional lecture-based classroom and involve them in "real-world" problem-solving opportunities of importance to their communities. Students apply theories learned in their classroom studies in a practical manner that generates a completed work product for one or more clients, while helping to cement in students' minds the usefulness of the theoretical concepts under study.

Virtual Community: An online forum or discussion group through which members may interact either in real time or asynchronously. Most virtual communities use discussion groups and message boards that are accessible online to all members. Members correspond by posting messages back and forth within the forum. Membership in a virtual community indicates that the user shares one or more common interests with others in the same forum.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 318-324, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Software & Systems Design

Building Secure and Dependable Online Gaming Applications

Bo Chen
Cleveland State University, USA

Wenbing Zhao
Cleveland State University, USA

INTRODUCTION

Online gaming has become a multibillion-dollar industry. The security and dependability of such games are critical for both the game providers and honest game players alike. Essential to all such applications is the use of random numbers; for example, random numbers are needed to shuffle cards. For obvious reasons, if the hands can be predicated, players could gain unfair advantages. The nature of this type of applications poses great challenges in increasing their availability while preserving their integrity (Arkin, Hill, Marks, Scjmod, & Walls, 1999; Viega & McGraw, 2002; Young & Yung, 2004).

Byzantine fault tolerance (BFT; Castro & Liskov, 2002) is a well-known technique to tolerate various malicious attacks to online systems and it often involves state machine replication (Schneider, 1990). However, state machine replication assumes that all replicas are deterministic, which is not the case for online gaming applications. In this article, we elaborate how we address this dilemma using an online poker application that uses a pseudorandom number generator (PRNG) to shuffle the cards as an illustrating example. We propose two alternative strategies to cope with the intrinsic application nondeterminism. One depends on a Byzantine consensus algorithm and the other depends on a practical threshold signature scheme. Furthermore, we thoroughly discuss the strength and weaknesses of these two schemes.

BACKGROUND

In this section, we provide a brief introduction of PRNG, the entropy concept, and the methods to collect and enhance entropy.

A PRNG is a computer algorithm used to produce a sequence of pseudorandom numbers. It must be initialized by a seed number and can be reseeded prior to each run. The numbers produced by a PRNG are not truly random because computer programs are in fact deterministic machines. Given the same seed, a PRNG will generate the same sequence of numbers. Consequently, if the seed is known to an adversary, then one can simulate the stream of random numbers (Young & Yung, 2004). Therefore, to make the random numbers unpredictable, it is important that the seeds to the PRNG cannot be guessed or estimated. Ideally, a highly random number that is unpredictable and infeasible to be computed is required to seed the PRNG in order to produce a sequence of random numbers.

The activity of collecting truly random numbers is referred to as collecting entropy by cryptographers (Young & Yung, 2004). Entropy is a measure of how much real randomness is in a piece of data, for example, using the outcome of coin flipping as 1 bit of entropy. If the coin toss is perfectly fair, then the bit should have an equal chance of being a 0 or a 1. In such a case, we have a perfect 1 bit of entropy. If the coin flip is slightly biased toward either head or tail, then we have something less than a bit of entropy. Entropy is what we really want when we talk about generating numbers that cannot be guessed. In general, it is often difficult to figure out how much entropy we have, and it is usually difficult to generate a lot of it in a short amount of time.

It is a common practice to seed a PRNG with the value of the local clock. Unfortunately, this is not a sound approach to preserve the integrity of the system, as described by Arkin et al. (1999) when telling about how they attacked a Texas Hold'em Poker online game. They show that with the knowledge of the first few cards, they can estimate the seed to the PRNG and subsequently predict all the remaining cards.

TOWARD SECURE AND DEPENDABLE ONLINE GAMING APPLICATIONS

In this section, we describe two possible strategies for building secure and dependable online gaming applications. One depends on a Byzantine consensus algorithm and the other depends on a practical threshold signature algorithm. These two algorithms are aimed at ensuring that all replicas use the same value to seed their PRNGs, with each replica taking entropy from its respective entropy source.

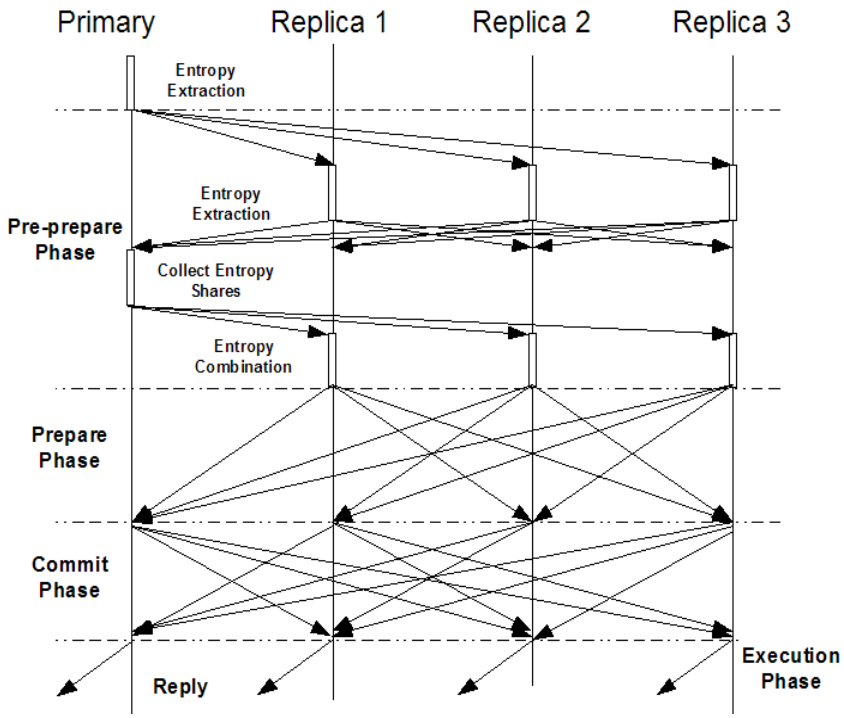
Byzantine Fault Tolerance

A Byzantine fault (Lamport, Shostak, & Pease, 1982) is a fault that might bring a service down or compromise the integrity of a service. A Byzantine faulty replica may use all kinds of methods to prevent the normal operation of a replicated service; in particular, it might propagate conflicting information to other replicas. To tolerate f Byzantine faulty replicas in an asynchronous environment, we need to have at least $3f+1$ replicas (Castro & Liskov, 2002). An asynchronous environment is one that has no bound on processing times, communication delays, and clock skews. Internet applications are often modeled as asynchronous systems. Usually, one replica is designated as the primary and the rest are backups.

We choose to use the well-known Byzantine fault tolerance algorithm developed by Castro and Liskov (2002). The BFT algorithm has three communication phases in normal operation. During the first phase, the *pre-prepare* phase, upon receiving a request from the client, the primary assigns a sequence number and the current view number to a message and multicasts this *pre-prepare* message to all backups. In the second phase, referred to as the *prepare* phase, a backup broadcasts a *prepare* message to the rest of the replicas after it accepts the *pre-prepare* message. Each nonfaulty replica enters into the *commit* phase, that is, the third phase, only if it receives $2f+1$ *prepare* messages (from different replicas) that have the same view number and sequence number as the *pre-prepare* message, then it broadcasts the *commit* message to all replicas including the primary. A replica commits the corresponding request after it receives $2f$ matching *commit* messages from other replicas. To prevent a faulty primary from intentionally delaying a message, the client starts a timer after it sends out a request and waits for $f+1$ consistent responses from different replicas. Due to the assumption that at most f replicas can be faulty, at least one response must have come from a nonfaulty replica. If the timer expires, the client broadcasts the request to all replicas. Each backup replica also maintains a timer for similar purposes. On expiration of their timers, the backups initiate a view change and a new primary is selected. In the BFT algorithm, a digital signature or an authenticator is employed to ensure the integrity of the messages exchanged, and a cryptographic hash function is used to compute message digests.

The above BFT algorithm must be modified to cope with intrinsic replica nondeterminism. The modified algorithm also consists of three phases, as shown in Figure 1. In the beginning of the first phase, the primary invokes the ENTROPY-EXTRACTION operation to extract its entropy and append the entropy to the *pre-prepare* message. It then

Figure 1. The adapted BFT algorithm to handle entropy extraction and combination



multicasts the *pre-prepare* message to the backups. Each replica records the primary's entropy from the *pre-prepare* message in its log and then invokes the ENTROPY-EXTRACTION operation to obtain its own share of entropy as well. Each backup then multicasts a *pre-prepare-update* message, including its share of entropy extracted. When the primary collects 2*f* *pre-prepare-update* messages from the backups, it constructs a *pre-prepare-update* message, including the digest of the 2*f*+1 entropy shares (2*f* received, plus its own) together with the corresponding contributor's identity, and multicasts the message.

Upon receiving the *pre-prepare-update* message from the primary, each replica invokes the ENTROPY-COMBINATION operation to combine the entropy from the 2*f*+1 shares. The outcome of the ENTROPY-COMBINATION operation ensures a highly random number due to the contributions from the nonfaulty replicas. The combined number is provably secure and will be used to seed the PRNG if the BFT algorithm terminates.

The second and third phases are similar to the corresponding phases of the BFT algorithm, except each replica will append the digest of the entropy set determined in the first phase to the *prepare* and *commit* messages. These two phases are necessary to ensure that all correct replicas agree on the same message total ordering and the entropy value despite the presence of Byzantine faulty replicas.

We now highlight the details of the ENTROPY-EXTRACTION and ENTROPY-COMBINATION operations.

Entropy-Extraction. The ENTROPY-EXTRACTION operation is based on software-based entropy collection. There are a number of techniques that can be used to extract the entropy, most of which are based on the timing of internal activities in a computer (Young & Yung, 2004). A well-known technique is called TrueRand, developed by Don Mitchell and Matt Blaze. The idea behind TrueRand is to gather the underlying randomness from idle CPUs by measuring the drift between the system clock and the generation of interruptions on the processor. Other frequently used techniques include recording network traffic as it comes into the server, timing how long it takes to seek the disk, and capturing kernel state information that changes often.

Entropy-Combination. The ENTROPY-COMBINATION operation combines the 2*f*+1 entropies the replica collected using the exclusive-or (XOR) operator (Young & Yung, 2004). This operation has several benefits.

First, it combines a number of weak sources of entropy to form an effective strong entropy source. Consider two entropy sources from coin flipping, and the case in which Source 1 results in head and Source 2 results in tail, or Source 1 results in tail and Source 2 results in head. Now consider that the probability for Source 1 to result in head is 10:16 and that for Source 2 is 12:16 (i.e., both are biased). If we combine these two sources, the probability of getting a head is 7:16. This shows that the coin flipping resulting from XORing the bits from the two sources is the same or better than the best flip in either of the two sources. Furthermore, the more sources we use, the higher entropy we get.

Second, the ENTROPY-COMBINATION operation eliminates any negative impact from malicious replicas. Among the 2*f*+1 entropy shares collected, there can be up to *f* of them coming from faulty replicas. The XOR operation guarantees that if at least one share comes from a good entropy source, the combined entropy is at least as good as that entropy share. This requirement is met because there are at least *f*+1 shares contributed by nonfaulty replicas. Any low-entropy or predicable shares generated by faulty replicas are virtually ignored.

Third, the ENTROPY-COMBINATION operation results in a single high entropy share used by all nonfaulty replicas, which ensures the consistency of the replicas when they are involved with intrinsically nondeterministic operations.

Threshold Signature

The other strategy to ensure consistent Byzantine fault tolerance replication for nondeterministic operations is based on threshold cryptography (Desmedt & Frankel, 1990; Deswarte, Blain, & Fabre, 1991). Threshold cryptography is a good way to distribute trust among a group of players to protect either information or computation (Zhou, Schneider, & Renesse, 2002).

A well-known secret sharing scheme (Shamir, 1979) is the (*k*, *n*) threshold digital signature scheme. In the (*k*, *n*) secret sharing scheme, a secret is divided into *n* sets and distributed to the same number of players. The secret can be reconstructed if *k* out of *n* players combine their shares. However, fewer than *k* players cannot collude to forge the secret. The (*k*, *n*) threshold digital signature scheme allows a set of servers to collectively generate a digital signature in a way similar to reconstructing a secret using the (*k*, *n*) secret sharing scheme.

In the (*k*, *n*) threshold digital signature scheme, a private key is divided into *n* shares, each owned by a player. A valid threshold digital signature can be produced if *k* players pool their shares. However, no valid signature can be generated by fewer than *k* players. Each player uses its private-key share to generate a partial signature on a message, and these partial signatures can be combined into a threshold signature on the message. The threshold signature can be verified using the public key corresponding to the divided private key.

The RSA Shoup scheme (Shoup, 2000) is one of the practical threshold digital schemes. In this scheme, a dealer generates a key pair and divides the private key into *n* shares at first. Each key share has a key verifier. Then the dealer distributes the message to be signed and the key shares to *n* players. Each player uses its key share to generate a partial signature on the message. Furthermore, each player sends the signed message with the verifier to a trusted server, which veri-

fies the signature shares and combines the partial signatures into a threshold signature verifiable by the public key.

In the following, we show how to integrate the threshold digital signature scheme with Byzantine fault tolerance for online gaming applications. The adapted BFT algorithm consists of three phases (under normal operation) for Byzantine agreement and an additional phase run at the beginning for key shares distribution. The Byzantine agreement algorithm works similar to the BFT algorithm except for the third phase, where each replica generates a partial signature (using its key share) to sign the client's message and piggybacks the partial signature to the *commit* message. Each replica combines the partial signatures into a threshold signature. The signature is then mapped into a number to seed the PRNG.

Despite the elegancy of the threshold signature, the algorithm, however, might not be practical in the Internet environment. First of all, it depends on a trusted dealer at the beginning to generate a key pair and divide the private key into several key shares, and it must also be responsible for distributing the key shares to all replicas. If the dealer is compromised, the entire system can be easily penetrated by the adversary. Furthermore, the threshold signature is computationally expensive, especially when generating the threshold signature. For example, for a 1,024-bit threshold signature, it usually takes 73.9ms on a PC equipped with a single 1.0GHz Pentium III CPU and 1.5 GB RAM (Rhea, Eaton, Geels, Weatherspoon, Zhao, & Kubiatowicz, 2003). Furthermore, the validity on the use of the threshold signature as the seed to the PRNG remains to be proven secure.

FUTURE TRENDS

PRNGs are not only used for online gaming applications, but are widely used in nearly every field in computers and networking. In particular, for cryptography, access to truly random numbers is extremely important. Even though there is moderate success in implementing PRNG, it remains vulnerable under cryptanalytic attacks and attacks against its internal state. Furthermore, it is easy to see that even if the PRNG is robust against many potential threats, once the seed is discovered, the numbers generated by the PRNG are no longer unpredictable. In light of this observation, more efforts should be engaged in how to gather and evaluate entropy in a secure and dependable manner. The research described in this article can be regarded as the first step toward this direction.

There are many open issues that remain to be resolved before we can confidently apply these techniques in practice. The most interesting research issue is how to maintain replica consistency. Common Byzantine fault tolerance techniques require deterministic execution of replicas despite the fact that all practical applications contain some degree of nondeterminism (Zhao, 2007), for example, clock values, CPU speed, multithreading, and so forth. (Note that these types of nondeterminism are not considered good entropy sources according to the cryptography standard, but the presence of these types of nondeterminisms nevertheless poses a big threat to maintaining replica consistency.) The Byzantine fault tolerance framework must be able to handle various nondeterministic applications in a systematic and efficient manner.

Another interesting research issue is regarding the performance of Byzantine fault tolerant applications. Besides the communication cost for achieving Byzantine fault tolerance, digital signing operations are computationally expensive. In a recent study (Chen, Hsiao, & Chen, 2004), it is shown that the cost of cryptographic operations can be reduced by using short secret keys of the elliptic curve cryptosystem and by enabling simultaneous signing.

CONCLUSION

In this article, we pointed out the threats to online gaming applications and presented two strategies that can be used to build secure and dependable online gaming applications. These strategies not only seek out solutions for gathering entropy to seed the PRNG used in such applications, but also intend to eliminate malicious intrusions to protect the seed and to maintain replica consistency. By applying these techniques, online gaming applications can ensure their service integrity (both the service providers and the innocent players are protected) and guarantee high availability despite the presence of Byzantine faults. Finally, we outlined some open research issues in this field.

REFERENCES

Arkin, B., Hill, F., Marks, S., Scjmod, M., & Walls, T. (1999). *How we learned to cheat at on-line poker: A study in software security.* Retrieved from http://www.developer.com/java/other/article.php/10936_616221

Castro, M., & Liskov, B. (2002). Practical Byzantine fault tolerance and proactive recovery. *ACM Transactions on Computer Systems, 20*(4), 398-461.

Chen, T., Hsiao, T.-C., & Chen, T.-L. (2004). An efficient threshold group signature scheme. In *Proceedings of the IEEE Region 10 Conference*, Chiang Mai, Thailand (Vol. B, pp. 21-24).

Desmedt, Y., & Frankel, Y. (1990). Threshold cryptosystems. In *Lecture notes in computer science* (Vol. 435, pp. 307-315).

Deswarte, Y., Blain, L., & Fabre, J. (1991). Intrusion tolerance in distributed computing systems. In *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Oakland, CA (pp. 110-121).

Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems, 4*(3), 382-401.

Luby, M. (1996). *Pseudorandomness and cryptographic applications.* Princeton University Press.

Rhea, S., Eaton, P., Geels, D., Weatherspoon, H., Zhao, B., & Kubiatowicz, J. (2003). Pond: The OceanStore prototype. In *Proceedings of the Second USENIX Conference on File and Storage Technology*, San Francisco (pp. 1-14).

Rivest, R., Shamir, A., & Adleman, M. (1978). A method for obtaining digital signatures and public key cryptosystems. *Communications of the ACM, 21*(2), 120-126.

Schneider, F. (1990). Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Computer Survey, 22*(4), 299-319.

Shamir, A. (1979). How to share a secret. *Communications of the ACM, 22*(11), 612-613.

Shoup, V. (2000). Practical threshold signature. In *Lecture notes in computer science* (Vol. 1807, pp. 207-223).

Viega, J., & McGraw, G. (2002). *Building secure software: How to avoid security problems the right way.* Addison-Wesley.

Young, A., & Yung, M. (2004). *Malicious cryptography: Exposing cryptovirology.* Indianapolis, IN: Wiley Publishing.

Zhao, W. (2007). Byzantine fault tolerance for nondeterministic applications. In *Proceedings of the Third IEEE International Symposium on Dependable, Autonomic and Secure Computing*, Columbia, MD (pp. 108-115).

Zhou, L., Schneider, F., & Renesse, R. (2002). COCA: A secure distributed online certification authority. *ACM Transactions on Computer Systems, 20*(4), 329-368.

KEY TERMS

Dependable System: A dependable system is one that is trustworthy to its users. It requires that the system be highly available (to legitimate users) while ensuring a high degree of service integrity.

Digital Signature: A digital signature aims to serve the same purposes as a real-world signature. A sound digital signature ensures that the sender of the digital signature can be authenticated, the sender cannot later repudiate that she of he has sent the signed message, and a receiver cannot forge a digital signature (without being detected).

Entropy: Entropy is a metric used to evaluate and describe the amount of randomness associated with a random variable.

Entropy Combination: It is the operation that combines a number of entropy shares into one. The combination is usually achieved by using the XOR operator. Entropy combination is an effective defense against adversaries that substitute a random value with a predictable one. The combined entropy is often of higher quality than each individual share.

Entropy Extraction: This is the operation that extracts entropy from a random variable (referred to as the entropy source). Entropy can be extracted using both software- and hardware-based methods.

Pseudorandom Number Generator (PRNG): A PRNG is a computer algorithm used to produce a sequence of pseudorandom numbers. It must be initialized by a seed number and can be reseeded prior to each run. The numbers produced by a PRNG are not truly random. Given the same seed, a PRNG will generate the same sequence of numbers.

Threshold Digital Signature: In the (k, n) threshold digital signature scheme, a private key is divided into n shares, each owned by a player. A valid threshold digital signature can be produced if k players combine their shares. However, no valid signature can be generated by fewer than k players. Each player uses its private key share to generate a partial signature on a message, and these partial signatures can be combined into a threshold signature on the message. The threshold signature can be verified using the public key corresponding to the divided private key.

Building Wireless Grids

Marlyn Kemper Littman
Nova Southeastern University, USA

INTRODUCTION

The accelerating implementation and remarkable popularity of sophisticated mobile devices, including notebook computers, cellular phones, sensors, cameras, portable GPS (Global Positioning System) receivers, and wireless handhelds such as PDAs (personal digital assistants), contribute to development of wireless grids. **Wireless grids** feature a flexible and adaptable cyberinfrastructure that supports coordinated and economical access to distributed resources and next-generation applications and services.

Generally, **wireless grids** are classified as ad hoc or standalone, and mixed-mode or hybrid. Ad hoc **wireless grids** enable diverse applications via MANETs (mobile ad hoc networks) and consist of mobile devices that operate in infrastructureless environments. Mobile network nodes process tasks and provide best effort delivery service to support wireless grid applications (Lima, Gomes, Ziviani, Endler, Soares, & Schulze, 2005). In the healthcare environment, for example, ad hoc **wireless grids** equipped with sensors monitor the status of critically ill patients and track the location of hospital equipment and supplies. Hybrid or mixed-mode **wireless grids** augment and extend the capabilities of wireline grids to remote locations; facilitate the shared use of resources and processing power; and consist of components ranging from supercomputers to distributed or edge devices such as very small satellite aperture terminals (VSATs) (Harrison & Taylor, 2006).

This chapter features an introduction to factors contributing to the development of present-day **wireless grids**. Wireless grid technical fundamentals, specifications, and operations are examined. Security challenges associated with safeguarding wireless grids are reviewed. Finally, the distinctive characteristics of innovative wireless grid initiatives are explored and research trends in the wireless grid space are described.

BACKGROUND

Established by **virtual organization**s (**VO**s), **wireline grids** facilitate trusted resource exchange in environments that cross multiple administrative domains. **Wireline grids** consist of substantial collections of shared networked components and resources for enabling reliable and dependable multimedia delivery; implementation of sophisticated bandwidth-intensive applications; scientific discovery in fields that include high-energy physics, medicine, astronomy, and bioinformatics; and e-collaborative problem resolution (Littman, 2006).

Wireline grids increasingly employ a high-performance **DWDM (Dense Wavelength Division Multiplexing)** cyberinfrastructure that supports wavelengths of light, or lambdas, on demand to facilitate reliable and dependable access to computational simulations, metadata repositories, large-scale storage systems, and clusters of supercomputers (Littman, 2006). Also called lambda-grids, **DWDM** grids, such as the TeraGrid, enable terabit and petabit transmission rates; teraflops and petaflops of compute power; seamless connectivity to feature-rich resources; and extendible grid and inter-grid services across multi-institutional distributed environments.

The popularity of multifunctional 3G (third generation) wireless technologies and devices and demand for anytime and anywhere access to grid resources are major factors contributing to design and implementation of **wireless grids** by **mobile dynamic virtual organizations (MDVOs)**. **MDVOs** are extensions of **VOs** that facilitate wireless grid deployment in infrastructureless and hybrid wireless environments. Wireless grid operations are dependent on network node mobility (Waldburger, Moraiu, Racz, Jahnert, Wesner, & Stiller, 2006). As a consequence, wireless grids are not as reliable as wireline grids in seamlessly and dependably supporting multimedia applications and services.

Wireless grid transmissions are impaired by signal fading, packet delay, and the absence of bandwidth to support applications requiring quality of service (QoS) guarantees. The performance of **wireless grids** is also affected by network node power consumption, the quality of the wireless medium, and the effectiveness of wireless grid security solutions. To counter these limitations, toolkits such **as GT4** (**Globus Toolkit version 4**); security mechanisms providing services such as authentication, authorization, data confidentiality, and data integrity; mobile agent systems; and WSs (Web Services) are utilized in building dependable wireless grid solutions.

CONSTRUCTING WIRELESS GRIDS

With **MDVO** implementation of wireless grids in new geographic environments, the numbers of wireless devices

and individuals that are served by wireless grids increase dramatically. A widely used toolkit supporting wireline and wireless grid implementations, **GT4** facilitates resource allocation, detection of connectivity, and location of specific resources in response to user requests (Littman, 2006). **GT4** also enables integration of wireless devices, applications, and services with low bandwidth and variable computational power into hybrid grid configurations. By working in conjunction with middleware architecture such as SIGNAL (Scalable Intergrid Network Adaptation Layers) and **GT4,** mobile devices in hybrid grid environments can support P2P (peer-to-peer) operations that enable each peer or node to function as a client and content provider. P2P implementations in wireless grid environments also aid in resolving problems associated with resource availability and battery power constraints.

Based on P2P, the P2P Discovery Protocol (P2PDP) coordinates distribution of grid tasks via MoGrid (mobility grid) architecture. MoGrid devices function as task processing nodes and interfaces to wireline grids (Lima et al., 2005). P2PDP separates requests for grid processes, such as file exchange and job execution, into a series of tasks that are processed concurrently, sequentially, or independently.

As a consequence of difficulties in enabling disconnected processes and asynchronous communications in wireless grid environments and the need to minimize the amount of energy consumed by each application, mobile agent technology is emerging as an effective enabler of reliable job execution in wireless grids (Zhang & Lin, 2007). In addition to supporting distributed applications, mobile agent architecture provides a framework for completion of reliable and dependable grid processes.

As an example, the Mobile Agent-based Collaborative Virtual Environment (MACVE) features agile mobile agent architecture for enabling 3-D (three-dimensional) applications in shared large-scale virtual environments. Distinguished by their ability to migrate autonomously among geographically distributed nodes to facilitate resource discovery, resource scheduling, and node monitoring, MACVE mobile agents also optimize grid performance and resource utilization across multiple **MDVO** domains (Zhang & Lin, 2007). Additionally, MACVE mobile agents also support development and management of VRML (Virtual Reality Modeling Language) 3-D applications, and enable load balancing and security services in wireless grid environments (Lin, Neo, Zhang, Huang, & Gay, 2007).

Another option for building wireless grid operations and facilitating dependable access to multimedia resources is utilization of network proxies that support localized storage and computational services. For example, MAPGrid (Mobile Applications on Grids) employs network proxies to ensure resource availability in wireline grid nodes that are situated near mobile devices. Proxies offload tasks and support intermittent proactive data caching to ensure the availability of resources for wireless grid multimedia applications (Huang & Venkatasubramians, 2007).

Wireless Grid Operations

Ad hoc and mixed-mode wireless grid functions are dependent on signal strength, protocols, and middleware for supporting dependable operations and performance (Li, Sun, & Ifeachor, 2005). Ad hoc **wireless grids** supported by a wireless cyberinfrastructure enable data transport between network nodes that are within transmission range of each other. Wireless links are temporary since grid nodes are limited in physical size, storage capabilities, and available bandwidth. As a consequence, algorithms and protocols, such as AODV (Ad Hoc Distance Vector) are used to facilitate dynamic routing between wireless grid nodes and support ad hoc grid functions. In contrast to ad hoc wireless grids, hybrid **wireless grids** are scalable and capable of providing faster transmission rates.

Problems in wireless grid operations typically stem from poor signal strength and cyberinfrastructure instability. Since mobile nodes join and leave wireless grids randomly, intermediate nodes also relay data (Li et al., 2005). Information exchange among mobile devices on **wireless grids** are adversely impacted by latency, inadequate numbers of nodes to support transmissions, multipath signal distortion, battery degradation, and fast handoffs. Inconsistencies in the power of wireless devices and lack of assurance of resource availability also contribute to delays in processing wireless grid applications (2005.) To counter these constraints, mobile agents and P2P (peer-to-peer) paradigms are utilized to foster robust, secure, and extendible wireless grid solutions.

Wireless Grid Specifications

WS (Web service) specifications consist of collections of open standards and protocols that support wireline and wireless grid construction and implementation of wireless grid applications. Developed by the Global Grid Forum (GGF), the **Open Grid Services Architecture** (**OGSA**) is based on a suite of WS standards for optimizing resource integration and resource management in evolving wireless grid environments. **OGSA** also promotes implementation of loosely coupled interactive services to facilitate grid operations, and defines common interfaces to support wide-scale grid initiatives (Harrison & Taylor, 2006).

Core components and design elements initially defined for **OGSA** are now integrated in the **Open Grid Services Infrastructure (OGSI).** The WS Resource Framework (WSRF) builds on OGSI functions originally developed by the GGF. **OGSI** specifies requirements for interface development, resource scheduling, and resource management to optimize wireline grid operations. Based on OGSI and mobile agent technology, Mobile **OGSI**.NET enables

integration of wireless devices into conventional wireline grid configurations by caching data to counter intermittent grid connectivity and, thereby, facilitating implementation of grid applications on resource-constrained wireless devices (Chu & Humphrey, 2004).

In addition to distributing tasks to mobile agents, MGSs (Mobile Grid Services) coordinate mobile agent communications and management. JADE (Java Agent Development Framework) and **GT4** provide a framework for supporting MGSs. MGSs also extend static grid services to dynamic environments and support resource control and accounting operations; coordination of mobile agent communications and management; and identification of improper resource utilization resulting from cyberincursions (Wong & Ng, 2006). The J2ME (Java2 Platform for Micro Edition) specifies standard grid APIs (application programming interfaces) for integrating wireless devices into wireline grid solutions.

A next-generation lightweight middleware toolkit, MAGE (Modular and Adaptable Grid Environment) utilizes reconfigurable component architecture with unified management interfaces to support scalable and adaptable grid operations in wireless environments as well (Kwon, Choi, & Choo, 2006). Featuring a service-oriented architecture (SOA) with customizable components, MAGE facilitates deployment of reconfigurable wireless grid operations including the removal, modification, and addition of grid components. SOA also enables wireless grid scalability and the interoperability of wireline and wireless grid applications.

In SP2A (Service-Oriented Peer-to-Peer Architecture) grids, resources are accessed through RPS (Resource Provision Services) (Amoretti, Zanichelli, & Conte, 2006). The RPS framework features a Java API that supports JXTA (juxtaposition) for P2P routing, WS deployment, and OWL-S (WSs-Ontology Web Language) utilization (Jabisetti & Lee, 2005). A semantic markup language for WSs, OWL-S establishes semantic service descriptions for autonomous WS operations, such as resource discovery, and facilitates WS execution in diverse grid environments. SP2A also defines modules for building service-oriented peers (SOPs) to foster resource sharing in **wireless grids**. Mobile agents built with the Java toolkit JGRIM (Java Generalized Reactive Intelligent Mobility) enable development of security services and service-oriented grid (SOG) applications that foster resource discovery, resource brokering, and application execution.

Wireless Grids in Action

Akogrimo (Access to Knowledge through the Grid in a Mobile World), GROW-Net (Grid Reconfigurable Optical Wireless Network), FWGrid (Fast Wired And Wireless Grid), and LOOKING (Laboratory for the Ocean Observatory Knowledge Integration Grid) are examples of innovative **wireless grids** in actions. The capabilities, applications, and services supported by these initiatives are examined in this segment.

Developed by a **MVDO**, **Akogrimo** facilitates next-generation wireless grid implementations by leveraging resources available via a large base of mobile devices across the European Union (Waldburger et al., 2006). **Akogrimo** utilizes 3G wireless technologies; IPv6 (Internet Protocol Version 6); mobile Internet architecture; and knowledge-related and semantics-driven **WSs** to prototype next-generation wireless grids. **Akogrimo** also enables e-learning, e-health, and disaster recovery applications; teleworking in the automotive, aerospace, and engineering industries; and development of privacy, security, and trusted services in the wireless grid environments.

GROW-Net features a hybrid wireline and wireless cyberinfrastructure consisting of a **DWDM** network backbone that works in concert with wireless mesh networks (WMNs) (Shaw, Gutierrez, Kim, Cheng, Wong, Yen, et al., 2006). A flexible solution for extending a wireline grid cyberinfrastructure to distributed locations, **GROW-NET** WMNs consist of static and mobile network nodes; feature multiple wireless access points; and use adaptive routing protocols and algorithms to optimize network performance.

Established by the University of California at San Diego, **FWGrid** supports development of grid middleware, distributed applications architecture, and a hybrid cyberinfrastructure to support current and emerging grid applications. The wireless component of the cyberinfrastructure facilitates access to video and images collected from wireless devices at rates ranging from 100 Mbps (megabits per second) to 1 Gbps (gigabit per second). The wired cyberinfrastructure component supports transmissions at rates ranging from 10 Gbps to 100 Gbps and features distributed computational clusters with teraflops of computational capabilities (McKnight, Howison, & Bradner, 2004).

Built on a mixed-mode wireline and wireless cyberinfrastructure, **LOOKING** interlinks ocean observatories into a knowledge grid that enables scientists to monitor oceanographic phenomena in real time. Data collected by undersea equipment, including motion, acoustical, biological, and optical sensors and sophisticated instrumentation, are transmitted to wireline grids at onshore institutions for analysis and assessment (Geer, 2006).

Research findings enable scientists to predict sea conditions; forecast impending natural disasters such as tsunamis; contain wildfires; and provide disaster recovery services. **LOOKING** employs a SOA and the WSRF to enable instrumentation management and support an array of applications across the distributed system of ocean observatories (Geer, 2006).

LOOKING functions comply with IEEE (Institute of Electrical and Electronics Engineers) 1451 wireless sensor interface standards (Geer, 2006). These specifications

establish uniform APIs that are independent of underlying physical media for enabling deployment of standards-compliant sensors in diverse wireless configurations including IEEE 802.11 WLANs (wireless LANs) and IEEE 802.15.1 WPANs (wireless personal area networks).

Wireless Grid Security

Mobile devices that form wireless grids are at risk to cyberincursions as a consequence of resource limitations, unreliable and lost connections, wireless link instability, and software incompatibilities. As with wireline grids, wireless grids enable access to distributed resources in shared, interconnected, and distributed computing environments, thereby, placing the security of these resources at risk (Littman, 2006). Therefore, **MDVOs** establishing wireless grids conduct risk assessments to determine vulnerabilities and support development of security policies, procedures, and mechanisms to safeguard the integrity of wireless grid operations. **MDVOs** also are responsible for promoting the use of encryption protocols such as the Lightweight Security Protocol (LiSP) (Park & Shin, 2004). LiSP features reliable cryptographic key distribution, a stream cipher for fast processing, and cryptographic hash algorithms to generate new keys to facilitate secure, pervasive, and reliable access to shared resources.

RESEARCH TRENDS

Advances in wireless technology contribute to development of wireless grids that support access to the rich mix of resources and services available in static or wired and wireless grids. For instance, small-scale **wireless grids,** consisting of motes or miniaturized transceivers that provide dynamic data on pollution levels, seismic activity, and changes in oceanic environment, are in development. In hospitals, small-scale wireless grids utilize sensors to monitor patients with conditions that include sleep apnea and congestive heart failure (Gaynor, Moulton, Welsh, LaCombe, Rowan, & Wynne, 2004). By contrast, large-scale grids, such as the Southern California Integrated GPS Network (SIGN) and the GPS Earth Observation Network System (GEONET), feature earth stations that generate terabits of data on applications that measure water quality and monitor air pollution (Aydin, Zhigang, Pierce, Bock, & Fox, 2007).

Formed by vehicles equipped with 3G wireless devices and technologies, vehicle-to-vehicle communications grids (VGrids) operate in conjunction with VANETs (vehicle ad hoc networks) (Gerla, Zhou, Lee, Soldo, Lee, & Marfia, 2006). VANETs are next-generation networks that support content sharing and P2P applications in VGrid environments. Designed to save lives and facilitate distributed traffic management services, VGrids provide real-time access to information on vehicular collisions, emergency evacuations, and disaster recovery services.

Approaches for enabling first responders to use **wireless grids** for emergency services in response to artificial disasters, such as terrorist attacks and natural disasters including floods, typhoons, hurricanes, and volcanic explosions, are in development as well. Capabilities of **wireless grids,** such as community resource grids (CRGs) in supporting medical assistance and seamless access to critical healthcare resources, are evaluated in testbed implementations.

CONCLUSION

The accelerating implementation of **wireless grids** reflects the popularity of mobile devices such as PDAs, smart phones, and laptops; the widespread adoption and convergence of wireless and cellular technologies; and demand for connectivity to wireless grid resources at any time and from any place. Challenges associated with wireless grid implementations include overcoming the adverse impact of atmospheric disturbances; reduction in power consumption, bandwidth and response time constraints; and limitations associated with the use of small wireless devices. Current research in the wireless grid space involves development of mechanisms for enabling dependable QoS guarantees; WMNs to support wireline and wireless grid applications; and policies and procedures to facilitate seamless, secure, and reliable applications and functions in ad hoc or infrastructureless and mixed-mode or hybrid wireless grid environments

REFERENCES

Amoretti, M., Zanichelli, F., & Conte, G. (2006). SP2A: A service-oriented framework for P2P-based grids. In *Proceedings of the 3rd International Workshop on Middleware for Grid Computing***.** Grenoble, France.

Aydin, G., Zhigang Q., Pierce, M., Bock, Y., & Fox, G. (2007). Building a sensor grid for real time global positioning system data. In *Proceedings of Workshop on Principles of Pervasive Information Systems Design in conjunction with Pervasive 2007*. Toronto, Ontario, Canada.

Chu, D., & Humphrey, M. (2004). Mobile OGSI.NET: Grid computing on mobile devices. In *Proceedings of the 5th IEEE/ACM international Workshop on Grid Computing*. Pittsburgh, PA, USA.

Gaynor, M., Moulton, S., Welsh, M, LaCombe, E., Rowan, A., & Wynne, J. (2004). Integrating wireless sensor networks with the grid. *IEEE Internet Computing, 8(*4), 32-39.

Geer, D. (2006). LOOKING at the other three-fourths of the

world, through ORION. *IEEE Distributed Systems Online, 7*(4), article number 0406-04005.

Gerla, M., Zhou, B., Lee, Y-Z., Soldo, F., Lee, U., & Marfia, G. (2006). Vehicular grid communications: The role of the Internet infrastructure. In *The Second Annual International Wireless Internet Conference.* Boston, MA., USA.

Harrison, A., & Taylor, I. (2006). Service-oriented middleware for hybrid environments. In *Proceedings of the 1st International Workshop on Advanced Data Processing in Ubiquitous Computing.* Melbourne, Australia.

Huang, Y., & Venkatasubramanians, N. (2007). Supporting mobile multimedia applications in MAPGrid. *Proceedings of the 2007 International Conference on Wireless Communications and Mobile Computing* (pp. 176-181). Honolulu, Hawaii, USA.

Jabisetti, N., & Lee, Y. (2005). OWL-S based autonomic services for grid computing. In *Proceedings of the 2005 IEEE International Conference on Web Services*, 826. Digital Object Identifier (DOI): 10.1109/ICWS.2005.89

Li, Z., Sun, L., & Ifeachor, E. (2005). Challenges of mobile ad-hoc grids and their applications in e-healthcare. In *Proceedings of the 2nd International Conference on Computational Intelligence in Medicine and Healthcare.* Lisbon, Portugal.

Lima, L., Gomes, A., Ziviani, A., Endler, M., Soares, L., & Schulze, B. (2005). Peer-to-peer resource discovery in mobile grids. In *Proceedings of the 3rd International Workshop on Middleware for Grid Computing.* Grenoble, France.

Lin, Q., Neo, H., Zhang, L., Huang, G., & Gay, R. (2007). Grid-based large-scale Web3-D collaborative virtual environment. In *Proceedings of the Twelfth International Conference on 3-D Web Technology* (pp. 123-132). Perugia, Italy.

Littman, M. K. (2006). Implementing DWDM lambda-grids. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (2nd ed.). Hershey, PA: IGI Global (formerly Idea Group, Inc).

McKnight, L., Howison, J., & Bradner, S. (2004). Wireless grids: Distributed resource-sharing by mobile, nomadic, and fixed devices. *IEEE Internet Computing, 8*(2), 2-10.

Messig, M., & Goscinski, A. (2007). Autonomic system management in mobile grid environments. In *Australian Symposium on Grid Computing and Research.* Ballarat, Australia.

Park, T., & Shin, K. (2004). LiSP: A lightweight security protocol for wireless sensor networks. *ACM Transactions on Embedded Computing Systems, 3*(3), 634-660.

Shaw, W-T., Gutierrez, D., Kim, K., Cheng, S-H., Wong, S-H., Yen, S-H., et al. (2006). GROW-Net – A new hybrid optical wireless access network architecture. In *Proceedings of the JCIS-2006.* DOI:10.2991/jcis.2006.316.

Waldburger, M., Moraiu, C., Racz, P., Jahnert, J., Wesner, S., & Stiller, B. (2006). *Grids in a mobile world: Akogrimo's network and business views. Technical Report No. 2006.05.* Zurich, Switzerland: University of Zurich Department of Informatics.

Wong, S-W., & Ng, K-W. (2006). Security support for mobile grid services framework. In *International Conference on Next Generation Web Services Practices* (pp. 75-82).

Zhang, H., & Lin, G. (2007). MACVE: A mobile agent based framework for large-scale collaborative virtual environments. *Presence: Teleoperators and Virtual Environments, 16*(3), 279-292.

KEY TERMS

Cyberinfrastructure: Advanced network platform that supports wireless and/or wireline grid research initiatives, applications, and experimentation. Used by the National Science Foundation (NSF) to describe next-generation grid initiatives.

Dense Wavelength Division Multiplexing (DWDM): Provides flexible and dynamic optical lightpaths on-demand to support extendible, scaleable, and reliable wireline grid and intergrid services and supports access to next-generation applications requiring high-bandwidth connections.

Global Toolkit Version 4 (GT4): A grid toolkit that supports grid interoperability and works in concert with the Web Services Resource Framework (WSRF).

Lambda: Lightpath or wavelength of light that interlinks two network endpoints.

Network node: An endpoint or redistribution point for data transport.

Web Services (WSs): Collections of protocols and open standards for enabling the convergence of WSs and wireless and/or wireline grid operations.

Web Services Description Language (WSDL): Defines an XML (extensible markup language) grammar for describing network services as collections of network endpoints that enable information exchange.

Wireless Grid: A cyberinfrastructure that interconnects wireless devices in ad hoc or infrastructureless and hybrid or mixed-mode wireline and wireless grid configurations.

Business Informatization Level

Ronaldo Zwicker
University of São Paolo – Brazil, Brazil

Cesar Alexandre de Souza
University of São Paolo – Brazil, Brazil

Antonio Geraldo da Rocha Vidal
University of São Paolo – Brazil, Brazil

INTRODUCTION

IT diffusion is central to the new economy and is reflected in a process of informatization of society and businesses. Although initially coined to represent the diffusion and adoption of information technology (IT) in all levels of society, the term informatization is also employed to represent the use of information technology resources in organizations. Weissbach (2003), for instance, defines *informatization* as being the process of gradual and increasing application of "planned and systematic use of IT penetrating the organization's functions". As pointed out by Lim (2001), the evaluation of an organization's *Informatization Level (IL)* is an important managerial concern. The author also points out the difficulties associated with this evaluation, stating that "this is not a simple problem because informatization includes many intangible factors such as the quality of information and the organization's culture". The purpose of evaluating a company's IL is to provide information for the organization to improve precisely its informatization level. It is also a means of benchmarking the efficacy and efficiency of IT investments in order to set up the baseline for improvement.

This topic depicts a measurement method for the IL of companies and shows results of its application in 830 Brazilian industries (Zwicker, Vidal, & Souza, 2005). The development of this method was based on the principle that IT results in companies are not obtained merely through investments and the implementation of systems but rather through its proper use in business processes. The proposed method extends the informatization dimensions proposed by Lim (2001), using the process-based view of the IT business value creation model proposed by Soh and Markus (1995) and the concept of "information systems coverage" proposed by Ravarini, Tagliavini, Buonanno, and Sciuto (2002).

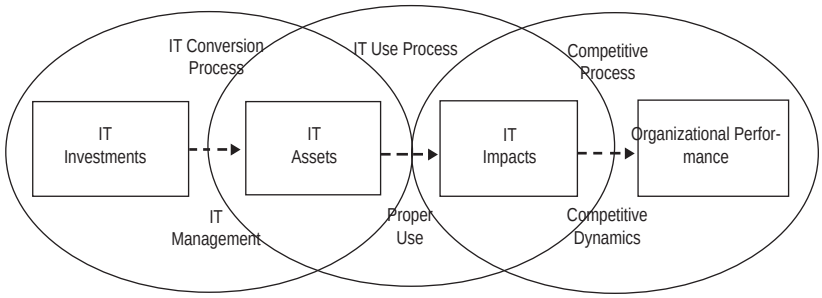
BACKGROUND

Accordingly to Hu and Quan (2005), there are four main visions in studies focusing the creation of *IT business value* through the use of the technology: the macroeconomic view, that believes that IT creates excess returns over other types of capital investments; the process-based view that believes that IT investments create competitive advantages by improving operational efficiency of intermediary processes; the resource-based view, that believes that IT investments create sustainable competitive advantage via unique, immobile, and path-dependent strategic resources and capabilities; and the digital option view that argues that IT investment creates values by giving options and flexibility for firms. Since the study is focused on the organizational level, the second and third views (process- and resource-based) are deemed more adequate.

Soh and Markus (1995) present a model that synthesizes concepts of other studies that also incorporate the process-based view (Grabowski & Lee, 1993; Lucas, 1993; Markus & Soh, 1993). Figure 1 represents the sequence of events and results associated to the process of obtaining organizational benefits from IT investments, according to this model. To obtain an improvement in the organization's performance by means of IT requires that IT impacts occur in the intermediary processes of the organization. However, the fact that impacts from intermediary processes were obtained is not sufficient to obtain organizational performance improvements, since this depends on external factors, such as the economic context and competition. These aspects comprise the "competitive process" of the model and must consider the requirements of the competitive dynamic in which the company is inserted.

On the other hand, to obtain the impacts of IT on the organization's processes requires that IT assets be available, that is: systems in operation, implemented infrastructure and people with suitable knowledge concerning the technology and its possibilities. IT assets constitute a combination of IT resources, applications, and the qualification of people

Figure 1. IT and business value creation (Soh & Markus, 1995)



(from the IT area and users). Human resources also include the partnership relationship with the users. Thus, IT assets are divided into tangible assets (hardware, software) and intangible assets (knowledge, relationship).

It is worth noting that the mere existence of IT assets does not necessarily imply that IT impacts will be obtained. It is necessary to consider the actual use of these assets, which comprises the "process of IT use", and meeting the requirements of "proper use" of these assets. The proper use refers to the effective application of the IT assets in the organization's activities and processes. In considering the appropriate use of IT, one must take into account: its extent (scope of business tasks performed with IT support), its intensity (volume of use), and the level of IT dependence that is imposed to the company.

Finally, the consolidation of IT assets calls for a compatible level of IT investments. However, the investments do not assure that effective assets will be obtained, since these investments can be made inappropriately. Weill (1992) defines the capacity of converting IT investments into IT assets as the "conversion effectiveness" and states that this results from the aspects of the organizational atmosphere that influence IT, the quality of IT management, and the company's commitment to IT. The effective transformation of IT investments into IT assets constitutes the *"process of IT conversion"* which, to be effective, broadly speaking, requires meeting IT management requirements.

INFORMATIZATION LEVEL MODEL

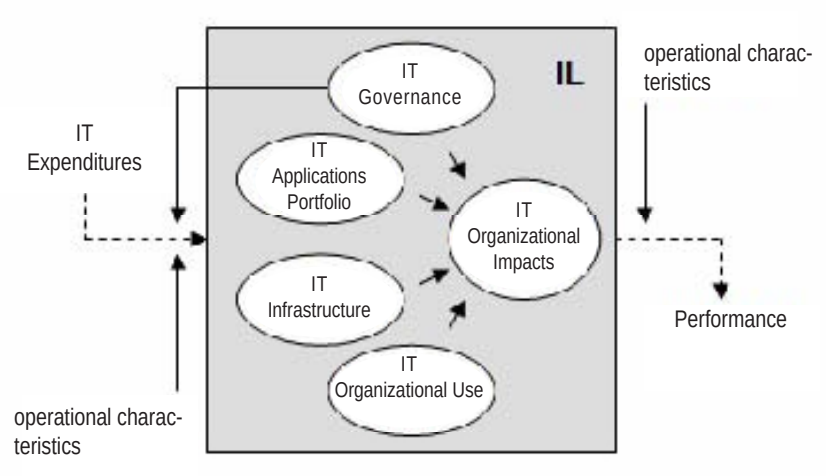
Based on the previous discussion, the concept of informatization, its dimensions, and the IL model adopted are established. Informatization can be defined as the managed process by which an organization continuously expands its IT assets and extends and deepens their appropriate use, aimed at improving the effectiveness and performance of its activities and processes. The five dimensions proposed for the IL measure in organizations are: (1) *IT Infrastructure*, related to fundamental IT assets; (2) *IT Applications Portfolio*, related to tangible information systems resources and intangible aspects of these resources; (3) *IT Organizational Use*, related to the extent and intensity of IT use in the organization; (4) *IT Governance*, related to the management of IT resources, the management of its use, and the planning and development of IT resources aligned with the organization's businesses; (5) *IT Organizational Impacts*, related to effectiveness and performance benefits for organizational activities and processes, achieved through the use of IT. The dimensions appear in the IL model at Figure 2.

In Figure 2, IT expenditures correspond to the sum of investments in IT and expenses with IT (IT expenses are related to monthly or periodic expenses like payroll, maintenance, and telecommunications). They are a necessary although insufficient condition for achieving certain levels of informatization, as indicated by the dotted arrow to the left of the figure. A better IL can contribute to the improvement of the organization's performance, although not necessarily so, as indicated by the dotted arrow to the right. The operational characteristics of the enterprise (for example, its size and the sector in which it operates) appear mediating the relationships between the IT expenditures, the IL, and the organization's performance. The characteristics of IT governance also interfere in the conversion of IT expenditures into a certain level of informatization. The dimensions "IT organizational use", "IT governance", "IT infrastructure", and "IT applications portfolio" work in combination, for the obtainment of "IT organizational impacts". The set of these five dimensions comprise the measurement structure proposed for the IL.

INFORMATIZATION LEVEL MEASUREMENT METHOD

The study gathered data from a sample of 830 Brazilian industrial companies that originated 66 indicator variables

Figure 2. Informatization level model



used to compute the IL. For each dimension, the data said respect to the following aspects (subdimensions):

IT Infrastructure

Infrastructure services, internal connectivity, and external connectivity were assessed through indicators like PCs by employee, number of connected PCs, Internet connection speed, and security strategies (Weill & Broadbent, 1998). *IT department infrastructure* was based upon the existence of a formal IT area, existence of an IT manager, number of IT tasks carried out, and indicators like IT personnel by PC (Lim, 2001).

IT Applications Portfolio

Systems integration was measured along a scale going from use of spreadsheets to fully-integrated ERP systems for each area of application. *Systems technical quality* was computed starting from the technical quality of systems in each area of application (sales, marketing, production, procurement, management, finance, and other IT applications).

IT Organizational Use

Extension of use and other applications use were assessed using the concept of "information systems coverage" (Ravarini et al., 2002). Several activities were defined for each area of application and the extension of use was measured considering the percentage of these activities held with information systems support. The same was done with a group of specific applications (Internet, CRM, SCM, CAD, BI), representing "other applications". *Dependency of use* was examined through the degree of dependency of the enterprise on the information systems in each of the considered activities. *Systems adequacy* was assessed through respondents' perception of the degree of adequacy of these systems to the activities they support.

IT Governance

IT planning and control used questions assessing the respondents' perception about IT strategic alignment and control of IT activities. *Users and executives participation and knowledge* used questions assessing the respondents' perception about the degree of participation of users and executives in IT planning and their knowledge about IT and systems in use in the company.

IT Organizational Impacts

Impacts of traditional applications and impacts of other applications were considered through questions assessing IT impacts on the organization (Mahmood & Soon, 1991). The questions were focused on the following items: sales increase, cost reduction, quality increase of products and services, delivery time reduction, and impacts on processes specific to each area of application and to the group of other applications (Internet, CRM, SCM, CAD, BI).

The IL and its dimensions are variables that are not directly observable. They constitute the latent variables of a structural equation model. The IL is modeled as a third order variable, measured by means of second order variables, which are measured by means of first order variables that

are finally measured trough indicators constructed from the observed variables. It was proposed as an initial model for IL and through the use of structural equation modeling (SEM) emerged the final model presented at Figure 3. The model was tested by means of a confirmatory factorial analysis (CFA), which derived the factorial loads shown in the figure and that confirms the appropriateness of the used indicators to the structure of this IL's model. The factorial loads reflect how well a specific dimension or subdimension reflects the IL.

The computation of the IL for each company is done starting from the factorial scores of each company calculated with SEM, and the obtained value is transformed into a scale from 0 to 100, whose limits are obtained from the minimum and maximum values of IL's obtained for the sample. Outliers are previously analyzed and handled. Moreover, for each one of the dimensions is obtained a specific value, also in the interval from 0 to 100 and that reflects the position of the company according to this dimension. The factorial scores of each company measure the relative position of the company in each one of these constructs. Values that correspond to aggregates, for example, aggregates by dimension, size, or total, are computed using the simple average of the companies ILs that compose the aggregate.

Figure 4 shows total IL average value along the companies' size. As was expected, the informatization level grows with size, staying the micro and small-size companies below the general average and the medium and large-size companies above the general average. The general average equals 51.0. The IL can also be analyzed observing its components in an individualized way, which is analyzing separately the several dimensions of the indicator. Figure 4 represents the averages for each company size, along the dimensions that compounds the IL. The dimension values accompany IL's general evolution, growing with the increase of the company size. The dimension that presents the smallest proportional increase along company size is the governance dimension. This may appear strange because the smaller companies tend to not own a formalized IT area, and therefore, the associate indicator for governance should be inexpressive in the smaller companies. However, the executives of the smaller companies are nearer to the computerization process and consequently tend to directly "care for" the aspects related to this inquiry.

However, the fact that IL of bigger companies is higher does not mean that there are no micro- or small-size companies with high IL. Figure 5 shows the IL distribution for each company size along 4 ranges: *low* with IL between 1 and 25; *regular* with IL between 26 and 50; *average* with IL between 51 and 75; and *high* with IL between 76 and 100. From this data, we can conclude, for example, that 3.3% of the micro and 15.5% of the small size companies hold a high informatization level.

Figure 3. CFA Model for the IL with standardized factor loadings

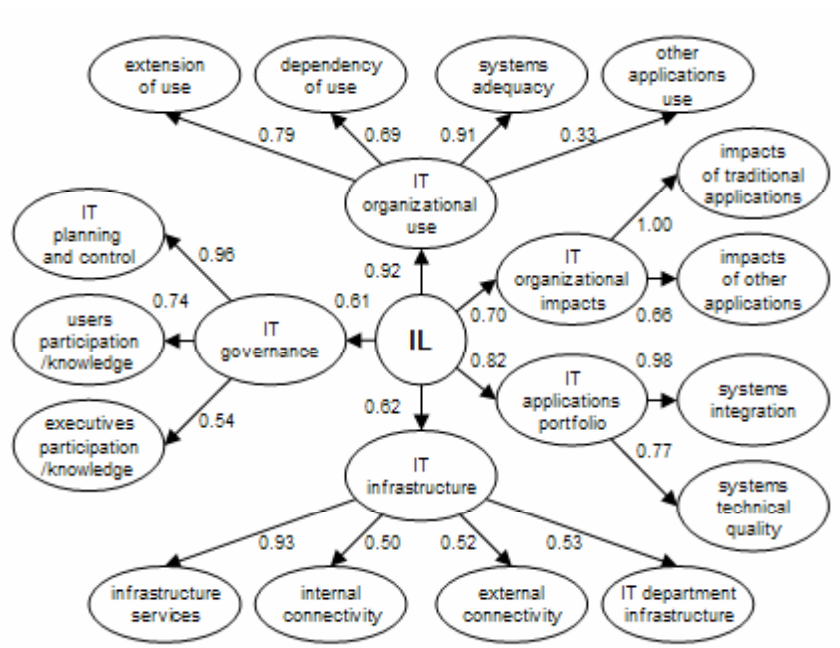


Figure 4. IL averages of each dimension and total per company size

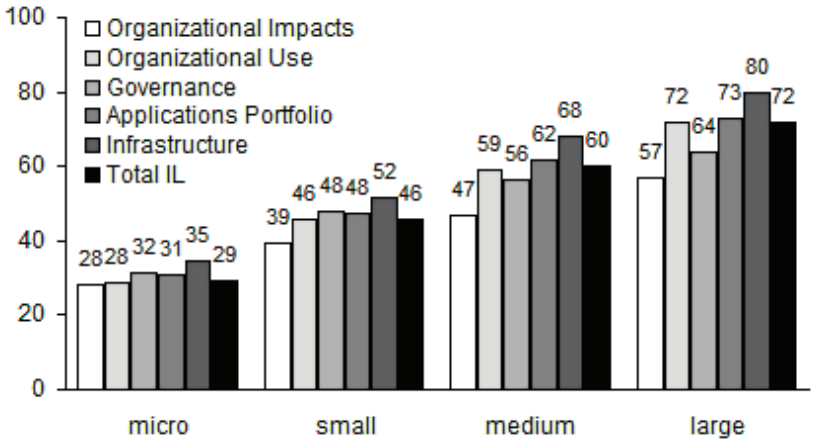


Figure 5. IL distribution per company size

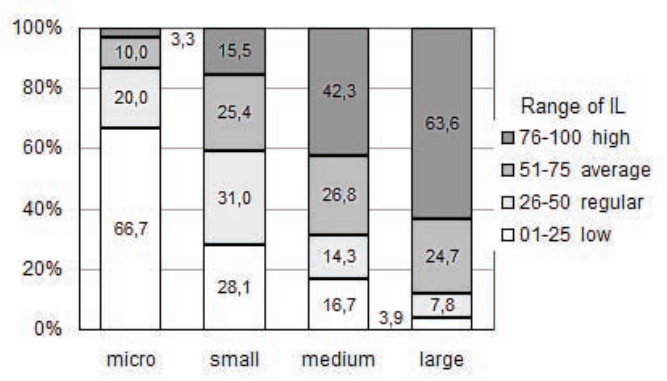


Figure 5 show that there still exists opportunity for IT overall improvements especially in micro-, small- and medium-size companies. The data of the study also showed that the micro- and small-size companies with high IL obtained that result investing proportionally less into IT than the large-size companies with similar IL. This occurs due to two factors. First, the sophistication of the technology needed for the computerization of large-size companies is greater, which conducts to more elevated investments and expenses. Second, at small-size companies, it is possible to reach more rapidly a greater percentage of employees along the computerization process because the number of persons in these companies is smaller. Also is conceivable that the computerization of micro- and small-size companies, in function of the relative volume of jobs they offer, contributes decisively to the "digital inclusion" of the whole society.

FUTURE TRENDS

Future perspectives include converting the IL measure into an auto-evaluation and learning tool for the companies and deepening the study of the relationship of the IL with other aspects of IT use in the companies, for example, (a) the relationship between IL and company size; (b) the relationship between IL and company's business process complexity; (c) the relationship between IL and IT investments; (d) the relationship between company size and the conversion capacity of IT expenditures into a greater IL.

These relationships refer to the question of IT adoption and the results that IT can and should propitiate. In this sense, the gathered data evidenced that the larger is the IL; the larger is the recognized impact to the performance of the companies. This perception is very clear in the case of the

larger companies because they own more people involved with IT use and maintain relationships with other companies through the technology. At a lesser degree, the same result is verified at the smaller companies. This corroborates Premkumar (2003) which establishes that the utility of the technology increases in the proportion at which more people and companies adopt it, therefore, network externalities also condition the increment of IT use and its success. This also agrees with McKeen and Smith (1993), which show that greater informatization levels reinforce the relation between IT use and the performance of the company. From the gathered data also was possible to recognize a "stage effect"; that is, to achieve significant IT benefits, it is necessary that a certain IT investments level, and therefore an informatization level, must be previously reached.

CONCLUSION

The IL includes the aspects considered relevant and that involve the intensity and quality of IT use in the companies. The IL is supported by the theory of IT use in organizations particularly on the value creation model for the businesses through information technology. The development of the measurement was performed seeking to fulfill the requirements of reliability and internal and external validity. Concerning the reliability and internal validity, we obtained adequate fit values, coherent values for the model's coefficients and adequate values for the reliability indexes. Also for external validity, the obtained results allowed comparisons to be made between companies of the same size, between different sizes and between different industrial sectors, producing results that are in line with what was expected.

This suggests that the rationale behind the IL can be generalised, and IL's computation can become a standard measurement methodology to evaluate IT use. Objectively, the IL measurement may contribute toward the identification and analysis of factors that can support companies to reach their proper informatization level. An indicator of the kind of IL can contribute to informatization processes as it supplies an auto-evaluation instrument and more including comparison parameters for the companies. The IL also can supply subsidies for other researches in the scope of the real value of the information technology and of the relation of the use of this technology with the issue of the productivity of the companies.

REFERENCES

Grabowski, M., & Lee, S. (1993) Linking information systems application portfolio and organizational strategy. In R.D. Banker, R.J. Kauffman, & A. M. Mahmood (Eds.), *Strategic information technology management.* Hershey, PA: Idea Group Publishing.

Hu, Q., & Quan, J. (2005) Evaluating the impact of IT investments on productivity: A causal analysis at industry level. *International Journal of Information Management, 25*(2), 39-53.

Lim, S. K. (2001) A framework to evaluate the informatization level. In W. V. Grembergen (Ed.), *Information technology evaluation: Methods & management.* Hershey, PA: Idea Group Publishing.

Lucas, H. C. (1993) The business value of information technology: A historical perspective and thoughts for future research. In R. D. Banker, R. J. Kauffman, & A. M. Mahmood (Eds.), *Strategic information technology management.* Hershey, PA: Idea Group Publishing.

Mahmood, M. A., & Soon, A. (1991) A comprehensive model for measuring the potential impact of information technology on organizational strategic variables. *Decision Sciences, 22*(1), 870-897.

Markus, M. L., & Soh, C. (1993). Banking on information technology: Converting IT spending into firm performance. In R.D. Banker, R. J. Kauffman, & A. M. Mahmood (Eds.), *Strategic information technology management.* Hershey, PA: Idea Group Publishing.

McKeen, J. D., & Smith, H. A. (1993) The relationship between information technology use and organizational performance. In R. D. Banker, R. J. Kauffman, & A. M. Mahmood (Eds.), *Strategic information technology management.* Hershey, PA: Idea Group Publishing.

Premkumar, G. (2003) A meta-analysis of research on information technology implementation in small business. *Journal of Organizational Computing and Electronic Commerce, 13*(2), 91-121.

Ravarini, A., Tagliavini, M., Buonanno, G., & Sciuto, D. (2002). Information systems check-up as a leverage for SME development. In S. Burgess (Ed.), *Managing information technology in small business: Challenges and solutions.* Hershey, PA: Idea Group Publishing.

Soh, C., & Markus, M. L. (1995) How IT creates business value: A process theory synthesis. In *Proceedings of the Sixteenth International Conference on Information Systems.* Amsterdam.

Weill, P. (1992) The relationship between investment in information technology and firm performance: A study of the valve manufacturing sector. *Information Systems Research, 3*(4), 307-333.

Weill, P., & Broadbent, M. (1998). *Leveraging the new infrastructure: How market leaders capitalize on information technology.* Boston: Harvard School Press.

Weissbach, R. (2003) Strategies of organizational informatization and the diffusion of IT. In M. Khosrow-Pour (Ed.), *Information technology & organizations: Trends, issues, challenges and solutions.* Hershey, PA: Idea Group Publishing.

Zwicker, R., Vidal, A. G., & Souza, C. A. (2005) Measuring the informatization level of businesses: A study of Brazilian industrial companies. *Proceedings of the Eleventh Americas Conference on Information Systems,* Omaha (pp. 315-323).

KEY TERMS

Information Systems Coverage: Extension of use, degree of dependency and adequacy of the information systems that a company owns.

Informatization: Managed process by which an organization continuously expands its IT assets and extends and deepens their appropriate use.

Informatization Level: An assessment of the effectiveness of the organizational use of IT.

Informatization Level Dimension: Conjunct of relevant aspects that participate in the informatization process and that are in some way related.

Informatization Level Model: Elements and relationships that structure the informatization level of a company.

IT Business Value Creation: The process of obtaining organizational benefits from IT investments.

IT Conversion Effectiveness: The capacity of converting IT investments into IT assets.

Business IT Systems Implementation

Călin Gurău
GSCM – Montpellier Business School, France

INTRODUCTION

The traditional channels of marketing are gradually being transformed by, or assimilated into, the global network represented by the Internet and modern information technology (IT) applications. Unfortunately, in most cases, the current IT systems are not fluid and dynamic enough to cope with ubiquitous customers who can contact the firm through a multitude of communication channels, such as mobile phones, Internet, or fax. The effective implementation of modern marketing strategies depends on the effective use of IT systems and procedures.

Internet-based technology can facilitate information dissemination, file transformation, data mining, and processing (Roberts, Raymond, & Hazard, 2005), which creates opportunities for the development and implementation of efficient *customer relationship management systems*. On the other hand, the new information technologies can also be used to increase the employees' satisfaction and productivity (Dorgan, 2003; Eichorn, 2004). Thus, the implementation and use of an efficient IT system for business and marketing activities becomes a fundamental task, which should be managed jointly by business specialists and IT professionals (Wierenga & Van Bruggen, 2000).

Unfortunately, these opportunities are hindered by many challenges at organisational or managerial levels, such as defining and restructuring the internal and the external sources of information, centralising the marketing database, and integrating the IT and marketing procedures at operational level.

Considering all these issues this paper attempts, on the basis of secondary data, to provide an overview of the main issues related with the implementation of IT systems in business organisations and the challenges related with the integration between information technology and marketing systems.

After a brief presentation of the previous research on this topic, the paper presents the stages of a gradual integration of IT systems in a business organisation and proposes a theoretical model

BACKGROUND

Considered a functional perspective, the main benefits of using modern IT systems for marketing operations are developed during three major phases: (1) automation, (2) information, and (3) transformation (Dedrick, Gurbaxani, & Kraemer, 2003).

1. **The First Stage:** IT systems are primarily used for automating manual systems of data recording and retrieving (Scott, Rosenbaum, & Jackson, 2003; Speier & Venkatesh, 2002). This level is particularly useful for improving the efficiency of routines, or simple tactical activities (Eli, Sundaram, & Chin, 2002).
2. **The Second Stage—Information:** The useful data are transformed through processing into relevant information for marketing operations and procedures (Ranchhod & Gurău, 2004). The information stage integrates the automated procedures developed in the previous phase, the data collected in the automation phase being scrutinised, selected, processed, and converted into business intelligence (information).
3. **The Third Stage—Transformation:** The company starts to adapt and use knowledge in order to enhance its strategic positioning. In this stage, the company will transform itself into a market-oriented, proactive organisation that uses IT systems in an integrated way to increase the effectiveness of every marketing operation (Roberts et al., 2005).

Many authors have emphasised the importance of IT systems for developing efficient *customer relationship management strategies* (Agrawal, 2004; Goldsmith 2004; Gurău, 2003; O'Malley & Mitussis, 2002; Plakoyiannaki & Tzokas, 2002; Roberts et al., 2005), and for employees' satisfaction (Dorgan, 2003; Eichorn, 2004). The level of IT integration in the organisational business structures and strategies is directly related with company's performance and profitability (Dedrick et al., 2003; Eichorn, 2004).

Unfortunately, in many organisations, the implementation of modern IT systems is a major source of tensions and problems, as many recent studies clearly demonstrate:

- The main reason for most B2B project failures is the incapacity of partners to implement a well-integrated technology infrastructure to support business processes (Meehan, 2002).
- A study of New Zealand corporations showed major difficulties in achieving the integration of business and

IT, both in the public and private sector (Navigate and Systems Planning Associates, 2002).

- A survey published by CFO Publishing Group indicated that 44% of chief financial officers indicate a weak alignment between IT and business strategies (Hoffman, 2003).
- An online survey showed that 91% of IT managers are aware of the necessity to integrate IT and business strategy, and 77% indicate that a poor understanding of business needs and objectives in a top barrier in the effective use of IT (Mejias, 2002).

The specialists have attempted to identify and propose solutions to these problems. A frequently used concept is that of *organisational* or *strategic alignment*, which emphasises the need to correlate the functioning of the IT system with business processes, in the context of organisational strategies (Roberts et al., 2005). Other authors emphasise the role of human resources in adopting, shaping, and enhancing the strategic use of IT for marketing operations (Dorgan, 2003; Speier & Venkatesh, 2002). By emphasising the role of organisational leaders, Eichorn (2004) constructs the concept of *internal customer relationship management*, and proposes the application of its principles as a solution for effective business-IT integration and performance. However, no study was able to synthesise all the aspects of *IT systems implementation* in modern business organisations.

The Integration Between IT Systems and Business Strategies

The introduction of modern IT systems, especially when based on Internet connectivity, requires the restructuring of information collection, archiving, and processing capabilities at the level of the entire organisation.

The recent development of communication technology forces the firm to redefine the sources of information. Customers can contact the company using multiple communication channels, such as mail; fax; fixed or mobile phone connections; or e-mails. The organisational structure should be able to accommodate all these flows of information and to introduce filter mechanisms that analyse the content and the level of urgency of the message, and then direct the communication to the relevant people within the organisation.

On the other hand, the complexity of the input data requires a centralised system of information storage that can be accessed simultaneously by various organisational departments. The centralisation of customer and organisational databases contradicts the traditional model of departmentalised databases, which often created redundancies and limited the interdepartmental communication and collaboration. In the present competitive environment, characterised by dynamism, complexity, and unpredictability, the speed of reaction and the capacity to work in multi-disciplinary teams are paramount for a company's survival and success.

The introduction of modern IT systems, and the restructuring of organisational processes and architecture, are often perceived as major cultural shocks by company's employees. Many studies have identified the resistance to change, and the concerns related with the use of novel IT systems as important barriers for the implementation of information technology in business organisations. This indicates that the implementation of a new IT system requires a change not only at functional level but also in the philosophy and culture of the firm. The adoption of a market-oriented, IT-based, business approach cannot succeed without the change of employees' mentality and system of values.

Another major problem in modern organisations is the interface between IT and marketing systems. Too often, the two organisational functions are independently structured, creating disparate subcultures within the same company, with different values, objectives, procedures, and approaches. The collaboration between IT and marketing becomes more difficult and complex as the culture gap widens and each group affirms and protects its primary importance in the organisational structure (Eichorn, 2004).

The adoption of a market-oriented approach by modern enterprises requires a closer collaboration between IT and marketing functions (Ranchhod & Gurău, 2004). The two functional departments need to share experience and information and to work together towards defining and realising common goals. Often, this approach is facilitated and coordinated by the top manager, who creates multi-divisional teams and establishes formal rules of cooperation, that are enforced at middle management level (Goleman, 2002).

Technology deployment represents the way in which companies plan and manage IT to enhance the marketing activities and procedures. From an organisational point of view, five main stages of this process can be defined as a gradual evolution of systems, processes, and procedures:

1. The strategic use of IT is focused on IT applications that support and enhance the competitive advantage of the company.
2. The IT management examines and improves internal IT related activities, such as the usage of new technologies, the development and adaptation of specific IT applications, or the degree of IT usage practised by the employees.
3. The enterprise information systems are restructured in order to align the IT strategy with the organisational structure and to manage more effectively the internal information networks.
4. The IT infrastructure is integrated in the company structure and its usage is formalised in order to manage more effectively the IT resources and capabilities of the organisation.

5. The IT system becomes centralised and its capabilities are shared by all the departments and the hierarchical levels of the organisation.

This process should be closely coordinated by the top management, taking into account a number of specific conditions, which are detailed in the five propositions presented hereafter. The restructuring of the company's systems and processes must balance novelty and continuity in order to facilitate employees' adaptation to the new structures and procedures (Ranchhod, Gurău, & Hackney, 2004):

Proposition 1

The implementation of IT systems in business organisations should take into account all organisational dimensions and structures (internal and external).

The implementation of modern IT systems should represent both a better adaptation to the evolving market environment, and a major enhancement of internal company capabilities. Before proceeding to the restructuring of the organisational architecture, the managers need to be aware of all the aspects involved in this process.

Proposition 2

The implementation of IT systems in business organisations is a multi-divisional team operation.

The complexity of novel IT systems, and the interaction between technology infrastructure and business procedures, makes the implementation process difficult and sensitive. Often, the senior manager's expertise cannot cover all the aspects of the implementation process. Although the manager needs to coordinate the integration effort, the responsibility for each implementation stage should be assumed by a multi-disciplinary team of specialists from different organisational departments. Not involving information technology professionals from the early stages of the implementation process might jeopardise the later integration of IT with key business systems and can create problems of compatibility between various applications.

Proposition 3

The implementation of IT systems in business organisations should be correlated with the level of skills, knowledge, and capabilities of the employees.

The decision to implement a new IT system should be accepted and shared by the company's employees in order to avoid future reactions of rejections. Besides a clear internal communication regarding the reasons and objectives of implementing the novel IT systems, the management of the firm has to establish training programs to familiarise the employees with the characteristics of the new IT tools and applications.

Proposition 4

The implementation of IT systems in business organisations must be compatible with, and enhance, the existing competitive advantage of the organisation.

The strategic alignment between IT systems and marketing strategies is a more complex process than simply translating the existing business processes into IT procedures. The alignment has to be made on different levels, in order to preserve and enhance the existing competitive advantage of the firm (between IT and marketing functions; between strategic objectives and tactic procedures; and between core IT applications and decision support systems).

Proposition 5

The implementation and functioning of IT systems in business organisations should be based on flexibility, transparency, and connectivity.

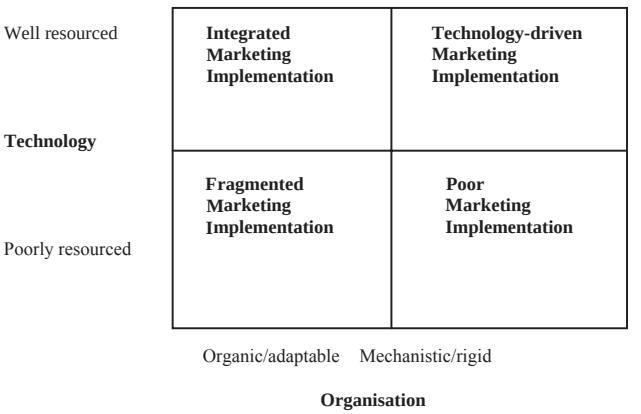
The implementation of an *IT-based marketing strategy* will impact and transform, in some measure, all the elements of the organisation. The strategic plan of this process needs to address and solve the problems of flexibility, transparency, and continuity between existing and future business structures and processes. A seamless integration of all IT operations and quick access to relevant data and market evaluations will enhance the capacity of the firm to react quickly to changes in the market structure and the external business environment.

The integration between IT and business systems is a continuous process that needs to be repeatedly evaluated and improved. In order to identify the company stance regarding the market strategy implementation and effectiveness, it is necessary to measure both the level of IT infrastructure as well as the internal culture/functioning of the organisation. By combining the technology context with the organisational structure, the matrix shown in Figure 1 can be developed (Ranchhod & Gurău, 2004).

Integrated Marketing Implementation

The companies from this quadrant have a well-balanced marketing implementation based on good IT systems and a dynamic/flexible organisation. Integrated systems offer high-value customer service, having a positive effect on customers' loyalty and company's profitability. Employees and technology work in unison, reinforcing their capabilities.

Figure 1. The influence of technology and organisational structure on the implementation of marketing strategies



Fragmented Marketing Implementation

These companies have excellent organisational capabilities, but their IT systems are basic and poorly integrated with business processes. These organisations need to invest more in IT infrastructure and to build management procedures for using the IT resources more effectively.

Technology-Driven Marketing Implementation

In this quadrant, companies tend to rely on technology to a greater extent. However, their internal processes lack flexibility and responsiveness, and for this reason, their sophisticated IT infrastructure is not fully used. These firms might use automated IT application to replace routine work, but do not use IT functions to improve their business intelligence and competitive advantage.

Poor Marketing Implementation

Companies in this quadrant have outdated organisational practices and possess poor technology systems. These firms need both organisational improvement and technology investment, which then need to be integrated at a strategic level. If well managed, the IT implementation process can be timely correlated with organisational restructuring, resulting in improved marketing capabilities and strategic focus.

The matrix can be used as an analysis tool by existing companies that can identify the main problems related with the implementation of their marketing strategy. The implementation of advanced technology can imbalance the firm and its market orientation if the organisational structure does not become more flexible and responsive. The procedural routines that are established for IT systems must not increase the rigidity of the organisational structure, but should be integrated into complex customer relationship management applications, providing the basis for reliable and responsive customer service operations.

FUTURE TRENDS

The complexity and the dynamism of markets, as well as the huge amount of information that needs to be accessed, selected, and processed requires the implementation of automated systems that can complement human skills and expertise. The combination between IT systems and flexible organisational structures will transform the marketing strategy into a knowledge management application, organically integrated into the model of learning organisation. The business organisation will therefore use advanced IT systems to create, use, and apply knowledge about markets and customers, transforming the firm into a flexible learning organisation. Within this organisational framework, new business strategies will be developed by *knowledge-based companies* (Dunn & Salazar, 2004):

1. **Knowledge Replication:** Which applies a verified organisational model into new markets, allowing the rapid expansion of the firm with reduced costs (e.g., MacDonald's);
2. **Knowledge Diffusion/Leveragability:** Based on the efficient management and diffusion of the existing knowledge within the organisation (e.g., Hewlett Packard);
3. **Knowledge Innovation:** Using the existing knowledge in new, innovative ways, in order to create superior value for customers (e.g., online recommendation systems);
4. **Knowledge Giveaway:** The knowledge developed by the firm is openly shared with the business community in order to reinforce the brand positioning of the company (e.g., Massachusetts Institute of Technology);
5. **Knowledge Commercialisation:** Based on the accumulated knowledge, the firm is able to commercialise consulting services to customers and other organisations (e.g., British Gas Plc).

Making a thorough analysis of their positioning and expertise, the modern organisations will be able to identify and apply the best strategies allowing them to take advantage of their specific market knowledge.

CONCLUSION

The implementation of IT systems in modern business organisations is a necessity imposed by the technologically driven evolution of the market. When fast decisions need to be made, the role of market intelligence becomes highly relevant and organisational systems need to offer up-to-date, relevant information to company's employees. On the other hand, as technology becomes more sophisticated, marketers need to understand its relationship with effective marketing strategies and to integrate organisational processes and IT infrastructure in a highly flexible architecture.

This study has outlined a series of important issues related with IT-business integration, especially related with the implementation process of novel IT systems. The complexity of this restructuring process cannot be solved with general solutions, each company having to design the best answer to its strategic problems. Although the theoretical model and the propositions presented in this paper are supported by secondary data evidence they need to be verified in real situations. On the basis of these theoretical findings, further research projects can be designed to collect, process, and analyse primary data regarding the problems experienced by managers implementing IT infrastructure and integrating it with existing organisational systems.

REFERENCES

Agrawal, M. L. (2004). Customer relationship management and corporate renaissance. *Journal of Services Research, 3*(2), 149-171.

Dedrick, J., Gurbaxani, V., & Kraemer, K. L. (2003). Information technology and economic performance: A critical review of the empirical evidence. *ACM Computing Surveys, 35*(1), 1-27.

Dorgan, M. (2003). Employee as customer: Lessons from marketing and IT. *Strategic HR Review, 2*(2), 10-11.

Dunn, D., & Salazar, A. (2004). Knowledge-based competitive advantage in the Internet age: Discovering emerging business strategies. *International Journal of Information Technology and Management, 3*(2/3/4), 246-258.

Eichorn, F. L. (2004). Applying internal customer relationship management principles to improve business/IT integration and performance. *Problems and Perspectives in Management, 4,* 125-148.

Eli, J., Sundaram, S., & Chin, W. (2002). Factors leading to sales force automation use: A longitudinal analysis. *Journal of Personal Selling & Sales Management, 22*(3), 146-156.

Goldsmith, R. E. (2004). Current and future trends in marketing and their implications for the discipline. *Journal of Marketing Theory and Practice, 12*(4), 10-17.

Goleman, D. (2002). *Primal leadership.* Boston: Harvard Business School.

Gurău, C. (2003). Tailoring e-service quality through CRM. *Managing Service Quality, 13*(6), 520-531.

Hoffman, T. (2003, October 21). CFOs cite poor alignment between IT, business. *Computerworld.* Retrieved November 15, 2005, from http://www.computerworld.com/managementtopics/management/project/story/0,10801,86303,00.html

Meehan, M. (2002, January 1). Users say lack of IT integration hurts B2B. *Computerworld, 36*(1), 8.

Mejias, E. (2002). IT—Business relationships survey. Retrieved November 15, 2004,from http://www.enterprise-works.com http://www.enterprise-works.com /servlet/queryITSurvey.class

Navigate and Systems Planning Associates. (2002). *The business—IT gap.* Retrieved November 15 2004, from http://www.navigate.co.nz/the.htm

O'Malley, L., & Mitussis, D. (2002). Relationship and technology: Strategic implications. *Journal of Strategic Marketing, 10*(3), 225-238.

Plakoyiannaki, E., & Tzokas, N. (2002). Customer relationship management: A capabilities portfolio perspective. *Journal of Database Marketing, 9*(3), 228-237.

Ranchhod, A., & Gurău, C. (2004). Qualitative issues in IT and organizational processes in implementing marketing strategies. *Qualitative Market Research: An International Journal, 7*(4), 250-256.

Ranchhod, A., Gurău, C., & Hackney, R. (2004). The challenge of cyber-marketing planning and implementation. *International Journal of Information Technology and Management, 3*(2/3/4), 141-156.

Roberts, M. L., Raymond, R. L., & Hazard, K. (2005). Strategy, technology and organisational alignment: Key components of CRM success. *Database Marketing & Customer Strategy Management, 12*(4), 315-326.

Scott, W., Rosenbaum, M., & Jackson, D., Jr. (2003). Keys to implementing productive sales force automation. *Marketing Management Journal, 13*(1), 1-13.

Speier, C., & Venkatesh, V. (2002). The hidden minefields in the adoption of sales force automation technologies. *Journal of Marketing, 66*(3), 98-111.

Wierenga, B., & Van Bruggen, G. (2000). *Marketing management support systems: Principles, tools and implementation.* Boston: Kluwer.

KEY TERMS

Chief Information Officer (CIO): An executive title in an organisation that designates the person responsible for developing and implementing the information systems.

Customer Relationship Management (CRM): A business model that includes specific techniques and methods for attracting customers and developing a long-term, company-customer relationship.

Decision Support Systems (DSS): A specific class of computerised information systems that supports business and organisational decision-making activities by providing facilities for data analysis, evaluation, and interpretation.

Internal Customer Relationship Management: A business theory advocating the application of CRM principles within business organisations among various functional departments.

Knowledge Management: The reuse and redeployment of accumulated knowledge from which organisational learning is manifested.

Organic Organisation: An organisation with a flexible structure that collects and uses outside knowledge.

Mechanistic Organisation: An organisation with a rigid hierarchical structure which implements highly formalised jobs/roles definitions.

Strategic Alignment: The extent to which the information systems strategy supports, and is supported by, the business strategy in an organisation.

Business Model Application of UML Stereotypes

Daniel Brandon, Jr.
Christian Brothers University, USA

OVERVIEW

The UML (Unified Modeling Language) has become a standard in design of object oriented computer systems (Schach, 2004). UML provides for the use of stereotypes to extend the utility of its base capabilities. In the design and construction of business systems, the use of stereotypes is particularly useful stereotypes, and this article defines and illustrates these.

UML STEREOTYPES

"Stereotypes are the core extension mechanism of UML (Lee & Tepfenhart, 2002). If you find that you need a modeling construct that isn't in the UML but it is similar to something that is, you treat your construct as a stereotype" (Fowler & Kendall, 2000). The stereotype is a semantic added to an existing model element, and diagrammatically it consists of the stereotype name inside of guillemots (a.k.a. chevrons) within the selected model element (Schach, 2005). The guillemot looks like a double angle bracket (<< … >>), but it is a single character in extended font libraries (Brown, 2002). Each UML model can have zero or many stereotypes; and each stereotype can define a set of tagged values, constraints, and possibly a special icon or color (Arlow & Neustadt, 2005). The UML defines about 40 of these stereotypes, such as "<<becomes>>", "<<include>>", and "<<signal>>" (Scott, 2001). However, these 40 standard stereotypes are not particularly useful in business models and do not add the meaning necessary for automatic code generation in a UML CASE tool.

One common general use of the stereotype is for a metaclass. A metaclass is a class whose instances are classes, and these are typically used in systems in which one needs to declare classes at run time (Eriksson & Penker, 1998). A similar general use is for powertypes. A powertype is an object type (class) whose instances are subtypes of another object type. Figure 1 shows an example of the use of stereotypes for powertypes (Martin & Odell, 1998). Another type of usage is in the "binding" of parameterized (template) classes historically in C++ and now in the latest version of Java (Oestereich, 1999).

USER-DEFINED STEREOTYPES FOR BUSINESS SYSTEMS

In the design of business systems we have found some stereotypes that were occasionally useful, and two stereotypes that are extremely useful. When defining stereotypes it is useful to describe (Eriksson & Penker, 1998):

1. On which (UML) element the user-defined stereotype should be based
2. The new semantics the stereotype adds or refines
3. One or more examples of how to implement the user-defined stereotype

Figure 1.

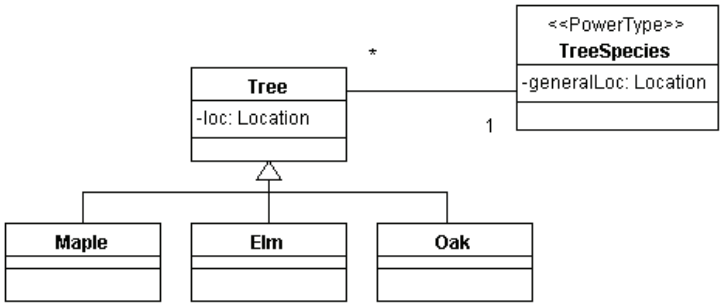


Figure 2.

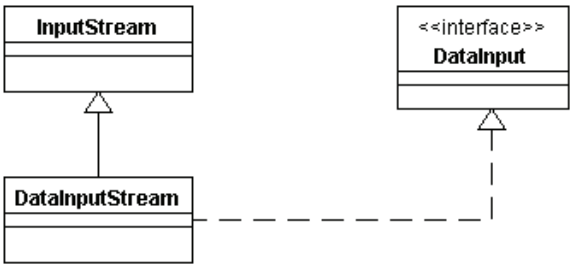
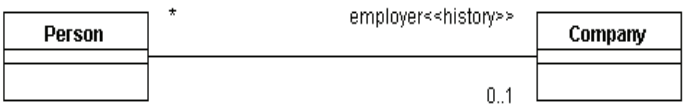


Figure 3.



An issue is where such definition takes place both within UML and in the UML software tool. If a modeling tool does not provide such built-in support (and most do not), many modelers put a note in the model or a reference to external documentation (Arlow & Neustadt, 2005).

One common use of stereotypes in business systems is for interfaces as found in Java or CORBA; this is shown in Figure 2. An interface typically has public functionality but not data (unless holding data for global constants). The class model element has been modified with the "<<interface>>" notation (Oestereich, 1999). This is commonly used for UML CASE products that do not have separate interface symbols or where these symbols do not allow data (i.e., global constants).

Another common stereotype usage in business systems is to clarify or extend a relationship. Figure 3 shows a stereotype called "history," which implies a "many" cardinality for history purposes, that is, each Person has zero or one current employers but may have many employers in terms of the employee's history. It may imply some common functionality upon code generation such as (Fowler & Kendall, 2000):

```
Company Employee::getCompany(Date);
```

Still another common stereotype usage in business systems is to define ancillary classes that are not part of the core business model, but are needed for a particular implementation. Boundary, controller, and entity stereotypes were predefined in UML (Oestereich, 1999) with special symbols in some products; however, the definition and use of these varies with author. Schach (2005) divides implementation classes into Entity classes, Boundary classes, and Control classes. Entity classes are part of the core business model and represent entities like people, organizations, transactions, places, and things. Boundary classes are used for communication between the software product and the actors such as I/O operations; and Control classes are used for algorithms and business rules. Each of these types of classes is designated with the stereotype notation.

CODE WRITING AND GENERATION

Most modern UML CASE (computer-aided software engineering) products can generate "skeleton" classes from the UML class diagrams and possibly other diagrams. For business systems design, we need to write the code for our classes (usually implemented in Java or C++) based on both the structural model (UML class diagram) and the dynamic model (UML activity diagram). This process is shown in Figure 4. It is very important that consistency between the two diagrams is achieved.

Many such CASE products allow the user to write his or her own "class generation scripts" in some proprietary scripting language or in a general scripting language (i.e., Python). With user-defined stereotypes, the user can modify the class generation script code to use his or her stereotypes as needed.

Figure 4.

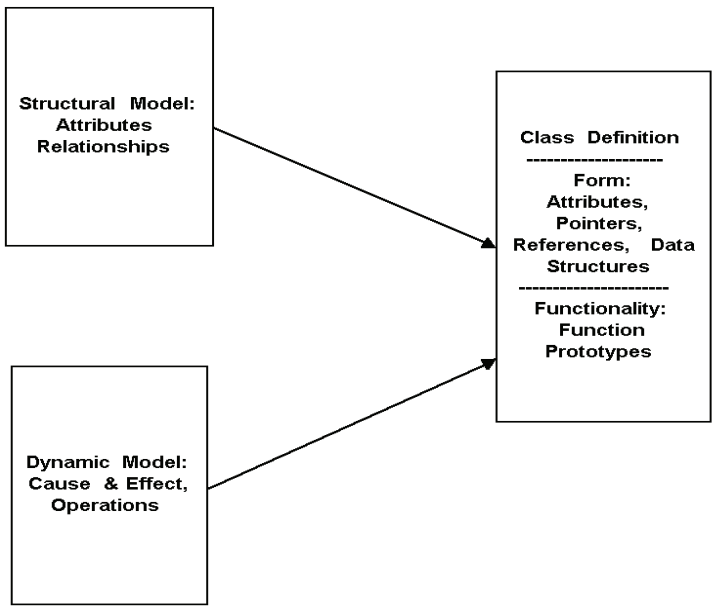
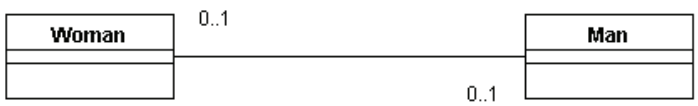


Figure 5.



RELATIONSHIP OBJECT TYPES

As an example, consider a simple association between two object types. Often these simple relationships need to be modeled as object types because these relationships have data content and/or functionality. Figure 5 shows a simple association between two object types representing the relationship "current marriage." If we need to maintain an attribute on each marriage (such as rating), then we can more effectively represent the relationship as an object type as shown in Figure 6. Here we use the "relationship" stereotype to indicate that this object type is a relationship; and the code generation can use a more appropriate class representation. Other authors have suggested other notations for relationship object types such as "placeholders" (Martin & Odell, 1998), and UML suggests using the dotted line from a standard object type (class) to the relationship line. But implementing these other diagramming techniques in code generation is difficult and has ambiguity problems.

ACTIVITY DIAGRAMS

A UML activity diagram is a state diagram in which most of the states are action states and most of the transitions are triggered by the completion of these action states. This is the case in most models of business systems. Activity diagrams identify action states, which we call operations (Martin & Odell, 1998), and the cause and effect between operations. Each operation needs to belong to an object type, at least for a C++ or Java implementation. Operations may be nested, and at some point in the design the operations need to be defined in terms of methods. The methods are the processing specifications for an operation and can be so specified in lower lever activity diagrams, pseudo code, or language specific code. Note that the term "methods" may cause some confusion here since in programming terminology, a method is a function defined within a class and it is invoked upon an object (unless it is a static method).

Figure 6.

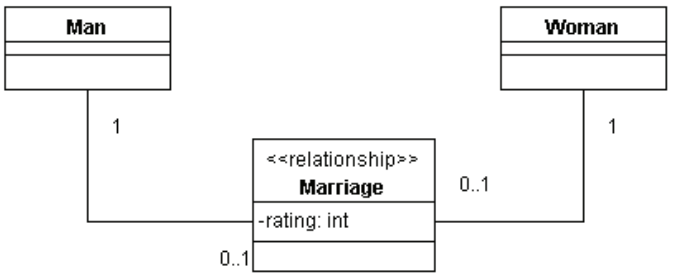
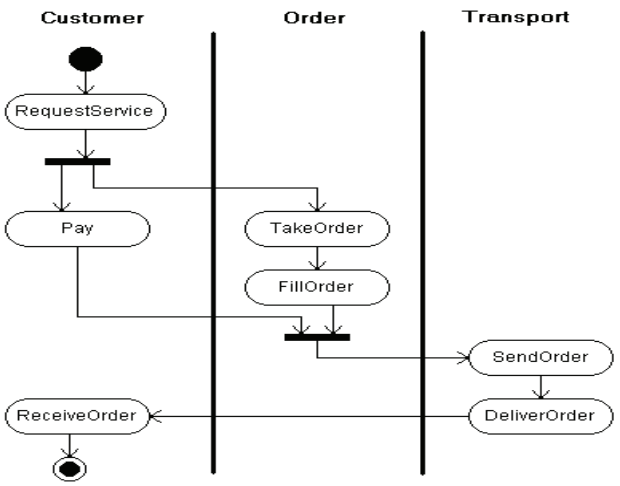


Figure 7.



Drawing Methodology

Figure 7 shows a typical UML activity diagram for a simple ordering process. The operations are represented in the ovals, and the arrows show the cause and effect scenario or the "triggers." In this diagram there are two "fork/join" model elements, and the use of "conditional branch states" is also common. Each of the operations must be associated with a particular object type. The standard way to do that in this UML type diagram is to use "swimlanes," and these are the vertical lines shown in Figure 7.

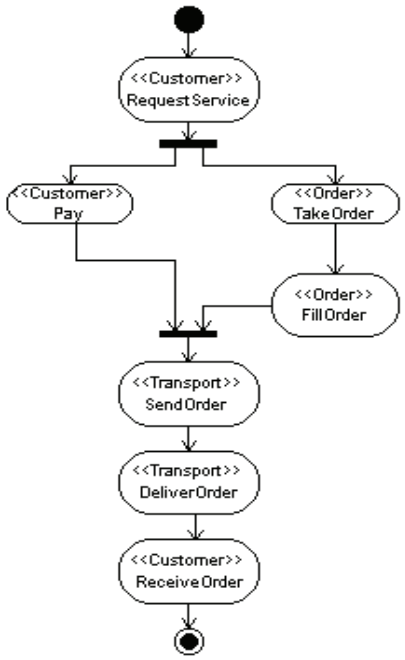
There are two problems with the standard representation as shown in Figure 7. The first problem is that as the system gets more complex (more object types and operations), it is very difficult to draw in swimlanes. The second problem is that code generation is very difficult in UML CASE products since you have to scan the geometry of the drawing to find out which operations lay in which swimlanes. A solution to the previous problems with standard UML activity diagrams is to use a stereotype for the operation element to indicate the object type (class) owning that operation. Figure 8 shows the same systems as Figure 7 drawn with the "operation owner" stereotype.

Model Consistency

The use of these UML stereotypes allows a greater degree of consistency checking of business models. A final business system design will involve several UML diagram types. For example, business systems typically have static structural diagrams (UML class diagram) and a dynamic diagram (UML activity diagram). These diagrams must be consistent with one another, in particular:

1. The object types (shown with the operation stereotype notation) that contain the operations in activity diagrams must be included on the structural diagrams.

Figure 8.



2. The operations shown in the activity diagrams (along with the object types identified with the stereotype notation) must be included as operations in the same object type on the structural diagrams.

For a UML business system example (including implementation in C++), the reader is referred to the full book chapter on this subject (Brandon, 2003).

FUTURE TRENDS

As UML becomes more accepted for general use in the design of business systems, we could expect to see more universal stereotypes being formulated. Eventually libraries of these stereotypes should become generally available and UML tool vendors would include support for these libraries in their products.

CONCLUSIONS

UML stereotypes can be very useful in designing business systems. For example, the use of a "relationship" stereotype is helpful in static structural models (UML class diagrams), and the use of an "operation owner" stereotype is most helpful in dynamic models (UML activity diagrams). These stereotypes aid in both the design/drawing phase and in the implementation (coding) phase of the overall system construction.

REFERENCES

Arlow, J., & Neustadt, I. (2005). *UML 2 and the unified process*. Addison Wesley.

Brandon, Jr., D. (2003). Use of UML stereotypes in business models. In L. Favre (Ed.), *UML and the unified process* (pp. 262-272). Hershey, PA: IRM Press.

Brown, D. (2002). *An introduction to object-oriented analysis*. John Wiley & Sons.

Eriksson, H-E., & Penker, M. (1998). *UML toolkit*. John Wiley & Sons.

Fowler, M., & Kendall, S. (2000). *UML Distilled*. Addison-Wesley.

Lee, R., & Tepfenhart, W. (2002). *Practical object-oriented development with UML and Java*. Upper Saddle River, NJ: Prentice Hall.

Martin, J., & Odell, J. (1998). *Object oriented methods—A foundation* (UML ed.). Upper Saddle River, NJ: Prentice Hall.

Object Domain. (2001). *Object Domain Systems Inc.* Retrieved from www.objectdomain.com

Oestereich, B. (1999). *Developing software with UML*. Addison Wesley.

Schach, S. (2004). *Introduction to object oriented analysis and design*. Irwin McGraw Hill.

Schach, S. (2005). *Object oriented and classical software engineering*. Irwin McGraw Hill.

Scott, K. (2001). *UML explained*. Addison-Wesley.

KEY TERMS

Activity Diagram: A UML diagram showing operations and triggers between operations; a diagram which shows system dynamics via cause and effect relationships. An Activity Diagram is a state diagram in which most of the states are action states and most of the transitions are triggered by the completion of these action states.

CASE: Computer-aided software engineering.

Class: A program construct representing a type of thing (abstract data type) that includes a definition of both form

(information or data) and functionality (methods); the implementation of the design concept of "object type."

Composition: A new class in an objected programming language that is composed of other classes.

Dynamic Model: A UML model describing dynamic behavior such as state changes, triggers, and object type operations.

Encapsulation: The ability to insulate data in a class so that both data security and integrity are improved.

Framework: A software foundation that specifies how a software system is to be built. It includes standards at all levels both internal construction and external appearance and behavior.

Function: A programming construct where code that does a particular task is segregated from the main body of a program; the function may be sent arguments and may return arguments to the body of the program.

Implementation: The code placed inside of methods. For some languages this code is pre-compiled or interpreted.

Include: Some code stored separately from the main body of a program so that this code can be used in many programs (or multiple places in the same program).

Inheritance: A feature of object oriented languages that allows a new class to be derived from another class (a more general class); derived classes (more specific classes) inherit the form and functionality of their base class.

Interface: The specification for a method ("what" a method does); how that function is called from another program. Interfaces are provided in source form as opposed to implementations, which are secure. This allows one to use a method without regard for "how" that method is coded. It also allows multiple implementations of the same interface.

Library: A group of functions and/or classes stored separately from the main body of the main program; an "include" file consisting of functions and/or classes.

Metaclass: A class whose instances are classes.

Method: A function defined inside of a class; a processing specification for an operation.

Object Type: A specification of a type of entity; both structure (attributes) and operations (functions) are specified for object types; the notion of object type is a design notion being implemented as a "class."

Operation: A process related notion in UML; operations cause state changes in objects.

Packages: Similar to a library, but just containing classes.

Patterns: A software library for a common business scenario. A framework may be a design framework (possibly expressed in UML) or an implementation framework (possibly in C++, Java, or PHP).

Polymorphism: The ability of object oriented programs to have multiple implementations of the same method name in different classes in an inheritance tree. Derived classes can override the functionality defined in their base class.

Relationship: A connection concept between object types. There are several types of relationships, including aggregation, composition, association, and inheritance (generalization/specialization).

Reuse: Reuse (software) is a process where a technology asset (such as a function or class) is designed and developed following specific standards, and with the intent of being used again.

Separation: The separation of what a method does (interface) from how the method does it (implementation).

Stereotype: The core extension mechanism of UML.

Structural Model: An UML model describing static structure (relationships and properties).

Trigger: One operation invoking another operation; a call from one method to another method within the same or different object type (class).

UML: Unified Modeling Language.

Business Models for Municipal Broadband Networks

Christos Bouras
University of Patras and Research Academic Computer Technology Institute, Greece

Apostolos Gkamas
Research Academic Computer Technology Institute, Greece

George Theophilopoulos
Research Academic Computer Technology Institute, Greece

Thrasyvoulos Tsiatsos
Aristotle University of Thessaloniki, Greece

INTRODUCTION

This article examines the most effective business model for the optimal exploitation of the currently developing broadband metropolitan area networks in various municipals around the globe. The proper exploitation strategy of the municipal broadband networks to be deployed could boost the demand for broadband connections and applications. The article describes the relevant, available business models in detail, including ways for broadband infrastructures' expansion, and deals with viability issues, regarding the managing authority which is responsible for the broadband metropolitan networks.

A business model, specifically in the current article, determines the way in which the exploitation of a metropolitan, community-owned, optical network will be effectuated. Municipalities may play a critical role in enabling the deployment of broadband infrastructures by the private sector (Government of Sweden, 2007):

- Placing open conduit under all freeways, overpasses, railway crossings, canals and bridges.
- Allowing over lashing of fiber on existing aerial fiber structures.
- Forcing existing owners of conduit, such as electrical companies, telephone companies, and so forth, to make 100% of their conduit accessible to third parties.
- Coordinate construction of all new conduits, especially by building entrances to minimize the "serial rippers" and make all such conduit open to third parties.

However, the development of such broadband infrastructures raises several questions regarding the business model that shall be used for their exploitation (e.g., what will be the role of the municipality, what will be the degree of government interventionism, how healthy competition is going to be promoted, how the network's viability is going to be ensured, etc.).

Therefore, this article intends to:

- Record international experience with respect to broadband business models for the exploitation of broadband infrastructures.
- Summarize the available business models and present, through comparative analysis, the advantages and disadvantages of each business model.

The remaining of this article is structured as follows: The next section presents the international experience in developing broadband metropolitan area networks in various municipalities around the globe. Next, the aticle presents and compares the available business models for the optimal exploitation of the broadband municipal networks, and presents the future trends in the area. Finally, the article is concluded.

BACKGROUND

In general, broadband metropolitan networks have been developed in municipalities along different parts of the globe. Pioneer countries, such as Canada and Sweden, present examples of how broadband infrastructures can reinforce the local economy and contribute in further development. International experience records various business models (OECD, 2003) on broadband infrastructures exploitation, and a few indicative ones are mentioned in the following paragraphs:

- **Demand aggregation model.** This model regards coordinating efforts, exerted by regional carriers and aiming at the aggregation of the demand for broadband

services. The regional carrier presents the aggregated demand as an attractive clientele basis to the service suppliers, with whom it negotiates the overall purchase of broadband services and the percentage ownership upon the infrastructure.

- **Open access/wholesale provider model.** According to this model, regional carriers and local communities, usually cooperating with an independent infrastructure provider, who offers wholesale prices (a public utility service, in principle), construct the fundamental broadband infrastructures (trenches, conduits, subterranean or aerial cables), incorporating a "public good" rationale, and based on the foreseen general needs, as is the case of roads and sewerage works.
- **Community-owned network with service provision model**. Regional carriers and local communities, usually cooperating with a local service supplier, or acting as broadband network service suppliers themselves, construct the fundamental broadband infrastructures and provide network wholesale or retail services, investing the resultant profits in the expansion of the infrastructure.

Remarkable efforts in Europe can be recorded in Ireland (www.enet.ie), Sweden (Stokab, www.stokab.se), Austria, The Netherlands (Kramer, Lopez, & Koonen, 2006) and Spain (LocalRet, http://www.localret.net/idiomes/english.htm). In the United States (U.S.), the cases of the State of Utah (UTOPIA, 2003; UTOPIA network, www.utopianet.org) and the city of Philadelphia (Wireless Philadelphia, 2005) are of great interest, concerning the successful application of business models for exploiting broadband metropolitan area networks. Besides from Europe and the U.S., remarkable efforts are tracked in other countries as well, such as Canada (CANARIE, www.canarie.ca) and New Zealand. Table 1 summarizes the features of business models of the most important of the aforementioned cases.

BUSINESS MODELS FOR MUNICIPAL BROADBAND NETWORKS

Important Aspects

A business model in our case determines the way in which the exploitation of a metropolitan, community-owned, optical network will be effectuated. Additionally, it determines the role of the municipality, the region and the private sector, the way healthy competition is going to be promoted, the

Table 1. Representative business models and their basic features

Features \ Business Models	Irish model	Stokab	LocalRet	UTOPIA	Philadelphia	CANARIE (Canada)
Public carrier						x
Local carrier (municipality, community, etc.)	x	x	x		x	x
Private carrier						x
Consortium			x	x		
Dark fibre network	x	x	x	x		
Last mile connections				x		
Government funding	x		x	x	x	x
Private support					x	x
Collocation facilities	x	x	x			
Leasing to telecommunication providers	x	x	x	x	x	
Supply of services		x				x

degree of involvement of the private sector and so forth. For example, competition is driving fixed and mobile players to invest in new technologies to reduce costs and position themselves in a converged environment (COM, 2006). Another example is that the liberalization of the local loop telecommunication infrastructure allowed the firms involved to behave more competitively and dropped broadband monthly fees to lower prices. Such an example is Sweden (Papacharissi & Zaks, 2006). Moreover, concerning the role of municipality, the analysis of Lehr, Sirbu, and Gillett (2004) shows that the case for a public role is complex and that the optimal policy is likely to depend critically on the type of wireless infrastructure that is being deployed, and the objectives for the system.

The business model aims to ensure the viability of the metropolitan community-owned optical network and to secure the resources for its operation, maintenance and expansion, while, at the same time, it aims to promote competition for offering better and cost effective services to the citizen (Henderson, Gentle, & Ball, 2005).

Figure 1 presents the three basic levels of a relevant business model (Hughes, 2003):

- The first level determines who (a private or public enterprise, etc.) exploits the network's passive equipment (channels, optical fibres, etc.).
- The second level determines who provides and exploits the active equipment of the network (switches, routers, etc.).
- The third level determines who offers access to the network, the services and the content.

"Open Access" and "Neutral Operator"

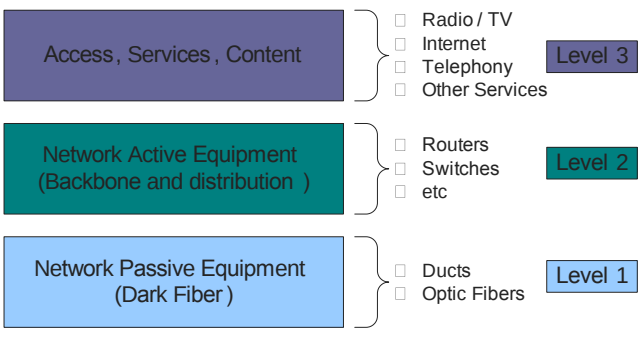
The attribution of different responsible carriers (municipality, private sector, etc.) to any one of the aforementioned levels of the business model leads to different business model scenarios indicating how public organizations and providers of infrastructures, equipment and services can cooperate for the consumer's benefit. Two basic features of the metropolitan optic fibre networks that do not really regard their construction as much as their funding, appropriate management and viability insurance are reflected in the concepts of "open access" and "neutral operator."

As far as the open access is concerned, the European Commission provides directions and guidelines that must rule the electronic communications between the member states of the EU. In particular, it is stated that the projects to be funded will have to be consistent and conforming to the new institutional framework for electronic communications, as well as to the rules concerning competition (issues of state aid and antitrust). Compliance to the competition rules constitutes an eligibility criterion for funding, while this has to be combined with the obligation for clear open access (Magnago, 2004). Specifically, funding has to be limited exclusively to infrastructure (i.e., installations of optic cables, channels, conduits, pylons, etc.) and equipment that is accessible to any telecommunication carrier and services supplier.

The infrastructure administrator will be liable to preserve the infrastructure character as an installation accessible to all carriers, supplying electronic networks and services, without discrimination. The role of the neutral operator (Monath, Cristian, Cadro, Katsianis, & Varoutas, 2003) is important because it has to:

1. Offer the network infrastructures' proprietors (on local, regional and national level) the possibility to increase their value and viability within a logical economic frame.
2. Reduce the needs for high initial investments on the part of the service suppliers and, at the same time, significantly increase the availability of economically accessible services on the part of the subscribers.

Figure 1. The basic levels of a business model



3. Be responsible for the observance and evolution of a revenue-sharing schema for all participating sides as well as for the continuous adaptation of the network's potential in accordance with the growing needs.
4. Act as an administrating entity, in general, ensuring the reliable operation of all cooperating parts (infrastructure proprietors, service suppliers and subscribers).

Equal Access Business Model

The target of this business model (Figure 2) is to ensure the equal access to the passive equipment of the network. More particularly:

- One entity is responsible for the first level, which offers cost-based access to the passive equipment of the network.
- In the second level, many providers are active and they offer access to active network equipment in a competitive environment.
- In the third level many providers are active and they offer broadband services to the end users in a competitive environment.

The entity responsible for the first layer constructs the network passive equipment and rents the passive equipment to one ore more networks providers. The network providers offer network services to services and content providers. And finally, services and content providers offer broadband services to the end users.

There are two important variations of this business model (PPPs orchestrated and Public Sector Telco), described in the following paragraphs.

The role of the first layer's entity is to motivate the competition in the above layers. This entity invests in passive equipment and, due to the nonprofit operation, offers the passive equipment in a cost-basis to the network providers. As a result, the market entry cost for a network provider is relatively low.

Full Public Control through Public-Private Partnerships (PPP)

In this business model (Figure 3), the municipal authority is responsible for all parts of the broadband network (passive equipment, active equipment, services). With this approach, the municipal ensures the full control in all levels through the participation in PPPs.

This business model can be used either when the municipal authorities are not ready to allow a temporary monopoly in broadband services or when there is law restrictions. In addition, this business model can be used when the service provides are not willing to invest (e.g., in rural areas).

A benefit of this business model is the simple administration due to the fact that only one organization is involved. The main disadvantage of this business model is the absence of competition. This results in limited options for the end user, as well as lack of pressure for price reduction. Finally, this business model requires the municipal authorities to operate a telecommunication network, an area in which municipal authorities have no experience.

PPPs Orchestrated

This is a variation of the equal access business model, which occurs when there is significant existing broadband

Figure 2. Equal access business model

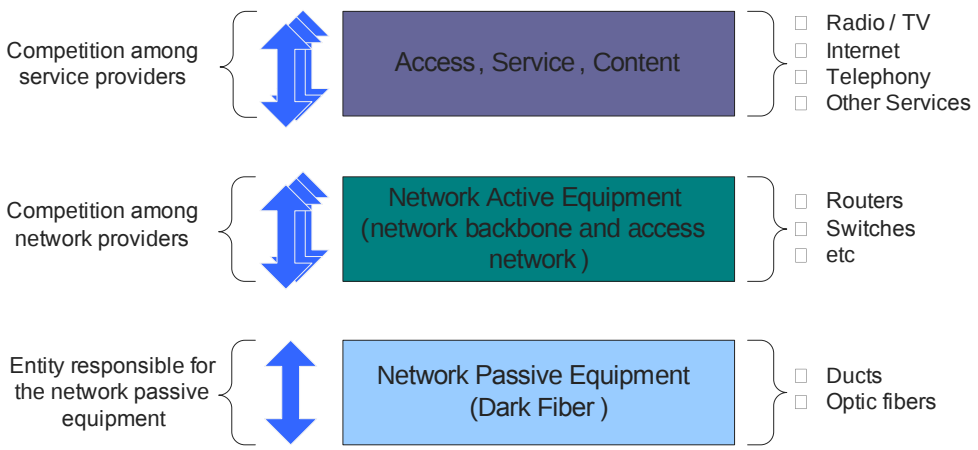
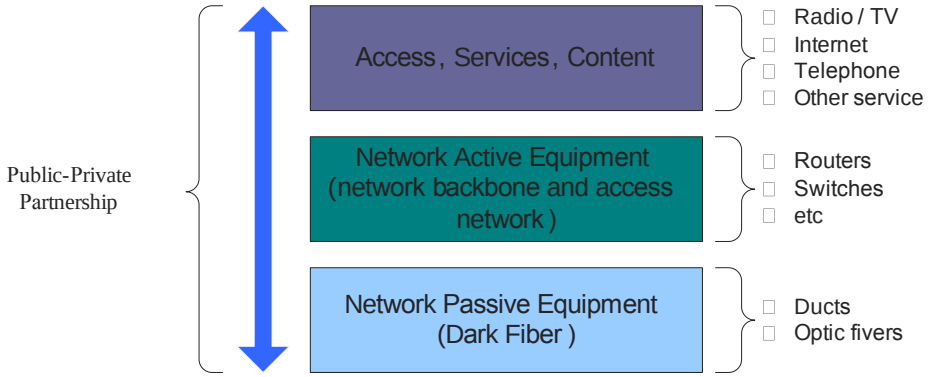


Figure 3. Full public control through PPPs



infrastructure in the area and major investments in new infrastructure is not necessary. In this scenario, the role of the local government may be to act as an orchestrator and, by bringing private organizations together, to ensure that existing assets are used to create a thriving market for broadband services.

This is usually performed by the local administrations and the owners of the existing infrastructure who create a joint venture to manage the passive infrastructure, as if it was a single asset. The active and access services layers are usually managed by one or more service providers on the basis of a partnership agreement with the joint venture.

In this scenario, it is common to have a single private company acting in the second layer (active equipment).

Public Sector Telco

This model is another variation of the equal access business model, where the public sector manages the passive and the active infrastructure, while competition among private companies is acting in the third level (services).

Sole Private Provider

According to this model, the operation and the management of the active network equipment and services are offered by a single private service provider. The network's passive equipment is owned by the public sector (e.g., municipality).

The advantage of this model (Heimgartner, Luke, Villa, & Johnston, 2005) is that the project becomes commercially viable at much lower levels of customer revenue. However, customers are unlikely to be offered as wide of range of services and will not benefit from the impact of competition on pricing. For these reasons the local government will often want to ensure that the monopoly is only offered as a temporary measure over a fixed term, during which time it hopes the sole service provider will generate sufficient numbers of customers to sustain a competitive market. Ob-

Figure 4. PPPs orchestrated

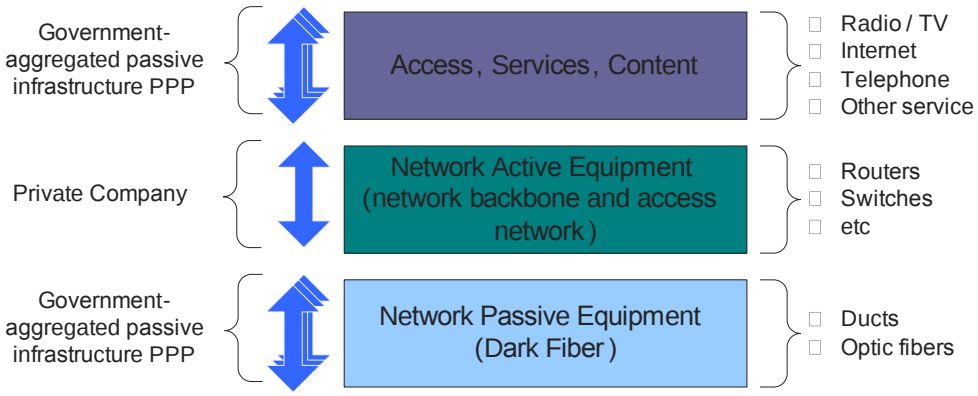


Figure 5. Public sector telco

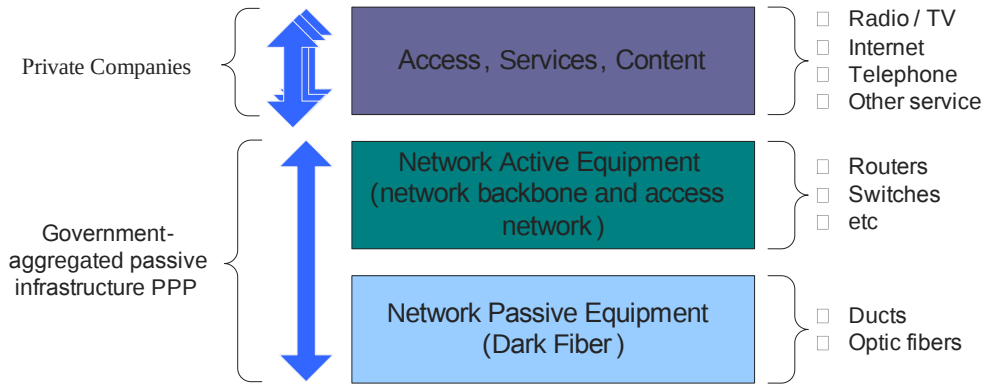
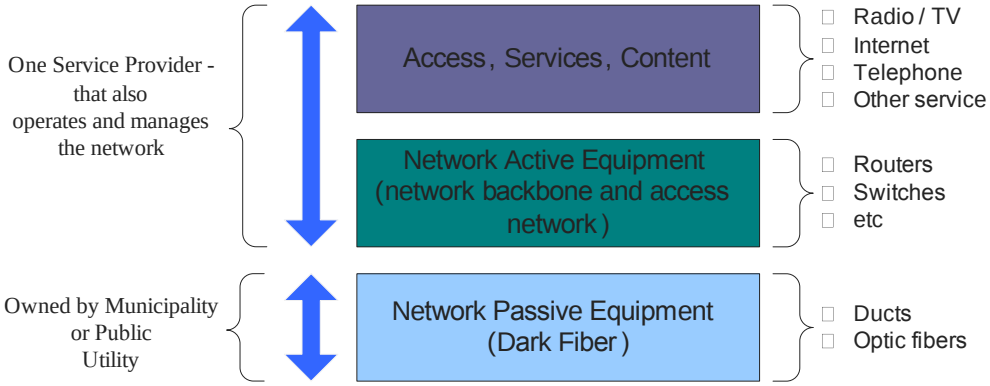


Figure 6. Sole private provider



viously, getting the length of this fixed period right is a key issue. Furthermore, transition to the equal access model at the end of the fixed period will require careful commercial and legal management.

FUTURE TRENDS

The neutral operator, in most cases an entity controlled by the municipal authorities, is of critical importance for the business models, because it:

- Secures financial viability of the owners of the infrastructure.
- Reduces the needs for high initial investments from the service providers and, at the same time, it increases considerably the availability of economically accessible services for the citizens.
- Is responsible for fair revenue sharing to all participants in the enterprising scheme.
- Plans and implements networks expansion.

In addition, the service providers should focus on providing economical and competitive services without caring for the development of the broadband infrastructure. Finally, as far as the end users are concerned, the selected business model ensures that they may choose between a number of services with financial and quality criteria. Until now, some research work has been presented concerning lessons learned from broadband development (Frieden, 2005).

All involved parts should bear in mind that once the broadband business model is applied and broadband infrastructures are deployed, quality of service and specific service level agreements (SLA) (Shin, Shin, & Han, 2004) for the provided services should be ensured.

Table 2. Comparison of local and regional models for broadband deployment

Model	Description	Advantages	Disadvantages
Equal Access	The municipal authorities develop and manage Level 1 (passive infrastructure). Both Level 2 and Level 3 are subject to competition.	• Public intervention at the lowest level of the business model (which, however, represents 70% of the cost of a new fixed network). • Market entry cost for a network provider is relatively low, due to leasing of the passive infrastructures on a cost basis.	• Entry barriers for network operators remain sizeable • Financial risk for the municipal authorities
Full Public control	All three levels are created and managed by the municipal authorities.	• Complete solution • Easier management of the whole "operation"	• Negative impact on competition in services and networks • Financial risk for the municipal authorities • Municipal authorities needs great technical and commercial expertise
PPPs orchestrated	Variation of the equal access business model. PPPs act in level 1, usually a single private company acts in level 2, while competition (many companies) acts in level 3.	Service providers may have a chance to enter the market, as part of the PPPs acting in level 3.	• Significant existing broadband infrastructure in the area is assumed • No competition in network active equipment (level 2) • Govermental interference both in levels 1 and 3
Public Sector Telco	Variation of the equal access business model, where the public sector acts in levels 1 and 2.	Simpler management of the network (both active and passive equipment), since it is performed by a single entity.	• Municipal authorities needs great technical and commercial expertise. • Negative impact on competition in active network equipment.
Sole Private Provider	The operation and the management of the active network equipment and services are offered by one private service provider, while a public utility or municipality manages level 1.	Cost-based leasing of passive infrastructures to the private company of levels 2 and 3.	• Negative impact on competition in services and networks • Fewer services to the customers

CONCLUSION

This article presents and compares the most important business models for the effective exploitation of the broadband municipal networks. The main objectives of such business models are the following:

- The passive network infrastructure may be used by a large number of service providers.
- The users have the choice of selecting one of the multiple services providers, according to their needs.
- Low operational expenditure (OPEX) and capital expenditure (CAPEX) must be ensured.
- Financial viability of all parts of the infrastructure must be achieved.
- The business model must motivate the competition for the benefit of the end users.

REFERENCES

COM. (2006, February 20). *Communications regulation and markets 2005* (11th report). Communication from the Commission to the Council, The European Parliament, The European Economic and Social Committee and the Committee of the Regions. European Electronic Commission of The European Communities Brussels.

Frieden, R. (2005). *Lessons from broadband development in Canada, Japan, Korea and the United States.* doi:10.1016/j.telpol.2005.06.002, 2005. Elsevier.

Government of Sweden. (2007). *The government and the government offices of Sweden, broadband for growth, innovation and competitiveness*. Retrieved May 28, 2008, from http://www.sweden.gov.se/download/9a39e612.pdf?major=1&minor=76048&cn=attachmentPublDuplicator_0_attachment

Heimgartner, A., Luke, M., Villa, N., & Johnston, P. (2005). *2010 broadband city: A roadmap for local government executives*. Cisco IBSG. Retrieved May 28, 2008, from http://www.cisco.com/web/about/ac79/docs/wp/2010/broadband/Broadband-City.pdf

Henderson, A., Gentle, I., & Ball, E. (2005). WTO principles and telecommunications in developing nations: Challenges and consequences of accession. *Telecommunications Policy, 29*(2-3), 205-221.

Hughes, G. (2003). Local & regional models for broadband deployment. In *Proceedings of eEurope: Broadband Local & Regional Best Practices Workshop.* Retrieved May 28, 2008, from http://europa.eu.int/information_society/eeurope/2005/doc/all_about/broadband/bb_regional/g_hughes.ppt

Kramer, R. D., Lopez, A., & Koonen, A. M. (2006, September 4-6). Municipal broadband access networks in the Netherlands—three successful cases, and how New Europe may benefit. In *Proceedings of the 1st International Conference on Access Networks*, Athens, Greece, (Vol. 267). New York: ACM. Retrieved May 28, 2008, from http://doi.acm.org/10.1145/1189355.1189367

Lehr, W., Sirbu, M., & Gillett, S. (2004, April 13-14). Municipal wireless broadband: Policy and business implications of emerging access technologies. *Competition in networking: Wireless and wireline*. London Business School.

Magnago, A. (2004). Open access—business models and operational costs. In *Proceedings of the Broadband Europe 2004.*

Monath, T., Cristian, N., Cadro, P., Katsianis, D., & Varoutas, D. (2003). Economics of fixed broadband access network strategies. *Communications Magazine, IEEE, 41*(9), 132-139.

OECD. (2003). *Policies for broadband development*. Recent OECD work on Broadband Committee for Information, Computer and Communications Policy, DSTI/ICCP(2003)13/FINAL/ADD/. Retrieved May 28, 2008, from www.oecd.org

Papacharissi, Z., & Zaks, A. (2006). Is broadband the future? An analysis of broadband technology potential and diffusion. *Telecommunications Policy, 30*(2006), 64-75, 70.

Shin, S.-C., Shin, S.-Y., & Han, S.-Y. (2004). Network performance monitoring system for SLA: Implementation and practices. In *Proceedings of the 6th International Conference on Advanced Communication Technology, 2004*, (Vol. 2, pp. 661-664, ISBN: 89-5519-119-7, Digital Object Identifier: 10.1109/ICACT.2004.1292952).

UTOPIA. (2003, November 26). *Utah's public-private fibre-to-the-premises initiative*. Utah Telecommunication Open Infrastructure Agency (UTOPIA) (White paper).

Wireless Philadelphia. (2005, February 9). *Wireless Philadelphia business plan.* Wireless Philadelphia Executive Committee. Retrieved May 28, 2008, from http://www.phila.gov/wireless/pdfs/Wireless-Phila-Business-Plan-040305-1245pm.pdf

KEY TERMS

Broadband: Broadband describes high-speed, high-capacity data communication making use of a wide range of technologies that often have diverse characteristics and seem appropriate for certain network scenarios and situations. There is no specific (international) definition or unique standard for broadband and the range of service speeds varies typically from 128 Kbps (or 200 Kbps according to the Federal Communications Commission—FCC—of United States) to 100 Mbps for broadband access. For the purpose of this article, we consider as broadband connection every connection which supports speeds greater than 200 Kbps.

Broadband Business Model: A business model determining the way in which the exploitation of a metropolitan, community-owned, optical network will be effectuated.

Broadband Network Passive Infrastructure: It is the physical infrastructure that is used to provide the broadband connectivity and may consist of fiber optic or copper cable.

Broadband Network Active Infrastructure: It consists of the elements used to transmit, forward and route informa-

tion data packets over fiber optic or copper cables. The main elements are switches and routers.

Broadband Services: They are the actual services offered to customers. Examples are: high speed Internet access (usually 10Mbit/s or higher); video telephony; video on demand; gaming portals; e-government and e-health services; Virtual Private Network services; video conferencing; Web hosting; data storage; video surveillance and so forth.

CAPEX: Acronym of the words Capital Expense. In the broadband networks it is the network implementation cost.

OPEX: Acronym of the words Operation Expense. In the broadband networks it is the operation and maintenance cost.

Business Process and Workflow Modeling in Web Services

Vincent Yen
Wright State University, USA

INTRODUCTION

In large organizations, typical systems portfolios consist of a mix of legacy systems, proprietary applications, databases, off-the-shelf packages, and client-server systems. Software systems integration is always an important issue and yet a very complex and difficult area in practice. Consider the software integration between two organizations on a supply chain; the level of complexity and difficulty multiply quickly. How to make heterogeneous systems work with each other within an enterprise or across the Internet is of paramount interest to businesses and industry.

Web services technologies are being developed as the foundation of a new generation of business-to-business (B2B) and enterprise application integration (EAI) architectures, and important parts of components as grid (*www.grid.org*), wireless, and automatic computing (Kreger, 2003). Early technologies in achieving software application integration use standards such as the common object request broker architecture (CORBA) of the Object Management Group (*www.omg.org),* the distributed component object model (DCOM) of Microsoft, and Java/RMI, the remote method invocation mechanism. CORBA and DCOM are tightly coupled technologies, while Web services are not. Thus, CORBA and DCOM are more difficult to learn and implement than Web services. It is not surprising that the success of these standards is marginal (Chung, Lin, & Mathieu, 2003).

The development and deployment of Web services requires no specific underlying technology platform. This is one of the attractive features of Web services. Other favorable views on the benefits of Web services include: a simple, low-cost EAI supporting the cross-platform sharing of functions and data; and an enabler of reducing integration complexity and time (Miller, 2003). To reach these benefits, however, Web services should meet many technology requirements and capabilities. Some of the requirements include (Zimmermann, Tomlinson & Peuser, 2003):

- *Automation Through Application Clients:* It is required that arbitrary software applications running in different organizations have to directly communicate with each other.
- *Connectivity for Heterogeneous Worlds:* Should be able to connect many different computing platforms.
- *Information and Process Sharing:* Should be able to export and share both data and business processes between companies or business units.
- *Reuse and Flexibility:* Existing application components can be easily integrated regardless of implementation details.
- *Dynamic Discovery of Services, Interfaces, and Implementations:* It should be possible to let application clients dynamically, i.e., at runtime, look for and download service address, service binding, and service interface information.
- *Business Process Orchestration Without Programming:* Allows orchestration of business activities into business processes, and executes such aggregated process automatically.

The first five requirements are technology oriented. A solution to these requirements is XML-based Web services, or simply Web services. It employs Web standards of HTTP, URLs, and XML as the lingua franca for information and data encoding for platform independence; therefore it is far more flexible and adaptable than earlier approaches.

The last requirement relates to the concept of business workflow and workflow management systems. In supply chain management for example, there is a purchase order process at the buyer's side and a product fulfillment process at the supplier's side. Each process represents a business workflow or a Web service if it is automated. These two Web services can be combined into one Web service that represents a new business process. The ability to compose new Web services from existing Web services is a powerful feature of Web services; however, it requires standards to support the composition process. This article will provide a simplified exposition of the underlying basic technologies, key standards, the role of business workflows and processes, and critical issues.

WHAT ARE "WEB SERVICES"?

The phrase "Web services" has been defined in many different ways (Castro-Leon, 2002; Ambrosio, 2002). In the working draft of Web Services Architecture (W3C, 2003), it is defined as:

"…a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards."

A simplified Web service architecture based on this definition is conceptually depicted in Figure 1.

Main features of Web services are that services (Burner, 2003):

1. Expose programmable application logic.
2. Are accessed using standard Internet protocol.
3. Communicate by passing messages.
4. Package messages according to the SOAP specification.
5. Describe themselves using WSDL.
6. Support the discovery of Web services with UDDI.
7. Are loosely coupled.

WEB SERVICES TECHNOLOGIES

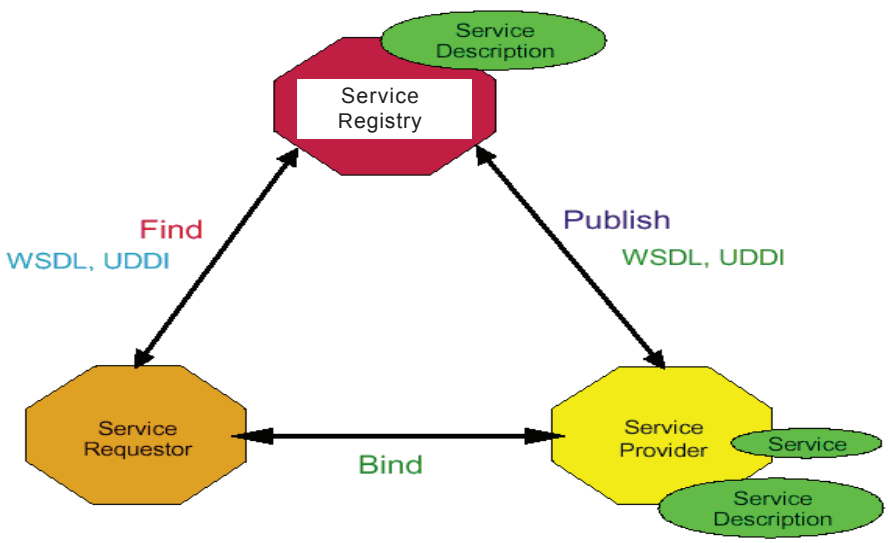
Three XML-based protocols—one for communication, one for service description, and one for service discovery—have become de facto standards (Curbera et al., 2002). They are:

- SOAP (the Simple Object Access Protocol) provides a message format for communication among Web services.
- WSDL (the Web Services Description Language) describes how to access Web services.
- UDDI (the Universal Description, Discovery, and Integration) provides a registry of Web services descriptions.

Another area of importance in Web services is the capability of constructing new composite Web services from existing Web services. Many standards in this area are being developed (Van der Aalst, 2003), for example, Business Process Execution Language for Web Services (BPEL4WS) by IBM and Microsoft (Fischer, 2002). It is not clear if there will be a common standard. However, regardless of the differences among vendor groups, the composition of Web services uses the concept of business processes and workflow management.

As noted earlier in this article, the development and deployment of Web services do not require a particular platform, nevertheless most Web services development is being accomplished today using either Microsoft .NET or

Figure 1. A simplified Web services architecture (W3C, 2003)



Sun Microsystems' J2EE specifications (Miller, 2003). It is not clear which of the two competing platforms is most suitable for the developers and their future directions (Miller, 2003; Williams, 2003).

THE ROLE OF BUSINESS PROCESSES AND WORKFLOW MODELING IN WEB SERVICES COMPOSITION

The need to integrate software components within and across companies is to economically re-use components for supporting business processes automation. A business process such as borrowing a book from a library may involve: 'check user identification', 'enter call numbers', and 'generate a receipt' activities. Each activity in the business process is a piece of work and is performed according to a sequence defined by business rules. That is, a business process contains a workflow (Allen, 2000).

The Workflow Management Coalition (Norin & Shapiro, 2002) defines the *business process* as:

"…a set of one or more linked procedures or activities which collectively realize a business objective or policy goal, normally with the context of an organizational structure defining functional roles and relationships."

Workflow management is further defined as:

"…the automation of a business process, in whole or part, during which documents, information, or tasks are passed from one participant to another for action, according to a set of procedural rules."

To understand complex business processes, workflow modeling techniques provide logical descriptions of the flow of activities that achieves the goal of the process. Companies compete to provide workflow-based tools for Web-service integration (Ganesarajah & Lupu, 2002). Such tools allow developers to use the concept of workflow to build complex business processes from Web services. Since a complex business process contains a large number of activities, managing the execution of these activities, (e.g., monitoring the workflow progress, sending activities to the right servers, and scheduling activities for execution) can be critical. The advent of the workflow management system or the workflow engine software is to serve this purpose.

Existing methods for creating business processes are not designed to work with cross-organizational components and Web services platforms (Peltz, 2003). This gives rise to multiple Web services standards for business process modeling and execution (Van der Aalst, 2003). However, regardless of standards, the composition of Web services requires the modeling (and construction) of a Web service from one party perspective and the interaction among each involved party. This is accomplished by using business process and workflow modeling in two aspects: 1) specifying internal details of the composition to offer an executable business process, and 2) planning the message sequences between parties and sources. The former is called *orchestration* and the latter is called *choreography* in creating business processes from composite Web services (Peltz, 2003).

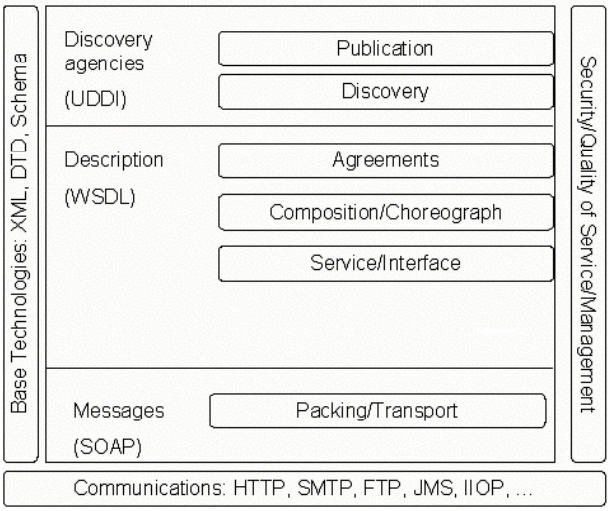
OUTSTANDING AND CHALLENGING ISSUES IN WEB SERVICES

Web service technology is still emerging, and researchers are still developing important functionalities for meeting technical and business requirements. Although some Web service standards such as SOAP, WSDL, and UDDI have been widely accepted, others such as Web services composition language are being developed by many organization and industry groups (Van der Aalst, 2003). These languages build upon the concept of workflow management systems to allow for orchestrating a group of required Web services in support of a business process. Due to real-world demands for more complex business problems integration, many companies and standards organizations have built extensions to the core specifications. Unfortunately, few of these extensions are interoperable. To resolve this problem and ensure application interoperability and extensibility, the W3C (2003) is developing a formal Web service architecture. There are different competing approaches (Kleijnen & Raju, 2003) to this problem, notably, the ebXML specifications suite, an independent technology for business-to-business integration (Patil & Newcomer, 2003). Additional challenges lie in XML and Web services security standards (Naedele, 2003). Obviously, the creation and support of Web service standards for mission-critical applications is a challenging, complex, and time-consuming process. Figure 2 adapted from W3C (2003) shows more completely the critical service areas of Web services.

A brief explanation of some Web service architectural components follows:

- *Packing/Transport:* This is the message layer that packs information into messages and transports it between parties.
- *Service/Interface:* Describes operational interfaces of a Web service.
- *Composition/Choreography:* Describes how services interact with each other in business processes.
- *Agreements:* The service level agreement defines the specific performance, usage, costs, and so forth, and the business level agreement defines a contractual

Figure 2. A Web service architecture



agreement between the two business partners in business engagement using Web services.

- *Security Management and Quality of Service:* Describes reliability and security characteristics.

Each component has varying degrees of achievement by industries and standard setting organizations. However, standards are not enough to realize the great expectations of Web services. Langdon (2003) describes several key inhibitors of Web services adoption, including the lack of complementary service providers for metering, accounting, and billing services; the lack of ready-to-use Web services from either internal sources or third parties; and the lack of third-party support of how to decompose an automation problem and how to deliver it.

CONCLUSION

Web services involve many technologies and standards. Although a very young field, it has made tremendous progress in the last two years. Software vendors like Microsoft's .NET and Sun's J2EE have tools for Web services development. However, the field has many challenging issues, with standards being one of them. This article provides an overview of Web services in general and the role of workflow in developing a Web services application in particular. Since workflow modeling in the composition of Web services utilizes knowledge of business processes, that is where MIS professionals should be actively involved.

REFERENCES

Allen, R. (2000). Workflow: An introduction. In L. Fisher (Ed.), *The workflow handbook 2001*. Workflow Management Coalition.

Ambrosio, J. (2002). Web services: Report from the field. *Application Development Trends, 9*(6).

Burner, M. (2003). The deliberate revolution. *ACM Queue, 1*(1), 28-37.

Chung, J., Lin, K. & Mathieu, R. (2003). Web services computing: Advancing software interoperability. *Computer, 36*(10).

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N. & Weerawarana, S. (2002). Unraveling the Web services web: An introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing, 6*(2), 86-93.

Ferris, C. & Farrell, J. (2003). What are Web services? *Communications of the ACM, 46*(6), 31.

Fischer, L. (2002). The WfMC heralds BPEL4WS standards for business process management industry. Retrieved from *xml.coverpages.org/WfMC-Heralds-BPEL4WS.html*.

Ganesarajah, D. & Lupu, E. (2002). Workflow-based composition of Web services: A business model or a programming paradigm? *Proceedings of the 6th International Enterprise Distributed Object Computing Conference*, Lausanne, Switzerland.

Kleijnen, S. & Raju, S. (2003). An open Web services architecture. *ACM Queue, 1*(1).

Kreger, H. (2003). Fulfilling the Web services promise. *Communications of the ACM, 46*(6), 29-34.

Langdon, C.S. (2003). The state of Web services. *Computer, 36*(7), 93-94.

Marin, M., Norin, R. & Shapiro, R. (Eds.). (2002). *Workflow process definition interface—XML process definition language*. Document No. WFMC-TC-1025. Document Status: 1.0 Final Draft.

Miller, G. (2003). .NET vs. J2EE. *Communications of the ACM, 46*(6), 64-67.

Naedele, M. (2003). Standards for XML and Web services security. *Computer, 36*(4), 96-98.

Patil, S. & Newcomer, E. (2003). ebXML and Web services. *IEEE Internet Computing, 7*(3), 74-82.

Peltz, C. (2003). Web services orchestration and choreography. *IEEE Computer, 16*(10).

Van der Aalst, W.M.P. (2003). Don't go with the flow: Web services composition standards exposed. *IEEE Intelligent Systems,* (January/February).

W3C. (2003, August). *Web services architecture.* W3C Working Draft. Retrieved from *www.w3.org/TR/2003/WD-ws-arch-20030808/*

Williams, J. (2003). J2EE vs. .NET. *Communications of the ACM, 46*(6), 59-63.

Zimmermann, O., Tomlinson, M. & Peuser, S. (2003). Perspectives on Web services. Berlin, Heidelberg, New York: Springer-Verlag.

KEY TERMS

EAI: Projects involving the plans, methods, and tools aimed at modernizing, consolidating, and coordinating the computer applications and data in an enterprise.

Grid Computing: A form of distributed computing that involves coordinating and sharing computing, application, data, storage, or network resources across dynamic and geographically dispersed organizations.

HTML (Hypertext Markup Language): A standard language for representing text, formatting specifications and hyperlinks.

HTTP (Hypertext Transfer Protocol): The standard for requesting and transmitting information between a browser and a Web server.

Java/RMI: A Java application programming interface known as remote method invocation.

Protocol: A set of rules and procedures governing transmission between two points in a network.

URL (Universal Resource Locator): A text string used as a reference to a Web resource. A URL consists of a protocol, a host name, and a document name.

XML (Extensible Markup Language): An extension of HTML that is being extensively used for transmitting data/information on the Internet.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 345-349,

Business Processes and Knowledge Management

John S. Edwards
Aston Business School, UK

INTRODUCTION

Knowledge has been a subject of interest and inquiry for thousands of years since at least the time of the ancient Greeks, and no doubt even before that. "What is knowledge" continues to be an important topic of discussion in philosophy.

More recently, interest in managing knowledge has grown in step with the perception that increasingly we live in a knowledge-based economy. Drucker (1969) is usually credited as being the first to popularize the knowledge-based economy concept by linking the importance of knowledge with rapid technological change in Drucker (1969). Karl Wiig coined the term knowledge management (hereafter KM) for a NATO seminar in 1986, and its popularity took off following the publication of Nonaka and Takeuchi's book "The Knowledge Creating Company" (Nonaka & Takeuchi, 1995). Knowledge creation is in fact just one of many activities involved in KM. Others include sharing, retaining, refining, and using knowledge. There are many such lists of activities (Holsapple & Joshi, 2000; Probst, Raub, & Romhardt, 1999; Skyrme, 1999; Wiig, De Hoog, & Van der Spek, 1997). Both academic and practical interest in KM has continued to increase throughout the last decade.

In this article, first the different types of knowledge are outlined, then comes a discussion of various routes by which knowledge management can be implemented, advocating a process-based route. An explanation follows of how people, processes, and technology need to fit together for effective KM, and some examples of this route in use are given. Finally, there is a look towards the future.

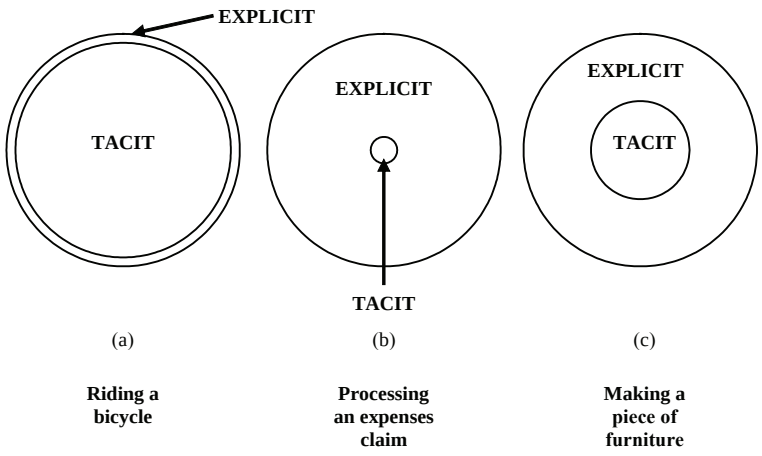
BACKGROUND

Types of Knowledge: Tacit and Explicit

Nonaka et al.'s book (1995) popularized the concepts of tacit and explicit knowledge, as well as KM more generally. They based their thinking on that of Michael Polanyi (1966), expressed most memorably in his phrase "we know more than we can tell."

It is, however, most important to realize that tacit and explicit knowledge are not mutually exclusive concepts. Rather, any piece of knowledge has both tacit and explicit elements, as shown in Figure 1. The size of the inner circle represents the proportion of tacit knowledge: the tacit core at the heart of the knowledge that we "cannot tell." Figure 1(a) shows a case where the knowledge is almost entirely tacit, as in riding a bicycle. Figure 1(b) shows mainly explicit knowledge where the tacit core is very small, for example

Figure 1. The relationship between tacit and explicit knowledge



how to process a claim for travel expenses in an organization. Figure 1(c) shows an intermediate case such as making a piece of furniture where substantial amounts of both tacit and explicit knowledge are involved.

KM Strategies

Hansen, Nohria, and Tierney (1999) identified that there are two fundamental KM strategies, codification and personalization. Codification concentrates more on explicit knowledge (often relying very heavily on information technology), personalization more on tacit knowledge. Again, it is important to realize that these are not mutually exclusive, and that a strategy combining elements of both is likely to be the most successful.

ROUTES TO IMPLEMENTING KM

Many organizations have found it difficult to implement knowledge management systems successfully. Identifying who is involved in knowledge management, what knowledge is being managed, and why is it being managed can be problematic. The routes they have attempted to follow can be put into five generic categories, which will now be described.

Knowledge World Route

A substantial amount of the literature on knowledge management addresses knowledge at the level of the whole organization, or in a "world of knowledge" that is not specifically linked to the activities that a particular organization carries out. On an abstract level, such discussion of knowledge management can be extremely valuable. However, it has weaknesses in terms of practical implementation. For example, it is necessary not only to understand how individuals learn, but also how they learn in a given organization, and how the organizational systems may help or hinder the individual's learning process. The same issue applies even more forcefully to group learning since the organization provides a crucial element of the group's context.

The practical focus in Nonaka et al. (1995) was very much on knowledge creation. As a result, organizations attempting to follow their principles for other aspects of KM, such as sharing or retaining knowledge, have sometimes found it difficult to make a specific connection from abstract ideas about knowledge to what the organization actually does, or could do, or should do. Often only the "why" is present, not the "who" or even the "what." Something more concrete is needed.

IT-Driven Route

This route assumes that the fundamental requirement is for the codification of as much knowledge as possible. Advocates of this approach sometimes refer to this as "extracting" the knowledge from the people who possess it; see for example Johannsen and Alty (1991). This is an inadvisable term to use for two reasons. First, it is logically incorrect; their knowledge is being shared, not extracted. The people still have the knowledge after the "operation" has taken place. Second, it gives the people the wrong impression—that their knowledge is being taken away. This is not a recipe to encourage their cooperation. In all but the smallest of organizations, such a codification task evidently requires IT support, and the thrust of this route is that once the "correct" form of IT support for managing knowledge has been chosen, it is simply a matter of a great deal of hard work.

This technology-driven route works well in a limited range of situations where the "what" questions are most important, for example, where the main KM task is managing the knowledge held by a company in the form of patents. In other circumstances, it may not achieve any improvement in knowledge management at all. One example of this from the author's experience is of a heavy manufacturing firm. Knowledge management in this organization was seen solely as an information systems issue; the KM group was part of the information systems department. The "solution" was seen in terms of the implementation of a knowledge sharing system based on Lotus Notes™. However, there was no real consideration as to who would share what knowledge with whom or for what specific purpose ("why"). Consequently, the eventual use of the installed IT was poor; the only really successful use was by the knowledge management project team itself, where the "who" and "why" questions had been properly addressed, as well as the "what" questions.

Functional Route

An alternative route that has the potential to address the "who," "what," and "why" questions is to organize the implementation around the existing organizational structure. The most commonly found structural elements intended to facilitate learning and knowledge sharing in organizations are departmental groupings based on functions. These have clear advantages in terms of what might be termed professional development and allegiance. Davenport and Prusak (1998) report examples of successful knowledge transfer between groups of surgeons, and groups of tunneling engineers, amongst others. However, this functional route also has the disadvantage that it encourages the compartmentalization of knowledge. This problem can only worsen over time, as specialisations multiply and sub-divide. In addition, professional divisions can actively prevent sharing of knowledge. It has, for example, taken decades for hospital doctors in the

UK National Health Service to allow other professionals such as pharmacists and physiotherapists to participate in decision-making about treatment of individual patients on an equal footing. On a wider scale, modern Western medical science has come to separate "diet" and "drugs," at least until the very recent past, in a way that Chinese medicine, for example, never has done. The problems of running an organization in this manner, and the "functional silos" mentality that tends to result, were recognized by authors such as Hammer (1990) as part of the business process re-engineering movement, when KM was in its infancy.

Therefore, although the functional route to implementation will allow some improvement in KM, progress may be limited by the characteristics of the existing structure, and in the worst cases (for example where transferring knowledge between functions is the greatest KM issue in the organization) this route may be counter-productive.

People-Centric Route

A people-centric route to KM is the essence of the Hansen et al. (1999) personalization strategy. By definition, such an approach, appropriately implemented, will answer all the "who" questions that might be involved in KM. Thus in organizations where there is general consensus on "what" knowledge is important and "why" it needs to be managed, such a route should prove effective.

However, as was mentioned in the previous sub-section, organizations have become increasingly diverse in their activities, and in the range of specialized knowledge that they need to access. This means that consensus even on what knowledge the organization has, never mind what is important, may be difficult to achieve. On the one hand, it may not be easy for a particular specialist to fully appreciate "what the organization does." Equally, even the most conscientious senior manager will find it literally impossible to understand all the expertise and knowledge possessed by the specialists in his or her organization. To repeat the quotation from Hewlett Packard CEO Lew Platt (Davenport et al., 1998, p. xii) "If HP knew what HP knows, we would be three times as profitable."

Business Processes Route

The managers in an organization have to translate the goals of any strategic program or initiative—whether on knowledge management or something else--into practical, implementable reality. In other words, to connect with "what the organization does." Various management thinkers have presented models of this, for example:

- Porter's value chain (Porter, 1985)
- Earl's view of core processes (Earl, 1994), the ones that are done directly for external customers
- Beer's "system ones" (Beer, 1985), the systems that make the organization what it is
- Core competences/competencies as espoused by Hamel and Prahalad (1994)

Although there are some significant differences between them, their common theme is that the effectiveness—indeed the competitive advantage—of organizations depends not on how they are structured, or on what resources they have, but on what they do. In the terminology of this article, this means their underlying business processes. Note that the term business processes is used throughout, but such processes exist equally in not-for-profit organizations.

- Business processes possess five characteristics that justify their use as a foundation for knowledge management in organizations.
- Business processes have identifiable customers, whether internal or external. Knowledge is of little relevance to the organization unless put to use for a customer of some kind.
- Business processes cut across organizational boundaries. Knowledge flows do not need to, and should not, obey the artificial boundaries within an organization.
- Business processes consist of a structured set of activities. Choosing the appropriate way to structure activities is an important part of the knowledge.
- Business processes need to be measured. Without some form of measurement as a comparison, knowledge cannot be validated.

While the parts of a business process are important, the overriding requirement is that the overall process works. Valid knowledge in an organizational context must take a holistic view.

An additional argument (Braganza, 2001) is that viewing knowledge management in terms of an organization's processes gives a much-needed demand-side view of knowledge. This is complementary to the supply-side view of knowledge that stems, for example, from considerations 'of data leading to information leading to knowledge'. Beer and Earl particularly concentrate on this demand-side perspective. Beer indeed goes even further, to include the informal processes and activities of the organization as well as the more formalized ones.

Completing this argument for a greater use of the business processes route, the knowledge that an organization requires must, logically, be related not just to what that organization does, but also to how it does it. Thus people in organizations should think about this knowledge, and how to manage it, by reference to that organization's business processes.

PEOPLE, PROCESSES, AND TECHNOLOGY

From the earlier discussion, it may be seen that, whichever route is chosen, effective KM requires the consideration of both tacit and explicit knowledge. The need is to coordinate people, processes, and technology successfully using some kind of KM system. It is important to realize that there is more to a KM system than just technology, and that any deliberate, conscious attempt to manage knowledge in an organization amounts to a KM system. The interaction of the three elements, people, processes, and technology, is shown in Figure 2.

Not only does a knowledge management system consist of more than technology, it is important to realize that the technology used to support KM does not have to be "KM software." Several studies have found that generic software such as e-mail or an Intranet may be at least as important as specific software (Edwards, Shaw, & Collier, 2005; Zhou & Fink, 2003).

KM BY A BUSINESS PROCESSES ROUTE

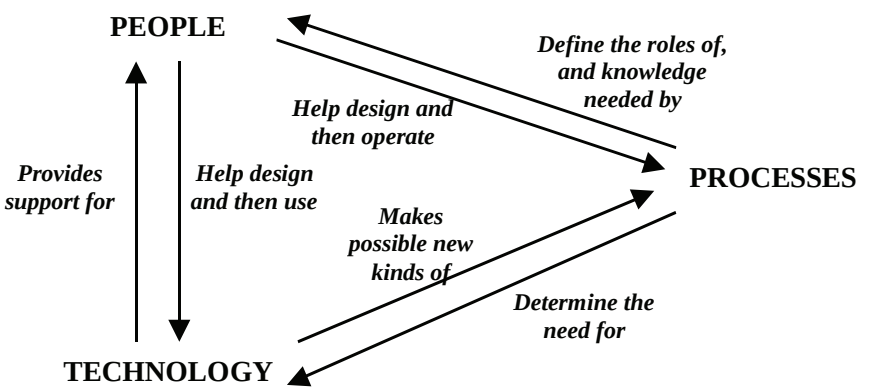
The business processes route to KM is becoming more widely adopted. The Singapore Ministry of Manpower (Fung, 2006) has used business strategies and processes to drive its blended KM approach rather than "plunging straight into analyzing the KM elements" (Fung, 2006, p. 31). Fung concluded that this dramatically increased the business relevance of the KM initiatives. Several organizations in the construction industry have used an approach based on incorporating KM within a process improvement model (Jeong, Kagioglou, Haigh, Amaratunga, & Siriwardena, 2006). Geisel (2005) gives examples of how the insurance industry in the USA and in Europe has been building knowledge into its business processes. Dayan and Evans (2006) consider KM as a way to support the process-focussed capability maturity model, and explain how this approach has been used in KM at Israel Aircraft Industries. Jambekar and Pelc (2006) describe a process-driven approach to KM in an anonymous company manufacturing electronic instruments.

Many of these examples involved substantial use of information technology. However, that does not have to be the case. The author's group has been working with a component manufacturer in the aerospace industry, whose KM initiative also has an explicit process focus. Typically, their manufacturing processes use a machine operated by one person. The operators' choice of the best way to retain and share knowledge does not use IT at all (except for a word processor). The agreed best operating procedure, with illustrations, is put on a laminated sheet of paper mounted near the machine, which includes the names of the people who had contributed to designing the procedure. A suitable pen is provided to annotate the laminated sheet. At regular intervals office staff come round to produce a revised version of any of the "standard operating sheets" that have been annotated.

FUTURE TRENDS

A further justification for the use of business processes as the foundation for implementing knowledge management is that they are now becoming part of the mainstream of management thought. For example, the latest version of the ISO9000 family of standards for quality management systems, including ISO9001: 2000, is constructed on the basis of a "process approach." The ISO9000 term realisation process is equivalent to Earl's core process or Beer's primary

Figure 2. People, processes, and technology in a KM system



activity as discussed earlier. Significantly, the latest editions of strategic management textbooks (Johnson, Scholes, & Whittington, 2004) typically discuss the business process view of organizations, whereas earlier ones did not.

It seems clear, therefore, that the business processes route to implementing knowledge management is likely to become more common in future, and that this will encourage the development of ever more appropriate information technology for supporting it. Equally, new information technologies will enable new types of process. Intelligent agents, smart cards, and camera phones, for example, all offer different possibilities for supporting KM which have only just begun to be considered.

CONCLUSION

This article has considered the implementation of knowledge management systems by looking at five different generic routes towards achieving it: knowledge world, IT-driven, functional, people-centric, and business processes. While each of these routes has some merits, it has been argued that the business processes route offers potential for the greatest integration between knowledge management and "what the organization does." It is thus likely to be increasingly common in the future.

REFERENCES

Beer, S. (1985). *Diagnosing the system for organisations*. Chichester: Wiley.

Braganza, A. (2001). Knowledge (mis)management…and how to avoid it. Information Management 2001. Olympia, London.

Davenport, T. H. (1993). *Process innovation: Reengineering work through information technology*. Boston: Harvard Business School Press.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Boston: Harvard Business School Press.

Dayan, R., & Evans, S. (2006). KM your way to CMMI. *Journal of Knowledge Management, 10*(1), 69-80.

Drucker, P. F. (1969). *The age of discontinuity*. London: Heinemann.

Earl, M. J. (1994). The new and the old of business process redesign. *The Journal of Strategic Information Systems, 3*(1), 5-22.

Edwards, J. S., Shaw, D., & Collier, P. M. (2005). Knowledge management systems: Finding a way with technology. *Journal of Knowledge Management, 9*(1), 113-125.

Fung, M. (2006). Breaking silos at Singapore Ministry of Manpower. *Knowledge Management Review, 9*(2), 30-33.

Geisel, R. W. (2005). The marriage of knowledge to business processes. *Business Insurance* (3), 18-20.

Hamel, G., & Prahalad, C. K. (1994). *Competing for the future*. Boston: Harvard Business School Press.

Hammer, M. (1990). Re-engineering work: Don't automate, obliterate. *Harvard Business Review, 68*(4 (July/August)), 104-112.

Hammer, M., & Champy, J. (1993). *Reengineering the corporation: A manifesto for business revolution*. London: Nicholas Brealey.

Hansen, M. T., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? *Harvard Business Review, 77*(2), 106-116.

Holsapple, C. W., & Joshi, K. D. (2000). An investigation of factors that influence the management of knowledge in organizations. *Journal of Strategic Information Systems, 9*, 235-261.

Jambekar, A. B., & Pelc, K. I. (2006). A model of knowledge processes in a manufacturing company. *Journal of Manufacturing Technology Management, 17*(3), 315-331.

Jeong, K. S., Kagioglou, M., Haigh, R., Amaratunga, D., & & Siriwardena, M. L. (2006). Embedding good practice sharing within process improvement. *Engineering, Construction, and Architectural Management, 13*(1), 62-81.

Johannsen, G., & Alty, J. L. (1991). Knowledge engineering for industrial expert systems. *Automatica, 27*(1), 97-114.

Johnson, G., Scholes, K., & Whittington, R. (2004). *Exploring corporate strategy: Text and cases* (7th ed.). Harlow: Financial Times Prentice Hall.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. New York and Oxford: Oxford University Press.

Polanyi, M. (1966). *The tacit dimension*. Garden City, NY: Doubleday.

Porter, M. E. (1985). *Competitive advantage: Creating and sustaining superior performance*. New York, London: Collier Macmillan.

Probst, G., Raub, S., & Romhardt, K. (1999). *Managing knowledge: Building blocks for success*. Chichester: Wiley.

Skyrme, D. J. (1999). *Knowledge networking: Creating the collaborative enterprise*. Oxford: Butterworth-Heinemann.

Wiig, K. M., De Hoog, R., & Van der Spek, R. (1997). Supporting knowledge management: A selection of methods and techniques. *Expert Systems with Applications, 13*(1), 15-27.

Zhou, A. Z., & Fink, D. (2003). Knowledge management and intellectual capital: An empirical examination of current practice in Australia. *Knowledge Management Research & Practice, 1*(2), 86-94.

KEY TERMS

Business Process: A structured, measured set of activities designed to produce a specified output for a particular customer or market. (Davenport, 1993, p.5)

Business Process Reengineering: The fundamental rethinking and radical redesign of business processes to achieve dramatic improvements in critical, contemporary measures of performance, such as cost, quality, service, and speed. (Hammer & Champy, 1993, p. 32)

Demand-Driven View of Knowledge: A view of knowledge stemming from the requirements of the organization; for example, what knowledge is needed to carry out a particular activity and how can it be applied?

Explicit Knowledge: Knowledge that has been (or can be) codified and shared with others.

Knowledge Management: Supporting and achieving the creation, sharing, retention, refinement, and use of knowledge (generally in an organizational context).

Knowledge Management Software: Software specifically intended for knowledge management, such as data mining and "people finder" software.

Knowledge Management System: A combination of people, processes and technology whose purpose is to perform knowledge management in an organization.

Supply-Driven View of Knowledge: A view of knowledge stemming from the knowledge itself rather than its uses. Often related to a continuum data-information-knowledge.

Tacit Knowledge: Knowledge that is difficult or impossible to express, except by demonstrating its application.

Business Relationships and Organizational Structures in E-Business

Fang Zhao
Royal Melbourne Institute of Technology, Australia

INTRODUCTION

In today's e-business, context, technology, customers, competitors, and partners can change rapidly. Technology can become obsolete in the blink of an eye and customers can appear and disappear with a keystroke. There are practically no barriers to new entrants (competitors) in an e-business world. Likewise, e-business partnerships and virtual organizations become ephemeral and opportunistic in nature. This article explores the dynamics of the changing nature, process, and practice of business relationships and network form of organizations in the cyberspace. It also identifies and discusses a series of management issues raised in the processes of e-partnerships and virtual organizations.

BACKGROUND

The virtual organization, which is actually a network form of organization, is a revolution in organizational design and has changed the definitions, boundaries, and forms of inter-organizational collaboration and partnerships. The network form of organizations is the defining business transformation of this generation (Hagel & Singer, 2000; Jin, 1999; Malone & Davidow, 1994). Cisco, for example, is a network of suppliers, contract manufacturers, assemblers, and other partners, which is connected through an intricate web of information technology. Seventy percent of Cisco's product is outsourced to its e-partners through Cisco's network (McShane & von Glinow, 2000).

As previously shown, virtual organizations rely on IT network and e-partnership. Theoretically, e-partnership refers to a partnership relying on electronic (information) technologies to communicate and interact amongst partners. In practice, the term e-partnership is mostly associated with e-commerce or e-business partnerships. It may take different forms and involve various partners from or between virtual enterprises and brick-and-mortar companies depending on the nature of e-business activities. It flourishes in particular in e-supply chains through adding electronic components to the business partnership across firms (O'Toole, 2003, Zhao, 2006). For example, in the manufacturing industry, the e-partners may include raw materials providers, component manufacturers, final assembly manufacturers, wholesalers, distributors, retailers, and customers (Cheng, Li, Love, & Irani, 2001). This supply chain may involve a number of hundreds or thousands of suppliers and distributors. The use of Internet and other electronic media and the introduction of inter-organizational information systems are constitutive to e-partnerships and lead to the growth of virtual organizations.

E-partnerships share some common characteristics with traditional inter-organizational partnerships (Segil, 2004). But they are different in many ways and thus require different strategies and structures to manage them. Bell (2001) and de Man, Stienstra, and Volberda (2002) studied the differences and found that e-partnerships are generally entrepreneurial and less planned in nature, must move at Web speed, require flexible network structure, and have a short lifespan.

E-partnerships as "a new breed of online alliances are fast emerging as the result of an incredible amount of Internet business in recent years" (Trask, 2000, p. 46). Entering an e-partnership is no longer a soft option but a vital need for gaining a competitive advantage and customer satisfaction in the trend of economic globalization (by globalization it means an increasing trend of economic integration worldwide). On the other hand, globalization is pushing companies to build informal network organizations, such as virtual organizations, that are able to work in a faster, cheaper and more flexible way. E-partnerships and virtual organizations are products of the globalization and IT advancement over the past decade and they have fundamental synergy between them. They interrelate and interact with each other in this digital era.

ADVANTAGES

The greatest advantage of e-partnership and virtual organization lies in the fact that they eliminate the physical boundaries of organizations, and that cross-functional teams and organizations are able to operate and collaborate across space and time by communicating with each other via electronic channels. The Internet becomes the most important interface between participating organizations, teams, and individuals. E-partnerships and virtual organizations enable businesses to sell and deliver products and services across the world in a more efficient way in terms of speed and cost. Amazon.

com, Priceline.com, and E*Trade are some of the successful e-businesses who have depended on, and maximized profits from, e-partnerships.

Other perceived benefits of e-partnerships and virtual organizations may include greater business opportunities, better integration of suppliers and vendors, better management information, lower operational costs, better market understanding and expanded geographical coverage (Damanpour, 2001). E-partnerships and virtual organizations may also offer the opportunity of consolidating resources of all partners and organizational flexibility as other forms of inter-organizational partnerships and alliances do.

In this rapidly changing competitive landscape, few organizations can rely on their internal strengths only to gain a competitive advantage in national and/or international markets. Inter-organizational collaborations, alliances, joint ventures, partnering, and the like are gaining unprecedented momentum, regardless of their organizational and management structures and styles and communication channels. An organization's resources are limited in one way or another. Forming a business e-partnership and taking a network form of organization is increasingly one of the most popular strategies available to an organization to take advantage of an Internet Highway on the one hand and share risks, capabilities, and revenue with partners on the other. The driving forces behind building an e-partnership and a virtual organization share a lot in common with those driving any other forms of inter-organizational collaborations. They include:

- To gain a competitive advantage or increase market share in national and/or global markets
- To share revenue and expand sales between merchant and partners
- To prevent competition loss
- To meet changing demands of customers and markets
- To gain core competencies from competitors (Dussauge & Garrette, 1999; Sierra, 1994; Trask, 2000)

KEY ISSUES

However, like e-business and e-commerce, e-partnership is also facing a range of issues related to the use of Internet as well as the reliance on inter-organizational interfaces. The key issues identified are:

- Challenges and risks of e-partnerships and virtual organizations
- Productivity and revenue sharing in e-partnerships and virtual organizations
- Transferring and sharing core competencies between participating organizations
- Power disparity
- Quality and effectiveness of communication

Addressing each of these issues has posed a formidable task in front of e-managers of various kinds of inter-organizational collaboration through electronic technologies and e-network. The following discussion explores each of these issues in detail.

Challenges and Risks

On the technological side, companies that are involved in e-partnerships must participate in external business relationships by using computer interactions. This forces e-managers to re-engineer their IT strategies and resources and re-think their ways of communicating and doing business with e-partners in a virtual environment. Main issues to be considered are IT infrastructure and managers' and operatives' knowledge and skills associated with e-business.

On the human resources' side, e-managers are surely confronting management complexities of making cooperation work. One of the biggest challenges is conflict and differences in organizational and country cultures and systems. Each organization has its own culture developed from its own particular experience, its own role and the way its owners or managers get things done (Hellard, 1995). In addition to the cultural differences at organizational level, multi-national e-partnerships encounter inevitably barriers caused by cultural differences between nations such as clashes between western and eastern cultures. Differences exist in systems including taxation systems, online intellectual property, and online trade and law. For example, EU member states must enact legislation to ensure that transfers of data outside their boundaries are allowed only to jurisdictions that can offer adequate protection of the data. The US believes that minimal domestic regulation would foster cross-border Internet trade (Damanpour, 2001). Managing the culture and system differences across organizations and across nations is one of the high agendas that challenge managers of e-partnerships and virtual organizations.

While the Internet and network organizations facilitate improved communication of data, information and knowledge, they give rise to issues and problems of privacy, data security, and intellectual property protection in the Internet. The information database created through Internet transactions may lead to legal disputes among e-partnerships over ownership of the IP and possible loss of the potential profit generated from the IP (Greif, 2000). Moreover, electronic research projects usually involve new technologies and innovative development, which creates a high level of technological, commercial, and legal risks for every organization involved.

Productivity and Revenue Sharing

The primary aim of building e-partnerships and virtual organizations is to generate more profit and achieve the best business results through taking advantage of online resources and extensive e-network. Trask (2000, p. 46) considered that "a well-designed revenue-sharing program may be the best and fastest way to generate online business." Revenue sharing becomes the most important issue in e-partnerships and virtual organizations when productivity increases and revenue goes up. The nature, timing, and amount of compensation (in the form of referral fees, royalty, and commission) together with the financial stability and honesty of commission reporting are core considerations of e-partners and crucial factors of success in sustaining e-partnerships.

Transferring and Sharing Core Competencies

According to Lei, core competencies comprise a company's specific and special knowledge, skills, and capabilities to stand out amongst competitors. They are intangible and an integrated part of a company's intellectual capital and un-tradable asset rather than legally protected intellectual property (Lei, 1997, p. 211). Inter-organizational collaboration provides an opportunity for participating organizations to acquire and absorb the core competencies from each other (Couchman & Fulop, 2000). This opportunity is particularly valuable for innovative business such as e-business. However, the greatest barrier is competitive concerns over information leakage. This is an unavoidable dilemma facing e-partnerships, which makes it difficult for e-partners to achieve the potential that IT technology can offer.

Power Disparity

It is normal that a decision-making body of inter-organizational collaboration and partnership is proportionately represented by participating organizations in terms of equity holdings in an online joint venture. It should be noted that due to difference in equity holdings, power disparity occurs and is likely to affect performance of inter-organizational collaboration, although division of power and responsibility has been clearly defined in legally binding agreements between e-partners.

Quality and Effectiveness of Communication

Networking and communication play a key role particularly in coordinating and liaising inter-organizational collaboration. Expanding e-networks and achieving effective communication amongst e-partners are a top priority. Like culture and commitment, communication is a soft outcome of a total quality partnership approach and the foundation for inter-organizational collaboration (Aggarwal & Zairi, 1998; Hellard, 1995; Rounthwaite & Shell, 1995). Effective networking and communication help to eliminate barriers to collaboration. Therefore, continuous improvement of quality and effectiveness of communication amongst partners is another key issue in the agenda of e-partnerships and virtual organizations.

OPTIONS

While sufficient support of IT infrastructure and resources are definitely important to successful e-partnerships and virtual organizations, reducing potential financial, commercial, and legal risks and effectively dealing with human and cultural factors exceed the complexities of technical setup and support in building e-partnerships and virtual organizations. Organizational culture and human resources are increasingly becoming important sources of competitive advantage, as they are difficult for competitors to imitate. How to foster a robust culture and capitalize on the core competence derived from e-partnerships is a crucial and tough issue for managers. Other critical success factors for e-partnerships and virtual organizations include:

- Level of accessibility, security, and compatibility of inter-organizational information systems
- Level of traffic in collaborative e-commerce activities
- Level of customer service and e-partner support service
- Level of transferring and sharing information and knowledge between e-partners
- Building and sustaining an effective virtual network structure amongst e-partners
- Level of individual and organizational commitment to e-partnerships
- Level of mutual trust, understanding, respect, and openness
- Level of corporate and business ethics and integrity
- Level of credibility of e-partners in relation to financial situation and business experience
- Level of mutual benefit through revenue sharing
- Effectiveness and efficiency of real-time commission reporting system
- Level of performance and productivity of e-partners
- Actively pursuing and sharing core competencies
- Willingness to share power and empower amongst e-partners
- Quality and effective networking and continuous improvement of communication (Singh & Byrne, 2005)

Achieving the best collaboration among e-partners requires more than tangible resources like IT infrastructure and support. Successful e-partnership needs a high level of intangible commitment and efforts to understand the needs and values of e-partners and customers. By resorting to a total quality partnership approach, it means that business ethics, integrity, honesty, trust, and sharing are required of e-managers of inter-organizational entities at the top and of the individuals and teams throughout the entire virtual organization (Aggarwal et al., 1998; Hellard, 1995; Rounthwaite et al., 1995). Disputes and conflicts caused by culture and system differences like those illustrated in this article could be reduced if e-partners could accept the differences and maintain a flexible and realistic attitude towards the differences.

FUTURE TRENDS

"The world has finally become a global village, not just in rhetoric, but in reality" (Hennessey, 2000, p.34). In today's corporate world, it will be more difficult for businesses to survive without joining an e-partnership and taking advantage of the Internet. The fast expansion of Amazon.com, travel.com, and the like through e-partnerships, online syndicates, and e-networks and their business success reinforce the importance of online strategic alliance. Corporate e-partnerships and network-based organizations will be a crucial factor and play a key role in the future development of online business activities. The future trends will be characterized by:

- More mature (rather than experimental) nature of e-commerce and e-business practices in terms of the scope, quality and credibility of online customer services and products
- More needs for devising and popularizing e-supply chains due to the needs for integrating the flow of information with the flow of goods (Kotzab, Skjoldager, & Vinum, 2003; van Hoek, 2001)
- Greater monopoly of the flow of e-commerce and e-business by bigger online syndicates like Amazon.com through building extensive online alliances with online retailers and suppliers (Werbach, 2000)
- More brick-and-mortar companies moving online to expand their business scope and capitalize on the abundance of the Internet
- Greater reliance on joint efforts across nations in online legislation to protect IP, security, and privacy of e-commerce and e-business
- Greater challenge for dot.com industries to achieve sustainability, due to a more uncertain economic environment and the increasing complexities of new technologies and the more globalized economy

CONCLUSION

Today's complex and volatile business world calls for changes and alternatives to old and conventional paradigm of organizational design and new ways of doing business with others. E-business becomes one of the most important forces shaping today's business. Virtual corporations and e-partnerships become increasingly popular in the perception of managers and in business operations. In such circumstances, it is important that e-business managers have an insightful knowledge and are well prepared to deal with the complexities of e-partnerships and virtual organizations. This article provides a better understanding of the crucial issues in cross-firm business processes and inter-organizational partnerships in the cyberspace.

Running inter-organizational partnerships implies multiplication of decision-making bodies from each participating organization and potential clash of interest and values amongst participants. As illustrated in this article, total quality partnership embodies the fundamental principles for managing collaborative partnerships including e-partnerships and can be developed and extended to help inter-organizational collaboration to achieve desired outcomes. However, managing e-partnership and virtual organization is more complex than managing intra-organizational collaboration and collaboration between brick-and-mortar companies due to the IT issues, human and cultural issues, and inter-organizational partnership issues as discussed in the article. Failure to consider the complexities of any of these issues will lead to a divorce of e-partnerships and collapse of virtual organizations.

REFERENCES

Aggarwal, A. K., & Zairi, M. (1998). Total partnership for primary health care provision: A proposed model—Part II. *International Journal of Health Care Quality Assurance, 11*(1), 7-13.

Bell, J. (2001). E-alliances: What's new about them? In A. P. de Man, G. M. Duysters, & V. Vasudevan (Eds.), *The allianced enterprise* (pp. 25-30). Singapore: Imperial College.

Cheng, W. L. E., Li, H., Love, E. D. P., & Irani, Z. (2001). An e-business model to support supply chain activities in construction. *Logistic Information Management, 14*(1/2), 68-78.

Couchman, P., & Fulop, L. (2000). Transdisciplinary research bodies: The changing nature of organizational networks and R & D in Australia. *Journal of World Business, 8*(2), 213-226.

Damanpour, F. (2001). E-business e-commerce evolution: Perspectives and strategy. *Managerial Finance, 27*(7), 16-32.

de Man, A. P., Stienstra, M., & Volberda, H. W. (2002). E-partnering: Moving bricks and mortar online. *European Management Journal, 20*(4), 329-339.

Dussauge, P., & Garrette, B. (1999). *Cooperative strategy: Competing successfully through strategic alliances*. New York: John Wiley & Sons.

Greif, J. (2000). Risky e-business. *Association Management, 52*(i11), 55.

Hagel, J., & Singer, M. (2000). Unbundling the corporation. In N. G. Carr (Ed.), *The digital enterprise: How to reshape your business for a connected world* (pp. 3-20). Boston: Harvard Business School.

Hellard, R. B. (1995). *Project partnering: Principle and practice*. London: Thomas Telford Publications.

Hennessey, A. (2000). Online bookselling. *Publishing Research Quarterly, 16*(i2), 34.

Jin, Z. (1999). Organizational innovation and virtual institutes. *Journal of Knowledge Management, 3*(1), 75-83.

Kotzab, H., Skjoldager, N., & Vinum, T. (2003). The development and empirical validation of an e-based supply chain strategy optimization model. *Industrial Management & Data Systems, 103*(5), 347-360.

Lei, D. T. (1997). Competence building, technology fusion, and competitive advantage: The key roles of organizational learning and strategic alliances. *International Journal of Technology Management, 14*(1), 208-237.

Malone, M., & Davidow, B. (1994). Welcome to the age of virtual corporations. *Computer Currents, 12*(1), 12-24.

McShane, L. S., & von Glinow, A. M. (2000). *Organizational behavior*. Sydney: Irwin McGraw Hill.

O'Toole, T. (2003). E-relationships—Emergence and the small firm. *Marketing Intelligence & Planning, 21*(2), 115-122.

Robbins, S. P., Bergman, R., Stagg, I., & Coulter, M. (2003). *Management* (3rd ed.). Sydney: Pearson Education Australia.

Rounthwaite, T., & Shell, I. (1995). Techniques: Designing quality partnerships. *The TQM Magazine, 7*(1), 54-58.

Segil, L. (2004). The eight golden rules of alliances. *Financial Executive Magazine & Business Week Online, 20*(9).

Sierra, M. C. D. L. (1994). *Managing global alliances: Key steps for successful collaboration.* Wokingham: Addison-Wesley Publishing.

Singh, M., & Byrne, J., (2005). Performance evaluation of e-business in Australia. *Electronic Journal of Information Systems Evaluation, 8*(1), 23-36

Trask, R. (2000). Developing e-partnerships. *Association Management, 52*(i11), 46.

van Hoek, R. (2001). E-supply chains—Virtually non-existing. *Supply Chain Management: An International Journal, 6*(1), 21-28.

Werbach, K. (2000). Syndication: The emerging model for business in the Internet era. In N. G. Carr (Ed.).*The digital enterprise: How to reshape your business for a connected world* (pp. 21-34). Boston: Harvard Business School.

Zhao, F. (2006). *Maximize business profits through e-partnerships*. Hershey, PA: Idea Group Publishing.

KEY TERMS

Brick-and-Mortar Organization: An organization located or serving customers in a physical facility as opposed to a virtual organization.

E-Business (Electronic Business): A comprehensive term used to describe the way an organization interacts with its key constituencies including employees, managers, customers, suppliers and partners through electronic technologies. It has a broader connotation than e-commerce because e-commerce is limited to business exchanges or transaction over the Internet only.

E-Partnership: A partnership relying on electronic (information) technologies to communicate and interact amongst partners. It is mostly associated with e-commerce or e-business partnerships.

E-Supply Chain: The physical dimension of e-business with the role of achieving base level of operational performance in the physical sphere (for more detail see van Hoek, 2001)

Intellectual Property (IP): A product of the intellect (intangible property) that has commercial value such as patents, trademarks, copyrights, etc.

Online Syndicate: Association of firms with a common interest formed to engage in e-business. Syndication has become an important e-business strategy of many companies.

Real Time: In the context of e-commerce, a real-time commission reporting system refers to a system in which a commission request is processed within milliseconds so that a commission report is available virtually immediately to an online salesperson.

Virtual Organization: A network form of organization that operates across space, time and organizational boundaries through an intricate Web of information technology.

Business Strategies for Outsourcing Information Technology Work

Subrata Chakrabarty
Texas A&M University, USA

INTRODUCTION

Firms pursue various strategies to exploit resources and capabilities and gain a competitive advantage (Porter, 1996). Interfirm relationships are collaborative agreements between organizations (Chakrabarty, 2006a; Whetten, 1981), and firms need to be careful in adopting suitable strategies to deal with interfirm relationships (Chakrabarty, 2007b). Interfirm relationships represent a sort of trade-off that organizations must make, whereby, in order to gain resources of other organizations, an organization must relinquish some its independence because the relationship also brings certain obligations with it (Whetten, 1981). Top management strategists might find their commitments to other firms as a sort of liability, and therefore, a serious evaluation of whether the benefits from the interfirm relationship outweigh the inevitable costs is needed before entering into interfirm relationships (Whetten, 1981).

Outsourcing is an Interfirm Relationship Between a Customer Firm and Supplier Firm

Work is outsourced to suppliers by a customer firm. A customer firm is therefore a firm that is in need of services, and a supplier firm is a firm that provides those services. The common synonyms for "customer" firm are either "client" firm or "buyer" firm. The common synonyms for "supplier" firm are either "vendor" firm, "consultant" firm, "third-party", or external service provider. This chapter will provide a useful summary of some strategies that customer firms can use for outsourcing information technology work to a supplier firm (Chakrabarty, 2006b, 2006c). For further information, readers are encouraged to refer to Chakrabarty (2006c) for real life case studies, and refer to Chakrabarty (2006b, 2007a, 2007b) for a deeper understanding of the advantages and disadvantages of various outsourcing strategies.

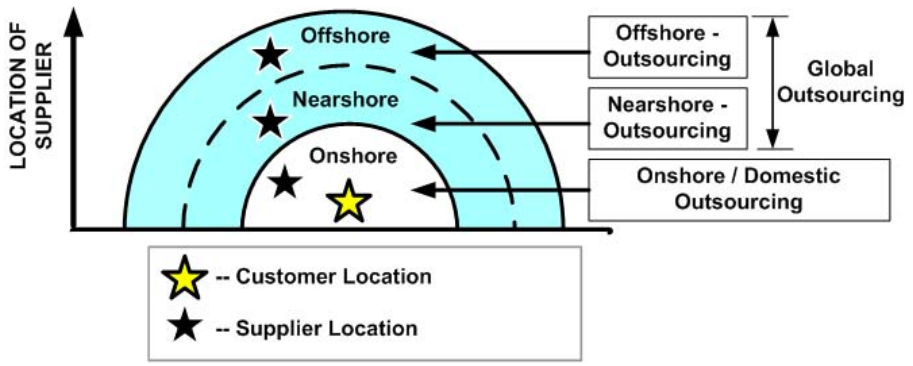
BACKGROUND

This section will provide some basic background information on outsourcing. Lacity and Hirschheim (1995) categorized the primary strategies of sourcing work into a continuum that ranges from total outsourcing at one extreme to total insourcing at the other extreme, and had selective sourcing as an intermediate strategy. *Total outsourcing strategy* is the strategy of a customer firm to outsource at least 80% of its information technology (IT) budget to suppliers. *Total insourcing strategy* (the opposite of outsourcing) is the strategy where a customer firm formally evaluates outsourcing but selects its own internal IT departments' bid over external supplier bids, and thereby allocates over 80% the IT budget to its internal IT department. *Selective outsourcing strategy* is the strategy whereby the customer firm opts to use suppliers for certain IT functions (representing around 20 to 60% of the overall IT budget, typically around 40%), and retains the remaining work for its internal IT department (Lacity & Hirschheim, 1995).

Further, Gallivan and Oh (1999), categorized the strategies for outsourcing on the basis of number of customers and suppliers into dyadic, multisupplier, cosourcing and complex outsourcing as follows. In a *dyadic outsourcing strategy*, there is just one customer and one supplier, that is, a customer firm uses only one supplier for a given activity, and the supplier in turn performs the given activity only for that customer firm. In a *multisupplier outsourcing strategy*, there is only one customer but many suppliers, that is, a customer firm uses many suppliers for a given activity. In a *cosourcing strategy*, there are many customers and only one supplier, that is, many customer firms jointly sign an outsourcing contract with a single supplier firm. In a *complex outsourcing strategy*, there are many customers and many suppliers; that is, it involves combining multiple customer firms and multiple supplier firms into a single contract (Gallivan & Oh, 1999).

Chakrabarty (2006b, 2006c) described how the location of the supplier to which work is outsourced can vary (see Figure 1). When a *domestic-outsourcing strategy* is adopted, both the customer and the supplier are located in the same country (this is also termed as *onshore-outsourcing*). In contrast, a customer and supplier can be located in different countries, and this known as a *global outsourcing strategy.* Though the term global outsourcing is widely referred to as offshore outsourcing, it can also be further classified into nearshore-outsourcing versus offshore-outsourcing. When a *nearshore-outsourcing strategy* is adopted, the chosen

Figure 1. Location of supplier in outsourcing



supplier located in a country that is geographically close to (but not the same as) the customer's country. When an *offshore-outsourcing strategy* is adopted, the chosen supplier is located in a country that is geographically far away from the customer's country. Time zones may also need to be factored during the formulation of strategy, because with improvements in communication technology and the need for 24x7 coordination of work, the time zones may be a bigger concern than geographical distance. We will now move on to more refined business strategies that can be used for outsourcing information technology work.

BUSINESS STRATEGIES FOR OUTSOURCING INFORMATION TECHNOLOGY WORK

Strategy of outsourcing selectively in a modular or flexible manner. A strategy often recommended to customer firms is that a selective set of information technology (IT) tasks need to be retained in-house based on the firm's own strengths and capabilities, and any remaining IT work that can be better performed by suppliers should be outsourced to the suppliers. *Selective outsourcing* is the strategy of outsourcing select IT tasks to suppliers, while retaining other IT tasks in-house (Lacity, Willcocks & Feeny, 1996). In selective sourcing, customer firms outsource between 20 to 60% of the IT budget to suppliers while still retaining a substantial amount of work for the internal IT department (Lacity & Hirschheim, 1995; see also Dibbern, Goles, Hirschheim & Jayatilaka, 2004, p. 10), and accordingly capitalizes on the strengths of both the internal IT department and the external suppliers. This is a flexible and modular form of outsourcing where work is broken down into multiple modules, of which, some are outsourced and some are retained in-house. This strategy of selective outsourcing has been given various other names such as *smart-sourcing, right-sourcing, flexible outsourcing*, and *modular outsourcing*.

Strategy of hiring multiple suppliers for an activity. Klotz and Chatterjee (1995) suggested that when a customer sources from two suppliers, it prevents the customer firm from being held by hostage by a monopolistic supplier, and it helps the customer firm to derive cost advantages due to competition among the suppliers. Currie and Willocks (1998) suggested the following three advantages of a *multiple-supplier outsourcing strategy*: (a) the customer firm is protected from being dependent on a single supplier, (b) the customer firm can use short-term contracts that may not be renewed with the same supplier (or combination of suppliers) and this encourages competition among the suppliers, and (c) the customer firm can focus on its core business while the suppliers manage and provide IT services. Such a strategy of multi-supplier outsourcing involves one-to-many relationships, indicating that one customer uses multiple suppliers with whom the division of labor is negotiated (Gallivan & Oh, 1999; see also Dibbern et al., 2004). Based upon the agreed division of labor, the various IT tasks are then jointly performed by the multiple suppliers, and this requires a cooperative environment among the suppliers, even though the suppliers are actually competing with each other for future business from the same customer (for case studies, see Chakrabarty, 2006c).

Strategy of contractually linking payments to realization of benefits - customer's performance determines supplier's revenue. A strategy where both the customer and supplier make upfront investments into a relationship and thereafter share both the risks and benefits is termed as a strategy of forming *benefit-based relationships* (Sparrow, 2003). Here, the customer firm makes its payments to the supplier depending on the specific benefits received. For example, if a customer can obtain potential business benefits by using the information technology services provides by a supplier, then the customer can establish a payment methodology

that links the payments to the supplier with the extent to which the customer benefits from the services. Hence, the supplier's earnings from the customer firm to which it is providing services is linked to the performance of the customer (Willcocks & Lacity, 1998).

This strategy is also termed as *business benefit contracting*, because it involves contracts that define the payments the customer will make whenever the customer earns excess revenues by using the supplier's services, and this arrangement essentially allows the sharing of both risks and rewards (Millar, 1994, as cited in Lacity & Hirschheim, 1995, pp. 4-5). The supplier provides services to the customer firm in manner that would ideally improve the customer firm's performance (Chakrabarty, Whitten & Green, 2007), and the customer evaluates the extent to which any improvement in its own performance is due to the supplier's contribution, and pays the supplier proportionately. Though such business benefit contracting has its advantages, it is often hard to adopt due to the challenges associated with negotiating and measuring the contractual criteria for sharing risks/costs and rewards/revenues (Lacity & Hirschheim, 1995).

Strategies of sharing risk and rewards using ownership and control structures. Novel ownership and control structures can be used to institutionalize the sharing of risk and rewards in two ways: (a) creating a new *joint venture* company where both the customer and supplier firms have ownership stakes, or (b) the customer firm can purchase share/equity for partial ownership of a supplier firm, and the supplier can similarly purchase share/equity for partial ownership of the customer firm (Currie & Willcocks, 1998; Sparrow, 2003; Willcocks & Lacity, 1998). These options are also known as *strategic alliances*.

The first strategy involves the customer and supplier firms creating and sharing ownership in a new *joint-venture* firm that has its own management team and IT employees, and the customer firm can outsource technology work to the joint venture company. Such joint venture companies enable the customer to gain access to new technical skills and resources, reorganize IT functions and processes and investigate new sources of revenue (Sparrow, 2003). Since the ownership of the new joint venture company is shared, the risks and rewards are also shared by the customer and supplier firms.

The second strategy involves *equity holding* deals, where the customer purchases enough shares of a supplier firm to partially own the supplier firm, and the supplier may also purchase enough shares of the customer firm to partially own the customer firm (Willcocks & Lacity, 1998). This automatically aligns the interests of both the customer and supplier firms, because each will benefit when the other performs well, and this arrangement motivates both the firms to share the risks and rewards.

Strategies for short-term requirements. A strategy for filling short-term labor demands is *body shop outsourcing*, whereby the customer goes shopping for "bodies" or human resources from suppliers. In other words, contract staff (such as programmers) are provided by a supplier, and these contract staff work at the customer firm's office and report directly to the customer firm's management executives (Lacity & Hirschheim, 1993). The contracted staff are therefore the supplier's paid employees who work under the supervision of the customer at the customer site.

Another strategy for getting temporary access to human resources for a short period of time is called *tactical outsourcing* (also known as *contracting-out* or *out-tasking*) (Sparrow, 2003). This strategy involves signing short-term outsourcing contracts with competent supplier firms who have the necessary technical skills to provide rapid solutions whenever the customer firm finds itself short of in-house employees to complete tasks in quick time.

Strategy of hiring a supplier for maintenance of technology assets. Most firms own a large amount of technology assets and infrastructure within their own facility (for example, hardware that needs maintenance or software that needs regular upgrades). A suitable strategy might be to hire a supplier who can offer the expertise and personnel to maintain the customer's technology assets and also lower the costs of maintaining these technology assets. That is, the customer owns the technology assets in the given facility, but hires a supplier to take over the operational control of these assets. This strategy is often termed as *facilities-management outsourcing* (Dibbern et al., 2004; Sparrow, 2003).

Strategy of outsourcing the process of setting up new offices/facilities abroad. Firms often need to expand their presence to new locations abroad, and this a challenge because the firm may not be knowledgeable about the processes of setting up an office/facility in the new location (Chakrabarty, 2007a). A suitable strategy to deal with this challenge is known as *managed offshore facilities' strategy*, whereby the customer firm outsources the process of creating its foreign subsidiary office to a supplier. Once the facility is up and running at the new location, the customer can take over the full ownership and control of the facility. At times, the customer firm retains the supplier for the long-term maintenance of the facility. A variant of the managed-offshore-facilities strategy is the *build-operate-transfer strategy*, whereby the supplier manages the process of creating the facility in the foreign location, and the customer firm has the option of taking full ownership by a specified date (i-Vantage, n.d.; Kobyashi-Hillary, M., 2004, p. 153). Hence, this outsourcing strategy has the potential to reduce many hassles for a firm that decides to set up its own subsidiary at a foreign location (for more details on the process of setting up a subsidiary abroad, see Chakrabarty, 2007a).

Strategies to strengthen the internal IT department. Though growth in the outsourcing of technology work is often assumed to be at the expense of the customer firm's internal IT department, a contrasting fact is that outsourcing

can also be used to strengthen the internal IT department (Green, Chakrabarty & Whitten, 2007). Firms sometimes undergo major transitions or technology overhauls in order to make use of newer technologies and bring in more efficiency. Suppliers can be used during this growth or maturation process of the customer's own IT department. For example, during a major changeover or transition, such as migration from a old technological platform to a modern technology platform, the customer firm can handover the management of the older systems to a supplier while the customer's IT department focuses on the transition to new technology. This is known as *transitional outsourcing* (Millar, 1994, as cited in Lacity & Hirschheim, 1995). A similar scenario whereby a certain work is outsourced to a supplier while the internal IT department transitions itself to a new set of skills is called a *transition-assistance* strategy (Wibbelsman & Maiero, 1994, as cited in Dibbern et al., 2004).

FUTURE TRENDS

Two distinct strategies that have gained prominence in recent times and are likely to be future trends are as follows. The use of a supplier that can provide teams at multiple locations, that is, a supplier team is at the customer site for coordination, while other skilled teams from the same supplier work at locations across the world at a lower cost. The renting of information technology services on a subscription basis.

Strategy of using suppliers with global capabilities. *Distributed consulting* implies that a supplier chosen by a customer has the ability to provide teams both at the customer's location and at the supplier's own location (Chakrabarty, 2006c). A *global delivery model* implies that the supplier can take advantage of the global talent pool and provide maximum value to the customer in terms of both quality and cost, by dividing the outsourced work into modules and distributing the modules to appropriate global locations (Infosys, n.d.).

Customer firms are increasingly adopting the strategy of offshore-outsourcing, that is, the chosen supplier is located in a country that is geographically far away from the customer's country (Chakrabarty, 2006c). This can be carried out more effectively by adopting a strategy of choosing a supplier that can provide supplier teams both at the customer's on-site location and at the supplier's offshore location. The supplier team at the customer site coordinates face-to-face with customer (Chakrabarty, 2006a), and the bulk of the outsourced work is carried out by the offshore supplier team (for case studies, see Chakrabarty, 2006c). Large IT service providers from India, such as TCS (http://www.tcs.com), Infosys (http://www.infosys.com), and Wipro (http://www.wipro.com), have incorporated such distributed consulting practices into their global delivery model (for case studies, see Chakrabarty, 2006c).

Strategy of accessing remotely hosted IT applications. One strategy that a customer firm can adopt for outsourcing information technology work is to rent the required service on a subscription basis (Chakrabarty, 2006b). Similar to the manner in which employees of a customer firm can access software applications installed on a LAN or data center within the customer firm, the customer firm's employees can also access a software application that is installed on a remote server under the control of the supplier. That is, the remote server resides at the supplier's data center and is accessed by the customer firm through a dedicated line, Internet, or extranet (Dewire, 2000). Hence, the suppliers develop, customize, install, and manage the software applications at the remote locations and host them for their customers over a suitable network or the Internet. Such suppliers are called *application service providers (ASP)* (Bennett & Timbrell, 2000; Susarla, Barua & Whinston, 2003), and this type of outsourcing strategy has been given various names such as *net-sourcing* (Kern, Lacity & Willcocks, 2002), *on-demand service, application utilities, real-time delivery* and *software-as-a-service* (Pring & Ambrose, 2004), all of which allow access to externally managed software applications.

CONCLUSION

Outsourcing is an interfirm relationship between a customer firm and supplier firm, where the customer firm is in need of services and the supplier firm provides those services. Since such interfirm relationships are essential for most businesses, and this chapter suggested that firms need to be careful in adopting suitable strategies. An array of strategies that can be used for both domestic and global outsourcing of information technology work were described, so that business managers can choose an appropriate strategy in order to get the best deal for their information technology needs.

REFERENCES

Bennett, C. & Timbrell, G. (2000). Application service providers: Will they succeed?. *Information Systems Frontiers, 2*(2), 195-211.

Chakrabarty, S. (2006a). A conceptual model for bidirectional service, information and product quality in an IS outsourcing collaboration Environment. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)*. Retrieved June 17, 2008, from http://doi.ieeecomputersociety.org/10.1109/HICSS.2006.7

Chakrabarty, S. (2006b). Making sense of the sourcing and shoring maze—The various outsourcing & offshoring alternatives. In H. S. Kehal & V. P. Singh (Eds.), *Outsourcing &*

offshoring in the 21st century – A socioeconomic perspective (1st ed., pp. 18-53). Hershey, PA: IGI Publishing.

Chakrabarty, S. (2006c). Real Life Case Studies of Offshore Outsourced IS Projects: Analysis of Issues and Socio-Economic Paradigms. In H. S. Kehal & V. P. Singh (Eds.), *Outsourcing & Offshoring in the 21st Century – A socio economic perspective* (1 ed., pp. 248-301). Hershey, PA: IGI Publishing.

Chakrabarty, S. (2007a). The journey to new lands: Utilizing the global IT workforce through offshore-insourcing. In P. Young & S. Huff (Eds.), *Managing IT professionals in the internet age* (1st ed., pp. 277-318). Hershey, PA: IGI Publishing.

Chakrabarty, S. (2007b). Strategies for business process outsourcing: An analysis of alternatives, opportunities and Risks. In J. Sounderpandian & T. Sinha (Eds.), *E-business process management: Technologies and solutions* (1st ed., pp. 204-229). Hershey, PA: IGI Publishing.

Chakrabarty, S., Whitten, D., & Green, K. W. (2007). Understanding service quality and relationship quality in IS outsourcing: Client orientation & promotion, project management effectiveness, and the task-technology-structure Fit. *Journal of Computer Information Systems, 48*(2), 1-15.

Currie, W. L. & Willcocks, L. P. (1998). Analyzing four types of IT sourcing decisions in the context of scale, customer/supplier interdependency and risk mitigation. *Information Systems Journal, 8*(2), 119-143.

Dewire, D. T. (2000). Application service providers. *Information Systems Management, 17*(4), 14-19.

Dibbern, J., Goles, T., Hirschheim, R., & Jayatilaka, B. (2004). Information systems outsourcing: A survey and analysis of the literature. *ACM SIGMIS Database, 35*(4), 6-102.

Gallivan, M. J. & Oh, W. (1999). Analyzing IT outsourcing relationships as alliances among multiple clients and vendors. In *Proceedings of the 32nd Annual International Conference on System Sciences*, Hawaii.

Green, K. W., Chakrabarty, S., & Whitten, D. (2007). Organisational culture of customer care: Market orientation and service quality. *International Journal of Services and Standards, 3*(2), 137-153.

Infosys (n.d.). *Global delivery model*. Retrieved June 17, 2008, from http://www.infosys.com/gdm/default.asp

i-Vantage. (n.d.). *Global insourcing services*. Retrieved June 17, 2008, from http://www.i-vantage.com/GlobalInsourcingServices.html

Kern, T., Lacity, M. C., & Willcocks, L. P. (2002). *Netsourcing: Renting business applications and services over a network*. New York: Prentice Hall.

Klotz, D. E. & Chatterjee, K. (1995). Dual sourcing in repeated procurement competitions. *Management Science, 41*(8), 1317-1327.

Kobyashi-Hillary, M. (2004). *Outsourcing to India: The offshore advantage*. Berlin, Germany: Springer-Verlag.

Lacity, M. C. & Hirschheim, R. A. (1993). Implementing information systems outsourcing: Key issues and experiences of an early adopter. *Journal of General Management, 19*(1), 17-31.

Lacity, M. C. & Hirschheim, R. A. (1995). *Beyond the information systems outsourcing bandwagon: The insourcing response*. Chichester: Wiley.

Lacity, M. C., Willcocks, L. P., & Feeny, D. F. (1996). The value of selective IT sourcing. *Sloan Management Review, 37*(3), 13-25.

Porter, M. E. (1996). What is strategy? *Harvard Business Review*, 61-78.

Pring, B. & Ambrose, C. (2004). Vendors vie for competitive position in ASP market. *Gartner research*

Sparrow, E. (2003). *Successful IT outsourcing*. London: Springer-Verlag.

Susarla, A., Barua, A., & Whinston, A. B. (2003). Understanding the service component of application service provision: An empirical analysis of satisfaction with ASP services. *MIS Quarterly, 27*(1), 91-123.

Whetten, D. A. (1981). Interorganizational relations: A review of the field. *Journal of Higher Education, 52*, 1-28.

Willcocks, L. & Lacity, M. (1998). *Strategic sourcing of information systems*. Chichester: Wiley.

KEY TERMS

Application Service Providing / Net-sourcing / On-Demand: Accessing remotely hosted information technology (IT) applications

Benefit Based Relationships / Business Benefit Contracting: Linking payments to realization of benefits; customer's performance determines supplier's revenue.

Body Shop Outsourcing: Using contract personnel.

Cosourcing: Many customers and only one supplier: Many customer firms jointly sign an outsourcing contract with a single supplier firm.

Distributed Consulting: Supplier has teams both at onshore and offshore.

Global Delivery: Large supplier delivering services from various global locations to customers at various global locations.

Managed Offshore Facilities: Outsourcing the process of setting up a subsidiary abroad.

Multisupplier Outsourcing / Dual Sourcing: A customer firm uses many suppliers for a given activity.

Nearshore-Outsourcing: Chosen supplier is located in a country that is geographically close to (but not the same as) the customer's country.

Offshore-Outsourcing (A Form of Global Outsourcing): Chosen supplier is located in a country that is geographically far away from the customer's country.

Onshore-Outsourcing / Domestic Outsourcing: Both customer and the supplier are located in the same country.

Outsourcing: Interfirm relationship between a customer firm and supplier firm, where the customer firm is in need of services and the supplier firm provides those services.

Selective / Smart / Right / Flexible / Modular Sourcing: Outsourcing and insourcing optimally and selectively; A customer firm uses suppliers for certain IT functions which represents between 20 and 60% of the IT budget (typically around 40%) and therefore retains substantial work for its internal IT department.

Tactical Outsourcing / Contracting-Out / Out-Tasking: Outsourcing for short term access to skilled professionals.

Transitional Outsourcing: Outsourcing during a major changeover; Helping the customer's IT department mature.

Business-to-Consumer Electronic Commerce in Developing Countries

Janet Toland
Victoria University of Wellington, New Zealand

Robert Klepper
Victoria University of Wellington, New Zealand

INTRODUCTION

Electronic commerce describes the process of buying, selling, transferring, or exchanging products, services, or information via computer networks including the Internet. In **business-to-consumer electronic commerce**, the sellers are organisations, and the buyers are individuals (Turban, Leidner, McLean, & Wetherbe, 2005). Business-to-consumer electronic commerce provides opportunities for less-developed countries to reduce transaction costs and bypass some of the intermediary linkages to connect to global supply chains (Molla & Licker, 2005). Though predictions vary, statistics seem to point to significant growth of the use of the Internet among businesses and consumers in **developing countries** in the next 10 years (Hawk, 2004). The focus here is to explore the potential for business-to-consumer electronic commerce in less-developed countries. The approach taken is to review the current worldwide usage of the Internet; to identify the factors necessary for **e-readiness**; to explore the barriers to business-to-consumer electronic commerce; and to identify strategies that can be adopted by both the public and private sectors to overcome these barriers.

By the end of 2003, developing countries accounted for more than one third of new Internet users worldwide. Though Internet access is rapidly increasing, most residents of developing countries still have no access to the Internet. For example, Internet access in Africa is less than 2% in a population of over 900 million, the lowest rate of access in the world (Dunphy, 2000; UNCTAD, 2004). Business-to-consumer electronic commerce in less-developed countries will grow in the future, but progress will be slowed by technological, cultural, economic, political, and legal problems (Davis, 1999; Enns & Huff, 1999). Differences in e-readiness and related **barriers to electronic commerce** will sustain substantial differences between regions of the world, between countries within regions, between urban and rural areas within countries, and between the genders and age groups. Despite the difficulties, when the basic communications infrastructure is available, options do exist to undertake business-to-consumer electronic commerce in less-developed countries.

BACKGROUND

Table 1 shows the number of Internet users in the major regions of the world reflecting vast differences in **e-readiness**. Less than 10% of the population in the developing regions of Africa, Latin America and the Caribbean, and Asia were using the Internet in 2004 as compared to regions such as North America, Europe, and Australasia where 30% or more used the Internet.

Africa

The **digital divide** is largest in Africa, with less than 2% of people having access to the Internet as compared to 50% in most advanced countries. There are some **business-to-consumer electronic commerce** success stories, mostly in the traditional handicrafts area, where the Internet offers the

Table 1. Internet users per 10,000 people, by region, 2005 (Adapted from http://www.internetworldstats.com/stats.htm)

REGION	USERS PER 10,000	% IN REGION (approx)
Africa	144	1.4%
Latin America & Caribbean	1,011	10.1%
North America	6,501	66.5%
Asia	738	7.4%
Europe	3,159	31.6%
Australia & New Zealand	4,735	47.4%
World	**1,385**	**12.7%**

opportunity for a niche player to access the global market of African Diaspora. All African capital cities now have local Internet connection available.

Latin America

Internet use in the region is dominated by Argentina, Brazil, Chile, and Mexico, who among them account for two thirds of Internet users in the region. **Business-to-consumer electronic commerce** is growing, with online car sales, consumer auctions, travel, computer hardware and software, and banking responsible for the highest revenue.

Asia

Among **developing countries**, Asia stands out as the leading user of **electronic commerce**. This is partly due to high population, but also because organizations tend to be more integrated into global trade flows than in other developing countries. Manufacturing enterprises in particular face pressure from their customers in developed countries to adopt electronic commerce. China offers the greatest potential electronic commerce market, and is now considered one of the top five nations in the world in terms of Internet use. While many Chinese are going online for the first time, less than 20% have done any shopping online (Hsu, 2003).

North America/Europe/Australasia

In the developed world, the **business-to-business electronic commerce** continues to grow faster than business-to-consumer, with Forrester Research (Johnson, Delhagen, & Yuen, 2003) forecasting that 26% of business-to-business sales in the United States will be traded online by 2006. **Business-to-consumer electronic commerce** has progressed significantly in some sectors particularly those offering digital products, such as software, music, and travel services.

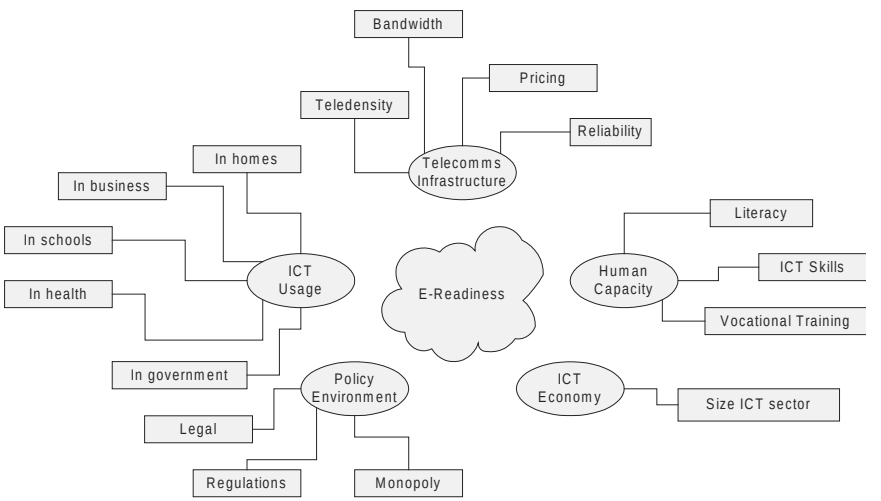
E-READINESS

An **e-readiness** assessment is an attempt to gauge how prepared a country is to benefit from information technology and electronic commerce. It is used to measure a country's ability to take advantage of the Internet as an engine of economic growth and human development (GIPI, 2001). An e-readiness assessment looks at infrastructure, the accessibility of information and communication technology (ICT) to the population at large, and the effect of the legal and regulatory framework on ICT use.

Over 15 different e-readiness assessment tools are currently available, and the assessments use a range of questionnaires, statistical methods, reports of best practice, and historical analysis (Bridges.org, 2002). Some tools look specifically at the e-economy and how ICT's can be used to improve the economy, whereas others are concerned with the broader picture, trying to measure the emergence of an e-society and assess how ICT's are improving social equality.

The assessment methods tend to include a common core of questions covering the areas of telecommunications infrastructure, levels of ICT usage throughout society, human

Figure 1. Factors commonly used to assess e-readiness



capacity in terms of educational level, the legal and regulatory environment, and the size of the ICT sector (InfoDev, 2001). Typical factors for consideration are shown in Figure 1.

The purpose of carrying out these assessments is to gather information that can assist with developing a strategy for ICT development.

BARRIERS TO ELECTRONIC COMMERCE IN LESS-DEVELOPED COUNTRIES

Business-to-consumer electronic commerce operates at three levels:

1. Information gathering and exchange (for example, a tourist may want to browse information about hotel accommodation);
2. Online transactions (for example, booking a hotel room);
3. Online payment, or paying for a hotel room online.

Most Internet use in **developing countries** is for e-mail; there is relatively little transactive use. Lack of e-readiness, or what might be called **barriers to electronic commerce**, such as low credit card use and poor logistics and fulfilment, inhibit fully developed electronic commerce (UNCTAD, 2002). Table 2 lists the main inhibitors of **electronic commerce** in developing countries. These have been grouped according to the five categories identified as important when considering e-readiness. An extra category has been introduced for general infrastructural issues, such as a poor transport network, that negatively impact the adoption of **business-to-consumer electronic commerce**.

As a result of the barriers, most electronic commerce in developing countries takes place between organizations. **Business-to-business electronic commerce** accounts for 95% of all electronic commerce in less-developed countries. Much of this is driven by large multi-national organizations operating in developing countries. Indigenous **electronic commerce** is usually small by comparison. Even within the realm of indigenous business-to-consumer electronic commerce, sales are often to offshore customers in the developed world, as is the case for the tourism, art, and handicrafts markets.

OPTIONS IN BUSINESS-TO-CONSUMER ELECTRONIC COMMERCE

Despite the barriers, **business-to-consumer electronic commerce** will grow in less-developed countries, albeit more slowly that in developed countries. Our focus is on options

Table 2. Barriers to electronic commerce in developing countries

TELECOMMUNICATIONS INFRASTRUCTURE	- Poor telecommunications service - Limited penetration of Internet - High cost of Internet connections
POLICY ENVIRONMENT	- Lack of competition in international telephone traffic - Lack of suitable regulatory environment
HUMAN CAPACITY	- Illiteracy - Low incomes - Poor customer service - Shortage of technical skills
ICT USAGE	- No tradition of mail order type shopping - Many small and medium size enterprises that lack capital for the development of electronic commerce - Lack of critical mass
ICT ECONOMY	Small ICT sector
GENERAL INFRASTRUCTURE	- Unstable power supplies - Inadequate payment systems and low credit card usage - Fulfilment problems occasioned by poor physical infrastructure (road, rail, air)

available to individuals and private sector enterprises for undertaking business-to-consumer electronic commerce.

Telecommunications Infrastructure

New technologies have the potential to bypass inadequate landline telecommunications infrastructure. **Developing countries** have the opportunity to leapfrog traditional copper and fibre-based landlines and move directly to leading-edge wireless technologies. Wireless technologies have taken off even in relatively low-income areas of the world, where prepaid cards allow access without having to pass a creditworthiness check. As of 2004, major wireless network projects have been completed in Shanghai and Fujian in China, and in Peru, Indonesia, and Ethiopia (Lancaster, 2003; Simon, 2004).

Policy Environment

The actions of government in government policy, the legal environment, the management of government owned and/or controlled telecommunications infrastructure, while important to electronic commerce success, are not our concerns here.

ICT Usage and ICT Economy

ICT usage is the aggregate of usage by individuals, businesses, and non-profit organisations. More businesses and more customers will be online if they know they can do transactions efficiently by electronic commerce. In addition, more businesses in less-developed countries will be online if they know they can reach customers across the globe efficiently. With sufficient demand from businesses and customers, the level of infrastructure represented by Internet service providers and facilities for Web site development and hosting and processing if electronic payments will also become available. These developments are interrelated and self-reinforcing.

In many less-developed countries, the evolution of ICT use for commerce will take place from the bottom up. Within the constraints mentioned earlier, the size of the ICT economy will grow with the individuals and businesses doing transactions on the Internet. In what follows, we briefly survey options for online store-type businesses:

- Hardware has become a commodity, and free, open source software tools are available for online store development and Web hosting (Klepper & Carrington, 2002). Even in countries with high tariffs on hardware, these costs are not the barrier they once were.
- Development of an online store is becoming easier for less-skilled developers as standard templates are now available on the Internet. Of course, customised stores require skilled labour, which is an issue we address next. Online store templates are widely available on the Internet and are increasingly available from Internet service providers in less-developed countries (Klepper & Carrington, 2002). Even greater availability can be expected in the future.
- Web site hosting is usually beyond the capabilities of a typical individual online store business, but hosting services are increasingly available from Internet service providers in less-developed countries. Online store businesses, particularly those aiming at a wider world market, can readily obtain hosting in developed countries; a 2003 survey of business Web sites in the nine least developed countries found that about half were hosted locally, and half offshore (Wresch, 2003).
- Most payments for online purchases are still accomplished by credit card in the developed world, an option that exists for a much smaller percentage of customers in **developing countries**. However, in cities and the portions of less-developed countries with sufficient population density, a variety of payment options are offered (Hawk, 2004). Cash payment can be made at the time of delivery, which occurs in a significant portion of such purchases in China and India, or an option of paying by a series of monthly cheques can be offered as in Latin America.
- Limitations on human capacity can be overcome. Specialized companies and consultants who build Web sites, provide Internet services, and host Web sites are already available in many developing countries (Wresch, 2003), and their numbers will grow as the demand for online transactions grow. The employment opportunities available in developed countries during the IT and dot com boom of the late 1990s have abated, and the skilled citizens of less-developed countries now have less incentive to migrate and those in developed countries are more likely to return to their native countries. Some research shows that small online businesses in **developing countries** often rely on skilled family members and friends for help on technical issues, and this is likely to continue (Utomo, 2001).

On the demand or customer side, access is growing through shared facilities like Internet cafes and through growing individual and family computer ownership.

Although the options outlined previously are still only available to higher income individuals living in cities in the developing world, the number of customers and the number of online businesses is expanding rapidly.

FUTURE TRENDS

Growth in the number of mobile telephone users worldwide has expanded from 50 million in 1998 to more than one and a half billion as of 2004 (Evans, 2004). At the end of 2003, Africa had more than 50 million mobile device users, and the majority of African nations had more mobile than fixed subscribers (Economist, 2005; UNCTAD, 2002). Similar trends have been observed in Latin America and Asia, where handheld devices enable users to overcome the difficulties caused by low fixed line penetrations. This explosive growth in **mobile telephony** in the developing world, particularly in Africa, offers an opportunity to overcome the problem of an inadequate telecommunications infrastructure. Mobile phones can provide an "Internet in your pocket".

Access to mobile phones is even more widespread than ownership numbers suggest. A telephone may be owned by one person in a village who runs a small business taking and receiving calls for neighbours. The World Bank estimates that 77% of the world's population now has the opportunity to link up to a mobile network (Atkins, 2005). The availability of prepaid cards means that no credit checks are necessary, removing a major barrier to participation. Mobile phones can be used to make payments for items as diverse as petrol, laundry, and soft drinks. This is a significant advantage because many less-developed countries' lack of a reliable credit card system is a real inhibitor of business-to-consumer electronic commerce.

Entrepreneurs in developing countries can learn by observing trends in the developed world, where it has been shown that companies using business-to-consumer electronic commerce to sell digital products, such as online travel, software, and financial services, are generating significant profits, while those selling tangible goods are not doing so well. In particular, tourism services such as hotels and travel agencies seem to transfer most successfully to electronic sales, and this has been observed across a number of countries (Wresch, 2003). The tangible products that sell least well electronically, are those that consumers want to "see and feel" such as clothing and furniture. Models for new business-to-consumer ventures should be developed with this information in mind.

CONCLUSION

The worldwide use of business-to-consumer electronic commerce is likely to remain uneven for some time. E-mail will continue to be the first priority of new users. Searching for information will also be important in the future. Growth in entertainment and online purchases will lag. **Business-to-consumer electronic commerce** growth will be largely an urban phenomenon, and rural areas will participate at much lower rates. The gaps between young and old and males and females, which have narrowed in the developed world, will persist much longer for people in less-developed countries.

The barriers to the development of business-to-consumer electronic commerce all have a common source in the conditions that create and sustain differences between the developed and developing world. These issues are beyond the scope of this brief comment on electronic commerce, but until these conditions change, less-developed countries will continually lag the developed world in electronic commerce, as they do in many, more important indicators of well-being such as health and education.

REFERENCES

Atkins, T. (2005, February 24). Digital divide narrowing fast, World Bank says. *Reuters*.

Bridges.org. (2002). *E-readiness assessment: Who is doing what and where*. Retrieved June 10, 2002, from http://www.bridges.org/ereadiness/where.html

Davis, C. H. (1999). The rapid emergence of electronic commerce in a developing region: The case of Spanish-speaking Latin America. *Journal of Global information Technology Management, 2*(3), 25-40.

Dunphy, H. (2000). *Report: African Internet use too low.* Retrieved October 31, 2000, from http://news.excite.com/news/ap/001030/19/africa-online

Economist. (2005, March 12-18). The real digital divide. (p. 11).

Enns, H. G., & Huff, S. L. (1999). Information technology implementation in developing countries: Advent of the Internet in Mongolia. *Journal of Global information Technology Management, 2*(3), 5-24.

Evans, R. (2004). Mobile phone users double since 2000. *Computerworld*. Retrieved April 18, 2005, from http://www.computerworld.com/printthis/2004/0,4814,98142,00.html

GIPI. (2001). *E-readiness guides*. Retrieved October 6, 2002, from http://www.gipiproject.org/readiness

Hawk, S. (2004). A comparison of B2C e-commerce in developing countries. *Electronic Commerce Research, 4*(3), 181-199.

Hsu, J. (2003). Reaching Chinese markets through e-commerce: A cultural perspective. In *Proceedings of the Fourth Annual Global Information Technology Management World Conference*, Calgary, Alberta, Canada (pp. 241-244).

InfoDev (2001). *E-readiness as a tool for ICT development. Annual Report.* Retrieved October 6, 2002, from http://www.infodev.org/library/WorkingPapers/Areaready.pdf

Johnson, C.A., Delhagen, K., & Yuen, E.H. (2003). US ecommerce overview: 2003 to 2008. *Forrester Research.* Retrieved August 22, 2005, from http://www.forrester.com

Klepper, R., & Carrington, A. (2002). Options for business-to-consumer electronic commerce in developing countries: An online store prototype. In S. Burgess (Ed.), *Managing technology in small businesses: Challenges and solutions*. Hershey, PA: Idea Group Publishing.

Lancaster, J. (2003). Village kiosks bridge India's digital divide. *Washington Post*, October 12, p. A1.

Molla, A., & Licker, P. S. (2005). Ecommerce adoption in developing countries: A model and instrument. *Information & Management, 42*, 877-899

Simon, S. J. (2004). Critical success factors for electronic services: Challenges for developing countries. *Journal of Global Information Technology Management, 7*(2), 31-53.

Turban, E., Leidner, D., McLean, E., & Wetherbe, J. (2005). *Information technology for management: Transforming business in the digital economy* (5th ed.). New York: John Wiley.

UNCTAD (2002). E-*commerce and development report 2002*. United Nations. Retrieved September 20, 2002, from http://www.unctad.org/ecommerce

UNCTAD (2004). *E-commerce and development report 2002*. United Nations. Retrieved September 20, 2002, from http://www.unctad.org/ecommerce

Utomo, H. (2001, June 10-12). The influence of change agents on IT diffusion: The case of SMEs in Indonesia. In P. Palvia & L. Chen (Eds.), *Proceedings of the Second Annual Global Information Technology Management World Conference*, Dallas, TX (pp. 176-179).

Wresch, W. (2003). Initial e-commerce efforts in nine least developed countries: A review of national infrastructure, business approaches and product selection. *Journal of Global Information Management, 11*(2), 67-78.

KEY TERMS

Business-to-Business Electronic Commerce (B2B): A business selling goods and/or services online to another business.

Business-to-Customer Electronic Commerce (B2C): A business selling goods and/or services online to private consumers.

Diaspora: Nationals of a country who have migrated and are scattered worldwide.

Digital Products: Goods or services that can be delivered directly over the Internet. Examples would be software, online travel, and financial services.

Less-Developed Countries (LDCs): Less-developed countries, commonly having a low standard of living, poor health, and inadequate education. Markets are less fully developed, and there is often a greater dependence on the primary sector of the economy.

Tangible Goods: Goods and services that have a physical form and cannot be delivered directly over the Internet.

Text Messaging (SMS): Short message service. The sending and receiving of short typed messages using mobile telephones.

Wireless Technologies: Technologies that communicate without landlines, that is, satellite, microwave, cellular radio, infrared. Common uses are pagers, cellular telephones, personal digital assistants, mobile data communications, and personal communications services.

CAD Software and Interoperability

Christophe Cruz
Université de Bourgogne, France

Christophe Nicolle
Université de Bourgogne, France

INTRODUCTION

Decisions taken during the conception phases in huge architectural projects influence a lot the cost and the schedule of the building construction. To ease this decision-making, many mock-ups have been used as a project prototype. This prototyping is useful to test and to improve the conception of projects. Nowadays, collaborative sites that appear on the Web greatly improve the flexibility of the framework's actors of a distant project [Aliakseyeu, Martens, Subramanian, Vrouble, & Wesselink, 2001; Balaguer & DeGennaro, 1996; Klinker, Dutoit, Bauer, Bayes, Novak, & Matzke, 2002). Digital mock-ups are used to represent future 3D elements of the final product. Digital mock-ups are known to be often employed in the architectural field. Indeed, the visualization of the future buildings in 3D by architects and engineers is a way to facilitate the testing of the choices, the scheduling of costs and processes, and the completion dates. In the architectural field, all types of activities have developed tools for special prototyping: structural analysis, thermal and fluidic networks, and so forth. Unfortunately, this development is completely chaotic. Sometimes existing tools in the same type of activity cannot exchange information. Moreover, information stored by tools is in most cases bound by a set of files that contain only geometrical descriptions of the building. Not every actor of a project has necessarily the same knowledge as the other actors to understand and to interpret information. Thus, the collaboration between the actors as well as the data interoperability seems to be difficult to evolve without a new kind of tool. The following section presents two examples of platforms using digital mock-ups to handle conception data. The section "Collaborative Web Platform" focuses on our solution through the presentation of the Active3D collaborative platform. The section "Interoperability Demonstration" presents the Active3D platform as a central point of collaboration with the help of use-cases examples. The last section concludes on the work being undertaken.

BACKGROUND

The collaborative work between distant actors on the same project improves the conception of a prototype by reducing the time between each update. A lot of CAD software packages were modified to allow virtual prototyping, but this was done independently of specific project requirements. Unfortunately, most of these solutions do not join together the essential capabilities of interaction and collaboration for the completion of an engineer project. To avoid this problem many projects were suggested. The project Cavalcade (Cavalcade, n.d.) is based on a distributed architecture, allowing several distant teams to collaborate on a conception, to test, to validate, and to exchange documents. Cavalcade provides a visual system of 3D visualization. Contrary to classical ideas on simulation tools, the virtual representation of a prototype concerns only the visual aspect of attributes of which the objects of the building are composed. These attributes are functions like "is a part of a subsystem" and documents like technical files or Web links. The 3D model becomes then a visual interface of information requests. Cavalcade aims to manage conception data. To exchange the models created with the help of CAD software, the developers of this software use specific format files for their requirements. The set of files that forms the conception of the project constitutes the digital mock-up. The 3D model of the conception object is generally integrated in this mock-up and a set of information allows management of the project by itself.

In addition, the organization of the engineering and design department must be reconsidered. To facilitate the pooling of data, a digital mock-up should be installed. The conception work is then immediately possible from the mock-up. Access to the last updated data avoids expensive errors related to the use of data not up to date. The sharing of conception data is obviously a requirement in order to accelerate the conception cycle. Several problems must be taken into account in the conception of a 3D collaborative platform.

The first problem concerns the choice of an information storage structure. There are two kinds of information storing: files and databases. In the field of 3D, the file formats are very numerous. Although the principal information of

these files is the geometrical representation in 3D of the objects, each kind of file has its own levels of abstraction. The higher the abstraction level of information is, the more semantics contains the file. This semantics is an additional knowledge on geometrical information, making it possible to re-use in a better way the geometrical file and its definitions. Databases ensure the storage of large quantities of information by structuring and indexing information. In general, the databases carry the subjacent semantics of information which they store. Indeed, the structures which receive information model information that they must contain, therefore these structures form metadata on information. The databases are, thus, of primary importance to organize information so that it becomes possible to search for relevant data in a vast set of information such as a file.

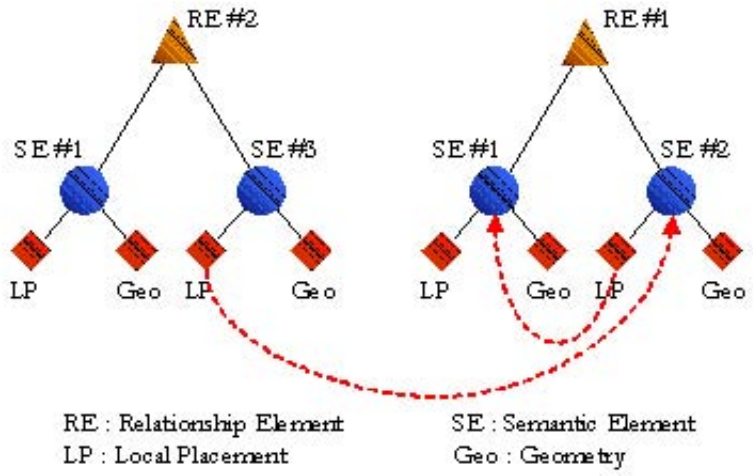
The second problem concerns the definition of an optimized 3D interface that allows a flexible and fast handling of stored information. Certain applications require at the same time a lot of memory and a minimum speed of execution. For instance, the computer-aided design often produces complex 3D geometrical models that have a very large size. Thus, the recurring problem in graphical application is the data visualization. Indeed, with the advent of design techniques, the growth of the volume to be computed is largely higher than the increase of the capacity of the graphic material. Consequently, a phase of optimization is necessary. It is located at the data model and at the data themselves. Many techniques of optimization and acceleration for interactive navigation were developed. These include calculations of visibility (Pearce, Partial, & Day, 2004), geometrical simplification (Hoppe, 1996; Pailot, Merienne, Frachet, & Nevfeu (2003), and the image-based representation (Christopoulos, Gaitatzes, & Papaioannou, 2003; Gortler, He, & Cohen, 1997; Levoy, & Hanrahan, 1996; Mark, McMillan, & Bishop, 1997). All these techniques were combined successfully to render architectural models (Funkhouser, Teller, Sequin, & Khorramabadi, 1996) and urban models (Wonka, Wimmer, & Sillion, 2001). The GigaWalk project is a rendering system that makes it possible to render CAD projects of more than 10 millions of polygons. The most striking example is the design project "DoubleEagle tanker" made up of more than four gigabytes of data, that is to say 82 million triangles and 127 thousand objects. The rate of calculation is 11 to 50 images per second which permits to navigate in real time through the digital mock-up after a time of approximately 40 hours pre-calculation (Baxter, Sud, Govindaraju, & Manocha, 2002). However, there is no perfect system. Each technique has advantages and disadvantages but certain combinations with precise conditions are very effective. These techniques described previously proved their reliability within a static framework, that is, the optimized scene is calculated once for several visualizations. In fact, the pre-computing time before the accessibility of the scene can sometimes take several days. The pre-computing of the optimized scenes is the major problem of all these techniques because if the 3D models evolve during this time then the management of the data synchronization must be taken into account. These synchronizations are not always possible because the complete structure of the scene can sometimes change. Other ways must be explored to allow for the visualization of a 3D scene to evolve during this time. Optimizations are always carried out in comparison to the geometry or the topology of the scene. On the one hand, the nature of the geometrical objects was not taken into account for computing optimization. The nature of the objects, thus, proves to be an undeniable way for research. This nature depends on the scene structuring but if it is limited to geometrical information then only geometrical optimizations are applicable. On the other hand, if information on the nature of the objects is indicated then this information provides a new way of research on the handling of geometrical information and their storage.

COLLABORATIVE WEB PLATFORM

Nowadays, the fundamental needs of all the actors in architectural engineering projects relate to a simple tool which allows a coordinated management of the actions carried out in a project. This tool must allow management of data generated during the lifecycle of the project through a 3D visualization of a digital mock-up and must also allow its access to all project actors through a collaborative Web platform. This section presents the ACTIVe3D-Build platform which makes it possible for the actors of a project, geographically dispersed - from the architect to the plumber - to exchange documents directly in a virtual environment during the lifecycle of a civil engineering project. A 3D visualization makes it possible for the actors to move around the building that is being designed and to obtain information on the objects. This section is divided into three parts. The first part presents the format used to describe the data. The second part deals with the data structuring. The third part presents the division method and exchange of data.

Data format: CAD software used in civil engineering projects models each building element by a set of vectors. In this formalism there is no semantic information on the objects that compose the building. Thus, there is no way to select automatically objects by their nature. To solve these problems the International Alliance for Interoperability (IAI) proposed a standard called IFC (Industrial Foundation Classes – http://www.iai-international.org/) which describes the representation of the objects that can be found in an architectural project. The IFC file format is a model which associates trade semantics with 2D/3D geometry for each element constituting the building. The addition of trade semantics makes it possible to limit the redundancies of information because it identifies instantaneously each element that the building is composed of for a faster

Figure 1. Example of direct and indirect links between several semantic elements



qualification of the building elements. The basic classes of the IFC include the description of the objects and provide a structure permitting the data interoperability between trade applications. For example, an IFC door is not simply a collection of lines and geometrical primitives identified as a door, but it is recognized such as "door" by the machine and has attributes corresponding to its nature. The adoption of this format by all the leaders of software CAD solves the problem of the interoperability of information between the various civil engineering professions.

Data structuring: The choice of the IFC format for the data structuring has many advantages but comprises also certain disadvantages. The study of the IFC shows the complexity of the links between the instances of relational classes and the instances of object classes. On this level there are two types of links between the objects. We call them direct and indirect links. The indirect links are defined by instances of relations, "Relationship Element" (RE: triangles - Figure 1). The direct links are defined by discontinuous red links between the instances of class resources (rhombuses - Figure 1) and an instance of "Semantic Element" classes (SE: circles - Figure 1). These indirect links are relationship elements. The instances of objects in our architecture are semantic elements. The instance resources are the attribute elements like the geometry (Geo) or the local placement (LP) and are structured as hierarchical trees.

These resources are of very diverse nature like the type of materials for a wall or the structural characteristics for a beam. Those are defined in the model IFC but we can also add other types of resource like Word® documents, requests on Web Services (elements of catalogue's suppliers), and so forth. These types of resources are added to the IFC via our structuring model of the IFC for the management of electronic documents. The IFC model defines only one type of direct link between two semantic elements. This link is the placement link used for the definition of the local placement between the graphic elements of scene 2D/3D. In the IFC model this link is called "IfcLocalPlacement". The whole local placement link forms a tree which is the graphic tree scene. Thanks to the definition of their semantics, the nature of the elements makes it possible to choose which elements are relevant for the actors of the project. Consequently, the elements are extracted from the platform only according to the needs.

Figure 2 shows a representation of the trade plumbing view of an IFC file. Figure 3 shows a representation of the architectural trade view. These trade views are textual information from which specific documents can be generated or associated (reports, information of management, etc). In the 3D scene all the geometrical forms defined in IFC trees are converted in the triangular model of surface (Ronfard & Rossignac, 1996). During this conversion, the 3D objects are associated to a GID. This GID is the general identifier used to identify each trade object of an IFC file. The GID is used to combine 3D visualization with the information stored in the database. Thanks to this database and the GID, all types of information can be combined to a trade object of a 3D scene.

Sharing and data exchange: The ACTIVe3D method: The mechanisms of management and handling of IFC files,

Figure 2. A 3D plumbing context view

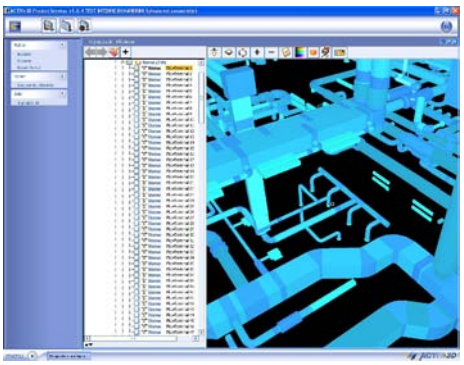
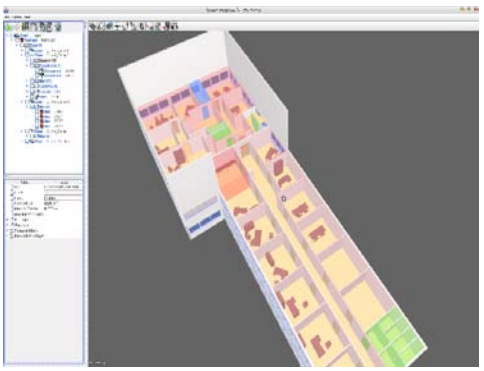


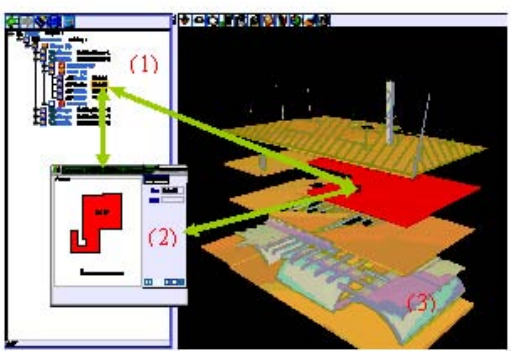
Figure 3. A 3D architectural context view



like the fusion of two files in one, the partial extraction of data from one file, visualization or storing, must take into account the multiple semantic values of the objects, which depend on the context of use. To achieve this goal, we defined a hierarchical structure of context called contextual view. The solution consists in reducing the complexity of a cyclic multiple-context graph in an acyclic mono-context graph.

Figure 4 presents the 3D scene management system which builds a specific user interface made up of a tree of composure (1), a 3D scene (2) and a technical chart (3) on a semantic element of the scene. The navigation between the elements is carried out using hypermedia links which associate a set of semantic elements to a trade object. In this case the trade object is a semantic element "Slab" (Slab - Figure 4). Certain contextual trees are generated dynamically by the system starting from IFC files. Others can be created specifically by the actors to structure their data according to their own format (starting from an IFC file or starting from existing trees).

Figure 4. Snapshot of the 3D scene management system



The principal tree is the geometrical contextual tree which contains the topological relations of the various objects. The resulting 3D scene corresponds to a particular trade view (Kim, Hwang, & Kim, 2002). This view corresponds to the actor trade association. This one is customized by the actor according to his needs, its rights, and the size of the data to be transmitted on the network. Starting from this interface, the actor can update the model while adding, modifying or removing part of the principal tree. The selection can be also carried out through the 3D scene by selecting the 3D objects. The following section shows the services available to the speakers of the project, underlining the collaborative aspect of the processes of the platform.

INTEROPERABILITY DEMONSTRATION

This section shows the use of the ACTIVe3D-Build platform as well as IFC files within the framework of civil engineering projects. In the first part of this section we will see the design phase of a building through actions made by several actors working on different CAD software. The second part of this section presents the technical study phase of the building for the structural and thermal validation of the building being designed.

Phase 1: Conception

The design of a building is divided into four phases. The first phase consists of bringing a ground statement of an old building to the platform. The second phase consists of defining an extension to this building. The third phase carries out the extension and the last phase defines the building's floors raised in the first intervention.

Viz'all: A solution for building statement via a pocket PC: Viz'all® is an automated solution of building statements, associating the use of a laser meter, a pocket PC, and software on a pocket PC. The principle consists in tracing by hand the sketch of the room on the touch screen of a pocket PC. After the connection and the deposit of the statement on the ground, the model of the building is updated and is available for all the other actors of the building restoration project. The other actors of the project can now visualize the building starting from the 3D interface of the ACTIVe3D-Build platform.

ADT, Mock-up enrichment: After the deposit of the ground statement, other actors of the project connect themselves to the platform to retrieve this statement and to enrich the mock-up. For this, the architect defines new spaces by using ADT (Autodesk Architectural Desktop). Once the architect has finished his updates on the model concerning the future building extension, he adds these new data to the platform.

ARCHICAD, Mock-up enrichment: Following the architect's updating of new spaces in the extension of the building, the engineers of the civil engineering connect themselves to the platform to collect the last information. These engineers work on ArchiCAD® from Graphisoft. When they have completed their work of building design, these new data are added to the digital building mock-up that is designed on the ACTIVe3D-Build platform.

ALLPLAN, Importation of file and finalization of the building: The second part of the project on this building is the rehabilitation of the existing building. For that, a team of engineers manages the design of this part. As the other teams do, this one is connected to the platform to extract information concerning the principal building. This team works with the software AllPlan® from Nemetschek Systems Inc. Once the updates have been carried out, the engineers put this new information on the platform.

Assessment: Thanks to the platform, a set of actors can exchange information about the building being designed and this between various types of CAD software. The IFC 2.x standard is used to format the data sent to each actor. All the data flows forward through the ACTIVe3D-Build platform because it allows each actor to have all data up-to-date, once they were placed on the server. The effectiveness of this exchange process and the centralization of information saves important time. Indeed, the data exchanges take place on a daily basis in design projects. Thus, the waiting for data updates can block the work of another team; therefore, the access to the up-to-date digital mock-up on the platform makes it possible to resolve emergencies more quickly.

Phase 2: Technical Studies

We saw in the previous section that it is possible to add information to the semantic elements of the digital mock-up of the building. This semantic information will be re-used thereafter in the creation processes of new data on the building using calculation software.

RobotBAT, Structural calculation: During the design of a building, the structures, like the beams and the columns, must be validated. Indeed, if the structures are too weak and they do not respect the standards, then the plans must be modified accordingly. RoboBAT is a structure calculation software. This software optimizes and validates the structures according to the national and the European standards about reinforced concrete, wood, steel, aluminium, and so forth. This software is able to import IFC data. Consequently, the structural engineering and design departments can validate information of the digital mock-up being designed on the ACTIVe3D-Build platform. To realize this study, the engineers must connect themselves to the platform and select all the elements concerning the structure. These elements are the walls, the slabs, the beams, etc. For that, they use the definition of the contextual tree "structural analysis".

BBS Slama, Thermal calculation: There are standards for the validation of heat exchanges between building spaces. The thermal module of the software CLIMA-WIN® made by the company "BBS Slama" makes it possible to carry out the calculation of heat loss "Th-D 1991" as well as the lawful coefficients of the buildings according to "ThBât/ThU rules" 2001. This software imports and exports IFC data. This allows to update and to validate the digital mock-up concerning the heat exchange.

Windesc, Calculation of Bill of Quantities: Windesc® is a tool of bill of quantities from the company ATTIC+ which imports IFC files. This software provides reports/ratios and estimates in connection with surfaces of walls, grounds, and so forth. This tool is essential to establish the cost revaluation of the building construction. The reports/ratios and the estimates carried out on the digital mock-up are then added to information concerning the building and the phase of design.

FUTURE TRENDS

CAD Software interoperability is a main issue that was resolved partially by the information system presented. But there still remains a second main issue, which is the semantic coherency between geometries and them semantic definition. Indeed, when users add information in the platform, this can be false. For instance, "a window is in a wall" is correct information but the different geometries can be located in a wrong place. Consequently, the window is not really in the wall. Future trends consist in defining a process to check the semantic coherency.

CONCLUSION

We have presented the importance of interoperability for civil engineering projects and how to manage it. Thus, we have presented the Active3D Web platform which makes it possible to associate semantics, 3D and documents. This method was adapted to the IFC to allow the semantic handling of the building components. Currently, we are developing a module of 3D acquisition connected to the platform to convert a cloud of points into a set of IFC semantic elements. For this, the knowledge contained in the IFC model is used to search 3D objects in a cloud of points.

REFERENCES

Aliakseyeu, D., Martens, J.B., Subramanian, S., Vroubel, M. & Wesselink, W. (2001). *Visual Interation Platfom.* In Interact 232-239. July, Tokyo, Japan.

Balaguer, J.F.& De Gennaro, S. (1996). VENUS: A virtual reality project at CERN. *Computer Graphics* 30, November 4, 40-48.

Baxter III, W. V., Sud, A., Govindaraju, N. K. & Manocha, D.(2002). *GigaWalk: Interactive Walkthrough of Complex Environments*. Eurographics Workshop on Rendering.

Cavalcade (n.d.) : http://vr.c-s.fr/cavalcade/index.html.

Christopoulos, D., Gaitatzes, A., & Papaioannou, G.(2003). Image-Based Techniques for Enhancing Virtual Reality Environments. 2nd International Workshop on ICT's, Arts and Cultural Heritage, November, Athens, Greece.

Funkhouser, T., Teller, S., Sequin, C. & Khorramabadi, D. (1996). The UC Berkeley System for Interactive Visualization of Large Architectural Models, Presence. *The Journal of Virtual Reality and Teleoperators*. Vol. 5, nb1, MIT Press, pp.13-44.

Gortler, S.J., He, L.W. & Cohen, M. F.(1997). Rendering layered depth images. *Technical Report MSTR-TR-97-09*, Microsoft Research.

Hoppe, H.(1996). Progressive meshes, *ACM SIGGRAPH*, 99-108.

Kim, B.H. Hwang, J. & Kim, Y.C. (2002). *The design of high-level database access method in a Web-based 3D object authoring tool*. The Fourth International Conference on Distributed Communities on the Web, 3-5 April, Sydney, Australia.

Klinker, G., Dutoit, A., Bauer, M., Bayes, J., Novak, V. & Matzke, D. (2002). Fata morgana – a presentation system for product design. *In International Symposium on Augmented and Mixed Reality (ISMAR).*

Levoy, M. & Hanrahan, P. (1996). Light field rendering in Computer Graphics. *SIGGRAPH 96 Proceedings*, 31-42.

Mark, W. R., McMillan, L. & Bishop, G.(1997). Post-rendering 3D warping. *In Proceedings of the 1997 Symposium on Interactive 3D Graphics. ACM SIGGRAPH.*

Paillot, D., Merienne, F., Frachet, J.P., & Neveu M. (2003). Triangulation et simplification de modèles surfaciques application à la visualisation temps réel. *GTMG, Aix en Provence*, 131-138.

Pearce Partial, D., & Day, A.M. (2004). Visibility for Virtual Reality Applications, *WSCG'2004*. February 2-6. Plzen, Czech Republic.

Ronfard, R. & Rossignac, J. (1996). Full-range Approximation of Triangulated Polyhedra. *Computer Graphics Forum* 15(3): 67-76.

Wonka, P., Wimmer, M. & Sillion, F. (2001). *Instant Visibility, EG'01 Proceedings, vol. 20(3),* Blackwell Publishing, A. Chalmers and T.-M. Rhyne, pp. 411-421.

KEY TERMS

Building Lifecycle: The lifecycle of a building is articulated in two parts. The first part is about the construction into a civil engineering project. The second part concerns the "use of the building" which deals with facilities management. Currently, these two parts are dissociated in the building management processes. The Teams which are concerned with the processes facilities management are rarely those who have participated in the construction of the building. The facilities management step often begins with a physical analysis of the building to obtain a numerical representation of this building in CAD software. To avoid information loss acquired during the construction of the building, it is necessary to develop a building information system at the beginning of its lifecycle.

CAD: (Computer Aided Design) The use of computer programs and systems to design detailed two- or three-dimensional models of physical objects, such as mechanical parts, buildings, and molecules.

Civil Engineering: Includes the planning, the designing, the construction, and the maintenance of structures and altering geography to suit human needs. Some of the numerous subdivisions are transportation; for instance railroad facilities and highways, hydraulics; like river control, irrigation, swamp

draining, water supply, and sewage disposal' and structures by example buildings, bridges, and tunnels.

IAI: The International Alliance for Interoperability founded in 1995 is an organization representing widely diverse constituencies from architects to software companies and building product manufacturers. The members promote effective means of exchanging information among all software platforms and applications serving the AEC+FM community by adopting a single Building Information Model (BIM).

Interoperability: It is the ability of several systems, identical or completely different, to communicate without any ambiguity and to operate together.

ISO: "International Organization for Standardization" is a network of the national standards institutes of 148 countries, on the basis of one member per country, with a Central Secretariat in Geneva, Switzerland, that coordinates the system. ISO is a non-governmental organization. *http://www.iso.org*

Mock-Up: A mock-up is usually a full-sized scale model of a structure which is used for demonstration, study or testing. A digital mockup is a 3D graphical model.

Challenges in Data Mining on Medical Databases

Fatemeh Hosseinkhah
Howard University Hospital, USA

Hassan Ashktorab
Howard University Hospital, USA

Ranjit Veen
American University, USA

M. Mehdi Owrang O.
American University, USA

INTRODUCTION

Modern electronic health records are designed to capture and render vast quantities of clinical data during the health care process. Technological advancements in the form of computer-based patient records software and personal computer hardware are making the collection of and access to health care data more manageable. However, few tools exist to evaluate and analyze this clinical data after it has been captured and stored. Evaluation of stored clinical data may lead to discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management. A common goal of the medical data mining is the detection of some kind of correlation, for example, between genetic features and phenotypes or between medical treatment and reaction of patients (Abidi & Goh, 1998; Li et al., 2005). The characteristics of clinical data, including issues of data availability and complex representation models, can make data mining applications challenging.

BACKGROUND

Knowledge discovery in databases (KDD) is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data (Adriaans & Zantinge, 1996; Han & Kamber, 2001). Data mining is one step in the KDD where a discovery-driven data analysis technique is used for identifying patterns and relationships in datasets. Recent advances in medical science have led to revolutionary changes in medical research and technology and the accumulation of a large volume of medical data that demands in-depth analysis. The question becomes how to bridge the two fields, data mining and medical science, for an efficient and successful mining of medical data.

While data analysis and data mining methods have been extensively applied for industrial and business applications, their utilization in medicine and health care is sparse (Abadi & Goh, 1998; Babic, 1999; Brossette, Sprague, Hardin, Jones, & Moser, 1998). In Ohsaki, Yoshinori, Shinya, Hideto, and Takahira (2003), the authors discuss the methods of obtaining medically valuable rules and knowledge on pre- and post-processing and the interaction between system and human expert using the data of medical tests results on chronic hepatitis. They developed the system based on the combination of pattern extraction with clustering and classification with decision tree and generated graph-based rules to predict prognosis. In Tsumoto (2000), the author focuses on the characteristics of medical data and discusses how data miner deals with medical data. In (Ohsaki et. al., 2007), authors discuss the usefulness of the interestingness measures for medical data mining through experiments using clinical datasets on meningitis. Based on the outcomes of these experiments, they discuss how to utilize these measures in postprocessing.

The data mining techniques such as Neural Network, Naïve Bayes, and Association rules are at present not well explored on medical databases. We are in the process of experimenting with a data mining project using gastritis data from Howard University Hospital in Washington, DC to identify factors that contribute to this disease. This project implements a wide spectrum of data mining techniques. The eventual goal of this data mining effort is to identify factors that will improve the quality and cost effectiveness of patient care.

In this article, we discuss the challenges facing the medical data mining. We present and analyze our experimental results on gastritis database by employing different data mining techniques such as Neural Network, Naive Bayes, and Association rules and using the data mining tool XLMiner (Shmueli, Patel, & Bruce, 2007; XLMiner, 2007).

MEDICAL DATA MINING: CHALLENGES

The application of data mining, knowledge discovery and machine learning techniques to medical and health data is challenging and intriguing (Abidi & Goh, 1998; Brossette et al., 1998; Cios & Moore, 2002). The datasets usually are very large, complex, heterogeneous, and hierarchical and vary in quality. Data preprocessing and transformation are required even before mining and discovery can be applied. Sometimes the characteristics of the data may not be optimal for mining or analytic processing. The challenge here is to convert the data into appropriate form before any leaning or mining can begin.

There are a number of issues that must be addressed before any data mining can occur. In the following, we overview some of the challenges that face the data mining process on medical databases (Tsumoto, 2000).

High Volume of Data

Due to the high volume of the medical databases, current data mining tools may require extraction of a sample from the database (Cios & Moore, 2002; Han & Kamber, 2001). Another scheme is to select some attributes from the database. In both approaches, domain knowledge can be used to eliminate irrelevant records or attributes in reducing the size of the database (Owrang, 2007).

Update

Medical databases are updated constantly by adding new results for lab tests and new ECG signals for patients. Subsequently, any data mining technique should be able to incrementally update the discovered knowledge.

Inconsistent Data Representation

Inconsistencies due to data entry errors are common problems. Inconsistencies due to data representation can exist if more than one model for expressing a specific meaning exits (e.g., the location of disease for Colitis, one application may enter (sigmoid, or rectum, etc.) and another may enter (measurements such as 20 cm, 30 cm, etc.)). Additionally, the data type does not always reflect the true data type. For example, a column with numerical data type can represent a nominal or ordinal variable encoded with numbers instead of a continuous variable. This plays an important role during statistical analysis (mean and variance).

Poor Integration

Health data is fragmented and distributed between hospitals, insurance companies and government departments. This poses a substantial challenge for data integration and data mining in terms of the confidence that can be placed in the result and the semantics of a derived rule. One can use common data dictionary and standards to integrate data from heterogeneous systems. The emergence of XML as a data standard is gaining wider acceptance and hence making integration fairly easy in the near future (Cios & Moore, 2002).

Number of Variables

The computational complexity is not linear for certain data mining techniques. In such cases, the time required may become infeasible as the number of variables grow. Techniques such as principle component analysis, available in the XLMiner data mining tool, can be used to reduce the dimensionality (number of variables) of the dataset but retain most of the original variability in the data (XLMiner, 2007). In addition, domain knowledge can be used to eliminate the irrelevant attributes from data mining consideration (Owrang, 2007).

Missing/Incomplete Data

Clinical database systems do not often collect all the data required for analysis or discovery. Some data elements are not collected due to omission, irrelevance, excess risk or inapplicability in a specific clinical context. For some learning methods such as logistic regression (XLMiner, 2007), a complete set of data elements may be required. Even when the methods accept missing values, the data that was not collected may have independent information value and should not be ignored. One possible approach for handling the missing data is to substitute missing values with most likely values (Han & Kamber, 2001; Tsumoto, 2000; XLMiner, 2007).

Noise

Medical databases include some noises. Therefore, data mining techniques should be less sensitive to noises (Han & Kamber, 2001).

Amount of Results

The quantity of output from many data mining methods is unmanageable. Association rule mining has been used in hospital infection control and public surveillance data (Brossette et al., 1998) and in Seplic Shock patient data (Li, Fu, He, & Chen, 2005). Too many rules have been found in both projects. Other problems include trivial and similar patterns observed in drug reaction data and in chronic hepatitis data (Ohsaki, Kitaguchi, Okamoto, Yokoi, & Yamaguchi, 2004).

Pattern Tuning

Many fast heuristic data mining methods need a lot of tuning and they are not easy for users to use (Li et al., 2005). For example, K-means clustering method can generate some very good results, but adjustment of parameters and initial setting are tedious for many users. Similar problems exist in the setting of support and confidence threshold for association rule mining (Han & Kamber, 2001).

Interestingness

Risk patterns are not in line with most data mining objectives. Most data mining algorithms aim to uncover the more frequent pattern. In medical applications, the risk patterns usually exist in a small population. For example, a very small percentage of people are HIV positive or develop cancer. In medical datasets, a pattern in the abnormal group would hardly be frequent because the abnormal cases are rare. Hence, this requires a special measurement of interestingness (Ohsaki et al., 2004).

Interpretation of Mining Result Set

One of the biggest challenges in mining medical data is interpreting the results from discovery vs. noise (Li et al., 2005). In general, it is difficult to interpret results from neural networks. Decision tree can be extended to rules, and their results are more straightforward to interpret. Some techniques do not work well on the skewed cases in medical data where the normal population greatly outnumbers the population of disease (Ohsaki et al., 2004).

DATA MINING TOOL: XLMiner

XLMiner (Shmueli et al., 2007; XLMiner, 2007) is the data mining tool that we used for our experiments. XLMiner is a comprehensive data mining add-in for Excel. It has extensive coverage of statistical and machine learning techniques for classification, prediction, affinity analysis and data exploration and reduction. For our experiments, we considered Naïve Bayes (for classification), Neural Networks (for Prediction) and the Association rules (for Affinity) data mining algorithms.

DATA PREPARATION

The medical dataset is from Howard University Hospital in Washington, DC. The GastroIntestinal (GI) Pathology data contains the records of patients from the years 2002 to 2006. The data set is represented in MS Exel work sheet and contains 5,700 records. The dataset consists of the following attributes: File number, Specimen Number, Specimen Date, HP-test, ELISA, Silver stain, cag-A+, Age, Sex, Race, Nationality, Type of Surgery, Diagnosis, Location of disease, Death, Date of death, Comments, Previous History, and Present Illness.

The data preparation stage involved two stages: data cleaning and transformation. The data cleaning step eliminated inconsistent data, for example, data type inconsistencies (some numeric fields contain alphanumeric character). The transformation aims at converting all the data fields to numeric fields as required by the Neural Network data mining algorithm. It is also required to extract a subset of the fields and records that do not contain missing values.

The typical errors include the following:

1) Data field HP-test had mixed case values (positive, pos., POSITIVE, pos, negative, neg, NEGATIVE, neg.)
2) Data field Sex had values (M,F, f, male, female)

We used the SQL Update operation of the MS-Access database software to handle the records with exception. For column HP-test, all combinations were updated to represent Positive/Negative values and the Sex field values updated to M and F.

Using the Naïve Bayes mining technique (Shmueli et al., 2007; XLMiner, 2007), columns that did not contain any data were rejected by XLMiner. For example, in our gastritis data, we had a lot of empty values for the HP-test attribute. Other attributes that had empty values include Race (had 4,185 values blank), Nationality (had 4,681 values blank), and Location of disease (had 534 values blank). We used the "Missing Data Utility" of the XLMiner to substitute "n/a" for columns that had missing values.

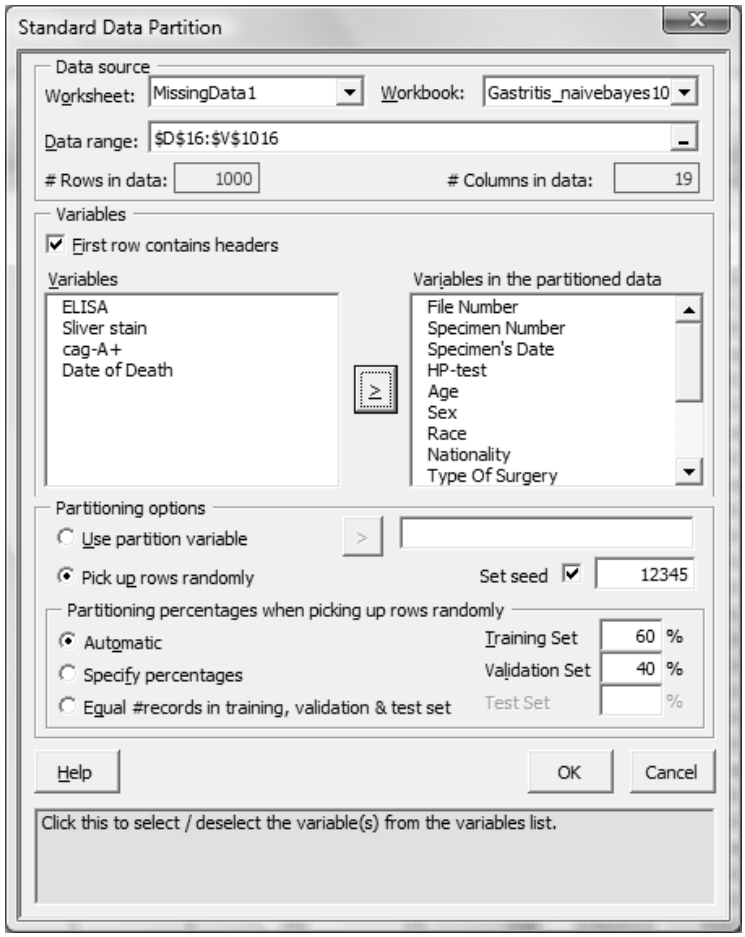
Once all NULL values were taken care of, we used the random partitioning of the XLMiner and created two mutually exclusive datasets, a training dataset comprising 60% of the total dataset, and a validation dataset of 40%, as shown in Table 1. These are the defaults for partitioning. The training dataset is used to train or build a model. For example, in a Neural Network model, the training dataset is used to obtain the network weights. Once a model is built on training data, you need to find out the accuracy of the model on unseen data, the validation dataset.

Upon selecting Naïve Bayes algorithm, we ran into distinctiveness issue as follows:

"Note: The Naive Bayes classification routine supports a maximum of 30 classes or distinct values for an Output variable."

In our original dataset, the attribute (or variable) Diagnosis contained 781 distinct values, and the attributes Type of

Table 1. Standard data partition in XLMiner



surgery contained 591 distinct values. For the experiments, we selected 1,000 records randomly. Still, we had 286 distinct values for the variable Diagnosis and 198 distinct values for the variable Type of Surgery. For Diagnosis, the values recorded had several ways to represent the same value or were spelling mistakes including the values "Mild chronic gastritis," "mild chronic gastritis." After looking at the data, we used similar grouping of values to narrow down the Diagnosis to:

1) acute and chronic gastritis
2) chronic active gastritis
3) mild chronic gastritis
4) moderate chronic gastritis
5) Chronic nonspecific gastritis
6) gastric ulcer

Type of Surgery attribute had similar issues with distinctiveness. The following are some of the possible values for Type of Surgery. Some of these values were the same due to the lack of standards in entering the data values.

1) EGD.bx (EGD.bx (EGD stands for Esophagus Gastro Dudenum and bx stands for biopsy)
2) EGD w biopsy (same as #1)
3) EGD,Gastrotomy
4) EGD.bx; colonoscopy
5) EGD.bx; polypectomy
6) Endoscopy
7) upper endoscopy,bx (same as #1)

MS-Access/SQL was used to fix issues with distinctiveness. There are some possible ways to avoid the aforementioned data issues by developing software/Web application to perform the following tasks, instead of manually entering the data:

1) Automatically filling is used that serves as counters for example Patient id, Specimen number, date and so forth.
2) Use drop down columns to select range data like Sex, Race, binary(positive or negative).

MEDICAL DATA MINING: EXPERIMENTS

There are many data mining techniques available for data classification, prediction, association analysis, and data exploration. In our experiments, we used Naïve Bayes and Association rules mining techniques. Using the gastritis dataset, Neural Networks classification rejected the majority of the input fields (17 of 19), as it does not support alphanumeric values.

In the following, we provide some of the detailed results of the experimental runs followed by a summary analysis, which includes some generated sample rules.

Gastritis Mining Using Naïve Bayes

RUN 1: Output Variable is Diagnosis. Input variables include: File Number, Specimen Number, Specimen's Date, HP-test, Age, Sex, Race, Nationality, Type of Surgery, location of disease, Death, Comments, Previous History, and Present Illness. Based on training data, the known probabilities for some of the diagnosis are shown in Table 2. Table 3 shows the diagnosis column with the highest probability for the top 2 records, upon using validation dataset.

XLMiner's Naïve Bayes data mining technique calculates the overall probability as shown above as well as conditional probabilities for other variables/attributes. In other words, Naïve Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variables. In our case, the output variable was Diagnosis. The target attribute Diagnosis was defined through each of the input variables. Only some of those rows with the highest probability are shown in Table 4.

We have done a similar run when output variable was the Type of Surgery. The intent of choosing a second output variable was to evaluate, if using the same data, we could observe a similar or different set of conditions. In the following, we provide the summary analysis of the runs and provide our views of the results.

Gastritis Data Mining Analysis Summary

In general, the interpretation of the results of mining over medical datasets requires significant domain expertise. In many cases, the discovered rules have been generally known, at least to the medical domain expert. Only in a few cases, the mining techniques have produced genuinely new knowledge. The hope for these experiments and analyses is to find the right data mining tool, the right mining technique, for the right dataset as well as to learn to properly set the data mining parameters (e.g., defining confidence threshold, data partitioning, etc.). Table 5 shows some of the rules generated

Table 2. Prior class probabilities for diagnosis

Class	**Prob.**
Mild chronic gastritis	0.001666667
acute and chronic gastritis	0.326666667
Chronic active gastritis 0	.445
chronic nonspecific gastritis	0.003333333
chronic nonspecific gastritis and duodenitis 0	.003333333
gastric ulcer	0.001666667
Mild chronic gastritis 0	.198333333
Moderate chronic gastritis 0	.008333333
nonspecific chronic gastritis	0.006666667

Table 3. Sample output for diagnosis column with the highest probability upon validation of the dataset

Predicted Class	Actual Class	Prob. for class acute and chronic gastritis	File Number
acute and chronic gastritis	acute and chronic gastritis	1	227530
acute and chronic gastritis	acute and chronic gastritis	1	763530

Table 4. Conditional probabilities for the output variables diagnosis, defined through other variables

	Classes-->	
Input Variables	chronic nonspecific gastritis	
	Value	Prob
HP-Test	Negative 1	

a. Diagnosis is defined through the variable HP-test

	Classes-->	
Input Variables	chronic nonspecific gastritis	
	Value	Prob
Sex	M	1

b. Diagnosis is defined through the variable sex

	Classes-->	
Input Variables	Moderate chronic gastritis	
	Value	Prob
Race	AA 1	

c. Diagnosis is defined through the variable race

	Classes-->	
Input Variables	Moderate chronic gastritis	
	Value	Prob
Type of Surgery	EGD.bx 0	.8

d. Diagnosis is defined through the variable type of surgery

by XLMiner using the Naïve Bayes and Association rules data mining techniques.

SUMMARY OF OBSERVATIONS

The Bayesian routine produced 35 rows from the validation output identifying records that were a match based on the training data. When the actual value from the training set matched the predicted value, the record was identified and expressed as a confidence percentage. Apart from the overall match for records that satisfy the criteria of the target output variable 1) Diagnosis or 2) Type of Surgery, conditional probability of each input variable vs. the target output variable was also recorded. Some of the observation shows that if HP-test is negative there is a likely chance of some form of gastritis. Of the available Race data (missing for many rows) it points to the fact that if Race =AA (AA stands for African American) there is a likely chance of gastritis. Also, the generated rules indicate that more female patients had some form of chronic gastritis.

In the gastritis dataset, redundant rules were generated due to redundant data (e.g., nonspecific chronic gastritis and chronic nonspecific gastritis are the same). Furthermore, with

Table 5. Sample rules generated by XLMiner using Gastritis data

XLMiner Technique	**Findings**
Naïve Bayes	**Using Diagnosis as the target Output variable**: 1. Chronic nonspecific gastritis: if HP-test is negative. Gastric ulcer: if HP-test is negative 2. Mild chronic active gastritis: if Age is 70 Gastric ulcer: if Age is 56 3. Chronic nonspecific gastritis: if Sex =M gastric ulcer: if Sex=F 4. Mild chronic active gastritis : if Race is AA **Using Type of Surgery as target Output variable:** 1. EGD, pancolo.bx: if HP-Test is negative upperendoscopy, bx: if HP-Test is negative panendo.bx: if HP-Test is positive 2. colon,bx: if Age is 31 upper endoscopy,bx: if Age is 35 pancolonoscopy.bx: if Age is 64 3. EGD w biopsy: if Sex=M upperendoscopy, bx: if Sex=M colon,bx: if Sex=F pancolonoscopy.bx: if Sex=F 4. EGD. Bx: if Race=AA
Neural Networks Classification	**Incompatible data.**
Association Rules	**No. of Rules generated = 246, TOP 5 Rules:** 1. IF Race=AA and Sex=F THEN Type of Surgery = EGD.bx has a confidence of 79% 2. IF Diagnosis =Mild chronic gastritis THEN HP-Test = Negative has a confidence of 76% IF Race=AA THEN Type of surgery= EGD.bx has a confidence of 74% 3. IF Diagnosis = acute and chronic gastritis THEN HP-Test = positive has a confidence of 64% 4. IF Race=AA and Type of Surgery=EGD.bx THEN Sex= F has a confidence of 63% 5. IF Sex=F and HP-Test=positive THEN Type of Surgery=EGD.bx has a confidence of 60%

Type of Surgery as target output, several irrelevant and trivial rules were generated, which indicates a relationship between Type of Surgery and HP-test. According to domain expert, the result of HP-test does not have any relationship with the Type of Surgery. Likewise, the Type of Surgery does not imply a particular value for the HP-test. The surgery (e.g., biopsy) needs to be done so that testing can be done. This problem indicates that we need the expertise of the medical experts in order to guide the discovery process in order to avoid generating irrelevant rules.

From the Association rules data mining technique, a total of 246 rules were generated using XLMiner. There were rules that were trivial, redundant, consistent with domain expert's views, as well as some that simply represented some facts about the data, for example, more female African Americans had gastritis.

Some of the findings from both approaches match while others differ as in the case of the fields Race or HP-test. In the Bayesian routine, if HP-test is negative there is a likely chance of some form of gastritis. This observation was

missing from the Association rules, contrary to the Race observations that were similar.

In our experiments, Naïve Bayes discovered more and better findings than Association rules. Some of the factors that contributed to this suggestion are summarized as follows:

1. The quality of the gastritis data set can be best described as below average or poor. Many columns had missing data, which is not conducive for Association rules. Association rules find interesting associations and relationships among large set of data items based on value conditions that occur frequently. The confidence factors from Association rules barely touched 70%. We had to lower the thresholds to identify interesting correlations.
2. A guided learning approach was used in Naïve Bayes technique. The prediction of the overall probability as well as conditional probability was performed using output variable and its correlation with each variable independent of other variables. In contrast, Association rules technique was minimally guided. The only option was to specify an output variable, and rule sets were presented.

FUTURE TRENDS

The following are other issues that need to be addressed before we could have a comprehensive and viable medical data mining environment (Cios & Moore, 2002; Li et al., 2005).

Poor Mathematical Characterization of Data

Business, financial, and scientific data can be easily modeled, transformed and applied formulas in contrast to medical data, whose underlying structure is poorly classified in mathematical terms. Medical data consists of images and free hand data with few constraints on vocabulary or image. In comparison, business and financial data have formal structures into which we can classify and organize data that may be modeled by Linear Regression, Neural Networks, and Naïve Bayes vs. medical attributes such as bloating, inflammation and swelling.

Health Professionals Interpretation

The doctors interpretation of tests conducted, imaging and other clinical data is generally written in free text that is very difficult to comprehend and difficult to standardize. This poses a great challenge for miners. Many times, professionals from the same field cannot agree upon interpreting a patient's condition but also use different names to describe the same disease.

Imaging and Complexity of Medical Data

With advances in medicine and imaging technology, the complexity of data has increased. Increasingly, clinical procedures are employing imaging as a tool of choice for diagnosis. There needs to exist efficient mining in databases of images which are much more difficult than mining purely numerical data. X-ray, cat scans and so forth become more prevalent and necessary diagnostic techniques, and the challenges to interpret and comprehend them in data mining will grow exponentially.

Ethical, Legal, and Privacy Issues

Clinical and medical data is primarily focused on humans and thereby becomes a primary target for abuse and misuse. Effective legal and ethical frameworks need to be in place to prevent their misuse.

CONCLUSION

This article focused on the characteristics of medical data and provided some experimental results on a gastritis medical database using the data mining tool XLMiner and data mining techniques of Naïve Bayes, Neural Network, and Association rule. It should be pointed out that the experimental results indicate that the discovered rules appear to be consistent with the domain experts' views. However, several meaningless rules were generated due to the wrong data entry. Although different data mining techniques of Naïve Bayes and Association rules produced some similar rules, there were rules generated in one technique and not in the other. Neural Network data mining technique could not be used, as most of the attributes were non-numeric.

The next step in this experiment is to use a bigger gastritis dataset along with more data cleaning. In addition, data mining techniques including K-means clustering and classification trees could be used (Shmueli et al., 2006). Most of the non-numeric values of the gastritis attributes need to be transformed to numeric values in order to be able to use the Neural Network mining technique of the XLMiner. Also, it might be worthwhile to group certain categories (e.g., the field Age) which may improve our chances of discovering less and meaningful rules. In the current scheme, each age entry was evaluated independently to find out the Diagnosis or Type of Surgery. However, one could group the age as 1-10 child,11-20 youth, 21-30 young adult, 30-50 adults, 50-65 old and 66 and above very old.

It would be beneficial to use other mining tools including Insightful Miner (Insightful Miner, 2007) in order to improve/verify the correctness and completeness of the discovered knowledge. Finally, domain knowledge (Owrang, 2007) can be used to reduce the computational complexity of data mining algorithms as well as to improve the quality of the discovered knowledge.

REFERENCES

Abidi, S. S. R., & Goh, A. (1998). Applying knowledge discovery to predict infectious disease epidemics. In H. Lee & H. Motoda (Eds.), *Lecture notes in artificial intelligence 1531- PRICAI'98: Topics in artificial intelligence*. Berlin: Springer-Verlag.

Adriaans, P., & Zantinge, D. (1996). *Data mining*. Addison-Wesley.

Babic, A. (1999). Knowledge discovery for advanced clinical data management and analysis. In P. Kokol et al. (Eds.), *Medical informatics Europe'99,* Ljubljana. Amsterdam: IOS Press.

Brossette, S. E., Sprague, A. P., Hardin, J. M., Jones, K. W. T., & Moser, S. A. (1998). Association rules and data mining in hospital infection control and public health surveillance. *Journal of American Medical Informatics Association, 5*(4), 373-381.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine, 26*(1).

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Morgan Kaufmann.

Insightful Miner. (2007). *Online documentation*. Retrieved May 31, 2008, from http://www.insightful.com/

Li, J., Fu, A. W., He, H., & Chen, J. (2005, August 21-24). Mining risk patterns in medical data. In *Proceedings of KDD'05*, Chicago, IL, USA.

Ohsaki, M., Kitaguchi, S., Okamoto, K., Yokoi, H., & Yamaguchi, T. (2004). Evaluation of rule interestingness measures with a clinical dataset on hepatitis. In *Proceedings of European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, (pp. 362-373).

Ohsaki, M., Yoshinori, S., Shinya, K., Hideto, Y., & Takahira, Y. (2003). A rule discovery support system for sequential medical data: The case study of a chronic hepatitis dataset. *SIG-KBS, 6*, 117-122.

Ohsaki, M., Hidenao, Tsumoto, Shusaku, Yokoi, Hideto, & Yamaguchi, Takahira (2007, November). Evaluation of rule interestingness measures in medical knowledge discovery in databases. *Artificial Intelligence in Medicine, 41*(3), 177-196.

Owrang O. (2007). Discovering quality knowledge from relational databases. In L. Al-Hakim (Ed.), *Information quality management: Theory and applications* (chap. III). Hershey, PA: Idea Group.

Shmueli, G., Patel, N. R., & Bruce, P. C. (2007). *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. Hoboken, NJ: Wiley InterScience.

Tsumoto, S. (2000). Problems with mining medical data. In *Proceedings of the 24th International Computer Software and Applications Conference (COMPSAC' 00*), **(p**p. 467-468).

XLMiner. (2007). *Online user guide*. Retrieved May 30, 2008, from http://www.xlminer.net/

KEY TERMS

Affinity: Association rules

Association Rules: Association rule mining finds interesting associations or correlation relationships among large set of data items. Association rules show attributes value conditions that occur frequently together in a given dataset. Association rules provide information of this type in the form of "if-then" statements.

Classification: Discriminant Analysis, Naïve Bayes, Neural Network, Classification Tree, Logistic Regression, K-Nearest Neighbor

Data Mining: Data mining is one step in the knowledge Discovery in Databases (KDD) where a discovery-driven data analysis technique, such as Naïve Bayes or Neural Networks or Association rules, is used for identifying patterns and relationships in data sets.

Data Mining Techniques: The general categories of the data mining techniques include:

Data Preparation: The data preparation stage involved two stages: data cleaning and transformation. The data cleaning step eliminated inconsistent data. The transformation aims at converting the data fields to numeric fields or vice versa.

Knowledge Discovery: Knowledge discovery in databases (KDD) is defined as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

Naïve Bayes Classification: Bayesian classifiers operate by saying "If you see a fruit that is red and round, which type

of fruit is it most likely to be, based on the training data set? In future, classify red and round fruit as that type of fruit."

Neural Networks Classification: Artificial neural networks process records one at a time, and "learn" by comparing their prediction of the record (which, at the outset, is largely arbitrary) with the known actual record. The errors from the initial prediction of the first record is fed back into the network, and used to modify the networks algorithm the second time around, and so on for many iterations.

Prediction: Multiple Linear Regressions, K-Nearest Neighbor, Regression Tree, Neural Networks

XLMiner: XLMiner is a comprehensive data mining add-in for Excel. It offers a variety of methods to analyze data. It has extensive coverage of statistical and machine learning techniques for classification, prediction, affinity analysis and data exploration and reduction.

Category: Electronic Business

Challenges of Interoperability in an Ecosystem

Barbara Flügge
Otto-Von-Guericke Universität Magdeburg, Germany

Alexander Schmidt
University of St. Gallen, Switzerland

INTRODUCTION

True e-enabled collaboration has been assessed for many years. With the growing reach of companies' business and cross-border trade, the entire ecosystem enterprises are embedded in is playing a crucial role to succeed. As ICT is a key driver for deploying true interoperability and integration among the participants of the ecosystem, actors with a lack of ICT knowledge, equipment, and implementation represent the vulnerable parts within the ecosystem. This article aims at providing an overview of challenges limiting business partners in an ecosystem to truly e-collaborate. Furthermore, it describes the key elements of e-enabled collaboration and interoperability ranging from the technical and business oriented to cross-organizational and cultural aspects.

BACKGROUND

There are two main directions that have been the basis for extensive research over the last decades to touch the ground for successful electronic collaboration (e-enabled collaboration). One direction led researchers to the field of organizational development. The other direction led to the field of ICT support and solutions initiating and facilitating collaboration models. An example of the initiation of collaboration models is the commencement of the e-commerce hype in the 1990s.

The magnifying glass that allows the focus on e-enabled collaboration is the set of key characteristics in these fields that are relevant to facilitate, change, or extend the level of e-enabled collaboration. The following paragraphs are focusing on what we explore by applying the magnifying glass.

Collaborative Environments

The point of origin that leads to the foundation of any ecosystem varies. We are assuming that the ecosystem is formed because of a common interest in conducting business successfully, competitively, and innovatively. Business transactions are executed to request, support, deliver, and exchange goods, services, and data. Each of the participants in the ecosystem contributes actively to the business purpose. They are ordering, delivering, supporting, producing, assembling, and selling goods, services, and data based on their roles and capabilities. Thus, the foundation of an ecosystem is related neither to a specific sector or region nor to the means that are required to run an ecosystem.

The ecological ecosystem is providing extensive research opportunities to analyze interactions, relationship building, and the evolution of organisms. The history of ecosystem research started with Sir Arthur Tansley (1935) when he introduced the term *ecosystem* based on Phillips' studies on complex organisms and the common term of *biotic communities* valuing similarities and boundaries of communities. He is comparing these terms to his own view of describing the changing vegetation, participants, and relationships.

Ritter, Wilkinson, and Johnston (2004) are focusing on the managerial value-related competences of organizations to steer, interact, and cooperate in a business-related network. Referring to Håkansson and Snehota's (1993) role of relationship building, any enterprise needs to broaden its business role by interacting and actively building relationships with its environment (Ritter et al.). In the work of Ritter et al., the environment of an enterprise is comprised of customers, "complementors," competitors and suppliers. Besides the given terms customer, competitor, and supplier, complementors are defined as "types of firms whose outputs or functions increase the value of their own outputs" (Ritter et al., p. 3). On the value side, Brandenburger and Nalebuff (1996) introduced the value net as a term to symbolize the dedicated purpose of realizing value in any given or created relationship among business partners.

Network-related research led to comparing studies. An extensive study conducted by Changizi, McDannald, and Widders (2002) examines the relevance of network size and the capability to grow in different networks such as ecological, technical, human-being, and urban networks. The number of participants joining the network is one of the positive effects that networks participants experience according to Farrell and Saloner (1985) and Reimers and Li (2005). We cautiously draft the analogy of ecological and business-oriented networks to ecosystems due to the fact that the capability to power play and act in a competi-

tive environment is determining the capability to grow and extend the given network from within. Networks that are not business-purpose driven like the Lego network in the case of Changizi et al. are excluded from that assumption.

Collaborating Participants

Further down the exploration of the ecosystem, the decomposition of the ecosystem requires a greater analysis of its relationships and participants. First, we are amplifying in the following the most relevant details of the participants of the ecosystem, representing enterprises, business partners, governmental institutions, and any other involved entity.

The success of any participant in an ecosystem is founded in the capability of the participants themselves in relation to the underlying purpose of establishing the ecosystem. Capability is determined by the level of activity of any participant, its interaction intensity within the ecosystem with one or various intentionally or unintentionally selected participants, the role and responsibility any participant is administering compared to the other participants, and the capabilities participants are offering to the ecosystem. Eisenhardt and Santos (2005) elaborate the key elements of organizational types that are relevant to conceptualize the boundaries of an organization. They distinguish four conceptions of boundaries: efficiency, power, competence, and capability. All four conceptions are main determinants relevant to an ecosystem.

Enlarging the view of the concept of boundaries, it is also relevant to the interaction capability of an organization with its participants in the ecosystem and with other ecosystems. However, efficiency and competence are the most important assets organizations need to enrich and increase the collaboration capability. According to Eisenhardt and Santos (2005), efficiency is required to minimize governance costs, including costs of conducting exchange with other ecosystem participants and those within the individual organization. Competence allows the organization to align its resources, skills, products, and services to outperform external opportunities and market expectations (Eisenhardt & Santos). Thus, five main characteristics of the ecosystem have been elaborated: (a) the ability to individually assign the purpose of an ecosystem to its components (participants), (b) the interactions (among and between participants), (c) the development process within an ecosystem (influencing the ongoing evolution), (d) the maturity and stability of an ecosystem and its components, and (e) the effects an ecosystem is causing in terms of results, measurements, changes in size, and composition.

Coming back to the initial details that are relevant to assess the entire reach of an ecosystem, in the second part of this article we focus on the relationships within an ecosystem. These are comprised of the flow of goods, services, and data. Matutinovíc (2002) applies the concept of flow networks, elaborating on ecological ecosystems and common patterns that address the flows and needs of organizational ecosystems. According to Matutinovíc, the purpose of any existing or planned relationship is based on the following parameters: competition, cooperation, and selection, creating feedback to the participants and positively forcing each to optimize its relationships. Given the fact that resources as outlined by Eisenhardt and Santos (2005) are one of the key parameters, organizations are constrained to keep their competency level high. Any resource optimization, including ICT and process optimization, is a key determinant of successful collaboration. The choreography of ICT solutions and processes will be outlined in the next section.

Role of ICT in E-Enabled Collaboration

The second direction researchers are concentrating on is the field of ICT support and solutions initiating and facilitating collaboration models. Various e-business and e-government initiatives have been formed to get a closer view on e-enabling. Those encompass nearly any business process and collaboration scenario in nearly any industry sector, optimizing any kind of organizational types such as multinational companies as well as small and medium-sized enterprises.

As outlined above, the ecosystem research in many cases does not take governmental institutions into account explicitly, excluding Eisenhardt and Santos' (2005) conception of power, whereas ICT-related research has been including the need of governmental support and e-enabling collaboration with governmental institutions. The dimension of ICT focuses on the technical understanding of collaboration, the applicability of applications, and the key concepts of interoperability according to Theling, Zwicker, Loos, and Vanderhaeghen (2005). There is a number of studies, such as *The European E-Business Report* (European Commission, 2005) and the UN report on e-government and e-inclusion (United Nations, 2005), that point out the need and urge of focusing on the core roles of ICT. ICT should strengthen the collaborative environments with more than Web site publishing or providing electronic media to enhance paper-based documents and business processes. One example supporting the need of ICT in the form of interorganizational systems and business integration is reflected in the case of Denmark (Bjørn-Andersen & Andersen, 2004). Another example is given in the U.S. residential mortgage industry where Markus, Minton, Steinfield, and Wigand (2006) call for the development and adoption of standardized business semantics and business processes to further enhance collaboration and accomplish further benefits.

APPROACHING E-ENABLED COLLABORATION BY DETERMINING THE INTEROPERABILITY FACTOR FOR ANY GIVEN ECOSYSTEM

The traditional and posttraditional context that has been discussed in the section above still leaves a growing community of business and governmental partners who seek for constant and reliable effects of collaboration beyond traditional one-to-one solutions. In the area of electronic customs, for example, business and governmental partners are confronted with an increasing demand for secure trade, and compliant and accessible data at any time for any business partner participating in trade processes.

Given the outlined example of global trade and facilitating the collaboration within ecosystems that are acting in the context of international trade, the limitations in literature and research have been mainly set by three factors. First, governmental institutions have not been considered as ecosystem participants and business partners. Second, collaboration models did not develop top-down cross-organizational business models but rather focused on individual views and perspectives. Third, many published interoperability-related concepts and standards reveal deficits in understanding and applying the interoperability factor: semantically unambiguous business process collaboration.

Modern business-oriented research should be urged to address the need for providing financial or any other measurable evidence to get an open ear in the commercial and governmental community. The need is to prove that a conception works in real-life environments. Added value will be provided by including governmental business cases as well as research on applying the interoperability factor across the ecosystem. The expected outcome hereby is the definition and the concept of an enterprise architecture framework transforming the dimensions of collaboration, namely, the ecosystem, the participants, cross-organizational relationships, and processes, into feasible interoperable solutions.

INTEROPERABILITY AS A MULTIFACET CONCEPT

If we consider interoperability as being the fundament for e-enabled collaboration, first of all it is necessary to specify the meaning as well as relevant aspects of the notion. While research in the past regarded interoperability solely from a technical perspective,[1] nowadays there is a wide consensus that this view needs to be extended. In literature, the most common distinction is made between technical, semantic, and organizational interoperability, even though their naming might slightly defer (cf. European Commission, 2004; Ministry of Economic Affairs and Communications, 2005). We will stick to this differentiation in this section.

Technical Interoperability

On the technical layer, interoperability is mainly concerned with technical issues of linking information systems and services, particularly the interoperability of hardware (infrastructure) and software. On this level, we are mainly concerned with the integration and exchange of data (via common protocols) between software used in different organizations, the definition of interfaces, and interconnection services. From a simply syntactical perspective, the challenge of technical interoperability is solved to a large extent nowadays. However, the semantic dimension remains an unsolved problem and, consequently, constitutes one of the pivotal areas of research in the near future.

Semantic Interoperability

Semantic interoperability denotes the ability of different applications and business partners to understand exchanged data in a similar way, implying a precise and unambiguous meaning of the exchanged information. On this level, standards (for data as well as messages) are the crucial factor for achieving interoperability. However, the problem faced by most organizations today lies with the multitude of standards complicating the decision regarding which standard to take within an ecosystem while different standards are used by different business partners. The heterogeneity of information systems and data models between companies limits the capability of integration and has led to isolated business models using proprietary standards and services (Legner & Wende, 2006). The main insufficiencies for realizing e-enabled collaboration on the semantic layer can be summarized as follows.

First, concerning the exchange and integration of data, there is still no common approach for structuring and naming data types that are the basis for assembling messages. Furthermore, the enrichment of business data with semantics, necessary to guarantee a common understanding of business information, is not satisfactorily solved. Semantics are essential for specifying the context in which data are used (so-called context drivers such as a specific industry, country, etc.) and, hence, for true interoperability. Such an approach has to be built up on a common grammar and library. The introduction of XML (extensible markup language) in the late 1990s raised hopes of seamlessly integrating heterogeneous IT landscapes becoming a de facto standard for the technical representation of business data and led to a bunch of XML-based data standards with the goal of increasing interoper-

ability. However, the industry-specific XML representations do not ensure that the expressed business information will be understood equally between all systems as XML does not detail how this information is modeled or named. A very promising approach attacking the problem of business data interoperability is the CCTS (core component technical specification) methodology for semantic data modeling on an syntax-independent level from UN/CEFACT (United Nations Centre for Trade Facilitation and Electronic Business) and ISO (International Organization for Standardization; Stuhec, 2005). Embedded in the UN/CEFACT e-business stack, the CCTS methodology enhances interoperability of the business data exchange with the help of two further components. First, reusable collaborative business processes in which the business information is commonly used and correctly interpreted are defined by the UN/CEFACT modeling methodology (UMM; cf. UN/CEFACT, 2006a). Second, the XML naming and design rules specification (cf. UN/CEFACT, 2006b) recommended by the UN/CEFACT allows the systematic transfer of the syntax-independent core components to XML schema and instances.

Second, on the message level, the semantic business data building blocks can be assembled to unambiguously interpretable business messages. Such messages can overcome the traditional problem of varying document structures, sizes, amounts, content in contained data fields, and so forth, which result from country-specific legal frameworks or different industry standards.

The result of applying such a syntax-neutral integrated approach is both an increased reusability and adaptivity to varying requirements, enabling interoperability and e-enabled collaboration and cooperation between business partners.

Organizational Interoperability

However, as already mentioned, interoperability needs to consider far more than just the technical dimension. Particularly, business-oriented problems, such as the coordination of business models and collaborative processes, constitute fundamental challenges on the road to truly e-enabled collaborations. We refer to this dimension as organizational interoperability. It is associated with the activities and processes of organizations and the (strategic) agreements between them.

The organizational dimension of interoperability comprises the alignment of business goals between business partners as well as their business processes in order to realize the collaboration of companies with differing internal structures and processes. The concept of public processes (cf. Legner & Wende, 2007) represents an approach that aims at facilitating the cross-organizational design of business processes, including the allocation of responsibilities, the decoupling of internal and external processes (in order to reduce the visibility of internal activities containing confidential information), the formal specification of interfaces, and the support of the alignment with multiple partners.

Together with the concepts for semantic interoperability, the public-process approach enables the realization of semantically unambiguous business process collaborations as the ultimate goal of nontechnical interoperability.

Interoperability Road Map

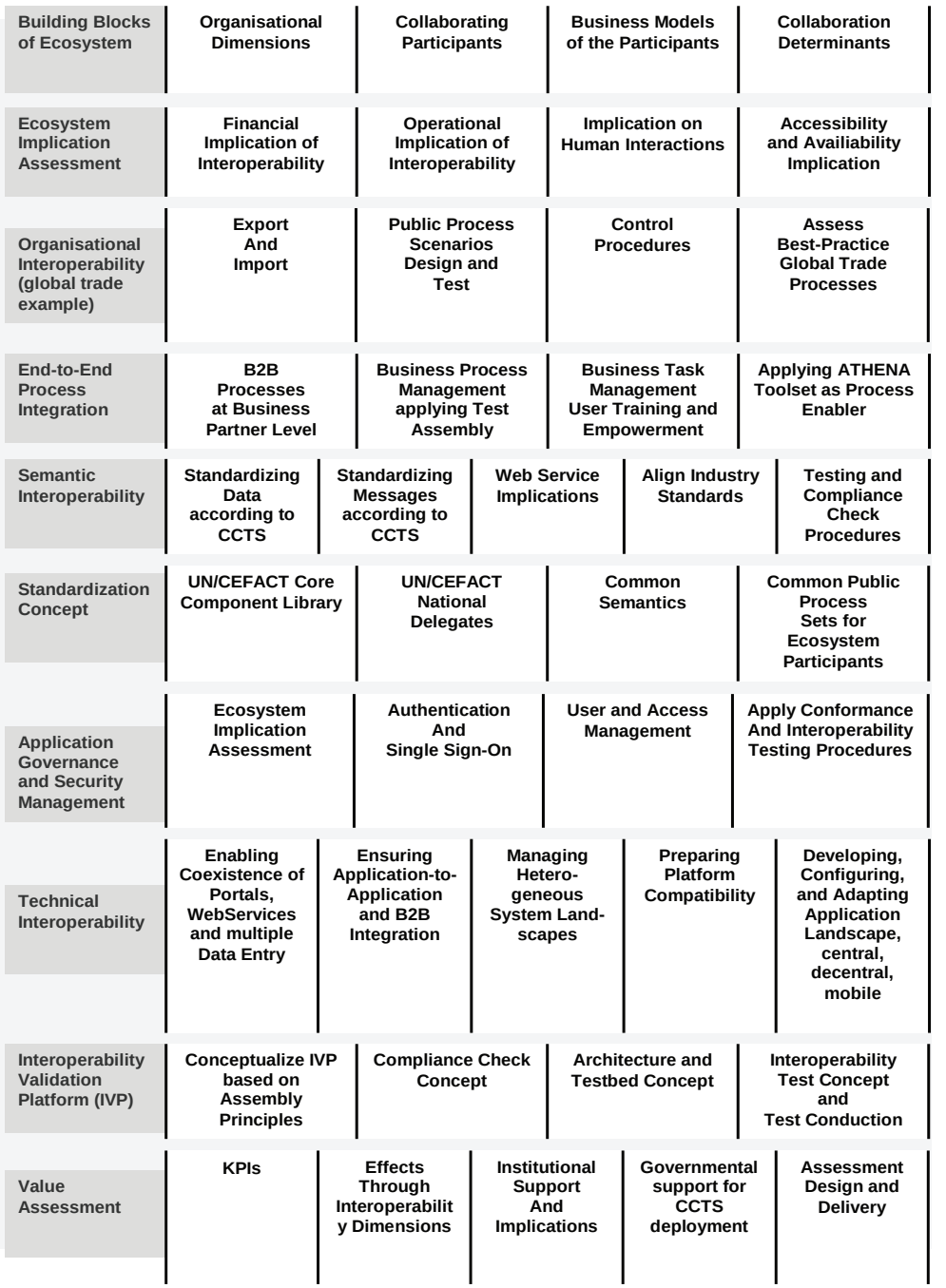
Reflecting the interoperability factor in the context of global trade, the following paragraph contains an example of how the road map to e-enabled collaboration may look like (Figure 1; Flügge, 2006). It also drafts the key components of that road map in the ecosystem. In the example of global trade, an ecosystem in the dairy-food industry is built by more than 20 participants, ranging from manufacturers such as farmers, logistics service providers, banks and insurance companies, governmental institutions such as customs and tax administration, and veterinarian authorities to service providers such as controlling companies, distributors, customs brokers, and port facilities.

The interoperability road map is comprised of a set of cornerstones seen and experienced as relevant to approach e-enabled collaboration. Each of the cornerstones covers an area that individually contributes to collaboration. They are the following: (a) organizational interoperability, (b) ecosystem implication assessment, (c) end-to-end process integration, (d) semantic interoperability, (e) the standardization concept, (f) application governance and security management, (g) technical interoperability, (h) interoperability validation platform, and (i) value assessment. Each of the cornerstones is built on a set of tasks that are highlighted in the road map and arranged next to the cornerstone they refer to.

FUTURE TRENDS

Research and development tasks in the interoperability and e-enabled collaboration areas will continue to be one of the main topics. Their success is a prerequisite for tomorrow's software engineering and development. There is a need in academic research to reflect on real business environments and their complexity on the one hand, and to transform that complexity into a research approach that allows academic partners to draw upon business needs and feedback, transferring those back into their research fields, on the other hand. Successful execution of ecosystems and collaboration is highly dependent on the standardization efforts. Research institutes and academic partners approaching collaboration and ecosystems from various perspectives and directions need to collaborate to truly assess all implications of collaboration.

Figure 1. Interoperability road map version 1.3 (Flügge, 2006)



To enable business partners to respond to innovative, commercial, and competitive objectives, researchers and practitioners should discuss the impact of value for each of the participants and the ecosystem itself. Key performance indicators to measure successful collaboration will be required to prove against the models as well as interoperability tools such as the outlined road map.

CONCLUSION

True e-enabled collaboration reaches beyond technical capabilities of software and middleware solutions. Any interaction in today's and tomorrow's business activities requires a common understanding of what needs to be shared, delivered, and exchanged and to what extent to fulfill a (common) busi-

ness purpose. The complexity in ecosystems requires a solid interoperability-grounded methodology. The methodology should provide guidelines to add collaboration as an out-of-the-box and built-in characteristic of any software solution that is offered in the market. In addition, the methodology should provide a framework with procedures suitable for any company regardless of its size, its industry focus, its technical competence, and its cultural context. Companies should not worry anymore about how they can technically content- and document-wise invite a new business partner to join the ecosystem.

REFERENCES

Bjørn-Andersen, N., & Andersen, K. V. (2004). *Diffusion and impacts of the Internet and e-commerce: The case of Denmark.*

Brandenburger, A. M., & Nalebuff, B. J. (1996). *Co-opetition.* Harper Collins.

Changizi, M. A., McDannald, M. A., & Widders, D. (2002). Scaling of differentiation in networks: Nervous systems, organisms, ant colonies, ecosystems, businesses, universities, cities, electronic circuits, and Legos. *Journal of Theoretical Biology, 218*(2), 215-237.

Eisenhardt, K. M., & Santos, F. M. (2005). Organizational boundaries and theories of organization. *Organization Science, 16*(5), 491-508.

European Commission. (2004). *European interoperability framework for Pan-European e-government services.*

European Commission. (2005). *The European e-business report: A portrait of e-business in 10 sectors of the EU economy.* Luxembourg.

Farrell, J., & Saloner, G. (1985). Standardization, compatibility, and innovation. *The RAND Journal of Economics, 16*(1), 70-83.

Flügge, B. (2006). *Interoperability roadmap based on collaboration scenarios in global trade (Version 1.3).* Unpublished manuscript.

Håkansson, H., & Snehota, I. (1993). *The content and function of business relationships.* Paper presented at the Conference on Industrial Marketing and Purchasing (IMP).

Institute for Electrical and Electronics Engineers (IEEE). (1990). *IEEE standard glossary of software engineering terminology.* Author.

Legner, C., & Wende, K. (2006). *Towards an excellence framework for business interoperability.* Paper presented at the 19th Bled eConference eValues.

Legner, C., & Wende, K. (2007). *The challenges of inter-organizational business process design: A research agenda.* Paper presented at the 15th European Conference on Information Systems (ECIS 2007).

Markus, M. L., Minton, G., Steinfield, C. W., & Wigand, R. T. (2006). Industry-wide information systems standardization as collective action: The case of the U.S. residential mortgage industry. *MIS Quarterly, 30*, 439-465.

Matutinovíc, I. (2002). Organizational patterns of economies: An ecological perspective. *Ecological Economics, 40*(3), 19.

Ministry of Economic Affairs and Communications. (2005). *Estonian IT interoperability framework.* Retrieved September 25, 2007, from http://www.riso.ee/en/files/framework_2005.pdf

Reimers, K., & Li, M. (2005). Antecedents of a transaction cost theory of vertical IS standardization processes. *Electronic Markets, 15*(4), 301-312.

Ritter, T., Wilkinson, I. F., & Johnston, W. J. (2004). Managing in complex business networks. *Industrial Marketing Management, 33*, 175-183.

Stuhec, G. (2005). *How to solve the business standards dilemma: The context driven business exchange.* Retrieved April 30, 2007, from https://www.sdn.sap.com/irj/servlet/prt/portal/prtroot/docs/library/uuid/a6c5dce6-0701-0010-45b9-f6ca8c0c6474

Tansley, A. G. (1935). The use and abuse of vegetational concepts and terms. *Ecology, 16*(3), 284-307.

Theling, T., Zwicker, J., Loos, P., & Vanderhaeghen, D. (2005). *An architecture for collaborative scenarios applying a common BPMN-repository* (Vol. 3543). Heidelberg, Germany: Springer Verlag.

United Nations. (2005). *Global e-government readiness report 2005: From e-government to e-inclusion* (No. UNPAN/2005/14). Department of Economic and Social Affairs Division for Public Administration and Development Management.

United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT). (2006a). *UMM meta model: Foundation module V1.0.*

United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT). (2006b). *XML naming and design rules: Version 2.0.* Retrieved May 3, 2007, from http://www.unece.org/cefact/xml/XML-Naming-and-Design-Rules-V2.0.pdf

KEY TERMS

Core Component Technical Specification (CCTS): CCTS represents a methodology for a semantically unambiguous definition of business information based on syntax-neutral and technology-independent building blocks that can be used for (semantic) data modeling. Therefore, it facilitates the reuse of existing data entities, increases semantic interoperability, and allows for an integration of vertical industry standards.

Cross-Organizational Processes: These are so-called public processes that are relevant to any business and governmental partner in a given ecosystem. Public processes reflect the common process elements that need to be visible, achievable, and executable by the participants of the ecosystem.

Ecosystem: The ecosystem is the real-life environment business and governmental partners form to interact, share, and execute goods, products, and services relevant to a common business purpose.

Interoperability: Interoperability is the capability to exchange and reuse information, messages, and documents between applications and business partners. As a multifaceted concept, it possesses three different dimensions: technical, semantic, and organizational interoperability.

Interoperability Factor: It is the maximum grade of applying the interoperable layers; refer to the term *interoperability*.

Interoperability Road Map: The interoperability road map is a tool set enabling companies and ecosystems to define a common denominator for interoperable solutions.

Semantic Interoperability: Semantic interoperability denotes the ability of different applications and business partners to understand exchanged data in a similar way, implying a precise and unambiguous meaning of the exchanged information.

United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT): The UN/CEFACT, as part of the United Nations Economic Commission for Europe (UNECE), aims at facilitating international transactions by simplifying and harmonizing the electronic exchange of information. The UN/CEFACT has long-lasting experience in developing e-business standards, amongst others ebXML and UN/EDIFACT.

ENDNOTE

[1] According to the Institute for Electrical and Electronics Engineers (IEEE, 1990), interoperability can be defined as "the ability of two or more systems or components to exchange information and to use the information that has been exchanged."

Characteristics and Technologies of Advanced CNC Systems

M. Minhat
The University of Auckland, New Zealand

X.W. Xu
The University of Auckland, New Zealand

INTRODUCTION

Computer Numerical Control (CNC) systems are the "backbones" of modern manufacturing industry for over the last 50 years and the machine tools have evolved from simple machines with controllers that had no memory and were driven by punched tape, to today's highly sophisticated, multiprocess workstations. These CNC systems are still being worked and improved on. The key issues center on autonomous planning, decision making, process monitoring and control systems that can adjust automatically to the changeable requirements.

Introduction of CNC systems has made it possible to produce goods with consistent qualities, apart from enabling the industry to enhance productivity with a high degree of flexibility in a manufacturing system. CNC systems sit at the end of the process starting from product design using Computer Aided Design (CAD) tools to the generation of machining instructions that instruct a CNC machine to produce the final product. This process chain also includes Computer Aided Process Planning (CAPP) and Computer Aided Manufacturing (CAM).

BACKGROUND

The development of an intelligent CNC controller architecture has been one of the main goals for both CNC manufacturers and end users. This new trend comes with a suite of new technologies and concepts. In today's manufacturing world, understanding the true meaning and implication of these technologies and characteristics is the top priority. Extensibility is the ease with which a third party is able to add capabilities to software or hardware. Interoperability is the ability of two or more systems or devices to exchange information and make use of it transparently. Portability is the ease with which application software can be transferred from one environment to another. Scalability is the ease with which an existing system's performance can be increased or decreased to suit different applications of different magnitude. The recent CNC controller development includes a digital signal processor (DSP) (Chang, 2007), virtual CNC (Erkormaz & Wong, 2007) and CNC system based on STEP-NC and Function Blocks (Wang, Xu, & Tedford, 2007).

STEP-NC (ISO 14649, 2003) is viewed to be an effective means of documenting task-level information in the CAD/CAPP/CAM/CNC manufacturing chain. The new standard of Function Block (IEC 61499, 1999) has the advantages of generating method-level data for the control unit. The kernel software proposed by Park, Kim, and Cho (2006) organized and managed various control software modules dynamically by using process and resource models. It enabled the CNC controller to be easily reconfigurable because the software modules can be plugged-and-played and be built modularly so that new features or number of control axes could be added easily. Bi, Yu, and Li (2006) introduced a new type of CNC, which is based on Sinumerik 840D and STEP-NC. This is a kernel of Intelligent Integrated Manufacturing System (I^2MS).

Open CNCs

Open CNC architecture can be understood as having standard hardware and software which permit system scalability, and ensure future performance enhancement. The development of an open CNC architecture entails the establishment of a type of software architecture that fits in with a "general" computer which is independent of a control vendor, plus a communication standard among computer hardware, an operation system and application software (Ambra, Oldknow Migliorini, & Yellowley, 2002; Pritschow et al., 2001).

Most of the advanced CNC controllers and supporting hardware have closed architecture designs which make it difficult, if not impossible, to incorporate advanced control schemes within the CNC itself as well as integrate with other manufacturing resources. Open architecture controllers are designed to remove this type of obstacles by creating a flexible control system that can be attached to a wide variety of machine tool systems in such a way that the original axis and spindle drive motors and supporting electronic interfaces can remain intact.

Work concerning open CNC architecture has been one of the main topic areas in CNC research activities since the mid-1980s, but it was not until the early 1990s that a few

Table 1. Some open-architecture controllers

Vendor and location	Product	Computer(s)	Operating systems(s)	Bus and net
Advanced Technology & Research Corp Burtonsville, Md.	RCS	Dual PCs	• Windows NT for GUI • Real-time kernel for control	• Ethernet • Sercos motion bus • Profibus distributed I/O
Bridgeport Machine Inc Bridgeport, Conn.	DX-32	Dual computers	• PC/DOS for GUI • Motorola 68K/real-time kernel for control	• Ethernet • Serial channel I/O
Cimetrix Inc. Provo Utah	CX3000	Single PC	• Window NT or Lynuxs • VMEbus version available	• Ethernet • PMAC motion card • Cognex or Datacube vision cards
Delta Tau Data Systems Inc Nortridge Calif.	PMAC-NC	Single PC	• Window 95	• PMAC motion card
GE Fanuc Inc. Charlottesville, VA.	MMC-4	PC front end	-	• Through Fanuc F-bus to CNC
	HSSB	PC front end	-	• Through high speed serial bus to CNC
Hewlett-Packard Co. Santa Clara, Calif.	OAC 500	Dual PCs	• Window NT for GUI • LynxOS for control	• Ethernet • Sercos motion bus • DeviceNet distributed I/O
ICON Industrial Controls Corps. Shreveport, La	MOS	Single PC	• Window 95 with real-time kernel	• Ethernet • Sercos motion bus • DeviceNet distributed I/O
Indramat (Div. of Rexroth Corp.) Hoffman Estates, III	MTC200	CNC coprocessor	• PC bus	• Discrete I/O cards
Manufacturing Data System Inc. Anna Arbor, Mich.	Open CNC	Single PC	• QNX real-time Unix	-
Siemens AG Erlangen, Germany	Sinumerik 840D	Dual PCs	• PC/DOS for Windows for GUI • RISC or PC with real-time kernel for control	• Nurbs for control • Siemens S7-300 PLC for I/O

commercial products were investigated and prototyped by some leading companies in the CNC industry. Table 1 summarizes some of the developed or prototyped open-architecture controllers (OSACA, 2007).

World-wide, three industrial consortiums have been active since the early 1990s. They are the Open System Environment for controllers (OSE) of Japan (Zhang, Wang, & Wang, 2003), OSACA of Europe, and OMAC consortium of the USA. Although different approaches are used, they all share a similar vision of using open-architecture controllers in replacement of the current closed CNC systems.

Adaptive/Adaptable CNCs

"Adaptive" or "adaptable" describes a high level of robustness in dealing with changes in a manufacturing environment such as machining process in a timely fashion. Both terms have been used to refer to similar situations. However, differences between "adaptive" and "adaptable" have been observed in that adaptable is often used to describe the science of a system, whereas adaptive is usually used to describe the enabling technologies for such a system. In other words, adaptable is at the system level, whereas adaptive is at the component level.

CNC machine tools require experienced and qualified programmers to operate. More often than not, they select and set operating parameters off-line based on their experience and knowledge, hence the tendency in using extremely proprietary machining parameters. This situation requires CNC machine tools to be adaptable to any changes that may occur and make the initial settings invalid. This is deemed necessary so as to protect the cutters, workpiece and machine tools from being damaged as well as achieve a desired productivity (Qin & Park, 2005; Soliman & Ismail, 2001). Adaptability is also an indispensable feature in an unmanned machining system.

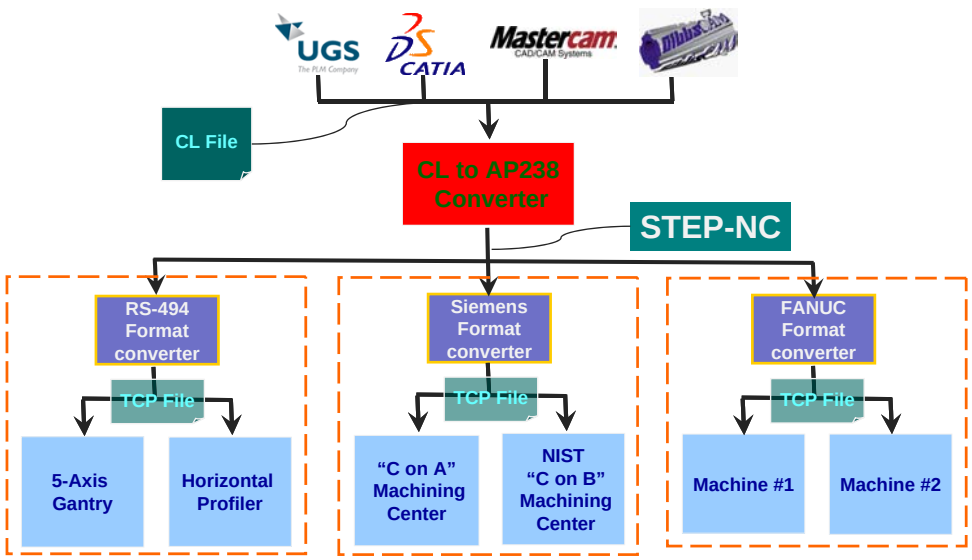
Various adaptive control technologies have been developed in the past 30 years. These include different feed-forward and feedback control mechanisms that enable automatic monitoring and adjustment of machining conditions (e.g., cutting speed, depth of cut and feed rate) in response to variations in operation performance. Various sensor-related technologies have been investigated (Liu, Zhang, & Tang, 2004; Morales-Menendez, Sheyla Aguilar, Rodriguez, Elizalde, & Garza Castanon, 2005). The adaptive control systems can be classified into two types, Adaptive Control with Constraints (ACC) and Adaptive Control with Optimization (ACO) (Cheng, Zhang, Hu, Wu, & Yang, 2001). ACC systems control and regulate one or more output parameters, typically cutting force or cutting power, to a set of limiting values by updating the machining parameters online and in real-time (Huh & Pac, 2003; Zuperl, Cus, & Milfelner, 2005). In the case of ACO, optimal goal(s) are set to be achieved by adjusting a number of controllable parameters.

More and more research effort has been spent on the development of adaptable kernel CNC software that can be easily taken up by system developers. The research work by Park et al. (2006) proposed several adaptable control architectures to enable the incorporation of new technologies into existing CNC controllers. Both kernel software and CNC control have adaptable functions. Kernel software-based controller consists of four distributed components, that is, intelligent control software modules, motion control software modules, kernel software, and a motion control card. The components can be easily replaced or adaptable to respond to any changes of the machining parameter.

Interoperable CNCs

Interoperability is often seen as an enabling technology in contemporary computer systems and their peripherals. Interoperability, in accordance with the definition proposed by the SEMI organization (Semiconductor Equipment and Material International), can be described by (i) interoperability for communication; (ii) interoperability for application services; and (iii) interoperability for interchangeability

Figure 1. Making CNC data interoperable



toward adaptability (Iung, Neunreuther, & Morel, 2001). Interoperability is now regarded as a key characteristic of the modern manufacturing system in its drive for a seamless product information flow and collaborative product development capability.

There are effectively two major issues that need to be addressed, namely product data interoperability and CNC machine tool compatibility. CNC machine tools' compatibility has been severely restricted by the way the machine control data is structured and generated. The ISO 6983 standard (i.e., G-codes) is being perceived as the contributor to the problem. This is due to the fact that the standard focuses on programming the path of the cutter centre location (CL) with respect to the machine axes. A specific post-processor has to be used for almost any individual machine tools, and of course the end result is that almost every single machine tool has to have its own specific machine control data. The consequences are the machine tools are not compatible and the entire CAD/CAPP/CAM/CNC process chain is broken.

There have been some efforts in making machine control data interchangeable, hence machine tools are made interoperable. Use of STEP-NC data model is one example. At the STEP-NC Forum in Orlando in 2005, four CAD/CAM systems (i.e., Unigraphics, Catia, GibbsCAM and MasterCAM) were used to generate CL part programs (Xu, Wang, & Rong, 2006). These CL data represent angular cutter motions in a CNC configuration-independent *I, J, K* way, with the assumption that the underlying machine tool controller will translate the *I, J, K* into machine specific five-axes angular configuration. Different CL-STEP-NC (AP-238) converters were developed to translate the CL file into a STEP-NC file (Figure 1). These converters, once embedded in the controllers, will make the STEP-NC file portable across the CNC machine tools, that is, in this particular case, the STEP-NC file has become neutral to all five-axis machines, be it a five-axis gantry CNC, "*C* on *A*" machining centre or "*C* on *B*" machining centre.

STEP-NC enabled interoperable CNC machining systems can be characterized as being capable of (i) seamless information flow; (ii) feature-based machining; (iii) autonomous CNC; (iv) fault tolerable; (v) networked (vi); scalable and (vii) portable.

Function blocks (IEC-61499) (Xu et al., 2006) are another emerging standard that can support interoperable manufacturing systems. Function block standard is being developed for distributed industrial process measurement and control systems. Because a function block can be viewed as a fundamental functional and executable unit, it is suitable to model machining information and can be designed to match individual machining features. CNC controllers can be built with function blocks as part of their device firmware or with function block libraries from which function blocks can be selected and downloaded. This renders a useful tool for developing interoperable CNC controllers and the control strategies.

Distributable CNCs

A distributable CNC system consists of several components distributed over a network of computers. These distributed components synchronize and coordinate their activities, via the communication network, to achieve a specified goal.

Figure 2. Distributed process planning

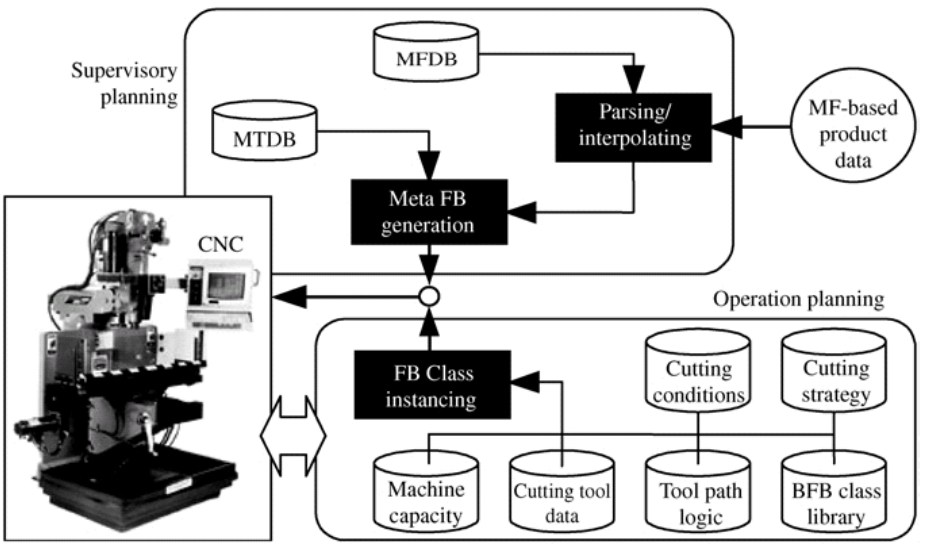
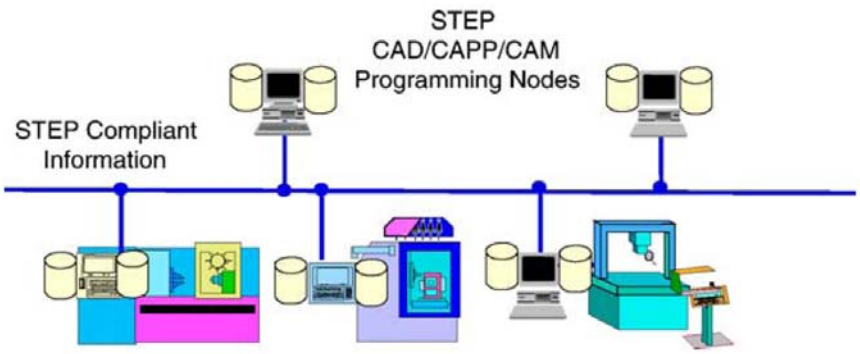


Figure 3. Distributed, STEP compliant NC machining



Such a system must be able to support an arbitrary number of processes, the distribution of which should be transparent to the user; provide an efficient communication facility; and be incorporated into a single virtual computer.

The key to a distributed CNC system is believed to be the ability of maintaining a constant availability of the functional components of the distributed operating system, for example system services and, more importantly, to improve the manageability of the distributed environment or network. The distributable functional components aim to help make the behaviour of the distributed system more predictable. Therefore, the infrastructure, that is, the set of control system services of the distributed operating system, is to be designed to include extra functionality to facilitate this task. To improve the quality of distributed software, the functional requirement of a distributed system infrastructure is needed and the problems inherent in a distributed operating system environment need to be identified.

Wang et al. (2006) proposed an architecture for Distributed Process Planning (DPP), using multi-agent negotiation and cooperation (Figure 2). It adopted technologies such as machining feature-based planning and function block-based control. Different from traditional methods, the proposed approach uses two-level decision-making, supervisory planning and operation planning. The former focuses on product data analysis, machine selection, and machining sequence planning, while the latter considers the detailed working steps of the machining operations inside each process plan and is accomplished by intelligent NC controllers. Through the nature of decentralization, the DPP shows promise of improving system performance within the continually changing shop floor environment.

Xu and Newman (2006) proposed a system model to support a scenario of "design anywhere/build anywhere" (Figure 3) (Liu, Yu, & Wang, 2005). It supports distributed manufacturing scenario through, for example, Ethernet connections to accomplish data collection, diagnostics and maintenance, monitoring and production scheduling on the same platform. This distributed feature is made possible by employing STEP as the universal data model across the CAD/CAPP/CAM/CNC process chain. Distributed CNC systems are often characterized as reconfigurable and modularized.

Reconfigurable CNCs

In general, any substantial changes that influence the production process in a manufacturing company will initiate reconfiguration of the manufacturing company's organisation. The principal causes may be a change of product demand, producing a new product on an existing system, or integrating new process technology into the existing system. In response to these changes, flexible manufacturing cell (FMC) and flexible manufacturing system (FMS) were developed and widely used.

More recently, the concept of reconfigurable manufacturing systems has been suggested and developed. Unlike FMCs and FMSs, they are capable of rapid changes in structure as well as hardware and software components in order to quickly adjust machining capacity and functionality in response to sudden changes in market or in regulatory requirements. The changing components may be machines and conveyors for entire production systems, mechanisms for individual machines, new sensors, and new controller algorithms.

A reconfigurable CNC system can be created by incorporating basic process components, hardware and software, which can be rearranged or replaced quickly and reliably to allow adding, removing, or modifying specific process capabilities, controls, software, or machine structure (Landers, Min, & Koren, 2001). This type of system will provide customized flexibility for a particular part family, and will be open-ended, so that it can be improved, upgraded, and reconfigured, rather than replaced.

The key characteristics of a reconfigurable CNC system can be summarized as follows,

- Integrability: Design systems and components for both ready integration and future introduction of new technologies. This requires a combination of hardware and software which can be rearranged in a rapid and reliable way.
- Convertibility: Allow quick changeover between existing products and quick system adaptability for future products.
- Diagnosability: Identify quickly the sources of quality and reliability problems that occur in large systems.
- Customization: Design the system capability and flexibility (hardware and controls) to match the application (product family).
- Modularity: Design all system components, both software and hardware, to be modular.

Modular CNCs

Modularity can be defined as the degree to which a product is composed of independent modules whose internal functionality is intact without any necessary interactions between them. A modular CNC system may consist of off-the-shelf units and it is nowadays considered a key attribute of any CNC systems to meet the need for mass customization. From the system point of view, modularized structure can improve its robustness, software maintenance, portability and reusability.

Although some of the modern CNC systems and the controllers are multiprocessor-based, they do not offer the flexibility to include sensing, monitoring and control modules. The type of the machine-specific components that can be modularized include, machine actuator mapping, servo control, and input/output interface. These components need to be designed independently so that they can be easily added to the controller, removed from the controller, or replaced by other components during system reconstruction. In a normal situation the integral designs result in complex parts that require more complicated and costly resources. Therefore, the modular CNC system is vital to be regarded as a strategy for effectively organizing complex products and processes. Having a modular feature allows a designer to control the degree to which changes in processes or requirements affect the product and, by promoting interchangeability, modularity gives designers more flexibility to meet these changing processes. It also allows for flexibility in function and flexibility in meeting end-user needs. In fact, modularity is a fundamental requirement for system interchangeability, openness, flexibility, scalability, portability and distributability. Different types of technologies have been deployed to achieve modularized CNC systems. Function Block is one of them. It is based on an explicit event-driven model and provides for data flow and finite-state automata-based control.

FUTURE TRENDS

Tomorrow's manufacturing world is believed to take a more distributed, decentralized and collaborated shape, the traces of which are already abundantly clear in many parts of the world. In fact, many factors help shape the future trends of advanced CNC systems due to the rapid development of manufacturing systems. The legacy of CNC control systems, or rather their development, has had a prolonged influence on today's CNC machine tools. Each CNC control system will not have its own software and hardware structure that is different from others, or else much of it will remain as a "black box" to the outside users.

Being proprietary in nature, these conventional CNC systems are limited in flexibility and robustness when there are needs to interface them with other systems. To unlock the unfulfilled potential, CNC machine tools will become more open, adaptable, interoperable, distributable, reconfigurable and modularized. The next generation of CNC systems can increase their capabilities and lower the cost by closing the loop of CAD/CAPP/CAM/CNC process chain.

CONCLUSION

Open CNCs have been extensively researched. Thank to the advances in modern control theories, adaptive control functions of these machine tools have been significantly improved. Interoperability is defined as the ability of two or more CNC systems to exchange information with no barriers. This is best described as having plug-and-play features (Wang et al., 2006). A key enabler for such a plug-and-play feature is portability, which is the ease with which machine control data can be transferred from one environment to another, or simply reused on a different machine. Distributability is needed to meet collaborative product development environment where machine control data have to be made transferable across the Internet or Intranet. Reconfigurability and modularity go hand in hand to give CNC machine tools additional features such extensibility and scalability, where extensibility is defined as the ease with which a third party is able to add capabilities to software or hardware, and scalability is the ease with which an existing system's performance can be adjusted to suit production demand.

STEP-NC and function blocks are two emerging standards that seem to hold the key for developing a generation of CNC system that is open, distributable, reconfigurable and modularized in structure, and adaptable and interoperable in functions. To a varying degree, both standards though unrelated, seem to complement each other. STEP-NC completes the entire product development chain by offering a STEP-compliant link between CAPP and CNC. It can therefore supply CNCs with a complete product model. STEP-NC is good at

supporting bi-directional information flow in CAD/CAM, data sharing over Internet, use of feature-based machining concept, modularity and reusability, and portability among resources. Function blocks, on the other hand, provide a useful tool for developing interoperable and modular CNC controllers and the control strategies. This is because a function block can be viewed as a fundamental functional and executable unit. Function blocks can be designed to match individual machining features and have needed algorithms and data embedded to decide the best cutting conditions and tool path once a machine and tool are selected. This makes function blocks best suited for CNC controls. In short, STEP-NC can be viewed as a "job-setter" (providing all the necessary information), whereas function blocks can be viewed as a "job-doer" (executing the machining commands) for a new generation of CNC systems.

REFERENCES

Ambra, C., Oldknow, K., Migliorini, G., & Yellowley, I. (2002). The design of a high performance modular CNC system architecture. In *Proceedings of the IEEE International Symposium on Intelligent Control*, (pp. 290-296).

Bi, J., Yu, T., & Li, Q. (2006). Special CNC based on advanced controller. *IFIP international federation for information processing* (Vol. 207, pp. 685-690).

Chang, Y.-F. (2007). DSP-based ignition delay monitor and control of an electro-discharge machining process. *Intelligent Automation and Soft Computing, 13*(2), 139-151.

Cheng, T., Zhang, J., Hu, C., Wu, B., & Yang, S. (2001). Intelligent machine tools in a distributed network manufacturing mode environment. *International Journal of Advanced Manufacturing Technology, 17*(3), 221-232.

Erkorkmaz, K., & Wong, W. (2007). Rapid identification technique for virtual CNC drives. *International Journal of Machine Tools and Manufacture, 47*(9), 1381-1392.

Huh, K., & Pak, C. (2003). Unmanned turning force control based on the spindle drive characteristics. *JSME International Journal, Series C: Mechanical Systems, Machine Elements and Manufacturing, 46*(1), 314-321.

IEC 61499. (1999). International Electrotechnical Commission, IEC TC65/WG6: 1999. *Function blocks for industrial process measurement and control systems, part 1: Architecture.*

ISO 14649-1. (2003). International Organization for Standardization, ISO 14649-1: 2003. *Data model for computerized numerical controllers, part 1: Overview and fundamental principles.*

Iung, B., Neunreuther, E., & Morel, G. (2001). Engineering process of integrated- distributed shop floor architecture based on interoperable field components. *International Journal of Computer Integrated Manufacturing, 14*(3), 246-262.

Landers, R. G., Min, B.-K., & Koren, Y. (2001). Reconfigurable machine tools. *CIRP Annals-Manufacturing Technology, 50*(1), 269-274.

Liu, T., Yu, T., & Wang, W. (2005). INC: A new type of computer numerical control. In *Proceedings of the 9th International Conference on Computer Supported Cooperative Work in Design*, (pp. 787-792).

Liu, Z. Q., Zhang, J. B., & Tang, Z. T. (2004). Intelligent error compensation in CNC machining through synergistic interactions among modeling, sensing and learning. In *Proceedings of the Materials Science Forum*, (pp. 178-182).

Morales-Menendez, R., Sheyla Aguilar, M., Rodriguez, G. A., Elizalde, F. G., & Garza Castanon, L. E. (2005). Sensor-fusion system for monitoring a CNC-milling centre. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 1164-1174).

OSACA. (2007). Retrieved May 28, 2008, from http://www.osaca.org

Park, S., Kim, S.-H., & Cho, H. (2006). Kernel software for efficiently building, re-configuring, and distributing an open CNC controller. *International Journal of Advanced Manufacturing Technology, 27*(7-8), 788-796.

Pritschow, G., Altintas, Y., Jovane, F., Koren, Y., Mitsuishi, M., Takata, S., et al. (2001). Open controller architecture—past, present and future. *CIRP Annals-Manufacturing Technology, 50*(2), 463-470.

Qin, Y., & Park, S. S. (2005). Robust adaptive control of machining operations. In *Proceedings of the IEEE International Conference on Mechatronics and Automation, ICMA 2005*, (pp. 975-979).

Soliman, E., & Ismail, F. (2001). A proposed adaptive current controller for CNC-milling systems. *AEJ-Alexandria Engineering Journal, 40*(3), 325-334.

Wang, L., Shen, W., & Hao, Q. (2006). An overview of distributed process planning and its integration with scheduling. *International Journal of Computer Applications in Technology, 26*(1-2), 3-14.

Wang, H., Xu, X. W., & Tedford, J. D. (2006). Making a process plan adaptable to CNCs. *International Journal of Computer Applications in Technology, 26*(1-2), 49-58.

Wang, H., Xu, X., & Tedford, J. D. (2007). An adaptable CNC system based on STEP-NC and function blocks. *International Journal of Production Research, 45*(17), 3809-3829.

Xu, X. W., & Newman, S. T. (2006). Making CNC machine tools more open, interoperable and intelligent—a review of the technologies. *Computers in Industry, 57*(2), 141-152.

Xu, X. W., Wang, L., & Rong, Y. (2006). STEP-NC and function blocks for interoperable manufacturing. *IEEE Transactions on Automation Science and Engineering, 3*(3), 297-307.

Zhang, C., Wang, H., & Wang, J. (2003). An USB-based software CNC system. *Journal of Materials Processing Technology, 139*(1-3 SPEC), 286-290.

Zuperl, U., Cus, F., & Milfelner, M. (2005). Fuzzy control strategy for an adaptive force control in end-milling. *Journal of Materials Processing Technology, 164-165*, 1472-1478.

KEY TERMS

Adaptable: Describes a high level of robustness in dealing with control system or machine's components changes in a manufacturing environment especially during machining process for future unmanned machining system.

Distributable: Consists of several components distributed over a network of computers involving Computer Aided Design (CAD), Computer Aided Process Planning (CAPP), Computer Aided Manufacturing (CAM) and Computer Numerical Control (CNC) try to synchronize and coordinate machining activities.

Extensibility: The facility by which a third party is able to add capabilities to software or hardware.

Interoperable: The ability of two or more computer systems to exchange information and make use of it transparently.

Interoperability: The ability of two or more computer systems to exchange information with no barriers and make use of it transparently.

Modular: The degree to which a product of CNC system components is composed of independent modules whose internal functionality is intact without any necessary interactions between them.

Open: Having standard hardware and software which permit system scalability, and ensure future performance enhancement; a vendor-neutrality and component-integrability; tool-neutral, and controller-neutral architecture.

Portability: The facility which applications software can be transferred to one environment from another, while maintaining its capabilities.

Reconfigurable: Is composed of multiple software and hardware, and incorporates basic process components which can be replaced, permitting disassembly, reassembly, rearrangement, or replaced quickly to the main system to allow adding, removing, or modifying specific process capabilities, controls, software, or machine structure.

Scalability: The facility which an existing system's performance can be increased or decreased in the application demand.

Chief Knowledge Officers

Richard T. Herschel
St. Joseph's University, USA

INTRODUCTION

Knowledge management (KM) refers to a range of practices used by organizations to identify, create, represent, and distribute knowledge for reuse, awareness, and learning across the organization. KM typically takes the form of programs that are tied to organizational objectives and are intended to lead to the achievement of specific outcomes such as shared intelligence, improved performance, competitive advantage, or higher levels of innovation.

Knowledge management focuses on developing and maintaining intellectual capital across the organization. It attempts to bring under one set of practices various strands of thought and practice relating to:

- Harnessing the effective use of data, information, and know-how in a knowledge-based organization and economy
- The idea of the learning organization
- Various enabling organizational practices such as communities of practice and corporate yellow page directories for accessing key personnel and expertise
- Various enabling technologies such as knowledge bases and expert systems, help desks, corporate intranets and extranets, and content management systems (Wikipedia, 2007).

Beginning in the 1990s, the person responsible for directing and coordinating these activities for organizations was oftentimes designated the chief knowledge office (CKO).

BACKGROUND

The role of a CKO was created and promoted by consultants in the late 1990s to develop a firm's knowledge infrastructure, to promote knowledge capture, storage, and distribution, and to act as a symbol that employees look to for guidance in a knowledge management culture. Bontis (2002) states that the CKO position was intended to help a firm to leverage its intellectual capital by:

- Promoting stability in a turbulent business environment
- Enabling the speedy delivery of products and services
- Creating high efficiency in the knowledge value chain by sharing of resources and realization of synergies
- Enabling the separation of work so that specialization is feasible

The CKO job description frequently encompassed a number of different responsibilities. For example, the CKO might be responsible for leading executive management to develop an enterprise knowledge strategy, validating this strategy across the enterprise, and then ensuring that its evolution complements and integrates with business strategy. The CKO might also be charged with setting priorities and securing funding for knowledge management programs as well as defining policies for security, usage, and maintenance of intellectual capital. Depending on the organizational culture, the CKO could also act as the chief advocate for KM as a discipline—walking and talking the program throughout the enterprise and assisting executives and senior management in building and communicating personal commitment and advocacy for KM (Davenport & Prusak, 1998).

Rarely did the CKO come from an information systems or human resource organization. In fact, CKO backgrounds were quite varied, though most had substantial experience with their firm and knowledge of the firm's industry. Whatever their background, CKOs were supposed to straddle business and information technology (IT) with a mandate to convince workers that it is good to share information and to work with IT to build applications to support such sharing (Earl & Scott, 1999).

In 2001, 25% of Fortune 500 companies had a CKO and 80% of Fortune 500 companies had a knowledge management staff. Forty-two percent of Fortune 500 companies anticipated appointing a CKO within the next three years (Flash, 2001).

While many organizations were enthusiastic about knowledge management programs, there were also firms that believed that a CKO function was not needed. Sometimes senior management felt that having a CKO was the wrong way to harness corporate know-how. Instead, they preferred a more grassroots approach, in which a team of knowledge management experts worked closely with--or were even as part of--the business units. The underlying rationale for this approach lay in the belief that by putting more control of knowledge management in the hands of end users, knowledge management would be an easier sell because knowledge sharing would be actively inculcated within business units.

Accordingly, these firms believed that centralizing knowledge management under a CKO would send out the wrong message (Cole-Gomolski, 1999).

In firms where CKOs did exist, Pringle (2003) notes that many of these survived by judiciously distancing themselves from the original "craze" while still exploiting knowledge management concepts. This oftentimes meant that CKOs didn't talk about knowledge management per se. Instead, the CKO pursued activities that encouraged employees to talk to one another or that allowed workers to reuse already existing materials or information. Pringle indicates that these CKOs typically imbedded knowledge management activities within performance management systems that gave staff members the incentive to learn and to share their expertise. That is, assessments of employee sharing of information efforts as well as demonstrable learning activities became an integral part of employee annual performance reviews.

CURRENT TRENDS

Quite a number of books and articles about knowledge management and CKOs were published in the late 1990s and the early 2000s. However, in 2007, while knowledge management is a concept still practiced and written about, new articles about the CKO entity are rare. The reason for this may be that either the desire for such a position with this title has diminished or that the knowledge management environment has evolved such that the need for a figurehead or program leader has been reduced.

Boothby (2007) believes that the CKO position itself still exists, but he also believes that the responsibilities of the position have changed. He states, that while the old definition of a CKO's job was to guide knowledge management, the new definition of the CKO's job is to empower knowledge workers. He argues that empowering workers means giving knowledge workers tools that make them more productive, which is operationalized as helping knowledge workers to communicate more effectively.

Boothby asserts that CKOs used to hire people who categorized everything and wrote complex taxonomies to organize knowledge. Moreover, this "traditional" CKO employed large, complex systems to create, capture, store, and distribute knowledge. That is, they looked for a standard approach that would satisfy the needs of their whole company.

The problem for the CKO today, according to Boothby, is that such an approach is not viable anymore. Needing to find one universal solution is a false constraint he asserts. He argues that an open Internet works just fine with multiple blogging systems, wiki systems, and open source programs and operating systems. In the current technology environment, large companies do not need one universal enterprise solution. Instead, Boothby states, large organizations probably need many different tools for different types of users and different types of problems.

In the past, knowledge management technology was oftentimes expensive, centralized, and coordinated. Today, knowledge management technology can be inexpensive, decentralized, and perfuse. Boothby notes that knowledge management technology can cost less than a tenth of the price of old systems. Moreover, with many systems today, users can generate their own software content (e.g., via Linux, blogs, wikis, etc). Boothby concludes that CKOs should stop focusing on what is ideal and allow any system, so long as it complies with some basic open Internet standards.

Boothby's arguments do assume that CKOs only seek technology solutions to knowledge management efforts. This is only partially true. In many organizations, CKOs are responsible for the sharing of both explicit and tacit knowledge and technology typically only addresses the former well.

However, Boothby's assertion that knowledge management technology is evolving appears affirmed by an Executive Report of the 2006 CKO Summit held at the Bath Priory in the United Kingdom (TFPL, 2007). Here, social computing, identified as blogs and wikis, was seen as the backbone of current knowledge sharing efforts. These technologies, combined with the use of search engines and document management systems, were seen by Summit participants as facilitating and diffusing knowledge transfer capabilities and better enabling knowledge harvesting. Moreover, social computing technologies were viewed as providing common and standardized information architectures for knowledge management programs, resulting in more active knowledge sharing activity.

The 2006 CKO Summit report also suggests that previously prescribed knowledge management leadership responsibilities have remained somewhat constant over time. The report indicates that the current issue for managers of knowledge management programs is to articulate the common framework for knowledge and information management for their organization. Articulated components of this framework include:

- Mission, vision, and objectives for shared services
- Governance (an agreed strategy for inter-organizational KM)
- KM vision and mission
- Operating model
- Information architecture
- Metrics/performance measures
- Delivery and benefits

What the Summit's report additionally makes apparent is the changing nature of knowledge sharing itself. In the earlier days of knowledge management, emphasis was placed on the need to manage "pull" technology. That is, organization's

had to promote sharing of information on existing platforms by encouraging employees to pull down information from repositories of data and information. Now, emphasis is being place on "push" technologies that allow employees and partners the opportunity to easily and readily publish information to platforms that are comparable to those used in the public domain. Hence, since sharing of explicit information is easy to do, management must now shift to helping to ensure that content generated is both valid and reliable.

Tacit knowledge, on the other hand, still appears to remain somewhat elusive in knowledge management circles. Current thinking argues that knowledge management programs should employ narratives and story telling to elicit this form of intellectual capital (e.g., Brown, Denning, Groh, & Prusak, 2005).

Remarkably, of the 28 participants at the 2006 CKO Summit (TFPL, 2007), not one held the title of CKO. In fact, only nine of the participants even had the words "knowledge management" in their job titles.

However, CKO positions do still exist. In England, for example, the CKO of the National Health Service is seen as a role and not a job, a role that has to make sure that the knowledge flows easily between all parts of the organization, thus helping it achieve it objectives more productively. Specifically, the responsibilities of the NHS CKO are:

- To coordinate the outputs of all those national bodies charged with the production of knowledge for patients and professionals
- To ensure that the knowledge derived from research is integrated with knowledge derived from the analysis of data and the knowledge derived from experience
- To ensure that all the Directorates of the Department of Health have their work supported by a National Knowledge Service
- To ensure that there is a national system for knowledge management in the mainstream of the NHS (National Knowledge Service, 2007).

Of note here is that the CKO's responsibilities make no reference to technology specifics. Instead it emphasizes knowledge coordination, sharing and integration.

Six Sigma recently hired a new CKO and their announcement also made no reference to technology issues. The responsibility of their CKO is to accelerate ongoing efforts to develop new, more efficient ways of delivering client value by pushing the envelope of business performance (Six Sigma, 2006).

FUTURE TRENDS

Is the CKO title diminishing in popularity? I suspect so. For example, at one Web site, the title of "CKO" is not even listed under knowledge management careers. However, this is not a sign that knowledge management efforts and programs are in any way declining. It is more likely that the concepts enabled by new technologies and encouraged by the more pervasive understanding and appreciation of what intellectual capital is, are now more widely understood and practiced throughout organizations. In fact, while the need for a specific recognizable symbol called the "CKO" probably has diminished, general emphasis on the importance of knowledge management tenets appears to have increased.

One indicator of this is that the number of academic programs in knowledge management is increasing. For example:

- Hong Kong Polytechnic University now offers a Master of Science and Postgraduate Diploma in Knowledge Management.
- Nanyang Technological University's School of Communication and Information in Singapore offers students a Master of Science in Knowledge Management degree.
- Kent State University in the United States promotes their certificate program in knowledge management and a Master of Science degree in Information Architecture and Knowledge Management.
- Saint Joseph's University in Philadelphia has developed three business intelligence programs, a field that is a subset of knowledge management (Herschel & Jones, 2005). These programs include a Masters in Decision & Systems Sciences, a Business Intelligence Certificate, and a Business Intelligence Minor.
- George Mason University's School of Public Policy lists a Master of Science in New Professional Studies in Knowledge Management.
- Walden University offers online Masters and Ph.D. programs that concentrate in knowledge management.

And organizations are also sponsoring in-house workshops to enhance employee understanding of knowledge principles so as to enhance organizational decision-making (e.g., Bank of England, 2006).

While none of the current educational activities emphasize or even mention the role of chief knowledge officers per se, their emergence indicates an increased public awareness of and sensitivity to the need to become more educated in knowledge management concepts, skills, and technologies.

Finally, there has been one recent study reaffirming the potential importance of CKO leadership. Chua and Lam (2005) conducted research designed to determine the effectiveness of a having a CKO. The findings confirmed that use of a CKO was a viable and effective instructional tool for imparting knowledge to study participants. Chua's and Lam's results indicate that having a CKO has a moderating

effect on study participants' attitude towards the experiment's subject matter. However, it can also be argued that it is not the existence of a CKO per se that is important, but rather having someone acting in the capacity of knowledge facilitator is key.

CONCLUSION

Today, the popularization and ease of use of e-mail, cell phones, wikis, and blogs have served to make the sharing of information a globalized phenomenon that is commonplace. And it is not just with text and voice-based systems that have become pervasive communication commodities. The creation and sharing of videos on a global platform are also becoming routine. The popularity of internet-based video platforms such as those provided at YouTube, AOL, and Yahoo have the potential to revolutionize the volume of tacit knowledge that is captured and distributed, especially since these technologies can be readily adapted to internal organizational applications. Of all the technologies used in knowledge management, video holds the most promise for enabling the efficient transformation of tacit knowledge to explicit knowledge via externalization (Herschel, Nemati, & Steiger, 2001, 2003).

In one of the first studies conducted on the CKO function, Earl et al. (1999) found that having a CKO was unlikely to be absolutely necessary for organizations to realize an effective knowledge management program: nor was it likely to be universally necessary. They conceded that some organizations would invest in aspects of knowledge management without appointing a CKO. They believed that knowledge management, like total quality management, would become embedded in organizations and that knowledge would become an obvious imperative source of value creation and competitiveness. In this scenario, they predicted that all members of the organization would own and drive knowledge management and, hence, the importance of and need for a CKO would decline. It seems that this prediction may be coming true.

However, while knowledge management technology, education, and concept awareness have generally increased, the need for organizational leadership remains important and constant. Appointing a CKO may no longer be seen as a prerequisite for galvanizing, directing, and coordinating knowledge management programs, but the need for knowledge management coordination probably remains viable and even essential. Intellectual capital, like any other form of capital, should be effectively managed.

REFERENCES

Bank of England. (2006). *Workshop: Enhancing a central bank's effectiveness through knowledge man*. Retrieved from http://www.bankofengland.co.uk/education/ccbs/courses/course25.htm

Bontis, N. (2002). The rising star of the chief knowledge officer. *Ivey Business Journal*, 20-25.

Boothby, R. (2007). *The chief knowledge officer's dilemma*. Retrieved from http://www.innovationcreators.com/2006/07/the_chief_knowledge_officers_d.html

Brown, J., Denning, S., Groh, K., & Prusak, L. (2005). *Storytelling in organizations: Why storytelling is transforming 21st century organizations and management*. Oxford, UK: Elsevier.

Chua, A., & Lam, W. (2005). Why KM projects fail: A multi-case analysis. *Journal of Knowledge Management, 9*(3), 6-17.

Cole-Gomolski, B. (1999). Knowledge "czars" fall from grace. *Computerworld, 33*(1), 13.

Davenport, T. H., & Prusak, L., (1998). *Working knowledge: How organizations manage what they know*. Cambridge: MA: Harvard Business School Press.

Earl. M., & Scott, I. (1999). Opinion: What is a chief knowledge officer? *Sloan Management Review, 40*(2), 29-38.

Flash, C. (2001). Who is the CKO?, Knowledge Management Magazine. February 20. Retrieved from http://kmleaders.com/KM%20Roles%20and %20Technological%20Issues%20_final_.pdf

Herschel, R., and Jones, N. (2005). KM & business intelligence: The importance of integration. *Journal of Knowledge Management, 9*(4), 45-55.

Herschel, R., & Nemati, H. (2001). Chief knowledge officers: Managing knowledge for organizational effectiveness. In Y. Malhorta (Ed.), *Knowledge management and business model innovation*, (pp. 414-425). Idea Group Publishing.

Herschel, R., Nemati, H., & Steiger, D. (2001). Tacit to explicit knowledge conversion: Knowledge exchange protocols. *Journal of Knowledge Management, 5*(1), 107-116.

Herschel, R., Nemati, H., & Steiger, D. (2003). Knowledge exchange protocols: A second study. *Journal of Information and Knowledge Management: Special issue of JIKMS on Knowledge Management in Context and Context for Knowledge Management, 2*(2), 153-163.

National Knowledge Service. (2007). *Chief knowledge officer*. Retrieved from http://www.nks.nhs.uk/nks_cko.asp

Pringle, D. (2003, January 7). Chief knowledge officers adapt to remain relevant, employed. *Wall Street Journal, 1*.

Six Sigma. (2006, October 17). *Mikel Harry to rejoin SSA as vice-chairman and chief knowledge officer*. Retrieved from http://www.isixsigma.com/library/content/n061017b.asp

TFPL. (2007). *CKO summits*. Retrieved from http://www.tfpl.com/thought_leadership/cko_summits.cfm

Wikipedia. (2007). *Chief knowledge officer*. Retrieved from http://en.wikipedia.org/wiki/Chief_knowledge_officer

KEY TERMS

Business Intelligence (BI): A business management term, which refers to applications and technologies which are used to gather, provide access to, and analyze data and information about their company operations.

Chief Knowledge Officer (CKO): A senior level executive responsible for managing a firm's knowledge management initiative.

Explicit Knowledge: Knowledge that can be expressed formally using a system of symbols, and can therefore be easily communicated or diffused. It is either object based or rule based.

Intellectual Capital: Can be divided into three categories:

- **Human Capital:** That in the minds of individuals: knowledge, competences, experience, know-how etc.
- **Structural Capital:** "That which is left after employees go home for the night": processes, information systems, databases etc.
- **Customer Capital:** Customer relationships, brands, trademarks, etc.

Knowledge Management: A program for managing a firm's intellectual capital by systematically capturing, storing, sharing, and disseminating the firm's explicit and tacit knowledge.

Knowledge Management Architecture: The technology and procedural platforms employed by a firm to enable knowledge capture, sharing, retention, and distribution.

Tacit Knowledge: Knowledge that is uncodified and difficult to diffuse. It is hard to verbalize because it is expressed through action-based skills and cannot be reduced to rules and recipes. Tacit knowledge is the same as implicit knowledge.

Category: Business Information Systems

A Classical Uncertainty Principle for Organizations

Joseph Wood
U.S. Army, USA

Hui-Lien Tung
Paine College, USA

James Grayson
Augusta State University, USA

Christian Poppeliers
Augusta State University, USA

W.F. Lawless
Paine College, USA

INTRODUCTION

After this article introduction, we review the prevailing theory of organizations, and what it means to organizational science and the new discipline of Quantum Interaction to have an uncertainty principle (ir.dcs.gla.ac.uk/qi2008; the corresponding author is one of the organizers). Further into the background, we review control theory for organizations and its importance to machine and human agents; we review the hypothesis for the uncertainty principle; and we review the status of the field and laboratory evidence so far collected to establish the uncertainty principle for organizations. Then we review future trends and provide the conclusion.

BACKGROUND

At the first Quantum Interaction conference, held at Stanford University in the spring of 2007, a panel addressed whether QI was relegated to being a metaphor or whether it could function as a working model that could be applied in an agent-based model to solve social problems like organizational decision making. Of the 24 papers presented at this inaugural conference, few put forth a working model with sufficient details to be falsified. We accept the challenge by proposing in this review a path forward to a working model.

Rieffel (2007) suggested that few advantages accrue from claiming that the quantum model is applicable to the social interaction when it is not, and few disadvantages from applying an uncertainty principle to demonstrate classical tradeoffs, as in the case of signal detection theory, or to demonstrate nonseparability when the tensor calculus fails to hold. In response, the model should lay the groundwork to demonstrate classical effects of the uncertainty principle for organizations.

As an example from common experience, movie entrepreneurs manipulate individuals en masse with entertainment exchanged for payment, as in the joint viewing of a Clint Eastwood movie where individual brains have been found to "tick collectively" (Hasson, Nir, Levy, Fuhrmann, & Malach, 2004). For organizational tradeoffs, the uncertainty principle means that under interdependence, the probability of applying sufficient attention to a plan or to execute it shifts uncertainty in an opposing direction, and vice versa, *iff* the state of interdependence continues (Note: the symbol *iff* means "if and only if").

The interdependent tradeoffs to control a system requires channels that enhance the ability of management to diminish the destructive interference from inside or outside of an organization. It means that tradeoffs form cross-sections that reflect defensive and offensive maneuvers to expand or limit the size of an organization. Tradeoffs mean that as perspectives shift, what is observed to change in an organization also shifts (Weick & Quinn, 1999); that illusions are fundamental to organizational hierarchies (Pfeffer & Fong, 2005) by driving or dampening feedback oscillations (Lawless, Whitton, & Poppeliers, 2008); and that tradeoffs explain why criteria for organizational performance has been intractable (Kohli & Hoadley, 2006).

We define illusions not as false realities, but as bistable interpretations of the same reality that can only be held simultaneously by neutrals while "true believers" drive neutrals to weigh one and then its opposing reality, for example, an ideology of nuclear waste cleanup or the concrete steps needed for cleanup. Single ideological views are usually

driven by strong-minded agents who we represent as forcing functions, *f(t)*, where the valence of each marginal element of fact they present to neutrals is represented by one bit of additional information. Illusions entangle only neutral agents not wedded to either competing view, where the valence of both views is represented by two bits of entangled information. Courting neutrals to decide outcomes moderates the heated debates between opposing drivers; when neutrals abandon the decision process, it becomes volatile and unstable (Kirk, 2003). Tradeoffs can reduce the effect of illusions by decreasing the volatility in organizational performance that produces "gridlock" (Lawless et al., 2008).

We define social influence as a form of social entanglement, which means that entangled elements can be manipulated together (von Bayer, 2004). Per Rieffel, a state $|\psi>$ is entangled when it cannot be written as the tensor product of single qubit states (p. 139). Here, we define interdependence from social influence as operating across neutral individuals as a superposition of waveforms composed of two or more simultaneous values that linearly combine under constructive interference such as rationalizing similar views into a single world view, or under destructive interference to disambiguate dissimilar views into the best concrete plan. Both interdependence and entanglement are fragile, do not always produce uniform effects, and experience rapid decay; the greater the clarity of an interdependent social situation (observation), the greater the uncertainty in the effect of social influence (action).

Establishing the uncertainty principle for organizational tradeoffs is not only important to move beyond the "quantum" as metaphor, but also because organizational theory has not progressed much beyond Lewin. Lewin himself has been blamed for putting too much attention on individual differences rather than an understanding of groups (Moreland, 2006), which remains elusive (Levine & Moreland, 1998). Instead of blaming Lewin, we attribute the problem to the recondite nature of tradeoffs; the greater the clarity of an interdependent social situation (observation), the greater the uncertainty in the effects from social influence (action).

MAIN FOCUS OF THE CHAPTER

Tradeoffs are inherent in the interdependence that exists in knowledge *iff* interdependence is nonseparable either at the level of information sources (e.g., the interdependence between static and dynamic visual perception; in Gibson, 1986), interdependent uncertainties, or interdependent contexts for decision-making (e.g., hierarchical framing effects). Organizations exist in states of interdependence (Romanelli & Tushman, 1994), characterized as a whole being different from the sum of its parts (Lewin, 1951).

Two of the goals for organizational science are to increase knowledge and to reflect associated uncertainties. A current goal of social science is to simulate human cognition. A unique contribution to these goals is to extend human cognitive simulation with a mathematical model of an organization(s) set within a system operating on knowledge interdependent with uncertainty. The ultimate goal is to design the control of a system of future human and artificial agents (in the military, warfighters and mobile machines advanced beyond present sensors, platforms like Predator-Global Hawk, and robots), or mixtures of both, but *iff* they are interdependent deciders operating under uncertainty. The system model can be used to study human organizations making decisions in marginal situations like mergers to address complex tasks under uncertainty. The primary characteristic of this interdependence is reflected in tradeoffs between coordinating social objects communicating to solve problems in states of uncertainty (Lawless & Grayson, 2004).

Mergers seem unlikely as a model because the explanations for mergers are controversial (Andrade & Stafford, 1999). Most researchers believe that mergers are a bad choice for a firm to consider because they often fail (e.g., Daimler merged with Chrysler in 1998 for $36 billion, only to sell it in 2007 for $7 billion). But mergers have been found to increase efficiency and market power in response to unexpected market shocks (Andrade, Mitchell, & Stafford, 2001). To protect against shocks, we have found that successful mergers, like SBC's with AT&T and Bellsouth increase stability (see below).

Mergers exemplify tradeoffs in nonseparable interdependent knowledge. However, mergers form forced cooperative systems that reduce internal and external information, a censorship that stabilizes systems, compared to the disambiguation and volatility under competition so easily observed by outsiders (Lawless & Grayson, 2004).

As an extreme tradeoff, organizations under central, command-driven or authoritarian leadership easily exploit consensus-seeking rules for decision making (Kruglanski, Pierro, Mannetti, & De Grada, 2006). Censorship under dictatorships reduces socio-political volatility in exchange for rigid control (May, 1973). Recent examples of censorship are found in news accounts of Myanmar's denials of village purges (Bhattacharjee, 2007); China's imprisonment of journalists; and Russian censorship of TV commentators. Censorship occurs in organizations within democracies, too; but when censored information is released, its volatility often forces attention to address the consequences (e.g., Sen, 2000, concluded that no modern democracy has ever suffered from famine).

Whether cooperation or competition increases social or individual welfare during decision making is the canonical tradeoff. Enforced consensus-seeking actions are predicated on a consensus world view, making knowledge more easily acquired *iff* the courses of action conform to a chosen world view, making them impractical for all actions except simple ones. In contrast, focusing on practical applications

under competition and uncertainty disambiguates complex actions for robots (Bongard, Zykov, & Lipson, 2006) and humans (Lawless et al., 2008); however, because it is driven by a polarization of at least two opposing viewpoints (bi-stable illusion), disambiguation less readily transforms into knowledge. These results have implications for the limits of knowledge.

As examples of *separable*, noninterdependent information, a meta-analysis of 30 years of research on self-esteem and academic or work performance, one of the most studied topics in social psychology, Baumeister, Campbell, Krueger, and Vohs (2005) found only a minimal association. They countered the prevailing belief by concluding that surveys of humans produced surprisingly limited information. Similarly, in a study with multiple regressions of USAF air-combat maneuvering (ACM) attempting to affirm the proposition that ACM educational courses improved air combat outcomes in machine space, we found no association (Lawless, Castelao, & Ballas, 2000). In the ACM study, we concluded that current machine "god's-eye-views" were limited to separable information. A god's-eye-view describes the situation where perfect information exists regarding the interactions occurring in machine space among artificial agents, humans or both (e.g., a computer's perfect access to the information produced by Swarm).

One well-used approach is to use game theory to model interdependence, its strength. But the weakness of game theory, including the quantum version by Eisert, Wilkens, and Lewenstein (1999), is the arbitrary value it assigns to cooperation and competition, exacerbated when the number of players is greater than two in complex and uncertain contexts (social vs. individual welfare). Kelley (1992) spent his career admittedly failing to link the expectations of subjects to the choices they later made while playing Prisoner Dilemma Games, only to conclude that individual subjects rarely acted as he or they expected once in a group. Game theory can be improved upon with an interdependent model of uncertainty determined by outcomes. For example, we have found that competition improves social welfare with practical decisions that feedback readily accelerates (Lawless, Bergman, & Feltovich, 2005). In contrast, we have found that gridlock is more likely under cooperative decision making because it is less able to challenge illusions (Lawless et al., 2008).

In the search for a classical organizational uncertainty principle, we have found in the field and confirmed in the laboratory a planning cognitive-execution tradeoff between consensus-seeking and majority rule decision making as citizen groups made decisions over complex issues like nuclear waste management (Lawless et al., 2005). In the field study, we looked at the decisions of all nine of the Department of Energy's Citizen Advisory Boards as they responded to DOE's formal request to support DOE's plans to speed the shipments of transuranic wastes to its repository in New Mexico (i.e., the WIPP facility; see www.wipp.energy.gov) as part of its mission to accelerate the cleanup of DOE facilities across the U.S. These nine boards were located at the DOE sites where the transuranic wastes were being removed and shipped to WIPP. DOE's plans were entailed in 13 concrete recommendations and explained to the various boards by DOE engineers (e.g., "DOE, in consultation with stakeholders and regulators, reexamine the categorization of TRU [transuranic] waste using a risk-based approach"). Consequently, 4/5 of DOE's majority-rule boards endorsed these recommendations, while ¾ of its consensus-ruled boards rejected them. In addition, the time spent in deciding for majority-ruled boards was about 1/4th the amount of time taken by the consensus-ruled boards. In a follow-on field study of consensus decisions by the Hanford Board and majority rule decisions at the Savannah River Site Board, we also found that consensus rule decisions produced a volatility that resulted in "gridlock" when the single world view of the board conflicted with DOE's vision (Lawless et al., 2008).

The results of these two field studies have many implications for advanced human-robot systems. Tradeoffs occur between cooperative and competitive approaches to decision making, suggesting an organizational uncertainty principle. Consensus-seeking is more likely to build on prior world views (e.g., situational awareness) at the expense of action; majority rule is more likely to produce practical decisions that will be enacted but under conflict, meaning that a shared world view is less likely to be retrieved. And consensus-seeking is more rational, but ill-suited to govern illusions, the very rule used by the DOE Boards to seek consensus precluded the rejection of any view no matter how bizarre (Bradbury, Branch, & Malone, 2003); majority-rule disambiguates illusions well, but is less suited to produce or unify rational perspectives (Lawless et al., 2008). But, and unexpectedly, cooperative decisions produced more anger between the DOE sponsor and DOE's Hanford Board. The probable cause was a conflict in the world views of these two organizations, with few neutrals available to resolve the conflict.

It may be that the cooperation under consensus-seeking is more volatile over the long run than the conflict generated in the short run from truth-seeking under majority rule. Stability arises from majority rule as decision drivers attempt to "entangle" neutrals into their view of reality by courting them on the rightness of their view, a confrontational approach between two drivers that dampens conflict by harnessing it to entangle neutrals into solving complex problems. The result is a tradeoff between decision processes that partly demonstrates the classical uncertainty principle for organizations.

To test whether this approach is on the right track, using multiple regressions, we have found tradeoffs among stock market sectors. In the tradeoffs among stock market sectors,

Table 1. Volatility and market leadership. In addition to the data shown, two columns representing two factors not shown but tested in the multiple regressions were "market value/revenue" and "market value/EBITDA."

	Beta	gross profit	EBITDA	ROE	mktcap/indcap
Southwest Air	0.32	2.78	1.36	7.97	0.244
Walmart	0.08	84.5	26.6	8.84	0.657
Starwood Hotels	0.26	1.64	1.35	29.12	0.243
Clear Ch Comm	0.69	4.42	2.37	8.57	0.543
Sears	0.19	15.19	3.59	12.67	0.254
Google	0.96	6.38	5.23	24.4	0.716
Suez	0.97	18.79	9.39	20.17	0.243
Boeing	0.71	11.09	6.27	27.87	0.739
Lockheed	0.19	3.93	4.62	35.12	0.181
Berkshire Hatha\	0.22	47.93	19.8	11.01	0.192
GE	0.44	89.3	34.6	18.94	0.597
United Air	3.49	2.72	1.47	4.29	0.097
99 Cents Stores	1.31	0.383	0.041	1.81	0.003
Siriius Sat	3.36	-0.061	-0.461	-21	0.119
Talbots	1.98	0.76	0.22	1.46	0.034
Gottschalks	1.96	0.241	0.029	2.82	0.002
Expedia	4.04	1.73	0.644	4.6	0.04
US Geothermal	4.66	-0.676	-0.001	-10.1	0.0003
Sequa	1.82	0.379	0.233	8.8	0.001
Taser	5.69	0.043	0.13	-4.28	0.0003
Hanover	1.59	2.64	0.309	9.91	0.003
Dynasil	7.19	0.002	0.001	22.61	0.00002

multiple regressions were computed for a broad cross-section of stock market firms. Selecting data about the leader and laggard in a market where data were available (see Table 1 below), the results were significant only for volatility (Beta) and market capitalization/industry capitalization (i.e., Beta = -4.34 * mktcap/indcap + 2.88, $p < .05$). We interpret this result to have two meanings: First, it indicated that one of the reasons organizations attempt to grow, organically or via mergers and acquisitions, is that it increases organizational stability, a finding that offers a novel explanation for mergers. Second, it supported the notion that the organizational uncertainty principle exists as a tradeoff between firm size and volatility (Frieden, 2004).

FUTURE TRENDS

In the future, we expect that the study of Fourier transform pairs and measures market power should continue to advance. We also believe that Monte Carlo simulations of the cognitive tradeoffs made by managers and their control of large, complex organizations will become increasingly important. If the Monte Carlo simulations with Gaussian-Fourier pairs prove to be as successful as we believe they will become, models can be constructed by using machine language (e.g., genetic algorithms) and Agent-based Models (e.g., with agents running in NetLogo or Repast).

CONCLUSION

The result of the multiple regressions is presented as the first evidence of the possible existence of an uncertainty principle for organizations. We laid the groundwork for this principle by reviewing the importance of moving beyond metaphor in applying the quantum model to the social interaction, especially to interdependence in the organization. In the future, we plan to explore the ramifications of this principle.

REFERENCES

Andrade, G.M.-M., Mitchell, M.L., & Stafford, E. (2001). *New evidence and perspectives on mergers*. Cambridge, MA: Harvard Business School (Working Paper No. 01-070). Retrieved May 31, 2008, from http://ssrn.com/abstract=269313

Andrade, G., & Stafford, E. (1999). *Investigating the economic role of mergers* (Working Paper No. 00-006). Cambridge, MA: Harvard Business School. Retrieved May 31, 2008, from http://ssrn.com/abstract=47264

Baumeister, R.F., Campbell, J.D., Krueger, J.I., & Vohs, K.D. (2005, January). Exploding the self-esteem myth. *Scientific American*.

Bhattacharjee, Y. (2007). News focus: Byanmar's secret history exposed in satellite images. *Science, 318,* 29.

Bongard, J., Zykov, V., & Lipson, H. (2006). Resilient machines through continuous self-modeling. *Science, 314,* 1118-1121.

Bradbury, J.A., Branch, K.M., & Malone, E.L. (2003). *An evaluation of DOE-EM public participation programs* (PNNL-14200). Richland, WA: Pacific Northwest National Lab.

Clark, W. (2002). *Waging modern war: Bosnia, Kosovo, and the future of combat*. PublicAffairs.

Cohen, L. (1996). *Time-freqency analysis: Theory and applications*. Prentice Hall.

Eisert, J., Wilkens, M., & Lewenstein, M. (1999). Quantum games and quantum strategies. *Physical Review Letters, 83*(15), 3077-3080.

Frieden, B.R. (2004). *Science from Fisher information*. New York: Cambridge University Press.

Gershenfeld, N. (2000). *The physics of information technology*. Cambridge, MA: Cambridge University Press.

Gibson, J.J. (1986). *An ecological approach to visual perception*. Hillsdale, NJ: Erlbaum.

Hagan, S., Hameroff, S.R., & Tuszynski, J.A. (2002). *Physical Review E, 65,* 1-11.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K.M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science, 304*, 438-441.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science, 303*, 1634.

Kelley, H.H. (1992). Lewin, situations, and interdependence. *Journal of Social Issues, 47*, 211-233.

Kirk, R. (2003). *More terrible than death. Massacres, drugs, and America's war in Columbia*. Public Affairs.

Kohli, R., & Hoadley, E. (2006). Towards developing a framework for measuring organizational impact of IT-Enabled BPR: Case studies of three firms. *The Data Base for Advances in Information Systems, 37*(1), 40-58.

Kruglanski, A.W., Pierro, A., Mannetti, L., & De Grada, E. (2006). Groups as epistemic providers: Need for closure and the unfolding of group-centrism. *Psychological Review, 113*, 84-100.

Lawless, W.F., Bergman, M., Louçã, J., Kriegel, N.N., & Feltovich, N. (2006a). *A quantum metric of organizational performance: Terrorism and counterterrorism, computational & mathematical organizational theory*. Springer Online. Retrieved May 31, 2008, from http://dx.doi.org/101007/s10588-006-9005-4

Lawless, W.F., Castelao, T., & Ballas, J.A. (2000). Virtual knowledge: Bistable reality and solution of ill-defined problems. *IEEE Systems, Man, & Cybernetics, 30*(1), 119-124.

Lawless, W.F., Bergman, M., & Feltovich, N. (2005). Consensus-seeking versus truth-seeking. *ASCE Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management, 9*(1), 59-70.

Lawless, W.F., & Grayson, J.M. (2004). A quantum perturbation model (QPM) of knowledge and organizational mergers. In L. van Elst & V. Dignum (Eds.), *Agent mediated knowledge management* (pp. 143-161). Berlin: Springer-Verlag.

Lawless, W.F., Wood, J., Everett, S., & Kennedy, W. (2006b). Organizational case study: Theory and mathematical specifications for an agent based model (ABM). In *Proceedings of Agent 2006*, DOE Argonne National Lab-University of Chicago, Chicago.

Lawless, W.F., Whitton, J., & Poppeliers, C. (2008). Case studies from the UK and U.S. of stakeholder decision-making on radioactive waste management. *ASCE Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management, 12*(2), 70-78.

Levine, J.M., & Moreland, R.L. (1998). Small groups. In D.T. Gilbert, S.T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (Vol. II, pp. 415-469). Boston: McGraw-Hill.

Mattiick, J.S., & Gagen, M.J. (2005). Accelerating networks. *Science, 307*, 856-8.

May, R.M. (1973/2001). *Stability and complexity in model ecosystems*. Princeton University Press.

Moreland, R.L. (1996). Lewin's legacy for small groups research. *Systems Practice (Special issue of the journal, edited by S. Wheelan, devoted to Kurt Lewin), 9*, 7-26.

Rieffel, E.G. (2007). Certainty and uncertainty in quantum information processing. In *Proceedings of Quantum Interaction: AAAI Spring Symposium*. Stanford University: AAAI Press.

Romanelli, E., & Tushman, M.L. (1994). Organizational transformation as punctuated equilibrium: An empirical test. *Academic Management Journal, 37*, 1141-66.

Sen, A. (2000). *Development as freedom*. Knopf.

Weick, K.E., & Quinn, R.E. (1999). Organizational change and development. *Annual Review of Psychology, 50*, 361-386.

KEY TERMS

Entanglement (quantum): Occurs when one quantum object enters into a (quantum) superposition with another, characterized by having a mutual influence on each other that requires a description of both with reference to the other whether or not spatially separated.

Fourier Pairs: The Fourier transform and its inverse form Fourier pairs; e.g., $f(t) \Leftrightarrow F(\omega)$.

Fourier Transforms: A representation of a signal received over time can be transformed into harmonic frequencies in the frequency domain. A uniform sine wave is transformed into a single frequency. The Fourier transform and its inverse are also known as harmonic analysis.

Interdependence: Occurs when two or more objects influence each other. Also, mutual sensitivity, mutual connectedness, or where an action on one object affects the other(s).

Social Influence: Occurs when an action on one or more individual(s) affects the other individuals in a group of two or more agents.

Tradeoffs: Occurs when one aspect of a phenomenon, such as its resolution, is improved while another aspect is lost or degraded.

Uncertainty Principle: Applies when a system with a pair of observables, such as the factors of action and observation, that are not independent, but rather interdependent, precluding a precise knowledge of both observables simultaneously.

A Classification of Approaches to Web-Enhanced Learning

Jane E. Klobas
University of Western Australia, Australia
Bocconi University, Italy

Stefano Renzi
Bocconi University, Italy
University of Western Australia, Australia

INTRODUCTION

The World Wide Web has become a mature platform for the support of learning at universities. Several patterns have emerged, both in the nature of use, and in understanding the conditions associated with successful adoption and application of web-enhanced learning (WEL). This article summarizes, in the form of nine scenarios, the ways in which the Internet is being used to enhance learning in traditional universities. It also discusses the changes needed if universities are to benefit more widely from WEL.

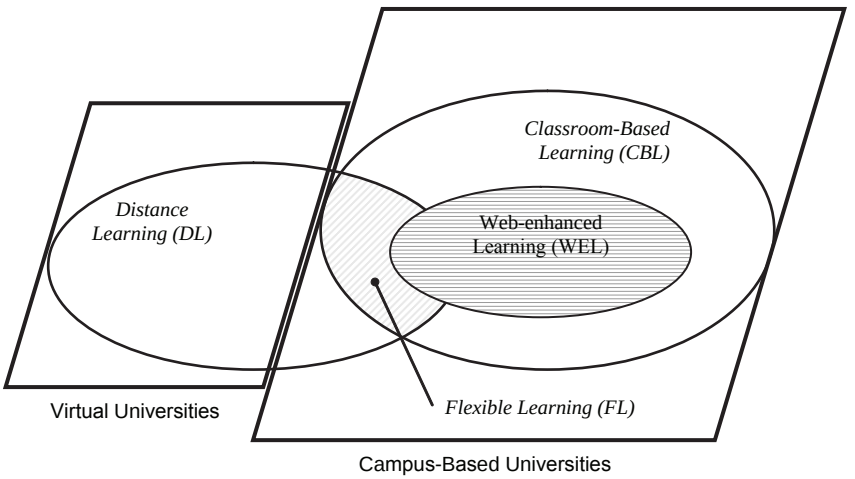
BACKGROUND

The Web is used by universities to make courses available to students who are distant from campus (*distance learning*, DL) and to enhance learning by students who attend courses on-campus (*web-enhanced learning*, WEL). Universities may be classified on the basis of the modes of learning that they offer. *Virtual universities* offer access to courses by DL only. Traditional, or *campus-based universities*, offer courses that are based on formal lessons held in classrooms or laboratories (*classroom-based learning*, CBL), but may also offer courses by DL, or *flexible learning* (FL), a combination of DL and CBL.

WEL is the use of the Web to enhance CBL in traditional universities. WEL provides students studying in the classroom with access to electronic resources and learning activities that would not be available to them in traditional classroom-based study. The simplest forms of WEL provide access to the Web from within the classroom, using the Web as a platform for real-time demonstration or as a digital library. More sophisticated forms of WEL blend activities in the classroom with Web-enabled learning activities that promote collaborative learning among students, even when they are distant from the classroom.

Figure 1 illustrates the relationship between the modes of learning offered by universities. WEL is represented as that portion of CBL that uses the Web to enhance learning.

Figure 1. The relationship between Web-enhanced learning (WEL) and other modes



When it is used to blend in-classroom and out-of-classroom activities, WEL shares the characteristics of DL and FL.

WEL differs from flexible learning in that the focus of the lesson remains the traditional classroom. With FL, classroom-based learning is mixed with learning at a distance. In the most common form of FL, *distributed learning* (also known as *blended learning* or *mixed mode learning*), students participate in formal lessons both in the classroom and at a distance, according to a schedule prepared by the instructor. Some flexible learning may be enhanced by use of the Web, for example, to provide discussion forums in which students studying at a distance and in the classroom may participate together, but use of the Web is not necessary for flexible learning.

This article is concerned with integration of online learning and classroom-based learning to achieve effective and manageable WEL for campus-based students. The focus is on change across a university system rather than in an individual classroom. We argue that WEL adds the most value when it is used to enable new forms of learning, and in particular, online collaborative learning by students working at a distance from the classroom as well as within it (Rudestam & Schoenholtz-Read, 2002). This value can only be obtained through attention at the institutional level to the organizational transformation required to implement, support, and sustain WEL (Bates, 2000).

WEL SCENARIOS

Nine distinct scenarios for use of WEL can be identified (Table 1, based on Klobas & Renzi, 2003). They can be divided into four groups: *information provision* scenarios, in which the Web is used to provide information to students and others outside the classroom; *classroom resource* scenarios, in which the Web is used to extend the classroom, either by providing access to resources in the classroom or by enabling lessons to be broadcast outside the classroom; *interactive learning* scenarios, which range from interactive virtual classrooms to the use of the Web to support collaborative learning among students working at a distance; and an *experimental* scenario, in which the Web is used to experiment with technology and pedagogy in ways not envisaged by the preceding scenarios. Any or all of the scenarios may be used alone or in combination, in a single course or by a single university.

Table 1. A hierarchy of WEL use scenarios

Scenario	Label	Use
INFORMATION PROVISION SCENARIOS		
1	Catalog	Provision of static, and primarily logistic, information about the course
2	Notice Board	Distribution of course materials in electronic form
3	Class Resource	Provision of additional materials and references in response to student and teacher experience in the course as it progresses
CLASSROOM RESOURCE SCENARIOS		
4	Classroom Resource	Use of the Web for demonstration or as a digital library during classroom sessions
5	Streaming Video	Broadcast of classroom sessions
INTERACTIVE LEARNING SCENARIOS		
6	Virtual Classroom	Synchronous interactive classroom sessions that include video and voice communication among instructors and students
7	Interactive Web	An interactive environment outside the classroom
8	CSCL	Computer Supported Collaborative Learning
EXPERIMENTAL SCENARIO		
9	Experimental	An experimental environment for innovative use of the Web

Information Provision

The first group of scenarios (1 to 3) represent incremental changes to traditional classroom-based learning: In these scenarios, the Web is used as an information delivery mechanism that provides students with some flexibility in the time and place with which they access some of the information required for the course. The scenarios range from simple publication of course catalog information to use of streaming video to permit students to 'attend' classes outside the classroom. The information and communications technology (ICT) infrastructure, training, and skills required for successful use of each scenario ranges from simple in the case of the *catalog* scenario to more complex in the case of the *streaming video* scenario.

The simplest, and most common, of WEL scenarios consists of provision of basic *catalog* information about a course: course description, list of textbooks, name of teacher(s), class schedule, allocated classroom(s), and examination procedures. Most university Web sites contain at least a subset of this information. This is a simple scenario to manage. The university needs a web server and the staff to maintain it. The information posted in such catalogs is often pooled or available from a single source. Because it is static, and seldom needs to be updated more than once a semester, the effort involved in maintaining currency is quite low.

In Scenario 2, *notice board*, teachers use the Web to distribute course materials in electronic form. Such material might include: educational material used by teachers in the classroom (slides, case studies, newspaper articles, site URLs related to the course content), past exam scripts and solutions, and official University communication. The content may be made available all at once before the course begins in the online equivalent of a coursebook, or from time to time during the life of the course (for example, lesson slides may be put online after each lesson).

Use of the Web in Scenario 3, *class resource*, is more dynamic than in Scenario 2. The teacher selects new material to add to the course Web site during the course, in response to questions asked, interest expressed, and other experiences of how the students are responding to the course as it is delivered. In courses with multiple classes (sections), each class may share the material available on the common course notice board, but may have quite different additional resources.

Effective adoption of Scenario 2 and Scenario 3 requires more extensive ICT infrastructure than Scenario 1 to permit students to access the course web sites on a regular basis. At Scenario 2, some universities make staff available to load materials to the site on behalf of the teacher, but at Scenario 3, the teachers need the IT and the training to be able to load their own materials. This level therefore marks a significant shift in the resources required to support WEL. At the same time, it marks a significant shift in the value added by WEL; at this level, WEL provides the opportunity to quickly and easily provide students with access to current material of direct relevance to their learning and experience in the course as it unfolds.

The Web as a Classroom Resource

In Scenario 4, *classroom resource*, the teacher may use the Web to access reference materials, presentations and demonstrations from sources outside the classroom. In this scenario, the Web provides convenient access to resources that might previously have been drawn from other sources or accessed in other ways by the teacher. While relatively simple for the teacher, this scenario requires provision of secure access to appropriate IT and Internet infrastructure from within the classroom.

Scenario 5, *streaming video*, requires more substantial investment in technology, including quality recording equipment in classrooms and the staff to operate and maintain it, high end servers, and high speed networks to all locations where the video may be watched. There are many systems available on the market. The University of Toronto has developed a system called ePresence which allows students to navigate inside the lesson and assists with management of an archive of lessons (Baecker, Moore & Zijdemans, 2003). Lectopia, developed by the University of Western Australia, offers a high level of automation in recording and publishing, making it unintrusive for teaching staff (Fardon, 2003). For effective use, teachers need to learn how to structure and present visual aids that will be suitable both in the classroom and for presentation by video. The primary uses of streaming video, to allow students the option of 'attending classes' from outside the classroom and to review the teacher's presentation of different material (Creighton & Buchanan, 2001), represent only an incremental change in the nature of education.

Learning Through Interaction

Use of the Web is more dynamic in the Interactive Learning scenarios, which involve interaction between the teacher and students, and among the students themselves.

In the WEL *virtual classroom* scenario (Scenario 6), the Web is used to transmit complete classroom lessons using synchronous video, voice, whiteboard, and other utilities. Teachers and students from different locations may share lesson components, ask questions, respond and interact with one another in a variety of ways. For campus-based students, a virtual classroom provides the opportunity to share classroom experiences with teachers and students in classrooms located on other campuses (Hiltz & Wellman, 1997). Universities

considering this option for WEL should weigh the considerable costs of investment in ICT infrastructure, training and course redesign against the return to students.

At Scenario 7, *interactive web*, the interactions are somewhat passive, based mainly on the use of course forums, resource contributions, self evaluation tests, delivery of assignments, and secure online exams. This is the most common application of online learning platforms. Teachers require considerable training and support to adopt this scenario effectively. Students also require preparation, both in use of the functions of the technology, and in how to use the provided functions to improve the quality of their course experience and learning.

A more complex interactive scenario is *CSCL* (Computer Supported Collaborative Learning, Scenario 8), an environment where at least the online component of teaching and learning is based primarily on interactions among students working in groups. This scenario includes collaborative group learning activities that go beyond those possible with simple course forums. Such activities may include group projects which involve sharing materials or preparation of joint documents. This scenario offers greater potential for improving the quality of learning at universities than any of the preceding scenarios.

Indeed, the power of WEL to change the quality of education is based on its potential to make collaborative learning possible. WEL makes a difference when it is used to enable students to learn collaboratively (Friedlander, 2002; Klobas & Renzi, 2003; Lammintakanen & Rissanen, 2003; Rudestam & Schoenholtz-Read, 2002). Students, themselves, describe the value of participation in learning communities of peers (Hamilton & Zimmerman, 2002), while educators claim that participation in collaborative learning not only results in better quality learning of course subject matter, but also in changes in the world view of the students and their capacity for lifelong learning and contribution to society (Klobas, Renzi, Francescato, & Renzi, 2002; Rudestam & Schoenholtz-Read, 2002). Furthermore, collaborative learning that makes a difference does not need expensive technologies. CSCL does not require the investment in ICT infrastructure of Scenario 6, and can be implemented with simple asynchronous conferencing software (Hazemi & Hailes, 2002; Hiltz & Turoff, 2002).

Experimental Scenario

The final scenario, *experimental* (Scenario 9), provides an environment for teachers to experiment with new applications of the Web in the classroom, new Web-based technologies, new educational strategies, and the interaction between these innovations. While current thinking focuses on CSCL as the most advanced form of WEL, the existence of an experimental scenario reminds us to be open to further changes in learning theory and technology. For example, the debate about how to assess both social interaction and learning in CSCL is still open and practical examples of assessment (Chan & van Aalst, 2004; Roberts, 2006) can still be considered as experimental. It also reminds us of the need to evaluate new approaches to learning associated with WEL.

FUTURE TRENDS

Universities across the globe, in developed and developing countries, have been quick to adopt technologies to support WEL, including the infrastructure to support widespread use of the Web as an information resource, and university-wide platforms to support online learning. But this rapid adoption of technology has had relatively little impact on the education of campus-based students (Middlehurst, 2003, as cited in Collis & Van der Wende, 2002; Observatory of Borderless Education, 2002). Those changes that have occurred have been incremental rather than transformational. The Web is most frequently used in a "distributive" mode to provide access to resources—as a substitute for, or complement to, notice boards, distribution of handouts, and use of the library—rather than to provide access to new forms of learning. Thus, the Web is being used to automate rather than to transform university education. Few attempts to go beyond this simple automation have been successful (Pollock & Cornford, 2000). Those universities that are making greater use of the Web are not distinguished from the others by virtue of their information technology infrastructure, but in terms of their focus on students, markets, and policy. Those looking to 'stretch the mould' of the future university emphasize flexibility in the location of learning, and have policies in place for quality, future markets, and costs and efficiency (Collis & Van der Wende, 2002) along with systems for training and supporting teachers (Trentin, 2006).

Radical changes in our approach to university education are therefore needed if universities are to benefit from WEL. Bates (2000) claims that "If universities and colleges are successfully to adopt the use of technologies for teaching and learning, much more than minor adjustments in current practice will be required. Indeed, the effective use of technology requires a revolution in thinking about teaching and learning." (p. v). That revolution, according to Rudestam & Schoenholtz-Read (2002), "demands a reexamination of our core beliefs about pedagogy and how students learn" (p. 4) based on theories of constructivism and collaborative learning (Leidner & Jarvenpaa, 1995). Change on this scale requires vision and leadership, as well as appropriate resources.

While much of the literature points to CSCL as a necessary part of the radical change, issues associated with implementing CSCL illustrate how difficult such a change might be. There is, for example, still no agreement on how CSCL can

be assessed in a way that is feasible for the teacher and the university. There continue to be calls for assessment practices to support the shift toward social constructivist educational theories (Chan & van Aalst, 2004).

Bates (2000) calls for "fundamental change in the way our higher education institutions are organized and managed" (p. 5) because "history dictates that the introduction of new technology is usually accompanied by major changes in the organization of work" (p. 1). This requires leadership, vision, policies, planning and evaluation that emphasize the educational goals of WEL rather than just its technological characteristics (Bates, 2000; Collis and Van der Wende, 2002; Friedlander, 2002; Klobas & Renzi, 2003; Pollock & Cornford, 2000; Surry, 2002). To this end, a group of Australian and British universities have developed a model for benchmarking the use of ICT in teaching and learning (Ellis & Moore, 2006).

While they have put the necessary ICT infrastructure in place, universities have paid relatively little attention to development of the human resources needed for successful adoption of WEL (Collis & Van der Wende, 2002). Financial investment in training and skill development for teachers, students, and technical staff is required. It is not surprising, given arguments for attention to pedagogy rather than technology, that successful adoption of WEL is associated with training to develop teachers' ability to design courses that use WEL to improve pedagogy rather than training that emphasizes the features of specific software (Klobas & Renzi, 2003).

Human resource issues associated with optimization and management of the ICT infrastructure also need to be addressed. This requires attention to engagement with the university's partners in the supply and maintenance of the ICT infrastructure for WEL (Pollock & Cornford, 2000). The ICT infrastructure for WEL involves several layers of technology (internal and external networks, servers, and applications), and successful WEL requires skillful coordination of suppliers (Klobas & Renzi, 2003).

Changes in the reward systems for university staff are necessary to support the changes in work demanded by WEL (Bates, 2000; Collis & Van der Wende, 2002; Klobas & Renzi, 2000, 2003; Surry, 2002). Such changes in reward systems require short term financial investment, but in the longer term are associated with changes in the nature and structure of work in the university.

CONCLUSION

WEL provides traditional universities with opportunities to enhance the quality of the education they provide to students on campus. While most current approaches to WEL involve incremental changes to classroom teaching, greater value is obtained through application of WEL to improve opportunities for collaborative learning among students. Success therefore requires attention–at the most senior levels–to educational values, financial and human resources, and transformation of educational processes and organizational structure, as well as to technology. WEL will become a natural part of the educational model of those universities with the management commitment and skill to implement and sustain the transformation required, while other universities may find it difficult to survive.

REFERENCES

Baecker, R. M., Moore, G., & Zijdemans, A. (2003). Reinventing the lecture: Webcasting made interactive. In *Proceedings of HCI International 2003* (pp. 896-900). Lawrence Erlbaum Associates.

Bates, A. W. T. (2000). *Managing technological change: Strategies for college and university leaders*. San Francisco: Jossey Bass.

Bento, R., & Schuster, C. (2003). Participation: The online challenge. In A. K. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 156-164). Hershey, PA: Information Science Publishing.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Chan, C., & van Aalst, J. (2004). Learning, assessment and collaboration in computer-supported environments. In J. W. Strijbos, P. A. Kirschner, & R. L. Martens (Eds.), *What we know about CSCL in higher education* (pp. 87-113). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Clulow, V., & Brace-Govan, J. (2003). Web-based learning: Experience-based research. In A. K. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 49-70). Hershey, PA: Information Science Publishing.

Collis, B., & van der Wende, M. (2002). *Models of technology and change in higher education: AniInternational comparative survey on the current and future use of ICT in higher education* (Report). Enschede, The Netherlands: University of Twente, Center for higher Education Policy Studies.

Creighton, J. V., & Buchanan, P. (2001). Toward the e-campus: Using the Internet to strengthen, rather than replace, the campus experience. *EduCause Review, 36*(2), 12-13.

Ellis, R. A., & Moore, R. R. (2006). Learning through benchmarking: Developing a relational, prospective approach to benchmarking ICT in learning and teaching. *Higher Education, 51*(3), 351-371.

Fardon, M. (2003). Internet streaming of lectures: A matter of style. In *Expanding the learning community: Meeting the challenges* (pp. 699-708). Proceedings of the EDUCAUSE in Australasia 03 Conference, 6-9 May 2003. Adelaide, South Australia.

Friedlander, L. (2002). Next generation distant learning. In F. Fluckiger, C. Jutz, P. Schulz, & L. Cantoni (Eds.), *4th International Conference on New Educational Environments,* Lugano, Switzerland, May 8-11, 2002 (pp. 3-6). Lugano: University of Applied Sciences Southern Switzerland and University of Southern Switzerland, Berne: Net4net.

Hamilton, S., & Zimmerman, J. (2002). Breaking through zero-sum academics. In K. E. Rudestam & J. Schoenholtz-Read (Eds.), *Handbook of online learning: Innovations in higher education and corporate training* (pp. 257-276). Thousand Oaks, CA: Sage.

Hazemi, R., & Hailes, S. (2002). Introduction. In R. Hazemi & S. Hailes (Eds.), *The digital university: Building a learning community*. London: Springer.

Hiltz, S. R., & Turoff, M. (2002). What makes learning networks effective? *Communications of the ACM, 45*(4), 56-59.

Hiltz, S. R., & Wellman, B. (1997). Asynchronous learning networks as a virtual classroom. *Communications of the ACM, 40*(9), 44-49.

Klobas, J. E. & Renzi, S. (2000). Selecting software and services for web-based teaching and learning. In A. K. Aggarwal (Ed.). *Web-based learning & teaching technologies: Opportunities and challenges* (pp. 43-59). Hershey, PA: Idea Group Publishing.

Klobas, J. E., & Renzi, S. (2003). Integrating online educational activities in traditional courses: University-wide lessons after three years. In A. K. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 415-439). Hershey, PA: Information Science Publishing.

Klobas, J. E., Renzi, S., Francescato, D., & Renzi, P. (2002). Meta-response to online learning / Meta-risposte all'apprendimento online. *Ricerche di Psicologia, 25*(1), 239-259.

Lammintakanen, J., & Rissanen, S. (2003). An evaluation of web-based education at a Finnish university. In A. K. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 440-453). Hershey, PA: Information Science Publishing.

Leidner, D. E., & Jarvenpaa, S. L. (1995). The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly, 19*(3), 265-291.

Middlehurst, R. (2003). Competition, collaboration and ICT: Challenges and choices for higher education institutions. In M. van der Wende & M. van der Ven (Eds.), *The use of ICT in higher education: A mirror of Europe*. Utrecht: Lemma.

Observatory of Borderless Education. (2002). *Online Learning in Commonwealth Universities*. Retrieved August 23, 2007, from obhe.ac.uk/products/reports/publicaccesspdf/OnlineLearningCUpartone.pdf

Palloff, R. M., & Pratt, K. (2002). Beyond the looking glass: What faculty and students need to be successful online. In K. E. Rudestam & J. Schoenholtz-Read (Eds.), *Handbook of online learning: Innovations in higher education and corporate training* (pp. 171-184). Thousand Oaks, CA: Sage.

Pollock, N., & Cornford, J. (2000). *Theory and practice of the virtual university.* Ariadne (24). Retrieved on August 23, 2007, from http://Web.ariadne.ac.uk/issue24/virtual-universities

Roberts, T. S. (2006). *Self, peer and group assessment in e-Learning*. Hershey, PA: Information Science Publishing.

Rudestam, K. E., & Schoenholtz-Read, J. (2002). The coming of age of adult online education. In K. E. Rudestam & J. Schoenholtz-Read (Eds.), *Handbook of online learning: innovations in higher education and corporate training* (pp. 3-28). Thousand Oaks, CA: Sage.

Sauter, V. L. (2003). Web design studio: A preliminary experiment in facilitating faculty use of the web. In A. K. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 131-154). Hershey, PA: Information Science Publishing.

Shapiro, J. J., & Hughes, S. K. (2002). The case of the inflammatory e-mail: Building culture and community in online academic environments. In K. E. Rudestam & J. Schoenholtz-Read (Eds.), *Handbook of online learning: Innovations in higher education and corporate training* (pp. 91-124). Thousand Oaks, CA: Sage.

Surry, D. W. (2002, April). *A model for integrating instructional technology into higher education.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Trentin, G. (2006). The Xanadu project: Training faculty in the use of information and communication technology for university teaching. *Journal of Computer Assisted Learning, 22*(3),182-196.

KEY TERMS

Blended Learning: *See mixed mode learning.*

Collaborative Learning: Learning that occurs through the exchange of knowledge among learners. Collaborative learning is a form of social learning.

Computer-Supported Collaborative Learning (CSCL): Collaborative learning that occurs via the medium of computer-based communication networks such as the Internet.

CSCL: *See computer-supported collaborative learning.*

Distributed Learning: *See mixed mode learning.*

Flexible Learning: Systems in which students may choose to complete some of their learning on-campus and some of their learning off-campus.

Mixed Mode Learning: Study that combines traditional face-to-face learning with learning at a distance in a structured program. The Web may be used to enhance learning during study by one or both of these modes. Mixed Mode is also known as Blended Learning and Distributed Learning.

Online learning Activities: Learning activities in which students interact with resources, or other students, or both, using the capabilities of the Internet or other computer-based communication networks.

Social Learning: Learning through social interaction with other people.

Web-Enhanced Learning (WEL): Use of the World Wide Web (Web) to provide students studying in the classroom with access to electronic resources and learning activities that would not be available to them in traditional classroom-based study. The simplest forms of WEL provide information about a course on the Web and access to the Web from within the classroom. More sophisticated forms of WEL blend activities in the classroom with Web-enabled online learning activities which promote collaborative learning among students even when they are distant from the classroom.

Classification of Semantic Web Technologies

Rui G. Pereira
University of Beira Interior, Portugal

Mário M. Freire
University of Beira Interior, Portugal

INTRODUCTION

Semantic Web is the name of the next generation World Wide Web, that has been recently proposed by Tim Berners-Lee and the World Wide Web Consortium (W3C)[1]. In this new Web architecture, information and Web services will be easily understandable and usable by both humans and computers. The objective is not to make computers understand the human language, but to define a universal model for the expression of the information and a set of inference rules that machines can easily use in order to process and relate the information as if they really understood it (Berners-Lee, 1998). Though, as the current Web provided sharing of documents among previously incompatible computers, the Semantic Web intends to go beyond, allowing stovepipe systems, hardwired computers, and other devices to share contents embedded in different documents. The most known architecture for Semantic Web is based on a stack of related technologies, each one being a whole research area by itself (Berners-Lee, Hendler, & Lassila. 2001; Pereira & Freire, 2005).

Accomplishment of the Semantic Web is considered a great challenge, not only due to the complexity of implementation but also because of the vast applicability in several areas. In spite of this, Semantic Web is still one of the most promising research areas among those which aim to define a new architecture for the Web.

Semantic Web goes far beyond previous information retrieval and knowledge representation projects, presenting a non-centralized way to represent and contextualize real-world concepts, unambiguously, for several areas of knowledge. Semantic Web-enabled machines will handle information at our communication level. It is clear that the ability to interpret reality is still very primitive, however, Semantic Web points a way towards machine interaction and learning (Pereira et al., 2005). Semantic Web will integrate, interact with, and bring benefits to most human activities. Its full potential will go beyond the Web to real-world machines, providing increased interaction between machines and with humans—smarter phones, radios, and other electronic devices. Semantic Web will bring a different kind of approach in the understanding of reality by the machines and will constitute a mark in the evolution of human knowledge (Pereira et al., 2005).

BACKGROUND

The arrival of many new Semantic Web technologies over the last few years reveals the acceptance and credibility of the Semantic Web architecture. Nevertheless, the fast appearance of these new technologies and the wide diversity of areas and forms where they can apply demand a high effort of knowledge updating about actual features used in Semantic Web technologies to all future developers. Due to the urgent need of eliminating a huge lack of information about recent Semantic Web technologies, we provide a classification in categories for 80 of more than 100 recent Semantic Web technologies that are accessible from the Web. Their main features are also presented.

This work is mainly based on information presented in Bizer (2005) and SemWebCentral (2005). We have classified the 80 Semantic Web technologies in 11 categories: Visualization, validation, conversion, annotation, brower, query, editor, integration, repository, API (Application Programming Interface), and reasoner. Next section provides a detailed description by category of the main existing Semantic Web technologies.

SEMANTIC WEB TECHNOLOGIES BY CATEGORY

In this section we present eleven tables with technologies grouped by each category and up to ten of their main characteristics:

- **Developer:** Developer name;
- **Release:** Latest release number and publication date;
- **License:** License under which the technology is distributed;
- **Language:** Language development used to develop the technology;
- **API-Paradigm:** API-Paradigm for the manipulation of RDF data;
- **Query-Languages:** Query-Languages used by technology;

Table 1. Main features of annotation technologies

	Developer	Release	License	Language	Input	Output	O.S.
AeroSWARM[2]	UBOT Team	-	-	-	HTML	OWL	Independent
OntoMat--Annotizer[3]	Handschuh, S., Braunv, M., Buerkle, C., Kühn, K., Meyer, L. and Krekeler, T.	0.8.2 Feb-05	LGPL	Java	N3 OWL RDF	N3 OWL RDF	Independent
PhotoStuff[4]	Mindswap	2.11 Mar-05	Mozilla	-	-	-	Independent
RIC[5]	Michael Grove (Mindswap)	3.0 Alpha	(Grove, 2002)	Java	-	-	Independent
SemanticWord[6]	Teknowledge Corporation	Alfa 1.0 Aug-04	-	Visual Basic	OWL	OWL	Windows NT/2000
Swangler[7]	STET	1.0.1 Apr-05	GPL	Java	OWL RDF	OWL	Independent

Table 2. Main features of API technologies

	Developer	Release	License	Language	API-Paradigm	Storage Model	Input	Output	O.S.
CARA[8]	Stefan Kokkelink	Pre0.001 Mar-01	GPL	Perl	Resource-centric	Memory	RDF/XML N-Triples	RDF/XML N-Triples	-
CODIP[9]	DARPA	0.9.0 Dec-04	BSD	Java	-	-	OWL UML XMI	OWL UML XMI	Independent
HAWK[10]	SWAT Lab, Lehigh University	1.1 Beta Dec-04	GPL	Java PL/SQL	-	-	OWL RDF	OWL RDF	Independent
Kazuki[11]	Self, T., Lerner, J., and Rager D.	1.2 Jun-04	BSD	Java	-	-	OWL	-	Independent
NG4J[12]	Chris Bizer Richard Cyganiak Rowland Watkins	0.4 Fev-05	BSD	Java	Statement-centric	Memory MySQL	RDF/XML N-Triples N3 TriX TriG	RDF/XML N-Triples N3 TriX TriG	Independent
OWL API[13]	Bechhofer, S., Volz, R., Kalyanpur, A., Crowther, P., Horan, B., Turi, D., and Lord, P.	1.4.2 Mar-05	LGPL	Java	-	-	OWL	OWL	Independent
OWL-S API[14]	Paolucci, M., Srinivasan, N. Softagents Group	0.1 beta Dec-04	LGPL	Java	-	-	OWL	OWL	Independent
OWL Semantic Search Services[15]	Bhanu Vasireddy John Li	0.1 beta Dec-04	-	Java Java-Script Prolog	-	-	-	-	Linux
Pyrple[16]	Sean B. Palmer	Jun-04	-	Python	Statement-centric	Memory	RDF/XML N3 N-Triples	RDF/XML N3 N-Triples	-
SOFA[17]	Alishevskikh, A. Mihalik, I. and Ganesh, S.	0.3 Mar-05	LGPL	Java	Resource-centric Ontology-centric	Memory JDBC compliant DB	OWL RDF	OWL RDF	Independent
Sparta[18]	Mark Nottingham	0.6	-	Python	Resource-centric	-	RDF/XML N3 N-Triples	RDF/XML N3 N-Triples	-

- **Model Storage:** Model storage for information;
- **Input:** Input formats that technology can use;
- **Output:** Output formats that technology can export;
- **O.S.:** Operation System that supports the technology.

Technologies had been grouped in accordance with its main characteristics. Despite this, some technologies could have been associated with more than one category.

Annotation Category

Annotation category is defined as the group of Semantic Web technologies that has as goal the manipulation of extra information associated with a particular point in a document. Main features of these technologies can be seen in Table 1.

API Category

API category is defined as the group of Semantic Web technologies that provides a set of commonly-used functions that can be use by others technologies. Main features of these technologies can be seen in Table 2.

Browser Category

Browser category is defined as the group of Semantic Web technologies that enables a user to display and interact with documents hosted by Web servers. Main features of these technologies can be seen in Table 3.

Conversion Category

Conversion category is defined as the group of Semantic Web technologies that has as a goal the changing of an entity of

Table 3. Main features of browser technologies

	Developer	Release	License	Language	Input	Output	O.S.
DumpOnt[19]	Moore, D., Kolas, D. Lerner, J., Dean, M. Blace, R., and Self, T.	1.2 Feb-04	BSD	Java	OWL RDF	HTML	Independent
FlinkCommands[20]	Barreau, G., Rocha, R. Machado, M., Gama, C. Abdalla, D., Gagnon, M. Oliveira, K., Anquetil, N.	-	GPL	Java	OWL	OWL	Linux
HyperDAML[21]	Lerner, J. Dean, M., and Self, T.	Jan-04	BSD	Java	OWL RDF	HTML	Independent
Object Viewer[22]	Jeremy Lerner Troy Self	Jul-04	BSD	Java	OWL RDF	-	Independent
OWL-P[23]	Mallya, A., and Desai, N.	1 Nov-04	-	Java	OWL	OWL	Independent
OWL-S IDE[24]	Naveen Srinivasan Softagents Group	0.1 beta Nov-04	-	Java	OWL XML	OWL	Linux Windows
Visual Variable-Depth Info Display[25]	Baoshi Yan	1.1 beta Set-04	Mozilla	Java Java-Script	OWL RDF XMI	HTML	Windows NT/2000

Table 4. Main features of conversion technologies

	Developer	Release	License	Language	Input	Output
java2owl-s[26]	Naveen Srinivasan Softagents Group	Beta Jun-04	-	-	Java	OWL-S
owl2dig[27]	Lei Zhang Jian Zhou	0.1 Jun-04	GPL LGPL	Java	OWL	DIG
OWL Converter[28]	Mindswap	1.2 Dec-03	-	Perl	DAML+OIL	OWL
Owl-s2uddi[29]	Naveen Srinivasan Softagents Group	Beta Jun-04	-	-	OWL-S	UDDI
SWeHG[30]	Semantic Computing Research Group	Out-03	-	Perl Prolog	RDF	HTML
Wsdl2owl-s[31]	ATLAS Group	1.0.1 Jan-05	-	Java	XML XML-Schema	OWL RDF

one datatype into another. Main features of these technologies can be seen in Table 4.

Editor Category

Editor category is defined as the group of Semantic Web technologies that can be used to make changes to documents of a particular type. Main features of these technologies can be seen in Table 5.

Integration Category

Integration category is defined as the group of Semantic Web technologies that has as a goal the combination of several components that are very useful when used together. Main features of these technologies can be seen in Table 6.

Query Category

Query category is defined as the group of Semantic Web technologies that allows interrogating and filtering described information in RDF/XML, N3, and OWL formats. Main features of these technologies can be seen in Table 7.

Reasoner Category

Reasoner category is defined as the group of Semantic Web technologies that has as a goal the manipulation of data in a way that can find new facts from existing data. Main features of these technologies can be seen in Table 8.

Repository Category

Repository category is defined as the group of Semantic Web technologies that has as a goal the storage and retrieval of Semantic Web metadata. Main features of these technologies can be seen in Table 9.

Validation Category

Validation category is defined as the group of Semantic Web technologies that can be used to check the consistency of semantic web documents. Main features of these technologies can be seen in Table 10.

Visualization Category

Visualization category is defined as the group of Semantic Web technologies that can be used to create graphical representations of RDF models and OWL ontologies. Main features of these technologies can be seen in Table 11.

CONCLUSION

This study presented a classification and assessment of the majority of existing Semantic Web technologies. It helps to establish a link between Semantic Web developers/researches and the actual Web users, in order to draw the attention of current Web users to available Semantic Web technologies. This interest is vital to the viability of the Semantic Web

Table 5. Main features of editor technologies

	Developer	Release	License	Language	Input	Output	O.S.
ORIENT[32]	APEX Lab	0.1.1 May-04	OSI Approved	Java	OWL RDF XMI	OWL RDF XMI	Independent
OWL Filetype Plugin For VIM[33]	Jeremy Lerner, Troy Self	Jan-04	BSD	-	OWL	OWL	-
OWL Mode For Emacs[34]	BBN Technologies	Beta Jan-04	BSD	-	-	OWL	Windows POSIX
Protégé[35]	Stanford Medical Informatics	3.0 Fev-05	Mozilla	Java	-	-	Independent
SMORE[36]	Mindswap	5.0 Apr-05	LGPL	Java	-	-	Independent
SWeDE[37]	BBN Technologies	1.0.2 Mar-05	BSD	Java	OWL	OWL	Independent
Swedt[38]	UBI	0.1 May-05	BSD	Java	XML	XML	Independent
SWOOP[39]	Mindswap	2.2 Mar-05	LGPL	Java	OWL RDF/XML N3 Turtle	OWL RDF/XML N3 Turtle	Independent

Table 6. Main features of integration technologies

	Developer	Release	License	Language	API-Paradigm	Query-Languages	Model Storage	Input	Output
4Suite[40]	Fourthought, Inc	1.0a4 Nov-04	Apache	Python	Statement-centric	Versa	Memory File PostgreSQL	RDF/XML	RDF/XML
Jena[41]	HP Labs Semantic Web Research	2.2 Jan-05	BSD	Java	Statement-centric Resource-centric Ontology-centric	RDQL SPARQL	Memory File Berkeley MySQL SQLite	RDF/XML N-Triples N3 Turtle	RDF/XML N-Triples N3 Turtle
KAON[42]	FZI WIM and AIFB LS3	1.2.7 Nov-04	LGPL	Java	Statement-centric Resource-centric Ontology-centric	KAON	Memory Any SQL2 compliant DB	RDF/XML	RDF/XML
PerlRDF[43]	Ginger Alliance	0.31 Mar-02	GPL Mozilla	Perl	Statement-centric Resource-centric Ontology-centric	Resource-centric	Memory File PostgreSQL	RDF/XML N3	RDF/XML N3
RAP[44]	Westphal, D., Bizer, C., Oldakowski, R., Gauß, T., Dawes, P., Grimnes, G., Köstlbacher, A., Auer, S., Smith, L., Lopez, L., Catanzani, R., Willy, S.	0.91 Dec-04	LGPL	PHP	Statement-centric Resource-centric Ontology-centric	RDQL	Memory ADOdb compliant Databases	RDF/XML N-Triples N3 GRDDL	RDF/XML N-Triples N3 GRDDL
RDF Gateway[45]	Intellidi-mension	2.2.2 Jan-05	Propri.	-	Model-centric Statement-centric Resource-centric	RDFQL	Memory File All OleDB compliant	RDF/XML N-Triples N3 Turtle	RDF/XML N-Triples N3 Turtle
Redland RDF[46]	Dave Beckett	1.0.0 Jan-05	GPL LGPL Apache	C	Statement-centric Resource-centric Ontology-centric	RDQL SPARQL	Memory File Berkeley MySQL SQLite PostgreSQL Microsoft SQLServer DB2	RDF/XML N-Triples Turtle RSS	RDF/XML N-Triples Turtle RSS
Wilbur[47]	Nokia Research Center	Nov-04	NOKOS	Common Lisp	Statement-centric Resource-centric	WQL	Memory	RDF/XML	RDF/XML

Table 7. Main features of query technologies

	Developer	Release	License	Language	API-Paradigm	Input	Output	O.S.
Cwn[48]	W3C	Jan-05	W3C	Python	Model-centric	N3 OWL RDF/XML	N3 OWL RDF/XML	Independent
Leigh University Benchmark[49]	SWAT Lab	1.1 Jun-04	GPL	Java	-	OWL	OWL	Independent
OWL-QL[50]	Robert Mccool	04-12-07 Dec-04	BSD	Java	-	-	-	Independent
OWLS-TC[51]	Klusch, M., Fries, B., and Khalid, M.	1.0 Apr-05	GPL	Java	-	-	-	Independent
ROWL[52]	Norman Sadeh	Jan-05	Propri.	Java	-	OWL RDF	-	-
Semantic Discovery Service[53]	Dan Mandell Sheila McIlraith	0.5 beta Jum-04	GPL	Java	-	OWL XML XML-Schema	-	MacOS Windows NT/2000 Linux
TAP[54]	Feigenbaum, E., Fikes, R., Guha, R., McGuinness, D., McIlraith, S., McCool, R., Miller, E., Brickley, D., Sundarajan, A., and Joly, K.	0.75 Apr-03	BSD	C	Model-centric Resource - centric	RDF/XML	RDF/XML	-

itself because it is known that even a very promising technology will not become a reality without the acceptance of its end-users. Finally, this study calls attention to the fact that the need for information exchange between different kinds of Semantic Web technologies is increasing every day and will prevail in the future. Most of current Semantic Web tools present different user interfaces and need constantly to map information resulting from each one. Therefore, it is crucial to create technologies that can be easily integrated among themselves.

REFERENCES

Berners-Lee, T. (1998). *What the Semantic Web can represent.* W3C (MIT, ERCIM, Keio). Retrieved May 5, 2005, from http://www.w3.org/DesignIssues/RDFnot.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*. Retrieved May 5, 2005, from http://www.scientificamerican.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21\&catID=2

Bizer, C., & Westphal, D. (2005). *Developers guide to semantic web toolkits for different programming languages.* Retrieved May 5, 2005 from http://www.wiwiss.fu-berlin.de/suhl/bizer/toolkits/02152005/

Golbeck, J., Alford, R., Baker, R., Grove, M., Hendler, J., Kalyanpur, A., et al. (2002, June). *Semantic Web tools from MIND SWAP*. Poster. International Semantic Web Conference (ISWC), Sardinia, Italy. Retrieved May 5, 2005, from http://iswc2002.semanticweb.org/posters/golbeck_a4.pdf

Guha, R., McCool, R., & Miller, E. (2003). Semantic search. In *WWW2003 - Proceedings of the 12th International Conference on World Wide Web* (pp. 700-709). ACM Press.

Halaschek-Wiener, C. (2004). *PhotoStuff overview*. MindSwap Research Lab. Retrieved May 5, 2005, from http://www.mindswap.org/2003/PhotoStuff/talks/psOverview.pdf

Hyvönen, E., Valo, A., Viljanen, K., & Holi, M. (2003). Publishing Semantic Web content as semantically linked HTML pages. In *Proceedings of XML Finland 2003 Conference*. Kuopio, Finland. Retrieved from May 5, 2005, http://www.cs.helsinki.fi/u/eahyvone/publications/xmlfinland2003/swehg_article_xmlfi2003.pdf

Kalyanpur, A., Parsia, B., & Hendler, J. (2005). A tool for working with Web ontologies. In *Proceedings of the International Journal on Semantic Web and Information Systems, 1*(1).

Table 8. Main features of API technologies

	Developer	Release	License	Language	API-Paradigm	Query-Languages	Model Storage	Input	Output	O.S.
Euler[55]	Jos De Roo	Fev-05	W3C	Java	Model-centric	N3QL	Memory	N3	N3	-
EulerMoz[56]	Doeblin, A., Roo, J., Alos, O., Hernandez, M., Sanchez, J.	Jan-05	W3C Mozilla	Java-Script	Model-centric	-	Memory	N3	N3	Indepen-dent
FaCT++[57]	Tsarkov, D. , Horrocks, I.	0.99	GPL	C++	-	-	-	-	-	Windows Linux
F-OWL[58]	Youyong Zou	-	BSD	Java Prolog	-	-	-	OWL RDF SWRL	OWL RDF	Linux
Hoolet[59]	-	-	LGPL	Java	-	-	-	OWL SWRL	-	Linux
Instance Store[60]	Turi, D., Bechhofer, S., Li, L., Lord, P. and Roberts, D.	1.4.1 Jul-04	LGPL	Java	-	-	-	OWL	-	Indepen-dent
Metalog[61]	Marchiori, M., Epifani, A., Trevisan, S., and Saarela, J.	2.1 Out-03	W3C	Python	Statement-centric Resource-centric	PNL	Memory File	RDF/XML N-Triples	RDF/XML N-Triples	-
OWLJess KB[62]	Joe Kopena	Jan-05	GPL	Java	Statement-centric Ontology-centric	Jess	Memory	RDF/XML	RDF/XML	-
Pellet[63]	Mindswap	1.1.0 Dec-04	W3C	Java	-	-	-	OWL RDF/XML	OWL RDF/XML	-
Pychinko[64]	Mindswap	0.1 Jan-05	Open Source	Python	Model-centric Statement-centric	N3	Memory	N3	N3	-
RdfLib[65]	Ian Davis James Carlyle	0.21 Jan-05	MIT	C#	Resource-centric Ontology-centric	-	Memory MySQL	RDF/XML N-Triples	RDF/XML N-Triples	-
Swish[66]	Graham Klyne	0.2.1 Fev-04	GPL	Haskell	Model-centric	-	Memory	N3	N3	-

Kogut, P., & Holmes, W. (2001, October 21). *AeroDAML: Applying information extraction to generate DAML annotations from Web pages*. First International Conference on Knowledge Capture (K-CAP 2001). Workshop on Knowledge Markup and Semantic Annotation, Victoria, BC. Retrieved May 5, 2005, from http://semannot2001.aifb.uni-karlsruhe.de/positionpapers/AeroDAML3.pdf

Marchiori, M. (2004). *Towards a people's Web: Metalog*. Retrieved May 5, 2005, from http://www.w3.org/People/Massimo/papers/2004/wi2004.pdf

Noy, N. F., Sintek, M., Decker, S., Crubezy, M., Fergerson, R. W. & Musen, M. A. (2001). Creating Semantic Web contents with Protege-2000. *IEEE Intelligent Systems 16*(2), 60-71.

Pereira, R.G., & Freire, M.M. (2005). Semantic Web. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (Vol. 2, pp. 917-974). Hershey, PA: Idea Group Reference.

Table 9. Main features of Repository technologies

	Developer	Release	License	Language	API-Paradigm	Query-Languages	Storage Model	Input	Output
3Store[67]	Riddoch, A., Gibbins, N. Harris, S., Beckett, D. Dawes, P.	2.2.18 Nov-04	GPL	C	Model-centric	RDQL OKBC	MySQL	-	-
Kowari[68]	Tucana Technologies	1.1.0 Dez-04	Mozilla	Java	Resource-centric Statement-centric Ontology-centric	iTQL	-	RDF/XML N-Triples	RDF/XML N-Triples
An Entry Sub-ontology of OWL Time[69]	Feng Pan	Dec-04	-	-	-	-	-	OWL	OWL
ParkaSW[70]	Mindswap	1.1b Apr-04	-	-	-	-	-	-	-
PySesame[71]	Tom Hoffman	0.1 Mar-04	LGPL	Python	-	-	-	-	-
RDFStore[72]	Alberto Reggiori	0.50 Aug-04	BSD	C	Model-centric	RDQL SPARQL	Memory File Berkeley	RDF/XML N-Triples XML	RDF/XML N-Triples XML
Sesame[73]	Aduna NLnet Foundation	1.1 Nov-04	LGPL	Java	Resource-centric Statement-centric	RDFS OWL-Lite	Memory File MySQL PostgreSQL Oracle Microsoft SQL Server	RDF/XML N-Triples N3 Turtle	RDF/XML N-Triples N3 Turtle
Time zone resource in OWL[74]	Feng Pan	Dec-04	LGPL	Python	-	-	-	OWL	OWL
YARS[75]	Andreas Harth Stefan Decker Hannes Gassert	Jan-05	BSD	Java	Model-centric	N3QL	-	RDF/XML N-Triples N3	RDF/XML N-Triples N3

Table 10. Main features of validation technologies

	Developer	Release	License	Language	Input	Output	O.S.
ConsVISor[76]	Versatile Information Systems	-	-	Java	OWL RDF	HTML OWL	Independent
OWL Validator[77]	Rager, D., Lerner, J., Self, T.	Jul-04	BSD	Java	OWL	Validation-Report XML	Independent
SWRL Validator[78]	Troy Self David Kolas	Nov-04	BSD	Java	SWRL	Validation-Report	Independent

Table 11. Main features of visualization technologies

	Developer	Release	License	Language	Input	Output	O.S.
IsaViz[79]	Emmanuel Pietriga	2.1 Out-04	-	Java	RDF/XML N3 N-Triples	RDF/XML N3 N-Triples SVG PNG	-
SVG-OWL Viewer[80]	Aditya Kalyanpur	3.0	BSD	Java	OWL	SVG	-
VisioOWL[81]	John Flynn	Jun-04	BSD	-	OWL RDF	Graph	Windows

Pereira, R.G., & Freire, M.M. (2006). *Integration of ontologies and semantic annotations with resource description framework in eclipse-based platforms withediting features for Semantic Web*.

SemWebCentral Web site. (2005). Retrieved May 5, 2005, from http://semwebcentral.org/index.jsp?page=home

Tallis, M. (2003, October 26). *Semantic word processing for content authors*. In Workshop Notes of the Knowledge Markup and Semantic Annotation Workshop (SEMANNOT 2003), Second International Conference on Knowledge Capture (K-CAP 2003), Sanibel, FL.

KEY TERMS

Ontology: The word *ontology* comes from the Greek *ontos* (being) and *logia* (written or spoken discourse). It is in use since Empedocles described the four elements – air, earth, fire, and water. In artificial intelligence, *ontology* is defined as a working model of concepts and interactions from a particular domain of knowledge, like medicine, mathematics, automobile repair, and so forth, which is used to easily describe the meaning of different contents that can be exchanged in information systems. Any *ontology* can be easily extended, refined, and reused by other *ontologies*, providing expressive representation for a wide diversity of real-world concepts.

OWL: The concept of Web Ontology Language, the most expressive of ontology languages currently defined for the Semantic Web. It has been developed by the W3C's Web Ontology Working Group and intended to be the successor of the DAML+May 5, 2005 OIL language. It is an extension of RDF Schema and a W3C Recommendation since 2004.

RDF: It is the concept of Resource Description Framework, a W3C Recommendation since 1999. It is a XML-based language that uses a triple-based assertion model and syntax to describe resources. RDF model is called "triple" because it can be described in terms of subject, predicate, and object, like grammatical parts of a sentence.

RDF API-Paradigm: There are four types of API-Paradigm for RDF:

- **Model-centric API:** Only allows loading, saving, and deleting whole RDF models.
- **Statement-Centric API:** RDF data is manipulated as a set of RDF triples each consisting of a subject, predicate, and object.
- **Resource-centric API:** Presents an RDF model as resources having properties.
- **Ontology-Centric API:** Adding direct support for the kinds of objects expected to be in an ontology: classes (in a class hierarchy), properties (in a property hierarchy), and individuals.

RDF Schema: Expresses a hierarchy class data model used for the classification and description of standard RDF resources. The role of RDF Schema is to facilitate the definition of metadata by providing a data model, much like many object-oriented programming languages, to allow the creation of data classes. It is a simple language that enables people to create their own RDF vocabularies in RDF/XML syntax.

Stovepipe Systems: A system where all the components are hardwired to only work with each other.

W3C Recommendation: A Recommendation W3C is interpreted by the industry and by the Web community, as being one synonym of normalization for the Web. Each Recommendation W3C is not more of the one than a steady specification developed by a Work group W3C (W3C Working Group) and reviewed by the members of the W3C. This type of recommendation promotes the interoperability of Web technologies from consensus gotten between the industry and the academy.

XML: The concept of Extensible Markup Language, a small set of rules in human-readable plaintext used to describe and share common structured platform-independent information. Its structure main components are elements and attributes of elements that are nested to create a hierarchical tree that can be easily validated. XML is *extensible* because, unlike HTML, anyone can define new tags and attribute names to parameterize or semantically qualify contents. It is a formal recommendation from W3C since 1998 playing an increasingly important role in the exchange of a wide variety of data on the Web.

ENDNOTES

1 World Wide Web Consortium (W3C) web site. Retrieved May 5, 2005 from: http://www.w3.org/
2 AeroText Semantic Web Automated Relation Markup (AeroSWARM). (Kogut & Holmes, 2001).
3 OntoMat-Annotizer. Retrieved May 5, 2005 from http://annotation.semanticweb.org/ontomat/index.html
4 PhotoStuff (Halaschek-Wiener, 2004).
5 RDF Instance Creator (RIC). (Golbeck et al., 2002).
6 SemanticWord. (Tallis, 2003).
7 Swangler. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/swangle/
8 CARA. Retrieved May 5, 2005 from http://cara.sourceforge.net/
9 Components for Ontology Processing (CODIP). Retrieved May 5, 2005 from http://codip.projects.semwebcentral.org/
10 HAWK. Retrieved May 5, 2005 from http://www.cse.lehigh.edu/~zhp2/hawk/readme.html
11 Kazuki. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/kazuki/
12 Named Graphs API for Jena (NG4J). Retrieved May 5, 2005 from http://www.wiwiss.fu-berlin.de/suhl/bizer/ng4j/
13 OWL API. Retrieved May 5, 2005 from http://sourceforge.net/projects/owlapi
14 OWL-S API. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/owl-s-api/
15 OWL Semantic Search Services. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/owl-semsearch/
16 Pyrple. Retrieved May 5, 2005 from http://infomesh.net/pyrple/
17 Simple Ontology Framework API (SOFA). Retrieved May 5, 2005 from https://sofa.dev.java.net/
18 Sparta. Retrieved May 5, 2005 from http://sparta-xml.sourceforge.net/
19 DumpOnt. Retrieved from: http://projects.semwebcentral.org/projects/dumpont/ (May 5, 2005)
20 FlinkCommands. Retrieved May 5, 2005 from http://flink.dcc.ufba.br/en/software/commands.html
21 HyperDAML. Retrieved May 5, 2005 from: http://projects.semwebcentral.org/projects/hyperdaml/
22 Object Viewer. Retrieved May 5, 2005 from: http://projects.semwebcentral.org/projects/objectviewer/
23 OWL Ontology for Protocols (OWL-P). Retrieved May 5, 2005 from: http://projects.semweb central.org/projects/owlp/
24 OWL-S IDE. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/owl-s-ide/ ()
25 Visual Variable-Depth Info Display. Retrieved from: http://www.wiwiss.fu-berlin.de/suhl/bizer/rdfapi/index.html
26 Java2owl-s. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/java2owl-s/
27 owl2dig. Retrieved May 5, 2005from http://projects.semwebcentral.org/projects/owl2dig/
28 Owl Converter. Retrieved May 5, 2005 from http://www.mindswap.org/~golbeck/code.shtml
29 owl-s2uddi. Retrieved May 5, 2005 from http://owl-s2uddi.projects.semwebcentral.org/
30 SWeHG. (Hyvönen, Valo, Viljanen, & Holi, 2003).
31 Wsdl2owl-s. Retrieved May 5, 2005 from http://www.daml.ri.cmu.edu/wsdl2owls/
32 Ontology engineeRIng ENvironmenT (ORIENT). Retrieved May 5, 2005 from http://apex.sjtu.edu.cn/projects/orient/
33 OWL Filetype Plugin For VIM. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/owl-vim/
34 OWL Mode for Emacs. Retrieved May 5, 2005 from http://owl-emacs.projects.semwebcentral.org/
35 Protégé. (Noy, Sintek, Decker, Crubezy, Fergerson, & Musen, 2001).
36 SMORE. Retrieved May 5, 2005 from http://www.mindswap.org/2005/SMORE/
37 Semantic Web Development Environment (SWeDE). Retrieved May 5, 2005 from http://owl-eclipse.projects.semwebcentral.org/
38 SWedt. (Pereira & Friere, 2006).
39 SWOOP. (Kalyanpur , Parsia, & Hendler, 2005).
40 4Suite. Retrieved May 5, 2005 from http://4suite.org/index.xhtml
41 Jena. Retrieved May 5, 2005 from http://jena.sourceforge.net/(
42 KAON. RetrievedMay 5, 2005 from http://kaon.semanticweb.org/
43 PerlRDF. Retrieved May 5, 2005 from http://www.gingerall.com/charlie/ga/xml/p_rdf.xml

[44] RDF API for PHP (RAP). Retrieved May 5, 2005 from http://www.wiwiss.fu-berlin.de/suhl/bizer/rdfapi/index.html

[45] RDF Gateway. Retrieved May 5, 2005 from: http://www.intellidimension.com/default.rsp?topic=/pages/site/products/rdfgateway.rsp

[46] Redland. Retrieved May 5, 2005 from http://librdf.org/

[47] Wilbur. Retrieved from http://wilbur-rdf.sourceforge.net/

[48] Cwm. Retrieved May 5, 2005 from http://www.w3.org/2000/10/swap/doc/cwm

[49] Lehigh University Benchmark. Retrieved May 5, 2005 from http://swat.cse.lehigh.edu/projects/lubm/index.htm

[50] OWL Query Language (OWL-QL). Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/owl-ql/

[51] OWL-S Service Retrieval Test Collection (OWLS-TC). Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/owls-tc/

[52] Rule Extension of OWL Mobile Commerce Lab (ROWL). Retrieved May 5, 2005 from http://www-2.cs.cmu.edu/~sadeh/mobilecomm.htm

[53] Semantic Discovery Service. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/sds/

[54] TAP. (Guha, McCool, & Miller, 2003).

[55] Euler. Retrieved May 5, 2005 from http://eulersharp.sourceforge.net/

[56] EulerMoz. Retrieved May 5, 2005 from http://sourceforge.net/projects/eulermoz

[57] FaCT++. Retrieved May 5, 2005 from http://owl.man.ac.uk/factplusplus/

[58] F-OWL. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/fowl/

[59] Hoolet. Retrieved from: http://owl.man.ac.uk/hoolet/

[60] Instance Store. Retrieved May 5, 2005 from http://instancestore.man.ac.uk/

[61] Metalog. (Marchiori, 2004).

[62] OWLJessKB. Retrieved May 5, 2005 from http://edge.cs.drexel.edu/assemblies/software/owljesskb/

[63] Pellet. Retrieved from: http://www.mindswap.org/2003/pellet/index.shtml

[64] Pychinko. Retrieved May 5, 2005 from http://www.mindswap.org/%7Ekatz/pychinko/

[65] RdfLib. Retrieved May 5, 2005 from http://www.semanticplanet.com/library?pagename =RdfLib.HomePage

[66] Swish. Retrieved May 5, 2005 from http://www.ninebynine.org/RDFNotes/Swish/Intro.html

[67] 3Store. Retrieved May 5, 2005 from http://sourceforge.net/projects/threestore/

[68] Kowari. Retrieved May 5, 2005 from http://www.kowari.org/

[69] An Entry Sub-ontology of OWL Time. Retrieved May 5, 2005 from http://entry-owl-time.projects. semwebcentral.org/

[70] ParkaSW. Retrieved May 5, 2005 from http://parkasw.projects.semwebcentral.org/

[71] PySesame. Retrieved May 5, 2005 from http://pysesame.projects.semwebcentral.org/

[72] RDFStore. Retrieved May 5, 2005 from http://rdfstore.sourceforge.net/

[73] Sesame. Retrieved May 5, 2005 from http://www.openrdf.org/

[74] Time zone resource in OWL. Retrieved May 5, 2005 from http://www.isi.edu/~pan/timezonehome page.html

[75] Yet Another RDF Store (YARS). Retrieved May 5, 2005 from http://sw.deri.org/2004/06/yars/yars.html

[76] ConsVISor. Retrieved May 5, 2005 from http://68.162.250.6/index.html

[77] OWL Validator. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/vowlidator/

[78] SWRL Validator. Retrieved May 5, 2005 from http://projects.semwebcentral.org/projects/swrl-val/

[79] IsaViz. Retrieved May 5, 2005 from http://www.w3.org/2001/11/IsaViz/

[80] SVG-OWL Viewer. Retrieved May 5, 2005 from http://www.mindswap.org/%7Eaditkal/svg_owl.shtml

[81] VisioOWL. Retrieved May 5, 2005 from http://web.tampabay.rr.com/flynn/VisioOWL/VisioOWL.htm

Category: Business Information Systems

Client Expectations in Virtual Construction Concepts

O.K.B Barima
University of Hong Kong, Hong Kong

INTRODUCTION

Meeting the expectations of clients through better service delivery has been a key concern of the construction industry over the years (Hui, 2005; Shen & Liu, 2004). One recommendation often suggested in recent studies to support the delivery of construction works to the construction client is the use of information and communication technology (ICT) (Weippert, Kajewski, & Tilley, 2003). In recent times the virtual construction concept has emerged where construction actors may rely on modern ICT tools to operate irrespective of time and space, to attain common value delivery goals in construction projects. For example, highly skilled construction parties may be in different physical geographic locations in the world, but they may use modern ICT tools to collaborate to achieve common project goals. The virtual construction concept has the potential to provide cost and time savings to the construction client, and it is also likely to play an important role in the delivery of construction works (Barima, 2003). A key party to the construction delivery process is the construction client, and it may be important to know the client's expectations in the use of the virtual construction concept. This knowledge may provide understanding on the potential expectations of construction clients and also assist construction service providers to improve on their value delivery systems to their clients. This chapter explores the potential expectations of construction clients in the virtual construction project environment. First, the background to this study is provided via review of previous literature, then the research methodology and key findings of this exploratory study are presented, before recommendations for future studies and the conclusions are given.

BACKGROUND

In recent years the customer (or client) has received attention in literature in various disciplines (Ellegaard, Johansen, & Drejer, 2003; Huang & Lin, 2002). Most of the studies have argued for paying attention to customers and their requirements, with the aim to either fulfill or exceed them, so as to create customer satisfaction (Huang & Lin, 2002; Winters, 2003). Recent perspectives on the customer appear to differ from traditional management perceptions, where there may be the orientation to focus on the internal transformation processes of the supplier.

In the construction industry the important role or needs of the construction client has also been directly or indirectly studied by scholars over the years (Briscoe, Dainty, Millett, & Neale, 2004; Hui, 2005; Kaya, 2004; Pries, Doree, Van Der Veen, & Vrijhoef, 2004; Shen & Liu, 2004; Winters, 2003). For example, Pries et al. (2004) have argued for client orientation in the construction industry. Briscoe et al. (2004) also suggested that construction clients are the influential drivers for innovation and performance improvement in the industry. According to a study by Pries et al. (2004) in spite of the arguments by scholars for client and market orientation in the construction industry, major industry leaders are still technology or project oriented.

Recent developments in the ICT sector in addition to changing perceptions have led to paradigm changes in the way businesses are executed (Barima, 2003). Varied management concepts have emerged, and one of these is the virtual concept in the construction industry, where actors may rely on modern ICT to operate independent of time and space to support the delivery of common goals (Barima, 2003). This model differs from traditional construction works delivery, which use physical delivery systems like face-to-face interactions, traditional mail delivery, and so forth.

The virtual concept is likely to play an important role in the construction industry in the future. However, research on the virtual concept (with a few exceptions see, e.g., Andresen, Christensen, & Howard, 2003; Rivard et al., 2004) have focused on the evolution of tools, software, pedagogic issues, and so on (Clayton, Warden, & Parker, 2002; Goh, 2005; Tse & Wong, 2004). As a relatively young research area little research has been done to explore the potential value delivery expectations of clients in the use of the virtual concept to support construction works delivery. Such exploration may improve understanding in this young area of research and also assist construction service providers to provide better service to the client. This chapter reports on an exploratory study on the potential client value delivery expectations in virtual construction projects. The next section provides the research methodology and findings of the study.

RESEARCH METHODOLOGY AND FINDINGS

Research Methodology

General

Although this report will emphasize the quantitative aspects of the study, it is worthy to note that a triangulated approach involving both qualitative and quantitative studies was used for this research (see, e.g., Mangan, Lalwani, & Gardner, 2004). This research approach was adopted to answer the research question on: what may construction clients potentially expect from service providers (or agents) in projects supported by the virtual concept. The mixed method approach provides (among others) the advantage of supporting the weaknesses of each of the mainstream research methods with the strengths of the other (Mangan et al., 2004).

After initial literature review, semi-structured in-depth interviews (1-2.5 hours long) were conducted with construction experts (in Hong Kong). The respondents were selected on purpose and deemed to have adequate knowledge on the subject. The interviews provided the opportunity to collate perspectives on what construction clients were likely to expect from their agents in the use of the virtual concept to support construction value delivery. The cross validated results of the qualitative studies were then used as key input for the subsequent exploratory quantitative studies.

The Sample

Stratified random sampling was used in the quantitative studies. The addresses of the potential respondents were obtained from the available professional addresses of the various construction-related associations (contractors; property developers; civil, architectural, mechanical/electrical engineering associations; and quantity surveyors in Hong Kong). The data collection method was via postal mail, and at the end of the data collection period 31 valid responses were received. This represents a response rate of about 8%. The respondents consisted of people deemed to have adequate knowledge of this study's subject, via built-in feedback questions in the scale (on their perceived knowledge of the subject).

The median experience of the respondents was about 16 years in construction. The modal single group (22.6% of the respondents) worked for property development companies. About 45% of the respondents worked for various consultancy companies (engineering, architectural, quantity surveying, facilities management, or multi-disciplinary), while 32.3% of the respondents worked for construction contracting or subcontracting companies.

Scale

The scale applied was a 5-point interval scale with potential scores which ranged from 1 (not important) to 5 (extremely important). Respondents were asked to provide scores on the perceived importance of the set of proposed items. The items were mainly generated from the qualitative studies. The items included in the scale may be broadly classified under: communications delivery and feedback systems; trust; tasks delivery skills and behavior for results delivery; and other management issues. The reliability (Cronbach alpha) of the 13-item scale used in the research was 0.9. This alpha value is greater than the minimum acceptable limit of 0.7, and hence the various subitems of the scale were accepted for further analysis (Herrmann, Tomczak, & Befurt, 2006).

Results of the study

General

The Statistical Package for the Social Sciences (SPSS) was used in the analysis of the data in this study. Unless otherwise stated all statistical tests were done at the 95% confidence level. Data analysis in this exploratory study also includes descriptive analysis, which uses the mean and standard deviations (SD). It must be noted that although the mean is a commonly used statistic, which could be understood easily, there is the tendency for mean scores to be affected by extreme values (Urdan, 2005).

Key Findings

The unit of analysis is the construction professional employed by various construction and client companies in Hong Kong. Table 1 shows a summary of the relative rankings of the proposed items (on the potential client demands in virtual construction project delivery) and their perceived mean scores and SDs. The prime perceived item is "enhanced communication delivery and skills in the virtual environment" with a mean of 4.1 and a SD of 0.71. The second ranked item is the display of "proven competence and dependability in delivering similar tasks in the virtual environment with minimum physical supervision" (mean = 3.93, SD=1.02). The third ranked item concerns the "proven assurance in protecting the shared information of other parties in the project" (mean=3.80, SD=0.81). Surprisingly placed at the fourth position is "enhanced productivity and effectiveness in results delivery" (mean=3.73, SD= 0.81).

The first and second ranked items are statistically not different from each other (t=-1.; df=29; p= 0.326, 2tail). The first and the third ranked items are marginally not significantly different (t=-1.964; df=29; p=0.06, 2tail). The second and third ranked items are also not significantly different from

Table 1. Summary of potential client expectations

Item	Description	Mean	SD*	Rank
a	Enhanced communication delivery and skills in the virtual environment.	4.10	0.71	1
b	Proven competence and dependability in delivering similar tasks/works in the virtual environment with minimum physical supervision.	3.93	1.02	2
c	Proven assurance in protecting the shared information of other parties in the project.	3.80	0.81	3
d	Enhanced productivity and effectiveness in product delivery.	3.73	1.05	4
e	Availing in real time accurate, timely, user-friendly project progress; cost; and other relevant project information.	3.63	0.81	5
f	Digital environment risk management capabilities and security consciousness.	3.57	0.73	6
g	Orientation to reasonably share information with other partners.	3.57	0.86	6
h	Enhanced communications and visualization of designed products.	3.57	0.77	6
j	Increased demands to leverage skills and best practices around the world coupled with cross-cultural delivery skills.	3.57	0.94	6
k	Proven corporate belief and commitment to digital tasks delivery and sustained capacity development.	3.50	0.82	10
l	Proven process efficiency and commitment to internal customer-focused quality orientation.	3.43	0.86	11
m	Costs and time reductions or savings.	3.40	1.07	12
n	Digital content quality delivery and customer-focused digital service delivery.	3.34	1.56	13

* SD (standard deviation)

each other (t=0.941; df=29, p=0.354, 2tail). There is, however, a statistically significant difference between the fourth and first ranked items (t=2.36; df =29; p=0.025, 2tail).

While the first ranked item is about communication delivery and skill in the virtual project environment, the second and third ranked items relate to issues which deal with trust. The fourth ranked item, however, concerns actual task delivery. Thus in other words softer items which could potentially enhance the delivery of actual construction tasks are perceived to be more important. It may however be better to perceive the items as intercorrelated. For example, the first ranked item "enhanced communication delivery and skills in the virtual environment" has a fairly strong and significant positive correlation with the fourth ranked item (enhanced productivity and effectiveness in product delivery) (r=0.591 at the 99% confidence level, 2tail). The first ranked item is also positively related to:

- The second ranked item (proven assurance in protecting the shared information of other parties in the project) (r= 0.397 at the 95% confidence level, 2tail).
- The third ranked item (proven competence and dependability in delivering similar tasks in the virtual environment with minimum supervision) (r=0.487 at the 99% confidence level, 2tail).

The correlations among the items may not necessarily imply causation, but they seem to suggest reasonable and meaningful relationships of association.

One important item (ranked fifth) is "availing in real time accurate and timely user-friendly project progress, cost, and other relevant information" (mean=3.63, SD=0.81). This fifth ranked item is seen as relatively more important than another item (h) which also concerns communication but is focused on enhanced product visualization. It may be noted that an "enhanced visualization and communication of designed products" (item h) could provide construction clients the opportunity to clearly see the nature of any proposed final construction products (in three or four dimensions). This visualization may enable clients to both appreciate and make effective contributions to the design of the product via clear descriptions of what they actually need.

Other issues like item j which deals with the use of international skills, best practice, and cross-cultural capabilities is in a joint sixth position with item f (digital environment risk management capabilities and security consciousness). The two items (j and f) are both seen as more important (in terms of the mean scores) than item k (proven corporate belief and commitment to digital tasks delivery and sustained capacity development).

The highest positive significant correlation among any two items in the 13-item scale was between the 10th ranked item k (proven corporate belief and commitment to digital tasks delivery and sustained capacity development) and the sixth ranked item g (orientation to reasonably share information with other partners) (r= .76 at the 99% confidence level). Thus, a very strong association exists between the expectations on the commitment by construction agents to digital task delivery, and the orientation of agents to reasonably share information with other partners.

Although two items (m and n) which concern "costs and time reduction or savings" and digital quality content and service delivery are in the respective positions of 12th and 13th, both of them have above average mean scores.

Implications of the Studies

Scholars have stressed the importance of trust and communication delivery in the general literature on the virtual concept (Hossain & Wigand, 2004). This exploratory study seems to suggest that apart from enhanced communication delivery and skills, other issues which concern trust (i.e., competence and dependability; ability to protect shared information of other parties) may also require comparable attention by service providers. Training all virtual construction participants/staff in effective virtual communication delivery and skills, as well as giving attention to other softer issues (on business efficacy) that concern trust, may be very useful. The study also appears to suggest that the expectation of construction clients on "cost and time savings," which may be potentially gained from the use of the virtual construction concept (Barima, 2003), is not seen as relatively more important than issues on enhanced communication and trust. The results of the study also seem to emphasize on the softer items which could potentially enhance the delivery of actual construction tasks. However, since all the proposed items in the study are perceived to be above average importance, a more useful approach may be via a robust focus (in which the items are seen as interrelated).

Limitations of the Studies

This study must be seen as exploratory in nature. Limitations like the size of the sample (and also nonresponse issues) used in the study do not make any statistical generalization useful, although statistically the sample size may be seen as large (Urdan, 2005). This exploratory study however provides important awareness on issues concerning trends and perceptions in this young area of research and practice.

FUTURE TRENDS

Future studies could focus on the evolution of robust virtual construction task delivery metrics to support construction service delivery improvement. Case studies may then be used across varying sizes/scales of construction project types to enhance understanding in this young area of research.

CONCLUSION

This article examined the potential expectations of construction clients for likely construction services in the virtual construction project environment. Although all the items in the study were seen as being above average importance, the study suggests that service providers may have to enhance communication delivery and skills, as well as display relevant behavior on soft issues which relate to trust. A better and useful position may be via adopting a robust systemic focus on the items (where the items are seen as interlinked). It is hoped that further studies could advance knowledge of the potential value delivery expectations in the use of the virtual concept to support construction works delivery.

REFERENCES

Andresen, J. L., Christensen, K., & Howard, R. W. (2003). Project management with a project web. *ITcon, 8,* 29-41.

Barima, O. K. B. (2003). An exploratory study of the usage of virtual project management in South Africa. *Proceedings of CIB TG 23 international conference*: *Professionalism in construction-culture of high performance* [CD ROM]. Hong Kong: International Council for Research and Innovation in Building and Construction.

Briscoe, G. H., Dainty, A. R. J., Millett, S. J., & Neale, R. H. (2004).Client-led strategies for construction supply chain improvement. *Construction Management & Economic, 22*(2), 193-201.

Clayton, M. J., Warden, R. B., & Parker, T. W. (2002). Virtual construction of architecture using 3D CAD and simulation. *Automation in Construction, 11*(2), 227-234.

Ellegaard, C., Johansen, J., & Drejer, A. (2003). Managing industrial buyer-supplier relations—The case for attractiveness. *Integrated Manufacturing Systems, 14*(4), 346-356.

Goh, B. H. (2005). IT barometer 2003: Survey of the Singapore construction industry and a comparison of results. *ITcon, 10,* 1-13.

Herrmann, A., Tomczak, T., & Befurt, R. (2006). Determinants of radical product innovations. *European Journal of Innovation Management, 9*(1), 20-43.

Hossain, L., & Wigand, R. T. (2004, November). ICT enabled virtual collaboration through trust [Electronic version]. *Journal of Computer-Mediated Communication, 10*(1). Retrieved December 14, 2005, from http://jcmc.indiana.edu/vol10/issue1/index.html

Huang, Y. S., & Lin, B. M. T. (2002). An empirical investigation of total quality management: A Taiwanese case. *The TQM Magazine, 14*(3), 172-180.

Hui, E. Y. Y. (2005). Key success factors of building management in large and dense residential estates. *Facilities, 23*(1-2), 47-62.

Kaya, S. (2004). Relating building attributes to end user's needs: "The owners-designers-end users" equation. *Facilities, 22*(9-10), 247-252.

Mangan, J., Lalwani, C., & Gardner, B. (2004). Combining quantitative and qualitative methodologies in logistics research. *International Journal of Physical Distribution & Logistics Management, 34*(7), 565-578.

Pries, F., Doree, A., Van Der Veen, B., & Vrijhoef, R. (2004). The role of leaders' paradigm in construction industry change. *Construction Management & Economics, 221*(1), 7-10.

Rivard, H., Froese, T., Waugh, L. M., El-Diraby, T., Mora, R., Torres, H., et al. (2004). Case studies on the use of information technology in the Canadian construction industry. *ITcon, 9,* 19-34.

Shen, Q., & Liu, G. (2004). Applications of value management in the construction industry in China. *Engineering, Construction and Architectural Management, 11*(1), 9-19.

Tse, T. C., & Wong, K. D. (2004). A case study of the ISO13567 CAD layering standard for automated quantity measurement in Hong Kong. *ITcon, 9,* 1-18.

Urdan, T. C. (2005). *Statistics in plain English* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Weippert, A., Kajewski, S. L., & Tilley, P. A. (2003). The implementation of online information and communication technology (ICT) on remote construction projects. *Logistics Information Management, 16*(5), 327-340.

Winters, P. M. (2003). What owners want from architects—And how to ensure that expectations are met. *Journal of Facilities Management, 2*(3), 276-284.

KEY TERMS

Client Orientation: Client orientation relates to the mental positioning matched by behavioral evidence of actors to deliver value to the client.

Client Value Expectations: Client value expectations refer to the potential characteristics of value which are expected to be delivered to the construction client.

Construction Paradigms Shift: Construction paradigms shift refers to the change in the traditional methods of thinking and execution of construction works.

Corporate Value Delivery Metrics: Corporate value delivery metrics are the collectively evolved measurable targets which may be used by the various parties of the construction value delivery process. They may be used to improve the construction delivery process.

Parties in a Construction Process: Various stakeholders who may have an interest in either the construction process or its product. Usually they may include construction contractors, consulting professionals (architects, engineers, quantity surveyors, etc.), the construction client, governments (if not the client), public or society, and so forth.

Virtual Construction Project: This is a construction project which is supported by the virtual concept.

Virtual Concept: Virtual concept refers to the reliance by actors on ICT tools to mimic the real world and operate independent of time and space to attain common goals.

Cluster Analysis Using Rough Clustering and k–Means Clustering

Kevin E. Voges
University of Canterbury, New Zealand

INTRODUCTION

Cluster analysis is a fundamental data reduction technique used in the physical and social sciences. It is of potential interest to managers in Information Science, as it can be used to identify user needs though segmenting users such as Web site visitors. In addition, the theory of Rough sets is the subject of intense interest in computational intelligence research. The extension of this theory into rough clustering provides an important and potentially useful addition to the range of cluster analysis techniques available to the manager.

Cluster analysis is defined as the grouping of "individuals or objects into clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters" (Hair, Black, Babin, Anderson, & Tatham, 2006). There are a number of comprehensive introductions to cluster analysis (Abonyi & Feil, 2007; Arabie, Hubert, & De Soete, 1994; Cramer, 2003; Everitt, Landau, & Leese, 2001; Gan, Ma, & Wu, 2007; Härdle & Hlávka, 2007). Techniques are often classified as hierarchical or nonhierarchical (Hair et al., 2006), and the most commonly used nonhierarchical technique is the *k*-means approach developed by MacQueen (1967). Recently, techniques based on developments in computational intelligence have also been used as clustering algorithms. For example, the theory of fuzzy sets developed by Zadeh (1965), which introduced the concept of partial set membership, has been applied to clustering (Abonyi & Feil, 2007; Dumitrescu, Lazzerini, & Jain, 2000). Another technique receiving considerable attention is the theory of rough sets (Pawlak, 1982), which has led to clustering algorithms referred to as rough clustering (do Prado, Engel, & Filho, 2002; Kumar, Krishna, Bapi, & De, 2007; Parmar, Wu, & Blackhurst, 2007; Voges, Pope, & Brown, 2002).

This article provides brief introductions to *k*-means cluster analysis, rough sets theory, and rough clustering, and compares *k*-means clustering and rough clustering. It shows that rough clustering provides a more flexible solution to the clustering problem, and can be conceptualized as extracting *concepts* from the data, rather than strictly delineated subgroupings (Pawlak, 1991). Traditional clustering methods generate *extensional* descriptions of groups (i.e., which objects are members of each cluster), whereas clustering techniques based on rough sets theory generate *intentional* descriptions (i.e., what are the main characteristics of each cluster) (do Prado et al., 2002). These different goals suggest that both *k*-means clustering and rough clustering have their place in the data analyst's and the information manager's toolbox.

BACKGROUND

k-Means Cluster Analysis

In the *k*-means approach, the number of clusters (*k*) in each partition of the data set is decided *prior to* the analysis, and data points are randomly selected as the initial estimates of the cluster centers (referred to as centroids). The remaining data points are assigned to the closest centroid on the basis of the distance between them, usually using a Euclidean distance measure. The aim is to obtain maximal homogeneity within clusters (i.e., members of the same cluster are most similar to each other) and maximal heterogeneity between clusters (i.e., members of different clusters are most dissimilar to each other).

K-means cluster analysis has been shown to be quite robust (Punj & Stewart, 1983). Despite this, the approach suffers from many of the problems associated with all traditional multivariate statistical analysis methods. These methods were developed for use with variables that are normally distributed and have an equal variance-covariance matrix in all groups. In most realistic data sets, neither of these conditions necessarily holds.

Rough Sets

The concept of rough sets (also known as approximation sets) was introduced by Pawlak (1982, 1991) and is based on the assumption that with every record in the information system (the data matrix in traditional data analysis terms), there is associated a certain amount of information. This information is expressed by means of attributes (variables in traditional data analysis terms) used as descriptions of the objects. For example, objects could be individual users in a study of user needs, and attributes could be characteristics of the users such as gender, level of experience, age, or other characteristics considered relevant. See Pawlak (1991) or Munakata (1998) for comprehensive introductions.

In rough set theory, the data matrix is represented as a table, the information system. The complete information system expresses all the knowledge available about the objects being studied. More formally, the information system is a pair, $S = (U, A)$, where U is a non-empty finite set of objects called the universe and $A = \{ a1, \ldots, aj \}$ is a non-empty finite set of attributes describing the objects in U. With every attribute $a \in A$, we associate a set Va such that $a : U \rightarrow Va$. The set Va is called the domain or value set of a. In traditional data analysis terms, these are the values that each variable can take (e.g., gender can be male or female, users can have varying levels of experience).

A core concept of rough sets is that of indiscernibility. Two objects in the information system about which we have the same knowledge are indiscernible. Let $S = (U, A)$ be an information system, then with any subset of attributes B, ($B \subseteq A$), there is associated an equivalence relation, INDA (B), called the B-indiscernibility relation. It is defined as:

$$\text{IND}A\ (B) = \{(x, x') \in U^2 \mid a \in B\ a(x) = a(x')\}$$

In other words, for any two objects (x and x') being considered from the complete data set, if any attribute a from the subset of attributes B is the same for both objects, they are indiscernible on that attribute. If $(x, x') \in$ INDA (B), then the objects x and x' are indiscernible from each other when considering the subset B of attributes.

Equivalence relations lead to the universe being divided into partitions, which can then be used to build new subsets of the universe. Two of these subsets of particular use in rough sets theory are the lower approximation and the upper approximation. Let $S = (U, A)$ be an information system, and let $B \subseteq A$ and $X \subseteq U$. We can describe the set X using only the information contained in the attribute values from B by constructing the B-lower and B-upper approximations of X, denoted $B_*(X)$ and $B^*(X)$ respectively, where:

$$B_*(X) = \{x \mid [x]B \subseteq X\} \text{ and } B^*(X) = \{x \mid [x]B \cap X \neq \varnothing\}$$

The set BN$B(X)$ is referred to as the boundary region of X, and is defined as the difference between the upper approximation and the lower approximation. That is:

$$\text{BN}B(\text{X}) = B^*(X) - B_*(X)$$

If the boundary region of X is the empty set, then X is a crisp (exact) set with respect to B. If the boundary region is not empty, X is referred to as a rough (inexact) set with respect to B. The important insight of Pawlak's work is his definition of a set in terms of these two sets, the lower approximation and the upper approximation. This extends the standard definition of a set in a fundamentally important way.

Rough Clustering

Rough clusters are a simple extension of the notion of rough sets. The value set (Va) is ordered, which allows a measure of the distance between each object to be defined, and clusters of objects are then formed on the basis of their distance from each other. An object can belong to more than one cluster. Clusters can then be defined by a lower approximation (objects exclusive to that cluster) and an upper approximation (all objects in the cluster which are also members of other clusters), in a manner similar to rough sets.

Let $S = (U, A)$ be an information system, where U is a non-empty finite set of M objects ($1 \leq i \leq M$), and A is a non-empty finite set of N attributes ($1 \leq j \leq N$) on U. The j^{th} attribute of the i^{th} object has value $R (i, j)$ drawn from the ordered value set Va.

For any pair of objects, p and q, the distance between the objects is defined as:

$$D(p,q) = \sum_{j=1}^{N} |R(p,j) - R(q,j)$$

That is, the absolute differences between the values for each object pair's attributes are summed. The distance measure ranges from 0 (indicating indiscernible objects) to a maximum determined by the number of attributes and the size of the value set for each attribute.

EXTENSIONS

The theory of rough sets continues to generate numerous edited books and conferences extending Pawlak's original insight into new areas of application and theory (e.g., An et al., 2007; Lin & Cercone, 1997; Polkowski & Skowron, 1998; Polkowski, Tsumoto, & Lin, 2000; Wang, Liu, Yao, & Skowron, 2003; Zhong, Skowron, & Ohsuga, 1999).

Rough *k*-Means Clustering

Lingras and West (2004) present a generalization of rough sets theory based on a relaxation of the basic equivalence relation (in fact most of the extensions of rough sets-based theory are based on this relaxation). They present a new technique combining k-means and rough set approaches by introducing a concept of upper and lower bounds to the k-means centroid. They applied this technique to a study of Web user behaviors in a first-year computer science class, and clearly identified three clusters of users—studious (continuous users), crammers (intermittent users, particularly before tests), and workers (users who mainly used access to

complete assignments). Lingras and West also found some overlaps between the clusters, a useful characteristic of all rough set-based clustering approaches.

Peters (2005, 2006) extended Lingras and West's technique to make it more compatible with the classical *k*-means algorithm when the upper and lower bounds converge, and to deal with outliers. He applied this improved technique to forest cover data and to cancer data. However, both the original technique and improved version have not solved the problem of the selection of initial parameters.

Rough Clustering Using an Evolutionary Algorithm

Voges and Pope (2004) developed a new technique for generating rough cluster descriptions using an evolutionary algorithm, with templates (Polkowski et al., 2000) as the building block of the data structure. The technique was applied to an analysis of a data set of perceptions of city destination image attributes (Voges, 2007) and was found to scale-up well to relatively large data sets (>6,000).

Clustering Different Data Types

Recently, rough clustering techniques have been applied to sequential data (Kumar et al., 2007) and categorical data (Parmar et al., 2007).

FUTURE TRENDS

Since Pawlak's original formulation of rough sets theory, theoreticians and practitioners have continued to explore and extend this rich field. Developments will continue in the underlying theory, the algorithms used for analysis, and the areas of application.

Developments in the underlying theory of rough sets include specifying an algebra of rough sets, formalizing rough reasoning, and extending rough topology. The underlying concepts will continue to be explored, including extending the definition of roughness through variable precision techniques. There will be a widespread utilization of fuzzy and rough sets as granulation sources within the field of Granular Computing, which provides a method of modeling that is far closer to the way human beings perceive their environment (Bello, 2008).

Analysis algorithms continue to be developed for rule extraction and the identification of reducts and cores. There will be an increasing use of hybrid applications, such as rough approximations of fuzzy sets and the use of rough set techniques in neural networks. New areas of applications continue to be explored, including language processing, video de-interlacing, image retrieval, face recognition, Web-based support systems, anomaly detection, text mining, and bioinformatics.

CONCLUSION

In rough clustering an object can belong to more than one cluster and therefore necessarily produce different solutions to *k*-means clustering, where an object can belong to only one cluster. This section briefly outlines a comparison between rough clustering and *k*-means clustering. A more detailed comparison can be found in Voges et al. (2002). See also Lingras and Huang (2005) for a comprehensive comparison of a number of clustering techniques.

In business, one of the most common applications of cluster analysis is the segmentation of a market by identifying homogeneous groups of buyers (Punj & Stewart, 1983; Wedel & Kamakura, 2000). Segmentation can also be applied to groups of users to assist in identifying user needs. In a market segmentation study that compared *k*-means and rough clustering analyses of shopping orientations of Internet shoppers, Voges et al. (2002) found that the two clustering techniques resulted in some clusters that were identified by both techniques, and some clusters that were unique to the particular technique used. The rough clustering technique also found clusters that were "refined" sub-clusters of those found by *k*-means clustering, and which identified a more specific sub-segment of the market.

Rough clustering produces more clusters than *k*-means clustering (Voges et al., 2002), with the number of clusters required to describe the data dependent on the interobject distance (D). It was found that the lower approximation of each cluster was dependent on the number of clusters selected for the solution. More clusters means an object has a higher chance of being in more than one cluster, which moves the object from the lower approximation to the boundary region and reduces the size of the lower approximation. This suggested that a number of factors needed to be considered when determining the best maximum value for D and the best number of clusters to include in the solution. A solution with too few clusters does not provide a useful interpretation of the partitioning of the data. On the other hand, too many clusters make interpretation difficult. In addition, the degree of overlap between the clusters needs to be minimized to ensure that each cluster provided information to aid in interpretation. Determining a good rough cluster solution requires a trade-off between these factors.

REFERENCES

Abonyi, J., & Feil, B. (2007). *Cluster analysis for data mining and system identification.* Basel: Birkhäuser.

An, A., Stefanowski, J., Ramanna, S., Butz, C.J., Pedrycz, W., & Wang, G. (Eds.). (2007, May 14-16). *Proceedings of the 11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing* (RSFDGrC 2007), Toronto, Canada. Berlin: Springer-Verlag.

Arabie, P., Hubert, L., & De Soete, G. (Eds.). (1994). *Clustering and classification.* River Edge, NJ: World Scientific.

Bello, R. (2008). *Granular computing: At the junction of rough sets and fuzzy sets.* Berlin: Springer-Verlag (Studies in Fuzziness and Soft Computing, vol. 224).

Cramer, D. (2003). *Advanced quantitative data analysis.* Philadelphia: Open University Press.

do Prado, H.A., Engel, P.M., & Filho, H.C. (2002). Rough clustering: An alternative to find meaningful clusters by using the reducts from a dataset. In J.J. Alpigini, J.F. Peters, J. Skowronek, & N. Zhong (Eds.), *Proceedings of the 3rd International Conference on Rough Sets and Current Trends in Computing* (RSCTC 2002). Berlin: Springer-Verlag.

Dumitrescu, D., Lazzerini, B., & Jain, L.C. (2000). *Fuzzy sets and their application to clustering and training.* Boca Raton, FL: CRC Press.

Everitt, B.S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). New York: Oxford University Press.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: Theory, algorithms, and applications.* Philadelphia: Society for Industrial and Applied Mathematics.

Hair, J.F., Black, B., Babin, B., Anderson, R.E., & Tatham, R.L. (2006). *Multivariate data analysis* (6th ed.). London: Prentice Hall.

Härdle, W., & Hlávka, Z. (2007). *Multivariate statistics: Exercises and solutions.* New York: Springer.

Kumar, P., Krishna, P.R., Bapi, R.S., & De, S.K. (2007). Rough clustering of sequential data. *Data and Knowledge Engineering, 63,* 183-199.

Lin, T.Y., & Cercone, N. (Eds.). (1997). *Rough sets and data mining: Analysis of imprecise data.* Boston: Kluwer.

Lingras, P., & Huang, X. (2005). Statistical, evolutionary, and neurocomputing clustering techniques: Cluster-based vs. object-based approaches. *Artificial Intelligence Review, 23,* 3-29.

Lingras, P., & West, C. (2004). Interval set clustering of Web users with rough K-means. *Journal of Intelligent Information Systems, 23,* 5-16.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L.M. Le Cam & J. Neyman (Eds.), *Proceedings of the 5th Berkeley Symposium on Mathematics, Statistics and Probability* (vol. 1, pp. 281-298). Berkeley: University of California Press.

Munakata, T. (1998). *Fundamentals of the new artificial intelligence: Beyond traditional paradigms.* New York: Springer-Verlag.

Parmar, D., Wu, T., & Blackhurst, J. (2007). MMR: An algorithm for clustering categorical data using rough set theory. *Data and Knowledge Engineering, 63,* 879-893.

Pawlak, Z. (1982). Rough sets. *International Journal of Information and Computer Sciences, 11,* 341-356.

Pawlak, Z. (1991). *Rough sets: Theoretical aspects of reasoning about data.* Boston: Kluwer.

Peters, G. (2005). Outliers in rough k-means clustering. In S.K. Pal et al. (Eds.), *PReMI* (pp. 702-707). Heidelberg: Springer-Verlag (LNCS 3776).

Peters, G. (2006). Some refinements of rough k-means clustering. *Pattern Recognition, 39,* 1481-1491.

Polkowski, L., & Skowron, A. (Eds.). (1998). *Proceedings of the 1st International Conference on Rough Sets and Current Trends in Computing* (RSCTC98). Berlin: Springer-Verlag.

Polkowski, L., Tsumoto, S., & Lin, T.Y. (Eds.). (2000). *Rough set methods and applications: New developments in knowledge discovery in information systems.* New York: Physica-Verlag.

Punj, G., & Stewart, D.W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20,* 134-148.

Voges, K.E. (2007). Rough clustering of destination image data using an evolutionary algorithm. *Journal of Travel and Tourism Marketing, 21,* 121-137.

Voges, K.E., & Pope, N.K.Ll. (2004). Generating compact rough cluster descriptions using an evolutionary algorithm. In K. Deb (Ed.), *Genetic and evolutionary computation.* New York: Springer-Verlag.

Voges, K.E., Pope, N.K.Ll., & Brown, M.R. (2002). Cluster analysis of marketing data examining on-line shopping orientation: A comparison of *k*-means and rough clustering approaches. In H.A. Abbass, R.A. Sarker, & C.S. Newton (Eds.), *Heuristics and optimization for knowledge discovery* (pp. 207-224). Hershey, PA: Idea Group.

Wang, G., Liu, Q., Yao, Y., & Skowron, A. (Eds.). (2003). *Proceedings of the 9th International Conference on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing* (RSFDGrC 2003). New York: Springer.

Wedel, M., & Kamakura, W.A. (2000). *Market segmentation: Conceptual and methodological foundations* (2nd ed.). Boston: Kluwer Academic.

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control, 8,* 338-353.

Zhong, N., Skowron, A., & Ohsuga, S. (Eds.) (1999). *New directions in rough sets, data mining, and granular-soft computing.* Berlin: Springer-Verlag.

KEY TERMS

Approximation Set: An alternative (and more technically correct) name for a rough set, which is defined by two sets, the lower approximation and the upper approximation.

Boundary Region: Those object that may or may not be in the approximation set. It is the difference between the upper approximation and the lower approximation. If the boundary region is empty, the set is said to be crisp. If the boundary region is not empty, the set is rough.

Cluster Analysis: A data analysis technique involving the grouping of objects into sub-groups or clusters so that objects in the same cluster are more similar to one another than they are to objects in other clusters.

***K*-Means Clustering:** A cluster analysis technique in which clusters are formed by randomly selecting *k* data points as initial seeds or centroids, and the remaining data points are assigned to the closest cluster on the basis of the distance between the data point and the cluster centroid.

Lower Approximation: In rough sets theory, one of the two sets used to define a rough, or approximate set. The lower approximation contains objects that are definitely in the approximation set.

Market Segmentation: A central concept in marketing theory and practice; involves identifying homogeneous sub-groups of buyers within a heterogeneous market. It is most commonly conducted using cluster analysis of the measured demographic or psychographic characteristics of consumers. Forming groups that are homogenous with respect to these measured characteristics segments the market.

Rough Classification: Finds mappings from the partitions induced by the equivalence relations in the condition attributes to the partitions induced by the equivalence relations in the decision attribute(s). These mappings are usually expressed in terms of decision rules. It performs the same type of classification function as discriminant analysis or logistic regression, where there is a known sub-grouping in the data set, which is identified by the decision attribute.

Rough Clustering: A simple extension of rough sets theory, analogous to traditional cluster analysis. The information table has no pre-existing subgroups, and clusters of objects are formed based on a distance measure. Clusters are defined by a lower approximation (objects exclusive to that cluster) and an upper approximation (all objects in the cluster that are also members of other clusters), in a manner similar to rough sets. An object can belong to more than one cluster.

Rough Set: The concept of rough, or approximation, sets was introduced by Pawlak and is based on the single assumption that information is associated with every object in an information system. This information is expressed through attributes that describe the objects; objects that cannot be distinguished on the basis of a selected attribute are referred to as indiscernible. A rough set is defined by two sets, the lower approximation and the upper approximation.

Upper Approximation: In rough sets theory, one of the two sets used to define a rough, or approximate set. The upper approximation contains objects that may or may not be in the approximation set. It can be formally defined as the union of the lower approximation and the boundary region. It is the complement of the set of objects definitely not in the set.

Clustering Algorithms for Data Streams

Christos Makris
University of Patras, Greece

Nikos Tsirakis
University of Patras, Greece

INTRODUCTION

The World Wide Web has rapidly become the dominant Internet tool which has overwhelmed us with a combination of rich hypertext information, multimedia data and various resources of dynamic information. This evolution in conjunction with the immense amount of available information imposes the need of new computational methods and techniques in order to provide, in a systematical way, useful information among billions of Web pages. In other words, this situation poses great challenges for providing knowledge from Web-based information. The area of data mining has arisen over the last decade to address this type of issues. There are many methods, techniques and algorithms that accomplish different tasks in this area. All these efforts examine the data and try to find a model that fits to their characteristics in order to examine them. Data can be either typical information from files, databases and so forth, or with the form of a stream. Streams constitute a data model where information is an undifferentiated, byte-by-byte flow that passes over the time. The area of algorithms for processing data streams and associated applications has become an emerging area of interest, especially when all this is done over the Web. Generally, there are many data mining functions (Tan, Steinbach, & Kumar, 2006) that can be applied in data streams. Among them one can discriminate clustering, which belongs to the descriptive data mining models. Clustering is a useful and ubiquitous tool in data analysis.

BACKGROUND

Data Mining and Knowledge Discovery

Classic algorithms handle small amounts of data and face up performance problems when data are huge in capacity. For example, a sorting algorithm runs efficiently with some megabytes of data but could have difficulties in running for some gigabytes of data. Many methods such as clustering and classification have been widely studied in the data mining community. However, a majority of such methods may not be working effectively on data streams. This happens because data streams provide huge volumes of data and at the same time require online mining, in which we wish to mine the data in a continuous fashion. Generally, there are many specific problems with traditional algorithms. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. In addition, it gives new opportunities for exploring and analyzing new types of data and for analyzing old types of data with new ways. Data mining is an integral part of knowledge discovery in databases (KDD). These two terms are often used interchangeably (Dunham, 2003). Over the last few years, KDD has been used to refer to a process consisting of many phases, while data mining is only one of these phases. Below are some definitions of knowledge discovery in databases and data mining (Fayyad, Piatetsky-Shapiro, &Smyth, 1996a, 1996b).

- **Knowledge discovery in databases (KDD):** Is the process for finding useful information and patterns in data.

Knowledge discovery in databases is a process that involves five different phases which are listed bellow (Dunham, 2003):

1. Data selection
2. Data preprocessing
3. Data transformation
4. Data mining
5. Data interpretation/evaluation

Data mining attempts to autonomously extract useful information or knowledge from large data stores or sets. It involves many different algorithms to accomplish different tasks. All these algorithms attempt to fit a model to the data. The algorithms examine the data and determine a model that is closest to the characteristics of the data being examined. These algorithms consist of three parts:

- **Model:** The purpose of the algorithm is to fit to the data.
- **Preference:** Some criteria must be used to fit one model over another.

- **Search:** All algorithms require some technique to search the data.

There are many different methods used to perform data mining tasks. These techniques not only require specific types of data structures, but also imply certain types of algorithmic approaches. Data mining tasks are generally divided into two different categories.

- **Predictive tasks:** These tasks predict the value of a particular attribute based on the values of other attributes. Predictive tasks include classification, regression, time series analysis and prediction.
- **Descriptive tasks:** Here, the objective is to derive patterns or relationships in data. Descriptive tasks include clustering, summarization, association rules and sequence discovery.

CLUSTERING

Clustering and other mining techniques have grasped the interest of the data mining community. It is alternatively referred to as unsupervised learning or segmentation. In broad strokes, clustering is the problem of finding a partition of a data set so that, under some definition of "similarity," similar items are in the same part of partition and different items are in different parts. These parts are called clusters. Items can be any type of data in any form. The most "common used" form of items in clustering are vectors. These vectors consist of *d*-dimensions in the Euclidian or generally metric space, so we talk about clustering in many dimensions. Some techniques meet some limitations in the value of *d* and most of the times *d* is the metric which differentiates algorithms.

Sometimes it is useful to use a threshold in order to specify that all objects in a cluster must be sufficiently close to one another. A more enlightening definition could be (Dunham, 2003):

- We have a set of alike elements. Elements from different clusters are not alike.
- The distance between points in a cluster is less than the distance between a point in the cluster and any point outside it.

A proportional process in data bases is segmentation where alike records are grouped together. Clustering is one of the most useful processes of data mining for cluster recognition and to define patterns and trends over data. It is similar to classification except that the groups are not predefined, but rather extracted by the specific distribution of the data.

Classification of Clustering Algorithms

There are various classes of clustering algorithms depending, each time, on the method for the clusters definition. The most common division of them is as *hierarchical* or *partitional*. In the first class of algorithms, we have a nested set of clusters. In addition, each level of hierarchy has a separate set of clusters. At the lowest level, each item is in its own unique cluster and at the highest level all items belong to the same cluster. Hierarchical clustering does not have as input the desired number of clusters. In the second class of algorithms, we meet only one set of clusters. Apart from these two classes of clustering algorithms, there are some recent studies that look at *categorical* data and are targeted to *large databases*. Algorithms targeted to large databases may adopt memory constraints by either *sampling* the database or using data structures which can be *compressed* or pruned to fit into memory regardless of the size of the database. Another approach of clustering algorithms is to further classify them based on the implementation technique used.

Data Streams

Traditional data bases store static data without the sense of time apart from the case the time is a part of the data. As time elapses the data explode and there is a need of online processing and analysis of data from many applications. This need made current methods deficient and processed data took another form and name. Data streams are data which change continuously over the time in a fast rate. There are many types of applications like network monitoring, telecommunications data managements, clickstream monitoring, manufacturing, sensor networks and many others, where data have the form of streams and not finite sum of data. In these applications users make continuous queries in contrast with the classic queries (one-time queries).

Models

We consider an input stream with x_i items that arrives sequentially and describes a signal X, a one-dimensional function X: $[1....N] \rightarrow R^2$. Models differ on how x_i's describe X and can be divided into three categories (Muthukrishnan, 2003):

- **Time series model:** This model is suitable for time series data such as observing the traffic at an IP link for a predefined time horizon (e.g., every 5 minutes).
- **Cash register model:** This is the most popular data stream model. It is suitable for applications such as monitoring IP addresses that access a Web server and source IP addresses that send packages to a link.
- **Turnstile model:** This is the most general model and it is appropriate in order to study fully dynamic situations where there are inserts as well as deletes, but it is often hard to get interesting bounds in this model.

The above models can be sorted according to their generality in decreasing order as follows: Turnstile, Cash Register and Time Series. More information about these models can be found in Gilbert, Kotidis, Muthukrishnan and Strauss (2001a). Finally, if we want to compute various functions on the signal at different times during the stream there are some different performance measures:

- The processing time per item x_i in the data stream.
- The space used to store the data structure on X_t at time t.
- The time needed to compute functions on X.

Applications

There are many applications that handle data streams. *Sensor networks* are used in many control applications that make complicated processes of filtering. These applications require aggregation and merging of multiple data streams in order to provide analysis from different sources of data. These applications usually collaborate with DSMSs (Data Stream Management Systems) in order to be controlled by a central mechanism, and the most common technique there is special purpose Query Languages. *Network monitoring* is a category of applications where systems do analysis as far as concerning the network traffic in a real time manner and are used to compute statistical methods in order to identify strange situations (Cranor et al., 2002; Gilbert, Kotidis, & Muthukrishan, 2001b; Sullivan & Heybey, 1998). There are DSMSs that provide long-running continuous queries which are required here. *Financial tickets* are another type of application where we meet online analysis of financial data like association discovery, trend recognition and prediction of future values. Finally, transaction analysis systems are systems that monitor all transactions that take place through Internet, telecommunications and banks. The aim of these systems, that handle data streams, is to find patterns of user behavior that are interesting and to recognize special and insecure user movements.

Some notable data stream managements systems (DSMS) are *Aurora* (Abadi et al., 2003) and *QuickSand* (Gilbert et al., 2001b). *Aurora* is a system for monitoring applications developed in consultation with defense, financial, and natural science communities. The core of the Aurora system consists of a large network of triggers. It deals with large numbers of data streams and users can build queries out of a small set of operators (called boxes) via a friendly user interface. Aurora consists of three components. A GUI environment where tuple structures and Aurora flow networks are defined, and also a server that executes an Aurora network. The inputs and outputs of the Aurora server are streams of tuples, delivered over TCP/IP sockets. Finally, Aurora has another GUI performance monitor that shows the quality of service being provided by the server at a given moment. The second system, *QuickSand,* gives solutions for monitoring and analyzing track data generated from large ISP networks. More precisely, it builds compact summaries of the track data called sketches at distributed network elements and centers. These sketches are able to respond well to queries that seek features that stand out of the data. Sketches use small space, they can be computed as data streams, and can be combined across distributed sites.

Clustering Data Steams

Clustering as mentioned above can be defined as the process of grouping a collection of N patterns into distinct segments or clusters based on a suitable notion of closeness or similarity among these patterns. As a procedure, it can be applied in many fields and in many daily problems. Some outstanding examples are clustering for understanding and clustering for usability. Good clusters show high similarity within a group and low similarity between patterns belonging to two different groups. The data stream model has been defined for new classes of application involving massive data being generated at a fast pace. Both data streams and online or incremental models are similar in that they require decisions to be made before all the data are available. For this purpose cluster analysis on data streams becomes more difficult, because the data objects in a data stream must be accessed in order and can be read only once or a few times with limited resources. There are some notable works in this area with algorithms that have been developed for data streams analysis. Many methods are based on traditional data mining techniques and provide ways to analyze this type of contemporary data. Clustering techniques are the most common techniques in this analysis phase as they can provide better and useful results to the analysts.

In the bibliography there is much research in streaming algorithms. Problems that can be solved in a small space when the data is a stream include (Guha et al., 2003): frequency estimation, norm estimation, order statistics, synopsis structures, time indexed data, and signal reconstructions.

One of the most interesting approaches to the clustering problem for data stream applications is (Aggarwal et al., 2003) where a framework is proposed for clustering evolving data streams. They divide the procedure of clustering into two phases, an online phase where periodically summary statistics are stored and an off-line phase which uses this summary statistics. The off-line phase is utilized by the analyst who can use a variety of inputs in order to provide a quick understanding of the broad clusters in the data stream. This two-phased approach provides the user with the flexibility to explore the nature of the evolution of the clusters over different time periods. One of the most important issues that this research is studding is the fact that a data stream cannot be revised over the time. Hence, the clustering algorithm must have a basic attribute, to be one pass algorithm. This means that this algorithm maintains a substantial amount of

information so that important details are not lost. K-means is the most common clustering algorithm for this purpose (MacQueen, 1967) and there is rich literature about this algorithm and its variants.

Clustering Click Stream Data

A special and interesting kind of data streams is click stream data. This kind of streams is centralized to user behavior. A visitor in a Web site is characterized by a set of one or more sessions. Each session is comprised by the sequence of pages visited, the time spent on each page, the information typed in and so forth. This information is huge when we are referring to popular Web sites, and at the same time this information is rapid and has the form of a stream. Consequently, clustering visitors based on such clickstream information can help a Web site to provide customized content for the users, suggest additional content from other user's preferences and generally separate users according to some similarity measures.

In Makris and Tsirakis (2006) the authors have proposed a model for clustering clickstream data. The main purpose of the model is to provide different kinds of data mining results from user behavior and to form this idea through a fast and efficient way. The model is based on three different states of data processing. First, click streams are being collected in real time manner and useful information is stored in a compressive way. A clustering procedure works on these stored data and provides clusters of users (short-term results), and finally a metaclustering procedure analyses the changes of these clusters over the time and provides long-term clustering results. The second step of the model is being designed in a way to support simultaneous data mining techniques such as different clustering algorithms of other mining techniques. With this opportunity, there is a variety of results and data inputs for the final metaclustering procedure.

In the field of Web usage mining, where the click stream clustering belongs to, there are many research studies. Clustering users based on their Web access patterns is the main concern and remains an active area. Among them marks out (Garofalakis et al., 1999) where the most popular techniques and algorithms for discovery Web, hypertext and hyperlink structure are reviewed. In Fu, Sandhu and Shih (1999) authors study the clustering of sessions based on browsing activities or access patterns on the Web, while Cadez et al. (2000) present a methodology for visualizing navigation patterns on a Web site using model-based clustering and describe a tool based on this technology. In Phoha, Iyengar, and Kannan (2002) the authors propose a new learning algorithm for fast Web page allocation on a server using the self-organizing properties of the neural network. One interesting aspect of mining clickstream data is models that predict user behaviour. Finally, in Gunduz and Ozsu (2003) the authors study the problem of modeling the behaviour of a Web user by introducing a similarity metric to find pair-wise similarities between user sessions. By partitioning user sessions based on that similarity metric, they propose a tree construction for representing clusters. This can be applied in Web sites with different structure.

FUTURE TRENDS

Clustering is a fundamental data analysis step and has been studied widely for decades in many disciplines. Considering the fact that the amount of data is continuously increasing in real life applications, efficient mining on data streams becomes a challenge to existing data mining techniques and algorithms. A wide research activity has taken place in the data mining area, but a few have been examined in the context of stream data clustering. Some open research issues are: detection of cluster-based outliers and semantic outliers in the data stream environment, new scalable clustering techniques that preserve privacy issues and ways to provide results with usability perspectives and finally, new techniques that deal with distributed data streams.

CONCLUSION

Efficient mining on data streams becomes a challenge to existing data mining algorithms, partially because of the high cost on both storage and time of distributed computation. Clustering data streams constitutes one of the most powerful data mining methods in data mining area. Only few of the techniques developed scale to support clustering of very large data sets. Data mining applications and the associated data sets have brought new clustering challenges. This article highlighted some of these issues and also described recent advances for successfully addressing them with great interest in click stream data.

REFERENCES

Abadi, D., Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Erwin, C., et al. (2003). Aurora: A data stream management system. In *Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'03),* San Diego, CA.

Aggarwal, C., Han, J., Wang, J., & Yu, P. P. S. (2003). A framework for clustering evolving data streams. In *Proceedings of the 2003 International Conference on Very Large Data Bases (VLDB'03),* Berlin, Germany.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2000). Visualization of navigation patterns on a Web site using model-based clustering. In *Proceedings of the Sixth*

International KDD Conference, (pp. 280-284).

Cranor, C.D.,Gao, Y., Johnson, T., Shkapenyuk, V., & Spatscheck, O. (2002). Gigascope: High performance network monitoring with an SQL interface. In *Proceedings of the ACM SIGMOD Conference, 2002.*

Dunham, M.H. (2003). Data *mining: Introductory and advanced topics*. Prentice Hall.

Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996a). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Fayyad, U.M., Piatetsky-Shapiro, G., & Smyth, P. (1996b). From data mining to knowledge discovery: An overview. *Advances in knowledge discovery and data mining.* Menlo Park, CA: American Association for Artificial Intelligence.

Fu, Y., Sandhu, K., & Shih, M. (1999, August). A generalization-based approach to clustering of Web usage sessions. In *Proceedings of WEBKDD 1999*, San Diego, CA, (pp. 21-38).

Garofalakis, M.N., Rastogi, R., Seshadri, S., & Shim, K. (1999). Data mining and the Web: Past, present and future. In *Proceedings of the 2nd International Workshop on Web Information and Data Management*, (pp. 43-47).

Ghosh, J. (2003). Scalable clustering methods for data mining. *Handbook of data mining*. Lawrence Erlbaum.

Gilbert, A.C., Kotidis, Y., & Muthukrishnan, S. (2001b). *QuickSAND: Quick summary and analysis of network data* (DIMACS Tech. Rep. No. 2001-43).

Gilbert, A.C., Kotidis, Y., Muthukrishnan, S., & Strauss, M. (2001a). *Surfing wavelets on streams: One-pass summaries for approximate aggregate queries*, VLDB 2001.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., & O'Callaghan, L. (2003). Clustering data streams: Theory and practice. *TKDE Special Issue on Clustering, 15*.

Gunduz, S., & Ozsu, M. (2003). A Web page prediction model based on click-stream tree representation of user behavior. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 535-540).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1965, (pp. 281-297). University of California Press.

Makris, C., & Tsirakis, N. (2006). *A model for clustering clickstream data* (manuscript).

Muthukrishnan, S. (2003). Data streams: Algorithms and applications. In *Proceedings of the 2003 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'03)*, Baltimore, MD, (pp. 413-413).

Phoha, V. V., Iyengar, S.S., & Kannan, R. (2002, December). Faster Web page allocation with neural networks. *IEEE Internet Computing, 6*(6), 18-26.

Sullivan, M., & Heybey, A. (1998). Tribeca: A system for managing large databases of network traffic. In *Proceedings of the USENIX Annual Technical Conference*.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Addison-Wesley.

KEY TERMS

ClickStream: Series of page visits and associated clicks executed by a Web site visitor when navigating through the site. As the user clicks on a link on a Web page, the action is logged inside the Web server, as well as possibly the Web browser, routers, proxy servers, and ad servers.

Clustering: Clustering is an algorithmic concept where data points occur in bunches, rather than evenly spaced over their range. A data set which tends to bunch only in the middle is said to possess centrality. Data sets which bunch in several places do not possess centrality. What they do possess has not been very much studied, and there are no infallible methods for locating the describing more than one cluster in a data set (the problem is much worse when some of the clusters overlap).

Data Bases: A database is a collection of information stored in a computer in a systematic way, such that a computer program can consult it to answer questions. The software used to manage and query a database is known as a database management system (DBMS). The properties of database systems are studied in information science.

Data Mining: Is the process of autonomously extracting useful information or knowledge from large data stores or sets. Data mining can be performed on a variety of data stores, including the World Wide Web, relational databases, transactional databases, internal legacy systems, pdf documents, and data warehouses.

Data Streams: An undifferentiated, byte-by-byte flow of data. A data stream can be distinguished in practice from a block transfer, although the moving of blocks could itself be considered a "stream" (of coarser granularity).

Knowledge Discovery: Is the process of finding novel, interesting, and useful patterns in data.

Synopsis Data Structures: Are data structures that use very little space, can be any data structures that are substantively smaller than their base data sets. The design and analysis of effective synopsis data structures offer many algorithmic challenges.

Web Mining: Is the application of data mining techniques to discover patterns from the Web. According to analysis targets, Web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.

Cognitive Research in Information Systems

Felix B. Tan
Auckland University of Technology, New Zealand

M. Gordon Hunter
University of Lethbridge, Canada

INTRODUCTION

The existence and significance of cognition in organizations and its influence on patterns of behaviour in organizations and organizational outcomes are increasingly accepted in information systems (IS) research (Barley, 1986; DeSanctis & Poole, 1994; Griffith, 1999; Griffith & Northcraft, 1996; Orlikowski, 1992, 1994 #208). However, assessing the commonality and individuality in cognition and eliciting the subjective understanding of research participants either as individuals or as groups of individuals remain a challenge to IS researchers (Orlikowski & Gash, 1994). Various methods for studying cognition in organizations have been offered - for example, clinical interviewing (Schein, 1987), focus groups (Krueger, 1988), discourse-based interviewing (Odell, Goswami & Herrington, 1983). This article proposes that cognition applied to making sense of IT in organizations can also be explored using Kelly's (1955) Personal Construct Theory and its methodological extension, the Repertory Grid (RepGrid). The RepGrid can be used in IS research for uncovering the constructs research participants use to structure and interpret events relating to the development, implementation, use and management of IS in organizations.

In the context of this article, cognition is considered to be synonymous with subjective understanding: "the everyday common sense and everyday meanings with which the observed human subjects see themselves and which gives rise to the behaviour that they manifest in socially constructed settings" (Lee, 1991, p. 351). Research into cognition in organizations investigates the subjective understanding of individual members within the organization and the similarities and differences in the understandings among groups of individuals (Jelinek & Litterer, 1994; Porac & Thomas, 1989). In IS research, it is the personal constructs managers, users and IS professionals use to interpret and make sense of information technology (IT) and its role in organizations. The discussion here outlines the myriad of ways the RepGrid can be employed to address specific research objectives relating to subjective understanding and cognition in organizations. It illustrates, from a variety of published studies in IS (see Table 1), the flexibility of the RepGrid to support both qualitative and/or quantitative analyses of the subjective understandings of research participants.

BACKGROUND

We propose to use a framework to facilitate this discussion (see Figure 1) that presents a two-dimensional view of the types of research using the repertory grid. The examples in Table 1 are mapped along these two dimensions.

Theory-Focused vs. Method-Focused

On one dimension, we distinguish research that applies Kelly's (1955) personal construct theory (theory-focused)

Figure 1. Distinguishing research using the repertory grid

Theory-Focused	Hunter (1997)[R]	Latta and Swigger (1992)[=]
Method-focused	Moynihan (1996)[R]	Phythian and King (1992)[=]

[R] Idiographic (i.e. individual interpretations – unique grids)
[=] Nomothetic (i.e. group interpretations – common grids)

from those applying the repertory grid method, without delving into the conceptual underpinnings of the theory (method-focused). When introduced some 45 years ago, the repertory grid technique served as the methodological extension of the personal construct theory. It operationalizes key aspects of Kelly's fundamental postulate and corollaries. IS researchers interested in the subjective understandings of individuals will find the repertory grid a powerful tool that permits the study of the individual's construct system and provides richer cognitive insights into research findings. For example, Latta and Swigger (1992) validated the use of the repertory grid for representing commonality of construing among participants regarding the design of intelligent user interfaces. The study lent strong support to the commonality corollary in grids, which can be confidently used to represent a consensus of knowledge around a problem domain. Hunter (1997) used the laddering technique to elicit what Kelly termed as super-ordinate constructs – constructs that are core to the individual's system of interpretation.

In contrast, there is research that has accepted Kelly's theory and employed the repertory grid solely as a data gathering technique. These works have employed the utility of the technique purely for its methodological strengths. Stewart and Stewart (1981) suggest, "At its simplest, Grids provide a way of doing research into problems – any problems – in a more precise, less biased way than any other research method" (pp. vii). These authors further contend that the repertory grid "….enables one to interview someone in detail, extracting a good deal of information … and to do this in such a way that the input from the observer is reduced to zero" (p. 5). Two of the examples in Table 1 have taken the method-focused approach to the use of the repertory grid technique. For instance, Moynihan (1996) was purely interested in using the repertory grid technique to collect data and to compare the results with the extant literature. Moynihan argued that the free-ranging responses resulting from the non-prohibitive nature of the technique permitted the participants to apply the "theories-of-action" (theories individuals use to guide their actions) they employ daily – resulting in the identification of themes and issues over and above the extant literature. In the studies by Phythian and King (1992), the repertory grid was used to explore the similarity and differences in the views between individual managers. No direct references were made to Kelly's personal construct theory, as the focus was to identify key factors influencing tender decisions and the relationships among these factors by interviewing two managers closely involved in such tender activities.

Qualitative vs. Quantitative

On the second dimension, we distinguish research that is either qualitative or quantitative. The identification of emerging themes from elicited constructs is common in a qualitative approach using the repertory grid. For example, Hunter (1997), when investigating how certain groups of individuals interpreted the qualities of "excellent" systems analysts, employed content analysis of the data gathered from individual interviews conducted using the repertory grid technique. The numeric component of the grid was only employed to conduct visual focusing at the end of each interview as a means of quickly assessing what had transpired during the interview and whether the research participant agreed with this initial assessment. Similarly, Moynihan (1996) employed the repertory grid technique as a method to elicit interpretations from research participants of what aspects were considered important when deciding upon an approach to adopt for projects to be conducted for external clients. Unique grids were developed for each research participant. Then the data were analyzed from a qualitative perspective via content analysis at the construct level, where emerging themes were identified and categorized. In these examples, the researchers took an open view toward gathering data and allowed themes or categories to emerge from the data as the investigation proceeded.

In contrast, the quantitative approach utilizes mathematical and/or statistical analyses of grid data (Daniels, Markoczy & de Chernatony, 1994). These techniques are commonly used to explore the structure and content of an individual's construct systems or make comparisons between groups of individuals (Ginsberg, 1989). This approach was adopted by two of the examples in Table 1. For instance, in Phythian and King (1992), statistical analyses (specifically, cluster analysis and correlation analysis) were conducted on individual and combined grids. These data were used to support the development of an expert support system. Similarly, Latta and Swigger (1992) applied cluster analysis and Spearman's rank order correlation to analyze the grids. The study revealed an overall correlation between the students' and the instructor's grids, promoting the utility of the repertory grid technique in modeling knowledge relating to the design of information systems.

Idiographic vs. Nomothetic

Within both the qualitative and quantitative perspectives, research using the repertory grid technique is either idiographic or nomothetic in nature. The idiographic approach focuses on the subjective experiences of the individual and presents results in expressions and terms used by the individual. The resulting grid is considered unique in that there are no common elements or constructs employed in the elicitation process. For example, in the study of systems analysts, each participant was asked to name up to six systems analysts with whom s/he had interacted. In this project, Hunter (1997) provided a role description (i.e., system analysts interacted with) and asked each participant to specify examples that fit this category. The analysts named were not common

Table 1. Examples of IS research using the RepGrid

Table 1. Examples of IS Research Using the RepGrid

	Hunter (1997)	Moynihan (1996)	Phythian and King (1992)	Latta and Swigger (1992)
Research Objectives	Explore the qualities of "excellent" systems analysts	Identify the situational factors considered in the planning/running of new systems development projects	Develop rules for an expert system to support customer tender evaluations	Validate the RepGrid in modeling communal knowledge regarding design of system interfaces
Research Perspective	Qualitative	Qualitative	Quantitative	Quantitative
Nature of RepGrid	Idiographic	Idiographic	Nomothetic	Nomothetic
Key Findings	Several themes considered as qualities of "excellent" systems analysts	Identified themes over and above literature. Differences in project managers' construction of project contexts	Identified key factors and rules influencing tender decisions. Expert system improved consistency	Commonality of constructions support the use of the RepGrid to model group knowledge
Research Design:				
Element Selection	Systems analysts with whom participant has interacted	Systems development projects on which participant has worked	Previous customer tender enquiries	Components of online bibliographic retrieval systems
	Elicited	Elicited	Supplied	Supplied
Construct Elicitation	Qualities of "excellent" systems analysts	Situational factors influencing risks in new systems projects	Key factors and rules influencing tender decisions	Attributes of system interface design
	Minimum context form (triadic sort) and laddering	Minimum context form (triadic sort)	Minimum context form (triadic sort) and laddering	Minimum context form (triadic sort) and supplied constructs
Linking	Rating	None	Rating (Grid) Ranking (Elements)	Rating
RepGrid Analysis	Content analysis Visual focusing COPE and VISA	Content analysis	Cluster analysis (FOCUS), correlation, mathematical modeling	Cluster analysis, correlation
Sample and Size	53 (users and IT professionals) from two insurance companies	14 systems development project managers	Two manager-experts involved in assessing tender enquiries	Instructor and students who completed an "information search and retrieval" course

among participants and as such the resulting grids were not common in terms of the elements used. Similarly, Moynihan (1988) asked each participating project manager to make a list of systems development projects s/he had worked on as a project manager. If the project manager named more than nine projects, s/he was then asked to choose the three that were most successful, the three that were the least successful, and three in between. Moynihan's research objective was to identify the situational factors project managers regard as important when planning new development projects and not to compare the subjective understandings of different project managers. As such, he did not supply a common set of systems development projects that would have permitted a comparative analysis of individual grids.

In contrast, research comparing the grids of individuals or groups of individuals requires different decisions to be made concerning the elements and constructs in the repertory grid process. This nomothetic approach necessitates the use of a common set of elements and/or constructs to permit comparisons to be made between grids (Easterby-Smith, 1980). Such research also tends to be quantitative in nature. For example, research to identify the similarities and differences in two managers' views on tender enquiry evaluations imposed a common set of tender enquiries as elements in the repertory grid process (Phythian & King, 1992). This permitted the comparison of the construct systems of the two managers based on their personal experiences of similar events. In another example, a set of constructs was elicited from an instructor and then supplied as common constructs for a group of students to evaluate against a prescribed set of elements representing the features of online bibliographic retrieval systems (Latta & Swigger, 1992). This permitted the commonality of construing among students and between students and the instructor to be tested. In these examples, the reason for the use of common components in the repertory grid process was to compare grids.

Finally, none of the examples discussed in this section approached their studies using both qualitative and quantitative approaches. This does not imply that the repertory grid cannot lend itself to both qualitative and quantitative analysis of the collected data.

CONCLUSION

This article has demonstrated the flexibility of the repertory grid technique to support both qualitative and/or quantitative approaches to investigating the subjective understanding of human subjects in complex and socially dynamic organizations. With this technique, both qualitative and quantitative approaches need not be mutually exclusive. For instance, Hunter (1997) approached his study of systems analysts from a grounded theory perspective. The data collected using the repertory grid were analysed to identify themes or qualities of "excellent" systems analysts. In an extension of this investigation, the characteristics of excellent systems analysts were statistically tested to highlight the similarities and differences in the subjective understandings of Canadian and Singaporean research participants (Hunter & Beck, 1996).

We would like to conclude by reiterating two important points. The first is a word of caution. The repertory grid should not be considered a panacea to investigating the subjective understanding of individuals in an organizational setting. It can be used in conjunction with other methods – as a means of validating other techniques or as a preliminary phase to further interpretive or positivist investigations. The second is that the personal construct theory is one of several theories in cognitive science (Berkowitz, 1978). The repertory grid is one of several cognitive mapping methods available to the IS researcher (Huff, 1990). This article was written in an attempt to stimulate interest in the IS research community of the need for more cognitive emphasis in our field. Hopefully, IS researchers will be encouraged to further reflect on the virtues of applied theories and methods that can deliver utilizable and consumable outcomes to research and practice in IS.

REFERENCES

Barley, S.R. (1986). Technology as an occasion for structuring: Evidence from observations of CT scanners and the social order of radiology departments. *Administrative Science Quarterly, 31,* 78-108.

Berkowitz, L. (1978). *Cognitive theories in social psychology.* New York: Academic Press.

Daniels, K., Markoczy, L., & de Chernatony, L. (1994). Techniques to compare cognitive maps. *Advances in Managerial Cognition and Organizational Information Processing, 5,* 141-164.

DeSanctis, G., & Poole, M.S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science, 5*(2), 121-147.

Easterby-Smith, M. (1980). The design, analysis and interpretation of repertory grids. *International Journal of Man-Machine Studies, 13,* 3-24.

Ginsberg, A. (1989). Construing the business portfolio: A cognitive model of diversification. *Journal of Management Studies, 26*(4), 417-438.

Griffith, T.L. (1999). Technology features as triggers for sensemaking. *Academy of Management Review, 24*(3), 472-488.

Griffith, T.L., & Northcraft, G.B. (1996). Cognitive elements in the implementation of new technology: Can less information provide more benefits? *MIS Quarterly, 20,* 99-110.

Huff, A.S. (1990). *Mapping strategic thought.* Chichester: John Wiley & Sons Ltd.

Hunter, M.G. (1997). The use of RepGrids to gather interview data about information systems analysts. *Information Systems Journal, 7*(1), 67-81.

Hunter, M.G., & Beck, J.E. (1996). A cross cultural comparison of 'excellent' systems analysts. *Information Systems Journal, 6,* 261-281.

Jelinek, M., & Litterer, J.A. (1994). Toward a cognitive theory of organizations. In C. Stubbar, J.R. Meindl & J.F. Porac (Eds.), *Advances in managerial cognition and organizational information processing* (pp. 3-41). Greenwich, CT: JAI Press.

Krueger, R.A. (1988). *Focus groups: A practical guide for applied research.* Newbury Park, CA: Sage Publications.

Latta, G.F., & Swigger, K. (1992). Validation of the repertory grid for use in modeling knowledge. *Journal of the American Society for Information Science, 43*(2), 115-129.

Lee, A.S. (1991). Integrating positivist and interpretive approaches to organizational research. *Organization Science, 2*(4), 342-365.

Moynihan, J.A. (1988). Current issues in introducing and managing information technology: The chief executive's perspective. In *Information technology for organisational systems*. Brussels-Luxembourg: Elsevier Science.

Moynihan, T. (1996). An inventory of personal constructs for information systems project risk researchers. *Journal of Information Technology, 11,* 359-371.

Odell, L., Goswami, D., & Herrington, A. (1983). The discourse-based interview: A procedure for exploring tacit knowledge of writers in nonacademic settings. In *Research on writing: Principals and methods* (pp. 220-236). New York: Longman.

Orlikowski, W.J. (1992). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science, 3*(3), 398-427.

Orlikowski, W.J., & Gash, D.C. (1994). Technological frames: Making sense of information technology in organizations. *ACM Transactions on Information Systems, 12*(2), 174-201.

Phythian, G.J., & King, M. (1992). Developing an expert system for tender enquiry evaluation: A case study. *European Journal of Operational Research, 56*(1), 15-29.

Porac, J.F., & Thomas, H. (1989). Competitive groups as cognitive communities: The case of Scottish knitwear manufacturers. *Journal of Management Studies, 26*(4), 397-416.

Schein, E. (1987). *The clinical perspective in fieldwork.* Newbury Park, CA: Sage.

Stewart, V., & Stewart, A. (1981). *Business applications of repertory grid.* UK: McGraw-Hill.

KEY TERMS

Cognition: Cognition is considered to be synonymous with subjective understanding, "the everyday common sense and everyday meanings with which the observed human subjects see themselves and which gives rise to the behaviour that they manifest in socially constructed settings" (Lee, 1991, p. 351).

Construct: Constructs represent the research participant's interpretations of the elements. Further understanding of these interpretations may be gained by eliciting contrasts resulting in bi-polar labels. Using the same example, research participants may come up with bi-polar constructs such as "high user involvement – low user involvement" to differentiate the elements (i.e., IS projects). The labels represent the CSFs of IS projects.

Elements: Elements are the objects of attention within the domain of investigation. They define the entities upon which the administration of the RepGrid is based. For example, to explore the critical success factors (CSFs) of IS projects, IS researchers can use IS projects as elements in the RepGrid.

Idiographic: The idiographic approach focuses on the subjective experiences of the individual and presents results in expressions and terms used by the individual. The resulting RepGrid is considered unique in that there are no common elements or constructs employed in the elicitation process across the sample.

Links: Links are ways of relating the elements and constructs. The links show how the research participants interpret each element relative to each construct. Further, the links reveal the research participant's interpretations of the similarities and differences between the elements and constructs.

Nomothetic: The nomothetic approach permits the comparison of RepGrids of individuals or groups of individuals. It necessitates the use of a common set of elements and/or constructs to permit comparisons to be made between RepGrids.

Repertory Grid (RepGrid): The RepGrid is a cognitive mapping technique that can be used to describe how people think about a phenomenon in their world. The RepGrid technique, for IS, entails a set of procedures for uncovering the personal constructs individuals use to structure and interpret events relating to the development, implementation, use, and management of IT in organizations. The RepGrid contains three components – elements, constructs and links.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 53-58, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Multimedia Technology

A Cognitively-Based Framework for Evaluating Multimedia Systems

Eshaa M. Alkhalifa
University of Bahrain, Bahrain

INTRODUCTION

Multimedia systems waltzed into the lives of students and educators without allowing anyone the time required for the development of suitable evaluation techniques. Although everyone in the field is aware that judging this type of teaching software can only come through evaluations, the work done in this regard is scarce and ill-organized. Unfortunately, in many of the cases the evaluation forms were just filled in by instructors who pretended to be students when they went through the tutorial systems (Reiser & Kegelmann, 1994). Nowadays, however, awareness of the impact of evaluation results on the credibility of the claims made is rising.

BACKGROUND

Ever since the early days, researchers recognized the existence of two main dimensions of multimedia evaluations. Formative evaluation is concerned with the program's functional abilities and efficiency, while summative evaluations are concerned with the effectiveness of the system in achieving its goals (Bloom, Hastings, & Madaus, 1971; Scriven, 1967).

Heller and Martin (1999) inform us that the evaluation question depends on the core subject from which it emerges, and they present a list of four subjects, namely, computer science, computer graphics, education, and human-computer interaction. They explain that if the core subject is computer science, then the research question concerns the technical requirements of the multimedia systems, including data compression, storage requirements, bandwidth, and data transmission. If the question is from computer graphics, then the focus is on speed of image rendering, representation of light, and creation of animation. If the question is from education, then media is evaluated in terms of its impact on teaching and learning along with attributes such as motivation, feedback, and information delivery. If the question is from human-computer interaction, then it is concerned with the use of multimedia in interface design focusing on issues that impact the interaction itself, such as screen design, the use of metaphor, and navigational strategies.

In retrospect, when examining their classification, we find that the main two dimensions are well covered. Formative evaluation is the focus of computer science and computer graphics as a whole, while summative evaluation is the main, but not only, focus of education and human-computer interaction. With education, we find that "motivation" does not conform to any of the two dimensions, while human-computer interaction requires both formative evaluation in addition to summative evaluation.

Researchers tested their systems through a summative evaluation frequently using a pretest and posttest where the first is taken prior to using the system and the second following the use of the system. Unfortunately, this type of testing has been plagued with no significant[1] differences in student grades when multimedia is compared to classroom lectures or to carefully organized, well-illustrated textbooks (Pane, Corbett, & John, 1996). Others widened the scope of their evaluation procedure by adding learning style questionnaires that targeted student-learning preferences and a subjective questionnaire that investigated motivation issues (Kinshuk, Patel, & Russell, 2000).

Disappointment in the results of pretests and posttests caused researchers to alter the main summation evaluation question. They wondered if the test is for the educational effects of interactivity versus lack of interactivity, or should one compare animation with textual media (McKenna, 1995). If Pane et al. (1996) were aware of the work done by Freyd (1987) who studied the cognitive effects of exposing subjects to a series of still images to find that they are equivalent in the reactions they elicit to being exposed to a moving picture, then perhaps they would not have asked whether animation is equivalent to a textbook with carefully set images of all stages.

The problem that emerged in the summation dimension is therefore the question itself. Tam, Wedd, and McKerchar (1997) proposed a three-part evaluation procedure that includes peer review, student evaluation, and pre- and post-testing. They were not able to get rid of the pretest and posttest evaluation, as it is the primary test for how much learning was achieved, and they still saw no significant differences in their results. In other words, they collected more subjective feedback from users, and this is not classified under summation evaluation.

At this stage, researchers recognized that evaluations did not target the appropriate level of detail that would enable them to detect differences that may exist in their results. Song, Cho, and Han (2000, 2001) presented empirical support that animation helps reduce the cognitive load on the

learner. They also showed that multimedia is more effective in teaching processes than in teaching conceptual definitions, while textual presentations are better at the latter. However, all this was done in very limited test domains that lacked the realistic world of an educational system. Al Balooshi and Alkhalifa (2002) presented such an educational system evaluation by showing that the student cognitive styles do affect how they learn and that only the correct research question is capable of detecting these differences.

The evaluation framework therefore has to be expanded to include the motivation measure that is tested by designers but does not conform to any predefined dimension and to include these more detailed research questions at the appropriate level of abstraction to detect the differences between the multimedia systems subject to evaluation.

A THREE-DIMENSIONAL FRAMEWORK FOR EVALUATION

All evaluations start with an evaluation question that compares the one thing to another. If the words used to classify the question are not distinctive enough, then research will not benefit from the findings or misclassify them. Motivating students, for example, does not necessarily imply that the educational impact will be influenced. Asking a question that is at a high level of abstraction, like if one media teaches as well as another without specifying the details of the design, difference, and materials, may also generalize to the degree that true benefits are shadowed or lost in the generalization. Consequently, a complete framework of evaluation is required to take into account all issues concerning the software and the learning process. Evaluation questions can be channeled into three main dimensions of evaluation that could then be subdivided into the various methods that form possible criteria that guides the evaluation process.[2]

The framework is composed of three main steps that will be explained through a case study of a Data Structure Tutorial System (DAST) that was developed and evaluated at the University of Bahrain (Alkhalifa & Al Balooshi, 2003).

The first dimension was first reviewed by distributing three instructor surveys based on a series of questions proposed by Caffarella (1987), to allow them to illuminate the various anomalies that may be present in the multimedia system design. A similar evaluation was distributed among students to allow them to evaluate the system subjectively.

The second dimension required a slightly more complicated process, which started with a pre-evaluation test of all students to ensure they were divided into groups of equivalent mean grades. Then the pretests and posttests were written to ensure that one set of questions mapped onto the next by altering their order while ensuring they included declarative questions that required verbalization

Table 1. A three-dimensional framework for evaluation

First Dimension: System Architecture

This dimension is concerned with the system's main modules, their programming complexity, and their interactions. Evaluation within this dimension should be performed in any or all of the following methods:

- Operation of the system as a whole is described and evaluated to show optimization techniques used, and so forth.
- Expert survey of the system filled by experts in the field or educators.
- User evaluations of the way the system works, to indicate if they ran into any errors or problems that have not been predicted by designers.
- Architectural design of the system is presented and evaluated against prior work and evaluations of similar systems to ensure that it benefits from the lessons learned.
- All business related issues such as cost analysis and portability, time frame required for mass production of similar systems, and so forth.

Second Dimension: Cognitive Impact

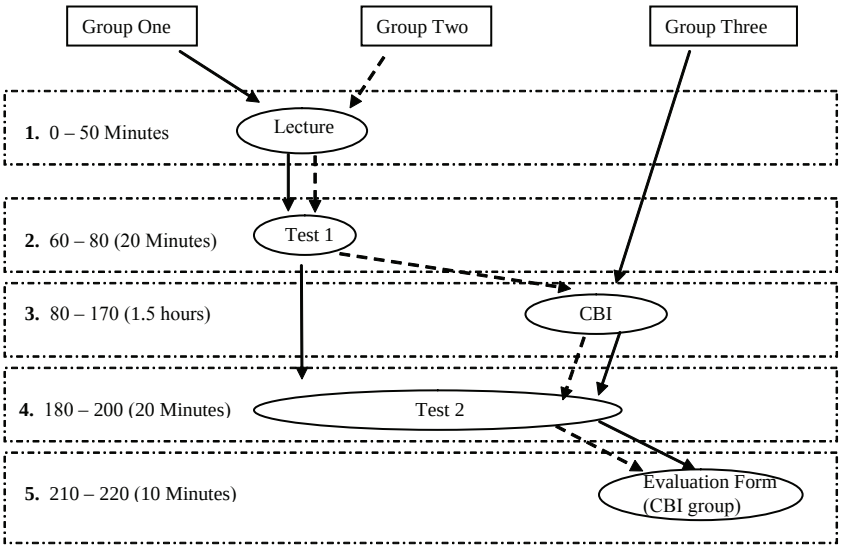
This dimension is concerned with assessing the benefits that could be gained by students when they use the system. Classically all the following methods must be measured using pretests and posttests of educational impact prior to and following use of the system. Here we find four areas of focus: two types of knowledge, cognitive traits, and the classification of the materials presented.

- Tests of declarative knowledge that required verbalization of what users understood or learned following the use of the system.
- Tests of procedural knowledge that tested if students understood how the concepts could be applied in novel situations.
- Tests of cognitive styles that impacted how well students learned from one approach to presenting information vs. an alternative approach.
- Tests of the alignment of the materials taught with the teaching style selected for that material. For example, is teaching mathematics more effective if it is interactive with live computation or if it is through textual presentation.

Third Dimension: Affective Measures

This dimension is mainly concerned with student opinions on the user friendliness of the system and allows them to express any shortcomings in the system. This could best be done through surveys where students are allowed to add any comments they wish freely and without restraints.

Figure 1. The evaluation procedure



of how students understood concepts as well as procedural questions that tested if students understood how the concepts could be applied. The evaluation procedure for students is shown in Figure 1.

ANALYSIS OF RESULTS

Results of the first dimension survey distributed to three peer experts showed the system earned an average rating of 5.33 on a scale of 0 to 6. Students of groups two and three generally gave ratings of around 4 to 5 on a scale from 0 to 6 with the highest for "The use of graphics, sound, and color contributes to the student's achievement of the objectives" and "The user can control the sequence of topics within the CBI program." The lowest score was 3.559 for "The level of difficult is appropriate for you." Therefore, it seems that the students in general enjoyed learning through the system, although they found the level of difficulty of the concepts presented challenging.

For the second dimension, student grades were first analyzed using the Analysis of Variance (ANOVA) test. This test allows the evaluation of the difference between the means by placing all the data into one number, which is F, and returning as a result one p for the null hypothesis. It also compares the variability that is observed between conditions to the variability observed within each condition. The statistic F is obtained as a ratio of two estimates of students' variances. If the ratio is sufficiently larger than 1, then the observed differences among the obtained means are described as being statistically significant. The term "null hypothesis" represents an investigation done between samples of the groups with the assumption that additional learning will not occur as a result of the treatment. In order to conduct a significance test, it is necessary to know the sampling distribution of F given the significance level needed to investigate the null hypothesis. It must be also mentioned that the range of variation of averages is given by the standard deviation of the estimated means.

The ANOVA test did show that there is a significant improvement in group two between the first test, which was taken after the lecture, and the second test, which was taken after using the system, which indicates that learning did occur during that phase for group two students. In order to be able to properly assess the amount of learning made by the CBI alone, one must compare student results to that of the classroom alone, to avoid the issue of confounding by assessing students who were exposed to the same material twice. To assess this, one may examine the scores by using the total average, which is approximately 10.5, calling this a borderline figure. Then a count of how many students scored above this borderline is given for each group as follows: group one has six students above this line, while group two has 11 students, and group three has 10, with all groups composed of the same total number of students (15). This shows that the third group, which took the CBI package alone, and the second group, which had both the classroom lecture and the CBI package exposure, are close. It also underlines how much the second group improved their test results after taking the CBI and in the same time showing that the first group had not improved much only with the lecture learning.

However, this was not sufficient to be able to pinpoint the strengths of the system. Therefore, a more detailed analysis was done of student performance in the individual questions

of test one and test two. Questions were placed carefully to test the various types of knowledge students may acquire. The largest differences were observed in the effects of the CBI on procedural knowledge with the question "*Using an example, explain the stack concept and its possible use*." In group two there was a significant improvement made from the pretest to posttest with F=58 and $\underline{p}$<.000. When comparing group one to group two students on the same questions, results turned out to be in favor of the classroom only group with an F=5.02 and $\underline{p}$ <.03. This result indicates that using the system following a classroom lecture reinforces student comprehension of how the concept relates to the real world, while teaching them through the CBI alone produces a worse effect than that in the classroom perhaps because it "limits" their imagination to the graphical representations shown in the system (Al Balooshi & Alkhalifa, 2002). Classroom lectures introduce students to the concepts allowing them all the freedom to select all types of applications, which is in some ways overwhelming. The use of the system, on the other hand, produces a safe haven to test their ideas and strongly pursue the examples they can imagine, which helps them arrive at a solid procedural understanding of the concepts.

For the third dimension, the same survey distributed to students reflected their enthusiasm to incorporate the use of this educational system into the course curriculum as is shown by achieving a rate of 5.085 on a scale of 6. They also voiced some comments requesting more examples that go more in depth into the theory to solve more complex questions that they face during the course. Other affective measures were included asking students what they feel about the course materials, level of difficulty and necessary prerequisites, and how much they learned from it.

FINE-GRAINED EVALUATION

The evaluation framework proposed here classifies the evaluation question as to focus on one or more of three specific dimensions. The first dimension focuses on the design issues to ensure avoiding asking the wrong question as was done in studies (Byrne, Catrambone, & Stasko, 1999; Lawrence, Badre, & Stasko, 1994) that did not take into account that a sequence of images are translated as animation (Freyd, 1987).

The second dimension evaluates multimedia educational software at a finer level of detail than what was previously followed. This was done carefully by asking clear questions of how groups differ from each other, causes of differences, and how to assess the level of learning attained using students' total average grade. More detailed results identified differences in procedural knowledge and not in declarative knowledge.

Last but not least, the same evaluation survey utilized for the first dimension contained measures for the motivational issues and how students rate the usefulness of the system from a subjective point of view.

FUTURE TRENDS

A number of researchers, including Magenheim and Scheel (2004), found this evaluation procedure credible enough to base future work upon it. On the other hand, we have renewed interest in the system architecture dimension as evident in the work done by Jones and Cockton (2004).

This is clear evidence that the ideas presented here are gradually finding their way to attract researcher attention to determine and clarify their research questions clearly and formulate them according to the context of the presented system and educational material without neglecting any of the factors that may influence results.

CONCLUSION

A three-dimensional framework is presented as a means to evaluating multimedia educational software in order to resolve the shortcomings of the current evaluation techniques. It differs from the other in that it adds a dimension that was previously tested but never given a title in addition to seeking a clearer more fine-grained analysis of the comparisons made and statistical analysis carried out. In other words, this framework offers a step-by-step guide of a credible evaluation technique of multimedia systems that is currently accepted by peer researchers.

REFERENCES

Al Balooshi, F., & Alkhalifa, E. M. (2002). Multi-modality as a cognitive tool. *Journal of International Forum of Educational Technology and Society: Special Issues: Innovations in Learning Technology, IEEE, 5*(4), 49-55.

Alkhalifa, E. M. (2005). Multimedia evaluations based on cognitive science findings. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 4) (pp. 2058-2062). Hershey, PA: Idea Group Reference.

Alkhalifa, E. M., & Al Balooshi, F. (2003). A 3-dimensional framework for evaluating multimedia educational software. In F. Al Balooshi (Ed.), *Virtual education: Cases in learning & teaching technologies* (pp. 195-209). Hershey, PA: Idea Group Publishing.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of learning*. New York: McGraw-Hill.

Byrne, M. D., Catrambone, R., & Stasko, J. T. (1999). Evaluating animation as student aids in learning computer algorithms. *Computers and Education, 33*(4), 253-278.

Caffarella, E. P. (1987). Evaluating the new generation of computer-based instructional software. *Educational Technology, 27*(4), 19-24.

Freyd, J. (1987). Dynamic mental representations. *Psychological Review, 94*(4), 427-438.

Heller, R. S., & Martin, C. D. (1999). Using a taxonomy to rationalize multimedia development. In *IEEE International Conference on Multimedia Computing and Systems (CMCS '99),* Florence, Italy (Vol. 2, pp. 661-665).

Jones, S., & Cockton, G. (2004). Tightly coupling multimedia with context: Strategies for exploiting multi modal learning in complex management topics. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT '04)* (pp. 813-815). Joensuu, Finland.

Kinshuk, Patel, A., & Russell, D. (2000). A multi-institutional evaluation of intelligent tutoring tools in numeric disciplines. *Educational Technology & Society, 3*(4). Retrieved May 14, 2006, from http://ifets.ieee.org/periodical/vol_4_2000/kinshuk.html

Lawrence, W., Badre, A. N., & Stasko, J. T. (1994). *Empirically evaluating the use of animation to teach algorithms* (Tech. Rep. No. GIT-GVU-94-07). Atlanta, GA: Georgia Institute of Technology.

Magenheim, J., & Scheel, O. (2004). Using learning objects in an ICT-based learning environment. In *Proceedings of E-Learn 2004, World Conference in E-Learning in Corporate, Government, Healthcare & Higher Education* (pp. 1375-1382). Washington, DC.

McKenna, S. (1995). Evaluating IMM: Issues for researchers. In *Occasional Papers in Open and Distance Learning,* No 17. Open Learning Institute, Charles Sturt University.

Pane, J. F., Corbett, A. T., & John, B. E. (1996). Assessing dynamics in computer-based instruction. In *Proceedings of the 1996 ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 197-204). Vacouver, B.C., Canada.

Reiser, R. A., & Kegelmann, H. W. (1994). Evaluating instructional software: A review and critique of current methods. *Educational Technology, Research and Development, 42*(3), 63-69.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Curriculum evaluation* (pp. 39-83). Chicago: Rand-McNally.

Song, S. J., Cho, K. J., & Han, K. H. (2000). The effectiveness of cognitive load in multimedia learning. In *Proceedings of the Conference of the Korean Society for Cognitive Science,* Seoul, Korea (pp. 93-98).

Song, S. J., Cho, K. J., & Han, K. H. (2001). Effects of presentation condition and content type on multimedia learning. In *Proceedings of the Third International Conference on Cognitive Science, ICCS 2001* (pp. 654-657). Beijing, China.

Tam, M., Wedd, S., & McKerchar, M. (1997). Development and evaluation of a computer-based learning pilot project for teaching of holistic accounting concepts. *Australian Journal of Educational Technology, 13*(1), 54-67.

KEY TERMS

Cognition: The psychological result of perception, learning, and reasoning.

Cognitive Load: The degree of cognitive processes required to accomplish a specific task.

Cognitive Science: The field of science concerned with cognition and includes parts of cognitive psychology, linguistics, computer science cognitive neuroscience, and philosophy of mind.

Cognitive Tool: A tool that reduces the cognitive load required by a specific task.

Declarative vs. Procedural Knowledge: The verbalized form of knowledge versus the implemented form of knowledge.

Learning Style: The manner in which an individual acquires information.

Multimedia System: Any computer delivered electronic system that presents information through different media that may include text, sound, video computer graphics, and animation.

ENDNOTES

1 This term is defined in the Analysis of Results section.

2 This framework is an enhanced version of a previously published framework (Alkhalifa, 2005) with most changes resulting from further work by either the author or other researchers who referenced the author's work.

Collaborative Virtual Environments

Thrasyvoulos Tsiatsos
Aristotle University of Thessaloniki, Greece

Andreas Konstantinidis
Aristotle University of Thessaloniki, Greece

INTRODUCTION

Computer supported collaboration is one of the most promising innovations to improve teaching, learning, and collaborating with the help of modern information and communication technology (Lehtinen & Hakkarainen, 2001). Continuous enhancements in computer technology and the current widespread computer literacy among the public have resulted in a new generation of users (less so in developing countries) that expect increasingly more from their e-learning experiences. To keep up with such expectations, e-learning systems have gone through a radical change from the initial text-based environments to more stimulating multimedia systems (Monahan, McArdle & Bertolotto, in press).

Generally a collaborative virtual environment (CVE) can be defined as a computer-based, distributed, virtual space or set of places. In such places, people can meet and interact with others, with agents (artificial intelligence), or with virtual objects. CVEs might vary in their representational richness from 3D graphical spaces, 2.5D and 2D environments, to text-based environments. Access to CVEs is by no means limited to desktop devices, but might well include mobile or wearable devices, public kiosks, and so forth (Churchill, Snowdon & Munro, 2001). CVEs are a subset of Virtual Environments (VEs) in that only VEs which support collaborative operations can be considered CVEs. The two primary uses of CVEs are for collaborative learning and/or collaborative work in either educational and/or professional environments.

Computer supported collaborative learning (CSCL) is an umbrella term for a variety of approaches in education that involve the joint intellectual effort by students or students and teachers and that require the use of computer and communication technology. Researchers (e.g., Ahern, Peck & Laycock, 1992; Bruckman & Hudson, 2001; Singhal & Zyda, 1990) have proven the effectiveness of collaborative learning compared to other educational practices (e.g., competitive or personalized learning), praising this method's way of aiding the acquisition of higher level cognitive abilities, problem solving abilities, ease in scientific expression and the development of communication, social and higher order thinking skills.

The most important advantages of using CSCL are discussed in Bruckman et al. (2002). It is mentioned that through CSCL teacher/student interactions become more balanced and that there is also some evidence to suggest that gender differences are reduced in online environments. In addition, students exhibit higher levels of attention and appear more honest and candid toward those in a position of authority. Learning becomes more student-oriented, thus increasing the likelihood that students will absorb and remember what they learn.

On the other hand, computer supported collaborative work (CSCW) is a generic term, which combines the understanding of the way people work in groups with the enabling technologies of computer networking, and associated hardware, software, services and techniques (Wilson, 1991). Although some authors consider CSCW and groupware as synonyms, others argue that that while groupware refers to real computer-based systems, CSCW focuses on the study of tools and techniques of groupware as well as their psychological, social, and organizational effects. For example, researchers Hiltz and Turoff (1993) conclude that the social connectivity of users who adopt a computer-mediated communication system increases notably. They also found a strong tendency toward more equal participation, and that more opinions tended to be asked for and offered.

The purpose of this chapter is to present a concise yet complete overview of collaborative virtual environments. In the following sections we will discuss the technological evolution of CVEs, their basic characteristics and architectures, and the tools and services integrated within them. Finally, there will be a brief mention of the design challenges facing CVE designers and of future trends with which CVE functionality will be extended.

BACKGROUND

The first virtual worlds were text based, in that their environments and the events occurring within them were described using words rather than images. Their primary use was for entertainment and specifically as fantasy role-playing games. Virtual worlds are often called MUDs (Multi User Dungeons) because MUD was the name of the first one to prosper. Its author was Roy Trubshaw. In 1989, TinyMUD was one of the first virtual worlds to focus on the social aspects of these environments. Users could create new locations and objects,

spending most of their time creating and talking about their creations. In 1990 MOO (MUD, Object Oriented), introduced a fully functional scripting language and allowed users of social-oriented virtual worlds to add not only objects, but also powerful functionality to the environment as it ran. MOO's descendents have found a niche in the educational world, as they are easy to use and can demonstrate the principles of programming to new users. Also in 1990, TinyMUSH, among other things, introduced event triggering and software automatons (known as puppets then and as agents today). In 1993, before the advent of the World Wide Web, MUDs constituted some 10% of the Internet (Bartle, 2004).

Text based collaboration started around 1990 with a system called "Reality Built for Two"; there is one system in 1987 by Sim et al. which can be classified as a Collaborative Virtual Environment, but was built using dedicated hardware for military training purposes (Joslin, Di Giacomo & Magnenat-Thalmann, 2004). It is interesting to note that CVE systems have been around long before the World Wide Web was invented, but have not been used as extensively by the general public for personal or commercial activities. This is possibly because of their complexity and base requirements being much more demanding, or possibly the content being much harder to create. Reality Built for Two (RB2) was a development platform for designing and implementing real-time virtual realities (Blanchard & Burgess, 1990). Development was rapid and interactive in RB2. Behavior constraints and interactions could be edited in real time with the system running. Changes made to interactions in the world were seen immediately in Virtual Reality (VR). The primary user input devices in use in RB2 were the DataGlove which allowed gestural and direct manipulation of the environment, and the Polhemus tracker for head tracking.

After 1990, the popularity of CVEs remained almost stable with the appearance of three to four new systems each year. A more substantial increase in popularity was observed in 1995 with the release of systems such as RING, Virtual Society, MASSIVE and SmallView. CVE popularity peaked in 1997 with new developments generally falling off ever since (Joslin et al., 2004). It seems 1997 can be seen as the point of maturity for CVEs. The decline of the scientific community's interest in the theoretical basis of CVEs has seen a rise in commercial CVE products today. Contemporary systems include Active Worlds (released in 1997), There (released in 2003, http://www.there.com/), I-maginer (http://www.i-maginer.fr), Workspace3D (http://www.tixeo.com), Second Life and Croquet.

The most successful CVE today seems to be Second Life with over four million total sign-ups. Released in 2003, the I nternet-based virtual world Second Life (SL, http://secondlife.com/) came to international attention via mainstream news media in late 2006 and early 2007. Users in SL can explore, meet other users, socialize, participate in individual and group activities and create and trade items

Table 1. The CSCW matrix

	Same Time (synchronous)	**Different Time (asynchronous)**
Same Place (collocated)	**Face-to-face interactions** – decision rooms, single display groupware, shared table, wall displays, room ware …	**Continuous task** – shift work groupware, project management, and so forth
Different Place (remote)	**Remote interaction** – video conferencing, instance messaging, chats/ MUDs/ virtual worlds, shared screens, multi-user editors, and so forth	**Communication and coordination** – email, bulletin boards, blogs, asynchronous conferencing, group calendars, workflow, version control, wikis, and so forth

or services from one another. Although signing up is free in SL, several activities such as the building of items and use of land are not.

On the other hand, Croquet (http://www.croquetconsortium.org/) is an open source cross platform 3D environment designed for rich interaction and simulation, with a combination of powerful graphics and multiuser collaborations (McCahill, 2004). Although Croquet started officially in 2001, it attracted most attention in 2006 with the release of the Croquet SDK Beta as open source.

ISSUES CONCERNING THE CHARACTERISTICS AND DESIGN OF CVES

Basic Characteristics

Computer supported collaboration systems can be categorized based on the spatial and temporal coexistence of the participants and their interactions, as shown in the CSCW matrix (Table 1) introduced in 1988 by Johansen (Baecker, 1995). Different forms of groupware, communication tools and collaboration activities can be organized in this way. In this matrix CVEs are located in the bottom left quadrant since they facilitate a remote yet synchronous form of collaboration.

VEs provide: (a) a shared sense of space, creating the illusion to the users that they are located in the same place, (b) a shared sense of presence, in relation to the virtual representation of the users that is commonly realized through anthropomorphic personas called avatars. Furthermore, VEs provide communication methods, such as gestures, text, and audio-visual cues. Finally, VEs support a sharing functionality, allowing participants to distribute workload between them (Singhal & Zyda, 1999) and, for example, cobrowse the Web and/or coauthor a document.

According to Bouras and Tsiatsos (2006), a CVE is an environment in which the users participating have different roles and privileges. The educational interactions (interactions which are intellectually beneficial) in the environment transform the simple virtual space into a communication space, meaning that multiple communication channels are available to the users, allowing them to interact in multiple ways with each other inside the virtual environment. Information in a CVE is represented in multiple ways that can vary from simple text to 3D graphics which usually visualize recognizable elements from the real world. On the pedagogical side, students are not passive users but can interact with each other and with the virtual environment which also supports the possibility of implementing multiple learning scenarios. On the technical side, the system that supports the CVE incorporates multiple technologies such as network architectures and 3D graphics engines.

Architectures

The network architectures that support CVEs can be categorized into: (a) client-server architectures, where the clients communicate their changes to one or more servers and these servers, in turn, are responsible for the redistribution of the received information to all connected clients and (b) peer-to-peer architectures, where the clients transmit their modifications and updates of the virtual world directly to all other connected clients (McGregor, Kapolka, Zyda & Brutzman, 2003). The client-server model is the simpler of the two, yet it cannot support high scalability as there is a central point of failure: the server. On the other hand, the peer-to-peer architecture's scalability is restricted by the network properties. It should be noted that hybrid solutions can be adopted, so as to cater for the specific needs and the type of the application that each system aims to support. There are hybrid architectures, which adopt the simple client-server model with peer-to-peer communication among groups of servers or with server hierarchies, where certain servers act as clients to others. Also, the client-server and peer-to-peer structures can be integrated into peer-server architectures, where some data packets are transmitted through certain nodes using peer-to-peer while others through a server.

When client server architecture is utilised to provide for the virtual environment of the CVE, it usually consists of layers. For example Bartle (2004) mentions that the most basic and fundamental layer handles things like memory management, parsing, data structures, input/output queuing, packet handling, and so forth. This layer makes available two foundation concepts: entities that make up the world and the association of input/output with some of those entities. The second layer up defines the physics of the virtual world, such as mass, movement, communication, and so forth. The next layer up adds world-specific concepts and functionality consequent of the physics defined in the previous layer. In the final layer actual data items define the individual world, differentiating it from all the worlds that could possibly be defined.

Collaborative Virtual Environments Tools and Services

In order to support the collaborative process, CVEs must be equipped with a multitude of tools and services catering for communication, file sharing, coauthoring of documents and cobrowsing of the Internet. Current collaborative learning environments offer many helpful tools which constitute and augment their pedagogical value. Typical tools and services

of CVEs include desktop conferencing, videoconferencing, electronic mail, blogs, forums, chat, meeting support systems, voice applications, workflow systems, group calendars and intelligent agents (Goebbels, Lalioti & Göbel, 2003). Based on the features and advantages of CVEs in general, these can be separated into five categories: tools for communication, management and coordination, teacher and student support, shared applications and avatar functions. These categories will be analyzed in the following paragraphs.

Email technology is the most broadly used method for text based asynchronous communication today (Ballesteros, 2006). In most contemporary CVEs access to email is usually available through an external application, or through an embedded Web browser. The same goes for discussion forums. An e-mail-based learning environment can be used as a very open system for spontaneous collaboration or it can be more organized, controlled, and tutored (Ahern et al, 1992). Other tools for communication include chat, videoconference, avatar gestures and VoIP. Many users deem the VoIP service necessary, since the ability to send sound messages strengthens the psychological sense of presence in the virtual world (Monahan et al., 2007), without increasing costs and bandwidth requirements.

Applications which can be shared between users include, but are not limited to, word processors, internet browsers, whiteboards, Computer Aided Design (CAD) and brainstorming tools. Also, many CVEs host links to multiple external sources of information, such as the Internet and third party applications.

Another group of tools concern the collaboration and interaction of teachers with students within virtual environments. The tools in this group are relevant with class management tasks and the evaluation of both the students and the teaching procedure in general. Other tools include voting and argumentation tools, RSS feeds and wikis.

For the efficient coordination of a collaborative process or a collaborative learning session, the way the users gain access to the common workspace is very important. Several methods have been employed including action key, attention focus and automatic agents. In the action key method, the user possessing the action key is the only one with access to the common workspace. The rest of the users can ask to obtain the key from the current owner. Attention focus is the ability to focus the attention of the students by the teacher. Here, the user who has access to the medium can lock the first person perspectives of all the other users to a certain object. Finally automatic agents are intelligent entities that can monitor students' progress and manage the system more efficiently. They can also be used to personalize the learning experience for each user tailoring the interface or the delivery of learning content to their individual preferences.

Inside a CVE users interact with the virtual world and its inhabitants through an avatar. Some of the basic operations that avatars can perform are interacting with objects and other users, navigating the virtual environment by walking, flying or teleporting, and communicating through gestures and facial expressions. Many CVEs allow the user to directly manipulate their avatar's appearance and thus enhance the feeling of trust and security between the members of a team.

Design Challenges

The design challenges faced by CVE developers vary. Some concern the technical requirements of the virtual environment, such as system responsiveness and network bandwidth, while others have to do with the users themselves or the type of interfaces that are realized.

Concerning technical issues, high system responsiveness is perceived as having a very positive impact on collaboration. Even downsizing the application in order to decrease the CPU load is recommendable. Good system responsiveness is guaranteed if all inputs and outputs are processed and rendered within less than 50 milliseconds (Goebbels et al., 2003). A typical issue is that of the virtual world slowing or stopping when new scene components are loaded. Given the user's expectation of free movement at all times, this suggests to the user that an error has occurred, or that the operation failed.

An inherent failure of multi-user applications (and therefore of CVEs also) is that they never provide precisely the same benefit to every group member (Grudin, 1991). For instance the fact that low achievers progressively become passive when collaborating with high achievers must be taken into account when organizing groups (Dillenbourg, Baker, Blaye & O'Malley, 1996). Another consideration is that a group of three is less effective because it tends to be competitive, whilst pairs tend to be more cooperative.

Application issues are generally concerned with the affordances of objects and the lack of help with the virtual environment itself. They are broad in nature, from problems with objects whose operation is not obvious, to wider topics such as how best to represent group services to group members.

Interaction and interface issues challenging designers have to do with some tasks being less "shareable" than others and that although users might find 3D interfaces enjoyable, recognizable, and memorable because they improve spatial memory use, they can also be distracting and confusing because of increased visual complexity.

Finally, empirical testing confirms that virtual reality systems induce physical symptoms and effects that need to be compensated for. In addition, more research is needed in order to assess social discomfort levels generated in a CVE and caused by participants working concurrently with real people and their avatars.

FUTURE TRENDS

Although the commercial success of CVEs has proven their effectiveness in entertainment, for real world organisational users there is the matter of fitness for purpose, and consequently, confidence in such novel technology. CVEs must show they can deliver safety-critical training to senior professionals within simulated real life working environments and through this lead to a validation by a recognized training and standards body as being of a suitable standard. Finally, CVEs must be accepted by the trainers, trainees and employers who will have to use them (Turner & Turner, 2002).

In order to meet the criteria listed previously, CVEs must enhance the sense of realism through advanced graphics capabilities and the incorporation of pioneering technologies such as haptic technology and brain computer interfaces. Strong utility of pure 3D interfaces for medical, architectural, product design, and scientific visualization means that interface design for pure 3D remains an important challenge.

Furthermore, server and network technology needs to be augmented in order to efficiently support large-scale applications. The term "large-scale" refers both to the data size (in terms of virtual space and graphics) as well as to the concurrent number of users. In conclusion, the same way the Internet is driven by standards, there should also be work towards the standardization of protocols for designing, developing and evaluating CVEs (Ballesteros, 2006).

CONCLUSION

In this chapter we presented some basic information concerning collaborative virtual environments. The areas presented were the evolution of CVEs in the last three decades, their characteristics and the fundamental architectures available for their development, as well as some significant issues that should be taken into account when designing a CVE. It seems that although many challenges remain to be overcome, the future social and educational use of CVEs will aid in engaging students in more meaningful ways than are typically seen in the classroom and in revolutionizing the pedagogical process.

REFERENCES

Ahern, T. C., Peck, K., & Laycock, M. (1992). The effects of teacher discourse in computer mediated discussion. *Journal of Educational Computing Research, 8*(3), 291-309.

Ang, K. H. & Wang, Q. (2006). A case study of engaging primary school students in learning science by using Active Worlds. In *Proceedings of the First International LAMS Conference: Design the future of learning*.

Baecker, R. M. (1995). Readings in human-computer interaction: Toward the year 2000. Morgan Kaufmann Publishers.

Ballesteros, I. L. (2006). Future and emerging technologies and paradigms for collaborative working environments. Information Society. European Commission

Bartle, R. A. (2004). *Designing virtual worlds.* New Riders Publishing, USA.

Blanchard, C. & Burgess, S. (1990). Reality built for two: A virtual reality tool. In *Proceedings of the 1990 Symposium on Interactive 3D Graphics* (pp. 35 – 36), Snowbird, Utah.

Bouras, C. & Tsiatsos, T. (2006). Educational virtual environments: Design rationale and architecture. *Multimedia Tools and Applications (MTAP), 29*(2), 153-173.

Bruckman, A. & Hudson, J. M. (2001). Disinhibition in a CSCL environment. In *Proceedings of Computer Support for Collaborative Learning (CSCL)* (pp. 629-630), Maastricht, Netherlands.

Bruckman, A., Elliott, J., & Adams, L. (2002). No magic bullet: 3D video games in education. College of Computing, Georgia Tech, Atlanta, GA, USA.

Churchill, E., Snowdon, D., & Munro, A. (2001). *Collaborative virtual environments: Digital places and spaces for interaction.* London: Springer-Verlag.

Dillenbourg, P., Baker, M., Blaye, A., & O'Malley, C. (1996). The evolution of research on collaborative learning. In E. Spada & P. Reiman (Eds.), *Learning in humans and machine: Towards an interdisciplinary learning science* (pp. 189- 211). Oxford: Elsevier.

Goebbels, G., Lalioti, V., & Göbel, M., (2003). Design and evaluation of team work in distributed collaborative virtual environments. In *Proceedings of the ACM symposium on Virtual Reality Software and Technology* (pp. 231-238), Osaka, Japan.

Grudin, J. (1991). Obstacles to user involvement in software product development, with implications for CSCW. *International Journal of Man-Machine Studies, 34*(3), 435-452.

Hiltz, S. R. & Turoff, M. (1993). *The network nation: Human communication via computer*. Cambridge: MIT Press.

Joslin, C., Di Giacomo, T., & Magnenat-Thalmann, N. (2004). Collaborative virtual environments – From birth to standardization. *Communications Magazine, IEEE, 42*(4), 28-33.

Lehtinen, E. & Hakkarainen, K. (2001). *Computer supported collaborative learning: A review*. Retrieved June 17, 2008, from http://tinyurl.com/226965

McCahill, M. (2004). Design for an extensible croquet-based framework to deliver a persistent, unified, massively multi-user and self organizing virtual environment. In *Proceedings of the Second Conference on Creating, Connecting and Collaborating through Computing*, Kyoto, Japan.

McGregor, D., Kapolka, A., Zyda, M., & Brutzman, D. (2003). Requirements for large-scale networked virtual environments. In *Proceedings of the 7th International Conference on Telecommunications ConTel 2003* (pp. 353-358). Zagreb, Croatia.

Monahan, T., McArdle, G., & Bertolotto, M. (in press). mCLEV-R: Design and evaluation of an interactive and collaborative m-learning application. *International Journal of Emerging Technologies in Learning.*

Muller, K. & Koubek, A. (2002). Collaborative and virtual environments for learning. In *Proceedings of the ACM SIG,* New Orleans, Louisiana.

Singhal, S. & Zyda, M. (1999). *Networked virtual environments: Design and implementation.* ACM Press.

Turner, P. & Turner, S. (in press). An affordance-based framework for CVE evaluation. *People and Computers XVI.*

Wilson, P. (1991). *Computer supported cooperative work: An introduction.* Kluwer Academic Pub.

KEY TERMS

Brain Computer Interface: A novel technology which allows the construction of a communication pathway between an organisms brain and a computer

Collaborative Virtual Environment (CVE): An extension of a networked virtual environment which aims at a collaborative task. CVEs aim to provide an integrated, explicit and persistent context for cooperation that combines both the participants and their information into a common display space. These objectives create the potential to support a broad range of cooperative applications such as training.

Computer Supported Collaborative Learning (CSCL): A variety of approaches in education that involve the joint intellectual effort by students or students and teachers and require the use of computer and communication technology.

Computer-Supported Cooperative work (CSCW): Is a computer-assisted coordinated activity carried out by groups of collaborating individuals.

Groupware: Is software that accentuates the multiple user environments, coordinating and orchestrating things so that users can "see" each other and yet not conflict with each other.

Haptic Technology: Refers to technology which enables users to interact with virtual environments through hand motions and receive feedback through the sense of touch

MUD Object Oriented (MOO): Refers to any MUD that uses object oriented techniques to organize its database of objects and allow players to create objects and alter their functionality.

Multiple User Dungeon (MUD): Was the name of the first VE to prosper and has been used ever since to refer to a multi-player computer game that combines elements of role-playing games, hack and slash style computer games and social chat rooms.

Virtual Environment (VE): A computer-generated simulation which is to some degree shared and persistent, allowing its users to interact with it and with each other in real time.

Virtual Reality (VR): A technology which allows a user to interact with a computer-simulated environment.

Combination of Forecasts in Data Mining

Chi Kin Chan
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

The traditional approach to forecasting involves choosing the forecasting method judged most appropriate of the available methods and applying it to some specific situations. The choice of a method depends upon the characteristics of the series and the type of application. The rationale behind such an approach is the notion that a "best" method exists and can be identified. Further that the "best" method for the past will continue to be the best for the future. An alternative to the traditional approach is to aggregate information from different forecasting methods by aggregating forecasts. This eliminates the problem of having to select a single method and rely exclusively on its forecasts.

Considerable literature has accumulated over the years regarding the combination of forecasts. The primary conclusion of this line of research is that combining multiple forecasts leads to increased forecast accuracy. This has been the result whether the forecasts are judgmental or statistical, econometric or extrapolation. Furthermore, in many cases one can make dramatic performance improvements by simply averaging the forecasts.

BACKGROUND OF COMBINATION OF FORECASTS

The concept of combining forecasts started with the seminal work 37 years ago of Bates and Granger (1969). Given two individual forecasts of a time series, Bates and Granger (1969) demonstrated that a suitable linear combination of the two forecasts may result in a better forecast than the two original ones, in the sense of a smaller error variance. Table 1 shows an example in which two individual forecasts (1 and 2) and their arithmetic mean (combined forecast) were used to forecast 12 monthly data of a certain time series (actual data).

The forecast errors (i.e., actual value – forecast value) and the variances of errors are shown in Table 2.

From Table 2, it can be seen that the error variance of Individual Forecast 1, Individual Forecast 2, and the Combined Forecast are 196, 188, and 150, respectively. This shows that the error variance of the combined forecast is smaller than any one of the individual forecasts and hence demonstrates an example how combined forecast may work better than its constituent forecasts.

Table 1. Individual and combined forecasts

Actual data (monthly data)	Individual Forecast 1	Individual Forecast 2	Combined Forecast (Simple Average of Forecast 1 and Forecast 2)
196	195	199	197
196	190	206	198
236	218	212	215
235	217	213	215
229	226	238	232
243	260	265	262.5
264	288	254	271
272	288	270	279
237	249	248	248.5
211	220	221	220.5
180	192	192	192
201	214	208	211

Note: The individual forecasts are rounded to the nearest integers as the actual data are in integers.

Table 2. Forecast errors and variances of errors

Errors of Individual Forecast 1	Errors of Individual Forecast 2	Errors of Combined Forecast
1	-3	-1
6	-10	-2
18	24	21
18	22	20
3	-9	-3
-17	-22	-19.5
-24	10	-7
-16	2	-7
-12	-11	-11.5
-9	-10	-9.5
-12	-12	-12
-13	-7	-10
Variance of errors = 196	**Variance of errors = 188**	**Variance of errors = 150**

Bates and Granger (1969) also illustrated the theoretical base of combination of forecasts. Let X_{1t} and X_{2t} be two individual forecasts of Y_t at time t with errors

$$e_{jt} = Y_t - X_{jt}, \quad j = 1,2$$

such that

$$E[e_t] = 0, \quad E[e_{jt}^2] = \sigma_j^2 \quad j = 1,2$$

and

$$E[e_{1t}e_{2t}] = \rho\,\sigma_1\sigma_2$$

where σ_j^2 is the error variance of the j^{th} individual forecast and ρ is the correlation coefficient between the errors in the first set of forecasts and those in the second set.

Consider now a combined forecast, taken to be a weighted average of the two individual forecasts,

$$X_{ct} = kX_{1t} + (1-k)X_{2t}.$$

The forecast error is

$$e_{ct} = Y_t - X_{ct} = ke_{1t} + (1-k)e_{2t}.$$

Hence the error variance is

$$\sigma_c^2 = k^2\sigma^2 + (1-k)^2\sigma_2^2 + 2k(1-k)\rho\,\sigma_1\sigma_2. \qquad (1)$$

This expression is minimized for the value of k given by

$$k_0 = \frac{\sigma_2^2 - \rho\,\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\,\sigma_1\sigma_2},$$

and substitution into equation (1) yields the minimum achievable error variance as

$$\sigma_{c0}^2 = \frac{\sigma_1^2\sigma_2^2(1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\,\sigma_1\sigma_2}.$$

Note that $\sigma_{c0}^2 < \min(\sigma_1^2, \sigma_2^2)$ unless either ρ is exactly equal to σ_1 / σ_2 or σ_2 / σ_1. If either equality holds, then the variance of the combined forecast is equal to the smaller of the two error variances. Thus, a priori, it is reasonable to expect in most practical situations that the best available combined forecast will outperform the better individual forecast—it cannot, in any case, do worse.

Makridakis et al. (1982), Makridakis and Hibon (2000), Makridakis and Winkler (1983), Newbold and Granger (1974), Terui and Van Dijk (2002), and Winkler and Makridakis (1983) have also reported empirical results that show that combinations of forecasts outperformed individual methods. However, Koning, Franses, Hibon, and Stekler

(2005) commented that one of the conclusions of the M3-competition (Makridakis & Hibon, 2000) that a combination of methods is better than that of the methods being combined was not proven.

Since Bates and Granger (1969), there have been numerous methods proposed in the literature for combining forecasts. However, the performance of different methods of combining forecasts varies from case to case. There is still neither a definitive nor a generally accepted conclusion that sophisticated methods work better than simple ones, including simple averages. As Clemen (1989, p. 566) commented, "In many studies, the average of the individual forecasts has performed the best or almost best." Others would agree with the comment of Bunn (1985, p. 152) that the Newbold and Granger (1974) study and that of Winkler and Makridakis (1983) "demonstrated that an overall *policy* of combining forecasts was an efficient one and that if an *automatic* forecasting system were required, [for example], for inventory planning, then a linear combination using a 'moving-window' estimator would appear to be the best overall."

DATA MINING AND COMBINATION OF FORECASTS IN INVENTORY MANAGEMENT

Data mining is the process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database.

The data mining process is deemed necessary in a forecasting system, and it is particularly important in combining forecasts for inventory demands. Errors in forecasting demand can have a significant impact on the costs of operating and the customer service provided by an inventory management system. It is therefore important to make the errors as small as possible. The usual practice in deciding which system to use is to evaluate alternative forecasting methods over past data and select the best. However, there may have been changes in the process, generating the demand for an item over the past period used in the evaluation analysis. The methods evaluated may differ in their relative performance over subperiods of the method that was best only part of the time, or in fact never the best method and perhaps only generally second best. Each method evaluated may be modeling a different aspect of the underlying process generating demands. The methods discarded in the selection process may contain some useful independent information. A combined forecast from two or more methods might improve upon the best individual forecasts. Furthermore, the inventory manager typically has to order and stock hundreds or thousands of different items. Given the practical difficulty of finding the best method for every individual item, the general approach is to find the best single compromise method over a sample of items, unless there are obvious simple ways of classifying the items, by item value or average demand per year, and so forth. Even if this is possible, there will still be many items in each distinct category for which the same forecasting method will be used. All of the points made on dealing an individual data series, as earlier, apply with even more force when dealing with a group of items. If no one individual forecasting method is best for all items, then some system of combining two or more forecasts would seem *a priori* an obvious approach, if the inventory manager is going to use the same forecasting system for all items.

The need for data mining in combining forecasts for inventory demands comes from selection of sample items on which forecasting strategy can be made for all items, selection of constituent forecasts to be combined, and selection of weighting method for the combination.

The selection of the sample items is a process of exploratory data analysis. In this process, summary statistics such as mean and coefficient of variation can be investigated so that the sample selected could represent the total set of data series on inventory demands.

The selection of constituent forecasts to be combined is the first stage of model building. The forecasting methods might be selected from popular time series procedures such as Exponential Smoothing, Box-Jenkins, and Regression over Time. One could include only one method from each procedure in the linear combination, as the three groups of methods were different classes of forecasting model and thus might contribute something distinct, whilst there was likely to be much less extra contribution from different methods within the same class. It might also be useful to attempt to tailor the choice of methods to particular situations. For example, Lewandowski's FORSYS system, in the M-competition (Makridakis et al., 1982), appears to be particularly valuable for long forecast horizons. Thus it might be a prime candidate for inclusion in situations with long horizons but not necessarily in situations with short horizons. It is also important to note that combining forecasts is not confined to combination utilizing time series methods, as was the case in this research. The desire to consider any and all available information means that forecasts from different types of sources should be considered. For example, one could combine forecasts from time series methods with forecasts from econometric models and with subjective forecasts from experts.

The second stage of model building is to select the weighting method for the combination of forecasts. The weighting method could be simple average or "optimal" weighting estimated by a certain approach, for instance, the constrained OLS method (Chan, Kingsman, & Wong, 1999b). Furthermore, the "optimal" weights obtained could either be fixed for a number of periods (Fixed Weighting) or re-estimated every period (Rolling Window Weighting)

(Chan, Kingsman, & Wong, 1999a, 2004). The selection process can then be done by comparing the different weighting methods with an appropriate performance measure. In the case of inventory management, the carrying of safety stocks is to guard against the uncertainties and variations in demand and the forecasting of demand. These safety stocks are directly related, or made directly proportional, to the standard errors of forecasts. If, as is usually the case, a stock controller is dealing with many items, it is the performance across the group of items that matters. Hence, the sum of the standard errors of the forecasts, measured by the sum of the root mean squared errors over all the items, can be used to compare the results between the different weighting methods. The ultimate aim is to find the one best overall method for the weighting process to use for all the items.

FUTURE TRENDS

The usual approach in practical inventory management is to evaluate alternative forecasting methods over a sample of items and then select the one that gives the lowest errors for a majority of the items in the sample to use for all items being stocked. The methods discarded in the selection process may contain some useful independent information. A combined forecast from two or more methods might improve upon the best individual forecasts. Studying how to select between methods and their combinations is an important direction of research. For example, the variance of the individual methods may be useful for how to select among methods and their combination. This gives us some insights into the process of data mining. There are a number of well-known methods in data mining such as clustering, classification, decision trees, neural networks, and so forth. Finding a good individual method from our tool kit to handle the data is clearly an important initial step in data mining. Then we should always bear in mind the power in combining the individual methods.

CONCLUSION

There are two kinds of direction to do the combination. The first one is basically a direct combination of the individual methods, such as simple average or "optimal" weighting. The other one is to classify our data first, and then select the weighting method. Classification is always an important aspect of data mining, and combination of forecasts sheds some new light on this. Another important message is that if we are dealing with large data sets, then it is not very worthwhile to find the "best" individual method. Obviously there may not be any best individual at all. A viable alternative is to find several sensible individual methods and then combine them as the final method. This approach will usually relieve much of our effort in finding the best individual method, as justified by the law of diminishing return. More recently, Hibon and Evgeniou (2005) conducted experiments on combining forecasts using 14 individual methods and all 3003 time series from the M3-competition (Makridakis & Hibon, 2000). Their empirical results indicate that,

…when we do not know which individual forecasting method is the best, selecting among combinations in practice leads to a choice that has, on average, significantly better performance than that of a selected individual method. Therefore the advantage of combinations is ... that selecting among combinations is less risky than selecting among individual forecasts. (p. 16)

REFERENCES

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Operational Research Quarterly, 20*(4), 451-468.

Bunn, D. W. (1985). Statistical efficiency in the linear combination of forecasts. *International Journal of Forecasting, 1*(2), 151-163.

Chan, C. K., Kingsman, B. G., & Wong, H. (1999a). The value of combining forecasts in inventory management—A case study in banking. *European Journal of Operational Research, 117*(2), 199-210.

Chan, C. K., Kingsman, B. G., & Wong, H. (1999b). A comparison of unconstrained and constrained OLS for the combination of demand forecasts: A case study of the ordering and stocking of bank printed forms. *Annals of Operations Research, 87*, 129-140.

Chan, C. K., Kingsman, B. G., & Wong, H. (2004). Determining when to update the weights in combined forecasts for product demand—An application of the CUSUM technique. *European Journal of Operational Research, 153*(3), 757-768.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583.

Hibon, M., & Evgeniou, T. (2005). To combine or not to combine: Selecting among forecasts and their combinations. *International Journal of Forecasting, 21*(1), 15-24.

Koning, A. J., Franses, P. H., Hibon, M., & Stekler, H. O. (2005). The M3 competition: Statistical tests of the results. *International Journal of Forecasting, 21*(3), 397-409.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., et al. (1982). The accuracy of extrapolation (time series) methods: Results of a forecasting competition. *Journal of Forecasting, 1*(2), 111-153.

Makridakis, S., & Hibon, M. (2000). The M3—Competitions: Results, conclusions and implications. *International Journal of Forecasting, 16*(4), 451-476.

Makridakis, S., & Winkler, R. L. (1983). Averages of forecasts: Some empirical results. *Management Science, 29*(9), 987-996.

Newbold, P., & Granger, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts (with discussion). *Journal of Royal Statistical Society, Series A, 137*(2), 131-149.

Terui, N., & Van Dijk, H. K. (2002). Combined forecasts from linear and non-linear time series models. *International Journal of Forecasting, 18*(3), 421-438.

Winkler, R. L., & Makridakis, S. (1983). The combination of forecasts. *Journal of the Royal Statistical Society, Series A, 146*(2), 150-157.

KEY TERMS

Combination of Forecasts: Combine two or more individual forecasts to form a composite one.

Constrained OLS Method: A method to estimate the "optimal" weights for combination of forecasts by minimizing the sum of squared errors as in a regression framework, and the weights are constrained to sum to one.

Data Mining: The process of selection, exploration, and modeling of large quantities of data to discover regularities or relations that are at first unknown with the aim of obtaining clear and useful results for the owner of the database.

Fixed Weighting: "Optimal" weights are estimated and are used unchanged to combine forecasts for a number of periods.

Highest Weighting: Use the individual forecast procedure that is given the highest weight in the fixed weighting method. This is not a combination. This method is equivalent to choosing the forecasting technique that is the best on the weight estimation period.

Rolling Window Weighting: "Optimal" weights are estimated in each period by minimizing the errors over the preceding *m* periods, where *m* is the length in periods of the "rolling window." The weights are then used to combine forecasts for the present period.

Simple Average Weighting: A simple linear average of the forecasts, implying equal weights for combination of forecasts.

Combining Local and Global Expertise in Services

Hannu Salmela
Turku School of Economics and Business Administration, Finland

Juha Pärnistö
Fujitsu Services, Finland

INTRODUCTION

Since the 1990s, services characterized by a considerable geographical distance between the service person and the customer have become increasingly commonplace. Banks and insurance companies are introducing call centers or service centers to complement, or even replace, the old regional service organization. In the information and communication technology (ICT) sector, companies such as Fujitsu and IBM provide part of the end-user support for their clients from a few centralized call centers. Telecommunications operators have established call centers to serve their customers in conducting basic business transactions. To a large extent, the change in the 1990s can be attributed to ICT development. As call centers and local offices have equal access to all the information, many of the services that previously had to be provided locally can now come from a call center. Furthermore, this decade will bring new technologies that will further enhance capabilities to serve customers over long distances. They will, for instance, provide increasingly rich media for interaction between customers and remote service personnel.

This article investigates factors that need to be considered when moving service production from regional offices to service centers. The empirical part of the study comprises a longitudinal analysis of the ways how Fujitsu Invia, a European IS company within Fujitsu Group, has transformed its service organization. The company has moved a long way from local, site-specific service units to national service centers, and ultimately to a few global centers that provide services to thousands of computer users worldwide. In retrospect, it can be said that the decision to centralize service production turned out to be very successful. However, the reasons why Fujitsu Invia decided to return part of the end-user support closer to customer sites illustrates the complexities associated with centralizing services that were previously produced locally.

BACKGROUND

The ability to centralize services appears to provide a cure for some of the traditional problems of service organizations. Managers of distributed service organizations are painfully aware of the difficulties to maintain an equal level of knowledge among all individual service persons in all regional offices. Centralizing the services to a call center or service center seems like an easy solution for ensuring that all customers receive equal service.

The more complex the services are, the more difficult it becomes to maintain equal knowledge in all regional offices. Hence, the analysis of forces for specialization among service staff is one of the key issues when considering the potential advantages of centralizing service production. Because of the special expertise necessary to solve the specific problems they encounter, professional service providers need a high level of specialization (Koelemeijer & Vriens, 1998). Factors that increase service complexity and thus pressure for specialization include (Mäkelin & Vepsäläinen, 1989), for instance,

1. diversity in customer needs and requests,
2. variety in the services available,
3. the number of situational factors that need to be considered,
4. uncertainty related to customer needs and circumstances,
5. the ability of the customer to define the services, and
6. the complexity of contracts used for governing the transactions.

In essence, complexity makes it difficult for a generalist service person to be able to handle all possible inquiries from all customers adequately. While generalists can deal with routine cases, specialists are needed to handle the difficult and unique cases. The main problem for producing complex services in regional offices is that it is difficult to maintain highly specialized knowledge in every regional office. The most forceful argument for establishing a service center is that the customer with a unique problem can talk

with a global specialist rather than with a local generalist. In addition, the cost savings that can be achieved in regional offices (office space, service personnel) are often sufficient to make the projects acceptable in terms of financial profitability measures.

In this chapter we suggest, however, that managers should also pay considerable attention to the opposite forces as well, that is, forces for providing local service. An obvious reason for providing services locally is that the service has a physical component and thus requires presence close to the customer. There are, however, many soft issues that may also make the customers prefer a local and personal service. In face-to-face discussions, information can be communicated with multiple cues like the sound of voice, facial expressions, and body language. Thus, the richness of communication media is very high (Daft & Lengel, 1986; Huang, Watson, & Wei, 1998). In this respect, the need to rely on conversations over phone or e-mail may have a negative impact on the quality of service.

A TYPOLOGY OF SERVICE ORGANIZATIONS

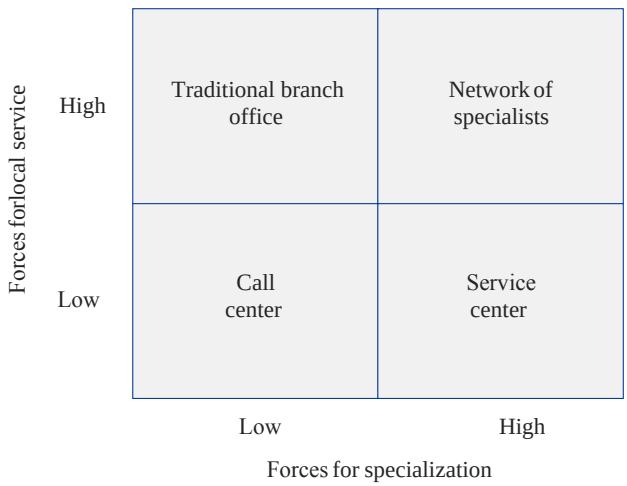
The typology that is suggested in this article is based on forces for providing local service and forces for specialization among service staff. Figure 1 provides a typology of four ideal types of service organizations, each designed to take into account the particular nature of service situations. Each organizational type has its strengths as well as its typical ways of using technology. More often than not, service organizations provide a mix of services for their customers. Some of the services provided may require local presence and some of them specialization. Thus, the objective is not to locate the whole company to a single quadrant. Rather, it is essential to identify all services that are provided to customers and to locate each of them to the right quadrant.

A *traditional branch office* is best suited to relatively simple service situations that require local presence. In simple services, the local generalist service persons are able to provide sufficient service quality. The quality and profitability of service is ensured by replicating and controlling precisely defined activity cycles, personal selling approaches, inventory control patterns, and counter display techniques (Quinn & Paquette, 1990). The possibilities for more stringent specialization are limited because a relatively small number of service personnel in one office have to deal with a full variety of customer inquiries.

In a *call center*, service personnel are centralized to one physical location. Part of the service personnel is moved from regional offices to a physical location to answer a definite set of requests from customers. The change is usually motivated by a managers' observation that some of the simple services are such that customers don't really expect to be served locally. The objectives for establishing a call center are a reduced cost of service, extended contact hours, and standardized quality. As the services provided are fairly simple, there is no need for specialization between service personnel: Any service person can solve any incoming customer inquiry.

A *service center* type of organization takes advantage of the fact that one centralized unit handling inquiries from

Figure 1. Simple typology of service organizations



a large number of customers allows greater specialization of service personnel. Each service person specializes in a particular type of customer problem, and incoming customer inquiries are routed to the best specialist in that area. The assumption is that a specialist with experience in solving similar problems with many companies can provide a more accurate service, particularly with regard to complex customer problems.

A *differentiated network* is the most challenging type of service organization as the service should be simultaneously local and specialized. Neither a local service office nor a centralized service center alone can provide the service. Finding a perfect organization in this quadrant is difficult. Often, organizations in this quadrant rely on a combination of traditional branch offices and centralized service centers (which deal with more complex cases). The potential for using sophisticated information and communication technologies is clearly highest in a differentiated network.

Even if the differentiated network may seem like a less attractive organizational solution, one should keep in mind that in service situations requiring both a local presence and specialization, there may be no alternative. In this situation, a service center may not be able to respond to local needs.

In fact, writings about multinational manufacturing enterprises often favor a differentiated network, which allows local responsiveness in different areas but possesses the control mechanisms to ensure global integration (Bartlett & Ghoshal, 1998; Nohria & Ghoshal, 1997; Prahalad & Doz, 1987). Rather than choosing a fully centralized or decentralized structure, the assets and resources of the company are widely dispersed but mutually supportive to achieve global-scale efficiency, the roles and responsibilities of organizational units are differentiated but interdependent to maximize national flexibility, and its knowledge and initiatives are linked through a worldwide learning capability that assures the efficient development and diffusion of innovations (Bartlett & Ghoshal).

With such practices, the organization provides a context for employees to create and share both tacit and explicit knowledge (Nonaka & Takeuchi, 1995; Polanyi, 1966). The relationships between individual actors are seen as a form of social capital that constitutes a valuable resource (Castells, 1996; Nahapiet & Ghoshal, 1998). In essence, it is asserted that creating and sharing knowledge in a geographically distributed organization is not only possible, but it may even be more effective than knowledge creation in one centralized point.

It seems apparent that many service organizations could benefit from similar practices. In fact, predictions of the impact of IT on service organizations have discussed the "spider's web" type of organization, where local offices are independent but are also able to use each other's knowledge resources as and when needed (Quinn & Paquette, 1990). Similarly, a recent study concluded that the competence to do global product development is both collective and distributed (Orlikowski, 2002).

DEVELOPMENT OF SERVICES IN FUJITSU INVIA

Fujitsu Invia (formerly ICL Invia) is a North-European vendor of information technology products and services. It belongs to Fujitsu Group, which operates in more than 100 countries and has almost 200,000 employees. Because the case describes developments mainly in the Fujitsu Invia, this brand name will be used below hereafter. According to a recently launched strategy, Fujitsu Invia has two main businesses: solutions and services. Solutions business refers to consulting, designing, and constructing information systems. Services business mainly involves operating customer's IT infrastructure as well as business applications. The case description focuses on service business and its evolution during the past few years.

Stage 1: All Services Produced Locally

Only a few years ago, Fujitsu Invia was mainly a hardware vendor. The main service was naturally hardware maintenance, which is typically a local service. The customers were provided with a service number to which they could call and report a hardware failure. The maintenance organization also delivered large installation projects, but not actual end-user support or administrative services (e.g., opening of a new e-mail address). These were typically delivered by the customer's own IT department. Maintenance was organized in regional business units that were quite independent. Customer-call management and the information system for managing service processes were typically shared by business units on a national level. Service coordination between countries was not really required, although certain issues such as pricing and service levels were negotiated on a corporate level.

Stage 2: Initiation of Centralized Service Production

In the second stage, customers became interested in acquiring relatively large service packages so that one vendor takes responsibility for a relatively large and logical part of the IT services. The change was initiated by customers who found managing IT and particularly the IT infrastructure increasingly difficult. These packages typically include a selection of IS services, for example (Weill & Broadbent, 1998),

- on-site support (i.e., solving end users' problems locally),

- help desk (i.e., a contact point where end users' requests are registered and escalated; also attempts to solve requests),
- hardware maintenance (i.e., fixing broken IT equipment),
- systems management (i.e., monitoring and operating systems typically from a centralized service center), and
- service management (i.e., customer reporting, maintaining contracts, improving service quality).

It was not possible to deliver these services by a maintenance organization, so in order to respond to this request, Fujitsu Invia established new business units. Centralized services (mainly help desk and systems management) are used to support the IT infrastructure. Centralization seemed like an obvious solution because of cost-effectiveness requirements and also because the technical expertise is easier to acquire and maintain in specialized teams. The aim was to minimize local service. Maintenance is the only local service. New technologies, like remote management, further improved the quality of service as end users no longer needed to describe the problems in detail. The help-desk specialist simply opened a remote connection to the end user's workstation, diagnosed the problem, and changed the parameters or trained the end user.

Stage 3: End-User Services Return to be Produced Locally

Although a centralized service center is probably the most cost-effective way to organize this type of services, there are also some problems. Customer organizations are not standardized really well even though the technological solutions are. Information systems can be installed in many ways, a situation that has implications on their operation. The quality of service perceived by end users is an even more significant issue. It seems that people prefer to be served locally. The barrier to call to a distant help desk and to an unknown specialist is rather formidable, even within a single nation. Furthermore, transferring the necessary knowledge about local environments to a help desk is not easy. Therefore, a local service organization is frequently required to manage situations like this. The end user easily feels that he or she has to call a help desk, which mainly escalates problems to a local organization. The added value of the service can be questioned.

Because of the problems described above, the help desk has been reorganized in some customer cases so that the first line is either integrated with a local service organization, or the help-desk first line is actually on the customer site. Some members of the on-site support team receive customer service requests, and other members undertake installations and problem solving. The roles within the team can be changed, which means that the on-site support team knows the local customer environment very well, and also the end users know on-site support people. This makes communication easier and the barrier to calling is lower. The solution rate has improved significantly in the pilot cases (in one case, from a level of 50% to 80%), which makes it possible to deliver the same service with fewer resources. At the same time, customer satisfaction improved (in the example case, from 6.5 to 8.4 when measured on a 1 to 10 scale). Their second-line and third-line support is still delivered from a centralized service center during the night and weekends.

Analysis of the Nature of Services

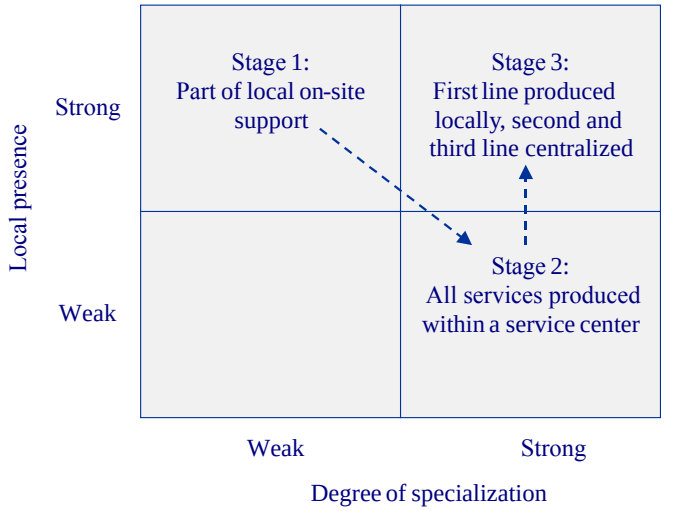
The most problematic service for Fujitsu Invia has been how help-desk services should be organized. What makes it problematic is that on the one hand, end users in customer companies prefer to contact local service persons instead of a distant call center. On the other hand, solving the problems often requires a specialized knowledge that the local service persons lack. The fact that the service should be both simultaneously local and also specialized has made the service difficult to organize, and the attempt to find the best means of organization has been a trial-and-error process (Figure 2).

In the first stage, the idea of providing help-desk services to end users was more or less informal. While end-user satisfaction for services was probably high, the customers felt that the cost of producing services locally was too high. Outsourcing these services to Fujitsu Invia and simultaneously centralizing them was considered an interesting option. The new technologies appeared to provide excellent tools for centralized production of help-desk services. When the service was centralized, however, more was learned about its nature. Service turned out to be more local than what the customers and Fujitsu Invia had expected. In the third stage, help-desk services for large-site customers are based on a combination of local and centralized services. While the experiences from this organizational form are highly positive, it is also more costly and can be applied only to customers with relatively large sites.

FUTURE TRENDS

The sociological theory of the postindustrial society was elaborated more than 3 decades ago (Bell, 1973; Touraine, 1969). In essence, the theory combined three postulates (Castells & Aoyama, 1994): (a) The source of productivity and growth lies in the generation of knowledge, (b) economic activity shifts from goods production to services delivery, and (c) the importance of occupations with a high information and knowledge content will increase.

Figure 2. Organizational arrangements for help-desk services in different stages



Existing postindustrial societies do not fully reflect all predictions made within this theory (Castells, 1996; Castells & Aoyama, 1994). Nevertheless, the significance of various types of services, for example, capital management services, services produced for industrial companies, and social and health services, is increasing. As firms in advanced economies have off-shored manufacturing jobs, the society as a whole has transformed toward a "service economy" model (Castells & Aoyama). Hence, the significance of services in general and knowledge-intensive services in particular is increasing.

The point raised in this paper is that information and communication technology will also induce a major transformation inside the service sector. Increasing telecommunications bandwidth together with its falling cost will enable new applications that are particularly useful for geographically dispersed production of services. In the future, the possibility to use wide computer screens for face-to-face interaction is likely to have a major impact on the way services are organized. Furthermore, the penetration of innovations such as integrated customer databases, e-services, global scheduling systems, call center technology, and virtual learning environments is likely to continue.

The emergence of call centers and service centers provides first visible evidence that the transformation is already taking place. However, as the Fujitsu Invia case illustrates, a call center can be too remote for providing services that require local presence. Hence, the organizational forms required in many services may be more complex than mere centralization of service production. The pressures for local responsiveness are likely to force service organizations to adopt similar complex network structures as some of the advanced manufacturing companies are already using.

This research was, however, based on only one project in one service organization. As service organizations both in the private and public sector have been slow in adopting new technologies, previous IS research has paid relatively little attention to these sectors. The need for further research about the ways in which organizational structures, processes, and information infrastructures are being transformed is evident.

CONCLUSION

This paper provides a simple typology of service situations that can be used when planning IT investments in service organizations. The use of the typology is illustrated through a longitudinal analysis of the ways in which Fujitsu Invia, a European IS service company, has transformed its service organization to better meet customer expectations and needs. Practicing managers can use the typology to analyze organizations' services, to investigate where services ought to be produced, and how ICT technology should be employed to support customer service situations. The authors hope that this article will be useful for people working in companies either providing a service business or aiming to enter a service business in an international context. For research, the typology, its dimensions, and the background theories identified provide some early steps on the way toward a more comprehensive understanding about the impact of ICT on service organizations.

REFERENCES

Bartlett, C. A., & Ghoshal, S. (1998). *Managing across borders: The transnational solution* (2nd ed.). Boston: Harvard Business School Press.

Bell, D. (1973). *The coming of postindustrial society: A venture in social forecasting.* New York: Basic Books.

Castells, M. (1996). *The rise of the network society*. Malden, MA: Blackwell Publishers Inc.

Castells, M., & Aoyama, Y. (1994). Paths towards the informational society: Employment structure in G-7 countries, 1920-90. *International Labour Review, 133*(1), 5-31.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

Huang, W., Watson, R. T., & Wei, K. (1998). Can a lean e-mail medium be used for rich communication? *European Journal of Information Systems, 7*, 269-274.

Huber, G. P. (1991). Organizational learning: The contributing processes and the literatures. *Organization Science, 2*, 88-115.

Koelemeijer, K., & Vriens, M. (1998). The professional services consumer. In M. Gabbott & G. Hogg (Eds.), *Consumers and services* (pp. 163-184). Chichester, England: John Wiley & Sons.

Mäkelin, M., & Vepsäläinen, A. (1989). *Palvelustrategiat: Palveluorganisaation kehittäminen ja tietotekniikka*. Jyväskylä: Gummerus kirjapaino Oy, Finland.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review, 23*(2), 242-266.

Nohria, N., & Ghoshal, S. (1997). *The differentiated network: Organizing multinational corporations for value creation*. San Francisco: Jossey-Bass Inc.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company: How Japanese companies create the dynamics of innovation*. Oxford, England: Oxford University Press.

Orlikowski, W. J. (2002). Knowing in practice: Enacting a collective capability in distributed organizing. *Organization Science, 13*(3), 249-273.

Polanyi, M. (1966). *The tacit dimension*. London: Routledge & Kegan Paul.

Prahalad, C. K., & Doz, Y. L. (1987). *The multinational mission: Balancing local demands and global vision*. New York: Free Press.

Quinn, J. B., & Paquette, P. C. (1990). Technology in services: Creating organizational revolutions. *Sloan Management Review, 33*(2), 67-78.

Sanchez, R., Heene, A., & Thomas, H. (1996). Introduction: Towards the theory and practice of competence-based competition. In R. Sanchez, A. Heene, & H. Thomas (Eds.), *Dynamics of competence based competition: Theory and practice in the new strategic management* (pp. 1-35). Exeter: Pergamon.

Storbacka, K., Strandvik, T., & Grönroos, C. (1994). Managing customer relationships for profit: The dynamics of relationship quality. *International Journal of Service Industry Management, 5*(5), 21-38.

Touraine, A. (1969). *La societe post-industrielle.* Paris: Denoel.

Weill, P., & Broadbent, M. (1998). *Leveraging the new infrastructure: How market leaders capitalize on information technology*. Boston: Harvard University Press.

KEY TERMS

Capabilities: Are repeatable patterns of action in the use of assets to create, produce, and/or offer products to a market (Sanchez et al., 1996).

Competence: An ability to sustain the coordinated deployment of assets in a way that helps a firm to achieve its goals (ability here is used in the ordinary language meaning "of a power to do something"; Sanchez, Heene, & Thomas, 1996).

Information: Refers to data that give meaning by reducing ambiguity, equivocality, or uncertainty, or data that indicate that conditions are not presupposed (Huber, 1991).

Knowledge: Refers to interpretations of information, know-how, and beliefs about cause-effect relationships (Huber, 1991).

An entity **learns:** If, through its processing of information, the range of its potential behaviors is changed. The information processing can involve acquiring, distributing, or interpreting information (Huber, 1991).

Perceived Service Quality: Refers to customers' cognitive evaluation of the service across episodes compared with some explicit or implicit comparison standard (Storbacka, Strandvik, & Grönroos, 1994).

Skill: Understood as a special form of capability, with the connotation of a rather specific capability useful in a specialized situation or related to the use of a specialized asset (Sanchez et al., 1996).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 457-463, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Communicability of Natural Language in Software Representations

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

In software engineering, separating problem and solution-level concerns and analyzing each of them in an abstract manner are established principles (Ghezzi, Jazayeri, & Mandrioli, 2003). A software representation, for instance, a model or a specification, is a product of such an analysis. These software representations can vary across a formality spectrum: informal (natural language), semi-formal (mathematics-based syntax), or formal (mathematics-based syntax and semantics).

As software representations become pervasive in software process environments, the issue of their communicative efficacy arises. Our interest here is in software representations that make use of natural language and their communicability to their stakeholders in doing so. In this article, we take the position that if one cannot communicate well in a natural language, then one cannot communicate via other, more formal, means.

The rest of the article is organized as follows. We first outline the background necessary for later discussion. This is followed by the proposal for a framework for communicability software representations that are created early in the software process and the role of natural language in them. We then illustrate that in software representations expressed in certain specific nonnatural languages. Next, challenges and directions for future research are outlined and, finally, concluding remarks are given.

BACKGROUND

The use of natural language in software is ubiquitous. In spite of its well-known shortcomings that are primarily related to the potential for ambiguity or limitations of automatic verifiability, surveys have shown (Berry & Kamsties, 2005) that the use of natural language (such as English) continues to play an important role in software representations. Agile software process environments such as Extreme Programming (XP) (Beck & Andres, 2005) tend to accentuate the use of "lightweight" models such as user stories (for input to software requirements and test cases) (Alexander & Maiden, 2004) that often depend exclusively on the use of natural language. In Literate Programming, natural language prose is used in documenting (explaining) the source code as if it were the work of literature to make it more readable to both humans as well as to machines.

In recent years, there has been an increasing emphasis on the quality of software representations that are created "early" from the point of view of control and prevention of problems that can propagate into later stages. The significance of expressing software requirements in natural language that minimizes ambiguity is emphasized in Berry and Kamsties (2005) whereas a model to study their quality is presented in (Fabbrini, Fusani, Gervasi, Gnesi, & Ruggieri, 1998). However, these efforts do not systematically address the issue of communicability in software representations.

USE OF NATURAL LANGUAGE IN SOFTWARE REPRESENTATIONS

Our understanding of communicability of a software representation is based on the following interrelated hypotheses:

- **Hypothesis 1.** Readability is a prerequisite to communicability, which in turn is a prelude to comprehensibility. The basis for this hypothesis is that if a user has difficulty accessing or deciphering a certain message, then that user will not be able to understand it in part or in its entirety either.
- **Hypothesis 2.** The comprehension of a given software representation in a nonnatural language takes place only when it is first internally translated into the user's natural language of choice. The basis for this hypothesis is that the mode of internalization of some knowledge is the conversion of explicit knowledge into the tacit knowledge. The natural language acts as a "proxy" in this internalization of knowledge inherent in a semiformal/formal software representation.

Using these as the basis, the discussion of software representations that follows rests on the framework given in Table 1.

Table 1 provides necessary but not sufficient conditions for communicability. We now discuss the elements of the framework in detail.

Table 1. A high-level view of communicability of a software representation making use of natural language

Entity	Natural Language Use in Software Representation	
Pragmatic goal	Communicability	Feasibility
External quality attributes	Readability, other	
Internal quality attributes	Secondary notation (labeling, typography)	

Communicability and Semiotics

Semiotics (Nöth, 1990) is concerned with the use of symbols to convey knowledge. From a semiotics' perspective, a representation can be viewed on three interrelated levels: syntactic, semantic, and pragmatic. Our concern here is the pragmatic level, which is the practical knowledge needed to use a language for communicative purposes.

Readability

For our purposes, readability (legibility or clarity) is the ease with which a stakeholder can interpret a piece of text character by character. In general, readability applies to both textual statements and nontextual constructs (say, a histogram), but we shall restrict ourselves to the former.

Secondary Notation

The *secondary notation* (Petre, 1995) is one of the cognitive dimensions and, in our context, defined as the stylistic use of the primary notation (that is, the normative syntax) for perceptual cues to clarify information or to give hints to the stakeholder. The nonmutually exclusive natural language secondary elements that affect the communicability of artifacts are labeling and typography.

- **Labeling.** Labels are comprised of names and, in general, names are not useful by themselves (Laitinen, 1996). For example, names such as `iIndex` for an integer variable reflect their type rather than their purpose or role. Use of the terminology of the application domain in text labels and metaphors (Boyd, 1999) makes it particularly easier for nontechnical stakeholders or users new to language extension to become familiar with the representation. Furthermore, these labels will be more readable and reduce possibilities of misinterpretations if they follow a *natural naming* scheme. Natural naming (Keller, 1990) is a technique initially used in source code contexts that encourages the use of names that consist of one or more full words of the natural language for program elements in preference to acronyms or abbreviations. For example, `QueueManager` is a combination of two real-world metaphors placed into a natural naming scheme.
- **Typography.** The positioning (layout, justification, space between words and lines) of text impacts readability in any document context. Of special concern in our case are the choice and the sequence of characters in the use of text that can affect readability. For example, the characters in a name like `O0lI1` are hard to distinguish and therefore difficult to read. The choice of fonts used for labeling depends on a variety of factors (serif/sans serif, kerning, font size, and so forth) that are important for legibility. Color can be used as an emphasis indicator and for discriminability in text. For example, by associating different colors with text, a stakeholder can be informed of the semantic similarity and differences between operations and attributes in two object classes.

Feasibility

By acknowledging that there are time, effort, and budgetary constraints on producing a software representation, we include the notion of feasibility as an all-encompassing factor to make the framework practical. There are well-known techniques such as Analytical Hierarchy Process (AHP) and Quality Function Deployment (QFD) for carrying out feasibility analysis, and further discussion of this aspect is beyond the scope of this article. Any feasibility analysis, however, also needs to be in agreement with the organizational emphasis on decision support for software engineering in general.

Use of Natural Language in UML

The Unified Modeling Language (UML) (Booch, Jacobson, & Rumbaugh, 2005) is a standard language for modeling the structure and behavior of object-oriented software. The issue of communicability of UML models has been addressed in (Kamthan, 2006).

Natural language enters UML models in several places, especially metadata in the form of annotation (using UML Note construct); class/object, operation, and attribute names; state and transition names; extensions (tags, stereotypes, constraints); and so on. Natural naming can be useful in these cases in general, but particularly when application domain-specific UML extensions are used, as even a UML-aware stakeholder may not be familiar with them. For example, it would be difficult for a reader to decide if the stereotypes `<<TP>>` and `<<TermProc>>` stand for terminal process or terminate process.

We now briefly discuss, from a UML viewpoint, two of the software representations that are typically created earliest in a software process.

Domain Models

A domain model is a vocabulary of the domain, including key concepts, their properties, and the relationships between the concepts necessary to describe the domain sufficiently enough for the purpose at hand. The UML class diagrams at a high level of abstraction (only class and attribute names) and relationships types (association, generalization, aggregation, and composition) are often used for constructing domain models (Larman, 2004), and these constructs are amenable proper use of secondary notation. To distinguish concepts, properties, and relationships different case schemes for their natural names can be adopted. For example, the concept `PaymentSystem --{isComposedOf}-->` the concept `CreditCardSystem` (with attributes/types `cardtype:String` and `valid:Boolean`).

Use Case Models

A use case reflects external observable behavior of the interaction of an actor (human, program) with the system (Kulak & Guiney, 2004). The development of a use case model and domain model go hand in hand (Larman, 2004). Actor names, use case names, and interaction (stimulus, response) message labels can all benefit from proper use of secondary notation. For example, actor names should represent roles rather than designations or positions of people, and use case names should reflect a domain concept rather than a system or interaction mode-dependent entity. Therefore, in a Course Registration System, we would have actor names like `RegisteredStudent` (rather than `JohnSmith`) and use case names like `CourseScheduler` (rather than `CourseInterfaceMenu`).

The natural language use in previous models can be generalized to other UML modeling contexts such as macro- and micro-architecture design models.

Use of Natural Language in XML

The Extensible Markup Language (XML) (Bray, Paoli, Sperberg-McQueen, Maler, & Yergeau, 2004) is a metamarkup mechanism that lends a suitable basis for concrete serialization syntax for expressing information within a software representation. XML is an exemplary of descriptive markup that defines the notion of a document in terms of the ordered hierarchy of content objects. The content objects (elements) of an XML document can be associated with properties (attributes). Both elements and attributes are labeled using mnemonic names, based on author discretion and usually inspired by the domain being addressed.

As an example, the following shows a fragment of a simple Requirements Markup Language (RML). The root element is `rml` (that is an acronym rather than a natural name), `id` and `lang` are attributes (that are abbreviations for identification and language, respectively), while names of other elements and attributes do follow natural naming.

```
<rml version="1.0" xml:lang="en">
 <requirement id="..." type="..." priority-level="...">
  <statement>...</statement>
  <rationale>...</rationale>
 </requirement>
</rml>
```

The descriptive markup in its source form, particularly when there are complex (usually nonlinear) structural relationships involved, is not considered very readable. This is ameliorated to a certain extent by rendering tools that highlight the encapsulated content and suppress the markup. (This, for example, is similar to user agent processing a HyperText Markup Language [HTML] document.) Specifically, XML does not have native support for presentation semantics and therefore relies on ancillary mechanisms such as the Cascading Style Sheets (CSS) for presenting XML documents on different devices and user agents.

Use of Natural Language in OWL

The declarative knowledge of a domain is often modeled using ontology. The Web Ontology Language (OWL) (Dean & Schreiber, 2004) is an ontology specification language designed for the Semantic Web.

An OWL ontology can be used as an alternate to the traditional UML-based domain models with better opportunities for expressiveness and reasoning of knowledge. A typical OWL ontology has classes and relationships between those classes. These classes and relationships can also have their own unique properties modeled in OWL by datatype and object properties, respectively. The labels of (some but not all) classes, relationships, and properties are author defined and could benefit from natural naming. Since OWL is based

Figure 1. A box structure presentation of a Z schema

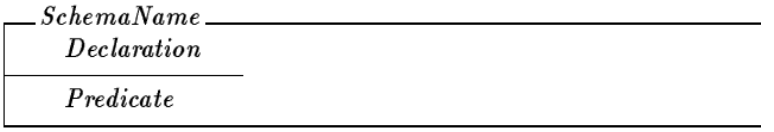


Figure 2. A fragment of the Birthday Book System in Z

$$[NAME, DATE]$$

$$\begin{array}{|l} \textit{BirthdayBook} \\ \hline known : \mathbb{P}\, NAME \\ birthday : NAME \pfun DATE \\ \hline known = \mathrm{dom}\, birthday \end{array}$$

on XML, it inherits both the advantages and limitations of the natural language use in it.

Use of Natural Language in Z

Z is a widely used formal specification language for describing sequential software systems, especially those that are safety critical. Z has traditionally used LATEX mathematically typesetting language for its syntax and relies on first-order predicate logic and typed set theory for its semantics.

The primary construct for structuring and reusing mathematical notation in Z is the schema. A schema is divided into two parts (Figure 1): (1) a declaration part (signature) and (2) a predicate part (axioms). The declaration part of the schema contains variables representing the before and after states, input and outputs. It consists of declarations of the form *s : T*, where *s* is a state variable of type *T*. The predicate part of the schema consists of predicate logic expressions that define the relationships between the declared variables. It defines the relationship between the before and after states of an operation. The names of state variables and operations are chosen by the author of the Z specification and can benefit from proper use of secondary notation.

Figure 2 illustrates part of the well-known example of the birthday book system (Spivey, 2001) in Z. It contains the type definitions and the state of the system modeled as a state space schema. The types `NAME` and `DATE` are defined as basic types. The variable known is a set of names that records the birth dates and birthday is a partial function from `NAME` to `DATE` . The invariant condition states that the variable *known* can be derived from the value of *birthday*. We note the use of natural naming here. For instance, the name `BirthdayBook` is more intuitive than the name `BB` , even though they are both syntactically legal under Z.

Abbreviating a Z schema can have a negative impact on readability. For instance, in Figure 2, the invariant condition states that the variable *known* can be derived from the value of *birthday*. Therefore, although it is possible to specify this part of the birthday book system without any reference to the variable *known*, introducing redundancy by giving a name to the concept *known* makes the specification more readable.

Challenges to Natural Language Use in Software Representation

There are a few challenges in the use of secondary notation pertaining to the use of natural language in software representations.

It is obvious that over use of natural naming will increase the file size of a software representation. Also, once it is known that stakeholders are aware of a certain acronym or abbreviation, natural naming may not be necessary and even counter productive if used. It is straightforward to automatically check for spelling errors in names but not for pseudonyms or homonyms.

As for typography, fonts specifically designed for presentation on paper may in general be hard to read on a computer screen. Often fonts created by the manufacturer of one operating system are not portable to others. Any use of color should take into account the variations in the interpretation of primary colors by computer monitors, contrast between background and foreground, the way people with color vision deficiency read text, and the possibility that diagrams may be printed on a black and white printer.

The automated use of tools eases much of the tedium involved in the task of improving the use of natural language in software representations. Refactoring (Wake, 2004) provides transformation methods for eradicating the undesirables from

software representations while preserving their semantics; and the support for it in modeling tools is on the rise. However, getting consensus on the notion of "semantics" has proven difficult and the process of refactoring "impurities" related to names can only be partially automated.

FUTURE TRENDS

The discussion of the previous section was limited to one natural language setting, namely English. As internationalization of software increases, exploring the role of non-English languages in software representations would be of interest. This is already a challenge if one needs to preserve natural naming in descriptive markup context. For example, an element or attribute name for Québec Déjà Vu has to be written as `Qu&eacute;becD&eacute;j&agrave;Vu`, or a variant thereof, which is not in the spirit of readability support in natural naming. This could even be more challenging in languages where pictographs rather than words are used for communication.

Besides the secondary notation, it would also be of interest to examine the impact of cognitive dimensions pertaining to language characteristics (such as the number of elements in the syntax, their discriminability, frequency of their use in software representations, steepness of the learning curve, and so forth) towards readability in special cases such as UML, XML/OWL, and Z.

Finally, it would be useful to consider the issues related to secondary notation, readability, and eventually communicability of heterogeneous specifications that contain interoperable fragments from different semiformal/formal languages. In such cases, XML could be used as uniform serialization syntax, however, the issue of a common basis for semantics is still a challenge.

CONCLUSION

The computer software stakeholder "family" has broadened over the years to include nondomain experts and noncomputer scientists/software engineers. If documentation was the "castor oil" of the early years of programming (Weinberg, 1998), representations are the castor oil of software engineering today.

The natural and formal representations of software *complement* each other and should be seen as such. We simply cannot abandon our natural language knowledge and skills for the sake of novelty. If the move from informality to formality in software representations is to reduce unpredictability, then we cannot afford "ad hoc-ness" to creep in.

In conclusion, the purpose of software representations from a stakeholder viewpoint is communication and natural language (still) matters. The burden of accomplishing that largely rests on the author. An optimal and systematic means of the use of secondary notation is essential for readability of software representations. Although there is no need to abandon abbreviations or acronyms in names, it is important to realize that succinctness has its trade-offs. A labeling scheme relying on natural naming, when deployed appropriately, provides a means for communicable names in software representations. A feasible plan for typography that takes the stakeholder environment into consideration only strengthens this.

REFERENCES

Alexander, I., & Maiden, N. (2004). *Scenarios, stories, use cases through the systems development life-cycle*. John Wiley & Sons.

Beck, K., & Andres, C. (2005). *Extreme programming explained: Embrace change* (2nd ed.). Addison-Wesley.

Berry, D. M., & Kamsties, E. (2005). The syntactically dangerous all and plural in specifications. *IEEE Software, 22*(1), 55-57.

Booch, G., Jacobson, I., & Rumbaugh, J. (2005*). The unified modeling language reference manual* (2nd ed.). Addison-Wesley.

Boyd, N. S. (1999). Using natural language in software development. *Journal of Object-Oriented Programming, 11*(9).

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., & Yergeau, F. (2004). *Extensible markup language (XML) 1.0* (3rd ed.). W3C Recommendation. World Wide Web Consortium (W3C).

Dean, M., & Schreiber, G. (2004). *OWL Web ontology language reference*. W3C Recommendation. World Wide Web Consortium (W3C).

Fabbrini, F., Fusani, M., Gervasi, V., Gnesi, S., & Ruggieri, S. (1998). Achieving quality in natural language requirements. *The 11th International Software Quality Week (QW'98)*, San Francisco, USA. May 26-29, 1998.

Ghezzi, C., Jazayeri, M., & Mandrioli, D. (2003). *Fundamentals of software engineering* (2nd ed.). Prentice Hall.

Kamthan, P. (2006). How useful are your UML Models? *The 2006 Canadian University Software Engineering Conference (CUSEC 2006)*, Montreal, Canada, January 19-21, 2006.

Keller, D. (1990). A guide to natural naming. *ACM SIGPLAN Notices, 25*(5), 95-102.

Kulak, D., & Guiney, E. (2004). *Use cases: Requirements in context* (2nd ed.). Addison-Wesley.

Laitinen, K. (1996). Estimating understandability of software documents. *ACM Software Engineering Notes, 21*(4), 81-92.

Larman, C. (2004). *Applying UML and patterns: An introduction to object-oriented analysis and design and the unified process* (3rd ed.). Prentice Hall.

Petre, M. (1995). Why looking isn't always seeing: Readership skills and graphical programming. *Communications of the ACM, 38*(6), 33-44.

Spivey, J. M. (2001). *The Z notation: A reference manual*. Prentice Hall.

Wake, W. C. (2004). *Refactoring workbook*. Addison-Wesley.

Weinberg, G. M. (1998). *The psychology of computer programming (silver anniversary ed.)*. Dorset House.

KEY TERMS

Cognitive Dimensions of Notations: A generic framework for describing the utility of information artifacts by taking the system environment and the user characteristics into consideration.

Descriptive Markup: A model of text that focuses on description of information using markup delimiters for consumption by both humans and machines.

Domain Model: A simplified abstraction from a certain viewpoint of an area of software interest.

Formal Specification: A software representation with well-defined syntax and semantics that is usually used to express software requirements or detailed software design.

Natural Naming: A technique for using full names based on the terminology of the application domain that consists of one or more words of the natural language instead of acronyms or abbreviations for elements in a software representation.

Ontology: An explicit formal specification of a conceptualization that consists of a set of terms in a domain and relations among them.

Software Representation: A representation of the system expressed at some level of abstraction, in some modality, and at certain level of formality.

Communication Integration in Virtual Construction

O.K.B. Barima
University of Hong Kong, Hong Kong

INTRODUCTION

In recent times the use of the virtual and allied concepts in the delivery of tasks in the construction industry has received attention in literature (Barima, 2003). However it appears there is the lack of integrated models to support the study and practice of communication in the physical world (the place) and the virtual realms (the space) in the construction industry. This chapter seeks to evolve frameworks which integrate communication paths in the traditional construction place with that of the construction space. This approach may provide an integrated perspective to support the study and practice of communication in the place and space in the construction industry. First, a review of traditional communication theories is done, and then integrated models of the construction place and space are presented, before an integrated matrix of potential communication paths across likely construction work scenarios is given. The suggestions for future studies and the conclusions to this chapter are preceded by a map of the potential communication conflicts (or agreements) in both realms (i.e., physical and virtual).

BACKGROUND

Communication Theories

The study and practice of communication is not a new science. It has been claimed that the formal origin of this science may be traced to an essay in the year 3000 BC (Bryant, 2004). In this essay advice was given to the son of an Egyptian king on how to communicate effectively (Bryant, 2004). Other notable historical contributors to the communication theory are Aristotle, who advanced the works of Heraclitus, and Socrates and Plato in rhetoric theory (Bryant, 2004).

Communication theory has evolved over the years since the days of Aristotle, and seven traditional communication typologies are now established in communications studies. The seven traditions are: (1) the semiotic, (2) sociopsychological, (3) cybernetic, (4) phenomenological, (5) sociocultural, (6) rhetorical, and (7) critical traditions (Craig, 1999; Griffin, 2003). There has been suggestions to add areas like ethical, economic, aesthetic, and so on to the seven communication traditions, however there seems to be a consensus that the seven traditions have the capacity to explain (to a greater extent) the research and practice of communication (Griffin, 2003). The seven traditions of communication theory may be summarized as follows (Craig, 1999; Griffin, 2003):

- Semiotic involves the study of signs. According to the semiotic tradition, meaning does not reside in words or other symbols, but it resides in people. Words, for example, are perceived to be arbitrary symbols with no latent meaning.
- In the sociopsychological tradition communication is seen from the perspective of interpersonal influence. Studies focus on the cause-and-effect factors in relationships, so as to understand which communication behavior will succeed.
- Cybernetic tradition perceives communication as the link which binds the components of a system together, and also as a means for processing information.
- The phenomenological tradition emphasizes (among others) the personal experience of communication and that of others via discourse.
- The sociocultural perspective of communication is very helpful in understanding the gaps in culture which may exist among parties. This tradition perceives communication to involve the creation and realization of social reality, and culture is created and recreated when people talk.
- The rhetorical tradition concerns (among others) the art of public speaking or practical art of discourse.
- The critical theorists are against the use of language to attain imbalances in power; the blind acceptance of scientific methods and empirical findings without criticisms. They theorize communication via discursive reflection.

In relating the summary of the seven communication traditions to the theme of this chapter, communication may be seen as the information processing glue which may connect actors in a construction project together. This process may involve emphasizing: the communication experience of construction actors; interpersonal influences; and the effective communication of meaning which resides in the actors and their effect on successful communication. Communication may also be put under the lens of critical thinking/analysis to advance any desired communication

objectives in a construction project. In a construction project the objectives may include effective results delivery across both the construction space and place. Another important communication aspect may concern ethical issues in the virtual construction project. Ethical issues may require critical attention in the communication processes in both the space and place. With the track of traditional communication studies and practice reviewed, the next section discusses recent developments in communication and the virtual construction project environment.

Communication and Virtual Construction Projects

Communication in the construction industry has been directly or indirectly studied by a number of scholars (Jaggar, Ross, Love, & Smith, 2001). Over the years most of the studies have focused on traditional communication systems which use media like face-to-face collocation (see e.g., Jaggar et al., 2001). However, recent revolutionary developments in the information and communication technology (ICT) sector have provided significant transformations in the manner in which construction project actors are able to communicate and share information. Concepts like virtual construction where construction actors may rely on ICT to function irrespective of time and space to deliver common goals have emerged (Barima, 2003). The two major means used to support communication in virtual construction environments are either via fixed or mobile ICT terminals. For example, fixed computer terminals may be used to access virtual construction project environments like project Web sites via the Internet (Andresen, Christensen, & Howard, 2003). Another access to the virtual construction environment may be via mobile ICT tools. Certain earlier studies on the virtual concept used or implied the use of fixed ICT terminals to access the virtual environment (Caneparo, 2001; Clayton, Warden, & Parker, 2002). This review will emphasize on the most recent trends in this field of communication, and this appears to be on the use of mobile computing in the construction industry. The next paragraph discusses the use of mobile computing in the construction industry.

In recent times scholars have given recognition to the important use of mobile and allied facilities to support the exchange of information in the construction industry (Johanson & Törlind, 2004; Kuladinithi, Timm-Giel, & Görg, 2004; Olofsson & Emborg, 2004; Rebolj & Menzel, 2004; Ward, Thorpe, Price, & Wren, 2004). Various proposals for the empirical use of mobile computing have been made to assist in fieldwork, partner integration, supervision, scheduling, and less formal specifications (Olofsson & Emborg, 2004; Rebolj & Menzel, 2004). Also the utility in the use of portable computing tools like mobile phones to access databases, CAD drawings, and hold video conferences among parties (either on construction sites or via remote means) without being burdened by any extra hardware have been reported (Kuladinithi et al., 2004). At the construction site level, the use of mobile devices to capture and store data for easy and timely access to the flow of information and also manage projects for cost reduction and performance improvement have also been examined (Ward et al., 2004). Despite their noted limitations like communications costs, bandwidth and coverage, and so forth (Johanson & Törlind, 2004; Kuladinithi et al., 2004), mobile computing also appears to hold promise for the future of communication in the construction industry just like the use of fixed terminals/tools.

The Need for Integration

Certain scholars have directly or indirectly noted the potential benefits of the virtual construction concept in task delivery (Barima, 2003; Savioja et al., 2003; Sulankivi, 2004). For example, the use of the virtual concept to support construction works delivery among dispersed parties may potentially provide cost and time savings (Barima, 2003). However, it is also essential to remember that the nature of construction work makes traditional communication media like rich physical face-to-face interaction very important. The two scenarios therefore necessitate the integration of communication processes in both the physical and virtual realms in the construction industry. This will remind, structure, and focus communication studies and practice within the construction industry. In reality the two may support each other, and the strengths of one may be leveraged to support the weaknesses of the other. Further, the rich lessons learned from the traditional systems (in particular) could also aid communication in the construction place and space. The next sections introduce models for the structured integration of the two worlds.

MODEL PRESENTATION

Construction Space and the Construction Place

Rayport and Sviokla (1999) have identified and classified two worlds where business communications may occur as the market place and market space. In a similar analogy construction activities (or value delivery mechanisms) may also occur within the construction place and construction space. Figure 1 demonstrates the two realms of operation. Each of the two realms (construction space and construction place) have two dimensions which may be characterized as mobile and fixed: locations or spaces (see Figure 1). The construction place may refer to the traditional physical places/sites where physical construction activities/delivery may occur, while the fixed and mobile construction space refer to the

use of ICT tools to support task deliveries. The boundaries of communication within the construction place (fixed and mobile) and the construction space (fixed and mobile) are shown in Figure 1.

As may be seen in Figure 1, diverse communication paths may be identified in the diagram. Communication in the fixed and mobile construction places are respectively shown as PF (1…n) and MP (1…w) across varied times (t1, t2, t3, etc.). In a similar manner potential communication across the mobile and fixed construction space are respectively shown as M (1…t) and F (1…z) across times t1, t2, t3 respectively. Communication may be via synchronous or asynchronous modes across varied times (t1, t2, t3, etc.) depending on any given context.

There may also be intra- or inter-boundary communication across the identified construction space and locations. For example, intra-boundary communication between parties in the fixed construction place may be limited to a particular location (e.g., site PF1); or across construction sites PF 1 and PF2 which are within walking distance of each other. Inter-boundary communication may occur between parties across each of the four boundaries. For example, this may occur from a mobile construction space M1 to: a fixed construction space F1, or a fixed construction place PF1, or a mobile construction place MP1. Other potential paths are also shown across all the identified boundaries.

Complex communication paths may also be potentially viable between communication objects at both the construction place and space. Objects may refer to humans or machines, or combinations of both. There may be transactions involving: one-to-one communication, one-to-many communication, many-to-many communication, and so forth either within or across the identified boundaries. For example, machine-to-machine, construction task interaction may occur in an electronic data interchange (EDI) transaction in the procurement of say construction materials, and so on. Human-to-human transaction may also, for example, occur between two parties within the construction space (say chat room).

There may also be differences in communication within (or across) the construction space and places, and not all communication types may suit specific types of information exchanges. Time may also become a critical dimension in the choice of any type of communication (i.e., synchronous/asynchronous) within the identified boundaries in the diagram (Figure 1) to suit any given context. Communication across the identified boundaries may also sometimes require critical thought. For example, what may be the effect of differences in culture, language, and so forth on the communication of meaning across potential diverse geographic distances? Ethical issues on potential use of information (which may be easily gathered and stored) may also be another dimension for critical consideration, especially in the space. The next section presents a matrix of potential communication paths across various scenarios and also potential conflicts (or agreements) in communication.

Matrix of Potential Communication Paths

Many types of communication may exist at different operational levels in the construction project process. Table 1 shows a matrix of potential communication paths across varying traditional levels and functional jobs within a construction project. This matrix may be combined with the concept of the construction space and place. As shown in Table 1 communication may be formal or informal, and may also occur within (or outside of) the boundaries of the construction project environment. For example, formal communication paths may be required for the execution of technical construction tasks, while informal communication may be needed for social and other purposes.

Figure 1. Communication paths in the construction space and place

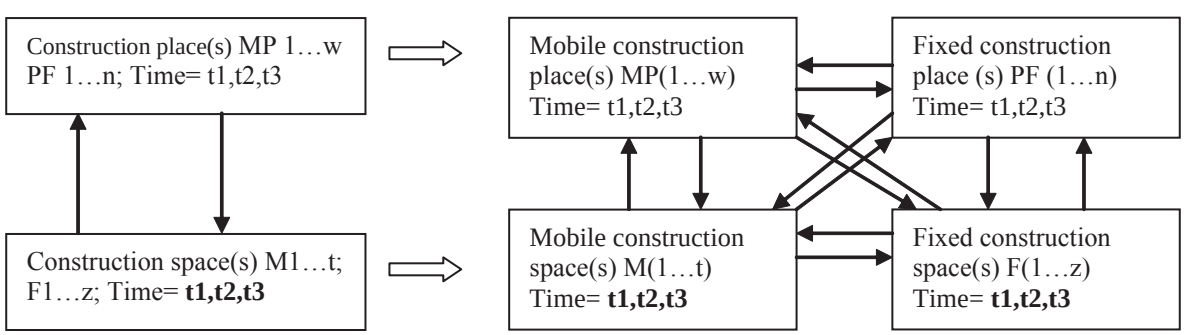


Table 1. Matrix for understanding communication paths between objects

	Project environment				Outside the project environment (ICT, business environment, etc)
	Formal (virtual/physical)		Informal operations		
Project level (1….n)	Function 1	Function (2…n)	Social, etc		
Strategic level	A	B	C	D	E
Management level	G	H	J	K	L
Operational level	M	N	O	P	Q
Other					

Figure 2. Map of potential conflicts (or agreements) in construction places and space

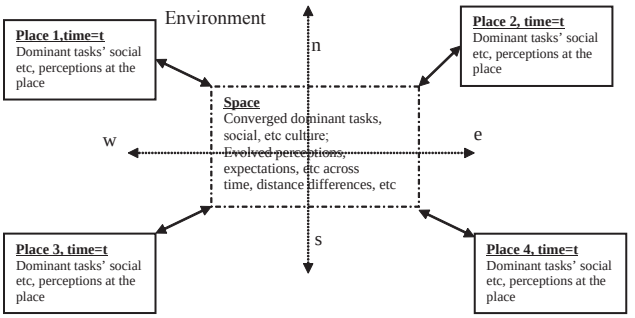


Table 1 also shows communication paths at the strategic, managerial, or operational levels for interactions within (and across) functional areas 1, 2,…n, and also the social context. Two types of paths are shown: (1) one in bold lines (to show physical communication interaction) and (2) the other in broken lines (to indicate virtual communication/interaction). For example, path AHO may refer to formal and informal communication in the construction space across the strategic, managerial, and operational levels for the different functional areas (1..2...n) and also the social context. Path distance AB may refer to say a machine-to-machine, EDI communication initiated at the strategic level across two functional groups. Path AE may refer to a communication path in the construction space between say a party from the strategic functional area 1 and another party not within the project's environment. This interaction along path AE may, for example, represent transactions for the procurement of goods using e-commerce. Path GL in a bold line may refer to physical face-to-face communication between people at the managerial level and other potential public stakeholders outside of the project's environment. Each of the square boxes in Table 1 also provide potential for tailoring information delivery to meet specific needs of users, and also grading access to certain types of information (in the space if necessary).

Potential Conflicts (or Agreements) in Communication

The diagram in Figure 2 shows the map of a typical example of potential communication conflicts (or agreements) across the construction places and space. In construction projects there are bound to be dominant local task cultures, language, legal, and perceptions/expectations. These perceptions and expectations may act as latent influence on the dispersed construction space from the local construction places 1, 2, 3, and 4. This latent influence may require effective management, training, and so forth to transform any potential chaos into a synergistic culture in the construction space for effective task delivery.

Figure 2 also shows feedback arrows between the indicated constructs to remind construction actors about the potential effect of the evolved culture (adopted/adapted) within the construction space on the indicated construction places. For example, a relatively easy access to information (and also potentially blurred communication paths) in the construction space may psychologically affect construction actors, who may operate in very rigid formal settings in the construction place. This psychological impact may continue in a dynamic iterative fashion throughout the project until dynamic equilibrium is attained. Another example may concern the sharing of technical information. Communication of task information via 3-D or 4-D graphics/modeling in the construction space may be used to support construction operations across the varying construction places. The use of 3-D or 4-D graphics/modeling in the construction space may potentially impact on operations at the construction places which may have the rooted culture of working with 2-D drawings, and so forth. There may be the need for effective change management to mitigate the effects of potential resistance (and even sabotage) in the use of 3-D or 4-D model applications in the affected construction places which may prefer working with 2-D drawings.

Other potential conflicts to consider may be technical units of measurement; operational laws; time zones; and economic, political, and ethical issues which may indirectly affect the construction space from the construction places 1, 2, 3, and 4. In summary the potential interaction between the environ-

ments of the physical distributed construction places and the evolved construction space may require critical attention and management to avoid most of the identified threats.

FUTURE TRENDS

The models in this article provide varied research opportunities. For example, future studies may be based on the models to consider how dominant local construction cultures in distributed geographic construction places could shape communication flow in the construction space. Focused case studies may also be used to explore the dynamics of why and how specific communication cultures may evolve in the construction space and place for task delivery in varied scales of construction projects. Lessons from such case studies could form good basis for the evolution of metrics to improve the integrated communication processes in the construction place and space for effective task delivery.

CONCLUSION

This article discussed the communication paths within the space and place in virtual construction projects and argued for the need to integrate both realms to aid structure and remind and focus communication theory and practice. Although communications in the construction space may usefully support construction task delivery, the nature of construction work makes traditional communication processes in the construction place also very important. Various models were proposed in the article to illuminate the concepts of integrated communication in the construction place and space. From an integrated perspective, the potential communication conflicts (or agreements) across the construction places and space on communication and task delivery processes were also explored. The integrated view provides a better understanding of the potential effect of latent occurrences in each world on the other. The study and practice of communication in the virtual construction project environment from an integrated (rather than discrete) perspective (of the space and place) therefore seems to be a useful route.

REFERENCES

Andresen, J. L., Christensen, K., & Howard, R. W. (2003). Project management with a project web. *ITcon, 8,* 29-41.

Barima, O. K. B. (2003). An exploratory study of the usage of virtual project management in South Africa. *Proceedings of CIB TG 23 international conference*: *Professionalism in construction-culture of high performance* [CD ROM]. Hong Kong: International Council for Research and Innovation in Building and Construction.

Bryant, J. (2004). Critical communication challenges for the new century. *Journal of Communication, 54*(3), 389-401.

Caneparo, L. (2001). Shared virtual reality for design and management: The Porta Susa project. *Automation in Construction, 10*(2), 217-227.

Clayton, M. J., Warden, R. B., & Parker, T. W. (2002). Virtual construction of architecture using 3D CAD and simulation. *Automation in Construction, 11*(2), 227-234.

Craig, R. T. (1999). Communication theory as a field. *Communication Theory, 9*(2), 199-161.

Griffin, E. (2003). *A first look at communication theory.* New York: McGraw-Hill.

Jaggar, D., Ross, A., Love, P. E. D., & Smith, J. (2001). Overcoming information opacity in construction: A commentary. *Logistics Information Management, 14*(5/6), 413-420.

Johanson, M., & Törlind, P. (2004). Mobility support for distributed collaborative teamwork. Special Issue: Mobile Computing in Construction. *ITcon, 9,* 355-366.

Kuladinithi, K., Timm-Giel, A., & Görg, C. (2004). Mobile ad-hoc communications in AEC industry. Special Issue: Mobile Computing in Construction. *ITcon, 9,* 313-323.

Olofsson, T., & Emborg, M. (2004). Feasibility study of field force automation in the Swedish construction sector. Special Issue:Mobile Computing in Construction. *ITcon, 9,* 297-311.

Rayport, J. F., & Sviokla, J. J. (1999). Exploiting the virtual value chain. In D. Tapscott (Ed.), *Creating value in the network economy* (pp. 35-51). Boston: Harvard Business School.

Rebolj, D., & Menzel, K. (2004). Mobile Computing in Construction (Editorial). Special Issue: Mobile Computing in Construction. *ITcon, 9,* 281-283.

Savioja, L., Mantere, M., Olli, I., Äyräväinen, S., Gröhn, M., & Iso-aho, J. (2003). Utilizing virtual environments in construction projects. Special Issue: Virtual Reality Technology in Architecture and Construction. *ITcon, 8,* 85-99.

Sulankivi, K. (2004). Benefits of centralized digital information management in multipartner projects. *ITcon, 9,* 35-63.

Ward, M., Thorpe, T., Price, A., & Wren, C. (2004). Implementation and control of wireless data collection on construction sites. Special Issue: Mobile Computing in Construction. *ITcon, 9,* 297-311.

KEY TERMS

3-D Graphics/Modeling: 3-D graphics/modeling refers to three dimensional perspectives of communicated graphics.

Communication Paths: Communication paths refer to the tracks of potential communications which may occur between communication objects (humans or computers).

Construction Place: The physical geographic locations where the construction activities take place.

Construction Space: The "imaginary" space (via ICT terminals/tools) where construction activities are executed as imitations of the real world scenarios by means of computers.

Fixed Construction Place: Fixed construction place refers to the bounded construction locations where construction activities take place.

Intra-Boundary Communication: Intra-boundary communication refers to communication within identified systemic boundaries.

Inter-Boundary Communication: Inter-boundary communication refers to communication across identified systemic boundaries.

Mobile Construction Place: Mobile construction place relates to the dynamic construction locations within the fixed locations where construction works are done.

A Comparison of Data Modeling in UML and ORM

Terry Halpin
Neumont University, USA

INTRODUCTION

The *Unified Modeling Language* (UML) was adopted by the Object Management Group (OMG) in 1997 as a language for object-oriented (OO) analysis and design. After several minor revisions, a major overhaul resulted in UML version 2.0 (OMG, 2003), and the language is still being refined. Although suitable for object-oriented code design, UML is less suitable for information analysis, since its graphical language provides only weak support for the kinds of business rules found in data-intensive applications, and its textual Object Constraint Language (OCL) is too technical for most business people to understand. Moreover, UML's graphical language does not lend itself readily to verbalization and multiple instantiation for validating data models with domain experts.

These problems can be remedied by using a *fact-oriented* approach for information analysis, where communication takes place in simple sentences, each sentence type can easily be populated with multiple instances, and attributes are avoided in the base model. At design time, a fact-oriented model can be used to derive a UML class model or a logical database model. *Object Role Modeling* (ORM), the main exemplar of the fact-oriented approach, originated in Europe in the mid-1970s (Falkenberg, 1976), and has been extensively revised and extended since, along with commercial tool support (e.g., Halpin, Evans, Hallock, & MacLean, 2003). Recently, a major upgrade to the methodology resulted in ORM 2, a second-generation ORM (Halpin 2005). Neumont ORM Architect (NORMA), an open source tool accessible online at *http://sourceforge.net/projects/orm*, is under development to provide deep support for ORM 2 (Curland & Halpin, 2007).

This article provides a concise comparison of the data modeling features within UML and ORM. The next section provides background on both approaches. The following section summarizes the main structural differences between the two approaches, and outlines some benefits of ORM's fact-oriented approach. A simple example is then used to highlight the need to supplement UML's class modeling notation with additional constraints, especially those underpinning natural identification schemes. Future trends are then briefly outlined, and the conclusion motivates the use of both approaches in concert to provide a richer data modeling experience, and provides references for further reading.

BACKGROUND

Detailed treatments of early UML use are provided in several articles by Booch, Rumbaugh, and Jacobson (Booch et al., 1999; Jacobson et al., 1999; Rumbaugh et al., 1999). The latest specifications for UML 2 may be accessed at *www.uml.org/*. The UML notation includes hundreds of symbols, from which various diagrams may be constructed to model different perspectives of an application. Structural perspectives may be modeled with class, object, component, deployment, package, and composite structure diagrams. Behavioral perspectives may be modeled with use case, state machine, activity, sequence, collaboration, interaction overview, and timing diagrams. This article focuses on data modeling, so considers only the static structure (class and object) diagrams. UML diagrams may be supplemented by textual constraints expressed in the Object Constraint Language (OCL). For detailed coverage of OCL 2.0, see Warmer and Kleppe (2003).

ORM pictures the world simply in terms of objects (entities or values) that play roles (parts in relationships). For example, you are now playing the role of reading, and this article is playing the role of being read. Overviews of ORM may be found in Halpin (2006, 2007b) and a detailed treatment in Halpin and Morgan (2008). For advanced treatment of some specific ORM topics, see Bloesch and Halpin (1997), De Troyer and Meersman (1995), Halpin (2001, 2002, 2004a), Halpin and Bloesch (1999), and Hofstede, Proper, and van der Weide (1993).

DATA STRUCTURES

Table 1 summarizes the correspondences between the main, high-level data constructs in ORM and UML. An uncommented "—" indicates no predefined support for the corresponding concept, and "†" indicates incomplete support. This comparison indicates that ORM's built-in symbols provide greater expressive power for capturing conceptual constraints in graphical data models.

Classes and data types in UML correspond to *object types* in ORM. ORM classifies objects into *entities* (UML objects) and *values* (UML data values—constants such as character strings or numbers). A *fact type* (relationship type)

Table 1. Comparison of the main data constructs in ORM and UML

ORM	*UML*
Data structures:	**Data structures**:
object type: entity type;	object class
value type	data type
— { use fact type }	attribute
unary fact type	— { use Boolean attribute }
2^+-ary fact type	2^+-ary association
objectified association (nesting)	association class
co-reference	qualified association †
Predefined Alethic Constraints:	**Predefined Constraints**:
internal uniqueness	multiplicity of ..1 †
external uniqueness	— { use qualified association } †
simple mandatory role	multiplicity of 1^+.. †
disjunctive mandatory role	—
frequency: internal; external	multiplicity †; —
value	enumeration, and textual
subset and equality	subset †
exclusion	xor †
subtype link and definition	subclass, discriminator, etc. †
ring constraints	—
join constraints	—
object cardinality	class multiplicity
— { use uniqueness and ring } †	aggregation/composition
—	initial value, changeability
Deontic Rules	—
User-Defined Textual Constraints	**User-Defined Textual Constraints**

† = incomplete coverage of corresponding concept

in ORM is called an *association* in UML (e.g., Employee works for Company). The main structural difference between ORM and UML is that ORM avoids *attributes* in its base models. Implicitly, attributes may be associated with roles in a relationship. For example, Employee.birthdate is modeled in ORM as the second role of the fact type: Employee was born on Date.

The main advantages of attribute-free models are that all facts and rules can be naturally verbalized as sentences, all data structures can be easily populated with multiple instances, models and queries are more stable since they are immune to changes that reshape attributes as associations (e.g., if we later wish to record the historical origin of a family name, a family name attribute needs to be remodeled using a relationship), nulls are avoided, connectedness via semantic domains is clarified, and the metamodel is simplified. The price paid is that attribute-free diagrams usually consume more space. This disadvantage can be offset by deriving an attribute-based view (e.g., a UML class model or a relational database schema) when desired (tools can automate this).

ORM allows relationships of any *arity* (number of roles). A relationship may have many readings starting at any role, to naturally verbalize constraints and navigation paths in any direction. Fact type readings use *mixfix* notation to allow object terms at any position in the sentence, allowing natural verbalization in any language. Role names are also allowed. ORM includes procedures for creating, verbalizing, and transforming models. The first step in creating a data

model is to verbalize relevant information examples—these "*data use cases*" are in the spirit of UML use cases, except the focus is on the underlying data.

In an ORM diagram, object types appear as named, soft rectangles, and roles appear as boxes connected by a line to their object type. A predicate appears as an ordered set of role boxes together with a predicate reading. Since role boxes are set out in a line, fact types may be conveniently populated with tables holding multiple fact instances, one column for each role. This allows all fact types and constraints to be *validated by verbalization as well as sample populations.*

While supporting binary and longer associations, UML uses Boolean attributes instead of *unary* relationships. For example, the fact instance expressed in ORM as "Person 'Sam Spade' smokes" would typically be rendered awkwardly in UML as "SamSpade: Person.isSmoker = true." To be business friendly, UML should support unary fact types directly (e.g., Room has a window, Person smokes, etc.).

In UML, each association has at most one name, and verbalization into sentences is practical only for infix binaries. Since roles for ternaries and higher arity associations are not on the same line, directional verbalization and multiple instantiation for population checks are ruled out. UML does provide object diagrams for instantiation, but these are convenient only for populating with a few instances.

Both UML and ORM allow associations to be objectified as first-class object types, called *association classes* in UML and *objectified* (or *nested) associations* in ORM. UML requires the same name to be used for the association and the association class, impeding natural verbalization, in contrast to ORM nesting based on linguistic nominalization (a verb phrase is objectified by a noun phrase).

CONSTRAINTS AND IDENTIFICATION SCHEMES

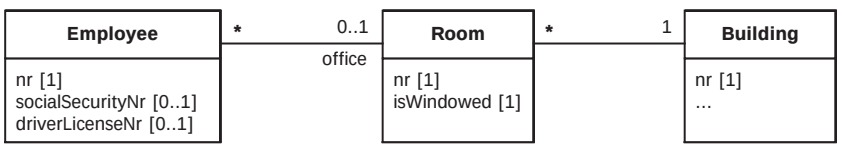
Business people communicate about things using value-based identification schemes, not memory addresses or hidden object identifiers, and hence these need to be supported in any conceptual model. This entails rich support for what ORM calls internal and external uniqueness constraints, which may be used as a basis for identification of entities. UML 2.0 includes only a restricted form of internal uniqueness (maximum multiplicity of 1) and an even weaker form of external uniqueness (via qualified associations). Moreover, UML 2.0 does not require any value-based identification scheme for a class.

For example, consider the UML class model in Figure 1. Classes are depicted as rectangles, with the class name in the top compartment, and attribute names in a lower compartment. Binary associations are depicted as lines connecting the classes involved. *Multiplicity constraints* may be applied to attributes and roles (association ends), using "1" for "exactly one," "0..1" for "at most one," and "*" for "zero or more.

Here, the attribute *multiplicity constraints* tell us that each employee has exactly one employee number and at most one social security number and at most one driver's license number, each room has exactly one room number and exactly one value for isWindowed (a Boolean attribute to indicate whether a room has a window), and each building has exactly one building number. The multiplicity constraints on the associations tell us that each employee has at most one office, each room is the office of zero or more employees, each room is in exactly one building, and each building houses zero or more rooms. What semantics are missing (allowing that appropriate association names can be added)?

To begin with, there is no declaration of how entities are identified by business people. In the business domain, employees are identified by their employee numbers, and rooms are identified by combining their local room number with the number of the building in which they are housed. The lack of the ability to declare these identification schemes prevents an understanding of these business rules, and excludes any formal way to populate the data structures with fact instances. Moreover, the uniqueness constraint that each social security number applies to at most one employee is lost, because UML cannot graphically declare the 1:1 nature of an association that is modeled as an attribute. Finally, in this business domain each employee is required to have

Figure 1. What semantics are missing?



either a social security number or a driver's license number or both, but this is not captured in the diagram.

One partial solution to this problem would be to add additional constructs in UML to apply uniqueness constraints to arbitrary combinations of attributes or binary association roles that may be considered properties of the same class. Perhaps some future version of UML might do this.

For comparison, an ORM schema for the same domain is shown in Figure 2. Entity types are depicted as named, soft rectangles. Value types appear as named, dashed, soft rectangles. Fact types are depicted as role sequences with at least one reading. Each role is depicted as a box connected to the object type whose instances play that role.

Here we have one unary fact type: Room has a window. The other fact types are binary. For example, "Employee has office in Room" and "Room is office of Employee" are forward and inverse readings of the same fact type. Role names may be added (e.g., office). Employees are identified by their employee number, and buildings by their building number. Simple identification schemes like these may be abbreviated in parentheses as shown. The bars over roles depict internal uniqueness constraints (e.g., **Each** Room is in **at most one** Building) and the solid dots are mandatory role constraints (e.g., **Each** Room is in **at least one** Building). The circled dot is an inclusive-or constraint (**Each** Employee has **some** SocialSecurityNr **or** has **some** DriverLicenseNr).

The external uniqueness constraint (circled bar) indicates that each room number and building combination applies to at most one room; the double bar indicates this provides the preferred identification scheme for rooms. If we instead introduced a simple identifier for Room, the external uniqueness constraint would be depicted as a single bar.

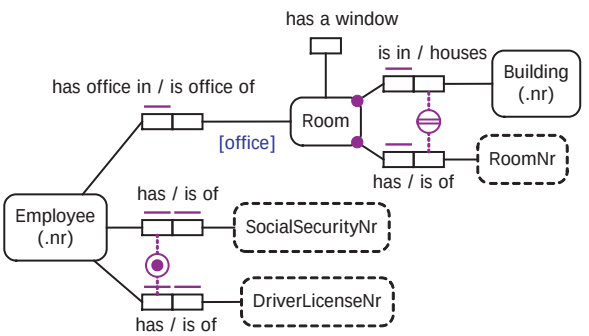
Given an *n*-ary association ($n > 2$), UML's multiplicity notation cannot express a mandatory role constraint on any association that has between 1 and $n-2$ mandatory roles. This is because multiplicity on one role is defined in terms of the other n–1 roles. This is fine for binary associations, but not for ternaries and beyond. For practical examples of such constraint patterns, see Halpin (2001).

ORM includes many other graphical constraint primitives that go far beyond those found in UML class diagrams, for example set-comparison (subset, exclusion, equality) constraints over compatible role sequences, ring constraints (asymmetric, intransitive, acyclic, etc.), and join constraints (Halpin & Morgan, 2008). In addition to these alethic constraints (which hold necessarily for each state of the business domain), ORM supports deontic rules (which ought to be obeyed but may be violated). This makes ORM far richer for capturing business rules, many of which are deontic in nature (Halpin, 2007a). Moreover, ORM's constraint primitives are far more orthogonal than those of UML.

UML and ORM both permit users to add other constraints and derivation rules in a textual language of their choice. UML suggests OCL for this purpose (Warmer & Kleppe, 2003). Although OCL is an unambiguous language, its mathematical syntax renders it unsuitable for validating rules with non-technical domain experts. For example, the inclusive-or constraint displayed earlier as a circled dot on the ORM diagram is expressed in OCL as follows in the context of the Employee class: {self.socialSecurityNr -> notEmpty() or self.driverLicenseNr -> notEmpty() }. Compare this with the automated ORM verbalization given earlier.

ORM's conceptual query language, ConQuer, is both formal and readily understandable to non-technical users, and its attribute-free nature makes it much more *semantically stable* than an attribute-based language such as OCL (Bloesch & Halpin, 1997; Halpin & Bloesch, 1999). Currently a number of Conquer-like languages are being developed to provide a single textual language for ORM that can be used for both models and queries. Recently an ORM foundation (*www.ormfoundation.org*) has been set up as a non-profit organization to promote the fact-oriented approach.

Figure 2. The ORM model captures the semantics lost in the UML model



FUTURE TRENDS

The OMG recommends its Model Driven Architecture framework as a way to facilitate the generation of software artifacts from high-level models, starting with a Computation Independent Model (CIM), and then moving down to a Platform Independent Model (PIM), and finally the Platform Specific Model (PSM) used in the actual implementation. For reasons given earlier, although useful for code design, UML currently has some shortcomings with regard to conceptual data analysis and the specification of business rules to be validated by business domain experts.

With a view to providing better support at the CIM level, the OMG recently adopted the Semantics of Business Vocabulary and Business Rules (SBVR) specification (OMG, 2007).

Like ORM, the SBVR approach is fact oriented instead of attribute based, and includes deontic as well as alethic rules. Many companies are now looking to model-driven development as a way to dramatically increase the productivity, reliability, and adaptability of software engineering approaches. It seems clear that both object-oriented and fact-oriented approaches will be increasingly utilized in the future to raise the percentage of quality code in software applications that can be generated from higher-level models.

CONCLUSION

UML class diagrams are often more compact than ORM models, and they can be adorned with implementation detail for engineering to and from object-oriented programming code. Moreover, UML includes mechanisms for modeling behavior, and its adoption by the OMG is helping it gain wide support in industry, especially for the design of object-oriented software.

ORM is based on a small set of easily mastered, orthogonal concepts, its attribute-free nature facilitates model validation by verbalization and population and conveys semantic stability, and its graphical constraint language can formally capture many business rules. UML modelers willing to learn ORM can get the best of both approaches by using ORM as a front end to their information analysis and then mapping their ORM models to UML, where ORM constraints with no UML counterpart can be captured in notes or formal textual constraints. This option will become more attractive once more software tools provide automatic transformation between ORM and UML.

REFERENCES

Bloesch, A., & Halpin, T. (1997). Conceptual queries using ConQuer-II. In D. Embley & R. Goldstein (Eds.), *Proceedings of the 16th International Conference on Conceptual Modeling* (ER'97) (pp. 113-126). Berlin: Springer-Verlag.

Booch, G., Rumbaugh, J., & Jacobson, I. (1999). *The Unified Modeling Language user guide.* Reading: Addison-Wesley.

Curland, M., & Halpin, T. (2007). Model driven development with NORMA. *Proceedings of HICSS-40.*

De Troyer, O., & Meersman, R. (1995, December). A logic framework for a semantics of object oriented data modeling. *Proceedings of OOER'95: Object-Oriented and Entity-Relationship Modeling* (pp. 238-249). Berlin: Springer-Verlag (LNCS 1021).

Falkenberg, E. (1976). Concepts for modelling information. In G. Nijssen (Ed.), *Modelling in data base management systems* (pp. 95-109). Amsterdam: North-Holland.

Halpin, T. (2001). Supplementing UML with concepts from ORM. In K. Siau & T. Halpin (Eds.), *Unified Modeling Language: Systems analysis, design and development issues*. Hershey, PA: Idea Group.

Halpin, T. (2002). Information analysis in UML and ORM: A comparison. In K. Siau (Ed.), *Advanced topics in database research* (vol. 1, pp. 307-323). Hershey PA: Idea Group.

Halpin, T. (2004a). Constraints on conceptual join paths. In T. Krogstie, T.A. Halpin, & K. Siau (Eds.), *Information modeling methods and methodologies.* Hershey PA: Idea Group.

Halpin, T. (2004b). Comparing metamodels for ER, ORM and UML data models. In K. Siau (Ed.), *Advanced topics in database research* (vol. 3, pp. 23-44). Hershey, PA: Idea Group.

Halpin, T. (2005). ORM 2. In R. Meersman et al. (Eds.), *On the move to meaningful Internet systems 2005: OTM 2005 workshops* (pp. 676-687). Berlin: Springer-Verlag (LNCS 3762).

Halpin, T. (2006). Object-role modeling (ORM/NIAM). In *Handbook on architectures of information systems* (2nd ed., pp. 81-103). Heidelberg: Springer-Verlag.

Halpin, T. (2007a). Modality of business rules. In K. Siau (Ed.), *Research issues in systems analysis and design, databases and software development* (pp. 206-226). Hershey, PA: IGI.

Halpin, T. (2007b). Fact-oriented modeling: Past, present and future. In J. Krogstie, A. Opdahl, & S. Brinkkemper (Eds.), *Conceptual modelling in information systems engineering* (pp. 19-38). Berlin: Springer-Verlag.

Halpin, T., & Bloesch, A. (1999). Data modeling in UML and ORM: A comparison. *Journal of Database Management, 10*(4), 4-13.

Halpin, T., & Morgan T. (2008). *Informational modeling and relational databases 2nd edition*. San Fransisco: Morgan Kauffman.

Halpin, T., Evans, K., Hallock, P., & MacLean, W. (2003). *Database modeling with Microsoft® Visio for enterprise architects.* San Francisco: Morgan Kaufmann.

Hofstede, A., Proper, H., & van der Weide, T. (1993). Formal definition of a conceptual language for the description and manipulation of information models. *Information Systems, 18*(7), 489-523.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The unified software development process.* Reading, MA: Addison-Wesley.

OMG. (2003). *OMG Unified Modeling Language specification version 2.0.* Retrieved from http://www.uml.org/

OMG. (2007). *Semantics of business vocabulary and business rules (SBVR).* Retrieved from http://www.omg.org/cgi-bin/doc?dtc/2006-08-05

Rumbaugh, J., Jacobson, I., & Booch, G. (1999). *The Unified Modeling Language reference manual.* Reading, MA: Addison-Wesley.

Warmer, J., & Kleppe, A. (2003). *The Object Constraint Language: Getting your models ready for MDA* (2nd ed.). Reading, MA: Addison-Wesley.

KEY TERMS

Arity: The number of roles in a fact type (unary = 1, binary = 2, ternary = 3, etc.). In ORM, fact types may be of arity 1 or more. In UML fact types (associations) may be of arity 2 or more.

Business Rule: A constraint or derivation rule that applies to the business domain. An alethic/deontic static constraint restricts the possible/permitted states of the business, and a dynamic constraint restricts the possible/permitted transitions between states. A derivation rule declares how a fact may be derived from existing facts, or how an object is defined in terms of existing objects.

Conceptual Schema: Specification of the structure of a business domain using language and terms easily understood by a non-technical, domain expert. A conceptual schema typically declares the fact types and business rules that are relevant to the business domain.

Elementary Fact Type: In ORM, an elementary fact is an atomic proposition that applies a logical predicate to a sequence of one or more objects of a given type; it cannot be split into smaller facts without information loss. An elementary fact type is a kind of elementary fact. For example: Person smokes; Person was born in Country; Person introduced Person to Person. In UML, a non-unary elementary fact type is an elementary association.

Entity Type: An entity is a non-lexical object that in the real world is identified using a definite description that relates it to other things (e.g., the Country that has CountryCode 'US'). Typically, an entity may undergo changes over time. An entity type is a kind of entity, for example, Person, Country. In UML, an entity is called an object, and an entity type is called a class.

Object Type: In ORM, an object is either an entity (non-lexical thing) or a value (lexical constant, such as a character string), and an object type is a kind of object (e.g., Person, CountryCode). In UML, the term "object" is restricted to entities (instances of classes), while the term "data value" is used for instances of data types.

Object-Role Modeling (ORM): A fact-oriented approach for modeling information at a conceptual level, using language that is easily understood by non-technical, domain experts. ORM includes rich graphical and textual languages for modeling facts and business rules, and provides procedures for creating conceptual models and transforming them to lower-level models for implementation.

Role: In ORM, a role is a part played in a fact type (relationship type). In UML, this is known as an association-end. For example, in the fact type Person works for Company, Person plays the role of employee, and Company plays the role of employer.

Unified Modeling Language (UML): Language adopted by the Object Management Group as a modeling language for object-oriented analysis and design of software systems. UML includes several sublanguages and diagram notations for modeling different aspects of software systems.

Completeness Concerns in Requirements Engineering

Jorge H. Doorn
INTIA, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina & Universidad Nacional de La Matanza, Argentina

Marcela Ridao
INTIA, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina

INTRODUCTION

The difficulties that software developers must face to understand and elicit clients' and users' necessities are widely known. The more complex the context of the problem, the more difficult the elicitation of software requirements becomes. Many times, requirements engineers must become themselves problem-domain experts during the acquisition of knowledge about the context of the application. The requirements engineering (RE) objective is to systematize the process of requirements definition (Maculay, 1993; Maté & Silva, 2005; Reubenstein & Waters, 1991) along with creating a compromise among clients and users with developers since they must both participate and collaborate together. The requirements engineering process consists of three main activities: elicitation, modeling, and analysis of the application domain (Kotonya & Sommerville, 1998; Sommerville & Sawyer, 1997). Later, requirements management deals with the changes in the requirements and the irruption of new ones.

RE provides methods, techniques, and tools to help requirements engineers elicit and specify requirements, ensuring their highest quality and completeness. However, the problem of completeness is a certain menace to requirements quality and casts a serious doubt on the whole RE process. Completeness is an unreachable goal, and to estimate the degree of completeness obtained at a certain step in the project is even more difficult. The requirements engineer faces a universe of discourse (UofD) that he or she will hardly ever fully know. This situation is not unique during the whole software development process since something similar happens while testing.

The use of statistical models to predict the number of defects in a software artifact was successfully introduced some time ago (Petersson, Thelin, Runeson, & Wohlin, 2003). In this article, the use of capture and recapture information is applied in the RE field in order to make an estimation of the number of undiscovered requirements after a requirements elicitation process.

The following section analyses the validation problem in RE. Then, a section describing the use of LEL (language-extended lexicon) and scenarios in requirements engineering is included. After that, the problem of estimating closed populations is studied. Later, the use of capture and recapture in the RE domain is introduced, and finally, some future work and conclusions are presented.

BACKGROUND

The completeness problem in software engineering and requirements engineering is very similar to others in many knowledge areas. Otis, Burnham, White, and Anderson (1978) introduced a method to estimate the size of a closed population of wild animals based on the data gathered during repetitive capture of specimens. This method has been extended to the area of software inspections by several authors (Biffl, 2003; Briand, El Emam, Freimut, & Laitenberger, 2000; Thelin, 2004; Petersson, Thelin, Runeson & Wohlin, 2003; Wohlin & Runeson, 1998).

Requirements validation has become a complex task mainly due to the kind of representation models used, which requires clients and users with special skills to understand them. As it is pointed out by several authors (Cysneiros & Yu, 2003; Sommerville & Sawyer, 1997), the requirements validation seldom discovers all defects, which may reach later stages in the software development process. It has been proven that the use of natural-language representation for requirements helps validation, especially when requirements are expressed using the client user's vocabulary (Leite & Franco, 1990). To be able to provide such representation, the requirements engineer should acquire the clients' and users' vocabulary. However, ambiguity is the main drawback of the natural-language approach (Berry & Kamsties, 2004; Jackson, 1995; Sommerville & Sawyer, 1997). The construction of a glossary of clients' and users' jargon helps reduce ambiguity and build requirements specification in an understandable language. Several experiences have

shown that a glossary of the clients' and users' vocabulary is, in itself, a source of information to elicit valuable UofD information (Ben Achour, Rolland, Maiden, & Souveyet, 1999; Oberg, Probasco, & Ericsson, 1998; Regnell, 1999; Rolland & Ben Achour, 1998; Weidenhaupt, Pohl, Jarke, & Haumer, 1998).

In this entry, an RE process that begins with the construction of an LEL as its first activity (Leite, Hadad, Doorn & Kaplan, 2000) is addressed in order to analyse the impact of completeness in it. In this process, the LEL construction is followed first by the building of scenarios to understand and model the current UofD, and later by the building of another set of scenarios to figure out how the future UofD could be and to model it. Finally, this process ends with the set of requirements of the software system to be developed.

THE PROCESS

The backbone of the process is to anchor every model in the UofD vocabulary. Knowledge acquired by means of observations, document reading, interviews, and other techniques is first modeled using LEL and later by means of scenarios (Leite et al., 2000). LEL and scenarios are verified for internal consistency and validated with the collaboration of clients and users. During verification and validation (V&V), completeness is a key issue since several steps during LEL and scenario inspections (Leite, Doorn, Hadad & Kaplan, 2005) and also some guidelines for validation are designed with completeness as their main target. However, this is far from being enough.

Language Extended Lexicon

Most relevant or peculiar words or phrases (named LEL symbols) of the UofD are included in the LEL. Every symbol is identified by its name (including synonyms) and by two descriptions: notion and behavioral response. The notion contains sentences defining the symbol and the behavioral response reflects how it influences the UofD. Figure 1 depicts the model used to represent LEL symbols.

Scenarios

Scenarios are used to understand the UofD first, and later to understand the problem and its functionality. Each scenario describes a specific situation of the UofD focusing on its behavior. The scenario model (see Figure 2) contains the following components: title, goal, context, resources, actors, episodes, and exceptions.

A scenario must satisfy a goal, which is reached by performing its episodes. Episodes represent the main course of action, but they may include variations or possible alternatives. Actors carry out episodes using resources. While performing episodes, an exception may arise, signaling an obstacle to goal achievement. The context is described detailing a geographical location, a temporal location, or preconditions. Context, resources, and episodes may have constraints, which are used to record restrictions of any kind. Constraints are used to characterize nonfunctional requirements.

Figure 1. Language-extended lexicon model

LEL: LEL is the representation of the symbols in the language of the application domain.
Syntax: $\{Symbol\}_1^N$

Symbol: It is an entry of the lexicon that has a special meaning in the application domain.
Syntax: $\{Name\}_1^N + \{Notion\}_1^N + \{Behavioral\ Response\}_1^N$

Name: It is the identification of the symbol. Having more than one name represents synonyms.
Syntax: Word | Phrase

Notion: It is the denotation of the symbol.
Syntax: Sentence

Behavioral Response: It is the connotation of the symbol.
Syntax: Sentence

Figure 2. Scenario model

Scenario: It is the description of a situation in the application domain.
Syntax: Title + Goal + Context + $\{Resources\}_1^N$ + $\{Actors\}_1^N$ + $\{Episodes\}_2^N$ + {Exceptions}

Title: It is the identification of the scenario.
Syntax: Phrase | ([Actor | Resource] + Verb + Predicate)

Goal: This is the aim to be reached in the application domain. The scenario describes the achievement of the goal.
Syntax: [Actor | Resource] + Verb + Predicate

Context: This is the geographical location, temporal location, or precondition.
Syntax: {Geographical Location} + {Temporal Location} + {Precondition}

Resources: They are the relevant physical elements or information that must be available in the scenario.
Syntax: Name + {Constraint}

Actors: These are persons, devices, or organization structures that have a certain role in the scenario.
Syntax: Name

Episode: It is a set of actions that details the scenario and provides its behavior.
Syntax:
<episodes> ::= <group> <group> | <episodes> <group>
<group> ::= <sequential group> | < nonsequential group>
<sequential group> ::= <basic sentence> | <sequential group> <basic sentence>
<nonsequential group> ::= **#** <sequential group> **#**
<basic sentence> ::= <simple sentence> | <conditional sentence> | <optional sentence>
<simple sentence> ::= <episode sentence>
<conditional sentence> ::= **IF** <condition> **THEN** <episode sentence>
<optional sentence> ::= **[** <episode sentence> **]**

Exception: An exception hinders the achievement of the scenario goal.
Syntax: Cause [**(**Solution**)**]

Constraint: It is a scope or quality requirement referring to a given entity. It is an attribute of resources, basic episodes, or subcomponents of context.
Syntax: ([Subject | Actor | Resource] + **Must** [**Not**] + Verb + Predicate) | Phrase

CLOSED-POPULATIONS ESTIMATION

The estimation method in closed populations of any nature such as wild animals, software bugs, and in this case LEL symbols is based on the idea of capturing specimens belonging to the population whose size is unknown more than one time. Initially, this method was applied to wild-animal populations (Otis, Burnham, White & Anderson, 1978). The strategy, sometimes called capture, mark, and recapture, or the tag recapture method, involves capturing and releasing specimens. If the same specimens are mostly captured over and over again, it can be concluded that the population is basically the captured set. On the other hand, when a few of the specimens are recaptured, it can be concluded that the population is actually very large. The impact of the capture-release procedure over the specimens should be as smooth and bloodless as possible.

Initial capture and recapture statistical models assumed that every single wild animal had the same probability to become trapped and that every trap or capture procedure had the same probability to capture specimens. Later, several models considering differences among wild animals and

among other qualitative factors were introduced. A basic model assuming no difference among animals or any other factor is currently identified as M_o. Models M_t, M_b, and M_h introduce the assumption of different sources of unequal probabilities. Model M_t allows capture probabilities to vary by time due to climate, trap location, or even capture procedure. Model M_b allows capture probabilities to vary by behavioral changes caused by previous captures. Finally, M_h allows capture probabilities to vary due to heterogeneity among individual animals.

Models combining more than one source of capture probability variation were also developed such as M_{tb}, M_{th}, M_{bh}, and M_{tbh}. For every one of these models, the independence among qualitative factors is assumed. In other words, the combined probability is determined by the product rule of Bayes' theorem (Papoulis, 1985).

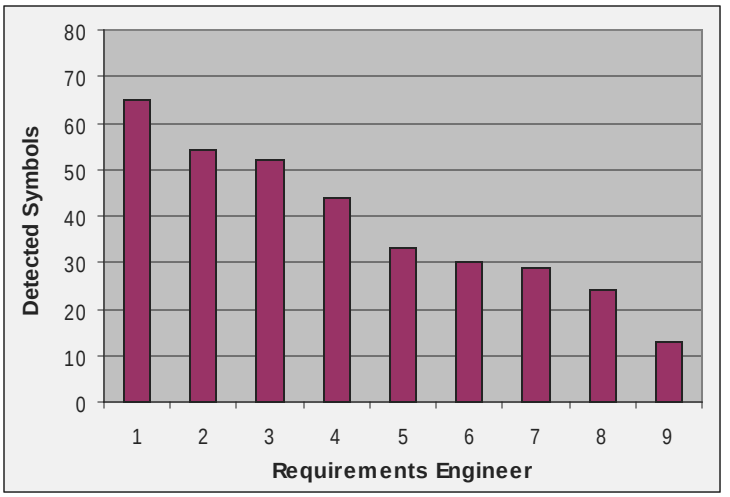
Researchers who applied capture and recapture methods to software inspections (Biffl, 2003; Briand, El Emam, Freimut & Laitenberger, 2000; Petersson, Thelin, Runeson & Wohlin, 2003; Thelin, 2004; Wohlin & Runeson, 1998) mapped the role of wild animals to software bugs and the role of traps or hunters to software inspectors. It became evident that the variation of capture probabilities due to bug behavioral changes caused by more than one inspection is senseless. Then models M_b, M_{bh}, M_{tb}, and M_{tbh} should be discarded using M_o, M_h, M_t, and M_{th} instead. This became so obvious that the authors formerly mentioned (Biffl, 2003; Briand, El Emam, Freimut & Laitenberger, 1997) did not even mention the factor behavior probability change.

CAPTURING REQUIREMENTS

When capture and recapture techniques are applied to the RE domain, the behavior probability-change factor should be carefully watched since it is not clear if it should or should not be disregarded. Actually, this depends upon the source of information used during knowledge elicitation and the acquisition technique applied. For example, if more than one requirements engineer reads the same document to pick LEL symbols, there is no way for such symbols to be influenced by previous readings. On the other hand, if several requirements engineers interview the same client or user during the LEL creation activity, it is conceivable that the interviewed person (who actually holds in his or her way of speaking the information about LEL symbols) may become influenced by previous interviews. This implies that the person being interviewed tends to change after the initial interview. He or she may develop a discomfort toward being interviewed over and over again about the same subject, or he or she may get more involved in the software project, supplying more and more useful information in consecutive interviews.

Preliminary experimental studies showed that the time factor (actually the difference among requirements engineers) behaves in a way similar to what happens in software inspections (Doorn & Ridao, 2003; Ridao & Doorn, 2006). In other words, in both cases, the time factor is more important than in what happened with wild animals. However, a notoriously more important fact became evident: Requirements engineers' probabilities and symbols' probabilities are far from being independent factors. This was not reported during capture and recapture experiences in software inspections.

Figure 3. Number of LEL symbols elicited by different requirements engineers



In other words, an M_{th} model cannot be applied using factor independence assumption. This phenomenon can be easily understood by paying attention to the fact that different people have different biases in their skills. Then it seems there is no such thing as the probability of a symbol to be picked by a requirements engineer and also there is no such thing as the probability of a requirements engineer to pick symbols. What there actually exists is the probability of a given requirements engineer to pick a specific symbol.

Figure 3 depicts the number of LEL symbols obtained by different requirements engineers in a replicated experiment based on a system dealing with savings plans for the acquisition of brand-new cars (Mauco, Ridao, del Fresno, Rivero, & Doorn, 1997; Rivero, Doorn, del Fresno, Mauco, Ridao, & Leonardi, 1998). It works through groups of people who pay monthly fees in order to obtain automobiles. They participate in monthly meetings in which a unit is adjudicated by a drawing or bidding by the participants. The system also includes arbitration in the event of renouncement or a participant's death, lack of payment of the monthly fees, insurance, and contracts with the makers, among others. It is obvious that the time factor must be carefully taken into account.

FUTURE TRENDS

The preliminary experiments should be extended to more study cases not only in LEL but in other models of the process. Metrics to define the granularity levels of the completeness studies should be defined and tested. It is not known whether dealing with the capture of notions or behavioral responses will improve or make the prediction ability of the models difficult. It is also unknown how the completeness of a model will influence the completeness of later models and, more important, over the completeness of the software system requirements.

CONCLUSION

Software requirements validation is a hard task not for the requirements actually understood and modeled, but for those that remain hidden. They will show up in the middle of the software development process with a notorious disturbance power, or they will be discovered when the software is put into service. Its functionality will not be as expected by clients and users.

Since some requirements will remain hidden because the requirements engineer has not seen them, there are few chances in which clients or users discover such missing software specification.

A capture and recapture strategy may help in reducing the amount of hidden requirements by giving the people involved a figure about how many requirements remain unmodeled and perhaps help in the development of better heuristics for the whole requirements engineering process.

REFERENCES

Ben Achour, C., Rolland, C., Maiden, N. A. M., & Souveyet, C. (1999). Guiding use case authoring: Results of an empirical study. In *Proceedings of the International Symposium on Requirements Engineering* (pp. 36-43). Limerick, Ireland: IEEE Computer Society Press.

Berry, D. M., & Kamsties, E. (2004). Ambiguity in requirements specification. In J. S. C. P. Leite & J. H. Doorn (Eds.), *Perspectives on software requirements* (pp. 7-44). Kluwer Academic Press.

Biffl, S. (2003). Evaluating defect estimation models with major defects. *The Journal of Systems and Software, 65*(1), 13-29.

Briand, L., El Emam, K., Freimut, B., & Laitenberger, O. (1997). Quantitative evaluation of capture recapture models to control software inspections. In *Proceedings of the Eighth International Symposium on Software Reliability Engineering* (pp. 234-244).

Briand, L., El Emam, K., Freimut, B., & Laitenberger, O. (2000). A comprehensive evaluation of capture-recapture models for estimating software defects contents. *IEEE Transactions on Software Engineering, 26*(6), 518-540.

Cysneiros, L. M., & Yu, E. (2004). Non-functional requirements elicitation. In J. S. C. P. Leite & J. H. Doorn (Eds.), *Perspectives on software requirements* (pp. 115-138). Kluwer Academic Press.

Doorn, J., & Ridao, M. (2003). Completitud de glosarios: Un estudio experimental. In *Anais do WER'03: Workshop en Engenharia do Requisitos, Paracicaba-SP*, Brazil (pp. 317-328).

Jackson, M. (1995). *Software requirements & specifications: A lexicon of practice, principles and prejudices.* Addison Wesley & ACM Press.

Kotonya, G., & Sommerville, I. (1998). *Requirements engineering: Processes and techniques.* John Wiley & Sons.

Leite, J. C. S. P., & Franco, A. P. M. (1990). O uso de hipertexto na elicitaçao de linguagens da aplicaçao. In *Anais de IV Simpósio Brasilero de Engenharia de Software*, Brazil (pp. 134-149).

Leite, J. C. S. P., Doorn, J. H., Hadad, G. D. S., & Kaplan, G. N. (2000). A scenario construction process. *Requirements Engineering Journal, 5*(1), 38-61.

Leite, J. C. S. P., Hadad, G. D. S., Doorn, J. H., & Kaplan, G. N. (2005). Scenario inspections. *Requirements Engineering Journal, 10*(1), 1-21.

Maculay, L. (1993). Requirements capture as a cooperative activity. In *Proceedings of the IEEE International Symposium on Requirement Engineering*, San Diego, CA (pp. 174-181). IEEE Computer Society Press.

Maté, J. L., & Silva, A. (2005). *Requirements engineering for sociotechnical systems.* Information Science Publishing.

Mauco, V., Ridao, M., del Fresno, M., Rivero, L., & Doorn, J. H. (1997). *Ingeniería de requisitos, proyecto: Sistema de planes de ahorro* (Tech. Rep.). Tandil, Argentina: ISISTAN, UNCPBA.

Oberg, R., Probasco, L., & Ericsson, M. (1998). *Applying requirements management with use cases.* Rational Software Corporation.

Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture on closed animal populations. *Wildlife Monograph, 62*.

Papoulis, A. (1985). *Probability, random variables, and stochastic processes.* McGraw Hill.

Petersson, H., Thelin, T., Runeson, P., & Wohlin, C. (2003). Capture-recapture in software inspections after 10 years research: Theory, evaluation and application. *The Journal of Systems and Software, 72*, 249-264.

Regnell, B. (1999). *Requirements engineering with use cases: A basis for software development.* Unpublished doctoral dissertation, Department of Communication Systems, Lund University.

Reubenstein, H. B., & Waters, R. C. (1991). The requirements apprentice: Automated assistance for requirements acquisition. *IEEE Transactions on Software Engineering, 17*(3), 226-240.

Ridao, M., & Doorn, J. (2006). Estimación de completitud en modelos de requisitos basados en lenguaje natural. In *Anais do WER'06: Workshop en Engenharia do Requisitos,* Rio de Janeiro, Brazil (pp. 151-158).

Rivero, L., Doorn, J., del Fresno, M., Mauco, V., Ridao, M., & Leonardi, M. C. (1998). Una estrategia de análisis orientada a objetos basada en escenarios: Aplicación en un caso real. In *Anais do WER'98: Workshop en Engenharia do Requisitos*, Maringá, Brazil (pp. 79-90).

Rolland, C., & Ben Achour, C. (1998). Guiding the construction of textual use case specifications. *Data & Knowledge Engineering, 25*, 125-160.

Sommerville, I., & Sawyer, P. (1997). *Requirements engineering: A good practice guide.* John Wiley & Sons.

Thelin, T. (2004). Team-based fault content estimation in the software inspection process. In *Proceedings of ICSE'04: 26th International Conference on Software Engineering* (pp. 263-272).

Weidenhaupt, K., Pohl, K., Jarke, M., & Haumer, P. (1998). Scenarios in system development: Current practice. *IEEE Software, 15*(2), 34-45.

Wohlin, C., & Runeson, P. (1998). Defect content estimations from review data. In *Proceedings of the 20th International Conference on Software Engineering* (pp. 400-409).

KEY TERMS

Capture and Recapture: It is a method to estimate the abundance of members of a given class in a closed environment. Specimens are captured, marked, and released.

Elicitation: This is the activity of acquiring knowledge of a given kind during the requirements engineering process.

Language Extended Lexicon: It is a semiformal model holding the most relevant words or phrases of the language of the application domain carrying a special meaning.

Requirements Engineering: It is an area of software engineering that is responsible for acquiring and defining the capabilities of the software system.

Requirements Management: It is the activity of following up on the evolution of the software system requirements set, and dealing with the allocation of requirements to specific pieces of the software and with the changes in the requirements along time.

Scenario: This is a semiformal model describing observable situations in the current UofD or guessed situations that would take place when the software system is put into operation.

Universe of Discourse: It is the environment in which the software artifact will be used. It includes the macrosystem and any other source of knowledge.

Validation: It is the activity of contrasting a model with the actual world. It should answer the question "Are we creating the right model?"

Verification: It is the activity of checking different parts of a model or different models among them. It should answer the question "Are we creating the model right?"

Complex Organizations and Information Systems

Leoni Warne
Department of Defence, Australia

Helen Hasan
University of Wollongong, Australia

Henry Linger
Monash University, Australia

INTRODUCTION

In modern organizations, information, and particularly knowledge, is known to be the most strategically important resource. The defining characteristics of modern organizational forms are purported to be flatter hierarchies, decentralized decision making, greater capacity for tolerance of ambiguity, permeable boundaries, capacity for renewal, self-organizing units, continual change, and an increasingly complex environment (Daft & Lewin, 1993; Warne, Ali, Bopping, Hart, & Pascoe, 2004). Yet, many systems that are developed to support organizational activities continue to fail at an alarming rate (Hart & Warne, 2005; Warne, 2002). Many explanations have been offered for such failures (e.g., DeLone & McLean, 1992; Fortune & Peters, 2005; Lyytinen & Hirschheim, 1987; Sauer, 1993; Warne, 2002), but contradictions and stresses continue to confound organizations and their use of information and communications technology (ICT).

The challenge for information systems (IS) research and practice is to articulate an organizational paradigm, including its structures, forms, and systems, that will enable the organization to be agile, innovative, and have the capacity to learn. This article discusses some of the parameters for a new contemporary model for organizations.

BACKGROUND

A modern paradigm for organizations needs to focus on their ability to support knowledge work practices that integrate thinking and doing (Burstein & Linger, 2003). Such practices address both the production of goods and services and the means of their production. Most importantly, such practices rely on the ability to remember and learn from the past and to use this learning to make sense of current situations. It is these practices that enable organizations to effectively compete in a rapidly changing environment through their ability to respond flexibly to internal and external demands. Such flexibility is derived from the dynamic of a network-centric organizational form, the shift to knowledge as a critical resource, the emphasis on learning, and a recognition and acceptance of complexity as the modern context of organizations.

The 'sensible organization' is an articulation of such an organizational paradigm. The concept of a 'sensible organization' is related to the sense-making view of organizations (e.g., Weick, 1995; Wiley, 1994; Cecez-Kecmanovic & Jerram, 2002). There are three significant levels of sense-making (see Linger & Warne, 2001): individual, organizational, and an intermediate level involving teams, groups, or units. Knowledge has traditionally been understood at the individual level. It is often said that "only people know," and individuals learn as they acquire knowledge from others. At the organizational level we use metaphors of 'organizational learning' and 'organizational memory' in the context of formal knowledge repositories, intranets, databases, and data warehouses that are invariably ICT based. The focus of most knowledge management initiatives is at this organizational level, while less attention has been paid to the collective knowledge at the intermediate level.

The focus of the sensible organization is the intermediate level since the informality, interactivity and adaptability of small teams defines a space for what is traditionally called 'common sense'. Within this space, teams are able to construct shared understanding and take action based on that understanding, amid the accountability and constraints of the formal enterprise. In this sense, teams represent the site of most innovation and creativity in organizations, and consequently where the challenges and potential of a sense-making approach are most apparent. Sensible organizations therefore encourage the emergence of self-directed teams interconnected in a network-centric configuration as described in Warne, Ali, and Hasan (2005b).

What is proposed by the sensible organization is not new but a return to past skills that have often been overtaken by the bureaucratization of the workplace—a process that, in many instances, is itself a result of ICT-driven change.

THE SENSIBLE ORGANIZATION IN CONTEXT

As organizations change in order to maintain their strategic and sustainable position in the broader society, they are adopting flatter forms that require substantial changes in the way people work. These changes are directed to supporting agile teamwork and coordinated group activity that is flexible but also well aligned with the desired organizationally defined outcomes. The sensible organization needs to be understood in the context of its structural and functional forms, and the interdependencies between these forms, in shaping the organization.

On the other hand, IS research will need to increase its understanding of these transformed organizational cultures in order to provide advice on managing organizations where uncertainty and complexity are the norm. Such understanding is a necessary prerequisite to the design and implementation of ICTs that are consistent with the sensible organization. For ICT systems to effectively support the sensible organization, the underlying architecture will need to appropriate social technologies (e.g., Pfaff & Hasan, 2006; Hasan, 2006a) in a manner that empowers knowledge workers and democratizes organizational information.

In order for IS research to understand the sensible organization, it is necessary to examine elements that characterize the sensible organization and its context.

Situating the Sensible Organization—The Complex Environment

Organizations are confronted by increasing complexity and a rapid rate of change (Robbins, 1990), where the nature of change is frequently revolutionary rather than evolutionary. This is exemplified by the impact of ICT, and the Internet in particular, on how organizations work and interact with their environment. The challenge for the sensible organization is to successfully manage this transformative environment as a network of complex entities and to adopt 'systems thinking' (Senge, 1994) in order to recognize and understand emerging patterns of this complex world.

Traditionally, organizations adopted a reductionist approach that attempted to summarize the dynamics, processes, and change that occurred in terms of the lowest common denominators and the simplest, yet most widely provable, applicable, and elegant explanations. For the sensible organization, it is more appropriate to view the world as a complex system that includes numerous elements, arranged in structures, that go through processes of change. These changes are neither describable by a single rule nor are reducible to only one level of explanation and often include features whose emergence cannot be predicted from their current specifications (Hasan, 2006b). This approach is consistent with IS, which intrinsically takes a socio-technical systems view of the situations it investigates as illustrated by 'soft systems methodology' (SSM) developed by Checkland (1991).

For the organization to make sense of a complex system, it needs to accommodate the system's inherent dynamics, including the ability to incorporate unanticipated and unforeseen features that emerge from that dynamic. This requires an innovative means of understanding the longitudinal changes to the organization and the possibilities open to the organization in the future. Frameworks like Snowden's (2002) Cynefin model are useful to reach such understanding.

The Sensible Organization as a Complex System

Complexity itself is characterized by a number of important properties such as self-organization, non-linearity, and emergence. Snowden's (2002) Cynefin framework is a model that presents organizations as a knowledge space with five domains: two domains of order, the known and the knowable; two domains of unorder, complexity and chaos; and the undesirable domain of disorder. Each domain has a different mode of community behavior, and each implies a different form of management, a different leadership style, and the adoption of different tools, practices, and conceptual understanding. For the sensible organization, the 'complex' domain is of particular interest with its characteristics of self-determination, emergence, and organic forms.

Importantly, sensible organizations are not limited to any domain, but exhibit, to a greater or lesser extent, characteristics of each domain. Sensible organizations are often more like ecosystems than machines with one domain predominating in any specific situation. Using the Cynefin model one is able to see how organizations and their information systems can simultaneously be mechanistic and organic. When confronting a complex and changing environment, the sensible organization replaces rational planning with processes that stimulate patterns of propitious emergent activity with an emphasis on sense-making, unstructured decision making, and shared situation awareness. The current reality is that organizational transformations will continue to be a permanent feature and therefore it makes sense to view the organization primarily from the perspective of the Complexity domain.

The Sensible Organization Infrastructure—Socio-Technical Systems to Support Complexity

The characteristic of IS that distinguishes it from other management fields in the social sciences is that it concerns the use of "artifacts in human-machine systems" (Gregor, 2002). Conversely, the characteristic that distinguishes IS

from more technical fields, such as software engineering, computer science, and information technology, is its concern for the human elements in organizational and social systems. A basic premise of this article is that all information system artifacts are essentially socio-technical in nature and continually evolve so that they acquire emergent properties, uses, and impacts. The term *socio-technical* effectively expresses the intricate relationships and interrelationships between the social and technical elements of any information system, particularly within organizations. Taking into account the properties of complex environments, Coakes (2002) describes the goal of socio-technical design as producing systems capable of self-modification, of adapting to change, and of making the most of the creative capacity of the individual for the benefit of the organization. Scholtz (2002) also sees the socio-technical perspective as valuing small independent work groups engaged in highly varied tasks, managing their own activities, and often supported by technology. These descriptions support the notion that socio-technical principles, and their application, help organizations to explore complexity in the human, organizational, and technical aspects of change (Coakes, 2002).

IS draws significantly on the uniqueness of computer-based ICT and their place in shaping recent human, social, and organizational history. The rapid evolution of ICT is continuing to transform firms from relatively simple entities, with solid boundaries and formal hierarchical structures, into a much more complex interconnected set of internal and external relationships. However, many of the anticipated benefits of ICT have not been realized, a reflection of the fact that, in a truly complex environment, patterns emerge that cannot be predicted. The phenomena emerging from developments in ICT, such as 'information overload', are merely symptoms of deeper issues that are less visible but that have greater impact: ICT/IS in organizations has taken away:

- the routine parts of work, and with them the time and space for reflection;
- the hands-on experience of work so that people lose touch with reality;
- much of the face-to-face social interaction so that opportunities for building trust and understanding are diminished; and
- recognition of the importance of, and support for, the invisible, pervasive, informal organization.

It is these issues that present the greatest challenges and opportunities for IS research and practice.

The significance of ICT that could support the less formal, social aspect of teamwork has long been a focus of the field of computer supported cooperative work (CSCW). CSCW explores a wide range of issues concerning cooperative work arrangements and ICT support for those arrangements (Bannon, 1992). Importantly, CSCW focuses on work arrangements grounded in the actual work being performed. When new social technologies (e.g., Wikis and blogs) that support cooperation (Prinz & Kock, 2007) are introduced into the workplace, and used appropriately by organizations, some of the original promise of ICT may be sensibly fulfilled.

INVESTIGATING THE SENSIBLE ORGANIZATION: EMPIRICAL STUDIES

The case for the sensible organization is based on a reinterpretation of the authors' research published over the past 10 years to draw informed lessons from the past to make sense of current trends in order to anticipate, and prepare for, a more sensible future. This research involved field studies of diverse organizations, and used approaches, methods, and theories that were at times developmental, interventionist, and participatory, involving action research arrangements and the collection, analysis, and interpretation of qualitative data. These approaches take a holistic and systemic view, and address the attributes that are the focus of the sensible organization. The studies focus on complexity, dynamics, changing work practices, blending the social with the technical and the cultural, as well as the social aspects of learning, growing, and adapting. The following is the collection of inter-related research projects undertaken by the authors and the focus of each project.

- *Social Learning in the Australian Defense Organization*—Identified the significance of the human dimensions of even the most technical and bureaucratic of organizations (e.g., Ali, Pascoe, & Warne, 2002; Linger & Warne, 2001; Warne, Ali, Pascoe, & Agostino, 2001; Warne, Ali, & Pascoe, 2003; Warne, Hasan, & Ali, 2005a).
- *Perceptions of Middle East Warfighters*—Highlights the importance of networks in modern agile organizations and the need to understand the human side of the network-centric view of organizations (e.g., Ali, 2006; Pascoe & Ali, 2006; Warne, 2006).
- *Weather Forecasting*—Demonstrated the importance of knowledge-based work practices to sense-making at the intermediate, team-based level of the organization (e.g., Iivari & Linger, 2000; Linger & Burstein, 2001).
- *Developing Web Communities*—Established an understanding of voluntary collaboration and natural ways of working together when civil society appropriates the Internet-enabled digital culture (e.g., Connery & Hasan, 2005; Hodgkinson & Hasan, 2006).
- *Corporate Wikis*—Explored the benefits and challenges of appropriating social technologies from the civil digital culture to support social learning in corporations (e.g., Pfaff & Hasan, 2006).

INVESTIGATING THE SENSIBLE ORGANIZATION: LESSONS LEARNED

The issues raised by these studies point to changes that are necessary for transformation to a sensible organization. The complex environment presents organizations with a very large range of problems and opportunities. Organizational response to this diversity of inputs is ideally to become agile, flexible, and adaptable by transforming their structures and processes. Such a response is consistent with Ashby's (1957) Law of Requisite Variety as the organization attempts to construct itself as an entity that can match the variety that it confronts in its environment. Thus, for example, a network-centric structure allows the organization to be more flexible than a more rigid hierarchical command and control structure in order to adapt itself dynamically to changes in the environment. The variety of the network-centric structure is derived from the ability of individual nodes to respond to the environment. These nodes are constrained only by the dynamics of the network. The network-centric structure is dependent on small, autonomous, self-directed, and self-coordinating groupings. Such groups are more apt at recognizing and understanding changes in the activity system for which they have responsibility and have the expertise and authority to act on that understanding. Such action, within a Network-centric structure, impacts on other groups, who in turn take action, and this provides the organization with the requisite variety to dynamically interact with its environment. The organization still preserves the capability to make sense of its situation from a broader, more abstract, and longer-term perspective, and to act on that understanding. This can be contrasted to the monolithic response of the control and command structure that requires considerable lead times in order to undertake the rational process of information gathering and decision making in order to reach its singular response. The response is predicated on the assumptions of "perfect" information and the ability to reach the optimal decision, often without regard to time constraints.

The studies highlight that organizations are transforming by adopting flatter hierarchical structure and combining these with a network-centric configuration. These transformations are challenging because culture change is imposed much more rapidly than it would normally occur. Managers are having to relinquish some of their traditional control to small, self-directed teams while workers need to increase their situational awareness in order to take on more responsibilities and exercise authority within a small, less prescribed group setting. This is a considerable change from the way they would have operated in the past and often there is little training or even understanding of the skills and capability that are needed.

In order to match their volatile environments, most enterprises today espouse the idea that they are a learning organization able to transform themselves as needed in a creative and informed manner. The studies articulate the critical role of social learning to the growth of modern enterprises and the importance of developing a culture of empowerment, trust, forgiveness, openness, commitment, and recognition. On the other hand, many of the knowledge work practices that are essential to the development of expertise and innovation in the socio-technical system are often invisible and are thus not acknowledged nor recognized by the organization. The studies also show that the necessary skills and capabilities required in a network-centric environment can be supported in a mix of face-to-face and online collaborative spaces and viewed holistically as socio-technical systems. Reinforcing the imperative to transform to a network-centric structure, the studies identified that effective teamwork was considered crucial in achieving operational goals. But with this transformation, integrity, maturity, adaptability, flexibility, competency and a sense of humor emerge as highly rated skills and qualities that are important for team members.

FUTURE TRENDS

The sensible organization targets the chaotic, complex environment of the agile, social, networked organization, and related communities, where members are knowledge workers engaged in activities that meld thinking and doing. This requires a new approach to developing supporting technologies. The focus of ICT needs to expand from organizational requirements and individual users to the support needed for employees working in self-directed groups and teams, as well as investigating the appropriation of social software and other emerging technologies to enable more democratic management of collective knowledge.

The exponential growth in the use of technologies like YouTube and Facebook suggest that social software needs to be channeled into organizational applications that support the creative energies of complex environments, encouraging the emergence of innovative new forms of working. Such transformations require a multi-perspective, multi-disciplinary approach together with holistic systems thinking. This is the new challenge for ICT developers.

Organizations need to reinvent their competitive advantage by introducing, or reintroducing, collaboration technologies and practices within and between their internal boundaries (Josserand, 2004). This requires blurring the distinction between organizational systems and those used in civil society, and the corporate appropriation of social technologies. It highlights the need to increase our understanding of the organizational cultures of community networks in civil society in order to understand how to manage organizations where uncertainty and complexity are the norm.

CONCLUSION

Large bureaucratic organizations are facing rapid and substantial changes that require new understandings, skills and the capability inherent in network-centric structures. Many organizations are adopting hybrids of a traditional hierarchy with a limited command and control structure that allow for the emergence of self-directed groups in a network-centric configuration. The sensible organization adopts network-centrism and encompasses the organizational, social and cultural as well as the technical aspects of working in these changing, hybrid environments.

Knowledge workers need the authority, skills, and capability to work effectively and creatively in teams as much as they need the traditional operational skills to do their job. In a sensible organization this creative and cognitive work needs to be overtly recognized and rewarded. Management needs to encourage sense-making and exploration rather than resort to traditional methods of setting objectives and measuring outcomes. A sensible organization is grounded in a culture that promotes social learning, community/team building, and the appropriation of new flexible technologies in emergent socio-technical systems.

In terms of the Cynefin framework, socio-technical systems, and particularly ICT, have a long history of supporting, and even automating, aspects of the *known* and *knowable* domains. The challenge now is to understand the sensible organization in order to design and construct the necessary socio-technical system that will enable the sensible organization to be effective in the *complex* and *chaotic* domains, while it remains efficient in the *known* and *knowable* domains.

Sense-making as a concept underpins a new approach to research. The sense-making approach allows sensible organizations to emerge from much of the chaos that has arisen from a general lack of understanding of the cultural, social, and personal effects of ICT, and more particularly in recent times, social technologies. The emergence of the sensible organization requires cultural investigations as part of requirements analysis to facilitate a sound understanding of existing work practices in terms of both social and functional processes and networks. Research needs to expand its focus to the development of a culture of cooperation and the creation of an environment for innovation through collaborative knowledge work. In this environment, civil society, as well as the government and corporate sector, need to be included as legitimate sites of research.

The sensible organization can define a new organizational paradigm within which socio-technical systems are not only focused on automating work, but also supporting the way it is done and enhancing productivity and the work experience itself**.**

ACKNOWLEDGMENT

The information in this article is largely derived from the authors' own research work and those of their colleagues, and from a paper they have co-authored entitled, "The Sensible Organization: A New Agenda for IS Research," presented at the International Conference on Information Systems (ICIS 2007) in Montreal in December 2007.

REFERENCES

Ali, I. (2006, September 26-28). Information sharing and gathering in NCW environment: Voices from the battlespace. *Proceedings of the 11th International Command and Control Research and Technology Symposium,* Cambridge, UK. Retrieved from http://www.dodccrp.org/events/11th_ICCRTS/iccrts_main.html

Ali, I., Pascoe, C., & Warne, L. (2002). Interactions of organizational culture and collaboration in working and learning. *Educational Technology and Society, 5*(2), 60-69.

Ashby, W.R. (1957). *An introduction to cybernetics*. London: Chapman & Hall. Retrieved from http://pcp.vub.ac.be/books/IntroCyb.pdf

Bannon, L. (1992, August). Perspectives on CSCW: From HCI and CMC to CSCW. *Proceedings of the International Conference on Human-Computer Interaction (EW-HCI'92) (pp. 148-158), St. Petersburg, Russia.*

Burstein, F., & Linger, H. (2003). Supporting post-Fordist work practices: A knowledge management framework for dynamic intelligent decision support. *Information Technology & People, 16*(3), 289-305.

Cecez-Kecmanovic, D., & Jerram, C.A. (2002). Sensemaking view of knowledge in organizations. *Proceedings of ECIS2002,* Gdansk, Poland.

Checkland, P. (1981). *Systems thinking, systems practice.* Chichester: John Wiley & Sons.

Coakes, E. (2002). Knowledge management: A sociotechnical perspective. In E. Coakes, D. Willis, & S. Clarke (Eds.), *Knowledge management in the sociotechnical world* (pp. 4-14). London: Springer-Verlag.

Connery, A., & Hasan, H. (2005). Social and commercial sustainability of regional Web-based communities. *Journal of Web-Based Communities, 1*(3), 246-261.

Daft, R.L., & Lewin, A.Y. (1993). Where are the theories for the "new" organizational forms? An editorial essay. *Organization Science, 4,* i-vi.

DeLone, W.H., & McLean, E.R. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research, 3*(1), 60-95.

Drucker, P (1959). *Landmarks of tomorrow.* New York: Harper.

Edmunds, A., & Morris, A. (2000). The problem of information overload in business organisations: A review of the literature. *International Journal of Information Management, 20,* 17-28.

Fortune, J., & Peters, G. (2005). *Information systems: Achieving success by avoiding failure.* Chichester: John Wiley & Sons.

Gregor, S. (2002). A theory of theories in information systems. In S. Gregor & D. Hart (Eds.), *Information systems foundations: Building the theoretical basis* (pp. 1-20). Canberra: ANU.

Hart, D., & Warne, L. (2005). Comparing cultural and political perspectives of data, information, and knowledge sharing in organizations. *International Journal of Knowledge Management, 2*(2), 1-15.

Hasan, H. (2006a). Innovative socio-technical systems for complex decision-making. *Proceedings of the IFIP8.3 DSS Conference,* London.

Hasan, H. (2006b, December). Design as research: Emergent complex activity. *Proceedings of the 17th Australasian Conference on Information Systems,* Adelaide.

Hodgkinson, A., & Hasan, H. (2006). Blending diverse community capability for regional development: The case of an e-commerce initiative for local indigenous artists. *Proceedings of 28th Annual Conference of the Australian and New Zealand Regional Science Association,* Beechworth, Australia.

Iivari, J., & Linger, H. (2000, August). The characteristics of knowledge work: A theoretical perspective. *Proceedings of the Americas Conference on Information Systems* (AMCIS'2000), Long Beach, CA.

Josserand, E. (2004). Cooperation within bureaucracies: Are communities of practice an answer. *M@n@gement, 7*(2), 307-339.

Linger, H., & Burstein, F. (2001). From computation to knowledge management: The changing paradigm of decision support for meteorological forecasting. *Journal of Decision Systems, 10*(2), 195-216.

Linger, H., & Warne, L. (2001). Making the invisible visible: Modelling social learning in a knowledge management context. *Australian Journal of Information Systems, 8,* 56-66.

Lyytinen, K., & Hirschheim, R. (1987). Information systems failures—a survey and classification of the empirical literature. *Oxford Surveys in Information Technology, 4,* 257-309.

Pascoe, C., & Ali, I. (2006, September 26-28). Network centric warfare and the new command and control: An Australian perspective. *Proceedings of the 11th International Command and Control Research and Technology Symposium,* Cambridge, UK. Retrieved from http://www.dodccrp.org/events/11th_ICCRTS/iccrts_main.html

Pfaff, C.C., & Hasan, H. (2006). Overcoming organisational resistance to using Wiki technology for knowledge management. *Proceedings of the 10th Pacific Asia Conference on Information Systems,* Kuala Lumpur, Malaysia.

Prinz, W., & Kock, M. (2007). Why CSCW research Web 2.0 and social software solve our problems anyhow! *Proceedings of ECSCW07,* Limerick, Ireland. Retrieved from http://lubnalam.wordpress.com/2007/06/18/workshop-why-cscw-research-web-20-and-social-software-solve-our-problems-anyhow/

Robbins, S.P. (1990). *Organization theory: Structure, design, and applications* (3rd ed). Englewood Cliffs, NJ: Prentice Hall.

Sauer, C. (1993). *Why information systems fail: A case study approach.* Alfred Waller.

Scholtz V. (2002). Managing knowledge in a knowledge business. In E. Coakes, D. Willis, & S. Clarke (Eds.), *Knowledge management in the socio-technical world* (pp. 43-51). London: Springer-Verlag.

Senge P. (1994). *The fifth discipline: The art and practice of the learning organization.* New York: Doubleday.

Snowden, D. (2002). Complex acts of knowing: Paradox and descriptive self-awareness. *Journal of Knowledge Management, 6*(2).

Warne, L. (2002). Conflict and politics and information systems failure: A challenge for information systems professionals and researchers. In S. Clarke (Ed.), *Socio-technical and human cognition elements of information systems* (pp. 140-135). Hershey, PA: Idea Group.

Warne, L. (2006, May 1-3). NetworkER centric warfare: Outcomes of the human dimension of future warfighting task. *Proceedings of the TTCP Symposium—Human Factors Issues in NCW,* Sydney, Australia. Retrieved from http://www.dsto.defence.gov.au/events/4659/

Warne, L., Ali, I., Bopping, D., Hart, D., & Pascoe, C. (2004). *The network centric warrior: The human dimension of network centric warfare (U).* DSTO CR-0373,

Defense Systems Analysis Division, ISL, Defense Science and Technology Organization, Department of Defense, Australia. Retrieved from http://dspace.dsto.defence.gov.au/dspace/handle/1947/3403

Warne, L., Ali, I., & Pascoe, C. (2003). *Social learning and knowledge management—a journey through the Australian Defense Organization: The final report of the enterprise social learning architectures task.* DSTO-RR-0257, Defense Systems Analysis Division, Information Sciences Laboratory, Department of Defense, Australia.

Warne, L., Ali, I., Pascoe, C., & Agostino, K. (2001). A holistic approach to knowledge management and social learning: Lessons learnt from military headquarters. *Australian Journal of Information Systems, 8*(special issue), 127-142.

Warne, L., Hasan, H., & Ali, I. (2005a). Transforming organizational culture to the ideal inquiring organization: Hopes and hurdles. In J.F. Courtney, J.D. Haynes, & D.B. Paradice (Eds.), *Inquiring organizations: Moving from knowledge management to wisdom* (pp. 316-336). Hershey, PA: Idea Group.

Warne, L., Ali, I., & Hasan, H. (2005b). The network centric environment viewed through the lens of activity theory. In G. Whymark & H. Hasan (Eds.), *Activity as the focus of information systems research* (pp. 117-140). Eveleigh, Australia: Knowledge Creation Press.

Weick, K.E. (1995). *Sensemaking in organizations.* Sage.

Wiley, N. (1994). *The semiotic self*. Cambridge: Polity Press.

KEY TERMS

Collective Activity Systems: The work of groups and teams can be viewed as collective activity systems which are carried out by people in support of their interpretations of their role, the opportunities and resources available to them, and the purpose for which the activity exists. Using the language of activity theory, this is both subjective, in the sense that it is a matter for individual interpretation, and objective, in the sense that the motives, purpose, and context are a vital part of the reality of human work. An activity is defined by the dialectic relationship between a subject (i.e., a person or small group of people) and the object of their work, which includes purpose, motive, and context. An activity both mediates and is mediated by the tools used and the social context of the work activity. This two-way concept of mediation implies that the capability and availability of tools mediates what can be done, and the tool in turn evolves to hold the historical knowledge of how a society works and is organized.

Complexity: Understood as a concept in various ways, yet not definitively defined in the literature. However, in recent times, comprehensive theories of complexity and chaos have become popular in many disciplines of both the natural and social sciences. These theories reflect the tension between the natural tendency for disorder to increase while humans strive to impose order by developing ever more interconnected systems. In business today success is no longer determined by a few single factors, but by systems with multiple interacting relationships. Thus complexity and chaos theories are being applied in organization science where both operations and management in human enterprises are becoming increasingly complicated. The response is frequently to impose greater planning, control, rules, and regulation. At some point, organizations reach a state of complexity where planning and control of mandated work-practices should give way to the provision of a supportive environment that allows innovation and creativity for problem identification and solutions to emerge. The former is likely to be exploitative and bureaucratic, while the latter can be networked and innovative.

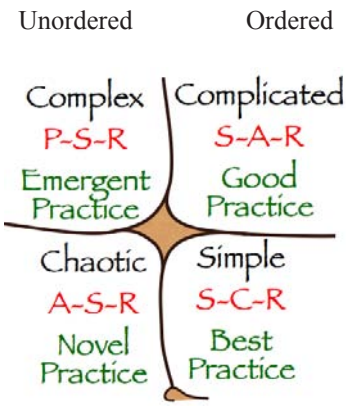
Cynefin (pronounced kun-ev'in)**:** The name of a sense-making framework proposed by Snowden (2002). It is a knowledge space with five domains setting the context for collective decision making which has been used in knowledge management as well as other applications including conflict resolution. The domains are, characterized by the relationship between cause and effect. The first four domains are:

1. **Simple or Known,** in which the relationship between cause and effect is obvious to all; the approach is to *Sense–Categorize–Respond*, and we can apply *best* practice.
2. **Complicated or Knowable,** in which the relationship between cause and effect requires analysis or some other form of investigation and/or the application of expert knowledge; the approach is to *Sense–Analyze–Respond*, and we can apply *good* practice.
3. **Complex,** in which the relationship between cause and effect can only be perceived in retrospect, but not in advance; the approach is to *Probe–Sense–Respond*, and we can sense *emergent* practice.
4. **Chaotic,** in which there is no relationship between cause and effect at systems level; the approach is to *Act–Sense–Respond*, and we can discover *novel* practice.

The fifth domain is **Disorder**, which is the state of not knowing what type of causality exists, in which state people will revert to their comfort zone in making a decision.

Cynefin Framework: Defined by the following figure

(adapted from Wikipedia).



Knowledge Work: The term 'knowledge worker' is attributed to Drucker (1959), who used it to describe someone who adds value by processing existing information to create new information that can be used to define and solve problems. The subsequent development of information and communications technologies has added new meaning to this concept. For the contemporary knowledge worker, managing the collective knowledge about his or her work is an integral part of the work itself, and thus is critical to the performance of the organization. Knowledge work reflects the self-directed work practices of individuals and teams, in almost every industry, who continuously engage in processes that create and exploit knowledge. The modern work activity system is located within a space defined by the doing, thinking, and communicating dimensions (Burstein & Linger, 2003).

Network Centricity/Centrism: Large bureaucratic organizations, and the people who work in them, are facing rapid and substantial changes which require new understandings, skills, and capabilities for the network-centric environment. In some of the early literature, the term 'network-centric' only referred to the connectivity achieved through technological networks, in particular the Internet and Web-enabled applications. However its connotation has expanded as ICT networks and applications are transforming the ways in which people gather, share and process information and knowledge, and consequently, on the ways they make decisions to act. This is having an impact in organizations: on their structures, their ways of working, on organizational learning, as well as on the ways people collaborate and form social networks. Many organizations are now hybrids of a traditional hierarchy, with a limited command and control structure, and a network-centric configuration allowing the emergence of self-directed groups. The domain of network-centrism now encompasses the organizational, social and cultural as well as the technical aspects of working in these changing, hybrid environments.

Requisite Variety: Principle proposed by Ashby (1957) suggesting that the internal diversity of any self-regulating system, such as an organization, must match the variety and complexity of its environment if it is to deal with the challenges posed by that environment. Diversity of knowledge and skill can provide a resource for innovation and learning, at all levels of organizational management. If the systems that regulate do not have enough (or requisite) variety to match the complexity of the regulated, then regulation will fail. The system will be out of control. If an organization is complicated or complex, it is likely to have plenty of variety; if it is simple (e.g., purely hierarchical), the variety is usually low and the organization will struggle with the current complex environment.

Sensible Organization: There are many things about work in today's organizations that just do not make sense. We observe contradictions, stresses, and tensions everywhere. The advancements of science and technologies, which promised to take away the drudgery of the human condition do not seem to have fulfilled their promise. We have replaced 'drudgery' with the new disease of 'affluenza'. We fondly reflect on the creativity and community spirit of pre-industrial age cottage industries and the subsequent de-humanization of the workforce in the assembly line of factories of the industrial age. The notion of 'sensible organization' is a return to the human and social values that have disappeared in the modern workplace. Sensible assumptions about most modern organizations, which have complex hybrid structures consisting of hierarchies and networks, is that they are often more like organic ecosystems than machines. Moreover, it makes sense to adopt the position that this mechanistic-organic hybrid is now a natural state of affairs and should not be resisted. Indeed this creates an ideal context for innovation, creativity, and growth—a context in which rational planning should give way to processes that stimulate patterns of propitious emergent activity with an emphasis on sense-making, unstructured decision making, and shared situational awareness.

Social Learning: Learning that occurs within or by a group, an organization, or any cultural cluster, and it includes:

- the procedures by which knowledge and practice are transmitted across different work situations and across time; and
- the procedures that facilitate generative learning that enhances the enterprise's ability to adjust to dynamic and unexpected situations, and to react creatively to them.

Social learning represents important processes that contribute to individuals' abilities to understand information, create knowledge from that information and share what they know. Social learning is therefore intrinsic to the factors in organizations that enhance and enable the assimilation, generation, sharing, and building of knowledge that transforms an organization into a learning organization.

Social Software (or Social Technology): A new civil digital culture has taken hold, in which so-called 'social' and/or 'conversational' technologies are providing unprecedented opportunities for everyday civil user activities. The attraction of these social technologies is their low cost, intuitive functionality, and connectivity. Social technologies provide computer-mediated environments that use applications such as Weblogs (blogs), Wikis, chatrooms, and various Web-based groupware systems. They support new forms of informal, network-centric interaction and activity between people, allowing and enhancing informal access to create and distribute information. These technologies empower ordinary people to have a global presence for business, political and social purposes. The new social technologies are tools of a rising digital democracy that provide users with a new flexibility and independence to support collective actions, knowledge sharing and decision making by self-directed groups. Social technologies, which support cooperative socio-technical systems, are being appropriated by enlightened enterprises which are transforming from traditional hierarchical structures to more network-centric configurations.

Category: Electronic Business

Complexity Factors in Networked and Virtual Working Environments

Juha Kettunen
Turku University of Applied Sciences, Finland

Ari Putkonen
Turku University of Applied Sciences, Finland

Ursula Hyrkkänen
Turku University of Applied Sciences, Finland

INTRODUCTION

Working environments are changing from the traditional model. An increasing amount of work takes place in networked and virtual environments which are not tied to one place and time. The work environment is defined "virtual," when the employee uses information and communication technology (ICT) for collaboration (Vartiainen, 2006). The planning of working conditions becomes challenging task for managers and ICT tool developers, because there is a lack of understanding the consequences of emerging virtual work.

The capacity of workers to percept and process information is burdened with the complexity and high demands of working life. Knowledge of the complexity factors of the overall work system is essential for an in depth understanding of human working capabilities and limitations (Kleiner, 2006). The complexity of work is usually considered as a factor related to the task. At the one end the task is creative and demanding and at the other end it is simple and routine-like. The expanded complexity concept also takes into account the working environment that can be different combinations of physical, virtual, social and cultural spaces.

The purpose of this article is to present a framework to analyse the complexity factors in networked and virtual working environments. The approach developed in this article is intended to be generic in order to be applicable to various kinds of organisations and networks for the purpose of management. It is important that the working conditions of workers can be planned in advance to provide workers with appropriate ICT tools and data networks to enable efficient cooperation in networks in a way that the workload can be limited to a sustainable level. The described framework is assessed using the case of the Turku University of Applied Sciences (TUAS).

BACKGROUND

Organisational Context of the Study

Networked and virtual work are analysed by applying the complexity approach to the Turku University of Applied Sciences. The strategic plan of the TUAS is to react to the changes in a flexible way (Kettunen, 2006, 2007; Kettunen & Kantola, 2006). The interaction of the institution is close with its operational environment. The purpose of the institution is to react to the changes in its environment in a flexible way and to increase its external impact on the region.

TUAS is a multidisciplinary higher education institution founded in 1992. The City of Turku owns the institution, which has 800 full-time employees. The TUAS has six faculties and a Continuing Education Centre. ICT is an important field of education and is combined with business, biotechnology, mechanical engineering, health care, performing arts, communication and many other subjects. Cooperation with other universities is active. One reason is that the ICT education and research of the University of Turku and Åbo Akademi University are located in the same ICT Building.

The TUAS has 9,500 degree students. The institution offers tuition mainly in Finnish but there are also degree programmes, modules and courses in other languages. Internationalisation is one of the focus areas of the institution. The TUAS has wide international networks. The institution has cooperation with several higher education institutions in Europe, Asia and the Americas. Five entire programmes are taught in English. The objective is to improve the students' ability to work in a global environment.

Networked and Virtual Environments

Figure 1 describes the dimensions of networked and virtual work. There are three modes for organising the communication and collaboration of work: traditional organisation,

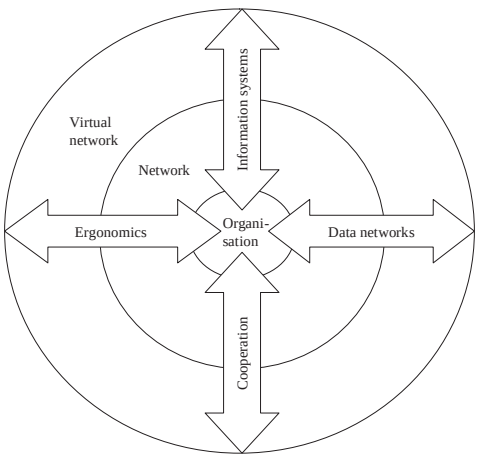
network and virtual network. The concept of the virtual network includes networked and traditionally organised work. Mobility may take place within or outside the organisation. Networked and virtual work can be analysed using the various dimensions that come across the organisation and networked and virtual environments.

Information systems have been typically planned for the organisation, but an increasing amount of information systems have also been designed for the cooperation in networked and virtual environments. Virtual work environments increase the complexity of the work and therefore, various approaches are required to analyse and design the well-being of workers and the performance of the overall work system.

Data networks and ICT reflect the needs of working environments. The traditional organisation-centred work is extended with the mobile work, networked cooperation in diverse locations and virtual systems, which increase the complexity of working environments. Working in these environments requires not only traditional network but also an increasing amount of wireless data networks.

Ergonomics as a discipline is concerned with interactions with human-machine systems. Ergonomics has played a vital role, for example, in the reduction of occupational injuries, improved performance, increased health and safety of workers and end-products. Pheasant (1996) has concluded that the objective of ergonomics is to achieve the greatest possible harmony between the product and its users in the context of the task performed. With complex sociotechnical systems the above is not enough because various subsystems exist and the importance of interactions between them has a significant role. Kleiner (2006) emphasises that the larger work systems have to consider when there is a need to understand human-technology interaction, capabilities and limitations better. Macroergonomics as a subdiscipline of ergonomics is concerned with human-organization interface technology and the optimisation of the overall work system, that is, design of the worker-job, worker-machine, and worker-software interfaces (Hendrick & Kleiner, 2001).

Figure 1. Dimensions of networked and virtual work



ORGANIZATIONAL EXPANSION TOWARD THE VIRTUALITY

Complexity Factors of Networked and Virtual Work

Vartiainen (2006) has described the complexity of working environments by following six dimensions: mobility, geographical dispersion of the workplaces, diversity of actors, asynchronous working time, temporary structure of the working groups and mediated interaction. These six dimensions form, in addition to task complexity, a set of requirements that can also be considered work load factors (Hyrkkänen, 2006). The complexity factors arise from the conjunction of a worker and the particular kind of working environment. We have further developed the complexity model of the networked and virtual work by exploring it taking into account the human and technology related enablers and limitations of virtual work.

Table 1 describes the general concepts of complexity factors for the traditionally organised, networked and virtual work. We suggest this classification as a framework, which can be used for example, as a framework for empirical studies, participatory design projects and consulting processes. The working environments are changing from the traditional model, where the place and time have an important role. An increasing amount of work takes place in networked and virtual environments which are not tied to one place and time. The planning of working conditions becomes challenging, because there is a lack of proper tools for analysing and managing sustainable and safe working conditions.

The information environments and systems can be classified as mechanical, organic and dynamic (Ståhle & Grönroos, 2000; Ståhle & Hong, 2002; Ståhle, Stahle, & Poyhonen, 2003). The information systems in the mechanical information systems increase the efficiency of internal processes and include thoroughly controlled information systems, which are based, for example, on the automation of routines. The organic information system emphasises dialogue, communication and sharing of experience-based tacit knowledge (Kim, Chaudhury, & Rao, 2002; Takeuchi & Nonaka, 2004). The dynamic information systems include information systems which produce innovations and services by self-organisation. The virtual systems, networking, net casting and different portals, are often in connection with

Table 1. General concepts of complexity factors in an organisation, network and virtual network

	Information systems	Data networks	Cooperation	Ergonomics
Organisation	• Mechanical	• Intranet	• Internal working groups	• Microergonomics
Network	• Organic	• Wireless local area networks	• Working in networks	• Sociotechnical systems • Macroergonomics
Virtual network	• Dynamic	• Internet	• Working in virtual environments	• Macroergonomics

more risky, chaotic and innovative environments and online networking platforms (Steinberg, 2006).

Data networks include typically Internet, local wireless networks and Intranet. Mobile devices enable workers access to wireless networks. The emerging and expanding wireless networks enable the use of the data network wirelessly. A wireless local area network (WLAN) enables communication between devices in a limited area. WLAN or Wireless Fidelity (WiFi) is becoming increasingly popular with the rapid emergence of portable devices. Other similar certifications such as Worldwide Interoperability of Microwave Access (WiMAX) offer longer ranges of radio waves.

Cooperation and collaboration is not only within working groups within organisations, but typically in networks and virtual environments. Cooperation in networks is typical in clusters, which are geographic concentrations of interconnected companies, specialised suppliers, service providers, firms in related industries, and associated institutions in particular fields that compete but also cooperate (Porter, 1998). Many of the advantages of clusters involve location-specific business services and networking. Some of the clusters are closely aligned to government and to public institutions (Denison, 2007; Graham, 2005). They include face-to-face contact, close and ongoing relationships and access to information via data networks. Some of the cooperation exists only in virtual networks. A typical example is the development of the Linux, which is one of the most prominent examples of open source development and free software.

Ergonomics can be classified as microerconomics, which focuses on the interfaces between the individual, organisation and other system elements and macroergonomics, which focuses on the design of overall work system. Organisational factors are related to the physical activity of workers, environmental factors and organisational complex factors. When the functioning of the networked working environment is under development, it is advisable to analyse it as a sociotechnical system, covering interaction between workers and technology (Carayon, 2006). Macroergonomics is an approach of work systems which attempts to achieve a fully harmonized work system at both the micro- and macro-ergonomic level by integrating principles and perspectives from industrial, work and organisational psychology (Kleiner, 2006). Typically the complexity of work increases when workers are moving to virtual work environments (Richter, Meyer, & Sommer, 2006).

Networked and Virtual Work at the TUAS

Table 2 describes the complexity factors of work at the Turku University of Applied Sciences. It is not an exhaustive description of all the characteristics of the traditionally organised, networked and virtual work of the institution, but it provides examples of complexity factors of the institution. The approach helps the management of the institution to analyse the complexity factors and take them into account in the work design and human resources planning.

The information systems include the devices and software used in the various information environments. Mechanical information systems include traditional systems such as accounting, personal administration and payroll systems, which are strictly tied to the processes and structures of an organisation. Organic information systems include, for example, query systems which are used to attain feedback from students and employers. They include also the library, project management and management information systems (Kettunen & Kantola, 2005). The dynamic information systems include the courses in the platform of virtual learning, Finnish virtual university portal and services of virtual libraries.

Data networks include, for example, Intranets and Internet. Wireless local area networks are expanding rapidly. An example of the wireless network is the SparkNet. It is

Table 2. Complexity factors of work at the TUAS

	Information systems	Data networks	Cooperation	Ergonomics
Organisation	• Desk computers • Internal phones	• Administration and study intranets	• Working groups in educational development and research	• Face-to-face development of safe, healthy and efficient tools and work environments
Network	• Laptops • Communicator • 3G telephones • VoIP	• WLAN/WiFi (SparkNet) • WiMAX	• Local, national and international networks of higher education and research	• Local, national and international mobility, geographical dispersion of workers and information technology
Virtual network	• Virtual learning environments	• Internet • Finnish University and Research Network (Funet)	• Studying and cooperating in virtual environments	• Communication and interaction are ICT mediated • Temporary and self-organised virtual teams

located in Turku and is the largest and most extensive wireless network solution in Finland with about 100,000 users. The coalition of wireless networks was established in 2003 by the local universities, the City of Turku, MasterPlanet Ltd. and a development company ICT Turku Ltd. The idea of SparkNet is based on exploiting existing network resources wirelessly. The members of the SparkNet coalition build pieces of a public Wireless Fidelity (WiFi) network instead of building their own WiFi network. SparkNet is easy and affordable for the students to pursue their studies wherever they want using Voice over Internet Protocol (VoIP), online courses, video conferencing (Skype) and other systems.

The cooperation and networks of the TUAS are primarily located in Southwest Finland, which is the second largest economic area after the capital of Finland. ICT, biotechnology and mechanical engineering are the strongest clusters of the region, which are engaged in the region in mutually beneficial cooperation with higher education, research, commerce and culture. This emphasises the important role of networking in the clusters of the region. The institution has also national networks with 20 traditional science universities, 28 professional-oriented universities of applied sciences, other education institutions and enterprises and other partner organisations. The annual number of cooperation contacts is 7,000-8,000 including education projects, applied research and development, practical training and administration.

Ergonomic methods are required when the safe, healthy and efficient tools or work environments are designed. The networked work of the TUAS has more enriched job characteristics such as local, national and international mobility, the geographical dispersion of workers and information technology. Typically, the cooperation and collaboration of international activities is complex due to various geographical locations, asynchronous work in different time zone and diversified cultural backgrounds. The workers in virtual environments have a larger amount of organisational tasks, more demanding learning requirements and level of participation than the workers in traditional jobs. Workers' communication and interaction are mainly ICT mediated. A worker participates in several virtual teams at the same time. Teams are often temporal and heterogeneous including people from different cultural backgrounds. Usually, the virtual work environment of the individual worker is not controlled by the managers; it is self-managed by the worker and this causes an additional dimension of complexity.

FUTURE TRENDS

A number of changes are taking place in global business and technology that lead to the increasing complexity of work

systems. Today, "system" in ergonomics means a broad range of working environments from the usage of a simple tool to the performing of a task in a complex sociotechnical environment (Carayon, 2006). Such complexity can be seen in the expansion of virtual organizations and mobile workers.

In the future, especially with complex work systems, a human-centred design approach has to be used as opposed to a technology-centred approach (Hendrick & Kleiner, 2001). Early observation of the system's complexity is an increasingly important managerial task in order to design a work system where the well-being of workers and the overall system performance are in balance. The earlier the input occurs, the greater the impact on costs and schedule. Redesigning an existing work system is expensive and time consuming.

There are some principles concerning how the human-centred design approach can be realised. Hendrick and Kleiner (2001) list three criteria what are essential for an effective work system design approach; (1) joint design, (2) a humanized task approach and (3) consideration of the organization's sociotechnical characteristics. Joint design means that the personnel subsystem and technological subsystem should be developed simultaneously and supported by employee participation throughout the entire design process. Humanized task approach is concerned with human functions and tasks in the work system prior to the decision to allocate tasks to workers or machines. The sociotechnical characteristics of the organisation should be assessed and integrated into the work system's design process. When an evaluation or development methodology fulfils the above mentioned three criteria it is human-centred and macroergonomic.

A working virtual team requires both ICT mediated and face-to-face interactions, at least as long as ICT tools do not include more sophisticated features. The functionality of a virtual team relies on the interactions between team members. A group of people who are working separated by distance or time and have a shared task to perform, is dependent on common tools, information and social dealings. Existing ICT tools such as mobile phones, e-mail and video conferencing are rather well fulfilling the needs of information transmission. However, they have limited features to mediate issues relating to social dealings, that is, when a new team member collaborates it is difficult to identify with the rest of team thorough ICT, because, for example, the development of trust needs a richer communication channel than text or voice only.

According to Edwards and Wilson (2004), when a virtual team is able to cross the organizational, cultural and functional boundaries of a single organization, it can deliver the combined skills and knowledge required to be more competitive in fast-changing markets. In addition, Wilson (2000) emphasises that we should study interactions not simply to design artefacts but to understand the interactions themselves in order to design more sophisticated ICT. These are the fundamental driving forces of the future ICT tool design.

CONCLUSION

This study presented a useful approach to analyse the complexity factors of networked and virtual work. The organisational environment was extended toward networked virtual work, where it is useful to analyse the different complexity dimensions. The dimensions of this study include information systems, data networks, cooperation and ergonomics. The approach can be used in empirical studies, participatory design projects and consulting. This approach can also be extended using other appropriate dimensions.

Networked and virtual work provides new challenges and we can no longer understand the behaviour of workers and their performance at work as we once did. For instance, we should now study interactions between different organisational dimensions, not simply to design new communication tools but to understand the interactions themselves in order to manage the overall work system.

When the functioning of the entire virtual working environment is under development, a human-system interaction and the entire work organization and sociotechnical system have to be taken jointly into consideration. Macroergonomics is an approach of work systems design which attempts to achieve a fully harmonized work system at both the macro- and micro-ergonomic level. In future, macroergonomics should be more common knowledge among managers in order to meet the design and development challenges of complex work environments. This means, for example, that workers should be more involved in the design and implementation of technology and new information and communication systems in organisations.

As organisations expand to virtual environments, the need for experience and special solutions will increase. However, supporting the work of virtual environments is a difficult assignment. Identifying complexity factors is a good start. The existing models of complexity are mainly based on the concept of traditional organisational work. This study contributes to the current discussion by offering one model for assessing the complexity factors of networked and virtual work.

REFERENCES

Carayon, P. (2006). Human factors of complex sociotechnical systems. *Applied Ergonomics, 37*(4), 525-535.

Denison, T. (2007). Support networks for rural and regional communities. In H. Rahman (Ed.), *Information and communication technologies for economic and regional developments*

(pp. 102-120). Hershey, PA: Idea Group.

Edwards, A., & Wilson, J.R. (2004). *Implementing virtual teams: A guide to organisational and human factors*. Abingdon: Gower.

Graham, G. (2005). Community networking as radical practice. *The Journal of Community Informatics, 1*(3), 4-12.

Hendrick, H.W., & Kleiner, B.M. (2001). *Macroergonomics: An introduction to work system design*. Santa Monica, CA: Human Factors and Ergonomics Society.

Hyrkkänen, U. (2006). Analysis of work load factors and well-being in mobile work. In M. Vartiainen (Ed.), *Workspace methodologies–studying communication, collaboration and workscapes* (Report 2006/3*)* (pp. 63-79). Espoo: Helsinki University of Technology, Laboratory of Work Psychology and Leadership.

Kettunen, J. (2006). Strategies for the cooperation of educational institutions and companies in mechanical engineering, *Journal of Educational Management, 20*(1), 19-28.

Kettunen, J. (2007). Strategies for the cooperation of higher education institutions in ICT. In H. Rahman (Ed.), *Information and communication technologies for economic and regional developments* (pp. 22-38). Hershey, PA: Idea Group Publishing.

Kettunen, J., & Kantola, I. (2005). Management information system based on the Balanced Scorecard, *Campus-Wide Information Systems, 22*(5), 263-274.

Kettunen, J., & Kantola, M. (2006). Strategies for virtual learning and e-entrepreneurship, In F. Zhao (Ed.), *Entrepreneurship and innovations in e-business: An integrative perspective* (pp. 107-123). Hershey, PA: Idea Group.

Kim, Y. J., Chaudhury, A., & Rao, H. R. (2002). A knowledge management perspective to evaluation of enterprise information portals. *Knowledge and Process Management, 9*(2), 57-71.

Kleiner, B.M. (2006). Macroergonomics: Analysis and design of work systems. *Applied Ergonomics, 37*(1), 81-89.

Pheasant, S. (1996). *Bodyspace, anthropometry, ergonomics and the design of work*. London: Taylor & Francis.

Porter, M. (1998). *On competition*. Boston: Harvard Business School Press.

Richter, P., Meyer, J., & Sommer, F. (2006). Well-being and stress in mobile and virtual work. In J.H.E. Andriessen & M. Vartiainen (Eds.), *Mobile virtual work: A new paradigm?* (pp. 13-44). Heidelber: Springer-Verlag.

Steinberg, A. (2006). Exploring rhizomic becomings in postcom crash networks. In F. Zhao (Ed.), *Entrepreneurship and innovations in e-business: An integrative perspective* (pp. 18-40). Hershey, PA: Idea Group.

Ståhle, P., & Grönroos, M. (2000). *Dynamic intellectual capital. Knowledge management in theory and practice.* Vantaa: WSOY.

Ståhle, P., & Hong, J. (2002). Managing dynamic intellectual capital in world wide fast changing industries. *Journal of Knowledge Management, 6*(2), 177-189.

Ståhle, P., Ståhle, S., & Pöyhönen, A. (2003). *Analyzing an organization's dynamic intellectual capital. System based theory and application*. Lappeenranta University of Technology.

Takeuchi, H., & Nonaka, I. (2004). *Hitotsubashi on knowledge management.* Singapore: John Wiley & Sons.

Vartiainen, M. (2006). Mobile virtual work, concepts, outcomes and challenges. In J.H.E. Andriessen & M. Vartiainen (Eds.), *Mobile virtual work: A new paradigm?* (pp. 13-44). Heidelber: Springer-Verlag.

Wilson, J.R. (2000). Fundamentals of ergonomics in theory and practice. *Applied Ergonomics, 31*(6), 557-567.

KEY TERMS

Cluster: Clusters are geographic concentrations of interconnected enterprises, specialised suppliers, service providers, firms in related industries, and associated institutions in particular fields that compete but also cooperate.

Human-Centred Approach: This is an approach to human-machine function and task allocation that first considers the capabilities and limitations of the human and whether the function or task justifies the use of a human.

Linux: Linux referes to Unix-like computer systems which have a Linux kernel. The source code of Linux is available to anyone to use, modify and redistribute freely. The Linux was originally developed by Linus Thorwalds.

Macroergonomics: The subdiscipline of ergonomics that focuses on the design of the overall work system. Conceptually, a top-down sociotechnical systems approach to the design of work systems and carry through of the overall work system design characteristics to the design of human-job, human-machine and human-software interfaces to ensure that the entire work system is fully harmonised.

Microergonomics: Those aspects of ergonomics that focus on the design of interfaces between the individual and other system elements, including human-job, hu-

man-machine, human-software and human-environment interfaces.

WiFi: Wireless Fidelity was originally a brand licensed by the WiFi Alliance. It describes the underlying technology of wireless local area networks based on the IEEE 802.11 specifications. It was developed for use in mobile computing devices, such as laptops, in local area networks.

WiMAX: Worldwide Interoperability of Microwave Access promotes conformance and interoperability of the IEEE 802.16 standard. WiMAX is a certification mark given to equipment that meets the certain conformity and interoperability tests.

WLAN: Wireless local area network is used to link two or more computers without using wires. WLAN uses spread-sprectrum technology based on radio waves to enable communication in a limited area.

Computational Biology

Andrew LaBrunda
GTA, Guam

Michelle LaBrunda
Cabrini Medical Center, USA

INTRODUCTION

It is impossible to pinpoint the exact moment at which computational biology became a discipline of its own, but one could say that it was in 1997 when the society of computational biology was formed. Regardless of its exact birthday, the research community has rapidly adopted computational biology and its applications are being vigorously explored.

The study and application of medicine is a dynamic challenge. Changes in medicine usually take place as a result of new knowledge acquired through observation and experimentation. When a tamping rod 1-inch thick went through Phineas Gage's head in 1848, his survival gave the medical field an unusual opportunity to observe behavior of a person missing their prefrontal cortex. This observation lead to the short-lived psychosurgical procedure known as a lobotomy, which attempted to change a person's behavior by separating two portions of a person's brain (Pols, 2001). Countless observations, experiments and mistakes represent how almost all medical knowledge has been acquired.

The relatively new field of computational biology offers a nontraditional approach to contribute to the medical body of knowledge. Computational biology is a new field combining biology, computer science, and mathematics to solve problems that are unworkable with traditional biological techniques. It includes traditional areas such as systems biology, molecular biology, biochemistry, biophysics, statistics, and computer science, as well as recently developed disciplines including bioinformatics and computational genomics. Algorithms, which are able to closely model biological behavior, validate the medical understanding of the observed processes and can be used to model scenarios that might not be able to be physically reproduced.

The goal of computational biology is to use mathematics and computer science to model biological systems on the molecular level. Instead of taking on large complex systems, computational biology is starting small, literally. Modeling problems in molecular biology and biochemistry is a far less daunting task. At a microscopic level, patient's characteristics drop out of the equation and all information behavior affecting is known. This creates a deterministic model which, given the same input, will always produce the same output. Some of the major subdisciplines of computational biology are computational genomics, systems biology, protein structure prediction, and evolutionary biology, all of which model microscopic structures.

COMPUTATIONAL GENOMICS

An organism's heredity is stored as deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Each of the storage methods contains a linear chain of finite elements called bases. In the case of DNA the chain is composed of four bases, adenine, cytosine, guanine, and thymine, and RNA is composed of four bases, adenine, cytosine, guanine, and uracil. This information can be representing in a computer through a series of linked-lists or arrays. The order of bases in DNA/RNA is 99.9% the same between members of the same species and genetic sequencing is the way of determining the order of the bases. Certain regions of DNA, called genes, are used by the body to create proteins. These proteins are used to construct and maintain the organism. Genes can be thought of as blueprint instructions for how to make each unique person. There are long stretches of DNA between the genes the function of which is not well understood. The sum of all the genetic information about an organism is called a genome. When a computer is applied to deciphering an organism's genome, this is known is computational genomics. Computational genomics are used to better understand and compare sequences, identify related organisms, and measure biodiversity.

There are two major challenges in genomic studies. The first is genetic sequencing, (determining the order of the bases that make a strand of genetic material) and the second is localizing the genes within the genome. Sequence comparison is probably the most useful computational tool for molecular biologists. Repositories containing hundreds of genomes have been established and are available for public access. A biologist now has the ability to compare a unique sequence of DNA with the already known genetic sequences from this massive repository. Prior to the application of computers to the problems genomics scientist had to manually attempt to align sequences using ill suited tools such as word processors.

By identifying the genome for several related organisms it is possible to compare the related species and see how genetic mutation could allow one organism to evolve while leaving the other species unchanged. This type of analysis provides scientists with a more accurate descriptive tool to identify differences between organisms rather than relying on physical taxonomy. It also gives scientist an opportunity to see how a species' genetic information changes over time. If a large number of species have a genetic mutation which increases their successfulness, this observation may indicate an environmental change. As an example, if three unrelated species in different parts of the world genetically mutate to grow longer or thicker hair, one could hypothesize that an ice age was beginning.

One of the most important modern molecular genetic advances is the sequencing of specific genes and computational biology is becoming of increasing importance in sequencing studies. In the future, medications may be tailor made to the needs of each individual based on their specific genetic makeup. One of the hurdles that must be crossed to reach this point is a cost effective and accurate computer sequencing technique (Mitchell & Mitchell, 2007). Standardizing the reporting system of gene sequencing would minimize error and give researches a common template from which data can be extracted.

Now, computers have the ability to perform searches using pair-wise or fuzzy logic algorithms. Fuzzy logic allows for the identification of nonintuitive relationships. Identification of such relationships can provide a great deal of insight. The sequence of a gene not only encodes genetic information, but also give clues as to the function of the gene (Gibas & Jambeck, 2001, pp 13-14). Studies are actively being done using fuzzy systems to model genetic processes (Ishibuchi & Nojima, 2006).

NEURAL NETWORKING AND FUZZY LOGIC

A fuzzy system is a control system utilizing a nondiscrete mathematical system to form loosely coupled decision making data structures. Fuzzy logic has many different applications. The most widely accepted solution for pattern matching is that of Neural Networking (NN). Neural networks are simplistic software models of brain function at a cellular level. The nervous system is composed of neurons, which receive input through dendrites and transmit output through axons, all of which communicate through synapses (small spaces between nervous system cells across which chemical messages are transmitted chemically). Hypothetical software models of neural processes can be applied to emulate some of the pattern matching abilities of the brain. For example, a series of lotteries forming words can be fed into a NN system. The NN first "learns" by observing the probability that a sequence of letters will occur will occur. As the NN processes additional input, nonlinear relationships emerge where new data uses a similar set of nodes, representing neurons, during processing. Fuzzy systems accept inputs between 0 and 1 representing a continuum of trueness. This is different from classical discrete computational systems, which only allow inputs of false (0) and true (1). Fuzzy neural-based systems try to maximize accuracy while minimizing complexity.

PROTEIN FOLDING

Proteins are molecules which comprise many of the structural and functional components of living things. They are made by stringing together amino acids. All of the proteins in the human body are made from combination of various quantities of only 20 different amino acids. As proteins are created (by ribosomes) they fold into three-dimensional structures and it is the 3-dimensional interaction of the amino acids that is important. Sometimes a protein may be composed of multiple three dimensional structures interacting to perform a specific function in the body. It is the three-dimensional structure of a protein that determines its function. Many diseases occur when just one amino acid is replaced for another, such as in sickle cell anemia.

Understanding the structure of proteins allows understanding of their function and may someday provide a new therapeutic target for people with disease caused by defects in protein structure.

Research is currently being done using computational biology to understand protein structure in the energy-producing portion of cells called mitochondria (Gabaldon, 2006). Mitochondria are unique in that they are all inherited maternally. Understanding mitochondrial protein structure helps us understand how normal cells function and allows for understanding of diseases caused by mitochondrial dysfunction. Classically, mitochondrial diseases are considered to be uncommon and include diseases such as Leber's hereditary optic neuropathy and Kearns-Sayre syndromes, but more recent evidence has linked mitochondrial function to common diseases such as type 2 diabetes, Parkinson's disease, atherosclerotic heart disease, stroke, Alzheimer's disease, and cancer (Baloyannis, 2006; Folmes & Lopaschuk, 2007). Understanding mitochondrial protein structure will likely provide insight into disease processes for a gamut of pathological processes.

Although no computer model is currently able to accurately replicate the processes of protein folding, the accuracy of models has been improving. Currently, a popular method of protein modeling is to look at the hydrophobic (water repelling) and hydrophilic (water loving) properties of each individual amino acid and modeling the protein structure on a two or three-dimensional grid lattice. The

hydrophobic properties of amino acids are thought to be the most important factor in determining the 3D structure of a protein and advances in lattice selection have allowed for improved accuracy in protein modeling (Böckenhauer & Bongartz, 2007). Surprisingly, accurate models can be generated using only hydrophobic/hydrophilic properties of amino acids when grid technology is used. Grid technology is a system by which multiple computers are linked together and simultaneously work to solve a problem.

Another modeling system has been developed in which "low energy" amino acids are brought together first followed by higher energy amino acids to create a 3-D model (Hockenmaier, Joshi, & Dill, 2007). This is not a perfect model, but represents one of the first steps in using computational biology to model protein folding. Before it is understood how proteins function in diseased states, research must first be done to understand how they function in healthy individuals.

COMPUTATIONAL EVOLUTION

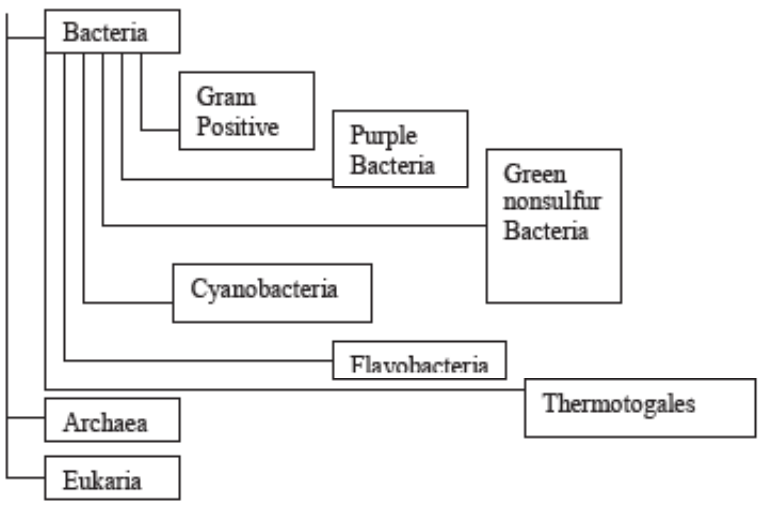
Computational evolution is a subspecialty of computational biology that uses computer technology and mathematics applied to evolutionary processes. Traditionally, evolutionary studies are retrospective. Scientists look backward at extinct species and study current species to try and understand the chronological development of the various species. The discovery of DNA has allowed a new level of understanding of these processes and more accurate classification of organisms. As computational genetics becomes more advanced, it will become clear not only physically where on a strand of DNA changes occurred creating new species, but trends will emerge. Computer mapping of these trends will allow predictions to be made about the evolution of future species (Banzhaf et al., 2006).

Scientific taxonomy emerged in the 18th century and has created a framework of nomenclature from which modern scientists work. This taxonomic classification is based largely on phenotypic relationships. In the modern era, one can extract genetic sequences and find relationships between organisms irrespective of the framework of the existing taxonomy. Often there is discordance between the historical taxonomic structure and the relationship determined through genetic analysis. For example, children are taught in grade school that there is a seven layered naming system for each organism (kingdom, phylum, class, order, family, genus, species). DNA studies provide evidence that there are three large groups of organisms, Bacteria, Archaea (an ancient form of bacteria) and Eukarya (everything else).

Sequencing the entire human genome is a landmark in scientific achievements; however, a future goal of the scientific community is to sequence the genome of every species starting with endangered species (Lesk, 2002, pp. 5-6). It will be understood why certain mutations are more likely than others. Information of this type will eventually allow a prospective study of evolution, meaning that predictions can be made based on where evolution will go in the future.

Maximum parsimony is a new concept in modeling evolution, recently formulated by Nakhleh and colleagues in 2005, and is one of the most popular methods for constructing a phylogenic tree (figure 1). This modeling technique considers as its basis that the edges of the tree should have the least amount of changes. The rational is that a new species has existed for fewer generations and therefore will exhibit less variety in any given set of genetic markers. A tree constructed

Figure 1. Example of a phylogenic tree



in this manner will model a system with the fewest number of mutations. While not without controversy, preliminary work applying maximum parsimony to phylogenic classification systems has yielded interesting results (Jin, Nakhleh, Snir, & Tuller, 2007). The parsimony trees have found genetic relationships between species which do not visibly appear similar but share key genetic similarities.

SYSTEMS BIOLOGY

Systems biology endeavors to understand the cause and effect of a biological system by studying and modeling the interaction of components. By modeling biological systems, one is better able to understand the complex biological relationships and fundamental behavior of living things. As simple models are designed, developed, tested, and validated, researchers will be able to model higher-level processes composed of model-proven subprocesses. This progress will lead to practical innovations in drug development and medical engineering.

One of the major challenges in studying systems biology is that the more complex the system being modeled the less likely the system will successfully represent actual behavior. Biological systems require physical models which are dynamic, continuous, and nondeterministic in nature. Unfortunately, these are the three most difficult characteristics to simulate. As an example, take the infection bacterial endocarditis, an infection of the inner lining of the heart. As the bacteria multiply the friction from blood flow could cause bacteria to dislodge from the heart valves and spread to other parts of the body (Durack, Lukes, Bright, & Duke, 1994). Correctly estimating when the bacteria will detach and how it might spread throughout the body is dependent on many factors. Among them would be the patients INR (International Normalized Ratio), detailed vein, artery, and blood flow characteristics as well as the bacteria's precise rate of reproduction. These nondiscrete values are extremely difficult to know with certainty given their propensity toward continuous change and difficulties in making precise measurements accurate enough to generating models.

At a microscopic level biological information is stored as DNA or RNA then translated into proteins. One of the most active areas of research currently undertaken in the field of systems biology is to model the processes involved in reading and executing the organic instructions that start with a piece of DNA and end in a protein. There are three key processes: 1) replication, the process by which DNA is copied, 2) transcription, which takes a DNA and produces RNA, and 3) translation which takes an RNA and produces a protein.

Replication is seen in cell division and is responsible for creating new copies of cells as old cells die and need to be replaced. For example, making new red blood cells or replacing skin cells. During replication, DNA untwists and the helix separates into two separate strands of DNA. Through a complex set of reactions, each half rebuilds their missing chemical complements, typically with few errors. Of the three processes, this one is the easiest to model.

Transcription uses DNA as a blueprint for creating RNA. Due to the different chemical makeup of RNA, it can exist as a single 3D strand rather than being restricted to the double helix form, as is DNA. Modeling the transcription process and the ensuing behavior of RNA has yet to be achieved. Not only is its transcription process difficult to model, but the type of RNA produced affects the function and behavior. There are three types of RNA that can be produced: messenger, transfer, and ribosomal. Each type of RNA has its own function and modeling challenges.

Translation is the process by which a protein is produced from a strand of RNA. The protein produced as a result of this process can be model to represent a linear chain amino acids as described in detail earlier. The physical 3-D structure of the protein is just as important to the proteins behavior as the chemical composition (Gibas & Jambeck 2001, pp 26-29). A great deal of research is ongoing to understand the physicochemical nature of protein.

Another challenge in systems biology is compiling the ever-growing amount of information into a usable format. New advances are made daily in protein structure, gene mapping and numerous other fields. As data is rapidly generated, the problem becomes how to organize, analyze and utilize the information. One of the goals of systems biology is to organize and catalog new information as it becomes available. These organizational systems must be flexible enough to absorb new types of information and integrate them with the old, but organized enough to separate relevant from irrelevant information. Grids technology is one of the computer resources being applied to systems biology (Strizh, Joutehkov, Tverdokhlebov, & Golitsyn, 2007).

Grid technology is a computer modeling system that allows higher throughput computing utilizing networked computers; essentially making all computers on a network one large distributed machine. This system allows large data sets to be analyzed by breaking them down into smaller ones and running them simultaneously on a number of separate networked computers. For this type of processing to be effective, problems must be discrete, uncoupled, and computable. Although grid computing has been around for years, its application in the field of medicine is only begging to be realized.

CONCLUSION

The field of computational biology is in its infancy. There are many problems that must be overcome before the benefits of the field can be tapped. Headway is being made in the modeling of complex interacting systems at the molecular level. As microsystems are better able to more accurately model biological processes, new tools will be developed to model higher-level systems using the micromodels as a foundation. The hope is one day to accurately model the entire body's biological processes.

Technologies and logic systems such as grid processing, fuzzy logic and maximum parsimony are finding patterns and relationships that couldn't be found through observation alone. The tools and technologies that are now being developed offer a new way to understand the structure of nature. Someday in the distant but foreseeable future, resource consuming experiments may move from the laboratory to the computer, allowing for optimal solutions to complex problems and novel therapeutic modalities.

REFERENCES

Baloyannis, S. (2006). Mitochondrial alterations in Alzheimer's disease. *Journal of Alzheimer's Disease, 9*(2), 119-126.

Banzhaf, W., Beslon, G., Christensen, S., Foster, J., Képès, F., Lefort, V., et al. (2006). From artificial evolution to computational evolution: A research agenda. *Nature Reviews, 7*, 729-734.

Böckenhauer, H., & Bongartz, D. (2007). Protein folding in the HP model on grid lattices with diagonals. *Discrete Applied Mathematics, 115*(2), 230-256.

Durack, D., Lukes, A., & Bright, D. (1994). Duke endocarditis service. New criteria for diagnosis of infective endocarditis: Utilization of specific echocardiographic findings. *The American Journal of Medicine, 96*, 200-209.

Folmes, C., & Lopaschuk, G. (2007). Role of malonyl-CoA in heart disease and the hypothalamic control of obesity. *Cardiovascular Research, 73*(2), 278-287.

Gabaldon, T. (2006). Computational approaches for the prediction of protein function in the mitochondrion. *American Journal of Physiology–Cell Physiology, 291*, C1121-C1128.

Gibas, C., & Jambeck, P. (2001). *Developing bioinformatics computer skills*. Sebastopol, CA: O'Reilly & Associates.

Hockenmaier, J., Joshi, A., & Dill, K. (2007). Routes are trees: The parsing perspective on protein folding. *Proteins: Structure, function and genetics, 66*(1), 1-15.

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. (2007). Inferring phylogenetic networks by the maximum parsimony criterion: A case study. *Molecular Biology and Evolution, 24*(1), 324-337.

Lesk, A. (2002). *Introduction to bioinformatics*. New York: Oxford University Press.

Mitchell, D., & Mitchell, J. (2007). Status of clinical gene sequencing data reporting and associated risks for information loss. *Journal of Biomedical Informatics, 40*(1), 47-54.

Nakhleh, L., Jin, G., Zhao, F., & Mellor-Crummey, J. (2005, August). Reconstructing phylogenetic networks using maximum parsimony. In *Proceedings of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005)*, (pp 93-102).

Nakhleh, L., Ruths, D., & Wang, L. (2005). RIATA-HGT: A fast and accurate heuristic for reconstructing horizontal gene transfer. In *Proceedings of the Eleventh International Computing and Combinatorics Conference (COCOON 05)* (pp. 84-93). Berlin, Heidelberg: Springer-Verlag.

Pols, H. (2001). An odd kind of fame: Stories of Phineas Gage. *Journal of the History of Medicine and Allied Sciences, 56*(2), 192-194.

Strizh, I., Joutchkov, A., Tverdokhlebov, N., & Golitsyn, S. (2007). Systems biology and grid technologies: Challenges for understanding complex cell signaling networks. *Future Generation Computer Systems, 23*, 428-34.

KEY TERMS

Computational Biology: A new field combining biology, computer science, mathematics and physics to model and understand complex biological processes at the molecular level.

Computational Evolution (Artificial Evolution): The application of computational and mathematical techniques to retrospectively and prospectively model evolutionary processes.

Computational Genetics: The application of computational biology to genetics.

DNA (Deoxyribonucleic Acid): Comprises the genetic material of humans and most other organisms. It can be thought of as the blueprint to create a unique organism.

Fuzzy Logic: Fuzzy logic breaks input into variables and assigns each input a probability of being correct on a scale of 0 to 1 with 0 being false This is different from classical

discrete computational systems which only allow inputs of false (0) and true (1).

Grid System: Utilization of networked computers to solve problems involving data sets too large to be handled by one computer.

Phylogenic Tree: A diagram illustrating how various species are interconnected.

Translation: The process by which a protein is produced from a strand of mRNA.

Computer Attitude and Anxiety

Pieter Blignaut
University of the Free State, South Africa

Andries Burger
University of the Free State, South Africa

Theo McDonald
University of the Free State, South Africa

Janse Tolmie
University of the Free State, South Africa

INTRODUCTION

Computers in the workplace are a given. Although the advantages of computers are well-known and proven, many people still try to avoid using them. It is extremely important to find out which factors influence the success of end-user computing. What are the reasons that some people excel on a computer while others have problems and even build up a resistance toward the use of computers?

This chapter provides a literature-based overview of computer attitude and computer anxiety as factors that influence a user's resistance, commitment, and achievement. A graphic model, according to which the interactions between computer attitude and anxiety, their causes, indicators, and impacts may be understood, is proposed. It is put forth that external strategies to deal with anxiety and a negative attitude are imperative to break down a snowballing effect of cause and effect and to ensure effective end-user computing.

BACKGROUND

Computer Attitude

Gordon Allport (1935) defined the concept of attitude, in general, as follows: "An attitude is a mental and neural state of readiness, organized through *experience*, exerting a directive or dynamic *influence* upon the individual's *response* to all objects and situations with which it is related" (p. 810). In other words, attitude is determined by experience and impacts upon the individual's behavior.

A person's attitude toward a computer is influenced by a variety of aspects, e.g., the social issues relating to computer use (Popovich et al., 1987), computer liking, computer confidence, computer anxiety or comfort (Delcourt & Kinzie, 1993; Loyd & Gressard, 1984a), achievement (Bandalos & Benson, 1990), usefulness, and value (Francis-Pelton & Pelton, 1996).

Computer Anxiety

According to Henderson et al. (1995) anxiety is viewed as "a drive that motivates the organism to avoid the stimulus for anxiety" (p. 24). This implies that an individual will avoid the use of a computer in the presence of computer anxiety and if possible.

Kaplan and Sadock (1998) referred to anxiety as "a diffuse, unpleasant, vague sense of apprehension, often accompanied by autonomic symptoms" (p. 581). Specifically, computer anxiety involves an array of emotional reactions, including fear, apprehension, uneasiness, and distrust of computer technology in general (Negron, 1995; Rohner & Simonson, 1981).

Computer anxiety is also influenced by a variety of aspects, e.g., general anxiety and confidence (Harrison & Rainer, 1992), computer liking (Chu & Spires, 1991; Loyd & Gressard, 1984b), impact of computers on society (Raub, 1981), equipment-related anxiety (Marcoulides, 1989), comfort and value (Violato et al., 1989), and corporate pressure.

The Relationship between Computer Attitude and Computer Anxiety

Computer anxiety is often included as a component of attitude (Delcourt & Kinzie, 1993; Loyd & Gressard, 1984a). Jawahar and Elango (2001) reported, however, that previous studies used the concepts of computer anxiety and negative attitudes toward computers interchangeably. Computer anxiety is, however, not solely responsible for a negative attitude. A person can have a negative attitude toward computers even though he or she is not overly anxious about

using them. This may be because of a negative experience, e.g., an apologizing clerk blaming an erroneous account statement on the computer.

Furthermore, attitude allows for both a negative and a positive grading, whereas anxiety is, by definition, either negative or absent.

MAIN THRUST OF THE CHAPTER: A MODEL FOR INTERACTION

In order to indicate the various influences on the mental states of computer attitude and computer anxiety and the effect they have on a user's ability to execute computer-related tasks effectively, a model for interaction was developed (Figure 1).

The model shows interaction on three levels of abstraction. The right-hand column resembles a typical flow diagram but with an adapted convention. It shows the sequence of mental and operational events when a user is confronted with a task to be done on the computer. The diamond symbols do not represent conscious decisions but rather indicate general user behavior as determined by the user's current levels of computer attitude, computer anxiety, knowledge, and pressure experienced.

As an example of how to read the flow diagram, consider a user who has to perform a computer task. If his or her level of computer anxiety is not above a specific critical level (*D1*), he or she has a positive attitude toward computer tasks (*D2*). If he or she knows how to perform the task (*D4*), he or she will do the task (*P2*). If the user's knowledge is inadequate, this person will go through a process of learning (*P1*) until he or she can do the task. If the anxiety level is high (*D1*), the user will only use the computer if forced to do so (*D3*), or else he or she will opt out of the task or do it without a computer (*P3*).

The middle column in Figure 1 indicates the user's current levels of computer anxiety and computer attitude. The influence that computer anxiety and computer attitude have on each other as well as their influence on user behavior and the processes of learning and task execution is indicated with curved arrows (*E5–E11*). It is also clear that task execution (computer experience, *P2*) impacts computer attitude and anxiety in return (*E12–E13*).

The left-hand column shows the external processes and factors that influence a user's levels of computer attitude and anxiety.

Further discussion in this chapter serves to substantiate the claimed influences from the literature.

Factors that Determine Computer Attitude

Several studies have been undertaken to explore potential factors associated with a positive attitude toward computers (Brodt & Stronge, 1986; Scarpa et al., 1992; Sultana, 1990; Schwirian et al., 1989; Bongartz, 1988; Burkes, 1991). Some of the factors that were considered were level of education, years of experience in the work environment, computer experience, age, gender, and job title (*E3* and *E13*). The only factor that was repeatedly, although not consistently, found to have a positive effect on computer attitude, was computer experience (*E13*).

Causes of Computer Anxiety

According to Torkzadeh and Angulo (1992), there are three perspectives of computer anxiety: psychological (*E1*), sociological (*E1*), and operational (*E1* and *E12*). From a psychological perspective, users may fear that they will damage the computer, feel threatened when having to ask younger workers for help, or feel that they are losing control because computers are perceived as a threat to one's power and influence. From a sociological perspective, people have a need for social contact with other people, and because computers can change existing social patterns, they find the situation unbearable. People may also have a fear of computers replacing them. From an operational point of view, people want to avoid embarrassment connected with their inability to type or to use the keyboard. An initially confident user might be disillusioned with the complexity and sophistication of computer systems and procedures after a first experience (*E12*).

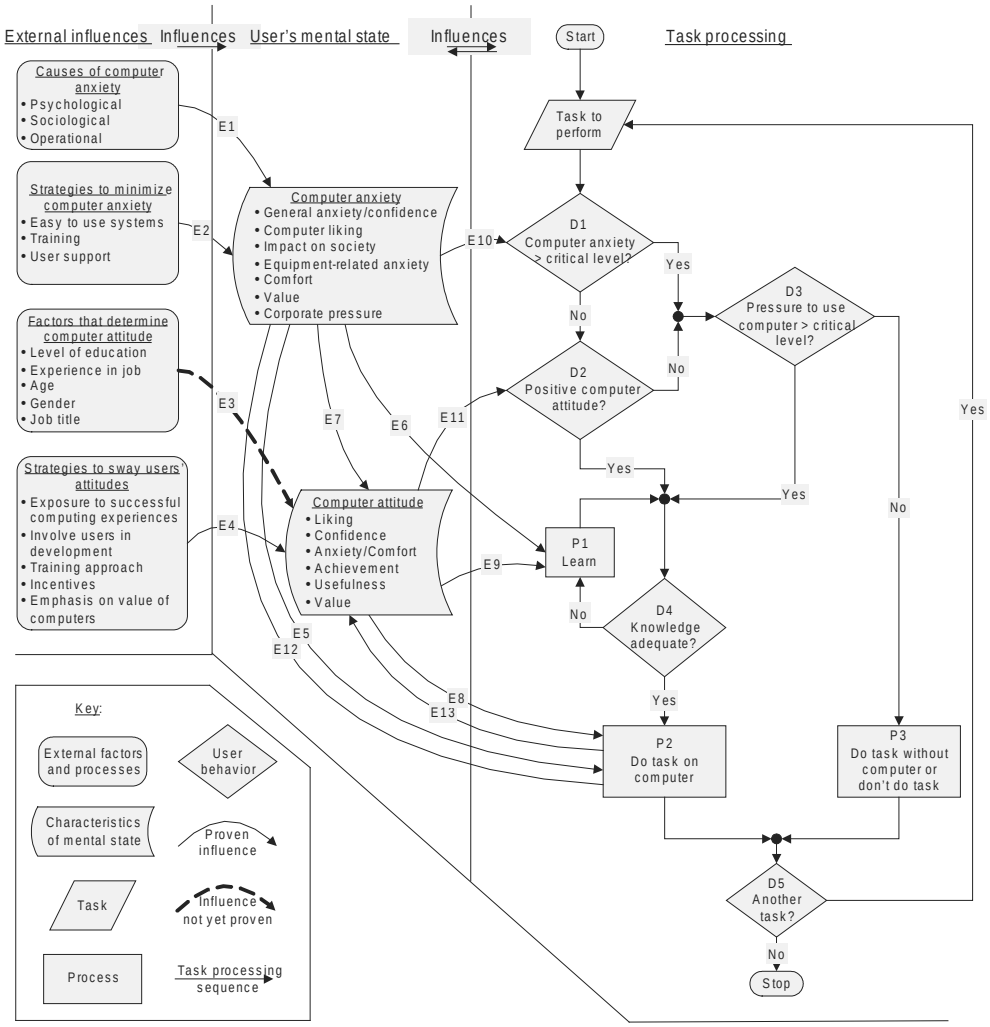
Effects of a Positive Attitude

According to Ngin et al. (1993), individuals with work excitement express creativity, receptivity to learning, and have the ability to see opportunity in everyday situations. Positive attitudes enhance the learning process (Shneiderman, 1980) (*E9*), specifically the motivation to learn and the ability to retain information in a given situation (Jawahar & Elango, 2001).

A negative attitude may lead to computer resistance (Sheiderman, 1980) (*D2*, *D3*, and *P3*), a phenomenon that can be found among experienced as well as inexperienced users (Negron, 1995). A negative attitude may even lead to defamation or sabotage of computer technology (Gibson & Rose, 1986).

A person's attitude toward computers and related technology could determine his or her performance with the technology and the satisfaction he or she draws from the

Figure 1. Model for interaction



experience (*E8*), although contradictory results are reported in the literature. Nickell and Pinto (1986) as well as Jawahar and Elango (2001) found a positive correlation between scores on a computer attitude scale and the final course grades of students enrolled in an introductory computer class. Other studies on students report an inverse relation between attitudes and performance (Hayek & Stephens, 1989; Marcoulides, 1988; Mawhinney & Saraswat, 1991), while still other researchers report no relationship between these two constructs (O'Quin et al., 1987; Kernan & Howard, 1990; Szajna & Mackay, 1995).

It was indicated above that scales to measure computer attitude differ with regard to the components they include. Jawahar and Elango (2001) also indicated that some tests measure attitudes toward working with computers, while others have components measuring general attitudes toward computers. These might be possible reasons for the contradictory results regarding the impact of computer attitude on computing performance. One cannot expect to obtain consistent results if the instruments used differ substantially from one another. This is one area in this field of research where agreement and subsequent standardization are yet to be achieved.

Impact of Computer Anxiety

Anxiety affects people's thinking, perception, and learning (Kaplan & Sadock, 1998). It also produces confusion and distortions of perception relating to time, space, people, and the meanings of events. These distortions usually have a negative effect on learning ability by lowering concen-

tration, reducing recall ability, and impairing the ability to make associations (*E6*).

Specifically, computer anxiety can be recognized by a fear expressed regarding present or future interactions with computers or computer-related technology, negative global attitudes about computers (*E7*), or self-critical internal dialogues during computer interactions (Rosen & Weil, 1995).

People with computer anxiety experience the incidence of physical, phobic symptoms, such as stomach cramps and cold sweats, as well as psychological symptoms, such as a resistance to use and a negative attitude toward a system (Shneiderman, 1980).

Computer anxiety seems to be a good predictor of computer achievement (*E5*). Marcoulides (1988) found that computer anxiety influenced the effectiveness with which college students could utilize the computer. Rosen et al. (1987) also found that computer-anxious undergraduate students enrolled in computer-based courses were twice as likely to drop out of the course and received lower course grades than nonanxious students.

Although anxiety usually has a negative connotation, there appear to be optimal levels of anxiety that help people to function effectively (Higgins, 1989) and that make them more alert and aware of what is going on (Lugo & Hershey, 1981; Beck & Emery, 1985) (*E5*).

Strategies to Sway Users' Attitudes

Previously, it was indicated that a negative computer attitude might lead to computer resistance. It is, therefore, important to identify strategies to enhance or sway users' attitudes toward computer use. Possible strategies are as follows (*E4*):

- Expose the users to successful and positive computer experiences (Yaghmaie et al., 1998).
- Make users aware of the fact that their computer literacy would be an advantage in any future work environment (Emmet, 1988).
- Involve users actively in the implementation of computer systems (Barry & Gibbons, 1990).
- Emphasize the ease with which work can be accomplished using a computer function (Marasovic et al., 1997).
- In training, stress the fact that computer use leads to desired outcomes (Marasovic et al., 1997).
- Include in training programs activities aimed at increasing self-efficacy (Henderson et al., 1995).
- Introduce incentives in order to convince and motivate users (Sultana, 1990).
- Consider the user's attitude and motivation when system designers design the user interface (Galitz, 1997).

Strategies to Minimize Computer Anxiety

Computer anxiety is something that will not just disappear on its own, but it is something that has to be dealt with. It is considered to be a temporal emotional state rather than a permanent personality trait and, therefore, can be remedied through positive computing experiences (Cambre & Cook, 1987). Possible strategies are as follows (*E2*):

- Easy-to-use computer systems (Appelbaum & Primmer, 1990)
- Basic training for the most routine tasks (Cambre & Cook, 1987; Flaughler, 1986; Lewis, 1988)
- Advanced training, which is essential to keep users interested as they progress (Appelbaum & Primmer, 1990)
- User support in the form of manuals, consultants, user groups, and computer specialists (Appelbaum & Primmer, 1990)

FUTURE TRENDS

It was argued above that the variety with regard to components and focus of the scales used by researchers to measure computer attitudes might be responsible for the contradictory results regarding the impact of computer attitude on computing performance. If a measuring scale for both computer attitude and computer anxiety can be standardized, much of the current confusion that exists in the literature could be solved. For example, the definition of computer comfort as it is given under "Terms and Definitions" below is merely the opposite of that for computer anxiety. This means that a standardization of measuring scales would include a consolidation of terminology.

CONCLUSION

The use of a graphic model provides an efficient way of understanding the interaction between computer attitude and computer anxiety, together with their causes, indicators, impacts, and strategies to deal with them.

From this model, it can be deduced that there are two-way interactions between computer attitude and computer anxiety on the one hand and computer task execution on the other. A vicious circle can be established where anxiety inhibits performance (*E5*), which in turn, produces a negative attitude toward the system (*E13*) and further slows the process of learning to use the system (*E9*).

External strategies can be applied to reduce computer anxiety and sway users' attitudes toward the use of computers (*E1–E4*). This will then lead to better achievement (E5

and *E8*), which will, in turn, lead to self-confidence (*E12*) and appreciation of the value of computers in general (*E13*). Less anxiety and a positive attitude will support the learning process (*E6* and *E9*), which will eventually result in better achievement.

Because of the proven effect of computer experience on attitude, moderate pressure can be applied to ensure that users do not avoid the use of computers and, in this way, never get the opportunity to gain experience (*D3*). This pressure can be explicit by means of assignment or by applying measures to create intrinsic motivation so that the user drives him- or herself.

REFERENCES

Allport, G. W. (1935). Attitudes. In C. M. Murchison (Ed.), *Handbook of social psychology* (pp. 796–834). Worcester, MA: Clark University Press.

Appelbaum, S. H., & Primmer, B. (1990). An RHx for computer anxiety. *Personnel, 87*(9), 8–11.

Bandalos, D., & Benson, J. (1990). Testing the factor structure invariance of a computer attitude scale over two grouping conditions. *Educational and Psychological Measurement, 50*(1), 49–60.

Barry, C. T., & Gibbons, L. K. (1990). Information systems technology: Barriers and challenges to implementation. *Journal of Nursing Administration, 20*(2), 40–42.

Beck, A. T., & Emery, G. (1985). *Anxiety disorders and phobias: A cognitive perspective*. New York: Basic Books.

Bongartz, C. (1988). Computer-oriented patient care: A comparison of nurses' attitudes and perceptions. *Computers in Nursing, 6*, 204–210.

Brodt, A., & Stronge, J. H. (1986). Nurses' attitudes toward computerization in a mid-western community hospital. *Computers in Nursing, 4*, 82–86.

Burkes, M. (1991). Identifying and relating nurses' attitudes toward computer use. *Computers in Nursing, 9*, 190–201.

Cambre, M. A., & Cook, D. L. (1987). Measurement and remediation of computer anxiety. *Educational Technology, 27*(12), 15–20.

Chu, P. C., & Spires, E. C. (1991). Validating the computer anxiety rating scale: Effects of cognitive style and computer courses on anxiety. *Computers in Human Behavior, 7*, 7–21.

Delcourt, M. A. B., & Kinzie, M. B. (1993). Computer technologies in teacher education: The measurement of attitudes and self-efficacy. *Journal of Research and Development in Education, 27*(1), 35–41.

Emmet, A. (1988). Overcoming computer resistance. *Personal computing, 7*(12), 80–98.

Flaughler, P. O. (1986). Computer training for nursing personnel. *Computers in nursing, 4*(3), 105–108.

Francis-Pelton, L., & Pelton, T. W. (1996). *Building attitudes: How a technology course affects preservice teachers' attitudes about technology*. Retrieved from *http://www.math.byu.edu/~lfrancis/tim's-page/attitudesite.html*

Galitz, W. O. (1997). *The essential guide to user interface design: An introduction to GUI design principles and techniques*. New York: John Wiley & Sons.

Gibson, S. E., & Rose, M. A. (1986). Managing computer resistance. *Personal Computing, 4*(5), 201–204.

Harrison, A. W., & Rainer, K. (1992). An examination of the factor structures and concurrent validities for the computer attitude scale, the computer anxiety scale, and the computer self-efficacy scale. *Educational and Psychological Measurement, 52*, 735–745.

Hayek, L. M., & Stephens, L. (1989). Factors affecting computer anxiety in high school computer science students. *Journal of Computers in Mathematics and Science Teaching, 8*, 73–76.

Henderson, R. D., Deane, F. P., & Ward, M. (1995). Occupational differences in computer related anxiety: Implications for the implementation of a computerised patient management information system. *Behaviour and Information Technology, 14*(1), 23–31.

Higgins, D. L. (1989). Anxiety as a function of perception: A theory about anxiety and a procedure to reduce symptoms to manageable levels. In M. D. Yapko (Ed.), *Brief therapy approaches to treating anxiety and depression* (pp. 245–263). New York: Brunner/Mazel, Inc.

Jawahar, I. M., & Elango, B. (2001). The effect of attitudes, goal setting and self-efficacy on end user performance. *Journal of End User Computing, 13*(2), 40–45.

Kaplan, H. I., & Sadock, B. J. (1998). *Synopsis of psychiatry: Behavioural sciences/clinical psychiatry* (8th ed.). Baltimore, MD: Lippencott Williams & Wilkins.

Kernan, M. C., & Howard, G. S. (1990). Computer anxiety and computer attitudes: An investigation of construct and predictive validity issues. *Educational and Psychological Measurement, 50*, 681–690.

Lewis, L. H. (1988). Adults and computer anxiety: Facts or fiction? *Lifelong Learning, 11*(8), 6–12.

Loyd, B. H., & Gressard, C. (1984a). The effects of sex, age, and computer experience on computer attitudes. *Association for Educational Data Systems, 18*(2), 67–77.

Loyd, B. H., & Gressard, C. (1984b). The reliability and factorial validity of computer attitude scales. *Educational and Psychological Measurement, 44*, 501–505.

Lugo, J. O., & Hershey, G. L. (1981). *Living psychology*. New York: Macmillan.

Marasovic, C., Kenney, C., Elliott, D., & Sindhusake, D. (1997). Attitudes of Australian nurses toward implementation of a clinical information system. *Computers in Nursing, 15*(2), 91–98.

Marcoulides, G. A. (1988). The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research, 4*, 151–158.

Marcoulides, G. A. (1989). Measuring computer anxiety: The Computer Anxiety Scale. *Educational and Psychological Measurement, 49*, 733–739.

Mawhinney, C. H., & Saraswat, S. P. (1991). Personality type, computer anxiety, and student performance. *Journal of Computer Information Systems, 8*, 110–123.

Negron, J. A. (1995). The impact of computer anxiety and computer resistance on the use of computer technology by nurses. *Journal of Nursing Staff Development, 11*(3), 172–175.

Ngin, P., Simms, L., & Erbin-Roesemann, M. (1993). Work excitement among computer users in nursing. *Computers in Nursing, 3*, 127–133.

Nickell, G. S., & Pinto, J. N. (1986). The computer attitude scale. *Computers in Human Behavior, 2*, 301–306.

O'Quin, K., Kinsey, T. G., & Beery, D. (1987). Effectiveness of a microcomputer-training workshop for college professionals. *Computers in Human Behaviour, 3*, 85–94.

Popovich, P. M., Hyde, K. R., & Zakrajsek, T. (1987). The development of the attitudes toward computer usage scale. *Educational and Psychological Measurement, 47*, 261–269.

Raub, A. C. (1981). *Correlates of computer anxiety in college students*. Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia.

Rohner, D. J., & Simonson, M. R. (1981). *Development of an index of computer anxiety.* Paper presented at the annual convention of the Association of Educational Communications and Technology, Philadelphia, PA.

Rosen, L. D., Sears, D. C., & Weil, M. M. (1987). Computerphobia. *Behavior Research Methods, Instruments, and Computers, 19*, 167–179.

Rosen, L. D., & Weil, M. M. (1995). Adult and teenage use of consumer, business, and entertainment technology: Potholes on the information highway? *Journal of Consumer Affairs, 29*(1), 55–84.

Scarpa, R., Smeltzer, S. C., & Jasion, B. (1992). Attitudes of nurses toward computerisation: A replication. *Computers in Nursing, 10*, 72–80.

Schwirian, P., Malone, J. A., Stone, V. J., Nunley, B., & Francisco, T. (1989). Computers in nursing practice: A comparison of the attitudes of nurses and nursing students. *Computers in Nursing, 7*, 168–177.

Shneiderman, B. (1980). *Software psychology*. Cambridge, MA: Winthrop.

Sultana, N. (1990). Nurses' attitudes toward computerization in clinical practice. *Journal of Advanced Nursing, 15*, 696–702.

Szajna, B., & Mackay, J. M. (1995). Predictors of learning performance in a computer-user training environment: A path-analytic study. *International Journal of Human–Computer Interaction, 7*(2), 167–185.

Torkzadeh, G., & Angulo, I. E. (1992). The concept and correlates of computer anxiety. *Behavior and Information Technology, 11*(2), 99–108.

Violato, C., Marini, A., & Hunter, W. (1989). A confirmatory factor analysis of a four-factor model of attitudes toward computers: A study of preservice teachers. *Journal of Research on Computing in Education*, Winter, 199–213.

Yaghmaie, F., Jayasuriya, R., & Rawstorne, P. (1998). Computer experience and computer attitude: A model to predict the use of computerized information systems. *Human Computer Interaction*, 895–899.

KEY TERMS

Computer Anxiety: A diffuse, unpleasant, and vague sense of discomfort and apprehension when confronted by computer technology or people who talk about computers.

Computer Attitude: A complex mental state that affects a human's choice of action or behavior toward computers and computer-related tasks.

Computer Comfort: The user does not experience any suffering, anxiety, pain, etc., when using a computer.

Computer Confidence: The user is confident that he or she would be able to master a required skill to solve a particular problem using a computer, e.g., learning how to use a specific facility of an application program or learning a programming language.

Computer Liking: The use of a computer to solve problems is enjoyable, stimulating, and even addictive.

Self-Efficacy: Measure of a person's belief and confidence that he or she can perform a certain task.

Usefulness: The user is convinced that the use of computers in the workplace is an efficient and effective means to solve problems.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 495-501, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Computer Music Interface Evaluation

Dionysios Politis
Aristotle University of Thessaloniki, Greece

Ioannis Stamelos
Aristotle University of Thessaloniki, Greece

Dimitrios Margounakis
Aristotle University of Thessaloniki, Greece

INTRODUCTION

The old computing is about what computers can do, the new computing is about what people can do. Ben Schneiderman, HCI Researcher (1997)

One of the most intriguing fields of human-computer interaction (HCI) involves the communication aspects of computer music interfaces. Music is a rich communication medium, and computer music is the amalgam of interface science and musical praxis forming a dynamic subset of HCI.

There are structural similarities between the job of a music composer and that of a *user interface designer* (although their objectives may be different). While sound has been used in general purpose interfaces as an *object*, its use has been deteriorated at a primary level, that of a signal-processing approach. However, music composition and performance are highly abstract human activities involving a semantic and a symbolic mechanism of human intellectual activity.

This article analyzes the unique problems posed by the use of computers by composers and performers of music. It presents the HCI predicates involved in the chain of musical interaction with computer devices, commencing from the abstract part of symbolic composition, then coping with usability issues of the graphical user interfaces (GUIs) implemented for musical scripting, and concluding to a synthesis stage which produces digitized sounds that enhance or replace original analog audio signals. The evaluation of HCI elements for computer music under the prism of usability aims at the development of new graphical tools, new symbolic languages, and finally better user interfaces. The advance in technology on this area creates the demand for more qualitative user interfaces and more functional and flexible computer music devices. The peculiarities of computer music create new fields in HCI research concerning the design and the functionality of computer music systems.

BACKGROUND

Computer Music Interfaces

In the early stages of the microcomputer evolution, various protocols had been developed in order to achieve interconnection between computers and instruments. The milestone of computer music proved however to be the musical instrument digital interface (MIDI), which is a communications standard used for transmitting musical performance information (Aikin, 2003). It was developed in 1983 in response to the increasing sophistication, and corresponding complexity, of commercial electronic instruments, especially synthesizers. Therefore, MIDI is a protocol specifying how electronic musical instruments may be controlled remotely. In brief, MIDI is a very successful and inexpensive protocol that has reshaped the computer music landscape. However, it cannot overcome easily its representation limitations, especially on alternative music notations. The common music notation (CMN) scheme along with the MIDI specification is Western music oriented. The problem with CMN has been taken into account in several works: Although CMN is supposed to furnish a model for traditional music in a European style, it is not absolutely supposed that this model is also convenient or suitable for music coming from outside of Western traditions (East Asia, Middle East countries, etc.) (Bellini, Barthelemy, Nesi, & Zoia, 2004). As a result, they are not able to clearly depict alternate musical forms and traditions.

Almost all music recordings today utilize MIDI as a key enabling technology for recording music. In addition, MIDI is also used to control hardware including recording devices as well as live performance equipment such as stage lights and effects pedals. Lately, MIDI has exploded onto the scene with its adoption into mobile phones. MIDI is used to play back the ring tones of MIDI capable phones. MIDI is also used to provide game music in some video games.

MIDI is almost directly responsible for bringing an end to the "wall of synthesizers" phenomenon in 1970-1980s rock music concerts, when musical keyboard performers

were sometimes hidden behind banks of various instruments. Following the advent of MIDI, many synthesizers were released in rack-mount versions, enabling performers to control multiple instruments from a single keyboard. Another important effect of MIDI has been the development of hardware and computer-based sequencers, which can be used to record, edit, and playback performances.

A number of music file formats have been based on the MIDI bytestream. These formats are very compact; often a file of only 10 kilobytes and can produce a full minute of music.

MIDI, albeit the dominant, is not the most expandable and modular interface. Also, other interfaces like the Synthesis toolKit Instrument Network Interface (SKINI) physical modeling interfaces have appeared (Cook, 1996). These interfaces are purely computer software inventions and lack the hardware orientation of MIDI. However, they are more adaptive in expressing alternate musical forms and interfaces.

Computer Music Languages

An *audio programming language* is a programming language specifically targeted to sound and music production or synthesis. Such languages are: ABC, ChucK, CMix, CMusic, Common Lisp Music (CLM), CSound, Haskore, HMSL, Impromptu, jMusic, JSyn, Loco, designed to be for sound what Logo is for graphics, Max/MSP, Music I, Music-N, Nyquist, OpenMusic, Pure Data, Real-time CMix, Soundscrape, SuperCollider, Q-Audio. Each of thoses languages has its own features and objectives. For instance, JSyn is used by JAVA programmers and makes use of simple methods, which are written in C language, for real-time audio synthesis.

Score Writing and Notation Creation

This category of interfaces consists of state-of-the-art, easy-to-use GUIs that provide ways to create, enter, edit, hear, view, lay out, and ultimately print music in staff notation. Usually these programs have complete control over every aspect of music printing and publishing. Generally, they are perceived as mature products, satisfying the musician in the same sense that a good word processing system satisfies the author enough to shift from handwriting to electronic processing.

However, their expression format is staff based, and therefore they can satisfy users' needs as long as the CMN can satisfy the expression of the melody accurately. A typical interface for CMN composition is shown in Figure 1. Figures 2 and 3 show two applications for music notation: Sibellius and Quitar Pro.

There are several methods used to enter music data into notation editors and sequencers.

An attractive method for keyboard players is to enter music by playing it on a MIDI keyboard. Most commercial notation editors allow this method. Unfortunately, automatic identification of rhythms is difficult, so the user must carefully check all notes in order to correct errors. Furthermore, each voice must be entered separately.

Optical music recognition (OMR), the musical equivalent of optical character recognition (OCR), has been used in building some music collections. Unfortunately, OMR is less accurate than OCR, and the scanned music must be carefully checked for errors, a process that often requires a considerable length of time (Rossant & Bloch, 2004).

The most common music entry method is by handling a GUI, using a mouse. The problem is that a mouse provides two-dimensional data entry, with horizontal and vertical coordinates. Music notation, however, is inherently three dimensional, with the horizontal dimension indicating time of note onset, the vertical dimension indicating the frequency of the note, and the shape of the note indicating rhythm, or note duration.

Musical Interfaces for Alternative Music Systems

For Western music users, or for systems that have adapted to CMN, there seems to be little or no problem. However, the world of music is not unified. Especially in the East, we do have alternate musical interfaces which use different semantics. A classical example is that of Byzantine music. This kind of music, apart from having a significant diachrony, serves also as an intuitively alternate interface, since it uses the notification methodology of ancient Delta systems (Margounakis & Politis, 2005).

In Figure 4 ARION is presented, which is a prototypal visual client, the first of its kind, that has served for composition with Ancient Greek music semantics (Politis, Vandikas, & Margounakis, 2005). ARION uses real-time physical modeling, voice reproduction techniques and provides ethnomusicology with an easy-to-use and functional interface for notation-based Ancient Greek music synthesis.

Recording Systems and Production Systems

This category of products, undoubtedly the flagship of the computer music industry, produces complete professional music recording systems. Usually they combine high resolution MIDI recording channels with audio recording in either 16- or 24-bit formats. This way they offer state-of-the-art multi-track, digital input capabilities and thus simulate and gradually replace classical analog recording studios in the meanwhile having the advantage of inherent communication with digital instruments. Although the hardware and

Figure 1. Entry level score writer, cakewalk's notation handling module

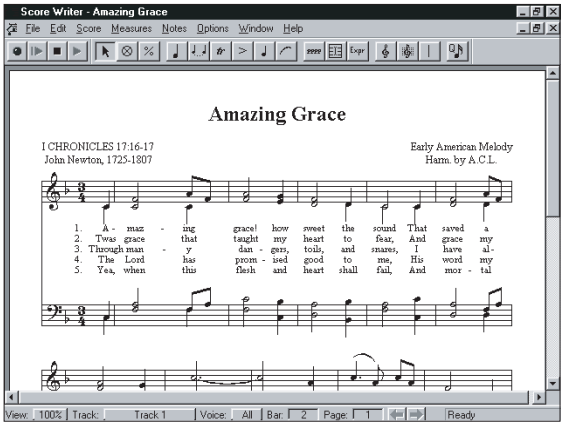
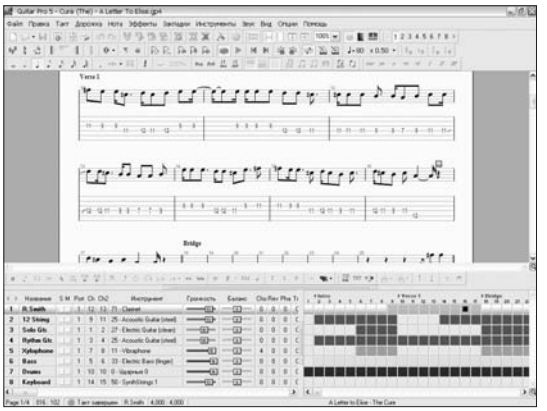


Figure 2. SIBELIUS *is a complete software for writing, playing, printing, and publishing music notation.*

Figure 3. Guitar Pro is a multi-track tablature editor for guitar, banjo, and bass.



the low-level capabilities of these systems rely heavily on the MIDI specification and therefore do not offer substantial improvements over the performance limitations of MIDI, their high-level perception, that is, their GUIs challenge the computer music community for the invention of new conceptually and functionally composing and synthesizing schemas. To a great extent this has been solved with the various plug-ins that accompany and complement the basic systems.

An important attribute of music recording and production software systems is the effective visualization of computer music predicates. Several plug-ins allow this visualization in a spatial and semantically polymorphic way. Another key element of these systems is that they also carry out tasks in real time. This fact makes them straightly comparable to hardware synthesizers. Aikin (2003) compares the advantages of computer-based instruments (software synths) to dedicated hardware (hardware synthesizers) and comes to the conclusion that although music software is an evolving dynamic scene, hardware instruments will always play a vital role in music synthesis.

In musical composition, the strength of these systems lies on the symbolic processing of CMN. However, these systems are not merely notation interfaces that perform via the MIDI specification. Good performance criteria should also apply.

In this field, the advances in HCI challenge for innovations beyond the limited, file-based, single-data-type applications. The momentum is towards models supporting richer data types, visualization paradigms, and distributed storage such as the model behind the rapidly evolving World Wide Web.

The multi-dimensionality of musical data begs for higher dimensions of control and representation impossible with the current "paper on a desktop" metaphor (Freed, 1995) (Figures 5, 6).

Waveform Processing Systems

This category of products, is related with the signal processing aspect of computer music. Although these systems allow the production of virtually *any* sound, they are not short-term composition systems, and they will not be considered.

PROBLEM FORMULATION

The use of computer music interfaces aims at producing melodic pieces. The instrument used in this case is a computer program, perhaps in conjunction with a keyboard hardware interface communicating via the MIDI-IN and OUT ports. Producing melodic lines is a matter of inspiration and not an arbitrary or disciplined procedure. In terms of HCI it means that the computer program used must have

Figure 4. The graphical user interface of ARION

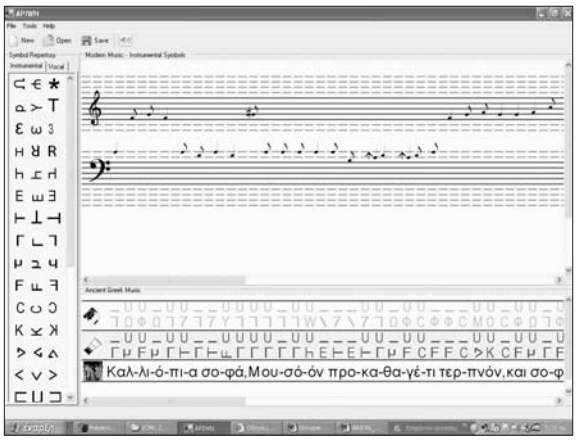


Figure 5. Typical recording systems interfaces implementing the recording console metaphor

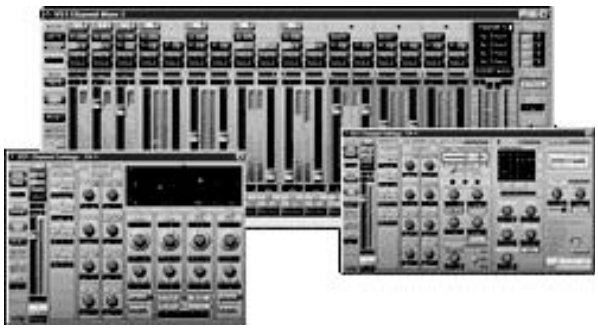


Figure 6. Orion Platinum 6 introduces the impulse response processor, an effect for applying reverbera-tion from captured impulse responses.

functionality and *usability* features that enable the user to record in symbolic form the music he has conceived. Usually, five criteria are used in order to evaluate the usability of an interface according to the ISO/DIS 9241-11 directive (ISO DIS 9241-11, 1996):

1. Learnability for the use of the new system. Five principles that affect learnability are: predictability, synthesizability, familiarity, generalizability, and consistency (Dix, Finlay, Abowd, & Beale, 2004).
2. Effectiveness, that is, the extent to which the intended goals of musical synthesis and composition are achieved. The effectiveness with which users employ a product to carry out a task is defined as a function of two components, the quantity of the task attempted by the users, and the quality of the goals they achieve.

   Effectiveness |= *f*(quantity, quality)

3. Efficiency, when used by experienced and trained users, that is, the amount of resources that have to be expended to achieve the intended goals. This criterion is more procedural than quantitative in computer music. In engineering, the term *efficiency* is uncontentiously understood as the ratio of useful energy output to energy input.
4. Satisfaction, in the sense of the extent to which the user finds the use of the product acceptable and desirable.
5. Capability to use the system from users not familiar with its musical categories and predicates after a long time.

In order to evaluate the performance of computer music systems on alternate musical interfaces a heuristic evaluation will be performed. According to Nielsen (1994) heuristic evaluation is a usability engineering method for finding the usability problems in a user interface design so that they can be attended to as part of an iterative design process. Heuristic evaluation involves having a small set of evaluators, experts in their field, examining the interface and judging its compliance with recognized usability principles (the "heuristics"). For each category of musical interface products, evaluation takes place according to the previously mentioned criteria.

RESULTS

The evaluation of specific computer music interfaces is based on the previously mentioned usability criteria. These criteria however are adjusted to the specific communication content of each interface. The evaluation is calibrated with

Table 1. Usability evaluation of computer music protocols on their ability to simulate alternate musical sounds

Protocol	Simulation	Interconnection	Expandability	Acceptability	Learnability
MIDI (Industry standard)	+/-	+	-	+	+
PM	+	+	+	-	+
Extended MIDI	+	+	+/-	+/-	+

Table 2. Usability evaluation of computer music software modules on their ability to track down alternate musical predicates and to produce adequate sounds

Product group	Alternate symbolic representation	learnability	Simulation effectiveness	Expandability	Efficiency	Satisfaction
CMN based	-	+/-	-	+/-	-	-
System based	-	+/-	+/-	+	+/-	+/-
Alternate	+	+	+	-	+	+

the following ratings of confidence whether a task can be performed:

- : weak confidence, +/- : plausible, + : strong confidence.

Interfaces

The usability criteria for the category of computer music protocols and specifications has to do mainly with the ability to simulate a broad range of musical data, to present them in an acceptable way, and to expand to alternative musical forms. An evaluation of some schemas is shown in Table 1.

The term *simulation* describes the ability to render musical sounds close to the real-time performance data.

The term *interconnection* implies the ability to communicate with other digital musical devices. The term *expandability* describes the ability to engulf alternate musical systems and events.

The term *acceptability* measures the propagation of the protocol to alternative musical traditions users. The term *learnability* implies how easily the users of a specific product learn to produce alternative musical sounds and predicates.

Score Writing and Production

In this combined category, the evaluation criteria are adjusted to the pool of computer music users attempting to compose not abstract CMN melodies but melodies which will be performed and propagated to listeners of alternate musical systems. Modern Greek Pop music has been taken into account.

Hardware incarnations of such systems were also considered; the basic criterion for their acceptance is the existence of a corresponding software module which can at least create notation or symbolic scripting of the performed music. For instance, if we have a keyboard performing Arab or Byzantine tunes, it is a prerequisite to have a software module that can write melodic lines according to this system. It is desirable but not obligatory for these systems to communicate.

The comparison of such systems is performed in Table 2. The well-known ISO/DIS 9241-11 (ISO DIS 9241-11, 1996) usability criteria are applied.

Some variations and extension of these criteria have to do with:

1. whether the system has room for symbolic representation of the alternate musical form;
2. whether the system is learnable for users expressed mainly in alternate forms and not in CMN;
3. whether the produced sound or the symbolic scripting of a melody are close to the alternate music predicates;
4. whether the system is modular and can cooperate with other computer music instruments and gadgets;
5. whether expert users of computer software and alternate music theory and practice can produce alternative music; and
6. whether the listeners of alternate music forms accept the audio result of the simulation.

FUTURE TRENDS

Virtual Music Environments

The concept of the virtual music environment (VME) is a generalization of a virtual music instrument, in which the virtual world itself acts as a source of multimodal (e.g.,

visual, audio, haptic) feedback, and at the same time, a place of interaction. Naturally, the two important issues in designing a VME are: (1) the display content and (2) the control interface. These two issues are in fact interrelated as music performance is seen as a closed loop system composed of the user and the VME. The control interface must be designed to be as natural and easy-to-use as possible for quick responses to the ongoing music, and the world around the user must contain the right information and convey it to the user in an intuitive manner for acceptable controllability. The display must also be "vivid" in the sense that it must leave the user with a strong musical impression so that one remembers the "essence" of the musical content. Several virtual music environments have been proposed (Broersen & Nijholt, 2002; Valbom & Macros, 2003).

Interaction with Computer Music Systems

In an effort for more natural interaction with the computer music systems, we need support for input devices with higher control bandwidth and dimensionality than the mouse that may lead to a faster, more natural, and more fluid style of interaction for certain tasks (Wu & Balakrishnan, 2003). There is also need to integrate new kinds of keyboards and a broader range of physical gestures and nonhuman control sources. Several works are related to new gesture input devices (Malik & Laszlo, 2004; Wilson, 2004; Wilson & Cutrell, 2005).

Many musicians find the interface (mouse, computer keyboard and/or synthesiser keyboard) less natural than the traditional pencil and manuscript, so alternatives are an active area of research. Pen-based systems for data entry are rapidly developing, driven by their popularity with users. As a consequence, pen-based systems for music tasks are designed (Phon-Amnuaisuk, 2004). A set of gestures for the pen entry of music was reported in 1996. One of the most important pen gestures is *Presto1* and *Presto2* by Presto.

Most electronic music controllers that have been created are based on existing acoustic instruments, such as the piano keyboard. Such electronic controllers have the obvious advantage of being used relatively easily by "traditionally" trained musicians. However, there is an emergence of whole new types of controllers and new ways of performing music (Masui, Tsukada, & Siio, 2004; Orio, Schnell, & Wanderley, 2001). Obviously, this also highlights the need for evaluation of these new types of devices (Wanderley & Orio, 2002).

CONCLUSION

Already several prototypal and research projects have been focusing on alternate music representation, authoring, and scripting. It is expected soon that the advances in GUI software engineering will enable the production of commercial products that can compete the more than a decade old Western music counterparts. Since the Western music interfaces have a very interesting evolution from the HCI point of view, a first performance evaluation on their outsourcing capabilities has been achieved.

REFERENCES

Aikin, J. (2003). *Software synthesizers.* San Francisco: Backbeat Books.

Bellini, P., Barthelemy, J., Nesi, P., & Zoia, G. (2004, September 13-15). A proposal for the integration of symbolic music notation into multimedia frameworks. In J. Delgado, P. Nesi, & K. Ng (Eds.) *Proceedings of 4th International Conference on Web Delivering of Music (WEDELMUSIC 2004)*, Barcelona, Spain (pp. 36-43). IEEE Computer Society Press.

Broersen, A., & Nijholt, A. (2002, September 9-12). Developing a virtual piano playing environment. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2002)*, Kazan, Russia (pp. 278-282).

Cook, P. (1996). *The SKINI interface*, Princeton. Retrieved December 2, 2005, from http://www.cs.princeton.edu/~prc

Dix, A., Finlay, J., Abowd, G. D., & Beale, R. (2004). *Human-computer interaction* (3rd ed.). Harlow, UK: Pearson Education Unlimited/Prentice Hall—Euope.

Freed, A. (1995). Improving graphical user interfaces for computer music applications. *Computer Music Journal, 19*(4), 4-5.

ISO DIS 9241-11 (1996). Ergonomic requirements for office work with visual display terminals (VDT)s—Part II guidance on usability.

Malik, S., & Laszlo, J. (2004, October 13-15). Visual touchpad: A two-handed gestural input device. In *Proceedings of International Conference on Multimodal Interfaces (ICMI '04)*, State College, PA (pp. 289-296).

Margounakis, D., & Politis, D. (2005, November 30- December 2). Producing music with N-delta interfaces. In *Proceedings of AXMEDIS 2005*, Florence, Italy (pp. 53-59).

Masui, T., Tsukada, K., & Siio, I. (2004). MouseField: A simple and versatile input device for ubiquitous computing. *UbiComp2004* (LNCS3205, pp. 319-328). Springer

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods.* New York:

John Wiley & Sons.

Orio, N., Schnell, N., & Wanderley, M. M. (2001, April). Input devices for musical expression: Borrowing tools from HCI. In *Proceedings of First Workshop on New Interfaces for Musical Expression (NIME01)—ACM CHI 2001*, Seattle, WA.

Phon-Amnuaisuk, S. (2004, September 15-16). Challenges and potentials in freehand music editing using pen and digital ink. In *The Fourth MUSICNETWORK Open Workshop*, Universitat Pompeu Fabra, Barcelona, Spain.

Politis, D., Vandikas, K., & Margounakis D. (2005, September 5-9). Notation-based Ancient Greek music synthesis with ARION. In Suvisoft Oy Ltd. (Ed.), *Proceedings of International Computer Music Conference ICMC 2005*, Barcelona, Spain (pp. 475-478).

Rossant, F., & Bloch, B. (2004). A fuzzy model for optical recognition of musical scores. *Fuzzy Sets and Systems, 141*, 165-201.

Schneidermann, B. (1997). *Designing the user interface—Strategies for effective human-computer interaction* (3rd ed.). MA: Addison-Wesley.

Valbom, L., & Macros, A. (2003, October-November). WAVE—AN audio virtual environment. In *Proceedings of 2nd International Workshop on ICTs, Arts and Cultural Heritage—Digital Art Technologies, Applications & Policy.* Foundation of the Hellenic World, Cultural Centre (Hellenic Cosmos), Athens, Greece.

Wanderley, M. M., & Orio, N. (2002, Fall). Evaluation of input devices for musical expression: Borrowing tools from HCI. *Computer Music Journal, 26*(3), 62-76.

Wilson, A. (2004, October 13-15). Touchlight: An imaging touch screen and display for gesture-based interaction. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, State College, PA (pp. 69-70).

Wilson, A., & Cutrell, E. (2005, September 12-16). FlowMouse: A computer vision-based pointing and gesture input device. In *Proceedings of INTERACT 2005*, Rome, Italy (pp. 565-578).

Wu, M., & Balakrishnan, R. (2003, November 2-5). Multi-finger and whole hand gestural interaction techniques for multi-user tabletop displays. In *Proccedings of ACM Symposium on User Interface Software and Technology*, Vancouver, Canada (pp. 193-202).

KEY TERMS

Alternative Music Systems: Music systems that are not conformed to Western music features (notation, scales, consonance, timbre, rhythm, etc.). Byzantine and Oriental music are two examples of such systems.

Computer Music (CM): A field of study that examines both the theory and application of new and existing technologies in the areas of music, sound design and diffusion, acoustics, sound synthesis, digital signal processing, and psychoacoustics.

Computer Music Languages: Programming languages specifically targeted to sound or music production and synthesis.

Human-Computer Interaction (HCI): The study of the interaction between people and computers.

Interface Design: The design of software applications with the focus on the user's experience and interaction.

Musical Praxis: Musical practice; exercise or discipline for a specific purpose or object (here music composition and performance).

Prototypal: Representing or constituting an original type after which other similar things are patterned.

Usability: The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. In order to evaluate usability, five criteria are used: (1) learnability, (2) effectiveness, (3) efficiency, (4) satisfaction, and (5) capability.

Computer-Aided Diagnosis of Cardiac Arrhythmias

Markos G. Tsipouras
University of Ioannina, Greece

Dimitrios I. Fotiadis
University of Ioannina, Greece, Biomedical Research Institute-FORTH, Greece, & Michaelideion Cardiology Center, Greece

Lambros K. Michalis
University of Ioannina, Greece & Michaelideion Cardiology Center, Greece

INTRODUCTION

In this chapter, the field of computer-aided diagnosis of cardiac arrhythmias is reviewed, methodologies are presented, and current trends are discussed. Cardiac arrhythmia is one of the leading causes of death in many countries worldwide. According to the World Health Organization, cardiovascular diseases are the cause of death of millions of people around the globe each year. The large variety and multifaceted nature of cardiac arrhythmias, combined with a wide range of treatments and outcomes, and complex relationships with other diseases, have made diagnosis and optimal treatment of cardiovascular diseases difficult for all but experienced cardiologists. Computer-aided diagnosis of medical deceases is one of the most important research fields in biomedical engineering. Several computer-aided approaches have been presented for automated detection and/or classification of cardiac arrhythmias. In what follows, we present methods reported in the literature in the last two decades that address: (i) the type of the diagnosis, that is, the expected result, (ii) the medical point of view, that is, the medical information and knowledge that is employed in order to reach the diagnosis, and (iii) the computer science point of view, that is, the data analysis techniques that are employed in order to reach the diagnosis.

BACKGROUND

Arrhythmia can be defined as either an irregular single heartbeat (arrhythmic beat), or as an irregular group of heartbeats (arrhythmic episode). Arrhythmias can affect the heart rate causing irregular rhythms, such as slow or fast heartbeat. Arrhythmias can take place in a healthy heart and be of minimal consequence (e.g., respiratory sinus arrhythmia), but they may also indicate a serious problem that may lead to stroke or sudden cardiac death (Sandoe & Sigurd, 1991). Ventricular arrhythmias may be categorized broadly as premature ventricular contractions (PVCs) and ventricular tachyarrhythmias, the latter including ventricular tachycardia (VT) and ventricular fibrillation (VF). Atrial fibrillation (AF) is the most prevalent arrhythmia in the western world, affecting 6% of the individuals over age 65 and 10% of those over age 80.

REVIEW OF THE PROPOSED METHODS

There are several aspects that can be addressed in order to review the proposed methods for computer-aided diagnosis of cardiac arrhythmias. The type of the diagnosis is the most important since cardiac arrhythmia is a very complex problem, having several different characteristics that need to be considered before reaching a safe diagnosis. Also, the electrocardiogram (ECG) analysis that is employed for this purpose is another important aspect. Finally, the data analysis and classification algorithms that are used define the accuracy and robustness of each approach.

Type of Diagnosis

Concerning the type of the diagnosis, two main approaches have been followed in the literature: (i) arrhythmic episode classification, where the techniques focus on the total episode and not on a single beat, and (ii) beat-by-beat classification, in which each beat is classified into one of several different classes related to arrhythmic behavior. Arrhythmic episode classification was performed in most of the methods proposed early in the literature, addressing mainly the discrimination of one or more of ventricular tachycardia (VT), ventricular fibrillation (VF), and atria fibrillation (AF) from normal sinus rhythm (NSR). More recent approaches mainly focus on beat-by-beat classification. In each case, a much larger

number of different types of cardiac arrhythmic beats are considered. A combination of these two different approaches has been proposed by Tsipouras (Tsipouras, Fotiadis, & Sideris, 2005), where beat-by-beat classification was initially performed and the generated annotation sequence was used in order to detect and classify several types of arrhythmic episodes.

Medical Data and Knowledge

In what concerns medical information, the main examination that leads to cardiac arrhythmia diagnosis is the ECG recording; thus, the majority of the methods proposed in the literature are based on its analysis. In the early studies, the ECG waveform was directly used for the analysis from it. However, more recent approaches are based on ECG feature extraction. In this case, features are mainly on the time and frequency domains in the early studies, while more complex time-frequency (TF) and chaos analysis are employed in the more recent studies, trying to access the nonstationary dynamic nature of the signal. Related to morphological features, QRS detection is the easiest to apply and the most accurate ECG processing method and thus, the most commonly used in the literature: almost all proposed methods include QRS detection in some stage of their analysis. Several other morphological features, inspired from the physiology of the ECG signal, have been employed in the proposed studies. However, the detection and measurement of all morphological features is seriously affected by the presence of noise, with the R peck being the least distorted. The identification of the importance of the heart rate variability has led to the development of methods based solely on the RR-interval signal and features that are extracted from it, for cardiac arrhythmia assessment.

Medical knowledge has been used in most of the proposed methods, mainly defining the features that carry sufficient information to be used for the classification. Rule-based medical knowledge has been employed in limited studies; instead mainly data-driven approaches have been developed. Thus, medical knowledge have also been employed for the generation of the annotation of the datasets, since data-driven approaches require an initial annotated dataset in order to be trained. In Tsipouras (Tsipouras, Voglis, & Fotiadis, 2007), a hybrid technique has been proposed in which both knowledge-based rules and training based on annotated data have been integrated into a single fuzzy model.

Data Analysis and Classification Algorithms

The general scheme that dominates the proposed methods for arrhythmia diagnosis is a two-stage approach: feature extraction from the ECG and classification based on these features. For the feature extraction stage, time and frequency analysis techniques have been gradually replaced by TF and chaos analysis. TF distributions and wavelet transform (WT) are commonly employed to measure the signal's energy distribution over the TF plane. In addition, combination between features originated from two or more different domains, has become a common practice. However, the major shift has emerged in the classification stage: the initial threshold-based techniques and crisp logic rules have been replaced with complex machine-learning techniques and fuzzy models. Most of the proposed approaches are based on several variations of artificial neural networks (ANNs) for the final classification and, in addition, hybrid approaches employing several ANNs and combining their outcomes have also been widely used. Support vector machines (SVMs) have also been incorporated, while fuzzy logic has been employed, either as part of the classification mechanism or for the definition of the fuzzy classifiers.

PROPOSED APPROACHES FOR COMPUTER-AIDED DIAGNOSIS OF CARDIAC ARRHYTHMIAS

Several researchers have addressed the issue of automated computer-based detection and classification of cardiac rhythms. We present methods that are considered in the literature as the most cited in the field. Thakor et al. (**Thakor,** Zhu, & Pan, **1990)** proposed a sequential hypothesis testing (SHT) algorithm for the detection of VT and VF. The detection is based on the evolution of a binary signal, generated using a threshold crossing interval (TCI) technique. The algorithm was tested in a database consisting of 170 ECG recordings (half from each category) having 8-sec length, and the results indicate that the detection was 94.12% for VT and 82.35% for VF after 4 sec, while the results increase to 100% for both arrhythmic types after 7 sec. An improvement of this approach was proposed by Thakor et al. (Thakor, Natarajan, & Tomaselli, 1994), named multiway SHT algorithm. In this case, three types of rhythm are considered, VT, supraventricular tachycardia (SVT), and NSR. The algorithm is evaluated in a dataset including 28, 31, and 43 segments for each type of rhythm and its accuracy is 98% after 1.6-, 5-, and 3.6-sec evolution time of the episode for VT, SVT, and NSR, respectively. Chen et al. (Chen, Clarkson, & Fan, 1996) applied the SHT algorithm, on the malignant arrhythmia subset of the MIT-BIH database (MITADB), resulting in much lower results. Thus, they have proposed a modification for this technique, employing dubbed blanking variability as basis for discrimination. Testing both the initial SHT technique and their modified approach to a subset of the database (30 episodes of VF and 70 episodes of VT), the classification accuracy improved from 84% to 95%. A similar approach

was proposed by Zhang et al. (Zhang, Zhu, Thakor, & Wang, 1999). Again, a binary signal is generated from the ECG recording using a threshold, and then the normalized complexity measured is calculated. Then, minimum, maximum, mean, and standard deviation features are extracted from it, for a specific window length, and are used for classification using statistical analysis. The classification is performed for three types of cardiac rhythm, VT, VF, and NSR. The dataset was the same as the one used in Thakor (1990), while 34 ECG recordings of NSR were added from the MITADB. A comparative study was performed with respect to the length of the window length, which varied from 1 to 8 sec. Results indicated 100% sensitivity and specificity for all categories when 7-sec window length is reached.

The problem of VF detection was also addressed by Alfonso and Tompkins (1995), using TF analysis of ECG recordings. They compared short-time Fourier transform (STFT) and two TF distributions, smoothed pseudo Wigner Ville distribution (SPWVD) and cone-shaped kernel distribution (CSKD), in order to identify the TF analysis that provided more discriminatory features for VF vs. NSR. Recordings of NSR were obtained from the MITADB, while VF recordings were extracted from the Stanley database. Based on these, the authors demonstrated that TF distributions can provide a more detailed inside of the VF than STFT. Khadra et al. (Khadra, Al-Fahoum, & Al-Nashash, 1997) proposed a more spherical approach for the arrhythmic episode classification problem, addressing the classification of VT, VF, atrial tachycardia (AT), and NSR. Their approach was also based on TF analysis, applying WT to ECG recordings in order to produce the TF representation, and measuring the signal's energy distribution in specific time intervals and frequency subbands (related to ECG physiology), thus producing three features. The final classification was provided using these features with a rule-based classifier. Forty-five ECG recordings of 2-sec length were selected from the MITADB and the ECG database of the electronic engineering department of the Yarmouk University (YUDB), with 12 associated with VF, 13 with VT, 12 with AT, and 8 with NSR, to evaluate the proposed technique and the obtained accuracy is 88.89%. A similar approach was proposed by Al-Fahoum and Howitt (1999). In this case, six features were extracted measuring the signal's energy distribution in specific time intervals related to ECG physiology and frequency subbands. Classification was performed using radial basis functions neural network (RBFNN), based on the leave-one-out evaluation strategy. The employed ECG recordings had 1-sec length, thus focusing on a single beat rather than a group of beats. For the evaluation of the method, 49 beats of VF, 49 beats of VT, 21 beats of AF, and 40 beats of NSR (in total 159 beats) were selected from the MITADB, the YUDB, and the Marquette medical systems ECG database. Comparative analysis was performed between nine different wavelets and the results indicated accuracy ranging from 81.1%-97.5%. TF analysis was also employed by Tsipouras and Fotiadis (2004) for the classification of ECG segments as normal or arrhythmic. In this study, time and TF analysis were used to extract features from the RR interval signal while classification was performed based on mixture-of-classifiers approach, using the combined results of several artificial neural networks (ANNs). In the time domain, six known heart rate variability (HRV) features were extracted, such as standard deviation of the RR intervals and standard deviation of the HRV, and all different combinations among them were used to train the ANNs. In the case of TF analysis, STFT and 18 time-frequency distributions were used to generate the TF representation and then, for each of them, features were extracted representing the signal's energy distribution over the TF plane and based on these, an ANN was trained. The evaluation was performed based on 32 RR intervals segments (corresponding to approximately 30-sec ECG recordings), extracted using all recordings from the MITADB. Classification sensitivity and specificity were 87.53% and 89.48%, respectively, for the time domain features, and 89.95% and 92.91%, respectively, for TF domain features.

Beat-to-beat classification was addressed by Simon and Eswaran (1997). In their work, analysis of each ECG waveform was performed using discrete cosine transform (DCT) and using the extracted coefficients as an input into a set of ANNs, each of them identifying a specific classification category, while the final classification was defined combining the outputs of all ANNs. The proposed method was used to identify beats belonging to five categories, while using data from the MITADB, the Glasgow database, and scanned waveforms, training and test sets were comprised, including 54 beats for training and 1,040 beats for testing. The evaluation of the method indicated an average sensitivity of 96.04%. Hu et al. (Hu, Palreddy, & Tompkins, 1997) have presented a method for beat classification, based on a mixture of experts approach, which is also a technique based on combining outcomes of ANNs. In this case, the ECG beat waveform was used as input for two ANN-based classifiers, one trained using data from several patients (global expert), and a second using data from a specific patient (local expert), and the combination of their outputs, using a gaiting network, provided the final classification. Although, the authors use a four-category dataset extracted from the MITADB, evaluation is based on the identification of only two categories, namely PVCs and normal (N) beats. Evaluation was based on 43,897 N beats and 5,363 PVC. The reported accuracy was 95.52%. Beat morphology and interval features were also used by Chazal et al. (Chazal, O'Dwyer, & Reilly, 2004) for cardiac beats classification. The extracted features are used with a linear discriminants classifier, based on maximum likelihood, to classify heart beats into five categories. The MITADB was used for training and evaluating the proposed method: 51,020 beats were used for training and 49,711 for testing, obtaining 86.2% accuracy.

Beat-by-beat classification was also addressed by Miniami et al. (Miniami, Nakajima, & Toyoshima, 1999), for supraventricular rhythm (SVR), ventricular rhythm (VR), and ventricular flutter/fibrillation (VFF). The spectrum of each QRS complex is accessed using FT, extracting five spectral components (features) fed into an ANN, which provides the final classification. The dataset included 700 QRSs, including 200, 100, and 200 for SVR, VR, and VFF, respectively, used for training the ANN and 100, 50, and 100 QRSs used for evaluation. Results indicated 90.33% average sensitivity and 95.33% average specificity. Dokur and Olmez (2001) also exploited beat-by-beat classification based on ANNs. Frequency domain features are extracted from the ECG beat waveform, using two different approaches, one based on FT and the second based on WT. Classification is performed for 10 categories of arrhythmic beats, adopted from the MITADB annotations. A hybrid ANN was trained for each of the two approaches for feature extraction, using 150 beats from each category for training while evaluation was performed using again 150 beats from each category. The obtained results indicated 65.4% and 88% classification accuracy for the FT and WT feature sets, respectively. Osowski and Lihn (2001) developed a method for ECG beat classification based on ANNs. In this case, statistical features were derived from each ECG beat waveform and, based on them, a fuzzy hybrid ANN was used to generate the final classification. The classification categories were 7, again originated from the MITADB annotations, while 4.035 beats were used for training and 3.150 for testing the method, which reported 96.06% classification accuracy. Ge et al., (Ge, Srinivasan, & Krishnan, 2002) proposed a method for classification of cardiac beats into six categories: NSR, artial premature contractions (APC), PVC, SVT, VT, and VF. The ECG waveform is analyzed using autoregressive (AR) modeling, and the estimated AR coefficients were used in a generalized linear model to classify the cardiac arrhythmias. Using 856 cardiac beats, obtained from the MITADB, the proposed approach presented 96.85% classification accuracy. In a second approach, Osowski et al. (Osowski, Linh, & Markiewicz, 2004) addressed this problem, using again statistical features, but also a second set of features based on Hermite basis functions (HBF) extraction, while SVMs were employed. The number of categories increased to 13 and the train and test sets were comprised from 6.690 and 6.095 beats, respectively. The average sensitivity was 93.72%, in the case of statistical features, 94.57% in the case of HBF features, and 95.91% when both classifiers were integrated.

Ham and Han (1996) proposed the discrimination of N beats and PVCs based on fuzzy adaptive resonance theory mapping (fuzzy ARTMAP). Their approach was based on the analysis of three features, extracted from each cardiac beat: two linear predictive coding coefficients and the mean square value of the QRS complex. Then, a fuzzy ARTMAP, which is an ANN classifier, was employed for the final discrimination. For training purposes, and to test the described method, 12,123 N beats and 2,868 PVCs were used, resulting in 99.88% sensitivity and 97.43% specificity. Fuzzy logic was also employed in the approach proposed by Weibien et al. (Weibien, Afonso, & Tompkins, 1999) for the same problem. In this case, features were extracted from each cardiac beat related to heart rate, morphology, and TF characteristics of the QRS, accessed using a filter-bank for analysis and a fuzzy rule-based classifier was employed for the final classification. For training, 8,089 N beats and 1,080 PVCs were used, and 77,374 and 5,900, respectively, for testing, reporting average sensitivity 74.6% and average positive predictivity of 66.5%. Wang et al. (Wang, Zhu, Thakor, & Xu, 2001) addressed the problem of classification of ECG records in three categories: VT, VF, and AF. The analysis was based on features representing the short-time generalized dimensions of multifractal analysis, while classification was provided by a fuzzy Kohonen network. The test set comprised from 60 ECG recordings of each category, 6- or 8-sec long, while the average sensitivity and specificity were 97.2% and 98.63%, respectively.

Acharya et al. (Acharya, Bhat, Iyengar, Rao, & Dua, 2003) presented a classification method based solely on the RR interval signal. Four features are extracted from the RR interval signal, the average heart rate, two features representing signals energy on specific frequency subbands and the correlation dimension factor, extracted from 2-D phase-space plot of the signal. Based on these, a fuzzy equivalence relation classifier is generated, for four classification categories: ischemic/dilated cardiomyopathy, complete heart block, sick sinus syndrome, and NSR. Two hundred and seventy six ECG segments were used for training the classifier and 66 for testing. The average sensitivity was 95%. Acharya et al. (Acharya, Kumar, Bhat, Lim, Iyengar, Kannathal, & Krishnan, 2004) also presented a similar approach for eight classification categories, the four mentioned previously, and additionally left bundle branch block, PVC, AF, and VF. In this case, three features from the RR interval signal were employed, the spectral entropy, geometry of the 2-D phase-space plot of the signal, and the largest Lyapunov exponent, while the classifier was the same. Two hundred and seventy nine ECG segments were used for training the classifier and 167 for testing; the average sensitivity was 85.36%. Another approach for arrhythmic beat classification and arrhythmic episode detection and classification which is also based on the RR-interval signal, is proposed by Tsipouras et. al. (Tsipouras, 2005). A three RR-interval sliding window is used in the arrhythmic beat knowledge-based classification algorithm. Classification is performed for four categories of beats: N, PVCs, VFF, and 2o heart block (BII). The beat classification is used as input of a knowledge-based

deterministic automato to achieve arrhythmic episode detection and classification. Six rhythm types are classified: ventricular bigeminy, ventricular trigeminy, ventricular couplet, VT, VFF, and BII. The achieved scores for all beats of the MITADB (approximately 110,000 beats) indicate high performance: 94% average accuracy for arrhythmic beat classification and 94% average accuracy for arrhythmic episode detection and classification. For the same arrhythmic beat classification problem, Tsipouras et al. presented a second approach (Tsipouras, 2007), based on fuzzy logic. A similar knowledge-based initial set of rules was employed, which was then transformed into a fuzzy model, and an optimization technique was used to tune its parameters. Comparative analysis was performed for different fuzzyfication functions and realizations of the fuzzy operators used in the fuzzy model. The same dataset was used for evaluation, indicating 96.5% average accuracy.

FUTURE TRENDS

Based on the presented works, there should be no doubt that several challenges remain concerning the computer-aided diagnosis of cardiac arrhythmias. From the medical point of view, a major limitation in all proposed methods is that they are based solely on the analysis of ECG recordings. Other medical information related to the patient, such as medication and medical history, are not included. Genomic data, such as single nucleotide polymorphisms related to cardiac disorders, can also have a significant impact on the diagnostic value of such systems if they are incorporated. From the computer science point of view, several of the proposed techniques are based on "black box" approaches for the classification, such as ANNs, which cannot provide interpretation of the produced diagnosis. This is a very important aspect, since providing a clear insight of their inner process for decision-making mechanism is essential for physicians in order to incorporate such systems in their clinical practice. Moreover, experts rely more on systems that contain, to some extent, established medical knowledge in their decision-making mechanism. Thus, machine learning techniques that can generate transparent classification mechanisms and can also integrate established medical knowledge are the key to creating computer-based arrhythmia diagnosis systems that will be widely used.

CONCLUSIONS

Several computer-based methods for the automated diagnosis of cardiac arrhythmias have been presented. The methods indicate that, under special conditions, they can provide the doctors with diagnosis of arrhythmias. However, most of the methods have been evaluated in existing datasets and not in clinical conditions, where usually the signals are noisy and not easily interpretable. It seems that methods based on knowledge, in general, do not provide with the best results, but the doctor could be helped since the decisions made are easily interpretable. All methods to become a real tool need further refinement in terms of facing real clinical conditions, real-time operation, and extensive evaluation.

REFERENCES

Acharya, U. R., Bhat, P. S., Iyengar, S. S., Rao, A., & Dua, S. (2003). Classification of heart rate data using artificial neural network and fuzzy equivalence relation. *Pattern Recognition, 36*, 61–68.

Acharya, U. R., Kumar, A., Bhat, P. S., Lim, C. M., Iyengar, S. S., Kannathal, N., & Krishnan, S. M. (2004). Classification of cardiac abnormalities using heart signals. *Medical & Biological Engineering and Computing, 42*, 288–293.

Afonso, V. X., & Tompkins, W. J. (1995). Detecting ventricular fibrillation. *IEEE Engineering in Medicine and Biology, 14*, 152–159.

Al-Fahoum, A. S., & Howitt, I. (1999). Combined wavelet transformation and radial basis neural networks for classifying life-threatening cardiac arrhythmias. *Medical & Biological Engineering and Computing, 37*, 566–573.

Chazal de, P., O'Dwyer, M., & Reilly, R. B. (2004). Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering, 51*, 1196–1206.

Chen, S. W., Clarkson, P. M., & Fan, Q. (1996). A robust sequential detection algorithm for cardiac arrhythmia classification. *IEEE Transactions on Biomedical Engineering, 43*, 1120–1125.

Dokur, Z., & Olmez, T. (2001). ECG beat classification by a hybrid neural network. *Computer Methods and Programs in Biomedicine, 66*, 167–181.

Ge, D., Srinivasan, N., & Krishnan, S. M. (2002). Cardiac arrhythmia classification using autoregressive modeling. *Biomedical Engineering OnLine, 1*.

Ham, F. M., & Han, S. (1996). Classification of cardiac arrhythmias using fuzzy ARTMAP. *IEEE Transactions on Biomedical Engineering, 43*, 425–430.

Hu, Y. Z., Palreddy, S., & Tompkins, W. J. (1997). A patient-adaptable ECG beat classifier using a mixture of experts approach. *IEEE Transactions on Biomedical Engineering, 44*, 891–900.

Khadra, L., Al-Fahoum, A. S., & Al-Nashash, H. (1997). Detection of life-threatening cardiac arrhythmias using wavelet transformation. *Medical & Biological Engineering and Computing, 35*, 626–632.

Minami, K., Nakajima, H., & Toyoshima, T. (1999). Real-time discrimination of ventricular tachyarrhythmia with Fourier-transform neural network. *IEEE Transactions on Biomedical Engineering, 46*, 179–185.

MIT-BIH (1997). *Arrhythmia Database CD-ROM*. Harvard-MIT Division of Health Sciences and Technology, 3rd Edition.

Osowski, S., & Linh, T. H. (2001). ECG beat recognition using fuzzy hybrid neural network. *IEEE Transactions on Biomedical Engineering, 48*, 1265–1271.

Osowski, S., Linh, T. H., & Markiewicz, T. (2004). Support vector machine-based expert system for reliable heartbeat recognition. *IEEE Transactions on Biomedical Engineering, 51*, 582–589.

Sandoe, E., & Sigurd, B. (1991). *Arrhythmia - A guide to clinical electrocardiology*. Bingen: Publishing Partners Verlags GmbH.

Simon, B. P., & Eswaran, C. (1997). An ECG classifier designed using modified decision based neural networks. *Computers and Biomedical Research, 30*, 257–272.

Thakor, N. V., Natarajan, A., & Tomaselli, G. (1994). Multiway sequential hypothesis testing for tachyarrhythmia discrimination. *IEEE Transactions on Biomedical Engineering, 41*, 480–487.

Thakor, N. V., Zhu Y. S., & Pan, K. Y. (1990). Ventricular tachycardia and fibrillation detection by a sequential hypothesis testing algorithm. *IEEE Transactions on Biomedical Engineering, 37*, 837–843.

Tsipouras, M. G., & Fotiadis, D. I. (2004). Automatic arrhythmia detection based on time and time-frequency analysis of heart rate variability. *Computer Methods and Programs in Biomedicine, 74*, 95–108.

Tsipouras, M. G., Fotiadis, D. I., & Sideris, D. (2005). An arrhythmia classification system based on the RR-interval signal. *Artificial Intelligence in Medicine, 33*, 237-250.

Tsipouras, M. G., Voglis, C., & Fotiadis, D. I. (2007). A framework for fuzzy expert system creation – application to cardiovascular diseases. *IEEE Transactions on Biomedical Engineering, 54*, 2089–2105.

Wang, Y., Zhu, Y.S., Thakor, N. V., & Xu, Y. H. (2001). A short-time multifractal approach for arrhythmia detection based on fuzzy neural network. *IEEE Transactions on Biomedical Engineering, 48*,989–995.

Weibien, O., Afonso, V. X., & Tompkins, W. J. (1999). Classification of premature ventricular complexes using filter bank features, induction of decision trees and a fuzzy rule-based system. *Medical & Biological Engineering and Computing, 37*, 560–565.

Zhang, X. S., Zhu, Y. S., Thakor, N. V., & Wang, Z. Z. (1999). Detecting ventricular tachycardia and fibrillation by complexity measure. *IEEE Transactions on Biomedical Engineering, 45*, 548–555.

KEY TERMS

Artificial Neural Network (ANN): An interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach to computation. It has the ability to learn from knowledge, which is expressed through interunit connection strengths, and can make this knowledge available for use.

Atrial Fibrillation (AF): Disorganized, high-rate atrial electrical activity.

Atrial Premature Contractions (APCs): Single or paired extrasystoles that originate in the atrials.

Electrocardiogram (ECG): Recording of the electrical activity of the heart.

Fuzzy Logic: Derived from fuzzy set theory dealing with reasoning that is approximate rather than precisely deduced from classical predicate logic.

Heart Rate Variability (HRV): The alterations of the heart rate between consecutive heartbeats.

Premature Ventricular Contractions (PVCs): Single or paired extrasystoles that originate in the ventricles.

QRS Detection: Procedure for detecting the QRS complexes in the ECG signal.

RR-Interval Signal: The signal representing the durations between consecutive R waves of the ECG.

Ventricular Tachyarrhythmias: Repetitive forms of three or more consecutive ventricular ectopic beats.

Computing Curriculum Analysis and Development

Anthony Scime
State University of New York College at Brockport, USA

INTRODUCTION

Information technology (IT) is an umbrella term that encompasses disciplines dealing with the computer and its functions. These disciplines originated from interests in using the computer to solve problems, the theory of computation, and the development of the computer and its components.

Professionals from around the world with similar interests in IT came together and formed international professional organizations. The professional organizations span the disciplines of computer engineering (CE), computer science (CS), software engineering (SE), computer information systems (CIS), management information systems (MIS), and information technology (IT) (Freeman & Aspray, 1999). Note that information technology is both an umbrella term and a specific discipline under that umbrella.

These organizations exist to promote their profession and one method of promotion is through education. So, these professional organizations defined bodies of knowledge around the computer, which have been formalized and shaped as model curriculums. The organizations hope that colleges and universities will educate students in the IT disciplines to become knowledgeable professionals.

Because of the common interest in computing, there is a basic theory and a common technical core that exists among the model curricula (Denning, 1999; Tucker et al., 1991). Nevertheless each of the model curricula emphasizes a different perspective of IT. Each fills a different role in providing IT professionals. It falls upon the colleges and universities to select and modify the corresponding curriculum model to fit their needs.

BACKGROUND

Currently, there are a number of model curricula for computing (Table 1). A Joint Task Force on Computing Curricula created by the Association for Computing Machinery (ACM), and the IEEE Computer Society (IEEE-CS) developed Computing Curricula 2001 (CC 2001). This model focuses on programs in theoretical and applied computer science with various areas of emphasis in all areas of computing including computer engineering (CE), the engineering of computer hardware, and computer science (CS), the theory and design of hardware and software (Computing Curricula, 2001).

The field of information systems (IS) can be divided into the management of information systems (MIS), the engineering of computer information systems (CIS) and the use of existing commercial software applications to solve organizational problems or information technology (IT). The Information Resource Management Association (IRMA) and the Data Administration Managers Association (DAMA) have a curriculum model for MIS known as the Information Resource Management (IRM) model. It takes a management of data approach to information systems (Cohen, 2000). For a strong accounting and management MIS orientation, the Information Systems Auditing and Control Foundation has developed an interdisciplinary curriculum known as the Information Systems Auditing at the Undergraduate and Graduate Levels (ISA) model (ISACF, 1998).

IS 2002 (Information Systems, 2002) is a model curriculum developed through the joint efforts of the ACM, the Association for Information Systems (AIS), and the Association of Information Technology Professionals (AITP). This curriculum model focuses on information systems development as well as on management (Gorgone, Davis, Valacich, Topi, Feinstein, & Longenecker, 2002). The Information Systems Centric Curriculum (ISCC '99) model was developed by a task force that included members from academe and industry. It is oriented to large-scale system design and implementation. The focus is on the construction of the tools necessary for information management (Lidtke, Stokes, Haines, & Mulder, 1999).

The IT education special interest group of the ACM (SIGSITE) has developed a curriculum proposal. This curriculum is oriented toward the use of computing applications to solve organizational problems (IT Curriculum Proposal – Draft, 2002). An existing IT model is the Organizational and End User Information Systems (OEIS) model developed by the Office Systems Research Association. It is aimed at IT support of end-users. (Office Systems Research Association, 1996).

The Software Engineering Institute (SEI) has developed a model that follows the engineering approach of design first in the construction of software for embedded, large, and critical systems. The model strongly suggests specialization in a specific application domain (Bagert, Hilburn, Hislop, Lutz, McCracken, & Mangal, 1999).

Table 1. IT professional organizations and curriculum models

Professional Organization	Curriculum Models
Association for Computing Machinery (ACM)	Computing Curricula 2001 (CC 2001)
	Information Systems 2002 (IS 2002)
Association for Information Systems (AIS)	Information Systems 2002 (IS 2002)
Association of Information Technology Professionals (AITP)	Information Systems 2002 (IS 2002)
Computing Society of the Institute of Electrical and Electronic Engineers (IEEE-CS)	Computing Curricula 2001 (CC 2001)
Data Administration Managers Association (DAMA)	Information Resource Management (IRM)
Information Resources Management Association (IRMA)	Information Resource Management (IRM)
Information Systems Audit and Control Association (ISACA)	Information Systems Auditing at the Undergraduate and Graduate Levels (ISA)
International Federation for Information Processing (IFIP)	Informatics Curriculum Framework 2000 (ICF-2000)
Office Systems Research Association (OSRA)	Organizational and End User Information Systems (OEIS)
Software Engineering Institute (SEI)	Software Engineering Institute (SEI)
Special Interest Group in Information Technology Education (SITE) of the ACM	Information Technology Curriculum Proposal (IT)
An independent group of Academics and Professionals	Information Systems Centric Curriculum ’99 (ISCC ’99)

Internationally, the International Federation for Information Processing (IFIP) in coordination with the United Nations Educational, Scientific and Cultural Organization (USESCO) has developed a framework within which schools can develop an IT curriculum (Mulder & van Weert, 2000). The Informatics Curriculum Framework 2000 (ICF-2000) considers the needs of developing countries for IT knowledgeable workers. These needs are balanced against the country’s educational resources. The result is a tailored IT curriculum based on the established models.

CONSIDERATIONS IN DEVELOPING A CURRICULUM

To select or develop a curriculum, a school needs to assess their objectives and capabilities in providing graduates to the IT work force. A school with a strong liberal arts tradition has a different philosophy than a technically oriented school. A large university may have schools of computing and business; the focus of each may produce different information technology professionals. Some schools prepare students for further study while others are oriented to the job market. Schools with an international or national focus have different objectives than schools providing entry-level professionals locally.

The resources of a school may limit the curriculum as well. Time is a critical resource. Some model curricula require a student begin studying information technology courses in the first semester, others require only 4 or 5 semesters of IT and begin in the 3rd year of study. IT disciplines vary on the requirement for facilities. Courses requiring hands-on hardware to test theory or practice application require laboratories similar to those in an electrical engineering department. Some curricula stress practical application in commercial settings. This requires the school have close working relationships with local sponsors. The interests of the IT faculty also have an impact on curriculum development. Ideally, IT departments are well balanced in faculty expertise. However, it is possible for a balanced department to have a greater interest in the development of IT artifacts versus the management of those artifacts. IT is a very large discipline, for one small department to be able to provide expertise in all facets of IT is unlikely.

Having considered the previously mentioned considerations, a school should also consider the role for which they are preparing graduates. Students require different knowledge dependent upon the role they will perform within IT. The

Table 2. Technical core and role emphasis by discipline and curriculum model

Discipline		CIS		MIS		IT		SE	CE	CS
Technical Core	**Model**	ISCC '99	IS 2002	IRM	ISA	IT	OEIS	SEI	CC 2001	CC 2001
	Role	CDM	CDS	C	DS	DS	CS	DM	CDM	CDM
Computer Literacy and Use of Software Tools			2R		1R 2E	1R 2E	1R 1E		1R	1R
Overview of IT and the Relationship to Business		2R	3R	2R 2E	1E	3E	2R 1E			
Computer Organization and Architecture			1R			3E			1R	2R 3E
Operating Systems and Systems Software									1R	1R
Programming, Algorithms and Data Structures		2R	1R	1R	1R	1R 2E		2R	3R	3R
Networking and Telecommunications		2R	1R	1E	1R 1E	1R 2E	1R		1R	
Systems/Software Analysis & Design		2R	2R	1R	2R 3E	3E	2R 3E	4R	3E	1R
Database and Information Retrieval		1R	1R	1R	1R 1E	1R 2E			1R	1R
Project Management			1R		2E	3E	1R	1R		
Intelligent Systems		1R		1R 1E	1E		1E		1R	1R
Social, Ethical & Professional Issues		1R					1E	1R	1R	
Internship/Capstone Project		1R					2E	1R	1R	1R

Key: C – Conceptualizer D – Developer M – Modifier S – Supporter xR – Number of Required Courses xE – Number of Elective Courses

curriculum studied helps determine the graduate's role. All information technology education consists of various technical, computer-oriented topics ranging from a theoretical understanding of computing, through the design and support of practical applications for complex computer systems. The depth of knowledge in these topics varies with the specific IT discipline and model curriculum (Scime, 2002b).

The fundamental knowledge in all the information technology disciplines involves the development, modification, support, conceptualization, and management of software and hardware artifacts (Freeman & Aspray, 1999; Information Technology Association of America, 1997). Developers work with programming and modeling languages. Developers need multiple courses in at least one of these areas. Modifiers need strength in programming or application tools. Supporters need to understand the end-users as well as the technology. Although conceptualizers typically have graduate degrees, at the undergraduate level this thinking begins with a strong theory component and by relating IT to the real world. Finally, IT managers are also conceptualizers by bringing together the IT professionals to satisfy an organization's need for information.

Not emphasizing the same common technical core of IT knowledge is what makes the IT disciplines differ from one another. Some areas of the core are emphasized in different model curricula, and provide an orientation of knowledge toward one or more of the professional roles. By considering the number of required and elective courses for an element of the technical core in a model the strengths of the model can be estimated. The relationship of the emphasis of tech-

nical topics, and the IT disciplines and professional roles supported by the curriculum models is of value in selecting a model as a beginning to program development (Table 2) (Scime, 2002b).

FUTURE TRENDS

The ACM and IEEE-CS are currently working on extending their model (CC 2001) to include other computing disciplines. They expect to create four volumes of curriculum models. These will include a model for computer engineering, software engineering, and information systems, as well as the current model for computer science. The software engineering work is currently in first draft. The computer engineering model is currently a preliminary draft. The information systems model is expected to closely match AIS/AITP/ACM's current IS 2002 model.

As computing becomes more and more ubiquitous, the use of the computer and its theoretical basis will continue to expand and infiltrate other disciplines. Expansion will manifest itself as the creation and clear definition of subdisciplines, such as CE, SE and IS are today. Infiltration is the inclusion of IT into the sciences, arts, and humanities, for example, the use of graphical information systems in earth science and criminal justice.

CONCLUSION

The constantly changing world of computing and the constantly changing world of business leads to the enviable weakness of computing education (Lidtke, Stokes, Haines, & Mulder, 1999). Each school needs to assess their educational philosophy and student needs to create the curriculum best for them. By closely following a model, a school's prospective students, student's potential employers, and graduate schools know the type of education received by the graduates. The school administration is assured that the IT department is providing a curriculum, which covers all of the central topics of IT and emphasizes the chosen information technology discipline (Scime, 2001, 2002a, 2002b).

Although the disciplines differ in emphasis, all businesses that use information (and they all do) will need information technologist from each of the disciplines. All are necessary, but not all need to be provided from the same source.

REFERENCES

Bagert, D.J., Hilburn, T.B., Hislop, G., Lutz, M., McCracken, M., & Mangal, S. (1999). Guidelines for Software Engineering Education Version 1.0 (Technical Report CMU/SEI-99-TR-032). Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Cohen, E. (Ed.) (2000). IRMA/DAMA curriculum model, IRMA, Hershey. Retrieved on November 15, 1999 from *http://gise.org/IRMA-DAMA-2000.pdf*

Computing Curricula 2001 (CC2001). (2000). The joint task force on computing curricula. *IEEE Computer Society and Association of Computing Machinery*, March 2000.

Denning, P.J. (1999, March). Our seed corn is growing in the commons. *Information Impacts Magazine*. Retrieved on September 19, 2000 from *http://www.cisp.org/imp/march_99/denning/03_99denning.htm*

Freeman, P., & Aspray, W. (1999). *The supply of information technology workers in the United States*. Washington, DC: Computing Research Association.

Gorgone, J.T., Davis, G.B., Valacich, J.S., Topi, H., Feinstein, D.L., & Longenecker, H.E., Jr. (2002). Model curriculum and guidelines for undergraduate degree *Programs in Information Systems Association for Information Systems*.

Information Technology Association of America (ITAA) (1997). Help wanted: The workforce gap at the dawn of a new century, Arlington, VA, (p.9).

ISACF (1998). ISACF Task Force for Development of Model Curricula in Information Systems Auditing at the Undergraduate and Graduate Levels, Academic Relations Committee and Research Board (1998). Model curricula for information systems auditing at the undergraduate and graduate levels. Information Systems Audit and Control Foundation.

IT Curriculum Proposal – Draft (2002). SITE Curriculum Committee. *Proceedings of the 2002 Conference for Information Technology Curriculum*, Rochester, NY, September.

Lidtke, D.K., Stokes, G.E., Haines, J., & Mulder, M.C. (1999). ISCC'99, An information systems-centric curriculum '99 program guidelines for educating the next generation of information systems specialists. In *Collaboration with industry*.

Mulder, F., and van Weert, T. (2000). ICF-2000 Informatics Curriculum Framework 2000 for Higher Education. Paris: UNESCO. Retrieved on February 7, 2004 from http://poe.netlab.csc.villanova.edu/ifip32/icf2000.htm

Office Systems Research Association (1996). Organizational & end-user information system curriculum model, OSRA. Retrieved on December 3, 2000, from *http://pages.nyu.edu/~bno1/osra/model_curriculum/*

Scime, A. (2001). Information systems draft accreditation criteria and model curricula. *Proceedings of the 18th Annual*

Information Systems Conference (ISECON 2001), Cincinnati, OH, November. Retrieved on November 10, 2003, from *http://colton.byuh.edu/don/isecon/*

Scime, A. (2002a). Information systems and computer science model curricula: A comparative look. Chapter 18 in A. Saber, S. Saber, & M. Dadashzadeh (Eds.), *Information technology education in the new millennium* (pp.146-158). Hershey, PA: IRM Press.

Scime, A. (2002b). Information technology model curricula analysis. Chapter 12 in E. Cohen (Ed.), *Challenges of information technology education in the 21st century* (pp.222-239). Hershey, PA: Idea Group Publishing.

Tucker, A.B., Barnes, B.H., Aieken, R.M., Barker, K., Bruce, K.B., Cain, J.T., Conry, S.E., Engel, G.L., Epstein, R.G., Lidtke, D.K., Mulder, M.C., Rogers, J.B., Spafford, E.H., & Turner, A.J. (1991). Computing Curricula 1991: Report of the ACM/IEEE-CS Joint Curriculum Task Force, Association of Computing Machinery.

KEY TERMS

Computer Engineering (CE): The engineering of computer hardware.

Computer Information Systems (CIS): Concerns information systems with an emphasis on information as an enterprise resource, and its design, development, implementation, and maintenance of information systems. Sometimes referred to as information systems.

Computer Science (CS): Hardware and software theory and design.

Computing Curricula 2001 (CC 2001): Developed by the Joint Task Force on Computing Curricula created by the ACM and the IEEE-CS. This model focuses on programs in theoretical and applied computer science (Computing Curricula, 2001).

Information Resource Management (IRM): Curriculum model 2000 of the IRMA and the DAMA focuses particularly on the disciples of information resource management and management information systems. It takes a management of data approach (Cohen, 2000).

Information Systems (IS): Use data to create information and knowledge to assist in operational, management, and strategic organizational decision-making. It is also an umbrella term for computer information systems, management information systems and information technology.

Information Systems 2002 (IS 2002): Model curriculum developed by the efforts of the ACM, AIS, and AITP. This curriculum model focuses on information systems development and management (Gorgone, Davis, Valacich, Topi, Feinstein, & Longenecker, 2002).

Information Systems Auditing (ISA) at the Undergraduate and Graduate Levels: Model developed by the ISACF Task Force for Development of Model Curricula. It is an interdisciplinary approach with a strong accounting and management orientation. (ISACF, 1998).

Information Systems Centric Curriculum (ISCC '99): Model developed by a task force that included members from academe and industry. It is oriented toward large-scale system design and implementation as opposed to automata theory and programming. The focus is on the construction of the tools necessary for information management (Lidtke, Stokes, Haines, & Mulder, 1999).

Information Technology (IT): Uses existing commercial software applications to solve organizational problems. Sometimes refer to as information systems. It is also the umbrella term for all the disciplines involved with the computer.

IT Curriculum Proposal: Being developed by SIGSITE, the curriculum is oriented toward the use of computing applications to solve organizational problems (IT Curriculum Proposal – Draft, 2002).

Management Information Systems (MIS): The management of information systems and data including management of the design, development, implementation, and maintenance of information systems. Sometimes referred to as information systems.

Organizational and End User Information Systems (OEIS): Model developed by the OSRA is aimed at IT support of end-users (Office Systems Research Association, 1996).

Software Engineering (SE): The engineering of software for embedded, large, and critical systems.

Software Engineering Institute (SEI) Model: Model developed at SEI that follows an engineering approach of design first. The model suggests specialization in a specific domain (Bagert, Hilburn, Hislop, Lutz, McCracken, & Mangal, 1999).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 508-512, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Concept-Oriented Programming

Alexandr Savinov
University of Bonn, Germany

INTRODUCTION

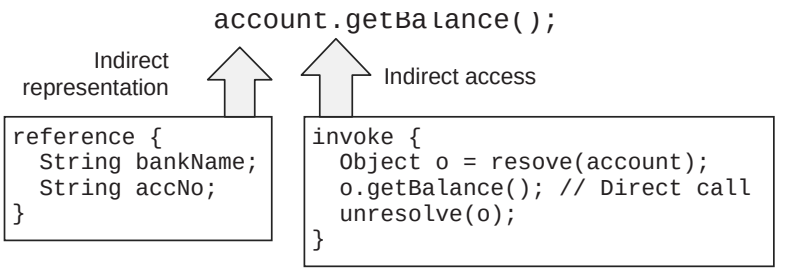
In the concept-oriented programming (CoP) (Savinov, 2005, 2007), the main idea is common to many other approaches and consists in raising the abstraction level of programming by introducing new language constructs and mechanisms. The distinguishing feature of CoP is that it aims at automating the way objects are represented and accessed (ORA). More specifically, one of the main concerns in CoP is modeling the format of object references and the procedures executed during object access.

For example, if we need to retrieve the current balance stored in a bank account object then we make the following simple method call: `account.getBalance()`. In object-oriented programming (OOP), it results in an *instantaneous* execution of the target method because this variable contains a *primitive* reference which is supposed to provide *direct* access to the represented object. In CoP, it is not so and everything depends on the format of the reference used to represent this account object. References in CoP have an arbitrary custom format defined by the programmer and hence objects are represented *indirectly* using abstract identifiers from a virtual address space. In this case, the real procedure executed during access depends on what is stored in the variable `account`. In particular, it may well happen that the account object is stored on a remote computer in another organization. Then, its reference can be rather complex and include such fields as `bankName` and `accNo` (Figure 1). Object access to such an indirectly represented account will involve many intermediate operations like security checks, transaction management, network packet transfer and operations with persistent storage. However, all these intermediate actions will be executed behind the scenes so that we have the illusion of instantaneous action. Then the programmer is still able to use the target objects as if they were local directly accessible objects, at the same time having a possibility to inject any intermediate code responsible for object representation and access (ORA).

References in CoP are as important as objects because both have arbitrary structure and behavior associated with them. If OOP deals with objects then CoP deals with both objects and references. The main role of references consists in representing objects, that is, they contain some data that makes it possible to access the object. Thus, references are *intermediate* elements which are activated each time the represented object is about to be accessed. For example, each time we read or write a field, or call a method, the object reference intercepts these requests and injects its own actions. Thus, any object access can trigger a rather complex sequence of intermediate actions which are executed behind the scenes. In large programs this hidden functionality associated with references can account for a great deal or even most of the overall complexity. The main task of CoP in this sense consists in providing adequate means for effectively describing this type of hidden intermediate functionality which has a cross-cutting nature. OOP does not provide any facilities for describing custom references and all objects are represented and accessed in one and the same way. CoP fills this gap and allows the programmer to effectively separate both concerns (Dijkstra, 1976): explicitly used business logic of objects and intermediate functions executed implicitly during object access.

The problem of indirect object representation and access can be solved by using such approaches as dynamic proxies (Blosser, 2000), mixins (Bracha & Cook, 1990; Smaragdakis & Batory, 1998), metaobject protocol (Kiczales et al.,

Figure 1. Indirect method call via custom references and intermediate operations



1991; Kiczales et al., 1993), remoting via some middleware (Monson-Haefel, 2006), smart pointers (Stroustrup, 1991), aspect-oriented programming (Kiczales et al., 1997) and others. However, CoP is the only technology that has been developed for precisely this problem and solves it in a principled manner. It is important that CoP generalizes OOP by providing a possibility of smooth transfer to the new technology.

BACKGROUND

Hierarchical Address Space

A concept-oriented program can be viewed as a set of nested spaces (Figure 2). Each space has one parent where it is identified by some local address. The parent space itself is identified by some address relative to its own parent and so on till the root. Thus, any element is identified by a sequence of local addresses where each next address identifies the next space. Such an identifier is referred to as a *complex address* while its constituents are referred to as *segments*. This structure is analogous to the conventional postal addresses where cities are identified by names within countries and streets have unique names within cities. For example, an element in the postal address space could be identified by the following complex address: `<"Germany," "Bonn," "University of Bonn">`.

An important consequence of such a design is that objects can interact only by intersecting intermediate space borders. An access request such as a method call or message cannot *directly* (instantaneously) reach its target. Instead, it follows some access path starting from the external space and leading to the internal target space (Figure 2). In order to access an element of the space it is necessary to resolve all segments of its complex reference starting from the high segment and ending with the low segment, which is the target object. The resolution procedure is responsible for locating the element identified by one segment in the context of the parent space. Thus, each intermediate border along the access path executes some special functions, which are triggered automatically as an access request intersects this border.

The same approach is used in CoP where objects are identified by complex references defined by the programmer rather than using primitive references. For example, if account reference consists of two segments—bank name and account number—then the balance could be obtained as usual by applying a method to this complex reference:

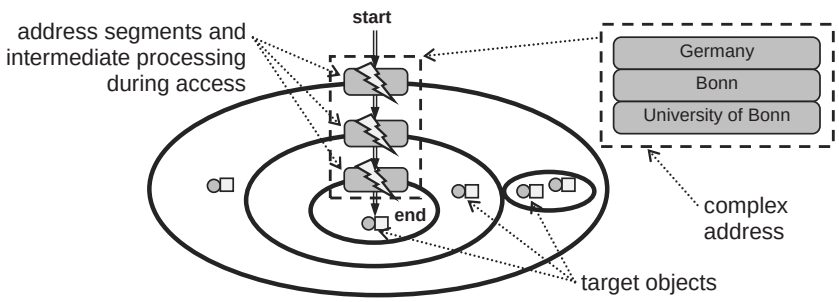
```
Account account = <"MyBank," "98765432">;
double balance = account.getBalance();
```

Since objects in CoP are represented by complex references each access requires several intermediate steps for locating the object. For example, in order to resolve the account object represented by its bank name and account number it is necessary to find the bank object and then to find the account object. Notice that the method applied to the reference is only the last step in this indirect access procedure. An important assumption of the concept-oriented approach to programming is that most of the functionality in large programs is concentrated on intermediate space borders. Target methods in this case account for a relatively small portion of the overall complexity. The goal of CoP in this sense can be formulated as providing support for describing such a hierarchical space at the level of the programming language rather than in middleware or hardware. The programmer then is able to describe an arbitrary *virtual address space* which serves as a container for objects. Such addresses are virtual because they are not directly connected with the real object position and hence they provide an additional level of abstraction.

References and Objects

In OOP, the programmer models objects by classes while all references have one and the same type. Thus we cannot influence how objects are represented and how they are

Figure 2. A program can be viewed as a hierarchical space



accessed. In CoP, the programmer deals with two types of things—references and objects—as opposed to OOP where only objects are considered. This means that any concept-oriented program consists of and manipulates custom objects and custom references both having arbitrary structure and behaviour. However, objects and references are not considered separately. Instead, they are thought of as two sides of one and the same thing. In other words, in CoP, there are no such things as an isolated reference and an isolated object—they can only exist within one element as two its sides or flavors. Thus any concept-oriented system consists of and manipulates object-reference pairs. The space of all elements is then broken into two parts: the *identity world* consisting of references and the *entity world* consisting of objects (Figure 3). For example, a city has two sides: city name and city object. A bank account also has two sides: account reference (with properties like account number), and account object (with properties like balance). The two worlds of references and objects can be viewed as cross-cutting concerns separated in a principled manner within CoP. The functions of objects and references are orthogonal but on the other hand these two elements cannot exist separately because they are two sides of one and the same thing.

Both references and objects have their own structure and functions defined by the programmer however their roles are different. References are intended to represent objects. They are always passed by-value while objects are passed by-reference. Thus references do not have their own locations and exist only in transient form. Reference represents part of reality which is directly comprehendible while object is a thing-in-itself which is not observable in its original form and hence is radically unknowable. So the only way to get information about an object or to interact with it consists in using its reference which stays between us and objective reality.

Elements of the program (reference-object pairs) exist in a hierarchical space where each element has a parent element as described in the previous section. This hierarchical structure is described by *inclusion relation*, that is, we say that an element is included into its parent element. For example, Bonn is included in Germany and accounts "456789" and "987654" are included in "MyBank" element identifying some bank.

Additionally, it is assumed that any element substitutes for some primitive element and this structure is described by means of *substitution relation*. For example, DNS computer names substitute for IP addresses and Java references substitute for memory handles. Finding the primitive reference substituted by this reference segment is referred to as *reference resolution*. The resolution procedure allows us to translate virtual address into real addresses and to get direct access to the represented entity.

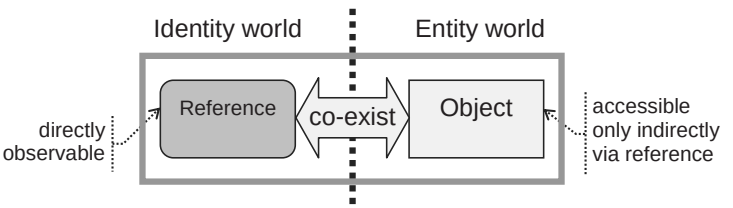
FROM CLASSES TO CONCEPTS

Concepts and Inclusion Relation

Any concept-oriented program manipulates reference-object pairs. Since both constituents have their own structure and behavior they can be described by means of conventional classes. This means that references are described by *reference classes* and objects are described (as usual) by *object classes*. However, since references and objects are two sides of one element these two classes are constituents of one construct, called *concept*. Thus, concept is defined as a pair of two classes: one reference class and one object class. If concept has empty reference class then it is equivalent to normal classes as used in OOP and therefore concepts can be viewed as a generalization of classes. For example, a concept of bank account could be defined as follows:

```
concept Account
 reference { // Reference class
   char[8] accNo;
   double getBalance() { ... }
   Person getOwner(Address address) { ... }
   ... // Other reference class members
 }
 object { // Object class
   double balance;
   double getBalance() { ... }
   Person getOwner(Address address) { ... }
   ... // Other object class members
 }
```

Figure 3. Two crosscutting concerns: The world of identities and the world of entities



Here the reference class is declared using keyword 'reference' while object class is marked by keyword 'object.' Concepts are used where classes are used in OOP, namely, for declaring types of variables, parameters, fields, return values and other elements of the program. In the above example method `GetOwner` takes one parameter of type `Address` and returns an element of type `Person`. However, since it is a concept-oriented program, these two types are names of concepts. As a consequence, any element stores a custom reference with the format defined in the corresponding concept rather than a primitive reference. For example, the return value might store a person unique number or full name and birthday depending on how concept `Person` defines its reference class.

When a concept is instantiated then two instances are created: an instance of the reference class, called reference, and an instance of the object class, called object. The reference is stored in the variable (in code) while the object is created in some storage (in data) like memory. Actually, objects can reside anywhere in the world because they are represented by *virtual* addresses. References are always stored and passed by-value while objects are stored and passed by-reference. In other words, references do not have their own references and their role consists in indirectly representing objects. Thus, we always manipulate references which indirectly represent objects.

It is important that a reference always intercepts any access to the represented object. This means that if we call a method which is defined in both reference class and object class then the reference method will be executed before object method. The reference method can then call its object method if necessary. In the following example, reference method `getBalance` simply calls its object counterpart (dual method) using keyword 'object' and the object method `getBalance` then returns the value of its field:

```
concept Account
 reference { // Reference class
   char[8] accNo;
   double getBalance() {
     print("===> Reference method");
     return object.getBalance();
   }
 }
 object { // Object class
   double balance;
   double getBalance() {
     print("---> Object method");
     return balance;
   }
 }
```

If we apply method `getBalance` to an account reference

```
Account account = findAccount(person);
account.getBalance();
```

then the following output will be produced:

```
$ ===> Reference method
$ ---> Object method
```

In OOP, any method invocation results in an immediate execution of its body defined in the object class. In CoP variables store custom data according to the reference class structure. Therefore, object method cannot start immediately and reference methods play a role of intermediate processing points which take control before the represented object can be reached. When an object method is executed it can use functions of its reference by means of keyword 'reference.' For example, object method `getBalance` from the previous example could print its account number as follows:

```
print("---> Account: "+reference.accNo);
```

Just as classes in OOP can inherit base classes, concepts are defined within a hierarchy. The difference is that concepts use *inclusion relation* for describing their hierarchy, that is, a concept can be included in a parent concept. Inclusion relation generalizes inheritance and its main distinguishing feature is that it describes a hierarchy of elements, that is, concept instances exist within a hierarchy at run-time. In other words, if classes in OOP are connected by means of IS-A relation, then concepts in CoP are connected via IS-IN relation. For example, one bank account may have many (internal) savings accounts which are described by concept `SavingsAccount` included in parent concept `Account`:

```
concept SavingsAccount in Account
 reference { // Reference class
   char[2] subAccNo; // Sub-account id
   double getBalance() { ... }
   ...
 }
 object { // Object class
   double balance;
   double getBalance() { ... }
   ...
 }
```

Here we use keyword 'in' to declare the parent concept. Sub-accounts are identified by two digits stored in a field of its reference class. However, it is important to understand that this identifier is local, that is, it uniquely identifies its objects only in the context of some parent account. In order to get a complete reference we need to store two segments: one parent reference consisting of eight digits and one child reference consisting of two digits. For example, if we get an account for some person:

```
SavingsAccount account = findAccount(person);
double b = account.getBalance();
```

then variable `account` will contain a *complex reference* consisting of two segments: `<"12345678," "01">`. High segment identifies the main account (concept `Account`) while low segment identifies some its sub-account (concept `SavingsAccount`). Thus concepts and inclusion relation allow us to control what is stored in variables and passed in parameters rather than only the object format in OOP. Notice that the programmer does not have to know where the objects really reside because variables store virtual addresses and finding the object is the task of the corresponding concept. In particular, we can still use variable `account` as if it contained a primitive reference for calling various methods like `getBalance` in the previous example. Such access may trigger rather complex intermediate behavior (security, transactionality, persistence, etc.) which is however hidden in the definition of concepts `Account` and `SavingsAccount`.

Reference Resolution

References in CoP are virtual addresses and hence they represent objects only indirectly by substituting for some primitive reference which provides direct access. Accordingly, in order to access the target object, it is necessary to resolve its indirect (virtual) reference, which is an instance of some concept reference class defined in the program. It is assumed that any reference class is responsible for resolving its instances (references) into primitive references. This resolution logic is implemented in a special method of the reference class, called *continuation method*. Whenever an object of this concept is about to be accessed, the compiler automatically calls the continuation method of its reference. The resolved reference is then used to directly access the object state and functions. For example, continuation method (called `continue`) could be implemented as shown in the following code fragment:

```
concept Account
 reference { // Reference class
   char[8] accNo;
   void continue() {
     print("===> Account: Start resolution");
     Object primitive = loadFromDb(accNo);
     primitive.continue(); // Proceed
     storeInDb(accNo, primitive);
     print("<=== Account: End resolution");
   }
 }
 object {
   double balance;
 }
```

Here continuation method of concept `Account` prints a message when it starts and loads the account object in memory from some database using its number as a primary key. Then it uses the obtained primitive reference (with the type `Object`) in order to pass control to the object and it is precisely the point where the target method is called or the object is otherwise accessed. When the access is finished, the continuation method stores the state of the object back to the database and finally prints a message. For example, the following output will be produced each time the account balance is read:

```
$ ===> Account: Start resolution
$ <=== Account: End resolution
```

Effectively, any access to the object is wrapped into the reference resolution procedure implemented in the continuation method however these actions will be executed implicitly. An advantage is that our code does not depend on how objects are represented and accessed, i.e., these concerns are separated. If later we change the format of references or the location of objects then the code that uses them will remain the same. Thus we can work with indirectly represented objects located anywhere in the world as if they were directly accessible local objects.

Normally concepts are defined as included in some parent concept and hence their objects are represented by complex references consisting of several segments. In this case each reference segment has to be resolved individually by its own continuation method. The resolution procedure in this case consists of several nested steps. It starts from resolving the very first (high) segment corresponding to the parent concept. Then it resolves the second segment and so on down to the last (low) segment corresponding to the type of the object. Each resolved primitive reference is stored on top of a special data structure called *context stack*. When the complex reference is completely resolved the context stack contains direct primitive references of all the object segments. It is important that when the next segment is resolved all its parent objects can be directly accessed using their primitive references from the context stack. Parent objects can be accessed using keyword 'super.' Notice however that one parent object can be shared among many child objects while in OOP for each child there is one parent. Normally parent object is used to store information about the location of its children.

For example, let us assume that an account has many internal savings accounts which are described by concept `SavingsAccount` included in concept `Account`. Parent concept `Account` can resolve its references as shown in the previous example by loading account objects from some persistent storage. Savings accounts have to resolve their own references (two digit number) in the context of their main account. In particular, we can store the mapping

between sub-account numbers and object primitive references in a field in the parent object as shown in the following code fragment:

```
concept Account
 reference { // Reference class
   char[8] accNo;
   void continue() {
     // Precisely as in the pervious example
   }
 }
 object {
   double balance;
   Map map; // Used by child objects
 }

concept SavingsAccount in Account
 reference { // Reference class
   char[2] subAccNo; // Sub-account id
   void continue() {
     print("  ===> SavingsAccount: Start resolution");
     Object primitive = super.map.get(subAccNo);
     primitive.continue();
     print("  <=== SavingsAccount: End resolution");
   }
 }
 object { // Object class
   double balance;
 }
```

Notice that here parent concept `Account` has one additional field `map` in its object class which stores the mapping from virtual sub-account identifiers to their primitive references. Child concept `SavingsAccount` accesses this field from its continuation method using 'super.' If a savings account needs to be accessed then its main account number will be resolved by concept `Account` and then its sub-account number will be resolved on the second step by concept `SavingsAccount`. For each such access the following output will be produced:

```
$ ===> Account: Start resolution
$   ===> SavingsAccount: Start resolution
$   <=== SavingsAccount: End resolution
$ <=== Account: End resolution
```

Dual Methods and Overriding

A concept can provide two definitions of one method in its reference class and object class, which are called *dual methods*. In addition, parent and child concepts can as usual define one and the same method. For example, if concept `SavingsAccount` is included in concept `Account` then both concepts can implement method `doAction` in their reference classes and object classes. Thus there are 4 implementations of this method in two concepts. Earlier we postulated that references intercept method calls to their objects but in this case objects are represented by two reference segments implementing method `doAction`. In order to resolve this ambiguity CoP follows the following principle: *parent reference methods have precedence over (override) child reference methods*. This means that if we apply a method to a multi-segment reference then its high segment will intercept this call by executing its version of this method. This principle is quite natural because for entering some internal space it is necessary to cross its external border. For example, in order to reach some city we have to intersect the country border. This principle of overriding is analogous to the mechanism of inner methods (Goldberg, Findler, & Flatt, 2004) used in the Beta programming language (Kristensen, Madsen, Moller-Pedersen, & Nygaard, 1987). In the following example concept `Account` implements method `doAction` in its reference class and hence it will be executed each time this method is applied to a reference of any sub-concept including `SavingsAccount`:

```
concept Account
 reference { // Reference class
   char[8] accNo;
   void doAction() {
     print("===> Account: Reference method");
     sub.doAction();
     print("<=== Account: Reference method");
   }
 }
 object { // Object class
   double balance;
   void doAction() {
     print("---> Account: Object method");
     super.doAction();
     print("<--- Account: Object method");
   }
 }
```

By means of this mechanism parent reference methods can override child reference methods. It is useful if it is necessary to control access to internal scope. Once a parent reference method got control it can decide how to continue. Normally, parent methods perform some actions and then delegate the method call further to its child reference. Child reference segment is available via special keyword 'sub' which is opposite to the conventional 'super' for accessing parent objects.

Applying a method to keyword 'sub' means that control is passed to the next segment of the complex reference. In our example reference method `doAction` of concept `SavingsAccount` will be called. This method can delegate request further to a possible child reference but in the example below it calls the same object method using keyword 'object:'

```
concept SavingsAccount in Account
 reference { // Reference class
  char[2] subAccNo; // Sub-account id
  void doAction() {
   print("  ===> SavingsAccount: Reference method");
   object.Action();
   print("  <=== SavingsAccount: Reference method");
  }
 }
 object {
  double balance;
  void doAction() {
   print("  ---> SavingsAccount: Object method");
   super.doAction();
   print("  <--- SavingsAccount: Object method");
  }
 }
```

Once we are in an object method, it is possible to call methods of parent object using keyword 'super' precisely as it is done in OOP. Here we use the dual principle formulated as follows: *child object methods have precedence over (override) parent object methods*. For example, object method `doAction` of concept `SavingsAccount` calls its parent object method (method of the context). Notice that parent objects are accessed directly using primitive references from the context stack. Here, the output produced during invocation of method `doAction` for a savings account object is shown:

```
$ ===> Account: Reference method
$   ===> SavingsAccount: Reference method
$   ---> SavingsAccount: Object method
$ ---> Account: Object method
$ <--- Account: Object method
$   <--- SavingsAccount: Object method
$   <=== SavingsAccount: Reference method
$ <=== Account: Reference method
```

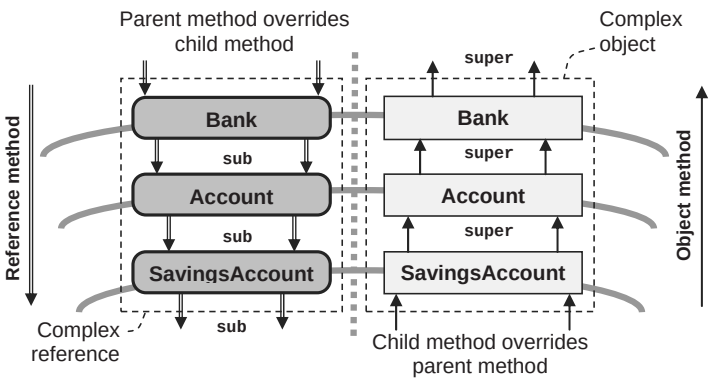
The sequence of access via dual methods is shown in Figure 4. It is assumed that concept `SavingsAccount` is included in concept `Account` which in turn is included in concept `Bank`. Thus the element has three reference segments (on the left) and three object segments (on the right). If some method is applied to such a complex reference then the parent reference segment intercepts it and then the same method of the child reference segment is called. Thus we can move down through reference segments using keyword 'sub.' At some moment the reference method can call its object method and the process switches to the right half of the diagram (entity world). Here the process propagates upward through the object segments using keyword 'super.'

FUTURE TRENDS

In future this approach will be developed in the direction of defining concrete programming languages and introducing elements of the concept-oriented programming in existing languages. In order to implement the described approach it is necessary to develop such mechanisms as complex references, context stack, reference length control and others. One general problem here is that there are two approaches to using concepts which are called CoP-I (Savinov, 2005) and CoP-II (Savinov, 2007, 2008; this article). Hence it is necessary to understand what are their advantages and disadvantages when used in programming languages.

Another direction for research consists in integrating this approach with the concept-oriented data model (CoM) (Savinov, 2006, 2009) which is based on the same principles. The challenging goal here consists in unifying programming with data modeling. This model uses the formalism of nested ordered sets when describing data semantics and such operations as projection and de-projection for data manipulations. However, identity modeling in CoM is completely analogical to CoP. In particular, data can be described by concepts and inclusion relation. Any data item is uniquely identified by its complex reference from a virtual address space which can be stored in other data items as a property.

Figure 4. Duality of method overriding



CONCLUSION

For the past several decades, programmers have been modeling things in the world with trees using hierarchies of classes and object-oriented programming languages. This abstraction works out pretty well because most of the world is hierarchical and in most cases things can be easily fit into a hierarchy. Yet, there is one serious problem with this approach: classes allow us to model only entities (objects) but not identities (references). Thus identity modeling and references were considered second class citizens in the area of computer programming, data modeling, analysis and design. The concept-oriented paradigm fills this gap by introducing concepts and inclusion hierarchy which generalize classes and inheritance. Now the world is still described using a hierarchy but such a model reflects the dual nature of things which consist of one entity and one identity. Such a generalization is informally analogous to introducing complex numbers in mathematics. Just as concepts in CoP, complex numbers have two constituents: a real part (object or entity) and an imaginary part (reference or identity). Yet these two constituents are always manipulated as one whole and this makes mathematical expressions much simpler and more natural. The same effect we get in programming when concepts are introduced: program code gets simpler and more natural because two sides or flavors are manipulated as one whole.

REFERENCES

Blosser, J. (2000). Explore the Dynamic Proxy API, *Java World*. Retrieved November 2000 from http://www.javaworld.com/javaworld/jw-11-2000/jw-1110-proxy.html

Bracha, G., & Cook, W. (1990). Mixin-based inheritance. In *Proceedings of the OOPSLA/ECOOP'90. ACM SIGPLAN Notices, 25*(10), 303-311.

Dijkstra, E.W. (1976). *A discipline of programming*. Prentice Hall.

Goldberg, D.S., Findler, R.B., & Flatt, M. (2004). Super and inner: together at last! *Proc. OOPSLA'04*, 116-129.

Kiczales, G., Rivieres, J., & Bobrow, D.G. (1991). *The art of the metaobject protocol*. MIT Press.

Kiczales, G., Ashley, J.M., Rodriguez, L., Vahdat, A., & Bobrow, D.G. (1993). Metaobject protocols: Why we want them and what else they can do. In A. Paepcke (Ed.), *Object-oriented programming: The CLOS Perspective*, (pp. 101-118). MIT Press.

Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Lopes, C., Loingtier, J.-M. et al. (1997). Aspect-Oriented Programming. In *Proceedings of ECOOP'97*, LNCS 1241, (pp. 220-242).

Kristensen, B.B., Madsen, O.L., Moller-Pedersen, B., & Nygaard, K. (1987). The Beta programming language. In *Research Directions in Object-Oriented Programming*. (pp. 7-48). MIT Press.

Monson-Haefel, R. (2006). *Enterprise JavaBeans*. O'Reilly.

Savinov, A. (2005). Concept as a generalization of class and principles of the concept-oriented programming. *Computer Science Journal of Moldova*, *13*(3), 292-335.

Savinov, A. (2006). Grouping and Aggregation in the Concept-Oriented Data Model. *In proceedings of the 21st Annual ACM Symposium on Applied Computing (SAC'06)*, Dijon, France, (pp. 482-486).

Savinov, A. (2007). *An Approach to Programming Based on Concepts*. Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, Technical Report RT0005, 49 pp.

Savinov, A. (2008). Concepts and concept-oriented programming. *Journal of Object Technology, March-April 2008 7*(3), 91-106.

Savinov, A. (2009). Concept-oriented model. In V. E. Ferraggine, J. H. Doorn, & L. C. Rivero (Eds.), *Encyclopedia of Database Technologies and Applications*, (2nd ed.). IGI Global (accepted).

Smaragdakis, Y., & Batory, D. (1998). Implementing layered designs with mixin-layers. *Proc. ECOOP'98*, (pp. 550-570).

Stroustrup, B. (1991). *The C++ Programming Language*, (2nd ed.). Addison Wesley.

KEY TERMS

Complex Reference: A sequence of references, called segments, where each next segment represents an object within the context of the previous segment. Complex references are used to represent objects in a hierarchical address space. The format of one segment is defined by the reference class of the corresponding concept.

Concept: A pair consisting of one object class and one reference class. Instances of the object class are referred to as objects and are passed-by-reference. Instances of the

reference class, called references, are passed by-value and represent objects.

Continuation Method: A special method of the reference class which is intended to resolve this reference by translating its field values into the substituted primitive reference which can be then used for direct access.

Context Stack: A stack of resolved references to parent objects (contexts). A reference on top of the context stack provides direct access to the target object and is obtained by resolving the last (low) segment of a complex reference. A reference at the bottom is obtained by resolving the first (high) segment of the complex reference.

Dual Methods: These have the same signature but different definitions in the object class and the reference class of a concept. Dual methods are used as usual by specifying their name and parameters.

Primitive Reference: A reference provided by the compiler depending on the run-time environment. Structure and functions of primitive references cannot be changed by the programmer and therefore it is assumed that they provide direct access to objects.

Reference Class: A class which is intended to describe structure and behavior of object identifiers. Its instances, called references, are passed by-value and indirectly represent objects by substituting for some primitive reference.

Concepts and Dynamics of the Application Service Provider Industry

Dohoon Kim
Kyung Hee University, Korea

INTRODUCTION: SOFTWARE AS A SERVICE

The enterprise intelligence through e-transformation is one of the cornerstones of the next-generation e-business era where the Internet constitutes the core business resource. Furthermore, the severe competitive landscape of e-business makes firms focus on their core capability and farm out staffing functions such as IT. Under this circumstance, enhancing intelligence and synergy through e-transformation will be accomplished by IT outsourcing via ASPs (application service providers). The ASP industry now provides an essential infrastructure for the Internet-based e-business transactions, thereby accelerating corporate e-transformation.

An ASP is generally defined as a third-party service firm that deploys, manages, and/or remotely hosts a software application through centrally located servers in a lease agreement. ASPs started their business by providing online application programs such as ERP (enterprise resource planning) and CRM (customer relationship management) solution packages to corporate customers. The first customers were small companies or local branches of multinational companies where IT outsourcing was the only option to deploy IT resources due to financial or regional constraints. As seen in these cases, the biggest merit of employing ASPs is that corporate customers do not have to own the applications and take responsibilities associated with initial and ongoing support and maintenance. Consequently, ASPs are differentiated from the existing IT services in that ASPs provide IT resources to multiple corporate clients on a one-to-many basis with a standardized service architecture and pricing scheme.

BACKGROUND: INDUSTRY VALUE CHAIN

The industry value chain does not allow a single service provider to control the entire service delivery process. Even if we confine our attention to the software delivery process in the value chain, the complexity does not reduce significantly. In order to deliver applications over the Internet, we need a mechanism to establish and maintain collaboration among independent functional divisions. Analysis of this nature of the value chain shows how the industry is likely to evolve and gives some insights into the strategic meaning of special types of convergence. In particular, we should point out two critical aspects of the value chain, which are required to survive in the market: a large customer base and stable relationship with other functional divisions. The structure of partnership among the players in the value chain is one of the major elements to classify emerging ASP business models. Table 1 summarizes key players in the ASP value chain.

There are a number of factors that are frequently cited as fueling or dashing the growth of the ASP market (Burris, 2001; Factor, 2002; Kim, 2002; Sparrow, 2003; Toigo, 2001). One of the striking characteristics observed so far is that immaturity of the industry is the most representative challenge in terms of the market factor: for example, the uncertainty as to whether existing and emerging ASPs are winning enough customers to validate an ASP business model for highly sophisticated enterprise applications. While some ASPs are gaining momentum with early adopters, there are many client companies that are unwilling to rent ERP applications due to the lack of trust in the industry itself in Korea (Kim & Choi, 2001). Moreover, it is security control and remote monitoring systems, SLA (service level agree-

Table 1. Key players in the ASP value chain model

- Software Vendors: including ISVs (independent software vendors), content providers (CPs), and so forth
- Network Infrastructure Providers: including telecommunication operators, ISPs (Internet service providers), and so forth
- Application Service Providers: as an intermediary or an organizer between software vendors and customers
- Individual and Corporate Customers: subscribers (end users) of the ASP services

Table 2. Drivers and challenges of the ASP industry

Category	Drivers	Challenges
Technology	◆ Reduce risk of technological obsolescence due to rapidly changing IT ◆ Provide a chance to utilize best-of-breed applications ◆ Avoid IT staffing shortage	◆ Unsolved security concerns ◆ Emerging, new technological requirements from the clients: e.g., SLA with client participation ◆ Unproved service reliability: e.g., network problems, system scalability and performance
Market	◆ Minimize up-front TCO (total cost ownership) ◆ Provide predictable cash flows	◆ Unproved client momentum ◆ Failure in giving clients sufficient trust due to unstable ASP industry

Table 3. ASP business models and capability profiles

Basic Types	Characteristics and Value-Added Components	Basic Capability
H-ASP (Horizontally Specialized ASP)	◆ Develop deep expertise within a given functional area (as opposed to one-stop shop): Substantial consulting services are possible ◆ ISV's need of partnership with systems integration and distribution companies ◆ Should be Web-based software provider ◆ Either own the software or develop proprietary integration in a specific field	◆ Well positioned to expand customer basis quickly ◆ Hard to copy the domain-specific knowledge
V-ASP (Vertically Specialized ASP)	◆ Industry-specific applications (in contrast to one-stop shop) ◆ Vertically oriented template methodology: easily deploy across multiple clients in the same industry	◆ Strong advantage in customized solutions ◆ Hard to copy the industry-specific knowledge
AIP (Application Infrastructure Provider)	◆ Originated from telecommunication company that owns networks and has operations experience ◆ Provide infrastructure management to ASPs ◆ Provide system management services including SLA ◆ Alleviate client concerns regarding network reliability, etc.	◆ High investment costs as an entry barrier: easy to protect their market share
XSP (Extended Service Provider)	◆ Provide total services from front end to back end with systems integration consulting ◆ Create new business process by rearranging suppliers and customers ◆ Help customers and even other service providers enter new markets, deploy services, and improve profitability easily while minimizing risk ◆ Build and integrate customized applications, thereby enabling clients to avoid the need to handle multiple ASP solutions	◆ Going back to one-stop-shop idea: Improved flexibility will be the core competitive edge for XSP

ment; Lee & Ben-Natan, 2002; Sturm, Morris, & Jander, 2000) management, and the global standardization process that should be further developed to support proliferation of ASPs. In the end will survive only a few successful ASPs that adapt themselves to the market requirements and take the most advantage of the competitive landscape.

ASP BUSINESS MODELS

The industry's short history raises the following questions. What changes will happen? Who will be the winners and losers? To answer these questions, Table 3 clarifies different types of the ASP business domains that are currently emerging. ASP's common value proposition to improve total benefits from IT outsourcing has been giving rise to various trials in designing the service delivery processes, each of which corresponds to a business model suggested in the table. Therefore, this classification plays a key role in identifying and analyzing the collaborative networking structure in the ASP value chains.

FUTURE TRENDS: INDUSTRY DYNAMICS AND EVOLUTION

The guiding principles of the industry evolution, which have been leading the industry to face proliferation of ASP business models, are summarized into (a) economies of scale through positive feedback from the market and (b) integration across the value chain for attaining cost reduction and differentiation.

First, it is the economies of scale or increasing return that serves as the core economic guiding principle for ASPs. A survey on the Korean IT outsourcing market reveals that, in terms of the TCO (total cost ownership) of a typical ERP package, IT outsourcing through ASPs enables clients to save roughly 20% of their up-front license fee and 80% of the implementation and maintenance service costs (Kim & Choi, 2001). Accordingly, ASPs that host these applications basically seek to lower this 80% portion of the TCO upon the notion of a one-to-many relationship between an ASP and its clients. An ASP is usually able to leverage standardized solutions across multiple clients. Attaining client momentum and reducing the overall costs per client are the major economic motivations for ASPs to compete with each other, thereby creating a positive feedback mechanism through network externality on the demand side. In sum, the competition keeps going for the expansion of a customer base or market share, which provides a good surrogate measure of profit for this case.

Second, the competitive landscape is also defined by the unique nature of a service system market where independent hardware and software resources are combined and reorganized into a new package in alignment with partners along the value chain and even a customer's business process. These offerings aim at designing a seamless and proprietary service delivery process to sharpen the competitive edge while raising the entry barrier. This essential feature of integration in the service delivery process makes the various possible business models reduce into the different types of service product combinations along the value chain as presented in Table 1. The integration, however, should be verified by achieving savings in TCO, though it is not easy to measure the amount of cost reduction by a certain partnership structure. Accordingly, the cutthroat competition fueled by business domain integration not only drives down the price to an acceptable market price, but also creates diverse market segmentations based on the service product differentiation.

Furthermore, increasing switching costs and rising entry barriers, two basic phenomena regarding the guiding principles and competitive landscape, are common to all the business models. As a result, efforts to penetrate into different market segments and build new customer relationships at the niche will inevitably run into strong resistance from the incumbents. Some events like technological breakthroughs will be required in order for a specific ASP model to consolidate another. Therefore, various business models will thrive over a period of time before some giant players in each business model emerge. Despite a unique coexistence of diverse ASP business models, some hypotheses on the industry structure change and evolution can be derived from those observations together with the guiding principles.

First, the general trend will be that the total number of ASPs in the industry will reduce since the customer base is not large enough to keep all the incumbent ASPs alive. Cash flows generated from the market give winners resilience to possible occasional failures and allow them to better manage risk by diversifying a portfolio of value components to open a new market niche. It is this kind of positive feedback loop (Arthur, 1989; Nelson & Winter, 1978) from the economies of scale that accelerates the exit of losers from the market and shapes the industry structure (Shy, 2002).

Second, the industry has been concentrating more around horizontally and vertically specialized ASPs than around the pure ASPs (that is, a simple partnership with an ISV). The primary concern of the emerging ASPs is to build some value-added components to the service architecture, thereby making it hard for competitors to replicate their business model and for customers to replace the current provider. However, reliance on third-party ISVs could make it more difficult to resolve underlying performance issues that have been the subject of customer scrutiny. On the other hand, looking up the capability profiles of the ASP business models, we can conclude that both H-ASPs and V-ASPs hold a dominant position from this standpoint. If some technical constraints such as SLA and security requirements come to rise to the surface, AIP will gain technological competitiveness since delegating the control of core enterprise applications to an external provider requires ASPs to prove their capability of reliable and stable operations.

Last, we predict that the rate-of-demand increase will affect the industry structure: the pattern of market segmentation, the market share of each ASP type, and so forth. The speed of the market expansion will affect ASPs' selection of competitive priorities.

CONCLUSION

The ASP industry will shape the future e-business transactions, providing a great flexibility in redeploying a firm's resources. Although the industry is currently at its early stage of the industry life cycle, much attention is now paid to vertical or domain-specific expertise and flexible capabilities in addition to the basic offerings. In order to assess both the market and the supply side, classified are emerging ASP business models together with some driving forces shaping the evolutionary path. Some careful observations disclosed that (a) the capability of an ASP model hinges on the differentiation of service products to a large degree and

(b) economies of scale play a key role in the dynamically evolving market mechanisms. ASPs that originally developed their proprietary solutions will be better positioned in terms of ultimate performance and scalability. Those ASPs will increase the chance to succeed in the market irrespective of how critical a given solution is to their client's day-to-day operations. Last, some technical factors that may affect the evolution path (for example, SLA regulation and security) should be considered in the near future.

REFERENCES

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *Economic Journal,* (99), 116-131.

Burris, A. M. (2001). *Service provider strategy: Proven secrets for xSPs.* Upper Saddle River, NJ: Prentice Hall.

Church, J., & Gandal, N. (2000, Spring). Systems competition, vertical merger and foreclosure. *Journal of Economics and Management Strategy*, (Vol. 9, pp. 25-51).

Factor, A. (2002). *Analyzing application service providers.* Upper Saddle River, NJ: Sun Microsystems Press.

Farrell, J., Monroe, H., & Saloner, G. (1998, Summer). The vertical organization of industry: Systems competition versus component competition. *Journal of Economics and Management Strategy*, (Vol. 7, pp. 143-182).

Harney, J. (2002). *Application service providers (ASPs): A manager's guide.* Boston, MA: Addison-Wesley.

Katz, M. L., & Shapiro, C. (1994, Spring). Systems competition and network effects. *Journal of Economic Perspectives,* (Vol. 8, pp. 93-115).

Kim, D. (2002). ASP and collaborative network infrastructure for global enterprise intelligence: An explanatory approach to identify prerequisites and challenges. In J. Chen (Ed.), *Global supply chain management* (pp. 166-170). Beijing, China: International Academic Publication.

Kim, M. S., & Choi, Y. C. (2001). *The current status of ASP market development.* Korea Information Society Development Institute, Seoul, Korea.

Lee, J. J., & Ben-Natan, R. (2002). *Integrating service level agreements: Optimizing your OSS for SLA delivery*. Indianapolis, Indiana: Wiley.

Nelson, R. R., & Winter, S. G. (1978). Force generating and limiting concentration under Schumpeterian competition. *Bell Journal of Economics,* (9), 524-548.

Shy, O. (2002). *The economics of network industries.* New York: Cambridge University Press.

Sparrow, E. (2003). *Successful IT outsourcing: From choosing a provider to managing the project.* London: Springer.

Sturm, R., Morris, W., & Jander, M. (2000). *Foundations of service level management.* Indianapolis, Indiana: Sams.

Toigo, J. W. (2001). *The essential guide to application service providers.* Upper Saddle River, NJ: Prentice Hall.

Zuscovitch, E., & Justman, M. (1995). Networks, sustainable differentiation, and economic development. In D. Batten, J. Casti, & R. Thord (Eds.), *Networks in action* (pp. 269-285). New York: Springer-Verlag.

KEY TERMS

AIP: An application infrastructure provider (AIP) is a type of ASP, which is usually originated from telecommunication operators that run their own networks and IDCs (Internet data centers). AIP focuses on server hosting and network infrastructure management for other ASPs and corporate clients, and provides value-added services based on its technology leadership, for example, online security and e-payment services.

ASP: An application service provider (ASP) is a third-party service firm that deploys, manages, and/or remotely hosts a software application through centrally located servers in a lease agreement.

Economies of Scale: Economies of scale are the achievement of lower average cost per unit through increased production, or the decrease in the marginal cost of production as a firm's extent of operations expands.

Horizontal ASP: Horizontal ASPs provide online applications for a specific business function such as human resource management, procurement, customer relations, and so forth.

IT Outsourcing: IT outsourcing is the outsourcing of enterprise information systems and management to computer manufacturers or software companies (the term outsourcing stems from using an outside resource). Companies can save purchasing cost, maintenance cost, and labor cost by outsourcing and paying for those services. Outsourcing has become a common practice in the US where companies are faced with uncertain returns of massive investment in IT resources.

SLA: A service level agreement (SLA) is a contract between a supplier and a customer that identifies (a) services supported at each of three layers—application, host (system), and network—(b) service parameters for each service, (c) levels of service quality, and (d) liabilities on the part of the supplier and the customer when service quality levels are not met.

Vertical ASP: Vertical ASPs provide online applications customized for a specific industry such as staff or operations scheduling for hospitals, material procurement for the steel industry, and so forth.

Value Chain: A value chain is a chain of activities in a group of collaborators who are designed to meet market demand. They are vendors involved in value chains across purchasing, procurement, manufacturing, warehousing, distribution, and sales of components, equipments, raw materials, and so forth to manage a series of resource and information flow.

XSP: An extended service provider (XSP) provides total IT-related services from online applications through maintenance and reengineering of IT resources to business process consulting for its clients. Success of the XSP model should presume the rapid proliferation of the ASP services in an overly complex market. If the ASP service demand grows explosively in a short period of time, the XSP model will debut in the market earlier on, increasing the possibility of XSPs dominating the industry as they have scale advantage in terms of cost.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 514-518, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Conceptual Commonalities in Modeling of Business and IT Artifacts

Haim Kilov
Stevens Institute of Technology, USA

Ira Sack
Stevens Institute of Technology, USA

INTRODUCTION

The proverbial communication gap between business and IT experts is mostly due to the fact that what is considered obvious to some business experts might not be obvious or even known to IT experts, and might substantially differ from what is considered obvious to other business experts. Thus, different stakeholders may understand the business domain and problems in tacit and quite different ways, and at the same time might be unaware of these differences. This leads to business-IT misalignment, and therefore to many serious information system failures, from life-threatening to financial or simply very annoying (loss of customers' trust and patience).

The article provides a concise overview of the topic, includes definitions of some essential concepts and constructs together with industrial examples of their use in modeling and in fostering business-IT alignment, and shows how a small subset of UML has been successfully used to represent the essential structure of a model.

BACKGROUND

Creating business and IT system models readable and understandable to all stakeholders is possible only if the system of concepts underlying such models is well-defined, understandable to the model *readers*, and does not require extensive and painstaking explanations. Because the experiences of different stakeholders are usually quite different, the fundamental underlying concepts should be invariant with respect to the specific business (or IT) domain of interest.

Fortunately, a *system* of simple and elegant *abstract* modeling concepts and constructs exists and has been stable for centuries. Precise definitions of its semantics come from exact philosophy, mathematics, programming, and systems thinking. It has been successfully used in theory, in industrial practice (including international standards such as the Reference Model of Open Distributed Processing (ISO/IEC, 1995)), and in teaching of business and IT modeling. It includes such generic concepts as system, model, abstraction, structure, relationship, invariant, state, action, behavior, conformance, type, composition, template, name in context, and so forth, thus providing a solid foundation for systems of appropriate domain-specific concepts, such as contract, trade, confirmation, derivatives, options trade, and so forth, for the financial domain.

Why are we reusing concepts from exact philosophy? As Mario Bunge observed, "all factual sciences, whether natural, social, or mixed, share a number of philosophical concepts…and a number of philosophical principles" (Bunge, 2001). Moreover, philosophers "won't remain satisfied with examples [...]; they will seek general patterns" (Bunge, 2004). The work of such outstanding systems thinkers as Mario Bunge and F.A. Hayek includes clear and concise definitions and descriptions of such concepts and of some fundamental patterns which are essentially the same as those formulated—independently—in the best IT-based sources. Specifically, the concept of a *relation* is indispensable for understanding a system:" so long as the elements... are capable of acting upon each other in the manner determining the structure of the machine, their other properties are irrelevant for our understanding of the machine" (Hayek, 1952), and "the structure of a system is the set of all the relations among its components, particularly those that hold the system together" (Bunge, 2003). The same kinds of generic relations hold together very different systems, thus providing a solid foundation for bridging the communication gap between business and IT experts.

E.W. Dijkstra observed decades ago (Dijkstra, 2007) that "it is the sole purpose of the specifications to act as the interface between the system's users and the system's builders. The task of "making a thing satisfying our needs" as a single responsibility is split into two parts– "stating the properties of a thing, by virtue of which it would satisfy our needs" and "making a thing guaranteed to have the stated properties." These considerations apply both to the "business side" and to the "IT side" of any application development (or purchase) project.

Some examples of simple and elegant generic specifications understandable to their readers include the relational data model (Codd, 1970), certain high-level programming languages (Wirth, 1995), and the Macintosh user interface. The underlying implementations may have been complex,

but it was of no importance to the users of these systems: the "internals" were not exposed. However, in too many instances businesses have been unnecessarily restricted by "requirements" imposed by inadequate IT systems. As a well-known example, consider restrictions imposed on data by IT systems that require a set of predefined manual codes (without any business meaning) and prohibit the use of business-specific aliases. These restrictions—a typical example of business-IT misalignment—still result in serious losses for businesses, although excellent IT-based solutions to these problems were described more than 30 years ago, for example, in Gilb and Weinberg (1977).

A SYSTEM OF REUSABLE ABSTRACT CONCEPTS

A business model ought to be abstract enough to be understood, and therefore, while being precise, it should not be excessively detailed. But it also *cannot* be too detailed: as observed by Hayek, in any complex system "of life, mind and society" it is possible to "determine only the general character of the resulting order and not its detail." The purpose of a (high-level) model of a complex domain is to "bring about an abstract order – a system of abstract relations – concrete manifestations of which will depend on a great variety of particular circumstances which no one can know in their entirety" (Hayek, 1985). Such a model is essential for strategic decision making; and because tactical and operational decision making are determined by strategic decisions, such a model becomes essential for any kind of business decision.

As early as 1605, Bacon noted that "amongst so many great foundations of colleges in Europe, I find strange that they are all dedicated to professions, and none left free to Arts and Sciences at large." This is what "systems thinking" is about, and it has been around, under different names, for millennia. It is based on mathematics and exact philosophy–areas of human endeavor that have also been around for millennia (see, for example, the eloquent presentation in (Russo, 2004)). Furthermore, if the business stakeholders did not state the requestes properties of the IT "thing" (also sometimes known as "business rules") for any reason–for example, they were never asked, "everyone knew" what these properties were supposed to be, or it was not known "how to ask"– then the developers make these properties up and often specified them only in their code (very often unreadable to anyone except–prehaps–the developers themselves), so that as a result the IT thing has an important but not too useful property "it does what it does".

William Kent presents (Kent, 1978) an approach in which asking apparently trivial questions (like "What is an information systems thing?" "What is a Real World thing?" "What does 'the same thing' mean?" "What is a name?" etc.) leads to substantial clarification of a business domain and of business problems. This is an essential component of the framework for business-IT alignment because understanding and articulation of questions and problems is much more important than answering or finding solutions: "deeper understanding of the real features of a problem ... is an essential prelude to its correct solution" (Johnstone, 1977). Understanding a problem is possible only when the business domain for that problem is demonstrably understood in the same manner by all stakeholders, in order for the (inevitably) different viewpoints of different stakeholders to be conceptually compatible. More specific questions asked about a business domain, such as "what is a bank?" and "what is a margin call?" proposed by Dines Bjørner (Bjørner, 2006), lead to understanding of a business domain and articulating that understanding. As one of many examples of failures due to inadequate understanding, consider *The Spectator*'s assessment (Vincent, 2007) of hedge fund modeling: incomplete models for which invariants were not made explicit and too much was left outside as tacit assumptions led to very expensive failures of quantitative hedge funds due to violations of these assumptions: "shares were moving in ways that hadn't been programmed into the computer models." No wonder that, as a result, *Financial Times* (Davies & Tett, 2007) described the "Flight to simplicity" and understandability requested by users or potential users of financial instruments who wanted to return to "old fashioned banking" sketched in the article in exactly the same manner as in Dunbar (1901); for a simple example, a credit card is conceptually the same as a letter of credit a century ago.

Clearly, using abstraction is a prerequisite for understanding and for separating the concerns of various business and IT stakeholders. But this is not sufficient: As observed by F.A. Hayek: "Until we have definite questions to ask we cannot employ our intellect; and questions presuppose that we have formed some provisional hypothesis or theory about the events" (Hayek, 1985). A domain model, or its fragment, is just such a provisional theory. If it does not exist yet for the specific domain of interest, we may try to use appropriate generic models that almost always do exist; and if no generic models exist, then we can try to compose such a model out of appropriate business patterns.

This concept of business patterns is not new at all. Adam Smith eloquently presented it about 250 years ago in *The Theory of Moral Sentiments*:

When a number of drawings are made after one pattern, though they may all miss it in some respects, yet they will all resemble it more than they resemble one another; the general character of the pattern will run through them all; the most singular and odd will be those which are most wide of it; and though very few will copy it exactly, yet the most accurate delineations will bear a greater resemblance to the most careless, than the careless ones will bear to one another.

Relationships—from fundamental such as composition to business-specific such as foreign exchange option trade—are well-known examples of business patterns.

The concepts of a structure in general and of a relationship in particular are central in analysis and understanding of any system, be it a business or an IT one; moreover, they are the basis of business-IT alignment. Fortunately, there are only a few generic relationships, and their definitions—coming from exact philosophy and mathematics—have been around for centuries, if not millennia. For example, in the third century BC, the Stoics explicitly distinguished between different kinds of composition based on different ways of emergent property determination (Russo, 2004).

Property determination is the essential aspect of characterizing the semantics of different types of interesting generic relationships.

The *composition* relationship is defined as "[a] combination of two or more [items] yielding a new [item], at a different level of abstraction. The characteristics of the new [item] are determined by the [items] being combined and by the way they are combined" (ISO/IEC, 1995). Whereas in some cases the values of emergent properties may be easily determined by a computer-based system (e.g., the total number of pages in a book composed of chapters), in more interesting cases the values of the emergent properties can be determined only by human valuation, which is always subjective (Mises, 1949) (e.g., the abstract of such a book, or a binary-valued emergent property "do I want to buy it?"). The article (Davies & Tett, 2007) also refers to inadequate determination of emergent properties of the exotic and not very transparent financial instruments, presumably due to inadequate models. Different kinds of composition help us to understand business systems better and to articulate that understanding: for example, the distinction between a traditional and a modern corporation (or a traditional and modern industry) (Drucker, 2001) may be illuminated as the distinction between a hierarchical and a nonhierarchical composition of parts of that corporation (Kilov, 2002).

A *subtyping* relationship is that between a supertype and its subtypes, such that each instance of a subtype has all properties of its supertype and some additional (subtype-specific) properties. For example, a foreign exchange option trade contract is a subtype of a derivative contract which, in turn, is a subtype of a trade contract which, in turn, is a subtype of a contract.

A *reference* relationship is that between a reference and maintained items, such that some properties of the maintained item are determined by the properties of its reference item. As an example (based on Bjørner, 2006), a seller's action is subtyped into a seller's follow-up and delivery to an unprepared customer, while the seller's follow-up (that may be subtyped further) has a reference relationship to the customer's confirmed order.

The three generic relationships are analogous to the three control structures in classical programming: sequence ("like" reference), selection ("like" subtyping), and iteration ("like" composition).

It is important to observe that, firstly, these relationships have been used to define both traditional businesses and IT systems, and, secondly, that the participants in these relationships need not be only "things" but may also be actions (also known as "processes" or "steps"). Thus, the structure of any system, of any (business or IT) domain, or any (business or IT) process may be defined using the same underlying fundamental constructs; generic relationships with precisely defined semantics. In particular, changing a business process (i.e., a partially ordered composition of steps), including (sub)components of the change implemented by means of IT systems, may be understood, specified, and explained to all interested parties using these constructs.

A business or IT model should be simple to read, but creating it is not easy at all. While modeling ought to be based on a solid theoretical foundation (some fragments of which were sketched above), "...the task of recognizing the presence in the real world of the conditions corresponding to ... our theoretical schemes is often more difficult than the theory itself... We cannot state simple, almost mechanical criteria by which a certain type of theoretical situation can be identified, but we have to develop something like a sense for the physiognomy of events" (Hayek, 1969).

The most important characteristics of a graphical language chosen to represent models or requirements are its simplicity and well-defined semantics. With respect to simplicity, the best approach is to follow the advice of E.W. Dijkstra for programming languages:

"...all knowledge of one's programming language can conveniently be formulated on a single sheet of paper and can be gotten across the limelight in about twenty minutes lecturing... only then real difficulties of understanding and solving real problems can be dealt with; that activity requires the ability to think effectively more than anything else (Dijkstra, 2007).

The same considerations apply to modeling languages, and even to a greater extent because the backgrounds of all programmers include (or are at least supposed to include) the same system of common fundamental concepts, while the backgrounds of business and IT stakeholders may be very different. The notation-independent and buzzword-invariant system of fundamental concepts used in systems thinking and described—independently!—by such apparently diverse authors as Bunge (2003, 2004), Codd (1979), Hayek (1969), Hoare (1975), Smith and Smith (1977) ,Weinberg (1982) and others, as well as in RM-ODP and in modeling texts such as Kilov (2000) and Kilov (2002), may be (and has been) explained to and understood by business and IT

stakeholders without any problems. With respect to well-defined semantics, too many diagrams "...appeared precise, but the meaning of the boxes and arrows was fuzzy: it was quite easy for two people to draw the same diagram with very different intent" (Parnas, 2001). Thus, in choosing the subset of UML (Unified Modeling Language) to represent understandable business and IT models, it is essential to concentrate on a very small, simple, and powerful subset of language constructs (not restricted by any technological considerations) having a well-defined semantics. Such a small subset of UML, with relationship semantics precisely defined as described above, exists and became an OMG standard: See www.omg.org/cgi-bin/doc?formal/2004-02-07.

FUTURE TRENDS

The problem of business-IT alignment has been around for decades and still has not been successfully solved. However, with the substantially increasing importance of all-pervasive information technology, a critical mass of users request simple, elegant, and usable IT systems. It has become acceptable and even fashionable to speak about going "back to basics," and the need for simplicity in IT emphasized by such computing science founders as Dijkstra and Hoare became acknowledged both by businesses and by some IT stakeholders: "Simplicity is the unavoidable price which we must pay for reliability" (Hoare, 1975).

In a market economy, the demand for a currently scarce type of a product or service, such as understandable IT systems, leads to additional design and development of products or services of that type. The strive for simplicity, together with appearance of elegant IT products and services (such as the well-known Google®, iPod®, iPhone®, and others) described not only in industrial publications but also in the *Wall Street Journal*, *The Economist*, *The Spectator*, and so forth, demonstrates that in future users, perhaps starting with consumers, but eventually—and soon—also business enterprises, will have more and better information technology choices. The foundations for such developments are already there.

CONCLUSION

The system of reusable abstract concepts, based on well-known ideas from mathematics, exact philosophy, and systems thinking and discussed here, provides a sound basis for understandable, complete, and rigorous models of businesses or IT systems. These models are based on the semantics of the appropriate domain rather than on existing or planned systems, products, or solutions. Therefore, they are an excellent—and essential—framework for business-IT understanding and alignment, including decision making demonstrably based on a solid foundation.

A business domain model clearly defines the structure of the domain. It is an abstract and precise road-map of a fragment of a business or IT system, or of a product, with appropriate refinements. Such business models have been successfully used for making demonstrably justified strategic, tactical, and operational decisions in all kinds of business and IT system environments, and thus for successful business-IT alignment (Kilov & Sack, 2007).

The reference list presents a lot of interesting and useful material for further reading. Textbooks like Kilov (2002) and Morabito, Sack, and Bhate (1999) provide a good starting point for becoming a good analyst able to write models understandable and usable by all stakeholders.

REFERENCES

Bjørner, D. (2006). *Software engineering* (Vols. 1-3). Springer-Verlag.

Bunge, M. (2001). *Philosophy in crisis: The need for reconstruction*. Amherst, NY: Prometheus Books.

Bunge, M. (2003). *Philosophical dictionary* (enlarged ed.). Amherst, NY: Prometheus Books.

Bunge, M. (2004). *Emergence and convergence: Qualitative novelty and the unity of knowledge.* University of Toronto Press.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377-387.

Codd, E. F. (1979). Extending the database relational model to capture more meaning. *ACM Transactions on Database Systems, 4*(4), 397-434.

Davies, P. J., & Tett, G. (2007, October 22). "A flight to simplicity:" Investors jettison what they do not understand. *Financial Times*.

Dijkstra, E.W. (2007). *E.W. Dijkstra archive*. Retrieved May 31, 2008, from http://www.cs.utexas.edu/users/EWD/welcome.html

Drucker, P. (2001). The next society. *The Economist, 8246*(361).

Dunbar, C. F. (1901). *Chapters on the theory and history of banking* (2nd ed., enlarged and edited). O. M. W. Sprague (Ed.). New York London: G.P. Putnams Sons.

Gilb, T., & Weinberg, G. (1977). *Humanized input*. Winthrop.

Hayek, F. A. (1952). *The sensory order*. London: Routledge & Kegan Paul.

Hayek, F. A. (1969). *Studies in philosophy, politics, and economics*. New York: Simon and Schuster.

Hayek, F. A. (1985). *New studies in philosophy, politics, economics and the history of ideas*. London: Routledge and Kegan Paul.

Hoare, C. A. R. (1975, June). Data reliability. In *Proceedings of the International Conference on Reliable Software, ACM SIGPLAN Notices*, (pp. 528-33).

ISO/IEC. (1995). *Open distributed processing—reference model, Part 2: Foundations* (ITU-T Recommendation X.902 | ISO/IEC 10746-2).

Johnstone, P. T. (1977). *Topos theory*. London: Academic Press.

Kent, W. (1978). *Data and reality*. North Holland. (Reprinted by 1st Books, 2000).

Kilov, H. (2000). Representing business specifications in UML. In K. Baclawski & H. Kilov (Eds.), *Proceedings of the 9th OOPSLA Workshop on Behavioral Semantics*, (pp. 102-111). Boston: Northeastern University.

Kilov, H. (2002). *Business models*. Prentice Hall.

Kilov, H., & Sack, I. (2007). Mechanisms for communication between business and IT experts. *Computer Standards and Interfaces*.

Mises, L., von. (1949). *Human action: A treatise on economics*. New Haven, CT: Yale University Press.

Morabito, J., Sack, I., & Bhate, A. (1999). *Organization modeling: Innovative architectures for the 21st century*. Prentice Hall.

Parnas, D. L. (2001). Software design. In D. M. Hoffman & D. M. Weiss (Eds.), *Software fundamentals*. Addison-Wesley.

Russo, L. (2004). *The forgotten revolution: How science was born in 300 BC and why it had to be reborn.* Springer-Verlag.

Smith, J. M., & Smith, D. C. P. (1977). Database abstractions: Aggregation and generalization. *ACM Transactions on Database Systems, 2*(2).

Vincent, M. (2007, October 10). When computers go crazy. *The Spectator.*

Weinberg, G. M. (1982). *Rethinking systems analysis and design*. Boston, Toronto: Little, Brown and Company.

Wirth, N. (1995). A plea for lean software. *IEEE Computer, 28*(2), 64-68.

KEY TERMS

Analysis: "breaking down a whole into its components and their mutual relations" (Bunge, 2003).

Composition: A relationship between a composite and several components, such that some (emergent) properties of the composite are determined by the components and by the way they are combined.

Generic Relationship: Composition, subtyping, or reference.

Reference: A relationship between a reference and maintained items, such that some properties of the maintained item are determined by the properties of its reference item.

Reference Model of Open Distributed Processing (RM-ODP): an international standard defining the semantics of fundamental concepts and constructs used for specification of any system independently of a specific methodology, technology, or tool.

Relationship: A collection of items together with the invariant referring to the properties of the items.

Subtyping: A relationship between a supertype and its subtypes, such that each instance of a subtype has all properties of its supertype and some additional—subtype-specific—properties.

Type: A predicate (ISO/IEC, 1995).

ENDNOTES

[a] Independent Consultant, and Stevens Institute of Technology (haimk@acm.org; hkilov@stevens.edu)

[b] Stevens Institute of Technology (isack@stevens.edu)

Consistent Queries over Databases with Integrity Constraints

Luciano Caroprese
DEIS Universitá della Calabria, Italy

Cristian Molinaro
DEIS Universitá della Calabria, Italy

Irina Trubitsyna
DEIS Universitá della Calabria, Italy

Ester Zumpano
DEIS Universitá della Calabria, Italy

INTRODUCTION

Integrating data from different sources consists of two main steps, the first in which the various relations are merged together, and the second in which some tuples are *removed* (or *inserted*) from the resulting database in order to satisfy integrity constraints. There are several ways to integrate databases or possibly distributed information sources, but whatever integration architecture we choose, the heterogeneity of the sources to be integrated causes subtle problems. In particular, the database obtained from the integration process may be inconsistent with respect to integrity constraints, that is, one or more integrity constraints are not satisfied. Integrity constraints represent an important source of information about the real world. They are usually used to define constraints on data (functional dependencies, inclusion dependencies, etc.) and have, nowadays, a wide applicability in several contexts such as semantic query optimization, cooperative query answering, database integration, and view update.

Since the satisfaction of integrity constraints cannot generally be guaranteed, if the database is obtained from the integration of different information sources, in the evaluation of queries, we must compute answers that are consistent with the integrity constraints. The following example shows a case of inconsistency.

Example 1: Consider the following database schema consisting of the single binary relation *Teaches (Course, Professor)* where the attribute *Course* is a key for the relation. Assume there are two different instances for the relations *Teaches*, D1={(c1,p1),(c2,p2)} and D2={(c1,p1),(c2,p3)}. The two instances satisfy the constraint that *Course* is a key, but from their union we derive a relation that does not satisfy the constraint since there are two distinct tuples with the same value for the attribute *Course*.

In the integration of two conflicting databases simple solutions could be based on the definition of preference criteria such as a partial order on the source information or a majority criterion (Lin & Mendelzon, 1996). However, these solutions are not generally satisfactory, and more useful solutions are those based on (1) the computation of "repairs" for the database, and (2) the computation of consistent answers (Arenas, Bertossi, & Chomicki, 1999).

The computation of repairs is based on the definition of minimal sets of insertion and deletion operations so that the resulting database satisfies all constraints. The computation of consistent answers is based on the identification of tuples satisfying integrity constraints and on the selection of tuples matching the goal. For instance, for the integrated database of *Example 1*, we have two alternative repairs consisting in the deletion of one of the tuples *(c2,p2)* and *(c2,p3)*. The consistent answer to a query over the relation *Teaches* contains the unique tuple *(c1,p1)* so that we do not know which professor teaches course *c2*. Therefore, it is very important, in the presence of inconsistent data, not only to compute the set of consistent answers, but also to know which facts are unknown and if there are possible repairs for the database.

BACKGROUND

Several proposals considering the integration of databases as well as the computation of queries over inconsistent databases have been provided in the literature (Agarwal, Keller, Wiederhold, & Saraswat, 1995; Arenas et al., 1999; Arenas, Bertossi, & Chomicki, 2000; Bry, 1997; Dung, 1996; Greco & Zumpano, 2000; Lin, 1996; Lin & Mendelzon, 1996; Lembo, Lenzerini, & Rosati, 2002; Lenzerini, 2002; Wijsen, 2003). Most of the techniques for computing queries over inconsistent databases work for restricted cases, and only

recently have there been proposals to consider more general constraints. This section provides an informal description of the main techniques proposed in the literature.

- Lin and Mendelzon (1996) proposed an approach taking into account the majority view of the knowledge bases in order to obtain a new relation that is consistent with the integrity constraints. The technique proposes a formal semantics to merge first order theories under a set of constraints.
  However, the "merging by majority" technique does not resolve conflicts in all cases since information is not always present in the majority of the databases, and, therefore, it is not always possible to choose between alternative values. Moreover, the use of the majority criteria involves discarding inconsistent data and hence the loss of potentially useful information.
- Arenas et al. (1999) introduced a logical characterization of the notion of consistent answer in a possibly inconsistent database. The technique is based on the computation of an equivalent query $T_\omega(Q)$ derived from the source query Q. The definition of $T_\omega(Q)$ is based on the notion of residue developed in the context of semantic query optimization.
  More specifically, for each literal B, appearing in some integrity constraint, a residue *Res(B)* is computed. Intuitively, *Res(B)* is a universal quantified first order formula that must be true, because of the constraints, if B is true.
  The technique, more general than the previous ones, has been shown to be complete for universal binary integrity constraints and universal quantified queries. However, the rewriting of queries is complex since the termination conditions are not easy to detect and the computation of answers generally is not guaranteed to be polynomial.
- Arenas et al. (2000) proposed an approach consisting in the use of a Logic Program with exceptions (LPe) for obtaining consistent query answers. An LPe is a program with the syntax of an extended logic program (ELP), that is, in it we may find both logical (or strong) negation ($\neg$) and procedural negation (not). In this program, rules with a positive literal in the head represent a sort of general default, whereas rules with a logically negated head represent exceptions. The semantic of an LPe is obtained from the semantics for ELPs, by adding extra conditions that assign higher priority to exceptions. The method, given a set of integrity constraints ICs and an inconsistent database instance, consists in the direct specification of database repairs in a logic programming formalism. The resulting program will have both negative and positive exceptions, strong and procedural negations, and disjunctions of literals in the head of some of the clauses, that is, it will be a disjunctive extended logic program with exceptions. As shown by Arenas et al. (1999), the method considers a set of integrity constraints, IC, written in the standard format $\vee_{i=1}^{n} P_i(x_i) \vee \vee_{i=1}^{m} (\neg Q_i(y_i) \vee \varphi$, where φ is a formula containing only built-in predicates, and there is an implicit universal quantification in front. This method specifies the repairs of the database, D, that violate IC, by means of a logical program with exceptions, Π^D. In Π^D, for each predicate P a new predicate P' is introduced, and each occurrence of P is replaced by P'.
  The method can be applied to a set of domain independent binary integrity constraints *IC*, that is, the constraint can be checked w.r.t. satisfaction by looking to the active domain, and in each *IC* appear at most two literals.
- Calì, Calvanese, De Giacomo, and Lenzerini (2002), Lembo et al. (2002), and Lenzerini (2002) proposed a framework for data integration that allows to specify a general form of integrity constraints over the global schema, and it is defined a semantics for data integration in the presence of incomplete and inconsistent information sources. Moreover, it is defined as a method for query processing under the previous semantics when key constraints and foreign key constraints are defined upon the global schema.
  Formally, a data integration system I is a triple $<G, S, M_{G,S}>$, where G is the global schema, S is the source schema, and $M_{G,S}$ is the mapping between G and S. More specifically, the *global schema* is expressed in the relational model with both key and foreign key constraints; the *source schema* is expressed in the relational model without integrity constraints; and the *mapping* is defined between the global and the source schema, that is, each relation in G is associated with a view, that is, a query over the sources. The semantics of a data integration system is given by considering a source database D for I, that is, a database for the source schema S containing relation r^D for each source r in S.
  Any database G is a *global database* for I, and it is said *legal* w.r.t. D if:
  - It satisfies the integrity constraints defined on G.
  - It satisfies the mapping w.r.t. D, that is, for each relation r in G, the set of tuples r^B that B assigns to r is a subset of the set of tuples $\rho(r)^D$ computed by the associated query $\rho(r)$ over D: $\rho(r)^D \subseteq r^B$.

  In this framework, the semantics of I w.r.t. a source database D, denoted $sem^D(I, D)$, is given in terms of a set of databases. In particular, $sem^D(I, D) = \{ B \mid B$ *is a legal global database for I, w.r.t. D}*. If $sem^D(I, D) \neq \emptyset$, then I is said to be consistent w.r.t. D.

In this setting, a query *q* posed to a data integration system *I* is a conjunctive query over the global schema, whose atoms have symbols in *G* as predicates. A tuple $(c_1, .., c_n)$ is considered an answer to the query only if it is a *certain* answer, that is, if it satisfies the query in every database that belongs to the semantics of the data integration system.

The *retrieved global database*, denoted by *ret(I, D)*, is obtained by computing for each relation *r* of the global schema r^D; the query $\rho(r)$ is then evaluated the query over the source database *D*. Note that the *retrieved global database* satisfies all the key constraints in *G*, as it is assumed that $\rho(r)$ does not violate the key constraints; thus, if *ret(I,D)* also satisfies the foreign key constraints, then the answer to a query *q* can be done by simply evaluating it over *ret(I,D)*. If it is the case that *ret(I,D)* violates the foreign key constraints, then tuples have to be added to the relations of the global schema in order to satisfy them.

Obviously in general there are an infinite number of legal database that are coherent with the retrieved global database, even if it is shown that there exists one, the *canonical database*, denoted *can(I,D)*, that represents all the legal databases that are coherent with the retrieved global database. Thus formally the answer to a query q can be given by evaluating can(I, D). Anyhow, the computation of the canonical database is impractical, as generally the database can be infinite; thus, Calì et al. (2002) defined an algorithm that computes the certain answers of a conjunctive query q without actually building can(I, D).

- Wijsen (2003) proposed a general framework for repairing databases. In particular the author stressed that an inconsistent database can be repaired without deleting tuples (*tuple-based* approach), but using a finer repair primitive consisting in correcting faulty values within the tuples, without actually deleting them (*value-based* approach).

Example 2: Suppose to have the following set of tuples reporting the dioxin levels in food samples:

Sample	Sample Date	Food	Analysis Date	Lab	DioxinLevel
110	Jan 17, 2002	poultry	Jan 18, 2002	ICI	normal
220	Jan 17, 2002	poultry	Jan 16, 2002	ICB	alarming
330	Jan 18, 2002	beef	Jan 18, 2002	ICB	normal

Dioxin Database

and the constraint:

$$\forall s, d_1, f, d_2, l, d(\text{Dioxin}(s, d_1, f, d_2, l, d) \rightarrow d_1 \leq d_2)$$

that imposes the date of analyzing a given sample cannot precede the date the sample was taken.

The first tuple in the Dioxin Database says that the sample 110 was taken on January 17, 2002, and analyzed the day after at the ICI lab, and that the dioxin level of this sample was normal. While the sample 110 respects the constraint, the sample 220 violates it. An inconsistency is present in the database, and the author claims to "clean" it in a way that avoids deleting the entire tuple, that is, acting at the attribute level and not at the tuple level.

Given an inconsistent database, a consistent answer can be obtained by letting the database in its inconsistent state, and by propagating in the answer the consistent portion of the database, that is, the set of tuples matching the query and satisfying the constraints. As the repair work is deferred until query time, this approach is called *late-repairing*. In this framework an alternative technique is proposed consisting in a *database transformation*. Given a satisfiable set of constraints Σ, that is, a set of finite constraints, and a relation I, apply a database transformation $h_\Sigma : I \rightarrow I$ such that for every query Q, $Q(h_\Sigma(I))$ yields exactly the consistent answer to Q on input I and Σ.

Observe that $h_\Sigma(I)$ is not necessarily a repair for I and Σ; it can be thought of as a "*condensed representation*" of all possible repairs for I and Σ that is sufficient for consistent query answering. The practical intuition is that an inconsistent database I is first transformed through h_Σ in such a way that the subsequent queries on the transformed database retrieve exactly the consistent answer; since databases are modified prior to query execution, this approach is called *early-repearing*.

Clearly for a given set of satisfiable constraints Σ, early and late repairing should yield the same set of consistent answers, hence $f_\Sigma(Q)(I)=Q(h_\Sigma(I))$, for every query and every relation.

A NEW TECHNIQUE FOR QUERYING AND REPAIRING INCONSISTENT DATABASES

Greco, Greco, and Zumpano (2001, 2003) and Greco and Zumpano (2000) proposed a general framework for computing repairs and consistent answers over inconsistent databases with universally quantified variables. The technique is based on the rewriting of constraints into extended disjunctive rules with two different forms of negation (negation as failure and classical negation). The disjunctive program can be used for two different purposes: compute "repairs" for the database, and produce consistent answers, that is, a maximal set of

atoms that do not violate the constraints. The technique is sound and complete (each stable model defines a repair, and each repair is derived from a stable model) and more general than techniques previously proposed.

Specifically, the technique is based on the generation of an extended disjunctive program LP derived from the set of integrity constraints. The repairs for the database can be generated from the stable models of *LP*, whereas the computation of the consistent answers of a query *(g,P)* can be derived by considering the stable models of the program $P \cup LP$ over the database *D*.

Let *c* be a universally quantified constraint of the form:

$$\forall X [B_1 \wedge ... \wedge B_k \wedge \text{not } B_{k+1} \wedge ... \wedge \text{not } B_n \wedge \phi \supset B_0]$$

then, *dj(c)* denotes the extended disjunctive rule

$$\neg B'_1 \vee ... \vee \neg B'_k \wedge B'_{k+1} \vee ... \vee B'_n \vee B'_0 \leftarrow (B_1 \vee B'_1), \ldots, (B_k \vee B'_k), (\text{not } B_{k+1} \vee \neg B'_{k+1}), \ldots, (\text{not } B_n \vee \neg B'_n), \phi, (\text{not } B_0 \vee \neg B'_0),$$

where B'_i denotes the atom derived from B_i, by replacing the predicate symbol *p* with the new symbol p_d if B_i is a base atom otherwise is equal to false. Let *IC* be a set of universally quantified integrity constraints, then $DP(IC) = \{ dj(c) \mid c \in IC \}$, whereas *LP(IC)* is the set of standard disjunctive rules derived from *DP(IC)* by rewriting the body disjunctions.

Clearly, given a database *D* and a set of constraints, *IC*, $LP(IC)_D$ denotes the program derived from the union of the rules *LP(IC)* with the facts in *D*, whereas $SM(LP(IC)_D)$ denotes the set of stable models of $LP(IC)_D$, and every stable model is consistent since it cannot contain two atoms of the form *A* and $\neg A$. The following example shows how constraints are rewritten.

Example 3: Consider the following integrity constraints:

$\forall X [p(X) \wedge \text{not } s(X) \supset q(X)]$
$\forall X [q(X) \supset r(X)]$

and the database D containing the facts p(a), p(b), s(a), and q(a).
The derived generalized extended disjunctive program is defined as follows:

$$\neg p_d(X) \vee s_d(X) \vee q_d(X) \leftarrow (p(X) \vee p_d(X)) \wedge (\text{not } s(X) \vee \neg s_d(X)) \wedge (\text{not } q(X) \vee \neg q_d(X)).$$
$$\neg q_d(X) \vee r_d(X) \leftarrow (q(X) \vee q_d(X)) \wedge (\text{not } r(X) \vee \neg r_d(X)).$$

The previous rules can now be rewritten in standard form. Let *P* be the corresponding extended disjunctive Datalog program. The computation of the program P_D gives the following stable models:

$M_1 = D \cup \{ \neg p_d(b), \neg q_d(a) \},$

$M_2 = D \cup \{ \neg p_d(b), r_d(a) \},$

$M_3 = D \cup \{ \neg q_d(a), s_d(b) \},$

$M_4 = D \cup \{ r_d(a), s_d(b) \},$

$M_5 = D \cup \{ q_d(b), \neg q_d(a), r_d(b) \}$ and

$M_6 = D \cup \{ q_d(b), r_d(a), r_d(b) \}.$

A (generalized) extended disjunctive Datalog program can be simplified by eliminating from the body rules all literals whose predicate symbols are derived and do not appear in the head of any rule (these literals cannot be true). As mentioned before, the rewriting of constraints into disjunctive rules is useful for both (1) making the database consistent through the insertion and deletion of tuples, and (2) computing consistent answers leaving the database inconsistent.

FUTURE TRENDS

As a future trend, an interesting topic consists in specifying preference criteria so that selecting among a set of feasible repairs the preferable ones, that is, those better conforming to the specified criteria. Preference criteria introduce desiderata on how to update the inconsistent database in order to make it consistent; thus they can be considered as a set of desiderata that are satisfied *if possible* by a generic repair. Therefore, informally a preferred repair is a repair that better satisfies preferences. Preliminary results have been published by Greco, Sirangelo, and Trubitsyna (2003).

CONCLUSION

In the integration of knowledge from multiple sources, two main steps are performed, the first in which the various relations are merged together, and the second in which some tuples are removed (or inserted) from the resulting database in order to satisfy integrity constraints.

The database obtained from the merging of different sources could contain inconsistent data. In this article we investigated the problem of querying and repairing inconsistent databases. In particular we presented the different techniques for querying and repairing inconsistent databases (Agarwal et al., 1995; Arenas et al., 1999; Greco & Zumpano, 2000; Lin & Mendelzon, 1996).

REFERENCES

Arenas, M., Bertossi, L., & Chomicki, J. (1999). Consistent query answers in inconsistent databases. *PODS Conference* (pp. 68-79).

Arenas, M., Bertossi, L., Chomicki, J. (2000). Specifying and querying database repairs using logic programs with exceptions. *FQAS Conference* (pp. 27-41).

Argaval, S., Keller, A. M., Wiederhold, G., & Saraswat, K. (1995). Flexible relation: An approach for integrating data from multiple, possibly inconsistent databases. *ICDE Conference* (pp. 495-504).

Baral, C., Kraus, S., Minker, J., & Subrahmanian, V. S. (1999). Combining knowledge bases consisting of first order theories. *ISMIS Conference* (pp. 92-101).

Bry, F. (1997). Query answering in information system with integrity constraints. In *IFIP WG 11.5 Working Conf. on Integrity and Control in Inform. System* (pp. 113-130).

Calì, A., Calvanese, D., De Giacomo, G., & Lenzerini, M. (2002). Data integration under integrity constraints. *CAiSE02 Conference* (pp. 262-279).

Dung, P. M. (1996). Integrating data from possibly inconsistent databases. *CoopIS Conference* (pp.58-65).

Grant, J., & Subrahmanian, V. S. (1995). Reasoning in inconsistent knowledge bases. *TKDE Conference* (Vol. 7, 177-189).

Greco, G., Greco, S., & Zumpano, E. (2001). A logic programming approach to the integration, repairing and querying of inconsistent databases. *ICLP Conference* (pp. 348-364).

Greco, G., Greco, S., & Zumpano, E. (2003). A logical framework for querying and repairing inconsistent databases. *TKDE* (15, 1389-1408).

Greco, G., Sirangelo C., Trubitsyna I., & Zumpano, E. (2003). Preferred repairs for inconsistent databases. *IDEAS Conference* (pp. 202-211).

Greco, S., & Zumpano, E. (2000). Querying inconsistent databases. *LPAR Conference* (pp. 308-325).

Lembo, D., Lenzerini, M., & Rosati, R. (2002). Incompleteness and inconsistency in information integration. *KRDB Conference,* Toulouse, France.

Lenzerini, M. (2002). Data integration: A theoretical perspective. *PODS* (pp. 233-246).

Lin, J. (1996). A semantics for reasoning consistently in the presence of inconsistency. *Artificial Intelligence, 86*(1), 75-95.

Lin, J., & Mendelzon, A. O. (1996). Merging databases under constraints. *Int. Journal of Cooperative Information Systems, 1*(7), 55-76.

Wijsen, J. (2003). Condensed representation of database repair for consistent query. *ICDT* (pp. 378-393).

KEY TERMS

Consistent Answer: A set of tuples, derived from the database, satisfying all integrity constraints.

Consistent Database: A database satisfying a set of integrity constraints.

Data Integration: A process providing a uniform integrated access to multiple heterogeneous information sources.

Database Repair: Minimal set of insert and delete operations that makes the database consistent.

Disjunctive Datalog Program: A set of rules of the form:

$$A_1 \vee \ldots \vee A_k \leftarrow B_1, \ldots, B_m, not\ B_{m+1}, \ldots, not\ B_n, \quad k+m+n>0,$$

where $A_1,\ldots, A_k, B_1,\ldots, B_n$ are atoms of the form $p(t_1,\ldots, t_h)$, p is a predicate symbol of arity h, and the terms $t_1,\ldots, t_h$ are constants or variables.

Inconsistent Database: A database violating some integrity constraints.

Integrity Constraints: A set of constraints that must be satisfied by database instances.

Constructionist Organizational Data Mining

Isabel Ramos
Universidade do Minho, Portugal

João Álvaro Carvalho
Universidade do Minho, Portugal

INTRODUCTION

Scientific or organizational knowledge creation has been addressed from different perspectives along the history of science and, in particular, of social sciences. The process is guided by the set of values, beliefs, and norms shared by the members of the community to which the creator of this knowledge belongs, that is, it is guided by the adopted paradigm (Lincoln & Guba, 2000). The adopted paradigm determines how the nature of the studied reality is understood, the criteria that will be used to assess the validity of the created knowledge, and the construction and selection of methods, techniques, and tools to structure and support the creation of knowledge. This set of ontological, epistemological, and methodological assumptions that characterize the paradigm one implicitly or explicitly uses to make sense of the surrounding reality is the cultural root of the intellectual enterprises. Those assumptions constrain the accomplishment of activities such as construction of theories, definition of inquiry strategies, interpretation of perceived phenomena, and dissemination of knowledge (Schwandt, 2000).

Traditionally, social realities such as organizations have been assumed to have an objective nature. Assuming this viewpoint, the knowledge we possess about things, processes, or events that occur regularly under definite circumstances, should be an adequate representation of them. Knowledge is the result of a meticulous, quantitative, and objective study of the phenomenon of interest. Its aim is to understand the phenomenon in order to be able to anticipate its occurrence and to control it.

Organizations can instead be understood as socially constructed realities. As such, they are subjective in nature since they do not exist apart from the organizational actors and other stakeholders. The stable patterns of action and interaction occurring internally and with the exterior of the organization are responsible for the impression of an objective existence. The adoption of information technology applications can reinforce or disrupt those patterns of action and interaction, thus becoming key elements in the social construction of organizational realities (Lilley, Lightfoot, & Amaral, 2004; Vaast & Walsham, 2005).

BACKGROUND

The Rational and Emotional Nature of Personal Knowledge

Individual knowledge is actively constructed by the mind of the learner (Kafai & Resnick, 1996).

We make ideas instead of simply getting them from an external source. Idea making happens more effectively when the learner is engaged in designing and constructing an external artifact, which is meaningful for the learner, and he or she can reflect upon it and share it with others. From this constructionist description of the learning process, we can emphasize several elements associated with the creation of knowledge, namely, *cognition*, *introspection*, *action*, *interaction*, and *emotion*.

Through *cognitive* processes, humans construct mental representations of external and mental objects. *Introspection* is a specific type of cognition that permits the personal inquiry into subjective mental phenomena such as sensory experiences, feelings, emotions, and mental images (Damásio, 1999; Wallace, 2000). Through *action* and *interaction*, we create our experiences of the world we live in. The effective construction of personal knowledge requires the building of relationships between concepts and other mental constructs, in profoundly meaningful experiences (Shaw, 1996). All human experience is mediated by *emotions*, which drive our attention and concentration in order to help us to process external stimuli and to communicate with others.

The Historical and Socio-Cultural Context of Knowledge

A social reality is a construction in continuous reformulation that occurs whenever social actors develop social constructions that are external and sharable.

By the mere fact that people interact, influencing each other's mental constructs, social reality is in constant reconstruction. In this context, learning of new concepts and practices is happening continuously, either intentionally or unintentionally.

Learning happens inside specific mental and social spaces, meaning that what a group can learn is influenced by:

- The concepts, schemata, values, beliefs, and other mental constructs shared by the group.
- All knowledge we create about external things, events, and relationships is based on and constrained by our mental constructs and the tools we use to experience that external world.
- The creation of knowledge is founded on the historical and socio-cultural context of its creators, providing a shared basis for the interaction inside a group. The continuous interaction of the group members, happening in a common environment, leads to similar mental constructs, a common interpretation of events, and the creation of shared meaning structures and external constructions, such as new tools that change how the external word is experienced.
- There is no viewpoint outside human subjectivity or historical and socio-cultural circumstances from which to study phenomena and to judge the inquiry process and the knowledge produced.

ODM AND KNOWLEDGE CREATION: PROBLEMS AND OPPORTUNITIES

ODM (organizational knowledge discovery) has been defined as the process of analyzing organizational data from different perspectives and summarizing them into useful information for organizational actors who will use that information to increase revenues, reduce costs, or achieve other relevant organizational goals and objectives (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Matheus, Chan, & Piatetsky-Shapiro, 1993).

Data mining is a sub-process of the knowledge discovery. It leads to the finding of models of consumer behavior that can be used to guide the action of organizational actors. The models are built upon the patterns found out among data stored in large databases that are backed by statistical correlations among that data. Those patterns are extracted by specific mechanisms called data mining algorithms.

Attached to the discourse around the data mining tools, there is the idea that in the future, new and more powerful algorithms will be developed that will be capable of finding more valuable patterns and models, independently from human subjectivities and limitations. If it ever becomes possible to integrate the knowledge of the relevant business domain into the system, the algorithm would be able to decide about the usefulness and validity of discovered patterns, correlations, and models, as well as to grow in sophistication by integrating these models in its knowledge of the business. The decision-making process would become extensively automated and guided by the objective reasoning of clear and rational rules implemented in a computer-based system.

However, this view has several drawbacks, namely:

1. Since all human knowledge has a tacit and non-expressible dimension, it will never be possible to integrate all relevant business knowledge in a repository to be analyzed by a data-mining algorithm.
2. The diversity of views about the business activities and their context is what allows for the emergence of organizational creativity and development and the challenge of taken-for-granted concepts and practices (Bolman & Deal, 1991; Morgan, 1997; Palmer & Hardy, 2000). The stored knowledge representations are those around which there is some degree of consensus. This is important for the stability of work concepts and practices and to support organizational cohesion. However, they may also trap organizational actors in those concepts and practices, even when evidence shows they are threatening organizational success.
3. The relevance of knowledge representations stored in organizational repositories changes according to changes in the socio-cultural circumstances that offer the context for making sense of the representations. Only the organizational actors can understand those contexts and are able to give meaning to knowledge representations.
4. It is still believed that decision-making is or should be an essentially rational process, guided by cognitive processes such as planning, resolution of problems, and creativity (Sparrow, 1998). However, recent experiments in neurobiology show that emotion is an integral part of reasoning and decision-making (Damásio, 1999). Thus, only organizational actors can make decisions. The full automation of the process is not a realistic objective.

Instead of the present focus on the technological side of ODM, it would be interesting to adopt a constructionist approach and to focus on the social process of knowledge construction that makes ODM meaningful. With this new focus on people and the way they create and share knowledge, the main concern would be to mobilize the knowledge of organizational actors so the whole organization can benefit from it. This concern is justified by the awareness that the organization, seen as a community, is more intelligent than each one of its members, including any of its leaders.

LEVERAGING KNOWLEDGE CREATION IN ORGANIZATIONS: SOME CONSTRUCTIONIST GUIDELINES FOR ODM

With ODM there is a special focus on knowledge about consumer behavior to support decision and action. ODM

Table 1. A summary of constructionist guidelines for ODM

Using data mining tools	Creating rich learning environments
Data mining results will support insight and creativity when organizational actors have enough time to reflect upon them, to change their actions accordingly, to learn with the consequences of their own and others' actions.	Work relationships must be strengthened in order to create the social cohesiveness needed for the ongoing production of shared constructions that engage the organization in developmental cycles.
Effective formal and informal communication must be fostered in order for it to become possible to discuss each others' interpretations of past and present experience in the light of perceived enduring changes in organizational agents' work context.	The construction of knowledge about customers' preferences and their future needs and reactions must be guided by the shared purposes of the specific communities of practice that constitute the organization.
Data mining tools may be used to seek dissonance between the usual representations of consumer behavior and the evidence of actual behavior. This dissonance usually triggers the need to reestablish consonance by adopting more effective work practices.	Organizational repositories, data mining tools, and the results of data mining are social artifacts that should be used to make ideas tangible, to negotiate meanings, and to facilitate communication between organizational actors.
The search and interpretation of patterns and models of consumer behavior should be guided by a multi-dimensional knowledge of the business domain, and work concepts and practices.	Knowledge representations were created and stored under specific historical and socio-cultural circumstances of which their readers must be aware in order to be able to understand relevance or inadequacy of those representations.

assists the organization in knowing the preferences of its customers and in anticipating their needs and reactions. The construction of this knowledge must be guided by the specific purposes of the several communities of practice that constitute the organization.

ODM and the knowledge it helps to create are social constructions. Repositories, data mining tools, and the resulting patterns, correlations, and models are social artifacts that should be used to make ideas tangible, to negotiate meanings, and to facilitate communication between organizational actors. As such, they may become catalysts for the development of shared knowledge about consumer behavior, when they are used in the contexts of meaningful projects.

Data mining systems may become empowering tools in the sense that they make viable the analysis of large organizational repositories of knowledge representations. These knowledge representations are social constructions that connect organizational actors to a common view of the business concepts and practices that shape their intentions and interactions. Problems in the performance of organizational tasks or in organizational adaptation to environmental changes may reside in the inappropriateness of knowledge representations or in the tools used to extract rules and patterns from them. Knowledge representations were created and stored under specific historical and socio-cultural circumstances of which their readers must be aware in order to be able to understand their relevance or inadequacy.

Table 1 summarizes the constructionist guidelines for ODM, grouping them in two categories:

- guidelines that should be considered for the creation of rich learning environments in which data mining systems are used as social artifacts that leverage continuous learning, and
- guidelines that should be considered when using a specific data mining tool.

These guidelines are given from constructionist theories developed and applied in areas such as psychology, education, and organization theory.

FUTURE TRENDS

According to the assumptions of the constructionist perspective, ODM should be designed to involve organizational actors in the social construction of something external and sharable. The designing of a marketing campaign, the making of a decision, and the transformation of work representations and practices are examples of social construction processes for which ODM could be viewed as relevant.

As a result of the process, the individual and shared knowledge will become more sophisticated, empowering the action of individuals and groups, and facilitating interaction. In this way, organizational actors consciously create cohesive and pluralist work environments, more prone to deal with problems and difficult decisions associated with consumer behavior. This perspective is more realistic than the traditional view of ODM as a process of making knowledge

neutral and independent of the knower and social contexts in which is created, in order to support decision-making processes idealized as inherently rational.

The tools used to support ODM fundamentally shape and define the process. Lack of appropriate tools impoverishes a social setting and makes social construction difficult. Future research is needed to study how current data mining tools support (1) the transformation of work practices and representations concerning consumer behavior and (2) the definition of effective practices to deal with opportunities and threats perceived in actual consumer behavior. It will also be important to create practical experiences of designing and implementing the ODM process in specific organizational settings so that learning from a constructionist perspective can be supported.

CONCLUSION

This article describes ODM as a process for the social construction of knowledge. As such, the focus changes from the technology used to discover patterns in the stored data to the human and social issues surrounding knowledge creation in organizations.

Managers should provide the resources and the conditions for the emergence of rich learning environments in which data repositories and data mining tools sustain collective cognitive processes such as memory, reasoning, language, and attention. In this way ODM becomes a key organizational process in the construction of organizational representations of external realities. These representations will guide organizational decision and action. In accordance with this view, the article provides a summary of constructionist guidelines for ODM to help managers leveraging knowledge creation in organizations.

REFERENCES

Bolman, L. G., & Deal, T. E. (1991). *REFRAMING ORGANIZATIONS: Artistry, choice, and leadership*. San Francisco: Jossey-Bass Publishers.

Damásio, A. (1999). *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery: An overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy (Eds.), *Advances in knowledge discovery and data mining* (pp. 1-34). Cambridge, MA: The MIT Press.

Kafai, Y., & Resnick, M. (Eds.). (1996). *Constructionism in practice: Designing, thinking, and learning in a digital world*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lilley, S., Lightfoot, G., & Amaral, P. (2004). *Representing organization: Knowledge, management, and the information age*. Oxford, UK: Oxford University Press.

Lincoln, Y. S., & Guba, E. G. (2000). Paradigmatic controversies, contradictions, and emerging confluences. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 163-188). Thousand Oaks, CA: Sage Publications.

Matheus, C. J., Chan, P. K., & Piatetsky-Shapiro, G. (1993). Systems for knowledge discovery in databases. *IEEE Transactions on Knowledge and Data Engineering, 5*(6), 903-913.

Morgan, G. (1997). *Images of organization*. Thousand Oaks, CA: Sage Publications.

Palmer, I., & Hardy, C. (2000). *Thinking about management*. London: Sage Publications.

Schwandt, T. A. (2000). Three epistemological stances for qualitative inquiry: Interpretivism, hermeneutics, and social constructionism. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 189-213). Thousand Oaks, CA: Sage Publications.

Shaw, A. (1996). Social constructionism and the inner city: Designing environments for social development and urban renewal. In Y. Kafai & M. Resnick (Eds.), *Constructionism in practice* (pp. 175-206). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Sparrow, J. (1998). *Knowledge in organizations: Access to thinking at work*. London: SAGE Publications.

Vaast, E., & Walsham, G. (2005). Representations and actions: The transformation of work practices with IT use. *Information and Organization, 15*(1), 65-89.

Wallace, B. A. (2000). The taboo of subjectivity: Toward a new science of consciousness. New York: Oxford University Press.

KEY TERMS

Constructionism: A set of theories that defines the human beings as active constructors of their own learning and development. This learning and development of knowledge happens more effectively when individuals are involved in the construction of something external, something that can be shared, or both.

Objective Social Reality: It has an independent existence from any account of it.

Objectivism: A set of theories that views true knowledge about external realities, and the process of its creation, as neutral and independent of the knowledge creator.

Rich Learning Environment: A learning environment in which the learner is empowered to create a strong connection with the reality of interest by directly experiencing with it in order to develop mental constructs that are deep, complex, pluralist, and emotionally rich.

Socially Constructed Reality: It is created through purposeful human action and interaction. This reality is shaped by the individual's subjective conceptual structures of meaning. It is reconstructed by the human interactions that support continuous reinterpretations and change of meanings. The social institutions are the means through which meanings are stabilized and the social reality assumes an objective appearance.

Social Constructions: External and sharable concepts, associations, artifacts, and practices that people actively develop and maintain in their social settings. An organization is an example of a social construction that interconnects its members in a specific social setting, in which many other social constructions are continuously being developed and maintained.

Constructivism in Online Distance Education

Kathaleen Reid-Martinez
Azusa Pacific University, USA

Linda D. Grooms
Regent University, USA

Mihai C. Bocarnea
Regent University, USA

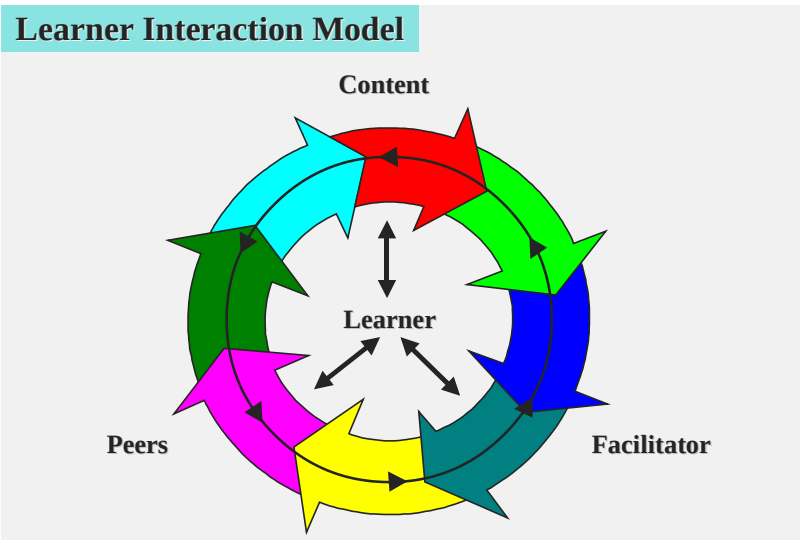
INTRODUCTION

The past two decades have ushered in a very pronounced gravitation toward a constructivist approach to teaching and learning in all realms of society and most particularly in the online distance education environment. Augmenting communication in and among those in the academic, business, and military communities, the exponential advancement of science and technology has availed vast amounts of information to virtually millions of people around the globe. In conjunction with this knowledge explosion has been a growing concern for the democratization of the learning process, with constructivism driving much of the educational agenda. This article examines the resurgence of this approach to teaching and learning, its convergence with rapidly changing technological advances, and how it forecasts future trends in online pedagogy.

BACKGROUND

While the constructivist method has been highly emphasized in the more recent literature (Jonassen, Davidson, Collins, Campbell, & Haag, 1995; Rovai, 2004; Tenenbaum, Naidu, Jegede, & Austin, 2001), it is not a new approach to learning. Presenting an early example, Socrates facilitated discourse with students asking directed questions to assist them in realizing the weaknesses in their logic and critical thinking. This enabled them to share in the responsibility of their learning through active participation while negotiating meaning in the creation of shared understanding. In contrast, over time, most professors in Western culture often served as primary repositories of information along with the scrolls and velum texts found in the limited number of physical libraries available to educators. This role included the important function of disseminating information, as well as assisting students in

Figure 1. © 2000, Grooms, L.D.



shaping and forming that knowledge. The lecture served as the quickest and easiest way to reach both small and large groups of individuals.

While the lecture method was the norm of information delivery for centuries in Western culture, the knowledge explosion of the 20th century demanded more active learner participation. In light of this constant and rapid flux of information and knowledge, students became lifelong learners compelled to use metacognitive skills to constantly evaluate and assimilate new material into their respective disciplines. As this implies, knowledge was no longer viewed as a fixed object; rather, learners constructed it as they experienced and co-created an understanding of various phenomena by collaborating and working with peers and professors as well as with the information. Based on the work of Kidd (1973), Long (1983), Moore (1989), and Palmer (1993), Grooms' (2000) Learner Interaction Model (see Figure 1) illustrates that in the constructivist culture, the learner perpetually interacts with these three components of learning.

Now, rather than strictly acquiring information, Duffy and Cunningham (1996) explicated that "learning is an active process of constructing…knowledge and…instruction is a process of supporting that construction" (p. 171). Critical in this process is recognizing the shifting role of the professor who becomes the *guide on the side* or content facilitator and is no longer the proverbial *sage on the stage* or content provider. The student's role also has changed from being a passive receiver of information to an active participant in the knowledge-making process (Weller, 1988), aligning with Bandura's (1977, 1994) concept of the autonomous learner, an important dimension of the constructivist model. Table 1 delineates these two approaches to learning.

Of special interest in Table 1 is the role of community. The constructivist approach recognizes that students do not learn strictly within the limited confines of an educational institution, but rather within the broader context of their personal lives. Consequently, the boundaries between the educational institution and the larger community become blurred, creating a unique set of challenges.

As people work collaboratively in the learning activities, they bring their own worldviews and experiences to each situation, often creating a plethora of perspectives. During this collaborative learning process, they must negotiate and generate meaning and solutions to problems through shared understanding. Thus, education moves from a single, solitary pursuit of knowledge to a collaborative learning community that shapes and informs responses to the environment. As noted by Fuller and Söderlund (2002), this challenges the common metaphor of the university as a self-contained village.

RAPIDLY CHANGING DISTANCE LEARNING TECHNOLOGIES

Over the years, educators have experimented with and successfully employed multiple media for distance learning. As early as the 18th century, print material was used and even today still serves an important role in distance education.

Table 1

Approaches to Learning		
	Traditional	**Constructivist**
Professor	Sage on the Stage	Guide on the Side
	Content Provider	Content Facilitator
Learner	Passive Recipient	Active Participant
Knowledge	Fixed Object	Fluid
Organization of Learning	Ordered & Structured	Open & Often Chaotic
Communication	Uni-directional	Multi-directional
Primary Resource	Text & Professor	Multiple Sources
Method	Lecture	Active Process
Media	Print	Blended
Format	Individualized	Collaborative
Activities	Goal-oriented	Problem-centered
Focus of Learning	Knowledge & Understanding	Application, Analysis, Synthesis, & Evaluation
Assessment	Recall	Alternative Assessment
Community	Educational Institution	Integrated with Life

After the 1930s, other media became significant with audio, including radio and audiotapes, and video, including public broadcasting, satellite, and cable, dominating much of the 20th century.

Much of this education was one-way based on a mass communication or one-to-many educational model. Basically, it was a rigid structure with information flowing in one direction, from the powerful and knowledgeable instructor reaching to a large group of students. It included elements of limited feedback through the use of such things as the penny post in the 19th century and the addition of telephone and fax in the 20th century. Limited opportunities for face-to-face interaction were also incorporated with some programs. Thus, much of distance learning during these times remained mainly non-interactive.

By the 1990s, the advent of the Internet presented new opportunities in distance education. The result was the evolution of a new type of collaborative learning, in which the potential for interaction between the professor and the learner increased exponentially with wide-area networks accommodating synchronous and asynchronous communication. While exploring computer-mediated activities of the online learning environment, Santoro (1996) highlighted three broad categories: (a) computer-assisted instruction, which allows the computer to serve as teacher by structuring information delivered to the human user, (b) computer-based conferencing, which includes e-mail, interactive messaging, and group conference support systems, and (c) informatics, which refers to online public access libraries and interactive remote databases. This proliferation of the Internet unlocked the door for educational institutions to reach beyond their four walls, making services accessible to students around the world through online activities.

In their work with the U.S. Department of Education, Waits and Lewis (2003) reported that in the 2000 to 2001 academic year, 56% of the nation's degree-granting 2- and 4-year institutions offered courses at a distance with another 12% planning to do so within the next year. Although the communication technologies of the 21st century—print, audio, video, and the Internet—cover a broad spectrum of distance education mediums, this exponential growth in science and technology has catapulted the Internet into rapidly becoming the preferred delivery platform. Researchers such as Cotton (1995) and others who have been tracking this information over the last decade along with scholars such as Bocarnea, Grooms, and Reid-Martinez (2006) continue to explore not only the trends in distance education, but also the understanding of and the issues involved in aligning the environment with student needs. Typical factors include (a) the characteristics of the discipline, (b) the degree of interactivity sought in the distance learning process, (c) learner characteristics, (d) instructor traits, (e) the expansiveness of the distance education initiative, (f) the desired level of accessibility and flexibility of the delivery platform, (g) the availability of technical support, and lastly (h) the potential for growth.

In addition to the global reach of the Internet, the lines *among* communication technologies have swiftly blurred. Today, in the convergence of technologies, computers, telephones, and cameras are no longer distinct entities, but can be found bundled into one small handheld gadget through the fusion of technology (McCain & Jukes, 2001). Through this fusion, communicating with students and colleagues has become more integrated, vastly expanding the means of feedback.

With such rapid technological advances, today's educators are dropped into what Jacque Ellul (1964) described as the intersection of tension between humanity and technology. This struggle with the latent and manifest, and intended and unintended consequences of technology exists as students and professors wrestle with new Web-enabled devices that seem to expand learning media at exponential rates. These include, but are not limited to, online courseware and portals, interactive media (edutainment), video and telephone conferencing, podcasting, vodcasting, phonecasting, instant messaging, blogging, moblogging, linklogging, vlogging, and photologging.

This new technology facilitates greater flexibility and customization in the learning process. For example, combining metatagging and object tagging with some of these media automates the process in highly specialized ways that create cost and human resource efficiencies. Other advances supporting increased efficiency in communities of practice include knowledge management, learning objects, and electronic performance support systems (EPSS).

Instructional designer Don Morrison (2004) demonstrated how the aforementioned learning channels can be established within parameters and policies that most appropriately align with the primary strengths and weaknesses of each medium. He noted that among others, cost, time constraints, delivery speed, and infrastructure help determine appropriate application. Morrison's work also pointed to ways in which educational models can be designed to marry traditional and online means of moving from the simple to the more complex methods of learning.

Citing that the mix of traditional and online learning depends upon the student, the context, the available channels, and the time constraints in which the education takes place (Morrison, 2004), one such marriage is blended learning. Although the typical example of blended learning evokes images of traditional classes combined with an online component, this is not the only alternative. While images of the traditional precipitate thoughts of coaching and mentoring, collegial relationships, seminar participation, and workshops, and the virtual environment conjures visions of referencing manuals and online communities, any combination of these means constitute blended delivery.

From blended to distance learning, these new electronic forms of communication have forced a paradigm shift in education. This move is most avidly seen in distance learning, where even the terminology has shifted from distance education to online or e-learning. This new term more clearly indicates the way in which learners can use computer-mediated hypertext multimedia communication to easily collaborate in a continuous integration of knowledge and social capital.

FUTURE TRENDS

As previously discussed, the rapid growth of technology continues to herald unprecedented opportunity for distance learning, and when wed to a constructivist approach, it presents opportunities for online pedagogy that can transcend traditional modes of education. From this marriage emerges three primary factors that define the new online pedagogy: (a) community building through networks that cross time and geographic boundaries, (b) structure through technology that manages, and (c) collaborative opportunities for shared knowledge and wisdom in response to the complexities of a global society (Reid-Martinez, 2006).

Community

As Bocarnea et al. (2006) note, today's technologies launch a new paradigm of online learning and pedagogy, which has the potential to be communal in nature. Primarily, these technologies allow for interaction between students and professors, students and peers, and the broader community in unprecedented ways. For example, Wojnar (2002) and Young (2002) confirm that students in Web-based courses have greater instructor access through e-mail and e-learning platforms than they would often have in traditional lecture halls.

In fact, Young (2002) highlights the differences between the boundaries embedded in his traditional face-to-face class and that which he encounters online. This suggests that guidelines and boundaries following good business practices are essential to prevent online instructors and students from feeling overwhelmed by the 24-7 opportunity for interaction. As this suggests, these technologies provide opportunities for networking and building strong virtual learning communities that can transcend geographic boundaries and extend far beyond the duration of the student's formal education.

In addition, this communal nature of the virtual learning environment provides opportunity for students to bring their local community context into the learning experience in direct ways as well as immediately allowing them to apply what they have learned through their study. For example, students in leadership programs can be employed full-time in leadership positions and take their learning experiences directly into their work environment through well-designed course assignments. The professor is no longer someone whom the student must wait to see in class later in the week, but rather is readily available in the e-learning platform to serve as consultant and mentor as the student applies the principles studied that week. The professor has become the "guide on the side." This triangulation of student-professor-content points to the need for well-designed learning experiences developed from a constructivist perspective to meet the challenges and needs of today's students. Indeed, uniting the new technology with this approach appears to be a marriage made in heaven for contemporary students working in a rapidly changing and highly demanding global environment.

Unfortunately, this opportunity for such rich learning experiences is not always the case. For example, Schweizer (1999) noted that many online courses have been cited as unsound, lacking pedagogical clarity and adequate design. While some scholars such as Wang and Newlin (2002) make certain that pedagogy guides the design of their online courses, all too often technology rather than pedagogy has driven the computer-mediated learning experience instead of simply serving as the delivery mechanism.

In fairness to contemporary educators, the rapid advancement of technology creates a moving-target challenge for course developers who often find themselves reacting to the technological advances rather than proactively establishing the technology's relationship to the learning process. As Bocarnea et al. (2006) observe, theory typically "follows technology in desperate attempts to describe the impact of an already existing and rapidly fading…technological reality" (p. 385).

This suggests that staying focused on strategy and content design remains the dominant challenge for curriculum developers. Online pedagogy, the science of and about online education, provides perspective to assist the developers in focusing and maintaining the balance necessary for creating excellent online learning experiences. Furthermore, online pedagogy allows designers to be proactive as they can be in constant reflection and introspection about the technological processes involved in online learning.

Structure

Heralded just over 14 years ago by Negroponte (1995), the information age is collapsing on itself as the amount of online information is becoming unbearable. After the scramble to have everything digitized, the primary challenge today is how to create meaningful knowledge from such massive amounts of data. The quality of knowledge in contrast to quantity drives the heart of this concern. In light of this overload, structure is essential to online knowledge development.

Related to the structure is the development of open-source initiatives. While most often understood as software that is

open for use and modification by the public, the phrase has become a recognized attribute ascribed to multiple endeavors, such as knowledge-building. The open-source nature of online initiatives pushes a new model for managing learning and knowledge-building through the communal process. It allows diverse individuals from various locations to combine information from multiple sources into distributed knowledge networks. Through this open-source structure, participants interact to share experiences and knowledge, thereby expanding their awareness of new concepts and differing approaches to problem solving as they modify the information in the open-source environment and redistribute it back to fellow participants (Bocarnea et al., 2006).

As this suggests, through interaction, participants build complex webs of knowledge in the open-source cyberspace. The technology provides the structure to create and maintain webs of knowledge, and it also grants ease of access globally to those in the public interested in that knowledge. In the process, knowledge is given away to others who in turn begin to use it in multiple ways while beginning the next evolution of knowledge development as they add to and transform the knowledge base they accessed through the open-source structure. With this transformation is the transference of power and control that becomes less centrist and more distributed globally.

Collaborative Knowledge-Building

As noted above, interaction is the key for the development of open-source knowledge-building. While scholars such as Cederblom and Paulsen (2001) posit learning as a behavioral change, others hold that it is simply when learners meet needs and establish goals for attaining knowledge (Ponton & Carr, 2000). Referring to this process as an implied contract, Keirns (1999), along with the above scholars, suggests that if online learning is used, structure is critical to allow students to advance in their knowledge.

The design of the online course becomes the principal structure to assure learners' goals are achieved. In that course design, structured interaction must be at the core. Not only does this interaction provide for knowledge development, but Blair (2002) claims that online threaded discussions can serve as primary places for faculty to prompt students' critical thinking. Furthering this thought in comparison with on-campus courses, she cites that online learners exemplify a greater degree of reflection and depth in their questions. This is likely because in most cases, online learners have more time to process and research their responses. Moreover, the delay in feedback in the asynchronous learning environment also allows learners to be less inhibited, to take greater chances, and thus to offer more in-depth analysis (Smith, Ferguson, & Caris, 2002).

One way of establishing structured interaction is through required dialogue participation (Klemm, 1998). As earlier noted, the online classroom creates greater access for participation (Cummings, 1998; Wojnar, 2002; Young, 2002), which Blair (2002) suggests often forges stronger relationships due to increased interaction frequency. This increased interaction also relates to higher learner commitment due to the socialization the learner goes through to be a participant in the knowledge-building process. Thus, learner perception of the degree of interaction plays an important role in student achievement, satisfaction, and course quality (Roblyer & Ekhaml, 2000).

Again, in the collaborative nature of the constructivist online culture, interaction perpetually occurs between learners and content, learners and instructors, and learners and peers, with each type of interaction reinforcing and fostering collaborative knowledge-building for both the learner and the faculty. With this in mind, online course designers must decide how to best structure courses capitalizing on this collaborative interaction.

CONCLUSION

As the above suggests, the advent of online learning as a new form of distance education has not just provided opportunity to disseminate information in a new medium, but it has radically adjusted the distance learning paradigm in terms of distribution methods, community-building, and pedagogy. The use of 21st century technology is rapidly closing the gap of the communication immediacy essential in developing communities of practice for knowledge-building. With their open-source networks, these new technologies encourage and actively support constructivist pedagogy in the new distance education paradigm. Most of all, distance education can now fulfill its greatest potential, which is to reach every learner who desires to participate in the knowledge-building process. The result is a democratization of education not previously seen, allowing for shifts in power and control throughout societies.

REFERENCES

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behavior* (Vol. 4, pp. 77-81). New York: Academic Press.

Blair, J. (2002). The virtual teaching life. *Education Week, 21*, 31-35.

Bocarnea, M. C., Grooms, L. D., & Reid-Martinez, K. (2006). Technological and pedagogical considerations in online learning. In A. Schorr & S. Seltmann (Eds.), *Changing*

media markets in Europe and abroad: New ways of handling information and entertainment content (pp. 379-392). New York: Pabst Science Publishers.

Cederblom, J., & Paulsen, D. W. (2001). *Critical reasoning: Understanding and criticizing arguments and theories* (5th ed.). Belmont, CA: Wadsworth.

Cotton, C. (1995). Time-and-place independent learning: The higher education market for distance learning emerges. *Syllabus, 8*(5), 37-39.

Cummings, J. A. (1998). Promoting student interaction in virtual college classrooms. *IHETS.* Retrieved March 6, 2007, from http://www.ihets.org/archive/progserv_arc/education_arc/distance_arc/faculty_papers_arc/1998/indiana2.html

Duffy, T. M., & Cunningham, D. J. (1996). Constructivism: Implications for the design and delivery of instruction. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 170-198). New York: Simon Schuster Macmillan.

Ellul, J. (1964). *The technological society* (J. Wilkinson, Trans.). New York: Vintage Books.

Fuller, T., & Söderlund, S. (2002). Academic practices of virtual learning by interaction. *Futures, 34*, 745-760.

Grooms, L. D. (2000). Interaction in the computer-mediated adult distance learning environment: Leadership development through online education. *Dissertation Abstracts International, 61*(12), 4692A.

Jonassen, D., Davidson, M., Collins, M., Campbell, J., & Haag, B. B. (1995). Constructivism and computer-mediated communication in distance education. *The American Journal of Distance Education, 9*(2), 7-26.

Keirns, J. (1999). *Designs for self-instruction: Principles, processes, and issues in developing self-directed learning.* Needham Heights, MA: Allyn & Bacon.

Kidd, J. R. (1973). *How adults learn.* Chicago: Follett Publishing Company.

Klemm, W. (1998). Eight ways to get students more engaged in online conferences: A blackboard tip sheet. Retrieved March 6, 2007, http://resources.blackboard.com/scholar/general/pages/ictraining/Eight_Ways_Engage_Conferences.pdf

Long, H. B. (1983). *Adult learning: Research and practice.* New York: Cambridge.

McCain, T., & Jukes, I. (2001). *Windows in the future: Education in the age of technology.* Thousands Oaks, CA: Corwin Press.

Moore, M. G. (1989). Editorial: Three types of interaction. *The American Journal of Distance Education, 3*(2), 1-7.

Morrison, D. (2004). *What do instructional designers design?* Retrieved January 27, 2004, from http://www.morrisonco.com

Negroponte, N. (1995). *Being digital.* New York: Alfred A. Knopf.

Palmer, P. J. (1993). *To know as we are known: Education as a spiritual journey.* San Francisco: Harper Collins.

Ponton, M. K., & Carr, P. B. (2000). Understanding and promoting autonomy in self-directed learning. *Current Research in Social Psychology, 5*(19). Retrieved June 14, 2003, from http://www.uiowa.edu/~grpproc/crisp/crisp.5.19.htm

Reid-Martinez, K. (2006). *What's that in your hand?* Presentation at the 2006 General Assembly, International Conference of Educators, Indianapolis, IN.

Roblyer, M. D., & Ekhaml, L. (2000, June 7-9). *How interactive are your distance courses? A rubric for assessing interaction in distance learning.* Paper presented to the Distance Learning Association Proceedings. Retrieved June 8, 2003, from http://www.westga.edu/~distance/roblyer32.html

Rovai, A. P. (2004.). A constructivist approach to online learning. *The Internet and Higher Education, 7*(2), 79-93.

Santoro, G. M. (1996). What is computer-mediated communication? In Z. L. Berge & M. P. Collins (Eds.), *Computer mediated communication and the on-line classroom* (Vol. 1, pp. 11-27). Cresskill, NY: Hampton Press.

Schweizer, H. (1999). *Designing and teaching an on-line course: Spinning your Web classroom.* Boston: Allyn & Bacon.

Smith, G. G., Ferguson, D., & Caris, M. (2002). Teaching over the Web versus in the classroom: Difference in the instructor experience. *International Journal of Instructional Media, 29*(1), 61-67.

Tenenbaum, G., Naidu, S., Jegede, O., & Austin, J. (2001). Constructivist pedagogy in conventional on-campus and distance learning practice: An exploratory investigation. *Learning and Instruction, 11*(2), 87-111.

Waits, T., & Lewis, L. (2003). *Distance education at degree-granting postsecondary institutions: 2000-2001* (NCES No. 2003-017). Washington, DC: U.S. Department of Education, National Center for Educational Statistics.

Wang, A., & Newlin, M. (2002). Predictors of performance in the virtual classroom. *T. H. E. Journal Online, 29*(10), 21-22, 26-28.

Weller, H. G. (1988). Interactivity in microcomputer-based instruction: Its essential components and how it can be enhanced. *Educational Technology, 28*(2), 23-27.

Wojnar, L. (2002). Research summary of a best practice model of online teaching and learning. *English Leadership Quarterly, 25*(1), 2-9.

Young, J. R. (2002). The 24-hour professor: Online teaching redefines faculty members' schedules, duties, and relationships with students. *The Chronicle of Higher Education, 48*(38), A31-A33.

KEY TERMS

Autonomous Learning: The process in which individuals take responsibility for their learning.

Collaborative Learning: The process in which individuals negotiate and generate meaning and solutions to problems through shared understanding.

Computer-Assisted Instruction: The computer serves as the "teacher" by structuring information delivered to the human user.

Computer-Based Conferencing: E-mail, interactive messaging, and group conference support systems.

Constructivism: An approach in which students share responsibility for their learning while negotiating meaning through active participation in the co-creation of shared understanding within the learning context.

Distributed Knowledge: Information dispersed throughout a community of practice and not held by any one individual.

Informatics: Online public access libraries and interactive remote databases.

Interaction: Mutual communicative exchange between individuals.

Category: Human Aspects of Technology

Constructivist Apprenticeship through Antagonistic Programming Activities

Alessio Gaspar
University of South Florida, Lakeland, USA

Sarah Langevin
University of South Florida, Lakeland, USA

Naomi Boyer
University of South Florida, Lakeland, USA

INTRODUCTION

Computer programming involves more than thinking of a design and typing the code to implement it. While coding, professional programmers are actively on the lookout for syntactical glitches, logic flaws, and potential interactions of their code with the rest of the project. Debugging and programming are therefore not to be seen (and taught) as two distinct skills, but rather as two intimately entwined cognitive processes. From this perspective, teaching programming requires instructors to also teach students how to read code rigorously and critically, how to reflect on its correctness appropriately, and how to identify errors and fix them.

Recent studies indicate that those students who have difficulties in programming courses often end up coding without intention (Gaspar & Langevin, 2007). They search for solved exercises whose descriptions are similar to that of the new problem at hand, cut and paste their solutions, and randomly modify the code until it compiles and passes the instructor's test harness. This behavior is further exacerbated by textbooks, which only require students to modify existing code, thus ignoring the creative side of programming. Breaking this cognitive pattern means engaging students in activities that develop their critical thinking along with their understanding of code and its meaning.

This article discusses constructivist programming activities that can be used in undergraduate programming courses at both the introductory and intermediate levels in order to help students acquire the necessary skills to read, write, debug, and evaluate code for correctness. Our constructivist apprenticeship approach builds on earlier field-tested apprenticeship models of programming instruction that successfully address the learning barriers of the new generations of novice programmers. We go one step further by realigning such approaches to the genuine difficulty encountered by students in a given course, while also addressing some pedagogical shortcomings of the traditional apprenticeship instructional practice. This is achieved by introducing a strong pedagogical constructivist component at the instructional level through so called antagonistic programming activities (APA). We conclude with a manifesto for a new multidisciplinary research agenda that merges the perspectives on learning found in both the computing education and evolutionary computation research communities.

BACKGROUND

Novice Programmers and their Learning Barriers

The study of the learning barriers encountered by novice programmers is critical to the computing education research community. Recent studies describing the misconceptions and preconceived notions held by novice programmers (Chen, Lewandowski, McCartney, Sanders, & Simon, 2007; Kolikant, 2005) indicate that these learning barriers evolve with each new generation of students. In this context, a phenomenon known as "programming without intention" has been identified as an attempt by students who encounter difficulties in programming to mechanize the programming thought process. Their heuristic boils down to the following: (a) reading the description of the program to write and look up available documentation (solved exercises, Google, Krugle, etc.) for another similar, already-solved exercise, (b) cutting and pasting the solution to that exercise as a starting point for the current assignment, and (c) compiling and running the program and, since it most likely does not do what is expected, modifying it. Due to the lack of understanding of the solution being reused and the lack of time devoted to understand the programming activity from the ground up (e.g., learn the syntax, learn the role of statements, learn when to use which), these modifications often boil down to a series of almost random changes until the program seems to execute according to the requirements.

This obviously random-based development approach has very little to do with programming and leaves students unable to explain why a particular statement is in their code. In some occurrences, students stated, "I have the code now for this assignment; I need to understand it." This indicates a complete reversion of the programming thought process leading from ideas to implementations. Instead, intentionality is lost, and statements are manipulated in an almost mechanical manner without second thoughts. Essentially, students are utilizing skills at the lower end of the knowledge framework by demonstrating cognitive functions that Bloom (1956) would have termed as knowledge or understanding with no ability to analyze, synthesize, or evaluate the programming process itself.

Criticizing this approach is, however, insufficient. Understanding what reinforces our students' belief that they are problem solving when developing code this way is what can really help us lead them to overcome this particular learning barrier. The nature of the exercises typically found in some introductory programming courses might be partly responsible for this situation. Often, novice programmers are only required to reuse already working programs and modify them slightly (under heavy guidance) to do something new. While analogical thinking is essential to the professional developer when learning new languages, technologies, and paradigms, it is not safe for it to be the only conceptual tool developed by students during their first programming experience. Creative thinking, critical thinking (e.g., debugging), and problem solving are all essential components of the programming thought process, which, if not given proper attention from the beginning, might fuel the misconception that programming is just a matter of pattern matching in a big book of existing solutions.

Leveraging Apprenticeship in Programming Courses

This learning barrier can be addressed by an apprenticeship model of teaching (Kolling & Barnes, 2004), which can take on several distinct forms. The most obvious one is instructor-led live coding: An instructor presents a problem to her or his students, lets them work on it for a definite time, and then introduces the solution. Instead of presenting students with a detailed explanation of the complete solution, the instructor builds the solution from scratch in front of his or her audience. This diverges from the usual instructional pattern, which leads students to build a dictionary of problem-solution pairs that were introduced in class. Such courses encourage students to memorize data in the hopes that they will be able to simply regurgitate it at the next exam. If a question dares differ from a previously solved problem in any significant way, they will then attempt to fit the memorized solution to this new problem by applying a couple of minor adjustments, which could be stumbled upon almost randomly. By developing the solution in front of the students, the instructor's teaching is aligned with the learning outcomes of the course: the programming thought process itself vs. its outcomes. This approach is clearly illustrated in the work of the BlueJ team and their textbook (Kolling & Barnes, 2004).

Other implementations of the apprenticeship model of teaching are closer to problem-based learning approaches; students are taught the programming thought process by applying it frequently to solve new problems from scratch. This learn-by-programming or learn-by-doing approach also leads students to realize the importance of creative and critical thinking in the programming activity while reducing the benefits of memorization-only or analogy-only strategies. In complement, these pedagogical strategies are often coupled with peer learning approaches (McDowell, Hanks, & Werner, 2003; Willis, Finkel, Gennet, & Ward, 1994).

These apprenticeship pedagogical strategies address the above-mentioned learning barriers by aligning the skills being practiced by students during exercise sessions with the authentic learning outcomes expected from an introductory programming course. This in itself complements nicely with constructive alignment theory (Biggs, 2003), which aligns assessment tools with expected learning outcomes.

From Apprenticeship to Constructivist Apprenticeship

Despite these significant pedagogical achievements, the apprenticeship model of instruction can be further improved from the instructional method perspective. Let us take a critical look at the above-mentioned apprenticeship activities: instructors demonstrating the programming thought process while solving a problem live, classmates developing code while other students play the role of a peer programming observer, students coding against each other in a game-based learning environment (e.g., Bierre, Ventura, Phelps, & Egert, 2006).

These activities are essentially instructivist in nature; students are presented with a problem, they work on it, and then the instructor (or their peer) corrects them or even develops a complete solution for them. Even though the thought process is the focus of the demonstration rather than the solution itself, the teaching process is mostly unilateral. The "sage on the stage" (or next seat) strikes again and leads students to adopt a rather passive attitude as they receive their instruction.

Besides the motivational or attention-span issues that such approaches can cause, the work invested by students to develop their own solution is completely ignored in the instructional process (a hallmark of instructivist pedagogies). They are therefore never corrected, improved, or even

leveraged to understand how the student learns. The student is instead required to replace his or her misconceptions with the correct ones. This process is not transformative and often leaves students wondering why their own solutions were incorrect. Down the road, students will commit errors when trying to apply a knowledge that was accepted on top of very blurry foundations.

In such a scenario, the quality of instruction is expected to make up for this lack of connection with the students' cognitive models. Its quality is based on the instructor's past experience with similar student populations or on the instructor's knowledge of studies published on traditional learning barriers encountered by sometimes significantly different student populations. While necessary, this knowledge too often fails to capture the authentic learning barriers encountered by students in the classroom being currently taught. Indeed, such studies do not account for the evolution of the learning barriers, cognitive models, and preconceptions with which each single generation of students arrives. If they did, the computing education research community would have certainly converged by now toward an optimal pedagogical strategy for novice programmers. This misleadingly portrays the pedagogy of programming research as a static problem that might end up being solved once and for all for the ages to come regardless of the evolution of students' learner profiles and the programming technologies and methodologies themselves.

We suggest that the next step in improving the apprenticeship model of learning and teaching resides in developing and replacing inherent instructivist dynamics with constructivist ones. This work led us to define constructivist apprenticeship as a general pedagogy. In the context of computing education research, this pedagogical strategy proved to be particularly suited for introductory and intermediate programming courses in which it could be implemented as antagonistic programming activities (Gaspar & Langevin, 2007).

CONSTRUCTIVIST APPRENTICESHIP AS A PROGRAMMING PEDAGOGY

Constructivist Apprenticeship and Antagonistic Programming Activities

While the constructive alignment theory (Biggs, 2003) aligns assessments with learning outcomes, constructivist apprenticeship completes it at two levels. First, it aligns the learning activities themselves (focus on programming thought process and apprenticeship) with these same learning outcomes. Second, it aligns, through constructivism, the pedagogy of instruction with the authentic learning barriers students encounter (as opposed to the assumed ones). In a programming course, such a pedagogical strategy can be implemented through programming activities that adhere to the following principles. First, the activity must develop programming skills and prevent students from reaching acceptable results through mere cut-and-paste strategies. Second, it must reduce the instructivist dimension of the learning experience by fostering a symmetrical dialog between instructor and student (or among students for peer learning). Last, it must reduce this instructivist dimension by de-emphasizing the importance of showing students the "right way." The last two principles led us to consider the benefits of situations in which students work against one another. This reflection led us to experiment with various classes of antagonistic programming activities.

APAs are programming activities aimed at honing the students' critical thinking and troubleshooting skills in order to develop authentic programming skills at the higher levels of Bloom's (1956) taxonomy. These active learning activities enable groups of students to work together but with antagonistic goals, in face-to-face or distance education settings, under the instructor's supervision or purely as peer learning. These activities also leverage constructivist teaching methods in so far that very little of the solution is communicated to students; instead, their errors are pointed out by the instructor or their peers, but left to be fixed by them. The following sections examine the core idea of each APA variant, provide application examples, and discuss their salient features.

APA 1: Student-led live coding (design-focused variant). Many instructors have already realized the benefits of developing solutions in front of their students (live coding). This apprenticeship teaching (Kolling & Barnes, 2004) successfully refocuses the teaching effort on the programming skills themselves instead of the memorization of finished code solutions. However, students are still passively taught the correct thought processes and accept it as they would have accepted a complete solution, while their own errors are still left out of the learning experience.

In response to the instructor-led live coding's shortcomings, one can let students solve problems in front of their classmates. This student-led live-coding variant requires, for each exercise, the teacher to pick a student who will operate the podium PC with a wireless keyboard (Hailperin, 2006). The student's work is projected, thus enabling the instructor to solicit classmates for corrections, improvements, and discussions about the thought process itself and not the complete finished solution only (unlike most peer learning pedagogies). Most importantly, the students' errors provide the instructor with an authentic understanding of his or her students' problems on which to base the lecturing and class

discussions. This marks an improvement compared to the way instructors usually base their pedagogy on their own past experience or published studies, which, while statistically significant, might relate to a population significantly different from their current students. This is also the trademark of a constructivist approach, which values and integrates the students' errors as part of their learning experience.

APA 2: Student-led live coding (alternative-focused variant). In the alternative-focused variant, one student is still given the wireless keyboard. However, while this student develops his or her own solution, others work independently on their own. This allows for more independence and therefore enables the emergence of radically different solutions while still enabling those having difficulty to peek at the chosen student's early work. During the activity, the instructor walks among students, helping them with their individual solutions, answering questions, and assessing the variety of strategies being implemented. After a set time, students are invited to look back at the chosen student's work and contribute comments, fixes, and suggestions for alternative approaches. The instructor can then lead a hands-on discussion while the code is being modified to evaluate the various strategies (good or bad) that were explored independently by students. The following exercise was given as such as an in-class activity for the fourth week of an intermediate programming course right after a recursion lecture:

> *Implement an iterative and recursive version of a function that will return how many times its (strictly positive integer) argument can be divided by 2 until you get a nonnull remainder. For instance, F(4) will display 2 time(s), F(5) will display 0 time(s), and F(6) will display 1 time(s).*

Students came up with a wide range of solutions including iterative ones (e.g., *while* loop), recursive ones building their results on call returns (expected as the closest of the lecture's examples), and recursive ones building their results during the call (using an extra parameter). This resulted in discussions about tail recursion optimization and detailed explanations (based on program execution stack diagrams) of each solution to ensure they were understood by all. Finally, students also generated unanticipated solutions involving global and static local variables, which motivated a minireview session on these topics. This variant illustrates how a slight protocol change can foster significantly different learning dynamics, which can be further supported by open-ended exercises.

APA 3: Test-driven peer code reviews. So far, APA relied on instructors to serve as a central hub, coordinating the students learning. In larger classes, it becomes difficult for a single instructor to coordinate many students while still allowing all of their voices to be heard. In such settings, peer learning strategies are generally more successful. The very principles of constructivist apprenticeship can also be leveraged in a peer learning context by enabling students to correct each other in a noninstructivist manner. Most peer learning pedagogies rely on collaboration between peers of equivalent levels. In practice, the peers might end up exchanging complete solutions unilaterally or, in a less extreme dysfunctional scenario, exchange corrected code fragments or partial solutions. Even if this teaching and learning dynamic is not unilateral, each peer's contribution too often ignores the others' and boils down to a "your code doesn't work, mine does" reaction. How can we switch this learning dynamic to something more balanced (both students get to develop their solution and improve them) and more constructivist (their individual attempts are matured in a satisfying solution rather than ignored in the acceptance of correct solutions)? The following exercise, adapted from the Javabat applets (Parlante, n.d.), illustrates how this can be done.

The students were paired and given a problem description:

The squirrels in Palo Alto spend most of the day playing. In particular, they play if the temperature is between 60 and 90 (inclusive). Unless it is summer, then the upper limit is 100 instead of 90. Using Raptor, write a flowchart which is going to ask the user to provide a temperature value (between 0 and 130) and a number summer which will be equal to 0 or 1. Depending on the values that were passed to you by the user, you will determine whether the squirrels are playing or not and display on the screen an appropriate message.

In each pair, both students developed their own solutions independently along with a test harness. A test harness, in this case, was simply a written list of tests to be performed on the programs along with their expected and observed outcomes. Once satisfied with their code, students applied their test harnesses to their peer's code. As they did so, they read the code itself, taking note of which tests failed and adding new tests to capture flaws they spotted in their peer's code. Then, they exchanged again their programs to improve them based on the failed tests. This process was reiterated until both programs were successful through each test harness.

The main pedagogical benefit of this APA is its constructivist nature; students instruct each other without exchanging code and without directly fixing errors (a more subtle form of instructivism). Instead, each peer can only adapt their test

harness to reflect the flaws they perceived in both programs. This allows students to develop their critical thinking and debugging skills without being shown how to improve directly. Misconceptions in students' minds, along with their incarnation as bugs in their respective programs, are therefore addressed without relying on an instructivist exchange.

The activity is also antagonistic; students "attack" each other's code through test harnesses. However, unlike most game-based educational activities (Bierre et al., 2006), students are not allowed to confuse "learning by blasting" with "learning to blast. In many competitive learning settings the task at hand (e.g., piloting a tank) ends up overpowering the educational outcome (e.g., learning to program). Code quality, sometimes even correctness, become secondary to performing well in the game. Depending on the game itself, the winning strategy can cost students their learning experience. In our activities, this obstacle is removed by having students compete on code correctness through test harnesses instead of programs' outcomes.

FUTURE TRENDS

Constructivist apprenticeship realigns the pedagogy of contents with learning outcomes and the pedagogy of instruction with the authentic learning barriers experienced by students, thus complementing the achievements of constructive alignment theory. From the applicability perspective, this principle is extremely flexible, and its benefit can reach beyond the programming courses (introductory and intermediate) we have been focusing our discussion on so far. Any course conveying a problem-solving skill to students can benefit from this pedagogical strategy (e.g., accounting, software engineering, algorithms design, etc.). In this expanded context, live coding and code peer review activities can be more broadly perceived as peer-reviewed problem solving. Besides its application to a wider range of courses and disciplines, constructivist apprenticeship has also the potential to serve as ambassador for a new interdisciplinary research agenda bringing together the understanding of learning dynamics from two apparently unrelated fields: evolutionary computation and computing education. The following example illustrates how APA can draw inspiration from and leverage evolutionary computation techniques to inform and improve educational practices.

Using test harnesses in APA allowed us to leverage constructivist peer learning dynamics. Instructors can use similar strategies to teach students by, for instance, crafting an array of values to fail a student's sorting algorithm. Such an approach convinces students of their errors without forcing a solution, which would fail to explain why their codes are not acceptable. Interestingly, similar strategies are used when coevolving artificial neural networks. The way students develop iteratively new test cases meant to fail their peer's code as it is being improved from its confrontation with previous test harnesses can be seen as a coevolutionary dynamic. Test harnesses and programs are defining their respective fitness (i.e., goodness) as a function of how they perform against one another. Similar coevolutionary dynamics can be found in predator-prey models, leading each species to improve under the selective pressure of their counterpart's own evolutionary drive. This dynamic has also been successfully applied to the design of artificial neural networks (Mayer, 1998). A series of neural networks are created randomly, thus forming a population that will be opposed to another population made of training samples, that is, a set of input-output pairs that represent the expected behavior neural networks are expected to learn. The two populations undergo an evolutionary computation transform mimicking the mutation, recombination, and selection scheme found in natural evolutionary systems (Holland, 1992). The quality of a neural network is then quantified in terms of the number of training samples it can successfully solve. Inversely, the quality of a training sample is measured by the number of neural networks it can cause to produce a wrong error. By improving the quality of one of the two populations, we improve the quality of the other, thus leading to a so-called coevolutionary scheme.

CONCLUSION

This article describes constructivist apprenticeship and its applications to programming courses through antagonistic programming activities. We stressed the potential of this approach to help students overcome a learning barrier characterized by a loss of intentionality when designing programs and its applicability to both instructor-focused and peer-learning scenarios. Ramifications and synergies with other educational theories were discussed as well as the potential for constructivist apprenticeship to benefit other computing courses, or even disciplines, that focus on teaching problem-solving skills. Besides the educational framework in which this approach has been developed, we see constructivist apprenticeship as a herald of an interdisciplinary research agenda bringing into the computing education research field an understanding of learning dynamics from the perspective of evolutionary computation researchers. As discussed, results in the latter can mature into and inspire improved pedagogical strategies.

REFERENCES

Bierre, K., Ventura, P., Phelps, A., & Egert, C. (2006). Motivating OOP by blowing things up: An exercise in cooperation and competition in an introductory java programming course. In *Proceedings of the 37th SIGCSE Technical Symposium on*

Computer Science Education (Vol. 38, No. 1, pp. 345-358). New York: ACM Press.

Biggs, J. (2003). *Teaching for quality learning at university.* Buckingham, United Kingdom: Open University Press/McGraw Hill Educational.

Bloom, B. S. (1956). Taxonomy of educational objectives: The classification of educational goals. In *Handbook I: Cognitive domain.* New York: David McKay Co, Inc.

Chen, T.-Y., Lewandowski, G., McCartney, R., Sanders, K., & Simon, B. (2007). Commonsense computing: Using student sorting abilities to improve instruction. In *Proceedings of the 38th SIGCSE Technical Symposium on Computer Science Education* (pp. 276-280). New York: ACM Press.

Gaspar, A., & Langevin, S. (2007). Restoring "coding with intention" in introductory programming courses. In *SIGITE 2007 Proceedings of the International Conference of the ACM Special Interest Group in Information Technology Education.* New York: ACM Press

Hailperin, M. (2006, July 10). Classroom programming. Message posted to the SIGCSE members mailing list, archived at http://listserv.acm.org/archives/sigcse-members.html

Holland, J. H. (1992). *Adaptation in natural and artificial systems.* Boston: MIT Press.

Kolikant, Y. B. D. (2005). Students' alternative standards for correctness. In *Proceedings of the 2005 International Workshop on Computing Education Research* (pp. 37-43). New York: ACM Press

Kolling, M., & Barnes, D. J. (2004). Enhancing apprentice-based learning of Java. In *Proceedings of the 35th SIGCSE Technical Symposium on Computer Science Education* (pp. 286-290). New York: ACM Press.

Langr, J. (2005). *Agile Java: Crafting code with test driven development.* Pearson.

Mayer, H. A. (1998). Symbiotic co evolution of artificial neural networks and training data sets. In *Lecture notes in computer science* (Vol. 1498, pp. 511-520). Springer Verlag.

McDowell, C., Hanks, B., & Werner, L. (2003). Experimenting with pair programming in the classroom. In *ACM SIGCSE Bulletin: Proceedings of the Eighth Annual Conference on Innovation and Technology in Computer Science Education, 35*(3), 60-64.

Pargas, R. P. (2006). Reducing lecture and increasing student activity in large computer science courses. In *Annual Joint Conference Integrating Technology into Computer Science Education: Proceedings of the 11th Annual SIGCSE Conference on Innovation and Technology in Computer Science Education* (pp. 3-7).

Parlante, N. (n.d.). *Online resource, JavaBat applets.* Retrieved October 11, 2007, from http://javabat.com/

Willis, C. E., Finkel, D., Gennet, M. A., & Ward, M. O. (1994). Peer learning in an introductory computer science course. In *Proceedings of the 25th SIGCSE Technical Symposium on Computer Science Education* (pp. 309-313). New York: ACM Press.

KEY TERMS

Antagonistic Programming Activities: These are programming learning activities meant to motivate students by leveraging competitive dynamics focused on scrutinizing, critiquing, improving, and troubleshooting classmates' code. These activities embody the constructivist apprenticeship principles as applied in both instructor-supervised and peer-learning contexts.

Apprenticeship: An apprenticeship is an educational approach historically employed to train crafts practitioners adapted to computing education as a way to teach programming skills through instructor-led demonstrations of from-scratch problem solving (instructor-led live coding).

Constructive Alignment Theory: It is an alignment of learning and teaching activities with the course outcomes and constructivist principles. It was introduced by Professor John Biggs (2003)

Constructivist Apprenticeship: This is a variant of the apprenticeship model of teaching that realigns the teaching practice to incorporate constructivist educational practices. It is applied to programming courses through antagonistic programming activities.

Evolutionary Computation: It is the field of research that deals with the design, and application to engineering problems (e.g., optimization, learning), of bio-inspired algorithms that embody the quintessential characteristics of natural evolutionary systems.

Student-Led Live Coding: It is an antagonistic programming activity in which a student's programming thought process is made visible to all in order to enable an apprenticeship learning based on the authentic learning barriers encountered by students (as opposed to the assumed ones).

Test-Driven Peer Code Review: It is an antagonistic programming activity that pairs students to work on a given exercise, yet allows them to develop their own solutions independently, thus ensuring the symmetrical involvement of

both peers. Solutions are then exchanged and test harnesses developed to fail the peer's code. These test cases are the only instructive information exchanged between peers, thus ensuring a constructivist learning dynamic.

Contactless Payment with RFID and NFC

Marc Pasquet
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Delphine Vacquez
ENSICAEN, France

Joan Reynaud
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Félix Cuozzo
ENSICAEN, France

INTRODUCTION

The radio frequency identification (RFID) reading technology enables the transfer, by radio, of information from electronic circuit to a reader, opened up some interesting possibilities in the area of e-payment (Domdouzis, Kumar, & Anumba, 2007). Today, the near field communication technology (NFC) opens up even more horizons, because it can be used to set up communications between different electronic devices (Eckert, 2005).

Contactless cards, telephones with NFC capacities, RFID tag have been developed in industry and the services (Bendavid, Fosso Wamba, & Lefebvre, 2006). They are similar, but, some major differences explain the specificity of these three applications and the corresponding markets. The label, or marker, is a small size electronic element that transmits, on request, its numerical identification to a reader.

The RFID identification makes it possible to store and recover data at short distance by using these miniature markers or labels (see Figure 1) associated to the articles to identify. The cost of the label is only few centimes. An RFID system is made of labels, readers connected to a fixed network, adapted software (collection of information, integration, confidentiality...), adapted services, and management tools that allow the identification of the products through packing.

Contactless smartcards (see Figure 2) contain a microprocessor that can communicate under a short distance with a reader similar to those of RFID technology (Khu-smith & Mitchell, 2002).

The originality of NFC is the fact that they were conceived for the protected bilateral transmission with other systems. NFC respects the standard[a] ISO-14443 (Bashan, 2003) and thus, can be used as a contactless card. It can be used as a contactless terminal communicating with a contactless card or another NFC phone (ISO-18092). Services available through NFC are very limited today, but many experiments are in progress and electronic ticketing experiences (subways and bus) started in Japan[b].

There are two types of NFC phones:

- The mono chip composed of only one chip for GSM services (called the SIM) and NFC services. In that case, an NFC service is dependent of the phone operator.
- The dual chip shows a clear separation of the two functions within two different chips. That completely

Figure 1. Some examples of RFI label

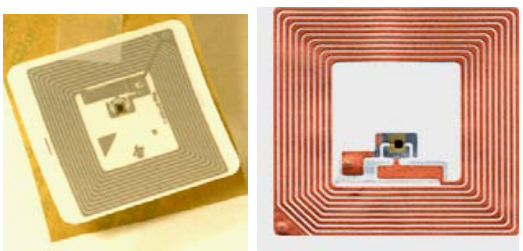


Figure 2. Example of a contactless bank card

isolates the operator and allows independent NFC services…

We define the technology standards, the main platforms and actors in the background section. The main trust develops some contactless payment applications, and analyses the benefits and constraints of the different solutions. The future trends section concerns the research and technology evolution in contactless payment applications.

BACKGROUND

The major interest of contactless cards is to facilitate access control, micropayment… Another interest refers to the usury of card; it is insensible to contact oxidation. We detail briefly the international standards that are involved in RFID and NFC.

Standards

ISO-14443

This standard is the international one for contactless smartcards operating at 13.56 MHz in close proximity of a reader antenna. This ISO norm sets communication standards and transmission protocols between a card and a reader to create interoperability for contactless smartcard products. Two main communication protocols are supported under the ISO-14443 standard: Type A and B. Other protocols were only formalized: Type C (Sony/Japan), Type D (OTI/Israel), Type E (Cubic/USA), Type F (Legic/Switzerland).

This norm is divided in four parts and treats Type A and Type B cards:

- ISO-14443-1 defines the size and physical characteristics of the antenna and the microchip;
- ISO-14443-2 defines the characteristics of the fields to be provided for power and bi-directional communication between coupling devices and cards;
- ISO-14443-3 defines the initialization phase of the communication and anticollision protocols;
- ISO-14443-4 specifies the transmission protocol.

ISO-14443 uses different terms to name its components:

- PCD: proximity coupling device (or reader);
- PICC: proximity integrated circuit card (or contactless card).

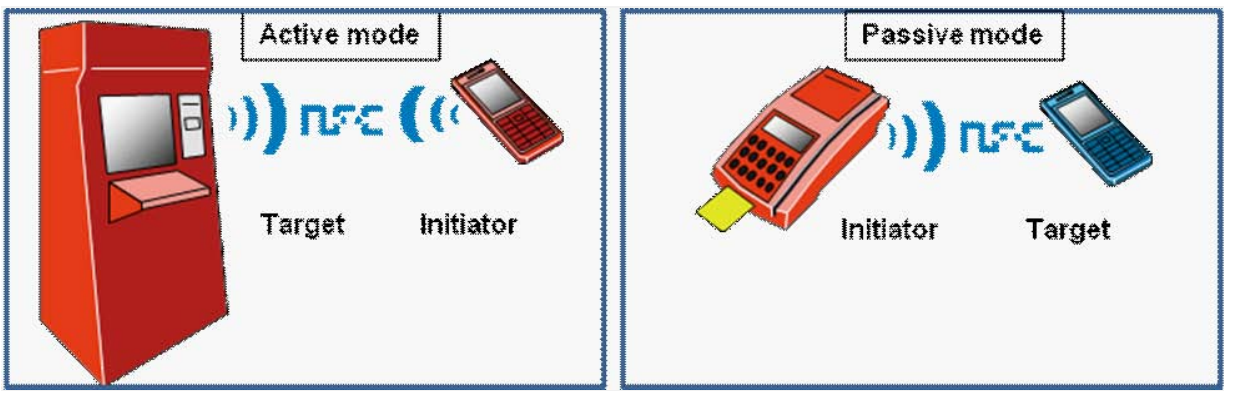
ISO-18092

NFC is a short-range (10 to 20 centimeters) wireless communication technology that enables the exchange of data between devices over a short distance. Its primal goal is the mobile phones usage. This open platform technology is standardized in ISO-18092 norm NFC Interface protocol-1[c]. In NFC technology, two communication modes exist: passive and active communication modes of NFC interface protocol to realize a communication network using NFC devices for networked products and also for consumer equipments (see Figure 3).

ISO-21481

The ISO-21481 standard (NFC interface protocol-2[d]) is derived from Ecma-356 (interconnection) standard. It specifies the selection mechanism of communication mode in order to not disturb communication between devices using ISO-

Figure 3. The two NFC communication modes



18092, ISO-14443 (contacless interface - proximity), and ISO-15693 (contacless interface - vicinity).

Application Platforms and Major Actors

There are major actors in the field of contactless applications; we distinguish two important platforms using the contactless technology: Mifare and FeliCa. This chapter does not focus on more details about these platforms technology, but is more about their applications.

Current actors in payment applications, namely MasterCard and Visa, stay alert, and intend to play a major role in future payment applications. They have already joined the movement and launch many developments over contactless payments. They begin to agree to a common communications protocol for contactless payment devices. This is based on the MasterCard PayPass™ protocol. MasterCard made the first step with a contactless credit card (see Figure 4) (Olsen, 2007).

The Visa PayWave technology is rather largely deployed within many European countries. They both intended the American market to future deployments (Turner, 2006).

Visa and MasterCard technologies comply with the EMV (Europay Mastercard Visa) standard. This standard defines the interoperation between smartcards and terminals for authenticating credit and debit cards. It defines strong security measures and provides a strong authentication along the process.

Mobile specifications are still in an early stage of development. Those who want to follow the development can do it at the EMVCo Web site[e]. Contactless cards that define the EMV standard over contactless communication does not differ so much with contact cards. The differences will be in the usages and applications.

Figure 4. Payment with a contactless card

MAIN FOCUS OF THE CHAPTER

The main focus of the chapter is an analysis of the benefits and limitations of RFID authentication for electronic payment (Tajima, 2007). This part deals with the particular constraints of banking (computation time, security…) for this kind of authentication process (Chen & Adams, 2004). The use of radio frequency and the small distance allows some security weakness that leads to security reinforcements.

Contactless Cards in Banking Applications

We have seen that MasterCard and Visa have an agreement to share a common transmission protocol and experimentation for the contactless payments by radio frequency in the points of sale.

Contactless payments, as conceived in the programs MasterCard PayPass and Visa PayWave[f], make it possible for the cardholders to carry out fast payments by a simple passage of their card in front of a terminal, thus, avoiding them giving their payment card to a merchant or handling cash. Contactless payments are much more practical for the consumers and are particularly adapted in environments of purchase where the speed is essential, like fast food, the gas station, but also theaters. They also offer new appropriate payments by using a card in unusual environments of purchase, like slot-machines or tolls. To make a payment, a user presents his/her card near the front of a terminal (a beep is emitted by the terminal). A request for an online authorization is sent. The payment is carried out.

There exist two types of PayPass cards:

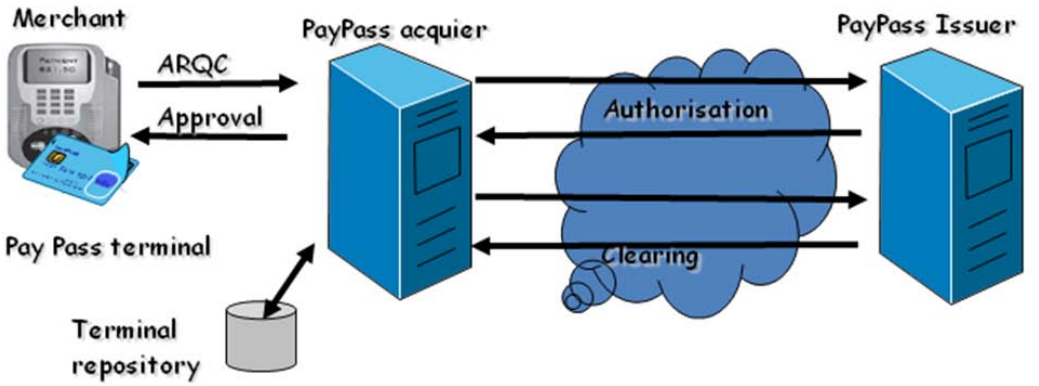
- Contactless with a magnetic stripe ;
- Contactless with a chip that is EMV compliant (dual-use card).

For the European market, Visa is planning on using RFID-enabled dual-use debit cards, based on its own Visa Contactless payment technology. It aims, in particular, at European countries already using EMV compliant cards. But, Visa is also understood to be in talks with mobile manufacturers to use NFC technology that will enable a phone to be used instead of a card.

For the US market, the contactless PayPass is not EMV compliant, so, the target is to limit the authorization requests (see Figure 5). How does it work?

- For a small amount (as for illustration <$30), an authorization of $30 is requested, then debited of the small amounts carried out;
- The payments lower than the authorization threshold, decrement of the preceding payments, are not online transaction object;

Figure 5. Delayed clearing and settlement



- As soon as the cumulated ceiling of preauthorization is reached, a new preauthorization is required.

When the amount exceeds $30, for example, an authorization request is always sent with the following exchange. The security of the transaction is guaranteed, first, by the use of some information stored in the card such as the card number, and second, through the secure transfer with the terminal by using the RSA algorithm: For the EMV countries, cards have contactless capability and EMV compliance. The transaction has another scheme.

NFC

NFC can be used as a terminal or a simple contactless card (Remédios, Sousa, Barata, & Osório, 2006). Before NFC, contactless applications, like payment or control access, were only implemented on cards. These developments were limited because the card has no battery power supply. Mobile phone is a possible solution in face of this problem because it is auto-provided with energy. Lots of new applications can be charged on an NFC phone, but not on a contactless card.

In opposition to credit card, mobile phone memory is not safe enough for storing secret data and critical applications. That is why these applications can be embedded on a separate chip. The security and confidentiality of data are ensured by encryption, which is handled by the chip itself.

The chip can be located in several places (Mallett, Millar, & Beane, 2006):

- Into the phone: The drawback is that the chip is not transferable; it is integrated in the phone. This system is named dual chip and has many advantages: telecom operators independence, security… This solution has been chosen by the ITEA research project called SmartTouch (SmartTouch, 2005) whose objective is to study and promote the use of NFC mobile phones for different applications including payment ;
- In the SIM card: The problem of this solution is that there is no standardization yet, and the telecom operator is responsible for all applications on its SIM card, thus responsible for the NFC part.

NFC mobile can be used as a terminal with a contactless card or another phone (ISO-18092), or it can be used as a contactless card (ISO-14443). Mobile screen, key pad, and connectivity features can be used to create more and more user friendly applications.

NFC Phone as a Contactless Credit Card

The NFC mobile phone can be used as a contactless credit card. In this case, NFC technology does not bring anything more than a contactless card (except for dematerialization of card, which is more useful).

An example of these applications is the proximity payment. The merchant, who has an NFC terminal, enters the amount of the transaction. The mobile owner then puts the phone near the terminal and information about his/her transaction are displayed on his/her mobile screen. If the owner agrees the transaction, he/she validates and enters his/her PIN code. He/she puts the phone on the terminal to send all information (number bank account…). He/she finally can take his/her ticket, since the transaction is closed.

In Japan, this payment solution is actually used. Lots of taxis adopted NFC payment system in their vehicle, which secures transactions for taxi and users do not need any cash money[g]. Some gas stations are already equipped for the NFC payments with mobile phone. Lots of research programs are based on NFC technology (Jaring, Törmänen, Siira, & Matinmikko, 2007).

FUTURE TRENDS

Future trends stress the different research topics that should participate to solve some still existing problems in contactless payment (Chen & Adams, 2004). We can consider that there are two main actors involved in the NFC payment with different objectives and limits.

The Mobile Phone Operators

An entity, either the mobile phone operator or a third-party vendor, sells the mobile phone and produces the correct information required to personalize the SIM. The responsibility of the personalization of the SIM is given to the mobile phone operator. The mobile phone operator prefers using a second chip than a single one because there is no exchange with the environment of the phone without it permits it (and more, it bills it). In that case, if a banking actor tries to modify the parameter of the SIM to implement the secrecy of the bank to allow the payment, it can do it only with the agreement of the mobile phone operator, who is the owner of the secrecy of the SIM.

That relation between the mobile phone operator and the bank can only be nowadays, with a one-to-one agreement that is incompatible with a generalization of that solution. Today, to start experimentations, the solution is to externalize the personalization to a specialized third partner well known by the mobile phone operators and by the banks. For each cardholder, the third party will receive the secret keys and software from the cardholder's mobile phone operator and the cardholder's bank, and will personalize the SIM with that information (Pasquet, Reynaud, & Rosenberger, 2008). That solution is tested by the three mobile phone operators and by five major banks in France in the Pegasus project on two French cities (Strasbourg and Caen).

In the future, it is necessary to modify the SIM architecture and to create some virtual shelters inside the SIM, protected by keys communicated by the SIM operator to the bank to allow the bank a remote personalization. The global platform specifications (new standard for smartcards infrastructure) are on the way to allow that secure remote personalization, but it will take more or less 2 or 3 years[h].

A second solution is possible with an NFC mobile phone; it is to not deal with a bank, and to pay a service by the costumer's mobile phone bill. These payment types, already used to buy bell rings for the customer phone, can also be used for NFC payments.

The Banks

Ten years ago, some banks tried to develop, with mobile phone manufacturers, some mobile phone with a special slot to insert banking cards[i]. They have given up that solution; very secured but expensive.

Another possibility is the use of NFC dual chip mobile phone where the SIM is completely separated from the NFC chip. The SIM is bought by the customer and installed in the mobile phone (Remédios et al., 2006).

The phone manufacturers are interested in that solution but the question is: which actors will commercialize the mobile phone if the operators disagree with the dual chip solution? Experimentations are in progress in few countries in the world (Finland, France…), but it will be difficult to convince the whole mobile phone operators to share the income of such solution.

Except that problem, the banks have just to personalize the NFC mobile phone. But, the phone must be in front of the antenna of the personalization equipment. This leads to two solutions: the banks can give the NFC mobile phone (within the partnership with a mobile phone operator and a manufacturer) to their clients (which model, which color…?), or the banks develop some personalization equipments and install them to personalize the NFC mobile phone of their clients. The two solutions impose some high investments for the banks.

CONCLUSION

Which technology is the best regarding traceability and security (identification and authentication)? As regard to the identification, the RFID, the smartcard, and the NFC, after the barcode, are today in competition. All actors develop their technology, until the moment when the aspect of universality or cost price is called into question by new considerations.

It appears that each technology is worth only according to the markets that are open for him/her, and for which industrial series allow a good profitability of use of the associated products. The convergence of the new networks will bring, by association with other technologies (cryptology, reduction in price and volume of the memories, particular modulations...), elements suitable to stimulate this dynamics.

REFERENCES

Bashan, O. (2003). *An introduction to the contactless standard for smartcards and its relevance to customers*. Retrieved from http://www.otiglobal.com/objects/ISO%2014443%20WP%204.11.pdf

Bendavid, Y., Fosso Wamba, S., & Lefebvre, L.A. (2006). Proof of concept of an RFID-enabled supply chain in a B2B e-commerce environment. *ACM International Conference on Electronic commerce, 156*, 564-568.

Chau, P. Y. K., & Poon, S. (2003). Octopus: An e-cash payment system success story. *Communications of the ACM archive, 46*, 129-133.

Chen, J. J., & Adams, C. (2004). Short-range wireless technologies with mobile payments systems. *ACM Proceedings of the 6th international conference on Electronic commerce, 60*, 649-656.

Domdouzis, K., Kumar, B., & Anumba, C. (2007). Radio-frequency identification (RFID) applications: A brief introduction. *Advanced Engineering Informatics, 21*, 350-355.

Eckert, C. (2005). Security issues of mobile devices. *Lecture Notes in Computer Science Security in Pervasive Computing, 3450*.

Fosso Wamba, S., Lefebvre, L. A., Bendavid, Y., & Lefebvre, E. (2007). Exploring the impact of RFID technology and the EPC network on mobile B2B eCommerce: A case study in the retail industry. *International Journal of Production Economics.* In Press.

Jaring, P., Törmänen, V., Siira, E., & Matinmikko, T. (2007). Improving mobile solution workflows and usability using near field communication technology. *Lecture Notes in Computer Science Ambient Intelligence,. 4794*.

Khu-smith, V., & Mitchell, C. J. (2002). Using GSM to enhance e-commerce security. In *Proceedings of the 2nd International Workshop on Mobile Commerce*, pp. 75-81

Mallett, C. T., Millar, W., & Beane, H. (2006). Perspectives on next generation mobile. *BT Technology Journal, 24*, 151-160.

Olsen, C. (2007). Getting the most out of EMV with contactless cards. *Card Technology Today, 19*(4), 10-11.

Pasquet, M., Reynaud, J., Rosenberger, C. (2008). Secure payment with NFC mobile phones in the SmartTouch Project. *The 2008 International Symposium on Collaborative Technologies and Systems*. In Press.

Remédios, D., Sousa, L., Barata, M., & Osório, L. (2006). NFC technologies in mobile phones and emerging applications. In *Information Technology For Balanced Manufacturing Systems, IFIP International Federation for Information Processing*, vol. 220.

Roberts, C. M. (2006). Radio frequency identification (RFID). *Computers & Security, 25*, 18-26.

Sauniere-Miallet, G. (2004). School opts for contactless payment. *Card Technology Today*.

SmartTouch. (2006). *ITEA Project*. Retrieved from http://www.smarttouch.org/

Tajima, M. (2007). Strategic value of RFID in supply chain management. *Journal of Purchasing and Supply Management.* In Press.

Turner, S. (2006). A world beyond contactless cards? *Card Technology Today, 18*(10), 10-11.

KEY TERMS

Contactless Cards: The contactless smartcards contain a microprocessor that can communicate under a short distance with a reader similar to those of RFID technology

EMV: Europay, MasterCard and Visa specifications. This is a standard for interoperation between smartcards and point of sale terminals and also automated teller machine.

Felica: That platform owned by Sony Corporation, originally proposed as ISO-14443 type C but refused, is now compliant with the ISO-18092.

IC: Integrated circuit Miniaturized electronic circuit also known as microcircuit, chip, or microchip.

Myfare: That platform owned by NXP semiconductors, is compliant with the ISO-14443 type A standard.

NFC: Near field communication A short-range high frequency wireless communication technology, an extension of the ISO-14443 proximity-card standard for mobile phones.

RFID: Radio frequency identification Automatic identification method relying on storing and retrieving data using devices called RFID tags.

Smartcards: Card equipped with a chip or integrated circuit card (ICC). It defines any pocket-sized card with embedded integrated circuits which can process information.

Tags: Miniature markers or labels emitting a unique number or other information.

ENDNOTES

[a] ISO-14443 http://www.otiglobal.com/objects/ISO%2014443%20WP%204.11.pdf

[b] http://www.engadget.com/2004/08/09/japan-airlines-and-docomo-plot-to-abolish-the-plane-ticket/

[c] NFC Interface Protocol-1 http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-340.pdf

[d] NFC Interface Protocole-2 http://www.ecma-international.org/publications/files/ECMA-ST/Ecma-352.pdf

[e] EMVCo: www.emvco.com

[f] PayWave in Switzerland

[g] NFC development in Japan http://www.slashphone.com/70/6644.html

[h] See: www.globalplatform.org

[i] In Finland, in Mars 2002, Sonera started a demonstrator of a «Sonera Shopper» service, which had worked until august 2002, in accord with Luottokunta, the card system operator in Finland. It was an instant payment with a Visa or MasterCard card.

Category: Global Information Technology

Contemporary Concerns of Digital Divide in an Information Society

Yasmin Ibrahim
University of Brighton, UK

INTRODUCTION

The social issue of the "digital divide" has courted much political and scholarly attention in the last decade. There is, however, less consensus over the origin of the term, even though it is generally associated with the advancement and diffusion of information technology. According to Jan Steyaert and Nick Gould (2004), the concept of the digital divide is believed to have gained media and academic currency in the mid-1990s. In 1998, the United Nations labelled the digital divide as a new type of poverty that was dividing the world (cf. Hubregtse, 2005). A UNDP (United Nations Development Programme) report in 1999 (cf. Norris, 2000) stated that "the network society is creating parallel communications systems" that increase the divisions between rich and poor nations (p.3). The term, in effect, captures the social inequality of access to technology, particularly the Internet, as well as the long-term consequences of this inequality for nations and societies.

The significance of the term is embedded within the notion of an information society, where information is an important component of the global economy in terms of production, development, and social enrichment of societies and nations. The diffusion of technologies, such as the Internet, has meant the surfacing of various social issues including technology's impact on society, its relationship with older media forms, and its immediate impact on people's social and political lives (Robinson, 2003, p. i). New technologies, such as the Internet, are seen as transforming the globe into an information society with the ability to promote new forms of social identity and social networks while decentralizing power (Castells, 1996, p. 2001). Robin and Webster (1999, p. 91), nevertheless, are of the view that the contextualization of the digital divide debates within the issue of information revolution is misleading, for it "politicises the process of technological development by framing it as a matter of shift in the availability of and access of information."

The term digital divide conveys the broader context of international social and economic relations and in particular, the centre-periphery power configuration marked by American dominance over the rest of the world (Chen & Wellman, 2004, p. 41). In fact, rhetoric and literature on technology and information have always emphasized this divide (see Galtung & Ruge, 1965), not to mention the debates that were sparked in the 1980s by UNESCO's proclamation of the New World Information Order (cf. Norris, 2000). The term has been analysed both at global and regional levels, and has involved the investigation of socioeconomic contexts, global governance, policy issues, as well as cultural elements. The analysis of the digital divide on a global level may entail comparisons of large regions, between developed and developing countries, and between rural and urban areas. In modern consciousness, the phrase captures the disadvantages and inequalities of those who lack access or refrain from using ICTs in their everyday lives (Cullen, 2003).

BACKGROUND

The imbalances between North and South in the field of communications and information were published in the Macbride Report in 1980, under the auspices of the United Nations Educational, Scientific, and Cultural Organization (UNESCO). The report concluded that there were stark discrepancies between industrialised and developing countries with regard to information flows and capacticites for active participation in the communication process (Modoux, 2002, p.2). The report was instrumental in the formation of a New World Information and Communication Order (NWICO), led by the United Nations and UNESCO to address the imbalances. With the appropriation of the NWICO as a "Cold War" agenda and the illumination of information and communication as a key tool of control and propaganda, the debates about the information and communication imbalances became subsumed under this climate of political hostility. In the 1970s and 1980s, the fear that development in the communication field might predominantly benefit the authoritarian regimes in the South (Modoux, 2002, p. 7) mediated much of the rhetoric and, as such, the global political context was important in situating debates on communication and information disparities.

The digital divide, as an issue, dominated the G8 summit in Okinawa in 2000, and has also dominated similar discussions at the first World Social Forum in Porto Alegre in Brazil and the Davos World Economic Forum (Menou, 2001, p. 112). In the same vein, the "World Bank has, from the early 1990s, published a number of reports on information technology and the Internet, stressing it as a major area

of concern for the world. Other global initiatives naturally include the World Summit on Information Society (WSIS) meetings in Geneva in 2003 and Tunis in 2005" (Luyt, 2006, p. 276). In July 2001 at the Genoa Summit, the G8, comprising the most highly industrialized countries, adopted a plan to clarify the role of information and development strategies and their contribution to the fight against poverty. In its agenda, the "Genoa Plan of Action" embraced intiatives aimed at "creating conditions such that everyone, in the years ahead, should be able to participate in the 'information society' and share its benefits." The agenda, as Luyt notes (2004, 2006), to position the digital divide as a global issue, has been shaped by powerful corporations, governments, and civil society organisations. International agencies such as the World Bank, UNDP, and ITU (International Telecommunications Union) have reiterated the need for central government, local government, nonprofit organizations, and the private sector to bridge the global divide.

According to the ITU report (2005), the digital divide, in the last 10 years, has been shrinking in terms of the number of fixed phone lines, mobile subscribers, and Internet users throughout the world. However, there remain significant disparities from nation to nation in terms of access to such technology. According to ITU estimates, some 8,000,000 villages-representing one billion people worldwide-presently lack connection to any kind of ICTs. Statistics also revealed that in 2004 fewer than 3 out of every 100 Africans used the Internet, compared with an average of 1 out of every 2 inhabitants of the G8 countries (Canada, France, Germany, Italy, Japan, Russia, the UK, and the US). In addition, in 2004 there were approximately the same total number of Internet users in the G8 countries as in the rest of the world combined. This translates into 429m Internet users in G8 countries and 444m users in non-G8 countries.

The digital divide is often measured by the degree of access to ICTs and the Internet. With the rapid proliferation of information and communication technologies, there is growing concern over the disproportionate number of users concentrated in developed countries. In 2001 for example, 169m Americans were online, accounting for 60% of the US population and 29% of the world's Internet population (Chen & Wellman, 2004, p. 40). According to a 2005 ITU report, the present digital divide not only refers to inequalities of access to telephones and the Internet, but also to mobile phones, RFID (radio-frequency identification), and sensors. The report stressed that far from there being a single digital divide, there is instead a terrain of varying levels of access to ICTs that may widen the gulf between developed and developing countries if the latter do not actively invest in these fields. Martin and Robinson (2004, p. 2) point out that researchers and policy makers agree that there are presently profound differences in Internet use across incomes, educational levels, races, and ages both in the US and other nations, and often the disagreement is over how long these differences will persist or what these trends will be.

THE MAIN ISSUES

Beyond the contemporary currency of digital divide, the unequal development between rich and poor nations in technology and science had been termed by Hans Singer as "international technological dualism" more than three decades ago (cf. Gudmunsdottir, 2005). The digital divide captures the relationship between the Internet and social inequality, and as the Internet becomes more important in society, those who remain off-line (Martin & Robinson 2007, p. 1). The term situates two meta-issues: on the one hand it focuses on the issue of access and connectivity, and on the other it ventures beyond access issues into media literacy and associated skills and on to issues of social cohesion, civic engagement, and participation (Sciadas, 2002, p. 4). The problem of global information imbalance is often seen beyond the technology paradigm, and is often equated with cultural hegemony (Kema, 2005).

The digital divide refers mainly to the division between the information rich and the information poor, whether they be individuals or societies. It is also common to deploy the term to divide the globe geographically, as in the "North-South" dichotomy or the "West and the rest" (Gudmunsdottir, 2005, p. 3). At a global level, the digital divide results from the fact there is a huge and growing gap between the more advanced countries and the rest regarding the size and intensity of their ICT applications (Menou, 2001, p. 112). According to Rowena Cullen (2001, p, 311), "the digital divide has been applied to the gap that exists in most countries between those with ready access to the tools of ICTs, and the knowledge that they provide access to, and those without such access or skills. This may then be attributed to socioeconomic factors, geographical factors, educational, attitudinal and generational factors." Van Dijk (1999) lists four barriers of access that can impact on digitial divide, and this can include mental access (i.e., the lack of interest), material access (i.e., the lack of infrastructure), skills access (i.e., lack of literacy), and usage acess, which refers to the ability to embrace opportunities to access technology. Others, such as Warschauer (2004), have categorised these impediments as human resources, social resources, digital resources, and physical resources. Similar to Van Dijk's categories, these refer to the lack of infrastructure, language barriers, media literacies and skills, and additionally they focus on the social resources such as the agencies offered through the context of the community, as well as institutions that can mediate policy and deployment of technology.

From a social constructionist perspective, Luyt argues (2006, p. 279; Sciadas, 2002) that the global digital divide is not a social or policy problem but one technological condition among many in a world with divisions of many kinds. According to Luyt, what makes the lack of access to ICTs a policy problem is the work of claim makers who have generated much publicity about the condition and the

negative consequences for those experiencing it. Carsten Fink and Charles Kenny (2003, pp. 16-17) argue that a widening absolute gap in per-capita ICT access does not necessaily imply that poor countries are falling behind. They call for researchers to look at relative rates of growth due to the fact that "if poor countries experience faster rates of growth in ICT usage and access levels, it is mathematically inevitable that in the short term the absolute gap may continue to widen even though they may surpass the rich world at some point." Comparing growth rates for users in rich and poor countries since the early 1990s till 2000, Fink and Kenny (2003, p. 17) conclude that the gap between the countries in terms of usage has been shrinking in relative terms and hence the most striking feature of the digital divide is not how large it is but how rapidly it is closing. For example, in the mid 1990s over 90% of Internet hosts were found in North America and Western Europe while Asia only had a share of 3% of the global Internet hosts. However, by 2001 Asia had 144 million users compared to 180 million and 155 million users in North American and Western Europe, respectively (cf. Hao & Chao, 2004)

As mentioned, the digital divide can equally refer to both the lack of resources and infrastructure as well as the lack of media proficiency (i.e., literacy) and support infrastructure (i.e., education, policies and communal networks), which can create disparities between nations in the information age. The disparity between the digitally empowered and disempowered is expected to intensify without some degree of intervention at the global level. Language issues can also present a barrier to democratizing the digital revolution. This is mainly due to the fact that only about half of the world's Internet users are native English speakers and about three-quarters of all Web sites are in English (cf. Chen & Wellman 2004, p. 42). Besides linguistic barriers, various other factors including socioeconomic status, gender, life stage, and geographical locations can affect people's use and appropriation of new media technologies (Chen & Wellman, 2004, p. 42, Stayert & Gould, 2004). Bearing these five factors in mind, De Han and Rijken (2002) emphasise that the digital divide is another aspect of social exclusion in societies where barriers to accessing new media are no different from barriers to accessing other needs such as health, education, and employment.

Various factors can influence the diffusion of technology in developing countries and these can include infrastructure, government policies, regulations, economic development, culture, and language (cf. Bazar & Boalch, 1997; Cullen, 2003). The digital divide in countries may then be influenced by and reflect social demographics as well as public policies and social issues. As such, the marginalised or those with less social mobility may require public policy interventions to close the gap. In the US, for example, Afro-Americans, Latinos, and North American Indians have been identified as needing public programmes to widen their inclusion (Cullen, 2001, p. 312; Norris, 2000). Martin and Robinson (2007, p.1-2) point out that income inequality is a distinctive economic barrier to Internet use. According to Paul Dimaggio et al. (cf. Martin & Robinson, 2007, p. 2), persons of higher socioeconomic status employ the Internet more productively and to greater economic gain than their less privileged but still connected peers. Analysts (cf. Chowdhury, 2004; Cullen, 2001) point out that while improved access to ICTs and the Internet are important, one should not assume that they alone can solve all the problems that accompany the digital divide within and between countries. For example, new figures from the European Union's Official Statistics body reveals that in 2004, on average, 85% of European students used the Internet compared with only one in eight retired people (eGov Monitor). Secondly, as Cullen (2001, p. 312, 2003) reiterates, new technology does not always replace the old, and in the process of coexisting, they may enhance the range of human experience even if users opt not to favour newer technologies over older ones.

Jos De Haan (2004, pp. 67-68) postulates that social inequality lies at the heart of the debate on the rise of the information society, but too often these discussions are limited to, and centre on, access without fully analysing the consequences of differences in IT access. He argues that Internet access is seen as binary (i.e., whether one is a user of the Internet or not), and this hinders earlier theoretical progress in understanding the influence of communication processes of social change. Instead, he suggests a multidimensional model that includes motivation, possession, and digital skills. Motivation looks at attitudes, interests, the will to use technology, and the fears surrounding it. Possession accounts for access not just in the home, but in other places a user may find himself or herself . The third component looks at the extent to which potential users are able to handle IT. As such, De Hann stresses the need to analyse cultural and attitudinal factors in shaping consumption and behaviour towards technology.

Kubicek (2004) points out that within the developing world the digital divide is growing, as there are discrepancies between urban and rural populations. In addition to geographical remoteness, lack of relevant content and lack of technological support can constitute barriers for the use of the Internet in disadvantaged communities (Chen & Wellman, 2004). The issue of the digital divide is also bound up with the availability of telecommunications infrastructure along with technical capacities (i.e., telephone lines, broadband, satellite services, etc.) Wiring up rural areas entails major investments of capital, and often governments may opt to liberalise the telecommunications industry to allow private investments in these sectors. Where telecommunications industries are privately owned there may be reluctance on the part of investors to inject capital into areas that do not maximize their revenue stream, and this may include rural areas.

In most Western countries, governments actively engage and devise policies to ensure citizens' access to and consumption of ICTs to enable participation in social, educational, and economic activities (Cullen, 2001, p. 113, 2003). The Clinton administration, in 2000, proposed a new plan in which tax breaks were offered to private companies to help close the digital divide between the haves and the have-nots. Clinton's programme also included other support infrastructure such as teacher training programs and Community Technology Centres. As such, public programs are deemed just as important as investing in technical infrastructure to ensure access to the Internet. Other Western European countries have also harnessed public and private resources to achieve increased Internet literacy and access. Rao et al. (cf. Norris, 2000) also point to private and public initiatives to include rural areas in South Asia through remote access, community centres, and Internet kiosks.

The rhetoric of the digital divide distracts the world community from concentrating on society's more urgent needs such as health care and education (Cullen, 2001, 2003; Menou, 2001). Cullen (2001, p. 312) stresses that bringing the Internet to Africa and associating it with better education and increased awareness "concentrates on the wrong end of the technology spectrum as the Internet in itself is not education, does not develop literacy or skills to access and interpret information found." A better alternative, Cullen (2001) suggests, would be to "use basic technologies to promote traditional forms of education, enhance the delivery of health care and improve animal husbandry and crop management. Conversely, Internet technologies can provide different forms of benefits as in the sharing of expertise in health and farming initiatives as well as in tourism, trade and e-commerce"(p. 114).

On the other hand, information technology is deemed crucial in enabling less developed nations to participate in the global market place, to create networks and connections with the global communities. As Norris (2000) points out, "the effect of the Internet in broadening and enhancing access to information and communication may be greatest in poorer nations as it can allow the developing world to enter global trading markets directly irrespective of the traditional barriers of distance, costs of advertising and the intermediate distribution chains" (p. 2). In this sense, the Internet can highlight the concerns of developing societies in the international arena by connecting disparate social movements with global civic society.

Measures to restrict access to technologies such as the Internet can take a number of forms, and these can include financial, technical, administrative, or legislative issues in different countries (Modoux, 2002, p. 5). China, for example, imposes strict censorship laws on the Internet, and users can face severe sanctions if they transgress these regulations. There were an estimated 137 million Internet users in China in 2007, which is second in number to the United States (cf. Fallows, 2007). China, like Singapore, is a paradox due to the fact that despite increasing Internet access it has harsh censorship laws. As such, increasing Internet penetration may not equate to equal access to information when compared to other states with more liberal policies. Digital divide can then manifest in a myriad of forms, and can be politicised differentially in the national and global context. Additionally, China, like other countries, has an urban-rural divide, with only 17% of the total Internet population accruing from rural areas in 2007 (Fallows, 2007).

The lack of infrastructure is a major obstacle to the diffusion of technology. For example, the lack of telephone lines and electricity in many villages in India and the shortage of telephone lines in main cities of Kazakhstan, Uzbekistan, Tajikistan, and Turkemenistan (cf. Hao & Chan, 2004) exercabate the discrepancies between populations in rural and urban areas.

FUTURE TRENDS

The digital divide underpins a degree of vulnerability for those who do not have access to technology, however, the debates become confounded when users, particularly those in the developed world, refuse to engage with technology even when access is not an issue. According to research firm Point Topic, 44% or 11.2 million households in the UK refuse Internet connections and out of these, over 40% have little or no intention of getting connected to the Internet, and as the number of Internet households increase, those that are left are increasingly resistant to its appeal (BBC). As De Haan (2004, p. 84) observes, at the heart of inequality in the information society lies the question of the differential behavioural consequences of inequalities in IT access.

Beyond the attitudinal issues, it will remain a challenge for the global community and less developed countries to keep the digital divide on the global agenda in the coming years. While there is consensus, the narrowing of the digital divide should be a global issue rather than one left to indigenous governments and to the political interests of powerful stakeholders who will continue to mediate the issues and concerns raised at the global level. With much concern over the dominance of the US in the governance of the Internet, developing countries have voiced concerns over the political economy of the global resource that is the Internet.

Beyond being a cliche of the differentials and inequalities in society, the term can be positively enforced in both policy making and global consciousness to enforce equity and deter exploitation. As observed by Fink and Kenny (2003, p. 20), the digital divide paradigm can be used to "promote good policy as in the regulation of private comptetition in information infrastructure provision to ensure improved ef-

ficiency, lower costs and increased access." Inversely, they warn that it can lead to countries having an ill-conceived or over-ambitious infrastructure target leading to excessive competition, and which thwart the objectives of national policies and imperatives.

The endeavour to close the gap between the digitally empowered and the disempowered will inevitably entail high costs, and often it has to be weighed against other serious and pressing issues such as poverty and disease; it is difficult to argue that the digital divide is a bigger concern or threat to the people in developing worlds.

CONCLUSION

The issue of the digital divide has been raised as a global concern in the last decade to highlight yet another type of social inequality that could have future consequences for the world. The concept is tightly aligned with the emergence of a postindustrial or information society in which information is seen as a key resource that enables new modes of production, commerce, and civic engagements. The assumption that underpins the digital divide is that societies would be disadvantaged socially, politically, and culturally if they did not respond by ensuring that their citizens have access to technology and to information. Interventions to remedy the problems of the digital divide would include investments in infrastructure as well as public programmes to promote media literacy and digital skills. In both developed and developing countries, these have involved private and public partnerships, as well as the liberalization of the telecommunications market to encourage private investments in infrastructure. The reliance on private-sector investments to build infrastructure has, in many instances, sustained inequalities between urban and rural areas. Additionally, in many countries, digital inequality has mirrored the patterns of socioeconomic inequality. The digital divide as a term, then, addresses both the lack of access as well as the long-term social consequences of not utilizing new technologies that create the potential for global connectivity, as well as new forms of social and political engagements. For developing countries, the term has often led to a false belief that it will remedy the existing problems of poverty or disease, mirroring old debates where technology is perceived to be a panacea for complex social problems. However, as more governments reconfigure politics and civic engagements through new technological platforms, the issue of the digital divide will involve not just literacy, but cultural and attitudinal factors that shape interactions with both technology and politics.

REFERENCES

Bazar, B., & Boalch, G. (1997). *A preliminary model of Internet diffusion within developing countries*. Retrieved 12/12/2007, from http://ausweb.scu.edu.au/proceedings/boalch/paper.html

BBC. (2006). *Digital divide could be deepening*. BBC News, 26 October 2006. Retrieved 10/10/2007, from http://news.bbc.co.uk/go/pr/fr/-/1/ht/technology/6085412.stm

Castells, M. (1996). *The rise of network society. The Information Age: Economy, society and culture*, vol. 1. Oxford: Blackwell.

Castells, M. (2001). *The Internet galaxy: Reflections on the Internet, business and society*. Oxford: Oxford University Press.

Chan, W., & Wellman, B. (2004). The global digital divide – Within and between countries. *IT & Society*, *1*(7), 39-45.

Chowdury, G. (2004). *Access to information in digital libraries: Users and digital divide*. International Conference on Digital Libraries, New Delhi, India, 27 February 2004. Retreived 27/09/2007, from http://eprints.cdlr.strath.ac.uk/2622/

Cullen, R. (2001). Addressing the digital divide. *Online Information Review*, *25*(5), 311-320.

Cullen, R. (2003). The digital divide: A global and national call to action. *The Electronic Library*, *21*(3), 247-257.

De Hann, J. (2004). A multifaceted dynamic model of the digital divide, *IT & Society*, *1*(7), 66-88.

De Haan, J., & Rijken, S. (2002). The Internet and knowledge gaps: A theoretical and empirical investigation. *European Journal of Communication*, *7*(1), 65-84.

eGov. (2005). Statistics reveal European digital divide. *eGov Monitor*, 11 November 2005. Retrieved 14/11/2007, from http://www.egovmonitor.com/node/3499

Fallows, D. (2007). *China's online population explosion; What it may mean for the Internet globally...and for the US*. Pew Internet and American Life Project, July 2007. Retrieved 22/12/2007, from http://www.pewinternet.org/pdfs/China_Internet_July_2007.pdf

Fink, C., & Kenny, C. (2003). W(h)ither the digital divide. *info 5*(6), 15-24.

Glatung, J., & Ruge, M. (1965). The structure of foreign news: The presentation of the Congo, Cuba, and Cyprus Crises in Four Norwegian Newspapers. *Journal of International Peace Research*, *1*, 64-91.

Guðmundsdóttir. C. (2005). *Approaching the digital divide in South Africa.* NETREED Conference, Beitostølen, Norway, 5-7 Dec 2005. Retrieved 12/12/2007, fromhttp://www.netreed.uio.no/conferences/conf2005/GretaGudmundsdottir.pdfudmundsdottir.pdf

Hao, X., & Chao, S. (2004). Factors affecting Internet development: An Asian survey. *First Monday, 9*,(2). Retrieved 12/12/2007, from http://www.firstmonday.org/issues/issue9_2/hao/index.html

Hubregtse, S. (2005). The digital divide within the European Union. *New World Library, 106*(1210/1211), 164-172.

ITU Internet Reports. (2005). *The Internet of things*. WSIS Web site. Retrieved 14/10/2007, from http://www.itu.int/wsis/tunis/newsroom/stats/The-Internet-of-Things-2005.pdf

Kema, I. (2005). Globalization and the development of underdevelopment of the Third World*. Journal of Third World Studies*, Spring 2005.

Kubicek, H. (2004). Fighting a moving target. *IT & Society, 1*(6), 1-19.

Luyt, B. (2004). Who benefits from the digital divide? *First Monday, 9*(8). Retrieved 17/10/2007, from http:firstmonday.org/issues/issue9_8luyt/index.html

Luyt, B. (2006). Defining the digital divide: The role of e-readiness indicators. *Aslib Proceedings: New Information Perspectives, 58*(4), 276-291.

Martin, S. P., & Robinson, J. P. (2004). The income digital divide: An international perspective. *IT & Society, 1*(7), 1-20.

Martin, S. P., & Robinson, J. P. (2007). The income digital divide: Trends and predications for levels of Internet use. *Social Problems, 54*(1), 1-22.

Menou, M. J. (2001). The global digital divide: Beyond hICTeria. *Aslib Proceedings, 53*(4), 112 – 114.

Modoux, A. (2002). *The 'Digital Divide' could lead to the creation of a gigantic 'cyber ghetto' in the developing countries.* Retrieved 12/12/2007, from http://www.itu.int/wsis/docs/background/themes/digital_divide/modoux.pdf

Norris, P. (2000). *The worldwide digital divide; Information poverty and development.* Paper Presented at the Annual Meeting for the Political Studies Association of the UK, London, 10-13, April 2000. Retrieved 10/10/2007, from http://ksghome.harvard.edu/~.pnorris.shorenstein.ksg/acrobat/psa2000dig.pdf

Robin, K., & Webster, F. (1999). *Times of the techoculture: From the information society to the virtual life.* London: Routledge.

Robinson , J. P. (2003). Introduction to issues 4 and 5; Digital divides: Past, present and future. *IT & Society, 1*(1), i-xiv.

Sciadas, G. (2002). *Unveiling the digital divide*. Canada: Ministry of Industry

Stayaert, J., & Gould, N. (2004). The rise and fall of digital divide. In J. Graham, M. Jones, & S. Hick, (Eds.), *Digital divide and back: Social welfare, technology and the new economy*. Toronto: University of Toronto.

Van Dijk, D. (1999). *The network society: Social aspects of new media*. Thousand Oaks, CA: Sage.

Warschauer, M (2004). *Technology and social inclusion: Rethinking the digital divide*. Cambridge, MA: The MIT Press.

WSIS. (n.d.). *What is the state of ICT access around the world*? WSIS Web site. Retrieved 14/10/2007, from http://www.itu.int/wsis/tunis/newsroom/stats/

KEY TERMS

Digital Divide: The separation between the information rich, or haves, and information poor, or have-nots.

ICTs: Information and communication technologies, which are seen as the driving force of globalization and new forms of connectivity.

Information Society: The transition from the modern and industrial age in which modes of production, exchange, and social capital are increasingly defined through information

Network Society: The rise of the information society will see the emergence of a network society in which information and technology will enable the formation of networks and strategic planning.

Category: IT Education

Contemporary Instructional Design

Robert S. Owen
Texas A&M University-Texarkana, USA

Bosede Aworuwa
Texas A&M University-Texarkana, USA

INTRODUCTION

This article discusses the principles of two qualitatively different and somewhat competing instructional designs from the 1950s and 1960s, *linear programmed instruction* and *programmed branching*. Our hope is that an understanding of these ideas could have a positive influence on current and future instructional designers who might adapt these techniques to new technologies and want to use these techniques effectively. Although these older ideas do still see occasional mention and study (e.g., Brosvic, Epstein, Cook, & Dihoff, 2005; Dihoff, Brosvic, & Epstein, & Cook, 2004), many contemporary instructional designers are probably unaware of the learning principles associated with these (cf., Fernald & Jordan, 1991; Kritch & Bostow, 1998; McDonald, Yanchar, & Osguthorpe, 2005).

BACKGROUND

An important difference between these instructional designs is associated with the use of feedback to the learner. Although we could provide a student with a score after completing an online multiple-choice quiz, applications that provide more *immediate feedback* about correctness upon completion of each individual question might be better. Alternatively, we could provide *adaptive feedback* in which the application provides elaboration based upon qualities of a particular answer choice.

Following is a discussion of two qualitatively different instructional designs, one providing immediate feedback regarding the correctness of a student's answer, the other providing adaptive feedback based on the qualities of the student's answer. Suitability of one design or the other is a function of the type of learner and of the learning outcomes that are desired.

SOME CLASSIC CONCEPTS OF INSTRUCTIONAL DESIGN AND OUTCOMES

Although the idea of non-human feedback would seem to imply a mechanical or electronic device, other methods could be used. Epstein and his colleagues, for example, have used a multiple-choice form with an opaque, waxy coating that covers the answer spaces in a series of studies (e.g., Epstein, Brosvic, Costner, Dihoff, & Lazarus, 2003); when the learner scratches the opaque coating to select an answer choice, the presence of a star (or not) immediately reveals the correctness of an answer. Examples of the designs discussed next are based on paper books, but they are easily adaptable to technologies that use hyperlinks, drop-down menus, form buttons, and such.

Linear Programmed Instruction

The programmed psychology textbook of Holland and Skinner (1961) asked the student a question on one page (the following quote starts on page 2) and then asked the student to turn the page to find the answer and a new question:

> A doctor taps your knee (patellar tendon)
> with a rubber hammer to test your _____.

The student thinks (or writes) the answer and turns the page to find the correct answer ("reflexes") and is then asked another question.

Questions or statements are arranged in sequentially ordered *frames* such as the previous single frame. A frame is completed when the student provides a response to a stimulus and receives feedback. Skinner contended that this method caused learning through *operant conditioning*, provided through positive *reinforcement* for stimuli that are designed to elicit a correct answer (c.f., Cook, 1961; Skinner, 1954, 1958).

Skinner (and others who use his methods) referred to his method as *programmed instruction*, which incorporates at least the following principles (cf., Fernald & Jordan, 1991; Hedlund, 1967; Holland & Skinner, 1961; Skinner, 1958; Whitlock, 1967):

- Clear learning objectives.
- Small steps; frames of information repeat the cycle of stimulus-response-reinforcement.
- Logical ordered sequence of frames.

- Active responding by a student who works at his/her own pace.
- Immediate feedback to the response in each frame with positive reinforcement for correct answers.

A technique in programmed instruction is to help the student a great deal at first, and then gradually reduce the cues in latter frames; this is called *fading* (Fernald & Jordan, 1991; Reiff, 1980). If correct responding suggests that a student is learning at a quick rate, *gating* can be used to skip over frames that repeat prior information (Vargus & Vargus, 1991). The programmer is expected to use information about student performance to make revisions; if the student is not succeeding, then it is due to a fault of the program, not to an inability of the student (Holland & Skinner, 1961; Vargus & Vargus, 1991).

Programmed Branching

Crowder (e.g., 1959, 1963) and others (e.g., Pressey, 1963) were critical of Skinner's approach, arguing that students not only learn from knowing a correct answer, but also learn by making mistakes. Crowder distinguished between his *automatic tutoring device* and the Skinner-type *teaching machine*, proposing that the automatic tutoring device is more flexible in allowing the student to receive an explanation when an error is made. Crowder (1959, pp. 110-111) provides an example of how this approach could be used in a programmed textbook:

> In the multiplication of $3 \times 4 = 12$, the number 12 is called the product and the numbers 3 and 4 are called the
>
> Page 15 quotients.
> Page 29 factors.
> Page 43 powers.

In this *programmed branching* method of Crowder, the student is taken to one of several possible discussions depending on the qualities of the answer.

While Skinner's design would be expected to work only when stimuli elicit correct answers, Crowder's design allows for mistakes and must be designed to anticipate particular mistakes. Crowder believed that this method caused learning through *cognitive reasoning*. Whatever answer is chosen by the student, the programmed textbook (or machine) makes a *branch* to a discussion associated with issues relevant to the answer that was chosen. This is followed by a return to the same question if the student had made an incorrect choice, or a jump to new a *frame* containing the next question if the student had made a correct choice.

Learning Outcomes

Many issues have been raised over the years about programmed instruction methods. Reiff (1980) discussed several criticisms:

- It does not take into consideration the sequence of development and readiness to learn (e.g., children of different ages or children vs. adults).
- It develops rote learning skills rather than critical thinking skills.
- Students can in some implementations cheat.
- The encouragement to respond quickly could develop bad reading habits.

Crowder's *programmed branching* design, which has received far less attention and study than Skinner's ideas, would seem to answer at least some of these criticisms. Crowder's design provides an explanation to both correct and incorrect answers, so the learner is not rewarded for cheating or working too quickly. Since the explanation is tied to the learner's thinking at the time a choice was made, Crowder's design would appear to be better to develop critical thinking skills, but might not be so good at developing rote learning skills. Crowder's design would appear to be better suited to students who have a greater readiness to learn, while perhaps not so well suited to a student who is at an earlier stage of learning a subject.

The previous discussion suggests that each of these designs is useful, but that each is useful in different kinds of situations and that the *learning outcomes* of each approach might be different. Skinner's teaching machine, for example, might be more useful in situations where students are learning lists and definitions. The automatic tutoring device, on the other hand, might be more useful when the student is already at a higher level of understanding whereby s/he can now use reasoning to derive an answer, or in situations where the student understands that there are degrees of right and wrong without concrete answers. The Skinner-type teaching machine might be better suited to "lower-order" levels of learning, while the Crowder-type automatic tutoring device might be better suited to "higher-order" levels of learning.

Although many ideas have been proposed with regard to a hierarchical perspective on "lower" and "higher" levels of learning, the most well-known, "Bloom's Taxonomy" (A Committee of College and University Examiners, 1956), originated in about the same timeframe as the ideas of Skinner and Crowder. "Bloom's Taxonomy" proposes that the objectives of learning lie on a hierarchical continuum: (1) knowledge of terminology and facts, (2) comprehension of translation and paraphrasing, (3) application, (4) analysis, (5) synthesis, and (6) evaluation.

"Bloom's Taxonomy" is actually only Part I of a two-part work. The previously mentioned first part is known as the *cognitive domain*. Part II (Krathwohl, Bloom, & Masia, 1964) focuses on the *affective domain*: (1) willingness to receive ideas, (2) commitment to a subject or idea, (3) feeling that an idea has worth, (4) seeing interrelationships among multiple ideas, and (5) the integration of ideas as one's own.

FUTURE TRENDS

Fernald and Jordan (1991) discussed several reasons as to why programmed instruction might have fallen out of use since the decades of the 1950s and 1960s:

- It was seen to dehumanize the teaching process.
- Educators feared that it might be too effective and threaten their jobs.
- The importance of the learning principles was not understood.
- Applications were often not effectively designed.

Technology, economics, and attitudes have since changed. As economics and student demand push us to use distance education methods, the first two arguments would seem to become more diminished in the future.

It is hoped that this article assists in diminishing the latter two arguments by introducing instructional designers to the principles discussed in this article and by encouraging instructional designers to create more effective designs with regard to appropriateness for a particular student audience and with regard to the type and level of learning outcomes that are desired. By better understanding the past, we can better affect the future.

Curiously, there has been less attention devoted to Crowder's ideas of adaptive feedback than to Skinner's ideas of immediate feedback and reinforcement. We continue see occasional research devoted to related issues, such as issues of immediate vs. delayed feedback (e.g., Brosvic et al., 2005; Dihoff et al., 2004; Kelly & Crosbie, 1997) or of allowing students to keep selecting answers from a multiple-choice set until the correct answer is finally discovered (Epstein et al., 2003). However, we still can only speculate with regard to conditions under which a Skinner-style of instructional design would be better and when a Crowder-style of design would be better. It is hoped that this article generates greater awareness of and use of these designs in new technologies, but also that greater interest in these ideas will stimulate more research into the learning mechanisms associated with them.

CONCLUSIONS

New technologies such as Web browsers now make it relatively easy for educators with the most modest of skills to present instructional frames in a linear sequential ordering or as branches that are dependent on the student's selection of answers from a list. In adapting some of these older ideas to newer technologies, we hope that instructional designers will be better equipped to select appropriate methods by considering:

- the student's level of readiness for learning
- the basis for learning when different instructional designs are used
- the qualitatively different kinds of learning outcomes that are possible with different instructional designs

REFERENCES

A Committee of College and University Examiners. (1956). *Taxonomy of educational objectives—The classification of educational goals, Handbook I: Cognitive domain*. New York: David McKay Company, Inc.

Brosvic, G. M., Epstein, M. L., Cook, M. J., & Dihoff, R. E. (2005). Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: Learning in the classroom. *The Psychological Record, 55*(3), 401-418.

Cook, D. L. (1961). Teaching machine terms: A glossary. *Audiovisual Instruction, 6*(1961), 152-153.

Crowder, N. A. (1959). Automatic tutoring by means of intrinsic programming. In E. Glanter (Ed.), *Automatic teaching, the state of the art* (pp. 109-116). New York: John Wiley and Sons, Inc.

Crowder, N. A. (1963). On the differences between linear and intrinsic programming. In J. P. DeCecco (Ed.), *Educational technology: Readings in programmed instruction* (pp. 142-152). New York: Holt, Rinehart, and Wilson.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning is enhanced by immediate but not delayed feedback. *The Psychological Record, 54*(2), 207-231.

Epstein, M. L., Brosvic, G. M., Costner, K. L., Dihoff, R. E., & Lazarus, A. D. (2003). Effectiveness of feedback during the testing of preschool children, elementary school children, and adolescents with developmental delays. *The Psychological Record, 53*(2), 177-195.

Fernald, P. S., & Jordan, E. A. (1991). Programmed instruction versus standard text in introductory psychology. *Teaching of Psychology, 18*(4), 205-211.

Hedlund, D. E. (1967). Programmed instruction: Guidelines for evaluation of published materials. *Training and Development Journal, 21*(2), 9-14.

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior*. New York: McGraw-Hill Book Company, Inc.

Kelly, G., & Crosbie, J. (1997). Immediate and delayed effects of imposed feedback delays in computerized programmed instruction. *The Psychological Record, 47*(4), 687-698.

Krathwohl, D. R., Bloom, B. S., & Masia, B. (1964). *Taxonomy of educational objectives—The classification of educational goals, Handbook II: The affective domain*. New York: David McKay Company, Inc.

Kritch, K. M., & Bostow, D. E. (1998). Degree of constructed-response interaction in computer-based programmed instruction. *Journal of Applied Behavior Analysis, 31*(3), 387-398.

McDonald, J. K., Yanchar, S. C., & Osguthorpe, R. T. (2005). Learning from programmed instruction: Examining implications for modern instructional technology. *Educational Technology Research and Development, 53*(2), 84-98.

Pressey, S. L. (1963). Teaching machine (and learning theory) crisis. *Journal of Applied Psychology, 47*(1), 1-6.

Reiff, J. C. (1980). Individualized learning through programmed materials. *Education, 100*(3), 269-271.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24*(2), 86-97.

Skinner, B. F. (1958). Teaching machines. *Science, 128*(3330), 969-977.

Vargus, E. A., & Vargus, J. S. (1991). Programmed instruction: What it is and how to do it. *Journal of Behavioral Education, 1*(2), 235-251.

Whitlock, G. H. (1967). Programmed learning: Some nonconfirming results. *Training and Development Journal, 21*(6), 11-13.

KEY TERMS

Adaptive Feedback: Immediate feedback in the form of an explanation or discussion that is tailored to the qualities of the student's answer.

Automatic Tutoring Device: A device that uses programmed branching and adaptive feedback. Learning results from cognitive reasoning.

Cognitive Reasoning: Learning through the process of thinking about an issue; the student learns new ideas and relationships by relating an issue to previously learned material.

Frame: A small piece of information or a statement to which the student is exposed, such as a page with a single question. In linear programmed instruction, a frame includes a stimulus, a response, and reinforcement (positive feedback).

Hierarchy of Learning: The concept that learning can be sequentially ordered along a continuum from lower-order to higher-order. "Bloom's Taxonomy" is one of many that have been proposed.

Linear Programmed Instruction: A design whereby a series of frames are presented to the student in a specific sequential order. The student actively responds to stimuli in each frame and receives immediate feedback to that response. Learning results through operant conditioning.

Operant Conditioning: Learning through immediate positive feedback (reinforcement) regarding the correctness of an answer; the student learns to respond in a particular way to a particular question or issue (stimulus). Fading can be used by gradually reducing stimulus cues in subsequent frames when material is repeated.

Programmed Branching: A method whereby the student is taken to one of several possible explanations or discussions depending on the qualities of an answer that is given to a question. Gating is a simple skip of frames that repeat prior information when a student's answers suggest that the material has been adequately learned.

Teaching Machine: A device that uses linear programmed instruction whereby frames present a question followed by feedback of the correct answer. Learning results from reinforcement of the student's correct answer.

Category: IT Education

Contemporary Issues in Teaching and Learning with Technology

Jerry P. Galloway
Texas Wesleyan University, USA
University of Texas at Arlington, USA

INTRODUCTION

To speak of contemporary issues in instructional technology is like counting wave crests in a stormy ocean: they are changing quickly all the time. New technologies and new issues present themselves daily. Educators struggle with both the instructional integration of computing and developing the skills and knowledge necessary to use technology effectively (Lipscomb & Doppen, 2005). Why, after over 30 years of having computers in schools, are educators still having such difficulties?

Today's population is much more accustom to electronics, yet knowledge is weak, concepts are misunderstood, and the difficulties of teaching with technology seem as serious and convoluted today as ever before. The great physicist and thinker, Richard Feynman, offered some critical comments about the challenges of educators. "What happens is that you get all kinds of statements of fact about education, about sociology, even psychology — all kinds of things which are, I'd say, pseudoscience" (Feynman, 1999, p. 242). Today, we understand "more about education [but] the test scores are going down…we just don't understand it at all. It just isn't working" (p. 243). Being critical of how the scientific method is applied to education, Feynman's comments highlight how the study of teaching and learning yields limited or questionable results. Teacher trainers take their best guess on how to prepare teachers to use technology.

BACKGROUND

Educational computing is a relatively new discipline compared to mathematics and science. While the earliest uses of computers might have been by departments of mathematics, it quickly became important for virtually all teachers to become computer literate. But what exactly that entails was not exactly clear (Galloway, 1985) for learning and in society (Beaty & Tucker, 1987).

Microcomputer technology, primitive by today's standards, lacked user-friendly applications, any sort of consistent user interface, or easy-to-use telecommunications and interconnectivity. There was an early division between those who learned to program computers vs. those who focused more exclusively on applications software. Conceptual development, improvement of problem solving, and higher-order thinking skills in computing have been directly linked to the inclusion of Logo programming (Allen, 1993; Battista, 1994; Borer, 1993; Dalton & Goodrum, 1991) and BASIC programming (Overbaugh, 1993). Yet, in spite of an overwhelming need to operate early microcomputers through programming, educators focused instead on the actions and procedural tasks of specific applications (Galloway & Bright, 1987).

With this as a foundation, decades of training have followed in which educators have tried to master new devices and software. So, how long does it take to reach a point of nationwide competency, to develop the protocols of effective use, to establish the knowledge of how best to learn computing? Compared to centuries of science and mathematics, perhaps our 30-plus years do not seem so long.

EDUCATORS LEARN COMPUTING: A PROBLEM OF PERSPECTIVE

Our collective perspective on what it means to learn computing affect what goals we pursue and how we proceed. For example, the use of rubrics or portfolios were not commonly emphasized in education 30 years ago. Today, they are an accepted or at least popular tool for preparing educators (Galloway, 2006; Rural School and Community Trust, 2001). Does this represent progress or perhaps just a symptom of changing fads? Is this a function of real knowledge or mere opinions? This is again reminiscent of a Feynman (1999) criticism, as he suggests that professionals 30 years ago have as much right to a correct opinion as we have today, "to equally unscientifically come to a conclusion" (p. 243)—even if wrong.

Preparing Teachers

It is unlikely that educators younger than their mid-40s graduated high school without having computers in their education. There has been, since the late 1970s, a continual focus on the needs of teachers to learn and adapt to a technology-based profession.

Our attempt over the years to change educators into computer-literate professionals essentially failed. Many

will argue the point, as clearly there are countless success stories. But, with the exception of the techies and innovative pioneers, educators across the profession a generation ago did not, have not changed their basic approach to integrate technology.

Compared to in-service classes, college courses, training, or other options, an overwhelming majority of teachers maintain that their primary methods of learning computing was through self-study and personal experimentation (Galloway, 1997). It can be argued that teachers must assume a responsibility for advancing their technological knowledge and be engaged learners.

When taking a computer class, one must go beyond the prescribed activities. For example, it is not likely that one would be assigned the experience of losing a file or opening a file with the wrong program. These frustrations can be a very necessary part of learning. Far too often educators are passive and restrict their involvement to occasional and discrete enrichment offered through someone else's initiative. Delays and intermittent and partial commitments inhibit learning.

As an analogy, when this author was young, rock-n-roll music was still the choice of the young, but grandfather did not relate and found it quite distasteful. In elevators in 1968, one would hear music from Lawrence Welk and such. This author believed that if elders could simply understand and learn about rock-n-roll and what the artists were attempting to express musically that society could change and the music would be accepted. Today, one is likely to hear McCartney, Dillon, The Beatles, Buddy Holly, or many of the other artists that were objectionable in those earlier years. One might think that, indeed, things changed.

However, the point is that this did not occur because the elders were influenced or convinced. The younger, rock-n-roll generation did not change anyone. The elders were not convinced. No metamorphosis occurred. The young simply grew older and brought their music with them. As the elders died off, the young with a new culture replaced the old.

The same seems true for the computer-using generation. Our efforts a generation ago were ineffective. We have simply waited around while a new generation grows older bringing their technology-based lifestyle with them. Until our children have time to take their place, today's teachers are still introduced to computing as beginners.

Training vs. Education

What do current educators expect from computer training? If we accept that it is difficult to teach someone who does not want to learn, what do students expect from their training? Unfortunately, the most popular notion in instructional technology is that teachers are to be trained, not educated. More than mere semantics, teaching tends to emphasize *showing teachers how to use* technology — rather than facilitating insight, understanding, and conceptual development. In-service programs and college curricula emphasize only what teachers are expected to use rather than what might develop good concepts. Omitting programming is a classic example where teachers as end users of software never see the construction process or design methods behind what they are supposed to learn. Today's design tools (for Web pages and such) are a modern example of where these issues still apply.

Teaching for conceptual understanding and higher-order thinking skills should not be only a part of teaching programming (Tu & Falgout, 1995), but also a fundamental goal of instruction for beginners in computing. Skills and even performance standards can still fail to generate important understandings, perspectives, concepts — integrated knowledge — that all contribute a fundamental and critical basis for problem solving and adaptability.

Focusing on conceptual development will still involve procedures and tasks just as focusing on discrete skills will likely yield some insights and discoveries. But instruction should yield a more complete, fundamental understanding of computing. Most programs and perspectives fail to recognize this important viewpoint and instead pursue skills and competencies to the detriment of understanding, insight, and problem solving.

It is common in other disciplines to speak of *education* rather than *training.* Conceptual development is often the primary focus in the study of science (Trumper, 1997). Even when the preparation of teachers is described in terms of training, science concepts are emphasized, not skills (Thompson & Schumacher, 1995). In spite of the procedures and skills inherent in science and mathematics, students are guided toward the development of a conceptual understanding as they are educated — not trained.

A training model targets activities and the software teachers will use. Much like an airline reservations clerk must learn the keystrokes and procedures for prescribed tasks, educational computing is similarly conceived. An education model, on the other hand, calls for activities and experiences that will yield a deeper kind of learning. Keystrokes and software familiarity would be incidental to the more important yield of experiences, much like those in science and mathematics, that develop understanding, concepts, problem solving, and critical thinking skills. An education-based program would provide experiences because of their educational value regardless of whether they are part of an anticipated skill set. Skill sets, tasks, and the procedural rituals of training will inevitably change and evolve far beyond the scope of any training experience.

Student teachers can be part of the problem as they, very often, prefer the training model. Contrary to any real value or longevity of such an approach, a more involved education presents an undesirable challenge. They prefer to simply be shown what to do. Guided tasks, prescribed procedures,

and discrete tasks are all a matter of doing, not becoming. However, an education calls for change.

Acquiring mindless task sequences is often viewed by educators as success. Improved teaching is then viewed as having more complete checklists for more tasks. This *recipe mentality* of discrete procedural rituals ignores the need for discovery learning, transfer, and adaptability, and could be responsible for continuing the inhibited progress of the past 30 years.

Integration

Limiting factors for the integration of technology include funding, professional development, support for experimentation, and inadequate technology planning (Mehlinger & Powers, 2002). As Galloway (1997) examined technology adoption, it was learned that effective usage is related to the combination of both professional and personal adoption of technology. Virtually no one used technology in their classrooms where personal adoption was not combined with professional use.

It has been said that teachers exist only for the children. They express the sentiment that student needs are the primary, if not the only mission of teachers. It is easy, however, to draw the wrong conclusions from such a self-evident premise. For instructional technology, consistent with this perspective, trends have been directed away from empowering teachers, focusing instead on classroom integration. This may seem justified, but a serious problem remains: it is not reasonable to teach non-computer users to use technology in the classroom. Educating teachers to become computer-using, technology-competent professionals would more likely yield classroom integration as a matter of natural consequence.

While there are training programs targeting classroom integration, what success can be had if teachers are not computer literate or have never adopted computing in their lives? In other words, one must adopt technology as a life-changing metamorphosis. The approach of the past, training over education and rituals over holistic adoption, may likely continue the inhibited progress in integration and technological mastery.

Learning and Working

Everyone agrees that students must be prepared for a technological future, but perceptions vary widely on exactly how to achieve that. This is a challenge of pedagogy.

Computers have been perceived as tools (Beaty & Tucker, 1987) and have been used in that fashion. An alternative view might suggest that the computer is not a tool at all. As a tool might be selected or discarded based on a particular need, technology is too often viewed as independent from everyday life powered-up if the need is sufficiently demanding. Instead, computers are perhaps best viewed as a complete *environment.* It is where we live, work, and play. It is the medium of our planning, our creativity, and an extension of both our short-term and long-term memory. This thinking places different expectations on educators than has traditionally been made. The notion that one can remain a non-computer-using person while merely executing discrete tasks as needed must change.

FUTURE TRENDS

How has learning changed and what is *learning* in the modern tech-based world? Learning is far too often viewed by teachers as a matter of acquiring information. Teachers deliver information to the students who in turn sit on it for a period of time only to hand it back to teachers again in some form of quiz or standardized test. If students regurgitate and return the information accurately, they are said to have *learned.* In fact, students assume this is what is expected of them and resist anything more personally demanding. The notion that they must change, must invent, or synthesize is foreign to them.

Learning is no longer about the acquisition of information. We have information. It is possible to find the average price of a hang guilder, the architect of the Brooklyn Bridge, design plans for a new home, or a translation of the Dead Sea Scrolls in mere minutes — all while sitting in a hotel lobby or even lying in bed in a dormitory room. Acquisition of information is neither the problem nor the goal. Learning to think is the real challenge. Education has become about skill development with demonstrable competencies rather than about becoming smarter or learning to think. Education, as distinct from training, should improve problem-solving abilities, critical thinking abilities, developing an understanding, learning to discriminate and make good choices, and developing a contextual intuition.

Computing students do not want to have to explore and discover, wanting instead to be shown how to execute procedures. Alluded to earlier, this amounts to distributing recipes for subsequent replication. Being ready for tomorrow's computing world depends on understanding, problem-solving skills, and the ability to adapt to the unknown, not on knowing procedures in some software program.

Technology continues to develop faster than anyone can sufficiently learn it. Merely being able to operate the functions and tools in a program is usually considered a success. But achieving a deeper knowledge of how best to adopt, integrate, and teach in a world of technology is quite a different thing.

The Distant Future

Does today's science fantasy help to create the reality of tomorrow? That is of course debatable, but at least imagi-

nation does its part on one side of the evolution equation. Clearly, our children will experience amazing and incredible advancements that today seem like science fiction.

So, extend your vision to consider the following: a kind of futuristic electronic *bubble* as a kind of spherical energy shell that would surround one's head and face. The shell might not be spherical and could instead extend vertically downward in front of the face and chest somewhat like a large energy shield in front of the body. Generally, it may be maintained and kept active throughout the waking hours.

The shell or *e-bubble* would be generated by a multi-functional microchip and would act like a virtual outer skin or electronic membrane extending perhaps 10 to 12 inches in front of the body. Perhaps the orientation of the membrane (up/down, sections, areas, etc.) could be determined by detecting and interfacing with an electrical field from the heart or brain, much like an electrocardiogram or electroencephalogram. This would be important to establish a directional configuration since a microchip might be located more on convenience or medical necessity. The e-bubble should maintain a functional orientation to the body. The membrane can function in sections, quadrants, or areas, as well as operating as a whole or singular entity.

The e-bubble would serve as a communications interface for all sorts of input and output in work, learning, and recreation. The microchip and power supply might be worn as external hardware, like in a necklace, belt buckle, or collar. Perhaps the electronic membrane might exist in the form of a mere hologram projection from specialized glasses. Today's military pilots see electronic projections of critical data superimposed on their natural view of their environment during flight. Some hardware today can feed visual information into one eye, leaving the other eye normal as the brain integrates the two views. The integration of visual and auditory hardware with the body has already begun its prosthetic progression from the separate and independent technologies of yesterday's cathode-ray tube (CRT) and today's plasma flat screens.

Eventually, the development of the e-bubble would evolve beyond the independent device carried in hand. Even with the convenience of a wristwatch or techno-necklace, such devices are external and thus their service to our lives is an add-on not truly integrated and natural. The e-bubble will evolve beyond a separate prosthetic to such a size and state that it becomes an implant no more intrusive than an inner ear replacement.

Images displayed in the membrane fields would include all of the variations we know of today: pictures, graphics, text, color, and of course, full-motion video. One can imagine that these fields appear from the back side, viewed by others, as opaque with no detail or instead translucent with imagery appearing in reverse.

One can imagine interactive fields for drawing, writing, or other tactile manipulations. Perhaps an image might be relocated in the field matrix to open or power another area for a secondary purpose. The various fields might provide a multi-tasking experience of work and play, business and entertainment, or integrated learning and study experiences.

It is really an extension of what has already occurred. Star Trek and many other sources of imagination today illustrate not just an exciting possibility but an inevitable reality.

CONCLUSION

Are we to continue passively as followers of our children or step up as leaders, which the nobility of our profession demands? The overwhelming theme of the past 30 years is that training for discrete tasks must be replaced by holistic adoption and education. Being successful at computing is not a function of memorized procedures or specific skill sets. Procedural rituals, however conveniently arranged or exhaustively accounted, cannot substitute for intuition, problem solving, and a deeper understanding of computing.

The future depends on adaptability and learning transfer, which again attest to the inadequacy of mere training. It will be an exciting future, but challenging to us all. To quote Peter Drucker, 20th-century business pioneer, "The best way to predict the future is to create it."

REFERENCES

Allen, J. (1993). The impact of cognitive styles on the problem-solving strategies used by preschool minority children in Logo microworlds. *Journal of Computing in Childhood Education, 4*(3-4), 205-217.

Battista, M.T. (1994). Research into practice: Calculators and computers: Tools for mathematical exploration and empowerment. *Arithmetic Teacher, 41*(7), 412-417.

Beaty, J.J., & Tucker, W.H. (1987). *The computer as a paintbrush.* Columbus, OH: Merrill.

Borer, M. (1993). *Integrating mandated Logo computer instruction into the second grade curriculum.* MS Practicum Report, Nova University.

Dalton, D.W., & Goodrum, D.A. (1991). The effects of computer programming on problem-solving skills and attitudes. *Journal of Educational Computing Research, 7*(4), 483-506.

Feynman, R.P. (1999). Richard Feynman builds a universe. In J. Robbins (Ed.), *The pleasure of finding things out: The best short works of Richard P. Feynman* (pp. 225-243). New York: Basic Books.

Galloway, J.P. (1985). *What is computer literacy? Proceedings of the Texas Computer Education Association Area IV Fall Conference,* Houston, TX.

Galloway, J.P. (1997). How teachers use and learn to use computers. in *Technology and teacher education annual journal, 1997.* Charlottesville, VA: Association for the Advancement of Computing in Education.

Galloway, J.P. (2006). Electronic portfolios for educators. *International Journal of Arts and Sciences, 1*(1), 10-13.

Galloway, J.P., & Bright, G.W. (1987). Erroneous conceptions of computing concepts. In J.D. Novak (Ed.), *Proceedings of the Second International Seminar: Misconceptions and Educational Strategies in Science and Mathematics* (vol. 1, pp. 206-219). Ithaca, NY: Cornell University.

Lipscomb, G.B., & Doppen, F.H. (2005). Climbing the stairs: Pre-service social studies teachers' perceptions of technology integration. *International Journal of Social Education, 19,* 2.

Mehlinger, H.D., & Powers, S.M. (2002). *Technology and teacher education: A guide for educators and policymakers.* Boston: Houghton Mifflin.

Overbaugh, R.C. (1993). *A BASIC programming curriculum for enhancing problem-solving ability.* Evaluative Report, Darden College of Education, Old Dominion University, USA.

Rural School and Community Trust. (2001). *Assessing student work.* Retrieved January 8, 2007, from *http://www.ruraledu.org/site/c.beJMIZOCIrH/b.1389103/apps/s/content.asp?ct=838177*

Thompson, G.W., & Schumacher, L.G. (1995). Implications of integrating science in secondary agricultural education programs. *Proceedings of the American Vocational Association Convention,* Las Vegas, NV.

Trumper, R. (1997). The need for change in elementary school teacher training: The case of the energy concept as an example. *Educational Research, 39*(2), 157-174.

Tu, J.-J., & Falgout, B. (1995). Teaching if-then structures: An integrated approach. *Learning and Leading with Technology, 23*(3), 26-28.

KEY TERMS

Computer Literacy: The ability to effectively use computer technology to solve problems and efficiently meet personal and professional needs.

Education: Contrary to mere training, the process engaging in supportive and generative experiences for acquiring the broader understanding and mastery.

Educational Computing: Full range of uses of computers pursuant to conducting the profession.

Instructional Technology: The broader field of studying the use or related issues of all technologies in education.

Integration: The effective, instructional use of technology in the classroom.

Learning: Contrary to the acquisition of mere facts, and more than acquiring discrete skills and competencies, learning is the development of knowledge, conceptual understanding, and critical thinking abilities in a prescribed context.

Science Fiction: Imagining future developments in technology and the human-machine interface, including living and working in virtual worlds.

Training: Limited and highly specific instruction for learning discrete tasks and procedural rituals.

Contemporary IT-Assisted Retail Management

C

Herbert Kotzab
Copenhagen Business School, Denmark

INTRODUCTION

Retailing can be defined in two ways, either as a set of functions that adds value to products/services that are sold to end users (functional understanding of retailing) or as a specific institution within a marketing channel that executes retail functions (institutional understanding). The functional view explains retailing as an exchange activity in order to connect a point of production with a point of consumption. These exchange processes refer to (see Kotzab & Bjerre, 2005):

- Marketing processes, including all activities that provide a customized set of products/services as demanded by customers/consumers (which is basically known as offering a customer-oriented assortment in terms of quality and quantity)
- **Logistics** processes, including all activities that help to transfer this specific set of products/services to the markets (such as transportation, breaking bulk and inventory management)
- Assisting processes, which refer to all activities that facilitate a purchase (such as credit function, promotion or advice function).

The orchestration of these functions leads to various types of retail formats such as store-based retailers (e.g., hypermarkets or category killers), non-store-based retailers (e.g., mail-order retailing or electronic commerce) and hybrid retailers (e.g., home delivery services) (Coughlan et al., 2006).

Retailing plays a vital role in today's economy, but many retailing companies face economic pressure as they operate predominantly in mature and stagnant markets (e.g. Seth & Randall, 2001). In order to face these specific challenges, retailing companies adapt strategies that allow them to gain economies of scale by offering highly customized solutions to their customers (see Table 1).

These strategies are built upon the latest developments in information technology (IT) and are therefore called IT-assisted retail management strategies. The following chapter presents an overview to contemporary IT-based retail business models and frameworks that show how IT has created a new mandate for retail management. IT is defined here as the hardware and software that collects, transmits, processes and circulates pictorial, vocal, textual and numerical data/information (e.g., Hansen & Neumann, 2005; Chaffey, 2004).

The following IT is of special interest in relation to IT-assisted retail management:

- Mobile data capturing terminals, light pens, bar code readers, EAN.UCC 128, labels, disks, chip cards, RFID, EPC, sensors to collect information
- Data base systems, tapes, CDs, DVDs, optical disks, Document Retrieval Systems to store information
- PCs, Information Retrieval, Decision support systems, Expert systems: MIS, EIS, MSS, ESS to process information
- Services (e.g., fax, email, EDI, web-EDI, FTP, WAIS, WWW, SMTP, TCP/IP, XML, VAN, GPS), networks (videoconferencing, teleconferencing, voicemail, ISDN, LAN, WAN, fiber optic, intra-, inter- and extranet) and devices (e.g., phones, TV, radio, fax machine, PC, PDA) to transmit information

The increasing use of these technological possibilities has led to major changes in the strategic management of distribu-

Table 1. Cornerstones of contemporary IT-based retail management (see Kotzab & Bjerre, 2005)

IT-based retail marketing strategies	IT-based retail logistics systems
• Re-engineered IT-driven retail formats, allowing for a customized shopping experience • Development of new retail channels, (e.g., Internet-based retail formats to address new customer segments) • Category management, in order to offer client-oriented sets of products, resulting from a joint-planning process with manufacturers based on real-time accessed client data	• The implementation of just-in-time-oriented replenishment systems by connecting the electronic point-of-sale- (EPOS) systems with the manufacturers' ERP-systems • The execution of IT-driven distribution center operations with no-inventory-holding transit terminal structures • The realization of Vendor-Managed-Inventory-Programs on a continuous replenishment basis to reduce inventory levels and to improve order cycles

tion channels as the layers are compressed and the distances between the first and last echelon of the channel are reduced (e.g., Porter, 2001 or Coughlan et al., 2006). Leading retailers are aware of these possibilities and have implemented customized POS-data based marketing strategies (IT-based retail marketing) and demand-synchronized replenishment systems (IT-based retail logistics).

BACKGROUND

IT-Based Retail Marketing Processes

Business practice shows a huge variety of IT-based retailing marketing strategies including the use of smart cards, theft prevention, self-check-out systems, web-kiosks and/or merchandise planning systems. The common goal of all these strategies is to obtain better information on consumer behavior in order to improve customer service. In that sense IT-based retail marketing affects all retail areas from the sales floor to the back offices (Kotzab et al., 2003a; Kotzab & Bjerre, 2005).

IT influences the layout and the atmosphere of a retail store by optimizing the link between sales productivity and consumer excitement (Nymphenburg, 2001) as the following examples show:

- Metro operates the future store concept that promotes technologically-driven innovations in IT-assisted retail marketing as offered by the combined use of RFID, electronic shelf labels, self check out systems, personal shopping agents, instore media such as info terminals, loyalty cards, personal shopping assistant for shoppers, personal digital assistant for employees and intelligent scales (e.g., Metro, 2003, Kotzab & Bjerre, 2005).
- Rewe Austria operates an outlet in Purkersdorf (nearby Vienna), where shoppers self register their purchases via self scanning devices (see Kotzab et al., 2003a). Rewe also uses the "communicating" shopping cart WATSON, which uses a technology based on radio frequency. Whenever passing a "labeled" shelf, the cart announces a message to the shopper (Atlas New Media, 2001).
- Since 2004, Spar-Austria has run a modern supermarket in Mattighofen (near the city of Salzburg) with self-check-out, cash-back terminals, instore-videos and intelligent scales (Spar, 2004).
- Zielpunkt/Plus of the German Tengelmann-Group installed self-check-out systems and cash-back terminals in one Viennese store (Weber, 2004).
- Carter & Lomas (2003) present the Sainsbury store in Hazelgrove (UK) and the Darty store in France, that both represent the state-of-the art of technology driven store layout.
- Weber (2006) reports on the experiences of Belgium Delhaize group which has used Wincor-Nixdorf handheld self-scanning devices since 1997 in their stores which allow customers to scan their items while they are shopping. According to Delhaize, 26 % of all sales are registered with those systems. Mobile self-scanning has allowed Delhaize to install quick shopping lanes within the stores in order to increase throughput times of customers.
- The French retailer Auchan is testing in France so-called scan & bag technology, which is an automated cash-desk system. Auchan operates 166 self-service cash-desks in different stores in Italy (Weber, 2006).

IT has also changed the organizational set up from hierarchical to hybrid/borderless arrangements such as category management (CM) (Gruen, 2002). CM is a joint retailer and manufacturer effort that involves managing product categories as business units and customizing them on a store-by-store basis to satisfy end-user needs (Dussart, 1998). The purpose is to identify those combinations of products that make up consumers' expectations. CM replaces traditional product focused strategies (e.g. brand management) and allows retailers and suppliers to faster react to shifts in the market place (Schröder, 2003). The increasing use of data warehousing and data mining approaches helps to use the scanner data more efficiently in order to establish customer-oriented assortments (Chen et al., 2000). Recently, Kahler & Lingenfelder (2006) were able to identify a strong relationship between the category value for money and store loyalty as CM incorporates the consumers' views and perceptions.

Finally RFID-technology is not only going to retail marketing but also logistics dramatically (Finkenzeller, 2003). Metro's future store concept shows that self scanning processes can be replaced by RFID which reduces waiting times for customers. Du Mont & Hoda (2006) present the case of the Japanese Mitsukoshi who uses RFID technology for an intelligent fitting room in order to enhance customer service. Especially industrial initiatives such as the EPC-Global, a joint venture between EAN International and the UCC, have developed so-called electronic product codes (EPCs), which will increase the diffusion of RFID-technology in the retail industry (see Jordan & Adcock, 2006). EPC is an RFID-based advanced UPC-bar code with the benefit of being able to identify articles at the item level uniquely (Verisign, 2004). The power of such a code can be illustrated by the following quote: "Using this EPC, members of the supply chain can thus identify and locate information about the manufacturer, product class, and instance of a particular product. Depending on the type of tag, EPC can be used to uniquely identify up to 268 million unique manufacturers, each with 16 million types of products. Each unique product can include up to 68 billion individual items, meaning the format can be used to identify hundreds of trillions of unique items" (Verisign, 2004, p.2).

IT-Based Retail Logistics Processes

Logistics in a retailing context refers to multi-echelon logistics systems with many nodes from the original supplier to the final store destination (Kotzab & Bjerre, 2005). The use of specific IT in retail logistics, such as EDI (e.g., EANCOM), barcodes (e.g., EAN/UCC), scanner technology, RFID-technology (e.g., EPC) and XML has converted traditional retail logistics systems into just-in-time-oriented lean retail supply chain management systems. A chain-wide use of technology allows harmonization and synchronization of logistics operations between retailers and their suppliers and has given retailers additional profitability as such systems operate on a pull instead of a push base. Consequently, the total bullwhip effect in such channels is reduced (Lee & Whang, 2002).

The major IT-assisted retail logistics processes are cross docking and continuous replenishment (Kotzab & Bjerre, 2005) on a retail level and collaborative planning forecasting and replenishment (CPFR) on a retail-supplier level (Skjoett-Larsen et al., 2003).

Cross docking is the meta-term for all IT-related flow-through activities within a distribution center that provide tailor-made deliveries on a just-in-time basis. Different vendors deliver full truckloads of their goods to a retailer's transit terminal (a re-engineered distribution center; also called a transshipment point). There, the goods are then consolidated and/or broken to vendor-integrated POS-required smaller delivery units (see Figure 1).

The basic idea of Cross docking is to avoid inventory at the distribution center level, which leads to a replacement of stock holding activities through sorting, transportation and handling activities, which are controlled by increased use of IT (e.g., EAN/UCC 128 in combination with EANCOM messages). The relevant literature offers various types of cross docking operations depending on whether vendors deliver pre-labeled units to the distribution center, full pallets or cases or customized mixed or non-mixed pallets (e.g. Napolitano, 2000).

While cross docking refers to IT-assisted logistics at a distribution center level, vendor managed inventory (VMI) or continuous replenishment (CRP) refer to all cooperative forms of inter-firm automated replenishment programs where the common goal is the automatic reinforcement of the supply of goods and the transfer of the burden of responsibility of storage from a retailer to a vendor.

Within any VMI/CRP setting, retailers re-transfer the inventory competence back to their vendors by agreeing on certain average inventory levels at distribution center level, service levels and/or other arrangements like the reduction or avoidance of out-of-stock-situations (Raman et al., 2001). Within VMI/CRP the former one-to-one relationship (where a seller and a buyer individually represented the goals of their companies') is replaced by inter-departmental, inter-organizational teams, which are responsible for the ordering process (Kotzab & Bjerre, 2005).

A further development in IT-based retail logistics can be seen in the use of CPFR that is defined by the Voluntary

Figure 1. Basic cross docking operation (Kotzab & Bjerre, 2005)

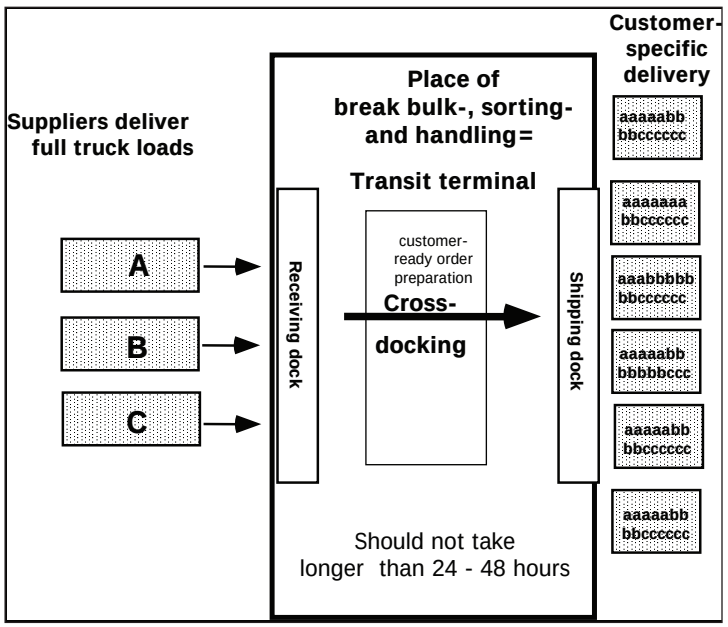


Figure 2. Basic continuous replenishment process in an ECR-environment (Glavanovits & Kotzab, 2002)

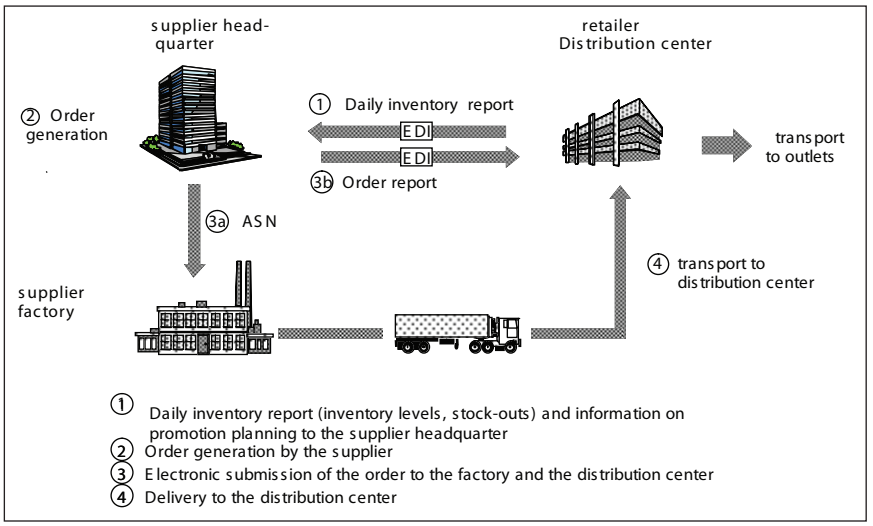


Table 2. CPFR-implementation guideline (see Kotzab et al., 2003b)

Step	Activity	Description
1	Develop front-end agreement	A front-end agreement is developed; criteria for success are established; identification of the CPFR project owners in the companies; financial reward and contribution system is agreed upon.
2	Create joint business plan	A joint Business Plan for the areas of collaboration is created Plans regarding advertising campaigns etc.
3-5	Sales forecast collaboration	The parties get together with each of their customer demand prognoses to establish a common prognosis. In case of deviation from forecast the partners meet to discuss deviations and to update the common forecast.
6-8	Order forecast collaboration	The partners share replenishment plans and discuss deviations and constraints.
9	Order generation	The reordering process/goods flow is initiated. Result data is discussed (POS, orders, shipments). Forecast deviation and stock level problems are identified and solved.

Interindustry Commerce Standards (VICS) as "a collection of new business practices that leverage the Internet and electronic data interchange in order to radically reduce inventories and expenses while improving customer service" (VICS, 1998a). This definition suggests that the Internet and electronic data interchange (EDI) are substantial prerequisites of CPFR.

VICS also crafted guidelines in the form of a nine-step model detailing how to implement CPFR (www.cpfr.org; VICS, 1998b; see Table 2).

These steps can be seen as a "cyclic and iterative approach to derive consensus based supply chain forecasts" (Fliedner, 2003, p. 15). The CPFR process builds to a large extent on the exchange of information among collaboration partners. This exchange of information can be carried out through the use of various technologies such as electronic data interchange (EDI), private networks or the Internet (XML). For the processing of information a large number of software programs have been developed to support the CPFR processes (e.g., Syncra, Logility, Manugistics; i2 Technologies, E-Millennium, E3, J.D. Edwards, Numetrix og Excentric, SAP, Oracle, Baan or Peoplesoft) (see Kotzab et al., 2003b).

FUTURE TRENDS

The Internet is going to revolutionize retailing. The challenge for e-tailing is logistics as the most prominent of Tesco.com shows. Consumer Direct Services (Corbae & Balchandani, 2001) will be more and more demanded, which will increase the number of single-purchase deliveries to the end-users' homes. The total home delivery market in the grocery industry is expected to be over 100 billion Euro, including over 900 million deliveries per year to over 35 million households.

IT-assisted home delivery concepts try to optimize the last mile to the consumer, which is the most expensive part of the total supply chain. There are plenty of home delivery systems, which either deliver the products directly to the home or where customers collect their purchased goods at certain pick-up-points (e.g., Pflaum et al., 2000, Punikavi et al., 2001). Tesco.com, however, goes a different way by serving consumer needs directly from their stores instead of installing special home-delivery distribution centres.

CONCLUSION

The chapter described the consequences of the use of IT in the field of retailing. It was shown how IT can be used in retail marketing to re-engineer retail formats in order to allow for a customized shopping experience. IT-assisted retail logistics refers to just-in-time-oriented demand synchronized delivery systems. The increased use of such systems will lead in the future to more hybrid retail management strategies, where the organizational borders between retailers, suppliers and even consumers will disappear.

REFERENCES

Atlas New Media (2001). *Watson, der sprechende Einkaufswagen (Watson, the talking shopping cart)*. Hamburg.

Carter, D., & Lomas, I. (2003). *Store of the Future, presentation at the 8th official ECR-Europe Conference*, Berlin, May 15, 2003.

Corbae, G., & Balchandani, A. (2001). Consumer Direct Europa - Erfolg durch Zusammenarbeit (Consumer Direct Europe—Success by cooperation). In D. Ahlert, J. Becker, P. Kenning, & R. Schütte, (Eds.), *Internet & Co. in Retail. Strategies, Business Models and Experiences. Series: Strategic Management fort Consumer goods industry and retail,* (pp.63-78). Berlin-Heidelberg: Springer.

Chaffey, D. (2004). *E-Business and E-Commerce Management* (2nd edition). Prentice Hall, Financial Times.

Coughlan, A., Anderson, E., Stern, L., & El-Ansary, A. (2006). *Marketing channels* (7th edition). Upper Saddle River, NJ: Prentice Hall.

Du Mont, S., & Hoda, S. (2006). *Exploiting information technology to drive business results—Today and tomorrow*. Presentation at the ECR-Europe Conference, Stockholm, Sweden, May 30, 2006.

Dussart, C. (1998). Category management: Strengths, limits and developments. *European Management Journal, 16*(1), 50-62.

Finkenzeller, K. (2003). *Fundamentals and applications in contactless smart cards and identification* (2nd edition). Wiley & Sons LTD.

Fliedner, G. (2003). CPFR: An emerging supply chain tool. *Industrial Management and Data Systems, 103*(1), 14-21.

Glavanovits, H., & Kotzab, H. (2002). *ECR Kompakt (ECR Compact)*. Wien: EAN-Austria.

Gruen, T. (2002). The evolution of Category Management. *ECR Journal International Commerce Review, 2*(1), 17-25.

Hansen, H.R. & Neumann, G. (2005). *Computer Science in Business and Management, 8th edition*. Stuttgart: UTB.

Jordan, P., & Adcock, C. (2006). *What is it? Where is it and where has it been?—EPC—The answer to your basic supply chain questions*. Presentation at the ECR-Europe Conference, Stockholm, Sweden, May 30.

Kahler, B., & Lingenfelder, M. (2006). Category management: When 1 + 1 = 3. *ECR-Journal, 6*(1), 64-69.

Kotzab, H., & Bjerre, M. (2005). *Retailing in a SCM-Perspective*. Copenhagen: CBS Press.

Kotzab, H., Schnedlitz, P., & Neumayer, K. (2003a). Contemporary IT-assisted Retail Management. In Joia, L. (Ed.), *IT-based management. Challenges and solutions* (pp.175-203). Hershey, PA: Idea Group Publishing.

Kotzab, H., Skjoett-Larsen, T., Andresen, C., & Thernoe, C. (2003): Logistics managers' perception and viewpoint to interorganizational supply chain collaboration by CPFR. In T. Spengler, S. Voss, & H. Kopfer (Eds.), *Logistics Management, Processes— Systems—Education*) (pp.65-78). Berlin: Springer.

Lee, H., & Whang, S. (2001). Demand chain excellence. A tale of two retailers, *Supply Chain Management Review, 5*(2), 40-46.

Metro (2003). Retrieved May 25, 2003, from www.future-store.org

Napolitano, M. (2000). *Making the move to cross docking*. Warehousing Education and Research Council.

Nymphenburg (2001). *Electronic media 'conquer' retail*. Retrieved August 14, 2001, from www.nymphenburg.de

Pflaum, A., Kille, C., Mirko, W., & Prockl, G. (2000). Home delivery services for groceries and consumer goods for every day usage in the internet—The last mile to the consumer from a logistical perspective.

Porter, M. (2001). Strategy and the Internet. *Harvard Business Review, 79*(3), 63-78.

Punakivi, M., Yrjölä, H., & Holmström, J. (2001). Solving the last mile issue: Reception box or delivery box? *International Journal of Physical Distribution & Logistics Management, 31*(6), 427-439.

Raman, A., DeHoratius, N., & Zeynep, T. (2001). Execution: The missing link in retail operations. *California Management Review, 43*(3), 136-152

Schröder, H. (2003). Category Management—Eine Standortbestimmung (Category Management, a positioning). In Schröder, Hendrik (Ed.), *Category Management. From Business Practice for Business Practice—Concepts—Cooperation—Experiences* (pp.11-38). Frankfurt am Main: LebensmittelZeitung.

Seth, A., & Randall, G. (2001). *The Grocers* (2nd edition). London, Dover: Kogan Page.

Skjoett-Larsen, T., Thernoe, C., & Andresen, C. (2003). Supply chain collaboration, *International Journal of Physical Distribution and Logistics Management, 33*(6), 531-549.

Spar (2004). *Premiere in Austrian grocery retailing in Mattighofen: Spar uses self-check out, cash-back and intelligent produce scales*. Press release from October 21, 2004.

Verisign (Ed.) 2004. *The EPC Network: Enhancing the supply chain*. Verisign.

Voluntary Interindustry Commerce Standards (VICS) (1998a). VICS Helps Trading Partners Collaborate to Reduce Uncertainty With CPFR Guidelines. CPFR Press Release October, 1998, http://www.cpfr.org/19981008.html

Voluntary Interindustry Commerce Standards (VICS) (1998b). *Collaborative Planning Forecasting and Replenishment*. Voluntary Guidelines.

Weber, B. (2004). *Zielpunkt with self scanning,* Lebensmittelzeitung, October 15.

Weber, B. (2006). *Delhaize commends self scanning,* Lebensmittelzeitung, February 2.

KEY TERMS

Barcodes: Simple form of optical character recognition, where information is encoded in printed bars of relative thickness and spacing. RFID combines this technology with radio frequency.

Continuous replenishment systems: Automated order retrieval systems which reduce out-of-stock situations at the point of sales by linking Electronic Point of Sales Systems with supplier factories.

Cross docking: Just-in-time flow through operations in a distribution center which transform incoming deliveries as fast as possible to customer specific outgoing deliveries.

Electronic Data Interchange (EDI): Meta-term for a multitude of different electronic message standards that allow a computerized and highly structured low error communication between computers. A "merge" between EDI and Internet technology can be recently observed by the upcoming of web-based EDI solutions, where on EDI-partner does not have to install EDI but use common web browsers to communicate via EDI.

Electronic Product Code (EPC): RFID-based product identification standard that developed from bar codes. EPC is managed by EPC Global Inc., which is a subsidiary of EAN.UCC. EPC numbers are able to accommodate all EAN.UCC keys.

Electronic Shelf Labels: Price tags that provide accurate pricing due to electronic linkage between the shelves and the checkout system. The technology is based on radio-frequency, infra-red and/or WLAN linking a master check-out with the shelves.

Intelligent Scale: A scale that is equipped with a special camera and identification software. Based on an object's structure, size, color and thermal image, the scale automatically recognizes the item, weighs it and prints out a price tag.

RFID (Radio Frequency Identification): Form of automated radio frequency based identification of objects. RFID systems consist of an antenna, a transceiver for reading radio frequency and to transfer information, a processing device and a transponder.

Scan & Bag: Special application of a self-check-out-system.

Scanner: Electronic devices that convert barcode information into digitized electronic images.

Self Check Out Systems: Self check out systems can occur at the end but also during shopping processes whenever cash-desk operations are 'outsourced' to consumers. In that case, consumers self register their items with specific scanning devices.

Virtual Shopping Assistant: Small mobile computer with a touch screen and bar-code scanner that can be installed to a shopping trolley and can serve as a personal shopping advisor (e.g., offering the customized shopping list).

Content-Based Image Retrieval

Alan Wee-Chung Liew
Griffith University, Australia

Ngai-Fong Law
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

With the rapid growth of Internet and multimedia systems, the use of visual information has increased enormously, such that indexing and retrieval techniques have become important. Historically, images are usually manually annotated with metadata such as captions or keywords (Chang & Hsu, 1992). Image retrieval is then performed by searching images with similar keywords. However, the keywords used may differ from one person to another. Also, many keywords can be used for describing the same image. Consequently, retrieval results are often inconsistent and unreliable.

Due to these limitations, there is a growing interest in content-based image retrieval (CBIR). These techniques extract meaningful information or features from an image so that images can be classified and retrieved automatically based on their contents. Existing image retrieval systems such as QBIC and Virage extract the so-called low-level features such as color, texture and shape from an image in the spatial domain for indexing.

Low-level features sometimes fail to represent high level semantic image features as they are subjective and depend greatly upon user preferences. To bridge the gap, a top-down retrieval approach involving high level knowledge can complement these low-level features. This articles deals with various aspects of CBIR. This includes bottom-up feature-based image retrieval in both the spatial and compressed domains, as well as top-down task-based image retrieval using prior knowledge.

BACKGROUND

Traditional text-based indexes for large image archives are time consuming to create. A domain expert is required to examine each image scene and describe its content using several keywords. The language-based descriptions, however, can never capture the visual content sufficiently because a description of the overall semantic content in an image does not include an enumeration of all the objects and their properties. Manual text-based annotation generally suffers from two major drawbacks: (i) content mismatch, and (ii) language mismatch. A content mismatch arises when the information that the domain expert ascertains from an image differs from the information that the user is interested in. When this occurs, little can be done to recover the missing annotations. On the other hand, a language mismatch occurs when the user and the domain expert use different languages or phrases to describe the same scene. To circumvent language mismatch, a strictly controlled set of formal vocabulary or ontology is needed, but this complicates the annotation and the query processes. In text-based image query, when the user does not specify the right keywords or phrases, the desired images cannot be retrieved without visually examining the entire archive.

In view of the deficiencies of text-based approach, major research effort has been spent on CBIR over the past 15 years. CBIR generally involves the application of computer vision techniques to search for certain images in large image databases. "Content-based" means that the search makes use of the contents of the images themselves, rather than relying on manually annotated texts.

From a user perspective, CBIR should involve image semantics. An ideal CBIR system would perform semantic retrievals like "find pictures of dogs" or even "find pictures of George Bush." However, this type of open-ended query is very difficult for computers to perform because, for example, a dog's appearance can vary significantly between species. Current CBIR systems therefore generally make use of low-level features like texture, color, and shape. However, biologically-inspired vision research generally suggests two processes in visual analysis: bottom-up image-based analysis and top-down task-related analysis (Navalpakkam & Itti, 2006). Bottom-up analysis consists of memoryless stimulus-centric factors such as low-level image features. Top-down analysis uses prior domain knowledge to influence bottom-up analysis. An effective image retrieval system should therefore combine both the low-level features as well as the high level knowledge so that images can be classified automatically according to their context and semantic meaning.

EXISTING CBIR SYSTEMS AND STANDARDS

The best-known commercial CBIR system is the QBIC (Query by Image Content) system developed by IBM (Flickner et al., 1995). Image retrieval is achieved by any combination of color, texture or shape as well as by keyword. Image queries can be formulated by selection from a palette, specifying an example image, or sketching a desired shape on the screen. The other well-known commercial CBIR systems are Virage (Gupta & Jain, 1997) which is used by AltaVista for image searching, and Excalibur (Feder, 1996) which is adopted by Yahoo! for image searching. Photobook (Pentland, Picard, & Sclaroff, 1996) from MIT Media Lab is the representative research CBIR system. Like QBIC, images are represented by color, shape, texture and other appropriate features. However, Photobook computes information-preserving features, from which all essential aspects of the original image can be reconstructed.

In 1996, the Moving Picture Experts Group (MPEG)–a working group (JTC1/SC29/WG11) of the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC)—decided to start a standardization project called MPEG-7 (Manjunath, Salembier, & Sikora, 2002). The aim is to provide a quick and efficient identification and management of multimedia content so that audio-visual information can be easily searched.

The MPEG-7 specifies the descriptors, description schemes and a description definition language. The descriptors is a representation of features at different levels of abstraction, ranging from low-level visual features like shape, texture and color to high level semantic information such as abstract concept and genres. The descriptor defines the syntax and semantics of the feature representation.

The description schemes specify the structure and semantics of the relationships between its components such as descriptors. The scheme provides a solution to model and describe content in terms of structures and semantics. The description definition language allows the creation of new description schemes as well as descriptors. This allows extension and modification of existing description schemes.

The MPEG-7 standard has eight parts. Of these, Part 3 specifies a set of standardized low-level descriptors and description schemes for visual content which includes shape descriptor, color descriptor, texture descriptor and motion descriptor. Note that the MPEG-7 specifies the descriptors only, and their extraction are not specified as part of the MPEG-7 standard.

CBIR METHODOLOGY

A CBIR system has three key components: feature extraction, efficient indexing and user interface:

- **Feature extraction:** Image features include *primitive features* and *semantic features*. Examples of primitive features are *color*, *texture*, and *shape*. Primitive features are usually quantitative in nature and they can be extracted automatically from the image. Semantic features are qualitative in nature and they provide abstract representations of visual data at various levels of detail. Typically, semantic features are extracted manually. Once the features have been extracted, image retrieval becomes a task of measuring similarity between image features.
- **Efficient indexing:** To facilitate efficient query and search, the image indices needed to be organized into an efficient data structure. Because image features maybe interrelated, flexible data structures should be used in order to facilitate storage/retrieval. Structures such as *k-d-tree*, *R-tree*, *R*-tree*, *quad-tree*, and *grid file* are commonly used.
- **User interface:** In visual information systems, user interaction plays an important role. The user interface consists of a query processor and a browser to provide an interactive environment for querying and browsing the database. Common query mechanisms provided by the user interface are: *query by keyword*, *query by sketch*, *query by example*, *browsing by categories*, *feature selection*, and *retrieval refinement*.

In *query by example*, the user specifies a query image (either supplied by the user or chosen from a random set), and the system finds images similar to it based on various low-level criteria. In *query by sketch*, the user draws a rough sketch of the image he/she is looking for, for example, with blobs of color at different locations, and the system locates images whose layout matches the sketch. In either case, features are first extracted automatically from this query image to form a query image signature. A matching with all other images in the archive is performed by measuring the similarity between their signatures. It can be seen that the matching result is heavily influenced by the choice of features.

BOTTOM-UP CONTENT-BASED RETRIEVAL IN THE SPATIAL DOMAIN

Content-based retrieval makes use of low level image features computed from the image itself for matching. Commonly used

features are *color*, *texture* and *shape* which can be obtained directly from the spatial domain representation.

Color Signature

Color is one of the most widely used features as it is relatively robust to translation and rotation about the angle of view (Deng et al., 2001). One of the often used color features is color histogram. It partitions color distribution into discrete bins and can be used to show the overall color composition in an image. The MPEG-7 specifies six color descriptors: color space descriptor, dominant color descriptor, scalable color descriptor, group of frames or group of pictures descriptor, color structure descriptor and color layout descriptor. Among these six descriptors, two are related to the color histogram. For example, the dominant color information is defined as, $F = \{(\vec{c}_i, p_i, \upsilon_i), s\}$, $i = 1,2,...,N$, where N is the total number of dominant colors, $\vec{c}_i$ is a vector storing color component values, p_i is the fraction of pixels in the image corresponding to color $\vec{c}_i$, υ_i is the color variance representing the color variation in a cluster surrounding the color $\vec{c}_i$ and s is a number representing the overall spatial homogeneity of the dominant colors in the image. The color structure information is obtained from the localized color histogram using a small structuring window so that local spatial structure of the color can be characterized.

Texture Signature

Texture is one of the basic attributes of natural images. Commonly used methods for texture characterization are divided into three categories: statistical, model-based and filtering approaches. Statistical methods such as co-occurrence features describe tonal distribution in textures (Wouwer, Scheunders & Van Dyck, 1999). Model-based methods such as Markov random field (Cross & Jain, 1983) provide description in terms of spatial interaction while filtering approaches including wavelet, Gabor-filters and directional filter-bank (DFB) characterize textures in frequency domain (Chang & Kuo, 1993; Manjunath & Ma, 1996). It has been shown that directional together with scale information is important for texture perception. As texture patterns can be analyzed at various orientations with multiple scales using Gabor-filters, good texture descriptions can be obtained. However, Gabor-filter involves nonseparable transform which is computationally expensive. Recently, Contourlet transform and multiscale directional filter bank have been proposed to solve this problem by combining the DFB with Laplacian pyramid (LP) (Cheng, Law, & Siu, 2007). Although the LP is somehow redundant, the combined approach is still computationally efficient while providing a high angular resolution. As a result, efficient texture descriptors can be obtained.

There are three texture descriptors in MPEG-7: a homogeneous texture descriptor, a texture browsing descriptor and an edge histogram descriptor. The homogeneous texture descriptor adopts Gabor-like filtering for the texture description. The texture browsing descriptor provides a perceptual texture characterization in terms of regularity, coarseness and directionality of the texture pattern. The edge histogram descriptor is obtained by analyzing spatial distribution of edges in an image.

Shape Signature

Shape is one of the key visual features used by human for distinguishing visual data. Compare with color and texture, shape is easier for users to describe in a query, either by example or by sketch. However, because shapes of natural objects in a 2D image can be obtained from different views of the same object, shapes can be rotated, scaled, or skewed. Hence, an effective shape representation should be rotation, translation and scaling invariant, as well as invariant to affine transform to address the different views of objects.

Two of the common approaches for shape representation are boundary-based and region-based approaches. The boundary-based approach works on the edges/outlines of the image while the region-based approach considers the entire regions. In fact, the MPEG-7 standard defines three shape descriptors: region-based shape descriptor, contour-based shape descriptor and 3D shape spectrum descriptor. The region-based descriptor is based on a complex 2D angular radial transformation which belongs to a class of shape analysis techniques using Zernike moments. The contour-based descriptor is based on extracting features such as contour circularity and eccentricity from the curvature scale-space contour representation. The 3D shape descriptor is based on a polygonal 3D meshes object representation.

Bottom-Up Content-Based Retrieval in the Compressed Domain

Features used in retrieval are often extracted from the spatial domain. This is in contrast to the fact that images are usually compressed using JPEG or JPEG 2000 to reduce their size for storage and transmission. Retrieving these kinds of compressed images then requires reconversion to the uncompressed spatial domain for feature extraction. This approach requires many decompression operations, especially for large image archives. To avoid some of these operations, it was proposed that feature extraction be done directly in the transformed domains.

JPEG employs the discrete cosine transform (DCT) for image compression (Pennebaker & Mitchell, 1993). Low-level features such as color, shape and texture have been proposed to be extracted directly by analyzing DCT coefficients. For example, DCT coefficients can be reorganized into a tree structure capturing the spatial-spectral characteristics for retrieval (Climer & Bhatia, 2002; Ngo, Pong, & Chin, 2001). JPEG 2000 is a new compression standard which compresses images using wavelets (Taubman & Marcellin, 2002). As wavelets provide a multiple resolution view of an image, several indexing techniques have been proposed to extract wavelet coefficients for coarse-to-fine image retrieval (Liang & Kuo, 1999; Xiong & Huang, 2002).

One of the major concerns in the compressed domain feature extraction is that they are domain specific. Different compression techniques result in different transformed coefficients. This implies that features that can be extracted in the compressed domain depend greatly on the compression scheme used. As a result, the retrieval system can only be used to retrieve images from a particular compression format. To extract features in the compressed domain irrespective of the compression format, a common framework called the subband filtering model has been proposed (Au, Law, & Siu, 2007). Using this subband model, the block-based DCT coefficients are concatenated to form structures that are similar to the wavelet subbands. This would allow similar features to be extracted for retrieval purposes. It has been proved that similar features can always be extracted in the JPEG and JPEG 2000 domains for retrieval, irrespective of the values of the compression ratio.

Top-Down-Based Retrieval Using Prior Knowledge

Although the visual mechanism is still not well understood, biological visual system research generally suggest two processes involved in visual analysis: bottom-up image-based analysis and top-down task-related analysis. Bottom-up analysis consists of extraction of low-level features described above. Top-down analysis refers to those high-level concepts that cannot be extracted directly from the image, but instead from the semantics of objects and image scenes as perceived by human beings. These conceptual aspects are subjective and are more closely related to users' preferences.

Low-level image features fail to represent high-level semantic image features because they are only the basic components to build cognitive features. A human always focus on the interesting part of an image based on what he or she is trying to look for. The cognitive features of an image play the most important role in the semantic understanding of images.

Whereas low level features are usually context-free, high level semantic features are context-rich. The relationship (spatial or conceptual) between image objects is a strong constraint for semantic image description. In Hare et al. (2006), a semantic space is used to describe the spatial relationships between image objects and hence the scene content. User's domain knowledge can be used to constrain such a semantic space, for example, a mountain scene would consists of blue sky on the top portion of the image, mountains in the middle portion of the image, and grassland or forest in the foreground.

To incorporate high level knowledge incrementally, *relevance feedback* was proposed to interactively refine retrieval results (Rui, Huang, Ortega, & Mehrotra, 1998). In particular, the user indicates to the retrieval system whether the particular retrieved result is "relevant," "not relevant" or "neutral." With the use of some learning algorithms such as Support Vector Machines, a set of possibly better retrieved results can be obtained after a few rounds of this feedback mechanism.

FUTURE TRENDS

A major challenge in CBIR research is to develop a semantic-sensitive retrieval system. Semantic similarity is an important image matching criterion in human and is related to how we interpret image content. An effective image retrieval system should combine both low-level features as well as high level semantic knowledge so that images can be classified according to their context and meaning. However, very little research has been done on high-level semantic features. A possible reason is that high level knowledge are difficult to define and formulate because they are highly subjective and are closely related to viewers' expectations of scene context. There are some attempts to segment an image into different regions using homogenous low level features and then combine them together with context information to form a model for image indexing and retrieval. However, meaningful segmentation is often difficult to obtain in practice.

The major problem in semantic-sensitive CBIR is to define and extract the semantic concept behind the image. The concept can be related to the image categories such as buildings and gardens. It can also be tailored to certain application domains, such as detecting human faces and intruders. Once the concept is extracted and associated with certain low-level features, content-based image retrieval becomes concept matching, which is semantically more meaningful than low-level features matching. Performing concept matching instead of low-level features matching could also speed up the retrieval process.

Machine learning techniques can be used to learn the high level semantic knowledge automatically. Relevance feedback has been proposed to improve the retrieval results. The major challenge here is to develop an effective and efficient algorithm for automatic concept learning and representation. Due to the fuzzy nature of semantic concepts,

probabilistic graphical models such as Bayesian network or relevance network could find applications here.

CONCLUSION

With the rapid growth of Internet and multimedia systems, the use of visual information has increased enormously such that image-based indexing and retrieval techniques have become important. In this article, an overview of CBIR is given. Current CBIR systems mainly extract low-level features such as color, texture, and shape from an image for image classification and retrieval. However, it is well known that these low-level features cannot capture image semantics. In contrast, human performs image searching by relying heavily on the semantic concepts behind the image. In order to close this semantic gap, much more research effort is needed to find ways to define and incorporate high level semantic knowledge onto the CBIR system. We have discussed two major challenges in this area, including (1) semantic concept formulation and extraction from images and (2) the development of effective machine learning algorithms for concept learning and representation.

REFERENCES

Au, K.M., Law, N.F., & Siu, W.C. (2007). Unified feature analysis in JPEG and JPEG2000 compressed domains. *Pattern Recognition, 40*(7), 2049-2062.

Chang, S.K., & Hsu, A. (1992). Image information systems: Where do we go from here? *IEEE Transactions on Knowledge and Data Engineering, 5*(5), 431-442.

Chang, T., & Kuo, C.C. J. (1993). Texture analysis and classification with tree-structured wavelet transform. *IEEE Transactions on Image Processing, 2*(4), 429-441.

Cheng, K.O., Law, N.F., & Siu, W.C. (2007). Multiscale directional filter bank with applications to structured and random texture classification. *Pattern Recognition, 40*(4), 1182-1194.

Climer, S., & Bhatia, S.K. (2002). Image database indexing using JPEG coefficients. *Pattern Recognition, 35*(11), 2479-2488.

Cross, G.R., & Jain, A.K. (1983). Markov random field texture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*(1), 25-39.

Deng, Y., Manjunath, B.S., Kenney, C., Moore, M.S., & Shin, H. (2001). An efficient color representation for image retrieval. *IEEE Transactions on Image Processing, 10*(1), 140-147.

Feder, J. (1996). Towards image content-based retrieval for the World Wide Web. *Advanced Imaging, 11*(1), 26-29.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., et al. (1995). Query by image and video content: The QBIC system. *IEEE Transactions on Computer, 28*(9), 23-32.

Gupta, A., & Jain, R. (1997). Visual information retrieval. *Communications of ACM, 40*(5), 71-79.

Hare, J.S., Sinclair, P.A.S., Lewis, P.H., Martinez, K., Enser, P.G.B., & Sandom, C. J. (2006). Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches. In *Proceedings of Mastering the Gap: From Information Extraction to Semantic Representation, 3rd European Semantic Web Conference*.

Liang, K.C., & Kuo, C.C.J. (1999). WaveGuide: A joint wavelet-based image representation and description system. *IEEE Transactions on Image Processing, 8*(11), 1619-1629.

Manjunath, B.S., & Ma, W.Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 837-842.

Manjunath, B.S., Salembier, P., & Sikora, T. (2002). *Introduction to MPEG-7: Multimedia content description language*. John Wiley & Sons.

Navalpakkam, V., & Itti, L. (2006). An integrated model of top-down and bottom-up attention for optimal object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 2049-2056).

Ngo, C.W., Pong T.C., & Chin, R.T. (2001). Exploiting image indexing techniques in DCT domain. *Pattern Recognition, 34*(9), 1841-1851.

Pennebaker, W.B., & Mitchell, J.L. (1993). *JPEG: Still image data compression standard*. Van Nostrand Reinhold.

Pentland, A. Picard, R.W., & Sclaroff, S. (1996). Photobook: Content-based manipulation of image databases. *International Journal Computer Vision, 18*, 223-254.

Rui, Y., Huang, T.S., Ortega, M., & Mehrotra, S. (1998). Relevance feedback: A powerful tool in interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology, 8*(5), 644-655.

Taubman, D.S., & Marcellin, M.W. (2002). *JPEG2000: Image compression fundamentals, standards and practice*. Kluwer Academic Publishers.

Wouwer, G.V.D., Scheunders, P., & Van Dyck, D. (1999). Statistical texture characterization from discrete wavelet

representations. *IEEE Transactions on Image Processing*, *8*(4), 592-598.

Xiong, Z., & Huang, T.S. (2002). Subband-based, memory-efficient JPEG2000 images indexing in compressed-domain. In *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, (pp. 290-294).

KEY TERMS

Bottom-Up Image Analysis: This refers to the use of low-level features, such as high luminance/color contrast or unique orientation from its surrounding, to identify certain objects in an image.

Compressed Domain Feature Analysis: This refers to the process of image signature extraction performed in the transform domain. Image features are extracted by analyzing the transform coefficients of the image without incurring a full decompression.

Content-Based Image Retrieval: This refers to an image retrieval scheme which searches and retrieves images by matching information that is extracted from the images themselves. The information can be color, texture, shape and high level features representing image semantics and structure.

Feature Descriptors: A set of features that is used for image annotation and indexing. The features can be keywords, low-level features including color, texture, shape, and high level features describing image semantics and structure.

High Level Semantics: This refers to the image context as perceived by humans. It is generally subjective in nature and greatly depends on user's preferences.

Image Retrieval System: A computer system for users to search images stored in a database.

Image Signature: This is the same as feature descriptors used for image annotation and indexing.

Keyword-Based Image Retrieval: This refers to an image retrieval scheme which searches and retrieves images by using metadata such as keywords. In this scheme, all images are annotated with certain keywords. Searching is then performed by matching these keywords.

Relevance Feedback: This provides an interactive way for humans to refine the retrieval results. Users can indicate to the image retrieval system whether the retrieved results are "relevant," "irrelevant" or "neutral." Retrieval results are then refined iteratively.

Spatial Domain Feature Analysis: This refers to the process of image signature extraction performed in the spatial domain. Image features are extracted by analyzing the spatial domain image representation.

Top-Down Image Analysis: This refers to the use of high level semantics, such as viewer's expectations of objects and image context, to analyze and annotate an image.

Content-Based Retrieval Concept

Yung-Kuan Chan
National Chung Hsing University, Taiwan, R.O.C.

Chin-Chen Chang
National Chung Cheng University, Taiwan, R.O.C.

INTRODUCTION

Because of the demand for efficient management in images, much attention has been paid to image retrieval over the past few years. The text-based image retrieval system is commonly used in traditional search engines (Ratha et al., 1996), where a query is represented by keywords that are usually identified and classified by human beings. Since people have different understandings on a particular image, the consistency is difficult to maintain. When the database is larger, it is arduous to describe and classify the images because most images are complicated and have many different objects. There has been a trend towards developing the content-based retrieval system, which tries to retrieve images directly and automatically based on their visual contents.

A similar image retrieval system extracts the content of the query example $\boldsymbol{q}$ and compares it with that of each database image during querying. The answer to this query may be one or more images that are the most similar ones to $\boldsymbol{q}$. Similarity retrieval can work effectively when the user fails to express queries in a precise way. In this case, it is no longer necessary to retrieve an image extremely similar to the query example. Hence, similarity retrieval has more practical applications than an exact match does.

Content-Based Image Retrieval Systems

In a typical content-based image retrieval system, the query pattern is queried by an example in which a sample image or sketch is provided. The system then extracts appropriate visual features that can describe the image, and matches these features against the features of the images stored in the database. This type of query is easily expressed and formulated, since the user does not need to be familiar with the syntax of any special purpose image query language. The main advantage is that the retrieval process can be implemented automatically (Chen, 2001). The scope of this article is circumscribed to image abstraction and retrieval based on image content.

Human beings have a unique ability that can easily recognize the complex features in an image by utilizing the attributes of shape, texture, color, and spatial information. Many researchers analyze the color, texture, shape of an object, and spatial attributes of images, and use them as the features of the images. Therefore, one of the most important challenges in building an image retrieval system is the choice and representation of the visual attributes. A brief overview of the commonly used visual attributes shape, texture, color, and spatial relationship will be illustrated as follows.

Commonly Used Image Features in Content-Based Image Retrieval Systems

Shape characterizes the contour of an object that identifies the object in a meaningful form (Gevers & Smeulders, 2000; Zhang & Lu, 2002). Traditionally, shapes are described through a set of features such as area, axis-orientation, certain characteristic points, and so forth. These systems retrieve a subset of images that satisfy certain shape constraints. In the shape retrieval, the degree of similarity between two images is considered as the distance between the corresponding points.

Color attribute may simplify the object's identification and extraction in the image retrieval (Galdino & Borges, 2000; Gevers & Smeulders, 2000). Color may provide multiple measurements at a single pixel of the image, and often enable the classification to be done without complex spatial decision-making. Any resulting difference between colors is then evaluated as a distance between the corresponding color points. The color-based retrieval system measures the similarity of the two images with their distance in color space.

Texture attribute depicts the surface of an image object (Yao & Chen, 2002; Zhang & Tan, 2003). Intuitively, the term refers to properties such as smoothness, coarseness, and regularity of an image object. Generally, the structural homogeneity does not come from the presence of a single color or intensity, but it requires the interaction of various intensities within a region.

Retrieval by spatial constraints facilitates a class of queries based on the 2-D arrangement of objects in an image (Chang Erland & Li, 1989; Chang & Li, 1988; Chang, Shi & Yan, 1987; Lee & Hsu, 1992). The query is composed by placing sketches, symbols or icons on a plane where every symbol or icon is predefined for one type of objects in an image. The relationships between the objects can be broadly

classified as either directional (also referred as projective) (Chang & Li, 1988; Chang, Shi & Yan, 1987) or topological (Lee & Hsu, 1992). Directional relationship is based on the relative location and the metric distance between two image objects. Topological relationships are based on set-theoretical concepts like union, intersection, disjunction and so forth. Spatial information is a higher-level attribute, which is increasingly more specific. For example, facial features are frequently presented in terms of spatial information (Sadeghi, Kittler & Messer, 2001).

Briefly, color attribute depicts the visual appearance of an image, characterized by the luminance and chrominance histograms of the image. Texture attribute refers to three components: bi-dimensional periodicity, mono-dimensional orientation, and complexity obtained through world decomposition. Shape attribute sketches the geometrical properties of objects in images. Spatial attribute represents the relative position relationships between objects of an image.

TYPICAL IMAGE RETRIEVAL SYSTEMS

This section briefly overviews the image retrieval systems based on the most commonly used image features: color, shape, texture, and spatial content.

The Color-Based Image Retrieval Systems

Generally, the color-based image retrieval system does not find the images whose colors are exactly matched, but images with similar pixel color information. This approach has been proven to be very successful in retrieving images since concepts of the color-based similarity measure is simple, and the convention algorithms are very easy to implement. Besides, this feature can resist noise and rotation variants in images.

However, this feature can only be used to take the global characteristics into account rather than the local one in an image, such as the color difference between neighboring objects in an image. For example, if a landscape image with blue sky on the top and green countryside at the bottom is employed as a query example, the system that retrieves the images with similar structures based on these global features often gives very unsatisfactory results. In addition, the color-based image retrieval system often fails to retrieve the images that are taken from the same scene in which the query example is also taken from under different time or conditions, for example, the images of a countryside taken at dusk or dawn under a clear or a cloudy sky. In another scenario, the same scene may be imaged by different devices. Using one image taken by one device as the query example may fail to find the same scene taken by other devices.

The Shape-Based Image Retrieval Systems

A shape-based image retrieval system is used to search for the images containing the objects, which are similar to the objects specified by a query. Since an object can be formed by a set of shapes in most cases (e.g., a car can be made of some little rectangles and circles), most similar objects have a high correlation in their set of shapes (Gevers & Smeulders, 2000; Zhang & Lu, 2002). The shape-based image retrieval system extracts the shapes of objects from images by segmentation, and classifies the shapes, where each shape has its own representation and variants to scaling, rotation, and transition.

Some criteria on shape representation and similarity measure for a well performing content-based image retrieval system should be achieved. Firstly, the representation of a shape should be invariant to scale, translation, and rotation. Secondly, the similarity measure between shape representations should conform to human perception; that is, perceptually similar shapes should have highly similar measures. Thirdly, the shape representation should be compact and easy to derive, and the calculation of similarity measure should be efficient.

However, how to locate and how to recognize objects from images is a real challenge. One of the obstacles is how to separate the objects from the background. Difficulties come from discrimination, occlusions, poor contrast, viewing conditions, noise, complicated objects, complicated backgrounds, and so forth. Moreover, the shape-based image retrieval system can only deal with the images that have simple object shapes. For complex object shapes, the region-based method has to build a binary sequence by using smaller grid cells, so that results that are more accurate can be obtained; nevertheless, the storage of indices and retrieval time may increase tremendously.

The Texture-Based Image Retrieval Systems

Literally, texture relates to the arrangement of the basic constituents of a material. In digital images, texture describes the spatial interrelationships of the image pixels. Texture similarity can often be useful in distinguishing the areas of objects in images with similar color, such as sky and sea as well as leaves and grass. Texture queries can be formulated in the manner that is similar to the color queries by selecting an example of desired textures from a palette, or by supplying an example query image. The system then returns the images which are most similar to the query example in texture measures.

Making texture analysis is a real challenge. One way to perform content-based image retrieval using texture as

the cue is by segmenting an image into a number of different texture regions and then performing a texture analysis algorithm on each texture segment. However, segmentation can sometimes be problematic for image retrieval. In addition, texture is quite difficult to describe and subject to the difference of human perception. No satisfactory quantitative definition of texture exists at this time.

The Spatial-Based Image Retrieval Systems

There are two kinds of spatial-based image retrieval systems: retrieval by spatial relationships (RSRs) and spatial access methods (SAMs). The RSR image retrieval system is to retrieve the images from a database that are similar to the query sample based on relative position relationships between the objects in the images. Hence, a physical image can be regarded as a symbolic image, each object of which is attached with a symbolic name. The centroid coordinates of the object with reference to the image frame are extracted as well. By searching for the logical images, the corresponding physical images can then be retrieved and displayed. Therefore, image retrieval can be simplified to the search of symbolic images.

Chang, Shi, and Yan (1987) used a 2D string representation to describe a symbolic image. Objects and their spatial relationships in a symbolic image can be characterized by a 2D string. An image query can be specified as a 2D string too. Consequently, the problem of image retrieval then turns out to be the matching of a 2D string. Subsequently, a great number of other image representations popped out that were derived from a 2D string, such as 2D G-string (Chang, Erland & Li, 1989), 2D B-string (Lee, Yang & Chen, 1992), 2D C-string (Lee & Hsu, 1992), and so forth. These representations adopt the description of orthogonal projection to delineate the spatial relationships between objects.

The SAM image retrieval systems are to manage large collections of points (or rectangles or other geometric objects) in the main memory or on the disk so that range queries can be efficiently answered. A range query specifies a region in the address space, requesting all the data objects that intersect it. They divide the whole space into several disjoint sub-regions, each with no more than P points (a point may represent a rectangle). P is usually the capacity of a disk page. Inserting a new point may result in further splitting a region. The split methods can be classified according to the attributes of the split (Gottschalk, Turney & Mudge, 1987).

Color attribute is most intuitive and straightforward for the user. Texture analysis systems are often developed to perform filtering in the transform domain in order to obtain feature images. The use of global color or texture features for the retrieval of images tends to be misleading, especially in homogeneous image collections. Though shape and texture are the essential visual attributes to derive potentially useful semantic information, there exists less understanding of the benefits to implement these attributes as compared to color, for efficient image retrieval. This approach apparently focuses on global frequency content of an image; however, many applications require the analysis to be localized in the spatial domain.

FUTURE TRENDS

Many visual attributes have been explored, such as color, shape, texture, and spatial features. For each feature, there exist multiple representations that model the human perception of the feature from different perspectives. There is a demand for developing an image content description to organize the features. The features should not only be just associated with the images, but also be invoked at the right place and the right time, whenever they are needed to assist retrieval.

Human beings tend to apply high-level concepts in their daily lives. However, most features, the current computer vision techniques automatically extracting from images, are low-level. To narrow down this semantic gap, it is necessary to link the low-level features to high-level concepts. On the high-level concept, it should allow the user to easily provide his or her evaluation of the current retrieval results to the computer. It is likely for different people to give an identical name; therefore, generating the representation by automatically extracting the objects from an original image is very difficult. Therefore, the spatial relationships between objects cannot be extracted automatically without human interaction with the current techniques of image understanding and recognition. More recent research emphasis is given to "interactive systems" and "human in the loop".

Due to the perception subjectivity of image content, it is difficult to define a good criterion for measuring the similarity among images. That is, the subjectivity of image perception prevents us from defining objective evaluation criteria. Hence, it is urgent to find an appropriate way of evaluating the system performance guiding the research effort in the correct direction.

Establishing a well-balanced large-scale test bed is an important task too. A good test bed must be huge in scale for testing the scalability (for multidimensional indexing), and be balanced in image content for testing image feature effectiveness and overall system performance.

Human beings are the ultimate end users of the image retrieval system. This topic has attracted increasing attention in recent years, aiming at exploring how humans perceive image content and how one can integrate such a "human model" into the image retrieval systems. Recently, more studies of human perception focus on the psychophysical aspects of human perception.

CONCLUSION

A successful image retrieval system requires the seamless integration of the efforts of multiple research communities. To achieve a fast retrieval speed and make the retrieval system truly scalable to large-size image collections, an effective multidimensional indexing module is an indispensable part of the whole system. The interface collects the information from the users and displays back the retrieval results to the users in a meaningful way. To communicate with the user in a friendly manner, the query interface should be graphics-based.

In the iconic image representation, an icon is used to represent an object in an image. The iconic image representation has two advantages. First, once the images in the database are analyzed, it is not necessary to analyze the image again in a query processing. Secondly, since the size of the symbolic image is much smaller than that of the original image, this representation can be well suited to be distributed to database environments where a large number of image transmissions between distant nodes are required.

Typically, color, shape and texture attributes provide a global description of images, but they fail to consider the meaning of portrayed objects and the semantics of scenes. The descriptions of objects and the relative positions among objects provide a spatial configuration and a logical representation of images. Combining approaches for content-based image retrieval systems could be considered as complementary.

REFERENCES

Chang, S.K., & Li, Y. (1988). Representation of multiresolution aymbolic and binary images using 2D H-strings. *Proceedings of IEEE Workshop Languages for Automation,* 190-195.

Chang, S.K., Erland, J., & Li, Y. (1989). The design of pictorial database upon the theory of symbolic projections. *Proceedings of the 1st Symposium on the Design and Implementation of Large Spatial Databases,* 303-323.

Chang, S.K., Shi, Q.Y., & Yan, C.W. (1987). Iconic indexing by 2D strings. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 9*(3), 413-328.

Chen, H.L. (2001). An analysis of image retrieval tasks in the field of art history. Information Processing and Management, *37*(5), 701-720.

Galdino, L.L., & Borges, D.L. (2000). A visual attention model for tracking regions based on color correlograms. *Proceedings of 8th Brazilian Symposium on Computer Graphics and Image Processing,* 36-43.

Gevers, T., & Smeulders, A.W.M. (2000). PicToSeek: Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing, 9*(1), 102-119.

Gottschalk, P., Turney, J., & Mudge, T. (1987). Two-dimensional partially visible object recognition using efficient multidimensional range queries. *Proceedings of IEEE International Conference on Robotics and Automation, 4,* 1582 -1589.

Lee, S.Y., & Hsu, F.J. (1992). Spatial reasoning and similarity retrieval of images using 2D C-string knowledge representation. *Pattern Recognition, 25*(3), 305-318.

Lee, Y., Yang, M.C., & Chen, J.W. (1992). 2D B-string knowledge representation and image retrieval for image database. *Proceedings of 2nd International Computer Science Conference Data and Knowledge Engineering: Theory and Applications,* 13-16.

Owei, V., & Navathe, S.B. (2001). Enriching the conceptual basis for query formulation through relationship semantics in databases. *Information Systems, 26*(6), 445-475.

Ratha, N.K., Karu, K., Chen, S.Y., & Jain, A.K. (1996). A real-time matching system for large fingerprint databases. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 799 -813.

Sadeghi, M., Kittler, J., & Messer, K. (2001). Spatial clustering of pixels in the mouth area of face images. *Proceedings of 11th International Conference on Image Analysis and Processing,* 36-41.

Yao, C.H., & Chen, S.Y. (2002). Retrieval of translated, rotated and scaled color textures. Pattern Recognition, *36*(4), 913-929.

Zhang, D., & Lu, G. (2002). Shape-based image retrieval using generic Fourier descriptor. *Signal Processing: Image Communication, 17*(10), 825-848.

Zhang, J., & Tan, T. (2003). Affine invariant classification and retrieval of texture images. *Pattern Recognition, 36*(3), 657-664.

KEY TERMS

Color Feature: Analyzing the color distribution of pixels in an image.

Geometric Hashing: The technique identifying an object in the scene, together with its position and orientation.

Query by Example: The image retrieval system where a sample image or sketch can be provided as a query.

Shape Feature: Characterizing the contour of an object that identifies the object in a meaningful form.

Spatial Feature: Symbolizing the arrangement of objects within the image.

Symbolic Image: Consisting of a set of objects; each object stands for an entity in a real image.

Texture Feature: Depicting the surface of an image object.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 564-568, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

A Content-Sensitive Approach to Search in Shared File Storages

Gábor Richly
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The article presents a novel approach to search in shared audio file storages such as P2P-based systems. The proposed method enables the recognition of specific patterns in the audio contents, in such a way it extends the searching possibility from the description-based model to the content-based model. The targeted shared file storages seam to change contents rather unexpectedly. This volatile nature led our development to use real-time capable methods for the search process.

The importance of the real-time **pattern recognition** algorithms that are used on audio data for content-sensitive searching in stream media has been growing over a decade (Liu, Wang, & Chen, 1998). The main problem of many algorithms is the optimal selection of the reference patterns (*soundprints* in our approach) used in the recognition procedure. This proposed method is based on distance maximization and is able to choose the pattern that later will be used as reference by the pattern recognition algorithms quickly (Richly, Kozma, Kovács & Hosszú, 2001).

The presented method called ***EMESE (Experimental MEdia-Stream rEcognizer)*** is an important part of a light-weight content-searching method, which is suitable for the investigation of the network-wide shared file storages. This method was initially applied for real-time monitoring of the occurrence of known sound materials in broadcast audio. The experimental measurement data showed in the article demonstrate the efficiency of the procedure that was the reason for using it in shared audio database environment.

BACKGROUND

From the development of the Napster (Parker, 2004), the Internet-based communication is developing toward the **application level network**s (**ALN**). On the more and more powerful hosts, various collaborative applications run and create virtual (logical) connections with each others (Hosszú, 2005). They establish virtual **overlay**, and oppositely to the older **client/server model** they use the **peer-to-peer** (**P2P**) communication. The majority of such systems deal with file sharing (Adar, Huberman, 2000), that is why their important task is to search in large, distributed shared file storages.

A large amount of effort is dedicated for improving their searching (Yang & Garcia-Molina, 2002) and downloading capability (Cohen, 2003; Qiu & Srikant, 2004), however, the searching is quite traditional, it is based on the descriptive metadata of the media contents, as the file name, content description, and so forth. Such method has an inherent limitation, since the real content remains covered and even in case of the mistake of the file description this search can fail. Oppositely to the widely used description-based seeking, the content-based searching has been the topic of the research, only. Its main reason is that the content and its coded representation have a huge variety that is why a comprehensive method has not developed yet.

The novel system, EMESE, is dedicated for solving a special problem, where a small but significant pattern should be found in a large voice stream or bulk voice data file in order to identify known sections of audio. The developed method is light-weight, meaning that its design goals were the fast operation and the relatively small computing power. In order to reach these goals, the length of the pattern to be recognized should be very limited, and the total score is not required.

This article deals mainly with the heart of the EMESE system, the pattern recognition algorithm, especially with the creation of the reference pattern, called *reference selection*.

THE PROBLEM OF THE PATTERN RECOGNITION

In the field of sound recognition there are many different methods and applications for specific tasks (Coen, 1995; Kondoz, 1994).

The demand for working efficiently with streaming media on the Internet increases rapidly. These audio streams may contain artificial sound effects besides the mix of music and human speech. These effects furthermore may contain signal fragments that are not audible by the ear. As a consequence, processing of this kind of **audio signal** is rather different from the already developed methods, as for example, the short-term predictability of the signal is not applicable.

The representation of digital audio signal as individual sample values lacks any semantic structure to help automatic identification. For this reason, the audio signal is transformed into several different orthogonal or quasi-orthogonal bases that enable detecting certain properties.

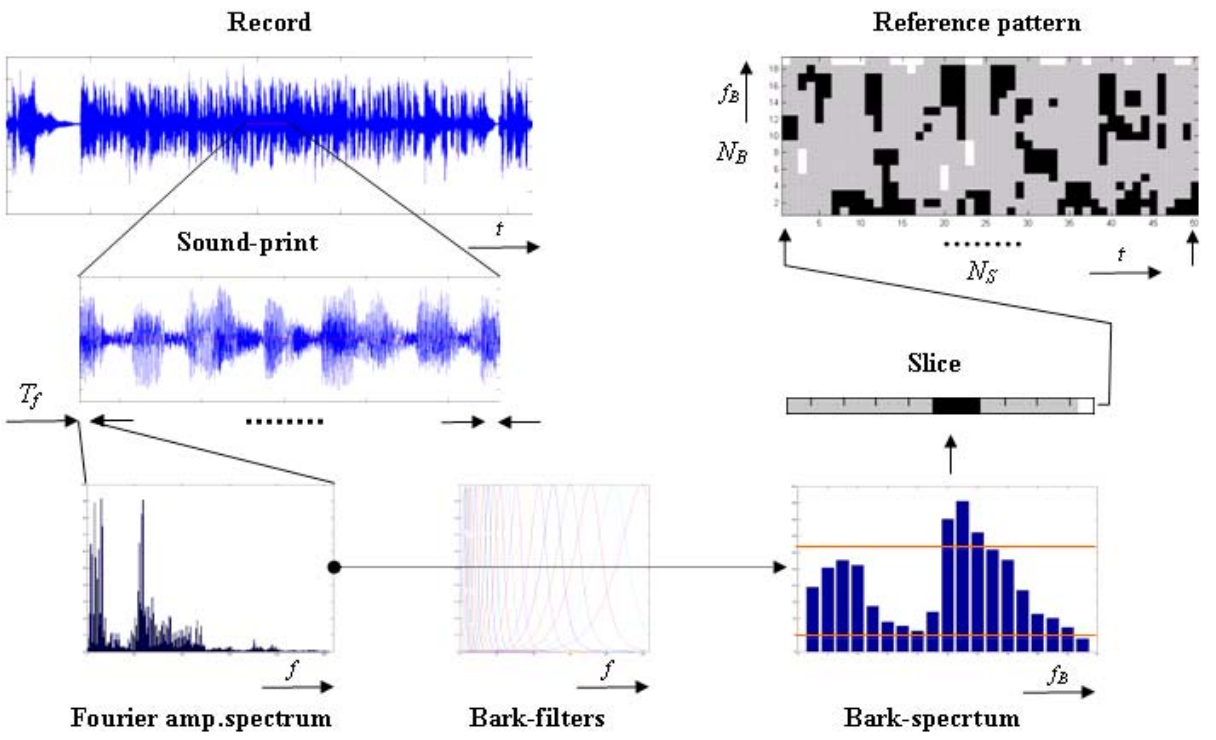
Already, there are solutions for classifying the type of broadcast on radio or television using the audio signal. The solution in Akihito, Hamada, and Tonomura (1998) makes basically a speech/music decision by examining the spectrum for harmonic content, and the temporal behavior of the spectral-peek distribution. Although it was applied successfully to that decision problem, it cannot be used for generic recognition purposes. The paper (Liu et al., 1998) also describes a scheme classifying method where the extracted features are based on the short-time spectral distribution represented by a bandwidth and a central frequency value. Several other features, for example, the volume distribution and the pitch contour along the sound clip, are also calculated. The main difficulty with this method is its high computation-time demand, so real-time monitoring is hardly possible, when taking the great number of references to be monitored into account.

A similar monitoring problem was introduced in Lourens (1990) and the used feature, a section of the energy envelope of the record signal (*reference*) was correlated with the input (*test*) signal. The demand on real-time execution drove the development of the recognition scheme introduced in Richly, Varga, Hosszú, and Kovács (2000), and Richly, Kozma, Kovács, and Hosszú (2001) that is capable of recognizing a pattern of transformed audio signal in an input stream, even in the presence of level-limited noise. This algorithm first selects a short segment of the signal from each record in the set of records to be monitored.

THE SOUND IDENTIFICATION IN THE EMESE

The reference selection algorithm needs a well understanding of the recognition method. The audio signal, sampled at f_s=16kHz is transformed into a spectral description because

Figure 1. The sound representation in the recognition system

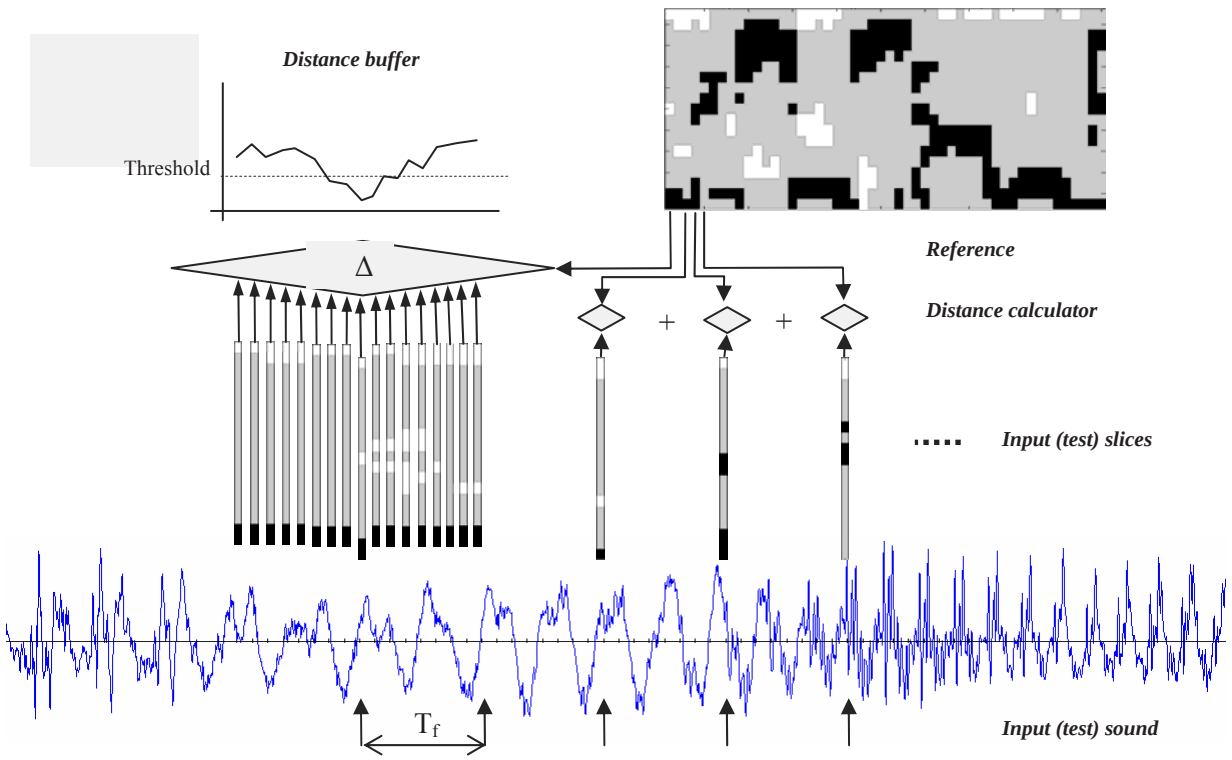


time-domain representation of sound is more sensitive to even light, non-audible distortions. A sound interval is transformed into a block of data, where the columns are vectors computed from a short, atomic section of the sound, where its spectral content can be assumed static. This section of time-domain data is called a *frame* ($N=256$ samples, $T_f=16ms$). First, the absolute value of the complex Fourier spectrum, the amplitude spectrum is computed from the frame. Then, amplitude values of neighboring frequencies are averaged to project the spectrum onto the Bark-scale, a non linear frequency scale. The reason for this is to speed up the later comparison stage by reducing the amount of relevant data and to include a well established emphasizing tool used in audio processing, the perceptual modeling of the human auditory system. As a result, we get a vector with $N_B=20$ values computed from the samples of a frame. In the next step, the vector is normalized and quantized. Two levels are determined in each transformed frame. The levels are the 10% and 70% of the peak value of the amplitude spectrum. We name the transformed frame a *slice* to indicate the applied column-vise recognition method of a reference. In every reference, there are N_S=50 slices of non-overlapping consecutive frames and the audio section, from which the reference was made, is called the soundprint of that specific record ($T_{SP} = T_f * N_S = 800ms$).

The scheme of the recognition algorithm is to grow the already identified parts of the reference patterns continuously, according to the input. This means that the algorithm takes a frame from the input signal, executes the previously described transformation series, and compares the resulting slice to the actual one of every reference. The actual slice is the first one in every reference initially and if it is decided to be similar to the slice computed from the input stream (a slice-hit occurs), the next, non-overlapping input slice will be compared to the next slice of that reference. If an input slice is decided to be non-similar, the actual slice of that reference is reset to the first one. The similarity is evaluated by calculating the weighted Manhattan-distance of the two slices. Manhattan-distance is the sum of the absolute element-vise differences in the slice vectors. This distance metric was chosen at the initial state of the project to avoid multiplication, since the system was developed for low-resource microcontroller and ASIC realizations. Weighting is introduced to emphasize the relevant differences in the used strong quantized environment.

For achieving more accurate alignment between the test and reference signals, the initial slice-hit in a reference is evaluated using a distance buffer, as demonstrated on Figure 2. In this circular-memory, the distances of that first reference slice to overlapping test slices are stored and the middle of the buffer is examined whether it contains the lowest value in the buffer. In case it does, and it also satisfies the threshold criteria, the identification of the reference proceeds to the next reference slice. This method intends to align the iden-

Figure 2. The synchronization mechanism



tification process to the "distance-pit" described in the next section (Richly, Kozma, Hosszú, & Kovács, 2001).

After successfully identifying the last slice of a reference, we successfully identified that record in the monitored input.

THE METHOD OF SELECTING THE REFERENCE PATTERNS

The selection algorithm uses the previously described weighted Manhattan-distance for measuring the similarity of the audio segments. In the vicinity of the reference's beginning, there have to be frames that vary a lot in the sense of the applied distance metric. This has to be fulfilled because the pattern recognition algorithm cannot synchronize to the given reference otherwise, since the record may appear anywhere in the monitored signal. This way a robust

Figure 3. The "distance-pit" around the reference position

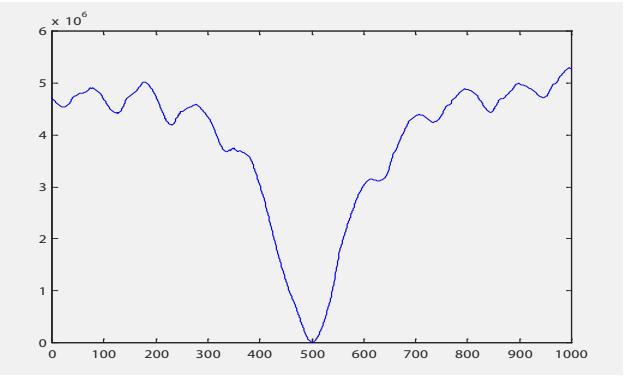
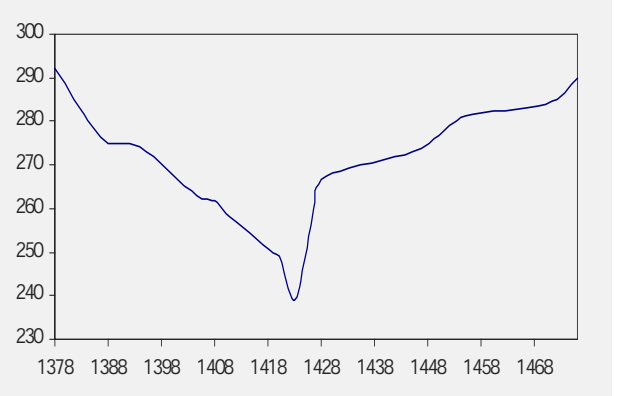


Figure 4. The width of the pits around the sound-print candidate



synchronization can be realized that is also successful in the presence of noise.

If we take a long sound-print (reference candidate) from a record to be monitored and calculate the distance of this section all along the record, then it can be observed that the distance function has a local minimum, a *pit* around the candidate's position (Richly, Kozma, Hosszú, & Kovács, 2001). This is demonstrated in Figure 3, where the x-axis shows which frame of the record is compared with the selected candidate, while the y-axis shows the Manhattan-distance values.

To achieve robust synchronization during the recognition, we must guarantee large Manhattan-distance between the candidate and its vicinity. This is assured if the slope of the pit, as showed on Figure 3, is as big as possible. For selecting the best distance-pit and its corresponding candidate section, we should determine the steepness of the pit-side. However, because it is generally not constant, so as an alternative we calculate the width at a given value. Figure 4 shows pit-width of 100 candidate sections, where the sections are extracted from the same record so that their first samples are consecutive in the record. In Figure 4 the horizontal axis is the sample position of the candidate in the record, while the vertical axis is the width of the pits at a record-adaptive level.

Our reference selection algorithm is based on the same principle, but since our pattern recognition method uses the first frame as kernel and grows from the first record, we observed this pit-width for one frame long candidates. The minimum value has to be found in this function without calculating every point of it.

We must also assure that our selected reference does not occur any more in the record again or in any other records. Using database terminology, we would say that the reference must be a key. To avoid unambiguous identification we must try to identify the selected reference in all the other records and if it is not a unique key then a new reference must be selected.

The exact solution would require us to compare every one of the reference-candidates to every other one. This would mean a lot of comparisons even in the case of a few records that could not be done in a conceivable time period.

In the presented algorithm we tried to keep the number of comparisons as low as possible. To do so we examine only the vicinity of the reference candidate in a region having the width w, where w is expressed in number of samples. Also we do not examine all possible reference candidates, only every 100th. The algorithm is listed in Table 1.

RESULTS

Using this algorithm we selected references from 69 advertisements that previously were recorded from a live Internet audio stream. During the tests, we monitored these 69 adver-

Table 1. The reference selection algorithm of the EMESE

1. In the first turn, the reference candidate is selected from the $\frac{w}{2}$th sample of the first region of the given record (The *region* is w=5000 samples). The first region begins on the first sample of the record, and it will define the first sample of the frame.
2. The frame is compared to all possible frames in the region according to the distance metric mentioned above. As a result we get $d(i)$, where $i=0\ldots w-N$, as shown on Figure 1.
3. The next region is selected $k*N$ samples forward in the record, and step 2 is repeated. We select further regions the same way and calculate the corresponding $d(i)$ until we reach the end of the record.
4. The steepest pit and the corresponding i_{opt} frame position in the record is selected examining all the $d(i)$ functions for the narrowest pit.
5. In the $k*N$ vicinity of position i_{opt} the frame with the narrowest distance-pit is determined using a gradient search algorithm. This is the lowest point of the function on *Figure 2*.
6. The reference consisting of N_R slices (transformed frames) is extracted from the record beginning with the frame selected in the previous step.
7. This reference is tested for uniqueness using the recognition algorithm. If the reference appears in the record more than once, not only at the correct position, then the next best reference must be selected in the previously described way.
8. The reference is then tried against all the other records, to filter out in-set false alarms. If the reference is found in any other record, step 7 is used for reference reselection.
9. The above steps are applied to all other records.

Figure 5. Percentage of patterns successfully identified by the recognition algorithm

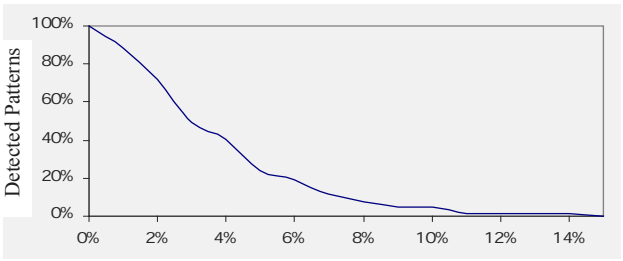
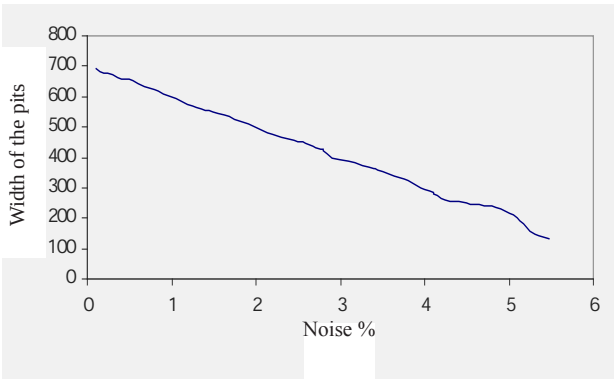


Figure 6. Result of the synchronization test



tisements that were broadcast in test Internet stream-media. We added white noise to the input signal to test the robustness of the system. The duration of the test was 48 hours. The recognition results are shown on Figure 5.

We also observed the performance of the system, namely how many references can be handled in real time. The computer used was equipped with a Pentium-II-350 MHz processor and 256 MB of RAM, and the maximum possible number of references was 258. If the record set to be monitored is added to, the reference selection for the added record must be performed, and the new references have to be checked for false alarms. If we detect a possible false alarm due to representative signal similarity the selection must be repeated for the whole set. This takes 525 minutes in case of 69 records. This is a worst-case scenario, and it should be very rare. The average selection time for every new record is 10 minutes.

The second test was a synchronization test. We selected 50 frames from the stream and we observed the dependency of the width of the pit on the level of the noise, and the noise level where the monitoring algorithm cannot synchronize to the frame. The result of this test is shown on Figure 6.

FUTURE TRENDS

Carrying out tests of various pattern recognition methods on live audio broadcasts showed that the success of identification process depends on the proper selection of the representative short segment. The position where this representative segment can be extracted is determined by the recognition algorithm. The selected references must be non-correlated to avoid false alarms.

The proposed method EMESE is an example of the Internet-oriented pattern recognition algorithms, which can be used for content-based search in media streams and files. The performed experiments checked how a monitoring system can synchronize to the stream under various conditions. The measured results proved the efficiency of the method however, more research efforts are necessary to reach a comprehensive content-based search method.

The Bark-spectrum as the base for the audio feature carries the possibility to easily process sampled audio with different sampling properties, since it hides the frequency resolution of the used Fast Fourier transform. However,

this generalization indicates further investigations to define the usable sound-print size and comparison thresholds. The reason for that is the unequal size of feature's frequency bands defined by the sampling frequency.

The system is designed to quickly change the parameters and thresholds for investigating the behavior of longer sound-prints, like a whole record. The idea is the application for automatic identification of even corrupted audio files.

CONCLUSION

The main directions of searching in shared file storages, especially the novel method based on the reference selection has been described. The proposed method was realized for an existing real-time recognition algorithm that was used on live audio streams to identify specific sound signals. The selection algorithm takes the properties of the recognition algorithm into account. The algorithm was tested on Internet media streams with a prerecorded signal set and we reached good results. Further tests should be carried out to determine the exact effect of the input noise level on the width of the distance-pit.

The light-weight pattern recognition methods presented earlier has proved that in Internet-wide environment it is possible to realize a fast recognition method, which does not require extraordinary CPU time or other computing resources. The increasing demand for searching in various media contents involves the popularity of the content-recognition tools. However, there are results, but further researches must be carried out in this field.

REFERENCES

Adar, E., & Huberman, B.A. (2000). Free riding on Gnutella. *First Monday, 5*(10). Retrieved June 6, 2005, from *http://firstmonday.org/issues/issue5_10/adar*

Akihito, M. A., Hamada, H., & Tonomura, Y. (1998). Video handling with music and speech detection. *IEEE Multimedia, July-September,* 16-25.

Coen, L. (1995). *Time-frequency analysis*. Prentice Hall.

Cohen, B. (2003, May). *Incentives build robustness in bittorrent*. Retrieved June 6, 2005, from http://bitconjurer.org/BitTorrent/bittorrentecon.pdf

Hosszú, G. (2005). Mediacommunication based on application-layer multicast. In S. Dasgupta (Ed.), *Encyclopedia of virtual communities and technologies* (pp. 302-307). Hershey, PA: Idea Group Reference

Kondoz, A. M. (1994). *Digital speech*. England: John Wiley & Sons.

Liu, Z., Wang, Y., & Chen, T. (1998). Audio feature extraction and analysis for scene segmentation and classification. *Journal of VLSI Signal Processing Systems for Signal, image and Technology*, 20(October), 61-79.

Lourens, J. G. (1990). Detection and logging advertisements using its sound. *IEEE Transactions on Broadcasting*, 36(3/September), 231-233.

Parker, A. (2004). *The true picture of peer-to-peer file sharing*. Retrieved June 8, 2005, from http://www.cachelogic.com

Qiu, D., & Srikant, R. (2004). Modeling and performance analysis of BitTorrent-Like peer-to-peer networks. *ACM SIGCOMM'04*, Portland, Oregon, USA, Aug. 30-Sept. 3 (pp. 367-378).

Richly, G., Kozma, R., Hosszú, G., & Kovács, F. (2001). A proposed method for improved sound-print selection for identification purposes. In N. Mastorakis, V. Mladenov, B. Suter, & L. J. Wang (Eds.), *Advances in scientific computing, computational intelligence and applications* (pp. 455-458). Danvers, USA: WSES Press (www.press.com).

Richly, G., Kozma, R., Kovács, F., & Hosszú, G. (2001). Optimised soundprint selection for identification in audio streams. *IEE Proceedings-Communications, 148*(5/October), 287-289.

Richly, G., Varga, L., Hosszú, G., & Kovács, F. (2000). Short-term sound stream characterization for reliable, real-time occurrence monitoring of given sound-prints. *MELECON2000*, Cyprus, May 29-31 (pp. 526-529, Vol.2).

Yang, B., & Garcia-Molina, H. (2002, July). Efficient search in peer-to-peer networks. *Proc. of IEEE International Conference on Distributed Computing Systems (ICDCS'02)*, Vienna, Austria (pp. 5-14).

KEY TERMS

Application Level Network (ALN): The applications, which are running in the hosts, can create a virtual network from their logical connections. This virtual network is also called ***overlay*** (*see later in the section*). The operations of such software entities are not able to understand without knowing their logical relations. The most cases this ALN software entities use the ***P2P model*** *(see later in the section)*, not the ***client/server*** *(see later in the section))* one for the communication.

Audio Signal Processing: It means the coding, decoding, playing, and content handling of the audio data files and streams.

Bark-Scale: A non-linear frequency scale modeling the resolution of the human hearing system. One Bark distance on the Bark-scale equals to the so-called critical bandwidth that is linearly proportional to the frequency under 500Hz and logarithmically above that. The critical bandwidth can be measured by the simultaneous frequency masking effect of the ear.

Client/Server Model: A communicating way, where one host has more functionality than the other. It differs from the ***P2P model*** (*see later in the section*).

Manhattan-Distance: The L_1 metric for the points of the Euclidean space defined by summing the absolute coordinate differences of the two points (|x2-x1|+|y2-y1|+…). Also known as city block or taxi-cab distance; a car drives this far in a lattice-like street pattern.

Overlay: The applications, which create an ***ALN*** *(see earlier in the section)* work together, and they usually follow the ***P2P communication model*** *(see later in the section)*.

Pattern Recognition: It means the procedure of finding a certain series of signals in a longer data file or signal stream.

Peer-to-Peer (P2P) Model: A communication way where each node has the same authority and communication capability. They create a virtual network, overlaid on the Internet. Its members organize themselves into a topology for data transmission.

Synchronization: It is the name of that procedure, which is carried out for finding the appropriate points in two or more streams for the correct parallel playing out.

Context-Aware Framework for ERP

Farhad Daneshgar
University of New South Wales, Australia

INTRODUCTION

Like many existing ERP models (e.g., Podolsky, 1998; Van Stijn & Wensley, 2001), the OOAB framework is also based on a widely accepted assumption that a corporate-wide information system consists of a set of potentially related subsystems; and as a result, information flows among these subsystems must be identified, and required resources planned, using an appropriate ERP methodology. However, up until now there existed no formalised framework that facilitates sharing of contextual knowledge in ERP processes. A unique attribute of the OOAB framework is that it treats ERP processes as a collaborative processes where various roles/actors collaboratively perform tasks in order to achieve a common overall goal. An object-oriented framework is presented in this article that facilitates sharing the contextual knowledge/resources that exist within ERP processes. Context is represented by a set of relevant collaborative semantic concepts or "objects". These are the objects that are localised/contextualised to specific sub-process within the ERP process.

BACKGROUND

From a purely object orientation perspective, a collaboration is defined as "the structure of instances playing roles in a behavior and their relationships" (OMG, 2001). The behaviour mentioned in this definition refers to an operation, or a use case, or any other behavioural classifier. This article provides an overview of a framework for analysing awareness requirements of the actors in ERP systems using an object-oriented awareness-based approach. A similar study was also conducted for developing a new version of this framework that takes into consideration the specific characteristics of virtual communities (Daneshgar, 2003). The proposed approach specialises the notion of collaboration and extends it to the ERP processes. This has roots in the activity network theory (Kaptilini et al., 1995) and is based on the fact that all ERP processes involve *multiple roles performing various tasks* using appropriate artefacts (e.g., departmental sub-systems, databases, etc.) in order to achieve both their local as well as the overall organization-wide goals. Conceptually speaking, this will justify a frame-based object-oriented approach to analysis and design for ERP processes (Turban & Aaron, 2001). The conceptual model of the proposed framework is made of the following components:

- a set of collaborative semantic concepts including roles, the tasks that these roles play within the process, and the artefacts that these roles use to perform various tasks within the process, and
- relationships among these semantic concepts.

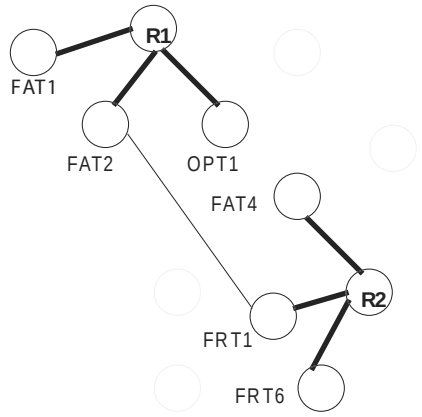
This conceptual model can then be mapped directly to an object model and be used as an analytical tool for identifying awareness requirements of the actors within the ERP process. The fact that ERP is treated as a collaborative process calls for a mechanism for maintaining awareness requirements of the actors involved in this collaboration. Furthermore, due to its object orientation, the framework is capable of encapsulating all complications and dependencies in sub/local processes within individual tasks as well as resources required to perform those tasks, further relieving the ERP management and the associated software.

OOAB FRAMEWORK

A domain-specific conceptual model of a hypothetical ERP process that resembles an object diagram is shown in Figure 1. Use of a domain-specific conceptual model instead of a straight object diagram is justified by the fact that the ontological foundation of the framework prevents growth of the objects and relationships indefinitely, and as a result using an object model may hide such ontology. In Figure 1 there are two roles: R1 and R2; six tasks: FAT1, FAT2, OPT1, FAT4, FRT1 and FRT6 (all shown by circles). It also shows various resources by straight lines connecting tasks and roles. These lines represent rich ontological relationship between a pair of semantic concepts. Each task object requires certain resources for achieving its local/departmental goal or purpose (called *process resource*), as well as certain other resources for achieving the collaborative organization-wide goals of the ERP process (called *collaborative resource*). In Figure 1, a line connecting a role vertex to a task vertex is a process resource, whereas a line connecting two tasks together is a collaborative resource.

According to the framework, effective knowledge and/or resource exchange among actors is closely related to the level of awareness as defined in the awareness model that each

Figure 1. A representation of an ERP collaborative process model



actor possess about the ERP process. These awareness levels are defined in terms of the collaborative semantic concepts used within the ERP conceptual model as shown in Figure 1. Details of the proposed methodology for identifying awareness requirements of actors in ERP process follow:

STEP 1. Develop an ERP Process Model similar to that in Figure 1.
FA: Financial Accounting sub-process/task
OP: Order Processing sub-process/task
CS: Customer Service subprocess/task
FR: Financial Reporting subprocess/task
T1…T6: <appear as postfixes indicating various tasks>

STEP 2. Measure the actual levels of awareness for each role on the process model using the awareness model. In order to measure this level of awareness the actor must be exposed to all the objects on the ERP process model, and be asked to identify those objects that s/he is aware of. Selected pool of objects are then used by an awareness model in order to arrive at a number reflecting the actual level of awareness associated with that role.

STEP 3: The actor's actual level of awareness is then is compared against the required level of awareness; the latter is a parameter, provided by the task that the actor performs within the process. The difference between these two levels of awareness constitutes the collaborative requirement of the actor for that particular task. Factors that affect the required level of awareness of a task include organisational culture, and the nature of task itself. Without possessing such awareness level the actor will not be able to collaborate with others optimally. A comparison between the actual level of awareness of the actor and the required level of awareness of the task will result in one of the following two outcomes:

1. The task's required level of awareness is either equal to, or less than, the role's actual level of awareness. This indicates that the role is qualified, or has sufficient level of awareness for taking up the task, and the OOAB framework cannot enhance collaboration any further.
2. The task's required level of awareness exceeds the role's actual level of awareness. This indicates potential for enhancing collaboration. To do so it will be necessary to put the missing objects within the focus of the actor in a way that s/he can perceive these objects, receive required awareness, and perform that particular task successfully. This will require additional resources in order to enhance the actor's awareness level. These required resources may include one or more of process resources, collaborative resources, and other communication resources, for example resources that provide awareness about other roles and other tasks within the ERP process.

IMPLEMENTATION ISSUES

One method for integration of the OOAB framework with the existing ERP systems is by developing an organisational infrastructure that provides business intelligence to the users of the ERP system by maintaining contextual knowledge that these users/actors require for effective collaboration within the ERP process. The writer is in the process of developing an expert system that provides expert advice for answering the following two specific questions:

(i) In terms of awareness requirements, is an actor capable of performing certain tasks within the ERP process?
(ii) If not, what objects need to be put within his/her focus in order to enable the actor to perform the task properly?

The ERP collaborative process of Figure 1 consists of 15 objects, including two roles, six subprocesses/tasks, six role artefacts and one task artefact. Within each of these objects is encapsulated all relevant contextual knowledge as well as pointers to relevant objects as determined by the process map. Each task possesses a set of attributes and relevant methods; and each method consists of a set of steps that corresponds to codes describing a codified knowledge. These attributes will indicate to which subprocess the task belongs to. This will enable an actor to play various roles within different subprocesses without being permanently linked to a specific subprocess, a factor that can remove some complexities in existing ERP implementations.

FUTURE TRENDS AND CONCLUSION

It was shown that the OOAB methodology can enhance existing ERP systems by formally adhering to a framework that facilitates knowledge sharing among various actors in ERP processes in the following ways: This article introduces a novel concept for measuring collaboration in ERP processes; a concept that is non-existent in current ERP literature. This measure is based on a conceptual model of collaboration that can be directly transformed into an object model. This measure is used to determine required level of awareness of various actors within the ERP process. As far as the author is aware, this is a novel concept that results in treating the organisation-wide ERP process as a single collaborative process that consists of multiple sub-processes, called tasks, that are linked through formal artefacts. These artefacts utilise certain resources in order to carry/deliver contextual collaboration knowledge that is required for effective collaboration. Knowledge sharing occurs through various interaction acts such as exchanging the artefacts, creation of artefacts, and updating artefacts. As a result of its object-orientation and collaborativeness of the ERP process, and contrary to the existing process-based ERP frameworks that assume fixed roles within the ERP process, the OOAB framework enables an actor to assume multiple roles within different ERP subprocesses. This will provide another avenue for sharing contextual knowledge for sub-processes/tasks. The interdependency issue among various subprocesses is also simplified by encapsulating this knowledge within the task objects and relevant artefacts.

By reducing granularity of the collaborative process model, the same model representation can be applied for internal representation of the subprocesses, allowing smooth transition of modelling components from subprocesses to the ERP process, and vice-versa. This in turn will reduce much of the existing complexities in designing ERP systems where the system is permanently responsible to maintain such linkages, rather than delegating such responsibility to various objects within the system.

Work is in progress for incorporating communication and coordination dimensions to the existing collaboration dimesion in order to provide complete analysis of groupware systems that maintain awareness requirements of the actors within the ERP processes (Daneshgar et al., 2004; Sundarraj et al., 2002).

REFERENCES

Abramson, B.D. (1998). Translating nations: Actor-network theory in/and Canada. *Canadian Review of Sociology and Anthropology, 35*(1), 1-20.

Daneshgar, F. (2003). Context management of ERP processes in virtual communities. In G. Grant (Ed.), *ERP & datawarehousing in organizations: Issues and challenges* (pp. 119-130). Hershey, PA: IRM Press.

Daneshgar, F., Ray, P., Rahbi, F., & Godar, C. (2004). Knowledge sharing infrastructures for teams within virtual communities. In M. Fong (Ed.), *e-Collaborations and virtual organizations.* Hershey, PA: IRM Press.

Kaptelinin, V., Kuutti, K., & Bannon, L. (1995). Activity theory: Basic concepts and applications. In Blumenthal et al. (Eds.), *Human-computer interaction. Lecture Notes in Computer Science.* Springer.

Object Management Group. (2001). *OMG Unified Modeling Language Specification – Version 1.4* (pp. 3-100).

Podolsky, M. (1998, July 20–24). An integrated approach to object-oriented modeling of business processes. *ECOOP'98 Workshop on Object-Oriented Business Process Modeling,* Brussels, Belgium.

Sundarraj, R.P., & Sarkis, J. (2002). Implementation management of an e-commerce-enabled enterprise information systems: A case study at Texas Instruments. In L. Hossein, J.D. Patrick & M.A. Rashid (Eds.), *Enterprise resource planning: Global opportunities & challenges* (pp. 133-148). Hershey, PA: IRM Press.

Turban, E., & Aarons, J.E. (2001). *Decision support systems and intelligent systems*. NJ: Prentice Hall International Inc.

Van Stijn, E., & Wensley, A. (2001). Organizational memory and the completeness of process modelling in ERP systems: Some concerns, methods and directions for future research. *Business Process Management Journal - Special Issue on Enterprise Resource Planning, 7*(3).

KEY TERMS

Action: A sequence of goal-directed steps.

Actual Level of Awareness: The awareness that a role actually possesses within the ERP process. Actual awareness is represented by an integer number ranging from zero to four, representing various levels of awareness. Actual awareness is a property of an actor who performs one or more roles within the ERP process.

Awareness: A specialised knowledge about the objects that leads an actor to an understanding of various aspects of the ERP collaborative process. It is defined and measured in terms of the semantic concepts (*task, role, process resource,* and *collaborative resource*) used in the map.

Awareness Model: A model that represents various levels of awareness. Level-0 awareness consists of the concepts that lead an actor to knowledge about all the tasks that an actor performs within the process.

Example: In Figure 1, level-0 awareness for the role "R1" is a sub-graph that consists of the tasks "FAT1," "FAT2," and "OPT1," as well as the process resources shown by thick lines connecting "R1" to these tasks. Level-1 awareness is a subgraph that consists of the R1's level-0 awareness subgraph, plus awareness about the concepts that leads an actor to knowledge about other related roles within the process. In Figure 1, level-1 awareness for the role "R1" is the sum of its level-0 awareness subgraph, plus one collaborative resource linking "FAT2" to "FRT1," plus "FRT1" itself, plus the process resources shown by thick lines connecting "FRT1" to "RT2," plus "RT2" itself. A role's level-2 awareness is its level-1 awareness, plus an awareness about all other (or, the rest of) roles within the process. Level-2 awareness is knowledge about the human boundary of the process. In Figure 1, there are no other roles that have not been known to the R1 already, and therefore the R1's level-2 and higher levels of awareness are irrelevant and identical to its level-1 awareness. However, for the sake of completeness, their definitions are presented.

A role's level-3 awareness is its level-2 awareness, plus awareness about all the interactions (represented by the process resources used/shared) that occur between any two roles within the process. And finally, level-4 awareness is the highest level of awareness that a role can have in any ERP process. It is defined as the knowledge about how everything fits together to form the ERP process. In other words, having this level of awareness will bring all the remaining concepts used in the ERP process model of Figure 1 within the focus of the role.

Collaborative Resources: An object representing a resource used/shared/exchanged by a pair of collaborating roles in order to perform certain simple tasks in collaboration with one another.

Process Resource: An object that represents a resource used by a role in order to perform a task in isolation from other tasks.

Required Level of Awareness: A property of a task. It represents the expected awareness from any actor who performs the task. Its value also ranges from 0 to 4.

Role: A set of norms expressed in terms of obligations, privileges, and rights assigned to an actor.

Task: An object with a set of attributes and actions to achieve a specific process goal using certain resource called *process resource.*

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 569-572,

Contingency Theory, Agent-Based Systems, and a Virtual Advisor

John R. Durrett
Texas Tech University, USA

Lisa Burnell
Texas Christian University, USA

John W. Priest
University of Texas at Arlington, USA

INTRODUCTION

In this article, we investigate the potential of using a synthesis of organizational research, traditional systems analysis techniques, and agent-based computing in the creation and teaching of a Contingency Theoretic Systems Design (CTSD) model. To facilitate understanding of the new design model, we briefly provide the necessary background of these diverse fields, describe the conceptualization used in the integration process, and give a non-technical overview of an example implementation in a very complex design environment. The example utilized in this article is a Smart Agent Resource for Advising (SARA), an intelligent multi-agent advising system for college students. To test all of the potential of our CTSD model, we created SARA utilizing a distributed instructional model in a multi-university, multi-disciplinary cooperative design process.

Just as a dynamic task environment forces an organization to compress its management structure and to outsource non-core activities in order to become flexible, a dynamic software development environment forces designers to create modular software. Until now, cooperative development paradigms were too complex to facilitate inter-organizational cooperative development efforts. With the increasing popularity of standards-based Web services, the development of pervasive computing technologies, and the advent of more powerful rapid application development languages and IDEs, this limitation has been removed. Our purpose in this research is twofold: first, to test the viability of using Contingency Theory (CT), a sub-discipline of Management Organizational Theory (OT), in an agent-based system; and second, to use these new technologies in creating a distributed instructional model that will allow students to interact with others in diverse educational environments. As an example implementation, we create a virtual advisor that will facilitate student advising in distributed environments.

In the following sections, we outline the background theories involved in the conceptualization of our design model. We start with the shifts in systems design techniques and how CT can be applied to them and to various Multi-Agent Systems (MAS) to allow Contingency Theoretic Systems Design (CTSD). Once the necessary background is in place, we briefly discuss our new eLearning approach to cooperative distributed education. Finally, the structure of the SARA is discussed.

BACKGROUND

Multi-Agent Systems

Agents and communication protocols form the basic components of a multi-agent system. Agents exchange messages according to a protocol of expected messages delivered in a communication language in which the message content and format adhere to a shared standard. Individual agents make decisions, which may include contacting other agents for information, and perform processing to satisfy their goals

An agent is commonly defined as a program or collection of programs that lives for some purpose in a dynamic environment and can make decisions to perform actions to achieve its goals. In other words, agents are goal-based programs that must deal with changing access to resources, yet run continuously. Like the best administrative assistants, agents know and adapt to their master. Individual agents may be conceptualized as having beliefs, desires, and intentions that can communicate with other agents to satisfy their goals. Multi-agent systems are those in which multiple agents (usually) cooperate to perform some task. Agents may be independently developed and allow the decomposition of a complex task into a collection of interacting agents that together solve some problem. It is not necessary that an individual agent "understand" the overall system goals or structure.

Agent communication can be viewed at four distinct levels. The first level is the expected protocol for exchanging sequences of messages, like a script. For example, when negotiating, the parties expect bids to be offered, rejected, and counter-offered. The second level relates to the content or mean-

ing of the messages. To enable inter-agent communication, an ontology is created. Examples of such concepts are things, events, and relationships. At the third level, a representation language defines the syntax for structuring the messages; The Knowledge Interchange Format (KIF) (Gensereth & Fikes, 1992) is one example. At the fourth level, an agent communication language (ACL) such as the Knowledge Query and Manipulation Language (KQML) or the Foundation for Intelligent Physical Agents (FIPA) ACL (Labrou, Finin, & Peng, 1999), defines message formats and message delivery. An example KQML message, in Sandia Lab's Java Expert System Shell (JESS) (Owen, 2004), that shows how an agent registers a service is shown below:

(register :sender student :receiver advisor :reply-with msg1 :language JESS :ontology SARA :content '(MajorCourses: Compliance Check Hours))

Just as human systems created to achieve complex goals are conceived of as organizations, multi-agent systems can be conceptualized as "organizations of agents". Individual components, whether human employees or software agents, need to be managed, guided toward a constructive goal, and coordinated toward the completion of the necessary individual tasks. In "empowered organizations", lower-level employees have the knowledge and authority to perform many tasks without the intervention of superiors. This conceptualization allows us to use well-established research from management organization theory (and Contingency Theory in particular) in creating guidelines for the design of agent-based systems.

ORGANIZATIONAL THEORY (OT)

While much of the background concerning OT is explained in the main chapter below, the following is a brief overview of the relevant research trends. OT examines an organization's structure, constituencies, processes, and operational results in an effort to understand the relationships involved in creating effective and efficient systems. A major division of OT, Contingency Theory (CT), postulates that no organization operates without constraints from environmental, personnel, technological, and informational influences (Andres & Zmud, 2001). This relationship is explained by the information processing theory (IPT) (Galbraith, 1973). IPT postulates that the more heterogeneous, unpredictable, and dependent upon other environmental resources a task is, the greater the information processing that the organization must be able to do in order to successfully accomplish it. As complexity and unpredictability increase, uncertainty increases due to incomplete information. As diversity of processes or outputs increases, inter-process coordination requirements and system complexity increase. As uncertainty increases, information-processing requirements increase. The basic premise of IPT is that the greater the complexity and uncertainty in the tasks in an organizational system, the greater the amount of information that the system must process (Galbraith, Downey, & Kates, 2001). A basic premise of our research is that this relationship is also true for information systems (Avgerou, 2001).

MAIN THRUST OF THE ARTICLE

Multi-Agent System Architectures Using CTSD

Contingency-theoretic system development (CTSD) adapts CT and IPT to the development and maintenance of software systems (Burnell, Durrett, Priest et al., 2002; Durrett, Burnell, & Priest, 2001, 2003). A business can organize employees in a number of different ways, for example by function or by project, and reorganize as the business environment changes. Software systems can benefit from this flexibility as well. The CTSD design approach is focused on design for maintainability, a crucial requirement for complex, dynamic systems.

Agent-based architectures are a means for structuring software systems that adhere to Contingency Theoretic principles. Each agent is viewed as an employee that has specific capabilities, responsibilities, and knowledge within an organization. Agents, like employees, are grouped into departments, as needed, to best satisfy the goals of the organization. Agents can communicate peer-to-peer within and across departments, and manager agents resolve conflicts and make resource allocation decisions.

Tightly interrelated tasks are grouped into one or more agents. Each of these groupings is referred to as a "software team", and parallels a department of employees that perform roughly equivalent jobs. For example, a set of agents that each handle one type of course requirement (e.g., lab, art appreciation) may be grouped into a team, where communication can occur quickly between these agents and with a "manager" agent that can resolve conflicts, exceptions, and course-independent tasks. An example agent in our system is encoded using JESS rules to check that student preferences (e.g., for afternoon courses) and constraints (e.g., no more than 12 hours per semester) are satisfied. Another agent offers heuristic advice as an actual advisor might. For example, a student may be able to enroll in 17 hours of math and science courses, but this may be strongly advised against, depending on the student's GPA and perhaps other factors.

Each agent in a multi-agent architecture has specific tasks to perform and communications requirements. Once an ontology and agent communication language has been specified, agents can be designed independently and integrated into the system to progressively add capabilities. Using CTSD principles, tasks that are dynamic and shared are grouped into support agents to enhance maintainability of the system. The

primary architectural decision is to separate knowledge based on two criteria: the degree of dynamicism and the degree of complexity. Dynamicism is exemplified by how frequently knowledge is expected to change over time, while complexity determines how abstract that knowledge is. A simple example describing the latter is the fact that "a student must be a junior with a GPA greater than 3.0 in all math and science courses". Dynamicism within the SARA domain was determined by advisor interviews, and analysis of degree requirements changes over time at three universities, both large and small and private and public.

Using the above CTSD conceptualizations in previous research (Burnell, Priest, & Durrett, 2002, 2003; Durrett, Burnell, & Priest, 2000), we have developed and tested the following CTSD guidelines for creating MAS:

1. *Describe business activity and identify tasks*: Allow management and developers to refine the overall purpose of the software being designed.
2. *Determine task predictability*: Since a basic premise of CT is that the control structure of a business process must match the environment in which it operates, we must identify the predictability of each task.
3. *Assign tasks to employees*: Once the level of predictability has been estimated for each task, the granularity of the employees being created can be determined and component designs finalized.
4. *Group employees into teams*: As with human organizations, our employees can be grouped along any of several dimensions, including task, workflow, product, manager, or communication requirements, as required by the operating environment.
5. *Identify communications needs:* Once teams are determined, the communication requirements of individual employees, and of teams, can be determined.
6. *Construct management groups*: In software systems operating in a dynamic environment, management is required only when employees are unable to handle events

.

We show the application of these guidelines in a distributed instructional environment in a multi-university, multi-disciplinary cooperative design process.

Multi-Agent System Architectures Using CTSD

Our example implementation of the above CTSD guidelines is SARA, an interactive tool intended to aid in the planning, scheduling, and advising process to complete a college student's degree. SARA, along with the instructional environment in which it was created, is described very briefly next; for more detailed information on SARA or our distributed instructional model please see any of the Burnell, Durrett, or Priest references in the reference section.

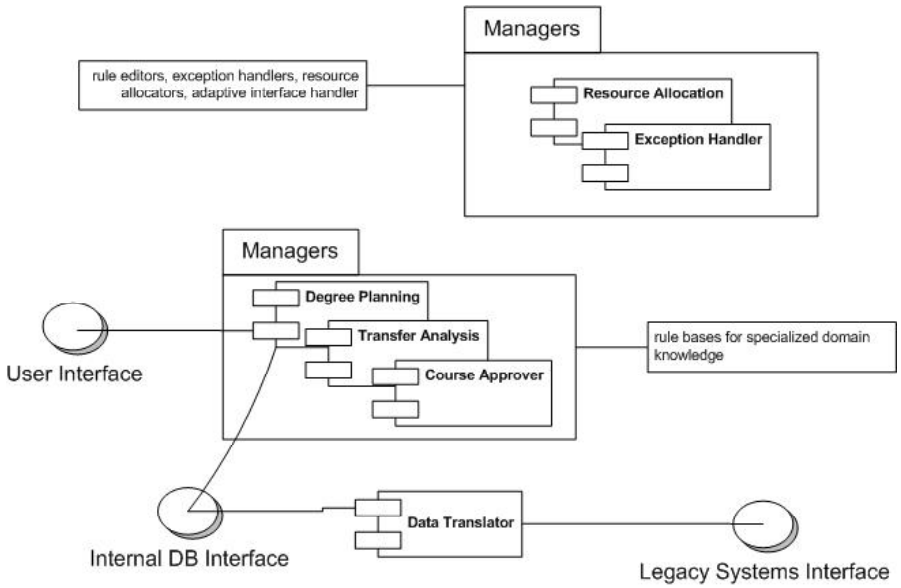
As described previously, CTSD is most effective in complex dynamic environments. To create this environment, SARA was designed in a collaborative education project among three major universities, the University of Texas at Arlington (UTA), Texas Christian University in Fort Worth (TCU), and Texas Tech University in Lubbock (TTU). Teams of students at UTA are in industrial engineering and software design courses. The TCU students are in computer science and have had coursework in artificial intelligence, and the TTU teams are in courses that study application servers and server-based JAVA databases. In creating SARA, responsibilities have been purposefully segregated to create the necessity of interdisciplinary cooperation. Current trends in industry are toward outsourcing in many major Fortune 500 companies (Fox, 2004). In order to do this effectively, the designer and the coders must communicate in an online forum and adapt to rapidly changing conditions. Therefore, we have attempted to emulate this CTSD type environment through our interdependent classes. Also, previous research has shown that a major hurdle to software development is the communication among the programmers, the users, the executives, and the domain experts. A major component of successful software development is overcoming this communication hurdle. We have attempted to emulate these trends in the design of SARA.

Design of SARA

SARA gives students the resources and knowledge to troubleshoot issues with their current degree plan. A profile of student preferences and a transcript of completed courses allow the system to generate customized degree plans. Students can also manually customize degree plans and create schedules course-by-course that are checked for requirements compliance. Errors in plans and schedules are explained so that the student can make corrections. Advisors can review degree plans on the system and send comments back to the students, without their having to physically meet. Thus, some of the most time consuming tasks in advising are eliminated (Priest, Burnell, & Durrett, 2002).

The current prototype of SARA that is depicted in the following diagrams and text is the work of several semesters work at all 3 universities. As a result of these cooperative efforts, a basic system prototype (shown in Figure 1) has been developed, initial system databases have been designed and implemented (shown in Figure 2), and smart user interfaces that utilize MAS (shown in Figure 3) were created. The user interfaces were designed primarily by the TCU teams, the basic CTSD-based analysis and thus overall system requirements were created by the UTA students, and the backend database design and application server support were provided by the TTU students.

Figure 1. Prototype architecture (from Priest et al., 2002)



The architecture depicted previously (in Figure 1) follows the CTSD guidelines discussed. Given the dynamic, complex environment in which the system will operate, we have segregated tasks into individual software teams and empowered the constituent agents with as much decision making ability as possible. Exceptional situations are handled by "manager agents" and by other exception handlers.

The prototype SARA database module depicted next (in Figure 2) was created to provide the flexibility required by the MAS system architecture. Initial implementations are on MySQL, and migration to IBM DB2 to provide for more system automation is planned.

The MAS-based user interfaces (one example of which is depicted in Figure 3) were created using data from TCU's course schedule. They allow students to do most of the routine work in degree planning, freeing up advising time for more meaningful discussions. Once proposed plans are created, they can be evaluated and approved off-line.

FUTURE TRENDS

Our future plans include broadening the scope of the project just described using an open source software development model. Our intent is to research the potential in open source design models for relieving some of the coordination issues inherent in a resource limited project such as ours. We hope that this new model will relieve some of the faculty coordination required and still allow (or force) students to cooperate with each other even though in different disciplines and locations.

CONCLUSION

In the project just described, we have developed a multi-agent system for college advising using the contingency-theoretic system development (CTSD) process. Multi-disciplinary teams of students at three universities were employed to create the system. Once the ontology and agent communication language are defined, individual agents are constructed incrementally. The resulting system is a hierarchy of intelligent agents that together provide support for course scheduling and degree planning and that are adaptable to changing environments. The CTSD process applied to the creation of intelligent multi-agent systems results in maintainable systems operating in complex, dynamic domains. As such domains become increasingly automated, training and support tools for distributed CTSD will be needed. We have created an approach for such training that has been successfully applied to SARA and other projects over the last three years. Support tools will need to be created that capture not only the system design, but provide knowledge-based analysis of designs based on the principles we have defined.

REFERENCES

Andres, H.P., & Zmud, R.W. (2001). A contingency approach to software project coordination. *Journal of Management Information Systems, 18*(3), 41-70.

Figure 2. SARA database entity diagram

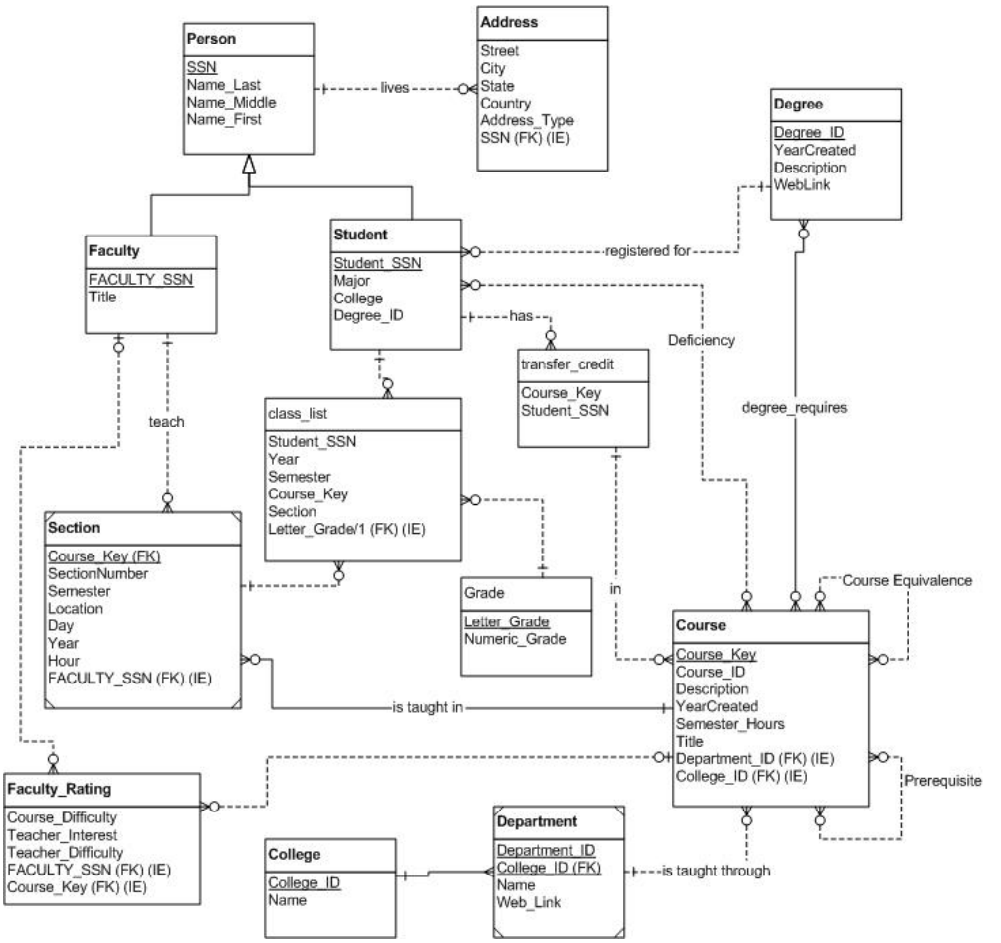
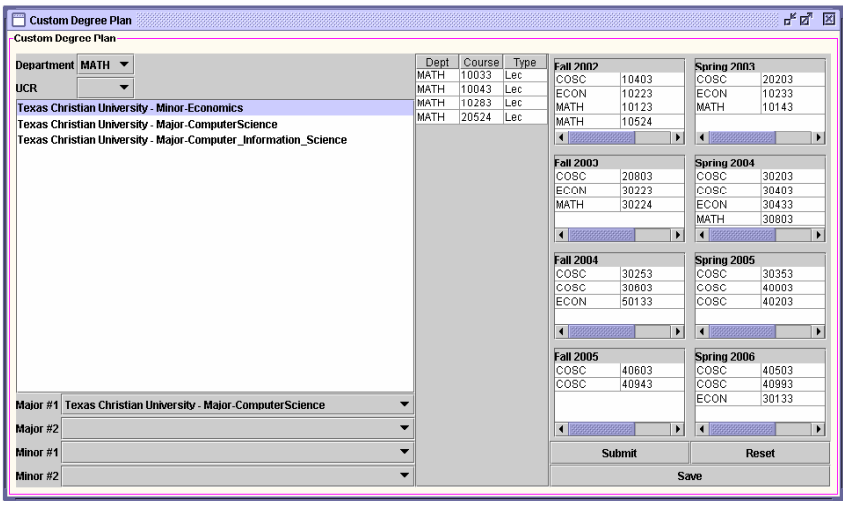


Figure 3. Custom degree planner interface



Avgerou, C. (2001). The significance of context in information systems and organizational change. *Information Systems Journal, 11*, 43-63.

Burnell, L.J., Durrett, J.R., Priest, J.W., et al. (2002). A business rules approach to departmental advising. Paper presented at the *Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference (FLAIRS-2002),* Pensacola Beach, FL.

Burnell, L.J., Priest, J.W., & Durrett, J.R. (2002). Teaching distributed collaborative software development. *IEEE Software, 19*(5), 86-93.

Burnell, L.J., Priest, J.W., & Durrett, J.R. (2003). Assessment of a resource limited distributed multidisciplinary process for teaching software development in a university environment. *ACM Inroads*.

Durrett, J.R., Burnell, L., & Priest, J. (2001, July 29-August 2). An organizational behavior approach for managing change in information systems. Paper presented at the *Proceedings of PICMET*, Portland, Oregon.

Durrett, J.R., Burnell, L.J., & Priest, J.W. (2000, November 2-8). Contingency theoretic methodology for agent-based, web-oriented manufacturing systems. Paper presented at the *SPIE: Photonics East 2000*, Boston, MA USA.

Durrett, J.R., Burnell, L.J., & Priest, J.W. (2003, December). A hybrid analysis and architectural design method for development of smart home components. *IEEE Personal Communications,* 2-9.

Fox, S. et al. (2004, March 8). Offshoring. *InfoWorld, 26,* Special Issue.

Galbraith, J., Downey, D., & Kates, A. (2001). *Designing dynamic organizations: A hands-on guide for leaders at all levels*. New York, NY: Amacom.

Galbraith, J.R. (1973). *Designing complex organizations*. Reading, MA: Addison-Wesley.

Gensereth, M.R., & Fikes, R.E. (1992). *Knowledge interchange format, version 3.0 reference manual*. Santa Clara, CA: Stanford University.

Labrou, Y., Finin, T., & Peng, Y. (1999). The current landscape of agent communication languages. *Intelligent Systems: IEEE Computer Society, 14*(2), 45-52.

Owen, J. (2004, March 15). Budget-minded BRMS: JESS and OPSJ are faster, cheaper, and harder to use. *InfoWorld, 26,* 24-26.

Priest, J., Burnell, L., & Durrett, J.R. (2002, May 19-22). SARA: Smart, agent-based resource for virtual advising. Paper presented at the *International Resource Managers Association*, Seattle, Washington.

KEY TERMS

Contingency Theoretic Software Development (CTSD): A new model for MAS design using tenets from CT and IPT. The CTSD design approach is focused on design for maintainability, a crucial requirement for complex, dynamic systems.

Contingency Theory (CT): A research branch of organizational theory that suggests that an organization's structure reflects its adaptation to the environment in which it operates. Hierarchical organizations operate best in stable, simple environments while flat, team-based organizations are better adapted to dynamic, complex task environments.

E-Learning: Any form of education or training that utilizes online media and remote connectivity for all or part of its curricula. This model includes both purely online courses and those in brick-and-mortar universities facilitated by email, the Internet, newsgroups, or chat.

Information Processing Theory (IPT): An explanation for the organization structure-environment relationship suggested by CT. IPT suggests that the information processing requirements dictated through interactions with the environment force certain structures in order to be efficient and effective.

Multi-Agent Systems (MAS): Multi-agent systems are those in which multiple agents (usually) cooperate to perform some task.

Ontology: An ontology is a well-defined set of concepts that are ordered in some manner to create an agreed-upon vocabulary for exchanging information.

Smart Agent: A program or collection of programs that lives for some purpose in a dynamic environment and can make decisions to perform actions to achieve its goals. Individual agents may be conceptualized as having beliefs, desires, and intentions that can communicate with other agents to satisfy their goals.

Software Team: Groups of agents which have tasks that are tightly interrelated; these teams roughly parallel a department of employees.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 64-69, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Electronic Commerce

Contributions of Information Technology Tools to Project's Accounting and Financing

R. Gelbard
Bar-Ilan University, Israel

J. Kantor
University of Windsor, Canada

L. Edelist
Bar-Ilan University, Israel

INTRODUCTION

"According to the Standish Group CHAOS Report 2003, each year in the USA there are approximately 175,000 projects in IT application development that spends $250 Billion. Among these, 31.1% of projects will be cancelled, 52.7% of projects will cost 189% of their original estimates, only 52% of required features and functions make it to the released product, and time overruns occur in 82% of the cases. In financial terms $55 billion dollars is wasted in these projects." (Madpat, 2005).

This chapter suggests an innovative platform to analyze software projects in order to overcome the difficulties that are shown through the statistics. The first layer of the platform is based on costing theories in order to handle the cost overruns. At the second layer are the project management tools, and on the third layer is the software engineering. The last two layers give the needed information on the project scope and the development efforts. Connecting those three layers gives a better perspective on the projects, which is the best platform for decision making.

Cost management of a project is defined by the PMBOK (project management body of knowledge) (PMI, 2004) as one of the nine core activities of projects management. This activity is defined as an assembly of processes that include planning, estimating, budgeting, and controlling of project costs so that the process will be executed within the budget framework that has been designated for it. However, although it defines costing as a core activity, it does not provide the methodologies for the application mode of the costing (Kinsella, 2002).

The challenge in project management is described as "the effective allocation of resources within the framework of time, cost and delineation constraints that are balanced against the quality demands and nature of relations with the customer" (Kerzner, 2003. p.5). Hence, cost management should be viewed as part of the project management challenge.

Software projects can be analyzed through software engineering tools, CASE (computer-aided software engineering tools), that assist in the analysis and characterization of the software project and in the evaluation and measurement of the work productivity in the project.

Cooper and Kaplan (1998) analyze the integration between costing systems and operational systems. The integration that Cooper and Kaplan introduce, like the classic costing methods, does not provide a response to the project structure and the features of a software project (such as estimation difficulties, risk management, and lifecycle). This chapter recommends integrating costing systems and operational systems of software projects; the projects management tools and the software engineering tools.

The data presented highlights the significance of costing and the difficulties in costing and estimating software projects. These difficulties derive both from the implementation's limitations of a costing solution in an intricate and changing technological environment (Wouters & Davila, 2004) and from the unique features of projects in general and software projects in particular. The characteristics that obstruct the solving of the costing problem include the project lifecycle that leads to changing work capacities over time (Kerzner, 2003), uncertainty levels and exposure to risk (Rajkumar & Rush, 2000), and a difficulty in defining an evaluation of the project scope.

Given all this, the conclusion that becomes clear is that there is an objective difficulty in establishing an accurate cost framework for the software project, especially prior to its detailed planning. Such planning is executed through software engineering tools. Those tools assist the analysis of the software project and the estimation and measurement of the project's work productivity (Liong & Maciaszek, 2005).

We have seen that cost management within the software project framework requires the combining of software engineering with the involvement of the development team. However, the development team's ability to be fully involved in the cost management process is limited. The develop-

ment team and projects managers function in monitoring the changing technological implementation throughout the project, and in the knowledge management of the project team. Hence, the amount of remaining time for the costing activity is small. Moreover, in order to accurately define the cost structure, there is a need for a costing model that includes, in addition to the direct costs of the project, also overhead costs in the organization. Project managers that work on the engineering and technical aspects of the system struggle with objectively defining and applying such a model (Wouters & Davila, 2004).

Given this, the chapter presents a model that allows the expression of each and every one of the cost's components (direct, indirect, risk, competitiveness), while it links three areas: project management, software engineering, and managerial accounting. The model will enable not only a retrospective analysis of the economic performances of the project/projects portfolio/software house, but also an in-advance evaluation of costs, economic feasibility, and economic risk level of the project/projects portfolio. The model introduces a new approach in the area of software project costing.

THEORETICAL BACKGROUND

This section presents the theoretical and practical foundation for the research model from several aspects. Our objective is to integrate models of software projects with three disciplines: software engineering, project management, and managerial accounting.

The first section reviews the foundation for the integration between the financial systems that serve the classic models of costing and other relevant systems.

The second section reviews the link between software engineering and project management. This section will emphasize the importance of the association between software engineering and project management tools as a managerial and costing necessity in the software project (or projects portfolio). In the third section, costing aspects are introduced and integrated. A basic integrative model for this association will be displayed, and we shall examine the extent of its compatibility with the costing need.

INTEGRATED COST SYSTEMS

In order to make important managerial decisions, detailed costing information is necessary. Detailed costing information is expected to include all types of costs that are required for manufacturing a product or providing a service. These are data that are based on financial systems and contain, in practice, costs that derive from the firms detailed income statement (and backup schedules). These data include the historical execution data and future estimations and forecasts (Needy, 2002).

Otley (2001) proposes an integration of accounting and financial data for obtaining execution and evaluation measures. It is suggested that these measures will be supplemented by information that is not financial. The need of nonfinancial information has evolved in recent decades out of the comprehension that the costing analysis is not sufficient for it being an outcome analysis: this has created the need for performance measurement. Such measurement includes the accounting information and costing logic with the incorporation of figures that are not financial. Various models (such as balanced scorecard) deal with performance measurement (Needy, 2002), and provide an additional layer for the need to execute integration between financial and operational systems.

Williams (2004) supports the integration approach in accordance with the viewpoint that a modern accounting system is supposed to supply a framework for strategic management of the company's resources. In order to realize this conception, Williams proposes a multidimensional construct that clusters information from the company's systems on customers base, activity areas, and more, for the purpose of forming an accounting system that facilitates planning, improvement, and control, analysis and regulation of resources, and enhancement of profitability. Such a system is based on integrative information from a number of systems or from the arrays DWH (data where house)/BI (business intelligence) in five areas: costs, assets, quality/service, time, and outputs.

The pioneers of the combining of financial and operational information are Cooper and Kaplan (1998), who developed the method of activity based costing (ABC) suggest, in light of the technological development of information systems, to define the integration between operational and financial systems for the purpose of building an accurate costing model.

SOFTWARE ENGINEERING AND PROJECT MANAGEMENT

The success of a software project depends on five software engineering areas that are related to each other: the development of the lifecycle of the software, process management, the model's configuration and language, software engineering tools, and project planning (Liong & Maciaszek, 2005. p.3). The combining between the formal tools of the software engineering and project management processes in its different stages has been proved by research as to result

in a positive contribution to the efficacy of the project, and as an improver of the adherence to costs, technical requirements, and the schedules that were allocated to the project (Verma & Barker, 2003). In this study, the researchers argue that following the combination of the areas, it is difficult to separate the contribution of the software engineering from the total contribution. Hence, an initial support may be assumed for the essentiality of the association between the areas. The integration's foundation that was displayed may be extended since it is impossible to separate the costing activity, which deals with financial values, from the total managerial activity, thus, also for the purpose of cost management (cost execution), the combination between software engineering and project management must be considered.

The combining of the areas is described as a comprehensive view of the software development process with the description of the products, resources, schedules, budgets, and the organizational structure that sustains the project. Such a framework provides not only the stages in the planning of the project's monitoring and controlling, but also strategies for managing project's risks (Rajkumar, 2003).

As an anchor for the costing model that will be presented in the study, a model that establishes, in a feasible way, the link between software engineering and the project manager's activities was chosen (Gelbard, Pliskin, & Spiegler 2002). The basis of this study is the mapping between the software engineering, which is implemented by CASE tools and project management tools. The mapping model allows a shift from an engineering description of the realized project, for example, as DFD (data flow diagram), to a description that includes also the project's schedule, which is usually portrayed through Gantt charts. The advantages of such mapping are as follows:

- The evaluation of the effort and costs over time while directly drawing from the engineering basis.
- Extended analysis of costs, time, and resources that also includes engineering parameters, such as complexity and demanded quality.
- Generic fittingness to every type of software project.
- A dynamic controlling of the project's process and evaluations of execution against the planning.
- The ability for "drill down" based both on functionality and intermediate products and on the level of the system's design.
- Integration with the development teams that allows reliable and accurate controlling and estimation.

Despite the model's advantages, it does not encompass project layers, such as an explicit allusion to the project's lifecycle and its risk management (including the cost of the risk). Moreover, the response is given at the single project level and does not reflect effects of project portfolio management, such as indirect costs, transportation of resources between projects, and the like.

On the other hand, the model's application allows an engineering and project insight in terms of time and cost of the project's constituents. Therefore, the gripping on to the model's advantages creates groundwork for the development of a costing model in software projects that will be built on the foundation of integration between the different areas.

COST MODELS AND METHODS

Detailed costing information is expected to include all types of costs that are required for manufacturing a product. Data based on financial systems, which contains costs, derived from the income statement and the estimation of the company's capital and assets, enclosed the historical execution data and future estimations and forecasts (Needy, 2002). Williams (2004) supports the integration approach according to the conception that a modern accounting system is supposed to supply a framework for strategic management of the company's resources. In order to realize this conception, Williams proposes a multidimensional construct that clusters information from the company's systems on customers base, activity areas, and more, for the purpose of forming an accounting system that facilitates planning, improvement and control, analysis and regulation of resources, and enhancement of profitability. Such a system is based on integrative information from a number of systems, or from the arrays DW (data warehouse), BI (business intelligent) in five areas: costs, assets, quality/service, time, and outputs. The pioneers of the combining of financial and operational information are Cooper and Kaplan, who developed the method of activity based costing (ABC) at the end of the eighties. Cooper and Kaplan (1998) suggest, in light of the technological development of information systems, to define the integration between operational and financial systems for the purpose of building an accurate costing model.

In light of this, establishment of integration conception required definition not only of an enterprise costing model, but also definition of interfacing between the different areas and systems; that is, interface between SE aspects-tools, financial aspects-tools, and PM tools.

Cost management is a term that is used for a wide description of short-term and long-term managerial activities that are involved in planning and controlling of costs (Horngren, Dater, & Foster 2000). Table 1 presents variety aspects of costing model in a technological projects environment.

Costs analysis within the framework of technological environment must be carried out with the understanding of

Table 1. Aspects of a costing model in a technological project environment

	Aspect	Description	Difficulties
1.	**Planning**	Costs estimation of the project and for each resource in the projects portfolio	Defining direct and indirect resources and their costs
2.	**Controlling**	Costs analysis for each project and executed task	Attributing in-reality-costs to each project's task
3.	**TimeLine**	Costs analysis over different time periods in planning and execution	Evaluating capacities of resources consumption over specified time periods
4.	**Tasks**	Identification and costing of project's tasks (WBS items)	Matching the costs to each of the project's components
5.	**Overhead Allocation**	A precise allocation of indirect costs	Determining the indirect cost generators in project's tasks
6.	**Risk management**	The inclusion of risk element and its value as part of the costing	Estimating risk on the basis of risk factors in the different tasks
7.	**Scenarios**	The ability to analyze alternative modes of action and costs	Defining assumptions and alternatives to the mode of cost's calculation
8.	**Profitability Analysis**	The understanding of the profit that derives from each of the projects and the whole projects portfolio	The inclusion of all the cost factors in the model

the project lifecycle. Kerzner (2003) portrays the distribution of the project's cost over the project's lifecycle:

- 5% - Conceptualization
- 10% - Feasibility study
- 15% - Preliminary planning
- 20% - Detail Planning
- 40% - Execution
- 10% - Testing and Commissioning

Tasks in each of these stages are described under the work breakdown structure (WBS). The WBS represents the required activities for the project's management in a hierarchical structure. For each component of the WBS, an evaluation of direct and indirect (overhead) costs must be included. Direct costs are divided to work's cost (usually work hours multiple hourly rate) and direct costs that are not work payment, such as travel, materials, and so forth. It is recommended that these costs will include managerial reserve as well (Jurison, 1999).

A reinforcement of the need to include the project's tasks (or the WBS components) in a costing model is intensified in the light of the cost estimations that are founded on work hours' evaluation. It has been argued (Ooi & Soh, 2003) that according to traditional approaches of software costing (time-based estimations), there may be a bending towards time planning without linking it to the specific task and the role player that performs it. Therefore, it is suggested to include the detailing of the tasks (Ooi & Soh, 2003) and/or an elaborate planning of the various project's resources as part of the costing model.

The advantages of the resources' cost analysis throughout activities/tasks: more detailed information for managers, monitoring abilities, analysis of resources' cost and allocation, and a more accurate ability of overhead allocation (Elnathan & Raz, 1999; Jahangir, 2003; Kinsella, 2002; Ooi & Soh, 2003).

Indirect costs (overhead costs) include all types of costs that cannot be attributed directly to a specific task in the project marketing and sales expenses, office supplies, buildings' cost, professional services, information systems, computerization infrastructure, and the like. These costs are only occasionally incorporated in the project planning, but they carry great influence on the profitability of the portfolio and the projects' pricing decisions (Horngren et al., 2000). These costs are described as one of the "major headaches" (Kerzner, 2003). However, in this context, it has been argued that the ability to control costs is largely dependent on the monitoring of these costs.

Table 2 summarizes costing methods according to financial and engineering literature. The table also presents the common evaluation of model compatibility in light of entire costing aspects.

- **Analogy:** Cost estimation based on previous experience, using case based reasoning techniques (CBR).

Table 2. Costing methods according to financial and engineering literature

Legend √ - Good compatibility X - No compatibility P - Partial compatibility * - Adjustments are required			**Planning**	**Controlling**	**Time Line**	**Task Resolution**	**Overhead Allocation**	**Risk Management**	**Scenarios**	**Profitability Analysis**
Software Eng.	Top Down	Analogy	√ *	P*	X	X	P*	X	X	X
		Parametric	√ *	P*	X	X	P*	P*	√	X
	Bottom Up	Function Points	√	X	X	√	P*	√*	√	X
		COCOMO II	√	P	X	√	P*	√	√	X
Costing		Target Costing	√	P	P	P*	P	X	P	√
		Standard Costing	√	P	√*	√*	P	X	√	P
		ABC	√	P	P*	X	√	X	√	√

The accuracy of this method ranges from -10% to +25% (Kerzner, 2003).

- **Parametric:** Cost estimation based on heuristics and thumb's rules (Jahangir, 2003). Similar to the analogy estimation method, a parametrical model is also based on accumulation of historical data of project costs. On the basis of these data, a mathematical model is defined for the prediction of costs (Kinsella, 2002). The level of accuracy of a parametrical model ranges on a wide scope of –25% to +75% (Kerzner, 2003).
- **Function Points:** A method that was first introduced in 1979 by Albrecht. Its objective is to assess the software system's size while using the user's requirements without direct dependence on the technological realization (Hale & Smith, 2001). The function points method is calculated in three steps, using the quantity and complexity of the functional components and the system attributes (Kemerer, 1993).
- **COCOMO** (constructive cost model): The model was first introduced in 1981 and since then, several modifications were made in order to suit fourth-generation languages, decrease in hardware costs, increase in QA levels, advanced and agile development methods. Current version, COCOMO 2.0 (Boehm, Clark, Madachy, Horowitz, Selby, & Westland, 1995), is not based upon line of codes, but on four submodels that match a spiral approach of software system development that are applied according to the stage of the lifecycle (The application-composition model, the early design model, the reuse model, and postarchitecture model).
- **Target costing:** Suits engineering framework in which there are several engineering activities simultaneously, and is utilized as a means for costs strategic management. The idea behind the method is that a product's cost must be based on the sum that can be received for it in the market, and in other words, the development cost should be the basis for the quantity and mode of investment in the development rather then the development's outcome.
- **Standard costing:** ascertains the cost framework while employing the amount of direct cost components and a standard price that was set for this unit. We shall formulate it concisely:

$$TotalCost = \sum_{i-1}^{n} Qty_i * StdP_i .$$

It should be accentuated that the standard price does not solely include the direct price of the component (price per working hour), and is intended to contain the meaning of the cost or the consumption of indirect resources (rent, computerization, etc.). In the calculation of the standard price, it is customary to rely on known performance data from the past (Horngren et al., 2000).

- **Activity based costing (ABC):** Is considered as one of the advanced models for predicting costs while incorporating managerial decisions. The model was developed in the eighties, and its main innovation is in the addition of nonfinancial elements to the cost-

ing model. The model is widely used in a variety of industries, such as agronomy, banking (Kao & Lee, 2001), and medicine. In the projects area, there is not much literature that discusses the application of ABC, however, there are a few studies that help to understand the method. These studies include the description of the method for software developing and assimilation (Ooi & Soh, 2003), the portrayal of the mode in which ABC can be taken on in projects (Elnathan & Raz, 1999), the implementation of ABC in favor of IT cost analysis in the organization, and a recommendation to include this model in PMBOK (Kinsella, 2002).

THE FRAMEWORK

The model proposes an approach to the integration between the financial systems that include the components of the financial reports and the software tools that serve the project management tools and software engineering tools, such as function points.

The model will enable not only an in-retrospect analysis of the economic performances of the project/projects portfolio/software house, but also an in-advance evaluation of costs, economic feasibility, and economic risk level for the project/projects portfolio. The model enables the execution of sensitivity analyses for the purpose of exploring alternatives of planning and organizing (resource utilization, resource transportation between projects, etc.).

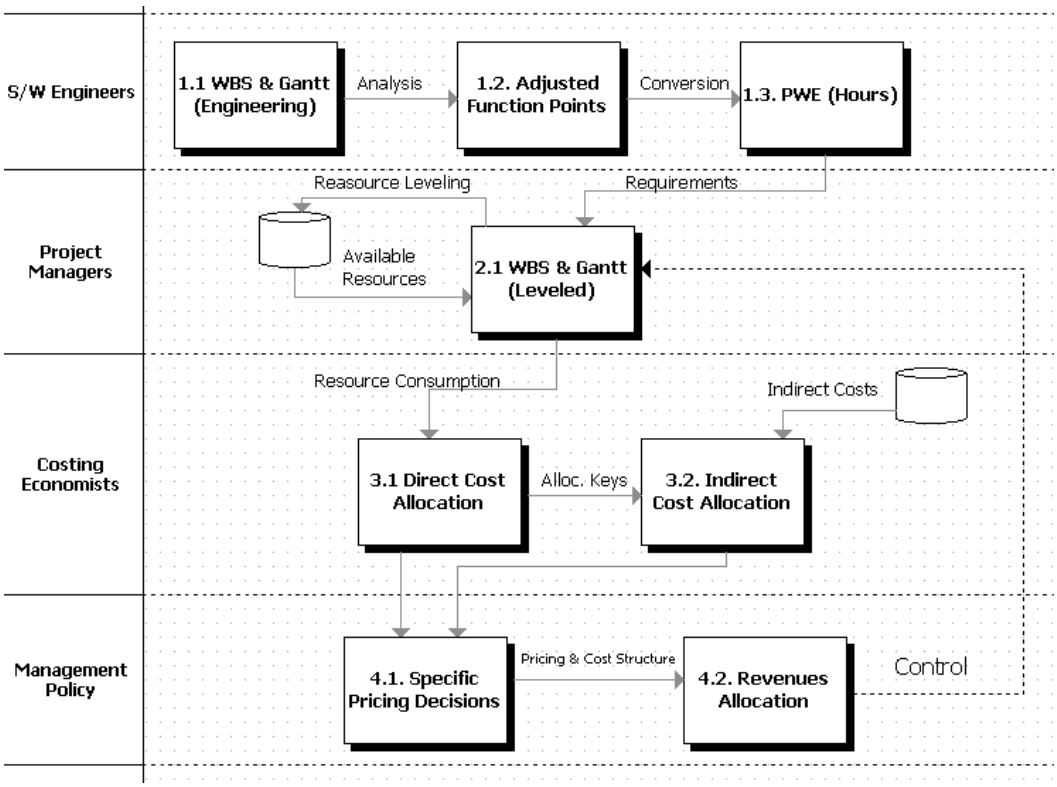
Costing Framework as a Business Process

In order to create a practical framework for costing, we have described the costing framework as a business process. The process combines four types of organization resources: software engineers (stages 1.1-1.3), project managers (stage 2.1), costing economists (stages 3.1-3.2), and managers (stage 4.1-4.2).

Figure 1 illustrates the costing activities and collaboration between the personnel:

The connection between the engineering planning and the project managing is based on the "translation" of the engineering planning terms to working hours, engineering complexity description (AFP – Adjusted Function Points) converted to developing efforts (working hours). In addition, on the top of the engineering planning (developing efforts and tasks composition), there are more constraints

Figure 1. Costing framework as a business process



that derived from the availability of the organization's resources. The project manager is testing the constraints, demands, and costs, and creates the optimal/ultimate GANT project. Those steps connected the managing and planning of the project to the engineering side that takes into account system complexity.

Thereafter, cost elements (that include direct cost calculations) are being introduced and overhead expenses are thereafter, added (indirect costs of the organization). The overhead expenses are based on the loading key that derives from the direct cost. At the end of this stage, we had for each project a "naïve" estimation, namely, a cost that does not depend on political factors (such as require profitability) that find expression at the last stage.

In the final stage, after building the costing structure for the whole project, we have management intervention (with the collaboration of planning factors). Within this framework, management can define the policy for each project according to the project type, the customer, and so on. Political decisions could influence the basics cost structure (costing decisions) and/or determine the profitability level that is necessary for the project (pricing).

This model reflects the integrative framework of the organization, the management, and the project, and raises questions about the effectiveness and applicability level of the model.

CONCLUSIONS

This work presents the costing of software projects as a business process that collaborates between entities in the organization.

Costing models usually have three stages: defining a costing object, cost tracing for direct costs, and cost allocation for indirect cost. Due to the complexity of software projects, these stages are not trivial to comply with. For example, the project life cycle (PLC) and the different tasks in it require the costing object to include several objects (to match the WBS). We have defined the following features for a costing model in order to fulfill its goals. These features are supported by the suggested framework we describe, yet are only partially supported by other known models. They include

- Planning -_understanding and estimating project activities that spread over the lifecycle of the project (project life cycle- PLC)
- Task analysis - The tasks (processes) in each of these stages are described under the work breakdown structure. The WBS represents the required activities for the project's management in a hierarchical structure. For each component of the WBS, an evaluation of direct and indirect (overhead) costs must be included.
- Direct resources_– Derived from engineering requirements, they are the work effort (like programmers) and material (like servers)
- Direct cost - Direct costs are divided into costs of labor (usually work hours multiple by hourly rate) and other direct costs such as materials. It is recommended that these costs will include a managerial reserve. (Jurison, 1999).
- Indirect costs - All types of costs that cannot be attributed directly to a specific task in the project (marketing and sales expenses, office supplies). These costs have a great influence on the profitability of the portfolio and the projects' pricing decisions (Horngren at al., 2000, p.148, 439). Indirect costs are described as one of the "major headaches" (Kerzner, 2003. p. 524). The ability to control costs is largely dependent on the monitoring of these costs.
- Control- An analysis of actual vs. budget and the ability to create forecasts based on planning and partial implementation.
- Timeline – Planning and control activities should take under consideration different types and volumes of activities during the lifecycle. Therefore, the project timeline becomes part of the costing model and its implementation.
- Profitability – Understanding the "bottom line" of the projects. Profitability analysis requires both an allocation of direct and indirect costs as well as revenue allocation.
- Risk analysis - The objective of risk analysis is to predict the value of uncertainty (the risk) that is involved in future project activity (Rajkumar, 2003). The literature shows that one of the challenges of a costing system is to introduce and represent risks as part of the costing process (Rajkumar & Rush, 2000).

The major point raised by these requirements is that the costing model should include a variety of cost information supported by different personnel. The advantage of resource cost analysis throughout the project life cycle is the creation of more detailed information for managers, better monitoring abilities, a more accurate analysis of resources, and better control over overhead allocation (Elnathan & Raz, 1999; Jahangir, 2003; Kinsella, 2002; Ooi & Soh, 2003).

REFERENCES

Boehm, B. W., Clark, B. K., Madachy, R., Horowitz, E., Selby, R. W., & Westland, C. (1995). Cost models for future software process: COCOMO 2.0. *Annals of Software Engineering, 1*, 57-94.

Cooper, R., & Kaplan, R. S. (1998). The promise - and peril - of integrated cost systems. *Harvard Business Review, July-*

August, 109-119.

Elnathan, D., & Raz, T. (1999). Activity based costing for projects. *International Journal Of Project Management, 17*(1), 61-67.

Gelbard, R., Pliskin, N., & Spiegler, I.(2002). Integrating system analysis and project management tools. *International Journal of Project Management, 20*(6), 461-468.

Hale, J., & Smith, R. (2001). An empirical study using task assignment patterns toimprove the accuracy of software effort estimation. *Software Engineering, IEEE Transactions, 27*(3), 264-271.

Horngren, C. T., Dater, S. M., & Foster, G. (2000). *Cost accounting* (10th ed.). Prentice Hall.

Howell, G., & Koskela, L. (2001). Reforming project management: The role of planning, execution and controlling. In *Proceedings of the 9th International Group for Lean Construction Conference* (pp.185-198).

Jahangir, M. (2003). Costing R&D projects: A bottom-up framework. *Cost Engineering, 45*(2), 12-17.

Jurison, J. (1999). Software project management: The manager's view. *Communication of the AIS, 2*(17), 1-57.

Kao, J., & Lee, T. (2001). Application of simulation technique to activity-based costing of agricultural systems: A case study. *Agricultural Systems,67*, 71-82.

Kemerer, C. F. (1993). Reliability of function points measurement: A field experiment. *Communications of the ACM , 36*(2), 85-98.

Kerzner, H. (2003). *Project management: A system approach to planning, scheduling and controlling*. Wiley & Sons Inc.

Kinsella, S. M. (2002). Activity-based costing: Does it warrant inclusion in a guide to the project management body of knowledge (PMBOK)? *Project Management Journal, 33*(2),49-56.

Liong, B. L., & Maciaszek, L.A. (2005). *Practical software engineering*. Addison-Wesley.

Madpat, S. (2005). Bridging organizational strategy & projects - An OPM3 insiders perspective. *The Milestone, PMI Memphis,* 5(10), 16-20.

Needy, A. (2002). *Business performance management - Theory and practice*. Cambridge University Press.

Ooi, G., & Soh, C. (2003). Developing an activity-based costing approach for system development and implementation. *The Data Base For Advances In Information Systems, 34*(3),.54-71.

Otley, D. (2001). Accounting performance measurement: A review of its purposes and practices. *International Journal of Business Performance Management, 3,* 245-260.

PMI. (2000,2004). *Project management body of knowledge (PMBOK).* PMI Institute.

Rajkumar, R., & Rush, C. (2000). Analysis of cost estimating processes used within a concurrent engineering environment throughout a product life cycle. In *7th ISPE International Conference on Concurrent Engineering* (pp.58-67).

Rajkumar , R. (2003). *Cost engineering: Why, what and how?. Decision Engineering Report Series*. Cranfield University Press.

Rolston, L. J., Grant, M. E., & Gardner, L.,L. (1994). Traditional vs. activity-based costing. In *1994 Winter Simulation Conference* (pp.10-50).

Sommerville, I. (2004). *Software engineering* (7th ed.). Addison-Wesley.

Stenzel, C., & Stenzel, J. (2005). An expert's perspective: A conversation with Gary Cokins. *Journal of Cost Management, 19*(1), 6-17.

Verma, D., & Barker, B.G. (2003). System engineering effectiveness: A complexity point paradigm for software intensive systems in the information technology sector. *Engineering Management Journal, 15*(3), 29.

Williams S. (2004). Delivering strategy business value. *Strategic Finance, 86*(2), 40-48.

Wouters, M., & Davila, A. (2004). Designing cost-competitive technology products through cost management. *Accounting Horizons, 18*(1), 13-26.

KEY TERMS

Activity Based Costing (ABC): A cost prediction model that has greatly improved the ability to predict a proper allocation of indirect costs among several activities and thereafter, between many products. Before using this model, one has to appreciate and understand the overall business (including its production and marketing). The model is not always applicable: a cost-benefit analysis is necessary before a final decision is made.

CASE/AMD Tools: Software tools (computer aided software engineering), also named AMD tools (analysis modeling and design), to assist the entire system life cycle (SLC). This includes the analysis phase, design phase, testing phase, and even maintenance phase. CASE /AMD tools support the

functional analysis phase using visual modeling notations, which can automatically be converted into code.

Cost Object: Anything for which cost data are desired. A cost object may be a product or a service. Cost object for projects can include a module, milestone, or a specific task. We recommend defining a cost object for a project as a WBS item.

Gantt Chart: A popular type of bar chart that illustrates a project schedule. Gantt charts illustrate the start and finish dates (ES, EF, LS, LF) of entire project elements. Project elements comprise the work breakdown structure (WBS) of the project. Gantt charts also show the dependency (i.e., precedence network) relationships between activities. Gantt charts can be used to show current schedule status using percentage completion shadings and a vertical "today" line.

Project Cost Management: Defined by the PMBOK (project management body of knowledge) as one of nine core activities of project management. This activity is described as the collection of processes including planning, estimating, and budgeting cost control, so that the project will carry out within the intended budgeting framework

Project Management: A wide discipline that includes the knowledge base and techniques for planning, controlling, and implementing projects. The PMI (Project Management Institute) defines it as follows: "Project management is the application of knowledge, skills, tools and techniques to project activities to meet project requirements."

Work Breakdown Structure (WBS): A technique for defining and organizing projects' tasks using a hierarchical tree structure. The first level contains one project outcome, and each level describes outcomes in details. Each level must include 100% of the total work (the sum of the work at the "child" level must equal 100% of the work represented by the "parent," and the WBS should not include any work that falls outside the actual scope of the project)

Creating Order from Chaos: Application of the Intelligence Continuum for Emergency and Disaster Scenarios

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Rajeev K. Bali
Coventry University, UK

INTRODUCTION

Recently, the world has witnessed several large scale natural disasters: the Tsunami that devastated many of the countries around the rim of the Indian Ocean in December 2004, extensive flooding in many parts of Europe in August 2005, hurricane Katrina in September 2005, the outbreak of Severe Acute Respiratory Syndrome (SARS) in many regions of Asia and Canada in 2003, and the earthquake disaster in Pakistan towards the end of 2005 . These emergency and disaster situations (E&DS) serve to underscore the utter chaos that ensues in the aftermath of such events, the many casualties and loss of life, not to mention the devastation and destruction that is left behind. One recurring theme that is apparent in all these situations is that irrespective of the warnings of the imminent threats, countries have not been prepared and ready to exhibit effective and efficient crisis management. This paper examines the application of the tools, techniques, and processes of the knowledge economy to develop a prescriptive model that will support superior decision making in E&DS and thereby enable effective and efficient crisis management.

BACKGROUND

Changing weather patterns, rapid urbanization, expansion of industry, not to mention development of air and ground transportation networks, population growth and migration, and recently – acts of terrorism , are associated with ever increasing frequency of major disasters involving multiple casualties (von Lubitz, Carrasco, Fausone, Gabbrielli, Kirk, Lary, and Levine, 2005). Emergency Healthcare management is a complex process which has to be tackled on various fronts (Beltrame, Maryni and Orsi, 1998; Kun and Bray, 2002). Such situations require effective crisis management capability, that is, pre-hospital and emergency/trauma, in-hospital medical services, firefighting, disaster-related law enforcement operations, and so forth, and superior decision making capabilities (von Lubitz, et al., 2005; von Lubitz and Wickramasinghe, 2005a; 2005b). Most of these services are governed by different local or national agencies, are subject to different rules and regulations, and develop independent operational plans. This in turn leads to the gathering and storing of data in disparate databases. However, given the interdependent nature of these elements, any decision making based on only one or a few of these data elements will logically provide only a partial picture and, thus, an inferior decision. Hence, it is necessary to collect multi-spectral data, analyze this data in aggregate to develop a complete picture if we are to truly support superior decisions. To do this effectively and efficiently, it is imperative to embrace the tools, techniques and processes of the knowledge economy (Liebowittz, 1999; Maier and Lehner, 2000; Shapiro and Verian, 1999; von Lubitz and Wickramasinghe, 2005b; Wickramasinghe, 2005; Wilcox, 1997; Zack, 1999). Advances in IT, coupled with the advent of Knowledge Management (KM), can facilitate better processes for efficient and effective healthcare (Dwivedi, Bali, James, Naguib, and Johnston; 2002).

MAIN FOCUS

The Intelligence Continuum consists of a collection of key tools, techniques, and processes of the knowledge economy; that is, including data mining, business intelligence/analytics and knowledge management which are applied to a generic system of people, process and technology in a systematic and ordered fashion (Wickramasinghe and Schaffer, 2005). Taken together, they represent a very powerful instrument for refining the data raw material stored in data marts and/or data warehouses and thereby maximizing the value and utility of these data assets. As depicted in Figure 1, the intelligence continuum is applied to the output of the generic information system. Once applied, the results become part of the data set that are reintroduced into the system and combined with the other inputs of people, processes, and technology to develop an improvement continuum. Thus,

the intelligence continuum includes the generation of data, the analysis of these data to provide a "diagnosis," and the reintroduction into the cycle as a "prescriptive" solution. In this way, continuous learning is invoked and the future state always builds on the lessons of the current state.

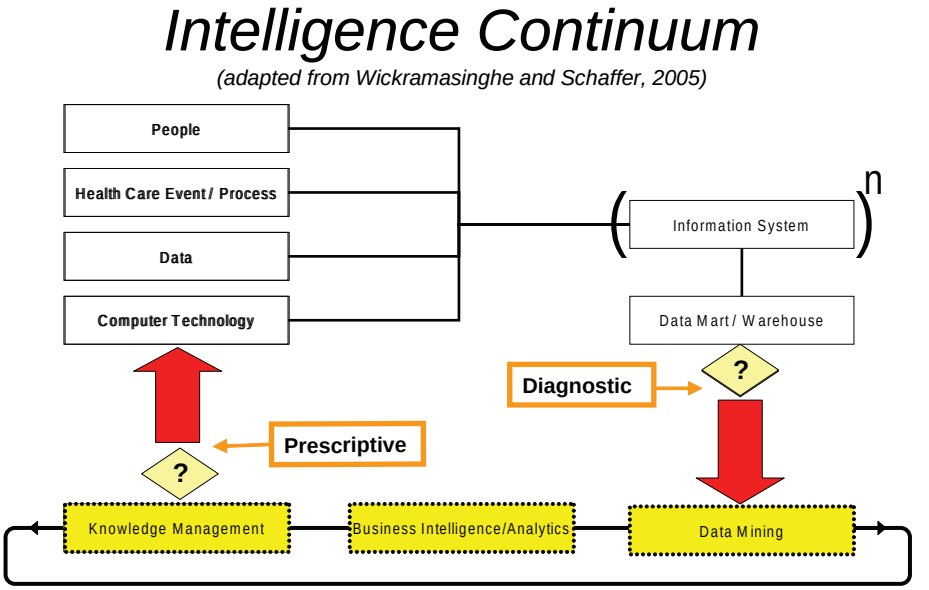
The key capabilities and power of the model are in analyzing large volumes of disparate, multi-spectral data so that superior decision making can ensue. This is achieved through the incorporation of the various intelligence tools and techniques which taken together make it possible to analyze all data elements in aggregate. Currently, most analysis of data is applied to single data sets and uses at most two of these techniques (Newell, Robertson, Scarbrough, and Swan, 2002; Nonaka, 1994; Nonaka and Nishiguchi, 2001; Schultze and Leidner, 2002; von Lubitz and Wickramasinghe, 2005b; Wickramasinghe, 2005; Wickramasinghe and Schaffer, 2005). Thus, there is neither the power nor the capabilities to analyze large volumes of multi-spectral data (ibid.). Moreover, the interaction with domain experts is typically non-existent in current methods. The benefits of applying the capabilities of the intelligence continuum to E&DS scenarios are profound indeed. E&DS scenarios are concomitant with complex, unstable, and unpredictable environments where the unknown or position of information inferiority prevails. Hence, these scenarios are chaotic and sub-optimal decision making typical results. In contrast, the tools and techniques of the intelligence continuum can serve to transform the situation of information inferiority to one of information superiority in real time through the effective and efficient processing of disparate, diverse, and seemingly unrelated data. This enables decision makers to make superior decisions which in turn lessen the chaos and facilitates the restoring of order. In order to appreciate the power of the intelligence continuum in such scenarios, it is necessary to briefly describe its key elements.

Data Mining

Due to the immense size of the data sets, computerized techniques are essential to help decision makers understand relationships and associations between data elements. Data mining is closely associated with databases and shares some common ground with statistics since both strive toward discovering structure in data. However, while statistical analysis starts with some kind of hypothesis about the data, data mining does not. Furthermore, data mining is much more suited to deal with heterogeneous databases, data sets, and data fields, which are typical of data in E&DS that contain numerous types of text and graphical data sets. Data mining also draws heavily from many other disciplines, most notably, machine learning, artificial intelligence, and database technology.

From a micro perspective, data mining is a vital step in the broader context of the knowledge discovery in databases (KDD) that transforms data into knowledge by identifying valid, novel, potentially useful, and ultimately understand-

Figure 1



able patterns in data (Adriaans and Zantinge, 1996; Bendoly, 2003; Cabena, Hadjinian, Stadler, Verhees, and Zanasi, 1998; Fayyad, Piatetsky-Shapiro, Smyth, 1996). KDD plays an important role in data-driven decision support systems that include query tools, report generators, statistical analysis tools, data warehousing, and on-line analytic processing (OLAP). Data mining algorithms are used on data sets for model building, or for finding patterns and relationships in data. How to manage such newly discovered knowledge, as well as other organizational knowledge assets, is the realm of knowledge management.

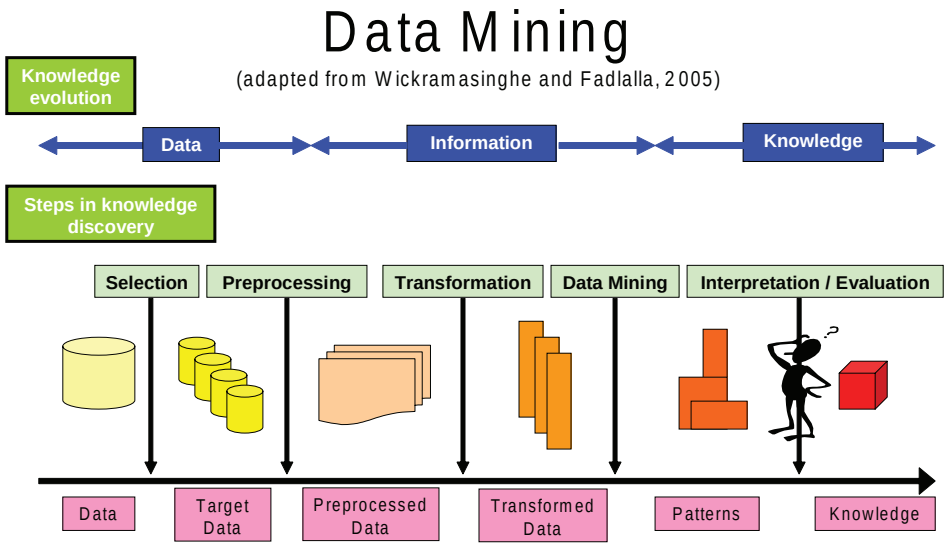
Figure 2 shows an integrated view of the knowledge discovery process, the evolution of knowledge from data to information to knowledge, and the types of data mining (exploratory and predictive) and their interrelationships. In Figure 2, all the major aspects connected with data mining are captured and by so doing the integral role of data mining to knowledge creation is emphasized. This is not normally explicitly articulated in the existing literature although the connection between data, information, and knowledge is often discussed (Becerra-Fernandez and Sabherwal, 2001; Choi and Lee, 2003; Chung and Gray, 1996; Holsapple and Joshi, 2002).

Data mining then, is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data (Fayyad, et al., 1996). It is essential to emphasize here the importance of the interaction with experts who always play a crucial and indispensable role in any knowledge discovery process in facilitating prediction of key patterns and also identification of new patterns and trends.

Business Intelligence/Analytics

Another technology-driven technique, like data mining connected to knowledge creation, is the area of business intelligence and the now newer term of business analytics. The business intelligence (BI) term has become synonymous with an umbrella description for a wide range of decision-support tools, some of which target specific user audiences (Wickramasinghe, 2005; Wickramasinghe and Schaffer, 2005). At the bottom of the BI hierarchy are extraction and formatting tools which are also known as data-extraction tools. These tools collect data from existing databases for inclusion in data warehouses and data marts. Thus, the next level of the BI hierarchy is known as warehouses and marts. Existing healthcare information systems are not generally designed to cater to new (data) needs (Anderson, 1997). Because the data come from so many different, often incompatible systems in various file formats, the next step in the BI hierarchy is formatting tools. These tools and techniques are used to "cleanse" the data and convert it to formats that can easily be understood in the data warehouse or data mart. Next, tools are needed to support the reporting and analytical techniques. These are known as enterprise reporting and analytical tools. On-line analytic process (OLAP) engines and analytical application-development tools are for professionals who analyze data and do, for example, business

Figure 2



forecasting, modeling, and trend analysis. Human intelligence tools form the next level in the hierarchy and involve human expertise, opinions, and observations to be recorded to create a knowledge repository. These tools are at the very top of the BI hierarchy and serve to amalgamate analytical and BI capabilities along with human expertise. Business analytics (BA) is a newer term that tends to be viewed as a sub-set of the broader business intelligence umbrella and concentrates on the analytic aspects within BI by focusing on the simultaneous analysis of patterns and trends in a given context (Wickramasinghe and Schaffer, 2005).

Knowledge Management

Knowledge Management is an emerging management approach that is aimed at solving the current business challenges to increase efficiency and efficacy of core business processes while simultaneously incorporating continuous innovation. Specifically, knowledge management through the use of various tools, processes, and techniques combines germane organizational data, information and knowledge to create business value, and enable an organization to capitalize on its intangible and human assets so that it can effectively achieve its primary business goals as well as maximize its core business competencies (Newell, et al., 2002; Nonaka, 1994; Nonaka and Nishiguchi, 2001; Schultze and Leidner, 2002; von Lubitz and Wickramasinghe, 2005b; Wickramasinghe, 2005; Wickramasinghe and Schaffer, 2005). The importance of knowledge management is confirmed by the increasing attention that the subject has received from both researchers and practitioners (Huang and Newell, 2003).

Broadly speaking, knowledge management involves four key steps of creating/generating knowledge, representing/storing knowledge, accessing/using/re-using knowledge, and disseminating/transferring knowledge (von Lubitz and Wickramasinghe, 2005a; Wickramasinghe, 2005b; Wickramasinghe and Schaffer, 2005).

Knowledge Management (KM) as a discipline is said not to have a commonly accepted or *de facto* definition. However, some common ground has been established which covers the following points. KM is a multi-disciplinary paradigm (Gupta, Iyer & Aronson, 2000) which often uses technology to support the acquisition, generation, codification, and transfer of knowledge in the context of specific organizational processes. Knowledge can either be tacit or explicit (explicit knowledge typically takes the form of company documents and is easily available, whilst tacit knowledge is subjective and cognitive). As tacit knowledge is often stored in the minds of healthcare professionals, the ultimate objective of KM is to transform tacit knowledge into explicit knowledge to allow effective dissemination (Bali, 2005).

KM initiatives should be incorporated in conjunction with the technological revolution that is occurring within healthcare organizations. A balance is required between organizational and technological aspects of the healthcare process (Dwivedi, et al. 2001a).

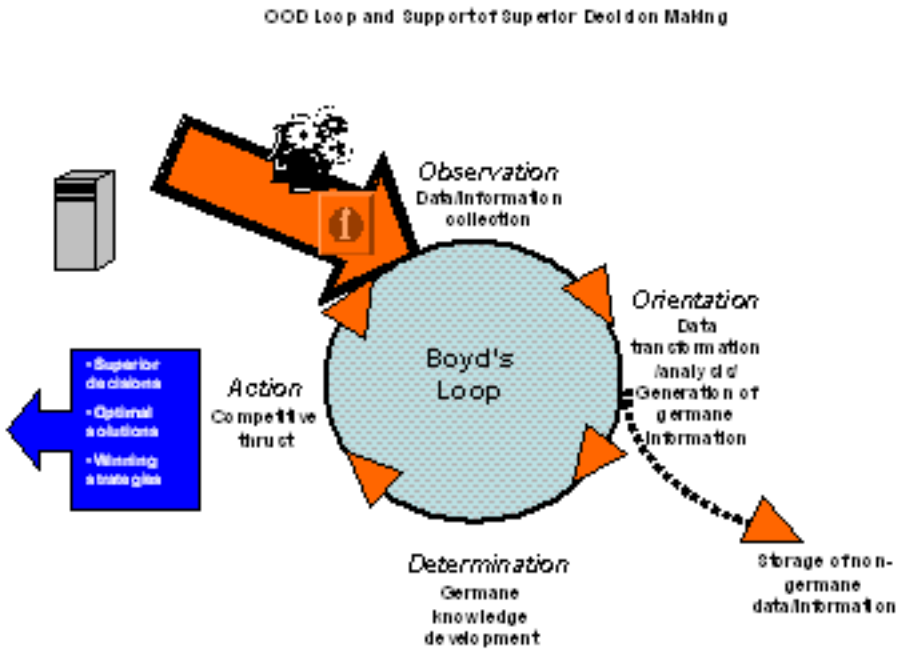
KM can enable the healthcare sector to successfully overcome the information and knowledge explosion by way of appropriate frameworks customized for healthcare institutions (Dwivedi, et al., 2001b, 2002a).

Knowledge Generation In Dynamic And Unpredictable Environments

Hierarchically, gathering of information precedes transformation of information into useable knowledge (Alavi and Leidner, 1999; Massey, Montoya-Weiss, and O'Driscoll, 2002). Hence, the rate of information collection and the quality of the collected information will have a major impact on the quality (usefulness) of the generated knowledge (Chang, et al., 2005). In dynamic and unstable environments, relative to the environment, the decision maker is in a position of tremendous information inferiority. In order to make effective decisions he/she must rapidly process seemingly irrelevant data and information into relevant and useable knowledge (Award and Ghaziri, 2004; Boyd, 1976; Courtney, 2001; Drucker, 1993; Newell, et al., 2002; Schultze and Leidner, 2002; Wickramasinghe, 2005). This necessitates a process perspective to knowledge management (von Lubitz and Wickramasinghe, 2005b; Wickramasinghe and von Lubitz, 2006). The cornerstone of such a perspective is the OODA Loop (Figure 3) which provides formalized analysis of the processes involved in the development of a superior strategy (Boyd, 1976, 1987; von Lubitz and Wickramasinghe, 2006).

The Loop is based on a cycle of four interrelated stages revolving in time and space: Observation, followed by Orientation, then by Determination, and finally Action. At the Observation and Orientation stages, multispectral implicit and explicit inputs are gathered (Observation) and converted into coherent information (Orientation). The latter determines the sequential Determination (knowledge generation) and Action (practical implementation of knowledge) steps. The outcome of the latter affects, in turn, the character of the starting point (Observation) of the next revolution in the forward progression of the rolling loop. The Orientation stage specifies the characteristics and the nature of the "center of thrust" at which the effort is to concentrate during the Determination and Action stages. Hence, the Loop implicitly incorporates the rule of "economy of force," that is, the requirement that only minimum but adequate (containment) effort is applied to insignificant aspects of competitive interaction. The Loop exists as a network of simultaneous and intertwined events that characterize the multidimensional action space (competition space), and both influence and are influenced by the actor (e.g., an organization) at the centre of the network.

Figure 3. (Adapted from von Lubitz and Wickramasinghe, 2005ab)



It is the incorporation of the dynamic aspect of the "action space" that makes the Loop particularly useful to environments that are inherently unstable and unpredictable, that is, medicine, business, war and emergency, and disaster scenarios (von Lubitz and Wickramasinghe, 2005a; 2005b; 2006).

FUTURE TRENDS

The need for more effective, superior crisis management techniques is clearly becoming apparent as we embark upon post crisis analysis for each of the more recent disasters from 9-11, to the Tsunami in December 2004, the floods in Europe, hurricanes Katrina and Rita and the earthquakes in Pakistan. This area is the focus for the emerging discipline of Operations Other than War (OOTW). Commonly considered as military, OOTW now becomes increasingly civilian-driven, and often executed as interventions in potentially unstable environments, or as management activities consequent to destabilizing events. In addition to the most obvious topic of terrorism, where risk assessment and management are prerogative to meaningful counteraction, problems of assessing and managing consequences of natural disasters, epidemic diseases, major industrial or transportation accidents, or humanitarian relief operations, become increasingly relevant. Key issues for OOTW include (Richards, 2004):

- Risk factors and their management
- Preparedness/readiness
- Political factors
- International organizations/national organizations/ NGOs in OOTW
- International cooperation
- Military/civilian interaction
- Law/law enforcement
- Healthcare and medical aspects
  - o telemedical operations
  - o medical logistics
  - o medical information networks in disaster operations
- ICT technology-based tools facilitating assessment and management of risk
  - o data mining/business analytics
  - o simulation/modeling
  - o training in complex synthetic environments
- Field operations and analysis of practical execution

As this nascent field evolves the above areas will form a central research focus for scholars in the near future.

CONCLUSIONS

Sound emergency management requires the ability to (Alexander, 2002, 2004; Marincioni, 2001):

1. Focus on solvable problems;
2. Priorities the elements of a problem in terms of how much progress can be achieved with each element in a small amount of time;
3. Delegate responsibility;
4. Manage the "span of control;"
5. Communicate clearly and rationally;
6. Keep a level head in a crisis; and
7. Make sound decisions.

However, when we analyze the recent natural disasters, a common recurring and unfortunate situation is that countries and regions are never as prepared and ready for the eminent disaster as they perhaps could have been. It is too late once the disaster strikes to have an organized and systematic fashion for contending with the aftermath. What is required is to be able to analyze past crises and develop appropriate lessons to apply to future events. In the advent of a health crisis, knowledge management is in a position to improve information sharing and coordination. The intelligence continuum model coupled with a process perspective of knowledge management appears to fill this void as it can be applied to existing and disparate data elements from past disasters in order to build a predictive model that can facilitate in the development of sound procedures and protocols to facilitate preparedness and readiness a priority so that *ex-ante* operations can, in fact, be more effective and efficient, decision making superior and order replace much of the chaos.

- Preparedness, unless based on broadly-based knowledge is useless. The development of appropriate preparedness is predominantly a strategic task that requires intimate knowledge of several aspects of the environment.
- Readiness, unless based on germane knowledge, is useless. Thus, coping with the sudden and unpredictable event requires the background of germane knowledge that will dictate the nature of the subsequent response. Readiness is, therefore, context dependent.
- Readiness is the most essential tool in response to, and containment of, an unexpected threat. While intuitively obvious, the practical development of readiness is not an easy task. Possession of knowledge is not equivalent to the ability to employ it under the stress of less-than-routine circumstances; that is, E&DS.

Hence in E&DS, what is needed is to be prepared and ready which, in turn, requires not only the possession of pertinent information and germane knowledge, but also the ability to apply it successfully; evoke superior decision making. Efficient flow of information is necessary in managing an outbreak (Kun and Bray, 2002). Hurricane Katrina serves to highlight how vulnerable and insufficient existing crisis management techniques are (CNN 2005a-c) as well as to underscore that developing better techniques through the utilization of critical data sources should be remedied immediately. The Intelligence continuum offers such a possibility.

REFERENCES

Adriaans, P. and Zantinge, D. (1996). *Data Mining*. Addison-Wesley.

Alavi, M. & Leidner, D. (1999). Knowledge Management Systems: Issues, Challenges and Benefits. *Communications of the Association for Information Systems.* Vol. 1 Paper #5.

Alexander, D. (2002). *Principles of Emergency Planning and Management*. Terra Publishings. London.

Alexander, D. (2004). Cognitive Mapping as an Emergency Management Training Exercise. *Journal of Contingencies and Crisis Management.* Vol.12, Dec., pp. 150-159.

Anderson, J.G. (1997). Clearing the way for physicians' use of clinical information systems,. *Communications of the ACM*, 40(8), pp.83-90.

Award, E. and Ghaziri, H. (2004). *Knowledge Management,* Prentice Hall, Upper Saddle River.

Bali, R.K. (2005) (Ed.) *Clinical Knowledge Management: Opportunities and Challenges*. IGP:USA.

Becerra-Fernandez, I. & Sabherwal, R. (2001). Organizational Knowledge Management: A contingency Perspective. *Journal of Management Information Systems* pp. 23-55.

Beltrame F., Maryni P., and Orsi G. (1998). On the Integration of Healthcare Emergency Systems in Europe: The WETS Project Case Study, *IEEE Transactions on Information Technology in Biomedicine*. Vol. 2, No. 2, pp.89-97.

Bendoly, E. (2003). Theory And Support For Process Frameworks Of Knowledge Discovery And Data Mining From ERP Systems. *Information & Management,* 40 pp.639-647.

Boyd, J.R. COL USAF, (1976). *Destruction and Creation, in R Coram "Boyd" Little*. Brown & Co, New York, 2002.

Boyd, J.R., COL USAF, (1987). In *Patterns of Conflict*, unpubl Briefing (accessible as *Essence of Winning and Losing*, http://www.d-n-i.net).

Brown, J.S., & Duguid, P. (2002). *The Social Life of Information*. Harvard Business School Press: Boston. pp. IX-328.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering Data Mining from Concept to Implementation*. Prentice Hall.

Chang Lee, K., et al. (2005). KMPI: Measuring Knowledge Management Performance. *Information & Management*. Vol 42 Issue 3 pp. 469-482.

Choi, B., & Lee, H. (2003). An empirical Investigation of KM styles and Their effect on Corporate Performance. *Information & Management*, 40 pp.403-417.

Chung, M. & Gray, P. Special Section: Data Mining. *Journal of Management Information Systems.* Vol. 16 No. 1, Summer, 1999 pp. 11–16.

CNN News. (September 2, 2005a). *The big disconnect on New Orleans. The official version; then there's the in-the-trenches version*.

CNN News. (September 2, 2005b). *New Orleans mayor lashes out at fedsNagin: "They are spinning and people are dying."*

CNN News. (September 2, 2005c). *Bush faults recovery efforts as "not enough."*

Courtney, J. (2001). Decision Making and Knowledge Management in Inquiring Organizations: Toward a New Decision-Making Paradigm for DSS. *Decision Support Systems Special Issue on Knowledge Management*, 31 p.17-38.

Drucker, P. (1993). *Post-Capitalist Society*. New York, Harper Collins.

Dwivedi, A., Bali, R.K., James, A.E. and Naguib, R.N.G. (2001a). Telehealth Systems, Considering Knowledge Management and ICT Issues. *Proc of the IEEE-EMBC 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBS), [CD-ROM], Istanbul, Turkey.

Dwivedi, A., Bali, R.K., James, A.E. and Naguib, R.N.G. (2001b). Workflow Management Systems, the Healthcare Technology of the Future? *Proc of the IEEE EMBC-2001 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (EMBS), [CD-ROM], Istanbul, Turkey.

Dwivedi, A., Bali, R.K., James, A.E. and Naguib, R.N.G. (2002a). The Efficacy of Using Object Oriented Technologies to build Collaborative Applications in Healthcare and Medical Information Systems., *Proc of the IEEE Canadian Conference on Electrical and Computer Engineering* (CCECE) 2002. Winnipeg, Canada, 2, 1188-1193.

Dwivedi A., Bali, R.K., James, A.E., Naguib, R.N.G., and Johnston D. (2002). Merger of Knowledge Management and Information Technology in Healthcare: Opportunities and Challenges. *Proceedings of the 2002 IEEE Canadian Conference*, pp.1194-1199.

Fayyad, Piatetsky-Shapiro, Smyth. (1996). From Data Mining to Knowledge Discovery: An Overview, in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy. *Advances in Knowledge Discovery and Data Mining*, AAAI Press / The MIT Press, Menlo Park, CA .

Gupta, B., Iyer, L.S., and Aronson, J.E. (2000). Knowledge management: practices and challenges. *Industrial Management & Data Systems*. Vol. 100, No.1, pp.17-21.

Holsapple, C., and Joshi, K. (2002). Knowledge Manipulation Activities: results of a Delphi Study. *Information & Management*, 39 pp.477-419.

Huang, J.C. and Newell, S. (2003). Knowledge Integration Processes and Dynamics within the context of cross-functional projects. *International Journal of Project Management*. Vol.21, 167-176.

Kun, G. L. and Bray, A. D. (2002). Information Infrastructure Tools for Bioterrorism Preparedness. *IEEE Engineering in Medicine and Biology*. Vol. 21, No.5, pp. 69-85.

Liebowittz, J. (1999). *Knowledge Management Handbook.* CRC Press, London.

Maier, R. & Lehner, F. (2000). Perspectives on Knowledge Management Systems Theoretical Framework and Design of an Empirical Study. *In Proceedings of 8th European Conference on Information Systems* (ECIS).

Marincioni, F. (2001). A cross-culture analysis of Natural Disaster Response: the Northwest Italy floods of 1994 compared to the U.S. Midwest Floods 1993. *Intl. J. of Mass Emergencies and Disasters.* Vol. 19 No. 2 pp.209-236.

Massey, A., Montoya-Weiss, M. and O'Driscoll, T. (2002). Knowledge Management In Pursuit of Performance: Insights From Nortel Networks. *MIS Quarterly.* Vol. 26 No. 3 pp. 269-289.

Newell, S., Robertson, M., Scarbrough, H. and Swan, J. (2002). *Managing Knowledge Work.* Palgrave, New York.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organizational Science.* 5:14-37.

Nonaka, I. and Nishiguchi, T. (2001). *Knowledge Emergence,* Oxford University Press, Oxford.

Richards, C., (2004). What is MOOTW?. *Proc. 1st Natl. Conf. on Operations Other than War* (Ed. J. Riess), CRC Press, Boca Raton (Fl), in press. Abridged version can be accessed at http://www.d-n-i.net/richards/what_is_MOOTW.doc.

Schultze, U. and Leidner, D. (2002). Studying Knowledge Management In Information Systems Research: Discourses And Theoretical assumptions. *MIS Quarterly.* Vol. 26 No. 3 pp. 212-242.

Shapiro, C. and Verian, H. (1999). *Information Rules*, Harvard Business School Press, Boston.

Shin, M. (2004). A Framework for Evaluating Economies of Knowledge Management Systems. *Information & Management*. Vol. 42 Issue 1 pp.179-196.

von Lubitz, D., Carrasco, B., Fausone, C.A., Gabbrielli, F., Kirk, J., Lary, M.J., Levine, H. (2005). Bioterrorism, medical readiness, and distributed simulation training of first responders. *In Community and Response to Terrorism (Vol. II): A Community of Organizations: Responsibilities and Collaboration Among All Sectors of Society* (J. Johnson, M. Cwiek, G. Ludlow, EDS), Praeger (Greenwood Publ. Group) Westport, CT, pp. 267-312.

von Lubitz, D. and Wickramasinghe, N. (2005a). Networkcentric Healthcare: Outline of Entry Portal Concept. In press *International Journal of Electronic Business Management*.

von Lubitz, D. and Wickramasinghe, N. (2005b). Creating Germane Knowledge In Dynamic Environments. In press *International Journal Innovation and Learning.*

von Lubitz, D. and Wickramasinghe, N. (2006). Technology and Healthcare: The doctrine of Networkcentric Healthcare Operations. In press *International Journal of Electronic Healthcare.*

Wickramasinghe, N. (2005). Knowledge Creation: A meta-Framework. In press *International J. Innovation and Learning.*

Wickramasinghe, N. and Fadlalla, A. (2005). Realizing Knowledge Assets in the Medical Sciences with Data Mining In Wickramasinghe, et al. Eds. *Creating Knowledge-Based Healthcare Organizations*, IDEA Group Hershey.

Wickramasinghe, N. and Schaffer, J. (2005). Creating Knowledge Driven Healthcare Processes With The Intelligence Continuum. In press *Intl. J. Electronic Healthcare.*

Wickramasinghe, N. and von Lubitz, D. (2006). *Fundamentals Of The Knowledge-Based Enterprise.* In press IDEA Group Hershey.

Wilcox, L, (1997). *Knowledge-based Systems as an Integrating Process*. In Liebowitz, J. & L. Wilcox. Eds. Knowledge Management and its Integrative Elements, CRC Press, New York.

Zack, M. (1999). *Knowledge and Strategy*, Butterworth Heinemann, Boston.

KEY TERMS

Data Mining and KDD Process: Knowledge discovery in databases (KDD) (and more specifically data mining) approaches knowledge creation from a primarily technology driven perspective. In particular, the KDD process focuses on how data is transformed into knowledge by identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, et al., 1996). From an application perspective, data mining and KDD are often used interchangeably.

Explicit Knowledge: Or factual knowledge, that is, "know what", represents knowledge that is well established and documented.

Germane Knowledge: The sum total of all information *plus* the ability to implement it constructively and purposefully in the dynamic and unstable environment.

Knowledge Spiral: The process of transforming the form of knowledge, and, thus, increasing the extant knowledge base as well as the amount and utilization of the knowledge within the organization.

Pertinent Information: Information structured data, grouped into coherent categories that are easily perceptible and understood.

Preparedness: The availability (prepositioning) of all resources, both human and physical, necessary for the management of, or the consequences of, a specific event or event complex .

Readiness: The instantaneous ability to respond to a suddenly arising major crisis (e.g. sudden slow-down in the manufacturing parts supply chain) that is based on the instantaneously and locally available/ un-prepositioned and un-mobilized countermeasure resources.

Tacit Knowledge: Or experiential knowledge, that is, "know how" represents knowledge that is gained through experience and through doing.

Creating Software System Context Glossaries

Graciela D. S. Hadad
Universidad Nacional de La Matanza, Argentina & Universidad Nacional de La Plata, Argentina

Jorge H. Doorn
INTIA, Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina
Universidad Nacional de La Matanza, Argentina

Gladys N. Kaplan
Universidad Nacional de La Matanza, Argentina & Universidad Nacional de La Plata, Argentina

INTRODUCTION

Requirements engineering (RE) is the area of software engineering responsible for the elicitation and definition of the software system requirements. This task implies joining the knowledge of the services that a software system can and cannot provide with the knowledge of clients' and users' needs (Jackson, 1995; Katasonov & Sakkinen, 2005; Kotonya & Sommerville, 1998; Sommerville & Sawyer, 1997; Sutcliffe, Fickas, & Sohlberg, 2006; Uchitel, Chatley, Kramer, & Magee, 2006). Frequently, this activity is done by people with a software engineering bias. The underlying hypothesis of this choice is that users' needs are easier to understand than the software's possible behaviors. This is not always true; however, this is the metacontext in which most RE heuristics and methodologies have been developed. Understanding clients' and users' needs is far more complex than merely interviewing selected clients and user representatives, compiling all gathered information in one document. Defining how to put into service a complex software system within an organization requires envisioning how the business process of the organization will be in the future from both points of view: software organization and business organization. This is the key of the RE commitment: to imagine how the future business process will be. This RE commitment requires a good knowledge about how the business process actually is. Understanding the software system's preexistent context basically means understanding the clients' and users' culture. In other words, this part of the RE is a learning process.

BACKGROUND

The importance language has in any culture should be noticed. Language is an organized system of speech by which people communicate with each other with mutual comprehension. Also, it is very important to note that by the words it contains and the concepts it can formulate, language is said to determine the attitudes, understandings, and responses in any society. Language, therefore, may be both a cause and a symbol of cultural differentiation (Fishman, 1999; Hall, Hawkey, Kenny, & Storer, 1986). Language reflects environment and technology: Arabic has 80 words for camels, while Japanese has more than 20 words for rice and Inuit has more than 20 words for snow and ice (Nettle & Romaine, 2000). Clients and users have several special words that they use when discussing their activities. The requirements engineer must pick and understand as many of these words as possible as a first step in understanding clients' and users' culture.

Glossaries have been used in software engineering with different purposes, such as data dictionaries in early database books (Codd, 1982) to document entities, attributes, relations, types, and services of databases. Thus, it provides a common understanding of all the system names to the developer team and later to the maintenance team. However, data dictionaries are also an important component of structured analysis, data recording, data storage, and the details of processes (Gane & Sarson, 1982; Senn, 1989), though authors like Gane and Sarson suggest that the real name for them should be *project guide* instead of *data dictionary*. These data dictionaries are created during analysis and also used during system design. They satisfy five objectives: to manage details, communicate common meanings, document system characteristics, help the analysis of details and changes, and locate errors and omissions of the system.

In this article an RE process beginning with the construction of a language-extended lexicon (LEL) as its first activity (Leite, Hadad, Doorn, & Kaplan, 2000) is addressed, and the structure and creation of this LEL is described.

GLOSSARY CREATION

The word *dictionary* was coined by Henry Cockeran in 1623, but the first known dictionaries belong to the seventh century B.C., and they contain the most important data of

the Mesopotamian culture (MSN Microsoft Corporation, 2006). The first dictionaries were catalogues of unusual, difficult, or confusing words and phrases since the common vocabulary was considered to have no need of an explanation or a definition. The oldest glossary comes from the second century A.D. and contains technical Greek words used by Hypocrites. It was in later centuries that a catalogue of all the words of a language was built: for Arabic. So, the origin of glossaries and dictionaries was to give definitions of words and phrases of a particular domain; then, they were extended to an entire language in lexical dictionaries. Nowadays, there are different types of dictionaries covering different necessities, like dictionaries of synonyms and antonyms, dictionaries of idiomatic usage, etymology dictionaries, encyclopedic dictionaries, bilingual dictionaries, glossaries in textbooks, dictionaries of ideologies, slang dictionaries, and dictionaries of neologisms, among many others.

Most relevant or peculiar words or phrases (named LEL symbols) of the universe of discourse (UofD) are included in the LEL. Every symbol is identified by its name (including synonyms) and two descriptions: notion and behavioral response. The notion contains sentences defining the symbol and the behavioral response reflects how it influences the UofD. Figure 1 depicts the model used to represent LEL symbols.

The LEL is created by filling the blanks in the LEL model (see Figure 1) using information obtained from the application domain. Intuition, supported with a good understanding of the LEL model, may be used to create the lexicon. Upon the experience and the skill of the authors, this may or may not lead to a well-conceived document. If so, apparently there is no need for heuristics. On the contrary, heuristics are needed, first to allow everyone to complete the process successfully and second to avoid weaknesses usually present in apparently good-quality LELs. Those weaknesses range from missing relevant symbols to the unnecessary insertion of some others, and the inclusion of excess of details in the symbol descriptions or the lack of them.

The lexicon creation process, depicted in Figure 2 using an SADT[1] model (Ross & Schoman, 1977), consists of five independent activities: (a) plan, (b) collect, (c) describe, (d) verify, and (e) validate.

As seen in Figure 2, the process shows a main stream composed of three tasks: plan, collect, and describe. There is a well-established feedback when the verification and validation activities take place. After verifying the LEL, the process returns to the *describe* activity, where corrections are made based on a DEO list.[2] After the *validate* activity, the process returns to the *collect* activity and/or the *describe* activity, depending on the validation DEO list, in order to make any necessary corrections. For easy reading, the SADT model does not show all the backtracking steps that may occur during the construction process. For instance, while describing a symbol, a wrongly assigned type may be discovered, thus a back step occurs in order to reclassify it (within the *collect* activity). Another example of going backward in the process could appear when a new term is identified while describing another. That is, the strategy is not at all a linear one. It is an iterative process where feedback is a constant mechanism. In addition to this continuous feedback, the main stream does not fully follow a cascade model since in practice its three main tasks may partially overlap. For instance, a symbol may

Figure 1. Language-extended lexicon model

LEL: It is the representation of the symbols in the application domain language.
Syntax: $\{Symbol\}_1^N$

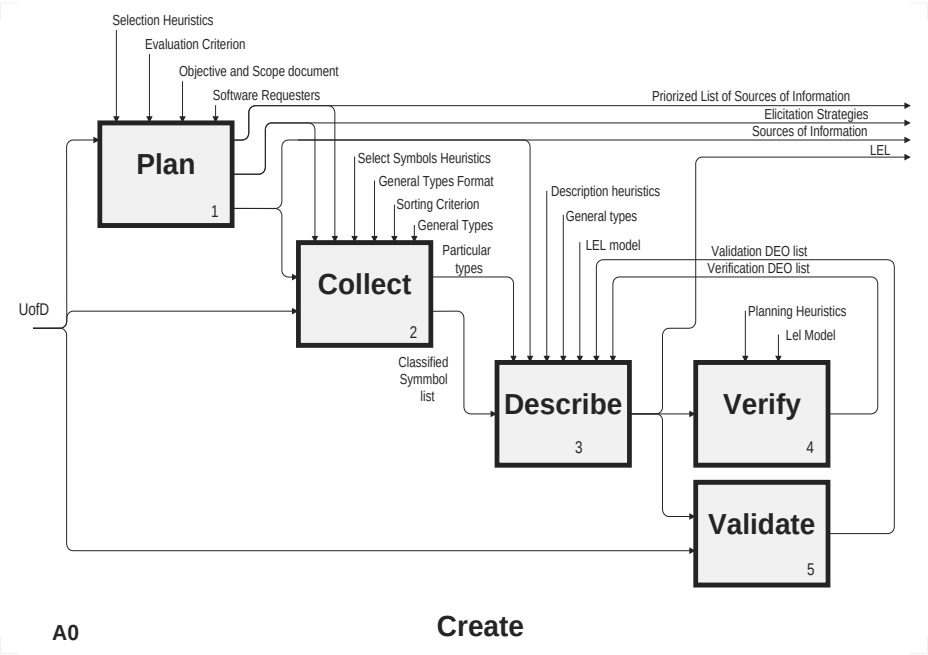
Symbol: It is an entry of the lexicon that has a special meaning in the application domain.
Syntax: $\{Name\}_1^N + \{Notion\}_1^N + \{Behavioral\ Response\}_1^N$

Name: This is the identification of the symbol. Having more than one name represents synonyms.
Syntax: Word | Phrase

Notion: It is the denotation of the symbol.
Syntax: Sentence

Behavioral Response: It is the connotation of the symbol.
Syntax: Sentence

Figure 2. SADT of the LEL creation process



be fully described while new sources of information should be identified to classify or describe others.

Plan. The *plan* activity basically consists of identifying the sources of information, evaluating them, and finally selecting the strategies to elicit symbols. To identify the sources of information, it is necessary to define the context where the RE process will take place. A mandatory source of information is the document of system objectives and scope (if it was written), or the requesters of the software system. The most reliable sources of information are documents and people, but some other relevant sources could be books about related themes, other clients' systems, and other systems available in the market. A source of information could be seen from several perspectives. One of the most salient perspectives is that of effectiveness, which classifies the information as either actual or formal. Formal information is about what should occur, but is not necessary to be put into practice or to be updated; actual information involves current practices or states, that is, what is actually in use. Accessing sources of information biased toward the formal point of view will create the important risk of developing a software system unable to deal with what actually happens in the business. On the other hand, ignoring formal sources of information does not allow the use of the software system as a tool for business process improvement. At this point, balancing the actual and formal points of view is almost impossible, but the requirements engineer must at least understand both.

Collect. The *collect* activity starts creating a candidate symbol list for accessing the sources of information by means of unstructured interviews, reading documents, or eliciting techniques appropriate for each source of information. The most important rules for choosing symbols are as follows.

- Pick exclusively words or phrases belonging to the application domain.
- Select words or phrases frequently used or highly repeated in documents.
- Select words or phrases meaningful in the application domain.
- Exclude too obvious words or phrases.
- Identify the full name no matter how long it is.
- An abbreviation or a partial name may be a synonym of a symbol with a long name.

Once the candidate list of symbols is available, every entry should be assigned to a class. In most cases, the basic classes of subject, verb, object, and state are useful. Subjects are active entities, such as persons, organizations, or software

systems. Objects are passive entities to which actions are applied. Verbs are entries representing actions that happen in the application domain, and states are conditions of a group of subjects, objects, or verbs.

Describe. Describing symbols defines their notions and behavioral responses based on the LEL model and the class to which they belong. In order to describe the symbols, the requirements engineers may use previous elicited knowledge, though often they should go back to the UofD to collect more information. In this case, it is recommended that they conduct structured interviews in order to ask clients and users about the meaning of the symbols. Nevertheless, other sources of information may well be used.

Below, some rules for describing symbols in the lexicon are itemised (description heuristics).

- A symbol must have at least one name, one notion, and one behavioral response.
- Every name of the symbol must be the one used in the application domain.
- Symbols used as synonyms in the application domain must share one entry in the LEL
- Symbols having a regular meaning must contain only the application domain sense.
- Notion and behavioral response must be described using simple and direct sentences.
- Each sentence should express only one idea.
- Each sentence should contain only one verb.
- Each sentence should make it easy to identify the perspective (formal or actual).
- If two symbols share a characteristic, it should be repeated in both entries.
- Every notion and behavioral response must have at least one reference to other symbols.
- References to other symbols should be enhanced (underlining, bolding, or any other way).
- Every symbol must be referenced at least by another symbol.
- A symbol's full name should be used when referenced by other symbols.

Connections among LEL entries should be stressed as much as possible using references to other symbols and reducing the use of vocabulary from outside the LEL.

Verify. Nowadays, inspections have been applied to requirements documents with great success (Leite, Hadad, Doorn, & Kaplan, 2005). Although the verification of the lexicon can be made by several techniques, an inspection variant based on Fagan's (1976) original proposal has been used. This technique provides specific heuristics to detect defects called discrepancies, errors, or omissions (DEOs) (Kaplan, Hadad, Doorn, & Leite, 2000). Each step in the heuristics is based on a defect-oriented form designed for a given type of defect and is accompanied with guides about how to fill the form and how to analyse what it is written in the LEL in order to maximise the finding of the defects.

Validate. While identifying and describing symbols, some degree of validation takes place. Later, a more structured validation activity is carried out allowing the requirements engineer to correct, ratify, or increase the knowledge about the application domain vocabulary. This usually consists of structured interviews or meetings with clients and users at their workplaces. The description of each symbol may be read to the interviewers who confirm, correct, make observations, or add missing information. Sometimes, instead of reading symbol descriptions during the interview, the engineer could give the interviewer a copy of the LEL in advance. Summarizing, the validation activity aims basically to do the following.

- Check notions and behavioral responses of symbols already described
- Ratify the definition of symbols
- Identify new symbols and synonyms

The validation activity generates a DEO list similar to the one produced at verification. It is then sent backward to the *collect* step and/or to the *describe* step to do the necessary corrections. Sometimes the feedback from validation may require identifying new sources of information.

FUTURE TRENDS

Even though the LEL has been used in many studies and practical application by researchers and practitioners, several aspects remain unknown and some questions remain unanswered. For example, there are no stopping rules defined. Different requirements engineers collect different symbols and different numbers of symbols; however, there is no known criterion about how to evaluate such differences.

Almost any LEL contains terms related by hypernym and hyponym relationships (Ureña López, García Vega, & Martínez Santiago, 2001). This symbol hierarchy can be observed among subjects, objects, and verbs. No benefits from this knowledge have been outlined yet.

CONCLUSION

Glossaries may be used in several stages during the software development process; however, their use during the RE phase is very convenient. Their use eases the understanding of the clients' and users' culture and allows creating all RE documents using their vocabulary.

Other documents produced in later phases of the software development process may also take advantage of the use of LEL.

Creating well-conceived system software context glossaries requires a good understanding of the model and following the heuristics and guidelines developed for this purpose.

Verification and validation of the glossary are key practices in the whole process since they will find out most of the existent defects, improving its quality.

REFERENCES

Codd, E. (1982). Relational databases: A practical foundation for productivity. *Communications of the ACM.*

Fagan, M. E. (1976). Design and code inspections to reduce errors in program development. *IBM Systems Journal, 15*(3), 182-211.

Fishman, J. (1999). *Handbook of language and ethnic identity.* Oxford, United Kingdom: Oxford University press.

Gane, C., & Sarson, T. (1982). *Structured system analysis: Tools and techniques.* McDonnell Douglas.

Hall, D., Hawkey, R., Kenny, B., & Storer, G. (1986). Patterns of thought in scientific writing: A course in information structuring for engineering students. *English for Specific Purposes, 5*(2), 147-160.

Jackson, M. (1995). *Software requirements & specifications: A lexicon of practice, principles and prejudices.* Addison Wesley, ACM Press.

Kaplan, G. N., Hadad, G. D. S., Doorn, J. H., & Leite, J. C. S. P. (2000). Inspección del léxico extendido del lenguaje. In *Proceedings of WER'00: III Workshop de Engenharia de Requisitos* (pp. 70-91).

Katasonov, A., & Sakkinen, M. (2005). Requirements quality control: A unifying framework. *Requirements Engineering Journal, 11*(1), 42–57.

Kotonya, G., & Sommerville, I. (1998). *Requirements engineering: Processes and techniques.* John Wiley & Sons.

Leite, J. C. S. P., Hadad, G. D. S., Doorn, J. H., & Kaplan, G. N. (2000). A scenario construction process. *Requirements Engineering Journal, 5*(1), 38-61.

Leite, J. C. S. P., Hadad, G. D. S., Doorn, J. H., & Kaplan, G. N. (2005). Scenario inspections. Requirements Engineering Journal, 10(1), 1, 21.

MSN Microsoft Corporation. (2006). *Encarta Encyclopedia 06.* Retrieved from http://encarta.msn.com/encyclopedia/761573731/Dictionary.html

Nettle, D., & Romaine, S. (2000). *Vanishing voices.* Oxford, United Kingdom: Oxford University Press.

Ross, D., & Schoman, A. (1977). Structured analysis for requirements definition. *IEEE Transactions on Software Engineering, 3*(1), 6-15.

Senn, J. A. (1989). *Analysis & design of information systems* (2nd ed.). McGraw-Hill Inc.

Sommerville, I., & Sawyer, P. (1997). *Requirements engineering: A good practice guide.* John Wiley & Sons.

Sutcliffe, A., Fickas, S., & Sohlberg, M. (2006). PC-RE: A method for personal and contextual requirements engineering with some experience. *Requirements Engineering Journal, 11*(3), 157-173.

Uchitel, S., Chatley, R., Kramer, J., & Magee, J. (2006). Goal and scenario validation: A fluent combination. *Requirement Engineering Journal, 11*(2), 123-137.

Ureña López, L. A., García Vega, M., & Martínez Santiago, F. (2001). Explotando las relaciones léxicas y semánticas de WordNet en la resolución de la ambigüedad léxica. In *VII Simposio Internacional de Comunicación Social,* Cuba (pp. 414-418).

KEY TERMS

Elicitation: Elicitation is the activity of acquiring knowledge of a given kind during the requirements engineering process.

Heuristic: It is a set of guidelines to help people to use others' experience to improve performance in a given task.

Language-Extended Lexicon: An LEL is a semiformal model holding the most relevant words or phrases of the language of the application domain carrying special meaning.

Requirements Engineering: It is an area of the software engineering that is responsible for acquiring and defining needs of the software system.

Software Engineering: It is the computer science discipline concerned with creating and maintaining software applications by applying technologies and practices from computer science, project management, engineering, application domains, and other fields.

Sources of Information: These include documents, key people, books, and so forth that can provide useful information about the matter being studied.

Validation: Validation is the activity of contrasting a model with the actual world. It should answer the question "Are we creating the right model?"

Verification: It is the activity of checking for the consistency of different parts of a model or different models among them. It should answer the question "Are we creating the model right?"

ENDNOTES

[1] The following is the notation of SADT: Boxes represent activities, left-pointing arrows represent input required by the activity, down-pointing arrows represent controls, up-pointing arrows represent mechanisms, and right-pointing arrows represent output from the activity.

[2] A DEO list contains the discrepancies, errors, and omissions discovered during the verification or validation activities, where suggestions for corrections are included.

Creating Superior Knowledge Discovery Solutions

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

INTRODUCTION

The information age has made information communication technology (ICT) a necessity for conducting business. This in turn has led to the exponential increase in the electronic capture of data and its storage in vast data warehouses. In order to respond quickly to fast changing markets, organizations must maximize these raw data and information resources. Specifically, they need to transform them into germane knowledge to aid superior decision-making (Wickramasinghe & von Lubitz, 2006). To do this effectively not only involves the analysis of the data and information but also requires the use of sophisticated tools to enable such analyses to occur. Knowledge discovery technologies represent a spectrum of new technologies that facilitate the analysis of data to find relationships from the data to finding reasons behind observable patterns (i.e., transform the data into relevant information and germane knowledge). Such new discoveries can have a profound impact on decision making in general and the designing of business strategies. With the massive increase in data being collected and the demands of a new breed of intelligent applications like customer relationship management, demand planning, and predictive forecasting, these knowledge discovery technologies are becoming competitive necessities for providing a high performance and feature rich intelligent application servers for intelligent enterprises.

Knowledge management (KM) tools and technologies are the systems that integrate various legacy systems, databases, ERP systems, and data warehouse to help facilitate an organization's knowledge discovery process. Integrating all of these with advanced decision support and online real time events enables an organization to understand customers better and devise business strategies accordingly. Creating a competitive edge is the goal of all organizations employing knowledge discovery for decision support (Thorne & Smith, 2000).

The following provides a synopsis of the major tools and critical considerations required to enable an organization to successfully effect appropriate knowledge sharing, knowledge distribution, knowledge creation, as well as knowledge capture and codification processes and hence embrace effective knowledge management (KM) techniques and advanced knowledge discovery.

BACKGROUND

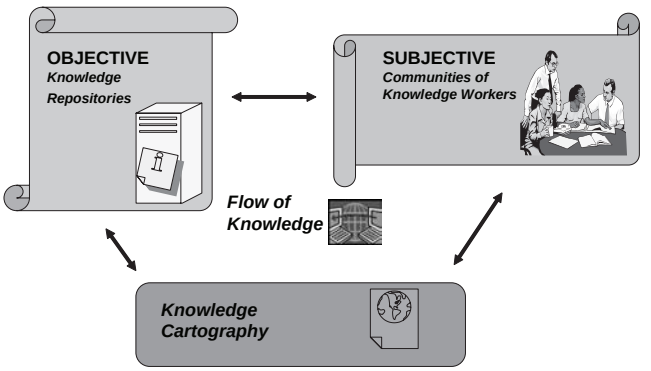
A necessary but not sufficient consideration to facilitate the generation of superior knowledge discovery solutions is the establishment of a sound KM infrastructure (Wickramasinghe et al., 2006). The KM infrastructure, in terms of tools and technologies, (hardware as well as software) should be established so that knowledge can be created from any new event or activity, which in turn will ensure that the extant knowledge base continuously grows (Wickramasinghe, Fadlalla, Geisler, & Schaffer, 2003; Wickramasinghe & Bali, 2006). The entire new know-how or new knowledge can only be created for exchange if the KM infrastructure is established effectively. Critical components of such a KM infrastructure include a repository of knowledge, and networks to distribute the knowledge to the members of organization and a facilitator system for the creation of new knowledge. Such a knowledge-based infrastructure will foster the creation of knowledge, and provide an integrated system to share and diffuse the knowledge in the organization (Srikantaiah & Koenig, 2000).

KNOWLEDGE ARCHITECTURE

Architecture, specifically the information technology architecture is an integrated set of technical choices used to guide an organization in satisfying its business needs (Weil & Broadbent, 1998). Underlying the knowledge architecture (Wickramasinghe, 2003; Wickramasinghe, 2005; refer to Figure 1) is the recognition of the binary nature of knowledge; namely its objective and subjective components. What we realize when we analyze the knowledge architecture closely, is that knowledge is not a clearly defined, easily identifiable phenomenon, rather it has many forms which makes managing it even more challenging (Schultz & Leidner, 2002; Wickramasinghe, 2005).

The knowledge architecture depicted in Figure 1 recognizes the two different, yet key aspects of knowledge; namely, knowledge as an object and a subject. By doing so, it provides the blue prints for an all encompassing knowledge management system (KMS). The pivotal function underlined by the knowledge architecture is the flow of knowledge. The

Figure 1. The knowledge architecture (Adapted from Wickramasinghe & Mills, 2001)



flow of knowledge is fundamentally enabled (or not) by the knowledge management system.

KNOWLEDGE MANAGEMENT SYSTEMS

Given the importance of knowledge, systems are being developed and implemented in organizations that aim to facilitate the sharing and integration of knowledge (i.e., support and facilitate the flow of knowledge). Such systems are called knowledge management systems (KMS) as distinct from transaction processing systems (TPS), management information systems (MIS), decision support systems DSS), and executive information systems (EIS) (Alavi & Leidner, 1999). For example, Cap Gemini Ernst & Young, KPMG, and Acenture all have implemented KMS (Wickramsinghe, 2003). In fact, the large consulting companies were some of the first organizations to realize the benefits of knowledge management and plunge into the knowledge management abyss. These companies treat knowledge management with the same high priority as they do strategy formulation, an illustration of how important knowledge management is viewed in practice (Wickramasinghe, 2003). Essentially, these knowledge management systems use combinations of the following technologies: the Internet, intranets, extranets, browsers, data warehouses, data filters, data mining, client server, multimedia, groupware, and software agents to systematically facilitate and enable the capturing, storing, and dissemination of knowledge across the organization (Alavi et al., 1999; Davenport & Prusak, 1998; Kanter, 1999). Unlike, other types of information systems, knowledge management systems can vary dramatically across organizations. This is appropriate if we consider that each organization's intellectual assets, intangibles, and knowledge should be to a large extent unique and thus systems enabling their management should in fact differ.

KNOWLEDGE MANAGEMENT TOOLS AND TECHNIQUES

KM tools and techniques are defined by their social and community role in the organization in (1) the facilitation of knowledge sharing and socialization of knowledge (production of organizational knowledge); (2) the conversion of information into knowledge through easy access, opportunities of internalization and learning (supported by the right work environment and culture); (3) the conversion of tacit knowledge into "explicit knowledge" or information, for purposes of efficient and systematic storage, retrieval, wider sharing, and application. The most useful KM tools and techniques can be grouped as those that capture and codify knowledge and those that share and distribute knowledge (Duffy, 2000, 2001; Maier, 2001).

Capture and Codify Knowledge

There are various tools that can be used for capture and codify knowledge. These include databases, various types of artificial intelligence systems including expert systems, neural networks, fuzzy logic, genetic algorithms, and intelligent or software agents.

Databases

Databases store structured information and assist in the storing and sharing of knowledge. Knowledge can be acquired from the relationships that exist among different tables in a database. For example, the relationship that might exist between a customer table and a product table could show those products that are producing adequate margins, providing decision-makers with strategic marketing knowledge. Many different relations can exist and are only limited by the human imagination. These relational databases help

users to make knowledgeable decisions, which is a goal of knowledge management. Discrete, structured information still is managed best by a database management system. However, the quest for a universal user interface has led to the requirement for access to existing database information through a Web browser.

Case-Based Reasoning Applications

Case-based reasoning (CBR) applications combine narratives and knowledge codification to assist in problem solving. Descriptions and facts about processes and solutions to problems are recorded and categorized. When a problem is encountered, queries or searches point to the solution. CBR applications store limited knowledge from individuals who have encountered a problem and found the solution and are useful in transferring this knowledge to others.

Expert Systems

Expert systems represent the knowledge of experts and typically query and guide users during a decision making process. They focus on specific processes and typically lead the user, step by step, toward a solution. The level of knowledge required to operate these applications is usually not as high as for CBR applications. Expert systems have not been as successful as CBR in commercial applications but can still be used to teach knowledge management.

Using I-Net Agents: Creating Individual Views from Unstructured Content

The world of human communication and information has long been too voluminous and complex for any one individual to monitor and track. Agents and I-net standards are the building blocks that make individual customization of information possible in the unstructured environment of I-nets. Agents will begin to specialize and become much more than today's general purpose search engines and "push" technologies.

Two complimentary technologies have emerged that allow us to coordinate, communicate, and even organize information without rigid, one-size-fits-all structures. The first is the Internet/Web technologies that are referred as I-net technology and the second is the evolution of software agents. Together, these technologies are the new-age building blocks for robust information architectures, designed to help information consumers find what they are looking for in the way that they want to find it. The Web and software agents make it possible to build sophisticated, well performing information brokers designed to deliver content, from multiple sources, to each individual, in the individual's specific context and under the individual's own control. The software agents supported with I-net infrastructure can be highly effective tools for individualizing the organization and management of distributed information.

Systems to Share and Distribute Knowledge

Computer networks provide an effective medium for the communication and development of knowledge management. The Internet and organizational intranets are used as a basic infrastructure for knowledge management. Intranets are rapidly becoming the primary information infrastructure for enterprises. An intranet is basically a platform based on Internet principles accessible only to members of an organization/community. The intranet can provide the platform for a safe and secured information management system within the organization, help people to collaborate as of virtual teams, crossing boundaries of geography and time. While the Internet is an open-access platform, the intranet, however, is restricted to members of a community/organization through multi-layered security controls. The same platform, can be extended to an outer ring (e.g., dealer networks, registered customers, online members, etc.), with limited accessibility, as an extranet. The extranet can be a meaningful platform for knowledge generation and sharing, in building relationships, and in enhancing the quality and effectiveness of service/support. The systems that are used to share and distribute knowledge could include group collaboration systems, groupware, intranets, extranets, and Internet, office systems, word processing, desktop publishing, or Web publishing.

FUTURE TRENDS

Just implementing a knowledge management system does not make an organization a knowledge based business. For an organization to become a knowledge-based business, several aspects must be considered. An organization that values knowledge must integrate knowledge into its business strategy and sell it as a key part of its products and services. To do this requires a strong commitment to knowledge management directed from the top of the organization. Furthermore, it is necessary for specific people, process, and technology issues to be considered. First, the knowledge architecture should be designed that is suitable given the context of a particular organization in its industry as well as the activities, products or services it may provide. From the knowledge architecture, it is important to focus on the organization's structure and culture and key people issues. Do the structure and culture support a knowledge-sharing environment or perhaps a more team focussed, sharing culture needs to be fostered. In addition, strategies need to be adopted for continuous training and fostering

of knowledge workers. Then, it is necessary to consider the processes of generating, representing, accessing and transferring knowledge throughout the organization. This also requires and evaluation of the technologies required enabling this. Finally, a knowledge based business should also enable organizational learning to take place so that the knowledge that is captured is always updated and current, and the organization is continually improving and refining its product or service as well as enhancing its extant knowledge base.

CONCLUSION

In today's hyper competitive business environment, organizations that can make superior and timely decisions have a greater chance of succeeding. This can only be done effectively and efficiently through the transformation of an organization's data and information assets into germane knowledge and relevant information. By utilizing the spectrum of intelligent technologies as well as embracing the tools and techniques of KM, organizations can successfully effect this required transformation and thereby create appropriate knowledge discovery solutions. Hence, the incorporation of KM and judicious adoption of intelligent technologies becomes a competitive necessity for all organizations.

REFERENCES

Alavi, M., & Leidner, D. (1999). Knowledge management systems: Issues, challenges, and benefits. In T. Davenport & L. Prusak (Eds.), *Communications of the association for information systems* (Vol. 1, Paper #5). Working Knowledge. Boston: Harvard Business School Press.

Duffy, J. (2001). The tools and technologies needed for knowledge management. *Information Management Journal, 35*(1), 64-67.

Duffy, J. (2000). The KM technology infrastructure. *Information Management Journal, 34*(2), 62-66.

Kanter, J. (1999). Knowledge management practically speaking. *Information Systems Management*, Fall.

Maier, R. (2001). *Knowledge management systems*. Berlin: Springer.

Schultze, U., & Leidner, D. (2002). Studying knowledge management in information systems research: Discourses and theoretical assumptions. *MIS Quarterly, 26*(3), 212-242.

Srikantaiah, T. K., & Koenig, M. E. D. (2000). ASIS monograph series. *Information Today*.

Thorne, K., & Smith, M. (2000). Competitive advantage in world-class organizations. *Management Accounting, 78*(3), 22-26.

Weill, P., & Broadbent, M. (1998). *Leveraging the new infrastructure*. Cambridge: Harvard Business School Press.

Wickramasinghe, N. (2005). The phenomenon of duality: A key to facilitate the transition from knowledge management to wisdom for inquiring organizations. In Courtney et al. (Eds.), *Inquiring organizations*. Hershey, PA: Idea Group Publishing.

Wickramasinghe, N. (2003). Do we practise what we preach: Are knowledge management systems in practice truly reflective of knowledge management systems in theory? *Business Process Management Journal, 9*(3), 295-316.

Wickramasinghe, N., & Mills, G. (2001). MARS: the electronic medical record system, The core of the Kaiser galaxy. *International Journal Healthcare Technology Management, 3*(5/6), 406-423.

Wickramasinghe, N., & Bali, R. (2006). Creating order from chaos: Application of the intelligence continuum for emergency and disaster scenarios. Forthcoming in *Encyclopaedia of Information Science and Technology*, Hershey, PA: Idea Group Publishing.

Wickramasinghe, N., & von Lubitz, D. (2006). *Knowledge-based enterprise: Theories and Fundamentals*. Forthcoming Idea Group.

Wickramasinghe, N., Fadlalla, A., Geisler, E., & Schaffer, J. (2003). Knowledge management and data mining: Strategic imperatives for healthcare. In *Proceedings of the 3rd Hospital of the Future Conference.*

KEY TERMS

KM Infrastructure: The tools and technologies (the specific tools required to capture and codify organizational knowledge, specific tools required to share and distribute organizational knowledge) that are required to support and facilitate KM in the organization. KM tools and technologies are the systems that integrate various legacy systems, databases, ERP systems, and data warehouse to help organizations to create, and use KM systems in the organization.

Knowledge: Knowledge is more comprehensive than data or information. It is a mix of experience, values, contextual information, expert insights, and grounded intuition that actively enables performance, problem solving decision-making, learning, and teaching.

Knowledge Architecture: The blue prints of subjective and objective knowledge, its flows and cartography of knowledge within the organization.

Knowledge as Object: This is when knowledge is conceptualized in the Lokean/Leibnitzian perspective and used to create efficient and effective solutions.

Knowledge as Subject: This is when knowledge is conceptualized in the Hegelian/Kantian perspective and used to create shared meanings and support sense making.

Knowledge Assets: The knowledge regarding markets, products, technologies, processes, and organizations, that a business owns or needs to own and which enable its business processes to generate profits, add value, etc.

Knowledge-Based Enterprises: Knowledge-based enterprises are those enterprises who derive the most value—from intellectual rather than physical assets. Knowledge-based enterprise is a firm that is fully embracing knowledge management and committed to fostering continuous learning.

Knowledge Management (KM): KM is the process through which organizations generate value from their intellectual and knowledge-based assets. Most often, generating value from such assets involves sharing them among employees, departments and even with other companies in an effort to devise best practices. KM is newly emerging, interdisciplinary business approach that involves utilizing people, processes, and technologies to create, store and transfer knowledge.

ENDNOTE

[1] This entry is a revised version of "An overview of knowledge discovery solutions for intelligent enterprises" By N. Wickramasinghe and S. Sharma Published By Idea Group in the previous edition of this encyclopaedia.

Credit Risk Assessment and Data Mining

André Carlos Ponce de Leon Ferreira de Carvalho
Universidade de São Paolo, Brazil

João Manuel Portela Gama
Universidade do Porto, Portugal

Teresa Bernarda Ludermir
Universidade Federal de Pernambuco, Brazil

INTRODUCTION

The widespread use of databases and the fast increase of the volume of data they store are creating a problem and a new opportunity for credit companies. These companies are realizing the necessity of making an efficient use of the information stored in their databases, extracting useful knowledge to support their decision-making process.

Nowadays, knowledge is the most valuable asset a company or nation may have. Several companies are investing large sums of money in the development of new computational tools able to extract meaningful knowledge from large volumes of data collected over many years. Among such companies, companies working with credit risk analysis have invested heavily in sophisticated computational tools to perform efficient data mining in their databases.

The behavior of the financial market is affected by a large number of political, economic, and psychological factors, which are correlated and interact among themselves in a complex way. The majority of these relations seems to be probabilistic and non-linear. Thus, these relations are hard to express through deterministic rules.

Simon (1960) classifies the financial management decisions in a continuous interval, whose limits are non-structure and highly structured. The highly structured decisions are those where the processes necessary for the achievement of a good solution are known beforehand and several computational tools to support the decisions are available. For non-structured decisions, only the managers' intuition and experience are used. Specialists may support these managers, but the final decisions involve a substantial amount of subjective elements. Highly non-structured problems are not easily adapted to the computer-based conventional analysis methods or decision support systems (Hawley, Johnson, & Raina, 1996).

BACKGROUND

The extraction of useful knowledge from large databases is named *knowledge discovery in databases* (KDD). KDD is a very demanding task and requires the use of sophisticated computing techniques (Brachman & Anand, 1996; Fayyad, Piatetsky-Shapiro, Amith, & Smyth, 1996). The recent advances in hardware and software make possible the development of new computing tools to support such a task. According to Fayyad et al. (1996), KDD comprises a sequence of stages, including:

- Understanding the application domain,
- Selection,
- Pre-processing,
- Transformation,
- Data mining, and
- Interpretation/evaluation.

It is also important to stress the difference between KDD and *data mining* (DM). While KDD denotes the whole process of knowledge discovery, DM is a component of this process. The DM stage is used as the extraction of patterns or models from observed data. KDD can be understood as a process that contains the previous listed steps. At the core of the knowledge discovery process, the DM step usually takes only a small part (estimated at 15-25%) of the overall effort (Brachman & Anand, 1996).

The KDD process begins with the understanding of the application domain, considering aspects such as the objectives of the application and the data sources. Next, a representative sample, selected according to statistical techniques, is removed from the database, preprocessed, and submitted to the methods and tools of the DM stage with the objective of finding patterns/models (knowledge) in the data. This knowledge is then evaluated regarding its quality and/or usefulness, so that it can be used to support a decision-making process.

Frequently, DM tools are applied to unstructured databases, where the data can, for example, be extracted from texts. In these situations, specific pre-processing techniques must be used in order to extract information in the attribute-value format from the original texts.

CREDIT RISK ASSESSMENT

Credit risk assessment is concerned with the evaluation of the profit and guaranty of a credit application. According to Dong (2006), the main approaches proposed in the literature for credit assessment can be divided into two groups: default models and credit scoring models. While default models assess the likelihood of default, credit scoring models assess the credit quality of the credit taker. This text covers credit scoring models.

A typical credit risk assessment database is composed of several thousands of credit applications. These credit applications can be related with either companies or people. Examples of personal credit applications are student loans, personal loans, credit card concessions, and home mortgages. Examples of company credits are loans, stocks, and bonds (Ross, Westerfield, & Jaffe, 1993).

Usually, the higher the value of the credit asked, the more rigorous is the credit risk assessment. Large financial institutions usually have whole departments dedicated to this problem.

The traditional approach employed by bank managers largely depends on their previous experience and does not follow the procedures defined by their institutions. Besides, several deficiencies in the dataset available for credit risk assessment, together with the high volume of data currently available, makes the manual analysis almost impossible. The treatment of these large databases overcomes the human capability of understanding and efficiently dealing with them, creating the need for a new generation of computational tools and techniques to perform automatic and intelligent analysis of large databases.

In 2004, the Basel Committee on Banking Supervision published a new capital measurement system, known as the New Basel Capital Accord, or Basel II, which implements a new credit risk assessment framework that supports the estimation of the minimum regulatory capital that should be allocated for the compensation of possible default loans or obligations (Basel, 2004; Van Gestel et al., 2006). The Basel Committee was created in 1974 by the central-bank of 10 countries. In 1988, the committee introduced a capital measurement system commonly referred to as the Basel Capital Accord. The first accord established general guidelines for credit risk assessment. The new accord, Basel II, stimulates financial institutions to adopt customized rating risk systems based on their credit transaction databases. As a consequence, DM techniques assume a very important role in credit risk assessment. They will allow the replacement of general risk assessment by careful analysis of each loan commitment.

Credit analysis databases usually cover a huge number of transactions performed over several years. The analysis of these data may lead to a better understanding of the customer's profile, thus supporting the offer of new products or services. These data usually hold valuable information, for example, trends and patterns, which can be employed to improve credit assessment. The large amount makes its manual analysis an impossible task. In many cases, several related features need to be simultaneously considered in order to accurately model credit user behavior. This need for automatic extraction of useful knowledge from a large amount of data is widely recognized.

USING DATA MINING FOR CREDIT RISK ASSESSMENT

DM techniques are employed to discover strategic information hidden in large databases. Before they are explored, these databases are cleaned. Next, a representative set of samples is selected. Machine learning techniques are then applied to these selected samples. The use of data mining techniques on a credit risk analysis database allows the extraction of several relevant pieces of information regarding credit card transactions.

The data present in a database must be adequately prepared before data mining techniques can be applied to it. The main steps employed for data preparation are:

- Preprocessing of the data to the format specified by the algorithms to be used;
- Reduction of the number of samples/instances;
- Reduction of the number of features/attributes;
- Features construction, which is the combination of one or more attributes in order to transform irrelevant attributes to more significant attributes; and
- Noise elimination and treatment of missing values.

Once the data have been pre-processed, machine learning (ML) techniques can be employed to discover useful knowledge. The quality of a knowledge extraction technique can be evaluated by different measures, such as accuracy; comprehensibility; and new, useful knowledge.

The application of data mining techniques for credit risk analysis may provide important information that can improve the understanding of the current credit market and support the work of credit analysts (Carvalho, Braga, Rezende, Ludermir, & Martineli, 2002; Eberlein, Breckling, & Kokic, 2000; Horst, Padilha, Rocha, Rezende, & Carvalho, 1998; Lacerda, de Carvalho, Braga, & Ludermir, 2005; Dong, 2006; Huang, Hung, & Jiau, 2006).

Credit analysis can be seen as a pattern classification task that involves the evaluation of the reliability and profitability of a credit application. If the labels associated with the data, characterizing each customer as either a good or a bad customer, are available, customers can be classified into these two classes. This is a binary classification problem. However, a credit scoring system does not need to be restricted to two classes. Financial institutions may have different categories for the customers, according to the provision required, for example. These categories usually form a ranking, with the best customers close to the top. Current economic indexes and changes in the customer profile may change his or her position in the ranking. In this case, a multiclass classification system should be used, preferably one that allows ranking classification.

In credit risk assessment, different misclassifications have different costs. However most of the existing data mining algorithms assume that the goal is to minimize the number of misclassification errors. Whenever different errors have different costs, several authors (Breiman, Freidman, Olshen, & Stone,1984; Turney, 1995) propose to minimize the conditional risk, which is the expected cost of predicting that *x* belongs to $Class_i$. The conditional risk is defined as: $Risk(Class_i / x) = sum_j P(Class_i / x) * C(i,j)$, where C(i,j) is the cost of predicting class i when the true class is j.

Different ML techniques have been used for credit risk assessment datasets. One of the studies found in the literature (Horst et al., 1998) compares the performance of neural networks and decision trees for credit risk assessment. Alternative approaches for the induction of Bayesian classifiers applied to credit scoring are investigated (Baesens, Egmont-Petersen, Castelo, & Vanthienen, 2002). Dong (2006) studies the influence of the metric distance in the performance of a case-based reasoning system used for credit scoring. A model based on support vector machines following the Basel II accord is proposed in Van Gestel et al. (2006). Another work using support vector machines highlights the importance of each input attribute selection (Yu, Lai, & Wang, 2006). The problem of imbalanced distribution of examples into the classes and its influence on different ML techniques is investigated in Huang et al. (2006).

Several of the most recent approaches are based on hybrid intelligent systems (HISs). In Mendes Filho, de Carvalho, and Matias (1997), multi-layer perceptron neural networks are designed by genetic algorithms to credit assessment. Fuzzy and neurofuzzy approaches for credit scoring are compared in Hoffmann, Baesens, Martens, Put, and Vanthienen (2002). Lacerda et al. (2005) show how radial basis function neural networks tuned for credit assessment problems can be evolved by genetic algorithms. Two approaches for learning fuzzy rules for credit scoring by using evolutionary algorithms are investigated in Hoffmann, Baesens, Mues, Van Gestel, and Vanthienen (2007). In another hybrid approach, a classifier based on multi-criteria linear programming is combined with independent analysis (Li, Shi, Zhu, & Dai, 2006).

CRITICAL ISSUES OF CREDIT ANALYSIS IN DATA MINING

There are, of course, a few problems that can arise from the use of data mining techniques for credit risk analysis. The main advantages and problems associated with this technology are briefly presented in Table 1. Current Internet technology may allow the invasion of company databases, which may lead to the access of confidential data. Besides, the stored data may be made available to other companies or used without the knowledge or authorization of the customer. On the other

Table 1. A summary of critical issues of credit risk analysis data mining

Data Mining Extraction of useful information from large Databases	**Privacy & Confidentiality Agreements** Addressing individual right to privacy and the sharing of confidential information
Concept Drifting When statistical properties of the target concept chance over time.	**Training of Employees** Credit analysis should be trained with the new tools available and shown the benefits of the new technologys
Inadequate information Applicants may have supplied incorrect information, intentionally or not	**User Ignorance and Perceptions** Lack of adequate understanding of the Data Mining and it usefulnes
Maintaining and integrity of data Maintaining up-to-date and accurate information on the databases	**Security** Maintaining secure and safe systems and keeping unauthorized user access out

hand, the extraction of useful information from credit risk analysis databases should result in benefits not only to the companies, but also to the customers, with the offering of better and more convenient credit products.

FUTURE TRENDS

A key issue in credit risk assessment is the non-stationary properties of the problem: interests change over time. In this context, methods that take drift into account can improve the generalization capability of the learning algorithms by adaptation of decision models to the most recent data (Gama, Medas, Castillo, & Rodrigues, 2004).

Credit-related data is being produced at growing rates. If the process is not strictly stationary (as in most real-world applications), the target concept could gradually change over time. For example, classes' profiles for good and bad clients may rapidly change. The ability to incorporate this concept drift is a natural extension for future DM systems. In real-time credit assessment, large chunks of data will be collected over time. Future credit assessment DM tools will have to deal with possible modifications in the existing classes. A natural approach for this new situation is the use of adaptive learning algorithms, where incremental learning algorithms take into account such concept drift.

Future work in this area should also take into consideration security issues, support to ubiquitous computing, and a better integration with current databases technology.

CONCLUSION

Several companies perform data mining on personal data stored in their databases. This is particularly true for credit risk analysis databases, where customers' personal data can be employed to improve the quality of credit assessment and support the offer of new credit products. Data mining systems usually employ artificial intelligence and statistical techniques to acquire relevant, interesting, and new knowledge for most applications, including credit risk analysis. Although one cannot avoid the use of these sophisticated techniques to support credit assessment, great care should be taken to guarantee privacy and that personal rights are not violated when working with private information.

A financial problem that shares several similarities with credit risk assessment is bankruptcy prediction. In the movement towards economic blocks, the growing globalization demands robust and reliable systems for banks' bankruptcy forecasting. This demand comes from different sources, like managers, investors, and government organizations. DM has also been successfully used for bankruptcy prediction, as can be seen in Martineli, Diniz, de Carvalho, Rezende, and Matias (1999), Atiya (2001), and Chakraborty and Sharma (2007).

REFERENCES

Atiya, A. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *IEEE Transactions on Neural Networks, 12*(4), 929-935.

Basel Committee on Banking Supervision. (2004). *International convergence of capital measurement and capital standards*. BIS.

Baesens, B., Egmont-Petersen, M., Castelo, R., & Vanthienen, J. (2002). Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo Search. *Proceedings of the International Conference on Pattern Recognition* (vol. 3, pp. 49-52).

Brachman, R., & Anand, T. (1996). *The process of knowledge discovery in databases: A human-centered approach* (pp. 37-57). Cambridge, MA: AAAI Press/The MIT Press.

Breiman, L., Freidman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth.

Carvalho, A., Braga, A., & Ludermir, T. (2005). Card users' data mining. *Encyclopedia of Information Science and Technology,* (I), 603-605.

Carvalho, A., Braga, A., Rezende, S., Ludermir, T., & Martineli E. (2002). Understanding credit card users' behaviour: A data mining approach. In Sarker, R.A., H.A. Abbass, & C.S. Newton (Eds.), *Heuristic and optimization for knowledge discovery* (pp. 240-261). Hershey, PA: Idea Group.

Chakraborty, S., & Sharma, S.K. (2007). Prediction of corporate financial health by artificial neural network. *International Journal of Electronic Finance, 1*(4), 442-459.

Dong, Y. (2006). A case based reasoning system for evaluating customer credit. *Journal of Japan Industrial Management Association, 57*(2), 144-152.

Eberlein, E., Breckling, J., & Kokic P. (2000). A new framework for the evaluation of market and credit risk. In G. Bol, G. Nakhaeizadeh, & K.-H. Vollmer (Eds.), *Datamining and computational finance* (pp. 51-67). Berlin: Physica-Verlag.

Fayyad, U., Piatetsky-Shapiro, G., Amith, S., & Smyth, P. (1996). *From data mining to knowledge*

discovery: An overview (pp. 1-34). Cambridge, MA: AAAI Press/The MIT Press.

Gama, J., Medas, P., Castillo, G., & Rodrigues, P.P. (2004). Learning with drift detection. *Proceedings of the 17th Brazilian Symposium on Artificial Intelligence* (pp. 286-295).

Giudici, P. (2003). *Applied data mining: Statistical methods for business and industry.* New York: John Wiley & Sons.

Hawley, D., Johnson, J., & Raina, D. (1996). Artificial neural systems: A new tool for financial decision making. In R. Trippi & E. Turban (Eds.), *Neural networks in finance and investing* (revised ed., pp. 25-44). New York: Irwin.

Hoffmann, F., Baesens, B., Martens, J., Put, F., & Vanthienen, J. (2002). Comparing a genetic fuzzy and a neurofuzzy classifier for credit scoring. *International Journal of Intelligent Systems, 17*(11), 1067-1083.

Hoffmann, F., Baesens, B., Mues, C., Van Gestel, T., & Vanthienen, J. (2007). Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European Journal of Operational Research, 177*(1), 540-555.

Horst, P., Padilha, T., Rocha, C., Rezende, S., & de Carvalho, A. (1998, May). Knowledge acquisition using symbolic and connectionist algorithms for credit evaluation. *Proceedings of the IEEE World Congress on Computational Intelligence* (WCCI'98), Anchorage, AK.

Huang, Y.-M., Hung, C.-M., & Jiau H.C. (2006). Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. *Nonlinear Analysis: Real World Applications, 7*(4), 720-747.

Lacerda, E.G.M., de Carvalho, A.C.P.F., Braga, A.P., & Ludermir, T.B. (2005). Evolutionary radial basis functions for credit assessment. *Applied Intelligence, 22*(3), 167-181.

Li, A., Shi, Y., Zhu, M., & Dai, J. (2006). A data mining approach to classify credit cardholders' behavior. *Proceedings of the 6th IEEE International Conference on Data Mining* (ICDMW) (pp. 828-832).

Martineli, E., Diniz, H., de Carvalho, A.C.P.F., Rezende, S., & Matias, A. (1999). Bankruptcy prediction using connectionist and symbolic learning algorithms. In A. Mustafa (Ed.), *Computational finance 1999* (pp. 515-524). New York: The MIT Press.

Mendes Filho, E., de Carvalho, A., & Matias, A. (1997). Credit assessment using evolutionary MLP networks. Decision technologies for computational finance. *Proceedings of the 5th International Conference on Computational Finance* (pp. 365-371), London.

Ross, S., Westerfield, R., & Jaffe, J. (1993). *Corporate finance.* New York: Richard D. Irwin.

Simon, H. (1960). *The new science of management decision.* New York: Harper & Row.

Turney, P.D. (1995). Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research, 2,* 369-409.

Van Gestel, T., Baesens, B., Van Dijcke, P., Garcia, J., Suykens, J.A.K., & Vanthienen, J. (2006). A process model to develop an internal rating system: Sovereign credit ratings. *Decision Support Systems, 42*(2), 1131-1151.

Vieira, A.S., Ribeiro, B., Mukkamala, S., Neves, J.C., & Sung, A.H. (2004, August/September). On the performance of learning machines for bankruptcy detection. *Proceedings of the 2nd IEEE International Conference on Computational Cybernetics* (pp. 323-328), Vienna.

Yu, L., Lai, K.K., & Wang, S. (2006). Credit risk assessment with least squares fuzzy support vector machines. *Proceedings of the 6th IEEE International Conference on Data Mining* (pp. 823-827).

Zhang, D., & Zhou, L. (2004). Discovering golden nuggets: Data mining in financial application. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, 34*(4), 513-522.

KEY TERMS

Consumer Credit: A loan to an individual to purchase goods and/or services for personal, family, or household use.

Credit: Delivery of a value in exchange of a promise that this value will be paid back in the future.

Credit Scoring: A numerical method of determining an applicant's loan suitability based on various credit factors such as types of established credit, credit ratings, residential and occupational stability, and ability to pay back loan.

Data: The set of samples, facts, or cases in a data repository. As an example of a sample, consider the field values of a particular credit application in a bank database.

Data Mining: The process of extracting meaningful information from very large databases. One of the main steps of the KDD process.

KDD: Process of knowledge discovery in large databases.

Knowledge: Defined according to the domain, considering usefulness, originality, and understanding.

Machine Learning: Sub-area of artificial intelligence that includes techniques able to learn new concepts from a set of samples.

Critical Realism as an Underlying Philosophy for IS Research

Philip J. Dobson
Edith Cowan University, Australia

INTRODUCTION

Many recent articles from within the information systems (IS) arena present an old-fashioned view of realism. For example, Iivari, Hirschheim, and Klein (1998) saw classical realism as seeing "data as describing objective facts, information systems as consisting of technological structures ('hardware'), human beings as subject to causal laws (determinism), and organizations as relatively stable structures" (p. 172). Wilson (1999) saw the realist perspective as relying on "the availability of a set of formal constraints which have the characteristics of abstractness, generality, invariance across contexts."

Fitzgerald and Howcroft (1998) presented a realist ontology as one of the foundational elements of positivism in discussing the polarity between hard and soft approaches in IS. Realism is placed alongside positivist, objectivist, etic epistemologies and quantitative, confirmatory, deductive, laboratory-focussed and nomothetic methodologies. Such a traditional view of realism is perhaps justified within the IS arena, as it reflects the historical focus of its use, however, there now needs to be a greater recognition of the newer forms of realism—forms of realism that specifically address all of the positivist leanings emphasised by Fitzgerald and Howcroft (1998). A particular example of this newer form of realism is critical realism. This modern realist approach is primarily founded on the writings of the social sciences philosopher Bhaskar (1978, 1979, 1986, 1989, 1991). The usefulness of such an approach has recently been recognized in the IS arena by Dobson (2001) and Mingers (2002).

BACKGROUND

Bhaskar's brand of realism (referred to by Searle, 1995, as a form of external realism) argues that there exists a reality totally independent of our representations of it; the reality and the "representation of reality" operating in different domains—roughly a transitive epistemological dimension and an intransitive ontological dimension. For the realist, the most important driver for decisions on methodological approach will always be the intransitive dimension—the target being to unearth the real mechanisms and structures underlying perceived events. Critical realism acknowledges that observation is value laden, as Bhaskar pointed out in a recent interview:

...there is no conflict between seeing our scientific views as being about objectively given real worlds, and understanding our beliefs about them as subject to all kinds of historical and other determinations. (Norris, 1999)

The critical realist agrees that our knowledge of reality is a result of social conditioning and thus cannot be understood independently of the social actors involved in the knowledge derivation process. However, it takes issue with the belief that the reality is a product of this knowledge derivation process. The critical realist asserts that "real objects are subject to value laden observation"; the *reality* and the value-laden *observation of reality* operate in two different dimensions, one intransitive and relatively enduring and the other transitive and changing.

An important aspect of a critical realist approach is that it not only provides direction on the characteristics and behaviour of the underlying objects of enquiry, but it also provides direction as to how to examine these objects. The philosophy is presented as an underlabourer to social enquiry in that it can help with "clearing the ground a little...removing some of the rubbish that lies in the way of knowledge" (Locke, 1894, p. 14). This integral and important role for philosophy in the enquiry process can help to avoid many potentially false pathways and avenues.

For example, Bhaskar (1979) presented fundamental difficulties with the way that prediction and falsification have been used in the open systems evident within the social arena. For the critical realist, a major issue with social investigation is the inability to create closure—the aim of "experiment" in the natural sciences. Bhaskar (1979) argued that this inability implies that theory cannot be used in a predictive manner and can only play an explanatory role in social investigations, because:

…in the absence of spontaneously occurring, and given the impossibility of artificially creating, closed systems , the human sciences must confront the problem of the direct scientific study of phenomena that only manifest themselves in open systems—for which orthodox philosophy of science, with its tacit presupposition of closure, is literally useless. In particular it follows from this condition that criteria for

the rational appraisal and development of theories in the social sciences, which are denied (in principle) decisive test situations, cannot be predictive and so must be exclusively explanatory. (p. 27)

As Mingers (2002) suggested, such an argument has specific ramifications with respect to the use of statistical reasoning to predict future results. Bhaskar (1979) argued that the primary measure of the "goodness" of a theory is in its explanatory power. From Bhaskar's perspective, predictive use of theories is not possible in open social systems, and therefore, predictive power cannot be a measure of goodness. From this point of view, theory acts as primarily an explanatory tool for explaining events in hindsight.

Critical realism uses abductive or retroductive reasoning as its main focus. Positivist approaches are associated more with deductive or inductive reasoning. Deductive reasoning is the fundamental reasoning of mathematics, whereby some statement "*p*" leads to implications "*q*"—a movement from the general to the particular. For example, the general claim that "all crows are black" moves to the particular inference that the next one seen will be black. For the crows example, retroductive or abductive reasoning follows from an observation of numerous black crows to a theory as to a mechanism to explain why crows are disposed to be black. As Mingers (2002) described:

We take some unexplained phenomenon and propose hypothetical mechanisms that, if they existed, would generate or cause that which is to be explained. So, we move from experiences in the empirical domain to possible structures in the real domain. This does not of itself prove that the mechanism exists, and we may have competing explanations, so the next step is to work toward eliminating some explanations and supporting others. (p. 300)

Outhwaite (1987) similarly suggested that the critical realist method involves "the postulation of a possible [structure or] mechanism, the attempt to collect evidence for or against its existence and the elimination of possible alternatives" (p. 58). The realist agrees that we have a good explanation when (a) the postulated mechanism is capable of explaining the phenomenon, (b) we have good reason to believe in its existence, and (c) we cannot think of any equally good alternatives. Such an explanatory target suggests that philosophical considerations must play an important role in the critical realist method, because such an approach often requires transcending, or speculating, perhaps nonobservable mechanisms and structures to explain perceived happenings. Such initial proposition is transcendental or metaphysical in its focus, and as such, any explanation or discovery made is seen to be fallible and extendable as knowledge grows. As Wad (2001) argued:

If we take explanation to be the core purpose of science, critical realism seems to emphasise thinking instead of experiencing, and especially the process of abstraction from the domains of the actual and the empirical world to the transfactual mechanisms of the real world. (p. 2).

This type of thinking is called transcendental by Bhaskar, in that it gives an important role to the crossing of the divide between the empirical and speculative activities of scientific work. As Wad pointed out, this is necessary because often the experienced world of events is not explainable in terms of the empirical facts but only by way of incorporating non-experienced mechanisms incorporated in objects that may be within or outside our domain of investigation.

RESEARCH IMPLICATIONS

Sayer (2000) contended: "Compared to positivism and interpretivism, critical realism endorses or is compatible with a relatively wide range of research methods, but it implies that the particular choices should depend on the nature of the object of study and what one wants to learn about it" (p. 19). As Mingers (2002) suggested, critical realism supports methodological pluralism in that it suggests that an external reality is open to multiple interpretations and understandings.

Yet, critical realism also has important things to say about the objects of enquiry in that it is an ontologically bold philosophy (Outhwaite, 1987, p. 34). It not only encompasses an external realism in its distinction between the world and our experience of it, but it also suggests a stratified ontology and a so-called depth realism in defining the objects that make up such a world. This concept suggests that reality is made up of three ontologically distinct realms: first, the empirical, that is experience; second, the actual, that is events (i.e., the actual objects of experience); and third, the transcendental, nonactual or deep, that is structures, mechanisms, and associated powers. This so-called depth realism proposes that "the world is composed not only of events and our experience or impression of them, but also of (irreducible) structures and mechanisms, powers and tendencies, etc. that, although not directly observable, nevertheless underlie actual events that we experience and govern or produce them" (Lawson, 1997, p. 8). Critical realism, in its use of retroduction, involves a movement from a surface phenomenon to a deeper causal thing; it involves the steady unearthing of deeper levels of structures and mechanisms.

The ontological complexity assumed by critical realism is, however, matched with a conservative epistemology heavily dependent on scientific argument. As Wad (2001) said, critical realism has little to say concerning practical advice:

One may get the feeling that critical realism develops a huge

superstructure of ontological and epistemological insights, but when it comes to practical research we are left with the usual methodological suspects, delivered by inductivists, positivist and empirical-analytical scientists. (p. 12)

As Stones (1996) argued, realist methodologies need to be able to account for the underlying ontological richness they implicitly assume. They also need to reflect the belief that any knowledge gains are typically provisional, fallible, incomplete, and extendable. Realist methodologies and writings, thus, must reflect a continual commitment to caution, scepticism, and reflexivity.

FUTURE TRENDS

Critical realism is becoming influential in a range of disciplines, including geography (Pratt, 1995; Yeung, 1997), economics (Lawson, 1997; Fleetwood, 1999), organization theory (Tsang & Kwan, 1999), accounting (Manicas, 1993), human geography (Sayer, 1985), nursing (Ryan & Porter,1996; Wainwright, 1997), logistics and network theory (Aastrup, 2002), and library science (Spasser, 2002). The application of critical realism within the IS field has been limited to date. Mutch (1995, 1999, 2002) has applied critical realist thinking in the examination of organizational use of information. In so doing, he commented on how difficult it is to apply such a wide-ranging and sweeping philosophical position to day-to-day research issues. Mingers (2001, 2002) examined the implications of a critical realist approach, particularly in its support for pluralist research. Dobson (2001, 2002) argued for a closer integration of philosophical matters within IS research and proposed a critical realist approach. Information systems are social systems, and it makes sense to apply a modern social philosophy such as critical realism to their examination.

CONCLUSION

In researching the social context within which information technology (IT) and IS operate, a modern social philosophy such as critical realism has considerable potential. It can provide useful insight into the type of (retroductive) questions that may be asked and also the means by which an examination can progress. The integrated nature of the philosophy encourages a consistency in research in that it recognizes the tripartite connections between ontology, methodology, and practical theory. As Archer (1995) argued:

The social ontology endorsed does play a powerful regulatory role vis-à-vis the explanatory methodology for the basic reason that it conceptualises social reality in certain terms. Thus identifying what there is to be explained and also ruling out explanations in terms of entities or properties which are deemed non-existent. Conversely, regulation is mutual, for what is held to exist cannot remain immune from what is really, actually or factually found to be the case. Such consistency is a general requirement and it usually requires two-way adjustment. (p. 17)

The required consistency between ontological and epistemological matters has led to the critical realist observation that many of the things that traditionally have been done in researching the social arena are actually inconsistent with the underlying nature of the social objects proposed (see, for example, Lawson, 1997, who presented some of the inconsistencies evident in traditional use of economic theory in the social arena). There are, however, few practical examples of the use of critical realism in the IS field and the obvious potential it has is not yet realized. While, as Wilson (1999) observed, the realist argument has shown a remarkable resilience and resourcefulness in the face of the "rise of relativism," and more practical examples of its application need to be developed. Carlsson (2003) examines IS evaluation from a critical realist perspective.

REFERENCES

Aastrup, J. (2002). *Networks producing intermodal transport*. Ph.D. dissertation, Copenhagen Business School.

Archer, M. (1995). *Realist social theory: The morphogenetic approach*. New York; London: Cambridge University Press.

Banville, C., & Landry, M. (1989). Can the field of MIS disciplined? *Communications of the ACM, 32*(1), pp. 48–60.

Bhaskar, R. (1978). *A realist theory of science*. Sussex: Harvester Press.

Bhaskar, R. (1979). *The possibility of naturalism*. Hemel Hempstead: Harvester Wheatsheaf.

Bhaskar, R. (1986). *Scientific realism and human emancipation*. London: Verso.

Bhaskar, R. (1989). *Reclaiming reality: A critical introduction to contemporary philosophy*. London: Verso.

Bhaskar, R. (1991). Philosophy and the idea of freedom. Oxford: Blackwell.

Carlsson, S.A. (2003). Advancing information systems evalution (research): A critical realist approach. *Electronic Journal of Information Systems Evalution*, (6)2, 11-20.

Dobson, P. (2001). The philosophy of critical realism—An opportunity for information systems research. *Information Systems Frontiers, (3)*2, 199-210

Dobson, P. (2002, January). Critical realism and IS research—Why bother with philosophy? *Information Research*. Retrieved from *http://InformationR.net/ir/*

Fitzgerald, B., & Howcroft, D. (1998). Towards dissolution of the IS research debate: From polarization to polarity. *Journal of Information Technology, 13*, 313–326.

Fleetwood, S. (Ed.). (1999). *Critical realism in economics: Development and debate*. London: Routledge.

Iivari, J., Hirschheim, R., & Klein, H. K. (1998). A paradigmatic analysis contrasting information systems development approaches and methodologies. *Information Systems Research, 9*(2), 164–193.

Kuhn, T. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: The University of Chicago Press.

Lawson, T. (1997). *Economics and reality*. London: Routledge.

Mingers, J. (2001). Combining IS research methods: Towards a pluralist methodology. *Information Systems Research, 12*(3), 240–259.

Mingers, J. (2002). Real-izing information systems: Critical realism as an underpinning philosophy for information systems. In *Proceedings of the 23rd International Conference on Information Systems* (pp. 295–303).

Norris, C. (1999). Bhaskar interview. *The Philosophers' Magazine, 8*, 34

Outhwaite, W. (1987). *New philosophies of social science: Realism, hermeneutics, and critical theory*. New York: St. Martin's Press.

Pratt, A. (1995). Putting critical realism to work: The practical implications for geographical research. *Progress in Human Geography, 19*(1), 61–74.

Ryan, S., & Porter, S. (1996). Breaking the boundaries between nursing and sociology: A critical realist ethnography of the theory–practice gap. *Journal of Advanced Nursing, 24*, 413–420.

Sayer, A. (1985). Realism in geography. In R. J. Johnston (Ed.), *The future of geography* (pp. 159–173). London: Methuen.

Sayer, A. (2000). *Realism and social science*. Thousand Oaks, CA: Sage.

Sayer, R. A. (1992). *Method in social science: A realist approach*. London: Routledge.

Searle, J. R. (1995). *The construction of social reality*. New York: Free Press.

Spasser, M. A. (2002). *Realist activity theory for digital library evaluation: Conceptual framework and case study. Computer Supported Cooperative Work* (Vol. 11, pp. 81–110). Dordrecht: Kluwer Academic Publishers.

Stones, R. (1996). *Sociological reasoning: Towards a past-modern sociology*. New York: Macmillan.

Tsang, E., & Kwan, K. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review, 24*(4), 759–780.

Wad, P. (2001). *Critical realism and comparative sociology*. Draft paper for the IACR conference, 17–19 August.

Wainwright, S. P. (1997). A new paradigm for nursing: The potential of realism. *Journal of Advanced Nursing, 26*, 1262–1271.

Wilson, F. (1999). Flogging a dead horse: The implications of epistemological relativism within information systems methodological practice. *European Journal of Information Systems, 8*(3), 161–169.

KEY TERMS

Author's Note: In philosophy, definitions become a basis for debate—they often reflect the area from which the author derives. Perhaps the following reflect realist origins.

Closed and Open Systems: A *closed system* is one restricted in such a way that laws have uniform effects. An *open system* is one that is not closed. Closed systems do not usually occur spontaneously in nature and generally require human intervention, such as in laboratory experiments (from www.raggedclaws.com/criticalrealism).

Critical Realism: The careful or critical application of the scientific approach to the social sciences.

Epistemology: The study of knowledge or how we come to know.

Ontology: The study of what exists.

Philosophy: "The critical examination of the grounds for fundamental beliefs and an analysis of the basic concepts employed in the expression of such beliefs" (*Encyclopaedia Britannica*, p. 388, *Micropedia*, Vol. 9, 1985) or the "rational, methodical, and systematic consideration of those topics that are of greatest concern to man" (*Macropedia*, Vol. 25, p. 742, 1985 edition).

Realism: The belief that there is a reality independent of our perceptions of it.

Transitive and Intransitive Dimensions: The *intransitive dimension* in the philosophy of science corresponds roughly to ontology, and the *transitive dimension* corresponds roughly to epistemology. Intransitive objects exist and act independently of our knowledge of them (except when we use our knowledge to intervene (see www.raggedclaws.com/criticalrealism).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 606-610, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Critical Realist Information Systems Research

Sven A. Carlsson
Lund University, Sweden

INTRODUCTION

The information systems (IS) field is dominated by positivistic research approaches and theories (Chen & Hirschheim, 2004). IS scholars have pointed out weaknesses in these approaches and theories and in response different strands of post-modern theories and constructivism have gained popularity—see, Lee, Liebenau, and DeGross (1997) and Trauth (2001). The approaches argued for include ethnography, constructivism, grounded theory, and theories like Giddens' structuration theory and Latour's actor-network theory. (We refer to these different research approaches and theories as "post-approaches" and "post-theories" when distinction is not required).

BACKGROUND

Although post-approaches and post-theories overcome some of the problems noted with positivistic approaches and theories, they have at least three major weaknesses and limitations. First, their fascination with the voices of those studied leads to IS research as mere reportages and local narratives which can lead to any narrative/reportage being as good as another narrative/reportage. Second, their focus on agency leads to ignoring the structural dimension—the agency/structure dimension is collapsed, leading to a flat treatment of the dimension. Third, their rejection of objectivist elements leads to problems when researching ICT-artifacts and ICT-based IS. For elaborate critique of post-approaches and post-theories, see Archer, Bhaskar, Collier, Lawson, and Norrie (1998).

An alternative to traditional positivistic models of social science as well as an alternative to post-approaches and post-theories is critical realism (CR). CR argues that social reality is not simply composed of agents' meanings, but that there exist structural factors influencing agents' lived experiences. CR starts from an ontology which identifies structures and mechanisms through which events and discourses are generated as being fundamental to the constitution of our natural and social reality. This article briefly presents CR and exemplifies how it can be used in IS research.

CRITICAL REALISM IN IS RESEARCH

CR has primarily been developed by Roy Bhaskar (1978, 1998) and can be seen as a specific form of realism. Good summaries of CR are available in Sayer (2000) and Archer *et al*. (1998) and key concepts and main developments are presented in Hartwig (2007). CR's manifesto is to recognize the reality of the natural order and the events and discourses of the social world. It holds that:

> *… we will only be able to understand—and so change—the social world if we identify the structures at work that generate those events and discourses … These structures are not spontaneously apparent in the observable pattern of events; they can only be identified through the practical and theoretical work of the social sciences.* (Bhaskar, 1989, p. 2)

Bhaskar (1978) outlines what he calls three domains: the *real*, the *actual*, and the *empirical*. The real domain consists of underlying structures and mechanisms, and relations; events and behavior; and experiences. The generative mechanisms, residing in the real domain, exist independently of but capable of producing patterns of events. Relations generate behaviors in the social world. The domain of the actual consists of these events and behaviors. Hence, the actual domain is the domain in which observed events or observed patterns of events occur. The domain of the empirical consists of what we experience; hence, it is the domain of experienced events. Bhaskar argues that:

> *…real structures exist independently of and are often out of phase with the actual patterns of events. Indeed it is only because of the latter we need to perform experiments and only because of the former that we can make sense of our performances of them. Similarly it can be shown to be a condition of the intelligibility of perception that events occur independently of experiences. And experiences are often (epistemically speaking) 'out of phase' with events—e.g. when they are misidentified. It is partly because of this possibility that the scientist needs a scientific education or training. Thus I [Bhaskar] will argue that what I call the domains of the real, the actual and the empirical are distinct.* (Bhaskar, 1978, p. 13)

CR also argues that the real world is ontologically stratified and differentiated. The real world consists of a plurality of structures and mechanisms that generate the events that occur.

CR has primarily been occupied with philosophical issues and fairly abstract discussions. In recent years attention has been paid to how to actually carry out research with CR as a philosophical underpinning—see Layder (1998), Robson (2002), Kazi (2003), and Pawson (2006). Gregor (2006) argues that five interrelated types of IS theory can be distinguished: (1) theory for analyzing, (2) theory for explaining, (3) theory for predicting, (4) theory for explaining and predicting, and (5) theory for design and action. The five types can be clustered into two main types: "traditional" natural/social research (first four types) and design science research (fifth type). This section briefly presents how CR can be used in the first four types of IS research and the next section addresses IS design science research based on CR.

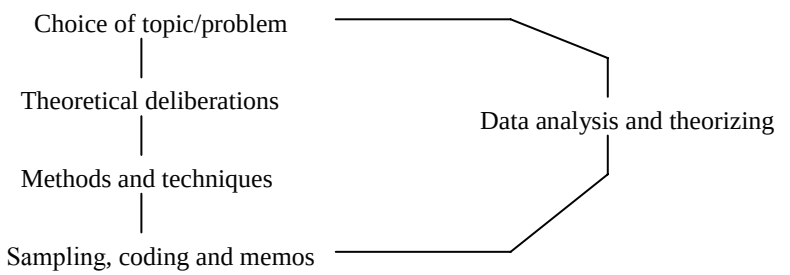
Bhaskar says that explanations (theories) are accomplished by the RRRE model of explanation comprising a four-phase process: (1) *Resolution* of a complex event into its components (causal analysis); (2) *Redescription* of component causes; (3) *Retrodiction* to possible (antecedent) causes of components via independently validated normic statements; and (4) *Elimination* of alternative possible causes of components." (Bhaskar, 1998). This is a rather abstract description of explanation (theory) development. Here we will instead use Layder's (1998) less abstract "adaptive theory." It is an approach for generating theory in conjunction with empirical research. It attempts to combine the use of pre-existing theory and theory generated from empirical data. Figure 1 depicts the different elements of the research process. There is not some necessary or fixed temporal sequence. Layder stresses that theorizing should be a continuous process accompanying the research at all stages. Concerning research design and methods, CR is supportive of: (1) the use of both quantitative and qualitative methods, (2) the use of extensive and intensive research design, and (3) the use of fixed and flexible research design.

To exemplify how CR and Layder's adaptive theory can be used in IS research addressing Gregor´s (2006) four first IS theory types, we will use a project on the use of executive information systems (EIS). The project was done together with Dorothy Leidner.[1] Here a new discussion of the research is carried out.

Layder's adaptive theory approach has eight overall parameters. One parameter says that adaptive theory "uses both inductive and deductive procedures for developing and elaborating theory." (Layder, 1998). The adaptive theory suggests the use of both forms of theory-generation within the same frame of reference and particularly within the same research project. We, based on previous EIS theories and Huber's (1990) propositions on the effects of advanced IT on organizational design, intelligence, and decision making, generated a number of hypotheses (a deductive procedure). These were empirically tested. From a CR perspective the purpose of this was to find patterns in the data that would be addressed in the intensive part of the study. We also used an inductive procedure. Although previous theories as well as the results from the extensive part of the project were fed into the intensive part, we primarily used an inductive approach to generate tentative explanations (theories) of EIS development and use from the data. The central mode of inference (explanation) in CR research is retroduction. It enables a researcher, using induction and deduction, to investigate the potential causal mechanisms and the conditions under which certain outcomes will or will not be realised. The inductive and deductive procedures led as to formulate explanations in terms of what mechanisms and contexts could lead (or not lead) to certain outcomes—outcomes being types of EIS use with their specific effects.

Another parameter says that adaptive theory "embraces both objectivism and subjectivism in terms of its ontological presuppositions" (Layder, 1998). The adaptive theory conceives the social world as including both subjective and objective aspects and mixtures of the two. In our study, one objective aspect was the ICT used in the different EIS and one subjective aspect was perceived effects of EIS use.

Figure 1. Elements of the research process (Layder, 1998, p. 29)



Two other parameters say that adaptive theory "assumes that the social world is complex, multi-faceted (layered) and densely compacted" and "focuses on the multifarious interconnections between human agency, social activities and social organization (structures and systems)" (Layder, 1998). In our study we focused the 'interconnections' between agency and structure. We addressed *self* (e.g., perceptions of EIS), *situated activity* (e.g., use of EIS in day-to-day work), *setting* (e.g. organizational structure and culture), and *context* (e.g., national culture and economic situation). Based on our data we hypothesized that national culture can affect (generate) how EIS are developed and used and how they are perceived. We also hypothesized that organizational 'strategy' and 'structure' as well as 'economic situation' can affect (generate) how EIS are developed and used and how they are perceived.

Our study and the results (theory) were influenced by, e.g. Huber's propositions, the 'theory' saying that EIS are systems for providing top-managers with critical information, and Quinns' competing values approach (Quinn et al., 2004). The latter theory was brought in to theorize around the data from the intensive (inductive) part of the study. Adaptive theorizing was ever present in the research process. In line with CR, we tried to go beneath the empirical to explain *why* we found what we found through hypothesizing the mechanisms that shape the actual and the events. Our study led to our argument that it is a misconception to think of EIS as systems that just provide top-managers with information. EIS are systems that support managerial cognition and behavior—providing information is only one of several means—as well as it can be one important means in organizational change. Based on our study, we "hypothesize" that "tentative" mechanisms are, for example, national culture, economic development, and organizational strategy and culture. We also hypothesized how the mechanisms together with different actors' decisions and actions, based on their desires, beliefs, and opportunities, lead to the development and use of different types of EIS. For example: (1) EIS use for personal productivity enhancement respectively EIS use for organizational change, and (2) EIS use for organizational change respectively EIS use for control and stability.

FUTURE TRENDS

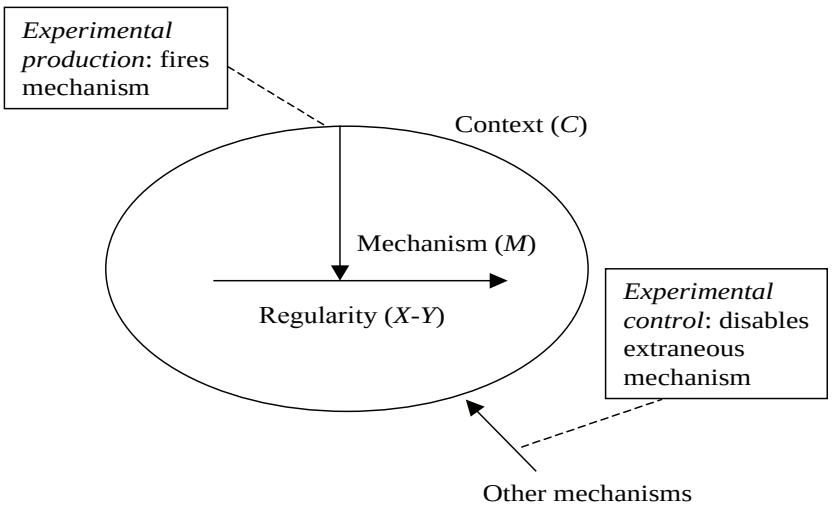
This section presents how CR can be used in IS design science research. The primary constituent community for the output of IS design science research is IS-professionals (Walls et al., 1992). This means primarily professionals who plan, manage and govern, design, build, implement, operate, maintain and evaluate different types of IS initiative and IS.

Using van Aken's (2004) classification we can distinguish three different types of designs an IS professional makes when designing and implementing an IS-initiative: 1) an *object-design*, which is the design of the IS intervention (initiative), 2) a *realization-design*, which is the plan for the implementation of the IS intervention (initiative), and 3) a *process-design*, which is the professional's own plan for the problem solving cycle and includes the methods and techniques to be used to design the solution (the IS intervention) to the problem. IS design science research should produce knowledge that can be used by the professionals in the three types of designs. Van Aken defines a technological rule as "…an instruction to perform a finite number of acts in a given order and with a given aim"; and a technological rule is "*a chunk of general knowledge, linking an intervention or artefact with a desired outcome or performance in a certain field of application*" (van Aken, 2004, p. 228). A technological rule is general, which for IS design knowledge means that a rule is a general prescription for a class of IS problems. Since a technological rule should be used by practitioners it should be applicable and actionable. Generally, the form of the technological rules is like "if you want to achieve A (outcome) in situation B (problem) and context C, then something like action/intervention D can help because E (reason)." "Something like action/intervention D" means that the rule is to be used as a design exemplar. A field-tested and grounded technological rule has been tested empirically and is grounded in science. Field-tested and grounded technological rules will in most cases be in the form of heuristics. This is consistent with CR's view on causality and means that the indeterminate nature of a heuristic technological rule makes it impossible to prove its effects conclusively, but it can be tested in context, which in turn can lead to sufficient supporting evidence (Groff, 2004).

Van Aken (2004) suggests that management design science research has much in common with CR-based evaluation research of social programs (Pawson & Tilley, 1997; Kazi, 2003). In line with CR-based evaluation research, the intention of IS design science research is to produce ever more detailed answers to the question of *why* and *how* an IS initiative works, *for whom*, and *in what circumstances*. This means that a researcher attends to how and why an IS initiative has the potential to cause the (desired) change. In this perspective, an IS design science (ISDS) researcher works as an experimental scientist, but not according to the logics of the traditional experimental evaluation research. Bhaskar states: "The experimental scientist must perform two essential functions in an experiment. First, he must trigger the mechanism under study to ensure that it is active; and secondly, he must prevent any interference with the operation of the mechanism. These activities could be designated as 'experimental production' and 'experimental control'." (Bhaskar, 1998). Figure 2 depicts the realist experiment.

ISDS researchers do not perceive that IS initiatives "work." It is the actions of different stakeholders and participants that make them work, and the causal potential of an IS initiative takes the form of providing the reasons and

Figure 2. The realist experiment (Pawson & Tilley, 1997)



resources to enable different stakeholders and participants to "make" changes. This means that an ISDS researcher seeks to understand why and how an IS initiative, for example, the implementation of an enterprise system works through understanding the action mechanisms. It also means that an ISDS researcher seeks to understand for whom and in what circumstances (contexts) an IS initiative works through the study of contextual conditioning.

ISDS researchers orient their thinking to context, mechanism, outcome pattern configurations (CMOCs). This leads to the development of transferable and cumulative lessons from ISDS research. A CMOC is a proposition stating what it is about an IS-initiative which works for whom in what circumstances. A refined CMOC is the finding of an evaluation of an IS initiative. Outcome patterns are examined from a "theory-testing" perspective. This means that an ISDS researcher tries to understand what the outcomes of an IS initiative are and how the outcomes are produced. Hence, the researcher does not just inspect outcomes in order to see if an IS initiative works, but analyzes the outcomes to discover if the conjectured mechanism/context theories are confirmed.

In terms of generalization, an ISDS researcher through a process of CMOC abstraction creates "middle-range" theories. These theories provide analytical frameworks for interpreting differences and similarities between classes and sub-classes of IS-initiatives. Given that the goal is to develop design theories and knowledge—to construct and test CMOCs explanations—for practitioners ISDS researchers need to engage in a learning relationship with IS practitioners.

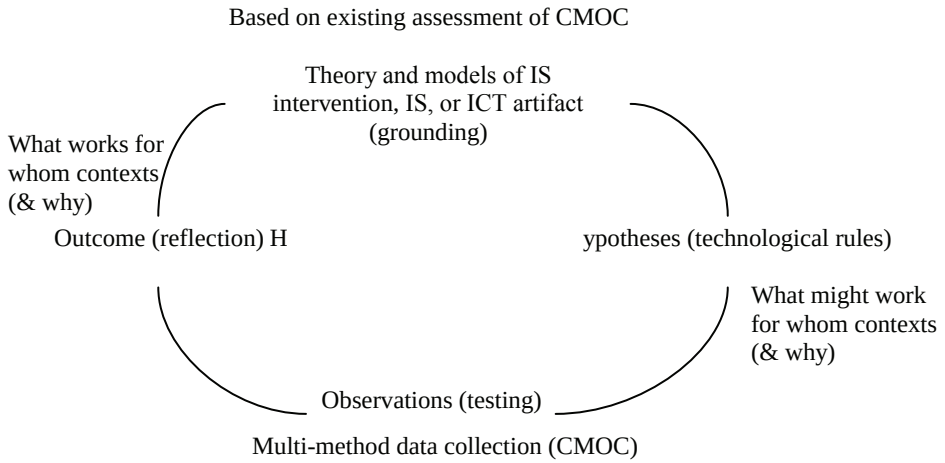
ISDS research based on the above can be carried out through an IS design science research cycle (Figure 3).

The starting point is theory and problems or issues. The research is driven by problems or issues. Problems or symptoms can be identified by practitioners or by researchers. For example, an organization can have the problem that their "ERP-projects are not leading to desired outcomes." The problems can also be identified through quantitative studies carried out by a researcher. For example, the researcher can analyze a data base containing use data for an IS and is looking for unwanted patterns. The theory includes propositions on how the mechanisms introduced by an IS intervention into a pre-existing context can generate (desired) outcomes. This entails theoretical analysis of mechanisms, contexts, and expected outcomes. This is the first step in developing technological rules and means that one tries to generate technological rules using our current knowledge, that is, grounding in theory. In general, the IS-researchers have been far from good in systematic reviews of research results. Pawson (2006) shows, from a critical realist perspective, how to do systematic reviews and make sense of a heterogeneous body of literature. Using Pawson's approach it should be possible to test and refine IS interventions. For example, it is possible to move away from the many one-off studies in the IS-field and instead learn from fields like medicine and policy studies on how to develop evidence-based IS design knowledge. Such a systematic review can be part of the starting point.

The second step consists of generating more specific "hypotheses." Typically the following questions would be addressed in the hypotheses: 1) what changes or outcomes will be brought about by an IS intervention (initiative), 2) what contexts impinge on this, and 3) what mechanisms (social, cultural, and others) would enable these changes, and which one may disable the intervention. In this step the technological rules are refined.

The third step is the empirical test. It is done through intervention and guided by theory and technological rules.

Figure 3. The information systems design science research cycle (based on Pawson & Tilley, 1997, and Kazi, 2003)



The step includes also the selection of appropriate data collection methods. ISDS research employs no standard research design formula. The base strategy is to develop a clear theory of IS initiative mechanisms, contexts and outcomes. Given the base strategy, an ISDS researcher has to design appropriate empirical methods, measures, and comparisons. In this step it might be possible to generate support of the IS intervention's ability to "change" reality. Based on the result from the third step, we may return to the IS intervention to make it more specific as an intervention of practice. Next, but not finally, we return to theory. The theory may be developed, the hypotheses and the technological rules refined, the data collection methods enhanced, etc. To develop the technological rules means that the cycle will be repeated. As said above most of the technological rules will be heuristic. Through multiple case-studies one can accumulate supporting evidence which can continue until "theoretical saturation" has been obtained. The researcher can be more or less active in the implementation (use) of the technological rules. The researcher can be very active and work like an action researcher, but can also be quite passive and work like an observer.

CONCLUSION

Although CR has influenced a number of social science fields, it is almost invisible in the IS field. CR's potential for IS research has been argued by, for example, Carlsson (2003, 2004, 2006), Dobson (2001), Mingers (2004), Mutch (2002), and Longshore Smith (2006). This article argued that CR can be used in IS research—behavioral and design science—to overcome problems associated with positivism, constructivism, and postmodernism.

REFERENCES

Archer, M., Bhaskar, R., Collier, A., Lawson, T., & Norrie, A. (Eds.).(1998). *Critical realism: Essential readings*. London: Routledge.

Bhaskar, R. (1978). *A realist theory of science*. Sussex: Harvester Press.

Bhaskar, R. (1998*). The possibility of naturalism* (3rd ed.). London: Routledge.

Bhaskar, R. (2002). *Reflections on meta-reality: Transcendence, enlightenment and everyday life*. London: Sage.

Carlsson, S. A. (2003). Advancing information systems evaluation (research): A critical realist approach. *Electronic Journal of Information Systems Evaluation, 6*(2), 11-20.

Carlsson, S. A. (2004). Using critical realism in IS research. In M.E. Whitman & A.B. Woszczynski (Eds.), *The handbook of information systems research* (pp. 323-338). Hershey, PA: Idea Group Publishing.

Carlsson, S. A. (2006). Towards an Information Systems design research framework: A critical realist perspective. In *Proceedings of the First International Conference on Design Science in Information Systems and Technology (DESRIST 2006)* (pp. 192-212).

Carlsson, S. A., Leidner, D.E., & Elam, J.J. (1996). Individual and organizational effectiveness: Perspectives on the impact of ESS in multinational organizations. In P. Humphreys, L. Bannon, A. McCosh, P. Migliarese and J.C. Pomerol (Eds.), *Implementing systems for supporting management decisions: Concepts, methods and experiences* (pp. 91-107). London: Chapman & Hall.

Chen, W., & Hirschheim, R. (2004). A paradigmatic and methodological examination of information systems research. *Information Systems Journal, 14*(3), 197-235.

Dobson, P. J. (2001). The philosophy of critical realism—An opportunity for information systems research. *Information Systems Frontier, 3*(2), 199-201.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611-642.

Groff, R. (2004). *Critical realism, post-positivism and the possibility of knowledge*. London: Routledge.

Hartwig, M. (Ed.). (2007). *Dictionary of critical realism*. London: Routledge.

Huber, G. P. (1990). A theory of the effects of advanced information technologies on organizational design, intelligence, and decision making. *Academy of Management Review, 15*(1), 47-71.

Kazi, M. A. F. (2003). *Realist evaluation in practice*. London: Sage.

Layder, D. (1998). *Sociological practice: Linking theory and social research*. London: Sage.

Lee, A. S., Liebenau, J., & DeGross, J. (Eds.). (1997). *Information systems and qualitative research*. London: Chapman & Hall.

Leidner, D. E., & Elam, J. J. (1995). The impact of executive information systems on organizational design, intelligence, and decision making. *Organization Science, 6*(6), 645-665.

Leidner, D. E., Carlsson, S. A., Elam, J. J., & Corrales, M. (1999). Mexican and Swedish managers´ perceptions of the impact of EIS on organizational intelligence, decision making, and structure. *Decision Sciences, 30*(3), 633-658.

Longshore Smith, M. (2006). Overcoming theory-practice inconsistencies: critical realism and information systems research. *Information and Organization, 16*(3), 191-211.

Mingers, J. (2004). Re-establishing the real: Critical realism and information systems. In J. Mingers & L. Willcocks (Eds.), *Social theory and philosophy for Information Systems* (pp. 372-406). Chichester, UK: Wiley.

Mutch, A. (2002). Actors and networks or agents and structures: Towards a realist view of information systems. *Organizations, 9*(3), 477-496.

Pawson, R. (2006). *Evidence-based policy: A realist perspective*. London: Sage.

Pawson, R., & Tilley, N. (1997). *Realistic evaluation*. London: Sage.

Quinn, R. E., Faerman, S. R., Thompson, M. P., & McGrath, M. R. (2004). *Becoming a master manager* (3rd ed.). New York: John Wiley & Sons.

Robson, C. (2002). *Real world research* (2nd ed.). Oxford, UK: Blackwell.

Sayer, A. (2000). *Realism and Social Science*. London: Sage.

Trauth, E. M., (Ed.).(2001). *Qualitative research in IS: Issues and trends*. Hershey, PA: Idea Group Publishing.

Van Aken, J. E. (2004). Management research based on the paradigm of design sciences: The quest for field-tested and grounded technological rules. *Journal of Management Studies, 41*(2), 219-246.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information systems design theory for vigilant EIS. *Information Systems Research, 3*(1), 36-59.

KEY TERMS

Constructivism (or Social Constructivism): Asserts that (social) actors socially construct reality.

Context-Mechanism-Outcome Pattern: Realist evaluation researchers orient their thinking to context-mechanism-outcome (CMO) pattern configurations. A CMO configuration is a proposition stating what it is about an IS initiative which works for whom in what circumstances. A refined CMO configuration is the finding of IS evaluation research.

Critical Realism: Asserts that the study of the social world should be concerned with the identification of the structures and mechanisms through which events and discourses are generated.

Empiricism: Asserts that only knowledge gained through experience and senses is acceptable in studies of reality.

Positivism: Asserts that reality is the sum of sense impression. In large, equating social sciences with natural sciences. Primarily using deductive logic and quantitative research methods.

Postmodernism: A position critical of realism and rejects the view of social sciences as a search for over-arching explanations of the social world. Has a preference for qualitative methods.

Realism: A position acknowledging a reality independent of actors' (incl. researchers') thoughts and beliefs.

Realist IS Evaluation: Evaluation (research) based on critical realism aiming at producing ever more detailed answers to the question of *why* an IS initiative works (better) for *whom* and in *what* circumstances (contexts).

Retroduction: The central mode of inference (explanation) in critical realism research. Enables a researcher to investigate the potential causal mechanisms and the conditions under which certain outcomes will or will not be realised.

Critical Success Factors for Distance Education Programs

Ben Martz
University of Colorado at Colorado Springs, USA

Venkat Reddy
University of Colorado at Colorado Springs, USA

INTRODUCTION

Distance education is playing an ever-growing role in the education industry. As such, it is prudent to explore and understand driving conditions that underlie this growth. Understanding these drivers and their corresponding concerns (Table 1) can help educators in the distance education field better prepare for the industry.

BACKGROUND

Distance education's primary driver is that it is the major growth segment in the education industry. In 1999, nearly 80% of the public, four-year institutions and over 60% of the public, two-year institutions offered distance education courses. Over 1.6 million students are enrolled in distance courses today. Over 90% of all colleges are expected to offer some online courses by 2004 (Institute of Higher Education Policy, 2000). Corporations envision online training warehouses saving large amounts of training dollars. Combined, the virtual education market and its sister market, corporate learning, are predicted to grow to over $21 billion by the end of 2003 (Svetcov, 2000).

A second major driver is employer expectations. Fundamental job market expectations are changing. Today, employees are not expected to stay in the same job for long periods of time; 20-plus year careers are not expected. The current modes of careers include multiple careers, combinations of part-time work in multiple jobs, telecommuting, leaving and re-entering into the full-time work force, switching jobs, and so forth, and today's employee easily accepts the need to maintain a level of knowledge current with the career demands (Boyatzis & Kram, 1999). To complement these changes in employer expectations, employees have begun to accept the need for life-long learning.

A third driver is the profit potential. Cost savings may be obtained and if significant enough may drive up demand and costs may be lowered. For example, elective classes that do not have enough students enrolled in them on-campus may pick up enough distance students to make teaching the course more feasible (Creahan & Hoge, 1998). A final driver is the institution's mission. Most educational institutions serve a geographical region, either by charter or mission, and a distance-learning program may be a practical method to help satisfy this strategic mission (Creahan & Hoge, 1998).

However, the "commercialization" of education raises its own concerns about the basic process of learning (Noble, 1999). For example, are there any problems fundamental to the distance environment because of limited social interaction?

Retention may be one such problem. Carr (2000) reports a 50% drop-out rate for online courses. Tinto (1975) compared the learning retention of distance groups with traditional groups and found that the social integration was a key factor in successful retention of traditional groups. Haythornthwaite et al. (2000) think they found another one. They looked at how social cues such as text without voice,

Table 1. Influences on the distance education industry

Table 1. Influences on the distance education industry	
Drivers	Concerns
Growth segment in education industry	Retention
Job market expectations	Fading Back
Life-long learning as an education paradigm	Less social learning
Profit center for educational institutions	Trust & isolation
Possible strategic competence	Impact of technology

voice without body language, class attendance without seating arrangements, and students signing in without attending Internet class impacted students "fading back." They found that the likelihood of students "fading back" is greater in distance-learning classes than in face-to-face classes. From the United Kingdom, Hogan and Kwiatkowski (1998) argue that the emotional aspects of this teaching method have been ignored. Similar concerns are raised from Australia, where technology has been supporting distance- teaching for many years, as Hearn and Scott (1998) suggest that before adopting technology for distance teaching, education must acknowledge the social context of learning. Finally, two other factors, trust and isolation, have been researched by Kirkman et al. (2002), whereby communication helped improve the measures of trust in students using the virtual environment.

By definition, the paradigm of distance education changes the traditional education environment by expanding it to cover geographically dispersed learning. In turn, this means that students will probably respond differently to this environment than they do to the traditional classroom. In addition, academic researchers have always been interested in explaining how people react to the introduction of technology. This body of work can be useful to the distance education environment.

Poole and DeSanctis (1990) suggested a model called adaptive structuration theory (AST). The fundamental premise of the model is that the technology under study is the limiting factor or the constraint for communication. It further proposes that the users of the technology, the senders and the receivers, figure out alternative ways to send information over the channel (technology). A good example here is how a sender of e-mail may use combinations of keyboard characters or emoticons (i.e., :) – sarcastic smile, ;) – wink, :o – exclamation of surprise) to communicate more about their emotion on a subject to the receiver.

Ultimately, the key to realizing the potential of distance education is trading off the benefits and the concerns to produce a quality product. In the new Malcolm Baldridge evaluation criteria, companies are asked to better show a program's effectiveness through customer satisfaction. In turn, Gustafsson et al. (2000) show customer satisfaction linked significantly to quality at Volvo Car Corporation. Finally, in their more broad analysis of well-run companies, Peters and Waterman (1982) deemed customer satisfaction as a key factor contributing to the companies' performance.

With these perspectives in mind, we suggest that these areas interact to identify satisfaction as one important measure of quality for distance education programs. Therefore, one of the key factors to a program's success will be the

Table 2. Questions that correlate significantly to satisfaction

ID	Question Statement	Correlation	
		Coef.	Sign.
16	I was satisfied with the content of the course	.605	.000
17	The tests were fair assessments of my knowledge	.473	.000
18	I would take another distance course with this professor	.755	.000
19	I would take another distance course	.398	.000
20	The course workload was fair	.467	.000
21	The amount of interaction with the professor and other students was what I expected.	.710	.000
22	The course used groups to help with learning	.495	.000
23	I would like to have had more interaction with the professor.	-.508	.000
26	The course content was valuable to me personally	.439	.000
28	Grading was fair	.735	.000
30	Often I felt "lost" in the distance class	-.394	.000
31	The class instructions were explicit	.452	.000
33	Feedback from the instructor was timely	.592	.000
34	I received personalized feedback from the instructor	.499	.000
36	I would have learned more if I had taken this class on-campus (as opposed to online)	-.400	.000
37	This course made me think critically about the issues covered.	.423	.000
38	I think technology (email, web, discussion forums) was utilized effectively in this class	.559	.000
39	I felt that I could customize my learning more in the distance format	.254	.001
42	The course content was valuable to me professionally	.442	.000
43	I missed the interaction of a "live," traditional classroom	-.341	.002
46	Overall, the program is a good value (quality/cost)	.258(1)	.017
LOHITECH	Aggregate of Yes votes in Q6 through Q15	.270(1)	.012
(1) While significant, the low correlation coefficient below .300 should be noted			

satisfaction of one of its key stakeholders – its students. If one can identify what helps satisfies students in a distance education environment, one has a better chance to develop a successful program.

THE RESEARCH STUDY

The distance program used in this study is one of the largest, online, AACSB-accredited MBA programs in the world (US News and World Report, 2001). The methodology used a questionnaire with a battery of 49 questions to gather the data. The questions were developed using the concepts and ideas from literature discussed earlier as a guide.

Once the subject identified his or her reference course, that subject's grade was obtained from administrative records and recorded. In addition, four other demographic questions gathered information on gender, number of courses taken, student status, amount of time expected to spend in the reference course, and the amount of time actually spent in the reference course (Martz et al., 2004).

Two sets of questions were used. The first set asked about the student's use of different technologies (i.e., chat, e-mail, streaming video, etc.) in the class and if used, how effective (five-point Likert: 1 = LO …. 5 = HIGH) did they believe the technology to be in helping them with the class. We created a new variable, LOHITECH, for analysis purposes. Using LOHITECH, respondents can be placed in one of two groups: one group that reported using three or less technologies, while the second group reported using four or more technologies in their reference class. The second set of questions asked students to rate (five-point Likert: 1 = Strongly Agree …. 5 = Strongly Disagree) their experience with the reference distance course against statements concerning potential influences for satisfaction. These questions associated a five-point rating scale to statements about the issues identified earlier. The order of the questions was randomly determined and the questionnaire was reviewed for biased or misleading questions by non-authors.

The questionnaire was sent to 341 students enrolled in the distance MBA program. In Fall 2002, the program served 206 students from 39 states and 12 countries. The majority of these students are employed full-time. The program used in this study has been running since Fall 1996 and has over 179 graduates. It offers an AACSB accredited MBA and its curriculum parallels the on-campus curriculum. Close to 33% of the enrolled students are female. The oldest student enrolled is 60 years old and the youngest is 22. The average age of all students enrolled is 35. Over 25 PhD qualified instructors participate in developing and delivering the distance program annually. Recently, the news magazine *US News and World Report* (2001) classified the program as one of the top 26 distance education programs.

There were 131 useable questionnaires returned. The students' final grade for their reference course was obtained and added to the questionnaire record as a variable. These were separated into two groups: 30 that had not yet taken a course and 101 that had completed at least one course. This second group, those students who had completed at least one course, provided the focus for this study.

RESEARCH RESULTS

Question 24, "Overall, I was satisfied with the course," was used as the subject's level of general satisfaction. The data set was loaded into SPSS for analysis. Table 2 shows that 23 variables, including LOHITECH, proved significantly correlated to satisfaction (Q24).

The large number of significant variables leads to the need for a more detailed analysis on how to group them (StatSoft, 2002). Kerlinger (1986, p. 590) suggests the use of factor analysis in this case "to explore variable areas in order to identify the factors presumably underlying the variables". An SPSS factor analysis was performed with a Varimax Extraction on those questions that had proven significantly correlated to satisfaction. All reliability coefficients (Cronbach Alpha) are above .7000 and all Eigenvalues are above 1.00, indicating an acceptable level for a viable factor (Kline, 1993; Nunnally, 1978). Finally, the five components explain 66.932% of the variance.

In summary, 22 variables from the questionnaire proved significantly correlated to satisfaction. A factor analysis of those 22 variables extracted five possible constructs. These

Table 3. Correlation of final constructs to satisfaction

Construct (Component: Loading)	Correlation	Significance
Professor Interaction (Q18: .576, Q21: .643, Q33: .794, Q34: .849)	.771	.000
Fairness (Q17: .722, Q20: .738, Q28: .626, Q31: .512)	.695	.000
Course Content (Q16: .596, Q26: .850, Q39: .689, Q42: .825)	.588	.000
Classroom Interaction (Q23: -.354, Q30: -.514, Q36: -.809, Q43: -.770)	-.515	.000
Technology Use & Value (LOHITECH: .508, Q19: .596, Q22: .542, Q37: .494, Q38: .478, Q46: .700)	.624	.000

constructs were labeled: Interaction with the Professor; Fairness; Content of the Course; Classroom Interaction; and Value, Technology & Learning, based upon the key characteristics of the underlying questions. Table 3 shows the results of combining the ratings for the questions in each construct and correlating each of them to satisfaction. As can be seen from the table, the constructs hold up well as five indicators of satisfaction.

FUTURE TRENDS

As mentioned earlier, the organization, the school in this case, is a key stakeholder in the success of a distance education program. The future success of distance programs depends largely on satisfying these critical success factors. Distance education courses and programs are not only used for providing an alternative delivery method for students but also to generate revenues for the offering unit/college/university. As the number of distance courses and programs increase at an exponential rate, the necessity to enhance quality and revenues also takes prominence. We conclude with a set of operational recommendations that can impact online program success (Table 4).

The data in this study indicate that a timely and personalized feedback by professors results in a higher level of satisfaction by students. The administrators therefore have to work closely with their faculty and offer them ways to enrich the teacher-student relationships. Paradoxically, a faculty member needs to use technologies to add a personal touch to the virtual classroom. For example, faculty should be encouraged to increase the usage of discussion forums, respond to e-mail within 24 to 48 hours, and keep students up-to-date with the latest happenings related to the course.

The data also indicate that good course content and explicit instructions increase student satisfaction in the virtual classroom. It may well be that this basically sets and manages the expectations for the distance student. This result suggests that faculty should have complete Web sites with syllabi and detailed instructions. In turn, this suggests that distance education administrators should focus their attention on providing faculty with support such as good Web site design, instructional designer support, test design, user interaction techniques, and so forth, appropriate for distance learning.

Since distance students' notion of value intertwines learning and technology, it is imperative that distance administrators offer, and faculty use, the available technology in the distance program. Technology in this case not only refers to the actual software and hardware features of the platform but also how well technology is adapted to the best practices of teaching. The results imply that if technology is available but not used, it lowers satisfaction. So, technology options that are not being used in a course should not appear available. For the program administrator, this would suggest adoption of distance platforms that are customizable at the course level with respect to displaying technological options.

CONCLUSION

This study attempts to identify potential indicators for satisfaction with distance education. A body of possible indicators was derived from the literature surrounding the traditional versus virtual classroom debate. A 49-question questionnaire was developed from the indicators and was administered to MBA students in an established distance education program. One hundred and one questionnaires from students with one or more distance classes were analyzed with the result that 22 variables correlated significantly to satisfaction. A factor analysis of the questionnaire data extracted five basic constructs: Professor Interaction, Fairness, Course Content, Classroom Interaction and Technology Use & Value. Several recommendations for implementing and

Table 4. Recommendations to increase online program success

1	Have instructors use a 24-48-hour turnaround for e-mail.
2	Have instructors use a 1-week turnaround for graded assignments.
3	Provide weekly "keeping in touch" communications.
4	Provide clear expectation of workload.
5	Provide explicit grading policies.
6	Explicitly separate technical and pedagogical issues.
7	Have policies in place that deal effectively with technical problems.
8	Provide detailed unambiguous instructions for coursework submission.
9	Provide faculty with instructional design support.
10	Do not force student interaction without good pedagogical rationale.
11	Do not force technological interaction without good pedagogical purpose.
12	Collect regular student and faculty feedback for continuous improvement.

managing a distance program were extracted from these constructs and discussed.

REFERENCES

Boyatzis, R.E., & Kram, K.E. (1999, Autumn). Reconstructing management education as lifelong learning. *Selections, 16*(1), 17-27.

Carr, S. (2000, February 11). As distance education comes of age the challenge is keeping students. *Chronicle of Higher Education.*

Creahan, T.A., & Hoge, B. (1998, September). *Distance learning: Paradigm shift of pedagogical drift?* Presentation at Fifth EDINEB Conference, Cleveland, OH.

Gustafsson, A., Ekdahl, F., Falk, K., & Johnson, M. (2000, January). Linking customer satisfaction to product design: A key to success for Volvo. *Quality Management Journal, 7*(1), 27-38.

Haythornthwaite, C., Kazmer, M.M., Robins, J., & Showmaker, S. (2000, September). Community development among distance learners. *Journal of Computer-Mediated Communication, 6*(1).

Hearn, G., & Scott, D. (1998, September). Students staying home. *Futures, 30*(7), 731-737.

Hogan, D., & Kwiatkowksi, R. (1998, November). Emotional aspects of large group teaching. *Human Relations, 51*(11), 1403-1417.

Institute for Higher Education Policy. (2000). *Quality on the line: Benchmarks for success in Internet distance education.* Washington, D.C.

Kerlinger, F.N. (1986). *Foundations of behavioral research* (3rd ed.). Holt, Rinehart & Winston.

Kirkman, B.L., Rosen, B., Gibson, C.B., Etsluk, P.E., & McPherson, S. (2002, August). Five challenges to virtual team success: Lessons from Sabre, Inc. *The Academy of Management Executive, 16*(3).

Kline, P. (1993). *The handbook of psychological testing.* London: Routledge.

Martz, W.B, Reddy, V., & Sangermano, K. (2004). Assessing the impact of Internet testing: Lower perceived performance. In C. Howard, K. Schenk & R. Discenza (Eds.), *Distance learning and university effectiveness: Changing educational paradigms for online learning.* Hershey, PA: Idea Group Publishing.

Noble, D.F. (1999). Digital diplomas mills. Retrieved November 28, 2002, from *http://www.firstmonday.dk/issues/issue3_1/noble/index.html*

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Peters, T.J., & Waterman, R.H., Jr. (1982). *In search of excellence.* New York: Harper and Row.

Poole, M.S., & DeSanctis, G. (1990). Understanding the use of group decision support systems: The theory of adaptive structuration. In J. Fulk & C. Steinfeld (Eds.), *Organizations and communication technology* (pp. 173-193). Newbury Park, CA: Sage Publications.

Rockart, J.F. (1979, March-April). Chief executives define their own data needs. *Harvard Business Review.*

Statsoft. (2002). Retrieved November 30, 2002, from *http://www.statsoftinc.com/textbook/stfacan.html*

Svetcov, D. (2000). The virtual classroom vs. the real one. *Forbes,* 50-52.

Tinto, V. (1975). *Leaving college.* University of Chicago Press.

US News and World Report. (2001, October). *Best online graduate programs.*

KEY TERMS

Classroom Interaction: The interaction that can only be achieved face-to- face in a classroom. For example, the real-time feedback of facial expressions is not (yet) available in a distance course and so would be considered "classroom interaction".

Concerns of "Commercialization": The negative factors that the implantation and use of distance education may create.

Course Content: The main themes covered in a course.

Critical Success Factors: The few key areas in which activities must "go right" so that a project of program succeeds (Rockart, 1979).

Exploratory Factor Analysis: A process used to identify statistically significant constructs underlying a set of data.

Fairness: A subjective term defining the level to which a student feels he or she was treated fairly by the professor with respect to the class, including but not limited to test questions, grading, schedule flexibility, and so forth.

Market Drivers for Distance Education: The key elements that seem to be driving the diffusion and usage of distance education in the marketplace.

Professor Interaction: The amount of communication (e-mail, phone calls, video, chat rooms, etc.) that occurs between a student and the professor.

Satisfaction Constructs for Distance Education: Five constructs identified that seem to help identify satisfaction in distance education programs.

Technology Use: The usage of a technology whether it be e-mail, chat rooms, automated tests, software, and so forth.

Technology Value: The user's benefits (perceived and actual) over the costs (perceived and actual) created by the use of technology.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 622-627, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Critical Success Factors for E-Health

Nilmini Wickramasinghe
Illinois Institute of Technology, USA

Jonathan L. Schaffer
The Cleveland Clinic, USA

INTRODUCTION

Within the umbrella of e-commerce, one area, e-health, has yet to reach its full potential in many developed countries, let alone developing countries. Each country is positioned differently and has varying potential and preparedness regarding embracing e-commerce technologies generally and e-health in particular. Given the macrolevel nature of many issues pertaining to the development of e-health (Alvarez, 2002), in order to be more effective in their e-health initiatives, it is important for countries to assess their potential, identify their relative strengths and weaknesses, and thereby develop strategies and policies to address these issues to effectively formulate and implement appropriate e-health initiatives. To do this effectively, it is valuable to have an integrative framework that enables the assessment of a country's e-health preparedness. This article serves to develop such a framework that can be applied to various countries throughout the globe, and from this generate an e-health preparedness grid. In so doing, we hope to facilitate better understanding of e-health initiatives and thus maximize their power.

BACKGROUND

Reducing health care expenditure as well as offering quality health care treatment is becoming a priority globally. Technology and automation have the potential to reduce these costs (Institute of Medicine, 2001; Wickramasinghe, 2000); thus, e-health, which essentially involves the adoption and adaptation of e-commerce technologies throughout the health care industry (Eysenbach, 2001; Wickramasinghe, Misra, Jenkins, & Vogel, 2006), appears to be a powerful force of change for the health care industry worldwide.

Health care has been shaped by each nation's own set of cultures, traditions, payment mechanisms, and patient expectations. Therefore, when looking at health systems throughout the world, it is useful to position them on a continuum ranging from high government involvement (i.e., a public health care system) at one extreme to little government involvement (i.e., a private health care system) at the other extreme, with many variations of a mix of private and public in between. However, given the common problem of exponentially increasing costs facing health care globally, irrespective of the particular health system one examines, the future of the health care industry will be partially shaped by commonalties such as the universal issue of escalating costs and the common forces of change including (a) empowered consumers, (b) e-health adoption and adaptability, and (c) a shift to focus on the practice of preventative- vs. cure-driven medicine. Additionally, there will be four key implications, namely, (a) health insurance changes, (b) workforce changes and changes in the roles of stakeholders within the health system, (c) organizational changes and standardization, and (d) the need for health care providers and administrators to make difficult yet necessary choices regarding practice management.

THE GOALS OF E-HEALTH

In order to develop a robust framework, it is imperative to understand the many goals of e-health. These goals, taken together, perhaps best characterize what e-health is all about (or what it should be about; *Journal of Medical Internet Research* [JMIR], 2003). Specifically, significant goals of e-health include the following.

Efficiency: One of the promises of e-health is to increase efficiency in health care, thereby decreasing costs. One possible way of decreasing costs would be by avoiding duplicative or unnecessary diagnostic or therapeutic interventions, through enhanced communication possibilities between health care establishments, and through patient involvement (Health Technology Center, 2000). The Internet will naturally serve as an enabler for achieving this goal in e-health.

Quality of care: Increasing efficiency involves not only reducing costs, and thus is not an end in and of itself, but rather should be considered in conjunction with improving quality, one of the ultimate goals of e-health. More educated consumers (as a result of the informational aspects of e-health) would then communicate more effectively with their primary care providers, which will, in turn, lead to better understanding and improved quality of care.

Evidence: E-health interventions should be evidence based in the sense that their effectiveness and efficiency should not be assumed but proven by rigorous scientific

evaluation and support from case histories. Web-accessible case repositories facilitate the timely accessibility of such evidence and thus help in achieving the necessary support of a diagnosis or treatment decision. The evidence-based medicine goal of e-health is currently one of the most active e-health research domains, yet much work still needs to be done in this area.

Empowerment of consumers and patients: By making the knowledge bases of medicine and personal electronic records accessible to consumers over the Internet, e-health opens new avenues for patient-centered medicine, enables patient education, and thus increases the likelihood of informed and more satisfactory patient choices (Umhoff & Winn, 1999).

Education of physicians and consumers: Online sources (continuing medical education for physicians, and health education and tailored preventive information for consumers) make it easier for physicians as well as consumers to keep up to date with the latest developments in the medical areas of their respective interests. This, in turn, is likely to have a positive impact on the quality of care vis-à-vis the use of the latest medical treatments and preventive protocols.

Extension of health care: Extending the scope of health care beyond its conventional boundaries, in both a geographical sense as well as in a conceptual sense, leads to enabling such techniques as telemedicine and virtual operating rooms, both of which are invaluable in providing health care services to places where it may otherwise be difficult or impossible to do.

Ethics: E-health involves new forms of patient-physician interaction and poses new challenges and threats to ethical issues such as online professional practice, informed consent, privacy, and security issues (Healthcare Advisory Board, 2002). However, this is not an intrinsic feature of e-health but rather a feature of the Internet technology, which is the foundation for all e-business initiatives; therefore, e-health along with e-government, e-insurance, e-banking, e-finance, and e-retailing must all contend with these ethical issues. Given the nature of health care, these issues could be more magnified.

Equity: To make health care more equitable is one of the aims of quality identified by the American Institute of Medicine (2001) generally and is one of the goals of e-health. However, at the same time there is a considerable threat that e-health, if improperly implemented and used, may deepen the gap between the "haves" and the "have-nots," hence the need for a robust framework to ensure the proper implementation of e-health initiatives. In particular, some of the key issues for equity revolve around broad access and familiarity with the technology.

PREREQUISITES FOR E-HEALTH

In order to actualize and thereby support the key goals of e-health presented above, it is necessary to have four critical prerequisites for any successful e-health initiative, namely, ICT architecture and infrastructure; standardized policies, protocols, and procedures; user access and accessibility policies and infrastructure; and finally government regulation and control. These will now be briefly discussed in turn.

ICT Architecture and Infrastructure

The ICT infrastructure typically includes phone lines, fiber trunks, submarine cables, T1, T3, OC-xx, ISDN (integrated services digital network), DSL (digital subscriber line), and other high-speed services used by businesses, as well as satellites, earth stations, and teleports. A sound technical infrastructure is an essential ingredient to the undertaking of e-health initiatives by any nation. Such infrastructures should also include telecommunications, electricity, access to computers, Internet hosts, ISPs (Internet service providers), and available bandwidth and broadband access. To offer good multimedia content and thus provide a rich e-health experience, one would require high bandwidth. ICT considerations are undoubtedly one of the most fundamental infrastructure requirements.

Networks are now a critical component of the business strategies for organizations to compete globally. Having a fast microprocessor-based computer at home has no meaning unless you have high-bandwidth-based communication infrastructure available to connect computers with the ISP. With the explosion of the Internet and the advent of e-commerce, global networks need to be accessible, reliable, and fast to participate effectively in the global business environment. Telecommunications is a vital infrastructure for Internet access and hence for e-commerce. One of the pioneering countries in establishing a complete and robust e-health infrastructure is Singapore, which is in the process of wiring every home, office, and factory to a broadband cable network that will cover 98% of Singaporean homes and offices (Wickramasinghe, 2007a).

Standardization Policies, Protocols, and Procedures

E-health by definition spans many parties and geographic dimensions. To enable such far-reaching coverage, significant amounts of document exchange and information flow must be accommodated. Standardization is the key for this. Once a country decides to undertake e-health initiatives, standardization polices, protocols, and procedures must

be developed at the outset to ensure the full realization of the goals of e-health. Fortunately, the main infrastructure of e-health is the Internet, which imposes the most widely and universally accepted standard protocols such as TCP/IP (transmission-control protocol/Internet protocol) and HTTP (hypertext transfer protocol). It is the existence of these standard protocols that has led to the widespread adoption of the Internet for e-commerce applications.

The shift to e-health by any country cannot be successfully attained without the deliberate establishment of standardization policies, protocols, and procedures that play a significant role in the adoption of e-health and the reduction of many structural impediments (Samiee, 1998).

User Access and Accessibility Policies and Infrastructure

Access to e-commerce is defined by the WTO (World Trade Organization) as consisting of two critical components: (a) access to Internet services and (b) access to e-services (Panagariya, 2000); the former deals with the user infrastructure, while the latter pertains to specific commitments to electronically accessible services. The user infrastructure includes the number of Internet hosts and number of Web sites, the number of Web users as a percent of the population as well as ISP availability and costs for consumers, the PC penetration level, and so forth. Integral to the user infrastructure is the diffusion rate of PCs and Internet usage. The United States and the United Kingdom have experienced the greatest penetration of home computers (Samiee, 1998). For developing countries such as India and China, there is, however, very low PC penetration and teledensity. In such a setting, it is a considerable challenge then to offer e-health since a large part of the population is not able to afford to join the e-commerce bandwagon. Countries thus have to balance local call charges, rentals, subscription charges, and so on; otherwise, the majority of citizens will find these costs a disincentive. This is particularly significant for developing and emerging nations where access prices tend to be out of reach for most of the population. Upcoming new technologies hold the promise to increase connectivity as well as the affordability level, and developing countries will need to seriously consider these technologies. In addition to access to PCs and the Internet, computer literacy is important; users must be familiar not only with the use of computers and pertinent software products, but also with the benefits and potential uses of the Internet and World Wide Web (Samiee).

Governmental Regulation and Control

The key challenges regarding e-health use include (a) cost effectiveness, that is, e-health must be less costly than traditional health care delivery, (b) functionality and ease of use, meaning it should enable and facilitate many uses for physicians and other health care users by combining various types and forms of data as well as be easy to use, and (c) e-health must be secure. One of the most significant legislative regulations in the United States is the Health Insurance Portability and Accountability Act (HIPAA; Protegrity, 2001).

Given the nature of health care and the sensitivity of health care data and information, it is incumbent on governments not only to mandate regulations that will facilitate the exchange of health care documents between the various health care stakeholders, but also to provide protection of privacy and the rights of patients. Some countries, such as China and Singapore, even control access to certain sites for moral, social, and political reasons while elsewhere, transnational data flows are hindered by a plethora of regulations aimed at protecting domestic technology and related human resource markets (Gupta, 1992; Samiee, 1998; Wickramasinghe, 2007a). Irrespective of the type of health care system, that is, whether it is 100% government driven, 100% private, or a combination thereof, it is clear that some governmental role is required to facilitate successful e-health initiatives.

KEY IMPACT OF E-HEALTH

The significance of the preceding four prerequisites for e-health initiatives will be modified by the impacts of IT education, morbidity, cultural and social dimensions, and the world economic standing as elaborated upon below.

Impact of IT Education

A sophisticated, well-educated population boosts competition and hastens innovation. According to Michael Porter (1990), one of the key factors to a country's strength in an industry is strong customer support. Thus, a strong domestic market leads to the growth of competition, which leads to innovation and the adoption of technology-enabled solutions to provide more effective and efficient services such as e-health and telemedicine. As identified earlier, the health consumer is the key driving force in pushing e-health initiatives. We conjecture that a more IT-educated health care consumer would then provide stronger impetus for e-health adoption.

Impact of Morbidity Rate

There is a direct relationship between health education and awareness and the overall health standing of a country. Therefore, a more health-conscious society, which tends to coincide with a society that has a lower morbidity rate, is more likely to embrace e-health initiatives. Furthermore, higher morbidity

rates tend to indicate the existence of more basic health needs (World Health Organization [WHO], 2003). Hence, treatment is more urgent than the practice of preventative medicine and thus e-health could be considered an unrealistic luxury; in some instances, such as when a significant percentage of a population is suffering from malnutrition-related diseases, e-health is even likely to be irrelevant at least in the short term. Thus, we conjecture that the modifying impact of the morbidity rate prioritizes the level of spending on e-health vs. other basic health care needs.

Impact of Cultural and Social Dimensions

Health care has been shaped by each nation's own set of cultures, traditions, payment mechanisms, and patient expectations. While the adoption of e-health, to a great extent, dilutes this cultural impact, social and cultural dimensions will still be a moderating influence on any countries' e-health initiatives. Another aspect of the cultural and social dimension relates to the presentation language of the content of the e-health repositories. The entire world does not speak English, so the e-health solutions have to be offered in many other languages. The e-health supporting content in Web servers and sites must be offered in local languages, supported by pictures and universal icons. This becomes a particularly important consideration when we look at the adoption and diffusion of evidence-based medicine as it will mean that much of the available evidence and case-study data will not be easily accessible globally due to language barriers.

Therefore, for successful e-health initiatives, it is important to consider cultural dimensions. For instance, an international e-commerce study by International Data Corp. indicates that Web surfing and buying habits differ substantially from country to country (Wilson, 1999), and this would then have a direct impact on people's comfort in using e-commerce generally and e-health in particular, especially as e-health addresses a more fundamental need. Hence, the adoption of e-health is directly related to one's comfort with using the technology and this in turn is influenced in a major way by cultural dimensions. Also connected to cultural aspects is the relative entrepreneurial spirit of a country. For example, Hofstede (1980) indicates that in a cultural context, Indians score high on "uncertainty-avoidance" criteria when compared to their Western counterparts. As a result, Indian nationals, for example, do not accept change very easily and are hostile toward innovation. This then would potentially pose a challenge to the start-up of e-health initiatives, whose success depends on widespread adoption for their technological innovations. Thus, we conjecture that fear of risk and absence of an entrepreneurial mind-set as well as other cultural and social dimensions can impact the success of e-health initiatives in a given country.

Impact of World Economic Standing

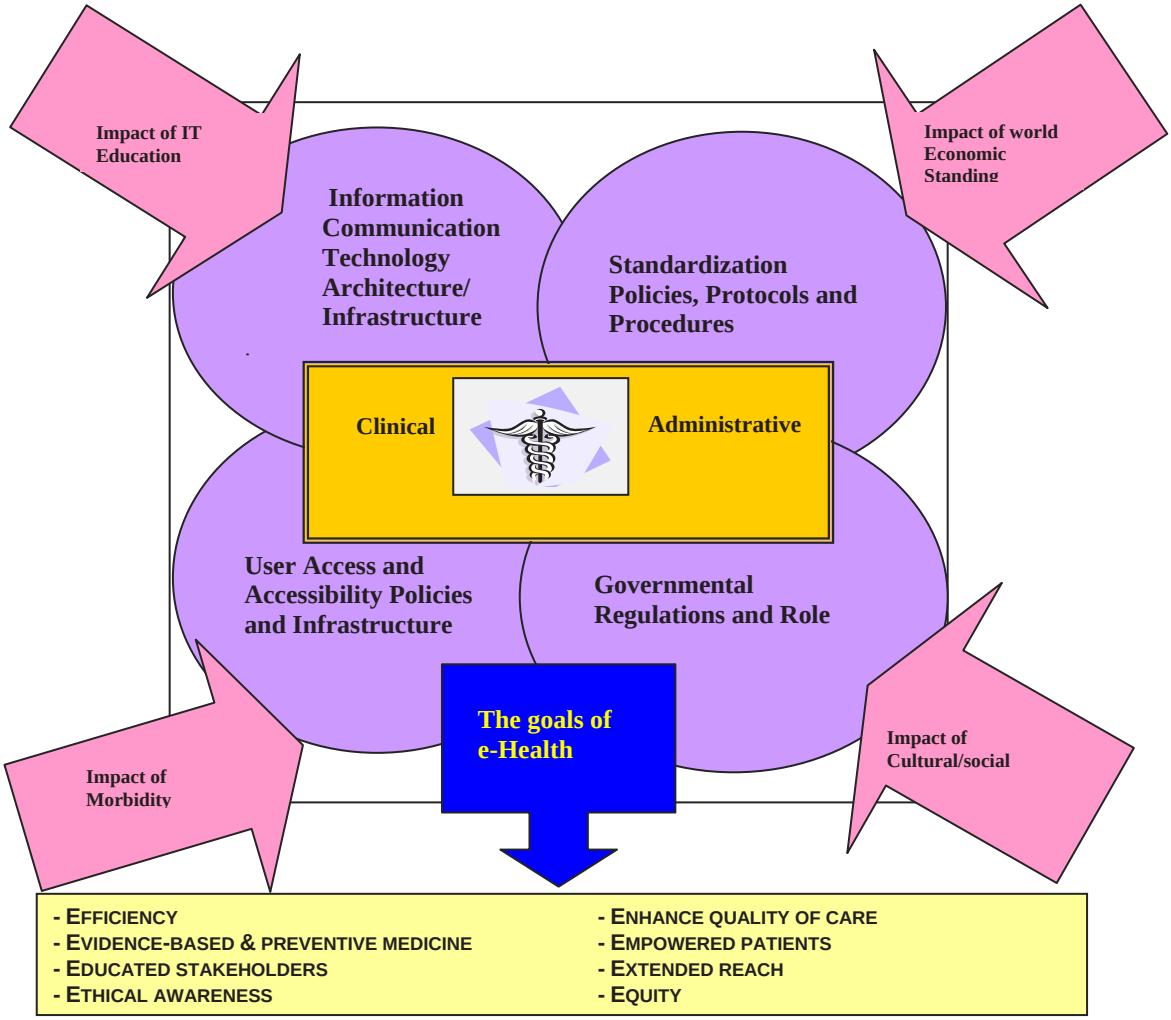
Economies of the future will be built around the Internet. All governments are very aware of the importance and critical role that the Internet will play on a country's economy. This makes it critical that appropriate funding levels and budgetary allocations become a key component of governmental fiscal policies so that such initiatives will form the bridge between a traditional health care present and a promising e-health future. Thus, the result of these initiatives would determine the success of effective e-health implementations and consequently have the potential to enhance a country's economy and future growth.

The World Economic Forum's global competitiveness ranking measures the relative global competitiveness of a country. This ranking takes into account factors such as physical infrastructure, bureaucracy, and corruption. Thus, we conjecture that when weak physical infrastructure is combined with high levels of bureaucracy and corruption, this will lead to significant impediments to the establishment of successful e-health initiatives.

A FRAMEWORK FOR ASSESSING E-HEALTH POTENTIAL

In order to understand numerous critical considerations to facilitate prudent decision making concerning any e-health initiative, it is most useful to have one integrative framework that not only brings together the eight goals of e-health but also the prerequisites and key impact aspects. We propose the framework shown in Figure 1 as such an integrative framework to assess the e-health potential and preparedness of countries as well as potential barriers for any particular e-health initiative. Health care polices are generally developed to a large extent at a macro, country level and thus we believe it is also necessary when looking at e-health to first take a macro perspective and analyze the level of the country in terms of embracing e-health. The framework highlights the key elements that are required for successful e-health initiatives and therefore provides a tool that allows analysis beyond the quantifiable data into a systematic synthesis of the major impacts and prerequisites. The framework contains four main prerequisites, four main impacts, and the implications of these prerequisites and impacts to the goals of e-health. By examining both the prerequisites and the impacts, it is now possible to assess the potential of a country and its preparedness for e-health as well as its ability to maximize the goals of e-health in a systematic and careful manner.

Figure 1. A framework for assessing a country's or region's e-health potential



FUTURE TRENDS

In developing an e-health initiative, a good first step for any country is to assess its standing with respect to the four prerequisites and four impacts discussed above. In this way it will be possible to evaluate its preparedness with respect to these parameters and consequently devise appropriate policies and strategies for an effective and successful e-health initiative. As e-health initiatives become more prevalent and mature, it will also become necessary to continually update and refine many aspects, most especially the knowledge and data content (Puentes, Bali, Wickramasinghe, & Naguib, 2007; Wickramasinghe, 2007b). To do this in a systematic fashion so that at all times it is possible to leverage the extant knowledge base, we anticipate that the application of the tools and techniques of knowledge management will begin to play a growing role in the future developments of all e-health initiatives, and we urge for more research in this area.

CONCLUSION

E-commerce, as noted by the United Nations secretary general's address, is an important aspect of business in the 21st century. No longer then is it a luxury for nations, but rather it is a strategic necessity in order for countries to achieve economic and business prosperity as well as social viability. One of the major areas within e-commerce that

has yet to reach its full potential is e-health. This is due to the fact that health care generally has been slow in adopting information technologies. Furthermore, there is a shortage of robust frameworks that may be used as guidelines for assessing countries' e-health preparedness and identifying the key areas and deficiencies that need to be addressed in order for successful e-heath initiatives to ensue. In addition, e-health is more than a technological initiative; rather, it also requires a major paradigm shift in health care delivery, practice, and thinking. We have attempted to address this gap by developing a framework that identifies the major factors involved in assessing the e-health preparedness of countries, thereby facilitating them in focusing their efforts on the relevant issues that must be addressed in order for successful e-health initiatives to follow (the goals of e-health are realized). An outcome from our analysis indicates that the relative health care system (i.e., whether government driven, public, or two tier) would appear to have less significance in establishing successful e-health initiatives. The first step in the development of any viable e-health strategy is to make an assessment of the current state of e-health preparedness and then to move to a state of higher preparedness. Finally, we note that with respect to our framework, other parameters also exist and could also be considered important, perhaps even as important as the ones we used. However, we believe that the framework will still function the same way (i.e., provide a useful tool for any country trying to determine and develop a successful e-health initiative) irrespective of the number of parameters; in this regard, we preferred simplicity over complexity.

REFERENCES

Alvarez, R. C. (2002). The promise of e-health: A Canadian perspective. *eHealth International, 1*(1), 4.

Eysenbach, G. (2001). *Journal of Medical Internet Research, 3*(2), e20.

Gupta, U. (1992). Global networks: Promises and challenges. *Information Systems Management, 9*(4), 28-32.

Healthcare Advisory Board. (2002). *Use of hospital Web sites to engender community loyalty.*

Health Technology Center. (2000). *A survey conducted for the Health Technology Center (HealthTech) by Harris Interactive in cooperation with Pricewaterhouse Coopers and the Institute for the Future (IFTF).* Retrieved from http://www.ncddr.org/cgi-bin/good-bye.cgi?url=http://www.healthtechcenter.org

Hofstede, G. (1980). *Culture's consequences: International work related values.* Beverly Hills, CA: Sage Publishing.

Institute of Medicine. (2001). *Crossing the quality chasm: A new health system for the 21st century.* Washington, DC: Committee on Quality of Health Care in America, Institute of Medicine, National Academy Press.

Journal of Medical Internet Research (JMIR). (2003). Retrieved from http://www.jmir.org

Panagariya, A. (2000). E-commerce, WTO and developing countries. *The World Economy, 23*(8), 959-978.

Porter, M. (1990). *The competitive advantage of nations.* New York: Free Press.

Protegrity. (2001). *Health Insurance Portability and Accountability Act (HIPPA) privacy compliance executive summary.* Author.

Puentes, J., Bali, R. K., Wickramasinghe, N., & Naguib, R. (2007). Telemedicine trends and challenges: A technology management perspective. *International Journal of Biomedical Engineering and Technology, 1*(1), 59-72.

Samiee, S. (1998). The Internet and international marketing: Is there a fit? *Journal of Interactive Marketing, 12*(4), 5-2.

Umhoff, B., & Winn, J. (1999, May 1). The healthcare profit pool: Who stands to gain and lose in the digital economy. *Health Forum Journal*.

Wickramasinghe, N. (2000). IS/IT as a tool to achieve goal alignment: A theoretical framework. *International Journal of Healthcare Technology Management, 2*(1/2/3/4), 163-180.

Wickramasinghe, N. (2007a). Critical success factors creating value driven e-business models in the Asia Pacific region. *International Journal of Services and Standards, 3*(2), 239-248.

Wickramasinghe, N. (2007b). Fostering knowledge assets in healthcare with the KMI model. International *Journal of Management and Enterprise Development, 4*(1), 52-65.

Wickramasinghe, N., Misra, S., Jenkins, A., & Vogel, D. (2006). The competitive forces facing e-health. *International Journal of Health Information Systems and Informatics, 1*(4), 68-81.

Wilson, T. (1999). Not a global village after all? Consumer behavior varies widely by country. *Internetweek, 792*, 13.

World Health Organization. (2003). Retrieved from http://www.emro.who.int/ehealth/

KEY TERMS

Efficiency: One of the promises of e-health is to increase efficiency in health care, thereby decreasing costs.

E-Health: It is health care delivery supported and enabled through the use of information systems and information technology, especially Web-based technologies.

Equity: This refers to making health care more equitable. In particular, some of the key issues for equity revolve around broad access and familiarity with the technology.

Evidence Based: E-health interventions should be evidence based in the sense that their effectiveness and efficiency should not be assumed but proven by rigorous scientific evaluation and support from case histories.

Framework: It is the conceptual structure used to solve a complex issue.

IT Infrastructure: It is the combination of hardware, software, networking, and telecommunications that forms the foundation for supporting IT capabilities in place.

Morbidity: This refers to either the incidence rate or to the prevalence rate of a disease.

Telemedicine: Telemedicine is the use of information systems and information technology to provide or facilitate the delivery of clinical-care evidence-based medicine. It also involves the application of uniform standards of evidence gained from previous cases to facilitate superior medical-practice outcomes.

C

Critical Trends, Tools, and Issues in Telecommunications

John H. Nugent
University of Dallas, USA

David Gordon
University of Dallas, USA

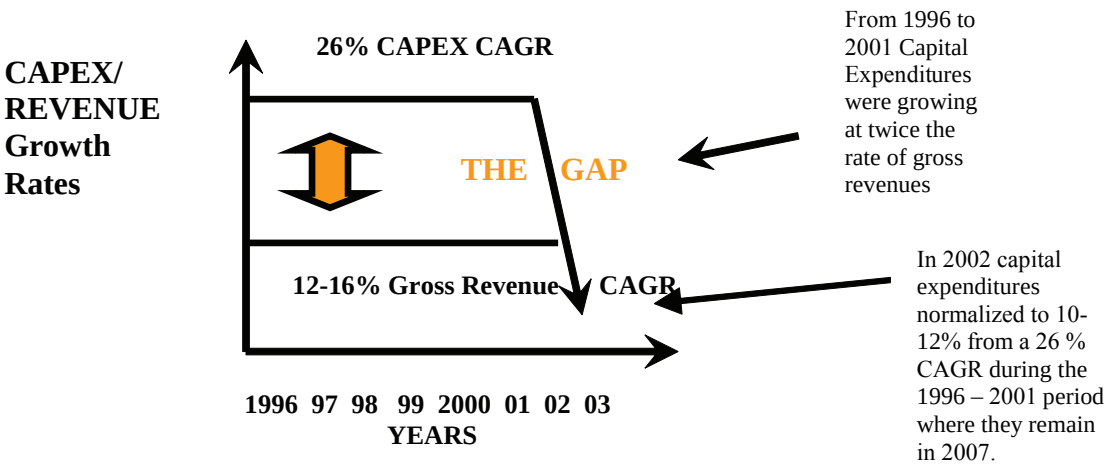
INTRODUCTION

As in all industries, in order to win in a market and set an appropriate strategy, it is important to know as much as possible about that market and have at one's disposal tools that will provide insight and competitive advantage when properly, collectively, consistently, and timely applied. This paper presents a series of powerful, but easy to use and understand, analytical and operational tools that deliver insight and competitive advantage to the telecommunications professional. Moreover, it should be stated that as with all good tools, the tools and models as presented herein transition across industry lines and are not limited to the telecommunications industry alone.

BACKGROUND

Starting in the 1990s, the telecommunications market appeared to experience unprecedented and unbounded growth with the advent of The Telecommunications Act of 1996. This growth was paralleled by a growth in capital equipment purchases (CAPEX) by network operators (see Figure 1). However, by the early 2000s, we saw a major market correction and the collapse of many firms that caught many industry professionals, bankers, and investors by surprise. The economic dislocations caused by the failure of so many telecommunications network providers were enormous. Hence, an examination was undertaken to see if tools and models existed that could provide significant insight into

Figure 1. The revenue capital expenditure growth rate comparisons



Source: Hilliard Consulting Group, Inc., 2007

changing market conditions. By examining these market dynamics and the fundamentals at play in the telecommunications space, it becomes apparent there are models and tools that provide insight as to the market's stage, and where it is likely to go next. Such a view is important to the investor, creditor, and operator alike in order to have a vision of the current and future market states so appropriate and timely decisions can be reached.

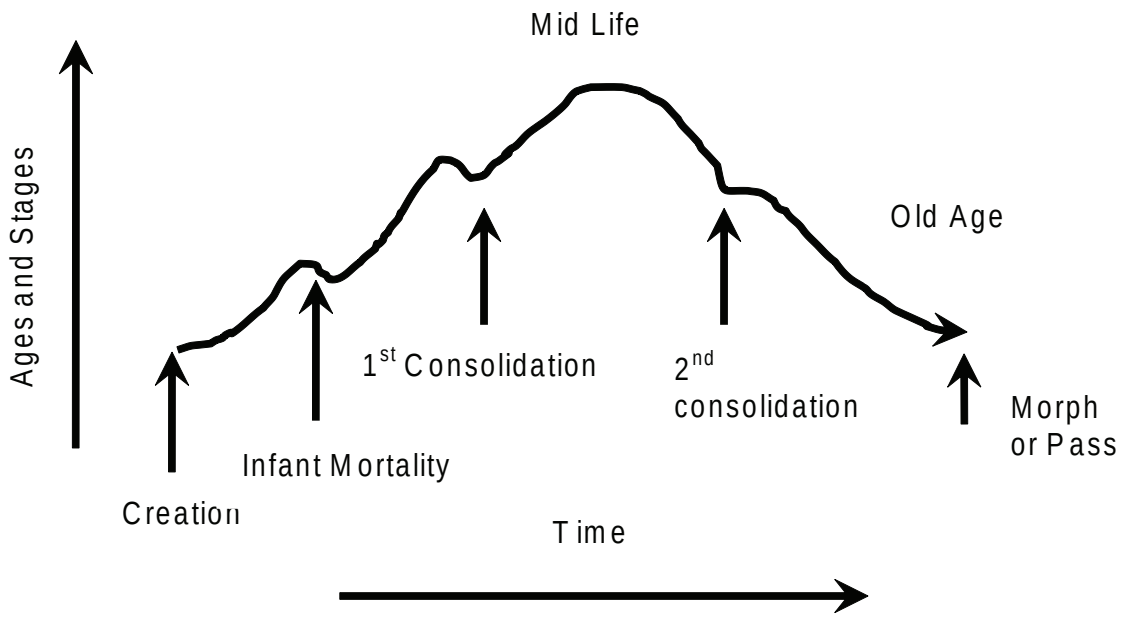
ANALYTICAL TOOLS AND MODELS

Because of the turmoil experienced in the telecommunications industry over the past decade, it is useful to view tools that can assist the telecommunications professional with understanding the market(s) and the trends at play. Looking at the telecommunications market from 1996 to 2007, it can be seen that the market exploded in the first half of this period with a 26% cumulative annual capital expenditure growth rate (CAPEX CAGR), collapsing in the latter part of this period (Hilliard, 2007; Lehman Brothers, 2000).

When capital expenditures so far outstrip the gross revenue growth rate, one knows this situation cannot continue unabated, and a return to a more normal state must take place. In order to discern approximately when a return to a more normal state will come about, one may examine the underlying market drivers (Nugent, 2001, 2003). Market drivers will often signal the size, breadth, and depth of a market.

Market Drivers: During the period of 1996-2003 several large drivers were evident. The first was identified as the Y2K driver. Here many firms determined it to be better, easier, and less costly and risky to replace versus remediate infrastructure equipment. But here it was known this driver would be satiated by 2000. A second major driver was The Telecommunications Act of 1996 (www.fcc.gov). This Act brought about the creation of many new telecom competitors that raised billions of dollars in the equity and debt markets that went on a spending spree. However, most of these firms had flawed business plans, and through competitive thrusts by the incumbents in the form of administrative delay, regulatory appeal, and litigation, these new entrants were literally bled dry via the consumption of cash in non-revenue producing activities such as regulatory appeals and litigation, and doomed to failure (Nugent, 2001, 2003). Understanding how significant incumbents fight and how they use the most strategic weapons of all – cash position and cash flow – the demise of these new incumbents could be foreseen.

Figure 2. The life cycle curve



Source: Hilliard Consulting Group, 2006

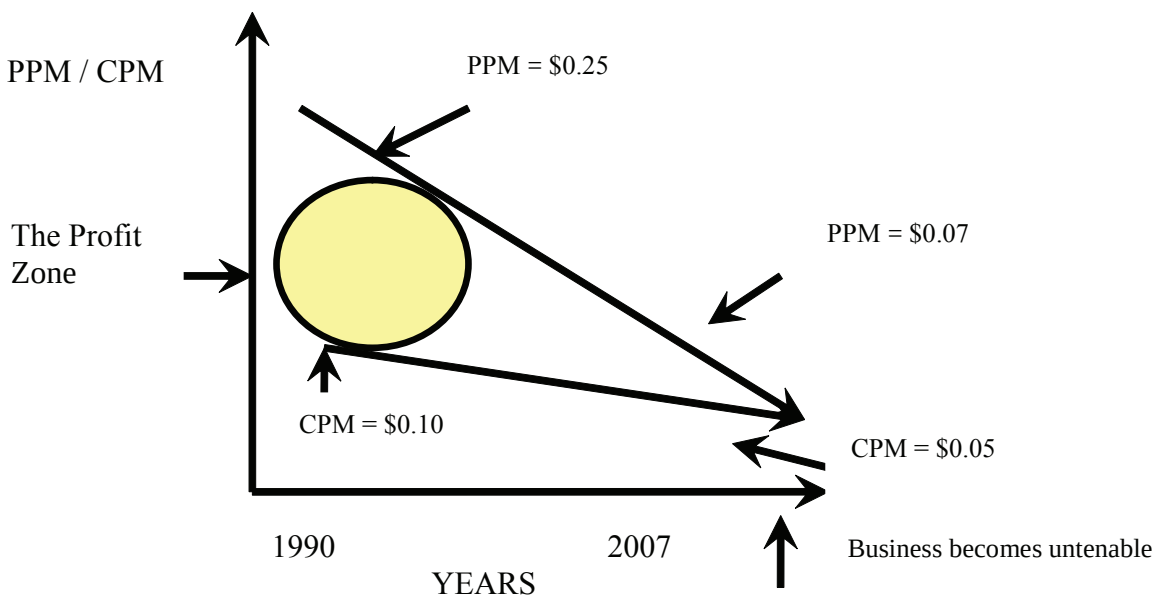
Another significant driver was the explosion in the number of wireless customers brought about by the "Digital One Rate" plan initiated by AT&T. Here wireless growth exploded from approximately 50 million subscribers to over 120 million in just several years. However, there are models that indicate this type of market satiates at approximately 50% of the overall population or 70% of the adult population (Nugent, 2003). In the United States, this satiation point is approximately 145 million narrowband voice subscribers – approximately where we are today. So this spending spurt on narrowband voice wireless Customer Premise Equipment (CPE) and infrastructure equipment could have also been estimated to end as the market approached satiation.

Hence, the telecommunications market downturn should not have been a surprise to anyone, as an understanding of the principal market drivers would have permitted an estimate of the market's size, breadth, depth, and duration.

Life Cycle Curve: Another important strategic understanding is the Life Cycle Curve. This Gaussian Curve (Bell Curve) is representative of all things, whether they are individuals, enterprises, nations, states, or civilizations. That is, each is created, grows rapidly, matures and passes. All such bodies, other than individuals, have the ability to change and adapt versus pass, but this almost never happens. Two companies that lasted longer than most, Stora Kopparbergs Bergslags AB, a Swedish trading company tracing its origins to the Middle Ages was acquired in 2002, and Kongo Gumi, in continuous operation since 560 AD, got into operating trouble in 2006 and was also acquired in November of that year. In fact, most of the leading companies last 40 years on average or less before the pass. This may be seen by looking back at any of the leading business lists of 40 years ago and seeing how many are still on the leading lists of today. Typically 80% of the entities on the first list will not appear on the second. Hence, an understanding of where one stands in its respective life cycle, as well as where its primary competitors, customers, and suppliers stand, is also important because leverage is changing all the time amongst this set. Moreover, to gain and sustain competitive advantage, one needs to constantly discern how and where to reallocate its assets. And the Life Cycle helps determine relative position.

The fundamentals of this model are that approximately 85+% of enterprises fail in the first five years of life, the Infant Mortality stage. The next two contractions represent major market consolidations (mergers and acquisitions). The first such consolidation takes place at approximately 50% market satiation; and the second consolidation at approximately 90+% market satiation. This understanding is important because it permits a right sizing of entity investments and business activity when the first market consolidation takes place, and provides clarity when one should exit its market

Figure 3. The minute margin squeeze model for the interexchange carrier (IXC) market



Key: PPM = Price Per Minute; CPM = Cost Per Minute Source: Hilliard Consulting Group, Inc. ©2007

if it cannot change the market. Moreover, it also hints that at and after the second consolidation point, an entity should only buy like kind assets at steep discounts versus large premiums. This is because in competitive markets at this point in time, dynamic growth and margins are gone.

Most recently we have seen several large 2nd Consolidation Stage purchases in the telecom arena where the industry is at the 90+% satiation point: AT&T's acquisition of Bell South and Cingular's acquisition of AT&T Wireless. In each case a large premium was paid for like kind assets late in the life cycle. This strategy is almost always costly in truly competitive markets, unless the enterprise can create a *de facto* duopoly where the two primary competitors no longer compete on price. The issue involves unit prices and unit costs, and the typical relationship between these functions. That is, as a function of time and competitive pressures, in competitive markets, unit prices decline faster than unit costs. A generic model of long distance unit prices and costs appears below:

This Unit Price Unit Cost Model above indicates that enterprise value, growth, and margins are greatest when the gap between unit prices and unit costs is greatest, The Profit Zone, and least when these functions (slopes) converge.

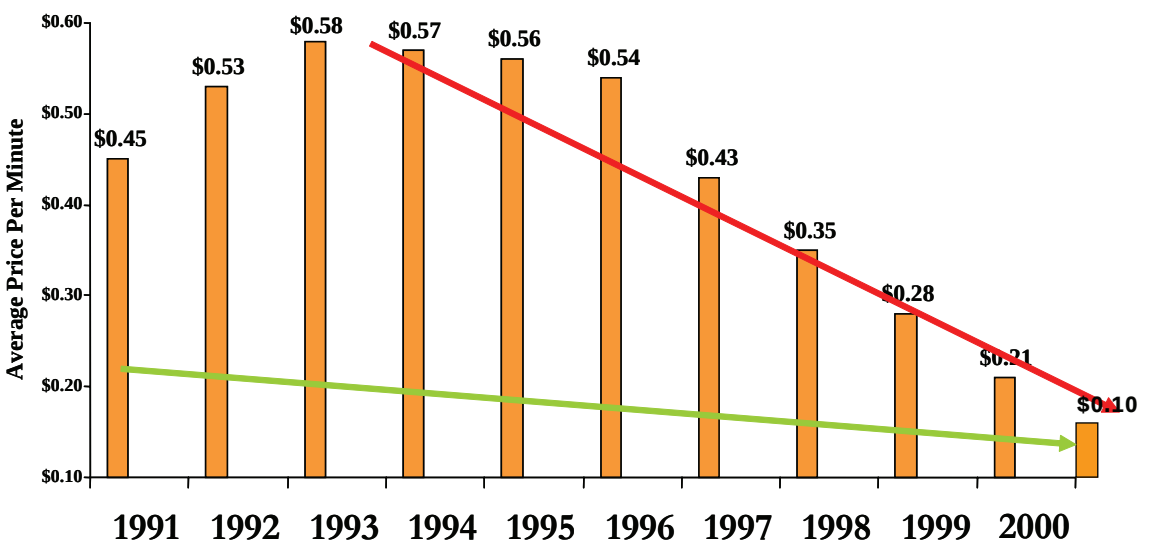
We see this same relationship in the United States narrowband wireless voice arena. Here, too, margin and growth are gone as the unit price/unit cost squeeze is on. Yet, we saw Cingular pay a large premium for AT&T Wireless late in life when margin and growth have abated. This will almost always create stress unless a *de facto* duopoly is created where the unit price degradation slope can be perverted as the two primary competitors no longer compete on price.

At a high level it is also important to understand where a market is today, and where it is going to be tomorrow. To help understand these conditions, a State, Gap, and Trend (SG&T) Analysis tool provides helpful insight (Hilliard, 2003; Wolford-Ulrich, 2004).

The development of a SG&T tool calls for a "one for one" transition (a "this to that" scenario over a period of time – there can be no ambiguities). Hence, a current and future state can be determined with some clarity.

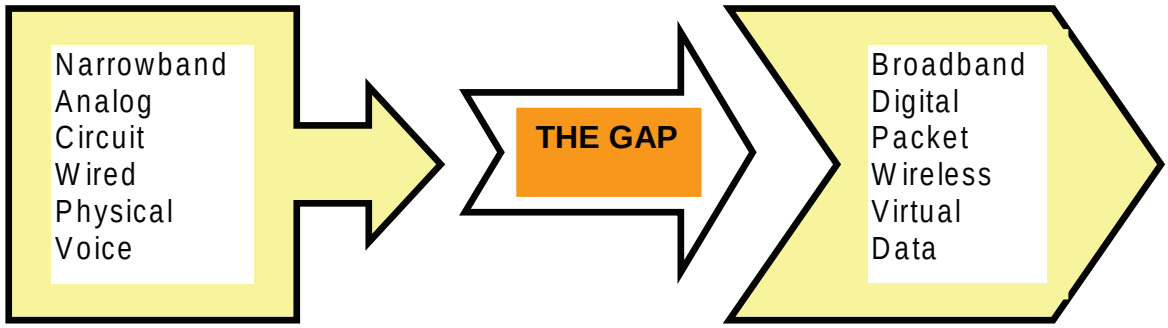
An examination of this SG&T tool presented in Figure 5 indicates that the telecommunications world is moving from a fixed, tethered, narrowband, analog, circuit-based world, to one principally comprised of mobile, wireless, broadband, digital, packet-based communications. This transition portends significant issues for land-based carriers whose assets principally are in big physical plant (central offices, switching facilities, tethered trunks and circuits, etc.). Yet in the two large telecom acquisitions mentioned above, we saw premiums being paid for yesterday's technology where markets are already satiated. This model further indicates that land-based carriers' assets are probably depreciating significantly faster than their balance sheets indicate. Supporting this premise is the decline in the number of residential landlines from approximately 168 million lines in 2001 to approximately 140 million residential landlines today (www.fcc.gov).

Figure 4. Wireless minute unit prices and costs



Source: FCC Annual Report on Wireless Industry, June 2001

Figure 5. State gap and trend analysis: Technology transition

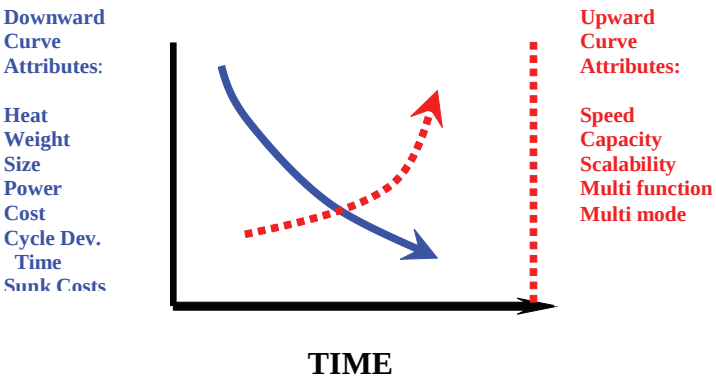


Source: Hilliard Consulting Group, Inc. © 2007

The issue of yesterday's assets and liabilities and the value shown in the financial statements will become apparent in the first quarter of 2008 when most SEC reporting companies must start reporting under FAS 157, Fair Value Accounting. Under this requirement, enterprises must disclose the difference in value from the value indicated on the balance sheet and their current value. For instance, Verizon carries on its books its wireless licenses at approximately $48 billion. In November, 2006 the FCC licensed approximately twice as much spectrum as Verizon holds for approximately $14 billion. This would indicate, under a market value approach, that Verizon's wireless licenses are worth significantly less than their carrying value by perhaps as much as $41 billion.

Moving from a macro model of market trends (SG&T) analysis in Figure 5, it can also be seen on a micro (tactical) level (Product Curve) what attributes successive telecommunications products must follow to win in future markets (Hilliard, 2006). Here, a Product Curve model is most helpful.

Figure 6. Product curve



Source: Hilliard Consulting Group, 2006

The Product Curve demonstrates that devices (network and CPE) need to become smaller, consume less power, weigh less, give off less heat, cost less, be developed in faster and faster cycle times, and have less in sunk development costs, while at the same time do more. They need to operate at faster speeds and higher capacities while performing more functions to win in future markets. The Product Curve also portends troubles for land-line carriers as it can be seen in not too many years, the central office of today will be displaced by a laptop wireless broadband tool of tomorrow. The SG&T Analysis and the Product Curve shown in Figures 5 and 6, respectively, only highlight some important attributes. There other numerous others that may, and should be added for a fuller comprehension of the industry.

To see the Product Curve in action, a comparison of the original Motorola "Brick" cell phone may be made with the sleek small wireless communication devices we use today. Here, we can see that the devices have become smaller, weigh less, consume less power, cost less, give off less heat, but do more. This model would have also clearly shown that Iridium and ICO had to fail because they each fought the attributes of the Product Curve: long cycle development times, high costs and prices, large bulky equipment, consumed a lot of power, gave off a lot of heat, etc. And, because the development cycle times were so long, alternate market winning technologies were developed that did follow the Product Curve attributes – GSM wireless solutions in the main.

Mix Shift Analyses are another way to discern important market changes (Hilliard, 2006). Here, many consulting firms forecast where we are today and likely will be tomorrow.

MIX SHIFTS

These Mix Shift analyses indicate that the telecommunications industry will move from a tethered to a wireless world in the relatively near future, while at the same time the mix of telecommunications traffic will shift from principally voice to principally data. This mix shift does not mean that voice traffic will decline; rather it indicates that data traffic will grow dramatically compared to voice over the period indicated. It is presently estimated that data is growing at 2+ exabytes a year (2 X 10 to the 18th power). Moreover, voice will largely become data as Voice over the Internet Protocol (VoIP) becomes more the norm. This trend is highlighted in the SG&A analysis above where we see a shift from circuit to packet, and narrowband to broadband communications. Hence, when we again look at the recent acquisitions mentioned herein, where yesterday's technology was principally acquired at a large premium, it is likely that the enterprise will be challenged to deal with this situation going forward.

OPERATIONAL TOOLS AND MODELS

Slope Analysis is the Converging/Diverging Gross Margin Analysis. Here, actual data from the Income Statement is plotted for several periods. Converging Gross Margins indicate increasing operational efficiency while diverging gross margins signify decreasing operational efficiency.

Table 1. Transport shift

Transport \ Year	2003	2008	2015
Landline	80%	50%	10%
Wireless	20%	50%	90%

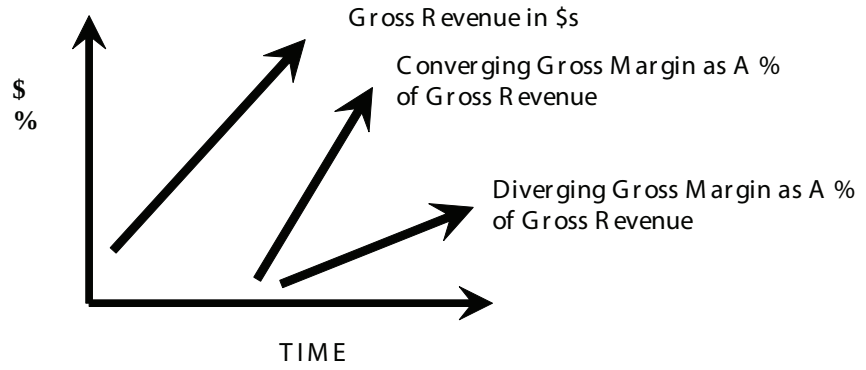
Source: Hilliard Consulting Group, Inc. 2006

Table 2. Mode shift

Service \ Year	1995	2020
Voice	90%	10%
Data	10%	90%

Source: Hilliard Consulting Group, Inc. 2006

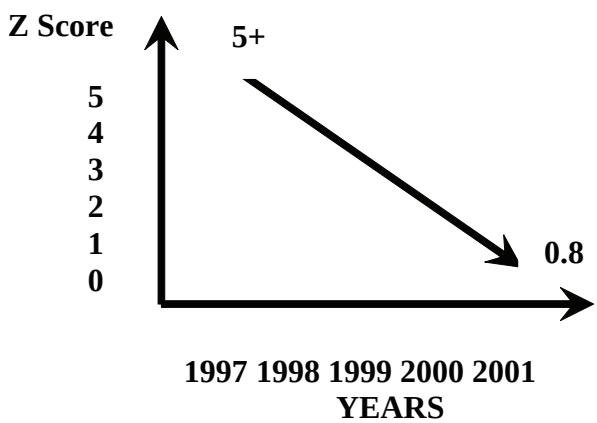
Figure 7. Converging/diverging gross margin analysis



Source: Hilliard Consulting Group, Inc. 2003

Examining AT&T's Initial Public Offering (IPO) for its wireless unit demonstrates the importance of this tool. The IPO for AT&T's wireless unit was well received and the stock price climbed immediately. However, a reading of the offering document would have shown that the Gross Margin decreased (diverged) by over 50% in the preceding annual period. This indicated significant operational troubles. By discerning this diverging gross margin and drilling down to determine the reasons, one would have discerned that AT&T's successful "Digital One Rate Offering" caught the company short, far short, of network capacity to support demand. Hence, AT&T had to go off-net and pay other carriers high fees to originate or terminate its traffic. This understanding would have highlighted additional issues facing AT&T Wireless: either take on more debt or dilute current shareholders further by issuing more stock in order to build more infrastructure.

Figure 8. AT&T Modified Z Score 1997 - 2001



Year	Score
1997	5+
1998	3.91
1999	1.87
2000	1.1
2001	.8

Source: Hilliard Consulting Group, Inc. 2006

One final operational tool is the Discriminant Function Algorithm (DFA) used to discern changes in corporate health on a prospective basis – Inflection Points. (Amdocs Corp., 2003; Slywotzky. Morrison, Moser, Mundt, & Quella 1999) This model uses Altman's Z Score algorithm for determining bankruptcy on a prospective basis (Altman, 1983). However, unlike Altman where he uses absolute scores, the DFA model only cares about changes in score - either positive or negative (Nugent, 2003). Moreover, the scale is changed in the DFA model from Altman's Bankruptcy Prediction model. As an example of what this tools yields on a prospective basis, AT&T is again viewed.

As can be seen, AT&T's Modified Z Score declined from over 5 in 1997 to under 1 by 2001. This is a dramatic negative decline. Yet, during much of this period, Wall Street was in love with this stock. Had one begun plotting this Z score in 1997, one could have discerned by 1999 that things were heading south long before others foresaw this decline. For instance, if Ericsson had been using this model in 1997 and beyond, it would have provided Ericsson early warning that it needed to re-address its marketing and sales strategy as its largest customer, AT&T, had a degrading corporate health.

FUTURE TRENDS

From our examination of the past and the tools that may be employed to glean a view of the future, it is apparent that major transitions are underway at the strategic level; namely, voice to data, circuit to packet, narrowband to broadband, wired to wireless, and physical to virtual. Such transitions will require new networks to support these changes. Moreover, at the product or tactical level, we see that new solutions will be required to be smaller, lighter, less costly, consume less power, give off less heat, weigh less, and be developed in shorter time cycles if they are to be successful in subsequent markets.

CONCLUSION

High level analytical and operational tools and models can assist the telecommunications professional in understanding the telecommunications market's characteristics, life cycles, trends, directions, limits, drivers, duration, and likely prospective performance. These tools demonstrate that wireless communications will follow the same life cycle characteristics as wired communications. Proper utilization of such tools collectively, consistently, and continually can lead to important and timely insights hopefully leading to competitive advantage based upon an early detection of changes in marketplace dynamics.

REFERENCES

Altman, Edward, I. (1983). *Corporate Financial Distress: A Complete Guide to Predicting, Avoiding, and Dealing with Bankruptc.y* John Wiley & Sons. Demonstrates the method by which to calculate declining financial condition via his Z Score Algorithm.

Amdocs Corporation (2003)., Demonstrates ages, stages and transitions in technology. *State, Gap & Trend Analysis*. www.amdocs.com

Argenti, J. (1976). *Corporate Collapse: Causes and Symptoms*. John Wiley & Sons. This book examines why enterprises fail.

AT&T Corporation. Provides many examples of strategic moves that create financial challenges going forward. www.att.com

FCC, Federal Communications Commission. Site provides market data on users and modes of communications. www.fcc.gov

Forbes. (Oct. 16, 2006), p.32. Article demonstrates the looming issues for major infrastructure companies relative to fair value reporting commencing in 2008.

Hilliard Consulting Group, Inc., McKinney, TX. Models first developed by Dr. J. Nugent, CPA, CFE, CISM, FCPA at the Hilliard Consulting Group, Inc. in 1999, and refined in 2001, and 2003.

Lehman Brothers (2000). Report on Capital Expenditures in the Telecommunications Industry. Study published for customers of the firm. Not publicly available.

Motorola Corporation. Presents a historical view of the Product Curve at work via a review of older compared to newer technologies. http://www.motorola.com

Nugent, J.H. (2001). *Telecom Downturn was No Surprise. Dallas Fort Worth TechBiz* , (September 10-18), p.22. Article highlights why all should have seen the return to a normal market stasis in the 2001 to 2002 period for the telecom industry. www.dfwtechbiz.com.

Nugent, J.H. (2003). *Plan to Win: Analytical and Operational Tools – Gaining Competitive Advantage*, McGraw-Hill, New York, NY, 2nd ed., 2003. Demonstrates in significantly more detail the application of the tools cited herein.

Slywotzky, A.J., & Morrison, D.J. (1997). *The Profit Zone*. Times Business, Random House. Book highlights the importance of operating in markets when the gap between unit prices and unit costs is largest, and why one should exit a market as these slopes converge if one cannot change that market.

Slywotzky, A.J., Morrison, D.J., Moser, T., Mundt, K.A., & Quella, J.A. (1999). *Profit Patterns*. Random House. Book highlights important trends in revenue generation.

The Telecommunications Act of 1996 – www.fcc.gov. This law was flawed in that it did not address how incumbents would take on new entrants. That is by administrative delay, regulatory appeal, and litigation.

The Wall Street Journal. (April 18, 2002). P. B5.

The Wall Street Journal. (November 2006). Article cited the acquisition of the longest continuously operating company by another. Kongo-Gumi in Osaka, Japan, had continuously operated since 560 AD, but got into operating trouble in 2006 and had to sell.

Wolford-Ulrich (2004). This piece discusses the major points of change during an enterprises existence. This piece used in conjunction with Altman's Modified Z Score analysis provides early warning of a growing Inflection point. www.inflectionpoints.com

KEY TERMS

CAGR: Cumulative Annual Growth Rate – The percent of growth from one annual period to the next.

CAPEX: Capital Equipment Expenditures usually measured as a percent of gross revenue.

Converging/Diverging Gross Margin Analysis: A Slope Analysis Tool used to plot actual sales and gross margin data for several periods in order to discern trends and likely outcomes. Measure operational efficiency or inefficiency.

CPE: Customer Premise Equipment (end-user equipment).

Discriminant Function Algorithm: A term developed for Edward Altman in describing his Z Score analytical tool in determining the likelihood of an enterprise going into bankruptcy on a prospective basis. Used as a method for determining inflection points – changes in corporate health versus for bankruptcy prediction.

Inflection Points: Significant changes in corporate performance.

IXC: Interexchange Carriers (long distance companies that transport inter Local Access Transport Area (LATA) traffic).

LATA: Local Access Transport Area – a geographic area defined by the FCC.

Minute Margin Squeeze: Also known as Unit Price/Unit Cost Model – see below.

Mix Shifts: Shifts in the market between major components usually requiring different technology or solutions.

Product Curve: This tool takes a micro view of transitioning requirements or attributes successive solutions it must adhere to in order to win in the future market place.

Slope Analysis: The plotting and visualization of certain operating functions in order to discern trends, often before others see them, thereby permitting alteration of strategies in order to gain competitive advantage.

State Gap & Trend Analysis: A tool used to present in a structured format current market or technology states as well as future states. This analysis requires a "one for one" transition – a "this to that" view. This model calls for no ambiguities. The perilous part of this tool is determining how to transition the Gap – where one has to be by when, with what.

Unit Price/Unit Cost Model: A Slope Analysis Tool used to plot actual unit prices and unit costs for several periods in order to discern future trends and likely outcomes.

Cross-Cultural Challenges for Information Resources Management

Wai K. Law
University of Guam, Guam

INTRODUCTION

Western organizations have led the globalization of business operations, especially in the deployment of multi-domestic strategy. The decentralized organizational control and customized operations support the fast penetration of huge global markets. Western management theory considers information the lifeblood of organization. The sharing of information lubricates the interlocking divisions within the organization, promoting the effective achievement of organizational goals with external business partners. However, in many regions of the world, information represents power, and managers often try to accumulate as much of it as they can while denying access to others (Oz, 2002). For others, the disclosure of information is considered a threat to the span of management control (Rocheleau, 1999). In some cases, administrators could be more interested in the scale of the information system and its associated budget, than the capability and functionality of the system (Kalpic & Boyd, 2000). These are examples of conflicting cultural values in a cross-cultural environment. The introduction of Western management approaches conflicts with regional administrative styles, diminishing the effectiveness of information systems (Raman & Watson, 1997; Shea & Lewis, 1996). Sensitivity to cultural differences has been recognized as an important factor in the successful global deployment of information systems. Minor information management issues potentially resolvable through improved communication in the west often manifest as major challenges in a cross-cultural environment.

BACKGROUND

The literature provided thorough coverage on designs, development, and implementation of computer-based information systems (CBIS). Numerous studies examined various systems-solutions for organization needs (McLeod, 1998; O'Brien, 2002). However, the projected value of information technology has been formulated based on a rough assessment of the possibilities without full appreciation of the limitations due to resistance to organizational and social changes (Osterman, 1991). Increasingly, management realized that massive deployment of information systems on a global basis, even with prudent management of the systems, has not been producing the desirable outcomes of value generation. Recent studies revealed the significant influence of cultures toward the success of transferring information technology beyond the Western world. National culture, organization culture, and MIS culture induced influence over the successful development and management of information resources (Hofstede, 1980; Raman & Watson, 1997). Shea and Lewis (1996) suggested the desirability of placing close attention to user absorptive rate in the transfer of new technology into a different cultural environment. It became apparent that adaptation of information system designs to new cultural environments was insufficient to guarantee successful implementation. User selection of technological features, driven by cultural preferences, could be a key factor for designing information systems in multi-cultural environments. Other studies reported the numerous obstacles of developing CBIS under various cultural settings, even with highly motivated leaders to support the deployment of information systems (Al-Abdul-Gader, 1999; Raman & Watson, 1997).

The information system function must enhance user effectiveness and efficiency in utilizing the information to improve value delivery for the organization. New challenges emerged as non-technical issues clouded the measurement of information system performance. A typical information system would be designed to provide information to users with common needs. Good data reports should contain all the required information with accurate representation of events. The reports needed to be generated in a timely fashion and in a format usable by the users (McLeod, 1998). However, individual users tended to value information systems for providing custom reports to meet individual needs in specific circumstances (Heeks, 1999). Inconsistent expectations in a cross-cultural environment crippled the effective management of information resources. Cultures carried different interpretations for timeliness, completeness, and relevancy of information.

Makeshift management decision generated new dynamics in several ways. In the spirit of promoting free information exchange, the department that owned the information system became obligated to provide information to others (Oz, 2002). However, the new responsibility seldom came with additional resources. The information owners became reluctant to supply information; doing so would take away

resource from other regular tasks (Davenport, 1997). Some managers shifted the data reporting responsibilities to other divisions, creating a bureaucratic nightmare for the users. Some ignored data requests, and others manipulated the data flows with respect to organizational politics (Oz, 2002; Rocheleau, 1999). Those working in the public sector faced the challenge of maintaining a delicate balance as they attempted to fulfill their responsibilities for both confidentiality and access to data (Duncan, 1999; Osterman, 1991). The problems would be more severe under a relationship-based culture where favors could not be declined.

Cultural backgrounds shaped the preferential information system model. In some cultures, managers would be intuitive, and feelings based, and have vague expectation for the performance of the information system. There would be more emphasis on group harmony and saving face than actual problem solving (Bjerke, 1999). Others would be more interested in meeting obligations, ignoring the source and validity of the reports. The controlling manager seeked a complex and broad information system providing qualitative data (Kluckhohn & Strodtbeck, 1961; Lane & DiStefano, 1992; Shea & Lewis, 1996). All these personality extremes co-exist in a cross-cultural setting, making it more challenging to design systems than a single culture environment. The perceived value of information resources became less predictable in cross-cultural environments.

CROSS-CULTURAL IRM CHALLENGES

The rapid expansion of Western influence on a global basis created an environment under the crosscurrents of Western corporate culture and regional cultures. In recent years, Western organizations have invested heavily in information technology (IT), turning it into an important tool, especially for the rapid expansion of business operations to global locations. Many organizations were surprised by the turbulence associated with the global deployment of information technology.

Management encountered new challenges as national workers joined the global team in serving customers from diversified cultural backgrounds. The national workers tended to hold on to their traditions, diluting the penetration of Western influence in the workplaces. The predominating regional workforce challenged Western corporate culture through their deep-rooted traditions and work habits. For example, a massive absenteeism could be expected on festival days, even without approved leaves or holidays. Timely arrival at a meeting could be accepted as up to several hours after the scheduled time. Mandated reports could be excused without penalty, and the uttermost concern to preserve group harmony over efficiency. Sometimes, this meant ignoring facts to restore stability and group harmony. Periodic acquisition of technology would be celebrated even without the appropriate infrastructure support, preventing the proper usage of the technology. Cross-cultural IRM issues emerged as significant challenges in cross-cultural environments.

Challenge 1: Information Resources Perception Challenges

Even as IT is transforming the world, a majority of the world's population still has limited understanding of information as a resource. The concepts of information resources escape the mind of even seasoned managers in the Western world. Potential cultural myopia requires great efforts to communicate the principles of information resources management.

Perception Challenge 1: Cultural Acceptance of IRM Practices

In a cultural environment that lacks appreciation for information resources, management must champion data planning and skillfully align the information support needs throughout the organization.

In cultures where gesture is more important than details, systematic failure to collect information would be accepted and forgiven. Information systems applications are often limited to payroll and accounting (Kalpic & Boyd, 2000). In some cases, the lack of adequate information is the key to assure continuing financial support, and the scale of the information system acquisition could be more important than the functionality. Organizations are unprepared to collect and store data to support meaningful decision support applications. A common pitfall is the underutilization of expensive information systems.

Perception Challenge 2: Information Resources Considered as Capital Expenditure

Funding is a necessary but insufficient condition for information resource development. The lack of proper organizational infrastructure dooms information system projects.

Deficiency in organizational data is often related to the lack of capital spending in IT resources. This perception underestimates the requirements for system analysis, data architecture development, data security and distribution, maintenance, technical support, and user training. Lack of organizational readiness stalls the deployment of information systems. Inflated expectation and uncoordinated usage of data services nullifies the value of the information systems. Erratic funding pattern destroys development projects, making it extremely difficult to retain technical personnel. Poor maintenance damages equipment and threatens data

integrity. Cultural managers eager to modernize without fully understanding the implications of information resources management eventually abandon their support for information resources.

Perception Challenge 3: Information Resources Development by Delegation

Information resources development without central coordination creates confusion and depresses the perceived value of information resources.

Management seeking an easy fix to the organizational data problem mandates data compilation by the functional divisions. Besides undermining information resources as a critical organizational asset, divisional managers tend to avoid the unfamiliar tasks by deferring to workers with little technical background. Unmotivated managers neglect data quality and resist data distribution. Administrator turnovers cause discontinuity in data resources development.

Perception Challenge 4: Information as a Freely Available Resource

The true value of information emerges from its effective distribution to end users. Information distribution is an expensive service.

Information resources can be compared to utility services such as water, the value of which relies on reliable and effective distribution with assured quality and sufficient supply. Often neglected is the accountability of the value contribution of information system, beyond periodic technical improvements. There is a need for benchmark studies to identify cost performance, as well as the critical roles of information resources within the organization.

Challenge 2: IT Transfer Challenges

The design objectives of information systems must expand from efficiency orientation to adaptive accommodation of cultural habits. It becomes desirable to allow and track dynamic modification of data processing procedures according to shifting organizational and cultural influences.

While a primary design objective of information systems is to facilitate efficient transaction processing, often the affected human system is slow to accept the implicit MIS culture embedded in the system design. Western cultures emphasize timeliness and accuracy, which are less important to other cultures (Kaufman-Scarborough & Lindquist, 1999; Straub, Loch, & Hill, 2001). For example, it often takes months to update databases from paper documents. Some users rely on the information system for information, while others insist on paper documents only. Hence circulation of multiple versions of reports is common depending on the sources of the reports. Parallel operations to accommodate parallel cultures generate organizational conflicts. Influential users and administrative interventions threaten the integrity of information systems. The full potential of information systems is suppressed in preference for cultural norms, and only system features that do not threaten cultural practices would be allowed to remain. Some local cultures put more emphasis on protecting family members than performance appraisal. The value of information is not as much for improving decision making, but to endorse group position, to preserve relationship, and to avoid embarrassment.

Challenge 3: Data Resources Management Challenges

Western culture encourages innovation, creativity, and the sharing of ideas and information. The same may not hold true under many cultures that adopt a historical perspective and value social stability to changes. Information resources could be jealously guarded in these cultures, and data distribution restricted.

Practical Challenge 1

There is a need for the clear definitions of data ownership and responsibilities for data acquisition, data quality control, and data distribution. This is especially challenging in cultural environments where the political attributes of information interfere with the communicative value of information.

In many Eastern cultures, credible information is deferred to leaders and elders with power and status. Political relationships dictate the availability of information and the accessibility to organizational data. This is in contrary to the basic assumptions of CBIS that promote the free exchange of information (Osterman, 1991; Oz, 2002; Rocheleau, 1999). The bureaucratic procedures for the approval of data usage defeat the designed roles of the information system. Fully developed database supports very limited applications. The lack of explicit system objectives coupled with the practice of delegating data management responsibility to the lowest level unskilled workers creates data integrity problems. For example, withholding information to gain and maintain power is acceptable among many Asian cultures. Openness would be considered a sign of weakness. It would be critical to formally establish the credibility, relevancy, and accessibility of data resources.

Practical Challenge 2

Management must meticulously plan data acquisition, data preparation, data distribution, and data usage, and fully understand the required organizational incentive and associated costs for maintaining information flow within the organization. This is especially important in cultural environments where data-driven decision-making is a new practice.

An uncoordinated approach to information resources management creates fragmented entities to process information for narrow applications. The fad of data-driven decision-making created a mad race for data reports using every available political connection. The result would be a great assortment of data reports with massive details. Inconsistency occurred among data reports depending on the data processing methods and storage formats. For example, a report from an off-line, static database in a remote office could be given equal credibility as a report generated from a current database from the data center. In a cross-cultural environment, influential individuals would compete to justify the merit of their reports from their cultural perspectives. The heated debates along with discrepancies among the reports frustrate the end users and lead to distrust toward the information systems for the inability to produce usable information reports. Regretfully, the information systems are seldom seriously designed for decision support.

Practical Challenge 3

Management must take leadership in establishing precise, formal data definitions, and communicate them to all potential data users, and those assigned roles in data distribution. This is especially important where mastery of languages, cultural predisposition, level of information literacy, and social attitude could strongly influence the group dynamic of data usage.

Technology evolution increasingly places information systems under the direct control of end users. However, end users often lack the technical expertise, and few are committed to the development of information resources. Events and samples are confused with statistics. Relaxed practices in standards and data definitions create issues in data validity and data quality. Potential information is lost when processed data replaces the raw data, while ignoring the time sensitivity of dynamic data. Time series data are deleted to preserve storage space. The information system is often blamed for the unfortunate chaos. For instance, technology facilitates individual users to maintain multiple copies of a database. Top management, unwilling to escalate cultural tension, ignored the potential seriousness of the data integrity issue. Improper database management practices yield different outcomes for identical requests for information from the different versions of the database. The absence of a single standard triggers disputes on the interpretation of data definitions according to the language understanding of the end users, while the actual data definition used by the data center to maintain the database is being rejected!

Challenge 4: Knowledge Sharing Challenges

Different cultures adopt different views toward knowledge, even among Western societies. Some cultures treat knowledge as an individual asset, a tradable resource. Intellectual property and usage rights become the main concern in knowledge sharing. Other cultures claim ownership of knowledge by the society and deny individual ownership of knowledge. A major concern would be censorship and control of information flow. Some cultures consider access to knowledge a privilege, given to individuals with the proper social ranks, while other cultures desire the widest distribution of knowledge to every member of the society. Information systems can be a threat to one society while high valued in another. Some cultures support openness, and others are conservative. Thus the emphasis could be on communication support in one society, while on preservation of traditions in another (Forstenlechner, 2005). Potential value clashes emerge in a cross-cultural setting when an individual raised in one social environment manages knowledge under a different social expectation. Knowledge sharing challenges are also expected with partnerships between organizations from different cultural background (Ford & Chan, 2003).

Challenge 5: Information Resources Accountability Challenges

As information resources emerge from a supportive role to become strategic assets, their management must also mature from the simple control measures for supplies. A greater challenge is to safeguard certain information resources as critical assets to be accessible only by trusted individuals.

Accountability Challenge 1

The increased complexity and frequency of usage of information reports is in reality a severe drain in budgetary resources, and management needs to develop a mechanism to track data usage and adjust resources appropriately. This could be more challenging under cultural environments that lack sophistication in information processing.

Modern management practices seeked opportunities to replace physical resources with information. When management failed to adjust budgets to support the in-

formation services, those affected would try every means to discontinue information services. On the other hand, uncontrolled access encouraged abuse, wasting valuable resources. Ethics, disciplined usage, and an understanding of information value supported the information practices in Western society. The problems would be crippling in cultures with different appreciation for information under different ethical standards. A local culture of generosity would insist on the free distribution of fully colored documents. Another practice has been the circulation of printed copies of e-mail to avoid offending anyone. These practices quickly deplete the budget for supplies.

Accountability Challenge 2

Management must take an active role in controlling the flow of organizational data, both within the organization and to the external environment. Management should consider endorsement of an official organizational data set to ensure consistency rather than leaving official data reports to random actions. This is especially important in cultural settings where it is impractical to correct public statements of social leaders regardless of facts.

In cultures where subordinates would not question the positions of leaders, information systems must implicitly support the decisions and public statements of the leaders (Gannon, 2001). In one example, officials of a local organization proposed an expensive marketing campaign pointing to decline in demand in the primary market. However, published data actually attributed the demand decline to the collapse of an emerging market. It would be an embarrassment to point out the omission, and the wrath of the society could be on those who allowed the facts to be publicized.

Accountability Challenge 3

Management must carefully orchestrate the deployment of information resources, with expected outcomes and supports. Strategic deployment rather than equal access should improve the value contribution of information resources.

While is it culturally sensitive to provide equal access to information resources, it is much more challenging to expect performance improvement with the availability of information resources. Management needs to carefully link value contribution to the deployment of information resources.

FUTURE TRENDS

The historical development of information systems has followed the model of a rational manager, with emphasis on openness, clear structure, innovative practices, and logical thinking. In regions where traditions and relationships resisted changes, information system designers must consider the needs of emotional decision-makers, with heavy emphasis on the concern to maintain social and cultural stability. Some cultures demand tight control of information flow, while other cultures are very casual about the absolute data quality. Some organizations integrate information systems as organizational backbone; others preferred to separate information in isolated pockets. Some prefer a simple information system, while others invest in a sophisticated intelligence system. Information systems for cross-cultural environments must deliver value to users with diversified backgrounds. Comparative study on information system features valued across cultural settings should improve the value delivery of the information system function.

CONCLUSION

Despite rapid technological development, information resource management is still a relatively new concept. Data reports preparation is often a laborious activity, and accepted practices and administrative preferences still drive decision-making. Organizations that anticipate increasing exposure to multi-cultural environments should allow longer time for organizational adjustment to technical development. Information systems originally developed as productivity tools for data processing, and report generation must undergo radical design evaluation to meet the diversified user expectations and information skills. Information resource managers must also carefully consider data ownership and data distribution issues. Cultural preferences and information values should be carefully considered in the justification of information services. Information system objectives should be clearly distinguished from information system capabilities, especially with different cultural interpretation of information value. Top management should play an active role in defining organizational data flow, with implementation of appropriate incentives. Special attention should be given to precise data definition, especially with a workforce with a different training background under different cultural and language settings. Lastly, it is critical to emphasize strict standards for data quality, due to differences in expectations for the information system performance.

REFERENCES

Al-Abdul-Gader, A. H. (1999). *Managing computer based information systems in developing countries: A cultural perspective*. Hershey, PA: Idea Group Publishing.

Bjerke, B. (1999). *Business leadership and culture: National management styles in the global economy*. Cheltenham, UK: Edward Elgar.

Davenport, T. (1997). *Information ecology: Mastering the information and knowledge environment*. New York: Oxford University Press.

Duncan, G. T. (1999). Managing information privacy and information access in the public sector. In G. D. Garson (Ed.), *Information technology and computer applications in public administration: Issues and trends* (pp. 99-117). Hershey, PA: Idea Group Publishing.

Ford, D. P. , & Chan, Y. E. (2003). Knowledge sharing in a multi-cultural setting: A case study. *Knowledge Management Research & Practice, 1*(1), 11-27.

Forstenlechner, I. (2005). The impact of national culture on KM metrics. *KM Review, 8*(3), 10.

Gannon, M. J. (2001). *Understanding global cultures: Metaphorical journeys through 23 nations*. Thousand Oaks, CA: Sage Publications.

Heeks, R. (1999). Management information systems in the public sector. In G. D. Garson, (Ed.), *Information technology and computer applications in public administration: Issues and trends* (pp. 157-173). Hershey, PA: Idea Group Publishing.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills, CA: Sage.

Kalpic, D., & Boyd, E. (2000). The politics of IRM: Lessons from Communism. In M. Khosrow-Pour (Ed.), *IT management in the 21st century* (pp. 72-73). Hershey, PA: Idea Group Publishing.

Kaufman-Scarborough, C., & Lindquist, J. D. (1999). Time management and polychronicity. *Journal of Managerial Psychology, 14*(3/4), 288-312.

Kluckhohn, F. R., & Strodtbeck, F. L. (1961). *Variations in value orientations.* New York: Row, Peterson and Company.

Lane, H., & DiStefano, J. (1992). *International organizational behavior.* Boston: PWS-Kent.

McLeod, R. M. Jr. (1998). *Management information systems*. Upper Saddle River, NJ: Prentice Hall.

O'Brien, J. A. (2002). *Management information systems: Managing information technology in the e-business enterprise.* New York: McGraw-Hill.

Osterman, P. (1991). Impact of IT on jobs and skills. In M. S. Scott Morton (Ed.), *The corporation of the 1990s: Information technology and organizational transformation* (pp. 220-243). New York: Oxford University Press.

Oz, E. (2002). *Management information systems*. Boston: Course Technology.

Raman, K. S., & Watson, R. T. (1997). National culture, information systems, and organizational implications. In P. C. Deans & K. R. Karwan (Eds.), *Global information systems and technology: Focus on the organization and its functional areas* (pp. 493-513). Hershey, PA: Idea Group Publishing.

Rocheleau, B. (1999). The political dimensions of information systems in public administration. In G. D. Garson (Ed.), *Information technology and computer applications in public administration: Issues and trends* (pp. 23-40). Hershey, PA: Idea Group Publishing.

Shea, T., & Lewis, D. (1996). The influence of national culture on management practices and information use in developing countries. In E. Szewczak & M. Khosrow-Pour (Eds.), *The human side of information technology management* (pp. 254-273). Hershey, PA: Idea Group Publishing.

Straub, D., Loch, K., & Hill, C. (2001). Transfer of information technology to developing countries: A test of cultural influence modeling in the Arab world. *Journal of Global Information Management, 9*(4), 6-28.

KEY TERMS

Cross-Cultural IRM: The special information resources management practices needed with the coexistence of more than one cultural influence in different segments of a society, or the simultaneous adoption of different cultural practices at work, social event, and family life.

Cultural Habit: Accepted behaviors within a group of people, sharing some common backgrounds, such as language, family heritage, education, living, and socializing environment.

Data Definition: An elaborate statement of the representation of each piece of data, its source, storage method, and intended usage.

Data Resources Management: The acquisition, organization, protection, maintenance, and selective distribution of organizational data.

Data-Planning: The projection of expected future need for data, with specifications on data sources, data collection

and storage, data processing and presentation, data distribution, and data security.

Information Resources: Resources required to produce information, including hardware, software, technical support, users, facilities, data systems, and data.

Information Resources Accountability: The activities related to tracking information resources usages, evaluation of information resources value contribution, and the monitoring of access to critical information resources.

IT Transfer: The introduction of an information communication technology to a new region of the world.

Knowledge Sharing: The activities relating to the exchange of meaningful information, along with interpretations and potential applications of the information.

Cross-Cultural Research in MIS

Elena Karahanna
University of Georgia, USA

Roberto Evaristo
University of Illinois, Chicago, USA

Mark Srite
University of Wisconsin-Milwaukee, USA

INTRODUCTION AND BACKGROUND

"Globalization of business highlights the need to understand the management of organizations that span different nations and cultures" (Srite et al., 2003, p. 31). In these multinational and transcultural organizations, there is a growing call for utilizing information technology (IT) to achieve efficiencies, coordination, and communication. However, cultural differences between countries may have an impact on the effectiveness and efficiency of IT deployment. Despite its importance, the effect of cultural factors has received limited attention from information systems' (IS) researchers. In a review of cross-cultural research specifically focused on the MIS area (Evaristo, Karahanna, & Srite, 2000), a very limited number of studies were found that could be classified as cross-cultural. Additionally, even though many of the studies found provided useful insights, raised interesting questions, and generally contributed toward the advancement of the state of the art in its field, with few exceptions, no study specifically addressed equivalency issues central to measurement in cross-cultural research. It is this methodological issue of equivalency that is the focus of this article.

METHODOLOGICAL ISSUES

Methodological considerations are of the utmost importance to cross-cultural studies, because valid comparisons require cross-culturally equivalent research instruments, data collection procedures, research sites, and respondents. Ensuring equivalency is an essential element of cross-cultural studies and is necessary to avoid confounds and contaminating effects of various extraneous elements.

Cross-cultural research has some unique methodological idiosyncrasies that are not pertinent to intracultural research. One characteristic that typifies cross-cultural studies is their comparative nature, i.e., they involve a comparison across two or more separate cultures on a focal phenomenon. Any observed differences across cultures give rise to many alternative explanations. Particularly when results are different than expected (e.g., no statistical significance, factor analysis items do not load as expected, or reliability assessment is low), researchers may question whether results are true differences due to culture or merely measurement artifacts (Mullen, 1995).

Methodological considerations in carrying out cross-cultural research attempt to rule out alternative explanations for these differences and enhance the interpretability of results (van de Vijver & Leung, 1997). Clearly, the choice and appropriateness of the methodology can make a difference in any research endeavor. In cross-cultural research, however, one could go to the extreme of classifying this as one of the most critical decisions. In this section, we briefly review such cross-cultural methodological considerations. Specifically, this section will address equivalence (Hui & Triandis, 1985; Poortinga, 1989; Mullen, 1995) and bias (Poortinga & van de Vijver, 1987; van de Vijver & Leung, 1997; van de Vijver & Poortinga, 1997) as key methodological concerns inherent in cross-cultural research. Then, sampling, wording, and translation are discussed as important means of overcoming some identified biases.

Equivalence

Achieving cross-cultural equivalence is an essential prerequisite in ensuring valid cross-cultural comparisons. Equivalence cannot be assumed a priori. Each cross-cultural study needs to establish cross-cultural equivalence. As such, equivalence has been extensively discussed in cross-cultural research, albeit using different terms to describe the phenomenon (Mullen, 1995; Poortinga, 1989).

To alleviate confusion created by the multiplicity of concepts and terms used to describe different but somewhat overlapping aspects of equivalence, Hui and Triandis (1985) integrated prior research into a summary framework that consists of four levels of equivalence: conceptual/functional equivalence, equivalence in construct operationalization, item equivalence, and scalar equivalence. Even though each level of equivalence is a prerequisite for the subsequent levels, in practice, the distinction between adjacent levels

of equivalence often becomes blurry. Nonetheless, the objective in cross-cultural research is to achieve all four types of equivalence. Hui and Triandis' (1985) four levels of equivalence are discussed as follows:

1. *Conceptual/functional equivalence* is the first requirement for cross-cultural comparisons and refers to whether a given construct has similar meaning across cultures. Furthermore, to be functionally equivalent, the construct should be embedded in the same nomological network of antecedents, consequents, and correlates across cultures. For instance, workers from different cultures may rate "supervisor is considerate" as a very important characteristic; however, the meaning of "considerate" may vary considerably across cultures (Hoecklin, 1994).
2. *Equivalence in construct operationalization* refers to whether a construct is manifested and operationalized the same way across cultures. Not only should the construct be operationalized using the same procedure across cultures, but the operationalization should also be equally meaningful.
3. *Item equivalence* refers to whether identical instruments are used to measure the constructs across cultures. This is necessary if the cultures are to be numerically compared.
4. *Scalar equivalence* (or full score comparability; see van de Vijver and Leung, 1997) occurs if the instrument has achieved all prior levels of equivalence, and the construct is measured on the same metric. This implies that "a numerical value on the scale refers to same degree, intensity, or magnitude of the construct regardless of the population of which the respondent is a member" (Hui & Triandis, 1985, p. 135).

Bias: Sources, Detection, and Prevention

To achieve equivalence, one has to first identify and understand factors that may introduce biases in cross-cultural comparisons. Van de Vijner and Poortinga (1997) described three different types of biases: construct bias, method bias, and item bias:

1. *Construct bias* occurs when a construct measured is not equivalent across cultures both at a conceptual level and at an operational level. This can result from different definitions of the construct across cultures, lack of overlap in the behaviors associated with a construct [e.g., behaviors associated with being a good son or daughter (filial piety) vary across cultures], poor sampling of relevant behaviors to be represented by items on instruments, and incomplete coverage of the construct (van de Vijver & Leung, 1997). Construct bias can lead to lack of conceptual/functional equivalence and lack of equivalence in construct operationalization.
2. *Method bias* refers to bias in the scores on an instrument that can arise from characteristics of an instrument or its administration (van de Vijver & Leung, 1997), which results in subjects across cultures not responding to measurement scales in the same manner (Mullen, 1995). Method bias gives rise to concerns about the internal validity of the study. One source of method bias is sample inequivalency in terms of demographics, educational experience, organizational position, etc. Other method bias concerns relate to differential social desirability of responses (Ross & Mirowsky, 1984) and inconsistent scoring across populations (termed "selection-instrumentation effects" by Cook and Campbell, 1979, p. 53). For instance, on Likert scales, Koreans tend to avoid extremes and prefer to respond using the midpoints on the scales (Lee & Green, 1991), while Hispanics tend to choose extremes (Hui & Triandis, 1985). Differential scoring methods may also arise if respondents from a particular culture or country are not familiar with the type of instrument being used.
3. *Item bias* refers to measurement artifacts. These can arise from poor item translation, complex wording of items, or items inappropriate for a cultural context. Consequently, item bias is best prevented through careful attention to these issues. Like method bias, item bias can influence conceptual/functional equivalence, equivalence of operationalization, and item equivalence.

Table 1 presents a summary of how the three types of bias can be prevented or detected. The next section discusses three important methods of bias prevention: sampling, wording, and translation. This article concludes by presenting a set of cross-cultural methodological guidelines derived by a committee of international scholars.

Sampling

Sampling decisions in cross-cultural studies involve two distinct levels: sampling of cultures and sampling of subjects (van de Vijver & Leung, 1997). Sampling of cultures involves decisions associated with selecting the cultures to be compared in the study. Many studies involve a convenience sample of cultures, typically ones where the researcher has preestablished contacts. Even though this strategy reduces the considerable costs of conducting cross-cultural research, it may hinder interpretability of results, particularly when no differences are observed across cultures (van de Vijver & Leung, 1997). Systematic sampling of cultures, on the other hand, identifies cultures based on theoretical considerations.

Table 1. Types of bias, prevention, and detection

	Detection	Prevention
Construct bias (focus: constructs)	• Informants describe construct and associated behaviors • Factor analysis • Multidimensional scaling • Simultaneous confirmatory factor analysis in several populations • Comparison of correlation matrices • Nomological network	• Informants describe construct and associated behaviors in each culture
Method bias (focus: administration procedures)	• Repeated administration of instrument • Method triangulation • Monomethod-multitrait matrix	• Sampling (matching, statistical controls) • Identical physical conditions of administering the instrument • Unambiguous communication between interviewer and interviewee • Ensured familiarity with the stimuli used in the study
Item bias (focus: operationalization)	• Analysis of variance • Item response theory • Delta plots • Standardized *p*-difference • Mantel–Haenszel procedure • Alternating least squares optimal scaling • Multiple group LISREL	• Wording • Translation

Typically, this involves selecting cultures that are at different points along a theoretical continuum, such as a cultural dimension. Random sampling of cultures involves selection of a large number of cultures randomly and allows for wider generalizability of results.

Most cross-cultural studies discussing sampling considerations, however, refer to sampling of subjects. Ensuring sample equivalency is an important methodological consideration in cross-cultural research, and it refers to the inclusion of subjects that are similar on demographic, educational, and socioeconomic characteristics. Sample equivalency can be achieved by either matching subjects across groups based on these background variables or statistically controlling for the differences by including such demographic variables as covariates in the cross-cultural comparisons (van de Vijver & Leung, 1997).

Wording and Translation

This is one of the key problems in cross-cultural methods, because in most cases, different cultures also have different languages. Even in cases when subjects from different countries are conversant with English, they may miss the nuances of the intended meanings in questionnaire items (e.g., British, Canadian, and American English all have unique terms). Researchers should ensure that measurement instruments keep the same meanings after translation. Moreover, a given latent construct should be measured by the same questionnaire items in different populations. In fact, researchers such as Irvine and Carrol (1980) made a convincing case for using factor-matching procedures to test for invariance of factor structures across groups before any quantitative analysis is performed.

To translate correctly, there is a need to translate to the target language. This needs to be performed by a native speaker of the target language. Then, the text must be back-translated to the original language, this time by a different native speaker of the original language. Brislin (1986) provided fairly complete guidelines for this process.

IMPLICATIONS

Judging by the issues described above, achieving cross-cultural equivalence is straightforward. However, it is also clear that many precautions can be taken to prevent construct, method, and item bias and thus increase the level of equivalence. These range from sampling, wording, and translation, to careful attention to administration procedures across cultures. A number of guidelines for cross-cultural research have been put forth by an international committee of scholars (Hambleton, 1994; van de Vijver & Hambleton, 1996). Even though the primary focus of these is on research on psychological and educational issues, these guidelines easily generalize to MIS research.

In addition to prevention, various statistical tools can assist in the detection of the various types of biases. In summary, similar patterns of functional relationships among variables need to be shown (Triandis, 1976). Moreover, multimethod measurement can help us to avoid the confound between the interaction of the method and groups studied and is unlikely to share the same statistical underlying assumptions, or even require strong conceptualization ability (Hui & Triandis, 1985). This idea is similar to the notions of multiple operationism and conceptual replication (Campbell & Fiske 1959). Hui and Triandis (1985) claimed that this may not be as difficult as one may think, as long as there is prior planning of research. As an example, Hui and Triandis (1985) mentioned that an instrument may be improved by proper translation techniques, and:

…then establish conceptual/functional equivalence as well as instrument equivalence by the nomological network method and by examination of internal structure congruence. After that, the response pattern method and regressional methods can be used to test item equivalence and scalar equivalence. (p. 149)

The major implication of methodological problems is complications in making valid inferences from cross-cultural data. Clearly, there are many problems with correctly inferring from data in a cross-cultural research project and attributing results to true cross-cultural differences. To do so, alternative explanations need to be ruled out. Establishing (and not merely assuming) the four levels of cross-cultural equivalence previously discussed in this article is a major step in this direction.

FUTURE DIRECTIONS AND CONCLUSION

Initial attempts at reviews of cross-cultural research in MIS (Evaristo, Karahanna, & Srite, 2000) show that, for the most part, MIS studies have refrained from testing theories across cultures, and when comparisons are made, they are often post hoc comparisons utilizing data from prior published studies in other countries. Clearly, this provides some insights into differences across cultures but suffers from a number of methodological shortcomings. In fact, the conclusions of Evaristo, Karahanna, and Srite (2000) were as follows:

In summary, we suggest that there are mainly three points where the MIS cross-cultural research is lacking: lack of theory base (testing or building); inclusion of culture as antecedents of constructs, and general improvement in methodologies used.

All three points are related, although to different extents, to methodological issues. The conclusion is that one critical issue that cross-cultural research in MIS needs to address before reaching the same level of sophistication and quality already attained by mainstream MIS research is to attend to methodological concerns. The current article is a step ahead in this direction and sets the stage for future research.

REFERENCES

Brislin, R. (1986). The wording and translation of research instruments. In *Field methods in cross-cultural research*. Lonner and Berry.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.

Evaristo, J. R., Karahanna, E., & Srite, M. (2000). *Cross-cultural research in MIS: A review*. Global Information Technology Management Conference, Memphis, TN.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10*, 229–244.

Hui, H., & Triandis, H. (1985). Measurement in cross-cultural psychology. *Journal of Cross-Cultural Psychology, 16*(2), 131–152.

Irvine, S., & Carrol, W. (1980). Testing and assessment across cultures: Issues in methodology and theory. In H. C. Triandis & W. Lonner (Eds.), *Handbook of cross-cultural psychology: Methodology* (pp. 181–244). Boston, MA: Allyn and Bacon.

Lee, C., & Green, R. (1991). Cross-cultural examination of the Fishbein Behavioral Intentions Model. *Journal of International Business Studies, 22*(2), 289–305.

Mullen, M. (1995). Diagnosing measurement equivalence in cross-national research. *Journal of International Business Studies,* (Third quarter), 573–596.

Poortinga, Y. H. (1989). Equivalence in cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*, 737–756.

Poortinga, Y. H., & van de Vijver, F. (1987). Explaining cross-cultural differences: Bias analysis and beyond. *Journal of Cross-Cultural Psychology, 18*(3), 259–282.

Ross, C. E., & Mirowsky, J. (1984). Socially desirable response and acquiescence in cross-cultural survey of mental health. *Journal of Health and Social Behavior, 25*, 189–197.

Srite, M., Straub, D., Loch, K., Evaristo, R., & Karahanna, E. (2003). Inquiry into definitions of culture in IT studies. In F. Tan (Ed.), *Advanced topics in global information management* (Vol. 2). Hershey, PA: Idea Group Publishing.

Triandis, H. (1976). Methodological problems of comparative research. *International Journal of Psychology, 11*(3), 155–159.

van de Vijver, F., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1*, 89–99.

van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.

van de Vijver, F., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(21–29).

KEY TERMS

Conceptual/Functional Equivalence: Refers to whether a given construct has similar meaning across cultures.

Construct Bias: Occurs when a construct measured is not equivalent across cultures both at a conceptual level and at an operational level.

Equivalence in Construct Operationalization: Refers to whether a construct is manifested and operationalized the same way across cultures.

Item Bias: Refers to measurement artifacts.

Item Equivalence: Refers to whether identical instruments are used to measure the constructs across cultures.

Method Bias: Refers to when subjects across cultures do not respond to measurement scales in the same manner. Bias in the scores on an instrument can arise due to characteristics of the instrument or its administration.

Multinational Corporation: A firm that has operations in multiple countries.

Scalar Equivalence: Refers to whether the construct is measured on the same metric across cultures. This occurs if the instrument has achieved all prior levels of equivalence, and the construct is measured on the same metric.

Transcultural Organization: A firm that operates across multiple cultures.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 59-63, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: IT Education

Cultural Diversity in Collaborative Learning Systems

Yingqin Zhong
National University of Singapore, Singapore

John Lim
National University of Singapore, Singapore

INTRODUCTION

Globalization makes cultural diversity a pertinent factor in e-learning, as distributed learning teams with mixed cultural backgrounds become commonplace in most e-learning programs, which can be study-based (schools and universities) or work-based (training units) (Zhang & Zhou, 2003). In these programs, collaborative learning is supported via computer-mediated communication technologies and instructional technologies. The primary goal of enhancing learning with technology aids, aligning with the goal of education at all levels, is to engage students in meaningful learning activities, which require learners to construct knowledge by actively interpreting, acquiring, and analyzing their experience (Alavi, Marakas, & Yoo, 2002). In accordance, meaningful learning requires knowledge to be constructed by the learners but not by the teachers. In this regard, collaborative learning, an activity where two or more people work together to create meaning, explore a topic, or improve skills, is considered superior to other individualistic instructional methods (Lerouge, Blanton, & Kittner, 2004). The basic premise underlying this is the socio-learning theory, which advocates that learning and development occur during cooperative socialization among peers and emerge through shared understandings (Leidner & Jarvenpaa, 1995). This highlights the criticality of the communication and collaboration pertaining to an individual's learning process. Since culture reflects the way one learns (Hofstede, 1997; Vygotsky, 1978), group members' cultural backgrounds play a significant role in affecting the collaborative learning process (Chang & Lim, 2005). Language, cognitive style, and learning style are some aspects of culture that concern collaborative learning in the short term.

Groups which have members of different cultural backgrounds are expected to be availed a wider variety of skills, information, and experiences that could potentially improve the quality of collaborative learning (Rich, 1997). In contrast, a group comprising members of similar backgrounds is vulnerable to the "groupthink" syndrome; when the syndrome operates, members could ignore alternatives, resulting in a deterioration of efficiency in making a group decision (Janis, 1982). Accordingly, it is conceivable that groups formed by members of different cultural backgrounds are inherently less prone to the "groupthink" syndrome. However, the advantages of cultural diversity in achieving meaningful collaborative learning are not easily realized, as the basic modes of communication may vary among different cultures and, in consequence, communication distortion often occurs (Chidambaram, 1992). Collaborative learning systems (CLS) are being increasingly researched owing to their potential capabilities and the associated new opportunities in supporting collaborative learning, in particular for distributed groups involving members of different cultural backgrounds (Alavi & Leidner, 2001). Collaborative learning systems provide the necessary medium to support interaction among learners, and therefore modify the nature and the efficiency of the collaborative learning activities (Mandryk, Inkepn, Bilezikjian, Klemmer, & Landay, 2001). The current article looks into how collaborative learning systems may better accommodate cultural diversity in e-learning groups. In addition, this article discusses pertinent issues regarding the role of a leader in building the common ground among learners in order to maximize the potential of collaborative learning systems when cultural diversity is present.

BACKGROUND

Collaborative learning is superior to individualistic instruction in terms of increase in individual achievement, positive changes in social attitudes, and general enhancement of motivation to learn, among other positive outcomes (Slavin, 1990). Learners tend to generate higher-level reasoning strategies, a greater diversity of ideas and procedures, more critical thinking, more creative responses, and better long-term retention when they are actively learning in collaborative learning groups than when they are learning individually or competitively (Schlechter, 1990). Growing interest in supporting the needs of collaborative learning, boosted by concurrent improvements in both computer mediated communication (CMC) and group support systems (GSS), has led to the emergence of the instructional technology known as collaborative learning systems. These are systems

implemented to provide computer-supported environments which facilitate collaborative learning. The importance of these systems lies fundamentally in their being a medium through which learners can cooperate with others.

Technology shapes the communication among users in terms of five media characteristics: symbol variety, parallelism, rehearsability, reprocessability, and immediacy of feedback (Dennis & Valacich, 1999). Symbol variety refers to the bandwidth that information can be communicated; parallelism is the number of concurrent conversations that a medium can support; rehearsability is the capability enabling users to modify a message before sending; reprocessability refers to the extent to which messages sent can be reprocessed during the communication; immediacy of feedback indicates whether a medium supports spontaneous feedback. In comparing collaborative learning systems and face-to-face setting in terms of three media characteristics—parallelism, rehearsability and reprocessability—the former outperforms the latter by embedding anonymity, text recording, and multiple access features; in terms of the other two media characteristics, symbol variety and immediacy of feedback, the situation is reversed (Dennis & Valacich, 1999).

Feather (1999) suggests that individuals will prefer learning in the virtual environment if they require more time to think about a question before answering, find it hard to speak out in a traditional class albeit possessing contributions, or like some degree of anonymity. Empirical evidence demonstrates that computer-mediated cooperative learning tended to have positive impacts on learners' performance and autonomy in controlling their learning pace (Salovaara, 2005; Yu, 2001).

MAIN THRUST OF THE ARTICLE

Potential of Collaborative Learning System in Accommodating Cultural Diversity

Culture is defined as the collective programming of the mind which makes the inhabitants of one country distinguishable from another (Hofstede, 1997). A heterogeneous group is one whose members are of different (national) cultural backgrounds while a homogeneous group has members of the same (national) cultural background. Hofstede (1997) has suggested four main cultural dimensions: individualism-collectivism, power distance, uncertainty avoidance, and masculinity-femininity. Hofstede's theory entails major cultural dimensions and seeks to explain the underlying causes of dissimilar behaviors in communication; indeed, different group behaviors are noted between heterogeneous and homogeneous groups (Stephan & Stephan, 2001). Members in an individualistic culture generally prefer loose ties with other peers during the collaboration process. In contrast, members in a collectivistic culture are typically more concerned with the common goal of the group and tend to prefer to work together.

A potential benefit of the collaborative learning systems is the support of diverse learning styles (Wang, Hinn, & Kanfer, 2001). Functions embedded in collaborative learning systems can enable more effective collaborative learning activities in heterogeneous groups by smoothing the communication process. In the face-to-face setting without technology aid, learners may feel the need to wait for others to express their ideas, by which time they may have either forgotten their own ideas or become less confident with these ideas; this phenomenon is called production blocking. Through embedding concurrent inputs by multiple users, collaborative learning systems offer a unique opportunity to eliminate production blocking, particularly as group size increases (Valacich, Jessup, Dennis, & Nunamaker, 1992). Moreover, text-based communication in these systems offers important features for communication that are radically different from the face-to-face setting. Group members' comments are recorded as text and they can be revisited repeatedly; such a feature is expected to enhance learning effectiveness as compared to oral communication, especially for non-native speakers, since no speaking has to take place (Herring, 1999). The communication support in collaborative learning systems has been suggested to be an effective tool in dealing with the lack of peer interaction in the classroom (Li, 2002). The underlying reason is that participation becomes more evenly distributed among members with computer-mediated interaction, while status and hierarchical structures become less important (Laughlin, Chandler, Shupe, Magley, & Hulbert, 1995).

Besides the communication difficulty mentioned previously, learners' uncertainty and anxiety form another challenge posted by cultural diversity in the face-to-face setting. In the absence of technological aid, when team members interact in the course of collaboration, uncertainty and anxiety of being in a heterogeneous group are likely to affect learners' communication with one another (Gudykunst, 1995), thus decreasing their performance. However, owing to the differences in communication process (as compared to face-to-face interaction), the rehearsability and the relatively lower degree of social presence embedded in collaborative learning systems are able to help the communication process in heterogeneous groups by lowering members' uncertainty and anxiety (Young, 2003). Therefore, the negative effects of cultural differences are reduced if not altogether eliminated by computer-aided systems, as learners of different cultures gain more accurate understanding of one another. Notwithstanding this, the diversity in terms of cultural values and experiences, earlier argued to be a strength, is not eroded. Also, the systems do not take the heterogeneous groups back to the "groupthink" situation which is more commonly present

in homogeneous groups. Thus, with the aid of collaborative learning systems, the potential strengths of heterogeneity can be optimized and cause the learners in heterogeneous groups to outperform those in homogenous groups. Yet, as far as satisfaction with the process is concerned, the heterogeneous groups and the homogeneous groups are not likely to differ; this is attributable to the overwhelming effect of collaborative learning systems which pervades the communication process, as well as minimizes the prominence of cultural diversity (Lim & Zhong, 2004).

Furthermore, learners in heterogeneous groups will conceivably have a more positive attitude toward collaborative learning systems usage as the systems make it easier for them to communicate with members of different cultural backgrounds—in comparison with their previous experience in face-to-face settings. The underlying reason is that learners in heterogeneous groups are more likely to be apprehensive toward oral communication; correspondingly, they perceive the text-based collaborative learning systems to be a more comfortable communication medium, an alternative to oral communication (Brown, Fuller, & Vician, 2004). Since members of homogeneous groups would not suffer communication barriers even in a face-to-face setting, they would not appreciate the benefits of collaborative learning systems to the extent their counterparts in heterogeneous groups would, relatively speaking.

Building Common Ground in Collaborative Learning Systems through Leadership

Collaborative learning systems have the potential to deal with the challenges introduced by cultural diversity; however, a mere focus on technology alone cannot guarantee an enhanced learning experience. Effective communication in collaborative learning systems hinges on establishing a common ground among members of a learning group (Cramton, 2002); common ground refers to the mutual understanding of the knowledge constructed during the learning process. Such mutual understanding is composed of not only the specific pieces of information but also the awareness that other members know the information. When common ground is achieved in a group, collaborative learning is more likely to be effective.

However, heterogeneous groups are inherently poor in establishing and maintaining such common ground; group members tend not to be able to understand or remember contextual information of others, and this may result in inaccurate understandings. Still worse, users from different cultural backgrounds may deem a certain technology (or a specific function) better suited for a given task. In general, collectivistic cultures, which prefer high-context communication (e.g., Asia), tend to perceive collaborative technologies to be a better fit for conveyance process in communication (e.g., composing messages, providing explanations, and carrying out convergence-oriented communications); on the other hand, individualistic cultures, which are leaning toward low-context communication (e.g., U.S.), tend to perceive collaborative technologies as a more appropriate tool for convergence process (e.g., making a group decision) (Massey, Montoya-Weiss, Hung, & Ramesh, 2001). This is also applicable to the context of collaborative learning systems, as the two primary processes in communication, conveyance and convergence, are inherently relevant in the collaborative learning activities. Conveyance process in collaborative learning refers to the exchange of information among learners; convergence process is the construction of shared meaning for information. In consequence, electronic means of communication embedded in collaborative learning systems may make it difficult to discover and resolve misunderstandings (e.g., uneven distribution of information among members) (Cramton, 2002).

We posit that leadership may facilitate the development and maintenance of common ground in collaborative learning systems. The leader of a learning group takes the responsibility for organizing a group, delegating assignments, coordinating information, and supporting the contributions of others (Hostager, Lester, Ready, & Bergmann, 2003). A leader facilitates group process by allowing varied views to be heard, providing information, probing for more information, and summarizing the progress the group is making toward its goals. The leader has to quickly recognize when a group wanders off and bring the participants back to the issue at hand.

The leaders should gain an adequate appreciation of the different reactions that technologies have evoked among members with different cultural backgrounds. Thereafter, he would be in a position to help facilitate in developing the group norms or guidelines concerning communications in a heterogeneous team. In formulating these rules or guidelines, the leader must recognize the influential role of cultural differences on users' perception. For example, learners of collectivistic cultural backgrounds should utilize features of the asynchronous discussion forum for their collaborative learning activities; these learners may prefer asynchronous groupware which allows more time to compose messages and express themselves. In particular, these rules or guidelines should define how, when, and which technologies should be used, and how the team will deal with conflict and make decisions based on members' perceptions and preferences. By doing so, the fit between the task and technologies used can be achieved, and eventually, this fit would enhance the common ground building among members in collaborative learning systems. All these activities carried out by a leader contribute toward forming a common ground and, in turn, lead to the achievement of meaningful learning.

FUTURE TRENDS

The current research has defined cultural diversity exclusively in terms of nationality being either heterogeneous or homogeneous. This approach can be used as a benchmark (or a foundation) for future studies, so as to develop greater conceptual understanding, or even a substantive theoretical model, of cultural diversity. Undoubtedly, there are other aspects that also deserve research attention to study the notion of degree of heterogeneity in groups.

Two examples are time and the type of cultural differences. First, cultural diversity may be affected by the time factor. Although cultural background of a person is mainly inherited from the society where he originates, it can change with time when he moves to a new society. Next, the concept of cultural diversity can be more precisely calibrated in terms of the extent of variety of cultures embedded in a given team. A heterogeneous group consisting of American and European members is in a sense less heterogeneous than one that comprises Japanese and Americans. Consequently, to utilize the potential benefit, collaborative learning systems should be incorporated into the learning process while taking into consideration the degree of heterogeneity.

Additionally, the effect of collaborative learning systems cannot be adequately discussed without considering other pertinent contextual factors that would help realize the potential of these systems in addressing cultural diversity in the context of e-learning. In this connection, leadership, common ground, conveyance process, and convergence process are the contextual factors identified in this article. They deserve to be further researched both in terms of breadth and depth.

CONCLUSION

Globalization and the paradigmatic shifts toward collaborative learning make cultural diversity a pertinent factor in e-learning. To provide meaningful collaborative learning experience to learners, collaborative learning systems facilitate heterogeneous groups by smoothing the communication process; they are, therefore, expected to play an important role in e-learning. This article addresses the conceptual and practical issues in invoking collaborative learning systems to support cultural diversity. By synthesizing the theoretical perspectives regarding cultural issues, the current article identifies the corresponding collaborative learning systems functions in overcoming the challenges alongside the conceptual expositions. In addition, it expounds on the role of leadership in building common ground among learners to utilize the potential of heterogeneous groups in e-learning.

REFERENCES

Alavi, M., & Leidner, D.E. (2001). Research commentary: Technology-mediated learning: A call for greater depth and breadth of research. *Information System Research, 12*(1), 1-10.

Alavi, M., Marakas, G.M., & Yoo, Y. (2002). A comparative study of distributed learning environments on learning outcomes. *Information Systems Research, 13*(4), 404-414.

Brown, S.A., Fuller, R.M., & Vician, C. (2004). Who's afraid of the virtual world? Anxiety and computer-mediated communication. *Journal of Association for Information Systems, 5*(2), 79-107.

Chang, K.T., & Lim, J. (2005). The role of information technology in learning: A meta-analysis. In D.D. Carbonara (Ed.), *Technology literacy applications in learning environment* (pp. 14-36). Hershey, PA: Idea Group.

Chidambaram, L. (1992). The electronic meeting room with an international view. In R.P. Bostrom, R.T. Watson, & S.T. Kinney (Eds.), *Computer augmented teamwork: A guided tour*. New York: Van Nostrand Reinhold.

Cramton, C.D. (2002). Finding common ground in dispersed collaboration. *Organizational Dynamics, 30*(4), 356-367.

Dennis, A.R., & Valacich, J.S. (1999). Rethinking media richness: Towards a theory of media synchronicity. In *Proceedings of 32nd Hawaii International Conference on System Sciences*. CA: IEEE Press.

Feather, S.R. (1999). The impact of group support systems on collaborative learning groups' stages of development. *Information Technology, Learning and Performance Journal, 17*(2), 23-34.

Gudykunst, W.B. (1995). Anxiety/uncertainty management (AUM) theory: Development and current status. In R.L. Wiseman (Ed.), *Intercultural communication theory* (pp. 8-15). CA: Sage Publications.

Herring, S.C. (1999). International coherence in CMC. In *Proceedings of 32nd Hawaii International Conference on System Sciences*. CA: IEEE Press.

Hofstede, G. (1997). *Cultures and organizations: Software of the mind*. New York: McGraw Hill.

Hostager, T.J., Lester, S.W., Ready, K.J., & Bergmann, M. (2003). Matching facilitator style and agenda structure in group support systems: Effects on participant satisfaction and group output quality. *Information Resources Management Journal, 16*(2), 56-72.

Janis, I.L. (1982). *Groupthink: Psychological studies of policy decisions and fiascoes* (2nd ed.). Boston: Houghton Mifflin Company.

Laughlin, P.R., Chandler, J.S., Shupe, E.I., Magley, V.J., & Hulbert, L.G. (1995). Generality of a theory of collective induction: Face-to-face and computer-mediated interaction, amount of potential information, and group versus member choice of evidence. *Organizational Behavior and Human Decision Processes, 63*, 98-111.

Leidner, D.E., & Jarvenpaa, S.L. (1995). The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly, 19*(3), 265-291.

Lerouge, C., Blanton, J.E., & Kittner, M. (2004). A causal model for using collaborative technologies to facilitate student team projects. *Journal of Computer Information Systems, 45*(1), 30-36.

Li, Q. (2002). Exploration of collaborative learning and communication in an educational environment using computer-mediated communication. *Journal of Research on Technology in Education, 34*(4), 503-516.

Lim, J., & Zhong, Y. (2004, May 23-26). Cultural diversity, leadership, and collaborative learning systems: An experimental study. In *Proceedings of 15th Information Resources Management Association International Conference*, New Orleans, LA.

Mandryk, R.L., Inkepn, K.M., Bilezikjian, M., Klemmer, S.R., & Landay, J.A. (2001, March / April). Supporting children's collaboration across handheld computers. In *Extend Abstracts of CHI, Conference on Human Factors in Computing System*, Seattle, WA.

Massey, A.P., Montoya-Weiss, M., Hung, Y.C., & Ramesh, V. (2001). Cultural perceptions of task-technology fit. *Communications of the ACM, 44*(12), 83-84.

Rich, M. (1997, August 15-17). A learning community on the internet: An exercise with masters students. In *Proceedings of Americas Conference on Information Systems*, Indianapolis, IN.

Salovaara, H. (2005). An exploration of students' strategy use in inquiry-based computer-supported collaborative learning. *Journal of Computer Assisted Learning, 21*, 39-52.

Schlechter, T.M. (1990). The relative instructional efficiency of small group computer-based training. *Journal of Educational Computing Research, 6*(3), 329-341.

Slavin, R.E. (1990). *Cooperative learning: Theory, research, and practice*. NJ: Prentice Hall.

Stephan, W.G., & Stephan, C.W. (2001). *Improving intergroup relations*. CA: Sage Publications.

Valacich, J.S., Jessup, L.M., Dennis, A.R., & Nunamaker, J.F. (1992). A conceptual framework for anonymity in electronic meetings. *Group Decision and Negotiation, 1*(3), 219-241.

Vygotsky, L.S. (1978). Mind in society: *The development of higher psychological processes*. MA: Harvard University.

Wang, X.C., Hinn, D.M., & Kanfer, A.G. (2001). Collaborative learning for learners with different learning styles. *Journal of Research on Technology in Education, 34*(1), 75-85.

Young, S.S.C. (2003). Integrating ICT into second language education in a vocational high school. *Journal of Computer Assisted Learning, 19*, 447-461.

Yu, F. (2001). Competition within computer-assisted cooperative learning environments: Cognitive, affective, and social outcomes. *Journal of Educational Computing Research, 24*(2), 99-117.

Zhang, D., & Zhou, L. (2003). Enhancing e-learning with interactive multimedia. *Information Resources Management Journal, 16*(4), 1-14.

KEY TERMS

Collaborative Learning Systems (CLS): Systems implemented to provide computer-supported environments which facilitate collaborative learning. Primarily, these systems serve as a medium through which learners can cooperate with others.

Common Ground in Collaborative Learning Group: The mutual understanding of the knowledge constructed during the learning process. Such mutual understanding is composed of not only the specific pieces of information but also the awareness that other members know the information.

Convergence Process in Collaborative Learning: The construction of shared meaning for information in collaborative learning activities.

Conveyance Process in Collaborative Learning: The exchange of information among learners in collaborative learning activities.

Cultural Diversity: The composition of members' (national) cultural backgrounds in a group. It is defined exclusively in terms of nationality, being either heterogeneous or homogeneous.

Heterogeneous Group: A group whose members are of different (national) cultural backgrounds.

Homogeneous Group: A group whose members are of the same (national) cultural background.

Category: IT Education

Cultural Issues in the Globalisation of Distance Education

Lucas Walsh
Deakin University, Australia

INTRODUCTION

This article discusses ongoing cultural challenges faced by distance education providers seeking to deliver programs of study transnationally. Focusing on a key period of distance education during the late twentieth century, this discussion begins by tracing the impact of global economic and technological developments, such as the growth of mega-university enrolments, privately owned education providers and the Internet. The 1990s saw intense interest in the use of Internet-based applications for distance learning and the subsequent arrival of important new actors in this marketplace, such as Blackboard and WebCT.

The author then examines some of the key cultural challenges arising from this convergence of economic, educational and technological dimensions of globalisation, such as the problematic use of models of independent learning in distance delivery.

Turning to future trends, three recent developments in the Internet pose significant challenges to these markets and approaches: open courseware and other initiatives seeking to provide open access to educational resources; the diffusion of user-generated applications, tools and environments; and the fragmentation of online information sources. These trends invite education providers to reflect on the cultural dimensions of distance education. It is argued that while new approaches to e-learning present new opportunities to enhance distance learning, certain key lessons from the 1990s should continue to inform the contemporary development of distance education.

BACKGROUND

The 1990s was a period of tremendous growth internationally in distance education, evident in the expansion of mega-universities, virtual campuses and Open University courses throughout the world. Distance education involves "the provision of programs of study which provide both content and support services to students who rarely, if ever, attend for face-to-face teaching or for on-campus access to educational facilities" (Cunningham et al., 1998, p. 23). Designed to appeal to students seeking greater choice in relation to the time and place of study, and to the mode and pace of learning, these programs tend to be taken by students who find on-campus attendance impractical due to factors such as geography, work and family commitments (Ryan, 1998). Open learning frameworks further allow students to enrol in off campus programs of study irrespective of their previous credentials (Cunningham et al., 1998).

Mega-universities were established throughout Asia in response to the new commercial realities of globalisation. At the time, a mega-university such as Sukhothai Thammathirat Open University of Thailand attracted around 250,000 students, while Indonesia's Universitas Terbuka had more than 350,000 enrolments. Other mega-universities in India, Korea and China had equally massive enrolments (International Centre for Distance Learning, 1995).

By the mid 90s, more than 2 million students from the Asia-Pacific region were enrolled in informal and formal distance education programs (Commonwealth of Learning, 1994; Latchem, 1997). Public universities across South East Asia were encouraged to develop distance education programs as a low-cost basis for mass education. Universiti Sains Malaysia and Hanoi Open University expanded their distance education programs, targeting potential students among working adults and those residing in more remote regions (Ziguras, 2000). Privately owned education providers within Asia also responded to the market potential of distance learning. Malaysia's first virtual university, Universiti Tun Abdul Razak (UNITAR), opened in 1999 and used technologies such as the Internet to teach students exclusively by distance mode (Ziguras, 2001).

UNITAR was one of a number of virtual universities seeking to exploit changing markets in distance learning. Institutions such as The University of Phoenix and Western Governors University were already providing distance education in North America. The first Web-based university courses emerged around 1995 (Bates, 2005). Like UNITAR, Western Governors University operates without a conventional home campus (Gilbert, 1996). While The University of Phoenix developed a flexible distance program for working adults, Western Governors University–a nonprofit online provider–sought to "expand the marketplace for instructional materials, courseware, and programs utilising advanced technology," as well as "identify and remove barriers to the free

functioning of these markets" (Noble, 1998, p. 361). With the growth in distance education in regions such as Asia, many Western education institutions saw an opportunity to capture potentially lucrative global education markets. Growth in the number courses offered via distance mode during this decade was significant. In 1997, 1,000 institutions throughout the world offered roughly 33,000 distance education courses and programs, a tenfold increase since 1991 (Latchem, 1997).

The latter 1990s was a watershed period of technological diffusion; namely, of the Internet. Important new actors, such as Blackboard and WebCT, entered the global education marketplace late in the decade and soon dominated the market for global provision of learning management services and systems. By 2001, WebCT had sold over 1 million student licences across 80 countries (Bates, 2005). During this period of uneven development, many transnational distance education providers struggled to effectively deliver educational courses and content within this confluence of market growth and changing technology.

CULTURAL CHALLENGES OF DISTANCE EDUCATION

Demand for distance education provided a powerful incentive for higher education, professional development and training providers to collaborate with commercial software developers to create Learning Management Systems (LMS) capable of efficient, low-cost delivery of course content and educational resources. Much of the early e-learning software was developed through these collaborations and then sold to other universities. LMS, such as WebCT, were developed during the latter 1990s to enable easier access to course materials, teaching tools and learning objects via Web-browsers. Designed to standardise online course development and simplify technological training and support at the deliverer's end, these systems came to dominate online provision in this education sector.

One of the major appeals of e-learning is the capacity of the Internet to enable geographically dispersed students to engage with their "virtual" classmates as part of an online community of independent learners. An important motivation for distance learning was the strategic interest of educational providers in using new technologies to grow student numbers and facilitate greater economies of scale in course-delivery. Reliant on Western content, many online providers failed to take into consideration the cultural dimensions of transnational delivery (Ziguras, 2000; 2001). The "one-size fits all" model underpinning transnational delivery presumed that Western standards, content and modes of delivery were "universally relevant and universally welcome" across different cultural settings (Patrick, 1997, p. 2).

During the 1990s, influential e-learning frameworks and software were developed in the English language (mainly in North America), and adopted Western models of learning. These were then used in other cultural settings without sufficient consideration of their appropriateness to the pedagogical and learning needs within those settings. To make distance courses simpler to use and more marketable for transnational delivery, distance courses were globalised to remove any cultural specificity of content. Seeking to provide curriculum and materials which "transcend local cultural and language barriers" and that are "relevant to learners wherever they happen to reside," many online courses were offered internationally but not modified to suit local sites of delivery (Bates & de los Santos, 1997, p. 49).

When delivering distance courses across different cultural settings, the models of independent learning underpinning distance education were in some instances problematic and in others, disastrous. During the 1980s, for example, Indonesia's Universitas Terbuka adopted a model of distance education based on the UK Open University. The Western approach to independent, self-directed learning that underpinned this model was unfamiliar to students and teachers from heteronymous Indonesian cultural backgrounds. Courses used text-based resources, which in an orally-based society characterised by low levels of reading and writing further contributed to the failure of this model (Dunbar, 1991).

The University of the South Pacific (USP) experienced similar problems when it adopted a distance education program. The program delivered resources to over 5,000 students spread across numerous islands, languages and cultures. The cultural backgrounds of this diverse group of learners were not taken into sufficient consideration in the course design. Like Universitas Terbuka, the program's emphasis on independent learning and use of the English language was unsuitable to Pacific peoples accustomed to learning from one another through physical interaction, observation and imitation, and through intimate relationships with their teachers (Thaman, 1997).

These examples highlight the kinds of cultural challenges associated with the globalisation of distance education; particularly the need to understand how the pedagogical and learning assumptions of a given distance framework impact upon different cultural settings. Courses via distance mode typically assume that students are self-motivated, self-directed learners (Cunningham et al., 1998), which may not be favourable to students (or teachers) from diverse cultural and educational backgrounds.

The cultural problems of the USP and Indonesian examples described above can, in part, be attributed to the dangers of a "quick fix" approach. Due to the somewhat *ad hoc* approach adopted by providers, implementation of technologically-assisted open and distance learning during the 1980s and 1990s was marred by a lack of coordination

or pooling of educational experience (Walker, 1997). While understanding of the pedagogical and cultural dimensions of e-learning in distance education improved significantly during the 1990s—particularly in areas that directly mediate learning, such as instructional design (Alexander & Blight, 1996; Henderson, 1994) and transnational delivery (Ziguras & Rizvi, 2001)—education providers, governments, telecommunications providers and industry struggled to collaboratively develop a common vision and understanding of the economic, social and cultural national and global benefits of distance learning during this time of often explosive and uneven development (Latchem, 1997).

FUTURE TRENDS

The market in distance learning continues to grow, both financially and in terms of global reach. It has been estimated that by 2015, the global online higher education market will be worth more than $69 billion (Hezel Associates, 2005). As a large proportion of this growth will consist of some form of distance education, not only will providers and educators face the kinds of issues described above; the market itself may be challenged by three developments in Internet-based learning: (1) the emergence of electronic publishing and open courseware initiatives seeking to provide open access to these educational resources; (2) the development of user-generated applications, tools and environments; and (3) the fragmentation of information sources and information units used for informal and institutionalised modes of learning.

During the last several years, a number of initiatives have developed that offer free access to course materials from all over the world. The Massachusetts Institute of Technology's (MIT) *OpenCourseWare* project enables open access to educational materials from over 1,400 MIT courses, including syllabi, lecture notes, exams, reading lists and video lectures. The idea of providing educators, students and self-learners with free access is part of MIT's efforts to reposition itself in a changing and competitive distance environment (Massachusetts Institute of Technology [MIT], 2006). Similarly, the Open University's *OpenLearn* project is also making educational resources freely available via the Internet in response to digital divide problems of availability and cost (BBC, 2006). Echoing other initiatives in the U.S. and Japan, *OpenLearn* reflects a shift away from the proprietary, "top-down" delivery models prevalent in the 1990s.

The growing development of open source and subsequent availability of free software packages, such as *Moodle*, further challenge proprietary, suite-based frameworks. In recent years, there have been signs of a shift from the centralised provision of content toward applications and services that enable users to take more control over how they access and share information. The rapid adoption and diffusion of user-generated Web sites such as wikis and Web logs reflect a broader shift in Internet-usage toward open-ended, user-driven, participatory online platforms to share, organise and repurpose different kinds of content for publication, subscription and linking across networks (Spivack, 2003; O'Reilly, 2005; Davis, 2005). A blog, for example, is in its most basic sense a personal journal consisting of "a hierarchy of text, images, media objects and data, arranged chronologically, that can be viewed in an HTML browser" so that it can be shared via the Web (Winer, 2003). By enabling users to develop, repurpose and customise their online educational resources and environments, these resources and platforms present opportunities for learning collaborations, knowledge construction and the localisation of culturally appropriate e-learning.

Tools such as wikis and blogs are designed for collaborative participation in the knowledge construction process, facilitating forms of personal knowledge mapping (Langreiter & Bolka, 2005). Wikipedia, for example, is a widely-used reference tool that enables documents to be written collaboratively using a simple Web-browser, and then continually revised, corrected and expanded by users. Issues of scholarship, authenticity and reliability aside, users of wikis can contribute to knowledge construction as consumers, creators and sharers of content. These applications and approaches broaden the scope for interaction and dialogue, inviting new opportunities for student participation that "is very different from traditionally assigned learning content. It is much less formal. It is written from a personal point of view, in a personal voice… what happens when students blog, and read reach others' blogs, is that a network of interactions forms - much like a social network" (Downes, 2005).

This movement away from the 1990s LMS paradigm favouring passive models of "content-consumption" and object-centred approaches is loosely referred to as "E-learning 2.0." Seeking to move beyond e-learning as a mode of delivery, this approach focuses more on learning actions, the use of interoperable *collections* of software applications. It also emphasises user-centred design and the nurturing of collaborative learning online (Mowbray, 2007). By interweaving approaches, online resources and applications, E-learning 2.0 suggests a movement in online education toward interoperating applications that enable creative learner-centred and potentially collaborative environments. Moving away from established top-down delivery frameworks which narrowly understand e-learning "as being a type of content, produced by publishers, organised and structured into courses, and consumed by students," this shift in e-learning seeks to avoid dependency on "an institutional or corporate application" in favour of "a personal learning center, where content is reused and remixed according to the student's own needs and interests" (Downes, 2005).

These approaches and tools for e-learning are not without certain familiar challenges. For example, the cultural frame of MIT courses is predominantly North American and, at the very least, requires varying degrees of contextualisation (e.g., by a local teacher) to be effective. English is arguably the *de facto* language of the international open source community. (Even though *Moodle's* online discussions take place in dozens of languages, the English-language nevertheless predominates.)

Much of the current enthusiasm for E-learning 2.0 needs to avoid the dangerous assumption that providing the technological tools for self-expression and creation will lead to empowerment and improved learning and teaching. Changing the emphasis from the design of learning content to the way it is used is a positive step; however, certain cultural and pedagogical issues need to be kept in mind. Langreiter and Bolka (2005) rightly caution that while content can be aggregated and tailored to the needs of individual learners, the task of consolidating and organising disaggregated content is shifted toward the learner. Using *Moodle* as its central platform, the *OpenLearn* project encourages learners "to become self-reliant, but also to use online communities to support their learning" (Lane, cited in BBC, 2006). Nevertheless, user-centred approaches require greater responsibility for content creation and maintenance. Not all teachers and learners will have the skills, technological literacy or resources to undertake these learner-centred responsibilities. While many of tomorrow's distance students will indeed be "digital natives" familiar to (if not proficient in) the use of the applications underpinning these environments, one could safely assume that they will continue to have a diverse range of skills, abilities and cultural backgrounds.

The emergent learning environments described above have their own structural biases and values that favour certain learning styles. It may well be that some students prefer to work independently and may not feel comfortable engaging in this kind of participatory process. However, implicitly privileging certain styles of independent learning will not be appropriate for all cultural learning styles.

Spurred by the spontaneous generation and dissemination of content by Web users, a diversification and expansion of information sources and educational resources is challenging conventional processes of information-gathering, research and access to content. The development of online learning may be characterised by a diversification of the types of educational resources available to learners in both informal and institutional settings. Where learners have traditionally consulted a single body of authoritative work (such as a printed book, journal article or teacher) students increasingly draw from a plethora of Web-based resources, such as online encyclopaedias, teaching materials, discussion forums, and RRS posts (Langreiter & Bolka, 2005).

While this process of fragmentation may be associated with problems such as the questionable use of unauthoritative sources and plagiarism, it is also stimulating educators to develop more diverse approaches to e-learning. Langreiter and Bolka highlight one benefit of fragmentation: information chunks created using a blogging application or a wiki are much easier to produce and maintain than larger networks. "Furthermore," they suggest, "disaggregated content - theoretically - can be re-aggregated to optimally suit an individual learner's preferences (instead of the needs of an idealised common denominator)" (2005, p. 80). Whereas conventional e-learning management systems tend to locate learners as relatively passive receivers of course-materials within a standardised framework, these tools offer new opportunities for distance education to become more responsive and culturally attuned to the different ways that students learn, and through which teachers have the potential to use technology more effectively according to their particular pedagogical strategies.

CONCLUSION

Cultural differences of learners (and teachers) have often been unacknowledged or treated as unproblematic in the design of software and content for online distance learning and delivery. Improved awareness of cultural issues in distance learning is informing the development and adaptation of learning programs and materials for cross-cultural delivery by acknowledging differences in communication and education cultures and incorporating this awareness into programs, resources and modes of delivery (Banks, 2006; Collis, 1996; Zaltsman, 2006). Effective, culturally sensitive and pedagogically appropriate distance learning requires understanding how the basic structure of courses, such as time allocation, assessment practices and pedagogy, and impact upon learners from different backgrounds by privileging certain values over others. In addition, education providers and developers of online distance learning programs, resources and software must be aware of how cultural biases may be built into the design of Websites and software, be it a digital repository, collaborative virtual learning environment, user-driven application or conventional LMS.

According to Sharma and Mishra (2006, p.1), "e-learning is the fastest growing sub-sector of a $2.3 trillion global education market." With this continued growth, there continues to be a danger that online distance learning will be driven by financial rather than culturally sensitive educational imperatives.

Following the unsteady and often culturally problematic growth of the global distance education marketplace during the 1980s and 1990s, recent online developments, such as user-generated software applications, open tools, free content and collaborative approaches present new opportunities and challenges. Nevertheless, the case studies from Indonesia and USP described above illustrate the importance of incor-

porating understandings of culture into any distance learning program, particularly when delivery is transnational. The mode of delivery may affect learners differently depending on the cultural styles of teaching and learning to which they are accustomed. For example, the collaborative and user-generated learning technologies discussed in the previous section will invariably be imbued with structural properties that may privilege some learners while excluding others previously schooled in different cultural milieus.

And while approaches such as E-learning 2.0 offer scope for flexibility in how learners negotiate distance learning and each other, the effectiveness of any technology in distance learning will continue to depend on how it is used rather than any particular feature of the medium itself (Bates, 2005; Inglis, Ling & Joosten, 2002).

REFERENCES

Alexander, S., & Blight, D. (1996). Technologies for the new millennium. *Technology in international education*. Sydney: University of Technology, Sydney and IDP Education Australia.

Banks, S. (2006, December 3-6). Collaboration for intercultural e-learning: A Sino-UK case study. In L. Markauskaite, P. Goodyear, & P. Reimann (Eds.), *Who's Learning? Whose Technology?: Proceedings of the 23rd Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education, Sydney, Australia,* (Vol. 1, pp. 71-77). Sydney University Press.

Bates, A. W. (2005). *Technology, e-learning, and distance education*. New York: Routledge Falmer.

Bates, A. W., & de los Santos, J. G. E. (1997). Crossing boundaries: Making global distance education a reality. *Journal of Distance Education,* 12(1-2), 49-66.

BBC. (2006, October 23). OU offers free learning materials. *BBC News*. Retrieved December 14, 2007, from http://news.bbc.co.uk/go/pr/fr/-/1/hi/education/6071230.stm

Collis, B. (1996*). Tele-learning in a digital world: The future of distance learning*. London: International Thomson Computer Press.

Commonwealth of Learning. (1994, November). COL in action. COMLEARN, 5(1).

Cunningham, S., Tapsall, S., Ryan, Y., Stedman, L., Bagdon, K., & Flew, T. (1998). *New media and borderless education: A review of the convergence between global media networks and higher education provision*. Canberra: Australian Government Publishing Service.

Davis, I. (2005, July 4). Talis, Web 2.0 and all that. *Internet Alchemy blog*. Retrieved December 14, 2007, from http://iandavis.com/blog/2005/07/talis-web-20-and-all-that

Downes, S. (2005, October). E-learning 2.0. *E-learn Magazine, 2005*(10). Retrieved December 14, 2007, from http://www.elearnmag.org/subpage.cfm?section=articles&article=29-1

Dunbar, R. (1991). Adapting distance education for Indonesians: Problems with learner heteronomy and a strong oral tradition. *Distance Education, 12*(2), 163-174.

Henderson, L. (1994). Reeves' pedagogic model of interactive learning systems and cultural contextuality. In C. McBeath & R. Atkinson (Eds.), *Proceedings of the Second International Interactive Multimedia Symposium,* (pp. 189-198). Perth: Promaco Conventions.

Hezel Associates. (2005). *Global e-learning opportunity for U.S. higher education*. Retrieved December 14, 2007, from http://www.hezel.com/globalreport/

Inglis, A., Ling, P., & Joosten, V. (2002). *Delivering digitally: Managing the transition to the knowledge media*. London: Kogan Page.

International Centre for Distance Learning. (1995). *Mega universities of the world: The top ten*. Milton Keynes: The Open University.

Langreiter, C., & Bolka, A. (2005). Snips and spaces: Managing microlearning. In T. Hug, M. Lindner, & P. A. Bruck (Eds.), *Microlearning: Emerging concepts, practices and technologies after e-learning, proceedings of microlearning 2005. Learning & working in new media,* (pp. 79-97). Innsbruck: Innsbruck University.

Latchem, C. (1996, September 30-October 2). Flexible and cost effective delivery around the world. In *Paper presented at the 10th Australian International Education Conference on Technologies for the New Millennium*. Sydney: IDP.

Latchem, C. (1997, May 12). A global perspective on flexible delivery. In *Paper presented at the Nuffic Seminar Virtual Mobility: New Technologies and Internationalisation*. The Netherlands: University of Twente.

Laurillard, D. (1994). How can learning technologies improve learning? *Law Technology Journal*, 3(2). Retrieved December 14, 2007, from http://www.law.warwick.ac.uk/ltj/3-2j.html

Massachusetts Institute of Technology. (2006). *MIT OpenCourseWare Web site*. Retrieved December 14, 2007, from http://ocw.mit.edu

Mowbray, M. (2007). Designing online learning communities to encourage cooperation. In P. Zaphiris & N. Lambropoulos

(Eds.), *User-centered design of online learning communities*, (pp. 102-121). Hershey, PA: Idea Group.

Noble, D. F. (1998). Digital diploma mills: The automation of higher education. *Science as Culture, 7*(3), 355-368.

O'Reilly, T. (2005, September 30). What Is Web 2.0? Design patterns and business models for the next generation of software. *O'Reilly* Web site (http://www.oreilly.com/). Retrieved December 14, 2007, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Patrick, K. (1997). Internationalising curriculum. In *Paper delivered at HERDSA, 97*. Melbourne, Australia: RMIT.

Ryan, Y. (1998). Time and tide: Teaching and learning online. *Australian Universities' Review, 41*(1), 14-19.

Sharma, R. C., & Mishra, S. (2006). Introduction. In R. C. Sharma & S. Mishra (Eds.), *Cases on global e-learning practices: Successes and pitfalls,* (pp. 1-11). Hershey, PA: Idea Group.

Spivack, N. (2003, December 10). Defining microcontent. *Nova Spivack's Web log*. Retrieved December 14, 2007, from http://novaspivack.typepad.com/nova_spivacks_weblog/2003/12/defining_microc.html

Thaman, K. H. (1997). Considerations of culture in distance education in the Pacific Islands. In L. Rowan, L. Bartlett & T. Evans (Eds.), *Shifting borders: Globalisation, localisation and open and distance education,* (pp. 23-36). Geelong: Deakin University Press.

Walker, J. H. (1997). Managing the digital revolution: A strategy to maximise the use of high capacity communications services in Australian education and training. *The new learning environment: A global perspective* [CD Rom]. USA: Pennsylvania State University.

Waters, M. (1995). *Globalization*. London: Routledge.

Winer, D. (2003, May 23). What makes a Web log a Web log? *Web logs at Harvard Law*. Hosted by the Berkman Center for Internet and Society at Harvard Law School. Retrieved December 14, 2007, from http://blogs.law.harvard.edu/whatMakesAWeblogAWeblog

Zaltsman, R. (2006). Communication barriers and conflicts in cross-cultural e-learning. In A. Edmundson (Ed.), *Globalized e-learning cultural challenges* (pp. 291-307). Hershey, PA: Idea Group.

Ziguras, C. (2000). *New frontiers, new technologies, new pedagogies: Educational technology and the internationalisation of tertiary education in South East Asia*. Research Report for Telstra Australia, prepared by Monash Centre for Research in International Education. Australia: Monash University.

Ziguras, C. (2001). Educational technology in transnational higher education in South East Asia: The cultural politics of flexible learning. *Educational Technology & Society, 4*(4), 8-18.

Ziguras, C., & Rizvi, F. (2001). Future directions in international online education. In D. Davis & D. Meares (Eds.), *Transnational education: Australia online* (pp. 151-164). Sydney: IDP Education Australia.

KEY TERMS

Distance Education: Provides programs of study and support services to students who do not wish to regularly attend face-to-face teaching or who find on-campus study impractical due to factors such as geographical constraints, family or work commitments.

E-Learning 2.0: Promotes online learning as a platform for personal learning through interoperable tools that enable reusable content to be authored, repurposed, mixed and shared according to students' particular needs and interests.

Globalisation: Globalisation is a process by which the impact of geographical constraints on cultural and social formations is diminished. Economic globalisation, which has arisen due to factors such as changing patterns of world trade and international mobility, is linked to the emergence of a seemingly universal ideology emphasising the need for flexible responsiveness to global markets.

Learning Management System (LMS): Educational software that enables the delivery and management of learning content and resources to students.

Moodle: An open-source online course management system (the name is derived from the acronym *Modular Object-oriented Dynamic Learning Environment*).

Web Log (Blog): A form of personal journal shared over the Web.

Wiki: Enables content to be written collaboratively using a simple Web browser that can be continually revised, corrected and expanded by its users.

Cultural Motives in Information Systems Acceptance and Use

Manuel J. Sanchez-Franco
University of Seville, Spain

Francisco José Martínez López
University of Granada, Spain

INTRODUCTION

Understanding the moderating factors that influence user technology acceptance and adoption in different contexts continues to be a focal interest in information systems (hereafter, IS) research. Moderating factors may account for both the limited explanatory power and the inconsistencies between studies (Sun & Zhang, 2006). Accordingly, based on a careful literature review, we believe that culture, defined as mental concepts influencing the relationships with other people, the environment and the concept of time (see Hofstede, 1991; Hall, 1989; Trompenaar, 1995), is an important moderating-factor; that is, culture constitutes "the broadest influence on many dimensions of human behaviour" (Soares, Farhangmehr, & Shoham, 2007).

Particularly, culture is a factor that has been shown to be significant but underresearched in recent studies of information-accessing behaviour. Nevertheless, there is increasing interest in the IS research literature in the impact of cultural differences on the development and use of information technologies (hereafter, IT) and IS. For example, the following authors identified cultural values as one of the influential factors on adoption of information and communication technology (hereafter, ICT): Bagchi, Cerveny, Hart, and Peterson (2003), Johns, Smith, and Strand (2003), Maitland and Bauer (2001) and Sørnes, Stephens, Saetre, and Browning (2004). Straub (1994) has used the uncertainty avoidance dimension to explain why the diffusion of information technologies differed in the USA and Japan. Watson, Ho, and Raman (1994) have also used the individualism-collectivism dimension to account for differences in the way Group Support Systems (GSS) affected group decisions in the USA and Singapore. Findings from Chau et al. (2002) illustrate how users from different countries differ in their perception of the purpose of Internet and, consequently, exhibit differences in their behaviours and general attitudes toward the Internet. Marcus and Gould (2000) examine a number of cultural dimensions and their possible impact on user-interface design (see also Barber & Badre, 2001; Del Galdo & Nielsen, 1996). Other authors, for example, explore cultural influences on technology development and innovation (Herbig, 1994), cultural influences on technology adoption (Straub, 1994), and culture as a factor in the diffusion of the Internet (Cronin, 1996; Goodman, Press, Ruth, & Rutkowski, 1994; Maitland, 1999). Finally, Veiga, Floyd, and Dechant (2001) suggest that perceptions of a technology's ease-of-use and usefulness are connected to an individual's broader system of belief, including culturally-sensitive beliefs.

Therefore, because of an anticipated large number of IS users from multiple cultures, research may systematically examine the acceptance and usage models and other models related to cross-cultural motives and beliefs. As Sun and Zhang (2006) suggest, these models have traditionally presented two limitations: (1) the relatively low explanatory power; and (2) inconsistent influences of the cross-study factors. Research may (1) focus on identifying the major cultural dimensions and their corresponding relationships with IS acceptance; and (2) examine the potential moderating effects that may overcome these limitations.

To sum up, culture's role within acceptance and usage model has been only recently investigated. Little research has systematically examined IS preferences of users related to cross-cultural design characteristics. Some researchers have done work in the area of culture and design, but results have been either inconclusive or unrelated to developing loyal users. In this sense, we deem it necessary to highlight several main starting questions. This would add to the few studies that take into account the individual and contextual factors in technology acceptance; specifically, a better understanding of how cultural differences could affect users' evaluations of IS can uncover ways of localising a global interface. While user-interfaces targeted to different cultures may not need to be completely different from each other, there might be some features that allow the targeted audience to feel *at home*.

BACKGROUND

In view of academic and theoretical perspective, the effects of culture on IS acceptance have been studied by researchers mostly based on Hofstede's (1980) cultural construct. It has also been shown to be stable and useful for numerous studies across many disciplines. First, Hofstede's dimensions assume culture falls along national boundaries and that the cultures are viewed as static over time. Second, Hofstede (1980) asserts that central tendencies in a nation are replicated in their institutions through the behaviour or practices of individuals. And, third, Hofstede's framework explicitly links national cultural values to communication practices; i.e., communication practices using ICT are central to our study (see Merchant, 2002; Samovar, Porter, & Jain, 1981; Stohl, 2001). Furthermore, Hofstede's model was important because it (a) organised cultural differences into overarching patterns, and (b) conducted the most comprehensive study of how values in the workplace are influenced by culture, which (c) facilitated comparative research and launched a rapidly-expanding body of cultural and cross-cultural research in the ensuing 20 years. Hofstede's (1980) cultural dimensions serve as the most influential culture theory among social science research, and has received strong empirical support. Hofstede, therefore, contributed the influential work in cross-cultural research.

Hofstede (1984, p. 51) defines culture as "the collective programming of the mind which distinguishes the members of one group from another"; and (b) proposes a series of four dimensions (a fifth was added later; that is, Confucian dynamism) that distinguishes between work-related values. The cultural dimensions are individualism-collectivism, power distance, uncertainty avoidance, and masculinity-femininity. Hofstede and Bond (1988) found an additional dimension, which is particularly relevant to Asian culture, Confucian dynamism (i.e., often referred to as long/short term orientation). These value dimensions, which distinguish national value systems, also affect individuals and organizations.

The present study, however, does not intend to examine the whole range of cultural dimensions influencing IS adoption. This article aims to restrict its focus on individualism and uncertainty avoidance. First, according to Hofstede's model, of the four dimensions, individualism vs. collectivism is the most common dimension used by researchers to understand the differences between two or more given cultures (see also Cohen & Avrahami, 2006). Furthermore, Hofstede's proposition confirms that an individualistic culture is also likely to be a low power-distance culture. Individualism is inversely related to the power distance dimension, which is -0.64 in Hofstede's original study, and -0.70 in the sample of teachers and -0.75 in that of students used in Schwartz's cross-cultural study (Schwartz, 1994; see also Gouveia & Ros, 2000). Power distance shows a pattern of correlations almost opposite to Hofstede's individualism (Hofstede, 1984). At least at a cultural level, individualism is the opposite of the acceptance of hierarchy and of ascribed social inequality. Therefore, we propose power distance index is dropped from explicit consideration here.

Second, with regard to the topic of this study, cultures have a different attitude toward uncertain or unknown matters (specifically, IS acceptance and usage by users from diverse cultures). The tolerance for ambiguity and uncertainty is expressed through the extent to which a culture resorts to written or unwritten rules to maintain predictability; for instance, the absence of physical contact with online partners emphasizes the role of perceived risk. Users in countries with a high score on uncertainty avoidance will thus be more risk-adverse and will not like making changes. For instance, Yeniyurt and Townsend (2003) found the uncertainty avoidance dimension, among other dimensions, to be negatively correlated with the adoption of ICT-based services such as Internet and PCs. In fact, uncertainty avoidance has the most direct bearing on preference for and use of communications media

Third, Bagchi et al. (2003) argued that «IT promote more cooperation at work, better quality of life and these values are espoused in nations with low MF (i.e., masculinity/femininity) index». However, as comment, «it could be argued equally well that in a country with high masculinity there would also be a positive attitude toward implementing ICT if these technologies improve performance, increase the chance of success and support competition, which are all key factors of a masculine culture». In this sense, Johns et al. (2003) included the individualism/collectivism and uncertainty avoidance dimensions only; these authors felt that achievement orientation (masculinity/femininity dimension) has a mixed impact on the use of technology. The masculinity/femininity dimension could thus have at least at the conceptual level a mixed impact on the ICT (see Kovacic, 2005). In this research, we also propose that masculinity/femininity dimensions are also dropped from explicit consideration.

INDIVIDUALISM AND UNCERTAINTY AVOIDANCE DIMENSIONS

Individualism/Collectivism

Individualism/collectivism focuses on the degree the society reinforces individual or collective achievement and interpersonal relationships. Hofstede (1980) argued that cultures high on individualism tend to promote individual decision-making over group consensus. Research has shown that individualistic users support individual identity, and they think that they should be self-sufficient; that is, they resist influence

attempts by higher-status individuals. In other words, status influence is likely to be low. Thus, the task and objectives is more important for them than the relationship.

In contrast, in societies emphasizing collectivism, the group becomes the primary source of an individual's identity and individuals seek approval, status and support through group affiliation. Collectivistic users are more group-oriented, and support the group identity over the individual identity (Chau, 1996). The relationships for the collectivistic users are more essential than the task to be completed. Concerns with respect to group welfare are emphasised, as aggregate interests tend to prevail over autonomous, individualistic ones (Hofstede & Bond, 1988).

Given these shortcomings, several studies have thus found that individualistic users show a greater degree of instrumental motivational orientation in comparison to collectivistic users. Individualism implies that social behaviour is established by personal goals and does not overlie the goals of the collective, while in collectivism the group is more important than the individual and the people in the group are ready to cooperate. Individualistic users (a) are motivated by achievement needs and (b) tend to exhibit more of certain values such as assertiveness, competitiveness and rationality. The core element of individualism is the assumption that individuals are independent of one another. In an individualistic culture, people therefore seem to be more innovative and trusting in exchange relationships with external parties (Veiga, Yanouzas, & Buchholz, 1993). On the contrary, the core element of collectivism is the assumption that groups bind and mutually obligate individuals.

Therefore, individualistic users could view the different IS as tools for performing tasks, whereas collectivistic users could view them as tools for socialization. In highly-individualistic societies, perceptions about the usefulness of the IS are likely to be based on beliefs about how they affect the individual's job performance. Because they believe in personal control and individual achievement, users in individualistic cultures accept task accomplishment as their personal responsibility. Beliefs about the usefulness are thus made based on the extent to which the IS are seen to enhance the task performance of individuals.

Individualistic users will try the different IS, even if they do not have a positive attitude toward using them, because they may provide productivity enhancement (i.e., usefulness). Consequently, individualistic users need to perceive them as being useful or they will not attempt to use it. It is expected that individualistic users' IS perceived usefulness may exert a more intense influence on determining the intention to use the Web. On the contrary, collectivistic users (related to lower instrumental usage) would tend to underestimate the perceived usefulness of the IS. Collectivistic users, as opposed to individualistic users, would not perceive them as being relatively useful.

Finally, when individualistic users engage in instrumental behaviour, they are motivated to perform their activities in an efficient, focused and timely manner, and with a minimum of irritation (adapted from Babin, William, & Griffinn, 1994). Accordingly, individualistic users do not want to be distracted from their tasks. IS that are easy to understand and use can thus be associated with being able to save effort and irritation. As we commented above, perceived ease-of-use influences individual attitudes through two mechanisms: self-efficacy and instrumentality. IS that save effort are correlated with increasing utility. Over time—that is, increasing self-efficacy as a basic determinant of perceived ease-of-use—the indirect effect of perceived ease-of-use (through perceived usefulness) becomes stronger (Watson et al., 1994). Therefore, perceived ease-of-use—as a factor facilitating task-performance and utility—will likely be weighted more strongly by individualistic and weak uncertainty avoidance users.

Uncertainty Avoidance

Hofstede (1991, p. 113) defines Uncertainty Avoidance Index (UAI) as "the extent to which the members of a culture feel threatened by uncertain or unknown situations." UAI focuses on the level of tolerance for uncertainty and ambiguity within the society, that is, unstructured situations. In a culture high on uncertainty avoidance, individuals are more likely to avoid acceptance and use new technologies because of the uncertainty and ambiguity involved. Moreover, IT may be considered inherently risky (Herbig, 1994). Individuals from high uncertainty avoidance societies attempt to reduce personal risk. For instance, empirical research in 11 European countries by Steenkamp, Ter Hofstede, and Dulton (1999) revealed that uncertainty avoidance scores, among other dimensions, is a strong cultural influence on user's innovativeness in general. It has been found that cultural uncertainty avoidance has a negative impact on users' innovativeness. Specifically, with regard to the topic of this study, cultures have a different attitude towards uncertain or unknown matters. The tolerance for ambiguity is expressed through the extent to which a culture resorts to written or unwritten rules to maintain predictability.

In strong uncertainty-avoidance societies, their members are encouraged to anticipate the future, create institutions establishing and reinforcing security and stability, and avoid or manage risk. Those members tend to "take time" for action until they acquire enough knowledge and information to reduce and resolve unclear and unstructured situations. Organizational members in strong uncertainty avoidance countries have a feeling of anxiety when encountering unfamiliar risks, deviant ideas, or conflicts in their work place. Moreover, in a culture high on uncertainty avoidance, individuals could be "comprehensive processors" who attempt to assimilate all available information before rendering judgment, while in a culture low on uncertainty

avoidance individuals could be "selective processors" who often rely on a subset of highly available and salient cues in place of detailed message elaboration (see also Meyers-Levy & Maheswaran, 1991; Meyers-Levy & Sternthal, 1991; Morris & Venkatesh, 2000). In contrast, in a weak uncertainty-avoidance society, members are encouraged to tolerate uncertainty, take risks, *take each day* as it comes; perhaps have low expectations and a fatalistic outlook. Members in weak uncertainty-avoidance countries tend to feel less uncomfortable in unclear and unstructured circumstances and are more likely to take risks in unfamiliar situations where encountering deviant and innovative ideas and behaviour with no rules. Achievement functions as a great motivational factor and it encourages those members to take actions in either familiar or unfamiliar situations.

Uncertainty avoidance societies could thus (a) hold lower perceptions of self-efficacy; (b) show more concern about the risks associated with technologies; and (c) experience higher levels of anxiety and more negative feelings; that is, the average risk propensity of an entire population of people tends to be higher for low uncertainty-avoidance cultures and lower for high uncertainty-avoidance cultures. *The uncertainty inherent in life is felt as a continuous threat which must be fought*. Hence, given the lower levels of perceived self-efficacy among uncertainty avoidance societies, users (a) could perceive greater risks, so (b) they will be more IS-usage averse. As Keil, Beranek, and Konsynski (2000) comment, decision makers tend to exhibit risk-averse behaviour when risk perception is high and risk-seeking behaviour when risk perception is low (Steenkamp et al., 1999) (e.g., Staw, Sandelands, & Dutton, 1981).

Hofstede (1980) argued that uncertainty avoidance is related to anxiety that could be the feeling-output when confronted with problems or challenges. The extent to which this occurs might negatively influence willingness to IS usage. That is to say, these societies–more risk averse and with a lower self-confidence–(a) will not engage in behaviours without previously adjusting their attitudes, and (b) will take more time to decide to try the different IS. Expanding the previous reasoning, attitude toward using IS will, therefore, be a relevant mediator between perceptions and intention to use them among uncertainty avoidance cultures.

Also, in general, once collectivist societies establish a positive attitude toward something, they tend to internalise it and take it into their in-group circle (Rice & Love, 1987; Pavlou & Chai, 2002). Thus, we would expect that members of a collectivist culture would want to maintain harmonious "IS<->user" relationships.

Furthermore, the high perceived risk associated with the IS-usage significantly reduces the weak uncertainty-avoidance societies' perception about (a) their self-efficacy in using it, (b) its perceived usefulness, and (c) its ease-of-use. These low evaluations of perceptions among weak uncertainty-avoidance societies can cause an increase in the salience of them in determining attitudes toward using the IS (adapted from Venkatesh & Morris 2000). Weak uncertainty-avoidance cultures may not be willing to accept a difficult and annoying interface.

To sum up, individuals from high uncertainty avoidance societies attempt to reduce personal risk, while being more likely to resist innovative ideas and conform to the rules. On the other hand, these individuals seek to avoid ambiguity and thus create rules for most possible situations. Members of a low uncertainty avoidance culture might be more broad-minded, without a lot of need for social approval, more prone to risk taking, with tolerance toward deviant behaviour and acceptance of innovative ideas.

FUTURE TRENDS

This study has future implications both in practice and in theory. It shows that cultural variables are relevant to IS acceptance and usage. Individualistic and weak uncertainty-avoidance scripts should thus focus on the following user-interface and design elements. On the one hand, work tasks, roles, and mastery, with quick results for limited tasks (aspects traditionally related to goal-directed activities); and, on the other hand, navigation-oriented to control and attention gained through games and competitions.

Likewise, in an attempt to develop positive attitudes among collectivistic and high uncertainty-avoidance cultures toward the IS, researchers and professionals might suggest and introduce courses and programs to gain more experience and self-efficacy and, in turn, higher optimal experiences; also, the creation of technical support programs which combine user service with precise, simple, understandable technical information, avoiding the use of jargon (i.e., clarity and courtesy). These policies (a) strengthen the user's perception of trustworthiness based on IS design, and (b) promote the progressive reduction of risk and technological anxiety, showing a determined willingness to understand and comprehend user's needs. The theoretical background provides evidence for culturally differentiated IS acceptance and usage.

Therefore, the study has an implication for diffusion theory, or adoption of IS; that is, findings may be taken as an operational basis for more intensive cultural adaptations of the IS. Our proposals justify the inclusion of moderating cultural-variables. These suggestions range from changes in pedagogy and perspective to making the computer a tool of collaboration between pairs or groups of users rather than individuals (adapted from Bryson & de Castell 1995; Littleton & Bannert 1999); and to giving collectivistic cultures a context in the IS-based experience. Therefore, while it seems that

these differences are perhaps more enduring than expected, they could certainly be controlled and rendered reasonable by the appropriate use of training-sessions.

CONCLUSION

Information systems offer unprecedented opportunities for world-wide access to information resources. Accordingly, the theoretical proposals presented in this article analyse the moderating effect of national culture (specifically, individualism and uncertainty avoidance dimensions) on the use of the IS. Differences among cultures in the ways in which they approach and interact with the IS are highly relevant to understanding how users use them in all settings.

Culture is one of the most relevant aspects of a user's personal and social context. Findings suggest that attitude formation is influenced by the objective characteristics of the IS, the extent of use, and individual and social users' differences. Studies continuously report that users are not always rational in selecting and using media and technologies, but attitudes toward and use of media and technologies are influenced by culture, norms, social contexts, or salient others (adapted from Fulk, Schmitz, & Schwartz, 1992; Rice & Love, 1987). Cultural differences (1) would be thus potentially critical to our understanding of IS acceptance and use; and (2) would play an important moderating role in determining how individuals make their decisions about adopting and using IS.

Cultural aspects need to be taken into account when developing IS that are especially to be used by a global audience. For instance, as Hermeking (2005) suggests, "a culturally well designed Web site may be defined as communicating the right information at the right place with the right layout in the right manner and in the right time according to the culture of each of its users."

Like any research, our study also has certain limitations. On the one hand, findings represent only preliminary tendencies. One of the main reasons for this is the high complexity and contingency of influences on IS design beyond cultural values and communication styles. On the other hand, Hofstede's method could thus be a significantly useful framework used in IS design. However, critics question the applicability of the dimensions to all cultures, emphasizing that "one can conjuncture that other types of samples might yield different dimensions and order of nations" (see Schwartz, 1994; Erez & Early, 1993; Soares et al., 2007). Most frequently, researchers question Hofstede's methodology and sample (e.g., Myers & Tan, 2002). Beyond age-based critiques, researchers also criticise Hofstede's dimensions for being data driven and not having a strong enough base in theory (e.g., Smith & Schwartz, 1997). Lastly, as Baack and Singh (2007, p. 182) summarise, "scholars point out that Hofstede's survey is specific to work values and may not apply to marketing research." Straub et al. (2002) argued that individuals may or may not identify with the national culture and they can show different cultural orientation even though they are in the same country. The social identity theory enables IS researchers to have a theoretical framework for studying at an individual level with a complimentary research perspective.

REFERENCES

Baack, D.W., & Singh, N. (2007). Culture and Web communications. *Journal of Business Research, 60*, 181-188.

Babin, B.J, William, R.D., & Griffinn, M. (1994, March). Work and/or fun: Measuring hedonic and utilitarian shopping value. *Journal of Consumer Research, 20*, 644-656.

Bagchi, K., Cerveny, R., Hart, P., & Peterson, M. (2003). The influence of national culture in information technology product adoption. In *Paper Presented at Proceedings of the Ninth Americas Conference on Information Systems,* (pp. 957-965).

Barber, W., & Badre, A.N. (2001). Culturability: The merging of culture and usability. In *Proceedings of the 4th Conference on Human Factors and the Web,* Basking Ridge, NJ, USA.

Bryson, M., & De Castell, S. (1998). New technologies and the cultural ecology of primary schooling: Imagining teachers as luddities in/deed. *Educational Policy, 12*(5), 542-567.

Chau, P.Y.K. (1996). An empirical assessment of a modified technology acceptance model. *Journal of Management Information Systems, 13*, 185-204.

Chau, P.Y.K., Cole, M., Massey, A., Montoya-Weiss, M., & O'Keefe, R.M. (2002). Cultural differences in consumers online behaviors. *Communications of the ACM, 45*(10), 138-143.

Cohen, A., & Avrahami, A. (2007). The Relationship between individualism, collectivism, the perception of justice, demographic characteristics and organisational citizenship behaviour. *The Service Industries Journal, 26*(8), 889-901.

Cronin, M. J. (1996). *Global advantage on Internet: From corporate connectivity to international competitiveness*. New York: Van Nostrand Reinhold.

Del Galdo, E., & Neilson, J. (1996). *International user interfaces*. New York: John Wiley & Sons.

Erez, M., & Early, P.C. (1993). *Culture, self-identity, and work*. New York: Oxford University Press.

Fulk, J., Schmitz, J.A., & Schwartz, D. (1992). The dynamics of context-behaviour interactions in computer-mediated communication. In M. Lea (Eds.), *Contexts of computer-mediated communication* (pp. 7-29). New York: Harvester Wheatsheaf.

Goodman, S. E., Press, L.I., Ruth, S.R., & Rutkowski, A.M. (1994). The global diffusion of the Internet: Patterns and problems. *Communications of the ACM, 37*(8), 27-31.

Gouveia, V.V., & Ros, M. (2000). The Hofstede and Schwartz models for classifying individualism at the cultural level: Their relation to macro-social and macro-economic variables. *Psicothema, 12*, 25-33.

Hall, E.T. (1989). *Beyond culture*. Garden City, NY: Doubleday.

Herbig, P.A. (1994). *The innovation matrix: Culture and structure prerequisites to innovation*. Westport, CT: Quorum Books.

Hermeking, M. (2005). Culture and Internet consumption: Contributions from cross-cultural marketing and advertising research. *Journal of Computer-mediated Communication, 11*(1). Retrieved December 12, 2007, from http://jcmc.indiana.edu/vol11/issue1/hermeking.html

Hofstede, G. (1980). *Cultures consequences: International differences in work related values*. Beverly Hills: Sage Publications.

Hofstede, G. (1984).Culture's consequences. *International differences in work-related values*. London: Sage.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind.* Berkshire: McGraw-Hill.

Hofstede, G., & Bond, M.H. (1988). The Confucian connection: From cultural roots to economic growth. *Organizational Dynamics, 16*(4), 4-21.

Johns, S. K., Smith, M., & Strand, C.A. (2003). How culture affects the use of information technology. *Accounting Forum, 27*(1), 84-109.

Keil, M., Beranek, P.M., & Konsynski, B.R. (1995). Usefulness and ease of use: Field study evidence regarding task considerations. *Decision Support Systems, 13*(1), 75-91.

Keil, M., Tan, B.C.Y., Wei, K.K., Saarinen, T., Tuunainen, V., & Wassenaar, A. (2000). A cross-cultural study on escalation of commitment behavior in software projects. *MIS Quarterly, 24*(2), 299-325.

Kovacic, Z. (2005). The impact of national culture on worldwide e-government readiness. *Informing Science, 8*, 143-158.

Littleton, K., & Bannert, M. (1999). Gender and IT: Contextualizing differences. In J. Bliss, R. Sääljö, & P. Light (Eds.), *Learning sites: Social and technical resources for learning* (pp. 171-182). Amsterdam: Pergamon.

Maitland, C. (1999). Global diffusion of interactive networks. *The Impact of Culture AI & Society, 13*, 341-35.

Maitland, C., & Bauer, J. (2001). National level culture and global diffusion: The case of the Internet. In C. Ess (Ed.), *Culture, technology, communication: Towards an intercultural global village* (pp. 87-128). Albany, NY: State University of New York Press.

Marcus, A., & Gould, E.W. (2000). Cultural dimensions and global Web user interface design: What? So What? Now What? In *Proceedings of the 6th Conference on Human Factors & the Web*, Austin, TX.

Merchant, J. E. (2002). Communicating across borders: A proposed model for understanding cross-cultural issues for the successful strategic implementation of information systems. In *Proceedings of InSITE 2002,* (pp. 1031-1040).

Meyers-Levy, J., & Maheswaran, D. (1991). Exploring differences in males and females processing strategies. *Journal of Consumer Research, 18*, 63-70.

Meyers-Levy, J., & Sternthal, B. (1991). Gender differences in the use of message cues and judgments. *Journal of Marketing Research, 28*, 84-96.

Morris, M., & Venkatesh, V. (2000). Age differences in technology adoption decisions: Implications for a changing workforce. *Personnel Psychology, 5*(3), 375-403.

Myers, M.D., & Tan, F.B. (2002). Beyond models of national culture in information systems research. *Journal of Global Information Management, 10*(1), 24-32.

Pavlou, P.A., & Chai, L. (2002). What drives electronic commerce across cultures? A cross-cultural empirical investigation of the theory of planned behaviour. *Journal of Electronic-commerce Research, 3*(4), 240-253.

Rice, R. E., & Love, G. (1987). Electronic emotion: Socio-emotional content in a computer-mediated communication network. *Communication Research, 14*(1), 85-105.

Samovar, L.A., Porter, R.E., & Jain, N.C. (1981). *Understanding intercultural communication*. Belmont, CA: Wadsworth.

Schwartz, S.H. (1994). Cultural dimensions of values: Toward an understanding of national differences. In U. Kam, H.C. Triandis, C. Kagitcibasi, S.C. Choi, & G. Yoon (Eds), *Individualism and collectivism: Theory, method, and application* (pp. 85-119). Newbury Park, CA: Sage.

Smith, P.B., & Schwartz, S.H. (1997). Values. In B. Segall & C. Kagitcibasi (Eds.). *Handbook of cross-cultural psychology: Social and behaviour application* (pp. 77-118). Massachusetts: Allyn and Bacon.

Soares, A.M., Farhangmehr, M., & Shoham, A. (2007). Hofstede's dimensions of culture in international marketing studies. *Journal of Business Research, 60*, 277-284.

Sørnes, J-O., Stephens, K.K., Sætre, A.S., & Browning, L.D. (2004). The reflexivity between ICTs and business culture: Applying Hofstede's theory to compare Norway and the United States. *Informing Science Journal, 7*, 1-30.

Staw, B., Sandelands, L., & Dutton, J. (1981). Threat-rigidity effects in organizational behavior: A multi-level analysis. *Administrative Science Quarterly, 26*, 501-524.

Steenkamp, J.B.E.M., Ter Hofstede, F., & Wedel, M. (1999). A cross-national investigation into the individual and national cultural antecedents of consumer innovativeness. *Journal of Marketing, 63*(2), 55-69.

Stohl, C. (2001). Globalizing organizational communication. In F. Jablin & L. Putnam (Eds.), *The new handbook of organizational communication (*pp. 323-375). Thousand Oaks, CA: Sage.

Straub, D., Loch, K., Evaristo, R., & Karahanna, E., & Strite, M. (2002). Toward a theory-based measurement of culture. *Journal of Global Information Management, 10*(1), 13-23.

Straub, D. W. (1994). The effect of culture on IT diffusion: E-mail and fax in Japan and the U.S. *Information Systems Research, 5*(1), 23-47.

Sun, H., & Zhang, P. (2006). The role of moderating factors in user technology acceptance. *International Journal of Human-Computer Studies, 64*, 53-78.

Trompenaar, F. (1995). *Riding the waves of culture. Understanding cultural diversity in business*. London.

Veiga, J.F., Floyd, S., & Dechant, K. (2001). Towards modeling the effects of national culture on IT implementation and acceptance. *Journal of Information Technology, 16*(3), 145-158.

Veiga, J.F., Lubatkin, M., Calori, R., & Very, P. (2000). Measuring organizational culture clashes: A two-nation post-hoc analysis of a cultural compatibility index. *Human Relations, 53*(4), 539-57.

Veiga, J.F., Yanouzas, J.N., & Buchholtz, A.K. (1993, June 1-4). Business practices: An exercise comparing U.S. and Russian managers. In Paper Presented at Proceedings of the Fifth Biennial International Management Conference of the Eastern Academy of Management, Berlin, Germany.

Venkatesh, V., & Morris, M. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behaviour. *MIS Quarterly, 24*(1), 115-139.

Watson, R.T., Ho, T.H., & Raman, K.S. (1994). Culture: A fourth dimension of group support systems research. *Communications of the ACM, 37*(10), 44-55.

Yeniyurt, S., & Townsend, J.D. (2003). Does culture explain acceptance of new products in a country? An empirical investigation. *International Marketing Review, 20*(4), 377-396.

KEY TERMS

Culture: Refers to the collective programming of the mind which distinguishes the members of one group from another.

Individualism (IDV): The degree to which individuals are integrated into groups. On the individualist side we find societies in which the ties between individuals are loose: everyone is expected to look after him/herself and his/her immediate family. On the collectivist side, we find societies in which people from birth onward are integrated into strong, cohesive in-groups, often extended families (with uncles, aunts and grandparents) which continue protecting them in exchange for unquestioning loyalty.

Masculinity (MAS): Refers to the distribution of roles between the genders which is another fundamental issue for any society to which a range of solutions are found.

Perceived Ease-of-Use (PEOU): Defined as the degree to which a person believes that using a particular IS would be free of effort.

Perceived Usefulness (PU): Defined as the degree to which a person believes that a singular IS would enhance his/her job performance.

Power Distance Index (PDI): The extent to which the less powerful members of organizations and institutions (like the family) accept and expect that power is distributed unequally.

Uncertainty Avoidance Index (UAI): Deals with a society's tolerance for uncertainty and ambiguity; it ultimately refers to man's search for Truth. It indicates to what extent a culture programs its members to feel either uncomfortable or comfortable in unstructured situations.

Culture and Anonymity in GSS Meetings

Moez Limayem
University of Arkansas, USA

Adel Hendaoui
University of Lausanne, Switzerland

INTRODUCTION

Managers spend a considerable part of their work time in meetings participating in group decision making. Group support systems (GSSs) are adopted in a variety of group settings—from within-organization team to multi-organization collaboration teams (Ackermann, Franco, Gallupe, & Parent, 2005)—to aid the decision-making process (Briggs, Nunamaker, & Sprague, 1998). A key characteristic of GSSs is anonymity, which improves various aspects of group performance, including improving group participation and communication, objectively evaluating ideas, and enhancing group productivity and the decision-making process (Nunamaker, Dennis, Valacich, Vogel, & George, 1991; Pinsonneault & Heppel, 1997; Postmes & Lea, 2000). Anonymity, as a distinct aspect of GSSs, was expected to increase productivity by reducing the level of social or production blocking, increasing the number of interpersonal exchanges, and reducing the probability of any one member dominating the meeting (Newby, Soutar, & Watson, 2003). For example, Barreto and Ellemers (2002) manipulated two aspects of anonymity separately: visibility of respondents (i.e., participants could or could not see who the other group members were) and visibility of responses (participants could or could not see the responses given by other group members). Results show that when group identification is low, anonymity manipulations affect group members' effort. Similarly, in their experiment, Reinig and Mejias (2004) found that anonymous groups produced more critical comments than identified groups did at the group level of analysis.

Numerous empirical findings have suggested that the use of anonymity and process structure in electronic brainstorming (EBS) generally promotes a positive effect on the number of ideas generated (Jessup, Connolly, & Galegher, 1990; Gallupe, Bastianutti, & Cooper, 1991) and quality of ideas achieved in decision making (Zigurs & Buckland, 1998). However, the anonymity function inherent in multi-workstation GSSs has been found to heighten conflict as members tend to communicate more aggressively because they tend to be more critical (Connolly, Jessup, & Valacich, 1990; Jessup, Connolly, & Tansik, 1990; Valacich, Jessup, Dennis, & Nunamaker, 1992), to have no effects on inhibition (Valacich, Dennis, & Connoly, 1994; Valacich et al., 1992), to increase group polarization (Sia, Tan, & Wei, 2002), and to have no effects on group performance (Valacich et al., 1994). Other studies show that, in terms of effectiveness, nominal brainstorming may be equal to (Gallupe et al., 1991; Cooper, Gallupe, Pollard, & Cadsby, 1998; Barki & Pinsonneault, 2001) or sometimes less than (Valacich et al., 1994; Dennis & Valacich, 1993) electronic brainstorming, indicating that at least as far as laboratory studies are concerned, empirical investigations have been inconclusive.

BACKGROUND

Ferraro (1998) provides a succinct definition of culture as follows: "Culture is everything that people have, think, and do as members of their society." Culture has been defined as the collective programming of the mind, which distinguishes the members of one group or category of people from another (Hofstede 1991; Tan, Watson, & Wei, 1995). Culture involves the beliefs, value system, and norms of a given organization or society, and can exist at national, regional, and corporate levels. In fact, even information systems theories and research are heavily influenced by the culture in which they were developed, and a theory grounded in one culture may not be applicable in other countries (Tan et al., 1995; Triandis, 1987). The theories explaining the effects of GSSs have come mainly from a North American perspective and may need adjustment for appropriate explanation of the same phenomenon in different contexts. Therefore, in order to incorporate a global dimension, theories and models that attempt to explain the effectiveness of technology will need to take into account the cultural background of the group being examined.

Hofstede (1991) identifies five dimensions of national culture based on his IBM study in 72 different countries:

- *Power distance* focuses on the degree of equality, or inequality, between people in a society. A high power distance ranking indicates that inequalities of power and wealth have been allowed to grow within that society. Similar societies—with high power distance—are more likely to follow a caste system that does not allow significant upward mobility of its citizens. A low power

distance ranking indicates that a society deemphasizes the differences between citizens' power and wealth. In these types of societies, equality and opportunity for everyone is stressed. Individuals in societies with low power distance cultures (e.g., the United States) may be more inclined to adopt technologies that reduce power distance (Reinig & Mejias, 2003). However, power distance effects can be helpful for some phases of group decision making but harmful for others (Tan, Watson, Wei, Raman, & Kerola, 1993).

- *Individualism* focuses on the degree in which a society reinforces individual or collective achievement and interpersonal relationships. This is opposed to *collectivism,* which implies a preference for a tightly knit social framework in which individuals can expect their relatives and clan to protect them in exchange for loyalty. A high individualism ranking indicates that individuality and individual rights are paramount within the society. Individuals in these societies may tend to form a larger number of looser relationships. A low individualism ranking typifies societies of a more collectivist nature with close ties between individuals. These cultures reinforce extended families and collectives where everyone takes responsibility for fellow members of their group. The people of collectivistic-culture societies (e.g., Hong Kong) tend to sustain group harmony and agreement, which exhibits less critical comments than those of individualistic-culture societies (e.g., the United States) in using group support systems (Reinig & Mejias, 2004). Likewise, Chinese participants, whose culture leans strongly toward collectivism, are more prone to follow the view of the majority, while Americans, whose culture leans strongly toward individualism, is less prone to follow the view of the majority (Zhang, Lowry, & Fu, 2006).
- *Masculinity* focuses on the degree in which the society reinforces, or does not reinforce, the traditional masculine work role model of male achievement, control, and power. On the contrary, *femininity* implies a preference for relationships, modesty, caring for the weak, and quality of life. A high masculinity ranking indicates that the country experiences a high degree of gender differentiation. In these cultures, males dominate a significant portion of the society and power structure, with females being controlled by male domination. A low masculinity ranking indicates the country has a low level of differentiation and discrimination between genders. In these cultures, females are treated equally to males in all aspects of the society.
- *Uncertainty avoidance* focuses on the level of tolerance for uncertainty and ambiguity within the society, that is, unstructured situations. A high uncertainty avoidance ranking indicates the country has a low tolerance for uncertainty and ambiguity. This creates a rule-oriented society that institutes laws, rules, regulations, and controls in order to reduce the amount of uncertainty. A low uncertainty avoidance ranking indicates the country has less concern about ambiguity and uncertainty, and more tolerance for a variety of opinions. This is reflected in a society that is less rule oriented, more readily accepts change, and takes more and greater risks.
- *Long-term orientation* focuses on the degree the society embraces, or does not embrace, long-term devotion to traditional, forward-thinking values. High long-term orientation ranking indicates the country prescribes to the values of long-term commitments and respect for tradition. This is thought to support a strong work ethic where long-term rewards are expected as a result of today's hard work. However, business may take longer to develop in this society, particularly for an "outsider." A low long-term orientation ranking indicates the country does not reinforce the concept of long-term, traditional orientation. In this culture, change can occur more rapidly, as long-term traditions and commitments do not become impediments to change.

It is interesting to note that power distance and individualism are found to be inversely related (Hofstede, 1991; Kim, Triandis, Kagitcibasi, Choi, & Yoon, 1994; Triandis, 1995). Many Western countries such as the United States, Great Britain, and Australia have been described as individualistic, low power distance cultures, while many Asian countries such as Hong Kong, Singapore, and China have been described as collectivistic, high power distance cultures (Hofstede, 1991).

More recently, Srite and Karahanna (2006) examined the influence of national culture on individual behavior and extended the Technology Acceptance Model by incorporating espoused national cultural values (masculinity/femininity, individualism/collectivism, power distance, and uncertainty avoidance) into the model. With respect to the impact of individualism/collectivism value on behavior for example, and because of the growing "virtualness" of collaborative teams, these authors call for further research investigating the acceptance of technologies used by teams composed of individuals from different national cultures.

CULTURE AND ANONYMITY IN GSS STUDIES

Although a GSS is a socio-technical system that involves not only computer and communication technologies but also a group of participants, culture was not specifically considered as an important dimension in the early studies of GSSs.

However, with globalization it is becoming increasingly important to adapt this tool to the cultural background of the organization or group that intends to use it effectively. These dimensions have been investigated in cross-culture GSS studies (such as Robichaux & Cooper, 1998; Tan, Wei, Watson, Clapper, & McLean, 1998; Tung & Quaddus, 2002; Watson, Ho, & Raman, 1994). Among the five dimensions, power distance and individualism have been shown to have impacts on group behavior and group outcomes (Tan et al., 1998; Watson et al., 1994). This is because the anonymity and simultaneous input features of GSSs support low power distance and individualistic cultural norms of desirable group behavior (Watson et al., 1994).

Watson et al. (1994) later provided empirical support for the inclusion of culture as a dimension of GSSs to add to DeSanctis and Gallupe's (1987) dimensions of group size, member proximity, and task type. Their study examined American and Singaporean cultures using GSSs, and the findings suggested that Singaporean groups tended to have a higher pre-meeting consensus and less change in consensus than the U.S. group. This may be explained with reference to the collectivist nature of Singaporean culture, as collectivists have a tendency towards group consensus (Mejias, Shepherd, Vogel, & Lasaneo, 1997).

Tan et al. (1995) suggested ways that different cultures can be studied with other important variables such as task type and group size. The study focused on finding a way to examine the robustness of previous and current GSS research across different cultures and to add a cultural perspective to existing GSS knowledge. Hofstede's dimension of power distance was examined in relation to GSSs, and the possible impacts of a GSS intervention in both high and low power distance countries were explored.

In studies examining only Singaporean groups (Tan et al., 1995), the use of a GSS resulted in a decreased impact of status and normative influences on decision making. These findings showed that change in consensus was greater in American groups than it was in than Singaporean groups, and influence was more equal in Singaporean groups than it was in American groups. The higher power distance of Singaporean groups may explain the differences between these two meeting outcomes, and the study supports the proposition that a GSS can overcome the effect of high power distance on group meetings.

A study comparing North American and Mexican groups participating in GSS sessions showed differences in terms of perception of consensus and satisfaction levels of group members (Mejias et al., 1997). American and Mexican groups were also studied for GSS effects on participation equity, with Mexican groups reporting higher participation equity levels than American GSS groups (Mejias et al., 1997). It was suggested that high power distance cultures benefit from GSSs, and that these findings indicate that culture has a significant bearing on crucial aspects of GSS meeting outcomes.

When members interact at different locations and at different times, GSS teams are considered virtual teams. Globalization implies that virtual collaborative teams are often made up of participants with different backgrounds (national culture, spoken language, and value system). Carte and Chidambaram (2004) suggested that GSS features can reduce the negative effects of cultural diversity (i.e., communication difficulties, misunderstandings, decreased cohesion, and increased conflict) early in the life of a culturally diverse team. Based on these findings, Staples and Zhao (2006) provided evidence that cultural diversity effects are different depending on the communication mode used between members (e.g., face-to-face vs. virtual-electronic) and specifically that the performance of heterogeneous teams using electronic channels of communication was higher than heterogeneous teams using face-to-face communication. The findings could have implications for the design and use of anonymity features in GSS-supported meetings.

Based on Hofstede (1991), Shin and Higa (2005) investigated the attitude of people with a particular cultural background toward different scheduling approaches. They found that face-to-face coordination led to higher group satisfaction than automated approaches did. Extending the implications of their study, their findings suggest that organizational cultures that show strong collectivism and social interactivity may hinder adoption and use of virtual technologies such as GSSs.

The feature of anonymity in GSSs facilitates group processes by moderating those participants who dominate group discussions, by hiding the identities of the participants to eliminate the influence of authority, and by removing the reliance on nonverbal cues in communication between group members. Some researchers have hypothesized that anonymity enhances group member participation by reducing inhibitions. For example, Wilson and Jessup (1995) found that groups interacting under anonymity generate more total comments and unique ideas, more ideas of higher rarity, and more critical comments than groups interacting without anonymity.

In his study of the effect of individuals' personality characteristics on their participation in a decision-making meeting, Hartmann (2001) investigated also the moderating effect by anonymity. It is found that anonymity allows the disagreeable persons (i.e., those who are less willing to help others) to participate more and provide more on-task comments with a high level of anonymity than they would with a low level of anonymity in a GSS-supported meeting. However, groups with agreeable individuals will perform better, but anonymity will allow the disagreeable individuals to participate more and therefore create more conflict, resulting in poor performance from the group.

Limayem, Khalifa, and Coombes (2003) conducted a study to explain the different effects of anonymity on the behavior of Hong Kong and Canadian groups during GSS

sessions. In the Hong Kong Chinese culture, group interactions tend to emphasize harmony, conformance, and reciprocal respect rather than openness and spontaneity. However, the Canadian group's culture, which frequently exhibits openness and spontaneity, will usually allow individuals to deviate from the norm. Anonymity was found to have more significant positive effects for Hong Kong groups. With anonymity, the performance of the Hong Kong group improved significantly in terms of number of contributions, quality of contributions, and perceived level of participation. No significant differences in the performance were found for the Canadian groups, except for the quality of contributions, which deteriorated with anonymity. A qualitative analysis of this negative effect revealed social loafing and lack of accountability as possible causes. This finding is consistent with the previous study that anonymity induces social loafing and flaming, reduces accountability, and ultimately may decrease participation (Er & Ng, 1995; Pinsonneault & Kraemer, 1990).

In a study exploring the usefulness of electronic brainstorming (a component of GSS), Dennis and Reinicke (2004) note that reinforcing the group culture is considerably more difficult, especially when brainstorming is anonymous. According to theses authors, adoption of such technologies depends on the relative importance of existing culture and power structures to key group members, because of the difficulty to identify and sanction those members who challenge current structures. Moreover, they suggest that if anonymity may improve performance, removing it may increase overall usefulness. They finally called for further research to predict adoption of GSS technologies with more refined versions of the Technology Acceptance Model.

In sum, studies investigating the use of GSS by people from different cultures have indicated that culture has a significant impact on GSS usage, and that cultural dimensions, such as those proposed by Hofstede (1991), have some relevance in explaining these differences. However, there is still uncertainty as to the specific impacts of culture on the performance of groups in anonymous GSS sessions, and therefore more must be done to clearly understand how different cultures respond to anonymity.

FUTURE TRENDS

Migrating to the era of virtual organization, it is more and more common that cross-organization project teams rely on GSSs to accomplish their daily tasks. Since team members are often from different cultural backgrounds, the optimal performance of the GSS relies on the cultural context in which it operates. Hence, the practical implications of cultural studies in GSS research are of significant value not only to facilitators of GSS, but also to users of groupware applications, as well as many other inter-organizational electronic communication systems.

In the same vain, anonymity in GSSs per se involves no value judgment at all. One cannot simply judge that it is a favorable or an unfavorable feature, especially when participants with different cultural backgrounds are involved. The use of anonymity should depend on the cultural context in which it is applied. For example, the use of anonymity in GSSs is likely to result in better meeting performance in groups that do not emphasize status hierarchies, conformance, mutual obligation, and reciprocity than in groups that emphasize on these qualities. In the former situation, anonymity could even lead to negative outcomes such as social loafing due to the reduction in motivation and effort that occurs when individuals work in anonymous groups. Conversely, it may be beneficial to use anonymity for GSS-supported groups with cultures that normally exhibit higher levels of conformance pressure and evaluation apprehension.

In a broader sense, GSS and groupware designers and developers should pay special attention to the implementation of anonymity features. For example, they could make it easier for users to turn these features on and off to accommodate the culture of the groups using the systems. Finally, facilitators should remember that studies suggest that culture influences participation in the GSS environment (Tung & Quaddus, 2002). Therefore, facilitators should study the culture of the group using the technology before blindly using anonymity to generate or evaluate ideas.

CONCLUSION

Culture is obviously an important factor affecting a group's response to anonymity in the GSS context. Cultural effect on group structure and evaluation apprehension is also an important consideration for designers, facilitators, and users of GSSs. Considering the rather rare number of studies investigating the impact of culture on GSSs, further research in this area is warranted. An interesting line of research in GSSs for the future would be to isolate the relative impact of anonymity in the context of different cultures engaged in different tasks and situations. The knowledge gained from this and other continuing studies will assist in the effective application of GSSs in increasingly diverse and global contexts.

ACKNOWLEDGMENT

The work described in this article was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. 9040564).

REFERENCES

Ackermann, F., Franco, L.A., Gallupe, B., & Parent, M. (2005). GSS for multi-organizational collaboration: Reflections on process and content. *Group Decision and Negotiation, 14*(4), 307-331.

Barreto, M., & Ellemers, N. (2002). The impact of anonymity and group identification on pro-group behavior in computer-mediated groups. *Small Group Research, 33*(5), 590-610.

Barki, H., & Pinsonneault, A. (2001). Small group brainstorming and idea quality is electronic brainstorming the most effective approach? *Small Group Research, 32*(2), 158-205.

Briggs, R.O., Nunamaker J.F., & Sprague, R.H. (1998). 1001 unanswered research questions in GSS. *Journal of Management Information Systems, 14*(3), 3-21.

Carte, T., & Chidambaram, L. (2004). A capabilities-based theory of technology deployment in diverse teams: Leapfrogging the pitfalls of diversity and leveraging its potential with collaborative technology. *Journal of the AIS, 5*(11-12), 448-471.

Connolly, T., Jessup, L.M., & Valacich, J.S. (1990). Effects of anonymity and evaluative tone on idea generation. *Management Science, 36*(6), 689-704.

Cooper, W.H., Gallupe, R.B., Pollard, S., & Cadsby, J. (1998). Some liberating effects of anonymous electronic brainstorming. *Small Group Research, 29*(2), 147-178.

Dennis, A.R., & Valacich, J.S. (1993). Computer brainstorms: More heads are better than one. *Journal of Applied Psychology, 78*(4), 531-538.

Dennis, A.R., & Reinicke, B.A. (2004). Beta versus VHS and the acceptance of electronic brainstorming technology. *MIS Quarterly, 28*(1), 1-20.

DeSanctis, G.L., & Gallupe, R.B. (1987). A foundation for the study of group decision support systems. *Management Science, 33*(5), 589-609.

Er, M.C., & Ng, A.C. (1995). The anonymity and proximity factors in group decision support systems. *Decision Support Systems, 14*(1), 75-83.

Ferraro, G.P. (1998). *The cultural dimensions of international business.* Englewood Cliffs, NJ: Prentice Hall.

Gallupe, R.B, Bastianutti, L., & Cooper, W.H. (1991). Unblocking brainstorms. *Journal of Applied Psychology, 76*(1), 137-142.

Hartmann, R.E. (2001, March). *Influence of personality type and anonymity on participation in a group support system.* MS Thesis, AFIT/GIR/ENV/01M-09. Air Force Institute of Technology, USA.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind.* London: McGraw-Hill.

Jessup, L.M., Connolly, T., & Galegher, J. (1990). The effects of anonymity on GDSS group process with an idea-generating task. *MIS Quarterly, 14*(3), 312-321.

Jessup, L.M., Connolly, T., & Tansik, D.A. (1990). Toward a theory of automated group work: The deindividuating effects of anonymity. *Small Group Research, 21*(3), 333-348.

Kim, U., Triandis, H.C., Kagitcibasi, C., Choi, S.C., & Yoon, G. (1994). Individualism and collectivism: Theory, methods and applications. Thousand Oaks, CA: Sage.

Limayem, M., Khalifa, M., & Coombes, J.M. (2003). Culture and anonymity in GSS meetings. In G. Ditsa (Ed.), *Information management: Support systems and multimedia technology* (pp. 150-161). Hershey, PA: Idea Group.

Mejias, R.J., Shepherd, M.M., Vogel, D.R., & Lasaneo, L. (1997). Consensus and perceived satisfaction levels: A cross cultural comparison of GSS and non-GSS outcomes within and between the United States and Mexico. *Journal of Management Information Systems, 13*(3), 137-161.

Newby, R., Soutar, G., & Watson, J. (2003). Comparing traditional focus groups with a group support systems (GSS) approach for use in SME research. *International Small Business Journal, 21*(4), 421-433.

Nunamaker, J.F., Dennis, A.R., Valacich, J.S., Vogel, D.R., & George, J.F. (1991). Electronic meeting systems to support group work. *Communications of the ACM, 34*(7), 40-61.

Pinsonneault, A., & Heppel, N. (1997). Anonymity in group support systems research: A new conceptualization, measure, and contingency framework. *Journal of Management Information Systems, 14*(3), 89-108.

Pinsonneault, A., & Kraemer, K.L. (1990). The effects of electronic meetings on group processes and outcomes: An assessment of the empirical research. *European Journal of Operational Research, 46*(2), 143-161.

Postmes, T., & Lea, M. (2000). Social processes and group decision making: Anonymity in group decision support systems. *Ergonomics, 43*(8), 1252-1274.

Reinig, B.A., & Mejias, R.J. (2003, January 6-9). An investigation of the influence of national culture and group support systems on group processes and outcomes. *Proceedings of the 36th Hawaii International Conference on System Sciences* (HICSS'03), Big Island, HI.

Reinig, B.A., & Mejias, R.J. (2004). The effects of national culture and anonymity on flaming and criticalness in GSS-supported discussions. *Small Group Research, 35*(6), 698-723.

Robichaux, B.P., & Cooper, R.B. (1998). GSS participation: A cultural examination. *Information & Management, 33*(6), 287-300.

Shin, B., & Higa, K. (2005). Meeting scheduling: Face-to-face, automatic scheduler, and email based coordination. *Journal of Organizational Computing and Electronic Commerce, 15*(2), 137-159.

Sia, C.-L., Tan, B.C.Y., & Wei, K.-K. (2002). Group polarization and computer-mediated communication: Effects of communication cues, social presence, and anonymity. *Information Systems Research, 13*(1), 70-90.

Srite, M., & Karahanna, E. (2006). The role of espoused national cultural values in technology acceptance. *MIS Quarterly, 30*(3), 679-704.

Staples, D.S., & Zhao, L. (2006). The effects of cultural diversity in virtual teams versus face-to-face teams. *Group Decision and Negotiation, 15,* 389-406.

Tan, B.C.Y., Watson, R.T., & Wei, K.K. (1995). National culture and group support systems: Filtering communication to damped power differentials. *European Journal of Information Systems, 4*(2), 82-92.

Tan, B.C.Y., Watson, R.T., Wei, K.K., Raman, K.S., & Kerola, P.K. (1993, January 5-8). National culture and group support systems: Examining the situation where some people are more equal than others. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (HICSS'93), Wailea, HI.

Tan, B.C.Y., Wei, K.K., Watson, R.T., Clapper, D.L., & McLean, E.R. (1998). Computer-mediated communication and majority influence: Assessing the impact in an individualistic and a collectivistic culture. *Management Science, 44*(9), 1263-1278.

Triandis, H.C. (1987). Individualism and social psychological theory. In C. Kagitcihbasi (Ed.), *Growth and progress in cross-cultural psychology* (pp. 78-83). Lisse, The Netherlands: Swets and Seitlinger.

Triandis, H.C. (1995). *Individualism and collectivism.* Boulder, CO: WestView.

Tung, L.L., & Quaddus, M.A. (2002). Cultural differences explaining the differences in results in GSS: Implications for the next decade. *Decision Support Systems, 33*(2), 177-199.

Valacich, J.S., Dennis, A.R., & Connoly, T. (1994). Idea generation in computer based groups: A new ending to an old story. *Organizational Behavior and Human Decision Processes, 57*(3), 448-468.

Valacich, J.S., Jessup, L.M., Dennis, A.R., & Nunamaker, J.F. Jr. (1992). A conceptual framework of anonymity in group support systems. *Group Decision and Negotiation, 1,* 219-241.

Watson, R.T., Ho, T.H., & Raman, K.S. (1994). A fourth dimension of group support systems. *Communications of the ACM, 37*(10), 45-55.

Wilson, J., & Jessup, L.M. (1995, January 3-6). A field experiment on GSS anonymity and group member status. *Proceedings of the 28th Annual Hawaii International Conference on System Sciences* (HICSS'95) (pp. 212-221), Maui, HI.

Zhang, D., Lowry, P.B., & Fu, X. (2006, January 4-7). Culture and media effects on group decision making under majority influence. *Proceedings of the 39th Annual Hawaii International Conference on System Sciences* (HICSS'06).

Zigurs, I., & Buckland, B.K. (1998). A theory of task/technology fit and group support systems effectiveness. *MIS Quarterly, 22*(3), 313-334.

KEY TERMS

Anonymity in GSS: The situation when participants' names are not made public in a GSS environment.

Culture: The collective programming of the mind, which distinguishes the members of one group or category of people from another.

Group Support System (GSS): Any combination of hardware and software that enhances groupwork.

Individualism: A preference for a loose-knit social framework in a society in which individuals are only supposed to take care of themselves and their immediate families. This is opposed to collectivism, which implies a preference for a tightly knit social framework in which individuals can expect their relatives and clan to protect them in exchange for loyalty.

Long-Term Orientation: The fostering of virtues oriented towards future rewards, in particular perseverance and thrift. Its opposite pole, short-term orientation, stands for the fostering of virtues related to the past and present, in particular, respect for tradition, preservation of 'face', and fulfilling social obligations.

Masculinity: A preference for achievement, heroism, assertiveness, and material success; as opposed to femininity, which implies a preference for relationships, modesty, caring for the weak, and quality of life.

Power Distance: The extent to which a society accepts the fact that power in institutions and organizations is unevenly distributed.

Uncertainty Avoidance: The degree to which a society feels threatened by uncertain and ambiguous situations, which leads its members to support beliefs promising certainty and to maintain institutions protecting conformity.

Current Network Security Technology

Göran Pulkkis
Arcada Polytechnic, Finland

Kaj Grahn
Arcada Polytechnic, Finland

Peik Åström
Utimaco Safeware Oy, Finland

INTRODUCTION

Network security is defined as "a set of procedures, practices and technologies for protecting network servers, network users and their surrounding organizations" (Oppliger, 2000, Preface). The need for network security is caused by the introduction of distributed systems, networks, and facilities for data communication. Improved network security is required because of the rapid development of communication networks. Network security is achieved by using software- and hardware-based solutions and tools.

BACKGROUND

This article gives a topical overview of network security technologies, that is, the topics are not covered in detail, and most topics are briefly introduced and left for further study. The main objective is to present "state-of-the-art" network security technologies and to stimulate discussion about related skills and education needed by network users, IT professionals, and network security specialists.

PROTECTION AGAINST MALICIOUS PROGRAMS

Malicious software exploits vulnerabilities in computing systems. Malicious program categories are (Bowles & Pelaez, 1992):

- **Host Program Needed:** Trap door, logic bomb, Trojan horse, and virus.
- **Self-Contained Malicious Program:** Bacteria and worm.
- **Malicious Software Used by an Intruder after Gaining Access to a Computer System:** Rootkit.

Threats commonly known as adware and spyware have proliferated over the last few years. Such programs utilize advanced virus technologies for the reason to gather marketing information or display advertisements in order to generate revenue (Chien, 2005).

Modern malicious programs (including adaware and spyware) employ anti-removal and stealth techniques as well as rootkits to hide and to prevent detection. Rootkits conceal running processes, files, or system data. This helps an intruder to maintain system access in a way, which can be extremely difficult to detect with known security administration methods and tools. Rootkits are known to exist for a variety of operating systems such as Linux, Solaris, and versions of Microsoft Windows. A computer with a rootkit on it is called a rooted computer (Hoglund & Butler, 2005; Levine, Grizzard, & Owen, 2006).

The ideal protection is prevention, which still must be combined with detection, identification, and removal of such malicious programs for which prevention fails. Protection software is usually called antivirus software, which is characterized by generations (Stephenson, 1993):

- **First Generation:** Simple scanners searching files for known virus "signatures" and checking executable files for length changes.
- **Second Generation:** Scanners using heuristic rules and integrity checking to find virus infection.
- **Third Generation**. Memory resident "activity traps" identifying virus actions like opening executable files in write mode, file system scanning, and so forth.
- **Fourth Generation:** Software packages using many different antivirus techniques in conjunction.

Anti-adware/spyware modules are usually integrated in these software packages.

Protection levels of modern antivirus software are:

- **Gateway Level Protection:** Consists of mail server and firewall protection. Viruses are detected and removed before files and scripts reach a local network.
- **File-Server-Level Protection:** Consists of server software. Viruses are detected and removed even before network users access their files/scripts.

- **End-User-Level Protection:** Consists of workstation software. Viruses undetected in outer defense lines are detected and removed. However, this level is the only antivirus protection level for data communication, which is end user encrypted.

All levels should be combined to achieve depth in antivirus defense. Virus definition databases should be automatically and/or manually updated.

Examples of antivirus and anti-spyware software are Ad-Aware, F-Secure Internet Security, and Norton AntiVirus.

FIREWALL TECHNOLOGY

Firewalls protect computers and computer networks from external security threats. Firewalls fall into four broad categories (Stallings, 2006):

- **Packet-Filtering Router:** Applies a software and/or hardware implemented filtering rule set to each incoming/outgoing IP packet and then forwards or discards the packet. Most TCP/IP routers support basic user defined filtering rules. A packet-filtering firewall can also be a stand-alone network link device, for example, a computer with two network cards.
- **Application-Level Gateway (Proxy Server):** Acts as an application level traffic relay, that is, traffic is filtered based on specified application rules. A typical application level gateway is a protocol oriented proxy server on a network link, for example, an HTTP proxy, a SMTP proxy, a FTP proxy, and so forth.
- **Circuit-Level Gateway:** Typically relays TCP packets from one connection to another without examining the contents. Traffic is filtered based on specified session rules such as when a session is initiated by a recognized computer.
- **Stateful Multilayer Inspection Firewall:** Traffic is filtered at three levels, based on a wide range of specified application, session, and packet filtering rules.

CRYPTOGRAPHIC TECHNOLOGY

Cryptographic network security technology consists of network security applications, network security system software, and cryptographic hardware.

Secure-Network-Level Data Communication

Secure-network-level data communication is based on the Internet protocol security (IPSec) protocol. Two computers in the same TCP/IP network implement end-to-end security through the network, when IPSec software is installed and properly configured in both computers. IPSec provides two operation modes:

- **Transport Mode:** Original IP headers are used.
- **Tunnel Mode:** New IP headers are created and used to represent the IP tunnel endpoint addresses.

IPSec is usually embedded in virtual private network (VPN) software. VPN provides secure LAN functionality in geographically distributed network segments and for Internet connected computers. Fundamental VPN types are:

- **Access VPN:** Secure connection to a LAN through a public TCP/IP Network.
- **Connection VPN:** Secure remote connection between two logical LAN segments through a public TCP/IP network.

IPSec and VPN functionality is included in Windows 2000/XP. Commercial VPN software products are F-Secure VPN+™, Nokia VPN, Cisco Security VPN Software, and so forth. Open source IPSec and VPN software is also available (Openswan Portal, 2006).

Middleware

Middleware is a software layer between the network and the applications for providing services like identification, authentication, authorization, directories, and security (Internet2 Middleware Initiative [I2-MI] Portal, 2006). Shibboleth is an example of open source authentication and authorization middleware (Shibboleth Project Portal, 2006). Commercial security middleware based on the SHH protocol is SSH Tectia Solution (2006).

Secure-Transport-Level Data Communication

Many network applications are based on the IETF transport layer security (TLS) standard (Dierks & Rescora, 2006). The TLS/SSL protocol is based on an established client-server TCP connection. Then both computers execute the SSL handshake protocol to agree on the cryptographic algorithms and keys for use in the actual data communication. TLS/SSL versions of common application level TCP/IP protocols are available (see Table 1).

VPN solutions can also be implemented using the TLS/SSL protocol and executed on the transport level. This technology, called SSL-VPN, provides VPN functionality to geographically distributed network segments and for Internet connected computers using a standard Web browser.

Table1. Secure application level protocols based on TLS/SSL (Oppliger, 2000, p. 135)

Secure protocol	Port	Description
HTTPS	443	TLS/SSL protected HTTP
POP3S	995	TLS/SSL protected POP3
IMAPS	993	TLS/SSL protected IMAP4
SMTPS	465	TLS/SSL protected SMTP
NNTPS	563	TLS/SSL protected NNTP
LDAPS	636	TLS/SSL protected LDAP

Open source SSL-VPN software can be downloaded from OpenVPN Portal (2005).

Web Security

Basic Web security features are access level security and transaction level security. Access level security is provided with firewalls, which guard against intrusion and unauthorized use. Transaction level security requires protocols for protecting the communication between a Web browser and a Web server. Proposed protocols are HTTPS, S-HTTP, and PCT (Pulkkis, Grahn, & Åström, 2003). HTTPS was originally introduced by Netscape for the Navigator browser. Presently HTTPS is an accepted standard supported by practically all Web browsers, while S-HTTP and PCT are seldom used.

E-Mail Security

E-mail traffic between e-mail servers is protected using the SMTPS protocol. Sessions between e-mail client programs and e-mail servers can be protected

- By using the mailbox access protocols POP3S and IMAPS.
- By embedding an e-mail client program in a HTTPS Web page.

E-mail content security requires solutions for signing and/or encrypting outgoing messages as well as for decryption and/or signature verification of incoming messages. These solutions can be adapted on

- Client level, by e-mail client program security extensions.
- Server level, by gateway security extension solutions.

The most widely used e-mail security extensions are PGP and S/MIME (see Stallings, 2006, Chap. 15). A commercial e-mail security extension solution is Utimaco Safeware's SecuE-Mail Gateway supporting OpenPGP (Open PGP) Alliance Portal, 2006) and S/MIME.

A current problem with e-mail is spam e-mail sent by some unknown party to a large number of recipients. Usually this spam e-mail has some commercial contents. Spam e-mail is also used to spread spyware and viruses. Server level solutions detect spam e-mail before they reach the e-mail server and client solutions, embedded in modern client level security suites, detect spam e-mail when e-mail reaches the e-mail client.

E-Commerce Security

There are three main e-commerce transaction categories:

- **Consumer-to-Business (C2B) Transactions:** Occur between a consumer and an electronic marketplace or a bank over public networks, usually over the Internet.
- **Business-to-Business (B2B) Transactions:** Called market link transactions. Here, businesses, governments, and other organizations conduct business using different electronic communication technologies.
- **Intraorganizational Transactions:** Also called market driven transactions for internal strategies by collecting outside information and by customer monitoring (Kalakota & Whinston, 1999).

Secure Electronic Transaction (SET), introduced by MasterCard and Visa, is a standard protocol for securing credit card transactions over insecure networks such as the Internet (Stallings, 2000). SET provides secure communication, trust based on X.509v3 digital certificates, and privacy based on strictly controlled access to sensitive information. The SET protocol was published in the late 1990s but has still only a small market share in existing implementations of C2B transactions.

HTTPS is presently a standard protocol for securing C2B transactions on the Internet. When a customer browses to an e-commerce Web page, then authentication of this Web page with a trusted X.509v3 certificate is required before any transactions occur. A typical transaction is then a SSL/TLS protected authorization of the customer to the e-commerce Web page to charge the cost of a purchase from the credit card account of the customer. For online transactions with bank accounts, customers have private HTTPS protected Web pages. When a customer browses to his/her private Web page, then mutual authentication of the customer and the customer's bank is required before any transactions occur. In this authentication the bank uses a trusted X.509v3 certificate, and the customer uses either a trusted X.509v3 certificate or a one-time password according to the requirements of the bank. In this case a typical transaction is an SSL/TLS protected authorization signed by the customer to transfer a specified amount of money from the customer's account to some other account. The signature is created with a trusted X.509v3 certificate or with a randomly chosen signature code according to the requirements of the bank.

For B2B transactions, the main technologies are RosettaNet XML, Electronic Data Interchange (EDI), and EDI over the Internet (EDIoI).

RosettaNet is a consortium of major information technology (IT), electronic components (EC), and semiconductor manufacturing (SM) vendors dedicated to the development and deployment of open e-commerce standards for B2B transactions in high tech supply chains. *RosettaNet Implementation Framework* (RosettaNet Implementation Framework: Core Specification, 2001) is an open common networked application framework defining a XML format for exchange of B2B documents. This framework includes S/MIME v2 for secure authentication, authorization, and confidentiality of B2B transacting.

The latest version of the international EDI standard can be downloaded from UN/EDIFACT Portal (2006). For B2B transacting via EDI, the trading partners must agree on

- what information is to be exchanged,
- which message standards are used, and
- the means of transportation (EDI network).

An EDI network consists of direct modem-to-modem data links between involved companies or is implemented by a third-party value-added network (VAN) service. The security of VANs is high, since they are private networks physically out of the reach for outsiders. Examples of present EDI VAN service providers are AT&T, British Telecom, IBM network, and General Electric Information Services (Kalakota & Whinston, 1999; Whiteley, 2000).

B2B transacting via EDIoI means EDI network implementation by the Internet. EDI transactions are implemented by Web browsing operations, which can be mutually authenticated and protected by the HTTPS protocol. To support the use of EDIoI, IETF has standardized the Electronic Data Interchange Applicability Statement 2 protocol (AS2), which uses S/MIME messaging for authentication and data confidentiality (Moberg & Drummond, 2005).

Secure Shell (SSH)

The secure shell (SSH), a secure remote connectivity protocol in TCP/IP networks, is a de-facto standard being further developed by one of the IETF Security Area Working Groups. Two SSH versions have hitherto been developed: SSH1 and SSH2. Commercial as well as open source SSH implementations are available (OpenSSH Portal, 2006; SSH Tectia Solution, 2006).

Wireless Security Software

A radio interface is by nature easy to access. Security threats are either passive or active attacks. Active attacks involve altering data streams. Passive attacks, on the other hand, include snooping on transmission. The most important security features are authentication, authorization, confidentiality, integrity, and availability. The corresponding software is included in the network.

WLAN security is built up around the security protocol 802.11i/WPA2 (Wi-Fi Protected Access). WPA2 was created to address problems with the security protocols WEP (wired equivalent privacy) and WPA. For authentication, IEEE 802.1X is used in current systems (Wi-Fi Protected Access, 2003).

WiMAX (WiMAX Forum, 2006) has adopted the DOCSIS BPI+ (Data Over Cable Service Interface Specification – Baseline Privacy Interface Plus) protocol. Authentication relies on PKM-EAP (privacy key management-extensible authentication protocol) and TLS (transport layer security). The CCMP (counter mode with cipher block chaining message authentication code protocol) protocol and the AES (Advanced Encryption Standard) algorithm are used for encryption.

In Bluetooth, there are three security modes handled by the security manager (Grahn, Pulkkis, & Guillard, 2002). A bonding process including pairing and authentication, and encryption based on the SAFER+ algorithm are implemented. Also a concept of trusted devices is applied.

ZigBee (ZigBee Alliance, 2006) uses basic security elements in IEEE 802.15.4. The AES is used to protect data. Any two devices must share a key for encryption and decryption. The public key encryption algorithm is based on ECC (elliptic curve cryptography).

The security features in a GSM network can be divided into three subparts: subscriber identity authentication, user and signaling data confidentiality, and subscriber identity confidentiality. In 3G systems security is based on what was

implemented in GSM. The encryption algorithm is stronger; the application of authentication algorithms is stricter, and subscriber confidentiality is tighter. The security principles are all incorporated into the authentication and key agreement (AKA) procedure (Grahn et al., 2002).

Secure Network Management

A protocol for secure network management, SNMPv3, was introduced in 1998 by IETF to address the lack of security in earlier SNMP versions. SNMPv3 incorporates authentication and encryption features to SNMP managers and access control features to SNMP agents (Stallings, 2000).

Secure DNS (DNSSEC)

The absence of trust in DNS host name resolution is a security hazard in all TCP/IP applications. To address this problem IETF formed a Working Group to develop the DNSSEC standard. The objective is to provide both authentication and integrity to DNS information. DNSSEC uses public key cryptography to sign DNS information (DNSSEC, 2006).

Secure Routing Software

Routing protocols and their hardware/software implementations in computer networks are usually open and functionally unprotected. A manifestation of an emerging recognition of routing security in the Internet community is the recently formed IETF Routing Area Working Group "Routing Protocol Security Requirements (rpsec)," which in October 2004 published an Internet Draft "Generic Threats to Routing Protocols" (Barbir, Murphy, & Yang, 2004).

Cryptographic Hardware

Cryptographic hardware is needed for data protection and also computational acceleration purposes. A piece of cryptographic hardware is usually used for both purposes, and it is called:

- a hardware security module (HSM Module), when the goal is to achieve data security
- a crypto co-processor or cryptographic accelerator chip, when the goal is improved computational efficiency

HSM Modules are used for:

- protection of sensitive cryptographic data structures like symmetric and private cryptographic keys
- secure generation and use of sensitive cryptographic data structures, such as:
  - one-time passwords with short validity time
  - cryptographic keys
  - irreproducible random numbers needed in key generation and for nonce generation in authentication protocols to prevent replay
- execution of key agreement protocols
- execution of encryption and decryption operations using symmetric and private keys

The cryptographic keys in HSM Modules are protected by pin codes or biometrically by digital fingerprint comparison and/or by digital voice recognition.

Examples of cryptographic hardware are:

- Smartcard chips. Smartcard types are:
  - Electronic Identity Cards (PKI Cards)
  - SIM, PKI SIM, USIM, and SWIM cards in mobile phones
- USB HSM tokens
- PC Card HSM Modules with PCMCIA or PCMCIA Express interface
- PCI card HSM Modules
- SecurID (Nystrom, 2000) and Digipass (Vasco Product Range, 2006) HSM Modules for generation and use of one-time passwords with short validity time
- TRNG (True Random Number Generator) devices for extraction of natural physical randomness for generation of irreproducible random numbers. TRNG devices are implemented by:
  - radiation counters
  - radio noise monitors
  - audio noise monitors
  - monitors of thermal noise in diodes, leaky capacitors, mercury discharge tubes, and so forth
- cryptographic processors/acceleration chips for execution of
  - symmetric encryption/decryption operations with DES, 3DES, AES, and so forth
  - RSA encryption/decryption operations in public key cryptography
  - arithmetics with discrete points on elliptic curves in elliptic curve cryptography
  - SHA-1 hashing

With smartcards, the most widely used cryptographic hardware, a smartcard reader is needed. With other cryptographic hardware no separate reader is needed. For true pin code security a smartcard reader with a dedicated keypad is necessary. Software required for accessing cryptographic tokens on smartcards is

- Device driver for communication with the smartcard through the used smartcard reader.

- PC/SC, a specification set released by an international consortium (PC/SC Workgroup, 2006) for integration with the operating system. In PC/SC a device manager keeps track of the cards and card readers connected to a computer.
- An Application Programming Interface (API) like PKCS#11, also called CrypTokI, or Microsoft Crypto API.

Public Key Infrastructure (PKI)

Network server authentication is usually based on the use of certified public key cryptographic key pairs. In network access software, such as SSH and VPN, a network user authentication option is based on the use of certified key pairs. The server or the network is authenticated by proving the ownership of the private key in a certified key pair. The Internet standard for key pair certification is presently X.509v3 (Public-Key Infrastructure [X.509] Working Group [pkix], 2006). An X.509v3 certificate is a public digitally signed data structure consisting of:

- the public key of a key pair
- the subject (=owner) of the key pair
- validity time of the certificate
- usage of the key pair
- issuer of the certificate

Also digital signatures are created and verified with certified key pairs in the X.509v3 standard. An X.509v3 certificate is signed by the private key in the key pair of the issuer. The public key in an X.509v3 certificate is trusted, if the issuer is trusted Certification Authority (CA). A PKI is hardware, software, people, policies, and procedures needed to issue, manage, store, distribute, use, and revoke X.509v3 certificates.

SECURITY ADMINISTRATION

Security administration uses intrusion detection software, vulnerability checking software, and software for security software management.

An intrusion detection system (IDS) monitors traffic in a network and/or user behavior in a host computer to identify possible intruders and/or anomalous behavior and/or misuse (Stallings, 2006). A distributed intrusion detection system coordinates and brings cooperation among several intrusion detection systems across a whole network. Standards to support such distributed intrusion detection systems are defined by the IETF Intrusion Detection Working Group (Stallings, 2000).

Major vulnerabilities are too short, easily guessed, or cracked passwords. A potential intruder could run a password cracker on the encrypted passwords stored in a network. System administrators can use cracking to disable usage of bad passwords.

Intrusion prevention requires regular scans for unnecessary open ports and other vulnerabilities like missing security patches.

Data encryption software protects data stored in networks using encryption. Encryption per user or per group of data stored in files and databases protects data contents from unauthorized access. Data encryption software examples are:

- Microsoft's Encrypting File System (EFS) technology for file and folder encryption on user level.
- Utimaco Safeware's SafeGuard LAN Crypt software for file and folder encryption on both user and group level. SafeGuard LAN Crypt supports encrypted network traffic for the encrypted data.

Network security software in host computers and in other network nodes like routers is often software managed. A management software example is F-Secure® Policy Manager™ for management of "not only antivirus solutions, but all critical network security solutions on all tiers of the network" (F-Secure Policy Manager, 2006).

DEVELOPMENT OF SECURITY SOLUTIONS

Antivirus protection programming skills require knowledge about self-modifying programs/scripts and about virus sensitive operating system features.

Firewall software programming skills are based on detailed knowledge of TCP/IP protocol stack implementation software.

The open source toolkit OpenSSL is available for TLS/SSL application design (The OpenSSL Project, 2005). OpenSSL is installed as a C function library. Also commercial development tools are available, for example (Certicom Security Builder SSL, 2006).

S/MIME e-mail extensions can with special toolkits be added to existing network software and be embedded in network software being developed. Freeware S/MIME v3 toolkits are (S/MIME Freeware Library, 2006) and the Mozilla S/MIME Toolkit. Phaos S/MIME Toolkit is a Java package (Phaos S/MIME, 2004) for secure messaging in Java applications.

IPSec software development is usually VPN software development. IPSec can be integrated in the networking software and/or hardware of a router/a computer node. Commercial IPSec developer toolkits are available, for example Certicom Security Builder IPSec (2006).

Program libraries for SSH protocol integration during network software design are also available; see, for example, Ganymed SSH-2 for Java (2005).

In smartcard application development usually some development kit is used. Microsoft offers a Smartcard Toolkit to be used together with visual programming tools.

DESIGN OF SECURE NETWORK SOFTWARE

Network security software implements security features. Other network software implements functionality and other features like usability, efficiency, simplicity, safety, dependability, reliability, and so forth. Security requirements for any network software include:

- absence of vulnerabilities and security holes
- secure interfaces

Security should be integrated in the network software life cycle starting from the specifications. The need to assess vulnerability and to react on security incidents should be proactively minimized before network software is used. A recent handbook for secure software design is available (Viega & McGraw, 2002).

FUTURE TRENDS

IPSec is integrated in the new version of the IP protocol, IPv6 (IP version 6, 2006). Thus IPSec is automatically included in the IP software in all nodes in future TCP/IP networks. Also DNSSEC and secure routing protocols will be included in the system software of future TCP/IP networks.

New wireless network protocols emerging are among others Wireless USB (WUSB) (Kolic, 2004) and ZigBee (ZigBee Alliance, 2006). WUSB will offer the same functionality as standard wired USB devices. ZigBee is a low-power, short-range, wireless technology. Both technologies will be used in networking solutions for home/industrial automation.

Wi-Fi Protected Access version 2 (WPA2) includes full 802.11i support in a WLAN (Wi-Fi Protected Access, 2003). WPA2 will replace RC4 with AES. It will also include the CCM protocol. The new standard implementation is hardware accelerated and will require replacement of most access points and some NICs (Network Interface Cards).

Session key agreements in future wired network, will be based on absolutely secure quantum cryptography protocols (Bennett, 1984), which are physically implemented by transmission of randomly polarized laser pulses in optical fibers (Stucki, Gisin, Guinnard, Ribordy, & Zbinden, 2002). Absolutely secure means that verified reception of a session key is also a proof that the same key has not been eavesdropped. Commercial Quantum key distribution technology is already available (id Quantique Portal, 2006).

CONCLUSION

Software and hardware solutions and tools are network security cornerstones. Today, network security technology is a large and complex rapidly expanding area. Network security software skills are needed by every computer and computer network user. This has profound implications on all education, since use of computer networks is inevitable.

Education for professional network security software skills should include:

- installation, configuration, and test use of all categories of available network security software/hardware solutions and products,
- source code inspection exercises of open source network security software solutions, and
- programming exercises and projects with TLS/SSL application development environments and cryptographic toolkits.

Network security software development skills are important in upper level network security education.

REFERENCES

Barbir, A., Murphy, S., & Yang, Y. (2004). *Generic threats to routing protocols*. IETF. Internet-Draft. Retrieved July 1, 2006, from http://www.ietf.org/internet-drafts/draft-ietf-rpsec-routing-threats-07.txt

Bennett, C. H., & Brassard, G. (1984). Quantum cryptography: Public key distribution and coin tossing. In *Proceedings of the IEEE International Conference on Computers, Systems, and Signal Processing* (p. 175). Los Alamitos, CA: IEEE Press.

Bowles, J., & Pelaez, C. (1992). Bad code. *IEEE Spectrum, 29*(8), 36-40.

Certicom Security Builder IPSec. (2006). Retrieved July 1, 2006, from http://www.certicom.com/index.php?action=product,sbipsec

Certicom Security Builder SSL. (2006). Retrieved July 1, 2006, from http://www.certicom.com/index.php?action=product,sbssl

Chien, E. (2005). *Techniques of adware and spyware* (White Paper). Symantec Security Response. Retrieved June 22,

2006, from http://www.symantec.com/avcenter/reference/techniques.of.adware.and.spyware.pdf

Dierks, T., & Rescora, E. (2006). *The transport layer security (TLS) protocol Version 1.1*. IETF. RFC 4346. Retrieved July 1, 2006, from http://www.ietf.org/rfc/rfc4346.txt

DNSSEC: DNS Security Extensions Securing the Domain Name System. (2006). Retrieved July 1, 2006, from http://www.dnssec.org

F-Secure Policy Manager. (2006). Retrieved November 14, 2006, from http://www.f-secure.com/products/fspm/

Ganymed SSH-2 for Java. (2005). Retrieved July 1, 2006, from http://www.ganymed.ethz.ch/ssh2/

Grahn, K., Pulkkis, G., & Guillard, J-S. (2002). Security of mobile and wireless networks. In *Proceedings of the Informing Science + IT Education Conference* (pp. 587-600). Cork, Ireland. Retrieved April 2, 2004, from http://ecommerce.lebow.drexel.edu/eli/2002Proceedings/papers/Grahn-152Secur.pdf

Hoglund, G., & Butler, J. (2005). *Rootkits: Subverting the Windows kernel.* Boston: Addison-Wesley.

id Quantique Portal. (2006). Retrieved July 1, 2006, from http://www.idquantique.com/

IETF. (2006). *The Internet Engineering Task Force*. Retrieved July 1, 2006, from http://www.ietf.org

IP Version 6 Working Group (ipv6). (2006). *IETF*. Retrieved July 1, 2006, from http://www.ietf.org/html.charters/ipv6-charter.html

Internet2 Middleware Initiative (I2-MI) Portal. (2006). Retrieved July 1, 2006, from http://middleware.internet2.edu/

Kalakota, R., & Whinston, A. B. (1999). *Frontiers of electronic commerce*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Kolic, R. (2004). *An introduction to wireless USB (WUSB)*. Retrieved July 1, 2006, from http://deviceforge.com/articles/AT9015145687.html

Levine, J. G., Grizzard, J. B., & Owen, H. L. (2006). Detecting and categorizing kernel-level rootkits to aid future detection. *IEEE Security & Privacy, 4*(1), 24-32.

Moberg, D., & Drummond, R. (2005). *MIME-based secure peer-to-peer business data interchange using HTTP, applicability statement 2 (AS2)*. IETF. RFC 4130. Retrieved July 1, 2006, from http://www.ietf.org/rfc/rfc4130.txt

Nystrom, M. (2000). *The SecurID(r) SASL mechanism*. IETF. RFC 2808. Retrieved June 22, 2006, from http://www.ietf.org/rfc/rfc2808.txt

OpenSSH Portal. (2006). Retrieved July 1, 2006, from http://www.openssh.org/

Openswan Portal. (2006). Retrieved July 1, 2006, from http://www.openswan.org/

OpenVPN Portal. (2005). Retrieved July 1, 2006, from http://openvpn.net/

Oppliger, R. (2000). *Security technologies for the Word Wide Web*. Boston; London: Artech House.

PC/SC Workgroup. (2006). Retrieved July 1, 2006, from http://www.pcscworkgroup.com

Phaos S/MIME. (2004). Retrieved July 1, 2006, from http://www.phaos.com/products/smime/smime.html

Public-Key Infrastructure (X.509) Working Group (pkix). (2006). *IETF.* Retrieved July 1, 2006, from http://www.ietf.org/html.charters/pkix-charter.html

Pulkkis, G., Grahn, K., & Åström, P. (2003). Network security software. In R. Azari (Ed.), *Current security management & ethical issues of information technology* (pp. 1-41). Hershey, PA: IRM Press.

RosettaNet Implementation Framework: Core Specification. (2001). Retrieved July 1, 2006, from http://xml.coverpages.org/RNIF-Spec020000.pdf

S/MIME Freeware Library (SFL). (2006). Retrieved July 1, 2006, from http://digitalnet.com/knowledge/sfl_home.htm

Shibboleth Project Portal. (2006). Retrieved July 1, 2006, from http://shibboleth.internet2.edu/

Stallings, W. (2000). *Network security essentials*. Upper Saddle River, NJ: Prentice-Hall.

Stallings, W. (2006). *Cryptography and network security* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

SSH Tectia Solution. (2006). Retrieved July 1, 2006, from http://www.ssh.com/products/tectia/

Stephenson, P. (1993). Preventive medicine. *LAN Magazine, 8*(11).

Stucki, D., Gisin, N., Guinnard, O., Ribordy, G., & Zbinden, H. (2002). Quantum key distribution over 67 km with a plug&play system. *New Journal of Physics, 4*(41), 1-8.

The OpenSSL Project. (2005). Retrieved July 1, 2006, from http://www.openssl.org

UN/EDIFACT Portal. (2006). Retrieved July 1, 2006, from http://www.unece.org/trade/untdid/welcome.htm

Vasco Product Range. (2006). Retrieved June 22, 2006, from http://www.vasco.com/products/range.html

Viega, J., & McGraw, G. (2002). *Building secure software*. Boston: Addison Wesley.

Whiteley, D. (2000). *E-commerce: Strategy, technologies and applications*. Cambridge: McGraw-Hill International (UK) Ltd.

Wi-Fi Protected Access. (2003). White Paper. Retrieved July 1, 2006, from http://www.wi-fi.org/white_papers/whitepaper-042903-wpa/

WiMAX Forum. (2006). Retrieved July 1, 2006, from http://www.wimaxforum.org/

ZigBee™ Alliance. (2006). Retrieved July 1, 2006, from http://www.zigbee.org

KEY TERMS

E-Mail Protocols: Simple mail transport protocol (SMTP) is a set of commands for transport of ASCII encoded e-mail messages. Post office protocol (POP3) retrieves new messages from a mailbox to a remote e-mail client. A remote e-mail client can simultaneously access several mailboxes on different mail servers with the Internet message access protocol (IMAP).

Internet Engineering Task Force (IETF): An open international community engaged in Internet architecture evolution (IETF, 2006). Working Groups in several topical areas develop technical drafts and Internet standards.

Internet Protocol Security (IPSec): The IPSec protocol suite is developed by an IETF Security Area Working Group. IPSec introduces a new TCP/IP protocol stack layer below IP. IPSec adds authentication and optionally encryption to transmitted data packets. Authentication ensures that packets are from the right sender and have not been altered. Encryption prevents unauthorized reading of packet contents.

Pretty Good Privacy (PGP): An e-mail extension used to encrypt/decrypt and cryptographically sign e-mail, as well as to verify e-mail signatures. Verification of a signature is a proof of sender identity and message authenticity.

Secure Multipurpose Internet Mail Extensions (S/MIME): A secure e-mail standard based on MIME. S/MIME, being further developed by an IETF Security Area Working Group, accomplishes privacy and authentication by using encryption/decryption, digital signatures, and X.509 certificates.

Simple Network Management Protocol (SNMP): An application layer TCP/IP protocol for management information exchange between network devices. SNMP includes two main software entity types: managers and agents.

Virus: Malicious code added to an executable file loaded to a computer and executed without the user's knowledge and consent. Computer viruses often copy and spread themselves to other computers in the same network.

Current Practices in Electroencephalogram-Based Brain-Computer Interfaces

Ramaswamy Palaniappan
University of Essex, UK

Chanan S. Syan
University of the West Indies, West Indies

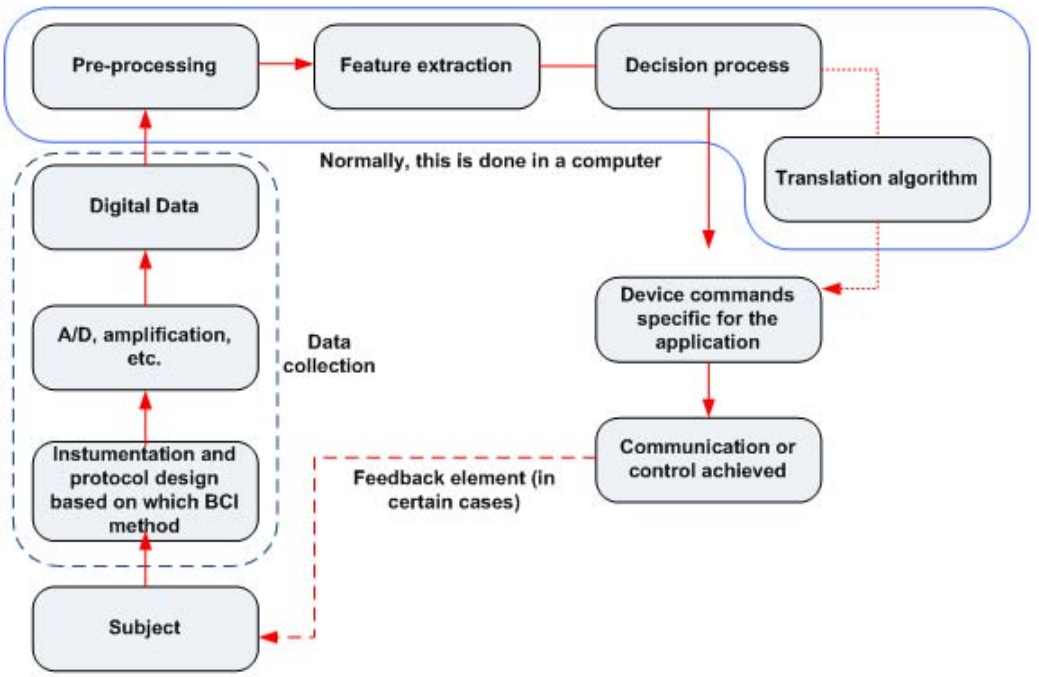
Raveendran Paramesran
University of Malaya, Malaysia

INTRODUCTION

Electroencephalogram (EEG) is the electrical activity of the brain recorded by electrodes placed on the scalp. EEG signals are generally investigated for the diagnosis of mental conditions such as epilepsy, memory impairments, and sleep disorders. In recent years there has been another application using EEG: for brain-computer interface (BCI) designs (Vaughan & Wolpaw, 2006).

EEG-based BCI designs are very useful for hands-off device control and communication as they use the electrical activity of the brain to interface with the external environment, therefore circumventing the use of peripheral muscles and limbs. Some current applications of BCIs in communication systems are for paralyzed individuals to communicate with their surroundings through character/menu selection and in device control such as wheelchair movement, prosthetics control, and flight and rehabilitative (assistive) technologies. For the general public, some of the possible applications are hands-off menu selection, flight/space control, and virtual reality (entertainment). BCI has also been applied in biometrics (Palaniappan & Mandic, 2007).

Figure 1. Main elements of general BCI system



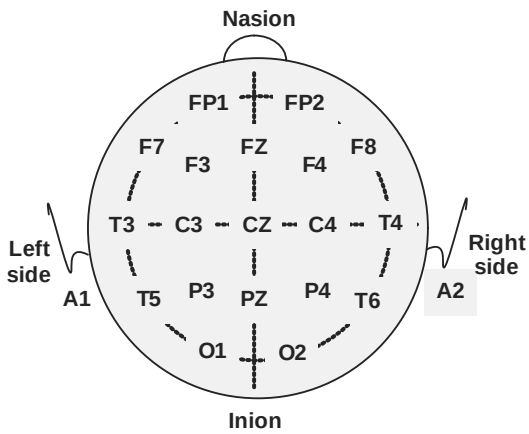
This research area is extremely exciting, and in recent times, there has been an explosive growth of interest in this revolutionary new area of science which would enable computers (and therefore any other reactive device) to be controlled by thought alone—the benefits for the severely disabled would be truly astonishing. For example, in 1990, there were less than 10 groups (mostly in the U.S.) with research interests in BCI; but this has grew to more than 130 groups worldwide in 2004 (Vaughan & Wolpaw, 2006). It is a multidisciplinary field comprising areas such as computer and information sciences, engineering (electrical, mechanical, and biomedical), neuroscience, and psychology. State-of-the-art BCI designs are still very primitive, but because of their potential to assist the disabled, there is an increasing amount of investment in their development.

This article will give an overview of the general elements in a BCI system and existing BCI methodologies, state the current applications of BCI devices in communication system and device control, and describe the current challenges and future trends in BCI technology.

BACKGROUND

In general, a BCI system comprises five stages: data collection, pre-processing, feature extraction, decision making (which includes translation algorithm[1]), and device command. Normally, the pre-processing, feature extraction, and decision-making stages are done using a computer, though a dedicated hardware could be designed for this purpose. Sometimes, these five stages can be simplified to just three: sensor, decoder, and actuator (Hochberg & Donoghue, 2006).

Figure 2. 10-20 electrode placement system



Data Collection Through Electrodes

Subjects will generate brain activity through an experimental paradigm that would depend on the particular BCI approach. The protocol to be followed by the subjects could be thinking about making imaginary movements, focusing on flashing characters on a screen, and so forth. This brain activity will be picked by electrodes (normally Ag/AgCl) placed on the scalp. The placement of electrodes commonly follows the 10-20 system (19 electrodes) or extensions of this system (32, 64, 128, or 256 electrodes). The recordings are normally referenced to the left and/or right mastoids. An example of the 10-20 electrode placement system is shown in Figure 2.

As the recorded signals are in the range of microVolts, amplifiers will be needed to amplify the multi-channel signals. These signals will then be sampled at a suitable frequency (a typical sampling frequency is 256 Hz) using an analogue-to-digital conversion device (nowadays with precision of 16-24 bits per channel). Currently, there are electrodes available that do the first-stage amplification in the electrode itself (which minimizes preparation time). In general, a single portable EEG signal acquisition unit is capable of amplification, sampling, and data transfer to the computer. Figure 3 shows an example of a subject using a BCI device.

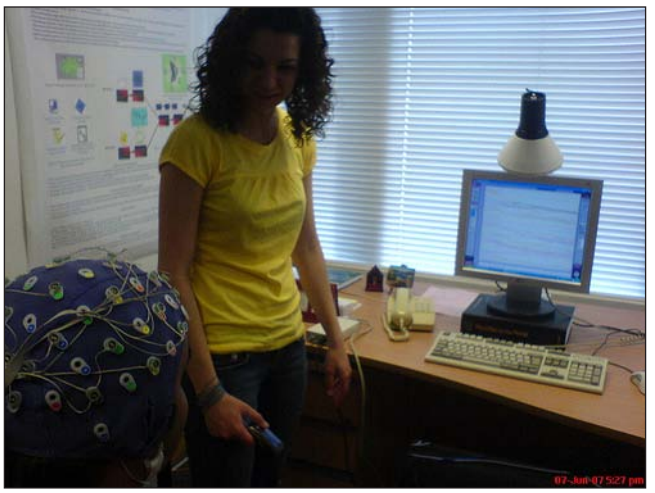
Pre-Processing

These digital EEG data normally contain a lot of noise (artifacts). Some examples of noise sources are 50/60 Hz power line interference, fluorescent lighting, baseline drift (low frequency noise), electrocardiogram (ECG), electromyogram (EMG), and random noise. Simple frequency-specific filtering is normally sufficient to reduce the narrow band noises such as the power line interference, baseline drift, and fluorescent lighting. However, more sophisticated methods such as principal component analysis (PCA) and independent component analysis (ICA) are popular to reduce ECG and EMG noises that have overlapping spectral information with EEG. Another common artifact that corrupts EEG signals is eye blinks; many techniques have been proposed to solve this problem (Thulasidas et al., 2004).

Feature Extraction

Though the raw EEG signal could be used by the next decision-making stage, very often features are extracted from these EEG signals. Depending on the EEG approach used in the BCI, the feature extraction approach would vary. For example, for the mental-activity-based BCI, autoregressive (AR) features have been used (Anderson, Stolz, & Shamsunder, 1998), where Burg's method (Shiavi, 1999) is the common procedure used to estimate the AR coefficients with

Figure 3. Example of a subject using a BCI device



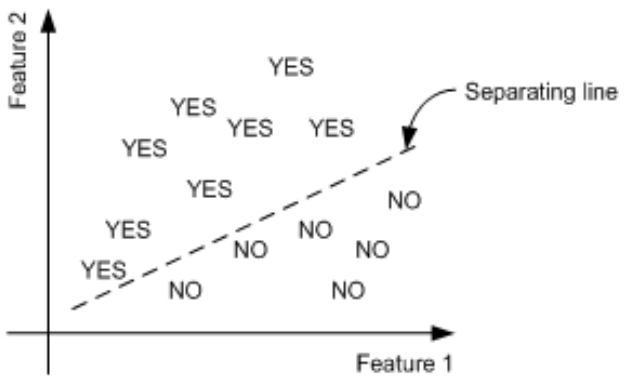
the model order chosen by Akaike Information Criterion (AIC) (Akaike, 1974). When there are inter-hemispheric differences, asymmetry ratio (Keirn & Aunon, 1990) provides to be a good feature extraction method. Some of the other common feature extraction methods are spectral analyses, voltage amplitude measurements, spatial filtering, and single neuron-separation (Wolpaw, Birbaumer, McFarland, Pfurtscheller, & Vaughan, 2002). The features could be in the time domain, such as the P300-evoked potential amplitude (Farwell & Donchin, 1988), or they could be in the frequency domain, using classical or modern spectral analyses—for example, mu or beta rhythm amplitudes (Pfurtscheller & Da Silva, 1999). Joint time-frequency features could also be used (Schalk, Wolpaw, McFarland, & Pfurtscheller, 2000).

Decision Making

This stage normally classifies the features from the previous stage into different categories based on the required device output. For example, for the P300 speller matrix paradigm (Donchin, Spencer, & Wijesinghe, 2000), this stage would classify the features into one of the categories representing the alphanumeric characters. There are several popular classification methodologies that have been explored in BCI: linear classifiers, non-linear classifiers, and Bayesian classifiers.

The most popular linear classifier in BCI design is the linear discriminant analysis (LDA), sometimes known as Fisher's LDA. This classifier uses hyperplanes to separate the

Figure 4. Simple linear separation of two classes: YES and NO



features into different classes (Duda, Hart, & Stork, 2001). For a two-class problem with two features, the boundary is simply a straight line (as shown in Figure 4). It should be noted that a simple binary switch ('YES', 'NO') such as the one shown is capable of generating more complex responses through, say, nested menu icon selection or by using some translation algorithm (such as Morse code).

For classifying several classes, the general strategy is multi-levels of 'one vs. the rest' classification, though one level of multi-class classification is also possible with the generation of several hyperplanes.

LDA is simple to use and in general gives acceptable levels of performance (Lotte, Congedo, Lecuyer, Lamarche, & Arnaldi, 2007), though EEG data is generally non-linear, but its low complexity makes it particularly suitable for online BCI systems. It has been used successfully in motor imagery-based BCI (Pfurtscheller & da Silva, 1999; Tsui, Vuckovic, Palaniappan, Sepulveda, & Gan, 2006), P300 speller matrix (Donchin et al., 2000), mental activity BCI (Huan & Palaniappan, 2004), and asynchronous BCI (Leeb et al., 2007).

The artificial neural network (ANN), typically the multilayer perceptron (MLP) architecture, is one of the most common non-linear classifiers employed in BCI designs (Garret et al., 2003). With enough neurons in the single hidden layer, MLP could approximate any continuous function, thereby being suitable for BCI designs. Typically, the back propagation and its modern variants have been used to train the MLP ANN (Palaniappan, 2004). Some of the other ANNs that have been used in BCI studies are Fuzzy ARTMAP (Palaniappan, Raveendran, Nishida, & Saiwaki, 2002) and learning vector quantization (Pfurtscheller, Flotzinger, & Kalcher, 1993).

A support vector machine (SVM) could be either a linear or non-linear classifier depending on which kernel function is used. SVMs have also been used successfully in several BCI designs: linear SVMs use linear decision boundaries (Rakotomamonjy, Guigue, Mallet, & Alvarado, 2005), while the non-linear ones use Gaussian or radial basis function kernels (Garret et al., 2003).

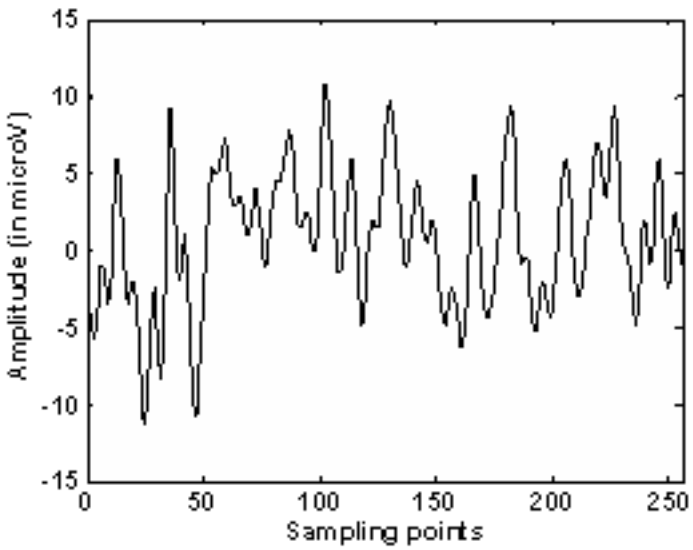
The nearest neighbor classifier is one of the simplest classifiers, but due to its computational complexity (as the distance of every test data from every training data must be computed), it has not seen much use in BCI except in a few studies (Palaniappan & Danilo, 2007; Ravi & Palaniappan, 2005).

Similarly, Bayesian classifiers are also not popular, though they were used successfully in motor imagery (Lemm, Schafer, & Curio, 2004) and mental activity BCI (Keirn & Aunon, 1990). Bayesian classifiers (typically the naïve version that assumes independence of features) use the prior class probability and conditional probability of the training data to estimate the maximum posterior hypothesis (i.e. class) of the test data.

Translation Algorithms

The translation algorithm translates the output of the classifier into meaningful information or command controls. For example, a sequence of mental activity denotes a particular command to move a wheelchair or some code to translate a sequence of imagined movements into English alphabets.

Figure 5. Example of a recorded EEG signal



However, this element is not present nor required for all BCI designs.

Sometimes, there is feedback from the device output to the subject. This will allow the subject to enhance its EEG output to increase the accuracy; this sort of feedback is common during the design of some BCIs like those using slow cortical potential (SCP).

OVERVIEW OF NON-INVASIVE BCI METHODOLOGIES

Basically, the BCI approaches could be either invasive or non-invasive. The non-invasive BCI methods using EEG, magnetoencephalogram (MEG), positron emission topography (PET), functional magnetic resonance imaging (fMRI), and optimal imaging (near-infrared spectroscopy (NIRS)) are more popular than the invasive one based on electrocorticogram (ECoG), though effective due to health hazards posed by the latter (as the electrodes are surgically implanted).

EEG-Based BCI

EEG is the brain's electrical activity recorded using electrodes attached to the scalp; it is the cumulative effect of thousands or more neurons (in the cortex) that are activated during mental processes. It is in the microVolts range due to high attenuation by skull and scalp (so for proper analysis, they have to be amplified). Figure 5 shows an example of a recorded EEG signal.

There are several methodologies for implementing EEG-based BCI: evoked potentials (typically from visual stimulus, though not always), better known as visual evoked potential (VEP) (Donchin et al., 2000; Wang, Wang, Gao, Hong, & Gao, 2006), mental activity (Palaniappan, 2006b), motor imagery (Wolpaw et al., 2002), and SCP (Mensh, Werfel, & Seung, 2002). The VEP approach could be further divided into P300-based VEP (Donchin et al., 2000) and steady state VEP (SSVEP) (Wang et al., 2006).

P300-Based VEP

VEP is a component in EEG that is evoked in response to an external visual stimulus like visualizing a picture or flash of light. The recorded signal consists of spontaneous EEG and VEP, where the spontaneous EEG is many times higher in amplitude as compared to VEP. Hence, measures like averaging from many trials are needed to obtain a reliable enough VEP. In recent years, principal component analysis (Palaniappan & Ravi, 2006) and independent component analysis (Palaniappan, 2006a) have been suggested for separating VEP from EEG in single trials.

P300 (or P3) is the third positive component in VEP (see Figure 6), and it is maximal in midline (like locations Cz, Pz, Fz, etc). It is evoked in a variety of decision-making tasks and, in particular, when a target stimulus is recognized. It is evoked around 300 ms after stimulus onset, and in general, P300 components encountered for BCI purposes are limited to 8 Hz.

Typically, the oddball paradigm is used to evoke P300 (Polich, 1991). In this paradigm, a target stimulus that oc-

Figure 6. Commonly encountered VEP components

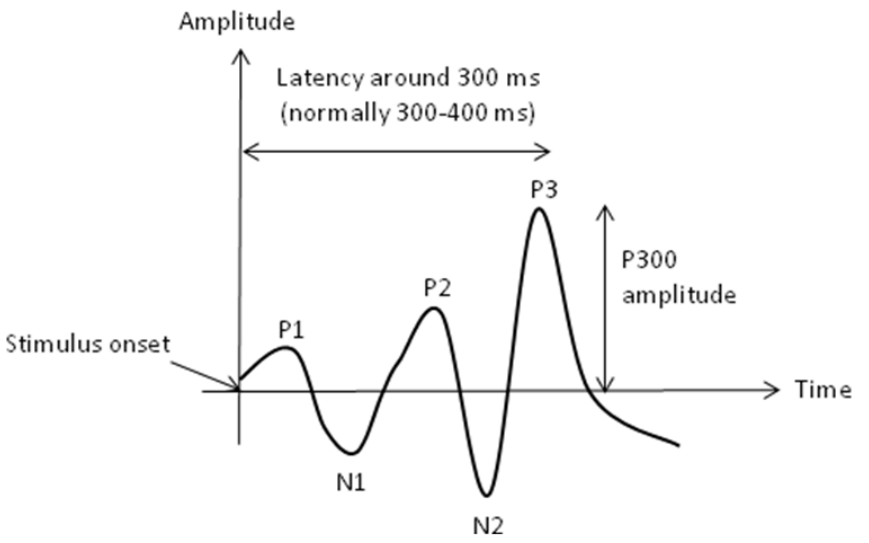
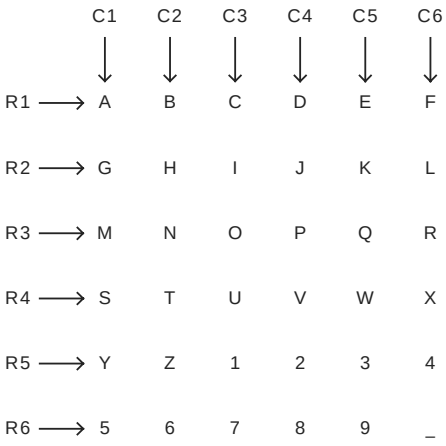


Figure 7. Onscreen display for speller matrix paradigm (Donchin et al., 2000)



curs infrequently compared to a non-target stimulus will evoke P300. In P300-VEP BCI designs (Donchin et al., 2000), a variation of this paradigm is used: spelling matrix or Donchin paradigm. In this paradigm, the screen consists of alphanumeric characters. Spontaneous EEG plus VEP is recorded when rows and columns flash, where each trial consists of 12 flashes and trials are repeated, typically 15, 20, or 40 times. Averaging from trials is performed to reduce unrelated spontaneous EEG from VEP. Normally, a low-pass filter (LPF) with cut-off at 8 Hz is used, and P300 peak detected around 300-400 ms (sometimes, other ranges like 300-500 or 400-600 ms are also used) and amplitude of P300 peak stored for analysis. The row and column containing the target (focused) character will have a higher P300 amplitude compared to the row or column that does not contain the target character. Figure 7 shows the onscreen display in this paradigm, while Figure 8 shows an example of real averaged VEP signals from 12 flashes obtained from a subject, where the focused target character is N (row 3, column 2). The abscissa is the amplitude of the VEP (in microVolts) and ordinate is the time (in seconds) after stimulus onset.

Mental-Activity-Based BCI

In this BCI design (Keirn & Aunon, 1990; Palaniappan et al., 2002; Palaniappan, 2006b), EEG signals are recorded when users think of different mental tasks covering a wide range of cognitive abilities (without any vocalizing or physical movements). Some examples of used mental tasks include:

- *Baseline*—Subjects relax and think of nothing in particular.
- *Letter Composing*—Subjects mentally compose a letter to someone.
- *Math*—Subjects do nontrivial multiplication problems, such as *42 x 18*.
- *Visual Counting*—Subjects visually imagined sequential numbers being written on a blackboard with the previous number being erased before the next number is written.
- *Geometric Figure Rotation*—Subjects imagine a figure being rotated around an axis.

These four active mental tasks (i.e., excluding baseline tasks) exhibit inter-hemispheric differences. For example: math tasks utilize more processes in the left hemisphere as compared to visual tasks that utilize more processes in the right hemisphere.

One useful measure to detect the activated hemisphere is through asymmetry ratio (AS) (Keirn & Aunon, 1990):

Figure 8. An example of real averaged VEP signals from 12 flashes (the focused target character is from row 3, column 2: N)

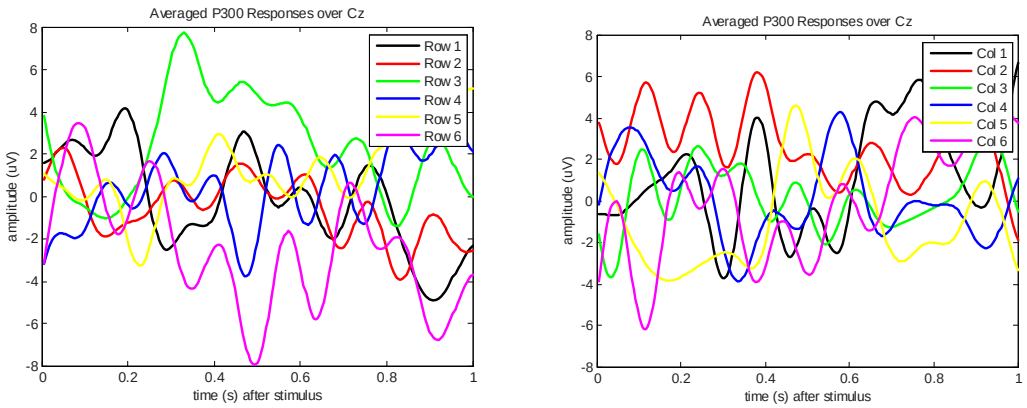
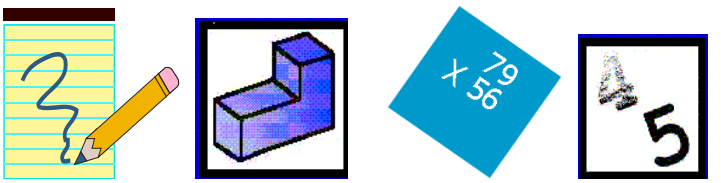


Figure 9. Some example of mental tasks

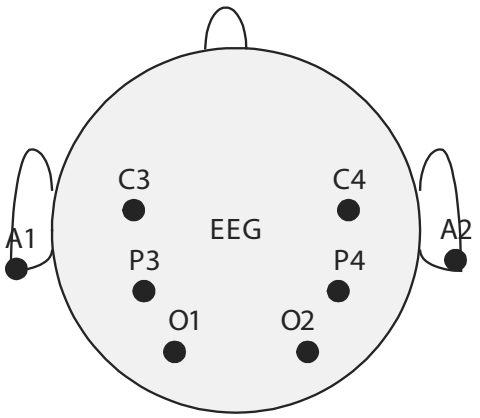


$$AS = \left[\frac{E_1 - E_2}{E_1 + E_2} \right], \quad (1)$$

where E_1 is the energy in one EEG channel in the left hemisphere and E_2 is the energy of EEG in another channel but in the right hemisphere. For example, for the electrodes shown in Figure 10, AS using channels (electrodes) O1 and O2 will be positive for maths activity as compared to object rotation (visual) activity. AS for baseline tasks will be near zero. So we can use baseline (B) and two tasks—maths (M) and object rotation (O)—to construct a communication method using a translation algorithm.

Here, translation algorithm translates the sequence of detected mental tasks into a command/output. For example, for wheelchair movement: task sequence MBO could denote 'turn left', while task sequence OBM could denote 'turn right'. For communication purposes, Morse code could be used to translate the sequence of mental tasks into alphabets. For example, letter I in Morse code is 'dot dot', so the sequence of mental tasks: OBOB or (MBMB) could denote letter I, where the baseline denotes the end and start of a new mental task (except for the first mental task). Figures 11 and 12 illustrate the use of this translation algorithm.

Figure 10. Electrode locations to illustrate AS



SSVEP BCI

SSVEP is the response due to a visual stimulus modulated at a frequency higher than 6 Hz. It is maximum in the visual cortex, specifically in the occipital region. In this design, each onscreen target flickers with a specific frequency, and a photic driving response in the brain causes the frequency (and its harmonics) to appear in SSVEP. In other words, the onscreen "flickers" drive a photic response (with corresponding frequencies) in the optical regions of the brain. So, determination of SSVEP frequency (through spectral analysis like Fourier methods) is enough to decide on the focused target. An example of the onscreen display for dialing telephone numbers is shown in Figure 13.

The flicker frequency selection can be up to 45 Hz; in general the higher frequency is better as this causes less strain on the eyes of the user (Wang et al., 2006). For example, for the onscreen telephone keypad as in Figure 13, the flicker frequencies could be in 1 Hz intervals: 1 (9 Hz), 2 (10 Hz), 3 (11 Hz), ……0 (16 Hz), # (17 Hz). In the work by Wang et al (2006), normalized fast Fourier transform (FFT) magnitude is computed from four seconds of SSVEP. Peak detection (above a certain threshold) is used to determine the frequency. Once detected, the number appears onscreen and the subject moves on to look at another number.

Generally, harmonics should be avoided (e.g., if 9 Hz is used for a block, integer multiples of this such as 18, 27, 36, or 45 Hz should not be used). However, recently Muller-Putz, Scherer, Brauneis, and Pfurtscheller (2005) showed that using harmonics could improve the detection of the focused target.

Motor Imagery BCI

In this approach, BCI is designed using imaginary movements, that is, when the subject imagines moving a limb (could be arm, leg, tongue, etc). An actual voluntary movement is composed of three phases: planning, execution, and recovery. But it is known that even during imaginary movement (motor imagery), there is the planning stage.

During planning, event-related desynchronization (ERD) and event-related synchronization (ERS) occur in μmeter[2] (alpha, 10-12 Hz) and beta (12-14 Hz) frequency ranges. ERD is the attenuation in EEG in primary and secondary

Figure 11. The requirement of space in addition to dot to construct the letter I (Palaniappan et al., 2002)

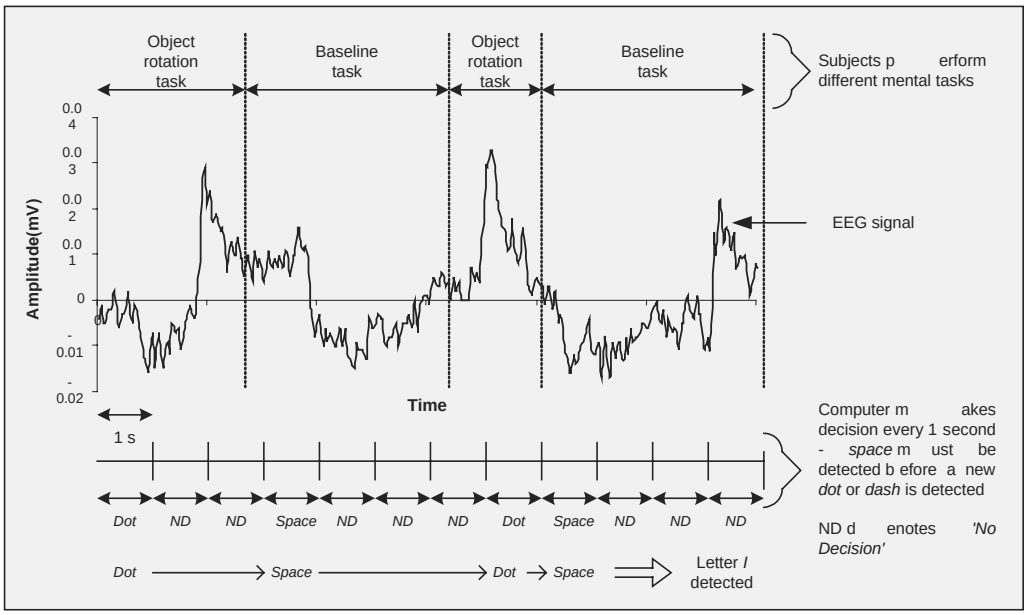
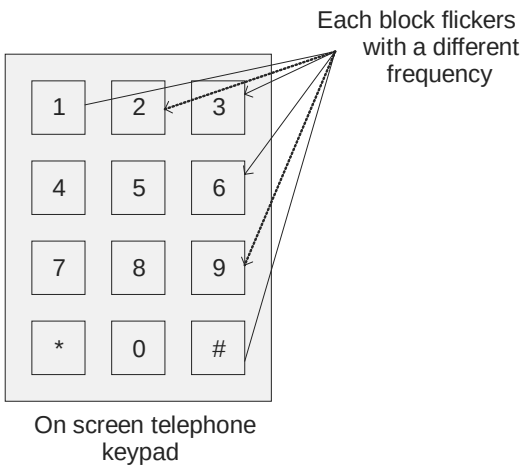


Figure 12. Schematic examples of the word 'EAT' constructed using tri-state Morse code scheme (Palaniappan et al., 2002)

Letter	Morse Code	Corresponding Mental Tasks
E	●	Letter, baseline
A	● —	Letter, baseline, count, baseline
T	—	Count, baseline

Figure 13. Onscreen display using SSVEP BCI design to dial telephone numbers



motor cortices during the preparatory stage which peaks at movement onset in contralateral hemisphere (e.g., left-hand motor imagery, ERD in the right side of the brain), while ERS is the corresponding EEG amplification but in an ipsilateral hemisphere. The ERD/ERS can be computed using power values by squaring the samples in the frequency ranges or using some power spectral density (PSD) measure (Pfurtscheller & Da Silva, 1999). In addition to these changes in alpha and beta frequency ranges, there is also the gamma burst, which is a sharp increase in EEG in the gamma (36-40 Hz) frequency range. Figure 15 shows an example of these ERD/ERS and gamma bursts.

One simple method to implement this paradigm is shown in Figure 16. Here, the subject imagines moving his or her left or right hand. Discrimination of these imagined movements can be used for a BCI design. This could be done by computing the PSD of EEG in C3, C4, and CZ. Next, the sum of PSD (SPSD) is computed. If $C3_{SPSD}$-CZ_{SPSD} > $C4S_{SPSD}$-CZ_{SPSD}, then it is a left-hand motor imagery; if $C4_{SPSD}$-CZ_{SPSD}

Figure 14. Normalized amplitude spectra corresponding to different data lengths. The 'number' focused here is '#', as there is a clearly defined peak at 17 Hz (Wang et al., 2006).

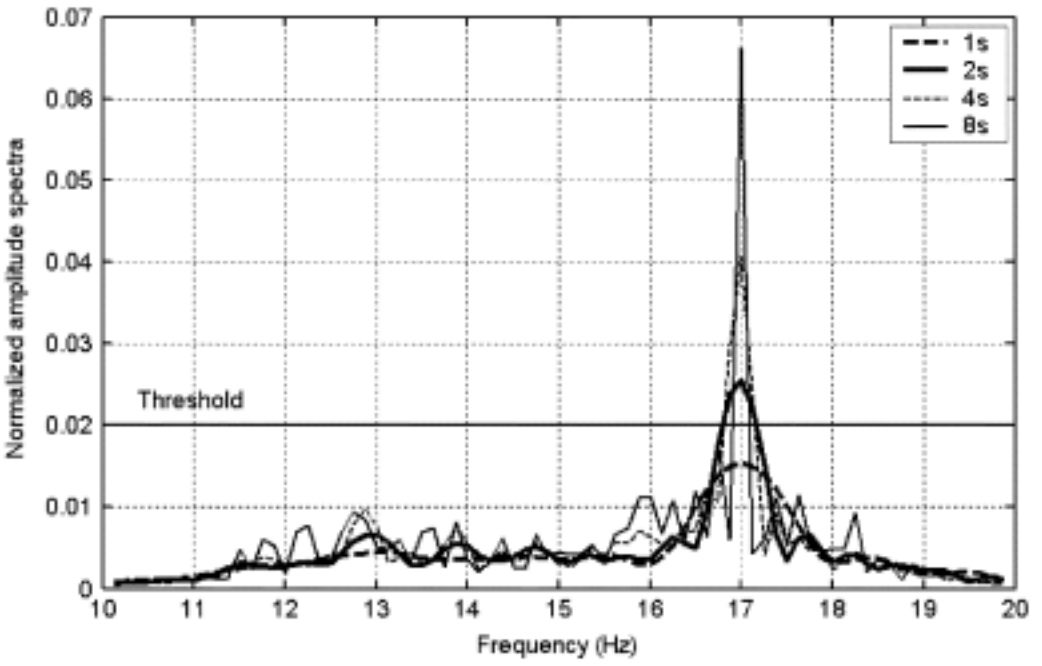
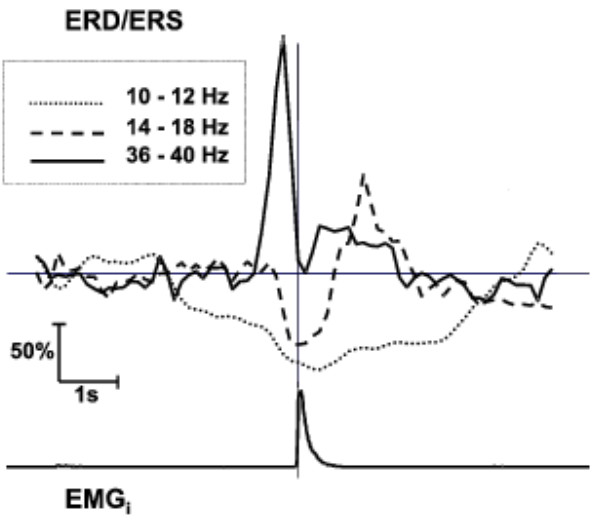


Figure 15. ERD/ERS and gamma bursts (Durka, Ircha, Neuper, & Pfurtscheller, 2001)



> $C3_{SPSD}$-CZ_{SPSD}, it is a right-hand motor imagery, and if $C4_{SPSD}$-$CZ_{SPSD} \approx C3_{SPSD}$-$CZ_{SPSD}$, then it denotes that no motor imagery has been detected.

SCP BCI

SCP is the potential shift in EEG (around 1-2 Hz) which can last several seconds. Humans can control SCP using feedback and a positive reinforcement mechanism, where the negativity SCP can be generated with tasks such as readiness to move or mobilization of resources for cognitive tasks, and the positivity SCP can be generated during execution of cognitive tasks or simply in inactive states. Self-regulation of SCP can be used to generate a binary signal or even menu/letter selection on screen for use in BCI designs. But the main problem with this approach is that it requires extensive training, typically a few months (Hinterberger et al., 2004).

Other Non-Invasive BCIs

Though the focus of this article is not on non-EEG-based BCIs, a short discussion on other current non-invasive BCIs would be useful and is given here.

PET-based BCI is where a tracer medium (radioactive isotope with short half life) is injected in blood. This tracer

Figure 16. ERD/ERS from left/right-hand motor imagery

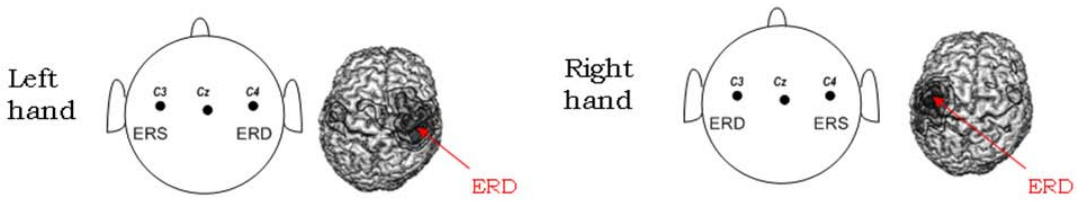
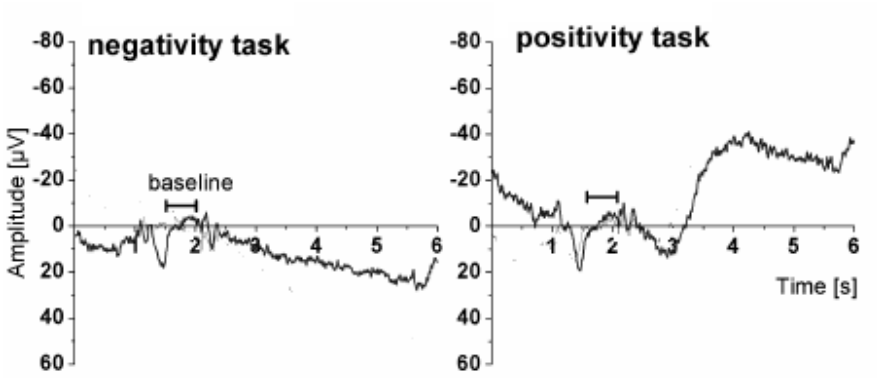


Figure 17. Negativity and positivity from SCP EEGs (Hinterberger et al., 2004)



medium moves in blood and decays by emitting a positron, and gamma photons are generated when the positron annihilates with an electron. A special scanner measures these gamma photons. So, changes in blood flow due to specific brain activity could be detected. For example, in a subject thinking of moving his or her right hand, brain activity will be detected in the left hemisphere, and vice versa. This detected brain activity could be translated for BCI applications. It is not popular as the equipment is too expensive and bulky where a cyclotron is needed for generation of tracer medium. It is also partially invasive (as it requires radioactive injection), though no surgery is involved. Furthermore, the waiting period is typically an hour before the system can be used and not suitable for continuous usage.

fMRI-based BCI detects oxygen-level changes in blood hemoglobin which generates magnetic resonance. The resulting image intensity variations show the brain activation areas, and this could be translated for BCI application. This is also not popular as it requires a bulky scanner (hence no mobility for subjects) and the vascular response is too slow—local response to this oxygen utilization occurs after a delay of approximately 1-5 seconds, with peaks at 4-5 seconds; in addition there is the possible detrimental effects due to prolonged exposure.

NIR-based BCI is a relatively new method that uses the near infra-red region of the electromagnetic spectrum (from about 800 nm to 2500 nm). The penetration is deeper into the skull as it uses NIR; NIR light is emitted and the reflection from the blood cells (specifically, oxygen level in hemoglobin) is used as a measure of blood flow. This blood flow denotes activation of brain areas (activation deduced from images with higher reflectance intensity). This procedure is sometimes known as an optical method, as NIR produces intensity images. Mobility is not limited (unlike fMRI) as the sensor-detector can be fixed on the skull, but its main hindrance is the slow vascular response (on the order of a few seconds). Also, though most NIR energy is reflected, a small portion of NIR energy might be absorbed by the brain cells and potentially be damaging in the long run.

MEG-based BCI is similar to EEG but uses measurements of magnetic fields rather than electric fields. The magnetic field generated by the brain is measured outside the scalp. The temporal resolution is better than EEG, but not popular due to high cost and difficulty in obtaining proper MEG readings as ultra-sensitive magnetic field detectors are needed. Mobility is also restricted, and shielding from other magnetic sources is also necessary which inhibits usage outside the lab environment.

BCI APPLICATIONS

One of the most important BCI applications is restoring control functions to those with motor impairments caused by progressive disorders such as amyotrophic lateral disorder (ALS), muscular dystrophy, and multiple sclerosis, and non-progressive disorders such as brainstem stroke, traumatic brain injury, spinal cord injury, and numerous orders that impair the neural pathways preventing proper muscle control. But BCI devices can also be developed for everyday use by the healthy population, though this is not its main use.

Individuals with Motor Disabilities

Severely affected individuals may lose all forms of voluntary muscle control. However, these individuals are able to survive with modern life technology support, but are completely 'locked-in' without any ability to communicate at all. The use of EEG-based BCI technologies can offer these individuals an alternative mode of communication. A typical example of a communication BCI system would be brain-controlled word processing software.

These technologies could also aid the disabled in restoring mobility (controlling wheelchair movement); environmental control (controlling TV, power beds, thermostats, etc); prosthetics control (motor control replacement, controlling artificial limbs); and rehabilitative (assistive) control—to restore motor control (strengthen or improve weak muscles).

Other Applications

The general population might also benefit from BCI devices. A simple example is environmental control (hands-off control), for example, control of external devices without using the external limbs (hand/legs). Other examples are in areas such as virtual reality (entertainment), for example, computer games like Mind Pacman (Krepki, Blankertz, Curio, & Müller, 2007). Recently, BCI has also been studied for use by astronauts in space, for example when they become temporarily paralyzed or to control devices in severely restricted extra terrestrial environments (Menon et al., 2007).

Biometrics is yet another recent application of BCI where BCI devices are used to identify an individual (Palaniappan, 2004) or authenticate an individual using thoughts alone (Palaniappan & Danilo, 2007). It is very useful for high-security individual identification applications, as EEG-based biometrics is more fraud resistant compared to other biometrics like fingerprints, iris scans, retina, ear shape, odor, and so forth (since thoughts cannot be forged!). EEG-based biometric verification is also more secure than personal identification numbers (PINs) and passwords, which can be easily compromised through shoulder surfing or other means. For example, the BCI device used to spell alphabets (Donchin et al., 2000) can actually be adapted to generate passwords that are based on thoughts alone.

CURRENT CHALLENGES

There are numerous current challenges for BCI designs. The following descriptions are not exhaustive but serve to illustrate some of the issues facing this research field that is still very much in its infancy.

It is not clear how the BCI designs will require adaptation for real users (like disabled people) and in real noisy environments. Most of the BCI research studies are conducted in laboratory environments where the condition is ideal. For example, the BCI experiments at the University of Essex are conducted in the BCI laboratory where there is special lighting with an electromagnetic radiation shield and under noise-free conditions. However, it should be remembered that the main BCI users will be disabled individuals, who will use the devices in a not ideal environment. So, it is important to realize that the high performances obtained in the laboratory are unlikely to be repeated in the real world. Another similar issue is that a large number of BCI studies are tested using healthy subjects (typically university students), and BCI systems may not perform up to expectation when used by disabled individuals and will require special adaptations.

A BCI system cannot be 'on' all the time. So we will need to separate algorithms to turn on or off the devices. In recent years, studies that combine the required BCI output and this on/off state recognition have been explored. These systems are known as asynchronous or self-paced BCI.

The current maximum information transfer rate for BCI systems is about 25 bits per second (Wolpaw et al., 2002), and even this—'still slow' for practical performance—is normally achieved only after extensive training or fine tuning. Improved systems are needed so that they are more accurate and give faster response. It may not be possible to train a disabled individual, and so systems that do not require any prior training and ease of use will be needed as well.

In general, most of the BCI systems are fine tuned for specific subjects. This results in individual BCI and not universal BCI, which is a problem as it may not be possible for fine tuning in disabled individuals. Therefore, universal BCIs should be explored more extensively.

Since BCI research is relatively new, it is not evident on the long-term effects of these systems on the user's health and if the performance would be stable across time.

Ethics is especially an issue of using (or testing) BCI with those already paralyzed. For example, it is difficult to decide if the permission from a responsible nearest relative of a completely disabled subject is sufficient or even appropriate to conduct experiments with the disabled subject.

FUTURE TRENDS

The solutions for all the issues discussed in the previous section will need to be explored—of course, some of these are already under consideration. For example, the use of active electrodes[3] simplifies the set-up and minimizes the preparation time. Also, dry electrodes (which do not require any wet gel to achieve the necessary impedance) are being investigated, as this will allow 24/7 use by subjects (especially the disabled).

Advancement in BCI algorithms related to signal preprocessing, feature extraction, and classification will need to be explored to maximize the accuracy while achieving a quicker response. Further, other more suitable paradigms should be explored for BCI designs.

CONCLUSION

Of the non-invasive BCI methods, EEG-based ones are the most common due to low cost and portability. Furthermore, only EEG has relatively short time constants (for rapid responses) compared to others like PET, fMRI, and NIRS (Wolpaw et al., 2002).

Each EEG-based BCI approach method has its own advantages and disadvantages. For example, some methods do not require prior user training like P300-VEP, but this method requires many trials, so has a slower response. Motor imagery BCI has faster response, but performance is not satisfactory without prior user training. Methods based on SSVEP have been shown to be successful using only one active channel (Wang et al., 2006), but require users to gaze at flashing blocks, which is only practical for short time use (typically a few minutes). Mental activity BCI tasks have fast response and do not require any visual interface, but performance is not stable over time. SCP-based BCI requires extensive training, but once mastered, performance is relatively stable.

Overall, EEG-based BCI is the most practical, portable, and cheap enough approach. A multidisciplinary approach involving clinicians (neurologists, physiatrists, psychologists), engineers, computer scientists, neuroscientists, and so forth is required to successfully design complete working BCI systems, which is probably realizable in a decade or so. With adequate engagements from the experts in these areas, BCI systems could be developed to provide an important mode of communication and control for those with severe disabilities, and in addition, to those without any disability for use in special environments.

ACKNOWLEDGMENT

The support of the following organizations in different parts of the work are acknowledged**:** University Malaya (Malaysia), Multimedia University (Malaysia), Ministry of Science (IRPA Grant, Malaysia), Nanyang Technological University (Singapore), Singapore–University of Washington Alliance, Essex University (UK), and European Space Agency.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716-723.

Anderson, C.W., Stolz, E.A., & Shamsunder, S. (1998). Multivariate autoregressive models for classification of spontaneous electroencephalogram during mental tasks. *IEEE Transactions on Biomedical Engineering, 45*(3), 277-286.

Donchin, E., Spencer, K.M., & Wijesinghe, R. (2000). The mental prosthesis: Assessing the speed of a P300-based brain-computer interface. *IEEE Transactions on Rehabilitation Engineering, 8*(2), 174-179.

Duda, R.O., Hart, P.E., & Stork, D.G. (2001). *Pattern recognition* (2nd ed.). New York: Wiley Interscience.

Durka, P.T., Ircha, D., Neuper, C., & Pfurtscheller, G. **(**2001). Time-frequency microstructure of event-related EEG desynchronization and synchronization. *Medical & Biological Engineering & Computing, 39*(3), 315-321.

Farwell, L.A., & Donchin, E. (1988). Talking off the top of your head: Toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology, 70,* 510-523.

Garrett, D., Peterson, D.A., Anderson, C.W., & Thaut, M.H. (2003). Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 11,* 141-144.

Hinterberger, T., Schmidt, S., Neumann, N., Mellinger, J., Blankertz, B., Curio, G., & Birbaumer, N. (2004). Brain-computer communication and slow cortical potentials. *IEEE Transactions on Biomedical Engineering, 51*(6), 1011-1018.

Huan, N., & Palaniappan, R. (2004). Neural network classification of autoregressive features from electroencephalogram signals for brain-computer interface design. *Journal of Neural Engineering, 1,* 142-150.

Keirn, Z.A., & Aunon, J.I. (1990). A new mode of communi-

cation between man and his surroundings. *IEEE Transactions on Biomedical Engineering, 37*(12), 1209-1214.

Krepki, R., Blankertz, B., Curio, G., & Müller, K. (2007). The Berlin brain-computer interface (BBCI)—towards a new communication channel for online control in gaming applications. *Multimedia Tools and Applications, 33*(1), 73-90.

Leeb, R., Friedman, D., Müller-Putz, G.R., Scherer, R., Slater, M., & Pfurtscheller, G. (2007). Self-paced (asynchronous) BCI control of a wheelchair in virtual environments: A case study with a tetraplegic. *Computational Intelligence and Neuroscience,* doi:10.1155/2007/79642.

Lemm, S., Schafer, C., & Curio, G. (2004). BCI competition 2003—data set iii: Probabilistic modelling of sensorimotor mu rhythms for classification of imaginary hand movements. *IEEE Transactions on Biomedical Engineering, 51,* 1077-1080.

Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of Neural Engineering, 4,* R1-R13.

Muller-Putz, G.R., Scherer, R., Brauneis, C., & Pfurtscheller, G. (2005). Steady-state visual evoked potential (SSVEP)-based communication: Impact of harmonic frequency components. *Journal of Neural Engineering, 2,* 123-130.

Menon, C., Negueruela, C., Millán, J.R., Tonet, O., Carpi, F., Broschart, M., Ferrez, P., Buttfield, A., Dario, P., Citi, L., Cecilia, L., Tombini, M., Sepulveda, F., Poli, R., Palaniappan, R., Tecchio, F., Rossini, P.M., & Rossi, D. (2006). Prospective on brain-machine interfaces for space system control. *Proceedings of the 57th International Astronautical Congress,* Valencia, Spain.

Mensh, B.D., Werfel, J., & Seung, H.S. (2004). BCI competition 2003—data set Ia: Combining gamma-band power with slow cortical potentials to improve single-trial classification of electroencephalographic signals. *IEEE Transactions on Biomedical Engineering, 51*(6), 1052-1056.

Palaniappan, R. (2004). Method of identifying individuals using VEP signals and neural network. *IEE Proceedings–Science, Measurement and Technology, 151*(1), 16-20.

Palaniappan, R. (2006a). Single trial visual event related potential extraction by negentropy maximisation of independent components. *WSEAS Transactions on Signal Processing, 2*(4), 512-517.

Palaniappan, R. (2006b). Utilizing gamma band spectral power to improve mental task based brain computer interface design. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14*(3), 299-303.

Palaniappan, R., & Mandic, D.P. (2007). Biometric from the brain electrical activity: A machine learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(4), 738-742.

Palaniappan, R., & Ravi, K.V.R. (2006). Improving visual evoked potential feature classification for person recognition using PCA and normalization. *Pattern Recognition Letters, 27*(7), 726-733.

Palaniappan, R., Raveendran, P., Nishida, S., & Saiwaki, N. (2002), A new brain-computer interface design using fuzzy ARTMAP. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 10*(3), 140-148.

Pfurtscheller, G., & da Silva, F.H.L. (1999). Event-related EEG/MEG synchronization and desynchronization: Basic principles. *Clinical Neurophysiology, 110*(11), 1842-1857.

Pfurtscheller, G., Flotzinger, D., & Kalcher, J. (1993). Brain-computer interface—a new communication device for handicapped persons. *Journal of Microcomputer Applications, 16,* 293-299.

Polich, J. (1991). P300 in clinical applications: Meaning, method and measurement. *American Journal of EEG Technology, 31,* 201-231.

Rakotomamonjy, A., Guigue, V., Mallet, G., & Alvarado, V. (2005). *Ensemble of SVMs for improving brain computer interface P300 speller performance* (pp. 45-50). Berlin: Springer-Verlag (LNCS).

Ravi, K.V.R., & Palaniappan, R. (2005). Leave-one-out authentication of persons using 40 Hz EEG oscillations. *Proceedings of the EUROCON 2005 Conference* (vol. 2, pp. 1386-1389), Belgrade, Serbia & Montenegro.

Schalk, G., Wolpaw, J.R., McFarland, D.J., & Pfurtscheller, G. (2002). EEG based communication and control: Presence of error potentials. *Clinical Neurophysiology, 1111,* 2138-2144.

Shiavi, R. (1999). *Introduction to applied statistical signal analysis* (2nd ed.). CA: Academic Press.

Thulasidas, M., Guan, C., Ranganatha, S., Wu, J.K., Zhu, X., & Xu, W. (2004). Effect of ocular artifact removal in brain computer interface accuracy. *Proceedings of the 26th Annual International Conference of the Engineering in Medicine and Biology Society* (vol. 6, pp. 4385-4388).

Tsui, C.S.L, Vuckovic, A., Palaniappan, R., Sepulveda, F., & Gan, J.Q. (2006). Narrow band spectral analysis for onset detection in asynchronous BCI. *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course* (pp. 31-31), Graz, Austria.

Vaughan, T.M., & Wolpaw, J.R. (2006). Guest editorial: Third international meeting on brain-computer interface technology. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14*(2), 126-127.

Wang, Y., Wang, R., Gao, X., Hong, B., & Gao, S. (2006). A practical VEP based brain-computer interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering, 14*(2), 234-239.

Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., & Vaughan, T.M. (2002). Brain-computer interfaces for communication and control. *Clinical Neurophysiology, 113,* 767-791.

KEY TERMS

Biometrics: Identification or authentication of the individuality of a person using the behavioral or physiological characteristics of the person.

Brain-Computer Interface/Brain-Machine Interface (BCI/BMI): Devices that use electroencephalogram signals to perform a communication or control action.

Electrodes (channels): Sensors normally made of Ag/AgCl that are used to record electroencephalogram.

Electroencephalogram (EEG): Brain activity obtained as recorded signals from the scalp using electrodes.

Mental Activity/Task: Any task that generates EEG for use in BCI designs.

Motor Imagery: Used in BCI designs where subjects imagine moving a limb.

P300 Component: The third positive component in visual evoked potential; normally evoked around 300 ms after stimulus onset.

Slow Cortical Potentials: The potential shifts in EEG (around 1-2 Hz), which can last several seconds.

Steady State VEP: A type of VEP caused by photic response (frequency following effect).

Visual Evoked Potential (VEP): An EEG component that is in response (i.e., evoked) by a visual stimulus modality.

ENDNOTES

1 This algorithm translates the decision system output into usable information or action.

2 A frequency range that is similar to alpha but occurs during actual movement/motor imagery.

3 Active electrodes have miniscule chips that are able to reduce noise caused by poor skin contact, thereby circumventing the necessity to clean the scalp prior to electrode attachment.

Customer Relationship Management and Knowledge Discovery in Database

Jounghae Bang
Kookmin University, Korea

Nikhilesh Dholakia
University of Rhode Island, USA

Lutz Hamel
University of Rhode Island, USA

Seung-Kyoon Shin
University of Rhode Island, USA

INTRODUCTION

Customer relationships are increasingly central to business success (Kotler, 1997; Reichheld & Sasser, 1990). Acquiring new customers is five to seven times costlier than retaining existing customers (Kotler, 1997). Simply by reducing customer defections by 5%, a company can improve profits by 25% to 85% (Reichheld & Sasser, 1990). Relationship marketing—getting to know customers intimately by understanding their preferences—has emerged as a key business strategy for customer retention (Dyche, 2002).

Internet and related technologies offer amazing possibilities for creating and sustaining ideal customer relationships (Goodhue, Wixom, & Watson, 2002; Ives, 1990; Moorman, Zaltman, & Deshpande, 1992). Internet is not only an important and convenient new channel for promotion, transactions, and business process coordination; it is also a source of customer data (Shaw, Subramaniam, Tan, & Welge, 2001). Huge customer data warehouses are being created using advanced database technologies (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Customer data warehouses by themselves offer no competitive advantages: insightful customer knowledge must be extracted from such data (Kim, Kim, & Lee, 2002). Valuable marketing insights about customer characteristics and their purchase patterns, however, are often hidden and untapped (Shaw et al., 2001). Data mining and knowledge discovery in databases (KDD) facilitate extraction of valuable knowledge from rapidly growing volumes of data (Mackinnon, 1999; Fayyad et al., 1996).

This article provides a brief review of customer relationship issues. The article focuses on: (1) customer relationship management (CRM) technologies, (2) KDD techniques, and (3) Key CRM-KDD linkages in terms of relationship marketing. The article concludes with the observations about the state-of-the-art and future directions.

BACKGROUND

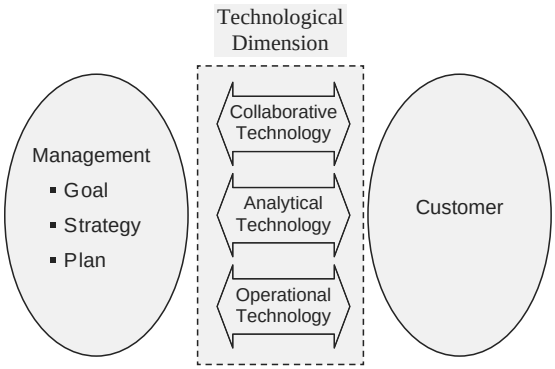
CRM Technologies

CRM is interpreted in a variety of ways (Goodhue et al., 2002; Winer, 2001; Wright, 2002). In some cases, CRM simply entails direct e-mails or database marketing. In other cases, CRM refers to CICs (customer interaction centers) and OLAP (online analytical processing), which is referred to as various types of online query-driven analyses for examining stored data. Overall, CRM can be seen as a core business strategy to interact with, create, and deliver value to targeted customers to improve customer satisfaction and customer retention at a profit. It is grounded in high quality customer data and enabled by information technology (Ang & Buttle, 2002).

Three core dimensions characterize buyer-focused CRM systems: customers, management, and technologies. *Customer* service and related issues must be included in the design, implementation, and operation of any CRM system. Organizations benefit from CRM only when such systems benefit their customers—using CRM merely as a sales or customer service solution is a recipe for failure (Davids, 1999). *Management's* articulation and tracking of customer relationship goals, plans, and metrics is an essential CRM component (Ang & Buttle, 2002; Greenberg, 2002). Successful CRM implementations rely on management goals, strategies, and plans that reflect customer commitment and promote a customer-responsive corporate culture at all levels of the organization (Ang & Buttle, 2002; Smith, 2001). *Technologies* for facilitating collaborative, operational, and analytical CRM activities are the manifest aspects of CRM (Goodhue et al., 2002).

Collaborative CRM systems refer to any CRM function that provides a point of interaction between the customer and the marketing channel (Greenberg, 2002). Web-based

Figure 1. Alignment of three dimensions of CRM



Internet, and in some cases mobile commerce systems, offer multiple "touch points" for reaching the customers. In employing the Web and mobile technologies, it is important to ensure that such technologies enhance older, preexisting channels (Johnson, 2002). *Operational CRM* systems refer to technologies that span the ordering-delivery cycle (Goodhue et al., 2002). Operational CRM is concerned with automating the customer-facing parts of the enterprise (Ang & Buttle, 2002). Since the sales process depends on the cooperation of multiple departments performing different functions, integration of all such functions is critical for operational CRM systems (Earl, 2003; Greenberg, 2002). *Analytical CRM* systems analyze customer data warehouses so that the firm can detect valuable patterns of customers' purchasing behavior. Off-line data mining of customer data warehouses as well as online analytical processing (OLAP) can aid in applications such as campaign management, churn analysis, propensity scoring, and customer profitability analysis (Goodhue et al., 2002). It is this component of CRM that has a clear linkage to KDD methods.

KDD Techniques

Since multiple data formats and distributed nature of knowledge on the Web make it a challenge to collect, discover, organize, and manage CRM-related customer data (Shaw et al., 2001), KDD methods are receiving attention in relationship marketing contexts (Fayyad et al., 1996; Mackinnon, 1999). Massive databases are commonplace, and they are ever growing, dynamic, and heterogeneous (Mackinnon & Glick, 1999). Systematic combining of data mining and knowledge management techniques can be the basis for advantageous customer relationships (Shaw et al., 2001).

KDD is defined as the process of data selection, sampling, pre-processing, cleaning, transformation, dimension reduction, analysis, visualization, and evaluation (Mackinnon, 1999). As a component of KDD (Fayyad et al., 1996), data mining can be defined as the process of searching and analyzing data in order to find latent but potentially valuable information (Shaw et al., 2001).

KDD constitutes the overall process of extracting useful knowledge from databases. It is a multidisciplinary activity with the following stages (Brachman, Khabaza, Kloesgen, Piatetsky-Shapiro, & Simoudis, 1996; Bruha, Kralik, & Berka, 2000; Fayyad et al., 1996):

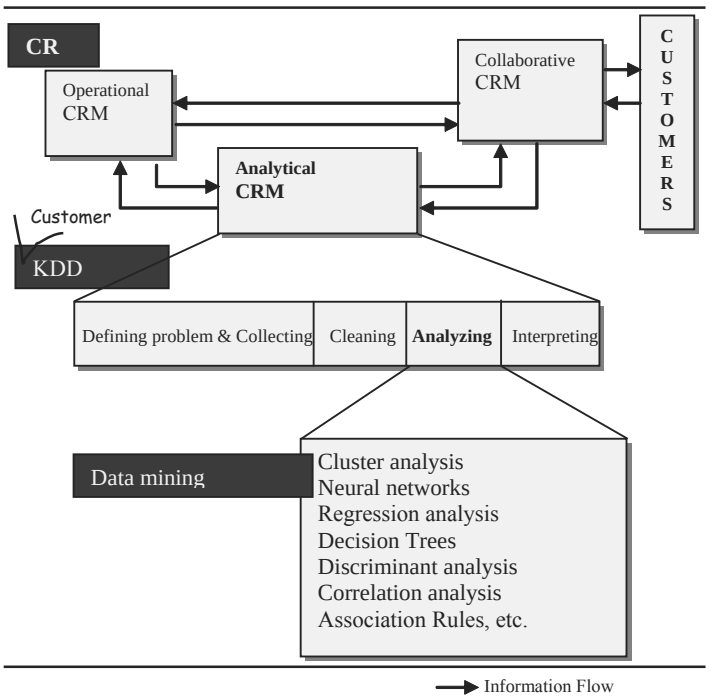
- Selecting the problem area and choosing a tool for representing the goal to be achieved
- Collecting the data and choosing tools for representing objects (observations) of the dataset
- Preprocessing of the data: integrating and cleaning data
- Data mining: extracting pieces of knowledge
- Post-processing of the knowledge derived: testing and verifying, interpreting, and applying the knowledge to the problem area at hand

In Web-based relationship marketing, three distinct categories of data mining have emerged: Web content mining, Web structure mining, and Web usage mining (Jackson, 2002). Web usage mining is also referred to as clickstream analysis (Edelstein, 2001). Valuable information hidden in the clickstream data of many e-commerce sites can provide sharp diagnostics and accurate forecasts, allowing e-commerce sites to profitably target and reach key customers (Moe & Fader, 2001). Therefore, many detailed studies have been conducted on Web usage mining. For example, Web access pattern tree (WAP-tree) mining is one of the sequential pattern mining techniques for Web log access sequences (Ezeife & Lu, 2005).

Such Web-based CRM systems require large, integrated data repositories and advanced analytical capability. Even though there are many success stories, Web-based CRM projects continue to be expensive and risky undertakings. OLAP refers to the various types of query-driven analysis for analyzing stored data (Berry & Linoff, 1997). Data mining and OLAP can be seen as complementary tools (Jackson, 2002). Both Web-based CRM systems and OLAP, in general, involve vast volumes of both structured and unstructured data. One common challenge with managing this data is to incorporate unstructured data into a data warehouse. Traditional database systems are not designed for unstructured data.

Research in KDD in general is intended to develop methods and techniques to process a large volume of unstructured data in order to retrieve valuable knowledge (which is "hidden" in these databases) that would be compact and abstract, yet understandable and useful for managerial applications (Bruha et al., 2000).

Figure 2. CRM and KDD process connection



STRENGTHENING CRM-KDD LINKAGES

Figure 2 explains the CRM-KDD linkage from a process point of view. As explained previously, the importance of gaining knowledge has been well recognized. In line with this notion, CRM starts with understanding customers and gaining in-depth knowledge about customers. Therefore, the intersection between KDD and CRM can be seen as the analytical CRM part of CRM systems and customer knowledge discovery in database process of overall KDD process as shown in Figure 2. Collaborative CRM systems help collect accurate information from customers, while operational CRM can capitalize on the result of the analyses. The problem definition stage of KDD process can be done also in the management dimension of CRM. Following the definition of KDD and data mining, techniques for data mining are included under the analysis stage of KDD.

Turning to the relationship marketing issues, with help of CRM and KDD technologies, database marketing and one-to-one marketing methods have come to the fore. Direct and database marketing methods can be regarded as powerful instruments to achieve CRM goals even though CRM is not a sub-task of direct and database marketers (Wehmeyer, 2005). The strategic goal of database marketing is to use collected information to identify customers and prospects as individuals and build continuing personalized relationships with them, leading to greater benefits for the individuals and greater profits for the corporation (Kahan, 1998). Database marketing anticipates customer behavior over time and reacts to changes in the customer's behavior. Database marketing identifies unique segments in the database reacting to specific stimuli such as promotions (McKim, 2002).

One-to-one marketing represents the ultimate expression of target marketing—market segments with just one member each—or at least one at a time (Pitta, 1998). It relies on a two-way communication between a company and its customers to enhance a true relationship and allows customers to truly express the desires that the company can help fulfill (Dyche, 2002). A promising solution to implementing one-to-one marketing is the application of data mining techniques aided by information technology. Data mining allows organizations to find patterns within their internal customer data. Whatever patterns are uncovered can lead to target segmentations. Armed with such information, organizations can refine their targets and develop their technology to achieve true one-to-one marketing (Pitta, 1998).

As an extension of one-to-one marketing, the concept of permission marketing is focused on seeking customers' consent about desired marketing methods. Customers not only need to be communicated with as individuals, but they themselves should also be able to stipulate how and when they wish to be approached (Newell, 2003). One-to-one and permission marketing rely heavily on information technology to track individual customers, understand their differences, and acknowledge their interaction preferences (Dyche, 2002).

Table 1. Customer relationship related data analysis and data mining tools

	CUSTOMER RELATIONSHIP MARKETING ISSUES		
	Database Marketing	One-to-One Marketing	Permission Marketing
Issue	Understanding customers with the database on customer behavior over time including reactions to changes	Communicating with customers as individuals Developing custom products and tailored messages based on customers' unspoken needs	Seeking customers' agreement about desired marketing methods
Challenge	Identifies unique segments in the database	Find patterns within the internal customer data. Track individual customers Understand their differences	Track individual customers Understand their differences Acknowledge their interaction preferences Stimulate the customer's response
Possible analysis	Segmentation	Classification Prediction	Classification Dependency Analysis
Data mining technique most likely used	Descriptive and visualization Cluster analysis Neural networks	Regression analysis Neural networks Decision Trees Discriminant Analysis	Descriptive and visualization Neural networks Regression analysis Correlation analysis Decision Trees Discrimination analysis Case-based reasoning Association Rules

Data mining methods allow marketers to sift through growing volumes of data and to understand their customers better. Shaw et al. (2001) introduced three major areas of application of data mining for knowledge-based marketing: (1) customer profiling, (2) deviation analysis, and (3) trend analysis. Also, Jackson (2002) noted that data mining can be used as a vehicle to increase profits by reducing costs and/or raising revenue. Some of the common ways to use data mining in customer relationship contexts include:

- Eliminating expensive mailings to customers who are unlikely to respond to an offer during a marketing campaign.
- Facilitating one-to-one marketing and mass customization opportunities in CRM.

In sum, many organizations use data mining to help manage all phases of the customer lifecycle, and CRM systems can benefit from well-managed data analysis based on data mining. Table 1 summarizes the relationship marketing issues, and includes the possible customer analyses and relevant data mining techniques.

Though data mining can be very useful in finding hidden patterns, there are important notions to keep in mind.

First, it is important to note that KDD and data mining involve continuous human-computer interactions (Firestone, 2005). Data mining is not like a simple easy button. It is the skills and background knowledge of humans that will make a difference in the performance of data mining and KDD. For example, Firestone (2005) argued that the quality of information in the data warehouse is critical. Records and information management managers can enhance the performance of KDD process with the ability to create information that can lead to new knowledge by interpreting and evaluating.

Second, a fit between information technology (IT) in an organization and organizational strategy and marketing requirement is required (Wehmeyer, 2005). Relationship marketing and CRM have been mainly discussed at a strategic level (Wehmeyer, 2005). Therefore, rather than separated marketing support technologies functioning only for itself, there should be a fit (Venkatraman, 1989) between IT support and marketing requirements at the operational as well as at the strategic level to design successful IT-enhanced marketing processes (Wehmeyer, 2005).

FUTURE TRENDS

Due to the advance of information technology, there are more opportunities to collect the data about customers. The Internet provides and promotes interactive communications

between the business and the customers, and it leads to increasing volume of rich data about customers. Based on the rich data collected, it is possible to successfully identify new prospective customers by using customer lifetime value to evaluate the current customers and match their profiles with those of new prospects (Wilson, 2006).

However, in interactive marketing contexts, customers are also able to "block out" the intrusive marketing actions, and therefore, appropriate depth and width of "permissions" should be obtained (Godin, 1999; Krishnamurthy, 2001). Therefore, understanding customers will become more critical as new information technology is being developed.

Furthermore, companies and customers can have opportunities to co-create products, pricing, and distributions. Information technology provides this opportunities by allowing companies to assess each customer individually and then to determine whether to serve that customer directly or via a third party, and whether to create an offering that customizes the product or standardizes the offering (Sheth, Sisodia, & Sharma, 2000). All the decisions to make should be based on thorough analyses of customer data and accurate knowledge generation about customers.

In the similar vein, it is projected that knowledge generation—business intelligence—and data mining technologies are integrated because of importance of business performance management (BPM), which is focused on metrics development and data gathering to measure performance (Firestone, 2005).

Not only knowledge generation, but also knowledge sharing and dissemination through the organization should be considered. Shaw et al. (2001) argued that ownership and access to the marketing knowledge, standards of knowledge interchange, and sharing of applications become critical. Ho and Chuang (2006) argued that it is important to establish a knowledge management and CRM mechanism that has a system, a plan, a classification feature, objectives, and an evaluation management mechanism to satisfy customers.

In various organizational environments, both managing KDD processes to generate customer knowledge and managing customer relationship based on the knowledge generated and shared through organization are challenges for the future.

CONCLUSIONS

This article offered a brief review of customer relationship issues. CRM systems consist of management, technology, and customer dimensions. CRM technologies are divided into three categories: analytical, operational, and collaborative CRM. As the importance of knowledge increases, KDD techniques are receiving attention as systematic processes to generate knowledge. Although CRM and KDD began separately, the two concepts have points of convergence. This article highlighted some of the intersections between the two.

Different relationship marketing issues have emerged, and these rely increasingly on CRM and KDD technologies, especially for in-depth analysis. Various data mining techniques and KDD processes exist and provide the right tools to solve relationship marketing problems. While companies are eager to learn about their customers by using data mining technologies, it is very difficult to choose the most effective algorithms for the diverse range of problems and issues that marketers face (Kim et al., 2002). Even though Table 1 illustrates the main relationship marketing issues, challenges, and the potential analytic data mining tools, it is the analyst who decides creatively which tool is appropriate for what in which situation and how to interpret the results.

From a process point of view as well, gaining customer knowledge becomes critical for managing customer relationships, and systematic knowledge generating processes are of great benefit. For effective customer-centric marketing strategies, the discovered knowledge has to be managed in a systematic manner.

REFERENCES

Berry, M. J. A., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. New York: John Wiley & Sons.

Brachman, R. J., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G., & Simoudis, E. (1996). Mining business databases. *Communications of the ACM, 39*(11), 42-48.

Davids, M. (1999). How to avoid the 10 biggest mistakes in CRM. *The Journal of Business Strategy, 20*(6), 22.

Dyche, J. (2002). *The CRM handbook: A business guide to customer relationship management*. Boston: Addison-Wesley.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. Association for Computing Machinery. *Communications of the ACM, 39*(11), 27.

Firestone, *J. M*. (2005). Mining for information gold. *Information Management Journal, 39*(5), 47.

Goodhue, D. L., Wixom, B. H., & Watson, H. J. (2002). Realizing business benefits through CRM: Hitting the right target in the right way. *MIS Quarterly Executive, 1*(2), 79-94.

Greenberg, P. (2002). *CRM at the speed of light: Capturing and keeping customers in Internet real time* (2nd ed.). Berkeley and London: McGraw-Hill.

Jackson, J. (2002). Data mining: A conceptual overview. *Communications of the Association for Information Systems, 8*(2002), 267-296.

Johnson, L. K. (2002). New views on digital CRM. *Sloan Management Review, 44*(1), 10.

Kim, E., Kim, W., & Lee, Y. (2002). Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems, 34*(2002), 167-175.

Krishnamurthy, S. (2001). A comprehensive analysis of permission marketing. *Journal of Computer-Mediated Communication, 6*(2).

Moe, W. W., & Fader, P. S. (2001). Uncovering patterns in cybershopping. *California Management Review, 43*(4), 106-117.

Moorman, C., Zaltman, G., & Deshpande, R. (1992). Relationships between providers and users of market research: The dynamics of trust within and between organizations. *Journal of marketing research, 24*(August), 314-328.

Newell, F. (2003). *Why CRM doesn't work: How to win by letting customer manage the relationship.* NJ: Bloomberg Press.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems, 31*(2002), 127-137.

Sheth, J. N., Sisodia, R. S., & Sharma, A. (2000). The antecedents and consequences of customer-centric marketing. *Journal of Academy of Marketing Science, 28*(1), 55.

Venkatraman, N. (1989). The concept of fit in strategy research. *Academy of Management Review, 14*(3), 423-444.

Wehmeyer, K. (2005). Aligning IT and marketing—The impact of database marketing and CRM. *Journal of Database Marketing & Customer Strategy Management, 12*(3), 243-256.

Wilson, R. D. (2006). Developing new business strategies in B2B markets by combining CRM concepts and online databases. *Competitiveness Review, 16*(1), 38-45.

KEY TERMS

Clickstream Data: Web usage data. A virtual trail that a user leaves behind while surfing the Internet. For example, every Web site and every page of every Web site that the user visits, how long the user was on a page or site.

Customer Relationship Management (CRM): A core business strategy that promotes interactions and creates and delivers value to targeted customers to improve customer satisfaction and customer retention at a profit. It is grounded in high quality customer data and enabled by information technology.

CRM Systems: Technological part of CRM. Computers and all other information technologies used to help CRM are included.

Data Mining (DM): The process of searching and analyzing data in order to find latent but potentially valuable information and to identify patterns and establish relationships from a huge database.

Electronic Commerce (E-Commerce): Any business done electronically. The electronic business where information technology is applied to all aspects of company's operations.

Knowledge Discovery in Databases (KDD): The process of data selection, sampling, pre-processing, cleaning, transformation, dimension reduction, analysis, visualization, and evaluation for the purpose of finding hidden knowledge from massive databases.

Online Analytical Processing (OLAP): Various types of online query-driven analyses for examining stored data. OLAP enables a user to easily and selectively extract and view data from different points-of-view.

Data Communications and E-Learning

Michael W. Dixon
Murdoch University, Australia

Johan M. Karlsson
Lund Institute of Technology, Sweden

Tanya J. McGill
Murdoch University, Australia

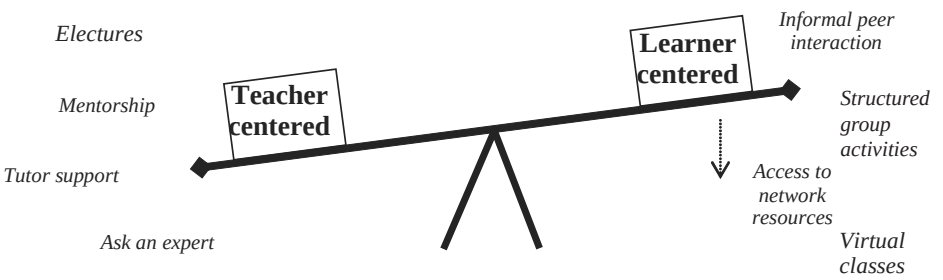
INTRODUCTION

Information and communications technology (ICT) has increasingly influenced higher education. Computer-based packages and other learning objects provide a useful supplement to students studying conventionally by illustrating aspects of the curriculum. Other packages are directed at aspects of course administration such as automated assessment (for example, see Randolph et al. (2002)). Initially such software and materials played only a supplementary role in course offerings, but this has changed rapidly. For example, Coleman et al. (1998) describe a successful early attempt to replace all lecturing with computer-aided learning. Remote delivery of courses also became a viable option because of the advent of the WWW. For example, Petre and Price (1997) report on their experiences conducting electronic tutorials for computing courses. Online education of various sorts is now routinely available to vast numbers of students (Alexander, 2001; Chen & Dwyer, 2003; Peffers & Bloom, 1999). Various terms have been used to label or describe forms of education supported by information technology. These include e-learning (e.g., Alexander, 2001; Campbell, 2004), Web-based learning (e.g. Huerta, Ryan & Igbaria, 2003; Khosrow-Pour, 2002), online learning (e.g., Simon, Brooks & Wilkes, 2003), distributed learning and technology-mediated learning (e.g., Alavi & Leidner, 2001); with e-learning probably the most commonly used term used to describe education and training that networks such as the Internet support.

E-learning has become of increasing importance for various reasons. These include the rise of the information and global economy and the emergence of a consumer culture. Students demand a flexible structure so that they can study, work and participate in family life at the same time (Campbell, 2004). This flexibility is reflected in alternative delivery methods that include online learning and Internet use. We have also become more sensitive to cultural and gender differences, and to the learning needs of the challenged. These needs may be addressed by e-learning (Campbell, 2004).

A number of studies have compared student learning and satisfaction between e-learning and traditional classroom teaching. In an early study, Hiltz and Wellman (1997) found that mastery of course material was equal or superior to that in the traditional classroom and that e-learning students were more satisfied with their learning on a number of dimensions. In particular, they found that the more students perceived that collaborative learning was taking place, the more likely they were to rate their learning outcomes as superior to those achieved in the traditional classroom. They did however identify some disadvantages to e-learning. These included ease of procrastination and information overload. More recently, Piccoli, Ahmad and Ives (2001) found that the academic performance of students in the two environments was similar, but that while e-learning students had higher levels

Figure 1. Categories of online activities



of self-efficacy, they were less satisfied with the learning process. Alexander's comment that "the use of information technology does not of itself improve learning" (Alexander, 2001, p. 241) perhaps highlights the fact that e-learning can be many things and that the intention to introduce e-learning is no guarantee of success.

The different types of teaching and learning activities that are made possible by the Internet are shown in Figure 1. Harasim and Hiltz (1995) divided these activities into two categories: learner or teacher centered. There is, however, no common agreement about which category is the best and many researchers argue for a mixture of learning activities, emphasizing group learning (Bento & Schuster, 2003; Klobas & Renzi, 2003). At the moment there still seems to be an overemphasis on teacher centered approaches, which hopefully will slowly change as a better knowledge of e-learning develops.

BACKGROUND

This article provides an illustration of blended e-learning by describing how we deliver and manage courses in a postgraduate degree in telecommunications management. We aim to foster learner centered education while providing sufficient teacher centered activities to counter some of the known concerns with entirely learner centered education. We use the Internet as the communication infrastructure to deliver teaching material globally and Lotus LearningSpace to provide the learning environment. While the primary aim of our approach is to enhance the student learning process, there are also other incentives that are consistent with this. The university is able to attract a more diverse range of students – those requiring flexibility of study and the other benefits of e-learning. Thus initiatives of this type can benefit the university while meeting the additional needs of students that are discussed in the introduction.

The use of learning and content management systems (LCMS) such as Blackboard, WebCT and Lotus LearningSpace have made e-course development less onerous for faculty. These systems provide a set of tools for publishing, communicating, and tracking student activity. Various guidelines have been suggested for evaluating and choosing software for e-learning (Klobas & Renzi, 2000). After establishing our requirements for a software tool for developing and delivering courses online, we evaluated various alternatives. The requirements that we identified included:

- Instructors should not have to program and debug HTML code;
- All courses should have the same professional look and feel without having to hire computer programmers to write special software, and students should always be presented with the same interface for all their courses;
- The software should be fully integrated (one software package should allow the instructor to do everything required, such as course development and course management);
- Professional support.

After evaluating various alternatives we choose Lotus LearningSpace (LS). Successful use of LS by instructors proved to be significantly less dependent on the technical knowledge of the instructor than was the case with some other popular LCMS. It allows the instructor to focus on the learning of the students rather than on creating and debugging HTML.

LS provides instant feedback to the students and instructor, and enables progress and problems that students encounter as they go through the curriculum to be monitored. Students also have a discussion area where they can ask questions and communicate with the instructor as well as with other students.

LS allows us to create distributed courses that students and instructors can access whether they are online or offline. Students are able to download material for a course onto their machine so they can go through the curriculum without having to have a direct Internet connection. Using the offline access method makes it easier for students to learn wherever they are located and for instructors to develop and manage course material and reduce critical network bandwidth requirements. Features that facilitate flexible student centered learning include:

- Schedule - provides students with a structured approach to assignments, materials, and assessments. Through the schedule, students can link to everything required to complete their course.
- MediaCenter - allows immediate and searchable access to all materials for the course as the instructor makes them available.
- CourseRoom - provides a discussion group facility, which hosts collaborative interchange between student groups and/or students and instructors.
- Profiles - helps students and instructors get to know their classmates to form productive teams and to network outside the course.

Features that facilitate course management include LS Central for course management and the Assessment Manager for setting up and tracking of students' progress in their courses.

USING LOTUS LEARNINGSPACE TO ENHANCE LEARNING OF DATA COMMUNICATIONS

Depending on course content and pedagogy the migration to e-learning can be more or less laborious. We have chosen one course in a telecommunication management degree to provide an example of how e-learning was implemented. Appropriately, this course is about networking technology.

The Data Communications Systems course provides an introduction to networking and networking devices focusing on Local Area Networks (LANs) and Wide Area Networks (WANs). Network design, Ethernet, ISDN, Frame Relay, and TCP/IP are introduced, including IP addressing and router protocols. There is a strong practical context that provides students with the opportunity to build, configure and problem solve in a multi-router network environment. This course also includes the first part of the Cisco Certified Network Associate (CCNA) curriculum. The Data Communications Systems course provides a good example of a difficult class to integrate with LS because of the external curriculum, which is contained on a separate Web server, and the external assessment requirement for students.

Students are required to log in to LS once they have selected the course. Students can access LS with a Lotus Notes Client or through a Web browser. The Lotus Notes Client allows students to work off-line by placing an image of what is on the LS server on their laptop or PC at home. This allows students to work while they travel and can reduce the amount of time they are required to be connected to the Internet. When they connect up to the Internet they can resynchronize their copy with the master copy located on a Lotus Notes Server; this flexibility is a major goal of e-learning. The students then use LS schedule to follow the schedule for the course (see Figure 2). Through the schedule, students can access the external curriculum, which resides on a separate Web server.

The course takes a blended learning approach with both face-to-face classes and online learning. This enables us to take advantage of the benefits of technology mediated learning, but does not risk losing the strengths of face-to-face teaching. Theory is presented to students in the form of online topics (Web-based), mini lectures, and laboratories. LS is used to integrate course material from a variety of sources, for example, the Cisco material that contributes to the CCNA. This kind of external material must be integrated seamlessly with local content so that students see a totally integrated system when it comes to the delivery of each course. The online teaching material combines Web based text and graphics to explain concepts (see Figure 3 for a sample screen). Short movies are also used to illustrate concepts that are difficult to communicate with just static text and graphics. The students are given aims and objectives at the start of each topic. The teaching material covers these topics in detail and at the end of each topic students have optional self-assessment quizzes that allow them to gauge their un-

Figure 2. Sample schedule screen

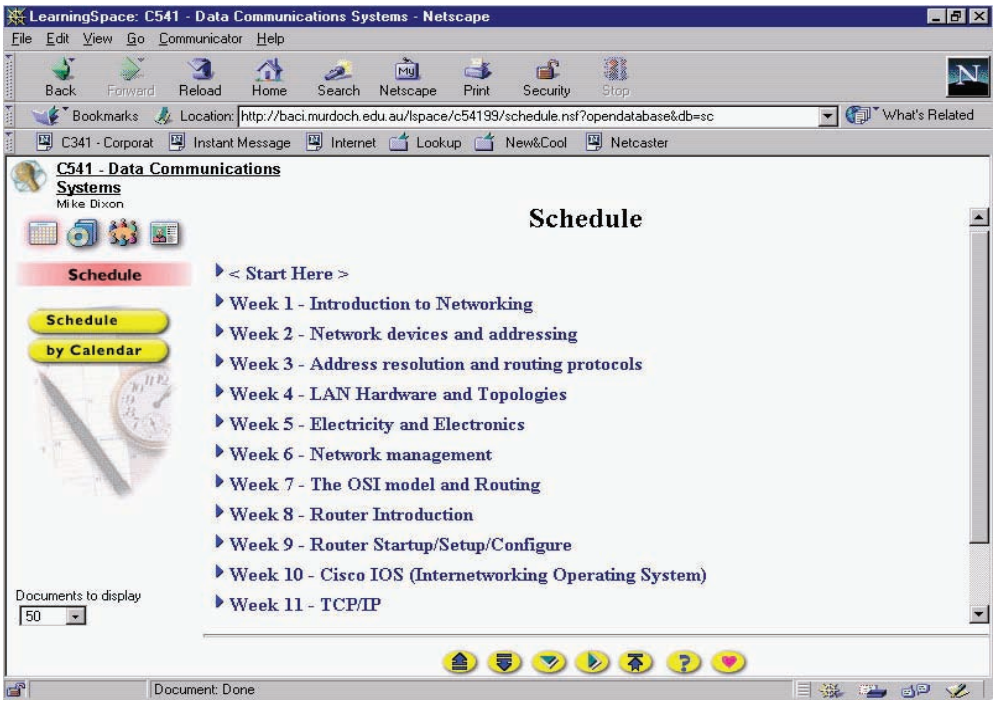
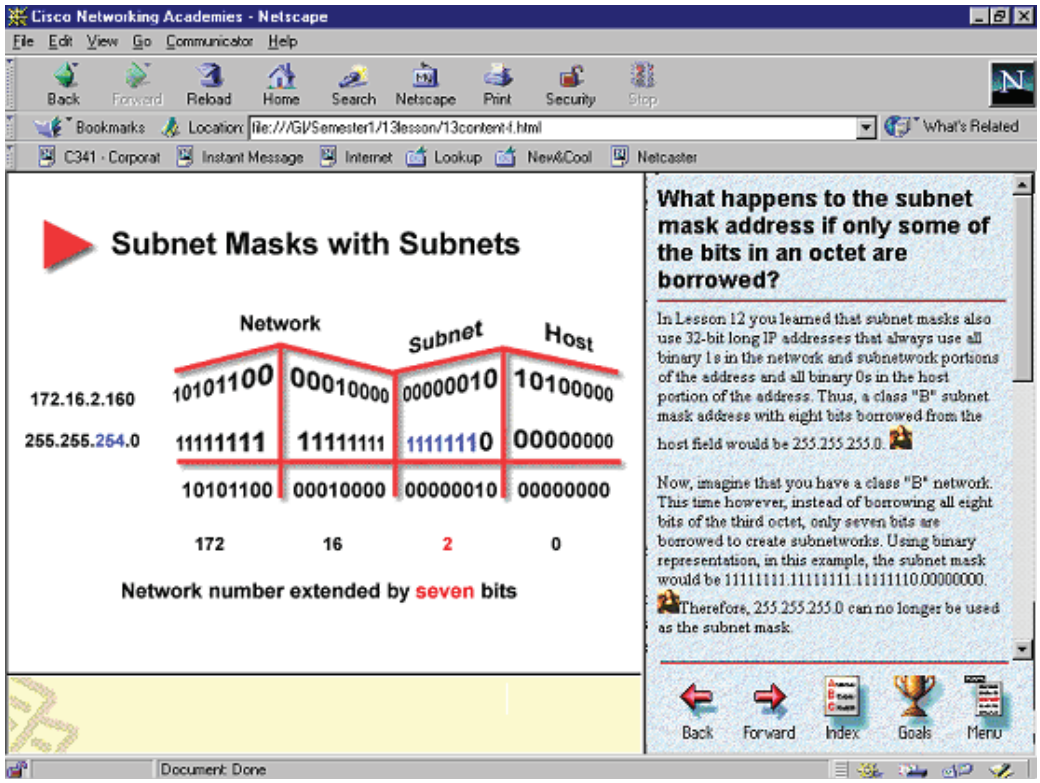


Figure 3. Sample course material screen



derstanding of the material. The multiple modes of content delivery cater to different cognitive styles of students.

Instructors also use these online quizzes to measure the understanding of the students before they attend classes. The instructor is able to identify students who are not keeping up to date with their work and also areas that students are having problems with. The mini lectures can then focus on the areas where students are having problems. The instructor can also discuss problem areas in the discussion group. Assignments are submitted online, graded locally and returned with the instructor's personal comments and grade.

There are two Cisco online exams for the course and students are required to pass these exams with a minimum score of 80%. Students are allowed to take the exam more than once. All students must also take a separate supervised written final exam to meet University requirements. The online Cisco exams are done on separate assessment servers, as they are part of the certification process. The Cisco Networking Academy results of all the students around the world are maintained in a database in Arizona so that instructors can analyze how their class has answered questions and compare the results of their students with those of students in other institutions around the world. The final exams for each course are taken locally in a more traditional way.

The course also includes on-campus practical work to prepare students to solve real-world networking problems. Students work in groups during the practical lab sessions. Students are given timed individual and group practical exams at the end of each major component. Students are allowed to take these exams as many times as they like but are required to pass all these exams.

These factors all contribute to facilitating student learning. Faster and more frequent feedback on the material keeps students more in touch with their progress. Testing and checkpoints with built in repetition are important for long-term retention and understanding of the material. The facility to continue to work with the teaching material until an 80% pass is achieved enhances performance. Students see important material multiple times so that their learning is reinforced and students are able to study wherever they are and still be part of the student community. The use of a virtual discussion group enhances the sense of community among the students and teachers. Combining a learner centered approach with LS allows us to achieve a quality course online.

FUTURE TRENDS

E-learning will continue to play an increasing role in education and training. Greater broadband access will enable delivery of richer content and greater interactivity. Convergence of

information technologies such as notebooks, phones and television and the development of pervasive computing will provide even greater flexibility to students. Educators and students are coming to understand that learning is lifelong and that technology is a valuable tool in supporting it.

CONCLUSION

E-learning is changing the face of university education. This article discussed an approach used to adopt a learner centered environment within an Internet based degree. As the course used to illustrate the issues in the article is about telecommunications, the Internet is a very appropriate medium for instruction. Students learn about telecommunications while accessing information through the Internet. Technology can provide flexibility of learning and hence enrich the learning experience and diversify the student mix.

REFERENCES

Alavi, M., & Leidner, D.E. (2001). Research commentary: Technology-mediated learning — a call for greater depth and breadth of research. *Information Systems Research, 12*(1), 1-10.

Alexander, S. (2001). E-learning developments and experiences. *Education + Training, 43*(4/5), 240-248.

Bento, R., & Schuster, C. (2003). Participation: The online challenge. In A. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 156-164). Hershey, PA: Information Science Publishing.

Campbell, K. (2004). *E-ffective writing for e-learning environments.* Hershey, PA: Information Science Publishing.

Chen, W.-F., & Dwyer, F. (2003). Hypermedia research: Present and future. *International Journal of Instructional Media, 30*(2), 143-148.

Coleman, J., Kinniment, D., Burns, F., Butler, T., & Koelmans, A. (1998). Effectiveness of computer-aided learning as a direct replacement for lecturing in degree-level electronics. *IEEE Transactions on Education, 41,* 177-184.

Harasim, L., Hiltz, S.R., Teles, L., & Turoff, M. (1995). *Learning networks - a field guide to teaching and learning on-line.* Cambridge, MA: The MIT Press.

Hitlz, S.R., & Wellman, B. (1997). Asynchronous learning networks as a virtual classroom. *Communications of the ACM, 40*(9), 44-49.

Huerta, E., Ryan, T., & Igbaria, M. (2003). A comprehensive Web-based learning framework: Toward theoretical diversity. In A. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 24-35). Hershey, PA: Information Science Publishing.

Khosrow-Pour, M. (Ed.). (2002). *Web-based instructional learning.* Hershey, PA: IRM Press.

Klobas, J., & Renzi, S. (2000). Selecting software and services for Web-based teaching and learning. In A. Aggarwal (Ed.), *Web-based learning and reaching technologies: Opportunities and challenges* (pp. 43-59). Hershey, PA: Idea Group Publishing.

Klobas, J., & Renzi, S. (2003). Integrating online educational activities in traditional courses: University-wide lessons after three years. In A. Aggarwal (Ed.), *Web-based education: Learning from experience* (pp. 415-439). Hershey, PA: Information Science Publishing.

Peffers, K., & Bloom, S. (1999). Internet-based innovations for teaching IS courses: The state of adoption, 1998-2000. *Journal of Information Technology Theory and Application, 1*(1), 1-6.

Petre, M., & Price, B. (1997). Programming practical work and problem sessions via the Internet. *ITiCSE 97 Working Group Reports and Supplemental Proceedings,* 125-128.

Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training. *MIS Quarterly, 25*(4), 401-427.

Randolph, G.B., Swanson, D.A., Owen, D.O., & Griffin, J.A. (2002). Online student practice quizzes and a database application to generate them. In M. Khosrow-Pour (Ed.), *Web-based instructional learning.* Hershey, PA: IRM Press.

Simon, J.C., Brooks, L.D., & Wilkes, R.B. (2003). Empirical study of students' perceptions of online courses. In T. McGill (Ed.), *Current issues in IT education.* Hershey, PA: IRM Press.

KEY TERMS

Blended Learning: E-learning used in conjunction with other teaching and learning methods.

Cisco Certified Network Associate (CCNA): A data communications industry certification.

Distributed Learning: Using a wide range of information technologies to provide learning opportunities beyond the bounds of the traditional classroom.

E-Course: Another term for an online course.

E-Learning: The use of new multimedia technologies and the Internet to improve the quality of learning.

Learning and Content Management Systems (LCMS): These systems provide a set of tools for publishing, communicating, and tracking student activity.

Learning Objects: Small (relative to the size of an entire course) instructional components that can be reused in different learning contexts. Learning objects are generally considered to be digital materials deliverable over the Internet.

Online Learning: An inclusive term for any form of learning supported by computer based training.

Pervasive Computing: Technology that has moved beyond the personal computer to everyday devices with embedded technology and connectivity. The goal of pervasive computing is to create an environment where the connectivity of devices is embedded in such a way that the connectivity is unobtrusive and always available.

Technology-Mediated Learning: Learning where the learner's interactions with learning materials, other students and/or instructors are mediated through information technologies.

Web-Based Learning (WBL): Use of Internet technologies for delivering instruction.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 685-690, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Data Dissemination in Mobile Databases

Agustinus Borgy Waluyo
Monash University, Australia

Bala Srinivasan
Monash University, Australia

David Taniar
Monash University, Australia

INTRODUCTION

The development of wireless technology has led to *mobile computing*, a new era in data communication and processing (Barbara, 1999; Myers & Beigl, 2003). With this technology, people can now access information anytime and anywhere using a portable, wireless computer powered by battery (e.g., PDAs). These portable computers communicate with a central stationary server via a wireless channel. Mobile computing provides *database applications* with useful aspects of wireless technology known as mobile databases.

The main properties of mobile computing include mobility, severe power and storage restriction, frequency of disconnection that is much greater than a traditional network, bandwidth capacity, and asymmetric communications costs. Radio wireless transmission usually requires a greater amount of power as compared with the reception operation (Xu, Zheng, Zhu, & Lee, 2002). Moreover, the life expectancy of a battery (e.g., nickel-cadmium, lithium ion) was estimated to increase time of effective use by only another 15% (Paulson, 2003). Thus, efficient use of energy is definitely one of the main issues.

Data dissemination (can also be called *data broadcasting*) is one way to overcome these limitations. With this mechanism, a mobile client is able to retrieve information without wasting power to transmit a request to the server. Other characteristics of data dissemination include: scalability as it supports a large number of queries; query performance which is not affected by the number of users in a cell as well as the request rate; and effective to a high-degree of overlap in the user's request. In this article, the terms data dissemination and data broadcasting are used interchangeably.

The ultimate challenge in data dissemination is to minimize *the response time* and *tuning time* of retrieving database items. Response time is the total of elapsed time required for the data of interest to arrive in the channel and the download time, while tuning time is the amount of time that a client is required to listen to the channel, which is used to indicate its energy consumption. In some cases, the response time is equal to the tuning time.

This article describes a state-of-the art development in data dissemination strategies in mobile databases. Several strategies for improving the query performance by disseminating data to a population of mobile users will be explained.

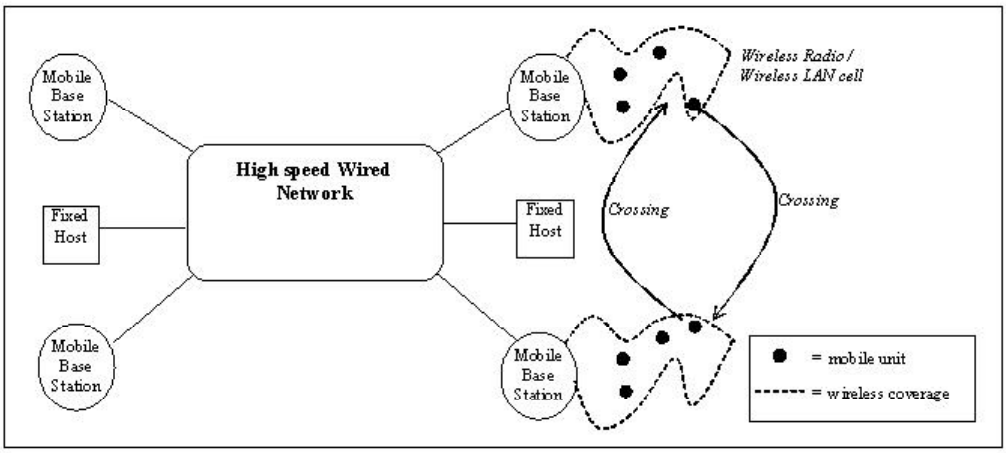
BACKGROUND

In general, each mobile user communicates with a mobile base station (MBS) to carry out any activities such as a transaction and information retrieval. MBS has a wireless interface to establish communication with the mobile client, and it serves a large number of mobile users in a specific region called a "cell". The number of mobile clients in a cell can be infinite. In mobile environment architecture, each MBS is connected to a fixed network as illustrated in Figure 1.

Mobile clients can move between cells while being active and this intercell movement is known as the handoff process (Trivedi, Dharmaraja, & Ma, 2002). Each client in a cell can connect to the fixed network via wireless radio, wireless local area network (LAN), wireless cellular, or satellite. Each of the wireless networks provides a different bandwidth capacity. However, this wireless bandwidth is too small compared with the fixed network such as asynchronous transfer mode (ATM) that can provide a speed of up to 155Mbps (Elmasri & Navathe, 2003).

Data dissemination refers to the periodic broadcasting of database items to mobile clients through one or more wireless channels (or also called broadcast channels), and the clients filter their desired data on the fly. Access to data is sequential. The behavior of the broadcast channel is unidirectional which means the server disseminates a set of data periodically to a multiple number of users. This mechanism is also known as the *push-mechanism* (Malladi & Davis, 2002; Yajima, Hara, Tsukamoto, & Nishio, 2001). It must be noted that data dissemination is different from the data replication mechanism. Conventional data replication distributes a set of database items to one or more identified clients according to a pre-determined requirement. However, data dissemination broadcasts the database items periodically to an unbounded

Figure 1. Mobile environment architecture



number of mobile clients, and the clients filter the data on air based on individual interest.

Figure 2 shows the mechanism of data dissemination. In this article, the term data item corresponds to database record or tuples, and data segment contains a set of data items. A complete broadcast file is referred to as a broadcast cycle. The terms mobile client, mobile computer, mobile unit, mobile user and client are used interchangeably.

DATA DISSEMINATION

Data dissemination schemes are classified into two categories: one is to minimize query response time, and the other minimizes tuning time.

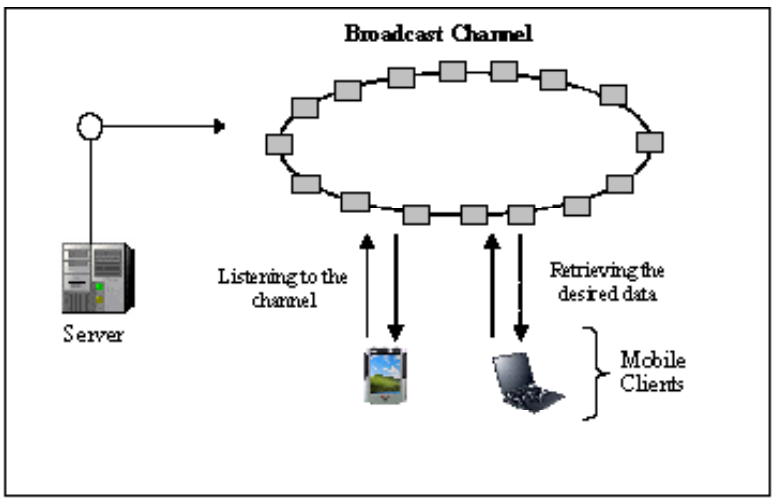
Minimizing Query Response Time

There are several data dissemination schemes, which include:

(i) Selection of Data Items to be Broadcast,
(ii) Non-Uniform Frequency Distribution of Broadcast Data Items,
(iii) Distribution of Data Items over Multiple Channels, and
(iv) Organization of Data Items.

These schemes aim to minimize the query response time by either reducing the waiting time for the desired data to arrive, or, both waiting and download time.

Figure 2. Data dissemination mechanism



a. Selection of Data Items to be Broadcast

A selection mechanism is designed to reduce the broadcast cycle length, which eventually reduces the query response time. During each broadcast cycle, additional items might be qualified as hot and some previously hot items might cool down, and therefore need to be replaced depending on the size of the cycle. A replacement algorithm is required to replace the cold data items with the new hot items. The server determines a set of appropriate database items to be broadcast, using information from queries received. Since hot items are those accessed by most clients, the access pattern of each client on the database items has to be derived from the clients back to the server. Finally, statistics will be compiled and the desired data items can be broadcast appropriately.

There are at least three replacement algorithms namely *Mean Algorithm*, *Window Algorithm*, and *Exponentially Weighted Moving Average* (*EWMA*) *Algorithm* (Leong & Si, 1997). These algorithms maintain a score for each database item to estimate the access probability of the next broadcast cycle. The scores are defined by measuring the cumulative access frequencies of each database item over the length of the observation period.

However, as the size of the database increases, the number of broadcast items may also increase accordingly. This situation will lead to an increase in response time. The next scheme can be used to improve the response time of the majority requests by manipulating the frequency of hot items to be broadcast.

b. Non-Uniform Frequency Distribution of Broadcast Data Items

The difference in bandwidth capacity between the downstream communication and upstream communication has created a new environment called the *Asymmetric Communication Environment*. In fact, there are two situations that can lead to communication asymmetry (Acharya, Alonso, Franklin, & Zdonik, 1995). One is raised as a result of the capability of physical devices. For example, servers have powerful broadcast transmitters, while mobile clients have little transmission capability. The other is due to the patterns of information flow in the application, such as the situation where the number of servers is far less than the number of clients. It is asymmetric because there is not enough capacity to handle simultaneous requests from multiple clients.

Broadcast Disk is an information system architecture, which utilizes multiple disks of different sizes and speeds on the broadcast medium. This architecture is used to address the above situations (Acharya et al., 1995). The broadcast consists of chunks of data from different disks on the same broadcast channel. The chunks of each disk are evenly scattered. However, the chunks of the fast disks are broadcast more frequently than the chunks of the slow disks. This is the opposite of a flat broadcast where the expected delay before obtaining an item of interest is the same for all broadcast items. With this differing broadcast frequency of different items, hot items can be broadcast more often than others. The server is assumed to have the indication of the clients' access patterns so that it can determine a broadcast strategy that will give priority to the hot items.

c. Distribution of Data Items over Multiple Channels

An alternative strategy to improve query response time is to distribute the broadcast data over more than one broadcast channel. Moreover, a certain pattern of distribution such as Data Stripping, Data Replication, and Data Replication can be used to handle obstacles like noise and signal distortion that may affect wireless transmission (Leong & Si, 1995). Figure 3 illustrates the three mechanisms.

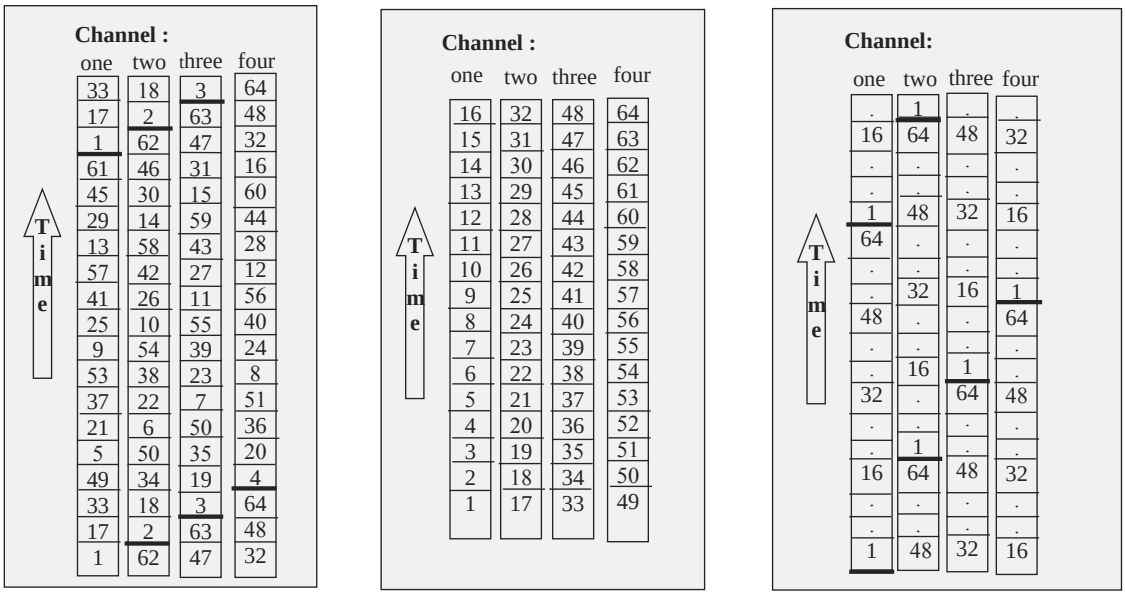
As shown in Figure 3(a), the data stripping mechanism broadcasts consecutive data items over a multiple channel array and the data items are broadcast at certain intervals to allow the client sufficient time to switch from one channel to another. The data partitioning mechanism in Figure 3(b) allows the database to be partitioned into a number of data segments, and each data segment is placed in a different channel. In Figure 3(c), the database is replicated across all channels, and is therefore called the data replication mechanism.

To avoid the effect of too large data items in a channel, a strategy to determine the optimum number of database items to be broadcast in a channel is needed (Waluyo, Srinivasan, & Taniar, 2003a). In this strategy, the query response time over an on-demand channel is used as a threshold point. Subsequently, the length of the broadcast cycle is split, and broadcast over multiple channels.

d. Organization of Data Items

The previous schemes are concerned with retrieving a single data item. However, in most cases multiple data items are involved. Organizing database items over the broadcast channel can be applied to reduce waiting time as well as download time. Traditional semantic query optimization is employed to reduce the download time when the application involves multiple entity types, and the query accesses related entities from different entity types (Si & Leong, 1999). The organization of broadcast data is designed to match the query access pattern from mobile clients as closely as possible. As the query initiated by each client varies, this problem is sometimes considered an NP-hard problem.

Figure 3. Data stripping, partitioning, and replication mechanism



(a) Data Stripping Mechanism (b) Data Partitioning Mechanism (c) Data Replication Mechanism

The difficulty is to decide the broadcast order in advance, even without much knowledge of any future query. In general, an access graph is needed to identify the optimal organization of database items over a channel. The access graph is used to represent the dependency of data items. Once the access graph is built, a certain algorithm is utilized to determine the best broadcast program. A cost model called the Semantic Ordering Model (SOM) can be used to determine the most efficient access graph (Si & Leong, 1999). SOM is defined into two models: namely, Simple and Extended SOM. Simple SOM considers only the relationship among entity types while Extended SOM incorporates historical query access patterns. The branch and bound like algorithm (Si & Leong, 1999), Heuristics algorithm (Hurson, Chehadeh, & Hannan, 2000), randomized algorithm (Bar-Noy, Naor, & Schieber, 2000), and Genetic algorithm (Huang & Chen, 2002, 2003) are some algorithms that can be used to identify the most effective organization of broadcast data items. The final broadcast program can be distributed over either a single or multiple channels.

Minimizing Tuning Time

A broadcast indexing scheme is needed to reduce the tuning time by providing accurate information for a client to tune in at the appropriate time for the required data (Lee, Leong, & Si, 2002). In this scheme, some form of directory is broadcast along with the data, and the clients obtain the index directory from the broadcast and use it in subsequent reads. The information generally also contains the exact time of the data to be broadcast. As a result, mobile clients are able to conserve the energy of their unit by switching to "*doze*" mode and back to "*active* mode" when the data is about to be broadcast.

In general, a broadcast indexing technique involves a trade-off between optimizing the client tuning time and the query response time. The consequence of minimizing one of them is the increase of the other. For instance, to minimize the response time is to reduce the length of broadcast cycles. In this case, the index can be broadcast once in each cycle but it will make the tuning time suffer since the client will have to wait for the index to arrive which happens only once in each broadcast cycle. On the other hand, when the index directory is frequently broadcast in each broadcast cycle to reduce the tuning time, the response time will be greatly affected due to the occupancy of the index in the cycle. Thus, it is necessary to find the optimal balance between these two factors.

Clustering index, non-clustering index, and multiple index methods are used to determine the index distribution over a broadcast channel (Imielinski, Viswanathan, & Badrinath, 1997). Clustering index refers to clustering of a record's attribute whenever all records that belong to the attribute have the same value consecutively. Non-clustering index defines a method for indexing the non-clustering attributes by partitioning each non-clustered attribute in the broadcast

Figure 4. Multi-level signature

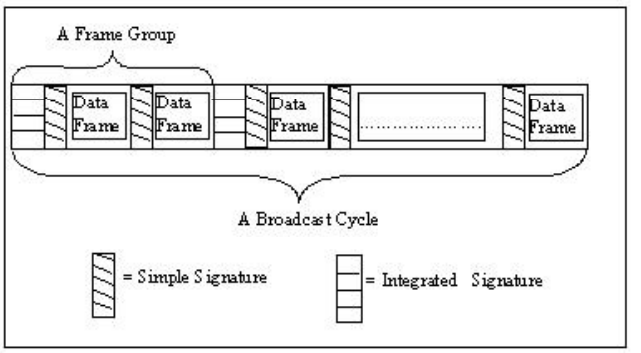
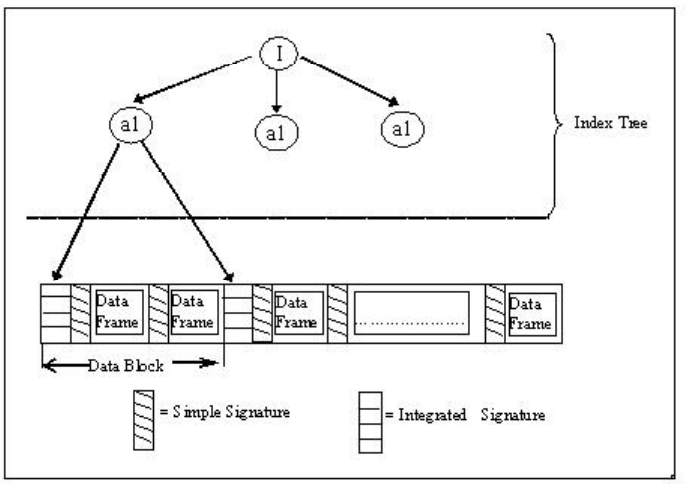


Figure 5. Hybrid indexing



cycle into a number of segments called *meta segments*. In multiple indexes, a second attribute is chosen to cluster the data items within the first clustered attribute.

Broadcast indexing can be *tree-indexing* based, *signature* based, or *hybrid indexing* that is the combination thereof. Tree-indexing based such as Global Index structure is utilised to minimise the index response time (Taniar & Rahayu , 2002; Waluyo, Srinivasan, & Taniar, 2003b). The design of the global index model is based on *B*+ tree structure; it incorporates an index channel, and the data items are broadcast separately in data channels.

A signature based index is derived by hashing the attribute values into bit strings followed by combining them together to form a bit vector or signature (Lee & Lee, 1996). The signature is broadcast together with the data on every broadcast cycle. This mechanism is also applied to the query initiated by the client. To process the query, mobile clients need to tune into the broadcast channel and verify the query signature with the data signature by performing a certain mathematical operation such as the "AND" operation. If the signature is not matched, the client can tune to the "doze" mode while waiting for the next signature to arrive. The main issue with this method is to determine the size of the signature as well as the number of levels of the signature. There are three different signature-based index algorithms, namely *simple signature*, *integrated signature* and *multilevel signature* (Lee & Lee, 1996).

In simple signature, the signature frame is broadcast ahead of each data frame. This makes the total number of signature frames the same as the data frames. An integrated signature is applied for a group of data frames and the signature is calculated accordingly. Figure 4 shows a combination

of these two algorithms, forming a multi-level signature. This technique is designed to interleave with data items in a single channel.

A hybrid indexing technique made up of index tree and signature is expected to outperform the single indexing techniques by integrating the advantages of the two techniques into a single operation (Lee, Hu, & Lee, 1998; Hu, Lee, & Lee, 1999). This technique is shown in Figure 5.

CONCLUSION

The inherent limitations and characteristics of mobile computing such as power, storage, asymmetric communication cost, and bandwidth, have become interesting challenges and research opportunities in the field.

This article describes main issues, and several approaches in data dissemination in mobile databases that have been derived from literature. We classify each scheme into two categories: one is to minimize query response time, and the other is to minimize tuning time. Broadcasting schemes that aim to minimize response time include:

(i) Selection of Data Items to be broadcast,
(ii) Non-Uniform Frequency Distribution of Broadcast Data Items,
(iii) Distribution of Data Items over Multiple Channels, and
(iv) Organization of Data Items.

To minimize the tuning time, broadcast indexing of data items is applied. Hybrid schemes that integrate the advantages of these approaches are also of great potential. These techniques can certainly be improved in a number of ways and a vast amount of research is continuing for this purpose.

REFERENCES

Acharya, S., Alonso, R., Franklin, M., & Zdonik, S. (1995). Broadcast disks: Data management for asymmetric communication environments. In *Proceedings of ACM Sigmod* (pp.199-210).

Barbara, D. (1999). Mobile computing and databases: A survey. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 108-117.

Bar-Noy, A., Naor, J., & Schieber, B. (2000). Pushing dependent data in clients-providers-servers systems. In *Proceedings of the 6th ACM/IEEE on Mobile Computing and Networking* (pp.222-230).

Elmasri, R., & Navathe, S.B. (2003). *Fundamentals of database systems* (4th ed.). Addison Wesley, U.S.A.

Hu, Q., Lee, W.C., & Lee, D.L. (1999). Indexing techniques for wireless data broadcast under data clustering and scheduling. In *Proceedings of the 8th ACM International Conference on Information and Knowledge Management* (pp.351-358).

Huang, J.-L., & Chen, M.-S. (2002). Dependent data broadcasting for unordered queries in a multiple channel mobile environment. In *Proceedings of the IEEE GLOBECOM* (pp.972-976).

Huang, J.-L., & Chen, M.-S. (2003). Broadcast program generation for unordered queries with data replication. In *Proceedings of the 8th ACM Symposium on Applied Computing* (pp.866-870).

Hurson, A.R., Chehadeh, Y.C., & Hannan, J. (2000). Object organization on parallel broadcast channels in a global information sharing environment. In *Proceedings of the 19th International Performance, Computing and Communications* (pp.347-353).

Imielinski, T., Viswanathan, S., & Badrinath, B.R. (1997). Data on air: Organisation and access. *IEEE Transactions on Knowledge and Data Engineering, 9*(3), 353-371.

Lee, D.L., Hu, Q., & Lee, W.C. (1998). Indexing techniques for data broadcast on wireless channels. In *Proceedings of the 5th Foundations of Data Organization* (pp.175-182).

Lee, K.C.K., Leong, H.V., & Si, A. (2002). Semantic data access in an asymmetric mobile environment. In *Proceedings of the 3rd Mobile Data Management* (pp.94-101).

Lee, W.C., & Lee, D.L. (1996). Using signature techniques for information filtering in wireless and mobile environments. *Journal on Distributed and Parallel Databases, 4*(3), 205-227.

Leong, H.V., & Si, A. (1995). Data broadcasting strategies over multiple unreliable wireless channels. In *Proceedings of the 4th Information and Knowledge Management* (pp.96-104).

Leong, H.V., & Si, A. (1997). Database caching over the air-storage. *The Computer Journal, 40*(7), 401-415.

Malladi, R., & Davis, K.C. (2002). Applying multiple query optimization in mobile databases. In *Proceedings of the 36th Hawaii International Conference on System Sciences* (pp. 294-303).

Myers, B.A., & Beigl, M. (2003). Handheld computing. *IEEE Computer Magazine, 36*(9), 27-29.

Paulson, L.D. (2003). Will fuel cells replace batteries in mobile devices? *IEEE Computer Magazine, 36*(11), 10-12.

Si, A., & Leong, H.V. (1999). Query optimization for

broadcast database. *Data and Knowledge Engineering, 29*(3), 351-380.

Taniar, D., & Rahayu, J.W. (2002). A taxonomy of indexing schemes for parallel database systems. *Distributed and Parallel Databases, 12*, 73-106.

Trivedi, K.S., Dharmaraja, S., & Ma, X. (2002). Analytic modelling of handoffs in wireless cellular networks. *Information Sciences, 148*, 155-166.

Waluyo, A.B., Srinivasan, B., & Taniar, D. (2003a). Optimal broadcast channel for data dissemination in mobile database environment. In *Proceedings of the 5th Advanced Parallel Processing Technologies*, LNCS, *2834*: 655-664.

Waluyo, A.B., Srinivasan, B., & Taniar, D. (2003b). Global index for multi channels data dissemination in mobile databases. In *Proceedings of the 18th International Symposium on Computer and Information Sciences*, LNCS, *2869*, 210-217.

Xu, J., Zheng, B., Zhu, M., & Lee, D.L. (2002). Research challenges in information access and dissemination in a mobile environment. In *Proceedings of the Pan-Yellow-Sea International Workshop on Information Technologies for Network Era* (pp.1-8).

Yajima, E., Hara, T., Tsukamoto, M., & Nishio, S. (2001) Scheduling and caching strategies for correlated data in push-based information systems. *ACM SIGAPP Applied Computing Review, 9*(1), 22-28.

KEY TERMS

Broadcast Channel: Unidirectional wireless channel to disseminate a set of database items periodically to multiple numbers of mobile users.

Broadcast Cycle: A complete broadcast file.

Data Dissemination/Broadcasting: Periodical broadcast of database information to mobile clients through one or more wireless channels.

Data Item: Database record or tuples.

Data Segment: A set of data items.

Mobile Base Station (MBS): Fixed host that has wireless interface for communicating with mobile clients.

Mobile Computing: The ability of mobile users to keep connected to the wireless network while traveling, and to access information such as news, weather forecast, email, and query to central database server.

Mobile Database: Mobile users connected to the wireless network and equipped with database application to conduct activity like transaction and information retrieval from central database server.

Query Response Time: The total elapsed time while waiting for the data of interest to arrive in the channel and downloading the data.

Tuning Time: The total time a mobile client must listen to the channel, which is used to indicate its energy consumption.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 691-697, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Data Mining

Sherry Y. Chen
Brunel University, UK

Xiaohui Liu
Brunel University, UK

INTRODUCTION

There is an explosion in the amount of data that organizations generate, collect, and store. Organizations are gradually relying more on new technologies to access, analyze, summarize, and interpret information intelligently. Data mining, therefore, has become a research area with increased importance (Amaratunga & Cabrera, 2004). Data mining is the search for valuable information in large volumes of data (Hand, Mannila, & Smyth, 2001). It can discover hidden relationships, patterns, and interdependencies and generate rules to predict the correlations, which can help the organizations make critical decisions faster or with a greater degree of confidence (Gargano & Ragged, 1999).

There is a wide range of data mining techniques, which has been successfully used in many applications. This article is an attempt to provide an overview of existing data mining applications. The article begins by explaining the key tasks that data mining can achieve. It then moves to discuss applications domains that data mining can support. The article identifies three common application domains, including bioinformatics, electronic commerce, and search engines. For each domain, how data mining can enhance the functions will be described. Subsequently, the limitations of current research will be addressed, followed by a discussion of directions for future research.

BACKGROUND

Data mining can be used to achieve many types of tasks. Based on the kinds of knowledge to be discovered, it can be broadly divided into supervised learning and unsupervised learning. The former requires the data to be pre-classified. Each item is associated with a unique label, signifying the class in which the item belongs. In contrast, the latter does not require pre-classification of the data and can form groups that share common characteristics (Nolan, 2002). To achieve these two main tasks, four data mining approaches are commonly used: classification, clustering, association rules, and visualization.

Classification

Classification, which is a process of supervised learning, is an important issue in data mining. It refers to discovering predictive patterns where a predicted attribute is nominal or categorical. The predicted attribute is called the class. Subsequently, a data item is assigned to one of the predefined sets of classes by examining its attributes (Changchien & Lu, 2001). One example of classification applications is to analyze the functions of genes on the basis of predefined classes that biologists set (see the section on "Classifying Gene Functions").

Clustering

Clustering is also known as *exploratory data analysis* (EDA) (Tukey, 1977). This approach is used in those situations where a training set of pre-classified records is unavailable. Objects are divided into groups based on their similarity. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups (Roussinov & Zhao, 2003). From a data mining perspective, clustering is an approach for unsupervised learning. One of the major applications of clustering is the management of customers' relationships, which is described in the section "Customer Management."

Association Rules

Association rules that were first proposed by Agrawal and Srikant (1994) are mainly used to find out the meaningful relationships between items or features that occur synchronously in databases (Wang, Chuang, Hsu, & Keh, 2004). This approach is useful when one has an idea of different associations that are being sought out. This is because one can find all kinds of correlations in a large data set. It has been widely applied to extract knowledge from Web log data (Lee, Kim, Chung, & Kwon, 2002). In particular, it is very popular among marketing managers and retailers in electronic commerce who want to find associative patterns among products (see the section on "Market Basket Analysis").

Visualization

The visualization approach to data mining is based on an assumption that human beings are very good at perceiving structure in visual forms. The basic idea is to present the data in some visual form, allowing the human to gain insight from the data, draw conclusions, and directly interact with the data (Ankerst, 2001). Since the user is directly involved in the exploration process, shifting and adjusting the exploration goals is automatically done if necessary (Keim, 2002). This approach is especially useful when little is known about the data and the exploration goals are vague. One example of using visualization is author co-citation analysis (see the section on "Author Co-citation Analyses").

DATA MINING APPLICATIONS

As previously discussed, data mining can be used to achieve various types of tasks, such as classification, clustering, association rules, and visualization. These tasks have been implemented in many application domains. The main application domains that data mining can support include bioinformatics, electronic commerce, and search engines.

Bioinformatics

In the past decade, we have been overwhelmed with increasing floods of data gathered by the Human and other Genome Projects. Consequently, a major challenge in bioinformatics is extracting useful information from these data. To face this challenge, it is necessary to develop an advanced computational approach for data analysis. Data mining provides such potentials. Three application areas, which are commonly presented in the literature, are described next.

Clustering Microarray Data

Unsupervised learning produces clustering algorithms, which are being applied to DNA microarray data sets. Many algorithms are available for clustering, such as k-means and hierarchical clustering. These algorithms have different strengths and weaknesses, so they may cause the lack of inter-method consistency in assigning related gene-expression profiles to clusters. To overcome this problem, Swift et al. (2004) proposed a consensus strategy to produce both robust and consensus clustering of gene-expression data and assign statistical significance to these clusters from known gene functions. Robust clustering is compiling the results of different clustering methods reporting only the co-clustered instances grouped together by different algorithms—that is, with maximum agreement across clustering methods. On the other hand, consensus clustering relaxes the full agreement requirement by taking a parameter, "minimum agreement," which allows different agreement thresholds to be explored. It is reported that this approach can improve confidence in gene-expression analysis.

Classifying Gene Functions

Biologists often know a subset of genes involved in a biological pathway of interest and wish to discover other genes that can be assigned to the same pathway (Ng & Tan, 2003). Unlike clustering, which processes genes based on their similarity, classification can learn to classify new genes based on predefined classes, taking advantage of the domain knowledge already possessed by the biologists. Therefore, the classification approach seems more suitable than clustering for the classification of gene functions.

Earlier studies focus on classification of gene functions based on a single source of data, for example, Kuramochi and Karypis (2001). Recently, heterogeneous sources of data have been adopted. For example, Deng, Geng, and Ali (2005) presented a hybrid weighted naive Bayesian network model for the prediction of functional classes of novel genes based on multiple sources of data, such as DNA sequences, expressions, gene structures, database annotations, and homologies. This model can also be used to analyze the contribution of each source of data toward the gene function prediction performance.

Identifying Phenotype Data

In the aforementioned two approaches, the genes are treated as objects, while the samples are the attributes. Conversely, the samples can be considered as the objects and the genes as the attributes. In this approach, the samples can be partitioned into homogeneous groups. Each group may correspond to some particular phenotypes (Golub et al., 1999). Phenotype is observable and physical characteristics of an organism. Over the past decade, growing interest has surfaced in recognizing relationships between the genotypes and phenotypes. Tracing a phenotype over time may provide a longitudinal record for the evolution of a disease and the response to a therapeutic intervention. This approach is analogous to removing components from a machine and then attempting to operate the machine under different conditions to diagnose the role of the missing component. Function of genes can be determined by removing the gene and observing the resulting effect on the organism's phenotype.

Electronic Commerce

The widespread use of the Web has tremendous impact on the way organizations interact with their partners and customers. Many organizations consider analyzing customers'

behavior, developing marketing strategies to create new consuming markets, and discovering hidden loyal customers as the key factors of success. Therefore, new techniques to promote electronic business become essential and data mining is one of the most popular techniques (Changchien & Lu, 2001). Data mining applications in electronic commerce include customer relationship management and market basket analysis.

Customer Relationship Management

For managing customers' relationships, a frequently used approach is to analyze their usage data in order to discover user interests, and then recommendations can be made based on the usage data extracted. A hybrid approach is often used for making recommendations. For example, Wang et al. (2004) used a hybrid method to develop a recommendation system for the cosmetic business. In the system, they segmented the customers by clustering algorithms to discover different behavior groups so that customers in same group have similar purchase behavior. For each group's customers, they then used the association rules to discover their purchase behavior. In addition, they scored each product for each customer who might be interesting in it with the collaborative filtering approach and the content-based filtering. They found that this approach could recommend not only the right product to the right person, but also the right product to the right person at the right time.

Furthermore, Liu and Shih (2005) proposed two hybrid methods for recommending products. These two methods incorporate the advantages of a weighted RFM-based method and a preference-based collaborative filtering method. The former employs association rule mining to identify recommendation rules from customer groups that are clustered according to customer lifetime value (CLV), including recency, frequency, and monetary. The latter is similar to the former, except that the preference-based method groups customers based on purchase preferences. The first proposed hybrid method is to group customers separately based on CLV and purchase preferences. Then, recommendation rules extracted from the weighted RFM-based method are used to recommend products to loyal customers; recommendation rules extracted from the preference-based collaborative filtering method are used to recommend products to less loyal customers. The second hybrid method is to group customers by considering both CLV and purchase preferences and then extract recommendation rules from each group to support recommendations. The experimental results indicated that the second hybrid method outperformed the first, especially when the CLV was weighted more heavily than purchase preferences.

Market Basket Analysis

Market basket analysis is a typical example among the various applications of association rules, and it aims at identifying associations between consumers' choices of different products (Giudici & Passerone, 2003). The data analysed in a market basket analysis usually consist of all purchasing transactions carried out by the consumers in a certain unit of time. The market basket analysis is to understand the association structures between the sales of the different products available. Once the associations are found, they may help planning marketing policies.

Time and location are two important issues for market basket analysis. To address these issues, Chen, Tang, Shen, and Hu (2005) proposed a method called store-chain association rules for a multi-store environment, where stores may have different product-mix strategies that can be adjusted over time. The format of the rules is similar to that of the traditional association rules. However, the rules also contain information on location and time where the rules hold. The rules extracted by the method may be applicable to the entire chain without time restriction, but may also be store and time specific. The results of the empirical evaluation showed that this method has an advantage over the traditional methods: when stores are diverse in size, product mix changes rapidly over time, and larger numbers of stores and periods are considered.

Search Engines

Data mining is of increasing importance for search engines. Traditional search engines offer limited assistance to users in locating the relevant information they need. Data mining can help search engines to provide more advanced features. According to current applications, there are three potential advantages: (a) ranking of pages, (b) improvement of precision, and (c) author co-citation analysis. These advantages are described next.

Ranking of Pages

Data mining identifies the ranking of the Web pages for a particular topic by analyzing the interconnections of a series of related pages. The PageRank (Kamvar, Haveliwala, Manning, & Golub, 2003) applies this approach to find pertinent Web pages. In the PageRank, the importance of a page is calculated based on the number of pages that point to it. This is actually a measure based on the number of backlinks to a page. A backlink is a link pointing to a page, rather than pointing out from a page. This measure is used to prioritize pages returned from a traditional search engine using keyword search. Google applies this measure to rank the search results. The benefit is that central, important, and authoritative Web

pages are given preferences. However, the problem is that it only examines the forward direction. In addition, a much larger set of linked documents is required.

Improvement of Precision

The problem of PageRank is that it is a purely link structure-based computation, ignoring the textual content. Therefore, the precision is low. On a narrowly focused topic, it frequently returns resources for a more general topic. IBM Almaden Research Centre continued develops Clever search engine (Chakrabarti et al., 1999). The main approach is to promote the precision by combining content with link information, breaking large hub pages into smaller units, and computing relevance weight for pages. User's log is the other source that can be used to improve the precision. Zhang and Dong (2002) developed a Chinese image search engine named eeFind by using Matrix Analysis on Search Engine Log (MASEL). The basic idea of MASEL is to use the query log to find relations among users, queries, and clicks on results. The relation between pages chosen after a query and the query itself provides valuable information. After a query, a user usually performs a click to view one result page.

Author Co-Citation Analyses

Author co-citation analysis (ACA) has been widely used as a method for analyzing the intellectual structure of science studies. It can be employed to identify authors from the same or similar research fields. Chen and Paul (2001) used visualization to perform such analysis, and 3D virtual landscape was applied to represent author co-citation structures. The most influential scientists in the knowledge domain appear near the intellectual structure's center. In contrast, researchers who have unique expertise tend to appear in peripheral areas. The virtual landscape also lets users access further details regarding a particular author in the intellectual structure, such as a list of the author's most-cited papers, abstracts, and even the full content of that author's articles.

FUTURE TRENDS

The previous three application domains demonstrate that data mining is a very useful technology and opens new opportunities for data analysis. However, there are still many limitations, which we need to be aware of and should be investigated in future work.

Preparation of Data

It is important to know that the effective implementation of data mining methods depends to a large extent on the availability and quality of the data. Therefore, a cleaning and data transformation step before analysis is usually needed. In addition, we have to understand the characteristics of each application and set clear data mining targets so that truly useful information can be produced with appropriate data mining techniques.

Good Communication

Data mining is an interdisciplinary research area, which involves experts from different domains. One of the significant problems for interdisciplinary research is the wide range and level of domain expertise that are present among potential users so it can be difficult to provide access mechanisms appropriate to all (Kuonen, 2003). Therefore, it is important to have good communications between data mining technologists and users, as well as communications among users from different domains.

Multimedia Mining

Basically, the Web content consists of a variety of media types such as text, image, audio, video, and so forth. Multimedia mining is used to mine the unstructured information and knowledge from these online multimedia sources. However, most current research in the area of data mining still focused on text data, and multimedia mining has received less attention than text mining (Kosala & Blockeel, 2000) and opens a new window for future research to explore.

Evaluation of Effectiveness

As described in the previous section, data mining enhances the functions of the applications in different domains. However, the effectiveness of these applications is still a question. Certainly, there is a need to conduct more empirical studies to verify the effectiveness and evaluate the performance of these applications. In particular, we need more evaluations from users' points of view. Better and more detailed evaluations can lead to better understanding of users' needs and help to develop more effective applications.

CONCLUSIONS

In summary, data mining can search for interesting relationships and global patterns from various types of resources. These relationships and patterns represent valuable knowledge about the objects that are reflected by many applications in the real world. In this article, we have given some background to data mining and have provided an overview of three application domains, including bioinformatics, electronic commerce, personalized environments, and search

engines. It should be noted that data mining has also been applied to other application domains, such as digital libraries, Web-based learning, health care, and so forth. This is another direction for future research to investigate what major functions are required for each application domain and to develop concrete criteria for the evaluation of their effectiveness. These works can be integrated together to generate guidelines, which can be used for commercial and research communities to select suitable data mining techniques. The ultimate goal is to enhance the functions and performances of these applications by exploiting the full potential of data mining techniques.

ACKNOWLEDGMENT

This study is in part funded by the Arts and Humanities Research Council in UK (Grant Reference: MRG/AN9183/APN16300).

REFERENCES

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In J. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Databases*, Santiago, Chile (pp. 478-499).

Amaratunga, D., & Cabrera, J. (2004). Mining data to find subsets of high activity. *Journal of Statistical Planning and Inference, 122*(1/2), 23-41.

Ankerst, M. (2001). Visual data mining with pixel-oriented visualization techniques. In *Proceedings of ACM SIGKDD Workshop on Visual Data Mining*, San Francisco.

Chakrabarti, S., Dom, B. E., Kumar, S. R., Raghavan, P., Rajagopalan, S., Tomkins, A., et al. (1999). Mining the Web's link structure. *Computer, 32*(8), 60-67.

Changchien, S., & Lu, T. (2001). Mining association rules procedure to support online recommendation by customers and products fragmentation. *Expert Systems with Applications, 20*(4), 325-335.

Chen, C., & Paul, R. J. (2001). Visualising a knowledge domain's intellectual structure. *Computer, 34*(3), 65-71.

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems, 40*(2), 339-354.

Deng, X., Geng, H., & Ali, H. (2005, August 8-11). Learning yeast gene functions from heterogeneous sources of data using hybrid weighted Bayesian networks. In *Proceedings of IEEE Computational Systems Bioinformatics Conference (CSB'05)* (pp. 25-34).

Gargano, M. L., & Ragged, B. G. (1999). Data mining—A powerful information creating tool. *OCLC systems services, 15*(2), 81-90.

Giudici, P., & Passerone, G. (2003) Data mining of association structures to model consumer behaviour. *Computational Statistics & Data Analysis, 38*(4), 533-541.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gassenbeck, M., Mesirov, J. P., et al. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science, 286*(5439), 531-537.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.

Kamvar, S., Haveliwala, T., Manning, C., & Golub, G. (2003, May 20-24). Extrapolation methods for accelerating PageRank computations. In *Proceedings of the 12th International World Wide Web Conference*, Budapest, Hungary (pp. 261-270). New York: ACM Press.

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 1-8.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM SIGKDD Explorations, 2*(1), 1-15.

Kuonen, D. (2003). Challenges in bioinformatics for statistical data miners. *Bulletin of the Swiss Statistical Society, 46*, 10-17.

Kuramochi, M., & Karypis, G. (2001, March 4-6). Gene classification using expression profiles: A feasibility study. In *Proceedings of the Second IEEE International Symposium on Bioinformatics & Bioengineering (BIBE 2001)* (pp. 191-201). Washington, DC: IEEE Computer Society.

Lee, K. C., Kim, J. S., Chung, N. H., & Kwon, S. J. (2002). Fuzzy cognitive map approach to Web-mining inference amplification. *Expert Systems with Applications, 22*(3), 197-211.

Liu, D., & Shih, Y. (2005). Hybrid approaches to product recommendation based on customer lifetime value and purchase preferences. *Journal of Systems and Software, 77*(2), 181-191.

Ng, S., & Tan, S. (2003). On combining multiple microarray studies for improved functional classification by whole-dataset feature selection. *Genome Informatics, 14*, 44-53.

Nolan, J. R. (2002). Computer systems that learn: An empirical study of the effect of noise on the performance of three classification methods. *Expert Systems with Applications, 23*(1), 39-47.

Roussinov, D., & Zhao, J. L. (2003). Automatic discovery of similarity relationships through Web mining. *Decision Support Systems, 35*(1), 149-166.

Swift, S., Tucker, A., Vinciotti, V., Marin, N., Orengo, C., Liu, X, & Kellam, P. (2004). Consensus clustering and functional interpretation of gene-expression data. *Genome Biology, 5*(11), R94. Retrived from http://genomebiology.com/2004/5/11/R94

Tukey, J. (1977). *Exploratory data analysis*. Addison-Wesley.

Wang, Y., Chuang, Y., Hsu, M., & Keh, H. (2004). A personalized recommender system for the cosmetic business. *Expert Systems with Applications, 26*(3), 427-434.

Zhang, D., & Dong, Y. (2002). A novel Web usage mining approach for search engine. *Computer Networks, 39*(3), 303-310.

KEY TERMS

Author Co-Citation Analysis: The analysis of how authors are cited together.

Bayesian Network: A directed acyclic graph of nodes representing variables and arcs representing dependence relations among the variables.

Bioinformatics: An integration of mathematical, statistical, and computational methods to organize and analyze biological data.

Collaborative Filtering: A technique that is used for making recommendations by computing the similarities among users.

Content-Based Filtering: A technique that involves a direct comparison between the content or attributes of a user's profile and the document to make recommendations.

Electronic Commerce: Commercial activities that facilitate the buying and selling of goods and services over the Internet.

Microarrays: A high-throughput technology that allows the simultaneous determination of mRNA abundance for many thousands of genes in a single experiment.

Multimedia Mining: A new field of knowledge discovery in multimedia documents, dealing with non-structured information such as texts, images, videos, audio, and virtual data.

Noisy Data: Errors in the data, due to the nature of data collection, measurement, or sensing procedures.

Search Engines: Web services that help search through Internet addresses for user-defined terms or topics.

Data Mining in Franchising

Ye-Sho Chen
Louisiana State University, USA

Grace Hua
Louisiana State University, USA

Bob Justis
Louisiana State University, USA

INTRODUCTION

Franchising has been a popular approach given the high rate of business failures (Justis & Judd, 2002; Thomas & Seid, 2000). Its popularity continues to increase, as we witness an emergence of a new business model, Netchising, which is the combination power of the Inter*net* for global demand-and-supply processes and the international fran*chising* arrangement for local responsiveness (Chen, Justis, & Yang, 2004). For example, *Entrepreneur* magazine—well known for its Franchise 500 listing—in 2001 included Tech Businesses into its Franchise Zone that contains Internet Businesses, Tech Training, and Miscellaneous Tech Businesses. At the time of this writing, 40 companies are on its list. Netchising is an effective global e-business growth strategy (Chen, Chen, & Wu, 2006), since it can "offer potentially huge benefits over traditional exporting or foreign direct investment approaches to globalization" and is "a powerful concept with potentially broad applications" (Davenport, 2000, p. 52).

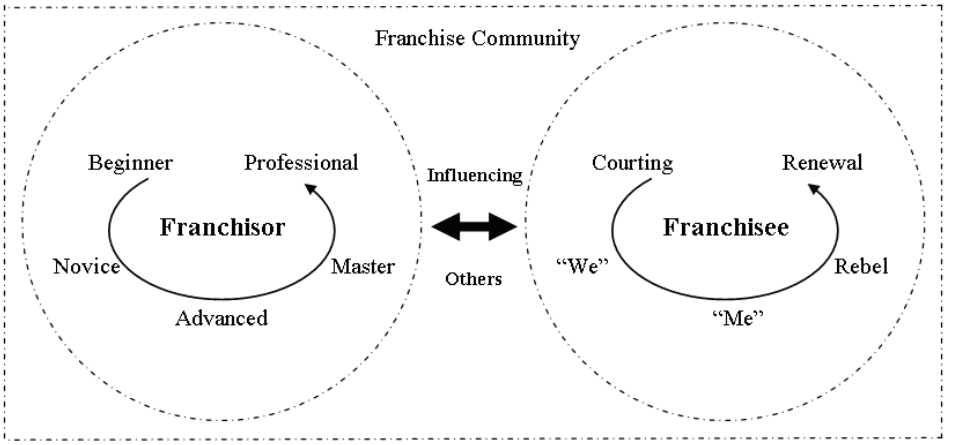
In his best seller, *Business @ the Speed of Thought*, Bill Gates (1999) wrote, "Information technology and business are becoming inextricably interwoven. I don't think anybody can talk meaningfully about one without talking about the other" (p. 6). Gates' point is quite true when one talks about data mining in franchise organizations. Despite its popularity as a global e-business growth strategy, there is no guarantee that the franchising business model will render continuous success in the hypercompetitive environment. This can be evidenced from the constant up-and-down ranking of the Franchise 500. Thus, to see how data mining can be "meaningfully" used in franchise organizations, one needs to know how franchising really works. In the next section, we show that (1) building up a good "family" relationship between the franchisor and the franchisee is the real essence of franchising, and (2) proven working knowledge is the foundation of the "family" relationship. We then discuss in the following three sections the process of how to make data mining "meaningful" in franchising. Finally, future trends of data mining in Netchising are briefly described.

FRANCHISING: THE FRANCHISOR/ FRANCHISEE RELATIONSHIP

Franchising is "a business opportunity by which the owner (producer or distributor) of a service or a trademarked product grants exclusive rights to an individual for the local distribution and/or sale of the service or product, and in return receives a payment or royalty and conformance to quality standards. The individual or business granting the business rights is called the *franchisor*, and the individual or business granted the right to operate in accordance with the chosen method to produce or sell the product or service is called the *franchisee*" (Justis & Judd, 2002, pp. 1-3). Developing a good "family" relationship between the franchisor and the franchisee is the key aspect of a successful franchise (Justis & Judd, 2002). Figure 1 describes how such a "family" relationship is built in the franchise community.

In Figure 1, the franchisor is expected to be flexible in dealing with business concerns to expedite the growth process. The learning process is incrementally developed through five stages (Justis & Judd, 2002): (1) beginner—learning how to do it; (2) novice—practicing doing it; (3) advanced—doing it; (4) master—teaching others to do it; and (5) professional—becoming the best that you can be. Once attaining the advanced stages of development, most preceding struggles have been overcome. However, further convoluted and challenging enquiries will arise as the franchise continues expansion. This is especially factual once the system reaches the "professional" stage, where various unpredicted and intricate problems could arise. Bud Hadfield (1995), the founder of Kwik Kopy franchise and the International Center of Entrepreneurial Development, aptly stated, "The more the company grows, the more it will be tested" (p. 156). To capture the learning process, a counterclockwise round arrow surrounding the franchisor is used to depict the increasing intensity of learning as the franchisor continues to grow. To understand how the "family" relationship is developed, one needs to know the five phases of franchisee life cycle (Schreuder, Krige, & Parker, 2000): (1)

Figure 1. Understanding how the franchisor/franchisee "family" relationship works



Courting—both the franchisee and the franchisor are eager with the relationship; (2) "we"—the relationship starts to deteriorate, but the franchisee still values the relationship; (3) "me"—the franchisee starts to question the reasons for payments related issues with the attitude that the success so far is purely of his/her own work; (4) rebel—the franchisee starts to challenge the restrictions being placed upon; and (5) renewal—the franchisee realizes the "win-win" solution is to continue teaming up with the franchisor to grow the system. Similar to the franchisor, a counter-clockwise round arrow surrounding the franchisee is used in Figure 1 to depict the increasing intensity of franchisee life cycle as the franchisee continues learning and growing.

As the franchisee progresses through the life cycle, the "family" relationship gradually develops an influencing process (Justis & Vincent, 2001), as depicted in Figure 1 with a bi-directional arrow: (1) working knowledge, proven abilities of expanding the franchise system profitably; (2) positive attitude, constructive ways of presenting and sharing the working knowledge; (3) good motivation, providing incentives for learning or teaching the working knowledge; (4) positive individual behavior, understanding and leveraging the strengths of the participants to learn and enhance the working knowledge; and (5) collaborative group behavior, having the team spirit to find the best way to collect, dissimilate, and manage the hard-earned working knowledge. By going through the processes of learning and influencing, both the franchisor and the franchisee gain the progressive working knowledge in the franchise community. The franchisor, the franchisee, and the franchise community in Figure 1 are surrounded with dashed lines, indicating that there is no limit to the learning process.

MANAGING FRANCHISE ORGANIZATIONAL DATA

There are many "touchpoints" within the franchise community where the franchisor and the franchisee can influence each other. Based on the customer service life cycle (CSLC) model, Chen, Chong, and Justis (2002) proposed a framework (Table 1) to harness the Internet to serve the customers for the franchising industry. The 11 sub-stages in Table 1 are based on two well-known franchising books by Justis and Judd (2002) and Thomas and Seid (2000). The model in Table 1 may be used as a comprehensive guide for a franchise to develop its Web site, especially at the stages of Requirements and Acquisition.

Table 1 also is a comprehensive framework for a franchise to model the data needed to serve its customers, that is, franchisees and their customers. A well-designed Internet strategy, often enabled by application service providers (Chen, Ford, Justis, & Chong, 2001), shall empower the franchisor and the franchisees to collect, use, renew, store, retrieve, transmit, and share the organizational data needed to do the collaborative work in the various phases of the CSLC model. Specifically, three types of data are needed:

- **Operational Data:** The daily activities at (1) the franchisor headquarters, including six major entity types: employees; business outlets owned by franchisees or companies; prospective franchisees; product development; suppliers (e.g., marketing agents, accountants, insurance providers, attorneys, and real estate agents); and government offices (e.g., taxes and worker compensation); and (2) the franchisee business outlet, including six major entity types: customers, employees, contacts with the headquarters, product inventory, suppliers, and government offices.

Table 1. The customer service life cycle model in franchising

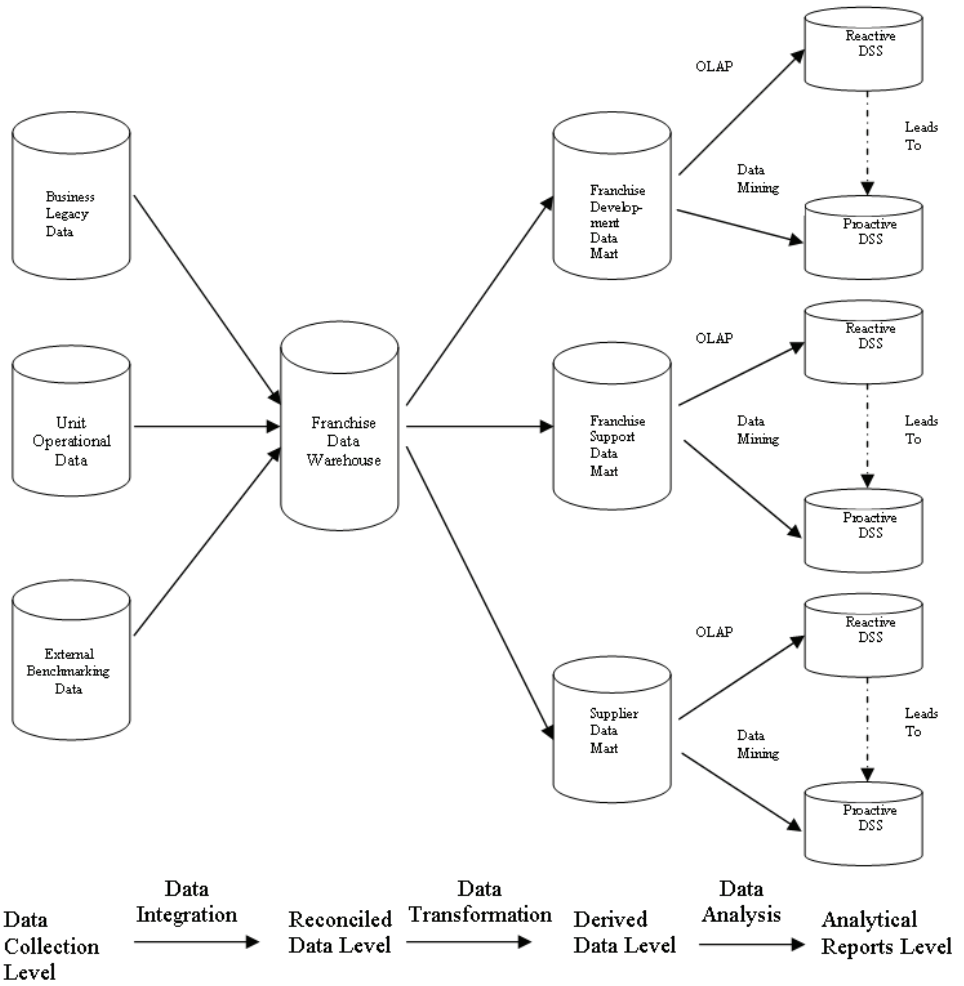
CSLC	Sub-stages	Example: Technology Strategies of WSI Internet (www.wsicorporate.com)
Requirements	Understanding How Franchising Works	Internet
	Investigating Franchise Opportunities	Internet • Global Gateway • Internet Solutions • Portfolio & Technologies • About Us • Franchise Opportunities • Experts Online • Interactive Online
	Obtaining Franchisee Prospectus	Internet • E-mail
	Making the Choice	Internet
Acquisition	Preparing Business Plan	Internet
	Financing the Franchised Business	Internet
	Signing the Contract	Internet
Ownership	Marketing & Promoting the Franchise Products/Services	Internet/Intranet/Extranet • Need a Website? • Live Call • Employment @ WSI • Hot News • WSI ICE Flash • Message from the President
	Managing the Franchise System	Internet Intranet • Serving Franchisees' Customers Extranet
	Building the Relationship between the Franchisor and the Franchisee • The Courting Phase • The "We"-Phase • The "Me"-Phase • The Rebel Phase • The Renewal Phase	Internet Intranet • Knowledge Centre • Training at Headquarters • Newsletter • Meetings • Toll-free Phone Line Extranet • Purchasing Cooperatives
Renewal or Retirement	Becoming a Professional Multi-unit Franchisee or Retiring from the Franchise System	Internet Intranet Extranet

- **External Data:** The relationship management activities in the franchise community, including three major entity types: the relationship with customers, the relationship with partners and suppliers, and the performance benchmarks in the industry.
- **Legacy Data:** the activities that have been working well or gradually adapted since the franchise system came into existence. Examples include (1) rewarding activities to the top performers among the franchisees; (2) efficient procedural activities for the employees at the headquarters supporting the franchisees; and (3) effective and friendly face-to-face activities for the field representatives to serve the franchisees at their outlets.

MANAGING FRANCHISE ORGANIZATIONAL INFORMATION

An architecture, adapted from Inmon (1996), of data mining in franchise organizations with respect to the franchisor/franchisee relationship management depicted in Figure 1 is shown in Figure 2. The architecture consists of four levels: (1) data collection level, holding operational, external, and legacy data collected from the franchise business environment; (2) reconciled data level, holding data warehouse data that are subject-oriented, integrated, time-variant, and non-volatile (Inmon, 1996); (3) derived data level, containing several data marts (e.g., franchisees, customers, competitors, and suppliers) derived from the data warehouse based on various franchisee/customer-centered segmentations; and (4) the analytical reporting level, producing various relationship performance reports (e.g., business outlet periodical summary, financial, scorecards, and mysterious shopping) for the decision makers using the decision support systems (DSS) for their decision making. To move from the data collection level to the reconciled data level, data integration is needed. It is a very time consuming process that involves the activities such as cleansing, extracting, filtering, conditioning, scrubbing, and loading. To move from the reconciled data level to the derived data level, data transformation is needed, which involves the activities such as replication,

Figure 2. An architecture of data mining in franchise organizations



propagation, summary, aggregate, and metadata. To move from the derived data level to the analytical reporting level, data analysis is needed, which involves two major activities: online analytical processing (OLAP) and data mining.

A typical OLAP analysis consists of pre-defined multi-dimensional queries. Some examples in franchising include:

- Show the gross margin by product category and by franchise outlets from Thanksgiving to Christmas in the last five years.
- Which franchise outlets are increasing in sales and which are decreasing?
- Which kinds of customers place the same orders on a regular basis at certain franchise outlets?
- How many franchisees did we lose during the last quarter of 2001, compared to 2000, 1999, and 1998?

Other OLAP activities include spreadsheet analysis, data visualization, and a variety of statistical data modeling methods. Since the query activities are pre-defined, we call the supporting systems reactive DSS.

Data mining, on the other hand, is used to identify hidden relationship patterns of the data residing in the data marts. Typical data mining modeling analysis can be classified into the following three categories:

- **Classification and Prediction:** Using techniques such as RFM (recency, frequency, and monetary), regression, decision tree, and neural network;
- **Association Rules:** Using techniques such as market basket analysis, correlation analysis, cross-sell analysis, and link analysis;
- **Cluster Analysis:** Using techniques such as partition, hierarchy, outlier, and density analysis.

Table 2, adapted from Delmater and Hancock (2001), shows that data mining techniques can be used to help serve franchisees' customers at the different stages of the CSLC model.

Since the data mining queries and related activities are not pre-defined, we call the supporting systems proactive DSS. A major drawback of proactive data mining is the fact that without vigilant preliminary examination of data characteristics, the mining activities may end in vain (Delmater & Hancock, 2001). In order to achieve higher success rate of data mining, we suggest (on the right side of Figure 2) that OLAP-based queries need to be conducted first. For example, one may find, through daily OLAP queries, that certain segments of customers buy certain products frequently. This pattern may lead to perform thorough and proactive analysis of the customer-product relationship. The results may help the company provide legendary service to its clients and generate higher profits.

MANAGING FRANCHISE ORGANIZATIONAL KNOWLEDGE

As mentioned in the discussions of Figure 1, the key for building the franchisor/franchisee "family" relationship is in the franchise organizational learning. In addition, there are five vital factors for a successful learning program: knowledge, attitude, motivation, individual behavior, and group behavior. Thus, working knowledge is the real foundation of a successful franchise "family" relationship. The working knowledge is structured in many forms of profiles that are embedded in the operational manuals of the franchise business processes. Table 3 gives some examples of those working knowledge profiles with respect to the CSLC business processes associated with the sub-stages in Table 1.

A working knowledge profile is developed when a certain task of the CSLC process is repeated several times with superior results. Consider the site profile used at the "marketing & promoting the franchise products/services" sub-stage in Table 3. The site profile is used to assist the new franchisee with locating a high-quality business site. Typically it is the real estate department at the franchisor headquarters that is responsible for the profile development. The site profile is unremittingly being tested and enhanced. Various OLAP/data mining analytical reports, monitoring the performance of the sites, are generated at the Analytical reports level shown in Figure 2. Based on those reports, the real estate experts and their teams are able to fine-tune the attributes and the parameters within the site profile. Most often, the corresponding data collection procedures in the CSLC sub-stage also need to be revised and perfected so that better report scorecards can be generated.

This process of enhancing the working knowledge profile will achieve its high peak when both the franchisor and the franchisees are arriving at the professional and renewal stage

Table 2. Franchisees' customers data mining using the CSLC approach

CSLC	Explanation	Data Mining Activities (and Techniques Used)
Requirements	Finding and reaching the customers	• Lead Generation • Market Analysis & Segmentation (Classification and Prediction) • Mining Web Site Visitors (Association Rules) • Text Mining Usenet Newsgroups (Cluster Analysis)
Acquisition	Selling to the customers	• Customer Acquisition Profiling • Customer Segmentation Strategy (Classification and Prediction) • Online Shopping Tracking (Association Rules) • Pricing Strategy (Association Rules) • Customer-centric Selling (Association Rules) • Text Mining Contact E-Mails (Cluster Analysis) • Scenario Notification (Association Rules)
Ownership	Satisfying the customers after the sales	• Customer Service • Inquiry Routing (Association Rules) • Text Mining E-Mails & Inquiries (Cluster Analysis) • Scenario Notification (Association Rules) • Staffing Level Prediction (Classification and Prediction)
Retirement or Renewing	Retaining the customers so that you can continue coming back	• Customer Retention • Sharper Customer Focus through Loyalty Program (Classification and Prediction) • Detecting Customer Complaints through Text Mining (Cluster Analysis) • Detecting Inappropriate Customer Services (Cluster Analysis) • Individual Customer Profiles (Classification and Prediction) • Scenario Notification (Association Rules)

Table 3. The CSLC model of franchise working knowledge

CSLC Sub-stages	Examples of Working Knowledge Profiles
Understanding How Franchising Works	• Lead Generation Profile • Website Visitor Profile
Investigating Franchise Opportunities	• Benchmark Profile • Successful Franchisee Profile
Obtaining Franchisee Prospectus	• Prospectus Profile
Making the Choice	• Competitor Profile
Preparing Business Plan	• Business Plan Profile
Financing the Franchised Business	• Financing Institute Profile • Non-traditional Franchising Profile
Signing the Contract	• Franchisee Profile
Marketing & Promoting the Franchise Products/Services	• Site Profile • Customer Profile • Product Profile
Managing the Franchise System	• Support Team Profile • Employee Profile • Supplier Profile
Building the Relationship between the Franchisor and the Franchisee	• Event Management Profile • Best Practices Profile • Crisis Management Profile
Becoming a Professional Franchisee or Retiring from the Franchise System	• Multi-unit Franchisee Profile • Co-branding Profile • Opportunities Profile • Social Network Profile

Figure 3. Working knwoledge repository in franchise organizations

		User Skill Levels				
		Beginner in the Courting Phase: Beginner Guide	Novice in the "We"-Phase: Practicing	Advanced in the "Me"-Phase: Doing	Master in the Rebel Phase: Teaching Others	Professional in the Renewal Stage: Improving and Leveraging
Working Knowledge Levels	Collaborative Team	Process of Influencing Others for Knowledge Sharing: Knowledge, Attitude, Motivation, Individual Behavior, and Group Behavior				
	Franchisee Outlet	Working Knowledge Profiles for Running the Franchisee Outlet: Customer Profile, Employee Profile, Product Profile				
	Franchisor Headquarters	Working Knowledge Profiles for Running the Franchisor Headquarters: Franchisee Profile, Site Profile, Product Profile, Employee Profile, Event Management Profile				
	Franchise Community	Working Knowledge Profiles for Relationship Management with the Community: Supplier Profiles, Community Profiles				

of growth. A significant phenomenon of being a professional franchisor and a renewal franchisee are their ability to leverage the assets of the hard-earned working knowledge profiles into dynamic capabilities and high-business-value-creation completive-advantage strategies (Chen, Seidman, & Justis, 2005; Chen, Yuan, & Dai, 2004). The new products or services coming out of the process of leveraging the working knowledge profiles may transform the franchise business into a more, sometimes surprisingly, profitable enterprise. The capability of leveraging the assets of franchise working

knowledge into profitable products or services is at the heart of a successful franchise.

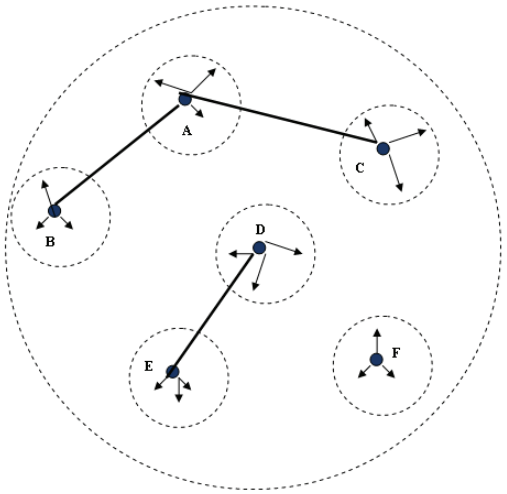
For instance, consider the site selection working knowledge at McDonald's. The Franchise Realty Corporation real estate business, a result of site selection asset leveraging, is the real moneymaking engine at McDonald's. This as can be evidenced from the following speech of Ray Kroc, founder of McDonald's, to the MBA class at the University of Texas at Austin in 1974: "Ladies and gentlemen, I'm not in the hamburger business. My business is real estate" (Kiyosaki, 2000, p. 85). In the book McDonald's: Behind the Arches (Love, 1995, p. 152), Ray Kroc commented further, "What converted McDonald's into a money machine had nothing to do with Ray Kroc or the McDonald brothers or even the popularity of McDonald's hamburgers, French fries, and milk shakes. Rather, McDonald's made its money on real estate…" McDonald's makes money out of real estate by leasing properties from landlords and then subleasing the stores to the franchisees. The professional franchisees, many of which are multiunit operators, can then focus on expending the business without worrying about finding good locations for the growth. This moneymaking real estate strategy is what separates McDonald's from other fast-food chains (David, 2003).

Knowledge repository systems, consisting of working knowledge profiles such as the one shown in Figure 3 (Chen, Hammerstein, & Justis, 2002), can be linked into the franchisor headquarters and the franchisee outlets for knowledge sharing and learning. Such a repository has two dimensions. First, there is a working knowledge level for the collaborative team, the franchisee outlet, the franchisor headquarters, and the franchise community. Second, there are user skill levels, including beginner in the courting Phase, novice in the "we"-phase, advanced in the "me"-Phase, master in the rebel phase (since the rebel ones tend to be those who know the system very well and are capable of influencing others to follow them), and Professional in the renewal stage of franchisee life cycle. The foundation of the framework is the working knowledge of the five crucial elements—knowledge, attitude, motivation, individual behavior, and group behavior—used by the collaborative team, to effectively influence others in building the franchise "family" relationship. The working knowledge profiles at the franchisee outlet, the franchisor headquarters, and the franchise community can be modularized according to user's level. An intranet-based curriculum of working knowledge modules can then be designed for the users to learn the working knowledge profiles effectively.

FUTURE TRENDS

The third industrial revolution, combining Internet technology with globalization, produces various new data mining opportunities for the growth of franchise organizations. For example, knowledge network applications, using data mining techniques such as social network analysis, can be developed to connect professional franchisees in the world. The goal is to enable the franchise system to venture into new global emerging markets, for example, China, through international franchising and develop innovative products/services through asset leveraging. This could be done because franchise capabilities, structured in the working knowledge repository shown in Figure 3, enable the professional franchisees to work with the franchisor to continuously improve and leverage the current franchise working knowledge. An example of knowledge networks of professional franchisees can be illustrated in Figure 4. There are six professional franchisees (A-F) in the figure with three clusters (A-C, D-E, and F) of knowledge networks. Each professional franchisee (a dot) has his/her personal knowledge network (arrows pointing out of the dot) tested and built over the years while doing day-to-day problem solving at the franchisee outlet. The knowledge network may include the customers' likes and dislikes, the kind of employees to hire, the competitors' and suppliers' pricing strategies, and the social needs in the local community. Each professional franchisee is surrounded with a circle with dashed lines, meaning there is no limit to the personal knowledge network. In order to solve the problems more effectively, professional franchisees may share with each other their approaches. Thus, clusters (connected dots) of knowledge network are formed for solving various problems more effectively (Chen, Justis, & Wu, 2006).

Figure 4. Knowledge networks of professional franchisees



CONCLUSIONS

Franchising has been popular as a growth strategy for small businesses; it is even more so in today's global and e-commerce world (Chen, Chen, & Wu, 2005). The essence

of franchising lies in managing the "family" relationship between the franchisor and the franchisee. In this article we showed that data mining plays an important role in growing and nurturing such a "family" relationship. Specifically, we discussed: (1) how franchise organizational data can be managed effectively using the methodology of customer service life cycle; (2) how franchise organizational information is deciphered from the customer-centered data using OLAP and data mining analytical techniques; and (3) how the franchise organizational knowledge is leveraged to grow the franchise system. The ability to continue leveraging the organizational knowledge assets based on the good "family" relationship is really what a franchise business is about.

REFERENCES

Chen, Y., Chen, G., & Wu, S. (2005). Issues and opportunities in e-business research: A Simonian perspective. *International Journal of E-Business Research, 1*(1), 37-53.

Chen, Y., Chen, G., & Wu, S. (2006). A Simonian approach to e-business research: A study in Netchising. In I. Lee (Ed.), *Advanced Topics in E-Business Research: E-Business Innovation and Process Management* (Vol. 1, pp. 133-161). Hershey, PA: Idea Group Publishing.

Chen, Y., Chong, P. P., & Justis, R. T. (2002). E-business strategy in franchising: A customer-service-life-cycle approach. In *Proceedings of the 16th Annual International Society of Franchising Conference*, Orlando, FL.

Chen, Y., Ford, C., Justis, R. T., & Chong, P. (2001). Application service providers (ASP) in franchising: Opportunities and issues. In *Proceedings of the 15th Annual International Society of Franchising Conference*, Las Vegas, NV.

Chen, Y., Hammerstein, S., & Justis, R. T. (2002). Knowledge, learning, and capabilities in franchise organizations. In *Proceedings of the Third European Conference on Organizational Knowledge, Learning, and Capabilities*, Athens, Greece.

Chen, Y., Justis, R., & Wu, S. (2006). Value networks in franchise organizations: A study in the senior care industry. In *Proceedings of the 20th Annual International Society of Franchising Conference*, Palm Springs, CA.

Chen, Y., Justis, R. T., & Yang, H. L. (2004). Global e-business, international franchising, and theory of Netchising: A research alliance of east and west. In *Proceedings of the 18th Annual International Society of Franchising Conference*, Las Vegas, NV.

Chen, Y., Seidman, W., & Justis, R. (2005). Strategy and docility in franchise organizations. In *Proceedings of the 19th Annual International Society of Franchising Conference*, London, UK.

Chen, Y., & Wu, S. (2006). E-business research in franchising (invited editorial preface). *International Journal of E-Business Research, 2*(4), i-ix.

Chen, Y., Yuan, W., & Dai, W. (2004, December 12-15). Strategy and nearly decomposable systems: A study in franchise organizations. In *International Symposium on "IT/IS Issues in Asia-Pacific Region, co-sponsored by ICIS-2004*, Washington, DC.

Davenport, T. (2000). E-commerce goes global. *CIO Magazine, 13*(20), 52-54.

David, G. (2003, March 30). Can McDonald's cook again? *Fortune Magazine*.

Delmater, R., & Hancock, M. (2001). *Data mining explained: A manager's guide to customer-centric business intelligence*. New York: Digital Press.

Gates, W. (1999). *Business @ the speed of thought*. New York: Warner Books.

Hadfield, B. (1995). *Wealth within reach*. Houston, TX: Cypress Publishing.

Inmon, W. H. (1996). *Building the data warehouse*. New York: John Wiley & Sons.

Justis, R. T., & Judd, R. J. (2002). *Franchising*. Houston, TX: DAME Publishing.

Justis, R. T., & Vincent, W. S. (2001). *Achieving wealth through franchising*. Holbrook, MA: Adams Media Corporation.

Kiyosaki, R. (2000). *Rich dad, poor dad*. New York: Time Warner.

Love, J. (1995). *McDonald's: Behind the arches*. New York: Bantam Books.

Schreuder, A. N., Krige, L., & Parker, E. (2000, February 19-20). The franchisee lifecycle concept—A new paradigm in managing the franchisee-franchisor relationship. In *Proceedings of the 14th Annual International Society of Franchising Conference*, San Diego, CA.

Thomas, D., & Seid, M. (2000). *Franchising for dummies*. New York: IDG Books.

KEY TERMS

Customer Service Life Cycle (CSLC): Serving customers based on a process of four stages: requirements, acquisition, ownership, and retirement. Many companies are using the approach to harness the Internet to serve the customers.

Data Mart: A small database with data derived from a data warehouse.

Data Mining: Analytical techniques used to find out the hidden relationships or patterns residing in the organizational data.

Data Warehouse: A database that is subject-oriented, integrated, time-variant, and non-volatile.

Franchisee: The individual or business that receives the business rights and pays the royalties for using the rights.

Franchisee Life Cycle: The stages a franchisee goes through in the franchise system: courting, "we," "me," rebel, renewal.

Franchising: A business opportunity based on granting the business rights and collecting royalties in return.

Franchisor: The individual or business that grants the business rights.

Franchisor/Franchisee Learning Process: The stages of learning, including Beginner, Novice, Advanced, Master, and Professional.

Franchisor/Franchisee Relationship Management: The vital factor for the success of a franchise, including knowledge, attitude, motivation, individual behavior, and group behavior.

Data Mining in Tourism

Indranil Bose
The University of Hong Kong, Hong Kong

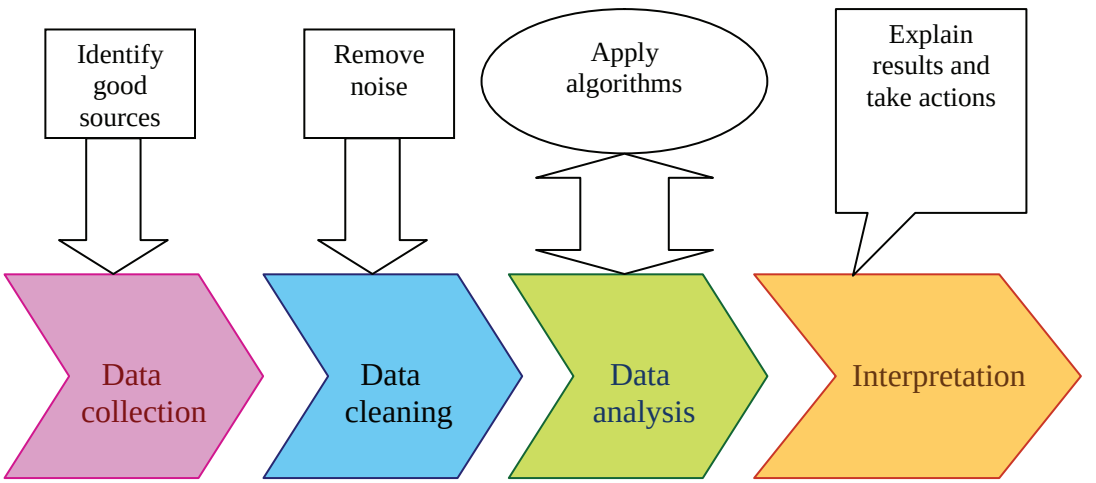
INTRODUCTION

Everyday, millions of people travel around the globe for business, vacations, sightseeing, or other reasons. An astronomical amount of money is spent on tickets, accommodations, food, transportation, and entertainment. According to World Travel and Tourism Council, travel and tourism represents approximately 11% of the worldwide gross domestic product (GDP) (Werthner & Ricci, 2004). Tourism is an information-based business where there are two types of information flow. One flow of information is from the providers to the consumers or tourists. This is information about goods that tourists consume such as tickets, hotel rooms, entertainments, and so forth. The other flow of information which follows a reverse direction consists of aggregate information about tourists to service providers. In this chapter we will discuss the second form of information flow about the behavior of tourists. When the aggregated data about the tourists is presented in the right way, analyzed by the correct algorithm, and put into the right hands, it could be translated into meaningful information for making vital decisions by tourism service providers to boost revenue and profits. Data mining can be a very useful tool for analyzing tourism-related data.

BACKGROUND

According to Tan, Steinbach, and Kumar (2006), "Data mining is the process of automatically discovering useful information in large data repositories" (p. 2). It uses machine learning and statistical and visualization techniques to discover and present knowledge in a form that is easily comprehensible to humans. Data mining involves four key steps: (1) data collection, (2) data cleaning, (3) data analysis, and (4) interpretation and evaluation. During data collection the most suitable data need to be collected from the most appropriate sources. There is often a need to consolidate data from a number of sources. Cleaning or cleansing is the process of ensuring that all values in a data set are consistent and correctly recorded (Hui, Pandey, Steinbach, & Kumar, 2006). Obvious data errors are detected and corrected, and missing data is replaced in this step. The third and the most important stage of data mining is the analysis of the data using known techniques. Usually the analysis is done using statistical- or machine-learning-based approaches. The choice of the technique depends on the type of problem and also the availability of appropriate data mining software (Bose & Mahapatra, 2001). The most difficult part in data mining

Figure 1. Different steps in data mining



is interpretation of results obtained during data analysis. The analysis results may show a high degree of accuracy, but unless the accuracy can be related to the context of the data mining problem, it is of no use. Once some meaningful interpretation of the data analysis results can be done, the final step is to take action(s) so that the gathered knowledge can be put into practical use. In the case of tourism data mining, the data mining process goes through the same four steps. Figure 1 identifies the four steps involved in data mining and is similar to the one used in Bose and Pal (in press).

TOURISM DATA MINING

In this section we discuss the different types of machine learning techniques and explain how they have been used for analyzing data related to tourism. Usually two types of machine learning activities are common in tourism—association learning and classification learning. In association learning, the learning method searches for associations or relationships between features of tourist behavior. For example, the algorithm may try to find out if tourists who are interested in shopping also prefer to stay near the center of a city. That is, there is no specific target variable in this type of data mining, and so this is popularly known as unsupervised learning. A second style of machine learning is classification learning. This learning scheme takes a set of classified examples from which it discovers a way of classifying unseen examples. This is a form of supervised learning, in which there is a specific target variable**.** For example, by using classification analysts may be interested to classify tourists into two groups—high spenders and low spenders for luxury items. In this case the target variable is expenditure on luxury items. Based on a set of demographic and other variables the classification algorithm will establish the specific attributes of a tourist that qualify them as a high spender or a low spender. Next, we describe the various machine learning techniques used in tourism data mining.

Artificial neural networks (ANN). ANNs are nonlinear predictive models that learn through training (Jain, Mao, & Mohinuddin, 1996). They are composed of interconnected neurons. Each neuron receives a set of inputs. Each input is multiplied by a weight. The sum of all weighted inputs determines the activation level. A very powerful algorithm that is used in training ANNs is called *backpropagation*. Here, weights of the connection are iteratively adjusted to minimize the error based on the difference between desired and actual outputs.

Clustering. This is the process of dividing objects into groups whose members are similar in some way(s) (Han, Kamber, & Tung, 2001). Although there are many clustering algorithms, the commonly used ones are exclusive clustering and distance-based clustering. In an exclusive clustering algorithm if a certain datum belongs to a definite cluster then it cannot be included in another cluster. In distance-based clustering, if two or more objects are "close" according to a given distance they are grouped into the same cluster. Self-organizing feature map (SOFM) is a special type of an ANN-based algorithm that performs clustering.

Rough sets (RS). RS theory is proposed to address the problem of uncertainty and vagueness in the classification of objects (Slowinski & Vanderpooten, 2000). It is founded on the hypothesis that every object is associated with some information, and objects that are associated with the same information are similar and belong to the same class. The first step of RS is discretization of independent attributes where numeric attributes are converted to categorical attributes. The second step is formation of reducts that provide same quality of classification as the original set of attributes. The last step is classification of unknown data based on decision rules and reducts.

Support vector machines (SVM). SVM classifies an input vector into known output classes. It starts with several data points from two classes and obtains the optimal hyperplane that maximizes the separation of the two classes. For nonlinearly separable data, it uses the kernel method to transform the input space into a high dimensional feature space, where an optimal linearly separable hyperplane can be constructed. Examples of kernel functions are linear function, polynomial function, radial basis function, and sigmoid function (Chang & Lin, 2001).

Use of Data Mining in Tourism

Tourism policy makers, retail business executives, directors of scenic spot management companies, and government organizations want to know the relationship between tourism activities and preferences of tourists so that they can plan for required tourism infrastructures, such as accommodation sites and transportation. They also need detailed analysis to help them make operational, tactical, and strategic decisions. Examples of these include scheduling and staffing, preparing tour brochures, and investments. Due to the need for this analysis, formal statistical techniques were introduced in tourism. However, statistical techniques suffer from the drawback that several assumptions about distributions of data have to be made before any analysis can be conducted. If these assumptions are violated there is no guarantee that the results will be valid. This limitation of statistical methods has prompted researchers to use machine-learning-based data mining for tourism data analysis. The three main uses of data mining techniques in the tourism industry are: (1) forecasting expenditures of tourists, (2) analyzing profiles of tourists, and (3) forecasting number of tourist arrivals. In the following sections examples are presented to demonstrate how data mining techniques are used to support these activities.

Forecasting Tourist Expenditures

ANNs with different architectures were built to forecast the tourist expenditures in the Balearic Islands by Palmer, Montano, and Sese (in press). The performances of different architectures were contrasted by comparing the mean absolute percentage error (MAPE) incurred, and it was shown that the forecasting accuracy of ANNs varied with the architecture. In another study that predicted shopping expenditures by tourists visiting Hong Kong, RS was used (Law & Au, 2000). Using nine explanatory variables and four levels in the dependant variable, the researchers devised 15 decision rules that could forecast shopping expenditures of tourists with 94% accuracy. The use of RS for predicting tourist expenditures in dining in Hong Kong was reported by Au and Law (2002). While using 10 explanatory variables and 18 induced decision rules, the researchers reported 83% accuracy in predicting dining expenditures of tourists.

Analyzing Profiles of Tourists

To forecast the profiles of tourists in Cape Town, South Africa, ANN was used by Bloom (2005). The researcher used survey data collected from 694 respondents who visited Cape Town in 2000-2001. The SOFM method was used for categorizing tourists into three groups: (1) vibrant and energetic (39.8%), (2) established and settled (34.3%), and (3) pleasure seekers (25.9%). SOFM was also used to identify segments in the market of West Australian senior tourists (Kim, Wei, & Ruys, 2003). Using three factors: (1) demographic, (2) travel motivation, and (3) concerns, the respondents were clustered into four groups: (1) active learners, (2) relaxed family body, (3) careful participant, and (4) elementary vacationer. Tourism development at any given place is often dependant on whether the local residents are in favor of it or not (Pérez & Nadal, 2005). A survey was conducted among 791 residents in the Balearic Islands, and clustering was used to categorize the local residents into (1) developmental supporters (11%), (2) prudent developers (26%), (3) ambivalent and cautious (24%), and (4) protectionists (20%).

Forecasting Arrivals of Tourists

ANNs have been a popular choice in forecasting tourist arrivals to a location. Backpropagation-based ANN has been used to forecast tourist arrivals from the United States to Durban, South Africa based on available data from 1992-1998 (Burger, Dohnal, Kathrada, & Law, 2001). ANNs were also used to predict the number of arrivals of Japanese tourists to Hong Kong using data on six explanatory variables, collected from a variety of sources over the period 1967-1996 (Law & Au, 1999). The ANN model yielded the lowest MAPE of 10.59% and performed significantly better than statistical models. The same study was repeated for forecasting Taiwanese tourist arrivals to Hong Kong using data from the same time period and the same variables (Law, 2000). Again, ANN yielded the best forecasting result with an MAPE of 2.76%. In a study that extended the previous studies by considering tourist arrivals from six different countries to Hong Kong (i.e., USA, Japan, Taiwan, Korea, UK, and Singapore) ANN again proved its supremacy over statistical methods. RS has also been used for travel demand analysis (Goh & Law, 2003). In this study, using data collected between 1985-2000 and with visitors to Hong Kong from 10 countries, the RS model was able to predict increases or decreases in the number of tourist arrivals with an accuracy of 87.2%. It was also found that volume of trade and GDPs were the most important predictor variables. Recently, the method of SVM has been used for predicting tourist arrivals to Barbados and has yielded a lower MAPE when compared to several statistical models (Pai & Hong, 2005).

Assessment of Data Mining for Tourism

In this chapter three specific applications of tourism data mining have been discussed. Other application areas of data mining include tour path planning using genetic algorithms (Yao, Huang, & Lee, 2003), where the itinerary of visiting multiple attractions is modeled as a traveling salesman person. Another novel application includes the use of ANN for identifying activities pursued by tourists at a popular beach in Australia (Green, Blumenstein, Browne, & Tomlinson, 2005). This is done by using ANN on images of beach scenes and segmenting these images to identify tourists in the scenes.

In general, it can be said that ANN is the most popular data mining technique in tourism. This is possibly because of the fact that ANNs can provide high performance in terms of accuracy for tourism data and do not seem to be affected much by the presence of noise. Clustering using SOFM is also a preferred technique for tourism data analysis due to its simplicity of operation. Another reason for the popularity of these two techniques is that prepackaged software that uses these techniques is readily available on the market and can be used for data analysis with very little training.

FUTURE TRENDS

From the previous discussion we can see that various data mining techniques have been used successfully for tourism-related data analysis in recent times. Most of the techniques have been employed using prepackaged software. Researchers have not devoted much effort to developing data mining software specifically dealing with tourism. There are some techniques such as Bayesian belief networks, case-based reasoning, and decision trees that have not been used much.

This may be due to the lack of availability of prepackaged software that uses these techniques. In the future, we expect to see tourism data mining using these techniques.

The other approach which we may see in the future is that of ensemble data mining in which more than one method is used for analyzing the data. When an ANN is used for forecasting tourist expenditures in a certain country, the performance may not be very good due to overtraining. However, if clustering is carried out before the ANN forecast, the investigation can be much more meaningful because the data is grouped and preprocessed to an extent. There may be opportunities to combine ANN with RS or SVM with clustering in a similar fashion.

CONCLUSION

It is only recently that tourism data mining has caught the attention of researchers and definitely much remains to be done. Only three main tourism data mining application areas have been identified in this paper. It is likely that more areas will come up in the future such as tourism recommendation analysis, tourism promotion response analysis, and so forth. Data mining holds much promise for improving these activities. Finally, only a limited number of data mining methods have been used so far. There are opportunities to use other popular data mining methods as well as ensemble data mining methods for analysis of tourism data.

REFERENCES

Au, N., & Law, R. (2002). Categorical classification of tourism dining. *Annals of Tourism Research, 29*(3), 819-833.

Bloom, J. Z. (2005). Market segmentation—A neural network application. *Annals of Tourism Research, 32*(1), 93-111.

Bose, I., & Mahapatra, R. K. (2001). Business data mining—A machine learning perspective. *Information & Management, 39*(3), 211-225.

Bose, I., & Pal, R. (in press). Predicting the survival or failure of click-and-mortar corporations: A knowledge discovery approach. *European Journal of Operational Research*.

Burger, C. J. S. C., Dohnal, M., Kathrada, M., & Law, R. (2001, August). A practitioners guide to time-series methods for tourism demand forecasting—A case study of Durban, South Africa. *Tourism Management, 22*(4), 403-409.

Chang, C.-C., & Lin, C.-J. (2001). LIBSVM: A library for support vector machines. Retrieved from http://www.csie.ntu.edu.tw/~cjlin/libsvm

Goh, C., & Law, R. (2003, October). Incorporating the rough sets theory into travel demand analysis. *Tourism Management, 24*(5), 511-517.

Green, S., Blumenstein, M., Browne, M., & Tomlinson, R. (2005). The detection and quantification of persons in cluttered beach scenes using neural network-based classification. In *Proceedings of the 6th International Conference on Computational Intelligence and Multimedia Applications,* 16-18 August, 303-308.

Han, J., Kamber, M., & Tung, A. (2001). Spatial clustering methods in data mining: A survey. In H. Miller & J. Han (Eds.), *Geographic data mining and knowledge discovery.* Taylor & Francis.

Hui, X., Pandey, G., Steinbach, M., & Kumar, V. (2006). Enhancing data analysis with noise removal. *IEEE Transaction son Knowledge and Data Engineering, 18*(3), 304-319.

Jain, A. K., Mao, J., & Mohinuddin, K. M. (1996, March). Artificial neural networks: A tutorial. *IEEE Computer, 29*(3), 31-44.

Kim, J., Wei, S., & Ruys, H. (2003, February). Segmenting the market of West Australian senior tourists using an artificial neural network. *Tourism Management, 24*(1), 25-34.

Law, R. (2000, August). Back-propagation learning in improving the accuracy of neural network-based tourism demand forecasting. *Tourism Management, 21*(4), 331-340.

Law, R., & Au, N. (1999, February). A neural network model to forecast Japanese demand for travel to Hong Kong. *Tourism Management, 20*(1), 89-97.

Law, R., & Au, N. (2000, June). Relationship modeling in tourism shopping: A decision rules induction approach. *Tourism Management, 21*(3), 241-249.

Pai, P.-F., & Hong, W.-C. (2005, October). An improved neural network model in forecasting arrivals. *Annals of Tourism Research, 32*(4), 1138-1141.

Palmer, A., Montano, J. J., & Sese, A. (in press). Designing an artificial neural network for forecasting tourism time series. *Tourism Management.*

Pérez, E. A., & Nadal, J. R. (2005, October). Host community perceptions a cluster analysis. *Annals of Tourism Research, 32*(4), 925-941.

Slowinski, R., & Vanderpooten, D. (2000, March/April). A generalized definition of rough approximations based on uncertainty. *IEEE Transactions on Knowledge and Data Engineering, 12*(2), 331-336.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining.* Boston, MA: Addison-Wesley.

Werthner, H., & Ricci, F. (2004, December). E-commerce and tourism. *Communications of the ACM, 47*(12), 101-105.

Yao, L., Huang, B., & Lee, D.-H. (2003). A criteria-based approach for selecting touring paths using GIS and GA. In *Proceedings of the IEEE Conference on Intelligent Transportation Systems,* 12-15 October, 2, 958-963.

KEY TERMS

Artificial Neural Networks (ANN): ANN is a pattern matching technique that uses training data to build a model and uses the model to predict unknown samples. It consists of input, output, and hidden nodes and connections between nodes. The weights of the connections are iteratively adjusted in order to get an accurate model.

Clustering: It is a technique that classifies instances into classes by calculating the distance between them and other instances. The instances that have the least distance between them are grouped into the same class. It is a type of unsupervised learning.

Decision Tree: It is technique for classifying data. The root node of a decision tree represents all examples. If these examples belong to two or more classes, then the most discriminating attribute is selected and the set is split into multiple classes.

Genetic Algorithms: These algorithms mimic the process of natural evolution and perform explorative search. The main component of this method is chromosomes that represent solutions to the problem. It uses selection, crossover, and mutation to obtain chromosomes of highest quality.

Mean Absolute Percentage Error (MAPE): MAPE is used as a figure of merit to identify whether a data mining method is performing well or not. The lower the MAPE, the better the performance of the data mining method.

Rough Sets (RS): RS is a technique used for classification. It is based on the notion of approximate class membership of an unclassified observation. The main steps in RS analysis include discretization, formation of reducts, and enumeration of rules.

Self-Organizing Feature Map (SOFM): SOFM is a data mining method used for unsupervised learning. The architecture consists of an input layer and an output layer. By adjusting the weights of the connections between input and output layer nodes, this method identifies clusters in the data.

Support Vector Machines (SVM): SVM is a data mining method useful for classification problems. It uses training data and kernel functions to build a model that can appropriately predict the class of an unclassified observation.

Tourism Recommendation System: Information technology may be used to make recommendations to tourists about sightseeing venues, restaurants, entertainment locations, shopping, and so forth. These systems will revolutionize future travel by making recommendations using data on demographics and preferences of tourists.

Data Streams as an Element of Modern Decision Support

Damianos Chatziantoniou
Athens University of Economics and Business, Greece

George Doukidis
Athens University of Economics and Business, Greece

INTRODUCTION

Traditional decision support systems (DSS) and executive information systems (EIS) gather and present information from several sources for business purposes. It is an information technology to help the knowledge worker (executive, manager, analyst) make faster and better decisions. So far, these data were stored statically and persistently in a database, typically in a data warehouse. *Data warehouses* collect masses of operational data, allowing analysts to extract information by issuing decision support queries on the otherwise discarded data. In a typical scenario, an organization stores a detailed record of its operations in a database, which is then analyzed to improve efficiency, detect sales opportunities, and so on. Performing complex analysis on these data is an essential component of these organizations' businesses. Chaudhuri and Dayal (1997) present an excellent survey on decision-making and online analytical processing (OLAP) technologies for traditional database systems.

In many applications however, it may not be possible to process queries within a database management system (DBMS). These applications involve data items that arrive online from multiple sources in a continuous, rapid and time-varying fashion (Babcock et. al., 2002). These data may or may not be stored in a database. As a result, a new class of data-intensive applications has recently attracted a lot of attention: applications in which the data is modeled not as persistent relations but rather as transient *data streams*. Examples include financial applications (streams of transactions or ticks), network monitoring (stream of packets), security, telecommunication data management (stream of calls or call packets), web applications (clickstreams), manufacturing, *wireless sensor networks* (measurements), RFID data, and others. In data streams we usually have "continuous" queries (Terry et. al., 1992; Babu & Widom, 2002) rather than "one-time." The answer to a *continuous query* is produced over time, reflecting the stream data seen so far. Answers may be stored and updated as new data arrives or may be produced as data streams themselves. Continuous queries can be used for monitoring, alerting, security, personalization, etc. Data streams can be either *transactional* (i.e., log interactions between entities, such as credit card purchases, web clickstreams, phone calls), or *measurement* (i.e., monitor evolution of entity states, such as physical phenomena, road traffic, temperature, network).

How to best model, express and evaluate complex queries over data streams is an open and difficult problem. This involves data modeling, rich querying capabilities to support real-time decision support and mining, and novel evaluation and optimization processing techniques. In addition, the kind of decision support over data streams is quite different from "traditional" decision-making: decisions are "tactical" rather than "strategic." Research on data streams is currently among the most active areas in database research community. Flexible and efficient stream querying will be a crucial component of any future data management and decision support system (Abiteboul et al., 2005).

BACKGROUND

The database research community has responded with an abundance of ideas, prototypes and architectures to address the new issues involved in data stream management systems (DSMS). STREAM is Stanford University's approach for a general-purpose DSMS (Arasu et. al., 2003); *Telegraph* and *TelegraphCQ* (Madden & Franklin, 2002: Chandrasekaran et. al., 2003) are prototypes focused on handling measurements of sensor networks, developed in Berkeley; *Aurora* is a joint project between Brandeis University, Brown University and MIT (Carney et. al., 2002) targeted towards stream monitoring applications; AT&T's Hancock (Cortes et. al., 2000) and *Gigascope* (Cranor et. al., 2003) projects are special-purpose data stream systems for network management; Tribeca (Sullivan, 1996) and NiagaraCQ (Chen, et. al., 2000) are other well-known projects from Telcordia and University of Wisconsin respectively. The objective of all these projects is to develop systems that can support the challenging analysis requirements of streaming applications.

Furthermore, a plethora of articles, papers and tutorials appeared recently in the research literature. Some of the most well known survey articles follow. Faloutsos (2004) discusses indexing and mining techniques over data streams; Koudas and Srivastava (2003), present the state-of-the-art

algorithms on stream query processing; Muthukrishnan (2003), reviews data stream algorithms and applications; Babcock et. al. (2002), present an excellent survey on data stream prototypes and issues; Garofalakis et. al. (2002), discuss various existing models and mining techniques over data streams.

APPLICATIONS

Stream applications span a wide range of everyday life. Real-time analytics is an essential part of these applications and becomes rapidly more and more critical in decision-making. It is apparent from the list of areas below that efficiently querying and processing data streams is a necessary element of modern decision support systems.

- **Telecommunications:** The telecommunications sector is undoubtedly one of the prime beneficiaries of such data management systems due to the huge amount of data streams that govern voice and data communications over the network infrastructure. Examples of stream analysis include fraud detection, real-time billing, dynamic pricing, network management, traffic monitoring and so on. Streams (calls, packets) have to be mined at real-time to discover outliers and patterns (fraud detection); correlated, joined and aggregated to express complex business rules (billing, dynamic pricing); and monitored—computing averages and min, max values over periods of time—to uncover unusual traffic patterns (network management).
- **Sensors:** Sensor technology becomes extremely widespread and it will probably be the next killer application: large number of cheap, wireless sensors attached to products, cars, computers, even sport players and animals, tracking and digitizing behavior, traffic, location and motion. Examples involve electronic property stickers (super markets, libraries, shopping carts, etc.), vehicle sensors (to electronically pay tolls, route traffic, set speed, etc.), and location-identification sensors (to report location, serve content, detect routes, etc.) A "sensor" world leads to a "stream" world. Millions of input data every few seconds need to be analyzed: aggregate (what is the average traffic speed), correlate (two products sell together), alert (quantity of a product is below a threshold), localize and monitor.
- **Finance:** Financial data streams come in many different forms: stock tickers, news feeds, trades, etc. Financial companies want to analyze these streams at real-time and take "tactical" business decisions (opposed to "strategic" decisions, associated to OLAP or data mining). For example, Charles Schwab wants to compute commission on each trade at real-time; Fidelity would like to route content in trades at real-time. Traderbot (www.traderbot.com) is a Web-based financial search engine that evaluates queries (both traditional and continuous) over real-time streaming data (e.g., "find all stocks between €20 and €200 where the spread between the high tick and the low tick over the past 30 minutes is greater than 3% of the last price and in the last 5 minutes the average volume has surged by more than 300%.")
- **Web management:** Large Web sites monitor Web logs (clickstreams) online to enable applications such as personalization, performance monitoring, and load balancing. Some web sites served by widely distributed web servers (e.g., Yahoo) may need to coordinate many distributed clickstream analyses, e.g. to track heavily accessed Web pages (e.g., CNN, BBC) as part of their real-time performance monitoring (Babcock et al., 2002}.
- **Network management:** Network traffic management systems monitor a variety of continuous data streams at real-time, such as packet traces, packet flows and performance measurements in order to compute statistics, detect anomalies, adjust routing, etc. The volume of data streams can be humongous and thus, query processing must be done very carefully.
- **Military:** One of the most interesting applications in military is battalion monitoring—where sensors are installed on every vehicle, human, etc.—having thousands of sensors reporting state in real-time. In these applications we want to know each time where vehicles and troops are. Examples include queries such as "tell me when three of my four tanks have crossed the front line" and "tell me when someone is pointing a gun at me."

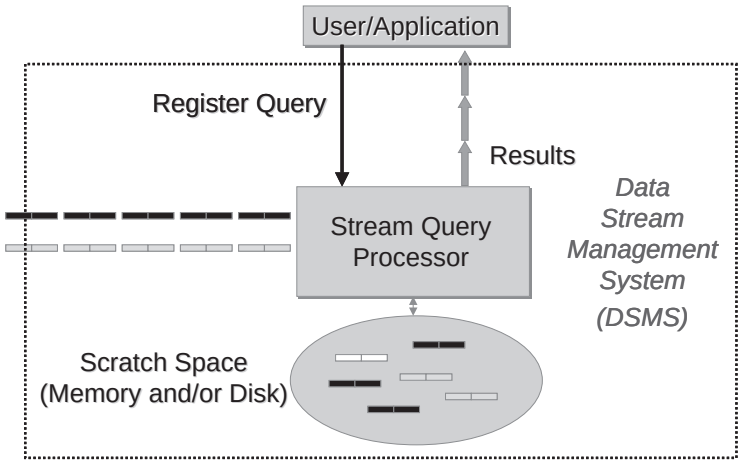
ISSUES AND CHALLENGES

Performing decision-making queries on top of data streams is a major challenge. For example, only one-pass algorithms are allowed (because data can be seen only once) and memory has to be managed very carefully (what to keep and what to discard). The need for a data stream management system comes in two forms: either the volume of data is huge and can not be stored in persistent relations (e.g., packet network traffic)—but still some data analysis has to be carried out, or an answer is required for a report at real-time (e.g., monitoring, alerting, fraud-detection.) As a result, DSMS are quite different in nature from traditional DBMS. Data is transient instead of persistent; queries may be "continuous" instead of "one-time"; processing techniques differ significantly, primarily due to main-memory management requirements. In Table 1 we list the differences between traditional database and data stream management systems (Babcock et al., 2002).

Table 1. Differences betweens DSMS and DBMS

DBMS	DSMS
Persistent relations	Transient streams
One-time queries	Continuous queries
Random access	Sequential access
"Unbounded" disk store	Bounded main memory
Only current state matters	History/arrival-order is critical
Passive repository	Active stores
Relatively low update rate	Possibly multi-GB arrival rate
No real-time services	Real-time requirements
Assume precise data	Data stale/imprecise

Figure 1. A generalized DSMS architecture



A generalized DSMS architecture is shown in Figure 1 (Babcock et al., 2002).

A user or application registers a (continuous) query with the stream system. As new stream data arrives, the answer of the registered query is updated. The stream query processor deals with the incoming data and has to decide on a significant number of important issues: should it shed data to keep up with the rate of the stream (leading to approximate answers)? How to best utilize main-memory space? How to handle multiple registered queries in an optimal way (maximizing resource utilization)? Finally, some scratch space may be required for intermediate computations.

The differences between DSMS and DBMS present several novel challenges for a system handling and analyzing data streams, both in terms of functionality and processing.

In terms of functionality (how to *use* data streams, i.e., reporting):

- Defining complex stream sessions (e.g., a network 'flow' or a finance 'burst')
- Representing nested stream structures (e.g., a sub-session within a session)
- Defining hierarchical statistical stream functions (similar to roll-up, drill-down in OLAP)
- Allowing multiple sources and/or multiple formats, possibly combine with persistent data
- Allowing user-defined functions—UDFs (e.g., a non-traditional financial computation)

In terms of evaluation (how to *process* data streams):

- Handling multiple, continuous, rapid, time-varying, ordered streams
- Having computations only in main-memory

- Queries may be continuous (not just one-time):
  - Have to be evaluated continuously as stream data arrives
  - Answer updated over time
- Queries may be complex
  - Beyond element-at-a-time processing
  - Beyond stream-at-a-time processing
  - Beyond traditional decision-making queries (scientific, data mining)
- Queries may involve nested structures
- Users may select for query answers to be
  - Exact (worst performance, sometimes not possible due to the stream rate)
  - Approximate (requires load shedding, (i.e. discarding data, better performance))

QUERYING AND EVALUATING DATA STREAMS

The challenges mentioned in the previous section lead to a number of major issues in querying and processing data streams. Understanding these aspects is crucial in developing a stream system. Different systems follow different methodologies in processing data streams (as presented in section 0).

Windows

Being able to formulate queries on top of data streams is a major component of any stream system. Most system prototypes extend SQL in some way in order to make it suitable for stream processing. For example, AT&T's Gigascope defines GSQL and Stanford's STREAM proposes CQL. Most stream query languages reference and produce both relations and streams.

Relations, Streams⟶***Stream Query Languages***⟶Relations, Streams

It is important, however, to transform infinite, unbounded streams to finite relations which then can be manipulated by a query processor. Defining *windows of streams* through some window specification syntax is the mechanism to turn (part of) a stream to a relation. For example, one could define a window as the one-thousand most recent stream tuples. Figure 2 shows graphically this concept.

Windows can be based on ordering attributes (e.g., a sliding or shifting window), tuple counts, explicit markers (called punctuations), etc. An application or a query may define more than one window and "join" them together.

Query Evaluation

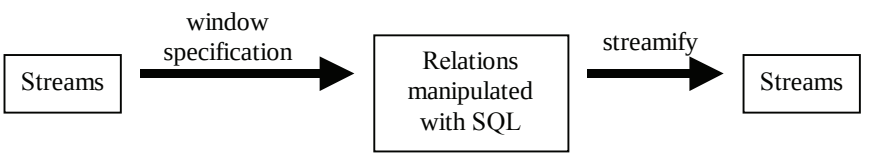
The unique characteristics of data streams lead to different evaluation algorithms and query processing: data elements arrive continuously, possibly at variable arrival rate; data streams are potentially unbounded in size; general retrieval of past elements is very difficult, if not impossible. As justly pointed out by Garofalakis et al. (2002), "you only get one look." This does not exclude the presence of conventional data stored in some traditional database system.

As a result, algorithms operate mainly in main-memory and involve one-pass techniques (or few-passes, if the input data stream can be organized in input blocks). Operators include—besides the traditional ones (selection, projection)—operations specific for combining and aggregating multiple streams, such as sliding window join and merging (Arasu et al., 2006). Optimization is very different than traditional DBMS. The goal in DSMS is to maximize the tuple output rate for a query (i.e., instead of seeking the least cost plan, seek the plan with the highest tuple output rate). Specialized data structures (synopses) that allow fast modifications and summarize efficiently information are frequently used. Finally, architectures are usually data-flow oriented (i.e., operators are assembled together in a workflow-style diagram).

Adaptive Query Processing

Traditional query processors use a request-response paradigm for query answering, where users pose a query, the optimizer finds the "best" query plan and proceeds with query evaluation. However, the introduction of continuous queries along with the possibility of a large number of registered queries with the stream manager changes this model significantly.

Figure 2. Turning streams to relations



An aspect of continuous query is the need for adaptability to change: unbounded queries will run long enough to experience changes in system and data properties as well as system workload during their run. A continuous query engine should adapt gracefully to these changes, in order to ensure efficient processing over time (Madden, Shah, et. al., 2002). The goal of an *adaptive query processing* system is to maximize utilization of storage and computation (in terms of operator sharing) of multiple simultaneous queries present in the system.

Exact vs. Approximate Answers

Traditional database systems compute exact answers for queries. However, there are cases that approximate answers as either "good enough" (e.g., histograms, sketches, traditional DBMS) or "the only possibility" (e.g., DSMS where main-memory limitations and one-pass algorithms reduce the class of queries with exact answers). High quality approximate answers are often acceptable instead of exact answers. Several approximation algorithms for data streams applications have been developed in recent years. This work has led to some general techniques for data reduction and synopsis construction, such as sketches, random sampling, histograms, and wavelets. The research on these summarization techniques led to some work on approximate query answering. An overview of stream algorithms for computation of synopsis structures appears in Guha (2005).

Load Shedding

Systems have a limit on how much incoming stream tuples can handle. As a result, when the arrival rate is too high (e.g., during spikes), queues will build up and the system will become overloaded. In this case, some *load shedding* (i.e., discarding some incoming stream data)—is necessary for the system to continue functioning properly (Babcock et al., 2004; Tatbul et al., 2003). In general, there are two kinds of load shedding, semantic and random. Semantic shedding is based on filtering, that is, there is a filter (a predicate) that is applied on each stream tuple with selectivity 1-p. Random shedding is based on dropping randomly incoming data with a probability p%, eliminating thus p% of the input load.

A stream system should be able to answer three questions: when, where and how much load shedding. The first issue—when load shedding should take place—requires constant monitoring of the system. The second one involves deciding the place in a query plan that load shedders will be added. For example, Aurora, a DSMS prototype, defines a stream query as a network (a workflow) of operators. Load shedding can be modeled as another operator, placed within this network. Dropping tuples as early as possible saves resources but there may be a problem later, if streams fan out to multiple queries. The last issue has to do with estimating of how much shedding is required (percent of random drops or filtering appropriately).

Mining Streams

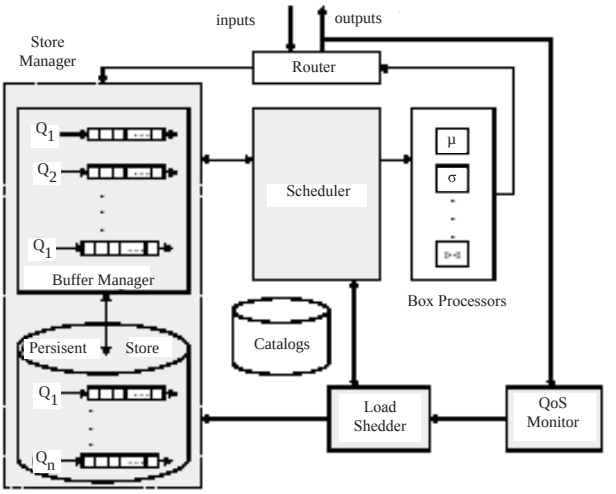
An orthogonal issue to processing streams is mining and indexing streams at real time (Garofalakis et. al., 2002). The goal of this topic is to find patterns in a sequence of measurements (financial, sensor, etc.) Besides the traditional mining patterns (classification, clustering and association rules), there exist many time series patterns that are of special interest in stream analysis (Faloutsos, 2004; Lerner et. al., 2004).

SYSTEM PROTOTYPES

The interest on data streams has led many research organizations (mainly in the U.S.) in the development of data stream prototypes. We list below some of the most well known systems.

- **Aurora:** Aurora is a joint project between Brandeis University, Brown University and MIT targeted towards stream monitoring applications. The processing mechanism is based on the optimization of a set of operators, connected in a data flow graph. Queries (three modes: continuous, ad-hoc and view) are formulated through a graphical "boxes and arrows" interface. Aurora's architecture is shown in Figure 3 (Carney et al., 2002).
  The designers of the Aurora stream system claim five unique features that distinguish it from other stream proposals: a workflow-orientation, a novel collection of operators, efficient scheduling, quality of service concerns and novel optimization methods.
  A workflow orientation is necessary due to the frequent preprocessing required for data streams. This way, users design an application (a workflow) with a "boxes" and "arrows" GUI and the system can optimize it more efficiently. The set of operators include "boxes" for filtering, mapping, aggregation (windowed) and join, similar to other stream systems. However, there are some distinct features: windowed operations have a timeout capability, the system deals with "out-of-order" messages (delayed messages), extensibility (user-defined functions) and re-sampling. Scheduling is a major component of Aurora system and it aims on reducing CPU cost of running the scheduler and maximizing the quality of service.
  An Aurora workflow is a dynamic object. When new applications get connected to the workflow, the number of boxes and arrows changes. Also adhoc queries that are run just once and then discarded have a similar effect. Furthermore, Aurora workflows may be quite

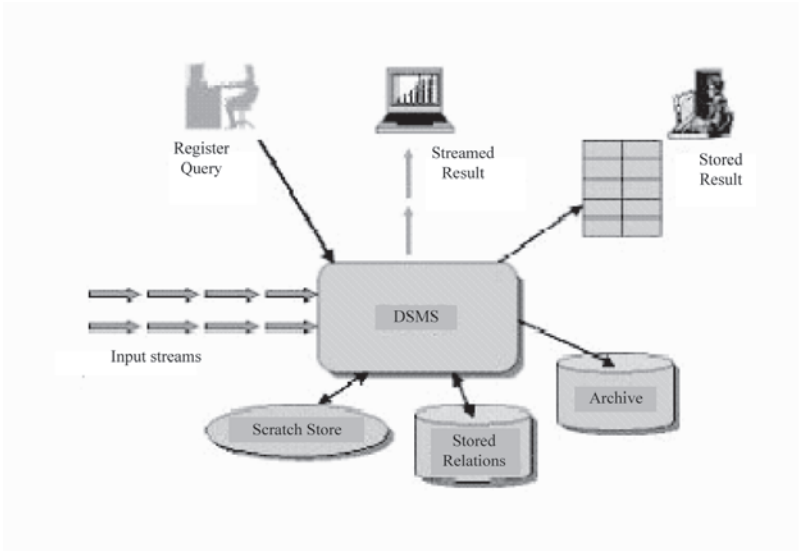
Figure 3. Aurora's architecture



large. As a result, Aurora performs only run-time optimization.

- **Gigascope:** AT&T's Gigascope project is a special-purpose data stream system for network applications, such as traffic analysis, intrusion detection, router configuration analysis, network monitoring and performance monitoring and debugging. The central component of Gigascope is a stream manager which tracks the query nodes that can be activated. Query nodes are processes. When they are started, they register themselves with the registry of the stream manager.
  Gigascope's query language, GSQL, is an SQL-like language with stream features. All inputs to and outputs of a GSQL query are data streams. Currently GSQL supports selection, join, aggregation and stream merge. A GSQL query is analyzed, translated to C/C++ code, and then executed. Optimization of GSQL queries consists of rearranging the query plan and generated-code optimizations. Another important optimization is pushing selections as far down as possible, even to the network level. To achieve this, the query processor of Gigascope "breaks" queries into high-level and low-level query nodes. Low-level query nodes are separate processes consisting of simple filtering conditions.
- **STREAM:** The Stanford stream data manager (STREAM) project at Stanford is developing a general-purpose DSMS for processing continuous queries over multiple continuous data streams and stored relations. The architecture of STREAM is shown in Figure 4 (Arasu et. al., 2003).
  The incoming input streams produce continuous data and drive query processing. Scratch store is used for the intermediate processing of the incoming data and can be stored in memory or on disk. An Archive storage is used for storing some (or all) of the data stream for possible off-line processing of expensive mining queries. Users or applications may register continuous queries and get the answers as output data streams or relational results that are updated over time. During the processing of continuous queries, traditional relational tables can be utilized. Currently STREAM offers a Web system interface through direct HTTP, and a Web-based GUI to register queries and view results.
  STREAM people developed a declarative query language, continuous query language (CQL) for continuous queries over data streams and relations. They model a stream as an unbounded, append-only bag of (tuple, timestamp) pairs, and a relation as a time-varying bag of tuples supporting updates, deletions and insertions. The key idea is to convert streams into relations using special windowing operations, perform transformations on relations using standard relational operators and then convert back (optionally) the transformed relational data into a streamed answer (Arasu et al., 2003). CQL uses SQL as its relational query language, SQL-99 amendments provide the window specification language, and it includes three relation-to-stream operators. When a continuous query specified in CQL is registered with STREAM, it is compiled into a query plan. The query plan is merged with existing query plans whenever possible, in order to share computation and memory.
- **TelegraphCQ:** TelegraphCQ is a prototype focused on handling measurements of sensor networks. Developed in Berkeley, it is based on adaptive and multi-query

Figure 4. Overview of STREAM architecture



processing and uses a mix of SQL and scripting programming languages for expressing complex queries. Its architecture is shown in Figure 5 (Chandrasekaran et al., 2003).

Telegraph, the predecessor of TelegraphCQ, consists of an extensible set of composable dataflow modules or operators that produce and consume records in a manner analogous to the operators used in traditional database query engines. The modules can be composed into multi-step dataflows, exchanging records via an API called Fjords (Madden & Franklin, 2002) that can support communication via either "push" or "pull" modalities.

TelegraphCQ supports continuous queries defined over relational tables and data streams with a rich set of windowing schemes (e.g., landmark and sliding windows).

Other stream projects include Tribeca, developed in Telcordia for network management; NiagaraCQ, a project of University of Wisconsin to monitor internet data (XML-format); Cougar and Amazon of Cornell University, aimed to monitor distributed sensor networks. We present in Table 2 a comparison table between Aurora, Gigascope, STREAM and Telegraph.

Figure 5. TelegraphCQ's architecture

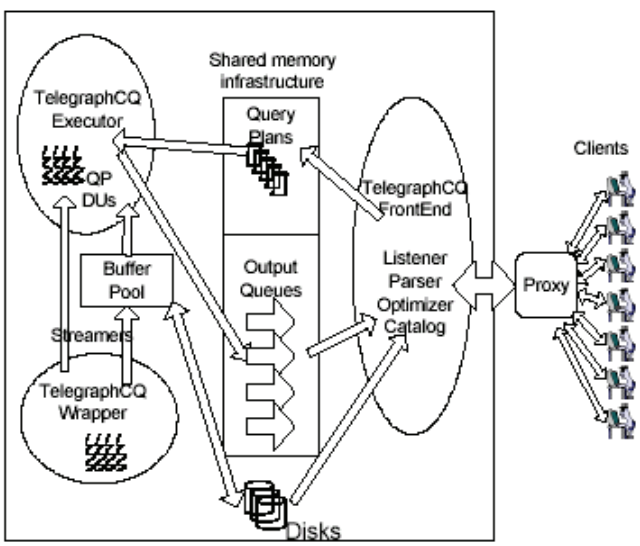


Table 2. Comparing different stream prototypes

Stream System	Input	Output	Query Language	Answers	Evaluation
Aurora	Relations, Streams	Relations, Streams	Operators (boxes and arrows)	Approximate	Run-time optimization
Gigascope	Streams	Streams	GSQL (SQL-like)	Exact	Splitting query to high- and low-level
STREAM	Relations, Streams	Relations, Streams	CQL (SQL-like)	Approximate	Static analysis, relational optimization
Telegraph	Relations, Streams	Relations, Streams	SQL-based, scripting	Exact	Adaptive query processing, multi-query

CONCLUSION

Equipping traditional database management systems with decision support capabilities has been studied extensively in the last decade. Building data warehouses, forming data cubes, creating data marts, carrying out data mining tasks and performing on-line analytical processing and multi-dimensional analysis are well-known subjects in the context of decision-making. However, an emerging trend in data management and data streams, is very different in nature from traditional database applications: queries are different, models and frameworks are special, query processing is peculiar and, subsequently, decision support is defined differently. The keywords in this domain are: real-time, continuous, and tactical decisions. A lot of progress has been recently made in this field by several research institutions to build various (either general or specific-purpose) data stream management systems with rich data analysis capabilities. Given the popularity of sensors and the need for real-time analytics, the ability of data analysis on top of data streams will be a crucial component of future data management and decision support systems.

REFERENCES

Abiteboul, S., Agrawal, R., Bernstein, P., Carey, M., & Ceri, S. et al. (2005). The Lowell Database Research Self-assessment. *Communications of the ACM, 48*(5), 111-118.

Arasu, A., Babu, S. & Widom, J. (2006). The CQL continuous query language: Semantic foundations and query execution. *The VLDB Journal, 15*(2), 121-142.

Arasu, A., Babcock, B., Babu, S., Datar, M., Ito, K., Motwani, R., Nishizawa, I., Srivastava, U., Thomas, D., Varma, R., & Widom, J. (2003). STREAM: The Stanford Stream Data Manager. *IEEE Data Engineering Bulletin, 26*(1), 19-26.

Babcock, B., Datar, M. & Motwani, R.. (2004). Load shedding for aggregation queries over data streams. In *Proceedings of International Conference on Data Engineering (ICDE)* (pp. 350-361).

Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and Issues in Data Streams. In *Proceedings of the 2002 ACM Symp. on Principles of Database Systems (PODS)* (pp.1-20).

Babu, S., & Widom, J. (2001). Continuous queries over data streams. *SIGMOD Record, 30*(3), 109-120.

Carney, D., Cetintemel, U., Cherniack, M., Convey, C., Lee, S., Seidman, G., Stonebraker, M., Tatbul, N., & Zdonik, S. (2002). Monitoring streams—A new class of data management applications. *28th International Conference on Very Large Databases (VLDB)* (pp.215-226).

Chandrasekaran, S., Cooper, O., Deshpande, A., Franklin, MJ., Hellerstein, J.M., Hong, W., Krishnamurthy, S., Madden, S., Raman, V., Reiss, F., & Shah, M. A. (2003). TelegraphCQ: Continuous dataflow processing for an uncertain world. *Conference on Innovative Data Systems Research*.

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record, 26*(1), 65-74.

Chen, J., DeWitt, DJ., Tian, F., & Wang, Y. (2000). NiagaraCQ: A Scalable Continuous Query System for Internet Databases. *ACM SIGMOD, Conference on Management of Data* (pp. 379-390).

Cortes, C., Fisher, K.,Pregibon, D., Rogers, A., & Smith, F. (2000). Hancock: A language for extracting signatures from data streams. *ACM SIGKDD, Conference on Knowledge Discovery and Data Mining* (pp. 9-17).

Cranor, C., Johnson, T., Spatscheck, O., & Shkapenyuk, V. (2003). Gigascope: A stream database for network applica-

tions. *ACM SIGMOD, Conference on Management of Data* (pp. 647-651).

Garofalakis, M., Gehrke, J., & Rastogi, R. (2002). Querying and mining data streams: You only get one look, a tutorial. In *Proceedings of the 2002 ACM SIGMOD International. Conference on Management of Data* (p. 635).

Guha, S., & Shim, K. (2005). Offline and stream algorithms for efficient computation of synopsis structures (Tutorial). In *Proceedings of the International Confernce on Management of Data* (p. 1364).

Faloutsos, C. (2004). Indexing and Mining Streams (Tutorial). In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (p. 969).

Koudas, N. & Srivastava, D. (2003). Data stream query processing: A tutorial. *29th International Conference on Very Large Databases (VLDB)* (p. 1149).

Lerner, A., Shasha, D., Wang, Z., Zhao, X., & Zhu, Y. (2004). Fast algorithms for time series with applications to finance, physics, music, biology, and other suspects (Tutorial). In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (pp. 965-968).

Madden, S. & Franklin, MJ. (2002). Fjording the Stream: An Architecture for Queries Over Streaming Sensor Data. In *Proceedings of the 2002 International Conference on Data Engineering (ICDE)* (pp. 555-566).

Madden, S., Shah, MA., Hellerstein, JM., & Raman, V. (2002). Continuously Adaptive Continuous Queries Over Streams. In *Proceedings of the ACM SIGMOD International. Conference on Management of Data* (pp. 49-60).

Muthukrishnan, S. (2003). Data Streams: Algorithms and Applications. *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (pp. 413).

Sullivan, M. (1996). Tribeca: A stream database manager for network traffic analysis. *22th International Conference on Very Large Databases (VLDB)* (p. 594).

Tatbul, N., Çetintemel, U., Zdonik, S., Cherniack, M., & Stonebraker, M. (2003). Load shedding in a data stream manager. *29th International Conference on Very Large Databases (VLDB)* (pp. 309-320).

Terry, D., Goldberg, D., Nichols, D., & Oki, B. (1992). Continuous queries over append-only databases. *31st International Conference on Very Large Databases (VLDB)* (pp. 309-320).

Zdonik, S., Stonebraker, M., Cherniack, M., Çetintemel, U., Balazinska, M., & Balakrishnan, H. (2003). The Aurora and Medusa projects. *IEEE Data Engineering Bulletin, 26*(1), 3-10.

KEY TERMS

Continuous Queries: The answer to a continuous query is produced over time, reflecting the stream data seen so far. Answers may be stored and updated as new data arrives or may be produced as data streams themselves.

Data Stream Management Systems (DSMS): A data management system providing capabilities to query and process data streams and store a bounded part of it.

Data Streams: Data items that arrive online from multiple sources in a continuous, rapid, time-varying, possibly unpredictable fashion.

Load Shedding: The discarding of input data by the DSMS when the input stream rate exceeds system capacity. It can either be semantic (based on certain semantic rules) or random.

Measurement Data Streams: Data streams representing successive state information of one or more entities, such as sensor, climate or network measurements.

Network Traffic Management: Monitoring a variety of continuous network data streams at real-time, such as packet traces, packet flows and performance measurements in order to compute statistics, detect anomalies and adjust routing.

Stream Window: A mechanism to extract a finite set of records from an infinite stream. This mechanism selects stream items based on time periods, counting, or explicit starting and ending conditions.

Transactional Data Streams: Data streams representing log interactions between entities, such as credit card transactions, phone calls and Web click streams.

Database Benchmarks

Jérôme Darmont
ERIC, University of Lyon 2, France

INTRODUCTION

Performance measurement tools are very important, both for designers and users of Database Management Systems (DBMSs). *Performance evaluation* is useful to designers to determine elements of architecture, and, more generally, to validate or refute hypotheses regarding the actual behavior of a DBMS. Thus, *performance evaluation* is an essential component in the development process of well-designed and efficient systems. Users may also employ *performance evaluation*, either to compare the efficiency of different technologies before selecting a DBMS, or to tune a system.

Performance evaluation by experimentation on a real system is generally referred to as benchmarking. It consists of performing a series of tests on a given DBMS to estimate its performance in a given setting. Typically, a *benchmark* is constituted of two main elements: a database model (conceptual schema and extension), and a workload model (set of read and write operations) to apply on this database, following a predefined protocol. Most *benchmarks* also include a set of simple or composite performance metrics such as response time, throughput, number of input/output, disk or memory usage, and so forth.

The aim of this article is to present an overview of the major families of state-of-the-art database benchmarks, namely, relational benchmarks, object and object-relational benchmarks, XML benchmarks, and decision-support benchmarks; and to discuss the issues, tradeoffs, and future trends in database benchmarking. We particularly focus on XML and decision-support benchmarks, which are currently the most innovative tools that are developed in this area.

BACKGROUND

Relational Benchmarks

In the world of relational DBMS benchmarking, the *Transaction Processing Performance Council (TPC)* plays a preponderant role. The mission of this non-profit organization is to issue standard benchmarks, to verify their correct application by users, and to regularly publish performance tests results. Its benchmarks all share variants of a classical business database (*customer-order-product-supplier*) and are only parameterized by a scale factor that determines the database size (*e.g.*, from 1 to 100,000 GB).

The *TPC* benchmark for transactional databases, TPC-C (TPC, 2005a), has been in use since 1992. It is specifically dedicated to On-Line Transactional Processing (OLTP) applications, and features a complex database (nine types of tables bearing various structures and sizes), and a workload of diversely complex transactions that are executed concurrently. The metric in TPC-C is throughput, in terms of transactions.

There are currently few credible alternatives to *TPC-C*. Although, we can cite the Open Source Database Benchmark (OSDB), which is the result of a project from the free software community (SourceForge, 2005). OSDB extends and clarifies the specifications of an older benchmark, AS^{3}AP. It is available as free C source code, which helps eliminate any ambiguity relative to the use of natural language in the specifications. However, it is still an ongoing project and the benchmark's documentation is very basic. AS^{3}AP's database is simple: it is composed of four relations whose size may vary from 1 GB to 100 GB. The workload is made of various queries that are executed concurrently. OSDB's metrics are response time and throughput.

Object-Oriented and Object-Relational Benchmarks

There is no standard benchmark for object-oriented DBMSs. However, the most frequently cited and used, OO1 (Cattel, 1991), HyperModel (Anderson, Berre, Mallison, Porter, & Schneider, 1990), and chiefly *OO7* (Carey, DeWitt, & Naughton, 1993), are *de facto* standards. These benchmarks mainly focus on engineering applications (*e.g.*, computer-aided design, software engineering). They range from OO1, which bears a very simple schema (two classes) and only three operations, to *OO7*, which is more generic and proposes a complex and tunable schema (ten classes), as well as fifteen complex operations. However, even *OO7*, the more elaborate of these benchmarks, is not generic enough to model other types of applications, such as financial, multimedia, or telecommunication applications (Tiwary, Narasayya, & Levy, 1995). Furthermore, its complexity makes it hard to understand and implement. To circumvent these limitations, the *OCB* benchmark has been proposed (Darmont & Schneider, 2000). Wholly tunable, this tool aims at being truly generic. Still, the benchmark's code is short, reasonably easy to implement, and easily portable. Finally, *OCB* has been extended into the Dynamic Evaluation Framework (DEF),

which introduces a dynamic component in the workload, by simulating access pattern changes using configurable styles of changes (He & Darmont, 2005).

Object-relational benchmarks such as BUCKY (Carey, DeWitt, & Naughton, 1997) and BORD (Lee, Kim, & Kim, 2000) are query-oriented and solely dedicated to object-relational systems. For instance, BUCKY only proposes operations that are specific to these systems, considering that typical object navigation is already addressed by object-oriented benchmarks. Hence, these benchmarks focus on queries implying object identifiers, inheritance, joins, class and object references, multivalued attributes, query unnesting, object methods, and abstract data types.

XML Benchmarks

Since there is no standard model, the storage solutions for XML (eXtended Markup Language) documents that have been developed since the late nineties bear significant differences, both at the conceptual and the functionality levels. The need to compare these solutions, especially in terms of performance, has lead to the design of several benchmarks with diverse objectives.

X-Mach1 (Böhme & Rahm, 2001), XMark (Schmidt, Waas, Kersten, Carey, Manolescu, & Busse, 2002), XOO7 (an extension of OO7) (Bressan, Lee, Li, Lacroix, & Nambiar, 2002) and XBench (Yao, Ozsu, & Khandelwal, 2004) are so-called application benchmarks. Their objective is to evaluate the global performances of an XML DBMS, and more particularly of its query processor. Each of them implements a mixed XML database that is both data-oriented (structured data) and document-oriented (in general, random texts built from a dictionary). However, except for XBench that proposes a true mixed database, their orientation is more particularly focused on data (XMark, XOO7) or documents (X-Mach1).

These benchmarks also differ in:

- the fixed or flexible nature of the XML schema (one or several Document Type Definitions or XML schemas);
- the number of XML documents used to model the database at the physical level (one or several);
- the inclusion or not of update operations in the workload.

We can also underline that only XBench helps in evaluating all the functionalities offered by the XQuery language.

Micro-benchmarks have also been proposed to evaluate the individual performances of basic operations such as projections, selections, joins, and aggregations, rather than more complex queries. The Michigan Benchmark (Runapongsa, Patel, Jagadish, & Al-Khalifa, 2002) and MemBeR (Afanasiev, Manolescu, & Michiels, 2005) are made for XML documents storage solution designers, who can isolate critical issues to optimize, rather than for users seeking to compare different systems. Furthermore, MemBeR proposes a methodology for building micro-databases, to help users in adding datasets and specific queries to a given performance evaluation task.

Decision-Support Benchmarks

Since decision-support benchmarks are currently a *de facto* subclass of relational benchmarks, the TPC again plays a central role in their standardization. *TPC-H* (TPC, 2005c) is currently their only decision-support benchmark. It exploits a classical *product-order-supplier* database schema, as well as a workload that is constituted of twenty-two SQL-92, parameterized, decision-support queries, and two refreshing functions that insert tuples into, and delete tuples from, the database. Query parameters are randomly instantiated following a uniform law. Three primary metrics are used in *TPC-H*. They describe performance in terms of power, throughput, and a combination of these two criteria.

Data warehouses nowadays constitute a key decision-support technology. However, *TPC-H's* database schema is not a star-like schema that is typical in data warehouses. Furthermore, its workload does not include any On-Line Analytical Processing (OLAP) query. *TPC-DS*, which is currently under development (TPC, 2005b), fills in this gap. Its schema represents the decision-support functions of a retailer under the form of a constellation schema with several fact tables and shared dimensions. *TPC-DS*' workload is constituted of four classes of queries: reporting queries, *ad-hoc* decision-support queries, interactive OLAP queries, and extraction queries. SQL-99 query templates help in randomly generating a set of about five hundred queries, following non-uniform distributions. The warehouse maintenance process includes a full ETL (Extract, Transform, Load) phase, and handles dimensions according to their nature (non-static dimensions scale up while static dimensions are updated). One primary throughput metric is proposed in TPC-DS. It takes both query execution and the maintenance phase into account.

As in all the other *TPC benchmarks*, scaling in *TPC-H* and *TPC-DS* is achieved through a scale factor that helps defining the database's size (from 1 GB to 100 TB). Both the database schema and the workload are fixed.

There are, again, few decision-support benchmarks out of the *TPC*, and their specifications are rarely integrally published. Some are nonetheless interesting. APB-1 is presumably the most famous. Published by the OLAP Council, a now inactive organization founded by OLAP vendors, APB-1 has been intensively used in the late nineties. Its warehouse dimensional schema is structured around four dimensions: *Customer*, *Product*, *Channel*, and *Time*. Its workload of ten queries is aimed at sale forecasting. APB-1 is quite simple and proved limited to evaluate the specifici-

ties of various activities and functions (Thomsen, 1998). It is now difficult to find.

Eventually, while the *TPC* standard benchmarks are invaluable to users for comparing the performances of different systems, they are less useful to system engineers for testing the effect of various design choices. They are indeed not tunable enough and fail to model different data warehouse schemas. By contrast, the *Data Warehouse Engineering Benchmark (DWEB)* helps in generating various *ad-hoc* synthetic data warehouses (modeled as star, snowflake, or constellation schemas) and workloads that include typical OLAP queries (Darmont, Bentayeb, & Boussaïd, 2005a). *DWEB* is fully parameterized to fulfill data warehouse design needs.

ISSUES AND TRADEOFFS IN DATABASE BENCHMARKING

Gray (1993) defines four primary criteria to specify a "good" *benchmark:*

1. *relevance:* the benchmark must deal with aspects of performance that appeal to the largest number of potential users;
2. *portability:* the benchmark must be reusable to test the performances of different DBMSs;
3. *simplicity:* the benchmark must be feasible and must not require too many resources; and
4. *scalability:* the benchmark must adapt to small or large computer architectures.

In their majority, existing benchmarks aim at comparing the performances of different systems in given experimental conditions. This helps vendors in positioning their products relative to their competitors', and users in achieving strategic and costly software choices based on objective information. These benchmarks invariably present fixed database schemas and workloads. Gray's scalability factor is achieved through a reduced number of parameters that mainly allow varying the database size in predetermined proportions. It is notably the case of the unique scale factor parameter that is used in all the TPC benchmarks.

This solution is simple (still according to Gray's criteria), but the relevance of such benchmarks is inevitably reduced to the test cases that are explicitly modeled. For instance, the typical *customer-order-product-supplier* that is adopted by the *TPC* is often unsuitable to application domains other than management. This leads benchmark users to design more or less elaborate variants of standard tools when they feel these are not generic enough to fulfill particular needs. Such users are generally not confronted with software choices, but are rather designers who have quite different needs. They mainly seek to evaluate the impact of architectural choices, or performance optimization techniques, within a given system or a family of systems. In this context, it is essential to multiply experiments and test cases, and a monolithic benchmark is of reduced relevance.

To enhance the relevance of benchmarks aimed at system designers, we propose to extend Gray's scalability criterion to *adaptability*. A performance evaluation tool must then be able to propose various database or workload configurations, to allow experiments to be performed in various conditions. Such a tool may be qualified as a benchmark generator, or as a tunable or generic benchmark. However, aiming at a better adaptability is mechanically detrimental to a benchmark's simplicity. This criterion, though, remains very important and must not be neglected when designing a generic tool. It is, thus, necessary to devise means of achieving a good adaptability, without sacrificing simplicity too much. In summary, a satisfying tradeoff must be reached between these two orthogonal criteria.

We have been developing benchmarks following this philosophy for almost ten years. The first one, the Object Clustering Benchmark (*OCB*), was originally designed to evaluate the performances of clustering algorithms within object-oriented DBMSs. By extending its clustering-oriented workload, we made it generic. Furthermore, its database and workload are wholly tunable, through a collection of comprehensive but easily set parameters. Hence, *OCB* can be used to model many kinds of object-oriented database applications. In particular, it can simulate the behavior of the other object-oriented benchmarks.

Our second benchmark is the *DWEB* data warehouse benchmark. *DWEB's* parameters help users in selecting the data warehouse architecture and workload they need in a given context. To solve the adaptability *versus* simplicity dilemma, we divided the parameter set into two subsets. Low-level parameters allow an advanced user to control everything about data warehouse generation. However, their number can increase dramatically when the schema gets larger. Thus, we designed a layer of high-level parameters that may be easily understood and set up, and that are in reduced number. More precisely, these high-level parameters are average values for the low-level parameters. At database generation time, the high-level parameters are automatically exploited by random functions to set up the low-level parameters.

FUTURE TRENDS

The development of XML-native DBMSs is quite recent, and a tremendous amount of research is currently in progress to help them become a credible alternative to XML-compatible, relational DMBSs. Several performance evaluation tools have been proposed to support this effort. However, research in this area is very dynamic, and new benchmarks will be needed to assess the performance of the latest discoveries. For instance, Active XML incorporates web services for data integration

(Abiteboul, Benjelloun, Manolescu, Milo, & Weber 2002). An adaptation of existing XML benchmarks that would exploit the concepts developed in TPC-App, could help in evaluating the performance of an Active XML platform.

No XML benchmark is currently dedicated to decision-support either, while many XML data warehouse architectures have been proposed in the literature. We are currently working on a benchmark called XWB, which is aimed at evaluating the performances of such research proposals. Furthermore, there is a growing need in many decision-support applications (e.g., customer relationship management, marketing, competition monitoring, medicine) to exploit complex data, that is, in summary, data that are not only numerical or symbolic. XML is particularly adapted to describe and store complex data (Darmont, Boussaïd, Ralaivao, & Aouiche, 2005b) and further adaptations of XML decision-support benchmarks would be needed to take them into account.

Finally, a lot of research also aims at enhancing the XQuery language, for instance with update capabilities, or with OLAP operators. Existing XML and/or decision-support benchmarks will also have to be adapted to take these new features into account.

CONCLUSION

Benchmarking is a small field, but it is nonetheless essential to database research and industry. It serves both engineering or research purposes, when designing systems or validating solutions, and marketing purposes, when monitoring competition and comparing commercial products.

Benchmarks might be subdivided in three classes. First, standard, *general-purpose benchmarks* such as the TPCs do an excellent job in evaluating the global performance of systems. They are well-suited to software selection by users and marketing battles by vendors, who try to demonstrate the superiority of their product at one moment in time. However, their relevance drops for some particular applications that exploit database models or workloads that are radically different from the ones they implement. *Ad hoc benchmarks* are a solution. They are either adaptations of general-purpose benchmarks, or specifically designed benchmarks such as the XML micro-benchmarks we described above. Designing myriads of narrow-band benchmarks is not time-efficient, though, and trust in yet another new benchmark might prove limited in the database community. Hence, the solution we promote is to use *generic benchmarks* that feature a common base for generating various experimental possibilities. The drawback of this approach is that parameter complexity must be mastered, for generic benchmarks to be easily apprehended by users.

In any case, before starting a benchmarking experiment, users' needs must be carefully assessed so that the right benchmark or benchmark class is selected, and test results are meaningful. This sounds like sheer common sense, but many researchers simply select the best known tools, whether they are adapted to their validation experiments or not. For instance, data warehouse papers often refer to TPC-H, while this benchmark's database is not a typical data warehouse, and its workload does not include any OLAP query. *Ad-hoc* and generic benchmarks should be preferred in such situations, and though trust in a benchmark is definitely an issue, relevance should be the prevailing selection criteria. We modestly hope this article will have provided its readers with a fair overview of database benchmarks, and will help them in selecting the right tool for the right job.

REFERENCES

Abiteboul, S., Benjelloun, O., Manolescu, I., Milo, T., & Weber, R. (2002). Active XML: Peer-to-Peer Data and Web Services Integration. *28th International Conference on Very Large Data Bases (VLDB 02)*. Hong Kong, China. 1087-1090.

Afanasiev, L., Manolescu, I.., & Michiels, P. (2005). MemBeR: A Micro-benchmark Repository for XQuery. *3rd International XML Database Symposium (XSym 05)*. Trondheim, Norway. LNCS. 3671, 144-161.

Anderson, T.L., Berre, A.G., Mallison, M., Porter, H.H., & Schneider, B. (1990). The HyperModel Benchmark. *International Conference on Extending Database Technology (EDBT 90)*. Venice, Italy. LNCS. 416, 317-331.

Böhme, T., & Rahm, E. (2001). XMach-1: A Benchmark for XML Data Management. *Datenbanksysteme in Büro, Technik und Wissenschaft (BTW 01)*, Oldenburg, Germany. 264-273.

Bressan, S., Lee, M.L., Li, Y.G., Lacroix, Z., & Nambiar, U. (2002). The XOO7 Benchmark. Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web. *VLDB 2002 Workshop EEXTT*. Hong Kong, China. LNCS. 2590, 146-147.

Carey, M.J., DeWitt, D.J., & Naughton, J.F. (1993). The OO7 benchmark. *ACM SIGMOD International Conference on Management of Data (SIGMOD 93)*. Washington, USA. 12-21.

Carey, M.J., Dewitt, D.J. & Naughton, J.F. (1997). The BUCKY Object-Relational Benchmark. *ACM SIGMOD International Conference on Management of Data (SIGMOD 97)*, Tucson, USA. 135-146.

Cattell, R.G.G. (1991). *An Engineering Database Benchmark*. The Benchmark Handbook for Database and Transaction Processing Systems, 1st Edition. Morgan Kaufmann. 247-281.

Darmont, J., Bentayeb, F., & Boussaïd, O. (2005a). DWEB: A Data Warehouse Engineering Benchmark. *7th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 05). Copenhagen, Denmark. LNCS.* 3589, 85-94.

Darmont, J., Boussaïd, O., Ralaivao, J.C., & Aouiche, K. (2005b). An Architecture Framework for Complex Data Warehouses. *7th International Conference on Enterprise Information Systems (ICEIS 05). Miami, USA.* 370-373.

Darmont, J., & Schneider, M. (2000). Benchmarking OODBs with a Generic Tool. *Journal of Database Management*. 11(3), 16-27.

Gray, J. (ed.). (1993). *The Benchmark Handbook for Database and Transaction Processing Systems, 2nd Edition*. Morgan Kaufmann.

He, Z., & Darmont, J. (2005). Evaluating the Dynamic Behavior of Database Applications, *Journal of Database Management*, 16(2), 21-45.

Lee, S., Kim, S., & Kim, W. (2000). The BORD Benchmark for Object-Relational Databases. *11th International Conference on Database and Expert Systems Applications (DEXA 00)*. London, UK. LNCS. 1873, 6-20.

Runapongsa, K., Patel, J.M., Jagadish, H.V., & Al-Khalifa, S. (2002). The Michigan Benchmark: A Microbenchmark for XML Query Processing Systems. *Efficiency and Effectiveness of XML Tools and Techniques and Data Integration over the Web, VLDB 2002 Workshop EEXTT*. Hong Kong, China. LNCS. 2590, 160-161.

Schmidt, A., Waas, F., Kersten, M., Carey, M.J., Manolescu, I., & Busse, R. (2002). XMark: A Benchmark for XML Data Management. *28th International Conference on Very Large Databases (VLDB 02)*. Hong Kong, China. 974-985.

SourceForge. (2005). *The Open Source Database Benchmark*, Version 0.19. http://osdb.sourceforge.net

Thomsen, E. (1998). *Comparing different approaches to OLAP calculations as revealed in benchmarks*. Intelligence Enterprise's Database Programming & Design. http://www.dbpd.com/vault/9805desc.htm

Tiwary, A., Narasayya, V., & Levy, H. (1995). Evaluation of OO7 as a system and an application benchmark. *Workshop on Object Database Behavior, Benchmarks and Performance*. Austin, USA.

TPC. (2005a). *TPC Benchmark C Standard Specification revision 5.6*. Transaction Processing Performance Council. http://www.tpc.org

TPC. (2005b). *TPC Benchmark DS (Decision Support) Draft Specification revision 32*. Transaction Processing Performance Council. http://www.tpc.org

TPC. (2005c). *TPC Benchmark H Standard Specification revision 2.3.0.* Transaction Processing Performance Council. http://www.tpc.org

Yao, B.B., Ozsu, T., & Khandelwal, N. (2004). XBench Benchmark and Performance Testing of XML DBMSs. *20th International Conference on Data Engineering (ICDE 04)*, Boston, USA. 621-633.

KEY TERMS

Benchmark: A standard program that runs on different systems to provide an accurate measure of their performance.

Database Benchmark: A benchmark specifically aimed at evaluating the performance of DBMSs or DBMS components.

Database Management System (DMBS): Software set that handles the structuring, storage, maintenance, update, and querying of data stored in a database.

Database Model: In a database benchmark, a database schema and a protocol for instantiating this schema, that is, generating synthetic data or reusing real-life data.

Performance Metrics: Simple or composite metrics aimed at expressing the performance of a system.

Synthetic Benchmark: A benchmark in which the workload model is artificially generated, as opposed to a real-life workload.

Workload Model: In a database benchmark, a set of predefined read and write operations or operation templates to apply on the benchmark's database, following a predefined protocol.

Database Integration in the Grid Infrastructure

Emmanuel Udoh
Indiana University – Purdue University, USA

INTRODUCTION

The capability of the Web to link information over the Internet has popularized computer science to the public. But it is the grid that will enable the public to exploit data storage and computer power over the Internet analogous to the electric power utility (a ubiquitous commodity). The grid is considered the fifth generation computing architecture after client-server and multitier (Kusnetzky & Olofson, 2004) that integrates resources (computers, networks, data archives, instruments, etc.) in an interoperable virtual environment (Berman, Fox, & Hey, 2003). In this vein, grid computing is a new IT infrastructure that allows modular hardware and software to be deployed collectively to solve a problem or rejoined on demand to meet changing needs of a user.

Grid computing is becoming popular in the enterprise world after its origin in the academic and research communities (e.g., SETI@home), where it was successfully used to share resources, store data in petabytes/exabytes, and ultimately lower costs. There are several reasons for the embrace of the enterprise grids. In the nineties, the IT world was confronted with the high cost of maintaining smaller, cheaper and dedicated servers such as UNIX and Windows. According to Oracle (2005), there was the problem of application silos that lead to underutilized hardware resources; monolithic and unwieldy systems that are expensive to maintain and change; and fragmented and disintegrated information that cannot be fully exploited by the enterprise as a whole. Various surveys put the average utilization of servers in a typical enterprise to often much less than 20% (Goyal & Lawande, 2006; Murch, 2004). But with the increasingly available cheaper, faster and affordable hardware such as server blades, and operating systems like the open source Linux, the IT world embraced grid computing to save money on hardware and software. With the growing importance of grid computing, it is easy to conjecture why many new terms have been coined for it. In the literature and industry, other terms used interchangeably for grid computing are utility computing, computing on demand, N1, hosted computing, adaptive computing, organic computing and ubiquitous computing (Goyal & Lawande, 2006; Murch, 2004; Oracle, 2005).

The grid is an all-encompassing, 21st century computing infrastructure (Foster, 2003; Joseph & Fellenstein, 2004) that integrates several areas of computer science and engineering. A database is an important component of the application stack in the industry and is increasingly being embedded in the grid infrastructure. This article focuses on integration of database grids or grid-accessible databases in the industry using Oracle products as examples. Vendors like Oracle and IBM are providing grid-enabled databases that are supposed to make enterprise systems unbreakable and highly available. Oracle has been in the forefront in this endeavor with its database products. In recognition of the significant new capabilities required to power grid computing, Oracle has named its new technology products Oracle 10g (g for grid). Oracle provides seamless availability through its database products with such features like streams, transportable tablespaces, data hubs, ultra-search and real application clusters. Although companies will not like to distribute resources randomly on the Internet, they will embrace enterprise database grids, as they embraced Internet in the form of Intranets. To the business world, database grids will help achieve high hardware utilization and resource sharing, high availability, flexibility, incrementally scalable low cost components and reduced administrative overhead (Kumar & Burleson, 2005; Kusnetzky & Olofson, 2004).

BACKGROUND

The grid technology is still evolving, and databases are increasingly being incorporated into its infrastructure. IBM has contributed immensely in this endeavor with its autonomic and grid computing projects, but Oracle is the clear-cut industry leader in enterprise database grids. Oracle, a member of the enterprise grid alliance (EGA) that promotes tools and standards for enterprise computing, has been preparing and targeting the grid market since the Oracle9i release with products like Oracle real application clusters (now morphed to automatic storage management system). This article will therefore discuss enterprise grid computing in terms of the features available in Oracle 10g database products.

The main features that clearly differentiate grid computing from other forms of computing architectures, such as client server or multitier, are virtualization and provisioning. Virtualization creates a logical view or abstraction that allows the pooling together of resources (e.g., data, computing power, storage capacity, and other resources) for consumption by other applications, while provisioning determines how to meet on demand the specific needs of consumers.

As consumers request resources through the virtualization layer (which breaks the hard-coded connection between providers and consumers (Oracle, 2005), provisioning guarantees resources are allocated for the request. To achieve these objectives, a grid implements a layered architecture as depicted in Figure 1a.

In a semblance of the generalized grid architecture, the Oracle grid builds a stack of its software in a virtual environment, as shown in Figure 1b. The bottom layer hosts the storage units such as a storage area network (SAN), while the next horizontal layer contains the infrastructure such as the hardware and software that create a data storage and program execution environment (infrastructure grid). The next layer, the application server, contains the program logic and flow that define specific business processes (application grid). The topmost layer (information grid) hosts applications such as user applications, enterprise resource planning and portal software that can be accessed over the network without the application being architected to support the device or network. This virtualized environment has a unified management structure as well as an infrastructure for security.

As depicted in Figure 1b, Oracle 10g software stack, which is configured to self-manage, acts as a single computing resource to the user even in a geographically distributed environment. This allows organizations to protect and optimize their investment in hardware and software and also access newer system features in a more reliable, powerful and scalable environment. To keep this unified structure manageable and also eliminate the application silo model, Oracle enterprise manager has a grid control that monitors, provisions, clones and automates even in heterogeneous environments (Khilani, 2005).

MAIN FOCUS

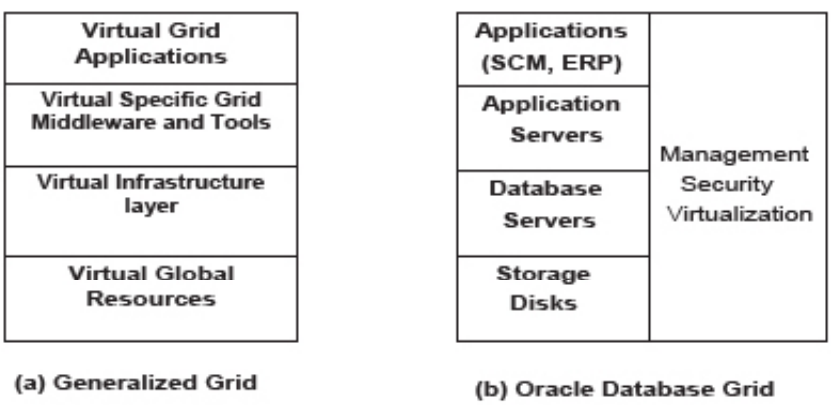
The evolving enterprise database grid brings substantial benefits to the industry but poses a major challenge in integration and adoption. With yearly global business volume in excess of $10 billion dollars, databases are critical components in the enterprise application stack. The existing database infrastructure in companies are aggregates of many years of investments in a wide range of interfaces and tools for performance and security. Noting all the hype about grids, it is natural that its adoption will be resisted by employees, who are unwilling to change existing processes and technology. Researchers believe that building grid-enabled database infrastructures from scratch is both unrealistic and a waste of effort and resources (Watson, 2003). Instead, existing database management systems should be integrated into the grid in an incremental manner, absorbing the existing investments without being disruptive. Oracle's grid-enabled database takes cognizance of this reality, and hence this article focuses on database grid integration and adoption vis-a-vis Oracle grid products.

At the department of computer science (CS), Indiana University–Purdue University, Fort Wayne, an Oracle database grid was introduced to aid the administration of the database program (for research and education). Experience from this program supports these transitioning steps for the integration and adoption of an enterprise database grid: identification, standardization, consolidation and automation.

Identification

Organizations have different IT infrastructures that may influence decisions to integrate their enterprises in database grids.

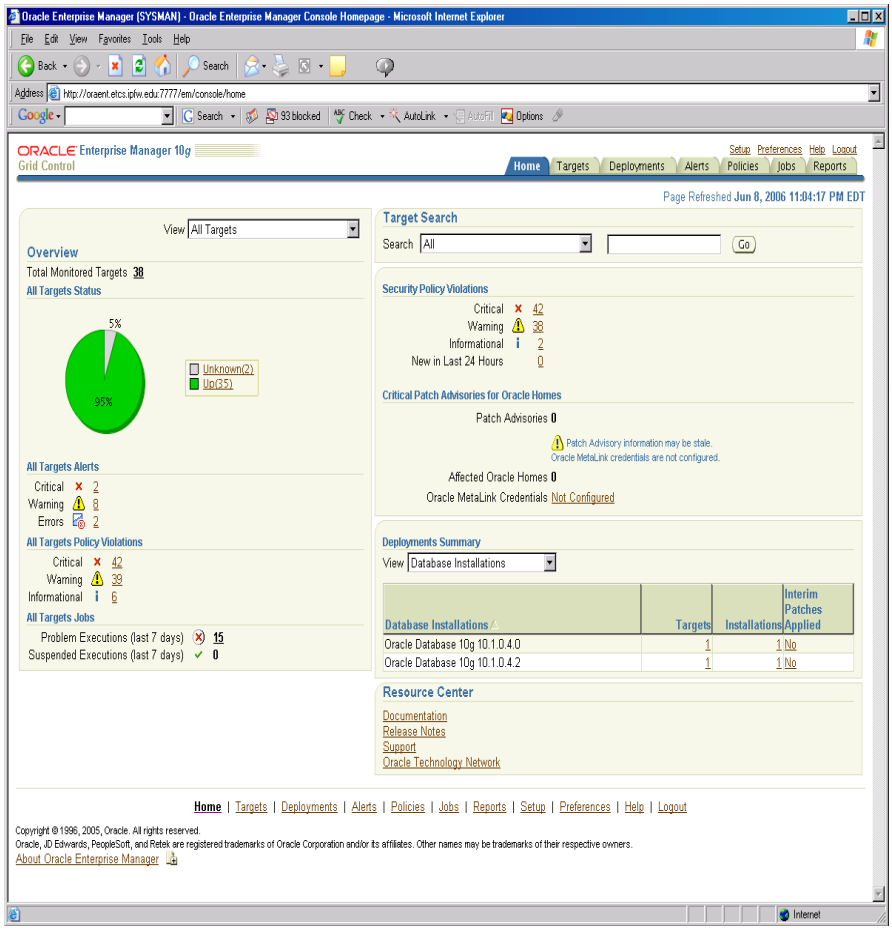
Figure 1. Structure of generalized grid and Oracle database grid



An approach is to identify a suitable activity or sweet spot that will facilitate the adoption of the grid, not necessarily in a big swoop but small scaled and incremental. Goyal and Lawande (2006) give some characteristics of a sweet spot, such as a task with a compelling use case that has broader acceptance in a company or a task with measurable goals and benefits on a small scale that could be executed within a realistic time frame. For instance, organizations may have periodic acquisition plans for software and hardware, a problem monitoring IT infrastructure, resource allocation or desire for improvement of business flows. In such situations, enterprises can acquire new, low cost modular servers during the cycle of hardware acquisition for database and application servers, deploy grid control as a solution for management constraints, or apply provisioning technologies such as transportable tablespaces and streams for resource allocation problems.

At the CS department, Fort Wayne, we were confronted with the problem of managing our Oracle servers (database, application server, storage and instrument) due to the absence of a DBA. This was a sweet spot that caused us to introduce Oracle database grid products. Oracle database grid control (ODGC–Figure 2) solved this management constraint based on the philosophy of managing many as one (Khilani, 2005; Kumar & Burleson, 2005). ODGC adopts a holistic management view to monitor diverse Oracle products, and also provision, clone and automate jobs, even providing with alerts and warnings in heterogeneous environments. Its self-monitoring features ensure that critical components of grid control are always available and functional (Khilani, 2005; Kumar & Burleson, 2005). ODGC is bundled with a management grid control console (client), a management service (middleware) and a management agent (backend) in a three-tiered architecture. It can monitor any node (target) with a management agent that collects host and database-related statistics. To add a new system to the management target list involves simply installing the management agent on that system. Thus, as an IT department grows, more serv-

Figure 2. Oracle grid control at the CS department, Fort Wayne



ers can be easily added for load balancing. In addition to the centralized management structure, the grid control provides a unified security/identity management interface based on the single sign-on technique, which allows a user to log on once to a network and then access multiple applications with a single password.

Standardization

As an attempt to achieve high productivity, organizations normally invest in the best of available technologies and processes, regardless of interoperability of the systems. This practice has caused the IT world to make huge financial outlay for interoperability (Chrabakh & Wolski, 2006). The grid approach supports using interoperable products as a way to curb variability and complexity prevalent in current data centers. According to Goyal and Lawande (2006), one way to achieve technology standardization is to limit the number of vendors and products that support industry standards for interoperability. This approach reduces the amount of resources for deployment and testing of single product lines and paves the way for easier future consolidation. While technology standardization focuses on reduction of heterogeneous hardware and software (vendors and products), process standardization is geared toward the architecture for development, deployment and management. Process standardization reduces the variability of a system life cycle and associated errors. In this vein, IT process standards such as ITIL could be adopted as a guide. This approach ensures that proven best management practice is observed in the life cycle of systems, thus streamlining activities and reducing cost.

To achieve standardization of the hardware and software at the CS department, Fort Wayne, the Oracle grid products are only operated on Linux servers (much cheaper than symmetric multiprocessor–SMP) and Linux operating systems (Red Hat). This ensures uniform maintenance of the hardware and software resources, thereby easing administration.

Consolidation

In consolidating technology and IT processes, organizations need to initiate a long term strategy to strengthen all aspects of grid-enabling IT infrastructure, by reducing fragmented data centers. A few data centers eases management in one place and encourages economy of scale. Consolidation supports the methodology of grid computing at every step of system development, deployment and management, for example, consolidating storage with integrated structure like a storage area network (SAN) improves performance. Furthermore, modular servers can be used to consolidate databases and application servers. According to Goyal and Lawande (2006), integration may be physical by moving data to one database or logical by using products like data hubs. This substantially reduces power, cooling, space requirements and management costs. Consolidation maximizes the gains of standardization. Currently, CS department at Fort Wayne is implementing a storage area network that will provide a storage subsystems to multiple hosts at the same time.

Automation

Automation of tasks and processes reduces human errors and lowers costs in an enterprise. After standardization and consolidation, automation is easier to implement. For instance, standardized operating systems and storage layouts can be automated, as scripts and automation tools can be easily deployed. However, best management practice favors automating simple, repetitive and laborious tasks at the onset. Such tasks can be identified beginning from initial deployments to maintenance of systems.

At the CS department, Fort Wayne, patches, alerts, testing, deployment of servers and maintenance can be scheduled for automation with the Oracle grid control. A valuable automation is that of patching, which allows downloads from the Oracle's Metalink Web site. This feature helps keep the systems in sync and more secured.

FUTURE TRENDS

Enterprise database grid computing is increasingly being adopted by major corporations. The acceptance of this trend will continue in the future as its benefits become more obvious. More than any other factor, the fact that vendors are increasingly offering grid-enabled technologies will encourage rapid adoption in the future. Furthermore, grid computing is evolving; and many relevant standards are currently being developed. For instance, there is currently a lack of established standards for dynamic and autonomic resource provisioning, which hinders vendor interoperability (Pollock & Hodgson, 2004; Wolski, 2003). Improvements in such standards will be incorporated in new grid products. However, the current lack of grid standards should not deter the adoption of grid computing. Enterprises can adopt grid methodology (a service-centric view on IT infrastructure: Yan & Chapman, 2006) and leverage their current investments, because companies are now creating new products that are grid-enabled (Dooley, Milfeld, Guiang, Pamidighantam, & Allen, 2006). Furthermore, there will be more developments in service-oriented architecture that will certainly impact grid applications. Ultimately, enterprise database grids will benefit from improvements in semantic grid, which will enable meaningful information processing or parsing of deep relationships between data without human intervention.

CONCLUSION

The emerging grid infrastructure will substantially reduce the cost of deploying and managing enterprise applications. With further grid product offerings and improvements in hardware and grid protocols, enterprise grid computing will be embraced by many organizations in the future. It is recommendable to adopt grid technology starting with a sweet spot and then incrementally integrating all aspects of the enterprise. Once the benefits of the grid are demonstrated on a small scale, scaling out can be initiated in subsequent operations. Staff resistance to adoption can be overcome through effective communication, team building and incentives to adaptation. Oracle offers a family of products that support database grid for an efficient data-driven business. Enterprise database grid is not a disruptive technology because it leverages existing investments and best practices in an organization.

REFERENCES

Berman, F., Fox, G., & Hey, T. (2003). The grid: Past, present, future. In F. Berman, G.C. Fox, & A.J.G. Hey (Eds.), *Grid computing* (pp. 51-63). New York: John Wiley & Sons.

Chrabakh, W., & Wolski, R. (2006). GridSAT: Design and implementation of a computational grid application. *Journal of Grid Computing, 4*(2), 177-193.

Dooley, R., Milfeld, K., Guiang, C., Pamidighantam, S., & Allen, G. (2006). From proposal to production: Lessons learned developing the computational chemistry grid cyberinfrastructure. *Journal of Grid Computing, 4*(2), 195-208.

Finkelstein, A., Gryce, C., & Lewis-Bowen, J. (2004). Relating requirements and architectures: A study of data-grids. *Journal of Grid Computing, 2*(3), 207-222.

Foster, I. (2003). The Grid: A new infrastructure for 21st century science. F. Berman, G.C. Fox, & A.J.G. Hey (Eds.), *Grid computing* (pp. 65-100). New York: John Wiley & Sons.

Goyal, B., & Lawande, S. (2006). *Enterprise grid computing with Oracle*. New York: McGraw-Hill.

Joseph, J., & Fellenstein, C. (2004). *Grid computing*. Upper Saddle River, NJ: Prentice Hall.

Khilani, A. (2005). *Oracle enterprise manager 10g grid control: Features for database management*. Retrieved December 12, 2007, from http://www.oracle.com/technology/tech/grid/index.html

Kumar, A.R., & Burleson, D. (2005). Easy *Oracle automation*. Kittrell, NC: Rampant Techpress.

Kusnetzky, D., & Olofson, C.W. (2004). *Oracle 10g: Putting grids to work*. IDC White Paper. Retrieved December 12, 2007, from http://www.oracle.com/technology/tech/grid/index.html

Murch, R. (2004). *Autonomic computing*. Upper Saddle River, NJ: Prentice Hall.

Oracle Inc. (2005). *Grid computing with Oracle*. Technical white aper. Retrieved December 12, 2007, from http://www.oracle.com/technology/tech/grid/index.html

Pollock, J.T., & Hodgson, R. (2004). *Adaptive information: Improving business through semantic interoperability, grid computing, and enterprise integration*. New York: John Wiley.

Watson, P. (2003). Databases and the grid. In F. Berman, G.C. Fox, & A.J.G. Hey (Eds.), *Grid computing* (pp. 363-384). New York: John Wiley.

Wolski, R. (2003). Experiences with predicting resource performance online in computational grid settings. *ACM SIGMETRICS Performance Evaluation Review, 30*(4), 41-49.

Yan, Y., & Chapman, B.M. (2006). Campus grids meet applications: Modeling, metascheduling and integration. *Journal of Grid Computing, 4*(2), 159-175.

KEY TERMS

Applications Grid: It shares and reuses application code but uses software technologies like service-oriented architectures that facilitate sharing business logic among multiple applications.

Grid Computing: A style of computing that dynamically pools IT resources together for use based on resource need. It allows organizations to provision and scale resources as needs arise, thereby preventing the underutilization of resources (computers, networks, data archives, instruments).

Information Grid: This grid shares information across multiple consumers and applications. It unlocks fragmented data from proprietary applications by treating information as a resource to be shared across the grid.

Infrastructure Grid: This grid pools, shares and reuses infrastructure resources such as hardware, software, storage and networks across multiple applications.

Provisioning: The allocation of resources to consumers on demand. A system determines specific need of the consumer and provides the resources as requested.

Semantic Web: Information processing model in which computers using resource description framework (RDF) and other technologies can explicitly associate meanings or parse relationships between data without human intervention.

Service-Oriented Architecture (SOA): This is a form of software design that allows different applications to interact in business processes regardless of specific technology like programming languages and operating systems

Silos/Islands of Applications/Computing: Condition whereby servers or computing resources are idle most of the time when the peak load is not reached. Such IT systems are not designed to share resources with each other, thus creating islands of information and computing infrastructure within a single enterprise.

Virtualization: A form of abstraction that provides location- and technology-transparent access of resources to the consumer. It decouples the tight connections between providers and consumers of resources, thus allowing sharing of the same resources by multiple users as needs arise.

Database Integrity Checking

Hendrik Decker
Universidad Politécnica de Valencia, Spain

Davide Martinenghi
Free University of Bozen/Bolzano, Italy

INTRODUCTION

Integrity constraints (or simply "constraints") are formal representations of invariant conditions for the semantic correctness of database records. Constraints can be expressed in declarative languages such as datalog, predicate logic, or SQL. This article highlights the historical background of integrity constraints and the essential features of their simplified incremental evaluation. It concludes with an outlook on future trends.

BACKGROUND

Integrity has always been an important issue for database design and control, as attested by many early publications (e.g., Bernstein & Blaustein, 1982; Bernstein, Blaustein, & Clarke, 1980; Codd, 1970, 1979; Eswaran & Chamberlin, 1975; Fraser, 1969; Hammer & McLeod, 1975; Hammer & Sarin, 1978; Nicolas, 1978, 1982; Wilkes, 1972); later ones are too numerous to mention. Expressing database semantics as invariant properties persisting across updates had first been proposed by Minsky (1974). Florentin (1974) suggested expressing integrity constraints as predicate logic statements. Stonebraker (1975) proposed formulating and checking integrity constraints declaratively as SQL-like queries.

Functional dependencies (Armstrong, 1974; Codd, 1970) are a fundamental kind of constraints to guide database design. Referential integrity has been part of the 1989 SQL ANSI and ISO standards (McJones, 1997). The SQL2 standard (1992) introduced the CHECK and ASSERTION constructs (i.e., table-bound and table-independent SQL query conditions) as the most general means to express integrity constraints declaratively (Date & Darwen, 1997). Since the 1990s, uniqueness constraints, foreign keys, and complex queries involving EXISTS and NOT became common features in commercial databases. Thus, arbitrarily general and complex integrity constraints can now be expressed and evaluated in most relational databases. However, most of them offer efficient support only for the following three simple kinds of declarative constraints:

- **Domain Constraints:** Restrictions on the permissible range of attribute values of tuples in table columns, including scalar SQL data types and subsets thereof, as well as options for default and null values.
- **Uniqueness Constraints:** As enforced by the UNIQUE construct on single columns, and UNIQUE INDEX and PRIMARY KEY on any combination of one or several columns in a table, preventing multiple occurrences of values or combinations thereof.
- **Foreign Key Constraints:** For establishing a relationship between the tuples of two tables, requiring identical column values. For instance, a foreign key on column emp of relation works_in requires that the emp value of each tuple of works_in must occur in the emp_id column of table employee, and that the referenced column (emp_id in the example) has been declared as primary key.

For more general constraints, SQL manuals usually recommend using procedural triggers or stored procedures instead of declarative constructs. This is because such constraints may involve nested quantifications over huge extents of several tables. Thus, their evaluation can easily become prohibitively costly. However, declarativity does not need to be sacrificed for efficiency, as shown by many methods of simplified integrity checking as cited in this survey. They are all based on the seminal paper (Nicolas, 1982).

SIMPLIFIED INCREMENTAL INTEGRITY CHECKING

A common idea of all integrity checking methods is that not all constraints need to be evaluated, but at most those that are possibly affected by the incremental change caused by database updates or transactions. Anticipating updates by patterns, most incremental integrity checking methods allow for simplifications of constraints to be generated already at schema compilation time. Such compiled simplifications are parametric conditions to be instantiated, possibly further optimized, and evaluated upon given update requests. For generating them, only the database schema, the integrity constraints, and the update patterns are needed as input.

Their evaluation, however, may involve access to the stored data at update time. Methods that generate compiled simplifications are described, for example, by Christiansen and Martinenghi (2006), Decker (1987), and Leuschel and De Schreye (1998). For unanticipated ad-hoc updates, the generation of simplifications takes place at update time. Optimizations for efficient evaluation of simplified constraints are addressed, for example, by Sheu & Lee (1987).

Simplifications can be distinguished by the database state in which they are evaluated. *Post-test* methods must evaluate their simplifications in the *new*, updated state, for example, Decker and Celma (1994), Grant and Minker (1990), Lloyd, Sonenberg, and Topor (1987), Nicolas (1982), and Sadri and Kowalski (1988). *Pre-test* approaches, for example, Bry, Decker, and Manthey (1988), Christiansen and Martinenghi (2006), Hsu and Imielinski (1985), McCune and Henschen (1989), and Qian (1988), only access the *old* state before the update, that is, they need not execute the update prematurely, since undoing an updated state if integrity is violated is costly. In case of integrity violation, the eagerness of pre-tests to avoid rollbacks is a clear performance advantage over post-tests.

For convenience, a finite set of constraints imposed on a database D is called an *integrity theory* of D. For a database D and an integrity theory IC, let $D(IC)$ = *satisfied* denote that IC is satisfied in D, and $D(IC)$ = *violated* that it is violated. Further, for an update U, let D^U denote the updated database. Any simplification method M can be formalized as a function that takes as input a database, an integrity theory and an update, and outputs either *satisfied* or *violated*. Thus, soundness and completeness of M can be stated as follows:

Let D be any database, IC any integrity theory, and U any update. Suppose that $D(IC)$ = *satisfied*. Then, an integrity checking method M is *sound* if the following holds:

If $M(D, IC, U)$ = *satisfied* then $D^U(IC)$ = *satisfied*.

It is *complete* if the following holds:

If $D^U(IC)$ = *satisfied* then $M(D, IC, U)$ = *satisfied*.

This formalism is applicable to most integrity checking methods in the literature. Many of them are sound and complete for significant classes of relational and deductive databases, integrity theories, and updates. Some methods, however, are only shown to be sound, that is, they provide sufficient conditions that guarantee integrity satisfaction of D^U, for example, Gupta, Sagiv, Ullman, and Widom (1994); further checking is required if these conditions are not satisfied. The main advantage is that the evaluation of $M(D, IC, U)$ is typically much simpler than that of $D^U(IC)$.

Most integrity checking methods can be described by distinguishing three (possibly interleaved) phases, namely the *generation*, *optimization*, and *evaluation* of simplified tests. Next, these phases, numbered I, II, III, are illustrated by an example.

EXAMPLE

Consider a relational database with tables for workers and managers, defined as follows:

```
CREATE TABLE(worker(CHAR name, CHAR department))

CREATE TABLE(manager (CHAR name)).
```

Suppose there is an integrity constraint requiring that no worker is a manager, expressed as a denial by the following SQL condition, which forms the body of a related SQL assertion:

```
NOT EXISTS (SELECT * FROM worker, manager WHERE
worker.name = manager.name).
```

If the number of workers and managers is large, then checking whether this constraint is violated or not can be very costly. The number of facts to be retrieved and tested is in the order of the cardinality of the cross product of worker and manager, whenever the constraint is checked. Fortunately, however, the frequency and amount of accessing stored facts can be significantly reduced when going through phases I -III. Beforehand, a possible objection at this stage should be dealt with.

SQL programmers might feel compelled to point out that the previous constraint is probably much easier checked by some trigger such as the following one in MS SQL Server syntax:

```
CREATE TRIGGER no_worker_manager ON worker FOR
INSERT : IF EXISTS

(SELECT * FROM inserted, manager WHERE inserted.name
= manager.name) ROLLBACK.
```

Its evaluation would only need to access manager and a cached relation inserted containing the row to be inserted to worker, but not the stored part of worker. However, it is easily overlooked that the sample integrity constraint entails that somebody who is promoted to a manager must not be a worker, thus necessitating a second trigger for insertions into manager. In general, each occurrence of each atom occurring in a constraint requires a separate trigger, and it is by far not always as obvious as in the simple previous example how they should look. Apart from being error-prone, hand-

coded triggers may also bring about unpredictable effects of mutual interactions that are hard to control. As opposed to that, the generic approach illustrated next can be fully automatized.

Now, let INSERT INTO worker VALUES ('Fred', 'sales') be an update. Then, phases I-III involve the following.

I. Generation

The constraint that no worker must be manager is clearly relevant for the given update and hence must be checked, but only for the newly inserted worker Fred. Any integrity constraint that is not relevant for insertions into the worker table need not be checked. For example, a constraint requiring that each department must have some least number of workers need not be checked for insertions but only for deletions in the worker table. Also, all constraints that do not involve the relation worker are not relevant.

Hence, as a result of phase I, the SQL condition

```
NOT EXISTS (SELECT * FROM worker, manager WHERE
worker.name = manager.name)
```

can be simplified to the following much less expensive expression:

```
NOT EXISTS (SELECT * FROM worker, manager WHERE

worker.name = 'Fred' AND worker.name = manager.
name).
```

If the worker table is involved in definitions of database views, then there may be implicit update consequences on such views. These, in turn, need to be run through I-III.

II. Optimization

Since the existence of a worker satisfying the subcondition worker.name = 'Fred' is assured by the update, the simplified condition in I can be further optimized to

```
NOT EXISTS (SELECT * FROM manager WHERE name
= 'Fred').
```

III. Evaluation

Evaluation of the query whether Fred is a manager means to look up a single fact in a stored relation. That, of course, is much less costly than evaluating all integrity constraints in their full generality, as would be necessary without having done I-III.

The previous example is very simple. However, by running phases I-III, the same proportions of simplification are obtained systematically for arbitrarily complex constraints.

FUTURE TRENDS

Future trends to be reckoned with for integrity checking can be grouped by the following thematic issues: growing demands, business rules, closing gaps between theory and practice, distributed and replicated databases, agents, extensions of the relational model, semi-structured data, XML, Web databases, and inconsistency tolerance. Editions in 1999 and 2003 of the SQL standard have tried to do justice to several of these trends by proposing norms for non-scalar data types, markup annotations, recursive view definitions, and triggers, which, however, have hardly been taken up uniformly by database vendors (cf. Gorman, 2001).

The importance of database integrity is likely to grow further, hand in hand with increasing concerns for the quality and reliability of data. Also, the declarativity of data integrity is bound to become more important because of the growing complexity of data and applications, the development and maintenance of which would otherwise become too troublesome. A success story about commercial use cases of declarative integrity constraints is the growing demand of business rules (Date, 2000; Ross, 2005). Another prospering application area, based on the use of integrity constraints in semantic query optimization (Grant & Minker, 1990), is consistent query answering, as indicated by Bertossi (2006).

Yet, commercial databases lag behind the state of the art of integrity checking in academia. Besides a lack of support for the simplification of ASSERTION statements, also the support for declarative constraints on views is scant, if not completely missing, in many database products. In general, technologies related to integrity constraints that are still to be transferred from theory to practice are aplenty (cf., e.g., Decker, 1998, 2002). Not only in practice, but also in theory, a void that continues to gape is related to simplified integrity checking for concurrent transactions. Locking policies and coordination algorithms, used to avoid or resolve conflicts and deadlocks of concurrent read-write accesses, may cause a considerable overhead. Performance is burdened even more by additional locks that may be necessary for integrity checking. A locking policy for checking the integrity of concurrent transactions and possible mitigations of the associated overhead are discussed by Martinenghi and Christiansen (2005).

Not only with regard to concurrency, but also in general, integrity support in decentralized networks of databases is even worse than for centralized systems. For instance, no declarative way of ensuring referential integrity across tables stored in different sites exists. This and other problems of integrity checking in distributed databases are addressed,

(e.g., Bright & Chandy, 2003; Ibrahim, 2002, 2006). Some relief that may make up for part of this deficit can be expected from a convergence of distributed databases and agents technology. An approach to handle integrity constraints in this context is described, for example, by Sadri, Toni, and Torroni (2002). Both theory and practice of integrity support need to advance also for replicated databases. Beyond distributed databases, there is the problem of the integrity of data on the Web, which certainly is going to receive more attention in the future; for a primer, see work by Aldana, Yagüe, and Gómez (2002).

In recent years, there has been a surge of interest in the design and control capacities of integrity constraints in XML-based data models and languages (Vianu, 2003). An initial focus on primary and foreign keys (Fan & Libkin, 2002) currently is being extended to more general kinds of constraints for XML (cf., e.g., Fan, 2005; Zhuge & Xing, 2005).

One of the reasons for the lack of support for database integrity in practice is an alleged inapplicability of constraint checking methods in the presence of inconsistencies. Simplification methods usually assume that, for checking integrity incrementally, the state before the update must satisfy integrity. Clearly, this assumption is not very realistic since hardly any real-life database of significant size can be expected to be completely free of any inconsistency. Fortunately, however, this assumption can be abandoned. As shown by Decker and Martinenghi (2006), many (though not all) methods provide simplified conditions that are necessary and sufficient for checking that all cases of integrity constraints that have been satisfied before the update will remain so afterward, even in presence of any number of violated cases. Ramifications of this result are expected to further boost the adoption of integrity checking technology in practice.

CONCLUSION

We have outlined basic concepts and principles for checking the semantic integrity of stored data that are more advanced than the usual SQL support for data integrity, but can be expressed in conventional SQL and processed by standard SQL engines. In the literature, integrity constraints are often represented in the language of first-order predicate logic. A systematic translation of constraints expressed in predicate logic into SQL is described by Decker (2002). There, also a fully automated and provably correct mechanism for translating declarative constraints into efficient triggers is specified. It takes advantage of simplifications as outlined previously. More details about general principles of integrity checking are documented by Martinenghi, Christiansen, and Decker (2006).

REFERENCES

Aldana, J., Yagüe, I., & Gómez, L. (2002). Integrity issues in the Web: Beyond distributed databases. In J. Doorn & L. Rivero (Eds.), *Database integrity: Challenges and solutions* (pp. 293-321). Hershey, PA: Idea Group Publishing.

Armstrong, W. (1974). Dependency structures of database relationships. In J. L. Rosenfeld (Ed.), *Proceedings of IFIP '74* (pp. 580-583). Stockholm, Sweden: North-Holland.

Bernstein, P., & Blaustein, B. (1982). Fast methods for testing quantified relational calculus assertions. In M. Schkolnick (Ed.), *Proceedings of SIGMOD'82* (pp. 39-50). Orlando, FL: ACM Press.

Bernstein, P., Blaustein, B., & Clarke, E. (1980). Fast maintenance of semantic integrity assertions using redundant aggregate data. In W. W. Armstrong et al. (Eds.), *Proceedings of the Sixth VLDB* (pp. 126-136). Montreal: IEEE-CS.

Bertossi, L. (2006). Consistent query answering in databases. *ACM SIGMOD Record, 35*(2), 68-77.

Bright, J., & Chandy, J. (2003). Data integrity in a distributed storage system. In J. R. Arabnia & Y. Mun (Eds.), *Proceedings of PDPTA'03* (pp. 688-694). Las Vegas: CSREA Press.

Bry, F., Decker, H., & Manthey, R. (1988). A uniform approach to constraint satisfaction and constraint satisfiability in deductive databases. In J. W. Schmidt, S. Ceri, & M. Missikoff (Eds.), *Proceedings of the First EDBT* (LNCS 303, pp. 488-505). Venice: Springer.

Christiansen, H., & Martinenghi, D. (2006). On simplification of database integrity constraints. *Fundamenta Informaticae, 71*(4), 371-417.

Codd, E. (1970). A relational model of data for large shared data banks. *Commun. ACM, 13*(6), 377-387.

Codd, E. (1979). Extending the database relational model to capture more meaning. *ACM Transactions on Database Systems, 4*(4), 397-434.

Date, C. (2000). *What, not how: The business rules approach to application development.* Boston: Addison-Wesley.

Date, C., & Darwen, H. (1997). *A guide to the SQL standard*. Boston: Addison-Wesley.

Decker, H. (1987). Integrity enforcement on deductive databases. In L. Kerschberg (Ed.), *Experts database systems* (pp. 381-395). Charleston, SC: Benjamin Cummings.

Decker, H. (1998). Some notes on knowledge assimilation in deductive databases. In B. Freitag, H. Decker, M. Kifer, & A. Voronkov (Eds.), *Transactions and change in logic*

databases (LNCS 1472, pp. 249-286). Schloss Dagstuhl, Germany: Springer.

Decker, H. (2002). Translating advanced integrity checking technology to SQL. In J. Doorn & L. Rivero (Eds.), *Database integrity: Challenges and solutions* (pp. 203-249). Hershey, PA: Idea Group Publishing.

Decker, H., & Celma, M. (1994). A slick procedure for integrity checking in deductive databases. In P. Van Hentenryck (Ed.), *Proceedings of the 11th ICLP* (pp. 56-469). Boston: MIT Press.

Decker, H., & Martinenghi, D. (2006). A relaxed approach to integrity and inconsistency in databases. In M. Hermann & A. Voronkov (Eds.), *Proceedings of the 13th LPAR* (LNAI 4246, pp. 287-301). Berlin; Heidelberg: Springer-Verlag.

Eswaran, K., & Chamberlin, D. (1975). Functional specifications of a subsystem for database integrity. In Douglass S. Kerr (Ed.), *Proceedings of the First VLDB* (pp. 48-68). Framingham, MA: ACM Press.

Fan, W. (2005). XML constraints: Specification, analysis, and applications. In H. Christiansen & D. Martinenghi (Eds.), *Proceedings of the 16th DEXA Workshop, LAAIC'05* (pp. 805-809). Copenhagen, Denmark: IEEE Computer Society.

Fan, W., & Libkin, L. (2002). On XML integrity constraints in the presence of DTDs. *Journal of the ACM, 49*(3), 368-406.

Florentin, J. (1974). Consistency auditing of databases. *The Computer Journal, 17*(1), 52-58.

Fraser, A. (1969). Integrity of a mass storage filing system. *The Computer Journal, 12*(1), 1-5.

Fuxman, A., & Miller, R. (2005). First-order query rewriting for inconsistent databases. In T. Eiter & L. Libkin (Eds.) *Proceedings of the 10th ICDT* (LNCS 3363, pp. 337-351). Edinburgh, UK: Springer.

Gorman, M. (2001). Is SQL really a standard anymore? *Whitemarsh Information Systems Corp.* Retrieved November 12, 2006, from www.ncb.ernet.in/education/modules/dbms/SQL99/issqlarealstandardanymore.pdf

Grant, J., & Minker, J. (1990). Integrity constraints in knowledge based systems. In H. Adeli (Ed.), *Knowledge engineering* (Vol. II) (pp. 1-25). New York: McGraw-Hill.

Gupta, A., Sagiv, Y., Ullman, J., & Widom, J. (1994). Constraint checking with partial information. In M. Yannakakis (Ed.) *Proceedings of the 13th PODS* (pp. 45-55). Minneapolis: ACM Press.

Hammer, M., & McLeod, D. (1975). Semantic integrity in relational database systems. In D. S. Kerr (Ed.), *Proceedings of the First VLDB* (pp. 25-47). Framingham, MA: ACM Press.

Hammer, M., & Sarin, S. (1978). Efficient monitoring of database assertions (abstract). In E. I. Lowenthal & N. B. Dale (Eds.), *Proceedings of SIGMOD'78* (p. 159). Austin, TX: ACM Press.

Hsu, A., & Imielinski, T. (1985). Integrity checking for multiple updates. In S. B. Navathe (Ed.), *Proceedings of SIGMOD'85* (pp. 152-168). Austin, TX: ACM Press.

Ibrahim, H. (2002). A strategy for semantic integrity checking in distributed databases. In J.-P. Sheu (Ed.), *Proceedings of the Ninth ICPDS* (pp. 139-144). Taiwan: IEEE Computer Society.

Ibrahim, H. (2006). Checking integrity constraints—How it differs in centralized, distributed and parallel databases. In H. Christiansen & D. Martinenghi (Eds.), *Proceedings of the 17th DEXA Workshop, LAAIC'06* (pp. 563-568). Karkow, Poland: IEEE Computer Society.

Leuschel, M., & De Schreye, D. (1998). Creating specialised integrity checks through partial evaluation of meta-interpreters. *Journal of Logic Programming, 36*(2), 149-193.

Lloyd, J., Sonenberg, L., & Topor, R. (1987). Integrity constraint checking in stratified databases. *Journal of Logic Programming, 4*(4), 331-343.

Martinenghi, D., & Christiansen, H. (2005). Transaction management with integrity checking. In K. V. Andersen, J. K. Debenham, & R. Wagner (Eds.), *Proceedings of the 16th DEXA* (LNCS 3588, pp. 606-615). Copenhagen, Denmark: Springer.

Martinenghi, D., Christiansen, H., & Decker, H. (2007). Integrity checking and maintenance in relational and deductive databases and beyond. In Z. Ma (Ed.), *Intelligent databases: Technologies and applications* (pp. 238-285), Hershey, PA: Idea Group Publishing.

McCune, W., & Henschen, L. (1989). Maintaining state constraints in relational databases: A proof theoretic basis. *Journal of the ACM, 36*(1), 46-68.

McJones, P. (1997). The 1995 SQL reunion: People, projects, and politics. *SRC Technical Note 1997 – 018*. Retrieved November 12, 2006, from www.mcjones.org/System_R/SQL_Reunion_95/sqlr95.html

Minsky, N. (1974). On interaction with databases. In R. Rustin (Ed.), *Proceedings of SIGMOD'74* (Vol. 1, pp. 51-62). Ann Arbor, MI: ACM Press.

Nicolas, J-M. (1978). First order logic formalization for functional, multivalued and mutual dependencies. In E. I.

Lowenthal & N. B. Dale (Eds.), *Proceedings of SIGMOD'78* (pp. 40-46). Austin, TX: ACM Press.

Nicolas, J-M. (1982). Logic for improving integrity checking in relational databases. *Acta Informatica, 18*, 227-253.

Qian, X. (1988). An effective method for integrity constraint simplification. In *Proceedings of the Ninth ICDE* (pp. 338-345). Los Angeles, CA: IEEE Computer Society.

Ross, R. (2005). *Business rule concepts* (2nd ed.). Business Rules Solutions Publishing.

Sadri, F., & Kowalski. R. (1988). A theorem-proving approach to database integrity. In J. Minker (Ed.), *Foundations of deductive databases and logic programming* (pp. 313-362). Morgan Kaufmann.

Sadri, F., Toni, F., & Torroni, P. (2002). An abductive logic programming architecture for negotiating agents. In S. Flesca, S. Greco, N. Leone, G. Ianni (Eds.), *Proceedings of the Eighth JELIA* (LCNS 2424, pp. 419-431). Cosenza, Italy: Springer.

Sheu, P., & Lee, W. (1987). Efficient processing of integrity constraints in deductive databases. *Future Generations Computer Systems, 3*(3), 210-216.

Stonebraker, M. (1975). Implementation of integrity constraints and views by query modification. In W. Frank King (Ed.), *Proceedings of SIGMOD'75* (pp. 65-78). San Jose, CA: ACM Press.

Vianu, V. (2003). A Web odyssey: From Codd to XML. *SIGMOD Record, 32*(1), 68-77.

Wilkes, M. (1972). On preserving the integrity of databases. *The Computer Journal, 15*(3), 191-194.

Zhuge, H., & Xing, Y. (2005). Integrity theory for resource space model and its application. In W. Fan, Z. Wu, & J. Yang (Eds.), *Proceedings of the Fourth WAIM* (LNCS 3739, pp. 8-24). Hangzhou, China: Springer.

KEY TERMS

Business Rule: Statement for defining or constraining the evolution of data pertaining to an enterprise's business. Business rules can be represented and enforced by integrity constraints.

Declarative vs. Procedural: Since the evaluation of declarative integrity constraints can be very costly, potentially troublesome procedural triggers and stored procedures are often used instead. The main thrust of this article is about reducing the cost of declarative integrity constraints.

Inconsistency Tolerance: A practically indispensable property. Integrity checking is inconsistency-tolerant if the invariance of all satisfied cases of integrity constraints across updates can be ensured, even if some other cases are violated.

Integrity: Semantic consistency, that is, the correctness of stored data with regard to their intended meaning, as expressed by integrity constraints. Not to be confused with namesake issues related to data security, serializability of concurrent transactions or sound failure recovery.

Integrity Checking: Systematic test for checking whether integrity constraints remain satisfied or become violated by some update.

Integrity Constraint: Integrity constraints are invariant properties of the database that evolve via updates. Often, it is convenient to state them as denials, that is, *yes/no* queries that are required to return the empty answer in each database state.

Integrity Enforcement: Actions taken to ensure that integrity remains satisfied across database updates. If integrity is violated by some update, then that update is rejected or some other corrective action is taken.

Integrity Satisfaction, Integrity Violation: Integrity is satisfied if each integrity constraint in the database schema, queried as a denial, returns the empty answer. Integrity is violated if any one of these integrity constraints returns a non-empty answer. Then, the database is also said to be inconsistent.

Semantic Integrity: The adjective "semantic" distinguishes a set of explicitly defined integrity constraints from structural constraints that are implicitly enforced by the used data model.

Simplification: A methodological approach to reduce complexity and costs of integrity checking. Also the simplified forms of integrity constraints generated by such methods are called simplifications.

Static vs. Dynamic Integrity Constraints: Static integrity constraints are semantic properties that are invariant across evolving database states. Dynamic integrity constraints refer explicitly to several (mostly consecutive) states or to their transitions, typically involving temporal or procedural constructs.

Trigger: An SQL statement that generates an action (e.g., an update or a reset) upon some anticipated event (e.g., an update request). Triggers are a procedural means to enforce integrity.

Database Support for M-Commerce and L-Commerce

Hong Va Leong
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

M-commerce (mobile commerce) applications have evolved out of e-commerce (electronic commerce) applications, riding on recent advancement in wireless communication technologies. Exploiting the most unique aspect inherent in m-commerce, namely, the mobility of customers, l-commerce (location-dependent m-commerce) applications have played an increasingly important role in the class of m-commerce applications. All e-commerce, m-commerce, and l-commerce applications rely on the provision of information retrieval and processing capability. L-commerce applications further dictate the maintenance of customer and service location information. Various database systems are deployed as the information source and repository for these applications, backed by efficient indexing mechanisms, both on regular data and location-specific data.

Bean (2003) gave a good report on supporting Web-based e-commerce with XML, which could be easily extended to m-commerce. An m-commerce framework, based on JINI/XML and a workflow engine, was defined by Shih and Shim (2002). Customers can receive m-commerce services through the use of mobile devices such as pocket PCs, PDAs, or even smart phones. These mobile devices together with their users are often modeled as mobile clients. There are three types of entities central to m-commerce and l-commerce applications: mobile device, wireless communication, and database. In this article, we focus our discussion on mobile-client enabled database servers, often referred to as mobile databases. Mobile databases maintain information for the underlying m-commerce and l-commerce applications in which mobile devices serve as the hardware platform interfacing with customers, connected through wireless communication.

Location is a special kind of composite data ranging from a single point, a line, a poly-line, to a shape defining an area or a building. In general, locations are modeled as spatial objects. The location of a static point of interest, such as a shop, is maintained in a database supporting spatial features and operations, often a spatial database (Güting, 1994). The location of a moving object, like a mobile customer, needs to be maintained in a moving object database (Wolfson, Sistla, Xu, Zhou, & Chamberlain, 1999), a database that supports efficient retrieval and update of object locations. To enable l-commerce, both spatial databases and moving object databases need to support location-specific query processing from mobile clients and location updates they generated.

The two major types of data access requirements for a mobile database are data dissemination and dedicated data access. Data dissemination is preferred, since it can serve a large client population in utilizing the high bandwidth downlink channel to broadcast information of common interest, such as stock quotations, traffic conditions, or special events. On the other hand, dedicated data access is conveyed through uplink channels with limited bandwidth. To disseminate database items effectively, the selected set of hot database items can be scheduled as a broadcast disk (Acharya, Alonso, Franklin, & Zdonik, 1995). Proper indexes can be built to facilitate access to broadcast database items (Imielinski & Badrinath, 1994). Redundancy can be included in data (Leong & Si, 1995) and index (Tan & Ooi, 1998) to combat the unreliability of wireless communication.

For dedicated data access, queries and updates to databases are transmitted from the client to the server. L-commerce services involve processing of location-dependent queries (Madria, Bhargava, Pitoura, & Kumar, 2000). The high frequency of updates to the location of moving objects calls for special indexing technique. The call-to-mobility ratio serves as a good indicator on the tradeoff of indexing mechanisms. The moving object databases should enable efficient execution of queries such as k-nearest neighbor, reversed nearest neighbor (Benetis, Jensen, Karĉiauskas, & Ŝaltenis, 2006), and nearest surrounder search (Lee, Lee, & Leong, 2006). In addition, they should support continuous queries (Prabhakar, Xia, Kalashnikov, Aref, & Hambrusch, 2002), such as continuous k-nearest neighbor, being executed continuously and returning location-dependent results (Lee, Leong, Zhou, & Si, 2005). Reversing the role of query and data, it is equally important to process data streams effectively (Babu & Widom, 2001) such as incoming sensor data streams (Mokbel, Xiong, Hammad, & Aref, 2005) for traffic monitoring in navigational applications.

A related and interesting research problem is the location privacy of a mobile client. For instance, the application server should not be able to deduce the exact location of Alice, when she raises a query to look for a nearest restaurant on the State Street. Yet, the information returned to Alice should enable her to determine the nearest restaurant. Location cloaking technique (Gedik & Liu, 2005) and location anonymizer (Mokbel, Chow, & Aref, 2006) would be used to ensure a

form of k-anonymity, such that Alice is indistinguishable from other k-1 clients around the State Street.

BACKGROUND

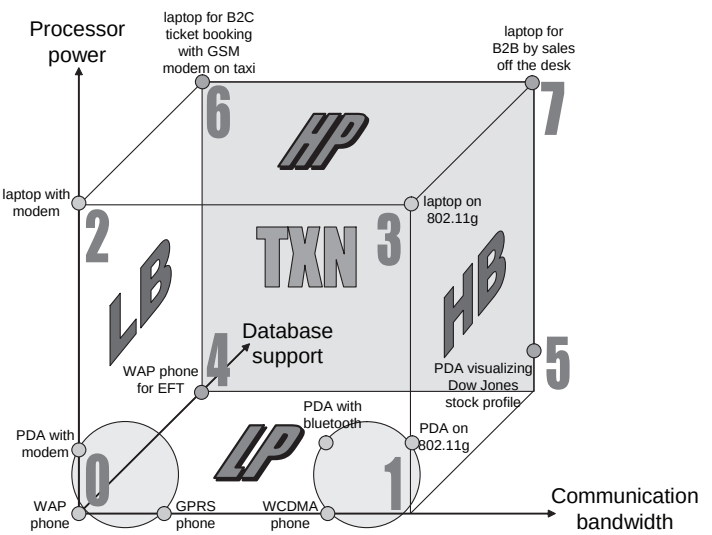
The three fundamental elements for m-commerce applications, namely, mobile device, wireless communication, and database support can be considered orthogonal. First, the variety of mobile devices differs vastly in computational power, ease of programming, interoperability of operating environments, and support for auxiliary devices. Some mobile clients are based on high-end laptops, while others are based on low-end PDAs or cellular phones. Second, wireless communication offers varying bandwidth and reliability, based on low-bandwidth and unreliable GSM connections, medium-bandwidth GPRS/EDGE and Bluetooth connections, or high-bandwidth 802.11g and WCDMA/CDMA2000 connections. Third, the database may be as primitive as a file system or simple relational database like MS Access, or as complex as the high performance Oracle with transactional and spatial data support. Transactions ensure a sequence of database operations to be executed consistently. This leads to a "cube"-like taxonomy as shown in Figure 1. The support of l-commerce requires a new location maintenance module at the database. However, for most practical applications involving the location of moving objects, transactional access is not required on the location, owing to the inherent imprecise nature of the changing location over time.

In Figure 1, the taxonomy for m-commerce and l-commerce support is displayed. Planes LP and HP represent the low computing power equipment and high computing power equipment respectively, whereas planes LB and HB reflect the availability of low and high communication bandwidth. With the availability of transactions in the TXN plane, this gives rise to eight different regions.

Region zero represents the support of standard file or simple database access from PDA connecting through low-speed modem or phone. Processing is basically performed at the server, since it is too expensive for clients to support complex mechanism. To reduce bandwidth consumption, information distillation/extraction (Cowie & Lehnert, 1996) may be performed to reduce the amount of information transmitted. Simple client/server data access paradigm suffices. Region one assumes an improved wireless network, with large scale WCDMA/CDMA2000 (3G) or small scale 802.11g (WiFi). Recent 802.16 (WiMAX) development and deployment lead to improved bandwidth in medium scale mobile environment. As a result, data access is more effective and conventional client/server data processing techniques can be adopted in a rather straightforward manner.

Region two corresponds to a mobile client with higher computational power. Information transmitted can be transcoded to reduce the bandwidth consumption. Interactive and intelligent mechanisms such as multi-resolution browsing (Leong & Si, 2005) can be employed. Database items are cached to combat the low communication bandwidth, unreliable communication, and frequent disconnection. Research work addressing this issue was pioneered by the Coda file system in 1992 (Satyanarayanan, 2002), in which files are cached by clients and updates made during client disconnection are reintegrated upon reconnection. Caching in an object-oriented database was studied by Chan, Leong, Si, and Wong (1999). Configurations in region three allow

Figure 1. Taxonomy on m-commerce and l-commerce support



easy access to data from server. With ample bandwidth and processing power, prefetching of database items allows the preparation for potential network disconnection (Jing, Helal, & Elmagarmid, 1999). Numerous research works on mobile data access have been conducted with respect to regions two and three.

Plane TXN represents the transactional equivalence of the four regions. Regions four and five involve the use of PDAs or phones to access databases in a transactional manner. Owing to the low device capability, the only effective mechanism is to execute the transaction at the server. Clients only implement the user interface, supplying the required data and displaying the result sets. Information distillation could be needed for the low bandwidth configurations in region four. The use of a proxy server helps to simplify the client/server design, since the proxy will be responsible for contacting different parties involved in the transaction. Finally, for regions six and seven, there are a lot more research potentials, since the client is more powerful in executing more complex algorithms. For instance, it would be appropriate to implement on the client a variant of the optimistic concurrency control protocol (Bernstein, Hadzilacos, & Goodman, 1987) for region six and two phase locking with lock caching (Franklin, Carey, & Livny, 1997) for region seven configurations.

In practical m-commerce applications, clients are moving around. This results in l-commerce applications, involving location-dependent queries (Madria et al., 2000). The configuration and taxonomy remain similar, except for better database support need. Research and application focus should be on the low communication bandwidth, as well as on client mobility. This corresponds to plane LB. Under such configurations, a large client population communicates with database servers over wireless communication channels (Alonso & Korth, 1993). A geographical information system component enables location-dependent queries to be resolved through geo-referencing (Choy, Kwan, & Leong, 2000). Moving object databases can keep track of the whereabout of clients (Wolfson et al., 1999) effectively.

STRONG DATABASE SUPPORT

M-commerce and l-commerce applications involve access to one or more databases, often being accessed concurrently by multiple clients. Location management is an element inherent in l-commerce. Efficient operations through moving object databases should be provided (Wolfson et. al., 1999). Transactions are executed to ensure the database consistency despite concurrent access. The correctness criterion of serializability on concurrent transactions can be enforced through concurrency control protocols. Concurrency control protocols in a client/server or a mobile environment can be classified according to their nature (Franklin et al., 1997; Jing et al., 1999). In m-commerce and l-commerce applications, simultaneous access to multiple databases, which are administered by different organizations, is a norm rather than an exception. One should provide consistent accesses to those multiple databases with a transaction-like behavior, known as global serializability (Breitbart, Garcia-Molina, & Silberschatz, 1992). The distributed activity accessing the multiple databases is called a global transaction. Tesch and Wäsch (1997) presented an implementation of global transactions on the ODMG-compliant multidatabase systems. Creating such a multidatabase system could involve a lot of coordination efforts, both at the system level and the enterprise managerial level. The execution cost of the global transactions, in terms of concurrency control and atomic commitment, can be high. Thus, global transactions have not been widely adopted, despite their usefulness and convenience. Rather, multiple subtransactions were commonly executed on individual databases without enforcing global serializability.

In a mobile environment, it is appropriate to relax the overly restrictive serializability correctness criteria. It is often acceptable for a mobile client not to see the updates made by concurrently executing mobile clients, in exchange for a faster execution and a higher probability of committing its transaction. This is particularly true about the location of a moving object, which is inherently imprecise. Recency and freshness of location value is considered more important. Isolated-only transactions (Satyanarayanan, 2002) were proposed to reduce the impact of client disconnection. N-ignorance (Krishnakumar & Bernstein, 1994), bounded inconsistency (Wong, Agrawal, & Mak, 1997), and update consistency (Shanmugasundaram, Nithrakashyap, Sivasankaran, & Ramamritham, 1999) were some of the common weaker forms of correctness criteria. In these approaches, they try to ignore some of the operations in a transaction or allow them to be executed out-of-order in some controlled manner.

Transaction processing throughput in a mobile environment can be improved by utilizing the broadcast bandwidth effectively. The database can be broadcast and transactions can be processed against the database items broadcast. This is very useful for read-only transactions (Pitoura & Chrysanthis, 1999) simply by tuning for a consistent set of database items over the broadcast. To enable update transaction processing, the hybrid protocol by Mok, Leong, & Si (1999) ensures serializability by performing validation for update transactions, and utilizing the uplink channel to request for additional database items not available over the broadcast. Consistency across database items is ensured through the use of timestamps. In update consistency (Shanmugasundaram et al., 1999), a mobile client is only required to see updates made at server consistent with the values it reads, without having to follow the same serialization order as those observed by other mobile clients; it can be enforced by the cycle-based algorithm.

With the embracement of the Internet computing paradigm, more and more enterprises are willing to publicize their databases as part of their drive toward B2B or B2C E-commerce. Under most cases, these databases can be accessed from outside the enterprise via a Web interface. The ability to access consistent information using global transactions becomes more practical and manageable. Although updates to databases are normally restricted across departments or enterprises, more and more databases become enabled for the execution of read-only transactions by external parties, through the provision of Web-services (Hoang, Kawamura, & Hasegawa, 2004). Weakly consistent global transactions could be executed based on a sequence of Web-service requests. Furthermore, the presence of a high proportion of read-only transactions even renders the concurrency control for global transactions far more efficient.

Under certain B2B setting, it would be advantageous to automate the workflow process across enterprises (Vonk & Grefen, 2003). It is important to ensure the transactional execution of a low level workflow process, albeit the more relaxed consistency requirement on higher level process. Naturally, the overall process can be modeled as a nested transaction, which is destined to be long-lived that global serializability could be too restrictive in delivering good performance. Instead, the adoption of special transactions to roll back unintended effects is more appropriate (Tesch & Wäsch, 1997). With respect to the B2C setting, mobile clients would normally only initiate local transactions or simple global transactions spanning across a small number of databases, rather than a long-lived workflow process. The limitation of communication bandwidth and occasional network disconnection implies a longer transaction execution cycle.

FUTURE TRENDS

Owing to the complexity of global transaction processing and the resource limitations of mobile clients, it is sensible to migrate the coordination effort to the proxy server and the Web server, thereby relieving the mobile clients from the complex processing. This is especially true in the context of regions four and five. As a result, there would be the decoupling of the transaction processing mechanism from the application logic, with basic transactional support at the proxy. The application logic can further be delivered conveniently through the adoption of mobile agents (Yau, Leong, & Si, 2003), preferably intelligent agents that can make sensible decision on behalf of the client, only reporting back the outcome and obtaining confirmation from the client. For instance, mobile agents can act on behalf of the client for event-driven transactions like stock selling transactions or auctioning. Support for global transactions can be provided primarily on the wired network through the agent.

Web services allow the bundling of higher level or more complex operations on the databases to be invoked by clients (Hoang et al., 2004), simplifying the need to ensure local serializability of transactions on a particular database. For regions six and seven configurations, more complicated control can be established at the clients, whose higher computing power can also be dedicated to filter for information more effectively, with reference to its local cache, and to provide value-added operations on the data. One could leverage on the computing power of these mobile clients for m-commerce and l-commerce, by engaging in the popular paradigm of peer-to-peer computing (Avancha, D'Souza, Perich, Joshi, & Yesha, 2003) and pervasive cooperative computing. For instance, group-based location reporting exploiting peer-to-peer computing power is effective in reducing the update cost for changing client locations (Lam, Leong, & Chan, 2007).

The location of a client in l-commerce is a piece of sensitive information that deserves protection. There is an increasing concern on the privacy issue of mobile clients, in the context of location-dependent service (Mokbel et. al., 2006) and data mining (Verykios et al., 2004). A server is able to deduce the movement and even spending pattern of a client. Privacy should be protected and the quality of service should not be undermined excessively. This is a highly challenging task to strike a good balance. Finally, effective visualization of the result set and navigation through the result sequence to decide on the next step is important, since low-end mobile devices are normally equipped with a relatively small display, a constraint that is only mitigated at a much slower pace than advancement in network bandwidth and processing power.

CONCLUSION

Database support is an important and fundamental issue in m-commerce and l-commerce applications. With respect to the vast difference in the power and capacity of mobile devices, the varying wireless communication bandwidth, and the new dimension of client mobility and network disconnection, adjustments need to be made to render appropriate database support to these applications. In this article, we gave a generic classification along the three major characteristics for m-commerce and l-commerce environments. The different issues on database support were surveyed and discussed. In the future, there should be a division of research efforts, in providing effective transactional support at the server or proxy through agents or Web services, while leveraging on the capability of the peer clients in information organization and presentation. There is also a serious quest for better value-added computing support and solutions, including data-intensive stream processing and client privacy preservation.

REFERENCES

Acharya, S., Alonso, R., Franklin, M., & Zdonik, S. (1995). Broadcast disks: Data management for asymmetric communication environments. In *Proceedings of ACM SIGMOD International Conference on Management of Data* (pp. 199-210). ACM.

Alonso, R., & Korth, H. (1993). Database system issues in nomadic computing. In *Proceedings of ACM SIGMOD International Conference on Management of Data* (pp. 388-392). ACM.

Avancha, S., D'Souza, P., Perich, F., Joshi, A., & Yesha, Y. (2003). P2P m-commerce in pervasive environments. *ACM SIGecom Exchanges, 3*(4), 1-9.

Babu, S., & Widom, J. (2001). Continuous queries over data streams. *SIGMOD Record, 30*(3), 109-120.

Bean, J. (2003). *Engineering global e-commerce sites: A guide to data capture, content, and transactions*. Morgan Kaufmann Publishers.

Benetis, R., Jensen, S., Karĉiauskas, G., & Ŝaltenis S. (2006). Nearest and reverse nearest neighbor queries for moving objects. *VLDB Journal, 15*(3), 229-249.

Bernstein, P. A., Hadzilacos, V., & Goodman, N. (1987). *Concurrency control and recovery in database systems*. Reading, MA: Addison-Wesley.

Breitbart, Y., Garcia-Molina, H., & Silberschatz, A. (1992). Overview of multidatabase transaction management. *VLDB Journal, 1*(2), 181-239.

Chan, B. Y. L., Leong, H. V., Si, A., & Wong, K. F. (1999). MODEC: A multi-granularity mobile object-oriented database caching mechanism, prototype, and performance. *Journal of Distributed and Parallel Databases, 7*(3), 343-372.

Choy, M., Kwan, M., & Leong, H. V. (2000). Distributed database design for mobile geographical applications. *Journal of Database Management, 11*(1), 3-15.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80-91.

Franklin, M. J., Carey, M. J., & Livny, M. (1997). Transactional client-server cache consistency: Alternatives and performance. *ACM Transactions on Database Systems, 22*(3), 315-363.

Gedik, B., & Liu, L. (2005). A customizable k-anonymity model for protecting location privacy. In *Proceedings of International Conference on Distributed Computing Systems* (pp. 620-629). IEEE.

Güting, R. H. (1994). An introduction to spatial database systems. *VLDB Journal, 3*(4), 357-399.

Hoang, P. H., Kawamura, T., & Hasegawa, T. (2004). Web service gateway—A step forward to e-business. In *Proceedings of IEEE International Conference on Web Services* (pp. 648-655). IEEE.

Imielinski, T., & Badrinath, B. R. (1994). Mobile wireless computing: Challenges in data management. *Communications of the ACM, 37*(10), 18-28.

Jing, J., Helal, A. S., & Elmagarmid, A. (1999). Client-server computing in mobile environments. *ACM Computing Surveys, 31*(2), 117-157.

Krishnakumar, N., & Bernstein, A. J. (1994). Bounded ignorance: A technique for increasing concurrency in a replicated system. *ACM Transactions on Database Systems, 19*(4), 586-625.

Lam, G. H. K., Leong, H. V., & Chan, S. C. F. (2007). Group-based location reporting with peer-to-peer clients. In L. T. Yang, & M. K. Denko *Handbook on Mobile Ad Hoc and Pervasive Communications*. American Scientific Publishers.

Lee, K. C. K., Lee, W. C., & Leong, H. V. (2006). Nearest surrounder queries. In *Proceedings of International Conference on Data Engineering* (pp. 85-94). IEEE.

Lee, K. C. K., Leong, H. V., Zhou, J., & Si, A. (2005). An efficient algorithm for predictive continuous nearest neighbor query processing and result maintenance. In *Proceedings of International Conference on Mobile Data Management* (pp. 178-182). IEEE.

Leong, H. V., & Si, A. (2005). Multi-resolution information transmission in mobile environments. *Mobile Information Systems: An International Journal, 1*(1), 25-40.

Leong, H. V., & Si, A. (1995). Data broadcasting strategies over multiple unreliable wireless channels. In *Proceedings of International Conference on Information and Knowledge Management* (pp. 96-104). ACM.

Madria, S. K., Bhargava, B. K., Pitoura, E., & Kumar, V. (2000). Data organization issues for location-dependent queries in mobile computing. In *Proceedings of International Conference on Database Systems for Advanced Applications* (pp. 142-156). Springer-Verlag.

Mok, E., Leong, H. V., & Si, A. (1999). Transaction processing in an asymmetric mobile environment. In *Proceedings of International Conference on Mobile Data Access* (pp. 71-81). Springer-Verlag.

Mokbel, M. F., Chow, C., & Aref, W. G. (2006). The new casper: Query processing for location services without

compromising privacy. In *Proceedings of International Conference on Very Large Data Bases* (pp. 763-774).

Mokbel, M. F., Xiong, X., Hammad, M. A., & Aref, W. G. (2005). Continuous query processing of spatio-temporal data streams in PLACE. *GeoInformatica, 9*(4), 343-365.

Pitoura, E., & Chrysanthis, P. K. (1999). Scalable processing of read-only transactions in broadcast push. In *Proceedings of International Conference on Distributed Computing Systems* (pp. 432-439). IEEE.

Prabhakar, S., Xia, Y., Kalashnikov, D. V., Aref, W. G., & Hambrusch, S. E. (2002). Query indexing and velocity constrained indexing: Scalable techniques for continuous queries on moving objects. *IEEE Transactions on Computers, 51*(10), 1124 -1140.

Satyanarayanan, M. (2002). The evolution of coda. *ACM Transactions on Computer Systems, 20*(2), 85-124.

Shanmugasundaram, J., Nithrakashyap, A., Sivasankaran, R., & Ramamritham, K. (1999). Efficient concurrency control for broadcast environments. In *Proceedings of ACM SIGMOD International Conference on Management of Data* (pp. 85-96).

Shih, G., & Shim, S. S. Y. (2002). A service management framework for m-commerce applications. *Mobile Networks and Applications, 7*(3), 199-212.

Tan, K. L., & Ooi, B. C. (1998). On selective tuning in unreliable wireless channels. *Data and Knowledge Engineering, 28*(2), 209-231. Elsevier.

Tesch, T., & Wäsch, J. (1997). Global nested transaction management for ODMG-compliant multi-database systems. In *Proceedings of International Conference on Information and Knowledge Management* (pp. 67-74). ACM.

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM SIGMOD Record, 33*(1), 50-57.

Vonk, J., & Grefen, P. (2003). Cross-organizational transaction support for e-services in virtual enterprises. *Journal of Distributed and Parallel Databases, 14*(2), 137-172.

Wolfson, O., Sistla, A. P., Xu, B., Zhou, J., & Chamberlain, S. (1999). DOMINO: Databases for moving objects tracking. In *Proceedings of ACM SIGMOD International Conference on Management of Data* (pp. 547-549). ACM.

Wong, M. H., Agrawal, D., & Mak, H. K. (1997). Bounded inconsistency for type-specific concurrency control. *Journal of Distributed and Parallel Databases, 5*(1), 31-75.

Yau, S. M. T., Leong, H. V., & Si, A. (2003). Distributed agent environment: Application and performance. *Information Sciences Journal, 154*(1-2), 5-21.

KEY TERMS

Continuous Query: A continuous query is a query, which is re-evaluated continuously. For example, the query "give me the most updated temperature" will return different readings depending on the current moment. Some continuous queries are also location-dependent. For instance, the query "show me the nearest gas station" will continually execute a location-dependent query. Advanced query processing technique is needed, in conjunction with moving object databases.

Geographical Information System: A geographical information system is an information system that stores and manipulates data for geographical entities such as streets, road junctions, railway, land-use, or even terrain. The data is associated with the location of the entities to allow fast geo-referencing.

Location-Dependent Query: A location-dependent query is a query whose results depend on the current location of the query issuer. For example, the query "which is the nearest gas station?" will return different gas stations depending on the current location of a driver.

Mobile Database: A mobile database is a database accessible to mobile clients. There are appropriate mechanisms to take into account of the limitation of the wireless bandwidth, the use of downlink broadcast channel, and the effect of client mobility.

Moving Object Database: A moving object database is a database that maintains efficiently the location information about moving objects, with proper indexing on the object location.

Serializability/Global Serializability: Serializability is the generally accepted correctness criterion for concurrent execution of transactions. The concurrent execution should produce the same effect and lead to the same database state as one possible sequential execution of the same set of transactions. Global serializability is the correctness criterion for concurrent execution of global transactions over many database systems. It is a stronger correctness criterion than serializability.

Spatial Database: A spatial database is a database that maintains spatial data, including the topology and relationship between points, lines and shapes, and supports spatial operations and queries, such as the area of a shape, the distance between two entities, whether a shape is covered by or next to another.

Transaction/Global Transaction: A transaction is a sequence of operations on a database that should appear as if it were executed non-interfered, even in the presence of other concurrent transactions. A transaction should satisfy the ACID properties, namely, atomicity, consistency, isolation, and durability. A global transaction is a distributed transaction that is executed on two or more database systems.

Decision Support Systems in Small Businesses

Yanqing Duan
University of Luton, UK

Mark Xu
University of Portsmouth, UK

INTRODUCTION

Decision support systems (DSSs) are widely used in many organisations (Arslan et al., 2004; Belecheanu et al., 2003; Dey, 2001; Gopalakrishnan et al., 2004; Lau et al., 2001; Puente et al., 2002). However, there is a common tendency to apply experience and techniques gained from large organisations directly to small businesses, without recognising the different decision support needs of the small business. This article aims to address the issues related to the development and the implementation of DSSs in small business firms. Our arguments are based on evidence drawn from a large body of DSS literature and an empirical study conducted by the authors in the UK manufacturing sector.

BACKGROUND

Early DSS were developed in parallel with management information system (MIS) in the 1970s. MIS is developed to primarily generate management information from operational systems, whilst DSS as defined by Gorry and Scott Morton (1971) is information systems that focus on supporting people in the unstructured and semi-structured decision-making process. A typical DSS consists of four main components: the database, the model base, the user interface and the users. Central to the DSS are the models and analytical tools that assist managers in decision making and problem solving. Concomitant with advances in the technology of computing, most DSS provide easy access to data and flexible control models with a friendly user interface design; some DSS also incorporate a variety of analytical tools and report/graphic generators. The main purpose of DSS is not to replace managers' ability to make decisions, but to improve the effectiveness of managers' decision making.

DSS in practice can hardly be separated from other types of computer-based systems, as it is often integrated with those systems, for example operational databases, spreadsheets, report generators, and executive support systems. Thus the boundary of DSS has now been extended, and DSS broadly refers to any computer-based information system that affects or potentially affects how managers make decisions. This includes data and model oriented systems, reporting systems, executive support systems, expert systems and group decision support systems.

The success and continued growth of small and medium sized enterprises (SMEs) are critically important to local and national prosperity, but their problems are not always accorded the same importance as those of larger organisations. Compared to the research devoted to large organisations on the use of information systems, SMEs have attracted much less attention. It is also the case that the problems inherent in providing support for small business management are more commonly studied from a social or economic viewpoint. Very few studies indeed have addressed decision support needs in the context of the use of information technology.

Managers of small businesses have often been disappointed with software packages because of the inability of these to adapt well to their needs (Heikkila et al., 1991). There are dangers in seeing small businesses as miniature versions of large businesses; many problems differ, and even similar problems require different solutions. Small enterprises normally have limited resources and less skilled managerial staff. They have higher failure risks and commonly do not have suitable access to the information they need.

DSS IMPLEMENTATIONS

Small business may represent a productive domain for attempts to introduce greater levels of computer-based decision support. Ray (1994) suggests that small business managers and their staff have positive attitudes towards the use of computers in business. Cragg and King (1993) report that many companies have plans to increase their use of computer applications, and found that the wish for better information was the motivating force in all case studies conducted. In the majority of the firms studied by Khan and Khan (1992), managers believed that a computerised system improved their performance in selected areas, but that there is still room for significant further development.

Gordon and Key (1987) point out that if small business managers' problem-solving skills are deficient in any of the critical areas of management decision-making, then they must improve those skills through the use of appropriate educational programmes, consultants, decision support tools,

or some combination of these. Unfortunately, the owner-manager (because of involvement in the day-to-day operation of the firm) has not the time, resource or expertise needed to evolve an appropriately analytical approach (Raymond et al., 1989, cited in Naylor & Williams, 1994). There would seem to be as strong a case for the potential benefits of DSS to the smaller business as for its larger counterpart, provided suitable software is available, and it is effectively used by the managers concerned.

Limited research has investigated the success factors for the use of information technology (including DSS) in small businesses (Delone, 1988; Lai, 1994; Raymond & Bergeron, 1992) and the design and development of specific DSSs for SMEs (Chaudhry et al., 1996; Houben et al., 1999). Some work has been done specifically to identify those areas that have not been adapted to DSS, but show potential for its introduction for the small business (Duan et al., 2002). Most research (Levy, 1999) indicates that computer use is still confined to operational activities, although a few studies (Naylor & Williams, 1994) found that some SMEs have realised the value of their information systems as decision support tools and had begun to use them for more complex activities. Other researchers suggest that there are many areas in which DSS can be better developed and utilised to help managers in critical decision-making processes, such as marketing, sales promotion, cash-flow management and customer services. It has been argued that small businesses can improve their organisational performance and increase their competitiveness with appropriate information systems (Levy et al., 1999). The increasing emphasis on competitiveness in small business has led to a new focus on the competitive advantage promised by appropriate use of information technology (Levy et al., 1999; Lin et al., 1993).

A study conducted within the UK manufacturing SMEs by Duan et al. (2002) shows that the extent of DSS use is generally limited and the use of DSS varies considerably among the firms surveyed. However, even where there was a reported low level of DSS use, managers' satisfaction was relatively high. The applications with which managers were most satisfied were: cash management, budget preparation and materials requirements planning. Despite the relatively low usage of DSS generally, the majority of SME managers indicated that they use computers personally to aid business decisions; this suggests that there is, at least, widespread use of desktop computing in managers' offices.

Regarding the inhibitors to the greater use of DSS, lack of staff time to analyse needs and identify solutions is the most significant factor identified. Lack of finance for systems purchase or development, lack of experience of systems development, lack of information on available DSS packages, and unavailability of appropriate software were other factors commonly cited (Duan et al., 2002).

DSS DEVELOPMENT METHODS

DSS for small businesses can be developed and implemented in different ways. Four routes were identified, such as:

- Off-the-peg - purchase of a commercially developed package;
- Bespoke - designed by a software house for the specific application;
- In-house - developed by the firm's own specialist staff;
- User - developed by managers as users.

Research (Duan et al., 2002) shows that the majority of DSS were purchased as commercially developed packages; other systems were developed by managers as users, developed by in-house specialists or developed as bespoke systems by software houses. In view of the normally limited resource base for IT development (Heikkila et al., 1991), it is not surprising that most small firms choose to purchase commercially developed, ready-to-use DSS software. By breaking down the development methods into three decision-making levels, it shows that commercial packages are more commonly used at the operational level (60%) than at the strategic level. In contrast, user-developed DSS are more commonly used at the strategic level than at the operational level.

Research on *in-house* and *user* development methods in small firms is scarce. The evidence from the Duan et al. (2002) survey suggests that small business managers are capable of developing their own DSS, and that a certain proportion do so. Research in Canada by Raymond and Bergeron (1992) found that user-developed DSS in small businesses are more successful than any developed by other means. A study by Lai (1994) in the USA, however, revealed no link between the method of system development and DSS success.

By far the most commonly used DSS in small manufacturing firms are commercial packages purchased off the shelf for operational decision making. The readiness of small business managers to purchase commercial packages, coupled with their recognition that DSS vendors provide valuable support, suggest that small businesses represent a market for software developers that will continue to grow. That many managers are developing their own systems to support strategic decisions might also suggest there to be a market opportunity here.

THE FUTURE NEEDS FOR DSS IN SMES

The research by Duan et al. (2002) attempts to identify the gaps between the current provision of DSS and small business managers' desired levels of DSS support. The findings reveal that:

- the current level of DSS usage is low;
- although DSS usage is limited, managers are generally satisfied with DSS they are using;
- the desired level of support is much higher than the current provision;
- the high standard deviations for current DSS use and desired levels of support indicate high variations among responses. The standard deviation of levels of satisfaction is lower than the other two variables; this suggests that there is less disagreement on this issue.

The study supports the argument that current DSS in small businesses are geared to operational rather than strategic decision making. It is evident that the low-level use of DSS found by Raymond in 1982 has not changed significantly. The desired level of support at the operational level is also much higher than that at the strategic level. Users appear to expect that DSS will provide the most benefit for operational decisions. This is perhaps as well, given the nature of the decision-making tasks at strategic level, involving complex and changing environments, high level of uncertainty and the need to include decision makers' personal intuition and judgement. The lower-level use of DSS and desired support for strategic decision making does not mean that there is no space for further improvement, however. Indeed, the fact that many managers are "going it alone" could mean that professional support will enhance strategic planning. Levy et al. (1998) report that one of their case study firms had been successful in integrating information systems into its business strategy and gained competitive advantages. However, computer support for strategic decisions is still a challenging area for future research and much effort is being expended to overcome the difficulties (Duan & Burrell, 1997; Li et al., 2000), yet again, in the context of the larger business.

CONCLUSION

Decision support systems are designed to assist managers to make more effective decisions. Potentially, they could provide great benefits to SME managers and enhance managers' decision-making capability. However, the extent of DSS implementation in small business is still low in general, although there is significant variation between different firms. The literature review of previous studies indicates that the situation has not changed significantly since Raymond's investigation in 1982, and the present study confirms this. Lack of staff time to analyse needs and identify solutions is the most significant factor holding firms back from adopting, or making further use of DSS. Use of DSS at the operational decision-making level is higher than at the strategic level. Small business managers are generally satisfied with the DSS they are using and are hoping for much better DSS support in the future. DSS development, particularly DSS for strategic decisions in small business, still represents both a challenge and an opportunity for DSS professionals and researchers.

DSS in SMEs are most commonly implemented by purchase of a commercial package, and only rarely by bespoke development. Most DSS are used for operational rather than strategic decision making. Those firms that do use DSS to support strategic decisions rely upon user-developed models.

Although DSS applications in SMEs are still relatively few in number, most DSS users report satisfaction with their systems. To reduce the gaps between current DSS provision and the managers' indicated needs, a greater focus on small business by DSS researchers and practitioners is required. Systems most likely to appeal to small business managers will have to be appropriate to their sector's needs, and capable of implementation with minimal user training. In conclusion, it can be said that the current situation in relation to DSS in small business is full of potential but requiring further professional support.

REFERENCES

Arslan, M., Catay, B., & Budak, E. (2004). A decision support system for machine tool selection. *Journal of Manufacturing Technology Management, 15*(1), 101-109.

Belecheanu, R., Pawar, K.S., Barson, R.J., Bredehorst, B., & Weber, F. (2003). The application of case based reasoning to decision support in new product development. *Integrated Manufacturing Systems, 14*(1), 36-45.

Chaudhry, S.S., Salchenberger, L., & Beheshtian, M. (1996). A small business inventory DSS: Design, development, and implementation issue. *Computers & Operations Research, 23*(1), 63-72.

Cragg, P.B., & King, M. (1993). Small-firm computing: Motivators and inhibitors. *MIS Quarterly, 17*(2), 47-59.

Delone, W.H. (1988). Determinants of success for computer usage in small business. *MIS Quarterly, 12*(1), 51-61.

Dey, P.K. (2001). Decision support system for risk management: A case study. *Management Decision, 39*(8), 634-649.

Duan, Y., & Burrell, P. (1997). Some issues in developing expert marketing systems. *Journal of Business and Industrial Marketing, 12*(2), 149-162.

Duan, Y., Kinman, R., & Xu, M. (2002). The use of decision support systems in SMEs. In S.S. Burgess (Ed.), *Managing information technology in small businesses: Challenges and solutions* (pp. 140-155). Hershey, PA: Idea Group Publishing.

Gopalakrishnan, B., Yoshii, T., & Dappili, S.M. (2004). Decision support system for machining centre selection. *Journal of Manufacturing Technology Management, 15*(2), 144-154.

Gordon, W.L., & Key, J.R. (1987). Artificial intelligence in support of small business information needs. *Journal of Systems Management, 38*(1), 24-28.

Gorry, G., & Scott Morton, M. (1971). A framework for management information systems. *Sloan Management Review, 13*(1), 55-70.

Heikkila, J., Saarinen, T., & Saaksjarvi, M. (1991). Success of software packages in small business: An exploratory study. *European Journal of Information Systems, 1*(3), 159-169.

Houben, G., Lenie, K., & Vanhoof, K. (1999). A knowledge-based SWOT analysis system as an instrument for strategic planning in small and medium sized enterprises. *Decision Support Systems, 26*(2), 125-135.

Khan, E.H., & Khan, G.M. (1992). Microcomputers and small business in Bahrain. *Industrial Management & Data Systems, 92*(6), 24-28.

Lai, V.S. (1994). A survey of rural small business computer use: Success factors and decision support. *Information & Management, 26*(6), 297-304.

Lau, H.C.W., Lee, W.B., & Lau, P.K.H. (2001). Development of an intelligent decision support system for benchmarking assessment of business partners. *Benchmarking: An International Journal, 8*(5), 376-395.

Levy, M., Powell, P., & Galliers, R. (1999). Assessing information systems strategy development frameworks in SMEs. *Information & Management, 36*(5), 247-261.

Levy, M., Powell, P., & Yetton, P. (1998, December). SMEs and the gains from IS: From cost reduction to value added. *Proceedings of IFIP WG8.2 Working Conference, Information Systems: Current Issues and Future Changes,* Helsinki (pp. 377-392).

Li, S., Kinamn, R., Duan, Y., & Edwards, J. (2000). Computer-based support for marketing strategy development. *European Journal of Marketing, 34*(5/6), 551-575.

Lin, B., Vassar, J.A., & Clark, L.S. (1993). Information technology strategies for small businesses. *Journal of Applied Business Research, 9*(2), 25-29.

Naylor, J.B., & Williams, J. (1994). The successful use of IT in SMEs on Merseyside. *European Journal of Information Systems, 3*(1), 48-56.

Puente, J., Pino, R., Priore, P., & de la Fuente, D. (2002). A decision support system for applying failure mode and effects analysis. *International Journal of Quality & Reliability Management, 19*(2), 137-150.

Ray, C.M. (1994). Small business attitudes toward computers. *Journal of End User Computing, 6*(1), 16-25.

Raymond, L. (1982). Information systems in small business: Are they used in managerial decisions? *American Journal of Small Business, 5*(4), 20-26.

Raymond, L., & Bergeron, F. (1992). Personal DSS success in small enterprises. *Information & Management, 22*(5), 301-308.

KEY TERMS

Database: A collection of related information. The information held in the database is stored in an organised way so that specific items can be selected and retrieved quickly.

Decision Support System (DSS): An interactive computer-based system, which helps decision makers utilise data and models to solve semi-structured to unstructured problems.

Executive Information System (EIS): A computer-based information delivery and communication system designed to support the needs of senior managers and executives.

Expert Systems: A computer-based system that performs functions similar to those normally performed by a human expert. It has a knowledge base, an inference engine and a user interface.

Group Decision Support Systems (GDSS): Information systems that support the work of groups (communication, decision making) generally working on unstructured or semi-structured problems.

Management Information System (MIS): A business information system designed to provide past, present, and future information appropriate for planning, organising and controlling the operations of an organisation.

Small and Medium Sized Enterprises (SMEs): The definition of SMEs varies in different countries. It is normally defined as having between 10 and 249 employees in the UK and Europe.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 754-758, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Decision-Making Support Systems

Guisseppi Forgionne
University of Maryland, Baltimore County, USA

Manuel Mora
Autonomous University of Aguascalientes, Mexico

Jatinder N. D. Gupta
University of Alabama-Huntsville, USA

Ovsei Gelman
National Autonomous University of Mexico, Mexico

INTRODUCTION

Decision-making support systems (DMSS) are computer-based information systems designed to support some or all phases of the decision-making process (Forgionne, Mora, Cervantes, & Kohli, 2000). There are decision support systems (DSS), executive information systems (EIS), and expert systems/knowledge-based systems (ES/KBS). Individual EIS, DSS, and ES/KBS, or pair-integrated combinations of these systems, have yielded substantial benefits in practice.

DMSS evolution has presented unique challenges and opportunities for information system professionals. To gain further insights about the DMSS field, the original version of this article presented expert views regarding achievements, challenges, and opportunities, and examined the implications for research and practice (Forgionne, Mora, Gupta, & Gelman, 2005). This article updates the original version by offering recent research findings on the emerging area of intelligent decision-making support systems (IDMSS). The title has been changed to reflect the new content.

BACKGROUND

Decision-making support systems utilize creative, behavioral, and analytic foundations that draw on various disciplines (Sage, 1981). These foundations give rise to various architectures that deliver support to individual and group DMSS users. The architectures, which are summarized in Table 1, include (a) classic systems (Alter, 1996) such as decision support systems (DSS), expert and knowledge-based systems (ES/KBS), executive information systems (EIS), group support systems (GSS), and spatial decision support systems (SDSS) and (b) new systems (Forgionne, 1991; Forgionne, Mora, Cervantes, & Gelman, 2002a; Gray & Watson, 1996; Mora, Forgionne, Gupta, Cervantes, & Gelman, 2003; Power, 2002; Turban & Aronson, 1998) such as management support systems (MSS), decision technology systems (DTS), integrated DMSS, data warehouse (DW)-based and data mining (DM)-based DMSS (DW&DM-DMSS), intelligent DMSS (i-DMSS), and Web-based DMSS or knowledge-management DMSS.

The architectures have been applied to various public and private problems and opportunities, including the planning of large-scale housing demand (Forgionne, 1997), strategic planning (Savolainen & Shuhua, 1995), urban transportation policy formulation (Rinaldi & Bain, 2002), health care management (Friedman & Pliskin, 2002), pharmaceutical decision making (Gibson, 2002), banking management (Hope & Wild, 2002), entertainment industry management (Watson & Volovino, 2002), and military situations (Findler, 2002). Applications draw on advanced information technologies (IT), such as intelligent agents (Chi & Turban, 1995), knowledge-based (Grove, 2000) and knowledge-management procedures (Alavi, 1997), synthetic characters (Pistolesi, 2002), and spatial decision support systems (Silva, Eglese, & Pidd, 2002), among others.

DMSS ACHIEVEMENTS

Once created, DMSS must be evaluated and managed. Economic-theory-based methodologies, quantitative and qualitative process and outcome measures, and the dashboard approach have been used to measure DMSS effectiveness. These approaches suggest various organizational structures and practices for managing the design, development, and implementation effort. Most suggestions involve much more user involvement and a larger role for nontraditional specialists during the technical design, development, and implementation tasks.

To gain further insights about DMSS achievements, challenges, and opportunities posed by the development, the

Table 1. Decision-making support systems architectures

Classic DMSS Architectures	Description	Main Decision-Making Phase Supported					DMSS' SUPPORT CHARACTERISTICS
		INTELLIGENCE	DESIGN	CHOICE	IMPLEMENTATION	LEARNING	
DSS	A DSS is an interactive computer-based system composed of a user-dialog system, a model processor and a data management system, which helps decision makers utilize data and quantitative models to solve semi-structured problems.			**A**			(A) What-if, goal seeking, & sensitivity analysis.
ES & KBS	An ES/KBS is a computer-based system composed of a user-dialog system, an inference engine, one or several intelligent modules, a knowledge base, and a work memory, which emulates the problem-solving capabilities of a human expert in a specific domain of knowledge.	**A**		**B**			(A&B) Symbolic pattern-based recognition; fuzzy data; how and why explanation facilities.
EIS	An EIS is a computer based system composed of a user-dialog system, a graph system, a multidimensional database query system and an external communication system, which enables decision makers to access a common core of data covering key internal and external business variables by a variety of dimensions (such as time and business unit).	**A**			**B**		(A&B) Key performance indicators (KPI's) in graphs and text tables; data exploring and searching through drill-down, roll-up, slice and dice and pivoting operations; networking communications to internal and external bulletin boards.
GSS	A GSS an integrated computer based system composed of a communication sub-system and model-driven DMSS (DSS), to support problem formulation and potential solution of unstructured decision problems in a group meeting.		**A**	**B**			(A) Idea generation through brainstorming facilities; pooling and display of ideas; generation of alternatives and criteria. (B) Preference models; voting schemes; conflict negotiation support.
SDSS	A SDSS a computer based system composed of a user-dialog sub-system, a geographic/ spatial database sub-system, a decision model sub-systems and a set of analytical tools, which enables decision makers to treat with situations based strongly on spatial data.	**A**		**B**			(A) Spatial data searching support; visualization tools for maps, satellite images, and digital terrains. (B) What-if analysis of scenarios, goal-seeking analysis, sensitivity analysis of decision variables upon spatial data.

continued on following page

Table 1. continued

Modern DMSS Architectures	Description	Main Decision-Making Phase Supported					DMSS' SUPPORT CHARACTERISTICS	
		INTELLIGENCE	DESIGN	CHOICE	IMPLEMENTATION	LEARNING		
MSS, DTS or I-DMSS	These systems are the result of the triple-based integration (i.e., DSS, EIS, and ES/KBS) and have the aim to offer a full support to decision maker in all phases of the DMP.	**A**	**B**	**C**	**D**		(A&D)	Visual data exploring through graphs; color codes and tables; data exploration with drill-drown, roll-up, slice, and dice, pivoting operations.
							(B)	Intelligent advice through AI-based capabilities to support the models selection task.
							(C)	Numerical modeling through available numerical-based models; what-if, goal seeking and sensitivity analysis.
DW & DM DMSS	DW&DM-DMSS are computer-based system composed of a user-dialog sub-system, a multidimensional database subsystem, and an on-line analytical processing (OLAP) component enhanced with knowledge discovery algorithms to identify associations, clusters, and classifications rules intrinsic into the data warehouse.	**A**					(A)	OLAP capabilities of aggregation, slice and dice; drill-down; pivoting; trend analysis; multidimensional query; graphics and tabular data support. Knowledge discovery patterns using statistical based, tree-decision or neural networks.
Web-DMSS & KM-DMSS	Web-DMSS & KM-DMSS are computer-based system composed of an user-dialog sub-system, a text &multimedia document storage subsystem and publishing/retrieval subsystem to preserve and distribute knowledge in the organization using intranets.	**A**				**B**	(A&B) Document publishing and retrieval facilities	
i-DMSS	Are computer based system composed of an user-dialog sub-system, a multidimensional database and knowledge base subsystem and a quantitative & qualitative processing sub-system enhanced all of them with AI-based techniques, designed to support all phases of the DMP.	**A**	**B**	**C**	**D**	**E**	(A&D)	Visual data exploring through graphs; color codes and tables; data exploration with drill-drown, roll-up, slice, and dice, pivoting operations.
							(B)	Intelligent advice through AI-based capabilities to support the models selection task.
							(C)	Numerical and qualitative modeling through numerical-based or symbolic models; what-if, goal seeking, and sensitivity analysis.
							(E)	Symbolic reasoning through knowledge-based models for explanations about how and why the solution was reached.

Table 2. DMSS achievements, challenges, and opportunities

DMSS Issue	Expert Collective Opinion
Key Achievements	The evolution of DMSS software and hardware; the implementation of DMSS in a variety of organizations; the creation of DMSS tailored design and development strategies
Research Issues and Practical Problems	Providing quality data for decision support; managing and creating large decision support databases; model management and model reuse; building knowledge driven DMSS; improving communication technologies; developing a uniform and comprehensive DMSS scheme; developing an effective toolkit; developing and evaluating a synergistic integrated DMSS; collecting insights about the neurobiology of decision support for managers' less structured work; the application of agent and object-oriented methodologies; developing DMSS though well-established methodologies
Core DMSS Architectural Concepts and Opportunities	Web technology; accessibility; security; effective data, idea, and knowledge management, possibly through the use of smart agents; effective model management; effective dialog management; EIS-like features; incorporation of basic and common DMSS functionalities; mobile computing; user-centric design.

original study compiled opinions from recognized leaders in the field (Forgionne, Gupta, & Mora, 2002b). The expert verbatim views are summarized in Table 2.

Expert Opinions

The expert opinion indicates that DMSS have been recognized as unique information systems. Collectively, these experts focus on the deployment of new and advanced information technology (IT) to improve DMSS design, development, and implementation. In their collective opinion, the next generation of DMSS will involve: (a) the use of portals, (b) the incorporation of previously unused forms of artificial intelligence through agents, (c) better integration of data warehousing and data mining tools within DMSS architectures, (d) creation of knowledge and model warehouses, (e) the integration of creativity within DMSS architectures, (f) the use of integrated DMSS as a virtual team of experts, (g) exploitation of the World Wide Web, (h) the exploitation of mobile IT, and (i) the incorporation of advanced IT to improve the user interface through video, audio, complex graphics, and other approaches.

Future opportunities, trends and challenges discerned by the experts include: (a) availability of DMSS packages for specific organizational functions, such as customer relationship management, (b) system functional and technical integration, consolidation, and innovation, (c) software tool cost education, (d) the creation of a technology role for the decision maker through the DMSS, (e) the integration of the decision maker into the design and development process, (f) developing effective design and development tools for user-controlled development, (g) accommodating the structural changes in the organization and job duties created by DMSS use, (h) developing new and improved measures of DMSS effectiveness, (i) incorporating the cognitive and group dimensions of decision making, (j) utilization of smart agents, (k) distribution of DMSS expertise through collaborative technologies, (l) incorporating rich data, information and knowledge representation modes into DMSS, and (m) focusing user attention on decisions rather than technical issues. Common themes suggested by this disparate expert opinion are (a) the DMSS should focus decision makers on the decision process rather than technical issues, and (b) DMSS development may require specialized and new IT professionals, and (c) there is need for a systematic and well-managed implementation approach.

Intelligent DMSS

Since most experts value artificial intelligence in decision making support, a historical review of the literature, covering the period 1980-2004, was conducted to examine the state of the intelligent DMSS (I-DMSS) concept (Mora et al., 2006). This history indicated that neural networks and fuzzy logic have become more popular than Bayesian/belief nets, and intelligent agents, genetic algorithms, and data mining have emerged as tools of interest.

In terms of the decision making process, the intelligence and choice phases have been the most supported phases. Over time, intelligence support has increased, while choice support has decreased. Within the intelligence phase, the problem recognition step has grown in popularity.

Among dialog user interface capabilities, text/passive graphics has remained the most used tool. Model management has been most often supported by knowledge-based methodologies and quantitative models. Knowledge-based models have been declining in importance, while quantitative models have been gaining popularity. Symbolic structured mechanisms, based on rule-based systems and fuzzy logic, and quantitative structured approaches, based on neural networks and data mining, have become the most popular data management tools.

FUTURE TRENDS

The historical analysis supports some of the expert opinion. Specifically, the reported record indicates that effort is underway to (a) increase DMSS processing capabilities through intelligent agents, fuzzy systems, and neural networks and (b) improve user-interface capabilities through multimedia and virtual environments. In short, the experts and literature on AI and DMSS implicitly recognize the relevance of improving the DMSS user interface, information and knowledge representations schemes and intelligent processing capabilities through the deployment of advanced IT.

CONCLUSION

In some ways, the DMSS field has not progressed very much from its early days. There is still significant disagreement about definitions, methodologies, and focus, with expert opinion varying on the breadth and depth of the definitions. Some favor analytical methodologies, while others promote qualitative approaches. Some experts focus on the technology, while others concentrate on managerial and organizational issues. There does not seem to be a unified theory of decision-making, decision support for the process, or DMSS evaluation. Moreover, achieving successful implementation of large-scale DMSS is still a complex and open research problem (Mora et al., 2002).

In spite of the diversity, opinions are consistent regarding some key DMSS elements. Most experts recognize the need for problem pertinent data, the role of the Internet in providing some of the necessary data, the need for system integration within DMSS architectures and between DMSS and other information systems, and the importance of artificial intelligence within DMSS processing. The historical record also supports the emerging importance of intelligent decision-making support and identifies quantitative-based methodologies as the growing form of intelligence. The DMSS concept also continues to be successfully applied across a variety of public and private organizations and entities. These applications continue to involve the user more directly in the design, development, and implementation process.

The trends will create DMSS that are technologically more integrated, offer broader and deeper support for decision-making, and provide a much wider array of applications. In the process, new roles for artificial intelligence will emerge within DMSS architectures, new forms of decision technology and methodology will emerge, and new roles will be found for existing technologies and methodologies.

As the evolution continues, many tasks that had been assigned to human experts can be delegated to virtual expertise within the DMSS. With such consultation readily available through the system, the decision maker can devote more effort to the creative aspects of management. Support for these tasks can also be found within DMSS. In the process, the decision maker can become an artist, scientist, and technologist of decision-making. The DMSS-delivered virtual expertise can reduce the need for large support staffs and corresponding organizational structures. The organization can become flatter and more project-oriented. In this setting, the decision maker can participate more directly in DMSS design, development, implementation, and management. Such changes will not occur without displacements of old technologies and job activities, radical changes in physical organizations, and considerable costs. As the reported applications indicate, however, the resulting benefits are likely to far outweigh the costs.

REFERENCES

Alavi, M. (1997). KPMG Peat Marwick U.S.: One giant brain. In *Creating a system to manage knowledge* (pp. 75-95). Harvard Business School Publishing (Case 9-397-108).

Alter, S. (1996). *Information systems: A management perspective*. Menlo Park, CA: Benjamin/Cummings.

Chi, R., & Turban, E. (1995). Distributed intelligent executive information systems. *Decision Support Systems, 14*(2), 117-130.

Findler, N. (2002). Innovative features in a distributed decision support system based on intelligent agent technology. In M. Mora, G. Forgionne, & J. Gupta (Eds), *Decision-making support systems: Achievements, challenges, and trends* (pp. 174-192). Hershey, PA: Idea Group Publishing.

Forgionne, G. (1997). HADTS: A decision technology system to support army housing management. *European Journal of Operational Research, 97*(2), 363-379.

Forgionne, G. (1991). Decision technology systems: A vehicle to consolidate decision-making support. *Information Processing and Management, 27*(6), 679-797.

Forgionne, G., Mora, M., Cervantes, F., & Gelman, O. (2002a, July 3-8). I-DMSS: A conceptual architecture for next generation of DMSS in the Internet age. In F. Adam, P, Brezillon, P. Humpreys, & J. Pomerol (Eds.), *Proceedings of the International Conference on Decision Making and Decision Support in the Internet Age (DSIAge02)* (pp. 154-165). Cork, Ireland.

Forgionne, G., Mora, M., Cervantes, F., & Kohli, R. (2000, August 10-13). Development of integrated decision-making support systems: A practical approach. In M. Chung (Ed.), *Proceedings of the AMCIS 2000 Conference* (pp. 2132-2134). Long Beach, CA, USA.

Forgionne, G., Gupta, J., & Mora, M. (2002b). Decision making support systems: Achievements, challenges, and opportunities. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 392-402). Hershey, PA: Idea Group Publishing.

Forgionne, G., Mora, M., Gupta, J. N. D., & Gelman, O. (2005). Decision-making support systems. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 759-765). Hershey, PA: Idea Group Publishing.

Friedman, N., & Pliskin, N. (2002). Demonstrating value-added utilization of existing databases for organizational decision-support. *Information Resources Management Journal, 15*(4), 1-15.

Gibson, R. (2002). Knowledge management support for decision making in the pharmaceutical industry. In M. Mora, G. Forgionne, & J. Gupta (Eds), *Decision-making support systems: Achievements, challenges, and trends* (pp. 143-156). Hershey, PA: Idea Group Publishing.

Glass, R., Ramesh, V., & Vessey, I. (2004). An analysis of research in computing discipline. *Communications of the ACM, 47*(6), 89-94.

Gray, P., & Watson, H. (1996, August 16-18). The new DSS: data warehouses, OLAP, MDD, and KDD. *Proceedings of the AMCIS Conference 1996*. Phoenix, AZ, USA.

Grove, R. (2000). Internet-based expert systems. *Expert Systems, 17*(3), 129-135.

Hope, B., & Wild, R. (2002). Procedural cuing using expert support system. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 101-119). Hershey, PA: Idea Group Publishing.

Mora, M., Cervantes, F., Gelman, O., Forgionne, G., Mejia, M., & Weitzenfeld, A. (2002). DMSS implementation research: A conceptual analysis of the contributions and limitations of the factor-based and stage-based streams. In M. Mora, G. Forgionne, & J. Gupta, (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 331-356). Hershey, PA: Idea Group Publishing.

Mora, M., Forgionne, G., Gupta, J., Cervantes, F., & Gelman, O. (2003, Sep. 4-7). A framework to assess intelligent decision-making support systems. In V. Palade, R. Howlett, & L. Jain (Eds.), *Proceedings of the 7th KES2003 Conference,* Oxford, UK, LNAI 2774 (pp. 59-65). Heiderberg, FRG: Springer-Verlag.

Mora, M., Forgionne, G., Gupta, J. N. D., Garrido, L., Cervantes, F., & Gelman, O. (2006). A strategic descriptive review of intelligent decision-making support systems research: The 1980-2004 Period. In J. N. D. Gupta, G. A. Forgionne, & M. Mora (Eds.), *Intelligent decision-making support systems: Foundations, applications, and challenges, series: Decision engineering* (pp. 441-462). Springer.

Pistolesi, G. (2002). How synthetic characters can help decision-making. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 239-256). Hershey, PA: Idea Group Publishing.

Power, D. (2002). Categorizing decision support systems: A multidimensional approach. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 20-27). Hershey, PA: Idea Group Publishing.

Rinaldi, F., & Bain, D. (2002). Using decision support systems to help policy makers cope with urban transport problems. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 86-100). Hershey, PA: Idea Group Publishing.

Sage, A. (1981). Behavioral and organizational considerations in the design of information systems and process for planning and decision support. *IEEE Transactions on Systems, Man, and Cybernetics, 11*(9), 640-678.

Savolainen, V., & Shuhua, L. (1995). Strategic decision-making and intelligent executive support system. In *Proceedings of the 12th International Conference on Systems Science* (pp. 285-295), Wroclaw, Poland.

Silva, F., Eglese, R., & Pidd, M. (2002). Evacuation planning and spatial decision making: Designing effective spatial decision support systems through integration of technologies. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 358-373). Hershey, PA: Idea Group Publishing.

Turban, E., & Aronson, J. (1998). *Decision support systems and intelligent systems* (pp. 20-23). Upper Saddle River, NJ: Prentice-Hall.

Watson, H., & Volonino, L. (2002). Customer relationship management at Harrah's Entertainment. In M. Mora, G. Forgionne, & J. Gupta (Eds.), *Decision-making support systems: Achievements, challenges, and trends* (pp. 157-172). Hershey, PA: Idea Group Publishing.

KEY TERMS

Data Warehousing-Data Mining (DW-DM) DMSS: Computer-based system composed of an user-dialog subsystem, a multidimensional database subsystem, and an

online analytical processing (OLAP) component enhanced with knowledge discovery algorithms to identify associations, clusters, and classifications rules intrinsic in a data warehouse.

Decision Making Support System (DMSS): An information system designed to support some, several or all, phases of the decision making process.

Decision Support System (DSS): An interactive computer-based system composed of a user-dialog system, a model processor and a data management system, which helps decision makers utilize data and quantitative models to solve semi-structured problems.

Executive Information System (EIS): A computer based system composed of a user-dialog system, a graph system, a multidimensional database query system and an external communication system, which enables decision makers to access a common core of data covering key internal and external business variables by a variety of dimensions (such as time and business unit).

Expert System/Knowledge Based System (ES/KBS): A computer-based system composed of a user-dialog system, an inference engine, one or several intelligent modules, a knowledge base and a work memory, which emulates the problem-solving capabilities of a human expert in a specific domain of knowledge.

Group Support System (GSS): An integrated computer based system composed of a communication sub-system and model-driven DMSS (DSS), to support problem formulation and potential solution of unstructured decision problems in a group meeting.

Intelligent Decision Making Support Systems (i-DMSS): Computer based system composed of an user-dialog sub-system, a multidimensional database and knowledge base subsystem, and a quantitative and qualitative processing sub-system enhanced with AI-based techniques, designed to support all phases of the decision making process.

Management Support Systems (MSS), Decision Technology Systems (DTS), or Integrated Decision Making Support Systems (I-DMSS): Systems that integrate DSS, EIS and ES/KBS to offer full support to the decision maker in all phases of the decision making process.

Spatial Decision Support System (SDSS): A computer-based system composed of a user-dialog sub-system, a geographic/spatial database sub-system, a decision model sub-systems and a set of analytical tools, which enables decision makers to analyze situations involving spatial (geographic) data.

Web-DMSS & Knowledge Management (KM)-DMSS: Computer-based system composed of an user-dialog subsystem, a text and multimedia document storage subsystem, and publishing/retrieval subsystem to preserve and distribute knowledge in intranet-supported organizations.

Delivering Web-Based Education

Kathryn A. Marold
Metropolitan State College of Denver, USA

INTRODUCTION

A decade of hindsight allows us to examine the phenomenon of Web-based course delivery and evaluate its successes and failures. When Web-delivered courses mushroomed from campuses in the 1990s, they were embraced by students, faculty, and administrators alike. The prospect of "electronic tutelage" (Marold, 2002), which allowed students through Web interface to take college courses for credit any time, any place (ATAP), was immediately popular with students. The interruptions of job and schedule changes, relocation, childbirth, failed transportation to campus, and so forth no longer necessitated an interruption in progress toward a degree. Likewise, faculty saw online teaching as an opportunity to disseminate knowledge and assess student progress according to their personal preferences, and to communicate personally with their students, albeit virtually. Administrators saw the revenue without physical classroom allocations as an immediate cash cow. In the beginning, there was satisfaction all around. Although this state of affairs was not necessarily universal, generally it could be concluded that Web-based education was a very good thing.

The Evolution Of Web-Based Course Delivery

Web-based education is a variation of distance learning: the content (college courses from an accredited North American institution, for purposes of this chapter) is delivered via the World Wide Web. The Web course content covers a quarter or semester of curriculum that the student must complete and prove a level of mastery within a given timeline. For the most part, Web-based courses use existing college curriculum and timelines. Web-based education is currently the most popular form of distance education. As educators are inclined to do, it was not long before they wanted to stand back and evaluate what they had created and determine the success of Web-delivered courses as a form of distance education. With McLuhanesque procedures, a glance in the "rear view mirror" was in order (McLuhan, 1964.) The results of many measures of success show that for *some* of the students, *some* of the time, in *some* situations, Web-based education is quite successful. Likewise, for many persons in many situations and in many phases of their formal education, Web-delivered education is *not* the answer.

BACKGROUND

The advent of the World Wide Web in the early 1990s promised a more effective, user-friendly form of Internet distance education. The graphical hypertext and, indeed, the hypermedia nature of the Web could enhance course delivery. Almost immediately, Web courses began to flourish. A new mode of delivery was firmly established.

Web-Based Education's Successes and Failures

Numerous publications have exposed problems associated with the Web-based form of distance education. The population taking the courses was sometimes the problem (Haga, 2001). The attrition and failure rate of Web-delivered courses was higher than the classroom arena (Terry, 2001). The content of the course could be problematic (Haga, 2002). The credibility of course credit achieved online was sometimes suspect (Moreno, 2000). The level of courses offered online was sometimes suspect (Marold, 2003). Research findings suggest the following conclusions concerning Web-based education (see Table 1).

There are almost as many reports of success with Web-based education as there are reports of failures. Students who are successful with Web courses tend to take more of them, sometimes as many as 90 hours of the 120 hours required for a bachelor's degree. There are now entire degrees offered online. The earliest research on Web-based education reported no statistical difference in final grades between Web-based groups and classroom groups (Mawhinney, 1998; Schulman, 1999).

The conclusion that Web-delivered education, like all distance education, is only appropriate for some students cannot be denied. It is obvious that Web courses are not going to go away. It is also undeniable that regardless of how enrollments in Web-based courses are screened, there will be students who enroll in Web courses that should not be in them. It has been shown time and again that some Web students enroll for all of the wrong reasons (Haga, Marold, & Helms, 2001). It is equally obvious that Web courses fill an enormous need for many students and are, therefore, very successful in many instances.

Table 1. Successes and failures of Web-based education

Positive	Negative
Survey level courses are the most successful.	The attrition and failure rates for upper level, analytical Web-based courses often reach 50%.
Courses at the 1-2 level of Bloom's taxonomy (Bloom, 1956) of learning immersion are more successful than those at the 3-5 level.	Students at the B and C level (the vast majority of students in any institution) are the most at risk for not completing and not passing Web-delivered courses.
Students with GPA of 3.5 or better are the most successful at completing and excelling in Web-delivered courses.	
Graduate level Web-delivered courses are more successful.	Web-delivered courses are a disaster for the passive learner without time management and independent study skills.
Courses delivered via a 3rd party distributor or a portal (such as *WebCt* or *Blackboard*) are more successful than self-hosted Web courses	Both students and faculty alike indicate time spent on an Internet delivered course is more than it would be on its classroom equivalent.
Internet students generally do better than their classroom counterparts on exams. Internet students generally do worse on projects than their counterparts on assigned projects.	
Analytical and problem solving courses are least successful.	
Web-delivered courses are a godsend for the highly motivated, independent learner.	
Final grades on Web-based education courses generally do not differ significantly from those earned in the classroom. Web-based courses are here to stay. They are an accepted, credible method of course delivery	

While the students who are at risk for failure in Web-based courses that are analytical and require problem solving are those students who are generally classified as mid-level achievers, taking prerequisite courses online seems to alleviate the risk slightly (Pence, 2003). The student group at the greatest risk is the mid-level achieving group, which in a normal distribution is the largest number of students in the class (Marold & Haga, 2004). Pence suggests some alleviating factors, such as taking the prerequisite course online from the same institution. This suggests that as students become more accustomed to the requirements and idiosyncrasies of online learning, the risk decreases. Experience makes a difference. In addition, the majority of students taking online courses indicate that they would take another online course, even though they perceive them to be more work than an equivalent classroom course. Despite attrition and failure rates that sometimes reach 50%, Web-based education is clearly a student favorite.

Tables 2 and 3 show some of the research results of a decade of Web-based education.

In the above research of two separate Web-based required computer information systems junior level courses in the same department of a large urban state school, student tests were higher in the Internet version, but their project scores (application of learning) were lower.

In Table 2, there were three different courses at freshman, sophomore, and junior levels, offered online as well as in the classroom, from the same department of a large urban institution. All three courses were survey-type courses with a heavy skills component. This research was done earlier than the study shown in Table 2. The test scores were also higher for the Internet sections, and the projects lower.

Table 2. Comparing Internet scores with classroom scores (Haga, 2001)

Telecommunications	Internet	Classroom
Projects average	60.8	83.5
Exams average	64.2	62.7
Visual Basic Programming		
Projects average	69	73.3
Exams average	72	71.3

Table 3. Comparing Internet scores with classroom scores (Marold, Larsen, & Moreno, 2002)

Introduction to Computers 1010	Internet	Classroom
Projects average	92.4	92.4
Exams average	78.2	73.5
Computer Applications for Business 2010	Internet	Classroom
Projects average	90.3	91.8
Exams average	77.5	70.01
Micro-based Software 3270	Internet	Classroom
Projects average	87.5	93.8
Exams average	77.7	67.8

** Revised classroom exams average for 2010*

Table 4. Successful Web-course characteristics

• More is not better. Simplicity, organization, and clarity are paramount.
• Frequent communication is important, both synchronous (chat, messaging) and asynchronous (e-mail, forum, discussion boards).
• Smaller modules and units work better.
• More frequent, less intricate assignments are better than complex comprehensive assignments due at the end of the course.
• Student uploaded projects and assignments work better than hard copy submissions.
• Individualized assignments are more likely to be submitted.
• The more choices in assignments and projects, the better.
• Proctored exams and secure quizzes are essential for course acceptance in the community and the workplace.
• Courses shorter in duration are more successful than ones that span semesters or quarters.
• Electronic grade books and regular feedback to student assignments submitted are essential.
• Samples of student work communicate expectations for assignments.
• Profiles or student Web pages for class members build community.
• Web links to helpful course content and enrichment material are appreciated, although not used extensively.
• More than one route to the same location in a course individualizes the course.
• Calendars of due dates and course announcements provide quick reference.

Although Web-based grades differ from classroom grades on individual projects and exams, generally over a decade of looking back, we see that final grades on courses completed on the Internet and in the traditional classroom do not differ significantly (Haga, 2002; Marold, 2002; Moreno, 2000; Presby, 2001).

DESIGNING AND DELIVERING WEB-BASED EDUCATION

It is heartening to know that the success of a Web-delivered course is not directly related to the faculty's skill in instructional design and Web publication (Marold, 2002). Templates and Web design software suffice for course design. Simple courses are just as successful as complex ones. Studies that count visits to various pages in Web courses show that many of the available pages are infrequently (and sometimes never) used by a majority of online students (Haga, 2001). The same is true for extensive instructions on how to take an online course, or lists of Web links available for extra information supplementing the course material. Factors that contribute to successful Web-delivered course are shown in Table 4. These are general conclusions based upon a decade of studies of Web-based education.

For every individual characteristic listed here, one undoubtedly can find research contradicting it. From their inception, Web delivered courses have been examined and reexamined. However, in general, these are factors that have been found to make a positive difference for faculty and designers of Web-based courses.

FUTURE TRENDS

As Web courses become more common and educators become more proficient at designing and deploying them, undoubtedly their success rate will improve. As students become more accustomed to taking Web courses, they will become more proficient at completing them. As McLuhan (1964) noted, it takes a period of adjusting to a new medium before we are entirely comfortable with it. Once Web courses become part of the "every day" of our existence, Web-based education may take its legitimate place alongside other modes of delivery.

CONCLUSION

In conclusion, Web-based education is working—for some. Entire degrees are obtainable online; students, faculty, administrators, and organizations accept credits earned online. For better or for worse, Web-based education is firmly entrenched in all areas of higher education. While Web-based education has not yet achieved the same level of success as classroom delivered instruction, it is part of most institutions' programs. The fact that Web-based education is truly working in a cost-effective manner for some of the student population some of the time, assures its continuance. Like the printing press of the late 1400s for mass distribution of knowledge, Web-based education provides mass distribution of knowledge in a new, effective way.

REFERENCES

Bloom, B.S. et al. (1956). *Taxonomy of educational objectives. Handbook I.: The cognitive domain*. New York: David McKay.

Haga, W., & Marold, K. (2002). Is the computer the medium *and* the message? A comparison of VB programming performance in three delivery modes. *International Business and Economic Research Journal, 1*(7), 97-104.

Haga, W., Marold, K., & Helms, S. (2001). Round 2 of online learning: Are Web courses *really* working? *Proceedings of the Twenty-Ninth Annual Conference of IBSCA*. Providence, RI.

Marold, K. (2002). The twenty-first century learning model: Electronic tutelage realized. *Journal of Information Technology Education, (1)*2, 113-123.

Marold, K., & Haga, W. (2003). The emerging profile of the on-line learner: Relating course performance with pretests, GPA, and other measures of achievement. *Proceedings of the Information Resources Management Association*. Philadelphia, PA.

Marold, K., & Haga, W. (2004). E-learners at risk: Midlevel achievers and online courses. In A. Appicello (Ed.), *Instructional Technologies: Cognitive Aspects of Online Programs.* Hershey, PA: Idea Group Publishing.

Marold, K., Larsen, G., & Moreno, A. (2002). Web-based learning: Is it working? In M. Khosrowpour (Ed.), *Web-based instructional technologies.* Hershey, PA: Idea Group Publishing.

Mawhinney, C. et al. (1998, October). Issues in putting the business curriculum online. *Proceedings of the Western Decision Sciences Institute*. Puerto Vallarta, MX.

McLuhan, M. (1964). *Understanding media*. New York: McGraw Hill.

Moreno, A., Larsen, G., & Marold, K. (2000). The credibility of online learning: A statistical analysis of IS course delivery at three levels. *Proceedings of the Western Decision Sciences Institute*. Maui, HI.

Pence, N.K. et al. (2003). An exploration of the impact of online delivery in prerequisite courses on CIS majors' course sequence. *Proceedings of the International Business and Economic Research Conference*. Las Vegas, NV.

Presby, L. (2001). Increasing productivity in course delivery. *T.H.E. Journal, 28*(7), 52-58.

Schulman, A., & Sims, R.L. (1999). Learning in an online format versus an in-class format: An experimental study. *T.H.E. Journal, 26*(11), 54-56.

Terry, N. (2001). Assessing enrollment and attrition rates for the online MBA. *T.H.E. Journal, 28*(7), 64-68.

KEY TERMS

ATAP: Any time, any place learning. A basic characteristic of Web-based education courses in that they are available to the student on a 24-7 basis.

Bloom's Taxonomy of Learning: A scale that represents an organization of learning levels (five levels) that are characterized by the student's immersion into the theory and application of principles of a course content.

Chat Sessions: Live discussions online with a variable number of participants in a Web-based class. They can be formal and led by the instructor, or they can be leaderless informal conversations. Chat sessions are synchronous.

Electronic Gradebooks: Maintaining a record of a student's progress in a Web-based education class by posting grades on the course Web pages. General gradebooks show all enrollees; personalized gradebooks can only be viewed by the individual student.

Electronic Tutelage: Learning of new complex concepts (sometimes called *scientific* concepts), not with the intervention of a physical tutor, but via electronically delivered materials; the basic theory behind Web-based education.

Synchronous and Asynchronous Communication: Synchronous communication via the Web is immediate communication, such as in chat or instant messaging. Asynchronous communication is delayed communication via the Web, such as threaded discussions, forums, or e-mail messages, where each participant does not have to be online at the same time.

Web-Based Education: A variation of distance learning; the content (college courses from an accredited North American institution, for purposes of this chapter) is delivered via the World Wide Web. The Web course content covers a quarter or semester of curriculum that the student must complete within a given timeline in return for course credit.

Web Links: Hyperlinks (hot links) to other Web sites that are embedded in the active pages of a Web-based education course.

Web Profiles: Short biographies of students enrolled in a Web-based education course. The profiles may contain Web links to students' own home pages and digitized photos of the students.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 786-790, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Democratic E-Governance

Ari-Veikko Anttiroiko
University of Tampere, Finland

INTRODUCTION

The changing role of the state and a managerialist view of the operations of public sector organizations gave rise to the idea of new public governance. Gradually more citizen-centered views of governance also emerged, reflecting a need to strengthen the role of citizens and communities in governance processes at different institutional levels. This development, especially since the mid-1990's, has been affected by new technologies, leading to a kind of coevolution of institutional arrangements and technological solutions that have paved the way for a better understanding of the potentials of democratic e-governance.

BACKGROUND

Discussion about governance has acquired new dimensions since the early 1990's due to the gradual erosion of the hierarchical, mainly state-centric bases of political power. The decline of the nation state and the rise of the regions and local governments as the new key players in coping with external challenges and imposing a political will within territorial communities is among the core topics. Also, after the second World War and the 1980's in particular, international organizations and regional institutions started to gain more power in international arena (Pierre, 2000).

Another widely discussed aspect of public governance relates to the functioning and ways of working of public sector organizations. The entire institutional landscape and the overall understanding of the role of public sector organizations has gradually changed practically everywhere in the world, thus fueling the discussion about public governance. One important governance agenda-setter was the OECD Public Management Committee (PUMA), which carried out work on this topic during the first half of the 1990's and as a synthesis published a policy paper entitled *Governance in Transition* in 1995 (OECD, 1995). OECD's policy lines have been more or less neoliberal, which means that governance issues were discussed and still are to a large extent within the framework of New Public Management (NPM). In essence, its message is that the approach to the management of public organizations and services needs to be based on managerialism and market-based coordination.

Contemporary understanding and use of the concept of governance has its roots in the changing role of the state and in a managerialist view of the operations of public organizations. These two discourses have been challenged by another approach, which could be called democratic governance. It emphasizes the interactions between citizens, political representatives and administrative machinery providing a special view of citizens' opportunities to influence and participate in governance processes.

DEFINITION OF GOVERNANCE

One of the reasons behind the revival of the concept of governance was the need to distinguish between the traditional, institutionally oriented conception of "government" and more dynamic and network-based ways of thinking and working in policy processes. Thus, *government* refers to the institutions and agents that perform the governmental functions, that is, to formal institutions of the state or those of decentralized territorial governments and their ability to make decisions and take care of their implementation, whereas *governance* is about the new modes and manner of governing within policy networks and partnership-based relations (Jessop, 1998; Kooiman, 1993; Pierre & Peters, 2000; Stoker, 1998).

The way the concept of governance is used here can be specified as "public governance", which aims to pursue collective interest in the context of intersectoral stakeholder relations. In this sense, governance refers to the coordination and the use of various forms of formal or informal types of nonhierarchically organized interaction and institutional arrangements in the policy-making, development and service processes to pursue collective interest (Anttiroiko, 2004).

E-TRANSFORMATION IN DEMOCRATIC GOVERNANCE

Informatization as an important side of the transformational aspect of governance profoundly affects the relationships of different actors, forms and channels of communication and interaction, and the entire fabric of network and partnership relations. The introduction of ICTs in the public sector in the 1960's in most of the advanced countries started to reshape their data processing activities, such as record keeping and financial administration. Electronic systems started to replace old manual systems. This picture changed dramatically in

the 1990's. At the core of this revolution was the Internet (Seneviratne, 1999).

Since the 1990's a need for reconstruction of technology along more democratic lines became apparent. New ICTs have a potential to restructure government and to strengthen democracy, and to create a closer relationship between public administration and citizens in particular. It has even been said that new ICTs applied by government contribute to the emergence of a different type of governance, that is, more "direct" government, as concluded by Pardo (2002).

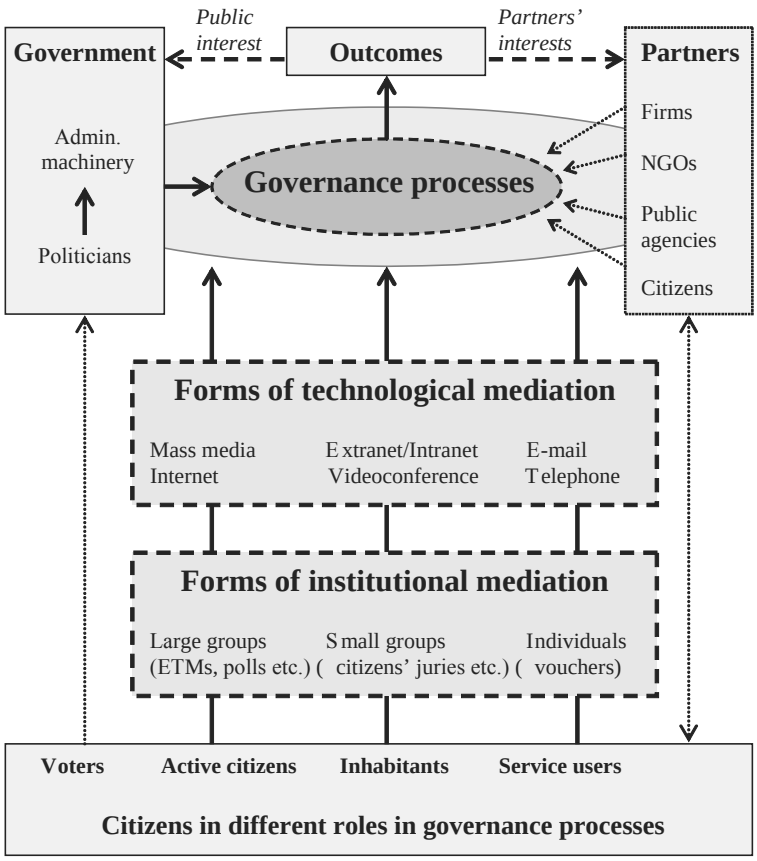
This development boils down to the idea of democratic e-governance, which combines three conceptual elements: *governance* as a process and activity area, *democracy* as an applied principle, and *information and communication technologies* as a tool. Democratic e-governance is a technologically mediated interaction in transparent policy-making, development and service processes in which political institutions can exercise effective democratic control and, more importantly, in which citizens have a chance to participate and effectively influence relevant issues through various institutionally organized and legitimate modes of participation (Anttiroiko, 2004). At a practical level democratic e-governance requires both institutional and technological mediation of civic and community interests in formal governance processes, as illustrated in Figure 1.

One of the expected strengths of citizen-centered democratic e-governance is its ability to combine a discursive public sphere with the decision-making sphere, and thus to eliminate hierarchical relations which characterize the contemporary representative systems of government.

METHODS OF DEMOCRATIC E-GOVERNANCE

There is nothing inherently "democratic" in governance. It can be and historically has been performed in various ways that cannot be called democratic. In the history of the institution of community governance the early modern times represent the era of elite control that since the 19th century began to transform into a conventional democratic mold having expression in the form of civic rights and representative system of government. This was followed by the rise of professionalism and managerialism in the 20th century.

Figure 1. Aspects of democratic e-governance (cf. Anttiroiko, 2004, p. 40)



The new phase that is emerging is in some descriptions and visions been called the era of citizen governance (Box, 1998; see also Hirst, 2000). Democratic governance requires that political institutions are capable of steering governance processes and affecting their outcomes, that citizens are given a chance to influence and participate in these processes in principle whenever they see fit, that governance processes themselves are made transparent, and that key actors are held accountable for their actions to political institutions and ultimately to society as a whole.

Institutional mediation tools of democratic governance are needed to facilitate civic involvement. A very fundamental aim is to supplement the representative system that is considered by many to be too hierarchical, inflexible, and distant from the point of view of ordinary citizens. Gross (2002) summarizes the basic requirements of e-democracy in the following way: citizens need to be able to access information, to discuss political issues, and to make decisions by voting electronically. Similar logic is followed in the UNDESA e-Participation Framework, which includes (a) increasing e-information to citizens, (b) enhancing e-consultation for deliberative and participatory processes, and (c) supporting e-decision making (UNDESA, 2005).

Conventional ways of political influence and participation in modern democracies include voting and campaigning in elections and being an active member of a political party or pressure group. Another category includes memberships of advisory committees or other bodies with a stake in policy processes and also various forms of client or customer involvement in implementation of public policies. More direct forms include voting in referendums and participation in the consultative or advisory bodies set up on an ad hoc basis. Lastly, there are various types of community group actions as well as political demonstrations that aim at changing public policy, and even various forms of civil disobedience (Birch, 1996). One way to systematize these forms on the basis of the degree of citizen influence is presented in Figure 2.

Figure 2 systematizes citizen influence within a continuum, the two extreme ends being a chance to obtain information and a direct political decision-making. The idea behind this is originally a rather old concept of eight rungs on a ladder of citizen participation which proceeds from nonparticipation through a certain degree of tokenism to the exercise of direct citizen control (Bishop & Davis, 2002).

Methods of democratic e-governance are based on the functions they serve in policy processes. On the basis of the functions of institutional and technological mediation tools, the methods applicable in e-governance can be presented as in Table 1 (Anttiroiko, 2004; see also Becker & Slaton, 2000; Gronlund, 2002; Gross, 2002; 6, 2001).

Figure 2. Continuum of citizen influence (applied from Bishop & Davis, 2002, pp. 20-21)

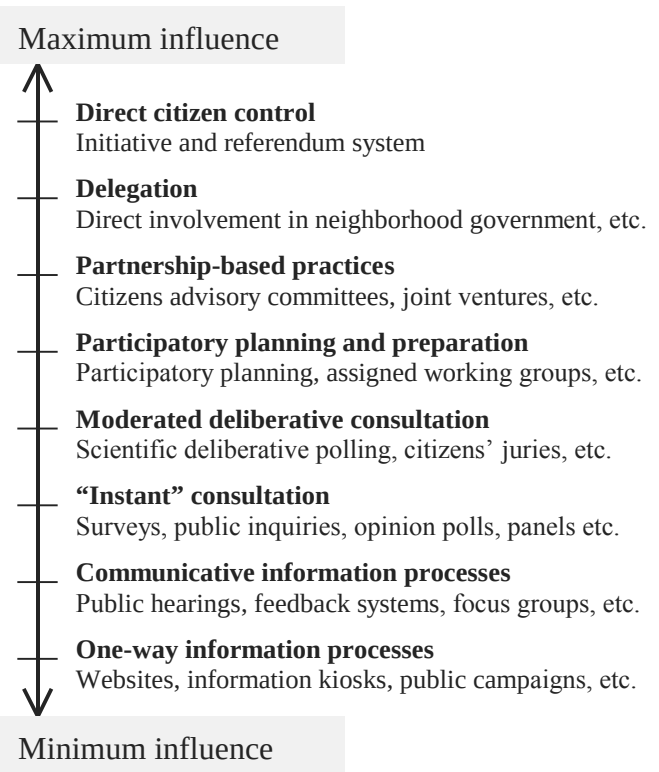


Table 1. Methods and forms of democratic e-governance

1. **Facilitating information processes**
• *Presenting, disseminating and sharing information (Websites, blogs, e-BBS, etc.)*
• *Collecting and processing data (e.g., database management tools and e-document management)*
• *Facilitating communicative or two-way information processes (e-mails and e-feedback systems)*
2. **Supporting communication and negotiation**
• *Facilitating discussion and interaction (electronic discussion forums, e-mails, mobile communication, etc.)*
• *Generating understanding and awareness (idea-generating tools, simulations, etc.)*
• *Facilitating citizen-expert interaction (e.g., consensus conferences)*
3. **Citizen consultation and involvement in preparation and planning**
• *Consultative referendum*
• *Moderated deliberative polling (scientific deliberative polling, electronic town meeting, etc.)*
• *Other forms of citizen consultation (e-citizens' juries, standing e-panels, etc.)*
• *Participatory planning*
• *Modeling decisions and advising on possible consequences (expert systems, decision support systems, etc.)*
4. **Community-based deliberation and participation**
• *Virtual communities and cyberassociations*
• *Community networks and local associations*
• *Local and neighborhood governments*
5. **Political transactions and decision-making**
• *Making proposals and initiatives (e-initiatives, e-petitions, etc.)*
• *Participating and voting in elections (e-electioneering, e-voting, etc.)*
• *Making decisions (initiative and referendum processes, including e-referendums)*
6. **Implementation and service processes**
• *Various forms of user democracy (e-feedback systems, e-vouchers, etc.)*

Since the 1970's small groups of academics, community organizers, activists, government officials, and media professionals have been experimenting with electronic media and ICTs. Since then new democratic practices have emerged and many experiments have been conducted in different parts of the world. Almost all of these have shown how interested ordinary people have been in the opportunity to participate, and how much they have appreciated being consulted and included in these processes (Becker & Slaton, 2000; see also Tsagarousianou, Tambini & Bryan, 1998).

FUTURE TRENDS

Democratization of governance is conditioned by such pervasive changes as globalization, technological development, new forms of social organization, and increased multiculturalism and individualism. It may be that a hybrid model of democratic governance is in the making, in which the new platforms and applications are evolving along with the societal and governmental structures.

Becker and Slaton (2000) have claimed that despite some setbacks genuine democracy will increase in degree and scope. This development is supported by such factors as greater influence of social movements, new methods of direct democracy, wider use of consensus building mechanisms, and new forms of e-enabled democratic political organization. These developments may give rise to new forms of democracy, which extend the application of the principles of democracy to nonhierarchically organized interaction and institutional arrangements in collective decision-making. Such a new aspect of democracy may be called network democracy.

A technological trend that is likely to have a deep long-term impact on the preconditions of democracy relates to ubiquity. The most important aspect of the emerging ubiquitous society is expected to be the new forms of interaction and transaction that are possible anywhere and at any time due to the utilization of networks and applications based on ubiquitous technologies. Thus, sooner or later various expressions of u-democracy will most likely be brought to the global democracy reform agenda.

To sum up, new technological mediation tools and the Internet in particular may prove vital in rethinking conceptions of democratic governance, giving rise to such new conceptions as ubiquitous network democracy. However, only modest democratic gains can be achieved through electronic means unless a radical redesign of democratic institutions is accomplished (Anttiroiko, 2003).

CONCLUSION

A new discourse about democratization of public governance reflects a gradual transition from the state-centric model of governance and managerial and market-based views of new public management to the politically oriented coalition-building and stakeholder-involving new public governance model that is rooted in the values of authentic democracy (cf. Barber, 1984; Becker & Slaton, 2000; Reddel, 2002). In this sense democratic e-governance is a technologically mediated interaction in governance processes in which special attention is paid to citizens' chances to participate and influence public policies.

At a practical level democratization of public governance may give rise to a network democracy and along with it technological development may provide fruitful soil for the development of ubiquitous democracy, which may converge into a new paradigm in democratic theory. Such a hybrid form of democracy may be characterized as ubiquitous network democracy, in which democratic principles are extended to nonhierarchical network-based governance with the help of ubiquitous technologies.

REFERENCES

6, P. (2001). E-governance. Do digital aids make a difference in policy making? In J. E. J. Prins (Ed.), *Designing e-government. On the crossroads of technological innovation and institutional change* (pp. 7-27). The Hague: Kluwer Law International.

Anttiroiko, A.-V. (2003). Building strong e-democracy. The role of technology in developing democracy for the information age. *Communications of the ACM, 46*(9ve), 121-128.

Anttiroiko, A.-V. (2004). Introduction to democratic e-governance. In M. Malkia, A.-V. Anttiroiko & R. Savolainen (Eds.), *e-Transformation in governance* (pp. 22-49). Hershey, PA: Idea Group Publishing.

Barber, B. (1984). *Strong democracy: Participatory politics for a new age*. Berkeley, CA: University of California Press.

Becker, T. & Slaton, C. D. (2000). *The future of teledemocracy*. Westport, CT: Praeger.

Birch, A. H. (1996). *The concepts and theories of modern democracy*. New York: Routledge.

Bishop, P. & Davis, G. (2002). Mapping public participation in policy choices. *Australian Journal of Public Administration, 61*(1), 14-29.

Box, R. C. (1998). *Citizen governance. Leading American communities into the 21st century*. Thousand Oaks, CA: Sage.

Gross, T. (2002). e-Democracy and community networks: Political visions, technological opportunities and social reality. In A. Gronlund (Ed.), *Electronic government: Design, applications & management* (pp. 249-266). Hershey, PA: Idea Group Publishing.

Gronlund, A. (Ed.) (2002). *Electronic government: Design, applications & management*. Hershey, PA: Idea Group Publishing.

Hirst, P. (2000). Democracy and governance. J. Pierre (Ed.), *Debating governance. Authority, steering, and democracy* (pp. 13-35). Oxford: Oxford University Press.

Jessop, B. (1998). The rise of governance and the risks of failure: The case of economic development. *International Social Science Journal, L*(1), 29-45.

Kolsaker, A. (2006). Reconceptualising e-government as a tool of governance: The UK case. *Electronic Government – An International Journal, 3*(4), 347-355.

Kooiman, J. (Ed.) (1993). Modern governance. *New government-society interactions*. London: Sage.

OECD (1995). *Governance in transition. Public management reforms in OECD countries*. Paris: Organisation for Economic Co-operation and Development.

Pardo, M. d. C. (2002). New information and management technologies for the 21st centrury public administration. In *Proceedings of the Workshop Report of Twenty-fifth International Congress of Administrative Sciences: Governance and Public Administration in the 21st Century: New Trends and New Techniques* (pp. 83-99). Brussels: IIAS.

Pierre, J. (2000). Introduction: Understanding governance. J. Pierre (Ed.), *Debating governance. Authority, steering, and democracy* (pp. 1-10). Oxford: Oxford University Press.

Pierre, J. & Peters, B. G. (2000). *Governance, politics and the state*. Houndmills: Macmillan.

Reddel, T. (2002). Beyond participation, hierarchies, management and markets: 'New' governance and place policies. *Australian Journal of Public Administration, 61*(1), 50-63.

Seneviratne, S. J. (1999). Information technology and organizational change in the public sector. G. David Garson (Ed.), *Information technology and computer applications in public administration: Issues and trends* (pp. 41-61). Hershey, PA: Idea Group Publishing.

Stoker, G. (1998). Governance as theory: Five propositions. *International Social Science Journal, L*(1), 17-28.

Tsagarousianou, R., Tambini, D., & Bryan, C. (Eds.) (1998). *Cyberdemocracy: Technology, cities and civic networks*. New York: Routledge.

UNDESA (2005). UN global e-government readiness report 2005: From e-government to e-inclusion. Department of economic and social affairs. Division for public administration and development management, United Nations. Retrieved June 13, 2008, from http://unpan1.un.org/intradoc/groups/public/documents/un/unpan021888.pdf

Walsh, L. (2007). Extending e-government and citizen participation in Australia through the Internet. *Encyclopedia of digital government* (Vol. II, pp. 812-818). Hershey, PA: Idea Group Reference.

KEY TERMS

E-Democracy: Electronic democracy (e-democracy) as a tool-oriented conception of democracy refers to new democratic practices in which ICTs and innovative institutional arrangements are utilized (cf. teledemocracy).

E-Governance: In the public sector context governance refers to coordination, interaction, and institutional arrangements which are needed to pursue collective interest in policy-making, development and service processes in the context of nonhierarchically organized stakeholder relations. Electronic governance or e-governance is technologically mediated communication, coordination, and interaction in governance processes.

E-Government: Electronic government (e-government) is government's use of information and communication technologies, particularly Web-based applications, to support responsive and cost-effective government by facilitating administrative and managerial functions, providing citizens and stakeholders with convenient access to government information and services, facilitating interaction and transactions with stakeholders, and providing better opportunities to participate in democratic institutions and processes.

Informatization: The unprecedented growth in the speed and quantity of information production and distribution and the increased role of ICT-assisted knowledge processes, systems, and networks in society.

Network: Networks are loose sets of actors who work together in order to promote their interests within a common operational framework, which is held together by some shared interests, reciprocity, and trust. In their most characteristic form networks are flexible ways of organizing activities that require competences of several independent actors.

Network Democracy: Innovative application of the principles of democracy in the nonhierarchical network-based public governance.

New Public Management (NPM): Neo-liberally oriented public management doctrine based on a market-oriented view stating that, instead of direct political control and hierarchies, public organizations should rely on indirect control—that is, market-based coordination—in the interaction between public organizations and their environments. It emphasizes the efficiency and effectiveness of public organizations, customer focus in provision of public services, and market-based conditioning frameworks, such as privatization, competition, and contracting out.

Teledemocracy: A normative theory on democracy that is dedicated to greater, improved, and more direct citizen participation and influence in all aspects of government. It is based on the notion of transformational politics, which emphasizes the necessity to fine-tune a democratic system to meet the requirements of an increasingly complex information society. This is evident in how it favors the utilization of new ICTs in democratic processes, in such forms as electronic town meeting, scientific deliberative polling, and new democratic use of the Internet. The most prominent academic developer and advocate of teledemocracy is Professor Ted Becker, Auburn University, Alabama.

U-Democracy: Ubiquitous democracy (u-democracy) refers to new forms of democracy in which ubiquitous technologies are utilized.

Departure of the Expert Systems Project Champion

Janice C. Sipior
Villanova University, USA

INTRODUCTION

This article discusses the expert system (ES) project champion by examining the experiences of Ciba-Geigy Corporation with an ES project, impeded by the departure of the project champion. The OpBright Expert System, developed to support the identification of appropriate optical brightener products by sales representatives, was intended to provide a competitive advantage through superior customer service. With the promotion and transfer of the vital force committed to the project's success, the ES encountered a stalemate. The difficulties in maintaining momentum for the ES without a project champion are discussed. Finally, suggestions are presented to guide organizations away from the same fate.

BACKGROUND

The role of project champion has been recognized as vital to successful project development since the time of Schon's (1963) seminal work. A project champion for information systems is defined as "a key *individual*, whose *personal* efforts in support of the system are critical to its successful adoption" (Curley & Gremillion, 1983, p. 206). The project champion, for ES projects in particular, is recognized as critical to the successful application of this technology (Hayes-Roth & Jacobstein, 1994; Sipior, 2000: Wong, 1996). Champions of ES projects differ from those of other projects due to the necessity to identify, document, and distribute knowledge and expertise, facilitating knowledge-sharing. The characteristics of project champions are discussed in the next section.

Formal Position

A project champion is frequently an executive from the area of application (Willcocks & Sykes, 2000), but may come from external organizations, such as a consultants or vendors (Thomas, 1999). Champions may be managers (Beath, 1991); or hold other formal positions (Mayhew, 1999; Pinto & Slevin, 1989; Thomas, 1999). Surprisingly, champions rarely come from formal IT functions (Martinsons, 1993; Willcocks & Sykes, 2000) and may even view IT managers as too conservative, adversaries to technological innovations, and even inept (Beath & Ives, 1988). Rather than being assigned to the role, interest and personal conviction to a project compel the champion to emerge (Pinto & Slevin, 1989; Schon, 1963). Formally appointing an individual could actually lead to his demise (Howell & Higgins, 1990). Once convinced, the champion exhibits an entrepreneurial spirit (Bolton & Thompson, 2000; Pinto & Slevin, 1989; Schon, 1963).

Leadership Qualities

The champion tends to go well beyond job responsibilities, and may even go against management directives (Beath, 1991; Curley & Gremillion, 1983). Champions are characterized as more than ordinary leaders. They exhibit transformational leadership behaviors (Howell & Higgins, 1990). Such leadership is particularly valuable for implementing systems intended to bring about organizational change (Beath, 1991; Landers, 1999), such as redefining responsibilities, realigning lines of authority, shifting power centers, and adjusting reward schemes. As knowledge repositories, ES certainly has the potential to invoke change of this nature.

Base of Power

Some level of power is held by champions (Mayhew, 1999; Pinto & Slevin, 1989), attributable to formal position or personal relationships. Diminished power can result in project failure (Scott & Vessey, 2002). Champions are perceived as influential or prestigious by organizational members (Curley & Gremillion, 1983). This perception by others may be the result of a planned influence strategy to attract followers (Schon, 1963). Such influence strategies include impression building, rational justification, assertion, or persuasive communication (Howell & Higgins, 1990). Although activities of champions may be intentionally fostered, their influence tactics are not always regarded in a positive light (Beath, 1991).

Visionary Perspective for Change

The champion is willing to put himself on the line, risking his reputation, to complete the project. The champion serves

as a visionary and directs his energies to bring about change to achieve that vision (Landers, 1999; Willcocks & Sykes, 2000). Primary among the influence strategies is persuasive communication (Sumner, 2000). The vision must be clearly communicated in order that others understand and support the vision (Kotter, 1995). An unrealistic or misunderstood vision can result in failure well after the project is underway (Royer, 2003).

A CASE STUDY OF THE PROJECT CHAMPION AT CIBA-GEIGY

Ciba-Geigy Corporation, an international chemical manufacturing firm headquartered in Basel, Switzerland, continually strives to gain market position by fostering their progressive image. Ciba-Geigy emphasizes customer service, especially important to the Dyestuffs and Chemicals Division. This division produces over 2,000 products including fabric dyes, optical brighteners, and industrial chemicals, representing approximately 20% of corporate sales. The OpBright Expert System, developed to support the identification of appropriate optical brightener products, was championed by the vice president (VP) of the division as providing benefits realizable from managing internal knowledge. The VP was convinced, as was found in previous research, that effective knowledge management can impact business performance (Alavi & Leidner, 1999; Hansen, Nohria, & Tierney, 1999; Zack, 1999). Included among the anticipated benefits of OpBright are gaining competitive advantage, faster response to customers, consistent quality customer service, training new salespeople, and managing product expertise, as discussed in the following sections.

Gain Competitive Advantage

As a leading dyestuffs and chemicals producer, Ciba-Geigy recognizes the value of IT as an important means for gaining competitive advantage. Continually striving to gain market position by fostering their progressive image, Ciba-Geigy has emphasized the need to utilize IT in direct marketing. The use of laptops by the sales force, championed by the VP, provides a highly visible means for projecting this image as well as enhancing sales force performance.

Respond More Quickly to Customers

Improved communication between the sales force and the division office, in terms of such factors as speed, receipt and response, and content completeness, was realized through the use of laptops by the sales force. For example, access to the online order processing inventory and sales service system enables sales representatives to complete a sales transaction more quickly, increasing employee productivity and providing more responsive and effective customer service. This taste of success led the VP to seek further improvement. In informal meetings with the corporation's computer vendor, IBM, the VP became convinced that customer support could be enhanced through the implementation of an ES, a recognized benefit of expert system applications (Mattei, 2001).

Provide Consistent Quality Customer Service

The VP had the insight to identify the importance of offering fast, expert advice regarding the appropriate use of optical brighteners for individual customer's applications at the time of on-site sales calls. Optical brighteners are used for a wide variety of end-products. For textiles, paper products, and detergents, optical brighteners are applied to enhance coloring, that is, to make "whites whiter and brights brighter." Non-textile applications include testing for leaks, such as those in automotive parts. Salespeople are thus challenged to make appropriate and specific recommendations concerning a wide range of applications, wherein the factors to consider can vary widely. The inability of a salesperson to answer customers' questions can result in delayed or lost sales. Recognizing this impact, the VP championed the expert system as a means of increasing sales profitability. Individual customer questions could be addressed on the spot while maintaining consistency and quality in responses.

Train the Sales Force

By managing and distributing knowledge about optical brightener product features, areas of application, and troubleshooting solutions, the sales force is able to develop a greater understanding of the optical brightener product line and technical characteristics. New salespeople benefit by having unconstrained access to a "technical expert". Sales force training is thereby enhanced during formal training sessions and while on the job.

Manage Critical Product Expertise

The VP envisioned an entire family of ES, for all optical brightener applications, would be developed in the future. The domain for the first ES was appropriately narrowed to include the application of optical brighteners to fabrics only. An expert in this area of application was identified. This individual has extensive experience with the optical brightener product category, having served as a customer support technician and troubleshooter for over 15 years. His knowledge about customer requirements, properties of fabrics, and application processes enabled him to recommend appropriate optical brightener products based on features of

those products. This critical product expertise was captured, documented, and can be distributed through OpBright. Such astute management can preclude the loss of valuable corporate knowledge and expertise resulting from normal turnover or absence (Mattei, 2001), promotion (Prietula & Simon, 1989), or retirement.

Momentum without a Project Champion

The VP classified the development of the ES as a research and development (R&D) effort. No formal cost/benefit analysis was performed. Even for R&D projects, a formal analysis is recommended (Brenner & Tao, 2001). However, the cost of this first-time project was viewed as an investment in experience to be applied to future areas of application, an argument commonly made by champions in securing funding (Sharpe & Keelin, 1998). The VP was convinced that the impact of this technology would far outweigh the initial investment.

OpBright encountered a stalemate shortly after its completion. The VP was promoted and transferred to another division. The incoming VP reviewed the OpBright expert system project and restated the objective to develop a *prototype* ES. The objective was met as the prototype had been developed. Without a replacement project champion, OpBright remains at a standstill.

FUTURE TRENDS

Lessons learned, from the Ciba-Geigy case study, to avoid pitfalls attributable to the departure of the champion are discussed in the following sections. These insights may serve to guide future research; focusing on identifying strategies organizations may employ to nurture the continuation of the role of project champion.

Incorporate Expertise Management in Strategic Planning

An assessment of areas of expertise critical to organizational processes and the potential for applying ES technology, should be included within the strategic planning process. Expertise management thereby becomes a formalized managerial activity (Sipior & Garrity, 1990). Incorporating expertise management in strategic planning is clearly unique to ES in particular, differentiating ES project champions from champions of technological innovations in general.

Secure the Support of Top Management

It is well recognized that securing the support of top management is a prerequisite to the long-term success of ES projects (Sipior & Volonino, 1991; Wong, 1996). Broad support for technological innovation from the top individual alone is insufficient unless it is translated into support for specific applications of ES technology by management levels below him.

Top management support is more likely if the ES fits with corporate strategy. By incorporating expertise management within strategic planning, this fit becomes more likely (Zack, 1999). In turn, this fit will garner broader organizational commitment (Meador & Mahler, 1990). To gain the necessary support, a results-oriented approach is preferable since it enables management to buy into the impact ES can have, rather than focusing on the technology itself (Sipior & Volonino, 1991). For management support to be on-going, the results should have continued benefit.

Secure the Support of the Next Generation of Top Management

Top management support is not sufficient, unless the support of the next generation of top management is secured (Kotter, 1995). If successors are ambivalent about the ES project, as was the case for Ciba-Geigy, they certainly will not take the initiative to understand this technology and maintain project momentum.

Recognize, Support, and Nurture a Project Champion

As discussed, a project champion tends to emerge, rather than be assigned to the role. Evidence suggests both the inability to nurture such individuals as well as instances wherein this individual has been successfully developed (Schon, 1963). When a champion does not emerge naturally, a company may be able to "*find* or *make* a champion for the system" (Curley & Gremillion, 1983, p. 207).

Formalize Project Measurement, Monitoring, and Follow-up

Formalization of expert system project measurement, monitoring, and follow-up evaluation can improve the probability of project success (Brenner & Tao, 2001). Project measurement should include an evaluation at each phase of the project: (1) the validity of the technology to be employed in terms of delivering its claimed capabilities, (2) the benefits to be derived by its application, and (3) the products and services to be produced by employing the technology (Brenner & Tao, 2001). For each phase, project characteristics should be evaluated. The type of idea that generated the project could be rated from high of three to low of zero for a technological advance, new or novel technology, new twist on a known technology, or no technological advance. Additional

characteristics to consider include expertise and capabilities gained, time saved, R&D dollars saved, intellectual property expected from the project, importance of the project objective, commercial impact such as increased sales revenue, technical leverage, and internal and external relationship building. Changes in the project should be monitored as the project advances. At project implementation, the estimates from the initial project evaluation should be compared to actual project performance data. Formalizing measurement, monitoring, and follow-up evaluation forces project accountability in terms of delivering the intended results.

CONCLUSIONS

The experiences of Ciba-Geigy underscore the critical importance of the role of project champion in ES development. Conversely, the loss of this individual can threaten the very existence of a project. Organizations should thus take heed and devote attention to harnessing the enthusiasm and drive of ES project champions for advantage, before it is too late to preclude the loss of momentum generated by these individuals. Future research may be directed toward exploring the reasons for failure of projects supported by champions, and identifying strategies organizations may employ to nurture the continuation of the role of project champion for ES.

REFERENCES

Alavi, M., & Leidner, D.E. (1999). Knowledge management systems: Issues, challenges, and benefits. *Communications of the Association for Information Systems, 1*(2), Article 1.

Beath, C.M. (1991). Supporting the information technology champion. *MIS Quarterly, 15*(3), 355-372.

Beath, C.M., & Ives, B. (1988). The information technology champion: Aiding and abetting, care and feeding. *Proceedings of the 21st Annual Hawaii International Conference on Systems Sciences IV* (pp. 115-123), Kailua-Kona, Hawaii, USA.

Bolton, B., & Thompson, J. (2000). A breed apart. *Director, 53*(10), 54-57.

Brenner, M.S., & Tao, J.C. (2001). *Research Technology Management, 44*(3), 14-17.

Curley, K.F., & Gremillion, L.L. (1983). The role of the champion in DSS implementation. *Information and Management, 6*, 203-209.

Hansen, M.T., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? *Harvard Business Review, 77*(2), 106-119.

Hayes-Roth, F., & Jacobstein, N. (1994). The state of knowledge-based systems. *Communications of the ACM, 37*(3), 27-35.

Howell, J.M., & Higgins, C.A. (1990). Champions of technological innovation. *Administrative Science Quarterly, 35*(2), 317-341.

Kotter, J.P. (1995). Leading change: Why transformation efforts fail. *Harvard Business Review, 73*(2), 59-67.

Landers, T.L. (1999). Are you a project champion? *Modern Materials Handling, 54*(2), 33.

Martinsons, M.G. (1993). Cultivating the champions for strategic information systems. *Journal of Systems Management*, 31-34.

Mattei, M.D. (2001). Using "expert systems" for competitive advantage. *Business and Economic Review, 47*(3), 17-20.

Mayhew, D.J. (1999). Business: Strategic development of the usability engineering function. *Interactions, 6*(5), 27-34.

Meador, C.L., & Mahler, E.G. (1990). Choosing an expert systems game plan. *Datamation*, (August), 64-69.

Pinto, J.K., & Slevin, D.P. (1989). The project champion: Key to implementation success. *Project Management Journal, 20*(4), 15-20.

Prietula, M.J., & Simon, H.A. (1989). The experts in your midst. *Harvard Business Review, 67*(1), 120-124.

Royer, I. (2003). Why bad projects are so hard to kill. *Harvard Business Review, 81*(2), 48-56.

Schon, D.A. (1963). Champions for radical new inventions. *Harvard Business Review, 41*(2), 77-86.

Scott, J.E., & Vessey, I. (2002). Managing risks in enterprise systems implementations. *Communications of the ACM, 45*(4), 74-81.

Sharpe, P., & Keelin, T. (1998). How SmithKline Beecham makes better resource-allocation decisions. *Harvard Business Review, 76*(2), 45-57.

Sipior, J.C. (2000). Expert system stalemate: A case of project champion departure. *Information Resources Management Journal, 13*(4), 16-24.

Sipior, J.C., & Garrity, E.J. (1990). The potential of expert systems for competitive advantage. *Proceedings of the Decision Sciences Institute*, San Diego, CA USA.

Sipior, J.C., & Volonino, L. (1991). Changing the image of expert systems from sensational to organizational tools. *Proceedings of the 24th Annual Hawaii International Conference on System Sciences*, Kauai, Hawaii, USA.

Sumner, M. (2000). Critical success factors in enterprise wide information management systems projects. *Proceedings of the 2000 ACM SIGCPR Conference on Computer Personnel Research*, Chicago, IL, USA.

Thomas, J. (1999). Share and share alike. *Logistics Management and Distribution Report, 38*(3), 44-48.

Willcocks, L.P., & Sykes, R. (2000). The role of CIO and IT function in ERP. *Communications of the ACM, 43*(4), 33-38.

Wong, B.K. (1996). The role of top management in the development of expert systems. *Journal of Systems Management,* 36-40.

Zack, M.H. (1999). Managing codified knowledge. *Sloan Management Review, 40*(4), 45-58.

KEY TERMS

Assertion: Authoritatively convince others of the project's potential benefits so they dedicate their efforts to the project.

Charismatic Behavior: Captivate others into believing in the project as the champion himself does.

Impression Building: Portray outcomes of endeavors as highly positive achievements to promote an image of competence and success.

Inspirational Behavior: Influence others by using emotional appeals, and vivid and persuasive images, to elevate their performance.

Intellectual Stimulation: Challenge others to aspire to imaginative use of their individual skills.

Persuasive Communication: Rely more on persistence, rather than clear and concise arguments, to attain agreement.

Rational Justification: Analyze how the project advances goals and objectives and upholds values of the organization.

Transformational Leaders: Inspire others, through charisma, inspiration, intellectual stimulation, and individualized consideration, to transcend their own self-interests for a higher collective purpose.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 797-801, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Deploying Pervasive Technologies

Juan-Carlos Cano
Technical University of Valencia, Spain

Carlos Tavares Calafate
Technical University of Valencia, Spain

Jose Cano
Technical University of Valencia, Spain

Pietro Manzoni
Technical University of Valencia, Spain

INTRODUCTION

Communication technologies are currently addressing our daily lives. Internet, fixed-line networks, wireless networks, and sensor technologies are converging, and seamless communication is expected to become widely available. Meanwhile, the miniaturization of devices and the rapid proliferation of handheld devices have paved the path towards pervasive computing and ubiquitous scenarios.

The term *ubiquitous and pervasive computing* refers to making many computing devices available throughout the physical environment, while making them effectively invisible to the user (Weiser, 1991). Thanks to advances in the devices' processing power, extended battery life, and the proliferation of mobile computing services, the realization of ubiquitous computing has become more apparent, being a major motivation for developing location and context-aware information delivery systems.

Strongly related to ubiquitous computing is *context-aware computing*. In context-aware computing, the applications may change or adapt their functions, information, and user interface depending on the context and the client's profile (Weiser, 1993). Many research centers and industries are actively working on the issues of context-awareness or more generally on ubiquitous computing (Baldauf, Dustdar, & Rosenberg, 2007). In particular, several proposals focus on smart spaces and intelligent environments (Harter, Hopper, Steggeles, Ward, & Webster, 1999; Kindberg et al., 2002; Smart-its, 2007), where it is expected that smart devices all around us will maintain updated information about their locations, the contexts in which they are being used, and relevant data about the users.

Clearly, contextual services represent a milestone in today's mobile computing paradigm, providing timely information anytime, anywhere. Nevertheless, there are still few examples of pervasive computing environments moving out from academic laboratories into our everyday lives. This occurs since pervasive technologies are still premature, and also because it is hard to define what a real pervasive system should be like. Moreover, despite the wide range of services and potential smart applications that can benefit from using such systems, there is still no clear insight about a realistic killer application.

BACKGROUND

Pervasive computing has been in development for more than 15 years. In this section we briefly review some of the most relevant prototypes.

Various companies are already working to extend wireless technologies that will seamlessly connect to other nearby devices. However, despite the wide range of services and potential smart applications that can benefit from using such tools, there is still no clear understanding about a realistic killer. One critical question that still needs to be addressed is the identification of business scenarios that can move ubiquitous computing from academic and research laboratories into our everyday lives.

Tourism was one of the first areas to yield the business application area for the development of such potential applications. To this end, context-aware services combined with content-oriented applications could exploit wireless technology to provide personalized tours that could guide and assist tourists in museums or historical sites. One on the earlier prototypes of a mobile context-aware tour guide is the Cyberguide project (Abowd et al., 1997). The Cyberguide prototype uses the current location of users to provide visitors with services concerning location and information. For indoor applications, Cyberguide uses infrared technology as a positioning solution. On the other hand, for outdoor applications, they replace the infrared positioning module with a GPS unit. Cyberguide presents an innovative architecture, which mainly focuses on the development of location-aware applications. However, further efforts are needed to improve on context awareness. Systems similar to Cyberguide have

also been proposed by other researchers, including the GUIDE (Davies, Mitchell, Cheverst, & Blair, 1998) project proposed at Lancaster University. Cyberguide and GUIDE were influenced by earlier location-aware works such as the PARCTab at Xerox PARC (Want, Hopper, Falcao, & Gibbons, 1992), the InfoPad project at Berkeley (Long, Kooper, Abowd, & Atkeson, 1996), and the Personal Shopping Assistant at AT&T (Asthana, Cravatts, & Krzyzanouski, 1994).

The CoolTown project (Kindberg et al., 2002) at HP Laboratories focuses on building ubiquitous computing systems by embodying Web technologies into the physical environment. The Websign project (Pradhan, Brignone, Cui, McReynolds, & Smith, 2004) is a component of the CoolTown research program which allows users to visualize services related to physical objects of interest. While Websign could be adapted to offer tourist guide services, its intended use is more general, providing user interactions for services associated with physical objects. The Rememberer tool (Fleck et al., 2002) is another interesting approach, which, similarly to the CoolTown project, chooses museums as an environment to implement context-aware applications. Rememberer is a tool that offers visitors of museums services to record their visits. Each record, which can be consulted after the user's visit, consists of a set of Web pages with multimedia data describing the visit. The location of the visitor is identified using infrared technology and RFID sensors. Other works related to the CoolTown project include Spasojevic, Mirjana, and Kindberg (2001) and Semper and Spasojevic (2002).

The Digital museum project (Sakamura, 1998) at Tokyo University uses smartcards to detect the proximity of visitors and then provide information about the exhibited objects. The information provided can be based on a static profile stored previously in the smartcard. Similar work has also been done by Davin and Ing (1999) where infrared infrastructure and wireless LAN connections were used for connectivity and location awareness respectively.

Cano, Ferrández, and Manzoni (2005) used a network simulator to evaluate the feasibility and performance of using Bluetooth as the underlying networking technology to establish context-aware services. They compare results obtained from simulation with those obtained from a real test-bed. Authors observed that simulation results show a much smoother behavior than those obtained in real experiments.

The Massachusetts Institute of Technology (MIT) has a project called Oxygen (Rudolph, 2001). They envision a future where ubiquitous computing devices are as freely available and easily accessible as oxygen is today.

More recently, many of the ubiquitous and pervasive proposals are characterized by their focus on real-world deployments. In fact, some of the most valuable experiences when dealing with pervasive systems come not only from the design and implementation of a particular system, but also from the experience of trying to move those systems out of the laboratory into the real world.

Thomas, Jakob, and Mads (2006) create an interactive and pervasive system to be used in hospitals. The authors highlighted how issues that seem to be trivial in the laboratory, such as calibrating location systems or finding places for interactive displays, might become major obstacles when deploying systems in real-world settings. Other interesting works related to smart and interactive spaces can be found in Fitton et al. (2005) and Oliver et al. (2006).

Overall, we can state that, although new technologies are emerging and a number of leading technological organizations are exploring pervasive computing, the most crucial objective is not necessarily to develop new technologies, but to find ways to integrate existing technologies with a wireless infrastructure.

EXPERIENCES ON DEVELOPING PERVASIVE SERVICES

Next we present our vision on ubiquitous computing by reporting our experiences building the BlueHospital system, a pervasive prototype that provides context-aware information and location-based services to clinicians on hospitals' recovery wards. BlueHospital leverages Bluetooth technologies and Java services to offer patient information to clinical personnel based on the patient's profile and the clinicians' preferences and requirements.

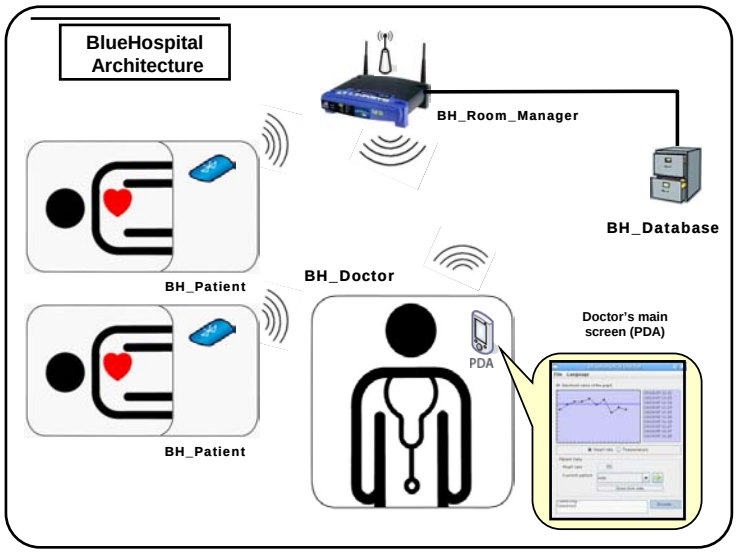
When developing a suite for a pervasive computing system, two main lessons were confirmed from our previous experience. First, we require a low-cost solution that offers computation and communication capabilities to an ordinary object. Second, minimizing power consumption and size are mandatory in order to make more apparent the realization of ubiquitous computing. To this end, we developed our own inexpensive platform prototype for ubiquitous computing, which has been implemented based on commercial off-the-shelf components.

The overall network architecture is based on the cooperation of an edge wireless network and a core wired network. The edge part is based on Bluetooth technology alone. The core network is based on a fixed 100 Mbps Ethernet local area network used to connect the edge infrastructure with the central database.

The system considers four types of entities: hospital patients (BH_Patient), room managers (BH_Room_Manager), clinical personnel (BH_Doctor), and the central database server (BH_Database_Server). Figure 1 shows a pictorial representation of the BlueHospital architecture.

A doctor provided with a Bluetooth-enabled PDA is the basic example of a mobile BH_Doctor entity. BH_Doctors are connected to the central database through the room man-

Figure 1. The BlueHospital system architecture



ager to receive information about patients, including their case history as well as comparable cases. Doctors can make diagnoses faster using all the information available at the central database, choosing the best therapy or medicine for each particular patient. There is a BH_Patient associated to each patient who has been admitted and allocated in a recovery ward. The BH_Patient connects to the room manager to register with the central database. The data being monitored consists of frequent measurement of body temperature and heart rate. There is a BH_Room_Manager associated to every recovery ward acting as a bridge between doctor and patient sensors to the central database and vice versa. Finally, the central database is able to generate precise patient profiles based on clinicians' requests and preferences.

Being a context-aware system, BlueHospital is able to provide valuable information to clinical personnel without any user interaction. When a clinician is visiting patients at recovery wards, his application will automatically search for the room manager device, which will offer any new information of interest about the patient. If a user wants to look up new information or introduce information in the system, he must send the request together with the user profile that was entered in the initial configuration process to the manager—that is, the user is a doctor, a nurse, and so forth. Knowing the user profile, the room manager can process the request combining the user profile with some additional information, and it will send it to the central data server. There the request is logged and processed, and a reply is returned to the room manager which relays it to the clinician. All the process takes place automatically, and clinicians can change their profile at any time to receive more details about the patients in their own language and adapted to their device (i.e., mobile phone, PDA, or laptop).

The BlueHospital system is also able to track clinical personnel and patients in our location- and context-aware infrastructure. Instead of using some commercial system that is perhaps more precise though expensive, we delegate to Bluetooth the location functionality. Since we only require location tracking of patients and clinicians within a room-level granularity, we developed a coarse-grained location system. All BH_Room_Manager devices in the recovery wards include a USB Bluetooth adapter with a range of up to 10 meters. When the manager discovers a new authorized Bluetooth-enabled device, it will accordingly update the tracking information within the BH_Database_Server. The proposed solution also allows using existing Bluetooth-enabled mobile phones owned by clinical personnel and patients for location tracking.

We also design our own BH_Patient device. It consists of a low-power microcontroller connected to a Bluetooth transceiver and a heart rate monitoring system through a serial UART interface, so that it can collect data from the measuring system and handle the wireless communication by sending the received data to the room manager device. In order to optimize battery life, the Texas Instruments MSP430 ultra-low-power family of microcontrollers has been selected because it is particularly well suited for power-constrained applications. A class 1 Bluetooth module has been used for wireless communication.

Concerning the BH_Room_Manager we adopt a low-cost solution based on WiFi routers that accommodate Bluetooth connectivity through an USB port, as well as a Fast Ethernet interface to connect to the central database.

Each operating ward is equipped with a BH_Room_Manager based on the ASUS WL-500g Premium wireless router (AsusTek-Computer, 2007), which contains—at a very reasonable price—almost every feature you may require to deploy a small computer network. When a connection comes in, the BH_Room_Manager will spawn a child process to deal with the mobile client's request. The child process will receive the client's profile, and it is sent to the central server. There, it will be logged and processed according to the client's profile. Eventually the required information will be sent back and passed on to the client. The parent process will carry on waiting for further client connections.

With respect to the BH_Doctor entity, we developed a PDA-oriented application using the standard Java APIs for Bluetooth wireless technology (JABWT). The application has been designed to be used by clinicians' PDAs.

The first step for each BH_Doctor application is initialization: the user must state his or her preferences as a basic set of input parameters, that is: (a) the type of device, (b) the preferred language, and (c) the user identification.

Once all the data is filled in and the user has been identified and authenticated, the application will search for information offered by the BH_Room_Manager through the Bluetooth inquiry process and service discovery protocol (SDP) searches. Once connected to the BH_Room_Manager, the application will send the stored profile to the room manager and it will eventually receive the information. Once all the information is received, the application will present it to the user. The user can then request new information (e.g., check in detail a given value) to select the period of time to be analyzed or a different patient.

Finally, the central BH_Database_Server stores in an SQL database all the information related to patients, as well as other information concerning the BlueHospital system.

The central database server has two main functions: (a) to attend connections for BH_Room_Managers requests, and (b) to manage the SQL database, that is, to handle all the information related to patients and clinicians, clinic schedules, and location tracking information. The BH_Database_Server starts and waits for a connection on the default server port. When it receives a connection request, a new process is spawned to attend it. If the connection request comes from a valid BH_Room_Managers, the data server will receive the user profile. The server will acknowledge each profile option received to provide a higher data consistency check. Once the whole profile is received, the server will log the request for security and for statistical data gathering. After logging completes, the data server will obtain (according to the received profile) the requested information from the SQL database. This information will be sent back to the room manager, which in turn will resend it to the BlueHospital application. Once the process of sending data to the room manager finishes, the connection will be dropped.

We developed both clients and server applications, providing routines to handle doctor requests about their patients, find empty recovery wards, and do profile logging and also content filtering. For example, the system may use the location tracking system to find a doctor, or it can notify a patient's ward about his or her medical treatment. Our system supports many similar scenarios aimed at providing awareness and enhancing communication.

FUTURE TRENDS

As we have previously shown, pervasive computing involves three converging areas of information and communications technology: computing devices, communications, and user interfaces. There are many visions for the future development of computing devices for pervasive systems from handheld units to near invisible devices with its own power supply. These devices will rely on some kind of wireless communication to act intelligently, being able to create the most effective form of connectivity in a given scenario. Communication can be achieved via both wireless and wired technologies, supporting switching between different networks. Finally, the user interface represents the union point between technology and human users. New devices are being developed for pervasive systems that will be able to sense information about users and the environment for advanced processing.

The appropriate combination of these three areas yield up to a wide range of applications, for example, healthcare, transport security, monitoring, tourism information, and smart networking, many of which may not yet have been identified. Emerging technologies supporting those applications include wearable computing, smart homes, intelligent environments, and augmented reality.

CONCLUSION

Research into the nature of pervasive computing has now been around for more than a decade. Nowadays pervasive applications exploit mobile wireless communication technologies to interconnect computing devices along with various sensing technologies, setting up a new kind of intelligent environment where applications can transparently search and use services without the users' intervention.

However, there are still engineering problems to be solved before the envisaged applications become a reality. Security problems, the cost of technologies, and fault tolerance issues represent only a small subset of the problems ahead. However, even though there are plenty of challenging problems to be solved, the expected reward will be great.

REFERENCES

Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R., & Pinkerton, M. (1997). Cyberguide: A mobile context-aware tour guide. *ACM wireless Networks, 3*(5), 421-433.

Asthana, A., Cravatts, M., & Krzyzanouski, P. (1994). An indoor wireless system for personalized shopping assistance. *Proceedings of the Workshop on Mobile Computing Systems and Applications.*

AsusTek-Computer. (2007). *ASUS WL500g premium wireless Internet router review.* Retrieved from *http://www.asus.com*

Baldauf, M., Dustdar, S., & Rosenberg, F. (2007). A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing, 2*(4).

Cano, J., Ferrández, D., & Manzoni, P. (2005). Evaluating Bluetooth performance as the support for context-aware applications. *Telecommunication Systems,* (5).

Davies, N., Mitchell, K., Cheverst, K., & Blair, G. (1998). *Developing a context sensitive tourist guide.* Technical Report Computing Department, Lancaster University, UK.

Davin, S., & Ing, L. (1999). Innovations in a technology museum. *IEEE Micro, 19*(6).

Fitton, D. et al. (2005). Rapid prototyping and user-centered design of interactive display-based systems. *IEEE Pervasive Computing, 4*(5).

Fleck, M., Frid, M., Kindberg, T., O'Brien-Strain, E., Rajani, R., & Spasojevic, M. (2002). Rememberer: A tool for capturing museum visits. *Proceedings of the Ubiquitous Computing International Conference.*

Harter, A., Hopper, A., Steggeles, P., Ward. A., & Webster, P. (1999). The anatomy of a context-aware application. *Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking* (MOBICOM).

Kindberg, T. et al. (2002). People, places, things: Web presence for the real world. *MONET, 7*(5).

Long, S., Kooper, R., Abowd, G., & Atkeson, C. (1996). Rapid prototyping of mobile context-aware applications: The Cyberguide case study. *Proceedings of the 2nd Annual International Conference on Mobile Computing and Networking.*

Oliver, S., Adrian, F., Nigel, D., Joe, F., Corina, S., & Jennifer, S. (2006). Public ubiquitous computing systems: Lessons from the e-campus display deployments. *IEEE Pervasive Computing, 5*(3).

Pradhan, S., Brignone, C., Cui, J., McReynolds, A., & Smith, M. (2004). *Websign: Hyperlinks from a physical location to the Web.* Retrieved from http://www.cooltown.hp.com/

Rudolph, L. (2001). Project oxygen: Pervasive, human-centric computing—an initial experience. *Proceedings of the 13th International Conference on Advanced Information Systems Engineering* (CAiSE).

Sakamura, K. (1998). Digital museum. *Journal of Information Processing Society of Japan, 39*(5).

Semper, R., & Spasojevic, M. (2002). The electronic guidebook: Using portable devices and a wireless Web-based network to extend the museum experience. *Proceedings of the Museums and the Web.*

Smart-its. (2007). *Interconnected embedded technology for smart artifacts with collective awareness.* Retrieved from http://www.smart-its.org/

Spasojevic, M., Mirjana, A., & Kindberg, T. (2001). *A study of an augmented museum experience.* Retrieved from http://www.cooltown.hp.com/

Thomas, H., Jakob, B., & Mads, S. (2006). Moving out of the lab: Deploying pervasive technologies in a hospital. *IEEE Pervasive Computing, 5*(3).

Want, R., Hopper, A., Falcao, V., & Gibbons, J. (1992). The active badge location system. *ACM Transactions on Information Systems, 10*(1).

Weiser, M. (1991). The computer for the 21st century. *Scientific American, 256,* 94-104.

Weiser, M. (1993). Some computer science problems in ubiquitous computing. *Communications of the ACM.*

KEY TERMS

Augmented Reality: A new technology that involves overlaying the real world with digital information. It will further blur the line between what is real and what is computer generated by enhancing what we see, hear, feel, and smell.

Bluetooth: A short-range low-power radio technology that allows multiple compatible devices to connect to each other to transmit voice and data.

Context-Aware Application: One of a set of applications that may change or adapt their functionality depending on the context, the client profile, and the user interface.

Fast Ethernet: A collective term for a number of Ethernet standards that carry traffic at the nominal rate of 100 Mbit/s.

Pervasive Computing: The next computing paradigm based on environments with information and communication technology—everywhere, for everyone, at all times.

Ubiquitous System: A system from which the personal computer has disappeared and it has been replaced by a multitude of wireless computing devices embodied in everyday objects.

Wearable Computing: Computing devices that have been scaled down for body-wear, being always available in a transparent manner.

WiFi: Short for wireless-fidelity; a logo from the WiFi Alliance that certifies network devices comply with the IEEE 802.11 wireless standards.

Deriving Formal Specifications from Natural Language Requirements

María Virginia Mauco
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

María Carmen Leonardi
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

Daniel Riesco
Universidad Nacional de San Luis, Argentina

INTRODUCTION

Formal methods have come into use for the construction of real systems as they help to increase software quality and reliability, and even though their industrial use is still limited, it has been steadily growing (Bowen & Hinchey, 2006; van Lamsweerde, 2000). When used early in the software development process, they can reveal ambiguities, incompleteness, inconsistencies, errors, or misunderstandings that otherwise might only be discovered during costly testing and debugging phases.

A well-known formal method is the RAISE Method (George et al., 1995), which has been used on real developments (Dang Van, George, Janowski, & Moore, 2002). One tangible product of applying a formal method is a formal specification. A formal specification serves as a contract, a valuable piece of documentation, and a means of communication among stakeholders and software engineers. Formal specifications may be used throughout the software lifecycle and they may be manipulated by automated tools for a wide variety of purposes such as model checking, deductive verification, animation, test data generation, formal reuse of components, and refinement from specification to implementation (van Lamsweerde, 2000). However, one of the problems with formal specifications is that they are hard to master and not easily comprehensible to stakeholders, and even to non-formal specification specialists. This is particularly inconvenient during the first stages of system development when interaction with stakeholders is very important. In practice, the analysis often starts from interviews with the stakeholders, and this source of information is heavily based on natural language as stakeholders must be able to read and understand the results of requirements capture. Then specifications are never formal at first. A good formal approach should use both informal and formal techniques (Bjorner, 2000).

The requirements baseline (Leite, Hadad, Doorn, & Kaplan, 2000), for example, is a technique proposed to formalize requirements elicitation and modeling, which includes two natural language models, the language extended lexicon (LEL) and the scenario model, which ease and encourage stakeholders' active participation. However, specifying requirements in natural language has some drawbacks related to natural language imprecision.

Based on the previous considerations, we proposed a technique to derive an initial formal specification in the RAISE specification language (RSL) from the LEL and the scenario model (Mauco, 2004; Mauco & Riesco, 2005a; Mauco, Riesco, & George, 2004). The technique provides a set of manual heuristics to derive types and functions and structure them in modules taking into account the structured description of requirements provided by the LEL and the scenario model. But, for systems of considerable size this manual derivation is very tedious and time consuming and may be error-prone. Besides, maintenance of consistency between LEL and scenarios, and the RSL specification is a critical problem as well as tracking of traceability relationships.

In this article, we present an enhancement to this technique, which consists in the RSL-based formalization of some of the heuristics to derive RSL types from the LEL. The aim of this formalization is to serve as the basis for a semiautomatic strategy that could be implemented by a tool. More concretely, we describe a set of RSL-based derivation rules that will transform the information contained in the LEL into abstract and concrete RSL types. These derivation rules are a useful starting point to deal with the great amount of requirements information modeled in the LEL, as they provide a systematic and consistent way of defining a tentative set of RSL types. We also present some examples of the application of the rules and discuss advantages and disadvantages of the strategy proposed.

BACKGROUND

In spite of the availability of other notations such a tables, diagrams, and formal notations, natural language is still

chosen for describing the requirements of a software system (Berry, Bucchiarone, Gnesi, Lami, & Trentani, 2006; Bryant et al., 2003; van Lamsweerde, 2000).

The language extended lexicon (LEL) and the scenario model are two well known natural language requirements models used and accepted by the requirements engineering community (Leite et al., 2000). The LEL aims at registering significant terms in the Universe of Discourse (UofD). Its focus is on the application domain language, rather than the details of the problem. It unifies the language allowing the communication with stakeholders. LEL is composed by a set of symbols that represent words or phrases that stakeholders repeat or emphasize. Each entry in the LEL has a name (and possibly a set of synonyms), and two descriptions: notion, which describes what the symbol is, and behavioral response, which describes how the symbol acts upon the system. Each symbol is classified as object, subject, verbal phrase, or state. Figure 1 shows an example of an object LEL symbol taken from the LEL of the milk production system (Mauco, 2004). Underlined words or phrases correspond to other LEL symbols. The scenario model contains a set of scenarios where each scenario describes a situation in the UofD. Scenarios are naturally linked to the LEL. This link is reflected by underlying LEL symbols every time they appear in a scenario description. Figure 2 shows an example of a scenario with all its components.

In order to take profit of natural language requirements specifications, it would be necessary to look at ways for mapping the conceptually richer world of requirements engineering to more formal designs on the way to a complete implementation (Nuseibeh & Easterbrook, 2000). Many works aiming at reducing the gap between the requirements step and the next steps of software development process have been published. Some of them describe, for example, different strategies to obtain object oriented models or formal specifications from requirements specifications (Bryant et al., 2003; Díaz, Pastor, Moreno, & Matteo, 2004; Juristo, Moreno, & Lopez, 2000; Lee & Bryant, 2002).

The RAISE method includes a large number of techniques and strategies for doing formal development and proofs, as well as a formal specification language, RSL, and a set of tools to help writing, checking, printing, storing, transforming, and reasoning about specifications (George, 2001, 2002; George et al., 1995). Usually the first RSL specification is

Figure 1. Field LEL symbol

FIELD
Notion
- Land where cows eat pastures.
- It has an identification.
- It has a precise location in the dairy farm.
- It has a size.
- It has a pasture.
- It has an hectare loading.
- It is divided into a set of plots.
- It has a list of previous plots.

Behavioral Response
- A dairy farmer divides it into a set of plots, separated by electric wires.
- Many different groups of cows can be eating in it simultaneouslv.

Figure 2. Feed a group

TITLE: Feed a group
GOAL: Register the daily ration given to a group of cows.
CONTEXT: It is done once a day. Pre: Group is not empty.
RESOURCES: Group Date Quantity of corn silage Quantity of Hay Quantity of concentrated food Feeding form
ACTORS: Dairy farmer
EPISODES:
- COMPUTE RATION.
- The dairy farmer records, in the Feeding form, the date and the quantities of corn silage, hay and concentrated food given to each cow in the group.
- COMPUTE PASTURE EATEN.

an abstract, applicative, and sequential one, which is later developed into a concrete specification, initially still applicative and then, imperative and sometimes concurrent. A specification in RSL is a collection of modules. A typical applicative module specification contains type, value, and some axiom definitions. Axioms may be used to constrain the values.

When using the RAISE method, writing the initial RSL specification is the most critical task because this specification must capture the requirements in a formal, precise way. But, at the beginning of the software development process, it would be better to use some kind of informal representations to allow stakeholders' participate actively in the requirements definition process. RSL specifications of many domains have been developed by starting from natural language descriptions containing synopsis, narrative, and terminology. The gap between these kind of descriptions and the corresponding RSL formal specification is big, and thus, it is difficult and not always possible to check whether the informal specification models what the informal description does and vice versa.

The requirements definition strategy we present in this article is an attempt to integrate natural language requirements models and formal specifications, which provides a way to fruitfully use all the information available after the requirements definition step. The RSL formal specification derived is the basis to start applying the steps of the RAISE Method, encouraging separate development, and step-wise development.

THE TECHNIQUE TO DERIVE THE RSL SPECIFICATION

As an attempt to reduce the gap between stakeholders and the formal methods world, we proposed a technique to derive an initial formal specification in RSL from natural language requirements models such as LEL and scenarios, which are closer to stakeholders' language (Mauco, 2004). The derivation of the specification is structured in three steps, not strictly sequential:

- **Derivation of types:** Derives a set of abstract as well as concrete types, which model the relevant terms in the domain. We perform the derivation of types in two steps. First we identify the types, and then we decide how to model them.
- **Definition of modules:** Produces a hierarchy of modules, which organizes all the types produced by the derivation of types step in order to obtain a more legible and maintainable specification. This hierarchy of modules can be represented using a layered architecture composed of three layers (specific layer, general layer, and middleware layer), where each layer is a set of RSL modules that share the same degree of generality. These modules would be later completed with the definition of functions, and probably they will be completed with more type definitions.
- **Derivation of functions:** Defines a set of functions that model the functionality in the UofD. Scenarios are the main source of information when defining functions, as they are natural language descriptions of the functionality in the domain. Functions are usually identified at the top level as scenarios help to generate them there. Functions at one level in the hierarchy of modules frequently have counterparts at lower levels, but with different parameters. For each function in the top-level module, we model the necessary functions in lower level modules in order to simplify the legibility and maintainability of the specification. The heuristics we propose help to identify and to model the functions by showing how to derive arguments and result types, how to classify functions as partial or total, and how to define function bodies by analyzing the components of scenarios.

At first, these heuristics have been applied manually. But, for systems of considerable size, this manual derivation is very tedious and time consuming and may be error-prone. Besides, maintenance of consistency between LEL and scenarios, and the RSL specification is a critical problem as well as tracking of traceability relationships. Then, we developed a RSL-based formalization of some of the heuristics to derive RSL types from the LEL (Mauco, Leonardi, Riesco, Montejano, & Debnath, 2005b). The aim of this formalization is to serve as the basis for a strategy that could be implemented by a tool. More concretely, we propose a set of RSL-based derivation rules that will transform the information contained in the LEL into abstract and concrete RSL types.

The RSL Derivation Strategy

In this section, we present a derivation strategy to derive the types of a RSL specification from the information modeled in the LEL. The strategy consists in taking a LEL description and parsing it into RSL abstract syntax (Figure 3), and then applying a set of RSL-based derivation rules to obtain an initial set of RSL types, modeled with a RSL map type expression (Figure 4).

To follow the principles proposed in the RAISE method, we organize the strategy in two steps:

- **Identification of types:** Considering LEL symbols classification, we define five derivation rules, which may produce as result abstract types, subtypes, or collections.
- **Development of types:** In order to remove underspecification, we propose to return to the LEL to

Figure 3. Formal definition of LEL in RSL

```
scheme LEL_DEF =
 class
  type
   LEL = Sym_name →m Name-list × Notion × BhResponse × Classification,
   Notion = Notion_entry-set, BhResponse = BhResponse_entry-set,
   Sym_name = Name,   /* name of a LEL symbol */
   Word == noun | verb | pronoun | preposition | adjective |
    adverb | conjunction | article,      /* any word that is not a LEL symbol name */
   Name = Text,      /* symbol name or symbol name synonym */
   Notion_entry = Nt*, Nt = Sym_name | Word,
   BhResponse_entry = Br-list, Br = Sym_name | Word,
   Classification == subject | obj | verb | state
  axiom
   ∀ s : Sym_name, lel : LEL •        /* sym is the first name in the list */
     s ∈ lel ⇒ let (names, not, bresp, clas) = lel(s) in names ≠ 〈〉 ⇒ s = hd (names) end,
   ∀ lel : LEL, s1, s2 : Sym_name •    /* Notion belongs to only one symbol in the  LEL */
     s1 ∈ lel ∧ s2 ∈ lel ∧ let (ns1, not1, br1, clas1) = lel(s1), (ns2, not2, br2, clas2) = lel(s2)
     in not1 = not2 ⇒ s1 = s2 end,
   ∀ lel : LEL, s1, s2 : Sym_name •   /* BhResponse belongs to only one symbol in  the  LEL */
     s1 ∈ lel ∧ s2 ∈ lel ∧ let (ns1, not1, br1, clas1) = lel(s1), (ns2, not2, br2, clas2) = lel(s2)
     in bresp1 = bresp2 ⇒ s1 = s2 end,
   /* each symbol has at least one entry in the notion and one entry in the behavioral response */
   ∀ s : Sym_name, lel : LEL •
     s ∈ lel ⇒ let (names, not, bresp, clas) = lel(s) in not ≠ {} ∧ bresp ≠ {} end,
    ...
 end
```

Figure 4. Map type expression for derived types

```
type
Type_id = Text,
Type_def = Text,
Types = Type_id →m Type_def
```

analyze the notion of the symbols that motivated the definition of an abstract type. Concretely, the rule defines attributes for abstract types that come from LEL symbols classified as subjects or objects.

It is important to remark that this strategy must be complemented with the participation of software engineers who will adjust and enhance the results obtained after the application of the derivation rules.

The RSL-Based Derivation Rules

This section describes the derivation rules as RSL functions. Each derivation rule specification contains a name, a brief natural language description, and the corresponding RSL specification. We illustrate the application of each rule with examples taken from a milk production system (Mauco, 2004). In some rules, we also mention how the result would have been if we had used the manual heuristics proposed in Mauco (2004).

Rule Name: IT1_abstract_type

a. **Description:** Each LEL symbol classified as subject or object becomes an abstract type.
b. **RSL Formal Specification of the Rule** (shown in Box 1).

Box 1. RSL formal specification of the rule IT1_abstract_type

```
IT1_abstract_type : L.Sym_name × L.LEL × Types →~ Types
IT1_abstract_type(sym, lel, types) ≡ types † [D.get_name(sym) ↦ ""]
pre sym ∈ lel ∧ let (names, not, bresp, clas) = lel(sym)
    in clas = L.subject ∨ clas = L.obj end
```

c. **Application of the rule to the case study:** Figure 1 shows the LEL symbol field, which was classified as object. By applying IT1_abstract_type rule we obtain the abstract type field.

Rule Name: IT2_abstract_type

a. **Description:** Each LEL symbol classified as a verbal phrase describing a "registration" behavior becomes an abstract type.
b. **RSL formal specification of the rule:** (shown in Box 2)
c. **Application of the rule to the case study:** Considering LEL symbol Assigns cow to a group, classified as verbal phrase, the rule finds its notion or behavioral response contains verbs describing a registration behavior. So, it is modeled as the abstract type Assigns_cow_to_a_group. Verbal phrases, which do not represent registration behavior, are considered later, when defining RSL functions.

Rule Name: IT3_subtype

a. **Description:** Each LEL symbol that references in its name another LEL symbol, previously defined as a type T, becomes a subtype of T.
b. **RSL formal specification of the rule:** (shown in Box 3
c. **Application of the rule to the case study:** Analyzing the name of the LEL symbol Dairy cow the rule finds another LEL symbol, Cow, previously modeled as an abstract type. Then, the following subtype expression is defined:

    - type
      Cow, /* already defined*/
      Dairy_cow = {|dc: Cow :- is_dairy_cow(dc)|}

Rule Name: IT4_subtype

a. **Description:** Each LEL symbol classified as state becomes a subtype if it references another LEL symbol in its name, or an abstract type if it does not.
b. **RSL formal specification of the rule:** (shown in Box 4)
c. **Application of the rule to the case study:** Our case study contains only three LEL symbols classified as states: Lactation, Pregnant, and On Heat. As none of them mentions in its name another LEL symbol, the application of rule IT4_subtype gives as result one abstract type for each.

Rule Name: IT5_collection

a. **Description:** Each LEL symbol classified as subject or object and whose name is in plural becomes a list.
b. **RSL formal specification of the rule:** (shown in Box 5)
c. **Application of the rule to the case study:** This rule cannot be applied in our case study because the LEL does not contain symbols with plural names.

It is common practice not to include in the LEL symbols defining a collection of another LEL symbol when the actions that could be applied to the collection are the classical ones such as adding, removing, or recovering elements. When applying the manual heuristics, subjects and objects whose names were in singular were analyzed to find out if they could have more than one instance in order to model the corresponding collection. The formalization of the heuristic as a fixed rule excludes this kind of analysis and then valuable information is missing. Thus, human intervention is unavoidably in order to correct these problems and enhance the final result.

Box 2. RSL formal specification of the rule IT2_abstract_type

```
IT2_abstract_type : L.Sym_name × L.LEL × Types →~ Types
IT2_abstract_type(sym, lel, types) ≡ types † [D.get_name(sym) ↦ ""]
pre sym ∈ lel ∧ let (names, not, bresp, clas) = lel(sym)
      in clas = L.verb end ∧ has_data_save(sym, lel)
```

Box 3. RSL formal specification of the rule IT3_subtype

```
IT3_subtype : L.Sym_name × L.LEL × Types →~ Types
IT3_subtype(sym, lel, types) ≡ types † [D.get_name(sym) ↦
      "{|s:" ^ main_type(D.get_name(sym), lel, types) ^ ":-is_" ^D.get_name(sym) ^ "(s)|}"]
pre sym ∈ lel ∧ let (names, not, bresp, clas) = lel(sym)
     in (clas = L.subject ∨ clas = L.obj) end ∧ (s : Text :- mention(sym, s, lel) ∧ s ∈ types)
```

Box 4. RSL formal specification of the rule IT4_subtype

```
IT4_subtype : L.Sym_name × L.LEL × Types →~ Types
IT4_subtype(sym, lel, types) ≡ if (s: Text :- mention(sym, s, lel) ∧ s ∈ types)
                    then types † [D.get_name(sym) ↦ "{|s:" ^ main_type(sym, lel, types)
                          ^ ":-is_" ^D.get_name(sym) ^ "(s)|}"]
                    else types † [D.get_name(sym) ↦ ""] end
pre sym ∈ lel ∧ let (names, not, bresp, clas) = lel(sym) in clas = L.state end
```

Box 5. RSL formal specification of the rule IT5_collection

```
IT5_collection : L.Sym_name × L.LEL × Types →~ Types
IT5_collection(sym, lel, types) ≡ types † [D.get_name(sym) ↦ D.singular(sym) ^ "*"]
pre sym ∈ lel ∧ let (names, not, bresp, clas) = lel(sym) in clas = L.subject ∨ clas = L.obj end
∧ D.is_plural(sym) ∧ (∃! tid : Type_id :- tid ∈ types ∧ tid = D.singular(sym))
```

Rule Name: DT1_comp_type

a. **Description:** Defines a record type expression for abstract types coming from LEL symbols classified as subjects or objects. Record components are nouns or LEL symbols mentioned in an entry of the notion of these kind of LEL symbols.
b. **RSL formal specification of the rule:** (shown in Box 6)
c. **Application of the rule to the case study:** This rule is the one that deals with the development of abstract types. For example, for the abstract type Field defined by rule IT1_abstract_type, rule DT1_comp_type completes its definition by defining the following record type expression:

```
type
  Field::
     pasture
     identification
     dairy_farm
     size
     hectare_loading
     set
     plots
     list
```

One of the main problems of this rule is that it misses noun groups. As the rule defines every noun as a potential record component, it generates more and sometimes inappropriate components. However, noun groups detection may be included following linguistic approaches (Díaz et al., 2004; Juristo et al., 2000).

FUTURE TRENDS

Concerning future work, we are analyzing the formalization of a set of derivation rules to define RSL functions and modules based on the manual heuristics presented in Mauco (2004). To define these kind of rules, we would have to formalize the scenario model in RSL. Besides, it would be necessary to incorporate linguistic approaches (Díaz et al., 2004; Juristo et al., 2000) to achieve a better processing of the natural language-based requirements models.

Traceability plays a crucial role in any software development process. The derivation rules we have proposed provide a way to track "traced-to" and "traced-from" relationships, and they also help to cope with the problem of maintaining consistency between LEL and scenarios, and the RSL specification. However, a more detailed and deeper analysis should be made.

It would also be interesting to have a tool to assist in the derivation process. This tool could be later integrated with the RAISE tools in order to have assistance in the RSL specification complete development process. Tool support for formal methods will be critical in the next 10 years (Bowen et al., 2006).

Natural language-oriented models are still widely used in requirements modeling due to their well-known advantages. This kind of requirements models have to be reinterpreted by software engineers into a more formal design on the way to a complete implementation. So, bridging the gap between requirements specifications and formal specifications continues to be one of the challenges for the future. In particular, in the context of the RAISE method, our proposal of a semiautomatic transformation to map LEL and scenarios knowledge into a RSL specifications is a first step into this direction.

CONCLUSION

To contribute to bridge the gap between requirements engineering and formal methods, we have presented a technique to be used in the first stages of development using the RAISE method. We proposed and defined a three-step process that could be applied to any domain to derive an initial formal

Box 6. RSL formal specification of the rule DT1_comp_type

```
DT1_comp_type : L.Sym_name × L.LEL × Types →~ Types
DT1_comp_type(sym, lel, types) ≡ let (names, not, bresp, clas) = lel(sym) in
          types † [D.get_name(sym) ↦ def_record(not, lel, types)] end
pre sym ∈ lel ∧ let (names, not, bresp, clas) = lel(sym)
                in clas = L.subject ∨ clas = L.obj end ∧ D.get_name(sym) ∈ types,

/*def_record defines the record type expression corresponding to a LEL symbol notion*/
```

specification in RSL from LEL and scenarios, two natural language models belonging to the requirements baseline. In this way, we could take profit of informal descriptions reducing the gap between them and the final RSL specification. Moreover, by using the three-step process we proposed, the effort to define complete requirements models is worth doing because, though partially, they could be later mapped onto a RSL formal specification.

In addition, we have described a proposal to define, in a semiautomatic way, an initial set of RSL types starting from the LEL. The RSL-based derivation rules we proposed may be implemented by a tool, thus allowing one to automatically obtain a first initial RSL specification currently obtained in a manual way. Derivation rules are a concrete automation of some of the manual heuristics of our three-step process, and then they involve fix decisions about certain modeling issues. They are a useful starting point to deal with the great amount of requirements information modeled in the LEL, as they provide a systematic and consistent way of defining a tentative set of RSL types. Though a manual derivation produces a better and more accurate specification definition as it allows one to cope with the power of expression of natural language, the specification obtained by applying the derivation rules could be later refined by a human, who will correct and complete it.

The technique we have developed gives, as a result, a set of modules hierarchically structured, aiming at increasing the maintainability and legibility of the specification. The hierarchy of RSL modules obtained can be mapped onto a layered architecture, which is the basis to start applying the steps of the RAISE method and provides the specific properties all its developments should have. The use of a layered architecture is particularly useful when designing complex systems because it facilitates and encourages not only reuse but also separate and step-wise development. Thus, the RSL initial specification derived could be later developed into a concrete one according to the steps provided by the RAISE method. With a concrete specification the SML translator (George, 2001) could be used in order to have a quick prototype and get a feeling of what the specification really does.

REFERENCES

Berry, B., Bucchiarone, A., Gnesi, S., Lami, G., & Trentani, G. (2006). A new quality model for natural languages requirements specifications. In *Proceedings of the 12th International Workshop on Requirements Engineering: Foundation for Software Quality*, Luxembourg.

Bjorner, D. (2000). *Software engineeering: A new approach.* From domains via requirements to software. Formal specification and design calculi. Dept. of Informatics and Mathematical Modelling, Technical University of Denmark. Retrieved August 2007, from http://www.it.dtu.dk/~db/s2000/notes.ps

Bowen, J., & Hinchey, M. (2006). Ten commandments of formal methods … Ten years later. *IEEE Computer*, 40-48.

Bryant, B. R., Lee, B., Cao, F., Zhao, W., Burt, C., Raje, R., Olson, A., & Auguston, M. (2003). From natural language requirements to executable models of software components. In *Proceedings of Monterey Workshop on Software Engineering for Embedded Systems: From Requirements to Implementation* (pp. 51-58).

Dang Van, H., George, C., Janowski, T., & Moore, R. (2002). *Specification case studies in RAISE*. Springer-Verlag.

Díaz, I., Pastor, O., Moreno, L., & Matteo, A. (2004). Una Aproximación Linguística de Ingeniería de Requisitos para OO-Method. In *Proceedings of IDEAS´2004: Workshop Iberoamericano de Ingeniería de Requisitos y Desarrollo de Ambientes de Software* (pp. 270-281). Perú.

George, C. (2002). *Introduction to RAISE*. UNU/IIST, Macau, Technical Report 249.

George, C. (2001). *RAISE tools user guide*. UNU/IIST, Macau, Research Report 227.

George, C., Haxthausen, A., Hughes, S., Milne, R., Prehn, S., & Pedersen, J. S. (1995). *The RAISE development method*. BCS Practitioner Series, Prentice Hall.

Juristo, N., Moreno, A., & Lopez, M. (2000). How to use linguistic instruments for object-oriented analysis. *IEEE Software* (pp. 80-89), May-June.

Lee, B., & Bryant, B. (2002). Automated conversion from requirements documentation to an object-oriented formal specification language. In *Proceedings of the 2002 ACM Symposium on Applied Computing* (pp. 932-936).

Leite, J. C. S. P, Hadad, G., Doorn, J., & Kaplan, G. (2000). A scenario construction process. *Requirements Engineering Journal, 5*(1), 38-61, Springer-Verlag.

Mauco, M. V. (2004). *A technique for an initial specification in RSL*. Master thesis. Facultad de Informática, Universidad Nacional de La Plata, Argentina.

Mauco, M. V., & Riesco, D. (2005a). Integrating requirements engineering techniques and formal methods. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 1555-1559). Hershey, PA: IRM Press.

Mauco, M. V., Leonardi, M. C., Riesco, D., Montejano, G., & Debnath, N. (2005b). Formalising a derivation strategy for formal specifications from natural language requirements models. In *Proceedings of the 5th IEEE International Sym-*

posium on Signal Processing and Information Technology (ISSPIT 2005) (pp. 646-651). Greece.

Mauco, M. V., Riesco, D., & George, C. (2004). Deriving an initial specification in RSL from natural language models. In *Proceedings of the 1st International Conference on the Principles of Software Engineering* (pp. 111-120). Argentina.

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: A roadmap. In *Proceedings of the Conference on the Future of Software Engineering, ACM* (pp. 35-46).

van Lamsweerde, A. (2000). Formal specification: A roadmap. In *Proceedings of the Conference on the Future of Software Engineering, ACM* (pp. 147-159).

KEY TERMS

Formal Methods: This term refers to the variety of mathematical modeling techniques that are applicable to computer system (software and hardware) design. Formal methods may be used to specify and model the behavior of a system and to mathematically verify that the system design and implementation satisfy system functional and safety properties. These specifications, models, and verifications may be done using a variety of techniques and with various degrees of rigor.

Formal Specification: It is the expression, in some formal language and at a some level of abstraction, of a collection of properties some system should satisfy. A specification is formal if it is expressed in a language made of three components: the syntax (rules for determining the grammatical well-formedness of sentences), the semantics (rules for interpreting sentences in a precise, meaningful way in the domain considered), and the proof theory (rules for inferring useful information from the specification.

RAISE Method: The RAISE method encompasses formulating abstract specifications, developing them to successively more concrete specifications, justifying the correctness of the development, and translating the final specification into a programming language, and it is based on a number of principles such as separate development, step-wise development, invent and verify, and rigor. RAISE is an acronym for "rigorous approach to industrial software engineering," and it gives its name to a formal specification language, the RAISE specification language, the associated method, and a set of tools.

RAISE Specification Language (RSL): It is a formal specification language intended to support the precise definition of software requirements and reliable development from such definitions to executable implementations. It supports specification and design of large systems in a modular way, and thus it permits separate subsystems to be separately developed. It also provides a range of specification styles (axiomatic and model-based; applicative and imperative; sequential and concurrent) as well as it supports specifications ranging from abstract (close to requirements) to concrete (close to implementations).

Requirements Baseline: It is a mechanism proposed to formalize requirements elicitation and modeling. It is a structure which incorporates descriptions about a desired system in a given application domain. Although it is developed during the requirements engineering process, it continues to evolve during the software development process. It is composed of five complementary views: the lexicon model view, the scenario view, the basic model view, the hypertext view, and the configuration view.

Requirements Engineering: It comprehends all the activities involved in eliciting, modeling, documenting, and maintaining a set of requirements for a computer-based system. The term "engineering" implies that systematic and repeatable techniques should be used to ensure that system requirements are consistent, complete, relevant, etc.

Requirements: They are descriptions of how the system should behave, or of a system property or attribute. They are defined during the early stages of a system development as a specification of what should be implemented. They should be statements of what a system should do rather than a statement of how it should do it.

Universe of Discourse (UofD): It is the overall context in which the software will be developed and operated.

Design and Applications of Digital Filters

Gordana Jovanovic Dolecek
INSTITUTE INAOE, Puebla, Mexico

INTRODUCTION

Digital signal processing (DSP) is an area of engineering that "has seen explosive growth during the past three decades" (Mitra, 2005). Its rapid development is a result of significant advances in digital computer technology and integrated circuit fabrication (Jovanovic Dolecek, 2002; Smith, 2002). Diniz, da Silva, and Netto (2002) state that "the main advantages of digital systems relative to analog systems are high reliability, suitability for modifying the system's characteristics, and low cost".

The main DSP operation is digital signal filtering, that is, the change of the characteristics of an input digital signal into an output digital signal with more desirable properties. The systems that perform this task are called digital filters. The applications of digital filters include the removal of the noise or interference, passing of certain frequency components and rejection of others, shaping of the signal spectrum, and so forth (Ifeachor & Jervis, 2001; Lyons, 2004; White, 2000).

Digital filters are divided into finite impulse response (FIR) and infinite impulse response (IIR) filters. FIR digital filters are often preferred over IIR filters because of their attractive properties, such as linear phase, stability, and the absence of the limit cycle (Diniz, da Silva & Netto, 2002; Mitra, 2005). The main disadvantage of FIR filters is that they involve a higher degree of computational complexity compared to IIR filters with equivalent magnitude response (Mitra, 2005; Stein, 2000).

For example let us consider an FIR filter of length N = 11 with impulse response

$$h(n) = \begin{cases} 0.8^n & for \quad 0 \le n \le 10 \\ 0 & otherwise \end{cases}, \tag{1}$$

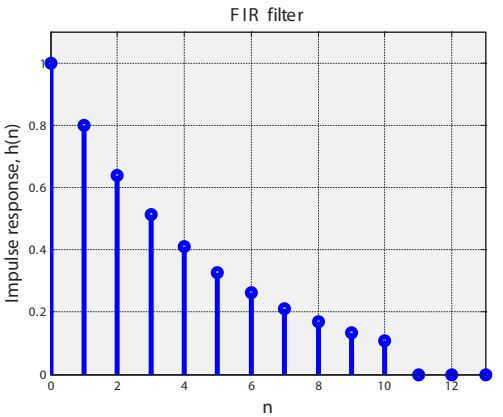
as shown in Figure 1a.

In Figure 1b the initial 20 samples of the impulse response of an IIR filter

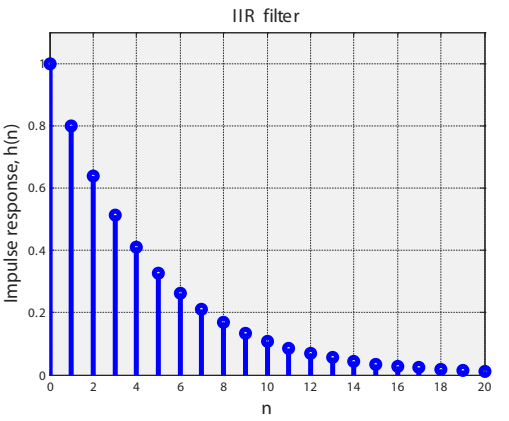
$$h(n) = \begin{cases} 0.8^n & for \quad 0 \le n \\ 0 & for \quad n < 0 \end{cases}. \tag{2}$$

are plotted.

Figure 1. Impulse responses of FIR and IIR filters



a. FIR filter



b. IIR filter

Equation 3.

$$y(n) = x(n) * h(n) = h(n) * x(n) = \sum_k h(k)x(n-k) = \sum_k x(k)h(n-k)$$

BACKGROUND

Digital Filters in Time and Transform Domain

The operation in time domain which relates the input signal $x(n)$, impulse response $h(n)$ and the output signal $y(n)$, is called the *convolution*, and is defined in Equation 3.

The output $y(n)$ can also be computed recursively using the following *difference equation* (Mitra 2005; Proakis & Ingle, 2003),

$$y(n) = \sum_{k=0}^{M} b_k x(n-k) + \sum_{k=1}^{N} a_k y(n-k), \quad (4)$$

where $x(n\text{-}k)$ and $y(n\text{-}k)$ are input and output sequences $x(n)$ and $y(n)$ delayed by k samples, and b_k and a_k, are constants. The order of the filter is given by the maximum value of N and M. The first sum is a *nonrecursive*, while the second sum is a *recursive* part. Typically, FIR filters have only nonrecursive part, while IIR filters always have the recursive part. As a consequence, FIR and IIR filters are also known as nonrecursive and recursive filters, respectively.

From (3) we see that the principal operations in a digital filter are multiplications, delays and additions. Using equation (3) we can draw the structure of the digital filter which is also known as a *Direct form* and is shown in Figure 2. More details about filter structures can be found for example in Mitra (2005).

The representation of digital filters in the transform domain is obtained using the *Fourier transform* and *z-transform*.

The Fourier transform of the signal $x(n)$ is defined as

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{j\omega n}, \quad (5)$$

where ω is digital frequency in radians and $e^{j\omega n}$ is an exponential sequence. In general case, the Fourier transform is a complex quantity.

The convolution operation becomes multiplication in the frequency domain,

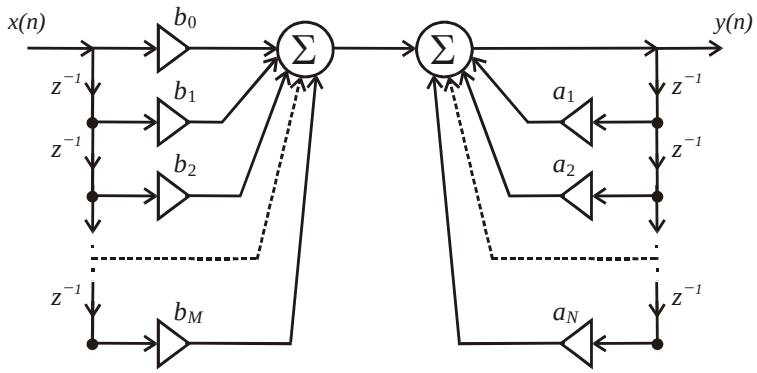
$$Y(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega}), \quad (6)$$

where $Y(e^{j\omega})$, $X(e^{j\omega})$, and $H(e^{j\omega})$, are Fourier transforms of $y(n)$, $x(n)$ and $h(n)$, respectively. The quantity $H(e^{j\omega})$ is called the *frequency response* of the digital filter, and it is a complex function of the frequency ω with a period 2π. It can be expressed in terms of its real and imaginary parts, $H_R(e^{j\omega})$ and $H_I(e^{j\omega})$ or in terms of its magnitude $|H(e^{j\omega})|$ and phase $\varphi(\omega)$,

$$H(e^{j\omega}) = H_R(e^{j\omega}) + jH_I(e^{j\omega}) = |H(e^{j\omega})|e^{j\phi(\omega)}. \quad (7)$$

The amplitude $|H(e^{j\omega})|$ is called the *magnitude response* and the phase $\varphi(\omega)$ is called the *phase response* of the digital filter. For a real impulse response digital filter, the magnitude response is an even function of ω, while the phase

Figure 2. Direct form structure



response is a real odd function of ω. In some applications, the magnitude response is expressed in the logarithmic form in decibels as

$$G(\omega) = 20\log_{10}\left|H(e^{j\omega})\right| \quad \text{dB} \quad , \tag{8}$$

where $G(\omega)$ is called the *Gain function*.

For the sequence $x(n)$, z-transform is defined as

$$X(z) = \sum_{n=\infty}^{\infty} x(n)z^{-n} \, , \tag{9}$$

where z is a complex variable. All values of z for which (9) converges are called the *region of convergence* (ROC).

Z-transform of the unit sample response $h(n)$, denoted as $H(z)$, is called *system function*. Using z-transform of the Equation (4) we arrive at

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}}. \tag{10}$$

The roots of the numerator, or the values of z for which $H(z)=0$, define the locations of the *zeros* in the complex z plane. Similarly, the roots of the denominator, or the values of z for which $H(z)$ become infinite, define the locations of the *poles*. Both poles and zeros are called *singularities*. The plot of the singularities in z-plane is called the *pole-zero pattern*. The zero is usually denoted by a circle (o) and the pole by a cross (x). An FIR filter has only zeros (poles are in the origin), whereas an IIR filter can have either both zeros and poles, or only poles, (zeros are in the origin). All poles of the linear-phase filter are in the origin and zeros are in ether in symmetrical positions in respect of the unit circle or on the unit circle. Its unit sample response has symmetry (Mitra, 2005; Proakis & Ingle, 2006). If the FIR filter does not have the linear phase (the unit sample response does not have symmetry) its zeros are not in symmetrical positions around the unit circle. This filter has all zeros inside the unit circle and is called a minimum-phase filter (Mitra, 2005). IIR filter which passes all frequencies without attenuation is called allpass filter. The zeros and poles of this filter are in a symmetrical positions relating to the unit circle. The filter is stable if all poles are inside the *unit circle* in z-plane. (FIR filters with the poles in origin, and IIR filters with the poles inside the unit circle.) More details about characteristics and applications of different FIR and IIR filters can be found in (Kuo, 2006; Lyons, 2004; Mitra 2005; Proakis & Ingle, 2003; Stearns, 2002; Smith, 2002, Weeks, 2006)

Transform of LP into HP Filter

Instead of designing a high-pass filter by brute force, we can transform it into a low-pass filter. We replace the desired cutoff frequencies of the high-pass filter ω_p and ω_s, by the corresponding low-pass specifications as follows:

$$\omega'_p = \pi - \omega_p$$
$$\omega'_s = \pi - \omega_s \, . \tag{11}$$

Given these specifications, a low-pass FIR filter can be designed. From this auxiliary low-pass filter, the desired high-pass filter can be computed by simply changing the sign of every other impulse response coefficient. This is compactly described as,

$$h_{HP}(n) = (-1)^n h_{LP}(n), \tag{12}$$

where $h_{HP}(n)$ and $h_{LP}(n)$ are the impulse responses of the high-pass and the low-pass filters, respectively.

Examples of Filtering

Example 1

In this example we consider a signal composed of two cosine signals $x_1(n)$ and $x_2(n)$, shown in Figures 3a, and 3b.

$$x_1(n) = \cos(0.2\pi n), x_2(n) = \cos(0.6\pi n).$$

$x(n)=x_1(n)+x_2(n)$

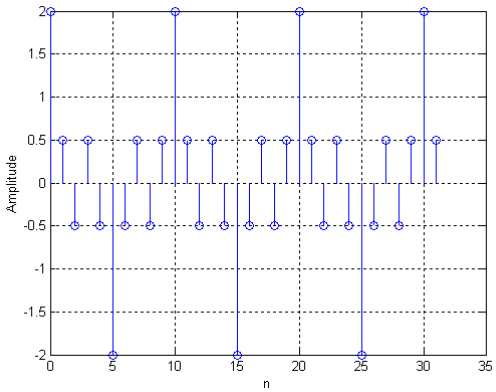
Two peaks at 0.2π and $0.6\,\pi$ in the spectral characteristic correspond to the cosine components x_1 and x_2, respectively.

Suppose we now apply low-pass (LP) filtering to the sum of these two cosine signals. The result of filtering is shown in Figures 3c and d. Notice that the second high pass cosine signal has been eliminated. To eliminate the low-pass cosine signal, we design the high-pass (HP) filter. The filtered signal is shown in Figures 3e and f.

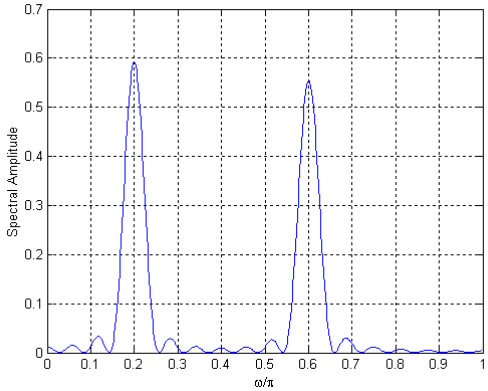
Example 2

The following figure presents an example of a speech signal (McClellan, Schafer & Yoder, 1998). We consider one part of the signal (the samples from 1,300 to 1,500), which is shown in Figures 4b and c. The low-pass filter which passes all spectral components below $0.25\,\pi$ and eliminates all spectral components higher than 0.3π (Figure 4c). The speech signal filtered by the filter is shown in Figure 4d. Notice that the resulting signal becomes smoother when higher frequencies

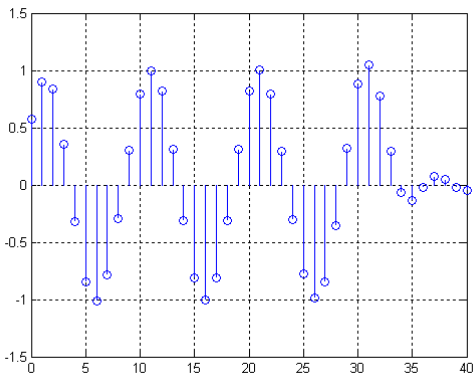
Figure 3. Sum of two cosine signals



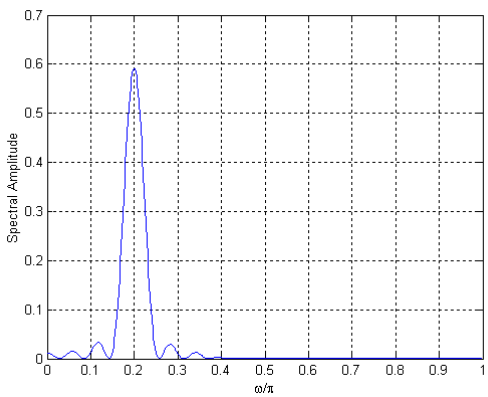
a. Time-domain



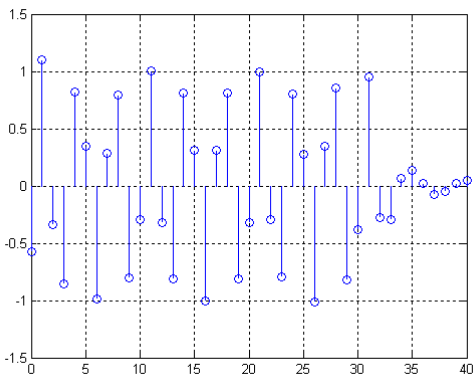
b. Frequency domain



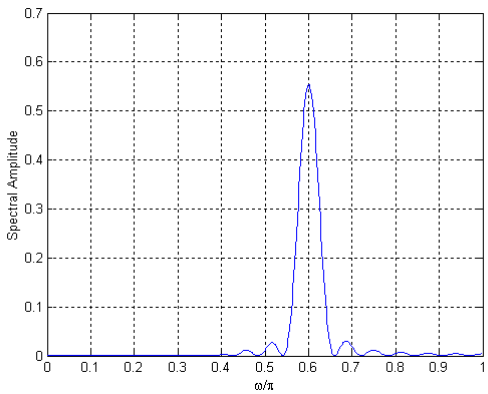
c. LP Time-domain signal



d. LP Frequency domain

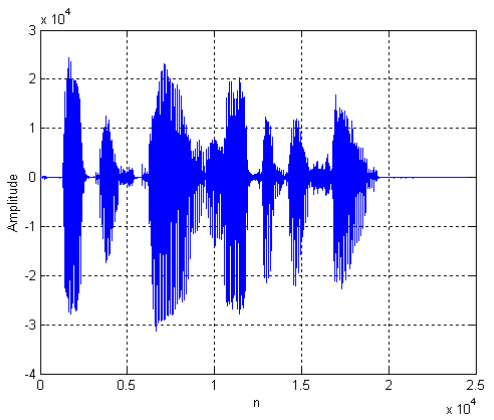


e. HP Time domain

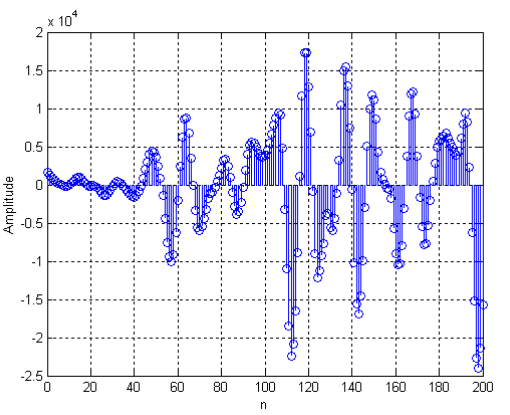


f. HP Frequency domain

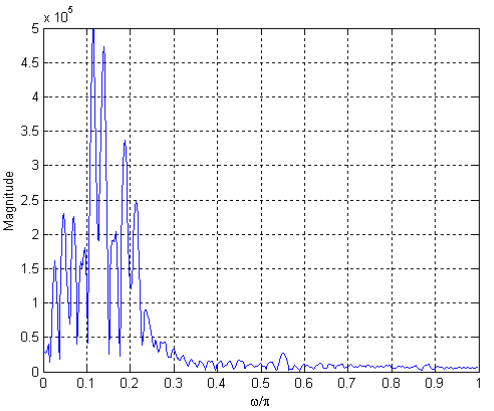
Figure 4. Sampled speech waveform



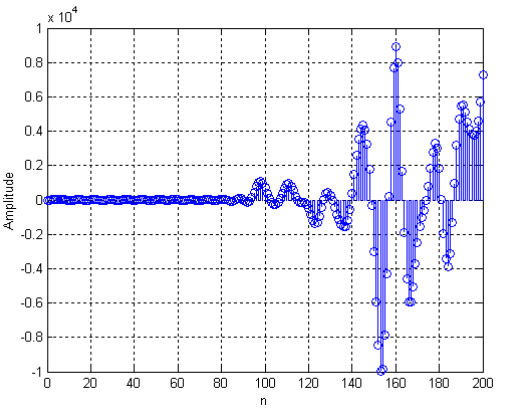
a. Speech signal



b. Part of the speech signal



c. Frequency domain



d. LP filtering

are eliminated. Therefore, low-pass filtering can be used to remove large fluctuations in the signal.

More details for digital filters for audio signal processing can be found in Meana (2007); Huang and Benesty (2004); Spanias, Painter, and Atti (2007).

Example 3

In this example we illustrate the effect of filtering of an image, generated in MATLAB, shown in Figure 5a. The noise is added to the image and the result is shown in Figure 5b. Two filters are applied to eliminate the noise. Figure 5c shows the result of applying a simple averaging filter, while Figure 5d shows the effect of applying a special filter called the median filter. Notice that the median filter is much better in removing noise.

More details about image signal processing can be found in Bose and Meyer (2006); Barnet (2007); Woods (2006); Bovik (2000).

FUTURE TRENDS

For years much effort has been made to reduce the complexities of the FIR filters (Chan, Tsui & Zhao, 2006; Izydorczyk, 2006; Lin, Chen & Jou, 2006; Macleod & Dempster, 2005; Maskell, Jussipekka & Patra, 2006; Xu, Chang & Jong, 2006). This field of research "continues to be in full of vigor as new design problems arise and innovative design techniques emerge" (Lu, 2006).

Another two direction of the research include IIR designs that are nonlinear and nonconvex and FIR filter design with certain structures that lead to nonconvex second-order or higher order design (Lu, 2006).

Another important trend is in design of variable digital filters which can be designed using either FIR or IIR filters (Yli-Kaakinen & Saramaki, 2006).

Figure 5. Removing the noise from the image

a. Image signal

b. The image with added noise

c. Filtering with averaging filter

d. Filtering with median filter

CONCLUSION

The digital filter changes the characteristics of the input digital signal in order to obtain the desired output signal. Digital filters either have a finite impulse response, (FIR), or an infinite impulse response, (IIR). FIR filters are often preferred because of desired characteristics, such as linear phase and no stability problems. The main disadvantage of FIR filters is that they involve a higher degree of computational complexity compared to IIR filters with equivalent magnitude response. In many applications where the linearity of the phase is not required, the IIR filters are preferable because of the lower computational requirements.

REFERENCES

Abeysekera, S. S. & Padhi, K. P. (2000). Design of multiplier free FIR filters using a LADF sigma-delta (Sigma-Delta) modulator. *2000 IEEE International Symposium on Circuits and Systems. Emerging Technologies for the 21st Century*. Proceedings IEEE, 2, 65-68.

Barner, K. E. (2007). *Nonlinear signal and image processing: Theory, methods, and applications*. Taylor & Francis.

Bose, T. & Meyer, F. G. (2003). *Digital signal and image processing.* New York: John Wiley & Sons.

Bovik, A. C. (2000). *Handbook of image and video processing*. Academic Press.

Chan, S. C., Tsui, K. M., & Zhaoo, S. H. (2006). A methodology for automatic syntheses of multiplier-less digital filters with prescribed output accuracy. In *Proceedings of the IEEE Conference, APCCAS* (pp. 61-64).

Coleman, J. O. (2002). Cascaded coefficient number systems lead to FIR filters of striking computational efficiency. In *Proceedings of the 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* (pp. 513-516). Piscataway, NJ.

Diniz, P. S. R., da Silva, E. A .B., & Netto, S. L. (2002). *Digital signal processing: System analysis and design*. Cambridge: Cambridge University Press.

Huang, Y. & Benestry, J. (Eds.) (2004). *Audio signal processing for next-generation multimedia communication systems.* Springer.

Ifeachor, E. C. & Jervis, B. E. (2001). *Digital signal processing: A practical approach* (2nd ed.). NJ: Prentice Hall.

Izydorczyk, J. (2006). An algorithm for optimal terms allocation for fixed point coefficients of FIR filters. In *Proceedings*

of the IEEE Conference on Circuits and Systems, ISCAS 2006 (pp. 609-611).

Jovanovic-Dolecek, G. (Ed.) (2002). *Multirate systems: Design and applications*. Hershey, PA: Idea Group Publishing.

Kuo, S. M. (2006). *Real-time digital signal processing: Implementations and applications*. Wiley.

Kuo, C. J., Chien, H. C., & Lin, W. H. (2000). Neighboring full-search algorithm for multiplierless FIR filter design. IEICE transactions on fundamentals of electronics communications & computer sciences. *Inst. Electron. Inf & Commun, 11*, 2379-2381.

Lin, M. C., Chen, H. Y., & Jou, S. J. (2006). Design techniques for high-speed multirate multistage FIR digital filters. *International Journal of Electronics, 93*(10), 699-721.

Lu, W. S. (2006). Digital filter design: Global solutions via polynomial optimization. In *Proceedings of the IEEE Conference, APCCAS* (pp. 49-52).

Lyons, R. G. (2004). *Understanding digital signal processing* (2nd ed.). Prentice Hall PTR.

Macleod, M. D. & Dempster, A. G. (2005). Multiplierless FIR filter design algorithms. *IEEE Signal Processing Letters, 12*(3), 186-189.

Meana, H. P. (Ed.) (2007). *Advances in audio and speech signal processing: Technologies and applications*. Hershey, PA: IGI Global.

Mitra, S. K. (2005). *Digital signal processing: A computer-based approach*. New York: McGraw-Hill, Inc.

Proakis, J. G. & Ingle, V. K. (2003). *A self-study guide for digital signal processing*. Prentice Hall.

Proakis, J. G. & Manolakis, D. K. (2006). *Digital signal processing* (4th ed.). Prentice Hall.

Smith, S. (2002). *Digital signal processing: A practical guide for engineers and scientists*. New York: Newnes.

Spanias, A., Painter, T., & Atti, V. (2007). *Audio signal processing and coding.* Wiley-Interscience.

Stein, J. (2000). *Digital signal processing: A computer science perspective*. New York: Wiley- Interscience.

Weeks, M. (2006). *Digital signal processing using MATLAB and wavelets*. Infinity Science Press.

White, S. (2000). *Digital signal processing: A filtering approach*. Delmar Learning.

Woods, J. W. (2006). *Multidimensional signal, image, and video processing and coding.* Academic Press.

Xu, F., Chang, C., & Jong, C. C. (2006). A new integrated approach to the design of low-complexity FIR filters. In *Proceedings of the IEEE Conference on Circuits and Systems, ISCAS 2006* (pp. 601- 604).

Yli-Kaakinen & Saramaki, T. (2001). A systematic algorithm for the design of multiplierless FIR filters. In *Proceedings of the 2001 IEEE International Symposium on Circuits and Systems* (pp. 185-188). Piscataway, NJ.

Yli-Kaakinen, & Saramaki, T. (2006). Approximately linear-phase recursive digital filters with variable magnitude characteristics. In *Proceedings of the 2006 IEEE International Symposium on Circuits and Systems* (pp. 5227-5230).

KEY TERMS

Convolution y(n)=x(n)*h(n): Time domain operation which relate the output of the digital filter y(n) with the input signal x(n) and the impulse response of the filter h(n).

Cutoff Frequencies: The frequencies which determine the passband (the frequencies which are passed without attenuation), and the stop-band (the frequencies which are highly attenuated).

Difference Equation: Time domain relation between the output and the input of digital filter in terms of coefficients which are characteristics of the filter. Generally contains recursive and nonrecursive parts.

Frequency Response $H(e^{j\omega})$**:** The discrete-time Fourier transform of the impulse response of the system is called the Frequency response. It provides a frequency-domain description of the system. In general, it has a complex value.

High-Pass Digital Filter: Digital filter which passes only high frequencies defined by the passband cutoff frequency and attenuates all frequencies from 0 to cutoff stopband frequency.

Impulse Response $h(n)$: The response of a digital filter to a unit sample sequence, which consists of a single sample at index $n = 0$ with unit amplitude.

Low-Pass Digital Filter: Digital filter which passes only low frequencies defined by the passband cutoff frequency and attenuates all high frequencies from the cutoff stopband frequency to π.

Magnitude Response $\left|H(e^{j\omega})\right|$: Absolute value of the complex frequency response.

Phase Response: Phase of the complex frequency response.

Singularities: Poles and zeros of system function. Poles of system function are zeros of its denominator while zeros are zeros of its nominator.

System Function: Z-transform of the impulse response of the filter. FIR filters has only the nominator, while an IIR filter has denominator or both nominator and denominator.

Design and Development of Communities of Web Services

Zakaria Maamar
Zayed University, UAE

INTRODUCTION

In the field of Web services (Benatallah, Sheng, & Dumas, 2003; Bentahar, Maamar, Benslimane, & Thiran, 2007; Medjahed & Bouguettaya, 2005), a community gathers Web services that offer similar functionalities. Hotel booking and car rental are samples of functionalities. This gathering takes place regardless of who developed the Web services, where the Web services are located, and how the Web services function to satisfy their functionalities. A Web service is an accessible application that can be discovered according to its functionality and then invoked in order to satisfy users' needs. In addition, Web services can be composed in a way that permits modeling and executing complex business processes. Composition is one of Web services' strengths as it targets user needs that cannot be satisfied by any single available Web service. A composite Web service obtained by combining available Web services may be used (Figure 1). The use of communities in composition scenarios offers two immediate benefits. The first benefit is the possibility of accelerating the search of Web services required to satisfy user needs by looking for communities rather than screening UDDI (universal description, discovery, and integration) and ebXML registries. The second benefit is the late execution binding of the required Web services once the appropriate communities are identified. Both benefits stress the need of examining Web services in a different way.

Current practices in the field of Web services assume that a community is static and Web services in a community always exhibit a cooperative attitude. These practices need to be revisited as per the following arguments. A community is dynamic: New Web services enter, other Web services leave, some Web services become temporarily unavailable, and some Web services resume operation after suspension. All these events need to be closely monitored so that inconsistent situations are avoided. Moreover, Web services in a community can compete on nonshareable computing resources, which may delay their performance scheduling. Web services can also announce misleading information (e.g., nonfunctional details) in order to boost their participation opportunities in composition scenarios. Finally, Web services can be malicious in that they can try to alter other Web services' data or operations.

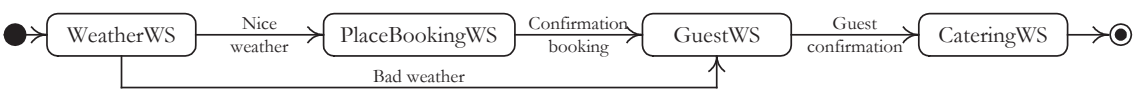
To look into ways of making Web services communities active, we describe in this article some mechanisms that would enable Web services among other things to enter a community, to leave a community after awhile, to reenter the same community if some opportunities loom, and to be rewarded for being part of a community. These mechanisms would be developed along three perspectives, which we refer to as the following.

- Community management: How do we establish or dismantle a new or existing community of Web services?
- Web services attraction and retention: How do we invite and convince new Web services to join a community? How do we retain existing Web services in a community?
- Interaction management: How are interactions between Web services regulated in a community? How do we deal with conflicts in a community?

BACKGROUND

The term *community* means different things to different people. In *Longman Dictionary*, community is "a group

Figure 1. Example of composition scenario



of people living together and/or united by shared interests, religion, nationality, etc." In the field of knowledge management, communities of practice constitute groups within (or sometimes across) organizations who share a common set of information needs or problems (Davies, Duke, & Sure, 2003). A community is not a formal organizational unit but an informal network of entities with common interests and concerns.

When it comes to Web services, Benatallah et al. (2003) define community as a collection of Web services with a common functionality, although these Web services have distinct nonfunctional properties. Medjahed and Bouguettaya (2005) consider community as a means to provide an ontological organization of Web services sharing the same domain of interest (Budak Arpinar, Aleman-Meza, Zhang, & Maduko, 2004). Finally, Maamar, Lahkim, Benslimane, Thiran, and Sattanathan (2007) define community as a means to provide a description of a desired functionality without explicitly referring to any concrete Web service (already known) that will implement this functionality at run time.

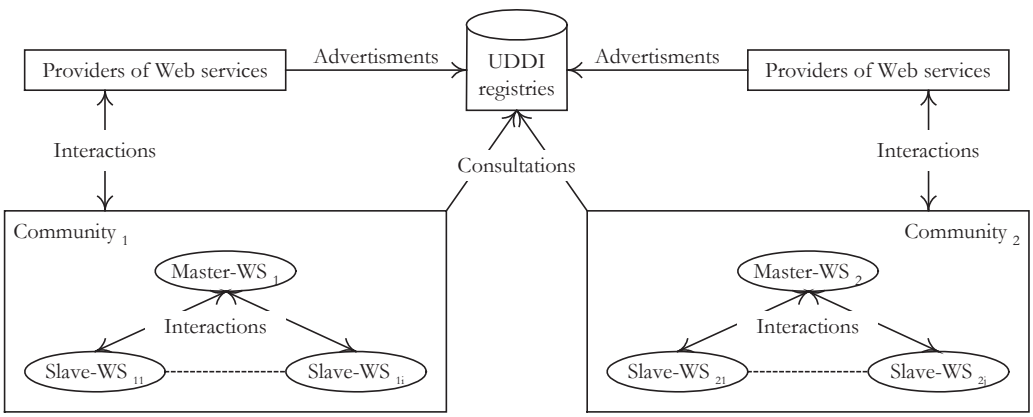
It is worth examining the similarity between a society of software agents and a community of Web services. A society is a group of agents of different types and capabilities that come together in order to collaborate and meet some common goals (Narendra, 2001). This does not apply in communities. Web services in a community do not collaborate. They rather compete to participate in composite Web services since they all offer the same functionality but in a different configuration. The collaboration for the sake of developing a composite Web service takes place at the community level where component Web services from independent communities work together. Each community contributes one Web service to the composite Web service.

CONCEPTS AND OPERATIONS IN COMMUNITIES

Figure 2 represents an environment of communities of Web services, providers of Web services, and UDDI (or ebXML) registries. Communities are dynamically established and dismantled according to protocols defined in the community management perspective. UDDI registries receive advertisements of Web services from providers. Several UDDI registries can be made available to providers for advertisement needs. In this context, competitor providers do not want to have their Web services posted on the same UDDI registries (Budak Arpinar et al., 2004).

The environment in Figure 2 offers some characteristics that need to be stressed. First, the regular way of describing, announcing, and invoking Web services is still the same, although Web services are now elements of communities. Second, the mechanisms that UDDI registries regularly offer in terms of announcing and discovering Web services are still the same. Finally, the selection of Web services out of communities is transparent to users and independent of the way these Web services are gathered into communities. A master component always leads a community. This component is itself implemented as a Web service in compliance with the rest of the Web services in a community, which are now denoted as slaves. A master-slave Web services relationship is framed in the interaction management perspective. One of the responsibilities of the master Web service is to attract Web services to join its community using rewards (Bentahar et al., 2007). This happens as part of the Web services attraction and retention perspective. As a result, the master Web service regularly screens UDDI registries so that it knows the latest advertisements of Web services.

Figure 2. Representation of communities of Web services



In a community, the master Web service is designated in two different ways. The first way is to have a dedicated Web service play the role of master for the time being in a community. This Web service is independently developed from other Web services that are advertised in UDDI registries. It is noted that the leader Web service in a community does not participate in any composition. As a result, this Web service is only loaded with mechanisms related to community management. The second way of designating a master Web service is to identify a Web service out of the Web services that reside in a community. This identification happens either on a voluntary basis or by election. Because of the temporary no-participation restriction of a master Web service in compositions, the nominated Web service is compensated by its peers. The call for elections in a community happens frequently so that the burden on the same Web services leading a community is minimized and hopefully avoided.

Community Management Perspective

A community gathers Web services with similar functionalities. This gathering is a designer-driven activity that includes two steps. The first step is to define the functionality (e.g., flight booking) of the community by binding to a specific ontology (Medjahed & Bouguettaya, 2005). This binding is crucial since providers use different terminologies to describe the functionality of their respective Web services. For example, flight booking, flight reservation, and air-ticket booking are all about the same functionality. The second step is to deploy the master Web service that will lead the community and take over multiple responsibilities. One of them is to invite and convince Web services to sign up for its community. The survivability of a community, that is, to avoiding dismantlement, depends to a certain extent on the status of the existing Web services in this community. Another responsibility is to check the credentials of Web services before they get admitted into a community. The credentials could be related to quality of service (QoS), protection mechanisms, interaction protocols, and so forth. Credential checking boosts the security level within a community and enhances the trustworthiness level of a master Web service toward the slave Web services.

Dismantling a community is another designer-driven activity, which happens upon request from the master Web service. This one oversees the events in a community such as the arrival of new Web services, the departure of some Web services, the identification of Web services to take part in composite Web services, sanctions on Web services because of misbehavior, and so on. When a master Web service observes first that the number of Web services in a community is less than a certain threshold and second that the number of participation requests in composite Web services that arrive from users over a certain period of time is also less than another threshold, then the community will be dismantled. Both thresholds are set by the designer. Web services to withdraw out of a community are invited to join other communities subject to assessing the functionality similarity with other existing communities' functionalities.

Web Services Attraction and Retention Perspective

Attracting new Web services and retaining existing Web services in a community fall into the responsibilities of the master Web service. A community could vanish if the number of Web services running in it drops below a certain threshold.

Attracting Web services requires the master Web service to consistently consult the different UDDI registries in looking for new Web services. These latter services could have recently been posted on UDDI registries or have had the description of their functionality changed. Changes in the functionality of a Web service raise challenges as this Web service may no longer be appropriate for a community. As a result, this Web service is invited to leave the community. When a candidate Web service is identified based on the functionality it offers, the master Web service interacts with its provider (Figure 2). The purpose is to ask the provider to register its Web service with the community of this master Web service. Some arguments to convince the provider include a high rate of participation of the existing Web services in composite Web services (a good indicator of the visibility of a community of Web services to the external environment and the reputation of Web services)(Maximilien & Singh, 2002), short response time when handling users' requests, and efficiency of the security mechanisms against malicious Web services.

Retaining Web services in a community for a long period of time is a good indicator of the following elements.

- Although Web services in a community are in competition, they expose a cooperative attitude. For instance, Web services have not been subject to attacks from peers in the community. This backs the security argument that the master Web service uses to attract Web services.
- A Web service is satisfied with its participation rate in composite Web services. This satisfaction rate is set by its provider. Plus, this is in line with the participation-rate argument that the master Web service uses to attract Web services.
- Web services know peers in the community that could replace them in the case of failure, with less impact on the composite Web services in which they are involved.

Web services attraction and retention shed light on a third scenario, which is when Web services are invited to leave a community. A master Web service could issue such a request upon assessment of the following criteria.

- The Web service has a new description of the functionality it provides. The description does not match the functionality of the community.
- The Web service is unreliable. On different occasions, the Web service failed to participate in composite Web services due to recurrent operation problems.
- The credentials of the Web service were "beefed up" to enhance its participation opportunities in composite Web services. Large differences between a Web services' advertised QoS and delivered QoS indicate performance degradation (Ouzzani & Bouguettaya, 2004).

Interaction Management Perspective

In a community, participation-related interactions in compositions between the master Web service and the slave Web services are framed using the contract-net protocol (CN; Smith, 1980). This protocol is built upon the idea of contracting and subcontracting jobs between two types of agents known as the initiator and the participant. At any time, an agent can be an initiator, a participant, or both. The sequence of steps in the contract-net protocol is as follows: (a) The initiator sends to participants a call for proposals in relation to a certain job to carry out, (b) each participant reviews the call for proposals and bids if interested (i.e., it is a feasible job), (c) the initiator chooses the best bid and awards a contract to that participant, and (d) the initiator rejects other bids.

Mapping the contract-net protocol onto the operation of a community occurs as follows. When a user (through some assistance; Schiaffino & Amandi, 2004) selects a community based on its functionality, the master Web service of this community is contacted in order to identify a specific slave Web service that will implement this functionality at run time. The master Web service sends all slave Web services a call for bids (CN_{step1}). Before contacting the master Web service, the slave Web services assess their status by checking their ongoing commitments in other compositions (CN_{step2}; Maamar, Benslimane, & Narendra, 2006). Only the slave Web services that are interested in bidding inform the master Web service. The latter screens all the bids before it chooses the best one (CN_{step3}). Afterward, the winning slave Web service is notified so it can get itself ready for execution when requested (CN_{step3}). The rest of the slave Web services that expressed interest but were not selected are notified as well (CN_{step4}).

FUTURE TRENDS

The design, management, and development of communities of Web services open up additional research venues that need to be looked into. Two are identified here, namely alliance development and OWL-S (ontology Web language for Web services) for community management.

Alliance Development

In *Longman Dictionary*, alliance is "an arrangement in which two or more countries, groups, etc. agree to work together to try to change or achieve something." One of the scenarios affecting the internal organization of a community is the setting up of alliances among Web services. An alliance is like a microcommunity whose development is triggered because of some mutual agreements between providers of Web services as part of their partnership strategy. Providers can join forces by referring to or recommending other peers' Web services and vice versa. Alliances constitute an attractive solution for exception handling. A Web service could be to a certain extent easily substituted with a peer in the same alliance before looking for another peer in other alliances in the same community. Similar to the dynamic nature of a community, an alliance has a dynamic nature as well: New alliances could be formed, new members could be admitted to as well as excluded from alliances, and some alliances could be either discarded or merged.

OWL-S for Community Specification

The specification of a community could happen using OWL-S. OWL-S organizes the description of a Web service along three categories: profile, process model, and grounding. The application of these categories in a community should happen as follows: what the community does (profile), how the community operates internally (model), and how the community accepts requests (grounding). In line with the profile, model, and grounding categories, a community could also be defined along the following three dimensions.

- Functional dimension: This describes the functionality associated with the community in terms of purpose, description ontology, and constraints. This dimension supports what needs to be done to advertise a community's functionality.
- Behavioral dimension: It describes the control flow of the functionality as depicted by the functional dimension. The control flow concerns task decomposition, chronology, and dependency. This dimension supports what needs to be done to achieve the functionality.
- Information dimension: It describes the data that are used throughout the performance of the functionality as

depicted by the behavioral dimension. This description includes the source of data, the semantic exchange of data, and the security of data. This dimension supports what needs to be provided to the functionality.

CONCLUSION

In this article, we discussed the mechanisms that should make communities of Web services active. Web services offering the same functionality are gathered into a single community in which a master Web service leads. This master is responsible for attracting new Web services to the community, retaining existing Web services in the community, and identifying the Web services in the community that will participate in composite Web services, among other things. The identification of these Web services happens in accordance with the contract-net protocol, which frames the interactions between the initiator and participant Web services. Dismantling a community could happen as well for various reasons, such as too small of a number of Web services residing in this community.

REFERENCES

Benatallah, B., Sheng, Q. Z., & Dumas, M. (2003). The self-serve environment for Web services composition. *IEEE Internet Computing, 7*(1).

Bentahar, J., Maamar, Z., Benslimane, D., & Thiran, P. (2007). Using argumentative agents to manage communities of Web services. In *Proceedings of the 2007 International Workshop on Web and Mobile Information Systems (WAMIS2007) held in conjunction with the IEEE 21st International Conference on Advanced Information Networking and Applications (AINA2007)*, Niagara Falls, Ontario, Canada.

Budak Arpinar, I., Aleman-Meza, B., Zhang, R., & Maduko, A. (2004). Ontology-driven Web services composition platform. In *Proceedings of the IEEE International Conference on E-Commerce Technology (CEC2004)*, San Diego, CA.

Davies, J., Duke, A., & Sure, Y. (2003). OntoShare: A knowledge management environment for virtual communities. In *Proceedings of the Second International Conference on Knowledge Capture (K-CAP2003)*, Sanibel Island, FL.

Maamar, Z., Benslimane, D., & Narendra, N. C. (2006). What can context do for Web services? *Communications of the ACM, 49*(12).

Maamar, Z., Lahkim, M., Benslimane, D., Thiran, P., & Sattanathan, S. (2007). Web services communities: Concepts & operations. In *Proceedings of the Third International Conference on Web Information Systems and Technologies (WEBIST2007)*, Barcelona, Spain.

Maximilien, M., & Singh, M. (2002). Concept model of Web service reputation. *SIGMOD Record, 31*(4).

Medjahed, B., & Bouguettaya, A. (2005). A dynamic foundational architecture for Semantic Web services. *Distributed and Parallel Databases, 17*(2).

Narendra, N. C. (2001). Flexible agent societies: Flexible workflow support for agent societies. In *Proceedings of the 2001 International Conference on Intelligent Agents Web Technologies and Internet Commerce (IAWTIC2001)*, Las Vegas, NV.

Ouzzani, M., & Bouguettaya, A. (2004). Efficient access to Web services. *IEEE Internet Computing, 8*(2).

Schiaffino, S., & Amandi, A. (2004). User-interface agent interaction: Personalization issues. *International Journal of Human Computer Studies, 60*(1).

Smith, R. (1980). The contract Net protocol: High level communication and control in distributed problem solver. *IEEE Transactions on Computers, 29*.

KEY TERMS

Alliance: An alliance is an arrangement in which two or more countries, groups, and so forth agree to work together to try to change or achieve something.

Community: It is a group of people living together and/or united by shared interests, religion, nationality, and so on.

Composition: It targets users' needs that cannot be satisfied by any single available Web service: A composite Web service obtained by combining available Web services may be used instead.

Contract-Net Protocol: It contracts and subcontracts jobs between two types of agents known as the initiator and the participant.

Ontology Web Language for Web Services (OWL-S): OWL-S supplies Web service providers with a core set of markup-language constructs for describing the properties and capabilities of their Web services in unambiguous, computer-interpretable form.

Universal Description, Discovery, and Integration (UDDI): It is a specification that provides a platform-independent way of describing services, discovering businesses, and integrating business services using the Internet.

Web Service: It is a software application identified by a URI whose interfaces and binding are capable of being defined, described, and discovered by XML (extensible markup language) artifacts, and that supports direct interactions with other software applications using XML-based messages via Internet-based applications.

Design and Implementation of Scenario Management Systems

M. Daud Ahmed
Manukau Institute of Technology, New Zealand

David Sundaram
University of Auckland, New Zealand

INTRODUCTION

Scenarios have been defined in many ways, for example, a management tool for identifying a plausible future (Porter, 1985; Schwartz, 1991; Ringland, 1998; Tucker, 1999; Alter, 1983) and a process for forward-looking analysis. A scenario is a kind of story that is a focused description of a fundamentally different future (Schoemaker, 1993), that is plausibly based on analysis of the interaction of a number of environmental variables (Kloss, 1999), that improves cognition by organizing many different bits of information (De Geus, 1997; Wack, 1985; van der Heijden, 1996), and that is analogous to a "what if" story (Tucker, 1999). It can be a series of events that could lead the current situation to a possible or desirable future state. Scenarios are not forecasts (Schwartz, 1991), future plans (Epstein, 1998), trend analyses, or analyses of the past. Schoemaker (1993) also explains that scenarios are for strategy identification rather than strategy development. Fordham and Malafant (1997) observe that decision scenarios allow the policymaker to anticipate and understand risk, and to discover new options for action. Ritson (1997) agrees with Schoemaker (1995) and explains that scenario planning scenarios are situations planned against known facts and trends, but deliberately structured to enable a wide range of options and to track the key triggers that would precede a given situation or event within the scenario.

In this article we propose an operational definition of scenarios that enables us to manage and support scenarios in a coherent fashion. This is then followed by an in-depth analysis of the management of scenarios at the conceptual level as well as at the framework level. The article goes on to discuss the realization of such a framework through a component-based layered architecture that is suitable for implementation as an n-tiered system. We end with a discussion on current and future trends.

BACKGROUND

The basic structure and behavior of the scenario is similar to the decision support system (DSS) components *model* and *solver* respectively. In information systems literature, a use case instance irrespective of transaction or decision context is considered as a scenario. But scenarios are primarily related to complex business change management processes; they might address semi-structured and unstructured decision problems. Hence we define scenario as a complex decision situation analogous to a model that is instantiated by data and tied to solver(s). In its simplest form, scenario is a complex combination of data, model, and solver.

Decision makers have been using the concepts of scenarios for a long time, but due to their complexity, their use is still limited to strategic decision-making tasks. Scenario planning varies widely from decision maker to decision maker, mainly because of lack of a generally accepted principle for scenario management. Albert (1983) proposes three approaches for scenario planning: expert scenario approach, morphological approach, and cross-impact approach. Ringland (1998) describes three-step scenario planning: brainstorming, building scenarios, and decisions and action planning. Schoemaker (1995) outlines a 10-step scenario analysis process. Huss and Honton (1987) identify three categories of scenario planning: intuitive logics, trend-impact analysis, and cross-impact analysis. These planning processes are useful but they are not entirely supported by the available decision support systems frameworks. Either or both of the existing scenario planning processes and the DSS frameworks needs to be modified for planning scenarios within DSS.

SCENARIO MANAGEMENT SYSTEMS

Few of the decision support system frameworks emphasize a lifecycle approach based fully featured scenario planning, development, analysis, execution, and evaluation environment. DSS components such as data, model, solver, and visualization have been extensively used in many DSS framework designs, but they did not consider scenario as a component of DSS. Scenario plays such an important role in the decision-making process that it is almost impractical to develop a good decision modeling environment while leaving out this component.

Figure 1. Scenario-driven decision system framework

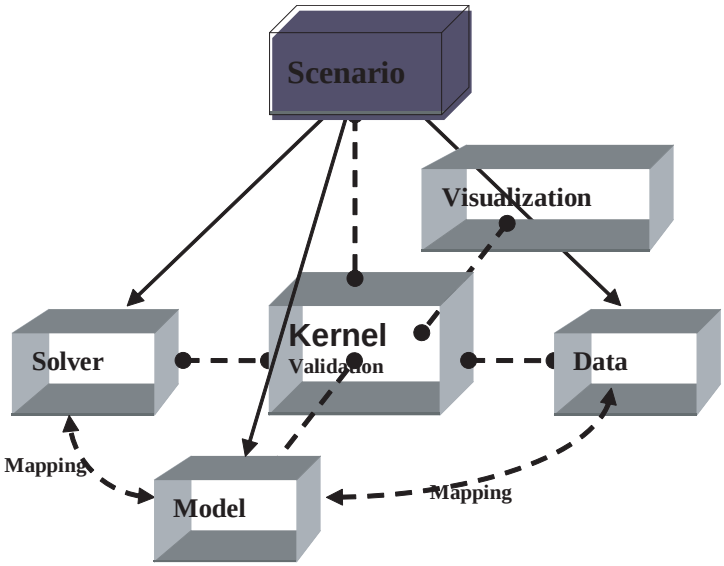
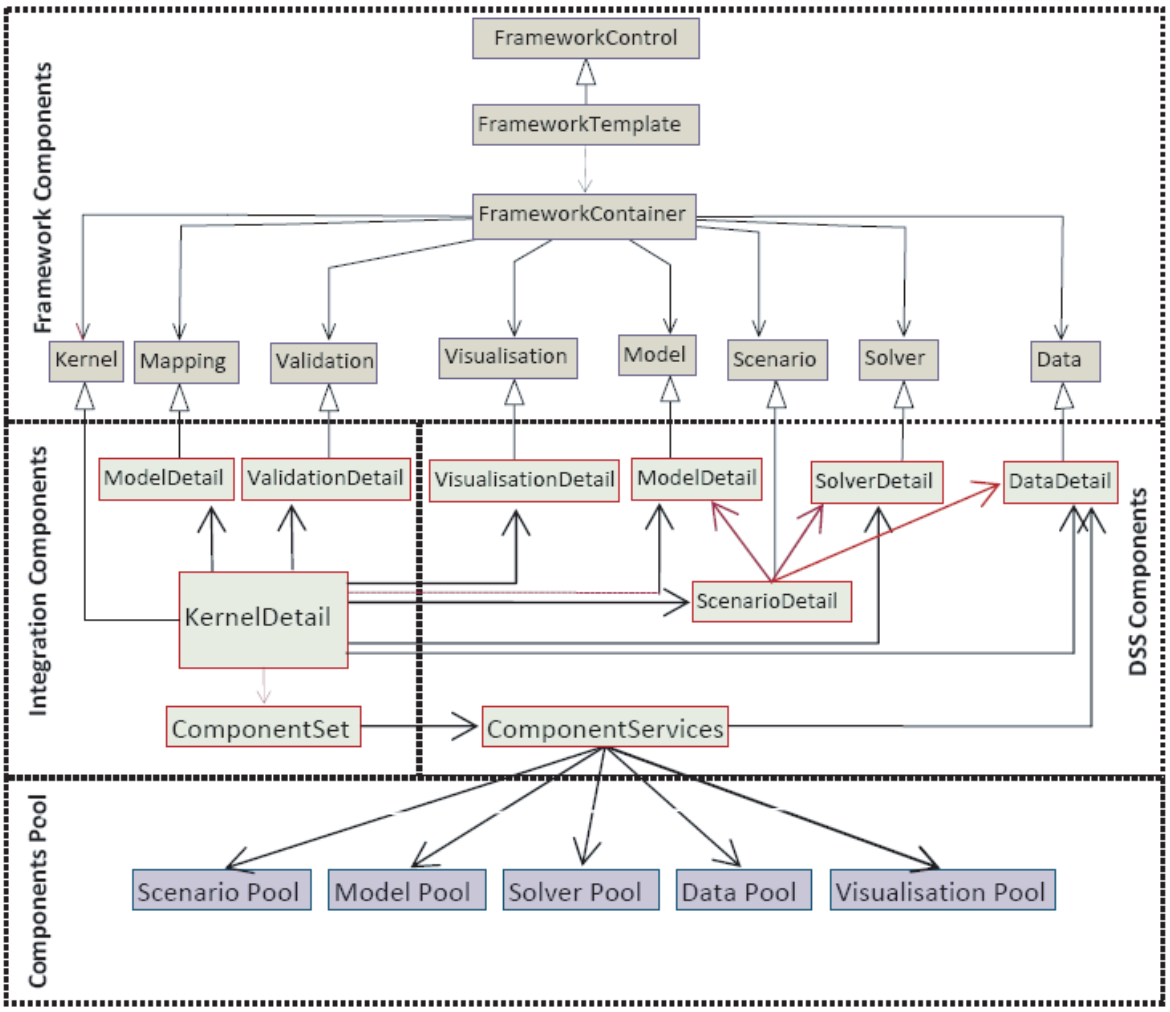


Figure 2. SDSSG system architecture



To overcome these shortcomings we propose a scenario-driven decision support systems generator (SDSSG) framework, as illustrated in Figure 1, to effectively manage scenarios and their lifecycle. The SDSSG components are categorized into decision support components (DSC) and integration components (IC). DSC includes the data, model, solver, scenario, and visualization components, and IC includes kernel, mapping, and validation components. The kernel is supported by a component set that interacts with a components pool that includes the data pool, model pool, solver pool, scenario pool, and visualization pool as outlined in SDSSG system architecture.

The system architecture showing the component is presented in Figure 2. While the architecture is generic enough that it can be used to support DSS, it has been extended keeping in mind the requirements of scenario management. The architecture supports scenario planning, development, analysis, execution, and evaluation. Figure 2 illustrates the inheritance, aggregation, and dependency relationship among the architectural components. However their attributes and behaviors are omitted for the sake of simplicity.

Components of the SDSSG Architecture

The design of the architecture roughly follows the framework described above and is broken up into four parts: framework components, integration components, DSS components, and components pools. The framework components are the highly abstract-level components, and integration and DSS components are concrete-level components. The components pool is independent from the SDSSG system components. The components pool contains five different types of components that are used for component state management.

We use object-oriented concepts for designing the decision support and integration components. The concepts of abstraction, encapsulation, inheritance, and polymorphism are basic concepts for component development. Modularity, sub-system, and packaging concepts have been used for managing the components. The fields, properties, methods, and events are used to develop the structure, behavior, and message services of the components while the class is used to develop the object, and the reuse of the object is ensured by designing proper component interfaces.

Framework Components

The framework components of the SDSSG architecture are composed of abstract-level model, solver, data, scenario, kernel, validation, and mapping components. These components are described under decision support components and integration components. In addition a FrameworkControl, a FrameworkTemplate, and a FrameworkContainer are also included to support the flexible replacement of decision support components at runtime.

Decision Support Components

The decision support components of the SDSSG architecture are DataDetail, ModelDetail, SolverDetail, VisualisationDetail, and ScenarioDetail. These components are created by inheriting the abstract-level framework components of the same type. ModelDetail refers to the real-life problems, DataDetail represents parameter instance of the model, and SolverDetail is the operation that can be executed on the model. The solver is the separated behavior(s) of the model from its structure. The scenario is the complex combination of the data, model, and solver. Visualization facilitates flexible presentation of data, executed model, or scenario. In addition we have used a ComponentServices component that is responsible for interaction with component pools for retrieval and update of components.

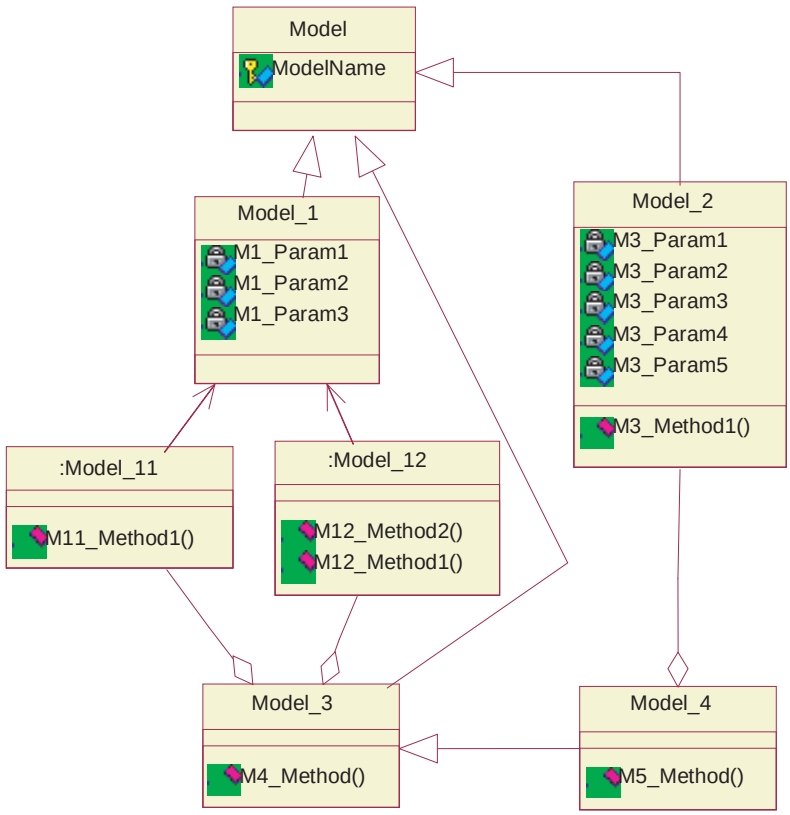
Model Component

The model component may contain many interrelated models as shown in Figure 3. The top-level abstract model has an association with the framework template. This model is detailed through inheritance and aggregation, and then associated with the kernel and scenario. Figure 3 shows that the concrete-level model is developed using other base-level models. In the process of detailing or developing concrete models, we use the concepts of abstraction, inheritance, and aggregation. The model can also contain an instantiated model as shown in Figure 3. It can be of the primitive type or the compound type (Geoffrion, 1987). Primitive-type models are directly derived using base data-type variables as well as executed model values of the base-level models. The compound-type model either inherits or aggregates the base models or may both inherit and aggregate as well as add some other independent parameters. The top-level model class is associated with the framework template while the concrete model classes are associated with the kernel.

Solver Component

The solver component is made up of algorithms for applying behaviors on the model instance. The design concept is almost similar to the design of the model component as discussed above. Each solver contains methods related to specific operations and is implemented through overloading, overriding, and shadowing techniques of object-oriented methodology.

Figure 3. A simple model hierarchy



Data Component

The data component is designed to represent facts of business operations and decision scenarios. The data component can be a subset of a database, a data table, or a record or a serialized object or a formatted file/document. It may be any or all of binary, xml, text, relational data, and ADO.NET object formats. It describes value and data type of discrete information. An abstract-level data object is detailed for its suitable use with the model and the scenario. This component is also used in other components—namely, kernel, ComponentSet, and ComponentServices.

Visualization Component

The visualization component is designed for presentation of model or scenario-executed data or text. Most of the visualizations are flexible and created at runtime.

Scenario Component

A unique problem is central to scenarios, but the instances and implementation environment could be diverse. As we have described, the scenario in its simplest form is a combination of data, model, and solver. The architecture allows generating a number of simple, pipelining, and aggregate scenarios. Scenario information can be saved to the scenario pool. Previously developed scenarios can be retrieved from the scenario pool, and the same can be customized using models and solvers. The scenarios can be used as specific DSSs or as complex data for input to the next level of model for further analysis. Different scenarios can be computed simultaneously, and sensitivity and goal-seek analysis can be done using different scenarios. The architecture is suitable for analyzing internally coherent scenarios or scenario bundles, and examining the joint consequences of changes in the environment for supporting the decision maker's strategy.

ComponentsServices Component

The ComponentServices component is designed to work with the component pool using universal data access technology (e.g., OLE DB, ODBC). This component works between the component set and the component pool. It contains several parameters and methods that support export, and import all sorts of component information to and from data, model, solver, and scenario pools. Some of the methods have been predefined, but they can be changed at the runtime to suit the changed component pool management system.

Integration Components

The integration component is composed of the details of kernel, validation, mapping, and ComponentSet. These components are described below.

Kernel and User Interface

The kernel is the integration tool of the system that integrates the decision support components (i.e., data, model, solver, visualization, scenario) and other associated integration components (such as component set, mapping, and validation). The class diagram of the kernel and the associated components are shown in Figure 4.

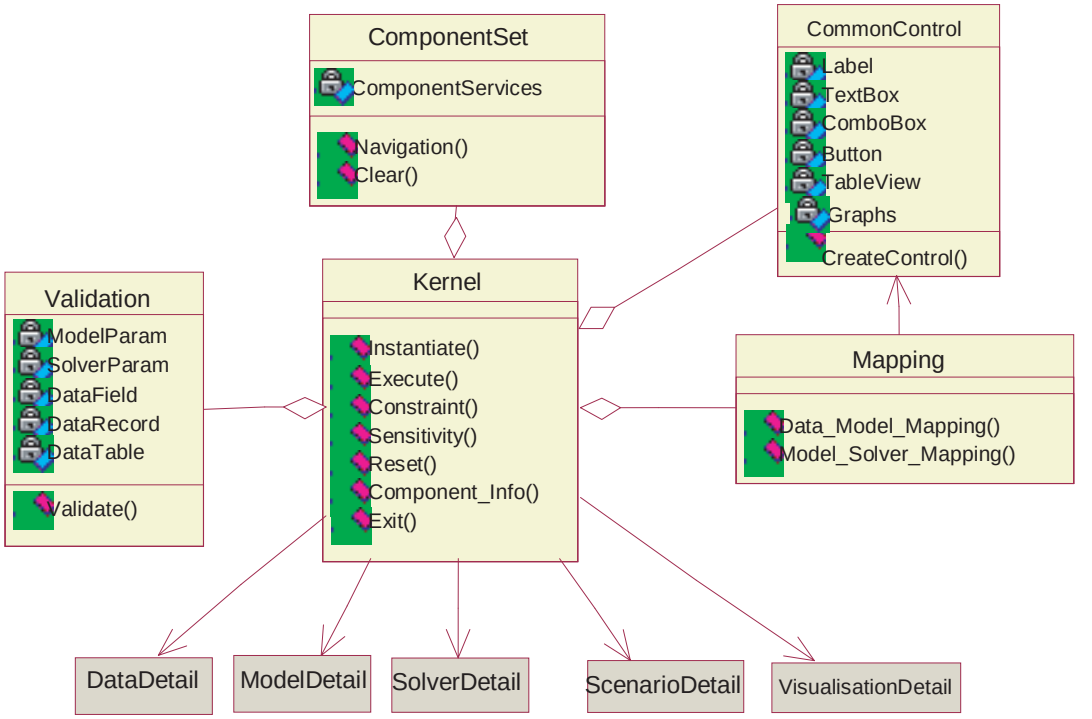
The kernel component works as the center of communication between and among the decision support components. Components or component instances are called inside the kernel when any member of the component is instantiated or invoked at runtime. The kernel is designed for both flexible coupling and tight coupling of data, model, solver, and scenario. The mapping component plays a vital role for flexible coupling. For the tightly coupled system, usage of data, model, and solver are predefined. The model understands which data is to be picked up on the basis of data and model activation to instantiate the model and which solver is to be called to execute the model.

In addition to activating and using the component functions, the kernel generates runtime user interfaces for communicating with and integrating with other components of the system with the support of CommonControl Class. This is a container of many controls.

Mapping Component

The mapping is a process-oriented generic component that enables the model component to communicate with data and

Figure 4. Kernel and its associated components



solver components properly while they are completely separate from one another. In SDSSG, the mapping component is composed of data-model and model-solver mapping. During the mapping, this component collects the name and data type of attributes of the selected data, model, and solver from the component set and presents them on a dynamically created user interface. This process is designed in such a way that the decision maker does not need to change any attribute name in the component set.

The mapping component facilitates the communication between two attributes having different names. The decision maker then arranges the attributes of the components for communication. The model is the center for mapping the attributes. The model attributes are fixed, and the user selects the data attributes for model-data mapping and selects the solver name and solver attributes for model-solver mapping. The mapping component also facilitates transformation of data types (e.g., from integer type to string type, or string type to double type, etc.). The system can automatically transform lower-level data types to higher-level data in a hierarchical data structure of the base data type (e.g., integer to double value), but any transformation from higher-level data types to lower-level data types must be done through explicit operation. This operational method for transformation can be built-in with the model, solver, or kernel.

Validation Component

The validation component is responsible for checking the input data type of the model and the solver after the mapping. In model-data validation checking, the data type of the model attribute is fixed and checks whether the data attributes are similar or convertible to the data type of the model attribute.

For model-solver validation, the data type of the solver attribute is fixed and checks whether the data types of the attributes of the model instance are similar or convertible to the data type of the solver attribute.

Component Set

The component set is an independent component that performs as a temporary repository of the component information and works as a data access layer with the support of the ComponentServices component. The component set can be filled with records from the data pool, model pool, solver pool, and scenario pool. A component set can be loaded with a set of records, model, solver, and scenario data. The component set is also responsible for uploading the data from the data grid to the user interface. There are two locations for uploading the data. One area is customized for a specific domain while the other area is a general interface that can be used for uploading a component table. In this way the framework is able to work with many databases and data tables. Once data are added to the component set, the component schema is also updated with the new data. The schema can easily be converted to an XML base. The component set can also be reconstructed from the XML base. Therefore, the system works without a database system or without having connection to the central DBMS system. This conversion facilitates the portability of the system.

Components Pool

The components pool contains the data pool, model pool, solver pool, and scenario pool, which are collections of domain data, model information, solver information, and scenario information respectively. The model pool, solver pool, and scenario pool can contain generalized as well as domain-specific model, solver, and scenario information. The records of model and solver pools are dependent on the structure of the models and solvers to be used in the system, while the scenario records are dependent on the structure of the scenario as well as the type of scenario to be analyzed. The method of the scenario component is a generalized method that is responsible for developing scenario instance from the information of the scenario pool, as well as using the data table, model pool, and solver pool. The scenario method is domain independent, but the scenario instance is domain specific. The component pool can be developed using a database language outside the decision support system. The component pool is fully separated and independent from the SDSSG application. This enables us to create, update, maintain, and manage the component pool without involving the application program. The SDSSG is capable of using the data tables that have the required data.

IMPLEMENTATION OF THE SDSSG ARCHITECTURE

The architecture can be implemented as a single-, two- or three-tiered architecture. Since object-oriented and component-based concepts are the central focus of our framework and architecture, Microsoft's .NET framework, was considered as the implementation platform and C# was used for implementing the system architecture. Component technologies (e.g., Dynamic Link Library, Component Object Model), database management systems (e.g., SQL 2000 Server, Microsoft Access), and extensible markup language (XML) were also used in building the system. The important features of the implementation are described in the following sections.

Component Independence and Interoperability

Each and every architectural component is designed as an independent unit using the .Net framework concept that supports the COM+ specification. They are compiled to intermediate language as DLL. The components are compiled to native code by a runtime compiler when they are called to execute a task. The components are independent until they are not actively participating in the process. The intermediate language-based components and their compilation at runtime make them interoperable with COM/COM+ components.

Runtime Code and Scenario Management

The user can write database access codes at runtime to work with different DBMSs. The construction of scenario from data, model, and solver from the component pool also depends on the runtime written command. The runtime-generated and executed scenarios are managed through a runtime scenario pool for developing upper-level scenarios, as well as sensitivity and goal-seek analysis.

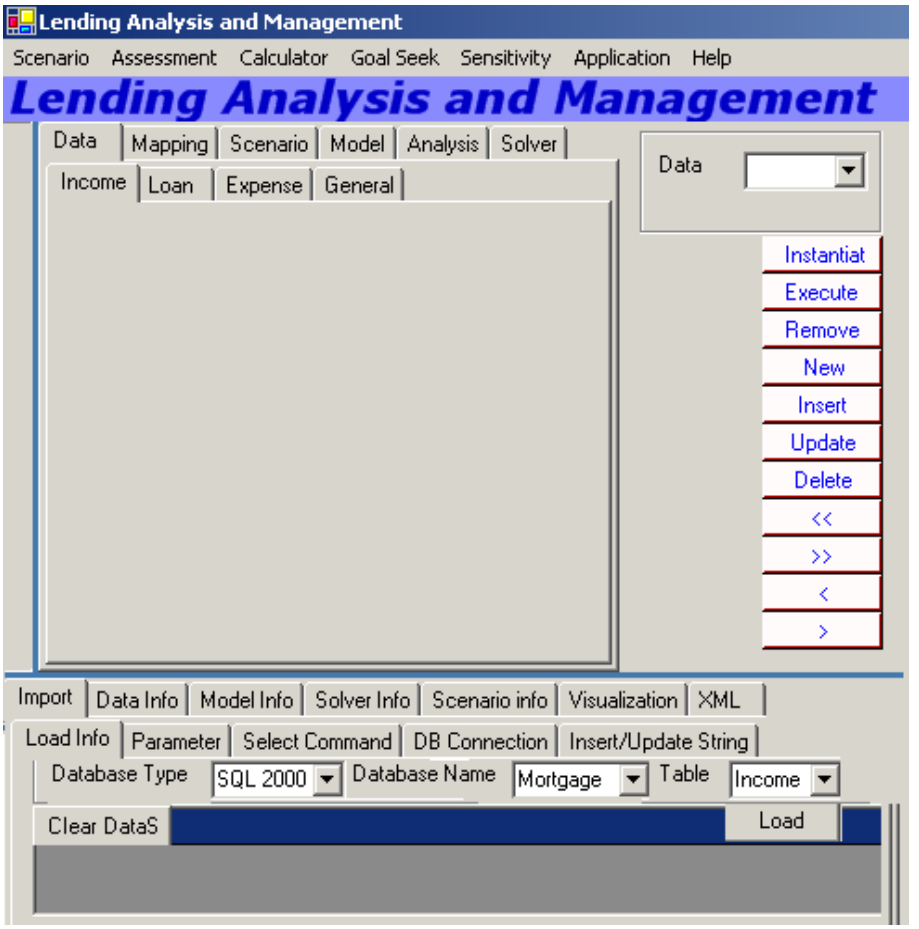
Component Versioning and Extensibility

The framework template is specifically designed for customization of the architecture by replacing its component(s) with different versions of the same component. So the architecture is extensible to accommodate multiple versions of a component.

Data Access and Component Set

The system is able to import data from a database system that has an ODBC or OLE DB interface, text data, or XML-

Figure 5. The SDSSG implementation in the mortgage domain



formatted data. It uses ADO.net architecture that supports the data structure independence. The data is managed through a temporary disconnected runtime repository named Component Set, which can hold multiple component tables at a time. The component performs as a runtime DBMS schema and supports defining relationships between the component tables that help in developing the scenario instance.

Presentation

The execution results of the scenarios are presented in dynamically runtime-created visualization for evaluation through comparison. The visualization can accept a replacement/update of a scenario. A new scenario can be added to the visualization or can be deleted from the visualization.

Ease of System Use

The implemented SDSSG system is suitable for both the naïve user as well as the DSS builder. The naïve user can easily use the semi-automated, tightly coupled system while the DSS builder can use a versatile, flexible-coupled system. Some of the process can be done using a flexible-coupled system while the rest of the process can be completed using a tightly coupled system.

Implementation Domain

The SDSSG framework, architecture, and implementation were tested within the context of the mortgage domain (see Figure 5). Specifically, we implemented affordability scenarios, lending scenarios (equal installments, reducing installments, interest only, etc.), and payment scenarios. Within each of these scenarios, we explored sensitivity and goal-seek analyses.

Once a base scenario has been developed, we explore a number of alternative scenarios, including the best-case and worst-case scenarios through sensitivity analysis. We used the system to compare multiple scenarios of similar type (homogeneous comparison) or different types (heterogeneous comparison) or both homogeneous and heterogeneous at a time in a single visualization.

The system was tested and evaluated for sensitivity analysis for refinancing from different lending sources, and increase or decrease of the interest rate, loan amount, initial payment, installment, and pay period. Apart from this we also explored sensitivity analysis on complex interlinked scenarios, which in turn were made up of sub-scenarios. The system supports complex analyses from the very lower-level scenarios to higher-level/aggregate scenarios. Scenarios analyzed from the bottom up may or may not satisfy the prime objective. In this circumstance, a top-down scenario analysis (goal-seek analysis) could bring to light the optimum acceptable scenarios.

FUTURE TRENDS

We have designed the SDSSG framework and architecture using the modeling-based DSS concept and implemented it using the modeling as well as process-based development methods. It supports scenario planning, scenario modeling, scenario development, and execution and decision support for the decision maker at runtime. This system integrates scenario with the model-based DSS, but it overlooks the integration of knowledge-based DSS into the SDSSG framework and architecture. The proposed process-based development method is suitable to incorporate the knowledge-based sub-system in the decision-making process. The knowledge sub-system or case-based reasoning helps the decision maker reference the past decision scenarios as well as the decisions. The integration of knowledge-based DSS with the SDSSG will be a unique decision support system that would help exhaustive scenario analysis as well as uniform learning throughout the organization.

CONCLUSION

The existing scenario planning and analysis systems are very complex, they are not user friendly, they cannot support multiple scenarios modeling at one time, and they do not provide any facilities for scenario evaluation or comparison between multiple scenarios. The scenario-based decision support process is mostly being used for developing corporate strategies. But scenarios are useful for tactical and operational level decisions as well.

This research develops a generic scenario-driven flexible decision support systems generator framework and architecture that supports complex decision-making processes. It supports sensitivity and goal-seek analysis. It uses decision support components, integration components, and component pools for scenario planning, development, and analysis. Scenario has been introduced as a new DSS component that is developed as a complex combination of other decision support components—namely, data, model, and solver. The proposed framework and architecture are domain and platform independent, component based, and modular. The architecture is composed of multiple layers, for example, component pool layer, data access layer, decision support services layer, integration layer, and user services layer. Each layer performs specific functions, which are suitable for implementation of the architecture as a single-, two-, or three-tiered system. Architectural components can be updated with new versions; they are interoperable with other

platforms. A prototype was developed using the framework and architecture. The implementation has been evaluated with the objectives and design goals in the context of the mortgage domain. Significant extensions to this framework and architecture have been proposed and implemented, while retaining the core constructs in other domains with different paradigms (Ahmed & Sundaram, 2007). Key principles from this research have formed the foundation stone in implementations in the field of sustainability modeling and reporting (Ahmed & Sundaram, 2008).

REFERENCES

Ahmed, M.D., & Sundaram, D. (2007). A framework for sustainability modeling and reporting. *International Journal of Environmental, Cultural, Economic and Social Sustainability, 3*(2), 29-40.

Ahmed, M.D., & Sundaram, D. (2008). Sustainability modelling and reporting: Integrative frameworks and architectures. In B.S. Sahay, J.N.D. Gupta, S. Kumar, & S. Kumar (Eds.). *Decision sciences and technology for globalization* (pp. 526-540). New Delhi: Allied.

Albert, K.J., (1983). *The strategic management handbook.* New York: McGraw-Hill.

Alter, S.L. (1980). *Decision support systems: Current practice and continuing challenge.* Reading, MA: Addison-Wesley.

De Geus, A. (1997). *The living company: Habits for survival in a turbulent business environment.* Boston: Harvard Business School Press.

Epstein, J.H. (1998). Scenario planning: An introduction. *Futurist, 32*(6), 50-51.

Fordham, D.P., & Malafant, K.W.J. (1997). The Murray-Darling basin irrigation futures framework. *Proceedings of the International Congress on Modelling and Simulation Conference* (MODSIM 97) (vol. 2, pp. 643-648).

Geoffrion, A. (1987). An introduction to structured modeling. *Management Science, 33*(5), 547.

Huss, W.R., & Honton, E.J. (1987). Scenario planning: What style should you use? *Long Range Planning,* (April).

Kloss, L.L. (1999). The suitability and application of scenario planning for national professional associations. *Nonprofit Management & Leadership, 10*(1), 71-83.

Porter, M. (1985). *Competitive advantage*. New York: The Free Press.

Ringland, G. (1998). *Scenario planning managing for the future.* New York: John Wiley & Sons.

Ritson, N. (1997). Scenario planning in action. *Management Accounting, 75*(11), 24-28.

Schoemaker, P.J.H. (1995). Scenario planning: A tool for strategic thinking. *Sloan Management Review, 36*(2), 25-40.

Schoemaker, P.J.H. (1993). Multiple scenario development: Its conceptual and behavioral foundation. *Strategic Management Journal, 14*(3), 193-213.

Schwartz, P. (1991). *The art of the long view.* New York: Doubleday.

Tucker, K. (1999). Scenario planning. *Association Management, 51*(4), 70-75.

van der Heijden, K. (1996). *Scenarios, the art of strategic conversation.* New York: John Wiley & Sons.

Wack, P. (1985). Scenarios, uncharted waters ahead. *Harvard Business Review.*

KEY TERMS

Aggregate Scenarios: The structure of different scenarios or results from multiple scenarios are combined/aggregated together to develop a more complex scenario.

Decision Support Systems/Tools: In a wider sense, can be defined as systems/tools that affect the way people make decisions. In our present context it is defined as systems that increase the intelligence density of data and support interactive decision analysis.

Goal-Seek Analysis: Accomplishes a particular task rather than analyzing the changing future. This goal-seek analysis is just a reverse or feedback evaluation where the decision maker supplies the target output and gets the required input.

Intelligence Density: The useful 'decision support information' that a decision maker gets from using a system for a certain amount of time, or alternately the amount of time taken to get the essence of the underlying data from the output.

Pipelining Scenarios: One scenario is an input to another scenario in a hierarchical scenario structure. In this type of scenario, the lower-level scenario can be tightly or loosely integrated with the higher-level scenario.

Scenario: A complex problem situation analogous to a model that is instantiated by data and tied to solver(s). A scenario can be presented dynamically using different visualizations. A scenario may contain other scenarios.

Sensitivity Analysis: Allows changing one or more parametric value(s) at a time and analyzes the outcome for the change. It reveals the impact on itself as well as the impact on other related scenarios.

Simple Scenarios: The simple scenario is not dependent on other scenarios, but completely meaningful and usable.

Design Levels for Distance and Online Learning

Judith V. Boettcher
Designing for Learning and the University of Florida, USA

INTRODUCTION

The importance of design for instructional programs — whether on campus or online or at a distance — increases with the possible combinations of students, content, skills to be acquired, and the teaching and learning environments.

Instructional design —as a profession and a process— has been quietly developing over the last 50 years. It is a multidisciplinary profession combining knowledge of the learning process, humans as learners, and the characteristics of the environments for teaching and learning. The theorists providing the philosophical bases for this knowledge include Dewey (1933), Bruner (1963), and Pinker (1997). The theorists providing the educational and research bases include Vygotsky (1962), Knowles (1998), Schank (1996), and Bransford, Brown, and Cocking (1999).

Instructional design offers a structured approach to analyzing an instructional problem and creating a design for meeting the instructional content and skill needs of a population of learners usually within a specific period of time. An instructional design theory is a "theory that offers explicit guidance on how to better help people learn and develop" (Reigeluth, 1999).

BACKGROUND

This entry describes a multi-level design process for online and distance learning programs that builds on a philosophical base grounded in learning theory, instructional design, and the principles of the process of change as reflected in the writings of the theorists listed above. This design model builds on traditional instructional design principles, as described by Gagne (1965), Dick & Carey (1989), and Moore & Kearsley (1996). It integrates the strategic planning principles and the structure of the institutional context as described in Kaufman (1992) and Boettcher & Kumar (1999), and also integrates the principles of technological innovation and the processes of change as described by E. M. Rogers (1995) and R. S. Rosenbloom (1998).

This entry describes a six-level design process promoting congruency and consistency at the institution, infrastructure, program, course, activity, and assessment level. It also suggests a set of principles and questions derived from that framework to guide the instructional design process.

SIX LEVELS OF DESIGN

Effective instructional design for online and distance learning benefits from instructional planning at six levels. Figure

Figure 1. Six levels of design for learning

Six Levels of Design	Design Responsibility	Sponsor/Leader	Design and Review Cycle
Institution	Entire campus leadership and community	Provost, CIO and Vice-presidents	3-5 Years
Infrastructure	Campus and Technology Staff	Provost, CIO and Vice-presidents	2-3 Years
Degree, Program	College/Deans/Faculty	Dean and Chairs	1-3 Years
Course	Faculty	Dept Chair	1-2 Years
Unit/Learning Activity	Faculty	Faculty and or Faculty team	1-2 Years
Student Assessment	Faculty	Faculty and or Faculty team	1-2 Years

1 summarizes these six levels of design, and identifies the group or individuals usually responsible for the design at that level and the length of the design cycle at each level. Ideally, the design at each of these six levels reflects philosophies of teaching and learning that are consistent with the institutional mission and consistent with the expectations of the students and society being served.

Level One: Institutional Design

The design work to be done at an institutional level is similar to the strategic planning and positioning of an institution. Institutional planning generally begins with an institution's current vision and mission statements and then proceeds through a data collection and input process that addresses a set of questions such as the following:

Institutional Questions:

- What programs and services comprise our primary mission? For whom?
- To what societal needs and goals is our institution attempting to respond?
- What life goals are most of our students working to achieve?
- What type of learning experiences are our students searching for?
- What changes in our infrastructure are needed to match our desired services, programs, and students?
- Does our institution have any special core competencies, resources, or missions that are unique regionally or nationally that might form the basis for specialized online and distance programs? What are the strengths of our mature faculty? Of our young faculty?

Level Two: Infrastructure Design

People often think that buildings, classrooms, Web applications, communication services, and servers are neutral as far as having an effect on teaching and learning. Nothing could be more misleading. Design of the infrastructure includes design of all the elements of the environment that impact the teaching and learning experiences of faculty and students and the staff supporting these experiences. It includes design of the following:

- Student services, faculty services, and learning resources.
- Design of administrative services, including admission processes, financial processes, and institutional community life events.
- Design of physical spaces for program launching events, hands-on, lab, or network gathering events, as well as celebratory graduation events.

Physical and Digital Plants

Infrastructure design for online and distance teaching and learning programs focuses on the design of the network and Web infrastructure. Infrastructures for online learning have offices, classrooms, libraries, and gathering spaces for the delivery and management of learning and teaching. However, these offices and classrooms are accessed through Web services, rather than through physical buildings. The good news about online infrastructures is that they support an unparalleled new responsiveness, feedback, and access for learning activities.

After almost ten years of building online campuses, we now know that a "digital plant" infrastructure is needed to support the new flexible online and distance environments. We know that this new digital plant needs to be designed, built, planned, maintained, and staffed. The infrastructure to support the new programs cannot be done with what some have called "budget dust" (McCredie, 2000). It is not nearly as easy or inexpensive as we all first thought. Some experts suggest that, a "full implementation of a plan for technology support on campus costs about the same as support of a library — approximately 5% of the education and general budget" (Brown, 2000).

Components of a Digital Infrastructure

What exactly is a digital plant infrastructure? One way of describing this infrastructure is to think of it in four major categories of personal communication tools, networks, hardware for servers, and software applications. A key component of the digital infrastructure is the group of individuals who make the systems work. This digital plant is shown in Figure 2 (Boettcher and Kumar, 2000).

Some of the questions that might be used to guide the development of the digital infrastructure follow.

Personal communication tools and applications:

- Will all students have their own computer? Their own laptop?
- Do we expect students all to be proficient with word processing applications, mail, Web applications, researching on the Internet? With collaborative tools and with one or more course management systems?

Networks that provide access to Web applications and resources and to remote, national, and global networks:

- What physical wired or wireless networks are needed to support Web applications, such as e-mail servers, directory servers, and Web application services?
- How often will higher bandwidths be needed for video conferencing for programs? For meetings? For downloading large files? For streaming video?

Dedicated servers and software applications that manage campus services:

- What types of interactive Web services will be provided? What hardware and software will be required?•What type of administrative systems and course management system will we use?
- What do we need to do to assure student, faculty, and staff accessibility from anywhere at anytime?

Software applications and services from external providers, such as research and library services that are licensed to the institutional community, Internet services, and outsourced services, such as network services:

- What licensed services are required and desired?
- What budget is required to support these services currently and into the future?

Technology decisions for students have always been part of the instructional design process for distance learning. A comforting way of thinking about the technology for the infrastructure design level is in terms of the generations of technologies used in distance learning. (Sherron and Boettcher, 1997). Distance learning was made possible with the widespread availability of technologies, such as the mail, radio, telephone, television, and audio and videocassettes. In the 21st century we simply have more technology and more choices.

Now let's look at the design of programs and courses. Design issues at these levels are principally the responsibility of the institutional academic leadership.

Level Three: Program Design

At the program level of design, instructional planners answer questions about the type of program to be offered, to whom, and over what period of time and at what cost. When venturing into new business areas, the following two guidelines are useful: (1) focus on programs that can leverage institutional core competencies and strengths, (2) plan a phased approach, gaining experience in delivering programs in one or two areas before launching others, and (3) recognize that online and distance learners generally are interested in achieving or completing an instructional goal that can assist in their current or future career path.

It is in the next four levels of design that the principles of Vygotsky are most applied, building on Vygotsky's (1962, 1978) view of the learner as a goal-oriented learner within a specific learning context using specific resources as directed by a teacher. These four core elements of all learning experiences provide a framework for the design process:

- The person doing the learning — the learner
- The person guiding and managing the learning —the faculty/teacher/mentor
- The content /knowledge/skill to be acquired/or problem to be solved
- The environment or context within which the learning experience occurs

Program Level Planning Questions:

Program planning design has four categories of planning — curriculum, design/development process, faculty, and student.

Curriculum questions:

- What is the degree or certificate program to be offered online? Will it be a full master's degree (10 to 16 courses), an undergraduate minor (four to six courses) or a certificate program (two to four courses)?
- What types of courses are envisioned? Will these courses be a fully developed "course in a box" with a minimal amount of interaction or a highly interactive and collaborative course requiring or using many online resources and applications?

Design and development questions:

- Who are the faculty who will design, develop and deliver the courses in the program?
- Who will lead the effort to develop the degree or certificate program for online or distance delivery? Which organization will be marketing the program?
- What course management system or similar Web tool will be used for the content management? What tools and resources will be available and supported for the interaction and collaboration activities?
- What is the schedule for design and development and delivery of courses and program? For the marketing and recruiting of the students?

Faculty questions:

- What training will be available to faculty as they transition to online teaching and learning programs?
- What tools and resources and support will be available to faculty?
- Will faculty have any released time or budget for teaching and learning resources in the new online or distance environment?
- What type of access to the network is recommended and available? Will dial-up be sufficient, or will DSL or cable access be recommended or required?

Student questions:

- Who are the students who will enroll in this course of study? How will we find them and market the program to them?
- What will our students bring to the program experience?
- What tools and resources will the student in this program or certificate program require or be likely to use?
- Where will the students be doing their learning and with what types of content resources and applications? What level of network access is required or recommended?

The question of technology access for students was particularly important in the mid-1990s, when technology access was relatively scarce. However, the latest data from the Campus Computing Study of 2002 suggests that more than 75% of all students own their own desktop or notebook computer (Green, 2002). If all students have their own computers and access to the Internet, this access greatly impacts the design of communication activities and course experiences.

Course Design — Level Four

Design at the course level is usually considered to be the responsibility of the faculty member. In online and distance courses, however, the stand-alone course is the exception rather than the rule. Most online and distance courses are part of a curriculum, certificate, or degree program. This means course-level design occurs within the context of the larger program and that many of the design decisions are made in collaboration with other faculty within the academic program or department. Faculty at the course level are primarily responsible for design decisions on content, objectives, student goals, learning experiences, and assessment for a particular course. Many of these questions for this design level parallel questions at the program level design. The following questions are more specific to a single course:

Course questions:

- Where does this course fit within the context of the degree or certificate program to be offered online? Is it an early course focused on core discipline concepts, peer discussions, and standard problems or a later course focusing on applications and complex scenarios?
- What is the core set of knowledge, skills, and attitudes/values to be acquired by the students?
- What is the set of content resources required and recommended? What content resources will students use to customize the learning experience for their needs and state of knowledge and personal interests?
- Will students be a cohesive cohort group?

Design and development questions:

- What types of instructional activities and experiences will support student learning of the knowledge, skills, and attitudes of the students?
- What course management system or similar Web tool will be used for the content management? For the interaction and collaboration activities?
- What is the schedule for design and development and delivery of this course?

Faculty questions:

- What training is needed for a faculty member to transition to online teaching and learning programs?
- What tools and resources and support will be needed to support the delivery of this course?

Student questions:

- Who are the students taking this course? What are their hopes and expectations? What future courses will depend on the knowledge, skills, and attitudes acquired in this course?
- What knowledge and expertise do the students bring to the course? What is their zone of proximal development (Vygotsky, 1978)?
- What types of teaching and learning strategies best suit the students in this course? What are the life style and learning styles of the students?
- When and where will the students be likely to gather for their collaborative work? When and where will they do their more self-study activities?
- Where will the students be doing their learning? What level of network access is required or recommended?

The next two design levels are within the course parameters and generally are the responsibility of the faculty designing the course.

Level Five: Unit/Learning Activity

Many of the design questions for the unit/learning activity level and the student assessment level are derived from a design model that focuses on integrating student life style and learning styles into instructional planning (Boettcher, 2003). Examples of cognitive learning style design questions include: "How do students process information?"; "How do students respond in their minds when challenged with new concepts and rich content structures?", "What knowledge do students bring to the learning experience?"

The life style of the learner is also addressed in these questions. Life style includes all the elements in a learner's current life situation. Where will the learner be working? Will they have a personal space where they can control sound, temperature, disturbances, and network access? Will they

have to "ask" their family if they can access the network? A life-style focus encourages analysis of the where, when, with whom, and with what resources the learner is going to be doing their learning work.

Learning work consists of constructing new knowledge, applying and integrating knowledge, and solving problems with that new knowledge. Mobile, wireless technologies enable learners to study anywhere at any time. Initially, the ability to study anywhere seemed to hold the promise of solving many problems associated with access to learning. However, we have not addressed the question of just when and just where this "anytime" is likely to occur.

Learning activity questions:

- What is the knowledge, skill or attitude that is the desired outcome of this learning activity?
- What kinds of problems can students solve now? What kinds of problems do we want students to be able to solve at the conclusion of the experience?
- What instructional strategy or experience will support the learner learning the desired knowledge, skill, or attitude?
- When, where, with whom, and with what resources is the instructional activity envisioned to occur?
- What role will the teacher/mentor/faculty play during this activity? Will the teacher be present physically or at a distance, synchronously or asynchronously?
- When will a learner know that he/she knows? What feedback or result from a problem being solved will make the learner's knowledge evident?

Level Six: Assessment of Student Learning

Assessment is fundamental to the design process. Assessment planning helps to balance the goals of learning effectiveness and teaching and learning efficiency. In productive, customized, enriched learning experiences, each learner begins with an existing, personal store of knowledge representation. Learning experiences expand and enrich that knowledge base in individual and personal ways. Productive learning experiences mean that all students complete a learning experience with an expanded, yet different, store of knowledge. The goal is that learners learn core principles so that the core principles can be effectively applied, but the particular way the knowledge base is constructed in individuals is unique.

What we can design into instructional planning is that all learners share some of the same experiences and that assessment focuses on the common learning that is achieved. Assessment can also provide for demonstration of knowledge and skills in more complex environments and for some elements of customized knowledge acquisition. Here are some selected questions for assessing student learning:

Assessment Questions

- How will the learners know the goals and objectives for the learning? It is good to plan for core concepts, practice of core concepts, and customized applications of concepts.
- Will learners be generating a set of their own goals for learning? How will the faculty mentor and learner communicate and agree on goals for learning, particularly for customized applications of concepts?
- In what ways and where will the students be evaluated and graded?
- How will students demonstrate their competency in concept formation? In solving problems?
- If we don't see the students on a regular basis, can we design ways to "see" their minds virtually through conversation and experiences?

The portfolio project within the National Learning Infrastructure Initiative (NLII) is also useful as an assessment of complex learning (www.educause.edu/nlii/keythemes/eportfolios.asp).

FUTURE TRENDS

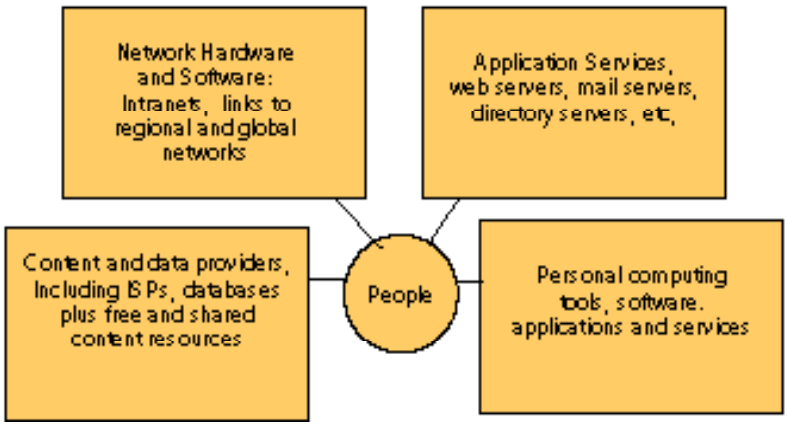
The process of instructional design is a professional task requiring knowledge of educational research, learning processes, and, increasingly, a respect for context, and a comfort level with innovation and instructional technologies. It is a labor-intensive task requiring interaction with content experts and administrators and institutional representatives. Much of the current instructional design processes focus on the analyses of needs and contexts for learning for learners, their tools and their resources — as a group. Future work in instructional design will focus more on the learner as an individual with unique knowledge structures and thus promote the needs for a rich contextual learning environment that is multi-leveled and customizable.

The design of instructional planning will become more of a priority as the demand for effective and efficient learning grows as a result of time pressures, budget pressures, and increasing demands for accountability in education.

CONCLUSION

These instructional design principles reaffirm the iterative nature of design work and the sharing of design work among the hierarchical groups of an institution. Instructional design, when done well, results in delighted and productive learners

Figure 2. Teaching and learning infrastructure - "digital plant"



and faculty pleased with their roles and their work. Consistently applied, instructional design principles keep teaching and learning focused on the who, when, where, how, and why of teaching and learning and help to ensure that the money and time invested in learning programs provide an appropriate return for individuals and for society. Instructional design is a powerful tool that moves teaching and learning into the science of learning and knowing.

REFERENCES

Boettcher, J.V. (2000). How much does it cost to put a course online? It all depends. In M.J. Finkelstein, C. Francis, F. Jewett, & B. Scholz (Eds.), *Dollars, distance, and online education: The new economics of college teaching and learning* (pp. 172-197). Phoenix, AZ: American Council on Education/Oryx Press.

Boettcher, J.V. (2003). Design levels for distance and online learning. In R. Discenza, C. Howard, & K. Schenk (Eds.), *Distance learning and university effectiveness: Changing educational paradigms for online learning.* Hershey, PA: Idea Group.

Boettcher, J.V., & Kumar, V.M.S. (2000, June). The other infrastructure: Distance education's digital plant. *Syllabus*, (13), 14-22.

Bransford, J.D., Brown, A.L., & Cocking, R.R. (1999). *How people learn. Brain, mind, experience, and school.* Washington, DC: National Academy Press. Retrieved from the World Wide Web at: *www.nap.edu/books/0309070368/html/*

Brown, D.G. (2000). Academic planning and technology. In J.V. Boettcher, M.M. Doyle, & R.W. Jensen (Eds.), *Technology-driven planning: Principles to practice* (pp. 61-68). Ann Arbor, MI: Society for College and University Planning.

Bruner, J.S. (1963). *The process of education.* New York: Vintage Books.

Business-Higher Ed Forum.(2003). *Building a nation of learners: The need for changes in teaching and learning to meet global challenges.* American Council on Education. Retrieved from the World Wide Web at: *www.acenet.edu/bookstore/pubInfo.cfm?pubID=285*

Dewey, J. (1933). *How we think* (1998 ed.). Boston, MA: Houghton-Mifflin.

Dick, W., & Carey, L. (1989). *The systemic design of instruction.* New York, Harper Collins.

Gagne, R. M. (1965). *The conditions of learning.* New York: Holt, Rinehart & Winston.

Green, K. C. (2002). *Campus computing, 2002.* Encino, CA: The Campus Computing Project. Retrieved from the World Wide Web at: *www.campuscomputing.net*

Kaufman, R. (1992). *Strategic planning plus : An organizational guide.* Thousand Oaks, CA: Sage Publications.

Knowles, M. (1998). *The adult learner: A neglected species*. Houston, TX: Gulf.

Moore, M.G., & Kearsley. G. (1996). *Distance education: A systems view.* Belmont, CA: Wadsworth.

Newell, H.A. (1996). *Sciences of the artificial.* Boston, MIT Press.

Pinker, S. (1997). *How the mind works.* New York: W.W. Norton.

Reigeluth, C.M. (1999). What is instructional-design theory and how is it changing? In C.M. Reigeluth (Ed.), *Instructional-design theories and models, volume II: A new paradigm of instructional theory* (pp. 5-29). Mahwah, NJ: Lawrence Erlbaum.

Rogers, E.M. (1995). *Diffusion of innovations.* New York, Free Press.

Rosenbloom, R.S. (1998). *Sustaining American innovation: Where will technology come from?* Forum on Harnessing Science and Technology for American's Economic Future, National Academy of Sciences Building, Washington, DC, National Academy of Science.

Schank, R.C. (1996). Goal-based scenarios: Case-based reasoning meets learning by doing. In D. Leake (Ed.), *Case-based reasoning: Experiences, lessons & future directions* (pp. 295-347). AAAI Press/The MIT Press.

Schrum, L. & Benson, A. (2002). Establishing successful online distance learning environments: Distinguishing Factors that contribute to online courses and programs. In R. Discenza, C. Howard, & K. Schenk (Eds.), *The design and management of effective distance learning programs* (pp. 190-204). Hershey, PA: Idea Group.

Sherron, G. T., & Boettcher, J. V. (1997). *Distance learning: The shift to interactivity.* CAUSE Professional Paper Series #17. Retrieved September 22, 2004 from the World Wide Web at: *www.educause.edu/asp/doclib/abstract.asp?ID=pub3017*

Vygotsky, L.S. (1962). *Thought and language.* (E. Hanfmann & G. Vakar, trans.) Cambridge: MIT Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes.* Cambridge: Harvard University Press.

KEY TERMS

Instructional Design: The process of analyzing the students, content, and intended context of an instructional program to provide detailed specifications for an instructional program or curriculum to achieve effective and efficient student learning within an affordable and accessible delivery format.

Instructional Design Theory: A "theory that offers explicit guidance on how to better help people learn and develop" (Reigeluth, 1999).

Instructional Strategy: An instructional strategy is a communication activity used to engage the learner in an educational experience and to assist the learner in acquiring the planned knowledge, skill, or attitude. Instructional strategies include lectures, discussions, reading assignments, panel presentations, study and media projects, problem analysis and solutions, field trips and assessment activities.

Learning Infrastructure: The set of physical and digital buildings, applications, services, and people that provide and support the environments for learning.

Learning Theory: A set of hypotheses or beliefs that explain the process of learning or acquiring knowledge and skill.

Online Course: A set of instructional experiences using the digital network for interaction, learning and dialogue. An online course does not require any face-to-face meetings in a physical location. Similar courses such as web-centric courses (also called hybrid or blended courses) are similar to online courses, but require regular scheduled face-to-face classes or meetings.

Zone of Proximal Development: This is a key concept in Lev Vygotsky's theory of learning. The Zone of Proximal Development (ZPD) is the "distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under the adult guidance or in collaboration with more capable peers" (Vygotsky, 1986).

ENDNOTE

Note: This article is an adaptation of the following book chapter. Boettcher, J.V. (2003). Design levels for distance and online learning. In R. Discenza, C. Howard, & K. Schenk (Eds.), *Distance learning and university effectiveness: Changing educational paradigms for online learning*. Hershey, PA: Idea Group.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 802-809, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Design Patterns from Theory to Practice

Jing Dong
University of Texas at Dallas, USA

Tu Peng
University of Texas at Dallas, USA

Yongtao Sun
American Airlines, USA

Longji Tang
FedEx Dallas Tech Center, USA

Yajing Zhao
University of Texas at Dallas , USA

INTRODUCTION

Design patterns (Gamma, Helm, Johnson, & Vlissides, 1995) extract good solutions to standard problems in a particular context. Modern software industry has widely adopted design patterns to reuse best practices and improve the quality of software systems. Each design pattern describes a generic piece of design that can be instantiated in different applications. Multiple design patterns can be integrated to solve different design problems. To precisely and unambiguously describe a design pattern, formal specification methods are used. Each design pattern presents extensible design that can evolve after the pattern is applied. While design patterns have been applied in many large systems, pattern-related information is generally not available in source code or even the design model of a software system. Recovering pattern-related information and visualizing it in design diagrams can help to understand the original design decisions and tradeoffs.

In this article, we concentrate on the issues related to design pattern instantiation, integration, formalization, evolution, visualization, and discovery. We also discuss the research work addressing these issues.

BACKGROUND

Formalization

Design patterns are typically described informally for easy understanding. However, there are several drawbacks to the informal representation of design patterns. First, informal specifications may be ambiguous and imprecise. They may not be amendable to rigorous analysis. Second, formal specifications of design patterns also form the basis for the discovery of design patterns in large software systems. Third, design patterns are generic designs that need to be instantiated and perhaps integrated with other patterns when they are applied in software system designs. There can be errors and inconsistencies in the instantiation and integration processes by using informal specifications. Finding such errors or inconsistencies early at the design level is more efficient and effective than doing it at the implementation level. In addition, it is interesting to know whether some of these processes are commutative at the design level (Dong, Peng, & Qiu, 2007b).

The initial work on the formal specification of architecture and design patterns has been done in Alencar et al. (Alencar, Cowan, & Lucena, 1996). The composition of two design patterns based on a specification language (DisCo) has been discussed in Mikkonen (1998). A formal specification approach based on logics is presented in Eden and Hirshfeld (2001). Some graphical notations are also introduced to improve the readability of the specifications. The structural and behavioral aspects of design patterns in terms of responsibilities and rewards are formally specified in Soundarajan and Hallstrom (2004). Taibi and Ngo (2003) propose specifying the structural aspect of design patterns in the first order logic (FOL) and the behavioral aspect in the temporal logic of action (TLA). Formal specification of design patterns and their composition based on the language of temporal ordering specification (LOTOS) is proposed in Saeki (2000).

Evolution

Change is a constant theme in software system development. Most design patterns describe some particular ways for future changes and evolutions. In this way, the designers can add or

remove certain design elements with minimal impact on other parts of the system. However, such evolution information of each design pattern is normally implicit in its descriptions. When changes are needed, a designer has to read between the lines of the document of a design pattern to figure out the correct ways of changing the design. Misunderstanding of a design pattern may also result in missing parts of the evolution process. It might be a disaster if a change causes any inconsistency, any violation of pattern constraints and properties, and consequently, a system crash. It is important to regularize, formalize, and automate the evolution of design patterns.

Design pattern evolutions in software development processes have been discussed in Kobayashi and Saeki (1999), where software development process is considered as the evolutions of analysis and design patterns. The evolution rules are specified in Java-like operations to change the structure of patterns. Noda (2001) consider design patterns as a concern that is separated from the application core concern. Thus, an application class may assume a role in a design pattern by weaving the design pattern concern into the application class using Hyper/J. Improving software system quality by applying design patterns in existing systems has been discussed in Cinnéide and Nixon (2001). When the user selects a design pattern to be applied in a chosen location of a system, automated application is supported by applying transformations corresponding to the minipatterns.

Visualization

When a design pattern is applied in a large system design, pattern-related information is normally lost because the information on the role a model element plays in the pattern is often not available. It is unclear which model elements, such as class, attribute, or operation, participate in the pattern. There are several problems when design patterns are implicit in software system designs. First, software developers can only communicate at the class level instead of the pattern level because they do not have pattern-related information in system designs. Second, each pattern often documents some ways for future evolutions, as discussed previously, that are buried in the system design. The designers are not able to change the design using relevant pattern-related information. Third, each pattern may preserve some properties and constraints. It is hard for the designers to check whether these properties and constraints hold when the design is changed. Fourth, it may require considerable efforts on reverse-engineering design patterns from software systems.

Early work on explicitly visualizing design patterns in UML has been investigated in Vlissides (1998), where all approaches surveyed can only represent the role a class plays in a pattern, not the roles of an attribute (or operation). They cannot distinguish multi-instances of a pattern either. Current approaches on visualizing design patterns can be categorized into two kinds, UML-based approaches (France, Kim, Ghosh, & Song, 2004; Lauder & Kent, 1998; Vlissides, 1998) and non-UML-based approaches (Mapdlsden, Hosking, & Grundy, 2002; Reiss, 2000). The UML-based approaches can be further divided into single-diagram (Vlissides, 1998) and multidiagram (France et al., 2004; Lauder & Kent, 1998).

Discovery

Design document is often missing in many legacy systems. Even the document is available; it may not exactly match the source code that may be changed and migrated over time. Missing pattern-related information may compromise the benefits of using design patterns. The applications of design patterns may vary in different layouts, which also pose challenges for recovering and changing these design pattern instances. It is important to effectively and efficiently recover the design pattern from the source code.

Several approaches have been proposed to discover a design pattern from either source code or design model diagrams, such as the UML. A review of these approaches has been presented in Dong (Dong, Zhao, & Peng, 2007d). Among them, Antoniol (2004) uses the abstract object language (AOL) as the intermediate representation for pattern discovery. Tsantalis et al. (Tsantalis, Chatzigeorgiou, Stephanides, & Halkidis, 2006) applies a graph matching algorithm to calculate the similarity of two classes in pattern and system. Machine learning algorithms, such as decision tree and neural network, have been applied to classify the potential pattern candidates in (Ferenc, Beszedes, Fulop, & Lele, 2005, Gueheneuc, Sahraoui, & Zaidi, 2004).

FROM THEORY TO PRACTICE

In this section, we present our approaches on the formalization, evolution, visualization, and discovery of design patterns. In addition to the theory of our approaches, we provide several tools for practical uses of our approaches.

Formalization

Over the past decade, we have applied several formal methods, such as first-order logic, temporal logic of action (TLA) (Lamport, 1994), Prolog, Calculus for Communicating System (CCS) (Milner, 1989), to specify design pattern structure and behavior. More specifically, we applied first-order logic to specify the structural aspect of a design pattern and the TLA to specify the behavior of each design pattern in Dong (Dong, Alencar, & Cowan, 2000). The structural aspect is described by predicates for describing classes, state variables, methods, and their relations. The

integration of two design patterns is the union of two sets of the predicates corresponding to the structures of the two patterns. We specify the behavioral aspect using the TLA since it is an axiomatic style of semantic definition suitable for describing both safety and fairness properties. We define the behavioral semantic of each design pattern in terms of TLA formulas. In addition to TLA, we used CCS to specify the behavioral aspect of design patterns in Dong (Dong, Alencar, & Cowan, 2006a), where we define the behavior of each pattern in terms of CCS processes. The objects and their communications in each pattern are represented in the processes and their communications. We define the interface, input/output messages, and actions of each process. We then defined the behavioral instantiation and integration based on the process definitions.

Formal specification allows describing the structural and behavioral aspects of design patterns more precise and concise. It also facilitates automated verification techniques, including model checking and theorem proving.

Model checking techniques typically include a model specification language that specifies a finite state model and a property specification language that defines the properties of a system in, for example, temporal logic. A model checker is a tool that explores the finite state model to match the properties. We have explored model checking techniques in Dong et al. (2006a). By specifying the properties of each pattern, we used a model checker to check them against the behavioral specification of the pattern. In this way, we can check the consistencies of the integration of design patterns and discover errors early at the design level.

Theorem proving is another verification technique that ensures the correctness of a design. By applying rigorous mathematical knowledge, formal model abstracts the structure and behavior of a design pattern, and the operations between them, which enables us to summarize, predict, prove, or exclude certain general properties of design pattern operations. Based on our definitions of the structural and behavioral aspects of deign patterns, we have proved several theorems related to the structural integrity, safety, and liveness properties in Dong et al. (2000). While manual proving theorems can be tedious and error-prone, we also explore the application of Prolog for deducing facts from a formal specification of patterns in Alencar et al. (Alencar, Cowan, Dong, & Lucena, 1999).

While each design pattern needs to be instantiated when it is applied in a system, it may also be integrated with other patterns to solve multiple design problems. It is interesting to know whether the instantiation and integration processes are commutative. Proving the conditions that these processes are commutative is important because it can save a lot of time of the designers on trial-and-error. In this way, we are able to predict the possible outcome of a design and to disclose potential problems. The commutability problem has been explored systematically under our formal model in Dong et al. (2007b).

Evolution

The evolution process of design patterns has been initially studied in Alencar et al. (1999), where Prolog is used to capture the structural evolution processes of design patterns. The structural aspect of a design pattern is described in terms of Prolog facts. Thus, the evolution of a design pattern application can be achieved by the addition or removal of new or old Prolog facts.

While there are many different ways to evolve design patterns, we classified them into two-level transformations, the primitive level and pattern level, in Dong et al. (Dong, Yang, Lad, & Sun, 2006b). The primitive level transformations include the addition/removal of an object-oriented modeling element, such as class, attribute, operation, association, generalization, aggregation, composition, realization, dependency. The pattern level transformations are a group of primitive level transformations that reappear in many design patterns. We categorized five pattern-level transformations:

1) simple addition/removal of an independent class and the corresponding relationships between this class and the classes in the original pattern;
2) addition/removal of one independent class with attributes and/or operations and the corresponding relationships between this class and the classes in the original pattern;
3) addition/removal of an attribute/operation in several classes consistently;
4) addition/removal of a group of correlated classes;
5) addition/removal of a group of classes and some attributes or operations in the classes involved in the original pattern instance.

With this classification of the evolution processes of design patterns, we are able to automate these evolutions with tool support. We used XMI to describe our two-level evolutions. Using an XMI file processor, design pattern evolutions can be automated by transforming from the original UML model of a design pattern to the destination UML model of the pattern. In particular, we explored two main model transformation techniques, XSLT and QVT, to automate the evolution processes.

By semiautomating the evolution process, our tool provides the following features: first, analyzing the legacy code and presenting in a visible manner the pattern-related system pieces that can be evolved and the possible evolutions for each system piece; second, when certain evolution is selected, input fields for the required information are

displayed to ensure no missing information and the integrity for the evolution; and third, ensuring the consistency by facilitating the reasoning.

Visualization

Our research of design pattern visualization is a UML-based approach (Dong, Yang, & Zhang, 2007c). We have extended the UML with a new profile for design pattern that defines new stereotypes (PatternClass, PatternAttribute, and PatternOperation) for tracking design patterns in UML diagrams. Each stereotype may be attached by a tagged value: role@name[instance]. The pattern-related constraints for stereotypes are defined based on OCL. These new stereotypes and tagged values are attached to a modeling element to explicitly represent the role the modeling element plays in a design pattern so that the user can identify the pattern in a UML diagram. Based on this profile, we also develop a Web service (tool), called VisDP, for explicitly visualizing design patterns in UML diagrams based on coloring and mouse movement. In this way, our tool can hide all pattern-related information and allows the user to visualize design patterns on demand. All pattern-related information is displayed only when requested.

Discovery

We propose a novel approach based on matrix and weight to discover design patterns from source code (Dong, Lad, & Zhao, 2007a). In particular, the system structure is represented in a matrix with the columns and rows to be all classes in the system. The value of each cell represents the relationships among the classes. For each specific relationship, there is a unique prime number associate with it. For example, the "Generalization" relationship can be prime number 2, and "Aggregation" can be prime number 3. The cell value for any two classes is the multiplication of each associated prime to the power of occurrence of the relationship. If two classes have both the "Generalization" and "Aggregation" relationships, the cell value is $2^1 \times 3^1 = 6$.

After we get the matrix representation of the system, we can apply direct matching method to find the pattern instance. The structure of each design pattern is also represented in another matrix. The discovery of design patterns from source code becomes matching between the two matrices. The direct matching results of the structural analysis may include false positive instances due to missing behavioral analysis. Our approach may proceed to check back the XML files or directly into the source code for behavior characteristics. Some patterns may be hard to distinguish since they have the same structural and behavioral characteristics. In such case, our approach also analyzes the semantic information of the pattern instances, like naming convention. We automated the structural, behavioral, and semantic analyses in our DP-Miner tool (Dong et al., 2007a).

We also apply template matching method to pattern discovery (Dong, Sun, & Zhao, 2008). The basic idea is to calculate the normalized cross-correlation value of two vectors f and g. Normalized cross correlation defines the $cos\theta$ value, where θ is the angle between vector f and g. The maximum value is 1 when f and g is an exact match, that is, $\theta = 0$. A two-dimensional matrix can be flatted into one-dimensional vector by appending the following rows values to the first row. If two matrixes are similar to each other, we are expecting to see a small angel between the two flatted vectors, in other words, a high normalized cross-correlation value. Our template matching approach encodes both pattern and system knowledge into two overall matrixes, and calculates their similarity score by cross correlation. We also implement a Web-based tool to facilitate the template matching for pattern recovery. The advantage of template matching approach is that it cannot only find the exact matches of pattern instances, but also identify their possible variants.

FUTURE TRENDS

Our future direction on formalization is to use our formal model and its derivation methodology in security-related applications, where the correctness of pattern operations is vitally important to the completeness of a successful system. We are also interested in combining the use of formal derivation together with model checking, to grasp a deep understanding of system properties, and gain more confidence on the system correctness.

We will characterize the constraints of evolutions of each design pattern, and provide techniques and tools for checking such constraints after evolutions. In addition, we are investigating the model transformation techniques based on other techniques to improve our current approach.

Our research on visualization can be extended from design patterns to architectures. We are interested in architecture visualization that can explicitly display critical architecture information in large software architecture.

Applying machine learning and data-mining methods are future trends for pattern discovery. In addition, we will continue to further optimize our tool for better performance. Furthermore, pattern discovery techniques may be also extended to architectural pattern discovery.

CONCLUSION

In this chapter, we present the research issues related to design patterns. These issues range from theory to practice,

including formalization, evolution, visualization, and discovery of design patterns. We discussed the research problems related to these issues and describe the existing solutions to these problems. In addition, we introduced our formal and automated engineering solutions to these problems. Future research directions are also pointed out. Design patterns are good designs that are at the heart of software development. Research work on software design is critically important to the success of software systems.

REFERENCES

Alencar, P. S. C., Cowan, D. D., Dong, J., & Lucena, C. J. P. (1999). A pattern-based approach to structural design composition. In *Proceedings of the IEEE 23rd Annual International Computer Software & Applications Conference* (pp. 160-165).

Alencar, P., Cowan, D. D., & Lucena, C. J. P. (1996). A formal approach to architectural design patterns. In *Proceedings of the Third International Symposium of Formal Methods Europe* (pp. 576–594).

Antoniol, G., Fiutem, R., & Cristoforetti, L. (1998). Design pattern recovery in object-oriented software. In *Proceedings of the 6th IEEE International Workshop on Program Understanding* (pp. 153-160).

Cinnéide, M. Ó., & Nixon, P. (2001). Automated software evolution towards design patterns. In *Proceedings of the International Workshop on the Principles of Software Evolution* (pp. 162-165).

Dong, J., Alencar, P. S. C., & Cowan, D. D. (2000). Ensuring structure and behavior correctness in design composition. In *Proceedings of the 7th Annual IEEE International Conference and Workshop on Engineering of Computer Based Systems* (pp. 279-287).

Dong, J., Alencar, P. S. C., & Cowan, D. D. (2006a). Automating the analysis of design component contracts. *Software – Practice and Experience, 36*(1), 27-71.

Dong, J., Lad, D. S., & Zhao, Y. (2007a). DP-Miner: Design pattern discovery using matrix. In *Proceedings of the Fourteenth Annual IEEE International Conference on Engineering of Computer Based Systems* (pp. 371-380).

Dong, J., Peng, T., & Qiu, Z. (2007b). Commutability of design pattern instantiation and integration. In *Proceedings of the First IEEE & IFIP International Symposium on Theoretical Aspects of Software Engineering* (pp. 283-292).

Dong, J., Sun, Y., & Zhao, Y. (2008). Design pattern detection by template matching. In *Proceedings of the 23rd Annual ACM Symposium on Applied Computing.*

Dong, J., Yang, S., Lad, D. S., & Sun, Y. (2006b). Service oriented evolutions and analyses of design patterns. In *Proceedings of the Second IEEE International Symposium on Service-Oriented System Engineering* (pp. 11-18).

Dong, J., Yang, S., & Zhang, K. (2007c). Visualizing design patterns in their applications and compositions. *IEEE Transaction on Software Engineering, 33*(7), 433-453.

Dong, J., Zhao, Y., & Peng, T. (2007d). Architecture and design pattern discovery techniques – A review. In *Proceedings of International Conference on Software Engineering Research and Practice* (pp. 621-627).

Eden, A. H., & Hirshfeld, Y. (2001). Principles in formal specification of object-oriented architectures. In *Proceedings of the 11th CASCON.* IBM Press.

Ferenc, R., Beszedes, A., Fulop, L., & Lele, J. (2005). Design pattern mining enhanced by machine In *Learning, 21st IEEE International Conference on Software Maintenance* (pp. 295-304).

France, R. B., Kim, D., Ghosh, S., & Song, E. (2004). A UML-based pattern specification technique. *IEEE Transactions on Software Engineering, 30*(3), 193-260.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software.* Addison-Wesley.

Gueheneuc, Y.-G., Sahraoui, H., & Zaidi, F. (2004). Fingerprinting design patterns. In *Proc. 11th Working Conf. on Reverse Eng. (WCRE'04)* (pp. 172-181).

Kobayashi, T., & Saeki, M. (1999). Software development based on software pattern evolution. In *Proceedings of the Sixth Asia-Pacific Software Engineering Conference* (pp. 18-25).

Lamport, L. (1994). The temporal logic of actions. *ACM Transactions on Programming Languages and Systems, 16*(3), 873–923.

Lauder, A., & Kent, S. (1998). Precise visual specification of design patterns. In *Proceeding of Third International Conference on Object-Oriented Programming* (pp. 114-143).

Mapdlsden, D., Hosking, J., & Grundy, J. (2002). Design pattern modeling and instantiation using DPML. In *Proceedings of the 40th International Conference of Object-Oriented Languages and Systems (TOOLS Pacific)* (pp. 3-11).

Meyer, B. (1992). Applying "design by contract". *IEEE Computer, 25*(10), 40–51.

Mikkonen, T. (1998). Formalizing design pattern. In *Proceedings of the 20th International Conference on Software Engineering* (pp. 115–124).

Milner, R. (1989). Communication and concurrency. *International Series in Computer Science*. Prentice Hall.

Natsuko, N., & Tomoji, K. (2001). Design pattern concerns for software evolution. In *Proceedings of the 4th International Workshop on Principles of Software Evolution* (pp. 158-161).

Reiss, S. P. (2000). Working with patterns and codes. In *Proceedings of the 33rd Hawaii International Conference on System Sciences* (pp. 8054-8054).

Saeki, M. (2000). Behavioral specification of GoF design patterns with LOTOS. In *Proceedings of the Seventh Asia-Pacific Software Engineering Conference* (pp. 408–415).

Soundarajan, N., & Hallstrom, J. O. (2004). Responsibilities and rewards: Specifying design patterns. In *Proceedings of the 26th International Conference on Software Engineering* (pp. 666–675).

Taibi, T., & Ngo, D. (2003). Formal specification of design pattern combination using BPSL. *Information and Software Technology, 45*(3), 157–170.

Tsantalis, N., Chatzigeorgiou, A., Stephanides, G., & Halkidis, S. (2006). Design pattern detection using similarity scoring. *IEEE transaction on software engineering, 32*(11), 896-909.

Vlissides, J. (1998). Notation, notation, notation. *C++ Report*, 1-6.

KEY TERMS

Design Patterns: Design patterns represent solutions to problems that arise when developing software within a particular context. Design patterns capture the static and dynamic structure and collaboration among key participants in software designs.

Design patterns are generic design pieces that need to be instantiated before uses. The instantiation of a design pattern describes the process of applying generic design pieces into a system design.

The integration of design patterns describes the process of composing multiple design patterns to solve a number of design problems. Design patterns can be integrated by overlapping common parts from different patterns or adding new relationships between parts from different patterns.

The formalization of design patterns is to apply rigorous methods to specify design patterns or to verify their properties. These formal methods include logic-based and process-based methods.

The evolution of a design pattern is a process to add or remove design elements to/from existing design pattern applications in a software system. It takes place when new requirements, platforms, technologies, or environments change and therefore software system need to be adapted to such change.

The visualization of design pattern provides techniques and tools for explicitly visualizing the instances of design patterns applied in a large software system design. These visualization techniques and tools can help software designers for tracing, identifying, and checking design patterns in the software system design, and making right design decision of applying design patterns.

Design pattern discovery techniques are used to recover design pattern instances applied in existing source code. It becomes a key issue for many research areas, such as reverse engineering and code refractory, because it helps for program comprehension and design visualization.

Designing Agents with Negotiation Capabilities

Jana Polgar
Monash University, Australia

SOFTWARE AGENTS TODAY

Agents are viewed as the next significant software abstraction, and it is expected they will become as ubiquitous as graphical user interfaces are today. Agents are specialized programs designed to provide services to their users. Multiagent systems have a key capability to reallocate tasks among the members, which may result in significant savings and improvements in many domains, such as resource allocation, scheduling, e-commerce, and so forth. In the near future, agents will roam the Internet, selling and buying information and services. These agents will evolve from their present day form - simple carriers of transactions - to efficient decision makers. It is envisaged that the decision-making processes and interactions between agents will be very fast (Kephart, 1998).

The importance of *automated negotiation systems* is increasing with the emergence of new technologies supporting faster *reasoning engines* and mobile code. A central part of agent systems is a sophisticated reasoning engine that enables the agents to reallocate their tasks, optimize outcomes, and negotiate with other agents. The *negotiation strategy* used by the reasoning engine also requires high-level inter-agent communication protocols, and suitable collaboration strategies. Both of these sub-systems – a *reasoning engine* and a *negotiation strategy* - typically result in complicated agent designs and implementations that are difficult to maintain.

Activities of a set of *autonomous agents have to be coordinated*. Some could be mobile agents, while others are static intelligent agents. We usually aim at decentralized coordination, which produces the desired outcomes with minimal communication. Many different types of *contract protocols* (cluster, swaps, and multiagent, as examples) and *negotiation strategies* are used. The evaluation of outcomes is often based on marginal cost (Sandholm, 1993) or game theory payoffs (Mass-Colell, 1995). Agents based on constraint technology use complex search algorithms to solve optimization problems arising from the agents' interaction. In particular, coordination and negotiation strategies in the presence of incomplete knowledge are good candidates for constraint-based implementations.

SELECTED NEGOTIATION AND REASONING TECHNIQUES

Negotiation space is determined by two components: *negotiation protocol* and *negotiation strategy*. The *negotiation protocol* defines the rules of behavior between the participants in terms of interactions, deals, bidding rules, temporal constraints and offers, as components of the protocol. Two agents must first agree on the negotiation protocol before any interaction starts.

The *negotiation strategy* is a specification of the sequence of actions the agent intends to make during the negotiation. Strategies should be compatible with the negotiation protocol. The focus of any negotiation strategy is to maximize outcomes within the rational boundaries of the environment. The classification of negotiation strategies is not an easy task since the negotiation strategy can be realized by any algorithm capable of evaluating outcomes, computing appropriate actions, and following the information exchange protocol.

The *negotiation mechanism* is the actual implementation of negotiation strategy and negotiation protocol. This field is evolving fast, with emergence of new agent platforms, wireless encounters and extended mobility.

Negotiation is a search process. The participants jointly search a multi-dimensional space (e.g., quantity, price, and delivery) in an attempt to find a single point in the space at which they reach mutual agreement and meet their objectives. The *market mechanism* is used for many-to-many coupling or interactions between participants. *Auctions* are more appropriate for one-to-many negotiations. The market mechanism often suffers from inability to efficiently scale down (Osborne, 1990) to smaller numbers of participants. On the other hand, one-to-many interactions are influenced by strategic considerations and involve integrative bargaining, where agents search for *Paretto efficient* agreements (tradeoffs).

NEGOTIATION STRATEGIES

Analytical Approach (Game Theory)

The principles of bargaining and negotiation strategies in multiagent systems have attracted economists. Early foundations and mathematical models were investigated by Nash (1950), and the field is still very active. The *game theory* is a collection of analytical tools designed to understand and describe bargaining and interaction between decision makers. Game theory uses mathematical models to formally express real-life strategies (Fudenberg, 1991; Osborne, 1994).

The high-level abstraction allows the model to be applied to a variety of situations. The model places no restrictions on the set of actions available to the player. With regard to mathematical models, there already exist many sophisticated and elaborated strategies for specific negotiation problems. The Contract Net Protocol (CNP) (Sandholm, 1993; Smith, 1980) represents the model of decentralized task allocation where agents locally calculate their marginal costs for performing sets of tasks. The pricing mechanism in Sandholm (1993) generalizes the CNP to work for both cooperative and competitive agents. In Zeng (1996), bilateral negotiation based on the Bayesian method is presented. It demonstrates the static nature of the model. The learning effect is achieved by using dynamic updates of a knowledge base, which is consulted during the negotiation process.

Most of the studies assume perfect rationality (flawless deduction, marginal costs are computed exactly, immediately and without computational cost), and the infinite horizon of strategic bargaining. These are not realistic assumptions. More advanced studies deal with coalition formation and negotiation strategies in the environment of multiple self-interested or cooperative agents with bounded rationality (Sandholm, 1993) and bargaining with deadlines.

Analytical approach has the advantage of stable and reliable behavior. The main disadvantage is the static nature of the model, resulting in potential predictability of the outcomes. The other problems are associated with the notion of perfect rationality.

Contracts in automated negotiations consisting of self-interested agents are typically designed as binding (impossible to breach). In cooperative distributed problem solving, commitments are often allowed to be broken based on some local reasoning. Frequently, the protocols use continuous levels of commitment based on a monetary penalty method (Sandholm, 1993). Unfortunately, the inflexible nature of these protocols restricts an agent's actions when the situation becomes unfavorable. The models that incorporate the possibility of decommitting from a contract with or without reprisals (Sen, 1994; Smith, 1980) can accommodate some changes in the environment and improve an agent's status. However, all of these protocols are somewhat restricting with respect to evolving, dynamic situations.

Evolutionary Strategies

With *evolutionary strategies*, the data used as the basis for negotiation, as well as the algorithm operating on the data, evolve. This approach provides more efficient learning, supports the dynamics of the environment, and is adaptable. However, only a few implementations have been attempted, and these have been of only simple negotiation strategies (Aridor, 1998). *Genetic algorithms* are probably the most common techniques inspired by evolution, in particular by the concepts of natural selection and variation. The basic genetic algorithm is derived from the hypothesis that the candidate solutions to the problem are encoded into "chromosomes". Chromosomes represent a solution or instance of the problem hand encoded into a binary string. The algorithm then operates on this binary string. It begins with a randomly generated set of candidate solutions. The set of candidate solutions is generated as a random string of ones and zeroes. Each chromosome is evaluated and the fitness of the chromosome could be the value of the objective function (or the utility if we want to maximize the outcome). A new population is created by selecting individuals to become parents. A thorough description of the genetic algorithm approach can be found in Goldberg (1989).

A very large amount of research has been carried out in the application of evolutionary algorithms to situations that require decisions. Examples include coalition games, exchange economies, and double auctions. This approach was inspired by the concept of variation and natural selection. The intelligent agents are modeled using classifier systems to select decisions. Although the recent research shows that multiagent systems of classifiers are capable of learning how to play *Nash-Markov equilibrium*, the current limitations of computational resources and the instability of "home-grown" implementations significantly constrain the nature of the strategies. The important question is what design and implementation techniques should be used to ease this conflict and to provide the resources required for genetic learning to operate in an unrestricted way. It is believed that the ability of agents to learn simple games would be beneficial to electronic commerce.

Constraint Agents

The potential of constraint-based agents is still to be fully realized and appreciated. One of the possible frameworks for constraint-based agents is outlined in Nareyek (1998). This framework considers agents as a means for simplifying distributed problem solving. An agent's behavior and the quality of solutions depend on the underlying action-task planning system. The recent results with some constraint planners and constraint satisfaction problems (CSP) indicate the potential advantages of this approach.

Agents operating with only partial knowledge of the surrounding environment are prime candidates for the implementation of reasoning using constraint solvers. Agents can share and exchange knowledge in a distributed fashion. The suggestion-correction process used in negotiation strategies corresponds to mapping constraint satisfaction problems to search algorithms and heuristics (Tsang, 1993). The planning and scheduling systems can be treated as CSPs that allow the use of constraint solvers to support the planning activity of an agent. However, similar to analytical tools, constraint solvers represent static models of CSP. In order to achieve flexibility in the decision-making process, the solver has to be able to incorporate an agent's essential characteristics:

Reactive behavior is characterized by the agent's ability to absorb new information and restrict or relax its actions. When constraint solvers face the relaxation of constraints, they recompute the entire problem. The relaxation in constraint planners and schedulers nearly always results in reduced stability and possibly reduced quality of solutions. Despite the recent implementation of constraint solvers in graphical applications (Borning, 1995; Sadeh, 1995), with real-time computation and stability, the constraint relaxation and adaptive behavior still represent some difficulties for practical applications.

An agent's *rational behavior* and fast reaction to the changes in its environment is difficult to support with constraint solvers. Some scheduling systems incorporate this idea and extend or replace deliberative planning with behavior rules (ILOG, 1996). The majority of existing constraint solvers compute the search for solutions off-line. An approach to eliminate this problem is the development of an anytime algorithm (Zilberstein, 1995) and constraint hierarchies (Zanden, 1996). Some CSPs are suitable for iterative local search techniques, such as annealing, taboo search or genetic algorithms (Kautz, 1996). Additional information on suitability can be found in Tsang (1993).

Representation of time in traditional planning systems is based on Allen's model of temporal relations (Allen, 1985). This representation does not seem to be suitable for multi-agent interaction with complex temporal relationships and commitment protocols. Problems arise with respect to concurrency and events that last over a period of time.

Social abilities mean interaction, negotiation, coordination, and/or cooperation among agents. A language or interaction protocol typically supports these social skills. The ultimate goal of cooperation and coordination is to reach a globally optimal solution independent of the language or protocol used. If we map the cooperation goals into distributed problem-solving strategies and let each agent play the role of a cooperating computational unit instead of an autonomous negotiator, it is then possible to deploy distributed constraint satisfaction problem-solving strategies.

Multiagent solutions to the distributed constraint problems require new types of design frameworks based on replaceable components that can accommodate the local autonomy of agents, several negotiation strategies, cooperative problem solving, online search for solution, and support for an agent's rational behavior.

Agent design must accommodate static agents as well as trends in mobile agents. The lightweight architecture required for mobile agents, and an efficient and flexible negotiation strategy, are not mutually exclusive. The main issue is to provide a framework that guarantees that the negotiating agent is not overloaded with the complex intelligence that may not be used. We see an agent as a lightweight core with the ability to "borrow" the intelligence it requires from the hosting environment.

The major problem with this component-based framework in which one or more constraint solvers are used as plug-in components arises from the difficulties associated with the description of the constraint satisfaction problem. The representation of the problem can be either under-constrained or over-constrained. In either case, the agent is dealing with incomplete knowledge and it cannot determine which of the generated solutions is optimal and suitable to use in further negotiation.

One of the most important criteria used to judge constraint solvers is their performance. In order to provide reactive and pro-active behavior, it is important for the solver to generate the solution quickly enough to maintain responsiveness. This becomes more difficult to achieve as the number of constraints and variables becomes larger.

A pre-built constraint solver may be able to maintain constraints efficiently. However, declarative definitions of constraints are difficult to use in high-level agent building tools.

There has been very little work in incorporating constraint solvers into reasoning and negotiation strategies of agent systems. The slow acceptance of constraint solvers seems to have been caused by four reasons:

1. Many constraint solvers only support a limited range of constraints.
2. It is difficult to understand and control constraint solvers. When the programmer sets up a set of constraints, it may not be obvious how the constraint solver will maintain them.
3. If the search space is large then online performance of constraint solvers may not be satisfactory for some application domains (e.g., Eaton, 1997).
4. Multi-agent systems are represented by highly distributed search space (distributed constraint optimization problems (DCOP)).

DCOP have been considered an important subject of research for multi-agent systems and multi-agent cooperative mediation (Mailler & Lesser, 2004). DCOP aim at finding optimal assignment to a set of variables dispersed over a

number of agents that have some interdependencies. So far, the descriptions of the problem proposed include distributed partial constraint satisfaction problem (Hirayama & Yokoo, 1997), distributed valued constraint satisfaction problem (Lemaitre & Verfaillie, 1997), and DCOP (Yokoo & Durfee, 1991). Typical problem with these asynchronous protocols is the exponential increase in communication. The protocol suggested in Mailler and Lesser (2004) allows the agents to overlap the context used for their local decision making. During the mediation session, the mediator agent computes a partial solution and recommends the value assignments to other agents.

CONCLUSION AND FUTURE TRENDS

Agents are fairly complex and ambitious software systems. They will be entrusted with advanced and critical applications, such as network administration, database and application monitoring, workflow/business process support, and enterprise integration. As such, agent-based systems must be engineered with valid software engineering principles and not constructed in an ad hoc fashion.

Analytical strategies are tools based on a static mathematical model to evaluate outcomes and generate appropriate action. With *evolutionary* or *genetic approaches*, the learning process is more effective and models are adaptable. The advances in *constraint technology* enable the design and implementation of planning and scheduling tasks to be treated as constraint satisfaction problems. The agent concept then can be used to support dynamic adaptation and collaboration of local or distributed problem-solving modules.

Design and implementation of any reasoning strategies and collaboration protocols can lead to complex systems, and these issues can also lead to computer code that is difficult to maintain. The protocols of interaction are complicated and they typically have many levels or states. If a reasoning engine is very large and complex, it restricts and slows an agent's mobility, rationality, and real-time responsiveness. In recent years, patterns of agent collaboration and negotiation, and some useful components and architectures have emerged. Reusable components and generative patterns can greatly improve the quality of agent designs and implementations.

Any agent, and mobile agents in particular, should have the capability to adapt their negotiation strategy and protocol according to the tasks at hand. In order to match tasks with the appropriate negotiation and collaboration strategies, the agent should have the ability and the means to select a negotiation strategy as a component and then "plug it in" for use.

Nwana (1998) discusses the contradiction between the research in agent coordination and reasoning, and the reality of implementing real applications. He suggests that we should continue with "borrowing and consolidation" using already established AI work. This evolutionary path requires adaptable design techniques to support the trends in AI. Instead of building specific negotiation strategies and protocol cores for each problem domain, we build agents with robust, adaptable core capabilities. The agents' negotiation strategy and protocols are then components or pattern-based building blocks that are "borrowed" and matched with the task at hand.

Future trends in the design of negotiating agents will undoubtedly track those in software engineering in general. Interesting and valuable future developments are expected to appear in the areas of separation of concerns and aspect-oriented programming (AOP). Object-oriented techniques, components and patterns have revolutionized software engineering. However, as software applications get more ambitious and complicated, object technology reaches its limits, especially due to its inability to address behavior and functionality that crosscuts an object hierarchy. Research completed by Kendall (1999) has considered how the separation of concerns and AOP can be used to improve the quality of agent designs. Role modeling is another key area for future research, and this area is also detailed in Kendall (1999).

REFERENCES

Allen, J.F., & Hayes, P.J. (1985). A common-sense theory of time. *Proceedings International Joint Conference on Artificial Intelligence* (pp. 528-531).

Aridor, Y., & Lange, D. (1998). Agent design patterns: Elements of agent application design. *Proceedings of the 2nd International Conference of Autonomous Agents* (on CD). IEEE.

Borning, A. (1995). The OTI constraint solver: A constraint library for constructing interactive graphical user interfaces. *Proceedings of the 1st International Conference on Principles and Practice of Constraint Programming* (pp. 624-628).

Eaton, P.S., Freuder, E.C., & Wallace, R.J. (1997). *Constraint-based agents: Assistance, cooperation, compromise.* Technical Report. Computer Science Department, University of New Hampshire.

Fudenberg, D., & Tirole, J. (1991). *Game theory.* Cambridge: MIT Press.

Goldberg, D.E. (1989). *Genetic algorithms in search, optimisation and machine learning.* Reading, MA: Addison-Wesley.

Hirayama, K., & Yokoo, M. (1997). Distributed partial constraint satisfaction problem. In G. Smolka (Ed.), *Principles and practice of constraint programming (CP-97), Lecture*

Notes in Computer Science, 1330, 222-236. Springer-Verlag.

ILOG. (1996). *ILOG SCHEDULER user's manual.*

Kautz, H., & Selman, B. (1996). Pushing the envelope: Planning, propositional logic, and stochastic search. *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI-96),* 1194-1201.

Kendall, E. (1999). Role model designs and implementations with aspect oriented programming. *Proceedings of the 1999 Conference on Object- Oriented Programming Systems, Languages, and Applications (OOPSLA'99)* (pp. 132-145). ACM Press.

Kephart, J.O., Hanson, J.E., Levine, D.W., Grosof, B.N., Sairamesh, J., Segal, R.B., & White, S.R. (1998, July 4-7). Dynamics of an information-filtering economy. *Proceedings of 2nd International Workshop on Cooperative Information Agents (CIA-98)*, Paris. Retrieved October 6, 2003, from *http://citeseer.ist.psu.edu/70606.html*

Lemaitre, M., & Verfaillie. (1997). An incomplete method for solving distributed valued constraint satisfaction problems. *Proceedings of the AAAI Workshop on Constraint and Agents* (on CD).

Mailler, R., & Lesser, V. (2004). Solving distributed constraint optimization problems using cooperative mediation. Retrieved October 4, 2003, from *ftp://mas.cs.umas.edu/pub/mailler-569.pdf*

Mass-Colell, A., Whinston, R., & Green, J.R. (1995). *Microeconomic theory.* Oxford University Press.

Nareyek, A. (1998). *Constraint-based agents.* Technical Report. German National Research Center for Information Technology, Berlin.

Nash, J. (1950). The bargaining problem. *Econometrica, 18,* 155-162.

Nwana, H.S., & Ndumu, D.T. (1998). A perspective on software agents research. *ZEUS Methodology Documentation.* British Telecom Laboratories.

Osborne, M.J., & Rubinstein, A. (1990). *Bargaining and markets.* Academic Press.

Osborne, M.J., & Rubinstein, A. (1994). *A course in game theory.* Cambridge: The MIT Press.

Sadeh, N.M., & Fox, M.S. (1995). Variable and value ordering heuristics for the job shop scheduling constraint satisfaction problem. *Technical Report CMU-RI-TR-95-39.* Carnegie Melon University.

Sandholm, T. (1993). An implementation of the contract net protocol based on marginal cost calculations. *Proceedings of the 12th International Workshop on Distributed Artificial Intelligenc*e (pp.256-262). Retrieved March 4, 2000, from *http://citeseer.ist.psu.edu/sandholm93 implementation.html*

Sen, S., & Durfee, E. (1994). The role of commitment in cooperative negotiation. *International Journal of Intelligent Cooperative Information Systems, 3*(1), 67-81.

Smith, R.G. (1980). The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on Computers, C-29*(12),1104-1113.

Tsang, E., & Borrett, P.K. (1993). *Foundation of constraint satisfaction.* Academic Press.

Yokoo, M., & Durfee, E.H. (1991). Distributed constraint optimization as formal model of partially adversarial cooperation. *Technical Report CSE-TR-101-91.* University of Michigan, Ann Arbor.

Zanden, B. (1996, January). An incremental algorithm for satisfying hierarchies of multi-way dataflow constraints. *ACM Transactions on Programming Languages and Systems, 18*(1), 30-72.

Zeng, D., & Sycara, K. (1996, August). How can an agent learn to negotiate? In J.P. Muller, M.J. Wooldgridge & N.R. Jennings (Eds.), *Intelligent agents III: Agent theories, architectures, and languages. Proceedings of European Conference on Artifical Intelligence'96.* LNCS. Springer.

Zilberstein, S., & Russell, S. (1995). Approximate reasoning using anytime algorithms. In S. Natarajan (Ed.), *Imprecise and approximate computation.* Kluwer Academic Publishers.

KEY TERMS

Agent: A program designed to provide specialized and well-defined services. An agent can be *static* – executing on the computer where it was installed, or *mobile* – executing on computer nodes in a network.

Bayesian Method: Means of quantifying uncertainty, based on the probability theory. The method defines a rule for refining a hypothesis by factoring in additional evidence and background information. It uses results of previous events to predict results of future events.

Evolutionary Game Theory: Study of equilibria of games played by population of players where the "fitness" of the players derives from the success each player has in playing the game. It provides tools for describing situations where a number of agents interact. Evolutionary game theory

improves upon traditional game theory by providing dynamics describing how the population will change over time.

Genetic Algorithm: Class of algorithms used to find approximate solutions to difficult-to-solve problems, inspired and named after biological processes of inheritance, mutation, natural selection, and generic crossover. Genetic algorithms are a particular class of evolutionary algorithms.

Negotiation Mechanism: The actual implementation of negotiation strategy and negotiation protocol.

Negotiation Strategy: Specification of the sequence of actions the agent intends to make during the negotiation.

Paretto Efficient Agreement: State of things in which no player (in game theory) is worse off than the others. It typically refers to the distribution of resources. The agreement could lead to cooperation of players.

Reasoning Engine: A computer program that enables the agents to negotiate with other agents and that involves negotiation strategy.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 810-815, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Designing Learner-Centered Multimedia Technology

Sandro Scielzo
University of Central Florida, USA

Stephen M. Fiore
University of Central Florida, USA

Haydee M. Cuevas
University of Central Florida, USA

INTRODUCTION

The ubiquitous use of information technology (IT) promotes a fast-paced and dynamic training environment with enormous potential for performance increases in a variety of domains. This reality has many important ramifications, including how best to incorporate multimedia IT into computer-based training (CBT). Well-designed CBT offers us tremendous potential to effectively and efficiently train the workforce, foster learning in academic environments, and improve performance over and above what is currently achieved. Following a learner-centered design approach, in this article, we present an in-depth look at the use of multimedia CBT, as it relates to aptitude-treatment interactions; that is, how various CBT designs can differentially interact with individual learner aptitudes, such as spatial and verbal ability, to influence training outcomes. The goal of this article is to emphasize the importance of learner-centered design when developing multimedia computer-based instructional material for the growing needs of many sectors of society.

BACKGROUND

CBT has long been touted as a cost-effective and efficient medium for instruction due to, in part, IT availability (e.g., McDermott, 2006). Over the past few decades, a number of theoretical frameworks have flourished, aimed at understanding how multimedia information is processed and how best to design CBT to maximize the amount of information retained from such instruction. For example, Paivio's (e.g., 1971) Dual Coding Theory underscored the importance of using multiple compatible modalities—such as text and images, which are parallel-processed within human working memory—in order to a) generate a strong encoding of processed information, and b) increase memory retrieval of the encoded information when compared to traditional methods employing only one modality (e.g., simply text, narrated lectures, etc.). However, the technological aspect of multimedia training was not yet developed and IT was still in its infancy in relation to CBT. Although IT was not yet widespread, the relevance of Paivio's theory to current CBT design is the focus on the human's capacity for processing multimodal information. Mayer (2001) later espoused this concept in his *Cognitive Theory of Multimedia Learning*, specifically addressing the manner in which CBT multimedia information is processed. Mayer and colleagues devised a number of learner-centered principles guiding CBT instructional design (e.g., Mayer, 1999, 2001) with the goal of helping society capitalize on the ubiquity of IT (e.g., Galvin, 2003; Najar, 1998).

The importance of following learner-centered design principles and guidelines becomes even more paramount for maximizing the potential of IT in a CBT environment. This emerging focus on learner-centered CBT design illustrates a crucial balancing act: on one hand, CBT design needs to capitalize on IT's power and availability; on the other hand, human cognitive limitations need to be considered when developing complex CBT. However, current learner-centered CBT design may be insufficient when individual learner aptitudes come into play. To further improve the efficiency of CBT design, it is pivotal to understand how individual learner characteristics can differentially influence training outcomes (e.g., Mayer, 2001). With this objective in mind, in this article, we present an empirical examination of the interaction between CBT design and individual aptitudes in relation to their impact on the training's learning outcomes and instructional efficiency.

DESIGNING EFFICIENT CBTS: THE MEDIATING ROLE OF INDIVIDUAL APTITUDES

Investigating the influence of specific learner aptitudes, such as spatial and verbal ability, on processing multimedia

information has been an integral component of our research (e.g., Cuevas, Fiore, Bowers, & Salas, 2004; Cuevas, Fiore, & Oser, 2002; Scielzo, Fiore, Dahan, Lopez, & Stafford, 2006; Scielzo, Cuevas, & Fiore, 2005; Scielzo, Fiore, Cuevas, & Klein, 2003). The importance of understanding the complex relationship between individual differences and multimedia information processing offers the opportunity to further refine CBT design guidelines by specifically looking at the manner in which design implementations (treatment) and individual differences (aptitudes) interact. As a result, a more precise learner-centered theory of CBT design can be obtained. Furthermore, evaluating the effectiveness of CBT design requires a number of training outcome measures including (a) assessment of learners' knowledge acquisition, and (b) the instructional efficiency of the CBT program itself, discussed next.

Learner-Centered Computer-Based Training and Assessment

Our approach to assessing learning in CBT utilizes increasingly complex measures of knowledge acquisition—from basic, factual knowledge to more complex integrative knowledge—in order to provide a more complete view of learning as it relates to performance. Our past research investigating CBT design found that the differential effect of training manipulations was often revealed via more complex knowledge assessment measures, that is, only those measures requiring learners to comprehend how various concepts *relate to one another* were able to successfully isolate the effects of training manipulations (e.g., Cuevas et al., 2004; Cuevas et al., 2002; Fiore, Oser, & Cuevas, 2000; Fiore, Cuevas, & Oser, 2003; Fiore, Cuevas, Scielzo, & Salas, 2002; Scielzo et al., 2003).

In addition, instructional efficiency (e.g., Kalyuga, Chandler, & Sweller, 1999; Paas & Van Merrienboer, 1999) has been shown to be an important CBT assessment technique that offers further insight into the effectiveness of varying training manipulations. Instructional efficiency combines standardized measures of knowledge performance and subjective mental workload (perceived cognitive effort during training or performance) to determine overall CBT efficiency; that is, instructional efficiency considers performance levels in relation to the subjective appraisal of how mentally tasking it was to achieve these levels of learning or performance. Specifically, instructional efficiency is a normalized index ranging from -1 to +1, with positive scores indicating higher instructional efficiency (i.e., mental effort exerted is less, relative to the standard effort required to achieve that level of performance) and negative scores indicating lower instructional efficiency (i.e., mental effort exerted is greater, relative to the standard effort required to achieve that level of performance). Baseline (or standard level of efficiency) is represented by zero. Overall, instructional efficiency is an important measure that provides further information to enable instruction system designers to more sensitively distinguish between CBT designs yielding similar levels of performance.

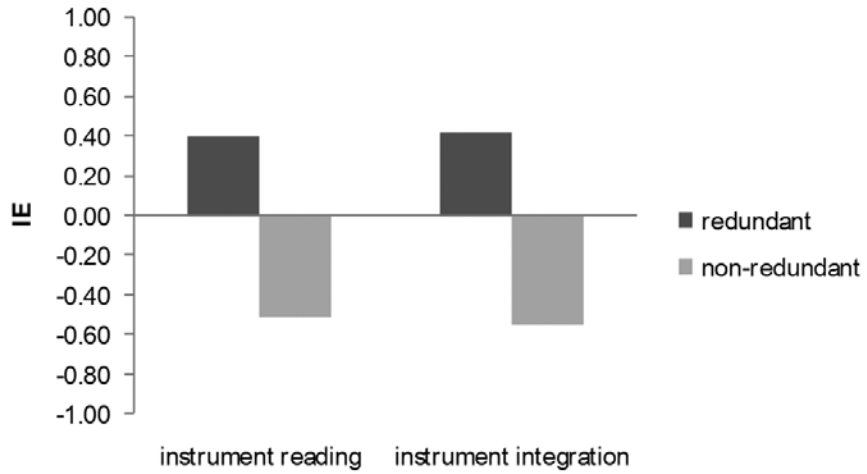
Finally, studies have shown that CBT design manipulations may interact with individual learner aptitudes to influence training outcomes. Aptitude-treatment interactions (see Snow, 1989), in relation to multimedia CBT have been documented to be particularly prominent for spatial (e.g., Chun & Plass, 1997; Mayer, 2001; Scielzo et al., 2006) and verbal (e.g., Cuevas et al., 2002; Chun & Plass, 1997; Mayer, 2001) ability. The next section briefly summarizes empirical findings that more explicitly highlight the mediating role of individual differences on the effect that a given CBT design may have on training outcomes.

Summary of Empirical Findings

Our research on complex multimedia CBT has investigated a number of design manipulations thought to differentially impact training outcomes in terms of knowledge acquisition and instructional efficiency. In our earlier work, individual differences, such as spatial and verbal ability, were covaried out to evaluate the more general effects of design manipulations. For example, when looking at the influence of the differential combination of modalities in a dual-coding paradigm (i.e., presenting one modality vs. two modalities), we found that the beneficial effects of including graphical representations to illustrate concepts presented in text extended to more complex training paradigms requiring the acquisition and integration of numerous concepts over multiple training modules (i.e., training rudimentary aspects of the principles of flight; Fiore et al., 2003). Specifically, results indicated that multimedia CBT combining text and graphical representations supported comprehension and integration in a complex training paradigm. Importantly, by covarying out spatial ability, this effect was isolated using a measure of knowledge integration requiring the ability to understand the relationship of inter-related concepts presented across multiple training modules. These findings, in part, suggest that dual-coding effects (i.e., the claim that encoding instructional material via two modalities improves information retrieval when compared to encoding material via only one modality) do extend to more complex multimedia CBT, but that these effects may only be diagnosable when using measures designed to assess the integration of concepts (Fiore et al., 2003; see also Cuevas et al., 2002).

Expanding upon this line of research, we investigated the effects of differential combination of modalities in a temporal paradigm (i.e., presenting modalities simultaneously or sequentially) and found that the principle of redundancy (Mayer, 2001), indicating that combining narration, text and animations simultaneously is detrimental on knowledge performance due to high levels of workload, did not extend

Figure 1. Instructional efficiency (IE) scores based on instrument reading and instrument integration measures



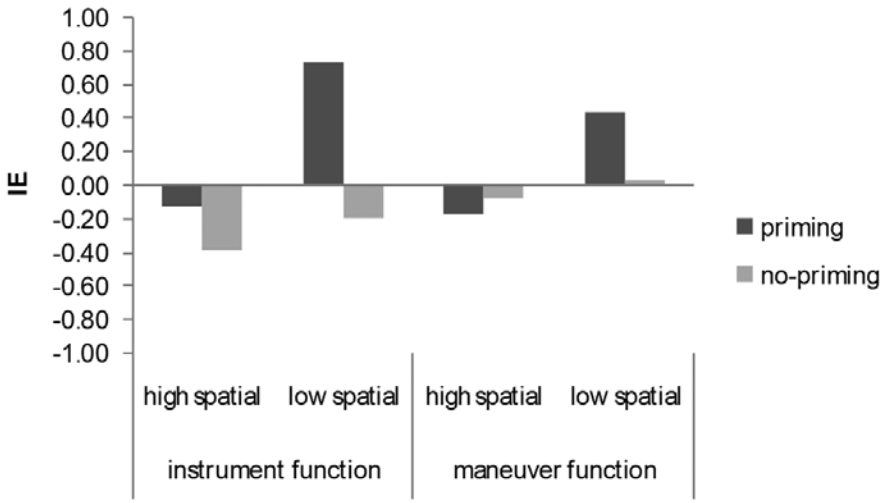
to more complex training paradigms such as the domain of aviation (Scielzo et al., 2003). Specifically, results indicated that multimedia CBT combining text, narration and animations can favorably enhance performance when the domain studied is complex. As in our earlier work, this effect was isolated using more complex measures of knowledge integration (instrument reading and instrument integration). These findings suggest that the principle of redundancy (i.e., the claim that multimedia CBT redundancy is detrimental to knowledge acquisition) may need to be revised to explain the differential training outcomes found across domains, ranging from relatively simple (e.g., Kalyuga et al., 1999, 2004) to relatively complex (Scielzo et al., 2003).

In terms of instructional efficiency, unpublished results of this study analyzed for this chapter revealed a similar trend, with a CBT redundant format yielding greater instructional efficiency (for details on the experimental methodology, see Scielzo et al., 2003). Specifically, univariate statistics showed a significant effect of redundancy on the two more complex measures of knowledge integration: instrument reading, F (1, 32) = 9.07, p = .005 (partial η^2 = .22), and instrument integration, F (1, 32) = 6.92, p = .013 (partial η^2 = .18) [a]. Both measures indicated that CBT instruction with three modalities (text, narration, and animations) was more efficient than instruction with two modalities (narration and animations) (see Figure 1).

In later research, we directly examined the mediating effect of individual differences in learner aptitudes on training outcomes. For example, Scielzo et al. (2005) investigated the effect of a training intervention (embedded content-free prompts asking learners to elaborate on the information presented) on overall retention of the training material and the CBT's instructional efficiency within the context of a dynamic distributed decision making task. Overall, the study found that the training intervention's effect was differentially affected by participants' verbal ability. Specifically, incorporating prompts into the CBT (e.g., asking participants to complete content-free question stems such as "*How would you use_____to_____?*") significantly improved retention of the information presented for participants with low verbal ability. No difference was found for participants with high verbal ability. In addition, when evaluating outcomes for participants with low verbal ability, only the CBT designed with the training intervention yielded positive instructional efficiency scores (i.e., higher performance was achieved with less perceived cognitive effort). When evaluating outcomes for participants with high verbal ability, the CBT's instructional efficiency was consistently positive across both training conditions. These results illustrate the importance of considering individual differences in learner aptitudes such as verbal ability in CBT design by showing that low-verbal participants may benefit from training manipulations that have little or no effect on high-verbal participants.

In another study using our "principles of flight" training paradigm, Scielzo et al. (2006) investigated the effect of text priming in a temporal context and the potential mediating effect of spatial ability. Priming was achieved by first showing textual information (written text), followed by a verbal narration with a respective visual animation. Priming was compared to a no-priming condition that presented the three modalities (text, narration, and animation) simultaneously. Overall, results showed that text priming yielded significantly higher performance than nonpriming for trainees whose spatial ability was low, while high spatial ability trainees favored from the nonpriming training implementation. Further analysis of this study's results, not yet published, examined

Figure 2. Instructional efficiency (IE) scores based on instrument function and maneuver function measures



the effects of priming vs. nonpriming on instructional efficiency[b]. Results showed a strong, but nonsignificant, trend in the expected direction on the two more complex knowledge integration measures (see Figure 2). Although not significant, the importance of this study similarly resides in highlighting the important role of aptitude-treatment interactions, and how different training interventions can lead to opposite results when factoring in the mediating effect of individual differences in learner aptitudes.

FUTURE TRENDS

The issues discussed in this chapter lead to broad theoretical and practical implications for future research. Theoretically, researchers need to more precisely define the mediating effects of individual differences in learner aptitudes (e.g., verbal and spatial ability) on CBT design and training outcomes to better understand the learning and knowledge acquisition process. Once a stronger theoretical foundation is established, a number of practical implications to address individual differences with learner-centered CBT design can emerge. For example, *intelligent tutoring systems* (e.g., Akras & Self, 2002) can be incorporated into CBT design to automatically adapt the instruction to match trainees' individual strengths and limitations. Another promising approach focuses instead on developing training aimed at improving specific learner aptitudes. For example, Rehfeld (2006) has shown that spatial ability can be improved via mental rotation training (see also Kass, Ahlers, & Dugger, 1998). Overall, future developments in CBT design will need to address individual differences to further their potential in terms of knowledge gains and overall instructional efficiency.

CONCLUSION

The overall objective of this article was to illustrate the important role of individual differences in training outcomes within CBT environments. The research findings presented in this article showed how learner aptitudes interact with CBT design manipulations to influence outcomes on the measures that are more indicative of higher levels of knowledge acquisition. Our findings also highlighted the relation between individual differences and the training program's instructional efficiency, showing that trainees with lower aptitudes benefited more from CBT learner-centered design (i.e., greater performance achieved with less perceived cognitive effort). Overall, the findings reported in this article support the need to better utilize information technology to design more effective computer-based training programs that can flexibly adapt to individual differences in learner aptitudes, and thereby, maximize overall learning outcomes.

ACKNOWLEDGMENT

Work on this article was partially supported by funding on Grant Number N000140610118, awarded to the second author from the Office of Naval Research (ONR). The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies, either expressed or implied of the ONR, Department of the Navy, Department of Defense, the U.S. Government, the University of Central Florida, or any of the organizations with which the authors are affiliated.

REFERENCES

AKHRAS, F. N., & SELF, J. A. (2002). Beyond intelligent tutoring systems: Situations, interactions, processes and affordances. *Instructional Science, 30*(1), 1-30.

Chun, D. M., & Plass, J. L. (1997). Research on text comprehension in multimedia environments. *Language Learning & Technology, 1*(1), 60-81.

Cuevas, H. M., Fiore, S. M., Bowers, C. A., & Salas, E. (2004). Fostering constructive cognitive and metacognitive activity in computer-based complex task training environments. *Computers in Human Behavior, 20*, 225-241.

Cuevas, H. M., Fiore, S. M., & Oser, R. L. (2002). Scaffolding cognitive and metacognitive processes in low verbal ability learners: Use of diagrams in computer-based training environments. *Instructional Science, 30*, 433-464.

Fiore, S. M., Cuevas, H. M., & Oser, R. L. (2003). A picture is worth a thousand connections: The facilitative effects of diagrams on task performance and mental model development. *Computers in Human Behavior, 19*, 185-199.

Fiore, S. M., Cuevas, H. M., Scielzo, S., & Salas, E. (2002). Training individuals for distributed teams: Problem solving assessment for distributed mission research. *Computers in Human Behavior, 18*, 729-744.

Fiore, S. M., Oser, R. L., & Cuevas, H. M. (2000). Knowledge structure instantiation: Implications for measurement and transfer. In *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomic Society*, (Vol. 1, pp. 292). Santa Monica, CA: Human Factors and Ergonomics Society.

Galvin, T. (2003). 2003 industry report. *Training Magazine, 40*(9), 19-45.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351-371.

Kalyuga, S., Chandler, P., & Sweller, J. (2004). When redundant on-screen text in multimedia technical instruction can interfere with learning. *Human Factors, 46*(3), 567-581.

Kass, S. J., Ahlers, R. H., & Dugger, M. (1998). Eliminating gender differences through practice in an applied visual spatial task. *Human Performance, 11*(4), 337-349.

Mayer, R. E. (1999). Multimedia aids to problem-solving transfer. *International Journal of Educational Research, 31*(7), 611-623.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

McDermott, J. T (2006). Computer-based flight simulation: A cost effective way for general aviation pilots to improve their instrument proficiency. *International Journal of Applied Aviation Studies, 6*(1), 155-163.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*(1), 156-163.

Najjar, L. J. (1998). Principles of educational multimedia user interface design. *Human Factors, 40*(2), 311-323.

Paas, F. G. W. C, Van Merrienboer, J. J. G., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills, 79*, 419-430.

Rehfeld, S. A. (2006). The impact of mental transformation training across levels of automation on spatial awareness in human-robot interaction. *Dissertation Abstracts International*, AAT 3242463.

Scielzo, S., Cuevas, H. M., & Fiore, S. M. (2005). Investigating individual differences and instructional efficiency in computer-based training environments. In *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*, (pp. 1251-1255). Santa Monica, CA: HFES.

Scielzo, S., Fiore, S. M., Cuevas, H. M., & Klein, J. L. (2003). Impact of multimedia presentation on knowledge acquisition for complex training. In *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society,* (pp. 2042-2044). Santa Monica, CA: HFES.

Scielzo, S., Fiore, S. M., Dahan, Y., Lopez, J., & Stafford, S. (2006). Computer based training and multimedia design: The role of spatial aptitudes in learning. In *Proceedings of the 50th Annual Meeting of the Human Factors and Ergonomics Society*, (pp. 1231-1235). Santa Monica, CA: HFES.

Snow, R. (1989). Aptitude-treatment interaction as a framework for research on individual differences in learning. In P. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences*. New York: W.H. Freeman.

KEY TERMS

Aptitude-Treatment Interaction: The interaction of learners' individual differences (i.e., aptitudes) with experimental manipulations (i.e., treatment). In the context of computer-based training, aptitude-treatment interactions illustrate the differential effect of design manipulations when factoring-in specific learners' aptitudes (e.g., verbal and spatial ability).

Cognitive Theory of Multimedia Learning: A theory of multimedia information processing for computer-based

training paradigms. This theory underscores the importance of supporting the three fundamental cognitive processes leading to learning: information selection, organization, and integration.

Computer-Based Training: The use of computers as a medium to convey instructional material. Typically, computer-based training replaces the instructor and allows learners to interact with multimedia content in a self-paced environment.

Dual-Coding Theory: A theory of learning that indicates the benefit from learning via two noncompeting modalities (in terms of accessing the same attentional resources) such as text and images. The dual-coding benefit refers to the increased probability of information retrieval when compared to encoding information with only one modality.

Individual Differences: Learners variability in terms of knowledge, skills, and attitudes. In a computer-based training environment, individual differences play a potential mediating role between training manipulations and learning.

Instructional Efficiency: An index of training efficiency that combines learners' subjective appraisal of mental workload in the training environment with learners' performance on knowledge testing. The instructional efficiency index ranges from -1 to +1, with negative scores indicating lower instructional efficiency, and positive scores indicating higher instructional efficiency. Baseline instructional efficiency is indicated by zero.

Learner-Centered CBT: A computer-based training interface designed to account for learners' individual differences by either adapting to the learners' aptitudes, such as intelligent tutoring systems, or by providing an interface designed to accommodate a specific learner aptitude (e.g., low verbal or spatial aptitude).

ENDNOTES

[a] The design was a 2 temporal (simultaneous vs. sequential) x 2 redundancy (redundant vs. nonredundant) between subjects design, with spatial ability as a covariate. Forty-three participants were trained on the principles of flight. GLM univariate analyses of covariance were used with an alpha level of .05.

[b] The design of the study was a 2 priming (priming vs. no-priming) x 2 spatial ability (high vs. low) between subjects design. Twenty-eight participants were trained on the principles of flight. GLM univariate analyses of covariance were used with an alpha level of .05 (see Scielzo et al., 2006).

Designing Web Systems for Adaptive Technology

Stu Westin
University of Rhode Island, USA

INTRODUCTION

For over a decade the term *digital divide* has been used to refer to discrepancies between various population segments in terms of access to information technologies. The digital divide is in opposition to the ideal of *equality of access* in which all citizens are afforded uniform access to information and information technology. Discussions on this topic seem to most often focus on such factors as race, income, education, geography, and the like. There is, however, a significant and growing group of "digital have-nots" that is frequently overlooked. This group comprises individuals who have some form of physical, sensory, and or mental disability. While the need for full enfranchisement of this group can be effectively argued on legal as well as ethical grounds, it can be shown to make sound business sense as well.

Consider this statistic from the most recent U.S. Census. A startling 21.8% of Americans above the age of 16 have at least one disability that results in a "substantial limitation" of one or more "major life activities." Examples of such disabilities are vision problems (3.5%), hearing problems (3.3%), difficulty using hands (3.0%), and learning disabilities such as dyslexia (1.4%) (U.S. Department of Commerce, 2000, pp. 62-63). Each of these disabilities carries negative consequences regarding accessibility to Web-based resources.

The prevalence of disability increases with age. For example, according to 2005 data, 12.1% of Americans in the age group 16-64 have at least one disability. The percentage jumps to 40.5% when considering those of age 65 and above (U.S. Department of Commerce, 2006, Table S1801). Much of this dramatic increase in occurrence is due to declining vision, hearing, and dexterity (Bergel, Chadwick-Dias, & Tullis, 2005; Fox, 2004; Loiacono, McCoy, & Chin, 2005; Steinmetz, 2006).

The youngest American baby boomers are now in their forties. The average age of the population of the U.S. and of most other developed nations will increase substantially over the next few decades, as will the concomitant prevalence of physical disability (Bergel et al., 2005). This demographic shift is due partly to the post World War II "population bubble," but it is also due to the tremendous increase in life expectancy in modern times (an increase of 30 years since 1900, according to U.S. Administration on Aging statistics) (Mosner & Spiezle, 2003). The segment of the American population comprising individuals of age 50 and above will grow from the current 38% to 47% by the end of the next decade (Moos, 2005).

Also growing dramatically is the average age of the workforce. Workers are delaying retirement for numerous reasons, while the rate at which younger workers enter the workforce is declining (Mosner & Spiezle, 2003). In an increasingly Web-oriented information-based economy, worker productivity hinges on accessibility to Web-based systems. This issue demands more attention as the age of the workforce (read prevalence of physiological impairments among workers) increases.

This article considers some of the issues surrounding accessibility to Web systems and services by individuals with imperfect abilities. It is argued that, beyond the moral and legal reasons for accommodating this group, there are numerous advantages for business and commerce that can be achieved.

BACKGROUND

As is the case with all technologies, the design and organization of Web content can greatly impact accessibility of that content by persons with certain physical or mental impediments or disabilities. Consider, for example, those individuals who have even minor mobility or dexterity problems. This might include persons of advanced age, as well as those who suffer from arthritis, rheumatism, Parkinsonism, effects of stroke, or similar maladies. For this group an activity as simple as clicking a particular hot-zone on an image map can be difficult, depending on the size and the complexity of the object. Even activities as common as using the scrollbar to move through the content of a Web page can be troublesome to individuals with motion impairments.

It should be noted that even conditions that are not considered disabling can negatively affect access to poorly designed Web content in certain circumstances. For example, about 8% of males worldwide are color deficient (often called color blindness). The vast majority of these individuals have problems discerning red or green. The prevalence and sever-

ity of this condition often increases with age. A commercial Web page that states *Products listed in red are currently out of stock* may convey little information to the color deficient electronic shopper.

The terms *assistive technology* and *adaptive technology* are used to describe technologies which are intended to help provide independence to disabled individuals. The two terms are often used interchangeably in the literature, but in the case of adaptive technology (AT), the focus is on providing access to products and systems which were initially designed for use by people who are not disabled. The Web is an example of such a system. Adaptive and assistive technologies, when applied to computing and information systems, are sometimes referred to as "electronic curb cuts." This term makes an analogy to the decades-old federally mandated removal of curbs at pedestrian crossing points to facilitate use by persons in wheelchairs.

In the case of the Web, the adaptive technological problems can be particularly vexing because of its stateless, two-tiered (i.e., client/server) architecture. That is, the adaptive technologies reside on the client side, but the Web content can be designed and served with no knowledge of how the AT is configured, or even that such is being used. Consequently, Web content that is designed without regard for such technologies can render the content useless for the end user.

An approach to Web content design that aimed at reducing or eliminating barriers to accessibility, and at facilitating the effectiveness of AT is said to be an *accessible Web design*. Related to this is the notion of *universal design* where the intent is to meet the needs of the broadest range of clients, regardless of their individual abilities, disabilities, circumstances, or environments. In the words of Mates (2006, ch. 2, sec. 2)

The Web page designer addressing universal design and accessibility is more concerned with information dissemination for all, rather than visual appeal for most. When designing the document, an attempt is made to make all the material displayed as accessible as possible, whether it is a menu item, graphic, or video clip. Creating accessible Web pages may not take additional money, just more time and consideration.

Adaptive technology on the Web can be as simple as tweaking your browser settings to display the largest text size, or to specify default colors for text and background. Most current Microsoft products, including IE, provide a set of options aimed at broadening accessibility. In the case of IE the user is able to prevent a Web page from overriding his or her choice of text colors, font styles, or font sizes. There is also an option to force the use of a local (i.e., user supplied) style sheet for rendering the presentation.

Users with poor vision or with learning disabilities, for example, will often configure their browsers to display oversized text in a non-serif style (e.g. Arial) with a high-contrast color scheme (e.g. black on white). Features which might be distracting, such as background images and italic type, may be removed and all text may be displayed in bold style to further enhance contrast. The implication of this with respect to Web content design is that the presentation rendered on the client side may be very different from that which is conceived in the mind of the designer. Any meaning or information that is presented only through color, text style, or the like may be lost. The principles of universal design are aimed at avoiding this.

While the above description of employing AT–modifying standard browser behavior through intrinsic features–is relevant to this discussion, more substantial accommodations are often made through third-party solutions. In this case, hardware or software devices and mechanisms are designed with a specific accommodation in mind. Two common examples of this are large-print access systems and screen-reading systems. These two types of AT are now briefly described within the context of Web-based content so that some of the complexities and special requirements of such systems can be understood. The intention here is to illustrate how the efficacy of such systems is affected by Web content design. A full discussion of third-party AT solutions is beyond the scope of this article.

WEB-BASED EXAMPLE: LARGE-PRINT ACCESS SYSTEMS

Large-print access systems usually comprise a screen magnification software component and may or may not include a special large monitor device suitable for handling the large display image. Large-print access systems are used to accommodate individuals with impaired vision, but with sufficient vision to discern shapes. The systems also are of aid to individuals with certain forms of cognitive or learning disabilities. These systems are expected to become more commonplace as our population ages.

With large-print access systems, the complication with respect to Web content accessibility stems from the fact that the systems can severely distort the visual presentation, and the nature of this distortion can vary substantially. In some cases, the entire page is magnified so there is relatively little geometric distortion, but only a portion of the page may be visible at one time, depending on the size of the display device. In extreme cases only one or two words may be displayed on the screen at a time.

Other software products of this type will enlarge only the textual content but leave graphic content untouched. Note that words contained within images may therefore be unreadable. Still other products will only magnify the area of

the screen that surrounds the mouse pointer, or will magnify an area of the screen and will move the magnified area down the page at a fixed rate. In this latter case, the user is forced to process the information from top to bottom, regardless of the design of the content.

WEB-BASED EXAMPLE: SCREEN-READING SYSTEMS

When a disability is such that large-print technology is unsuitable–say in the case of blindness, severe dyslexia, or literacy problems–a screen-reading system may be employed. The intent of these systems, also known as audio- or speech-output systems, is to allow the user to "listen" to whatever content is presented on the screen. Output is in the form of a simulated voice which is driven by screen-reading software. Screen-reading systems are practical for any computing application where text-based output is generated. When the output moves beyond a pure text presentation of words, the efficacy of the systems can suffer if the information is improperly structured.

In the case of Web applications screen-reading system software relies heavily on punctuation and on the source markup tags to determine how textual content should be presented aurally. For example, proper use of the <H1> and the <H2> tags on a page would be required for the software to accurately interpret content as a primary heading or as a secondary heading. Text size or presentation format would not be a factor in this decision. Similarly, appropriate use of the <UL> tag followed by a series of <LI> tags would be required for the software to present the content aurally as a bulleted list of items rather than as a sequence of paragraphs, each headed with a dot icon. As for graphic elements, for example, charts, icons, animations, background images, button images, and so forth, the screen-reader relies exclusively on the ALT (alternate text) attribute or on the LONGDESC (long description) attribute of the image element.

Designing Web content that will be accurately interpreted by this technology can be exigent. A discussion of the challenges and caveats can be found in Asakawa (2005). In common Web-design practice, validation of Web-based applications usually extends to the point where it is determined that the content is visually correct. Speech-output technology requires that the content be syntactically correct (e.g., just because something appears as a bulleted list when viewed in a browser, doesn't mean that it will be interpreted as such by a screen reader).

A mouse is a visually driven device, so mice are not used with screen-reading systems. All input is provided through the keyboard, and focus is given to individual hyperlinks and form objects with the Tab key. The visual cues that normally guide the sighted user as to the requirements of input or the interpretation of output are totally irrelevant in this environment. It is vital, therefore, that all of the visual guides within Web content be supplemented with commensurate textual information so that equivalent aural guides can be provided by the system.

Graphic elements that are meant to provide information content must be coded differently from those that are intended to provide only visual enhancement. The former group should include tagged alternate text that provides as much of the information content as is practical. The latter group should have a null alternate text description (ALT="") to prevent the screen reader from announcing the presence of an "unspecified graphic element" to the user. Chart and graph entities should include a full summary of the content as a linked, plain text description (LONGDESC attribute or "d-link").

With highly interactive applications, such as those involving e-commerce, it is vital that the traditional visual cues be supplemented with equivalent textual information. The disabled individual will be tabbing through and "hearing" the individual items on an interactive Web form rather than viewing the entire form and configuring or filling in the relevant form objects with the aid of a mouse. The overall demands of the application (e.g., which fields are required and which are optional) as well as the semantic of each individual item (e.g., what is the implication of selecting this checkbox?) must be clearly detailed for the screen-reader software.

DESIGNING FOR AT: STANDARDS AND GUIDELINES

Depending on the situation, accessibility of Web content may be legally mandated. In the U.S., the determining factor is whether the site is within the scope of Section 508 of the Rehabilitation Act of 1973. This section requires that information systems utilized by U.S. Government agencies be accessible to people with disabilities. Both in-house-developed and outsourced systems fall under the purview of Section 508. For the first 25 years of its existence the principles of Section 508 were largely unenforceable and ignored. That changed in 1998 with amendments to the act instituted by President Clinton. The 1998 amendments to Section 508 provided technical standards which are unambiguous and enforceable.

The intent of Section 508 is surely noble. The motivation is to provide equality of access to information resources. The expectation is that federal employees who are disabled, as well as disabled members of the public at large, are to have access to federal information services at the same level as their nondisabled counterparts. Many other nations (e.g., Australia, Canada, England, Portugal) have similar legislation in place (Lazar & Greenidge, 2004).

While Section 508 applies to all information technology systems developed or procured by government agencies,

Subsection 1194.22 has arguably been the biggest burden for agencies. This subsection has to do with Web-based, both Internet and intranet, applications and systems. The difficulty surrounding Subsection 1194.22 is understandable in considering that the accommodations legislated by Section 508 must be available to the general public. The agencies responsible for Section 508 compliance have no control over the hardware or software choices or configurations utilized by the public clients of the Web systems that they host.

Subsection 1194.22 includes 16 specific guidelines (labeled "a" through "p") that must be met for full compliance. §1194.22(a), for example, reads as follows. *A text equivalent for every non-text element* [of the Web content] *shall be provided (e.g., via 'alt', 'longdesc', or in element content*). As another example, §1194.22(c) reads *Web pages shall be designed so that all information conveyed with color is also available without color, for example from context or markup*. The full list of guidelines is available at http://www.section508.gov/index.cfm?FuseAction=Content&ID=12#Web.

Often confused with Section 508 is WCAG (Web Content Accessibility Guidelines). This set of guidelines aimed at reducing barriers to accessibility was introduced in 1999 by the World Wide Web Consortium (W3C) as a part of its Web Accessibility Initiative. This W3C initiative commenced in 1997 and its intended purpose is best described by Tim Berners-Lee in the inaugural press release as follows. "The W3C is committed to removing accessibility barriers for all people with disabilities, including the deaf, blind, physically challenged, and cognitive or visually impaired. We plan to work aggressively with government, industry, and community leaders to establish and attain Web accessibility goals" (World Wide Web Consortium, 1997). Unlike Section 508, however, WCAG is not a legal mandate.

WCAG comprises 14 basic guidelines aimed at ensuring accessibility. An example is Guideline 1 which states *Provide equivalent alternatives to auditory and visual content.* Each of the 14 general principles is accompanied by a set (1-10) of numbered *checkpoints* describing how the guideline would be applied in specific application scenario examples. Each checkpoint is assigned a *priority* from 1 to 3. To denote impact on resulting accessibility, priorities 1, 2 and 3 are ascribed the respective tags *must satisfy, should satisfy,* and *may address*. In terms of WCAG conformance, Web content can conform at one of three levels, as shown in Table 1. The full WCAG specification can be found at http://www.w3.org/TR/WAI-WEBCONTENT/.

The motivations behind the two standards, (WCAG and Subsection 1194.22) are similar. That is, to eliminate barriers to accessibility to Web content. Their orientations differ slightly, however, because Section 508 focuses exclusively on disabilities, while WCAG also recognizes environmental and equipment factors such as noisy and poorly-lit surroundings, and hands-free applications. For this and other reasons, full compliance with one standard does not necessarily indicate full compliance with the other. A detailed comparison of the standards is available at http://www.jimthatcher.com/sidebyside.htm#s508View.

Several automated tools to evaluate accessibility and compliance to standards are available. One of the earliest and most recognized free tools is *Bobby* (now called *WebXACT*, http://webxact.watchfire.com/). A full discussion of such automated tools, including their strengths, limitations, and overall efficacy can be found in Ivory, Mankoff, and Le (2003).

Despite the availability of published standards, and of automated tools to evaluate compliance to those standards, accessible content on the Web remains more the exception than the rule. Depending on the type of Web site, violations of principles of accessibility range between 70-98%, except for college and university sites which fare better (Lazar, Dudley-Sponaugle, & Greenidge, 2004; Loiacono et al., 2005). Even in the case of federal and federal-contractor Web sites, those that are mandated under federal law to be Section 508 compliant, only 23% have been found to meet that legal obligation (Loiacono et al., 2005). Perhaps even more discouraging, Web sites have been shown to actually *decline* in accessibility over time, regardless of the category of site (Asakawa, 2005; Hackett, Parmanto, & Zeng, 2003; Lazar & Greenidge, 2006).

Table 1. WCAG conformance requirements

Conformance Level	Requirements
Level AAA	Web content satisfies all checkpoints of Priority 3 and above.
Level AA	Web content satisfies all checkpoints of Priority 2 and above.
Level A	Web content satisfies all checkpoints of Priority 1.

EMBRACING AT: THE BUSINESS JUSTIFICATION

Whether legally mandated or not, at least partial compliance with Section 508, with WCAG, or with both is a responsible business practice. The case for accessible Web design is easy to make on ethical grounds alone. It is difficult to argue that Web content made available to some should not be made available to all, regardless of ability, disability, or situation. Each step toward further compliance represents a potential increase in reach of an organization's Web-based messages and services. An organization's employees also stand to realize advantages of increased accessibility to Internet and intranet based content in the workplace. This becomes especially pertinent as the workforce ages.

As a side benefit, accessible design practices are known to simplify the structure of Web pages and Web sites much as adhering to accepted software development practices and standards leads to improved software design and performance. The end result of both of these efforts is fewer errors and simplified maintenance, leading to reduced cost.

An additional, and often unexpected, benefit of accessibility guideline compliance is that the design principles greatly facilitate accessibility of Web content by devices and technologies which are unrelated to disability. Examples of these technologies are PDAs, handheld computers, Web-enabled phones, automobile-based PCs, audio browsers, and the like. It turns out that the special requirements of these client-side devices are the same as those of most adaptive technologies; specifically, the requirement that the content be independent from display issues. Still another benefit is that the efficacy of the modified Web content is increased in less than perfect work environments such as conditions of poor lighting, or when using older monochrome or LCD screens.

When the physical cuts in the curbs of roadways were federally mandated decades ago to allow wheelchair access, the general population realized unanticipated benefits. These physical curb cuts also facilitated the use of bicycles, roller blades, baby strollers, shopping carts, and skateboards. Similarly, the "electronic curb cuts" can simplify the use of numerous Web-enabled technologies by the population at large.

Where B2C e-commerce is concerned, the disabled population–one of the largest minorities that exists–represents a potential market niche that is largely untapped or undertapped. Despite the heterogeneity with respect to demographic and disability factors, this market segment exhibits numerous features that should be of interest to e-tailers.

Of all adults who use the Internet, disabled adults spend twice as much time online, an additional 10 hours per week, than the nondisabled (Taylor, 2000). This former group is also substantially more likely (48% vs. 27%) to recognize the Internet as factor that "significantly improves the quality of their lives." The difference becomes more pronounced (56% vs. 6%) when considering individuals 65 and older (Taylor, 2000). Slicing this market segment across other demographic factors reveals additional economic enticements. Granted, in almost all cases, home Internet access, as well as general PC usage is lower among disabled individuals compared to the nondisabled. However, the discrepancy diminishes as both income and education rise (U.S. Department of Commerce, 2000). In other words, it is the wealthiest and the most educated of the disabled population who are in the best position to engage in Web-based commerce.

Considering the age factor, which correlates with the incidence of physical impairment and with reliance on technology accommodation, the numbers present a clear message. The U.S. market segment comprising those of age 50 is currently the fastest growing segment; it will increase from the current 38% to 47% by 2020 (Moos, 2006). Households in this age group currently spend twice as much as younger adults and control $750 billion in discretionary income (Moos, 2006). Furthermore, the sharp growth in interest in online banking and electronic shopping by elders is projected to only increase because this group has been shown to be as enthusiastic as their younger counterparts once they get online (Fox, 2004).

The segment of the population that relies on special accommodation appears, therefore, ready and willing to embrace the B2C e-commerce paradigm. They may be ready and willing, but are they *able*? The answer to this question may depend on the degree to which accessibility features are included in commercial Web sites and Web services. Unfortunately, accessibility in retail and service Web sites is as low as 17% (Loiacono & McCoy, 2004).

FUTURE TRENDS

It is difficult to predict what the future holds for Web content accessibility. Based on clear demographic and cultural trends, the need for special technology accommodation will only increase as the population ages and as the Web becomes more engrained in the fabric of our society. In light of these trends, the current level of conformance among federal (23%) and commercial (17%) Web sites is discouraging at best. This situation is exacerbated by the fact that Web systems show a tendency to decline in accessibility over time. As the intricacies of Web content continue to rise (e.g., advanced multimedia formats) we should expect that it will become increasingly difficult to meet the accessibility requirements of all Web users.

On the positive side, the inclusion of accessibility features in familiar software packages and systems is becoming more and more common. For example, Windows XP contains standard accessibility features such as *StickyKeys* and *SoundSensor* that can be easily toggled on and off to

meet the needs of Web users with diverse abilities. IE can also be easily reconfigured through several accessibility options. Furthermore, latest versions of popular Web development systems, for instance *FrontPage* and *Dreamweaver*, contain functionality that assists in development of accessible Web content. Compliance evaluation systems such as *Bobby* (*WebXACT*) are also becoming more powerful and more popular in the Web-development community. These software trends facilitate accessibility from both the user and the developer end.

CONCLUSION

Beyond the legal obligations that exist, arguments for parity of access to Web-based content by individuals with imperfect abilities are easy to make on ethical grounds. Furthermore, there seem to be numerous business advantages related to accessibility conformance of Web systems. There are potential productivity gains from the aging workforce that can be realized through accommodations in Internet and intranet work environments. There are potential performance and maintenance benefits with regard to server-side systems, as well as simplified integration of newer client-side technologies. From the B2C perspective, there is the potential for expanded audience reach and for increased share of an underexploited market niche.

Unfortunately, recent studies have shown that conformance to accessibility standards is sorely lacking on the Web, and that existing Web sites actually *decrease* in accessibility over time. Overall, it appears that little attention is being paid to issues such as Section 508 or WCAG conformance. There is an obvious need for training and awareness of these issues within the Web development community. Web content should be designed from the start with accessibility features in mind, and this awareness should be carried through the entire life of the applications, if we are to protect a significant sector of our society from increasing marginalization.

REFERENCES

Asakawa, C. (2005). What's the Web like if you can't see it? In *Proceedings of the 2005 International Cross-Disciplinary Workshop on Web Accessibility*, (pp. 1-8).

Bergel, M., Chadwick-Dias, A., & Tullis, T. (2005). Leveraging universal design in a financial services company. In *ACM SIGACCESS Accessibility and Computing*, (Vol. 82, pp. 18-24).

Fox, S. (2004, March 25). *Older Americans and the Internet*. Pew Internet & American Life Project. Retrieved December 14, 2007, from http://www.pewinternet.org/PPF/r/117/report_display.asp

Hackett, S., Parmanto, B., & Zeng, X. (2003). Accessibility of Internet Web sites through time. In *Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 32-39).

Ivory, M., Mankoff, J., & Le, A. (2003). Using automated tools to improve Web site usage by users with diverse abilities. *IT & Society, 1*(3), 195-236.

Lazar, J., Dudley-Sponaugle, A., & Greenidge, K. (2004). Improving Web accessibility: A study of Web master perceptions. *Computers in Human Behavior, 20*(2), 269-288.

Lazar, J., & Greenidge, K. (2006). One year older, but not necessarily wiser: An evaluation of home page accessibility problems over time. *Universal Access in the Information Society, 4*(4), 285-291.

Loiacono, E., & McCoy, S. (2004). Web site accessibility: An online sector analysis. *Information Technology & People, 17*(1), 87-101.

Loiacono, E., McCoy, S., & Chin, W. (2005, January/February). Federal Web site accessibility for people with disabilities. *IT Pro*, 27-31.

Mates, B. (2006). *Adaptive technology for the Internet: Making electronic resources accessible to all* (online version). American Library Association. Retrieved December 14, 2007, from http://www.ala.org/ala/ProductsandPublications/editions/adaptivetechnology.htm

Moos, B. (2005, December 11). Ads target empty nests, full wallets. *WFAA.Com* Retrieved December 14, 2007, from http://www.wfaa.com/sharedcontent/dws/bus/stories/121105dnbusboomertising.29c7e45.html

Mosner, E., & Spiezle, C. (2003). *The convergence of the aging workforce and accessible technology*. Redmond, WA: Microsoft Press.

Steinmetz, E. (2006). Americans with disabilities: 2002. *Current populations reports* (pp. 70-107). Washington, DC: U.S. Census Bureau.

Taylor, H. (2000, June 7). *How the Internet is improving the lives of Americans with disabilities*. The Harris Poll #30, Harris Interactive. Retrieved December 14, 2007, from http://www.harrisinteractive.com/harris%5Fpoll/index.asp?PID=93

U.S. Department of Commerce. (2000). *Falling through the net: Toward digital inclusion*. Retrieved December 14, 2007, from http://search.ntia.doc.gov/pdf/fttn00.pdf

U.S. Department of Commerce. (2006). *2005 American community survey*. Retrieved December 14, 2007, from http://www.census.gov/acs/www/index.html

World Wide Web Consortium. (1997, May 6). *World Wide Web consortium (W3C) launches international Web accessibility initiative: W3C leads program to make the Web accessible for people with disabilities*. Retrieved December 14, 2007, from http://www.w3.org/Press/WAI-Launch.html

KEY TERMS

Adaptive/Assistive Technology: Technology that is aimed at providing independence to individuals with disabilities. The terms *adaptive* and *assistive* are often used interchangeably, but there is a fine distinction. Assistive technologies stand on their own, while adaptive technologies provide accessibility to existing mechanisms and systems which might otherwise be inaccessible.

Bobby (WebXACT): The most well-known tool to evaluate Web sites for compliance to Section 508 and WCAG standards. Bobby was launched in 1995, but it was re-implemented as WebXACT in May of 2005. WebXACT remains a free service available at http://webxact.watchfire.com/. Despite the change, it is still often referenced through its original *Bobby* handle.

Digital Divide: The concept that there exist inequities in access to public information and information technologies by certain segments of the population. These segments are defined by such attributes as income, race, education, and disability.

Electronic Curb Cuts: A term referring to adaptive technologies that are aimed at computing and information technologies. The term makes an analogy to the federally mandated removal of curbs at crosswalks to facilitate wheelchair access. Unexpected benefits of such modifications are often called the *curbcut advantage*.

Equality of Access/Equity of Access: The hypothetical ideal situation in which all citizens are afforded full and equal access to public information and information technology, regardless of situation, status, or ability.

Large-Print Access Systems: A type of adaptive technology that is intended to provide screen-based access to persons with impaired vision or with certain cognitive disabilities. These systems are based on enlarging the screen image in some way.

Screen-Reading Systems: Adaptive technologies that are designed to produce the speech output equivalent of text-based information. These systems can be efficacious with Web content only if the content is designed properly.

Section 508: The section of the Rehabilitation Act of 1973 that deals with information technologies and systems. This section requires that all information systems utilized by the U.S. Government agencies be accessible to disabled individuals.

Subsection 1194.22: The portion of Section 508 that covers Web-based (both Internet and intranet) applications. (See *Section 508*).

Universal Design: An approach to designing Web content where the intent is to provide access to the broadest range of clients, regardless of individual abilities, disabilities, circumstances, or environments.

WCAG (Web Content Accessibility Guidelines): A set of design guidelines championed by the World Wide Web Consortium aimed at increasing accessibility to content on the Web. These guidelines consider poor user environments as well as users of varying physical or cognitive abilities.

Developing a Web Service Security Framework

Yangil Park
University of Wisconsin – LaCrosse, USA

Jeng-Chung Chen
National Cheng-Kung University, Taiwan

INTRODUCTION

Web Service (WS) is an open standard software component that uses *Extensible Markup Language* (XML) functions to access and exchange data via networks in communicating with other WSs. In business transactions, Web Service Description Language (WSDL) is used to describe data and deliver all parameters, return values, and types. However, the convenience of using *WSDL* in business transactions also lets hackers snoop and analyze data easily. In addition, the lack of Web Service standards at present makes the *security* issue even more serious in business transactions using WS. This article proposes a high-level *security* model for the *security* problems of the applications. Unlike other studies in the field, this study is dedicated to provide a total solution consisting of technological, organizational, and managerial aspects when using WS. Therefore, the understanding and development of business behaviors is essential in this study. It first introduces current uses of WS. Then definitions of WS, WS security, and WS policy are reviewed. Finally, a WS security model is proposed and explained in the following examples.

BACKGROUND

The Internet is becoming a global common platform where organizations and individuals communicate with each other to carry out various commercial activities and to provide value-added services (Wang, Huang, Qu, & Xie,, 2004). According to the World Wide Web Consortium (W3C, 2004), Web service means that an application program can be described and invoked, and made use of Uniform Resource Identifier (URI) to distinguish via XML. The *application program interfaces* define ways of contacting and supporting other application programs in order to urge directly through the protocol conforming to the Internet with the information of *XML* form. Web service technology allows users to customize services according to their own needs. This allows businesses to interact more accurately and efficiently with customers, cooperative enterprises, and suppliers.

According to a recent Gartner survey, 10 percent of midsize businesses cited using Web services for some production applications, while 47 percent of midsize stated that they plan to deploy Web services (Browning & Anderson, 2004). Also, Gartner Dataquest predicted that Web services will grow from $56 billion worldwide in 2003 to $283 billion worldwide in 2007 (Varbusiness, 2005). By then, Web services will take hold as a competitive differentiator in business relationships and product innovation (Andrews, 2003; Fensel and Bussler, 2002). Enterprises that want to remain competitive will need to use Web services to provide commonly requested data to their partners (Andrews, 2003) and, therefore, Web service technology will no longer offer a competitive advantage to enterprises. It is necessary for them to become competitive (Wiseth, 2004).

When an enterprise has some basic Web services, the high-level functional demands such as the service *security*, the service composition, and the service semantics will increase, and they are critical to the success of deploying Web services (Wang, et al., 2004). Presently, service-oriented architectures use the Web services to work on the business transaction based on the *Web Service Description Language* (WSDL). During the process of digital data delivery hackers are capable of obtaining the parameters. This data can be decoded and analyzed, which could cause a threat to a business. Communication over Web services is done by using a *Simple Object Access Protocol* (SOAP) that is associated with other programs that are built on *XML*. *SOAP* transfers everything over the HTTP, allowing data to pass through firewalls via a TCP port. This enables information to travel through firewalled ports, but this kind of firewall penetration also adds another *security* concern.

Industry observers have said the biggest obstacle to the wider adoption of Web services has been security concerns (Geer, 2003). Thus, security is critical to the adoption of Web services by enterprises, but, as it stands today, the Web service framework does not meet basic *security* requirements (Wang, et al., 2004). With this in mind, this study attempts to develop and implement a web services *security* model and discuss the essential security problems of this structure.

WEB SERVICE SECURITY FRAMEWORK

Web Service

Web services are a relatively new and emerging technology, they allow enterprises to share application logic and data using the standardized data and messaging formats, namely *XML* and *SOAP*. Web services can be accessed through the Internet and are based on existing communication protocols, such as HTTP. However, a standard definition of Web services has yet to be resolved. According to W3C (2004), "a Web service is a software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-processable format (specifically *WSDL*)." In addition, IBM (n.d.) defines "Web service is a new breed of Web application that is self-contained and self-describing, and which can provide functionality and interoperation ranging from the very basic to the most complicated business and scientific processes." Gartner Group (Natis, 2003) defines Web services as a software module that represents a business function (or a business service) and can be accessed by another application (a client, a server, or another Web service) over public networks using generally available ubiquitous protocols and transports (i.e. SOAP over HTTP).

From the above-mentioned definitions, one can understand that the purpose of the Web services is providing an interoperable interface for application-to-application interaction over public networks. The Internet protocols and standards on which Web Services are based are:

- *Simple Object Access Protocol* (SOAP) that enables communications among Web services.
- Web Services Description Language (WSDL) is the *XML* language that providers use to describe their Web Services.
- Universal Description, Discovery and Integration (UDDI) directories enable brokers to register, categorize, and list Web services and requesters to find them.

In traditional application development, programmers spend a great amount of time to tell an application to find another application, and this connection may require maintenance over the course of its lifetime; again, using human application developers. Web services provide certain secure protection in terms of quality, security, data integrity, and complicated transaction. Table 1 presents the comparison of traditional applications and Web services.

Web Service Security

Security is considered the biggest obstacle to general adoption of Web services (Geer, 2003). Cross-enterprise exchange of information over the Internet is vital but may have *security* implications. *Security* issues over the Internet are important, because the Internet is an insecure and non-trustable public network infrastructure, prone to malicious attacks by professional and amateur intruders (Wang, et al., 2004). According to O'Neill (2002) and O'Neill, White, & Watters, (2003), Web services security challenges are as follows:

- The challenge of *security* based on the end user of a Web Service: In order for users to gain access to a

Table 1. Comparison of traditional applications and Web services

Traditional application	Web services
Centralized	Decentralized
Contained and controlled	Open and unmonitored
Limited, defined user base	Unknown, unlimited user base
Secure (risk minimized)	Exposed (open to random events)
Proprietary	Shared
Fixed, well-defined, compiled	Built dynamically, on-the-fly
Incremental scale based on known demand	Unlimited scale, based on unknown, unpredictable demand
Staged, periodic changes	Continuous, *ad hoc* changes
Implementation technologies are all the same or compatible	Heterogeneous in implementation technologies

(Modify from Ratnasingam, 2002; Yang, 2002)

Web service where logical access control is required, there must be some forms of authentication, such as a username combined with a password, or a digital certificate (secured by a pass phrase) and based on the information to make authorization decisions.

- The challenge of maintaining *security* while routing between multiple Web services: When SOAP message routing between Web services, the requirements for confidentiality can be implemented using SSL between Web services.
- The challenge of abstracting *security* from the underlying network: Web service security not only relies on Web security, but also includes *SOAP* services security, such as *SOAP* messages to be secured between Web services.

In addition, as the incidents of attacks increases, the threats of Web service *security* are emerging. These threats can roughly divide into two parts as follows: (King, 2003; Morrison, 2004; Ratnasingam, 2002; Slewe and Hoogenboom, 2004; Treese, 2002):

- About network *security*:
  1. A shift from generic attacks to more sophisticated and well targeted attacks.
  2. Increasing propagation speed and volume of virus attacks.
  3. Increasing speed of the release new of viruses and attacks after the detection of a *vulnerability.*
  4. Increase in identity or federating identity fraud.
  5. Network eavesdropping.
  6. Unauthorized access.
  7. The *security* of the network in which the runtime is deployed.
- The security of information that is shared between corresponding parties at run time:
  1. Parameter manipulation.
  2. Disclosure of configuration data.

Standards and Technologies in Web Service Technology

When using Web services as the transaction platform in enterprises, the *security* issues will often be put forward and regarded as the credibility that queries the Web services, and influences the development of the Web services. Technology vendors and standards organizations have recognized the risks posed by Web services *security vulnerabilities*. In standards bodies and in various consortia, they are working to formulate new technologies and better practices designed to make Web services more secure. To date, there are numerous standards under development that aim to address a broad range of issues, such as privacy, trust, and reliability of Web services.

- WS-Security
  WS-*Security* was originally designed by VeriSign, Microsoft, and IBM before being submitted to the Organization for the Advancement of Structured Information Standards (OASIS). It provides a way for Web services to attach *security* data to the header of *SOAP* messages, and works with several different security models via *SOAP* extensions.
- *XML* Key Management Specification (XKMS)
  It was originally designed by VeriSign, Microsoft, and Web Methods before being submitted to the W3C organization. XKMS does offer a simplified approach to integrating public key management capabilities with applications.
- *XML* Access Control Markup Language (XACML)
  The XACML was developed by the Organization for the Advancement of Structured Information Standards (OASIS) in 2003. XACML is designed to express access control rules in *XML* format, and integrate seamlessly with SAML and, also, with *XML* Digital Signature.
- Extensible Rights Markup Language (XrML)
  XrML can provide the service about DRM, Metadata, content management, and content transmitting of the digital content.
- Secure Assertion Mark-Up Language (SAML)
  The SAML was developed by OASIS in 2002. SAML defines an *XML* schema that allows trust assertions (*authentication*, authorization, or attribute) representation in XML and request/response protocols to perform *XML authentication*, authorization, and attribute assertion request.
- *XML Encryption* (XML Enc)
  The XML *Encryption* was developed by W3C in 2002. XML *encryption* provides a model for *encryption*, decryption, and representation of full XML documents, single *XML* elements in an XML document, contents of an XML element in an XML document, and arbitrary binary content outside an XML document.
- XML Digital Signature (XML Dig)
  The *XML* Digital Signature was developed by W3C in 2002. It defines syntax and processing rules for representing digital signatures how to digitally sign XML contents and how to represent the resulting information according to an *XML* schema.

Information Security Policy

In an ideal world, configuration, operation, and management of *security* functionality should allow to specify business

Table 2. Security policy and technology trust mechanisms

Security Policy	Technology trust mechanisms
Information Custodianship	Authentication; Authorization
Physical Access Security	Authentication; Authorization
Information Security Administration Functions	Authentication; Authorization
Logon Security	Authentication
Transaction Controls and Database Security	Integrity; Authentication; Content inspection; Confidentiality; Authorization; Non-repudiation
Agency Security Management	Monitoring; Auditing; Exception management
Information Recovery	Backup and recovery procedures
Data Exchange Agreements	Authentication; Authorization
Vendor/Contractor Agreements	Authentication; Authorization
Employee/Agent Responsibilities	Authorization; Exception management (Monitoring and alerting); Authentication
Sensitive information	Authorization; Authentication

oriented terms by non-specialist staff (Kearney, Chapman, Edward, Gifford, & He, 2005). Such tasks should be performable by staffs involved in business operations rather than technical experts. Therefore, the person responsible must be able to lay down rules that are meaningful both to a human being and to the Web service management software. Such rules are known as policy (Kearney, et al., 2004). Web services *security* provides developers, application architects, and security professionals the need to build *security* policies and strategies from the ground up in a Web Services environment.

According to Panko's (2004) definition, the *security* policies specify at a broad level what should be done in terms of security. In addition, Rees, Bandyopadhyay, & Spafford, (2003) defines that the *security* policies are generally high-level, technology neutral, and concern risks and set directions and procedures. The *security* policy should be applied to all in existence and future technology infrastructures including the provision of management standards for information departments and non-information departments in enterprises to follow.

Table 2, summarized from literature and cases, shows how a *security* policy can be implemented by applying technology trust mechanisms (Natoli, 1997). By doing so, it is expected to increase the enterprise *security* when using Web services to communicate with each other.

Web Service Security Model

To provide a comprehensive model of *security* functions and components for Web services, the integration of currently available processes and technologies with the evolving *security* requirements of future applications is required. It demands unifying concepts requiring solutions to both technological (secure messaging) and business processes (policy, risk, and trust). It also requires coordinated efforts by platform vendors, application developers, and network and infrastructure providers and customers. Therefore, we suggest that the development of Web services *security* model categorizes into three aspects: organization, technology, and management (see Figure 1). *Security* capabilities of Web services in this model are: authorization, confidentiality, *authentication*, integrity, content inspection, and *encryption*.

The security model shown in figure 2 details how to secure critical components such as Web services, application servers, and networks.

The model includes three parts: external, Web Service *Security* Integration (WSSI), and internal. External application, Web service, or user can request/response business internal Web services or application servers via firewall. Management console can set enterprise *security* policies to re-evaluate whether the enterprise's *security* policy can satisfy business requirements in the future technology infrastructures. In addition, *authentication*, authorization, confidentiality, content inspection, *encryption*, routing, and integrity mechanisms need to protect messages and transactions, including how to implement and communicate those mechanisms using

Figure 1. Development of Web service security model

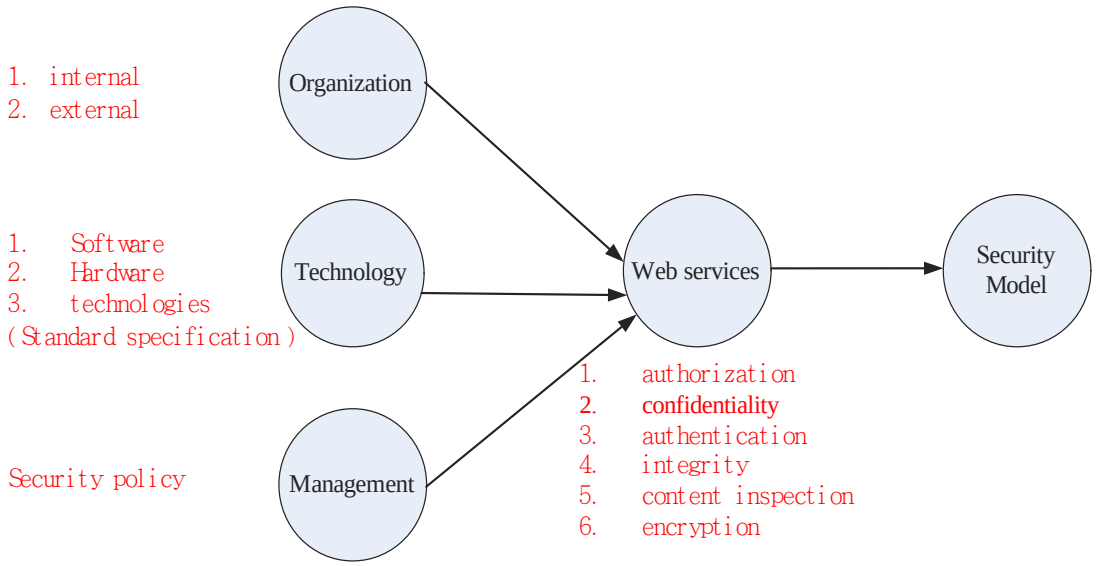
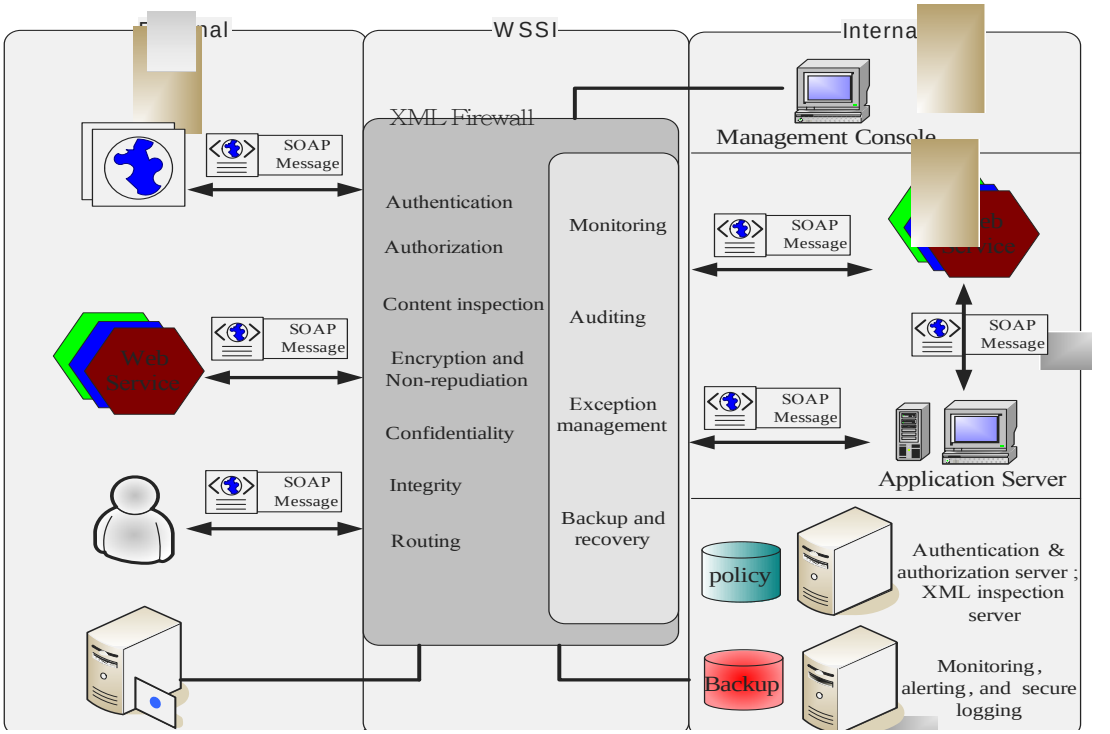


Figure 2. Web service security model



WS-Security, *XML Encryption*, *XML* Signature, SAML, WS-Routing, and XACML. The importance of auditing, monitoring, and exception management is also covered to ensure that data access and system execution are accurate, and data backup can protect data from irreparable loss or damage and reduce enterprise's risks.

FUTURE TRENDS

Web services are becoming a very serious part of development for businesses, which is why Web service *security* plays such an important role. The *security* infrastructure for Web services is growing rapidly due to the demand for privacy of users; sensitive information of users should be protected by *security* policies and a strong security system. The goal for this field of research is to provide a safer environment for users so that this service can be used without the worries of data leakage.

This calls for a significant amount of research that needs to be done in the Web services *security*, such as authorization and *authentication* in heterogeneous environments. For instance, understanding and analyzing similarities of distributed applications design would help to design more secure Web services applications.

CONCLUSION

Security requires all parties and systems to work in show to build a good protection, thus, secure Web services entail a macro solution. This paper suggests a high level solution for the security problems of the application to business-to-business behavior. We discussed the essential security problems of this structure, and then laid out solutions to the currently faced problems. The study also discussed the different platforms and policies of Web services security; however, further understanding of the many kinds of *security* infrastructures is suggested. The suggested *security* model shows that security policies and technology trust mechanisms are working successfully to prevent improper accesses and protect their assets. These security issues are a very important subject in the business-to-business communication sector due to possibility of great losses if data leakage exists.

REFERENCES

Andrews, W. (2003). Predicts 2004: Web Services. *Gartner research*. Retrieved May 20, 2005, from the World Wide Web: http://www4.gartner.com/DisplayDocument?ref=g_search&id=416001.

Browning, J.A., & Anderson, R.P. (2004). Adoption of Web Enablement Can Improve SMB Business. *Gartner research*. Retrieved May 20, 2005, from the World Wide Web: http://www4.gartner.com/DisplayDocument?ref=g_search&id=428824.

Fensel, D. & Bussler, C. (2002). The Web Service Modeling Framework WSMF, Electronic Commerce. *Research and Application*, 1, 113-137.

Geer, D. (2003). Taking Steps to Secure Web Services, *IEEE Computer*, 36(10), 14-16.

IBM. (n. d.). Standards and Web services. Retrieved May 23, 2005, from the World Wide Web: http://www-128.ibm.com/developerworks/webservices/standards/.

Kearney, P. (2005). Message level security for web services. *Information Security Technical Report*, 10(1), 41-50.

Kearney, P., Chapman, J., Edward, E., Gifford, M. & He, L. (2004). An overview of Web Services security. *BT Technology Journal*, 22(1), 27-42.

King, S. (2003). Threats and Solutions to Web Services Security. *Network Security*, 2003(9), 8-11.

Morrison, K. S. (2004). Security behind the Firewall: The Challenge for Web Services. *Business Integration Journal*, 77-79.

Natis, Y. V. (2003). Service-Oriented Architecture Scenario. *Gartner research*. Retrieved May 23, 2005, from the World Wide Web: http://www.gartner.com/DisplayDocument?ref=g_search&id=391595.

Natoli, J. (1997). Information Security Policy. Retrieved June 4, 2005, from the World Wide Web: http://www.oft.state.ny.us/policy/tp_971.htm.

O'Neill, M. (2002). Is SSL Enough Protection for Web Services? *eAI Journal*, 31-33.

O'Neill, M., White, A. & Watters, P. A. (2003). *Web services security, U.S.A.*. McGraw-Hill. Osborne, Media.

Panko, R. R. (2004). *Corporate computer and network security, U.S.A.*. Prentice Hall.

Ratnasingam, P. (2002). The importance of technology trust in Web services security. *Information Management & Computer Security*, 10(5), 255-260.

Rees, J., Bandyopadhyay, S. & Spafford, E. H. (2003). PFIRES: A Policy Framework for Information Security. *Communications of the ACM*, 46(7), 101-106.

Slewe, T. & Hoogenboom, M. (2004). Who will rob you on the digital highway. *Communications of the ACM*, 47(5), 56-60.

Treese, W. (2002). XML, WEB SERVICES, AND XML, netWorker, 6(3), 9-12.

Varbusiness, (2005). http://www.varbusiness.com/sections/research/research.jhtml.

W3C. (2004).Web Services Architecture, Retrieved May 20, 2006, from the World Wide Web: http://www.w3.org/TR/ws-arch.

Wang, H., Huang, J.Z., Qu, Y. & Xie, J. (2004). Web services: problems and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 1(3), 309-320.

Wiseth, K. (2004, July/August). Scoring with Web Services. *ORACLE Magazine*, 30-39.

Yang, A. (2002). Web Services Security. *eAI Journal*, 19-23.

KEY TERMS

B2B (Business-to-Business): Refers to one business communicating with or selling to another.

HTTP (HyperText Transfer Protocol): Protocol used to transfer hypertext requests and information between servers and browsers.

Port: A number from 0 through 1023 used to identify a network service on an IP network.

Security: Freedom from risk or danger; safety.

Security Policy: A security policy is a generic document that outlines rules for computer network access. It determines how policies are laid out and some of the basic architecture of the company security environment.

TCP (Transmission Control Protocol): TCP ensures that all data arrive accurately and 100% intact at the other end.

Web Service: An open standard software component that uses XML functions to access and exchange data via networks to communicate with other Web Services.

Web Service Description Language (WSDL): Used to describe data and deliver all parameters, return values, and types.

Developing an Effective Online Evaluation System

Martha Henckell
Southeast Missouri State University, USA

Michelle Kilburn
Southeast Missouri State University, USA

David Starrett
Southeast Missouri State University, USA

INTRODUCTION

As with any new program, the chance of failure runs high and distance education, in comparison with the longevity of traditional education, is considered relatively new. Still, distance education appears to be here to stay. In fact, a 2000 market survey found that over 94% of all colleges were either offering or planning to offer distance education courses (Twigg, 2001). With this much interest and popularity, the need for policies to regulate distance education program practices should be recognized by all participating institutions of higher education (Czubaj, 2001). While students appear to be more focused on the conveniences that distance education provides, universities are more attentive to the need for offering a valid learning alternative. Higher education enrollments have shown upward movement and this has, to a degree, been attributed to the adult learners' interest in, and availability of, distance education (Boettcher, as cited by Worley, 2000). Change in the enrollment demographics and the offering of distance education programs stimulates the need for new decisions by academic administrators for quality and accreditation purposes (Shea, Motiwalla, & Lewis, 2001; Tricker, Rangecroft, Long, & Gilroy, 2001).

One of the first steps toward ensuring success of distance education programs is identifying the requirements of all those involved. Student needs are to receive a quality education; faculty needs are to have at their disposal (and to use) the knowledge and means to provide this education; and institution needs are to assess that students receive a quality education and to provide faculty with the resources for student educational needs to be met. One of the problems that could harm distance learning or prevent it from being all that it can be is the lack of a good evaluation system. The focus of this article will be to identify and describe, from the literature, the components of an effective evaluation system. Armed with this information, administrators will be able to make better program decisions.

BACKGROUND

The need for information in any decision-making process is crucial. The newness of distance education makes the need for related information even more critical. One of the most popular methods for amassing information in higher educational settings is by performing evaluations. According to Patton (1997), education has a long history of using evaluations. Users of this data have their own purposes in mind. Students are seeking affirmation that the course contains relevant content, the instructor teaches effectively, and the course will help them reach their long-term goals (McKeachie, 1996; Spencer & Schmelkin, 2003; Willis, 1993). Faculty will have access to feedback that can help guide them in their teaching. Job performance reviews can be gleaned from either an administrator or student evaluation of faculty (Algozzine et al., 2004; Chen & Hoshower, 1998; Halpern & Hakel, 2003; McKeachie, 1996; Spencer & Schmelkin, 2002; Willis, 1993). Critical to institutional administrators is the collection of information that relates to whether or not institutional strategic goals are being accomplished. Decisions as to the potential development of a distance program (Willis, 1993) and changes to support programs (i.e., bookstore, tutoring, etc.) that support this program can be made. Academic administrators use evaluation data as one means to judge teaching performance (Emery, Kramer, & Tian, 2003; Neumann, 2000; Willis, 1993). Whether appropriate or not, decisions on tenure and promotion are frequently based on this information (Algozzine et al., 2004; Chen & Hoshower, 1998; Halpern & Hakel, 2003; McKeachie, 1996; Spencer & Schmelkin, 2002; Willis, 1993). Regardless of the reason for information collection, quality information can be gathered only with the use of a quality instrument. Reliability and validity of the information is always in the forefront of concerns when conducting an evaluation (Scanlan, 2003; Griffin, Coates, McInnis, & James, 2003; Marshall, 2000; Regalbuto, 1999; Achtemeier, Morris, & Finnegan, 2003). To fur-

ther perpetuate this problem, unless faculty believe in the validity of the information collected, change is not likely to occur (Reid & Johnston, 1999); unless students believe their responses will provide a reward, less-than-valid response may be supplied (Chen & Hoshower, 1998). Differences associated between distance and traditional courses can hinder the desired outcome of validity, emphasizing the evidence that an alternate evaluation instrument is required. Despite the distinctiveness of distance education, many universities continue to use traditional course student evaluation instruments to evaluate distance learning courses (Achtemeier et al., 2003). To increase the reliability and validity of evaluation data, an evaluation instrument designed to represent distance education uniqueness would be required (Henckell, 2007). At the very least, alterations or amendments are required to take a well-designed traditional evaluation instrument and make it valid for evaluating distance education courses (Holcomb, King, & Brown, 2004; Shuey, 2002; Willis, 1993). A system contains parts that, when placed together, represent and share a relationship to the whole or what Marshall (2000) describes as a model. As with traditional courses, student evaluations are a vital part of the system for assessing distance education programs. Information collected from student evaluations should not stand alone. Administrative reviews are also necessary to provide a more accurate picture of performance. With each type of evaluation, there is the need to review the components of the evaluation process and what can positively or negatively affect these events. With the recommendations provided in this article, changes can be made to perfect the components used in an evaluation system. Improvements to current evaluation systems will hopefully lead to a greater buy-in of the system by students, faculty, and administrators.

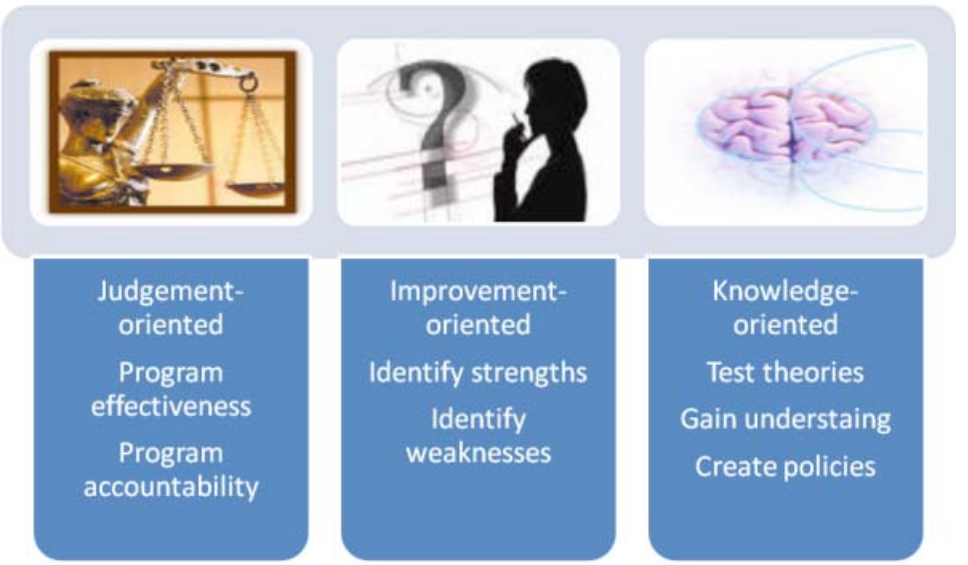
COMPONENTS OF A COMPREHENSIVE DISTANCE EDUCATION EVALUATION SYSTEM

Involved in the building of an evaluation system is an evaluation plan. This plan must recognize purposes and rationale of an evaluation and identify how, what, and when to evaluate (Henckell, 2007). Evaluation methods, styles, and strategies can then be determined (Robson, 2000). University administrators, academic administrators, faculty, and students are the four parties that should be included in all evaluation systems of distance education courses (Willis, 1993).

How to Evaluate

First and foremost, the purpose of the evaluation must be identified in order to know the right information for decision making will be present. The cynosure of an evaluation, according to Patton (1997), is its intended use. Data gathered

Figure 1. Intended use



from the process can be utilized to relinquish judgments, expedite improvements, and create knowledge. Patton's judgment-oriented evaluation could be used to focus on program effectiveness and accountability. His second evaluation type, improvement oriented, could be used to enhance programs by identifying strengths and weaknesses. Both of these evaluation types required that a decision or action follow the use of these instruments. In contrast, the intent of the third evaluation type is to simply increase knowledge. Uses of the knowledge-oriented evaluation include testing theories, gaining understanding, or creating policies. For representation of these intended uses, see Figure 1. Regardless of which type of evaluation is used, if the collection of valid information fails to be used, one has to question the value of the exercise.

When and What to Evaluate

Society has held a high regard of universities for years; the public, in more recent times, has demanded more accountability (Sutherland, 2000). Evaluations are one of the main methods used to prove courses are being monitored and institutional standards are maintained. Two of the most frequent evaluation activities are the end-of-course student evaluation and an evaluation conducted by an academic administrator. As part of the evaluation process, Benigno and Trentin (2000) recommend that students enrolled in a distance education course complete a pre-course survey. This procedure would allow instructors to discover distinct characteristics of the student such as previous experience with distance education, the student's learning environment that may adversely affect the student, and technology skills. Vrasidas, Zembylas, and Chamberlain (2003) agreed with Benigno and Trentin but suggested also assessing student perceptions and attitudes of online courses. In addition to evaluating the students prior to course activity, three other evaluations have been recommended to take place throughout the course. During the later part of the second or third week, it is thought that students should have had enough course experiences to be able to provide valid information on the course, instructor, and medium (Henckell, 2007). What is crucial about conducting the evaluation during this time is collecting data prior to the course drop deadline. This may be the only opportunity to gather information regarding issues that forced the students to drop the course. Evaluation results are possibly skewed if information is provided by only the successful students (Phipps & Merisotis, 1999). While this early evaluation can aid in improving teaching, mid-semester evaluations also work toward this purpose. At mid-semester, more detailed and useful information is available (Laverie, 2002). Using this timeline allows time for the instructor to make improvements, benefiting the currently enrolled students. As practice has demonstrated, the third evaluation time recommended is slated for the end of the course. An overall impression can be provided during this time, but the downside is that the timing is not very favorable for accurate responses. Most students are extremely overwhelmed at this time and not in the best frame of mind or attitude to respond in a manner that may provide an authentic picture of course experiences. By conducting and combining the results of mid- and end-semester evaluations, a more valid representation of the course will transpire. To prove or refute student evaluation responses, another source for information should be sought, and academic administrators, using the right tools and processes, should be able to meet this need. Academic administrators experience a more difficult time evaluating faculty teaching distance education courses. Faculty evaluations generally require a visit to the classroom for observations and the writing of an evaluation report that sometimes includes results from scores on faculty presentation skills, professionalism, material coverage, media usage, and general comments (Tobin, 2004). Challenges for evaluating courses at a distance must be surmounted. Questions administrators must now seek answers to during the distance education evaluation process include:

1. How is a classroom visit to occur if the course is asynchronous?
2. What preparations are needed to review class discussions?
3. How can the evaluator ascertain the classroom discussion quality and the instructor's involvement?
4. How can the instructor's demeanor be evaluated?
5. Where, how often, and what should constitute a visit in the course Web site?
6. Is more multimedia required for the online instruction?
7. And most importantly, how can one evaluate an online course when one has never experienced the process? (Tobin, p. 76, as cited by Henckell, 2007)

Regalbuto (1999) also provided valuable and pertinent questions for use by academic administrators when assessing distance education. He suggested:

1. Was innovation present in the teaching style?
2. Was the learning competence equal or superior to that of a traditional classroom?
3. Were the students engaged in the material?
4. Were there interactions between professors and their students, and between the students themselves?
5. Was technical support readily available?
6. For online programs that are more extensive, such as entire degree programs, are the signs of academic maturity present?

Administrators who have never experienced the online process have a responsibility to immerse themselves in distance education literature. To get a true understanding of distance courses, administrators should go one step further and audit the type of distance education course they will be responsible for evaluating. Communication plays a very important role in distance education. As to how to evaluate asynchronous type courses, reviewing communications that occur between students and faculty can provide valuable clues as to the events and activities that occurred during the course, as well as how well the events and activities were received. Communication can also provide information as to the instructor's demeanor, whether students appeared to be engaged in the material, and whether or not academic maturity was present. Course materials, as well as lesson plans, can be reviewed in response to whether innovation was present in the teaching style or more multimedia should be required. Comments by the students regarding this subject would also assist in this determination. Special elements present in distance education courses require that the traditional course evaluation system be altered in order to prevent the loss of valuable information. The recommended evaluation of items known to affect student success are listed as (Benigno & Trentin, 2000):

"(a) individual characteristics; (b) level of participation; (c) collaborative and content message analysis; (d) interpersonal communication; (e) available support resources (i.e., bookstore, technology, registrar, etc.); (f) reaction to the methodological approach (as opposed to seeking judgment as to whether the student thought it was a correct instructional method); (g) usefulness of the learning material (as opposed to seeking judgment as to whether the student thought it was the correct learning material); (h) learning environment (i.e., local, virtual, social, etc.); (i) communication through technological means; and (j) value placed on the course as opposed to value placed on traditional learning." (as cited by Henckell, 2007, p. 70)

FUTURE TRENDS

Distance education grew rapidly after technology advancements (Spooner et al., 1999; Worley, 2000; Miller & King, 2003; Holcomb et al., 2004), and it has been envisioned that distance education will continue to experience growth (Allen & Seaman, 2004; Phipps & Merisotis, 1999). With such rapid growth, there has been little time devoted to creating policies to ensure quality programs. Now that programs are established, issues are becoming apparent and more universities are taking the time to create policies that will make the programs stronger. The development and use of an evaluation system is one policy, but others are needed. Future research is needed to identify what policies are being developed and the negative or positive effect of these policies on the programs, once in place. Evaluations serve a variety of purposes, and when the information collected is acted upon, students, faculty, academic administrators, institutional administrators, and distance education programs benefit. One of the most desired results of conducting evaluations is for course and content improvement. Research is needed on whether or not change is occurring based on evaluation results and methods employed by administrators to ensure that the evaluation results are being used.

CONCLUSION

Applied research has shown distinguishable differences between traditional and distance education courses that warrant the development of similar, yet separate, evaluation systems. In order for an evaluation system to meet its intended use, those involved must believe in the system. More than an end-of-course evaluation is needed to meet the challenges and uniqueness of distance education. The development of a distance learning evaluation system or, at a minimum, the revamping of the traditional system is warranted. Information provided in this article can be used to assist the reader in building a more comprehensive evaluation system that will, in turn, help improve distance education programs.

REFERENCES

Achtemeier, S.D., Morris, L.V., & Finnegan, C.L. (2003). Considerations for developing evaluations of online courses. *Journal of Asynchronous Learning Networks, 7*(1), 1-13.

Algozzine, B., Beattie, J., Bray, M., Flowers, C., Gretes, J., Howley et al. (2004). Student evaluation of college teaching: A practice in search of principles. *College Teaching, 52*(4), 134-141.

Allen, I.E., & Seaman, J. (2004). *Entering the mainstream: The quality and extent of online education in the United States, 2003 and 2004.* Needham, MA: Sloan-C.

Benigno, V., & Trentin, G. (2000). The evaluation of online courses. *Journal of Computer Assisted Learning, 16,* 259-270.

Bogdan, R.C., & Biklen, S.K. (1998). *Qualitative research for education: An introduction to theory and methods* (3rd ed.). Needham Heights, MA: Allyn & Bacon.

Chen, Y., & Hoshower, L.B. (1998). Assessing student motivation to participate in teaching evaluations: An application of expectancy theory. *Issues in Accounting Education, 13*(3), 531-548.

Czubaj, C.A. (2001). Policies regarding distance education. *Education, 122*(1), 119-122.

Emery, C.R., Kramer, T.R., & Tian, R.G. (2003). Return to academic standards: A critique of student evaluations of teaching effectiveness. *Quality Assurance in Education, 11*(1), 37-46.

Fowler, F.C. (2000). *Policy studies for educational leaders.* Upper Saddle River, NJ: Prentice Hall.

Griffin, P., Coates, H., McInnis, C., & James, R. (2003). The development of an extended course experience questionnaire. *Quality in Higher Education, 9*(3), 259-266.

Halpern, D.F., & Hakel, M.D. (2003). Applying the science of learning to the university and beyond: Teaching for long-term retention and transfer. *Change, 35*(4), 36-41.

Henckell, M. (2007). *Evaluating distance education: The student perspective.* Doctoral Dissertation, University of Missouri, USA.

Holcomb, L.B., King, F.B., & Brown, S.W. (2004). Student traits and attributes contributing to success in online courses: Evaluations of university online courses. *Journal of Interactive Online Learning, 2*(3), 1-17.

Laverie, D.A. (2002). Improving teaching through improving evaluation: A guide to course portfolios. *Journal of Marketing Education, 24*(2), 104-113.

Marshall, G. (2000). Models, metaphors and measures: Issues in distance learning. *Education Media International, 37*(1), 2-8.

McKeachie, W. (1996). *The professional evaluation of teaching: Student ratings of teaching.* Occasional Paper No. 33, American Council of Learned Societies, USA.

Miller, T.W., & King, F.B. (2003). Distance education: Pedagogy and best practices in the new millennium. *International Journal of Leadership in Education, 6*(3), 283-297.

Neumann, R. (2000). Communicating student evaluation of teaching results: Rating interpretation guides (RIGS). *Assessment & Evaluation in Higher Education, 25*(2), 121-134.

Patton, M.Q. (1997). *Utilization-focused evaluation* (3rd ed.).Thousand Oaks, CA: Sage.

Phipps, R., & Merisotis, J. (1999). What's the difference? A review of contemporary research on the effectiveness of distance learning in higher education. *Change, 31*(3), 12-17.

Regalbuto, J. (1999, December 7). *Teaching at an Internet distance: The pedagogy of online teaching and learning.* Retrieved February 5, 2005, from http://www.vpaa.uillinois.edu/reports_retreats/tid_report.asp?bch0

Reid, D.J., & Johnston, M. (1999). Improving teaching in higher education: Student and teacher perspectives. *Educational Studies, 25*(3), 269-281.

Robson, J. (2000). Evaluating on-line teaching. *Open Learning, 15*(2), 151-172.

Scanlan, C.L. (2003). Reliability and validity of a student scale for assessing the quality of Internet-based distance learning. *Online Journal of Distance Learning Administration, 6*(3). Retrieved February 5, 2005, from http://www.westga.edu/~distance/ojdla/fall63/scanlan63.html

Shea, T., Motiwalla, L., & Lewis, D. (2001). Internet-based distance education—the administrator's perspective. *Journal of Education for Business, 77*(2), 112-117.

Shuey, S. (2002). Assessing online learning in higher education. *Journal of Instruction Delivery Systems, 16*(2), 13-18.

Spencer, K.J., & Schmelkin, L.P. (2002). Student perspectives on teaching and its evaluation. *Assessment & Evaluation in Higher Education, 27*(5), 397-409.

Sutherland, T. (2000). Designing and implementing an academic scorecard. *Accounting Education News,* (Summer), 11-14.

Tobin, T. (2004). Best practices for administrative evaluation of online faculty. *Online Journal of Distance Learning Administration, 7*(2). Retrieved February 5, 2005, from http://www.westga.edu/~distance/ojdla/summer72/tobin72.html

Tricker, T., Rangecroft, M., Long, P., & Gilroy, P. (2001). Evaluating distance education courses: The student perception. *Assessment & Evaluation in Higher Education, 26*(2), 165-177.

Twigg, C.A. (2001). *Quality assurance for whom? Providers and consumers in today's distributed learning environment.* Retrieved April 14, 2005, from http://www.center.rpi.edu/PewSym/mono3.html

Vrasidas, C., Zembylas, M., & Chamberlain, R. (2003). Complexities in the evaluation of distance education and virtual schooling. *Educational Media International, 40*(3/4), 201-208.

Willis, B. (1993). *Distance education: A practical guide.* Englewood Cliffs, NJ: Educational Technology.

Worley, R.B. (2000). The medium is not the message. *Business Communication Quarterly, 63*(3), 93-103.

KEY TERMS

Applied Research: Research efforts that seek findings that can be used directly to make practical decisions about or improvements in programs and practices to bring about change with more immediacy (Schein, as cited by Bogdan & Biklen, 1998).

Distance Education: Education or training courses delivered to remote (off-campus) sites via audio, video (live or prerecorded), or computer technologies, including both synchronous (i.e., simultaneous) and asynchronous (i.e., not simultaneous instruction) (Distance Education, 2003).

Evaluation: The systematic determination of merit, worth, and significance of some object (Stufflebeam, as cited by Fowler, 2000).

Evaluation System: A devised system that outlines in a plan what, when, and how courses are to be assessed (Benigno & Trentin, 2000; Robson, 2000).

Improvement-Oriented Evaluation: Formative evaluation directed toward improving what is evaluated (Patton, 1997).

Judgment-Oriented Evaluation: Summative evaluation in which judgments are made on value or worth (Patton, 1997).

Knowledge-Oriented Evaluation: Evaluation used to test theories, gain a better understanding, and create policies (Patton, 1997).

Policy: "Policy as a chain of decisions stretching from the statehouse to the classroom is a by-product of [many] games and relationships; no one is responsible for the whole thing" (Firestone, as cited by Fowler, 2000).

Student Evaluations: Forms specifically designed to measure observed teaching styles or behaviors (Wright & Neil, as cited by Chen & Hoshower, 1998). Student evaluations are typically administered at the end of the course (Algozzine et al., 2004; Neumann, 2000).

Traditional Course: Course with no online technology used; content is delivered in writing or orally (Allen & Seaman, 2004).

Developing the Enterprise Architect Perspective

Brian H. Cameron
The Pennsylvania State University, USA

INTRODUCTION

Enterprise systems design, implementation, and integration are focal points for business and information technology. Businesses must change processes, environments, and technologies as organizations strive to become more integrated and break down traditional silos of information systems and responsibility. These challenges require a new type of technical professional: one with the training and perspective of an enterprise architect with general technical expertise as well as business strategy and planning skills. Some college and university programs have risen to this challenge in recent years, and the joint ACM/Association for Information Systems Task Force developed the MSIS curriculum model to establish the fundamentals of enterprise information systems in response to the increasing demand for university-trained graduates in an information economy (Gorgone, Gray, & Feinstein, 2000). Recently, the Association for Open Group Enterprise Architects called for industry and academia to work together to craft new enterprise systems curricula that are relevant to today's global business environment and developed from the perspective of an enterprise architect.

Today's globally competitive environment requires technical professionals to move beyond technical expertise and contribute to the strategy and development of dynamic IT systems that are able to support changing business objectives. To be prepared to meet such expectations, IT students must have broad experience in the design, implementation, and integration of such systems. This education is typically offered in a layered fashion, teaching students about databases, networks, and applications in different courses devoted to single topics (Nickerson, 2006). While this method allows universities to assign faculty with specific expertise to particular courses, it does not adequately prepare students for the work environment of the enterprise architect, where all of these different layers must be combined to support and align with business strategy. Students trained in a specific, narrow layer may fail to anticipate certain trends or requirements, such as a database designer overlooking the need for remote replication (Nickerson).

To meet this need, many information technology programs are incorporating enterprise systems curricula for senior students. These courses are often referred to as "capstones" in the curriculum, and must focus on a wide variety of educational goals including understanding the enterprise as a whole, understanding how technology can provide a competitive advantage, learning to design complex integrated systems, learning concepts underlying technical systems integration, learning how to assess the requirements of an integrated system, and learning how enterprise architecture design is practiced as a profession.

BACKGROUND

Enterprise architecture education is particularly important when trying to meet current business objectives. Several prestigious consulting groups, including IBM and Forrester, have noted a major shift in most technology-centric businesses since 2005 toward service-oriented architectures (SOAs; Boyle & Strong, 2006; Seethamraju, 2007). An SOA is the practice of sequestering the core business functions into independent services that typically do not change frequently. These services can then be combined to create composite applications that can be easily reconfigured to meet the changing needs of the organization. This new paradigm in enterprise systems development and integration highlights the demand for enterprise architects who can understand and align business goals with a technical strategy and architecture capable of supporting current and future needs. SOA does not represent the entire scope of responsibilities of the enterprise architect—it is simply one method of the overall goal of aligning the strategic vision of the business with its information technology infrastructure (Cannon, Klein, Koste, & Magal, 2004; Davis, 2004; Mulder, Lidtke, & Stokes, 1997).

In spring 2007, the Information Technology Association of America (ITAA) identified the need to double the number of graduates in science, technology, engineering, and math over the next 10 years to maintain U.S. information technology competitiveness. Specifically, ITAA (2007) identified "a commitment to the use of information technology to solve real customer problems now and in the future" as a primary goal of the U.S. education system—higher education in

particular. The organization is committed to enhancing IT education through better understanding of the IT workforce, and frequent assessment of the IT needs of industry.

The lack of well-educated IT workers is further emphasized when considering recent surveys predicting significant shortages in IT workers on the horizon. Despite the off-shoring of certain technology jobs, a large number of organizations in the United States are currently deficient in properly trained IT workers. A survey of Washington Trade Group members (over 14,000 companies) indicated that 36% of member companies had open technology jobs: *open* meaning the position has been posted and unfilled for more than 3 months (Barrett, 2007). The most common explanation for the open positions among executives interviewed is a lack of business literacy. In other words, applicants for the position are not sufficiently well rounded in business and technology. Most of these unfilled positions are for an employee who can interact with various groups within the organization, manage technology projects, analyze business needs and translate those needs into a technical solution, and become an effective bridge between functional business units and the technologists. In short, thousands of U.S. companies are in need of employees with the background, skills, and perspective attributed to the enterprise architect.

To meet the needs established by industry, information technology curricula must produce well-rounded students who have a broad enterprise-wide understanding of a variety of IT concepts from databases to networks, to data storage and management. IT firms are looking for employees who can engage the organization at a high level, define comprehensive requirements for large projects, design solutions, and be able to easily develop expertise in multiple areas of the company (Marshall & Roadknight, 2001; Sanders, 2004). This is no small task, and it necessitates a significant restructuring of many of the IT curricula in place today.

CHALLENGES TO ENTERPRISE SYSTEMS CURRICULA

Meeting the educational needs of enterprise-systems-related courses is difficult enough, but faculty and administrators in higher education are also plagued with paperwork and committees when attempting to implement new courses, content tracks, and areas of study. More significantly, university faculty are faced with a variety of concerns when attempting to produce and promote new curricular changes. On top of the challenge of mastering new content, many universities have an arduous approval process in place for any new class, making the task of linking a new course to an existing curriculum even more difficult (Helps, 2006). Most significant of all, the delivery of pedagogically sound content specific to information technology is problematic. Students must be prepared to engage rapidly developing equipment and practices by the completion of a degree, but ready access to equipment and content to meet these needs is extremely difficult. Universities cannot afford to adopt equipment at the same rate large companies are able to, making it difficult to offer a course on a topic like enterprise systems integration that will remain relevant and up to date (Davis, 2004; Prigge, 2005; Tompsett, 2005).

Beyond the challenge of specific courses, the landscape of enterprise information systems instruction in higher education covers a wide variety of interpretations. With no parent organization to make decisions about what is appropriate content for an information technology curriculum, individual colleges and universities are freely creating very disparate curricula. A 2005 survey of IT programs in colleges and universities around the United States showed that while many institutions placed unique emphasis on different aspects of information technology, all offered courses on networking, database construction and management, and software applications (including operating systems; Helps, 2006). Each of these parent topics in IT could easily be a curriculum of its own.

With such wide ground to cover with respect to content areas in information technology, capstone courses within the discipline are extremely challenging. Student preparation entering into these courses is often widely varied. These courses often take the form of an enterprise systems integration topic, or some other closely related topic (Suchan, Blair, Fairfax, Goda, Huggins, & Lemanski, 2006; Tetard & Patokorpi, 2005). It is at this point in an educational program that students have developed a broad-enough skill set to begin understanding the relationships between different areas of IT to one another and to the enterprise as a whole.

These capstone classes are often an ideal situation for academic-industry partnerships (Courte & Bishop-Clark, 2005; Turk-Bicakci & Brint, 2005). A few universities attempt to begin industry partnerships early in the academic program, but according to Courte and Bishop-Clark, partnerships involving more senior students tend to have higher rates of return (industry partners are interested in repeating the experience the following year) and more often lead to internships and job placements. Pedagogically, this industry interest in advanced students offers an opportunity to put students in situations that expose them to current technologies and problems within an industry setting.

The traditional method of teaching enterprise-systems-related topics at the college level would almost certainly involve the use of case studies to articulate relationships between technologies and practices. These case studies are beneficial to a student because they offer significant context to a real-life problem and afford the student an insider perspective on the subject. While this seems ideal, case studies cannot be written at the rate at which industry moves forward, rendering a specific case study more meaningless and outdated each semester. Industry engagement allows students

to work on projects designed with cooperating companies. Students receive the most hands-on training possible with relevant contexts and scenarios (Cameron, Knight, & Semmer, 2005; Harman, 2001).

THE PILLARS OF INFORMATION TECHNOLOGY

Helps' (2006) survey of information technology curriculum content identified three major areas of focus for most institutions: networking, databases, and applications. Presumably, a fully comprehensive education would offer a student significant training in all three areas. These areas map reasonably to industry definitions of information technology architecture as laid out by the five pillars of the modern IT architectural components suggested by EMC Corporation (Van Sickle et al., 2007), but not completely. Van Sickle et al.'s five pillars include databases, networking, software applications, operating systems (which many universities and colleges include under the umbrella of applications), and storage.

Van Sickle et al. (2007) argue that storage is a fifth pillar in the modern enterprise systems architecture that higher education has, for the most part, missed in curricular designs. The reasons for this lack of appreciation for the importance of this topic in modern IT and MIS (management information system) curricula are many, but mainly stem from a general lack of appreciation for the importance of storage-related topics in the modern corporate technology architecture. Storage as a topical area of study encompasses a wide range of concepts, topics, and issues including technologies and protocols, the evaluation of technical options based on business requirements, architectural design, systems management and governance, performance considerations, information management, data recovery, security, and emerging issues and technologies.

Furthermore, in modern enterprise architecture, each of the five pillars is interrelated with other pillars of IT. The enterprise architect cannot consider one pillar without at least some consideration of the other pillars. For example, databases require an expertise of their own, but some degree of expertise in networking allows a database administrator to better plan a database through knowledge of how users will access it. Likewise, expertise in storage allows the database administrator to better utilize resources by understanding the storage environment where the database resides. Network specialists can improve throughput with some expertise in storage technologies and through knowledge of the storage demands and capabilities of existing systems. Storage experts can customize their systems for better performance with expertise in the databases and applications stored on these systems.

This partial mapping shows that higher education is generally on the right track with IT education, but is often not doing a complete job. Several pillars of IT are commonplace in information technology curricula (databases, applications, networking), but these are often taught within their own curricular tracks or simply as independent courses of varying complexity, with universities expecting that a student will take lower level courses in each topic and then choose a specific pillar (for example, networking) to specialize in through advanced coursework.

The enterprise systems architect must be well versed in all of the pillars of IT, and higher education must develop curricula that foster this perspective in students (Catanio, 2005; McGann, Frost, Matta, & Huang, 2007). This means creating a curriculum to support comprehensive courses on enterprise systems design, implementation, and integration that teaches students to build systems encompassing networking, storage, databases, and applications (including operating systems or a lower level course specifically on operating systems) all in the context of alignment with business objectives and corporate strategy.

IMPORTANCE OF THE STORAGE PILLAR

While storage spending in industry (and in higher education, for that matter) continues to skyrocket, with respect to both spending and information production and retention (Gantz, 2007; Sun, 2005; "VUB Bank Improves Storage Performance," 2006), colleges and universities lag woefully behind in curriculum implementation. This can be due to several factors including the rapidly changing landscape of the storage industry (making it difficult to keep up with in an academic setting), an absence of instructional materials on the subject, or a lack of appreciation on the part of faculty of the importance of the topic in industry today. At the time of this writing, there is a plethora of practitioner-focused articles on storage-related topics, but no adequate textbooks or other instructional materials commercially available. Companies are forced to train storage professionals on the job rather than being able to hire university graduates with relevant knowledge and skills ("SNIA Announces Storage Networking Courses," 2005; Van Sickle et al., 2007).

In the mid to late 1990s, when universities began deciphering what content would be appropriate for an information technology curriculum, industry spending was predominantly on network equipment and software applications (McDonald, Rickman, McDonald, Heeler, & Hawley, 2001). This was quickly followed by interest in databases (Lynch, Carbone, Arnott, & Jamieson, 2002), and the early information technology curricula were framed. It has been a decade since those content areas were selected for IT in higher education, and most universities still hold tightly to them today. Industry definitions of core competencies for information technologists have evolved in that time. While

many universities have begun teaching operating systems from an experiential standpoint or as part of their application content, storage remains relatively ignored.

The primary problem with the disregard for storage is that the topic remains critical to enterprise systems education whether curricula support it or not. As a result of the absence of storage courses in higher education, enterprise architects must either go into the workforce with no understanding of storage (and thus, significantly unprepared to meet a growing industry need), or enterprise systems integration instructors must abandon weeks of their course to cover storage—thereby undermining the students' training in the integration of the five pillars of information technology.

Storage is of significant importance to industry because the value of information (and the virtual space it consumes) continues to climb. We have moved on from bookkeeping and asset-management tasks to business-to-business multimedia, video on demand, and voice and data integration. The number of e-mail messages alone has grown from 9.7 billion per day in 2000 to more than 35 billion messages per day in 2007. Within those e-mails we embed a variety of media and file types, forcing a focus on information sharing rather than server-centric data storage. Material has to be shared via storage networking environments to meet current information needs ("Information Lifecycle Management," 2006; Mesabi Group, 2006]. In addition, the increased storage and information management demands related to the Health Insurance Portability and Accountability Act (HIPAA), Sarbanes-Oxley, and other government-mandated regulations have created an enormous demand for enterprise storage and information management solutions.

To meet growing storage needs, the industry has introduced a selection of storage solution alternatives, each addressing specific data storage and management needs (Duplessie, 2006). Direct attached storage (DAS) systems attach storage drives directly to servers, network attached storage (NAS) environments are made up of specialized servers dedicated to storage, storage area networks (SANs) are highly scalable and allow hosts to implement their own storage file systems, and content addressable storage (CAS) systems are mechanisms for storing and retrieving information based on content rather than location. Because the storage needs of all organizations are growing exponentially today, huge investments are made each year in storage-related hardware, software, and skilled employees to design and navigate through these complex enterprise solutions.

CHALLENGES OF TEACHING STORAGE

The rapidly increasing need for storage professionals calls for an innovative capstone component in information technology education, drawing together the many aspects of storage technology and linking them to core information technology concepts. Students must be able to design, build, and manage storage architectures, as well as strategically plan for an organization's storage and information management needs.

Understanding the landscape of storage requires a broad skill set. Storage is not simply hard drives with data on them. The design and development of storage technologies require understanding technologies from legacy systems to emerging technologies, including knowledge management, disaster recovery, data replication, and application-aware resource management, as well as the business needs that storage networking addresses. Clearly, storage is a complex topic that highlights the relationships between each of Van Sickle et al.'s (2007) five pillars of IT: While each is unique, they go hand in hand, such as storage and networking, or storage and databases, and so forth.

As important as how storage courses are taught is when they are taught. Storage is one of the five pillars of information technology, and universities would do well to treat it as such. Courses and modules on enterprise storage-related topics throughout an educational curriculum are needed today in order to properly develop the enterprise architect demanded by industry ("SNIA Announces Storage Networking Courses," 2005). The implication of adding storage to the existing curriculum is significant. Additional classes on any topic are always a challenge for a solid curriculum as it usually means other classes must be dropped or modified to make space in the overall credit load. Capstone courses, such as systems integration, must also be changed to model the realistic expectations of a systems integration professional in industry. An integration architect must plan and coordinate all aspects of the information architecture for an organization, which includes storage as an independent item of consideration. Information storage architecture is a unique pillar of IT because it has unique requirements and attributes related to but not found in the other more traditional pillars. Likewise, a capstone enterprise systems integration course should seek to bring all five pillars of IT together for students, making distinct courses in networking, databases, applications, operating systems, and storage prerequisites for such a capstone course.

FUTURE TRENDS

Given the clear demand for graduates with the perspective of an enterprise architect in industry today and for the foreseeable future, and the rapidly growing demands for storage over that same time period, enterprise systems curricula clearly need to make adjustments to accommodate courses on enterprise systems integration as well as redesigning or creating necessary supporting courses.

Enterprise systems curricula must reflect comprehensive, broad education in relevant topics to prepare students for occupations in IT. This means introductory courses on networking, databases, applications, and storage, necessitating a minimum of four independent courses. Students will also require training on organizations and business processes, project management, and systems design. At an advanced level, enterprise systems curricula will need courses specific to enterprise integration (people, processes, and technology), teaching students to connect in relevant, meaningful ways what they have learned in the supporting courses.

CORPORATIONS LEADING THE CHARGE FOR STORAGE EDUCATION

The need for storage education has been publicized for several years (Morgenstern, 2003; Ruddlesden, 2005; Trelwyn, 2004). With virtually no institutions of higher education hearing the call, corporations have begun leading the charge for storage education. Hewlett-Packard and McData are among industry leaders actively encouraging colleges and universities to include storage education in their curriculum (Trelwyn). Professional organizations, such as the Storage Networking Industry Association (SNIA), are also encouraging storage education, offering course requirement suggestions and certifications.

EMC Corporation has recently launched the Academic Alliance Program (Van Sickle et al., 2007) to partner in a variety of ways with academia. The company is teaming up with university faculty to determine how to best create courses to educate students in storage-related topics within the context of a larger curriculum in information technology or related areas. The company also offers course content for storage education, as well as simulation materials and a current industry perspective on storage. To date, EMC is the only corporation stepping up in such an organized manner to work with academia to address the industry need for the inclusion of storage and information management topics and courses in IT, MIS, and other technology curricula.

CONCLUSION

With the clear importance of the enterprise architect's perspective in industry, and the subsequent increase in demand for skilled enterprise systems graduates, colleges and universities wishing to remain competitive in their educational offerings will have to make significant overhauls to their IT curriculum to accommodate these demands. Despite the difficulty of incorporating a complement of foundational courses in the five areas that comprise the modern enterprise IT architecture, IT and MIS curricula must strive to meet the increasing demand for employees who can address information systems and technologies at an organization-wide level. With IT budgets increasingly being spent on storage-related projects and components, a competitive educational institution must offer foundational education in all of the five areas suggested by Van Sickle et al. (2007).

While the challenges to implementing new curricula are significant, and the development of storage-specific courses is especially difficult due to the lack of instructional materials, academic partnerships with industry can help a college or university make significant progress toward completely supporting enterprise systems education. In addition to the creation of appropriate, relevant instructional content, these alliances also afford students industry engagements that are both highly effective for learning and helpful in ensuring skill sets that will help information technology students be better prepared for the business environment they will face upon graduation.

REFERENCES

Barrett, R. (2007). Worker shortage in the making? *JSonline.* Retrieved August 5, 2007, from http://www.jsonline.com/story/index.aspx?id=307642

Boyle, T. A., & Strong, S. E. (2006). Skill requirements of ERP graduates. *Journal of Information Systems Education, 17*(4).

Cameron, B. H., Knight, S. C., & Semmer, J. F. (2005). Strategies for experimental learning: The IT consulting model. Innovative methods for industry partnerships. In *Proceedings of the Sixth Conference on Information Technology Education: SIGITE '05.*

Cannon, D. M., Klein, H. A., Koste, L. L., & Magal, S. R. (2004). Curriculum integration using enterprise resource planning: An integrative case approach. *Journal of Education for Business, 80*(2).

Catanio, J. T. (2005). Extension of the IT curriculum: Developing LaSalle's IT graduate certificate program partnered with industry. In *Proceedings of the Sixth Conference on Information Technology Education: SIGITE '05.*

Courte, J., & Bishop-Clark, C. (2005). Strategies for making connections with industry: Creating connections. Bringing industry and education together. In *Proceedings of the Sixth Conference on Information Technology Education: SIGITE '05.*

Davis, C. H. (2004). Enterprise integration in business education: Design and outcomes of a capstone ERP-based

undergraduate e-business management course. *Journal of Information Systems Education, 15*(3).

Dede, C. J. (1986). The implications of emerging technologies for the value-oriented curriculum. *Momentum, 17*(3).

Dougherty, J. P., Dececchi, T., Clear, T., Richards, B., Cooper, S., & Wilusz, T. (2002). ITiCSE 2002 working group report: Information technology fluency in practice. *ACM SIGCSE Bulletin, 35*(2).

Duplessie, S. (2006). Storage networking: Back to basics. *Computerworld.*

Gantz, J.F. (2007). *The expanding digital universe: A forecast of world wide information growth through 2010.* IDC.

Gorgone, J. T., Gray, P., & Feinstein, D. (2000). MSIS 2000: Model curriculum and guidelines for graduate degree programs in information systems. *Communications of the Association for Information Systems, 3*(1).

Harman, G. (2001). University-industry research partnerships in Australia: Extent, benefits, and risks. *Higher Education Research & Development, 20*(3).

Helps, C. R. G. (2006). IT education: Curriculum development. Instructional design theory provides insights into evolving information technology technical curricula. In *Proceedings of the Seventh Conference on Information Technology Education SIGITE '06.*

Information lifecycle management: A discipline, not a product. (2006). *Datamonitor.*

Information Technology Association of America. (2007). Workforce & education. *Business Development.* Retrieved August 2, 2007, from http://www.itaa.org/workforce/

Light, A., & Strayer, W. (2000). Determinants of college completion: School quality or student ability? *The Journal of Human Resources, 35*(2).

Lynch, K., Carbone, A., Arnott, D., & Jamieson, P. (2002). A studio-based approach to teaching information technology. In *Proceedings of the Seventh World Conference on Computers in Education.*

Marshall, I. W., & Roadknight, C. M. (2001). Management of future data networks. *IEEE Transactions on Networking.*

McDonald, M., Rickman, J., McDonald, G., Heeler, P., & Hawley, D. (2001). Practical experiences for undergraduate computer networking students. *Journal of Computing Sciences in Colleges, 16*(3).

McGann, S. T., Frost, R. D., Matta, V., & Huang, W. (2007). Meeting the challenge of IS curriculum modernization: A guide to overhaul, integration, and continuous improvement. *Journal of Information Systems Education, 18*(1).

Mesabi Group. (2006). *Why SMI-S compliance is key to efficient storage management.* Author.

Morgenstern, D. (2003a). *Missing from the resume: SAN higher education.* Retrieved August 1, 2007, from http://findarticles.com/p/articles/mi_zdewk/is_200309/ai_ziff59296

Morgenstern, D. (2003b). *Storage education still on hold.* Retrieved August 1, 2007, from http://findarticles.com/p/articles/mi_zdewk/is_200310/ai_ziff108565

Mulder, M. C., Lidtke, D., & Stokes, G. E. (1997). Enterprise enhanced education: An information technology enabled extension of traditional learning environments. In *Proceedings of the 28th SIGCSE Technical Symposium on Computer Science Education SIGCSE '97.*

Nickerson, J. V. (2006). Teaching the integration of information systems technologies. *IEEE Transactions on Education, 49*(2).

Prigge, G. W. (2005). University-industry partnerships: What do they mean to universities? *Industry & Higher Education, 19*(3).

Ruddlesden, R. (2005). The need for data storage education. *IT Observer.* Retrieved August 1, 2007, from http://www.it-observer.com/articles/909/the_need_data_storage_education/

Sanders, L. (2004). Strategies for teaching something new. *Science Scope, 28*(1).

Seethamraju, R. (2007). Enterprise systems (ES) software in business school curriculum: Evaluation of design and delivery. *Journal of Information Systems Education, 18*(1).

SNIA announces storage networking courses. (2005). *Channel Times: Newsline of the IT Industry.*

Suchan, W. K., Blair, J. R. S., Fairfax, D., Goda, B. S., Huggins, K. L., & Lemanski, M. J. (2006). IT education: Faculty development. Faculty development in information technology education. In *Proceedings of the Seventh Conference on Information Technology Education SIGITE '06.*

Sun. (2005). *The business case for storage consolidation.* Author.

SUNY Fredonia. (2007). *ITAB renovation.* Retrieved August 1, 2007, from http://www.fredonia.edu/ITS/ITAB/labrenovation20000.asp

Tetard, F., & Patokorpi, E. (2005). A constructivist approach to information systems teaching: A case study on a design course for advanced-level university students. *Journal of Information Systems Education, 16*(2).

Tompsett, C. (2005). Reconfigurability: Creating new courses from existing learning objects will always be difficult. *Journal of Computer Assisted Learning, 21.*

Turk-Bicakci, L., & Brint, S. (2005). University-industry collaboration: Patterns of growth for low- and mid-level performers. *Higher Education, 49.*

Van Sickle, E., Cameron, B. H., Groom, F., Mallach, E., Dunn, D. B., Rollins, R., et al. (2007). Storage technologies: An educational opportunity. In *Proceedings of the Eighth Conference on Information Technology Education SIGITE '07.*

VUB bank improves storage performance and reliability, builds a platform for future growth with HP StorageWorks solution. (2006). *Case Study Forum.*

KEY TERMS

Enterprise Architect: An enterprise architect (EA) takes a company's business strategy and defines an IT systems architecture to support that strategy.

Enterprise Architecture: Enterprise architecture is a comprehensive framework used to manage and align an organization's business processes, IT software and hardware, local and wide area networks, people, operations, and projects with the organization's overall strategy.

Enterprise Systems Integration: It is a discipline that combines processes and procedures from systems engineering, systems management, and product development for the purpose of developing large-scale, complex systems that involve hardware and software and may be based on existing or legacy systems coupled with totally new requirements to add significant functionality.

Information Technology: IT includes all matters concerned with the furtherance of computer science and technology and with the design, development, installation, and implementation of information systems and applications. An information technology architecture is an integrated framework for acquiring and evolving IT to achieve strategic goals.

Sarbanes-Oxley (SOX): Administered by the Securities and Exchange Commission (SEC) in 2002, SOX regulates corporate financial records and provides penalties for their abuse. It defines the type of records that must be recorded and for how long. It also deals with falsification of data.

Service-Oriented Architecture: A service-oriented architecture is essentially a collection of services. These services communicate with each other. The communication can involve either simple data passing or it could involve two or more services coordinating some activity. Some means of connecting services to each other is needed.

Storage Networking: It is the practice of creating, installing, administering, or using networks whose primary purpose is the transfer of data between computer systems and storage elements and among storage elements.

Developing Trust in Virtual Teams

Niki Panteli
University of Bath, UK

INTRODUCTION

During the last few years, there has been an increasing acknowledgment of the importance of trust in business interactions within the management and organizational literatures (e.g., Kramer & Tyler, 1996; Mayer, Davis, & Schorman, 1995; Rousseau, Sitkin, Burt, & Camerer, 1999). Trust, as a positive and confident expectation in the behavior of another party (Cook & Wall, 1980; Currall & Judge, 1995), enables cooperation and becomes the means for complexity reduction, even in situations where individuals must act under uncertainty with ambiguous and incomplete information. Therefore, it is not surprising that in the current age of global and digital economy and virtuality (Shepherd, 2004), there has been an overwhelming interest in trust. Motivated by the need to better understand trust in the digital era, this paper views the case of global virtual teams in commercial business organizations.

BACKGROUND

Trust has received significant recognition as a phenomenon worthy of detailed study in organizational and management studies (Dirks & Ferrin, 2001). In organizations, individuals must often act under uncertainty with ambiguous and incomplete information. This lack of explicit knowledge introduces risk and thus the requirement for trust. Accordingly, trust is defined as the willingness of a party to be vulnerable to the actions of another party (Mayer et al., 1995) based on a state of a positive, confident, though subjective, expectation regarding the behavior of somebody or something in a situation that entails risk to the trusting party (Baba, 1999; Cook & Wall, 1980; Currall & Judge, 1995).

Numerous scholars agree that trust is highly beneficial for the functioning of organizations. Trust "is at the heart of knowledge exchange" (Davenport & Prusak, 1998, p.35). High levels of trust are also key to effective communication (Dodgson, 1993) as they "improve the quality of dialogue and discussions … [that] facilitate the sharing of … knowledge" (Ichijo, von Krogh, & Nonaka, 2000, p.200), and committed relationships (ibid). The centrality of trust is further accentuated by its absence: "mistrust ... makes success harder to attain" (Kanter, 1994, p.105) as it weakens relationships, increases dependence on less information, compromises rational and unprejudiced analysis and exploration, and undermines learning (Luhmann, 1979). Furthermore, it has been recognized that if trust is not prominent, this may lead to dissatisfaction, absenteeism, and even intention to quit (Cunningham & MacGregor, 2000). At the inter-organizational level, trust also plays a vital role since it is found to affect the degree of cooperation among participating parties (Grabowski & Roberts, 1998; Newell & Swan, 2000). This is particularly important for virtual organizations. The business motivation for virtual arrangements is the potential for increased value-added and competitive advantage from the enhanced knowledge stock and core competencies, which are deemed to accrue to such networks (Alavi &Leidner, 2001).

Clearly, there is little dispute over the significance of trust in the organizational literature. However, there seems to be little agreement on how trust is developed and maintained in both the traditional and the virtual organizational literature.

In the traditional literature on trust where face-to-face communication is the norm, trust develops as the degree of familiarity with other people increases; i.e., the more we get to know others, the more likely it is that we trust them (Lewicki & Bunker, 1995,1996). Lewicki and Bunker (1996) take the view that trust varies over time and takes on a different character at the various stages (early, developing, and mature stages) of a relationship, as we not only begin to feel more comfortable with other people as we spend more time with them, but also as our knowledge of their integrity and competence improves. Based on this view, Lewicki and Bunker (1996) suggest three categories of trust, each corresponding to a different stage of the relationship:

- Calculus-Based Trust, the type of trust that is grounded in the rewards to be derived from pursuing and preserving the relationship or in the fear of punishment for violating trust within the relationship;
- Knowledge-Based Trust that assumes that the more information one has about others, the more able one is to predict their actions; and
- Identification-Based Trust, the type of trust that is characterized by mutual understanding among all parties to the point that each can effectively act for the other.

These types of trust are "linked in a sequential iteration in which the achievements of trust at one level enables the development of trust at the next level" (p. 119).

Familiarity with other people has also been identified as an important antecedent of trust development in virtual teams. According to Handy (1995), for trust to develop in virtual environments there is a need for constant face-to-face communication. As he puts it: "paradoxically, the more virtual an organization becomes, the more its people need to meet in person" (Handy, 1995, p.46). This view has also been reinforced by Lipnack and Stamps (1997, p.226): "if you can drop by someone's office, see first-hand examples of prior work, and talk with other colleagues, you can more easily evaluate their proficiency." Researchers have already argued that the lack of proximity impersonalizes trust (Nandhakumar, 1999), while the virtual context of a geographically dispersed workforce may constrain or even impede rich information exchange[1] since communication becomes highly computer-mediated (Davenport & Pearlson, 1998). It follows, therefore, that trust based on familiarity with other individuals could not be easily developed in virtual settings.

In the following section, the challenges of developing trust in a virtual team setting are discussed by drawing upon the findings of existing empirical research.

Trust and Virtual Teams: Empirical Findings

While trust has been identified as a key feature for the success of virtual interactions, empirical research in this area has remained limited. Jarvenpaa and Leidner (1999) have conducted one of the most detailed research projects into studies on trust and virtual teams thus far. Their eight-week study of 75 teams of university students, each consisting of four to six members, highlighted significant differences in the behaviors and strategies between high- and low-trust teams and supported the existence of swift trust; this type of trust presumes that roles are clear and that each team member has a good understanding of others' roles and responsibilities (Meyerson, Weick, & Kramer, 1996).

However, trust is not always swift. Tucker and Panteli (2003) have illustrated the significance of shared goals and power in influencing trust development; these factors were not identified in the context of university settings as the tasks are often well-articulated in advance while power differentials, which could influence the degree of inter-dependence among members, are not significant in the case of university students. In business environments, however, power differentials prevail. Power, defined as the capability of one party to exert an influence on another to act in a prescribed manner, is often a function of both dependence and the use of that dependence as leverage (Rassingham, 1999). Indeed, power is an important contextual factor that affects trust (Hart & Saunders, 1997) in that it suggests the existence of a unilateral dependency or an imbalanced relationship (Allen, Colligan, Finnie, & Kern, 2000).

Accordingly, within a business environment where conflict and power differentials prevail, building trust is not always a swift process. Instead, it is found that the process of jointly constructing team goals holds significant value as it may provide the "glue" to hold team members together long enough to enable trust development.

Shared goals are and should be a key characteristic of virtual teams. They could provide a means to developing a common sense of identity for team members that can be of particular benefit to those global virtual teams who meet infrequently or perhaps not at all. These benefits include the establishment of a foundation upon which to build trust and minimize the use of coercive power in pursuit of a collaborative and productive relationship. However, the study finds that even though shared goals are important for the success of virtual teams, these should not be taken for granted. Indeed, goals may not be shared either because they do not exist at all, or because team members have not become aware of them, have their own priorities, or share different interpretations of the team's role. Furthermore, this study has also shown that the construction of shared goals is often not a one-off activity, but rather it is a process that requires the ongoing participation of all parties involved. Though this could be a time-consuming, iterative, and difficult process, these findings allow us to argue that it is far better to invest in it and as up front in the project as possible than deal with the vicious, destructive, downward spirals that result from team members with conflicting goals and poor levels of trust.

In considering power within virtual teams, there is an increasing recognition in the literature that knowledge is indeed power and that teams are often formed to create knowledge through combination and exchange. Within these teams, the team member with power at any given time is the one with the most relevant knowledge at that time. Tucker and Panteli (2003) found that in high-trust teams power differentials do not disappear; rather, power shifts from one member to another throughout the life cycle of a project depending on the stage and requirement of each stage.

Further to the issues of shared goals and power, Tucker and Panteli (2003) found support for the need for face-to-face interaction. However, the opportunities to meet face-to-face have been severely limited by economic pressures and, more recently, terrorist attacks. Under these circumstances, those virtual teams that work well tend to undertake regular communications via synchronous, "live" computer-mediated communication (CMC) such as the telephone and videoconferencing systems. Participants confirmed that synchronous media offered more feedback and therefore facilitated understanding more effectively than asynchronous technologies such as voicemail and e-

mail. The use of asynchronous technologies was, however, regularly used for documenting and recording agreements and providing brief, simple updates to work progress. The teams that worked well also included a social and fun element in their interactions that appeared to help in creating a stronger shared social context.

Table 1 details the common features and behaviors observed within the global virtual teams studied in Tucker and Panteli (2003, p.91).

FUTURE TRENDS

The increasing pressure to perform in a short period of time with unfamiliar people and within a computer-mediated environment makes it imperative to study not only the type of trust but also how trust is formed and developed in a virtual team context.

It is readily acknowledged that what has been attempted here is only an exploration of contingencies to provide a better understanding of trust within the virtual team environment. There is no doubt that this is only the beginning of our understanding of trust in a virtual context.

Virtual interactions, however, exist at multiple levels—between individuals, individuals and organizations, and between organizations. Nandhakumar, Paneli, Powell, and Vidgen (2004, p.79) have introduced a digital era interaction (DEI) matrix to support exploration of trust relationships at these different levels.

The digital era interaction (DEI) matrix in Figure 1 indicates areas where digital era developments might be expected with consequent implications for trust. It views trust in two dimensions allowing an exploration of its nature in organization-to-organization settings, organization-to-individual and individual-to-organization interactions, and at an individual-to-individual level. In this paper, trust at the individual-to-individual level (and, more specifically, employee-to-employee) was explored. Thus, although the individual-to-organization (I2O) quadrant is currently the least developed of the four, it may prove to be an interesting sector in the longer term. Consumer-to-Business (C2B) is another growing area that allows consumers to organize themselves and use their stronger bargaining power to obtain a better price. As it was put: "The matrix is not exhaustive (for example, it does not include E2C, employee to consumer), but it does cater for the principal electronic relationships that currently exist and highlights ones that will be significant in the future" (p.79).

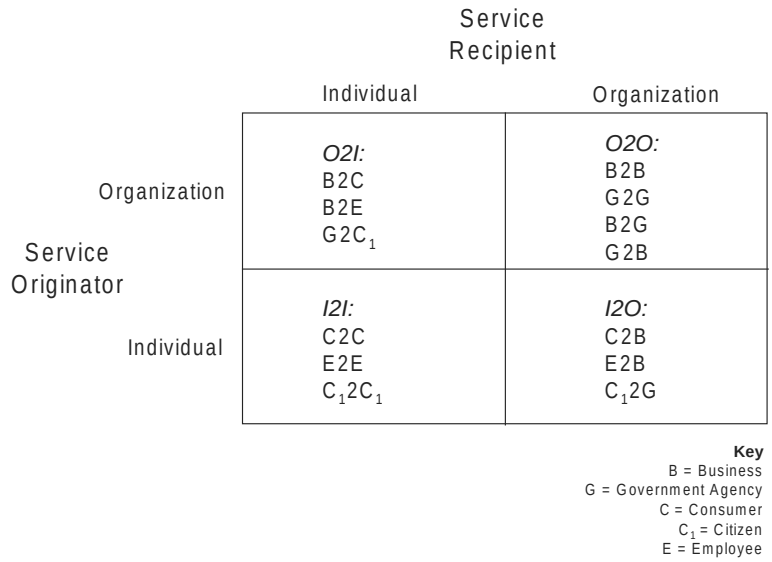
CONCLUSION

This paper reinforces arguments in the existing literature on the significance and complexity of trust dynamics in building effective virtual teams. It defined trust and outlined its central role in virtual interactions while also identifying some of the challenges involved in developing trust in the digital era, arguing that trust is necessary but not sufficient for promoting effective, collaborative virtual interactions using empirical findings. The paper illustrates the significance of shared goals and power in influencing trust development. It has also become apparent that while the agreement of shared goals provides a mobilizing force for the members of global virtual teams, the process of developing these goals holds significant value in terms of the exchange of information, learning, improving understanding, and an opportunity to

Table 1. Differences between high-trust and low-trust global virtual teams

High-Trust Global Virtual Teams	***Low-Trust Global Virtual Teams***
Factors related to Shared Goals: *Awareness of shared goals *Take time to build shared goals *Open debate for shared goals up front *Team-based goals have primacy	**Factors related to Shared Goals:** *Lack of awareness of shared goals *Lack of shared goals *Opinions of others not considered *Individual goals take primacy
Factors related to Power: *Availability of facilitators *Facilitators focus on win-win *Recognition of knowledge as power *Recognition that power moves; power in many places *Power differentials are minimized	**Factors related to Power:** *Power battles *Coercion *Misunderstandings and conflict of interests *Use of positional power *Perception of 'I have Power'
Communication: *Face-to-face where possible *Regular synchronous CMC *Social interaction	**Communication:** *Asynchronous CMC * Time difference matters *Little or no social interest

Figure 1. Digital era interactions (DEI)



demonstrate trustworthiness. Repositioning trust in this way increases our chances of making sense of complex virtual, computer-mediated situations, and puts us in closer touch with the challenges of developing trust. In doing so, the paper carries forward an important debate in the digital era.

REFERENCES

Alavi, M. & Leidner, D.E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Allen, D., Colligan, D., Finnie, A., & Kern, T. (2000). Trust, power and inter-organisational information systems: The case of the electronic trading community TransLease. *Information Systems Journal, 10*, 21-40.

Baba, M. (1999). Dangerous liaisons: Trust, distrust, and information technology in American work organizations. *Human Organization, 58*(3), 331-346.

Cook, J. & Wall, T. (1980). New work attitudes measures of trust: Organizational commitment and personal need fulfilment. *Journal of Occupational Psychology, 53*(1), 39-52.

Cunningham, J.B. & MacGregor, J. (2000). Research note: Trust and the design of work: Complementary constructs in satisfaction and performance. *Human Relations, 53*(12), 1575-1591.

Currall, S. & Judge, T. (1995). Measuring trust between organization boundary role persons. *Organization Behaviour and Human Decision Processes, 64*(2), 151-170.

Daft, R.L. & Lengel, R.H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

Davenport, T.H.& Pearlson, K. (1998). Two cheers for the virtual office. *Sloan Management Review*, Summer, 51-65.

Davenport, T.H. & Prusak, L., (1998). *Working knowledge: How organizations manage what they know.* Cambridge, MA: Harvard Business School Press.

Dirks, K.T. & Ferrin, D.L. (2001). The role of trust in organizational settings. *Organization Science, 12*(4), 450-467.

Dodgson, M., (1993). Learning, trust and technological collaboration. *Human Relations, 46*(1), 77-95.

Grabowski, M. & Roberts, K.H. (1998). Risk mitigation in virtual organizations. *Journal of Computer-Mediated Communication, 3*(4), June.

Handy, C. (1995). Trust and the virtual organization. *Harvard Business Review*. May-June, 40-50.

Hart, P. & Saunders, C. (1997). Power and trust: Critical factors in the adoption and use of electronic data interchange. *Organization Science, 8*(1), 23-42.

Ichijo, K., von Krogh, G., & Nonaka, I., (2000). Knowledge enablers. In G. von Krogh, J. Roos, & D. Kleine (Eds.), *Knowing in firms: Understanding, managing and measuring*

knowledge, pp. 173-203. London: Sage Publications.

Jarvenpaa, S.L. & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organization Science, 10,* 791-815.

Kanter, R.M., (1994). Collaborative advantage: Successful partnerships manage the relationships, not just the deal. *Harvard Business Review,* July-August, 98-108.

Kramer, R.M. & Tyler, T.R. (1996). Whither trust. In R.M. Kramer & T.R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research*. Thousand Oaks, CA: Sage Publications.

Lewicki, R.J. & Bunker, B.B. (1995). Trust relationships: A model of trust development and decline. In B. Bunker & J. Z. Rubin (Eds.), *Conflict, cooperation and justice*.San-Francisco: Jossey-Bass.

Lewicki, R.J. & Bunker, B.B. (1996). Developing and maintaining trust in working relationships. In R.M. Kramer & T.R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research*. Thousand Oaks, CA: Sage Publications.

Lipnack, J. & Stamps, J. (1997). *Virtual teams: Reaching across space, time, and organizations with technology*. New York: John Wiley & Sons.

Luhmann, N. (1979). *Trust and power*. London: John Wiley and Sons.

Mayer, R.C., Davis, J.H., & Schorman, F.D., (1995). An integrative model of organizational trust. *Academy of Management Journal, 20*(3), 709-734.

Meyerson, S., Weick, K.E., & Kramer, R.M. (1996). Swift trust and temporary groups. In R.M. Kramer & T.R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research*. Thousand Oaks, CA: Sage Publications.

Nandhakumar, J. (1999). Virtual teams and lost proximity: Consequences on trust relationships. In P. Jackson (Ed.), *Virtual working – Social and organizational dynamics*. London: Routledge.

Nandhakumar, J., Panteli, N., Powell, P., & Vidgen, R., (2004). Trust in the digital era. In N. Mylonopoulos, N. Pouloudi, & G. Doukidis, *Social and economic transformation in the digital era*. Hershey, PA: Idea Group Publishing.

Newell, S. & Swan, J. (2000). Trust and inter-organizational networking. *Human Relations, 53* (10), 1287-1328.

Rassingham, P. (1999). Risks in low trust among trading partners in electronic commerce. *Internet Research: Electronic Networking Applications and Policy, 10*(1), 56-62.

Rousseau, D., Sitkin, S., Burt, R., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review, 23*(3), 393-404.

Shepherd, J. (2004). Why the digital era? In N. Mylonopoulos, N. Pouloudi, & G. Doukidis, *Social and economic transformation in the digital era*. Hershey, PA: Idea Group Publishing.

Tucker, R. & Panteli, N. (2003). Back to basics: Sharing goals and developing trust in global virtual teams. In N. Korpela, R. Montealegre, & A. Poulymenakou (Eds.), *Organizational information systems in the context of globalization.* Boston, MA: Kluwer Academic Publishers.

KEY TERMS

Computer-Mediated Communication: Communication that is facilitated using information technologies such as email, videoconferencing, teleconferencing.

Power: The ability to influence others.

Power Differentials: The existence of imbalanced power relationships.

Shared Goals: Goals that articulate what the teams stand for and their shared vision.

Social Interactions: A chain of interrelated messages that include a social and fun element and contribute to increasing familiarity among participants.

Trust: A state of a positive, confident though subjective expectation regarding the behavior of somebody or something in a situation that entails risk to the trusting party.

Virtual Teams: A group of geographically dispersed individuals who work on a joint project or common task and communicate electronically.

ENDNOTE

[1] This view corresponds to the media richness theory that argues that electronic media are lean (Daft & Lengel, 1986).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 844-848, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Diffusion of E-Learning as an Educational Innovation

Petek Askar
Hacettepe University, Turkey

Ugur Halici
Middle East Technical University, Turkey

INTRODUCTION

Most of the discussions related to education are about technological innovations. Indeed as Rogers (1995) stated, we often use the word "innovation" and "technology" as synonyms. Technology is regarded as an agent of change in educational settings, and a quick analysis of the educational projects all over the world shows us that it is not possible to define a future vision of education without technology, especially e-learning, which brings two important concepts together: technology and learning. Therefore as a form of distance learning, e-learning has become a major instructional force in the world.

Besides the technological developments, the last two decades have brought a tremendous increase in knowledge in education, particularly in learning. The emerging views of learning which should be taken into consideration for every learning environment could be stated as follows: personalized, flexible, and coherent (learning is connected to real-life issues); not bounded by physical, geographic, or temporal space; rich in information and learning experiences for all learners; committed to increasing different intelligences and learning styles; interconnected and collaborative; fostering interorganizational linkages; engaged in dialogue with community members; accountable to the learner to provide adaptive instructional environments (Marshall, 1997).

WWW is an environment that fits the new paradigm of learning and facilitates "e-learning" which faces a challenge of diffusion. Diffusion is defined by Rogers (1995) as the process by which an innovation is communicated through certain channels over time among the members of a social system. Therefore the adoption of WWW as a learning environment is influenced by the following set of factors: 1) the individuals' perception of the attributes of e-learning, 2) the nature of the communication channels, 3) the nature of the social system, and 4) the extent of the change agents' efforts in the e-learning. These are the variables that affect the diffusion of e-learning in the schools and countries.

E-LEARNING AND INSTRUCTIONAL DESIGN

E-learning not only opens up new ways of learning and teaching, but also leads to a new way of thinking and organizing learning content. Collaborations among different stakeholders cause new standards for design of knowledge on the Internet. In traditional computer-based instruction, content comes in units called courses. However a new paradigm for designing instruction, grounded in the object-oriented notion of computer science, is called "learning objects."

Learning object is defined by the Learning Technology Standards Committee (2002) of the Institute of Electrical and Electronics (IEEE) as any entity, digital or non-digital, that can be used, re-used, or referenced during technology-supported learning. The features of learning objects are self-contained, interactive, reusable, and tagged with metadata. By the use of learning objects, one can learn just enough, just in time, and just for them. Learning objects can be considered a movement within the field of e-learning, one aimed at the componentization of learning resources, with a view to reusability (Duchastel, 2004).

The idea of educational software as a package is becoming outdated and making way for learning objects as a new way of designing instructional materials. In designing learning objects, the studies on multiple representation of knowledge become important since people have different learning styles and strategies. The associations between these two constructs are the main focus of the new instructional design principles. Therefore, the development of learning objects and the way of creating teaching units are well suited for what we call the Information Age.

A representation of knowledge could be decomposed into its parts, where the parts are far from arbitrary. Then they can be used and reused in a great variety of combinations, like a child's set of building blocks. Every combination is meaningful and serves as an instructional whole. Holland (1995) compares building blocks to the features of the human face. The common building blocks are: hair, forehead, eyebrows, eyes, and so on. Any combination is different and may never appear twice. This analogy could be true of

e-learning platforms, where learning objects are put together to make up a meaningful whole, which we call instructional materials.

The five fundamental components of instructional design process are learners, content, objectives, methods, and assessment. Hence, for a systematic instructional design of a subject matter, the basic steps are: learner characteristic identification, task analysis, objectives, content sequencing, instructional strategies, message design, and instructional delivery and evaluation.

The awareness of learner differences with respect to entry competencies, learning styles and strategies, motivation, and interest are critical. However it is difficult to accomplish this task by using ongoing approaches. Indeed, new technologies, if used properly, enable us to make the lessons more individualized. The Learning Objects Metadata Working Group (IEEE, 2002) stated its goal as: to enable computer agents to automatically and dynamically compose personalized lessons for an individual learner. This leads a paradigm shift to approaches to instructional design. As Wiley (2001) stated, a problem arose when people began to actually consider what it meant for a computer to "automatically and dynamically compose personalized lessons." It seems that the idea of learning objects is challenging, but opens to new concepts, strategies, and research areas in the instructional design process.

E-LEARNING AND SCHOOL MANAGEMENT

For most of the last two decades, technology has been implemented in schools, and its potential to change the educational systems has been argued for. There are tremendous efforts to encourage the integration of computers and Internet into schools. However, in one of the diffusion studies conducted by Askar and Usluel (2001), two paths to the adoption of computers are presented. One path is related to the use of technology in the school management system; the other one is related to the use of technology in the teaching and learning process. For many reasons the rate of adoption of computers in management applications is quicker than the learning-teaching applications. Indeed the concerns related to use of computers in the teaching-learning process are still at the awareness stage. On the other hand, the need for using computers and the Internet for management purposes is more relevant and seems more convenient for the current school system.

Educators assert that the central purpose of school management systems should be to improve instructional program quality. In light of this idea, a typical configuration of a Web-based school management system designed—taking this idea into consideration—includes administration, assessment, and communication. The features are: student enrollment, attendance, registration, test scores, grades and other record-keeping tasks, formative and summative evaluation, and feedback to parents and teachers about student learning and performance. In addition, new online management systems include item- banking capability for adaptive testing and online learning modules.

E-LEARNING AND THE COMMUNITY

The modern world requires individuals and communities to be able to continually develop and utilize different skills and knowledge. There is growing consensus among OECD countries that modern economies cannot afford a significant number of uneducated people (OECD, 2000). However, education systems throughout the world are ill equipped to address individual and community learning needs. The existing school system is not flexible for those who for some reason left school early.

Distance education is a recognized solution all over the world for bridging the learning and education divide between the educated and poorly educated. It gives people the opportunity to continue their formal education. Despite the initial concerns that distance education might be lower in quality than traditional method of schooling, many forms of distance education are gaining acceptance (Belanger & Jordan, 2000). Therefore distance education is receiving positive attention from governments as a solution to the educational problem mentioned above.

Also, the trend towards lifelong learning is universal. The transformations taking place in all societies require an increasing participation of individuals, an ability to innovate and solve problems, and a capacity to learn and go on learning (Mayor, 1994). Moreover, the term "open learning" is used to lower barriers that stand in the way of individuals and communities wishing to engage in different learning opportunities.

One of the solutions for the above mentioned problems is learning centers, which are flexible learning organizations and which serve the learning needs of the individuals and communities. A school that is well equipped and organized could be opened during non-traditional school hours. Therefore, schools as learning centers can be critical resources to meet the growing need for distance education students and other community members. However, in highly centralized education systems, it is very difficult to organize schools for those other than the registered students. The rules and regulations for conventional school become real barriers for open learning environments.

FUTURE TRENDS IN TECHNOLOGY

While adopting the current technology for enhancing teaching and learning, advances in micro- and nano-technology push the limits of miniaturization, and of minimizing the costs and power of microelectronic components and microsystems. Explorations of alternative materials are expected to allow organic flexible materials for displays, sensors, and actuators so that they can be placed anywhere and can take any shape. Furthermore it is expected that not only PCs, but also all our surroundings, will be interfaced. Instead of only " writing and reading" in human-computer interaction, all senses are to be used intuitively. Information search will be context-based instead of "word " based. Mobile and wireless devices will be used not only for voice transfer, but also for full multimedia (IST WP, 2002)

As information and communication technologies change, open systems and services are to be developed in support of ubiquitous, experiential and contextualized learning, and virtual collaborative learning communities improving the efficiency and cost-effectiveness of learning for individuals and organizations, independent of time, place, and pace. Next-generation learning solutions are expected to combine cognitive and knowledge-based approaches, with new media having intelligence, virtual and augmented reality, virtual presence, and simulation (ISTC, 2001).

CONCLUSION

Our current educational system is highly resistant to change. Making a client-based change rather than a technological based one will be the most important innovation to accomplish for the educational change. While technology is pushing the limits of e-learning environments, special care should be taken in educational and organizational frameworks. The stakeholders of the systems are students, teachers, principals, learners, and community. Their attitudes, needs, and expectations from e-learning are important issues for the change process. The innovation adoption variables of relative advantage, compatibility, visibility, ease of use, results demonstrability, and triability should be considered by school administrators seeking to increase the rate of adoption of e-learning within their organization (Jebeile & Reeve, 2003).

Complexity is another issue to be considered. Fullan (1991) defines complexity as the difficulty and extent of change required of the individuals responsible for implementation. Therefore, there should be an emphasis on simplifying the process of e-learning, while moving from approaches based on knowledge transfer to systems based on the dynamic construction and user-friendly exchange of knowledge between learners, teachers, and learning communities.

The educational community, as the end user of e-learning systems, should be given the opportunity of observing and trying the e-learning systems. Awareness or being informed about the innovation is the key factor for changing the negative attitudes or beliefs. It is known that if people see the implementation and results of innovation, they are more likely to adopt them for their usage. Unfortunately, the benefits of e-learning are not well known and well recognized by all relevant stakeholders. Therefore, a comprehensive and systematic awareness campaign is needed to speed up the rate of adoption.

REFERENCES

Askar, P. & Usluel, Y. (2001, March 5-10). Concerns of administrators and teachers in the diffusion of IT in schools: A case study from Turkey. *Proceedings of the 12th International Conference of Society for Information Technology and the Teacher Education,* Orlando, Florida. Retrieved from *www.aace.org/dl/index.cfm/fuseaction/View/papered/3970*

Belanger, F. & Jordan, D.H. (2000). *Evaluation and implementation of distance learning: Technologies, tools and techniques.* Hershey, PA: Idea Group Publishing.

Duchastel, P. (2004) Learning objects and instructional design. *Interactive Technology and Smart Education, 1*(1), 67-70.

Holland, J.H. (1995). *Hidden order: How adaptation builds complexity*. Perseus Books.

IEEE Learning Technology Standards Committee. (2002). *Draft standard for learning object metadata.* Retrieved November 22, 2002, from *ltcs.ieee.org/wg12/LOM_1484_12_1_V1_Final_Draft.pdf*

IST WP. (2002, December 17). *IST priority workprogramme 2003-2004, information society technologies.* Retrieved from *fp6.cordis.lu/fp6/call_details.cfm?CALL_ID=1*

ISTC. (2001, October 17). *Technology supported learning, ISTC–Information Society Technologies Committee.* Final Report from the Working Party on Education and Training, Luxembourg. Retrieved from *www.proacte.com/downloads/eandt/ISTC-CONsolidated-report-et.DOC*

Jebeile, S. & Reeve, R. (2003) The diffusion of e-learning innovations in an Australian secondary college: Strategies and tactics for educational leaders. *The Innovation Journal, 8*(4), 1-21.

Marshall, S.P. (1997). Creating sustainable learning communities for the twenty-first century. In F. Hesselbein et al. (Eds.), *The organization of the future* (pp. 177-188). San Francisco: Jossey-Bass.

Mayor, F. (1994). Lifelong learning for the 21st century. *Proceedings of the 1st Global Conference on Lifelong Learning*, Rome, UNESCO, DG/94/39. Retrieved November 17, 2002, from *unesdoc.unesco.org/ulis/dgspeech other.html*

OECD/National Centre for Adult Literacy. (2000). *Learning to bridge the digital divide (schooling for tomorrow)*. France: OECD Publications.

Rogers, E.M. (1995). *Diffusion of innovations*. New York: The Free Press.

Shepherd, C. (2000). Objects of interest. Retrieved November, 18, 2002, from *www.fastrak- onsulting.co.uk/tactix/Features/perfect_etutor.htm*

Wiley, D.A. (2001). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In D.A. Wiley (Ed.), *The instructional use of learning objects*. Bloomington, IN: Association for Educational Communications and Technology. Retrieved February, 12, 2004, from *wiley.ed.usu.edu/articles.html*.

KEY TERMS

Diffusion: The process by which an innovation is communicated through certain channels over time among the members of a social system.

Distance Education: A type of formal education in which the majority of the instruction, which is transmitted through technology, occurs when student and instructor are not in the same place.

Innovation: An idea, practice, or object that is perceived as new by an individual or other unit of adoption.

Instructional Design: The systematic process of translating general principles of learning and instruction into plans for instruction and learning.

Learning Object: Any entity, digital or non-digital, that can be used, re-used, or referenced during technology-supported learning.

Learning Styles: The ways in which a person takes in and processes information and experiences.

Rate of Adoption: The relative speed with which an innovation is adopted by members of a social system.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 849-852, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

A Diffusion-Based Investigation into the Use of Lotus Domino Discussion Databases

Virginia Ilie
University of Kansas, USA

Craig Van Slyke
Saint Louis University, USA

Hao Lou
Ohio University, USA

John Day
Ohio University, USA

INTRODUCTION

Some information and communication technologies (ICT) that support groups have become tightly engrained in the fabric of organizational life, but others have not been as widely adopted (Orlikowski, 1993). This is true of both businesses and educational institutions. For example, many professors and students regularly interact through e-mail. In contrast, groupware systems are not as widely used.

In this article, we present research that uses diffusion of innovation (DOI) theory (Rogers, 1995) as the lens through which to study factors that impact intentions to use a groupware application in a higher education setting. This research investigates the following research question: *Are adopters' perceptions of the characteristics of groupware technology related to their intentions to use the technology?*

Organizations are increasingly making use of ICT to enable distance learning for their employees and, in some cases, customers (Dick, Case, Ruhlman, Van Slyke, & Winston, 2006). Furthermore, numerous academic institutions are implementing and supporting collaborative technologies to support student learning. For instance, Cordon, Anaya, Gonzalez, and Pinzon (2007) report on implementation of a virtual learning center to support learning of 4,000 students in Spain. Leung and Li (2006) describe efforts to create an e-learning environment in Hong Kong. High student dropout rates and low student satisfaction with e-learning remain major drawbacks in such implementations. Despite the presence of online discussion boards, sometimes students feel that there is little interaction in Web-based learning (Chatterjea, 2004). Avoiding failure in distance learning efforts requires better understanding of e-learners and their perception of ICT-based learning technologies.

BACKGROUND

Groupware technology is designed to facilitate the work of groups. This technology may be used to communicate, cooperate, coordinate, solve problems, compete, or negotiate. While traditional technologies such as the telephone qualify as groupware, the term is ordinarily used to refer to a specific class of technologies relying on modern computer networks, such as e-mail, newsgroups, videophones, or chat.

Figure 1. Groupware classification

	Same Time "Synchronous"	**Different Time** "Asynchronous"
Same Place "Co-Located"	Group decision support systems, voting, presentation support	Shared computers
Different Place "Distance"	Videophones, chat	Discussions, e-mail, workflow

Groupware technologies are typically categorized along two primary dimensions, time and place (Johansen, 1988), as shown in Figure 1. In this study, we investigate Lotus Domino discussion database (DDB), an asynchronous groupware product designed to be used by users "any time and any place."

The DDB is one of the Lotus Notes groupware applications made available to Web browsers via the Domino HTTP server technology. One may think of a DDB as an informal meeting place where the members of a workgroup can share ideas and comments. Like a physical meeting, each member of the workgroup "listens" to what others have to say, and can voice his or her own opinion.

Users have the ability to simply browse through discussion topics and responses contributed by others. This is particularly useful for new workgroup members who need to become oriented to important issues regarding the group. The history of any discussion is preserved in the discussion database and is presented as a discussion thread. Figure 2 illustrates a threaded discussion.

In a threaded discussion, users can either respond to an existing discussion thread or create a new discussion thread by posting a new topic. Posted items can also be edited and deleted by the author. Among the most important benefits of such groupware systems is to extend learning beyond the classroom (Day, Lou, & Van Slyke, 2004). Discussion databases can expand learning between students and faculty, and between students themselves by encouraging interaction and reflection on a topic.

An additional benefit comes from the fact that online interactive discussions may promote higher-order learning. For example, online debates often require synthesis of knowledge, which represents higher-order learning (Hazzari, 2004). In asynchronous online discussions, students have more time to reflect and synthesize their knowledge; there is less time pressure to respond quickly than there would be in a classroom setting. In addition, higher levels of certain cognitive processes may occur with online learning than with traditional classroom interactions (Heckman & Annabi, 2006).

MAIN FOCUS OF THE CHAPTER

DOI theory serves as the theoretical basis for this study. DOI theory is concerned with how the use of an innovation spreads throughout a social system (Mahajan, Mueller, & Bass, 1990). Diffusion theory has been applied to a wide range of technologies, including information and communication technologies such as groupware.

An often studied area related to innovation adoption is the impact of adopters' perceptions of the characteristics of an innovation on its adoption (Gatignon & Robertson, 1985; Lancaster & Taylor, 1986; Rogers, 1995). It is important to note that it is the potential adopters' *perceptions* of these characteristics rather than an expert's objective assessment of how an innovation rates on these characteristics

Figure 2. Discussion thread in a Domino discussion database

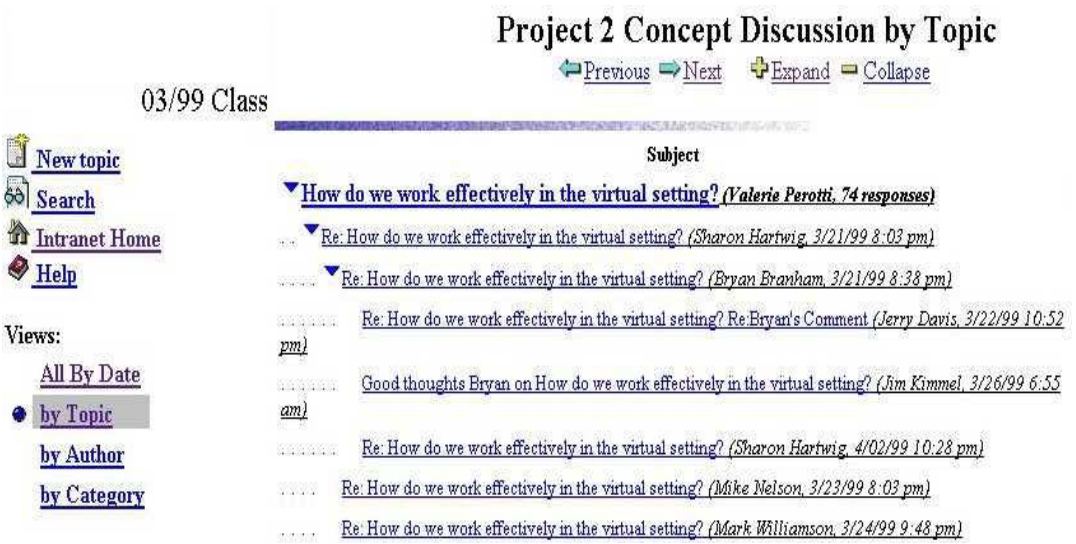


Table 1. Perceived innovation characteristics

Characteristic	Definition	Conceptual References	Empirical References
Relative Advantage	Degree to which an innovation is seen as being superior to its predecessor	Rogers, 1995	Van Slyke, Belanger, & Comunale, 2004
Complexity	Degree to which an innovation is seen by the potential adopter as being relatively difficult to use and understand	Rogers, 1995	Cooper & Zmud 1990; Teo, Tan, & Wei, 1995; Van Slyke et al., 2004
Compatibility	Degree to which an innovation is seen to be compatible with existing values, beliefs, experiences, and needs of adopters	Rogers, 1995	Taylor & Todd, 1995; Van Slyke et al., 2004
Trialability	Based on adopters' perceptions of the degree to which an innovation can be used on a trial basis before confirmation of the adoption must occur	Rogers, 1995	Teo et al., 1995; Chong & Pervan, 2007
Result Demonstrability	Degree to which the results of using an innovation are perceived to be tangible	Moore & Benbasat, 1991	Ilie, Van Slyke, Green, & Lou, 2005
Visibility	The perception of the actual visibility of the innovation itself as opposed to the visibility of outputs	Moore & Benbasat, 1991	Agarwal & Prasad, 1997; Ilie et al., 2005
Voluntariness	Degree to which use of an innovation is perceived as being of freewill	Moore & Benbasat, 1991	Agarwal & Prasad, 1997; Van Slyke, Dick, & Case, 2005; Anderson, Schwager, & Kerns, 2006

that impacts the diffusion rate (Lancaster & Taylor, 1986; Rogers, 1995).

Rogers (1995) lists five perceived characteristics of an innovation that can help explain its adoption: (1) relative advantage, (2) compatibility, (3) complexity, (4) trialability, and (5) observability. Other constructs of interest to this study include voluntariness, result demonstrability, and visibility. Table 1 provides definitions for the innovation characteristics included in this study along with references for both conceptual and empirical studies related to each construct.

The research model derived from the relationships between intention to use and each of the factors discussed in this section is shown in Figure 3. Intention to use DDB is the dependent variable of interest. The expected direction of each relationship is indicated on the corresponding path.

A series of testable hypotheses can be developed from the research model and the operationalizations of the constructs of interest. These are shown below.

H1: Student perceptions of the characteristics of DDB technology are related to their intention to use the technology.

Several sub-hypotheses can be derived from H1, all of which are similar in wording. H1n shows the general form for each hypothesis. Table 2 shows the hypothesis number, scale of interest, and hypothesized relationship direction for each sub-hypothesis.

H1n: Student perceptions of the X characteristic of DDB technology as measured by the X scale are positively/negatively related to their intention to use the technology.

In order to test the research model, a voluntary survey was administered to business students at a major Midwestern university. Only students enrolled in a course where Domino

Figure 3. Research model

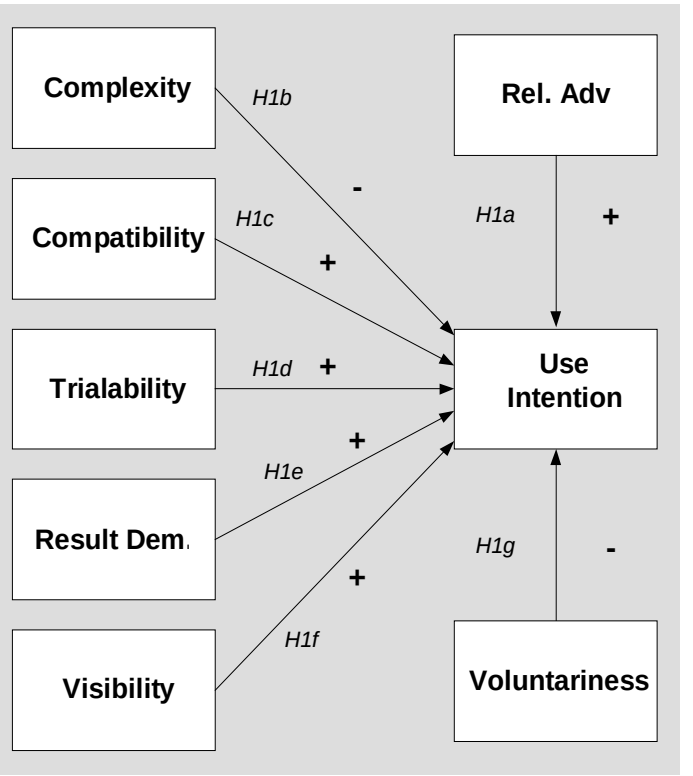


Table 2. Hypotheses

Hypothesis	Characteristic (X)	Direction
H1a	Relative Advantage	Positive
H1b	Complexity	Negative
H1c	Compatibility	Positive
H1d	Trialability	Positive
H1e	Result Demonstrability	Positive
H1f	Visibility	Positive
H1g	Voluntariness	Negative

Table 3. Sample characteristics (descriptive)

Characteristic	Median	Mean	Std. Dev.
Age	*22.5*	*25.9*	*7.28*
Work experience	*3.0*	*5.5*	*7.19*
Computer experience	*8.0*	*8.5*	*4.13*

Table 4. Sample characteristics (frequencies)

Characteristic	Frequency
Gender	
- Female	98
- Male	88
Class	
- Sophomore	4
- Junior	41
- Senior	139
- Graduate	2
Prior E-Mail Use	172
Prior Web Use	172
Prior Word Processor Use	164

was made available to the students as part of the course were included. A total of 186 surveys were completed. Tables 3 and 4 summarize the respondents' characteristics.

Measures

Most items for the measurement instrument were adapted from an instrument developed and validated by Moore and Benbasat (1991). Items were measured using a seven-point Likert-type scale. Cronbach's (1970) alpha values ranged from 0.77 to 0.96, indicating acceptable levels of reliability. The scales were further validated by performing a series of factor analyses, with one analysis performed for each scale. In all cases, the analysis indicated that the scale items associated with a given construct loaded on a single factor, indicating that each set of scale items measures a single construct. Factor loadings were all acceptably high.

Regression analysis was used to test the hypotheses. The data met the assumptions underlying the use of regression. An initial regression analysis revealed that no demographic variables were significant predictors of use intentions. Therefore none of the covariates were included in the final model.

The model variables were then analyzed, resulting in the following regression equation:

$$UI = 0.375 + 0.383RA + 0.335CP - 0.170CX + 0.012TR + 0.207RD - 0.027VI + 0.044VO$$

The F statistic for this equation is significant, indicating that the overall model is significant ($F = 36.313$, 7/178 df, $\alpha < .001$). The adjusted R^2 value for the model is 0.656, indicating that model can account for 65.6% of the sample variation in intention to use.

Next, the significance of individual terms were analyzed. These results are shown in Table 5.

FUTURE TRENDS

As new technologies and methods continue to emerge, there are a number of trends that warrant watching. The proliferation of mobile devices may impact the way in which education is conducted. Today's students seem tethered to their iPods and mobile phones. As the displays and input technologies on these devices improve, it may become possible to leverage their popularity by allowing students to use them for learning-oriented interaction. This would truly make learning anytime, anyplace. Our results indicate that designers of such systems and change agents interested in promoting their use should pay particular attention to promoting the relative advantage of these systems. Further, pointing out how the mobile, ubiquitous nature of such systems fits well with students' lives may improve compatibility beliefs, leading to greater adoption.

Table 5. Hypotheses test results

Construct	Hypothesis	Significance	Support
Relative Advantage (RA)	H1a	< 0.001	Supported
Complexity (CX)	H1b	0.035	Supported
Compatibility (CP)	H1c	< 0.001	Supported
Trialability (TR)	H1d	0.824	Not supported
Result Demonstrability (RD)	H1e	0.014	Supported
Visibility (VI)	H1f	0.601	Not supported
Voluntariness (VO)	H1g	0.475	Not supported

Students are not a homogeneous group, however. While younger users are accustomed to communicating through mobile devices, the same may not be true of non-traditional learners. Non-traditional learners are increasingly seeking new degrees and continuing education, so it is unwise to ignore their perceptions, especially since non-traditional students may hold favorable perceptions of distance learning (Stafford & Lindsey, 2007). In the future, it will become increasingly important to understand how to use ICT to meet the needs of diverse learner groups.

Because of this need, in the future it will be important to enable agile learning systems that can easily adapt to students' needs and preferences (Chatterjea, 2004). Such adaptation will positively impact perceptions of relative advantage and compatibility. We are already seeing the emergence of such systems, such as that reported by Leung and Li (2006), who use profiling methods to deploy personalized course materials.

CONCLUSION

ICT supporting "outside the classroom" interaction holds great potential for facilitating both traditional and distance learning. Our research shows that students' perceptions of these technologies are critical to their acceptance, which seems to agree with similar studies (e.g., Rentroia-Bonito, Jorge, & Ghaoui, 2006). Mandating use may lead to compliance through superficial use, but faithful use may be highly dependent on students' perceptions of relative advantage and compatibility. Open-ended comments from survey participants seem to support this thinking. Students pointed out a number of advantages to DDB, such as more frequent and timely instructor feedback and the ability to submit assignments online. Also, students pointed out that DDB facilitated easy communication among team members, another advantage of DDB.

Students also made comments related to compatibility. Recent research indicates that there must be congruence between technology that supports learning and the student's conception of the "best way to learn" (Hornik, Johnson, & Wu, 2007). Interestingly, students who had good working conditions at home (such as high-speed Internet access and a quiet workspace) were used to working at home and found DDB to be more compatible than those without suitable home conditions. Preferred communication style also seemed to be important to compatibility beliefs. Several students mentioned that DDB allowed them to interact, even though they were reluctant to speak up in class. In contrast, others said they preferred interacting personally. It is important to realize that students are not "one size fits all," so flexibility in a learning environment is important (Leung & Li, 2006; Chatterja, 2004).

While technology is a very useful tool and should be used in the classroom, both designers and faculty users should not forget the "human touch" (Lee, Tan, & Goh, 2004). In other words, according to Lee et al. (2004), IT tools are only effective when they reinforce good teaching!

REFERENCES

Agarwal, R., & Prasad, J. (1997). Role of innovation characteristics and perceived voluntariness in the acceptance of information technologies, *Decision Sciences, 28*(3), 557-582.

Anderson, J., Schwager, P., & Kerns, R. (2006). The drivers for acceptance of tablet PCs by faculty in a college of business. *Journal of Information Systems Education, 17*(4), 429-440.

Chin, W., & Gopal, A. (1995). Adoption intention in GSS: Relative importance of beliefs. *Data Base Advances, 26*(2 & 3), 42-61.

Chatterjea, K. (2004). Asynchronous learning using a hybrid learning package: A teacher developmental strategy in geography, *Journal of Organizational and End-User Computing, 16*(4), 37-54.

Chong, S., & Pervan, G. (2007). Factors influencing the extent of deployment of electronic commerce for small- and medium-sized enterprises. *Journal of Electronic Commerce in Organizations, 5*(1), 1-29.

Cooper, R., & Zmud, R. (1990). Information technology implementation research: A technology diffusion approach, *Management Science, 36*(2), 123-139.

Cordon, O., Anaya, K., Gonzalez, A., & Pinzon, S. (2007). Promoting the use of ICT for education in a traditional university: The case of the virtual learning center of the University of Granada. *Journal of Cases on Information Technology, 9*(1), 90-107.

Cronbach, L. (1970). *Essentials of psychology testing*. New York: Harper and Row.

Day, J., Lou, H., & Van Slyke, C. (2004). Instructors' experiences with using groupware to support collaborative project-based learning. *Journal of Distance Education Technologies, 2*(3), 11-25.

Dick, G., Case, T., Ruhlman, P., Van Slyke, C., & Winston, M. (2006). Online learning in the business environment. *Communications of the AIS, 17*, 895-904.

Gatignon, R., & Robertson, T. (1985). A propositional inventory for new diffusion research. *Journal of Consumer Research, 11*, 849-867.

Hazari, S. (2004). Strategy for assessment of online course discussions. *Journal of Information Systems Education, 15*(4), 349-356.

Heckman, R., & Annabi, H. (2006). How the teacher's role changes in online case study discussions. *Journal of Information Systems Education, 17*(2), 141-150.

Hornik, S., Johnson, R.D., & Wu, Y. (2007). When technology does not support learning: Conflicts between epistemological beliefs and technology support in virtual learning environments. *Journal of Organizational and End-User Computing, 19*(2), 23-46.

Ilie, V., Van Slyke, C., Green, G., & Lou, H. (2005). Gender differences in perceptions and use of communication technologies: A diffusion of innovation approach. *Information Resources Management Journal, 18*(3), 16-31.

Johansen, R. (1988). *Groupware: Computer support for business teams.* New York: The Free Press.

Lancaster, G., & Taylor, C. (1986). The diffusion of innovations and their attributes: A critical review. *Quarterly Review of Marketing, 11*(4), 13-19.

Lee, C.S., Tan, D.T.H., & Goh, W.S. (2004). The next generation of e-learning: Strategies for media rich online teaching and engaging learning. *Journal of Distance Education Technologies, 2*(4), 1-17.

Leung, E.W.C., & Li, Q. (2006). Distance learning in Hong Kong. *International Journal of Distance Education Technologies, 4*(3), 1-5.

Mahajan, V., Muller, E., & Bass, F.M. (1990). New product diffusion models in marketing: A review and directions for research. *Journal of Marketing, 54*(1), 1-26.

Moore, G., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Orlikowski, W. (1993). Learning from Notes: Organizational issues in groupware implementation. *Information Society, 9*(3), 237-250.

Rogers, E. (1995). *Diffusion of innovations.* New York: The Free Press.

Rentroia-Bonito, M.A., Jorge, J., & Ghaoui C. (2006). Motivation to e-learn within organizational settings: An exploratory factor structure. *International Journal of Distance Education Technologies, 4*(3), 24-35.

Stafford, T., & Lindsey, K. (2007). IP teleconferencing in the wired classroom: Gratifications for distance education. *Journal of Information Systems Education, 18*(2), 227-232.

Taylor, S., & Todd, P. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research, 6*(2), 144-176.

Teo, H., Tan, B., & Wei, K. (1995, December). Innovation diffusion theory as a predictor of adoption intention for financial EDI. *Proceedings of the 16th Annual International Conference on Information Systems* (pp. 155-165), Amsterdam.

Van Slyke, C., Belanger, F., & Comunale, C. (2004). Factors influencing the adoption of Web-based shopping: The impacts of trust. *Database for Advances in Information Systems, 35*(2), 32-49.

Van Slyke, C., Dick, G., & Case, T. (2005, December). Factors influencing distance learning intentions. *Proceedings of the 2005 International Conference on Informatics Education Research*, Las Vegas.

KEY TERMS

Compatibility: Degree to which an innovation is seen to be compatible with existing values, beliefs, experiences, and needs of adopters.

Complexity: Degree to which an innovation is seen by the potential adopter as being relatively difficult to use and understand.

Diffusion of Innovations: How the use of a new idea, product, or technology spreads throughout a social system.

Discussion Database: An asynchronous groupware product designed to be used by users "anytime and anyplace." The product allows threaded discussions among members of a group.

Groupware: Information and communication technologies designed to facilitate the activities of cooperative workgroups.

Relative Advantage: Degree to which an innovation is seen as being superior to its predecessor.

Result Demonstrability: Degree to which the results of using an innovation are perceived to be tangible.

Voluntariness: Degree to which use of an innovation is perceived as being of freewill.

Digital Asset Management Concepts

Ramesh Subramanian
Quinnipiac University, USA

INTRODUCTION TO DIGITAL ASSET MANAGEMENT

"DAM. Looks like something you might say if you couldn't find a photograph you needed for a front-page story. But DAM—digital asset management—is actually designed to preempt such frustrated outbursts. In an age when oodles of media, including print, images, video and audio, are stored in computers rather than file cabinets, newspapers and other groups need a way to organize, manipulate and share those media quickly and easily." (Grimes, 1998)

Dramatic changes have occurred on the corporate front in the last few years, as more and more businesses have started to conduct commerce on the Internet. New business concepts and products are being developed on a daily basis. The accent is on speed, and changes occur quickly – daily, hourly or even minute-to-minute. Two major facets of these changes are:

1. Large amounts of data are created and stored in digitized forms in organizations, and
2. New "digital products" are created.

As more and more information is created in electronic form, organizations are faced with the following problems:

- The volume of digital data has become cumbersome to manage and reuse (Sharples, 1999).
- Organizations have struggled to reduce cycle time, maintain brand consistency, and coordinate cross-media publishing as well as one-to-one marketing efforts.
- The number of digital assets that an organization may manage has exploded.
- Gistics, a California-based research firm that has studied media asset management for several years, estimates that approximately 30% of all media assets in organizations are misplaced, and then reworked or duplicated.

A 2001 Frost and Sullivan market indicator report by Subha Vivek forecasts tremendous future growth in the U.S. digital media management market (Vivek, 2001). The three market segments that will be affected represent the capture, storage and access, and distribution of digital media, respectively.

The promise of digital asset management has attracted a lot of commercial enterprises and software research laboratories, and several products have been introduced commercially in the last few years. However, due to the "newness" of the field, there is not much academic research literature in the field. A good source of academic thought in this field can be found in the online proceedings of the Annenberg DAM Conference, held at the Annenberg School of Communication, University of Southern California in 1998 (Annenberg DAM Conference, 1998).

BACKGROUND: DIGITAL ASSET MANAGEMENT (DAM) CONCEPTS

This section is adapted from our earlier paper on the subject (Subramanian & Yen, 2002).

A. Definition

A *digital asset* is any asset that exists in a digitized form, and is of intrinsic or commercial value to an organization. *Digital asset management* can be defined as a set of processes that facilitate the search, retrieval, and storage of digital assets from an archive.

B. Basic Features of DAM

The basic features of any DAM system include: storage, search and retrieval, and "thumbnail browsing" (Rosenblatt, 1998). A good DAM system will also include the ability to perform object check-in and check-out.

Other desirable features include:

- Integration of the DAM system with content creation applications on the desktop.
- Enterprise features, that is, features that are necessary for a digital media management system to be useful in a large-scale deployment at a large media company (i.e., an industrial strength, scalable database).
- The ability of a DAM system to have a user interface that can function in a cross-platform environment (e.g., the Java language from Sun Microsystems, and the development of XML technology).

- The ability to extend the functionality of the DAM system through programming interfaces.

C. What are the basic differences between DAMs and standard data management systems?

One might argue that DAMs are not much different from currently available database management systems that facilitate the search, retrieval, and storage of data. However, DAMs are different in their potential to address four key problems that pertain to the creation, storage, search and dissemination of multi-media data. According to Hilton (Hilton, 2003), those four issues are:

1. Asset mining: New and sophisticated methods for mining multidimensional, multi-media data stores, which can result in the creation of dynamic, "on-demand" digital products
2. Automation: Automated classification and retrieval systems
3. Managing intellectual property (and associated security issues)
4. Engagement: New GUIs and other data manipulation methods as well as collaboration tools

Somani, Choi and Kleewein distinguish traditional data management systems from "content management systems" that handle digital assets, communications and content such as documents, intellectual property, rich media, e-mail and Web data, and discuss the differences between the two types of systems in the following areas (Somani et al., 2002):

1. Data federation to provide in-place access to existing data
2. An expressive data model that accommodates data from very disparate sources
3. Search over metadata and data

Somani et al. then propose an architecture for integrating the two, that is, data and content management systems. A detailed discussion of the architecture is beyond the scope of this article.

DAM SYSTEM ARCHITECTURE

Figure 1 uses a three-tiered architecture of the generic DAM system architecture to show the process followed during an asset creator's session and a client's query session.

In the *asset creation flow*, the Asset Creator creates an asset, which could be in any digital format, and provides the asset and its associated information to the Asset manager. The Asset manager converts the information associated with the asset into an XML metadata format, builds the appropriate data type definitions, and passes the information and the asset to the Metadata manager. The Metadata manager manages the organization of the Metadata Store, which is a database containing meta information on the assets. Appropriate links between the metadata and the actual assets are created and maintained here. The Metadata Store contains information about the digital assets, typically in an XML DTD format. The assets are passed by the Metadata manager to the File manager, which is responsible for the check-in and check-out of the assets into and out of the Digital Asset Store, which is a combination of file systems and databases.

In the *query processing flow*, the Client asks a query. The Query processor parses the query and sends the user information as well as the parsed query to the Metadata manager, which maintains the metadata for the assets. The metadata include not only information about the asset but also information on who is allowed to access the asset. After this information is retrieved from the Metadata store, a message is sent back to the Query processor by the Metadata manager. The message passed may either be a refusal to access the asset, or an affirmation that the requested asset is being retrieved.

The metadata component acts as a store, search, retrieve and security tool, managed by the Metadata manager.

ADDITIONAL ARCHITECTURAL DETAILS

A. An Open and Distributed Architecture

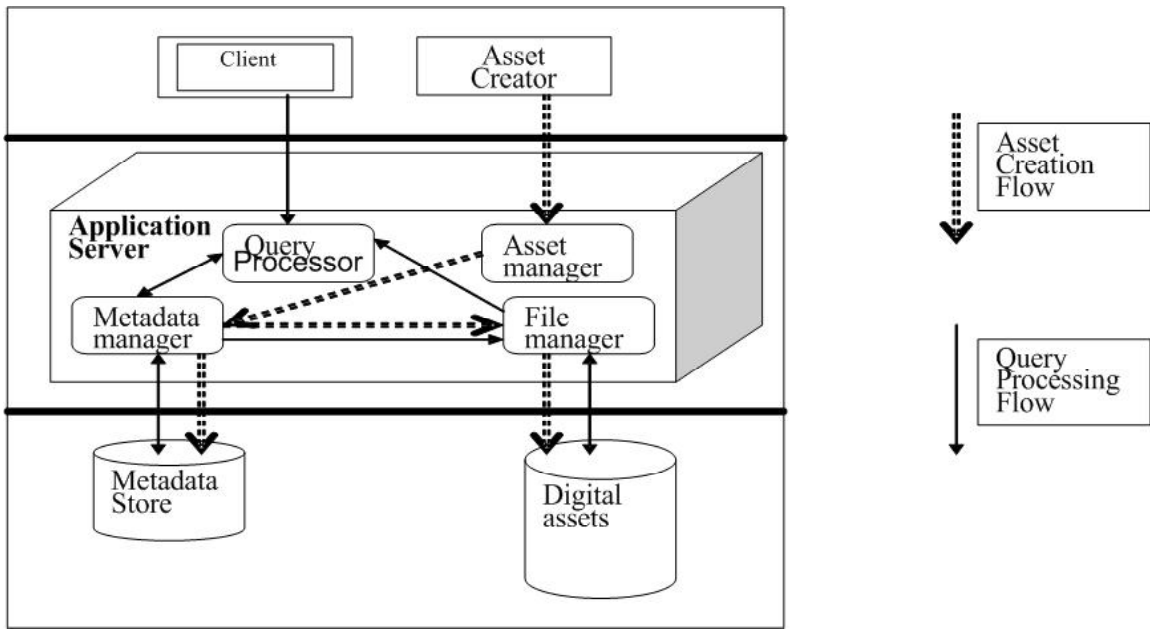
The key to any digital asset management system is to create an open and distributed architecture. A well designed DAM system first should provide the ability for people to take an asset, repository or archive and be able to customize it into their environment and extend it to their existing system and other new systems. The architecture should allow for the following features:

1. Scaling
2. User Interface and Custom Interface
3. File Management and Asset Association
4. Platform Independence and Transportability

B. Representation and Identification of Digital Assets

1. Representation Issues and Addressable Unit: The three categories of digital asset representation issues are:

Figure 1. Architecture for DAM system



Production Workflow, Creative Workflow, and the Addressable Unit (Romer, 1998). Production workflow is the ability for metadata to be added throughout the life cycle of digital asset handling (captioning, or cataloging), with appropriate knowledge management capabilities to establish standardization and consistency. Production workflows often deal with known items. Creative workflows are more discovery-oriented, hence more volatile and browse intensive. This workflow is characterized by the need to do many interactive searches and temporarily store candidate assets until a final decision can be made. A good DAM system must allow the designer or user to leverage this valuable production knowledge to help search for what one is looking for, by defining appropriate addressable units within a particular digital asset such as a digital video movie.

2. Identification Issues and Metadata Creation: Organizations are becoming increasingly aware that library science techniques need to be a very integral part of the process of identifying digital assets. DAM systems need to have a very sophisticated metadata model that will store and classify property information, details, names and descriptions, and anything else that the user defines and can be put on top of his/her digital assets. One promising method for classifying digital assets is through the use of "controlled vocabulary". This technique addresses how the hierarchical key words work, so that one can take a system, like the key word structure from the Library of Congress, and apply it to assets inside one's database.

C. Searching Digital Assets

1. "Top Down" Approach: Usually provides specialized searching routines that recognize the properties of the objects being searched. It is also referred to as "Content-based analysis". Image understanding and pattern recognition are all technologies that automatically process images based upon certain inherent properties of the image itself. Some examples are the color, texture and shape recognition systems from Virage (*http://www.virage.com*) or IBM's QBIC application (Faloutsos et al., 1994; Flickner et al., 1995). "Search extenders" facilitate content-based searching of digital files (e.g., content base searching of textural images). Bulldog's Search Extender™ technology allows the search mechanism to recognize the properties of various file types and implement specialized searching routines. They allow users to do context searching, such as finding the name "Bulldog" very close to information about Sun Microsystems inside the database. Equivalent search extender technology exists for images, video and audio.

2. "Bottom Up" Approach: A method is to design image representation schemes that will match the goals of the search process used for retrieval. The search subsystem was designed with decision support in mind.

TAXONOMY OF DAM SYSTEMS

We have categorized the DAM systems based on two criteria. The first criterion is based on how many people DAM systems need to serve, the price range and their complexity level. In this criterion, DAM can be divided into three categories (Rosenblatt, 1998):

1. The Enterprise category, to give an organization the capabilities to be scalable and with industrial strength. It typically costs in the millions of dollars.
2. The Mid-range category, where the system is purpose-built for a specific asset management purpose. They typically cost in the one hundred to five hundred thousand dollar range.
3. The Low-end category, for basic, low-end solutions that are useful for small installations. These systems cost typically cost around $50K. These are systems that just have basic features, and they run on work group class servers.

The second criterion we use is based on several representative functions of the DAM.

1. The Image Cataloging Feature: With image cataloging feature, systems capture low-resolution thumbnails.
2. The Object Repository Feature: With object repository feature, systems can capture not only thumbnails and metadata, but also high-resolution media files.
3. Content Management Back End for Web Sites Feature: This provides service/editorial environment with content management for the Web. The content can be in different file formats, such as text, graphics, multimedia, XML, and SGML documents.
4. Intelligent Cataloguing and Indexing Feature: This intelligent cataloguing and indexing feature has Web-based interface with intelligent navigation tools, allowing the user to search and organize images, video, audio and text documents easier and more efficiently.

PROTECTION AND DISTRIBUTION ISSUES OF DIGITAL ASSETS

Digital Assets Protection

The rapid growth of multimedia manipulation tools and the wide availability of network access lead to the convenience of digital data processing, delivery and storage. Powerful digital facilities can produce a large amount of perfect digital asset copies in a short period of time. The advanced compression technology also contributes to make the digital content compacted into a small data stream to facilitate transmission. These advantages benefit content users, definitely to raise concerns from content creators and content owners if the intellectual property right cannot be enforced successfully (Su et al., 1999). This is especially true in the present time, which has seen the rapid proliferation of peer-to-peer online file sharing systems such as Napster and Kazaa (Subramanian & Goodman, 2003). Yet, it has been noted by several researchers that senior corporate executives rarely pay a whole lot of attention to computer security and the security of digital assets. Austin and Darby discuss the types of threats that a company is apt to face and propose eight processes, ranging from deciding how much protection each asset receives, to insisting on secure software, to rehearsing a response to a security breach (Austin & Darby, 2003). We discuss the subject of digital assets protection from three different viewpoints, that is, legal, technical and non-technical means to ensure protection of digital assets:

1. Legal Methods: Patents, Copyrights, and Trademarks:
   A patent is a grant of exclusive right that is given by the government to anybody who files an application for patent protection. It confers upon the applicant the right to exclude everyone else from making, using or selling that patented invention. The term of protection offered by a patent is shorter than that of a copyright by a good deal. The term of protection for a patent is generally 20 years from filing throughout the world. Under the U.S. patent system, that protection is available for any invention. Anything that can be conceived of by humans and put into practice in a useful way that is new and novel can be subject to patent protection. This includes any creation of a human's mind, such as drugs, pharmaceuticals, computer software, inventions, processes, methods of manufacture, chemical compounds, electronic devices, and so forth.
   A copyright protects an original work of authorship. The term of protection for a copyright under the U.S. law is "life of the author and fifty years for the case of a work made in the course of employment by an employee working for an employer". Under certain conditions, the protection period for the latter case could be 75 years from publication or 100 years from creation, whichever expires first. There is a proposal currently in the U.S. Congress to extend the term of copyright protection by an additional 20 years, to bring it up to the level that is guaranteed in the European Union.
   A trademark is a right that attaches to a marking used to identify goods and commerce. It helps serve as a

guarantee of the source of those goods for consumers, so you know where they come from and you may have expectation as to the quality associated with that mark. Trademarks may be registered with the USPTO; they may also be registered with state authorities. Many states have their own trademark registries. Trademarks can also arise by the operation of common wall.

Copyright laws can be separated into the Anglo-American and the European copyright laws. The Anglo-American laws are followed by America, Great Britain and other British colonies. They adopt a pragmatic, market-oriented approach towards copyrights. On the other hand, under European laws, the copyright is seen as a basic fundamental human right. The economic incentives that it provides are secondary to insuring the protection of someone's personality.

2. Technical Methods for Intellectual Property Protection: Encryption and watermarking are the two most important digital assets content protection techniques. Encryption protects the content from anyone without the proper decryption key. There are two types of encryption existing: symmetric (secret-key) mechanism and asymmetric (public-key) mechanism. Symmetric mechanism uses the same security key to "lock" and scramble a digital file and to recover a bit-exact copy of the original content at the destination. Asymmetric encryption employs dual keys. The sender encrypts the digital with the recipient's public key, and the recipient decrypts it with his or her private key (Cravotta, 1999). Watermarking, on the other hand, further protects the content by embedding an imperceptible signal directly into the content for ownership declaration, play control or authentication (Su et al., 1999).
3. Non-technical methods for intellectual property protection: "This has to do with our providing products at a value above and beyond that provided by those who would pirate our products" (Griffin, 1998).

Digital Assets Distribution

In order to manage the asset owner and the user, the owner of a digital asset must be able to determine the usage scenarios that are appropriate for the asset. The scenarios are summarized by the notion of placement, size, duration, and extent. License management refers to the legal terms and conditions that apply to the specific usage scenario.

CONCLUSION AND FUTURE DIRECTIONS

In the era of e-commerce, digital asset management is emerging as a very hot topic of study for both practitioners as well as researchers. In this article we discuss different concepts and issues of DAM, which include but are not limited to the components and architecture of DAM systems, its basic and desirable features, the taxonomy of DAM systems, and protection and distribution sub-systems of DAM. There are several open issues for research in DAM. A list of these issues includes but is not limited to modeling the storage and organization of digital assets, the digital assets valuation, pricing, rights and licensing models, and methodologies for optimal search and retrieval of digital assets. In a future study, we plan to address some of these open issues.

REFERENCES

Annenberg DAM Conference. (1998). *http://www.ec2.edu/dccenter/dam/*

Austin, R.D., & Darby, C.A.R. (2003, June). The myth of secure computing. *Harvard Business Review, 81*(6).

Cravotta, N. (1999, March). Encryption: More than just complex algorithms. *EDNMagazine*. *http://www.reed-electronics.com/ednmag/index.asp?layout=article&articleid=CA56697&rid=0&rme=0&cfd=1*

Faloutsos et al. (1994). Efficient and effective querying by image content. *Journal of Intelligent Information Systems, 3,* 231-262.

Flickner et al. (1995, September). Query by image and video content: The QBIC system. *Computer, 28*(9), 23-31.

Griffin, J. (1998). Annenberg DAM Conference 1998. *Transcripts of the session "Auditing & Valuation in Different Industries: Licensing & Delivery to Consumers". http://dd.ec2.edu/1998/dam98_2a_transcript.html*

Grimes, B. (1998, November/December). Digital asset management 101. *TechNews, 4*(6). *http://www.naa.org/technews/tn981112/editorial.html*

Hilton, J.L. (2003, March/April). Digital asset management systems. *EDUCAUSE Review, 3*(2).

Romer, D. (1998). Annenberg DAM Conference 1998. *Transcripts of the session "Asset Representation in the Creative Process: Bringing Media to Life". http://dd.ec2.edu/1998/dam98_2b_transcript.html*

Rosenblatt, W. (1998). Annenberg DAM Conference 1998. *Transcripts of the Session "Storage / Retrieval / Presentation". http://dd.ec2.edu/1998/dam98_1b_transcript. html*

Sharples, H. (1999, November). Sights set on site. *Graphic Arts Monthly,* 52-54.

Somani, A., Choy, D., & Kleewein, J.C. (2002). Bringing together content and data management systems: Challenges and opportunities. *IBM Systems Journal, 41*(4), 686.

Su, Po-Chyi et al. (1999, April 25-28). Digital image watermarking in regions of interest. *Proceedings of the IS&T Image Processing/Image Quality/Image Capture Systems (PICS),* Savannah, Georgia. *http://biron.usc.edu/~pochyisu/pochyi_files/main.htm*

Subramanian, R., & Goodman, B. (2003). Peer-to-peer corporate resource sharing and distribution with mesh. *Proceedings of the 14th IRMA International Conference.*

Subramanian, R., & Yen, M. (2002). Digital asset management: Concepts and issues. In A. Gangopadhyay (Ed.), *Managing business with electronic commerce: Issues and trends*. Hershey, PA: Idea Group Publishing.

Vivek, S. (2001). DAM: It's consolidation time. Frost and Sullivan "Market Indicator" Report. *https://www.frost.com/prod/servlet/market-insight.pag?docid=RCOL-4ZEMMP&ctxht= FcmCtx4&ctxhl=FcmCtx5&ctxixpLink=FcmCtx5 &ctxixpLabel= FcmCtx6* (requires account).

KEY TERMS

Addressable Unit: A specific unit within a particular digital asset such as a digital video movie.

Asset Creator: Anyone who creates an asset, which could be in any digital format, and provides the asset and its associated information to the asset manager.

Asset Manager: The asset manager converts the information associated with the asset into an XML metadata format, builds the appropriate data type definitions, and passes the information and the asset to the metadata manager.

Creative Workflow: These are more discovery-oriented, hence more volatile and browse intensive. This workflow is characterized by the need to do many interactive searches and temporarily store candidate assets until a final decision can be made.

Digital Asset: A digital asset is any asset that exists in a digitized form, and is of intrinsic or commercial value to an organization.

Digital Asset Management (DAM): Digital asset management can be defined as a set of processes that facilitate the search, retrieval, and storage of digital assets from an archive.

Digital Asset Store: A combination of file systems and databases.

Metadata Manager: The metadata manager manages the organization of the metadata store, which is a database containing meta information on the assets.

Metadata Store: The metadata store contains information about the digital assets, typically in an XML DTD format.

Production Workflow: It is the ability for metadata to be added throughout the life cycle of digital asset handling (captioning, or cataloging), with appropriate knowledge management capabilities to establish standardization and consistency. Production workflows often deal with known items.

Search Extenders: Methods that facilitate content-based searching of digital files.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 864-869, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Global Information Technology

From Digital Divides to Digital Inequalities

Francesco Amoretti
University of Salerno, Italy

Clementina Casula
University of Cagliari, Italy

INTRODUCTION

Concerns about inequalities deriving from the penetration of new information and communication technologies (ICTs) have only recently become a widely debated issue in industrial societies. Until the 1980s the diffusion of ICT was mainly considered a matter of technological innovation regarding selected fields and limited territorial areas (such as the military and academic centers in the U.S.). Gradually, scholars started to point to the rise of an *information society* based on the production of information as the crucial resource to manage coordination and control of increasingly interconnected organizational systems (Masuda, 1981; Beniger, 1986; Toffler, 1990). The expression offered an alternative to the otherwise negative definitions used by scholars since the 1970s to identify changes occurring in Western democratic societies ('post-capitalism', 'post-industrialism', 'post-materialism', etc.) (Touraine, 1969; Bell, 1973).

The debate over the information society, enthusiastically greeted by some authors (Negroponte, 1995) and critically observed by others (Castells, 1996, 2001; May, 2002; Mattelart, 2003), witnessed since the mid-1990s widespread success in public and political debates (Thomas, 1996). In front of the fast and capillary diffusion of ICTs virtually to all sectors of private and public life, most Western countries' governments and international organizations have inserted within their policy agendas a reference to the unavoidability, if not desirability, of a radical shift to the new information age. The rhetoric accompanying those discourses often presents the expansion of the ICT sector—and especially the Internet—as offering citizens returns at both the individual and collective level, in the form of greater access to goods and services, increased levels of social and civic participation, and wider economic and working opportunities for all. Presented as a crucial means to participate in the new global information society, ICTs become recognized as a resource that should be fairly distributed among citizens, albeit on the basis of different arguments (ranging from social equity to economic efficiency or global development concerns), often leading to opposite conclusions on the scope for redistributive interventions (Strover, 2003; Selwyn, 2007).

BACKGROUND

The first framing of the issue of digital inequalities was in terms of a *digital divide* indicating the distance between the 'haves' and the 'have nots' in access to ICTs, mainly with reference to the Internet (NTIA, 1999). At the territorial level the analysis focused on the wide gap existing between most industrialized and Third World countries in access to ICTs (*global digital divide*), although it was also applied at the sub-national level (with reference to the disadvantages of peripheral or rural areas): in both cases existing gaps were mainly explained with reference to the unequal spatial distribution of socio-economic wealth between centers and peripheries. The measurement of *social digital divides* (i.e., gaps between the different social groups) was mainly considered within countries, because of the peculiarities of different socio-institutional contexts.

The widely registered existence of unequal access to ICTs was differently interpreted. We can draw an ideal type, analytically distinguishing three positions on the debate, ranging from a more to a less optimistic view, and consequently drawing an increasingly active role for policy intervention: (a) ICT diffusion may particularly favor disadvantaged groups; (b) ICT distribution becomes increasingly even with their diffusion; and (c) ICT diffusion follows and reinforces existing inequalities.

The first position holds an optimistic view both in terms of equal access to ICTs and the opportunities that free market developments in the ICT sector may offer in terms of redressing inequalities actually structuring societies at different levels. An example of this position can be found in the so-called *leap-frog hypothesis,* referring to the fact that poor countries investing in the latest ICTs may develop into an information society skipping some of the difficult stages (in terms of political, economic, and social problems) faced by developed countries that heavily invested in older information technology (Butler, 1999; James, 2001). Another example is offered by those arguments stressing the *death of distance* created by ICTs, in terms of improving the quality of life, particularly for disadvantaged individuals or communities in terms of offering new services (empowering inhabitants of peripheral or disadvantaged areas, women having to reconcile work and family) and enhancing civic participation (the *mobilization hypothesis* refers to the role of ICTs in helping

citizens actually marginalized from their political system to get informed, organize, or engage in public life).

The second position interprets the diffusion of ICTs as following a 'natural path', already undertaken by other mass media (telephone, radio, TV). Initially access to the new technology is restricted to an *élite*, with a large divide between the 'haves' and the 'have-nots', but with time its increasing penetration within society progressively reduces gaps. The *normalization thesis* is based on the idea that a series of factors linked with the development of technology (increasingly lower costs, user-friendly access, differentiated contents) will create a saturation in the market allowing 'have-not' groups to access innovation. From this view, the role played by processes of liberalization of the ICT sector, reducing the digital divide due to increased competition in the telecommunication market, is often praised (OECD, 2001; Dutta & Mia, 2007). However a limited contribution to governmental action is also allowed as far as it enhances this path, through policy measures aiming to facilitate the development of the ICT sector (infrastructure building, facilitated connections in schools and other public institutions, introduction of digital alphabetization in education programs) and market liberalization (regulatory actions to grant free competition).

The third position, identifiable with the *stratification thesis,* argues that there is a strong positive relation between distribution of access to ICTs and the main social inequalities related to different variables: economic wealth, education, gender, location, age, and ethnicity (Norris, 2001). This perspective revives the *knowledge gaps model* (Tichenor, Donohue, & Olien, 1970), in that it argues that the segments of population with the higher socio-economic status tend to have easier access to ICTs, and thus to knowledge, than those with a lower status and that, with time, this increases distance between the two groups. From this perspective governmental action will be fundamental in devising policies to reduce digital divides, specifically targeted for disadvantaged groups (for instance with specific services for SMEs or rural communities, or vocational training courses for women or the elderly). The inclusion of the 'digital divide' issue in the policy agendas of most Western governments during the 1990s has also been related to the wider emphasis on social inclusion legitimized by 'New-Left' governments at that time in power in the wealthiest countries (Selwyn, 2004).

THE INCREASING COMPLEXITY OF DIGITAL INEQUALITIES STUDIES

The differences between the positions enounced reveal an increasingly rich and complex picture of the debate on inequalities and ICTs, as their diffusion further develops and a body of data and literature goes more thoroughly into the matter (Van Dijk & Hacker, 2003).

The first position has the merit of having emphasized the relevance of investing in innovation, information, and knowledge as crucial sectors for a development considered not only in terms of growth, but also as a fairer distribution of resources at the global and social level. In so doing, it builds upon initial enthusiasm on the potentials of ICT, considered as an intrinsic democratic medium because of the logic of its networked structure, leading to a reduction in social differences and space-time coordinates. However the rhetoric over the democratic potential of the rising information age must be tempered by the reality that its effects are benefiting a limited part of the world's population (OECD, 2001; Norris, 2001). Findings also suggest that the countries with strongly developed information societies are those that invested heavily in early communications infrastructures and reaped economic benefits that propelled them forward in development (Howard, 2007).

The fact that the age curves usually show increasing access of younger generations to ICTs, due to the lowering of diffusion costs of the technology and the growing digital alphabetization of societies, has been interpreted as a confirmation of the 'normalization thesis' (Moschella & Atkinson, 1998). The thesis, arguing that the diffusion of technologies is followed by spillover effects and positive externalities, builds on the analogy with the case of radio and television. However, while the abilities needed to use the 'old media' are quite intuitive, they are more demanding in the case of ICTs. To avoid falling in a technological determinism fallacy, the study of digital inequalities also must consider how, beyond access, the use of ICTs relates to the different social groups and institutional contexts considered (Wilson, Wallin, & Reiser, 2003; Carter Ching, Basham, & Jang, 2005).

The third position contributed to the identification of a plurality of digital divides related with social inequalities, showing that they need to be considered from a multidimensional perspective (Bertot, 2003). However, its reference to an idea of 'strong media' has as its main drawback the assignment of a relatively passive role to social actors, which does not allow acknowledgment of the relevance of exceptions (as in the case of gender divides, the movement of *cyberfeminism* or data relative to women's overtaking in some countries; Plant, 1996; Tsaliki, 2001). Those difficulties seem to be related to the main limits of the initial framing of the debate on inequalities and ICTs. On the one hand id the reference only to 'access' to ICTs, considering information as something physical that one can easily accumulate or redistribute, as in the mathematical model of communication (Shannon & Weaver ,1949). On the other hand is the consideration of inequalities only in terms of 'haves' and 'have nots', referring to an ideal concept of 'simple equality' as a distributive condition inadequate to address the issue of inequality with reference to a plurality of social spheres (Walzer, 1983).

These critiques have prompted researchers to broaden the frame defining the issue inequalities and ICTs. This has been pursued, on the one hand, by considering the evolution of information and communication technologies as a collective and shared process embedded within different socio-institutional contexts; and on the other hand, by defining the issue of inequalities related to ICTs not as a social problem in itself, but when determined by the distribution of other material or immaterial resources. From this wider perspective research has addressed the issue of the many origins and consequences of differences in ICTs, not only in terms of access but also with reference to use, emphasizing how it is affected by a variety of factors such as level of autonomy, adequate cognitive and linguistic skills, and social network support. At their turn, those factors are related to individual characteristics (motivation, age, gender, race, IQ), the specific arenas of social interaction (family, school, labor market), and the wider socio-institutional and political context (location, presence of a cosmopolitan culture, openness of political regime, public investments in R&D), offering actors both incentives and constraints defined by contexts of opportunities as well as patterns of social inequalities (DiMaggio & Hargittai, 2001; Katz & Rice, 2002; De Haan, 2004; Milner, 2006).

The widening of the theoretical frame defining the issue of inequalities and ICTs has already provided new, interesting insights from empirical research on the topic, although contributions are usually limited in their scope, and investigate practices and use, focusing on case studies or using specific research designs. Studies on the relation between ICTs and democracy have shown how public policies defined within the discursive framework of 'e-democracy' are later translated in actions more interested in containing the decline of popular consensus than in launching a process of democratic reform (Amoretti, 2006) or in favoring the diffusion of alternative sources of information (Kalathil & Boas, 2003). Other evidence suggests that, even in front of closing 'digital divides' in the case of access to ICTs, the analysis of differences in the practices linked to the use of ICTs reveals a relation with material, cognitive, and social resources (De Haan, 2004). For instance, gender divides in the use of ICTs have been related to the 'symbolic violence' inbred in technology practices in different social spheres (school, household, work) linked to naturalized male-dominated cultures (Casula & Mongili, 2006). Thus, forms of 'cyber-phobia', sometimes preventing women or the elderly to use computers, might also be read as processes of self-exclusion responding to an already defined social hierarchy. The accumulation of advantages following the acquisition of digital resources (following the so-called *Matthew effect*) is thus reconsidered as part of a complex social process, rather than a separate trend, where resources related to ICTs (the 'digital capital') also interlink with other (economic, cultural, social) resources identified as crucial to the creation of social stratification in contemporary society (Bourdieu, 1979; Coleman, 1990). This also explains the relevance given by some authors to the availability for the individual of social networks, which can help him or her in gradually overcoming fears in access to ICTs, diversifying their use and solving technical problems (DiMaggio, Hargittai, Celeste, & Shafer, 2003; Sartori, 2006).

Developments within the academic debate on the issue of digital inequalities have been followed by modest changes in the political discourse, still bound to the 'digital divide' formulation. This has to do with the fact that the increasing complexities of the debate (offering alternative views, definitions, and datasets discouraging the achievement of definitive answers and 'one best way' prescriptions) do not easily combine with the 'political-electoral-cycle' logic, driving politicians to consensus-building rhetoric and short-term outcome actions.

FUTURE TRENDS

To date, research on ICTs and inequalities has mostly concentrated on access to the Internet and possession of computers; as partly anticipated, this approach is increasingly deemed as in need of further elaboration. Along with recent developments in the 'multimedia' sector, the field of analysis should be revised and broadened also with reference to the medium itself. In this regard particularly promising for future research seems the case of mobile phones, both because of their increasing integration of different services and their wide diffusion crossing social stratification.

The shift of the debate from the issue of access to that of use of ICTs, as well as the recognition of their 'embeddedness' in different socio-institutional contexts, suggests that further efforts are needed to tackle digital inequalities. Differently from the policies promoting access, focused on the provision of equal access for all, those for use do not aim to equalize behavior but rather to regulate new problematic areas related to the use of ICTs according to democratic values. In this regard, for instance, new instruments are actually being devised in most democratic countries in order to update the protection of rights (such as that to privacy, freedom of speech, intellectual property), or to define different forms of preventing, controlling, and fighting against the illegal uses of ICTs (from fraudulent commercial practices to paedo-pornography or terrorism).

CONCLUSION

Until recently, concerns over digital inequalities have been mainly addressed in the public debate in terms of the existence of a 'digital divide', identifying unequal distributions of individual access to ICTs. However, developments of research in the field have shown both the existence of a plurality of

digital divides, typically related to social stratification, and the closing up of others as technologies diffuse or due to specific policy action. The latest contributions further enriched the debate, shifting the analysis of inequalities from access to use of ICTs, considering the factors that both at the individual and socio-institutional level may affect the possibility of exerting an aware and proper utilization. In this respect, the role of governments becomes relevant not only in favoring investments relative to the ICT sector, but mostly in devising new instruments regulating the field of ICTs to protect citizens' rights and enabling them to benefit from their use. Because of its shifting nature and increasing complexity, the issue of digital inequalities seems bound in the next future to constant empirical investigations and theoretical redefinitions.

REFERENCES

Amoretti, F. (2006). La rivoluzione digitale e i processi di costituzionalizzazione Europei. L'e-democracy tra ideologia e pratiche istituzionali. *Comunicazione Politica, VII*(1), 49-75.

Barney, D. (2004). *The network society.* Cambridge: Polity Press.

Bell, D. (1973). *The coming of post-industrial society.* New York: Basic Books.

Beniger, J.R. (1986). *The control revolution: Technological and economic origins of the information society.* Cambridge, MA: Harvard University.

Bertot, J.C. (2003). The multiple dimensions of the digital divide: More than technology "haves" and "have nots." *Government Information Quarterly, 20,* 182-191.

Bourdieu, P. (1979). *La distinction. Critique sociale du jugement.* Paris: Editions de Minuit.

Butler, D. (1999). Internet may help bridge the gap. *Nature, 397*(6714), 10-11.

Carter Ching, C., Basham, J.D., & Jang, E. (2005). The legacy of the digital divide: Gender, socioeconomic status, and early exposure as predictors of full-spectrum technology use among young adults. *Urban Education, 40,* 394-411.

Castells, M. (1996). *The rise of the network society.* Cambridge: Blackwell.

Castells, M. (2001). *Internet galaxy.* Oxford: Oxford University Press.

Casula, C., & Mongili, A. (2006). *Donne al computer.* Cagliari: CUEC.

Coleman, J.S. (1990). *Foundations of social theory.* Cambridge: Belknap Press.

De Haan, J. (2004). A multifaceted dynamic model of the digital divide. *IT& Society, 1*(7), 68-88.

DiMaggio, P., & Hargittai, E. (2001). *From the 'digital divide' to 'digital inequality': Studying Internet use as penetration increases.* Working Paper 15, Center for Arts and Cultural Policy Studies, Princeton University, USA.

DiMaggio, P., Hargittai, E., Celeste, C., & Shafer, S. (2003). *From unequal access to differentiated use: A literature review and agenda for research on digital inequality.* Working Paper 29, Center for Arts and Cultural Policy Studies, Princeton University, USA.

Dutta, S., & Mia, I. (Eds.). (2007). *The global information technology report 2006-2007: Connecting to the networked economy.* New York: Palgrave Macmillan.

Howard, P.N. (2007). Testing the leap-frog hypothesis: The impact of existing infrastructure and telecommunications policy on the global digital divide. *Information, Communication and Society, 10*(2), 133-157.

James, J. (2001). Bridging the digital divide with low-cost information technologies. *Journal of Information Science, 27*(4), 211-217.

Kalathil, S., & Boas, T.C. (2003). *Open networks, closed regimes: The impact of the Internet on authoritarian rule.* Washington, DC: Carnegie Endowment for International Peace.

Katz, J.E., & Rice, R.E. (2002). *Social consequences of Internet use: Access, involvement and interaction.* Cambridge, MA: MIT Press.

Masuda, Y. (1981). *Information society as post-industrial society.* Bethesda, MD: World Future Society.

Mattelart, A. (2003). *The information society: An introduction.* London: Sage.

May, C. (2002). *The information society: A skeptical view.* Malden: Polity Press.

Milner, H.V. (2006). The digital divide: The role of political institutions in technology diffusion. *Comparative Political Studies, 29*(2), 176-199.

Moschella, D., & Atkinson, R.D. (1998). *The Internet and society. Universal access, not universal service.* Policy Report, Progressive Policy Institute, USA.

Neef, D. (Ed.). (1998). *The knowledge economy.* Boston: Butterworth-Heinemann.

Negroponte, N. (1995). *Being digital*. New York: Vintage Books.

Norris, P. (2001). *Digital divide: Civic engagement, information poverty and the Internet worldwide.* Cambridge: Cambridge University Press.

NTIA (National Telecommunications and Information Administration). (1999). *Falling through the Net: Defining the digital divide.* Retrieved from http://www.ntia.doc.gov

OECD. (2001). *Understanding the digital divide.* Paris: OECD Publications.

Plant, S. (1996). Beyond the screens: Film, cyberpunk and cyberfeminism. In S. Kemp & J. Squires (Eds.), *Feminisms*. Oxford: Oxford University Press.

Schramm, W.L. (1964). *Mass media and national development: The role of information in the developing countries.* Stanford, CA: Stanford University Press.

Sartori, L. (2004). *Il divario digitale: Internet e le nuove disuguaglianze sociali.* Bologna: Il Mulino.

Selwyn, N. (2004). Reconsidering political and popular understandings of the digital divide. *New Media and Society, 6*(3), 341-362.

Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communication.* Urbana: University of Illinois Press.

Strover, S. (2003). Remapping the digital divide. *The Information Society, 19,* 275-277.

Thomas, R. (1996). *Access and inequality.* In N. Heap, R. Thomas, G. Einon, R. Mason, & H. Mackay (Eds.), *Information technology and society.* London: Sage.

Tichenor, P., Donohue, G., & Olien, C. (1970). Mass media and differential growth in knowledge. *Public Opinion Quarterly, 34,* 158-170.

Toffler, A. (1990). *The third wave.* New York: Batman Books.

Touraine, A. (1969). *La société postindustrielle.* Paris: Denoël.

Tsaliki, L. (2001). Women and new technologies. In S. Gamble (Ed.), *The Routledge companion to feminism and postfeminism.* London: Routledge.

Van Dijk, J., & Hacker, K. (2003). The digital divide as a complex and dynamic phenomenon. *The Information Society, 19,* 315-326.

Walzer, M. (1983). *Spheres of justice: A defence of pluralism and equality.* Oxford: Basil Blackwell.

Wilson, K.R., Wallin, J.S., & Reiser, C. (2003). Social stratification and the digital divide. *Social Science Computer Review, 21,* 133-143.

Young, J.R. (2001). Does digital divide rhetoric do more harm than good? *Chronicle of Higher Education, 48*(4).

KEY TERMS

Cyberfeminism: Feminist movement interpreting the evolution of cybernetics as allowing the development of a culture in which inequalities are eradicated and traditional gender relations and stereotypes are defied (for instance, through the experimentation with gender identities or the creation of sisterhood networks on the Internet), empowering women and marking a shift away from their traditional symbolic representation as technologically ignorant.

Information Society: Expression used by scholars since the 1970s, but gaining popularity 20 years later, identifying in the expansion and ubiquity of information and communication technologies in contemporary societies the rise of a new age centered on information and thus influenced by the network morphology (along which information flows) in the organization of time and space as well as other social spheres, from the local to the global level.

Knowledge Gaps Thesis: The position arguing that, in the penetration of mass media within a social system, the segments of a population with higher socioeconomic status will retrieve information faster than those with a lower socioeconomic status; as a result of this positive feedback, with time the knowledge gap between the two social segments will increase rather than decrease. Originally applied to the diffusion of the TV, it has recently been revived by the *stratification thesis* in the case of ICTs.

Leap-Frog Hypothesis: An optimistic view of the potential of information and communication technologies (ICTs) in enhancing a fairer global development, referring to the fact that poor countries investing in the latest ICTs may develop into an *information society,* skipping some of the difficult stages (in terms of political, economic, social problems) faced by the more developed countries that heavily invested in older ICTs.

Matthew Effect: An expression introduced by sociologist R. Merton (with reference to the Evangelist's sentence "unto everyone that hath shall be given") to indicate a process of social accumulation, and lately also applied to the debate on inequalities and ICTs. Refers to a mechanisms of accumulation of advantages, in science as in other occupational spheres, when certain individuals or groups repeatedly receive resources and rewards, enriching them at an accelerating rate and leading in the medium-long term to stratification and élite formation.

Mobilization Hypothesis: An optimistic view fn the potential of information and communication technologies (ICTs) in terms of enhancing social and political inclusion, which tends to consider ICTs as an intrinsically 'democratic' means that may help citizens actually marginalized from their political system to get informed, organize, or engage in public life.

Normalization Thesis: The position arguing that inequalities in access to information and communication technologies are mainly linked to a first stage of penetration of the technology—the normal path—but gradually decreasing in the later stages when the technology is adopted widely across society because it becomes cheaper, easier, and more effective; in contrast to the *stratification thesis,* this view posits an optimistic scenario where also less privileged social groups can fully participate in the *information society.*

Stratification Thesis: The position arguing that inequalities in access to information and communication technologies are mainly linked to the relatively stable hierarchies positioning individuals and groups within a social system, allowing some of them more privileged position than others. In contrast to the *normalization thesis,* this position takes back the *knowledge gaps* stance, stressing the crucial role of information as a source of power in contemporary society to explain access to information as a new source of social, economic, and political differentiation.

Digital Game-Based Learning in Higher Education

Sauman Chu
University of Minnesota, USA

INTRODUCTION

Since the 1970s, computer games have become more and more a part of our culture. For the past 20-30 years, some studies have shown that certain aspects of computer games may have potential benefits in a learning environment. For example, games may increase motivation, collaboration and competition, as well as provide an effective inquiry-based framework (Squire, 2002). With the extreme success of the gaming industry in recent years, the potential for using computer games as a teaching tool in higher education is being increasingly explored. Game-based learning has been used in the military, medicine, and physical education quite successfully. However, the rules and guidelines of incorporating digital game-based learning into education are still quite open and exploration of possibilities is greatly encouraged.

BACKGROUND

To understand the concept of a digital game, and be able to create a game, one must understand the components of a game. In other words, what are the elements that comprise a game?

Definition of a Game

Salen and Zimmerman's Rule of Play: Game Design Fundamental (2004) provides a comprehensive discussion of the definition of a game. The book reviews and compares eight different models of game definition by David Parlett, a game historian; Clark Abt, also a game historian; Johann Huizinga, an Anthropologist; Roger Caillois, a Sociologist; Bernard Sutis, a Philosopher; Chris Crawford, a computer game designer; Greg Costikyan, a game designer and writer; and Elliot Avedon and Brian Sutton-Smith, both scholars of play and games. Each of these scholars/professionals provides his own framework of the elements that comprise a game. These elements can be defined and grouped into 15 categories: (1) proceeds according to rules that limit players; (2) conflict or contest; (3) goal/outcome-oriented; (4) an activity, process or event; (5) involves decision-making; (6) not serious and absorbing; (7) not associated with material gain; (8) outside ordinary life; (9) creates special social groups; (10) voluntary based; (11) uncertain quality; (12) make-believe or representational; (13) inefficient; (14) resources and tokens; and (15) a form of art.

Based on their analysis, Salen and Zimmerman provide their own definition of a game: "A game is a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome. The key elements of this definition are the fact that a game is a system, players interact with the system; a game is an instance of conflict, the conflict in games is artificial, rules limit player behavior and define the game, and every game has a quantifiable outcome or goal" (Salen & Zimmerman, 2004, p. 83). Therefore, one can assume that digital games should include these components, and that digital games are like every other kind of game.

Definition of Digital Game

The major characteristic of a digital game that differs from other games is that the game itself is composed of computer hardware and software. According to Salen and Zimmerman (2004), there are four characteristics that summarize the qualities and capabilities of digital games: immediate but narrow interactivity, manipulation of information, automated complex system, and networked communication.

1. **Immediate but Narrow Interactivity:** One of the common elements of a digital game is that it can offer immediate and interactive feedback. However, the interaction that the player can have with a computer is quite narrow. For instance, one's interaction with a home computer is usually restricted by using a mouse, keyboard, screen, and speakers.
2. **Manipulation of Information:** Digital games are usually filled with various kinds of visual information such as text, images, video, audio, as well as 2D and 3D animations. Besides this visual information, digital games are also capable of handling internal logic, player interactivity, or even hiding the information and only revealing it under certain circumstances.
3. **Automated complex system:** Digital games can automate complicated procedures that are created to facilitate the play process. In most nondigital game environments, a player's direct input is necessary in

order to move the game forward. In a digital game, however, the program can automate these procedures without direct input from a player.

4. **Networked Communication:** Digital games can facilitate communication between players. For example, players can use e-mail, text chat, and audio communication that are digitally mediated. The major advantage of this form of communication is that it offers (the players) the ability to communicate over long distances, and to share a complex social space with many other participants.

The Relationship of Digital Game and Learning

Learning is powerful when it is personally meaningful, experiential, social, and epistemological (Shaffer, Squire, Halverson, & Gee, 2005). The focus of leaning is the interconnection between tools, resources, activities, and works that are composed in a particular learning environment. From a constructivist perspective of learning, knowledge is constructed when learners play an active role in the learning process by exploring, manipulating, and interacting within the learning environment. This reflects that learners have more control in the learning process, and that the environment must provide the structure to foster learners (Dickey, 2006).

The virtual worlds that are created by digital games are what make video games a powerful learning environment. In a virtual world, words and symbols are grounded in contexts; things and objects co-exist and correlate to each other. Through the experience of playing or the process of problem solving, learners can understand complex concepts by connecting abstract ideas with real problems. Games then integrate knowing and doing (Shaffer et al., 2005). In addition, games provide a goal-oriented environment that encourages the player to understand and solve problems through the provided tools. Therefore, the focus of gaming is reaching goals and solving problems, rather than simply learning facts and information.

Gee (2005) stated that the domain of knowledge is composed of ways of doing, being and seeing. Pivec, Dziabenko, and Schinnerl (2004) suggested that computer games can support and facilitate the learning process by providing a variety of presentations, and creating opportunities to apply knowledge within a virtual world. Playing games requires deep thought and complex problem solving skills (Gee, 2004). In the game environment, learners are encouraged to combine knowledge from different areas, and choose a solution. Learners are required to make decisions at certain points in a game in order to proceed, and they can experiment with how the outcome of a game may change based on their decisions. Games are simulations with a goal-oriented structure in which the learner has a definite objective and desired outcome. Salen and Zimmerman (2004) suggested that what makes games so appealing is that they give users meaningful choices.

Stoney and Oliver (1998) suggested eight attributes of an interactive multimedia leaning environment that affect the motivation and engagement for adult learners: immersion, reflection, play and flow, collaboration, learner control, curiosity, fantasy, and challenge. There seems to be a close relationship between playing and learning. Rieber (1996) indicated that play is an important part of a child's psychological, social, and intellectual development because play is motivating, and it involves make-believe thinking. Digital games allow players to think, talk, and act in new ways in which they take on representational or make-believe roles that are otherwise unattainable to them (Shaffer et al., 2005). In addition, computer games enhance learning through visualization and experimentation. Visualization has tremendous value in computer games because it is the main cognitive strategy for discovery and problem solving.

COMPUTER GAMES IN HIGHER EDUCATION

The Roles of Digital Game in Education

"Games and education, education and games, learning, play, theory, whatever the permutation, whatever the words, this is a deeply popular topic, almost a movement if some hierophants are to be believed" (Sefton-Green, 2005, p.441). The current learning generation is the key influence on this game movement. Microsoft's investment in MIT for the project, the education arcade, demonstrates a commitment to developing and researching games and education.

Digital game-based learning has been introduced in various educational settings for the purpose of creating a more engaging environment for the learners or so called "computer natives." The learning generation today is extremely game literate (Dekanter, 2004). David (1997) found an increasing demand from learners for greater interactivity in learning materials. A complex level of interactivity is required to stimulate learners' engagement. Squire (2005) stated that e-learning educators, in particular, spend a significant amount of time building learning environments from games. The question is not whether educators should use games to support learning, but how educators can use games more effectively as educational tools.

The integration of technology into the classroom allows educators to explore new technology-mediated spaces and environments for teaching and learning. The challenge in education is how to form a learning environment that can take advantage of the power and potential of the virtual world. Most educational games are not grounded in learning theory

and they lack underlying research (Shaffer et al., 2005). As Dickey (2006) suggested, new models and methods must be explored and used in order to create effective digital learning materials that will enhance learners' engagement and motivation.

Exploration of the potential uses for digital games in education should not be seen as a failure of the educational system. Nor should it create an expectation that gaming is the ultimate solution for students' learning issues. The purpose of analyzing the theory and conceptual framework of computer games is to examine possibilities for enhancing students' learning abilities. Robertson (2002) indicated that computer games have enormous educational potential if they are used responsibly and appropriately. Computer games are not intended to replace traditional teaching methods, but rather, to provide an additional method for transferring knowledge to others.

Research on Digital Game-based Learning

Lawson (2003) suggested that an adult social play gaming approach could help landscape architecture students obtain key knowledge and skills for land-use decision making. A study conducted by Amory, Naicker, Vincent, and Adams (1999) identified the type of game that was most appropriate for their teaching environment at a college level, and investigated game elements that students found useful within various types of games. Results showed that students rated logic, memory, visualization, and problem solving as the most essential game elements. Adventure games were the most highly preferred type of game. Students also indicated that sounds, graphics, story lines, and the use of technology are important motivators. Numerous memory retention studies have shown that game-based learning has better outcomes, and students are very much in favor of learning from interactive games (Pivec et al., 2004).

Another study conducted by Mann, Eidelson, Fukuchi, Nissman, Robertson, and Jardines (2002) used an interactive game-based learning module related to breast problems for training medical school students in surgery education. The game was created as a problem-based learning experience in which students were required to collect needed information to solve problems. The overall mean test scores among 32 student participants increased after the participants played the module. Participants also agreed that the game approach was helpful for imparting knowledge about breast disease.

Early research indicated that games intrinsically motivate users due to the presence of challenges, fantasy, curiosity, novelty, and complexity (Malone, 1981). In other words, although some games might not be educational, they are at least motivating. Squire (2005) suggested that future studies on educational games should examine how different players experience different games, as well as the relationship between those experiences and learning.

Problems of Digital Game-Based Learning

Edutainment software is defined as the combination of electronic games and the goal of achieving an educational purpose. Okan (2003) suggested that this approach has a long-term harmful effect to learners because it creates an expectation that learning should always be colorful and fun, and that the acquisition of knowledge cannot be a serious endeavor. He argues that education is concerned with the development of cognitive structures, and that technology is only a medium. However, the edutainment approach suggests that if students are not enjoying themselves, they are not learning. Therefore, if learning feels difficult, it is viewed as an obstacle to overcome. This trivializes the learning process.

Okan (2003) also addresses that the fact that currently, most attention is on how to use games as a tool for increasing students' motivation and engagement within the learning context. However, no attention has been paid to the fundamental issue of the impact of games on bringing about a change in the definition of learning. There is limited academic literature that explores the benefits of game-based learning. In fact, there is little existing theory in game design even though game technology is advancing (Smith & Mann, 2002). Druckman (1995) stated that game-based learning is not a superior learning method, because although games enhance motivation, and increase students' interest in the subject matter, it is very unclear as to whether games are an effective teaching method.

A further suggestion by Okan pointed out that animation is only beneficial for learning if the learner engages in active cognitive processing. Further research needs to be conducted on the effectiveness of using animation within the learning environment; such research should be based on well supported cognitive theory.

FUTURE TRENDS

There are potential subject areas where game-based learning could be explored. Specifically, these areas include (1) game-based learning to assist people with learning disabilities, (2) game-based learning for younger children, (3) computer games that deal with social and cultural issues, and (4) how games assist the learning of foreign languages. Future studies to examine the potential for applying computer games in education should be grounded in learning theory. Usability testing should also be done to provide necessary data that will ultimately help to develop guidelines for the

development of educational games. Most recently conducted studies have focused on investigating and measuring how much information students have learned from using computer games. More studies could be conducted to examine what subject matter or disciplines would benefit substantially from applying computer games as a teaching method.

CONCLUSION

Recent research studies and writings address the advantages and benefits of using a computer game model as a teaching strategy. Positive outcomes have been discovered, and students' learning responses to this approach are certainly encouraging. Game-based learning in higher education will be an important teaching and research focus in years to come. The MacArthur Foundation initiated its 5-year funding program (total of $50 million dollars) in 2006 to support projects that examine digital media and learning. Several proposals that have been funded by the MacArthur Foundation are related to computer game-based learning (MacArthur Foundation, 2006). With this new wave of using gaming technology as a teaching approach, it is my hope that significant issues such as gender-bias and cultural stereotyping in commercial computer game designs will be discussed and addressed.

REFERENCES

Amory, A., Naicker, K., Vincent, J., & Adams, L. (1999). The use of computer games as an educational tool: Identification of appropriate game types and game elements. *British Journal of Educational Technology, 30(4), 311-321.*

David, G. (1997). Integrated development and production tools for building hypermedia coursework and interactive scenarios. In *Proceedings of the ED-MEDIA 97 Conference.*

DeKanter, N. (2005). Gaming redefines interactivity for learning. *TechTrends, 49*(3), 26-31.

Dickey, M. (2006). Girl gamers: The controversy of girl games and the relevance of female-oriented game design for instructional design. *British Journal of Educational Technology, 37*(5), 785-793.

Druckman, D. (1995). The educational effectiveness of interactive games. *Simulation and gaming across disciplines and cultures: ISAGA at a watershed.* Thousand Oaks CA: Sage.

Gee, J. (2004). *What video games have to teach us about learning and literacy.* New York: Palgrave Macmillan.

Gee, J. (2005). What would a state of the art instructional video game look like? *Innovate 1*(6). Retrieved December 7, 2007, from http://www.innovateonline.info/index.php?view=article&id=80

Lawson, G. (2003). Ecological landscape planning: A gaming approach in education. *Landscape Research, 28*(2), 217-223.

MacArthur Foundation. (2006, December). *Building the field of digital media and learning*. Retrieved December 7, 2007, from http://www.digitallearning.macfound.org/site/c.enJLKQNlFiG/b.2029199/k.BFC9/Home.htm

Malone, T. W. (1981). What makes computer games fun? *Byte, 6*(12), 258-277.

Mann, B., Eidelson, B., Fukuchi, S., Nissman, S., Robertson, S., & Jardines, L. (2002). The development of an interactive game-based tool for learning surgical management algorithms via computer. *The American Journal of Surgery, 183*, 305-308.

Okan, Z. (2003). Edutainment: Is learning at risk? *British Journal of Educational Technology, 34*(3), 255-264.

Pivec, M., Dziabenko, O., & Schinnerl, I. (2003). Aspects of game-based learning. In *Proceedings of I-KNOW 03, the Third International Conference on Knowledge Management*. Retrieved December 7, 2007, from http://www.fhjoanneum.at/zml/publikationen.asp?jahr=2003&typ=P&lan=DE

Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research & Development, 44*(2), 43-58.

Robertson, J. (2002). Develops communication skills and literacy skill with game play. In *Proceedings of Game On—the Conference Exploring the Potential of Computer Games in Learning*, (pp. 20-21).

Salen, K., & Zimmerman, E. (2004). *Rules of play: Game design fundamentals*. Cambridge, MA: The MIT Press.

Sefton-Green, J. (2005). Changing the rules? Computer games, theory, learning, and play. *Discourse: Studies in the Cultural Politics of Education, 26*(3), 411-419.

Shaffer, D., Squire, K., Halverson, R., & Gee, J. (2005). Video games and the future of learning. *Phi Delta Kappan, 87*(2), 104-111.

Smith, L., & Mann, S. (2002). Playing the game: A model for gameness in interactive game based learning. In *Proceedings of the 15th Annual Conference of the NACCQ,* New Zealand.

Squire, K. (2002). Games to teach. In *Proceedings of Game On—the Conference Exploring the Potential of Computer Games in Learning*, (pp. 25-27).

Squire, K. (2005). Changing the game: What happens when video games enter the classroom*? Innovate, 1*(6). Retrieved December 7, 2007, from http://www.innovateonline.info/index.php?view=article&id=82

Stoney, S., & Oliver, R. (1998). Interactive multimedia for adult learners: Can learning be fun? *Journal of Interactive Learning Research, 9*(1), 55-81.

KEY TERMS

Animation: An illusion of movement that is created through a rapid display of a sequence of drawings which shows a continuous action.

Constructive Learning: Constructivism describes learners as active participants in knowledge acquisition. Learners engage in knowledge restructuring, manipulation, re-invention and experimentation. Understanding is constructed by the learner. Therefore, knowledge becomes meaningful and permanent.

Digital/Computer Natives: Refers to individuals who grew up with direct access to digital media such as computers, cell phones, and electronic games. Electronic media is part of this group's culture, and they are native speakers of the digital language.

Digital Immigrants: Refers to individuals who have needed to adapt to the digital environment after they were born. Examples include individuals who have needed to learn about the Internet or e-mail later on in their lives.

Edutainment: Edutainment is a type of entertainment which provides information that is both educational and entertaining at the same time.

Game-Based Learning: Game-based learning includes elements of competition, engagement, and immediate reward. Players should receive immediate feedback—for example, scoring—when a goal is accomplished. A game-based learning environment allows students to compete with one another or work collaboratively; it provides a level of challenge that motivates students' learning; and it provides a storyline that will help students engage in activities.

Usability Testing: Usability testing is a way of measuring and observing the degree to which people are able to use a product for its intended purpose. Usability testing is done in a controlled environment, and the objective is to discover errors in the product and identify areas that need improvement.

Digital Identity in Current Networks

Kumbesan Sandrasegaran
University of Technology, Sydney, Australia

Xiaoan Huang
University of Technology, Sydney, Australia

INTRODUCTION

The daily activities of humans and business are increasingly depending on the usage of (digital) identity for interaction with other parties and for accessing resources. Current networks use a number of digital-identity management schemes. For example, in the public switched telephone network (PSTN), a telephone number is simply used as the digital identity of a user. Most of the digital-identity management schemes are effective only within their networks and have limited support for interoperability. In the hybrid network environment of next-generation networks (NGNs), new digital-identity management models are expected to be proposed for digital-identity management.

The rest of the chapter is focused on the introduction of digital identity, the digital-identity schemes used in current telecommunication networks, and the future trends.

BACKGROUND

What is Digital Identity?

We define the identity of an individual as the set of information known about that person. With the development and widespread use of digital technologies, humans have been able to communicate with each other without being physically present. Digital identity is the means that an entity (another human or machine) can use to identify a user in a digital world. The aim of digital identity is to create the same level of confidence and trust that a face-to-face transaction would generate. Some selected definitions for digital identity are as follows.

Digital ID World ("What is Digital Identity?" 2003):

"A Digital Identity is the representation of a human identity that is used in a distributed network interaction with other machines or people. The purpose of the Digital Identity is to restore the ease and security human transactions once had, when we all knew each other and did business face-to-face, to a machine environment where we are often meeting each other for the first time as we enter into transactions over vast distances."

Field Elliot (2002)

"A Digital Identity is an assurance by one end of a digital conversation (such as a Web Services transaction) that the other end of the conversation is being conducted on behalf of a specific human, company, or other entity."

Composition of Digital Identity

Digital identity is comprised of two basic elements: the actual identity of the entity (something that can be observed by human senses), and the credentials or what are used to prove the identities. Credentials can take the following forms (Reed, 2002).

- **Something that the entity knows:** An example is a password or any piece of knowledge that the entity knows.
- **Something the entity has or possesses:** An example would be a magnetic swipe card used for entry into a room, elevator, or so forth.
- **Something the entity is:** Examples of parts of an entity include fingerprints and eye scans. These attributes are the most difficult to copy or impersonate.

Profile

A profile consists of data needed to provide services to a user once his or her identity has been verified. A user profile could include what an entity can do, what he or she has subscribed to, and so on. Profiles are important to digital identity as they represent records and other data about users that can be stored external to the actual entity itself.

Usage of Digital Identity

Authentication

One of the important uses of digital identity is authentication, where an entity must prove digitally that it is the entity that it claims to be ("What is Digital Identity?" 2003). It is at this stage that the credentials of digital identity are used.

The simplest form of authentication is the use of a user name and a corresponding password.

Authorisation

Once an entity is authenticated, a digital identity is used to determine what that entity can do. This is where the profile of a digital identity is required. For example, while both an administrator and a user are authenticated to use a computer, the actions that each may do with that resource are determined by the authorisation.

Accounting

Accounting involves the recording and logging of entities and their activities within the context of a particular organisation, Web site, and so forth. Effective accounting processes enable an organisation to track unauthorised access when it does occur.

DIGITAL IDENTITY IN CURRENT NETWORKS

Digital Identity in Mobile Networks

One of the evolution paths of mobile networks is from the global system for mobile communications (GSM) to general packet radio service (GPRS), and to the universal mobile telecommunication system (UMTS). Figure 1 shows the architectures of the GSM, GPRS, and UMTS networks. In GSM/GPRS, users connect to the mobile core networks via the base-station subsystem. The mobile switching centre (MSC) and visitor location register (VLR) in the core network are the main entities used for the CS domain. The serving GPRS support node (SGSN) and gateway GPRS support node (GGSN) are the main entities used for the PS domain. In UMTS, the new radio network controller (RNC) takes charge of connecting users with the SGSN or MSC and VLR. A number of RNCs form the UMTS terrestrial radio access network (UTRAN).

Digital-Identity Composition in Mobile Networks

There are three broad aspects of digital-identity composition in mobile networks.

Identity and Communication Management

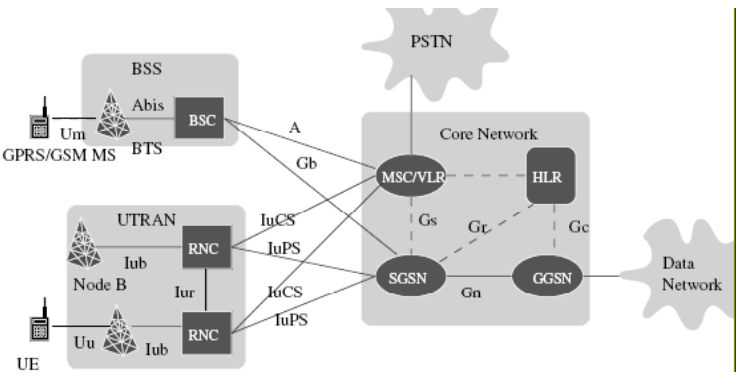
In identity and communication management, each subscriber has to be uniquely identified. The unique addressing codes are described below (Kaaranen, 2001).

- **International Mobile Subscriber Identity (IMSI)**
  It is a unique and confidential identity for the mobile subscriber (MS) and is the same in GSM, GPRS, and UMTS. The structure of IMSI is shown below (Pandya, 1997):

  IMSI = MCC + MNC + MSIN,

  where
  MCC = mobile country code (3 digits),
  MNC = mobile network code (2 digits),
  and
  MSIN = mobile subscriber identity number (normally 10 digits).
- **Mobile Subscriber International ISDN (Integrated Services Digital Network) Number (MSISDN)**
  The MSISDN is used for service separation. A subscriber may have several services provisioned and

Figure 1. GSM, GPRS, and UMTS network architectures (Lin & Chlamtac, 2001)



activated with only one IMSI. The MSISDN consists of three parts:

MSISDN = CC + NDC + SN,

where
CC = country code (one to three digits),
NDC = national destination code (one to three digits), and
SN = subscriber number.

- **Temporary Mobile Subscriber Identity (TMSI)**
  Due to security reasons, it is very important that the IMSI is always transferred in ciphered mode during signaling across the air interface so as to protect the digital identity of the user. For this purpose, the GSM system uses a 4-byte TMSI number.
- **International Mobile Equipment Identity (IMEI)**
  IMEI is a number uniquely identifying the user's mobile equipment hardware. The network may ask the user equipment (UE) to identify itself with an IMEI number either in the context of every transaction or occasionally as defined by the network operator (Kaaranen, 2001).
- **Mobile Subscriber Roaming Number (MSRN)**
  The MSRN is a number used for call routing purposes in the mobile terminated call.
- **Packet Temporary Mobile Subscriber Identity (P-TMSI)**
  The P-TMSI is a temporary identity used in the PS domain for the same purpose as TMSI in the CS domain. The P-TMSI is a set of random-format numbers consisting of 32 bits and has limited validity in terms of time and area. The P-TMSI is allocated by the SGSN.
- **P-TMSI Signature**
  The P-TMSI signature, which may be sent to the MS by the SGSN, consists of three octets. It can be used to prove that the corresponding P-TMSI returned by the MS is the one allocated by the SGSN.
- **Temporary Logic Link Identity (TLLI)**
  The TLLI is derived from the P-TMSI received in the MS. It is used to identify a mobile user on the radio path.
- **Access Point Name (APN)**
  The APN is used to identify the GGSN in the PS domain. It will be translated into an IP (Internet protocol) address of the GGSN for roaming purposes.

Mobility Management

One of the important functions of a mobile network is to keep track of the approximate location of a mobile user for the purpose of routing incoming calls. The following identities are used for mobility management.

- **Cell Global Identity (CGI)**
  To globally separate cells from each other, the CGI is used:

  CGI = MCC +MNC + LAC + CI (cell identity).

  The CGI value consists of the country of the network (MCC), the network within the country (MNC), the location area in the network, and finally the cell number within the network.
- **Location Area Identity (LAI)**
  The location area is used in the circuit-switched domain. It consists of a group of cells. The minimum is one cell and the maximum is all the cells under one VLR. An LAI is used to uniquely identify an LA.
- **Routing Area Identity (RAI)**
  An RA is very similar to the LA. An RA is the area where the MS may move without performing a routing area update. The reason why these two definitions coexist is to cater to the possibility that an MS can have either circuit or packet connection.
- **UTRAN Registration Area Identity (URAI)**
  For mobility management in 3G (third generation) UMTS, the URA is introduced. As a subscriber moves into the geographical range of an RNC serving area, the subscriber is allocated into the serving URA and the information in the network databases has to be updated (Third-Generation Partnership Project [3GPP], 2005).

Security Identity

In GSM, the identification key (Ki) is used for authentication purposes. The Ki is allocated at the subscription time together with the IMSI. It is confidential and never transmitted over the network (Bates, 2001).

In UMTS, the authentication mechanism uses a master key K. This is a permanent, secret private key with a length of 128 bits that is never transferred outside of the USIM and AuC (Kaaranen, 2001; 3GPP, 2004).

Digital Identity Storage in Mobile Networks

- **Subscriber Identity Module (SIM)**
  The SIM card represents the main hardware used for identification purposes in GSM. It contains authentication algorithms and the application protocol of GSM, and also the subscription and user-specific data (Vedder, 2001). In UMTS, the SIM card has been developed to use the universal subscriber identity module (USIM).

- **Home Subscriber Sever (HSS)**
  The HSS is the main entity in which permanent subscriber information is stored. The HSS contains the following information about the subscriber for the purposes of authentication, authorisation, location information, naming and addressing resolution, and so forth (3GPP, 2005).
  - Identification, numbering, and addressing information of the user
  - User security information for authentication and authorisation
  - User profile information
  - User location information at the intersystem level
- **Home Location Register (HLR)**
  The HLR can be seen as a subset of the HSS. It contains permanent data of the subscribers. One subscriber can be in only one HLR.
- **Authentication Centre (AuC)**
  The AuC contains parameters (e.g., secret keys, the cipher key, etc.) that the VLR uses for security activities. The AuC has only one associated HLR and communicates only with it (Bates, 2001).
- **VLR**
  The VLR database contains temporary copies of the active subscribers who have performed a location update in its coverage area. The identity information includes the user's IMSI, TMSI, MSISDN, MSRN, old LAI, and so on. The VLR also contains supplementary service parameters from the HLR specific to a mobile subscriber.
- **Equipment Identity Register (EIR)**
  The EIR maintains the security information about the mobile equipment hardware, which is the IMEI.
- **SGSN**
  The SGSN stores subscriber data required for a number of functions such as mobility management, packet data transfer, and so forth. The information includes the IMSI, subscription information, P-TMSI, location information, and so on.
- **GGSN**
  The GGSN stores subscriber data received from the HLR and the SGSN. The data are needed to handle originating and terminating packet data transfer.

Digital-Identity Usage in Mobile Networks

Security

The usage of digital identity for security is mainly in the authentication procedures. The algorithm A3 and the identity of Ki that corresponds to the IMSI are used for authentication in GSM (3GPP, 2000). In UMTS, the possibility of an attacker setting up a phony network has been removed by means of a mutual or two-way authentication procedure.

Mobility Management

When the old and new location areas are different, the MS initiates a location area update procedure to inform the network in the CS domain; this is typical in GSM. The user's TMSI and the LAI are used in the procedure (3GPP, 2000). In the PS domain, when the old and new routing areas are different, the MS initiates a routing area update procedure. The routing area update is typical in GPRS. The P-TMSI of the user and the RAI are used for this purpose. There are both a location area update and a routing area update in UMTS (3GPP, 2002).

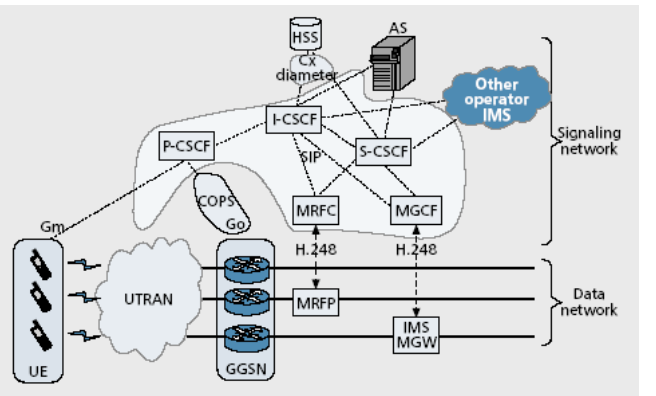
Internet Protocol Multimedia Subsystem (IMS)

IMS is the emerging subsystem of UMTS for providing users with IP multimedia services. The architecture of IMS is shown in Figure 2.

The IMS architecture is organised in two networks: the control network and the transport network. The main element in the control network is the call session control function (CSCF). It is responsible for session control, and AAA and charging support. There are three types of CSCF in IMS: a serving CSCF (S-CSCF), a proxy CSCF (P-CSCF), and an interrogating CSCF (I-CSCF).

The S-CSCF is located in the home network of the MS. The session control of multimedia services is handled by the S-CSCF. In addition, it is responsible for authenticating

Figure 2. Architecture of the IMS (Marquez, Rodriguez, Valladares, de Miguel, & Galindo, 2005)



the users by communicating with the HSS/AuC. A P-CSCF acts on behalf of a roaming mobile terminal in the visited IMS. It is responsible for redirecting the messages of a UE to the home network. An I-CSCF is a firewall for the messages coming into a home network. It is also responsible for selecting an S-CSCF in the home network for the UE (Marquez et al., 2005).

Composition of Digital Identity in IMS

- **IP Multimedia Private Identity (IMPI)**
  The IMPI is a unique global identity for the IMS user. It has similar functions as the IMSI in GSM. During registration, the initial request of the user should contain the user's IMPI. Each IP multimedia service identity module (ISIM) stores one IMPI. The S-CSCF and HSS store the user's IMPI during registration (3GPP, 2005).
- **IP Multimedia Public Identity (IMPU)**
  The IMS user will also be allocated one or more IMPUs by the home network operator. Other parties can know the subscriber by its IMPUs. One ISIM has one or more IMPUs (3GPP, 2005).

Usage of Digital Identity in IMS

P-CSCF Discovery

The UE needs to discover the P-CSCF no matter if the user is in the home network or roaming in a visited network. After receiving both the domain name and IP address of a P-CSCF, the UE may initiate communication with the IMS.

Registration Procedure

After discovering the corresponding P-CSCF, the user can initiate the registration procedure of the IMS. The user's IMPI is used for registration purposes (Yi-Bing, Ming-Feng, Meng-Ta, & Lin-Yi, 2005).

Digital Identity in WLAN

In IEEE (Institute of Electrical and Electronics Engineers) 802.11, the wireless local area network (WLAN) systems primarily focus on the physical and link layers and have limited mobility when compared to mobile networks. Therefore, the digital identities in WLAN are not as complex as in mobile networks. No specific mechanisms have been defined for the IP layer and above layers. The format of digital identities in these layers varies and depends on the specific protocol being used.

The WLAN architecture shown below has several stations (STAs) that are connected to access points (APs). The APs are interconnected to form a local network or are connected to external fixed networks. The stations and APs that are within the same radio coverage form a basic service set (BSS), which is the basic building block of a WLAN. A collection of BSSs can form a single network called an extended service set (ESS).

Digital-Identity Format in WLAN

In WLAN, the media access control (MAC) address is normally adopted to identify some entities of WLAN, for example, the AP or BSS. The receiver's or transmitter's MAC address can also be used as the digital identity.

Figure 3. The IEEE 802.11 architecture (Yan, 1998)

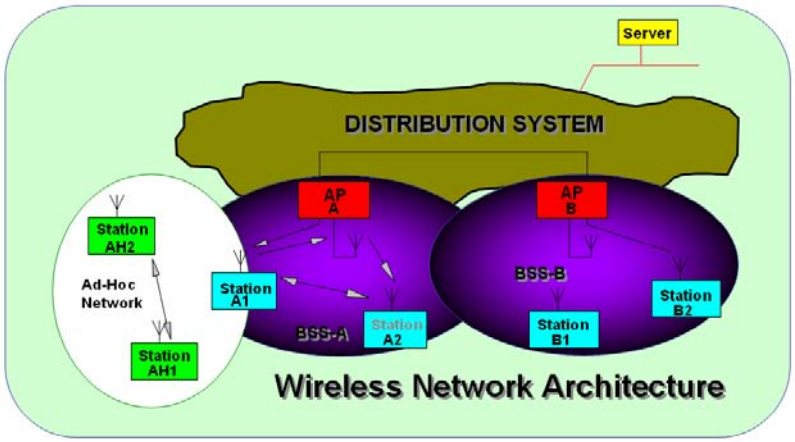
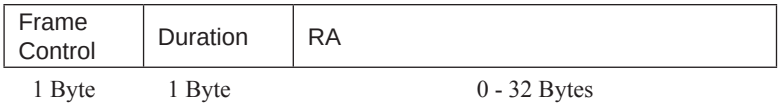


Figure 4. Format of the SSID information element



SSID

The service set identifier (SSID) can be up to 32 bytes in length, and if the length is 0, the SSID is a broadcast SSID.

In IEEE 802.11, the SSID is used to identify all entities of a WLAN or ESS.

Usage of Digital Identity in WLAN

Authentication

In IEEE 802.11 specifications, there are two basic mechanisms for the authentication of WLAN clients: open authentication and shared-key authentication. The WEP key is used as the client's identity.

Open authentication relies on the preconfigured WEP key on the client and the AP. If the keys mismatch, the clients are not allowed to communicate.

Shared-key authentication is more robust by using a challenge. If the AP can retrieve the original challenge that was previously sent by decrypting the frame from the client, the identity of the client is validated (Roshan & Leary, 2004).

In addition, MAC-address authentication is also supported by many vendors.

Encryption

Data encryption mechanisms are based on cipher algorithms (Roshan & Leary, 2004). IEEE 802.11 provides the encryption of data by using the WEP algorithm.

Mobility Management

Generally, the mobility management solutions of WLAN can be classified into the following two categories.

- Layer 2 solutions (link-layer solutions)
- Layer 3 solutions (network-layer solutions)

In Layer 2, scanning is employed by wireless clients to find a suitable AP by identifying the SSID. Besides the AP discovery mechanisms, the roaming process of Layer 2 mobility management mainly relies on the registering and updating of the client's MAC address in the APs (Roshan & Leary, 2004).

In IEEE 802.11 WLAN, the client must perform Layer 2 roaming before it can use Layer 3 roaming. The roaming between different domains can be enabled by network-layer solutions, such as mobile IP (MIP). MIP is a protocol for accommodating node mobility within the Internet. It enables nodes to change their points of attachment to the Internet without changing their IP addresses (Perkins, 2002).

Digital Identity in PSTN

Identities in PSTN are much simpler compared to mobile networks. The wireline network in PSTN and dedicated loop to a user makes it naturally more secure than mobile networks, which use the air interface. Furthermore, the fixed nature of PSTN users results in far less mobility management functionality than in mobile networks.

PSTN has several characteristics relevant to digital identity.

1. Simplified identity: The amount of user information that needs to be sent and stored for a PSTN is minimal.
2. Security concerns: The simplicity of PSTN, as well as its dependence on wiring, makes it less susceptible to security breaches.
3. Simple technology: It is difficult to introduce new services or increase the variety of data types that PSTN can support. Therefore, simple identity is sufficient to be used in PSTN.

Digital-Identity Format in PSTN

The phone numbers used in PSTN can be considered as simple digital identities. They are referenced according to area (national, state or province, household).

ITU-T International Numbering Plan

The numbering format was defined in the ITU-T Recommendation E.164 and has the following three parts.

- Country code
- National destination code
- Subscriber number

The CC is also referred to as the international access code. It is comprised of one to three digits. The world number zones are defined in the first digit.

The length of the NDC and the SN can vary according to the needs of the country, but should not exceed 15 digits.

Digital identity in mobile communications (e.g., GSM, GPRS, UMTS, and IMS) has been designed carefully to provide users with anonymity, security, and mobility. Consequently, it has become more complex than digital identity in other networks. Due to the advanced nature of digital identity in mobile networks, it is expected that most of this digital identity's features will be retained in NGN.

FUTURE TRENDS

Trends in mobility, security, and convergence are driving the current telecommunications networks toward a single all-IP-based network, which is referred to as NGN. All of the traditional telecommunication networks and the emerging networks will be integrated and become a single NGN in the near future. Users should have ubiquitous access to multiple services anywhere and anytime (ITU-T, 2004). In the hybrid network environment of NGN, digital-identity information needs to be exchanged across different network technologies. However, most of the digital-identity management schemes used in current networks are effective only within their networks and have limited support for interoperability. For example, the identities used in one network cannot be used in others or the interconnection between them causes a lot of errors or latency. Therefore, new digital-identity management models have to be proposed for digital-identity management in NGN.

CONCLUSION

In this article, we have defined the concept of digital identity and described the basic usage of it in authentication, authorisation, accounting, and so forth. Thereafter, the digital identities that are used in current networks including wireless cellular networks (such as GSM, GPRS, UMTS, and IMS), WLAN, and PSTN have been introduced. The composition, storage, and usage of the digital identities in these networks are analyzed and compared. Some of the advanced features of digital identity in current networks will be retained in NGN. However, new digital-identity management models are required to meet the requirements of digital identity in NGN.

REFERENCES

Bates, R. J. (2001). *GPRS: General packet radio service.* New York: McGraw-Hill.

Field-Elliot, B. (2002). *Identity and the digital estate.* Retrieved from http://www.digitalidworld.com/modules.php?op=modload&name=News&file=article&sid=54

ITU-T. (2004). *NGN-related recommendations: ITU-T, Study Group 13 NGN-WD-87.*

Kaaranen, H. (2001). *UMTS networks: Architecture, mobility, and services.* Chichester, United Kingdom: John Wiley & Sons.

Lin, J. Y.-B., & Chlamtac, I. (2001). *Wireless and mobile network architectures.* New York: Wiley.

Marquez, F. G., Rodriguez, M. G., Valladares, T. R., de Miguel, T., & Galindo, L. A. (2005). Interworking of IP multimedia core networks between 3GPP and WLAN. *IEEE Wireless Communications, 12*(3), 58-65.

Pandya, R. (1997). Numbers and identities for emerging wireless/PCS networks. *IEEE Personal Communications, 4*(3), 8-14.

Perkins, C. (2002). *IP mobility support for IPv4.* Retrieved from http://www.faqs.org/rfcs/rfc3220.html

Reed, A. (2002). *The definitive guide to identity management.* Retrieved from http://realtimepublishers.com

Roshan, P., & Leary, J. (2004). *802.11 wireless LAN fundamentals* (1st ed.). Cisco Press.

Third-Generation Partnership Project (3GPP). (2000). Security related network functions (Release 1999). In *Digital cellular telecommunications system (Phase 2+).* Author.

Third-Generation Partnership Project (3GPP). (2002). Architectural requirements for Release 1999 (Release 1999). In *Technical specification group services and systems aspects.* Author.

Third-Generation Partnership Project (3GPP). (2004). 3G security: Security architecture (Release 6). In *Technical specification group services and systems aspects*. Author.

Third-Generation Partnership Project (3GPP). (2005a). IP multimedia subsystem (IMS). In *Technical specification group services and system aspects.* Author.

Third-Generation Partnership Project (3GPP). (2005b). Network architecture (Release 6). In *Technical specification group services and systems aspects.* Author.

Third-Generation Partnership Project (3GPP). (2005c). *UTRAN overall description (Release 6).* In *Technical specification group services and systems aspects.* Author.

Vedder, K. (2001). The subscriber identity module: Past, present and future. (chap. 13). John Wiley & Sons Ltd.

What is digital identity? (2003). *Digital Identity World.* Retrieved from http://www.digitalidworld.com/local.php?op=view&file=aboutdid_detail

Yan, R. (1998). *Wireless data.*

Yi-Bing, L., Ming-Feng, C., Meng-Ta, H., & Lin-Yi, W. (2005). One-pass GPRS and IMS authentication procedure for UMTS. *IEEE Journal on Selected Areas in Communications, 23*(6), 1233-1239.

KEY TERMS

Accounting: It involves the recording and logging of entities and their activities within the context of a particular organisation, Web site, and so forth. Effective accounting processes enable an organisation to track unauthorised access when it does occur.

Authentication: It is the process where an entity must prove digitally that it is the entity that it claims to be.

Authorisation: It is the process that is used to determine what an entity can do once the entity is authenticated.

Credentials: They are objects that people use to prove their identities in an authentication process.

Digital Identity: It is the means that an entity (another human or machine) can use to identify a user in a digital world. The aim of digital identity is to create the same level of confidence and trust that a face-to-face transaction would generate.

NGN: A single all-IP-based network that integrates the current telecommunications networks. Users of NGN should have ubiquitous access to multiple services anywhere and anytime.

Profile: It consists of data needed to provide services to users once their identities have been verified. A user profile could include what an entity can do, what it has subscribed to, and so on.

Digital Knowledge Management Artifacts and the Growing Digital Divide: A New Research Agenda

Ioannis Tarnanas
Kozani University of Applied Science, Greece

Vassilios Kikis
Kozani University of Applied Science, Greece

INTRODUCTION

That portion of the Internet known as the World Wide Web has been riding an exponential growth curve since 1994 (Network Wizards, 1999; Rutkowski, 1998), coinciding with the introduction of NCSA's graphically based software interface Mosaic for "browsing" the World Wide Web (Hoffman, Novak, & Chatterjee 1995). Currently, over 43 million hosts are connected to the Internet worldwide (Network Wizards, 1999). In terms of individual users, somewhere between 40 to 80 million adults (eStats, 1999) in the United States alone have access to around 800 million unique pages of content (Lawrence & Giles, 1999), globally distributed on arguably one of the most important communication innovations in history.

Yet even as the Internet races ambitiously toward critical mass, some social scientists have begun to examine carefully the policy implications of *current* demographic patterns of Internet access and usage (Hoffman & Novak, 1998; Hoffman, Kalsbeek, & Novak, 1996; Hoffman, Novak, & Venkatesh, 1997; Katz & Aspden, 1997; Wilhelm, 1998). Looming large is the concern that the Internet may not scale *economically* (Keller, 1996), leading to what Lloyd Morrisett, the former president of the Markle Foundation, has called a "digital divide" between the information "haves" and "have-nots." For example, although almost 70% of the schools in this country have at least one computer connected to the Internet, less than 15% of classrooms have Internet access (Harmon, 1997). Not surprisingly, access is not distributed randomly, but correlated strongly with income and education (Coley, Cradler, & Engel 1997). A recent study of Internet use among college freshman (Sax, Astin, Korn, & Mahoney 1998) found that nearly 83% of all new college students report using the Internet for school work, and almost two-thirds use e-mail to communicate. Yet, closer examination suggests a disturbing disparity in access. While 90.2% of private college freshman use the Internet for research, only 77.6% of students entering public black colleges report doing so. Similarly, although 80.1% of private college freshman use e-mail regularly, only 41.4% of students attending black public colleges do.

Further, although numerous studies (e.g., CyberAtlas, 1999; Maraganore & Morrisette, 1998) suggest that the gender gap in Internet use appears to be closing over time and that Internet users are increasingly coming from the ranks of those with lower education and income (Pew Research Center, 1998), the perception persists that the gap for race is not decreasing (Abrams, 1997).

We now raise a series of points for further discussion. We believe these issues represent the most pressing unanswered questions concerning access and the impact of the digital divide on the emerging digital economy. This article is intended to stimulate discussion among scholars and policymakers interested in how differences in Internet access and use among different segments in our society affect their ability to participate and reap the rewards of that participation in the emerging digital economy. In summary, we have reviewed the most recent research investigating the relationship of race to Internet access and usage over time. Our objective is twofold: (1) to stimulate an informed discussion among scholars and policymakers interested in the issue of diversity on the Internet, and 2) to propose a research agenda that can address the many questions raised by this and related research.

BACKGROUND

Laugksch (1999) pointed out that scientific literacy has become an internationally well-recognized educational slogan, buzzword, catchphrase, and contemporary educational goal. The same applies to the case of the digital divide. Courtright and Robbin (2001) contend that "the metaphor of the digital divide" has become part of the national discourse of the United States, an abstract symbol that condenses public concerns about social inequality and evokes hopes for solutions related to the use of information technology. In addition, "the digital divide is a potent resource whose symbolic properties and communicative power have activated a wide array of participants in the policy debates about how to create a more just society."

According to Hoffman (2001; cf. Arquette, 2001), the term *digital divide* was first used by Lloyd Morrisett who vaguely conceived of a divide between the information-haves and have-nots. However, the divide herein mainly is a gap of PC penetration in the early days of the Apple II in 1980 (Arquette, 2001). The term then grasped the public's attention with the issuance of the first National Telecommunications and Information Administration (NTIA) survey on Internet adoption and use in the United States in 1994 with the catchy title, *Falling Through the Net.* Since then, numerous articles, either popular or academic, on this issue have been published. According to a convenient sample of newspapers, journal articles, newswires, and similar mass media sources in the Lexis-Nexis database from January 1999 to December 2000 (Arquette, 2001), the increasing rate of digital divide-related articles hits almost 3,000%.

In developing countries, the digital divide is receiving similar social saliency. A quick search with the keywords "digital divide" in one of Greece's leading news Web sites *Daily Online* (*www.in.gr*), shows that at least 500 articles somehow related to this term are available. In July 2001, a high-level forum on public understanding of information technology with the special topic of "Pay Attention to the Digital Divide" was held in Greece. A wide range of representatives, including governmental officials, information technology (IT) experts, educators, social scientists, and media practitioners, presented their viewpoints and comments on this issue. The digital divide has been incorporated into daily conversational discourse.

Ironically, while the term *digital divide* has frequently appeared in varied contexts, including academic writings, both the connotative and denotative meanings of it are confusingly incoherent. The presence of other similarly prevalent terminologies, such as digital equality, information equality, e-development, network readiness, and so forth, add confusion. People seem to debate on the issue without a shared understanding of what is meant by the digital divide. As Arquette (2001) contends, the entire researcher community is plagued by a lack of definitional clarity of the concepts such as digital divide: "Each researcher assumes other researchers use the same definitional frameworks for these terms while in fact there is no such shared meaning in nomenclature" (p. 3).

While the comment of Arquette (2001) mainly refers to the phenomenon in the English-speaking world, the use of its minority counterpart of the term *digital divide* is also in a similar situation. For example, among more than 30 articles collected by the book *Pay Attention to the Digital Divide in Developing Countries* (Leng, 2002), no consistent conceptual definition is available across the writings. While some are talking about the Internet penetration divide among different social groups categorized by age, occupation, and educational level, others refer to the concept as an uneven development of e-infrastructure among different areas or nations. So, whenever the term digital divide is confronted, the following question can always be raised: *In terms of what*?

This article intends to introduce a new approach of operationalizing the digital divide from the perspective of developing countries. We first briefly review different definitional perspectives of the term *digital divide.* Then a detailed introduction of the National Informatization Quotient (NIQ) is presented which will be employed as the operational definition of the informatization level of a region. Finally we will investigate the geographical digital divide in developing countries in terms of NIQ.

CONCEPTUAL REVIEW

Conceptual definition involves verbal descriptions of the essential properties that are to be included in the intended meaning of a concept. In research practice, it often involves specifying the essential dimensions of a concept (McLeod & Pan, 2002, p. 62). On the other hand, o*perational definition* involves procedures by which a concept is to be observed, measured, or manipulated. It details the rules, specific steps, equipment, instruments, and scales involved in measuring a concept (p. 65). In this section, we will briefly review the multiple conceptions around digital divide.

Digital divide is a fresh term not unfamiliar to communication scholars (Zhu, 2002). As early as 1970, a theory called *knowledge gap* (Tichenor, Donohue, & Olien, 1970) was developed which was one of the most active inquiry fields thereafter in communication studies. The supposition of knowledge gap mainly concerns the different knowledge possession through mass media by social groups with varied social-economic status. In the 1980s, with the development of information and communication technologies (ICTs), especially with the wide application of PCs in diverse contexts, a divide between the information-haves and have-nots was sensitively observed and warned (Compaine, 2001). Since the early 1990s, digital divide has gradually become a convenient label, or more precisely, a metaphor (Courtright & Robbin, 2001), in describing the inequality of possessing and using ICTs, especially the Internet connectedness.

The first group of definitions varies on the concrete referents of what 'digital' means. In a narrow sense of the definition, digital divide particularly referred to the inequality of Internet access and use among different social groups or localities. The U.S. Department of Commerce's (1995, 2001) *Falling Through the Net* reports represent the most influential version of the stream. Zhu (2002) also takes Internet penetration as the sole indicator of what 'digital' means in his construction of the digital divide index (DDI), while taking age, sex, education, and occupation collectively as the categorizing factors. In short, in this stream of definitions, digital divide is operationalized to Internet

access/penetration divide categorized by demographics and social status factors.

However, to many people, the term *digital* means a wide range of ICTs other than the Internet. Arquette (2001) labeled it as the concept *fit disjuncture* in the studies of digital divide — that is, to measure global digital equality in terms of teledensity or Internet penetration. Employing the so-called Information Intelligence Quotient (IIQ) analytical framework, he uses ICT infrastructure rather than a single ICT, such as the Internet or telephony, as the subject of the "digital."

A second clue of conceptualizing the digital divide basically focuses on the meaning of "divide." Many different analytical perspectives on this concern are available. Jackel (2001) exemplifies some of these:

- A macro-level-comparison of the so-called First and Third world or a comparison of rich and poor countries;
- a comparison of differences in modern societies according to the level of integration in the labor market;
- a comparison of differences in modern societies according to education groups, gender, and age — that is, more general a comparison of generations;
- a specification of differences in modern societies according to communication skills;
- a comparison of different diffusion curves as a result of differing demands.

As can be seen, the dimensions identified by these perspectives are noticeably diverse.

Synthesizing the prior research on digital divide, Arquette (2001) proposed an organizational framework based on three dimensions of digital divide: ICS infrastructure, access, and use. ICS infrastructure refers to the technophysical means by which voice, audio, video, and/or data communication circulates. The operationalization of the dimension involves the specification of 16 indicators, including telephony penetration (wire line and wireless), Internet hosts, and costs of calls. The second dimension is ICS access, which focuses on the ability of persons interested in using the infrastructure (regardless of that infrastructure quality or quantity) to gain access to the infrastructure. Nineteen indicators are developed to operationalize the dimension.

The third dimension of digital divide that Arquette (2001) specifies is ICS use. Use-based conceptualizations of digital divide are conceived in terms of how people employ the technologies. Another 19 indicators are developed to measure the situation of this dimension of digital divide. In summary, IIQ is an aggregate meta-analytic framework for assessing the state of digital divide among different nations or regions.

A notable point implied by the IIQ is its consideration of the dimension of ICT use. In fact, 'access is not enough' is becoming a recognizable consensus (e.g., Blau, 2002; Jackel, 2001; Nagaraj, 2002). In other words, merely connecting people and computers will not bridge the digital divide (Blau, 2002), and there's digital discrimination among the information haves, too (Nagaraj, 2002). Jackel (2002) labels the divide among ICT haves as the second level of "divide."

NIQ is a composite index comprising 20 indicators in six dimensions. It is the operational definition of the National Informatization Level (NIL). In the remaining part of the article, the digital divide is discussed in terms of this NIL, which is operationally defined as NIQ. The six dimensions of NIQ are:

1. **The development and application of information resources (IR):** The indicators under this umbrella term include *radio and TV broadcasting hour/per 1,000 people, bandwidth per person, telephone use frequency per person,* and *total capacity of Internet database.*
2. **Information network construction (IN):** There are four components in this dimension, including *total length of long distance cable, microwave channels, total number of satellite stations,* and *number of telephone lines per 100 people.*
3. **The application of information technologies (IT):** The indicators for this dimension include *number of cable TV stations per 1,000 people, number of Internet users per one million people, number of computers per 1,000 people, number of TV sets per 100 people, e-commerce trade volume,* and *proportion of investment in the information industry by enterprises to the total fixed investment.*
4. **Information industry development (II):** There are two indicators designed to reflect the situation of this dimension: *added value contributed by the information industry to the total GDP* and *contributions made by the information industry to the total GDP increase.*
5. **Human resources of informatization (HR):** There are two indicators for this dimension: *proportion of university graduates per 1,000 people* and *information index which refers to the proportion of expenditure other than fundamental consumption to the total expenditure.*
6. **The environment for informatization development (EI):** Two indicators are designed to measure the situation of the dimension: *proportion of expenses for research and development of the information industry to the country's total budget in R&D* and *proportion of investment on the infrastructural development of the information industry to the country's total investment in capital construction.*

Compared to other index used to reflect the digital divide or ICT development in a country or region, NIQ has several characteristics:

1. It is a multi-dimensional composite index. Therefore, NIQ is a comprehensive reflection of the state informatization level rather than the development of some particular ITs.
2. As for its application in assessing digital divide, the divide it evaluates is a geographical divide rather than informatization divide among different social groups or divides defined by other factors.
3. The index covers a wide range of the aspects regarding the informatization development. Particularly, NIQ emphasizes the importance of information industry in its structure of dimensions. The proportion of indicators related to information industry is notably high, which reflects the fact that NIQ will be a guideline for the promotion and adoption of IT in developing countries.

DEVELOPING A RESEARCH AGENDA

We now raise a series of points for further discussion. We believe these issues represent the most pressing unanswered questions concerning access and the impact of the digital divide on the emerging digital economy.

Computers in the Home

While previous research has shown that inequalities in Internet access in schools persist (Educational Testing Service, 1997; Sax et al., 1998), the research reviewed here suggests that inequalities in Internet access at home may be even more problematic. The role of access to the Internet at home needs to be much more clearly understood (Abrams, 1997). Caucasians are more likely to have access to the Internet and to have ever used the Web than minorities, and these gaps appear to be *increasing* over time. Probing deeply, we have discovered that among recent Web users, who by definition have access, the gaps in Web use have been *decreasing* over time. Over time, there appear to be no or only slight differences between caucasians and minorities in how recently they had used the Web, how frequently, or in their length of time online. Gaps in general Web access and use between different minorities and caucasians appear to be driven by whether or not there is a computer present in the home. Access to a personal computer, whether at home, work, school, or somewhere else, is important because it is currently the dominant mechanism by which individuals can access the Internet. We believe that access translates into usage. Overall, individuals who own a home computer are much more likely than others to use the Web. This suggests that programs that encourage home computer ownership (e.g., Roberts 1997) and the adoption of inexpensive devices that enable Internet access over the television should be aggressively pursued, especially for minorities.

Morrisette (1999) forecasted that by the year 2003, over half of all households in the United States would have access to the Internet, but that PC penetration could stall at 60% of households. Research is necessary to understand what motivates individual-level adoption of home computers and related technologies, as well as Internet adoption, both within and outside the home. Additionally, research is required to understand the long-term impact of home computer ownership on Internet access and use. Katz and Aspden (1997) investigated the role of social and work networks in introducing people to the Internet. The dominant three ways people were originally introduced to the Internet were (1) taught by friends or family, (2) learned at work, and (3) self-taught. Formal coursework was the *least* often mentioned way people were introduced to the Internet. Long-term Internet users were most likely to have learned at work; for recent Internet users, friends/family and self-taught were equally important. These results reinforce the importance of the presence of a computer at home, or the opportunity to access the Web from locations other than the home, in stimulating Web use.

Insight into the importance of reducing this gap in Web use between caucasians and African-Americans is provided by Anderson and Melchior's (1995) discussion of *information redlining.* Information redlining signifies the relegation of minorities into situations where satisfying their information needs is weighed against their economic and social worth. From the minority point of view, this is both an access issue and a form of discrimination. The new technologies of information are not simply tools of private communication as a telephone is or tools of entertainment as a television is; they provide direct access to information sources that are essential in making social choices and keeping track of developments not only in the world at large, but also within their immediate neighborhoods. Unless the neighborhoods are properly served, there is no way out of information redlining for most of these disadvantaged groups. Research on this topic is warranted.

There are also interesting differences in media use between caucasians and minorities that also deserve further probing. For example, although the rate of home PC ownership among minorities is flat or even decreasing, the rates of cable and satellite dish penetration are increasing dramatically for minorities. At a minimum, these results suggest that minorities may make better immediate prospects than caucasians for Internet access through cable modems and satellite technology.

Web Use Outside of the Home

In addition to gaps in home computer ownership, the implications of differential Internet access at locations outside the home, including school, the workplace, and other locations, needs to be clearly understood. Research suggests that additional access points stimulate usage. Further research is necessary to understand the impact of multiple access points on Web use, particularly for individuals who have no access at home. Public-private initiatives such as Bell Atlantic's efforts in Union City and Bill Gates's announcement of a $200 million gift to provide library access to the Internet are a step in the right direction (Abrams, 1997). It has also been noted that "community networks and public access terminals offer great potential for minority communities" (Sheppard, 1997). Further, the recent rollout of e-rate funds (Schools and Libraries Corporation, 1998) provides a significant opportunity for researchers to understand the factors important in stimulating Web usage among those least likely to have access.

School Web Use

The role of Web access in the schools, compared to other locations, needs to be clearly understood. Students enjoy the highest levels of Internet access and Web use, especially when there are computers in their households. However, caucasian students are still more likely than minority students to have access and to use the Internet, and these gaps persist over time. Indeed, our findings closely parallel statistics comparing student Internet use at private universities and minority public colleges (Sax et al., 1998). As a recent report by the Educational Testing Service (1997) makes clear:

- "There are major differences among schools in their access to different kinds of educational technology."
- "Students attending poor and high-minority schools have less access to most types of technology than students attending other schools."
- "It will cost about $15 billion, approximately $300 per student, to make all the our schools 'technology rich'."

This is five times what we currently spend on technology, but only 5% of total education spending. Anderson and Melchior (1995) cited lack of proper education as an important barrier to technology access and adoption. Access to technology does not make much sense unless people are properly educated in using the technologies. Our data do not speak to the quality of the hardware/network connections or the quality of information technology education that is provided by school.

Differences in Search Behavior

Reasons for the gap between different minorities and caucasians in Web search behavior need to be clearly understood. Such differences could have important implications for the ultimate success of commercial efforts online. Caucasian Web users are more likely to report searching for product- or service-related information than minorities. One possibility is that despite sites such as NetNoir and Black Entertainment Television, general purpose search agents may not be perceived as an effective way to locate Web content that is compelling to minority users (New Media Week, 1997). This suggests the development of search engines and portals targeted to the interests of racial/ethnic groups.

Shopping Behavior

There appear to be no differences between different minorities and caucasians in the incidence of Web shopping. Is this because race does not matter for "lead users" who are most likely to shop, or is this because commercial Web content better targets racial and ethnic groups than does non-commercial Web content? Previous research (Novak, Hoffman, & Yung, 1999) suggests that more skill is required to shop online than to search. However, as noted above, caucasians are more likely to search for information online than are minorities. More generally, consumer behavior in the commercial Web environment is complex and only weakly understood. Further research is needed to explore fully the differences in consumer behavior on the Web and their implications for commercialization.

Multicultural Content

Studies investigating the extent of multicultural content on the Web are needed. Another possibility for the gap between different minorities and caucasians in Web search behavior is that there is insufficient content of interest to different minorities. *Interactive Daily* (1997) claimed that "while there are about 10 million sites on the Web, there are fewer than 500 sites targeted" to different minorities. However, others have commented on the multicultural diversity of the Web. Skriloff (1997) reported, "There are thousands of Web sites with content to appeal to other ethnic groups. Many of these sites are ready-for-prime time with high quality content, graphics, and strategic purpose."

Community Building

Are there different cultural identities for different parts of cyberspace? Schement (1997) notes that by the year 2020, major U.S. cities such as Los Angeles, Chicago, and New

York will have increasingly divergent ethnic profiles and will take on distinctive cultural identities. An important question is whether there are divergent ethnic profiles for areas of cyberspace. While the questions in the three IDSs do not allow us to directly address this issue, our analyses provide some preliminary evidence of divergent ethnic profiles for various Web usage situations. For example, minorities appear to be more likely to use the Web at school and at other locations, and in some cases are more likely to use the Web at work. How much of this is driven by the lack of a PC in the home and how much by other factors we have yet to hypothesize and investigate?

In addition to facilitating community building at the global level, the Web also facilitates neighborhood-level community building. Schwartz (1996) discusses how the Internet can be used as a vehicle for empowering communities. Anderson and Melchior (1995) raise the issue of the ways in which telecommunications can be used to strengthen communities. Thus, we should expect to find neighborhood Web sites emerging as an important aspect of cyberspace, and that these Web sites will parallel the ethnic profiles of the corresponding physical communities.

Income and Education

Income matters, but only after a certain point. Household income explains race differences in Internet access, use, home computer ownership, and PC access at work. In terms of overall access and use, higher household income positively affects access to a computer. But at lower incomes, gaps in access and use between caucasians and minorities existed and were increasing. Research is necessary to determine the efforts most likely to be effective to ensure access for lower-income people, especially minorities. The situation is different with education. As with income, increasing levels of education positively influence access, Web use, PC ownership, and PC access at work. However, caucasians are still more likely than minorities to have access to and use the Internet, and own a home computer, and these gaps persist even after controlling for educational differences.

Tags to Allow a Higher Degree of Social Networking

Since the later half of 2004, the communities on the Internet have embraced a practice of manually tagging content with metadata, which produced a phenomenon described as folksonomies (Mathes, 2004). Until now, metadata has been known as an element attributed to librarians and information architects to manage large pools of data such as collections of books, archives, and so forth. Now, however, metadata has become rather trendy in organizing personal digital information and artifacts by using simple keywords without hierarchies. Whereas in the domain of information architecture and retrieval, the artifacts stored in databases and information systems are usually described with conventional metadata according to a standard with controlled vocabularies and taxonomies, the new trend of folksonomies allowed users and user communities to describe and classify digital artifacts by using the keywords that they felt were suitable — hence the term *folksonomy,* a combination of folk and taxonomy. These keywords, called tags, are now commonly used by many applications for organizing and tagging pictures, bookmarks (URL), blog entries, Web feeds, and other digital content items on the Internet.

Social Content and Social Context

A plethora of recent online tools use tags for the above listed reasons. One main purpose of these tools is to allow people to organize their personal knowledge artifacts (content), whereas the networking and interlinking of people through networks gives us the context where this takes place. The terms *social content* and *social context* are used by social networking specialist Stowe Boyd (2004). By social content, he means any information that people create about themselves to share with others such as preferences, postin gs, or manifests of relationships (e.g., self-proclaimed such as *Friend of a Friend*, *www.foaf project.org*). This could be extended to contain all the digital artifacts that people create both for formal and informal learning.

On the other hand, social context centers on a person's heterogeneous social networks; people are known to have a variety of networks, some being through their social lives, whereas others might be based on interests in professional lives, in hobbies, and such. How to capitalize on this multitude of networks and contacts for the purpose of facilitating social networking is a challenge for current applications. Creating social protocols as we experience them in real life does not translate easily to a form of a software application. Social context in this article has been extended to explain how the tools and applications listed below facilitate the creation of networks trying to imitate the reallife social protocols that users have.

Tools for Knowledge Artifacts and Social Networking

The following is a short summary of some current applications that fall into the category of tools for creation of knowledge artifacts and social networking. They are explained briefly by using the terms *social content* and *social context* as explained above. Blogs and other end user authoring tools are not mentioned in this list, as they are outside the scope of this article. A good reference on the subject can be found by Downes (2005).

Rojo.com (www.rojo.com)

Short description: RSS feed reading and discovery with commenting and sharing though keywords and established groups.

Social content: RSS feeds: find, track, read, and share feeds through tags.

Social context: Allows setting up groups, also sharing through tags. Recommends links though Rojo.

43 Things (*www.43things.com*)

Short description: A site that provides an area where people can write their goals, become inspired by others, and share their process, as well as learn from others how to achieve goals.

Social content: Lists of life goals, desired things to achieve, and places that people plan and wish to visit. Also "have done this" and user profile.

Social context: Connections are built between people who have listed similar aims or desires in order to have a peer group to support one another. Connections can also be made between people who want to achieve some goal and the ones who have already done that in order to give guidance and support.

360° Yahoo! (360.yahoo.com)

Short description: Yahoo!'s service for blogging and networking, allows sharing all types of artifacts, also the ones on external services, with public or restricted groups.

Social content: Own profile, blog, photos, local review s, friends, music, lists of favorite books, movies, music, TV shows, and groups. A personalized page collects different artifacts either from Yahoo! Services or external ones.

Social context: Groups (private, friends, friends of a friend, etc.) can be created and categorized, and artifacts can be shared with groups through Web feeds, notifications on a messenger, and so forth.

The policy implication needs to be carefully considered. To ensure the participation of all people in the information revolution, it is critical to improve the educational and social networking opportunities for minorities. How this might best be achieved is an open research question.

CONCLUSION

The consequences to civil society of the digital divide in Internet use are expected to be severe (Beaupre & Brand-Williams, 1997). Just as Liebling (1960) observed for the freedom of the press, the Internet may provide for equal economic opportunity and democratic communication, but only for those with access. The united world economy may also be at risk if a significant segment of our society, lacking equal access to the Internet, wants the technological skills to keep national firms competitive.

Personal knowledge management in the context of the digital divide is probably an issue that is less discussed nowadays, but one that will gain more public interest as learning management systems and eportfolios are rolled out in education on a wider scope. Being able to organize one's own knowledge as the information revolution progresses — to create personal repositories of knowledge artifacts, and to enhance and enforce new ways of learning through social networks — will be one of the future challenges (classify, link, share, recommend, distribute). New efforts should be put into researching the importance of personal knowledge management in the context of decreasing the digital divide and its potential for enhancing social aspects of bridging the digital divide with ICTs. It could be predicted that social software as described in this chapter plays a major role, even driving the development.

The broad policy implications for these findings should not be overlooked. By identifying those who are truly in need, policymakers can prudently and efficiently target support to these information disadvantaged. Only when this point is reached can *all* those who desire to access online services possibly be accommodated. However, connectivity to all such households will not occur instantaneously; rather, there is a pivotal role to be assumed in the new electronic age by the traditional providers of information access for the general public — the public schools and libraries. These and other "community access centers" can provide, at least during an interim period, a means for electronic access to all those who might not otherwise have such access. Policy prescriptions that include public "safety nets" would complement the long-term strategy of hooking up all those households that want to be connected to the online services.

REFERENCES

Abrams, A. (1997). Diversity and the Internet. *Journal of Commerce,* (June 26).

Anderson, T.E., & Melchior, A. (1995). Assessing telecommunications technology as a tool for urban community building. *Journal of Urban Technology, 3*(1), 29-44.

Arquette, T.J. (2001, September 15). *Assessing the digital divide: Empirical analysis of a meta-analytic framework for assessing the current state of information and communication system development.* Unpublished Draft, Department of Communication Studies, Northwestern University, USA.

Beaupre, B., & Brand-Williams, O. (1997). Sociologists predict chasm between black middle-class, poor will grow. *The Detroit News,* (February 8).

Blau, A. (2002). Access isn't enough. *American Libraries,* (June/July), 50-52.

Boyd, S. (2004, January). *The barriers of content and context.* Retrieved from http://www.darwinmag.com/read/010104/context.html

Compaine, B.M. (Ed.). (2001). *The digital divide. Facing a crisis or creating a myth?* Cambridge, MA: MIT Press

Courtright, C., & Robbin, A. (2001, November 15-17). Deconstructing the digital divide in the United States: An interpretive policy analytic perspective. *Proceedings of the International Association of Media and Communication Research and International Communication Association Symposium on the Digital Divide,* Austin, TX.

Educational Testing Service. (1997). *Computers and classrooms: The status of technology in U.S. schools.* Retrieved from http://www/ets.org/research/pic/compclass.html

eStats. (1999, May 10). *Net market size and growth: U.S. Net users today.* Retrieved from http://www.emarketer.com/estats/nmsg_ust.html

Jackel, M. (2001, November 15-17). Inclusion, exclusion and the diversity of interests. Is digital divide an adequate perspective? *Proceedings of the International Association of Media and Communication Research and International Communication Association Symposium on the Digital Divide,* Austin, TX.

Interactive Daily. (1997). More different minorities plan to go online. *Interactive Daily,* (February 18).

Katz, J., & Aspden, P. (1997, October 6). Motivations for and barriers to Internet usage: Results of a national public opinion survey. *Proceedings of the 24th Annual Telecommunications Policy Research Conference,* Solomons, MD.

Laugksch, R.C. (2000). Scientific literacy: A conceptual overview. *Science Education, 84*(1), 71-94.

Liebling, A.J. (1960). *The New Yorker, 36*(105).

Mathes, A. (2004, December). *Folksonomies — cooperative classification and communication through shared metadata.* Retrieved from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

McLeod, J.M., & Pan, Z. (2002). *Concept explication and theory construction.* School of Journalism and Mass Communication, University of Wisconsin–Madison, USA.

Morrisette, S. (1999, January). *Consumer's digital decade.* Retrieved from http://www.forrester.com/

Nagaraj, N. (2002). The other divides. *Businessline,* (April 24).

New Media Week. (1997). BET, Microsoft sees potential in African-American Audience. *New Media Week,* (March 3).

Nielsen Media Research. (1997). *The Spring '97 CommerceNet/Nielsen Media Internet demographic survey, full report* (vols. I & II). Author.

Novak, T.P., Hoffman, D.L., & Yung, Y.F. (1999). Modeling the flow construct in online environments: A structural modeling approach. *Marketing Science.*

Nowak, M. (2005, August). *One login to bind them all.* Retrieved from http://www.wired.com/news/privacy/0,1848,68329,00.html

Roberts, R.M. (1997). Program lowers costs of going online; families can get break on equipment. *The Atlanta Journal and Constitution,* (June 19).

Rutkowski, A.M. (1998, February). *Internet trends.* Retrieved from http://www.ngi.org/trends.htm

Sartori, G. (1984). Guidelines for concept analysis. In G. Sartoni (Ed.), *Social science concepts: A systematic analysis* (pp. 15-85). Beverly Hills, CA: Sage.

Sax, L.J., Astin, A.W., Korn, W.S., & Mahoney, K.M. (1998). *The American freshman: National norms for fall 1998.* Retrieved from http://www.acenet.edu/news/press_release/1999/01January/freshman_survey.html

Schement, J.R. (1997). Thorough Americans: Minorities and the new media. *Proceedings of the Aspen Institute Forum.*

Schools and Libraries Corporation. (1998, November 23). *First wave of e-rate funding commitment letters sent.* News Release, Schools and Libraries Corporation, USA.

Schwartz, E. (1996). *NetActivism: How citizens use the Internet.* Sebastopol, CA: O'Reilly & Associates.

Sheppard, N. (1997). Free-Nets reach out to communities' needs. *The Ethnic NewsWatch,* (April 30).

Skriloff, L. (1997). Out of the box: A diverse Netizenry. *Brandweek,* (February 17).

Tichenor, P.J., Donohue, G.A., & Olien, C.N. (1970). Mass media and differential growth in knowledge. *Public Opinion Quarterly, 34,* 158-170.

U.S. Department of Commerce. (1995). *Falling through the net: A survey of the have nots in rural and urban America.* Retrieved from http://www.ntia.doc.gov/ntiahome/fallingthru.html

U.S. Department of Commerce. (2001). *Falling through the net: Toward digital inclusion.* Retrieved from http://www.esa.doc.gov/fttn00.pdf

KEY TERMS

Artifacts: Tools to manage one's personal knowledge artifacts by categorizing them using novel approaches such as tags and ratings, but also more conventional forms of metadata.

Digital Divide: A gap between the information "haves" and "have-nots."

Folksonomy: A combination of folk and taxonomy.

Knowledge Gap: The different knowledge possession through mass media by social groups with varied social-economic status.

Personal Knowledge Management: Storing personal information online, for example favorite recipes, which allow people not only to access them from any computer connected to the Internet, but also arrange them using personally meaningful keywords.

Social Networking: A phenomenon described as a community consisting of a collection of individuals and the linkages among them.

Digital Literacy and the Position of the End-User

Steven Utsi
K.U.Leuven, Belgium

Joost Lowyck
K.U. Leuven, Belgium

INTRODUCTION

As an educational setting, the traditional classroom fails to meet the learner's need for suitable skills to learn with educational software. The development of digital learning skills in school curricula challenges designers of educational software. A useful starting point of research in this domain is the study of literacy, both in its traditional and new forms (Tyner, 1998). It is a powerful background for research on the interaction of learners with educational software platforms. A "platfom" is a particular software package, designed for educational use.

BACKGROUND

Both in school and society, the skill to comprehend and handle printed course materials is essential. Literacy has since long been a vital skill for functioning adequately in an industrial society (see e.g. Marvin, 1984).

An Emerging Plural Notion of Literacy

The International Adult Literacy Survey (IALS) describes literacy as a broad range of information processing skills in relation to written or printed language. Traditional literacy is defined as follows (OECD, 1997, p. 2):

"Using printed and written information to function in society,to achieve one's goal and to develop one's knowledge and potential."

However, traditional literacy is increasingly evolving into a new, plural literacy that refers to making sense of meaningful content in more complex and technological environments (Erstad, 1998). The growing importance of images and of communication technologies has a cultural backlash that even transforms the nature of literacy. Gee (1990) opened up so-called "New Literacy Studies" (NLS). He defends a socio-cultural approach of literacy (p. 153):

"Literacy is the mastery of, or fluent control over, secondary discourse."

While primary discourse pertains to infant face-to-face interaction of children with trusted figures (parents, family, and others), secondary discourse develops through contact with public life and its social and cultural conventions. Secondary literacy is in itself a plural concept: a multitude of social institutions and commitments to public life invade an adult's life and are as many "literacies" to master. As Walter (1999, p. 34) points out:

"The existence of multiple literacies, none more valid than the next, but each specific to a culturally-defined community."

According to this plural notion of literacy, literacy can be neither neutral nor universal, since all literacy includes social and cultural conventions that shape a particular type of "literacy". Visual literacy, for instance, complements traditional literacy and claims a unique position in today's school curriculum. Debes (1969) first mentioned "visual literacy". According to visual literacy, a specific "image" language supports communication. In traditional language, words support verbal communication. Visual literacy may not only be a means of communication, but also a way of thinking (Hortin, 1983). Thinking visually, then, means the ability to think and learn in terms of images. And children's acquisition of skills to work effectively and efficiently with educational software has to underpin this recent position of a new and full interpretation of literacy.

Undoubtedly, it is of prime importance to analyse the nature of skills necessary to take full advantage of today's learning opportunities. In a visual oriented culture the acquisition of new reading and writing skills is indispensable, e.g. the analysis and composition of images. Indeed, literacy supposes an active intervention in a social and cultural context. Avgernou and Ericson (1997) define visual literacy as a group of skills that make it possible for an individual to understand and use visuals for intentional communication with others. This concerns different target groups, for instance primary school pupils or even impaired children.

During the last decade, a wide array of "literacies" relating to information and communication technologies (ICT) surfaced: media literacy (Hobbs, 1998; Potter, 1998), electronic literacy (Maylath, 1993), multimedia literacy (Kellner, 1998), computer literacy (Guthrie & Richardson, 1995; Peha, 1995), and digital literacy (Gilster, 1997). This evolution accompanies the expansion of IT to ICT. Indeed, communication is now a central feature of technological environments, clearly depending on both "traditional" and "new" literacies (Plowman & Stephen, 2003):

"(...) the flexible and sustainable mastery of a repertoire of practices with the texts of traditional and new communication technologies via spoken language, print and multimedia."

The overarching notion "information literacy" denotes the ability to access, retrieve, manage, and use information relevant to an identified need for information (Kuhltau, cit. in Campbell, 1994). Originally, information literacy was limited in scope to computer information. The progress of computer sciences and, more generally, the use of ICT in a wide array of domains broadened its meaning into library skills, computer skills, thinking skills, and critical reading skills.

Media literacy pertains to communication through and critical analysis of a diversity of media; it is the end user's ability to navigate both effectively and efficiently and to keep track of position in electronic media, while "criss-crossing the landscape" (Spiro, R. J., Feltovich, R. L., Jacobson, M. J., & Coulson, R. L., 1991). Gilster (1997, p. 1) defines digital literacy as follows:

"(…) the ability to understand and use information in multiple format from a wide range of sources when it is presented via computers."

Computer literacy is the ability to integrate information and build a personal knowledge base. Both electronic literacy (e-mail reading skills) and multimedia literacy (technical multimedia skills) are building blocks of more general "computer" literacy. Electronic and multimedia literacy explain, for instance, the comprehension of hypertext.

When comparing different "literacies", two observations are important. First, critical analysis, interpretation, and processing of information are attributed to media literacy and digital literacy. The processing and integration of information (computer literacy) and technical skills (electronic and multimedia literacy) have to be critically evaluated by computer users. Secondly, without the notion of traditional and visual literacy, none of the newer forms of literacy can be understood. Indeed, media and digital literacy acquire meaning for users through similar basic mechanisms as traditional and visual literacy. Literacy education elucidates implicit messages, ideological content or even idiosyncratic intentions designers may embed in software packages. On the other hand, the study of ICT related literacies informs software designers of problems encountered by learners with educational software platforms. Traditional issues are accessibility of information and user interface design.

Current Research Questions

The "literacies" debate is a theoretical starting point. Empirically, the detection of specific skills that explain interaction with educational software -digital literacy- is a first research path. These skills have to be integrated in the school curriculum and are treated as abilities underlying new "literacies". Before any application of theoretical insights, a primary research question has to pertain to the relationship between "operational skills" (searching, clicking, and/or dragging screen and user interface objects) and content comprehension in educational software. Is retrieval of information influenced by the mastery of operational skills?

Moreover, information can be represented through text, visualization, or talk. Does the integration of these different symbol systems in educational software alter the typical linear end-user interaction with the computer screen interface? The most common pattern of software use is sequencing interface screen after interface screen in a so-called linear fashion. Clicking hotspots and exploring additional in-depth layers of screens, providing e.g. background information, are seldom spontaneous actions. This type of research question addresses conditions that facilitate "switching content" in -for instance- an educational software package fitted out with hotspots and hyperlinks. The content of an educational platform can for example be organized in an adventure game with hyperlinks, combined with an illustrated encyclopaedia supporting the game with textual and verbal background information. A related question points to the relationship between switching content and retrieving or remembering information afterwards. Is switching detrimental to retrieval of information or does it on the contrary support memory?

Research with 3rd and 4th graders using a multimedia comic strip about World War II (see Utsi & Lowyck, 2002) revealed end-users to anticipate crucial events: they look for objects in the interface screens that most probably will play a crucial role in the next few screens. Mere reactions to audio-visual events in interface screens steadily fade, while searching, clicking, and/or dragging objects become increasingly well-considered throughout the user-interface interaction. Throughout the process, visual literacy gradually changes from superficial use of visual cues to deeper comprehension of educational content. Thus, visual literacy is an essential condition for meeting the educational goals. Multimedia literacy skills are effortlessly acquired on the spot: clicking and dragging objects pose no problem. When first confronted with a new educational software platform, visual literacy seems narrowed to multimedia skills, like

clicking and dragging, but gradually visual literacy opens up again to more thoughtful, content driven interaction with all the platform's trimmings.

FUTURE TRENDS

Future studies in this research field can make use of the cognitive load model of Kalyuga, Chandler, and Sweller (2000), who suggest a trade-off effect between text and visual cues. Cognitive load can be defined as the amount of mental resources necessary for information processing. High cognitive load requires the end-user to spend extra memory resources to deal with new incoming information. Accordingly, text and visual cues compete for attention of users since the cognitive span of learners is limited, while effort needed for both task completion -clicking one's way through the platform's interface screens- and content comprehension requires maximum investment of cognitive resources and attention (see also Kirschner, 2002). The trade-off of text and visual cues (Utsi & Lowyck, 2002) yields a clear distinction between low performing pupils clinging to visual cues and the ones who use the full range of novel experiences and skills offered in the educational platform.

Processing and integrating information from different sources is a digital and "computer" (Guthrie & Richardson, 1995; Peha, 1995) literacy skill that may acquire its place in school curricula in due time. Recent and future studies can provide input for implementing research findings with regard to digital literacy in educational software. Like any other type of ICT, educational multimedia runs through a cycle of design, development, and implementation. The shift, in the strip story mentioned above, from the user's shallow reactions to visual stimuli toward anticipation of events is a leading theme. This anticipation can be triggered by visual cues or by story line comprehension. Pupils seem to encounter difficulties to cope with the different layers of multimedia information: text, pictures, motion, music, and sound. They hardly break up the linear nature of the "storyline": background information is scarcely accessed. A core concern while designing, developing, and implementing educational media is the trade-off of textual and visual information in educational multimedia, hampering the break up of linear content material. Textual and visual symbol systems compete for the attention of learners, involved in learning with educational software.

Designers of educational software need to try and ease the integration of text and visuals. Throughout the realization of educational software, care should be taken that the precise role of hyperlinks and hotspots is clear to the user. Indeed, available cognitive span is highly occupied by an educational platform using different symbol systems, such as textual and visual information. The end user's decision to access an additional content layer in the platform needs to be an informed decision. If the learner does not know where a hyperlink or a hotspot will lead to, he may neglect to select the hyperlink, because he is not aware of the relevance of this information for the task at hand. The type and relevance of information hyperlinks and hotspots can be realized through embedded, formal features (for instance glowing or blinking) that signal the information value of a hyperlink or hotspot.

Furthermore, educational software platforms presenting information and exercises in a rousing setting, for instance in an adventure game context, need to ensure that a balanced arch of task tension is built up throughout the platform. Task tension is the "track" of the learner to reach intermediary goals that eventually lead to a successful mastery of an assignment. Regularly, the learner has to be attended to the end goal, for example retrieving a tool for repairing a space ship. This tool has to turn up visually from time to time or nearing it may be signalled by a characteristic sound. Ideally, this arch of task tension implies a series of pauses: moments for the learner to relax and explore course content from another point of view, for instance through the access of background information. Ultimately, tension is discharged in a culminating accomplishment of the assignment or end goal. Pauses can be created through opening up rigid linear progress from one content element or exercise to another and/or and through filling in the "gaps" with interesting, but distracting information or modest tasks. In an educational adventure game, for instance a journey through outer space, a simple, but distracting lottery game can be integrated, without any relevance for the end goal of the adventure game: retrieving a tool for repairing a space ship. Educational material that is presented in a serial, linear fashion without consideration of the limited processing capacities of young learners is doomed to be incompletely and thus inadequately processed.

Information sticks better when the end user's attention is triggered by repeating screen elements that suggest an optimal route across specific content elements offered at a learner's initial level. The design, development, and implementation of educational ICT need to balance the signalling function of visual cues, routing pupils' clicking behaviour, and the effort of learners to look for content, relevant for reaching goals. In educational software, this can for instance be achieved via a "road map", lending learners a hand in orienting themselves on the way to their goal. Small screen elements cue the learner's attention and break up the linear character of traditional course material, if presented in a low task tension context.

CONCLUSION

Digital literacy is a baseline set of skills for successfully coping with a complex, often technological world, holding multiple media messages. In line with traditional reading

and writing training, digital literate learners may also cope successfully with new ways of communicating. Digital literacy comprises technical computer skills (searching, clicking, and/or dragging) and visual cue awareness that triggers reading and deepened understanding of information. Support of the user-platform interaction needs to be embedded in the design, development, and implementation of educational software platforms. This interaction is more than merely user-friendly, but it is challenging the in-depth understanding of the information at hand. Cognitive load is an important constraint for unconcerned use of educational software. Apart from hampering processing and transferring information to memory, cognitive load undermines spontaneous exploration across content layers. The linear nature of most actual course material can be broken up by the users' inquisitiveness, but only if they decide to do so and if actions to switch to parallel content layers in an educational software platform do not imply a cognitive burden. The integration of visuals and other media types, text and sound, has to be well-considered. Consistent appearance and use of hyperlinks and hotspots, an arch of tension with some moments to pause, and repeating screen elements are important in designing an appealing and effective educational software platform.

REFERENCES

Avgerinou, M. & Ericson, J. (1997). A review of the concept of visual literacy. *British Journal of Educational Technology, 28*, 280-291.

Campbell, B. S. (1994). *High school principal roles and implementation themes for mainstreaming information literacy instruction.* Unpublished doctoral dissertation, University of Connecticut.

Debes, J. (1969). The loom of visual literacy. *Audiovisual Instruction, 14* (8), 25-27.

Erstad, O. (1998). Media literacy among young people: Integrating culture, communication and cognition. In B. Höijer & A. Werner (Eds.), *Cultural cognition: New perspectives in audience theory* (pp. 85-101). Göteborg: Nordicom.

Gee, J. P. (1990). *Social linguistics and literacies: Ideology in discourses*. New York: The Falmer Press.

Gilster, P. (1997). *Digital literacy*. New York: Wiley.

Guthrie, L. F., & Richardson, S. (1995). Language arts: Computer literacy in the primary grades. *Educational Leadership, 53*(2), 14-17.

Hobbs, R. (1998). Literacy in the information age. In J. Flood, S. B. Heath, & D. Lapp (Eds.), *Handbook of research on teaching literacy through the communicative and visual arts* (pp. 7-14). New York: Simon and Schuster Macmillan.

Hortin, J. (1983). Visual literacy and visual thinking. In L. Burbank & D. Pett (Eds.), *Contributions to the study of visual literacy* (pp. 92-106). Blacksburg, Virginia: IVLA Inc.

Kalyuga, S., Chandler, P., & Sweller, J. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology. 92*(1), 126-136.

Kellner, D. (1998). Multiple literacies and the critical pedagogy in a multicultural society. *Educational Technology, 48*(1), 103-122.

Kirschner, P. (2002). Cognitive load theory: implications of cognitive load theory on the design of learning. *Learning and Instruction, 12*, 1-10.

Marvin, C. (1984). Constructed and reconstructed discourse: Inscription and talk in the history of literacy. *Communication Research, 11*(4), 563-594.

Maylath, B. (1993). Electronic literacy: What's in store for writing and its instruction? Paper presented at the *Annual Meeting of the Conference on College Composition and Communication* (44th, March 31-April 3), San Diego, CA,

OECD (1997). *Literacy skills for the knowledge society: Further results from the International Adult Literacy Survey*. Paris: OECD/HRD Canada.

Peha, J. M. (1995). How K-12 teachers are using networks. *Educational Leadership, 53*(2), 18-25.

Plowman, L., & Stephen, C. (2003). A "benign addition"? Research on ICT and pre-school children. *Journal of Computer Assisted Learning, 19*, 149-164.

Potter, W. J. (1998). *Media literacy*. London: Sage.

Spiro, R. J., Feltovich, R. L., Jacobson, M. J., & Coulson, R. L. (1991). Cognitive flexibility, constructivism, and hypertext: Random access instruction for advanced knowledge acquisition in ill-structured domains. *Educational Technology, 31*, 24-33.

Tyner, K. (1998). *Literacy in a digital world*. Wahwah, N.J.: Lawrence Erlbaum

Utsi, S., & Lowyck, J. (2002). Empirical validation of the concept "multimedia literacy". Paper presented at the *SIG Instructional Design of the European Association for Research on Learning and Instruction* (EARLI) (Erfurt, 27th - 29th June).

Walter, P. (1999). Defining literacy and its consequences in the developing world. *International Journal of Lifelong Education, 18*(1), 31-48.

KEY TERMS

Children: Tutees, enrolled in primary school.

Cognitive Load: Amount of mental resources necessary for information processing.

Educational Software: Software packages, supporting specific goals in the education of target groups, e.g. primary school tutees or impaired children.

End-User: Tutee, working with dedicated educational software packages.

Impaired Learners: learners, hampered by physical of psychological deficiencies.

Instructional Design: Lay-out of an optimal integration of educational content and interface lay out of end-user software.

Learning: Cognitive processing and integration of new educational content, if possible induced through exercises or games.

Literacy: Operational and cognitive skills, necessary to work effectively and efficiently with educational software.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 875-879,

Digital Video Broadcasting Applications for Handhelds

Georgios Gardikis
University of the Aegean, Greece

Harilaos Koumaras
University of the Aegean, Greece

Anastasios Kourtis
National Centre for Scientific Research "Demokritos", Greece

INTRODUCTION

Following the success and wide adoption of the European Digital Video Broadcasting for Terrestrial (DVB-T) standard for digital terrestrial television, numerous coordinated research efforts on digital broadcast technology resulted in the recent standardization of Digital Video Broadcasting for Handeld Devices (DVB-H). The new specification aims at defining the physical and link-layer level of a digital broadcast network for Internet protocol (IP) datacasting services. At its core, DVB-H is based on DVB-T but it is more oriented in mobile and stationary reception by handheld devices. This article attempts a brief though thorough overview of the new technology, its technical aspects, and its new application perspectives.

BACKGROUND

During the mid-1990s, the MPEG-2 Transport Stream (TS) (International Organization for Standardization [ISO], 1996) was accepted worldwide as baseband format for digital television networks. Its structure allows the transmission of encoded digital video and audio streams, along with IP data, organized in a statistical Time Division Multiplex (TDM). The need for an efficient physical layer arose, which would deliver the MPEG-2 TS to the end-user terminals via the "difficult" terrestrial channel.

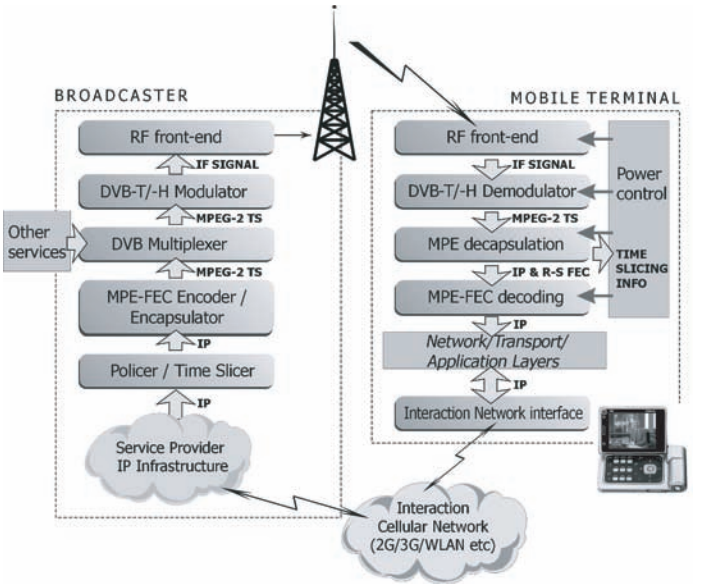
Several research efforts have been conducted around the world to optimize the physical layer for terrestrial digital television (DTV). North America adopted the ATSC A/53 system, developed by the Advanced Television Systems Committee (ATSC) in 1995, based on 8-VSB modulation. In Japan, the Association of Radio Industries and Businesses developed in 1998 the Integrated Services Digital Broadcasting-Terrestrial (ISDB-T) specification for the same purpose.

In Europe, DVB-T was standardized by the European Telecommunications Standards Institute (ETSI) in 1997 as a transmission system designed and optimized for terrestrial DTV configurations. Although it was initially designed for stationary use, DVB-T also presented an outstanding performance in mobile reception (Stare, 1998), where it outclassed ATSC (Wu, Pliszka, et al., 2000). To further support the perspective of mobile DTV, ETSI introduced in 2004 the DVB-H specification (*Digital Video Broadcasting (DVB); Transmission System for Handheld terminals (DVB-H),* 2004). DVB-H substantially comprises of a set of extensions to DVB-T which are oriented to handheld use. DVB-H inherits all the benefits of its predecessor and adds new, mobile-oriented features, focusing on IP datacasting and including better mobility and handover support, adaptive per-service error protection and power saving capabilities. At present, DVB-H is the dominant open standard in its field, and compliant systems are being deployed around the world, including Europe, the United States, and China. A strong competitor of DVB-H is Terrestrial Digital Multimedia Broadcasting (T-DMB), a standard developed in Korea and Japan, based on the European Digital Audio Broadcasting (DAB). A third player in the field of handheld DTV is Media Forward Link Only (MediaFLO), a U.S. proprietary technology developed by Qualcomm, which is gaining ground in North America.

The ETSI specification defines DVB-H as a "broadcast transmission system for datagrams." Like DVB-T, it specifies the physical and link layers, along with the service information. A DVB-H-compliant broadcast platform consists substantially of a DVB-T chain, including all the enhancements introduced by the new specification (Figure 1). Since a broadcast platform has no native support for interactivity, an IP-based cellular infrastructure (like WLAN, 2G/3G) can be employed complementarily to enable for fully interactive applications.

It must be clarified that most of the innovative features of DVB-H, as explained in the next section, are implemented on the link layer and do not affect the DVB-T physical layer. This allows the new technology to inherit all the benefits

Figure 1. Block diagram of a DVB-H system



of its predecessor, including flexible transmission schemes providing from 5 up to 32Mbps of capacity, excellent multi-path performance, due to the use of OFDM (Orthogonal Frequency Division Multiplexing), use of TV bands UHF using 8 MHz channels, and SFN-based operation.

In order to use the link-layer features of DVB-H, it is assumed that the useful payload to be conveyed consists of IP-datagrams (or other network layer datagrams) which are transmitted within the MPEG-2 TS, encapsulated according to the Multi Protocol Encapsulation (MPE) protocol. With this assumption, DVB-H becomes a totally IP-oriented system and does not support native MPEG-2 audiovisual streams. It is however feasible (although not recommended), that DVB-H services can coexist with traditional, DVB-T, MPEG-2-based DTV programs within the same multiplex.

DVB-H TECHNICAL INNOVATIONS AND APPLICATION PERSPECTIVES

The new features of DVB-H were introduced taking into consideration three principal issues in mobile use: (1) handover/mobility, (2) varying signal reception conditions, and (3) limited battery time. DVB-H innovations can be summarized as follows:

DVB-H innovations at physical layer (extensions to DVB-T)

- **4K FFT Mode:** Native DVB-T operates at two modes (8K and 2K), referring to the number of carriers within the OFDM spectrum. 8K mode (6817 carriers) provides a longer symbol period, having a very good performance in large Single Frequency Networks (SFNs) due to better tolerance in long echoes. However, it is unsuitable for fast-moving receivers, since it is very vulnerable to Doppler shift, having relatively small intercarrier spacing. On the other hand, 2K mode (1705 carriers) provides improved Doppler performance but its behavior in SFN networks is poor. DVB-H introduces the 4K mode (3409 carriers) as a trade-off, combining good mobile reception with acceptable performance in small and medium SFNs.
- **Additional TPS Signalling:** Transmission Parameter Signalling (TPS) bits within the OFDM symbol carry additional DVB-H related information to enhance and speed up service discovery. TPS also carries cell-specific information, which assists the handover procedure in mobile receivers.
- **In-depth Symbol Interleaver:** The DVB-T symbol interleaver requires a certain buffer size both in the transmitter and the receiver. When switching from 8K mode to 4K or 2K, the buffer required for the process falls to ½ and ¼ respectively, since the size of the symbol (in bits) also decreases. DVB-H exploits the unused buffer by increasing the interleaving depth by a factor of 2 (4K) and 4 (2K), thus increasing tolerance to impulse interference.

DVB-H innovations at link layer

- **Time Slicing:** A basic issue in handheld operation is the limited battery time. This issue is of particular importance in terrestrial DVB reception, where the receiver/demodulator/demultiplexing/decapsulation chain consumes typically 1W. The time slicing fea-

ture of DVB-H aims at reducing the average power consumption by allowing the terminal to know when to expect data and to switch off the receiving chain when not needed (i.e., when the transmitted data are of no interest to the specific receiver). At the broadcaster, prior to encapsulation, IP data belonging to a certain stream are organized in TDM bursts. During encapsulation, each IP section is tagged with a "delta-t" value, which informs the receiver about the time interval until the next burst. This information allows the receiver to switch off until the next burst of data arrives (Figure 2). Practically, the duration of one burst is in the range of several hundred milliseconds, whereas the powersave time may amount to several seconds. A typical power saving up to 90% is expected, whereas this figure depends on the number and the bit rate of the IP services that the terminal is "listening" to.

- **Multi Protocol Encapsulation-Forward Error Correction(MPE-FEC):** DVB-T includes two layers of error-protection coding, namely a transport-level Reed-Solomon and an inner convolutional coder. These methods protect the transport stream as a whole and have been proven to be very effective. DVB-H introduces an additional FEC layer, prior to encapsulation, which can be applied on a per-stream basis. The MPE-FEC method organizes the IP datagrams in a table, column-by-column, and then protects each row of the table with a Reed-Solomon overhead, as shown in Figure 3. IP datagrams are then separately encapsulated and transmitted from the FEC data. The latter can be discarded by FEC-ignorant receivers, thus making the method backwards compatible. Puncturing, applied on the useful data or the FEC overhead, can result in either stronger or weaker coding, respectively. MPE-FEC allows the broadcaster to apply a different level or protection on each broadcast IP service, depending either on the importance of the service, and/or on the reception conditions of the terminal(s) to which the service is targeted. Intensive testing of DVB-H, which was carried out by DVB member companies in autumn of 2004, showed that the use of MPE-FEC can result in a coding gain of some 7 dB over DVB-T (Kornfeld & Reimers, 2005).

The efficiency and flexibility of DVB-T, enhanced with the new, mobile-oriented features of DVB-H, open virtually innumerable application perspectives for the digital broadcasting market. Validation efforts include services based on anywhere-anytime access, with portable devices and mobile terminals in cars, trains, and other transportation media with very high success. The report of the validation phase is extensively presented in *Digital Video Broadcasting (DVB); Transmission to Handheld Terminals (DVB-H) Validation Task Force Report* (2005).

Very promising service scenarios are expected via the synergy between digital broadcasting and cellular networks (Rauch & Kelleler, 2001; Xu, Tonjes, 2000). Hybrid DVB-H/cellular handheld terminals are soon to appear in the market (Figure 4). At the moment, as the official DVB-H site reports (The DVB Project Office, 2006) 33 companies worldwide

Figure 2. The principle of time slicing

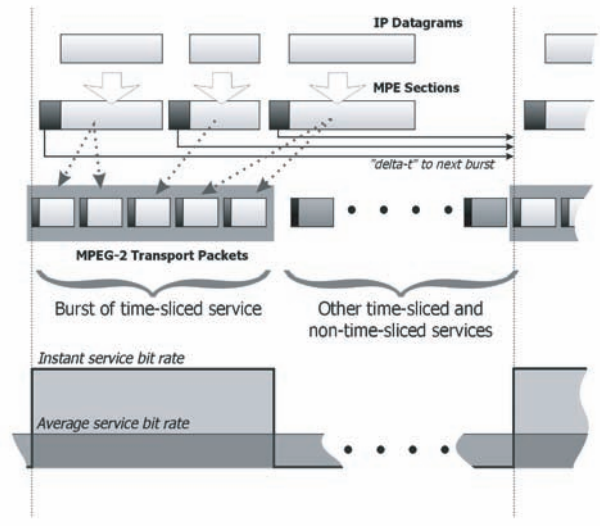


Figure 3. IP data protection via MPE-FEC

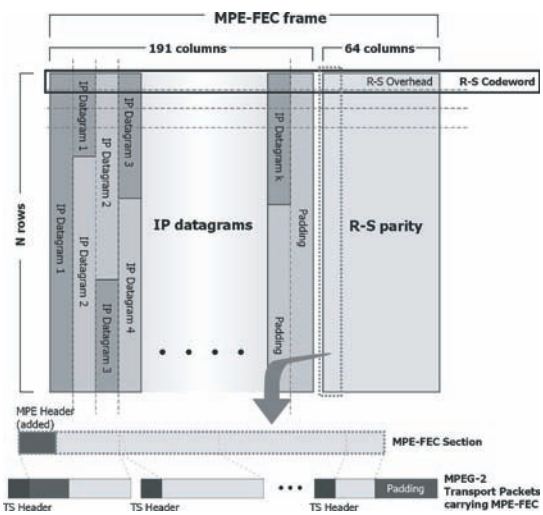


Figure 4. Experimental hybrid DVB-H/cellular handheld terminal

are finalizing their DVB-H receivers and 42 manufacturers have DVB-H headends commercially available.

Service scenarios which can exploit the capabilities of the new broadcast technology can be discriminated according to the degree of interactivity they require: *Noninteractive* (broadcast) applications are unidirectional, as data are broadcast to the terminals, allowing only for local pseudo interactivity between the user and the terminal. Applications with *low interactivity* include the occasional transmission of small blocks of data back to the broadcaster, via the interaction cellular network (e.g., Tele-voting via SMS or IP). *Fully interactive* applications require a constant, bidirectional, asymmetric flow of data between the broadcaster and the terminal.

Future use cases of DVB-H include, but are not limited to, the following scenarios:

- **Digital Television Broadcasting:** DVB is the fundamental use of every DVB platform. The innovation introduced by DVB-H is that DTV programs are no longer limited to MPEG-2 encoding, but are conveyed over IP using state-of-the-art encoding protocols, like MPEG-4 or H.264/AVC. Early DVB-H tests of video streams encoded at these formats showed that a pleasant viewing experience can be achieved on a handheld device at common intermediate format (CIF) resolution using a rate of around 300 kbps. This means that a 10 Mbps downlink can accommodate more than 30 simultaneous programs. Moreover, the use of MPE-FEC gives the broadcaster the potential to prioritize the various DTV programs, assigning a different degree of error protection to each of them. Small hard disks or flash memory modules incorporated in the DVB-H terminals can be employed to add video recording capabilities.
- **Scrambled DTV Transmission:** Currently, DVB platforms utilize transport-oriented proprietary conditional access (CA) methods for scrambling pay-TV content. In DVB-H, where video streams are conveyed over IP, security mechanisms can be elevated to the network layer. All state-of-the-art authentication and security mechanisms designed for IP networks can be used for encrypting data, including IPsec, and multicast key management.
- **Push/caching of DTV Content (News, Weather Forecast, Sports Flash, etc.):** "Idle" DVB-H terminals can work in the background to receive broadcast multimedia content and store it locally. The user can then access the content and view it off-line, whenever appropriate. For example, a citizen going to work may use the handheld terminal when waiting for the bus to view the latest news, or a traveler can watch a cached movie during travel. This feature can be employed at no operating cost for the broadcaster as no bandwidth-per-user is required. Indeed, broadcast platforms like DVB-H are extremely cost- and spectrum-efficient when the same content is to be distributed to a large target group. This is not the case with 3G-based TV streaming, which is very expensive since additional bandwidth must be allocated for every user joining the service.
- **Message Alerts:** A multicast-based alert service can enable users to stay informed about events of their interests such as a "breaking news" event, a goal which was scored, or an unusual stock fluctuation. A text message can be accompanied with audiovisual content. Since the common downlink is used for all customers, this service can also be provided at no cost.
- **Enhanced Interactivity DTV Programs:** The foundation of all DVB-H services over IP and the use of an interaction channel via a cellular network allows for new, truly interactive mobile DTV services, including televoting, e-shopping, participation in quiz shows, questionnaire filling, via the easy-to-use handheld devices, all over IP. The built-in cameras of hybrid terminals can also enable the real-time transmission of pictures/sound/video to the broadcaster, thus enabling a fully interactive television. The *mobile active viewer* scenario envisages that the citizens act as journalists, by providing live feeds—when needed—to the broadcasters, giving instant, on-site news coverage.
- **Push/Caching of Web Content:** A broadcaster may decide to allocate a portion of the DVB-H bandwidth for multicasting popular Web content to an unlimited number of terminals for time-shifted, off-line use, as Stare (2002) suggests. Such content may include electronic newspapers, traffic reports, stock quotes, or entertainment guides. Given that the memory capacity

of the handheld terminals is constantly increasing, it is possible for the broadcaster to allocate, for example, 2 Mbps for this type of service, providing the customers with a complete Web site of 300 pages (assuming an average page size of 50 Kbytes) at their hands after only 1 minute of transmission! The users are unaware of the caching procedure and are experiencing a very high *virtual bandwidth* as they access the content offline. Time slicing can be employed along with proper service information to enable the terminal to save memory and precious battery time by caching only the information in which the user is interested in.

- **Full On-Demand Access to Data and Multimedia Content:** If the content in which the user is interested in is neither broadcast nor cached, the hybrid cellular-broadcast topology can be used to retrieve the data on demand. By sending the requests/acknowledgments via the cellular network and receiving the data via the DVB downlink (hybrid asymmetric access), several Mbps of download rate can be achieved (Gardikis, Kormentzas, Xilouris, Koumaras, & Kourtis, 2005). An interesting approach is that of the load sharing between the broadcast and the cellular network: When many users request the same block of information, the delivery is performed over DVB-H so that all users benefit from the common downlink. If there are only a few requests, the data is delivered to each user via the cellular network. In the latter case, the download rate can be much lower, but the wasting of the DVB capacity is avoided (Cosmas, Itegaki, Cruickshank, 2003).
- **DVB-H Service Continuity Using the Cellular Interaction Network:** In the case that the user roams outside the DVB-H coverage area, a tight broadcast/cellular synergy could enable for the continuity of the DTV service by routing it exclusively via the cellular network at lower quality and, presumably, at higher cost.
- **Emergency Systems:** In the case of a widespread emergency situation (e.g., a natural disaster or a massive terrorist attack), where cellular networks usually collapse, DVB-H can realize a low-cost and always-on backup broadcast system. A centralized authority can use a DVB-H transmitter to broadcast encrypted or unencrypted material with high error protection to ambulances, police cars, and so forth.

FUTURE TRENDS

In order for all the aforementioned services, along with many more to come, can be deployed in a uniform basis across different countries, a lot of standardization effort is to be devoted in the near future regarding the architecture and software issues of DVB-H platforms. Towards this direction, the DVB Convergence of Broadcasting and Mobile Services (CBMS) working group has set an initial framework for use cases and services of DVB-H (*IP Datacast over DVB-H: Use Cases and Services,* 2005a). The same group is also responsible for specifying the video and audio formats, the Electronic Service Guide (ESG), and the content protection aspects of the DVB-H standard. It has also provided guidelines for the ESG information flow, defined interfaces among the various network entities and illustrated the way the components in IP Datacast over DVB-H work together (*IP Datacast over DVB-H: Architecture*, 2005b).

There is also a lot of work to be done in the migration of the state-of-the-art features of IP into the world of DVB. The transfer of the benefits of IPv6 in a DVB-H platform can result in a unified architecture for providing dynamic addressing, mobility/handover support, and increased security.

CONCLUSION

This article outlined the innovative features of the DVB-H technology and demonstrated how they can be exploited via numerous service scenarios, tailored to suit the capabilities of a hybrid broadcast/cellular network. The enhancements introduced by DVB-H, combined with the efficiency of the DVB-T-based physical layer and the synergy with an IP-based cellular network for interaction, opens innumerable application perspectives to this new digital broadcasting system. By offering high data rates, very good mobile performance, flexibility, and interactivity, DVB-H brings the dream of *mobile broadband access* much closer to realization.

ACKNOWLEDGMENTS

The DVB-H-related research effort from which this paper was derived is carried out within the PYTHAGORAS II research framework, jointly funded by the European Union and the Hellenic Ministry of Education.

REFERENCES

Cosmas, J., Itegaki, T., Cruickshank, L., et al. (2003). Converged DVB-T and GPRS service scenarios and application production tools. *Proceedings of the DTV Workshop, CONFTELE 2003,* Aveiro, Portugal (pp. 5-8).

Digital Video Broadcasting (DVB): Transmission System for Handheld terminals (DVB-H). (2004). *European Standard, ETSI EN 302 304 v.1.1.1.*

Digital Video Broadcasting (DVB): Transmission to Handheld Terminals (DVB-H). (2005). Validation Task Force Report. *ETSI TR 102 401.*

Gardikis, G., Kormentzas, G., Xilouris, G., Koumaras, H., & Kourtis, A. (2005, July 18-20). Broadband data access over hybrid DVB-T networks. *Proceedings of the 3rd Conference on Heterogeneous Networks (HET-NETs), 05.* Ilkley, UK.

International Organization for Standardization (ISO). (1996). *Generic coding of moving pictures and associated audio information (MPEG-2) Part 1: Systems.* (ISO/IEC 13818-1). Geneva, Switzerland: ISO.

IP Datacast over DVB-H: Use Cases and Services. (2005a). *DVB Document A097.*

IP Datacast over DVB-H: Architecture. (2005b). *DVB Document A098.*

Kornfeld, M., & Reimers, U. (2005, January). DVB-H—The emerging standard for mobile data communication. *EBU Tech. Rev.*

Rauch, C., & Kelleler, W. (2001, February). Hybrid mobile interactive services combining DVB-T and GPRS. *Proceedings of the European Personal Mobile Communication Conference (EPMCC) 2001.* Vienna, Austria.

Stare, E. (1998). Mobile reception of 2K and 8K DVB-T signals. *Proceedings of the International Broadcasting Convention, 98* (pp. 473-478).

Stare, E. (2002, June). *Hybrid broadcast-telecom systems for spectrum efficient mobile broadband Internet access.* Retrieved June 28, 2006, from http://www.s3.kth.se/signal/edu/seminar/01/Mobile.Broadband.Internet.Access.pdf

The DVB Project Office. (2006). *DVB-H, the global mobile TV.* Retrieved June 28, 2006, from http://www.dvb-h-online.org

Wu, Y., Pliszka, E., et al. (2000, June). Comparison of terrestrial DTV transmission systems: The ATSC 8-VSB, the DVB-T COFDM and the ISDB-T BST-OFDM. *IEEE Trans. on Broadcasting, 46*(2), 101-113.

Xu, L., Tonjes, R., Paila, T., Hansmann, W., Frank, M., & Albrecht, M. (2000, November). DRIVE-ing to the Internet: Dynamic radio for IP services in vehicular environments. *Proceedings of the 25th Annual Conference on Local Computer Networks (LCN'00),* Florida.

KEY TERMS

Digital Video Broadcasting (DVB): A family of standards specifying technology for the global delivery of digital television and data services. The DVB project is an industry-led consortium committed to the standardization process.

European Telecommunications Standards Institute (ETSI): An independent organization, whose aim is to produce telecommunications standards for European and global use. Among ETSI achievements is the deployment of the DVB family of standards.

FFT Mode: The modulation mode which refers to the number of orthogonal subcarriers within an OFDM frame. DVB-T employs two FFT modes (8K, 2K), while DVB-H adds a 4K mode.

MPEG4/H.264: State-of-the-art protocols for digital video compression. They can achieve remarkable video quality even at bit rates of a few hundreds of Kbps.

Multi Protocol Encapsulation (MPE): An adaptation protocol which undertakes the framing and fragmentation of IP datagrams to be injected in MPEG-2 Transport Packets so that they can be conveyed over a DVB platform.

TDM Burst: A block of contiguous data which belongs to the same IP stream and is transmitted within a DVB-H Time Slice.

Ultra High Frequency (UHF): The frequency band between 300 MHz and 3 GHz. The TV-dedicated part of UHF (470-806 MHz) is divided into 8-MHz channels.

Digital Watermarking Techniques

Hsien-Chu Wu
National Taichung Institute of Technology, Taiwan

Hei-Chuan Lin
National Taichung Institute of Technology, Taiwan

INTRODUCTION

In recent years, services on the Internet have greatly improved and are more reliable than before. However, the easy downloads and duplications on the Internet have created a rush of illicit reproductions. Undoubtedly, the rights of ownership are violated and vulnerable to the predators that stalk the Internet. Therefore, protection against these illegal acts has become a mind-boggling issue.

Previously, artists and publishers painstakingly signed or marked their products to prevent illegal use. However with the invention of digital products, protecting rightful ownership has become difficult. Currently, there are two schemes to protect data on the Internet. The first scheme is the traditional **cryptography** where the important data or secret is to be encrypted by a special process before being transmitted on the Internet. This scheme requires much computational process and time to encrypt or decrypt. On the other hand, the second scheme is **steganography** where the important message or secret is hidden in the digital media. The hidden data is not perceptible by the **human visual system** (**HVS**). The digital **watermarking** technique is an application of **steganography** (Chang, Huang, & Chen, 2000; Chen, Chang, & Huang 2001). In order to safeguard copyrights and rightful ownerships, a representative logo or watermark could be hidden in the image or media that is to be protected. The hidden data can be recovered and used as proof of rightful ownership.

The **watermarking** schemes can be grouped into three kinds, largely, dependent on its application. They use the fragile watermark, semi-fragile watermark, and robust watermark, respectively (Fabien, Ross, & Markus, 1999). Fragile watermarks are easily corrupted when the watermarked image is compressed or tampered with. Semi-fragile watermarks can sustain attacks from normal image processing, but are not robust against malicious tampering. Fragile and semi-fragile watermarks are restricted in its use for image authentication and integrity attestation (Fridrich,2002; Fridrich, Memon, & Goljan, 2000). For the **robust watermarking**, it is always applied in ownership verification and copyright protection (Fridrich, Baldoza, & Simard, 1998; Huang, Wang, & Pan, 2002; Lu, Xu, & Sun, 2005; Solanki, Jacobsoen, Madhow, Manjunath, & Chandrasekaran, 2004). Some basic conditions must be followed: (1) Invisibility: the watermarked image must look similar to its original and any difference invisible to the **human visual system**. (2) Undetectable: the watermark embedded in the image must not be easily detectable by computing processes or statistical methods. (3) Safety: watermark is encrypted and if accessed by a hacker; cannot be removed or tampered with. (4) Robustness: the watermark is able to withstand normal and/or illegal manipulations, such as compression, blurring, sharpening, cropping, rotations and more. The retrieved watermark is perceptible even after these processes. (5) Independence: the watermark can be retrieved without the original image. Last but not the least, (6) Efficiency: the watermarked image should not require large storage and must also allow for a comparable-sized watermark to be hidden in the media.

The proposed method is a **VQ-based watermark** technique that depends on the structure of a **tree growth** for grouping the codebook. The scheme is robust. That is, the watermark is irremovable and also can withstand normal compression process, tampering by compression or other malicious attacks. After these attacks, the watermark must be recovered with comparable perceptibility and useful in providing proof of rightful ownerships.

BACKGROUND

The **watermarking** schemes are classified into the methods of the **spatial domain** and the **frequency domain**, respectively (Chang, Huang, & Chen, 2000; Chen, Chang, & Huang, 2001; Zhao, Campisi, & Kundur, 2004). To hide a watermark in the **frequency domain**, an image has to be transformed from a **spatial domain** into its **frequency domain**. This scheme requires many computations and time to embed/retrieve the watermarks. Meanwhile in the **spatial domain**, the watermark can be directly embedded into the pixels values. The algorithms for embedding and recovering are simple. Traditionally, the scheme involves hiding the watermark bits in the **least significant bits** (**LSB**). More literature on the **LSB** technique can be found in Chan and Cheng (2004) and Chang, Hsiao, and Chan (2003). This scheme is not robust. The watermark is easily corrupted after compression.

VQ-Based Watermarking

VQ is a low bit-rate image compression technique. It is simple and easy to encode and decode. Suppose an original image I is partitioned into small non-overlapped blocks with $m \times m$ pixels. Each block contains m^2 pixels. Before **VQ** encoding, block vectors are trained dispersedly and uniformly from several images. The trained set of the block vectors is called codebook. Each block vector in the codebook is called a codeword. In the **VQ** encoding phase, the codeword closest to the encoded block is chosen and stored as an index value in a table. The procedure is repeated for all the blocks in the image. In **VQ** decoding, indexes from the index table will be used to find the corresponding codeword in the codebook and to recover the image (Chang, Huang, & Chen, 2000; Poggi & Ragozini, 2001; Wu & Shih, 2004).

VQ-based watermarking is an effective **steganography** scheme in **spatial domain** and several algorithms have been proposed in different literatures (Wu & Chang, 2004; Huang, Wang, & Pan, 2002). Many of them rely on modifying codewords to achieve the purpose of watermark embedding. But it makes more distortion of images. Lu et al. (2005) used another scheme in illustrated as follows:

1. Input an image X with size $M \times N$, watermark W with size $M_W \times N_W$.
2. In **VQ** encoding process, X is divided into vectors x's with size

   $$\frac{M}{M_W} \times \frac{N}{N_W},$$

   then find the closest to encoded block vectors from the codebook, yield and record index $y(m, n)$ to table Y. Equation is shown as follows:

   $$Y = VQ(X),\ y(m, n) = VQ(x(m, n). \quad (1)$$

3. Compute the variances of $y(m, n)$ of Y after **VQ** encoding vectors by this equation.

   $$\sigma^2(m,n) = \left(\frac{1}{9}\sum_{i=m-1}^{m+1}\sum_{j=n-1}^{n+1} y^2(i,j)\right) - \left(\frac{1}{9}\sum_{i=m-1}^{m+1}\sum_{j=n-1}^{n+1} y(i,j)\right)^2. \quad (2)$$

   And obtain polarites P:

   $$P = \bigcup_{m=0}^{(M/M_W)-1}\ \bigcup_{n=0}^{(N/N_W)-1} p(m,n), \text{ where}$$

   $$p(m,n) = \begin{cases} 1, & if\ \sigma^2(m,n) \geq Threshold \\ 0, & otherwise. \end{cases} \quad (3)$$

   W_P will be the watermark W reordered by random function with $key1$.

   $$key2 = W_p \oplus P \quad (4)$$

   After the **VQ** decoding, the reconstructed image and the secret key will be the protection of the original image.

In Lu et al. (2005), the authors also point to the problems of the algorithm: (1) The watermarks are not really embedded into the image. A secret key is only produced by composing the **VQ** index table and watermark messages, so users can retrieve the watermark from the original image where no other watermark had been embedded in it. (2) If the codebook is public then the user can also embed another watermark in the watermarked image without any modification.

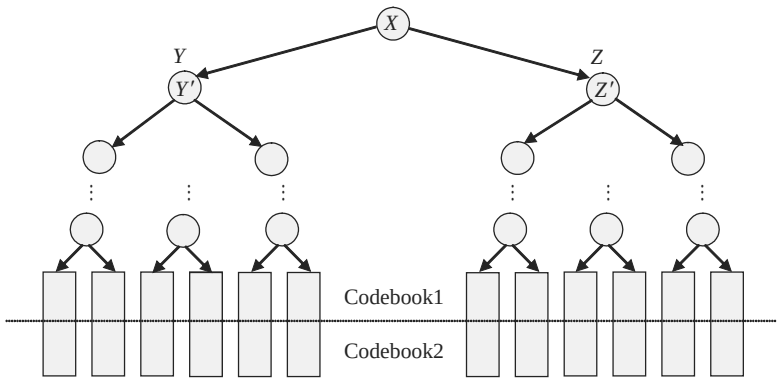
THE PROPOSED SCHEME

The proposed method in this article embeds the watermark bit stream into the **VQ** encoding codes directly without modifying the **VQ** codewords anymore. Meanwhile, the watermarked image quality is controlled by **VQ** compression technique. We can use the existing codebook or produce a new one. In the next section, we demonstrate the proposed codewords grouping method by applying the tree growing structure. The proposed watermark embedding and retrieving processes are illustrated in the following sections.

Grouping Codewords by Proposed Tree Growing Structure

The scheme for grouping the codewords is based on a tree growing structure. First, codewords from a codebook are classified into groups by tree growing structure as described in Algorithm 1. Assume $X(x_1, x_2, ..., x_{m \times m})$, the centroid of the codebook, is the root of tree. Let $Y(y_1, y_2, ..., y_{m \times m})$ and $Z(z_1, z_2, ..., z_{m \times m})$ individually be two nodes of two branches such that $y_i = x_i + k$, $z_i = x_i - k$ and k is a constant. Let each codword belong to the nearest branch node and separate all the codewords into two groups. Computation is repeated to the new centroids Y' and Z', and let Y' and Z' be the new nodes in the sub-tree which will grow the same way as before until members of each sub-tree are equal or less than 2. Each group of codewords should be close to each other and their members are distributed to subcodebook 1 and subcodebook 2. The
steps are described in detail in Algorithm 1.

Figure 1. Grouping codewords by tree growing structure



Algorithm 1: Grouping Codewords by Tree Growing Structure

Input: A codebook C containing codewords $c_1, c_2, ..., c_n$.

Output: A new codebook C' contains 2 subcodebooks C_0 and C_1.

Step 1: Compute the centroid of the codebook, and present it as $X(x_1, x_2, ..., x_{m\times m})$. Also, let X be the root of the tree.

Step 2: Use $X(x_1, x_2, ..., x_{m\times m})$ to grow two vectors Y, Z and satisfy $y_i = x_i + k$, for each $y_i \in Y$, if $y_i > 255$ then $y_i = 255$; $z_i = x_i - k$, for each $z_i \in Z$, if $z_i < 0$ then $z_i = 0$.

Step 3: Classify the vectors belonging to the nearest vectors of the Y and Z groups, and obtain the new centroids Y' and Z'.

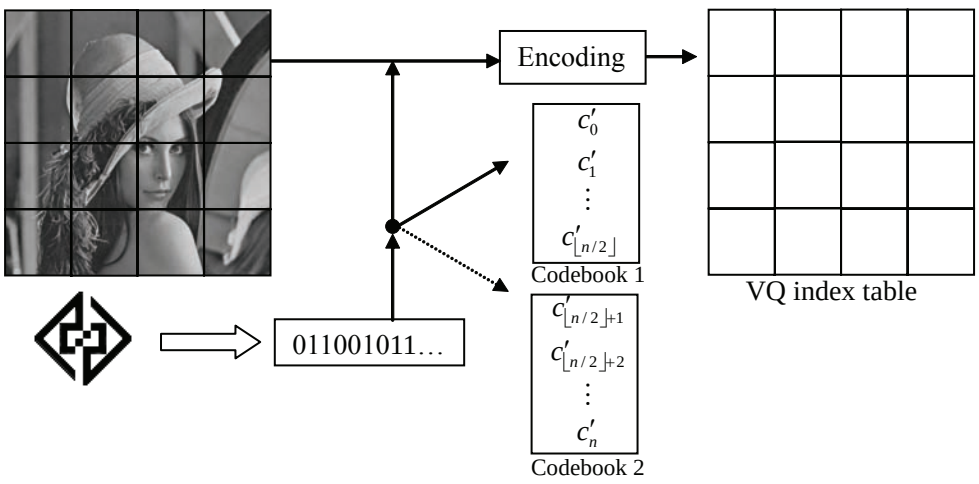
Step 4: Let Y' and Z' be the new root of each sub-tree and repeat **Step 2** and **Step 3** until each group member is equal or less than 2 and very close to each other.

Step 5: Partition the two closest vectors of each group into subcodebook 1 and subcodebook 2. Reindex the two sucodeboks $C_0 = (c'_0, c'_1, \cdots, c'_{\lfloor n/2 \rfloor})$ and $C_1 = (c'_{\lfloor n/2 \rfloor+1}, c'_{\lfloor n/2 \rfloor+2}, \cdots, c'_n)$

Embedding the Watermark

First, the watermark is transformed into a bit stream and let watermark $W = b_1 b_2 ... b_k$. During **VQ** compression, the original image is partitioned into $m\times m$ pixels blocks. The encoding order is then decided randomly to increase the

Figure 2. Embedding a watermark during VQ encoding



security of the embedding watermark. When encoding the first block, and embedding b_1, the closest codeword located in the binary tree is searched from one of the two subcodebooks according to the value of b_1. If the value of b_1 is "0" then search the codeword from C_0. Otherwise, if b_1 is "1" then search the codeword from C_1. The index is then stored into the index table. The process is repeated for the rest of the blocks until all the watermark bits are embedded.

Retrieving the Watermark

When the copyright of an image is in doubt, the embedded watermark can be retrieved and used as proof of its origin. In the retrieval process, the image is read from the index table by the random order using the same seed used in the embedding process. The values "0" or "1" are retrieved for the watermark, *W*, based on whether the index value is smller than $\lfloor n/2 \rfloor$ or otherwise. After all of the watermark bits are retrieved and the watermark is recovered to its original form, the recovered watermark will be used to proof the image's rightful ownership.

EXPERIMENTAL RESULTS

In the experiment, the original image is a grayscale 512×512 image "Lena" as shown in Figure 4(a). The watermark is a black and white 64×64 image as shown in Figure 4(b). Using the proposed **tree growth** scheme, Figure 4(c) shows the watermarked image with PSNR 31.22 dB. By visual

Figure 3. Retrieving watermark during VQ decoding

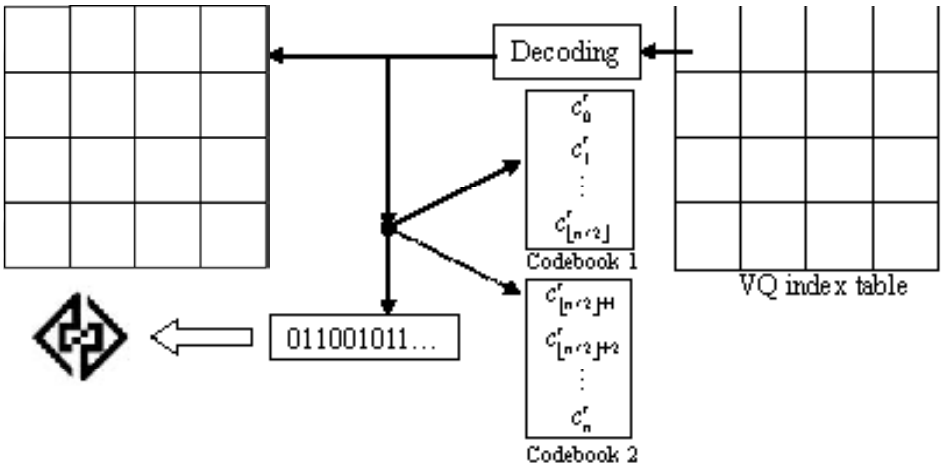


Figure 4. Hiding a watermark in Lena

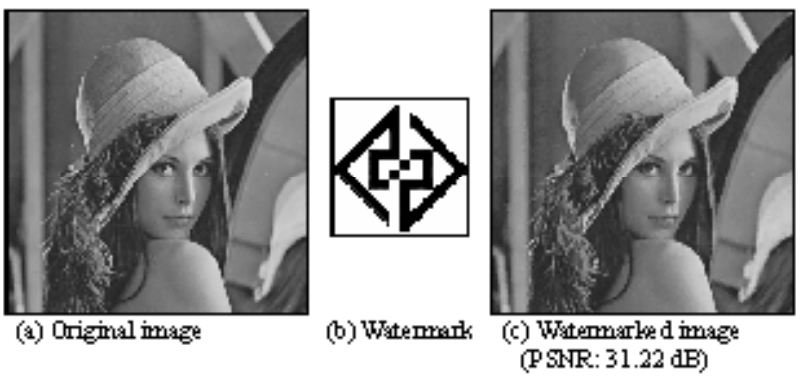
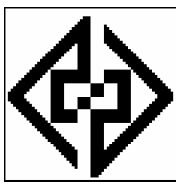


Figure 5. Retrieved watermark with correct bit rate 100%



perception, the watermarked image is similar to its original in Figure 4(a). This means that **tree growth** based hiding did not distort the image.

The retrieved watermark is as shown in Figure 5. Its correct bit rate is 100%, that is, the retrieved watermark is the same as the original watermark in Figure 4(b).

The watermarked images are processed by normal and/or illegal manipulations, such as cropping (shown in Figure 6), blurring (shown in Figure 7), sharpening (shown in Figure 8) and resizing (shown in Figure 9), respectively. The retrieved watermark is perceptible even after these processes. The percentages of retrieved watermarks with correct bit rate are shown in Table 1.

In order to show the **robustness** of the **watermarking** technique, we also do similar experiments using the same watermark and the same size of codebook with 4096 codewords as Lu et al. (2005). The comparative results from the experiments are shown in Table 1. We apply the normalized hamming similarity (NHS) to compute the effectiveness of the twoschemes.

$$NHS = 1 - \frac{H(W, W')}{A_W \times B_W}, \tag{5}$$

where W is defined as a watermark, and W' is the retrieved watermark from the watermarked image that might be tampered.

Comparing with the scheme of Lu et al. (2005), the proposed method can resist against various temper. Some results are better than Lu et al. (2005) and some are not, as shown in the results from the experiments. However, the key point in our method is that the watermark is robustly embedded in the image. A hacker cannot remove the watermark except to destroy the image. The watermarked image could be sent directly on the Internet and does need not an extra key at all. The algorithm and codebook can also be open to the public and the watermark will still be safe. In Lu et al.'s (2005) method, the watermark is not embedded in the image. It only produces a secret key composed of the **VQ** index table and watermark message. The image has to be sent with the secret key. It goes without saying that there would exist two problems. They are: 1) users can retrieve the watermark from the original image without any watermark embedded in it; 2) if the codebook is public then the user can also embed another watermark in the watermarked image without any modification.

Figure 6. The watermarked images are under cropping attacks

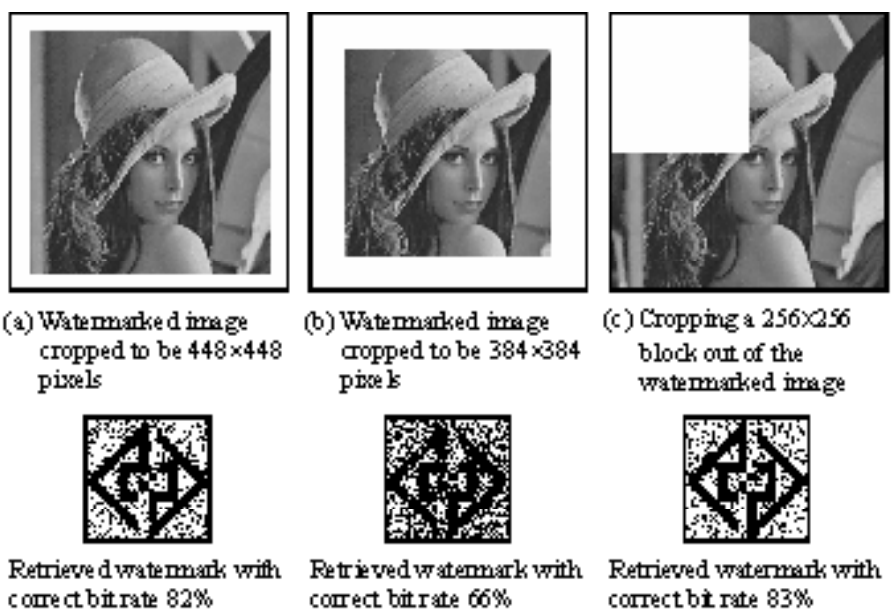


Figure 7. The watermarked images are manipulated by different blurring times

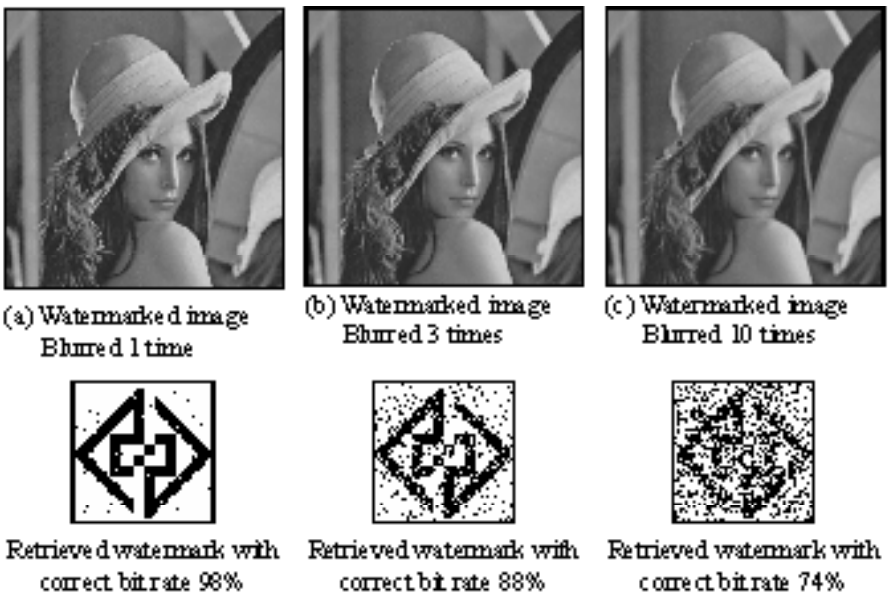


Figure 8. The watermarked image is manipulated by sharpening

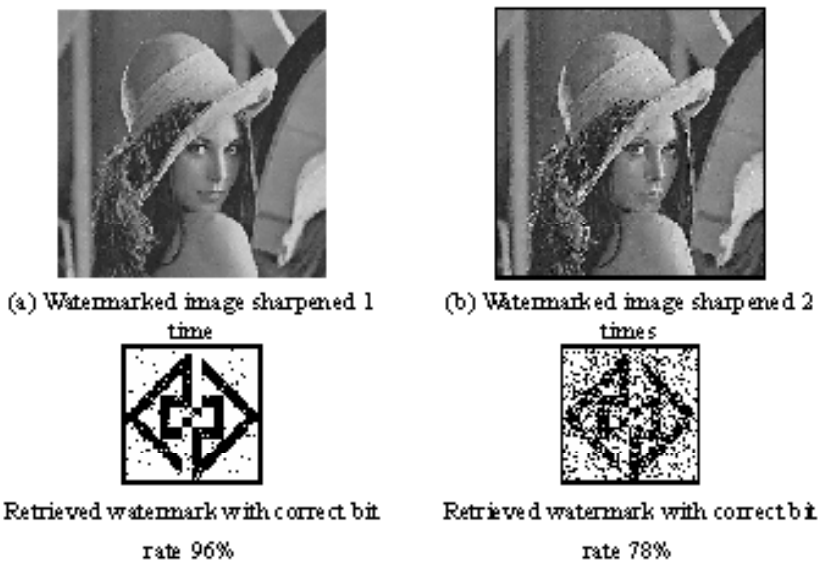


Figure 9. The watermarked images are manipulated by resizing

Table 1. Results for the experiment

Operations	Lu et al.'s method	The proposed method
VQ compression using codebook with 4096 codeword and producing VQ recovered watermarked image (PSNR in dB)	33.74 dB	35.215 dB
Cropping a 256×256 block out of the watermarked image (NHS)	0.887	0.894
Watermarked image blurring (NHS)	0.954	0.973
Watermarked image sharpening (NHS)	-	0.957
Watermarked image contrasting (NHS)	0.926	0.760
Watermarked image brightening (NHS)	0.71	0.747
Watermarked image resized to 256×256 (NHS)	-	0.924
Watermarked image resized to 1024×1024 (NHS)	-	1.000

FUTURE TRENDS

The performance of the proposed **watermarking** technique is stable and reliable in dealing with the ownership disputes of digital media. The capacity for hiding can be further enhanced if the codewords are classified into more groups

CONCLUSION

The proposed method embeds a watermark into an image during **VQ** encoding. The watermark can be retrieved during decoding back correctly. The requirement for space storage is low since **VQ** compression is used. Furthermore, very little computation is needed and time is saved. The proposed method is efficient, robust, and can be used to safeguard rightful ownership. The hidden watermark is safe and not easily detectable by computing processes or statistical methods. Since the watermark is randomly embedded by a secret key, illegal users will have a difficult time hacking even if the watermark algorithm is known. As proven in the experimental results, the watermarked image has a high PSNR of 31.22 dB, is not distorted, and the hidden watermark is confined to invisibility. The retrieval process does not require the original image and the retrieved watermark can be used to prove the rightful ownership of an image.

REFERENCES

Chan, C. K., & Chengm, L. M. (2004). Hiding data in images by Simple LSB substitution. *Pattern Recognition, 37*, 469-474.

Chang, C. C., Hsiao, J. Y., & Chan, C. S. (2003). Finding optimal least significant bit substitution in image hiding by dynamic programming strategy. *Pattern Recognition, 36*, 1583-1595.

Chang, C. C., Huang, K. F., & Chen, T. S. (2000). *Electronic imaging techniques*. Taiwan: Flag Information Co., Ltd.

Chen, T. S., Chang, C. C., & Huang, K. F. (2001). *Digital image processing techniques*. Taiwan: Flag Information Co.

Fabien, A. P., Ross, J. A., & Markus, G. K. (1999). Information hiding: A Survey. *Proceedings of the IEEE Special Issue on Protection of Multimedia Content, 87*(7), 1062-1078.

Fridrich, J. (2002). Security of fragile authentication watermarks with localization, *Proceeding SPIE Photonic West, 4675, Electronic Imaging. Security and Watermarking of Multimedia Contents* (pp. 691-700). San Jose, CA.

Fridrich, J., Baldoza, A. C., & Simard, R. J. (1998). Robust digital watermarking based on key-dependent basis functions. *Proceedings of the 2nd Information Hiding Workshop, Lecture Notes in Computer Science, 1525* (pp. 143-157). New York: Springer-Verlag.

Fridrich, J., Memon N., & Goljan, M. (2000). Further attacks on Yeung-Mintzer fragile watermarking scheme. *Proceedings on SPIE Photonic West, Electronic Imaging, Security and Watermarking of Multimedia Contents* (pp. 428-437). San Jose, CA.

Huang, H. C., Wang, F. H., & Pan, J. S. (2002). A VQ-based robust multi-watermarking algorithm. *IEICE Transactions on Fundamentals, E85-A* (7), 1719-1726.

Lu, Z. M., Xu, G. D., & Sun, H. S. (2005). Multipurpose image watermarking algorithm based on multistage vector quantization. *IEEE Transactions on Image Processing, 14*, 822-831.

Poggi, G., & Ragozini, A.R.P. (2001). Tree-structured product-codebook vector quantization. *Signal Processing: Image Communication, 16*, 421-430.

Solanki, K., Jacobsen, N., Madhow, U., Manjunath, B. S. & Chandrasekaran, S. (2004). Robust image-adaptive data hiding using erasure and error correction. *IEEE Transactions on Image Processing, 13*(12), 1627-1639.

Wu, H. C., & Chang, C. C. (2004). Embedding invisible watermarks into digital images based on side-match vector quantization. *Fundamenta Informaticae,*1001-1017.

Wu, Y. T., & Shih, F. Y. (2004). An adjusted-purpose digital watermarking technique, *Pattern Recognition, 37*, 2349-2359.

Zhao, Y., Campisi, P., & Kundur, D. (2004). Dual domain watermarking for authentication and compression of cultural heritage images. *IEEE Transactions on Image Processing, 13*, 430-449.

KEY TERMS

Image Compression: An image is compressed by VQ compression or other compression techniques to maintain a smaller sized image that requires less storage and, at the same, maintains a perceptible undistorted appearance when in decompressed mode.

Robustness: Refers to sustainability in an image that is not easily distorted by normal or malicious manipulations.

Steganography: Associated to a data hiding technique for hiding an important message or secret data into digital

media. The hidden data is not visible, and the image with the hidden data is called a stego-image.

Tree Growing Structure: The codebook for an image is trained based on the structure of a **tree growth**. Training starts from the root of the tree that grows into two branches and two for each branch and multiple off. The centroid at each node is computed until the tree grows to a certain level where the codewords are the closest that would give the maximal trained codebook.

Vector Quantization: A scheme for image compression that maps blocks of the image to a codebook and its index table.

Watermark: A digital logo, mark, symbol, or data stream that is hidden in an image by a **watermarking** technique to protect the rightful ownership of a media data.

Watermarked Image: An image that has one or more hidden watermarks.

Category: IT Education

Distance Education Initiatives Apart from the PC

José Juan Pazos-Arias
University of Vigo, Spain

Martín López-Nores
University of Vigo, Spain

INTRODUCTION

Developed countries have long been interested in distance education. This interest is growing due to the advance toward a global economy, because education is commonly regarded as the best way to maintain a region's competitiveness. Thus, we have recently witnessed a great development of *e-learning* (taken as a synonym for Web-based learning, or learning through an Internet-enabled computer) to the point that using the Internet to deliver educational material has practically displaced the early initiatives based on postal mail, radio, or television.

The initial evolution of the Internet led to envisaging a massive adoption of e-learning solutions. However, as proved by data from Internet World Stats (http://www.internetworldstats.com), the penetration of the Internet in homes has been rather limited (around 35% in Europe and 67% in the USA), so it follows that the penetration of e-learning has been limited too. This is indeed one consequence of the so-called *digital divide*, that is, the separation between people who make frequent use of the information technologies and those who have no access to them or, even having access, lack the necessary knowledge to use them.

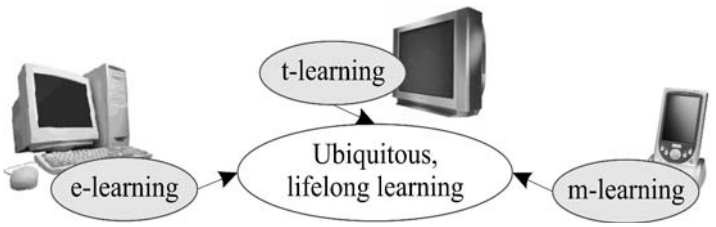
A divide in the access to technology can lead to inequalities in the access to knowledge and education, posing risks of social exclusion. To prevent that, public administrations have launched large-scale initiatives, like the *World Summit on the Information Society* and the *i2010 plan*, that aim at making technology available to everyone, at anytime and from anywhere. As a cornerstone, these initiatives promote the development of access platforms different from the PC, with special interest in harnessing the interactive features of devices that have attained greater penetration in society. This includes the new digital TV set-top boxes, which bear the term *t-learning*, and the modern mobile devices (e.g., mobile telephones and media players), which set the foundations for *m-learning*. The vision, as represented in Figure 1, is that the information technologies, combined with suitable pedagogical and andragogical approaches, will enable a scenario of ubiquitous and lifelong learning, freeing people from time and place constraints, and offering flexible learning opportunities to individuals and groups.

This article describes technical, methodological, and educational issues that make t-learning and m-learning substantially different from previous works on e-learning. We also review developments in both areas to finally discuss problems that may be the subject of much research in the near future.

BACKGROUND

Next, we discuss the evolution of the motivations and scopes of t-learning and m-learning. It will be seen that many pieces

Figure 1. Convergence of approaches to distance learning



of work have had disparate views of the learning paradigms that should be pursued over the new technological media.

Motivations and Scope of T-Learning

Television has been present in nearly every home for decades, getting to be so familiar that everyone feels comfortable using it. Thus, it may become an entry point to the information society, especially for the social sectors that are more reluctant to have contact with technology. Actually, the idea of using the television for educational purposes dates back to the 1950s, but its potential to disseminate knowledge remained underexploited. However, the advent of interactive Digital TV (IDTV) by the late 1990s opened an unprecedented range of possibilities, which made Lytras, Lougos, Chozos, and Pouloudi (2002) predict that "the possibility of broadcasting data and interactive applications jointly with the audiovisual contents will have far-reaching implications in education." T-learning then arose as an approach to exploiting the advances in IDTV technologies to deliver interactive applications that would promote learning and problem solving by requiring active involvement from the viewers (Zhao, 2002).

At first, there were discrepancies regarding the very conception of t-learning, with two opposite perspectives.

- On the one hand, some authors (Russell et al., 2004) argued for simply providing an interface to the same e-learning services running on the Internet, porting existing solutions to a new execution platform (the IDTV set-top box).
- On the other hand, as noticed in Pazos-Arias et al. (2006), most of the experiences launched up to 2005 merely consisted of adding interaction capabilities to the TV programs, promoting the concept of "edutainment" (education and entertainment).

The first approach has been practically abandoned for neglecting various IDTV-specific features that advise against making t-learning a direct translation of the models devised for e-learning (Lekakos & Chorianopoulos, 2006). First, there are technical factors like the limited interactivity achievable with a remote control, the reduced amount of text that can be readable on a TV screen, or the low computing power of a set-top box. Furthermore, it is clear that many potential IDTV users have a lower level of predisposition to learn new technologies than Internet users. Finally, the many years of analogue TV have consolidated a passive attitude from the users, plus a conception of television as an entertainment medium. Thereby, as claimed by the second approach, it is now accepted that an effective t-learning strategy should lean on entertainment to lure people into education, and deliver interactive applications that guide the users through audio and video contents (Chorianopoulos & Lekakos, 2007; Trindade, do Vale, & Pedroso, 2006). Notwithstanding, the kind of edutainment envisaged at first has evolved into two distinct philosophies in the design of t-learning services, which may be seen as the reverse of one another.

- By *pure edutainment* (meaning education that entertains), we now refer to educational services whose central axis is a TV program, enhanced with interactive learning elements that furnish pedagogical added value.
- The term *entercation* (entertainment that educates) refers to educational services designed around an interactive learning element that is supplemented with audiovisual material for the sake of amusement.

In sum, we can delimit the scope of t-learning halfway between the mere entertainment provided by the TV programs and the formalities of e-learning. Interactivity provides a major advantage with regard to the traditional TV programs because it makes the learning experience more engaging, for example, by letting the user influence the presentation of contents, evaluate his or her knowledge through assessment tests, participate in competitions synchronized with TV shows either individually or as representative of a group (Sperring & Strandvall, 2006), and so forth.

Motivations and History of M-Learning

Grounded on the requirement to have the learners in front of the PC, the e-learning models cannot meet the requirements of the modern lifestyle. M-learning arose to embrace the initiatives that harness the educational possibilities of mobile devices that have attained massive penetration in society, like mobile telephones, media players, and PDAs (personal digital assistants). The vision is common that mobile computing will enable the greatest level of time and space flexibility, together with unprecedented possibilities to adapt to individual learners' needs. However, this idea has borne disparate realizations.

The first m-learning developments focused on providing access through mobile telephones to existing e-learning platforms using technologies like wireless application protocol (WAP) to browse hypermedia and short messaging system (SMS) to deliver textual notifications (Garner, Francis, & Wales, 2002), and using specialized navigators and e-mail (Savill-Smith & Kent, 2003). In general, there were no attempts to think of m-learning as a distinctive approach, and research focused on technical issues like delivering content over wireless networks (Vedula & Han, 2003), adapting content and interfaces to make the same courses accessible through different devices (Bandelloni and Paternò, 2004), or maintaining the learning activities during periods of disconnection (Trifonova & Ronchetti, 2005).

In a posterior stage, many claims arose from the educational community to develop m-learning more from the

pedagogical and andragogical perspective, with a focus on exploiting the highly fragmented attention of users on the move. This philosophy led to guidelines on how to create and organize content (Vavoula, 2004): structuring content in short learning modules (from 30 seconds to 10 minutes), supporting nonlinear follow-ups, and so on. Within these guidelines, we witnessed a proliferation of experiments that assessed m-learning in a variety of contexts (Kukulska-Hulme & Traxler, 2005), and its scope grew to embrace both formal and informal, intentional and unintentional learning activities: accessing academic courses as above (Alexander, 2004), downloading localized information during museum visits (Papadimitriou, Komis, Tselios, & Avouris, 2006), and taking pictures to aid retention of knowledge from a seminar or surgical operation (Parks & Dransfield, 2006), among others. Accordingly, the notion of learning was broadened to include "the needs that emerge when one strives to overcome any problem in everyday activity," and the range of devices for m-learning was once and for all extended from just mobile phones to "any device that is small, autonomous and unobtrusive enough to accompany us in every moment in everyday life, and that can be used for some form of learning: digital cameras, smart phones, MP3 players, etc" (Trifonova & Ronchetti, 2005). This enumeration should now be expanded to leave place for ongoing advances in wearable computers.

As the major distinctive feature of m-learning, many authors are working on mechanisms to provide contextualized learning experiences, i.e., to take into account where the user is and what he/she is doing in order to recommend the best suited educational activity. In this regard, we can cite studies to manage context using geographical positioning (Ogata, 2006) and radio-frequency identifiers (Baggetun & Wasson, 2006). Also, terms like *adaptivity* and *context awareness* already appear regularly in the m-learning literature, but there is yet little evidence that m-learning applications harness these capabilities or fulfill novel pedagogical aspirations.

THE CURRENT SCENARIO OF DISTANCE LEARNING

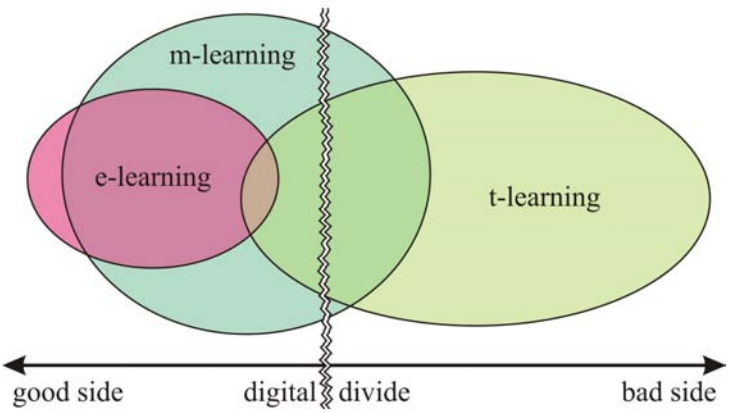
The works cited in the preceding section, together with many others in literature, already provide sufficient practical feedback to consolidate the scopes of both t-learning and m-learning with regard to e-learning. The relations can be represented as in Figure 2.

To begin with, it is noticeable that e-learning is mostly circumscribed so as to enhance the training of people on the good side of the digital divide since it is mostly sustained by (a) companies looking for a quick and affordable way to keep employees educated on the latest developments in their fields, and (b) universities offering certificate programs via the Internet. Accordingly, e-learning has consolidated as a suitable approach for formal and intentional educational settings, targeting people who engage purposefully in well-defined methodologies to achieve certain curricula. The role played by these learners is active, promoting the design of user-driven interaction schemes wherein the applications respond to the users' actions.

Due to the usage models linked to the PC, e-learning is unsuitable for informal and accidental types of learning, which may be defined as:

the process whereby individuals, often in an unintended or unexpected manner, acquire attitudes, skills and knowledge from daily experience and the educative influences and resources in their environment, from family and neighbors, from work and play, from the market place, the library and the mass media. (Conner, 2005)

Figure 2. The relative scopes of e-learning, t-learning, and m-learning



This definition supposedly applies to over 75% of adult learning, and obviously to an even greater share in children (Organization for Economic Cooperation and Development [OECD], 2006). T-learning is called to fill this huge gap, especially targeting the people who have not been involved with formal or semiformal education for many years, and those who have a poor technological background. Interactions in this case are predominantly *media driven* to let the users take part in a reduced number of decisions. Nevertheless, t-learning is also available for people on the good side of the divide, who may select a formal or informal learning option depending on their interests in individual topics.

Finally, m-learning may be regarded as a medium to provide increased learning opportunities to people with a solid technological background, plus the so-called *digital natives*, that is, people who have grown up accustomed to using electronic widgets (Prensky, 2001). Targeting this audience, m-learning covers the wide spectrum of applications in distance education opened by the latest pilots and trials, to such an extent that it could be perceived as a superset of e-learning. The only e-learning services that fall out of the scope of m-learning are those for which the educational material cannot be adequately perceived and/or handled with a mobile device due to limitations in screen size and/or input facilities. Undoubtedly, though, the most distinctive feature of m-learning lies within its possibilities to furnish a *context-driven* approach to interactivity, capable of adapting the learning experiences to the user's evolving attention, skills, knowledge, and location.

Having said that, the goal of achieving convergence of the three distance learning paradigms (as depicted in Figure 1) is still hampered by incompatibilities among the different technologies employed for learning management tasks: student monitoring, content description, organization and retrieval, and so forth. The reason is that the efforts in the standardization of distance learning have been mostly focused on e-learning. However, recent studies have shown that many solutions originally devised for e-learning, specifically, those bundled in the *sharable content object reference model* (SCORM) (ADL, 2004), can be successfully applied to t-learning and m-learning as well, requiring minor adaptations or extensions to cater for their peculiarities (Low, 2007; Pazos-Arias et al., 2006). These works reveal that it is possible to build a full-featured and consistent whole reusing existing solutions.

FUTURE TRENDS

One of the topics that will presumably be the subject of much research in the following years is that of *personalization*, aimed at enabling learning experiences suited to the interests, needs, and capabilities of the different users. The SCORM standards, together with domain-specific norms, provide the foundations to develop intelligent tutoring systems that match the descriptions of the educational resources available, records of TV watching habits, learning histories, and so on to discover the most suitable educational resources for each individual. In this regard, as suggested by Blanco-Fernández, Pazos-Arias, Gil-Solla, Ramos-Cabrer, & López-Nores (2007), *semantic reasoning* techniques have the potential to deliver the most engaging and effective experiences, but their application in this area is still incipient. The same happens with research on the automatic assembly of services tailored to users' interests, which aims at boosting reusability to lower the costs of producing educational material. Finally, closely related to personalization, there are still more open problems than working solutions in adapting the learning experiences to the user's context and changing attention. This issue is mostly linked to m-learning, but also touches t-learning to a remarkable extent, for example, to differentiate when the user is watching TV alone or accompanied, or to change the learning material offered as the user zaps different channels. Advances in these topics require contributions from such fields as electronics, artificial intelligence, software engineering, human-computer interaction, and even human physiology to bring distance learning progressively closer to emerging areas like *ambient intelligence* (Remagnino, Foresti, & Ellis, 2005) and *ubiquitous computing* (van't Hooft & Swan, 2006).

CONCLUSION

During the last decade, we have witnessed a great development of e-learning solutions, taken as a synonym for learning through an Internet-enabled personal computer. However, there is growing evidence that e-learning fails to reach sizable social sectors, especially people who are reluctant to have contact with the new technologies and those who have not been involved with educational activities for many years. To narrow the effects of the digital divide, the research community is devoting significant effort to develop alternative learning platforms that solve the accessibility, usability, and availability limitations of the PC. The bet is placed on exploiting the interactive features of digital TV receivers and the modern mobile devices. We have seen that these initiatives serve to fill in different gaps left uncovered by e-learning, and that, overall, the goal of making learning opportunities available at any time, from any place and for everyone is progressively becoming a reality.

REFERENCES

ADL. (2004). *Sharable content object reference model.* Retrieved December 24, 2007, from http://www.adlnet.org

Alexander, B. (2004). Going nomadic: Mobile learning in higher education. *EDUCAUSE Review, 39*(5), 28-35.

Baggetun, R., & Wasson, B. (2006). *MOTEL: A location-based framework for virtual geo-tagging in higher education.* Paper presented at the Fifth World Conference on M-Learning, Banff, Canada.

Bandelloni, R., & Paternò, F. (2004). Migratory user interfaces able to adapt to various interaction platforms. *International Journal of Human Computer Studies, 60*(5), 621-639.

Blanco-Fernández, Y., Pazos-Arias, J., Gil-Solla, A., Ramos-Cabrer, M., & López-Nores, M. (2007). *Overcoming weaknesses of current personalization techniques by Semantic Web technologies.* Paper presented at the Second International Conference on Metadata and Semantics Research, Corfu, Greece.

Chorianopoulos, K., & Lekakos, G. (2007). Learn and play with interactive TV. *ACM Computers in Entertainment, 5*(2), 1544-1574.

Conner, M. (2005). *Informal learning.* Retrieved December 24, 2007, from http://agelesslearner.com/intros/informal.html

Garner, I., Francis, J., & Wales, K. (2002). *An evaluation of the implementation of a short messaging system (SMS) to support undergraduate students.* Paper presented at the European Workshop on Mobile and Contextual Learning, Birmingham, United Kingdom.

Kukulska-Hulme, A., & Traxler, J. (2005). *Mobile learning: A handbook for educators and trainers.* New York: Routledge.

Lekakos, G., & Chorianopoulos, K. (2006). *Interactive digital TV: Technologies and applications.* Hershey, PA: IDEA Group.

Low, L. (2007). *M-learning standards report.* Retrieved December 24, 2007, from http://e-standards.flexiblelearning.net.au/

Lytras, M., Lougos, C., Chozos, P., & Pouloudi, A. (2002). *Interactive television and e-learning convergence: Examining the potential of t-learning.* Paper presented at the European Conference on E-Learning, Uxbridge, United Kingdom.

Ogata, H. (2006). *Supporting ubiquitous language learning by linking RFID tags and videos.* Paper presented at the Fifth World Conference on M-Learning, Banff, Canada.

Organization for Economic Cooperation and Development (OECD). (2006). *Recognition of non-formal and informal learning.* Retrieved December 24, 2007, from http://www.oecd.org

Papadimitriou, I., Komis, V., Tselios, N., & Avouris, N. (2006). *Designing PDA-mediated educational activities for a museum visit.* Paper presented at the International Conference on Cognition and Exploratory Learning in Digital Age, Barcelona, Spain.

Parks, M., & Dransfield, M. (2006). *Mo-blogging: Supporting student learning whilst in health care practice settings.* Paper presented at the Fifth World Conference on M-Learning, Banff, Canada.

Pazos-Arias, J., López-Nores, M., García-Duque, J., Gil-Solla, A., Ramos-Cabrer, M., Blanco-Fernández, Y., et al. (2006). ATLAS: A framework to provide multiuser and distributed t-learning services over MHP. *Software: Practice and Experience, 36*(8), 845-869.

Prensky, M. (2001). Digital natives, digital immigrants. *On the Horizon, 9*(5).

Remagnino, P., Foresti, G., & Ellis, T. (2005). *Ambient intelligence: A novel paradigm.* Berlin, Germany: Springer.

Russell, T., Varga-Atkins, T., & Smith, S. (2004). *Enticing learners: Rethinking the relationship between e-learning via DiTV and via the Internet*. Paper presented at the Second European Conference on Interactive Television, Brighton, UK.

Savill-Smith, C., & Kent, P. (2003). *The use of palmtop computers for learning: A review of the literature.* Retrieved December 24, 2007, from http://www.m-learning.org/knowledge-centre/m-learning-research.htm

Sperring, S., & Strandvall, T. (2006). *Viewers' experiences of a TV quiz show with integrated interactivity.* Paper presented at the Fourth European Conference on Interactive Television, Athens, Greece.

Trifonova, A., & Ronchetti, M. (2005). *Hoarding content in m-learning context.* Paper presented at the Third IEEE Conference on Pervasive Computing and Communications, HI.

Trindade, D., do Vale, D., & Pedroso, L. (2006). *Digital TV and distance learning: Potentials and limitations.* Paper presented at the 36th ASEE/IEEE Frontiers in Education Conference, San Diego, CA.

TV-Anytime Forum. (2003). *TV-Anytime specification series* (ETSI standard TS 102 822).

van't Hooft, M., & Swan, K. (Eds.). (2006). *Ubiquitous computing in education: Invisible technology, visible impact.* Mahwah, NJ: Lawrence Erlbaum Associates.

Vavoula, G. (2004). *KLeOS: A knowledge and learning organisation system in support of lifelong learning.* Unpublished doctoral dissertation, University of California.

Vedula, I., & Han, R. (2003). *A distributed software system architecture for wireless peer-to-peer collaborative learning.* Paper presented at the Third IEEE International Conference on Advanced Learning Technologies, Athens, Greece.

Zhao, L. (2002). *Interactive television in distance education: Benefits and compromises.* Paper presented at the International Symposium on Technology and Society, Raleigh, NC.

KEY TERMS

Context-Driven Interaction: It is an approach to the design of interactive services that adapts their content and functionality to users' evolving context, including attention, interests, needs, and location.

Digital Learning Divide: This is a social phenomenon caused by inequalities in the access to knowledge and education that arise from familiarity with the information and communication technologies.

Edutainment: It is a pedagogical and andragogical orientation based on developing services around entertaining material, enhanced with learning elements that furnish educational added value.

Entercation: It is a pedagogical and andragogical orientation based on developing services around a learning element that is supplemented with audiovisual material for the sake of amusement.

Media-Driven Interaction: This is an approach to the design of interactive services that caters for passive consumption habits, letting the applications guide the users through audiovisual contents.

M-Learning: M-learning is a distance learning approach based on delivering interactive and possibly contextualized educational material to mobile devices.

T-Learning: T-learning is a distance learning approach based on accessing interactive educational material of a predominantly audiovisual nature, primarily within the home and using devices that bear usage habits similar to those of watching television.

Category: IT Education

Distance Education Teaching Methods in Childcare Management

Andreas Wiesner-Steiner
Berlin School of Economics, Germany

Heike Wiesner
Berlin School of Economics, Germany

Petra Luck
Liverpool Hope University, UK

INTRODUCTION

The cultural and technical history of e-learning scenarios can be traced back to traditional forms of distance studies, CD-Rom learning programmes, audio-programmes or educational TV. But other than these forerunners, two closely related myths often shape policy towards ICT and education: the irresistible power of globalisation and the determining effect of technology. Both views present the success of e-learning throughout the education system as inevitable. The space left for practitioners in higher education is either to embrace the new media or to watch its inevitable unfolding. In this paper we take a critical stance towards that perspective and suggest that the shape and learning effect of new media in higher education is contested and evolves in communities of practice. No technologies are neutral and it is more appropriate to speak of economic, technological and societal features as interactively fostering the importance of e-learning through *distributed actions* (Rammert, 2002). From such a perspective, e-learning is perceived as a co-product of didactically and technically situated features (Wiesner-Steiner, Wiesner, & Schelhowe, 2006) that foster and enable but don`t determine human learning through the use of digital technologies. Main characteristics are:

- Interactive and multimedial design of content
- Learning via digital networks
- Netbased communication

The EU-Leonardo-project "European Enhancement of Early Years Management Skills—EEEYMS" (http://www.eeeyms.org/) was intended to enhance employability of people employed in the Early Years Childcare management sector by providing access to a high level qualification in line with the emerging industry requirements. This was achieved by developing distance learning materials available via the World Wide Web and other forms of media including CD-Rom`s, specific to the employment area which is also aligned to a degree pathway, and will be available within Europe. It was further achieved by the creation of a European network association for childcare to ensure sustainability after the project is complete. EEEYMS provides an accredited route for the attainment of a relevant degree level qualification for careers and managers within the childcare sector, and assist in attracting suitable people into this employment sector to meet the childcare demand over the next 10 years. With ODL materials, the project enhances employment opportunities and career status for a still predominantly female workforce. Research suggests that the increased status and professionalisation obtained through the availability of a high level qualification will make the industry more attractive to male employees. EEEYMS thus provided higher level qualification to people disadvantaged in the labour market and those who faced discrimination in accessing training due to disability, geographical location or family commitments. The use of ICT systems was thus thought to enhance knowledge and learning experience *and* the employability factors, as the knowledge will be directly transferable to the work environment.

The primary target group was that of childcare professionals actively working in the sector or entering this profession, where a niche in the market exists for a relevant specific degree award. EEEYMS thus wanted to attract more women into managerial positions, while encouraging more men to enter the profession by providing a credible award.

Because empirical evidence on the increase of e-learning-efficiency is both difficult and important, external evaluation of the EEEYMS e-learning modules via surveys has been an integral part of the entire project. The aim here was to include a more objective, independent feedback at every stage of the programme. According to the projects aims, the evaluation was conducted following the principle of gender mainstreaming (Wiesner, Kamphans, Schelhowe, Metz-Göckel, Zorn, Drag, Peter, & Schottmüller, 2004) and considering intercultural inclusion-aspects (Zorn & Wiesner-Steiner, 2006).

The article is divided into three main sections. After introducing the use of VLE and a problem based learning

approach, we discuss the effects of group work, the use of technology and the main learning experiences. As a result we come up with an overview of critical sociotechnical issues of distance learning materials.

BACKGROUND

In the development of e-learning for the early years sector through the EEEYMS partnership these key issues emerged: the importance of the use of a suitable VLE in delivering the learning programme, the use of problem-based learning (PBL) to enhance student motivation through collaboration, the need of IT skills development and the role of context as it relates to student success.

The VLE in use is Granada's "Learnwise". This VLE has as one of its technical features collaborative "forums" in which participants take part in asynchronous discussion in small teams and work on specific management and education problems The partnership decided that these forums would provide a prime vehicle for student support through "encouraging active learning", shifting from didactic to "facilitative teaching" or "building online communities" (Armitage, Brown, & Jenkins, 2001)

The stated aim of the EEEYMS project is that early years practitioners will develop knowledge and understanding of the educational and management issues pertinent to their sector, and that they will also develop the requisite skills to critically analyse, evaluate and apply this knowledge. As professional knowledge requires functioning knowledge that can be put to work immediately, most module designers for EEEYMS choose to adopt a "problem-based learning" approach.

Problem based learning simulates everyday learning and problem solving. Knowledge is acquired in a working context and is put back to use in that context. The learning and assessment on the programme will be aligned (Biggs, 1999) to learners everyday work experiences. Participants learn the skills for seeking out the required knowledge when the occasion arises during the process. They are motivated immediately by the interaction with a 'real' problem and are active early in the process.

Although on-line participants face time constraints as working practitioners and as parents with family responsibilities, the use of media-communicated communication has been used to build successful collaborative learning. As Salmon (2000) asserts, the Internet can change concepts of space and time: "*Working and learning with others who happen to live in a particular locale may become less important than finding shared professional and personal interests in online environment" (p. 492).*

The EEEYMS project aimed to provide learning opportunities at degree level, so that practitioners can develop the requisite skills to critically analyse, evaluate and apply knowledge. A large body of literature support the motivational aspects of collaboration on learning (Johnson & Johnson, 1989; Sharan & Shaulov, 1990). Wenger (1999) also offers a perspective on learning that emphasises social learning processes within *communities of practice* where individuals engage in the negotiation of meaning and the mutual construction of knowledge. The EEEYMS participants often refer to this "community of practice" when expressing the relevance of the tasks to the everyday practice.

The issue of gender was also pertinent as with the exception of one male EEEYMS participant, all others were female. For example, a study by Kirkup and von Prumm (1990) comparing the experiences of women adult distance learners in Germany and the UK points to a pattern of preference for shared learning.

This type of social-technical interaction, learning and decision making is expected in the workplace today and this approach should ultimately therefore promote a desire for and ability to partake in 'life long learning'.

Meisalo, Lavonen, and Juuti (2005) also emphasise the importance of Web based community formation for off-campus participants in their study of primary teachers taking a science education course. Dron (2005) in his paper on the construction of e-learning environments to cater for the needs of diverse learners utilises Michael Moore's theory of transactional analysis. For Moore (1980), distance is a pedagogical more than a physical phenomenon, and transactional distance measures the amount and nature of dialogue. Transactional distance is said to be low when there is a lot of dialogue between learners and teacher, but where transactional distance is high, teachers often provide a highly structured learning experience. The use of PBL appears to ensure that student autonomy flourishes and dialogue is high not only between student and teacher, but also student and student.

This importance of web based community and the need to maintain a low transactional distance through constant dialogue appears to be a critical outcome of the EEEYMS project. Donohue (2002) analyses the challenges of teaching the target group for EEEYMS online, as the Early Years sector is characterised by "low tech/ high touch". While many Early Years Managers and Practitioners might only have little involvement with high tech equipment such as computers in their work place settings, much of their practice is concerned with managing relationships with colleagues, children and families. Donohue (2002) suggests the use of learning approaches aiding the building of a community of practitioners such as collaborative knowledge construction and group work. The evaluation results discussed now show that it has successfully utilised learning approaches to mirror that "high touch."

PARTICIPANT EXPERIENCES WITH GROUP WORK AND E-LEARINNG TECHNOLOGY

In our view, interlinking the scopes of didactics, evaluation and technology can help to increase the user's (long dated) commitment to e-learning modules. If electronic learning tools are perceived as "didactical actors" that not only bear their own action potential but influence and redirect participants belief systems and agency (Wiesner-Steiner, Wiesner & Schelhowe, 2006), new relations between learners and *didactical technology* come into focus. The results of the external evaluation thus mark important "passage points" for technical *and* didactical implementations of e- learning modules.

Methodically we used semi-standardized questionnaires that consist of a combination of closed yet multiple choice questions and open questions that leave room for participants to explain their more subjective learning experiences. Interpretation was done by means of the content analysis (Gläser & Laudel, 2004; Mayring, 2000).

GENERAL PARTICIPANT EXPERIENCES

Asking for general experiences, we could identify three points:

- Time management, intensity of tasks, and work amount mark the most difficult aspects.
- Although quite challenging, the modules "were well designed" and "offered something for everybody"—learning outcomes were met throughout all the modules.
- Group work and tutors play a crucial and positive role in all modules.

GROUP WORK

Throughout our evaluation of EEEYMS, the overall importance of group work in a VLE became evident. Group work activities are not just a work form among others. For e-learning modules, they proved to be the work form par excellence! Group work was not only generally important, it does enhance group commitments in virtual communities. Two of the most stunning *social* remarks about electronic group work thus stated: *"Without group work I don`t think I would have managed the course."* (...) *"I feel as if I know my group members better without having the physical get-togethers than in previous studies."* Accordingly, students also pointed to the advantage of a shared workload, especially for part-time students or people who have to work full time and have a family life. They also mentioned that informal phone and e-mail contacts have been used in group work more often than informal chat rooms. Moreover, they found that e-learning group work offers different perspectives on an issue, allowing for a more holistic image and approach of the tasks. The relevance of group work can be summed up as follows:

- Electronic group work needs blended learning with face-to-face meetings at the beginning, "everybody had a face".
- Being part of the same group in different modules helps for getting used to different learning styles.
- Changing groups can constrain learning processes.

Both the combination of face-to-face with electronic group communication and the importance of group continuity mark important points for the appropriation of different learning styles and the development of social commitment. Didactically applied, they can improve the general learning experiences mentioned previously.

Although most EEEYMS students agree that group work plays a very important role in e-learning modules, they also addressed some risks:

- "The team leader's position is sometimes confusing because of individual's different aims—sometimes ignoring other needs."
- "You have to monitor yourself all the time in order to try to avoid communication misunderstandings."

Group members have to learn to work in a team. But as we know from everyday life and work practices, teams can be organized in more hierarchical or more symmetrical ways, depending on the members and social dynamics of a certain group as well as on the specific tasks and contexts. Within e-learning environments, students have to organize their group work mostly on their own. Moreover, they have to deal with social aspects like leadership and communication. Participants also found that group work is helpful not only for dealing with certain tasks but to navigate through both the technical and learning requirements of a module. Due to that, group work also functions as a method to downsize the drop-out quote in two directions—dropping out because of difficult (social) tasks or dropping out because of technical problems. But do the online students use the communication offers given by Learnwise? Studies in the area of e-learning and knowledge management systems conclude that communication offers are not used very often if they are not designed in a didactically carefully fashion. This is clearly perceived by EEEYMS participants as one of the most positive aspects of the e-learning modules. Thus group work—experienced as a new way of collaborative learning and in combination

with team support by tutors—was perceived as *the* most positive aspect. Nonetheless, group work proved both important and challenging, creating mutual dependencies as well as commitments. It is important to note that both aspects can be perceived as very positive, depending on the student's experiences with team support and technical support.

Group work was also useful in cases when technology did not work as expected, trying to solve or bypass technical problems as a group. Because group work also works in informal, non-electronic ways, an important point stressed by the students is the unique form of electronic group work. Electronic group work is seen as a learning process in itself, requiring both commitment and easily accessible technology. In sum, the importance of electronic group work for learning processes was highlighted for all Hope modules, not the least because group work and electronic learning tools build learning communities and communities of practice (Lave & Wenger, 1991). We thus recommend that the concept of technically mediated group work should mark one of the central aspects for future e-learning modules and should be integrated into didactical approaches.

TECHNOLOGY

In accordance with the role of group work, we also evaluated the role of technology in e-learning environments:

- Electronic systems are great as long as they are running.
- E-sources where sometimes difficult to find and time consuming.
- Forums and chat rooms for group work were useful for communication among group members and contacts with tutors.
- Online sources were most useful in combination with supplementary materials (CD-Rom), group work and tutorial help.
- Tutors play an important role as they mediate between the requirements, technical possibilities and social dynamics of the e-learning modules.
- Tutors were given excellent credits, have been responsible "even at simple questions", accessible most of the time and (often) replied promptly.

For EEEYMS participants, websites and journals became high-rated whereas library resources seem to be "out". Nonetheless, in terms of provided materials the students often felt that hard-copy handouts gave more safety than other material. This has to do with technical problems that are perceived as very time-consuming, i.e. access-problems that occurred while trying to use e-journals, informal chat rooms with slow responding action or Web links with passwords (with the exception of tutor-directed Web links). CD-ROMs and module handbooks thus became important when access to those materials failed. Moreover, they were also important in the modules`introductory phases. In addition, the role of the tutor became quite important in cases when technology failed, forcing tutors to organize a new or alternative learning environment. The facilitation and encouragement of electronic communication by tutors marks another important evaluation point. We thus recommend that for future e-learning modules, tutors should be especially trained in supporting online communities and group activities.

LEARNING

Not only did EEEYMS participants learn something new, but they became able to translate their new knowledge into their own professional contexts. In addition, the following features of their collective learning experiences point to the close interplay between didactical, social and technical e-learning issues:

- Discovering the unique style of e-learning;
- Discovering group work as a learning experience;
- Learning of IT skills and time management;
- Discovering the advantages of problem based learning;
- Discovering the possibility to study while working and being "old age";
- Discovering the possibility to choose one's own learning time (look at resources, listen to CD lectures, etc.);
- Discovering many ways to act towards the same aims;
- Discovering appropriate and comprehensive modules for day care managers";
- Discovering excellent experiences in communicating and learning with colleagues from different countries; and
- Discovering that children have the same basic demands and affairs in spite of cultural differences.

FUTURE TRENDS

In summary, e-learning is clearly perceived as a new type of learning. Nonetheless, we might only use 20% of the possibilities of e-learning. To some extent, that mirrors the development of television where at the beginning the actors did act like actors in a stage play and were not aware of the new technology and its influence on their performance. The same thing could be true for e-learning, when virtual group work and electronic learning tools are offered and mediated in a didactically careful fashion, creating new forms of lifelong learning. As our informants mentioned, this process can be initiated and improved by

Table 1. A summary of critical issues for future distance learning materials

• Time management, intensity of tasks and social context (work amount, family) create learning problems. • Integration of working duties and job related perspectives into the content of teaching and into the structure of e-learning modules is important (e-learning adds the load to daily work!). • Interaction between didactical and technological issues within sociotechnical support: tutors and group work play a crucial role in e-learning-modules as they mediate between the requirements, technical possibilities/problems and social dynamics of the e-learning modules. • Electronic group work needs to be discovered as a learning experience of its own. • Group work bears risks and opportunities, is at the same time socially challenging and creates new learning experiences. • Intercultural aspects are linked to technical aspects: Time differences, quality of technology at hand and language barriers (academic language) can create communication problems between participants from different countries. • Online-sources are most useful in combination with supplementary materials (CD-Rom) and group work.

- Strengthening of group commitments;
- Clarification of submission procedures;
- Research guidance;
- Obligatory contributions to the forums and chat rooms;
- Language diversity;
- Less academic language;
- Small weekly activities with one large assignment at the end of a module; and
- Reflections on different professional backgrounds and qualification levels.

CONCLUSION

Observing participant experiences with group e-learning technology, group work, and learning itself sharpens the view on how the shape and learning effects of new media in higher education is contested and evolves in communities of practice. Due to that, we assessed e-learning as a mixture of didactically and technically situated and mediated features. Carefully applied, these features foster and enable human learning. Although successfully evaluated for the EEEYMS case, distance and e-learning technologies are not free of risks and controversies. In this respect, Table 1 summarized important issues of distance learning materials, issues that are at the same time technical, didactical and social. Throughout these issues, participants clearly articulated how working in groups "kept them going". The availability of synchronous and asynchronous communication tools utilised by the e-learning cohort, has supported that desire or preference for shared learning, without the many barriers perceived by "regular" learning.

REFERENCES

Armitage, S., Brown, T., & Jenkins, M. (2001). *Management and implementation of virtual learning environments: A UCISA funded survey.* UCISA.

Biggs, J. (1999). *Teaching for quality learning at University*. SRHE, Open University.

Donohue, C. (2002). It's a small world: Taking your first steps into online teaching and learning. *Childcare Information Exchange, 9*, 20-25.

Dron, J. (2005). Control termites and e-learning, *IADIS International Conference Web Based Communities*, 23-25.

Gläser, J., & Laudel, G. (2004). *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen.* Verlag für Sozialwissenschaften.

Johnson, D. W., & Johnson, R. T. (1989). *Cooperation and competition: Theory and research*. Interaction Book Company.

Kirkup, G., & von Prummen, C. (1990). Support and connectiveness: The needs of women distance education particpants. *Journal of Distance Education*. Retrieved from http://cade.athabascau.ca.vol5.2/7_kirkup_and_von_prummer.html

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge University Press.

Mayring, P. (2000). *Qualitative Inhaltsanalyse. Grundlagen und Techniken*. Deutscher Studien Verlag.

Meisalo, V., Lavonen, J., & Juuti, K. (2005). *A case study on a group of unqualified primary teachers taking a science education course in a Web based environment.* IADIS International Conference Web Based Communities.

Moore, M. G. (1980). Independent study, Redefining the discipline of adult education. In Boyd & Apps (Eds.), *Redefining the discipline of adult education*, (pp. 16-27). San Francisco.

Rammert, W. (2002). *Technik als verteilte Aktion. Wie technisches Wirken als Agentur in hybriden Aktionszusammenhängen gedeutet werden kann.* Technical University Technology Studies Working Papers, TUTS-WP-3.

Salmon, G. (2000). Computer mediated conferencing for management learning at the Open University. *Management Learning, 31*(4), 491-502.

Sharan, S., & Shaulov, A. (1990). Cooperative learning, motivation to learn and academic achievement. In S. Sharan (Ed.), *Co-operative learning: Theory and research* (pp. 173-202).

Wenger, E. (1999). *Communities of practice: Learning, meaning, and identity*. Cambridge University Press.

Wiesner, H., Kamphans, M., Schelhowe, H., Metz-Göckel, S., Zorn, I., Drag, A., Peter, U., & Schottmüller, H. (2004). *Leitfaden zur Umsetzung des Gender Mainstreaming in den "Neuen Medien in der Bildung – Förderbereich Hochschule"*. Bremen – Dortmund.

Wiesner-Steiner, A., Wiesner, H., & Schelhowe, H. (2006). Technik als didaktischer Akteur: Robotik zur Förderung von Technikinteresse. In C. Gransee (Ed.), *Hochschulinnovation. Gender-Initiativen in der Technik. Reihe: Gender Studies in den Angewandten Wissenschaften. Gender Studies & Applied Sciences* (pp. 89-115). LIT-Verlag Hamburg.

Zorn, I., & Wiesner-Steiner, A. (2006). Technologies for inclusion: Design for intercultural virtual communities. *ICDML Proceedings of the first International Conference on Digital Media and Learning, Bangkok,* Thailand (pp. 71-77).

KEY TERMS

E-Learning: Electronically supported learning; all kinds of learning with the use of digital media for the presentation or distribution of learning materials and/or the support of communication.

Gender: Refers to the social gender role or the social gender qualities; everything that is typically assiociated with men and women within a certain culture.

Gender Mainstreaming: Refers to the political idea of gender equality on all levels of society.

ODL: Open distance learning.

Sociotechnical: Emerging from management and economic research as well as from the sociology of science and technology, the term refers to the mutial dependency and intertwining of social and technical features within companies, production systems or social systems at large.

VLE: Virtual learning environment.

Category: IT Education

Distance Learning Overview

Linda D. Grooms
Regent University, USA

INTRODUCTION

The knowledge explosion, the increased complexity of human life, and the ubiquitous nature of technology coupled with the globalization of the marketplace herald the need to embrace the most effective methods and formats of teaching and learning. Currently providing powerful educational opportunities, the science and technology of distance learning continues to multiply at unprecedented rates. Where just a short time ago traveling from village to village verbally disseminating knowledge was the only process of training those at a distance, today many eagerly embrace the rapidly expanding synchronous and asynchronous delivery systems of the 21st century. So what exactly is distance learning?

In very simplistic terms, distance learning is just that: learning that occurs at a distance (Rumble & Keegan, 1982; Shale, 1990; Shale & Garrison, 1990) or that which is characterized by a separation in proximity and/or time (Holmberg, 1974, 1977, 1981; Kaye, 1981, 1982, 1988; D. J. Keegan, 1980; McIsaac & Gunawardena, 1996; M. Moore, 1983; M. G. Moore, 1973, 1980, 1989a, 1989b, 1990; Ohler, 1991; Sewart, 1981; Wedemeyer, 1971). In his 1986 theory of transactional distance, Michael Moore (Moore & Kearsley, 1996) defined distance not only in terms of place and time, but also in terms of structure and dialogue between the learner and the instructor. In this theory, distance becomes more pedagogical than geographical. As structure increases, so does distance. As dialogue increases, distance declines, thus accentuating the need for interaction in the distance learning environment. Saba (1998) furthered this concept, concluding,

the dynamic and systemic study of distance education has made "distance" irrelevant, and has made mediated communication and construction of knowledge the relevant issue.... So the proper question is not whether distance education is comparable to a hypothetical "traditional," or face-to-face instruction, but if there is enough interaction between the learner and the instructor for the learner to find meaning and develop new knowledge. (p. 5)

To facilitate greater interaction in the geographically and/or organizationally dispersed distance environment, today, individuals most often use some form of technology to overcome the barrier of separation, affording institutional and learner opportunity to transcend intra- and inter-organizational boundaries, time, and even culture. By definition, the paradigm of distance learning revolutionizes the traditional environment (Martz & Reddy, 2005); however, even with this change, learning, which involves some manner of interaction with content, instructor, and/or peers, remains at the core of the educational process.

Although imperative in both environments, these three types of interaction seem to be at the hub of the ongoing traditional-vs.-distance argument. Traditionalists often fear that with anything other than face-to-face instruction, interaction somehow will decrease, thus making learning less effective, when in reality, numerous studies have revealed no significant difference in the learning outcomes between traditional and distance courses (Russell, 1999). In fact, distance courses have been found to "match conventional on-campus, face-to-face courses in both rigor and quality of outcomes" (Pittman, 1997, p. 42). Despite these findings, critics still abound.

Two distinguishing characteristics of the nontraditional environment—individualized learning and flexibility—often arouse suspicion and caution among traditionalists (Grooms, 2000). Many are convinced that with any form of study outside the confines of the typical brick and mortar, "every vestige of intellectual rigor [will] disappear into oblivion....[These skeptics interpret] individualized learning as individualized isolation, especially from faculty, and they look on flexibility as no more than a synonym for escape from regulation and responsibility" (Gould, 1972, p. 9). Inherently, they fear loss of interaction.

In contrast, with their introduction of equivalency theory, Simonson, Schlosser, and Hanson (1999) accentuated the concept of equivalency as "central to the widespread acceptance of distance education" (p. 72), thus supporting Keegan's (1989) call for parity in quality, quantity, and status. Furthermore, recognizing the need to bring integrity and prestige to the field, Shale and Garrison (1990) suggested building a framework based not on isolation but upon interdependence, which would imply that distance learning would merely become an alternative method for delivering traditional content. This begs the question of how distance learning has evolved.

BACKGROUND

As previously mentioned, distance learning has been with us in one form or another virtually since the creation of time.

For years, itinerant teachers traveled from village to village verbally disseminating information to those hungry for knowledge; however, the invention of Guttenberg's printing press in 1440 made possible serious distribution of learning to larger numbers of people.

Capitalizing on this broader use of print media, correspondence study became a popular form of distance education, the first record of which was in 1728 when Caleb Philipps advertised the introduction of shorthand (Battenberg as cited in Baath, 1980; & Holmberg, 1986). Often conjuring thoughts of isolation and autonomy, this record of instruction mirrored those images. In fact, in this account, there was no mention of interaction of any type other than what was inherent with the content.

Over a hundred years later in his 1833 Swedish advertisement, although not directly stated, Meuller's offer to study composition seems to be the first to imply some form of exchange between the student and teacher. More definitively, in 1840, the most acknowledged root of distance learning explicitly employing learner-instructor interaction began in the United Kingdom. Using passages from the Bible, Isaac Pitman taught shorthand (Baath, 1980; Holmberg, 1974; Kaye, 1988; Rumble, 1986), but this time, once learners transcribed these passages, they were returned for correspondence with the teacher via the penny post, thus some call it postal teaching (Dewal, 1988).

As evidenced in these early days of pure correspondence education, any offered guidance transpired through some form of dispatched communication such as the mail (Wedemeyer, 1971), and student contact, even with the instructor, was not necessarily encouraged. This is clearly seen in Keegan's (1980) classic article "On Defining Distance Education," where he documented that in its strictest sense, pure correspondence study advocates specified that "students enrol [sic] with them because they 'want to be left alone'" (p. 31).

As distance learning evolved, learner-instructor interaction became increasingly important, thus catapulting the first of two paradigm shifts. While many recognized the significantly positive impact of the distance learning interactive component (Cookson, 1989; Grooms, 2000, 2003; Robinson, 1981), others such as Daniel and Marquis (1979) accentuated the importance of getting the right independence-interaction mixture. Further stressing this need for learner-instructor interaction, Holmberg (1982) directly confronted the pure correspondence model when he concluded that "any postgraduate distance study must have a truly communicative character if more is meant than merely providing reading lists and odd comments on students' work" (p. 259).

Print remained the primary mode of distance learning until the 1920s and '30s when the introduction of radio broadcasts soon followed by television and satellite delivery systems initiated the labor pains for the birth of the current online technological revolution. Prior to the advent of the World Wide Web (WWW) in the early 1990s, interaction continued to transpire primarily between the learner and content, with occasional interaction between the learner and the instructor through such means as telephone and videoconferencing. The second paradigm shift was on the horizon.

THE CURRENT STATE OF AFFAIRS

To be embraced, any new mode or method of education must do more than merely emulate the status quo. The virtual environment of the 21st century claims to do just that. While offering flexibility from traditional proximity and time constraints (Barnes & Greller, 1994; Harasim, 1990; Hiltz & Johnson, 1990; Kaye, 1989; M. Moore, 1983), computer-mediated communication (CMC) (Harasim, 1993; Kaye, 1989; McIsaac & Gunawardena, 1996) serves as an excellent participation equalizer. Coupled with unprecedented technological advances (Graham, Allen, & Ure, 2005; Osguthorpe & Graham, 2003), the line between traditional face-to-face learning and that which occurs at a distance becomes increasingly blurred.

While multiple studies have indicated there is no significant difference between distance and traditional learning effectiveness, the geographical dispersion of people, shifting market conditions, and rapid technological changes continue to compel transformation in the way we do business both in the marketplace and in the halls of academe. Promising to deliver increased access, quality, and efficiency of learning in an ever-growing competitive market (Benoit, Benoit, Milyo, & Hansen, 2006), the technology of higher education alters teaching and learning (Kapitzke, 2000) and thus instructor and student roles (Stadtlander, 1998).

Learning is no longer dispatched through print or even audio or video, but rather it is now mediated through synchronous (interactive/real time) or asynchronous (delayed interaction) means. Regardless of technology's sophistication, the most critical consideration must always be to align the task, the delivery method, and the delivery format.

Distance Learning Delivery Methods

Almost 150 years following the advent of postal teaching and the first record of any form of learner-instructor interaction, Linda Harasim (1989), a pioneer in the online classroom, clearly differentiated three delivery methods that she believed distinguished traditional, distance, and online education: *one-to-many*, as in the traditional lecture method when one instructor addresses many students; *one-to-one*, as in the tutorial method; and *many-to-many*, a collaborative process with students learning from each other, with or without an instructor. In the first method, learners are mere passive recipients of knowledge and information, whereas in the latter two, they are actively involved in the learning

process. A clear shift in the role of the instructor transpires from information dispenser to one who facilitates an environment where knowledge "emerges from active dialogue among those who seek to understand and apply concepts and techniques" (Hiltz, 1990, p. 135).

Although the traditional face-to-face environment is time-place dependent, it allows for the implementation of all three delivery methods contingent upon the instructional task. In contrast, the distance and online environments are time-place independent and mediated, facilitating flexibility and reflective response; however, in like manner they also align method with context. When appropriate, distance learning uses the *one-to-one* (e.g., print media such as programmed instruction) or *one-to-many* (e.g., audio and video teaching) methods, while online classes additionally employ the *many-to-many* concept, which forces the second paradigm shift: the need for peer interaction.

Distance Learning Delivery Formats

Distance learning formats typically corral into four arenas: print, audio, video, or digital, with all serving as viable options. Apart from media or technology differences, communication within these formats can either be one-way with the learner taking a passive role or two-way with the learner interacting with either the instructor or peers.

Today, using syllabi, texts, and instructor notes, the medium of print remains a significant component in distance learning, while telephones and audio conferencing permeate our 21st century culture. Although video technology also engulfs our society, the rapid proliferation of the Internet has provided a plethora of learning opportunities, both synchronous and asynchronous, through such means as instant messaging and white boards, or through e-mail and threaded discussions. Regardless of the format, learning remains the ultimate goal. Table 1 delineates some of the various formats at our disposal.

More recently, many educational institutions have turned to a blended learning approach, combining some form of the traditional with the digital. Osguthorpe and Graham (2003) summarize this approach very well:

Those who use blended approaches base their pedagogy on the assumption that there are inherent benefits in face-to-face interaction (both among learners and between learner and instructor) as well as understanding that there are some inherent advantages to using online methods in their teaching. Thus the aim of those using blended learning approaches is to find a harmonious balance between online access to knowledge and face-to-face interaction. The balance…will vary for every course [based upon the] instructional goals, student characteristics, instructor background, and online resources….No two courses will be exactly the same….[The goal] is to ensure that the blend involves the strengths of each type of learning environment and none of the weaknesses. (p. 228)

Citing three typical reasons for using a blended approach—more effective pedagogy, increased convenience and access, and increased cost effectiveness (Graham et al., 2005)—it is easy to see why progressively more institutions are entertaining the thought of implementing such a format.

Further blurring the line between traditional and distance learning, the blending and convergence of technologies continue to increase the fluidity and ubiquity of education. Today, learners already have almost instant access to a plethora of information via the Internet. With Webcams and microphones, they can freely see and talk with instructors and peers, maintaining the dialogue heralded earlier as necessary for effective distance learning. With handheld computers, learners—no longer desk- or library-bound—can travel freely around the globe without ever missing a class. With the future in mind, McCain and Jukes (2001) posited, "at a certain point, the boundaries between reality and virtual reality will collapse because of the increased sophistication and transparency of these powerful, fused technologies" (p. 60). So, where do we go from here?

FUTURE TRENDS

In a 1965 *Electronics* magazine article, Gordon Moore, cofounder and CEO of Intel, the world's largest silicon chip manufacturing company, predicted that the number of components on the integrated circuit board would double every 12 to 18 months. While this timeframe was amended in 1975 to every 24 months (Kanellos, 2005), it has been accurate in excess of 40 years. At present, what has become known as Moore's Law shows no signs of slowing down, although in a 2000 interview with *Time* magazine, Moore did acknowledge that at some point in the next two or three generations, this exponential rate of growth would perhaps slow down to doubling every 5 years. Even then, we are quite a distance from reaching the apex of technological discovery, which history reveals directly affects distance learning.

Negroponte (1995) poignantly prophesied, "Distance means less and less in the digital world. In fact, an Internet user is utterly oblivious to it. On the Internet, distance often seems to function in reverse" (p. 178). Supporting this prediction, science and technology have triggered physical distance to become truly irrelevant. So, what does the future hold? Where will distance learning be in 5, 10, or even 20 years from now? Some would postulate that we are only limited by our imaginations.

Table 1. Distance learning delivery formats

MEDIA	PASSIVE One-Way	INTERACTIVE Two-Way	
		Synchronous Real Time	**Asynchronous** Delayed
PRINT	Texts Syllabi Instructor notes Study guides Workbooks Fax		
AUDIO	Radio Audiotape CD-Rom Voice mail	Telephone Audio conferencing with or without a bridge	
VIDEO	Videotape DVD Film Cable, broadcast, and digital television (one-way audio and video)	Satellite videoconferencing (two-way audio and one-way video) Microwave television conferencing Digital videoconferencing (two-way audio and video) Internet videoconferencing	
DIGITAL	CAI (computer-assisted instruction), using the computer as a self-contained teaching machine CD-Rom Streaming video Blogging Linklogging Moblogging Photologging Vlogging Weblogging Podcasting Phonecasting Vodcasting	Chat rooms Shared white boards Instant messaging	CMI (computer-mediated instruction) E-mail Bulletin boards (threaded discussions, newsgroups) Listservs Wikis Web-based instruction (e.g., Web CT, Blackboard)

CONCLUSION

Whether one views distance as geographical or pedagogical, as the above suggests, the technological explosion of the 21st century provides unprecedented opportunity to render distance virtually irrelevant. If Moore's Law holds fast and technology continues to double every 18 to 24 months, the exponential growth of distance learning will continue to catapult educators into uncharted territory. The question remains: Will there still be a place for the staunch traditionalists or will education be so radically transformed that traditional education is hardly recognizable? Regardless of the media (print, audio, video, or digital), method (*one-to-many*, *one-to-one*, or *many-to-many*), or format (passive or interactive), interactive learning continues to remain at the core of the distance educational process.

REFERENCES

Baath, J. A. (1980). *Postal two-way communication in correspondence education: An empirical investigation.* Malmo, Sweden: LiberHermods.

Barnes, S., & Greller, L. M. (1994). Computer-mediated communication in the organization. *Communication Education, 43*, 129-142.

Benoit, P. J., Benoit, W. L., Milyo, J., & Hansen, G. J. (2006). *The effects of traditional versus Web-assisted instruction on learning and satisfaction.* Columbia, MO: University of Missouri Graduate School.

Cookson, P. S. (1989). Research on learners and learning in distance education: A review. *The American Journal of Distance Education, 3*(2), 22-34.

Daniel, J. S., & Marquis, C. (1979). Interaction and independence: Getting the mixture right. *Teaching at a Distance, 14,* 29-44.

Dewal, O. S. (1988). Pedagogical issues: Distance education. *Prospects, 18*(1), 63-73.

Gould, S. B. (1972). Prologue: Prospects for non-traditional study. In S. B. Gould & K. P. Cross (Eds.). *Explorations in non-traditional study* (pp. 1-12). San Francisco: Jossey-Bass.

Graham, C. R., Allen, S., & Ure, D. (2005). Benefits and challenges of blended learning environments. In *Encyclopedia of information science and technology* (Vol. 1, pp. 253-259). Hershey, PA: Idea Group Publishing.

Grooms, L. D. (2000). Interaction in the computer-mediated adult distance learning environment: Leadership development through online education. *Dissertation Abstracts International, 61*(12), 4692A.

Grooms, L. D. (2003). Computer-mediated communication: A vehicle for learning. *International Review of Research in Open and Distance Learning, 4*(2). Retrieved from http://www.irrodl.org/index.php/irrodl/article/view/148/709

Harasim, L. (1989). On-line education: A new domain. In R. Mason & A. Kaye (Eds.), *Mindweave: Communication, computers and distance education* (pp. 50-62). New York: Pergamon Press.

Harasim, L. M. (1990). Online education: An environment for collaboration and intellectual amplification. In L. M. Harasim (Ed.), *Online education: Perspectives on a new environment* (pp. 39-64). New York: Praeger.

Harasim, L. M. (1993). Networlds: Networks as social space. In L. M Harasim (Ed.), *Global networks: Computers and international communication* (pp. 15-34). Cambridge, MA: The MIT Press.

Hiltz, S. R. (1990). Evaluating the virtual classroom. In L. M. Harasim (Ed.), *Online education: Perspectives on a new environment* (pp. 133-169). New York: Praeger.

Hiltz, S. R., & Johnson, K. (1990). User satisfaction with computer-mediated systems. *Management Science, 36*, 739-764.

Holmberg, B. (1974). *Distance education: A short handbook.* Malmo, Sweden: Hermods.

Holmberg, B. (1977). *Distance education: A survey and bibliography.* London: Kogan Page.

Holmberg, B. (1981). *Status and trends of distance education.* London: Kogan Page.

Holmberg, B. (1982). Distance study at the post-graduate level: Graduate study at a distance requires greater attention to communication with the student. In J. S. Daniel, M. A. Stroud, & J. R. Thompson (Eds.), *Learning at a distance: A world perspective* (pp. 258-260). Edmonton, Canada: Athabasca University, International Council for Correspondence Education.

Holmberg, B. (1986). *Growth and structure of distance education.* Wolfeboro, NH: Croom Helm.

Kanellos, M. (2005, April 1). FAQ: Forty years of Moore's law. *TechRepublic.* Retrieved March 25, 2007, from http://articles.techrepublic.com.com/2100-1035_11-5647824.html

Kapitzke, C. (2000). The sociability and spatiality of online pedagogy and collaborative learning in an educational media and technology course. *Educational Technology & Society, 3*, 344-441.

Kaye, A. (1981). Media, materials and learning methods. In A. Kaye & G. Rumble (Eds.), *Distance teaching for higher and adult education* (pp. 48-69). London: Croom Helm.

Kaye, A. (1982). Using the media for adult basic education. In A. Kaye & K. Harry (Eds.), *Using the media for adult basic education* (pp. 9-29). London: Croom Helm.

Kaye, A. (1988). Distance education: The state of the art. *Prospects, 18*, 43-54.

Kaye, A. (1989). Computer-mediated communication and distance education. In R. Mason & A. Kaye (Eds.), *Mindweave: Communication, computers and distance education* (pp. 3-21). New York: Pergamon Press.

Keegan, D. (1989). Problems in defining the field of distance education. In M. G. Moore & G. C. Clark (Eds.), *Readings in principles of distance education* (pp. 8-15). University Park, PA: The Pennsylvania State University.

Keegan, D. J. (1980). On defining distance education. *Distance Education, 1*(1), 13-36.

Martz, B., & Reddy, V. (2005). Critical success factors for distance education programs. In *Encyclopedia of information science and technology* (Vol. 1, pp. 622-627). Hershey, PA: Idea Group Publishing.

McCain, T., & Jukes, I. (2001). *Windows in the future: Education in the age of technology.* Thousands Oaks, CA: Corwin Press.

McIsaac, M. S., & Gunawardena, C. N. (1996). Distance education. In D. H. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 403-437). New York: Simon & Schuster Macmillan.

Moore, G. (2000, June 19). Gordon Moore Q & A. *Time, 155*(25), 99.

Moore, G. E. (1965, April 19). Cramming more components onto integrated circuits. *Electronics, 38*(8), 114-117.

Moore, M. (1983). The individual adult learner. In M. Tight (Ed.), *Adult learning and education* (pp. 153-168). London: Croom Helm.

Moore, M. G. (1973). Toward a theory of independent learning and teaching. *Journal of Higher Education, 44*, 661-679.

Moore, M. G. (1980). Independent study. In R. D. Boyd, J. W. Apps, & Associates (Eds.), *Redefining the discipline of adult education* (pp. 16-31). San Francisco: Jossey-Bass.

Moore, M. G. (1989a). Editorial: Three types of interaction. *The American Journal of Distance Education, 3*(2), 1-7.

Moore, M. G. (1989b, May). *Effects of distance learning: A summary of the literature.* Paper presented for Congress of the United States Office of Technology Assessment, Washington, DC.

Moore, M. G. (1990). Correspondence study. In M. W. Galbraith (Ed.), *Adult learning methods: A guide for effective instruction* (pp. 345-365). Malabar, FL: Robert E. Krieger Publishing.

Moore, M. G., & Kearsley, G. (1996). *Distance education: A systems view.* Boston: Wadsworth Publishing.

Negroponte, N. (1995). *Being digital.* New York: Alfred A. Knopf.

Ohler, J. (1991). Why distance education? *The Annals of the American Academy of Political and Social Science, 514*, 22-34.

Osguthorpe, R. T., & Graham, C. R. (2003). Blended learning environments: Definitions and directions. *The Quarterly Review of Distance Education, 4*(3), 227-233.

Pittman, V. (1997). Distance education exchange. *The Journal of Continuing Higher Education, 45*(2), 42-43.

Robinson, B. (1981). Support for student learning. In A. Kaye & G. Rumble (Eds.), *Distance teaching for higher and adult education* (pp. 141-161). London: Croom Helm.

Rumble, G. (1986). *The planning and management of distance education.* New York: St. Martin's Press.

Rumble, G., & Keegan, D. (1982). Introduction. In G. Rumble & K. Harry (Eds.), *The distance teaching universities* (pp. 9-14). London: Croom Helm.

Russell, T. L. (1999). *The no significant difference phenomenon as reported in 355 research reports, summaries and papers: A comparative research annotated bibliography on technology for distance education.* Raleigh, NC: North Carolina State University.

Saba, F. (1998). Is distance education comparable to "traditional education"? *Distance Education Report, 2*(5), 5.

Sewart, D. (1981). Distance teaching: A contradiction in terms? *Teaching at a Distance, 19*, 6-18.

Shale, D. (1990). Toward a reconceptualization of distance education. In M. G. Moore, P. Cookson, J. Donaldson, & B. A. Quigley (Eds.), *Contemporary issues in American distance education* (pp. 333-343). New York: Pergamon Press.

Shale, D., & Garrison, D. R. (1990). Education and communication. In D. R. Garrison & D. Shale (Eds.), *Education at a distance: From issues to practice* (pp. 23-29). Malabar, FL: Robert E. Krieger Publishing Company.

Simonson, M., Schlosser, C., & Hanson, D. (1999). Theory and distance education: A new discussion. *The American Journal of Distance Education, 13*(1), 60-75.

Stadtlander, L. M. (1998). Virtual instruction: Teaching an online graduate seminar. *Teaching of Psychology, 25*(2), 146-148.

Wedemeyer, C. A. (1971). Independent study. In L. C. Deighton (Ed.), *The encyclopedia of education* (Vol. 4, pp. 548-557). New York: Macmillan Company & the Free Press.

KEY TERMS

Correspondence Learning: A form of distance learning using dispatched or one-way interaction.

Equivalency: Distance learning that possesses equality with learning experienced in the face-to-face venue.

Face-to-Face Learning: Learning that is time-place dependent.

***Many-to-Many*:** A collaborative process with students learning from each other with or without an instructor.

***One-to-Many*:** The lecture method; one instructor to many learners.

***One-to-One*:** The tutorial method; one instructor to one learner.

Time-Place Dependent: Education that transpires in the same location at the same time.

Time-Place Independent: Learning that does not rely on proximity or time.

Traditional Study: Face-to-face learning.

Distributed Construction through Participatory Design

Panayiotis Zaphiris
City University, London, UK

Andrew Laghos
City University, London, UK

Giorgos Zacharia
MIT, USA

INTRODUCTION

This article presents an empirical study of an online learning community that collaborates with the course design team under the Participatory Design methodology. The different phases of this methodology were implemented using a four-stage participatory design process (Zaphiris & Zacharia, 2001):

1) building bridges with the intended users,
2) mapping user needs and suggestions to the system,
3) developing a prototype, and
4) integrating feedback and continuing the iteration.

We took advantage of the online and distributed nature of the student community to asynchronously design, implement, and study the course. We carried out the participatory design methodology by following the Distributed Constructionism pedagogical theory. During the different phases of the design process, we measured the student participation and the changes in their behavior when new design elements were introduced. We conclude that the most important element of this course was our discussion board, which helped us to promote student collaboration and the identification of the key community users who can participate productively in Participation Design activities.

There are three main sections to this article. After defining the key terminology, our Participatory Design approach is presented and its linkage to the Distributed Constructionism pedagogical theory specified. The article ends with ideas for future research and a set of conclusions.

BACKGROUND

Participatory Design

Participatory design (PD) refers to a design approach that focuses on the intended user of the service or product, and advocates the active involvement of users throughout the design process. PD is often termed as the "Scandinavian Challenge" (Bjerknes, Ehn & Kyng, 1987), since it was researchers from Scandinavian countries who pioneered its use in information systems development (Blomberg & Henderson, 1990; Bodker, Gronbaek & Kyng, 1993; Ehn, 1988).

User involvement is seen as critical both because users are the experts in the work practices supported by these technologies and because users ultimately will be the ones creating new practices in response to new technologies (Ellis, Jankowski & Jasper, 1998).

Blomberg and Henderson (1990) characterize the PD approach as advocating three tenets:

- The goal is to improve the quality of life, rather than demonstrate the capability of technology.
- The orientation is collaborative and cooperative rather than patriarchal.
- The process is iterative since PD values interactive evaluation to gather and integrate feedback from intended users.

By involving the users in the design process, the designers also gain knowledge of the work context, so that the new technology explicitly incorporates the values, history, and context of the work system (Ehn, 1988). The users take part in the entire design, implementation, and decision-making processes. Their involvement ensures that their activities are taken into account. Also by participating in the design, the users have a sense of "ownership" (Brown & Duguid,

2000), and the final system will have an increased user acceptance.

Distributed Constructionism

Simply put, Constructionism can be thought of as "learning-by-making" (Papert, 1991). It is both a theory of learning and a strategy for education (Papert, 1993). It focuses on the construction of a system rather than the information that will be used. The theory views computer networks as a new medium for construction, not as an information distribution channel. By embedding construction activities within a community, new ways for students to learn arise (Papert, 1993). Based on Piaget's constructivist theories, people don't get ideas, they make them. Learning is an active process where people construct knowledge from their experiences (Resnick, 1996).

Distributed Constructionism (Resnick, 1996) extends the Constructionism theory (Papert, 1991, 1993) to knowledge-building communities, where the online learning community (instead of one student) collaboratively constructs knowledge artifacts (Resnick, 1996). Distributed Constructionism asserts that "a particularly effective way for knowledge-building communities to form and grow is through collaborative activities that involve not just the exchange of information but the design and construction of meaningful artifacts" (Resnick, 1996). The three major activities of DC, within the context of an online learning community, are (Resnick, 1996):

- *Discussing Constructions:* Students discuss their constructions during the design, implementation, evaluation, and reiteration phases.
- *Sharing Constructions:* Web-based systems allow students to share their constructions and make them part of the shared knowledge.
- *Collaborating on Constructions:* The community can use online communication to collaborate on the design and development of the knowledge artifacts.
- *Distributed Constructionism:* Was enhanced among the users of the system, due to the iterative structure of our Participatory Design approach. Both the learning experience of the users and the content and functionality of the course itself were enhanced by the knowledge artifacts that were contributed to the course.

DESIGN APPROACH AND COURSE EVOLUTION

In this section, a case study applying the theories presented in the previous section is described.

Our focus has been to design an online learning community around a Computer Aided Language Learning (CALL) course. We believe that online interaction and community would increase users' motivation, commitment, and satisfaction with the online course. The Participatory Design methodology blends nicely with our goal. In particular, involving users during system development is thought to lead to greater user commitment, acceptance, usage, and satisfaction with the system (Baroudi, Olson & Ives, 1986).

In the design phase of the online course, we implemented PD as a four-step process (Ellis et al., 1998).

1) Building Bridges with the Intended Users

This step opened lines of communication between intended users and the development team. Specifically, this step involved the initialization of a multidisciplinary development team, identifying key groups of end users, and creating new methods of communication with users.

The development team in this project came out of the Kypros-Net (2002) group. Through their involvement in Cyprus and Greece related projects, they had longstanding relations with the intended user community.

The intended users have been especially people of the Greek Diaspora, travelers to Cyprus and Greece, and other Greek-speaking areas, and people who are generally interested in the Greek culture and language or languages in general. In our case, bridges with the intended users were build through our years of work at providing information about Cyprus through the Web pages of Kypros-Net, which primarily attracts the same user population as our intended Greek language online course.

2) Mapping User Needs and Suggestions to the System

Our conceptual design model has been "to design an effective online Greek language course that can build and sustain an online learning community of students." Based on the questions and inquiries we received from our users, we tried to match their needs (they wanted an easy-to-follow, both elementary and advanced course that they could attend at their own pace) with our conceptual design model.

3) Developing a Prototype

The project consists of 105 audio files, which were originally recorded as Radio lessons in Modern Greek for English speakers in the 1960s. The lessons were retrieved from the archives of the Cyprus Broadcasting Corporation, digitized in Real Audio 5.0 format, and published online through the course. Although, an optional textbook accompanied the original radio lessons, the online lessons were designed as a complete standalone course. We used several tools to assist students with the lessons, including an online Eng-

lish-Greek-English dictionary, a Greek spell checker, and a Web-based discussion board. The discussion boards served as the foundation for creating a community of online students and enhanced the learning experience with Distributed Constructionism.

4) Integrating Feedback and Continuing the Cycle

Feedback from our users and suggestions are continuously incorporated into our design through a series of additions and corrections. For example, we were asked to add an online notes section and to encode some files again because they were corrupted.

An important element in the participatory design methodology is the direct involvement of the users in all stages of the design process. We kept the users involved by participating in the discussion boards, and sharing with them design and development plans for the course.

The students of the audio courses included people with no knowledge of Greek language, bilingual members of the Greek Diaspora, as well as high-school professors of non-Greek language. These students created an open online community whose collaboration has boosted the learning experience of the whole community. The Web-based discussion board has proven to be the most constructive tool for the students' learning experience and the main source of feedback for the maintainers of the project. The experiences shared on the discussion board included tricks and tips on how to record the audio files, installation of Greek fonts, learning methodologies, and questions about the Greek language itself that arise from the lessons. The experienced users (some of them were retired teachers of foreign languages) had taken a lead role in the vast majority of the threads on the discussion board, answering most of the questions and encouraging the beginners to study the lessons further. They have also become the communication interface between the maintainers of the project and the community's requests.

At some point, the users started exchanging, through e-mail, written notes taken by the experienced users. They also used the discussion board to announce the availability of their personal notes. This behavior suggests that we must provide (and we did) the users with the capability to post their notes on the project's site.

The students had initiated Distributed Constructionism themselves. The course designers only provided technical support to facilitate the students' construction activities.

Discussing the Constructions

The course designers offered to provide publishing access to the online course to whomever wanted to contribute their material. Five users asked to be given access. Consequently, the five users, along with the two course designers, constituted the Participatory Design team. The PD team solicited contributions from the user community. The users suggested that they should transcribe the audio lessons, and compile verb lists, vocabulary lists, and grammatical notes for each lesson.

Sharing the Constructions

All the user contributions were shared in the common area of the online course. The user members of the PD team regularly posted notices on the discussion board about new material for the course. Also, other less active users chose to offer contributions for the course, by posting on the discussion board, rather than contacting the PD team. In their study on student involvement in designing an online foreign language course, Zaphiris and Zacharia (2002) state that the discussion board proved to be the most constructive tool for the students' learning experience and the main source of feedback for the maintainers of the project.

Collaborating on the Constructions

The user members of the PD team did not include any native speakers of Greek. They were all learning the language through the online course, and at that stage, they primarily depended on the audio lessons. In order to ensure the quality of the new material before publishing them on the course Web site, the user members of the PD team implemented a peer review process. A group of seven users, which included the five central user PD team members, reviewed and corrected all the material before posting them on the Web site. Each of the seven users offered to transcribe a number of the 105 audio lessons, and two of them also offered to provide verb and vocabulary lists. However, all materials were posted in a private area first, reviewed by the seven user members of the PD team, and posted on the Web site, when the five PD users were satisfied with the quality. Then the two PD course designers, who were both native Greek speakers, would go over the already published material and make sure it was correct. Most of the mistakes we had to correct were spelling mistakes, and we rarely had to correct grammatical mistakes.

Two months after the Distributed Constructionism effort started, students of the audio lessons managed to transcribe 81 out of the 105 lessons, correct them through the peer review process among themselves, and post them on the project's Web site. Six months later, the students had transcribed and peer reviewed all 105 lessons.

The knowledge constructed attracted significant user attention. The access to the audio lessons, the language tools, and the total access of the message board and the notes pages all kept increasing exponentially (Zaphiris & Zacharia, 2001). However, once we allowed our users to publish their own notes, there was a dramatic shift of traffic from the message board to the notes pages. In our view this is due to the fact

that the users did not need to visit the discussion board any more to find out where other users had posted their notes. All the content was already aggregated and organized in a central location.

The course's popularity is apparent from the fact that the course currently has more than 25,000 registered students who actively participate in an online community that evolved around the course.

FUTURE TRENDS

Future work on this specific project will focus on a non-virtual, face-to-face participatory design team. Like the previous PD team, key stakeholders (teachers of Greek in the Diaspora, students, administrators, and designers) will work together, participating and interacting throughout the whole iterative design process. They will once again collaborate on the content and functionality development, peer review, and publish content contributions.

We believe that by encouraging the active involvement of the users, the product developed will be more enjoyable, more usable, and most importantly, more catered to their specific needs and requirements.

Also we anticipate that the expected benefit of this face-to-face PD team versus the virtual PD team will be that everyone involved will feel more like a team and have a stronger relationship with each other. Since the PD team will be a face-to-face one, communication will be better and there will be fewer misunderstandings or misinterpretations, and finally the collaboration results should be more immediate, and the final product more usable and acceptable by all the stakeholders.

From our analysis of existing literature, we observed that there is a need for additional research in areas like ethnography in participatory design and the application of our proposed methodology to new domains. As new delivery e-learning technologies are constantly emerging, research into Distributed Constructionism with the latest technology also remains important. Finally, evaluations of case studies of PD and DC will be very useful to case-specific applications of the theories.

CONCLUSION

By facilitating Distributed Constructionism in the iteration phase of a Participatory Design methodology, we enhanced the learning experience in our Web-based training. A questionnaire evaluation (Zaphiris & Zacharia, 2001) shows that the end system received high usability ratings from the users. Therefore, Distributed Constructionism enhanced the learning experience of both the PD team and the more passive users.

The students who participated actively in the design of the course also played a central role in the discussion board, answering other students' language questions, helping students to overcome technical problems, and helping them to find other resources to enhance their learning of the Greek language. These observations are with agreement the underlying goals of Participatory Design, which was an integral part of the development of this specific course.

Furthermore, the results of the analysis of the user questionnaire and the server logs shows that the final product (the course) meets—to a very large extent—the expectations and needs of the whole user population of this specific course. We believe that the direct involvement of the users in the development of the course helped in designing a more usable course that enhanced the learning of our users, and provided them with an enjoyable and rewarding experience.

REFERENCES

Baroudi, Olson & Ives (1986). An empirical study of the impact of user involvement on system usage and information satisfaction. *CACM, 29*(3), 232-238.

Bjerknes, G., Ehn, P. & Kyng, M. (Eds.). (1987). *Computers and democracy — A Scandinavian challenge.* Aldershot: Gower.

Blomberg, J.L. & Henderson, A. (1990). Reflections on participatory design: Lessons from the Trillium experience. *Proceedings of CHI'90* (pp. 353-359). Seattle, WA: ACM Press.

Bodker, S., Gronbaek, K. & Kyng, M. (1993). Cooperative design: Techniques and experience from the Scandinavian scene. In D. Schuler & A. Namioka (Eds.), *Participatory design: Principles and practices* (pp. 157-175). Hillsdale, NJ: Lawrence Erlbaum.

Brown, J.S. & Duguid, P. (2000). *The social life of information*. Boston: Harvard Business School Press.

Ehn, P. (1988). *Work-oriented design of computer artifacts.* Hillsdale, NJ: Lawrence Erlbaum.

Ellis, R.D., Jankowski, T.B. & Jasper, J.E. (1998). Participatory design of an Internet-based information system for aging services professionals. *The Gerontologist, 38*(6), 743-748.

Kypros-Net Inc. (2002). *The world of Cyprus.* Retrieved December 4, 2002, from *www.kypros.org*

Nielsen, J. (1993). *Usability engineering*. Chestnut Hill, MA: AP Professional.

Papert, S. (1991). Situating construction. In I. Harel & S. Papert (Eds.), *Constructionism* (pp. 1-12). Norwood, NJ:

Ablex Publishing.

Papert, S. (1993). *The children's machine: Rethinking school in the age of the computer.* New York: Basic Books.

Perlman, G. (1999). *Web-based user interface evaluation with questionnaires.* Retrieved December 4, 2002, from *www.acm.org/~perlman/question.html*

Resnick, M. (1996). *Distributed Constructionism.* Retrieved December 4, 2002, from *Web.media.mit.edu/~mres/papers/Distrib-Construc/Distrib-Construc.html.*

Zaphiris, P. & Zacharia, G. (2002, September). Student involvement in designing an online foreign language course. *Proceedings of the British HCI Conference* (Volume 2, pp. 170-173), London.

Zaphiris, P. & Zacharia, G. (2001, October 23-27). User-centered evaluation of an online modern Greek language course. *Proceedings of the WebNet 2001 Conference,* Orlando, FL.

KEY TERMS

Computer Aided Language Learning (CALL): The use of computers in learning a language.

Distributed Constructionism (DC): An extension of the Constructionism theory to knowledge-building communities, where the online learning community (instead of one student) collaboratively constructs knowledge artifacts.

Ethnography: The branch of anthropology that provides scientific description of individual human societies.

Human-Computer Interaction: The study, planning, and design of what happens when humans and computers work together.

Participatory Design (PD): A design approach that focuses on the intended user of a service or product, and advocates the active involvement of users throughout the design process.

Pedagogy: The activities of education or instructing or teaching.

User-Centered Design: Puts the user into the center of the software design process.

Web-Based Training (WBT): Anywhere, anytime instruction delivered over the Internet, or a corporate intranet to learners.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 902-906, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Distributed Geospatial Processing Services

Carlos Granell
Universitat Jaume I, Spain

Laura Díaz
Universitat Jaume I, Spain

Michael Gould
Universitat Jaume I, Spain

INTRODUCTION

The development of geographic information systems (GISs) has been highly influenced by the overall progress of information technology (IT). These systems evolved from monolithic systems to become personal desktop GISs, with all or most data held locally, and then evolved to the Internet GIS paradigm in the form of Web services (Peng & Tsou, 2001). The highly distributed Web services model is such that geospatial data are loosely coupled with the underlying systems used to create and handle them, and geospatial processing functionalities are made available as remote, interoperable, discoverable geospatial services.

In recent years the software industry has moved from tightly coupled application architectures such as CORBA (Common Object Request Broker Architecture—Vinoski, 1997) toward service-oriented architectures (SOAs) based on a network of interoperable, well-described services accessible via Web protocols. This has led to *de facto* standards for delivery of services such as Web Service Description Language (WSDL) to describe the functionality of a service, Simple Object Access Protocol (SOAP) to encapsulate Web service messages, and Universal Description, Discovery, and Integration (UDDI) to register and provide access to service offerings. Adoption of this Web services technology as an option to monolithic GISs is an emerging trend to provide distributed geospatial access, visualization, and processing. The GIS approach to SOA-based applications is perhaps best represented by the spatial data infrastructure (SDI) paradigm, in which standardized interfaces are the key to allowing geographic services to communicate with each other in an interoperable manner. This article focuses on standard interfaces and also on current implementations of geospatial data processing over the Web, commonly used in SDI environments. We also mention several challenges yet to be met, such as those concerned with semantics, discovery, and chaining of geospatial processing services and also with the extension of geospatial processing capabilities to the SOA world.

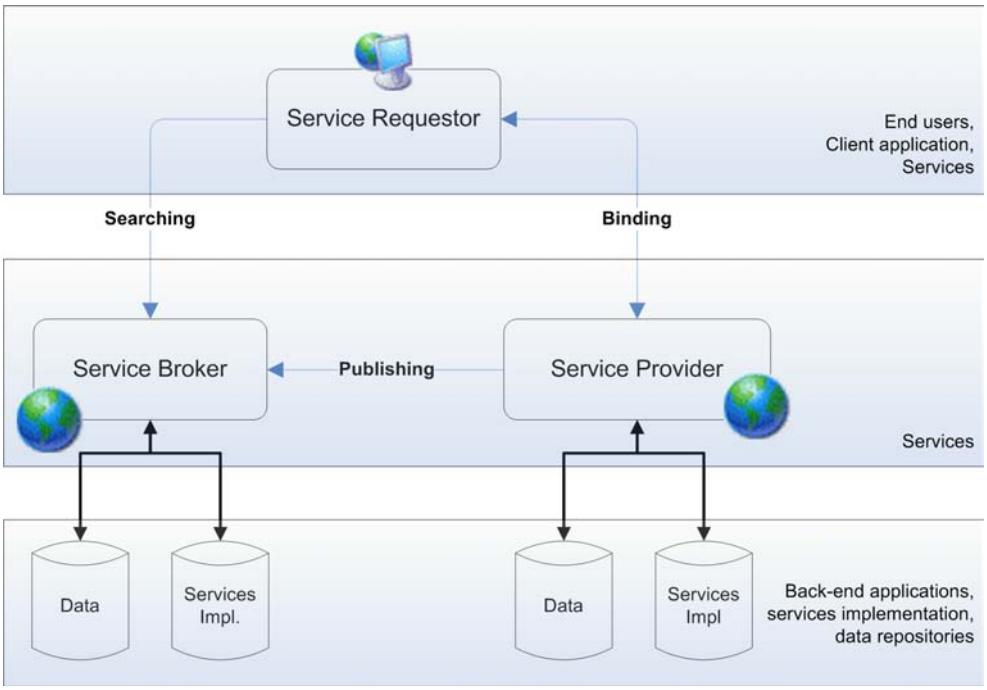
BACKGROUND

Service-Oriented Architecture

A Web service is an executable program available on the Internet. Services are the basic units for creating distributed applications in the context of SOAs. As Papazoglou (2008) stated, SOA is an architectural style to design service-centric applications relying on published and discoverable interfaces. Web services are, by definition, loosely coupled (independent units) and are well described (interface description contains functional properties), thereby promoting one of the goals of SOA: enabling interoperability or the ability of software components to interact with minimal knowledge of the underlying structure of other components (Sheth, 1999). Interoperability is achieved by using standard interfaces (SOA does not focus on the concrete implementations of components) and also by decomposing an application's functionality into modular and flexible services. Such building-block services can be published, discovered, aggregated, reused, and invoked using standard protocols and specifications, independently of the specific technology used to create each component. Essentially SOA introduces a new philosophy for building a pyramid of distributed applications where Web services can be published, discovered, and bound together to create more complex value-added services (Alameh, 2003; Lemmens et al., 2006).

Figure 1 illustrates some of the roles and operations in SOA-based applications. There are three different main SOA roles: service provider, service requestor, and service broker. Each SOA role interacts with others utilizing three basic operations: publication, search, and binding. The service provider publishes service descriptions to the service broker. The service requester searches the required services by querying the service broker and then consumes (binds to) them. Note that often the role of service requester is assigned both to end users (and client applications) and to other services. The latter makes use of two key mechanisms in SOA: service reuse and service chaining to create new,

Figure 1. Roles and operations in SOA



complex, value-added services from simpler, discoverable services. In this sense, services can play the role of service requestor and service provider.

The OWS Service Framework

Within the GIS community, the Open Geospatial Consortium (OGC)—an international industry consortium created in 1994 to develop consensus-based open standards and specifications to support the exchange, sharing, and processing of geospatial data—has adopted a general set of interfaces for a wide range of geospatial Web services (ISO 19119, 2005). Table 1 lists a sample of key OGC Web Services (OWS) categorized as defined in ISO 19119.

These OWS services fall into five categories as follows:

- *Application services* are client-side applications that provide an entry point for end users to find and access geospatial data and services. Among the notable

Table 1. Examples of OGC Web Services

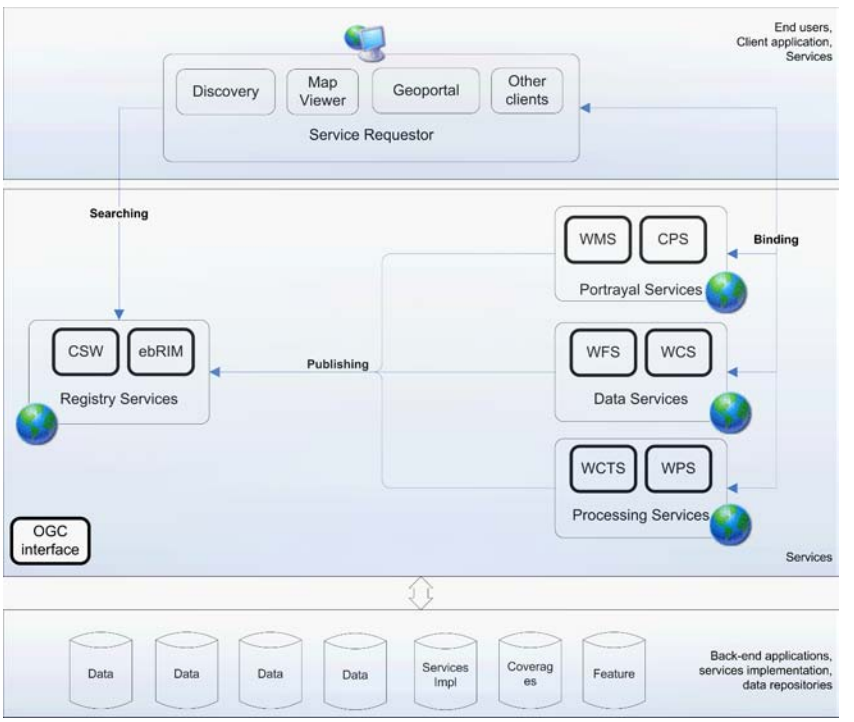
Service Category	Service Name
Application Services	Discovery Application Services Map Viewer Application Services Sensor Web Application Services Geoportal (one-stop portal)
Registry Services	Catalog Service (CSW)
Data Services	Web Feature Service (WFS) Web Coverage Service (WCS)
Portrayal Services	Web Map Service (WMS) Coverage Portrayal Service (CPS)
Processing Services	Web Coordinate Transformation Service (WCTS) Geocoder Services Gazetteer Services Route Determination Services Web Processing Services (WPS)

examples of application services are geoportals (Bernard, Kanellopoulos, Annoni, & Smits, 2005), which in turn may integrate other client-side application services such as discovery services and map viewers. Examples of geoportals may be found at *http://www.geodata.gov and http://geoportal.jrc.it/*.

- *Registry services* (often called catalog services) are a special kind of service that offers end users a common mechanism to register, search, and access discoverable geospatial data and services.
- *Data services* are the basic geospatial services that serve geospatial data to application services. Examples of data services include the Web Feature Service (WFS), which filters and retrieves vector format representations of geospatial features and feature collections encoded in Geographic Markup Language (GML) (Cox, Daisay, Lake, Portele, & Whiteside, 2002), and the Web Coverage Service (WCS), which provides access to client-specific continuous coverage or image datasets.
- *Portrayal services* may be also considered a specialized data service that produce rendered data such as portrayed maps, perspective views of terrain, annotated images, and so on. Examples are the Web Map Service (WMS) that dynamically produces spatially referenced maps of client-specified criteria from one or more geographic datasets, returning the map views in well-known image or graphics formats.
- *Processing services* essentially transform geospatial data to produce new data or actionable information. Examples are the Web Coordinate Transformation Service (WCTS), which transforms the geographic coordinates of feature (map) or coverage (imagery) data from one coordinate reference system (CRS) to another; and the Gazetteer Service, which provides location geometries for specified geographic names. In order to help standardize access and binding to processing services, the OGC created the Web Processing Service (WPS) specification (Schut, 2007), which describes the interfaces needed in order to offer generic geospatial processing services over the Internet, as described in the following section.

Spatial data infrastructures (SDIs) were designed to share existing geospatial data (most held by the public sector) and make them widely accessible and available at the lowest possible cost, where and when they are needed (Granell, Gould, Manso, & Bernabé, 2008). An SDI can be thought of as a network of interoperable Web services to facilitate basic geospatial data (e.g., a digital topographic map) and customized information (e.g., a daily forest fire risk map) and services. Figure 2 summarizes the conceptual

Figure 2. The OWS Service Framework (adapted from Percival, 2003; Yang & Tao, 2006)



SDI architecture that may be interpreted as a traditional three-tier client-middleware-server model, where GIS applications (clients) seek geospatial data content (servers) that are discovered and then possibly transformed or processed by intermediary services (middleware) before results are presented back to the client tier. The presentation layer in Figure 2 includes the application services, whereas the middleware layer contains data services, registry services, portrayal services, and processing services.

Beyond the three-tier model, however, under the SOA perspective, the previous SDI architecture also may be interpreted using the Web services 'publish-find-bind' triangle model (Papazoglou, 2008), as shown in Figure 1. In this context, the OGC proposed the OWS Service Framework (OFS) as the common set of interfaces required for enterprise-wide interoperability within and beyond the GIS community (see Figure 2). Following this framework, geospatial data content (and service) offers are published to registry services, which are later queried to discover (find) the data or services, and finally the client application binds to (consumes or executes) them. In this sense, the adoption of a common geographical data model expressed in GML and standardized OGC specifications constitutes one ingredient to achieve geospatial data integration and interoperability in the wider sense (Díaz, Granell, & Gould, 2008a). The next section will focus on OGC specifications for geospatial (Web) processing services.

DISTRIBUTED GEOSPATIAL PROCESSING SERVICES

Although OGC has already proposed specifications under the processing services category (see Table 1), these are devoted primarily to performing specific and well-defined processing functions. A substantial leap ahead in the domain of processing services was the recently released OGC Web Processing Service (WPS) specification (Schut, 2007), which was designed to encapsulate generic geoprocessing operations over the Internet. The WPS specification allows any piece of geospatial processing code to be published and accessed as if it were a common OWS service (WMS, WFS, etc.). This section focuses on this new specification and describes some emerging open source frameworks that support the implementation of distributed geospatial processing services as defined by the OGC WPS.

OGC Web Processing Service Interface

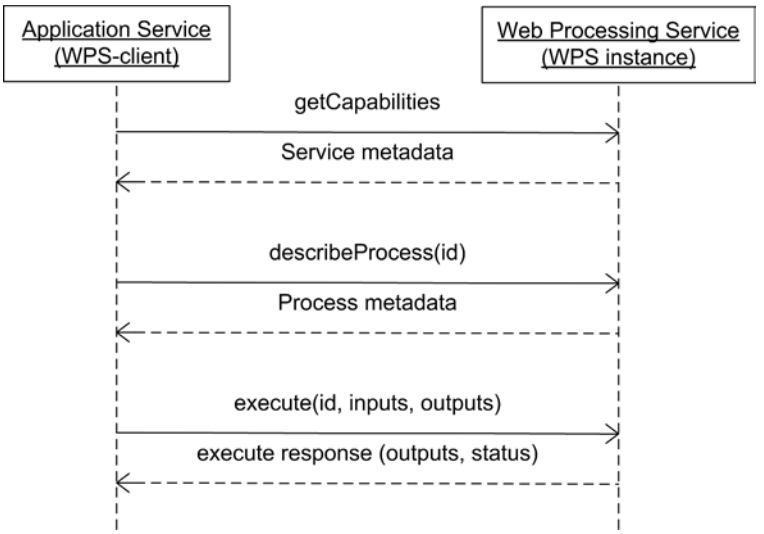
OGC WPS specification provides the service interface definitions to specify a wide range of geospatial processing tasks as geospatial Web services in order to distribute over the Internet many of the functionalities (computation, analysis, etc.) common in today's desktop GIS applications. Geospatial processing services can be considered as being similar to collections of operations in a software component library in the sense of preexisting components that deliver some concrete functionality. The main difference is that WPS can be accessed remotely and can be reused in many different scenarios. This can be achieved by creating accessible libraries of geospatial processing algorithms under the appearance of geospatial Web service chains (Alameh, 2003; Lemmens et al., 2006).

The OGC WPS provides access to calculations or models that operate on spatially referenced data. The data required by the service can be available locally or delivered across a network using data exchange standards such as GML or Geolinked Data Access Service (GDAS). The calculation can be as simple as subtracting one set of spatially referenced numbers from another (e.g., determining the difference in influenza cases between two different seasons) or as complicated as a global climate change model. While most OGC specifications and standards are devoted to geospatial data abstraction, access, and integration, the OGC WPS specification is focused on geospatial data processing of heterogeneous data sources. The main steps in this process are to identify the spatially referenced data required by the calculation, initiate the calculation, and manage the output from the calculation so that it can be accessed by the client. The OGC WPS specification is targeted at both vector and raster data processing.

The basic operational unit of the OGC WPS is the notion of process—a geospatial operation with inputs and outputs of a defined type. This means that a given WPS instance (a concrete WPS service running) may offer one or various operations (or processes) as normal Web services do. Figure 3 shows how a WPS client communicates with a WPS instance, issuing three types of requests. A request can be sent to the WPS instance via HTTP GET with parameters provided as Key-Value Pairs (KVP) or via HTTP POST, with parameters supplied in a XML document. These three types of requests are:

- *getCapabilities:* First, a WPS instance receives a KVP getCapabilities request (which is common for all OWS services) and simply responds with an XML document, containing metadata such as server provider, contact information, general description, and a list of contained geoprocessing operations (processes) offered by the queried WPS instance.
- *describeProcess:* A WPS client selects a process identifier from the getCapabilities response and performs a describeProcess request, either as a KVP or as an XML document. The WPS instance responds with an XML document containing needed information for the solicited process, such as input and output parameter

Figure 3. Synchronous interaction between a WPS-compliant client and a WPS service instance



names and types, so that the WPS client may later build the execute request.

- *Execute:* The WPS client eventually requests the execution of a geospatial operation, with all required input data by invoking the execute method as an XML document request. The WPS instance then runs the operation and returns the results, informing also of its status.

Implementations

This section summarizes some relevant and interesting open source frameworks that currently support one or both available versions (0.4 or 1.0) of the OGC WPS specification.

Cepický and Becchi (2007) introduce the *Python Web Processing Service* (PyWPS), an open source python framework that implements the OGC WPS specification version 0.4.0 (*http://pywps.wald.intevation.org/*). PyWPS includes native support for GRASS (Geographic Supported Analysis Support System, *http://grass.itc.it*) GIS, as well as with the R Project for Statistical Computing (*http://www.r-project.org/*). GRASS GIS is a well-known, powerful GIS tool for geospatial data management and analysis, image processing, graphics/maps production, geospatial modeling, and visualization, while the R Project is a free software environment developed for statistical computing. It is important to highlight that PyWPS allows developers to make native connections to both GRASS GIS and R Project commands, wrapping (or encapsulating) them as contained processes in a given WPS service. This capability fosters the proliferation of distributed geospatial processing services in new domains (environmental, hydrological, etc.) in which distributed geospatial processing services previously were not so easily implemented.

The *Tigris WPSint* implementation (*http://wpsint.tigris.org/*) is an open source Java plug-in for Spring—a Java framework for developing Web applications—to support the OGC WPS version 0.4. This implementation was initially developed by Peter Schut and colleagues during an OGC Interoperability Experiment in order to define initial interfaces and XML schemas for geospatial geoprocessing services, leading then to the first release of the OGC WPS specification in 2005. Contrary to PyWPS, the Tigris WPSint implementation has recently added support for SOAP and WSDL, key Web service components. This feature helps to converge SOA-based services and OGC-based services because both kinds of services may be combined to build heterogeneous service chains since both are described using the same service interface (WSDL).

The *52N Web Processing Service* (52N WPS) is an open source Java framework developed by the 52 North Open Source Initiative (*http://www.52north.org*) that enables the deployment of WPS services. It features a pluggable and extensible architecture for processes and data encodings based on the notions of repositories, which provide dynamic access to the embedded functionality of the WPS already registered in the framework (Foerster, 2006). The current release provides the first attempt to support both the GRASS GIS framework and the WSDL specification. In this sense, the 52N WPS follows the path chosen by PyWPS and Tigris WPSint to support GRASS GIS commands and WSDL/SOAP interfaces respectively, reinforcing the idea that both characteristics are crucial for a widespread use of distributed geospatial processing services in SOA contexts. The benefit of 52N WPS implementation is that it integrates

both capabilities, although they are not (yet) as mature as in the previous two implementations.

The *Deegree* project (*http://www.deegree.org*) is an open source Java framework that implements the OGC WPS as well as traditional OGC services such as WMS, WFS, WCS, and so on. The benefit of Deegree is that it provides the most extensive implementation of OGC standards; however, unfortunately, their WPS implementation seems less mature when compared with the previous WPS implementations. Examples of WPS services using the Deegree project have been reported by Kiehle (2006).

FUTURE TRENDS

The OGC WPS services have been tested in different contexts (Friis-Christensen, Lutz, Ostländer, & Bernard, 2007; Foerster & Shäffer, 2007; Díaz et al., 2008b), illustrating that it is possible to combine several geospatial processing services for accessing, processing, and visualizing data within an SDI. However, many open issues remain regarding the structure and use of the OGC WPS specification itself (Michael & Ames 2007). Other technical and architectural design limitations that constraint the usability, flexibility, and scalability of applications based on distributed geospatial processing services also remain (Friis-Christensen et al., 2007).

One of the most essential problems in implementing distributed geospatial processing services is the overall service chain performance when distributed data sources are involved. This is the case when large processing tasks are performed over the network, because of network bandwidth, data transportation, and data validation. Historically a critical factor of distributed processing has been the network capability or network bandwidth. As GIS resources (inputs and outputs) are by nature large data files, the network bandwidth will always be a limiting factor for successfully distributed geospatial processing. Apart from the bandwidth factor, data transportation and validation (parsing of geospatial data used for the processes) may dramatically increase the response time to users as well. Friis-Christiensen et al. (2007) propose the use of asynchronous messaging to address time-consuming requests. In asynchronous messaging the WPS instance does not return immediately the process results, but rather it responds some time later in a different communication session. This means that the WPS client would not be waiting while the WPS instance is processing a request, but instead it would monitor the process and retrieve the results once the WPS instance has either finished or reported a failure; this essentially means processing results off-line.

Finally other open and challenging issues are enumerated that need further research:

- semantically enriching the descriptions of geospatial processing services by means of geo-ontologies and semantic descriptions that will help to clarify meanings when searching and combining geospatial processing services;
- creating alternative architecture designs and methodologies for chaining geospatial processing services, including in mobile computing contexts;
- creating a mechanism for improving discovery of geospatial processing services;
- using transactional processes;
- improving security; and
- introducing performance and novel techniques for overcoming data transportation issues.

CONCLUSION

The future scenario for geospatial Web services may not ever reach a wholly automated service chaining for a set of self-describing geospatial Web services; however in the near term, semi-automated solutions will emerge to assist users in solving geographical problems with remote services. The geospatial Web services listed in Table 1 mainly deal with the delivery of data instead of advanced processing which is performed online. More heterogeneous, complex geospatial processing services will need to be specified in order to distribute functionalities common in desktop GISs and frameworks such as GRASS GIS and Project R, and make them available over the Internet. The first steps towards distributed, advanced geospatial processing services online are outlined by the recently published OGC Web Processing Service (Schut, 2007), which provides interface specifications to enable geospatial Web services to support a wide range of geospatial processing operations, by creating accessible libraries of geospatial processing algorithms under the appearance of geospatial Web services.

Future research efforts in distributed geospatial processing services should involve new mechanisms for enhancing description and discovery of geospatial processing services, as well as new methodologies for improving composition of geospatial processing services in mobile contexts.

REFERENCES

Alameh, N. (2003). Chaining geographic information Web services. *IEEE Internet Computing, 7*(5), 22-29.

Bernard, L., Kanellopoulos, I., Annoni, A., & Smits, P. (2005). The European geoportal—one step towards the establishment of a European spatial data infrastructure. *Computers, Environment and Urban Systems, 29*(1), 15-31.

Cepický, J., & Becchi, L. (2007). Geospatial processing via Internet on remote servers—PyWPS. *OSGeo Journal, 1*(May).

Retrieved from http://www.osgeo.org/journal/volume1

Cox, S., Daisay, P., Lake, R., Portele, C., & Whiteside, A. (Eds.). (2002). *OpenGIS Geography Markup Language (GML) version 3.0.* Retrieved from http://www.opengeospatial.org/standards/gml

Díaz, L., Granell, C., & Gould, M. (2008a). Spatial data integration over the Web. In V.E. Ferraggine, J.H. Doorn, & L.C. Rivero (Eds.), *Encyclopedia of database technologies and applications* (2nd ed.). Hershey, PA: Information Science Reference.

Díaz, L., Granell, C., & Gould, M. (2008b). Case study: Geospatial processing services for Web-based hydrological applications. In J.T. Sample, K. Shaw, S. Tu, & M. Abdelguerfi (Eds.), *Geospatial services and applications for the Internet.* Berlin: Springer-Verlag.

Foerster, T. (2006). An open software framework for Web service-based geo-processing. *Proceedings of the Free and Open Source Software for Geospatial* (FOSS4G 2006), Lausanne, Switzerland.

Foerster, T., & Shäffer, B. (2007). A client for distributed geo-processing on the Web. *Proceedings of International Symposium on Web and Wireless Geographical Information Systems* (W2GIS 2007) (pp. 252-263), Cardiff, Wales. Berlin: Springer-Verlag (LNCS 4857).

Friis-Christensen, A., Lutz, M., Ostländer, N., & Bernard, L. (2007). Designing service architectures for distributed geoprocessing: Challenges and future directions. *Transaction in GIS, 11*(6), 799-818.

Granell, C., Gould, M., Manso, M.A., & Bernabé, M.A. (2008). Spatial data infrastructure. In H. Karimi (Eds.), *Handbook of research on geoinformatics.* Hershey, PA: Information Science Reference.

ISO 19119. (2005). *Geographic information services.* ISO Technical Committee 211 in Geographic Information/Geomatics.

Kiehle, C. (2006). Business logic for geoprocessing of distributed geodata. *Computers & Geosciences, 32*(10), 1746-1757.

Lemmens, R., Wytzisk, A., de By, R., Granell, C., Gould, M., & van Oosterom, P. (2006). Integrating semantic and syntactic descriptions to chain geographic services. *IEEE Internet Computing, 10*(5), 42-52.

Michael, C., & Ames, D.P. (2007). Evaluation of the OGC web processing service for use in a client-side GIS. *OSGeo Journal, 1.* Retrieved from http://www.osgeo.org/journal/volume1

Papazoglou, M.P. (2008). *Web services: Principles and technology.* Essex: Pearson Education.

Peng, Z.-R., & Tsou, M.-H. (2001). *Internet GIS: Distributed geographic information services for the Internet and wireless networks.* Hoboken, NJ: John Wiley & Sons.

Percivall, G. (Ed.). (2003). *OGC reference model, open geospatial consortium* (doc. no. 03-040). Retrieved from http://portal.opengeospatial.org/files/?artifact_id=3836

Vinoski, S. (1997). CORBA: Integrating diverse applications within distributed heterogeneous environments. *IEEE Communications Magazine, 45*(2), 46-55.

Schut, P. (Ed.). (2007). *OpenGIS Web processing service, version 1.0.0.* Retrieved from http://www.opengeospatial.org/ standards/wps/

Sheth, A.P. (1999). Changing focus on interoperability in information systems from system, syntax, structure to semantics. In M.F. Goodchild, M.J. Egenhofer, R. Fegeas, & C.A. Kottman (Eds.), *Interoperating geographic information systems* (pp. 5-30). Norwell, MA: Kluwer Academic.

Yang, C.P., & Tao, C.V. (2006). Distributed geospatial information. In S. Rana & J. Sharma (Eds.), *Frontiers of geographic information technology* (pp. 103-120). Berlin: Springer-Verlag.

KEY TERMS

Geography Markup Language (GML): An XML grammar defined by OGC to express geographical features. To help users and developers to structure and facilitate the creation of GML-based application, GML provides *GML profiles* that are XML schemas that extend the very GML specification in a modular fashion. A GML profile is a GML subset for a concrete context or application, but without the need for the full GML grammar, simplifying thus the adoption of GML and facilitating its rapid usage. Some common examples of GML profiles that have been published are *Point Profile,* for applications with point geometric data, and *GML Simple Features Profile,* supporting vector feature requests and responses, as in the case of a WFS.

Geospatial Processing Service: Similar to operations in a software library in the sense that these services are preexisting software components that deliver any geospatial processing functionality over the Internet.

ISO/TC211: ISO Technical Committee 211 in Geographic Information/Geomatics is in charge of establishing a set of standards for digital geographic information concerning objects or phenomena that are directly or indirectly associated with a location relative to the earth.

Open Geospatial Consortium (OGC): An international industry consortium participating in a consensus process to develop publicly available interface specifications. OGC members include government agencies, commercial companies, and university research groups.

Service: Functionality provided by a service provider through interfaces (paraphrased from ISO 19119).

Service Broker: Publishes service descriptions and is queried by the service requestor in order to discover suitable services that meet requestor needs.

Service Metadata: Metadata describing the operations and geographic information available at a particular instance of a service (paraphrased from ISO 19119).

Service Provider: Provides software applications as Web services, creating functional descriptions and making them available in public registries.

Service Requestor: Requires certain requirements and needs that are fulfilled by one or more Web services available over the Internet.

Distributed Systems for Virtual Museums

Míriam Antón-Rodríguez
University of Valladolid, Spain

José-Fernando Díez-Higuera
University of Valladolid, Spain

Francisco-Javier Díaz-Pernas
University of Valladolid, Spain

INTRODUCTION

The Internet has meant a social revolution, changing forever the way we communicate and how we access to the information. The growing expansion of technology and the development of easier applications have given as a result the high level of popularity achieved by Internet related services, especially the World Wide Web. Using a hypertext system, Web users can select and read in their computers information from all around the world, with no other requirement than an Internet connection and a browser. For a long time, the information available on the Internet has been a series of written texts and 2D pictures (i.e., static information). This sort of information suited many publications, but it was highly unsatisfactory for others, like those related to objects of art, where real volume and interactivity with the user, are of great importance. Here, the possibility of including 3D information in Web pages makes real sense.

As we become an increasingly visual society, a way to maintain heritage is to adapt museums to new times. The possibility of not only visiting and knowing the museums nearby but also enabling anybody to visit the building from their homes could be enabled. This would imply the incorporation of the virtual reality (Kim, 2005; Vince, 2004), although today only a few museums allow this kind of visit via Internet. In virtual reality, human actions and experiences that interact with the real world are emulated although, obviously, with some limitations. With virtual reality, the user could walk, examine, and interact with the environment, in contrast to traditional media like television that present excellent graphics but lack interactivity. Although this is not a new idea, it is achieving a wider expression due to the availability of software standards like VRML and X3D. VRML, virtual reality modeling language (Carey, Bell, & Marrin, 1997) is a widespread language for the description of 3D scenes and WWW hyperlinks (an analogy of the HTML for virtual reality). X3D, Extensible 3D (Web3D Consortium, 2004) is the successor of VRML, it is intended to be the universal interchange format for integrated 3D graphics and multimedia. VRML/X3D are, perhaps, most interesting to Internet users eager to discover new interesting sites on the Internet, and for the people that use it like a hobby, but those could also allow us to see a 3D artifact from any angle and perspective, to turn it in any way, manipulate it (Lepouras & Vassilakis, 2005; Petridis et al., 2005)—something totally forbidden in a real museum.

This work deals with the design of a system, which allows this interactive Web access to works of art in 3D, as a step in a research project dealing with the design and implementation of a virtual and interactive museum in 3D on the Web. Also, all the associated information like history, architectural data, archaeological data, and culture will be available at the click of a mouse.

BACKGROUND

Several museums around the world are already committed to a strong Web presence and many others will adopt one very soon. Dynamic museum leaders understood that the increasing number of internauts requires special attention from museums: Internet—and CD-ROM's—represent new media that will challenge museum communication strategies.

According to Proença, Brito Ramalho, and Regalo (1998):

Two distinct Web approaches are being adopted by the museums. Some regard their presence on the Web as another way to publicize the museum and to promote their activities; others use the Web as a powerful resource to achieve their purposes: to conserve, to study and to display.

The most common attitude is to consider the Web as a simple sum of the different kinds of information already in use by museums—especially printed information—but gathered in a global structured way. These data include a museum description and a list of activities and collections, where a typical Web page structure contains: collections and exhibitions, visit planning and conditions, new acquisitions, projects and activities, museum organizational schemes, and

educational programs. Several museums on the Web follow this approach. Among them it may be worth a visit to the Museo Arqueológico Nacional of Madrid (http://man.mcu.es/), online Picasso Project (http://csdll.cs.tamu.edu:8080/picasso/), the Asian Art Museum of San Francisco (www.asianart.org), the Museum of Modern Art (www.moma.org), and the Library of Congress Vatican Exhibit (www.ibiblio.org/expo/vatican.exhibit/exhibit/Main_Hall.html); this site has a good image quality, but with a traditional structure to present the exhibition themes.

Some museums demonstrate greater innovation in their Web presences; they have temporary exhibitions online, promote virtual visits and access to their databases, present technical information for museums professionals and researchers, keep available information about previous activities and exhibitions, and organize links to related sites. For these museums, the Web is also an exhibition and a presentation medium that must be integrated in the communication policy of the museum. Among them, it may be worth a visit to the Musée des Beaux Arts de Montréal (www.mbam.qc.ca/en/index.html), the Museum of Anthropology at University of British Columbia (www.moa.ubc.ca/), and the Museo del Prado (http://museoprado.mcu.es/home.html).

Latest advances are becoming popular 3D (plus color) scanners, which allow the measurement of 3D artifacts such as art works (Gómez García-Bermejo, Díaz Pernas, & López Coronado, 1997; Rocchini, Cignoni, Montani, Pingi, & Scopigno, 2001). After measuring, a 3D plus color model from the real object can be obtained. 3D scanning technology has been adopted in a number of recent projects in the framework of cultural heritage. Just to give an example, we may cite the Digital Michelangelo Project of the Stanford University (Levoy et al., 2000) or the acquisition of a section of the Coliseum in Rome. Unfortunately, a detailed 3D (plus color) model of a free form object usually requires a great amount of data. This data can hardly pass through the Web, even when using compression. Therefore, additional reduction of transmission requirements is desirable.

A few years ago, some image-based modeling and rendering techniques were developed making it possible to simulate photo-realistic environments. One of the most popular image-based modeling and rendering techniques is the virtual reality modeling language/extensible 3D (*VRML/X3D*). *VRML/X3D* (Carey et al., 1997; Web3D Consortium, 2004) became an open standard for the delivery of 3D models over the Internet. It combines both geometry and runtime behavioral descriptions into a single file that has a number of different file formats available for it. Last specifications have incorporated latest advances in security (encryption) and speed (compression) based on years of feedback from the VRML97 development community (Li & Kuo, 1998; Matsuba & Roehl, 1999; Taubin, Horn, Lazarus, & Rossignac, 1998).

Using these techniques, some systems allow us to see art works in 3D (Cignoni, Montani, Rocchini, & Scopigno, 2001), while others allow a virtual walk through the rooms of some real buildings such as The Virtual Living Kinka Kuji Tempers (Refsland, Ojika, & Berry, 2000), some have reconstructed scenario such as the Historic Villages of Shirakawa-go (Hirayu, Ojika, & Kijima, 2000), or some imaginary buildings such as Virtual Museum of Helsinki (www.virtualhelsinki.net/museum).

The main feature of our system is that users may walk through a three-dimensional (3D) representation of the whole Fabio Neri's Palace, the building where the Museum of Valladolid is located, viewing its collections, and seeing pictures in 2D and archaeological objects in 3D, together with information about them. To allow all this, an architecture of interactive dynamic Web pages has been designed (Díez-Higuera & Díaz-Pernas, 2002). In order to capture 3D information, we have used the laser acquisition system developed by the Industrial Telematic Group of Telecommunications Engineering School of Valladolid (Gómez et al., 1997). These data, together with 2D images and information files, are compressed and stored in a remote server, and can be retrieved over the Internet. Rather than transmitting a high-resolution object, our system at the client end allows users to selectively retrieve images at specific resolutions. This selective retrieval is achieved by implementing a client-server communication protocol. Information is accessed through intuitive exploration of the site and therefore each session varies depending on both the participant and the path chosen. In this manner, the visitor becomes familiar with the virtual museum in much the same way as they would become familiar with the physical museum. Users may identify particular areas of interest, which may be revisited using familiar routes or accessed via browsing.

Within this framework, a distributed telecommunications system for the remote accessing of multiple virtual environments is implemented. This system is divided into the virtual worlds' systematization and the information's distribution. The objective of the virtual worlds' systematization is not only reducing the development time of new museums but extending the systems' useful life by easing the edition of those virtual worlds already deployed. The distribution of heterogeneous information among different servers gives the users the possibility of visiting multiple online museums within a same virtual reality environment (homogeneous looks) and getting in touch among them in order to contribute with their active participation in the creation of information sharing communities. The server will know at all times the location of the user inside the system to facilitate the communication between different end users. This is possible by an easy-to-use IRC software (Mutton, 2004; Oikarinen & Reed, 1993).

DESCRIPTION OF THE SYSTEM

Figure 1 shows the general architecture of the system. It has two main parts: the dynamic Web pages system based on the *Microsoft Internet Information Server*, which embraces the virtual visit to the 3D museum and the access to data and its 3D visualization; and the platform of telematic services, which implements the server-client architecture, allowing the transmission of three-dimensional and colorimetric data of the objects in several resolutions.

Figure 2 shows the basic communication structures between clients and servers. Server to server communication is done using RMI (*remote method invocation*) while the information exchange between client (applet) and server goes through sockets. This makes necessary the elaboration of an application protocol, as it is shown in Figure 3. The protocol will start working once the browser has loaded the Web page and the corresponding applet. The server application communicating with the applet will be in charge of collecting all clients' requests and sending them, through RMI, to the request serving application, either locally or remotely.

Client-Server Architecture

Servers are used to store the huge amount of scene data that is being requested by the user, while the client is designed to interact with the user and retrieve the necessary scene data from the server. Java programming language has been chosen for the implementation of the server-client architecture. The reason is that Java allows the introduction of executable codes in Web pages, and, therefore, giving the desired portability, security and interactivity to the system.

Figure 1. Global architecture of the proposed system

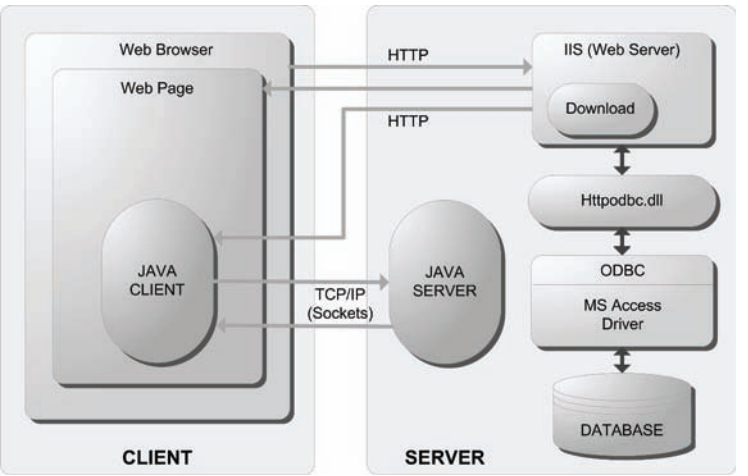


Figure 2. Basic communication structures between clients and server

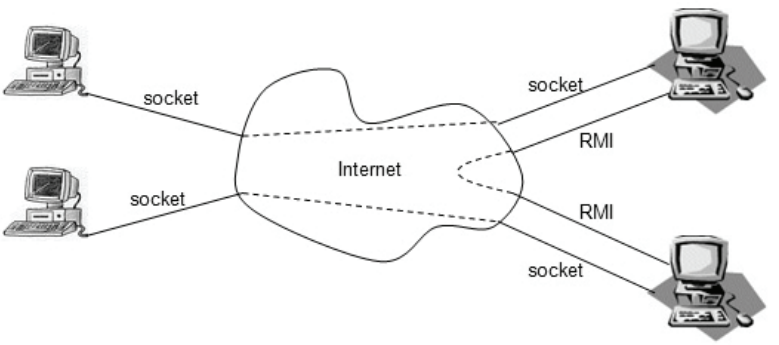
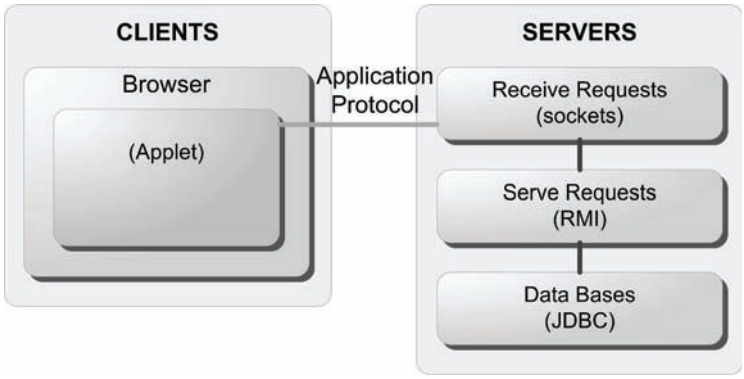


Figure 3. Client-server and server-server communication protocol



Dynamic Web Pages System

A dynamic Web pages system has been implemented to give access to the database so the information required by the user about any object can be shown. This system has also been used for a 3D virtual walk through the museum.

Each section, or room, implemented in the museum becomes a specific *VRML/X3D* file (sometimes, even a single room is divided in several *VRML/X3D* files). Using smart nodes from *VRML/X3D* language, which activate any element in a 3D universe when the user clicks on it, the file stored in the URL of the *VRML/X3D* file code shall be loaded, interconnect these files. Once the user is in a *VRML/X3D* file (containing any of the rooms where both 2D objects--pictures--and 3D ones are displayed) he or she can walk around the room at will: approach any object to have a first impression and, if wanted, click on it to acquire information about it (author, period, technique, etc.), as well as to visualize it in 3D (or with a higher resolution still image if the object is a picture). When the user clicks on any object, the dynamic Web pages system starts to work, giving access to the database and bringing the required information.

Format of the Data Used by the System

VRML/X3D is the working format for the 3D display of the objects, as it is standard and commonly used in the Web (Walsh & Bourges-Sevenier, 2000). First of all, 3D plus color data must be acquired from art works. The corresponding object model, obtained from the acquired data, could then be directly expressed in *VRML/X3D* format. However, the raw format of the data in our database is not *VRML/X3D*. Instead, it is the specific one given by the laser acquisition system (Gómez et al., 1997). This system, starting from a real object, gives a *.pie* file with the three-dimensional data, plus another three files (*.r, .g, and .b*) with its colorimetric information. Those files are converted into ASCII files using software developed by the industrial telematic group; these new files are *.pts*, with the Cartesian coordinates of all the points as read from the object, plus its colors, and *.tri*, with the triangular distribution of points for a certain resolution (which allows the construction of faces for the display of the object).

The servers store the last couple of files: one *.pts* file for each object, and several *.tri* (one for each different resolution). From them, and given a certain resolution, the server obtains the files needed by the client to reconstruct the *VRML/X3D* file. It is important to notice that there is no duplication of information in the server, as points and colors appear only once in the *.pts* file and the set required by the client at each moment is obtained from it. There are only several files of triangles, as the distribution of points on the faces is different for each resolution. Moreover, three-dimensional and colorimetric information is sent only once to the client. The result is that this design, as a whole, improves the efficiency of the system.

Also, we have developed an alternate solution, which allows lower transmission requirements. In short, we allow art works to be requested in different levels of detail (LOD). User begins with the lowest LOD model, and requests a set of progressively increasing LOD models as his interest on the object increases. We benefit from this by building the different LOD models in particular way: each LOD model is obtained from the immediately higher LOD model, by just picking some of its points. In this way, when the users ask for a higher LOD model, the whole model transmission is no more required. Instead, the new model can be reconstructed by adequately merging new points into the previously existing model. Unfortunately, this strategy is not implemented in actual *VRML/X3D*; so we have implemented it by using a dedicated Java client. Basically, when the user asks for a superior LOD model, only additional 3D points, and a new faces description, are sent. This faces description must then be translated by the client. This is done by means of a LUT

(*Look-Up Table*), which indicates actual position of the new points in a local *VRML/X3D* file.

User Interface

The user interface module is based on a Web page template in which the virtual world is represented on a part of the screen, while on the rest the name of the content plus the options that the user can select are left. In the main part of the page, visiting users can get to move all through the building, its rooms and exhibits, and once inside them, the user can select a picture or a 3D artifact in order to get information about it (Figures 4, 5, 6, and 7). Furthermore, users will be able to access all the available options using the applet on the left, features like accessing different museums, logging as a registered user which will allow the system to know information and the location of all users connected to it, thus facilitating the communication among them, for example using an IRC software (Figures 8, 9), conducting distributed information searches, and more. People in charge of the management will be able to include links to new museums as well as modifying existing ones (including new pieces or roving exhibits) as shown in Figure 10.

Figure 4. Initial Web page of the virtual museum: View of Fabio Neri's Palace

Figure 5. General view from the patio

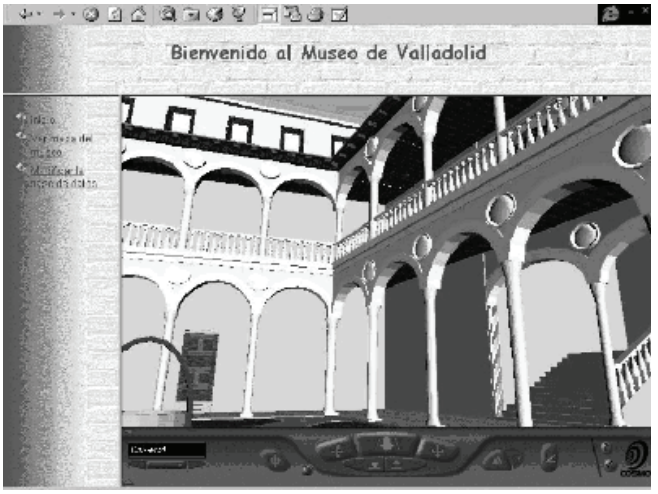


Figure 6. Initial view of the virtual room

Figure 7. An example of 3D artifact

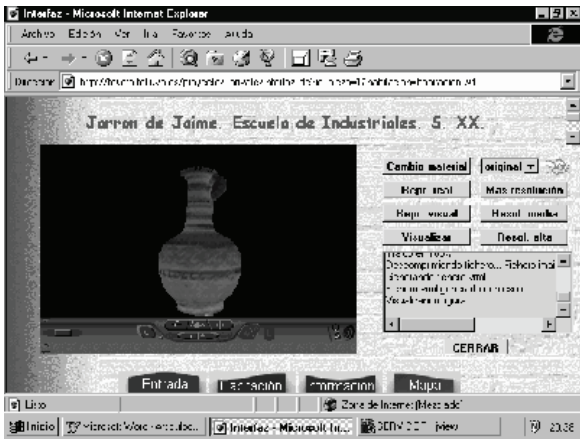


Figure 8. Sample access to the info of a connected user

Figure 9. Use example of the IRC developed to get in touch the users of the system

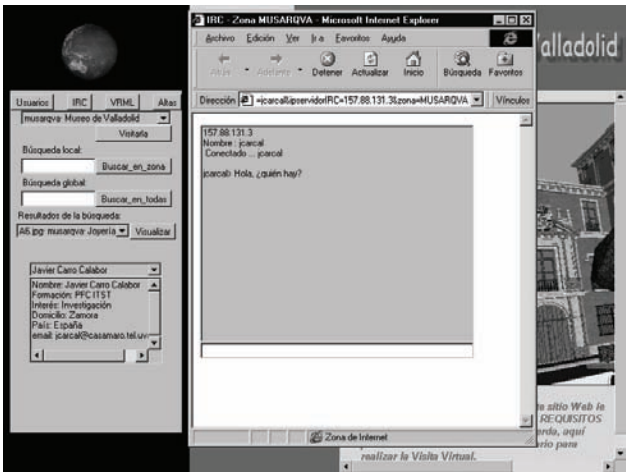
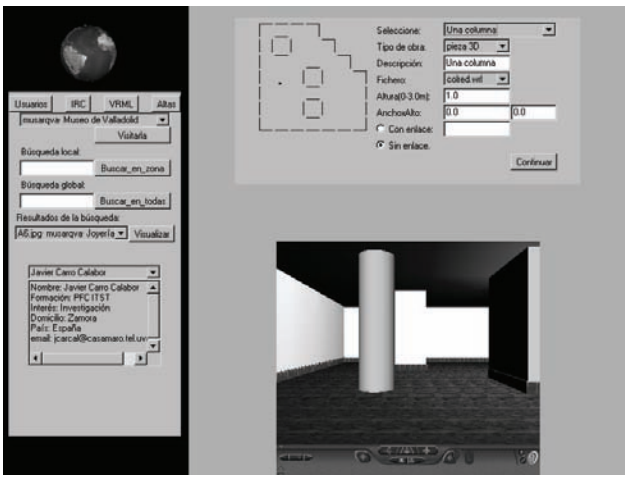


Figure 10. Tool developed to create, manage and modify virtual reality environments



FUTURE TRENDS

Our system marks the advent of a new and interactive form of documenting, referencing, and archiving heritage structures. Future work involves advanced development of this system as follows:

- Include dynamic elements like 3D surround sound, voice-overs for history, culture, and other information along with traditional audio pieces to make a complete user-friendly interactive interface.
- Increase the number of works of art and information in texts in order to give a more detailed presentation of Archaeology in Spain.
- Incorporate accurate details by using the latest photogrammetric techniques.
- Depict "as was," "as is," and "as it could be" transitions for studies specific to architecture, conservation, games for children and light simulations.

CONCLUSION

Our project resulted in the following observations:

- We achieved a fairly accurate 3D model of a complex heritage structure using archaeological orthographic projections as reference material.

- Construction of a virtual museum is possible, allowing users to examine, conduct research, or navigate any part of the building at their own convenience where they not only can see photos, and 3D objects, and even have the opportunity to play with them.
- Homogeneous access to a set of museums, either self-managed or managed by a third entity, is offered by this application, allowing for a greater content diversity.
- Fast and easy access to the information kept in different virtual museums, even if those museums are deployed in distinct servers, and so, allowing access to distributed information.
- User can navigate through the virtual environment of a museum and, in any moment, can request information from other museums according to a search term.
- An appropriate management of users' information allows the system to know the number of visitors inside the museum at any moment, allowing users to be in contact which each other—using a custom IRC software—and this way the museum not only becomes a space for collecting and showing valuable pieces but also a catalyst to create information sharing communities.
- The developed system is easily updatable and scalable. Any room's content can be modified and even new rooms can be added or removed to any museum.
- The viability of the system has been demonstrated, as well as its correct operation in the net in the particular case of the Museum of Valladolid.

REFERENCES

Carey R., Bell, G., & Marrin, C. (1997, April). The virtual reality modeling language. ISO/IEC DIS 14772-1. Last Update: December 2003 by Web3D Consortium. Retrieved from http://www.web3d.org/x3d/specifications/vrml/ISO-IEC-14772-VRML97/

Cignoni, P., Montani, C., Rocchini, C., & Scopigno, R. (2001, July). Acquisition and management of digital 3D models of statues. In *Proceedings of the 3rd International Congress on Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin, Alcalá de Henares (Spain)* (pp. 1159-1164).

Díez-Higuera, J. F., & Díaz-Pernas, F. J. (2002). VRML-based system for a three-dimensional virtual museum. In T. K. Shih (Ed.). *Distributed multimedia databases: Techniques & applications* (pp. 306-317). Hershey, PA: Idea Group Publishing.

Gómez G. J., Díaz Pernas, F. J., & López Coronado, J. (1997, April). Industrial painting inspection using specular sharpness. In *Proceedings of the International Conference on Recent Advances in 3-D Digital Imaging and Processing (IEEE Press)* (pp. 335-338). Otawa, Canada.

Hirayu, H., Ojika, T., & Kijima, R. (2000). Constructing the historic villages of Shirakawa-go in virtual reality. *IEEE Multimedia, 7*(2)**,** 61-64.

Kim, G. J. (2005). *Designing virtual reality systems*. Springer.

Lepouras, G., & Vassilakis, C. (2005, June). Virtual museums for all: Employing game technology for edutainment. *Virtual Reality, 8*(2), 96-106.

Levoy, M., Pulli, K., Curless, B., Rusinkiewicz, S., Koller, D., Pereira, L., Ginzton, M., Anderson, S., Davis, J, Ginsberg, J., Shade, J., & Fulk, D. (2000, July 24-28). The digital Michelangelo project: 3D scanning of large statues. *Comp. Graph. Proceedings, Annual Conference Series (Siggraph '00), ACM SIGGRAPH* (pp. 131-144), 2000. Addison Wesley.

Li, J. K., & Kuo, C. J. (1998, June). Progressive coding of 3D graphics models. In *Proceedings of IEEE, 86*(6), 1052-1063.

Matsuba, S. N., & Roehl, B. (1999, Spring). Bottom, thou art translated: The making of VRML dream. *IEEE Computer Graphics and Applications, 19*(2), 45-51.

Mutton, P. (2004, July). *IRC Hacks*. O'Reilly.

Oikarinen, J., & Reed, D. (1993). *RFC 1459: Internet relay chat protocol*. Network Working Group.

Petridis, P., White, M., Mourkousis, N., Liarokapis, F., Sifniotis, M., Basu, A., & Gatzidis, C. (2005, March). Exploring and interacting with virtual museums. In *Proceedings of Computer Applications and Quantitative Methods in Archaeology* (*CAA*), Tomar, Portugal.

Proença, A., Brito, M., Ramalho, T., & Regalo, H. (1998). Using the Web to give life to Museums CD-ROM. In *Proceedings of the Museums and the Web*. Toronto, Canada.

Refsland, S. T., Ojika, T., & Berry Jr., R. (2000). The living virtual Kinka Kuji temple: A dynamic environment. *IEEE Multimedia Magazine, 7*(2), 65-67.

Rocchini, C., Cignoni, P., Montani, C., Pingi, P., & Scopigno, R. (2001). A low cost scanner based on structured light. *Computer Graphics Forum. Eurographics 2001 Conference Proceedings, 20*(3), 299-308.

Taubin, G., Horn, W. P., Lazarus, F., & Rossignac, J. (1998, June). Geometry coding and VRML. In *Proceedings of IEEE, 86*(6), 1228-1243.

Vince, J. (2004). *Introduction to virtual reality*. Springer.

Walsh, A. E., & Bourges-Sevenier, M. (2000, September) Core Web3D. Pearson Education.

Web3D Consortium. (2004). Extensible 3D (X3D) - Part 1: Architecture and base components. ISO/IEC 19775-1:2004. Last Update: November 2005. Retrieved from http://www.web3d.org/x3d/specifications/ISO-IEC-19775-X3DAbstractSpecification/Part01/Architecture.html

KEY TERMS

Dynamic HTML: A collective term for a combination of new HTML tags and options, style sheets and programming, which enable you to create Web pages that are more interactive and faster to download.

Internet Relay Chat (IRC): A form of instant communication over the Internet. It is mainly designed for group (Many-to-many) communication in discussion forums called channels, but also allows one-to-one communication.

Java: A platform-independent programming language, produced by Sun Microsystems. Java is built as a method to provide services over the WWW. With Java, a Web site provides a Java application (called an applet) which is downloaded by the client and executed on the client machine. Java is specifically built so that an application can be run on any kind of system.

Virtual Museum: A collection of digitally recorded images, sound files, text documents, and other data of historical, scientific, or cultural interest that are accessed through electronic media. A virtual museum does not house actual objects and therefore lacks the permanence and unique qualities of a museum in the institutional definition of the term.

Virtual Reality (VR): The use of computer modeling and simulation to enable a person to interact with an artificial three-dimensional visual or other sensory environment. VR applications immerse the user in a computer-generated environment that simulates reality through the use of interactive devices, which send and receive information and are worn as goggles, headsets, gloves, etc.

VRML (Virtual reality modeling language): A programming language for the creation of virtual worlds. Using a VRML viewer, you can take a virtual tour of a 3D model building, or manipulate animations of 3D objects. Hyperlinks to other sites and files can be embedded in the world you visit.

X3D: Extensible 3D (X3D) is a software standard for defining interactive Web- and broadcast-based 3D content integrated with multimedia. X3D is the successor to the Virtual Reality Modeling Language (VRML), the original ISO standard for Web-based 3D graphics (ISO/IEC 14772). It improves upon VRML with new features and a componentized architecture that allows for a modular approach to supporting the standard.

A Duplicate Chinese Document Image Retrieval System

Yung-Kuan Chan
National Chung Hsing University, Taiwan, R.O.C.

Yu-An Ho
National Chung Hsing University, Taiwan, R.O.C.

Hsien-Chu Wu
National Taichung Institute of Technology, Taiwan, R.O.C.

Yen-Ping Chu
National Chung Hsing University, Taiwan, R.O.C.

INTRODUCTION

An optical character recognition (OCR) system enables a user to feed an article directly into an electronic computer file and translate the optically scanned bitmaps of text characters into machine-readable codes; that is, ASCII, Chinese GB, as well as Big5 codes, and then edits it by using a word processor. OCR is hence being employed by libraries to digitize and preserve their holdings. Billions of letters are sorted every day by OCR machines, which can considerably speed up mail delivery.

The techniques of OCR can be divided into two approaches: template matching and structure analysis (Mori, Suen & Yamamoto, 1992). The template matching approach is to reduce the complexity of matching by projecting from two-dimensional information onto one; the structure analysis approach is to analyze the variation of shapes of characters. The template matching approach is only suitable for recognizing printed characters; however, the structure analysis approach can be applied to recognize handwritten characters.

Several OCR techniques have been proposed, based on statistical, matching, transform and shape features (Abdelazim & Hashish, 1989; Papamarkos, Spilioties & Zoumadakis, 1994). Recently, integrated OCR systems have been proposed, and they take advantage of specific character-driven hardware implementations (Pereira & Bourbakis, 1995). OCR generally involves four discrete processes (Khoubyari & Hull, 1996; Liu, Tang & Suen, 1997; Wang, Fan & Wu, 1997):

1. separate the text and the image blocks; then finds columns, paragraphs, text lines, words, and characters;
2. extract the features of characters, and compare their features with a set of rules that can distinguish each character/font from others;
3. correct the incorrect words by using spell checking tools; and
4. translate each symbol into a machine-readable code.

The duplicate document image retrieval (DDIR) system transforms document formatted data into document images, then stores these images and their corresponding features in a database for the purpose of data backup. The document images are called duplicate document images. When retrieving a duplicate document image from the database, users input the first several text lines of the original document into the system to create a query document image. Then the system figures out the features of the image, and transmits to the users the duplicate document image whose image features are similar to those of the query document image (Nagy & Xu, 1997).

Some approaches have been proposed for the DDIR system. Doermann, Li, and Kia (1997) classified and encoded character types according to the condition that four base lines cross each text line, and uses the codes as the feature of the document image. Caprari (2000) extracted a small region from one document, assigned this region to the template (signature generation), and then scanned this template over a search area in another document. If the template also appears in the second document (signature matching), the two documents are classified as duplicates. Angelina, Yasser, and Essam (2000) transformed a scanned form into a frameset composed of a number of cells. The maximal grid encompassing all of the horizontal and vertical lines in the form is generated; meanwhile, the number of cells in the frameset, where each cell was created by the maximal grid, was cal-

culated. Additionally, an algorithm for similarity matching of document framesets based on their grid representations is proposed too. Peng, Long, Chi, and Siu (2001) used the size of each component block containing a paragraph text image in a duplicate document image and its relative location as the features of the duplicate document image.

The approaches mentioned previously are only suitable for stating the characteristics of an English document image. The characteristics of Chinese characters are quite different from those of English ones, and the strokes and shapes of Chinese characters are much more complicated than those of English characters. Chan, Chen, and Ho (2003) provided a line segment feature to represent a character image block and presented a duplicate Chinese document image retrieval (DCDIR) system based on this feature. The purpose of this short article is to give a brief overview of the duplicate Chinese DDIR systems.

BACKGROUND

Traditional information retrieval methods use keywords for textual databases. However, it is difficult to describe an image using exact information, and defining manually keywords is tedious or even impossible for a large image database. Moreover, some non-text components cannot be represented in a converted form without sufficient accuracy. One solution is to convert a document into digital images; meanwhile, some methods are applied to extract the features of the images. Based on the feature, some document images with database satisfying query requirements are returned.

A duplicate document image retrieval (DDIR) system has to own the following properties (Doermann, Li, & Kia, 1997):

- **Robust:** The features should be reliably extracted even when the document becomes degraded.
- **Unique:** The extracted features can distinguish each document image from others.
- **Compact:** The storage capacity required to hold the features should be as small as possible.
- **Fast:** The system needs a quick response with an answer to the query.
- **Scalable:** As more documents are processed, the size of the database could grow to tens of millions.
- **Accurate:** The system should accurately response with an answer, which satisfies the query requirement.

Unfortunately, many DDIR systems are vulnerable to poor qualities of document images, such as the scale, translation, rotation, and noise variants. Because of different resolution setup of a scanner, the same image may be scanned to become two images with different sizes. We call this phenomenon the scale variant. When an image is added with a great amount of noises, it may be regarded as a different image from the original one. It is named a noise variant image of the original one. In a particular document, images with rotation and translation variants may be generated owing to placing the document on different orientation angles or on different positions on a scanner. The variants mentioned previously will cause many troubles in feature extracting and image matching stages. They should be removed in advance.

A CHINESE DDIR SYSTEM

Many techniques about the DDIR system have been proposed (Caprari, 2000; Doermann, Li, & Kia, 1997; Peng, Chi, Siu, & Long, 2000; Peng, Long, Chi, & Siu, 2001). Since an English document mostly consists of approximately **70** commonly-used characters which contain **52** uppercase as well as lowercase English letters and punctuation marks, the classification and encoding procedure based on the feature of these characters' font types are possible. However, these techniques are only suitable for duplicate English document images, but not for duplicate Chinese document image retrieval (DCDIR) because the number of different Chinese characters is about **45,000**. What is more, the shapes of Chinese characters are complex, and many different characters have similar shapes to each other. Hence, there are several major problems with Chinese character recognition, that is, Chinese characters are distinct and ideographic, the size of a character is large, and there exist many structurally similar characters (Amin & Singh, 1996; Chan, Chen, & Ho, 2003).

It is necessary to develop a feature offering an excellent identification capability to classify Chinese characters by only using a little extra memory space. To reduce the extra memory space, it is feasible to segment a duplicate document image into blocks, each of which contains a set of adjacent characters, and then to extract the features from the blocks. Since the number of the blocks in a duplicate document image is much smaller than that of the characters in an identical duplicate document image, the feature dimensions are reduced greatly; however, its identification capability is lessened.

I. DCDIR System

The proposed duplicate document image retrieval system approximately includes three parts — image preprocessing, database creation, and document retrieval. This section will introduce these three parts in details.

A. Image Preprocessing

When scanning a document to generate a duplicate document binary image, the position of the document on the

scanner may be misplaced so that the duplicate document image may become inclined. Figure 1(a) shows an original image and Figure 1(b) is its duplicate document image that appears inclined. The inclined condition of a document image may lead to inconvenience to users and cause the errors in extracting its image features. Peng et al. (2000) used a correlation-based method to detect the inclination of an image, and then applied an interpolation technique to turn the image back according to the detected inclination. The DCDIR system will use this technique to turn the inclined document image back. Figure 1(c) is the duplicate document image after adjusting the inclination.

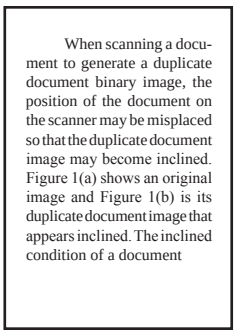
As in Figure 1(c), after turning back the duplicate document image, the frame of the duplicate document image will become inclined. It is necessary to cut off the border blank of the document image. While removing the border blank, the system starts scanning the duplicate document image from the left-top pixel. Then, in the order from left to right and top to bottom, each pixel is scanned until one certain black pixel ***P*** is found. Finally, all pixels locating on the lines, which are prior to the line containing ***P*** are removed to cut off the top border blank of the document image. By using the same method, the bottom, left, and right border blanks of the document image are removed as well. Figure 1(d) demonstrates the final duplicate document image after cutting off the border blanks of the document image as illustrated in Figure 1(c).

B. Database Creation

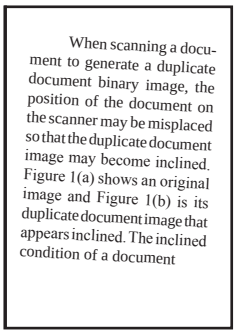
After that, the DCDIR system extracts the character image features from the duplicate document image ***I*** in which its border blanks have been cut off and the system stores the features in the database. Before extracting the character image features of ***I***, the system first performs the text line segmentation on ***I*** to make every line image block contain only the complete image of one certain line text in the original document. Then, the system segments out all character image blocks from each previously segmented line image block so that every character image block contains only one Chinese character. Finally, the feature of each character image block is then extracted.

Figure 1. Normalization of an inclined document image

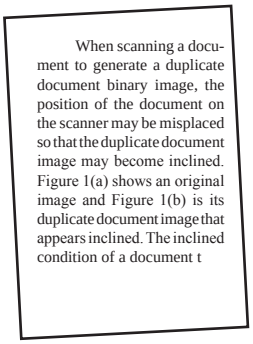
(a) Original document



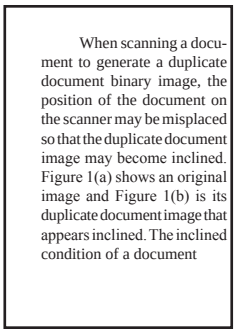
(b) Duplicate document image after scanning



(c) Duplicate document image after inclination adjusting



(d) Duplicate document image after cutting off the border blanks



Concerning the steps of segmenting line image blocks from a duplicate document image, first, all of the black pixels in the duplicate document image are projected in horizontal direction onto a projection vertical axis. The length of a black section on the projection vertical axis is just the height of the corresponding text line containing those character images whose black pixels are projected onto the black section.

Next, all of the black pixels in every line image block is projected in vertical direction onto a certain projection horizontal axis. In this case, the distribution borders of the black pixels and the white pixels on the projection horizontal axis are the locations of the left and the right boundaries of the character image blocks. On the projection horizontal axis, the length of a black section is just the width of the corresponding character image block whose black pixels are projected onto the black section.

The sizes of most Chinese characters are close. When the height of a certain character image block ***CB*** is smaller than three-fourths of the average height of all character image blocks in the document image, and the width of ***CB*** is also smaller than three-fourths of the average width of all character image blocks, the system will then regard the character in ***CB*** as a noise, and then remove it.

After that, three horizontal scanning lines are drawn on each character image block. These three horizontal scanning lines are respectively located at ***1/4×H***, ***2/4×H*** and ***3/4×H*** character heights in the block. Here ***H*** represents the height of the character image block. According to the ratio of the total number of the black pixels to that of the white pixels, which the scanning line goes through, an encoding process is executed to reduce the memory space required to store the feature of the character image block. The way of encoding is shown as follows

$$X_i = \begin{cases} 0, \text{ if } D_{i,b} \times m > D_{i,w} \\ 1, \text{ if } D_{i,b} \times m \le D_{i,w} \end{cases} \text{, for } i = 0, 1, \text{ and } 2.$$

In this equation, $D_{i,w}$ and $D_{i,b}$ are respectively the total numbers of the white pixels and the black pixels that the ***i***-th scanning line passes through, and ***m*** is the weight (a given constant value) for the ratio from the total numbers of the black pixels and the white pixels on the scanning line. Thus, each character image block can be represented by a three-bit ($X_0X_1X_2$) code; we name the code the feature code of the character image block. There are ***8*** different binary codes ***000***, ***001***, ***010***, ***011***, ***100***, ***101***, ***110***, and ***111*** corresponding to decimal feature codes ***0, 1, …,*** and ***7*** respectively.

Because the resolution setup of a scanner may be different, the same original document may be scanned to become duplicate document images of different sizes. This proposed feature adopts the ratio from the number of black pixels and white pixels, so the feature encoding will not be affected due to the scale variant of images. Moreover, the desired duplicate document image and the query document image are both from the same original document. Therefore, the problem that the font type and style of the characters in the query document are different from those in the desired duplicate document image will not occur in this system.

C. Document Retrieval

Let $\boldsymbol{Q} = q_1q_2\ldots q_l$ be the feature code of query document image $\boldsymbol{I}_q$, and the length of the feature code be ***l***. Next, the system extracts the first feature codes with the length ***l*** from every duplicate document image $\boldsymbol{I}_d$ in the database. Let the extracted feature codes be $\boldsymbol{D} = d_1d_2\ldots d_l$. Then, the system compares the corresponding bit pair q_i and d_i between ***Q*** and ***D*** from left to right, respectively. When $q_i = d_i$, the system adds ***1*** to the value of ***S***. The final value of ***S*** is the similarity between $\boldsymbol{I}_q$ and $\boldsymbol{I}_d$. Finally, the duplicate document image with the largest similarity value is found out.

II. Experiments

Experiment 1 is to explore the constant weight ***m***. As for different values of ***m***, the character image blocks of ***5401*** in commonly used Chinese characters among Big5 codes, are categorized into eight groups each of which corresponds to one feature code. Table 1 shows the number of members in each group for ***m = 2***, ***3***, and ***4***, where σ_c^2 is the variance of the number of members of the eight groups. The experimental results shows that when ***m = 3***, σ_c^2 is minimal. This means, when ***m = 3***, all Chinese character image blocks are most uniformly mapped to various kinds of feature codes. The next experiment will set ***m = 3***.

Experiment 2 is to investigate the performance of the DCDIR system. This experiment scans each page of the book "朝花夕拾、呐喊" with ***336*** sheets to become images by a scanner. This experiment rescans ***101*** sheets of the book to generate the query document images. Here, the first ***L*** text lines of the ***101*** document sheets are respectively used as the contents of the query document images. Table 2 shows the experimental results. The average searching time is approximately ***8*** seconds for each query.

FUTURE TRENDS

After a paper document is used over a long period, the document may be stained or worn out, so that its contents may be indistinct. How to develop an effective image feature insensitive to the rotation, scale, translation, and noise variations is an important task in the future. Many documents are printed or handwritten texts, are multilingual, or are composed of basic composition style and font for text. A future document image indexing method should own the robustness of the above variants.

Table 1. Results of the first experiment

Feature code \ m	2	3	4
000	372	1253	2266
001	596	773	763
010	262	387	402
011	813	525	312
100	337	564	628
101	817	591	324
110	390	374	262
111	1798	918	428
S_c^2	220549	74523	387785

Moreover, the following topics have been explored, and will continue to be researched. The first is to find or to match the instances of a document image with known content in a database. The techniques can be applied to maintaining database integrity by eliminating duplicates and retrieval itself. The second is to index image captions and to establish a relationship between the content and the images they describe. Then the caption can be a valuable tool of their duplicate document images. Ideally, a duplicate detection algorithm can find both exact duplicates which have just the same content, and partial duplicates, which have a large percentage of their text in common. Locating exact duplicates could reduce the storage required for a large database. Finding partial duplicates will allow users to easily find other versions of a given document.

REFERENCES

Abdelazim, H.Y., & Hashish, M.A. (1989). Automatic reading of bilingual typewritten text. *Proceeding of VLSI and Microelectronic Applications in Intelligent Peripherals and their Application Network,* 2.140-2.144.

Amin, A., & Singh, S. (1996). Machine recognition of hand-printed Chinese characters. *Intelligent Data Analysis*, *1*, 101-118.

Angelina, T., Yasser E.S., & Essam A.E.K. (2000). Document image matching using a maximal grid approach. *Proceedings of SPIE on Document Recognition and Retrieval IX,* 4670, 121-128.

Caprari, R.S. (2000). Duplicate document detection by template matching. *Image and Vision Computing, 18*(8), 633-643.

Chan, Y.K., Chen, T.S., & Ho, Y.A. (2003). A duplicate Chinese document image retrieval system based on line segment feature in character image block. In S. Deb (Ed.), *Multimedia systems and content-based image retrieval, (pp.14-*23). Hershey, PA: Idea Group Publishing.

Doermann, D., Li, H., & Kia, O. (1997). The detection of duplicates in document image databases. *Image and Vision Computing, 16*(12-13), 907-920.

Khoubyari, S., & Hull, J.J. (1996). Font and function word identification in document recognition, *Computer Vision and Image Understanding, 63*(1), 66-74.

Liu, J., Tang, Y.Y., & Suen, C.Y. (1997). Chinese document layout analysis based on adaptive split-and-merge and qualitative spatial reasoning. *Pattern Recognition, 30*(8), 1265-1278.

Mori, S., Suen, C., & Yamamoto, K. (1992). Historical review of OCR research and development. *Proceedings of the IEEE, 80*(7), 1029-1058.

Table 2. Results of Experiment 2

Text lines / experimental results	*L*=3	*L*=5	*L*=6	*L*=7	*L*=10
Correctly finding out desired images	99	100	100	101	101
Accuracy rate (%)	98.0	99.0	99.0	100	100

Nagy, G., & Xu, Y. (1997). Bayesian subsequence matching and segmentation. *Pattern Recognition Letters, 18*(11-13), 1117-1124.

Papamarkos, N., Spilioties, I., & Zoumadakis, A. (1994). Character recognition by signature approximation. *International Journal of Pattern Recognition and Artificial Intelligence, 8*(5), 1171-1187.

Peng, H., Chi, Z., Siu, W.C., & Long, F. (2000). PageX: An integrated document processing software for digital libraries. *Proceedings of the International Workshop on Multimedia Data Storage, Retrieval, Integration, and Applications, Hong Kong,* (pp.203-207).

Peng, H., Long, F., Chi, Z., & Siu, W.C. (2001). Document image template matching based on component block list. *Pattern Recognition Letters, 22*(9), 1033-1042.

Pereira, N., & Bourbakis, N. (1995). Design of a character driven text reading system. *Proceedings of "SPIE", San Jose, California,* (p.6-9).

Wang, A.B., Fan, K.C., & Wu, W.H. (1997). Recursive hierarchical radical extraction for handwritten Chinese characters. *Pattern Recognition, 30*(7), 1213-1227.

KEY TERMS

Character Segmentation: The technique that partitions images of lines or words into individual characters.

Content-Based Image Retrieval (CBIR): The technique of image retrieval based on the features automatically extracted from the images themselves.

Duplicate Document Detection: The technique to find the exact duplicates, which have exactly the same content, or partial duplicates which have a large percentage of their text in common.

Duplicate Document Image Retrieval (DDIR): A system for finding the image-formatted duplicate of documents from a database.

MFU: Most frequently used characters in a character set with enormous members.

Optical Character Recognition (OCR): The technique of automatically translating the content of an image formatted document into text-formatted materials.

Skew Correction: The technology detects and compensates for the inclination angle of a document image.

Template Matching: The approach involves designing template masks, which are capable of detecting incidences of the relevant feature at different orientations.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1-6, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Dynamic Taxonomies for Intelligent Information Access

Giovanni M. Sacco
Università di Torino, Italy

INTRODUCTION

End-user interactive access to complex information is a key requirement in most applications, from knowledge management, to e-commerce, to portals. Traditionally, only access paradigms based on the retrieval of data on the basis of precise specifications have been supported. Examples include queries on structured databases and information retrieval. There is now a growing perception that this type of paradigm does not model a large number of search tasks, such as product selection in e-commerce sites among many others, that are imprecise and require exploration, weighting of alternatives and information thinning. The recent debate on findability (Morville, 2002) and the widespread feeling that "search does not work" and "information is too hard to find" shows evidence of the crisis of traditional access paradigms.

New access paradigms supporting exploration are needed. Because the goal is end-user interactive access, a holistic approach in which modeling, interface and interaction issues are considered together, must be used and will be discussed in the following.

BACKGROUND

Four retrieval techniques are commonly used: (a) information retrieval (IR) systems (van Rijsbergen, 1979), also search engines; (b) queries on structured databases; (c) hypertext/hypermedia links and d) static taxonomies, such as Yahoo!.

IR systems exhibit an extremely wide semantic gap between the user model (concepts) and the model used by commercial retrieval systems (words). This leads to a significant loss of relevant information (Blair & Maron, 1985), and to poor user interaction because query formulation is difficult and no or very little assistance is given. In addition, because results are presented as a flat list with no systematic organization, no exploration is possible. Database queries require structured data and are not applicable to situations in which information are textual and not structured or loosely structured. Exploration is usually limited to sorting flat result lists according to different ordering criteria.

Hypermedia techniques (Groenbaek & Trigg, 1994) have become pervasive and support exploration. However, they do not support abstraction so that exploration is performed one-document-at-a-time, which is quite time consuming. Building and maintaining nontrivial hypermedia networks is very expensive.

Traditional taxonomies are based on a hierarchy of concepts that can be used to select areas of interest and restrict the portion of the infobase to be retrieved. They are easily understood by end-users, but they are not scalable for large information bases (Sacco, 2006b), so that the average number of documents retrieved becomes rapidly too large for manual inspection.

A more recent approach is the Semantic Web (Berners-Lee, Hendler, & Lassila, 2001). Although one of the driving forces behind it is retrieval, the general semantic schemata proposed are intended for programmatic access and are known to be difficult to understand and manipulate by the casual user. User interaction must be mediated by specialized agents, which increases costs, time to market and decreases the transparence and flexibility of user access.

DYNAMIC TAXONOMIES

Dynamic taxonomies (Sacco, 1987, 2000), also called *faceted search systems*, are a general knowledge management model based on a multidimensional classification of heterogeneous data items and are used to explore/browse complex information bases in a guided yet unconstrained way through a visual interface.

The intension of a dynamic taxonomy is a taxonomy designed by an expert. This taxonomy is a concept hierarchy going from the most general to the most specific concepts. A dynamic taxonomy does not require any other relationships in addition to *subsumptions* (e.g., IS-A and PART-OF relationships). Directed acyclic graph taxonomies modeling multiple inheritance are supported but rarely required.

In the extension, items can be freely classified under n (n>1) concepts at any level of abstraction (i.e., at any level in the conceptual tree). The multidimensional classification required by dynamic taxonomies is a generalization of the monodimensional classification scheme used in conventional taxonomies and models common real-life situations. First, items are very often about different concepts: for example, a news item on September 11th, 2001 can be classified under "terrorism," "airlines," "USA," and so forth. Second, items to be classified usually have different features, "perspectives"

or facets (e.g., Time, Location, etc.), each of which can be described by an independent taxonomy.

In dynamic taxonomies, a concept C is just a label that identifies all the items classified under C. Because of the subsumption relationship between a concept and its descendants, the items classified under C (items(C)) are all those items in the *deep extension* of C, that is, the set of items identified by C includes the *shallow extension* of C (i.e., all the items directly classified under C) union the deep extension of C's sons. By construction, the shallow and the deep extension for a terminal concept are the same. This set-oriented approach implies that logical operations on concepts can be performed by the corresponding set operations on their extension, and therefore the user is able to restrict the information base (and to create derived concepts) by combining concepts through all the standard logical operations (and, or, not).

A fundamental feature of this model is that dynamic taxonomies can find all the concepts related to a given concept C: these concepts represent the conceptual summary of C. Concept relationships other than subsumptions are inferred on the basis of empirical evidence through the extension only, according to the following *extensional inference rule*: two concepts A and B are related if there is at least one item *d* in the knowledge base which is classified at the same time under A or under one of A's descendants and under B or under one of B's descendants. For example, we can infer an unnamed relationship between *terrorism* and *New York*, if an item classified under *terrorism* and *New York* exists. At the same time, because *New York* is a descendant of *USA*, also a relationship between *terrorism* and *USA* can be inferred.

The extensional inference rule can be easily extended to cover the relationship between a given concept C and a concept expressed by an arbitrary subset S of the universe: C is related to S if there is at least one item d in S which is also in items(C). Hence, the extensional inference rule can produce conceptual summaries not only for base concepts, but also for any logical combination of concepts. In addition, because it is immaterial how S is produced, dynamic taxonomies can produce summaries for sets of items produced by other retrieval methods such as database queries, shape retrieval, and so forth, and therefore access through dynamic taxonomies can be easily combined with any other retrieval method.

Dynamic taxonomies are defined in terms of conceptual descriptions of items, so that heterogeneous items of any type and format can be managed in a single, coherent framework. Finally, because concept C is just a label that identifies the set of the items classified under C, concepts are language-invariant, and multilingual access can be easily supported by maintaining different language directories, holding language-specific labels for each concept in the taxonomy.

Exploration

The user is initially presented with a tree representation of the initial taxonomy for the entire knowledge base. The initial user focus F is the universe, that is, all the items in the information base. In the simplest case, the user selects a concept C in the taxonomy and zooms over it. The *zoom* operation changes the current state in the following way:

1. Concept C is used to refine the current *user focus* F, which becomes F∩items(C). Items not in the focus are discarded.
2. The tree representation of the taxonomy is modified in order to summarize the new focus. All and only the concepts related to F are retained and the count for each retained concept C' is updated to reflect the number of items in the focus F that are classified under C'. The *reduced taxonomy* is derived from the initial taxonomy by pruning all the concepts not related to F, and it is a conceptual summary of the set of documents identified by F, exactly in the same way as the original taxonomy was a conceptual summary of the universe. In fact, the term *dynamic taxonomy* indicates that the taxonomy can dynamically adapt to the subset of the universe on which the user is focusing, whereas traditional, static taxonomies can only describe the entire universe.

The retrieval process can be seen as an iterative thinning of the information base: the user selects a focus, which restricts the information base by discarding all the items not in the current focus. Only the concepts used to classify the items in the focus and their ancestors are retained. These concepts, which summarize the current focus, are those and only those concepts that can be used for further refinements. From the human computer interaction point of view, the user is effectively guided to reach his goal by a clear and consistent listing of all possible alternatives, and, in fact, this type of interaction is often called *guided thinning* or *guided navigation.* Such an iterative refinement terminates when the number of items in the focus is sufficiently small for manual inspection. In order to assist the user in deciding whether a simple concept expansion or a zoom operation is required, each concept label usually shows a count of all the items classified under it, that is, the cardinality of items(C) for all C's.

Dynamic taxonomies can be integrated with other retrieval methods in two basic ways. First, focus restrictions on the dynamic taxonomy can provide a context on which other retrieval methods can be applied, thereby increasing the precision of subsequent searches. Second, the user can start from an external retrieval method, and see a conceptual summary of the concepts that describe the result. Concepts in this summary can be used to set additional foci. These

two approaches can be intermixed in different iteration steps during a single exploration.

Figures 1 to 5 show how the zoom operation works. Figure 1 shows a dynamic taxonomy: the upper half represents the intension with circles representing concepts; the lower half is the extension, and documents are represented by rectangles. Arcs going down represent subsumptions; arcs going up represent classifications. In order to compute all the concepts related to C, we first compute, in Figure 2, all the documents classified under C (that is, the deep extension of C, items(C)) by following all the arcs incident to C and its descendants, H and I, from the extension: items(C)={ c, d }. All the items not in the deep extension of C (Figure 3) are removed from the extension. In Figure 4, the set of all the concepts under which the documents in items(C) are classified, B(C), is found by following all the arcs leaving each element in the set: B(C)={ F, G, H, I }. Because of the inclusion constraint implied by subsumption, if items(C) denotes the set of documents classified under C and C' is a descendant of C in the taxonomy, then items(C')⊆items(C) (Sacco, 2000), or, equivalently, a document classified under C' is also classified under C. Hence, the set of concepts related to C is given by B(C) union all the ancestors of all the concepts in B(C), that is, the set of all concepts related to C is {F, G, H, I, B, C, A}. Finally, in Figure 5, all the concepts not related to C are discarded, thus producing a reduced taxonomy that fully describes all and only the items in the current focus.

Figure 1. A dynamic taxonomy: The intension is above, the extension below. Arrows going down denote subsumptions, going up classification.

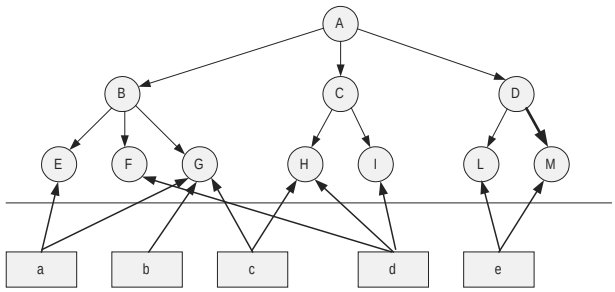


Figure 2. Focusing on concept C: Fnding all the items classified under C, that is, the deep extension of C

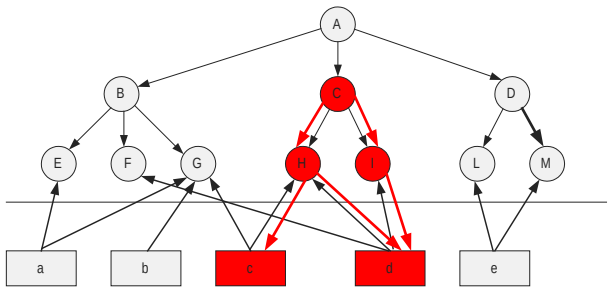


Figure 3. All the items not classified under C are removed

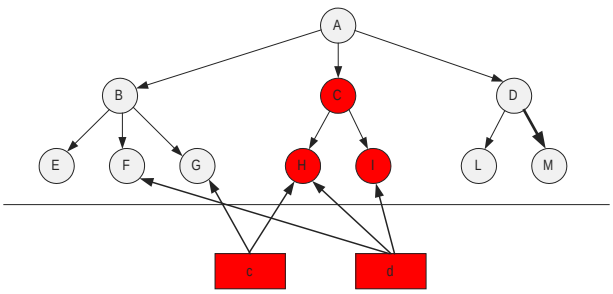
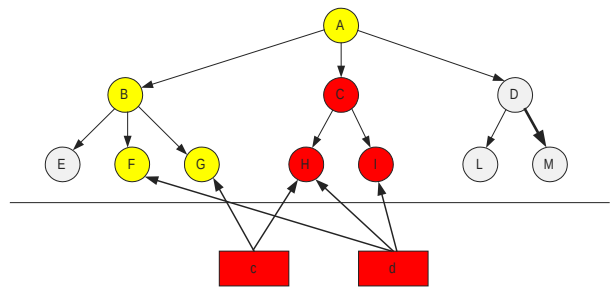


Figure 4. All the concepts under which the items in the focus are classified and (because of subsumptions) their ancestors are related to C

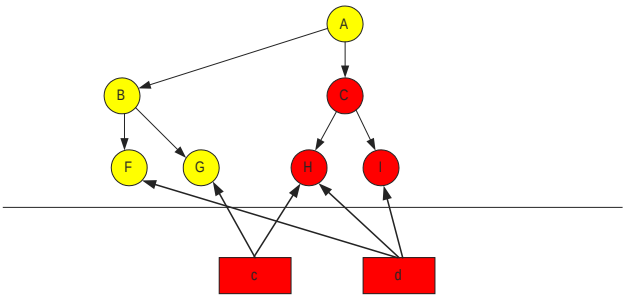


Advantages

The advantages of dynamic taxonomies over traditional methods are dramatic in terms of convergence of exploratory patterns and in terms of human factors. The analysis by Sacco (Sacco, 2006a) shows that 3 zoom operations on terminal concepts are sufficient to reduce a 10 million item infobase, described by a compact taxonomy with 1,000 terminal concepts organized according to 10 independent facets, to an average 10 items. Experimental data on a real newspaper corpus of over 110,000 articles, classified through a taxonomy of 1,100 concepts, reports an average 1,246 documents to be inspected by the user of a static taxonomy vs. an average 27 documents after a single zoom on a dynamic taxonomy.

Dynamic taxonomies only require the concept of a taxonomic organization and the zoom operation, which seems to be very quickly understood by end-users. Usability tests on a corpus of art images were conducted by Yee, Swearingen, Li, and Hearst (2003). Despite slow response times, access through a dynamic taxonomy was shown to produce a faster overall interaction and a significantly better recall than access through text retrieval. Perhaps more important are the intangibles: the feeling that one has actually considered all

Figure 5. The reduced taxonomy: All concepts not related to the current focus are pruned



the alternatives in reaching a result. Although few usability studies exist, the recent, widespread adoption by e-commerce portals, such as Yahoo!, Lycos, Bizrate, and so forth, strongly supports this initial evidence.

The extensional inference rule as a device to derive concept relationships has important implications on conceptual modeling. First, it simplifies taxonomy creation and maintenance. In dynamic taxonomies, no relationships in addition to subsumptions are required, because concepts relationships are automatically derived from the actual classification. For this reason, dynamic taxonomies easily adapt to new relationships and are able to discover new, unexpected ones. By converse, in traditional approaches, only the relationships among concepts explicitly described in the conceptual schema are available to the user for browsing and retrieval, so that all possible relationships must be anticipated and described: a very difficult if not helpless task.

Second, because dynamic taxonomies synthesize compound concepts, they need usually not be represented explicitly, so that the main cause of the combinatorial growth of traditional taxonomies is removed. Sacco (2000) developed guidelines that produce taxonomies that are compact and easily understood by users. Some are similar to basic faceted classification (Ranganathan, 1965; Hearst, 2002), at least in its basic form: the taxonomy is organized as a set of independent, "orthogonal" subtaxonomies (facets or perspectives) to be used to describe data. As an example of faceted design guidelines, consider a compound concept such as "19th century French paintings." It can be split into its **facets**: a Location taxonomy (of which France is a descendant), a Time taxonomy (of which the nineteenth century is a descendant) and finally an Art taxonomy (of which painting is a descendant). The items to be classified under the compound concept will be classified under Location>France, Time>19th century and Art>Painting instead. The extensional inference rule establishes a relationship among these concepts and the compound concept can be recovered by zooming on any permutation of them.

In a conventional classification scheme, such as Dewey indexing (Dewey, Mitchell, Beall, & Matthews, 1997), in which every item is classified under a single concept, a number of different concepts equal to the cartesian product of the terminals in the three taxonomies has to be defined. Such a combinatorial growth either results in extremely large conceptual taxonomies or in a gross conceptual granularity (Sacco, 2000). In addition, faceted design coupled with dynamic taxonomies makes it simple to focus on a concept, for example, 19th century, and immediately see all related concepts such as literature, painting, politics, and so forth, which are recovered through the extensional inference rule. In the compound concept approach, these correlations are unavailable because they are hidden inside the concept label.

Additional advantages include the uniform management of heterogeneous items of any type and format, easy multilingual access and easy integration with other retrieval methods. Dynamic taxonomies do not support reasoning beyond the extensional inference rule, and are therefore less powerful than general ontologies. However, they can be directly manipulated by users without the mediation of specialized agents and represent a quicker, less costly and more transparent alternative.

APPLICATIONS

The main industrial application is currently e-commerce. Assisted product selection is a critical step in most large-scale e-commerce systems (Sacco, 2003) and the advantages of dynamic taxonomies in user interaction are so significant as to justify the restructuring of well-established e-commerce portals: current examples include Yahoo, Lycos, Bizrate, and so forth.

However, dynamic taxonomies have an extremely wide application range and a growing body of literature indicates that their adoption benefits most search tasks. In addition to e-commerce, e-auctions and e-catalogs, key areas include e-government (Sacco, 2005a, 2005c), human resources and job placement portals (Berio, Harzallah, & Sacco, 2007), news portals, art and museum portals (Yee et al., 2003; Hyvönen, Saarela, & Viljanen, 2004), medical guidelines (Wollersheim & Rahayu, 2002) and diagnostic systems (Sacco, 2005b), among others. An additional area is multimedia databases, where dynamic taxonomies can be used to integrate access by conceptual metadata and access by primitive multimedia features (color, texture, etc.) into a single, coherent framework (Sacco, 2004).

A growing number of Web-based commercial systems based on dynamic taxonomies exist. Among these, Knowledge Processors, Endeca, i411 and Siderean Software.

FUTURE TRENDS

Current applications of dynamic taxonomies are relatively simple and use shallow taxonomies to access small to medium databases. However, the widespread adoption by e-commerce portals will make the model familiar to a large number of users and will make them aware of its significant benefits. For this reason, we expect widespread adoption and applications to evolve according to four coordinates:

1. New and emerging application areas, such as the ones reviewed above;
2. Larger information bases. Although the advantages of dynamic taxonomies over traditional search technologies are evident even with tens of documents, the superior convergence of this method makes it almost a forced choice for large multimillion item databases, which we expect will become the dominant application area in the near future;
3. Richer semantics. Current e-commerce applications have a relatively poor semantic content and require very simple taxonomies, but potential application areas can be much more complex and demanding in the design of multilevel taxonomies with thousands of concepts; and
4. Different architectures. Dynamic taxonomies are currently deployed in traditional Web sites with pull technology. Currently, push implementations are being experimented (Sacco, 2005e) and we expect early adoption in P2P networks, where access is currently based on a primitive form of text retrieval and no exploration capabilities are present. A significant use of dynamic taxonomies in mobile equipments is also expected, because the access model, coupled with a careful design of the taxonomy, is especially suited to low-resolution devices.

In order to achieve these results, continuing research is required. A review of research issues can be found in Sacco (2006b). They can be grouped into three main areas:

1. **Automatic classification and data modeling:** Dynamic taxonomies define how an existing classified corpus can be accessed and explored, but how documents are actually classified is left undefined. This is by design, because different types of data items (text, audio, video, etc.) need different classification strategies. Current research includes automatic text classification (Dakka, Ipeirtis, & Wood, 2005) and automatic classification from structured data (Sacco, 1998). Other research addresses the problem of specifying valid term composition rules in faceted taxonomies for textual information (Tzitzikas, Analyti, & Spyratos, 2005).

   As regards data modeling, recent investigations (Sacco, 2005d) suggest that dynamic taxonomies can be automatically derived from semantically rich conceptual schemata and used as a user-centered front-end to complex information. Extensions to the current model, in order to account for a fuzzy (Zadeh, 1965) classification, in which a document can be classified under several concepts with different probabilities, are also being investigated (Sacco, 2004).
2. **Human factors:** Because of its user-centered approach, human factors play a central role in the model, and especially in critical issues such as the presentation and manipulation of the taxonomy, where several alternatives exist (see Yee et al., 2003 vs. Sacco, 2000, 2004).
3. **Implementation:** The zoom operation and the subsequent reduction of the corpus taxonomy must be performed in real time because a slower execution would severely impair the sense of free exploration that the user of dynamic taxonomy systems experiences. Special data structures and evaluation strategies must be used (Sacco, 1998). In addition, distributed and federated architectures need to be investigated because centralized architectures are not always appropriate, because of organization needs and of performance and reliability bottlenecks.

CONCLUSION

Exploratory browsing applies to most search tasks: an extremely wide application range going from multilingual portals, to e-commerce, e-auctions, e-government, human resources management, CRM, and so forth. In this context, dynamic taxonomies represent a dramatic improvement over other search and browsing methods, both in terms of convergence and in terms of full feedback on alternatives and complete guidance to reach the user goal. For these reasons, we expect dynamic taxonomies to become pervasive in the short period, and to replace or integrate traditional techniques in a growing number of applications.

REFERENCES

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Comm f the ACM, 8*(3), 289-299.

Berio, G., Harzallah, M., & Sacco, G. M. (2007). Portals for integrated competence management. In A. Tatnall (Ed.), *Encyclopedia of portal technology and applications*. Hershey, PA: Idea Group.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, May 17*, 35-43.

Dakka, W., Ipeirotis, P. G., & Wood, K. R. (2005). Automatic construction of multifaceted browsing interfaces. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05),* (pp. 768-775).

Dewey, M., Mitchell, J. S., Beall, J., & Matthews, W. E. (1997). *Dewey decimal classification and relative index* (21st ed.). OCLC.

Groenbaek, K., & Trigg, R. (Eds.). (1994). Hypermedia. *Communications of the ACM, 2*, 37.

Hyvönen, E., Saarela, S., & Viljanen, K. (2004). Application of ontology techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium (ESWS 2004),* (pp. 92-106). Springer-Verlag, LNCS 3053.

Morville, P., (2002, April 29, 2002). *The age of findability. Boxes and Arrows*. Retrieved December 9, 2007, from http://www.boxesandarrows.com/archives/002595.php

Ranganathan, S. R. (1965). *The colon classification. In S. Artandi (Ed.), Rutgers series on systems for the intellectual organization of information (Vol. 4)*. New Jersey: Rutgers University Press.

Sacco, G.M. (1987). Navigating the CD-ROM. In *Proceedings of the International Conference on Business of CD-ROM*.

Sacco, G. M. (1998). *Dynamic taxonomy process for browsing and retrieving information in large heterogeneous data bases*. US Patent 6,763,349; also, Italian Patent 01303603.

Sacco, G. M. (2000). Dynamic taxonomies: A model for large information bases. *IEEE Transactions on Knowledge and Data Engineering, 12*(2) , 468-479.

Sacco, G. M. (2003). The intelligent e-sales clerk: The basic ideas. In *Proceedings of the INTERACT'03 -- Ninth IFIP TC13 International Conference on Human-Computer Interaction*, (pp. 876-879).

Sacco, G. M. (2004). Uniform access to multimedia information bases through dynamic taxonomies. In *Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering,* (ISMSE '04), (pp. 320-328).

Sacco, G. M. (2005a). No (e-)democracy without (e-)knowledge. In *E-government: Towards electronic democracy, International Conference of IFIP TCGOV 2005,* Bolzano, (pp. 147-156). Springer Lecture Notes in Computer Science 3416.

Sacco, G. M. (2005b). Guided interactive diagnostic systems. In *Proceedings of the 18th IEEE International Symposium on Computer-based Medical Systems (CBMS '05),* (pp. 117-122).

Sacco, G. M. (2005c). Guided interactive information access for e-citizens. In *EGOV05: Proceedings of the International Conference on E-government, within the Dexa Conference Framework* (pp. 261-268). Springer Lecture Notes in Computer Science 3591.

Sacco, G. M. (2005d, November 11). *Discount semantics: Modeling complex data with dynamic taxonomies* (Tech. Rep.). Università di Torino.

Sacco, G. M. (2005e). DBWorld Xtended: Semantic dissemination of information through dynamic taxonomies. In *Proceedings of the 5th International Conference on Knowledge Management, I-KNOW05*. Graz, J. of Universal Computer Science: Springer-Verlag.

Sacco, G. M. (2006a, June). Analysis and validation of information access through Mono: Multidimensional and dynamic taxonomies. In *Proceedings of FQAS 2006: 7th International Conference on Flexible Query Answering Systems*, Milano. Springer Lecture Notes on Artificial Intelligence 4027.

Sacco, G. M. (2006b, August). Some research results in dynamic taxonomy and faceted search systems. In *Proceedings of the SIGIR'2006 Workshop on Faceted Search,* Seattle, WA, USA.

Tzitzikas, Y., Analyti, A., & Spyratos, N. (2005). Compound term composition algebra: The semantics. *Journal on Data Semantics, 2*, 58-84.

van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworths, London.

Wollersheim, D., & Rahayu, W. (2002). Methodology for creating a sample subset of dynamic taxonomy to use in navigating medical text databases. In *Proceedings of the IDEAS 2002 Conference*, (pp. 276-284).

Yee, K-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. In *Proceedings of the ACM CHI 2003*, (pp. 401-408).

Zadeh, L. (1965). Fuzzy sets. *Information Control, 8*, 338-353.

KEY TERMS

Extension, Deep: Of a concept C, denotes the shallow extension of C union the deep extension of C's sons.

Extension, Shallow: Of a concept C, denotes the set of documents classified directly under C.

Extensional Inference Rule: Two concepts A and B are related if there is at least one item *d* in the knowledge base which is classified at the same time under A (or under one of A's descendants) and under B (or under one of B's descendants).

Facet: One of several top level (most general) concepts in a multidimensional taxonomy. In general, facets are independent and define a set of "orthogonal" conceptual coordinates.

Subsumption: A subsumes B if the set denoted by B is a subset of the set denoted by A ($B \subseteq A$)

Taxonomy: A hierarchical organization of concepts going from the most general (topmost) to the most specific concepts. A taxonomy supports abstraction and models subsumption (IS-A and/or PART-OF) relations between a concept and its father. Tree taxonomies can be extended to support multiple inheritance (i.e., a concept having several fathers).

Taxonomy, Monodimensional: Taxonomy where an item can be classified under a single concept only

Taxonomy, Multidimensional: Taxonomy where an item can be classified under several concepts

Taxonomy, Reduced: In a dynamic taxonomy, a taxonomy, describing the current user focus set F, which is derived from the original taxonomy by pruning from it all the concepts not related to F.

User Focus: The set of documents corresponding to a user-defined composition of concepts; initially, the entire knowledge base.

Zoom: A user interface operation, that defines a new user focus by OR'ing user-selected concepts and AND'ing them with the previous focus; a reduced taxonomy is then computed and shown to the user.

E-Book Technology in Libraries

Linda C. Wilkins
University of South Australia, Australia

Elsie S. K. Chan
Australian Catholic University, Australia

INTRODUCTION

The shift towards electronically mediated texts entails major structural issues for libraries and the publishers and aggregators who supply them. Stakeholders within the digital supply chain are struggling to reconceptualize the book as an artefact (Esposito, 2003). Academic and scholarly libraries are at the forefront of these changes. In this article we review some recent developments in the technology underpinning e-books, introduce some of the key players, and review influences affecting uptake.

BACKGROUND

Libraries have traditionally played a key role in providing access to and disseminating information across a community. That role has now extended to facilitating access to innovative technologies. Technological improvements such as Amazon's Look inside the Book technology and Google Print offering pages of a book on the Web drive accelerate demand for access ("The Economist," 2005). Such developments have generated pressures on libraries to make research output more widely available through search engines and open access mechanisms which in turn result in rising accessibility of research material via downloads and citations (Rosenzweig, 2005). Electronic access in the form of electronic journal subscriptions, e-books, and databases has resulted in the rapid and continuing evolution of library facilities. Many public libraries have experimented with a variety of e-book devices while academic libraries have overseen a dramatic shift in the percentage of their budget allocation dedicated to the provision of digital resources.[1] Hence, a review of the current provisions of electronic resources in libraries appears to be timely.

ENGAGING WITH E-BOOK TECHNOLOGY

The term *e-book* is unsatisfactory in many respects. In the case of traditional print books, users can immediately understand and identify elements belonging to book technology. By contrast, the term e-book does not explain either the form or its operations (for further information see Lynch, 2001, p. 125). As a generalised term it was initially applied to three types of appliances: e-book, e-tablet, and Personal Digital Assistant (PDA). Only their design, purpose, and size distinguished them from software book readers. Table 1 lists devices available in 2005.

Reader software can be categorised by e-book formats. Abobe, HTML, and Microsoft readers are some examples of e-book formats. Table 2 lists different formats of e-book software.

E-Books in Public Libraries

The period between 1999 and 2001 saw a surge in e-book reader trials in libraries in the United States, Canada, Denmark, Norway, and Australia.[2] Trials of e-book reader hardware in Australia, Canada, and the U.S. recorded overwhelmingly positive responses from pilot group users and librarians alike (Wilkins, Coburn, Burrows, & Loi, 2001). In Australia, user participants in the Brisbane Public Library e-book reader pilot study loved the compact, portable nature, adjustable font size, dictionary, and search and bookmark functions of dedicated e-book readers while librarians were keen to showcase the new technology. Many saw the readers as an opportunity to expose their community to technology as it "came down the pipeline" (Glencoe Public Library Illinois, USA, as cited in, Wilkins et al., 2001, pp. 252-253).

A generic feature of rapid technological change is the proliferation of designs, many of which will inevitably fail over time (Bijker, 1995). At a time when proprietary platforms dominated the marketplace, librarians found it difficult to make decisions about which device to go with, which text format to choose, and what copyright arrangements to take up. Eventually those librarians who had pioneered e-book reader technology trials in multi-year pilot projects found themselves tied to restrictive access models with exclusive proprietary book file format.[3] Unique formats that can become obsolete at any moment slow uptake and support for e-book reader hardware. The limited range of titles available on the available hardware also meant their appeal to borrowers faded over time (Lynch, 2001). By 2005 these early adopters found that "dedicated readers with pre-loaded content were

Table 1. e-book reader device (adapted from eBookMall (2005a)

Device	Example Image	eBook Format	Weight in oz.	Size	Screen Description
Gemstar eBook		Gemstar eBook	17	largish	4.75" x 3" Monochrome Back lit Touch screen
Handspring Visor		Palm Reader Mobipocket	5.4-6.9	smallish	about 3" x 4", some color, some not
hiebook		hiebook	8.8	115.4 x 146 x 17 mm	Back lit 480 X 320 px touch screen LCD display
Palm		Palm Reader Mobipocket	4-6	4.82" x 3.1" x .87"	Advanced LCD with backlight
Pocket PC		Microsoft Reader Mobipocket	6-16	Depends on device	Reflective or Transflective LCD, 16+ colors

not what patrons most wanted in an eBook" (M. Williams, personal communication, November 30, 2005).

The introduction of software book readers that run on general purpose computers and require no additional financial outlay for a separate hardware device or book reading appliance has effectively turned the desktop or laptop into book reading appliances. Many of those users accessing texts electronically on general purpose computers appear to be doing so in academic and research library settings.

E-books in Academic and Research Libraries

Our customers expect electronic content. (Woodward, 2005, p. 3)

Academic libraries are particularly well suited to the e-market with their large, expensive, and rapidly dated reference books that are costly to weed out (Michael, 2005). While all librarians have a professional commitment to efficient document delivery and place priority on offering content that their users require, librarians in research settings experience significant additional pressures to seek optimal methods for providing access, disseminating, receiving, and reporting publications. All students have now come to expect free Internet access; on-campus and part-time students require remote and 24/7 access via private ISPs. Researchers place priority on speed, timing, and knowing the latest in cutting-edge technology.[4] Institutional requirements for the library to provide information in a cost-effective way add to these pressures from competing constituencies.

Electronic delivery systems appear to offer solutions to many of these needs and requirements. Online search functions, easier navigation, the ability to cut and paste, well-organised and up-to-date materials, convenience (no carrying of books), paper saving, and lower levels of physical maintenance are all attractive features of e-book provision. Academics also appreciate that requirements for accessing e-books encourage students to copy with appropriate acknowledgement.

These features of the research constituency favour electronic delivery and hence offer a partial explanation for the dramatic growth in e-resources as a proportion of academic library budgets. Underpinning the drive for uptake is the fact that electronic publishing has transformed the book into a

Table 2. e-book formats (adapted from eBookMall, 2005b)

Format	Advantages	Reader Software	Navigation	Platforms
Adobe Reader	Cross-platform compatibility, printable, single or double page view	Get Adobe Reader	Library, Table of contents, Chapter links, bookmarkable	Windows PC, Macintosh, Palm
eBookMall	Dedicated reader for eBooks, carry titles with you		Library, Table of contents, bookmarks	Gemstar & Rocket eBook devices
hiebook	Dedicated reader for eBooks, includes lots of other programs	hiebook	Library, Table of contents, bookmarks	Hiebook devices
HTML	Easy to use, customizable, can be read on anything with a browser	Get Microsoft Internet Explorer Netscape Now! MOBIPOCKET Reader	Hypertext links	Windows PC, Macintosh, Linux, Unix, Palm, Pocket PC, eBookMan
Instant eBook	No special reader software required (only Internet Explorer), easy to use	Get Microsoft Internet Explorer	Hypertext links	Windows PC
Microsoft Reader	ClearType Display Technology, book-like reading environment, bookmarks and annotations	eBooks are Here!	Library, Table of contents, Chapter links, bookmarkable	Windows PC, Pocket PC
MOBIPOCKET	Can be used on any PDA	MOBIPOCKET Reader	Library, Table of contents, Chapter links, bookmarkable	Palm, Pocket PC, eBookMan, Windows PC
palm	Great eBooks for your PDA	palmreader MOBIPOCKET Reader	Library, Table of contents, Chapter links, bookmarkable	Palm OS, Pocket PC, Handheld PC, Windows CE, Windows PC

digital product and hence into a market offering. The potential for growing this market underpins the considerable economic interest in e-books among publishers and aggregators.

Key Players

The following list introduces some of the e-book suppliers to libraries currently active in this market.

- **EBL (www.eblib.com):** EBL offers around 18,000 titles from around 90 publishers. EBL was launched in 2004 by Ebooks Corporation (ebooks.com) and developed in collaboration with academic publishers and libraries in Australia, the U.S., the UK, and Europe. EBL's distinctive access requirements were designed primarily for academic and research libraries. Lending models include options for multiple concurrent use, unlimited access, and short-term circulation. As well as, individual e-book chapters can be set aside for reserve lending or be included within course packs. Academic publishers associated with EBL are Taylor and Francis, Oxford University Press, Cambridge University Press, Kluwer, and World Scientific Press. Distributors Blackwell's Book Services and Dawson Books have signed partnership agreements regarding content for North America, the UK, and Australia. EBL e-books are to be integrated into Collection Manager, Blackwell's Web-based collection development and acquisition system. Librarians' workflow requirements are catered for by EBL's publisher interface (Pi), which offers an Extensible Markup Language (XML) service to expedite entry of new electronic titles and requires only an ISBN entry.

- **Proquest Safari Computer eBooks (www.proquest.com):** Connection to the Proquest Safari Computer eBooks Web site is through a personal computer with registration via Library card. Books can be checked out and read online using the Web browser. Titles can be changed monthly with accurate usage data allowing removal of unused titles. Individual chapters can be downloaded to the user's personal computer (PC). Prices are weighted based on book demand. Safari publishes mainly information technology (IT) books—a category frequently selected by librarians undertaking e-book trials who perceive this audience as more accepting of their delivery mechanisms (Michael, 2005). Safari offers outright purchase rather than subscription as well as the option of a small starter package. Safari's purchase model has become popular with a number of librarians in research-based settings (Cox, 2004). As a low risk option it has proven particularly attractive to librarians seeking to trial e-books within budget constraints (H. Pearsall, personal communication, November 17, 2005).
- **NetLibrary (http://www.netlibrary.com):** A pioneer in the delivery of e-books to universities — is a subsidiary of the huge OCLC library cooperative, the world's largest bibliographic database built and maintained collectively by librarians http://www.oclc.org/worldcat/about/default.htm (retrieved 30th June 2006). Netlibrary offers a collection of some 90,000 e-book titles, from more than 300 publishers via a purchase-based model. Customers accessing NetLibrary e-books include 5,500 libraries and organizations.
- **ebrary (http://www.ebrary.com):** ebrary's subscription model offers more than 60,000 e-book titles from more than 200 publishers, accessed by some 500 libraries in 60 countries.
- **Questia (http://www.questia.com):** Questia offers 50,000 e-book titles, more than 119,000 journal articles, and over 159,000 newspaper articles and claims to be the world's largest online library of books and journal articles.

FUTURE TRENDS

International multi-year pilot studies, focus groups, and case studies conducted with publishers, librarians, and library users between 2001 and 2005 revealed a number of concerns that continue to have an impact on e-book uptake in libraries.

- **Competing Constituencies:** Instead of joining forces in the digital era, libraries and publishers have been fighting a battle sometimes described by librarians as the fight between good and evil (Vigen & Paulson, 2003). Frequently at issue is the fact that libraries aim to address community needs on the basis of sharing, borrowing, and recycling—facilities traditionally offered to patrons on a nonpaying basis. Publishers' transaction models for e-books accommodate the concept of controlled sharing only with considerable difficulty (Hoorebeck, 2003).
- Within the research community itself, libraries must also cater to diverse stakeholders. Students welcome the additional features e-texts offer. They want and expect uninhibited access to information in multiple formats where and when they require it. Academics and researchers, by contrast, place far greater priority on the *quality* of content (Michael, 2005).
- Within the library community, archivists represent another constituency concerned that the move to e-resources leaves unanswered questions about preservation issues and continuity of access over time.
- **A Plethora of Standards and Devices:** Despite attempts by the Open E-book Forum (now The International Digital Publishing Forum[5]) a variety of proprietary standards still exist rendering most e-books compatible only with certain devices. The lack of an agreed standard implies that an agreeable machine to deliver books to a mass audience has not yet arrived on the scene (Turney, 2005).
- **Content:** Low levels of currency and relevance to reader requirements and the time lag between printed and electronic versions of texts have presented serious drawbacks to the spread of e-books, particularly in research collections.
- **Authentication:** Products developed by providers and distributors are often not geared to procedures that libraries use for controlling access to e-resources. Agents offering competitive "deals," cast librarians in the unfamiliar and time-consuming role of negotiating terms, generating increased work loads for librarians.
- **Work Flows:** Collection managers find that making new monthly selections can generate considerable additional work for some of the library staff (Abbott & Kelly, 2004). Payment methods can also be comparatively cumbersome and a disincentive for uptake. Clearly for e-books to succeed, selecting and purchasing them needs to be as easy as ordering and buying from the campus bookstore (Abbott & Kelly, 2004).
- **Budgetary Pressures:** Research resources are provided or supported by major national institutions and organisations, and by many local bodies including universities, libraries, and archive offices. These bodies are all experiencing funding pressures and challenges in balancing electronic and nonelectronic resource provision (British Academy Report, 2005).

CONCLUSION

The e-book delivery supply chain as it currently exists has been described as "ad hoc and fragmented, lacking in leadership and coordinated strategy" (British Academy Report, 2005). Despite these problems, some signs of progress do exist. Major publishers have now come to the realisation that they cannot afford to not have a fully developed e-books strategy (Turney, 2005). Google's commitment to an e-book scanning project in partnership with leading universities in the U.S. and the UK represents a major push to address the dearth of quality material on the Web. These developments have accelerated adaptations in publishers' business models and supply chains to facilitate e-book uptake (Rosenblatt, 2004).

Other drivers are also emerging. We have already discussed market forces favouring digital delivery of information to libraries. Governments are now also under increasing fiscal pressure to maximise return on public investment. Hence, barriers to the free flow of information and the costs these barriers represent to publicly funded research are beginning to draw attention at the national level (for UK see Harnad, 2003; for Australia see Department of Education, Science and Training [DEST], 2004). Policy and/or legislative moves to lift these barriers may well stimulate further progress in diffusion of e-book technology.

REFERENCES

Abbott, W., & Kelly, K. (2004, February 3-5). Sooner or later!—Have e-books turned the page? Paper presented at the V*ALA 2004: Breaking Boundaries: Integration and Interoperability. Twelfth Biennial Conference and Exhibition,* Melbourne Convention Centre. Retrieved June 30, 2006, from http://www.vala.org.au/vala2004/2004pdfs/46AbbKel.PDF

ARL Statistics 2001-2002. (2004). *Expenditure trends in ARL libraries.* Retrieved December 1, 2005, from http://www.arl.org/stats/arlstat/graphs/2004/aexp04.pdf

Bijker, W. (1995). *Of bicycles, bakelites, and bulbs: Toward a theory of sociotechnical change*. Cambridge, MA: MIT Press.

British Academy Report. (2005). *E-resources for research in the humanities and social sciences—A British academy policy review*. Retrieved December 1, 2005, from http://www.britac.ac.uk/reports/eresources/report/index.html

Cox, J. (2004, October). E-book challenges and opportunities. *D-Lib Magazine, 10*(10). Retrieved December 1, 2005, from http://www.dlib.org/dlib/october04/cox/10cox.html

Department of Education, Science and Training (DEST). (2004). *Backing Australia's ability—Building our future through science and innovation.* Retrieved December 1, 2005, from http://backingaus.innovation.gov.au/

ebookMall. (2005a). *eBook device comparisons*. Retrieved December 1, 2005, from http://www.ebookmall.com/knowledge-collection/device-comparisons.htm

ebookMall. (2005b). *eBook formats*. Retrieved December 1, 2005, from http://www.ebookmall.com/knowledge-collection/format-comparisons.htm

Esposito, J. J. (2003, March). *The processed book.* Retrieved December 1, 2005, from http://firstmonday.org/issues/issue8_3/esposito/index.html

Harnad, S. (2003). Electronic preprints and postprints. *Encyclopaedia of Library and Information Science,* Marcel Dekker. Retrieved December 1, 2005, from http://www.ecs.soton.ac.uk/~harnad/Temp/eprints.htm

Hoorebeck, M. (2003). e-books, libraries and peer to peer file-sharing. *Australian Library Journal, 52*(1). Retrieved June 30, 2006, from http://alia.org.au/publishing/alj/52.2/

Kennewell, S. (2005). Fewer dollars, fewer books, but many more words: The state of campus libraries. *Campus Bookseller and Publisher 85*(2), August. Mebourne, Australia: Thorpe-Bowker.

Lynch, C. (2001). *The battle to define the future of the book in the digital world*. Retrieved June 30, 2006, from http://www.firstmonday.org/issues/issue6_6/lynch/index.html

Michael, R. (2005, August). What price ebooks? *AustralianBookseller and Publisher 85*(2)*.* Melbourne, Australia: Thorpe-Bowker.

Rosenblatt, B. (2004). Public libraries offer new digital formats. *The Seybold Report: Analysing Publishing Technologies, 24*(3), 15-19.

Rosenzweig, R. (2005). Should historical scholarship be free? *American Historical Association.* Retrieved December 1, 2005, from http://www.historians.org/Perspectives/issues/2005/0504/0504vic1.cfm

The Economist. (2005, December). *Pulp Friction, 377*(8452), 63-64.

Turney, D. (2005, August). Generation e. *Australian Bookseller and Publisher 85*(2). Melbourne, Australia: Thorpe-Bowker.

Vigen, J., & Paulson, K. (2003). *E-books and interlibrary loan: An academic centric model for lending.* Retrieved December 1, 2005, from http://www.nla.gov.au/ilds/abstracts/VigenJ.pdf

Wilkins, L., Coburn, M., Burrows, P., & Loi, D. (2001). The trials of technology: The Brisbane e-book reader trial and focus group. In B. Cope & D. Kalantzis (Eds.), *Print and electronic text convergence: Technology drivers across the book production supply chain, from creator to consumer* (pp. 223-265). Altona Victoria, Australia: Common Ground Publishing.

Woodward, H. (2005, October). OUP pushes e-book agenda. *Information World Review, 217,* 3.

ACKNOWLEDGMENTS

Special thanks to Mike Williams, Indianapolis-Marion County Public Library USA for information about e-book reader pilot studies; and to Heather Pearsall, Library Manager (Electronic Services), Australian Catholic University (ACU) for the information on e-book uptake at the ACU.

KEY TERMS

Amazon.com: Amazon's "Search Inside the Book" feature introduced in 2003 and now available as a feature for half of all books sold, allows people to use keywords to search text inside books. The firm has plans to introduce the "Amazon Pages" program in 2006, enabling the purchase of online access to text—up to and including an entire work, and "Amazon Upgrade" a program to access books electronically that have been shipped in printed form.

E-Book Supply Chain: A sequence of activities and organisations involved in producing and delivering a good or service. An IT supply chain—as required for e-books—is the flow of resources into and out of the firm's IT operations.

Electronic Book (e-Book): An e-book is an electronic (or digital) version of a book that can be downloaded to computers or handheld devices.

Electronically Mediated Text E-Book Reader or Device: An e-book reader can be a software application for use on a computer, such as Microsoft's free Reader application, or a book-sized computer that is used solely as a reading device, such as Nuvomedia's Rocket e-book. Users can purchase an e-book on diskette or CD, but the most popular method of getting an e-book is to purchase a downloadable file of the e-book (or other reading material) from a Web site (such as Barnes and Noble) to be read from the user's computer or reading device. Generally, an e-book can be downloaded in 5 minutes or less.

Google Book Search: Launched in 2004, the Google Print Library Project digitises and makes online texts searchable from the university library collections at Harvard, Stanford, Michigan, Oxford, and the New York Public Library giving locations for purchase or borrowing. Copyright issues are yet to be resolved. Database access will only be via Google's search engine.

Open E-Book Forum: The International Digital Publishing Forum (IDPF), formerly the Open eBook Forum is the trade and standards association for the digital publishing industry (http://www.idpf.org/).

ENDNOTES

1 U.S. figures show a 227% increase in serial expenditure since 1986 or by an average of 7.7% per annum (ARL Statistics, 2001-2002). In Australia, academic libraries such as those of the Australian Catholic University and the University of Western Australia have overseen similar marked growth in their expenditure on e-book subscriptions (H. Pearsall, personal communication, November 17, 2005; Kennewell, 2005).

2 For example, Maricopa County Public Library in Phoenix, Arizona had 100 Rocket e-Book Pro readers available for loan in 2001 (Wilkins et al., 2001).

3 In Table 2 we show that the majority of e-book reader formats currently available are no longer restricted to proprietary platforms.

4 For example, on an individual basis, members of medical faculties typically rate as the highest users of electronic resources (Kennewell, 2005).

5 The URL of the International Digital Publishing Forum is www.idpf.org

Category: Electronic Business

E-Business Systems Security in Intelligent Organizations

Denis Trček
Jožef Stefan Institute, Slovenia

INTRODUCTION

Security as we perceive it today became a topic of research with the introduction of *networked information systems*, or networked ISs, in the early 1980s. In the mid-1990s the proliferation of the Internet in the business area exposed security as one of the key factors for successful online business, and the majority of efforts to provide it were focused on technology. However, due to lessons learned during this period, the paradigms have since changed, with increasing emphasis on human factors. It is a fact that security of networked ISs is becoming part of the core processes in all e-business environments. While data is clearly one of the key assets and has to be protected accordingly, ISs have to be highly integrated and open. Appropriate treatment of these contradictory issues is not a trivial task for managers of contemporary intelligent organizations. It requires new approaches, especially in light of new technologies.

BACKGROUND

Proper management of security in *e-business systems* requires a holistic methodology that can be viewed on three planes: technology, organization, and legislation (Trček, 2006). ISs security management starts with the identification of threats and threats analysis. A typical approach is based on risk probability and derived damage estimates (Raepple, 2001). Following this, the approach differs according to the plane:

- The technological plane takes into account machine-related interactions. This plane is about deployment of appropriate *security services* that are based on *security mechanisms*. To become operational, key management issues (i.e., handling of cryptographic algorithms' keys) have to be resolved. Finally, human-to-machine interactions have to be addressed carefully.
- The organizational plane takes human resources management into account. It emphasizes the organizational issues and socio-technical nature of contemporary IS, where various modern methodologies play a central role.
- In parallel, it is necessary to address legal issues. Not only national, but also international legislation in this area is becoming increasingly broad and complex. Each and every security policy has to take this into account, especially cryptography regulations, digital signature issues, privacy issues, and intellectual property rights.

METHODOLOGICAL APPROACHES TO E-BUSINESS SYSTEMS

From the technological point of view, the prevention of threats is achieved by use of security mechanisms and security services (ISO, 1995). Mechanisms include symmetric and asymmetric cryptographic algorithms, for example, AES (Foti, 2001) and RSA (RSA Labs, 2002); one-way hash functions such as SHA-1 (Eastlake & Jones, 2001); and physical mechanisms. For devices with weak processing capabilities like smart-cards, elliptic curve-based systems such as ECDSA (ANSI, 1998) can be used. Regarding physical security, using cryptographic algorithms one can only reduce the amount of data that have to be physically protected, but physical protection cannot be eliminated.

To ensure that a particular public key indeed belongs to the claimed person, a trusted third party called *certification authority*, or CA, has to be introduced. The CA issues *public key certificates* that are digitally signed electronic documents, which bind entities to the corresponding public keys (certificates can be verified by CA's public key). CA also maintains certificate revocation lists, or CRLs. These should be checked every time a certificate is processed in order to ensure that a private/public key is still valid. The de iure and de facto standard for certificate format is X.509 standard (ITU-T, 2000).

By use of security mechanisms, the following security services are implemented:

- **Authentication:** Ensures that the peer communicating entity is the one claimed.
- **Confidentiality:** Prevents unauthorized disclosure of data.
- **Integrity:** Ensures that any modification, insertion, or deletion of data is detected.

Table 1. Summary of basic security related elements—technological plane

• **Security Mechanisms:** Symmetric and asymmetric algorithms, one-way hash functions, physical mechanisms • **Security Services:** Authentication, confidentiality, integrity, non-repudiation, access control, auditing • **Security Infrastructure:** Public key infrastructure, commercial off-the-shelf solutions (firewalls, intrusion detection systems), technologies IPSec, SSL, S/MIME, EAP, RADIUS, and WPA

- **Access Control:** Enables authorized use of resources.
- **Non-Repudiation:** Provides proof of origin and proof of delivery, where false denying of the message content is prevented.
- **Auditing:** Enables detection of suspicious activities and analysis of successful breaches, and serves as evidence when resolving legal disputes.

To enable these services, a certain infrastructure has to be set up. It includes a registration authority (RA) that serves as an interface between a user and CA, identifies users, and submits certificate requests to CA. In addition, a synchronized time base system is needed for proper operation, along with a global directory for distribution of certificates and CRLs. All these elements, together with necessary procedures, form a so-called *public key infrastructure* or PKI (Arsenault & Turner, 2002).

To provide security, mostly commercial off-the-shelf solutions are used. Such solutions typically include firewalls, which are specialized computer systems that operate on the border between the corporate network and the Internet, where all traffic must pass through these systems (Cheswick & Bellovin, 1994). Further, real-time intrusion detection systems (Kemmerer & Vigna, 2002) are deployed for detecting acts that differ from normal, known patterns of operation, or for detecting wrong behavior. Further, IPSec (Thayer, Doraswamy, Glenn, 1998), which is a security enhancement for IP protocol, is becoming a norm to prevent masquerade, monitoring of a communication, modification of data, and session overtaking. IPSec is suitable for virtual private networks, or VPNs, where one can establish secure private networks using public networks such as the Internet. Further, secure sockets layer protocol, or SSL (Freier, Karlton, & Kocher, 1996), provides a common security layer for Web and other applications and is available by default in Web browsers. It provides authentication, confidentiality, and integrity with the possibility of negotiating crypto primitives and encryption keys. Last but not least, secure/multipurpose Internet mail extensions standard, or S/MIME (Ramsdell, 1999), is often deployed as security enhancement for ordinary e-mail. It provides authentication, confidentiality, integrity, and non-repudiation.

Increased penetration of wireless communications into business environments brings new topics on the agenda that are specific to *wireless security* (Miller, 2001). In this area it is nowadays common to deploy physical/link level security by using IEEE 802.1X standard (IEEE, 2004). 802.1X is a framework for authentication, access control, and key-exchange. Its authentication deploys Extensible Authentication Protocol, or EAP (Aboba, Blunk, Vollbrecht, Carlson, & Levkowitz, 2004), which is usually used together with RADIUS server (Rigney, Willens, Rubens, & Simpson, 2000) for remote authentication and accounting. In addition, a derivative of 802.1X, called WiFi Protected Access, or WPA, is very common. Despite these technologies, wireless networks are inherently more vulnerable than wired networks, thus wireless access points are often put outside corporate firewalls.

However, even superior technological solutions will be in vain, if the complementary organizational and legal issues are not treated properly. Therefore, the second plane that is concentrated on organizational issues through human resources management has to be properly covered by *security policy*. In addition, the third plane has to be taken into account to assure that all efforts are aligned with existing legislation.

The first standard in the area of security policies was BS 7799 (BSI, 1995), with its most current international successor being ISO (2005). This standard plays a central role as far as security policy management is concerned. However, to implement successfully security policy, it is essential to support managers of contemporary intelligent organizations with appropriate techniques. The organizational plane is characterized by a complex interplay between human factor and technology. The two constituent parts are coupled in many ways, such as by interactions. A large number of these interactions form various feedback loops. There are also soft factors that have to be taken into account, for example, human perception of various phenomena like trust. Therefore, to support decision making properly with regard to security, one has to deal with physical and information flows. Additionally, decisions are often to be made in circumstances where there is not enough time or resources to test decisions in a real environment; often such checks are not possible at all. Therefore, support from computer simulations is highly desirable.

The methodology that can be used to support the resolution of the previously mentioned problems is *business dynam-*

Table 2. A summary of advanced security related elements—organizational and legal planes

• **Security Policy:** An organization's document covering its networked information systems security from technological, organizational, and legal plane • **Business (or System) Dynamics:** A modeling methodology for socio-technical systems • **Business Intelligence:** Enabling integral capturing of data with knowledge extraction to support decision making

ics (Sterman, 2000). It enables qualitative and quantitative modeling of contemporary IS. Using this methodology, one starts with the identification of variables that are relevant to system behavior and defines the boundary distinguishing between endogenous and exogenous variables. Variables are connected by causal links, which have positive polarity if increase/decrease of input variable results in increase/decrease of output variable. In a case when increase/decrease of input variable results in decrease/increase of output variable, the polarity is negative. Some variables are of a stock nature (also called levels), and these introduce persistency (inertia) into the system. They decouple inflows and outflows, and thus present a kind of buffer or absorber. When building a model of a system by this approach, so-called causal loop diagrams are obtained, which provide a very useful means for management of e-business systems security. Finally, causal loop diagrams are upgraded with appropriate equations and tuned using real data to enable quantitative analysis, that is, computer simulation.

Summing up, security policy is the main document about risk management in every organization, and its introduction and maintenance has to be adequately supported by use of modern business intelligence techniques. Using causal loop diagramming, decision makers can get a holistic perspective on their systems, but to verify and properly adjust these models, they have to link them to appropriate sources to obtain real-time data for simulations. The resulting architecture provides an important tool for information security management of intelligent organizations with emphasis on human factors that are critical in the context of security policy. However, due to increasingly expanding and complex legal issues in the area of e-business systems security, no security policy shall be treated solely as a techno-organizational issue. In fact, security policy is becoming primarily a legal document and should be treated as such. A lawyer or a legal expert should be involved in its derivation and maintenance process.

FUTURE TRENDS

Emerging new paradigms will be covered in this section. From the technology plane perspective these paradigms include objects, components, mobile code (computing), and intelligent agents. Every code (and object) can be treated as an electronic document. The creator defines its initial data and behavior (methods) and, optionally, signs it. The signature on the code gives a user the possibility to be assured of proper functioning of this object, where the problem is analogous to that of ensuring authentication and integrity for ordinary electronic documents. When considering intelligent mobile agents that are objects that satisfy certain conditions (Griss, 2001), the security paradigm is reversed. Agents operate in unpredictable environments and have to be protected from malicious hosts. These important issues have yet to be resolved.

With regard to organization plane, business dynamics can be further enhanced by linking it to *business intelligence* (Ortiz, 2002) in order to achieve an appropriate architecture for qualitative and quantitative management of security in intelligent organizations. The basis is operational security-related data that come from various sources: general host logs, router logs, application logs, phone and fax logs, physical security system logs, and so forth. These operational, legacy data have to be transformed in the next step, which means preparation of data for a security data warehouse. The exact data that have to be captured in a security warehouse are defined in line with the causal loop diagram model of an organization. This also implies that such a solution will have to be tailored to the needs of each particular organization.

Finally, an interesting trend is emerging where business dynamics is complemented by intelligent agents-based simulations (Trček, 2006). Business dynamics operates at the aggregates level and is top-down oriented, while agents-based simulations operate at the level of individuals and are bottom-up by nature. Thus using both methodologies as an input to business intelligence enables further improvement of security policy related decision making.

CONCLUSIONS

Even in the era of intelligent organizations, security management cannot avoid technical foundations. It is a fact that classical, cryptography-based approaches form the core of all security management so that each IS security has to start with addressing these issues. However, this basis serves as a starting point for further development of methodolo-

gies for risk management that are concentrated on human resources. Experience shows that human factors play an increasingly important role. Taking this into account and due to the emergence of business intelligence, it is possible to further support the management of IS security. Using business dynamics, intelligent agents, and business intelligence techniques, decision-makers can obtain data in real time and simulate the effects of their decisions in advance. This way they are armed with additional tools for successful protection of their ISs through appropriate security policies. And finally, increasingly expanding and complex legislation is becoming a fact in the area of e-business systems security, and thus requires proper attention.

REFERENCES

Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., & Levkowitz, H. (2004). *Extensible authentication protocol. RFC 3748*. Reston, VA: IETF.

ANSI. (1998). *The elliptic curve digital signature algorithm (ECDSA). X9.62 standard*. Washington, DC: ANSI.

Arsenault, A., & Turner, S. (2002). *Internet X.509 public key infrastructure roadmap. PKIX draft standard*. Reston, VA: IETF.

BSI. (1995). *Code of practice for information security management. British standard 7799*. London: British Standards Institute.

Cheswick, W., & Bellovin, S. (1994). *Firewalls and Internet security*. Reading, MA: Addison-Wesley.

Eastlake, D., & Jones, P. (2001). *Secure hash algorithm–1. RFC 3174 standard*. Reston, VA: IETF.

Foti, J. (Ed.). (2001). *Advanced encryption standard. FIPS 197 standard*. Washington, DC: Department of Commerce.

Freed, N. (1996). *Multipurpose Internet mail extensions. RFC 2045 standard*. Reston, VA: IETF.

Freier, A. O., Karlton, P., & Kocher, P. C. (1996). *Secure sockets layer protocol (version 3)*. Mountain View, CA: Netscape Corp. Retrieved from wp.netscape.com/eng/ssl3/index.html

Griss, L. M. (2001). Accelerating development with agent components. *IEEE Computer, 5*(34), 37-43.

IEEE. (2004). *Port based network access control. IEEE 802.1X*. Piscataway: IEEE.

ISO. (1995). *IT, open systems interconnection: Security frameworks in open systems. IS 10181/1 thru 7*. Geneva: ISO.

ISO. (2005). *Code of practice for information security management. ISO 17799 standard*. Geneva: ISO.

ITU-T. (2000). *Public key and attribute certificate frameworks. X.509 standard*. Geneva: ISO.

Kemmerer, R. A., & Vigna, G. (2002). Intrusion detection: A brief history and overview. *IEEE Computer, Security & Privacy, 35*(5), 27-30.

Miller, S. K. (2001). Facing the challenge of wireless security. *IEEE Computer, 34*(7), 16-18.

Ortiz, S. (2002). Is business intelligence a smart move? *IEEE Computer, 35*(7), 11-15.

Raepple, M. (2001). *Sicherheitskonzepte fuer das Internet*. Heidelberg: dpunkt-Verlag.

Ramsdell, B. (1999). *S/MIME message specification. Standard RFC 2633*. Reston, VA: IETF.

Rigney, C., Willens, S., Rubens, A., & Simpson, W. (2000). *Remote authentication dial in user service (RADIUS). RFC 2865*. Reston, VA: IETF.

RSA Labs. (2002). *PKCS–RSA cryptography standard, v 2.1*. Bedford: RSA Security.

Sterman, J. D. (2000). *Business dynamics*. Boston: Irwin-McGraw-Hill.

Thayer, R., Doraswamy, N., Glenn, R. (1998). *IP security document roadmap. RFC 2411*. Reston, VA: IETF.

Trček, D. (2006). *Managing information systems security and privacy*. Berlin/New York: Springer.

KEY TERMS

Business Intelligence: Deployment of (usually artificial intelligence-based) techniques such as online analytical processing (OLAP) and data mining to analyze information in the operational data sources.

Certification Authority (CA): An authority trusted by one or more users to create and assign public key certificates.

E-Business System: An organized, structured whole that implements business activities, which are based on electronic technologies, methodologies, and processes.

Networked Information Systems: An IS that is strongly integrated in a global network. From the technological point of view, the difference between the IS and the global network is blurred; however, it exists from the administrative point of view.

Public Key Certificate: The public key of a user, together with some other information, rendered unforgeable by encryption with the private key of the CA that issued it.

Public Key Infrastructure: The infrastructure capable of supporting the management of public keys able to enable authentication, encryption, integrity, or non-repudiation services.

Security Mechanism: A basis for a security service—using particular security mechanism (e.g., cryptographic algorithm), the security service is implemented.

Security Policy: Documented procedures that focus on the organization's management of security; it is about information confidentiality, integrity, and availability of resources.

Security Service: A service provided by an entity to ensure adequate security of data or systems.

Wireless Security: Security assurance for communicating information in electro-magnetic media over a distance through a free-space environment.

E-Collaboration in Organizations

Deborah S. Carstens
Florida Institute of Technology, USA

Stephanie M. Rockfield
Florida Institute of Technology, USA

INTRODUCTION

Organizations are shedding conventional work team structures in favor of virtual team structures that are increasing in popularity (Lee-Kelley, Crossman, & Cannings, 2004). E-collaboration enables collaboration between individuals not constrained by geographical distance or time. The emergence of the virtual team concept provides organizations with an alternate approach to managing work and individuals that are geographically separated (Gatlin-Watts, Carson, Horton, Maxwell, & Marltby, 2007). An advantage of virtual teams is that organizations can tap into resources rapidly to create a specialized work team that acts like a team, works like a team but doesn't look like a typical team because team members may not be co-located (Stough, Eom, & Buckenmyer, 2000). E-collaborative technologies such as computer-based conferencing systems are of critical importance to the success of a virtual team (Arnison & Miller, 2002). In the absence of water-cooler philosophizing, virtual teams rely on technology to build trust between team members, resulting in greater synergy and ultimately team success in carrying out work tasks (Arnison & Miller, 2002; Stough et al., 2000). The article focus is on technological and organizational aspects of e-collaboration occurring today and forecasted for tomorrow. The specific topics addressed are e-collaboration in organizations, e-collaboration in organizations of today, specific e-collaboration success factors and future trends of e-collaboration in organizations of tomorrow.

BACKGROUND

Collaboration is simply described as individuals working together while sharing information (Yen, Wen, Lin, & Chou, 1999). E-collaboration is merely taking collaboration to an electronic level. The expansion of the Internet has created opportunities to increase business collaboration, resulting in enhanced information sharing while reducing the amount of uncertainty in decision-making resulting in better profits (Rudberg, Klingenberg, & Kronhamn, 2002). E-business consists of more than buying and selling of goods and services on the Internet, as it also entails the servicing of customers and collaborating with business partners. Information management both internally and externally is an increasing concern for organizations as paper-based systems can be very slow, prone to error and difficult to update. With a growing interest in e-business solutions that facilitate information sharing between organizations, organizations within a supply chain are looking to achieve greater synergies with e-businesses and specifically participation in e-collaboration. By integrating e-collaboration in supply chain services through an electronic marketplace, companies are able to work together more efficiently through sharing vital information to assist in supply chain activities without the implementation of expensive EDI networks. With the globalization in business, e-collaboration has become almost a requirement for an organization to successfully compete in the marketplace in terms of optimizing productivity, quality and ultimately profits (Yen et al., 1999).

E-COLLABORATION IN ORGANIZATIONS OF TODAY

Contemporary workplaces are allowing employees to work from home and other off-site locations which are changing the view of traditional organizational work teams (Rudberg et al., 2002). Over the last decade, organizations are becoming flatter resulting in the need to increase the number of virtual teams that bring internal and external people with diverse disciplines together (Dustdar, 2005). These virtual teams survive based on a combination of mobile and fixed people, devices and applications. E-collaboration technologies influence how people process, manage and manipulate information, which is useful for both co-located and virtual work teams (Rudberg et al., 2002). E-collaboration provides a mechanism for virtual teams to accomplish work tasks while providing the foundation for productive team work accomplished through nontraditional means. Team-based structures in workplaces are common and in great demand because work teams are often comprised of individuals with different backgrounds due to the nature of work being multidisciplinary as well as global. In working on joint projects between businesses, there can be a reduction in the risks, making innovation less costly for any one business with regard to new ventures.

There are several e-collaborative technologies available today that allow organizations to have instant communication within a business, between businesses or between businesses and consumers (Rudberg et al., 2002). The range of activities that e-collaborative technologies support is e-mail to videoconferencing to file sharing, regardless of geographical time and distance. Groupware commonly consists of e-mail, computer-based conferencing systems, collaborative writing, programming and drawing systems (Stough et al., 2000; Yen et al., 1999). Groupware enables virtual teams to experience more traditional group life because team members can interact with each other by actually seeing each other and jointly manipulating information real-time. Groupware can also be called Teamware, as the focus of both is in assisting groups or teams in successfully performing work, whether helping a face-to-face team with brainstorming anonymously or helping distributed teams see each other through videoconferencing (Yen et al., 1999). Technology is useful for the exchange and sharing of information (Yen et al., 1999). There are many different products available which integrate several features such as e-mail, calendaring, instant messaging, multiperson chats or discussions, mobile, wireless, social software, team collaboration, e-forms, enterprise-wide discussion forums, team spaces, Web-conferencing, instant remote access control, instant desktop sharing, two-way file sharing, online meetings and Web-content management (Bal & Teo, 2001).

However, technology alone will not constitute virtual team success. Teams that switch from a collaboration to an e-collaboration environment will likely experience a transformation process that involves complex team dynamics and work processes that at first may create new challenges. For example, virtual teams may need to find alternative ways to form trust between team members, as water cooler philosophizing as a way for team members to form bonds don't exist as easily in virtual teams (Arnison & Miller, 2002; Stough et al., 2000). Early literature of the 1930s and 1940s approached conflict as dysfunctional for group dynamics, resulting in researchers studying causes and prevention of conflict (Passos & Caetano, 2005). Later literature discussed other views such as the human relations' approach, where conflict was viewed as natural or even positive for teams to experience, resulting in studies to identify situations where conflict was either good or bad for team dynamics. Conflict is another aspect of teams where virtual teams have an added challenge when resolving. Molleman (2005) suggests those individual team members' attributes such as age, skills or personality traits can be noticeable, which can lead to subgroups and possibly result in conflict among the subgroups. Group cohesion, mutual trust and team efficacy can be shared team characteristics, provided e-collaborative technologies aid in relationship building for teams (Bal & Teo, 2001; Molleman, 2005).

ACHIEVING SUCCESS TODAY AND TOMORROW WITH E-COLLABORATION

E-collaboration technologies enable virtual teams or even face-to-face teams to be more productive and therefore successful. When considering a virtual team, it is important to also consider all of those factors which contribute to the team's success. With most virtual teams being separated by geographical time and distance, it is not surprising that a team's success can be greatly influenced by the availability and capability of technology (Stough et al., 2000). Aside from specifically e-collaboration technologies, team members that have access to different knowledge databases also have greater opportunities to have successful collaborative efforts in developing new ideas, products, markets, strategies, organizational designs and visions (Shani, Sena, & Stebbins, 2000). Technology assists in all collaborations, making it easier for individuals with different backgrounds to come together and virtually resulting in higher creativity and innovation. Technologies such as the Internet, e-mail, groupware, video-conferencing, cellular phones and intranets will continue to aid in information sharing and communication between team members (Arnison & Miller, 2002).

There are several identified sources that lead to creative performance as being resource availability, leadership, group size, group cohesiveness, communication patterns and group diversity (Molleman, 2005; Shani et al., 2000). Furthermore, the structure of group interaction within teams significantly impacts team members' creativity. The work on a team can be delegated best when team members are aware of each other's skill set even when tasks are unclear (Cruz, Perez, & Ramos, 2007). Technology has been a contributor leading to improvements in communication for virtual team members through enabling more closely knit work teams (Arnison & Miller, 2002). Of equal importance, the virtual team must be able to keep up with any new technologies utilized to aid team performance. Training in e-collaborative technologies may be a necessity for the vitality of virtual teams. Technologies have drastically improved, yet the mobility of technology is still in need of being improved, as employees need to have the ability to work anytime anywhere in order for an organization to compete globally in the marketplace.

Virtual teams must have a purpose and vision to ensure that every team member is focused and understands the goals of the team. If a virtual team lacks productive interactions, the effectiveness of the team could be compromised (Arnison & Miller, 2002). Because virtual teams may not have a chance to check-in with their team members as often as face-to-face teams, team members must be provided with clear goals to help the team in achieving a cohesive contribution (Arnison & Miller, 2002; Bal & Teo, 2001; Molleman, 2005). Team members in a virtual team as well as a face-to-face team must have an identity within their team through having

an understanding of exactly how they fit within their team structure (Arnison & Miller, 2002). Also, teams must have an understanding of the role played by every member to assist teams toward achieving their goals. Without clarity in the role played by an individual on a virtual team as well as the roles played by others on the team, poor interaction among members could occur because of a lack of understanding of how each team member's performance impacts the other members and ultimately the virtual team's goals.

Poor interaction can also lead to the lack of trust among team members, resulting in reduced opportunities for cohesiveness. Duarte and Snyder (1999) identified three factors that assist virtual teams in building trust among team members. The three factors are performance and competence, integrity and concern for the well-being of others. If team members view each other as high performers and as competent, trust can be built. Integrity can be preserved due to various collaboration technologies that have anonymity features of team member's input. Bal and Teo (2001) identify that trust can be built through ensuring that all team members have understanding of the team's purpose and process for continuous improvements. Virtual teams that are self-managed teams are unlike traditional work groups in that the leaders can be more informal leaders and the leadership role may even be divided among many of the team members (Yukl, 2002). Virtual teams can be easily compared to self-managed teams due to the definite and distinctive independence that exists in virtual teams caused from the distance between team members (Arnison & Miller, 2002). Strong leadership is not only essential but also the key for the successful operations of a purely virtual team. Leadership is of even further importance when the skills and knowledge within the team are highly diverse. Individuals need to stay motivated in performing work and this may even be more crucial for virtual teams (Stough et al., 2000). Therefore, even small wins should not go unnoticed by the leadership of the virtual teams. Virtual teaming effectiveness measures should be identified and utilized as an aid not only to virtual team leaders but team members as well to contribute toward keeping the team focused and on task. Furthermore, there should be shared rewards among team members for milestones being accomplished.

FUTURE TRENDS

Rudall and Mann (2007) discuss how computing is shifting to be more cognizant of the importance of a relationship between computing and our environments, taking into consideration both social and technological aspects. Duarte and Snyder (1999) identified how different modes of communication such as audio, video or data assist with different types of tasks. For instance, technical or interpersonal conflicts are not easily resolved through e-mail. Therefore, e-collaborative technologies of the future should build on past research to ensure that new features optimize team interactions to enhance productivity and quality of the team output.

The 2010 Internet access network is anticipated to have an improved capacity from the currently available spectrum, as it will begin to move away from a cellular-only system to one that integrates broadcast, cellular, cordless, wireless LAN, point-to-multipoint and fixed access technologies (Reynolds, 2003). With the expansion of smart systems, e-collaborative technologies will benefit through virtual team members being empowered through digital environments that have intelligence built-in to the design (Rudall & Mann, 2007). For instance, digital environments are being predicted to have the ability to be aware of people's presence and context. Furthermore, the new digital environments would have capabilities such as being sensitive, adaptive and responsive to individuals' needs, habits, gestures and emotions of those individuals' part of the environment. These smart systems will eventually achieve a level of maturity resulting in computers communicating back and forth with each other, taking into account the needs of surrounding individuals as well as the environment while being completely hidden within environments.

The potential for enhanced smart systems with embedded intelligence has resulted in new and emerging fields such as the combining of pervasive computing, cognitive intelligence and expansion of software-intensive systems (Rudall & Mann, 2007). Pervasive computing can be described as computing that is embedded into professional or personal environments providing seamless computing anytime and anywhere. Whereas, cognitive intelligence refers to computing that is able to act as intelligent agents that not only understand individuals' mental states but also have the capability to socialize with individuals, much like individuals interact with each other. There are a variety of technologies being researched such as pervasive computing, energy autonomy and networking. The Coopers project focuses on integrated traffic management with a vision that vehicles connect through continuous wireless communications with road infrastructures in order to exchange data and information that would be relevant for specific road segments. This information would increase overall road safety through enhancing traffic management. Ultimately, future e-collaboration technologies must assist the growing number of virtual teams by not merely making mobility of content but also making mobility of context (Dustdar, 2005). Content is described as any information (work procedures, work documents, etc.) that teams need to have access to, whereas context refers to the ability to trace and support dynamic relationships between people, information and team processes.

CONCLUSION

Successful virtual teams must rely on several factors such as e-collaboration technologies, highly skilled team members and strong leadership (Stough et al., 2000). Virtual teams should have access to technologies that go beyond groupware products that focus on mobility of content but also technologies that assist with mobility of context that integrates managing work, people and their processes (Dustdar, 2005). These factors provide the setting and appropriate groundwork for success (Stough et al., 2000). Collectively, these factors support activities such as setting clear goals, coordinating and negotiating with others, building trust among team members, planning and managing work processes, enhancing decision-making skills, contributing to relationship building and performing management functions to include budgeting and scheduling (Bal & Teo, 2001; Molleman, 2005; Stough et al., 2000). Virtual teams as well as face-to-face teams must be recognized and rewarded by their organization for their successes (Stough et al., 2000). E-collaborative technologies of today appear on first glance as groundbreaking, but the e-collaborative technologies of the future haven't even begun to break ground.

REFERENCES

Arnison, L., & Miller, P. (2002). Virtual teams: A virtue for the conventional team. *Journal of Workplace Learning, 14*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/13665620210427294

Bal, J., & Teo, P.K. (2001). Implementing virtual teamworking: Part 2—a literature review. *Logistics Information Management, 14*(3), 208-222.

Cruz, N.M., Perez, V.M., & Ramos, Y.F. (2007). Transactive memory processes that lead to better team results. *Team Performance Management, 13*(7/8), 192-205.

Duarte, D.L., & Snyder, N.T. (1999). *Mastering virtual teams*. San Francisco: Jossey-Bass.

Dustdar, S. (2005). Architecture and design of an Internet-enabled integrated workflow and groupware system. *Business Process Management Journal, 11*(3), 275-290.

Gatlin-Watts, R., Carson, M., Horton, J., Maxwell, L., & Maltby, N. (2007). A guide to global virtual teaming. *Team Performance Management, 13*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/13527590710736725

Lee-Kelley, L., Crossman, A., & Cannings, A. (2004). A social interaction approach to managing the "invisibles" of virtual teams. *Industrial Management & Data Systems, 104*(8), 650-657.

Molleman, E. (2005). The multilevel nature of team-based work research. *Team Performance Management, 11*(3/4), 113-124.

Passos, A.M., & Caetano, A. (2005). Exploring the effects of intragroup conflict and past performance feedback on team effectiveness. *Journal of Managerial Psychology, 20*(3/4), 231-244.

Reynolds, P. (2003). A vision of the Internet in 2010. *Campus-Wide Information Systems, 20*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/10650740310491289

Rudberg, M., Klingenberg, N., & Kronhamn, K. (2002). Collaborative supply chain planning using electronic marketplaces. *Integrated Manufacturing Systems, 13*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/09576060210448170

Rudall, B.H., & Mann, C.J.H. (2007). Smart systems and environments. *Kybernetes, 36*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/03684920710747066

Shani, A.B., Sena, J.A., & Stebbins, M.W. (2000). Knowledge work teams and groupware technology: Learning from Seagate's experience. *Journal of Knowledge Management, 4*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/13673270010336602

Stamm, B.V. (2004). Collaboration with other firms and customers: Innovation. *Strategy & Leadership, 32*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/10878510410535727

Stough, S., Eom, S., & Buckenmyer, J. (2000). Virtual teaming: A strategy for moving your organization into the new millennium. *Industrial Management & Data Systems, 100*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/02635570010353857

Yen, D.C., Wen, J., Lin, B., & Chou, D.C. (1999). Groupware: A strategic analysis and implementation. *Industrial Management & Data Systems, 99*. Retrieved May 30, 2008, from http://www.emeraldinsight.com/10.1108/02635579910243879

Yukl, G. (2002). *Leadership in organizations* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

KEY TERMS

Cognitive Intelligence: Cognitive intelligence refers to computing that is able to act as intelligent agents that not only understand individuals' mental states but also have the capability to socialize with individuals much like individuals interact with each other.

Collaboration: Collaboration occurs when two or more individuals work together to accomplish a work goal.

E-Collaboration: E-collaboration occurs when two or more individuals work together with the use of technology to assist in accomplishing a work goal.

E-Collaborative Technologies: E-collaborative technologies consist of any form of computing that is used to assist a virtual team in accomplishing a work goal.

Virtual Teams: Virtual teams are work teams where a minimum of one team member is separated from other team members due to geographical time and distance.

Groupware: Groupware is a type of e-collaborative technology developed solely for the purpose of supporting virtual teams in communicating and information sharing to accomplish their work goals.

Pervasive Computing: Pervasive computing can be described as computing that is embedded into professional or personal environments to provide seamless computing anytime and anywhere.

E–Commerce Taxation Issues

Mahesh S. Raisinghani
TWU School of Management, USA

Dan S. Petty
North Texas Commission, USA

INTRODUCTION

This article is designed to give the reader a balanced perspective on some of the issues surrounding the current discussions related to state and local taxation of Internet access fees and sales transactions. It attempts to express the issues being discussed and presents several viewpoints. The proponents of Internet taxation are searching for technological and administrative system to meet their goal. After much deliberation, the Advisory Commission on Electronic Commerce released its final recommendations to Congress in April 2000. Major emphasis is being placed on simplification, neutrality, avoiding double taxation and accepting the existing tax rules with no new taxes.

The United States economy has benefited tremendously by e-commerce. This escalation has created numerous highly skilled jobs, providing the consumer with goods and services at competitive prices. The Internet Tax Fairness Coalition and many other groups feel that implementing taxes on the Internet transaction can have an adverse affect on the businesses. According to the Supreme Court of United States, a vendor has a sales tax obligation only when the buyer and seller are in the same state or has a physical presence (nexus) in the buyer's state. These coalitions feel that entry barriers for new and old companies, who have yet to exploit the e-commerce, will slow the growth in this sector. With over 30,000 taxing jurisdictions, tax collection and payment can be a complex process. Many street retailers collect at a single rate, and prepare and file a single tax return at one place. Taxation of online transactions would require the vendor to identify and send forms to all taxing jurisdictions. Under the present circumstances, the ever-changing maze of state and local tax policies makes application of a single Internet transaction tax policy virtually impossible.

The complicated, complex and ever changing maze of state and local tax policies and laws make application of a sensible, fair and easily understood Internet transaction tax policy virtually impossible under the present circumstances. James Plummer, a policy analyst at Consumer Alert wrote, "Nefarious new taxes and regulations will kill many new start-up e-businesses before they even start up; denying consumers their chance to find the specialized products and services for their needs" (Plummer, 2000).

The anti-tax community and coalitions have a strong adversary in the National Governor's Association. The State is worried that the brick and mortar stores are jeopardized by the popularization of Internet commerce, which is tax-free. The Governors suggest that government tax policy offers a competitive advantage to Internet stores. Major brick and mortar retailers such as Sears and Wal-Mart are concerned that if unresolved, this issue may gain much public resistance, thus making the taxing of e-commerce politically impossible.

BACKGROUND

The United States Congress enacted The Internet Tax Freedom Act in 1998, imposing a three-year moratorium on new Internet taxation. It also established the Advisory Commission on Electronic Commerce to address the issues related to Internet taxation (Advisory Commission on Electronic Commerce, 2000).

The Advisory Commission has representatives from state and local governments and e-commerce industry. It is to conduct a study of federal, state, local and international taxation and tariff treatment of transactions using the Internet and Internet access, and other comparable sales activities. The Commission's recommendations are to be submitted to Congress no later than April of 2000. Based on testimonies, the Commission is reviewing barriers imposed in foreign markets on U.S. property, goods or services engaged in Internet and its impact on U.S. consumers, and ways to simplify federal, state and local taxes imposed on telecommunications services.

The National Governor's Association's Perspective

Today, 46 states have a sales tax of some sort. All of the 46 states that have a sales tax also have what is called a complementary use tax. Consumers pay the sales tax when they buy goods and services in their own state and use tax when they buy from other states. This strategy avoids double taxation. When the consumer buys from an out of state merchant, such as mail order or the Internet, tax is collected and sent to the consumer's state only if the merchant has a nexus in the consumer's state. According to U.S. Supreme Court 1967, National Bellas Hess and 1992

Quill decisions, the merchant is not required to collect the use tax and remit it to the state of residence of the consumer. Consumers then are responsible for paying taxes on goods they purchase through mail-order catalogues and over the Internet. This subsidizes one category of businesses at the expense of their competitors.

The Governors have suggested a streamlined sales tax system for the 21st century. Some of the features of the governor's proposed streamlined system include:

- Maintain the current definitions of nexus and eliminate collection of state and government taxes
- Simplify the current system and without any federal government intervention
- Eliminate the cost of compliance, tax returns and payments and tax audits
- Eliminate tax-rate monitoring and implementation.
- Eliminate risks for sellers exercising reasonable care

The states would implement uniform laws, practices, technology applications, and collections systems to achieve the goals and results. These goals, when implemented, would achieve the first step of the streamlined system. The second step would be for all state and local governments to adopt the same classification systems, definitions and audits

The overall concept of the streamlined system is to reduce the costs and burden of sales tax compliance by shifting sales tax administration to a technology oriented business model operated by trusted third parties (TTPs).

THE E-COMMERCE COALITION PERSPECTIVE

The e-Commerce Coalition is a broadly based national coalition dedicated to providing sound policy information on electronic commerce taxation, and includes AOL, Bank One, Cisco Systems, and the like. Can e-commerce step up the process of making each state responsible for administration of its own tax system and simplification? Time is of the essence because of the speed at which this industry is growing and changing.

GOVERNOR JAMES GILMORE'S PERSPECTIVE

Governor James Gilmore of Virginia is Chairman of the Advisory Commission on Electronic Commerce. Governor Gilmore submitted a proposal to the Commission on November 8, 1999 entitled "No Internet Tax" (Gilmore, 1999).

The Governor believes that American public policy should embrace the Internet and the borderless economy it creates.

- Prohibit all sales and use taxes on business-to-consumer interactions and protect companies from unfair taxes imposed due to their virtual presence
- Amend the Tax Freedom Act to prohibit all taxes on Internet access
- Abolish the federal 3% excise tax on telephone service
- No international tariffs or taxes on e-commerce

E-COMMERCE TAXATION FROM AN INTERNATIONAL PERSPECTIVE

The problem with international taxation is essentially that it is likely not possible to govern well when there is no international government to create an appropriate incentive structure to induce and compel good behavior (Bird, 2003). Molina and Michilli (2003) discuss how e-commerce is emerging in European regions. For instance, the region of Veneto has implemented "Bollo Auto" — the payment of car taxes through a digital network making use of lottery terminals located in the popular tobacco shops (*tabacchinos*). Since July 1, under the aegis of "leveling the playing field," the European Union (EU) has been imposing a value-added tax (VAT) on digital goods — namely games, music, and software — downloaded from non-EU companies via the Internet by EU citizens (Pappas, 2003). McLure (2003) concludes that any failure to apply value-added tax (VAT) to electronic commerce crossing borders between EU member states and other countries should not affect the value added-tax liability of registered traders, even if the reverse charge rule (taxation in the hands of recipients) is not applied. The study notes that the sales of digital content to consumers and unregistered traders that constitutes a minuscule fraction of purchases by households and unregistered traders (given the extremely low level of small-business exemptions) is problematic.

Li (2003) provides a technical and policy analysis of the Canadian Goods and Services Tax (GST) in the context of e-commerce and suggests some options for reform. Even though the GST has had a bad reputation in Canada and its integrity is now threatened by growing online cross-border shopping, based on the revenue potential of the GST, a replacement is highly unlikely, and a cleaned-up or reformed GST is more practical. Thus the government should take advantage of the opportunity presented by e-commerce to reform the GST.

In an application of the substance over form doctrine to the international e-commerce taxation issue, Ngoy (2003) proposes an approach consisting of applying what is called

here the permanent establishment (PE) function test to e-commerce infrastructures in order to see whether they qualify for being fiscally treated as PE, if they pass the concerned test. The study concludes that some of them substantially have the same function as the category of office PE, and they should be fiscally treated as this category of PE no matter the form they have.

SUMMARY OF OPTIONS FOR RESOLUTION OF INTERNET TAXATION ISSUES

The European Union has worked out a system where all 15 members impose hefty value-added taxes and all retailers must collect the tax for all sales within the union. United States negotiators are working with groups internationally to come to some understanding regarding these complex issues (Landers, 2000). Senator Ron Wyden and Representative Christopher Cox state that they would work to extend the e-tax moratorium for 5 more years and permanently bar all access taxes (Business Week, 2001).

On the other hand, the taxing entities feel that the online retailers should have to collect the same taxes as brick and mortar businesses. They have signed on to the Streamlined Sales Tax Project, whereby they have agreed to pattern their tax system from a model code (The Wall Street Journal, 2001). Four states, California, Massachusetts, Virginia and Colorado, who are leading technology states, put forth the argument that taxing Internet access and commerce would harm the growth sector of their economies.

Congress has taken the first steps to make sure states cannot impose sales taxes on Internet access, use or content and cannot impose multiple or discriminatory new taxes. Congress has also encouraged and supported the work of the Streamlined Sales Tax Project so states can continue to receive sales tax revenues from Internet transactions, just as they would if the transactions took place in a brick and mortar location. Retail businesses that enter into Internet commerce will be required to collect sales taxes on Internet sales in accordance with the rules that state develops. Sales taxes will remain solely under the jurisdiction of state and local governments for the foreseeable future (Goold, 2003).

The Streamlined Sales Tax Agreement (SSTA) sets uniform definitions and other standards that will make it easier for retailers to collect tax from out-of-state purchasers. The SSTA would make it easier for states to collect taxes on Internet purchases — an extra-territorial money grab strikingly similar to the EU's plan. Although 31 states have already approved the new SSTA, the agreement is voluntary since remote Internet and mail-order sellers are still not legally obligated to collect any tax from out-of-state consumers. However the new Simplified Sales and Use Tax Act of 2003 introduced in Congress by Representative Ernest Istook (R-OK) and a companion bill sponsored by Senator Mike Enzi (R-WY) would effectively make the SSTA mandatory (Pappas, 2003; Rankin, 2003).

Most of the 45 states in the United States that impose sales and use taxes consider the advent and expansion of e-commerce the greatest threat to their financial stability since catalog merchandising. The Constitution's Commerce Clause grants Congress the power to regulate commerce among the several States. In Quill Corp. v. North Dakota, the Supreme Court revisited its long-standing rule that a state could not establish nexus with a remote seller unless the remote seller was physically present in the state. The U.S. Supreme Court reaffirmed that nexus cannot be established in the absence of a remote seller's physical presence in a state (Trealease & Storum, 2003).

FUTURE TRENDS

After reviewing some of the information available, it appears that an interim solution might evolve from the final report and recommendations of the Advisory Commission on Electronic Commerce, including an extension of the moratorium on taxes on Internet access. The infant industry argument is that a tax hurts an emerging industry that still needs a small boost to continue expanding and developing new products and new technology. Critics of Internet taxation charge that it is also unconstitutional and unfair. However, "unfair" is also what land-based retailers might cry, as they claim that Internet retailers have an undue advantage. For both catalog and Internet sales, states want to be able to tax those purchasers who live in their state (Jossi, 2003). Although there are serious and complex issues, it appears that a resolution can be constructed that will be favorable to consumers and businesses alike, including technology being applied to the collection process, standardization of tax systems, and state and local governments being responsible to pay the costs of newly developed and technologically sophisticated collection systems. In February 2003, high-profile retailers such as –INCLUDEPICTURE "http://proquest.umi.com/images/common/circlei3.gif" * MERGEFORMATINET —Wal-Mart, –INCLUDEPICTURE "http://proquest.umi.com/images/common/circlei3.gif" * MERGEFORMATINET —Target, Toys R Us, Marshall Field's, and –INCLUDEPICTURE "http://proquest.umi.com/images/common/circlei3.gif" * MERGEFORMATINET —Mervyn's began voluntarily collecting sales taxes from their online customers. However, substantial nexus is the keystone of a state's taxing jurisdiction over remote sellers. If substantial nexus does not exist between the state and a remote seller (whether an Internet seller or not), the state may not validly impose tax liability on the remote seller or require it to collect and remit such tax. Perhaps, just as the Internet added new confusion to the

area of sales and use taxation, emerging technology-based alternatives may eventually help Internet sellers efficiently carry some of the burdens of multistate tax compliance (Trealease & Storum, 2003). A system currently being used in Europe contains most of the features that states would find necessary for the proper and efficient collection of their taxes. The European Union has worked out a system where all 15 members impose hefty value-added taxes and all retailers must collect the tax for all sales within the union. That is, an Internet purchase made in Germany by a customer in Portugal gets taxed at the VAT rate for Portugal; the German seller collects it. United States negotiators are working with groups internationally to come to some understanding regarding these complex issues (Landers, 2000). State and local governments have acknowledged that their system of sales and use taxes must change in a substantial manner if they are to remain viable in the 21st century.

CONCLUSION

Seven criteria have been laid out for use in designing an acceptable cyber tax system. The system should be equitable and simple, ensure user confidence, prevent tax evasion and economic distortion, maintain a fair balance among countries, and not introduce a new form of taxation (Lee & Hwangbo, 2000). Without the assurance of a uniform nationwide approach, even the most sophisticated technological solution will collapse. It is critical to resolve the e-commerce taxation issue by finding a feasible way to implement a multi-state system for collecting taxes from literally hundreds of tax jurisdictions across the country. The Streamlined Sales Tax Project has been launched by some 30 state governments "to develop a radically simplified sales and use tax system that eases the burden of state use and tax compliance for all types of retailers, particularly those operating on a multi-state basis" (Rankin, 2000). The outcome will have long-term consequences for U.S. retailing, and, according to some, for the American system of government itself.

REFERENCES

Advisory Commission on Electronic Commerce. (2000). Retrieved April 29, 2000, from *http://www.ecommerce commission.org/FAQs.htm*

Bird, RM. (2003). Taxation and e-commerce. *The Canadian Business Law Journal, 38*(3), 466.

BusinessWeek, 49 (2001, February 19). The other tax battleground of 2001: The Internet.

The e-Freedom Coalition. (2000). Retrieved April 29, 2000, from *http://www.policy.com/news/dbrief/dbrief arc453.asp*

Gilmore, J., III. (1999). No Internet tax proposal. Retrieved April 18, 2000, from *http://www.ecommercecommission.org/proposal*

Goold, L (2003). Point, click, tax? *Journal of Property Management, 68*(5), 20.

The Internet Tax Fairness Coalition. (2000). Retrieved April 29, 2000, from *http://www.nettax.fairness.org/facts*

Jossi, F. (2003). The taxing issue of e-commerce. *Fedgazette, 15*(6), 9.

Kyu Lee, J., & Hwangbo, Y. (2000, Winter). Cyber consumption taxes and electronic commerce collection systems: A canonical consumer-delivered sales tax. *International Journal of Electronic Commerce, 4*(2), 6-82.

LaGesse, D. (2000). Governor George W. Bush. *The Dallas Morning News,* 1D.

Landers, J. (2000). Internet tax issues. *The Dallas Morning News,* 1D.

Li, J. (2003). Consumption taxation of electronic commerce: Problems, policy implications and proposals for reform, 38(3), 425.

McLure, Jr., C.E. (2003). The value added tax on electronic commerce in the European Union. *International Tax and Public Finance, 10*(6), 753.

Molina, A., & Michilli, M. (2003). E-commerce innovation in the Veneto region: Sociotechnical alignment in the context of a public administration. *International Journal of Entrepreneurship and Innovation Management, 3*(4), 415.

National Governor's Association. (2000). Retrieved April 29, 2000, from *http://www.nga.org/internet/overview.asp*

National Governor's Association. (2000). Retrieved April 29, 2000, from *http://www.nga.org/internet/facts.asp*

National Governor's Association. (2000). Retrieved April 29, 2000 from *http://www.nga.org/internet/proposal.asp*

National Tax Association Communications and Electronic Commerce Tax Project Final Report, vi-vii. (1999, September 7).

Ngoy, J.M. (2003). Is international e-commerce an HIV tax issue? *International Journal of Services Technology and Management, 4*(1), 53.

Pappas, M. (2003). Europe's global tax. *Foreign Policy, 139,* 92.

Parrish, R.L. (1999). *Tandy/Radio Shack corporation comments to advisory commission on electronic ecommerce.* E-mail correspondence forwarded by the author.

Plummer, J. (2000). Consumer alert. Interview. Retrieved April 29, 2000, from *http://www.policy.com/news*

Rankin, K. (2000, August). Race against time: Seeking a net sales tax solution. *ECWorld,* 26-28.

Rankin K. (2003). Tax-free Internet sales may be ending soon. *DSN Retailing Today, 42*(22), 9.

Releaser, N.T., & Storum, L.A. (2003). The gathering storm: State sales and use taxation of electronic commerce. *Corporate Taxation, 30*(3), 9, 16.

The Wall Street Journal. (2001, March 7). States at odds over Web taxes, B3.

Wyld, D.C. (2003). Don't shoot the Internet. *Computerworld, 37*(47), 21.

KEY TERMS

Double Taxation: When the same taxable item is taxed more than once by either the same or by different government agencies, there is said to be double taxation. The juridical type of double taxation happens when comparable taxes are imposed by two or more taxing jurisdictions on the same taxpayer in respect of the same taxable income or capital.

E-Commerce: Conducting commercial transactions on the Internet, where goods, information or services are bought and then paid for.

Moratorium: Temporary suspension of payments due under a financial or tax agreement. For example, a governmental body may offer a tax moratorium as an incentive to entice a business to locate and/or start up operations in its jurisdiction.

Nexus: The general concept of some connection or link to the taxing jurisdiction. In the U.S., jurisdiction for levying taxes has a Constitutional basis.

Sales Tax: An excise tax imposed on the transfer of goods, typically at retail. States vary as to whether the tax is imposed on the seller of goods or on the buyer; however, sales taxes are almost universally collected from the purchaser at the time of sale.

Use Tax: A complementary or compensating tax imposed by all states that impose a sales tax. Use taxes are typically charged on the "storage, use, or consumption" of goods in the taxing state. Liability to remit use taxes usually falls on the buyer of taxable property or services. Since it is administratively difficult to compel individual self-assessment of use taxes, most of those taxes will go uncollected unless the states can compel sellers to collect them. Significantly, a state may impose use tax collection responsibilities on Internet sellers if they have nexus with the state. The use tax is intended to stem the erosion of the sales tax base when a state's residents purchase taxable goods or services from sellers located outside of the state.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 957-961, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

E-Contracting Challenges

Lai Xu
CSIRO ICT Centre, Australia

Paul de Vrieze
CSIRO ICT Centre, Australia

INTRODUCTION

A decade ago, IT — through its innovations in business process reengineering — led the way in breaking down the inefficiencies within companies. Firms in the new millennium now face relentless pressure to perform better, faster, cheaper, while maintaining a high level of guaranteed results. Firms must thus focus on their core competencies and outsource all other activities. Working with a partner, however, requires breaking down the inefficiencies between organizations and coping with frequent change across the entire end-to-end value chain. In this new world of collaborative commerce and collaborative souring, a standard business process is simply inadequate. Using e-contracts to build new business relationships and to fulfill e-contracts through the Internet are important trends. E-contracting is however not a new concept. The history of e-contracting can be reviewed from legal and technology aspects.

Over the last 20 years or so, a growing body of research in artificial intelligence has focused on the representation of legislation and regulations (Sergor, 1991). As specific regulations, contracts are used to regulate the actions of two- or multi-party interactions. Gardner (1987) has developed contract formation rules. Her work concerns legislation about the nature of exchanges that lead to contractual relations. The ALDUS project and Legal Expert project investigated drafting the Sale Goods contract (ALDUS, 1992) and the United Nations Convention on contracts for the international sale of goods (Yoshino 1997, 1998), respectively. Detailed information on developing logic-based tools for the analysis and representation of legal contracts can be found in Daskalopulu (1997, 1999).

The law regards contracts as collections of obligations; research in this area includes automated inference methods, which are intended to facilitate application of the theory to the analysis of practical problems. The purpose of a legal e-contracting system is to clarify and expand an incomplete and imprecise statement of requirements into a precise formal specification.

In the early 1990s, the development of EDI (electronic data interchange) was a significant movement for electronic commerce. EDI was considered a term that refers solely to electronic transactions and contracts (Justice Canada, 1995). EDI requires an agreement between trading partners that not only dictates a standard data format for their computer-to-computer communications, but also governs all related legal issues of EDI usage. In 1987, the first set of EDI rules was named the Uniform Rules of Conduct for Interchange of Trade Data by Teletransmission (UNCID, 1987). In 1990, the American Bar Association (ABA) published a Model Trading Partner Agreement and Commentary, together with an explanatory report (Winn & Wright, 2001). In 2000 IBM submitted to OASIS (for standardization) the first example of an XML-based EDI TPA language, called Trading Partner Agreement Markup Language (tpaML).

While the EDI standard introduced efficient communication channels between companies, its implementation was not widely accepted due to its high installation costs, lack of flexibility, and technological limitations (Raman, 1996). With the development of the Internet, electronic contracting began to be interpreted in broader terms. In this new view, an e-contract is not only used as a legally binding agreement between a buyer and seller, but it can also used across different workflow systems to cross different organizational business processes (Koetsier, Grefen, & Vonk, 1999; Kafeza, Chiu, & Kafeza, 2001; Cheung, Chiu & Till, 2002) to integrate different Web services (Cheung et al., 2002, 2003). E-contracting has become synonymous with business integration over electronic networks.

BACKGROUND

New technologies, the Internet, and other networks have changed business environments and provided the trading processes in e-business more efficiency. Legal regulations, such as the European directive for electronic signatures (EU Directive, 2000) and national e-commerce regulations, have set up a framework for using electronic contracts in business. Concepts of e-contracts under the network environment definitely have different characteristics than the concepts for traditional paper contracts. Whereas a paper-based contract document is a static view on the obligations, an e-contracting system could monitor the responsibilities of each contractual party and the performance of the obligations.

In a networked environment, the definition of the concept of e-contracts can be emphasized as "a contract is a guarantee" or "contracts build new business collaborations between contractual partners" (Xu, 2004a). First, the contract provides a guarantee to all contractual partners according

to the clauses of the signed contract and relevant laws. An agreement between consumers and retailers in B2C commerce is a typical example of "a contract is a guarantee." The agreement provides protection to both consumers and retailers. Second, contractual partners build a business relationship using a contract such as an "arm's length transaction." Two (or multi) parties who used different workflows can cooperate using e-contracts to support business automation (Koetsier et al., 1999; Kafeza et al., 2001). Web service composition can also be implemented using e-contracts. There also exist some e-contract applications that actually cover both sides' concepts. For instance, Trading Partner Agreement (TPA) in ebXML provides a guaranteed business exchange with a certain quality. It also specifies a long-term business relationship/collaboration between partners to conduct the business. It is important to realize though that the concept of e-contract has only a partial overlap with the concept of a paper contract. Both have features that do not have their representation in the other one.

Under electronic communications, e-contracting processes have their unique characteristics. The result of e-contracting, the contract, is a semi-structured document, which stored in any format (e.g., MS Word, PDF, or XML, etc.). Most e-contracts contain semi-structured information, such as XML-based words, sentences, clauses, or meta-information. Furthermore, some e-contracts have the legal status of digital documents. Depending on whether networks are used during the e-contract establishment stage, e-contracts can be created online (i.e., through networks) or electronically without networks. The collaboration in the contract formation phase can be asynchronous (e.g., by e-mail) or synchronous (e.g., through online collaboration). Moreover, the e-contracting process can be finished on a shared platform (e.g., e-marketplace) or be interconnected between contractual partners. The ownership of the contracting platform can thus be with a third party (ASP) or with the contractual partners. The e-contract can be fulfilled online (e.g., digital goods, services) or off-line (e.g., physical goods). As e-contracts serve different purposes, different opportunities will bring extra values during the e-contracting stages. For example, in the contract execution/performance stage, extra monitoring information can be provided by different messages over networks. This is a significant difference with traditional contracting.

In short, e-contracting can protect contractual partners in electronic environments, reduce time-to-contract, and reduce process costs. It can also provide new opportunities on contract management, contract content re-use, and contract monitoring. Benefits of e-contracting are:

- Avoiding errors,
- re-using content after closing,
- reducing time-to-contract,
- providing machine-processable document,
- minimizing risks in a contractual agreement for ad-hoc business relationships over public networks (such as the Internet), and
- reducing contract management costs.

E-CONTRACTING

Although there exist different descriptions for the e-contracting process (Milosevic & Bond, 1995; Goodchild, Herring, & Milosevic, 2000), the general e-contracting process includes two stages: contract establishment (contract formation) and contract enactment (contract performance or contract fulfillment) (Xu, 2004a; Angelov, 2005). E-contracting activities such as identifying, checking, and validating of contractual parties, negotiation, and validation contract are included in the stage of contract establishment. The contract enactment is further separated into two phases: performance and post-contractual activities. Monitoring of contract performance and compensation activities belongs to the contract performance phase, while contract enforcement may be involved in both the contract performance and post-contractual activities.

The research area of electronic contracting focuses on negotiation of the terms and conditions of the contract and the monitoring of contract performance (Lee, 1998). Contract negotiation is described as the process in which contracting parties come to a mutual agreement on the contract content. Contract negotiation can be performed with or without the help of a third party. There are three critical aspects for the negotiation of a contract (Burgwinkel, 2002a). First, the subject of the contract needs to be defined exactly. Second, the legal validity is formulated. Third, the price and conditions of each clause need to be negotiated in relation to the quality of deliverables and the quality of services, and in relation to the legal terms.

Contract monitoring is the process of observing the activities performed by the parties for the purpose of proactive imminent contract violations or detecting contract violations. To prevent undue costs, it is important for the contractual parties to monitor the performance of the other collaborating parties, especially if the transactions are business critical. The monitoring of contract performance can be split into two parts divided by the occurrence of an anomalous action (Xu & Jeusfeld, 2003; Xu, 2004a). The part preceding the occurrence of anomalous actions is called the proactive monitoring of contract performance. The part following it is called the reactive monitoring of contract performance. In the proactive monitoring stage, anomalous actions can be avoided and anticipated before contract violation occurrence — for example, by warning about impending deadline violations. In the reactive monitoring stage, anomalous actions can be detected, the partners who are responsible for the violations need to be identified, the relevant partner needs

to be compensated, and unsolvable disputes can be stored for future, human-involved resolution.

Contract enforcement is the process of persuading the noncompliant party to perform corrective actions. Contract enforcement can be performed in three ways: proactively (through constraints provided in the contract), reactively (via auxiliary corrective measures aiming at minimizing the deviations from the contract), and post-contractually (by constraining future activities of that company in this domain) (Angelov, 2005).

In order to format and fulfill a contract electronically, e-contracts must be managed. There are different views to look at contract management. From a contract platform view, contract management includes:

- A single repository for all contracts, related documents, and information to users;
- searching, reporting, and reusing capabilities to access all information in contracts and attachments;
- capabilities partners to track and monitor key performance indicators (KPIs) and performance over the contract execution, and use this information to target improvement actions and to determine preferred status, rankings, and so forth;
- maintenance of different versions of contracts, automatically reconcile changes to terms and clause language, and compare different versions;
- clause and template library to capture standard and alternate clauses along with guidelines; and
- alerts and reminders to inform contract partners of any upcoming dates, events, and milestones.

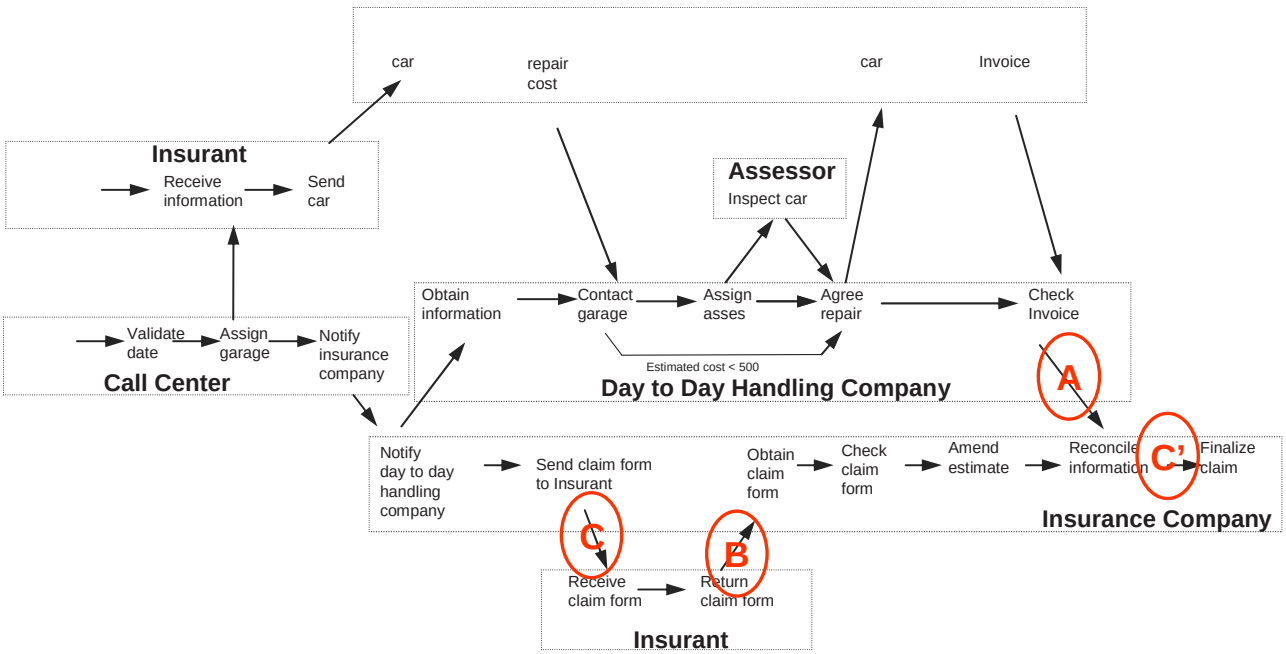
From one contractual party point of view, both its supplier and customer contracts need be managed. Moreover, the interrelation between these internal and external obligations, rights, and penalties must be synchronized.

MULTI-PARTY CONTRACT CASE AND ISSUES

We provide a multi-party contract case to explain the existing issues in multi-party contract execution. The case outlines the manner in which a car damage claim is handled by a car insurance company. The contract parties work together to provide a service level that facilitates efficient claim settlement. The parties involved are a call center, a day-to-day handling company, a group of garages, and an association of assessors. The call center is responsible for registering the insurant information, suggesting an appropriate garage (most times a close-by garage is assigned), and notifying the insurance company about the insurant's claim. The day-to-day handling company coordinates and manages the operation on a day-to-day level on behalf of the insurance company. A group of garages is assigned to assess car damages and to repair damaged cars for an insurant, who has bought car insurance from the car insurance company. The assessors conduct the physical inspections of damaged vehicles and agree upon repair figures with the garages.

The general processing of a claim in the car insurance case is as follows (see Figure 1): The insurant phones the call center using a toll-free phone number to give notification of

Figure 1. Process of car insurance case



a new claim. The call centre will register the information, suggest an appropriate garage, and notify the car insurance company, which will check whether the policy is valid and covers this claim. After the car insurance company receives this claim, the car insurance company sends the claim details to the day-to-day handling company. The car insurance company will send a letter to the insurant to ask for a completed claim form. The day-to-day handling company will agree upon repair costs if an assessor is not required for small damages; otherwise, an assessor will be assigned. The assessor will check the damaged vehicle and agree upon repair costs with the garage. After receiving an agreement for repairing the car from the day-to-day handling company, the garage will then commence repairs. After finishing repairs, the garage will issue an invoice to the day-to-day handling company, which will check the invoice against the original estimate. The day-to-day handling company returns all invoices to the car insurance company. After the car insurance company also receives the completed claim form from the insurant, the payment is processed. In the whole process, if the claim is found invalid, all contractual parties will be contacted and the process will be stopped.

There are many potential contract violations in this case; for example, after sending invoices to the day-to-day handling company, the garage does not get money back from the car insurance company. It could be caused, for example, by:

- The day-to-day handling company, because it does not forward the invoices to the car insurance company (Figure 1, with mark A);
- the insurant, because the insurant did not return the completed claim form to the car insurance company (Figure 1, with mark B);
- the car insurance company, because the car insurance company forgot to send the claim form to the insurant or simply because it did not pay the garage in time (Figure 1, with mark C and C'); or
- any combination from above.

Contract Specification

A multi-party contract is specified as a set of actions, a set of commitments, and a commitment graph of a contract. An action is specified as *action = (name, sender, receiver, deadline)*. For example, Action (A_agreeRepairCar,L,G",3) describes agreement of the day-to-day handling company for the garage to repair the car and that the day-to-day handling company has three days to respond.

A commitment is specified as *commitment = (name, sender, receiver, n, {(a_1, u_1), (a_2, u_2), . . . , (a_n, u_n) : ai — A, ui — U})*. An example of a commitment, *C_repairService* is specified as *(C_repairService,G, P,{(A_sendCar, tr), (A_estimateRepairCost, fi), (A_agreeRepairCar, tr), (A_repairCar, fi)})*. The garage will offer the repair service to the insurant. After the insurant sends his or her car to the garage (action *A_sendCar* has a trigger attribute), the garage estimates the repair costs (action *A_estimateRepairCost* has a finish attribute). After the garage receives an agreement from the day-to-day handling company about the repair costs (action *A_agreeRepairCar* has a trigger attribute), the garage repairs the car (action *A_repairCar* has a finish attribute).

Due to space limitation, readers are referred to Xu (2004a) and Xu, Jeusfeld, and Grefen (2005) for more details on multi-party contract specification and formal reasoning.

FUTURE TRENDS

Several technological, business, and legal hurdles must be overcome to establish e-contracting as a collaboration technology.

Multi-Party E-Contract vs. Multiple Bilateral E-Contracts

Little research has been done on multi-party contracts. Almost all research on multi-party contracts tries to break down a multi-party contract into a number of bilateral contracts (Xu et al., 2005). The semantics of a multi-party relationship are not always the same as multiple binary relationships. Only in some cases is it possible to see multi-party contracts as multiple binary contracts without losing valuable information. However, as more multi-party relations develop between companies, more contracts will be in force that would result in loss of information and increased complexity if relationships got hidden. The issues of how to represent or model multi-party e-contracts, how to identify the responsible partner(s) for a contract violation, and how to provide extra services for multi-party contracting (e.g., proactive monitoring functions) are critical in the area of e-contracting.

E-Contracting Challenges on Modeling and Representing Contractual Relationships

Existing contract models, such as the e-contracting logic model (Lee, 1998), aim to improve both expressiveness and inferential capabilities of the contracts. The model proposed in Weigand and Xu (2001) focuses on task allocations and process coordination. The proactive monitoring contract model (Xu, 2003, 2004a) and multi-party contract model (Xu, 2004b) provide contract models for different objects. These models represent trading contracts.

A model for e-contracts needs to present an integrated view on both the regular contracts and the technical agreement to be able for e-contracts to be fulfilled over networks. XML-based e-contracts have been investigated by several

research projects: SeCo (Greunz, Schopp, & Stanoevska-Slabeva, 2002), COSMOS (Merz, Griffel, Boger, Weinreich, & Lamersdorf, 1999), OCTANE (Kühne, Jungemann-Dorner, & Lam, 2001), and eLEGAL (Carter, White, Hassan, Shelbourn, & Baldwin, 2002). An XML-based electronic contracting editor is currently developed and can be deployed in different contract domains (Burgwinkel, 2002b). It enables all collaborating parties to create contracts from existing templates, to exchange and negotiate them, and finally to sign them over the Internet. The contract is a domain-specific XML document. The semantics of legal clauses can be expressed using clause types and structuring rules for legal documents. The e-contracts can be signed electronically using XML signatures. Ponton X/E, a visual XML editor, was developed by the German Ponton consulting company. The possibilities and limitations of XML-based e-contracting are discussed in Burgwinkel (2002b). Applying Semantic Web technology and business rules for e-contracting is studied in Grosof (2001).

There are some specific communities that publish standards and rules for the formation of contracts, such as the International Chamber of Commerce's model international sale contract (*www.modelcontracts.com*) and the Swiss IT association.

E-Contracting Challenges on Negotiation

Electronic environments have a great impact on contract negotiation. The benefits of e-negotiation could reduce the need for face-to-face meetings and traveling, improve the quality of agreements and mutual satisfaction, and reduce the damage caused by lasting and unresolved disputes. Contracts can be negotiated by exchange of messages over networks or using a negotiation platform. Studies on e-mail negotiation have shown that trading partners behave differently in electronic negotiation than they would in a face-to-face meeting (Shell, 2001). E-negotiation is a challenging task from the legal as well as from the business point of view. New laws and new legal frameworks for e-contracting such as building e-notaries are needed. User acceptance, security, and confidentiality are also important challenges of e-negotiation.

E-Contracting Challenges on Monitoring

A considerable amount of recent research and industrial application effort has concentrated on the provision of standards for automated establishment and subsequent implementation of electronic contracts. In general terms, an e-contract between multiple business partners contains some statements about their business relationship — in particular, on their physical and informational interactions over networks. One purpose of such an e-contract is to distinguish expected and acceptable behavior from behavior violating some clause in the multi-party contract. During the contract fulfillment, all necessary messages will be exchanged over networks. Those messages, events, or actions of contractual partners can be used to provide extra information for (proactive) monitoring services. Most of the current work focuses on the automation of contracting processes, rather than the development of services for contract fulfillment support, such as monitoring. Value-adding services — such as proactive monitoring and detecting the party (or parties) responsible for a contract violation in multi-party contract performance — are the new opportunities as well as the key issues for realizing a trustworthy e-commerce environment.

E-Contracting Challenges on Contract Management

When a company is engaged in many contractual business relationships, all using different e-commerce standards, management of the relations between the contractual parties and their e-commerce service providers becomes especially interesting. For each contractual party, it is important to monitor the performance of their providers, especially if the transaction is business critical, such as a payment service with a guaranteed payment.

Besides, from one contractual party's perspective, challenges for contract management are being able to achieve the maximum benefit of a contractual party, pre-calculating the cost of the contract violation, and trying to reduce the potential costs.

Currently, some contract management applications are offered by Dicarta and Oracle.

CONCLUSION

Electronic contracting affects the traditional roles and attitudes of sales and purchasing departments as well as lawyers. Growing IT support for contract management will provide assistance in handling the growing complexity of contractual relationships.

As an important component for trusted e-business in a global environment, e-contracting will help to reduce time to contract, improve the collaboration between the trading partners, and reduce minimize financial and legal risk. However, deployment of e-contracting has great challenges from the technical, business, and legal view. To handle the complexity of the setup and operation of contractual relations based on e-contracts, further research is needed.

REFERENCES

Angelov, S. (2005). *Foundations of B2B electronic contracting.* Eindhoven: Technische Universiteit Eindhoven.

ALDUS. (1992). *The ALDUS project: Artificial Legal Draftsman for Use in Sales.* Brussels: ESPRIT Commission.

Burgwinkel, D. (2002a). Decision support in electronic contract management. *Proceedings of the International Conference on Decision Marking and Decision Support in the Internet Age.*

Burgwinkel, D. (2002b). Improving contract management in e-business: Business and legal analysis of electronic contracts. In P. Schubert & U. Leimstoll (Eds.), *Proceedings of the 9th Research Symposium on Emerging Electronic Markets 2002* (pp. 95-102).

Carter, C., White, E., Hassan, T., Shelbourn, M., & Baldwin, A. (2002, November). Legal issues of collaborative electronic working in construction. *Proceedings of the Institution of Civil Engineers, Civil Engineering, Special Issue Two: Information Technology — The Key to Collaboration* (pp. 10-16).

Cheung, S.C., Chiu, D.K.W., & Till, S. (2002). A three-layer framework for cross-organizational e-contract enactment. *Proceedings of the Web Service, E-Business and Semantic Web Workshop with CAiSE'02.*

Cheung, S.C., Chiu, D.K.W., & Till, S. (2003). Data-driven methodology to extending workflows to e-services over the Internet. *Proceedings of the 36th Hawaii International Conference on System Sciences.*

Chiu, D.K.W., Cheung, S.C., Karlapalem, K., Li, Q., & Till, S. (2002). Workflow view driven cross-organizational interoperability in a Web-service environment. *Proceedings of the Web Service, E-Business and Semantic Web Workshop with CAiSE'02.*

Daskalopulu, A., & Sergot, M.J. (1997). The representation of legal contracts. *AI & Society, 11*(1/2), 6-17.

Daskalopulu, A. (1999). *Logic-based tools for the analysis and representation of legal contracts.* PhD Thesis, Imperial College London, UK.

Department of Justice Canada. (1995). *A survey of legal issues relating to the security of electronic information.* Technical Report, Department of Justice, Canada.

European Parliament and the Council. (2000). Directive 1999/93/EC of the European Parliament and of the Council of 13 December 1999 on a community framework for electronic signatures. *Official Journal of the European Communities, 43*(L13), 12-20.

Gardner, A. (1987). *An artificial intelligence approach to legal reasoning.* Cambridge, MA: MIT Press.

Goodchild, A., Herring, C., & Milosevic, Z. (2000, June 5-6). Business contracts for B2B. In H. Ludwig, Y. Hoffner, C. Bussler, & M. Bichler (Eds.), *Proceedings of the CAISE*00 Workshop on Infrastructure for Dynamic Business-to-Business Service Outsourcing,* Stockholm, Sweden.

Greunz, M., Schopp, B., & Stanoevska-Slabeva, K. (2000). Supporting market transactions through XML contracting container. *Proceedings of the 6th America Conference on Information Systems* (AMCIS 2000).

Grosof, B. (2001). Representing e-business rules for the Semantic Web: Situated courteous logic programs in RuleML. *Proceedings of the Workshop on Information Technologies and Systems* (WITS), New Orleans, LA.

Kafeza, E., Chiu, D.K.W., & Kafeza, I. (2001). View-based contracts in an e-service cross-organizational workflow environment. *Proceedings of the 2nd International Workshop on Technologies for E-Service.*

Koetsier, K., Grefen, P., & Vonk, J. (1999). *Contract model.* Technical Report Deliverable D4b, Cross-Organizational/ Work, Crossflow ESPRITE/28635.

Kühne, N., Jungemann-Dorner, M., & Lam, T. (2001). Open contracting transactions in the new economy — OCTANE project. *Proceedings of eBusiness and eWork 2001,* Amsterdam.

Lee, R. (1998). Towards open electronic contracting. *Electronic Markets, 8*(3), 3-8.

Merz, M., Griffel, F., Boger, M., Weinreich, H., & Lamersdorf, W. (1999). Electronic contracting im Internet. In R. Steinmetz (Ed.), *GI/ITG-Konferenz Kommunikation in Verteilten Systemen* (KIVS'99) (pp. 314-325). Berlin: Springer-Verlag.

Milosevic, Z., & Bond, A. (1995). Electronic commerce on the Internet: What is still missing? *Proceedings of the 5th Annual Conference of the Internet Society* (INET'95), Honolulu, HI.

Raman, R. (1996). *Cyber assisted business — EDI as the backbone of electronic commerce.* EDI-TIE B.V.

Sergor, M.J. (1991). The representing legislation as logic programs. In Hayes, Michie, & Richards (Eds.), *Knowledge-based systems and legal applications.* Academic Press.

Shell, R. (2001). Electronic bargaining: The perils of e-mail and the promise of computer-assisted negotiations. In S. Hoch & H.G. Kunreuther (Eds.), *Wharton on making decisions.* New York: John Wiley & Sons.

UNCID. (1987). *Uniform rules of conduct for interchange of trade data by teletransmission.* Author.

Weigand, H., & Xu, L. (2001). Contracts in e-commerce. *Proceedings of the 9th IFIP 2.6 Working Conference on Database Semantic Issues in E-Commerce Systems* (DS-9).

Xu, L., & Jeusfeld, M.A. (2003). Pro-active monitoring of electronic contracts. *Proceedings of the 15th Conference on Advanced Information Systems Engineering* (CAiSE 2003). Berlin: Springer-Verlag (LNCS 2681).

Xu, L. (2004a). *Monitoring multi-party contracts for e-business.* PhD Thesis, Tilburg University, The Netherlands.

Xu, L. (2004b). A multi-party contract model. *ACM SIGecom Exchanges, 5*(1), 13-23.

Xu, L., Jeusfeld, M., &.Grefen, P. (2005). Detection tests for identifying violators of multi-party contracts. *ACM SIGecom Exchanges, 5*(3), 19-28.

Yoshino, H. (1997). Legal expert project. *Journal of Advanced Computational Intelligence, 1*(2), 83-85.

Yoshino, H. (1998). Logical structure of contract law system for constructing a knowledge. Base of the United Nations convention on contracts for the international sale of goods. *Journal of Advanced Computational Intelligence, 2*(1), 2-11.

KEY TERMS

Business Collaboration: A set of activities or processes that lead to the accomplishment of an explicitly shared business goal by coordinated business parties; involves at least two autonomous business parties.

Contract: A legally binding exchange of promises or agreement between parties that the law will enforce.

Contract Enforcement: The process of persuading the noncompliant party to perform corrective actions.

Contract Monitoring: The process of observing the activities performed by the parties, knowing the state of contract execution and detecting contract violations.

Contract Violation: Refers to breaking or failing to comply with a term of the contract by a party.

E-Contract: A contractual agreement, represented as digital information and signed with digital signatures of the parties.

E-Contracting: The processes of formatting and negotiating of contracts electronically, and also monitoring the contract performance over networks.

E-Negotiation: Negotiation over networks where interested parties resolve disputes, agree upon courses of action, bargain for individual or collective advantages, and/or attempt to craft outcomes that serve their mutual interests.

eCRM Marketing Intelligence in a Manufacturing Environment

Aberdeen Leila Borders
Kennesaw State University, USA

Wesley J. Johnston
Georgia State University, USA

Brett W. Young
Georgia State University, USA

Johnathan Yehuda Morpurgo
University of New Orleans, USA

INTRODUCTION

This article examines the issue of electronic customer relationship management (eCRM) in a manufacturing context. ECRM has been described as the fusion of a process, a strategy, and technology to blend sales, marketing, and service information to identify, attract, and build partnerships with customers (Bettis-Outland & Johnston, 2003; Jaworski & Jocz, 2002). Although some customers still pay a premium for face-to-face or voice-to-voice interaction in today's high-tech world, through external (e.g., advertising) and internal (e.g., word-of-mouth) influence, the diffusion of the use of eCRM to build and sustain customer loyalty as a firm's strategy is on the rise. Manufacturers use the knowledge of their customers' needs and preferences to manage profitable customer interactions. This increased use of eCRM as a new manifestation (technological consolidation) of firmly established customer relationship management techniques has been shown to improve customer relationships and enhance customization (Kennedy, 2006).

BACKGROUND

Before the 1930s, the production era in which firms pushed to be the "provider" of products, whether customers needed, wanted, or could afford them was prevalent. From the 1930s to the 1960s, the selling era dictated the commerce arena in which salespersons were encouraged to make sells, regardless of costs. The onset of the 1960s to the 1990s portrayed the infancy of the marketing era in which the marketing concept (or satisfaction of the customer) laid the historical foundation of eCRM. From the 1990s, the partnering era has predominated and some functions previously performed by marketing have become absorbed into other functional areas such as manufacturing. Today, value lies within customer relationships that are satisfying to both the customer and the company. Applying principles to marketing that are enabled by technology with huge reservoirs of data, empowered by insight, and informed by history can provide the ROI that firms expect. Customer satisfaction is, in many cases, considered on par with revenue and profits as the performance metric by which companies are measured (Wu & Wu, 2005).

VALUE OF ONLINE SPACE

In reaching the maximization of value, businesses must define value, deliver it, and communicate it to customers. Value to the customer is his or her perception of the use of a product or service in relation to expectations. Managing the flow and not just the manufacturing process itself allows manufacturers to reduce non-value adding activities. Having the tools and enablers in place to integrate and automate processes allows all organizations to have marketing performance data (Freeland, 2003). Drivers of value differ in physical (off-line) and online places. Krishnamurthy (2003) viewed online operations as either profit centers (sources of income) or loss centers (offered as service to consumers). The 4 Ps (product, price, place, and promotion) primarily drive physical places. Online, the 6 Cs are the drivers—commerce, content, communication, connectivity, community, and computing (Krishnamurthy, 2003). *Commerce* describes the selling of products from the manufacturing, distribution, and retail firms to customers. Included in this category are the large businesses buying from other businesses in electronic marketplaces. *Content* is applicable to the news publishers (e.g., CNN, *New York Times, etc.)*, e-books, or companies using the Internet to educate their customers (e.g., Procter and Gamble at Crest.com). *Communication* involves Web-based seminars, Internet company meetings, and e-mail-based customer service. *Connectivity* refers to

the interconnections that employees and users have through the use of the Internet or other knowledge management tools. *Community* is portrayed through special user groups. *Computing* is manifested through tools such as mapping software, tracking software, and other portfolio management tools that empower customers. Manufacturers build online trust and commitment and potentially increase their value to customers by designing interactive Web sites (Merrilees, 2002). As businesses expand internationally, *culture* needs consideration as a seventh "C" as companies develop online experiences for their customers (Sigala, 2006). An organization's attitude is now its lifeblood. Integrating online operations with physical operations and leveraging company assets provides synergy between physical and online stores—a key to effective eCRM.

DIGITALITY

Digitality refers to the proportion of a company's business that is online (Krishnamurthy, 2003). The digitality of a business lies between zero and one. A business that is completely online with no physical components has a digitality of one. An example is one in which all employees telecommute, digital products such as software are sold, and customers communicate directly with the company's Web site. Alternatively, businesses with no representation in the online space have a digitality of zero. Most manufacturing firms would have digitality close to zero, except those that have incorporated online activities like Dell or Boeing.

MANUFACTURING PROCESSES

Manufacturing processes are the most likely places for sources of innovation and are probably 10 years ahead of service or customer-facing processes (Dixon & Duffy, 1990). Although speed of production still reigns in importance in manufacturing processes, the quality of the manufactured product, the flexibility to manufacture different types of products, reliable, predictable adherence to manufacturing timetables, and lowering of the cost and price of products must be matched against the marketing, engineering, and manufacturing capabilities for firms to become world-class competitors in the eCRM world (National Center for Manufacturing Services, 1990).

Typically, sales, manufacturing, and logistics are tightly woven. In coordinating manufacturing with sales, companies attempt to manufacture products and quantities to customer specifications and to minimize delays in delivery. This process has been described as the "lean production system" in automobile manufacturing (c.f. Davenport, 1993). In consumer foods processing, sales and manufacturing are driven to retail, wholesale, and distribution outlets by consumer demand. A common eCRM tool used is the salesperson's handheld computer that assists with the aggregation of store-level data, enabling linkages to materials and inventory systems, logistics, and sales departments.

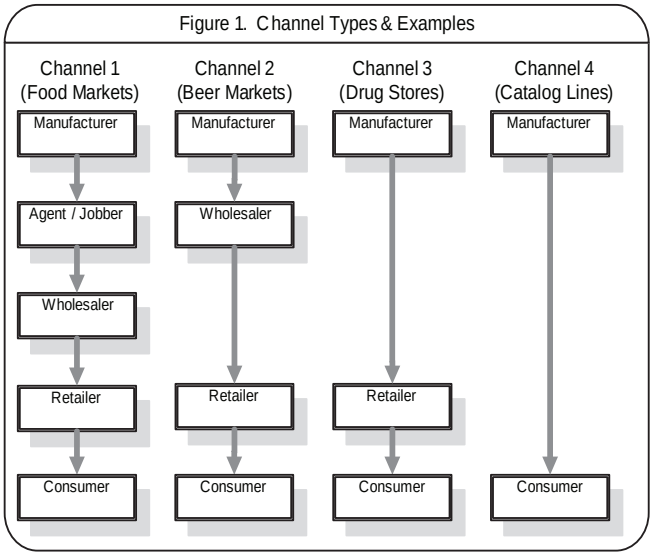
Equipment maintenance is another key area in which knowledge and information must be shared in a manufacturing environment to avoid downtime or scheduling and resource requirements conflicts. Radical changes, even lofty customer-initiated improvements, have to be phased in incrementally due to interfaces with legacy systems and logistical concerns in manufacturing arenas, regardless of the company's eagerness to be customer responsive. Many companies innovating with eCRM to coordinate the procurement and delivery of goods, on the outbound logistics side, find it advantageous to use just-in-time delivery or electronic data interchange to shorten the order-to-delivery cycle (Borders, Johnston, & Rigdon, 2001). Intense global competition and less loyal but more sophisticated customers are demanding a growing corporate emphasis on just-in-time marketing to make it faster, cheaper, and better to get products into the marketplace (Freeland, 2003). Manufacturers had to change from merely managing the direct labor content and variances in component costs to optimizing throughput and quality. ECRM data, along with the appropriate integrated technology speeds execution and improves information delivery to the point of need. From this workbench, the different stakeholders have information tailored to meet their needs. Finished goods customization, in some instances, is created only to fill customer orders and to ship goods to the customer, eliminating the need for warehousing. Customers with great bargaining power relative to their suppliers often initiate influence tactics that force suppliers to deliver on rapid, short terms (Borders, 2006).

DISINTERMEDIATION

In many cases, the Internet has become another sales channel with complementary features for bricks-and-mortar stores. The role of intermediaries is to facilitate buyer-seller transactions. Understanding disintermediation (the process of eliminating intermediates) requires familiarity with traditional business-to-business channels. Figure 1 illustrates business-to-business channels 1, 2, and 3.

In Figure 1, channel 1 is common in food markets, channel 2 in beer markets, and channel 3 in drugstore lines. Channel 4 is pervasive in the catalog line. Because the Internet makes it easy and relatively inexpensive to interact with customers directly, disintermediaries must prove their value. Channel 4 represents a business-to-consumer (B2C) disintermediated channel in which brokers or infomediaries prevail, as follows:

Figure 1. Channel types and examples



B2C intermediaries:

- Brokers (facilitate transactions, charge a fee on transactions)
- Infomediaries (serve fulfillment roles, act as filters between companies and consumers, sell market research reports, and help advertisers target their ads (e.g., ETrade), virtual malls that help consumers buy from a variety of stores, etc.)

Other disintermediated channels are consumed by metamediary figures to encourage procurement activity. These channel members offer consumers access to a variety of stores or provide transaction services and include search agents (e.g., My Simon, that helps consumers compare different stores); bounty brokers that charge a fee to locate a person, place, or idea, (e.g., Bounty Quest); or advertisement-based businesses (pop-ups, banners, other Internet linkages). In disintermediated channels, the major challenges that manufacturing firms face include building online traffic and sustaining customer loyalty and derived demand.

DERIVED DEMAND

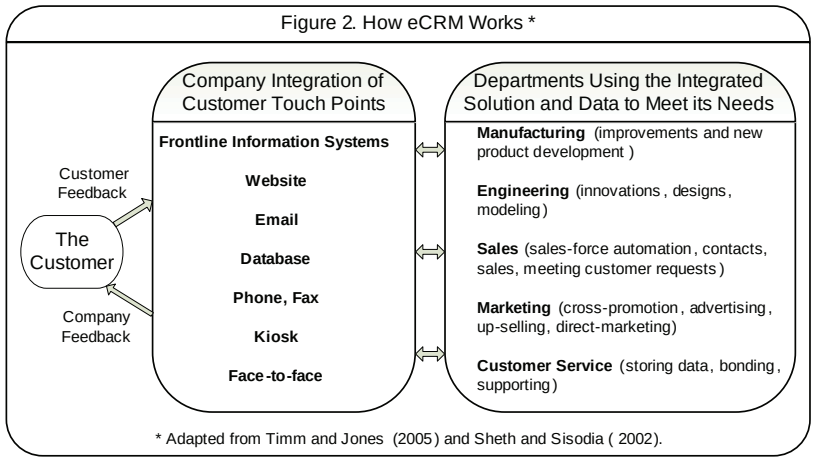
Once a manufacturer builds customer traffic, switching costs for the customer increase. The provision of detailed personal and payment information, and "mental transaction costs" relating to trust and privacy by using too many vendors, can be costs that customers will choose to avoid. In employing eCRM, firms should use the appropriate mix of low-tech, middle tech and high tech tools to acquire customers and satisfy them, and the appropriate metrics should be in place for low-profit or high-profit customers.

Krishnamurthy (2003) recommended that firms steer clear of too much eCRM customization and characterized online customers as simplifiers, surfers, bargainers, connectors, routiners, and sportsters. *Simplifiers* respond positively to easier and faster on- than off-line experiences. *Surfers* like to spend a lot of time online, thus companies must have a huge variety of products and constant updates. *Bargainers* are looking for the best price. *Connectors* like to relate to others. *Routiners* want content. *Sportsters* like sports and entertainment sites. Manufacturers target and build positive brand associations through opinion leaders that identify and attract customers most economically (Krishnamurthy, 2003). This subset of market segmentation, also described as "customer profiling," is conducted through data mining or optimization searches from consumer data as well as internal information systems (Padmanabhan & Tuzhilin, 2003). Once these customers begin using the products, they serve as role models and influencers of others in their communities and subcultures in the persuasion of purchasing products.

HOW ECRM WORKS

Successful e-commerce operations involve three elements (Figure 2)—Internet technology, a business model, and marketing (Krishnamurthy, 2003). ECRM describes an Internet-enabled system that leverages the power of the Web to deliver the best possible customer experience. The cost per customer runs from a low of $5 for individual customers to a high of $6,244 for business customers (Chatham,

Figure 2. How eCRM works



2001). Figure 2 shows how eCRM works by connecting the customer to different departments that need and make use of an integrated data solution.

Effective eCRM solutions create synergy among sales, marketing, and customer service activities. With multiple customer contact points, time frames, and systems, customer conversations can be disruptive and content lost, if they are not managed properly. eCRM consolidates customer information, such as personal data, preferences, inquiries, and order information and eliminates customers being given the run-around from department to department. eCRM is a supportive technology, whereby timely, targeted, personalized information and solutions are possible for customers.

ONLINE KNOWLEDGE

The quality of the online customer relationship is influenced by the *interactivity* between the customer and the firm, as mediated by the computer (Merrilees, 2002). Interactivity is described in terms of communication as follows:

1. The ability to address someone
2. The ability to gather information and responses from someone
3. The ability to re-address the individual with a unique response (Deighton, 1996)

Internationally, as well as in the United States, utilization of online knowledge (and often proactive database research) on the part of manufacturers has increased customer satisfaction and demand, as well as produced changes in organizational structures of firms (Stefanou, Saramaniotis, & Stafyla, 2003). It is prudent to use a database to target different consumers for different direct mail campaigns. However, junk mail (online or off-line) should not be directed at customers merely because the firm has customer contact information. On the other hand, some customers such as "surfers" welcome any contact with the firm. What might be right for a financial institution may be entirely wrong for a small appliances manufacturer. eCRM allows for the creation of the type of relationship that best suits and satisfies the customer's needs. When the information is "siloed" or stored in different places and is not easily accessible or compatible, there is a call for eCRM.

eCRM is embraced through automated quote generations where customers can provide specifications for products (e.g., computers, sales referral tracking, travel expense reporting, account management, and sales activity). By studying Web site traffic, companies are in better shape for advertisement placement, budget preparation, and demographic campaign management.

ECRM AS A SOLUTION

"The true business of every company is to make and keep customers" (Drucker, 1954). The Web has become an outstanding channel for delivery of customer assistance and effective customer support activities. In an eCRM environment, the Web allows customers to troubleshoot their own problems with a computerized assistant guide. In a manufacturing environment, when moving customer support to the Web, a firm might start with a static Web page, similar to an electronic version of a marketing brochure, or an information base of common questions and answers. From the company's perspective, these "low-assistance delayed communication" forms are the least expensive to implement. Self-serve personalized answers such as tracking status, personalized pages, and real-time data are also eCRM

features tailored to meet customers' needs. Self-service can be provided through quick, simple, and accurate global 24/7 representations. Moving past the self-serve common-answer category is a self-learning knowledge base. Manufacturers that encourage customers to use online knowledge to answer their own questions share in the wealth of information gained from service resolution, abandonment rates, average site connect time, and they are able to better serve their customers by improving content site and contact information (Timm & Jones, 2005).

When customers cannot find the answers they need from self-serve sites, they can then turn to e-mail in which customer service representatives respond. In comparing transaction costs per channel, Timm et al. (2005) contended that firms have come to realize that the average transaction cost for Web self-service and interactive voice response (IVR) self-service, ($0.24 and $0.45, respectively) compares more favorably than e-mail and phone customer care ($5.00 and $5.50, respectively).

THE BRICKS-AND-CLICKS BUSINESS MODEL IN THE ECRM MANUFACTURING ENVIRONMENT

In several manufacturing areas, it makes sense to combine online capabilities (clicks) with the advantages of traditional stores (bricks). The Internet allows an additional channel for customers to reach businesses and, in turn, through which businesses can reach their customers. For several product categories, such as furniture and apparel, individuals like to touch, feel, order, or try on the product before buying. Manufacturers set up "virtual dressing rooms" that provide customers the opportunity to try on a dress or shirt or other product before ordering it. Therefore, the bricks-and-clicks models can assimilate the experience of the product. The Internet could be used to locate the store or the variety of merchandise and to keep track of the status of the order.

Partnerships with brick-and-mortar stores can help alleviate delivery problems by providing another outlet for a customer to pick up a delivery that may have been placed online. Likewise, brick-and-mortar outlets that assort several items for shipping, and reduce the cost of shipment per item can aid in returning items to manufacturers. Functionality that makes it convenient for the customer is the motivation behind several eCRM efforts. Deshpandé and Farley (2002) agreed that the appropriate level of customer attention [market orientation] should be what the *customer* thinks it should be.

WHEN A MANUFACTURER MUST CHOOSE

When manufacturers diversify to create new revenue streams (by creating online tangential sales), they expose themselves to new levels of competition and vulnerability. Firms enhance the shopping experience by remembering relevant information about customers that can establish switching costs. They then use the positive brand-associated name to introduce other products through e-mail alerts, and they provide recommendations.

When implementing eCRM, economies of scale can be garnered when spreading costs across greater categories and leveraging the same brand name and customer base. In the electronics business, manufacturers have stringent requirements for retailers on how they will display and sell their products. Thus, after becoming an authorized dealer, one can get lower prices, money for cooperative advertising, and the right to sell warranties. Improved cross marketing, copromotion programs, and customer acquisition are provided when manufacturers such as metal companies, thermoplastics companies, and automobile and airplane manufacturers participate in *buy-centric markets,* whereby large, influential buyers find a place where small and fragmented sellers can sell their goods.

EFFECTIVENESS OF ECRM

The hallmark of eCRM is perceived personalization. It promotes cross selling of additional products, and up selling of more profitable products. Determining the needs and wants of customers and satisfying these needs better than the competition while keeping and strengthening the customer relationship bond is the business of eCRM (Kalwani & Narayandas, 1995). ECRM allows firms to solve the problem (competence), apologize for inconvenience (care), and offer a peace token such as waiving fees (comfort).

FUTURE TRENDS

An interesting aspect of some manufacturers' eCRM strategy is to implement programs similar to Amazon's associates or affiliates program. The thrust behind these programs is to have small sites generate traffic by having content with a link to Amazon or other manufacturers. The originating site receives a commission for referred purchases and a smaller commission for other purchases made by the customer. Amazon's brand enhances the smaller site's presence. Partnering with other businesses that sell products the manufacturer does not is a profitable revenue stream (Timm et al., 2005).

Manufacturing operations and other backroom processes are now absorbing previous frontline customer-facing functions. Large numbers of food products previously sold in bulk to retailers are now being shipped and sold by manufacturers in smaller packages or multiple size packages, as customers demand (Sheth & Sisodia, 2002). Although many agree that eCRM is a business strategy to select and manage the most valuable customer relationships, business processes are aligning with customer strategies to build customer loyalty and increase profits over time (Rigby, Reichheld, & Schefter, 2002). Quantitative approaches to eCRM market targeting and optimization are advancing rapidly. As new methods in this area are indexed, more powerful customer data systems are emerging (Padmanabhan et al., 2003).

Increasingly enabled by new technologies and applications, eCRM is the current generation of marketing intelligence tools for understanding how interactions and relationships with customers influence consumer behavior. While the hallmark of eCRM is perceived personalization, the next wave of marketing intelligence moves from customer "relationship" management to customer "experience" management (CEM). Negroponte (1995) predicted the onset of this capability when he asserted that everything is made to order, and information is extremely personalized. He purported that individualization is the extrapolation of narrowcasting—you go from a large to a small to a smaller group, ultimately to the individual. CEM is focused on defining and understanding the individual characteristics of a customer and having the ability to manufacture and sell that individual extraordinarily customized goods and services. By using real-time customer tracking enabled through the use of radio frequency devices (RFID), global positioning systems (GPS), video, portable shopping devices (PSDs), and analysis of clickstream data, CEM has become a reality (Burke, 2005).

CONCLUSION

Three areas of importance that cannot be neglected in eCRM manufacturing processes are the leveraging of database technology, the value of frontline information systems, and the importance of having the right employees (Sheth et al., 2002). Leveraging database technology leads to better targeting of and maintenance of customers. The frontline information systems (FISs) approach allows customers and frontline employees to be at the cutting edge of information technology that supports relationships and directly impacts customer satisfaction. Having employees that are responsive, courteous, professional, and competent improves a firm's productivity, profitability, and employee and customer satisfaction. Considering the operating costs of serving customers over time and the costs of customer turnover, manufacturers may find the value from customer loyalty, behavior, and satisfaction far exceeds the costs of implementing and maintaining appropriate eCRM systems (Jones & Sasser, 1995; Kundisch, Wolfersberger & Kloepfer, 2001).

REFERENCES

Bettis-Outland, H., & Johnston, W. J. (2003). Electronic customer relationship management (eCRM) in a business-to-business marketing setting. In T. Reponen (Ed.), *Information technology enabled global customer service*. Hershey, PA: Idea Group Publishing.

Borders, A. L. (2006). Customer-initiated influence tactics in sales and marketing activities. *Journal of Business and Industrial Marketing, 21*(6), 361-375.

Borders, A., Johnston, W., & Rigdon, E. (2001). Beyond the dyad: Electronic commerce and network perspectives. *Industrial Marketing Management, 30*(2), 199-206.

Burke, R. R. (2005). The third wave of marketing intelligence. In M. Krafft & M. Murali (Eds.), *Retailing in the 21st Century: Current and future trend*s (pp. 113-125). Heidelburg, Germany: Springer.

Chatham, B. (2001). CRM: At what cost? *The Forrester Report*. Retrieved from http://www.forrester.com/ER/Research Report/0,1338,11224,FF.html

Davenport, T. H. (1993). *Process innovation: Reengineering work through information technology.* Boston: Harvard Business School Press.

Deighton, J. (1996). The future of interactive marketing. *Harvard Business Review, 74*(6)*,* 150-151.

Deshpandé, R., & Farley, J. (2002). Looking at your world through your customer's eyes: Cross-national differences in buyer-seller alliances. *Journal of Relationship Marketing, 1*(3/4), 3-22.

Dixon, J. R., & Duffey, M. R. (1990). The neglect of engineering design. *California Management Review, 32*(2, Winter), 9-23.

Drucker, P. (1954). *The practice of management.* New York: Harper and Row.

Freeland, J. G. (2003). *The ultimate crm handbook.* New York: McGraw-Hill.

Jaworski, B., & Jocz, K. (2002, September/October). Rediscovering the customer. *Marketing Management, 11*(5), 22-27.

Jones, T. O., & Sasser, W. E., Jr. (1995). Why satisfied customers defect. *Harvard Business Review, 77*(6), 88-99.

Kalwani, M., & Narayandas, N. (1995). Long-term manufacturer-supplier relationships: Do they pay off for supplier firms? *Journal of Marketing, 59*(1), 1-16.

Kennedy, A. (2006). Electronic customer relationship management (eCRM): Opportunities and challenges in a digital world. *Irish Marketing Review, 18*(1/2), 58-68.

Krishnamurthy, S. (2003*). E-commerce management: Text and cases*. Cincinnati, OH: Thomson Southwestern.

Kundisch, D., Wolfersberger, P., & Kloepfer, E. (2001). Enabling customer relationship management: Multi-channel content model and management for financial eServices. *Journal of Marketing Management, 3*(11), 91-104.

Merrilees, B. (2002). Interactivity design as the key to managing customer relations in e-commerce. *Journal of Relationship Marketing, 1*(3/4), 111-125.

National Center for Manufacturing Services. (1990). *Competing in world-class manufacturing: America's twenty-first century challenge.* Homewood, IL: Richard D. Irwin.

Negroponte, N. (1995). *Being digital*. New York: Knopf.

Padmanabhan, B., & Tuzhilin, A. (2003). On the use of optimization for data mining: Theoretical interactions and eCRM opportunities. *Management Science, 49*(10), 1327-1343.

Rigby, D., Reichheld, F., & Schefter, P. (2002). Avoid the four perils of CRM. *Harvard Business Review, 80*(2), 101-108.

Sheth, J. N., & Sisodia, R. S. (2002). Marketing productivity issues and analysis. *Journal of Business Research, 55*(5), 349-362.

Sigala, M. (2006). Culture: The software of e-customer relationship management. *Journal of Marketing Communications, 12*(3), 203-223.

Stefanou, C. J., Saramaniotis, C., & Stafyla, A. (2003). CRM and customer-centric knowledge: An empirical research. *Business Process Management Journal, 9*(5), 617-634.

Timm, P. R., & Jones, C. G. (2005). *Technology and customer service: Profitable relationship building.* Upper Saddle River, NJ: Pearson-Prentice Hall.

Wu, I. L., & Wu, K. W. (2005). A hybrid technology acceptance approach for exploring e-CRM adoption in organizations. *Behaviour & Information Technology, 24*(4), 303-316.

KEY TERMS

Commerce: The selling of products from the manufacturing, distribution, and retail firms to customers.

Connectivity: The interconnections that employees and users have through the use of the Internet or other knowledge management tools.

Customer Profiling: Selecting customers you want to find, going after them, and keeping them.

Digitality: The proportion of a company's business that is online.

Disintermediation: The process of eliminating intermediaries in the channels of distribution.

eCRM: The fusion of a process, a strategy, and technology to blend sales, marketing, and service information to identify, attract, and build partnerships with customers.

Interaction Management: An entire system that monitors customer communications at every possible contact point, regardless of source (i.e., Web, telephone, fax, e-mail, kiosks, or in person).

Knowledge Base: An online repository of information that represents the collective wisdom regarding a product or service.

Education for Library and Information Science Professionals

Vicki L. Gregory
University of South Florida, USA

INTRODUCTION

Libraries and information centers today are very different places from those that existed at the beginning of the 20th century, and very different as well from the libraries of only 25 years ago. Education for library and information science has striven to keep pace with all the myriads of changes. Within the last 100 years, fortunately and necessarily in order to retain its relevance, professional library education and practice has evolved from the centrality of teaching and writing the "library hand" to providing modern curricula such as services for distance learners and Web-based instruction using course management systems such as Blackboard, WebCT, and so forth. Along the way, the library profession has often been first not only to accept but also to adopt and apply the technological innovations now common to modern civilization. One of the newest trends involves the "I-Schools" where information is taught as the overarching discipline with librarianship just one of the programs in a larger college offering programs in informatics, information science, information architecture, knowledge management, and so forth. Throughout, library and information science educators have paved the way to the acceptance of innovation in libraries and information centers by instructing students to use and apply new technologies.

BACKGROUND

The revolutionary changes over the past 25 years in the educational curriculum for schools of library and information science, which are necessitated by the exponential expansion of computer-based technologies, require an almost constant and continuous reexamination of the skills and expertise needed to be acquired by the next generation of librarians. Although much has changed in libraries, the core of who we are and what we are truly remains the same. Librarianship is and will continue to be a profession devoted to bringing users and information together, as effectively and efficiently as possible. To meet that ideal, librarians have used technology to enhance and create services. In addition, it is important to meet emerging educational needs of our increasingly multicultural and diverse society. Librarians have recognized that changing expectations and lean budgets require organizations to call upon the talents of everyone (Butcher, 1999). And, librarians have become more engaged in teaching and research in order to serve the needs of users better (Bahr & Zemon, 2000).

THE I-SCHOOLS

The I-Schools are a group that has been coming together over the last few years of schools/colleges that are taking a broad approach to the study of information. The deans and faculty of those schools held their first conference in September of 2005 to explore the similarities and differences among the schools present and basis upon which to build a foundation for a new type of College. The I-Schools include a number of schools with traditional LIS programs such as Syracuse, Pittsburgh, Rutgers, and others, plus a number of other programs such as the College of Information Sciences and Technology at Pennsylvania State University. John King describes the I-Schools thusly:

The I-School movement is made up of novel academic programs that embrace new intellectual and professional challenges in a world awash in information. I-Schools move beyond traditional programs, while building on the intellectual and institutional legacies of those programs. I-Schools straddle the academy's ancient engagement with information and the contemporary challenges of ubiquitous information affecting all aspects of society. (King, 2006)

The I-Schools bring together a variety of disciplines that are scattered among different colleges and departments into one college of information. Some areas of what is traditionally in computer science departments are also a part of this movement (Carroll et al., 2006). The proximity and interaction among these programs by being brought together in a new college should lead to an enriched research and teaching environment.

IMPORTANCE OF PEOPLE AND PEOPLE SKILLS

Computer technologies and communication systems have had an undeniable impact on society as a whole and our

profession, but it is also critical to remember the importance of the individual and of the need for interpersonal skills in our profession, which at its heart remains basically a "people profession." We harness technology for a reason—to promote learning and the dissemination of information—and we do not simply revere technology for its own sake. With the aid of computer specialists, we could design the best information system imaginable, but unless it operates in a manner that is accessible to people, nobody will use it. The ability of librarians (whether through collecting, organizing, or retrieving information) to act as intermediaries between users and the world's information resources will, in my opinion, never become outdated (Gilbert, 1998). In sum, the rapid changes in all types of libraries and the burgeoning of new technologies for librarians to learn, while increasing the amount of information that students need to have under their "academic belts" if they are to enter successfully into a library career, nevertheless remain rooted in the need to carry out the traditional librarian roles—though hopefully faster, cheaper, smarter, and more effectively.

PREPARING STUDENTS IN TRADITIONAL AREAS OF LIBRARY RESPONSIBILITIES

The traditional heart and sole of a library is and remains, of course, its collections—from the time of the great Alexandrine library of the Classical era, libraries have been, in essence, civilization's repositories of learning, and hence the materials through which learning is transmitted down the generations. Current students preparing for the future (and indeed the present) electronic library cannot be permitted to overlook the continued, lasting importance of print publications in the library's carrying out of its role, but they by necessity must be equipped to deal with the rapidly expanding world of digital materials. Thus, collection development courses must reflect an appropriately balanced approach, emphasizing the latest technology not as an end in itself, but rather as simply another tool to use in addressing the problems arising in acquiring adequate resources for a library collection in whatever format is most appropriate for the particular library and the "task at hand" (Thornton, 2000).

As librarians and information professionals go about the process of acquiring electronic information resources in carrying out their collection development role, they must also continue to recognize and care about the important questions that have always concerned libraries respecting questions of future accessibility and preservation of library resources. Electronic materials with their typical provision to libraries only through a licensing regime rather than through outright purchase present altogether different problems for the library than do print materials. Collection development and preservation must remain an important part of the library school curriculum no matter how dominated the library may become with electronic materials (Kenney et al., 2002).

In most conceptions of the libraries of the future, reference librarians may expect to continue to play many of the same reference roles that they have traditionally performed in interacting with their library's users. Reference librarians will continue to serve in an intermediary role to assist users in finding needed information and providing important "value-added" services through the production of instructional materials and guides to information resources. However, many of these functions, out of necessity, will be performed in media other than those that have been traditionally utilized. Collaboration and instruction may be expected to take place in a Web-based "chat" environment or by e-mail rather than through a face-to-face meeting over the reference desk (Abels, 1996; Domas White, 2001).

Reference librarians of the future must therefore acquire teaching skills as well as informational skills. They will need to be able to teach information literacy skills as students discover that just finding some online information on a topic and pushing the "print" or "download" button is not enough. In the electronic information world, librarians must be prepared to evaluate resources in a somewhat more in-depth way than was necessary when they could often depend upon refereed print journals for the majority of their information (Grassian & Kaplowitz, 2001).

In addition to all the vagaries involved with the classification and cataloging of traditional print materials, today technical services librarians will have to be prepared to cope with all the exponential varieties and forms that electronic resources may take. Technical services professionals are increasingly dealing with so many different formats and kinds of materials that may defy classification and are often not traditionally cataloged; other approaches, such as indexing and abstracting techniques and the development of in-house library-constructed databases, as well as Webliographies, may be undertaken as methods of organizing the access and retrieval process.

Future graduates planning a career in the technical services areas should place a much greater focus than is presently typically allowed for in most library school curriculums on the technological aspects of information provision. Concurrently, library and information science schools need to take steps to provide for the programs and/or the courses that will include building student skills in document creation for the digital library environment. Unfortunately, all this cannot be allowed to serve as a replacement for the traditional knowledge and skills involved in cataloging and classification. As a minimum, students will need to gain a hands-on knowledge of the architecture of the infrastructure and databases behind a digital library. This means that LIS schools must develop additional specific courses, rather than trying to make room in the already overstuffed basic

"organization of knowledge" classes that most schools currently offer (Vellucci, 1997).

DISTANCE EDUCATION

In the foreseeable future, it is probable that more and more instruction will be provided in a distance mode utilizing Web delivery, videoconferencing, and other technological means of providing instruction. A burden on many LIS faculty members at present is how to adapt a course, originally designed for a face-to-face classroom encounter, to a Web-based encounter. Although the goals, objectives, and major assignments for a class might remain the same, the overall means of delivery puts more pressure on faculty members to devise new ways of delivering material (Gregory, 2003). Both virtual and print reserve materials may become problematic as distance from the home site increases, although the increasing amounts of electronic materials held by academic libraries have been making access to library materials easier for students at a distance.

Faculty members have become increasingly innovative in finding ways to bring some feeling of community to students involved in distance education. Examples include PowerPoint slides with a voice-over by the instructor, text and voice chat over the Web, and the construction of virtual community sites, often in the same software environment as the classes (Gregory, 2006; Nicholson, 2005).

Compounding the traditional instructional component, there is the additional element of computer support on a 24 hour, seven days a week basis (Young, 2002). Increasingly, when something goes wrong with the computer on a student's end, the faculty member is expected to be able to do computer troubleshooting over the telephone or by e-mail. It is common for programs and universities to provide technical support, but even so the faculty member usually gets caught up in the technical support problems, obviously much more so than when the class is taught in the traditional manner (Carey & Gregory, 2002; Newton, 2003). Of course, when the academic computing staff person or the faculty member is unavailable, the next major organization on the campus that fields these questions is most likely the library. Librarians must be able to deal with technical, computing, or network issues and attempt to aid the beleaguered student (or faculty member). So although these issues primarily affect the teaching of library and information studies classes, they also have a major impact on the services demanded of the library (Barron, 2003).

FUTURE TRENDS

For the past 10 years or so, a rift has occurred between the information science and librarianship sides of the field. The I-School movement is one manifestation of that divide, but the concern of library practitioners is the other (Stofle & Leeder, 2005). Michael Gorman, president of the American Library Association for 2005-2006, took on the side of the practitioners in a series of meetings and presentations. His major points were that LIS education had become too concentrated on information and technology and not enough attention was being spent on the traditional core of librarianship (Gorman, 2005). While others have disputed Gorman's position (Dillon and Norris 2005), this debate seems to be a point of contention that will continue for some time.

Is "information" itself a discipline/field of study, or is it just a part of librarianship and other traditional academic disciplines? Does any one field of study own "information?" Only time will tell the outcome.

CONCLUSION

The rapidly changing requirements in the educational curriculum of schools of library and information science resulting from the exponential expansion of computer-based technologies naturally result in a re-examination of the knowledge and skills that need to be acquired by the next wave of library and information professionals. Skills in the use of new technologies are not only important in professional work, but also in the education process itself, as more and more LIS courses are being offered via the Web with faculty and students utilizing course management software.

REFERENCES

Abels, E. G. (1996). The e-mail reference interview. *RQ, 35*(3), 345-358.

Bahr, A. H., & Zemon, M. (2000). Collaborative authorship in the journal literature: Perspectives for academic librarians who wish to publish. *College and Research Libraries, 61*(5), 410-419.

Barron, D. D. (Ed.). (2003). *Benchmarks in distance education: The LIS experience*. Westport, CT: Libraries Unlimited.

Butcher, K. (1999). Reflections on academic librarianship. *Journal of Academic Librarianship, 25*(5), 350-353.

Carey, J. O., & Gregory, V. L. (2002). Students' perceptions of academic motivation, interactive participation, and selected pedagogical and structural factors in Web-based distance education. *Journal of Education for Library and Information Science, 43*(1), 6-15.

Carroll, J. M., Dourish, P., Friedman, B., Kurosu, M., Olson, G. M., & Sutcliffe, A. (2006). Institutionalizing HCI: What

do I-schools offer? In Conference on Human Factors in Computing Systems (pp. 17-20). New York: ACM Press.

Dillon, A., & Norris, A. (2005). Crying wolf: An examination and reconsideration of the perception of crisis in LIS Education. *Journal of Education for Library and Information Sciences, 46*(4), 280-298.

Domas White, M. (2001). Digital reference services: Framework for analysis and evaluation. *Library & Information Science Research, 23*(3), 211-231.

Gilbert, B. (1998). The more we change, the more we stay the same: Some common errors concerning libraries, computers, and the Information Age. In M. T. Wolf, P. Ensor, & M. A. Thomas (Eds.), *Information imagineering: Meeting at the interface* (pp. 219-227). Chicago: ALA.

Gorman, M. (2005). Why library education matters. *American Libraries, 36*(7), 5.

Grassian, E. S., & Kaplowitz, J. R. (2001). *Information literacy instruction: Theory and practice.* New York: Neal-Schuman.

Gregory, V. L. (2003). Student perceptions of the effectiveness of Web-based distance education. *New Library World, 104*(10), 426-433.

Gregory, V. L. (2006). Virtual communities: A way to connect students in an internship program. *International Journal of Learning, 12*. Retrieved June 8, 2006, from http://www.Learning-Journal.com

Kenney, A. R., McGovern, N. Y., Botticelli, P., Entlich, R., Logaoze, C., & Payette, S. (2002). Preservation risk management for Web resources. *D-Lib Magazine, 8*(1). Retrieved December 10, 2003, from http://www.dlib.org/dlib/january02/kenney/01kenney.html

King, J. L. (2006). Identity in the I-School movement. *Bulletin of the American Society for Information Science and Technology, 32*(4). Retrieved June 8, 2006, from http://www.asis.org/Bulletin/index.html

Newton, R. (2003). Staff attitudes to the development and delivery of e-learning. *New Library World, 104*(1193), 312-425.

Nicholson, S. (2005). A framework for technology selection in a Web-based distance education environment: Supporting community-building through richer interaction opportunities. *Journal of Education for Library and Information Sciences, 46*(3), 217-233.

Stofle, C. J., & Leeder, K. (2005). Practitioners and library education: A crisis of understanding. *Journal of Education for Library and Information Sciences, 46*(4), 312-319.

Thornton, G. A. (2000). Impact of electronic resources on collection development, the roles of librarians, and library consortia. *Library Trends, 48*(4), 842-856.

Vellucci, S. L. (1997). Cataloging across the curriculum: A syndetic structure for teaching cataloging. *Cataloging and Classification Quarterly, 24*(1/2), 35-39.

Young, J. R. (2002). The 24 hour professor: Online teaching redefines faculty members' schedules, duties, and relationships with students. *The Chronicle of Higher Education, 38*(May 31). Retrieved June 7, 2006, from http://chronicle.com/weekly/v48/i38/38a03101.htm

KEY TERMS

Collection Development: The portion of collection management activities that has primarily to do with selection decisions.

Collection Management: All the activities involved in information gathering, communication, coordination, policy formulation, evaluation, and planning that result in decisions about the acquisition, retention, and provision of access to information sources in support of the needs of a specific library community.

Course Management Systems (CMS): Computer software system that provides a course shell with a number of integrated tools that may include chat software, threaded discussion board, online grade books, online testing, and other classroom functions.

Digital Libraries: Organized collections of digital information.

Distance Education: A planned teaching and learning experience that may use a wide spectrum of technologies to reach learners at a site other than that of the campus or institution delivering the course.

Information Literacy: An integrated set of skills and the knowledge of information tools and resources that allow a person to recognize an information need and locate, evaluate, and use information effectively.

I-Schools: A group of schools/colleges that focus on the discipline of information with programs in information architecture, information technology, information science, knowledge management, librarianship, and other disciplines associated with the production, organization, and use of information.

Videoconferencing: Conducting a conference between two or more participants at different geographical sites by using computer networks to transmit audio and video data.

The Effect of Sound Relationships on SLA's

AC Leonard
University of Pretoria, South Africa

INTRODUCTION

End users have expectations regarding services and *support,* and the quality thereof, provided by the supplier. They compare their expectations to the received service to assess the service quality (Coye, 2004).

In order to ensure that the service supplied by the service provider meets the expectations of end users, a successful *service level agreement (SLA)* is required. Quality *SLA's* clearly define, amongst many other elements, the commitments and responsibilities of the IT service provider and end users within the *service delivery* processes (Larson, 1998). One method of measuring the success of SLA's is by using service metrics with regard to the availability, reliability, serviceability, response, and user satisfaction of the SLA (Larson, 1998). Therefore, the success of the SLA depends on a clear, common understanding of the services and service quality between the service provider and end users. Furthermore *commitment*, trust, and cooperation between all parties is necessary to achieve success with SLA's (Hiles, 1994). However, in this paper it is argued that all these soft issues can only form a basis when sound relationships are established and maintained between the IT service provider and end users (Leonard, 2002).

This paper aims to determine how the establishment of a sound IT-*end user* relationship can add value to the SLA for both the IT service provider and the end users, and increase the success of SLA's.

PROBLEM BACKGROUND AND RESEARCH APPROACH

According to Parish (1997), the benefit of an SLA is that the identification of accountability in the service delivery process can be determined more easily, even when more than one service provider is involved in the process. Therefore, *commitment* from service providers and end users in the service delivery process can be determined. Secondly, the SLA will promote a focus on the quality of service required by *end users* to support their business needs. Thirdly, it will enable the service provider to clearly identify the key service needs of the end user organization to ensure that the business operations of the end user organization are operating at optimal level. Finally, a successful SLA will enable the service provider and the end users to implement correct service metrics to monitor the quality of service, which will enable them to perceive any service problems in advance and implement contingent plans. According to Lehr and McKnight (2002) the service metrics can be described as the commitments from the IT service provider to guarantee the quality of service delivered to end users at the agreed service level stated in the SLA.

According to Hiles (1994), the reasons for unsuccessful SLA's are insufficient service definition, poor measurement of service quality, inconveniently large documentation of SLAs, a lack of mutual understanding and, most commonly, a lack of commitment from the *end users*.

According to Leonard (2002), *commitment* and mutual understanding form, amongst other *soft issues*, the basis of any sound relationship between end users and IT professionals. Leonard (2002) states that an *IT-end user relationship* consists of physical and abstract elements which impact on the soundness of the IT-end user relationship. It is, therefore, argued that the elements of the abstract dimension of a sound IT-end user relationship contribute to the successfulness of the SLA. The most important elements that play a role in this regard, are a supportive culture, commitment, and cooperation (Leonard 2002). On the contrary, it is argued that a lack of commitment from *end users* and IT service providers will result in a poorly drafted SLA due to an unclear service definition and a lack of proper service quality metrics. Pratt (2003) has indicated that poorly drafted service elements in the SLA will create a poor picture of the services provided by the service provider from the point of view of the end users, resulting in an unsuccessful SLA.

An interpretive research approach was followed, taking into consideration the important principles for interpretive research as stated by Klein & Myers (1995) and Sahay et al. (1994). Apart from doing a theoretical study of the field, employees of about 15 different companies were approached to respond to a number of questions. The feedback on these questions, and the theory of IT-end user relationships and SLA's serve as basis for the arguments followed in this paper.

The next section starts with a discussion on service level agreements, followed by a brief discussion of the theory of IT-end user relationships. The paper concludes with a proposed conceptual framework showing how sound relationships enhance the worthiness of SLA's.

THE ESTABLISHMENT OF SLA'S

According to Larson (1998), a *service level agreement (SLA)* is a formal contract between the IT service provider and the business unit within an organization. The SLA provides a common understanding of the quality of service that the IT service provider will provide. It also helps create reasonable expectations amongst end users at a specific business unit. Apart from defining the standard of service quality and setting customer expectations, the SLA outlines the role of the *end users* and the role of the service provider. Therefore, an SLA will enable the end users to be fully aware of the *service delivery* capabilities and limitations of the service provider, while the service provider will understand the expectation and IT service needs of the end users. This common understanding about the service delivery between the IT service provider and end users is an important component for establishing a successful IT service provider-end user relationship as indicated by Smith (1996).

Once there are common interests and mutual understandings between the end users and IT service provider, it will enable the IT service provider to include the services and quality of service needed by end users in the draft SLA. According to Pratt (2003), the SLA will only be of value if the IT service provider has a clear understanding of the end user organization's core business operations and business needs.

Tonks and Flanagan (1994) state that the *SLA* should focus on adding value to the *service delivery* process for the service provider and the end users. Therefore, it should not be regarded as a contract to penalize the service provider should he/she fail to deliver the service. Instead, it should be used as a condition to set the desired service required from the service provider and how the quality of service will be measured and reported to the end users.

DEFINING AN IT-END USER RELATIONSHIP

According to Leonard (2002) an *IT-end user relationship* consists of two dimensions, namely, a *physical dimension* and an *abstract dimension*. The physical dimension describes those elements required to enable contact between IT and its end users, whereas the abstract dimension describes the *soft issues* (such as trust, commitment) of a relationship. These two dimensions enable one to fully describe the holistic nature of such a relationship and encapsulate the important elements of a *support*-oriented organization, namely mutuality, belonging, and connection, as mentioned by Pheysey (1993) in her book *Organizational Cultures*.

HOW THE ELEMENTS OF THE PHYSICAL AND ABSTRACT ELEMENTS IMPACT ON THE PROCESSES OF SLA CONSTRUCTION AND USE

In this section a brief description is given of the elements of both the *physical* and *abstract* dimensions which have the most significant impact on the worthiness of SLA's. In this regard, research has indicated that *people*, *technology*, and *procedures* are the most important elements from the physical dimension, whilst *knowledge base*, *supportive culture*, *commitment*, *cooperation*, and *holistic nature* are the most important elements from the abstract dimension.

Elements of the Physical Dimension

According to Leonard (2002) a sound *IT-end user relationship* consists of all the responsible people involved at a given time. "Responsibilities are negotiated and shared …" (Referring to the work of Dahlbom and Mathiassen (1993)). In terms of the *people* element, it follows that sound relationships will have a positive impact on the development of an *SLA*. Furthermore, the parties involved communicate through communication technology (e.g., video conferencing technology, helpdesk) as indicated by Leonard (2002). This communication technology also enables the IT service provider and end users to communicate with each other during the development and implementation of the *SLA*, which should have a positive impact on the development process of an SLA. Finally, communication technology will improve the response performance of the SLA as the IT service provider can attend to requests from end users and reply back to them with lesser turn-around time (Larson 1998).

The development of an *SLA* will be affected by existing policies and procedures. On the other hand, a new SLA can also introduce new policies and standards into the end user company. Therefore, any procedures (which form part of the physical dimension of an IT-end user relationship) will ensure that the implementation of the SLA does not compromise the business operations of the end user organization.

Elements of the Abstract Dimension

The impact of a sound *IT-end user relationship* will enable end users to have a highly affective *commitment* with the service provider. End users who have such commitment are less likely to switch to a new service provider (Mattila, 2004). Therefore, there is a large possibility that the end users will continue using the service of the current service provider or develop new SLA's with them instead of searching for a new provider in case of poor services or support.

According to Leonard (2002) a sound *IT end-user relationship* has a knowledge base. This enables end users to understand the limitations/trade-offs of their desired service in terms of availability, reliability, pricing, and serviceability. Furthermore, it enables end users to participate more effectively during the SLA development process.

A supportive culture exists during sound *IT-end user relationships* (Leonard, 2002), which establishes a mutual understanding and support between the IT service provider and end users. This supportive culture allows the IT service provider and *end users* to work together as a team striving to gain mutual benefits. According to Smith (1996), the success of managing services is that the end user must have a clear understanding of the IT *service* provider's capabilities. The IT service provider must also understand the expectations of end users, and both parties should appreciate the limitation of their partner organization. Therefore, the supportive culture will enable the IT service provider to clearly understand the expectations and needs of end users. On the other hand, end users will understand the capabilities of the IT service provider and services offered by them. This will result in a well-defined *service* definition between the end users and the IT service provider.

According to Leonard (2002), a sound relationship also implies that cooperative behavior exists between all role players. This means that an IT *service* provider and its end users will work together towards gaining positive future results. The *service* defined in an SLA is affected by the changes in service needs of the end user organization. *Service* needs of the *end user* organization are affected by the changes in the markets. A successful SLA should therefore be modifiable, to accommodate the service needs of end users (Larson, 1998). Therefore, if the IT service provider and the end users have a sound IT-end user relationship, they can cooperate to modify the SLA and re-define the service deliverables and service definitions in the SLA to accommodate the changes in service needs of the end users.

An IT-end user relationship has an holistic nature as indicated by Leonard (2002). The holistic nature emphasizes the cooperative behavior and support culture in the IT-end user relationship, which also enhances the successfulness of the SLA.

A sound IT-end user relationship has *commitment* from both the end users and IT service provider (Leonard, 2002), which enables both parties to work towards the same goal to gain mutual benefit and appreciation of each other's capabilities and limitations. According to Hiles (1994), a lack of commitment is one of the primary reasons for the failure of SLA's. Therefore, commitment from both the end user and IT service provider has a positive impact on the success of SLA development.

ANALYSIS OF FEEDBACK RECEIVED FROM PRACTITIONERS

The purpose of this part is to interpret/analyze some of the feedback received from different role players (IT professionals and end users) in the private sector using SLA's. The responses given by the participants are given literally. In other words, no changes have been made to the grammar of the feedback. Only spelling mistakes were corrected. Furthermore, only those parts of responses that could be regarded as of interest to the phenomenon under investigation, are cited. The responses are also not given in a specific order:

- Sound relationships between the two groups bring understanding of the problems, challenges and needs of both sides. It therefore encourages the service providers to give their best and the users to gauge their expectations, thus reducing on conflict and misunderstanding.
- I believe sound relationships with the users will result in the IT department knowing the user's needs better, which will help in developing SLA's with the user's needs at the center of their development.
- If there is not a sound relationship, the two/more parties involved tend to do just enough. I speak from personal experience that the relationship influences the effectiveness of SLA's in business. If the parties involved do not have a sound relationship, even well developed SLAs are not as effective as they should be.
- Sound relations ensure that SLA communicate effectively to both IT and end user departments. This also ensures that both parties expectations are aligned.
- Unlike traditional purchasing contracts, SLAs require all parties to have a higher understanding of the customer's business requirements. This understanding can only be obtained through a sound relationship. In addition, an SLA is actually a contractual embodiment of desired or existing business relationships. The purpose of an SLA is to ensure business continuity, necessitating a deep understanding of the existing business processes. An SLA is an organic document in that it changes with changing business requirements.
- Sound relationships and SLA are closely related in a sense that without sound relationships between IT department and business units, then it will be difficult to have well defined SLAs and without SLA, then the relationship might get sour due to different expectations of the outcome of services that are provided by IT departments for other business units and the users of the service will not have sufficient information necessary to understand and use the services. This

might also lead to poor relationships between different departments.

- I do believe that a link exists between a sound relationship and an SLA. The reason for this is that in order to know the what, where, and when of a *service,* a sound relationship should be built between a service buyer and a service provider. Getting all partners involved through sound relationships of trust, commitment, and the like, will improve the quality of the service level agreement and ensure that all its interdependencies are not left out.

INTERPRETATION/ANALYSIS OF THE ABOVEMENTIONED CITATIONS

The abovementioned citations not only indicate the importance of sound "working relationships" between IT professionals and *end users*, but also regard it as an indispensable piece of "equipment" or "tool" which is necessary to *support* all participants in addressing the needs of end users or business units. The citations also refer to important aspects regarding the establishment and maintenance of sound relationships. First of all, the culture gap is addressed, and, as such, the importance of a common knowledge base, which will surely help all participants of a given IT-end user relationship to "realize" what is expected of them. In terms of how sound relationships would impact on the construction and operations of an SLA, everyone refers to the fact that if sound relationships do not exist or are not maintained, trust between all role players would be in jeopardy. In such a case SLA's would not be honored.

From the above analysis, it is clear that the elements of the physical and abstract dimensions have a direct impact on the quality and operational success of SLA's. This impact is illustrated in the following conceptual framework.

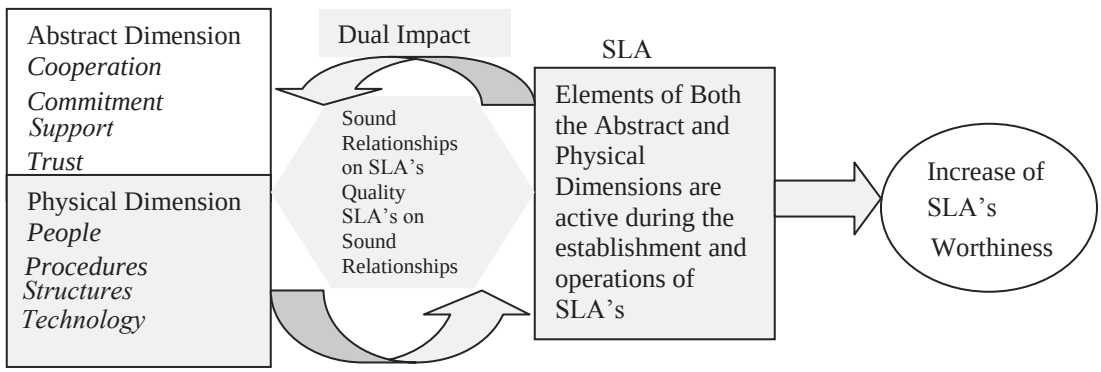
FUTURE TRENDS

It is clear that in the future much more research needs to go into the management of relationships and service level agreements to get a better understanding of the nature of the impact relationships have, not only on the operational value of SLA's but also on the behavior of people. In this regard the role and impact of organizational culture cannot be ignored. Leonard (2002) lists several examples of cultural differences between end users and IT professionals that lead to poor relationships which, on their part, impact negatively on the construction and operation of service level agreements. Therefore, in terms of the future of IT service management it is clear that cultural differences should also become part of the picture.

CONCLUDING SUMMARY

In this paper the impact of sound *IT-end user relationships* on the construction and operations of *SLA*'s is discussed. To illustrate this, the most important elements of the *physical* and *abstract* dimensions that play an important role in this regard are discussed. It was argued that each of these elements exist during the processes necessary to negotiate and construct SLA's of a high quality.

Figure 1. The dual impact of sound IT-end user relationships on SLA operations



Based on the definition and theory of sound *IT-end user relationships,* one can argue that the elements of the physical and abstract dimensions enhance the communication, effective commitment, and trust between the IT service provider and end users. On the other hand, this ensures that all role players strive to gain mutual benefit for all parties involved.

A conceptual framework is also proposed illustrating how sound *IT-end user relationships* enhance the worthiness of *SLA's*. This enhancement takes place in terms of the SLA development process as well as all operations concerning the SLA. Furthermore, this enhancement should have a dual effect in the sense that it not only supports the SLA design and operations process, but all the elements of the physical and abstract dimensions should, in turn, be enhanced.

REFERENCES

Coye, R. W. (2004). Managing customer expectations in the service encounter. *International Journal of Service Management,* vol.15 no.1, pp. 54-71.

Dahlbom B, & Mathiassen L. (1993). *Computers In Context: The Philosophy and Practice of Systems Design*. Blackwell Publishers, Cambridge UK.

Gale Group. (2001). *Management Summary, Service Level Agreements*, p1.1, September 2001.

Gale Group. (2001. *Objectives, benefits and costs of SLA, Service Level Agreements*, p3.1, September 2001.

Hamaker, S. & Hutton, A. (2003). *Principles of Governance, Information Systems Control Journal*, vol. 3, ISACA.

Hiles, A. N. (1994). *Service Level Agreements: Panacea or Pain?* The TQM Magazine, vol. 6 no. 2, pp.14-16.

Jain, G., Singh, D., and Verma, S. (2002). Service level agreements in IP networks, *Information Management & Computer Security*, vol. 10 no. 4, pp. 171-177.

Larson, K. D. (1998). The role of service level agreements in IT service delivery. *Information Management & Computer Security*, vol. 6 no.3, pp. 128-132.

Lehr, W. and McKnight, L. W. (2002). Show me the money: contracts and agents in service level agreement markets. *Info - The journal of policy, regulation and strategy for telecommunications*, vol.4 no.1, pp.24-36.

Leonard, A. C. (2002). A conceptual framework for managing relationships between all participants during IT service and support activities. *SA Journal of Industrial Engineering*, vol. 13 no. 2, pp.81-96.

Mattila, A. S. (2004). The impact of service failures on customer loyalty. *International Journal of Service Management*, vol.15 no.2, pp. 134-149.

Parish, R. J. (1997). Service level agreements as a contributor to TQM goals. *Logistics Information Management*, vol.10 no. 6, pp. 284-288.

Pratt, K. T. (2003). Introducing a service level culture. *Facilities*, vol. 21 no.11, pp.253-259.

Smith, R. (1996). Business continuity planning and service level agreements. *Information Management & Computer Security*, vol.3 no.3, pp.17-19.

Tonks, P. and Flanagan, H. (1994). Positioning the Human Resource Business Using Service Level Agreements. *Health Manpower Management*, vol. 20 no. 1, pp.13-17.

KEY TERMS

Abstract Dimension: The abstract dimension describes the soft issues of such a relationship.

IT/End-User Relationship: A relationship between IT and the end user consists of two dimensions, namely, a physical dimension and an abstract dimension. The physical dimension describes those elements that are necessary in order to enable contact between IT and its end users; whereas, the abstract dimension describes the soft issues of a relationship.

Physical Dimension: The physical dimension describes those elements that are necessary in order to enable contact between an IT professional and its end users.

SLA: Service level agreement.

Soft Issues*:* Those elements that describes how a person behaves under specific circumstances. Examples of such soft issues are trust, commitment, support, cooperation, and so forth.

Effective Leadership of Virtual Teams

David Tuffley
Griffith University, Australia

INTRODUCTION

Geographically dispersed project teams collaborating in virtual environments face a range of challenges in the successful completion of IT development projects. This is particularly the case when the project teams are nonhomogenous, comprising multidisciplinary members with a range of skills, professional orientations and cultural backgrounds. Of interest to the global enterprise are those leadership mechanisms and attributes that may serve to optimize team functioning.

With an increasing portion of the estimated US$600,000,000,000 (Cusamano, 2004) global software industry being performed by virtual teams, and with the mechanics and dynamics of virtual team operations being a relatively new area of study, the significance of the problem can be firmly established.

Virtual teams, and the leadership thereof, is therefore a significant aspect of the global software development industry. Yet as Cusamano (2004) asserts, it is the *business* itself (and the processes therein), not the technology that determines the success or failure of the organizations that produce the software.

BACKGROUND

The past 50 years have seen a remarkable proliferation of what might be termed the *global enterprise*, organizations that transcend national borders and extend across the globe. Commercial organizations in industrialised economies have increasingly established international networks of subsidiaries and affiliates with which to pursue a global agenda, taking advantage of economies of scale and effort. This trend inevitably leads to the advent of distributed work environments and the consequent formation of multidisciplinary virtual teams (teams that operate across different time and physical space).

Collaborative technologies (messaging and discussion forums, audio and video conferencing, as well as knowledge portals, business directories, Web cams) are assumed to facilitate team functioning in virtual environments, yet it is nonetheless important that we examine the broad issue of team work processes and optimising. The building of functional social networks in virtual environments can be a difficult task, particularly on an international scale. The respective cultures of the team members are a significant factor. Other factors include physical environments, information technology support, communication policies and procedures, as well as leadership.

VIRTUAL TEAMS

Distinguishing Virtual Teams From Conventional Teams

Bell and Kozlowski (2002), quoting a widely cited earlier study by Townsend, DeMarie, and Hendrickson (1998) define virtual teams as:

Groups of geographically and/or organizationally dispersed co-workers that are assembled using a combination of telecommunications and information technologies to accomplish and organizational task.

Virtual teams can therefore be distinguished from conventional teams in two fundamental ways; their *spatial proximity* and the *communications technologies* employed.

When contrasting Townsend et al.'s (1998) definition of virtual teams with that of conventional teams (Humphrey, 2000), we see that the Humphrey definition offers a good general purpose view of what a team is:

A team consists of:

1. *At least two people, who*
2. *Are working toward a common goal/objective/mission, where*
3. *Each person has been assigned specific roles or functions to perform, and where*
4. *Completion of the mission requires some form of dependency among group members.*

Operational Definition of Virtual Team

It might be reasonable, therefore, to combine these definitions:

A virtual team consists of:

1. At least two mutually interdependent people, who
2. Are geographically dispersed, and who

3. Are working toward a common goal/objective/mission, where
4. Each person is assigned specific roles or functions to perform, and where
5. Communication is facilitated by a combination of telecommunications and information technologies to work toward the completion of the project/mission.

Leadership

Until "kings were philosophers or philosophers were kings" there will be injustice in the world. (Plato)

The classical period of ancient Greece produced concepts and modalities that have become the foundation of western civilization. In relation to leadership studies the philosopher Plato (427-347 BC) in his renowned dialogue *The Republic* outlined certain enduring leadership principles that Western administrative thinking has based itself upon (Takala, 1998). Plato developed systematic administrative thinking for the efficient running of the city-sate (polis) which over time allowed the evolution of democracy. Plato described in detail the appropriate relationship between the state and individual citizens. This relationship was so close that it was not possible to think of a citizen living outside of his state (Takala, 1998). The purpose of this state is to educate people to become "good." The state is like the human body in which parts complement each other and act harmoniously. In terms of organizational theory, Plato would be regarded as a premodern functionalist.

Distinguishing Leaders and Managers

The terms leader and manager are sometimes used interchangeably, adding to the ambiguity surrounding the study of leadership. Yet studies of administrative science usually find the terms differentiated. How is this done?

According to Takala (1998) what they have in common is the ability to get things done. We then distinguish them by managers being a kind of instructor who puts pieces together and manages the "things." A manager is primarily concerned with making an organization function by evolving routines that serve the ongoing and sometimes changing purposes of the organization.

Leadership Qualities of Great Groups

Bennis and Beiderman (1997) discuss at length the leadership qualities required in Great Groups. They observe that group leaders can vary widely. There can be facilitators, doers, contrarians. Leaders are catalytic completers; taking on roles that nobody else plays and that are needed for the group to achieve its goal. They have an intuitive understanding of the "chemistry" of the group and the dynamics of the work process. Furthermore, they encourage dissent in the establishment and maintenance of a shared vision. They can distinguish between healthy, creative dissent and self-serving obstructionism.

Bennis and Beiderman (1997) identify four behavioral traits of effective group leaders:

1. **Provide direction and meaning:** Group members are kept up-to-date on what is important and why their work makes a difference.
2. **Generate and sustain trust:** The group has trust in itself and its leadership. This allows members to accept dissent and tolerate the turbulence of the group process.
3. **Display a bias toward action, risk taking, and curiosity:** A sense of urgency and willingness to risk failure to achieve results.
4. **Are purveyors of hope:**. Find tangible and symbolic ways to demonstrate that the group can overcome difficulties.

Personality Traits and Competencies of Effective Leaders

Bennis (1994) in a wide-ranging study determined that effective leaders display four distinct personality traits, and five specific competencies, the sum of which tends to manifest in strong and effective leadership. Personality traits include guiding vision, passion, integrity, and daring (Bennis, 1994). The competencies are technical competence, interpersonal skills, conceptual skills, judgment, and character (Bennis, 1999a). No pairing order is implied by this table, as it is a listing only.

Bennis (1999a) asserts that it is character that is the essential element determining a leader's effectiveness, saying "leaders rarely fail because of technical incompetence" but more so for lack of character. (Bennis, 1999b)

Table 1. Personality traits and competencies of effective leader (Bennis, 1994, 1999a)

Personality Traits	Competencies
Guiding vision	Technical competence
Passion	Interpersonal skills
Integrity	Conceptual skills
Daring	Judgment
	Character

Strong character can manifest in positive and negative ways, as the lessons of history inform us. Strong character makes for a strong leader, but character can be strong and negative/destructive. Offerman, Hanges, and Day (2001) relates that a person's character will be determined by the sum total of his or her values. Offerman et al. (2001) identified the source of an employee's dissatisfaction and disillusionment as the particular values held by leaders and the actions that these values motivate.

Underlying Qualities of Effective Leaders

The qualities that inspire people to persevere in the face of great difficulty, and that engender trust and a sense of worth among team members are not always readily identifiable. They are qualities that are not easily detected, but that are found in the best of leaders.

Champy (2003) identifies these underlying qualities as:

- **Empathy:** Macaluso (2003) suggests that empathy is the secret weapon of corporate success, an indispensable quality for any successful leader. Empathy is described as the ability to see the world through another's eyes, to experience it as they would, or "To walk a mile in another's shoes."
- **Personal responsibility:** Effective leaders accept that the circumstances in which they find themselves are largely the result of their own previous actions. They do not blame others (Macaluso, 2003).
- **Openness to discovering truth:** Effective leaders fearlessly search for truth, knowing that sometimes the truth will not be pleasant to face (Macaluso, 2003).

Transformational vs. Transactional

Zhang, Fjermestad and Tremaine (2005) identify two parallel dimensions of leadership: *transformational vs. transactional,* and *participative vs. directive*. These have been derived from a body of foundational work in the area of leadership styles in a virtual team context.

On the Transformational/Transactional dimension we see the *Transformational* element as comprising four behavioural components (Bass, 1985; Bass, Avolio, & Goodheim, 1987; Lowe, Kroeck, & Sivasubramaniam, 1996):

- **Charisma or idealized influence:** The leader engenders in the members a sense of pride, respect, faith and respect, together with a sense of purpose/mission.
- **Individualized consideration:** The leader manifests a deep concern for the well-being of the members, and provides mentoring.
- **Intellectual stimulation:** The leader stimulates members to think in original ways, emphasising the triumph of reason over irrationality, and challenging established ways of thinking.
- **Inspirational motivation:** The leader creates high standards, communicating high expectations.

Continuing with the Transformational/Transactional dimension we see the *Transactional* element as comprising three behavioural elements (Bass, 1985; Bass et al., 1987; Lowe et al., 1996):

- **Contingent reward:** The leader rewards performance on the basis of it having fulfilled prescribed obligations.
- **Management-by exception:** The leader ensures the standards are met.
- **Management-by-exception (passive):** The leader adopts a laissez-faire attitude until noncompliance of standards has occurred.

Leadership of Virtual Teams

The concept and practice of distributed work is not new, and in fact enjoys a long and colourful history, as discussed by O'Leary, Orlikowski and Yates (2002) in their extended case study of the Hudson Bay Company from 1670 to 1826. Yet it has been the advent and subsequent advances in communications technology that has been a critical enabler of the development of this organisational form and practice (Ahuja, Carley, & Galletta, 1997).

It has been observed (Cascio & Shurygailo, 2003) that distributed teams (or virtual teams as they might be called) face particular problems in relation to leadership. Organizational and management research has focused intensively on the issue of leadership, as seen in a previous section, yet there is relatively little research done thus far on the emerging challenge of leadership in virtual teams (Cascio & Shurygailo, 2003).

Leadership Challenges for Virtual Teams

An in-depth study into the typology of virtual teams, and the implications therein for effective leadership, is found in Bell and Kozlowski's (2002) work. This work proposes 11 distinct challenges for the leadership of virtual teams.

Bell and Kozlowski (2002) identify four broad categories of leadership challenge in virtual teams; (a) temporal distribution, (b) boundary spanning, (c) life cycle and (d) member roles. The categories are described by Bell and Kozlowski (2002) as shown in Table 2. Table 3 elaborates the 11 propositions relating to leadership challenges in vir-

Table 2. Bell and Kozlowski's (2002) four categories of leadership challenge in virtual teams

Category	Description
Temporal Distribution	Virtual teams operating in real-time use rich, synchronous communication media and temporal entrainment to effect performance management.
Boundary spanning	Individualized consideration for and performance management of team members who span different functional areas, organizations or cultures.
Member Roles	Members holding multiple roles within and across virtual teams.
Lifecycle	Performance management effectiveness is improved when team membership is stable and ongoing, allowing time for relationships to be established and developed.

Table 3. Bell and Kozlowski's (2002) 11 propositions of leadership challenge in virtual teams

Category	Leadership challenge
Temporal Distribution	Distributed virtual teams are more likely to use synchronous, richly textured communications media.
Temporal Distribution	Effective virtual team leaders are more likely to develop substitutes for face-to-face contact.
Temporal Distribution	The more complex the virtual project, the more likely it will be performed in real time, not distributed time.
Boundary spanning	The more complex the task, the more likely the team will be distributed.
Boundary spanning	Virtual team boundaries will be less permeable in complex projects where established operating procedures and stable relationships are needed.
Boundary spanning	Effective team leaders are likely to create proactive performance management functions, AND be good at using technology to provide members with team development experiences.
Boundary spanning	Effective leaders are good at evaluating the effectiveness of self regulation mechanisms, AND that these developmental functions will be more difficult to implement across multiple boundaries.
Boundary spanning	More complex projects are likely to require stable team membership.
Member Roles	More complex projects are likely to require clearly defined singular roles for members.
Member Roles	Multiple roles and boundaries are likely to make performance management more difficult, AND effective leaders are more likely to clearly specify roles and role interrelationships, particularly in more complex projects.
Lifecycle	Discrete life cycle of virtual projects will be experienced integrated difficulty with establishing performance regulating functions, AND leaders will therefore focus on the most critical issue of establishing effective working relationships with members.

Table 4. Characteristics of empirical studies of leadership in virtual teams (Adapted from Misiolek (2006) and Dube & Pare (2004)

Authors	Main research method	Theoretical perspective
Balthazard et al. (2004)	Lab experiment	Shared leadership; leadership style; transformational and transactional leadership
Cogburn et al. (2002)	Quasi-experimental field study	Behavioural; two-factor theory
Connaughton & Daly (2004)	Interviews	Implicitly behavioural
Hoyt & Blascovich (2003)	Lab experiment	Transformational and transactional leadership
Authors	**Main research method**	**Theoretical perspective**
Kayworth & Leidner (2002)	Field experiment	Behavioural; behavioural complexity theory; trust
Pauleen (2003)	Case study	General theoretical discussion
Pauleen (2004)	Interviews & two 10-week action learning sessions + grounded theory analysis	General theoretical discussion with focus on relationship-building and trust
Piccoli & Ives (2000); Piccoli et al. (2004)	Field experiment	Team control structure; self-managing teams
Sarker et al. (2002); Nicholson et al. (2002)	Field experiment	Emergent leadership; propose new theoretical model incorporating culture, communication, technical ability, trust, gender, performance, and client location
Sudweeks & Simoff (2005)	2 case studies	Behavioural; implied two-factor theory; emergent leadership
Tyran et al. (2003)	Field experiment	Behavioural; two-factor theory; emergent leadership
Weisband (2002)	Field experiment	Behavioural; two-factor theory; group awareness
Yoo & Alavi (2004)	Field experiment + grounded theory analysis of transcripts of team interactions	Behavioural; two-factor theory; emergent leadership

tual teams outlined by Bell and Kozlowski (2002). They are grouped into the four categories discussed above.

Summary of Empirical Studies of Leadership in Virtual Teams

Dube and Pare (2004) surveyed virtual team characteristics published in empirical studies. Misiolek (2006) used this as a basis for further investigation into leadership aspects of virtual teams. The combination of these two sources, plus additional investigation, results in the table below. It summarizes the broad sweep of theoretical perspectives developed in these empirical studies.

FUTURE TRENDS AND CONCLUSION

Effective leadership of virtual teams in the world of tomorrow will be facilitated by increasingly rich communications media that allows collaboration between individuals as if they were in the same physical location. Broadband communications technologies such as fiber optics promises to deliver

the capability to create virtual environments rich enough with subtle detail to make this possible. The commercial potential for organizations to develop such technologies is high, ensuring a vibrant and competitive market for such products. This is potentially a major benefit.

The qualities of a good leader remain constant, whether they operate in the same space or in virtual space. The challenge for the leaders of tomorrow will be to negotiate successfully with the emerging collaborative technology to make the best use of it.

Such qualities have been displayed by notable leaders throughout history, are being displayed by effective leaders today, and can reasonably be expected to be displayed by the leaders of tomorrow, extending into the far distant future. These qualities are functions of human nature that have co-evolved during millions of years of human evolution. Indeed, the human capacity to collaborate to solve problems is a defining aspect of the human species, and is responsible in large part for our phenomenal success as a species. Implicit to this ability to collaborate is the need for someone to facilitate that collaboration (a leader).

Human kind stands today on the threshold of a major step in an evolutionary history stretching back 5 million years. We are making the transition from operating in a physical environment only, to operating in a hybrid physical-virtual environment, with the trend toward increasingly virtual environments. Humans evolved the ability to adapt themselves to a wide range of physical environments, practically the whole world from the Equator to the Poles, and beyond into space. This distinguishes humans from all other species. Having exhausted the physical environment in this unceasing expansion, human kind is now developing the technology to create virtual worlds in which to live and work. Leaders in these virtual worlds will be those that combine traditional leadership qualities with the ability to make these virtual worlds seem real.

REFERENCES

Ahuja, M. K., Carley, K., & Galletta, D. F. (1997). *Individual performance in distributed design groups: An empirical study.* In Paper presented at the SIGCPR Conference, San Francisco, (p. 165).

Balthazard, P., Waldman, D., Howell, J., & Atwater, L. (2004). Shared leadership and group interaction styles in problem-solving virtual teams. In *Proceedings of the 37th Hawaii International Conference on System Sciences*.

Bass, B. (1985). *Leadership and performance beyond expectations*. New York: The Free Press.

Bass, B., Avolio, B., & Goodheim, L. (1987). Biography and the assessment of transformational leadership at the world class level. *Journal of Management, 13*, 7-19.

Bell, B.S., & Kozlowski, S.W. (2002). A typology of virtual teams: Implications for effective leadership. *Group and Organisational Management, 27*(1), 14-19.

Bennis, W. (1994). *On becoming a leader, what leaders read 1* (p. 1). Perseus Publishing.

Bennis, W. (1999a). The leadership advantage. *Leader to Leader*, 12, 12.

Bennis, W. (1999b). Five competencies of new leaders. *Executive Excellence*, 16(7), 4-5.

Bennis, W., & Beiderman, P. (1997). *Organizing genius: The secrets of creative collaboration*. Addison-Wesley.

Cascio, W., & Shurygailo, S. (2003). E-leadership and virtual teams. *Organizational Dynamics, 31*, 362-376.

Champy, J. (2003). The hidden qualities of great leaders. *Fast Company Magazine, 76,* 2.

Cogburn, D.L., Zhang, L., & Khothule, M. (2002). Going global, locally: The socio-technical influences on performance in distributed collaborative learning teams. In *Proceedings of the 2002 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement Through Technology,* (pp. 52-64). Port Elizabeth, South Africa: South African Institute for Computer Scientists and Information Technologists.

Connaughton, S.L., & Daly, J.A. (2004). Leading from afar: Strategies for effectively leading virtual teams. In S.H. Godar & S.P. Ferris (Eds.). *Virtual and collaborative teams: Process, technologies, and practice* (pp. 49-75). Hershey, PA: Idea Group.

Cusamano, M.A. (2004). *The business of software: What every manager, programmer, and entrepreneur must know to thrive and survive in good times and bad.* New York: Free Press.

Dube, L., & Pare, G. (2004). The multifaceted nature of virtual teams. In D.J. Pauleen (Ed.), *Virtual teams: Projects, protocols, and practices* (pp. 1-39). Hershey, PA: Idea Group.

Hoyt, C.L., & Blascovich, J. (2003). Transformational and transactional leadership in virtual and physical environments. *Small Group Research, 34*(6), 678-715.

Humphrey, W.S. (2000). *Introduction to the team software process* (p. 19). Reading, MA: Addison-Wesley.

Kayworth, T., & Leidner, D. (2002, Winter). Leadership effectiveness in global virtual teams. *Journal of Management Information Systems, 18*, 7-40.

Lowe, K., Kroeck K., & Sivasubramaniam, N. (1996). Effectiveness correlates of transformational and transactional leadership: A meta-analytic review of the MLQ literature. *Leadership Quarterly, 7*, 385-425.

Macaluso, J. (2003). *Harnessing the power of emotional intelligent leadership. The CEO Refresher* (p. 2).

Misiolek, N. (2006). *Patterns of emergent leadership in distributed teams*. Unpublished doctoral dissertation, School of Information Studies, Syracuse University, Syracuse, NY.

Nicholson, D., Sarker, S., & Sarker, S. (2002). Ingredients of effective leadership in information systems development project teams: An exploratory study. In *Proceedings of the Eighth Americas Conference on Information Systems*. Retrieved December 8, 2007, from http://aisel.isworld.org/pdf.asp?Vpath=/amcis/2002/&PDFPath=022206.pdf

Offerman, L.R., Hanges, P.J., & Day, D.V. (2001). Leaders, followers, and values: Progress and prospects for theory and research. *The Leadership Quarterly, 12*, 129-131.

O'Leary, M., Orlikowski, W. J., & Yates, J. (2002). Distributed work over the centuries: Trust and control in the Hudson's Bay Company, 1670-1826. In P. Hinds & S. Kiesler (Eds.), *Distributed work* (pp. 27-54). Cambridge, MA: MIT Press.

Pauleen, D.J. (2003). Leadership in a global virtual teams: An action learning approach. *Leadership & Organization Development Journal, 24*(3), 153-162.

Pauleen, D.J. (2004). An inductively derived model of leader-initiated relationship building with virtual team members. *Journal of Management Information Systems, 20*(3), 227-256.

Piccoli, G., & Ives, B. (2000). Virtual teams: Managerial behavior control's impact on team effectiveness. In *Proceedings of the Twenty first International Conference on Information Systems,* (pp. 575-580). Atlanta: Association for Information Systems.

Piccoli, G., Powell, A., & Ives, B. (2004). Virtual teams: Team control structure, work processes, and team effectiveness. *Information Technology & People, 17*(4), 359-379.

Sarker, S., Grewal, S., & Sarker, S. (2002). Emergence of leaders in virtual teams. In *Proceedings of the 35th Hawaii International Conference on System Sciences.*

Sudweeks, F., & Simoff, S.J. (2005). Leading conversations: Communication behaviours of emergent leaders in virtual teams. In *Proceedings of the 38th Hawaii International Conference on System Sciences.*

Takala, T. (1998). Plato on leadership. *Journal of Business Ethics, 17*, 785-798.

Townsend, A.M, DeMarie, S.M., & Hendrickson, A.R. (1998, August). Virtual teams and the workplace of the future. *Academy of Management Executive, 12*, 17-29.

Tyran, K.L., Tyran, C.K., & Shepherd, M. (2003). Exploring emergent leadership in virtual teams. In C.B. Gibson & S.G. Cohen (Eds.), *Virtual teams that work: Creating conditions for virtual team effectiveness* (pp. 183-195). San Francisco: Jossey-Bass.

Weisband, S. (2002). Maintaining awareness in distributed team collaboration: Implications for leadership and performance. In P.J. Hinds & S. Kiesler (Eds.), *Distributed work* (pp. 311-333). Cambridge, MA: MIT Press.

Yoo, Y., & Alavi, M. (2004). Emergent leaders in virtual teams: What do emergent leaders do? *Information and Organization, 14*(1), 27-58.

Zhang, S., Fjermestad, J., & Tremaine, M. (2005). Leadership styles in virtual team context: Limitations, solutions and propositions. In *Proceedings of the 38th Hawaii International Conference on System Sciences*.

KEY TERMS

Charisma: The ability to develop or inspire in others an ideological commitment to a particular point of view.

Collaborative Technologies: Technology that allows people to interact effectively in virtual environments. Includes messaging and discussion forums, audio and video conferencing, knowledge portals, business directories, and Web cams.

Directive Leadership: Providing and seeking compliance with directions for accomplishing a problem solving task.

Empathy: The ability to see the world through another's eyes, to experience it as they would. An essential leadership quality.

Global Enterprise: An emerging phenomena facilitated by communications technology in which multinational organizations extend their operations globally, effectively removing themselves from the control of any one jurisdiction.

IPPD: Integrated Product and Process Development (a body of knowledge).

Integrated Team: A group of people with complementary skills who collaborate to deliver specified work products. An integrated team may be either colocated or distributed. Contrast with Virtual Team (below).

Laissez-Faire: From the French "allowed to be." Refers in this context to the management style where employees function best when left alone.

Participative Leadership: The equalization of power and sharing of problem solving with followers by consulting them before making a decision.

Transformation Leadership: Combining four dimensions; charisma, individualized consideration, intellectual stimulation, and inspirational motivation.

Transactional Leadership: Combining three dimensions; contingent reward, management-by exception, management-by-exception (passive).

Virtual Team: Group of geographically or organizationally dispersed coworkers that are assembled using a combination of telecommunications and information technologies to accomplish an organizational task.

Effective Learning Through Optimum Distance Among Team Members

Bishwajit Choudhary
Information Resources Management Association, USA

INTRODUCTION

For several years, researchers have argued that too much closeness or distance among the team members inhibits intellectual debate and lowers the quality of decision-making. In fact it is often said that if two people always agree, then one is useless and if they always disagree, then both are useless. While too much "closeness" leads to copycat attitude, too much "distance" among the team members results in incompatibility. Creating teams in which the members experience "optimum distance" is not easy.

In this backdrop, we have identified certain gaps in the contemporary organizational learning theories and developed conceptual constructs and conditions that are likely to cause optimum distance in teams.

BACKGROUND

Organizational learning (OL) gained currency when interpreting market information ahead of competitors was seen as a source of competitive advantage (DeGeus, 1988). Organizations increasingly realize the need to maintain a right degree of balance between exploiting the existing and exploring new knowledge base (Cox, 1993; Jackson et al., 1995). Concepts such as double loop learning (Argyris, 1977) and generative learning (Senge, 1990) have underlined the need for innovation and creativity in learning processes.

Research in organizational networks has primarily focused on knowledge creation at organizational levels (Nonaka et al., 1994). Almost all the analyses of networks have focused on inter-organizational groupings (Van De Ven & Walker, 1984). Andersen et al. (1994) define a business network as a set of two or more inter-connected business relationships and claim that the parties in networks have traditionally been shown to come from the same industry.

MAIN THRUST OF THE ARTICLE

In spite of pioneering attempts to conceptualize OL, lately, the researchers have expressed concerns. Ritcher (1998) remarks that the current literature does not adequately explore the dynamics of learning process. Nonaka et al. (1995) claim that "There is very little research on how knowledge is actually created *and hence there is a need to understand the dynamics of knowledge creation*" (italics added).

Alter and Hage (1993) have argued that new theories should be developed to encompass knowledge creation as a result of inter-firm collaboration. Macdonald (1995) claims that the current theories have neglected external-to-firm factors. The aim of OL should be to enhance innovation and not learning merely for the sake of it (Nonaka et al., 1994). D'Aveni (1995) argues that businesses need breakthrough innovations through industry-oriented learning processes and adequately respond to the dynamic external environment.

We now summarize the critical overview of the OL literature presented previously:

- Absence of external-to-firm factors in OL processes.
- Unclear conceptualization of optimum distance in teams.

WHAT IS OPTIMUM DISTANCE?

We delve deeper into OL processes by understanding the factors that constitute perceived distance among the team members by defining the relevant concepts.

Member Distance (MD)

Inkpen (1988) argues that in inter-organizational teams, distrust among members from the participating firms (who perceive each other as competitors) inhibits learning. We believe that this distrust among the team members is the result of the so-called "member distance". Member distance (or MD) reflects overall differences among the members due to objective factors (e.g., members' experience and education) and subjective factors (e.g., members' behavior, values and personality).

Extending Inkpen's (1988) classification of inter-organizational teams, we propose three team compositions comprising managers from:

- Different departments within the same firm (cross-functional teams).

- Same industry-sector but different firms (forums comprising partners).
- Different industries, but similar department (e.g., co-ordination forums for inter-sector policies or standards body, etc.).

Knowledge Distance (KD)

Managers in different industries need to know some basic industry-specific issues. For example, in the banking sector, managers need the knowledge of payment systems, customer support, and so forth, while in the telecommunication sector, managers need the knowledge of communication networks, mobile devices and so on. Knowledge distance (KD) conceptualizes industry-specific knowledge differences among managers from different sectors.

Professional Distance (PD)

Prolonged working and dedicated experiences within a specific department can influence managers' behavior at the workplace. Zuboff (1988) cites several examples (showing the impact of automation on employees' behavior). We refer to job-specific behavioral differences among managers as professional distance. Stated formally, professional distance (PD) comprises intuitive and often subjective personality differences among managers from different departments.

We now summarize some important observations on KD and PD:

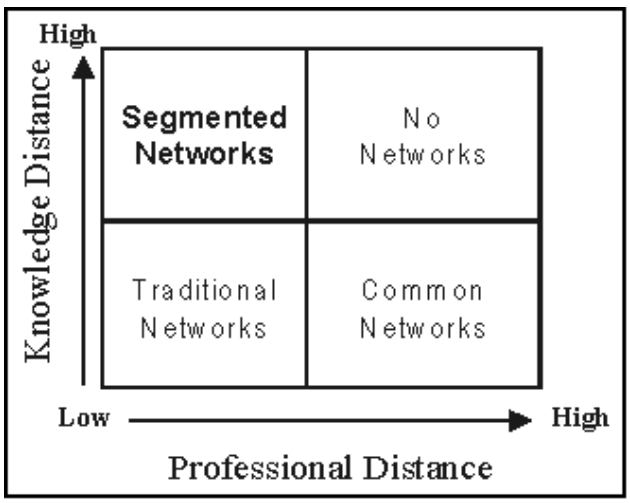
- KD captures dissimilarities among managers due to external-to-firm and knowledge specific factors. PD conceptualizes department-specific, behavioral differences among managers.
- KD represents the member-distance at a macro (inter-sector or firm) level. PD represents subjective and more complex personality-based differences at a micro (or department) level.
- Since PD depends on the department dynamics (which impact managers' behavior at the workplace), PD will be low between managers of similar departments even if they come from different industries. KD in a team will be low only when the managers come from firms within the same industry.
- The unit of analyses of learning processes in an individual (manager).

Since the large firms usually have a number of specialized departments, we may conceptualize such departments as micro-level "professional personality domains" and firms as "macro-level knowledge domains". KD and PD can then be used to conceptualize different team compositions as shown in Figure 1.

It is evident that the traditional networks occur when both KD and PD are low. These are the internal-department teams and found in all organizations. We refer to them as the "traditional networks".

When integrating knowledge from different departments is needed (in projects, for example), cross-functional teams are often created. In such teams, members have different behavioral approaches to problem solving (hence, high PD). However, their behavioral approaches are often complementary as well given their respective dedicated experience in different functional areas (as technology, business, finance and so on). However, as the managers in cross-functional teams come from the same firm, they share similar broad-based knowledge on industry-level issues (hence, low KD). In addition to low KD, the managers also

Figure 1. Balancing member distance for effective learning in various teams



share common objectives as they work for the same firm. Since such teams are very common, we refer to them as the "common networks".

When KD and PD are both high, no teams can (and should) be formed. Such teams would be difficult to manage and lack common interest areas. We do not discuss this scenario further.

SEGMENTED NETWORK

Varying knowledge-based competencies and similar professional personalities mark the conditions in this team. The presence of high KD creates knowledge-based differences, while low PD implies that members essentially share similar behavior at workplace. This so-called optimum balance between members' distance and closeness is achieved through knowledge-based differences but behavior-based similarities. High KD (through task-oriented conflicts) will lead to higher quality decisions as well (Schweiger et al., 1986) as well.

Segmented network (SN) is hence a team with conditions for optimum distance. It can typically be a forum to create best practices and inter-sector policies (e.g., standards, benchmarks, quality requirements). Use of macro (firm and industry) level and micro (department and individual) level factors facilitate our conceptual understanding of the relation between manageability of knowledge flow and team compositions. Such cross-sector forums are emerging fast (e.g., standards bodies, benchmarking forums, professional services (as consulting, law) firms). Although the participants in SNs come from different sectors, they usually serve the same (or same type of) user base, putting customers at the center-stage.

FUTURE RESEARCH

We now summarize the limitations of this work and identify areas of future research.

Measuring KD and (especially) PD may not be easy, but must be done one way or another. Conceptualizing learning processes with the assumptions of discrete activity is not realistic in a network sector where the value-creating activities are highly information-intensive and overlapping. We also ignored the impact of hierarchy and power on team dynamics for the sake of simplicity. Finally, the challenge in managing segmented networks would be high-level political issues and in turn, that would test the leadership skills of the team's coordinator.

CONCLUSION

By using the conceptual constructs as KD and PD we tried to plug the gaps in the contemporary organizational learning research. As the impact of external forces on firms' performance becomes increasingly pronounced, organizations will need learning processes that are inter-sector and not merely internal to the firm. Segmented networks fulfill this very important premise. Past research also supports our view that high degree of task-oriented conflicts (or high KD, as is the case in segmented networks) increases the quality of decision (Schweiger et al., 1986, 1989). Teams with a high degree of task-oriented conflicts are easier to manage than and may indeed be preferred to teams with a high degree of personality-oriented conflicts. Manageability of teams is a key issue.

Note that the common networks (or the cross-functional teams) also carry the important element of balance within them and their usefulness is well known. SNs offer a similar possibility at an inter-sector level, thereby complementing the internal-to-firm focus of the common networks. SNs are novel and manageable. Inter-sectors forums are growing especially fast in the network industries and so is the need for OL theories to conceptualize and understand SNs. In the network industries (such as banking, telecom sectors), interoperability of systems and procedures is a key success factor.

REFERENCES

Alter, C., & Hage, J. (1993). *Organizations working together.* Newbury Park, CA: Sage Publications Inc.

Anderson, J.C., Håkansson, H., & Johanson, J (1994, October). Dyadic business relationships within a business network context. *Journal of Marketing, 58,* 1-15.

Argyris, C. (1977, September/October). Double loop learning in organizations. *Harvard Business Review, 55,* 115 – 125.

Cox, T.H. (1993). *Cultural diversity in organizations: Theory, research and practice.* San Francisco: Barrett-Koehler.

D' Avneni, R.A. (1995). Coping with hypercompetition: Utilizing the new 7S's framework. *Academy of Management Executive, 9* (3), 45-60.

Dodgson, M. (1993, March-April). Organizational learning. *Harvard Business Review, 75,* 375-394.

Inkpen, A. (1998). Learning, knowledge acquisition, and strategic alliances. *European Management Journal, 16*(2), 223-229.

Jackson, S, May, K.E., & Whitney, K. (1995). Understanding the dynamics of diversity in decision-making teams. In R.A Guzzo & E. Salas (Eds.), *Team effectiveness and decision-making in organizations* (pp. 204-261). San Francisco: Jossey-Bass.

Macdonald, S. (1995, September/October). Learning to change: An information perspective on learning in the organization. *Organization Science, 6*(5), 557-568.

Nonaka, I., Byosiere, P., Borucki, C.C., & Konno, N. (1994). Organizational knowledge creation theory: A first comprehensive test. *International Business Review*, *3*(4), 337-351.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company.* New York: Oxford University Press.

Richter, I. (1998). Individual and organizational learning at the executive level. *Management Learning, 29*(3), 299-316.

Schweiger, D.M, Sandberg, W.R.,& Ragan, J.W. (1986). Group approaches for improving strategic decision making: A comparative analysis of dialectical inquiry, devil's advocacy, and consensus. *Academy of management Journal, 29*(1), 51-71.

Schweiger, D.M., Sandberg, W.R., & Rechner, P (1989). Experimental effects of dialectical inquiry, devil's advocacy and consensus approaches to strategic decision-making. *Academy of Management Journal, 32,* 745-772.

Senge, P.M. (1990). *The fifth discipline: The art and practice of the learning organizations.* New York: Doubleday.

Van de Ven, A.H., & Walker, G. (1984). The dynamics of inter-organizational coordination. *Administrative Science Quarterly, 29*(4), 598-621.

Zuboff, S. (1988). *In the age of smart machine-The future of work and power.* Basic Books.

KEY TERMS

Knowledge Distance (KD): Conceptualizes industry-specific knowledge differences among managers from different sectors.

Member Distance (MD): Reflects overall differences in approach among the members due to objective factors (e.g., members' experience etc.) and subjective factors (e.g., members' values etc.).

Optimum Distance (OD): Appropriate degree of closeness (or distance) among the team members achieved through the combination of KD and PD, leading to the so-called creative tension.

Professional Distance (PD): Comprises behavioral differences among managers from different departments.

Segmented Network (SN): Inter-sector teams with high KD and low PD, leading to conditions for the so-called optimum distance.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 976-979, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Effective Virtual Teams

D. Sandy Staples
Queen's University, Canada

Ian K. Wong
Queen's University, Canada

Ann-Frances Cameron
HEC Montréal, Canada

INTRODUCTION

Virtual teams are now being used by many organizations to enhance the productivity of their employees and to bring together a diversity of skills and resources (Gignac, 2005; Majchrzak, Malhotra, Stamps, & Lipnack, 2004), and it has been suggested that this will become the normal way of working in teams in the near future (Jones, Oyund, & Pace, 2005). Virtual teams are groups of individuals who work together from different locations (i.e., are geographically dispersed), work at interdependent tasks, share responsibilities for outcomes, and rely on technology for much of their communication (Cohen & Gibson, 2003). While the use of virtual teams is more common in today's organization, working in these teams is more complex and challenging than working in traditional, collocated teams (Dewar, 2006), and success rates in virtual teams are low (Goodbody, 2005). This article suggests best practices that organizations and virtual team members can follow to help their virtual teams reach their full potential.

In this article, virtual team best practices are identified from three perspectives: organizational best practices, team leadership best practices, and team member best practices. Ideas for best practices were identified from three sources: six case studies of actual virtual teams (Staples, Wong, & Cameron, 2004); the existing literature on virtual teams; and the existing literature on traditional (i.e., collocated) teams and telecommuting (i.e., research on virtual work at the individual level).

Table 1. Organizational best practices for effective virtual teams

• Carefully select team members for diversity
• Supply the team with sufficient resources, support, and information technology tools
• Develop human resource policies that reward team efforts and stimulate virtual team performance
• Provide the team with an appropriate level of autonomy
• Use standard processes and procedures
• Develop an organizational culture that stimulates the sharing of information

ORGANIZATIONAL BEST PRACTICES

There are six best practices that organizations that employ virtual teams should follow. Table 1 contains a list of these practices, each of which is explained next.

Carefully Select Team Members for Diversity

The distributed nature of virtual teams allows a diversity of backgrounds, experiences, ideas, thoughts, abilities, and perspectives to be assembled within a single team. Organizations forming virtual teams should take advantage of this, selecting team members with diverse backgrounds and skills. The importance of team diversity was identified in both the case studies and the traditional team literature (e.g., Bettenhausen, 1991; Cohen, 1994). In particular, research has shown that diversity provides information-processing benefits to teams such that they are more effective at their tasks (Dahlin, Weingart, & Hinds, 2004). Working on a diverse team can also be more rewarding, interesting, and fun as team members get the opportunity to learn about new cultures and interact with people beyond their own work location.

Supply Sufficient Resources, Support, and Information Technology (IT) Tools

Organizations have to supply virtual teams with sufficient resources including financial resources, time, facilities, hardware, software, information technology (IT) support, communication channels, technical equipment, and proper training (Jones et al., 2005). The traditional team literature suggests that team building activities and training members

how to work in teams are important because they ensure that employees develop the knowledge required to contribute to organizational performance (Cohen, 1994). In virtual teams it is especially difficult for team members to get to know one another. Thus, organizations may need to provide extra resources for extensive team building exercises.

Since virtual teams often need to communicate electronically, appropriate IT tools, training on how to use available IT and communication systems, and readily-available technical support are also important to virtual teams (Duarte & Snyder, 2001; Fisher & Fisher, 2001; O'Hara-Devereaux & Johansen, 1994; Pinsonneault & Boisvert, 2001; Staples et al., 2004). High quality systems for audioconferencing are essential, as are e-mail systems and systems for storing, accessing, and sharing electronic files. Collaborative whiteboard tools are very useful, as they support the interactive sharing of information during electronic meetings. Electronic voting and brainstorming tools can also be useful, depending upon the nature of the team's task. Instant messaging is a powerful tool that enhances the presence and connectedness of remote parties, and its informal nature enhances social interaction. Blogging tools give people the ability to share their views (Jones et al., 2005), possibly helping others understand their situation and perspective.

Develop Human Resource Policies that Stimulate High Virtual Team Performance

Policies must be designed in such a way that virtual team members are recognized, supported, and rewarded for their work (Duarte & Snyder, 2001; Jones et al., 2005). Providing team-based (rather than individual performance-based) rewards to team members can increase team cohesiveness, motivation, and effectiveness (e.g., Cohen, Ledford, & Spreitzer, 1996; Hertel, Konradt, & Orlikowski, 2004; Lawler, 1986, 1992). Since virtual team members are not seen every day in a central office, it is also possible that they may be overlooked for promotional opportunities (Duarte & Snyder, 2001). Therefore, special career development opportunities, such as job rotations and opportunities to present to team sponsors/executive groups, should be created for virtual team members so that this "out of sight, out of mind" phenomenon does not occur (Jones et al., 2005; Pinsonneault & Boisvert, 2001).

Provide the Team with Appropriate Autonomy

Consistent with traditional team research (Cohen & Bailey, 1997), virtual team members interviewed in the case studies reported that little involvement from senior management was usually preferred over hands-on management, as long as the organization still provides the necessary funds and resources. Worker autonomy is shown to have clear benefits such as enhanced worker attitudes and performance (Stewart, 2006). Organizations should give team members the power to take action and make decisions while still providing the team with the information it needs to make sound business decisions (Cohen, 1994). Organizations should provide information on processes, quality, customer feedback, business results, competitor performance, and organizational changes.

Use Standard Processes and Procedures

The use of standard processes and procedures, such as having a project charter, can reduce the time needed for team start-up and may eliminate the need for unnecessary reinvention of operating practices every time a new team is created (Duarte & Snyder, 2001; Gignac, 2005). For virtual teams that rarely meet face-to-face, standard communication procedures and policies are extremely important so that norms and expectations are clear (Duarte & Snyder, 2001; Fisher & Fisher, 2001; Grenier & Metes, 1995). An initial face-to-face team meeting can allow team members to develop communication norms and agreements on how they are going to work together. An experienced virtual team facilitator for the start-up phase can be valuable to ensure a team starts with a strong foundation (Gignac, 2005).

Develop an Organizational Culture that Stimulates the Sharing of Information

Sharing information effectively is critical to virtual team success (Gignac, 2005; Jones et al., 2005). Organizational culture influences how individuals in an organization behave and, thus, plays a large role in determining how well a virtual team functions. Therefore, organizations should work to build norms and values that promote communication and the sharing of information (Goodbody, 2005). The traditional team research also identified the importance of having a supportive culture. Organizations should create a cooperative work environment where norms are established that reinforce and support team behaviors such as sharing information, responding appropriately to team members, and cooperating (Bettenhausen, 1991), as such an environment is critical for effective team performance (Tjosvold, 1988).

TEAM LEADERSHIP BEST PRACTICES

There are seven best practices relating to the leadership and management of the virtual team. Table 2 contains a list of all seven team leadership practices, each of which is explained next.

Table 2. Management and team leader best practices for effective virtual teams

• Set goals and establish direction
• Recognize and manage the diversity within the team
• Provide feedback via coaching and modelling
• Build trust through open communication, honest behavior, and delivering on commitments
• Empower the team
• Motivate the team
• Use appropriate leadership styles at appropriate times

Set Goals and Establish Direction

The virtual team literature strongly suggests that effective leaders understand the importance of defining a vision for the virtual team (Fisher & Fisher, 2001; Grenier & Metes, 1995; Lipnack & Stamps, 1997; O'Hara-Devereaux & Johansen, 1994). According to Lipnack and Stamps (1997), a predictor of virtual team success is the clarity of its purpose and vision. To succeed, teams must turn their purpose and vision into action by assigning roles and responsibilities (Fisher & Fisher, 2001; Grenier & Metes, 1995; O'Hara-Devereaux & Johansen, 1994; Pinsonneault & Boisvert, 2001) and setting clear objectives goals (Jones et al., 2005; Lipnack & Stamps, 1997).

In the virtual team case studies, 64% of team members recognized this need to carefully set clear goals and realistic timelines. To accomplish this, management and team leaders can first develop a "roadmap" with realistic timelines that are compatible with the expectations of senior management. Next, the critical path through the project should be identified. Based on this path, major milestones should be set. Whether or not it affects them directly, all team members should be frequently reminded of the next milestone so that the team is aware of their interdependencies and how their work affects others. Focusing on milestones and deliverable dates will help members keep the "big picture" in mind when working on their individual tasks. Successful virtual teams are those that, with the help of a focused manager, are consistently able to meet milestones within the allotted time.

Recognize and Manage the Diversity within the Team

Virtual teams are typically more diverse than traditional teams since team members are drawn from multiple locations (Gibson & Cohen, 2003). Information diversity is a benefit for a team since it brings wider knowledge and expertise to bear on the team's task. However, some types of diversity can potentially impede teamwork and communication. People from different countries and cultures may have different values, expectations, language skills, communication norms, and appearances (Goodbody, 2005; Jones et al., 2005). If not managed well, these differences can easily lead to misunderstandings and the formation of team sub-groups, which fracture the team and impede sharing and performance. Managers must help the team recognize the diversity within the team, and openly discuss individual differences and expectations such that new team-level expectations are established (Gibson & Cohen, 2003). Also, diversity of location often means working across multiple time zones, which creates scheduling challenges. Managers should ensure that time differences are visible, respected, and that one group is not favored at the expense of the rest of the team.

Provide Feedback via Effective Coaching and Modeling

Team leaders need to provide members with timely feedback about their performance so team members know what they can do to continuously improve their performance (Duarte & Snyder, 2001). A manager's ability to provide remote employees with advice and help can increase the effectiveness of the remote employees (Staples, 2001). In virtual teams, this may require getting informal input from various people who interact with team members both within and outside of the organization. Virtual leaders also need to model appropriate virtual work habits (Jones et al., 2005; Staples, Hulland, & Higgins, 1999). To accomplish this, the telecommuting literature suggests managers should keep remote employees well-informed of organizational activities, provide regular feedback on performance and progress, establish well-structured and constant communications, and be available at hours that fit with work routines of remote employees (Pinsonneault & Boisvert, 2001).

Build Trust through Open Communication, Honest Behavior, and Delivering on Commitments

Establishing trust within a virtual team is critical (Gignac, 2005: Jones et al., 2005). Without trust, productivity suffers as team members spend time playing politics instead of working on real business issues (Fisher & Fisher, 2001). To build trust, it is important for team leaders to communicate openly and frequently with team members. The use of instant messaging (IM) can help do this (Jones et al., 2005). Perhaps the single most important variable that affects trust is honesty. Leaders who demonstrate openness about their actions will find that members respond with sincerity. How a leader listens and

communicates with his or her team members is very much related to team effectiveness and trust (Cohen & Bailey, 1997). Listening to team members and visibly keeping commitments increases trust, whereas broken promises diminish it (Fisher & Fisher, 2001). Effective leader communication is especially important to the performance of teams that are geographically dispersed (Cummings, 2006).

Empower the Team

This best practice is related to the organizational practice of more generally providing autonomy to teams. Team leaders also have to provide the appropriate level of autonomy, setting overall team goals and establishing direction while allowing individual team members to decide how they carry out their own specific tasks (Cohen & Bailey, 1997; Fisher & Fisher, 2001). Leaders who trust team decisions can give the members a sense of ownership. This approach is particularly important in a virtual team environment where geographic separation makes micromanagement and direct observation impractical.

Motivate the Team

In a virtual team environment where tasks may appear unconnected, the "big picture" is not always easy to visualize, making it difficult for employees to remain committed to the project. Thus, team leaders can play a key role in keeping virtual team members motivated. Motivation can be stimulated by making the importance of the team's task clear (such that passion for the team's cause is created) and by demonstrating how the project will result in significant outcomes for the individual team members (Fisher & Fisher, 2001; Staples et al., 2004). By linking team success to individual success and opportunities, team members will be highly motivated to succeed on the project. Visibly celebrating achievements can also build motivation and enthusiasm (Goodbody, 2005).

Use Appropriate Leadership Style

Over a quarter of case study team members (Staples et al., 2004) reported that appropriate leadership at the *appropriate time* was one of the key elements of a successful virtual team. Case study participants suggested that during the initial phases of the project, the appropriate leader is one who can "whip up enthusiasm" to motivate the team and create a sense of team spirit. During the later stages, the effective leader is someone who is "getting the right people together and keeping everybody on task and keeping everything going." Therefore, the style and activities of team leaders have to be appropriate for the team's development stage and needs at that particular time.

Table 3. Team member best practices for effective virtual teams

• Communicate effectively
• Have the necessary skill sets
• Be highly motivated
• Be supportive of other team members
• Be action-oriented

TEAM MEMBER BEST PRACTICES

Suggestions for what makes individual members of virtual teams effective include specific behaviors, as well as attitudes and beliefs. The five general characteristics of effective virtual team members are listed in Table 3 and described next.

Communicate Effectively

Research in traditional teams, virtual teams, and telecommuting recognizes that the ability to communicate effectively is a critical skill (Cohen, 1994; Goodbody, 2005; Jones et al., 2005; Pinsonneault & Boisvert, 2001; Staples, 2001). Eighty-four percent of team members interviewed in the case studies also recognized the importance of effective communication in building a successful team. Communication involves transferring ideas, sharing information, listening and internalizing the ideas of others, and notifying team members of any problems or issues. This can be challenging in a virtual team where face-to-face communication and impromptu meetings are infrequent, if not impossible. To solve this problem, virtual team members suggest working hard to keep lines of communication open, and developing or finding the right communications tools that make up for the loss of face-to-face time and provide for informal interactions. For example, team members can use e-mail or IM as a "virtual coffee pot or water cooler" around which personal conversations can occur or as an alternative medium to reach other team members (Cameron & Webster, 2004; O'Hara-Devereaux & Johansen, 1994). These informal interactions can also help to create team spirit.

In addition, team members themselves have to be responsive, quickly returning telephone calls and responding to e-mails, even if it is just to say, "I don't have time right now but I'll get back to you in two days with the answer." Recipients can also confirm that the message was received and ensure that the major points in the message were understood. Setting communication norms such as these helps to avoid misunderstanding that can occur when communicating electronically. IM tools also have the ability to indicate availability (i.e., online, temporarily away, off-line, etc.), potentially enhancing the presence a virtual team member projects to others and his/her awareness of the other team members' availability (Jones et al., 2005).

Have the Necessary Skill Sets

The effectiveness of a team depends on the collective knowledge and skills of its members. In order to make good decisions, team members need the appropriate knowledge and skills (Cohen, 1994; Jones et al., 2005). Specific skills mentioned by the virtual team members in the case studies included the ability to organize effectively, a strong competency in an individual's functional area of responsibility, adequate technical skills to use the information and technology tools available, and good time management skills. Specific social skills mentioned in the virtual team literature include learning how to negotiate creatively, mediating online disputes, and making new members of the team feel included (Grenier & Metes, 1995).

Be Highly Motivated

In some cases, a particular practice may exist at both the management and individual level. As described previously, motivating is an activity that must be performed by team leaders. The self-managing nature of many virtual teams means that *self*-motivation and *self*-discipline are also essential. Virtual team members must be able to work independently and be motivated to make appropriate decisions. This is made possible by having clear goals and responsibilities, and having high personal commitment and motivation to the team, along with having the resources and information needed to do the job (Lipnack & Stamps, 1997). The ability to self-motivate is important since virtual team members are often working far apart from their other team members. Virtual workers should also have a low preference or need for social interaction or have the ability to fulfill this need outside of the work team (Pinsonneault & Boisvert, 2001).

Be Supportive of Other Team Members

The way in which team members interact with each other influences team effectiveness (Cohen, 1994; Jones et al., 2005). Supporting team members involves working together with a sense of energy and team spirit and sharing ideas and expertise to help others (Cohen, 1994). Several dimensions of a supportive team emerged during the case study interviews. First, team members felt that it was important to recognize when someone else did a good job, and to congratulate or thank them accordingly. Second, interviewees sought a respectful team environment where members were not afraid to openly discuss ideas. Third, interviewees reported that the ability to get along with other team members was an essential quality of an effective virtual team member.

Be Action-Oriented

Interviewees felt that individuals should have an action-oriented approach when participating in a virtual team. Case study members of one team described a top performing virtual team member as someone who is a "doer," is "proactive," "uses an entrepreneurial approach," and "looks for solutions." One virtual team member stated that successful virtual team members are those who "organize their thoughts into actions or proposals that get a good buy" or influence the rest of the group.

FUTURE TRENDS AND OPPORTUNITIES

As global competition increases and trading barriers decline, the creation of teams comprised of members from many different locations is expected to become the norm. With this virtual team growth, knowing the best practices for team members, leaders, and organizations with virtual teams becomes very important. While some useful best practices are explained previously, some questions are left unanswered. These questions create many opportunities for future research such as:

- Which of the best practices are most critical for team effectiveness?
- Does the impact of certain practices on effectiveness vary depending on the task, organizational context, and/or the national and cultural background of the team members? How?
- Does one set of practices (i.e., individual, managerial, or organizational) take precedence such that those practices have to be in place before the other practices have a positive effect?
- How does an organization ensure that best practices are followed?
- Can training programs be developed for managers and leaders, and for members of virtual teams? What should be in these training programs, and how should they be delivered?
- Can policies be developed and norms established in organizations such that supportive practices, which research suggests lead to effective virtual work, are followed? How can this be done most effectively?
- Do the best practices vary for hybrid teams, in which some members are co-located and others are working remotely?

CONCLUSION

The ideas presented in this article should help organizations create and maintain more effective virtual teams. Individual virtual team members, virtual team leaders, and organizations can all follow best practices that contribute to virtual team effectiveness. Given the growing use of virtual teams in organizations today, there is a need to more deeply understand what makes a virtual team effective. We hope we have made a contribution in that direction, and we look forward to other researchers and practitioners answering some of the additional questions posed.

REFERENCES

Bettenhausen, K. L. (1991). Five years of group research: What we have learned and what needs to be addressed. *Journal of Management, 17*(2), 345-381.

Cameron, A. F., & Webster, J. (2004). Unintended consequences of emerging technologies: Instant messaging in the workplace. *Computers in Human Behavior, 21*(1)*,* 85-103.

Cohen, S. G. (1994). Designing effective self-managing work teams. In M. M. Beyerlein, D. A. Johnson, & S. T. Beyerlein (Eds.), *Advances in interdisciplinary studies of work teams, series of self-managed work teams* (Vol. 1, pp. 67-102). Greenwich, CT: JAI Press.

Cohen, S. G., & Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management, 23*(3), 239-290.

Cohen, S. G., & Gibson, C. B. (2003). Putting the team back in virtual teams. Paper presented at the 18th Annual Conference of the Society for Industrial/Organizational Psychology, Orlando, FL.

Cohen, S. G., Ledford, G. E., & Spreitzer, G. M. (1996). A predictive model of self-managing work team effectiveness. *Human Relations, 49*(5), 643-676.

Cummings, J. N. (in press). Leading groups from a distance. In S. Weisband (Ed.), *Leadership at a distance: Interdisciplinary perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates.

Dahlin, K. B., Weingart, L. R., & Hinds, P. J. (2005). Team diversity and information use. *Academy of Management Journal, 48*(6)*,* 1107-1123.

Dewar, T. (2006). Virtual teams—Virtually impossible? *Performance Improvement, 45*(5), 22-25.

Duarte, D. L., & Snyder, N. T. (2001). *Mastering virtual teams: Strategies, tools, and techniques that succeed*. San Francisco: Jossey-Bass Inc.

Fisher, K., & Fisher, M. D. (2001). *The distance manager: A hands on guide to managing off-site employees and virtual teams*. New York: McGraw-Hill.

Gibson, C. B., & Cohen, S. G. (2003). The last word: Conclusions and implications. In C. B. Gibson & S. G. Cohen (Eds.), *Virtual teams that work: Creating conditions for virtual team effectiveness* (pp. 403-421). San Francisco: John Wiley & Sons, Inc.

Gignac, F. (2005). *Building successful virtual teams*. Norwood, MA: Artech House.

Goodbody, J. (2005). Critical success factors for global virtual teams. *Strategic Communication Management, 9*(2)*,* 18-21.

Grenier, R., & Metes, M. (1995). *Going virtual*. Upper Saddle River, NJ: Prentice Hall, Inc.

Hertel, G., Konradt, U., & Orlikowski, B. (2004). Managing distance by interdependence: Goal setting, task interdependence, and team-based rewards in virtual teams. *European Journal of Work and Organizational Psychology, 13*(1), 1-28.

Hofstede, G. (1997). *Cultures and organizations: Software of the mind*. New York: McGraw-Hill.

Jones, R., Oyund, R., & Pace, L. (2005). *Working virtually: Challenges of virtual teams*. Hershey, PA: Cybertech Publishing.

Lawler, E. E. (1986). *High-involvement management: Participative strategies for improving organizational performance*. San Francisco: Jossey-Bass & Associates.

Lawler, E. E. (1992). *The ultimate advantage: Creating the high involvement organization*. San Francisco: Jossey-Bass & Associates.

Lipnack, J., & Stamps, J. (1997). *Virtual teams: Reaching across space, time, and organizations with technology.* New York: John Wiley & Sons.

Majchrzak, A., Malhotra, A., Stamps, J., & Lipnack, J. (2004). Can absence make a team grow stronger? *Harvard Business Review, 82*(5), 131.

O'Hara-Devereaux, M., & Johansen, R. (1994). *Global work: Bridging distance, culture & time*. San Francisco: Jossey-Bass Inc.

Pinsonneault, A., & Boisvert, M. (2001). The impacts of telecommuting on organizations and individuals: A review

of the literature. In N. J. Johnson (Ed.), *Telecommuting and virtual offices: Issues & opportunities* (pp. 163-185). Hershey, PA: Idea Group Publishing.

Staples, D. S. (2001). Making remote workers effective. In N. J. Johnson (Ed.), *Telecommuting and virtual offices: Issues & opportunities* (pp. 186-212). Hershey, PA: Idea Group Publishing.

Staples, D. S., Hulland, J. S., & Higgins, C. A. (1999). A self-efficacy theory explanation for the management of remote workers in virtual organizations. *Organization Science, 10*(6), 758-776.

Staples, D. S., Wong, I. K., & Cameron, A. F. (2004). Best practices for virtual team effectiveness. In D. Pauleen (Ed.), *Virtual teams: Projects, protocols and processes* (pp. 160-185). Hershey, PA: Idea Group Publishing.

Stewart, G. L. (2006). A meta-analytic review of relationships between team design features and team performance. *Journal of Management, 32*(1), 29-55.

Tjosvold, D. (1988). *Working together to get things done: Managing for organizational productivity*. Lexington, MA: Lexington Books.

KEY TERMS

Blog: A blog (short for Weblog) is a personal online journal, containing the author's views and reflections on some topic about which he/she chooses to write.

Communication Norms: In the context of virtual teams, communication norms are typical routines and expectations for communicating within a virtual team using the communication media that the team has available to it (e.g., electronic communication such as e-mail or instant messaging, telephone, etc.). Responding within a day to all e-mails, even if just to say, "I'm busy but will get to this tomorrow," is an example of a virtual team communication norm.

Hybrid Team: A team (i.e., a group of individuals who work on interdependent tasks and who share responsibility for outcomes) in which some team members work in the same location and other members work remotely.

Instant Messaging (IM): Information systems that enable team members to exchange real time electronic messages and presence information (e.g., I'm off-line, online, busy, away, on the phone, etc.). Logs of the interaction can be captured, and some systems allow files to be exchanged.

Organizational Culture: The collective programming of the mind that distinguishes the members of one organization (or part of an organization) from another (Hofstede, 1997)

Remote Worker: An individual who works at a different location than his/her co-worker and/or manager. That person is remote, in terms of physical presence from his/her colleagues.

Team Diversity: The combined variety of skills, backgrounds, experiences, ideas, thoughts, abilities, and perspectives that individuals bring to their team.

Team Effectiveness: The ability of a team to perform its tasks on time, on budget, and with acceptable quality, as well as the satisfaction, motivation, and commitment of the team members.

Telecommuting: The practice of working from the employee's home instead of physically commuting to a company office location. The connection to the office is done via telecommunications, rather than physically commuting (i.e., travelling) to the office.

Traditional Team/Collocated Team: A traditional team is a group of individuals who work on interdependent tasks, who share responsibility for outcomes, and who work together at the same location (i.e., their office/work area is in the same general location).

Virtual Coffee Pot/Virtual Water Cooler: Using electronic communication (such as e-mail or instant messaging) to conduct informal interactions (personal or non-work conversations) that would normally be discussed around the office water cooler or coffee areas in a face-to-face work environment.

Virtual Team: A group of individuals who work together from different locations (i.e., are geographically dispersed), work at interdependent tasks, share responsibilities for outcomes, and rely on technology for much of their communication.

Effectiveness of Web Services: Mobile Agents Approach in E-Commerce System

Kamel Karoui
University of Manouba, Tunisia

Fakher Ben Ftima
University of Manouba, Tunisia

INTRODUCTION

With the development of the Internet, the number of people buying, selling, and performing transactions is expected to increase at a phenomenal rate. The emergence of e-commerce applications has resulted in new net-centric business models. This has created a need for new ways of structuring applications to provide cost-effective and scalable models.

Mobile Agents (MA) systems are seen as a promising paradigm for the design and implementation of distributed applications, including e-commerce. MA are also useful in applications requiring distributed information retrieval because they move the location of execution closer to the data to be processed. While MA have generated considerable excitement among the research community, they have not been applied into a significant number of real applications.

Web services (WS) are emerging as a dominant paradigm for constructing distributed business applications and enabling enterprise-wide interoperability. A critical factor to the overall utility of WS is a scalable, flexible and robust discovery mechanism; an application can be built by integrating multiple services together to make a more efficient service. WS represent a major development in the e-commerce sector. They enable companies to capitalize on their existing architecture by making their application services accessible via the Internet.

The application of MA and WS technologies to e-commerce will provide a new way to conduct business-to-business (B2B), business-to-consumer (B2C), and consumer-to-consumer transactions (C2C) and facilitate the communication between heterogeneous environments.

In this article, we first focus on these two technologies of actuality and show their integration in an e-commerce system. Second, we present different kinds of interaction between MA and WS and study their effect on application performance. We also study an example that illustrates an e-commerce system including three categories of transactions:

-Shopping transactions: a customer delegates one MA for research and purchase of articles online. The MA will interact with available WS to find the article and its best price.

-Salesman transactions: to valorize their products, WS will invoke MA to make publicity for the customers.

-Auction transactions: for this type of transaction, a MA (respectively a WS) can sell and buy a product from/to others MA (WS) by auction.

Finally, we conclude with a discussion on our inferences and their implications.

This work is structured as follows:

Section "background" reviews the notions of e-commerce system, WS and MA paradigms. Section "Web services and mobile agents' technologies on e-commerce system" presents the integration of these two paradigms on the e-commerce system. In section "performance evaluation," we evaluate the performances of our approach and we study an illustrated example in the section "a case study." The section "future trends" presents our future perspectives and we end this work with the "conclusion" in the last section.

BACKGROUND

E-Commerce

E-commerce can be viewed as a set of processes that support commercial activities within an information network (Chaves, Martins, Monteiro, & Boavida, 2002). These activities produce information about products, events, services, suppliers, consumers, publicists, transactions, advanced search algorithms, transactional security, authentication, and so forth. In brief, e-commerce entails the development of a business vision, supported by information technology with the goal of enhancing efficiency within the process of trade (Adam, Dogramaci, Gangopadhyay, & Yesha, 1999). The fact that this technology is so fast, transactions require less human interaction and a greater reliance on autonomous software agents (Chaves, Simões, & Monteiro, 2003).

Web Services

WS are a new kind of Web application. They are self-contained, self-describing and modular applications that can be published, located, and invoked across the Web. WS perform functions, which can be anything from simple requests to complicated business processes. Once a WS is

deployed, other applications (and other WS) can discover and invoke the deployed service (Lemahieu, 2001). WS can significantly increase the Web architecture's potential, by providing a way of automated program communication, discovery of services, and so forth. Therefore, they are the focus of much interest from various software development companies (WSAP, 2000).

Mobiles Agents

MA are software programs that can travel autonomously from host to host to perform one or more tasks on behalf of a user. They can communicate (and even negotiate) with other agents and hosts. The MA paradigm proposes a new approach for designing applications in open and heterogeneous distributed environments (Nwana, Rosenschein, Sandholm, Sierra, Maes, & Guttman, 1998). Several application areas can benefit from the adoption of the MA technology: It can support electronic commerce transactions and help in information gathering, filtering, and negotiation. MA solutions provide mobility, autonomy and easy personalisation (Guttman, Moukas, & Maes, 1998).

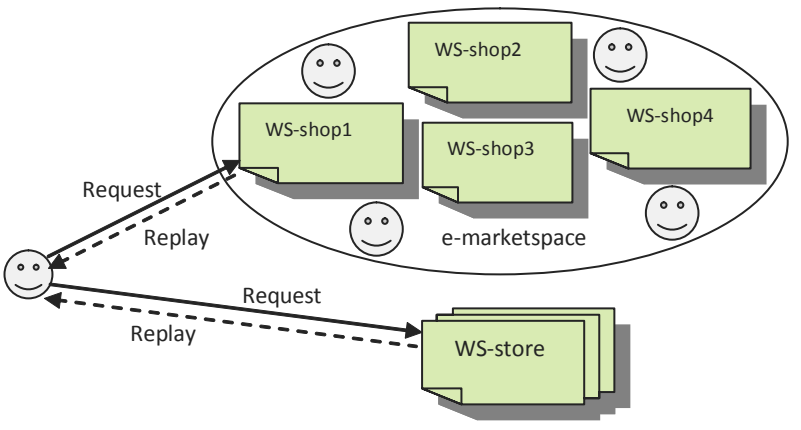
WEB SERVICES AND MOBILE AGENTS TECHNOLOGIES ON E-COMMERCE SYSTEM

E-commerce covers any form of business or administrative transaction or information exchange that is executed using any information and communication technology. It refers especially to the commercial activities conducted on the Internet. E-commerce systems not only provide commercial information, such as product price and features, but also facilitate various commercial actions, such as buying, selling, and negotiation. The popularity of software agents, in the execution of tasks related to information filtering, mapping of people with similar interests and automation of repetitive behaviours is well known (Maes, Guttman, & Moukas, 1999). It is thus without surprise that agent-based technology is seen as the one that will revolutionize e-commerce in the way it is seen today, promising a new and innovative approach in the way transactions are processed, whether they are business-to-business, business-to-consumer or even consumer-to-consumer. Using MA represents an important leap in the development of first generation (static) agent systems (El Falou & Bourdon, 2004). The possibility of working off-line, thus saving network resources, is one of the main advantages. There is no need to keep a connection active while a transaction is processed.

To achieve a result, a program tasked (e.g., a MA) can use Web services as support for its computation or processing. The program can discover Web services and invoke them fully automated. Hence, it becomes a service requester. If the Web services have a cost attached, the program knows when to search for a cheaper service and knows all the possible payment methods. Furthermore, the program might be able to mediate any differences between its specific needs and a Web service that almost fits (Bernard, 2000).

In e-commerce, this translates into automatic cooperation between enterprises (Shaw, 2000). Any enterprise requiring a business interaction with another enterprise can automatically discover and select the appropriate optimal Web services relying on selection policies. They can be invoked automatically and payment processes can be initiated. Any necessary mediation applied is based on data and process ontologies and the automatic translation of their concepts into each other (Papaioannou & Edwards, 1998; Sandholm, 1999).

Figure 1. Shopping transaction



The idea of this work is to construct a simple and efficient e-commerce system that helps customers find the best possible offer for their needs. MA plays an important role in this process, as they represent the users in their interaction with the ever-growing marketplace represented by WS. We classify interactions between MA and WS (Karoui & Ftima, 2008), in the e-commerce system into three categories: shopping transactions, salesman transactions and auction transactions.

Shopping Transactions

For this type of transaction, MA purchases in e-marketplaces (a collection of e-shops represented by WS) on behalf of their owner according to user-defined specifications. This model of e-commerce uses a customer-driven marketplace. A typical shopping agent may compare features of different products by visiting several online stores and report the best choice to its owner. The MA carries the set of features to be considered and their ideal values as specified by its owner. The MA is given one or more WS to visit and may dynamically visit other WS based on subsequent information. One example of a system that implements shopping agents is illustrated in Figure 1, where agents deal with many WS to find a product.

Salesman Transactions

For this type of transaction, MA behave like a travelling salesman who visits customers to sell his wares. This model of e-commerce uses a supplier driven marketplace and is particularly attractive for products with a short shelf-life. A supplier (WS) creates and dispatches a MA to potential buyers by giving it a list of stores (WS-s) or e-marketplaces (WS-m) to visit. The MA carries with it information about available stock and the product price. A schema implementing a salesman agent is illustrated in Figure 2.

Auction Transactions

In this type of transaction, MA can bid for and sell items in an online auction on behalf of their owners (a user or a WS). In the presence of multiple auction hosts (e-marketplace, e-store…), MA can be used for collecting information across them. An agent can make a decision to migrate to one of them dynamically, depending on the amount of information transmitted, latency, and so forth. Some advantages of using MA include allowing disconnected operation of auction agents, reducing network traffic, and facilitating quicker response during auction. One example of a system that implements mobile auction agents is illustrated in Figure 3.

PERFORMANCE EVALUATION

The quantity of information exchanged between the different WS, the type of application as well as the characteristics of the network are decisive parameters for the choice of the technology to be implemented. In this section, we will show the advantages of our system compared to the client/server model.

Figure 2. Salesman transaction

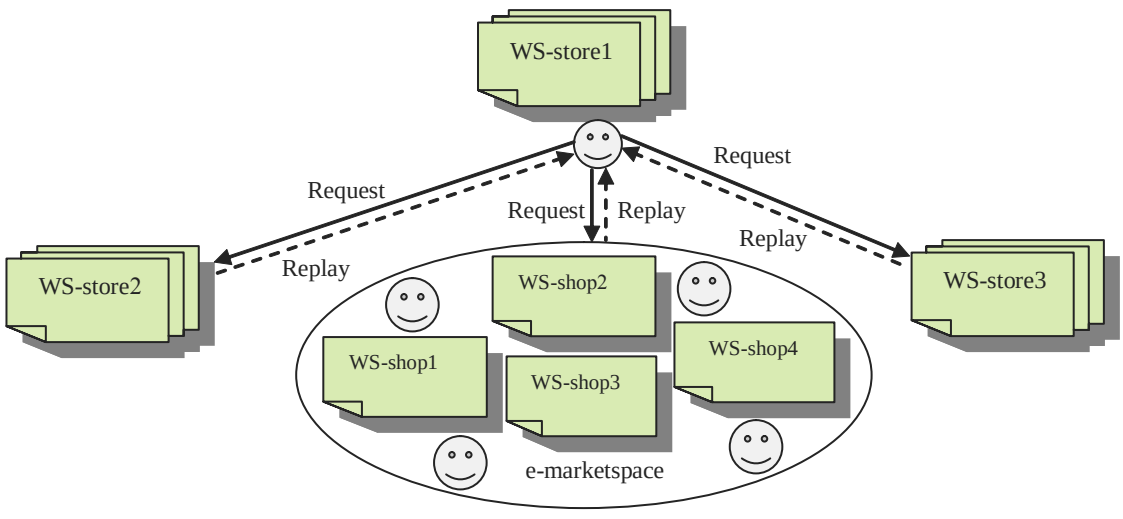
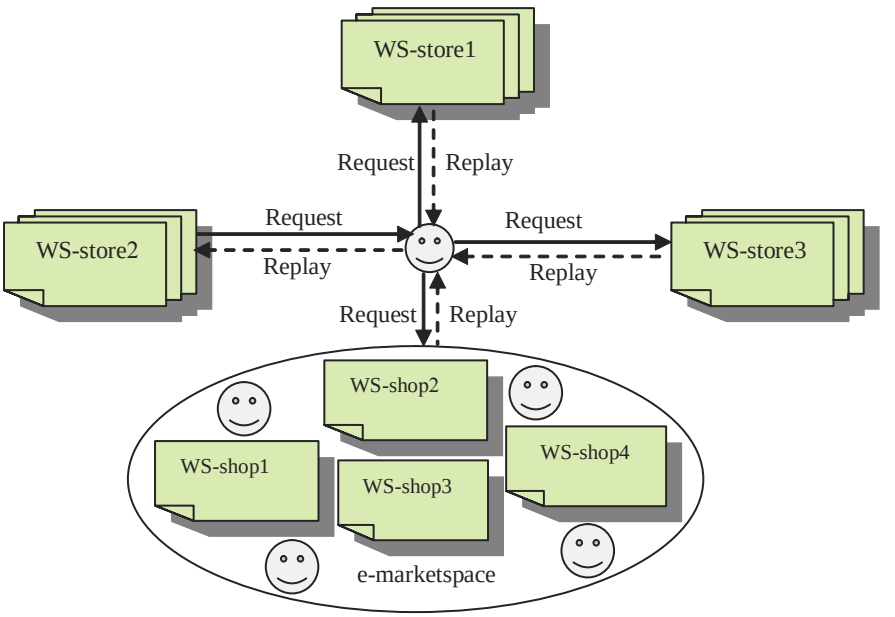


Figure 3. Auction transaction



By using the client/server technology, the different components of an e-commerce system communicate with a permanent link for requests and answers. Exchanges proceed mostly by a manager machine that redirects requests toward the concerned components.

The use of the MA-WS technology will allow:

- To make the circulation of information between the various components of the e-commerce system easier: MA move from a WS (host) to another without a well-defined route.
- To accomplish the request tasks asynchronously contrary to the client/server technology that requires a permanent link for the execution of their tasks.
- To lighten the bandwidth use: The client/server technology consumes a big part of the bandwidth due to the big volume of requests exchanged. MA encapsulate the request and migrate from a WS (host) to another when necessary.
- To reduce the number of the exchanged requests as the MA moves from host to host without every time returning the results to the original host.
- To reduce the requests execution time: by studying the communication's parameters between hosts, requests will be executed faster with the MA-WS (see details on the next paragraph).

Requests Execution Time Reduction

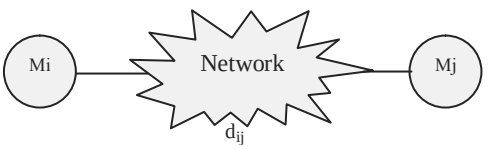
We will compare the two approaches in the case of an exchange between two hosts M_i and M_j joined by a network with a bit rate d_{ij} (bits/s). In order to compute the transmission time (Halsall, 2002), we need to divide the size of the request (bits) by d_{ij} (Figure 4).

Client/server approach:

The client sends requests to the server across the link d_{ij}. Q_{ijk} is the size of the client's request "k," A_{ijk} the size of the server's answer "k" and N is the number of exchanged requests between M_i and M_j. The transmission time is (El Falou & Bourdon, 2006):

$$T_{CS} = \sum_{k=1}^{N} \frac{A_{ijk}}{d_{ij}} + \sum_{k=1}^{N} \frac{Q_{ijk}}{d_{ij}} = \sum_{k=1}^{N} \frac{(A_{ijk} + Q_{ijk})}{d_{ij}} \quad (1)$$

Figure 4. Communication between two hosts



If we suppose that (A_{ij1}= A_{ij2}=… $A_{ijk)}$ and (Q_{ij1}= Q_{ij2}=… Q_{ijk}). The expression (1) becomes:

$$T_{CS} = \frac{N \times (A_{ij} + Q_{ij})}{d_{ij}} \quad (2)$$

MA Approach

The client creates a MA on M_i and sends it toward M_j. At the end of the interaction, the MA turns back to the client on Mi with the result. The transmission time corresponds to the elapsed time between the dispatching of the MA by the client and the reception of the result by the MA, knowing that an agent on a server M_i is composed of two elements: processing part (S) and data part (D_i). The global size of the agent on M_i is S+ D_i (respectively S+ D_j on server M_j). The transmission time is:

$$T_{MA} = \frac{S + D_i}{d_{ij}} + \frac{S + D_j}{d_{ij}} = \frac{2 \times S + D_i + D_j}{d_{ij}} \quad (3)$$

A comparison between the expressions (2) and (3) allows us to say that the more the value of N increases, the more the MA-WS approach is beneficial.

$$\frac{T_{MA}}{T_{CS}} < 1 \Rightarrow \frac{2 \times S + D_i + D_j}{A_{ij} + Q_{ij}} < N \quad (4)$$

A CASE STUDY

We have chosen a typical e-commerce application (Sohn & Yoo, 1998), that of a user organizing a complete holiday trip. We have produced a prototype using the three strategies mentioned. We didn't take into account the security parameters (for buying and selling owners) during the elaboration of our prototype. We have used the Aglets platform for MA implementations (Lang & Oschima, 1998). It is implemented in Java and provides support for mobiles objects and autonomous MA. We have also implemented this application with the Client/server model to evaluate our prototype. The reservation steps are the following (Figure 5):

Step 1: The user sends a MA.

The user wants to organize a complete holiday trip, including flight reservation, hotel booking, car rental, day trip arrangements, and so forth. For these different parts of his holiday, he sends a customized MA to the network that will find the best bargains. Obviously, the MA will have to exchange information to be able to jointly organize the holiday according to the overall requirements of the user (Figure 5-Link 1). After the MA is sent to the network, the user can disconnect from the network.

Step 2: The MA travels and collects data.

The MA travels from host (WS) to host; only the agent parameters and collected data are transferred between the agent platforms (they constitute the personalization and customization of the code). The MA will collect offers from different WS and other information at each platform (Figure 5-Link 2); this kind of interaction illustrates the interaction MA-WS outlined previously.

Step 3: The MA conducts a payment transaction.

The MA will decide or need to conduct a payment transaction nearby a bank. This latter (the bank represented by a WS-b) will then ask the cooperation of other agents in order to verify the client's identity (e.g., his digital signature). After checking on this information, the bank validates the operation and the MA finishes its transaction (Figure 5-Link 3); this kind of interaction illustrates the interaction WS-MA outlined previously.

Step 4: The MA conducts its proper transactions.

At the end of its research, the MA will return to its platform. The data possibly includes transactions conducted by this MA; to guarantee the best offer for the customer, a MA can carry out its own transactions (Figure 5-Link 4 & 5); these transactions are hybrid type interactions. This kind of interaction illustrates the interaction WS- MA and MA-S outlined previously.

Step 5: The user consults the result.

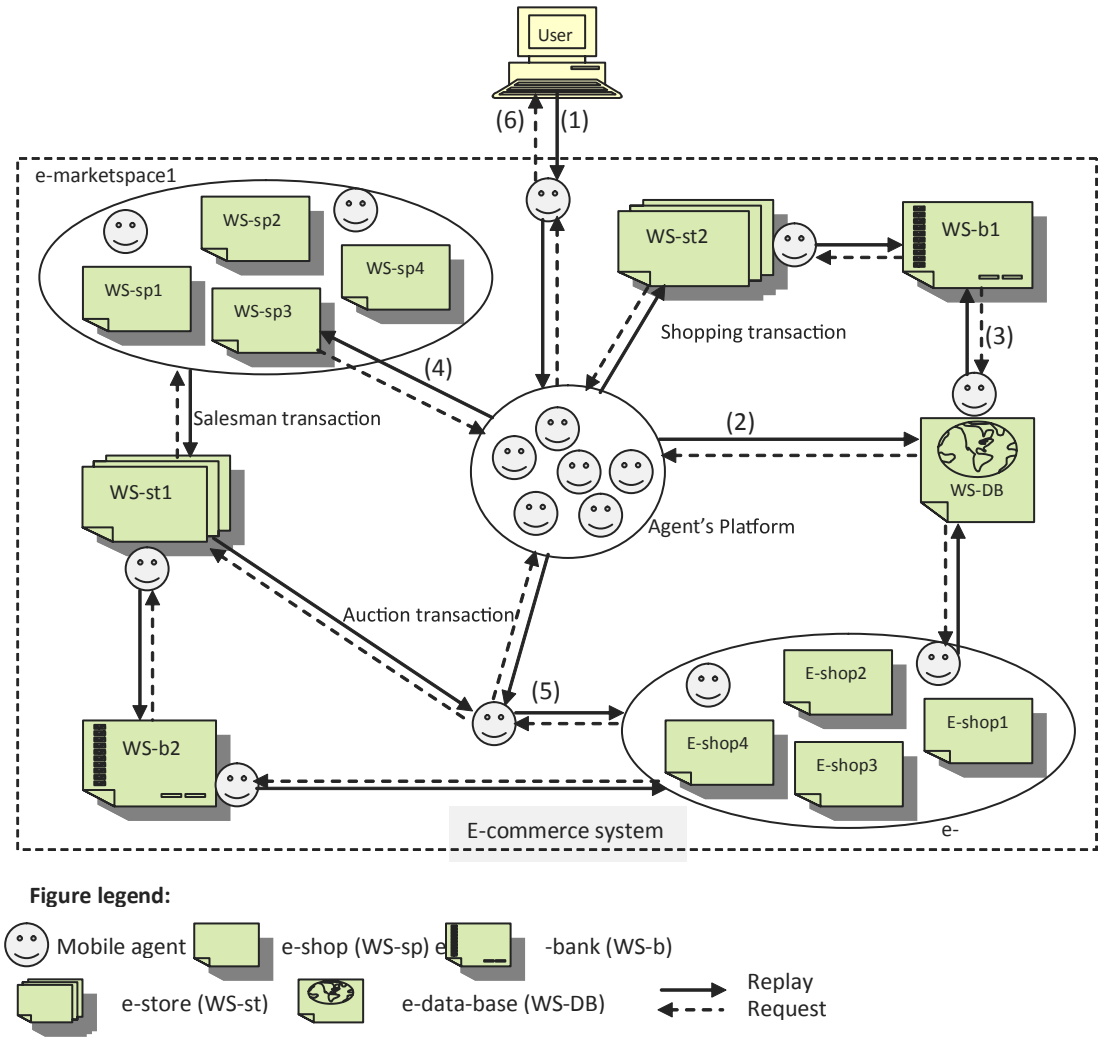
When the users reconnect to the network, they can request the status of their MA and retract the returned agents from the network (Figure 5-Link 6). They can then disconnect again and interpret the collected data of the agents.

Measures and Interpretations

To evaluate this application, we have implemented it with both approaches: MA-WS and Client/server. For our example, we restrict to following steps of the holiday trip:

Holiday trip = flight reservation + hotel booking + car rental.We suppose that the links joining the three hosts have the same bit rate (56 kbits/s) and the global size of the MA is 20 kbits.

Figure 5. E-commerce system



Client/Server Approach

$$\text{Flight reservation} = \frac{A_{ij} + Q_{ij}}{d_{ij}} = \frac{64{,}36 + 52{,}12}{56} = 2{,}08\text{s}$$

$$\text{Hotel booking} = \frac{A_{ik} + Q_{ik}}{d_{ik}} = \frac{117{,}52 + 113{,}20}{56} = 4{,}12\text{s}$$

$$\text{Car rental} = \frac{A_{in} + Q_{in}}{d_{in}} = \frac{36{,}52 + 32{,}36}{56} = 1{,}23\text{s}$$

According to (1), the three steps are fulfilled in 7,43s.

MA Approach

$$\text{Flight reservation} = \frac{S + D_i}{d_{ij}} = \frac{20 + 52{,}24}{56} = 1{,}29\text{s}$$

$$\text{Hotel booking} = \frac{S + D_k}{d_{jk}} = \frac{20 + 117{,}2}{56} = 2{,}45\text{s}$$

$$\text{Car rental} = \frac{S + D_n}{d_{kn}} = \frac{20 + 37{,}68}{56} = 1{,}03\text{s}$$

According to (3), these even steps are fulfilled in 4,77s.

It is clear that our approach is more interesting in comparison with that of the Client/server on the theoretical and experimental plans.

FUTURE TRENDS

In this work, we have compared the two approaches according to the transmission time. In future work, we will study the effectiveness of our approach using other criterias as, for example, propagation time, bandwidth, and so forth. This work is under realization on theoretical as well as experimental plans.

CONCLUSION

In this article, we have proposed the integration of MA-WS in an e-commerce system. In effect, usually e-commerce systems are based on client/server model. Then, we have listed the advantages of this approach in comparison with the client/server model. Also, we have compared both approaches according to the transmission time and we have implemented an application based on both approaches, from which we have calculated the transmission time of each of them. Acquired results confirm theoretical study that we accomplished.

REFERENCES

Adam, R., Dogramaci, O., Gangopadhyay, A., & Yesha, Y. (1999). *Electronic commerce, technical, business and legal issues*. Prentice Hall.

Bernard, G. (2000). Apport des agents mobiles à l'exécution répartie. ISYPAR'00.

Chaves, I., Martins, H., Monteiro, E., & Boavida, F. (2002). A secure e-commerce platform to enable the worldwide use of standards. In *Proceedings of the 1er Congreso Iberoamericano de Seguridad Informatica*, Morelia Michoacán, Mexico, (pp. 18-22).

Chaves, I., Simões, R., & Monteiro, E. (2003). *Electronic delivery under a secure e-commerce environment. Technolegal aspects of information society and new economy: An overview*. Formatex Information Society Book Series.

El Falou, S., & Bourdon, F. (2004). Programmation répartie et agents mobiles. SETIT.

El Falou, S., & Bourdon, F. (2006). Agent mobile et recherche d'information sur le Web: Une solution basée sur le MDP. RFIA.

Guttman, R., Moukas, A., & Maes, P. (1998). Agent-mediated electronic commerce: A survey. *Knowledge Engineering Review, 13*(2), 143-147.

Halsall, F. (2002). *Data communications computer network and open system* (4th ed.). Addison-Wesley.

Karoui, K., & Ftima, F.B. (2008). Interaction mobile agents—Web services. *Encyclopedia of multimedia technology and networking*. Hershey, PA: IGI Global.

Lange, D.B., & Oschima, M. (1998). *Programming and developing Java mobile agents with aglets*. Addison-Wesley.

Lemahieu, W. (2001). Web service description, advertising and discovery: WSDL and beyond. In J. Vandenbulcke & M. Snoeck (Eds.), *New directions in software engineering*. Leuven University Press.

Maes, P., Guttman, R., & Moukas, A. (1999, March). Agents that buy and sell: Transforming commerce as we know it. *Communications of the ACM*.

Nwana, H., Rosenschein, J., Sandholm, T., Sierra, C., Maes, P., & Guttman, R. (1998). Agent-mediated electronic commerce: Issues, challenges, and some viewpoints. In *Proceedings of the Workshop on Agent Mediated Electronic Truding*, Minneapolis/St Paul, MN, USA, (pp. 189-196). ACM.

Papaioannou, T., & Edwards, J. (1998). Mobile agent technology in support of sales order processing in the virtual enterprise. In *Proceedings of the 3rd IEEE International Conference on Information Technology for Balanced Auromarion Systems in Manufacturing*, Prague, Czech Republic, (pp. 275-288). Kluwer Academic.

Sandholm, T. (1999). Unenforced e-commerce transactions. *IEEE Internet Computing, 1*(6), 47-54.

Shaw, M. (2000). Electronic commerce: State of art. *Handbook on electronic commerce* (chap. 1, pp. 3-24). Berlin: Springer-Verlag.

Sohn, S., & Yoo, K.J. (1998). An architecture of electronic market applying mobile agent technology. In *Proceedings of the 4th IEEE Symposium on Computers and Communications*, Athens, Greece, (pp. 359-364). IEEE Computer Society.

WSAP. (2000). *Web service activity proposal*. Retrieved May 28, 2008, from http://www.w3.org/2001/10/ws-activity.html

KEY TERMS

Aglets: It is a java-based mobile agent platform and library for building mobile agents-based applications.

Client/Server: It is a distributed computing model in which client applications request services from server. Clients and servers typically run on different computers interconnected by a computer network.

Distributed Application: It is an application composed of distinct components running in separate runtime environments, usually on different platforms connected via a network.

E-Commerce: It is the buying and selling of goods and services on the Internet, especially the World Wide Web.

Hybrid Interaction: It is a mixture of MA-WS interactions and WS-MA interactions.

MA-WS Interaction: It is an interaction in which a mobile agent invokes a Web service for a request execution

Mobile Agent: It is a mobile software entity that can migrate from one host to another in order to satisfy client requests.

Web Service: It is a paradigm that allows interaction between distant applications via Internet independently of their platforms and languages.

WS-MA Interaction: It is an interaction in which a Web service invokes a mobile agent for a request execution.

Effects of Extrinsic Rewards on Knowledge Sharing Initiatives

Gee Woo Bock
Sungkyunkwan University, Korea

Chen Way Siew
IBM Consulting Services, Singapore

Youn Jung Kang
Sungkyunkwan University, Korea

INTRODUCTION

In the field of motivation, incentives are seen as a means of motivating people. Incentives are usually applied in the form of a scheme, such as piece-rate and fixed-rate monetary rewards. Since the field of knowledge management involves a certain measure of motivation, a number of organizations have used incentives to encourage their employees to share knowledge. Research to date concerning the role of incentives in knowledge sharing seems to contradict one another. Furthermore, when an incentive is sufficiently large, some individuals are inspired to increase their performance to reflect the incentive received (London & Oldham, 1976).

Along with this negative disposition, intrinsically motivated individuals would experience a deterioration of such motivation due to the introduction of incentives, thus jeopardizing the whole knowledge sharing initiative (Deci, Koestner, & Ryan, 1999; Jordan, 1986).

Some research (Bock & Kim, 2002; O'Dell & Grayson, 1998) has suggested a trigger effect that comes from implementing incentives. Empirical evidence concerning the long-term effects of incentives in the field of knowledge sharing is also lacking (Fossum, 1979; O'Dell & Grayson). This research seeks to consolidate the many different views of past research, investigating areas that are lacking. Is it possible to consolidate the different views of incentives in knowledge sharing? Are there differences between having fixed-rate, piece-rate, or no incentive schemes in knowledge sharing initiatives? Do incentives exhibit a triggering effect in motivating individuals to share their knowledge? Would the removal of incentives after the trigger period affect a knowledge sharing initiative? Will the continual increase of incentives remain effective in the long term for knowledge sharing initiatives? These research questions will be answered as the article progresses.

BACKGROUND

This research into the effects of extrinsic rewards on knowledge sharing initiatives encompasses a number of constructs. These constructs were grouped into three sections—knowledge sharing, the introduction of incentives in knowledge sharing, and overcoming past research limitations—and are as described below.

Knowledge Sharing

Knowledge is defined to be a justified belief that enables effective action through the increase of an entity's capacity (Nonaka, 1994). It is considered to be a vital part of an organization's resources. In the resource-based view (Barney, 1991), resources that are valuable, are rare, lack substitutes, and are imperfectly imitable, such as knowledge, offer a source of sustained competitive advantage. In order for an organization to exploit its knowledge, there is a need for the management of knowledge. According to von Krogh (1998), knowledge management is the process of identifying and leveraging the knowledge within an organization so as to help maintain its competitiveness.

Organizations are able to manage knowledge through the use of specialized information systems: knowledge management systems. Knowledge management systems are also referred to as knowledge repositories, shared knowledge bases, or knowledge-based systems, and can include bulletin-board systems (BBS) as well as online forums that archive users' posts.

In the field of knowledge management, the process of transferring knowledge (i.e., knowledge sharing) is considered to be of utmost important. Knowledge sharing is defined as the voluntary process of transferring or disseminating knowledge from one person to another person

or group in an organization (Nelson & Cooprider, 1996). If there were no knowledge transfer activities, the field of knowledge management would not exist. For the purpose of this research, the process of knowledge sharing is taken from the viewpoint where it is empowered by technology through the use of knowledge management systems.

Incentives in Knowledge Sharing

In the knowledge sharing field, incentives are used as a means to an end, easing individuals into parting with their knowledge (Ba, Stallaert, & Whinston, 2001; von Krogh, 1998). There exist costs in the preparation of knowledge for sharing purposes and individuals may not share unless they are duly compensated. Should the benefit exceed the cost, individuals will share knowledge (Constant, Keisler, & Sproull, 1994). These costs arise from lost work time, reduced power and influence, as well as the extra effort needed to articulate knowledge into a comprehendible form. Incentive schemes are the means implemented in organizations to compensate for these costs.

Incentives are usually administered in the form of a structured scheme commonly known as an incentive scheme. Schemes are structured according to the needs of the organization, guided by the purpose for which it would be used and the personnel it is directed at (Jennergren, 1980). The functions of incentives, in addition to inciting action, include affecting the individual's goals and intentions, suggesting to varying degrees goals or intentions, and aiding in the ensuring of an individual's commitment to various goals (Dobmeyer, 1972).

For the purpose of this research, extrinsic rewards would follow the definition provided by Deci et al. (1999) and would specifically imply monetary rewards. Monetary rewards are able to trigger action because "it can provide outcomes that satisfy physiological and psychological needs" (Stajkovic & Luthans, 2001, p. 581). When it comes to extrinsic rewards, a number of researchers have either found a negligible or negative relation between incentives and knowledge sharing. Although individuals interviewed by Bock and Kim (2002) prior to conducting a survey seem to place an emphasis on extrinsic rewards, the result of the study found a negligible relation between rewards and knowledge sharing activities. They justified this result based on motivation literature such as that of Herzberg (1968). Extrinsic rewards do not motivate, but move. Individuals move to avoid the punitive effects from both extrinsic rewards and punishment of the carrot-and-stick philosophy. Once the work environment changes and the carrots are no longer desirable, extrinsic rewards would lose its effectiveness (Levinson, 1973).

In fact, most experienced employees regard knowledge sharing as part of their work responsibilities, and thus hold a negative perception toward the introduction of extrinsic rewards (Constant et al., 1994). The presence of extrinsic rewards can attract nonintrinsically motivated individuals to participate in knowledge sharing (Davenport, Prusak, & Wilson, 2003). The presence of such individuals could prove disruptive to the knowledge sharing initiative as they would share knowledge of low or no quality for the sole purpose of attaining the reward. Alongside this, extrinsic rewards would simultaneously decrease the motivation of individuals who are intrinsically motivated (Deci et al., 1999; Jordan, 1986). The presence of extrinsic rewards would change their focus to that of the reward (Kerr, 1999). They would dispense all of their efforts in pursuit of the rewards, thus affecting their perception of the task at hand (Kreps, 1997; Meyer, 1975; Pfeffer, 1998). A negative perception of the task arises because if they have to be bribed to perform it, the task must be something that they would not otherwise perform (Kohn, 1993).

Overcoming Past Research Limitations

Past research with regard to extrinsic rewards and knowledge sharing were limited in a number of ways. Bock and Kim (2002) and O'Dell and Grayson (1998) also made mention of the possibility of extrinsic rewards having a triggering effect on knowledge sharing initiatives. Empirical evidence is also lacking when it comes to the long-term usage of incentive schemes. O'Dell and Grayson mention that "explicit rewards and incentives go only so far" (p. 168) while Fossum (1979) mentions "reward receipt did not lead to higher performance in subsequent periods, whether appropriately or inappropriately administered" (p. 586). The types of incentive schemes used are frequently generalized, failing to differentiate between the types of schemes: piece-rate, fixed-rate, as well as the absence of an incentive scheme. Piece-rate monetary incentive is defined as paying individuals for each unit produced predetermined amounts of money (Stajkovic & Luthans, 2001), while fixed-rate incentive is the fixed payment of a predetermined sum to individuals for their participation in a task (London & Oldham, 1976).

In order to address these limitations—view disparity, trigger effect, long-term effects, and incentive-schemes differentiation—four research questions were synthesized to guide this research.

- Is it possible to consolidate the different views of incentives in knowledge sharing? This question seeks to address the disparity in views.
- Are there differences between having fixed-rate, piece-rate, or no incentive schemes in knowledge sharing initiatives? Here we address the lack of comparison in this area.
- Do incentives exhibit a triggering effect in motivating individuals to share their knowledge? We want to

validate this trigger nature of incentives with empirical evidence.

- Will the continual increase of incentives remain effective in the long term for knowledge sharing initiatives? This question is asked to investigate the effectiveness of long-term incentive schemes.

METHODOLOGY

In order to investigate the above research questions, six groups of participants are classified. This can be simplified into a 2x3 matrix research design, as shown in Table 1. As its name implies, fixed-rate rewards do not have increasing or decreasing amounts throughout the experiment.

Research Hypotheses

London and Oldham (1976) found that a group provided with fixed-rate extrinsic rewards actually performed worse than that of a group without rewards. A group that does not receive any extrinsic rewards would be intrinsically motivated. With regard to this experiment, the participants from Experimental Group A were informed about their fixed-rate rewards and were expected to perform better than the control group throughout the experiment.

On the other hand, a piece-rate incentive scheme would provide individuals with a goal to work toward. The amount of rewards that an individual receives corresponds to the amount of knowledge shared. Fossum (1979) supported this view when he found that there was no performance difference between a group with piece-rate and a group without extrinsic rewards in the long term. With regard to this experiment, the participants from Experimental Group B were informed about their piece-rate rewards and their performance increase was expected to trickle off toward the experiment's end.

Kohn (1993) as well as Wasko and Faraj (2000) found that extrinsic rewards do not secure anything but temporary compliance from individuals. When individuals comply in the short run, it was able to help trigger knowledge sharing activities. This trigger effect has been suggested in the work of Bock and Kim (2002) as well as O'Dell and Grayson (1998). With regard to this experiment, the participants from Experimental Group B, C, and E are expected to show a performance increase over the control group in the initial stages of the experiment due to the triggering effect of extrinsic rewards.

Herzberg (1968) mentions that it is a myth to think that increasing rewards would offset the long-term negative effects. Coupled with the trigger effects mentioned earlier, even the continuous increase of rewards would only kick-start the initiative. Continuously increasing rewards would not only falter just as with any other long-term incentive schemes, but is also not cost effective. With regard to this experiment, the participants from Experimental Group E would initially perform better than the control group, but increasing rewards do not help in the long term.

The hypotheses for this study are as a shown in Table 2.

Experiment

In order to empirically test the above-mentioned hypotheses, an experiment was conducted. Experimental research was chosen because the research questions sought to observe the changes in knowledge sharing behavior through the manipulation of extrinsic rewards.

The IVLE (Integrated Virtual Learning Environment) online forum, the medium that facilitates students' knowledge sharing activities in the School of Computing (SoC) of the National University of Singapore (NUS), allowed students to share their knowledge with ease, and all knowledge shared would be archived for the whole semester. A module in programming was chosen as the platform to conduct the experiment because the likelihood of students sharing their knowledge on the IVLE online forum is high. Participants were randomly assigned to one of five different groups that

Table 1. Experiment's groups and 2 x 3 matrix research design

Trends / Scheme	**Constant**	**Increasing**	**Decreasing**
Fixed-Rate	Experimental group A. Fixed-rate rewards	Not Applicable	
Piece-Rate	Experimental group B. Piece-rate rewards	Experimental group D. Increasing piece-rate rewards	Experimental group C. A period of piece-rate rewards followed by a period of no rewards

Table 2. Hypotheses

Hypotheses
H1: Fixed-rate extrinsic rewards do not significantly affect knowledge sharing activities.
H2: Piece-rate extrinsic rewards do not significantly affect knowledge sharing activities in the long term.
H3: Extrinsic rewards exhibit a triggering effect towards knowledge sharing activities.
H4: Increasing extrinsic rewards above a certain threshold would no longer be effective in influencing knowledge sharing activities

consisted of one control group and four experimental groups. Depending on the group that they were in, participants were given varying amounts of extrinsic rewards, and the amount of knowledge sharing that occurred was monitored. Senior students who have previously taken this module were asked about their opinions regarding the amount of rewards that would spur them on to share knowledge. For fixed-rate rewards, the amount of $5.00 was estimated to be half the average amount that students receiving piece-rate rewards will achieve. As for the group with increasing rewards, the piece-rate reward of $1.00 was recorded for a period longer than 2 weeks to place emphasis on the middle amount paid to the students.

Data Collection

Depending on which group students were in, they were individually informed by e-mail at the appropriate time that they would be rewarded for shared knowledge. The knowledge shared by students is classified into six categories, of which only two would be rewarded. First and foremost, knowledge is classified into explicit, implicit, or a combination of both as listed in Table 4. Second, knowledge is categorized as relevant or irrelevant. The two categories of knowledge that will be rewarded are the relevant implicit knowledge and the combination of explicit and implicit knowledge that is relevant, as shown in Table 3. Sample knowledge from all the relevant categories and a single irrelevant piece of knowledge are listed in Table 4.

Data Analysis and Result

As can be seen from Figure 1, the graph pattern of the control group seemed to mainly hover below 5. It starts off slow and experiences multiple peaks and dips as it moves along, reaching its highest point at the sixth week.

Experimental Group A (fixed-rate-rewards group) seemed to initially follow closely the pattern of the control group, albeit at a higher level. Experimental Group B (piece-rate-rewards group) hovered close to the level of the control group but followed a different pattern. Experimental Group D (increasing-piece-rate-rewards group) seemed to oppose the pattern of the control group for the first four weeks.

The control group started their knowledge sharing activities later than the other groups. Experimental Group A (fixed-rate-rewards group) actually started well, surpassing the levels of the control group as well as the piece-rate-rewards group. Experimental Group B (piece-rate-rewards group) started higher than both the control group as well as the deferred-piece-rate-rewards group, but was comparatively much lower than that of other groups. Experimental Group C (initial-piece-rate-rewards group) had continuously increasing amounts of shared knowledge up until the third week. Finally, although Experimental Group E (increasing-piece-rate rewards group) started off with the lowest amount of piece-rate rewards, it is the group with the highest starting point.

From the data analyzed, it was found that all hypotheses were supported. Although Figure 1 showed that the fixed-rate group shared much knowledge initially before tapering off, statistical results showed otherwise. Fixed-rate rewards

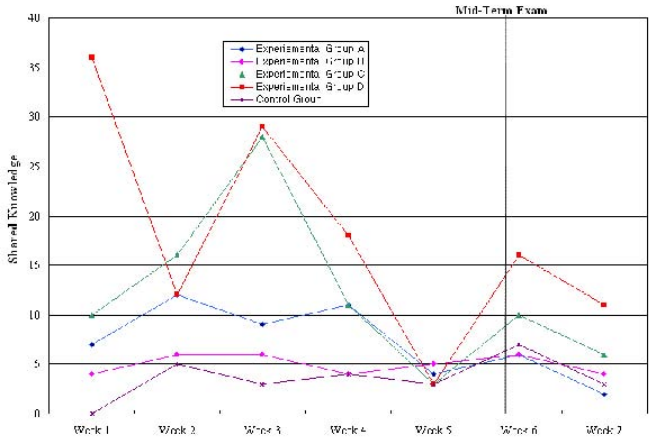
Table 3. Categories of knowledge

<table>
<tr><th>Type
Classification</th><th>Explicit</th><th>Explicit
+ Implicit</th><th>Implicit</th></tr>
<tr><td>Relevant</td><td rowspan="2">Not Reward</td><td colspan="2">Rewarded</td></tr>
<tr><td>Irrelevant</td><td colspan="2"></td></tr>
</table>

Table 4. Table sample knowledge

Classification	Sample Knowledge
Relevant Explicit + Implicit Knowledge	... This is what you can do... [Replace] all **String concatenations by StringBuffer operations.** (Remember the tip from the Resources page in the lab website) and we can use StringBuffer [right]...<Lab submission timings>
Relevant Explicit Knowledge	(test run information)<Lab submission timings>
Relevant Explicit Knowledge	...[Now] you can use binary search...[Find] a match 03... [Then] traverse either up or down to the first similar match or the last similar match respectively and print all the matches... [So]now you don't have to worry about [whether] the matches are sorted in the order of matric no. [because]they would be...
Irrelevant Knowledge	...[Valentines'] coming...[Thinking] of [what] to buy even more stress... [1] came across this site www.pinkbottles.com ... [They] sell beautiful romantic message bottles...

Figure 1.



did not significantly influence knowledge sharing activities throughout the 7 weeks of observation. Trigger effects seemed to be exhibited by the piece-rate, initial-piece-rate, and increasing-piece-rate groups. Following the trigger period, the initial-piece-rate group dropped in the amount of knowledge shared. A number of those within this group that shared much knowledge said that they were initially spurred on by the extrinsic reward. Since shared knowledge becomes a public good, most people would choose to consume rather than contribute to the shared knowledge base.

From the results of this study, it might seem that extrinsic rewards do not help at all with a knowledge sharing initiative. The future of extrinsic rewards in knowledge sharing may not be as bleak as it seems. Rewards are only able to explain 9% of variances in knowledge sharing activities. Rewarding knowledge sharing activities when the organizational culture does not support it would bring about cynicism among employees.

FUTURE TRENDS

This research has shown that it is possible to reconcile the disparity in past research, having shown that incentives do exhibit both facilitating and inhibiting views: It all depends on how a scheme is implemented. Continued research along these lines would bring about meaningful results that could easily be ported for practitioners' usage. This study has shown that extrinsic rewards do exhibit a triggering effect, and this can be exploited to aid in successful knowledge sharing initiatives.

When implementing an incentive scheme, organizations should be aware of the type of schemes used. One should utilize a piece-rate scheme, coupling it with proper organizational norms to aid in achieving a successful knowledge sharing initiative. Recognize that the continual increase of extrinsic rewards does not work if the scheme is used by itself. In addition to not being cost effective, it could fairly well jeopardize the whole initiative.

CONCLUSION

This study set out to build on the foundations of past research, fusing together the differing views of past literature. Based on the results shown, it is believed that this study has achieved this goal. Incentives exhibit both facilitating and inhibiting nature. What is needed is for organizations to utilize them so as to maximize the facilitating and minimize the inhibiting factors. It is important for organizations to properly design a performance-dependent incentive scheme that appropriately rewards employees. This scheme has to be coupled with an organizational culture that emphasizes the collective wellness of the organization. Do not allow an incentive scheme to proceed, especially for the long term if a knowledge sharing culture is not initially embedded in the organization. This would only result in detrimental effects.

REFERENCES

Ba, S., Stallaert, J., & Whinston, A. B. (2001). Introducing a third dimension in information systems design: The case for incentive alignment. *Information Systems Research, 12*(3), 225-239.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99-120.

Bock, G. W., & Kim, Y. G. (2002). Breaking the myths of rewards: An exploratory study of attitudes about knowledge sharing. *Information Resources Management Journal, 15*(2), 14-21.

Constant, D., Keisler, S., & Sproull, L. (1994). What's mine is ours, or is it? A study of attitudes about information sharing. *Information Systems Research, 5*(4), 400-421.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge.* Boston: Harvard Business School Press.

Davenport, T. H., Prusak, L., & Wilson, H. J. (2003). Who's bringing you hot ideas and how are you responding? *Harvard Business Review, 81*(2), 58-64.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). The undermining effect is a reality after all: Extrinsic rewards, task interest, and self-determination: Reply to Eisenberger, Pierce, and Cameron (1999) and Lepper, Henderlong, and Gingras (1999). *Psychological Bulletin, 125*(6), 692-700.

Dobmeyer, T. W. (1972). A critique of Edwin Locke's theory of task motivation and incentives. In H. L. Toci, R. J. House, & M. D. Dunnette (Eds.), *Managerial motivation and compensation.* East Lansing, MI: MSU Business Studies.

Fossum, J. A. (1979). The effects of positively and negatively contingent rewards and individual differences on performance, satisfaction, and expectations. *Academy of Management Journal, 22*(3), 577-589.

Herzberg, F. (1968). One more time: How do you motivate employees? *Harvard Business Review, 46*(1), 53-62.

Jennergren, L. P. (1980). On the design of incentives in business firms: A survey of some research. *Management Science, 26*(2), 180-201.

Jordan, P. C. (1986). Effects of an extrinsic reward on intrinsic motivation: A field experiment. *Academy of Management Journal, 29*(2), 405-412.

Kohn, A. (1993). Why incentive plans cannot work. *Harvard Business Review, 71*(5), 54-63.

London, M., & Oldham, G. R. (1976). Effects of varying goal types and incentive systems on performance and satisfaction. *Academy of Management Journal, 19*(4), 537-546.

Nelson, K. M., & Cooprider, J. G. (1996). The contribution of shared knowledge to IS group performance. *MIS Quarterly, 4*, 409-429.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

O'Dell, C., & Grayson, C. J. (1998). If only we knew what we know: Identification and transfer of internal best practices. *California Management Review, 40*(3), 154-174.

Stajkovic, A. D., & Luthans, F. (2001). Differential effects of incentive motivators on work performance. *Academy of Management Journal, 4*(3), 580-590.

von Krogh, G. (1998). Care in knowledge creation. *California Management Review, 40*(3), 133-153.

Wasko, M. M., & Faraj, S. (2000). "It is what one does": Why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems, 9*(2-3), 155-173.

KEY TERMS

Explicit Knowledge: It is knowledge that has been captured and codified into manuals, procedures, and rules, and is easy to disseminate.

Fixed-Rate Incentive: It is a scheme that pays individuals predetermined amounts of money for each unit produced.

Implicit Knowledge: It is knowledge that can be expressed in verbal, symbolic, or written form but has yet to be expressed.

Incentive or Rewards: They are events or objects external to the individual that can incite action.

Irrelevant Knowledge: Knowledge that is erroneous and incomplete, unrelated to the course work, does not answer the question of the knowledge seeker, and was previously shared by another student.

Knowledge Management Systems: These are a class of information systems developed to support and enhance the organizational processes of knowledge creation, storage and retrieval, transfer, and application.

Knowledge Sharing: It is the voluntary process of transferring or disseminating knowledge from one person to another person or group in an organization.

Piece-Rate Incentive: It is a scheme that pays a predetermined sum to individuals for their participation in a task.

Relevant Knowledge: It is knowledge that is correct and complete, related to the course work of students, answers the question posed by another student, and was not previously shared before by another student.

Efficient Multirate Filtering

Ljiljana D. Milić
University of Belgrade, Serbia

INTRODUCTION

A *multirate filter* can be defined as a digital filter in which the input data rate is changed in one or more intermediate points. With the efficient multirate approach, computations are evaluated at the lowest possible sampling rate, thus improving the computational efficiency, increasing the computation speed, and lowering the power consumption. Multirate filters are of essential importance for communications, image processing, digital audio, and multimedia. The role of multirate filtering in modern signal processing systems is threefold: Firstly, they are used whenever there is a need to preserve the signal properties when connecting two systems operating at different sampling rates. Secondly, multirate techniques are used for constructing filters with stringent spectral constraints that are very difficult, even impossible, to be solved otherwise. Thirdly, multirate filters are used in constructing multirate filter banks.

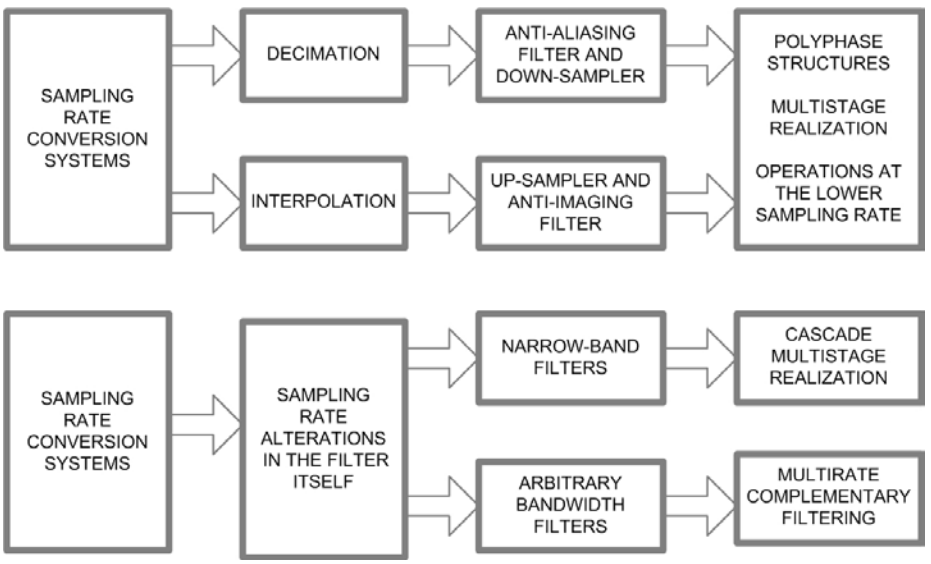
BACKGROUND

Efficient multirate filtering techniques have been developed during the past three decades for implementation of digital filters with stringent spectral constraints (Ansari & Liu, 1993; Bellanger, 1984, 1989; Crochiere & Rabiner, 1981, 1983; DeFata, Lucas & Hodgkiss, 1998; Fliege, 1994; Harris, 2004; Hentchel, 2002; Milić & Lutovac, 2002; Milić, Saramäki & Bregović, 2006; Mitra, 2006; Proakis & Manolakis, 1996; Vaidyanathan, 1990, 1993; Zelniker & Taylor, 1994).

MULTIRATE FILTERING TECHNIQUES

Multirate filtering is one of the best approaches for solving complex filtering problems when a single filter operating at a fixed sampling rate is of a very high order. With a multirate filter, the number of arithmetic operations per second

Figure 1. An overview of multirate filtering techniques



is considerably reduced. The multirate technique is used in filters for sampling rate conversion where the input and output rates are different, and also in constructing filters with equal input and output rates. For multirate filters, FIR (finite impulse response) or IIR (infinite impulse response) transfer functions can be used. An FIR filter easily achieves a strictly linear phase response, but requires a larger number of operations per output sample when compared with an equal magnitude response IIR filter. Multirate techniques significantly improve the efficiency of FIR filters that makes them very desirable in practice.

Figure 1 depicts an overview of different multirate filtering techniques.

Polyphase Realization

Polyphase realization is used to provide an efficient implementation of multirate filters. A *polyphase structure* is obtained when an Nth order filter transfer function is decomposed into M polyphase components, M<N. For FIR filters, polyphase decomposition is obtained simply by inspection of the transfer function (Crochiere, & Rabiner, 1983; Fliege, 1994; Harris, 2004; Mitra, 2006; Proakis & Manolakis 1996; Vaidyanathan, 1993). For multirate IIR filters, several approaches to polyphase decomposition have been developed (Bellanger, Bonnerot & Coudreuse, 1976; Crochiere, & Rabiner 1983; Drews & Gaszi, 1986; Krukowski & Kale, 2003; Renfors & Saramäki, 1987; Russel, 2000).

Multirate Filters for Sampling Rate Conversion

Filters are used in decimation to suppress aliasing, and in interpolation to remove imaging. The performance of the system for sampling rate conversion is mainly determined by filter characteristics. Since an ideal frequency response cannot be achieved, the choice of an appropriate specification is the first step in filter design.

Reducing the sampling rate by a factor of M is achieved by omitting every M-1 sample, or equivalently keeping every Mth sample. This operation is called down–sampling. In order to avoid aliasing, a low-pass *anti-aliasing filter* before down-sampling is needed. Therefore, a *decimator* is a cascade of an anti-aliasing filter and a down-sampler. To increase the sampling rate (interpolation by factor L), L-1 zeros are inserted between every two samples (up-sampling). An interpolation filter has to be used to prevent imaging in the frequency band above the low-pass cutoff frequency. An *interpolator* is a cascade of an up-sampler and an *anti-imaging* filter.

The efficiency of FIR filters for sampling rate conversion is significantly improved using the *polyphase realization*. Filtering is embedded in the decimation/interpolation process and a *polyphase structure* is used to simultaneously achieve the interpolation/decimation by a given factor but running at a low data rate.

Due to the *polyphase multirate implementation*, the number of arithmetic operations in linear-phase FIR filters is decreased by a factor M (or L). An effective method, which leads to high efficiency for a high-order FIR filter is proposed in Muramatsu & Kiya (1997). Efficient decimation and interpolation for the factor M=2 (L=2) is achieved with FIR half-band filters since the number of constants is a half of the filter length.

Very sharp filters with reduced computational efficiency can be achieved by combining the multirate approach and frequency response masking techniques (Lim & Yang, 2005).

Polyphase IIR filters require lower computation rates among the known decimators and interpolators (Renfors & Saramäki, 1987). If a strictly linear phase characteristic is not requested, an IIR filter is an adequate choice. Moreover, an IIR transfer function can be designed to approximate a linear phase in the pass-band (Jaworski & Saramäki, 1994; Lawson, 1994; Surmo-Aho & Saramäki, 1999). An IIR decimator or interpolator is particularly useful in applications that cannot tolerate a considerably large delay of an adequate FIR decimator or interpolator. For a restricted class of filter specifications, an attractive solution based on all-pass subfilters can be used leading to very efficient implementation (Krukowski & Kale, 2003; Renfors & Saramäki, 1987). The most attractive solution is an IIR half-band filter implemented with two all-pass subfilters (Johansson & Wanhammar, 1999; Krukowski & Kale, 2003; Milić & Lutovac, 2002; Renfors & Saramäki, 1987). For a rational conversion factor L/M a very efficient decomposition of IIR filter is proposed in Russel (2000).

There are many applications requiring the sampling rate conversion between arbitrary sampling rates. The sampling rate alteration by an arbitrary factor can be viewed as the computation of the new sample values at arbitrary time instants between the existing samples. There are many valuable contributions in the literature, which concentrate on the various solutions of this important problem (Harris, 2004; Mitra, 2006; Vesma & Saramäki, 2007).

Multirate Filters with Equal Input and Output Rates

Digital filters with sharp transition bands are difficult, sometimes impossible, to be implemented using conventional structures. A serious problem with a sharp FIR filter is its complexity. The FIR filter length is inversely proportional to transition–width and complexity becomes prohibitively high for sharp filters (Crochiere & Rabiner, 1983; Fliege, 1994; Mitra, 2006; Proakis & Manolakis, 1996; Saramäki, 1993; Vaidyanathan, 1993). In a very long FIR filter, the

finite word-length effects produce a significant derogation of the filtering characteristics in fixed-point implementation (Mitra, 2006). IIR filters with sharp transition bands suffer from extremely high sensitivities of transfer function poles that make them inconvenient for fixed-point implementation (Lutovac, Tošić & Evans, 2000). In many practical cases, the multirate approach is the only solution that could be applied for the implementation of a sharp FIR or IIR filter. Thus, to design a multirate narrowband low-pass FIR or IIR filter, a classical time-invariant filter is replaced with three stages consisting of: (1) a low-pass anti-aliasing filter and down-sampler, (2) a low-pass kernel filter, and (3) an up-sampler and low-pass anti-imaging filter (Crochiere & Rabiner, 1983; Fliege, 1994; Milić & Lutovac, 2002; Mitra, 2006). The total number of coefficients in a multirate solution is considerably lower than the number of coefficients of a single rate time invariant filter.

Multistage Filtering

For decimators and interpolators, and for multirate narrowband filters, additional efficiency may be achieved by cascading several stages, each of them consisting of an anti-aliasing filter and down-sampler for decimation and an up-sampler and an anti-imaging filter for interpolation (Fliege, 1994; Milić & Lutovac 2002; Mitra, 2006). Design constraints for subfilters are relaxed if compared to an overall filter. Hence, by using the multistage approach, the total number of coefficients is significantly reduced when compared with the single stage-design. The effects of finite word-length in subfilters are low in comparison with the single-stage overall filter. When a decimation/interpolation factor is expressible as a power-of-two, the application of half-band filters improves the efficiency of the system.

Multirate Complementary Filters

This method can be used in designing filters with any pass-band bandwidth. The multirate techniques are included to reduce the computational complexity. Using the complementary property, the multirate, narrow pass-band filter designs can be used to develop high-pass and low-pass filters with wide pass-bands (Fliege, 1994; Mitra, 2006; Ramstad & Saramäki, 1990). When the output of a low-pass multirate filter is subtracted from the delayed replica of the input signal, the result is a wideband high-pass filter. The delay has to be selected to exactly equal the group delay of the multirate filter. For a low-pass wideband filter the multirate narrowband high-pass filter has to be used.

Efficient FIR filters with an arbitrary bandwidth can be designed using multirate and complementary filtering (Fliege, 1994; Johansson & Wanhammar, 2002; Ramstad & Saramäki, 1990). The overall design is evaluated by cascading complementary multirate filtering two-ports composed of two series branches and one parallel branch. The cascade is terminated with a simple kernel filter. One series branch of the cascade is a *decimator* (filter and down-sampler), while the other is an *interpolator* (up-sampler and filter). The parallel branch is a delay. The most efficient solution is obtained when *half-band filters* are used in the cascade.

Recently, the complementary filtering approach is extended to IIR filters (Johansson, 2003). The overall filter makes use of an IIR filter as a kernel filter, the periodic all-pass filters for constructing complementary pair, and linear phase FIR filters for the sampling rate alterations.

Half-Band Filters

Half-band filters are basic building blocks in multirate systems. A *half–band filter* divides the basis band of a discrete-time system in two equal bands with symmetry properties. The FIR filters are most often used as half-band filters. For a linear-phase FIR half-band filter, half of the constants are zero valued when the filter order is an even number (Mitra, 2006; Saramäki, 1993). A half-band IIR filter can have fewer multipliers than the FIR filter for the same sharp cutoff specification. An IIR elliptic half-band filter when implemented as a parallel connection of two all-pass branches is an efficient solution (Milić & Lutovac, 2002, 2003). The main disadvantage of elliptic IIR filters is their very nonlinear phase response. To overcome the phase distortion one can use optimization to design an IIR filter with an approximate linear phase response (Surma-Aho & Saramäki, 1999), or one can apply the double filtering with the block processing technique for real-time processing (Lutovac & Milić, 2000; Powel & Chau, 1991).

For the appropriate usage of digital filter design software in half-band filter design, it is necessary to calculate the exact relations between the filter design parameters in advance (Milić & Lutovac, 2002, 2003). The accurate FIR half-band filter design methods can be found in Saramäki (1993), Vaidyanathan & Nguen (1987), and Wilsson & Orchard (1999). For the IIR half-band filter design see Milić & Lutovac (2002, 2003), Mitra, (2006), and Schüssler & Stefen (1998).

Complementary Filter Pairs

Complementary filter pairs are used to split the input signal in two adjacent bands, and also are of importance for constructing complex multirate systems and filter banks. A low-pass/high-pass filter pair can be designed to exhibit strictly complementary, all-pass complementary, power complementary, or magnitude complementary properties (Mitra, 2006; Vaidyanathan, 1993). This solution benefits the possibility of implementing the complementary filter pair at the cost of a single FIR (IIR) filter (Fliege, 1994;

Mitra, 2006; Vaidyanathan, 1993). In the most applications, half-band filter pairs with the crossover frequency located in the middle of the base band are used to divide the frequency band in two equal sub-bands. It was shown recently that a *complementary IIR filter pair* with the arbitrary crossover frequency can be easily obtained by simple transformation of the start-up half-band prototype filter pair (Damjanović, Milić, & Saramäki, 2005, Milić & Saramäki, 2003; Milić, Damjanović & Nikolić, 2006). This new filter pair retains the complementary properties of the start-up half-band filter pair.

Multiplierless Solutions

The efficiency of *multirate filters* is significantly improved by simplifying arithmetic operations. This is achieved by replacing a multiplier with a small number of shifters-and-adders. Generally, implementing multiplierless design techniques in subfilters, at the cost of a slight derogation of filtering performances, increases the efficiency of the overall multirate filter.

For instance, one can use the optimization technique (Yli-Kaakinen & Saramäki, 1999 and 2007), multiple constant multiplication (MCM) technique (Dempster & Macleod, 1999), or design based on EMQF (Elliptic Minimal Q-Factors) transfer functions (Milić & Lutovac, 1999, 2003; Lutovac & Milić 2000). It was shown recently (Gustafsson & Dempster, 2004; Gustafsson, Johansson, Johansson & Wanhammer, 2006; Eghbali, Gustafsson, Johansson & Löwenberg, 2007) that multiple constant multiplication (MCM) technique is an efficient way to reduce the computational workload in the polyphase decimators and interpolators. An effective method for designing the multiplierless highly-selective FIR half-band decimators and interpolators was proposed by Saramäki and Yli-Kaakinen (2006). A well-known solution for large conversion factors in decimation is a cascaded integrated comb (CIC) filter, which performs multiplierless filtering (Harris, 2004; Hentchel, 2002; Hogenauer, 1981; Jovanović-Doleček & Mitra, 2005; Mitra, 2006).

Future Trends

The rapid development of new algorithms and new design methods in the area of multirate filtering has been influenced by the advances in computer technology and software development. Although the existing literature on the subject is very large, the multirate filtering is an open area of research. The future trends coincide with the general demand for decreasing the computational complexity, increasing the computational speed, and lowering the power consumption. In the case of multirate filters, those goals can be achieved by following several directions: developing new optimization methods for constructing high-performance multirate filters; integrating the multirate filtering and the frequency-response masking techniques; and developing new methods for simplifying implementation structures of subfilters and of overall multirate systems.

CONCLUSION

The multirate filtering techniques are widely used in sampling rate conversion systems, and for constructing filters with equal input and output sampling rates. Various multirate design techniques provide that the overall filtering characteristic is shared between several simplified subfilters that operate at the lowest possible sampling rates. Design constraints for subfilters are relaxed if compared to a single rate overall filter. Hence, by using the multistage approach, the total number of coefficients is significantly reduced. As a consequence of the reduced design constraints, the effects of quantization (finite word-length effects) in subfilters are decreased. Multirate filters provide a practical solution for digital filters with narrow spectral constraints that are very difficult to solve otherwise.

REFERENCES

Ansari, R. & Liu, B. (1993). Multirate signal processing. In Sanjit. K. Mitra and James F. Kaiser (ed.), *Handbook for Digital Signal Processing*. (pp. 981-1084). John Wiley, PA: Idea Group Publishing.

Bellanger, M.G., Bonnerot, G. & Coudreuse, M. (April, 1976). Digital filtering by polyphase network: application to sample-rate alteration and filter banks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24, 109-114.

Bellanger, M. (1984, 1989). *Digital processing of signals: Theory and practice*. New York: John Wiley & Sons, Inc.

Crochiere, R.E. & Rabiner, L.R. (1981). Interpolation and decimation of digital signals - A Tutorial Review. *Proceedings of the IEEE*, 78, 56-93.

Crochiere, R.E. & Rabiner, L.R. (1983). *Multirate digital signal processing*. Englewood Cliffs: Prentice-Hall, Inc.

Damjanović, S., Milić, L. & Saramäki, T. (2005). Frequency transformations in two-band wavelet IIR filter banks. *Proceedings of the IEEE Region 8 International Conference on "Computer as a Tool"*, EUROCON 2005. 87-90.

DeFata, D.J., Lucas, J.G. & Hodgkiss, W.S. (1988). *Digital signal processing: A system design approach*. New York: John Wiley & Sons, Inc.

Dempster, A.G. & Macleod, M.D. (1995). General algorithms for reduced-adder integer multiplier design. *Electron. Letters*, 31, 261-264.

Drews, W. & Gaszi, L. (1986). A new design method for polyphase filters using all-pass sections. *IEEE Transactions on Circuits and Systems*, 33, 346-348.

Eghbali, A., Gustafsson, O., Johansson, H. & Löwenberg, P. (2007). On the complexity of multiplierless direct and polyphase FIR filter structures. *Proc of the 5th International Symposium on Image and Signal Processing and Analysis, ISPA 2007*, 200-205.

Fliege, N.J. (1994). *Multirate digital signal processing*. New York: John Wiley & Sons, Inc.

Gustafsson, O. & Dempster, A.G., (2004). On the use of multiple constant multiplication in polyphase FIR filters and filter banks. *Proc. Nordic Signal Proc. Symp*. 53-56.

Gustafsson, O., Johansson, K., Johansson, H. & Wanhammar, L. (2006). Implementation of polyphase decomposed FIR filters for interpolation and decimation using multiple constant multiplication techniques. *Proc. 2006 Asia Pacific Conference on Circuits and Systems*. 926-923.

Harris, F. J., (2004). *Multirate signal processing for communication systems*. Upper Saddle River: Prentice Hall PTR.

Hentchel, T. (2002). *Sample rate conversion in software configurable radius*. Morwood, MA: Artech House, Inc.

Hogenauer, E.B. (1981). An economical class of digital filters for decimation and interpolation. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 29, 155-162.

Jaworski, M. & Saramäki, T. (1994). Linear phase IIR filters composed of two parallel all-pass sections. *Proc of IEEE Int. Symposium on Circuits and Systems*, 2, 537-540, London, U. K..

Johansson, H. (2003). Multirate IIR filter structures for arbitrary bandwidth. IEEE Transactions on Circuits and Systems-I: *Fundamental Theory and Applications*, 50, 1515-1529.

Johansson, H. & Wanhammar, L. (1999). High-speed recursive filter structures composed of identical all-pass subfilters for interpolation, decimation, and QMF banks with perfect magnitude reconstruction. *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, 46, 16-28.

Johansson, H. & Wanhammar, L. (2002). Design and implementation of multirate digital filters. In Gordana Jovanović-Doleček, (Ed.), *Multirate Systems: Design & Applications (pp. 257-292)*. Hershey, PA: Idea Group Publishing.

Jovanović-Doleček, G. & Mitra, S.K. (2005). A new two-stage sharpened comb decimator. *IEEE Transactions on Circuits and Systems – I: Regular Papers*, 52, 1416-1420.

Krukowski, A. & Kale, I. (2003). *DSP system design: Complexity reduced IIR filter implementation for practical applications*. Boston: Kluwer Academic Publishers.

Lawson, S.S. (1994). Direct approach to design of PCAS filters with combined gain and phase specification. *IEEE Proceedings of Vision, Image and Signal Processing*, 141, 161-167.

Lim, Y.C. & Yang, R. (2005). On the synthesis of very sharp decimators and interpolators using the frequency-response masking technique. *IEEE Transactions on Signal Processing*, 53, 1387-1397.

Lutovac, M.D. & Milić, L. D. (2000). Approximate linear phase multiplierless IIR half-band filter. *IEEE Signal Processing Letters*, 7, 52-53.

Lutovac, M.D., Tošić, D.V. & Evans, B.L. (2000). *Filter design for signal processing using MATLAB and Mathematica*. Upper Saddle River, New Jersey: Prentice Hall.

Milić, L., Damjanović, S. & Nikolić, M. (2006). Frequency transformations of IIR filters with filter bank applications. *Proc. APCAS 2006 IEEE Asia Pacific Conference on Circuits and Systems*, 1053-156. Singapore.

Milić, L.D. & Lutovac, M.D. (1999). Design of multiplierless elliptic IIR filters with a small quantization error. IEEE Transactions on Signal Processing, 47, 469-479.

Milić, L.D. & Lutovac, M.D. (2002). Efficient multirate filtering. In Gordana Jovanović-Doleček, (ed.), *Multirate Systems: Design & Applications*. (pp. 105-142). Hershey, PA: Idea Group Publishing.

Milić, L.D. & Lutovac, M.D. (2003). Efficient algorithm for the design of high-speed elliptic IIR filters. *International Journal of Electronics and Communications*, 57, 255-262.

Milić, L.D. & Saramäki, T. (2003). Three classes of IIR complementary filter pairs with an adjustable crossover frequency. *Proc. IEEE Int. Symp. Circuits Syst. ISCAS 2003*, 4, 145–148.

Milić, L.D., Saramäki, T. & Bregović, R. (2006). Multirate filters: An overview. *IEEE Asia Pacific Conference on Circuits and Systems APCCAS 2006*, 914 - 917.

Mitra, S.K. (2006) *Digital signal processing: A computer based approach*. (Third edition). New York: The McGraw-Hill Companies, Inc.

Muramatsu, S. & Kiya, H. (1997). Extended overlap-add and save methods for multirate signal processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 45, 2376-2380.

Powel, S. & Chau, M. (1991). A technique for realizing linear phase IIR filters. *IEEE Transactions on Signal Processing,* 39, 2425-2435.

Proakis J.G. & Manolakis D.G. (1996). *Digital signal processing: Principles, algorithms, and applications*. London: Prentice Hall.

Ramstad, T.A. & Saramäki, T. (1990, May). Multistage, multirate FIR Filter structures for narrow transition-band filters. *Proc. 1990 IEEE Int. Symp. Circuits and Systems*. (pp. 2017 – 2021). New Orleans, Louisiana.

Renfors, M. & Saramäki, T. (1987, January). Recursive Nth-band digital filters - Part I: Design and properties. *IEEE Transactions on Circuits and Systems,* 34, 24-39.

Russel, A.I., (2000). Efficient rational sampling rate alteration using IIR filters. *IEEE Signal processing Letters*, 7, 6-7.

Saramäki, T. (1993). Finite impulse response filter design., Chapter 4 in Handbook for Digital Signal Processing. Edited by S. K. Mitra and J. F. Kaiser, John Wiley and Sons, New York, pp. 155 – 277.

Saramäki, T. & Yli-Kaakinen, J. (2006). A novel approach for synthesizing multiplication-free highly-selective FIR half-band decimators and interpolators, *Proc. 2006 Asia Pacific Conference on Circuits and Systems*. 922-925.

Schussler, H.W. & Stefen, P. (1998). Halfband filters and Hilbert transformers. *Circuits Systems Signal Processing*, Vol. 17, No. 2, 137-164.

Surma-Aho, K. & Saramäki, T. (1999, July). A systematic technique for designing approximately linear phase recursive digital filters. IEEE Transactions on Circuits and Systems-II, 46, 956-963.

Vaidyanathan, P.P. (1990). Multirate digital filters, filter banks, polyphase networks, and applications: A Tutorial. *Proceedings of the IEEE*, 78, 56-93.

Vaidyanathan, P.P., (1993). *Multirate systems and filter banks.* Englewood Cliffs, NJ: Prentice Hall.

Vaidyanathan, P.P. & Nguen, T.O. (1987, March). A trick for the design of FIR half-band filters. *IEEE Transactions on Circuits and Systems,* 34, 297-300.

Vesma, J. & Saramäki, T. (2007). Polynomial-based interpolation filters—Part I: Filter synthesis. *Circuits, Systems & Signal Processing*. 26, 115–146.

Wilsson, Jr., A.N. & Orchard, H.J. (1999). A design method for half-band FIR filters. *IEEE Transactions on Circuits and Systems-I: Fundamental Theory and Applications,* 45, 95-101.

Yli-Kaakinen, J. & Saramäki, T. (1999). Design of very low-sensitivity and low-noise recursive filters using a cascade of low-order lattice wave digital filters. *IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing*, 46, 906-914.

Yli-Kaakinen, J. & Saramäki, T. (2007). A systematic algorithm for the design of lattice wave digital filters with short-coefficient word length. *IEEE Transactions on Circuits and Systems-I: Regular Papers*, 54, 1838-1851.

Zelniker, G. & Taylor, F.T. (1994). *Advanced digital signal processing: Theory and application.* New York, NJ: Marcel Dekker.

KEY TERMS

Decimation: Decreasing the sampling rate. Decimation process consists of filtering and down-sampling.

Down-Sampling: Discarding every M-1 samples (retaining every Mth sample).

FIR Filter: A finite impulse response digital filter.

Half-Band Filters: A low-pass or high-pass filter that divides the basis band in two equal bands, and satisfies prescribed symmetry conditions.

IIR Filter: An infinite impulse response digital filter.

Interpolation: Increasing the sampling rate. Interpolation consists of up-sampling and filtering.

Multirate Filter: A digital filter, which changes the input data rate in one or more intermediate points in the filter itself.

Multistage Filtering: Cascade of filters and decimators (interpolators and filters).

Polyphase Decomposition: Decomposition of a transfer function in M (L) polyphase components that provide sequential processing of the input signal at the lower sampling rate.

Up-Sampling: Inserting L-1 zeros between every two samples.

From E-Governance Towards E-Societal Management

Nicolae Costake
Certified Management Consultant, Romania

INTRODUCTION

The purpose of this article is to contribute to the definition of the still emerging strategic concept of e-societal management (e-SM) as a key component of the information society (IS). It is a continuation of articles published in the *Encyclopedia of Digital Government* (Antiroikko, 2006; see also, Costake, 2008a, 2008b) which describe the concepts of:

(1) e-government (e-Gvt) as a set of e-services provided by the public administration (i.e., executive authority) to citizens and organizations, and
(2) e-governance (e-G) as including also the set of e-services provided by the judicial and legislative authorities of the state.

The two quoted chapters contain the historical perspectives mainly between the mid-1990s (e.g., G7 Conference on Information Society, 1995) and the mid 2000s (e.g., the EU's i2010 Program, 2006). The above definition of e-G may be interpreted also as equivalent to the e-SM at the country level and below, provided that the actions of the three authorities converge to assure continuous and sustainable socio-economic development. On the other hand, the new century started with increased global threats and also opportunities (e.g., the development of the IS). Both suggest the importance of the societal management, as well as the difference between *enterprise management,* which aims to achieve performance within a given national and international societal environment, and *societal management,* which aims to assure the societal environment best supporting developments of the economy and civilization.

It follows that the historical development of e-SM is connected to the rather slow evolution of SM and the rapidly developing IS, including e-Gvt and e-G. Its start can be considered the UN General Assembly's Declaration on Computer and Development (UN, 1968), followed by the UN's World Summit for the IS (e.g., WSIS, 2005), which adopted requirements for national and international e-SM (implicitly addressed). E-procurement for public acquisitions, the "Trans-European Administration Network" recommended for the EU by Bangemann et al. (1994), and the start of the Single Euro Payment Area project in 2002 (see SEPA, 2007, which contains complete chronology and content) are examples of other relevant milestones. E-SM is a concept built on those of socio-economic system (SES) and societal management. Many opinions were expressed.

This article begins by sketching a model of the socio-economic system as a foundation; selects, with the risk of being subjective, a number of relevant positions taken by individual institutional and authors; describes e-SM and its associated issues and trends; and proposes conclusions.

BACKGROUND

The general model of the global SES, suggested in Figure 1, includes natural resources (a component of the natural system) and the human activities system (HAS), layered above the natural system. In the absence of the HAS, the natural system was stationary, its biologic components being in a dynamic equilibrium (their growth and levels limited by the available resources). The HAS also transforms data into information, and information into knowledge, thus accumulating experience. It also enforces the concept of ownership, developing the economic activity. Some frequently mentioned risks and issues for the global SES are mentioned in Table 1.

Table 1 suggests that societal management cannot be restricted to the national domain, mankind is faced with possible irreversible dangerous processes, and the permitted duration for decision and action is shortening. The obvious conclusion is the need for e-SM.

The model of a generic SES is presented in Figure 2 as having two subsystems: (1) a societal operational subsystem (SOS) in which social, energy, material, financial, and informational processes based on natural resources and artifacts take place in communities, enterprises (non-financial and financial), markets, NGOs, and other organizations; and (2) a societal management subsystem (SMS), the institutions of which generate regulations and actions necessary for the proper organization, functioning, and development of the SES, including defense of the system's domain, protection of public order and of property, mandatory education, social security, solving conflicts and breaches of regulations (assures the necessary homeostasis), assuring relationships with the external environment, and so forth. The two subsystems are interconnected and connected with the external environment by flows of energy/matter/products/services, financial means, information, and personnel. SES may be basic (such as household), organization, territorial community

Table 1. Frequently mentioned global risks and issues

Risk and/or Issue	Comments and/or Mitigation
Global warming (Gore, 2006; Kyoto 2007)	Measures agreed at the Kyoto conference, but still not implemented by the highly polluting countries
Deterioration of the environment (Gore, 2006)	De-forestation, pollution of water, destruction of the ozone layer, destruction of species
Exhaustion of natural hydrocarbon energy reserves due to the increase of their consumption (Laherrère, 1998)	Implies changing the energy orientation of many present technologies, for example: use of renewable energies, green and zero-waste technologies (Greyson, 2007), increase of the efficiency of energy generators and consumers, and also possibly controlled thermonuclear power generation
Overpopulation: total human population may exceed the capacity to be supported by the natural resources (e.g., P&P, 2007)	Mitigation implies, for example: (a) genetic engineering of biological natural resources (b) industrialized food (c) conversion of sea water into drinking water
Endemic local wars, other armed conflicts (some with religious character), and terrorism	A large amount of resources is spent in weaponry and wars and/or other use of armed forces, the global arsenal being sufficient to destroy life on the planet
Global societal division, by large gaps*: (i) between rich and poor countries; (ii) between rich and poor people; (iii) digital divide (BECTA, 2001)	See also the world model with three economies: natural, developing, and advanced (e.g., Hart, 1997) and applicable theories, for example of the Open Society (e.g., Soros, 1998; Dror, 2002)
Deterioration of human behavior (e.g., Huxley, 1958; Lorenz, 1973)	Naisbit (1984) and Toffler (1970) underlined changes and the need for adaptation to change

** This issue can be also considered as a gap between long-term and general interests against short-term interests oriented on gaining as much as possible, as quickly as possible.*

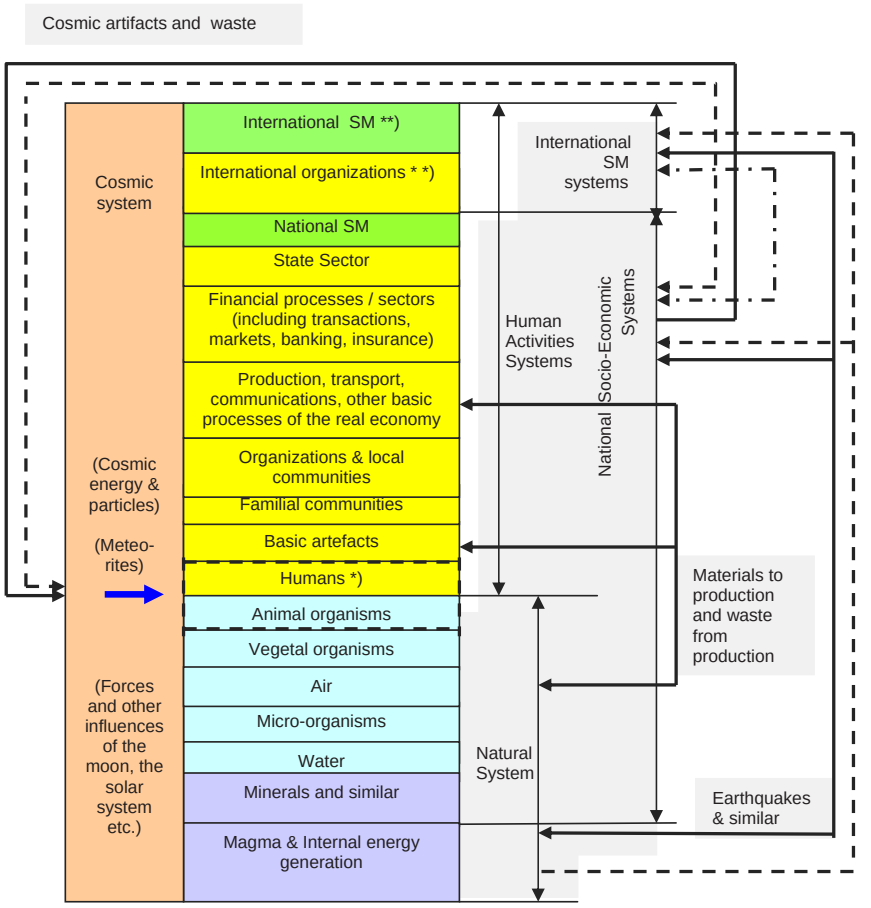
(such as municipality), a group of organizations, national, supranational, international, or global. As the national and supranational SESs raise most SM issues, they are first for discussion in this article.

There is a very large bibliography on SM and e-SM. As a comprehensive referencing is impossible, just a few examples are quoted. They are classified in streams and sub-streams:

Theoretical (some references are already included in Tables 1 and 2)

- *Global Governance:* Biermann (2006) proposes principles, research, and challenges such as "adaptive state." Heylinghen (2007) foresees the future "networked society."
- *SM Theory and Models:* Dror (2002) proposes capacity to govern in condition of global transformations, concluding also of the need for an independent societal feedback. Greer (2005) proposes an explanation why civilizations grow and then collapse due to unsustainable use of resources. Situngkir (2003) proposes an automatic control system analogy with Montesquieu's model of the powers in the state.
- *E-G Research:* E-Gvt (e.g., Lenk & Traunmuller, 2000; Scholl, 2003) => e-G (e.g., Lenk, 2003; Traunmuller & Wimmer, 2004) => Innovative e-G (e.g., Bicking & Wimmer, 2006). This sub-stream also proposes

Figure 1. Layered model of the global SES (Legend: flows of ⟶ matter/energy/products/services, - - → information, -·-·→ financial means)



** Above the level of other animal organisms*
*** Including supranational*

technical targets for 2020 and so on. See also Gupta, Kumar, and Bhattacharya (2005) and Klischewski and Scholl (2006).

Political Studies and Recommendations

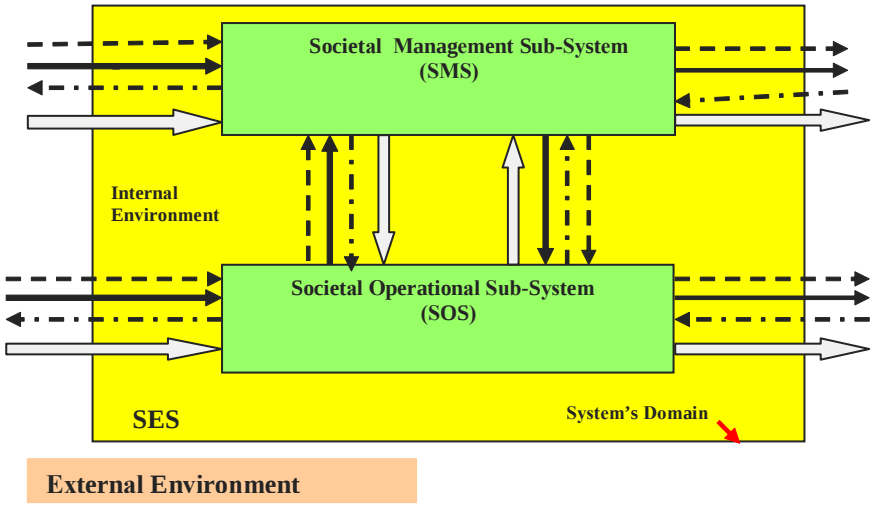
- *Specific Studies by International Organizations:* The OECD (2001) considers that as traditional governance becomes ineffective, the new governance will make changes in allocating and structuring power. EICTA (2004) formulates recommendations to the European Commission regarding the development of the European ICT industry.
- *Internationally Adopted Action Plans:* WSIS and i2010 Action Plans.

Practical, Generating Institutional Changes and/or Newly Implemented Solutions and/or Information Systems

There is vast literature at local and national/federal levels, a very small sample being about Canada and the United States (e-Govt, 2005), China (Lovelock & Urel, 2002), Malaysia (Tahir, 2005), and Singapore (Lee Boon Yang, 2007). The EU has a supra-national SES with all the components of an SMS:

- *A Legislative Authority* produces normative acts, to be mandatorily implemented by the member countries. One example is the draft directive of the infrastructure for spatial information (INSPIRE, 2004).

Figure 2. Generic model of a SES (⟹ flow of personnel; the significances of the other arrows is the same as in Figure 1)



- *The Legislative and the Judicial Authorities* (see Justice, 2007) are supported by the legislative information system and database (EUR-Lex, 2007).
- *An Executive Authority* offers many examples: drafting of European Programs, action plans, workshops, projects, studies, and reports for Development of the Information Society (e.g., IDABC, 2007, 2008). They succeeded to produce EU information systems such as the Schengen border control (e.g., Migrationsverket, 2001), the Single Euro Payment Area (SEPA, 2007), the EU network for implementing European policies (e.g., TESTA, 2005), and the European Interoperability Framework (Gartner, 2007). One recent example is reported by Undheim (2008), and general economic studies were also launched (e.g., MODINIS, 2006). In recent years, assessments of potential benefits from e-SM appeared, for example, approximately 1% GDP from e-procurement for public acquisitions (Timmers, 2005) and approximately 5% GDP from reducing the administrative burden of businesses (Reading, 2007). The French e-administration (ADELE, 2004) is an example of a high-return-on-investments national program. The EU's Executive Authority is supported also by Eurostat, and the Observatories for IT (EITO, 2007) and for e-government (now accessible via IDABC > eGovernment practice).

There are many points of view. Except e-G research, the theoretical works and political studies are mainly sociologically, economically, and politically oriented. Only a few mention the possible contribution of cybernetics (e.g., automatic or artificial closed loops) or the necessary informatic support; they do not consider the importance of the IS for SM. The practical approaches bring tangible results but do not include vision based on a comprehensive socio-economic model. Most do not discuss key performance indicators (KPIs) and do not foresee actions against two sources of losses, particularly in developing and transitional countries: underground economy and corruption.

THE EMERGING E-SOCIETAL MANAGEMENT

e-SM can be defined as ICT-enabled SM, assuring the integration and interoperability of the public information systems for:

- friendly online serving citizens and organizations;
- increasing the performance of the enterprises by supporting e-business and minimizing the administrative burdens;
- maximizing the performance of the socio-economic development of the SES; and
- complying to a set of constraints, such as laws, including free-market competition, democracy, and human rights.

The main dangers are:

- the accumulation of personal power (hence needing

autonomous feedback);

- the bureaucratization (hence needing the reduction of administrative burden including the single entry of an information element into the general public information system);
- the underground economy (hence needing a public finance management-integrated information system);
- corruption (needing, e.g., e-procurement for public acquisitions, specific legislation, plus an information system); and
- poor quality of strategic decisions (hence needing use of ICT including decision support tools and public societal feedback).

There are two direct consequences: a structure of the societal management subsystem and a specific architecture of the e-SM information system. They can be complemented by criteria for e-SM. The resulting necessary structure of the societal management sub-system is represented in Figure 3 as performing four functions (executive authority, judicial authority, legislative authority, and public societal feedback authority). The symbolic person (e.g., president or monarch), the ombudsman (if applicable), the spiritual authority (acting on the population), flows of personnel, and flows exchanged with the international environment were omitted, and other main flows only were drawn, not to complicate excessively the diagram. The legislative authority assures the regulations at the national and local levels. The judicial authority assures the homeostasy of the SES. The executive authority assures the current management (including drafting laws and international accords, implementation/enforcement of law and judicial decisions, provision of services to population and organizations, redistribution of revenues collected or generated by owned resources, etc.). The proposed public societal feedback authority (PSFA) refers to a group of activities usually considered components of the executive authority: e-voting (component of e-democracy), official statistics (generating information formatted as statistical tables and knowledge, i.e., statistical mathematical models and *caeteris paribus* mathematical models), and court of accounts. The PSFA closes the loop for permanent improvement. The resulting necessary high-level architecture of the e-SM information system is described in Figure 4. It is in line with the European best practice in the field (e.g., Vanvelthoven., 2005; IDABC, 2007, 2008). A sample of possible criteria for e-SM is suggested in Table 2, which also summarizes proposals formulated in previous papers (e.g., Costake 2006a).

FUTURE TRENDS AND FURTHER RESEARCH OPPORTUNITIES

The increasing planetary problems can be a key factor for the promotion of e-SM. A number of further research opportunities can be suggested, such as:

Figure 3. Necessary structure of the societal management subsystem (1, 2, 3, 4 are connectors, to avoid the drawing of many lines. The meaning of the arrows is the same as in Figures 1 and 2)

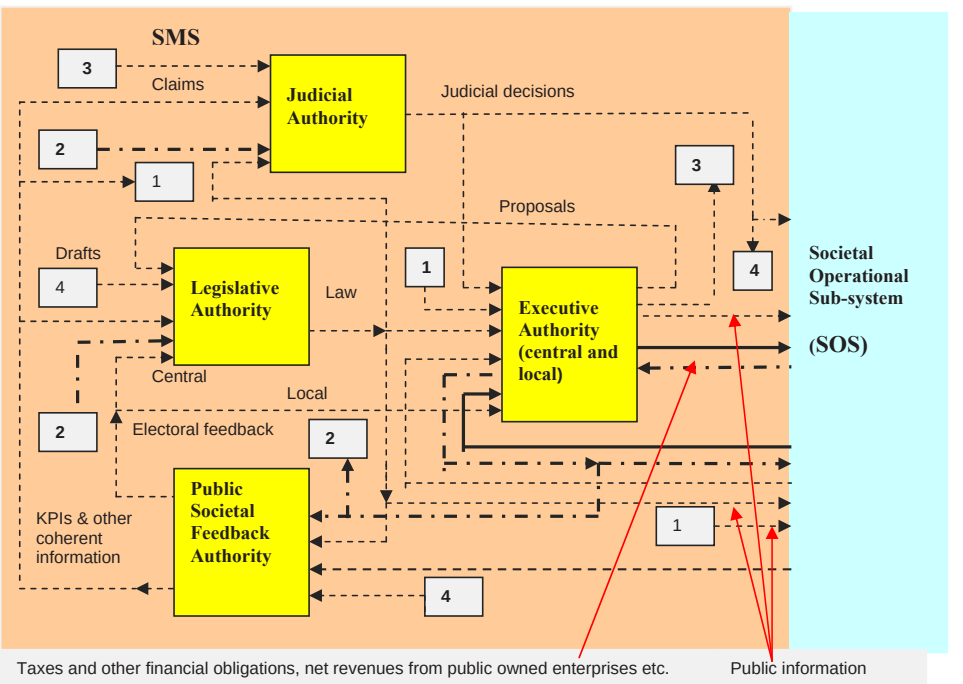
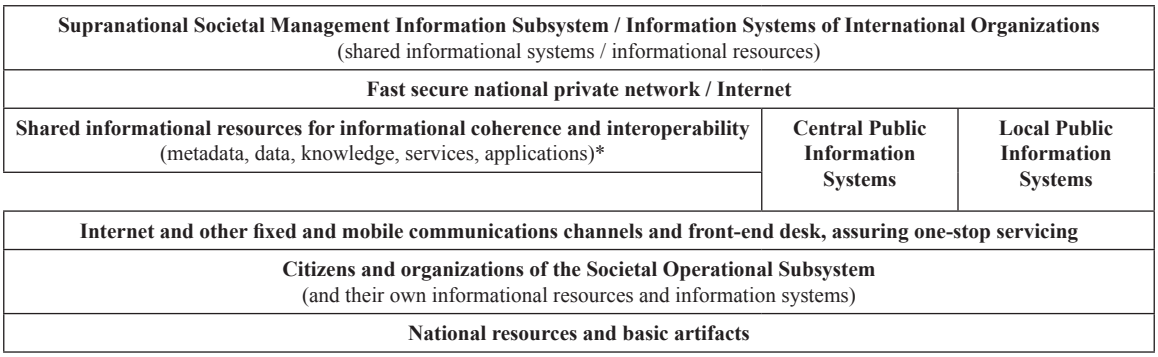


Figure 4. High-level architecture of the e-SM information system with five levels



** Can be hosted by public data centers*

(1) formulating societal management as a scientific discipline complementing the classical enterprise management science;
(2) exhaustively identifying potential sources of costs, benefits, and losses in the SM and e-SM, and solutions for achieving performance in sustainable growth;
(3) selecting the minimal number of KPIs capable to describe the performance of e-SM;
(4) creating blueprints for e-SM information systems at the national/federal and super-national/international levels;
(5) clustering countries and corresponding recommended specific actions at the national, international, and supranational levels;
(6) developing e-SM strategies at the national and supranational levels; and
(7) formulating a system engineering theory of e-SM, including a comprehensive model of the SES at various levels, identifying automatic closed loops and their effects, as well as possible informational artificial (new) virtuous circuits.

CONCLUSION

The following two conclusions are proposed:

(1) E-SM implies:
  (a) technical re-engineering of the activities of the public institutions (see Table 2);
  (b) specific functional structure (see Figure 3) and architecture of the supporting information system (see Figure 4); and
  (c) a new culture of the public sector in which every organizational entity is a component of a subsystem which supports and encourages:
    - development of very friendly e-services to citizens and organizations;
    - the development of the SES to the best medium and long-term interests of its stakeholders; and
    - a set of values, of a set of objectives and constraints, in a responsible and accountable manner, characterized and followed by KPIs.

(2) The economic benefits of e-SM have three sources:
  (a) better SM through better decisions;
  (b) savings (e.g., elimination of redundant collection of data and data processing/storage, economies of scale, replacement of obsolete legacy ICT, etc.); and
  (c) minimization of losses (including underground economy and corruption, in the countries where applicable).

REFERENCES

ADELE. (2004). *Administration electronique.* Retrieved from http://ec.europa.eu/information_society/index_en.htm

Table 2. Criteria for e-SM

Criterion	Examples of Requirements
General	
Conception of the target e-SM	Vision, general objectives, strategic planning, standards, KPIs, cost-benefit analysis, etc. (see note below)
Specific e-SM infrastructure	As shown in Figure 4
Performance of the governance	Minimization of commands, in favor of creating new feedback informational circuits and/or tuning existing automatic closed loops
Legislative Authority	
Methodology	Use of electronic documents and archives, shared legislative database, computerized standard workflow(s) and validation of drafts, shared knowledge base defining the general concepts and basic procedures and standards
KPIs	For example, annual total number of valid normative documents, mean time between amendments, compliance to the agreed international and supranational legislation
Judicial Authority	
Integrated judicial information system	Central database and data warehouse for uniquely identified cases, used by courts, prosecutor offices, other investigating organizations, and penitentiaries; interfaces to the other public information systems; standard workflow using e-documents and e-archives; computer-aided support for detecting incompatibilities and conflicts of interest
KPIs	For example, duration of cases (from start of investigation until final solution) per components and categories of cases, number of rotations of dossiers, degree uniformity of solutions per similar cases and circumstances (cluster analysis)
Executive Authority	
Integrated interoperable information systems (central, sectoral, and territorial components)	Information systems assuring e-services to people and organizations, protection of human rights, administration of public properties, optimization of taxation and similar obligations, specialized e-services and information systems supporting e-business, minimization of administrative burdens for citizens and organizations, coherent decentralization of decisions, anti-underground economy and corruption information system(s)
Current management of the SES	Compliant to the requirements of the target e-SM: assuring the necessary continuity and, at the same time, integrating change and technical and technological progress, protecting against concentration of power in the political and economic sense, protecting against consequences of possible disasters
KPIs	For example, GDP/capita, distribution of incomes (e.g., Gini coefficient), life expectancy at birth, use of ICT, population with higher use of ICT (%), proportion of population per economic sectors, economic growth correlated with (un)employment and foreign debt, territorial distribution of values added per capita per economic sector, total state budget/GDP, degree of e-inclusion
Public Societal Feedback Authority	
Functioning of the official statistics	Collection of relevant data only if not possible to obtain by interfacing to public information systems, determination of KPIs and comparisons to planned values, generation of coherent tabular, graphic and cartographic statistical information and mathematical models (statistical knowledge) + *caeteris paribus* extrapolations
Functioning of the electoral information system	Online general and local e-voting free from electoral fraud, participative democracy information system
Functioning of auditing information systems	Audit of public finance and specialized auditing information systems (e.g., anti-trust, protection of consumers, etc.)
KPIs	Accuracy, timeliness, integrity

Note: Moon, Welch, and Wong (2005) analyzed pushing and pulling factors on e-government performance and various KPIs. The World Bank uses E6 governance indicators, based on data and surveys (World Bank, 2007b): voice and accountability; political stability and absence of violence; government effectiveness; regulations quality; rule of law, control of corruption. An example of a synthetic indicator is a proposed societal management index (Costake, 2007): SMI = ((GDP/capita [k EUR])/ (Rate of unemployment [%])) (corruption perception index)*

> IDABC > Factsheets > France > Strategy

Antiroikko, A.V. (Ed.). (2006). *Encyclopedia of digital government.* Hershey, PA: Information Science Reference.

Bangemann, M., de Benedetti, P., Davis, E. da Fonesca, Gyllenhamar, C. et al. (1994). *Europe and the global information society.* Recommendations to the European Council, Brussels, Belgium.

BECTA. (2001). *The 'digital divide': A discussion paper.* Retrieved from http://www.becta.org.uk/page_documents/research/digitaldivide.pdf

Bicking, M., & Wimmer, M. (2006). *E-government research in Europe: Disciplinary understanding and state of play from eGovrtd 2020.* Retrieved from http://www.egovrtd2020.org/EGOVRTD2020/navigation/events/conferences/Paper_EGOV2006

Biermann, F. (2006). *Earth system governance. The challenge of social science.* Working Paper No. 19, Global Governance, The Netherlands.

Costake, N. (2007). From e-governance to e-societal management. A next challenge. Proceedings of Eastern Europe E-Gov Days 2007, Prague.

Costake, N. (2008a). General requirements for digital government. In A.V. Antiroikko (Ed.), *Encyclopedia of digital government* (pp. 98-110). Hershey, PA: Information Science Reference.

Costake, N. (2008 b). From e-government to e-governance. In A.V. Antiroikko (Ed.), *Encyclopedia of digital government* (pp. 58-66). Hershey, PA: Information Science Reference.

Dror, Y. (2002). *The capacity to govern.* Retrieved from http://www.futurecasts.com/book%20review%204-03.htm,

E-Govt. (2005). *E-government experience in the U.S. and Canada: How relevant is it to developing countries?* Retrieved from http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTINFORMATIONANDCDTECHNOLOGIES/EXTEDEVELOPMENT/0,,contentMDK:20483017~pagePK:148956~piPK:216618~theSitePK:559460,00.htm

EICTA. (2004). *EICTA position on guidelines for future EU's policy to support research.* Retrieved from ftp://ftp.cordis.europa.eu/pub/era/docs/eicta_position.pdf

EITO. (2007). *Homepage.* Retrieved from http://www.eito.org

EUR-Lex. (2007). *European Commission: A to Z.* Retrieved from http://ec.europa.eu/atoz_en.htm#E

G7. (1995, February 25-26). G7 Information Society Conference. Retrieved from http://ec.europa.eu/archives/ISPO/intcoop/g8/i_g8conference.html

Gartner. (2007). *Preparation for update European interoperability framework 2.0—final report.* Retrieved from http://ec.europa.eu/idabc/servlets/Doc?id=29101

Gore, A. (2006). *An inconvenient truth.* Translated into Romanian by C.O. Tabarcea and edited by RAO International Publishing Co., 2007.

Greer, J.M. (2005). *How civilizations fall: A theory of catabolic collapse.* Author.

Greyson, J. (2007). *An economic instrument for zero waste, economic growth and sustainability.* Retrieved from http://www.sdinnovation.co.uk/Resources/GreysonZEROWASTEfinal.doc

Gupta, M.P., Kumar, P., & Bhattacharya, J. (2005). *Government online. Opportunities and challenges.* New Delhi, India: Tata McGraw–Hill.

Hart, S. (1997). Beyond greening: Strategies for a sustainable world. *Harvard Business Review,* (January-February), 67-78.

Heylighen, F. (2007). The global superorganism: An evolutionary-cybernetic model of the emerging network society. Retrieved from http://209.85.135.104/search?q=cache:kl9XdfUS2xAJ:pespmc1.vub.ac.be/papers/Superorganism.pdf+Cybernetic+Model+Society&hl=en&ct=clnk&cd=1&gl=uk&lr=lang_en

Huxley, A. (1958). *Brave new world revisited.*

i2010. (2006). i2010 e-government action plan: Accelerating e-government in Europe for the benefit of all commissions of the European communities. *Proceedings of COM(2006)173,* Brussels, Belgium.

IDABC. (2007). *Interoperable delivery of European e-government services to public administrations, businesses and citizens.* Retrieved from http://ec.europa.eu/information_society/index_en.htm > IDABC > (Projects/e-Procurement/e-Government Practice/Interoperability etc.)

IDABC. (2008). Shaping a strategy for the future: The IDABC work program. SINeRGY, (January), 3-4.

INSPIRE. (2004). Directive of the European Parliament and of the council establishing an infrastructure for spatial information in the community. *Proceedings of COM(2004)516,* Brussels, Belgium.

Jensen, I. (2001). *The Leontief open production model of input-output analysis.* http://online.redwoods.cc.ca.us/instruct/darnold/laproj/Fall2001/Iris/lapaper.pdf

Justice. (2004). *Homepage.* Retrieved from http://ec.europa.

eu/justice_home/index_en.htm

Klischewski, R., & Scholl, H.J. (2006). Information quality as a common ground for key players in e-government integration and interoperability. *Proceedings of the 39th Hawaii International Conference on System Sciences.*

Kyoto. (2007). *Kyoto accord.* Retrieved from http://mindprod.com/environment/kyoto.html

Laherrère, J.H. (1998). *The evolution of the world's hydrocarbon reserves.* Retrieved from http://www.google.co.uk/search?as_q=Hydrocarbon+world+reserves&hl=en&num=20&btnG=Google+Search&as_epq=&as_oq=&as_eq=&lr=lang_en&cr=&as_ft=i&as_filetype=&as_qdr=all&as_occt=any&as_dt=i&as_sitesearch=&as_rights=&safe=images[

Lee Boon Yang. (2007). *Singapore's e-government—sharing our experience.* Retrieved from http://www.megegovconflisbon.gov.pt > Presentations and Speeches > Presentations and Speakers

Lenk, K. (2003). E-government in Europe. The state of affairs. *Proceedings of the EGOV 2003 International Conference,* Prague, Czech Republic.

Lenk, K., & Traunmuller, R. (2000). Perspectives on electronic government. In F. Galindo & G. Quirchmayr (Eds.), *Proceedings of the Advances in Electronic Government Working Conference of the International Federat of Information Processing W.G. 8.5 and Center for Computers and Law University of Zaragoza* (pp. 11-26).

Lovelock, P., & Ure1, J.E. (2002). *E-government in China.* Retrieved from http://www.trp.hku.hk/publications/e_gov_china.pdf

Lorenz, K. (1973). *Die acht todsuenden der zivilisierten menschheit Piper.* Munchen.

Maslow, A. (2000). A business reader: A collection of Abraham Maslow's works (preface, pp. 3-4). New York: John Wiley & Sons.

Masuda, Y. (1981). The information society as post-industrial society. Bethesda, MD: World Future Society.

Migrationsverket. (2001). The Schengen information system. Retrieved from http://www.migrationsverket.se/infomaterial/bob/sokande/eu/sis_en.pdf

MODINIS. (2006). *Presenting eGEP main findings.* Retrieved from 3rdWorkshop Vienna_Conference_eGEP_presentation.pdf.

Moon, J., Welch, E., & Wong, W. (2005). What drives global e-governance? An exploratory study at the macro level. Proceedings of the 18th Hawaii International on System Sciences.

Naisbitt, J. (1984). Megatrends. New York: Warner Books.

OECD. (2001). Governance in the XIst century. Retrieved from http://www.oecd.org/dataoecd/15/0/17394484.pdf

Reading, V. (2007). The political challenge of simplifying and opening-up e-government. *Proceedings of the International Conference on Advancing E-Government,* Berlin.

Scholl, H.J. (2002). E-government: A special case of ICT-enabled business process change. Proceedings of the 36th Hawaii International Conference on System Sciences.

SEPA. (2007). Homepage. Retrieved from http://www.europapaymentscouncil.org

Situngkir, H. (2003). Powers of the governmental state as feedback control dynamic system. *Journal of Social Complexity, 1*(1), 1-11.

Soros, G. (1998). *Crisis of global capitalism: Open society endangered.* Retrieved from http://www.ciaonet.org/conf/cfr08

Tahir, M.S.B. (2005). *Reaping e-government opportunities in meeting global trends—e-governance: Beyond e-government.* Retrieved from http://www.eivc.org/uni/Uploads/Admin/Mr.%20Mohd%20Suhaimi%20Mohd%20Tahir.pdf

TESTA. (2005). *TESTA.* Retrieved from http://ec.europa.eu/idabc/servlets/Doc?id=19933

Timmers, P. (2005). The impact of e-government. *Proceedings of the 2nd Open Workshop on eGEP.*

Toffler, A. (1970). *Future shock.* New York: Random House.

Traunmuller, R., & Wimmer, M. (2004). E-government: The challenges ahead. In R. Traunmueller (Ed.), *Electronic government* (pp. 1-6). Berlin: Springer-Verlag.

UN. (1968, December 20). Resolution 2458 (XXIII). *Proceedings of the 23rd ONU General Assembly,* New York.

Undheim, T.A. (2008). Best practices in e-government—on a knife-edge between success and failure. *European Journal of E-Practice,* (2), 23-46.

Vanvelthoven, P. (2005). Information sharing for better public services. *Proceedings of the EGOV 2005 Ministerial Conference,* Manchester.

World Bank. (2007a). *World Bank governance & anti-corruption.* Retrieved from http://go.worldbank.org/GOHQB2VP40

World Bank. (2007b). *World Bank worldwide governance*

indicators 1996-2006. Retrieved from http://info.worldbank.org/governance/wgi2007/sc_chart.asp

WSIS. (2005).*World Summit on the Information Society.* Retrieved from http://www.itu.int/wsis/index.html

KEY TERMS

E-Administration: Synonym to local e-SM (e-SM of communities). Its use is not recommended, because it may suggest a domain of the executive authority only, unless the meaning in the national language is more general.

E-Governance: ICT-enabled management of an SES whose domain is limited to a national/federal one, including e-government as one component, and not necessarily including the executive (strategic) management.

E-Government: ICT-enabled management of the executive authority of a local or national SES. Its main content is made up of the e-services provided to people and organizations (obviously has the highest political content for a government which may be changed every four or five years).

E-Societal Management: ICT-enabled management of the SES whose domain may include an international component and covers the entire management content, including strategic management.

Human Rights: Rights defined in the UN Declaration of 1948, updated by similar international documents.

Public Institution: Organization whose mission is to serve the SES and is financed at least in large part by public money.

Public Sector: System of public institutions, and enterprises owned or managed by public institutions. This system may have many levels: international, regional, national (central, sectoral, or territorial), usually hierarchically organized. It covers executive (e.g., central and local governments, public finance, etc.), legislative (e.g., Parliament, etc.), homeostatic (e.g., Judiciary, etc.), and societal feedback (e.g., electoral, statistic, auditing, etc.) oriented institutions.

E-Government and Digital Divide in Developing Countries

Udo Richard Averweg
eThekwini Municipality and University of KwaZulu-Natal, South Africa

INTRODUCTION

The transition of the global economy from an industrial focus to one based on knowledge and information presents numerous opportunities and challenges to countries, especially those in the developing world (Cape IT Initiative, 2003). The government sector (and especially the local government sector) needs to embrace information and communication technologies (ICTs) that enable it to operate more efficiently and communicate better with its citizens.

ICTs encompass all technologies that facilitate the processing and transfer of information and communication services (United Nations, 2002). Many factors affect how local governments (i.e., municipalities) in developing countries access ICTs. In order to bridge the digital divide—which separates the technology 'haves' from the technology 'have nots'—it is necessary to gauge where citizens are in terms of ICT adoption, that is, their e-readiness. E-readiness can be defined in terms of availability of ICT infrastructure, the accessibility of ICT to the general citizen population, and the effect of the legal and regulatory framework on ICT use in, for example, an e-government strategy.

eThekwini Municipality (2003), in the city of Durban in the developing country of South Africa, sees the e-government strategy and its Web site at *http://www.durban.gov.za* as important management tools for improved citizen service delivery and communication. The objective of this article is to report, as an example, on the survey of ICT and information needs of a selected metropolitan municipal area (eThekwini Municipality in South Africa). Such a report maybe useful to other municipalities in developing countries for their e-government strategies.

This article is organized as follows. The background to e-government and the digital divide are discussed. eThekwini Municipality in South Africa is then described. The research goals are outlined, the research method and data gathering are discussed, the survey results and discussion are given, and future trends for implementing an e-government strategy in municipalities in developing countries are suggested. Finally, a conclusion is given.

BACKGROUND TO E-GOVERNMENT AND THE DIGITAL DIVIDE

Nowadays governments around the world are embracing electronic government. In a broad sense, e-government can be defined as the process by which government communication and administration processes are made available using ICTs. All such technologies can improve service and output in the same way that they have revolutionized work and leisure of lives—the main difference being that e-government programs recognize that not all citizens have equal access to technology and need to be implemented accordingly. In the literature it is also recognized that e-government in the developing world must accommodate certain unique conditions, needs, and obstacles. E-government gives citizens access to relevant information and makes government more accountable to its citizens. Ultimately, e-government aims to enhance access to and delivery of government services to benefit citizens (Pascual, 2003).

In the same way that there are social and economic divides between poor and rich countries, in the field of ICTs there are also divides between those who can access and use ICT to gain the associated benefits and those who do not have access to the technology or cannot use it for one reason or another (Bridges.org, 2002). These digital divides exist between countries ('international divide') and between groups within countries ('domestic divide'). In looking at the difference in access between developed and developing countries, Gumucio-Dagron (2003) notes that the "divide has never been only a 'digital' or technological divide. It is a social, economic and political fracture." The divide between technology 'haves' and technology 'have nots' is significantly wide.

eTheKwini MUNICIPALITY IN SOUTH AFRICA

eThekwini Municipality's population is 3.09 million citizens (Statistics South Africa, 2001) within the eThekwini Municipal Area (EMA). The population is an amalgamation of racial and cultural diversity. The Black African community comprises 68.3%, Coloured citizens 2.8%, Asian citizens

19.9%, and White citizens 9.0% (Statistics South Africa, 2001). In the EMA 51.9% of the population are female and 48.1% are male.

eThekwini Municipality's Long Term Development Framework (LTDF) maps out the strategic vision for eThekwini Municipality during the next 20 years. The essence of the LTDF "is to achieve a balance between meeting basic needs, strengthening the economy and developing people skills and a technology base for the future" (eThekwini Municipality, 2007b). eThekwini Municipality has a capital budget of ZAR4.2 billion (approximately €0.42 billion) and an operating budget of ZAR12.90 billion (approximately €1.29 billion) for the 2007/2008 financial year (see *www.durban.gov.za*).

Durban is currently rationalizing its wide area network (WAN) infrastructure to streamline and enhance service delivery. Broadband access delivered via fiber optic, wireless, and power lines are leveraged by information and collaboration portals as well as offering EMA citizens services via fixed and mobile devices. The WAN's wireless component is being extended to offer access to municipal libraries, clinics, and other eThekwini Municipality facilities. ICT and Web sites can be seen as effective mechanisms to access municipal information and developmental information in general. South African Web sites which seek a local and global reach must cater for the digital divide that exists between the technological 'haves' and 'have nots' (Averweg, Barraclough, & Spencer, 2003). Bridging the digital divide in the EMA is not the end but the beginning to bring positive changes in the development of a municipal information society (MIS). An MIS is the innovative use of ICT to improve the internal operation of a municipality, as well as its communication and collaboration with citizens, the private sector, and civil society in a municipal area.

RESEARCH GOALS

eThekwini Municipality embarked on an initiative to understand the needs of its users and non-users in utilizing ICT as a tool to improve service delivery and establishing effective media communication between itself and its constituencies. This article reports on these initiatives and findings from two surveys conducted in the EMA.

Research Method and Data Gathering

Two survey instruments (hereinafter referred to as 'ICT Status Survey' and 'Library Survey') were developed to gauge EMA citizens' ICT and information needs. The first survey tool (ICT Status Survey) represents an attempt to obtain a snapshot of the ICT status of EMA citizens; the second tool (Library Survey) focused on citizens' ICT and information needs over a broad spectrum.

The ICT Status Survey and Library Survey instruments and their associated survey methods are now described.

ICT Status Survey Instrument and Survey Method

This survey instrument comprised two sections: (1) general information and (2) citizen's information needs. During May 2003, the survey instrument was administered face-to-face to 465 EMA citizens by seven temporary staff members under the auspices of an eThekwini Municipal official. The duration of each interview was approximately 10 minutes. The selected sample was on a random basis to gather quantitative data to develop qualitative information. Interviews were conducted at EMA customer service offices and municipal libraries. The requirement for effective e-government requires a good understanding of the cultural or social background of its end users (citizens in its communities). The citizen survey thus focused on the e-readiness of EMA citizens to 'tap' into the new methods of communication for e-government.

Library Survey Instrument and Survey Method

The Library Survey focused on library usage, citizens' needs and expectations of library services and facilities, and/or reasons for non-usage. This survey was undertaken by the research organization, Urban-Econ, based in Durban, South Africa. For the purpose of this article, the results of the ICT-related requirements of citizens' needs from the Library Survey are considered complementary to the ICT Status Survey.

During June 2002 the survey instrument was administered face-to-face by librarians to 471 library users in different age categories. This was undertaken at selected municipal libraries in the EMA in accordance with the established sample profile for the various socio-economic groups. Two experienced fieldworkers conducted 144 interviews at pre-

Table 1. Race grouping and gender of respondents surveyed

Race Grouping	Percentage (%) of Male Respondents	Percentage (%) of Female Respondents
Black	49.0%	51.0%
Asian	52.9%	47.1%
Coloured	42.9%	57.1%
White	36.8%	63.2%
Average	**48.2%**	**51.8%**

Table 2. Computer experience by race grouping of respondents surveyed

Race Grouping	Percentage (%) of Respondents Who Have Some Computer Experience	Percentage (%) of Respondents Who Have NO Computer Experience
Black	46.8%	53.2%
Asian	77.5%	22.5%
Coloured	64.3%	35.7%
White	85.7%	14.3%
Average	**58.7%**	**41.3%**

Table 3. Occupation status by socio-economic group of respondents surveyed (adapted from eThekwini Municipal Libraries, 2002)

Occupation Status	Percentage (%) Distribution of Occupation Status by Socio-Economic Groups				
	Rural Low	Urban Low	Urban Middle	Urban Upper	Total
Studying/ scholar	58.6%	53.8%	29.8%	25.0%	43.6%
Employed	20.7%	17.9%	31.6%	25.0%	23.4%
Self-employed	10.3%	5.1%	3.5%	0.0%	4.8%
Not working	10.3%	20.5%	24.6%	11.7%	19.8%
Pensioner	0.0%	2.5%	10.5%	33.3%	8.5%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

selected households, which ensured sample representivity of the surveyed areas. The results from this survey are contained in the "eThekwini Municipal Libraries: User and Non-User Survey" report, dated August 26, 2002, and used to inform the author's study.

SURVEY RESULTS AND DISCUSSION

From the ICT Status Survey, Table 1 reflects the race grouping and gender of respondents surveyed.

A comparable gender composition was reported in the Library Survey.

From the ICT Status Survey, Table 2 reflects the computer experience by race grouping of respondents surveyed.

From Table 2, an average of 58.7% of citizens reported that they have some computer experience. Computer experience by White citizens is relatively high (85.7%), followed by Asian citizens (77.5%). Black citizens reported the least computer experience (46.8%).

From the Library Survey, Table 3 reflects the occupation status by socio-economic group of respondents surveyed.

From Table 3, the need for information is greatest among students/scholars (43.6%), followed by employed citizens (23.4%). Unemployed (not working) citizens also indicated a significant need for information (19.8%).

From the ICT Status Survey, Table 4 reflects the computer literacy by socio-economic group of respondents surveyed.

From Table 4, 20.7% of citizens reported that they had no computer literacy, while 53.7% of citizens indicated

Table 4. Computer literacy by socio-economic group of respondents surveyed (adapted from eThekwini Municipal Libraries, 2002)

Computer Literacy	Percentage (%) Distribution of Computer Literacy by Socio-Economic Group				
	Rural Low	Urban Low	**Urban Middle**	**Urban Upper**	**Total**
Little bit	13.8%	34.6%	24.6%	12.5%	25.5%
No	31.0%	24.4%	3.3%	25.0%	20.7%
Yes	55.2%	41.0%	66.7%	62.5%	53.7%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 5. Respondents surveyed with Internet access who visited eThekwini Municipality's Web site

Percentage (%) of Respondents Who Have Visited eThekwini Municipality's Web Site	Percentage (%) of Respondents Who Have NOT Visited eThekwini Municipality's Web Site
16.7%	83.3%

Table 6. Internet usage patterns of respondents surveyed (adapted from eThekwini Municipal Libraries, 2002)

Reason for Internet Usage	Percentage (%) of Internet Usage by Library Users	Percentage (%) of Internet Usage by Library Non-Users
E-mail	42.2%	20.8%
Study assignments	43.6%	7.6%
School projects	35.1%	34.0%
Games	18.6%	21.5%
General research	9.5%	18.8%
Other	1.6%	2.15%
Total	**100.0%**	**100.0%**

Table 7. Information needs by socio-economic group of respondents surveyed (adapted from eThekwini Municipal Libraries, 2002)

Information Required	Percentage (%) Distribution of Information Needs by Socio-Economic Group				
	Rural Low	Urban Low	Urban Middle	Urban Upper	Total
Business	13.8%	6.4%	1.5%	4.2%	8.0%
Community services	17.2%	16.7%	8.8%	20.8%	14.9%
eThekwini Municipality	0.0%	3.8%	3.5%	0.0%	2.7%
Health	20.7%	3.8%	5.3%	0.0%	6.4%
International	24.1%	17.9%	21.1%	8.3%	18.6%
Local government	3.4%	2.6%	3.5%	0.0%	2.7%
National government	3.4%	10.3%	14.0%	16.7%	11.2%
No response	0.0%	5.1%	10.5%	20.3%	8.0%
Study projects	17.2%	28.2%	21.1%	25.0%	23.9%
Other	0.0%	5.1%	1.8%	4.2%	3.2%
Total	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

Table 8. Delivery mechanisms for receiving information about eThekwini Municipality

Delivery Mechanism	Percentage (%) of Respondents
Telephone	4.3%
Post Office	69.6%
Municipal customer service office	7.6%
Municipal publication	13.3%
Community meeting	4.4%
Internet	0.1%
School/tertiary	0.7%
Total	**100.0%**

that they are computer literate. Lesame (2005) reports that in South Africa, "[m]any schools make use of computers in the quest to create a computer literate society by teaching educators and learners computer skills required in the information society."

From the ICT Status Survey, Table 5 reflects the respondents surveyed with Internet access who visited eThekwini Municipality's Web site (*www.durban.gov.za*).

From Table 5, 16.7% of citizens reported that they had visited the Web site at *www.durban.gov.za*. The networking capability offered by the Internet and related technologies have the potential to transform structures and operations of government organizations. OleKambainei and Sintim-Misa (2003) argue:

"There is also a need to promote the use of ICT to provide better, cheaper and faster government services and information electronically, increase citizens' participation in decision-making and facilitate good governance (e-governance). To accomplish this, an effort should be made to develop comprehensive and active Web sites for governments."

From the Library Survey, Table 6 reflects the Internet usage patterns of respondents surveyed.

From Table 6, the Internet was used by library users mostly for study assignments (43.6%). This was closely followed by e-mail (42.2%) and school projects (35.1%). In the case of library non-users, the Internet was mostly used for school projects (34.0%).

From the Library Survey, Table 7 reflects the information needs by socio-economic group of respondents surveyed.

From Table 7, citizens' information needs focus on study projects (23.9%). This was followed by international infor-

mation (18.6%) and then community services information (14.9%). The need for business information (8.5%) and local government information (2.7%) rank relatively low.

From the ICT Status Survey, Table 8 reflects the delivery mechanisms for receiving information about eThekwini Municipality.

From Table 8, a significant number of citizens receive information about eThekwini Municipality via the *MetroBeat* publication (69.6%). The *MetroBeat* is delivered by the South African Post Office to citizens' post boxes.

The aim is not simply to deliver services electronically (e.g., Internet, short message service (SMS)) in the EMA, but to encourage its citizens to start learning about the Internet via its Web site and thereafter make use of the Internet for other services. The eThekwini Municipality Integrated Development Plan 2010 and Beyond report states that eThekwini Municipality's strategic commitment is to develop a smart city:

...to bridge the digital divide in eThekwini and to become a hub of information diffusion, as well as a centre for economic growth and integration. Bridging the digital divide will reduce the gap between those who have access to information and communication technology, and those who do not have access for socio-economic or infrastructure reasons. (eThekwini Municipality, 2007a)

It is argued that similar commitments to bridging the digital divide exist in other municipalities in developing countries.

The two surveys (ICT Status Survey and Library Survey) focused on establishing a better understanding of how ICT can contribute to eThekwini Municipality's citizen service delivery and development communication. From the eThekwini Municipality surveys, generic future trends are now presented which may be useful to other municipalities in developing countries for their e-government strategies.

FUTURE TRENDS

From this research, some future trends for implementing an e-government strategy in other municipalities in developing countries are suggested:

- Physical Access: All citizens should have equal access to the services to which they are entitled. Access to communications and the Internet are cornerstones of an MIS. Digital inclusion cannot be achieved without providing all citizens access with affordable ICT appliances to the information highway.
- Appropriate Technology: Network evolution has thus become an imperative for electronic service delivery (ESD). The challenge is to chart an appropriate course of network deployment which does not perpetuate Gumucio-Dagron's (2003) 'new apartheid' syndrome.
- Affordability: The communication mechanism should be affordable to citizens. Furthermore, public services should be provided economically and efficiently in order to give citizens the best possible value for their money.
- Human Capital: Training and re-skilling will be necessary. Since many citizens lack ICT skills and/or do not have access to desktop or laptop PCs, demonstrations on how to effectively utilize ICTs (e.g., e-mail/SMS) must be provided.
- Relevant Content: The content developed must be locally relevant to its constituency, especially in terms of language. Averweg et al. (2003) suggest that a Web site must facilitate access by end users not familiar with Internet norms and whose home language is not English. To bridge the digital divide through e-government, e-government must be relevant to citizens.
- Integration: ICT must not act as a further burden to citizens' lives. ICT should be integrated into priority sectors of the EMA economy and into citizens' daily lives. Okpaku (2003) suggests "ICTs have become so important to virtually all aspects of life…to systems of governance, that they have become fundamental to basic life."
- Socio-Economic Factors: A government has a responsibility for the well-being of its employees that cannot be ignored as new ICTs are introduced. The socio-economic status of citizens (end users) should be considered. By evaluating their e-readiness, this will determine the usability of e-government tools.
- Political Will: Government is a political organization in which elected officials are ultimately accountable to the voters (citizens), and this accountability should be acknowledged.
- Democratization of Society: Worldwide there has been an explosion in projects and initiatives—on a global, national, and (most often) local level—to exploit the potential of the Internet to draw citizens into civic participation and thereby enhance democratic participation (Tsagarousianou, 1998).
- Legal and Regulatory Framework: Government regulations affecting technology use and changes that need to be made to create an environment that fosters ICT usage must be considered. The effect of legal and regulatory frameworks on ICT use should be geared to facilitate the knowledge and information economy growth in the digital age in the developing world.

CONCLUSION

Given its location at the grassroots of any democratic society, a municipality faces the greatest challenge of all spheres of government when it comes to service delivery. While citizens interact with government at all levels, it is the services provided by local government with which citizens are most familiar and have the highest expectations. In response to the demands placed on municipalities in developing countries, e-government has a vital role to play in transforming the traditionally paper-based, paper-laden operations of municipalities' back offices and through streamlining the interaction between municipalities and their citizens. It is about how citizens change in relating to 'their' municipality and the degree to which e-government changes citizens relating to each other in an MIS.

To seek building an ICT capacity without a solid foundation of research and development is nothing but building a skyscraper in quicksand (Okpaku, 2003). Further research needs to be conducted by municipalities in developing countries regarding developmental issues that may impact their e-government development strategies. Should such research be undertaken, municipalities in developing countries will be able to ensure they adopt appropriate ICT infrastructures for their citizens, thereby improving ESD, operating more efficiently, communicating better with citizens, and narrowing the digital divide.

REFERENCES

Averweg, U.R., Barraclough, C.A., & Spencer, A.F.O. (2003, December 8-10). Towards creating a municipal information society: The development of 'eThekwini Online' in South Africa. *Proceedings of the World Forum on Information Society* (WFIS), Geneva, Switzerland.

Bridges.org. (2002). *Taking stock and looking ahead: Digital divide assessment of the city of Cape Town.* Retrieved from http://www.bridges.org/capetown

Cape IT Initiative. (2003). *First census of Western Cape ICT companies.* Retrieved from http://www.citi.org.za

eThekwini Municipality. (2003). *Integrated development plan 2003-2007.* Retrieved from http://www.durban.gov.za/council/transformation/download.htm

eThekwini Municipality. (2007a, July). *Integrated development plan, 2010 and beyond.* 2007-2008 Review, 1-120, Corporate Policy Unit, eThekwini Municipality, South Africa.

eThekwini Municipality. (2007b). *Medium term budget 2007/2008 to 2009/2010.* Unpublished Report, 1-110, eThekwini Municipality, South Africa.

eThekwini Municipal Libraries. (2002, August 26). *User and non-user survey, 2002.* Unpublished Report, URBAN-ECON, South Africa.

Gumucio-Dagron, A. (2003). *Take five: A handful of essentials for ICTs in development.* Retrieved from http://www.geocities.com/agumucio/ArtTakeFive.html

Lesame, Z. (2005). The social and economic aspects of the Internet. In N.C. Lesame (Ed.), *New media technology and policy in developing countries* (pp. 207-215). Pretoria: Van Schaik.

Okpaku, J.O. (2003). *Information and communications technologies as tools for African self-development*. ICT Task Force Series 2, United Nations ICT Task Force.

OleKambainei, E., & Sintim-Misa, M.A. (2003). Info-communication for development in Africa. In J.O. Okpaku (Ed.), *Information and communication technologies for African development* (ch. 9). ICT Task Force Series 2, United Nations ICT Task Force.

Pascual, P.J. (2003). E-governance. *Proceedings of the UNDP–Asia-Pacific Development Information Program, World Summit on the Information Society, Geneva 2003–Tunis 2005,* Kuala Lumpur, Malaysia.

Statistics South Africa. (2001). *Census 2001 digital census atlas.* Retrieved from http://gis-data.durban.gov.za/census/index.html

Tsagarousianou, R. (1998). Back to the future of democracy? New technologies, civic networks and direct democracy in Greece. In R. Tsagarousianou, D. Tambini, & C. Bryan (Eds.), *Cyberdemocracy: Technology, cities and civic networks.* London: Routledge.

United Nations. (2002). *Towards a knowledge-based economy. Regional assessment report.* New York: Author.

KEY TERMS

Digital Divide: A social, economic, and political fracture.

E-Governance: Refers to a government's inventiveness to electronically govern areas under its jurisdiction.

E-Government: The process by which government communication and administration processes are made available using information and communication technologies (ICTs).

E-Readiness: Defined in terms of availability of ICT infrastructure, the accessibility of ICT to the general citizen and business organization population, and the effect of the

legal and regulatory framework on ICT use in, for example, an e-government strategy.

Electronic Service Delivery (ESD): A method of delivering services and conducting business with customers, suppliers, and stakeholders to achieve local government developmental goals of improved customer service and business efficiency.

Information and Communication Technologies (ICTs): An umbrella term for a range of technological applications such as computer hardware and software, digital broadcast technologies, telecommunications technologies, and electronic information resources.

Municipal Information Society (MIS): A term used for the innovative use of ICT to improve the internal operation of a municipality, as well as its communication and collaboration with citizens, the private sector, and civil society in a municipal area.

E-Government and E-Democracy in the Making

Birgit Jaeger
Roskilde University, Denmark

INTRODUCTION

The development of electronic or digital government (e-government) has varied throughout the world. Although we give it the same name, we know from different studies that, for example, the concept of Information Society can be interpreted in different ways in different cultural settings (Jaeger, Slack, & Williams, 2000; Sancho, 2002). This article provides a general outline of the development of e-government in the West and is primarily based on European and Scandinavian experiences.

It is only possible to give an introduction to e-government if we can define what we are talking about. E-government is still a rather new concept, but most people agree that e-government includes the following features:

- E-government is based on information and communication technologies (ICTs).
- E-government is taking place in public administration.
- E-government concerns electronic ways to perform all kinds of internal administrative tasks.
- E-government also concerns the communication between the public administration and the citizens and other actors in the surrounding society (Jaeger, 2003: 50).

BACKGROUND

Based on the first part of this definition the history of e-government starts in the beginning of the 1960s when the magnetic tape replaced the punched card. During the 1960s and 1970s big central databases were built and were run on big mainframe computers. The databases mostly contained administrative data from fields where the law and regulations were clear and there was a large amount of data to process. In this period, large registers were formed, and software systems for the government of the economy including salaries, taxes and pensions were developed. These activities were often run centrally and the results were delivered to the relevant authority on paper.

When we turn to the second feature in the definition, we have to include the development of the public administration as well. During the 1980s and 1990s most Western countries had experienced a profound modernization of their public administration. At first, this modernisation was marked by reforms that have since been collectively labelled New Public Management. According to Rhodes (1997), New Public Management involves two different types of initiatives, the first of which relates to the management itself. These initiatives include a focus on management by objectives, clear standards, and evaluations of the quality of service, while at the same time granting greater attentiveness towards the users of the public service in question. The other type of initiative deals with the introduction of economic incentive structures. This involves the dissection of the public administration in demarcated services, contracting out some services, and other services are sought arranged in competitive-like situations by establishing quasi-markets in which the consumers of the services are provided with an opportunity to choose between different services.

These alterations have had a more or less unintended consequence—the emergence of new policy networks around the provision of public services (Heffen, Kickert & Thomassen, 2000; Rhodes, 1997; Stoker, 1998). These policy networks draw new agents into the management of the tasks in question, including agents from the business community as well as from civil society. Now we see private companies carrying out publicly-commissioned services. We also see civic groups in the local community; NGOs, sports clubs or interest organizations take over different tasks of more social and carrying kind, which were earlier defined as public. (Again we have to be aware of different traditions in different countries but especially in the Scandinavian countries; many of these tasks have been defined as public whereas in other Western countries the family and local community have played a much bigger part in taking care of these activities.) These agents are now engaged in relations with the public administration in collective, binding policy networks. This general development is often described in terms of the transformation of public sector regulation from government to governance.

These reforms differ from country to country (Rhodes, 1999), but the general picture is that these reforms have had great impacts on the way the public administrative tasks are preformed. Here the development of e-government plays a significant role. Many of these reforms would have been very difficult to realize without ICTs. An example of this is the decentralization of administrative tasks from town halls

to public institutions in Denmark. This reform was based on the use of PCs and the development of an internal electronic network between the town hall and all the public institutions. Today we describe it as the start of the development of the intranet in the authority in question (Jaeger, 2003).

This development has continued and today we have a wide range of different software systems for all kinds of administrative tasks. These include electronic archives, systems for handling electronic documents, systems for consideration of different cases and so forth. Garson (2000) provides an overview of this field as well as a review of the literature.

Rather early on, it became clear that the development of e-government was not just a question of the design of an information system and its implementation in an organization. Thus, over the years, a lot of effort has been put into developing methods for the process of design and implementation (Bødker, Kensing, & Simonsen, 2000). Based on analyses of different failures, it was acknowledged that it is very important to draw on the experience of the potential users in the design process. Otherwise, it is easy to produce systems that do not fit their needs. The experience also showed that the implementation of the system is very important if the organization is going to harvest the benefits of the system. A lot of parameters have to be taken into consideration in this process. The staff has to be informed and drawn into the process; it is likely that some training is needed; the way to organize the consideration of cases has to be carefully examined; and it is perhaps necessary to draw in other competences than those that already exist in the organization. The role of the users is not simply a question of starting to use the technology. Studies have shown that users have to domesticate a new technology before they are able to utilize it (Lie & Sørensen, 1996; Silverstone, Hirsch, & Morley, 1992). In this process they can also shape the technology and make it fit to the conditions under which it is utilized (Jaeger, 2005b; Oudshoorn & Pinch, 2003). The inclusion of all these factors is important if one wishes to ensure that the design and implementation of a new information system is to be a success for the public administration.

During the last couple of decades, public authorities on different levels have developed their own information systems for performing their tasks. This has lead to a situation where many public agencies are unable to communicate electronically because they use different technological protocols and standards. Thus there is a need today for developing common standards for electronic communication between public agencies at different levels. In recent years, this has become a barrier to the development of e-government and therefore a large amount of resources is now spent on solving this problem.

E-DEMOCRACY

Nevertheless, public authorities do not only communicate internally or with other public agencies, they also communicate with citizens, private companies and other users of public services. This is the last of the above listed features of e-government. With the introduction of the World Wide Web in the 1990s, the public authorities were given a tool for this external communication. During the late 1990s and since, most public agencies have developed their own Web site where they place a lot of information, and electronic forms citizens have to fill out to apply for a public service and so forth.

Also in this area of the e-government, we find different kinds of development. In a study of the development of digital cities in Europe (Bastelaer, Henin, & Lobet-Maris, 2000; Williams, Stewart, & Slack, 2005), it became clear that in some cities [e.g., Copenhagen (Jaeger, 2002)], the web site was developed as a part of the e-government and interpreted as a tool for communication between the public authorities and the citizens, while in other cities [e.g., Amsterdam, (Van Lieshout, 2001)], the web site was developed as a tool for communication between citizens and did not involve the public authorities very much.

In terms of definitions, this is the most debated aspect of e-government. Some people interpret a public web site only as being a tool for administrative tasks for use between the public agency and the citizens, while others see the web site as a place for debate and a tool for democracy as well. The first group defines the objectives of e-government as a way to rationalize public administration and increase its efficiency, thus democratic debate should, in their understanding, not be a part of a public authority's web site but should be developed as something else—as e-democracy. The last group defines the objectives of e-government as a tool for all the tasks a public authority has and, consequently, also a tool for the democratic process. In this understanding, the democratic use of ICTs should be developed side by side with the administrative use.

Whether e-democracy is developed as an integrated part of e-government or as a special service, it is the least developed use of ICTs. To further this development, it is necessary to define what kind of democracy the technology should support. Without going into a theoretical discussion of the concept of democracy, it is possible to state that we have at least two different kinds of democracy: representative democracy and participatory democracy. Representative democracy is what the Western world mostly defines as democracy, and functions through a parliament where all the citizens in a country have the possibility to elect some politicians to represent them. Participatory democracy is defined as the wide range of activities in which citizens participate in the political process. It

is important to remember that these two kinds of democracy are not mutually exclusive. In most Western countries, the two kinds of democracy exist side by side.

Technology will be developed differently in accordance with the type of democracy it should support (Hoff, Horrocks, & Tops, 2000). If e-democracy is going to be a tool for the representative democracy, it can be used for electronic voting or online referendums (Hauge & Loader, 1999). Or it can be used to support the political parties in their dialog with the voters (Löfgren, 2001). If e-democracy is going to be a tool for participatory democracy, it should be used to support political debate [e.g., in a local community or in other political processes, (Hoff & Storgaard, 2005)]. Then it is a question of making web sites containing information about the political process and chat rooms or newsgroups where people can discuss different political subjects and where politicians have an opportunity to argue for their opinions.

In recent years, there have been a row of small-scale experiments where ICTs have been tried out as a tool for e-democracy (Torpe, Nielsen, & Ulrich, 2005). Some of these have been promising, but many have more or less failed (especially the experiments with political debates on the Internet). Even if some of these experiments have been promising, we still lack a convincing example of large-scale use of ICTs in e-democracy.

THE POLITICIANS AND E-GOVERNMENT

The development of e-government has a great impact on the people who are affiliated to the public administration. During a Danish study of the role of politicians in the development of e-government (Jaeger, 2005a) it became clear that they play a very limited part in this development. It is mostly the civil servants who decide how to apply ICTs in e-government. Hesitancy among politicians towards becoming involved in the development of e-government and determining the means by which ICT is utilized in administration is largely due to their role conception, for example, whether or not they perceive involvement in such concrete, technical tasks, as being a part of their role.

The role of the politician builds on a sharp distinction between politics and administration. The politician is to represent the people and put forward grand political visions for societal development. Conversely, the administrator's role is to serve the politicians by carrying out their visions and generally administrating society on the basis of the politically defined framework. The means by which administrators implement and administer political decisions is regarded as a technical matter. The administrators can therefore choose the methods and instruments they consider to be most suitable and efficient in the given situation. According to this perception of the division of roles, it is clearly up to the administrators to determine how e-government should be shaped.

This distribution of roles between politicians and administrators has long historical roots and has been working well for generations. But due to the shift in the public administration from government to governance, the politicians have to change their role if they want to maintain their position as governors of the public sector. Sørensen (2002) describes the new requirements of politicians under governance regulation as different means of exercising meta-governance (Kooiman, 2000). The requirements of politicians exercising meta-governance concern the creation of frames to allow other actors to participate in the process. The role of the politicians as meta-governors is then to specify the competence that various actors have to perform, and to make decisions. In order to get this range of independent actors to work in the same direction, the politicians are also responsible for standing forth to offer political leadership capable of creating meaning via a common understanding of the general goals.

The development of e-government can serve as a possibility for the politicians to fulfill the requirements for exercising meta-governance. The manner in which public authorities are presently coordinating their utilization of ICTs will determine the electronic infrastructure these authorities must use internally and with external actors for many years to come. Accordingly, e-government can be configured in a way that supports the politicians as meta-governors. The use of ICTs can make it easier for the politicians to communicate with the external actors, and by doing so make the framework for cooperation visible just as they can use the technology for debates that are important in their efforts to create common understanding and meaning. But this development will not happen by itself. The politicians must perceive e-government as a dimension of the institutional framework for network regulation, just as they must conceive of technology as part of their work to create meaning and identity, as well as in their work with the construction and support of various policy networks. At the same time, the involvement of politicians in the e-government design process could probably lead to a further development of e-democracy where the democratic use of ICTs would receive greater emphasis.

FUTURE TRENDS

The potential of e-government is not yet fully discovered. In the coming years, we will see new ways of utilizing ICTs, which will enhance the performance of e-government. The field of communication between different public agencies especially can prove to increase the effectiveness of e-government if the developers succeed in agreeing to common standards for communication and the exchange of information. For the time being the buzzword for this development is: ICT architecture.

Furthermore, the field of a safe and accountable identification of the user has great potential for increased effectiveness.

With the successful development of one it will be possible to design a long row of forms, which require a safe identification of the user, and as a result shift a lot of work from the civil servants in the public administration to the individual user of public services. At the same time, an accountable identification of the user will make it possible for the public authority to give the citizens access to the information about them and to follow the consideration of their own cases. By doing this, it will be possible to have a much more open administration than the one we know today.

CONCLUSION

E-government is here to stay. Even though we still lack the results from a general study, which proves that e-government is a more efficient way to perform public administration and establishes how much money e-government has saved, some case studies have shown that a public administration using ICTs at least in these cases can perform more tasks, with the same amount of staff, serving a bigger population than the public administration without using ICTs (Jaeger, 2003, pp. 107-110).

The further success of the development of e-government is of course also dependent on the attitude of the citizens. If they do not have access to the technology, if they do not know how to utilize a computer, if they do not accept public services in an electronic way, or if they do not trust the electronic services but are afraid of misuse, then it will not be possible to realize the potentials in e-government. In this way, it becomes a task for the public authorities to prevent a digital divide among the citizens.

It is still rather unclear what will happen to the development of e-democracy. As it looks today, it is unlikely that e-democracy will be developed as an integrated part of e-government. On the other hand, the many small experiments with different forms for e-democracy point to an independent development where the democratic potentials of ICTs are tried out and new applications are developed. This situation is even more likely to take place if the politicians revise their role and start to interpret the development of e-government as a part of their role.

REFERENCES

Bastelaer, B. V., Henin, L, & Lobet-Maris, C. (2000). Villes virtuelles. *Entre Communauté et Cité. Anslyse de cas*. Paris: L'Harmattan.

Bødker, K., Kensing, F, & Simonsen, J. (2000). *Professional IT-inquiry—The goundation of sustainable IT-utilization*. Copenhagen: Samfundslitteratur.

Garson, D. G. (Ed.). (2000). *Handbook of public information systems*. New York: Marcel Dekker, Inc.

Hauge, N. B., & Loader, B. D. (Eds.). (1999). *Digital democracy. Discourse and decision making in the information age*. London: Routledge.

Heffen, O. V., Kickert, W. J. M., & Thomassen, J. J. A. (2000). *Governance in modern society. Effects, change and formation of government institutions*. Dordrecht: Kluwer Academic Publishers.

Hoff, J., Horrocks, I., & Tops, P. (Eds.). (2000). *Democratic governance and new technology*. London: Routledge.

Hoff, J., & Storgaard, K. (Eds.). (2005). *Information technology and democratic innovation—Citizen participation, political communication and public administration*. Frederiksberg: Forlaget Samfundslitteratur.

Jaeger, B. (2002). Innovations in public administration: Between political reforms and user needs. In J. Sundbo, & L. Fuglsang (Eds.), *Innovation as strategic reflexivity* (pp. 233-254). London: Routledge.

Jaeger, B. (2003). *Local authorities on the Net. Roles in e-government*. Copenhagen: DJØF Publishing.

Jaeger, B. (2005a). Digital visions—The role of politicians in transition. In V. Bekkers & V. Homburg (Eds.), *The information ecology of e-government* (pp. 107-125): IOS Press.

Jaeger, B. (Ed.). (2005b). *Young technologies in old hands—An international view on senior citizens' utilization of ICT.* Copenhagen: DJØF Publishing.

Jaeger, B., Slack, R., & Williams, R. (2000). Europe experiments with multimedia: An overview of social experiments and trails. *The Information Society*, *16*(4), 277-301.

Kooiman, J. (2000). Societal Governance: Levels, models and orders of social-political interaction. In P. J. (Ed.), *Deating governance. Authority, steering and democracy* (pp. 138-166). Oxford: Oxford University Press.

Lie, M., & Sørensen, K. H. (Eds.). (1996). *Making Technology our own? Domestication technology into everyday life*. Oslo: Scandinavian University Press.

Löfgren, K. (2001). Political *parties and democracy in the information age. The cases of Denmark and Sweden.* Unpublished Ph.D., Copenhagen University, Copenhagen.

Oudshoorn, N., & Pinch, T. (Eds.). (2003). *How users matter. The co-construction of users and technologies.* Cambridge, Mass.: The MIT Press.

Rhodes, R. A. W. (1997). *Understanding governance—Policy networks, governance, reflexivity and accountability.* Philidelphia: Open University Press.

Rhodes, R. A. W. (1999). *Understanding Governance: Comparing public sector reform in Britain and Denmark* (Workingpaper 17/1999). Copenhagen.

Sancho, D. (2002). European national platforms for the development of the Information Society. In J. Jordana (Ed.), *Governing telecommunications and the new information society in Europe* (pp. 202-227). Cheltenham: Edward Elgar.

Silverstone, R., Hirsch, E. & Morley, D. (1992). Information and communication technologies and the moral economy of the household. In R. Silverstone & E. Hirsch (Eds.), *Consuming technologies. Media and information in domestic spaces* (pp. 15-31). London: Routledge.

Stoker, G. (1998). Governance as theory: five propositions. *International Social Science Journal*, March 1998 (155), 17-28.

Sørensen, E. (2002). *Politicians and the network democracy. From sovereign politician to meta-governor*. Copenhagen: DJØF Publishing.

Torpe, L., Nielsen, J. A., & Ulrich, J. (2005). *Democracy at the Net. Publicity, participation and digital communication*. Aalborg: Aalborg Universitetsforlag.

Van Lieshout, M. (2001). Configuring the digital city of Amsterdam: Social learning in experimentation. *New Media & Society, 3*(2), 131-156.

Williams, R., Stewart, J., & Slack, R. (2005). *Social learning in technological innovation. Experimenting with information and communication technologies*. Cheltenham Glos: Edward Elgar.

KEY TERMS

Accountable Identification: A way to identify a person in an electronic interaction and to give legal status to electronic documents. Different technologies have been tried out (e.g., chip cards and digital signatures).

Digital City: Usually a web site, which is centered on a city, where public authorities, business and citizens can communicate and exchange information.

E-Democracy: Depending on what type of democracy it should support, ICT can be used for electronic voting, online referendums, or to support the political parties in their dialog with the voters. It can also be used to support political debate in a local community or in other political processes.

E-Government: Is based on ICT, taking place in public administration, concerns electronic ways to perform administrative tasks, and the communication between the public administration and the citizens.

Meta-Governance: A way to govern independent policy networks. The politicians have to create the frames that make it possible for other actors to participate in the policy process. They have to specify the competence of the various actors and to create meaning via a common understanding of the goals for the performance of the network.

New Public Management: Includes initiatives which relate to management of the public administration (e.g., management by objectives, clear standards, and evaluation of the quality of service). It also includes initiatives that deal with the introduction of economic incentive structures (e.g., outsourcing of public tasks and establishing of quasi-markets for public services).

Policy Network: Are centered on the provision of public services and include, beside the public administration, agents from the business community as well as from civil society (e.g., NGOs, sports clubs or interest organizations). These agents are engaged in interdependent relations with the public administration.

E-Learning Adaptability and Social Responsibility

Karim A. Remtulla
University of Toronto, Canada

INTRODUCTION

The global, knowledge-based economy is causing rapid change when it comes to workforce composition and the nature and character of work itself. At the same time, 'e-learning' is increasingly positioned as the panacea for workplace learning needs for a transforming workplace and the global, knowledge-based economy (Industry Canada, 2005; Rohrbach, 2007). In this information age of intense political, social, technological, and environmental upheaval, do organizations bear any social responsibility towards their employees when mandating workplace learning from their employees through e-learning?

The International Organization for Standardization (ISO, 2007a) specifies four key areas that all organizations need to pay heed to for 'social responsibility' to be accomplished: "environment; human rights and labor practices; organizational governance and fair operating practices; and, consumer issues and community involvement/society development" (para. 6). Accordingly, given the criteria of "organizational governance and fair operating practices," this article argues for e-learning adaptability as a burgeoning social responsibility in the workplace, when thinking about workplace learning, by discussing: (a) the workforce diversity, and other workplace changes, that increasingly challenge the current approaches to e-learning at work; and then, (b) highlights the e-learning adaptability framework (Remtulla, 2007) as one methodology to assess and enable e-learning adaptability to meet this social responsibility for the benefit of a global workforce.

BACKGROUND

Diversity at Work

Skills shortages are becoming more severe in advanced economies (OECD, 2005; Rohrbach, 2007). This is primarily due to the fact that the European Union and North America are facing an aging workforce, a dwindling youth cohort, and declining birth rates, simultaneously, resulting in a smaller workforce in the future to fuel the needs of mega-corporations. This means that the workforce is not only becoming more demographically diverse, but also more multicultural, because immigration from developing countries will count for most of the labor force growth in advanced economies in the near future.

Workers are also becoming more multifaceted. To remain competitive, workers are assuming personal responsibility for their learning and upskilling. One outcome of this is mass underemployment as workers bring with them an increasing range of talents to each new job. Many workers' skills and knowledge already far exceed the career opportunities available to them and their employers' ability to use these skills despite demanding it of their workers to get work in the first place (Mirchandani, 2003).

Other Workplace Changes

Work is becoming more homogeneous when it comes to tasks and responsibilities. One widely accepted reason for this is the influence of international standards bodies that promulgate systems to harmonize various job tasks across various industries and regions. Well-known examples of this are the ISO, International Electrotechnical Commission (IEC), and International Telecommunication Union (ITU). In turn these standards become a form of accountability on the job, mandating everyone to act in accordance with these 'international' standards. These international standards bodies work together and construct uniformity as necessary and universally beneficial (ISO, 2006).

Jobs are also becoming more normalized around certain competencies and behaviors with respect to 'high skills'. This comes from a pervasive belief that high-skilled work and competencies, based on knowledge and continuous innovation, are universally tantamount to business continuity and profitability (Rohrbach, 2007).

Social Responsibility, Workplace Learning, and E-Learning

Given the knowledge-based economy and corresponding workplace changes, e-learning is being promoted as the 'grand solution' for workplace learning, ushering in an era of anytime education and anywhere access to knowledge (Gasco, Llopis, & Gonzalez, 2004; Pollitt, 2005). However, the expression 'e-learning' is at present associated with a number of definitions that take a highly limited view of this form of workplace learning, such as:

- "A wide set of applications and processes, such as Web-based learning, computer-based learning, virtual classrooms and digital collaboration. It includes the delivery of content via Internet, intranet/extranet (LAN/WAN), audio- and videotape, satellite broadcast, interactive TV and CD-ROM." (DeRouin, Fritzsche, & Salas, 2005, p. 920)
- "The use of Internet technologies in order to provide a wide range of solutions that might improve knowledge and performance." (Andrade et al., 2005, p. 658)

As exemplified by these definitions, the dominant focus on e-learning remains almost exclusively on the issues of instructional design, hardware, or software. They focus on the mechanics and not the people, nor learning. Similarly, workplace learning professionals respond by tailoring their programs and practices to support these same homogenizing, normalizing, and standardizing trends in jobs, skills, and competencies and to this fixation on the 'technology' (Gagnon & Doray, 2005; Remtulla, 2007). However, what of the changing workforce? The changing nature of the workforce necessitates some acknowledgment of the needs of the global workforce and their unique circumstances.

The relevance and urgency for such acknowledgment in the implementation of mass, workplace learning interventions like e-learning were identified as a social responsibility as far back as the late-1990s, when the global, knowledge-based economy began to unequivocally impact the daily lives of individuals at work (as described earlier). This is echoed, for example, in the following passage from "Adult Education: The Hamburg Declaration—The Agenda for the Future" (UNESCO-UIE, 1997):

The development of the new information and communication technologies brings with it new risks of social and occupational exclusion for groups of individuals and even businesses which are unable to adapt to this context. One of the roles of adult education in the future should therefore be to limit these risks of exclusion so that the information society does not lose sight of the human dimension. (p. 6)

Yet, as noticed from the above definitions of e-learning and approaches to workplace learning, the acknowledgment of the needs of the global workforce remains elusive. Recognizing the lack of the 'human dimension' in organizational standards, the ISO (2007b) is already working on the development of an international standard (ISO 26000) providing guidelines for social responsibility; it is scheduled for release by 2010. Ziva Patir, chair of the ISO Technical Management Board, sums up the inequities in the current situation this way:

Our traditional role was to promote the standardization of products, services, processes, materials and systems. Then we evolved by developing standardized tools for management practice and now we are evolving further to develop standards that address the human aspects.

Today, in the light of ISO strategic vision for 2005-2010, we understand that everything is interconnected and one can no longer differentiate between software and hardware, between product and service, between management tools and the values of the organization. ISO has developed a policy to ensure the global relevance of our work, and today there are few areas more relevant than social responsibility (SR). (p. 3)

The significance of e-learning, as an issue of social responsibility in acknowledging the workplace learning needs of the global workforce, will become paramount in the future, as already noted by the efforts of the ISO and its global partner organizations like the United Nations Educational, Scientific, and Cultural Organization (UNESCO).

THE E-LEARNING ADAPTABILITY FRAMEWORK

To assist organizations in meeting their social responsibility through organizational governance and fair organizational practices, a multiperspectival framework may be one approach for assessing e-learning adaptability that brings together the elements of the knowledge-based economy, workplace changes, hardware, software, instructional design, skills, competencies, workplace learning, and the global workforce into a socially responsible and cohesive methodology. Such a framework represents a more 'socially responsible' alternative to current hardware, software, and instructional design-only based approaches to e-learning because this framework looks not just at how e-learning influences the global workforce, but also how the cultural and the social variability of the global workforce influence e-learning and workplace learning through their needs, motivations, and attitudes.

The e-learning adaptability framework (Remtulla, 2007) comprises a media perspective, a genre perspective, and a learning perspective, to allow for a multiperspectival take on e-learning in the workplace based on context, culture, and community. These three perspectives are further aligned along an *adaptability continuum*: 'media' at one extreme (which considers workplace context); 'genre' (which takes into account user communities in the workplace); and finally, 'learning' (which concerns notions of culture and how people learn differently). When taken together as a continuum, these perspectives represent an interconnected, mutually symbiotic, multiperspectival framework as a socially responsible methodology that potentially provides

better support in dealing with workplace learning and the challenges of a global workforce that remain unrecognized by current (mechanistic) approaches to e-learning.

Context, Learning, and a Media Perspective

Existing 'e-learning as hardware' approaches answer to the trend in the homogenizing of jobs in the workplace, but not necessarily to the contexts surrounding these jobs. These approaches effect the supplanting of face-to-face workplace learning by hardware (Remtulla, 2007). The premise is straightforward: greater variety and functionality in hardware for distribution and access of information equates to better workplace learning, regardless of context. Terms like 'computer-based instruction' and 'Web-based training' indicate the format of the content and the mode of its availability. By ignoring context, the focus stays on access and information systems and the storage and retrieval of information. Decisions about e-learning and workplace learning now have more to do with investment in and/or deployment of hardware and less to do with the learning needs of a diverse, multifaceted, global workforce.

A media perspective increases e-learning adaptability by neutralizing the *homogenizing* effect of hardware-only approaches to e-learning for workplace learning. This perspective goes beyond hardware-only approaches by placing e-learning in context. It prevents the misperception that e-learning is nothing more than distribution and access to information; a challenge readily addressed through expenditure on more hardware. A media perspective first looks to context, and then to what channels or formats may be made available for distribution and access given this context. In other words, this perspective assesses e-learning adaptability as complementary to contexts in the workplace, rather than implementing hardware as a substitute for them.

Communities, Learning, and a Genre Perspective

Current 'e-learning as software' approaches speak to the normalizing of skills and competencies in the knowledge-based economy, but not necessarily the multitude of communities that individuals interact with at work. These approaches bring about the erasing of difference between jobs through commercial and off-the-shelf software and applications solutions (Remtulla, 2007). Again, the formula is straightforward: greater standardization in software and applications equates to better workplace learning. This happens with standardized, one-size-fits-all software applications as a prescription for predetermined knowledge gaps.

In any organization, it is not uncommon to have many user communities. As this multifaceted, global workforce grows, the likelihood of individuals affiliating and associating with different communities at work will also likely increase. On the other hand, the e-learning software in question is likely to be more *generic* than is actually needed by the user groups and may not be entirely responsive for the particular gender, race, age, and/or other personal and professional communal needs of the users, especially when dealing with a global workforce. Decisions about e-learning now become more about obtaining the most cost-efficient software and less about the learning needs of a diverse, multifaceted, global workforce.

A genre perspective increases e-learning adaptability by countering the *normalizing* effect of software-only approaches. 'Genre' of software involves the intended users and *their* relation to one another (Bezemer & Kress, 2008). This perspective shifts the spotlight away from the media and hardware, and onto software and applications to determine which user communities will ultimately benefit from e-learning for workplace learning and why.

A genre perspective offers additional advantages to and appends a media perspective when pondering e-learning adaptability for workplace learning. It acknowledges the existence of different user groups and their competing requests and circumstances (Graham, 2004; Munro & Rice-Munro, 2004), whereas the media discussion limits the entire e-learning adaptability debate to a discussion about context. A genre point of view also accepts the fragmentation and partitions within communities of users in the workplace. In other words, this perspective assesses e-learning adaptability as fostering cross-cultural communication, community building, and interaction among user groups, rather than installing standardized software as a substitute for them.

Culture, Learning, and a 'Learning' Perspective

A 'learning' perspective on e-learning considers pedagogy—the principles and practices of instruction that make workplace learning possible. When it comes to the global workforce and workplace changes, this viewpoint may be most relevant to assessing e-learning adaptability because it, in fact, undergirds even the media and genre perspectives. It encompasses notions of knowledge and circumstances, and of *how and why* people come to learn things.

This perspective exposes the *universalizing* paradigm of present 'instructional design only' approaches. This universalized worldview is based on five implicit assumptions common to modern workplace learning programs and practices (Johansen & McLean, 2006; McLean, 2006):

(1) Western views and values of learning apply to all workers.

(2) The content and method of teaching are optimal when consistent.
(3) Learning is a personal, solitary, and cognitive undertaking.
(4) All workers are equal when it comes to access and capacity to learn.
(5) Learning must align with economic goals.

All this seems to ignore almost entirely the workplace learning needs and motivations of the non-Western workers who will populate more and more of the global workforce in the future.

The advantage that a learning perspective has over both media and genre perspectives when it comes to e-learning adaptability is that culture is brought squarely into the conversation. At the same time, it does not disregard either context or community, but goes beyond these to look at culture—that is *learning and the learners.* In other words, this perspective assesses e-learning adaptability as enabling a culture of organizational learning (Jensen & Markussen, 2007) and knowledge-building among workers, across contexts and communities, and for a global workforce, rather than using universalized instructional design as a substitute for them.

A Socially Responsible Definition for E-Learning

As such, this chapter now proposes the following definition of e-learning that includes acknowledgment of the needs of the global workforce and the implementation of e-learning through a socially responsible methodology: *E-learning is a process of workplace learning that utilizes information and communications technologies to promote competencies and skills for a diverse workforce in an adaptive manner that is sensitive to context, community, and culture in the workplace.*

FUTURE TRENDS

Organizational stakeholders need to be more vigilant in gaining a better grasp over the needs and complexities of workplace learning for a global workforce and workplace changes when envisioning the future of e-learning. The focus of attention must move away from hardware, software, and instructional design, and towards the global workforce and their contexts, communities, and cultures in the workplace. More powerful software and hardware is not the whole answer. They do not necessarily translate into learning or accomplishing social responsibility while global workers' needs go unresolved.

As such, further research is also needed on e-learning adaptability from a standpoint that incorporates multiple perspectives (i.e., media, genre, and learning) to bring about workplace learning. Additional innovation in assessment criteria and measurement tools and methods that link e-learning adaptability with organizational governance and fair organizational practices are also recommended. E-learning adaptability will become even more imperative through a growing, global workforce that is increasingly diverse; that may not be reflecting workplace priorities, accountabilities, performance, and productivity; and that may be *not learning* despite intensive investment in e-learning and extensive assessments of hardware, software, or instructional design.

CONCLUSION

Current approaches to assessing e-learning in the workplace focus on hardware, software, or instructional design. They speak to homogenizing, normalizing, and universalizing trends in the workplace when it comes to work, skills, and competencies in the knowledge-based economy. Hardware is not a substitute for context. Software cannot take the place of community. Instructional design cannot replace culture.

Given the global workforce and workplace changes, a multiperspectival framework to look at e-learning adaptability—one that brings context, community, and culture together with e-learning for workplace learning—now becomes essential. The e-learning adaptability framework comprises a continuum of three perspectives on e-learning adaptability: (a) a media perspective, which assesses e-learning adaptability as complementary to the contexts of the workplace; (b) a genre perspective, which assesses e-learning adaptability as fostering cross-cultural communication, community building, and interaction of user groups; and (c) a learning perspective, which assesses e-learning adaptability as enabling a culture of organizational learning and knowledge building among workers. When taken together as an interconnected, mutually symbiotic, and multiperspectival methodology, these perspectives represent a socially responsible framework that potentially provides better support in clarifying the workplace learning implications of a global workforce from e-learning, and vice versa.

REFERENCES

Andrade, J., Ares, J., García, R., Rodríguez, S., Seoane, M., & Suárez, S. (2005). A pedagogical overview on e-learning. In R. Khosla (Ed.), *KES 2005* (pp. 658-664). Berlin: Springer-Verlag.

Bezemer, J., & Kress, G. (2008). Writing in multimodal texts: A social semiotic account of designs for learning. *Written*

Communication, 25(2), 166-195.

DeRouin, R.E., Fritzsche, B.A., & Salas, E. (2005). E-learning in organizations. *Journal of Management, 31*(6), 920-940.

Gagnon, L., & Doray, P. (2005). Corporate training and the knowledge society: A re-examination of factors influencing participation.

Gasco, J.L., Llopis, J., & Gonzalez, M.R. (2004). The use of information technology in training human resources: An e-learning case study. *Journal of European Industrial Training, 28*(5), 370-382.

Graham, G. (2004). E-learning: A philosophical enquiry. *Education & Training, 46*(6/7), 308-314.

Industry Canada. (2005). *The Canadian education and training industry. Commercial education and training.* Retrieved September 8, 2005, from http://strategis.ic.gc.ca/epic/internet/incet-ecf.nsf/en/ok01770e.html

ISO (International Organization for Standardization). (2006). *ISO and world trade. Overview of the ISO system.* Retrieved August 15, 2006, from http://www.iso.org/iso/en/aboutiso/introduction/index.html#two

ISO. (2007a). *Future ISO 26000 standard on social responsibility reaches positive turning point.* Retrieved March 1, 2008, from http://www.iso.org/iso/pressrelease.htm?refid=Ref1049

ISO. (2007b). *Participating in the future international standard ISO 26000 on social responsibility*. Geneva, Switzerland: ISO Central Secretariat.

Jensen, C.B., & Markussen, R. (2007). The unbearable lightness of organizational learning theory: Organizations, information technologies, and complexities of learning in theory and practice. *Learning Inquiry, 1*(3), 203-218.

Johansen, B.-C.P., & McLean, G.N. (2006). Worldviews of adult learning in the workplace: A core concept in human resource development. *Advances in Developing Human Resources, 8*(3), 321-328.

McLean, G.N. (2006). Rethinking adult learning in the workplace. *Advances in Developing Human Resources, 8*(3), 416-423.

Mirchandani, K. (2003). Immigrants matter: Canada's social agenda on skill and learning. *Convergence, 37*(1), 61-68.

Munro, R.A., & Rice-Munro, E.J. (2004). Learning styles, teaching approaches, and technology. *Journal for Quality and Participation, 27*(1), 26-32.

OECD (Organization for Economic Cooperation and Development). (2005). *The measurement of scientific and technological activities: Guidelines for collecting and interpreting innovation data: Oslo manual* (3rd ed.). Retrieved January 10, 2008, from http://stats.oecd.org/glossary/detail.asp?ID=6864

Pollitt, D. (2005). E-learning delivers management skills to Ford's North American dealers. *Training & Management Development Methods, 19,* 6.39-36.42.

Remtulla, K. (2007). E-learning and the global workforce: Social and cultural implications for workplace adult education and training. In K. St. Amant (Ed.), *Linguistic and cultural online communication: Issues in the global age* (pp. 276-305). Hershey, PA: IGI Global.

Rohrbach, D. (2007). The development of knowledge societies in 19 OECD countries between 1970 and 2002. *Social Science Information, 46*(4), 655-689.

UNESCO-UIE (Institute for Education). (1997, July 18). Adult education: The Hamburg declaration—the agenda for the future. *Proceedings of the CONFINTEA 5th International Conference on Adult Education.* Retrieved October 31, 2006, from http://www.unesco.org/education/uie/confintea/pdf/con5eng.pdf

KEY TERMS

Community: The numerous professional, bureaucratic, social, gendered, racial, cultural, and other groups that individuals affiliate and associate with in the workplace.

Context: The shifting and overlapping political, social, technological, and environmental spaces that individuals occupy in the workplace.

Culture: The common values and paradigms that individuals share with each other across the various communities and contexts they encounter in the workplace.

E-Learning Adaptability Framework: A multiperspectival framework for assessing e-learning for a global workforce based on a continuum of three perspectives: media, genre, and learning.

Homogenizing: To make uniform and constant in function and intention.

Knowledge-Based Economy: "An expression coined to describe trends in advanced economies towards greater dependence on knowledge, information and high skill levels, and the increasing need for ready access to all of these by the business and public sectors" (OECD, 2005, para. 71). Also called *knowledge economy.*

Normalizing: To conform to a standard or widely accepted norm.

Universalizing: To generalize to all people, circumstances, and situations.

Workplace Learning: The skills and competencies that individuals gain knowledge of for work through higher and tertiary education and (re)certification; through workplace learning interventions, both formally organized or informally through coaching and mentoring between coworkers; or by means of other self-directed efforts both inside and outside of the workplace.

Electronic Government and Integrated Library Systems

Yukiko Inoue
University of Guam, Guam

INTRODUCTION

Twenty First Century Government is enabled by technology—policy is inspired by it, business change is delivered by it, customer and corporate services are dependent on it, and democratic engagement is exploring it. Technology alone does not transform government, but government cannot transform to meet modern citizens' expectations without it (Cabinet Office, 2005, p. 3).

According to the E-Government Readiness Ranking Report (United Nations, 2005), in 2005 the United States was the world leader followed by Denmark, Sweden, and the United Kingdom; and in 2004 the Republic of Korea, Singapore, Estonia, Malta, and Chile were also among the top 25 "e-ready" countries. The Ranking Report further emphasizes that 55 countries, out of 179, which maintained a government Web site, encouraged citizens to participate in discussing key issues of importance, and that most developing country governments around the world are promoting citizen awareness about policies, programs, approaches, and strategies on their Web sites—thus making an effort to engage multi-stakeholders in participatory decision-making.

Indeed, one of the significant innovations in information technology (IT) in the digital age has been the creation and ongoing development of the Internet—Internet technology has changed rules about *how* information is managed, collected, and disseminated in commercial, government, and private domains. Internet technology also increases communication flexibility while reducing cost by permitting the exchange of large amounts of data instantaneously regardless of geographic distance (McNeal, Tolbert, Mossberger, & Dotterweich, 2003). In Hirsch's (2006) words, "The Internet has finally achieved the convergence dream of the 1970s and everything that can be canned in digital form is traveling the Net" (p.3).

BACKGROUND

For hundreds of years, American government agencies have collected and provided data and information (such as statutes and regulations, court decisions, votes by Congress, and the records of hearings) for both citizens and government. In fact, American Federal Government has adapted progressive computer and telecommunication technologies both operationally and in policy to harness computing power to improve government performance and enhance citizen access to government and other information services and resources since the development of Internet technology—from the initial steps to establish the Internet in the late 1960s (originally ARPANET) to the establishment of the National Information Infrastructure (NII) and National Performance Review (NPR) initiatives in 1993 (Aldrich, Bertot, & McClure, 2000).

In the early 1990s, city governments began to use e-mail, listserv, and the World Wide Web (WWW or Web) to deliver information and services; and by the end of the 1990s, Web-based services were already an integral and significant part of Electronic Government (e-government) (Ho, 2002). E-government, simply defined as utilizing the Internet and Web for delivering government information and services to citizens, refers to the use by government agencies of IT (such as Wide Area Networks, the Internet, and mobile computing) that have the ability to transform relations with citizens, businesses, and other arms of government (AOEMA, 2006). In essence, e-government could enable citizens to interact and receive services from the federal, state, or local government "24 hours a day," "7 days a week"; it has taken promising steps to deploy e-government services, but much remains to be done, both in implementing e-government services and in developing new technologies and concepts (AOEMA, 2006). All the government activities essentially arise from a mixture of motives intertwined, principally in the interests of *efficiency*, *information access and provision*, and *democracy* (Hirst & Norton, 1999).

And in the electronic information age, the traditional roles of the Federal Depository Library Program (FDLP) libraries in selecting, acquiring, organizing, and providing access to and services for government information are more important than ever (Jacobs, Jacobs, & Yeo, 2005).

ELECTRONIC GOVERNMENT

Key Dimensions

E-government incorporates four key dimensions that reflect the functions of government itself, but the last dimension

is oftentimes slighted because it is mostly invisible to the public (Dawes, 2002): (1) *e-service* (the electronic delivery of government information, programs, and services often over the Internet); (2) *e-democracy* (the use of electronic communications to increase citizen participation in the public decision-making process); (3) *e-commerce* (the electronic exchange of money for goods and services, such as citizens paying taxes and utility bills, and renewing vehicle registration); and (4) *e-management* (the use of IT to improve the management of government from streamlining business processes to maintaining electronic records to improving the flow and integration of information). According to Dawes (2002), New York City is clearly in the first tier of jurisdictions in its development of e-government (NYC.gov). Also the city's Web site has been recognized nationally for its design and usability; especially after "September 11" (that is, "the 2001 terrorist attack on America"), the website has been used with great resourcefulness and flexibility to provide New Yorkers with the best information available.

Objectives/Innovations

Using information and communication technologies (ICTs), e-government promotes more *efficient* and *effective* government, but it is not a shortcut to economic development, budget savings, or efficient government. E-government is a *process* (thus called "e-volution") and often a struggle that presents costs and risks, both financial and political (AOEMA, 2006). However, e-government is not simply the process of moving existing government functions to an electronic platform but calls for "re-thinking" the way government functions are carried out today to improve some processes, involving four objectives/innovations (AOEMA, 2006) (see Table 1).

In addition to the above objectives/innovations, the Department of Labor (DOL), for instance, ensures that federal employees with disabilities are able to use IT to do their jobs, and that members of the public with disabilities who interact with DOL will be able to use IT to access information on equal footing with people who do not have disabilities, in reference to Sections 504 and 508 of the Rehabilitation Act of 1973 (as amended) and published standards (DOL, n.d.).

Security and Privacy

Through the Internet and Web revolution, more effective, convenient, and flexible e-government services are happening, yet *security* and *privacy* are the two critical issues in e-government. Building secure e-government systems requires a careful balancing between providing convenient access and appropriately monitoring permission and, in reality, technology-based solutions are still at their infancy and the existing alternatives consist essentially of enforcing privacy by law or self-regulation of operational practices ("Security and privacy," 2002).

Successfully implementing e-government does require a level of trust on the part of all transacting parties, and e-government security and privacy protection activities address the protection of the government assets involved in e-government. For example, DOL (n.d.) has developed a comprehensive cyber security program in accordance with the Federal Legislation and Policies, which include the Federal Information Security Management Act of 2002 and Privacy Act of 1974. Accomplishments by DOL include: (1) developing system security plans for major applications, general support systems, and financial systems; (2) developing an enhanced computer security awareness training

Table 1. E-government objectives/innovations

Government To Citizen (G2C):
- Providing one-stop, online access to information and services to individuals (citizens should be able to find what they need quickly and easily)
- Disintermediation of civil service staff—delivering services directly to citizens

Government To Business (G2B):
- Reduced burdens on business by providing one-stop access to information to facilitate business development
- Skilled, IT-literate, and flexible citizens for the labor market

Government To Employee (G2E):
- The ability to easily gather information from the field
- Access to important applications and content

Government To Government (G2G):
- Reducing the fractured nature of individual department and agencies, and moving towards "joined-up" government
- Changing the culture of the civil service from reactive to proactive

plan; and (3) issuing the systems development life-cycle methodology integrated IT security into each phase of the project's life cycle.

INTEGRATED LIBRARY SYSTEMS

In the United States, there are deeply rooted values that a democracy requires of an informed citizenry, that government must be accountable to its citizens, and that citizens must have full, free, and easy access to information about the activities of their government; therefore, these "values" have led to the creation of the Government Printing Office (GPO) and FDLP (Jacobs et al., 2005). GPO began operations in accordance with Congressional Joint Resolution 25 of June 23, 1860 as part of the legislative branch of the Federal Government, and it operates under the authority of the Public Printing and Documents Chapters of Title 44 of the U.S. Code (GPO, 2003). To organize, manage, and disseminate the Federal Government's information, the Depository Library Act of 1962 established FDLP, which is an important "partnership" between government and libraries to link the public with government information and resources.

New Online Catalog

According to GPO (2006), *Franklin* (franklin.gpo.gov)—the new online catalog for the National Bibliography of American Government Publications (a comprehensive index of public documents from all three branches of the Federal Government)—is being prepared for its official launch: *Franklin*, which is a component of a modernization plan to replace older legacy systems with GPO's recently acquired state-of-the-art integrated library system, will contain more than 500,000 records dating from July 1976 to the present and will become more far reaching in the future. Through the office of information dissemination programs, GPO disseminates the largest volume of American government publications and information in the world (*GPO Access*, 2005). In the future, GPO will first focus on digitizing legislative and regulatory material that expands the coverage of the most popular *GPO Access* databases (GPO, 2006).

Free Online Access

The traditional mission of GPO (that is, printing, and selling publications) is being made increasingly irrelevant by technological change (Jacobs et al., 2005). GPO was therefore required to disseminate government information with the passage of the GPO Electronic Information Access Enhancement Act of 1993; after this Act became law, the amount of electronic information in the program grew rapidly; in 1995, Public Law 103-40 authorized *GPO Access*—a free online access point to many government databases (Finchum, 2004). In 1993, Congress mandated that the Congressional Record and the Federal Register be made available online for free; as a consequence, there are about one million downloads per day from the Federal Register (Drake, 2006). As Drake further notes, 92% of all Federal Government documents are available online, and the remaining 8% represent specialized publications such as large maps; in 2005, 50% of all government publications were born "digital" and will never be printed by the Federal Government; however, there are now questions about the best way to keep them in perpetuity.

Additionally, the Library of Congress is midway through its ten-year, $100 million National Digital Information Infrastructure and Preservation Project, designed to develop digital preservation strategies; in 2005, the National Archives awarded Lockheed Martin a $308 million contract to develop ways of preserving diverse e-government records (Manes, 2006).

FUTUTE TRENDS

Information Society in 2015

E-government programs, indeed, affect the social, political, and economic fabric of a country, and national e-government programs have progressed steadily around the world since 2001; despite numerous creative initiatives, however, most national e-government programs still focus on *information*, rather than interactive services (International Telecommunication Union or ITU, 2006). Concerning the information society in 2015, ITU predicts as follows:

The next decade will witness innovation and growing convergence between the largely unregulated computing and consumer electronic industries and the more heavily regulated communications and media industries, posing new challenges for policy design and regulation. "The information society in 2015 initiative" considers future scenarios for technological convergence, the desegregation of sector-specific infrastructure from services and content, and the impact of global connectivity, content, and service providers. This initiative, based on possible scenarios, will consider how national, regional, and international policy and regulatory frameworks need to adapt. It will also consider whether shifts in further policy and regulatory regimes are likely to be led by individual users, industry groups or by government mandate—or by a combination of approaches (p. 2).

Moreover, by improving service delivery and associated costs, and by enhancing communication between citizens and government, e-government enables constituents to access information and services from home, which reduces traffic flow, and improves the environment. Sources of government

information are thus rapidly transitioning from paper format to electronic format. It should be noted, however, that e-government and regulatory policy making capabilities are only beginning to emerge and, in the future, it can be expected to fall within the four main areas (i.e., IT, agency management of rule making, public involvement in the rulemaking process, and regulatory compliance); these areas of the research can be enhanced through coordinated research efforts that involve perspectives from both social sciences and information sciences (Coglianese, 2004). Further, there has been a growing appreciation for the importance of doing interdisciplinary research—thus involving ecology, economics, sociology, psychology, political science, public policy, and urban design and planning (Waddell & Borning, 2004). Finally, five challenges identified by Dawes (2002) may continuously hold the future of e-government: (1) comprehensive strategy; (2) integration of information and services; (3) privacy and data sharing; (4) dynamic use of the Web; and (5) partnerships and other organizational networks.

The Once and Future FDLP

The mission of FDLP is to disseminate information about the government activities to over 1,250 Federal Depository Libraries in the 50 states, the District of Columbia, and U.S. territories (*GPO Access*, 2005). The shift to digital production of government information does not change the need for FDLP libraries and five criteria are those that FDLP has been meeting for decades and will be meeting for the future (Jacobs et al., 2005, p. 201): (1) information is available and fully functional to all without charge; (2) information is easy to find and use; (3) information is verifiably authentic; (4) information is preserved for future access and use in a distributed system of digital depository libraries; and (5) privacy of information users is ensured so that citizens can freely use government information without concern that what they read will be subject to disclosure or examination. To access the electronic information, and to make information easy to find and use, librarians must remain current with computer and Internet skills, in addition to knowledge of the legal system and the legislative process (Yang, 2002); thus "libraries must act as expert service providers, rather than warehousing physical collections" (James, 2003, p. 19).

In order to achieve a government information system including the aforementioned five criteria, each of the following stakeholders has to play a significant role (Jacobs et al., 2005): "government agencies" have obligation to collect and compile information for the public; "GPO" should be able to help an agency conform to preservation and dissemination standards and facilitate the creation of metadata; "Congress" should adequately fund the dissemination of information and avoid redundancy and inefficiency; "FDLP libraries" should be able to select government information, acquire digital files, preserve them, and organize them through integration into collections of other information—thus providing access to and service for that information; the role of the "National Archives and Records Administration" is long-term preservation of the record of government; and the "private sector" adds value to government information by re-packaging, re-organizing, and re-distributing the information.

CONCLUSION

Modern governments with serious transformational intent see technology as a "strategic asset" and not just a tool. The real challenge ahead is not just to "do IT better" in the context of the past models for delivery of pubic services; it is also about "doing IT differently" to support the next phase of public service reform—building services which are *more* joined-up, *more* personalized, *more* efficient, and *more* effective in terms of policy outcome (Cabinet Office, 2005).

More and more people are turning to the Internet for information, and there is an increasing emphasis on electronic dissemination of government information. At the same time, electronic information must be preserved against corruption and loss, and there must be ways of ensuring that information created today can be used tomorrow (Jacobs et al., 2005). Another important issue is related to the fact that the Internet's open nature remains an ideal arena for dissemination of misinformation. The question is how to deal with false information on the Web and how to decide whether a source is reliable, even though each individual is ultimately responsible for his or her use of technology and for decisions taken based on information gathered from the Web.

In summary, *power*, *control*, *security*, and *regulation* are the pieces of the "newborn" paradigm in the electronic information age of today (Hirsch, 2006). Certainly, more than ever, each country needs a strategy to position itself in the new economy and define the regulatory framework of e-government applications in accordance with its own goals. That is why Hirsch's inquiry is so important: What will be the copyright enforcement criteria in the Internet era?

REFERENCES

Aldrich, D., Bertot, J. C., & McClure, C. R. (2002). E-government: Initiatives, Developments, and Issues. *Government Information Quarterly, 19*(4), 349-355.

AOEMA (Asia Oceania Electronic Marketplace Association). (2006). *E-government: Definitions and objectives*. Retrieved May 3, 2006, from http://www.aoema.org/E-Government/Definitions_ and_Objectives.htm

Cabinet Office. (2005). *Transformational government enabled by technology*. Retrieved May 5, 2006, from http://www.cio.gov.uk/transformational_government/implplan

Coglianese, C. (2004). Information technology and regulatory policy. *Social Science Computer Review, 22*(1), 85-91.

Dawes, S. S. (2002). *The future of e-government*. Retrieved May 3, 2006, from http://www.ctg.albany.edu/publications/reports/future_of_egov?chap

DOL (Department of Labor). (n.d.). *U.S. Department of Labor in the 21st century.* Retrieved May 4, 2006, from http:www.dol.gov/_sec/e_government_plan/p23_security_privacy.htm

Drake, A. Miriam. (2006). Collaboration, Competition, and Controversy. *Information Today, 23*(3), 1-2.

Finchum, T. (2004). Browse topics: Government information webliographies. In F. Baudino, L. Nardis, S. G., Park, & C. J. Ury (Eds.), *Brick and click libraries symposium Proceedings* (pp. 125-129). Owens Library, Northwest Missouri State University.

GPO (Government Printing Office). (2003). *About the U.S. Government Printing Office*. Retrieved July 2, 2004, from http://www.gpoaccess.gov/about/index.html

GPO (Government Printing Office). (2006). *GPO update for ALA*. Retrieved May 1, 2006, from www.access.gpo.gov/su_docs/fdlp/events/ala_update06.pdf

GPO Access. (2005). *About the Federal Depository Library Program*. Retrieved May 1, 2006, from http://www.gpoaccess.gov/fdlp.html

Hirsch, C. (2006, March). *Global communications newsletter (a publication of the IEEE Communications Society*). Retrieved May 2, 2006, from www.comsoc.org/pubs/gcn

Hirst, P., & Norton, M. (1999). *Electronic government: Information technologies and the citizen*. Retrieved May 1, 2006, from www.parliament.uk/post/egov.htm

Ho, A. T-K. (2002). Reinventing local governments and the e-government initiative. *Public Administration Review, 62*(4), 434-444.

ITU (International Telecommunication Union). (2006). *Results of a questionnaire on possible topics for new initiatives workshops in 2006*. Retrieved July 23, 2006, from www.itu.int/osg/spu/ni/ITUNewInitiativesQuestionnaireResults2006.pdf

Jacobs, J. A., Jacobs, J. R., & Yeo, Shinjoung. (2005). Government Information in the Digital Age: The once and future federal depository library program. *The Journal of Academic Librarianship, 31*(3), 198-208.

James, B. R. (2003). New Directions for the FDLP. *A Quarterly Journal of Government Information, 31*(3/4), 17-20.

Manes, S. (2006). Keeping our bits about us. *Forbes, 177*(4), 60-62.

McNeal, R. S., Tolbert, C. J., Mossberger, K., & Dotterweich, L. J. (2003). *Innovating in digital government in the American States, 84*(1), 52-70.

Security and privacy in digital government application. (2002). Harrisonburg, VA: Commonwealth Information Security Center (CISC), James Madison University. Retrieved June 10, 2004, from http://www.cisc.jmu.edu/research/bouguettayal.html

United Nations. (2005). *Global e-government readiness report 2005: From e-government to e-inclusion*. United Nations publication.

Waddell, P., & Borning, A. (2004). A case study in digital government. *Social Science Computer Review, 22*(1), 37-51.

Yang, Z. Y. (2002). An assessment of education and training needs for government documents librarians in the United States. *Journal of Government Information, 28*(4), 425-439.

KEY TERMS

ARPANET: Advanced research projects agency network that the Internet has roots in, developed by the Department of Defense.

Electronic Information Access Enhancement Act: *GPO Access* is a free service funded by the Federal Depository Library Program and has grown out of Public Law 103-40, known as the Government Printing Office Electronic Information Enhancement Act of 1993.

Information Age: The current era, characterized by the shirt from an industrial economy to an information economy and the convergence of computer and communication technology.

Information and Communication Technologies (ICT): Includes telecommunications technologies (such as telephony, cable, satellite, and radio) and digital technologies (such as computers, information networks, and software).

Internet: World's largest network, a worldwide collection of networks that link together millions of businesses, governments, educational institutions, and individuals using modems, telephone lines, and other communications devices and media. Also called the Net.

Listserv: An automatic mailing list server developed by Eric Thomas for BITNET in 1986, and a program that automatically sends messages to multiple e-mail addresses on a mailing list.

National Information Infrastructure (NII): A futuristic network of high-speed data communications links that eventually will connect virtually every facet of our society.

National Performance Review (NPR): A management reform initiatives established by the national administration to identify ways to make the government work better and cost less.

Wide Area Networks (WAN): A network that extends over a long distance. Each network site is a node on the network. The largest WAN in existence is the Internet.

World Wide Web (WWW): Worldwide collection of electronic documents on the Internet that have built-in hyperlinks to other related documents. Also called the Web.

Electronic Marketplace Support for B2B Business Transactions

Norm Archer
McMaster University, Canada

INTRODUCTION

Information systems that link businesses for the purpose of inter-organizational transfer of business transaction information (inter-organizational information systems or IOIS) have been in use since the 1970s (Lankford & Riggs, 1996). Early systems relied on private networks using *electronic data interchange* (EDI) or *United Nations* EDIFACT standards for format and content of transaction messages. Due to their cost and complexity, the use of these systems was confined primarily to large companies, but low cost Internet commercialization has led to much more widespread adoption of IOIS. Systems using the Internet and the *World Wide Web* are commonly referred to as B2B (business to business) systems, supporting B2B electronic commerce.

Technological innovations have led to several forms of B2B Internet implementations, often in the form of online exchanges or electronic marketplaces (Wang et al., 2005). These are virtual marketplaces where buyers and sellers exchange information about prices, products, and service offerings, and negotiate business transactions. They are major components of the supply chains that they support. In addition to substituting proprietary lines of communication, emerging technologies and public networks have also facilitated new business models and new forms of interaction and collaboration in areas such as collaborative product engineering or joint offerings of complex, modularized products. During the years 1999-2001, a number of online exchanges were introduced, but many of these failed (Gallaugher & Ramanathan, 2002) due mainly to an inability to attract participating business partners, but also because potential participants and their business partners did not perceive enough value added through the significant investment they required. Those that have survived are often owned by companies or consortia that are also exchange customers or suppliers.

The objective of this overview is to describe the evolution and the characteristics of B2B Internet implementations, and to discuss management considerations, the evaluation, and adoption of B2B applications, and the technical infrastructure supporting these systems. We also indicate some of the open issues that remain as the technology and its adoption continues to evolve.

BACKGROUND

Although there are many classification schemes available for B2B online exchanges (Choudhury 1997; Kaplan & Sawhney 2000), we will use a more generic and functional focus, with three categories: sell-side, buy-side, and neutral/market-type applications (Archer & Gebauer, 2001). Early B2B sell-side applications featured online catalogs, made available to the Internet community by distributors and manufacturers, often complemented by features such as shopping baskets and payment functionality. Many now provide customized and secure views of the data, based on business rules from contract agreements with individual customers. In some cases, buying processes of the customers are supported, including features such as approval routing and reporting. While sophisticated applications exist to support collaborative functionalities among the participating organizations, such as forecasting or configuration of complex products, many sell-side systems handle only the simpler transactions such as maintenance, repair, and operation (MRO) supplies. Recently, features that support collaboration have become more widely available through both the "vertical" links of supply chain management and the "horizontal" links of buying groups that can represent suppliers or buyers (Wang & Archer 2007a).

Buy-side applications support procurement, moving order processes closer to the end user, and alleviating structured workloads in functional departments such as purchasing and accounts payable. For smaller companies, an affordable alternative is to work through hosted solutions using Internet browsers to access procurement functionality provided by a third party vendor or application service provider (ASP). Functionalities beyond the automation of highly structured procurement processes include production tendering and multi-step generation requests for proposals, as they are relevant for the procurement of freelance and management services. Interfacing purchasing systems to internal systems such as enterprise resource planning systems (ERP) makes it possible to automate commonly used transactions, thus greatly increasing processing speed and reducing costs. Buy-side solutions that involve long-term inter-organizational relationships are typically set up by the purchasing organization, which then controls catalog content, data format, and backend system functionality. Benefits include

a reduction in maverick buying, and freeing purchasing and accounts payable personnel from clerical work to handle more strategic tasks. Suppliers typically benefit from long-term relationships, and in many cases, the relationships between the buyer and its suppliers were in place long before online operations commenced.

The third group of applications, often referred to as B2B electronic marketplaces or hubs, can either bring together multiple buyers and sellers on an ad hoc basis involving various types of auctions, for example, or support more permanent relationships (a many-to-many equivalent to IOIS) (Wang & Archer, 2004). Those that have been more successful are likely to have been sponsored by a consortium (e.g., *GlobalNetXchange*, in the retail industry, sponsored by buying organizations, and *Global Healthcare Exchange* in the healthcare industry, sponsored by selling organizations). They may feature auctions, electronic catalogs, collaborative functionalities, and auxiliary value added functions such as industry news and online forums. The initiator typically controls the catalog content, aggregates supplier input, and provides additional functionality and standardized data access to buyers (Wang & Archer, 2007). These marketplaces may eliminate the need for market participants to link directly to their business partners, circumventing costly value added EDI network services. Their business models typically include service charges based on transaction volume and setup costs. They provide a standard for suppliers to deliver catalog content, increase flexibility if they support access to suppliers and customers outside pre-established relationships, and create customer value through competitive pressure. Participation in such marketplace solutions may also provide a low cost alternative for SMEs (small and medium enterprises), but SME adoption of such solutions is typically driven by their larger business partners who wield significant market power (Archer et al., 2003).

MANAGEMENT CONSIDERATIONS

A market assumes an intermediary role that supports trade between buyers and suppliers, including (Bailey & Bakos, 1997): (a) matching buyers and sellers, (b) ensuring trust among participants by maintaining a neutral position, (c) facilitating market operations by supporting certain transaction phases, and (d) aggregating buyer demand and seller information. Supporting the marketplace through an electronic exchange has characteristics of (Bakos, 1991): (1) cost reductions, (2) benefits increase with the number of participants, (3) potential switching costs, (4) capital investments but economies of scale and scope, and (5) significant uncertainties in benefits. Many of the management issues of B2B electronic commerce systems relate to the need to coordinate decisions and processes among multiple firms, often through differences in business processes, information systems, business models, and organizational cultures.

Early transaction cost theory recognized markets and hierarchies as the two main methods of governance for coordinating flows of goods and services. Markets such as stock exchanges coordinate the flow through supply and demand forces with price as the main coordination vehicle. Hierarchies such as production networks consist of predetermined relationships among customers and suppliers, and rely on managerial decisions to coordinate flows. There are many intermediate forms of governance such as network organizations and strategic alliances (Gulati, 1998). A common theme among all these governance structures is collaboration among the participants, but the level of collaboration varies. These levels can be described as cooperation, coordination, and collaboration (Winer & Ray, 1994). In cooperation, there is little sharing of goods, services, or expertise; coordination requires mutual planning and open communication among participants, who share resources; collaboration involves deeply synergistic efforts that benefit all parties. Collaboration at different levels between buyers and sellers are emphasized by online exchanges, but this can also take place separately among buyers and among sellers (Wang & Archer, 2004b). A recent survey of 89 online exchanges that offer collaboration services identified a range of collaboration functionalities, including "vertical" supply chain collaboration through collaborative fulfillment, private catalogues, product life cycle management, and supply chain coordination and integration. Additionally, "horizontal" functionalities may be offered in the form of buying groups, which can be classified as dealer-type, exchange-catalogue, exchange-negotiation, supplier-initiated, and buyer initiated (Wang et al., 2007b).

The growth of outsourcing arrangements and more cooperative, integrated long-term inter-organizational relationships with a relatively small number of preferred suppliers can be termed a "move to the middle" (Clemons et al., 1993). This can result in the adoption of collaborative functionalities such as CPFR (collaborative planning, forecasting, and replenishment), which may be used to support joint initiatives between large retail customers and suppliers (Holmstrom et al., 2002; Wang et al., 2004a). Distribution of market power is often an overriding factor. For example, auto manufacturers, as theirs is a concentrated industry, are likely to adopt an approach that involves long-term collaborative relationships among business partners rather than the short-term market-driven relationships that traditionally characterized this industry. On the other hand, relationships among companies in fragmented industries such as construction are typically short-term, with low levels of trust, and transactions such as online procurement are more likely to be through B2B tendering and auctions (Stein et al., 2003).

EVALUATION AND ADOPTION

The task of evaluating an electronic exchange becomes difficult when network effects are taken into account (benefits from participating are usually positively related to the number of participants). As a result of complications such as strategic necessity, dependence on the commitment of business partners, additional risk, external effects etc., the evaluation of electronic exchange adoption is much more complex than for systems deployed within organizations (Gebauer & Buxmann, 2000). B2B market mechanisms focus on four factors that favor one market mechanism over another: degree of fragmentation, asset specificity, complexity of product description, and complexity of value assessment (Mahadevan, 2003). These have a significant impact on the choice of an appropriate market mechanism for B2B interactions.

From the organizational perspective of setting up a successful B2B application, there are initiators and (potential) participants. Initiators bear the majority of the cost and risk, but on the other hand also enjoy the majority of the benefits, and they typically decide on technology infrastructure, type of systems used, corporate identity, representation of partners, and selection of participants. Success of the system depends on the participation of a critical mass of business partners. Supplier participation considerations in a buy-side solution, for example, include investments necessary to prepare and upload catalog data, integration with backend systems, training of staff, and adjustments of business processes. Depending on individual arrangements, benefits include reduced time and costs for order processing, improved customer service, increased customer reach in a globalized marketplace, and an increase in revenues from long-term and trusted customer relationships. Neutral intermediaries in such markets face a difficult balancing task, as they have to be careful to satisfy suppliers as well as buyers, and a business model must be chosen that will attract the desired participants.

Although B2B electronic marketplaces or exchanges have received a great deal of attention by researchers, their rate of adoption by business has not been high. While their aggregate transaction growth rate is growing, they were (in 2004) outranked by at least a factor of 10 in transaction volume by EDI installations, which are still firmly in place in many large corporations (Jakovljevic, 2004). By utilizing EDI over the Internet, companies and organizations also benefit from the ability to facilitate seamless bridging between XML and EDI that can now co-exist on the same infrastructure, and use common protocols to handle electronic procurement, invoicing, and logistics information. With the need to electronically exchange volume-intensive catalogue and product specification information, organizations can reduce significantly the high cost of this exchange by using Internet EDI. In addition, major system vendors (e.g., *IBM, SAP, Oracle*), that already provide enterprise systems to their long-standing corporate clients, compete with electronic marketplaces by providing an array of specialty software-based services that support supply chain and customer relation management interactions between business partners. These proprietary systems can represent obstacles to the development of effective IOIS between companies, but this problem is slowly being reduced due to consolidation of former vendor rivals, and more attention to system interfaces among these competitors.

SMEs can and often do handle B2B transactions through ecommerce solutions without fully automated transaction management systems, through hosted procurement applications. Supply side solutions with Web access can also be used as parallel and partially automated channels for larger businesses that wish to deal with small suppliers or customers. In practice, SMEs execute small numbers of transactions and may not wish to make the investment in resources, training, and internal integration required to link to their business partners (Archer et al., 2003). New technologies are beginning to make affordable solutions available to an increasing number of small and medium sized businesses and thus facilitate their integration in global supply chains. This opens a new range of possibilities for tracing product origins along the supply chain, resulting in improved logistics, better product security, and reduced health, safety and security threats (e.g., in food supply and distribution).

Motivations for joining online exchanges include (Gebauer & Raupp 2000): (a) coercion (through market power), (b) long-term commitment to business relationships and reduction of associated uncertainty, (c) subsidies to support system installations for potential business partners, and (d) general system improvements that result in improved efficiencies and effectiveness. SMEs that link to online exchanges are most likely to be motivated through pressure from their larger partners and by long-term commitments. Many use alternative interactions utilizing a combination of manual and online functions. For example, a medium-sized value-added retailer might use ad hoc purchasing procedures such as searching the Web for catalogue information on major suppliers, and then use the telephone to negotiate prices and delivery schedules (Archer et al., 2003).

TECHNICAL INFRASTRUCTURE

Software products to support B2B interactions are continuing to mature as more complex functions are added such as collaborative planning, forecasting, and replenishment, negotiation and decision support, and procurement and asset management of complex and highly customizable items and systems (Lamont, 2005; Paul et al., 2003). Linking data from many different sources, including legacy systems, through Web services (Iyer et al., 2003) has had many successes supported by the diffusion of extensible markup language

(XML). These two technologies can help to reduce start up and transaction costs as compared to traditional EDI systems. Based on XML and Web technology, a variety of standards have emerged on a more general or sectoral basis, that may assist in: automated recognition of supported business transactions, negotiation, contracting and processing of the deal, creating online dispute resolution mechanisms, signing and encrypting the contents transmitted using the Web, and more general issues such as Internet governance. The relevant international standards are being created through the *ebXML*[1] association. Technical issues are complicated by the critical role of security and confidentiality in inter-organizational settings, particularly when using public networks such as the Internet as compared to the private networks that were traditionally used for EDI implementations. This is further complicated if transactions involve international trade, unless the organizations involved subscribe to internationally recognized standards. Overcoming these obstacles may require significant investments in software, training, business process reengineering, technical support, and time, all of which favor larger corporations.

Transaction complexity is important in the adoption decision, and is determined by factors such as the number of sub-processes and organizational units that are involved, as well as their possible interactions, interdependencies, and relationships with the process environment (Gebauer et al., 2000). This in turn depends on the type of goods or services. Acquiring indirect or non-production supplies and services is the least complex type of transaction, followed by direct goods, and capital goods and other types of ad hoc purchases tends to be the most complex. A high degree of automation is economically viable only for high volume, less complex transactions. As complexity increases and volume decreases, human intervention is more likely to be needed to handle exceptions and ad hoc transactions.

B2B applications can have major impacts on inter-organizational business processes, depending on the level of IOIS integration required (Stelzer, 2001). After planning and designing the system business model and infrastructure, a careful plan of how to implement it, how to train employees, and how to adapt business processes, is the next step towards a successful project (Archer et al., 2001). Business partner adoption, catalog management, and integration with a heterogeneous system of backend applications are frequently listed as major stumbling blocks. For example, an organization could start out by reengineering and then automating a formerly inefficient process that causes long lead times and possibly frequent complaints, such as management approval of end user requests. As a next step, putting together an online catalog that contains the offerings of preferred suppliers can be useful as a first step to reduce "maverick" buying outside pre-established contracts. While the exact steps will depend on the situation within the individual firm, the stepwise approach will also allow frequent adjustments during the project planning and implementation process, including the addition of new requirements. The adoption of a B2B eCommerce solution is a strategic company decision and it is important to evaluate the potential overall impact of this innovation on the firm before proceeding (Warkentin & Bajaj, 2003), since it may require substantial reengineering to become effective (Barnes et al., 2004).

FUTURE TRENDS

There is little doubt that rapid growth will continue in the relative value of B2B transactions handled through electronic commerce solutions, especially as they become less costly and easier to implement in SMEs. However, most of the growth in such offerings is likely to be in sell-side or buy-side online exchanges, supported by infrastructures provided by major system vendors. There has been significant supplier resistance to participating in neutral market-type exchanges. Although these exchanges offer the greatest theoretical benefit because of the potential for standardized linkages among participating companies, and the collaborative functionalities offered by such systems, this does not outweigh resistance from suppliers who see declining profit margins due to transaction cost payments and competitive price bidding. Meanwhile, EDI systems continue to link many large companies, due to their reluctance to give up related investments or to change business processes to accommodate the newer solutions.

CONCLUSION

A clear understanding of the possibilities of emerging technologies is crucial to take advantage of new opportunities in the B2B marketplace. There have been failures in this environment due to a lack of consideration of the wide range of technical, managerial, and economic issues involved. No widely adopted frameworks have been developed to assist in the choice of level of integration or in reengineering boundary spanning business processes. Although technology continues to develop, it is still immature in many areas. In particular, the integration with current IT infrastructures is often extremely complex and difficult to justify for medium to low transaction rates. Meanwhile, B2B applications continue to evolve, changing the rules of the game in subtle ways, but providing fruitful areas for new development and related research.

REFERENCES

Archer, N., & Gebauer, J. (2001). B2B applications to support business transactions: Overview and management considerations. In M. Warkentin (Ed.). *Business-to-business*

electronic commerce: Challenges and solutions (pp. 19-44). Hershey, PA: Idea Group Publishing.

Archer, N., Wang, S., & Kang, C. (2003). Barriers to Canadian SME adoption of Internet solutions for procurement and supply chain interactions. MeRC Working Paper #5. Hamilton, Canada, McMaster eBusiness Research Centre.

Bailey, J., & Bakos, Y. (1997). An exploratory study of the emerging role of electronic intermediaries. *International Journal of Electronic Commerce, 1*(3), 7-20.

Bakos, J. Y. (1991). A strategic analysis of electronic marketplaces. *MIS Quarterly, 15*, 295-310.

Barnes, D., Hinton, M., & Mieczkowska, S. (2004). Managing the transition from bricks-and-mortar to clicks-and-mortar: A business process perspective. *Knowledge and Process Management, 11*(3), 199-209.

Choudhury, V. (1997). Strategic choices in the development of interorganizational information systems. *Information Systems Research, 8*(1), 1-24.

Clemons, E. K., Reddi, S. P., & Row, M. C. (1993). The impact of information technology on the organization of economic activity: The "move to the middle" hypothesis. *Journal of Management Information Systems, 10*(2), 9-35.

Gallaugher, J. M., & Ramanathan, S. C. (2002). Online exchanges and beyond: Issues and challenges in crafting successful B2B marketplaces. In M. Warkentin (Ed.), *Business to business electronic commerce: challenges and solutions* (Chapter III). Hershey, PA: Idea Group Publishing.

Gebauer, J., & Buxmann, P. (2000). Assessing the value of interorganizational systems to support business transactions. *International Journal of Electronic Commerce, 4*(4), 61-82.

Gebauer, J., & Raupp, M. (2000). Zwischenbetriebliche elektronische katalogsysteme: Netwerkstrategische gestaltungoptionen und erfolgsfaktoren (Interorganizational electronic catalogs: Strategic options and success factors) - in German. *Informatik Forschung und Entwicklun, 15,* 215-225.

Gulati, R. (1998). Alliances and networks. *Strategic Management Journal, 19*(4), 293-317.

Holmstrom, J., Framling, K., Kaipia, R., & Saranen, J. (2002). Collaborative planning forecasting and replenishment: New solutions needed for mass collaboration. *Supply Chain Management, 7*(3), 136-145.

Iyer, B., Freedman, J., Gaynor, M., & Wyner, G. (2003). Web services: Enabling dynamic business networks. *Communications of the Association for Information Systems, 11,* 525-554.

Jakovljevic, P. J. (2004). EDI versus XML: Working in tandem rather than competing? Retrieved August 13, 2007, from http://www.technologyevaluation.com

Kaplan, S., & Sawhney, M. (2000). E-hubs: The new B2B marketplaces. *Harvard Business Review, 78*(3), 97-100.

Lamont, J. (2005). Collaborative commerce revitalizes supply chain. *KM World, 14*, 16-18.

Lankford, W. M., & Riggs, W. E. (1996). Electronic data interchange: Where are we today? *Journal of Systems Management, 47*(2), 58-62.

Mahadevan, B. (2003). Making sense of emerging market structures in B2B e-commerce. *California Management Review, 46*(1), 86.

Paul, J., Withanachchi, S., Mocker, R. J., Gartenfeld M. E., Bistline, W., & Dologite, D. G. (2003). Enabling B2B marketplaces: The case of GE global exchange services. In M. Kosrow-Pour (Ed.), *Annals of cases on information technology* (Vol. 5, pp. 464-486). Hershey, PA: Idea Group Publishing.

Stein, A., Hawking, P., & Wyld D. C. (2003). The 20% solution? A case study on the efficacy of reverse auctions. *Management Research News, 26*, 1-20.

Stelzer, D. (2001). *Success factors of electronic marketplaces: A model-based approach*. Ilmenau, Germany: Technische Universitat Ilmenau: 25.

Wang, S., & Archer, N. (2004a). Strategic choice of electronic marketplace functionalities: A buyer-supplier perspective. *Journal of Computer Mediated Communication, 10*(1).

Wang, S., & Archer, N. (2004b). Supporting collaboration in business-to-business electronic marketplaces. *Information Systems and e-Business Management, 2*(2/3), 271-288.

Wang, S., & Archer, N. (2007). *Business-to-business collaboration through electronic marketplaces: An exploratory study*. MeRC Working Paper #18. Hamilton, Canada, McMaster University.

Wang, S., Zheng, W., & Archer, N. (2005). The impact of Internet-based electronic marketplaces on buyer-supplier relationships. *Journal of Internet Commerce, 4*(3), 41-68.

Warkentin, M., & Bajaj, A. (2003). Continuous demand chain management: A downstream business model for e-commerce. In J. Mariga (Ed.), *Managing e-commerce and mobile computing technologies*. Hershey, PA: Idea Group Publishing.

Winer, M., & Ray, K. (1994). *Collaboration handbook: Creating, sustaining, and enjoying the journey*. Saint Paul, MN: Amherst H. Wilder Foundation.

KEY TERMS

Application Service Provider (ASP): An ASP is a service company that can support and relieve a firm from the daunting challenges of finding, hiring, inspiring, and training technical personnel to manage an application in-house. An ASP provides software applications on a pay-per-use or service basis via the Internet and leased lines.

Collaborative Planning, Forecasting, and Replenishment (CPFR): CPFR is a global, open, and neutral business process standard for value chain partners to coordinate the various activities of purchasing, production planning, demand forecasting, and inventory replenishment, in order to reduce the variance between supply and demand and share the benefits of a more efficient and effective supply chain.

Electronic Data Interchange (EDI): A standard used to govern the formatting and transfer of transaction data between different companies, using networks such as the Internet. As more companies are linking to the Internet, EDI is becoming increasingly important as an easy mechanism for companies to share transaction information on buying, selling, and trading. ANSI (American National Standards Institute) has approved a set of EDI standards known as the X12 standards. Due to the growing influence of international trade, EDIFACT, a standard developed by the United Nations and used primarily in non-North American countries, is being merged with X12 into a worldwide standard.

Enterprise Resource Planning (ERP): A business management system that can integrate all facets of the business, including planning, manufacturing, sales, and marketing, through a common database. As the ERP methodology has become more popular, software applications have been developed to help business managers implement ERP in business activities such as inventory control, order tracking, customer service, finance and human resources.

Extensible Markup Language (XML): Document type definitions that can be used to specify or describe various types of objects. When a set of these is used on the Web to describe product information, it is referred to as cXML or commerce XML. It works as a meta-language that defines necessary information about a product, and standards are being developed for cXML in a number of industries, performing a function similar to that of EDI for non-Web-based systems. ebXML is a set of international standards being developed through the international organization OASIS (organization for the advancement of structured information standards), recognized by the UN. It standardizes the exchange of Web catalog content and defines request/response processes for secure electronic transactions over the Internet. The processes include purchase orders, change orders, acknowledgments, status updates, ship notifications and payment transactions.

Inter-Organizational Information System (IOIS): (Sometimes referred to as an IOS). An automated information system, built around computer and communication technology, that is shared by two or more companies. It facilitates the creation, storage, transformation, and transmission of information across a company's organizational boundaries to its business partners.

Maintenance, Repair, and Operations (MRO): Supplies and services purchased for use internally in the company, often referred to as indirect or non-production supplies and services (such as office supplies, computer equipment and repairs, cleaning supplies, etc.) These tend to be low unit cost, low volume, off-the-shelf purchases.

Small and Medium Enterprise (SME): The definition of small and medium enterprises varies from country to country. If the definition is based on number of employees, SMEs in the U.S. have from 1 to 499 employees. The dividing line between a small and medium business is variously defined as being either 50 or 100 employees.

ENDNOTE

[1] http://www.ebxml.org/

Electronic Payment

Marc Pasquet
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Sylvain Vernois
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Wilfried Aubry
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Félix Cuozzo
ENSICAEN, France

INTRODUCTION

Money has two main forms nowadays: the fiduciary money (coins, banknotes…) and the scriptural one (electronic or virtual). To pay goods, both are used. The electronic money, one specific form of the scripting money, is more and more used everywhere in the world. Electronic payment has many particularities: specific infrastructure, equipment, and software, new forms of regulations, technical agreements, normalizations, fraud limitations…

The objective of this chapter is to present a general overview of electronic payment. The background section presents its historical evolution. In the main thrust, the chapter focuses first on the general architecture of electronic payment. Second, different authorization mechanisms for the processing of the banking transaction and for fraud prevention are detailed. Future trends stress the different research topics that should be investigated, especially concerning the SEPA program (Single Euro Payments Area), which will harmonize bank payment systems in Europe through 2012.

Figure 1. Exchange of goods

BACKGROUND

Exchanging goods is the basis of commerce. From the origin with barter (Menger, 1892) to the introduction of a valuable commodity to give the goods a value, men managed to build a trustful organization, creating the money, and later the banks and laws, to protect this new financial structure.

Paper money has been developed since the 17th century as an object without any real value but with a financial value given by trust in the emitter: the bank (see Figure 3).

Nowadays, money is often dematerialized in the payment process (Schafer, Konstan, J.A., Riedl, 2001). This can particularly happen (see Figure 4):

- when the purchase is done: an electronic payment occurs between the customer and the merchant ;
- when banks want to clear the positions they have between each other.

E-payment allows exchanging scriptural money electronically, by the use of an "identifier" that can be associated not

Figure 2. Transaction using social money

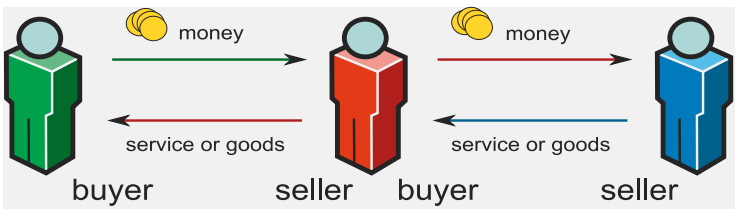


Figure 3. Transaction using paper money

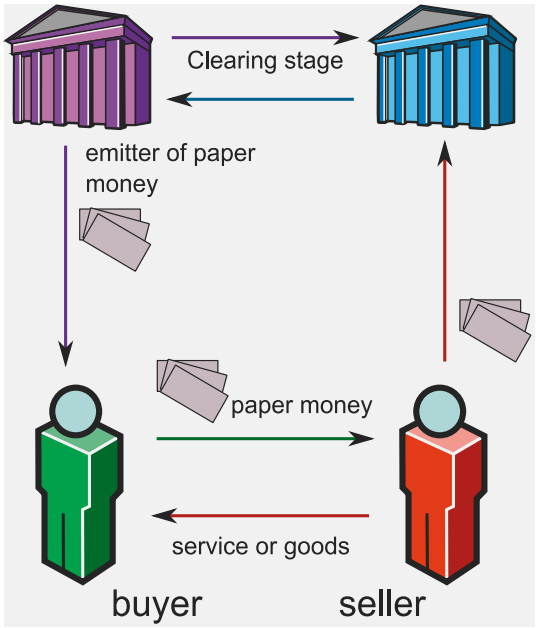
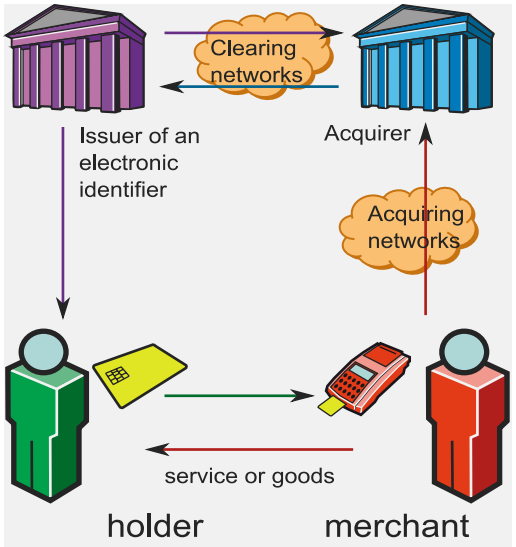


Figure 4. Transaction by electronic payment



only with a bank user or a fidelity account, but also with an electronic purse (off-line or online) or an anonymous account.

The most common media is a plastic card that includes different technologies (magnetic stripe, microprocessor chip, contactless…) allowing different services and security levels. Other objects and technologies can be used, such as customer's fingerprint, thanks to biometrics.

E-payment can be divided into two families:

- Face-to-face payment, where the customer and the merchant are physically in the same place ;
- Distant payment, where the two participants do not meet each other (mail orders, phone orders, and now electronic commerce via Internet) (see Figure 5).

With the development of recent communication technologies, new payment channels have been developed. Multimodality has become a bet for all participants in the payment process, from banks, merchants, to communication network operators.

Gradually, payment cards are allowing more services than their initial function. Open card development platforms (GlobalPlatform, 2006) of specific card specification, such as EMV (**Europay MasterCard and Visa**), are vectors of development of multiapplications cards. That brings new difficulties to solve, like the need of imperviousness between the applications embedded in a smartcard.

Historically, technical and functional payments architectures have mostly been driven by needs and industries. Some are "open" to allow different implementations to be interoperable:

- ISO/IEC norms, such as 8583 (ISO 8583) for message format or 7816 for smartcards (ISO/IEC 7816) ;
- Open specification, such as EMV specification (EMVco, 2006) that proposed an international specification for payment and multiapplication cards.

Others are kept private, to be used only by a restricted number of participants, such as:

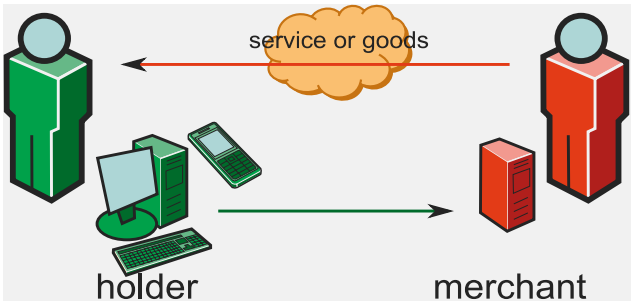
- Communication protocols between servers and terminals ;
- French smartcards specification (B4/B0') ;

However, results coming from research have allowed many evolutions. For example, electronic payment is possible thanks to cryptographic advances: the security aspects of payment are very important to accept the way the transaction is concluded. From the elementary algorithm of Lühn allowing checking the validity of card number to DES, RSA (Rivest, Shamir, & Adleman, 1978), or elliptic curves security algorithms, all aspects are interesting in the payment domain.

MAIN FOCUS OF THE CHAPTER

This section presents the state of the art on the way we can create and settle a transaction nowadays. The first part sets out the necessary exchanges for a full electronic transaction between the actors of the transaction. The second section

Figure 5. Distant payment



focuses on the authorization mechanism. The following part deals with the treatment of the transaction in a bank computer system, pointing out two schemes of transaction. Fraud prevention is the topic of the last section.

- Structure of the electronic transactions exchange

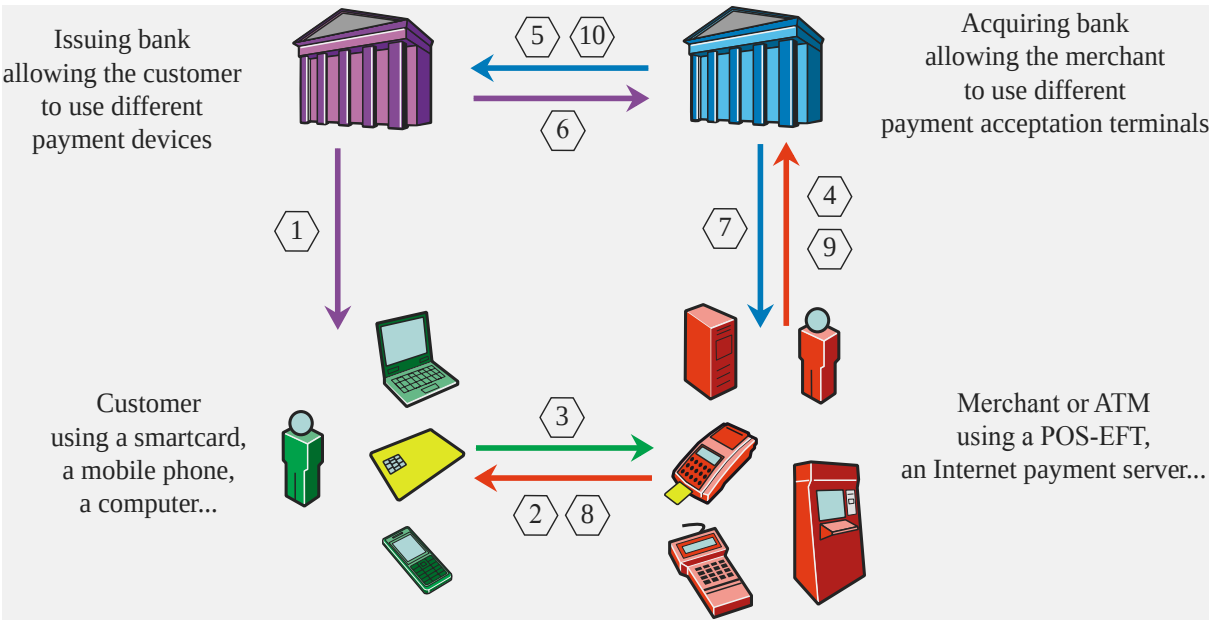
The technical mechanism of an electronic transaction is built on an old financial scheme: the trade. This involves many actors, as shown in Figure 6.

The issuing and the acquiring banks are, respectively, the banks that own the account of the customer and the merchant. Whatever is the relationship between the customer and the merchant (buying in a shop or ordering via the Internet), the scheme is always the same.

The different steps followed during a transaction are:

1. All the banks mainly propose to their account holders the use of an alternative way of payment like debit or credit card, NFC mobile phone... Thus, the first step is when the bank puts an electronic payment device at its customer's disposal;
2. Once the customer decided to buy something, the merchant starts the transaction using its point of sale or its payment server for a distant transaction;
3. The payment device is presented to the terminal that should accept this way of payment;
4. According to the type of device and terminal, each of them tries to authenticate the other. To carry out this authentication, the terminal can ask for an authorization

Figure 6. Electronic payment model



of the transaction. In many countries, this authorization is not automatically required. The authorization request is sent to the acquiring bank, which routes the message to the issuing bank;
5. The issuing bank is the only one able to make the decision to approve or refuse the transaction. The authorization request can travel through private, national, or international financial networks;
6. The issuing bank verifies the authentication of the electronic device holder (verification of the PIN code). Then, it consults all the information it owns about the accounts, the device, and the holder. Thus, the bank decides to accept or reject the transaction;
7. The answer to the authorization is sent back to the point of sale terminal. If the authorization is rejected, the transaction simply ends uncompleted;
8. With a positive answer, the transaction can be fully completed, giving the customer a receipt of the transaction;
9. Next step is the daily collection of all the transactions stored in the memory of the point-of-sale terminal;
10. Final steps are the clearing and the settlement of the transactions between the banks. The accounts of the customer and the merchant are debited and credited of the right amount.

This scheme can be shown as an off-line electronic transaction, except the authorization request, which is not compulsory. The cash withdrawal with ATM is a slightly different electronic transaction. Indeed, the ATM is online, so the authorization is compulsory and the transaction is instantaneously sent to the acquiring bank.

Authorization Mechanism

The authorization has mainly two interests during an electronic transaction:

1. By requesting an authorization, the acquiring bank makes sure that the transaction will be cleared properly;
2. The second interest is to verify that the electronic device and its holder are reliable. Regarding the EMV specification for an electronic transaction (EMVco, 2006), we can simply notice that the authorization is an online request. By this mechanism, we remove some of the weaknesses of the off-line process. Many sensible data (card number, encrypted PIN block…) join the authorization request. Thus, the issuing bank is fully able to verify all those data in order to strengthen its decision about the transaction.

As told previously, the electronic transaction is an off-line process, and requesting an authorization is not compulsory. It is the role of the electronic payment device (smartcard, NFC device…), the merchant, and the point-of-sale terminal to decide if the authorization is necessary for a specific transaction:

- The merchant, by pressing a specific key on its terminal, can decide to request an authorization;
- The electronic payment device has several data to make a decision:
  1. At random;
  2. On deviation from "standard" buying patterns calculated by the processor;
  3. According to internal parameters (floor and ceiling values);
  4. According to a specific programmed behavior (some chips are set up for compulsory authorization on every transaction).
- The point-of-sale terminal receives parameters each time it collects its transactions. Those parameters are used to decide if an authorization is necessary. By this means, the acquiring bank is able to supervise the activity of the merchant.

The authorization can also be compulsory for some specific actions like the cash withdrawal or Internet EMV payment (Van Herreweghen & Wille, 2000).

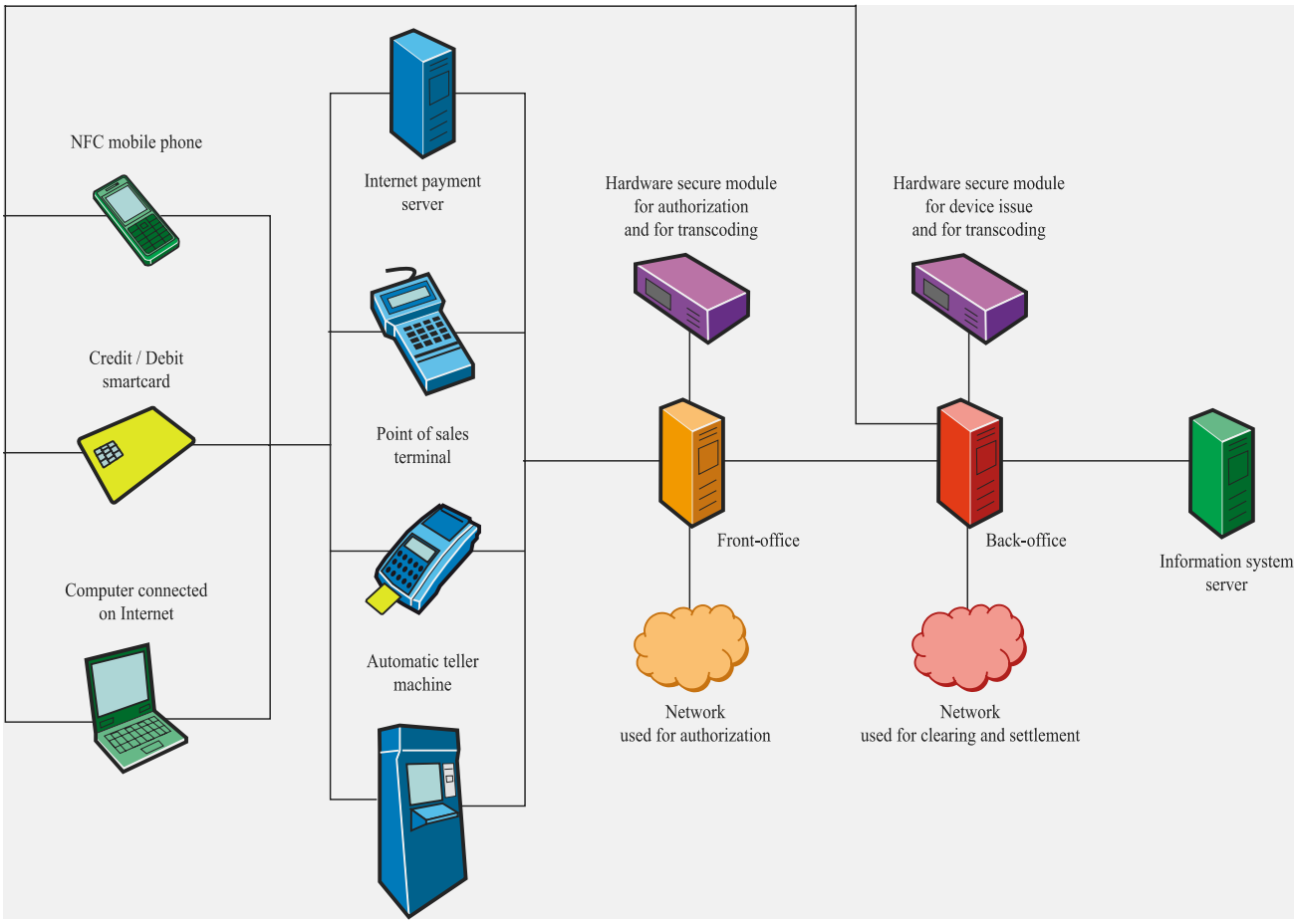
How to Watch Over and Prevent the Fraud?

Electronic commerce is a point of interest for fraud. Obviously, fraud can happen on every single part of the electronic commerce architecture (networks, devices, terminals…). Figure 7 gives an overall illustration of the real infrastructure working in the banks.

In France, an observatory about security for payment cards analyzes, every year, the evolution of the fraud. In its 2006 report (Observatory of the security cards payment, 2007), it is important to notice that the fraud rate in France, for the smartcard payments, is 0.064%. This includes many types of fraud. Most of them are absolutely not technical frauds (the use of stolen or lost cards, the use of non-upstart cards, or the use of stolen card numbers, especially for Internet payments). Only 18% of the fraud is made with falsified or counterfeited cards.

France is an interesting country because the debit smartcard is now used for more than 13 years. It is observed that the fraud rate is nearly divided by three as soon as a country widely uses smartcards for payments and cash withdrawal. For example, the United Kingdom reduced its fraud rate from 0.33 % in 1991 to 0.095 % in 2006, mainly by using EMV smartcards (APACS, the UK payments association 2007). That is one of the elements that decided the SEPA project to choose the use of the EMV specification for the

Figure 7. Technical bank infrastructure



electronic payment devices in its area (EMVco, 2005, European Payments Council, 2006). Many industrial answers to this problem are available. The first element is the setting of opposition lists. In those lists, either the authorized devices or the unauthorized ones are referenced. The main problem is the limitation of the size of the memory in the terminal and thus, the limitation of the size of the stored opposition list.

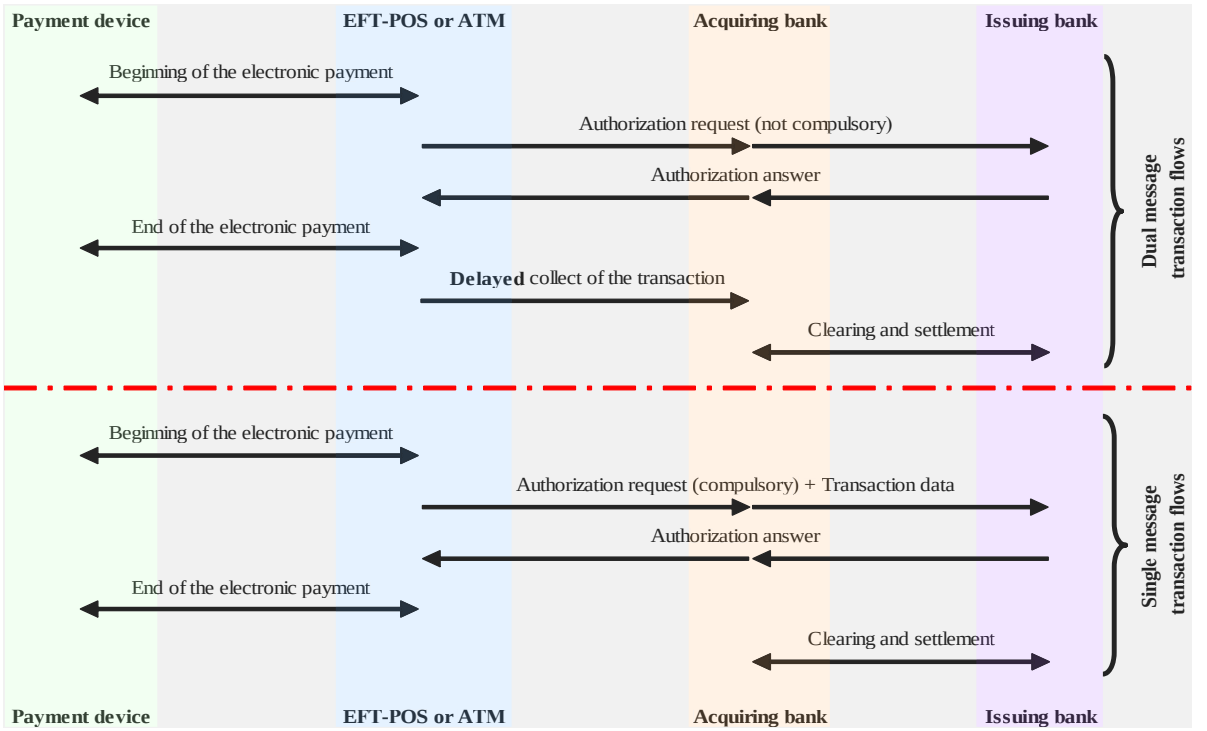
The opposition function is not enough to prevent frauds. It is also necessary to evolve the transaction treatment by decreasing the time taken for the treatment. This time lag let spaces for many frauds to happen. What seems to be a better solution is to have a full online and real-time treatment. Of course, this real-time treatment needs new structures:

- Low-cost networks accessible to the merchant to keep its terminal online;
- Compulsory authorization;
- Full treatment of the authorization (no use of delegation);
- Real-time collection of the transaction;
- Fraud control servers able to analyze, as best as possible, each transaction in real time (so it will be possible to stop a fraudulent transaction before it is ended). As for example, SAS and HSBC recently implemented a solution of real-time fraud management that is able to monitor 100% of the banks credit card transactions from more than 30 million accounts in real time. The fraud detection time is less than 30 milliseconds.

Visa (Visa Debits Processing Service, 2007) and MasterCard use, for cash withdrawal in some countries and with some banks, a function called "single message," in opposition to the "dual message" used in the architecture exposed in the former sections (see Figure 8). The principle consists in sending only one message, including all the data necessary for the authorization, the clearing, and the settlement of the transaction. Thus, there is no need to collect the transaction afterwards, which can be treated in real time just after the payment is finished.

Steve Mott (Mott, 2005), a former MasterCard executive and Norman G. Litell (Litell, 2005), a former vice president

Figure 8. Single and dual message transaction flows



of strategic planning at Visa USA, disagree about the use of the real time. Norman Litell thinks that we first should try to widely use the best secured payment solution available now, like EMV smartcards for proximity payments, and solutions like Verified by Visa or MasterCard SecureCode for distant payments (MasterCard SecureCode 2007, Verified by Visa, 2007) before choosing a full real-time architecture; certainly harder to set up and more expensive. He also pointed out that it is more important to focus on four points to well manage a financial transaction:

- Authenticating the party making the transaction;
- Assuring that proper authorization has been received from that party;
- Assuring the accuracy of all data elements required for the transaction (account number...);
- Assuring that funds will be available in the target account to settle the transaction once it clears.

FUTURE TRENDS

If the banking terminals evolved little during the last 10 years, this evolution is accelerating rapidly in many directions, and the electronic payment will be very different in 10 years from today, mainly in Europe, carried on by the SEPA.

The SEPA is a project founded by the EPC (European Payments Council) and the ECB (European Central Bank). The purpose of the SEPA is to set up a payment architecture allowing all the organizations and private individuals of the European zone unified to carry out payments within this zone, and that with all the advantages of a domestic payment (speed, low cost, and safety of the transaction). This initiative is justified with an economic aim: indeed, the strong costs of the European country-to-country transactions constituted a barrier to the development of a European banking market (European Payments Council, 2006). In particular, it did not allow a nondistorted and free competition, such as the Europe Council recommends. In the long-term, there will be no more distinction between national payments and cross-border payments. The working program will proceed from January 1, 2008 to December 31, 2012, end of deployment of SEPA in all Europe.

These will be the main changes compared to the previous situations:

- The generalization of the use of smartcards (EMV compliant);
- The replacement of the TLV messages (protocol ISO 8583), by XML messages (protocol UNIFI ISO, 20022) for all the exchanges;
- The creation of new point of interaction adapted to the new services (payment of invoice on ATM, GSM

recharging on POS…) and integrated in new networks (IPV6 for ATM, ADSL for POS, GSM for both…);

- The contactless payments with two main supports: contactless smartcards and NFC mobile phone. These devices are much more practical for consumers, and are particularly adapted in purchase environments, where the speed is essential, like in fast food, gas station, small trade supermarkets, and cinemas. The contactless smartcard and the NFC mobile phone become: transport tickets, access badge, payment system (Chanson & Cheung, 2002).
- The development of the biometric authentication. The increase of calculation and storage capacities makes it possible to consider that biometric data processing is now possible with a single chip, placed on a smartcard. Associated with contactless technology, one can imagine a growing use of biometrics for the individual authentication (Jain & Pankanti, 2006): local access control, or network access control (Hyeonjoon & Taekyoung, 2007).

CONCLUSION

The electronic payment is a young technique of only 30 years, but that technique allows everybody to realize a payment everywhere in the world with a secure process and with a very low risk of fraud. For example, even if there exist three big markets: US, Europe, and Asia, and if the three are little different in term of services (debit, credit, smartcard…), a cardholder can realize a transaction in one area with his account located in another area.

It is possible to consider electronic payment as a universal service like Internet, telephone… ; where the role of two big international actors is very important: the card schemes MasterCard and Visa. Those two companies represent the banks, and have a major impact on the electronic payment field in terms of: regulation, technical agreements, normalization, architecture, fraud fight… , but the individual banks have a very important role, in particular, in terms of development of new services. The researchers are so involved in the future trends because that sector knows a very quick evolution.

For the banking sector, the investments were, and continue to be, very high to fight against the fraud and to develop new services for customers (distant payment for e-commerce, bill payments on ATM, contactless payment…). The electronic payments will continue to grow to become, in a short term, the major system of payment in the world, before the cash.

REFERENCES

APACS, the UK payments association. (2007). *Fraud, the facts 2007: The definitive overview of payment industry fraud and measures to prevent it*. Retrieved from http://www.cardwatch.org.uk/

Chanson, S. T., & Cheung, T.-W., (2002). Design and implementation of a PKI-based end-to-end secure infrastructure for mobile e-commerce. *World Wide Web archive,* vol. 4, pp. 235 – 253. Hingham, MA: Kluwer Academic Publishers.

EMVco. (2005). *Common payment application (CPA) specification*. Retrieved from http://www.emvco.com/specifications.asp

EMVco. (2006). *Recommendations for EMV processing for industry-specific transaction types*. Retrieved from http://www.emvco.com/

European Payments Council. (2006). *Single Euro payment area cards framework, version 2.0.* Retrieved from http://www.europeanpaymentscouncil.eu/

Furnell, S. M., & Karweni, T., (2000). Security implications of electronic commerce: A survey of consumers and businesses.= *Internet Research*, *9*(5), 372-382.

GlobalPlatform. (2006). *Card specification, version 2.2*. Retrieved from http://www.globalplatform.org/specificationview.asp

Hyeonjoon, M., & Taekyoung, K. (2007). Biometric person authentication for access control scenario based on face recognition. *Lecture Notes in Computer Science, 4th International Conference on Universal Access in Human-Computer Interaction, 5*, 463-472.

ISO/IEC 7816 - *Identification cards -- Integrated circuit cards*

ISO 8583 - *Financial transaction card originated messages -- Interchange message specifications*

Jain, A. K., & Pankanti, S. (2006). A touch of money [biometric authentication systems]. *IEEE Spectrum, 43*, 22-27.

Litell, N. G. (2005). Real-time debit: A debate. *Digital Transactions*. Retrieved from http://www.digitaltransactions.net/files/1105cover.doc

MasterCard SecureCode. (2007). Retrieved from http://www.mastercard.com/securecode/

Menger, C. (1892). On the origins of money. *Economic Journal*, *2*, 239-55.

Mott, S. (2005). Real-time debit real soon, please. *Digital Transactions*. Retrieved from http://www.digitaltransactions.

net/files/ecommerce-3405.doc

Nabi, F. (2005). Secure business application logic for e-commerce systems. *Computers & Security, 24*, 208-217.

Observatory of the security cards payment. (2007). *Rapport annuel d'activité 2006*. Retrieved from http://www.banque-france.fr/observatoire/rap_act_fr_06.htm

Rivest, R., Shamir, A., & Adleman, L. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM, 21*, 120–126.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery, 5*, 115–153.

Torres, J., Izquierdo, A., Ribagorda1, A., & Alcaide, A. (2005). Secure electronic payments in heterogeneous networking: New authentication protocols approach. *Lecture Notes in Computer Science, Computational Science and Its Applications – Internet Communications Security (WICS) Workshop, 3482*, 729-738.

Van Herreweghen, E., & Wille, U. (2000). Risks and potentials of using EMV smartcards for Internet payments. *USENIX Workshop on Smartcard Technology*, Chicago, pp. 163-173.

Verified by Visa. (2007). Retrieved from http://www.visaeurope.com/personal/onlineshopping/verifiedbyvisa/

Visa Debits Processing Service. (2007). *Visa/Plus ATM Network*. Retrieved from http://www.visadps.com/products/visa_plus_atm_network.html

KEY TERMS

Acquirer: A financial institution having a business relationship with merchants, retailers, and other service provides to process their plastic card transactions.

Authorization: The process whereby a merchant requests permission for the card to be used for a particular transaction amount.

Automated Teller Machine (ATM): A computerized self-service device permitting the cardholder to withdraw cash from their account and access other banking services.

Card Issuer: A bank issuing payment or credit cards to its customers.

Card Scheme(s): Card schemes set the business rules that govern the issue of the payment cards that carry their logo. (Examples: Visa, MasterCard, American Express, Diners Club.)

Counterfeit Card: A device or instrument that has been printed, embossed, or encoded so as to purport to be a legitimate card, but which is not genuine.

Electronic Commerce: Transactions that are conducted over an electronic network where the buyer and merchant are not at the same physical location.

Europay MasterCard and Visa (EMV): The internationally agreed standards for chip payment cards. EMV standards are maintained by EMVCo.

E-Libraries and Distance Learning

Merilyn Burke
University of South Florida-Tampa Library, USA

INTRODUCTION

With the explosion of distance learning, academic libraries have had to change to meet the needs of their faculty, staff, and students. The ACRL (Association of College & Research Libraries) presented guidelines to help librarians manage these changes. The proliferation of articles on this topic points to the rapid acceptance of this form of education. This rapid expansion has offered interesting challenges such as providing equitable services for all students, and greater assistance to faculty in supporting their classes. How libraries respond to these challenges will impact the success or the failures of these programs.

BACKGROUND

Historically, distance learning or distance education began as little more than "correspondence courses," which promised an education in one's own home as early as 1728 (Distance Learning, 2002). By the 1800's the concept of distance education can be found in England, Germany, and Japan (ASHE Reader on Distance Education, 2002).

In 1933, the world's first educational television programs were broadcast from the University of Iowa and in 1982, teleconferencing began, (Oregon Community Colleges for Distance Learning, 1997) often using videotaped lectures, taped-for-television programs and live programming, adding a human dimension. Students and faculty were now able to interact with each other in real time; enhancing the learning process by allowing student access to teachers across distances.

By 2006, e-learning is incredibly mainstream, no longer relegated to a sideline position in higher education; e-learning earned its own berth in U.S. News & World Report with an annual guide not unlike the college & university edition. (U.S. News & World Report, October 16, 2006). Not only has learning gone online, so have the textbooks and other information sources. Libraries must respond to the pressures and needs of these students or become irrelevant.

ACADEMIC LIBRARIES AND DISTANCE LEARNING

Distance learning can be defined by the fact that the student and the instructor are separated by space. The issue of time is moot considering the technologies that have evolved allowing real time access. Today, universities around the world use various methods of reaching their remote students. With the use of technology, access becomes possible, whether it is from campuses to remote sites, or to individuals located in their own homes or even the dorms on campus.

The development of course instruction, delivered through a variety of distance learning methods (e.g., including Web-based synchronous and asynchronous communication, e-mail, and audio/video technology) has attracted major university participation (Burke, Levin, & Hanson, 2003). These electronic learning environment initiatives increase the number of courses and undergraduate/graduate degree programs being offered without increasing the need for additional facilities.

During the 2000-2001 academic year, the NCES (National Center for Education Statistics) estimated in the United States alone there were 3,077,000 enrollments in all distance education courses offered by 2-year and 4-year institutions with an estimated 2,876,000 enrollments in college-level, credit-granting distance education courses, with 82% of these at the undergraduate level. (Watts, Lewis, & Greene, 2003, p. 4). Further, the NCES reported that 55% of all 2-year and 4-year U.S. institutions offered college-level, credit-granting distance education courses, with 48% of all institutions offering undergraduate courses, and 22% of all institutions at the graduate level (ibid, p. 4). It is clear that distance education has become an increasingly important component in many colleges and universities, not only in the United States, but also worldwide.

Although educational institutions create courses and programs for distance learners, they often omit the support component that librarians and accrediting organizations consider critical. It is recommended that courses be designed to ensure that students have "reasonable and adequate access to the range of student services appropriate to support their learning" (WICHE, Western Interstate Commission for Higher Education). Further, course should incorporate

information literacy skills within the course or in class assignments to ensure skills for lifelong learning (American Library Association, 1989; Bruce, 1997). In addition, the Association of College & Research Libraries, ACRL, issued guidelines for distance learning library services that were approved in June, 2004 that update various guidelines that were developed beginning in 1963 for "extension students" (Guidelines for Distance Learning Library Services, ALA 2006.)

Distance learning (DL) students are unlikely to walk into the university's library for instruction on how to use the resources, from print to electronic journals, as well as services such as electronic reserves and interlibrary loan. The elements of any successful distance-learning program must include consideration of the instructors and the students, both of whom have needs that must be examined and served.

With imaginative use of technology, libraries have created "chat" sessions, which allow 24/7 access to librarians who direct students to the resources that are available online or through interlibrary loan. In addition, librarians assist faculty in placing materials on electronic reserve so that their students can access the materials as needed. Libraries have become willing to provide mail services and desktop delivery of electronic articles to their distance learning students and, when that is not possible, refer their students to local libraries to take advantage of the interlibrary loan system. Online tutorials have been created to help students learn how to access these resources, while other libraries have specific departments that assist their distance education students and faculty. The role of the library in this process is one of support, both for the students and the faculty.

Of all of the "traditional" library functions such as materials provision, electronic resources, and reciprocal borrowing available to the distance learner, there remains a significant gap in service, that of reference. Although chat lines and other 24/7 services are available, these services simply do not provide the DL student the same quality of service that the on-campus student gets when he or she consults with a librarian in person. Newer versions of distance learning course software provide external links to resources, but do not yet include reference service by email and live chat sessions with librarians in their basic packages. It will continue be the responsibility of the library to make these services easily available and known to the distant learner whose contact to the institution may not include information about the library and its resources. Proactive planning by the library with those who are responsible for distance education can ensure that the students are made aware of what is available for them in the library.

Recently, libraries have been looking at e-commerce business models as a functional way to serve their clientele in reference services, as today's "customers" are savvier and businesses have become more sophisticated in responding to customer's needs. Libraries can use these models to provide the service for DL's whose level of skills has risen with the increased use of the Internet. Coffman (2001) discusses the adaptation of such business tools as customer relations' management (CRM) software such as the Virtual Reference Desk, Webline, NetAgent, and LivePerson. These programs are based upon the "call center model," which can queue and route Web queries to the next available librarian. A quick visit to the LSSI Web site (Library Systems and Services, L.L.C, http://www.lssi.com) allows a look into the philosophy of offering "live, real-time reference services." LSSI's "virtual reference desk" allows librarians to "push" Web pages to their patron's browser, escort patrons around the Web, and search databases together, all while communicating with them by chat or phone" (www.lssi.com). Many of these systems provide the capability to build a "knowledge base" that can track and handle a diverse range and volume of questions. These collaborative efforts, with a multitude of libraries inputting the questions asked of them and creating FAQs (frequently asked questions lists), provide another level of service for the distance learner (Wells & Hanson, 2003).

These systems have great potential, and while they show tremendous possibilities, they need more work to make them more functional for library use. Chat sessions are problematic when the patron is using his or her phone line to connect to the computer, and libraries must look to the emerging technology to find solutions to such issues to prevent becoming obsolete.

Another direction is the development of "virtual reference centers," which would not necessarily have to be located in any particular physical library. Current collaboratives among universities have created consortial reference centers accessible anywhere and anytime. The reference center librarian could direct the student to the nearest physical resource or to an online full-text database based upon the student's educational profile (e.g., university, student status, and geographic location). Although the physical library may indeed become a repository for books and physical items, the reference component may no longer be housed within that particular building.

An example of support is Toronto's Ryerson Polytechnic University (Lowe & Malinski, 2000) infrastructure, which is based upon the concept that, in order to provide effective distance education programs and resources, there must be a high level of cooperation between the university, the departments involved, and the library. At Ryerson, the continuing education department studied what types of support the students needed and identified technical, administrative, and academic help as three major areas of concern. Technical help was assigned to the university's computing services, administrative help was available on the Web and through telephone access, and academic help included writing centers, study skill programs, and library services. Ryerson's philosophy encompassed the concept that synchronization of all these components would assist in making the student's

experience richer and give the student a higher degree of success. Their report shows an interesting view of librarians working to redefine their roles and participate in an important and exciting reference service to their distance learning population. (http://www.ryerson.ca/continuing/distance/)

The library and the distance education unit worked to provide connectivity to resources that were important to the classes being taught online or at-a-distance. It is these types of library involvement that can make distance learning an even more successful and enriching experience. When a university system, as a whole, embraces a collaboration of all its components, both the students and the university reap the rewards.

FUTURE TRENDS

As distance learning continues to flourish, research will be needed to examine the effective implementation and ongoing management of distance education. While several issues emerge as salient such as the social aspects of communication in the networked environment, and the integrity of Web-based course resources, it is the role of libraries in support of distance education that must be considered. Recent advances in groupware technologies have enhanced an individual's ability to stay connected for both work and social exchange through the use of synchronous and asynchronous remote communication and the previous concern of isolation has been all but forgotten (Li, 1998; Watson, Fritz et al., 1998). However, the increased use of technology suggests that formal and extensive training on both distance technology and team communications are necessary (Venkatesh & Speier, 2000).

Libraries, often overlooked in this process, are working to be far more assertive in the distance learning process. Libraries can be a center of technical and administrative help along with the traditional academic role that they have normally held. The growing DL field allows librarians to re-define their roles, and request monies for advanced technological necessary to become as "virtual" as the classes being taught. In addition, to serve the ever-increasing DL population, library education must now include the course work that will provide future librarians the training necessary to serve this ever-expanding population.

CONCLUSION

Distance education will continue to grow. In order to support this educational initiative, academic libraries must establish a supporting framework and commitment to those services traditionally provided by libraries such as lending books and answering reference questions in person or by telephone, plus new services such as "live chat" and desk top delivery of articles that are unique to the virtual environment. Faculty and students in distance learning courses should be able to depend on the academic library for their resources and services, and the library must be able to deliver materials to students or assist them in finding alternate sources in a timely manner, otherwise the students and faculty will seek other sources of materials. Libraries need to be able to identify and assist their DL students. Help desks, chat rooms, blogs, email programs, and live reference all contribute to the support of the distance learning programs. Since DL students may never visit a library's physical facility, it is important to provide information on how best to access the library virtually.

Faculty members also require library support for their courses. For example, materials may be scanned or digitized and placed on the Web, in a content management program, or videos may be "streamed" for online access. In order to digitize and make these items accessible, faculty members need information on the correct use of copyrighted materials. It is also important to put into place an action plan to implement a program for distance learning and a method for assessing that program once it is in place.

REFERENCES

American Library Association. (1989). *Presidential committee on information literacy*. Final Report. Chicago: The Association.

Bruce, C. (1997). *Seven faces of information literacy*. Adelaide, South Australia: AUSLIB Press.

Burke, M., Levin, B. L., & Hanson, A. (2003). Distance learning. In A. Hanson & B. L. Levin (Eds.), *The building of a virtual library* (pp. 148-163). Hershey, PA: Idea Group Publishing.

Coffman, S. (2001). Distance education and virtual reference: Where are we headed? *Computers in Libraries, 21*(4), 20.

Distance Education (2006, September). In Wikipedia, the free encyclopedia. Retrieved on December 14, 2006, from http://en.wikipedia.org/wiki/Blog

Distance Learning. (2002). *1728 advertisement for correspondence course*. Retrieved March 8, 2002, from http://distancelearn.about.com/library/timeline/bl1728.htm

Guidelines for Distance Learning Library Services. (2006). *American Library Association*. Retrieved from http://www.ala.org/acrl/resjune02.html

Kingsbury, A., & Galloway, L. (2006, October 16). *Education online*. U.S. News & World Report, Volume 141, issue 14, p. 62-72. Retrieved from http://www.usnews.com/usnews/edu/elearning/articles/1007classtech.htm

Li, F. (1998). Team-telework and the new geographical flexibility for information workers. In M. Igbaria, & M. Tan (Eds.), *The virtual workplace* (pp. 301-3118). Hershey, PA: Idea Group Publishing.

Lowe, W., & Malinksi, R. (2000). Distance learning: Success requires support. *Education Librarie*s, *24*(2/3), 15-17.

McConnell Funding Project Final Report: "A Digital Reference Service for a Digital Library: Chat Technology in a Remote Reference Service". Diane Granfield, Principal Investigator, May 15, 2002. (www.ryerson.ca/library/ask/McConnell.pdf)

Oregon Community Colleges for Distance Learning. (1997). *The strategic plan of the Oregon community colleges for distance learning, distance learning history, current status, and trends*. Retrieved March 8, 2003, from http://www.lbcc.cc.or.us/spoccde/dehist.html

Sittler, R. L. (2005). Distance education and computer-based services: The opportunities and challenges for small academic libraries. *Bookmobiles and Outreach Services, 8*(1), 23-35. Retrieved November 19, 2006, from Library Literature & Information Science database.

Venkatesh, V., & Speier, C. (2000). Creating an effective training environment for enhancing telework. *International Journal of Human Computer Studies, 52*(6), 991-1005.

Watson Fritz, M., Narasimhan, S., & Rhee, H. (1998). Communication and coordination in the virtual office. *Journal of Management Information Systems, 14*(4), 7-28.

Watts, T., Lewis, L., & Greene, B. (2003). Distance education at degree-granting postsecondary institutions: 2000-2001. Washington, D.C.: National Center for Education Statistics. [NCES 2003-017]. [Also available as an electronic http://nces.ed.gov/pubs2003/2003017.pdf].

Wells, A. T., & Hanson, A. (2003). E-reference. In A. Hanson & B. L. Levin (Eds.), *The building of a virtual library* (pp. 95-120). Hershey, PA: Idea Group Publishing.

WICHE (Western Cooperative for Educational Telecommunications). Balancing Quality and Access: Reducing State Policy Barriers to Electronically Delivered Higher Education Programs. [Electronic document]. Retrieved September 2, 2003 from http://www.wcet.info/projects/balancing/principles.asp

KEY TERMS

Asynchronous Communication: Is when messages are exchanged during different time intervals (e.g., e-mail).

Blog: A blog is a Web site where entries are made in journal style and displayed in areverse chronological order. Blogs often provide commentary or news on a particular subject. A typical blog combines text, images, and links to other blogs, Web pages, and other media related to its topic. The ability for readers to leave comments in an interactive format is an important part of many blogs. Most blogs are primarily textual although some focus on photographs, videos, or audio (podcasting), and are part of a wider network of social media.The term "blog" is derived from "Web log." "Blog" can also be used as a verb, meaning to maintain or add content to a blog.

Chat: A realtime conferencing capability, which uses text by typing on the keyboard, not speaking. Generally, between two or more users on a local area network (LAN), on the Internet, or via a bulletin board service (BBS).

CRM (Customer Relationship Management): This term refers to how a company interacts with its customers, gathers information about them (needs, preferences, past transactions), and shares this data within marketing, sales, and service functions.

Desktop Delivery: Using electronic formats to send articles to users.

Distance Learning/Distance Education: Taking courses by teleconferencing or using the Internet (together with e-mail) as the primary method of communication.

Electronic Reserves: The electronic storage and transmission of course-related information distributed by local area networks (LANs) or the Internet. Also known as e-reserves, in addition to displaying items on a screen, printing to paper, and saving to disk are often allowed.

Internet: A worldwide information network connecting millions of computers. Also called the Net.

Link-Rot: The name given to a link that leads to a Web page or site that has either moved or no longer exists.

Next Generation Internet (NGI): Currently known as Abilene, the next generation Internet refers to the next level of protocols developed for bandwidth capacity, quality of service (QOS), and resource utilization.

Real-Time: Communication, which is simultaneous; see Synchronous.

Social Aspects of Communication: A social process using language as a means of transferring information from one person to another, the generation of knowledge among individuals or groups, and creating relationships among persons.

Streaming Video: A technique for transferring data as a steady and continuous stream. A browser or plug-in can start displaying the data before the entire file has been transmitted. Synchronous and asynchronous communication: Synchronous communication is when messages are exchanged during the same time interval (e.g., Instant Messenger™).

Virtual Library: More than just a means of collocating electronic resources (full-text materials, databases, media, and catalogues), a virtual library also provides user assistance services such as reference, interlibrary loan, technical assistance, etc.

Voice Over Internet protocol (VoIP): A protocol that enables people to use the Internet as the transmission medium for telephone calls.

Web (World Wide Web): A global system of networks that allows transmission of images, documents, multimedia using the Internet.

E-Logistics: The Slowly Evolving Platform Undrepinning E-Business

Kim Hassall
University of Melbourne, Australia

INTRODUCTION

By 1998, arguably some four years after the Internet's general user beginnings, many commentators did not doubt that Internet based home shopping was on its way to revolutionize our lives. At the margin, it certainly allowed us another purchasing channel and for many retailers some 5% to 12% of differing goods is now done through an "e-store" or "e-marketplace". (Visser & Hassall, 2005). However, by 2001 a range of major e-business summits, perhaps very notable being the 44 nation OECD hosted e-transport and e-logistics summit in Paris (June, 2001), was beginning to demolish the euphoria of B2C. In its basic state, B2C was a very marginal business. But what of B2B? Yes, it is a bigger sector but how were the business rules and logistics strategies shaping up for network design, e-marketplace use, and logistic fulfilment changing when compared to the rapidly evolving B2C environment? The ICT sector rapidly began to assemble a host of B2B applications for Supply Chain Management and despite the "tech wreck" occurring towards the end of 2001, these highly expensive suites of products found some traction over the next three to four years. So, initially, the development of large logistics software packages such as I2, Baan, Descartes, and so forth, were offerings that the B2B sectors availed themselves of. However, besides the ICT developments in the B2B space, the evolution of new logistics strategies would prove themselves to be good, bad, and various shades in between, when examining the full end to end (E2E) e-business operations. Since 2001, a tide of interest has turned towards the adoption of fit for purpose e-logistic models to support the end to end functionality of e-business. Hassall (2003) describes a detailed survey for the international Postal Authorities as to what new e-logistics and e-business strategies should be developed. These ranged from new householder delivery choices, to global e-marketplaces being developed. Why this survey was important was because the global postal authorities are the largest combined B2C operator and also a growing B2B logistics supplier.

The Tools of E-Logistics

The staple of the world's logistics is activated by orders generated by the use of the phone and the fax machine. This is true for small/medium enterprises (SMEs), small

Table 1. The tools for e-logistics

E-logistics Tool	Description
e-ordering	Via Web, e-market, auction, collaborative system, etc.
EDI Requirements	Optional – dependent on contract requirement
Activated order/Shipment number	Usually an imperative requirement. Various Generators
Activation of Logistics services -order/pick/pack -despatch -transport, etc.	Activation of a specific set or single operation from warehouse, transport operator, delivery agent etc through shipper, broker, customs agent, and/or sub-contractor or own fleet
Barcode or RFID scan	Optional – dependent on requirement
Track and Trace capability	Optional – dependent on requirement
Call Centre CRM ability	Optional – dependent on requirement
Automatic Logistic Performance calculator	A rare but powerful tool. Can save many hours per week in evaluating if functionality is available.
Client Accounting	Commonly an e-market and portal offering,
Quarterly reporting	Specified financials/service performance or customers, etc.

Source: Hassall (2005)

office/home office (SOHOs) and Medium Enterprises (MEs) involved in B2C, b2b (small business to small business) or b2B (small business to large business). In many ways it will be the customer requirements that eventually force the smaller enterprises into adopting the use of further enhanced Web based products so that the information flow and reporting of their product orders or dispatches can feed customer or client information systems. B2B logistic contracts will often have a predefined set of software systems in place for reporting, monitoring, and accounting. Usually these will be more expensive than the suite of systems that the SMEs, SOHOs, and so forth, will have at their disposal.

The above list of e-logistic options is a list of capabilities that either the customer may require, or the logistics supplier offers. It would be quite unusual for many major 3PLs (Third Party Logistic Providers) to supply all of these capabilities unless directed to, usually by the decree of a major client. However, a subset of these strategies ought to be examined by the supplier or the e-logistics provider fulfilling the service.

The Evolution in B2C Logistics

The evolution of B2C from the Christmas mishaps in 1999 to now has been to achieve a cheap and successful delivery by the delivery agent. This statement is true but another dimension to the home delivery is trying to minimize the problems associated with product returns, and products being taken back to the delivery depot. That is, home delivery is also aware of the problems of "reverse logistics," which range from 2% returns for household chemicals to 50% returns for magazines. (Bayles, 2001). Reverse logistics is a large cost burden and, in fact, integral to the physical and environmental cost of the B2C operation. (Sarkis, Meade, & Talluri, 2004). Generally, the full planning and operational capability required for reverse logistics has even spawned several specialist providers in this area. (Poirier & Bauer , 2001). However, is a better way to minimize the reverse logistic operations to have the customer pick up the item? This may minimize some aspects of reverse logistics, but it may not be a winner in the area of customer satisfaction. Certainly delivering to a retail agent is a large cost benefit for the delivery agent. One drop of a hundred parcels to a retail agent is a lot cheaper than attempting delivery to one hundred households. But perhaps the delivery dump at the retail partner is not the choicest alternative for the majority of customers.

New strategies outlined in Table 2 are, for example, the electronic home parcel box (Number 2) which is just progressing beyond the R&D stage. In Europe this method of delivery is being discussed in regard to new planning regulations and this strategy may be a significant strategy within ten years. One way retailers are experimenting with loading for household delivery is directly out of their normal retail premises, not from a distribution centre. This Strategy (Number 8) is employed by such retailers as Tesco. This strategy may negate the need for a separate loading centre but what happens when 100 commercial vehicles arrive to load at the same time slot? Answer: Severe queuing and a valuable loss of time for the delivery agents.

However, for a wholesaler with a diverse enough range of products, an entire retail operation could, in theory, be by-passed in a home shopping environment. The wholesaler takes orders, picks the orders from a central warehouse, then undertakes the delivery of these orders directly to households from the warehouse. The benefits to customers could see a substitute to a retail price which would be now made up of a wholesale price, plus a transport cost, and a small margin. This could be cheaper than the retail purchase price. Many major retailers, however, offer both services to span both the customer shopping and home delivery requirements.

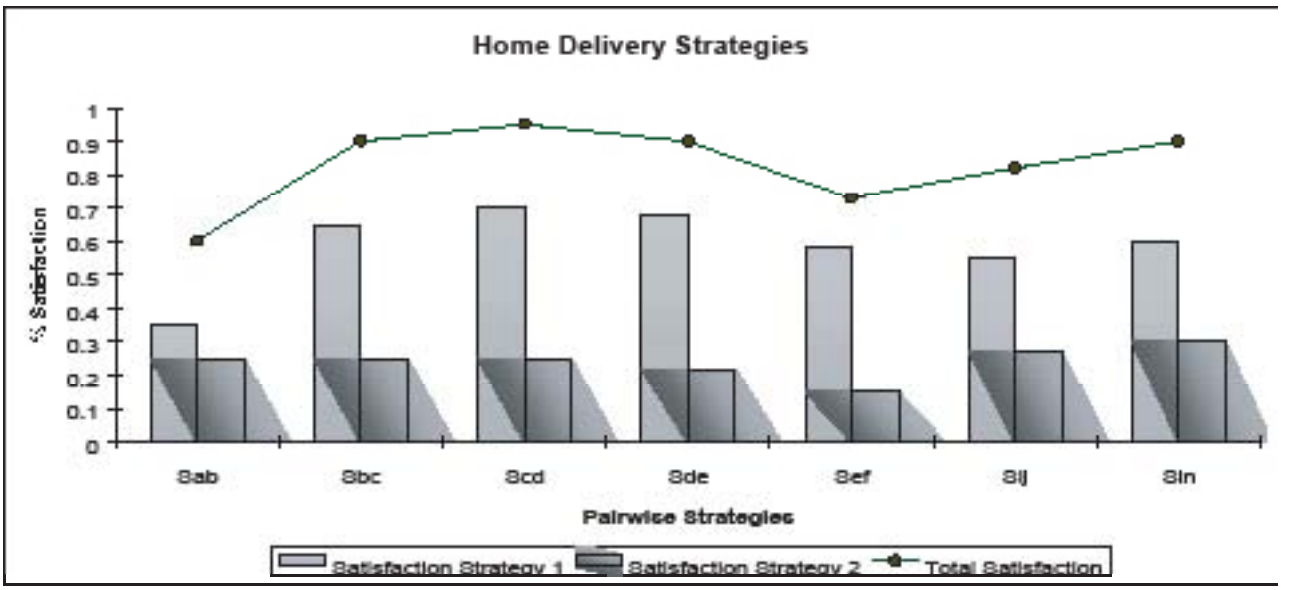
Customer Fit in the New B2C Pairwise Strategy Models

Some of these new strategies, listed in Table 2, are geared towards a high degree of cost minimization for the delivery operator. This inherently may not be a bad thing, however, where does the customer rate in the strategy? More importantly, customer response surveys may indicate exactly what proportion of the customers are happy and unhappy with the offered delivery strategies. If the surveyed service response rating exceeds 85% or 90%, then that one strategy may be worth keeping for that delivery agent for that class of home-shopping products. One survey conducted in 1998 (Hassall, 2000) suggested that about 12% of electronic order forms allowed for alternative delivery instructions. Limiting alternative delivery strategies will hardly reflect a high level of customer satisfaction. However, allowing a limited set of delivery options may add significantly to the home shoppers' level of satisfaction.

Figure 1 suggests that the combination of "hypothetical generic pairs" of strategies will give at least an equivalent level of cumulative satisfaction. Seven pairs of strategies are hypothetically displayed. For any particular retailer and delivery agent it is a matter of examining what pairwise options are feasible and customer friendly. It may only be that two single and two pairwise options are feasible. Perhaps even one direct hop strategy and one pairwise strategy is feasible for the particular commodity purchased, but it is certainly part of the operation that the retailer is aware of the customers' highest preferences for the particular delivery options either offered or not offered. For example, the delivery of backyard furniture may very much limit the feasibility of selecting from at most a small number of the 13 options listed in Table 2.

Why are pairwise delivery strategies important? As stated above pairwise strategies can reflect higher levels of customer

Figure 1. Pairwise home delivery strategies and satisfaction ratings



(Source: Hassall, 2001b)

satisfaction, but also they can be used to lower the marginal delivery cost of a single strategy. Another benefit for pairwise delivery strategies is that they can add a considerable level of security for particular commodities. For example, a business consultant takes advantage of a discount sale on a new desktop computer model. Instructions are to attempt after hours home delivery on a particular evening. If a computer agent sent the computer to the domestic householder upon nobody at home it will be returned to a safe storage site for customer pickup. For small items, a return to a local 24-hour convenience retail agent, or to a retail agent near a day office may be an alternative. However, boxed computers are not a small item, they are valuable as well, and a small retail agent may not wish to store such a large single cubic consignment. This pairwise strategy, therefore, is performed at smaller cost than attempting second delivery, or telephoning to arrange a second delivery.

One recent international survey of B2C delivery strategies across 16 countries suggested that the offering of a diversity of delivery options was the second most important strategic issue facing incumbent B2C operators (Hassall, 2003).

Reynolds (2004) describes six e-fulfilment models, however, when compared to Table 2, these are somewhat more macro in their descriptions.

B2B Fulfillment and the Appropriate E-Logistic Solutions

B2B is simple in concept but the sub-classifications of B2B are quite large. Many B2B transactions do not generate freight *per se* and this is why generally e-logistics is far more relevant to e-business generating physical product.

At the OECD/ECMT e-transport and e-logistics Summit (Paris, 2001), Nemoto, Visser, & Yoshimoto (2001) described B2B as "too big" a concept for the logistical requirements of B2B. Instead classifications such a G=Government, S=Shippers, L=Logistics Operators should be subsets of the B2B space. The only classification to have become recognized in its own rite is the Governmental category G. In fact the e-business of Government has become a specialist study area in its own rite.

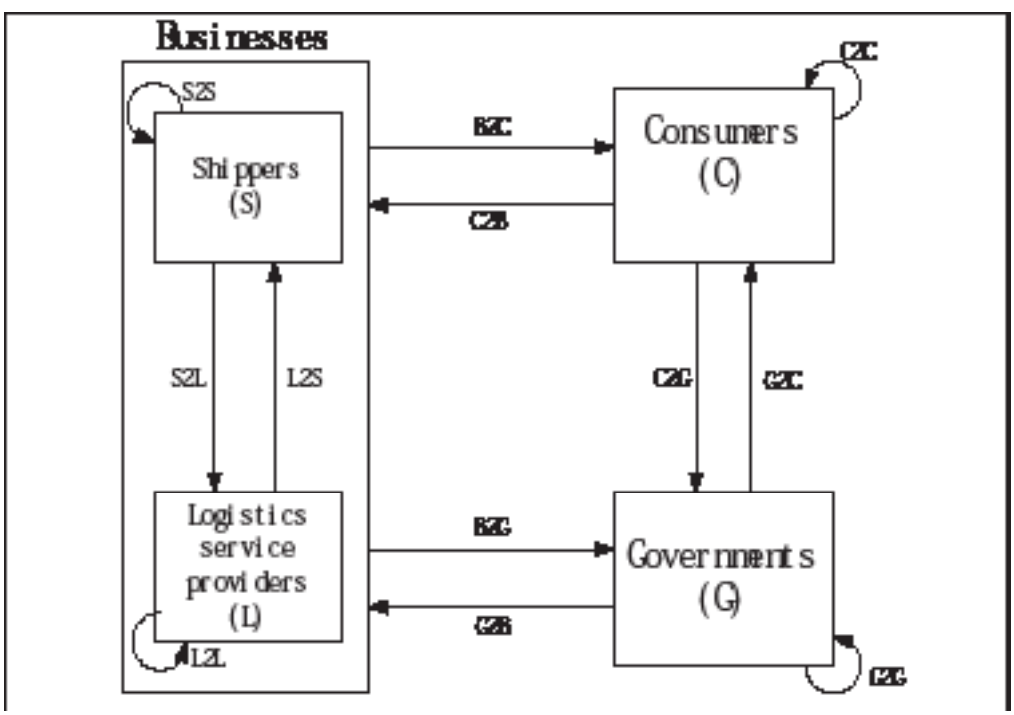
The logistic fulfilment for B2B is most often performed as a well defined set of pre-arranged business rules that are implemented between the customer and the 3PL or fulfilment agent. That is, the software requirements, service performance level, reporting levels, and invoicing are all covered by contractual arrangements. Enterprise Requirements Planning systems have been designed for the B2B space (Reynolds, 2001) but many are exceedingly costly. It will be interesting as to whether smaller "best of breed" technologies will

Table 2. B2C householder delivery strategies

13 STRATEGIES for Household delivery	
1.	Attempt 1st delivery, phone follow up for 2nd attempt.
2.	Attempt delivery to a home parcel box.
3.	Attempt home delivery, failure redirected to retail agent for customer pick up.
4.	Continual household attempt at delivery.
5.	Customer pick up from retail key-hole or kiosk site.
6.	Customer pick up from secure depot storage or common parcel box.
7.	Delivery agent to retail agent by direct drop.
8.	Delivery agents loads orders direct from retail site not a specialized distribution hub.
9.	Delivery to public provider parcel delivery box.
10.	Initial delivery to preferred post office of choice.
11.	Optional flexible delivery strategies as stated on the customer order form.
12.	Phone booking for initial delivery slot.
13.	Slotted after hours delivery.

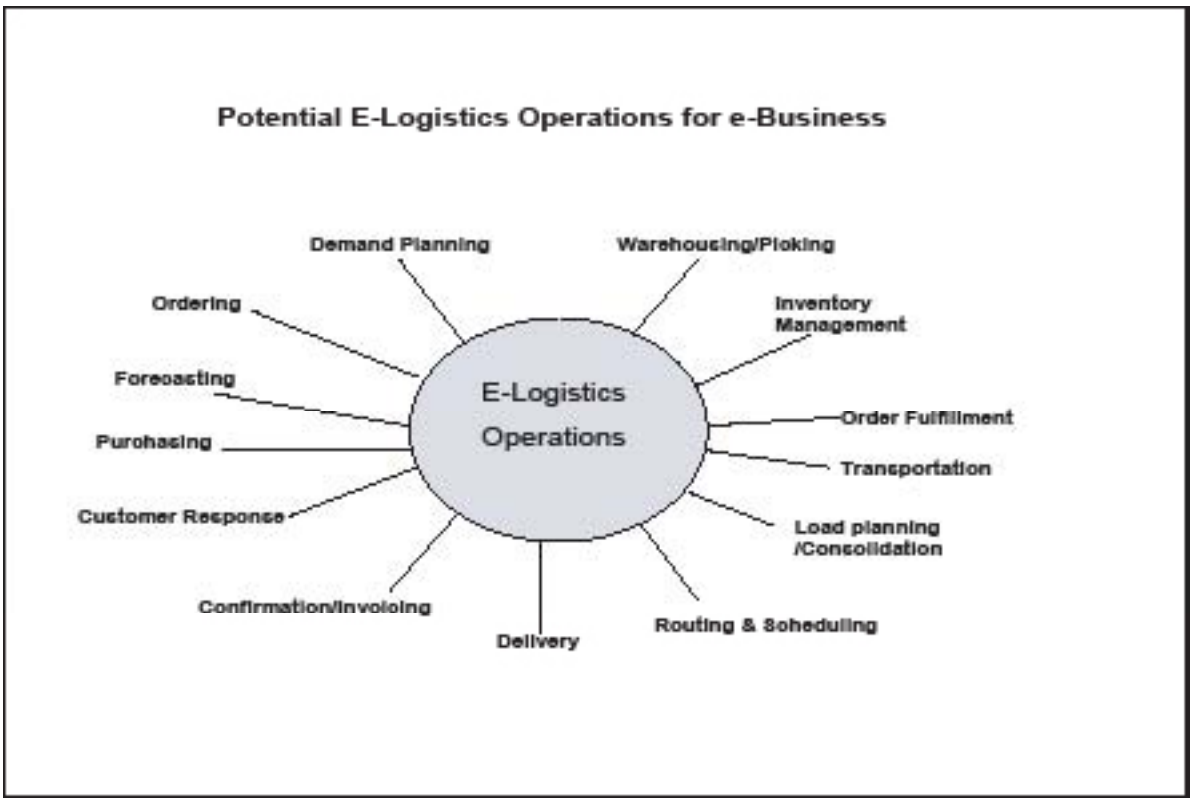
(Source: Hassall, 2002, revised.)

Figure 2. The proposed expanded B2B and B2C environments



(Source: Nemoto, Visser,& Yoshimoto, 2001)

Figure 3. Potential operational components of e-logistics



(Source: Hassall, 2001a)

replace these mega ERP systems that were being developed some five years ago. This is especially true as the requirement for many e-logistic applications such as Radio Frequency Identification (RFID), Track and Trace, Barcode reading, vehicle routing optimization and fleet management were often not part of these mega systems.

In fact, the build to scale E2E (end–to-end) solutions, or assembly of a suite of smaller "best of breed" solutions for MEs (Medium Enterprises), will be an attractive applications business space over the next five years for logistics operators to buy from.

Figure 3 depicts the several of the dozen or more end-to-end logistics and supply chain functions that can be streamlined through either individual or aggregated Web based e-logistic applications and platforms, through which e-business can be supported and fulfilled.

E-Logistics: Is it all about Visibility, Validation, and Verification? The New Question

What makes the difference between traditional logistics and e-logistics? This is a crucial question which may be better understood from examining the several definitions of e-logistics. Current definitions refer to logistics support for Web ordering. However, it may be suggested that true "e-logistics" operations are generated from Web based/electronic applications that are activated by the order, guided through the process stages until fulfilment, invoicing, payment, and reporting occurs. If one or more electronic logistic tools are invoked across the order to delivery, then there are the three properties of visibility, validation, and verification that can be seen as the electronic "value adds" for the e-logistics chain

as opposed to the "phone/fax" orders and delivery used by traditional logistics. In fact the property of Visibility is the most significant difference between e-logistics and traditional logistics, at least according to this author.

What Does the Crystal Ball Hold?

Around June 2001, somewhat coincidentally, when the OECD hosted a 44 country summit on "e-transport and e-logistics", a wave of reality was emerging such that the business models for the fulfilment of e-business had to be rethought. E-stores and e-markets were proving that they were not very "logistics savvy." The area of Business to Consumer (B2C) was very much a high focal point and is arguably still, in France, the Netherlands, Germany, Italy, and many other European countries. B2C was "doing it tough with very low, or negative, profit margins" even though over a dozen fulfilment strategies had evolved in the B2C space. The bigger Business to Business (B2B) brother, although it was seen as being several factors larger in volume and value than B2C operations, when viewed from an e-logistics perspective, it was somewhat harder to enunciate the generalised logistics activities.

However, B2B in one perspective looks after itself. If the customer requires specific levels of service, reporting, software collaboration, and so forth, then the logistics supplier has a choice, either develop, or link into the capability, or risk losing the contract. The customer needs also to be aware what, specifically, the logistical services from a third party, or own fulfilment divisions, involve, and at what cost they can be provided. Many B2B relationships may just be coming to grips with a few of the available e-logistic tools and how to both price and leverage off their use. However, many major B2B relationships are still totally supported phone and facsimile machines. It is more than possible that the future evolution of very cheap public domain software for planning, routing, reporting, and for performing even "track and trace" through various mechanisms, will encourage not only the Medium Enterprise (ME) logistics operators, but even quite large 3PLs, the opportunity to offer a much larger suite of e-logistic functionality to their clients at very cheap prices.

CONCLUSION

The absolute difference between e-logistics and traditional logistics is the electronic "visibility" of particular steps in the chain, from order to confirmation of delivery and onward to invoice payment. The electronic logistic audit trail also lends itself to "verification and validation" of orders, pick pack instructions, standards of delivery services, and any alternative delivery instructions. In fact, the full range of services taken up in the e-logistics chain can be a very powerful input activity subset into the Activity Based Costing models (ABC) for that particular chain. This also assists what many logistic companies, B2C or B2B, are not fully cogniscent of, and that is their full attributable activity-based chain costs. A comprehensive knowledge of impact of these operationally based costs can very much make the difference between profit and failure.

This difference between traditional and e-logistics has not been widely examined in any a more meaningful way than by suggesting that e-logistics "may" be highly Web and applications enabled, unlike traditional logistics. But traditional logistics offerings that might only depend on the phone and facsimile machine may also be supported by very good databases, service response teams, and accounting and business systems. These systems may be very functional and successful. The differences between the streams of e-logistics and traditional logistics, is a fertile ground for future research which should also canvass the opinions of the actual logistic and transport operators.

REFERENCES

Bayles, Deborah L. (2001). *E-Commerce Logistics & Fulfillment*, Prentice Hall PTR, New Jersey.

Hassall, K. (2000). The B2C revolution. *Supply Chain Review*, 9(4),41 – 44, Publishing Services Australia, Brisbane.

Hassall, K. (2001a). Trends and Hindrances in e-logistics: An Australian Perspective. *Proceeding of the OECD conference on e-Transport, Paris*, 2001.

Hassall, K. (2001b).The Evolution of B2C. *Supply Chain Review*, 10(4), 42 – 50, Publishing Services Australia, Brisbane.

Hassall, K. (2003). The E-Logistics Challenge for the Post Office: A Phoenix egg or an Ostrich Egg? *Universal Postal Union (United Nations), Postal Technology Branch*, Bern. http://www.e-thematic.org/download/The%20e-Logistics%20Challenge%20for%20the%20Post%20Office.pdf

Hassall, K. (2005). *Smart Supply Chain Conference*, Technical Seminars, Sydney. (presentation) http://www.smartsupplychain.com.au/seminars.cfm

Nemoto, T., Visser, J., & Yoshimoto, Y., (2001). Impacts of Information and Communication Technology on Urban Logistics Systems. *Proceedings of the OECD Summit on e-Transport and e-Logistics, OECD*. Paris 2001. Http://www1.oecd.org/cem/online/ecom01/Nemoto.pdf

Poirier, Charles C., & Bauer, Michael J., (2001). *E-Supply Chain*. Berrett-Koehler Publishers Inc., San Francisco.

Reynolds, J. (2001). *Logistics and Fulfillment for e-Business*. CMP Books, New York.

Reynolds, J. (2004). *The Complete E-Commerce Book, 2nd Edition*, CMP Books, New York.

Rowlands, P. (editor). (2000- current). *E-logistics: Magazine*. Spice Court Publications.

Sarkis, J., Meade, L.M., & Talluri, S. (2004). E-logistics and the Natural Environment. Supply Chain Management. *An International Journal*, 9 (4),303-312, Emerald Group Publishing Limited.

Visser, J., & Hassall, K. (2005). The Future of City Logistics: Estimating the Demand for Home Delivery in Urban Areas. *Proceedings of the 4th City Logistics Conference Langkawi:* Kyoto University.

www.fulfilmentonline.ac.uk

www.elogmag.com

KEY TERMS

3PL (IIIPL): Third party logistics provider. A diversified provider of logistics services that may include: warehousing, freight forwarding, longhaul and shorthaul transport , storage, inventory management, returns, tracking, performance monitoring, selected documentation, and so forth.

Electronic Data Interchange (EDI): Computer to computer interchange between two or more companies so such entities can enter a range of standard forms such as purchase orders, bills of lading, invoices, forward orders, stock replenishment, and so forth.

E-Logistics: The following definitions are variations on the theme that e-logistics utilises Web based tools in the support of e-business. Some definitions are fuller than others. The timing in the emergence of these definitions is also relevant.

1. **Finnish (translation)**
   E-logistics can be defined as the application of Internet based technologies to traditional logistics processes.
2. **German: (Translation)**
   E-Logistics: Web based applications and services dealing with the efficient transport, distribution, and storage of products along the supply and demand chain.
3. **French: (Translation)**
   E-Logistics: A collection of the new logistic management practices for the Internet.
4. **Forbes:** "E-logistics" describes three core back-end processes required to get an order from the "buy" button to the bottom line: warehousing, delivery, and transportation, and customer interaction (usually handled through a call center where customers can ask questions, place orders, check on order status, and arrange for returns). In many cases, a different vendor handles each of these three separate functions. Managing them all successfully and simultaneously requires an in-depth understanding of each discipline. Integrating them with one another and with a corporation's existing systems is even tougher. Source: www.forbes.com/specialsections/elogistics.
5. "The electronic activation of a set of physical logistic activities with associated electronic information flows that support e-Business." (Hassall, 2005) (presentation) http://www.smartsupplychain.com.au/seminars.cfm.

ICT: Information and Communication Technology.

IV PL (4PL™): Fourth Party Logistics Provider. The operation of end-to-end coordination and management of a logistics chain by direction to 3PL operators and often including the use of large scale IT systems that drive management, reporting, and scheduling activities. IV PL operators need not operate any 3PL activities themselves.

Radio Frequency Identification: Active or passive tag or print based "label" that can be pre-programmed with specific data input fields. These fields emit data from active tags or can reflect embedded data for inactive tags when pulsed with specific electronic signals.

Reverse Logistics: The process of collecting, moving, storing used, damaged, or outdated products and/or packaging from end users.

Tecso Model: Strategy for replenishment of grocery orders for delivery from existing supermarket stores rather than from a specialized, non retain pick/pack urban depots. Used by Tesco stores.

Emergence Index in Image Databases

Sagarmay Deb
Southern Cross University, Australia

INTRODUCTION

Images are generated everywhere from various sources. It could be satellite pictures, biomedical, scientific, entertainment, sports and many more, generated through video camera, ordinary camera, x-ray machine, and so on. These images are stored in image databases. Content-based image retrieval (CBIR) technique is being applied to access these vast volumes of images from databases efficiently. Some of the areas, where CBIR is applied, include weather forecasting, scientific database management, art galleries, law enforcement, and fashion design.

Initially image representation was based on various attributes of the image like height, length, angle and was accessed using those attributes extracted manually and managed within the framework of conventional database management systems. Queries are specified using these attributes. This entails a high-level of image abstraction (Chen, Li & Wang, 2004). Also there was feature-based object-recognition approach where the process was automated to extract images based on color, shape, texture, and spatial relations among various objects of the image.

Recently combining these two approaches, efficient image representation and query-processing algorithms, have been developed to access image databases. Recent CBIR research tries to combine both of these above mentioned approach and has given rise to efficient image representations and data models, query-processing algorithms, intelligent query interfaces and domain-independent system architecture.

As we mentioned, image retrieval can be based on low-level visual features such as color (Antani, Rodney Long & Thoma, 2004; Deb & Kulkarni, 2007; Deb & Kulkarni, 2007a; Ritter & Cooper, 2007; Srisuk & Kurutach, 2002; Sural, Qian & Pramanik, 2002; Traina, Traina, Jr., Bueno, & Chino, 2003; Verma & Kulkarni, 2004), texture (Antani et al., 2004; Deb & Kulkarni, 2007a; Zhou, Feng & Shi, 2001), shape (Ritter & Cooper, 2007; Safar, Shahabi & Sun, 2000; Shahabi & Safar, 1999; Tao & Grosky, 1999), high-level semantics (Forsyth et al., 1996), or both (Zhao & Grosky, 2001).

But most of the works done so far are based on the analysis of explicit meanings of images. But image has implicit meanings as well, which give more and different meanings than only explicit analysis provides. In this paper we provide the concepts of emergence index and analysis of the implicit meanings of the image which we believe should be taken into account in analysis of images of image or multimedia databases.

BACKGROUND

Concepts of Emergence

A feature of an image which is not explicit would be emergent feature if it can be made explicit. There are three types of emergence: computational emergence, thermodynamic emergence and emergence relative to a model (Cariani, 1992). We would use the latter one in our chapter.

Whenever we shift our focus on an existing shape, in other words an image, new shape emerges. The representation of the new shape is based upon our view of the original shape. The new shape emerges as we change our view of the original shape. This is the most important idea of emergence. Two classes of shape emergence have been identified: embedded shape emergence and illusory shape emergence (Gero, year unknown; Gero & Maher, 1994). These procedures could be based on geometrical, topological, or dimensional studies of the original shape.

Model of Emergence

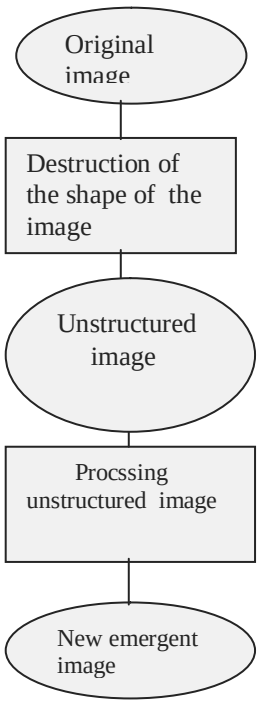
To extract emergent shape from an image, first we have to destroy the original shape of the image. This would give us an unstructured image. Now we take the unstructured image and find out the extra or implicit meaning out of it, in addition to the original meaning, and this process gives rise to emergent image with implicit meaning making explicit and emergent image would be generated. This can be defined in a model as follows (Gero & Yan, 1994):

Definition of Emergence Index

Image retrieval where the hidden or emergence meanings of the images are studied and based on those hidden meanings as well as explicit meanings, where there is no hidden meaning at all, an index of search is defined to retrieve images is called emergence index.

When images are retrieved based on textual information then various parameters and descriptions might define the input and the images of the database. Whenever there would

Figure 1 . Model of emergence



be symmetry of parameters and descriptions, the image could be retrieved. As mentioned earlier, in CBIR, color, texture, and shape are widely used as index to retrieve images. But in our studies, we can find the hidden meanings of the images and whenever those hidden meanings match with the input given, although the original image may not match at all with the input, we can retrieve that image.

When an input would come in the form of an image, the image could be studied based on features, constraints, variables and domains and converted into parametric form. Then the image database would be accessed and each image would be interpreted considering the items mentioned earlier, and also emergence and converted into parametric form like the input image. Whenever there would be a match between parameters of the input and the images of the database, these records would be selected. In other words, indexing would be decided by the outcome of emergence which means more meaningful images could be found hidden in an image which would otherwise not be understood.

Many images of the database may not have any apparent similarities with the input, but emergence could bring out the hidden meaning of the image and could establish similarities with the input image. So emergence outcomes of the images would form the index structure of the search.

Analyses of Works Done

Attempts have been made to give rise to symbolic representation of shape where shape is defined as

$$S = \{N; \text{Constraints}\}$$

where N is the cardinality, that is, the number of infinite maximal lines constituting shape S and the constraints limit the behaviors or properties resulting from the infinite maximal lines, based upon which particular shape is defined. Lines have been defined as Ik, Ij, and so on with their intersection as Ikj. Then topological, geometric, and dimensional properties are defined (Gero, 1992). Also symmetry has been found through the corresponding relevant positions of the lines and coordinates of one shape with that of the other and in the process, emergence of the shapes are studied (Jun, 1994).

There is no direct approach to solve the problem of emergent index other than the ones mentioned earlier. Only there is an indirect approach where this conception has been applied. In a model named Copycat involving computer programs, the program makes all possible sets of consistent combinations of pairings once all plausible pairings have been made. In other words, it gives rise to something explicit which were implicit earlier which is the essential feature of emergence phenomenon (Mitchell & Hofstadter, 1994).

MAIN FOCUS OF THE CHAPTER

We attempt to study the problem of image query where a query made would be searched through the database to select those records where a similar shape has been found. But in addition to that we pick up records based on the emergence phenomena where the query input may not have an apparent match in a particular image of the database, but emergence phenomena could give rise to a similar structure in the same image and as such this image should be selected as a query result. For example, a square with single diagonal can be observed as two triangles. So whenever search intends to find a triangle this image which apparently is much different than triangle would be selected because of emergence.

We calculate emergence index of images of image databases based on features, constraints, variables, domains, and emergence.

Various mathematical tools that could be used in the definition of the image:

- Geometric property
- Topological property
- Dimensional property
- Statistical properties

Structure of Emergence Index

Emergence indexes can be defined out of five factors

$$EI = f(D,F,V,C,E)$$

Where EI stands for emergence index, D for domain where the image belongs, F for features, V for variables which can define the feature's constraints under which the features are defined, C for constraints and E for emergence characteristics of images.

We believe any image, static or in motion, could be expressed semantically in terms of the above mentioned five parameters (Deb & Zhang, 2001).

Application of Emergence Index in Geographic Location

If we have a map of a geographic location like the one below, then we find there are three streets, namely, STREET1, STREET2 and STREET3. There is a park between STREET1 and STREET2 and HOUSE1, HOUSE2, HOUSE3, HOUSE4 are four houses on STREET2.

We also notice that STREET1, STREET2 and STREET3 form a triangle surrounding PARK. In normal map interpretation this may not surface. But when hidden shape is searched we get a triangle. This is the emergence outcome of the search. This would help us to locate the places more accurately by referring to the triangle in the map. Also if there is an input in the form of a triangle, then this image, although a map, would be selected because of emergence.

FUTURE TRENDS

Quite a few models of CBIR are now commercially available like QBIC, Virage, Excalibur, Attrasoft, and others. Noncommercial models developed by universities and research institutions are also available. But they do approximate matches between inputs and objects of image database. A query in most of the currently available CBIR systems is submitted in the form of an image and images similar to this particular image are selected and retrieved from the image database based on color, texture, shape, and spatial locations. Thorough and meaningful image segmentation, which is essential for accurate image retrieval, is still a problem. Also finding the semantic meanings out of an image from low-level features like color, shape, texture, and spatial locations and connect it to high-level features like chair, table, car, house, and so on is also another unresolved problem.

As far as CBIR based on emergence index is concerned, we have conducted preliminary experiments to show how it works (Deb & Kulkarni, 2007). We developed algorithms for retrieval of images from image database based on a given input image using emergence index. We plan to conduct more experiments based on automated image segmentation and retrieval from a query image by finding the semantic and hidden meanings of the target image in the database.

CONCLUSION

Emergence is a phenomenon where we study the implicit or hidden meaning of an image. We introduced this concept in image database access and retrieval of images using this as an index for retrieval. This would give an entirely different search outcome than ordinary search where emergence is not considered as consideration of hidden meanings could change the index of search. We discussed emergence, emergence index and approach as to how to apply this concept in image retrieval and preliminary experiments to retrieve images based on emergence index in this chapter.

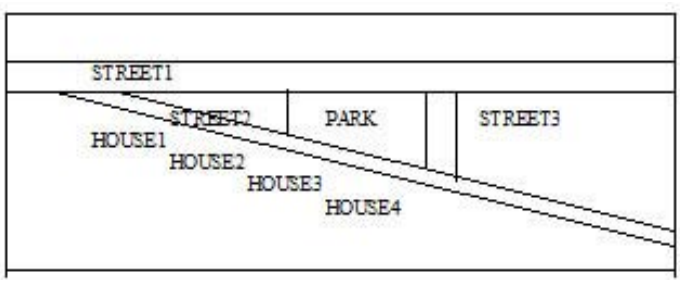
REFERENCES

Antani, S., Rodney Long, L., & Thoma, G. R. (2004). Content-based image retrieval. Large Biomedical Image Archives, *MEDINFO 2004*, 829-33 .

Cariani, P. (1992). Emergence and artificial life. In C. Langton, C. Taylor, J. D.

Farmer, & S. Rasmussen (Eds.), *Artificial life II* (pp. 775-797). Reading: Addision-Wesley.

Figure 2. Geographic location



Chen, Y., Li, J., & Wang, J. Z. (2004). *Machine learning and statistical modeling approaches to image retrieval.* New York: Kluwer Academic Publishers.

Deb, S. & Kulkarni, S. (2007). Human perception based image Retrieval using emergence index and fuzzy similarity measure. In *Proceedings of the Third International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP07)* (pp. 359-363). Melbourne, Australia.

Deb, S. & Kulkarni, S. (2007a). Content-based image retrieval with emergence index using fuzzy logic. In *Proceedings of the 5th International Conference on Advances in Mobile Computing and Multimedia (MoMM2007)*. Jakarta, Indonesia.

Deb, S. & Zhang, Y. (2001). Emergence index structure in image retrieval. *Tamkang Journal of Science and Engineering, 4*(1), 59-69.

Forsyth, D., et al. (1996). Finding pictures of objects in large collections of images. Report of the NSF/ARPA Workshop on 3D Object Representation for Computer Vision, 335

Gero, J. S. (1992). *Shape emergence and symbolic reasoning using maximal lines*. Unpublished notes, Design Computing Unit, Department of Architectural and Design Science, University of Sydney, Sydney

Gero, J. S. (Year unknown). *Visual emergence in design collaboration.* Key Center of Design Computing, University of Sydney.

Gero, J. S. & Maher, M. L. (1994, September). Computational support for emergence in design. In *Proceedings of the Information Technology in Design Conference*, Moscow.

Gero, J. S. & Yan, M. (1994). Shape emergence by symbolic reasoning. *Environment and Planning B: Planning and Design, 21*, 191-212

Jun, H. J. (1994). Emergence of shape semantics in CAD system. Unpublished doctoral thesis, Design Computing Unit, Department of Architectural and Design Science, University of Sydney, Sydney.

Mitchell, M. & Hofstadter, D. (1994). The copycat project: A model of mental fluidity and analogy-making. *Fluid concepts & analogies: Computer models of the fundamental mechanisms of thought*. New York: BasicBooks.

Ritter, N. & Cooper, J. (2007). Segmentation and border identification of cells in images of peripheral blood smear slides. In *Proceedings of Thirtieth Australasian Computer Science Conference (ACSC2007)* (pp. 161-169). Ballarat, Australia.

Safar, M., Shahabi, C., & Sun, X. (2000). Image retrieval by shape: A comparative study. In *Proceedings of IEEE International Conference on Multimedia and Exposition (ICME)*, USA

Shahabi, C. & Safar, M. (1999). Efficient retrieval and spatial querying of 2D objects. In *Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMCS99)* (pp. 611-617). Florence, Italy.

Srisuk, S. & Kurutach, W. (2002). An efficient algorithm for face detection in color images. In *Proceedings of 6th Joint Conference on Information Sciences* (pp. 688-691). Research Triangle Park, North Carolina.

Sural, S., Qian, G., & Pramanik, S. (2002). A histogram with perceptually smooth color transition for image retrieval. In *Proceedings of 6th Joint Conference on Information Sciences* (pp. 664-667). Research Triangle Park, North Carolina.

Tao, Y. & Grosky, W. (1999). Delaunay triangulation for image object indexing: A novel method for shape representation. In *Proceedings of the Seventh SPIE Symposium on Storage and Retrieval for Image and Video Databases* (pp. 631-942). San Jose, California.

Traina, A. J. M., Traina, C., Jr., Bueno, J. M., & Chino, F. J. T. (2003). Efficient content-based image retrieval through metric histograms. *World Wide Web Internet and Web Information Systems, 6*, 157-185.

Verma, B. & Kulkarni, S. (2004). Fuzzy logic based interpretation and fusion of colour queries. *Journal of Fuzzy Sets and Systems, 147*(1), 99-118.

Zhao, R. & Grosky, W. I. (2001). Bridging the semantic gap in image retrieval. *Distributed multimedia databases: Techniques and applications*. Hershey, PA: Idea Group Publishing.

Zhou, P., Feng, J. F., & Shi, Q. Y. (2001). Texture feature based on local fourier transform. In *Proceedings of the International Conference on Image Processing*, (Vol. 2, pp. 610-613). Thessaloniki, Greece.

KEY TERMS

Computational Emergence: Here it is assumed computational interactions can generate different features or behaviors. This is one of the approaches in the field of artificial life.

Content-Based Image Retrieval: In this kind of retrieval, symmetry between input image and images of database are established based on contents of the images under consideration.

Embedded Shape Emergence: In embedded shape emergence, all the emergent shapes can be identified by set theory kind of procedures on the original shape under consideration. For example, in a set S= {a,b,c,d,e}, we can find subsets like S1={a,b,c}, S2={c,d,e}, S3={a,c,e}, and so on.

Emergence Index: Image retrieval where the hidden or emergence meanings of the images are studied and based on those hidden meanings as well as explicit meanings, where there is no hidden meaning at all, an index of search is defined to retrieve images is called emergence index.

Emergence Relative to a Model: In this case, deviation of the behavior from the original model gives rise to emergence.

Illusory Shape Emergence: In illusory shape emergence, where contours defining a shape are perceived even though no contours are physically present. Here set theory procedures are not enough and more effective procedures have to be applied to find these hidden shapes.

Thermodynamic Emergence: This is of the view that new stable features or behaviors can arise from equilibrium through the use of thermodynamic theory.

Emerging Online E-Payment and Issues of Adoption

Qile He
University of Bedfordshire Business School, UK

Yanqing Duan
University of Bedfordshire Business School, UK

INTRODUCTION

Due to the rapid growth of e-commerce, the physical boundaries between parties in business transaction have been eliminated by the fast and convenient network connection. Nevertheless, most payment still has to be carried out offline by conventional methods. Over the last decade, a large number of online payment solutions have been developed, but many are still remained at the trial stage; while others are competing with each other; some even failed to reach a customer acceptance stage before developers quit the business. Reasons of slow acceptance are technological, but more importantly, societal. Developers are struggling in pushing increasingly secure and convenient technological solutions to the public. On the other side, users are seeking the balance between benefits and the risks of using online e-payment, which prolongs the process of wider acceptance. This article offers a brief introduction to typical online e-payment instruments and classifications of existing payment systems. It intends to provide researchers and developers with a clearer view on e-payment by comparing various existing systems. The article also attempts to shed light on the issue of social acceptance and adoption of online e-payment.

BACKGROUND

Online e-payment refers to the process of finance or payment mainly using Internet as a medium. Making payment electronically is not new. Long before online e-payment has been introduced, financial institutions had established Automated Clearing House (ACH) to clear money transfers electronically. The electronic financial infrastructure has promoted introduction of various Electronic Fund Transfer (EFT) solutions. In a broader sense, electronic payment includes those based on private networks, such as ATM, credit card payment, POS (point-of-sale) and other payment over proprietary networks. The rapid development of B2B, B2C and C2C e-commerce gave rise to the development of new payment solutions over the open network. Thanks to the rapid development of ICTs and e-commerce, a vast number of online e-payment solutions have been introduced. Conventional payment methods and concepts like credit card and cheque have been extended and modified to incorporate online transactions. New schemes of payment like electronic currency and smart cards have been introduced for online payment. Many e-payment systems share similar characteristics or developed upon similar protocols or payment infrastructures. Despite numerous attempts aimed at offering innovative alternatives, credit and debit cards payment based on the existing payment network and procedures remains the main payment instrument for online transactions. For instance, a report showed that in the UK some 90% of online purchases are made by credit card and debit card, although the amount only represents 3% of all card payments (Allen, 2003).

Many have realized that the limited acceptance of online e-payment is by and large a chicken and egg problem. Diffusion of online e-payment is limited by the unavailability of payment solutions accepted by wide range of transactions. Moreover, the lack of market-wide diffusion limited development of more integrated online payment solution (Allen, 2003). The phenomenon has attracted attention of researchers to investigate the factors hinder the wider acceptance of online e-payment (Abrazhevich, 2001b; He, Duan, Fu, & Li, 2006). To have a better understanding of the issue of acceptance, the characteristics of different online payment systems, and technological and social issues associated with their implementation need to be clarified.

EMERGING ONLINE E-PAYMENT TECHNOLOGIES

In response to the rapid development of e-commerce and the security requirements of the online e-payment, research groups, financial institutions and commercial firms have developed a number of online e-payment solutions since the last decade. Table 1 provides a list of some typical online e-payment methods with examples.

Table 1. Various online e-payment systems

Type of Payment	Example	Advantages	Disadvantages
Credit/Debit card	SET (Cyber Cash)	Card information not transmitted; Multiple-layer authentication.	Complexity
	SSL	Simplicity; Easy to setup; Using existing financial infrastructure.	Non-anonymity; Transaction cost relatively high.
Electronic Cheque	FSTC E-check	Able to handle large value payment.	Extra infrastructure needed; Need hardware devices to store electronic chequebook.
	NetBill	Simplicity; Able to handle small amount payment; Payer remains unknown to the payee.	Reliance on a central server; Server load limits number of participants.
	NetCheque	Choices of various NetCheque servers; Scalability.	Need to establish a hierarchy of servers.
Electronic Currency	Ecash	Payer anonymity remained.	Rely on single Ecash bank; Large database needed to store serial numbers.
	NetCash	Different currency servers allowed; Scalability.	Limited anonymity to users; Large database needed to prevent double spending.
	Millicent	Capable of handling micropayment; Communication efficiency; Low transaction cost.	Limited customer anonymity; Broker and vendor need to be trusted.
Smart Card Payment	Mondex	Portability; Anonymity; Secured fund transfer.	Need extra hardware devices.
	NACHA ISAP	Simplicity; Using existing financial infrastructure.	Need extra hardware devices; Non-anonymity.
Centralized Account System	Yahoo!Direct; PayPal; First Virtual; Nochex; iTransact	Double blind system; Payer and payee remain unknown to each other.	Lack of integrated system; Users need to register with various accounts.

Credit Card and Debit Card Payment

Extended from the conventional MOTO (Mail Order Telephone Order) transaction, credit and debit card as means of online e-payment has been widely explored. Thanks to the already developed payment schemes and established public acceptance, credit and debit card payment systems are most widely used online payment methods at present. Because cardholders need to provide card details on the Internet, which is subjected to attacks from hackers and fraudsters, security of the credit and debit card payment over the Internet is a common concern. To ensure the security, two major Internet standards have been introduced, namely SET and SSL.

The basic principal of SET (Secured Electronic Transaction) is to use digital certificate consists of a **private key** for encrypting data or documents and a corresponding **public key** for reading the data, so that the information is transmitted to the identified parties and is secured from external parties. SET creates secured dialogue between cardholder, merchant, and acquiring bank using digital certificates, which substitute credit card number during the transaction. However, the problem with SET is the complexity of the system, in that the transaction parties have to install independent pieces of software working together to accept the certificate and ensure the authentication of each party. SET also needs the support of a complex certification authority hierarchy, and requires the cardholders to go through a registration process in order to be issued with a certificate.

SSL (Secure Socket Layer) uses encryption technology to secure any dialogue taking place between buyer and

merchant across a "socket" interprocess communication mechanism. SSL allows encrypted traffic between a Web server and client using public key and private key technology. Due to its simplicity (only the merchant's Web server needs to be authenticated through a once-off and straightforward procedure, and the client remains unauthenticated), SSL is now widely used by online banks and online retail stores. However, there is no protection over the amount of the transaction, number of transactions, the legitimacy of the cardholder, or the creditworthiness of the card. It also can't prevent merchant from examining or tampering with payer's information.

Electronic Cheque

An **electronic cheque** is a document containing fields identical to those on a paper cheque with appropriate digital signatures being added when the cheque is first issued by the payer and also when it is endorsed by the payee. The **digital signature** is a piece of electronically recorded data representing fingerprint, which is then encrypted using a secret key. The digital signature can be signed only by the holder of the secrete key. Parties want to verify the digital signature can use the public-key to testify the validity of the fingerprint. The basic idea of electronic cheque is that payer who possesses an electronic chequebook signs the cheque with digital signature and passes it through secure e-mail or SSL enabled communication to the payee, who will then endorse and send the cheque to the payer bank to be settled for payment transfer.

The wide acceptance of the digital signature depends on the establishment of a network of cooperating certification authorities and bodies referred to as Public-key Infrastructure (PKI) (Li & Wang, 2003). Moreover, to handle e-cheque payments, banks need to provide some new infrastructures from existing systems. All these implementation rules inevitably increase the complexity of the system and its rapid adoption. However, compared with other online payment instruments, e-cheque is considered to be suitable to B2B and B2G **e-commerce**, as the electronic cheque has the potential to handle larger amount transactions.

Electronic Currency

Compared to credit card or cheque, conventional cash has the advantages of wider acceptability, guaranteed payment, no transaction charges, and anonymity (O'Mahony, Peirce, & Tewari, 2001). To resemble these advantages of cash, electronic currency was invented. **Electronic currency** was defined as "pre-paid products, in which a record of the funds or value available to a consumer is stored on an electronic device in the consumer's possession" (BIS, 2001, p. 1). To use e-currency, a payer has to prepay the fund in stored value card or in prepaid wallet software in order to issue a certificate. Once issued the electronic money represents the value, which may be spent with merchants who deposit the certificates in their own accounts or spend the currency elsewhere. Early trials of electronic currency include Ecash launched by DigiCash Inc. and NetCash developed by USC-ISI. The former employs Ecash wallet software to sore electronic coins issued by the same Ecash bank. The later allows payer and payee to use different currency servers to obtain or verify electronic coins. Another typical electronic currency called Millicent was developed by Compaq in the late 1990s to handle micropayment for information products over the Internet. Under the Millicent system, the customer purchases electronic currency called "broker scrip" from a broker which will be exchanged into "vendor scrip" when the customer decides to purchase a product from a vendor.

Smart Card Payment

Smart card is regarded as an ideal form of e-payment device, given its massive information storage capability and the embedded security function. Mondex is one of the typical smart card payment instruments owned by MasterCard and NatWest Bank of UK. Because the encryption software is stored in the smart card and authentication techniques are used, no central processing is required. Only two participants' cards are involved in the transaction. Consequently, no central records can be kept or interrogated, and the anonymity is maintained. Because the transaction cost by smart card is relatively low, this makes it suitable for small value payment online. Another typical smart card-based payment called Internet Secure ATM Payments (ISAP) is piloted by NACHA (a nonprofit electronic payment trade organization representing a large number of financial institutions) in late 2000 (O'Mahony et al., 2001). ISAP payment uses debit card embedded with smart card chip to store secret signing key. Instead of using a PIN number to authenticate the card payment, payers use digital signature, which is to be verified by the bank. The problem with both these systems is the need to have a card reader device installed to the user's computer, which increases the setting up cost of both payment methods.

Centralized Account Systems

The basic idea of **centralized account system** (also known as Third Party Handled System) is that a payer stores credit to the centralized account operated by an independent company. A payer who has been authenticated online notifies the centralized account system to make payment from its online account. The payee will receive verifications from the centralized account system and will notify the receipt of payment. Communications between payer, centralized

account system, payee and bank are normally secured by SSL protocols. In such a way, the centralized account system serves as a bridge between parties to allow payer spending credits from the online account to pay for goods and services to the merchants or individual payee. Typical examples of centralized account systems are PayPal and Yahoo!Direct.

For commercial vendors, the basic model of centralized account systems is less efficient because the payee has to verify the payment for each transaction. Therefore, to handle larger amount of transactions, Application Programming Interfaces (APIs) are introduced to integrate payment system with the merchant's Web site. APIs redirect customer from merchant Web site to the centralized account system server, which settles the payment and makes notifications to the merchant. Typical examples of such systems are Nochex and iTransact, which allow merchants to accept online credit card payment from customers, without bearing the cost and complexity of setting up online credit card payment system. Centralized account system is highlighted because the account information of both payer and payee remains unknown to each other during the transaction. However, because such instrument requires users to have an account with the same payment system, payers and payees need to have several accounts with different systems.

CLASSIFICATIONS OF ONLINE E-PAYMENT

Online vs. Off-Line

One of the classification criteria is whether the authorization server is involved during the payment process (Asokan, Janson, Steiner, & Waidner, 1997). The system is online when it involves an authorization server (usually as part of the issuer or acquirer) in each payment. On the other hand, off-line payments involve only payer and payee during the payment process. Compared to off-line payment systems, online payment systems are considered to be more secure, because an authorization party is involved in validating and monitoring the payment process. However, the increased communication also increases the complexity and commutation cost of the system. Typical examples of online payment systems include First Virtual, NetCheque, NetBill, Ecash, NetCash, and PayPal. Millicent and Mondex are two examples of off-line payment systems.

Specific Hardware vs. General-Purpose Hardware

Online e-payment systems could also be classified by specific hardware-based and general-purpose hardware-based (Ferreira & Dahab, 1997). Specific hardware refers to e-payment systems, in which users need the support of certain hardware during the payment process. FSTC E-check and smart card-based systems fall into this category, as payers need to have either a chequebook storing device or card reader device to make the payment. On the other hand, the general-purpose hardware payment system relies on the common computer to complete the payment, although some complementary software might be required to complete the payment. Examples are SSL, NetBill, NetCheque, Ecash, NetCash, Millicent, and PayPal.

Macropayments vs. Micropayments

Under the online payment scenario the value of payment could be very small, say one tenth of a cent to buy a single Web content. Thus, payment by macropayment systems, such as credit/debit card-based instruments, FSTC E-check, NetCash, and centralized account systems could be very cost ineffective. Therefore, micropayment systems are needed to carry out small value payments. Some payment instruments are designed particularly for micropayments, such as Millicent. There are also payment systems capable of handling small value payment, such as Mondex and NetBill.

Token-Based vs. Account-based

Despite the above-mentioned features, the role of financial institutions in the process of online payment is also an important issue. Currently, financial institutions are still the most important intermediaries for fund storage, clearing and transfer in most payment systems. Therefore, how bank accounts are incorporated into payment systems is an important characteristic to differentiate online e-payment systems. According to Abrazhevich (2001a), online e-payment could be classified as token-based or account-based. Account-based systems refer to payment systems in which money are represented by numbers in bank accounts and these numbers are transferred between parties in an electronic manner over computer networks. Accordingly, most credit and debit card-based systems, as well as electronic cheque payment and centralized account systems, fall into this category. Token-based systems are those that allow participants to exchange electronic tokens in the transaction, as the payment instrument itself carries the value. Most electronic currency systems, such as Ecash, NetCash, and Millicent, as well as some of the smart card based system, such as Mondex, fall into the category of token-based systems.

ONLINE E-PAYMENT ADOPTION AND FUTURE DEVELOPMENT

Among most of the emerging online e-payment solutions, security and reliability are of greater concern to developers. Many systems have sacrificed the simplicity for security purposes. Consequently, complex authentication procedures are introduced, more parties are involved, and more infrastructures are established. The intention is to reduce security bugs and to provide more complete and feasible online payment solutions. Nevertheless, the success of any online e-payment is not determined by perfection of the system but by wide market acceptance. Adoption of online e-payment is a more critical issue to be faced by vendors and developers.

Table 2. Requirements of merchants and customers over online e-payment

Requirements	Merchants	Customers
Security	√	√
Integrity	√	√
Low operational cost	√	
Compatibility to existing system	√	√
Reliability	√	√
Information richness	√	
Privacy	√	√
Flexibility		√
Convenience		√
Speed of settlement	√	√
User-friendly/ease of Use		√

Table 3. Possible risks of using online e-payment

Risks	Merchants	Customers
Security deficiencies	√	√
Lack of integrity	√	√
Overly dependent on system providers	√	
Sunk cost involved in implementation	√	
Lack of compatibility to existing system	√	
Lack of reliability	√	√
Losing privacy		√

Table 4. Barriers to online e-payment adoption

Internal Barriers		External Barriers	
Merchants	Customers	E-payment providers	Legal
Security concerns	Security concerns	Lack of cooperation	Lack of legal infrastructure
Reluctant to give up old system	Risk perceptions	Lack of interoperability between schemes	Too much restrictions
Lack of adequate HR	No access to services	Lack of standardization	
Inadequate technology infrastructure	Inadequate skills	Too many immature competing systems	
Financial resources constraints	Lack of bank account		

Due to the fast development of ICTs and increasingly integrated financial infrastructures, technology is no longer the main obstacle to online e-payment development. Nowadays, customers and merchants are facing the choice of many competing payment solutions. Developers need to be aware that the wide acceptance of online e-payment is more importantly a societal issue, which involves entities such as merchants, customers, system providers, financial institutions, and regulation bodies. Given users are ultimate forces determining the success of any e-payment systems, their behaviors and attitudes need to be examined more carefully before complicated systems are introduced (Hayashi & Klee, 2003; Lawson & Todd, 2003). In various e-commerce, merchants and customers are the most important entities and direct users. Their expectations and requirements are shaping the development and social acceptance of online e-payment (see table 2).

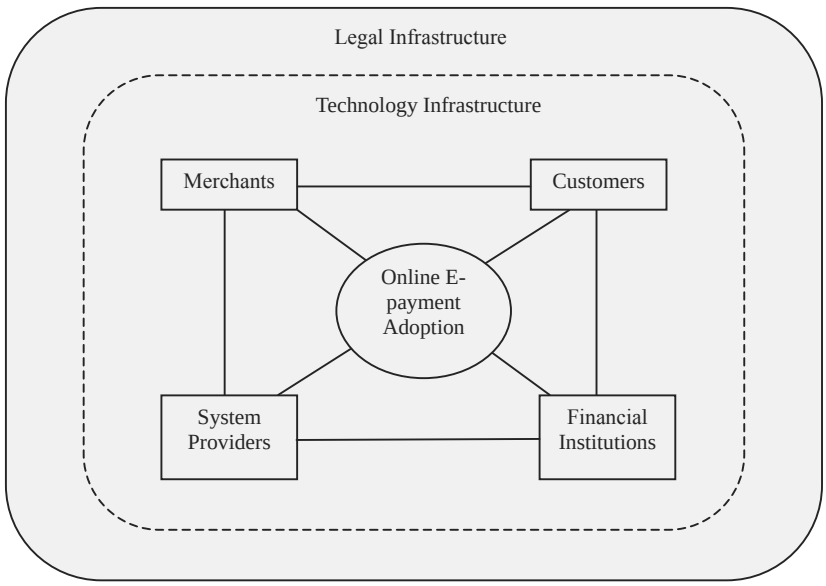
Expectations of both groups, however, are not the same, and sometimes are even conflicting to each other. For instance, merchants' expectation of acquiring more customer information is contradicting with customer's need of privacy (Heng, 2004). Such unbalanced expectations impose extra complexity to the development of online e-payment. At the early stage of e-payment adoption, the confidence of users is quite vulnerable (Thomas, 2003). Developers need to be aware that if the expectations of both business users and customers are not met, the benefits e-payment promised could easily turn into risks (see table 3). To meet the expectations of users, developers need to seek balance between various requirements of users. For example, developers need to be careful whether simplicity should be sacrificed to meet the high reliability and security standards.

Within barriers of adoption arise from internal and external of potential users (see table 4), collaboration of e-payment providers is a critical issue (ePSO, 2006). Plenty of evidence shows that too many competing payment systems damaged users' loyalty to e-payment and slowed down the adoption (Trombly, 2002). E-payment providers are suggested to cooperate more to improve the interoperability of different systems and the standardization of technologies (Allen, 2003). Moreover, a moderate legal framework is also needed to not only provide protection to e-payment users, but also give the market enough freedom for further development.

Obviously, the issue of successful development and adoption of online e-payment is too complex to be handled by a single entity. As shown in Figure 1, the wide acceptance of online e-payment needs the social wide collaboration from all parities involved, that is, system providers, financial institutions, merchants, customers, and regulation bodies. Moreover, successful development is bounded and also supported by technological and legal infrastructures. To overcome barriers to wider acceptance, e-payment developers need to have a more comprehensive view to the issue and seek more social wide collaboration.

Specifically, to avoid the vicious circle from occurring, in which merchants would not offer e-payment schemes if few customers use them, while customers would not use e-payments if few merchants accept them, developers are suggested to provide secure, simpler and more robust

Figure 1. Conceptual model of online e-payment adoption



payment solutions, which bridge various payment schemes to allow for the interoperatability of systems. It is worth noting that payment systems which use a maximum of the existing financial as well as technology infrastructures, are more likely to succeed.

CONCLUSION

This article introduced the emerging online e-payment. Different payment systems were classified according to various criteria. These classifications could be used as guidelines for both users and developers to have a better understanding of existing online payment solutions, and also the possible trend of future development. This article suggested that successful deployment and acceptance of online e-payment is not only a technological issue, but also a social issue. Behaviors and attitudes of both consumers and merchants need to be fully understood. Moreover, collaborations between various parities are needed to promote wider acceptance and further development of online e-payment.

REFERENCES

Abrazhevich, D. (2001a). Classification and characteristics of electronic payment systems. In K. Bauknecht, S. K. Madria, & G. Pernul (Eds.), *Proceedings of EC-Web 2001,* (pp. 81-90). Springer-Verlag.

Abrazhevich, D. (2001b). Electronic payment systems: Issues of user acceptance. In *Proceedings of eBusiness and eWork 2001*. Venice, Italy: IOS Press.

Allen, H. (2003). Innovations in retail payments: E-payments. *Bank of England Quarterly Bulletin, 43*(4), 428-438.

Asokan, N., Janson, P. A., Steiner, M., & Waidner, M. (1997). *Electronic payment systems*. Retrieved December 9, 2007, from http://citeseer.nj.nec.com/cache/papers/cs/1440/http:zSzzSzwww.semper.orgzSzinfozSz211ZR019.pdf/asokan-96electronic.pdf

BIS. (2001). *Survey of electronic money developments*. Retrieved December 9, 2007, from http://www.bis.org/publ/cpss48.pdf

ePSO. (2006). *Report on retail payment innovations, European Central Bank*. Retrieved December 9, 2007, from http://epso.intrasoft.lu/papers/Report-Retail-payment-innovations-2005.pdf

Ferreira, L., & Dahab, R. (1997). *A scheme for analyzing electronic payment systems*. Retrieved December 9, 2007, from http://citeseer.nj.nec.com/cache/papers/cs/15625/http:zSzzSzwww.dcc.unicamp.brzSz~lucasfzSzartigoszSzacsac98.pdf/decarvalhoferreira98scheme.pdf

Hayashi, F., & Klee, E. (2003). Technology adoption and consumer payments: Evidence from survey data. *Review of Network Economics, 2*(2).

He, Q., Duan, Y., Fu, Z., & Li, D. (2006). An innovation adoption study of online e-payment in Chinese companies. *Journal of Electronic Commerce in Organizations, 4*(1), 48-69.

Heng, S. (2004). E-payments: Modern complement to traditional payment systems. *Deutsche Bank Research, 44*(May 6).

Lawson, R., & Todd, S. (2003). Consumer preferences for payment methods: A segmentation analysis. *International Journal of Bank Marketing, 21*(2), 72-79.

Li, H., & Wang, Y. (2003). Public-key infrastructure. In W. Kou (Ed.), *Payment technologies for e-commerce* (pp. 39-70). Berlin: Springer-Verlag.

O'Mahony, D., Peirce, M., & Tewari, H. (2001). *Electronic payment systems for e-commerce* (2nd ed.). Boston: Artech House.

Thomas, D. (2003). Retailers suffer as attack hits e-payment provider. *Computer Weekly*, *4*.

Trombly, M. (2002). Cost savings and collaboration drive B2B E-payments. *Computerworld, 36*(42), 41.

KEY TERMS

Electronic Cheque: Cheque like electronic payment technologies using digital signature and certification technology to authorize and endorse the payment.

Electronic Currency: Prepaid product resembles the conventional cash, in which a record of the funds or value is stored on an electronic device in the consumer's possession.

Online E-Payment: Electronic payment technologies which allow money to be transferred over the open network, such as the Internet.

Public-Key Cryptography: An asymmetric encryption technology designed to secure communication between individual entities. The sender encrypts the message using public-key of the intended receiver, and the receiver then decrypts the message using its secret key.

Secure Socket Layer (SSL): An encryption technology based on public-key and private-key cryptography, which ensures secured dialogue between payer and merchant over the Internet.

Secured Electronic Transaction (SET): A set of technology protocols based on digital certificates, which identifies the parties involved in the payment transaction and protects the dialogue between parties.

Smart Card: A plastic card with an integrated circus chip securely embedded in the card. It can store 100 times more information than traditional magnetic cards in a form that cannot be copied.

Category: Software & Systems Design

E-Negotiation Support Systems Overview

Zhen Wang
National University of Singapore, Singapore

John Lim
National University of Singapore, Singapore

Elizabeth Koh
National University of Singapore, Singapore

INTRODUCTION

In this fast moving global working environment, negotiators are benefiting from the pervasive application of computers and networks in the workplace. There is an increasing usage of E-negotiation Support Systems (ENS) in both internal and external negotiations. ENS are computer systems that help negotiators achieve better agreements by enhancing their information processing capabilities and communication with other parties. Recent empirical research on ENS has shown that the employment of ENS facilitates the improvement of the negotiation process and outcome (e.g., Delaney, Foroughi, & Perkins, 1997; Goh, Teo, Wu, &Wei, 2000; Rangaswamy & Shell, 1997). This article identifies the key areas of ENS research, the corresponding constructs, findings and challenges. Finally, it proposes an integrative framework of ENS research for future research.

BACKGROUND

Negotiation

Despite being a common task for managers, negotiation is challenging, complex and effort demanding. Negotiations have been studied from many perspectives including sociology, psychology, political science, economics, applied mathematics, computer science and artificial intelligence. Negotiation is "a process in the public domain where two parties, with supporters of various kinds, attempt to reach a joint decision on issues in dispute" (Gulliver, 1979, p. 79). It is a special form of communication that centers on perceived incompatibilities and focuses on reaching mutually acceptable agreements (Putnam & Roloff, 1992). Through a negotiation, two or more parties can resolve conflicts and enter into contracts (Walters, Stuhlmacher, & Meyer, 1998).

Theoretical Perspectives of Negotiation

In the literature of negotiation research, the descriptive model and the prescriptive model form two major schools. While the descriptive model focuses on the process of negotiation, the prescriptive model emphasizes the outcomes of negotiation.

The descriptive model of negotiation is widely studied in social behavior science, sociology, and psychology. Based on sociological and psychological theories of learning and joint decision-making, the descriptive model seeks to describe what actually happens in a negotiation process (Weigand, Schoop, de Moor, & Dignum, 2003). Researchers within this stream focus on individual differences (Hausken, 1997), contextual characteristics of negotiation, situational determinants (Pruitt & Rubin, 1986), and cognitive processes of judgment, behavior, and outcomes in negotiation (Thompson, 1990).

The prescriptive model of negotiation, in contrast, stems from the studies of Game Theory, social psychology and organizational behavior. Its fundamental assumption is axiomatic rationality, where participants will always choose the options that are in their best interests according to the particular quality measurement instrument chosen. It is normative in the sense that it prescribes what negotiators should do to achieve the desired results (Weigand et al., 2003). While the theoretical objective of the prescriptive model is to predict the processes and outcomes of negotiation, the practical goal is to help people negotiate more effectively (Raiffa, 1982).

E-Negotiation Support Systems

Since the 1960s, when computer models were first employed for the support of individual negotiation, interest has been growing on the possibility of using computer technology and information systems to support negotiations (DeSanctis & Gallupe 1987). Today, a number of decision-aiding techniques

are employed by the decision support component, such as information control component for data storage and retrieval, representational aids, decision aiding techniques and models, and inference capabilities. The development of the Internet and multimedia benefited ENS as various communication channels could be deployed to enhance information exchange in the negotiation.

The types of ENS vary from no computer mediation assistance at all to fully automated computer arbitration. Based on the fundamental differences in the design and functionality of ENS, ENS can be classified into two categories (Rangaswamy & Starke, 2000): (1) preparation and evaluation systems, which provide negotiation decision support before or during a negotiation; (2) process support systems, which provide the negotiators with the means to communicate with each other, and computer mediation or arbitration mechanisms.

KEY AREAS OF ENS RESEARCH

To guide future research, there is a need to build a broader theoretical framework of ENS (Bui, 1994). In order to formulate the ENS research framework, this article reviews the following four major research areas:

1. The *e-negotiation support systems*; specifically, the system components.
2. The negotiation *process*; specifically, the interaction and cognition process.
3. The negotiation *outcomes*; specifically, the contract-related outcomes and process-related outcomes.
4. The *context* of ENS negotiation; specifically, the negotiator characteristics, the task nature and the culture.

E-Negotiation Support Systems

Conceptually, ENS consist of two subcomponents (Lim & Benbasat, 1993): the decision support systems (DSS) component and the electronic communication component. *DSS* help to refine the negotiators' objectives and enhance the capability of information processing and complex problem analyses, so that more efficient and balanced outcomes may be achieved. The use of the *electronic communication channel* increases the level of perceived commitment and trust in the other party. As a consequence, agreements may be reached with less time and effort spent.

In addition to the traditional ENS mentioned above, the autonomous *negotiation agent* is becoming popular (Beam & Segev, 1997). Instead of human negotiators, negotiation agents prepare and negotiate on behalf of their human "clients," especially in cases where the negotiation tasks are well-structured. Governed by computational rules, these agents may include a concession model with general strategies of concession in multiple-issue negotiations (Matwin, Szapiro, & Haigh, 1991), a case-based reasoning to plan and support negotiations (Sycara, 1990), and a genetic algorithm-based learning technique (Oliver, 1997). Negotiation agents could bring significant benefits, such as time savings, avoiding unnecessary cognitive limitations, lowering transaction costs, and increasing the efficiency of settlements (Rangaswamy & Starke, 2000).

Negotiation Process

Interaction Process-Communication

Negotiation, after all, is a special kind of communication (Putnam & Roloff, 1992). It is a dynamic process characterized by information exchange, persuasion, and joint problem solving. Communication in negotiation serves four primary functions (Tutzauer, 1989): (1) a vehicle for transmitting and accepting offers; (2) a means for conveying information; (3) a mechanism for shaping the relationship between the bargainers via argumentations; and (4) a lens for uncovering outcomes.

Research on communication in negotiation has followed two lines (Neale & Northcraft, 1991). One group studies the effects of communication on outcomes, specifically the content and style; while the other group investigates the determinants of communication tactical choices. In the context of facilitated negotiation using ENS, visual and audio channels are proposed to support the communication process. Media richness, synchronicity of the communication channels, and multilingual support are popular research areas.

Interaction Process-Negotiation Strategies

One of the important findings in negotiation strategy research is the Dual Concern Model (Pruitt & Rubin, 1986). This predicts that bargaining outcomes depend on the negotiator's concern for self profits (assertiveness) and concern for the opponent's welfares (cooperativeness). Table 1 presents the summary of the negotiation strategy and outcome predicted by this model.

The Dual Concern Model provides insights to the question of how to achieve a more effective negotiating tactic. This model suggests that successful integrative bargaining requires both high concerns for the value of one's own outcome as well as for the other's welfare. However, in real life, negotiators may fail to achieve that due to human limitations and resource constraints. These potential problems could be solved with the assistance of ENS.

Cognitive Process

Negotiator cognitions research focuses on what goes on in the mind of a negotiator. Negotiator cognitions can be classified into three classes (Neale & Northcraft, 1991):

Table 1. The dual concern model: Strategy and outcome

	High self concern		Low self concern	
	Strategy	**Joint Outcome**	**Strategy**	**Joint Outcome**
High other concern	Problem solving (Collaborative)	High	Yielding (Accommodating)	Low
Low other concern	Contending (Competitive)	Moderate	Inaction (Avoiding)	Low

negotiation planning, information processing, and affect. *Negotiation planning* refers to how an individual identifies systematic mechanisms for developing and implementing bargaining strategies. In the literature, proposed antecedents of planning include level of aspiration, integrative potential, role, and so forth. *Information Processing* refers to how an individual intakes and compiles the environmental inputs in order to make sense of the negotiation context. The identified information processing strategies include framing, anchoring and adjustment, overconfidence, and reactive devaluation. *Affect* refers to the ways an individual expresses his or her emotion and needs in a negotiation.

Negotiation Outcomes

Contract-Related Outcomes

Joint outcome or joint utility measures the efficiency of a negotiation. It is measured by the sum of total multi-attribute utility scores of negotiators from all parties for the final agreement. It hence provides a measure of the total utility of a negotiated settlement. It serves the same purpose as the efficient frontier.

Contract balance measures the fairness of the negotiation outcome. It is computed by the absolute value of the differences between the total utility scores achieved by each negotiation party. The agreement is balanced, if the score equals zero. It serves the same purpose as the Nash solution.

Process-Related Outcomes

Satisfaction is an important measure of negotiation outcomes (King & Hinson, 1994). It is essential that negotiators are satisfied enough with the negotiation process and outcomes to warrant further business. Good performance in a negotiation, improved relationship with opponents, and an enjoyable interacting process are ways to increase the negotiators' satisfaction.

Perceived control refers to the sense of ownership in the negotiation process. When negotiators are able to reach consensus, they prefer to do it by themselves rather than by relying on external assistance (Hiltrop & Rubin, 1982). Jones (1988) found that in high-conflict treatments, bargainers tended to ignore the computer suggestions in favor of their own solutions, even though they were often not as good as the options suggested by computer systems.

Negotiation Context

Negotiator Characteristics

Almost every negotiation involves people, thus it is essential to understand negotiator characteristics in the study of ENS. In the negotiation literature, gender (King & Hinson, 1994), personalities (Ford, 1983), intelligence, negotiation experience, communication style (Simintiras & Thomas, 1998) are the noted factors shaping the negotiation process and outcome (Lewicki, Saunders, & Minton, 2003).

Negotiation Task

The bargaining orientation of a negotiation task is an essential input to formulate strategies and tactics in the negotiation process (Kersten, 2001). A negotiation task can be distinguished through the dimensions of integrative and distributive tasks (Walton & McKersie, 1965). Distributive tasks are win-lose situations (Fisher & Ury, 1981) that lead negotiators to focus on "slicing the pie" (Thompson, 2000). In contrast, integrative tasks provide an opportunity for the negotiation parties to integrate their interests. Negotiators might achieve a win-win agreement by expanding the pie.

Nearly every negotiation is conducted under certain time limitations. *Time pressure* influences people's behaviors (Druskat & Kayes, 2000). The time allocated in the negotiation is defined as either relatively short (high pressure) or long (low pressure). Researchers have categorized response processes due to time pressure into three types (Stuhlmacher & Champagne, 2000): (1) the process toward agreement

would be accelerated due to the strong time pressures in the form of fixed deadlines (Druckman, 1994; Maule & Mackie, 1990); (2) negotiators may process information selectively with time pressure (Svenson & Edland, 1987); and (3) differential combination of information would be generated as well.

Structural Factors

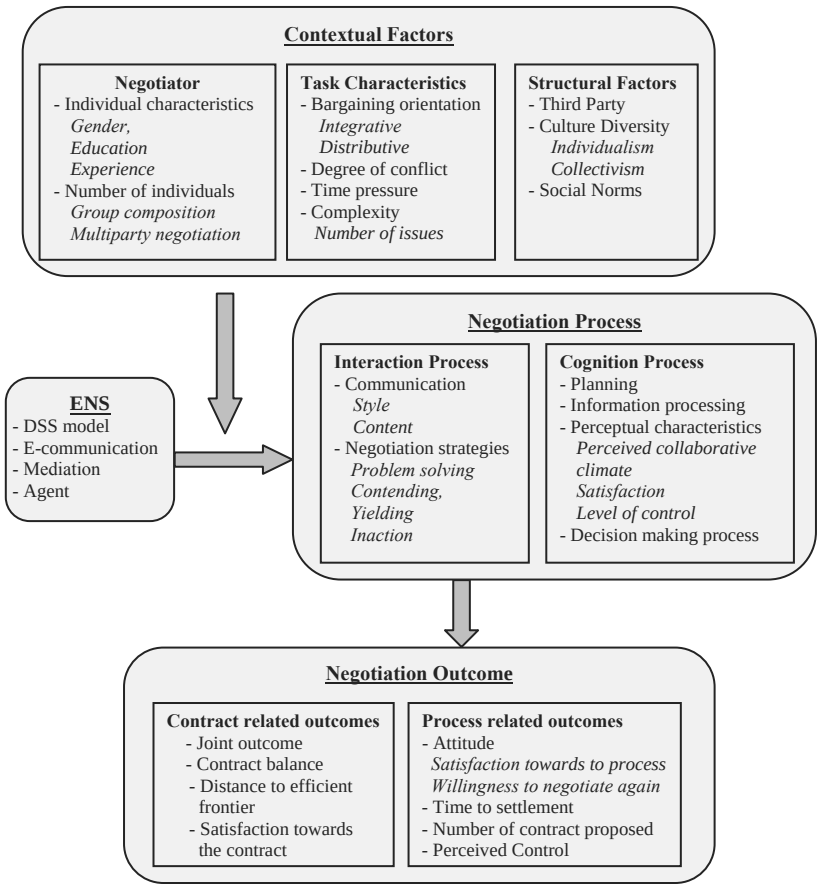
The *availability of third parties* may be another source that influences the negotiation process and outcome (Wall & Blum, 1991). In most of the cases, the third party would resolve the dispute in the negotiation, if the negotiating parties fail to do so. However, the third party may also have his or her own objectives (Zartman & Touval, 1985). Research has shown that negotiators alter their behavior with the presence of a third party (e.g., Neale, 1984). This effect would vary depending on the types and behavior of third parties. Third parties can be grouped into four basic types: mediator, arbitrator, conciliator and consultant (Fisher, 1990).

Culture can be defined as an entity consisting of interrelated parts, such as knowledge, beliefs, art, morals, customs and any other capabilities acquired by members of a society. A negotiator can only behave according to his or her cultural background and experiences. One conceptual dimension of culture is collectivism and individualism (Hofstede, 1991). This dimension refers to the extent to which a culture fosters its people to promote a group's well-being and to endorse the desire of self over a group. It reflects the relationship between an individual and his/her fellows. Thus, collectivists have norms and focuses about negotiation outcomes that are different from individualists (Wall & Blum, 1991). These norms and focuses are likely to influence the negotiation processes.

Research Framework for ENS

Figure 1 presents a proposed framework of ENS by integrating both prescriptive and descriptive perspectives of negotiation research. It summarizes the four categories reviewed above. Under each category, key variables are identified. This framework may provide researchers with a direction for future ENS research.

Figure 1. Proposed framework of e-negotiation support systems research



FUTURE TRENDS

E-Negotiation Technology

ENS can provide different levels of support to negotiators, such as DSS only, e-communication only, or full support. Researchers from the fields of computer science, engineering and economics might develop new algorithms and processing models of DSS to enhance the effectiveness and efficiency of ENS. Continuous effort on examining the variants of the e-communication component, such as video conferencing, is another direction (Yuan, Head, & Du, 2003). Although research on negotiation agents is still at the exploratory stage, its potential growth is both empirically and practically valuable. There is a need to develop an infrastructure and a negotiating protocol in which negotiators, DSS, and agents can work together for value creation (Lo & Kersten, 1999). Moreover, the role of ENS in a negotiation forms another interesting research direction. In addition, developing user-friendly interfaces by incorporating the findings of negotiator characteristics is another direction in the research domain of human computer interaction.

Interactive and Cognitive Behavior in Negotiation Process

Negotiation is a dynamic process characterized by information exchange, persuasion, and joint problem solving. There are two important components in the negotiation process: interaction process and cognition process. More studies are needed to reveal the dynamic between these two types of behavior. This helps to understand how a person's internal cognition leads to his or hers interaction behavior in negotiation.

In the ENS context, more effort is required to open the "black box" of negotiation process by exploring the negotiators' communication style and negotiation strategy. With greater knowledge in this area, researchers and designers might improve the ENS to better facilitate information exchange, increase the possibility to close a deal, and promote a collaborative negotiation atmosphere. The study of negotiation strategies could also contribute toward developing the concession model for negotiation agents and the design of negotiation training courses.

Contextual Factors in Negotiation Research

In the literature of information systems, social-psychology, computer science, and human computer interaction, three groups of contextual factors have been identified: negotiator characteristics, task characteristics and structural factors. These contextual factors are likely to moderate the effects of ENS on the negotiation process. There are a few studies that have visited these factors, specifically negotiator characteristics. More research is needed to understand the role of contextual factors in ENS facilitated negotiation. Potential findings can help system designers to better customize the ENS to satisfy different needs and fit varying situations. Nonetheless, group negotiation is prevalent in today's business world. The impacts of group formation, culture diversity, and third parties are areas that require more investigation.

Negotiation Outcomes

Contract-related outcomes are the measures reflecting the effectiveness and efficiency of the final agreement. Most empirical studies of ENS have examined these factors, such as joint outcome, contract balance, and the distance to efficient frontier. However, process-related outcomes have not been sufficiently studied; examples include perceived commitment of opponent, perceived collaborative atmosphere, and perceived control. These factors are essential to understand negotiator satisfaction and could consequently contribute to the knowledge of ENS adoption and diffusion. Case studies and longitudinal research are needed to understand these issues.

CONCLUSION

Having reviewed the influential theories and models in negotiation literature, we propose a framework for ENS research by synthesizing the research from both descriptive and prescriptive perspectives. These two theoretical perspectives are not mutually exclusive, but complementary. Together, they represent a holistic negotiation picture, which enables researchers to better understand the causes, processes and outcomes of negotiation. Four areas of interests are discussed: negotiation support systems, the negotiation process, contextual factors and the negotiation outcome. We also provide a list of important variables for each area. Some potential research directions and managerial implications are drawn from the proposed framework.

This article has some limitations as well. Firstly, the variables listed here are not exhaustive, but we aim to suggest a direction for research. Secondly, some of the variables proposed may be difficult to measure, for instance, communication content. Thirdly, in this framework, we have used constructs together with variables. This may lead to confusion. However, there is always a trade-off between the simplicity and the concreteness of a framework. When we try to provide a more complete view, we tend to include more variables. Thus, to simplify the framework, we used

some constructs instead. Lastly, we did not elucidate in detail all the relationships among the variables due to the page limit.

REFERENCES

Beam, C., & Segev, A. (1997). Automated negotiations: A survey of the state of the art. *Wirtschaftsinformatik, 39*(3), 263-268.

Brehmer, B. (1976). Social judgment theory and the analysis of interpersonal conflict. *Psychological Bulletin, 83*, 985-1003.

Bui, T.X. (1994). Evaluating negotiation support systems: A conceptualization. In *Proceedings of the Twenty-Seventh HICSS*, (pp. 316-324).

Delaney, M.M., Foroughi, A., & Perkins, W.C. (1997). An empirical study of the efficacy of a computerized negotiation support system. *Decision Support Systems, 20*(3), 185-197.

DeSanctis, G., & Gallupe, R.B. (1987). A foundation for the study of group decision support systems. *Management Science, 33*(5), 589-609.

Druckman, D. (1994). Determinants of compromising behavior in bargaining: A meta-analysis. *Journal of Conflict Resolution, 38*, 507-556.

Druskat, V.U., & Kayes, D.C. (2000). Learning vs. performances in short-term project teams. *Small Group Research, 31*(3), 328-353.

Fisher, R. J. (1990). Needs theory, social identity and an eclectic model of conflict. In J. Burton (Ed.), *Conflict: Human needs theory* (pp. 89-112). London: Macmillan.

Fisher, R., & Ury, W. (1983). *Getting to yes negotiating agreement without giving in*. New York: Penguin Books.

Ford, D.L. (1983). Effects of personal control belief: An explanatory analysis of bargaining outcomes in inter-group negotiations. *Group and Organization Studies, 8*, 113-125.

Foroughi, A., Perkins, W.C., & Jelassi, M.T. (1995). An empirical study of an interactive, session-oriented computerized negotiation support system. *Group Decision and Negotiation, 4*(6), 485-512.

Goh, K.Y., Teo, H.H., Wu, H.X., & Wei, K.K. (2000). Computer-supported negotiations: An experimental study of bargaining in electronic commerce. In *Proceedings of the 21st ICIS*, Australia, (pp. 104-116).

Gulliver, P. H. (1979). *Disputes and negotiations—a cross-cultural perspective*. Academic Press.

Hausken, K. (1997). Game-theoretic and behavioral negotiation theory. *Group Decision and Negotiation, 6*, 511-528.

Hiltrop, J.M., & Rubin, J.Z. (1982). Effects of intervention mode and conflict of interest on dispute resolution, *Journal of Personality and Social Psychology, 42*, 665-672.

Hofstede, G. (1991). *Cultures and organizations—software of the mind*. UK: McGraw-Hill.

Jones, B.H. (1988). *Analytical negotiation: An empirical examination of the effects of computer support for different levels of conflict in two-party bargaining*. Doctoral dissertation, Indiana University.

Kersten, G. E. (1997). Support for group decisions and negotiations: An overview multi-criteria analysis. *J. Climaco*, 332-246. Heilderberg: Springer-Verlag.

Kersten, G.E. (2001). Modeling distributive and integrative negotiations: Review and revised characterization. *Group Decision and Negotiation, 10*, 493-514.

King W.C., & Hinson, T.D. (1994). The influence of sex and equity sensitivity on relationship preferences, assessment of opponent, and outcomes in a negotiation experiment. *Journal of Management, 20*(3), 605-624.

Lewicki, R.J., Saunders, D.M., & Minton, J.W. (2003). *Negotiation* (4th ed.). Boston: McGraw-Hill.

Lim, L.H., & Benbasat, I. (1993). A theoretical perspective of negotiation systems. *Journal of Management Information Systems, 9*(3), 27-44.

Lo, G., & Kersten, G.E. (1999). Negotiation in electronic commerce: Integrating negotiation support and software agent technologies. In *Proceedings of the 5th Annual Atlantic Canadian Operational Research Society Conference.*

Matwin, S., Szapiro, T., & Haigh, K. (1991). Genetic algorithms approach to a negotiation support system. *IEEE Transactions on Systems, Man, and Cybernetics, 21*(1), 102-114.

Neale, M.A. (1984). The effect of negotiation and arbitration cost salience on bargainer behavior: The role of arbitrator and constituency in negotiator judgment. *Organizational Behavior and Human Performance, 34*, 97-111.

Neale, M.A., & Northcraft, G.B. (1991). Behavioral negotiation theory: A framework for conceptualizing dyadic bargaining. *Research in Organizational Behavior, 13*, 147-190.

Oliver, J.R. (1997). A machine learning approach to automated negotiation and prospects for electronic commerce. *Journal of Management Information Systems, 13*(3), 83-112.

Pruitt, D.G. (1981). *Negotiation behavior*. New York: Academic Press.

Pruitt, D.G., & Rubin, J.Z. (1986). *Social conflict: Escalation, stalemate, and settlement*. New York: Random House.

Putnam, L.L., & Roloff, M. E. (1992). *Communication and negotiation*. Newbury Park, CA: Sage.

Raiffa, H. (1982). *The art and science of negotiation*. Belknap Press of Harvard University Press.

Rangaswamy, A., & Shell, G.R. (1997). Using computers to realize joint gains in negotiations: Toward an electronic bargaining table. *Management Science, 43*(8), 1147-1163.

Rangaswamy, A., & Starke, K. (2000). Computer-mediated negotiations: Review and research opportunities. *Encyclopedia of microcomputers* (Vol. 26). New York: Marcel.

Simintiras, C.A., & Thomas, H.A. (1998). Cross-cultural sales negotiations: A literature review and research propositions. *International Marketing Review, 15*(1), 10-28.

Stuhlmacher, A.F., & Champagne, M.V. (2000). The impact of time pressure and information on negotiation process and decisions. *Group Decision and Negotiation, 9*, 471-491.

Svenson, O., & Edland, A. (1987). Change of preferences under time pressure: Choices and judgments. *Scandinavian Journal of Psychology, 28*, 322-330.

Sycara, K.P. (1990). Negotiation planning: An AI approach. *European Journal of Operational Research, 46*, 216-234.

Thompson, L. (1990). Negotiation behavior and outcomes: Empirical evidence and theoretical issues. *Psychological Bulletin, 18*(3), 515-532.

Thompson, L. (2000). *The mind and heart of the negotiator* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Tutzauer, F. (1992). The communication of offers in dyadic bargaining. In L.L. Putnam & M.E. Roloff (Eds.), *Communication and negotiation* (pp. 67-82). Newbury Park, CA: Sage.

Wall, J. A., & Blum, M.W. (1991). Negotiations. *Journal of Management, 17*(2), 273-303.

Walters, A.E., Stuhlmacher, A.F., & Meyer, L.L. (1998). Gender and negotiator competitiveness: A meta-analysis. *Organizational Behavior and Human Decisions Process, 76* (1), 1-29.

Walton, R.E., & McKersie, R.B. (1965). *A behavioral theory of labor negotiation: An analysis of a social interaction system* (2nd ed.). Ithaca, NY: ILR Press.

Weigand, H., Schoop, M., de Moor, A., & Dignum, F. (2003). B2B negotiation support: The need for a communication perspective. *Group Decision and Negotiation, 12*(1), 3-29.

Yuan, Y.F., Head, M., & Du, M. (2003). The effects of multimedia communication on Web-based negotiation. *Group Decision and Negotiation, 12*, 89-109.

Zartman, I.W., & Touval, S. (1985). International mediation: Conflict resolution and power politics. *Journal of Social Issues, 41*(2), 27-45.

KEY TERMS

Cognitive Information Processing: Refers to how an individual absorbs and compiles the environmental inputs in order to make sense of the negotiation context.

Contract Balance: Is the absolute value of differences between the total utility scores achieved by each negotiation party.

Descriptive Model of Negotiation Research: A research stream focusing on the process of negotiation.

E-Negotiation Support Systems (ENS): Are computer systems that enable negotiators to achieve better agreement by enhancing the negotiators' information processing capability and communication with the other party.

Joint Outcome: Refers to the efficiency of the negotiation. It is measured by the sum of the total multi-attribute utility scores from all the negotiators in the final agreement.

Negotiation: Is a process where two or more parties attempt to reach a joint decision on issues in dispute.

Negotiation Planning: Refers to a process where an individual identifies systematic mechanisms for developing and implementing bargaining strategies.

Prescriptive Model of Negotiation Research: A research stream focusing on the outcomes of negotiation.

Energy Management in Wireless Networked Embedded Systems

G. Manimaran
Iowa State University, USA

INTRODUCTION

Real-time systems have undergone an evolution in the last several years in terms of their number and variety of applications, as well as in complexity. A natural result of these advances, coupled with those in sensor techniques and networking, have led to the rise of a new class of applications that fall into the distributed real-time embedded systems category (Loyall, Schantz, Corman, Paunicka, & Fernandez, 2005; Report, 2006). Recent technological advancements in device scaling have been instrumental in enabling the mass production of such devices at reduced costs. As a result, applications with a number of internetworked embedded systems have become prominent. At the same time, there has been a need to move from stand-alone real-time unit into a network of units that collaborate to achieve a real-time functionality. Extensive research has been carried out to achieve real-time guarantees over a set of nodes distributed over wired networks (Siva Ram Murthy & Manimaran, 2001). However, there exist a number of real-time applications in domains, such as industrial processing, military, robotics and tracking, that require the nodes to communicate over the wireless medium where the application dynamics prevent the existence of a wired communication infrastructure. These applications present challenges beyond those of traditional embedded or networked systems, since they involve many heterogeneous nodes and links, shared and constrained resources, and are deployed in dynamic environments where resource contention is dynamic and communication channel is noisy (Report, 2006, Loyall et al., 2005). Hence, resource management in embedded real-time networks requires efficient algorithms and strategies that achieve competing requirements, such as time sensitive energy-efficient reliable message delivery. In what follows, we discuss some applications in this category, and discuss their requirements and the research challenges.

Safety-critical mobile applications running on resource-constrained embedded systems will play an increasingly important role in domains such as automotive systems, space, robotics, and avionics. The core controlling module in such mission critical applications is an embedded system consisting of a number of autonomous components. These components form a wireless (ad hoc) network for cooperatively communicating with each other to achieve the desired functionality. In these applications, a failure or violation of deadlines can be disastrous, leading to loss of life, money, or equipment. Hence, there arises a need to coordinate and operate within stringent timing constraints, overcoming the limitations of the wireless network. For example, robots used in urban search and rescue missions cooperate together and with humans in overlapping workspaces. For this working environment to remain safe and secure, not only must internal computations of robots meet their deadlines, but timely coordination of robots behavior is also required (Report, 2006). Other such medium-scale distributed real-time embedded applications include target tracking systems that perform surveillance, detection, and tracking of time critical targets (Loyall et al., 2005), or a mobile robotics application where a team of autonomous robots cooperate in achieving a common goal such as using sensor feeds to locate trapped humans in a building on fire. Other more passive applications include the use of networked embedded systems to monitor critical infrastructure such as electric grids (Leon, Vittal, & Manimaran, 2007). These applications need to meet certain real-time constraints in response to transient events, such as fast-moving targets, where the time to detect and respond to events is shortened significantly. In surveillance systems, for example, communication delays within sensing and actuating loops directly affect the quality of tracking. While providing real-time guarantees is the primary requirement in these applications, mechanisms need to exist to meet other crucial system needs such as energy consumption and accuracy (Rusu, Melhem & Mosse, 2003). In most cases, there are tradeoffs involved in balancing these competing requirements.

BACKGROUND

The typical architecture in a distributed real-time embedded system consists of several processor-controlled nodes interconnected through one or more interconnection networks. The system software running on each node enables the execution of one or more concurrent tasks that are activated by the arrival of triggering events generated by the external environment, a timer, or arrival of a message from another task. A response to an event generally involves several tasks to be executed on different nodes, and several messages to be

exchanged in the network. The tasks on the same node may share data and resources using synchronization mechanisms present in shared memory systems, and also interact with tasks on other nodes by exchanging messages using the services provided by the communication subsystem. For the proper functioning of the whole system, each individual task, as well as all the messages exchanged, need to be completed before specified deadlines.

The workload in the majority of the distributed embedded real-time applications is similar to those found in traditional real-time systems comprising of periodic and aperiodic tasks. Periodic tasks form the base load invoked at regular intervals, while aperiodic tasks include the transient load generated in response to alarm or an external environment stimuli. However, one can expect stronger cooperation between the internetworked units in more dynamic and complex systems inducing richer communication patterns than simple periodic messages. For distributed real-time embedded system, the primary requirement is that there is an end-to-end timing requirement that needs to be met. This implies that there exists a set of messages with complex precedence constraints that need to be exchanged between the networked nodes before some deadline. Hence, one needs to characterize the different message communications and computations that are possible, and perform a preruntime analysis to guarantee, a priori, that all the task deadlines will be met. Moreover, in a distributed real-time system, the ability to meet task deadlines largely depends on the underlying task allocation, and hence, we need a preruntime task allocation algorithm that takes into consideration the real-time constraints. Intertask communication significantly influences the response time of these distributed applications and hence, the design needs to account for the effect of delays imposed by the communication network and precedence constraints imposed by the communicating tasks during task allocation. Since the inherent nature of many of the discussed applications precludes the use of wired networks, wireless networks are commonly used in such applications.

The wireless medium is inherently unreliable due to characteristics such as fading and interference. Hence, to guarantee that tasks should meet timing constraints, it becomes necessary to develop techniques that characterize the unreliability in the network channel, and take them into account while making transmission scheduling decisions. Energy management is another crucial aspect for internetworked embedded devices. These devices contain not only radio and computer components, but also complete system functionalities, such as networking functions across all levels of the protocol stack. Energy savings and allocation among these modules will affect the life time of these battery-powered devices. Energy management also needs to be considered, together with other constraints in size, real-time requirements, functionalities, and network connectivity.

In summary, the combination of temporal requirements, limited resources and power, networked system architectures, time-varying wireless channel, and high reliability requirements presents unique challenges

(Loyall et al., 2005; Report 2006).The end goal of most of the research in this area is to devise efficient resource management algorithms for energy-constrained and highly dynamic wireless networks in order to support end-to-end system requirements that are comparable to their wireline counterparts.

MAIN FOCUS OF THE CHAPTER

Energy management is one of the key issues in the design and operation of networked embedded systems, which involves energy management at the system level considering both computing and communication subsystems. For embedded computing, there are well known techniques, such as dynamic voltage scaling (DVS) (Aydin, Melhem, Mosse, & Alvarez, 2004; Shin, Kim, & Lee, 2005) and dynamic power adaptation (DPM), that have been exploited by intertask and intratask scheduling algorithms. For wireless communication, techniques such as dynamic modulation scaling (DMS) (Raghunathan, Schurgers, Park, & Srivastava, 2002), dynamic code scaling (DCS), power adaptation (Raghunathan, Pereira, Srivastava, & Gupta, 2005), and adaptive duty cycling have been employed for minimizing energy consumption. These techniques essentially provide energy-time tradeoff, that is, the lesser the time taken for execution of tasks or transmission of messages, the higher the energy consumed.

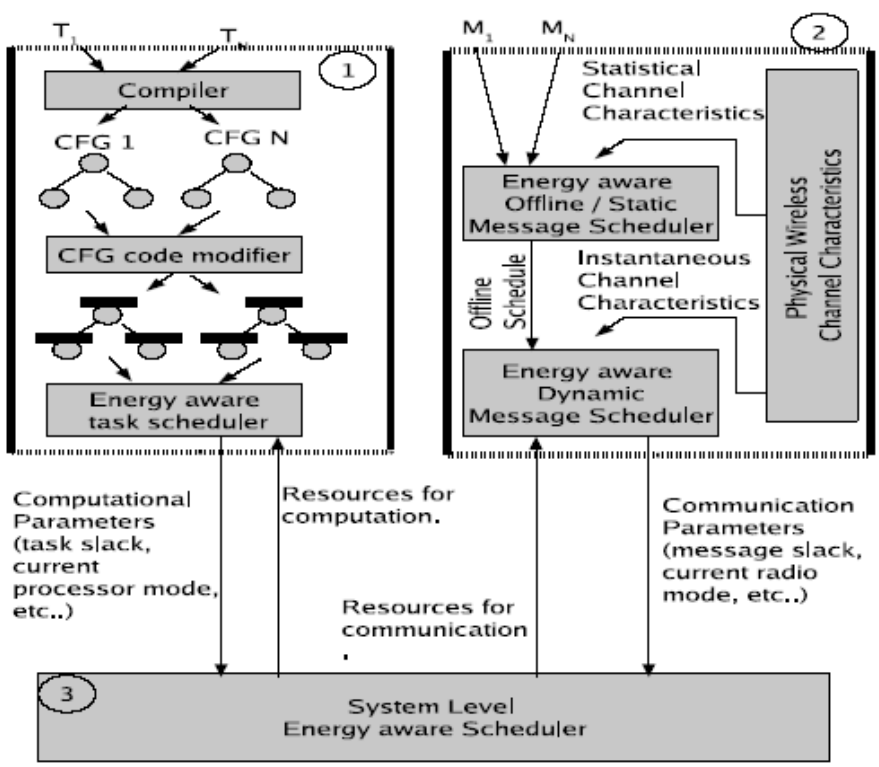
Our research contributions are in the design of a comprehensive energy management framework, with associated off-line and online scheduling algorithms, for networked embedded systems. In system-level energy management (Kumar, Sudha & Manimaran, 2007; Unsal & Koren, 2003), the fundamental question to be answered is how much of the available slack be allocated to each of task execution and message transmission. In general, the computation energy (for a CPU cycle) consumed is much lesser than the communication energy (for a bit of data transmission) for currently available technologies. Therefore, allocating as much slack as possible for communication energy optimization sounds appealing on the surface. However, our analysis shows that there is a diminishing return when the transmission time is increased beyond a certain threshold, with coding taken into account (Kumar et al., 2007). Therefore, the slack should be allocated in a balanced manner between computing and communication subsystems, considering current energy levels of tasks and messages and the channel condition. In our research, we considered DVS and DMS for energy

optimizations in computing and communication subsystems, respectively.

The major challenge in performing energy management in networked embedded systems lies in estimating the exact workload required by the application. The exact workload determines the least power mode that the device can operate at, while meeting the deadlines. In case of local computation, the workload refers to the task execution times, which exhibit a wide variation from their worst-case estimates. Most of the existing research speculates the task execution times. On the other hand, in the case of messages, the workload refers to the number of retransmissions required over a wireless link for a successful transmission. In our research, we consider both real-time constraints and channel conditions (reliability) while achieving energy efficiency of the networked embedded system. The proposed energy-aware resource management approach, shown in Figure 1, has the following three key components: computing subsystem energy management, communication subsystem energy management, and system-level energy management.

Energy Management at Computing Subsystem: This deals with the energy-aware real-time scheduling of tasks on a local node. Specifically, the goal is to minimize the processor energy consumption while meeting all the task deadlines. Specifically, we have designed cross-layer task scheduling algorithms that exploit intratask information, such as path locality information (if available) and run-time branching information, at the intertask scheduling level. We have designed a generic scheme, called *early basic block execution,* that aims at minimizing the energy consumption by reducing the nondeterminism in the workload; this is achieved by exploiting the control flow graph (CFG) of each task at the intertask level. The basic idea is as follows (Kumar, Sudha & Manimaran, 2006): "whenever the current task generates a slack due to a shorter branch execution, the early execution algorithm uses this slack to execute the basic blocks of the other ready tasks rather than using the entire available slack for slowing down the processor for the current task, with the objective of knowing other tasks' branching decisions which would otherwise be known at a later point of time." By performing such early execution of the basic blocks, the proposed algorithm builds a better picture of the workload at an earlier point in time. This will be exploited to scale the voltage/frequency appropriately across tasks, as opposed to within a task. This approach can acquire a much better idea of the future workload (branching decisions) than

Figure 1. Schematic of system-level energy management with computing and communication subsystems



a crude speculation, and hence, has the potential to offer significantly higher energy savings. These algorithms can be employed in networked as well as stand-alone embedded systems. The performance of such an algorithm depends on the nature of the workload and the overhead incurred by the algorithm itself.

Energy Management at Communication Subsystem: This deals with the energy-aware real-time scheduling of the internode messages over the wireless medium that is prone to phenomenon, such as fading, noise, and interference. Specifically, given a set of messages, each with a source and a destination, the goal is to transmit them, with the objective of minimizing the energy consumption of the communication subsystem while meeting all the message deadlines with a given probability of success. Due to the fading and noisy nature of the wireless channel, it is not feasible to guarantee 100% reliability.

We propose to estimate the channel condition using past feedback from the receivers, and based on the channel condition estimation/prediction, design efficient message transmission strategies, namely determining appropriate power level, modulation format, and coding scheme, for a given set of messages, such that they are successfully transmitted by the deadlines with maximum energy efficiency. We propose to include error-control coding in the energy consumption consideration, in addition to the modulation adaptation, as in DMS and power adaptation. By first quantifying the reliability (i.e., message success probability) using error-exponent and outage probability, we study the problem of allocating available slacks for communication among the messages in a way that maximizes the energy reduction.

Energy Management at the Networked System Level: In a typical networked embedded application, each node performs some local computation (task), and communicates the results (message) to a remote node in the network. Both task and message deadlines must be guaranteed in order to provide end-to-end deadline guarantees. In order to minimize the total energy consumption while guaranteeing the deadlines, the algorithm needs to optimally distribute the available slack among different tasks and messages. Task utilizes the slack to perform DVS, while the message uses the slack to perform DMS, or any other similar technique that trades off time for energy. In general, the computation energy is much less than the communication energy. Therefore, allocating maximum slack to communication subsystem sounds appealing on the surface. However, as can be seen from a more refined analysis, with coding taken into account, there is diminishing returns when the transmission time is increased beyond a certain threshold. Therefore, there should be a balance between the computing subsystem and the communication subsystem in slack distribution. For specific instances of the problem, we have validated such tradeoffs through theoretical analysis, considering the transmission time, wireless channel condition, and different overheads encountered in practice. Based on the results of our analysis, we have designed efficient energy-aware slack distribution algorithms that consider related tasks and messages in an integrated manner (Kumar et al., 2007). The challenge is to design distributed algorithms for slack distribution.

The optimization metric for the system-level energy-management algorithm depends on the requirements of the underlying application. Sample metrics include minimizing the total energy of the entire system, minimizing the maximum energy of the nodes in the system; at a higher level, metrics include maximizing the life time of the networked embedded system satisfying the coverage, connectivity, real-time, and/or reliability properties. The resource management consists of static and dynamic scheduling of tasks and messages. In static scheduling, the workload is periodic and the schedule is constructed off-line for the given workload on the target system platform. There are two options here: (1) construct an energy optimized feasible schedule using energy unoptimized feasible schedule as the input; (2) construct an energy optimized feasible schedule using the workload and target networked platform as inputs. While the first problem assumes a feasible schedule is produced by an existing distributing real-time scheduling algorithm, the second problem does not rely on such a schedule and hence, is a harder problem than the first one. In dynamic scheduling, two options exist: (1) reclaim both static and dynamic slacks and use them for energy optimization through a dynamic slack distribution algorithm. The static slacks are the ones that were left in the schedule as "holes," and the dynamic slacks are the ones that are created due to situations such as when the actual computation time of a task is less than its scheduled worst-case computation time, or when the actual transmission time (including retransmissions) is less than the worst-case transmission time of a message. In dynamic scheduling, efficient means to keeping track of the slack is critical for achieving high energy performance, and moreover, dynamic slack reclamation and distribution should be done in a distributed manner (with less or no coordination among nodes) without leading to any deadline violation anomalies.

FUTURE TRENDS

This work opens up several avenues for further research in the emerging area of networked embedded systems, which include the following: (1) Real-time networked embedded system architectures and resource management algorithms: (2) Cross-layer algorithms for compute-communication energy management; (3) Cross-layer algorithms for sense-compute-communication energy management; (4) Implementation of such architectures and algorithms for real-world applications; (5) Specific research on energy-management algorithms include: (a) designing static system-level energy-

aware scheduling algorithms taking task set and the network architecture as inputs, as opposed to taking a given feasible schedule as input; (b) studying the tradeoff involving other energy optimization techniques (e.g., DPM vs. DMS, DVS vs. DMS+DCS), and designing static and dynamic slack allocation algorithms for these specific instances; (c) addressing all these research problems in the context of multihop networks with structured topologies (e.g., tree and mesh) and arbitrary topologies.

CONCLUSION

Real-time embedded systems play a prominent role in a variety of applications ranging from medical sensors in the human body to signaling sensors in war fields. The consumer domain of the embedded devices is large and ever increasing. A natural result of this trend coupled with those in sensor technologies and wireless communications have led to the rise of a new class of systems, called the networked embedded systems. Energy management is of the key issues in the design and operation of such systems, which involves energy management at the system level considering both computing and communication subsystems. This chapter advocated cross-layer algorithms for energy-aware resource management in networked embedded systems, considering an integrated workload of tasks and messages and their respective power management techniques. The fundamental question to be answered is how much of the available slack be allocated to each of task execution and message transmission. The answer lies in analyzing the characteristics of tasks and messages, considering their current energy levels, deadlines, and the channel condition. The research highlighted in this chapter opens up several research directions in wireless networked embedded systems.

REFERENCES

Aydin, H., Melhem, R., Mosse, D., & Alvarez, P. M. (2004). Power-aware scheduling for periodic real-time tasks. *IEEE Transaction on Computers, 53*(5), 584-600.

Kumar, G. Sudha, A., & Manimaran, G. (2006). *Cross-layer algorithms for energy management in real-time embedded systems*. **Technical Report, Iowa State University.**

Kumar, G. Sudha, A., & Manimaran, G. (2007). **Energy-aware scheduling of real-time tasks in wireless networked embedded systems. In** *Proceedings of IEEE Real-Time Systems Symposium* **(pp. 15-24).**

Leon, R. A., Vittal, V., & Manimaran, G. (2007). Application of sensor network for secure electric energy infrastructure. *IEEE Transactions on Power Delivery, 22*(2), 1021-1028.

Loyall, J. P., Schantz, R. E., Corman, D., Paunicka, J. L., & Fernandez, S. (2005). A distributed real-time embedded application for surveillance, detection, and tracking of time critical targets. In *Proceedings of IEEE Real Time and Embedded Technology and Applications Symposium (RTAS)* (pp. 88-97).

Raghunathan, V., Pereira, C.L., Srivastava, M. B., & Gupta, R. K. (2005). Energy-aware wireless systems with adaptive power-fidelity tradeoffs. *IEEE Transactions on VLSI Systems, 13*(2), 211-225.

Raghunathan, V., Schurgers, C., Park, S., & Srivastava, M. B. (2002). Energy-aware wireless microsensor Networks. *IEEE Signal Processing Magazine, 19*(2), 40-50.

Report. (2006). *Coordination of safety-critical mobile real-time embedded systems*. Workshop on Research Directions for Security and Networking in Critical Real-Time and Embedded Systems, San Jose, CA, USA, TCD-CS-2006-16. Retrieved from http://www.cs.tcd.ie/publications/tech-reports/reports.06/TCDCS-2006-16.pdf

Rusu, C., Melhem R., & Mosse D. (2003). Maximizing rewards for real-time applications with energy constraints. *IBM Journal of R&D, 46*(5/6).

Shin, D., Kim, J., & Lee, S. (2005). Intratask voltage scheduling on DVS-enabled hard real-time systems. *IEEE Transactions on CAD, 24*(9).

Siva Ram Murthy, C., & Manimaran, G (2001). *Resource management in real-time systems and networks*. MIT Press.

Unsal, O. S., & Koren, I. (2003). System-level power-aware design techniques in real-time systems. *Proceedings of the IEEE, 91*(7), 1055-1069.

KEY TERMS

Cross-Layer Algorithms: Two or more layers of the system (e.g., computing and communication layers) work synergistically to achieve the stated objective of the system.

Embedded System: Computing system that is a core part of a large system to achieve sense, process, and actuation capabilities.

Energy-Aware Resource Management: The goal of minimizing energy consumption in the system.

Energy-Time Tradeoffs: This refers to the tradeoff involving time to execute a task or to transmit a message vs.

the amount of energy consumed. Lesser the time taken for execution of tasks or transmission of messages, the higher the energy consumed.

Real-Time Workload: The workload consists of a set of tasks and messages that have precedence relations among themselves, and each task/message has specific deadline before which the execution/transmission must be completed.

System-Level Resource Management: The goal of achieving system-level resource management objectives as opposed to making subsystem level optimizations.

Wireless Embedded Network: A set of embedded nodes connected through a wireless network; the wireless channel is time-variant.

Enhancing Workplaces with Constructive Online Recreation

Jo Ann Oravec
University of Wisconsin-Whitewater, USA

INTRODUCTION

Organizations have become more permeable— integrating more influences from the outside world— as participants engage in such online diversions as trading stocks, engaging in multiplayer games, or viewing images of their children in daycare. Ready availability of these activities has brought the potential for abuse but also new opportunities. Constructive uses of online recreation and play can enhance many workplaces (especially high-tech and information-saturated ones) and perhaps ultimately make them more productive. This article proposes that these complex issues be resolved through participatory approaches, involving workgroups in discussions as to what constitutes "constructive recreation" as well as in development and dissemination of effective and fair policies. This discourse can also ultimately increase levels of trust among team members and between employees and management.

BACKGROUND

Issues concerning the boundaries between work and play have provided continuing struggles for managers and employees. Workplaces have become more "porous" and permeable— integrating more influences from the outside world— as individuals engage in such online diversions as trading stocks, playing games, or viewing images of their children in daycare. Everyday workplace life is becoming more diverse and chaotic. Although many organizational roles today demand high levels of creativity and mental flexibility, they can also fail to provide the means through which individuals can gain fresh perspectives. In the "information age," playful, exploratory, and spontaneous interaction can also facilitate the exchange of ideas for tackling workplace problems. Managers who expect employees not to use the Internet for some amount of off-task activity severely misjudge the nature of workplace life— which is solidly infused in online interaction. Depriving employees of opportunities for Internet recreation in some cases excludes the possibility of nearly any form of diversion from assigned responsibilities.

Workplace use of the Internet for activities that are not directly authorized by management is often considered as the "theft" of human and computer resources, while construed as a just reward by employees (Lim, 2002). Even though many managers consider the personal use of the Internet as an ethical lapse (Greengard, 2000), the "moral high ground" concerning these issues is not entirely clear. Much of the rhetoric and advertising copy associated with workplace computing incorporates recreational imageries and motifs, which can send misleading signals to employees. A number of individuals have already had significant experience combining work with online recreation; convincing them that hard work cannot be combined with online play is thus a tough sell. Telecommuters returning to organizational settings are often not entrusted with the autonomy to engage in online breaks at appropriate times— latitude they take for granted when doing the same tasks in their home offices. Many young people became comfortable with computing through video games and online interpersonal interaction and took online breaks during their demanding college studies (Colkin & George, 2002). Individuals must find ways to cope psychologically with increased pressures on the job (Weil & Rosen, 1997) and management should explore creative but feasible ways to assist them in these efforts.

Wireless Internet applications add more complexities, further increasing the porousness of organizations and making employees' access to recreation less dependent on systems controlled by their managers. Daniels (2000) reports how wireless technologies (such as PDAs with Internet access) are used within meetings to amuse and distract participants, often resulting in productivity losses. Since wireless technologies are still in the early stages of adoption in many organizational contexts, placing severe restrictions on their use (and penalties for misuse) could be counter-productive. Personal computers became familiar workplace additions in the 1980s in part because of their use for gaming, an activity that encouraged employees of a variety of ages and backgrounds to explore the various dimensions of the devices and to become more comfortable with them.

If engaged in constructively, online recreation can aid in awakening creativity and increasing wellbeing, just as appropriate and timely face-to-face diversions have restored employees' energies over the past decades. However, some individuals may not be able to deal with online recreation constructively. They indeed will use it in ways that affect their organizations and themselves negatively, just as some

individuals cannot perform adequately on the job for other reasons. Forms of "positive discipline" can be utilized if employees choose to exceed reasonable, agreed-upon limits; implementing such discipline "requires that the supervisor and employee work together to correct the problem behavior" (Guffey & Helms, 2001). Managers and employees should strive together to harness online recreation toward positive ends, rather than condemning or seeking to stifle it completely.

WHAT IS "CONSTRUCTIVE RECREATION"?

Online recreation has already served many supportive purposes in organizations; games can be used to help decrease computer anxiety as well as encourage experimentation and the early stages of learning (Kendall & Webster, 1997; Oravec, 1999; Webster & Martocchio, 1992). What would make online recreation optimally beneficial to individuals, project teams, and the organization as a whole? To start the discussion: recreation is "constructive" when it is in synch with pending work responsibilities, allowing individuals to use time not consumed by workplace demands in ways that equip them to face future tasks with greater energy and expanded perspectives. Constructive recreation is also in keeping with technological constraints, as exemplified by the organizations that allow online recreation but place limits during certain hours to avoid system overload (Verton, 2000). Policies established are developed in participatory ways, and are disseminated broadly. Constructing ways of assigning tasks and evaluating employees so that significant and meaningful measures of productivity are involved can lessen an emphasis on the "surface" behavior of employees. Other characteristics of constructive recreation initiatives include:

- ***fostering flexibility:*** A major impetus behind constructive recreation initiatives is facilitating the rapid adaptation of individuals to changing circumstances. Constructive recreation affords individuals the means to maintain their flexibility in workplace environments that place increasing demands on their capacities to withstand change.
- ***manifesting sensitivity to cultural concerns:*** Workplace recreation is also "constructive" to the extent in which it is responsive to the overall culture of the organization and sensitive to the needs and values of other organizational participants (including freedom from harassment). Requirements of project team members in terms of scheduling are especially critical to recognize since the synchronization and sustained involvement of everyone are required during critical periods.
- ***providing stimulation and refreshment:*** Along with its other aspects, recreation is constructive if it provides intellectual and psychological stimulation or support, the sustenance often needed to take on tough challenges. "Reclaimed moments" that individuals spend in such activity can allow them to reestablish senses of control in otherwise stressful and constraining contexts. Ability to access such recreation and thus momentarily escape can provide a safety valve for those who face unyielding situations or put in long work hours, thus putting the porousness of today's Internet-supported workplaces to good use.

FUTURE TRENDS

The value of recreation and play in adult realms is not well understood. Play has been given an assortment of definitions in the academic and research literatures (with examinations in the fields of social psychology, philosophy, and anthropology); it is often considered in both its adult and child modes as a "cognitive and symbolic act that is fundamental to the human representational process" (Myers, 1999). Across species as well as cultures, play has been shown to help individuals prepare for the unexpected by presenting varying streams of novel or challenging situations (Spinka, 2001). Play is generally considered as a support for children's intellectual and social development, but its role in adult lives is less clear. Research initiatives on what kinds of recreation and play are most efficacious in different workplace environments—as well as on individual and group "play styles"—could enlighten constructive recreation efforts (although they cannot be expected to provide definitive results).

Simulation is indeed an aspect of play that has some direct implications for employee readiness in the workplace, and it has received some research treatment (Myers, 1999). Michael Schrage's (1999) *Serious Play* examines how simulations expand the intellectual capacities of knowledge workers; forms of online play may equip individuals to utilize an organization's "serious" computer simulations more effectively, thus reinforcing skills applicable in many workplace contexts. Many powerful simulation games with societal or political themes are widely available to the public and have considerable audiences; the Sims series and other popular single- and multiplayer games have been used to entertain and educate in a variety of contexts (Pillay, Brownlee & Wilss, 1999).

Constructive recreation initiatives will also be a part of many organizational efforts to build cohesion. Managers have often used organizationally sanctioned recreation as a perquisite, a bonus for acceptable conduct. It has served as an extension of the workplace, providing a form of "social capital" (part of the "glue" that holds the at-work community together). Through the past century, many organizations

have sponsored picnics and celebrations with the strategy of increasing workplace cohesion (Putnam, 2000).

As employees (including many white collar as well as knowledge workers) telecommute or put in long and irregular hours, the adhesive that binds organizations has been increasingly conveyed through electronic channels. However, it is unclear what kinds of online activity can foster social capital (Uslaner, 2000). Just as human resource experts struggled early in the twentieth century to integrate face-to-face recreation into workplace contexts, organizations should attempt similar feats in online realms, thus making online recreation a shared and open resource rather than a secretive endeavor. Unlike many early human relations experiments, the recreational activities involved should be developed in a participatory (rather than patriarchal) fashion. Whether organization-approved fantasy football, discussion group and collaborative filtering forums, joke-of-the-day contests, or other recreations are ultimately successful will depend on how they fit into everyday working experiences.

CONCLUSION

Can we indeed construct a "level playing field"? As workplaces have evolved, so have the issues that have divided employers and managers. Conflict has ensued for decades on an assortment of matters relating to the quality of worklife, often leading to dysfunctional confrontations. Today, employees who guess wrong about online recreation standards— or choose to violate them— often pay large penalties, even being demoted or fired. Some managers have devised negative sanctions for these infringements far more severe than those applied to comparable face-to-face interaction. Office workers paging through paper catalogues in idle minutes rarely face the harsh penalties that those caught shopping online often encounter.

Hard-line positions against forms of online recreation may be required in some instances and directly related to important organizational goals. For instance, air traffic controllers should be expected to keep focused on landing real airplanes rather than escape into fantasy games during assigned hours. However, some hard-line restrictions can reflect fear or lack of understanding of online realms. Management may assume that online recreation will foster or encourage Internet addiction or related concerns. "Internet addiction" has become a widely identified syndrome, although its medical underpinnings are still in question (Beard, 2002; Oravec, 1996, 2000).

Ambiguities concerning online work and play in virtual realms are increasingly adding complexities to these issues (Broadfoot, 2001). It is often difficult to tell which Web sites are related to business needs and which are recreational; many have dual purposes, combining amusement with news and other serious pursuits. Slashdot.org has humorous material as well as valuable technical commentary, and abcnews.com has stories on upcoming movies as well as current economic results. Helpful intelligent agents (some with cartoon-like manifestations) can add levity to everyday tasks. Surfing the Internet for an answer to a question or fiddling with various programs can interfere with productive effort, as individuals dwell on technological nuances. Managers and employees need to deal not only with recreational concerns but also with broader issues of how to integrate computing into workplaces in ways that are engaging yet productive. However, online recreation should not be exploited as a means to keep individuals glued to workstations for indefinite periods in lieu of reasonable work schedules and functional work-life balances.

Solutions as to how to couple online work and play are emerging in organizations that are tailored to specific workplace contexts. Managers and employees are gaining important experience in resolving these issues as individuals perform activities away from direct supervision via mobile computing or virtual office configurations. Managers are learning how to perform their functions without direct employee surveillance. Employees are learning higher levels of self-discipline and the skills of balancing on-line work and play— just as they have learned to balance face-to-face schmoozing with task orientation in the physical world. Thus setting severe restrictions on online recreation can serve to slow down the process of understanding how to migrate the organization into virtual realms and establish trust. Responsibility and respect for others in these realms can be difficult to acquire, and many employees will indeed need direction.

Allowing for reasonable and humane amounts of online recreation can indeed have considerable advantages, both for the individuals involved and the organization as a whole. It can serve to open blocked creative channels and possibly relieve stress as well. Online recreation can also extend the limits of individuals' working days by providing extra dimensions to workplace activity. Rather than going through the emotional labor of looking busy, employees can utilize spare moments on the job in recharging their mental batteries. Constructive use of recreation will require a number of changes, such as increases in managerial flexibility and employee empowerment (Boswell, Moynihan, Roehling & Cavanaugh, 2001; Kanter, 2002). Organizational participants must learn how to handle the distractions and opportunities of increasingly porous workplaces, with their many external influences. Education and training can be useful in these initiatives: novice employees can be aided to couple work and recreation in ways that increase overall effectiveness. Constructive recreation strategies can bring these complex matters into the open, rather than allow them to be objects of rumor and fear.

Forms of online diversion are already becoming integral elements of everyday workplace life, often serving to human-

ize and enhance organizations. Negotiation and discourse on constructive recreation issues can increase mutual trust and respect concerning online as well as face-to-face activity. With effort on everyone's part, the constructive use of online recreation can help the entire organization work harder and play harder.

REFERENCES

Beard, K. (2002). Internet addiction: Current status and implications for employees. *Journal of Employment Counseling, 39*(1), 2-12.

Boswell, W., Moynihan, L., Roehling, M., & Cavanaugh, M. (2001). Responsibilities in the 'new employment relationship': An empirical test of an assumed phenomenon. *Journal of Managerial Issues, 13*(3), 307-328.

Broadfoot, K. (2001). When the cat's away, do the mice play? Control/autonomy in the virtual workplace. *Management Communication Quarterly, 15*(1), 110-115.

Colkin, E., & George, T. (2002, March 25). Teens skilled in technology will shape IT's future. *InformationWeek, 881,* 72-73.

Daniels, C. (2000, October 30). How to goof off at your next meeting. *Fortune, 142*(10), 289-290.

Greengard, S. (2000). The high cost of cyberslacking. *Workforce, 79*(12), 22-23.

Guffey, C., & Helms, M. (2001). Effective employee discipline: A case of the Internal Revenue Service. *Public Personnel Management, 30*(1), 111-128.

Kanter, R. (2002). Improvisational theater. *MIT Sloan Management Review, 43*(2), 76-82.

Kendall, J., & Webster, J. (1997). Computers and playfulness: Humorous, cognitive, and social playfulness in real and virtual workplaces— introduction to the special issue. *DATA BASE, 28*(2), 40-42.

Lim, V. (2002). The IT way of loafing on the job: Cyberloafing, neutralizing and organizational justice. *Journal of Organizational Behavior, 23*(5), 675-694.

Myers, G. (1999). Simulation, gaming, and the simulative. *Simulation & Gaming, 30*(4), 482-490.

Oravec, J. (1996). *Virtual individuals, virtual groups: Human dimensions of groupware and computer networking.* New York: Cambridge University Press.

Oravec, J. (1999). Working hard and playing hard: Constructive uses of on-line recreation. *Journal of General Management, 24*(3), 77-89.

Oravec, J. (2000). Internet and computer technology hazards: Perspectives for family counselling. *British Journal of Guidance and Counselling, 28*(3), 309-324.

Pillay, H., Brownlee, J., & Wilss, L. (1999). Cognition and recreational computer games: Implications for educational technology. *Journal of Research on Computing in Education, 32*(1), 203-217.

Putnam, R. (2000). *Bowling alone: The collapse and revival of American community.* New York: Simon & Schuster.

Schrage, M. (1999). *Serious play.* Cambridge, MA: Harvard Business School.

Spinka, M. (2001). Mammalian play: Training for the unexpected. *Quarterly Review of Biology, 76*(2), 141-169.

Uslaner, E. (2000). Social capital and the net. *Communications of the ACM, 43*(12), 60-64.

Verton, D. (2000, December 18). Employers OK with e-surfing. *Computerworld, 34*(51), 1-2.

Webster, J., & Martocchio, J. (1992). Microcomputer playfulness: Development of a measure with workplace implications. *MIS Quarterly, 16*(2), 201-226.

Weil, M., & Rosen, L. (1997). *TechnoStress: Coping with technology @ work @ home @ play.* New York: John Wiley & Sons.

KEY TERMS

Flexible Workplace: Organizational settings that can quickly take external and internal changes into account in their processes.

Internet Addiction: Use of the Internet and network resources that undermines the fulfillment of some of an individual's basic human needs.

Organizational Policies: Openly-stated, officially-sanctioned rules for organizational resource usage and other kinds of organization-related conduct.

Participatory Management: Management in which the input of employees as well as managers is thoughtfully taken into account in setting organizational policies and developing organizational structures.

Play: Activities in which individuals and groups engage that stimulate various aspects of personal and social functioning without necessarily being related to particular utilitarian outcomes.

Simulation Games: Games in which important aspects of a system are modeled so that game participants can engage in activities and deal with events that are comparable to those that system participants would encounter.

Social Capital: Social closeness, mutual knowledge, and cohesion that are a product of a wide assortment of different kinds of informal, volunteer, and partially-structured social interactions.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1070-1074, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Enterprise Resource Planning (ERP) Maintenance Metrics for Management

Celeste See-pui Ng
Yuan-Ze University, R.O.C.

INTRODUCTION

A typical packaged software lifecycle, from the client-organization perspective, is packaged software selection followed by implementation, installation, training, and maintenance (that includes upgrades). Traditional software maintenance has been acknowledged by many researchers as the longest and most costly phase in the software lifecycle. This fact is no exception in the ERP packaged software maintenance context (Moore, 2005; Whiting, 2006).

According to Ng, Gable, & Chan (2002, pg. 100) ERP maintenance is defined as "post-implementation activities related to the packaged application software undertaken by the client-organization from the time the system goes live (i.e., successfully implemented and transported to the production environment), until it is retired from an organization's production system, to keep the system running; adapt to a changed environment in order to operate well; provide helps to the system users in using the system; realize benefits from the system (best business processes, enhanced system integration, cost reduction); and keep the system a supported-version and meet the vendor's requirements for standard code. These activities include: implementing internal change-requests (initiated by an ERP-using organization's system users and IT-staff); responding or handling user-support requests (initiated by an ERP-using organization's system users); upgrading to new versions/releases (introduced by the vendor); and performing patches (support provided by the vendor)."

In order to achieve the abovementioned maintenance objectives of keeping the ERP system running, adapting the system to a new operating environment, and ensuring the system up to the vendor's requirement for standard code; and realizing benefits such as competitive advantages from the system, the IT department staff has to collect some metrics or relevant data on patches and modifications done to the ERP system so that they can know or can tell the status and the performance of their maintenance activities. The authors in Fenton (1991), Fenton & Pfleeger (1997), and Florac (1992), agree that software maintenance data are useful for planning, assessment, tracking, and predictions on software maintenance. Although, there is a lot of literature on ERP, we find almost no literature on ERP maintenance metrics. Thus, this text is meant to provide some fundamental metrics on ERP patches and modifications which could be useful for ERP maintenance management in order to answer questions on the state of their ERP system, their patch implementation costs, and the ongoing maintenance costs for their previous modification or custom development.

BACKGROUND: METRIC

The *IEEE Standard Glossary of Software Engineering Terminology* (1990) defines a metric as a "quantitative measure of the degree to which a system, component, or process possesses a given attribute." Based on the definition, this text interprets a metric as being derived from data, and as quantifiable, meaningful, and used for strategic, tactical and/or operational purposes. Data (or data item), in turn, is defined as a quantitative indication of the extent, amount, dimension, capacity, size, or characteristic of particular attributes of a task or activity in a process. It can be collected using forms (e.g., change request form, change report, software engineering report), interviews (with the users, testers, programmers, analysts, managers), and via computerized systems (e.g., the in-built change management system in ERP, change request database). A goal/question/metric (GQM) paradigm is a systematic way of collecting predefined data, with intended goal(s), and the associated sets of predetermined questions, in order to derive the anticipated measurable metrics. Basili and Weiss (1984) advocate this methodology for collecting valid data. In GQM, which is also known as the top-down approach, the timing (in terms of the software life cycle or activity), interviewees, and reasons for collection are all predetermined.

The literature reports that successful use of measurement/metric program avoids recurring errors (Ebert, Dumke, Bundschuh, & Schmietendorf, 2005), improves software maintenance processes at Burrough Corporation (Rombach & Ulery, 1989), and Motorola (Smith, 1993), and improves product quality at AT&T (Fenton & Pfleeger, 1997).

On the flip side, a known metric – together with the context for its interpretation – can determine what data might be collected (Rombach & Ulery, 1989). There are three main purposes of metrics: assessment, prediction, and control.

Table 1. Main characteristics and purpose of three main metric categories

Metric category	Main characteristic	Purpose	Example of ERP maintenance metric
Assessment	Informative (about person, process, object); may be used in decision making or for controlling purposes	Retain knowledge	Programmer ID, problem description, description of changes, issues of consideration, patch ID
Prediction or decision-making	Conclusive; most likely derived from the assessment and control metrics; usually describes what should be done	Make decision, planning and estimation	Estimated maintenance time, quotation for a maintenance request, action to be taken, maintenance request type, projected availability
Control	Indicative (indicating that something needs to be done); most likely used to pinpoint that a particular decision needs to be made; usually requires data to be collected over a period of time; usually has some attached baseline value	Monitor performance, track progress, identify problem	Problem status, approved by, accepted by, maintenance request ID, time of problem occurrence, resolution impact

Table 2. Application of software metrics in practice

Purpose	Case name	Metrics used	Goal
Assessment	Hewlett-Packard (Wood, 2003).	Event chronology, problem symptoms, diagnostic information, release version information	Problem analysis and resolution
Decision-making	NASA's Mission Operations Directorate (Stark, Durst, & Vowell, 1994)	Previous project delivery rate	Estimate a test schedule
	Hewlett-Packard to (Grady, 1994).	Defect trend	Determine time to release a product
Control and monitoring	NASA's Mission Operations Directorate (Stark, et al., 1994)	Earned value management technique: project cost and schedule	Monitor project cost and schedule performance
		Defect density	Track quality in subsystem, efficiency in testing, and backlog of fault
	Hewlett-Packard (Grady, 1994)	Code size and time	Monitor project progress
	Bull's Arizona (Weller, 1994)	Effort, resources, product size, estimated completion date and defect detected	Manage project and improve project planning
	Siemens (Paulish & Carleton, 1994)	Defect rate (i.e. number of defect/ product size)	Measure performance
		Product size and effort (i.e. product size/effort)	Project productivity
		Productivity gain, error detection rate, and reduction in time to market	Measure software process improvement

Table 1 shows the main characteristics and objectives of these three metric categories. In addition to this, their application in practice is provided in Table 2.

APPLYING METRICS IN THE ERP MAINTENANCE (AND UPGRADE) CONTEXT: AN IMPLICATION FOR MANAGEMENT

The following discussion focuses on the practical application of the data items or software metrics. As stated earlier, metrics are derived from the data items collected from the activities that produce them. The metrics covered in this section are: (1) the state of the current system relative to the vendor's expectation, (2) the number of modifications or custom developments likely to be affected by each patch implementation, (3) patch implementation costs, (4) effort needed to reapply a previous modification or custom development, and (5) ongoing maintenance costs for modification requests. As shown later, these metrics are developed by manipulating (summation, division, multiplication, etc.) simple data items that are directly collected. Maintenance managers can use these metrics to manage maintenance activities for tactical and operational management decision-making. Table 3 provides the notations used in the abovementioned five metrics.

In addressing the manager question of: How up-to-date is my ERP system compared to my vendor's standard code?

With the aim of achieving economies-of-scale, many organizations "batch" the patches instead of implementing them as they arrive. Thus, in order to measure how up-to-date or state-of-the-art an installed ERP system is relative to the vendor's patch introduction for a given version, one needs to know how many patches have already been implemented, that is:

$$\textit{State-of-the-art} = \frac{R_A}{R_P} \text{X } 100\% \quad (1)$$

In addressing the manager question of: On average, how many modifications are affected by a patch implementation and what does it cost to implement a patch?

Table 3. Notations for metrics

Context	Notation	Metrics description	How it is measured?
User organization	k	Probability that a modification or custom development will be affected by future patch implementation	Done by studying the 'description of changes', 'related changes' and 'issues of consideration' related to that modification
	L	Labor rate	[a direct data item]
Previous modification	M	Number of modifications and/or custom developments	Summing up all the previous modifications based on previous record on it
	T_I	Effort required to conduct impact analysis for a modification	[a direct data item]
	T_R	Effort required to reapply a modification	[a direct data item]
	$T_{A\&T}$	Effort required to apply and conduct complete testing for a modification	[a direct data item]
Patch	N	Number of patch projects	Summing up all the previous patch projects based on previous record on it
	p_i	Number of patches in the *i*-th project	Each project may have different number of patched implemented at a time. This is usually depending on economy-of-scales and/or urgency of some patches
	T_{Pi}	Average effort required to implement a patch in the *i*-th project	Dividing the total effort in the *i*-th project by the total number of patches in the *i*-th project
	R_P	Patch introduction rate	Number of patches introduced by the vendor per year
	R_A	Patch implementation/application rate	Number of patches implemented by the client-organization per year

The *IEEE 1219 – Standard for Software Maintenance* (1998) suggests that cost and benefit analysis is required in order to determine the feasibility of a maintenance request, and to quantify the long-term or ongoing cost of a maintenance request. Patch implementation effort is believed to increase, to some degree, depending on the number of modifications already made to an ERP system (Ng, 2001). Patch implementation involves replacing some portion of the existing customized ERP code with the vendor's standard code. It can overwrite existing modifications or custom developments. However, not all modifications or custom developments are necessarily affected by each patch implementation. It is observed in a study by Ng (2001) based on an upgrade experience that only certain portion of them are affected. (Note that although a patch implementation is a small scale of an upgrade implementation, both of them could have overwritten effects on modifications done on existing system.) Therefore, the number of previous modifications or custom developments probably affected by each patch implementation, *m*, is:

$$m = k \text{ X } M \tag{2}$$

As a result,

Patch implementation costs = average number of patches per project X *average effort per patch* X *labor rate* + *effort required to reapply previous modifications* X *labor rate, or:*

$$= \frac{\sum_{i=1}^{N} p_i}{N} \text{ X } \frac{\sum_{i=1}^{N} T_{P_i} \text{ x } p_i}{\sum_{i=1}^{N} p_i} \text{ X } L + \sum_{i=1}^{m} T_R \text{ x } L \tag{3}$$

In addressing the manager question of: How much efforts is needed to reapply previous modification, and what are the ongoing maintenance costs for a modification request?

As modifications or custom developments may be overwritten during patch implementation, extra effort is needed in order to check whether such overwriting has occurred. If the affected modifications and custom developments are still needed, they may have to be reapplied to ensure that they can operate as they did before the patch(es) were implemented. Although activities such as analysis, design, and coding do not need to be repeated, complete testing (involving integration, validation, and system testing) for the previous modifications is required. These additional efforts generate the so-called ongoing maintenance costs for modification request. Hence, the major effort to reapply a single previous modification or custom development, T_R, is:

$$T_R = \quad T_I \quad + \quad T_{A\&T} \tag{4}$$

The greater the number of patch projects that must be carried out in an existing system, the greater the likelihood of ongoing maintenance costs associated with a modification in that system. Assuming that all modifications and custom developments will be affected by the patch and that the vendor will incorporate these modifications and custom developments into the new version. they are no longer needed after the upgrade. A simple formula for ongoing maintenance costs for a modification is as follows:

Ongoing maintenance costs = Number of patch project X *effort to reapply the modification* X *labor rate*

$$= \quad N \text{ X } T_R \text{ X } L \tag{5}$$

Thus, with these metrics, maintenance manager will be in a better position to charge the ongoing maintenance cost of the modification to the respective user department. Alternatively, this information can be used to convince the system user department to delay or forego the maintenance, based on a more accurate assessment of total perceived benefits and estimated costs.

FUTURE TRENDS

Many organizations start to recognize the importance of collecting ERP system's performance metrics such as inventory level, operational costs, schedule compliance, and on-time delivery (Jutras, 2007). Like other software systems, ERP system is not a panacea for all. ERP is built to be generic in order to meet a wide range of clients needs. Thus, while some clients try to change their business processes, some change the standard code in order to incorporate their unique business processes where they perceive critical for competitive advantages. Also, although ERP software provides a comprehensive set of business applications that can serve different departments under a single system, most companies still maintain their idiosyncrasy legacy systems and some buy best-of-breed packaged software from different vendors. These are because of unwillingness or resistance to replace a system that is still working, and the packaged software does not meet their requirements or is simply not good enough. Besides, there could also be owing to economy slowdown, wanting to avoid vendor lock-in, waiting for the Web-application (erpwire.com, 2006b) and mobile-application (erpwire.com, 2006a) to mature, and so forth. However, multi-system leads to integration problem. As a consequence, time and effort need to be invested in order to make disparate systems talk to each other. Some perceive that the answer to this is the Web and mobile application. But, as far as Web application is concerned, there are a number of issues yet to consider and address, that is, cost, benefit, network bandwidth and

infrastructure, threats and risks, security, and ethics. Also, developing mobile ERP application can be very complicated due to non-standardization in mobile devices (e.g., various sizes and shapes), and small screen display; it's not simply the broadband or bandwidth boost issue. Moreover, in light of growing customer base of the ERP system, it is valuable to investigate how ERP maintenance and upgrade could be better managed. How patches implementation can be better queued and implemented to achieve economy-of-scale? What and how process model will best describe activities in ERP maintenance and upgrade?

REFERENCES

Basili, V. R., & Weiss, D. M. (1984). A Methodology for Collecting Valid Software Engineering Data. *IEEE Transactions on Software Engineering, 10*(6), 728-738.

Ebert, C., Dumke, R., Bundschuh, M., & Schmietendorf, A. (2005). *Best Practices in Software Measurement: How to Use Metrics to Improve Project and Process Performance*. Berlin Heidelberg: Springer-Verlag.

erpwire.com. (2006a). *The Advancement of Wireless Technology in ERP*. erpwire.com. Retrieved January 16, 2008, from the World Wide Web: http://www.erpwire.com/erp-articles/wireless-erp.htm

erpwire.com. (2006b). *What Are the Facilities Offered by Web Enabled ERP services?* erpwire.com. Retrieved January 16, 2008, from the World Wide Web: http://www.erpwire.com/erp-articles/web-enabled-erp.htm

Fenton, N. E. (1991). *Software Metrics: A Rigorous Approach*. London: Chapman & Hall.

Fenton, N. E., & Pfleeger, S. L. (1997). *Software Metrics: A Rigorous & Practical Approach* (2nd Ed.). Boston, MA: PWS Publishing Company.

Florac, W. A. (1992). *Software Quality Measurement: A Framework for Counting Problems and Defects.* (Technical Paper CMU/SEI-92-TR-22). Pittsburgh, Pennsylvania: Software Engineering Institute (SEI), Carnegie Mellon University.

Grady, R. B. (1994). Successfully Applying Software Metrics. *IEEE Computer, 27*(9), 18-25.

IEEE. (1990). *IEEE Standard Glossary of Software Engineering Terminology - IEEE Std 610.12-1990*. New York: The Institute of Electrical and Electronics Engineers.

IEEE. (1998). *IEEE Standard for Software Maintenance, IEEE Std 1219-1998*. New York: Institute of Electrical and Electronics Engineers.

Jutras, C. (2007). *ERP in SMB: Exploring Growth Strategies* (Report). Aberdeen Group.

Moore, J. (2005). Negotiating ERP Maintenance Contracts. *Chemical Engineering Progress, 100*(5), 16.

Ng, C. S. P. (2001). A Decision Framework for Enterprise Resource Planning Maintenance and Upgrade: A Client Perspective. *Journal of Software Maintenance and Evolution: Research and Practice, 13*(6), 431-468.

Ng, C. S. P., Gable, G. G., & Chan, T. (2002). An ERP-client Benefit-oriented Maintenance Taxonomy. *Journal of Systems and Software, 64*(2), 87-109.

Paulish, D. J., & Carleton, A. D. (1994). Case Studies of Software-Process-Improvement Measurement. *IEEE Computer, 27*(9), 50-57.

Rombach, H. D., & Ulery, B. T. (1989). Improving Software Maintenance Through Measurement. *Proceedings of the IEEE, 77*(4), 581-595.

Smith, B. (1993). Six-sigma Design. *IEEE Spectrum, 30*(9), 43-47.

Stark, G., Durst, R. C., & Vowell, C. W. (1994). Using Metrics in Management Decision Making. *IEEE Computer, 27*(9), 42-48.

Weller, E. F. (1994). Using Metrics to Manage Software Projects. *IEEE Computer, 27*(9), 27-33.

Whiting, R. (2006). *Lower-Cost Options Free IT From Software Maintenance Fees*. InformationWeek. Retrieved January 15, 2008, from the World Wide Web: http://www.informationweek.com/shared/printableArticle.jhtml?articleID=192300361

Wood, A. P. (2003). Software Reliability. *Computer, 36*(8), 37-42.

KEY TERMS

ERP Modification: A type of maintenance request which results in changes being made to the existing ERP (standard) code and custom objects being created.

ERP Software Maintenance Data (or Data Item): ERP software maintenance quantitative indication of the extent, amount, dimension, capacity, size, or characteristic of particular attributes of a task or activity in a process.

ERP Software Maintenance Metric: ERP software maintenance measure that is derived from data, and is quantifiable, meaningful and used for strategic, tactical, and/or operational purposes.

ERP Vendor's Standard Code: Program code that is

without any modification or custom development at all.

ERP Vendor's Supported-Version: An ERP version that is still supported by the ERP vendor. This means that the vendor will provide patches for bug fixes and minor enhancements for this version.

Ongoing Maintenance Costs: Additional costs, besides the initial implementation costs, incurred as a result of upkeeping some software code or objects created in previous maintenance request.

Patch Implementation Costs: Costs incurred in implementing (usually) a batch of patches provided by the ERP vendor. As custom code or previous modifications may be overwritten while implementing the patches, usually these costs also include the costs of reapplying previous modifications.

Enterprise Resource Planning and Integration

Karl Kurbel
European University - Frankfurt (Oder), Germany

INTRODUCTION

Enterprise resource planning (ERP) is a state-of-the-art approach to running organizations with the help of comprehensive information systems, providing support for key business processes and more general, for electronic business (e-business). ERP has evolved from earlier approaches, in particular, materials requirement planning (MRP) and manufacturing resource planning (called MRP II) in the 1980s. The focus of MRP and MRP II was on manufacturing firms. The essential problem that MRP attacked was to determine suitable quantities of all parts and materials needed to produce a given master production schedule (also called a "production program"), plus the dates and times when those quantities had to be available. Application packages for MRP have been available from the 1960s on. In the beginning, they were mostly provided by hardware vendors like IBM, Honeywell Bull, Digital Equipment, Siemens, etc. MRP was later expanded to *closed-loop MRP* to include capacity planning, shop floor control, and purchasing, because as Oliver Wight (1884) puts it: "Knowing what material was needed was fine, but if the capacity wasn't available, the proper material couldn't be produced" (p. 48).

The next step in the evolution was *MRP II (manufacturing resource planning)*. According to the father of MRP II, Oliver Wight, top management involvement in the planning is indispensable. Therefore, MRP II expands closed-loop MRP "to include the financial numbers that management needs to run the business and a simulation capability" (Wight, 1984, p. 54).

Enterprise resource planning (ERP) has its roots in the earlier MRP II concepts, but it extends those concepts substantially into two directions. ERP takes into account that other types of enterprises than those producing physical goods need comprehensive information system (IS) support as well, and even in the manufacturing industry, there are more areas than those directly related to the production of goods that are critical for the success of a business.

BACKGROUND OF ERP

The key issue of ERP is integration (Langenwalter, 1999). While stand-alone solutions—sometimes quite sophisticated information systems—for various areas of a business have been available before, ERP takes a holistic approach. Instead of isolated views—on procurement, on manufacturing, on sales and distribution, on accounting, etc.—the focus is now on integrating those functional areas (Scheer & Habermann, 2000). The need for integrated systems has been recognized by many, but Germany-based SAP AG was the first to put them into reality. SAP's early success as worldwide market leader comes largely from the fact that this company actually designed and implemented business-wide integrated information systems. The lack of integration of information systems has created a variety of problems. The most serious ones are the following:

- Redundancy (i.e., the same information is stored and maintained several times)
- Inconsistency (i.e., information about the same entity stored in different places is not the same)
- Lack of integrity (i.e., databases where such information is stored are not correct)

Mistakes, wrong decisions, and additional work are some of the consequences resulting from these problems. Consider, for example, data about customers. Such data are often entered and maintained in a sales and distribution information system (customer orders), then again in the dispatching system (delivery orders), and perhaps once more in a financial accounting system (invoices). Not only is this redundant and means additional work, but also the same attributes may even stand for different things. For example, an "address" field in the sales and distribution system may represent the address of the customer's procurement department, whereas "address" in the dispatching system is the place where the goods have to be delivered.

Integration of information systems can be considered from several perspectives: from the data, the functions, the operations, the processes, the methods, and the software perspectives. The most important aspects are data integration, operations integration, process integration, and software integration:

- **Integration** of *data* means that data models and databases are unified so that all departments of an enterprise use the same data entities, with the same values.

- **Integration** of *operations* requires connecting individual operations, or steps of a business process, with preceding or succeeding operations, respectively.
- **Integration** of *processes* means that interfaces between different business processes are explicitly considered (e.g., connections between order processing and flow of material control).
- **Integration** of *software* means that different programs (e.g., information systems for different business functions, can run together and use each other's data and operations.

Those aspects of integration have always been considered important requirements for effective business information processing, but how does one actually obtain enterprise-wide integrated information systems?

Because most organizations have been using information systems in various business areas for quite some time, one way is to integrate those stand-alone systems subsequently. This approach has been discussed and practiced under the concept of "software reengineering," often related to the term "legacy systems" for the information systems to be integrated (Miller, 1997; Seacord et al., 2003).

The other approach to obtain integrated information systems is obviously to start developing them from scratch. In such a situation, information structures can be modelled and designed on the drawing board in an enterprise-wide manner, at least in theory. Practical experiences have shown that developing comprehensive information systems for all areas of a business is a giant task. That is why such systems have rarely been developed as individual solutions. Not only is the investment needed very high, but also manpower and know-how to develop such systems are often beyond the means of a single company. Therefore, comprehensive integrated information systems have mostly been developed by dedicated software and consulting companies. In the 1970s and 1980s, those systems were named with rather general terms, like standard packages or integrated business information systems, until the terms "enterprise resource planning" and "ERP system" emerged in the 1990s. In fact, the term "enterprise resource planning" has been coined by the software industry and not by academia.

Today there is a common understanding of what the term stands for. The definition used in this article is as follows: An enterprise resource planning system (ERP system) is a comprehensive information system that collects, processes, and provides information about all parts of an enterprise, automating business processes and business rules within and across business functions partly or completely.

Alternatively, an ERP system may be defined as a set of integrated information systems rather than as one system. This depends on the perspective of the viewer. For the user, an ideal ERP system will behave like one enterprise-wide information system, with one database, and one common user interface. Nevertheless, such a system may be composed of many subsystems and many databases, as long as they are well integrated.

COMPONENTS OF AN ERP SYSTEM

Horizontal and Vertical Views of Enterprise Resource Planning

An ERP system integrates information, processes, functions, and people into one coherent system (Brady et al., 2001). Such a system supports all horizontal business functions and all vertical levels of a business (operational, tactical, and strategic). Figure 1 illustrates this view in a simplified information systems pyramid. Each component may be seen as a functional subsystem. In a horizontal perspective, systems are integrated along the value chain. The vertical direction asks for integration of operative systems with their corresponding value-oriented accounting systems; reporting and controlling systems; analysis and management information systems; and long-term planning and decision support systems (Scheer, 1994, p. 5).

A typical ERP system provides components like the ones shown in Figure 1, arranged and extended in one way or another. As an example of integrated information systems, the mySAP ERP system is described subsequently.

An Example: SAP ERP

SAP ERP has evolved from SAP R/3 which is still the most frequently installed ERP system. SAP ERP is based on SAP NetWeaver as technology platform (SAP, 2006b). Encompassing all levels of the pyramid, SAP ERP is logically structured into the following modules (short descriptions are taken from SAP, 2006a, 2006b):

Analytics

- Strategic enterprise management—Supports the top level of the pyramid in Figure 1: integrated strategic planning, performance monitoring, business consolidation, and stakeholder communication; provides tools for planning and executing the strategies: balanced scorecard, value-based management, financial statement planning, risk management, investment planning, and more
- Business analytics (financial, operations, workforce analytics)—Supports managers with methods and tools for financial and management reporting, financial planning, budgeting and forecasting, profitability management, product and service cost management,

Figure 1. Integrated business information systems (Scheer, 1994, p. 5)

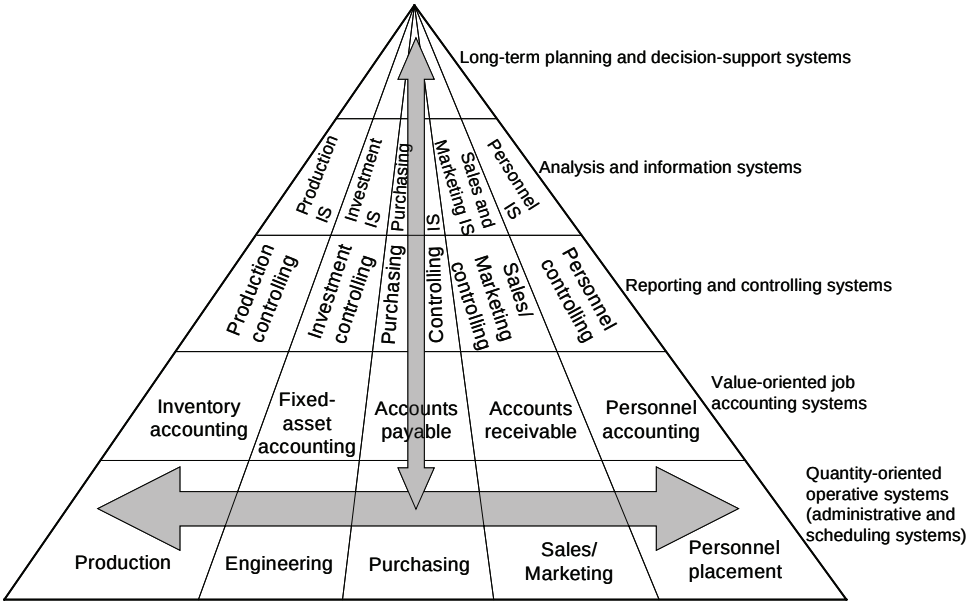
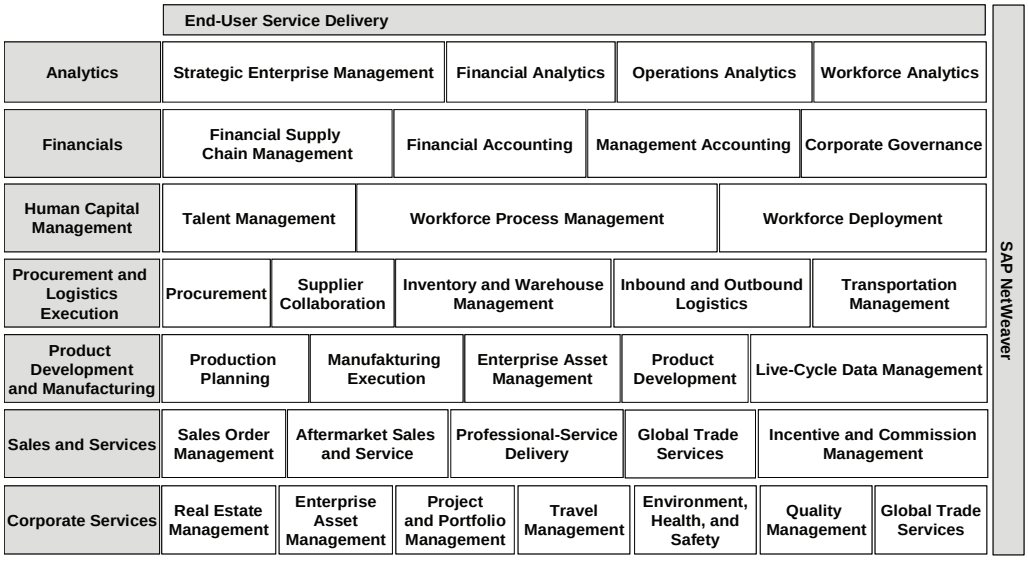


Figure 2. Application domains and modules of SAP ERP (SAP, 2006b)



overhead cost management, working capital and cash-flow management, etc.; provides analytical functions for procurement, inventory and warehouse management, manufacturing, transport, sales, customer service, quality management, enterprise asset management, program and project management, and more. [In Figure 1, those functions, tools, and methods belong mostly to levels 2 (analysis and information systems) and 3 (reporting and controlling information systems).]

Financials

The Financials module supports several application areas that are located on level 4 of the pyramid:

- **Financial accounting:** Processing of incoming and outgoing payments, cash flows; provides general ledger, accounts receivable, accounts payable, fixed assets accounting, inventory accounting, tax accounting, financial statements, and more; helps to monitor financial transactions; supports business analysis through combined planning, reporting, and analysis of competitive measures
- **Managerial accounting:** Provides profit center accounting, cost center and internal order accounting, project accounting, product cost accounting, profitability accounting; supports investment management, revenue and cost planning, transfer pricing, etc.
- **Financial supply chain management:** Supports financial collaboration within the enterprise and its business networks; provides credit management, cash and liquidity management, treasury and risk management, and more
- **Corporate governance:** Helps to ensure that the objectives, rules and regulations under which the corporation is directed and controlled are being followed and met.

Human Resources

The human resources module is comprehensive, including functionalities of the operative, administrative level (level 5) and the value-oriented level (level 4).

- **Talent management:** Supports recruiting and talent management, performance management, compensation management for various modes (e.g., performance- and competency-based pay)
- **Workforce process management:** Provides the central repository for employee data; integrates the information with other SAP business applications, especially Financials and Operations; supports time and attendance processing (planning, managing, and evaluating the working times and activities of internal and external employees); handles working-time provisions determined by companies themselves, by standard agreements, or required by law; handles all payroll processes, supports current legal regulations and collective agreement specifications, and ensures compliance with regulatory changes
- **Workforce deployment:** Provides project resource planning, resource and program management (i.e., resource management, project portfolio management, project execution, and skills management) and specific solutions for retail personnel and call centers.

Procurement and Logistics Execution

This module supports the bottom level of the pyramid through quantity-oriented operative subsystems for daily operations, including support for planning and execution.

- **Purchase order management:** Provides conversion from demands to purchase orders, issuance, and confirmation of purchase orders; supports purchasing of materials and services (for example, subcontracting for components)
- **Supplier cooperation:** Helps to streamline information supply and communication with and among suppliers
- **Inventory and warehouse management:** Comprises warehousing and storing (warehouse-internal movements and storage of materials) and managing physical inventory for the company's own stocks (periodic, continous, etc.)
- **Inbound and outbound logistics:** Supports inbound processing (all the steps of an external procurement process that occur when the goods are received) and outbound processing (all steps to p**repare and ship goods to their destination)**
- **Transportation management:** Provides transportation planning (routing, carrier selection, etc.) and execution (shipment orders), freight costing, and legal services

Product Development and Manufacturing

With the help of this module, engineering and design, creating relevant product data, and planning and executing manufacturing operations are supported.

- **Product planning:** Provides typical MRP (materials requirement planning) functionality: computing quantities and due dates for production orders and purchase requisitions through lead-time scheduling,

depending on buffers, operation times, lot-sizing rules and so on

- **Manufacturing execution:** Supports the process of capturing actual production information from the shop floor to support production control and costing processes; supports a variety of concepts: make-to-order, repetitive manufacturing, flow manufacturing, shop-floor manufacturing, lean manufacturing, process manufacturing, and batch manufacturing
- **Product development and life-cycle management:** Cover the life-cycle of product related master data such as product structures, routings and documents from invention to phase-out. Provide document management, product structure management (including bills of materials), recipe management, integration of CAD (computer-aided design), PDM (product data management), and GIS (geographical information system) data, and more.

Sales and Services

The Sales and service module addresses the customer focusing processes like selling products and services and providing aftermarket services.

- **Sales order management:** Supports quotation and order management (creating and processing orders, including pricing and scheduling orders for fulfillment) including inquiries and follow-up orders; provides mobile-sales, billing, and contract-management functionalities
- **Aftermarket sales end service:** Helps to manage a customer's product configuration (installation and configuration management, including definition of the product hierarchy, management of serial numbers, measurements, document management, and engineering change management), service contracts, planned services, warranties, and so on
- **Professional-service delivery:** Provides capabilities for selling, planning, delivering and billing project-based services
- **Incentive and commission management:** Helps to design incentive compensation plans, calculate variable compensation (e.g., direct sales commissions), carry out evaluations of performance and cost results, and more.

Corporate Services

The Corporate Services module provides comprehensive support for resource-intensive corporate functions including the following:

- **Real estate management:** Provides tools to support real-estate property acquisition and disposal, property portfolios, functions to help users lease and manage the real estate portfolio, etc.
- **Project and portfolio:** Provides functions for project (project structures, costs, budgets, workforce and resource planning, scheduling activities, etc.), project execution (monitoring project progress analysis/earned value analysis, progress tracking, etc.)
- **Travel management:** Provides functions for travel request and pre-trip approval, travel planning and online booking, travel expense management, services for mobile staff, and more
- **Environment, health, and safety:** Supports a variety of functions for product safety, handling of hazardous substances, transportation of dangerous goods, waste management, etc.; allows companies to take preventive care of their employees' health; schedules medical examinations and testing for workers; manages emissions for air, water, and soil; monitors and controls plant emission sources.
- **Quality management:** Supports quality engineering according to the ISO 9000 standard, quality assurance and control (quality inspections, statistical process control, traceability, etc.), quality improvement (audit management according to ISO 19011, problem/complaint management, corrective and preventive action, etc.)

This overview of components of the SAP ERP system illustrates the wide range of functions that support ERP nowadays. The reader interested in details of ERP functionality is encouraged to study the products' Web sites. Descriptions of the above functions can be found in SAP's electronic and printed documentation (e.g., SAP, 2006a, 2006b). Other vendors' ERP systems are outlined below.

THE MARKET FOR ERP

ERP systems have been around for about two decades now. All large and medium-size companies use such systems today, and more small companies are catching up. Well-known ERP systems in use include R/3 Enterprise, and SAP ERP (by SAP AG, Germany), Oracle's E-Business Suite (by Oracle, USA), Microsoft Dynamics, and infor: COM (by Infor Global Solutions, formerly Germany, now USA). Well-known ERP systems such as iBaan Enterprise, J.D. Edwards, PeopleSoft, Navision and many more have disappeared, either bought by competitors or gone up into other systems of the vendor.

Names and vendors are changing, as there is plenty of dynamic in the ERP market. Mergers and acquisitions and an ongoing market concentration can be observed year by year.

Although there is still a fairly large number of ERP vendors, a handful of them dominate the ERP market worldwide. The global market leader by far is SAP AG from Walldorf (Germany). Followers are Oracle, Microsoft, Infor, and Sage.

FUTURE TRENDS

While ERP continues to be the core of any integrated business software, the focus has shifted toward advanced user support, like Business Intelligence (Biere, 2003) and Knowledge Management (Davenport et al., 1998; Earl, 2001), and inter-organizational support of electronic business, in particular, supply chain management and customer relationship management. As those areas are closely related to ERP, all major vendors have extended their systems to support them as well. *Customer Relationship Management (CRM)* is an approach to develop a coherent, integrated view of all relationships a firm maintains with its existing andpotential customers (Laudon, 2007, p. 59). Nowadays, many channels are available for enterprises and customers to be in contact with each other: retail stores, telephone, e-mail, electronic shopping on the Web, mobile devices, etc. CRM systems try to consolidate customer information from all those channels and integrate the firm's diverse customer-related processes.

The major focus of ERP is to support the internal business processes of an organization. However, business activities do not end at the limits of one's own company. A natural extension of ERP is, therefore, s*upply chain management* (SCM; Ayers, 2001). SCM looks at the organization's business partners, in particular at the suppliers and their suppliers. In addition, many methodological and technical shortcomings of ERP have been removed or at least improved in SCM. Those improvements have been discussed in the literature under "Advanced Planning and Scheduling" (e.g., Meyr et al., 2002) and were implemented in SCM solutions by SCM vendors.

One example of such improvements is the use of optimization methods, like linear programming, mixed- integer programming, constraint propagation, and heuristics like genetic algorithms to solve production and distribution planning problems. Another example is pegging, i.e., creating, maintaining, and evaluating relationships between purchasing orders, production orders, transportation orders, and customer orders across entire supply networks worldwide.

CONCLUSION

ERP is a comprehensive approach to running organizations with the help of computer-based information systems. Such systems, so-called ERP systems, support all major areas of a business. ERP systems are integrated systems, with respect to information, processes, functions, and people. Usually, those systems are very large, developed by specialized software firms. The worldwide market leader is, by far, Germany-based SAP AG, with its current systems R/3 Enterprise and SAP ERP. Except for small enterprises, almost all of today's companies use ERP systems for their ongoing business.

Extensions of ERP systems can be observed at the former limits of ERP: in CRM, the focus is on dedicated customer support; the focus of SCM is to help organizations plan worldwide supplier–buyer networks and to act successfully within such networks; business intelligence enables enterprises to behave in an "intelligent" way (e.g., addressing the most promising customers well-aimed); knowledge management helps to formalize and preserve important knowledge in an organization; and information system support for key areas, like the before-mentioned areas, is provided both by ERP vendors and by specialized software and consulting firms

REFERENCES

Arnold, R. S. (1993). *Software reengineering*. Los Alamitos, CA: IEEE Computer Society Press.

Ayers, J. B. (2001). *Handbook of supply chain management*. Boca Raton, FL: St. Lucie Press.

Biere, M. (2003). *Business intelligence for the enterprise*. Upper Saddle River, NJ: Prentice Hall.

Brady, J., Monk, E. F., & Wagner, B. J. (2001). *Concepts in enterprise resource planning*. Boston: Course Technology.

Davenport, T. H., Delong, D. W., & Beers, M. C. (1998). Successful knowledge management projects. *Sloan Management Review, 39*(2), 43–57.

Earl, M. (2001). Knowledge management strategies: Toward a taxonomy. *Journal of Management Information Systems, 18*(1), 215-233.

Langenwalter, G. A. (1999). *Enterprise resources planning and beyond: Integrating your entire organization*. Boca Raton, FL: Saint Lucie Press.

Laudon, K. C., & Laudon, J. P. (2007). *Management information systems. Managing the digital firm* (10th ed.). Upper Saddle River, NJ: Prentice Hall.

Meyr, H., Wagner, M., & Rohde, J. (2002). Structure of advanced planning systems. In H. Stadtler & C. Kilger (Eds.), *Supply chain management and advanced planning* (2nd ed.) (pp. 99-104). New York: Springer.

Miller, H. (1997). *Reengineering legacy software systems*. Woburn, MA: Digital Press.

SAP. (2006a). *MySAP ERP—Solution overview*. Walldorf, Germany: SAP AG.

SAP. (2006b). *MySAP ERP business maps*. Retrieved July 2, 2006, from http://www50.sap.com/solutions/businessmaps/

Scheer, A. -W. (1994). *Business process engineering—Reference models for industrial companies* (2nd ed.). Berlin: Springer.

Scheer, A. -W., & Habermann, F. (2000). Making ERP a success. *Communications of the ACM, 43*(5), 57-61.

Seacord, R. C., Plakosh, D., & Lewis, G. A. (2003). *Modernizing Legacy Systems: Software technologies, engineering processes, and business practices*. Reading, MA: Addison-Wesley.

Wight, O. W. (1984). *Manufacturing resource planning: MRP II. Unlocking America's productivity potential* (revised ed. 1984). New York: John Wiley & Sons.

KEY TERMS

Data Integration: Unifying data models and databases so that all departments of an enterprise use the same data entities, with the same values.

Enterprise Resource Planning (ERP): The current state-of-the-art approach to running organizations with the help of information systems that provide support for key business processes and, more general, for electronic business (e-business).

ERP System (Enterprise Resource Planning System): A comprehensive information system that collects, processes, and provides information about all parts of an enterprise, automating business processes and business rules within and across business functions, partly or completely.

Operations Integration: Creating logical connections of individual operations, or steps of a business process, with preceding or succeeding operations, respectively.

Process Integration: Defining and automating interfaces between different business processes explicitly.

SAP: Software company based in Walldorf (Germany); market leader in ERP software worldwide. SAP is an abbreviation of the company's German name "Systeme, Anwendungen, Produkte in der Datenverarbeitung" (systems, applications, products in data processing).

Software Integration: Connecting different programs so that they can run together and use each other's data and operations.

Entrepreneurship in the Internet

Christian Serarols-Tarrés
Universitat Autònoma de Barcelona, Spain

INTRODUCTION

The increasing development of information technologies (IT) has significantly affected both firms and markets. IT is currently changing the world in a more permanent and far-reaching way than any other technology in the history of mankind (Carrier, Raymond, & Eltaief, 2004). A new economy, where knowledge is the most important strategic resource, is forcing firms to review their traditional routines and take advantage of the tools able to create new value.

Nowadays, there are two types of firms using this new IT. On the one hand, firms with physical presence (traditional companies) use the Internet as a new distribution channel or alternatively as a logical extension of their traditional business. On the other hand, there are dotcoms, Internet start-ups, or *cybertraders* (European Commission, 1997), which have been specifically conceived to operate in this new environment.

A number of scholars have attempted to explain the creation of new ventures from many different theoretical perspectives (economics, psychology, and population ecology among others) and have also offered frameworks for exploring the characteristics of the creation process (Bhave, 1994; Carter, Gartner, & Reynolds, 1996; Gartner, 1985; Shook, Priem, & McGee, 2003; Veciana, 1988; Vesper, 1990; Webster, 1976). However, despite the growing literature in this area, few studies have explored the process of venture creation in dotcom firms.

Cyberentrepreneurship is still in its emergent phase, and there is more to know about the phenomenon and the elements of the venture creation process (Carrier et al., 2004; Jiwa, Lavelle, & Rose, 2004; Martin & Wright, 2005). What are the stages they follow to create their firms? This article attempts to answer this question. First, we analyse the entrepreneurial process of a new firm's creation. Second, we shed some light on how this process is applied by cyberentrepreneurs in starting their businesses based on an in-depth, multiple case study of eight entrepreneurs in Spain.

BACKGROUND

Process of New Venture Creation

A framework for describing new venture creation integrates four major perspectives in entrepreneurship, for Gartner (1985): (1) the individuals involved in the creation of the new venture, (2) the activities undertaken by those individuals during the creation process, (3) the organisational structure and strategy of the new venture, and (4) the environment surrounding the new venture.

According to Gatewood, Shaver, and Gartner (1995) the *venture creation process* is defined as "the process that takes place between the intention to start a business and making the first sale"(p. 374). Much of the research on process of venture creation has assumed a linear, unitary process, composed of a set of activities, beginning with the recognition of a business opportunity and culminating with the first sale (Galbraith, 1982; Kazanjain & Drazin, 1990; Liao, Welsch, & Tan, 2005; Shane & Venkataraman, 2000). However, other authors have included activities occurring after the founding of the venture or its first sales (Bhave, 1994; Shook et al., 2003; Veciana, 2005; among others). For example, Veciana (1988, 2005) includes a consolidation stage where the entrepreneur squeezes out undesirable partners to establish his leadership and guarantee the survival of the firm. Another typical approach to study the process of venture creation is to examine the activities, key milestones, the frequency, and time of those activities (Carter et al., 1996; Gatewood et al., 1995; Kaulio, 2003). Empirical explorations (Hansen & Bird, 1997; Reynolds & Miller, 1992) have found that no one pattern or sequence of events is common to all emerging organisations. Despite this evidence, a recent exploratory study on the entrepreneurial process of creating a firm on the Internet (Carrier et al., 2004) has revealed that the cyberentrepreneurs had gone through basically the same stages, though they belonged to different industries.

According to Baker, Miner, and Eesley (2003) there are two approaches when studying the founding process in entrepreneurship research. First, a design-then-execution framework that assumes a mainly linear process in which start-up intentions and gestation typically lead to the creation of a plan. We refer to this model as *design-precedes-execution* (DPE). In contrast to this model they describe an improvisation framework, where design and execution of the start-up converge. In this case, founders may plunge into the start-up process, designing the firm as they create it.

Figure 1. Comparing the core stages in the start-up process

Pre-venture stage	Organisation stage	Financial jeopardy	Introduction of the product	Squeezing out partners	Outcome stage
Venture idea	Set up operations	Prototypes and channels established	Produce the product	Gain control by the entrepreneur	Survival

Webster (1976)

Proof of principle	Prototype	Model-shop	Start-up
Create a solution: configuration of the business idea	Refine developed technology	Produce and test a number of models	Produce the product (volume production)
Develop nascent proprietary technology	Produce the first prototypes		First sales

Galbraith (1982)

Gestation	Creation	Launching	Consolidation
Childhood	Look for a business opportunity	Create a team	Survival
Antecedents and professional knowledge	Create a solution: configuration of the business idea	Obtain and organise the means	Squeezing out partners
Incubator	Evaluate this opportunity	Develop the product/service	All under control
Precipitation condition	Write the business plan	Find out financial aid	
Decision of creating a new venture	Formal/legal constitution of the firm	Launch the product/service	

Veciana (1988, 2005)

Opportunity recognition	Technical set-up and organisation creation stage	Exchange stage
Business idea	Garner resources	Market feedback

Bhave (1994)

Entrepreneurial intent	Opportunity search	Decision to exploit	Explotation
Individual intent to create a venture based on perceptions of feasibility and desirability, and propensity to act upon opportunities	Search the business opportunity based on individual's alertness to new opportunities and past experiences	Decision to exploit based on risk propensity, motives and attitudes	Finding the resources, planning, networking, selling

Shook et al. (2003)

Business idea	Market needs	Identification business opportunity	Feasibility	Search for support	Venture creation
The initial vision or idea is generated	Determine the needs of different potential customers for the products	Identify opportunities, propose innovative solutions to market needs	Develop prototypes, write a business plan, or find contracts	Gather all the needed resources	Formal/legal constitution of the firm and first sales

Carrier et al. (2004)

Models of Venture Creation, a Sequence of Events

A number of scholars have offered numerous stage models of venture creation process. This approach implies that an additive combination of events will lead to the creation of a new firm. Yet there is little empirical evidence that either validate or fail to validate the linear model (Liao et al., 2005). An in-depth revision of the different stage models in the literature review reveals some similarities that should be noticed (see Figure 1). We can summarise all different stages in the aforementioned *stage models* in just three main steps:

1. *Concept (or gestation) stage* is where the idea of the new venture is set up. This stage is influenced by entrepreneur's background, and it culminates with the identification of a business opportunity. The precipitation condition, the business opportunity, and the incubator play a principal role in this stage.
2. *Planning stage* is where a sequence of steps is essential to coordinate in detail what resources will be needed to produce a product or to offer a service. Once the entrepreneur has detected the business opportunity and has refined it, the next step is to plan how he/she is going to carry on with the idea. This is the stage where the entrepreneur has to transfer what he/she has "in mind" to a plan. The business plan plays a major role in this stage.
3. *Implementation stage* is when the business begins to run. In this stage, all the plans are ready, and the firm begins to operate. Whatever has been planned is now being put into practice. The team is created, resources are gathered, financial aid is obtained, and the product or service is launched.

It is important to notice that these models vary from one author to another, stressing the step they consider most important. However, there are common features that make one able to find a core process. In fact, these stages are not quite well differentiated, with one step usually ending before the other begins.

Table 1. Technical specifications of the study

Sample segmentation criteria
1. Proportion of sales turned over on Internet > 95%.
2. Age of operation: > 3 years to ensure that they had all moved through basic start-up phases.
3. Belongs to already-existing business group: subsidiaries of already-existing groups were eliminated.
4. Main activity of firm: attempt to include widest range of activities, at most one firm per activity.

DISCUSSION

Via multiple case studies, it was hoped to provide an in-depth exploration of each cyberentrepreneur and give rich insights into the entrepreneurial process in such firms. Four sample segmentation criteria were used (see Table 1), and it was done a Web site analysis of the preselected firms.

The cyberentrepreneurs in the sample were asked to provide a detailed history of the process of venture creation. This process involved the activities from the time they first had the intention to start a business until they finally launched the firm and had the first sales. The information was coded, analysed, and the process was reconstructed.

Gestation Stage

It was interesting that most of the entrepreneurs had in mind the idea to create a firm but they had not detected an opportunity first. Yet, they actively began to search for a business opportunity when the Internet became widespread for commercial purposes in Spain in the late 1990s. They all realised that this new distribution channel had many advantages in comparison to others, they would not need big investments nor high operation costs to create and run a venture. Another factor that helped them to detect an opportunity was the fact that they were all aware of the potential offered by the Internet by either working in the field of IT, or doing research on the topic.

The approaches for the opportunity detection process took different forms in all cases. First, there is a wide group of cyberentrepreneurs that tried to combine their hobbies with the potential of the Internet. For example, A.com detected a gap in the Spanish digital photography retailer's market on the Internet. This type of photography was emerging, thus there was the need for a Web site that could aggregate the offer and the demand for these products. The owner of H.com was really interested in the cultural industry, especially in theatre and concerts. He had a broad network of friends working in this industry and he detected a gap in the market for selling last-minute seats at low cost.

Second, there are a couple of cyberentrepreneurs that were struck by a signal from the market, namely the need to facilitate the access to information. This eventually evolved in the possibility of developing software that would, automatically and in real time, help to match the offer and demand for a certain type of content. The owner of G.com invented the first software platform for content aggregation purposes.

Planning Stage

We identify a set of firms that were not founded through a process involving design or planning followed by execution. In these founding processes, the design of the new business occurred at the same time that the first implementation steps were being taken. We have labelled these founding processes as *improvisational*. For example, the owner of B.com did not conduct any attempt to design his business before beginning its execution, neither formally nor informally, nor did he systematically survey the potential market for his products. Instead, he began by legally constituting his firm and launched his Web site, intending to test it with real customers.

Secondly, there is a group of cyberentrepreneurs that spent a lot of time identifying market needs and the industrial structure of the product/service they were planning to sell. Their approaches to clearly define these business opportunities took different forms in all cases: market surveys, brainstorming, and order studies to marketing research companies, research with universities, and so forth. For example, the owner of C.com extensively researched the off-line market for holiday tourist rentals because there was hardly anything on the Internet at that time. He surveyed many of the existing Catalan tourist housing renting agencies offering them to take part in his booking platform. He was able to collect information about margins, tour operators, distribution channels, and so forth. He also went to the tourist faculty at a university to gather information about the customers and the sector. He ended up with a very good notion of the market needs and the structure of the industry.

Implementation Stage

In this stage, three main behaviours are identified. First, there is a group of entrepreneurs that essentially provided all the resources by themselves, both financial and labour. The owner of F.com went straight into the business with his own resources and only after 3 years of operation, he detected the need to write a business plan. In his opinion, the creation of the firm was a sort of prototype to test its success in the market.

Another group of entrepreneurs decided to involve other people or other entrepreneurs in their projects when they searched for support. This venture creation process could be labelled as *team founding*. C.com had to find partners to fund the business, and to design and execute his technological platform. He gave them some shares of the company in exchange for their contributions. The owner of D.com also had to involve some key partners from the organisation he was currently working for Centro Superior de Investigaciones Científicas (CSIC). This case is somehow unique in comparison to the rest, because D.com had been using facilities and knowledge from CSIC to develop his products. All this, and the fact that CSIC is a public research centre, provoked the involvement of CSIC in the project. Thereby D.com can be considered as a spin-off from the CSIC.

Finally, although E.com can not be considered as a team founding because the firm was already created, the owner also had to enter into a partnership agreement with a technological firm to design and execute his e-commerce portal. In exchange of this portal the owner of E.com gave 35% of the shares to this new partner.

Redefinition Stage

Based on the analysis of the data from the selected case studies, a new phase, called *redefinition stage*, has been detected. In this stage, the firm is already operating, sales are being conducted, and a lot of effort is put on receiving feedback from the market. This phase begins with the first sales and culminates with the redefinition of business opportunity and the adjustments that need to be done to the organisation for its survival. While reviewing stage models literature, we hardly find a redefinition stage. Serarols, Del Aguila, and Padilla (forthcoming) stress the importance of a redefinition stage in digital firms. Based on case studies, they observed that those entrepreneurs who had not redefined their businesses rapidly had many difficulties in surviving and some even died. The cyberentrepreneurs questioned for this study present different behaviours in this stage.

First, there is a group of cyberentrepreneurs that had to redefine their business opportunity because their current products/services did not match customer's needs or customers were not willing to pay for them. For example, D.com had to deal with the conceived idea that Spanish software is lesser quality than foreign software. Moreover, D.com was developing technology that was too advanced for the current Spanish market. Therefore, they had to put off their initial idea of focusing on developing advanced artificial intelligence (AI) software in favour of providing technological consultant services to generate revenues. The owner of G.com realised that advertisers were not investing their money in the Internet after the widely publicised failure of new e-businesses in late 2000/early 2001 so he had to adapt his initial idea of providing a customised e-paper to end users. The founding team of G.com adapted their developed technology and offered online press-clipping, professional services to businesses. Yet, this change enabled the firm to survive. Both enterprises were able to successfully adapt their business models and obtained financial aid from venture capital which led them to grow in sales and size.

A second group of two cyberentrepreneurs has achieved a considerable success, but as the Spanish market was not as big as to generate a critical demand for their products they had to diversify their products to grow. For H.com, selling low-fare tickets for concerts and the theatre was a way to penetrate the market. Soon after their first sales, they introduced

new complementary services based on customer's needs. Those services were budget flights, restaurants vouchers, hotel booking services, and last minute holiday packages. Three years after its foundation, H.com was selling over 10 million euros per year.

Appendix A presents a new venture creation model that reflects how each phase was applied by the cyberentrepreneurs studied. This model does not pretend to be a linear, unitary process because our analysis suggests at least two different patterns approaching new venture creation process in cyberentrepreneurships. First, there is a group of cyberentrepreneurs (B.com, E.com, and F.com) that have followed the following sequence: concept—implementation—planning/redefinition. Another group (A.com, D.com, G.com, and H.com) have followed this pattern: concept—planning—implementation—redefinition.

FUTURE TRENDS

This study has also generated several future research opportunities. First, future research should examine the conditions in which potential entrepreneurs discover opportunities without an active search. The Austrian School considers search unnecessary while some scholars view opportunity search as a natural activity of entrepreneurs (Shane, 2000). Our results suggest that some of the cyberentrepreneurs have created their firms without an active search for an opportunity while others have extensively searched for an opportunity. More work needs to be done on the exploration process that precedes and generates formal recognition of an Internet-based business opportunity. This type of information will certainly be of interest to researchers working on business idea exploration, and to entrepreneurship trainers.

Second, focus may be directed to the nature of start-ups and how it affects the venture creation process. For example, are there real differences in the process of start-ups according to the type of firm being created? It can also be useful to investigate how the venture creation process differs across technology-based and nontechnology-based firms, as well as to detect whether industry characteristics influence them.

Finally, future research should aim to investigate whether the environment could have influenced the redefinition stage. We have to bear in mind that many of these entrepreneurs lived the failure of new e-businesses in late 2000/2001.

CONCLUSION

The results of this research throw some interesting light on a new entrepreneurial form (cyberentrepreneurship) that is likely to become much more widespread with the advent of the new economy. The most important contribution of the paper certainly lies in its observation of the cyberentrepreneurship and the process that guides the creation of dotcom firms and its proposed model incorporating the practices observed.

The process of starting a new venture can be studied stressing a sequence of different steps to follow, or without a structured step-by-step sequence focussing on the key milestones that influence the process. When studying the process as a sequence of events, some similarities come up among different authors. Those similarities help us classify those stages into just three: (1) concept, (2) planning, and (3) implementation. Although a clear sequence of events does not exist one after the other.

While analysing the process of new venture creation in cyberspace, an additional stage is detected: redefinition. This stage begins with the first sales and culminates with the redefinition of business opportunity and the adjustments that need to be done to the organisation for its survival.

Overall, our results suggest that venture creation processes are exceedingly more complex and fluid than we have presumed. Venture creation is more than an orderly, unitary, and progressive path that consists of an accumulating series of events. Although the cyberentrepreneurs appear to have gone through similar stages, many different sequences might be observed. In the proposed model, at least two different patterns approaching new venture creation process can be identified.

REFERENCES

Baker, T., Miner, A. S., & Eesley, D. T. (2003). Improvising firms: Bricolage, account giving and improvisational competencies in the founding process. *Research Policy, 32,* 255-276.

Bhave, M. P. (1994). A process model of entrepreneurial venture creation. *Journal of Business Venturing, 9*(3), 223-242.

Carrier, C., Raymond, L., & Eltaief, A. (2004). Cyberentrepreneurship: A multiple case study. *International Journal of Entrepreneurial Behaviour & Research, 10*(5), 349-363.

Carter, N. M., Gartner, W. B., & Reynolds, P. D. (1996). Exploring start-up event sequences. *Journal of Business Venturing, 11*(3), 151-166.

European Commission. (1997). *A European initiative in electronic commerce*. Retrieved October 5, 1999, from http://cordis.europa.eu/esprit/src/ecomcom.htm

Galbraith, J. (1982). The stage of growth. *Journal of Business Strategy, 3*(1), 70-79.

Gartner, W. B. (1985). A framework for describing the phenomenon of new venture creation. *Academy of Management Review, 10*(4), 696-706.

Gatewood, E. J., Shaver, K. G., & Gartner, W. B. (1995). A longitudinal study of cognitive factors influencing start-up behaviours and success at venture creation. *Journal of Business Venturing, 10*(5), 371-391.

Hansen, E. L., & Bird, B. J. (1997). The stages model on high-tech venture founding: Tried but true? *Entrepreneurship Theory and Practice, 22*(4), 111-122.

Jiwa, S., Lavelle, D., & Rose, A. (2004). Netrepreneur simulation: Enterprise creation for the online economy. *International Journal of Retail & Distribution Management, 32*(12), 137-150.

Kaulio, M. A. (2003). Initial conditions or process of development? Critical incidents in the early stages of new ventures. *R&D Management, 33*(2), 165-175.

Kazanjain, R., & Drazin, R. (1990). A stage contingent model of design and growth for technology based ventures. *Journal of Business Venturing, 5*(3), 137-150.

Liao, J., Welsch, H., & Tan, W. L. (2005). Venture gestation paths of nascent entrepreneurs: Exploring the temporal patterns. *Journal of High Technology Management Research, 16,* 1-22.

Martin, L. M., & Wright, L. T. (2005). No gender in cyberspace? *International Journal of Entrepreneurial Behaviour & Research, 11*(2), 162-178.

Reynolds, P., & Miller, B. (1992). New firm gestation: Conception, birth, and implications for research. *Journal of Business Venturing, 7*(5), 405-417.

Serarols, C., Del Aguila, A. R., & Padilla, A. (forthcoming). Exploring the socio-demographic characteristics of the e-entrepreneur. An empirical study on Spanish ventures. In F. Therin (Ed.) *Handbook of Research on Technoentrepreneurship.* Edward Elgar Publishing.

Shane, S. (2000). Prior knowledge and the discovery of entrepreneurial opportunities. *Organization Science, 11*(4), 448-469.

Shane, S., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review, 25*(1), 217-226.

Shook, C. L., Priem, R. L., & McGee, J. E. (2003). Venture creation and the enterprising individual: A review and synthesis. *Journal of Management, 29*(3), 379-399.

Veciana, J. M. (1988). Empresario y proceso de creación de empresas. *Revista Econòmica de Catalunya*, 8.

Veciana, J. M. (2005). La creació d'empreses. Un enfocament gerencial. Col·lecció d'estudis econòmics, nº 3, Servei d'estudis "La Caixa", Barcelona (1ª Edició).

Vesper, K. H. (1990). *New venture strategies (*Rev. ed.*)*. Englewood Cliffs, NJ: Prentice Hall.

Webster, F. (1976). A model for new venture initiation. *Academy of Management Review, I* (1), 26-37.

KEY TERMS

Cyberentrepreneurs: The individual that creates a firm that is essentially founded upon e-commerce, and whose main activities are based on the exploiting networks using Internet, intranets, and extranets.

Cybertraders (dot-coms): These are companies without significant physical presence, which exclusively operate on the Internet. Using IT, they specialise in a market niche or in cost reduction. Most of them are small companies, and they are only known because of their Webs, though others (Amazon, Cdnow, etc.) have turned into companies of recognized prestige.

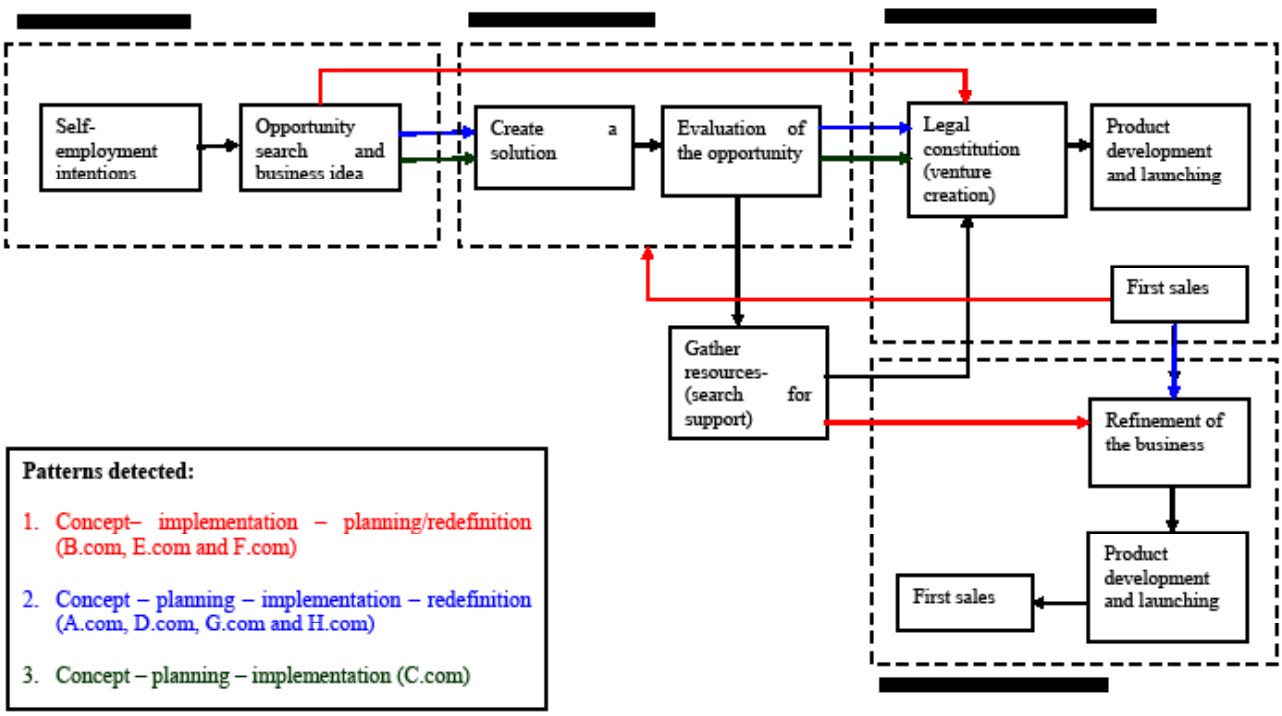
Redefinition Stage: This stage begins with the first sales and culminates with the redefinition of business opportunity and the adjustments that need to be done to the organisation for its survival.

Stage Model: A linear sequence or steps involved in the starting-up through the establishment of a chain of events, cause-and-effect phenomenon that should be followed to create the business.

Web Aggregator: Web aggregator is an entity that can transparently collect and analyse information from multiple Web data sources.

Venture Creation Process: The process that takes place between the intention to start a business and making the first sale.

APPENDIX A.



Patterns detected:

1. Concept– implementation – planning/redefinition (B.com, E.com and F.com)
2. Concept – planning – implementation – redefinition (A.com, D.com, G.com and H.com)
3. Concept – planning – implementation (C.com)

Category: Electronic Business

Envisaging Business Integration in the Insurance Sector

Silvina Santana
Universidade de Aveiro, Portugal

Vítor Amorim
I2S Informática-Sistemas e Serviços, Portugal

INTRODUCTION

Data, information and knowledge are the heart of the insurance business. Each policy is composed of a set of data that can vary substantially. Risk management is a complex process that implies the availability of rich and accurate information and knowledge. In our fast moving world, connectivity and articulation between insurance industry players is therefore mandatory. Information and communication systems and technology (ICST) can provide this connectivity, allowing insurance partners to become closer and able to reach better negotiation, reducing response time and costs and probably creating new business opportunities (Strazewski, 2001).

Insurance intermediaries (brokers and agents) are important players in this scenario. They act as consultants operating independently from insurance companies, being specialists in providing services to their clients, gathering the best solutions thanks to their vast knowledge of insurance companies' products. Consequently, they achieve the best insurance contracts at the least cost (APROSE, 2005a, 2005b).

Being a great value-adding activity, insurance mediation is also very complex. To operate in an effective and efficient way, intermediaries need to establish a good connection with all entities in the industry and electronic business can help insurance intermediaries' business model in both business-to-business (B2B) and business-to-consumer (B2C) dimensions. In B2B, intermediaries establish relations with insurance companies, agents, banks and official entities. In B2C, intermediaries establish relations with their clients, giving them all the necessary assistance in a customized and fast way, since the first contact and during the policy's whole life cycle, offering the best solutions according to their needs.

However, in spite of all the apparent and potential benefits, intermediaries are not grasping all the advantages that electronic business can provide. This definitely relates to a very important issue, the integration level between the different players' information systems.

Analysing the situation from the intermediary perspective, this article exposes the problems faced by intermediaries and insurance companies all over the world when trying to integrate their business electronically and how these can be overcome so that partners can fully benefit from the opportunities here identified. The methodology used includes a deep case study involving a Portuguese intermediary having a significant level of integration with an insurance company. Results are compared with situations reported in other countries, leading to the conclusion that most of the problems and barriers here identified are being experienced worldwide. Conclusions bring significant implications for information science and technology (IS&T) and add important contribution and knowledge to research in this area.

BACKGROUND

This case study involves three entities: an insurance intermediary, an insurance company and a software house specialized in the development of technological solutions for the insurance industry.

A deep study of the software house clients highlighted the case of an insurance intermediary having a significant level of integration with an insurance company. Interviews with representatives from the different companies have shown that in spite of strong and continued technological and financial efforts the parties were still far from having reached a satisfactory solution. Legal, technical and organizational issues seem to be burdening the integration process, maintaining an inefficient status quo and preventing the parties from grasping the desired benefits. It was then decided to deepen the analysis in order to foster knowledge in the area and assist in the development of proper solutions.

The intermediary in our case study uses GIS® Agents & Brokers (developed by I2S Informática-Sistemas e Serviços SA: www.i2s.pt) for managing its own business, but also to transfer data to/from the insurance company Web site. Any break in the electronic data flow, internally or between parties, adds additional costs at several levels. Information systems must support the integration of processes across the extended value chain.

The insurance company in this study has an Internet public site which allows the intermediary and the general public to make online proposals and product simulations and knowing available products. A username and a password allow partners to access a restricted area where they can find, for example, product manuals and rates, proposal forms and products' general conditions. In the same area, the same username and password give access to a new restricted area of the extranet named Broker Information System, where it is possible, among other operations, to view online several brokers' portfolio data (clients, policies and claims) and to access a file transfer area.

This file transfer area allows the intermediary to receive/send data from/to the insurance company on a periodic basis (might be daily). The application GIS® Agents & Brokers includes a set of data transfer modules that, according to the insurance company format, can integrate data from the files received into the GIS® Agents & Brokers database and generate new ones from this database.

Besides communicating with the insurance company, GIS® Agents & Brokers can also import/export files from/to banks and export data to agents. When the intermediary needs to integrate data with some defined clients, it is possible to develop a customized application. In spite of sending/receiving data to/from the ISP (Instituto de Seguros de Portugal (Portuguese Insurance Institute): www.isp.pt), a regulatory body, there is no data exchange that interacts directly with GIS® Agents & Brokers.

The study revealed the existence of an additional integration module used exclusively between the insurance company and some of its brokers using GIS® Agents & Brokers. Data concern claims and only flow from the insurance company to the intermediary, without integrating with GIS® Agents & Brokers. The claims module processes data sent by the insurance company and allows data browsing by means of an Internet browser.

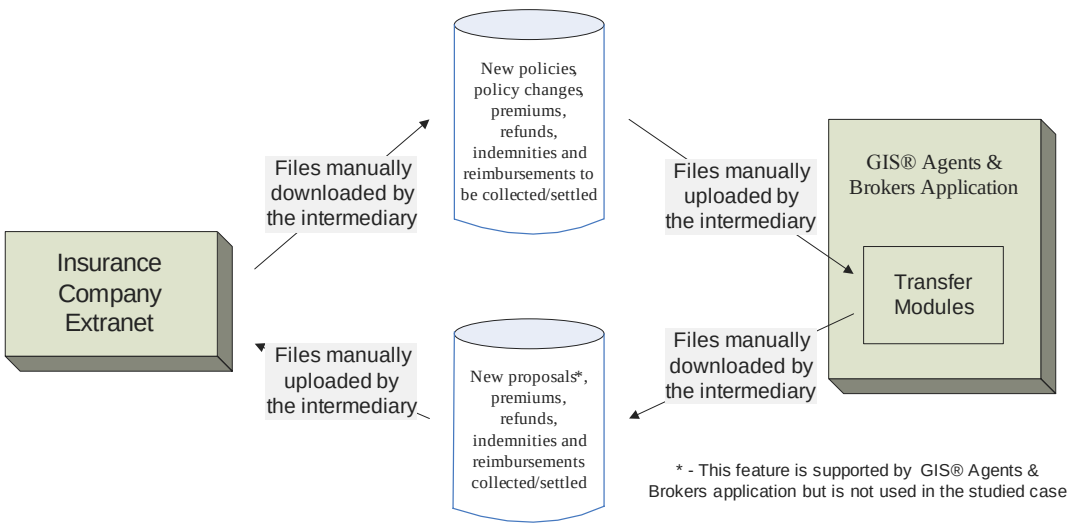
For detailed information and technical aspects, impossible to address here due to space restrictions, please refer to Amorim and Santana (2007) where you may find diagrams representing processes of interaction between the several entities and the electronic business platforms; data flow to/from the several transfer modules which integrate the application GIS® Agents & Brokers; and business terminology.

ELECTRONIC DATA EXCHANGE AND INTEGRATION VIA THE INTERMEDIARY APPLICATION

GIS® Agents & Brokers includes a set of modules which allow data to be integrated to and extracted from the intermediary information system. As depicted in Figure 1, by means of file manipulation each transfer module integrates data supplied by the several entities or extracts data to be supplied to them. The exchanged files are composed of plain text formatted in columns. There is no use for XML. Initially, files were supplied to/by the several modules in floppy disks. Nowadays, there are alternatives, namely the use of e-mail or data download/upload from/to a private extranet area.

The several modules correspond to mediation process data needs and their functioning can be summarized as follows:

Figure 1. Data exchange supported by GIS® Agents & Brokers between an insurance company and an intermediary



- **Extraction of new proposals:** If the intermediary enters new proposals in the system, a module will extract data to be supplied to the insurance company system. In the case under analysis, this module is not being used because the insurance company system cannot store the proposal number sent by the intermediary system or because the intermediary sends proposals that do not match the validation rules of the insurance company. Because frequently partners do not have the same calculation and validation rules, the insurance company may have to adjust some values so that the proposal is in conformity with the products being sold.
- **Integration of new policies:** After analyzing and accepting proposals, the insurance company issues the corresponding policies. Policies' data would be integrated in the intermediary system by a specific module, but in this situation such is not possible. As the insurance company system cannot integrate the proposal number produced by GIS® Agents & Brokers, the proposal and the policy numbers are different and the policy data supplied to the intermediary system cannot be integrated straightforwardly. Data is received by a different module that organizes it in order to be manually checked and integrated. Furthermore, it may happen that some changes have been made by the insurance company to some of the data fields provided in the proposal (such as to the number of characters allowed) and that these changes are not according to the intermediary rules. In such cases, the intermediary may opt to ignore the changes and preserving the data already available in its database. To establish the necessary correspondence, there is a module with a functionality named matching that allows, for each policy of the insurance company, to select the intermediary proposals that might correspond to that policy based on a data set (e.g., the policyholder name, the insurance start date or the sum insured). Manually, the intermediary matches the policy to the corresponding proposal. If proposals have not yet been entered in the intermediary system, he or she can directly accept the data to be integrated. However, if the intermediary does not enter the proposals in the system, then he or she will have to manage them manually (in paper, for example) and there is one situation where this can be quite complicated and inefficient: when the intermediary works with several sellers/agents. Without registering commercial data, the intermediary cannot properly control the production of each seller or agent. Besides that, when the proposal is sent to the insurance company, there is no indication of who has produced it. There is only indication that the proposal has been originated in this intermediary. As there are no proposals registered in the intermediary's system, after policy integration the intermediary has to manually introduce data related to the seller/agent that had produced each proposal. With proposals entered directly in the system, the intermediary does not have to care with this aspect, as commercial data has already been properly registered.
- **Integration of policy changes:** When there are changes to policies (e.g., changes to the address or to the sum insured), the insurance company system supplies the corresponding data to be integrated.
- **Integration of transactions:** Data concerning transactions (premiums, refunds, indemnities and reimbursements) to be collected/settled by the intermediary are supplied by the insurance company. This way, the intermediary can address the policyholders in order to receive/pay what is necessary. If the policyholders go to the insurance company branches instead of addressing the intermediary to pay/receive what is necessary, the insurance company system supplies the resulting data to the intermediary to be integrated. This avoids double collection or payment by the intermediary.
- **Extraction of transactions:** Besides other elements, there is a module that extracts data from transactions collected/settled by the intermediary to be supplied to the insurance company.

When a claim occurs, the insurance company system supplies the corresponding data to be processed by the intermediary claims module so that the intermediary can follow its development. This module is specific to the insurance company studied and does not integrate with GIS® Agents & Brokers.

In some particular situations, intermediary clients may need to supply or receive, in electronic format, a large volume of data. These cases can lead to the development of customized transfer modules.

If the intermediary collects premiums by bank debit, he or she can use another module which allows bank file extraction and integration. If the intermediary works with agents, there is a module that extracts data concerning premiums, refunds, indemnities and reimbursements that agents can collect/settle. If the agents have GIS® Agents & Brokers, then, with another module, they can integrate the data supplied by the intermediary module; otherwise, the agent system must have a functionality that can integrate the data received.

In this case, the intermediary does not exchange electronic data with banks, agents or the ISP. The intermediary also does not use the extranet (GIS® Web) which is part of GIS® Agents & Brokers, and that would allow an integrated view of the contract by the intermediary clients and agents. The product is not being used because the intermediary did not know about its existence.

Every module receiving data from the insurance company has functionalities which allow testing data quality before integrating it in the intermediary application. It is possible to simulate data integration to avoid abnormal data entry. The integration process has the same validation features as the simulation process. In both simulation and integration, the modules produce error printings and indicate what data would be/were loaded successfully. The simulation stage is optional and must be manually selected by the intermediary.

Intermediaries also have to download manually from the insurance company Web site the files that are to be loaded into their systems by the integration modules and have to upload manually the files that were produced by their systems and are to be integrated into the insurance company system.

Several barriers to the adhesion to electronic business have been identified:

- the intermediary and the insurance company have their own visions of business, organizing their own products in a particular and perhaps unique way;
- the integration of data between systems becomes more complex, time-consuming and prone to mistakes due to the lack of normalization, forcing data conversion so that the parties can dialog;
- lapses in data handling, the lack of validation rules and delays in data extraction decrease data quality and, in the limit, can make electronic integration unfeasible;
- parties implement only business facets presenting the best cost/benefit relation;
- some of the processes require manual intervention of the intermediary, originating breaks in the data flow; and
- paper is used in varied circumstances of this business, be it for legal impositions, or due to data validation/checking procedures, hampering complete dematerialization of business.

However, there are many advantages in the adhesion to electronic business in this industry, clearly assumed by the parties involved. These can be found in the following aspects:

- the intermediary becomes much more autonomous from the insurance company;
- the client/agent becomes much more autonomous from the intermediary;
- all benefit from the possibility of using up-to-date forms;
- manual data entry is minimized, with consequent reduction of human mistakes;
- the automatic processing of huge volumes of data in reduced time becomes possible;
- normally, cost reduction is attained;
- partners will be able to provide a better service to the final client.

FUTURE TRENDS

In the insurance industry, the key to electronic business evolution at electronic data integration level consists in the development of a normalized structure where the insurance companies' products can be published. Besides defining data representations, such a structure must also define the rules that relate and validate them. Data could then flow between systems because the rules that support them are the same, avoiding situations where valid data for one entity is not valid to another.

For some time now, the insurance industry has used several Service-oriented Architecture (SOA) approaches to enable electronic business. SOA "is a technical framework that allows faster, lower cost sharing of data and processing power across a heterogeneous IT infrastructure" and "is platform-neutral and uses standard data types, structured text-based messages, and open transport protocols" (Celent & Sun, 2006, p. 3). These text messages are typically written in Extensible Markup Language (XML) and SOA can be used both internally, over an internal TCP/IP network, or externally, over the Internet, in which case it is often referred to as a "Web service" (Celent & Sun, 2006, p. 3).

Web services technology has been advertised as an industry-transforming means of enabling communications among insurers, their business partners and end-customers (O'Donnel, 2004, p. 1). Even if security, performance and standards issues have led to cautious adoption of Web services, the technology brought many benefits for insurers working to integrate applications within the enterprise.

However, in spite of the effort put in the definition of standards, several national standards still prevail (Chesher & Kaura, 1998; CEN, 2003). Not even the problems raised in the past due to the existence of different standards for Electronic Data Interchange (EDI) have prevented the most recent developments at technological level from originating a number of standards that, directly or indirectly, compete with each other. Taking XML as an example, there are several standards competing to ebXML which is a standard developed by the Organization for the Advancement of Structured Information Standards (OASIS) and the United Nations Centre for Trade Facilitation and Electronic Business (UN/CEFACT).

Many efforts are being made to overcome the situation. In 1983, 21 insurance companies in the United States founded an organization named Insurance Value Added Network Services (IVANS) that operates as a communications network for the insurance industry aiming at using technology to increase sales and improve support to clients (IVANS, 2007a).

In 2000, in partnership with a software house from the insurance area, IVANS began the development of a solution for real-time communications between insurance companies and intermediaries named Transformation Station™ (Applied Systems, 2005). This solution is based on the standards defined by the Association for Cooperative Operations Research and Development (ACORD) and supplies, through the Internet, a Web Services communication infrastructure between intermediaries and insurance companies. Intermediaries now have access to a functionality named Single-Entry, Multiple-Company Interface (SEMCI), meaning that through a single data entrance point, they are able to interact with the interfaces of multiple insurance companies. For some types of products, they can, in real time, get the prices, view policies' data and even perform some processing, like claim handling (Fitzpatrick, 2005; IVANS, 2004, 2007b; McKenna, 2005).

Currently, the Council of Insurance Agents & Brokers (CIAB), along with several industry partners, is working on a new "groundbreaking electronic insurance exchange that will give brokers and agents the ability to consolidate their account information, enter the data once and send all material—or parts of it—to a variety of insurers" (CIAB, 2007, p. 15). This exchange is expected to be operational in 2008 (CIAB, 2007).

According to Speer (2007, p. 1), "the exchange takes the real-time, comparative rating, single-entry, multiple-company interface idea a step further." Along with providing producers of all sizes with the ability to perform "once and done" transactions with any number of carriers simultaneously, it works as a content aggregator, making exchange participants' data available in a variety of forms to subscribers. Many believe that this new approach will overcome past failures (Arvidson, 2006; Chordas, 2007; Maciag, 2006), and succeed because it is broker-oriented and broker-driven (Arvidson, 2006).

SEMCI has been the Holy Grail for the brokerage industry (Fitzpatrick, 2005) that intermediaries have waited for more than 20 years (Yates, 2004). The results of the 2006 ACORD-User Groups Information Exchange (AUGIE)-Agency Technology Survey that involved more than 7,500 agency, brokerage and wholesaler professionals show how important SEMCI is in the industry (AUGIE 2006). From the respondents, 48.9% answered that "learning and using various company proprietary systems" was their greatest challenge in supporting automation, while 46.6% quoted "duplicate data entry" as their major automation time waster in doing business today. For 70.6% of the respondents, in the 2 previous years, ease of doing business with a carrier has become an increasingly important factor in the decision to place business with that carrier.

However, despite all technological promises, finding systems that work well for all involved has proved difficult (Chordas, 2007, p. 1). There are still too many proprietary insurer forms and Web sites (Maciag, 2006), and examples of failed implementations abound. In 2006, six leading Lloyd's underwriters came together to develop the system Kinnect to transfer data and documents electronically. "With the loss of 70 million [pounds sterling] (about U.S. $137 million), the project disbanded after Lloyd's withdrew its support, saying the approach was no longer suited to the market" (Chordas, 2007, p. 1). The G6 initiative, that joins six Lloyd's insurers (known as managing agents), followed on the heels of Lloyd's unsuccessful electronic communications system. This system uses ACORD data standards (Chordas, 2007).

Overall, the results seem short from expectations. "The reality of most carriers is that critical costumer and product information is spread across disparate systems," "key channels for serving costumers cannot act in a coordinated fashion, and many core processes are still performed manually." It seems that "existing IT architectures are one of the biggest roadblocks to enhanced efficiency within the industry" (Tulloch, 2005, p. 1).

The insurers' ability to react to the industry fast changes is being hindered by the industry's preponderance of aging core legacy systems. Mainframe systems have run the business and undergone massive investment over decades, but it is now clear that they present many challenges for companies, namely increasing difficulty of integration to new applications, inhibiting data mastery and re-use and impeding reengineering and improvement of process and workflows (Itemfield, 2005).

Most large and mid-size insurers are moving forward by extending the use of critical legacy systems, while adopting ACORD industry document and transaction standards. The results from Celent's third annual survey of senior IT executives, conducted in the fall of 2005, reveals that insurers are now "more inclined to believe an acceptable solution is available in the marketplace. Most large firms are still inclined to assemble best-of-breed components, while many smaller companies prefer end-to-end solutions" (Celent, 2005). It is clear that insurers are making selective investments in new applications for the back and front office and specific core functions; however, it seems that integration issues still lag behind.

Maintaining legacy systems while trying to cope with industry transaction standards can be difficult, taking into consideration current solutions. Legacy systems must be prepared to support new front ends, information exchange with a wide range of internal applications and XML for Internet-based applications, such as with business partners or online services. On the other hand, companies adopting ACORD XML standards must decide for the use of an external data exchange service such as IVANS or in-house solutions. In-house solutions need to address a number of problems, namely due to frequent changes and implementation issues with ACORD XML. "While versions are intended to be backwards compatible, real world complexities such as data

elements being renamed or deprecated between versions, or proprietary implementations make use of current fixed parsers for ACORD cumbersome and expensive. Further complicating use of ACORD is the need to detect versions of ACORD XML and determine how to handle data elements that are not consistent across versions, for example, mapping from a higher to a lower version of the standard." Managing data quality issues in these environments becomes "costly and often limits the usefulness of enterprise data" (Itemfield, 2005, p. 2).

Several players are positioning themselves in this promising market, offering SOA solutions. SOA can improve access to legacy systems, integrate multiple sales channels and improve insurers' clients self-service, among other things. However, the industry is still in the early phases of adopting SOA and the internal efforts have led to some frustration. Tulloch (2005), reporting the results of a study, refers that "nearly one-third of companies polled believe service-oriented architectures are not performing as well as expected. Of those, almost a quarter actually believe that SOA has increased the complexity of their IT systems" (p. 2). Tulloch (2005) defends that "this has more to do with the approach taken, and internal skills that organizations have at their disposal, than with the results that can be realized from a well-designed and implemented SOA," but also recognizes that "with the modularity of a component-based approach, functionality to impact business can be delivered next quarter, not next year" (p. 2).

Fortunately, the industry now understands that "the big-bang, all-at-once approach to SOA implementation has been largely ineffective" (Conz, 2007, p. 1). It seems that the insurers that are seeing the most encouraging returns on their SOA efforts are those that took an incremental approach. New endeavours will certainly benefit from best practices regarding implementation and organization. A recent report published by Celent (2007) claims that large insurers have already realized integration efficiencies ranging from 10% to 49% from their investments in SOA.

Insurance business processes are complex and contract data may be very different from one insurance product to another. It is possible to reach a limit where certain contracts are unique, that is, they are completely different from any other, leading to data that is merely descriptive and can only be interpreted by humans. Setting aside these cases, insurance electronic data format normalization will be an important evolution to electronic business because the computer system of each intermediary will only need to know how to import and export data based on a format common to all insurance companies.

Besides that, with sufficient trust in exchanged data, intermediary and insurance company businesses' integration would be superior by discarding the manual validation stage. Data could freely flow from one system to another without human intervention. In addition, the absence of human intervention could lead to frequent or even real time data integration. However, the file exchange process must be more automated.

In the ideal situation, manual data entry would be done in only one system, and data would be sent automatically to the other system to be integrated. A new proposal could be entered in the intermediary system and later be electronically integrated in the insurance company system. Working with only one interface that would be their own systems, the intermediaries could manage data right from the beginning and the insurance company or the intermediaries themselves would not have to replicate the previous manual work. This way, all parties (client, intermediary and insurance company) could grasp the productivity gains attained through single-entry, multiple-company interface.

Alternatively, if the insurance company does not have means to electronically integrate the intermediary's data, then the intermediary could directly enter the data into the insurance company system (because it is the intermediary who has the data) and, automatically, these data would be sent to the intermediary system to be electronically integrated. Any delay in sending or integrating will be harmful to the intermediaries because the data they need to run their business is not in their systems. However, with this alternative, the intermediaries have the inconvenience of having different data entry interfaces according to the several insurance companies they work with, increasing the potential for human error.

In either of the suggested hypothesis data integration must evolve in a way to support an increasing data flow between the intermediary and the insurance company until total business integration is reached.

The first step toward overall business integration is the national adoption, by the insurance industry, of one data exchange standard like the one developed by the e-business Expert Group for the Insurance Industry (eEG7), that is part of the Centre Européen de Normalisation and Information Society Standardization System (CEN/ISSS), or the one developed by ACORD, for example. XML and a Service-oriented Architecture based on Web Services should be used to integrate all insurance players. XML is currently being widely adopted by data exchange standards and Web Services can provide an interface that simplifies integration between several systems.

Yet, the evolution cannot be restricted to the integration of data between intermediary and insurance company. Data integration between intermediary and agents must be substantially improved so that both systems are completely synchronized in all business data without data entry replication. On the other hand, the intermediary application must support electronic data extraction to regulatory bodies in the required formats, avoiding the manual process currently

performed by the intermediary. It is also fundamental to bet heavily on extranets, both in insurance companies and intermediaries.

CONCLUSION

To operate in an effective and efficient way, intermediaries need to achieve great connectivity and articulation with all insurance business entities. Electronic business can provide this connectivity and articulation, bringing the different parties together, reducing response time and costs, serving clients better and creating new business opportunities.

In spite of being relatively advanced in this domain, the intermediary and the insurance company here analysed still operate low dematerialized processes in which paper, manual data entry and validation routines are a daily reality.

This situation is due to different visions of the business, to the lack of normalization and poor data quality, to some specificities of the insurance company information system, to the unfavourable cost/benefit relation, to the existence of manual processes and to the continued use of paper.

However, there are many advantages in the adhesion to electronic business in this industry, clearly assumed by the parties. Manual data entry is minimized, with consequent reduction of human mistakes. The automatic processing of huge volumes of data becomes possible in a short time. The intermediary becomes much more autonomous from the insurance company, the client/agent becomes much more autonomous from the intermediary and all benefit from the possibility of using up-to-date forms. Cost reduction is normally attained and partners become able to provide a better service to the final client.

REFERENCES

Amorim, V., & Santana, S. (2007). Business integration in the insurance sector: The intermediary side of the question. In M. Cunha, B. Cortes, & G. Putnik (Eds.), *Adaptive technologies and business integration: Social, managerial and organizational dimensions* (chap. VI, pp. 118-136). Hershey, PA: Idea Group.

Applied Systems. (2005). *IVANS and applied systems to combine knowledge and technology for industry communication*. Retrieved May 31, 2008, from http://www.appliedsystems.com/products/transformation/News/ts10-12-2000.htm

APROSE. (2005a). *Porque deve contactar um mediador de seguros?* Retrieved May 31, 2008, from http://www.aprose.pt/publico/mediador/porquecontactar.php

APROSE. (2005b). *Qual o mediador adequado?*. Retrieved May 31, 2008, from http://www.aprose.pt/publico/quemediador/mediadoresadequados.php

Arvidson, C. (2006). In search of brokers' holy grail. *Leaders Edge Magazine*, April 2005. Retrieved May 31, 2008, from http://www.ciab.com/TemplateMagazine.cfm?Section=Editorial&MagazineTimeFrame_ICID=135&MagazineYear_ICID=168

AUGIE. (2006). *Agency technology survey—executive summary*. Retrieved May 31, 2008, from http://www.acordadvantage.org/augie/Survey_Exec_Summary.pdf

Celent. (2005). *Insurance CIO/CTO pressures, priorities, projects, and plans in 2006: Survey results*. Retrieved May 31, 2008, from http://www.celent.com/PressReleases/20051214(2)/CIOCTO.htm

Celent. (2007). *Web services and SOA in insurance 2007: Realizing and communicating the business Value*. Retrieved May 31, 2008, from http://www.celent.com/PressReleases/20070320/InsWSSOA.htm

Celent & Sun. (2006). *Driving business value with SOA in insurance—joint white paper with Sun Microsystems*. Retrieved May 31, 008, from http://www.sun.com/solutions/documents/white-papers/fn_soawp.pdf

CEN. (2003). *CEN/ISSS report and recommendations on key e-business standards issues 2003-2005*. Retrieved May 31, 2008, from http://www.cenorm.be/cenorm/businessdomains/businessdomains/isss/activity/reportfinal.pdf

Chesher, M., & Kaura, R. (1998). *Electronic commerce and business communications*. London: Springer-Verlag.

Chordas, L. (2007, May 1). Exchange place: Two global technology initiatives are bringing together agents, brokers and carriers. *Best's Review.* Retrieved May 31, 2008, from http://goliath.ecnext.com/coms2/gi_0199-6684000/Exchange-place-two-global-technology.html

CIAB. (2007). *The council of insurance agents & brokers—2007 year in review.* Retrieved May 31, 2008, from http://www.ciab.com/TemplateRedirect.cfm?template=/ContentManagement/ContentDisplay.cfm&ContentID=7224

Conz, N. (2007, May 18). *SOA adopters discuss best practices. Insurance & Technology*. Retrieved May 31, 2008, from http://www.insurancetech.com/showArticle.jhtml?articleID=199905005

Fitzpatrick, M. (2005, April). Retool time. *Leaders Edge Magazine*. Retrieved May 31, 2008, from http://www.ciab.com/Content/ContentGroups/Leaders_Edge_Magazine2/2005/April/Retool_Time.htm

Itemfield. (2005). *Next generation business integration solutions for insurance*. Retrieved May 31, 2008, from http://www.itemfield.com/pdf/Insurance_Solutions.pdf

IVANS. (2004). *Transformation station white paper*. Cincinnati, OH: IVANS. Retrieved May 31, 2008, from http://www.ivans.com/whitepapers/Tran%20Station%20White%20Paper.pdf

IVANS. (2007a). *About us*. Retrieved May 31, 2008, from http://www.ivans.com/main.asp?secname=ABOUT%20US

IVANS. (2007b). *Transformation station*. Retrieved May 31, 2008, from http://www.ivans.com/main.asp?secname=SOLUTIONS&subname=E-Commerce&offername=Transformation%20Station

Maciag, G. (2006). *The electronic insurance exchange*. Retrieved May 31, 2008, from http://www.acordceo.org/2006/12/the_electronic_.html

McKenna, S. (2005, December 1). SEMCI—not yet ready for prime time. *Insurance Networking News*. Retrieved May 31, 2008, from http://www.insurancenetworkingnews.com/protected/article.cfm?articleId=3710

Speer, P. (2007, March 1). *Brokers cast a wide net with electronic exchange. Insurance networking news: Executive strategies for technology management*. Retrieved May 31, 2008, from http://www.accessmylibrary.com/coms2/summary_0286-29984316_ITM

Strazewski, L. (2001, January). A winning combination. *Rough Notes, 144*(1). Retrieved May 31, 2008, from http://proquest.umi.com/pqdweb?did=67193062&sid=15&Fmt=4&clientId=23852&RQT=309&VName=PQD

Tulloch, S. (2005). *Insurer heal thyself*. Retrieved May 31, 2008, from http://www.ebizq.net/executive_corner/topics/ind_sol/features/6555.html?&pp=1

Yates, J. (2004, August). SEMCI building blocks are in place. *National Policyholder P & C, 108*(30). Retrieved May 31, 2008, from http://proquest.umi.com/pqdweb?did=682379891&sid=13&Fmt=4&clientId=23852&RQT=309&VName=PQD

KEY TERMS

Association for Cooperative Operations Research and Development (ACORD): ACORD (Association for Cooperative Operations Research and Development) is a global, nonprofit insurance association whose mission is to facilitate the development and use of standards for the insurance, reinsurance and related financial services industries.

Electronic Business: The electronic business concept is used to describe business conducted by electronic means, often over the Internet. It may be understood as a combination of business strategies and distributed processes that use technology to manage the electronic transfer of data and information between business partners. It may concern the exchange of structured and unstructured data and information.

eXtensible Markup Language (XML): XML is a set of standards that specify how to structure a text-based document for communication between two computers for any number of purposes. It allows incorporating metadata in the message to be exchanged, so the data may be transmitted and understood by the receiving party.

Insurance E-Business Expert Group (EG7): eEG7 is the European forum for the development of e-business standards for electronic communication in the insurance sector. It is one of the e-business Board for European Standardization (eBES) groups. eEG7 aims at facilitating the transfer of information between policyholders, professional intermediaries (agents, brokers), insurers and other involved parties. Its standards support the placing and administration of insurance contracts, claims handling and accounting.

Insurance Intermediary or Insurance Broker: Insurance brokerage is a remunerated activity whose main goals are to facilitate the settling of insurance contracts and assist them over their lifecycles. An insurance intermediary or insurance broker is a consultant operating in the insurance sector, independently from any insurance company. He specialises in providing services to their clients, and gathering the best solutions thanks to their vast knowledge of insurance companies' products.

Service-Oriented Architecture (SOA): The Service-oriented Architecture (SOA) is an approach to Enterprise Architecture where each major element is presented as a "service." A SOA solution materializes in a distributed computer environment with a high level of interoperability between the existing systems and eased integration of new functionalities. It allows combining and recombining software components, which is expected to reflect on business processes' flexibility.

Single-Entry Multiple Company Interface (SEMCI): Single Entry Multiple Company Interface (SEMCI) is a computer system based on service-oriented architecture (SOA). SEMCI acts as an interface which connects the agent information system to the information systems of multiple insurance companies.

ERP and the Best-of-Breed Alternative

Joseph Bradley
University of Idaho, USA

INTRODUCTION

Enterprise resource planning (ERP) systems are off-the-shelf software systems that claim to meet the information needs of organizations. These systems are usually adopted to replace hard-to-maintain legacy systems developed by IS departments or older off-the-shelf packages that often provided only piecemeal solutions to the organization's information needs. ERP systems evolved in the 1990s from material requirements planning (MRP) systems developed in the 1970s and manufacturing resources planning (MRPII) systems developed in the 1980s. ERP systems serve the entire organization, not just material or manufacturing planning. One advantage of ERP is that it integrates all the information for the entire organization into a single database.

Implementation of ERP systems has proven expensive and time consuming. Failed and abandoned projects have been well publicized in the business press. ERP systems are "expensive and difficult to implement, often imposing their own logic on a company's strategy and existing culture" (Pozzebon, 2000, p. 105).

Most firms utilize a single software vendor for the complete ERP system throughout their organizations. The integrated nature of ERP software favors this single-vendor approach. An alternative strategy adopted by some firms is the best-of-breed approach, where the adopting organization picks and chooses ERP functional modules from the vendor whose software best supports its business processes. Organizations adopting best of breed believe that this approach will create a better fit with existing or required business processes,

Table 1. Some functions available in SAP R/3 (Source: Davenport, 1998)

Financials	**Operations and Logistics**
Accounts receivable and payable	Inventory management
Asset accounting	Material requirements planning
Cash management and forecasting	Plant maintenance
Cost element and cost center accounting	Production planning
Executive information systems	Project management
Financial consolidations	Purchasing
General ledger	Quality management
Product-cost accounting	Routing management
Profitability analysis	Shipping
Profit-center accounting	Vendor evaluation
Standard and period-related costing	
Human Resources	**Sales and Marketing**
Human resources time accounting	Order management
Payroll	Pricing
Personnel planning	Sales management
Travel expenses	Sales planning

reduce or eliminate the need to customize a single-vendor solution, and reduce user resistance. Jones and Young (2006) found that 18% of companies used this approach to select ERP software packages.

This article examines what the best-of-breed strategy is, when it is used, what advantage adopting companies seek, examples of best-of-breed implementations, and differences in implementation methods.

BACKGROUND

ERP implementation projects can be distinguished from other IT projects by three characteristics (Somers, Ragowsky, Nelson, & Stern, 2001). First, ERP systems are "profoundly complex pieces of software, and installing them requires large investments in money, time and expertise" (Davenport, 1998, p. 122). Second, ERP packages may require the user to change business processes and procedures, may require customization, and may leave the firm dependent on a vendor for support and updates (Lucas, Walton, & Ginsberg, 1988). Finally, adopting firms are usually required to reengineer their business processes. Implementation projects must be managed as broad programs of organizational change rather than a software implementation (Markus & Tanis, 2000; Somers et al., 2001).

ERP systems include functionality for basic business processes based on the vendor's interpretation of best practices. However, the selected functionalities do not generally match the existing business processes of all organizations and may not be the best practices for a particular organization.

Typical ERP functions from SAP R/3, a major ERP vendor, are shown in Table 1. SAP R/3 modules provide a wide range of functional solutions, however, with the wide range of potential ERP customers, some organizations may not be a good fit. With the best-of-breed strategy, organizations can pick and choose the ERP modules from whichever vendor provides the best fit with their business processes and possibly reduce the amount of reengineering of business processes required, hence reducing the level of employee resistance.

BEST OF BREED IN INFORMATION SYSTEMS

The term *best of breed* was originally used in information systems literature to describe a situation where individual departments are allowed to install systems that best meet their needs rather than adhere to a corporate standard.

Acquisition costs are lower when all departments use the same software systems because of joint-purchase benefits such as volume discounts and other economies of scale. However, other costs may offset these savings. Unit document costs may be higher on a single-vendor approach compared to best of breed. Costs to translate and reformat data may be excessive. Switching costs may differ depending on the system chosen (Dewan, Seidmann, & Sunderesan, 1995).

BEST-OF-BREED ERP

ERP vendors design systems that are "purported to represent best practice and a more competitive business model." However, organizations interested in adopting ERP argue that "ERP software functionality is often lacking, the implicit business model does not represent their own and therefore reengineering business processes in line with this presents major difficulties" (Light & Holland, 2001, p. 217).

Single-vendor packages seem to have strengths in a particular functional area. PeopleSoft is known for exceptional human resource modules and Oracle has a reputation for exceptional financial modules.

Best-of-breed solutions provide an alternative strategy to enable organizations to implement ERP when a single vendor may not provide the functionality that the adopter requires or when modules from different vendors may provide a better match with existing or required business processes than a single-vendor solution. Lack of feature-function fit may be due to the design of most ERP systems for discrete manufacturing. Many organizations have specialized business processes common to their industry that may not be solved by the best practices embedded into single-vendor ERP systems. Various modules may not support process manufacturing industries, such as food processing and paper manufacturing; project industries, such as aerospace; or industries that manufacture products with dimensionality, such a clothing or footwear (Markus & Tanis, 2000).

While providing the additional needed functionalities, the best-of-breed approach complicates integration. With a single-vendor ERP system, "the whole package is designed for data compatibility" (Grant & Tu, 2005). With best-of-breed implementations, middleware is usually needed to link the various modules and databases: "The chance of being able to arrive at the same levels of integration as with an ERP system is very low, but this may be worth accepting as a means of saving the cost and pain associated with ERP implementation" (Payne, 2002).

Although most firms select a single vendor, a survey of Fortune 1000 firms found that 18% of the respondents chose ERP packages based on best-of-breed criteria; 32% of respondents used a combination of packages, such as SAP, PeopleSoft, Oracle, Baan, JD Edwards, Lawson, Adage, and SSA/CT (Jones & Young, 2006).

Although little empirical research has been done on best-of-breed ERP implementations, the information shown

Table 2. Best of breed vs. single-vendor ERP (Source: adapted from Light, Holland, & Wills, 2001)

BEST OF BREED	SINGLE-VENDOR ERP
Organization requirements determine functionality	Vendor determines functionality
Context approach to business process reengineering (BPR)	Clean-slate approach to BPR
Good flexibility in process redesign due to choice of components	Limited flexibility on process redesign
Reliance on numerous vendors distributes risk	Single vendor may increase risks
IT department requires multiple skills sets to deal with multiple software sources	Single skills set required in IT
Capabilities may be retained or enhanced with unique combinations of vendor packages and custom components	Distinctive capabilities may be impacted in common business process throughout industry
Need for flexibility and competitiveness is acknowledged up front	Flexibility and competitiveness may be constrained
Integration of applications may be time consuming and upgrades can be complicated	Integration of applications is precoded into system and maintained in upgrades

in Table 2 is the result of a single case study of a business referred to as Global Entertainment (Light & Holland, 2001). The best-of-breed approach clearly gives the adopting organization the ability to determine functionality rather than accepting a single vendor's determination of best practices. The adopting organization can select packages that best support its existing or desired business processes.

The adoption of a single-vendor ERP may cause the loss of competitive advantage. The single-vendor software may require the adopting organization to abandon its existing business processes, which may have created competitive advantage for the organization in favor of ERP-vendor-defined best practices. The organization's competitors may have also adopted these vendor-defined best practices, leaving the organization without a competitive advantage.

The best-of-breed approach may also spread the risk of the failure of an ERP vendor. If a vendor drops out of the ERP market for any reason, only a part of the system will be affected, not the entire system.

On the negative side, the best-of-breed approach requires more knowledge and skills in the adopting firm's IT department to support multiple packages. IT staff must be trained to support software from multiple vendors and maintain integration software as packages are upgraded.

Integration of modules from more than one vendor may be time consuming and costly. Upgrades to any of the vendor packages can cause complexities not encountered with a single-vendor approach.

EXAMPLES OF BEST OF BREED

Examples of best-of-breed selection and implementation projects may provide insight into the reasons best of breed remains a viable option for firms selecting ERP systems. Example 1 is a simple best-of-breed project with only two vendors. Example 2 represents a more complex best-of-breed project involving five different vendor packages.

Example 1

This example presents a relatively simple use of a best-of-breed solution. In 1995, a Houston-based energy services company embarked on a major systems effort. Legacy systems were accounting oriented and provided little operating information. Y2K (year 2000) problems were the catalyst for proceeding with the project.

With extensive help from the consultants on the project, Oracle was selected for the financial part of the new system. Oracle was considerably less expensive than other systems the company considered and the majority owner of the company implemented Oracle several years earlier. The decision was complicated by the lack of an Oracle module to support process manufacturing.

To resolve Oracle's inability at the time to support process manufacturing, the company adopted a process manufactur-

ing package from Datalogix called Global Enterprise Manufacturing Management Systems (GEMMS). The interface software between Oracle and GEMMS presented many technical implementation problems, which were resolved by the vendors and implementation consultants. While these technical integration problems caused some delay, they were not major issues from the point of view of the company.

During the course of the implementation project, a risk of the best-of-breed strategy occurred. Oracle acquired Datalogix in the midst of the project. Instead of helping resolve any problems by consolidating the software in one vendor, the acquisition exacerbated the problem because of a postacquisition exodus of Datalogix personnel.

The best-of-breed approach did not significantly contribute to cost overruns or delays in implementation. Overall, the project is regarded as a success. The problems caused in the integration of the two vendor solutions impacted mainly the software vendors and consultants, although company personnel had to coordinate the parties in reaching solutions. Project success can be attributed to the existence of a champion, training, and use of a project manager with both ERP and project management experience (Bradley, 2005).

Example 2

This example involves three divisions of a major defense contractor based in southern California. A series of acquisitions had left it with several nonintegrated mainframe systems supplemented by personal computer based point solutions. Prior to the ERP project, each division was pursuing different approaches to its information systems needs. One division was in the process of implementing an out-of-the-box Baan ERP system, the second division was using a heavily customized WDS system, and the third had no IT infrastructure. The corporate parent of these divisions wanted a single solution to control costs and leverage its purchasing power on the purchase of the software. This goal was complicated because the divisions had been highly independent of parent control and no single vendor offered a solution that met the needs of all three divisions. Standard packages failed to meet the needs of the bulk of their business. The parent was faced with balancing the benefits, cost savings, and standardization with the flexibility and independence of the divisions.

In 1996, when the project started, no single vendor provided a solution to the company's needs. A joint oversight team from the three divisions selected WDS (now Manugistics) as the core, but incorporated best-of-breed solutions such as PeopleSoft for human resource management, Oracle for financials, TIP QA for quality control, and Matrix One for product data management.

The best-of-breed approach did not cause significant problems according to the project manager; however, this project was his first ERP implementation, so he had little basis for comparison. The project was completed 6 months late and was $2 million over budget, although a contingency budget had addressed possible overruns.

FUTURE TRENDS

The Gartner Group coined the term *ERP II* to describe the shift in ERP from an enterprise information base focusing on back-office transaction processing within one organization to moving information across the supply chain ("Taking the Pulse of ERP," 2001). Others refer to this process as extended enterprise systems. ERP II or extended enterprise systems include customer relationship management (CRM) applications that accumulate information to better serve customers, supply chain management (SCM) applications that manage materials and services from acquisition to delivery to the customer, and e-business applications to enable the organization to reach customers over the Internet. Davenport and Brooks (2004) refer to basic ERP as infrastructural capability and SCM modules as strategic capabilities, raising the question of "whether to implement the infrastructural capabilities first, strategic capabilities first, or both simultaneously." They believe that while the infrastructural capabilities of ERP "provide very little in the way of real business value…they are critical to long-term internal and external integration." What they lack are short-term cost savings and competitive advantage. SCM and other extended enterprise systems modules have a payoff in terms of competitive advantage.

Best-of-breed terminology, while originally used for ERP software, is also applied to ERP II software. Major ERP vendors have concentrated on the transaction processing or back-office procedures in ERP. Independent vendors have concentrated on the add-ons that constitute ERP II. Davenport and Brooks (2004) point out that this situation is changing. Mainstream ERP vendors are adding SCM and other functionalities to their packages, avoiding some of the integration issues, but smaller vendors "have the edge in state-of-the-art functionality." This observation means that the best-of-breed strategy will continue to be a viable ERP II strategy choice.

CONCLUSION

The best-of-breed strategy will continue to be a viable option in both ERP and ERP II for organizations that attempt to create competitive advantage by assembling a custom ERP system rather than adopting the same off-the-shelf, "vanilla" ERP systems adopted by their competitors. While most adopting organizations will stick with the full integration of a single-vendor package, firms willing to take risks to obtain competitive advantage will assemble a best-of-breed solution that enhances their ability to serve their customers.

Example 1 demonstrates how a best-of-breed strategy can compensate for the lack of a desired functionality in one vendor's package by supplementing it with a second package. Example 2 shows how the strategy can blend together the diverse needs of three operating divisions to arrive at a solution acceptable to all parties. While one of these examples was clearly successful and the other had overruns in both cost and time, the best-of-breed option did not seriously hamper either of these implementations.

Both implementation strategies will continue to be complex and costly. Best of breed may provide an avenue to reduce the reengineering required at the cost of increased integration problems. The single-vendor approach may result in more reengineering of business processes but avoids the integration issues.

The best-of-breed implementation strategy is an area of ERP systems that has not been fully explored. More research of this promising alternative implementation strategy is needed.

REFERENCES

Bradley, J. (2005, August 11-14). Are all critical success factors created equal? *Proceedings of Eleventh Americas' Conference on Information Systems*, Omaha, NE (pp. 2152-2159).

Davenport, T. H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review, 76*, 121-131.

Davenport, T. H., & Brooks, J. D. (2004). Enterprise systems and the supply chain. *Journal of Enterprise Information Management, 17*(1), 8-19.

Dewan, R., Seidmann, A., & Sunderesan, S. (1995). *Strategic choices in IS infrastructure: Corporate standards versus "best of breed" systems.* Paper presented at the ICIS, Amsterdam.

Gelinas, Sutton, & Fedorowicz. (2004). *Business processes and information technology.* Thomson-South-western.

Grant, D., & Tu, Q. (2005). Levels of enterprise integration: Study using case analysis. *International Journal of Enterprise Information Systems, 1*(1), 1-22.

Jones, M. C., & Young, R. (2006). ERP usage in practice. *Information Resources Management Journal, 19*(1), 23-42.

Light, B., & Holland, C. P. (2001). ERP and best of breed: A comparative analysis. *Business Process Management Journal, 7*(3), 216-224.

Lucas, H. C., Jr., Walton, E. J., & Ginsberg, M. J. (1988). Implementing packaged software. *MIS Quarterly*, 537-549.

Markus, M. L., & Tanis, C. (2000). The enterprise experience: From adoption to success. In R. W. Zmud (Ed.), *Framing the domains of IT research: Projecting the future through the past.* Cincinnati, OH: Pinnaflex Educational Resources, Inc.

Pan, S. L. (2005). Customer perspective of CRM systems: A focus group study. *International Journal of Enterprise Information Systems, 1*(1), 65-88.

Payne, W. (2002). The time for ERP? *Work Study, 51*(2/3), 91-93.

Pozzebon, M. (2000). *Combining a structuration approach with a behavioral-based model to investigate ERP usage.* Paper presented at AMCIS 2000, Long Beach, CA.

Somers, T. M., Ragowsky, A. A., Nelson, K. G., & Stern, M. (2001). *Exploring critical success factors across the enterprise systems experience cycle: An empirical study* (working paper). Detroit, MI: Wayne State University.

Taking the pulse of ERP. (2001, February). *Modern Materials Handling*, pp. 44-51.

KEY TERMS

Best of Breed: It is a combination of ERP software provided by more than one vendor and legacy systems designed to meet the needs of an organization in a manner superior to the single-vendor ERP approach.

Business Processes: "A business process is a set of business events that together enable the creation and delivery of an organization's products or services to its customers" (Gelinas, Sutton, & Fedorowicz, 2004).

Customer Relationship Management (CRM): These are software packages that enable a business to develop knowledge of their customers' needs and buying patterns. These systems "focus on the integration of externally based customer data for the organization to pursue more customer-oriented activities like targeted advertising, one-on-one marketing, customer retention and building a real-time integrated view of the customer" (Pan, 2005).

Enterprise Resource Planning (ERP) System: An off-the-shelf accounting-oriented information system that is designed to meet the information needs of most organizations. ERP systems enable an organization to procure, process, and deliver customer goods or services in a timely, predictable manner. These systems are complex and expensive information tools that have proven difficult and time consuming to implement.

ERP II or Extended Enterprise Systems: ERP II opens ERP systems beyond the enterprise level to exchange information with supply chain partners and customers. ERP II extends beyond the four walls of the business to trading partners. Typically, ERP II includes CRM packages, SCM packages, and e-business packages.

Integration: Integration is generally defined as "the bringing together of related components to form a unified whole….The primary concern of integration is 'oneness' and 'harmony' between user, technology, and the environment" (Grant & Tu, 2005, p. 8). Grant and Tu propose a taxonomy of ERP integration ranging from the lowest level, system specification integration, to global integration, which deals with "issues of language, time difference, culture, politics, customs, management style." Their proposed Level II deals with system-user integration at both the ergonomic and cognitive level. Level III deals with the integration of islands of technology throughout the firm.

Legacy Systems: They are transaction processing systems designed to perform specific tasks, or systems that have become outdated as business needs change and the hardware and software available in the marketplace have improved.

Manufacturing Resources Planning (MRPII): MRPII extends MRP by addressing all resources in addition to inventory. MRPII links material requirements planning with capacity requirements planning avoiding over- and under-shop-loading typical with MRP.

Material Requirements Planning (MRP) Systems: They are processes that use bills of materials, inventory data, and a master productions schedule to time-phase material requirements, releasing inventory purchases in a manner that reduces inventory investment yet meets customer requirements.

Supply Chain Management (SCM): These software packages exchange information with supply chain partners to order and track the procurement of goods and services. SCM can be viewed in four basic categories (Davenport & Brooks, 2004): supply planning tools, demand planning tools, plant scheduling tools, and logistics systems. A newer functionality in SCM is collaborative planning, forecasting, and replenishment (CPFR). In CPFR, "supply chain partners exchange not only orders and shipment notices, but sales plans and production forecasts with each other, so that they can synchronize their respective processes more fully."

Category: Business Information Systems

ERP Systems' Life Cycle: An Extended Version

Cesar Alexandre de Souza
University of São Paulo – Brazil, Brazil

Ronaldo Zwicker
University of São Paulo – Brazil, Brazil

INTRODUCTION

The 1990's witnessed an impressive growth of Enterprise Resource Planning (ERP) systems in the market of corporate IT solutions, and now they are an important component of IT architecture in many companies. The ERP systems are introduced in companies following well-defined stages, namely the stages of decision, selection, implementation, stabilization and utilization. This last stage (utilization) is also characterized by the development of an organized effort to continuously ensure that the ERP system meets business needs regarding functionality, performance, availability, and to control operation costs, at the ERP management stage. This chapter presents aspects involved in each stage of this life cycle, based on the referenced bibliography.

BACKGROUND

Enterprise Resource Planning (ERP) systems are integrated information systems acquired as commercial software packages that aim supporting most of the operations of a company. Markus and Tanis (2000) define them as commercial packages that enable the integration of data coming from transactions-oriented information systems and from the various business processes throughout the entire organization. Examples of ERP systems found on the market are the SAP ERP of the German company SAP and Oracle Applications of the American Oracle. Some authors present and describe characteristics that allow differentiating ERP systems from systems developed within the companies and from other types of commercial packages (Markus & Tanis, 2000; Souza & Zwicker, 2001). These characteristics may be summarized as:

- ERP systems are commercial software packages;
- They include standard models of business processes;
- They are integrated information systems and use a corporate data base;
- They have a large functional scope;
- They require adjustment procedures to be used in a given company.

When deciding to use ERP systems, companies hope to achieve manifold benefits, like business processes integration, increased of control of operations, technological updating, IT cost reduction and access to information in real time for decision making. However, there are also problems to be considered, with implementation failures being reported (Barker & Frolick, 2003). Table 1 synthesizes benefits and difficulties of ERP systems mentioned by many authors (Bancroft, Seip & Sprengel, 1998; Davenport, 1998), and relates them to ERP systems' characteristics.

Although the initial focus of ERP systems was the integration of the internal value chain of large industrial companies, they are now evolving to a wider scope, including interenterprise integration features (McGhaughey & Gunasekaran, 2007). In many cases, the ERP systems became the basis upon which companies begun to develop other initiatives such as: customer relationship management (CRM), supply chain management (SCM) and business intelligence (BI). ERP systems are also now present in a growing number of companies in financial and service sectors and several vendors are now focusing small and medium companies as their target for market expansion.

ERP SYSTEMS LIFE CYCLE MODEL

The life cycle of information systems represents the various stages through which a project of development and utilization of information systems passes through. In its traditional form, the systems development life cycle encompasses project definition, system study, design, programming, installation, and post-implementation stages.

In the case of use of commercial software packages, these stages may differ. For instance, in the system study stage the focus is not on obtaining a detailed system specification from the users for programming the system, but instead, verifying the functionality of the many choices available from vendors, against a set of requisites from the users that will guide system adaptation or customization.

Like any commercial software package, ERP systems exhibit differences in their life cycle regarding traditional systems development projects. But because of their large functional scope and the integration between its various

Table 1. ERP systems benefits and difficulties

Characteristics	*Benefits Sought*	*Possible Difficulties*
Commercial Package	- IT costs reduction - Focus on company's core activities - Technological updating - Backlog reduction	- Supplier dependence - Lack of knowledge on the package - Loss of previous systems functionalities
Best Practice Business Models	- Knowledge on best practices - Access to other companies' experiences	- Need to adjust the company to the package - Need to change business procedures - Need of consulting for implementation
Integrated Information System	- Greater control on the company's operation - Real time access to data and information - Elimination of interfaces between isolated systems - Improvement of information quality - Synchronization between activities of the value chain	- High implementation complexity and costs - Difficulty to update the system as it requires agreement among various departments - One module not available may interrupt the functioning of the others - Resistance due to increase of demands to the areas responsible for data input
Corporate Data Base	- Standardization of information and data definitions - Elimination of discrepancies between information of different departments - Information quality improvement - Access to information for the whole company	- Cultural change of the view of "owner of the information" to that of "responsible for the information" may cause resistance to change - Responsibilities attribution on files shared between areas - Overload of the data base causing performance problems
Great Functional Scope	- Maintenance elimination of multiple systems - Standardization of practices - Reduction of training costs - Interaction with a single supplier	- Dependence upon a single supplier - If the system fails the entire company may stop - Support difficulties in the stabilization phase

modules, these differences are deepened. Some authors present models for the ERP systems' life cycle (Esteves & Pastor, 1999; Markus & Tanis, 2000; Souza & Zwicker, 2001). The main features of these many models are summarized in Figure 1, which includes the stages of decision and selection, implementation, stabilization, and utilization. The ERP management stage is included as an addition to the utilization stage and as an extension to the traditional ERP life cycle model, and is described next, along with the other stages.

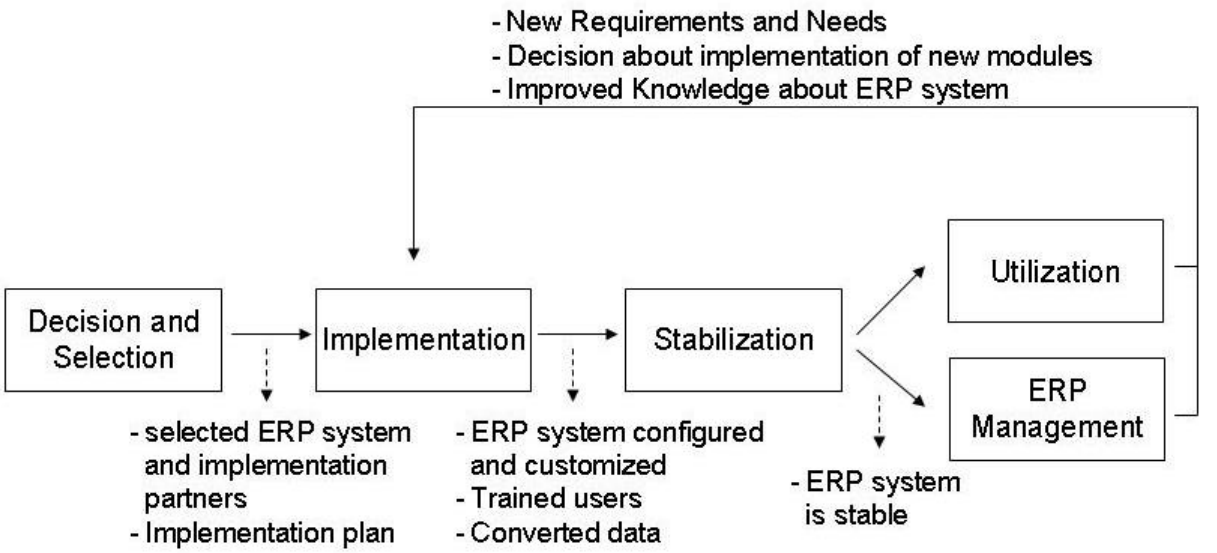
ERP Systems Decision and Selection

At the decision and selection stage the company decides to implement an ERP system as an IT solution and chooses the vendor. A series of issues must be taken into account at this stage. For instance, Davenport (1998) analyzes the decision from the point of view of the compatibility between the organization and the characteristics of the ERP systems. Hecht (1997) presents criteria that may help in this choice: adjustment of the package's functionality to the requisites of the company, technical architecture of the product, implementation costs, quality of post-sales support, and financial health of the vendor and its vision for the future of the package. The main product of this stage is a detailed implementation plan, where the modules to be implemented, the implementation approach, the project schedule, the implementation team and responsibilities are defined. Also, it is very common to hire a consulting company during the implementation project, for tasks that range from development of customizations to full responsibility for project management, depending on the case and the knowledge available in the company.

ERP Systems Implementation

Implementation comprises the second stage of the ERP systems' life cycle. The implementation of an ERP system may be defined as the process by which the system's modules

Figure 1. ERP systems life cycle model (adapted from Souza & Zwicker, 2001)



are put into operation within a company. Implementation entails adjustment of the business process to the system, configuration and eventual customizing of the system, loading or conversion of initial data, hardware and software configuration and training of users and managers. This stage encompasses the tasks ranging from the end of implementation plan's concept to the beginning of the operation.

The implementation stage is reported to be the most critical of all (Bingi, Sharma & Godla, 1999). Difficulties are mainly due to organizational changes that imply in changes of tasks and responsibilities of individuals and departments and transformations in the relationships between different departments. In an ERP system implementation it is generally pursued the optimization of the global processes of the company, which may cause as a counterpart changes in the activities in most of the departments involved. The need of intense participation and commitment of the company's top management and the requirement of permanent communication among the involved units is brought about by the size and complexity of this change and of the conflicts it may generate among those involved.

Adjusting the ERP system to company's processes is part of the implementation stage and is achieved by adjustment of parameters (configuration) or software customization (development of programs to modify or complement existing functions). At the implementation stage, the decision on how to start the operation of the ERP system (the "go-live") is also important. All modules may start operating in all divisions or plants of the company simultaneously (*big bang)* or in one division or plant after the other (*small bangs)*. The start may also occur in *phases* (one or some modules start operating in one division or plant after the other, also called *roll-out*). The approach used to start the ERP system operation is an important decision in its implementation project as it greatly affects the configuration of the system, the allocation of resources and the management of the project and its risks. It also plays a decisive role at all stages of ERP system's life cycle.

ERP systems' integration entails difficulties for the ERP systems implementation stage. These difficulties are related to three types of changes in the way people do their work (Souza & Zwicker, 2001):

(1) The integration transfers to departments that produce the information the responsibility to insert it properly. This includes data used by other departments only (for instance, typing of an accounting bill in a production entry) and as a consequence the users feel that their tasks are increased.
(2) Information must be inserted into the system at the best-suited moment for the process and not at the best-suited moment for a specific department. Thus, there is a need to change the way which tasks are carried out and other departments begin to demand the information they rely upon.
(3) The activities of a department become transparent to all others and this has the inconvenience to require "explanations" for everything it does.

Training of end users for working with an integrated system is an important consideration for the success of the implementation process.

Stabilization

In the first moments after the beginning of the ERP system operation there is a critical stage for the project's success: the stabilization. At this moment, the ERP system, that until then was only an abstraction, gains reality and starts to be part of company's and people's daily lives. This is when the highest amount of energy, be it managerial or technical, is required. It is a stage in which problems that could not have been easily detected at the implementation stage become apparent. This is a particularly critical stage, as the company is already relying upon the system for its activities and which causes major pressure for quick solution of problems. The length of this period depends on the company and takes about eight weeks (Zwicker & Souza, 2004).

The exact characterization of this stage is related to the operation starting mode chosen by the company. If operation of the ERP system started by means of a *big bang*, the stage of stabilization can be clearly distinguished from those of implementation and utilization. However, in companies that implement the modules in phases, or even in *small-bangs*, the stabilization stage is less characterized and merges with the implementation stage of the remaining modules. It can be stated that the stabilization stage in the case of implementation by phases starts with the operation of the first module and ends only when the last module implemented, in the last locality of the company, becomes stabilized. This longer implementation and stabilization time in general entails loss of focus of the project and may be viewed as a risk factor for the implementation in phases (for further details on the influence of ERP *go-live* approach on its life cycle, see Zwicker & Souza, 2004).

Utilization

Finally, at the fourth stage the system starts to belong to day-by-day operations. This does not mean that all its possibilities of usage are already known, nor that they are properly equated. Orlikovski and Hofman (1997) report on the difficulty within a company to know beforehand all use possibilities of the new information technologies. This knowledge is only achieved after a certain period of continued use of the technology, through the ideas that emerge during the utilization process. Therefore, the stage of utilization feeds back the stage of implementation with new possibilities and needs that may be solved through new modules, parameter adjustments or software customizing. In the case of implementation in phases, the already implemented modules may impose restrictions upon new modules caused by already defined parameters or customizations. Markus and Tanis (2000) point out that it is only in this stage that the organization is finally able to ascertain the benefits (if any) of its investment in the ERP system, due to continuous business improvement, additional user skill building, and post implementation benefits assessment. Also at this point, companies begin to consolidate their process reviews and effectively employ a model of integrated process management.

This stage can also be viewed as an evolutionary process. For instance, Botta-Genoulaz and Millet (2005) present a model to asses the ERP use maturity in three stages: mastery of the software, it is, the acquisition of the technical and operational knowledge about the ERP system; improvement, when the company has enough knowledge to effectively use the ERP systems for processes redesign and integration; and evolution, when it is possible to effectively align the ERP use to business strategies.

ERP Management

After its implementation, the ERP system becomes a critical component of the integrated management of diverse company areas and of supply chain management, which demands compliance with extreme availability and performance requirements. According to Zwicker and Souza (2006), ERP management encompasses the set of actions undertaken to ensure meeting the business needs, the performance, availability and control of the maintenance and operation costs. ERP systems management includes the activities of development (implementation and evolution of the system), operation (keep the system operation within the specified parameters of performance and availability), support (user services) and planning. Many of these tasks may be performed by the IT staff in cooperation with users (for example, the first instance of problem resolution may be done by users with good knowledge of the ERP system), as well as with outsourced personnel and consultants.

Esteves and Pastor (1999) also point out an interesting phase of the life cycle of ERP systems, the "retirement" stage, in which the ERP system would be substituted for other solution or solutions. What is being observed is that companies go through great efforts and costs each time a new version upgrade is needed (McMahon, 2004). Every time the decision to upgrade or not is posed to the organization it is also an opportunity to analyzing other ERP vendors or technologies (Kremers & van Dissel, 2000).

FUTURE TRENDS

The ERP Market has changed considerably since the 1990's. After a period of huge expansion based on sales for new clients who were manly large companies, the market stabilized and a series of mergers and acquisitions occurred. Vendors are now looking for new ways of increasing their revenues. Although the idea of one single system for the whole enterprise has not been reached yet, ERP vendors are incorporating new functionalities that comprise front-office and external integration tasks in their systems. Also,

companies in sectors not traditionally related to ERP, like finance, services and universities are beginning to implement it (McGhaughey & Gunasekaran, 2007). Another trend is the focus on the small and medium enterprise (SME) market. The implementation of whole ERP systems in these companies poses great challenges to vendors, due to the lack of financial, human and knowledge resources these companies usually present. Vendors are also struggling to innovate and incorporate new technologies and trends, like collaboration, business process management, service-oriented architectures (SOA), and business intelligence (Wailgum, 2007).

There are also challenges related to the management of IT in an ERP system context. An important option many ERP users are adopting is participating in user groups. This can be understood as an alternative strategy to obtain the knowledge needed to keep the ERP system up to date and better aligned with business needs.

CONCLUSION

This chapter presented an extended model for ERP systems life cycle that tries to encompass the complexities involved in implementing ERP systems in companies. To achieve this purpose, several aspects relating to the decision and selection, implementation, stabilization, utilization, and ERP management stages were presented.

The recommendation for companies that are deciding for ERP systems utilization and for companies that are already in the implementation stage is the careful analysis of the difficulties associated which each ERP systems' life cycle stage. With a better knowledge about these difficulties, the process can be improved though better planning and better action on the inherent difficulties of such an organizational change. One critical factor for success is to avoid it to be handled as an IT project (Willcocks & Sykes, 2000). Dedication and involvement of top management, strong participation of users and change management were aspects considered essential for the success of ERP systems implementation projects (Bingi, Sharma & Godla, 1999).

REFERENCES

Bancroft, N. H., Seip, H., & Sprengel, A. (1998). *Implementing SAP R/3: How to introduce a large system into a large organization*. Greenwich: Manning.

Barker, T. & Frolick, M. (2003). ERP implementation failure: A case study. *Information Systems Management*, 43-49.

Bingi, P., Sharma, M., & Godla, J. (1999). Critical issues affecting an ERP implementation. *Information Systems Management, 16*(3), 7-44.

Botta-Genoulaz, V. & Millet, P.-A. (2005). A classification for better use of ERP systems. *Computers in Industry, 56*, 573–587.

Davenport, T. H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review, Jul/Aug*, 121-131.

Esteves, J. M. & Pastor, J. A. (1999). An ERP life-cycle-based research agenda. In *Proceedings of the First International Workshop in Enterprise Management and Resource Planning: Methods, Tools and Architectures – EMRPS'99*, Venice, Italy.

Hecht, B. (1997). Chose the right ERP software. *Datamation*, March.

Kremers M. & van Dissel, H. (2000). ERP system migrations. *Communications of the ACM, 43*(4).

Markus, M. L. & Tanis, C. (2000). The enterprise system experience: From adoption to success. In Zmud, R. (Ed.), *Framing the domains of IT research: Glimpsing the future through the past.* Cincinnati, CT: Pinnaflex.

McGhaughey, R. E. & Gunasekaran, A. (2007). Enterprise resource planning (ERP): Present, past and future. *International Journal of Enterprise Information Systems, 3*(3), 23-35.

McMahon, S. (2004). Beating the clock on ERP upgrades. *Datamation*, June 7.

Orlikovski, W. J. & Hofman, J. D. (1997). An improvisational model for change management: the case of groupware technologies. *Sloan Management Review, Winter*, 11-21.

Souza, C. A. & Zwicker, R. (2001). ERP systems' life cycle: Findings and recommendations from a multiple-case study in Brazilian companies. In *Proceedings of the 2001 Conference of Business Association of Latin American Studies - BALAS 2001*, San Diego.

Wailgum, T. (2007). Under pressure, ERP giants struggle to innovate. *CIO UK, November*, 12.

Willcoks, L. P. & Sykes, R. (2000). The role of the CIO and IT function in ERP. *Communications of the ACM, 43*(4), 33-38.

Zwicker, R. & Souza, C. A. (2004). SAP R/3 Implementation approaches: A study in Brazilian companies. In L.K. Lau (Ed.), *Managing business with SAP: Planning, implementation and evaluation.* Hershey: Idea Group Publishing.

Zwicker, R. & Souza, C. A. (2006). A model for ERP systems management: An exploratory study in companies using SAP R/3. In *Proceedings of the Twelfth Americas Conference on Information Systems – AMCIS,* Acapulco, Mexico.

KEY TERMS

ERP Systems: Integrated information systems purchased as commercial software packages with the aim of supporting most operations of a company.

System Development Life Cycle: The various stages through which a project of development and utilization of information systems passes.

ERP Systems' Life Cycle: The various stages through which a project of introducing an ERP system in a company passes through.

ERP Decision and Selection Stage: Stage at which the company decides to implement an ERP system and chooses the supplier.

ERP Implementation Stage: Stage of an ERP project at which the ERP system's modules are put into operation.

ERP Stabilization Stage: The first weeks after the beginning of an ERP system operation in the company.

ERP Utilization Stage: Stage of an ERP project at which the system starts to belong to the day-by-day operations of the company.

ERP Management: The group of actions carried out to ensure the fulfillment of business requirements, the performance, the availability and the control of maintenance and operation activities of ERP systems.

Big-Bang Approach: Implementing all modules of an ERP system in all locations or plants of the company simultaneously.

Phased Approach: Implementing the modules on an ERP system in sequential steps, comprising one or more modules in one or more locations or plants, until the system is completely installed. It is also called roll out.

Establishing the Credibility of Social Web Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

In recent years, there has been a steady shift in the *nature* of Web applications. The vehicle of this transition of Web applications is *us*, the people. The ability to post photographs or videos, exchange music snippets with peers, and annotate a piece of information, are but a few exemplars of this phenomenon. Indeed, the pseudonym Web 2.0 (O'Reilly, 2005) has been used to describe the apparent "socialization" of the Web.

In spite of the significant prospects offered by human-centric Web applications, the mere fact that virtually *anyone* can set up such applications claiming to sell products and services or upload/post unscrutinized information on a topic as being "definitive," raises the issues of credibility from a consumers' viewpoint. Therefore, establishing credibility is essential for an organization's reputation and for building consumers' trust.

The rest of the article is organized as follows. We first provide the background necessary for later discussion. This is followed by the introduction of a framework within which different types of credibility in the context of human-centric Web applications can be systematically addressed and thereby improved. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

In this section, we present the fundamental concepts underlying credibility and present the motivation and related work for addressing credibility within the context of Web applications.

Basic Concepts of Credibility of Web Applications

For the purposes of this article, we will consider credibility to be synonymous to (and therefore interchangeable with) believability (Fogg & Tseng, 1999).

The concept of credibility can be classified based upon the types of user interactions with a Web application. A user could consider a Web application to be credible based upon direct interaction with the application (*active credibility*), or consider it to be credible in absence of any direct interaction but based on certain pre-determined notions (*passive credibility*). There can be two types of *active credibility*, namely *surface credibility*, which describes how much the user believes the Web application based on simple inspection, and *experienced credibility*, which describes how much the user believes the Web application based on first-hand experience in the past. There can be two types of *passive credibility*, namely *presumed credibility*, which describes how much the user believes the Web application because of general assumptions that the user holds, and *reputed credibility*, which describes how much the user believes the Web application because of a reference from a third party.

Related Work on Credibility of Web Applications

The issue of the credibility of Web applications has garnered attention in recent years from diverse viewpoints and this has lead to theoretical (Fogg, 2003; Metzger, 2005) and empirical (Consumer Reports WebWatch, 2005) studies pertaining to the credibility of both commercial and non-commercial Web applications.

There have been some partial efforts in addressing the credibility of Web applications. A set of guidelines for improving the credibility of Web applications have been presented (Fogg, 2003). However, these guidelines are stated in such a fashion that they can be open to broad interpretation, do not always present the relationships among them, and are stated at such a high-level that they may not always be practical or may be difficult to realize by a novice user.

A general framework for addressing the credibility of Web applications has been proposed previously (Kamthan, 2007; Kamthan, 2008). This article presents an adaptation as well as a modest extension of these works.

A SYSTEMATIC APPROACH TOWARDS THE CREDIBILITY OF WEB APPLICATIONS

In this section, we consider approaches for understanding and improving active and passive credibility.

Stakeholders and Credibility of Web Applications

We identify two broad classes of stakeholders with respect to their *roles* in relationship to a Web application: a *producer* (such as the provider or an engineer) is the one who owns, finances, develops, deploys, or maintains the Web application, and a *consumer* (such as a novice or expert user) is the one who uses the Web application for some purpose.

We then assert that credibility is a *perceived* quality attribute with respect to the stakeholders of a Web application. Indeed, we view credibility as a *contract* between a producer and a consumer. This contract can have ethical, legal, and/or moral implications.

Addressing Active Credibility of Web Applications

We consider a Web application to be an interactive information system and adopt semiotics (Shanks, 1999; Stamper, 1992) as the theoretical basis for communication of information. The active credibility of Web applications is viewed as a qualitative aspect and is addressed indirectly from the perspective of semiotics (Table 1).

We now discuss each of the components of Table 1 in detail.

Identification of Semiotic Levels

The first column of Table 1 addresses semiotic levels. We are particularly interested in the communicative properties of the representations of a Web application, which in semiotics we can view on six interrelated levels: physical, empirical, syntactic, semantic, pragmatic, and social.

We focus only on the quality concerns at the last two levels: at the *pragmatic level* the interest is in the utility of the representations to its stakeholders, while at the *social level* the interest is in the manifestations of social interaction among stakeholders with respect to the representations.

Decomposition of Semiotic Levels and Assignment of Quality Attributes

The second column of Table 1 draws the relationship between semiotic levels and corresponding quality attributes.

Since each semiotic level is rather high to be tackled directly, we decompose it further into quality attributes that are widely-known and relevant. Not all attributes corresponding to a semiotic level are on the same echelon, and therefore they are placed at different tiers. We contend that the quality attributes included are necessary but make no claim of their sufficiency. Also, the quality attributes are not necessarily mutually exclusive, and this dependency can be either favorable or unfavorable (Wiegers, 2003). We note that some of the quality attributes are classical and relevant in a desktop environment but they get amplified, and in certain cases exacerbated, in a networked environment.

Specifically, credibility belongs to the social level and depends on the layers beneath it. The quality attributes aesthetics (presentation), legality, privacy, security, and transparency (of the producer) also at the social level depend upon the quality attributes accessibility and usability at the pragmatic level, which in turn depend upon the quality attributes comprehensibility, interoperability, performance, readability, reliability, and robustness also at the pragmatic level.

We discuss only the entries in the social level in some detail. The sensitivity part of visual perception is strongly related to aesthetics as it is close to human senses. The artistic expression plays an important role in making a Web application "attractive" to its customers beyond simply the functionality it provides. It is critical that the Web application be legal (for example, is legal in the jurisdiction it operates and all components it makes use of are legal); takes steps to respect user's privacy (for example, does not abuse or share user-supplied information without permission); takes steps to secure itself (for example, in situations where financial transactions are made). The provider must take all steps to be transparent with respect to the user (for example, not include misleading information such as the features of products or services offered, clearly label promotional

Table 1. A semiotic framework for active credibility of Web applications

<table>
<tr><th>Semiotic Level</th><th>Quality Attributes</th><th colspan="2">Means for Assurance and Evaluation</th><th>Decision Support</th></tr>
<tr><td rowspan="2">Social</td><td>Credibility</td><td rowspan="4">Process-Oriented:
Inspections, Testing

Product-Oriented:
Training, Guidance</td><td rowspan="4">Tools</td><td rowspan="4">Feasibility</td></tr>
<tr><td>Aesthetics, Legality, Privacy, Security, Transparency</td></tr>
<tr><td rowspan="2">Pragmatic</td><td>Accessibility, Usability</td></tr>
<tr><td>Comprehensibility, Interoperability, Performance, Readability, Reliability, Robustness</td></tr>
</table>

content, make available their contact information including physical address, policies regarding returning/exchanging products, and so on).

Next, we separate the semiotic quality attributes and the means for addressing those.

Means for Active Credibility Assurance and Evaluation

The third column of Table 1 lists the means for assuring and evaluating active credibility. We note that the mapping between the aforementioned pragmatic and social quality attributes and the means for addressing them is many-to-many.

We now briefly discuss two product-oriented means, namely training and guidance for assuring the active credibility of a Web application.

Training

Often, the courses related to the Web offered at institutions tend to focus primarily on the manipulations of the "popular" (and moving target) client- and/or server-side technologies-of-the-day. The result is that the students tend to learn more about "technology hacks" rather than the fundamentals of analysis and design necessary towards a *systematic* approach to the large-scale development of Web applications.

Apart from an exposure to a comprehensive technical background in Web engineering (Mendes & Mosley, 2006), the toolbox of a prospective professional Web engineer should include several other aspects: means of precisely identifying user classes, user preferences, and their needs; understanding of quality attributes specific to the domain of the Web application and their social manifestations; appropriate use of standards for both the process and the product; ability of journalistic writing, including the ability of balancing information with other types of media (related to marketing such as advertisements); training in informed and balanced decision making in order to analyze the trade-offs and decide amongst different design approaches, or between the use of early and established technologies; basic knowledge of issues related to legal issues such as those related to intellectual property rights (IPR) and licensing; and basic knowledge of financial issues (such as those related to merchant accounts and payment systems) in the lieu of support for commercial transactions.

Guidance

We consider guidelines and patterns as two "bodies of knowledge" based on past experience and expertise that can serve as aids for structured guidance.

The guidelines encourage the use of conventions and good practice. They could serve as a checklist with respect to which an application could be heuristically and, to certain extent, automatically evaluated. There are guidelines available for addressing accessibility (Chisholm, Vanderheiden, & Jacobs, 1999) and usability (Nielsen, 2000) of Web applications. However, guidelines suffer from certain limitations: they may seem rather general at times, often do not discuss trade-offs as a consequence of their application or relationships among them, and tend to assume a certain level of knowledge of the domain and therefore are more suitable for an expert than for a novice.

A pattern provides a conceptually reusable and proven solution to a recurring problem in a given context. It has been pointed out (Friedman, 2005) that identifying patterns for the design of Web applications could be useful towards improving the credibility of these applications. Indeed, patterns for Web applications have begun to appear (Van Duyne, Landay, & Hong, 2003), and a judicious use of patterns can tackle many of the pragmatic and social quality attributes in Table 1. However, there are certain caveats in the adoption of patterns: there is an evident cost involved in adaptation of patterns to new contexts; since the mapping between patterns and quality attributes is many-to-many, their selection may not be trivial; and there is always a distinct possibility that for a given problem, there simply may not be any suitable pattern available.

We next briefly discuss two process-oriented means, namely inspections and testing for evaluating the active credibility of a Web application.

Inspections

Inspections are a rigorous form of auditing based upon peer review that, when practiced well, can help evaluating some of the quality attributes at both pragmatic and social levels in Web applications. These are aesthetics, comprehensibility, legality, privacy, readability, and transparency. Inspections could, for example, assess if the presentation of information appears "professional", determine "sufficiency" of contact information, or decide what information is/is not considered "promotional."

Since inspections is a means for *static* verification, it can evaluate in rather limited form (if at all) the quality attributes that by necessity require some form of "dynamism" or real-world use. These include accessibility, interoperability, performance, reliability, robustness, security, and usability.

In spite of the usefulness of inspections in early defect detection, adoption can depend on the level of organizational process maturity, their effectiveness lies strongly on the reading technique deployed, and entail an initial cost overhead of training each participant in the structured review process followed by the logistics of checklists, forms, and reports involved.

Testing

Testing is a means for *dynamic* verification and is usually supported by most Web application development processes. The attributes of accessibility, interoperability, performance,

reliability, robustness, security, and usability can to a large extent be tested (semi-)automatically using tools or with the help of actual users.

However, not all quality attributes at either pragmatic or social levels in a Web application can be tested automatically. For example, it is not possible to completely test a Web application for aesthetics, comprehensibility, legality, privacy, readability, or transparency (like producer's intent) using tools; human inspection would be necessary for checking and determining the level of support of these quality attributes. Thus, inspections and testing do not replace but *complement* each other.

Tools

There are various tools that can help improve quality concerns at technical and social levels, manually, semi-automatically, or automatically. For example, they can help us detect security breaches, inform us of absence of privacy metadata, report violations of accessibility guidelines, or suggest image sizes favorable to the performance on the Web.

However, state-of-the-art tools can be expensive, although this situation is changing with the rise of open source software (OSS). They also may not always be applicable.

Decision Support

The last column of Table 1 acknowledges that the activities of assurance and/or evaluation must be realizable in practice.

The providers of Web applications take into account organizational constraints of time and resources (personnel, infrastructure, budget, and so on) and external forces (market value, competitors, and so on), which compels them to make quality related decisions that, apart from being sensitive to credibility, must also be feasible. For example, an a priori guarantee that a Web application will be credible to *all* users at *all* times in situations that they can find themselves in, is simply impractical.

The feasibility analysis is evidently related to decision making and could be a part of the overall Web application project planning activity. Further discussion of this aspect is beyond the scope of this article.

Addressing Passive Credibility of Web Applications

In this section, we briefly look into the case of passive credibility, specifically reputed credibility.

Like in the real-world contexts, Web applications could be audited for quality in general and credibility in particular. Indeed, WebTrust and TRUSTe are two relevant initiatives in the direction of addressing reputed credibility.

We acknowledge that the perceptions related to presumed credibility may be one of the most difficult to tackle. There are no absolute guarantees but the following could be helpful for presumed credibility assurance of a Web application: personalizing the application to user context, making organizational policies explicit, and appropriately labeling the nature of content as per the requirements of the Internet Content Rating Association (ICRA).

Scope of Credibility of Web Applications

We note that credibility is not a "universal" concern. The credibility of a Web application is a concern to a user if there is an associated cost (say, in terms of lost time, effort, or money) that is outright unacceptable to the user.

That the credibility of a Web application is a concern may also depend on the purpose of the interaction and the role played by the user. For example, credibility may be a lesser concern to a user if (s)he is casually browsing a gossip column on a movie artist than if (s)he is filling out an annual tax return form.

We also note that credibility is *not* a quality attribute that is absolute with respect to users or with respect to the Web application itself. We contend that for a Web application to be labeled as non-credible there must exist at least a part of it that is labeled non-credible based on the aforementioned classification by at least one user at some point in time. For example, a user may question the credibility of information on a specific product displayed on a specific "Web page" within a Web application.

FUTURE TRENDS

The work presented in this article can be extended in a few different directions, which we now briefly discuss.

It is known that, when applied judiciously, standards can contribute towards quality improvement. Indeed, credibility has recently been a topic of interest in standards for Web applications such as the IEEE Standard 2001-2002. However, awareness and broad use of these standards among engineers is yet to be seen.

Due to the unique nature of Web applications, any initiative towards addressing the credibility of Web applications should take place within the auspices of a development process that is that is sufficiently *agile*. This would require that the current agile methodologies for the development of Web applications evolve to provide support for credibility in general and for pragmatic and social quality attributes discussed in this article in particular.

The semantic Web has recently emerged as an extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning (Hendler, Lassila, & Berners-Lee, 2001). At the highest level of this infrastructure is the issue of trust. For

example, ontologies are central to the semantic Web and their development and subsequent use is based on mutual trust among the stakeholders. For the sustainability of the Web architecture, it is critical that the social Web and the semantic Web co-exist and evolve harmoniously. However, the "human" aspects of the semantic Web remain largely unaddressed. A natural extension of the discussion on credibility of the preceding section could be within the context of semantic Web applications.

CONCLUSION

By shifting from a collective of computers towards a community of people, the Web is becoming a symbiotic means of contribution, participation, and collaboration. The consumer concerns of credibility and the extent to which they are addressed will remain a key determinant towards the success of this paradigm.

Although there have been many advances towards enabling the technological infrastructure of the Web in the past decade, there is much to be done in addressing the social challenges, including user perceptions and expectations.

In conclusion, if credibility is important to an organization, it needs to be considered as a *first-class* concern, rather than an afterthought, from inception to conclusion of a Web application development process. Addressing the credibility in a systematic and feasible manner is one step in that direction.

REFERENCES

Chisholm, W., Vanderheiden, G., & Jacobs, I. (1999). *Web content accessibility guidelines 1.0.* W3C Recommendation. World Wide Web Consortium (W3C).

Consumer Reports WebWatch. (2005). *Leap of faith: Using the Internet despite the dangers. Results of a national survey of Internet users for consumer reports WebWatch.* A Consumer Reports WebWatch Research Report. October 26, 2005.

Fogg, B. J. (2003). *Persuasive technology: Using computers to change what we think and do.* Morgan Kaufmann Publishers.

Fogg, B. J., & Tseng, S. (1999). *The elements of computer credibility.* The ACM CHI 99 Conference on Human Factors in Computing Systems, Pittsburgh, USA, May 15-20.

Friedman, B. (2005). *Credibility by design.* Internet Credibility and the User Symposium, Seattle, USA, April 11-13.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The semantic Web. *Scientific American, 284*(5), 34-43.

Kamthan, P. (2007). Towards a systematic approach for the credibility of human-centric Web applications. *Journal of Web Engineering, 6*(2), 99-120.

Kamthan, P. (2008). Addressing the Credibility of Web Applications. In: *Encyclopedia of Internet Technologies and Applications*. M. Freire, & M. Pereira (Eds.). IGI Global, 23-28.

Mendes, E., & Mosley, N. (2006). *Web engineering*. Springer-Verlag.

Metzger, M. (2005). *Understanding how Internet users make sense of credibility: A review of the state of our knowledge and recommendations for theory, policy, and practice*. Internet Credibility and the User Symposium, Seattle, USA, April 11-13.

Nielsen, J. (2000). *Designing Web usability: The practice of simplicity.* New Riders Publishing.

O'Reilly, T. (2005). *What is Web 2.0: Design patterns and business models for the next generation of software.* O'Reilly Network, September 30, 2005.

Shanks, G. (1999). *Semiotic approach to understanding representation in information systems.* Information Systems Foundations Workshop, Sydney, Australia, September 29.

Stamper, R. (1992). *Signs, organizations, norms and information systems.* The Third Australian Conference on Information Systems, Wollongong, Australia, October 5-8.

Van Duyne, D. K., Landay, J., & Hong, J. I. (2003). *The design of sites: Patterns, principles, and processes for crafting a customer-centered Web experience.* Addison-Wesley.

Wiegers, K. E. (2003). *Software requirements* (2nd ed.). Microsoft Press.

KEY TERMS

Credibility Engineering: The discipline of ensuring that a system will be perceived as credible by its stakeholders, and doing so throughout the life cycle of the system.

Delivery Context: A set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Quality: The totality of features and characteristics of a product or a service that bear on its ability to satisfy stated or implied needs.

Quality Model: A set of characteristics and the relationships between them that provide the basis for specifying quality requirements and evaluating quality of an entity.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Semiotics: The field of study of signs and the communicative properties of their representations.

Web Application: A specific to a domain Web site that behaves more like an interactive software system rather than a catalog: it will in general require programmatic ability on the server-side and may integrate/deploy additional software for some purpose (such as dynamic delivery of resources).

Web Engineering: A discipline concerned with the establishment and use of sound scientific, engineering and management principles and systematic approaches to the successful development, deployment, and maintenance of high-quality Web applications.

E-Technology Challenges to Information Privacy

Edward J. Szewczak
Canisius College, USA

INTRODUCTION

The collection of personal information by electronic technology (e-technology) and the possibility of misuse of that information are primary reasons why people limit their use of the Internet and are even limiting the success of e-commerce (Szewczak, 2004). Various uses of e-technology that collect and/or disseminate personal information include corporate and government databases, e-mail, wireless communications, clickstream tracking, and PC software. The main challenge to personal information privacy is *the surreptitious monitoring of user behavior on the Internet without the user's consent and the possible misuse of the collected information resulting in financial and personal harm to the user*. Our focus is primarily on Internet use in the United States of America, though clearly e-technology is global in nature and poses challenges and issues for societies around the world.

BACKGROUND

Concerns about the collection of personal information are ongoing. The results of a 1998 survey conducted by Louis Harris & Associates, Inc. revealed that worries about protecting personal information ranked as the top reason people generally are avoiding the Web (Hammonds, 1998). The misuse of credit card data for activities such as identity theft is a major concern (Stop thieves from stealing you, 2003; www.Truste.org/articles/holiday_shopping.php). A survey of 1068 consumers conducted by primary monitor Truste found that many people were skeptical of giving their personal information to online businesses. Seventy-five percent of respondents reported not liking to register at Web sites they visit. Fifteen percent of respondents refused to register at all. Forty-three percent of respondents stated they did not trust companies to not share their personal information. Of those respondents who did share personal information, 25% said they were less than impressed with the return on the information they provided (www.Truste.org/articles/quarterly_index1.php). A 2005 California HealthCare Foundation and the Health Privacy Project Poll found that 67% of national respondents are concerned about the privacy of their personal medical records (www.epic.org/privacy/medical/polls.html).

Dhillon and Moores (2001) reported that the selling of personal information by companies to third parties was the top privacy issue as identified by IS executives. However, failed Internet companies such as Boo.com, Toysmart.com, and CraftShop.com have either sold or have tried to sell customer data that may include phone numbers, credit card numbers, home address, and statistics on shopping habits, even though they had previously met Internet privacy monitor Truste's criteria for safeguarding customer information privacy. The rationale for the selling was to appease creditors (Sandoval, 2000). Even financially healthy companies realize there are advantages to be gained in the selling of collected customer information. Buyers include other businesses as well as the U.S. government. The Departments of Justice, State, and Homeland Security spend millions of dollars annually to buy commercial databases that track American citizens' finances, phone numbers, and biographical data. Often these data are accepted at face value without further evaluation for accuracy (Woellent & Kopecki, 2006). Companies such as Amazon, Ebay, and Google have opened up access to their databases to other companies for free in hopes these companies will develop new products and services that are organized around their database systems (Schonfeld, 2005).

In his excellent study on privacy in the information age, Cate (1997) adopted the definition of privacy as "the claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others" from Westin (1967, p. 7). Westin/Cate's definition is interesting because it allows for flexibility in discussing privacy within the context of the Internet. Whereas many people worry about divulging personal information electronically, other people seem more than willing to give it away, trading their personal information for personal benefits such as free shipping and coupons (Kuchinskas, 2000). Personalized service is the main benefit. A Web site can save a shopper time and money by storing and recalling a user's tastes and buying habits (Baig, Stepanek, & Gross, 1999). ISPs are willing to allow Web users cheaper access to the Internet provided the users are amenable to having their online behavior tracked for marketing purposes by specialized software (Angwin, 2000). If users are not amenable to having their personal information shared among company subsidiaries, at least one company

has warned that discounts for offerings such as high-speed Internet service will be taken away (Lazarus, 2004).

TECHNOLOGICAL CHALLENGES TO PRIVACY

Corporate and Government Databases

The practice of gathering personal information about customers and citizens by corporations and governments is well established. Software is available that is dedicated to analyzing data collected by company Web sites, direct-mail operations, customer service, retail stores, and field sales. Web analysis and marketing software enable Web companies to take data about customers stored in large databases and offer these customers merchandise based on past buying behavior, either actual or inferred. It also enables targeted marketing to individuals using e-mail. Governments routinely collect personal information from official records of births, deaths, marriages, divorces, property sales, business licenses, legal proceedings, and driving records. Many of the databases containing this information are going online (Bott, 2000).

These company and government databases are often not adequately secure against various threats to their integrity. Companies such as Choicepoint, Bank of America, Time Warner, Acxiom, and CardSystems Solutions have experienced breaches of security wherein unauthorized access of customer information such as names, addresses, Social Security numbers, credit and debit card numbers, and driver's license numbers occurred (Carrns, 2005; Perez & Brooks, 2005). Other companies blame checkout software that improperly stored credit card data for security breaches (Bank, 2005). For some companies, the threat to data security is internal rather than external (Yuan, 2005).

The deregulation of the financial services industry has made it possible for banks, insurance companies, and investment companies to begin working together to offer various financial products to consumers. Personal financial information that was kept separate before deregulation can now be aggregated. In fact the ability to mine customer data is one of the driving forces behind the creation of large financial conglomerates. Services can be offered to customers based on their information profiles. Large credit bureaus such as Equifax and Trans Union have traditionally been a source of information about a person's credit worthiness. Their databases contain information such as a person's age, address, and occupation. Credit bureaus have begun to sell personal information to retailers and other businesses (Big browser is watching you!, 2000). Some mutual fund companies have disclosed information such as customer name, home address, and account numbers on a U.S. government Web site in response to a Securities and Exchange Commission regulation, leaving customers vulnerable to identity theft and other crimes (Maremont, 2005).

Like personal financial information, medical information is for most people a very private matter. Despite this fact, there is a wealth of personal medical data in government and institutional databases. As *Consumer Reports* (Who knows your medical secrets, 2000, p. 23) notes, "[t]he federal government maintains electronic files of hundreds of millions of Medicare claims. And every state aggregates medical data on its inhabitants, including registries of births, deaths, immunizations, and communicable diseases. But most states go much further. Thirty-seven mandate collection of electronic records of every hospital discharge. Thirty-nine maintain registries of every newly diagnosed case of cancer. Most of these databases are available to any member of the public who asks for them and can operate the database software required to read and manipulate them."

Much of personal health information that is available to the public is volunteered by individuals themselves, by responding to 800 numbers, coupon offers, rebate offers, and Web site registration. Much of the information is included in commercial databases like Behavior-Bank sponsored by Experian, one of the world's largest direct-mail database companies. This information is sold to clients interested in categories of health problems, such as bladder control or high cholesterol.

E-Mail

In a survey of 840 U.S. companies, 60% said they use some type of software to monitor employees' e-mail activities (Tam, White, Wingfield, & Maher, 2005). Despite the fact that most companies had policies alerting employees that they were subject to monitoring, some had fired employees based on evidence collected during monitoring (Seglin, 2000). Hackers can also be a problem. Programs can be surreptitiously installed that monitor a user's keystrokes. The keystrokes can be sent across the Internet to a computer that logs everything that is typed for later use (Glass, 2000).

Employee's invasion of privacy claims have not been upheld in the United States courts, which argue that, since employers own the computer equipment, they can do whatever they want with it (McCarthy, 2000). However, use of Google's Gmail could present an information privacy issue of another sort. Gmail searches for certain words in a user's incoming messages, then displays text ads related to the words. There is a potential for many messages accumulated on Google's servers to be combined to produce a profile of an individual that could be accessed by, say, a government law enforcement agency or a criminal organization (Jesdanun, 2005; Wildstrom, 2004, 2006).

Wireless Communications

Wireless advertising poses a host of challenges for privacy advocates. Wireless service providers know customers' names, cell phone numbers, home and/or office addresses, and the location from where a customer is calling, as well as the number a customer is calling. Each phone has a unique identifier that can be used to record where in the physical world someone travels while using the cell phone (Petersen, 2000). The Federal Communications Commission requires cell phone service providers to be able to identify the location of a caller who dials 911, the emergency number. Since a cell phone service provider can track the location of a 911 call, it can track the location of any other call as well using global positioning system (GPS) technology. Mobile tracking firm Yora uses Nextel GPS phone software to create "geofences" technology that sets off an alarm at the office when field workers go to off-limit locations (Charny, 2004).

Wi-Fi networks provide ample opportunity for data breaches. Wi-Fi networks access points function as invitations to other wireless devices (such as laptop computers) to hook up and create a session. Because wireless networking is designed to be simple to install and easy to use, the various devices do not automatically distinguish between an authorized user and an unauthorized intruder (Bulkeley, 2004). Even with security systems using keys and encryption, Wi-Fi networks will remain vulnerable to sophisticated hackers (Crockett, 2004).

Radio frequency identification (RFID) systems use microchips (tags) that transmit/receive data to/from a reader device. Tags have been used by railroads and the U.S. Department of Defense to track inventory, in payment cards such as Mastercard PayPass, on individual items at Walmart, in library books, and in U.S. passports. Data collected from multiple tags can be aggregated in a database and associated with an individual using a tracking number (The end of privacy, 2006). RFID systems may also be a security challenge for companies as well. Corporate espionage involving tapping into warehouse readers or scanning goods as they leave a distribution point may constitute a serious security breach (Clayburn, 2004).

Clickstream Tracking

Tracking employee behavior on the Internet has become common practice. Software programs have been specifically designed to monitor when employees use the Internet and which sites they visit (McCarthy, 1999). Internet companies monitor Internet user behavior by a number of means, primarily to gather data about shopping and buying preferences with a view toward developing "user profiles." The primary means is the creation and use of cookies.

Cookies are text files created by a Web server and stored on a user's hard disk. A cookie is a set of fields that a user's computer and a server exchange during a transaction. The server may change or suppress the contents of a cookie it has created. When a user connects to a Web site, the browser checks the cookies on the hard drive. If a cookie matches the site's address, the browser uploads the cookie to the Web site. With the information contained in the cookie, the site can run programs that personalize site offerings and/or track the user's activity while online (www.cnil.fr/traces). Cookies come in two varieties: first party and third party. First party cookies are stored on a user's PC by the actual company site being visited. Third party cookies are stored by an outside company such as a Web analytic firm used by online retailers. A user is viewed as more inclined to trust a first-party cookie since it came from a known source (Kesmodel, 2005).

PC Software

Various kinds of PC software may pose a threat to a user's privacy. Usually the user has not explicitly asked for the software to be downloaded and is not aware the software is actually running on the PC.

Spyware installs itself surreptitiously on computers when a user downloads other programs, then tracks each click the user makes (Hagerty & Berman, 2003). A species of spyware called adware tracks Web surfing or online buying so marketers can send a user targeted ads (which are usually unsolicited). Other spyware may steal user passwords or credit card information (Gutner, 2004).

Phishing uses e-mail to direct unsuspecting users to legitimate-looking Web sites (such as a trusted bank) where the user is asked to provide personal information about themselves such as passwords and account numbers. Phishers team up with international criminal syndicates to fence stolen data (Kay, 2004).

Pharming has succeeded phishing. Users are redirected to an imposter Web page even though the user enters the correct address into a browser. The results may be the same as in phishing (Delaney, 2005).

FUTURE TRENDS

The issue of personal information privacy and the Internet continues to be debated within the community of Internet users. Privacy is a social issue, generally speaking. How the personal information privacy debate is ultimately resolved will be decided by the values inherent in a given society. Since the position of privacy advocates differs so markedly from the position of technology growth advocates, and since privacy issues have been addressed in court and precedents established in state and common law, it seems likely that the personal information privacy debate will be resolved in the world's legislatures and resulting laws enforced in the courts

(Lessig, 1999). The U.S. has seen the passage of the Health Insurance Portability and Accountability Act that prohibits unauthorized disclosures of personal medical information, punishable by a fine up to $250,000 and 10 years in jail (Lueck, 2003). The Identity Theft Penalty Enhancement Act prescribes stiff prison terms for people who use identity theft to commit other crimes (Kay, 2004). Other nations will have to adopt their own measures to ensure the privacy of their citizens' personal information. Since the Internet is a global technology, perhaps a global effort will be needed to adequately address these issues.

CONCLUSIONS

Various uses of e-technology that collect and/or disseminate personal information include corporate and government databases, e-mail, wireless communications, clickstream tracking, and PC software. Clearly the challenges to personal information privacy posed by the various forms of e-technology are not the result of the technology itself. Rather it is the uses of the technology that pose the threat to the integrity of personal information privacy. In particular, the surreptitious monitoring of user behavior without the user's consent and the possible misuse of the collected information pose the biggest threats. The various nations of the world must come to grip with these challenges in their own way, perhaps working in unison. Otherwise the personal information privacy of their citizens will be continuously at risk.

REFERENCES

Angwin, J. (2000, May 1). A plan to track Web use stirs privacy concern. *The Wall Street Journal*, p. B1f.

Baig, E. C., Stepanek, M., & Gross, N. (1999, April 5). Privacy. *Business Week*, pp. 84-90.

Bank, D. (2005, April 27). Stores blame checkout software for security breaches. *The Wall Street Journal*, p. B1f.

Big browser is watching you! (2000, May). *Consumer Reports*, pp. 43-50.

Bott, E. (2000). We know where you live, work, shop, bank, play … and so does everyone else! *PCComputing, 19*(5), 80-100.

Bulkeley, W. M. (2004, December 7). Wireless mischief. *The Wall Street Journal*, p. B1f.

Carrns, A. (2005, August 3). Trial highlights vulnerabilities of databases. *The Wall Street Journal*, p. B1f.

Cate, F. H. (1997). *Privacy in the information age*. Washington, DC: Brookings Institution Press.

Charny, B. (2004). *Big boss is watching*. Retrieved November 1, 2006, from news.com.com./Big+boss+is+watching/2100-1036_3-5379953.html

Clayburn, T. (2004). Watching out. Retrieved November 1, 2006, from www.informationweek.com/story/showArticle.jhtml?articleID=17603415

Crockett, R. O. (2004, January 19). For now, Wi-Fi is a hacker's delight. *Business Week*, p. 79.

Delaney, K. J. (2005, May 17). 'Evil twins' and 'pharming.' *The Wall Street Journal*, p. B1f.

Dhillon, G. S., & Moores, T. T. (2001). Internet privacy: Interpreting key issues. *Information Resources Management Journal, 14*(4), 33-37.

Glass, B. (2000, June 6). Keeping your private information private. *PC Magazine*, pp. 118-130.

Gutner, T. (2004, October 4). What's lurking in your PC? *Business Week*, pp. 108-109.

Hagerty, J. R., & Berman, D. K. (2003, August 27). New battleground in Web privacy war: Ads that snoop. *The Wall Street Journal*, p. A1f.

Hammonds, K. H. (1998, March 16). Online insecurity. *Business Week*, p. 102.

Jesdanun, A. (2005, July 18). Google's rapid expansion prompting privacy concerns. *The Buffalo News*, p. B8f.

Kay, R. (2004). Phishing. Retrieved November 1, 2006, from www.computerworld.com/security/story/0.10801,89096,00.html

Kesmodel, D. (2005, September 12). When cookies crumble. *The Wall Street Journal*, p. R6.

Kuchinskas, S. (2000, September). One-to-(N)one. *Business 2.0*, 141-148.

Lazarus, D. (2004). *Privacy is going to cost you.* Retrieved November 1, 2006, from www.sfgate.com/cgibin/article.cgi?files/c/a/2004/11/26/BUGHDA11CU1.DTL

Lessig, L. (1999). *Code and other laws of cyberspace*. New York: Basic Books.

Lueck, S. (2003, March 19). Tough new law helps to guard patient privacy. *The Wall Street Journal*, p. D1f.

Maremont, M. (2005, March 23). New privacy leak: Some mutual funds reveal clients' data. *The Wall Street Journal*, p. A1f.

McCarthy, M. J. (1999, October 21). Now the boss knows where you're clicking. *The Wall Street Journal*, p. B1f.

McCarthy, M. J. (2000, April 25). Your manager's policy on employee's e-mail may have a weak spot. *The Wall Street Journal*, p. A1f.

Perez, E., & Brooks, R. (2005, May 3). For big vendor of personal data, a theft lays bare the downside. *The Wall Street Journal*, p. A1f.

Petersen, A. (2000, July 24). Coming to phone screens: Pitches, privacy woes. *The Wall Street Journal*, p. B1f.

Sandoval, G. (2000, July 1). Sensitive data on customers being sold by failed e-retailers. *The Buffalo News*, p. A9.

Schonfeld, E. (2005, April). The great giveaway. *Business 2.0*, April, 81-84.

Seglin, J. L. (2000, August). Who's snooping on you? *Business 2.0*, August, 200-203.

Stop thieves from stealing you. (2003, October). *Consumer Reports*, pp. 12-17.

Szewczak, E. J. (2004). Personal information privacy and EC: A security conundrum? In M. Khosrow-Pour (Ed.), *E-commerce security: Advice from experts* (pp. 88-97). Hershey, PA: Idea Group Publishing.

Tam, P., White, E., Wingfield, N., & Maher, K. (2005, March 9). Snooping e-mail by software is now a workplace norm. *The Wall Street Journal*, p. B1f.

The end of privacy? (2006, June). *Consumer Reports*, pp. 33-39.

Westin, A. F. (1967). *Privacy and freedom*. New York: Atheneum.

Who knows your medical secrets. (2000, August). *Consumer Reports*, pp. 22-26.

Wildstrom, S. H. (2004, May 3). Google's Gmail is great—But not for privacy. *Business Week*, p. 30.

Wildstrom, S. H. (2006, February 20). Your data, naked on the net. *Business Week*, p. 24.

Woellent, L., & Kopecki, D. (2006, May 29). The snooping goes beyond phone calls. *Business Week*, p. 38.

Yuan, L. (2005, June 1). Companies face data breaches from inside, too. *The Wall Street Journal*, p. B1f.

KEY TERMS

Clickstream Tracking: The use of software to monitor when people use the Internet and what sites they visit.

Cookies: Text files created by a Web server and stored on a user's hard disk that contain data about which Web sites have been visited.

Global Positioning System (GPS): A satellite-based data system that works with a computer chip embedded in a cell phone to identify the location of the user anywhere in the world.

Identity Theft: The stealing and use of a person's identity through the acquisition of personal information without that person's knowledge or permission.

Personal Information: Information about, or peculiar to, a certain person or individual.

Pharming: Users are redirected to an imposter Web page, even though the user enters the correct URL into a browser (see **Phishing**).

Phishing: A user's e-mail program is used to direct the user to a legitimate-looking Web site where the user is asked to provide personal information about himself or herself such as passwords and account numbers.

Privacy: The claim of individuals, groups, or institutions to determine for themselves when, how, and to what extent information about them is communicated to others (Westin, 1967).

Radio Frequency Identification (RFID) System: Microchips (tags) are used to wirelessly transmit/receive data to/from a reader device.

Spyware: Software that installs itself on computers when programs are downloaded and that tracks each user click, usually without the user's knowledge or permission.

Ethical Issues in Conducting Online Research

Lynne D. Roberts
University of Western Australia, Australia

Leigh M. Smith
Curtin University of Technology, Australia

Clare M. Pollock
Curtin University of Technology, Australia

INTRODUCTION

The rapid growth of the Internet has been accompanied by a growth in the number and types of virtual environments supporting computer-mediated communication. This was soon followed by interest in using these virtual environments for research purposes: the recruitment of research participants, the conduct of research, and the study of virtual environments. Early research using virtual environments raised a number of ethical issues and debates. As early as 1996 a forum in the *The Information Society* (volume 12, issue 2) was devoted to ethical issues in conducting social science research online. The debate has continued with more recent collaborative attempts to develop guidelines for ethical research online (Ess & AoIR ethics working committee, 2002; Frankel & Siang, 1999). In this article we explore contemporary ethical issues associated with conducting research online.

BACKGROUND

The basic principles of ethical research with humans are integrity, respect, beneficence, and justice (National Health & Medical Research Council, 2006). Based on these principles many professional associations provide ethical guidelines, or codes, for the conduct of research. Guidelines and legislation vary across disciplines and across countries. However, these codes have typically been developed for use in off-line settings, prior to consideration of research being conducted online[1]. While these codes contain guiding principles for research generally, the translation of these principles into actions for conducting research in virtual environments is open to interpretation. The process of translating ethical guidelines into ethical practice online involves a deliberation of the options available to the researcher and the likely impact on research participants, their communities, and the research process. Central concerns in this process are maintaining respect for individuals, their online identities, and the ownership of words.

PUBLIC VS. PRIVATE SPACE

Research online can take place within a range of virtual environments that vary in terms of purpose, synchronicity, access, number of users, and norms. A major issue in developing ethical research procedures for use within a particular virtual environment is determining whether the setting represents a private or public "space." Various attempts have been made to distinguish between the public and the private in virtual environments (see, e.g., Lessig, 1995), but little agreement has been reached. There are currently no clear guidelines for researchers on what constitutes private vs. public space in virtual environments, yet the distinction is important, as it affects the rights of participants to be advised of the research and to give or withhold their informed consent.

The defining of public vs. private space cannot be reduced to the single dimension of accessibility to the virtual environment. Interactions that occur within publicly accessible virtual environments may be perceived by participants to be private. Newsgroups can be accessed without restriction, yet newsgroup postings can be, and frequently are, high in self-disclosure and are perceived by many users to be private (Witmer, 1997). Similarly, support groups on sensitive issues may be conducted in publicly accessible sites with participants adhering to norms of confidentiality and privacy (Elgesem, 2002).

Some ethical codes exempt naturalistic observations and archival research from requiring informed consent where no harm or distress is likely to come to those researched and where their confidentiality is protected. It has been argued that the decision not to inform members of online groups about research conducted on the group has the advantage of the research being "unobtrusive" (Langer & Beckman, 2005). Others, while acknowledging the benefits of naturalistic observation, regard this approach as placing researchers in a position "little better than spies" (Bakardjieva & Feenberg, 2001, p. 234). King (1996) highlighted the potential for psychological harm to members of online groups where research is conducted and published without the prior knowledge and informed consent of participants.

Where there has been the expectation of privacy within a group (however misinformed that expectation may be) the individual may feel violated upon hearing of, or reading, the results of that research.

Where the presumption is made that online communication occurs in public space simply because it is accessible without restriction, an anomaly may result in how research participants are treated in equivalent settings in online and off-line research. For example, research on support groups off-line requires the informed consent of research participants, while similar research online may occur without the knowledge or informed consent of the participants, on the grounds that all postings are public documents (see, e.g., Salem, Bogat, & Reid's 1997 study of a depression support group). Despite the inequities this raises, in a recent review of psychological research conducted online (Skitka & Sargis, 2006) it was noted that "most Institutional Review Boards are concluding that online postings represent the public domain and that researchers do not need to obtain informed consent to use this material" (p. 549).

Table 1 summarizes possible dimensions against which the public/private nature of a virtual environment can be assessed. Virtual environments where all dimensions fall on the left-hand side of the continua may be deemed as public environments for research purposes and subject to guidelines for research in public settings. We recommend virtual environments where all dimensions are on the right be deemed private environments, requiring informed consent from research participants. The difficulty arises with the majority of settings that do not fall clearly into public or private spaces. Researchers do not have the right to define virtual environments as public or private to meet their own research needs (Waskul & Douglass, 1996). Rather, account should be taken of the size and nature of the online forum and the intrusiveness of the study. Consideration should be made of the likely effect of the request to conduct research and the research itself on research participants and their communities. The process of requesting consent to research may in itself alter group dynamics (Sixsmith & Murray, 2001).

INFORMED CONSENT

Research conducted in virtual environments that have been conceptualized as private settings requires the informed consent of research participants. Obtaining informed consent in virtual environments is more problematic than in off-line research, as participants are frequently geographically dispersed. In addition, research participants may be reluctant to divulge details of off-line identities required for the signing of consent forms. Further, it is difficult to verify factors that may affect an individual's ability to provide informed consent, such as age, mental competency, and comprehension of risk (Skitka & Sargis, 2006).

A range of options has been suggested for obtaining informed consent in online research (Bruckman, 1997; Flicker, Haans, & Skinner, 2004; Jacobson, 1999; Kralik, Warren, Price, Koch, & Pignone, 2005; Roberts, Smith, & Pollock, 2004; Smith & Leigh, 1997), and these have been summarized in Table 2. Selection of a method for obtaining informed consent will necessarily be dependent upon the type of virtual environment, the level of anonymity required by research participants, and their access to high-level computing facilities. Regardless of the method used, the information about the research should be presented in a format that the research participants can keep and refer back to at any time before, during, or after their research participation.

Table 1. Dimensions of public and private space in virtual environments

Accessibility:	Accessible to all	→	Restricted membership
Users' perceptions:	Public	→	Private
Community statement:	Research permitted	→	Research prohibited
Topic sensitivity:	Low	→	High
Permanency of records:	Public archives	→	Private logs only

Table 2. Methods of obtaining informed consent in online research

	Format of information	***How consent obtained***
Signed consent:	Hard copy or electronic	Post, fax, or e-mail
Implied consent:	Electronic	Gateway WWW page E-mail consent Logging of consent Use of password protected site

Table 3. Factors that decrease the anonymity afforded by pseudonyms

• Use of name, derivation of name, or nickname
• Use of same pseudonym across virtual environments with differing requirements for identification
• Self-disclosure
• Active seeking of identifying information by others

Table 4. Levels of anonymity (site, pseudonym, and quotations)

• Identify site, use online pseudonym, and directly quote
• Identify site, use pseudonym of online pseudonym, and directly quote
• Identify site, use pseudonym of online pseudonym, and paraphrase
• Do not identify site, use online pseudonym, and directly quote
• Do not identify site, use pseudonym of online pseudonym, and directly quote
• Do not identify site, use pseudonym of online pseudonym, and paraphrase

Skitka and Sargis (2006) reviewed Internet-based psychological research conducted over a two-year period. In almost two thirds (62.5%) of studies reviewed, the researchers clearly obtained informed consent from research participants; 12.5% did not obtain informed consent, and it was unclear in the remaining 25% whether informed consent was obtained or not. Of concern, deception was used in 18% of studies, with less than half indicating that debriefing had been provided to research participants.

Consideration should also be given to seeking the cooperation of community gatekeepers[2] and advising the community of the research being undertaken. Advising communities of a research project requires the public identification of the researcher. In some circumstances, the decision to research within a particular virtual environment may be made after the researcher has been either an active participant or "lurker" within that environment. We recommend that researchers make their researcher status overt as soon as the research process begins. This may include identifying as a researcher in pseudonyms (Roberts et al., 2004), descriptions (Allen, 1996), or objects (Reid, 1996); linking between research and social identities (Roberts et al., 2004); and posting information about the research.

Advising communities of a research project may take ongoing effort in public virtual environments without membership boundaries. Identifying oneself as a researcher once within an online group does not mean that absent or future members of the group are also informed of the researcher's role (Sixsmith & Murray, 2001). There may be a need to re-identify researcher status and restate and clarify the role of the researcher on an ongoing basis.

PROTECTING ANONYMITY AND CONFIDENTIALITY

Individuals typically adopt a pseudonym (or pseudonyms) for use in virtual environments, providing a level of anonymity. Some studies include pseudonyms in reports and publications on the basis that they provide information relevant to the study (Langer & Beckman, 2005). While it has been argued that research involving pseudonymous characters is exempt from regulations governing human subjects as "true" (off-line) identities are not known (Jacobson, 1999), there are often links between the individual and their pseudonyms that decrease the level of anonymity a pseudonym provides (Allen, 1996; Bruckman, 2002; Jacobson, 1996). These are presented in Table 3. The combination of these factors means that researchers cannot assume that pseudonyms provide adequate protection for off-line identities. The degree of anonymity conferred in virtual environments does not reduce the ethical requirements for researchers to protect the anonymity of research participants and virtual interaction settings (Waskul & Douglass, 1996).

Protecting the anonymity[3] of the individual extends to protecting the anonymity of their pseudonym(s), as representations of the individual online. Pseudonyms themselves gain reputations over time (Bruckman, 2002). Researchers can provide varying levels of protection to research participants' anonymity (see Table 4). The practice of replacing existing pseudonyms with other pseudonyms in research materials confers little additional protection to the existing pseudonym when text searches can be used to identify source documents (Allen, 1996) or where the individual has a distinctive, recognisable writing style (Markham, 2005).

Further, other community members may seek to identify disguised identities and may share this information with others (Bruckman, 2002).

In addition to protecting the anonymity of research participants in research reports, the data collected need to be kept secure in order to protect confidentiality. Maintaining the security of data collected in computer-mediated research poses unique difficulties. Confidentiality may be breached at the site of data collection, during transmission of data, or in the storage of data. Sites at which data is collected may not be secure and may be subject to surveillance by gatekeepers, "hackers" (Rhodes, Bowie, & Hergenrather, 2003), other computer users, or others physically present in the location at the time (Kralik et al., 2005). Confidentiality of data may be breached during data transmission where another party intercepts data (Nosek, Banaji, & Greenwald, 2002). This may include the Internet service provider of the research participant or researcher. Employers may also monitor employees' e-mail (Sipior & Ward, 1995; Weisband & Reinig, 1995). Confidentiality of data may be breached during storage by hackers, employers, or as a result of "open records" legislation (Pittenger, 2003).

Online researchers need to provide the most secure forms of data collection, transmission, and storage possible, aiming to minimize the risks of unauthorized persons gaining access to research data at any stage of the research process. The procedures used to ensure this will differ according to the type of virtual media used. Using a "perspective of reasonableness," Kralik et al. (2005) argue that the chances of a breach of security are small and need to be weighed against possible gains from the research.

OWNERSHIP OF WORDS

The ownership of electronic messages has been contested. It is still unclear whether the individual who authored a message, the community to which it was sent, or anyone who has access to the message is the owner of the electronic message. Electronic postings may be considered original works protected by copyright, although this has not yet been legally tested (Sixsmith & Murray, 2001). If informed consent is not obtained to use electronic messages, copyright provisions suggest that they are subject to "fair dealing" for research purposes, and should be attributed to the author (Australian Copyright Council, 2001). Researchers who neither obtain informed consent, nor reference the material they are quoting, risk violating both ethical and copyright standards. With the consent of the research participant, quotes in research may be attributed to the individual, their online pseudonym, or used anonymously. Respect for research participants is demonstrated through asking, and abiding by, their preferences for anonymity, pseudonymity, or identification. However, this is not without its problems. Individual preferences for identification may not be consistent with the norms or wishes of the community. There can be tensions between respecting copyright entitlements of individuals and protecting the privacy of other participants within a virtual environment. Roberts et al. (2004) highlighted the potential for negative impacts on the privacy of other virtual environment members when one research participant's work is fully attributed, including information on the virtual environment.

RETURNING RESEARCH FINDINGS TO THE COMMUNITY

A requirement in some ethical codes is to provide research participants with information about the outcome of the research. In addition to being a requirement, this can demonstrate respect for the individuals who participated in the research. A summary of research findings can be provided to research participants in hardcopy or electronic format. Where research participants are reluctant to provide contact information that may link their online and off-line identities, the summary can be placed on a Web site or sent through the messaging system of the virtual community (Roberts et al., 2004).

FUTURE TRENDS

Rapidly changing technologies will result in the development of an increasing range of virtual environments and tools that may be used for research purposes. The precise characteristics of these new virtual environments and tools may vary greatly from those available today. Concern has already been raised over the potential for participant harm when non-human agents used in the research process evoke negative responses or unwanted arousal (Palomares & Flanagin, 2005). Privacy issues have been raised in relation to Web-cams, which have the potential to capture data from both consenting research participants and other non-consenting individuals in their proximity (Palomares & Flanagin, 2005).

Decisions made regarding research methodologies and designs in virtual environments need to be embedded in the basic principles of ethical research and may require multi-layered considerations. Before conducting research within each new type of environment or with a new type of research tool, researchers will need to address the intrusiveness of the proposed research, the perceived privacy of the research setting, the vulnerability of the community, the potential for harm to individuals and/or the community, and how confidentiality will be maintained and intellectual property rights respected in their research proposals (Eysenbach & Till, 2001). This requires a consideration of the likely impacts of the research on both research participants and the

communities in which the research is conducted. It should be guided by researchers' knowledge and adherence to the "Netiquette" and social norms of the virtual environments concerned. Guided by the principles outlined in their ethical codes, researchers will need to develop ethically defensible strategies that balance the needs of research participants and their online communities, offering protection to both, while providing the validity assurances required by the readers/users of the research.

CONCLUSION

Our approach to the conduct of ethical research in virtual environments is based on a human research perspective, explicitly recognizing that communication online is conducted by individuals who interact via their online identities. Our focus is therefore on individuals rather than texts. We privilege the rights of individuals to make informed consent about whether or not traces of their interaction online (e.g., logs or postings) can be used for research purposes. We believe this approach is consistent with general ethical guidelines for human research in the social sciences. Alternative perspectives to the conduct of ethical research in virtual environments place a stronger emphasis on the cultural production of texts and performance (Bassett & O'Riordan, 2002; White, 2002) reflecting calls for an ethical pluralism in Internet research that recognizes a range of ethical perspectives as legitimate (Ess, 2002).

Regardless of the perspective adopted, all research should comply with the principles of ethical research as outlined in the relevant professional association's code of ethics or by Institutional Review Boards. In the absence of specific guidelines for online research and where review committees are unfamiliar with online research issues (Keller & Lee, 2003), we recommend that researchers be guided by the principles outlined in their code, adapting the guidelines for use in virtual environments as necessary. In addition to protecting the rights of research participations, researchers have a social responsibility to ensure online research is conducted in an ethically defensible manner to maintain the viability of virtual environments as research media (Colvin & Lanigan, 2005).

REFERENCES

Allen, C. L. (1996). What's wrong with the "golden rule"? Conundrums of conducting ethical research in cyberspace. *The Information Society, 12*(2), 175-187.

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Retrieved October 29, 2003, from http://www.apa.org/ethics/code2002.html

Australian Copyright Council. (2001). *Information sheet G53. Copying for research or study.* Retrieved October 22, 2003, from http://www.copyright.org.au/

Bakardjieva, M., & Feenberg, A. (2001). Involving the virtual subject. *Ethics and Information Technology, 2*(4), 233-240.

Bassett, E. H., & O'Riordan, K. (2002). Ethics of Internet research: Contesting the human subjects research model. *Ethics and Information Technology, 4*(3), 233-247.

Bruckman, A. (2002). Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology, 4*(3), 217-231.

Bruckman, A. S. (1997). MOOSE crossing: Construction, community and learning in networked virtual world for kids (Doctoral dissertation, MIT Media Lab). *Dissertation Abstracts International, DAI-A 58/11*, 4241.

Colvin, J., & Lanigan, J. (2005). Ethical issues and best practice considerations for Internet research. *Journal of Family and Consumer Sciences, 97*(3), 34-39.

Elgesem, D. (2002). What is special about the ethical issues in online research? *Ethics and Information Technology, 4*(3), 195-203.

Ess, C. (2002). Introduction. *Ethics and Information Technology, 4*(3), 177-188.

Ess, C., & AoIR ethics working committee. (2002). *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee.* Retrieved October 22, 2003, from http://www.aoir.org/reports/ethics.pdf

Eysenbach, G., & Till, J. E. (2001). Ethical issues in qualitative research on internet communities. *British Medical Journal, 323*(1), 1103-1105.

Flicker, S., Haans, D., & Skinner, H. (2004). Ethical dilemmas in research on Internet communities. *Qualitative Health Research, 14*(1), 124-134.

Frankel, M. S., & Siang, S. (1999). *Ethical and legal aspects of human subjects research on the Internet: A report of a workshop June 10-11, 1999*. Retrieved October 22, 2003, from http://www.aaas.org/spp/dspp/sfrl/projects/intres/main.htm

Jacobson, D. (1996). Contexts and cues in cyberspace: The pragmatics of names in text-based virtual realities. *Journal of Anthropological Research, 52*(4), 461-479.

Jacobson, D. (1999). Doing research in cyberspace. *Field Methods, 11*(2), 127-145.

Keller, H. E, & Lee, S. (2003). Ethical issues surrounding human participants research using the Internet. *Ethics & Behavior, 13*(3), 211-219.

King, S. (1996). Researching Internet communities: Proposed ethical guidelines for the reporting of the results. *The Information Society, 12*(2), 119-127.

Kralik, D., Warren, J., Price, K., Koch, T., & Pignone, G. (2005). Methodological issues in nursing research: The ethics of research using electronic mail discussion groups. *Journal of Advanced Nursing, 52*(5), 537-545.

Langer, R., & Beckman, S. X. (2005). Sensitive research topics: Netnography revisited. *Qualitative Market Research, 8*(2), 189-203.

Lessig, L. (1995). The path of cyberlaw. *Yale Law Journal, 104*(7), 1743-1755.

Markham, A. N. (2005). The methods, politics, and ethics of representation in online ethnography. In N. K. Denzin & Y. S. Lincoln (Eds.), *The Sage handbook of qualitative research* (3rd ed.) (pp. 793-820). Thousand Oaks, CA: Sage Publications Ltd.

National Health and Medical Research Council. (2006). *National statement on ethical conduct in research involving humans (revised)*. Commonwealth of Australia. Retrieved November 17, 2006, from http://www.nhmrc.gov.au/publications/_files/e35.pdf

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the Internet. *Journal of Social Issues, 58*(1), 161-176.

Palomares, N. A., & Flanagin, A. J. (2005). The potential of electronic communication and information technologies as research tools: Promises and perils for the future of communication research. In P. J. Kalbfleisch (Ed.), *Communication yearbook 29* (pp. 147-185). Mahwah, NJ: Lawrence Earlbaum Associates.

Pittenger, D. J. (2003). Internet research: An opportunity to revisit classic ethical problems in behavioral research. *Ethics & Behavior, 13*(1), 45-60.

Reid, E. (1996). Informed consent in the study of online communities: A reflection on the effects of computer-mediated social research. *The Information Society, 12*(2), 169-174.

Rhodes, S. D., Bowie, D. A., & Hergenrather, K. C. (2003). Collecting behavioural data using the World Wide Web: Considerations for researchers. *Journal of Epidemiology and Community Health, 57*(1), 68-73.

Roberts, L. D., Smith, L. M., & Pollock, C. M. (2004). Conducting ethical research online: Respect for individuals, identities, and the ownership of words. In E. Buchanan (Ed.), *Readings in virtual research ethics: Issues and controversies* (pp. 159-176). Hershey, PA: Idea Group Inc.

Salem, D. A., Bogat, G. A., & Reid, C. (1997). Mutual help goes online. *Journal of Community Psychology, 25*(2), 189-207.

Sipior, J. C., & Ward, B. T. (1995). The ethical and legal quandary of e-mail privacy. *Communications of the ACM, 38*(12), 48-54.

Sixsmith, J., & Murray, C. D. (2001). Ethical issues in the documentary data analysis of Internet posts and archives. *Qualitative Health Research, 11*(3), 423-432.

Skitka, L. J., & Sargis, E. G. (2006). The Internet as psychological laboratory. *Annual Review of Psychology, 57*(1), 529-555.

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers, 29*(4), 496-505.

Waskul, D., & Douglass, M. (1996). Considering the electronic participant: Some polemical observations on the ethics of online research. *The Information Society, 12*(2), 129-139.

Weisband, S. P., & Reinig, B. A. (1995). Managing user perceptions of email privacy. *Communications of the ACM, 38*(12), 40-47.

White, M. (2002). Representations or people? *Ethics and Information Technology, 4*(3), 249-266.

Witmer, D. F. (1997). Risky business: Why people feel safe in sexually explicit online communication. *JCMC, 2*(4). Retrieved March 19, 1997, from http:// jcmc.huji.ac.il/vol2/issue4/witmer2.html

KEY TERMS

Computer-Mediated Communication: Communication between two or more individuals that occurs via computer networks. Computer-mediated communication may be text, audio, graphics, or video based and occur synchronously (in "real time") or asynchronously (delayed).

Informed Consent: An individual's freely given consent to participate in research based on information provided by the researcher(s) about the research, possible risks associated with the research, and the voluntary nature of participation.

Informed consent must be obtained without coercion or undue influence.

Netiquette: The etiquette, or social rules, associated with communicating online. Netiquette may vary across virtual environments.

Private Space: Off-line, private space refers to geographical areas that are not for general or public use (e.g., your home). Online, the term private space is commonly used to refer to virtual environments, or parts of virtual environments that have restrictions on who may access them.

Pseudonym: The fictitious name adopted for use within a virtual environment. An individual may consistently use the same pseudonym or adopt several pseudonyms for use within and between virtual environments.

Public Space: Off-line, public space refers to geographical areas that are accessible to the general public (e.g., streets). Online, the term public space is commonly used to refer to virtual environments that do not have restrictions on access.

Virtual Identity: Representation of the individual in a virtual environment. The form of representation varies across virtual environments and may range from a pseudonym only (Internet relay chat), a pseudonym combined with a character description (multi-user dimensions), through to graphical representations, or avatars, in graphics-based environments. An individual may have multiple virtual identities.

ENDNOTES

1 Over time as codes are updated, consideration of research conducted online may be included. See, for example, the revised *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association, 2002).

2 Gatekeepers can ease researchers' access to a virtual community. While gatekeepers can influence the ease of communicating with community members, ultimately each community member has the right to choose whether or not they will participate in research, irrespective of gatekeepers' views.

3 We note, however, that the professional responsibility of the researcher sometimes overrides that of anonymity and confidentiality (e.g., reportable offences or subpoenaed research records).

Ethics of New Technologies

Joe Gilbert
University of Nevada Las Vegas, USA

INTRODUCTION

Information processing has been done through telling stories, drawing on cave walls, writing on parchment, printing books, talking on telephones, sending messages via telegraphs, broadcasting on radio and television, processing data in computers, and now by instantaneous network dissemination. Since the mid-1990's, personal computers have been the instrument of choice for sending and receiving information, and for processing much of it. The technology is the latest in a long series, but social issues involved have not really changed. Issues of content (is it true? obscene?), ownership (whose picture/text/idea? whose parchment/telephone system/computer?), and impact (anti-government, anti-social, harmful to children) appear today just as they did hundreds or thousands of years ago.

BACKGROUND

New technologies enable people to do new things (send 20 copies of a memo at once) or to do old things in new ways, such as storing files (Freeman & Soete, 1997). Improvements in technology that are incremental do not usually introduce major social issues, but radical innovations frequently present new kinds of social opportunities and threats (Brown, 1997). Ethics is the branch of philosophy that studies interpersonal or social values and the rules of conduct that follow from them. Ethics deals with questions of how people should treat each other on a basic level (Berlin, 2000). It considers such issues as rights and duties and fairness or justice. Because ethics concerns itself with fundamental rules, its applications to specific new technologies might require both knowledge of the new technology and reasoning about its possible applications based on established principles of ethics (Burn & Loch, 2001; Halbert & Ingulli, 2002).

Philosophers have pondered and written about issues of ethics for thousands of years. Some of their writings on this subject continue to be read and debated generation after generation (LaFollette, 2000). Three basic approaches have been most common and most accepted in discussions of ethics.

- Utilitarianism maintains that the ethical act is the one that creates the greatest good for the greatest number of people.
- Rights and duties maintain that the ethical act is the one that acknowledges the rights of others and the duties which those rights impose on the actor.
- Fairness and justice hold that the ethical act is the one that treats similarly situated people in similar ways with regard to both process and outcome.

ETHICS AND TECHNOLOGY

John Stuart Mill and Jeremy Bentham are the two philosophers most closely associated with utilitarianism. This view of ethics puts a high value on results, and holds that we must consider whether and to what degree our actions will bring pain or pleasure not only to ourselves but to all others who will be impacted by what we do (Frey, 2000; Mill & Bentham, 1987). A utilitarian would argue that the harm done to many individuals and businesses by viruses and worms far outweighs any happiness brought to their authors, and thus creating and disseminating such code is unethical. Similarly, a utilitarian analysis of music file-sharing would consider whether widespread free file-sharing might result in composers and artists deciding that it is not in their financial interest to continue writing and performing music. If this result occurred, not only the composers and artists but also their listeners would end up suffering harm that might outweigh the good that they enjoy from free file-sharing. Finally, a utilitarian analysis would favor products and policies that increase the spread of computer literacy and availability, since the Internet can bring great good to its users and computer literacy and availability makes such use possible.

Many philosophers have written about rights and duties (Sumner, 1987). The basic idea of this approach is that individuals do have rights, and that these rights are, practically speaking, worthless unless someone or some group has a corresponding duty. Thus, if I have a right to privacy, you have a duty not to monitor my every move (Kelly & Rowland, 2000). There are four basic sources of rights, and we will consider each in turn.

Human rights are possessed by every human, simply by virtue of being human. Among these rights are the right to live (not to be randomly killed), to be told the truth, to own property, and to basic dignity (Ignatieff, 2001). Among these, the one that most often causes confusion is the right to be

told the truth. Humans could not interact with each other in any meaningful way if lying and truth-telling were equally valid. Promises, contracts, and interpersonal relations all depend on the fact that the default setting for conversation is truth-telling. This does not mean that everyone always will or even should tell the whole truth all the time. It does, however, mean that we can and do start with an assumption of truth-telling (Bok, 1999). A right to property, whether physical or intellectual, means that others have a duty not to take or use my goods without compensating me.

Since property rights are human, they apply whether a given country's laws regarding such things as copyright and intellectual property are specific on a given issue or not. Music companies and movie studios, on behalf of individual artists, have a right to control and charge for distribution of their products. This right imposes a duty on individuals not to take such property without paying for it and recognizing the terms of distribution. Similarly, software companies have a right to charge for and control the distribution of their intellectual property. They paid programmers to develop a software product; others have a duty to respect the rights to this intellectual property.

Some rights are given to individuals by law. These citizen rights come by reason of membership in a community (nation, state, county, etc.). The right of citizens of the United States to free speech is not recognized by some other countries. Typically, dictatorships grant few citizen rights to those under their rule. These rights often coincide with human rights (right to live, to property, etc.) but frequently go beyond basic human rights. Copyright, as it exists in the United States, is not recognized equally in all countries. This is why it is important that the basic right to own property is a human right—it is valid whatever the laws of a particular jurisdiction.

A third source of rights is position. Policemen may apprehend and incarcerate suspected criminals. CEO's can speak for their companies on many issues. Purchasing agents can spend a company's money on goods or services within some limits. Managers can set rules for computer usage at work. People have these rights not just because they are human, or because they are particularly wise or knowledgeable, but because of the position they occupy. Since individuals have these rights, others have duties to respect the rights and follow their direction.

The fourth and final basic source of rights is by contract. Individuals or organizations can agree to contractual relations that create rights and impose duties that would not otherwise exist. If I agree to pay a certain amount of money each month in order to use an online service, I have a duty to pay and the service provider has a duty to make the service available to me under the terms of the contract.

The third basic approach to ethics is fairness and justice: it is ethical to be fair and unethical to be unfair. It is not fair that some individuals should purchase software and others obtain it free through sharing or piracy. It is fair for those who invest time, talent and money in producing software to be paid for the products resulting from their efforts and investments by all of those who use them, not just by some. Issues of fairness sometimes arise in the area of using computer technology for purposes of employee monitoring (Alder, 1998). In general terms, fairness involves treating similarly situated people in similar ways with regard to both process and outcome. However, justice is sometimes defined as equality, and at other times, as based on contribution, on needs and abilities, or on maximum freedom (Velasquez, 2002).

An issue that often arises in considering fairness and justice is the question of which individuals or groups are similarly situated. In the sense that all who access the Internet can view unrestricted sites, all who access the Internet are similarly situated. In the sense that some who access the Internet may choose to view pornography and others may choose not to (even inadvertently), we have at least two groups that are not similarly situated. Using this approach, one might argue against unrestricted availability of pornography on the Internet, but in favor of restricted access to Internet pornography. All who receive e-mail might be viewed as similarly situated. Spam reaching all e-mail accounts thus reaches similarly situated people. However, if most individuals who receive e-mail do not wish to receive spam, then this group (the unwilling) might be seen as not similarly situated with those who do wish to receive it. Such an argument could serve as the basis for something like an e-mail equivalent of the do-not-call list recently introduced for telemarketing.

The different approaches to ethics often produce the same result. If we consider the issue of hacking or gaining unauthorized access to another's system, utilitarianism concludes that more harm than good results from this activity. Those whose system is wrongfully accessed are faced with revising controls, checking to see what harm if any has been done, and correcting any problems caused by the hacking. Only the hacker gains. Those who have created or purchased the system have a right to limit access; the hacker has a duty to respect this right. It is not fair or just that some people go through the appropriate authorization to access or use a system while others hack into it. Thus from all three perspectives, hacking as defined can be judged unethical. If one does not accept the basic premises of the prevailing capitalist system, however, a defense of hacking can be devised (Halbert, 1994).

When the three approaches provide different results, rights and duties usually prevails as a way of determining whether an act is ethical, because rights are so basic. However, this is not always the case. In American copyright law, there is a "fair use" provision that allows an individual to make one copy for personal use of a copyrighted article without obtaining permission of the copyright holder. Whether this copy is made from a printed article on a photocopy machine

or downloaded from a computer, the same basic principle applies (Halbert & Ingulli, 2003). A utilitarian analysis suggests that this provision allows the greatest good for the greatest number of people, because single-copy permissions for personal use would involve excessive transaction costs to both users and copyright owners. One could argue from a rights and duties perspective that the copyright owner has a right (and the user a corresponding duty) to payment for each and every copy made of the material, even for a single copy. Fairness and justice would suggest that the "fair use" exception is ethically acceptable as long as each owner and user play by the same rules.

FUTURE TRENDS

Each of the three views can be used to analyze issues regarding new information technology and its applications (Gilbert, 2001). Because information technology has developed so quickly in the last quarter century, and appears to be poised for continued rapid development, a good deal of ethical analysis is and will be needed of specific questions and issues. The basic approaches are clear from a long history of ethical theory. The specific issues are and will be fresh, but those concerned with using information technology ethically can find answers to their questions. The collecting, storing, transmitting and analysis of information will continue to be central to both commerce and society. Issues of ownership, access, privacy and social impact will continue to concern individuals and society.

It seems safe to assume that laws will evolve and court decisions will help to illuminate legal issues involved with information processing technology. Individual managers, technicians and citizens will ponder and debate the ethics of various uses of this technology. Basic philosophical approaches will remain the same; their application to individual situations will continue to require that thoughtful individuals work through the journey from abstract principles to particular applications and decisions.

CONCLUSION

Much of the discussion of right and wrong concerning information technology is based either on personal opinion or on legal interpretations of such topics as intellectual property rights and individual privacy. Philosophers have thought about, discussed, and written about right and wrong for thousands of years. While the history of philosophy has unfolded before the invention of electronic information technology, issues such as privacy, property ownership, truthfulness, and government intrusion on individual liberties have been the subjects of ethical inquiries for well over two thousand years. In attempting to cope with social issues raised by new technologies, the best thoughts of many generations of humans can be usefully brought to bear on current controversies. Doing this requires both knowledge of the discipline of ethics and at least some knowledge of new technologies and their social impacts. The application of such knowledge connects us with our ancestors who wrote on cave walls, told stories over campfires, printed books, and used the other means of processing information described at the beginning of this article.

REFERENCES

Alder, S. (1998). Ethical issues in electronic performance monitoring: A consideration of deontological and teleological perspectives. *Journal of Business Ethics, 17*, 729-743.

Berlin, I. (2000). The pursuit of the ideal. In H. Hardy & R. Hausheer (Eds.), *The proper study of mankind: An anthology of essays* (pp. 1-16). New York: Farrar, Strauss & Giroux.

Bok, S. (1999). *Lying: Moral choice in public and private life* (Updated ed.). New York: Vintage Books.

Brown, J.S. (Ed.). (1997). *Seeing differently: Insights on innovation*. Boston: Harvard Business Review Books.

Burn, J., & Loch, K. (2001). The societal impact of the World Wide Web—Key challenges for the 21st century. In G. Dhillon (Ed.), *Social responsibility in the information age: Issues and controversies* (pp. 12-29). Hershey, PA: Idea Group Publishing.

Freeman, R., & Soete, L. (1997). *The economics of industrial innovation* (3rd ed.). Cambridge, MA: MIT Press.

Frey, R. (2000). Act-Utilitarianism. In H. LaFollete (Ed.), *The Blackwell guide to ethical theory* (pp. 165-182). Malden, MA: Blackwell Publishing.

Gilbert, J. (2001). New millenium; new technology; same old right and wrong. In G. Dhillon (Ed.), *Information security management: Global challenges in the new millennium*. Hershey, PA: Idea Group Publishing.

Halbert, T. (1994). Computer technology and legal discourse. *Murdoch University Electronic Journal of Law*, 2, May.

Halbert, T., & Ingulli, E. (2002). *Cyberethics*. Cincinnati, OH: West Legal Studies in Business.

Halbert, T., & Ingulli, E. (2003). *Law and ethics in the business environment* (4th ed.). Mason, OH: West Legal Studies in Business.

Ignatieff, M. (2001). *Human rights as politics and idolatry*. Princeton, NJ: Princeton University Press.

Kelly, E., & Rowland, H. (2000, May-June). Ethical and online privacy issues in electronic commerce. *Business Horizons*, 3-12.

LaFollette, H. (Ed.). (2000). *The Blackwell guide to ethical theory*. Malden, MA: Blackwell Publishers.

Mill, J.S., & Bentham, J.(1987). *Utilitarianism and other essays*. London: Penguin Books.

Sumner, L. (1987). *The moral foundation of rights*. Oxford: Clarendon Press.

Velasquez, M. (2002). *Business ethics: Concepts and cases* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

KEY TERMS

Citizen Rights: Those rights that an individual has by virtue of being a member of a government unit (country, state, province, etc.). They vary from government unit to government unit.

Contract Rights: Those rights that an individual has by reason of a valid contract that imposes duties on the other contracting party or parties. They are enforceable under legal systems, but are not the same as citizen rights.

Duties: The correlative of rights, since rights by their nature impose duties.

Ethics: The study of social or interpersonal values and the rules of conduct that follow from them.

Fairness and Justice: The philosophical view that the moral act is the one that treats similarly situated people in similar ways with regard to both process and outcome.

Human Rights: Those rights that all humans have simply by reason of being human, without regard to an individual government unit's laws.

Philosophy: The study of basic principles including what and how we know, rules for language and reasoning, and the basis for social interaction.

Position Rights: Those rights that an individual has by reason of the position that he or she occupies, such as police officer, chief financial officer, or parent.

Rights and Duties: The philosophical view that the moral act is the one that recognizes the rights of others and the duties that those rights impose on the actor.

Utilitarianism: The philosophical view that the moral act is the one that results in the greatest good or happiness for the greatest number of people.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1121-1124, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Evaluating Computer-Supported Learning Initiatives

John B. Nash
Stanford University, USA

Christoph Richter
University of Hannover, Germany

Heidrun Allert
University of Hannover, Germany

INTRODUCTION

The call for the integration of program evaluation into the development of computer-supported learning environments is ever increasing. Pushed not only by demands from policy makers and grant givers for more accountability within lean times, this trend is due also to the fact that outcomes of computer-supported learning environment projects often fall short of the expectations held by the project teams. The discrepancy between the targets set by the project staff and the outcomes achieved suggests there is a need for formative evaluation approaches (versus summative approaches) that facilitate the elicitation of information that can be used to improve a program while it is in its development stage (c.p., Worthen, Sanders & Fitzpatrick, 1997). While the call for formative evaluation as an integral part of projects that aim to develop complex socio-technical systems is widely accepted, we note a lack of theoretical frameworks that reflect the particularities of these kind of systems and the ways they evolve (c.p., Keil-Slawik, 1999). This is of crucial importance, as formative evaluation will only be an accepted and effective part of a project if it provides information useful for the project staff. Below we outline the obstacles evaluation faces with regard to projects that design computer-supported learning environments, and discuss two promising approaches that can be used in complimentary fashion.

BACKGROUND

According to Worthen et al. (1997), evaluation is "the identification, clarification, and application of defensible criteria to determine an evaluation object's value (worth or merit), quality, utility, effectiveness, or significance in relation to those criteria." In this regard evaluation can serve different purposes. Patton (1997) distinguishes between judgment-, knowledge- and improvement-oriented evaluations. We focus on improvement-oriented evaluation approaches. We stress that evaluation can facilitate decision making and reveal information that can be used to improve not only the project itself, but also outcomes within the project's target population. The conceptualization of evaluation as an improvement-oriented and formative activity reveals its proximity to design activities. In fact this kind of evaluative activity is an integral part of any design process, whether it is explicitly mentioned or not. Accordingly it is not the question if one should evaluate, but which evaluation methods generate the most useful information in order to improve the program. This question can only be answered by facing the characteristics and obstacles of designing computer-supported learning environments.

Keil-Slawik (1999) points out that one of the main challenges in evaluating computer-supported learning environments is that some goals and opportunities can spontaneously arise in the course of the development process and are thus not specified in advance. We believe that this is due to the fact that design, in this context, addresses ill-structured and situated problems. The design and implementation of computer-supported learning environments, which can be viewed as a response to a perceived problem, also generates new problems as it is designed. Furthermore every computer-supported learning experience takes place in a unique social context that contributes to the success of an intervention or prevents it. Therefore evaluation requires that designers pay attention to evolutionary and cyclic processes and situational factors. As Weiss notes, "Much evaluation is done by investigating outcomes without much attention to the paths by which they were produced" (1998, p. 55).

For developers designing projects at the intersection of information and communication technology (ICT) and the learning sciences, evaluation is difficult. Evaluation efforts are often subverted by a myriad of confounding variables, leading to a "garbage in, garbage out" effect; the evaluation cannot be better than the parameters that were built in the project from the start (Nash, Plugge & Eurlings, 2001). Leaving key parameters of evaluative thinking out

of computer-supported learning projects is exacerbated by the fact that many investigators lack the tools and expertise necessary to cope with the complexity they face in addressing the field of learning.

We strongly advocate leveraging the innate ability of members of the computer science and engineering communities to engage in "design thinking" and turn this ability into a set of practices that naturally becomes program evaluation, thereby making an assessment of the usefulness of ICT tools for learning a natural occurrence (and a manifest activity) in any computer-supported learning project.

Design-Oriented Evaluation for Computer-Supported Learning Environments

There are two approaches that inherently relate themselves to design as well as to evaluation. Therefore they are useful tools for designers of computer-supported learning initiatives. These two perspectives, discussed below, are scenario-based design and program theory evaluation. Both approaches assume that the ultimate goal of a project should be at the center of the design and evaluation discussion, ensuring a project is not about only developing a usable tool or system, but is about developing a useful tool or system that improves outcomes for the user. Beyond this common ground, these approaches are rather complementary to each other and it is reasonable to use them in conjunction with one another.

Scenario-Based Approaches

Scenario-based approaches are widely used in the fields of software engineering, requirements engineering, human computer interaction, and information systems (Rolland et al., 1996). Scenarios are a method to model the universe of discourse of an application, that is, the environment in which a system, technical or non-technical, will be deployed. A scenario is a concrete story about use of an innovative tool and/or social interactions (Carroll, 2000). Scenarios include protagonists with individual goals or objectives and reflect exemplary sequences of actions and events. They refer to observable behavior as well as mental processes, and also cover situational details assumed to affect the course of actions (Rosson & Carroll, 2002). Additionally it might explicitly refer to the underlying culture, norms, and values (see Bødker & Christiansen, 1997). That said, scenarios usually focus on specific situations, only enlighten some important aspects, and generally do not include every eventuality (e.g., Benner, Feather, Johnson & Zorman, 1993).

Beside their use in the design process, scenarios can also be used for purposes of formative evaluation. First of all, as a means of communication, they are a valuable resource for identifying underlying assumptions regarding the program under development. Stakeholder assumptions might include those related to instructional theories, the learner, the environmental context, and its impact on learning or technical requirements. Underlying assumptions such as these are typically hidden from view of others, but easily developed and strongly held within individuals developing computer-supported learning environments. Scenarios help to reveal the thinking of designers so that others can participate in the design process and questionable assumptions can come under scrutiny. The use of scenarios also allows identification of pros and cons of a certain decision within the design process. In this vein Carroll (2000) suggests employing "claim analysis." Claims are the positive or negative, desirable and undesirable consequences related to a certain characteristic of a scenario. Assuming that every feature of a proposed solution usually will entail both positive and negative effects helps to reflect on the current solution and might provoke alternative proposals. The analysis of claims is thereby not limited to an intuitive ad hoc evaluation, but also can bring forth an explicit hypothesis to be addressed in a subsequent survey.

Program Theory Evaluation

Program theory evaluation, also known as theory-based evaluation, assumes that underlying any initiative or project is an explicit or latent "theory" (or "theories") about how the initiative or project is meant to change outcomes. An evaluator should surface those theories and lay them out in as fine detail as possible, identifying all the assumptions and sub-assumptions built into the program (Weiss, 1995). This approach has been promoted as useful in evaluating computer-supported learning projects (Strömdahl & Langerth-Zetterman, 2000; Nash, Plugge & Eurlings, 2001) where investigators across disciplines find it appealing. For instance, for designers (in mechanical engineering or computer science), program theory evaluation reminds them of their own use of the "design rationale." And among economists, program theory evaluation reminds them of total quality management (TQM). In the program theory approach (Weiss, 1995, 1998; Chen, 1989; Chen & Rossi, 1987), one constructs a project's "theory of change" or "program logic" by asking the various stakeholders, "What is the project designed to accomplish, and how are its components intended to get it there?" The process helps the project stakeholders and the evaluation team to identify and come to consensus on the project's theory of change. By identifying and describing the activities, outcomes, and goals of the program, along with their interrelationships, the stakeholders are then in position to identify quantifiable measures to portray the veracity of the model.

Theory-based evaluation identifies and tests the relationships among a project's inputs or activities and its outcomes via intermediate outcomes. The key advantages

to using theory-based evaluation are (Connel & Kubisch, 1995; Weiss, 1995):

- It asks project practitioners to make their assumptions explicit and to reach consensus with their colleagues about what they are trying to do and why.
- It articulates a theory of change at the outset and gains agreement on it, by all stakeholders reducing problems associated with causal attribution of impact.
- It concentrates evaluation attention and resources on key aspects of the project.
- It facilitates aggregation of evaluation results into a broader context based on theoretical program knowledge.
- The theory of change model identified will facilitate the research design, measurement, data collection, and analysis elements of the evaluation.

Both scenario-based design and program theory stress the importance of the social context while planning computer-supported environments. They also represent means to facilitate the communication among the stakeholders and urge the project team to reflect their underlying assumptions in order to discuss and test them. Furthermore, both approaches are particularly suitable for multidisciplinary project teams. Scenarios and program logic maps are not static artifacts; they are a starting point for discussion and have to be changed when necessary. With these similarities there are also differences in both approaches. The major difference between them is that program theory offers a goal-oriented way to structure a project, while scenario-based design proffers an explorative approach that opens the mind to the complexity of the problem, alternatives, and the diversity of theories that try to explain social and socio-technical process. That is, scenario-based design highlights the divergent aspects of project planning, and evaluation program theory stresses the convergent aspects. Program theory evaluation helps to integrate each scenario, decision, and predefinition into the whole process. Scenarios force users not just to use terms, but to give meaningful descriptions. They force users to state how they actually want to instantiate an abstract theory of learning and teaching. This helps to implement the project within real situations of use, which are complex and ill structured. Program theory helps to focus on core aspects of design and prevent getting 'lost in scenarios'. Scenarios and program theory evaluation can be used in an alternating way. Thereby it is possible to use both approaches and improve the overall development process.

The program theory of an initiative can be a starting point for writing scenarios. Especially the interrelations between the goals and interrelation between ultimate goal and inputs can be described with a scenario. The scenario can help to understand how this interrelation is meant to work and how it will look in a concrete situation. Scenarios on the other hand can be used to create program theory by pointing out main elements of the intended program. They can also be used to complete already existing program theory by presenting alternative situations of use. For developers of computer-supported learning environments, scenario-based design and program theory represent complementary approaches, which when used together or separately, can add strength to the implementation and success of such projects.

FUTURE TRENDS

It is clear that formative evaluation will become more important in the future, and it will be especially crucial to think about how to integrate evaluation into the design process. Essentially, designers will need to answer the question "Why does the program work?" and not just "Did it work?" It becomes obvious that the design of a computer-supported learning environment, like the development of any other complex socio-technical system, is a difficult process. In fact the necessity for changes in the original plan is practically preordained due to the ill-structured and situated nature of the domain. The mere act of engaging in a design process suggests that designers will engage in planned as well as evolutionary, unplanned activities. Therefore it is important that the project designers use methods that support divergent thinking and methods that support convergent processes. While scenario-based design and program-theory evaluation represent complementary views on the design and evaluation of computer-supported learning environments that can facilitate these processes, there is still room for improvement.

CONCLUSION

Formative evaluation is an important means to ensure the quality of an initiative's outcomes. Formative evaluation directed towards improvement of an initiative can be understood as a natural part of any design activity. While this is widely recognized, there is still a lack of program evaluation frameworks that reflect the uniqueness of the design process, the most crucial of which is the inherent ambiguity of design. In spite of great inspiration portrayed by project teams, usually manifested by visions of a certain and sure outcome, no project can be pre-planned completely, and mid-course corrections are a certainty. Scenario-based design and program theory evaluation provide a theoretical foothold for projects in need of collecting and analyzing data for program improvement and judging program success.

In sum, scenario-based design and program theory hold many similarities. The major difference between them is that program theory offers a goal-oriented way to structure a project, while scenario-based design provides an explorative approach that opens the mind to the complexity of the

problem, alternatives, and the diversity of theories that try to explain social and socio-technical process.

Scenario-based design highlights the divergent aspects of project planning, and evaluation program theory stresses the convergent aspects. For developers of computer-supported learning experiences, scenario-based design and program theory represent complementary approaches, which when used together or separately can add strength to the implementation and success of ICT learning projects.

REFERENCES

Benner, K.M., Feather, M.S., Johnson, W.L. & Zorman, L.A. (1993). Utilizing scenarios in the software development process. In N. Prakash, C. Rolland & B. Pernici (Eds.), *Information system development process* (pp. 117-134). Elsevier Science Publishers.

Bødker, S. & Christiansen, E. (1997). Scenarios as springboards in design. In G. Bowker, L. Gaser, S.L. Star & W. Turner (Eds.), *Social science research, technical systems and cooperative work* (pp. 217-234). Lawrence Erlbaum.

Carroll, J.M. (2000). *Making use: Scenario-based design of human-computer interactions.* Cambridge: MIT Press.

Chen, H.T. (1989). Issues in the theory-driven perspective. *Evaluation and Program Planning, 12,* 299-306.

Chen, H.T. & Rossi, P. (1987). The theory-driven approach to validity. *Evaluation Review, 7,* 95-103.

Connell, J.P. & Kubisch, A. (1995). Applying a theory of change approach to the evaluation of comprehensive community initiatives: Progress, prospects, and problems. In K. Fulbright-Anderson et al. (Eds.), *New approaches to evaluating community initiatives. Volume 2: Theory, measurement, and analysis.* Washington, DC: Aspen Institute.

Keil-Slawik, R. (1999). Evaluation als evolutionäre systemgestaltung. aufbau und weiterentwicklung der paderborner DISCO (Digitale Infrastruktur für computerunterstütztes kooperatives Lernen). In M. Kindt (Ed.), *Projektevaluation in der lehre—multimedia an hochschulen zeigt profil(e)* (pp. 11-36). Münster, Germany: Waxmann.

Nash, J.B., Plugge, L. & Eurlings, A. (2001). Defining and evaluating CSCL evaluations. In A. Eurlings & P. Dillenbourg (Eds.), *Proceedings of the European Conference on Computer-Supported Collaborative Learning* (pp. 120-128). Maastricht, The Netherlands: Universiteit Maastricht.

Patton, M.Q. (1997). *Utilization-focused evaluation* (3rd Edition). Thousand Oaks, CA: Sage Publications.

Rolland, C., Achour, C.B., Cauvet, C., Ralyté, J., Sutcliffe, A., Maiden, N.A.M., Jarke, M., Haumer, P., Pohl, K., Dubois, E. & Heymans, P. (1996). *A proposal for a scenario classification framework.* CREWS Report 96-01.

Rosson, M.B. & Carroll, J.M. (2002). Usability engineering: Scenario-based development of human-computer interaction. San Francisco: Morgan Kaufmann.

Strömdahl, H. & Langerth-Zetterman, M. (2000). *On theory-anchored evaluation research of educational settings, especially those supported by information and communication technologies (ICTs).* Uppsala, Sweden: Swedish Learning Lab.

Weiss, C. (1995). Nothing as practical as good theory: Exploring theory-based evaluation for comprehensive community initiatives for children and families. In J. Connell et al. (Eds.), *New approaches to evaluating community initiatives: Concepts, methods, and contexts.* Washington, DC: Aspen Institute.

Weiss, C. (1998). *Evaluation research: Methods for studying programs and policies.* Englewood Cliffs, NJ: Prentice-Hall.

Worthen, B.R., Sanders, J.R. & Fitzpatrick, J.L. (1997). *Program evaluation—alternative approaches and practical guidelines* (2nd Edition). New York: Addison Wesley Longman.

KEY TERMS

Computer-Supported Learning: Learning processes that take place in an environment that includes computer-based tools and/or electronically stored resources. CSCL is one part of this type of learning.

Evaluation: The systematic determination of the merit or worth of an object.

Formative Evaluation: The elicitation of information that can be used to improve a program while it is in the development stage.

Program: A social endeavor to reach some predefined goals and objectives. A program draws on personal, social, and material resources to alter or preserve the context in which it takes place.

Program Theory: A set of assumptions underlying a program that explains why the planned activities should lead to the predefined goals and objectives. The program theory includes activities directly implemented by the program, as well as the activities that are generated as an response to the program by the context in which it takes place.

Scenarios: A narrative description of a sequence of (inter-)actions performed by one or more persons in a particular context. Scenarios include information about goals, plans, interpretations, values, and contextual conditions and events.

Summative Evaluation: The elicitation of information that can be used to determine if a program should be continued or terminated.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1125-1129, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

E

Evaluating UML Using a Generic Quality Framework

John Krogstie
IDI, NTNU, SINTEF, Norway

INTRODUCTION

According to Booch, Rumbaugh, and Jacobson (2005), developing a model for an industrial strength software system before its construction is increasingly regarded as a necessary activity in information systems development. The use of object-oriented modeling in analysis and design started to become popular in the late 80s, producing a large number of different languages and approaches. Over the last 10 years, UML (OMG, 2006a) has taken a leading position in this area.

In this chapter, we give an overview assessment of UML using a generic evaluation framework. We will first present the evaluation framework. We will then evaluate the language quality of UML before pointing to the future direction and potential of UML.

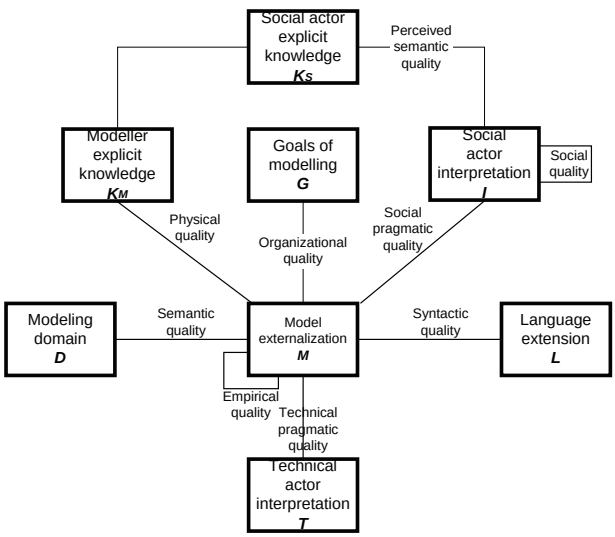
BACKGROUND

Earlier, we developed a framework for understanding and assessing quality of models and modeling languages (Krogstie & Sølvberg, 2003; Krogstie, Sindre, & Jørgensen, 2006).

The main concepts of the framework and their relationships are shown in Figure 1 and are explained next. Quality has been defined referring to the correspondence between statements belonging to the following sets:

- G, the goals of the modeling task
- L, the language extension (i.e., the set of all statements that are possible to make according to the graphemes, vocabulary, and syntax of the modeling languages used)
- D, the domain (i.e., the set of all statements that can be stated about the situation at hand)
- M, the externalized model itself
- Ks, the relevant explicit knowledge of those being involved in modeling. A subset of these is actively involved in modeling, and their explicit knowledge is indicated by K_M.
- I, the social actor interpretation (i.e., the set of all statements that the audience thinks that an externalized model consists of)
- T, the technical actor interpretation (i.e., the statements in the model as "interpreted" by modeling tools)

Figure 1. Framework for discussing the quality of models



The main quality types are indicated by solid lines between the sets and are described briefly next:

- **Physical quality:** The basic quality goals on the physical level is that the knowledge K of the domain D has been externalized, and internalizeability, that the externalized model M is available.
- **Empirical quality** deals with predictable error frequencies when a model is read or written by different users, coding (e.g., shapes of boxes) and HCI-ergonomics for documentation and modeling-tools. For instance, graph layout to avoid crossing lines in a model is a mean to address the empirical quality of a model.
- **Syntactic quality** is the correspondence between the model M and the language extension L.
- **Semantic quality** is the correspondence between the model M and the domain D. This includes validity and completeness.
- **Perceived semantic quality** is the similar correspondence between the audience interpretation I of a model M and his or hers current knowledge K of the domain D.
- **Pragmatic quality** is the correspondence between the model M and the audience's interpretation and application of it (I). We differentiate between social pragmatic quality (to what extent people understand and are able to use the models) and technical pragmatic quality (to what extent tools can be made that interpret the models).

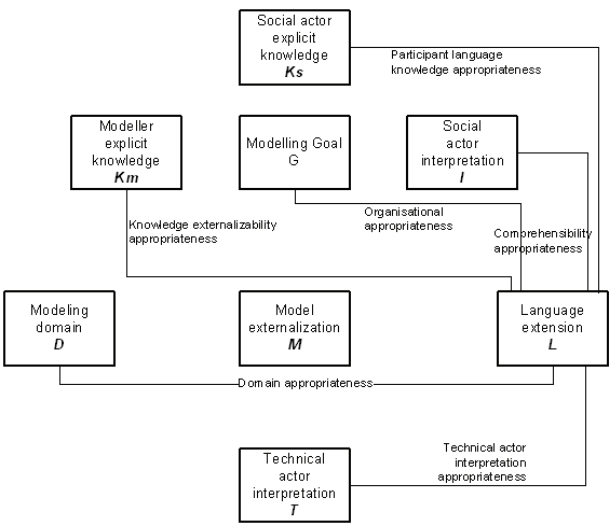
The goal defined for social quality is agreement among audience members' interpretations I.

The organizational quality of the model relates to that all statements in the model contribute to fulfilling the goals of modeling (organizational goal validity), and that all the goals of modeling are addressed through the model (organizational goal completeness).

Language quality relates the modeling language used to the other sets. Six quality areas for language quality are identified with aspects related to both the language metamodel and the notation as illustrated in Figure 2.

- **Domain appropriateness:** This relates the language and the domain. Ideally, the conceptual basis must be powerful enough to express anything in the domain, not having what Wand and Weber (1993) term construct deficit. On the other hand, you should not be able to express things that are not in the domain (i.e., what is termed construct excess) (Wand et al., 1993). Domain appropriateness is primarily a mean to achieve physical quality, and through this, to achieve semantic quality.
- **Participant language knowledge appropriateness** relates the social actors' explicit knowledge to the language. Participant language knowledge appropriateness is primarily a mean to achieve physical and pragmatic quality.
- **Knowledge externalizability appropriateness:** This area relates the language extension to the participant knowledge. The goal is that there are no statements in the explicit knowledge of the modeler that cannot be expressed in the language. Knowledge externalizability appropriateness is primarily a mean to achieve physical quality.

Figure 2. Language quality in the quality framework



- **Comprehensibility appropriateness** relates the language to the social actor interpretation. The goal is that the participants in the modeling effort using the language understand all the possible statements of the language. Comprehensibility appropriateness is primarily a mean to achieve empirical and pragmatic quality.
- **Technical actor interpretation appropriateness** relates the language to the technical audience interpretations. For tool interpretation, it is especially important that the language lend itself to automatic reasoning. This requires formality (i.e., both formal syntax and semantics being operational and/or logical), but formality is not necessarily enough, since the reasoning must also be efficient to be of practical use. This is covered by what we term analyzability (to exploit any mathematical semantics) and executability (to exploit any operational semantics). Different aspects of technical actor interpretation appropriateness are a mean to achieve syntactic, semantic, and pragmatic quality (through formal syntax, mathematical semantics, and operational semantics).
- **Organizational appropriateness** relates the language to standards and other organizational needs within the organizational context of modeling. These are means to support organizational quality.

Whereas knowledge externalizability appropriateness and organizational appropriateness is closely linked to particular organizations and modelers involved, for a more general evaluation, we can use the other four categories.

OVERVIEW OF EVALUATION

The evaluation is structured according to the four areas of language quality previously presented. Before presenting the evaluation, we will position UML in relation to the sets of the quality framework.

- **Domain:** According to OMG (2006a), UML is a language for specifying, visualizing, constructing, and documenting the artifacts of software systems, as well as for business modeling and other non-software systems. In other words, UML is meant to be used in analysis of business and information, requirement specification, and design. UML is meant to support the modeling of (object-oriented) transaction systems, real-time, and safety critical systems. As illustrated in Favre (2003), UML is used and adapted for a number of different specific areas.
- **Language:** We have based the evaluation on UML (version 2.0, (Booch et al., 2005, OMG, 2006a)).

The sets "Knowledge," "Model," and "Interpretation" must be judged from case to case in the practical application of a modeling language and tools. Also when it comes to weighting the different criteria against each other, this must be done in the light of the specific modeling task, such as has been done by (e.g., Krogstie & Arnesen, 2004; Nysetvold & Krogstie, 2006).

Due to the limitation on the length of a paper of this kind and the breadth of the evaluation, we will only have room for presenting some of the major results. See Krogstie (2003) for a more detailed description of using the framework for evaluating UML.

Domain Appropriateness

Looking briefly on the coverage of the seven main modeling-perspectives in information systems modeling (Krogstie et al., 2003), we find:

- UML has a very good support for modeling according to an object-oriented perspective, especially for design.
- The structural perspective is also well supported, although not as well as in languages made specifically for this purpose.
- The behavioral perspective is supported particularly through statecharts.
- The functional (process) perspective is supported on a high level through use case modeling, a language that has been highly criticized for not being well-defined (Hitz & Kappel, 1998). Whereas use-cases are meant for requirements modeling, activity diagrams can be used for simple procedural description by showing control flow and the production of data or objects in a process flow. Changes to the activity diagrams are introduced in UML2.0 to improve the modeling of business processes. The lack of traditional dataflow in activity diagrams has been noted as a problem. Note that another upcoming standard in this area, BPMN (BPMN 2006) has been taken over by OMG, although there are no immediate plans for including BPMN in UML.
- The actor-role perspective is partly covered using the collaboration diagrams. Using roles in sequence diagrams or "swimlanes" in activity diagrams, we also get a role-oriented view, but there is no intuitive way to represent organizational and group-structures and relationships in UML.
- There are fundamental problems with expressing certain constraints in any OO modeling framework (Høydalsvik & Sindre, 1993). Temporal and deontic constraints are hard to express. Whereas these kind of concepts are not included in UML, a new initiative within OMG, SBVR (OMG, 2006b), includes such

things as deontic operators. For the moment, there are no plans for including this in UML. There is no support for modeling of goal-hierarchies (Mylopoulos, Chung, & Tu, 1999).
- The language-action perspective, which is most useful for the analysis of businesses, is not supported.

A meta-model of UML is defined (using UML), and there exist extension mechanisms to make the language more applicable in specific domains, although when creating such profiles, it is not possible to describe the semantics of the extensions in a formal manner.

Most of UML is first useful during design. These parts of the language should not be used in analysis and requirements specification, even in areas where the transition from analysis to design is "seamless." (There is much evidence, especially for business systems, that this transition is far from seamless even when using object-oriented modeling in all domains (Davis, 1995; Høydalsvik et al., 1993). Proper guidelines for avoiding this are not consistently provided, and there is limited support for avoiding using analysis and design concepts in the same model.

There are also mismatches between underlying basis and external representation. In sequence diagrams, for instance a number of concepts are semantically vacant (Morris & Spanoudakis, 2001). Some of these problems are addressed in UML 2.0. We also find examples of concrete constructs in the UML meta-model, with no representation in the notation (e.g., namespace and model).

Participant Language Knowledge Appropriateness

It can be argued that for those being familiar with the main OO-modeling concepts and main OO modeling-languages, the core of UML should not represent a too steep learning curve. Almost all CS and IS-degrees now include courses where UML is lectured and used. On the other hand, we have noted the complexity of the language above. The large number of constructs in UML is partly due to its diverse diagramming techniques (Siau & Cao, 2001). Constructs in use case diagrams are very different from constructs in class diagrams. Class diagrams are also very different from activity diagrams or Statecharts. This diversity causes problems. First, there are more constructs to learn. Second, the knowledge and experience gained in using one diagramming technique cannot easily be transferred to another diagramming technique in UML. The definition of language units and increments within units introduced in UML 2.0 will hopefully help in practical application of the language.

Participant Comprehensibility Appropriateness

UML can be argued to be overly complex with 233 different concepts (in UML 1.4) (Castellani, 1999). A similar number of concepts can be found in UML 2.0. In Siau et al. (2001) a detailed comparison between UML and other modeling approaches are presented. Although UML has more diagramming techniques when compared to object-oriented methods such as OMT, OOAD, and Shlaer/Mellor, each of the diagramming techniques in isolation is no more complex than techniques found in these methods. On the overall method level on the other hand, UML stands out noticeably, being most complex according to most of the metrics. On the other hand, Erickson and Siau (2004) point out that the actual usage of a large number of the concepts within UML is limited.

With so many concepts, it is not surprising that some redundancy and overlap is witnessed.

Examples of lacking symbol differentiation is found, for example, that both classes and objects are shown using rectangles. On the other hand, since the text-style of the object gives an indication of this, this is a minor problem.

UML contains many possibilities of adding (often small) adornments to the models. Such small adornments are often difficult to see and comprehend.

A uniform use of symbols is not adhered to. An example is that different symbols are used for a role if it is external to the system (pin-man) or internal (rectangle).

Technical Actor Interpretation Appropriateness

The UML-syntax is rigorously defined and the language is described through a meta-model made in the structural model of UML with accompanying descriptions of the semantics. Using UML to model UML means that some of the definitions are circular, and this leaves UML (formally speaking) undefined. This would be unproblematic if most practitioners already understood the meaning of the core concepts (classes, inheritance, and associations) that are involved. In UML 2.0 a formal (operational) action language have been included to support a wider repertoire of modeling techniques. Only some parts of UML have been provided a formal (operational) semantics. So-called semantic variation points are included in UML 2.0 to clearly indicate where the semantics is not rigorously defined. In any case, the UML meta-model only describe the structural aspects of the language, and not the behavioral aspects.

FUTURE TRENDS

Modeling as a general technique within information systems development has received increasing interest over the last years, and will continue to be of high importance, for example, in connection to the OMG MDA approach and business process support. Due to its strong support, UML is probably the best general modeling language to adopt as a basis for object-oriented development if one is not already using another language with good tool support that one is satisfied with. Most of the accidental problems such as inconsistencies in the language-descriptions found in earlier version of UML seem to be addressed in UML 2.0, but there are still concerns. UML 3.0 is already planned, but it is difficult to judge what this will look like. Judging from the time it has taken to finalize UML 2.0, we expect to have to relate to this for a number of years.

CONCLUSION

UML has been developed and refined over the last 10 years, but there are still some major weaknesses with the approach. Even if it has not been possible to agree on a standard process, outline process guidelines need to be included--even if the best that can be done is to describe a number of alternatives. Particularly problematic is the logical/physical confusion in the UML-definition. As discussed by Davis (1995), there are fundamental differences between the models related to analysis, design, and requirement specification. What our investigation has also illustrated is that although there is a perceived need to extend the expressiveness and formality of the language, the language has several weaknesses regarding comprehensibility appropriateness, and is already looked upon as difficult to comprehend. The distinction between infrastructure and superstructure provided for UML 2.0 tries to address this. Looking at the accepted 1000-plus-page proposal for UML2.0 superstructure alone does not give us much hope that the whole of UML will be particularly much easier to learn and use, as pointed out also in France, Ghosh, Dinh-Trong, and Solberg (2006).

REFERENCES

Booch, G., Rumbaugh, J., & Jacobson, I. (2005). *The unified modeling language: User guide* (2nd ed.). Addison-Wesley.

BPMN. (2006). *Business process modeling notation*. Retrieved from http://www.bpmn.org

Castellani, X. (1999). Overview of models defined with charts of concepts. In E. Falkenberg, K. Lyytinen, & A. Verrijn-Stuart (Eds.), *Proceedings of the IFIP8.1 Working Conference On Information Systems Concepts (ISCO4)* (pp. 235-256); An Integrated Discipline Emerging September 20-22 , Leiden, The Netherlands.

Davis, A. (1995). Object-oriented requirements to object-oriented design: An easy transition? *Journal of Systems and Software, 30*(1/2), 151-159.

Erickson, J., & Siau, K. (2004). Theoretical and practical complexity of unified modeling language: Delphi study and metrics analyses. In *Proceedings of International Conference on Information Systems (ICIS)* (pp. 183-194).

Favre, L. (2003). *UML and the unified process*. IRM Press.

France, R. B., Ghosh, S., Dinh-Trong, T., & Solberg, A. (2006). Model-driven development using UML2.0: Promises and pitfalls. *IEEE Computer*, February, 59-66.

Hitz, M., & Kappel, G. (1998). Developing with UML--Some pitfalls and workarounds. In J. Bézivin, & P. A. Muller (Eds.), *UML'98—Beyond the notation* (pp. 9-20). Mulhouse, France: Springer-Verlag.

Høydalsvik, G. M., & Sindre, G. (1993, September). On the purpose of object-oriented analysis. In A. Paepcke (Ed.), *Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications (OOPSLA'93)* (pp. 240-255). ACM Press.

Krogstie, J. (2003). Evaluating UML using a generic quality framework. In L. Favre (Ed.), *UML and the unified process* (pp. 1-22). IRM Press.

Krogstie, J., & Arnesen, S. (2004). Assessing enterprise modeling languages using a generic quality framework. In J. Krogstie, K. Siau, & T. Halpin (Eds.), *Information modeling methods and methodologies.* Hershey, PA: Idea Group Publishing.

Krogstie, J., Sindre, G., & Jørgensen, H. (2006). Process models representing knowledge for action: A revised quality framework. *European Journal of Information Systems*, *15*, 91-102.

Nysetvold, A. G., & Krogstie, J. (2006). Assessing business process modeling languages using a generic quality framework. In K. Siau (Ed.), *Advanced topics in database research.* Hershey, PA: Idea Group Publishing.

Morris, S., & Spanoudakis, G. (2001). UML: An evaluation of the visual syntax of the language. In *Proceedings of HICSS 34*.

Mylopoulos, J., Chung, L., & Tu, E. (1999). From object-oriented to goal-oriented requirements analysis. *Communications of the ACM, 42*(1), 31-37.

OMG. (2006a). *Unified modeling language v 2.0*. Retrieved from http://www.omg.org

OMG. (2006b). *Semantics of business vocabulary and rules interim specification*. Retrieved June 3, 2002, from http://www.omg.org/cgi-bin/doc?dtc/

Siau, K., & Cao, Q. (2001). Unified modeling language (UML)—A complexity analysis. *Journal of Database Management,* 26-34, Jan-Mar.

Wand, Y., & Weber, R. (1993). On the ontological expressiveness of information systems analysis and design grammars. *Journal of Information Systems*, *3*(4), 217-237.

KEY TERMS

Analysis Model: A model developed to learn all aspects of a problem domain to determine the best way to solve a specific set of user needs.

Design Model: A model developed to represent the optimal technical solution of a specified user need (as represented in a requirements model).

Model: An abstraction represented in a modeling language.

Modeling Language: A language (i.e., a set of symbols, and rules for how to combine these symbols) to represent knowledge relevant in (information systems) development.

Requirements Model: A model to represent the external requirement to a system without taking into account how the system looks inside.

UML (Unified Modeling Language): A general-purpose visual modeling language that is used to specify, visualize, construct and document the artifacts of a software system.

Visual Modeling Language: A diagrammatic modeling language (i.e., where the models made in the language are 2-dimensional diagrams).

Evaluating Computer Network Packet Inter-Arrival Distributions

Dennis Guster
St. Cloud State University, USA

David Robinson
St. Cloud State University, USA

Richard Sundheim
St. Cloud State University, USA

INTRODUCTION

The past decade could be classified as the "decade of connectivity"; in fact, it is commonplace for computers to be connected to an LAN, which in turn is connected to a WAN, which provides an Internet connection. On an application level this connectivity allows access to data that even five years earlier were unavailable to the general population. This growth has not occurred without problems, however. The number of users and the complexity/size of their applications continue to mushroom. Many networks are over-subscribed in terms of bandwidth, especially during peak usage periods. Often network growth was not planned for, and these networks suffer from poor design. Also, the explosive growth has often necessitated that crisis management be employed just to keep basic applications running. Whatever the source of the problem, it is clear that proactive design and management strategies need to be employed to optimize available networking resources (Fortier & Desrochers, 1990). This is especially true in today's world of massive Internet usage (Zhu, Yu, & Doyle, 2001).

BACKGROUND

Obviously, one way to increase network bandwidth is to increase the speed of the links. However, this may not always be practical due to cost or implementation time. Furthermore, this solution needs to be carefully thought out because increasing speed in one part of a network could adversely effect response time in another part of that network. Another solution would be to optimize the currently available bandwidth through programming logic. Quality of service (QOS) and reservation bandwidth (RB) are two popular methods currently being utilized. Implementation of these optimization methods is rarely simple and often requires a high degree of experimentation if they are to be effectively configured (Walker, 2000). This experimentation can have detrimental effects on a live network, often taking away resources from mission critical applications. The client/server model so popular in Internet communication is a prime example of a system that can benefit from an analytical modeling strategy (Postigo-Boix, Garcia-Haro, & Melus-Moreno, 2005).

THE BENEFITS OF SIMULATION

Therefore, the most efficient way to ascertain the potential benefit and derive baseline configuration parameters for these optimization methods is through simulation or mathematical modeling. Simulation can be very effective in planning a network design. For example, what if network link number three was increased to 10Gbs? Would workstations on that link experience an improvement in response time? What would happen to workstations on the other part of the total network? Another approach to ascertain if a given network will exceed its capacity is based on network calculus (Cruz, 1991; Le Boudec, 1998). In this method the characteristics (such as speed, maximum packet size, and peak rate) of the network architecture are analyzed, and performance bounds are defined. The goal then is to devise control/management programs (such as QOS and RB) that will keep the workload within those defined bounds. There are numerous applications of this control/management logic, such as Cruz (1995), Cruz and Tsai (1996), Firoiu, Le Boudec, Towsley, and Zhang (2002), and Vojnovic and Le Boudec (2002). Recent work has focused on integrated service on WANs across service providers (Cruz & Santhanam, 2000). These control/management programs have proved very effective under a variety of circumstances, but are influenced by the packet inter-arrival rate as well. Therefore, if a network designer is contemplating invoking one of these options, simulation could be used to test how the option in question would improve performance on his or her system, provided an adequate method could be found to describe

the distribution within that network. Simulation has been used for many years in network design; however, the time and cost of its use have often been prohibitive. In recent years, new windows-based point and click products such as Comnet III (and its successors Network & Simscript) have eliminated the drudgery and the cost of writing simulations via a command line interface (CACI, 1998). Under Comnet III the appropriate devices are selected, connected together, and their characteristics defined. There is still a limiting factor in this process: the definition of the distribution of the packet inter-arrival rates. The theoretical model often used to describe computer networking is the Poisson. This model may have been adequate for some of the first single tier, single protocol networks. However, it lacks validity in describing the total stream in today's hierarchically complex multi-protocol networks. In the classical Poisson process model (such as M/M/1), when the number of arrivals follows a Poisson probability distribution, then the time between arrivals (inter-arrival time) follows a decaying exponential probability distribution. A number of studies confirm that the actual inter-arrival distribution of packets is not totally exponential as would be expected in the classical model (Guster, Robinson, & Richardson, 1999; Krzenski, 1998; Partridge, 1993; Vandolore, Babic, & Jain, 1999). However, in light of recent changes in networking regarding line speed and number of connected hosts that in effect have tripled in magnitude, the Poisson model is being reevaluated. Specifically, Karagiannis, Molle, and Faloutsos (2004) found that the stream as a whole may not be exactly Poisson, but some of it components might fit quite well. Those cases involved sub-second streams and large multi-second streams.

The inter-arrival distribution selected can have a major impact on the results of the simulation (Guster, Safonov, Hall, & Sundheim, 2003; Guster, Sohn, Robinson, & Safonov, 2003). In a study by Krzenski (1999) that analyzed the simulated performance on a shared Ethernet network, 12 different inter-arrival distributions were tried within the same simulation problem. These included the gamma distribution, which is a generalization of the exponential distribution, allowing for a modal inter-arrival time (the most commonly occurring time between arrivals) to be moved out away from the very short, nearly instantaneous time occurring with the exponential distribution. Another distribution among the 12 was an integer distribution, whereby equal probabilities are assigned to different values that are equally spaced throughout the possible inter-arrival times. Among the 12 distributions, there were vast discrepancies in the results. For example, the number of collision episodes varied from 310 with a gamma distribution to 741 with an integer distribution. These results further support the need to have the correct distribution in simulations designed to provide design and management feedback about computer networks. The frustration of the past work and the need for additional research is best summarized by Partridge (1993, p. 3):

… We still do not understand how data communication traffic behaves. After nearly a quarter of a century of data communication, researchers are still struggling to develop adequate traffic models. Yet daily we make decisions about how to configure networks and configure network devices based on inadequate models of data traffic. There is a serious need for more research work on non-Poisson queuing models.

A number of different strategies have been employed in the development of models used to describe packet inter-arrival rates (Guster, Litvinov, Richardson, & Robinson, 2002). Perhaps the most valid is to record all of the packet arrival times for the time period desired and use that to generate the distribution. The advantage of this strategy is accuracy, but it often requires massive amounts of data to be recorded and processed. To lessen this burden, often a representative sample from the time period is used. However, validating the sample period is often difficult, especially if the file size is not large. Known distributions have been used with limited success (Guster & Robinson, 1994, 2000; Guster, Robinson, & Juckel, 2000). For simple networks, exponential distributions provide some promise; however, they fail to deal with the intricacies of complex multi-protocol networks. Tabular distributions, in which one column describes the interval and a second column describes the probability of a value from that interval occurring, offer a moderate degree of accuracy, but they take time to derive, and their sophistication is related to the number of rows included. Regression and ANOVA have been used in some cases but lack the ability to describe the peaks and valleys associated with packet arrival data. Time series deals with these variations better but still lacks the sophistication needed and requires relatively complex models to even come close (Guster & Robinson, 1993). Packet trains are very effective in describing packet traffic from a single session such as telnet (Vandolore et al., 1999) but lack the complexity to deal with multiple concurrent sessions on the same network. However, the basic idea of breaking the total stream into subparts has been improved using a multilevel traffic approach. Karagiannis, Papagiannaki, and Faloutsos (2005) have been able to devise a classification schema based on the applications that generate them and have been able to classify 80-90% of traffic with more than 95% accuracy.

FUTURE TRENDS

Two non-Poisson queuing models have offered a degree of promise. One method involves viewing the observation interval as containing several independent Poisson processes rather than as a single exponential distribution. In a study by Guster et al. (1999), actual data were analyzed and shown to contain three Poisson processes of differing characteristics. During the first phase activity was increasing. During the

Figure 1. Frequency counts of packet inter-arrival times for Ethernet text data

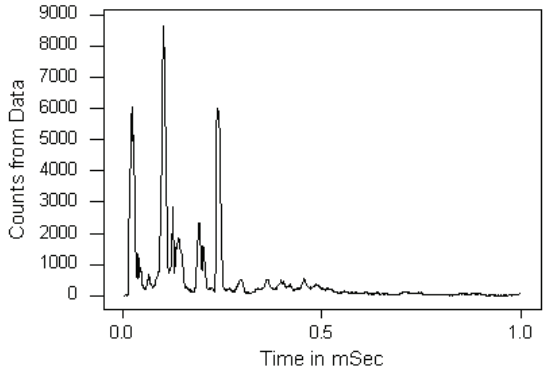
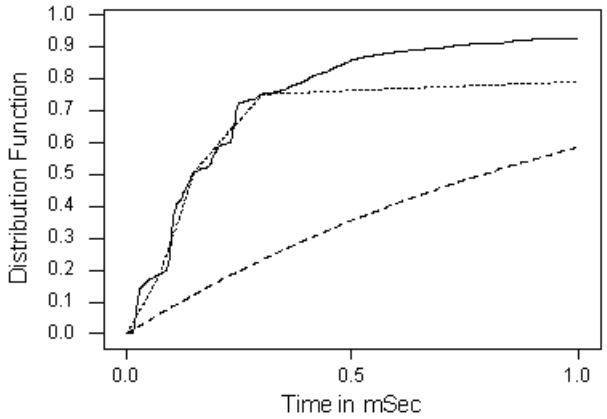


Figure 2. Distribution function for the actual data, the Poisson model, and Markov model



second phase activity was decreasing. The last phase followed a classical Poisson model. For each phase, a power law process model was fit to the data, indicating the nature of the changing traffic intensity (b>1 – increasing intensity, b<1 – decreasing intensity, b=1 – constant intensity). These data were taken over a 24-hour period. The three phases had widely different levels of traffic intensity. In a shorter time frame, for example 10 minutes, one is less likely to see differences in intensity that dramatic. Thus, the power law process model is most appropriate for longer time frames. A second strategy focuses on the influence any given data point has on later data points. In other words, does knowing the magnitude of the packet inter-arrival rate at any point in the time interval make it easier to predict the next inter-arrival time? Historically, there has been a tendency to apply statistical treatments such as regression, ANOVA, or time series analysis to categorize or forecast inter-arrival trends (Frieberger, 1972). These methods have proved to be limited in accuracy in modeling inter-arrival rates and require large, truly representative databases to calculate these values. Therefore, the attractiveness of methodologies such as Markov chains that use a more independent technology is apparent (Robert & Le Boudec, 1996). Specifically, a Markov chain is a tool for modeling how processes behave over time. We classify the process (in this case, the inter-arrival times of packets) into categories called "states." The assumption behind Markov chains is that the probability for what state will be observed next depends only on the previous state. In this context, a process depends only on the most recent inter-arrival time to determine the probability of subsequent inter-arrival times increasing or decreasing (Guster & Robinson, 1994, 2000). To provide the reader with a visual description of packet inter-arrival times from a complex multi-protocol network, a frequency plot is provided in Figure 1. It is the massive peaks and valleys from 0 to about 0.25 milliseconds that make modeling packet inter-arrival times so difficult. There are fairly great discrepancies in the distribution functions of the various distributions. For example, using the actual data from Figure 1, two attempts to model the cumulative distribution of the inter-arrival times are graphed with the actual distribution. Figure 2 depicts the actual data with the solid line, the Poisson model with long dashed line, and the Markov model with the short dashed line. From the data in Figure 2 it is clear that the Markov model fits quite a bit better than the classic Poisson model. To illustrate the effect not having the correct distribution can have on the results of a simulation, a 10-minute sample of packet traffic (packet inter-arrival times) was fed into a database inquiry simulation program. The delay in milliseconds from inquiry to response from that database was recorded. From that sample a tabular distribution was devised and placed into the same simulation and the results recorded. Then the same simulation was run three more times using exponential, lognormal, and normal distributions with the appropriate means and standard deviations derived from the original 10-minute sample. The normal distribution is the classical mound shaped probability distribution used in many statistical applications. The lognormal distribution is the result of exponentiating values from a normal distribution, giving a more realistic, asymmetric distribution for the inter-arrival times. Both the normal and lognormal distributions, like the gamma distribution, have modal values away from the very short, nearly instantaneous time occurring with the exponential distribution. The results are depicted in Figure 3. The tabular distribution was the closest to the actual, but a slight overestimate. All three of the other distributions were overestimates by quite a significant magnitude. From these results it is clear that the validity of any network simulation involving packet traffic is dependent upon using the appropriate distribution. This process can be

Figure 3. Work station average delay observed by varying the packet inter-arrival distribution

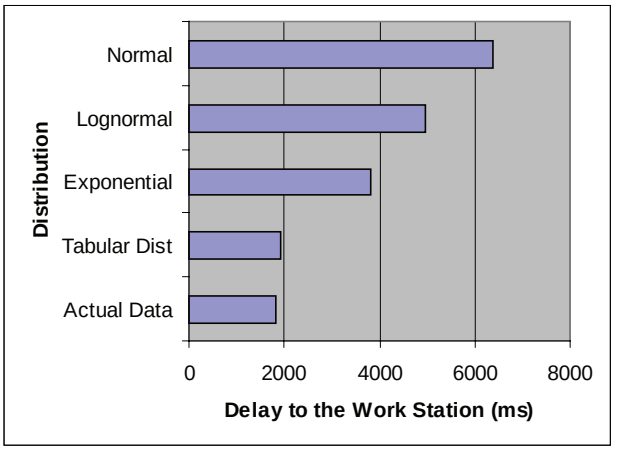


Figure 4. Frequency counts of packet inter-arrival times for Ethernet graphics data

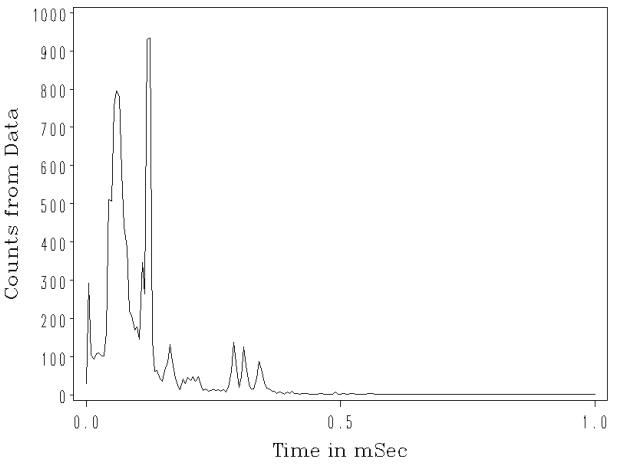


Table 1. Magnitude of initial and steady state values for text-based packets

Data Set	Throughput		Intensity	
	Initial	Steady-State	Initial	Steady-State
8 sessions	3,900	2,340	26	9.1
64 sessions	31,863	26,334	211	73.8
128 sessions	78,328	71,609	452	171.5

Table 2. Magnitude of initial and steady state values for graphics-based packets

Data Set	Throughput		Intensity	
	Initial	Steady-State	Initial	Steady-State
64 session	1458823	2000156	4705	2699
128 session	1751417	3003212	5649	4079

made more sophisticated by classifying the traffic and instead of using a single general distribution employing several specific distributions. For example, Guster, Sundheim, and Safonov (2005) classified traffic as either text- (telnet, ftp, ssh) or graphic- (http, https) based. As would be expected, they found different distributions and packet intensities when the two categories were compared. Figure 4 depicts the distribution of graphical oriented traffic, which can be compared to Figure 1 herein to provide a visual comparison of the two categories. In both cases spikes occur, but not the same number or in the same place. Furthermore, to provide some idea of the difference in intensity, Tables 1 and 2 depict the initial and steady state throughput in bytes for both text- and graphics-based distributions. In Table 1 the results of a workload generated by 64 Web client sessions is displayed. Whereas, in Table 2 the results of a workload generated by 128 Web client sessions is displayed. In the text data there is a very high intensity in the beginning of the session that tapers off after a steady state is reached. This is related to address resolution and forming connections. In the graphic data, it appears there is less "start up" traffic. However, it still takes place but is small in comparison to the massive

screen downloads that follow. A quick examination of the steady state throughput levels reveals that at the same number of client sessions the graphical data is about 70 times greater at the 64 session level and 40 times greater at the 128 session level.

CONCLUSIONS

As the need to make the most of available network bandwidth increases, the importance of having valid simulation techniques available to test optimization methodologies increases. The key to this validity is an inter-arrival distribution that is truly representative of the actual data. This article explored several alternatives in selecting this distribution. First, the actual data obviously offer the best accuracy but are often impractical due to their massive size. Furthermore, selecting a representative sample is often challenging due to widespread variation over time exhibited by most network traffic. Second, the exponential distribution (or any other known distribution), which according to queuing theory should be appropriate, does not fit the data well in a number of studies. Third, the power law process has limited value in relatively short time interval studies. It does, however, offer more promise in data sets involving very large time spans. Fourth, the Markov process has exhibited somewhat promising results. In fact, the visual plots (Figure 2) reveal a much closer fit than the Poisson model. However, Markov models have displayed limitations in the middle time range (.3-5 milliseconds), which make them far from a perfect choice (Guster, Litvinov, et al., 2002). Therefore, much additional work is needed. A simple visual inspection of Figure 1 reveals the complexity of the data and subsequently the difficulty in determining a mathematical model that would truly represent them. The results exhibited by the Markov process are encouraging, and studies that examine more sophisticated Markov related models should be encouraged. Fifth, recent work has shown that breaking the main distribution into several sub-distributions based on the application or service offers promise. For example, the difference in intensity between graphics and text services was illustrated herein in Tables 1 and 2. The importance of obtaining the appropriate distribution for network traffic should not be underestimated. Without it, simulations designed to ascertain network design efficiency, network performance, and the effectiveness of software optimization techniques generate invalid results.

REFERENCES

CACI. (1998). *Comnet III reference manual*. La Jolla, CA: CACI Products Inc.

Cruz, R. (1991). A calculus for network delay. *IEEE Transactions on Information Theory, 37,* 114-131.

Cruz, R. (1995). Quality of service guarantees in virtual circuit switched networks. *IEEE Journal of Selected Areas in Communication, special issue on Advances in the Fundamentals of Networking, 13*(6), 1048-1056.

Cruz, R., & Santhanam, A. (2000). A composable service model for lossy network elements. In *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS),* Geneva, Switzerland (Vol 4, pp. 97-100).

Cruz, R., & Tsai, J. (1996). COD: Alternative architectures for high speed packet switching. *IEEE/ACM Transactions on Networking, 4*(1), 11-21.

Firoiu, V., Le Boudec, J., Towsley, D., & Zhang, Z. (2002). Theories and models for Internet quality of service. *Proceedings of the IEEE, 90*(9), 1565-1591.

Fortier, P. J., & Desrochers, G. R. (1990). *Modeling and analysis of local area networks.* Boca Raton, FL: CRC Press.

Frieberger, W. (1972). *Statistical computer performance evaluation.* New York: Academic Press.

Guster, D., Litvinov, S., Richardson, M., & Robinson, D. (2002). A comparison of stochastic models for the interarrival times of packets in a computer network. In K. van Slooten (Ed.), Optimal information modeling techniques (pp. 248-257), Hershey, PA: IRM Press.

Guster, D., & Robinson, D. (1993, May). *The application of Box-Jenkins time series analysis to performance problems in computer networks.* Paper presented at the Information Resources Management Association Conference, Salt Lake City, UT.

Guster, D., & Robinson, D. (1994, April). *Markov chains as a predictor of performance decay in a PCbased LAN environment.* Paper presented at the Small College Computing Symposium, Winona, MN.

Guster, D., & Robinson, D. (2000, November). *Using Markov chains to analyze the inter-arrival distributions of ATM and Ethernet traffic in computer networks.* Paper presented at the 2000 DSI Annual Meeting, Orlando, FL.

Guster, D., Robinson, D., & Juckel, A. (2000, May). *Differences in the inter-arrival rate distributions between ATM and high-speed Ethernet and their implications on computer network performance.* Paper presented at the 11th Annual Information Resource Management Association International Conference, Anchorage, AK.

Guster, D., Robinson, D., & Richardson, M. (1999). Application of the power law process in modeling the inter-arrival times of packets in a computer network. In D. Dufner & O.

Kwon (Eds.), *Proceedings of the 30th Annual Meeting of the Midwest Decision Sciences Institute* (pp. 76-78). Springfield, IL: Decision Sciences Institute.

Guster, D., Safonov, P., Hall, C., & Sundheim, R. (2003). Using simulation to predict performance characteristics of mirrored hosts used to support WWW applications. *Issues in Information Systems, 4*(2), 479-485.

Guster, D., Sohn, C., Robinson, D., & Safonov, P. (2003). A comparison of asynchronous transfer mode (ATM) and high-speed Ethernet and the network design implications to a business organization. *Journal of Information Technology and Decision Making, 2*(4), 683-692.

Guster, D., Sundheim, R., & Safonov, P. (2005). Analysis of end-user services and their potential load on the network. *Journal of Academy of Business and Economcs, 4*(3), 56-73.

Karagiannis, T., Molle, M., & Faloutsos, M. (2004). A nonstationary Poisson view of Internet traffic. In *IEE INFOCON*. Hong Kong. Retrieved from http://www.caida.org/publications/papers/2004/infocom/

Karagiannis, T., Papagiannaki, K., & Faloutsos, M. (2005). BLINC: Multilevel traffic classification in the dark. *ACM SIGCOM* (pp. 229-240). Philadelphia, PA: ACM Press.

Krzenski, K. (1998). *Analysis of the predictive process request-response modeling in a hypermedia environment.* Unpublished master's thesis, St. Cloud State University, St. Cloud, MN.

Krzenski, K. (1999). *The effect of varying the packet inter-arrival distribution in the simulation of Ethernet computer networks.* Unpublished graduate research project, St. Cloud State University, St. Cloud, MN.

Le Boudec, J. (1998). Application of network calculus to guaranteed service networks. *IEEE Transactions on Information Theory, 44*(3), 1087-1096.

Partridge, C. (1993). The end of simple traffic models. Editor's note. *IEEE Network, 7*(5), 3.

Postigo-Boix, M., Garcia-Haro, J., & Melus-Moreno, J. (2005). Analytical model to optimize the cost of resource reservations in a client-server scenario. *IEICE Transactions on Communications, 88*(7), 2879-2886.

Robert, S., & Le Boudec, J. (1996). On a Markov modulated chain with pseudo-long range dependences. *Performance Evaluation, 27-28,* 159-173.

Vandolore, B., Babic, G., & Jain, R. (1999). *Analysis and modeling of traffic in modern data communications networks.* A paper submitted to the Applied Telecommunication Symposium. Retrieved from http://scholar.google.com/scholar?hl=en&lr=&q=Vandolore%2C+B.%2C+Babic%2C+G.%2C+%26+Jain%2C+R.+1999

Vojnovic, M., & Le Boudec, J. (2002). Stochastic bound on delay for guaranteed rate nodes. *IEEE Communications Letters, 6*(10), 449-451.

Walker, J. (2000). *Testing and tuning QoS for network policies.* Technical Paper. Net IQ Corporation. Retrieved from http://download.netiq.com/Library/White_Papers/TestingAndTuningQoSForNetworkPolicies.pdf

Zhu, X., Yu, J., & Doyle, J. (2001). *Generalized source coding and optimal Web layout design* (Tech. Rep. No. CIT-CDS-00-001). Pasadena, CA: Caltech CDS.

KEY TERMS

Inter-Arrival Distribution: The probability density function that describes likely and unlikely inter-arrival times for packets.

Inter-Arrival Time: The amount of time that elapses after the receipt of a packet until the next packet arrives.

Markov Chain: A model that determines probabilities for the next event, or "state," given the result of the previous event.

M/M/1 Model (Exponential/Exponential with One Server): The queuing model that assumes an exponential distribution for inter-arrival times, an exponential distribution for service times, and a single server.

Packet: A finite stream of bits sent in one block with header and trailer bits to perform routing and management functions.

Packet Intensity: The number of packets transferred per second across a computer network.

Power Law Model: A generalization of the classical Poisson model, allowing for changes in the intensity of the arrivals.

Quality of Service (QOS): A method of marking certain packets for special handling to ensure high reliability or low delay.

Reservation Bandwidth: Reserving a portion of the available bandwidth for a given protocol or application, which ensures its network access will not be adversely affected by massive network traffic generated by other protocols or applications.

Throughput: The number of bytes transferred per second across a computer network.

Evolution of Post-Secondary Distance Education

Iwona Miliszewska
Victoria University, Australia

INTRODUCTION

Distance education is an increasingly common educational alternative, as well as a key contributor to the newly competitive landscape in higher education. Once regarded as an experimental alternative outside mainstream university education, distance education has attained new levels of legitimacy and expansion and has grown into a higher education industry of its own. This article discusses the history and transformation of distance education to create a framework for the sequence of events that have contributed to the distance education movements and shaped modern post-secondary distance education programs.

The article outlines the evolution of post-secondary distance education from its inception to the present: its progression from informal programs offered by individual providers to a well-organised formal educational alternative; its purpose and characteristics; its expansion and internationalisation; and the various forces that have shaped its growth. While noting that technology has its limitations—it can facilitate teaching but not replace it—the article highlights the crucial role that advancements in technology have played in propelling the evolution of distance education, and points to the role of technology in blurring the conceptual divide between distance and traditional education.

BACKGROUND

Although there is no universal consensus on the origin of distance education, most researchers trace its roots to the emergence of correspondence education in the mid-nineteenth century in Europe (Great Britain, France, Germany) and the United States (Matthews, 1999; Peek, 2000; Ponzurick, France, & Logar, 2000). It was the English educator Sir Isaac Pitman who foresaw a need to deliver instruction to a student population that was limitless in comparison to the traditional classroom, and reach out to students in various locations (Matthews, 1999).

In the early years, distance education was dominated by individual entrepreneurs who worked alone; later, organised formal education institutions emerged, such as Sir Isaac Pitman Correspondence Colleges in Britain, and a school in Berlin to teach language by correspondence (Holmberg, 1995; *Simonson, Smaldino, Albright, & Zvacek,* 2000). At the same time, universities in Great Britain, such as Oxford and Cambridge, began to develop extension services. This university extension movement included not only traveling lectures, but also a system of correspondence education (Holmberg, 1995). In the United States, the earliest instance of distance education dates back to 1728 when an advertisement in a Boston newspaper offered weekly shorthand lessons by mail (Gilbert, 2001).

While initially, distance learning was envisioned as:

a way to serve students who lacked access to a complete education, whether due to insufficient resources, geographic isolation, or physical disabilities, it evolved to become a viable way to supplement programs and support innovation, rather than being merely a better-than-nothing alternative to doing without. (Weinstein, 1997, p. 24)

While some scholars identify Pitman as the initiator of correspondence education, other researchers recognise American educator William Rainey Harper as the pioneer of modern post-secondary correspondence teaching (Mood, 1995). Harper helped organise the Chautauqua College of Liberal Arts (New York), the first institution to receive, in 1883, official recognition of correspondence education; from 1883 to 1891, the college was authorised to grant academic degrees to students who successfully completed work through correspondence education and summer workshops.

DISTANCE EDUCATION: EVOLUTIONARY PERSPECTIVE

Growth in Distance Education Programs

The number of distance education programs has increased steadily from the mid-nineteenth century. For nearly 200 years, correspondence education was the primary means of distance education delivery, but in the late 1960s distance education reached a turning point with the introduction of a multimedia approach to its delivery; in addition to print, programs were also delivered through radio, television, audio, and video materials (Matthews, 1999).

In 1969, the Open University was established in the United Kingdom. This institution had a tremendous impact on distance education because it used a multimedia approach to teaching. The British Open University pioneered distance education on a massive international scale and, together with other open universities, helped raise the profile of distance education. For example, in Germany, the FernUnversität in Hagen was founded in 1974 (Matthews, 1999). However, the most dramatic growth of distance education programs has occurred from the 1980s until the present time. Since the mid 1990s, distance education programs have further transitioned into computer-based formats that enable the programs to be delivered fully or in part through the Internet. By the mid 1990s, nearly 25% of the colleges and universities in the United States offered degrees and certificates exclusively through distance education programs; the number grew to almost 58% 5 years later (Matthews, 1999).

In Asia, open and distance education has experienced an unparalleled growth, especially since the early 1990s, as the demand for learning in the region outstripped the capacity of traditional delivery methods by universities. Currently, almost all the countries in Asia have at least an open university. These universities command huge student populations, and of the 11 mega-universities (student enrolment of at least 100,000) worldwide, seven are located in Asia: China, India, Indonesia, South Korea, Thailand, Iran and Turkey (Shive & Jegede, 2001).

In Australia, in the past few decades, postsecondary education has developed an increasingly international orientation as the government encouraged universities to export their courses and import students (Marginson, 2004). Consequently, the export of Australian education now constitutes Australia's third largest services export after tourism and transport, and is judged to be Australia's fastest growing export sector (AVCC, 2005).

Forces Driving Distance Education

One of the major contributors to the dramatic growth of distance education has been technology. Advances in technology, including computer conferencing, interactive media, digital technologies, and the Internet have transformed the world into a borderless educational arena (Frantz & King, 2000). The new technologies significantly increase the reach of distance provision; they enable content to be current; they allow students to interact with instructors and with each other at any time; and, they open up a global market. The technologies not only offer new and better ways of communicating at a distance, but also have the potential to reduce the fixed costs of education (Taylor, 2001).

In addition to advances in technology, there are several other forces driving distance education including: the arrival of the Information Age, changing demographics, changing work and social patterns, declining government funding for further education, and competition in the educational market.

The transition from the Industrial Age to the Information Age has brought about appreciation of intellectual capital, which is now regarded as a valuable commodity. Cunningham et al. (2000) pointed out that the arrival of the Information Age heralded a new conception of knowledge. While previously, knowledge was of importance to an educated elite, and was applicable to a limited range of professions, its present cachet is much broader: it applies to a wide workforce, and it encompasses a variety of skills including "*thinking* skills, teaming capacity, and communication skills" (Cunningham et al., 2000, p. 21). Thus, knowledge workers represent a growing proportion of today's workforce, and the value of intellectual capital drives the demand for continuing education and emphasises a shortened lifespan of knowledge (Cunningham et al., 2000).

The explosion of knowledge, one of the consequences of the Information Age, also promotes distance education. There is a proliferation of new information: "in the past, information doubled every ten years; now it doubles every four years" (Aslanian, 2001, p. 6). It is no longer possible to *know everything*, even about one specialized discipline, so the aim of education must be *learning to learn* (Cunningham et al., 2000). Therefore, education can no longer be regarded as preparation for work, but rather as a lifelong effort to ensure employability rather than employment.

In view of these changing demands on the workforce, employees and employers alike increasingly regard adequate training as a valuable commodity; for employees "the opportunity for training is becoming one of the most desirable benefits any job can offer" (Cetron & Davies, 2005, p. 43); and, employers view "employee training as a good investment" (Cetron & Davies, 2005, p. 49). Thus, some of the changes underpinning the growing demand for lifelong learning "will demand short accelerated programs, well-suited for online delivery, and portfolio credentials" (Howell, Williams, & Lindsay, 2004); this, in turn, will drive the growing demand for distance education.

Changing demographics are also a driving force in distance education (Jones, 2001). High school leavers now represent only one type of tertiary student. Another type, increasingly growing in importance, is composed of adult learners, referred to by Cunningham et al. (2000) as *earner-learners*, who have paid jobs and seek postsecondary qualifications to maintain and enhance their careers. In addition, the importance of lifelong learning has shifted: it can no longer be regarded as a "discretionary personal investment; it has become an essential personal investment as people scramble to bolster their credentials in a volatile global work place" (Jones, 2001, p. 109). Lifelong learners represent a large and rapidly growing student body and demand relevant and accessible continuing professional development programs (Jones, 2001).

Because of the decreasing number of employees, and increasing demands on the ones that remain in the organisation, it has become increasingly difficult for employees to be released for training. This has sparked a trend to have educational programs delivered to companies, especially in global corporations. In Australia, the Coles Myer Institute is an example of a corporate education model. Established in 2003, it is a partnership between Coles Myer and Deakin University. The Institute provides Coles Myer employees, located across the organisation's 2,000 plus sites throughout Australia, with integrated vocational and professional development courses, and pathways to higher education awards (Walker, 2005).

Another factor contributing to the expansion of distance education is the rising cost of living and the tightening labor market: these factors have resulted in an increased number of two-income families. For many, sacrificing one income to return to studies is not an option. In addition, there is an increasing need to balance academic endeavors with work and family commitments. Thus, students with families and in the workforce demand programs that would fit their lifestyles; conventional time- and place-dependent education is not usually suitable. Bates (2000) points out that such students will particularly look for educational programs with *personally relevant content* that could be obtained through small specialized learning units.

Declining funds also drive distance education opportunities. Governments are increasingly reluctant to fund the growing demand for further education, so institutions of higher education are driven toward "for profit" education. They expect that students will be attracted to distance education programs, as they will be willing to pay for the opportunity to study while not being restricted by location or time (Bates, 2000).

Competition is another driving force. The corporate world sees the potential in the educational market and challenges universities by providing alternative programs to meet the rapidly growing demand. Middlehurst (2003) identifies the following categories of commercial provider and provision: corporate universities, private and for-profit providers, media and publishing businesses, and educational services and brokers.

Many corporations, especially large ones such as McDonalds, Ernst & Young, or Lufthansa, are developing corporate universities; at present, there are more than 2000 of such initiatives worldwide (Middlehurst, 2003). New private higher education institutions have also emerged recently on the distance education market as a result of a growing demand for foundation-level higher education (learners in the 18-25 age group), and for continuing and specialist education. These institutions usually provide programs in business, engineering, information technology, and teacher training to the niche market of working adults (Middlehurst, 2003; Ryan, 2002).

In addition, there has been a growth in the activities of commercial companies supporting online infrastructure of universities, including the Provincial Radio and TV Universities in China, or BBC's alliance with the Open University in the UK. Publishing companies, such as Pearson and Thomson Learning, are also involved in supporting educational providers, and developing new initiatives. While universities supply learning, assessment, and accreditation services, the publishers contribute their expertise in marketing, distribution, and content and electronic delivery systems (Middlehurst, 2003).

Finally, there has been huge growth in educational brokers over the recent years (Cunningham et al., 2000). The brokers, of whom Learnerdirect in the UK is an example, mediate between learners, companies and providers. They provide learners with access to study materials through conveniently located learning centers. Corporations are also promoting distance education course design, and course management tools (Middlehurst, 2003). This marketing effort further increases competition and applies additional pressure on the nonprofit university sector to provide distance education opportunities.

Effect of Technology on Distance Education

Over the years, changes in technology generated several significant milestones that affected the distance education market in terms of scale and delivery. Having examined the milestones, Sherron and Boettcher (1997) define four generations of distance education technologies according to five characteristics: (1) media and technologies, (2) communication features, (3) student characteristics and goals, (4) educational philosophy and curriculum design, and (5) infrastructure.

The first generation includes the period between early to mid twentieth century when print, radio, and broadcast television prevailed. Those media involved one-way communication as information was passed from teachers to students; there was no interaction among students, and minimal interaction between students and teachers. In addition, the radio and television broadcasts were time-dependent (Sherron & Boettcher, 1997).

The advent of the VCR and cable television in the early 1960s heralded the beginning of the second generation. The milestone that distinguished the second generation from its predecessor was the removal of time dependency: the broadcast portion of a distance education program was no longer tied to predetermined times (Sherron & Boettcher, 1997). In addition, videocassettes with their stop and rewind options gave learners control over the learning material: lectures could be interrupted and reviewed (Gunawardena & McIsaac, 2005). However, this generation still afforded

little interaction among students and between students and teachers.

The third generation arrived by the mid 1980s together with the personal computer and two-way videoconferencing. Two milestones separated this generation from the previous ones: one, the new technologies made it possible to communicate increasingly complex and large amounts of information to students; and two, they enabled interaction among students, and between students and teachers.

The growth in technological advancements accelerated significantly during the 1990s with the use of computer-mediated learning technologies; for example, two-way interactive video; Web-based asynchronous communication; and online or off-line Internet Web-based instruction (Ponzurick et al., 2000). It was also the beginning of the fourth generation. The fourth generation signified yet another milestone namely, increased interactivity among students, between students and teachers, and between students and content thanks to high-speed networks and more sophisticated software. Consequently, the amount and types of information that can be communicated significantly increased, and the exchange of information took significantly less time (Sherron & Boettcher, 1997).

The rapid dissemination of communication networks, computer-mediated learning technologies, and multimedia technologies in distance education provided the means for the introduction of more effective resources for learning than was previously possible. These technologies facilitated a shift in the relationship between teaching and learning in the distance education context: learning became the focus, rather than teaching; assisting students to become independent learners became the major goal; and the role of a teacher was transformed to that of a mentor and facilitator. This approach teaches learners how to learn, and provides them with the resources to enable them to learn in a manner which is relevant to their own intellectual and social circumstances (Gunawardena & McIsaac, 2005).

The current landscape of distance education includes a wide spectrum of technologies, spanning all generations. Although the technologies of the first generation have been surpassed by several other generations, they continue to play a considerable role in distance education. Radio and television are viable options in developing countries such as India and China where the infrastructure to support more recent technologies has yet to be developed (Middlehurst, 2003). In addition, print "remains a very important support medium for electronically delivered distance education" (Gunawardena & McIsaac, 2005, p. 365). In Thailand (e.g., at the Sukhothai Thammathirat Open University) print materials and audio cassettes are used as core media supplemented by educational radio and television programs, limited computer-assisted instruction, and face-to-face tutorials (Brahmawong, 2001). Even in the more developed Asian countries, for instance at the Korea National Open University, printed textbooks, satellite television, and audiocassettes remain the main teaching medium; videoconferencing systems, and the Internet are supplementary (In-sung, 2001). And, while much is being said about "potential uses of information and communication technologies, the major delivery methods for most distance education programs in Asia still rely heavily on print and the postal system, the "correspondence education" of yesteryear" (Shive & Jegede, 2001, p.11). In developed countries, radio and television are used extensively by institutions, such as the British Open University and FernUniversität in Germany, to deliver programs to a large number of learners (Gunawardena & McIsaac, 2005).

FUTURE TRENDS

The distinction between distance and traditional education will continue to fade due to advancements in interactive multimedia technologies such as automated response systems and interactive multimedia online, which allow for individualised and collaborative learning. These technologies also enable the creation of virtual communities in traditional settings. Consequently, all interactions with teachers, course content, learning activities, assessment, and support services will be delivered online even for campus-based students (Taylor, 2001).

Global connectivity and *networking* will replace *separation* and *distance* as the main characteristics of distance education. Students can increasingly participate in cooperative learning activities through computer networks. Thus, global classrooms may have participants from various countries interacting with each other at a distance (Gunawardena & McIsaac, 2005). Even a decade ago, Hall (1995) suggested that the descriptor *distance learning* was becoming less and less relevant with respect to distance programs and students, and proposed that *connected learning* might be a more accurate descriptor, reflecting the impact that technology has on distance education pedagogy.

Individualization and student-centred learning will increasingly underpin the distance education practice. This will be partly achieved through the provision of specially designed learning resources, with which the students will engage in an interactive manner. Consequently, the teacher's role will increasingly shift from one of a provider of information to one of a mentor and facilitator of learning. Thus, distance education teachers will discover that they need to develop a new set of skills if they are to be effective educators, which has obvious professional development implications.

Although technology is a central part of many distance education programs, it is important to remember that technology is just the method of facilitating connection and conveying content; technology is not the focus of the learning endeavour. According to Weinstein:

… the human touch cannot be delivered remotely. Distance learning technologies are intended to support an integrated program, not replace it. Balancing virtual and real interaction will be one of the key educational challenges as we enter the 21st century… (Weinstein, 1997, p. 25)

In short, any attempt to use technology to significantly reduce human contact within the learning environment would result in significant quality losses.

CONCLUSION

The expansion of distance education is set to continue to be driven by the growing needs of life-long learners, political and economic pressures, advances in technology, and changing conceptions of the nature of education. Distance education is evolving and changing so rapidly that no one can accurately predict its future. However, distance education seems to be moving away from delivering education at a distance, and toward providing a convenient educational alternative to students not even geographically separated from their teachers. The emphasis seems to be shifting away from the very term *distance* that used to define distance education, toward the term *connected* education, an education that connects learners with teachers, and learners with learners.

REFERENCES

Aslanian, C. (2001). Adult students today. New York: The College Board. Cited in *Paradigm, 13*(3). Retrieved December 8, 2007, from http://www.mercyhurst.edu/graduate/paradigm-newsletter-pdf/fall2001.pdf

AVCC (Australian Vice Chancellors' Committee). (2005, January). *Report*. Retrieved December 8, 2007, from http://www.avcc.edu.au/documents/publications/stats/International.pdf

Bates, T. (2000). *Distance education in dual mode higher education institutions: Challenges and changes.* Retrieved December 8, 2007, from http://bates.cstudies.ubc.ca/papers/challengesandchanges.html

Brahmawong, C. (2001). Thailand. In O.J. Jegede & G. Shive (Eds.), *Open and distance education in the Asia Pacific Region* (pp. 220-236). Hong Kong: Open University of Hong Kong Press.

Cetron, M.J., & Davies, O. (2005). Trends shaping the future: Technological, workplace, management, and institutional trends. *The Futurist, 39*(3), 37-50.

Cunningham, S., Ryan, Y., Stedman, L., Tapsall, S., Bagdon, K., Flew, T., & Coaldrake, P. (2000). *The business of borderless education* (pp. 18-23). Canberra: DETYA.

Frantz, G., & King, J. W. (2000). The distance education learning systems model (DEL). *Educational Technology, 40*(3), 33-40.

Gilbert, S. (2001). *How to be a successful online student.* New York: McGraw-Hill.

Gunawardena, C.N., & McIsaac, M.S. (2005). Distance education. *Association for Educational Communications and Technology*, Chapter 14, 355-395.

Hall, J. (1995). The convergence of means. *Educom Review, 30*(4), 42-45.

Holmberg, B. (1995). *Theory and practice of distance education* (2nd ed.). London: Routledge.

Howell, S.L., Williams, P.B., & Lindsay, N.K. (2004). Thirty-two trends affecting distance education: An informed foundation for strategic planning. *Online Journal of Distance Learning Administration, 6*(3). Retrieved December 8, 2007, from http://www.emich.edu/cfid/PDFs/32Trends.pdf

In-sung, J. (2001). Korea. In O.J. Jegede & G. Shive (Eds.), *Open and distance education in the Asia Pacific Region* (pp. 103-130). Hong Kong: Open University of Hong Kong Press.

Jones, G.R. (2001). Bridging the challenges of transnational education and accreditation. *Higher Education in Europe, 26*(1), 107-116.

Marginson, S. (2004). National and global competition in higher education. *The Australian Educational Researcher, 31*(2), 1-28.

Matthews, D. (1999). The origins of distance education and its use in the United States. *The Journal (Technological Horizons in Education), 27*(2), 54-61.

Middlehurst, R. (2003). The developing world of borderless higher education: Markets, providers, quality assurance and qualifications. In *Proceedings of the Conference on the First Global Forum on International Quality Assurance, Accreditation and the Recognition of Qualifications*. Paris, UNESCO, (pp. 25-39).

Mood, T.A. (1995). *Distance education: An annotated bibliography*. Eglewood, CO: Libraries Unlimited.

Peek, R. (2000). A distance learning reality check. *Information Today, 17*(2), 30.

Ponzurick, T.G., France, K., & Logar, C.M. (2000). Delivering graduate marketing education: An analysis of face-to-face versus distance education. *Journal of Marketing Education, 22*(3), 180-187.

Ryan, Y. (2002). Emerging indicators of success and failure in borderless higher education. *Observatory on Borderless Higher Education*. Retrieved December 8, 2007, from http://www.obhe.ac.uk/products/reports/pdf/February2002.pdf

Sherron, G., & Boettcher, J. (1997). *Distance learning: The shift to interactivity*. CAUSE Professional Paper Series #17. Boulder, CO: CAUSE. Retrieved December 8, 2007, from http://www.educause.edu/ir/library/pdf/PUB3017.pdf

Shive, G., & Jegede, O.J. (2001). Introduction. In O.J. Jegede & G. Shive (Eds.), *Open and distance education in the Asia Pacific Region* (pp. 1-26). Hong Kong: Open University of Hong Kong Press.

Simonson, M., Smaldino, S., Albright, M., & Zvacek, S. (2000). Teaching and learning at a distance: Foundations of distance education. Upper Saddle River, New York: Prentice Hall.

Taylor, J. C. (2001, June). Fifth generation distance education. Higher education series (Rep. No. 40). Canberra: DETYA.

Walker, S. (2005). Keynote address: Education sector response. *CEDA conference, lifelong learning: Challenges of an aging workforce*. Retrieved December 8, 2007, from http://www.deakin.edu.au/vc/presentations/CEDA-presentation-11.pdf

Weinstein, P. (1997). Education goes the distance: Overview. *Technology & Learning*, *17*(8), 24-25.

KEY TERMS

Asynchronous: Communication in which interaction between parties does not take place simultaneously.

Distance Education: The separation of student and learner in space or time, the use of educational media to unite teacher and learner and carry program content, and the provision of two-way communication between teacher, tutor, educational institution and the learner.

Educational Program/Course: A set of units/subjects, that lead to an academic qualification, for example, a degree.

Multimedia: Any document that uses multiple forms of communication, such as text, audio, or video.

Network: A series of points connected by communication channels in different locations.

Online: Active and prepared for operation; also suggests access to a computer network.

Two-Way Communication: A form of transmission in which both parties are involved in transmitting information: common forms include telephone conversations, instant messaging, and computer chatroom communication.

Executive Judgment in E-Business Strategy

Valerie Baker
University of Wollongong, Australia

Tim Coltman
University of Wollongong, Australia

INTRODUCTION

One of the main strategic challenges for organizations today is to effectively manage change and stay competitive in the future. Change appears to be the only constant in contemporary business and is present in every industry and in every country (Brown & Eisenhardt, 1998). Moreover, the key area of importance, current within many organizations, is how to effectively leverage technology within such a complex and dynamic business environment (Sauer & Willcocks, 2003). The alignment or fit approach, which has its roots in contingency theory, has long been promoted as the way to get high returns from technology investment. However, the realization of advantage from the Internet and related e-business technology investment has long been a source of frustration for corporate executives. Impressive performance returns by companies such as Dell Computers, Cisco Systems and General Electric illustrate that returns can be achieved by linking the Internet and related e-business technologies to firm strategy. These companies have shown that successful management of their IT investments can generate returns as much as 40% higher than those of their competitors (Ross & Weill, 2002). Yet, many executives view the Internet and related e-business technologies with intense frustration. They recollect investment in the great speculative bubble of the 1990s and excessive expenditure on year 2000 (Y2K) compliant systems (Keen, 2002). They recall high profile examples of botched enterprise resource planning (ERP) systems that have consistently run over time and budget and report that customer-relationship management (CRM) initiatives were largely a flop (Reinartz & Chugh, 2002). Unfortunately, it is not yet clear how firms should go about capturing the potential that exists in e-business, as few normative frameworks exist to guide practitioner investment.

BACKGROUND

One area of scholarly activity where consistent advances have been made regarding the determinants of firm performance is in structural contingency theory. Here, the contingency factor (i.e., environment-structure) has enabled predictions to be made in a relatively unambiguous manner (Donaldson, 1995). Applied to an e-business setting, contingency theory argues that performance increases can be expected whenever information technology is applied in an appropriate and timely way, in harmony with business, environmental and organizational conditions. Consider a typical scenario where an executive wants to make a strategic investment in information systems. They have two choices: (1) a system to support backend operations using ERP technology, and (2) a CRM support system. How do they prioritise between these competing investments? Contingency literature would argue that it depends upon the organization's strategy and decision-making information requirements (Chandler, 1962; Child, 1972; Galbraith & Kazanjian, 1986). Manufacturing excellence strategies associated with companies like Carrefour or Ford Motor Company would get greater value from ERP systems. Customer intimacy strategies at companies like CitiBank or IBM Global Services would benefit most by customer feedback systems.

As simple as this observation may appear, the application of alignment has proven elusive. Despite 20 years of effort and investment in consulting advice, CIOs are still struggling with the same set of alignment problems. A recent survey by CIO Insight (Patterson, 2001) highlights the point that only 34% of organizations considered the link between their IT priorities and their enterprise strategy to be "strong." While these statistics reflect the difficulties of coordinating complex organizations, they provide evidence that most managers are not using the basic tools of alignment that have been developed over several decades of research.

Priem and Cycyota (2000) equate the process of alignment between IT strategies and business goals with executive judgment. The literature regarding judgment theory argues that firm success can be explained by the judgments executives make concerning the current state of the environment and the vision of the organization. In uncertain times, where market pressures and time constraints dominate the business landscape senior manager's perceptions, skill and vision often form the basis on which strategic choices regarding IT investments are made. For example, it takes little more than a browsing of the management section of the local bookstore—blazoned with titles such as Inside the Minds: Leading CEOs—or a visit to the local news agent to pick

up a recent copy of Forbes, Fortune or Business Week to recognize the importance that publishers and managers place on the philosophies and actions of even some of the least successful or most unlikely of management leaders. Perhaps more relevant is that often the appointment of "higher quality CEOs" leads to immediate stock market reactions and greater long term performance. One such example was the reappointment of Steve Jobs as CEO of Apple Computer. Jobs has been widely praised for his skill in judging the commercial potential of convergent Internet technologies and his return to the company was considered instrumental in its reversal of bad fortunes (Stevens, 1997).

The corollary here is that judgment is an essential skill for setting the overall direction of the organization. In turbulent environments, often the context of e-business, quick trade offs need to take place, as the strategic direction of the firm enables it for the future. This being the case management discretion becomes increasingly important, as decisions are made "on the fly" with little information or understanding of the decision problem. Management play a vital role in "trading off" elements of organization control, that is, structure for better adaptation, a view supported by complexity theory (Brown & Eisenhardt, 1998). This theory views strategy as a process which constantly changes, and thus needs a type of structure or execution method that is dynamic and will allow the organization to be ready for the future.

Thus, although judgment appears to be important to organizational success, scholars have largely ignored executive intentions and no empirical link between executive choices and firm outcomes has been established. Instead, strategic outcomes are presumed to be due to strategic choice (Preim & Harrison, 1994). This omission may account in some part, for why practitioners continue to pay little attention to the large amount of published work concerning the antecedents of strategy and performance. This concern provided the motivation for a special issue of the Academy of Management Journal (AMJ, 1998, p.746) that sought greater understanding of the way knowledge is transferred between academics and practitioners. The issue again surfaced in a recent issue of the Academy of Management Executive, providing evidence that practitioners still typically turn to sources of information other than academics or the scientific literature when searching for ways to improve performance (Ford, Duncan et al., 2003).

FUTURE TRENDS

Clearly, we need greater understanding of the conditions which lead executives to make strategic choices if we are to develop research that has an impact on practitioners.

Existing research into the change process and the implementation of e-business related technology is limited because it fails to measure the link between strategic choices and firm outcomes. As we have suggested, the judgments that executives make provides important insight into how IT strategic change or e-business change is approached given different situations and organization contexts.

Peterson (2002) suggests that it is the processing of information and the judgments that are made by top management that leads to critical decisions being made about how firms deal with IT-related strategic change. As the business environment rapidly changes, the variance in possible outcomes ranges from failure to unparalleled success. These differences can largely be explained by the "mythical relationship between technology ecology, human nature, decision cycles, IT and the speed and veracity of their interactions" (Peterson, 2002, p.485). Executives process information about these relationships and form critical strategic judgments regarding the future direction of their organization through its e-business strategy.

Managers face conditions such as dynamic markets, casual ambiguity and path dependence that make it extremely difficult to predict the outcomes of their IT strategic investments. As this illustration suggests, it is imperative that managers have in place strategies to cope with changes as they occur. Faced with external environmental changes (e.g., new rates of Internet adoption, killer mobile commerce applications, etc.), managers need to be able to adjust their strategic choices accordingly "just as water shapes itself according to the ground, an army should manage its victory in accordance with the enemy. Just as water has no constant shape, so in warfare there are no fixed rules and regulations" (Sun Tzu in Hussey, 1996, p.208)

What Sun Tzu highlights is the requirement that strategies be flexible in order to manage strategic change. Mintzberg, Ahlstrand, and Lampel (1998) describe this as an emergent strategy, where rather than pursuing a strategy, an organization makes decisions based on the situation, effectively testing the market as they go.

Thus strategic decisions regarding IT management need to be a mixture of both deliberate and emergent strategies. "Real-world strategies need to mix these in some way: to exercise control while fostering learning" (Mintzberg et al., 1998, p.11). The importance of strategic alignment between the organization and its environment becomes even more critical given recent environmental turbulence and the evolving importance of technology and e-business to competitive advantage.

Executive Judgment and Strategic Alignment

Priem and Cycyota (2000) state that understanding judgments by strategic leaders is essential to determine the role of mental processes in strategy development and how these strategies and processes affect firm performance. They suggest that a number of theoretical platforms commonly found in the strategy literature provide a solid platform from which we

can examine strategic judgement. The "fit" or "alignment" paradigm is perhaps one of the most pervasive in strategy. Good strategy requires at a minimum alignment with changing external conditions. In simple terms, the proposition is that there is an organizational structure that fits the level of contingency factor whether it is environmental uncertainty, organizational characteristics, technological characteristics or strategy design interdependence so that an organization in fit creates significant and positive implications for performance. This idea that fit between organization structure and contingency factor leads to superior performance has been empirically supported in both qualitative and quantitative studies (Donaldson, 1995). Given this distinguished history, it might reasonably be expected that executives would frequently make decisions based on the principles of organization congruence (Priem, 1994).

Most early theories of structural contingency focused on how the fit between bivariate variables (i.e., structure-environment alignment, strategy-structure alignment or strategy-environment alignment) are associated with increased firm performance. However, information technology (IT) is becoming an important substitute for organization structure in modern organisation (Sauer & Willcocks, 2003). For example, Oracle's ability to transform itself into an Internet-enabled business would not have been possible without an appropriate technology base. The wrong technology base would have made such an initiative a massive technological and organizational challenge because of the custom integration required. Emerging evidence indicates that structure and technology are complementary. Where structures create boundaries for management control, technology permits those boundaries to be traversed thereby enabling more complex commercial activities to be effectively integrated and managed.

EMPIRICAL INVESTIGATION OF STRATEGIC JUDGMENT

The "integrative framework" developed by Lee (1989, 1991) in a series of papers regarding the management of information systems provides a suitable approach to the study of judgment. Lee's integrative framework formally presented in his 1994 paper combines three levels of "understandings": the subjective, the interpretive and the positivist. According to Lee, the three understandings are "far from being mutually exclusive and irreconcilable"; in fact, "they may be utilised as mutually supportive and reinforcing steps in organisational research". Priem and Cycyota (2000) also support this view by claiming that both qualitative and quantitative studies are necessary to increase our understanding of strategic judgement.

In the case of e-business, qualitative work can be useful in exploratory investigation that may highlight issues more formal approaches may miss. For example, case studies of the most spectacular strategic information systems initiatives Baxter Healthcare and American Airline's SABRE indicate that these IT/e-business systems were largely accidental success stories (Clemons, 1986). However, these subjective and interpretive studies cannot test hypotheses adequately, because of the close contact needed with research subjects and the resulting small sample size. Quantitative studies provide the crucial positivist link that complements exploratory work in a way that can generate more widely generalizable insights. Notable examples, include the study of IT's contribution to performance in the retail industry (Powell & Dent-Micallef, 1997), and the investigation of organizational antecedents to e-business adoption (Srinivasan, Lilien, & Rangaswamy, 2002).

The following sections focus on measurement techniques, which can be used to examine individual judgment. These techniques can be grouped into two categories: (1) composition methods, and (2) decomposition methods (Priem & Harrison, 1994).

Composition Methods

Composition methods focus on the processes that underlie individual judgments. Composition involves methods such as verbal protocol analysis, information searches and cause mapping to gather interpretive information from executives about the processes that lead them to make certain judgments. These types of techniques would be useful in identifying the variables that executives use in their strategic decision making, but which are not included in current management theory (Priem & Harrison, 1994).

Decomposition Methods

Decomposition techniques focus on the interactions that take place surrounding the judgment itself. The technique requires that the variables or judgment attributes be known a priori. The substantive nature of those variables must come from existing strategy theory (Priem & Harrison, 1994), and contingency theory provides an excellent starting point.

In this case, decomposition methods are required to focus on executive choices in response to a series of decision scenarios (i.e., behavioural simulations regarding the environment, firm structure and the strategy making process). The variance in executive choice is evaluated against these factors of interest, which can be manipulated across scenarios, using conjoint or choice analysis techniques.1 Figure 1 shows the way we can manipulate the important choices outlined in structural contingency theory. Paired comparisons (or stated preferences) are collected and then used to evaluate direct and interaction effects. In this way, we reveal the direction and strength of the three factors considered central to contingency theory. For example, an executive faced with a stable environment would lean towards

a planned strategy and decentralized structure according to contingency prescriptions. Respondent rankings on each path reflect the perceived utility respondents have for each combination of variables.

This combination of composition and decomposition techniques enables one to test whether the prescriptions of at least one well-known theory (i.e., contingency theory) influences executive judgment. The extent to which these prescriptions are already "obvious" to, or widely known by, practising executives will shed new light on the role of judgment in IT strategy and change.

CONCLUSION

We have outlined the importance of executive judgment to the strategic choice process and its particular relevance to the study of e-business, where environmental turbulence increases the relevance of managerial discretion. By separating the outcomes from the actual decision choice, we can begin to more fully understand how these strategic choices influence firm performance. Until now the process of strategic choice has largely been treated as a "black box" where it is assumed that measured outcomes are the result of deliberate choices.

One of the reasons contingency theory has become so popular is that it provides managers with prescriptive advice regarding which configurations lead to higher performance. Further examination of executive judgments to ascertain whether executives are making decisions based on the idea of alignment or fit is required. This will help in our understanding of whether material taught by academics in business schools is actually being used by students in industry. Are executives using the ideas of alignment or fit as prescribed in theory or are they making judgments based on other factors? The answer to this question has important implications for relevance and improving the linkage between theory and practice.

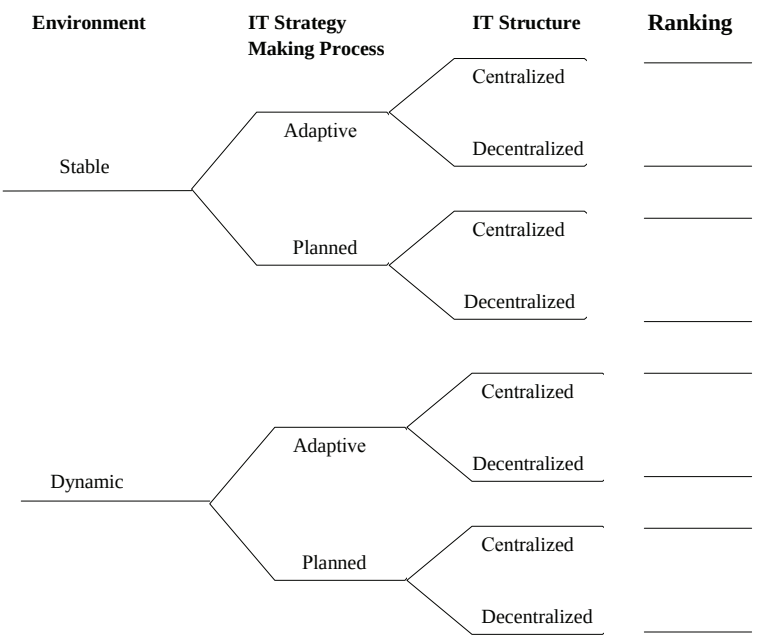
Understanding the processes that occur in strategy development will lead to greater knowledge of the decisions that executives make in uncertain environments and hyperturbulent contexts. This understanding is important if we want to develop e-business related research that is applicable to practitioners. It is this type of research that will guide executives in the strategic management of change and allow them to gain advantage from leverage their investments in e-business technology.

REFERENCES

A Special Research Forum Call for Papers: Knowledge Transfer between Academics and Practitioners. (1998). *Academy of Management Journal, 41*(6), 746.

Bharadwaj, A. (2000). A resource based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly, 24*(1), 169-196.

Figure 1. Judgment evaluation survey



Brown, E., & Eisenhardt, K.M. (1998). *Competing on the edge strategy as structured chaos*. Boston: Harvard Business School Press.

Chandler, A. (1962). *Strategy and structure*. Cambridge: M.I.T. Press.

Child, J. (1972). Organisation structure, environment and performance. *Sociology, 6*, 1-21.

Clemons, E.K. (1986). Information systems for sustainable competitive advantage. *Information & Management, 11*(3), 131-137.

Donaldson, L. (1995). *Contingency theory.* Aldershot, England: Dartmouth Publishing Company

Ford, E.W., Duncan, J.W. et al. (2003). Mitigating risks, visible hands, inevitable disasters, and soft variables: Management research that matters to managers. *Academy of Management Executive, 17*(1), 46.

Galbraith, J.R., & Kazanjian, R.K. (1986). Organizing to implement strategies of diversity and globalization: The role of matrix designs. *Human Resource Management, 25*(1), 37.

Henderson, C., & Venkatraman, N. (1999). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal, 38*(2&3), 472-482.

Hussey, D. (1996). A framework for implementation. *The Implementation Challenge,* New York: John Wiley & Sons

Keen, P., (2002). Getting value from IT. *Sydney University*, 19 August.

Lee, A.S. (1989). A scientific methodology for MIS case studies. *MIS Quarterly 13(1*), 33-50.

Lee, A.S. (1991). Integrating positivist and interpretative approaches to organisational research. *Organization Science, 2*(4), 342-365.

Mintzberg, H. (1994). *The rise and fall of strategic planning*. Prentice Hall.

Mintzberg, H., Ahlstrand, B., & Lampel, J. (1998). *Strategy safari: A guided tour through the wilds of strategic management*. New York: The Free Press.

Patterson, S. (2001). The truth about CRM, CIO Magazine, May 1st *http://www.cio.com/archive/050101/truth. content. html*

Peterson, J.W. (2002). Leveraging technology foresight to create temporal advantage. *Technological Forecasting and Social Change, 69*, 485-494.

Powell, T.C., & Dent-Micallef, A. (1997). Information technology as competitive advantage: The role of human, business, and technology resources. *Strategic Management Journal, 18*(5), 375-405.

Priem, R.L. (1994). Executive judgment, organizational congruence, and firm performance. *Organization Science, 5*(3), 421-437.

Priem, R.L., & Cycyota, C. (2000). On strategic judgement. In M. Hitt, R. Freeman & J. Harrison (Eds.), *Handbook of strategic management*, Blackwell.

Priem, R.L., & Harrison, D.A. (1994). Exploring strategic judgment: Methods for testing the assumptions of prescriptive contingency theories. *Strategic Management Journal, 15*(4), 311-324.

Reinartz, W.J., & Chugh, P. (2002). Learning from experience: Making CRM a success at last. *International Journal of Call Centre Management*, April, 207-219.

Ross, J.W., & Weill, P. (2002). Six decisions your IT people shouldn't make. *Harvard Business Review, 80*(11), 84.

Sauer, C., & Willcocks, L. (2003). Establishing the business of the future. *European Management Journal, 21*(4), 497-508.

Srinivasan, R., Lilien, G.L, & Rangaswamy, A. (2002). Technological opportunism and radical technology adoption: An application to e-business. *Journal of Marketing 66*(3), 47-61.

Stevens, A. (1997). Deja blue. *Industry Week, 246*(21), 82-88.

Teece, D. J., Pissano, G., & Sheun, A. (1997). Capabilities and strategic management. *Strategic Management Journal, 18*(7), 509-533.

KEY TERMS

Contingency Theory: A meta-theory, which argues that firm performance is defined by the environment-strategy-structure relationship, where the organization's strategy is contingent on the external environment and the organization structure is contingent on the firm's strategy.

E-Business Technology: Any technology, which enables an organization to conduct business electronically, with the overall aim of improving firm performance.

Executive Judgment: A decision that an executive makes, when they do not have a full understanding of the decision problem, based on their mental models of the environment and their vision for the organization.

External Environment: Factors which are external to an organization, such as new technology or product developments, changing rates of market growth, which an organization must respond to.

Fit/Alignment: Terms used to explain the relationship between IT and strategy. The IT strategy should work in synergy with the organizations strategy. These terms have their roots in the meta-theory contingency theory.

Strategic Choice: The choices that executives make which impact on the strategic direction of the organization. These choices exist as the intended strategies of the organization.

Strategic Decision-Making: The process of making important decisions (usually made by the top management team) to put executive choices into action by implementing strategies.

ENDNOTE

[1] There are a number of conjoint analysis methods, which can be used to test executive judgments. Each of the methods uses a variation of regression to decompose an executive's judgment. The most appropriate for evaluating executive judgments is metric conjoint analysis. For example, Priem (1994) used metric conjoint analysis to examine the judgments of CEO's in manufacturing firms. The outcome of this research was that the executives in manufacturing firms often make contingent judgments, regarding key strategy variables.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1149-1154, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Explicit and Tacit Knowledge: To Share or Not to Share

Iris Reychav
Bar-Ilan University, Israel

Jacob Weisberg
Bar-Ilan University, Israel

INTRODUCTION

The question of whether or not it is "worthwhile" for employees to share their knowledge has received a great deal of attention in the literature, which focuses on the technological factors that motivate knowledge sharing (Duffy, 2000). However, the ethical aspect regarding the question of knowledge ownership is discussed in only a partial way in Wang's (2004) model, where he examines employees' desire to share (or not to share) the knowledge they possess. This internal conflict is based on employees' having to choose between their own personal interests and their ethical understanding about organizational ownership of all employee-based knowledge. This article will elaborate on and examine the implications of knowledge sharing at the individual level. Employees, who manage to find the balance between their own personal interests and their ethical understanding about organizational ownership of employee-based knowledge, will engage in a high rate of knowledge sharing activities in the organization.

Goals of Managing Organizational Knowledge Sharing. An organization's desire to manage its knowledge sharing activities is based on the need to capture, catalog and store the organization's knowledge and transform it into knowledge that is both easily and immediately accessible to the organization and its members (Gupta & Govindarajan, 2000). The goal of knowledge sharing is to support and encourage the creation, transference, application and use of knowledge within the organization (Reychav & Weisberg, 2005). Scholars, researchers and practitioners alike express an increasing interest in the subject of organizational knowledge sharing between the individual employee and the organization, and among employees themselves (Almashari, Zairi, & Alathari, 2002).

Types of Knowledge. One of the classifications of organizational knowledge differentiates between two types of knowledge: *explicit knowledge* and *tacit knowledge* (Polanyi, 1958); *explicit knowledge* represents the knowledge that is accessible to all organization employees, while *tacit knowledge* represents the personal knowledge possessed by individual employees. Organizations seek to obtain employees' tacit knowledge and convert it into explicit knowledge, which can then be easily transferred to the organization's technological systems and networks. In this manner, the knowledge is distributed throughout the entire organization (Inkpen & Dinur, 1998; Ruppel & Harrington, 2001), thereby increasing the organization's human capital (its employees).

Conflicts of Interest. Organizations invest in developing their human capital (Nahpiet & Ghoshal, 1998). As a result, employees expand their knowledge and expertise in order to create a personal competitive advantage within the organization and the market (Carlile, 2002). Knowledge is a resource and individuals who possess knowledge use it to acquire positions of power and control both within the organization and outside of the organization. Therefore, organizations that attempt to gain their employees' knowledge (mainly of the tacit type) and make it accessible may, in the process, create a conflict of interests between the individual who possesses the knowledge and the organization that is interested in acquiring this knowledge (Storey & Barnett, 2000).

Hence, the main question is: Why would employees be motivated to share their personal knowledge with the organization at the risk of losing their relative power and advantage over the organization and the market? This question is even more complicated in light of the employee's other conflicting considerations: the understanding that the organization has ownership rights over the personal knowledge the employee acquires while employed by the organization, conflicting with employees' desire to realize their own personal interests by achieving a position of power/status.

The advancement and development of information and communication technologies have expanded organizations' formal capabilities regarding knowledge transference (Jarvenpaa & Staples, 2001). These technologies have also served to encourage the organization to centralize and control their information, based on the perception that knowledge belongs to the organization (Brynjolfsson, 1994). However, as soon as the organization is interested in and begins to operate according to norms and procedures, including the use of technological tools to transfer organizational knowledge, conflicts may arise within the organization at the employee level. These conflicts most often involve employees who possess a significant amount of organizational knowledge.

These employees may experience a conflict when they realize that they have to disseminate and store their knowledge within the organization in order to achieve a competitive advantage in the market at the risk of losing their own personal power and status within the organization (Jarvenpaa & Staples, 2001).

On the one hand, the accepted norm regarding knowledge sharing in organizations claims that all knowledge, such as ideas, processes, innovations, documentation, and computer programs developed and created by employees during their time of employment with the organization, belongs to the organization rather than to the individuals who initiated it (Constant, Kiesler, & Sproull, 1994).

This concept isn't stated in any contract or agreement signed between the employee and the organization, but is implied and understood through the organization's ethical values system, based on the idea that organizational knowledge is an asset that belongs to the organization.

On the other hand, the tacit knowledge residing within the sole control of an individual employee is considered personal knowledge of a specific type, as it is based on the subjective understanding, intuition, feelings, ideals, experience, values and emotions rooted in the individual (Polanyi, 1966, p.7).

BACKGROUND

Explicit knowledge, characterized by structured and constant knowledge, may be documented and distributed through technological systems and networks (Duffy, 2000; Martensson, 2000). Technological tools, such as the Internet, computerized libraries, documentation systems, and electronic group applications, all contribute to the transfer of knowledge within organizations (Tampose, 1996). However, the main contribution of these technological tools is to increase the compatibility of diverse organizational factors by reducing physical and personal constraints (DeLong, 1996). Even so, the existence of technological tools in an organization doesn't necessarily mean that employees will utilize them for the purpose of sharing knowledge. The extent of the actual use of technological tools depends upon the extent of the employees' motivation to utilize the technology (O'Dell & Grayson, 1998). Therefore, organizations have begun to understand that technology isn't a complete solution to knowledge sharing problems, because the way to promote knowledge sharing is to focus on the direct factors that effect employee behavior (Poole, 2000).

Ethics and Knowledge Sharing. Employees' ethical perceptions regarding the organization's ownership of its employees' knowledge affects their tendency to share their explicit knowledge with the organization. On the other hand, employees' willingness to share their tacit knowledge is based on their personal interests and the social and economic benefits they receive in exchange for sharing knowledge (Constant et al., 1994). Employees' transfer of tacit knowledge in the organization, based on their experience and expertise, may stem from a variety of considerations regarding personal interests, according to which employees believe that their explicit knowledge belongs to themselves, rather than to the organization. Organizations frequently engage in knowledge transfer that relates to technical aspects of the organization; this serves as an example of standards or documents that characterize products or services produced by the organization. This type of knowledge transfer is an accepted ethical activity of knowledge owned by the organization.

Knowledge: Organizational or Employee Ownership? The term "ownership" is widely used in the fields of law (Boyer, 1981), philosophy (Locke, 1978) and psychology (Markus, 1984). The distinction between organizational and employee ownership of knowledge was first presented by Jarvenpaa and Staples (2001).

1. **Organizational ownership of knowledge and expertise-related assets:** An employee's understanding regarding organizational ownership of knowledge-based assets mainly relates to the employee's explicit knowledge, which may be easily identified by the organization. Social Identity Theory can explain employees' ongoing behavior regarding the transfer of knowledge to the organization as being based on employees' desire to fulfill the goals of the organization they belong to (Tyler, 1999). Employees view their colleagues as necessary sources of information upon which completion of organizational tasks depend. Therefore, coworkers turn to one another in order to receive direction and guidance. Employees' feelings of belongingness and identification about their organization affect their level of knowledge sharing within the organization.
2. **Employee ownership of knowledge and expertise-related assets:** Employees' willingness to share their knowledge with the organization can be explained by two classical psychological theories:
    a. **Association theory:** Numerous psychological studies dealing with employee ownership of knowledge are based on Association Theory (Heider, 1958), which states that the individual who worked to create the knowledge asset and who, ultimately, controls the knowledge in the present time and also in the past, is the one who "owns" that knowledge asset. This theory suggests that because employees create the knowledge, they also own that knowledge. A good example of this differentiation between ownership and knowledge is conceptualized by

the subject of ownership rights and organizations' intangible assets. In the U.S., intangible assets are listed under the name of the individual employee who created the idea; however, the organization reserves the right to use this asset freely, without incurring any additional obligation toward that employee. From a legal standpoint, the organization's use of the knowledge asset in no way undermines the ownership rights of employees who have created that knowledge. On the contrary, employees' personal ownership of the knowledge they possess or create serves to increase their own feelings of self-worth as well as their sense of value and contribution to the organization.

b. **Exchange theory:** The idea of exchange relationships among organizational members was first presented in Homan's (1958) study. This study defined a *social exchange relationship* as a tangible activity or activities between at least two parties, which cannot be clearly identified and is based on the receipt of mutual reward. Furthermore, the idea of exchange relationships became a leading principle in psychological theory (Thibaut & Kelley, 1959) and was used to predict behavior by noting the behavioral characteristics of parties participating in exchange relationships.

Additional studies expanded the idea of exchange relationships to a more general approach, which included relationships between an individual and others within the same social network (Emerson, 1962). The idea of "exchange" crystallized into what Blau (1964) called Exchange Theory, which discusses employees' expectations to receive economic and social benefits as a reward for their contribution in exchange relationships with other organization members (Blau, 1964).

The Theoretical Model. According to the Association Theory described above, it seems that the employees, who are the owners of the knowledge, will hesitate to share their knowledge with the organization. However, from the employees' point of view, understanding the results that stem from sharing their knowledge may serve to increase their knowledge sharing behavior in order to achieve maximum benefits, in accordance with Blau's Exchange Theory.

This article presents a two-phase, multidimensional model (see Figure 1), which describes the considerations that motivate employees' willingness to share knowledge, actual knowledge-sharing activities, and the resulting outcomes at the employee level. In addition, employee involvement in the organizational knowledge sharing process is also discussed.

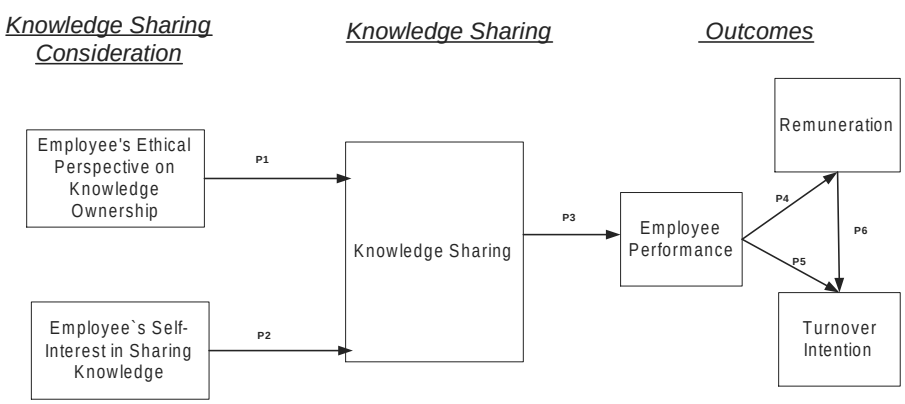
Phase 1: Knowledge-Sharing Variables

In its first phase, the model attempts to examine two seemingly contrasting affects:

A. The employee's ethical perception of organizational ownership regarding the knowledge created within the framework of the organization.
B. The effects of the employees' personal interests on their willingness to share knowledge with the organization.

Knowledge Ownership and Personal Interests. The organization desires to manage all organizational knowledge in such a way as to ensure its immediate accessibility to all

Figure 1. The impact of employee ethical perception and self interest in knowledge sharing on organizational outcomes



organization members. In contrast, employees express an interest to use the organizational knowledge they possess to realize personal goals such as achieving personal power/status within the organization.

The connection between employees' personal ownership of knowledge and that of the organization reveals that the more employees believe in their personal ownership of the knowledge asset, the more their trust regarding organizational ownership of that same knowledge increases (Jarvenpaa & Staples, 2001). In other words, the combination of the ethical perception about organizational ownership of employee knowledge and employees' personal interests affects employees' behavior regarding their desire to help achieve organizational goals.

Therefore, we suggest the following proposition:

Proposition 1: Employees' perception that knowledge acquired within the organization belongs to the organization is positively related to their intention to share knowledge.

Employees' willingness to share knowledge may be based on personal interests that focus more on protecting both their source of power/status and their advantageous position within the organization. Employees develop economic and social exchange relationships with the organization in an attempt to maximize their own personal interests within the organization (Culnan & Armstrong, 1999; Kim & Mauborgne, 1998). Therefore, we suggest the following proposition:

Proposition 2: Employees' feelings that they can maximize their own personal interests by sharing knowledge within the organization are positively related to their intention to engage in knowledge sharing.

Organizational Knowledge Sharing Processes. There are two processes by which employees share knowledge; these two processes take place simultaneously (Reychav & Weisberg, 2005). The first is the transfer of knowledge from the organization to its members (**knowledge transfer**), while the second is **knowledge exchange** among organization members.

Knowledge sharing processes were further expanded into three knowledge-sharing categories (Reychav & Weisberg, 2006), which include:

1. The transfer of explicit knowledge from the organization to its members: The explicit knowledge transfer process takes place within the framework of a one-directional learning process between the organization and the employee and among organization members;
2. The transfer of explicit knowledge from employees to the organization: The sharing of explicit knowledge is considered an ethical activity that takes place among organization members out of their commitment to their workplace (Jarvenpaa & Staples, 2001, p. 153); and
3. The conversion of tacit knowledge into explicit knowledge and its dissemination among organization members: The tacit knowledge transfer process takes place among organization members and is based on the conversion of tacit knowledge into explicit knowledge. This process reflects employees' considerations concerning their personal interests (Wang, 2004). When employees' feelings regarding knowledge ownership are positive, and when they identify their own personal interests as being similar to those of the organization, their belief that the knowledge they possess actually belongs to the organization increases (Jarvenpaa & Staples, 2001).

Phase 2: Knowledge Sharing Outcomes

In its second phase, the model suggests the implications of knowledge sharing at the individual level and as it relates to the organization's accumulated knowledge as a whole. In the current article, we propose an expansion of Wang's (2004) model, which focuses solely on the factors that contribute to employee knowledge sharing. In an organization, the more knowledge employees share, the more it has a direct effect on achieving better individual performance and, as a result, better organizational performance.

Knowledge Sharing and Performance. Prahalad and Hamel's (1990) studies have focused on the link between organizational knowledge and performance and consider organizational knowledge to be one of the most important developing abilities an organization can possess. Organizational knowledge is an intangible asset (Itami & Roehl, 1987) and can significantly affect organizational performance (Schoemaker, 1992).

The transfer of knowledge among an organization's business units significantly contributes to improving their performance (Darr, Argote, & Epple, 1995; Szuanski, 1996). The knowledge gathered through correct knowledge management in an organization leads to improved organizational performance. These improvements then serve as a way to measure how efficiently the organization is utilizing its knowledge (Dess & Shaw, 2001; Egan, Yang, & Bartlett, 2004). Therefore, we suggest the following proposition:

Proposition 3: Employees' level of engagement in knowledge sharing activities is positively related to their performance level.

From the organization's point of view, explicit knowledge sharing has a positive affect on organizational performance (Almashari et al., 2002). This type of knowledge sharing is based on an ethical values system, which includes standards

founded on trust and commitment between the organization and its members. This ethical values system is also clearly defined within the framework of the official work contract signed by both the organization and the employee. From the employees' point of view, considerations relating to personal interests may prevent employees from sharing their knowledge, especially in a competitive work environment where an employee's performance is judged in comparison with the performance of other employees. This strategy attempts to motivate employees to improve their performance in comparison with that of their coworkers, in exchange for future economic or social remuneration from the organization (Wang, 2004).

Performance and Remuneration. Receiving remuneration based on high performance levels is linked to high productivity and quality levels (MacDuffie, 1995). Applebaum, Bauley, Berg and Kallenberg (2002) found a positive connection between receiving remuneration for successful performance and developing a sense of commitment and trust on the part of the employee. This link between an employee's participation in organizational processes and receiving remuneration is presented at several levels: personal, group, and intergroup (Bartol & Srivastava, 2002). Therefore, we suggest the following proposition:

Proposition 4: Employees' performance is positively related to the remuneration they will receive.

Performance and the Intention to Leave. Today's knowledge-based organizations concentrate on identifying "knowledge workers" who can help develop and strengthen organizational relationships (connections through which knowledge can be distributed and that contribute toward improving employee performance within the organization (Capelli, 2000). Employees' increased involvement in improving organizational performance prevents the feelings of responsibility often experienced by employees toward the organization that motivate their desire to help other employees and the organization in general (Blau & Boal, 1987). Therefore, we suggest the following proposition:

Proposition 5: Employees' performance level is positively related to their intention to leave the organization.

Remuneration and the Intention to Leave. Receiving economic and social remuneration both affect employees' tendencies to leave in the following ways:

1. Receiving material remuneration: Organizations that adopt and engage in profit-sharing programs with their employees are known to have a smaller turnover rate (Azfar & Danniger, 2001).
2. Receiving nonmaterial remuneration: Nonmaterial remuneration is expressed in appreciation and promotion within the organization, which may lead to increased employee involvement and a decrease in his or her intention to leave (Faraj & Sproull, 2000). Therefore, we suggest the following proposition:

Proposition 6: Employees' remuneration is positively related to their intention to leave.

FUTURE TRENDS

Great importance is placed on understanding the conflict existing between the organization and the employee and the organization's attempt to help employees find a positive balance between their personal interests and the interests of the organization.

In the future, organizations will need to take the importance of the "human factor" in management more into consideration. As a result, there will be an increased tendency to allocate the necessary resources needed to preserve and protect the organization's "human capital." As a result, this human capital will have a more significant bearing on employees' organizational knowledge sharing.

The current article presented an empirical study comprised of direct factors motivating employees to engage in knowledge sharing activities within the organization. These factors relate to the combination of personal interests and ethical perceptions regarding organizational ownership of employees' knowledge. We believe this combined understanding can lead to increased knowledge sharing within organizations, while the documentation and retrieval of explicit knowledge on the part of the employee will also increase.

Future research studies should attempt to examine the effects of culture, industry, Hi-Tech and Low-Tech on different firms' willingness to share their knowledge with other firms in the market. Collaborative knowledge sharing among firms may be the basis for significant technological breakthroughs and meaningful changes in the social and economic values systems we are currently familiar with.

CONCLUSION

"Is it truly worthwhile for employees to share their knowledge with their organization?" This question has been widely discussed in the literature that focuses on knowledge management and sharing in organizations. In particular, the emphasis has been on motivating employees to share their knowledge by utilizing technology and computerized tools and systems (Hoff & Weenan, 2004; Hulpic, Pouloudi, &

Rzevski, 2002). Nevertheless, it appears that technological tools have a rather insignificant and indirect effect when it comes to motivating employees to create and transfer knowledge (O'Dell & Grayson, 1998). The current article combined employees' perceptions about organizational ownership over employees' knowledge, on the one hand, and employees' personal interests, on the other, as having an influence on knowledge sharing within the organization.

Knowledge retention is based on the ethical behavior of those individuals who share information, ideas, suggestions, expertise and experience. Knowledge sharing among organization employees is identified as an ethical and necessary activity in order to complete business activities and tasks within the organization. The employee's decision to interact with other members is directed by the organization's ethical values system, which encourages employees to share their knowledge in order to preserve the organization's accepted values system.

Considerations regarding personal interests and whether or not knowledge sharing is worthwhile motivate employees to share their knowledge with the organization and their coworkers, based on a desire to receive economic or social remuneration. The distinction described in the current article between knowledge sharing processes and the factors that motivate an employee's considerations and decisions based on ethical aspects and personal interests is necessary in defining appropriate strategies for managing organizational knowledge-sharing processes.

REFERENCES

Almashari, M., Zairi, M., & Alathari, A. (2002). An empirical study of the impact of knowledge management on organizational performance. *Journal of Computer information Systems, 43*(2), 74-82.

Appelbaum, E., Bailey, T., Berg, P., & Kalleberg, A.L. (2002). Shared work-valued care: New norms for organizing market work and unpaid work. *Economic and Industrial Democracy, 23*(1), 125-132.

Azfar, O., & Danninger, S. (2001). Profit-sharing, employment stability, and wage growth. *Industrial and Labor Relations Review, 54*(3), 619-630.

Bartol, K. M., & Srivastava, A. (2002). Encouraging knowledge sharing: The role of organizational reward systems. *Journal of Leadership & Organizational Studies, 9*(1), 64-76.

Blau, P. (1964). *Exchange and Power in social life.* New York: John Wiley & Sons.

Blau, G.J., & Boal, K.B. (1987). Conceptualizing how job involvement and organizational commitment affect turnover and absenteeism. *Academy of Management Review, 12*(2), 288-300.

Boyer, R.E. (1981). *Survey of the law of real property.* St. Paul, MN: West.

Brynjolfsson, E. (1994). Information assets, technology, and organization. *Management Science, 40*(12), 1645-1662.

Capelli, P. (2000). A market-driven approach to retaining talent. *Harvard Business Review, 78*, 103-113.

Carlile, P.R. (2002). A pragmatic view of knowledge and boundary objects in new product development. *Organization Science, 13*(4), 442-456.

Constant, D., Kiesler, S., & Sproull, L. (1994). What's mine is ours, or is it? A study of attitudes about information sharing. *Information Systems Research, 5*(4), 400-421.

Culnan, M.J., & Armstrong, P.K. (1999). Information privacy concerns, procedural fairness and impersonal trust: An empirical investigation. *Organization Science, 10*(1), 105-115.

Darr, E.D., Argote, L., & Epple, D. (1995). The acquisition, transfer, and depreciation of knowledge in service organizations: Productivity in franchises. *Management Science, 41*(11), 1750-1762.

DeLong, D. (1996). *Implementing knowledge management at Javelin Development Corporation: Case study.* Boston: Ernst & Young Center for Business Innovation.

Dess, G.G., & Shaw, J.D. (2001). Voluntary turnover, social capital, and organizational performance. *The Academy of Management Review, 26*(3), 446-456.

Duffy, J. (2000). Knowledge management: To be or not to be? *Information Management Journal, 34*(1), 64-67.

Egan, T.M., Yang, B., & Bartlett, K.R. (2004). The effects of organizational learning culture and job satisfaction on motivation to transfer learning and turnover intention. *Human Resource Development, 15*(3), 279-301.

Emerson, R.M. (1962). Power-dependence relations. *American Sociological Review,* (27), 31-41.

Faraj, S., & Sproull, L. (2000). Coordinating expertise in software development teams. *Management Science, 46*, 1554-1568.

Gupta, A.K., & Govindarajan, V. (2000). Knowledge flows within multinational corporations. *Strategic Management Journal, 21*, 473-496.

Heider, E. (1958). *The psychology of interpersonal relation.* New York: John Wiley & Sons.

Hlupic, V., Pouloudi, A., & Rzevski, G. (2002). Towards an integrated approach to knowledge management: "Hard," "soft" and "abstracts" issues. *Knowledge and Process Management, 9*(2), 90-102.

Homans, G.C. (1958). Social behavior as exchange. *American Journal of Sociology*, (62), 597-606.

Hooff, B., & Weenen, L. (2004). Committed to share: Commitment and CMC use as antecedents of knowledge sharing. *Knowledge and Process Management, 11*(1), 13-24.

Inkpen, A.C. & Dinur, A. (1998). Knowledge management processes and international joint ventures. *Organization Science, 9*(4), 454-468.

Itami, H., & Roehl, T. (1987). *Mobilizing invisible assets*. Cambridge, MA: Harvard University Press.

Jarvenpaa, S.L., & Staples, D.S. (2001). Exploring perceptions of organizational ownership of information and expertise. *Journal of Management Information Systems, 18*(1), 151-183.

Kakabadse, N.K., Kakabadse, A., & Kouzmin, A. (2002). Ethical considerations in management research: A "truth" seeker's guide. *International Journal of Value-based Management, 15*(2), 105-138.

Kim, W.C., & Mauborgne, R. (1998). Procedural justice, strategic decision making, and the knowledge economy. *Strategic Management Journal, 19*(4), 323-338.

Locke, J. (1978). Second treatise of government. In C.B. Macpherson (Ed.), *Property: Mainstream and critical* positions (pp. 15-27). Canada: University of Toronto Press.

MacDuffie, J. (1995). HR bundles and manufacturing performance: Organizational logic and flexible production systems in the world automobile industry. *Industrial and Labor Relations Review, 48*, 197-221.

Marakas, G.M. (1999). *Decision support systems in the twenty-first century.* Englewood Cliffs, NJ: Prentice Hall.

Martensson, M. (2000). A critical review of knowledge management as a management tool. *Journal of Knowledge Management, 4(*3), 204-216.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review, 23*, 242-266.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company.* UK: Oxford University Press.

O'Dell, C., & Grayson, C.J. (1998). If only we knew what we know: Identification and transfer of internal best practices. *California Management Review, 40*(3), 154-174.

Polanyi, M. (1958). *Personal knowledge: Towards a post critical philosophy*. IL: University of Chicago Press.

Polanyi, M. (1966). *The tacit dimension.* London: Routledge & Kegan Paul.

Poole, A. (2000). The view from the floor—what km looks like through the employee's lens. *Knowledge Management Review, 3*, 8-10.

Prahalad, C.K., & Hamel, G. (1990). The core competence of the corporation. *Harvard Business Review, 68*(3), 79-93.

Quinn, J.B, Philip, A., & Sydney, F. (1996). Managing professional intellect: Making the most of the best. *Harvard Business Review, 74*(2), 71-81.

Reychav, I., & Weisberg, J. (2005). Human capital in knowledge creation, management and utilization. In D.G. Schwartz (Ed.), *Encyclopedia of knowledge management* (pp. 221-229). Hershey, PA: Idea Group.

Reychav, I., & Weisberg, J. (2006). From learning organization to organization learning. *The International Journal of Knowledge Culture and Change Management, 5*(9), 53-62.

Ruppel, C.P., & Harrington, S.J. (2001). Sharing knowledge through intranets: A study of organizational culture and intranet implementation. *IEEE Transactions on Professional Communication, 44*(1), 37-52.

Schoemaker, P.J.H. (1992). How to link strategic vision to core capabilities. *Sloan Management Review, 34*(1), 67-81.

Smith, M.E., & Lyles, M.A. (2003). *The Blackwell handbook of organizational learning and knowledge management*. UK: Blackwell.

Storey, J., & Barnett, E. (2000). Knowledge management initiatives: Learning from failure. *Journal of Knowledge Management, 4(*2), 145-156.

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal, 17*, 27-43.

Tampoe, M. (1996). Motivating knowledge workers—the challenge for the 1990s. In P.S. Myers, (Ed.), *Knowledge management and organizational design* (pp. 179-190). Boston: Butterworth-Heinemann.

Thibaut, J.W., & Kelley, H.H. (1959). *The social psychology of groups*. New York: John Wiley & Sons.

Tyler, T.R. (1999). Why people cooperate with organizations: An identity-based perspective. In B.M. Staw & R. Sutton (Eds.), *Research in organizational behavior* (pp. 201-246). Greenwhich, CT: JAI Press.

Wang, C.C. (2004). The influence of ethical and self-interest concerns on knowledge: An empirical study. *International Journal of Management, 21*(3), 370-381.

KEY TERMS

Ethics: Ethics permits a system of moral standard or values (Wang, 2004). Ethical concerns permeate all human actions and interaction that arise in connection with core values as honesty or justice (Kakabadse, Kakabadse, & Kouzmin, 2002).

Human Capital: Human Capital is a combination of employee's education, training, experience (Becker, 1964).

Knowledge: Knowledge is an organized combination of ideas, rules, procedures, and information (Marakas, 1999, p.264).

Knowledge Management: Knowledge Management is an economic view of the strategic value of organizational knowledge that facilitate the acquisition, sharing and utilization of knowledge (Smith & Lyles, 2003, p.12)

Knowledge Sharing: Knowledge Sharing is an exchange or transfer process of facts, opinions, ideas, theories, principles and models within and between organizations include trail and error, feedback and mutual adjustment of both the sender and receiver of knowledge (Szulanski, 1996).

Organizational Knowledge: Organizational knowledge is equated with professional intellect (Quinn, Philip, & Sydney, 1996). Organizational knowledge is a metaphor, as it is not the organization but the people in the organization who create knowledge.

Organizational Norms: Organizational Norms is a set of rules for human behavior in the organization. Organizational Norms regard information sharing as usual, correct and socially expected work place behavior (Constant et al., 1994, p.404; Jarvenpaa & Staples, 2001, p.153).

Tacit Knowledge: Tacit knowledge is knowledge that has not been formalized or cannot be formalized or made explicit (Nonaka & Takeuchi, 1995). Tacit knowledge is based on the individual's subjective insights, intuitions and is deeply rooted in actions experience, ideals, values and emotions (Polanyi, 1966).

Exploiting Context in Mobile Applications

Benou Poulcheria
University of Peloponnese, Greece

Vassilakis Costas
University of Peloponnese, Greece

INTRODUCTION

Pervasive computing is nowadays becoming a reality, exploiting the capabilities offered by both computing infrastructure and communication facilities. The pervasive computing environment encompasses a multitude of diverse devices, operating systems, protocols, and standards. It includes mobile devices such as cellular phones, smart phones, PDAs, and handheld computers for information access, smart cards, and smart labels for identification and authentication, smart sensors, and actuators that perceive the surroundings and react accordingly. Voice technologies such as automatic speech recognition (ASR), text to speech (TTS) and VoiceXML enable the construction of convenient user interfaces and Web services are a key mechanism for interoperability. Wireless wide area networking allows long distance communication through cellular radio while wireless local and personal area networking and standards such as the Wi-Fi, Bluetooth and IrDA allow short distance communication through radio waves and infrared beams.

In the mobile and pervasive computing environment, software engineering should not treat diversity and mobility as problems to overcome, but seek methods of which it could take advantage instead. In these environments, the selection of purpose-oriented and timely information, tailored to user preferences and media characteristics will ensure optimised information delivery. To this end, the context—the information that surrounds the human-computer interaction—plays a key role and is rapidly changing in mobile settings, and the understanding of it is indispensable for application designers in order to choose, capture and exploit it. The importance of the context is to use it to make context-aware applications, that is, those applications that are interested in who, where, when and what, in order to determine why the situation occurs and adapt their behavior accordingly.

BACKGROUND

The Concept of Context

An important dimension of mobile computing is "mobility", which is primarily concerned about people moving in space and doing their personal, social and professional activities in a wide temporal space. The informative support of the mobile user can be accomplished through terminals, which are movable and operate regardless of the location and time and offer wireless access to information and services. Although mobility—spatial and temporal—is an important aspect of mobile computing, it constitutes only one dimension of the "context".

The term "context" is defined as "the interrelated conditions in which something exists or occurs" in Merriam-Webster's dictionary. In the domain of context-aware computing, researchers have defined context as "location, identities of nearby people and objects, and changes to these objects" (Schilit & Theimer, 1994), "location, identities of the people around the user, the time of day, season, temperature, etc." (Brown, Bovey, & Chen, 1997) and "knowledge about the user's and the device's state, including surroundings, situation and location" (Schmidt & Laerhoven, 2001).

Dey and Abowd (1999) propose a more generic definition according to which context is any information that can be used to characterize the situation of an entity. An entity is a person or object that is considered relevant to the interaction between a user and an application, including the user and the application themselves. Dey's definition is more comprehensive and generic and makes it easier for an application designer to enumerate the context for a given application and choose the appropriate desirables.

The contextual information can be classified, according to which entity it concerns, into the following categories (Schilit, Adams, & Want, 1994):

- **User Context:** User identity, location, collection of nearby people, social situation, activity, user's profile, and so forth.

- **Computing Context:** Hardware characteristics, software characteristics, network connectivity, communication bandwidth, nearby resources such as printers, displays and other devices and so forth.
- **Physical Context:** Lighting, noise level, temperature, humidity, and so forth.
- **Time Context:** Time of the day, week, month, season of year, time zone, and so forth.

Orthogonally to these classifications, context can be divided into two broad classes: *primary* and *secondary context*. Primary context derives directly from sensors or information sources while secondary context is inferred from the primary context. For example the name of a city is a secondary context because it derives from GPS coordinates through a relation mechanism.

Context can be also distinguished according to a range of temporal characteristics and it can be classified as *static* or *dynamic*. Static context does not change very quickly (or at all; e.g., a person's date of birth), while dynamic context does (e.g., the location of a person who is driving). When applications are not only interested in the current state of the context (*present context*), but past context is of importance too, *context histories* are stored; in some situations there is a necessity for context prediction (*future context*).

The interactions between context sources and context sinks can be characterized as "context push", when the context sources update periodically context information in context sinks and "context pull", when the sinks demand information from the context sources.

The types of context, according to the manner of its acquisition, can be divided in three categories:

- **Sensed context:** This context is acquired from the environment by means of physical or software sensors (identity, temperature, time).
- **Derived context:** This kind of contextual information is computed (for example the name of the city from GPS coordinates).
- **Context explicitly provided:** The context that the user provides explicitly (for example the entries in the user profile).

Defining Context-Aware Applications

Context-awareness is a concept that consists of two notions: the notion of perceiving the *context* and the adaptivity that derives from the awareness. Adaptivity or adaptability is defined as the ability of a service/application to react to its environment and change its behavior according to the context. Context-aware computing was first introduced by Schilit and Theimer (1994) as the use of software that adapts according to the location of use, the collection of nearby people and objects, as well as to changes to such elements over time. Fickas, Korteum, and Segall (1997) define context-aware applications (called environment-directed) as applications that monitor changes in the environment and adapt the operation according to predefined or user-defined guidelines. Dey and Abowd (1999) characterize a system as context-aware if it uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task.

The functions that a context-aware application should implement (Schilit et al., 1994) are:

- **Proximate selection:** A user interface-level technique where the nearby located objects are emphasized or otherwise made easier to choose.
- **Automatic contextual reconfiguration:** A process of adding new components, removing existing components, or altering the connections between components due to context changes.
- **Contextual information and command:** Queries on contextual information can produce different results according to the context in which they are issued. Similarly, context can parameterize "contextual commands".
- **Context-triggered actions:** "If-then" rules used to specify how context-aware systems should be adapted.

Dey and Abowd (1999) propose that the features that context-aware applications may support are:

- **Presentation** of information and services to the user or use context to propose appropriate selections.
- **Automatic execution** of a service according to context changes.
- **Tagging of context** to information for later retrieval.

Related Work

The first context-aware system was the Active Badge System developed at Olivetti Research Lab. In case an employee on duty was not in his office, this would direct phone calls to the closest appliance according to the employees' location in the office environment. An evolution of this system is the ParcTab system, which was developed at the Xerox Palo Alto Research Center, relied on PDAs to offer a range of context-aware office applications (Schilit et al., 1994). Besides the above projects, a number of efforts have resulted in context-aware applications that can be categorized as follows (Dockhorn, 2003):

- **Conceptual frameworks:** They focus on the architectural aspect of context-aware systems and introduce methods of gathering, interpreting and disseminating

context to the interested parties. Examples of this approach are the projects: Context Toolkit by Georgia Institute of Technology (Dey, Abowd, & Salber, 2001) and Cooltown by Hewlett-Packard (Kindberg et al., 2002).

- **Service platforms:** They allow the rapid creation, deployment, and dynamic discovery of services, in order to provide the appropriate functionality according to user context. Examples of service platforms are the M3 architecture from the University of Queensland (Indulska, Loke, Ratotonirainy, Witana, & Zaslavsky, 2001) and the Platform for Adaptive Applications from the Lancaster University (Efstratiou, Cheverst, Davies, & Friday, 2001).
- **Appliance environments:** They try to support interoperability among collections of appliances. Examples of such environments are the Ektara environment from MIT (DeVaul & Pentland, 2000) and the Universal Information Appliance from IBM (Eustice, Lehman, Morales, Munson, Edlund, & Guillen, 1999).
- **Computing environments:** They provide context-aware computing that decouples users from devices and enables applications to perform tasks on behalf of the user. Projects of this category are the PIMA by IBM (Banavar, Beck, Gluzberg, Munson, Sussman, & Zukowski, 2000) and Portolano by the University of Washington (Esler, Hightower, Anderson, & Borriello, 1999).

AN ARCHITECTURE FOR CONTEXT-AWARE MOBILE APPLICATIONS

The Requirements

The dynamic environments, in which the mobile applications operate, constitute a challenge in order to propose a new software architecture that exploits the context and facilitates the design of applications which enhance the user's mobile experience with suitable services.

The requirements that have to be met by such an application are:

- **Capture** contextual information from sensors and users.
- **Store** part of the contextual information for later exploitation.
- **Interpret** information, at an abstract level, in order for it to be more meaningful.
- **Transit** the information to the application modules.
- **Adapt** the application behavior according to the context.

The main goal of our approach is the adaptation process, which consists of two components: first, the design of an architecture which supports adaptation at run-time and secondly, the provision of design of the core application itself in order to be context-aware. One evident question concerning the adaptation, is "which are the real adaptation tasks that a context-aware mobile application has to implement?" Placing the end-user in the middle of the concern, thus employing a user-centered approach (Vredenburg, Isensee, & Righi, 2001), the adaptation operation (Kappel, Retschitzegger, & Schwinger, 2001) should consist of the following:

- **Content adaptation:** Determine which content elements are hidden or revealed according to context changes.
- **Operation adaptation:** The choosing of and switching between the functional components of a service/application according to the context.
- **Presentation adaptation:** The tailoring of the system's user interface and interactive behavior to the individual needs of the user and system capabilities.

Another issue with regard to the adaptability that should be decided is "where the adaptation process to the context takes place". It could occur (1) within the application and it is *called laisser-faire adaptation*; (2) exclusively outside the application, called *application transparent adaptation*; and (3) within the application as well as out of it, called *application-aware adaptation* (Satyanarayanan & Ellis, 1996). Our approach adopts the application-transparent adaptation, since mixing context management code with the application code renders the application code more complex, difficult to maintain and impossible to reuse. This indicates the decoupling of context-independent activities of the application from the contextual concerns, as well as the separation of contextual data from the application's data.

In the following paragraphs, we first present context modelling, which is a required underpinning for any context-aware application and subsequently present our proposed framework that satisfies the above requirements.

Context Modeling

One of the major issues in context-aware applications is the representation of context, in such a way that facilitates the context reasoning and sharing. Insofar a lot of context models have been proposed, some based on proprietary representation schemes and others on more formal data models.

The existing context models belong to one or more of the following categories:

- **Key-value models:** The simple key-value data structures, which were first introduced by Schilit, Theimer,

and Welch (1993) in PARC's mobile computing environment. This representation provides a fast and easy way to set and update context but it is only feasible in situations where (key, value) pairs obtained from the environment exactly match those in the model; for example, the pair (temperature, 50) matches itself only, while it does not match the pair (temperature, 50.1). Moreover, this model lacks any structure and is thus difficult to manage when the number of keys and/or values increases.

- **Web-based model:** In this context model, each entity (person, place, or thing) corresponds to an unstructured web page, retrieved via a URL. The Cooltown project (Kindberg et al., 2002) relies on this model, which is intended to be used by humans rather than by applications.
- **Markup schemes models:** Using markup languages—such as XML—is another method to model contextual information in a flexible and structured manner. Profiles are a typical example of this kind of context modeling approach. For example, CC/PP (composite capabilities preferences profile) is an RDF-based framework (resource description framework) to describe user preferences and devices' capabilities and characteristics. However the usage of profiles for context representation becomes difficult when the relationships and constraints of context are complex. The models under this category are extensions to the above standard, with no fixed hierarchy and try to cover the higher dynamics of contextual information.
- **Object-oriented models:** They are based on an object-oriented approach in which context information is structured around a set of entities. This modeling often uses a variation of the Entity Relationship model and stores the contextual information in relational databases. Although the object-oriented models employ the benefits of encapsulation and reusability, they are usually designed for a specific context domain.
- **Ontology-based models:** Ontologies are believed to be the most suitable model for the representation and reasoning of context information, for the following reasons (Mokhtar, Fournier, Georgantas, & Issarny, 2005): (1) they enable knowledge sharing through a common set of concepts, (2) they allow efficient reasoning so as to deduce high-level from low-level context and (3) they enable interoperability.

The Context-Aware Architecture

The overall proposed system architecture is depicted in Figure 1, while its operation and individual modules are discussed in the following paragraphs.

Context Manager

The *context manager* is the system module that is responsible for the capturing of the context from different sources, the interpretation of it in a higher level format, the storage of the context and the distribution of it to the *adaptation manager*.

The *context gatherer* subsystem is in charge of the capturing of context either from hardware sensors (e.g., location sensors, identification sensors, motion sensors, etc.) or software sensors (processing power, available hardware components, storage systems, software components, current time, etc.) or through the import of relevant profiles. These profiles could be:

- **User profile:** It contains information about the user and his/her desired application's features. User-pertinent data could be his/her name, date of birth, address, occupation, hobbies, available technical equipment, billing plans, activities over time, etc. Computer-relevant data could be application-independent, like font size, preferred language, content characteristics (e.g., text, audio or video), or application- dependent, like the kind of stocks for which the user wishes to be informed when price changes occur.
- **User agent profile:** It describes the capabilities of devices e.g. display size, memory, operating system, browser characteristics, etc.
- **Network profile:** It describes network characteristics like supported bandwidths, bearer type and so forth.
- **Application profile:** It holds application-specific information, which can be used for dynamic composition of modules from distinct services.

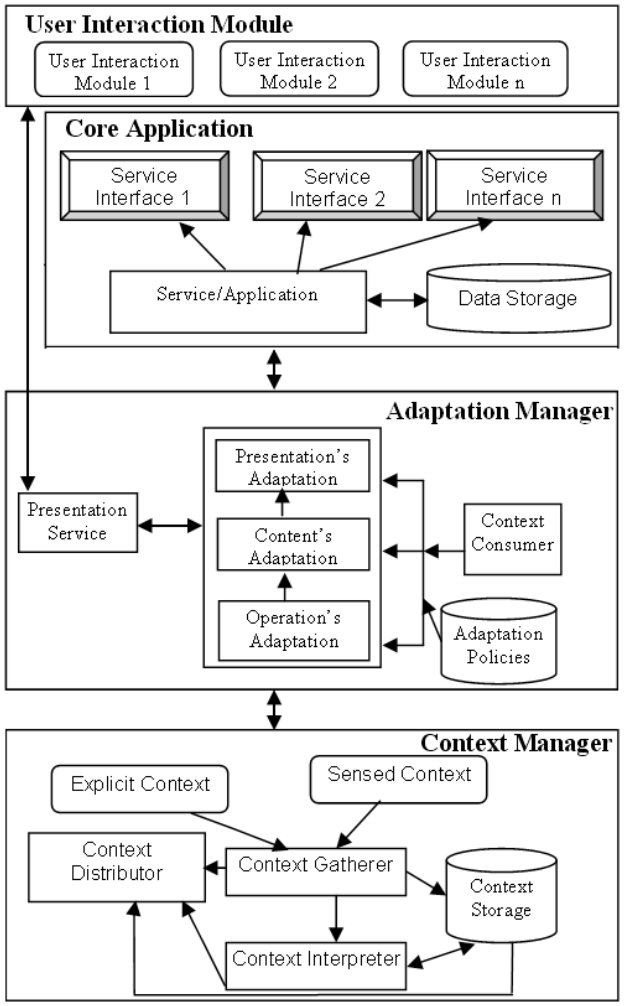
The *context interpreter* gathers information from *context gatherer* and *context storage* and conducts aggregation and inference activities. The resulting context is either passed to the *context distributor* or stored to the *context storage* for later retrieval.

The *context distributor* makes contextual information uniformly available to the *adaptation manager*. It provides information in two different modes: *request-response* and *event triggered*, supporting the "push" and "pull" character of the context. In *request-response* mode it grants context only on explicit request while in *event triggered* mode the provision of context is fired from specific events.

Core Application

The *core application* consists of a number of services that materialize the desired functionality. These services are not context-aware and their behavior remains the same in the different context situations. Furthermore these services expose their functionalities in a number of distinct and predefined

Figure 1. System architecture



alternatives called *service interfaces* (e.g. multimedia or plain text version) in order to support the diverse preferences during the operation adaptation process.

Adaptation Manager

The adaptation process occurs explicitly in the boundaries of the *adaptation manager* and could be accomplished through three concrete steps: the operation, the content and the presentation adaptation. The *adaptation policie*s repository contains the adaptation rules according to which the adaptation process takes place (e.g., *If* condition *x* is met *then* message *y* is triggered). The *presentation service* is the particular instance of a specific adaptation operation which is forwarded to the *user interaction module3.*

The *operation adaptation module* carries out the service's composition, which is the process of collecting the suitable application services in the desired version. These services are matched according to the adaptation rules, which depend on the context situation. The context data are passed as input to this process.

The *content adaptation module* materializes the *content adaptation* that follows the *operation adaptation*. It makes the choice of the appropriate content, the language, the compression of data etc, in order to suit the client's display characteristics, network capabilities and user preferences.

The *presentation adaptation module* conducts the *presentation adaptation* that occurs after the *content adaptation*. It makes the transcoding and the modality transformation with a view to satisfying the clients' claims.

The *context consumer* is the bearer of the context to the *adaptation manager*. Additionally the services subscribe to it, in order to be informed by the "push" notifications upon context changes.

User Interaction Module

The user interaction module presents data and selection choices to the end user in different modalities. The adaptive user interfaces have been produced as a result of the adaptation process, considering both individual needs and changing conditions in the application environment.

In many existing projects in the area of context-aware computing, capturing and implementation of context have been made in an ad-hoc and per application manner. In the proposed architecture the management of the context is encapsulated in a single module (context manager) and it may be reused from any other application; moreover, it can evolve to include other aspects of context that have not yet been foreseen and determined. The implementation of context adaptation logic (adaptation manager) can be supported by utilizing the aspect-oriented paradigm and reflective programming techniques, which assist in managing cross-cutting concerns and dynamically discovering and invoking functionality, respectively.

FUTURE TRENDS

Although research has addressed a number of issues related to context understanding, modeling, reasoning and knowledge sharing, further investigation of them is needed to clarify the relevant concepts. Best practices for analysts to elucidate user requirements pertinent to context adaptability and frameworks for designers and implementers that will facilitate the development of context-aware applications should also be surveyed. In the implementation phase, in particular, the ability to separate handling of context-awareness from application logic is highly desirable, since it will decrease the complexity of application development and will make incorporation of context-aware aspects to existing, non context-aware applications easier.

Currently, context-aware services have not been adopted by the masses to the initially anticipated extent. The reasons for this lag need to be surveyed, and feedback from the existing or potential user groups should be collected and analyzed. Social and legal issues as well as privacy and security concerns related to context awareness should be also considered.

CONCLUSION

In this paper, we have discussed the issues of context and context-awareness in the provisioning of mobile services, identified context categories, presented modeling and acquisition techniques, as well as methods of context exploitation. We have presented a generic architecture for supporting context-aware applications, which provides facilities for managing context, performing functional adaptation and tailoring the user interface to suit the current context parameters. The profound understanding of context will facilitate the process of choosing, managing and utilizing it, in order to provide context-aware applications, those ones that deliver the correct service, to the correct user, at the correct place and time, and in the correct format, with the minimum distraction of the user.

REFERENCES

Banavar, G., Beck, J., Gluzberg, E., Munson, J., Sussman, J., & Zukowski, D. (2000). Challenges: An application model for pervasive computing. *Proceedings 6th Annual International Conference on Mobile Computing and Networking* (pp. 266-274).

Brown, J., Bovey, D., & Chen, X. (1997). Context-aware applications: From the laboratory to the marketplace. *IEEE Personal Communications, 4*(5), 58-64.

DeVaul , W., & Pentland, A. (2000). *The Ektara architecture: The right framework for context-aware wearable and ubiquitous computing applications.* MIT Technical Report.

Dey, A., & Abowd, G. (1999). *Towards a better understanding of context and context-awareness.* Technical Report 99-22, Georgia Institute of Technology.

Dey, A., Abowd, G., & Salber, D. (2001). A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human-Computer Interaction Journal, 16*(24), 97-166.

Dockhorn Costa, P. (2003). *Towards a services platform for context-aware applications.* Master Thesis, University of Twente, The Netherlands.

Efstratiou, C., Cheverst, K., Davies, N., & Friday, A. (2001). An architecture for the effective support of adaptive context-aware applications. *Proceedings of 2nd International Conference in Mobile Data Management* (pp. 15-26).

Esler, M., Hightower, J., Anderson, T., & Borriello, G. (1999). Next century challenges: Data-centric networking for invisible computing. *ACM/IEEE International Conference on Mobile Computing and Networking* (pp. 256-262).

Eustice, F., Lehman, J., Morales, A., Munson, C., Edlund, S., & Guillen, M. (1999). A universal information appliance. *IBM Systems Journal, 38*(4), 575-601.

Fickas, S., Korteum, G., & Segall, Z. (1997). Software organization for dynamic and adaptable wearable systems. *Proceedings of the 1st IEEE International Symposium on Wearable Computers* (pp. 56-63).

Indulska, J., Loke, S., Ratotonirainy, A., Witana, V., & Zaslavsky,A. (2001). An open architecture for pervasive systems. *Proceedings of the 3rd International Working Conference on Distributed Applications and Interoperable Systems* (pp. 175-188).

Kappel, G., Retschitzegger, W., & Schwinger, W. (2001). Modeling ubiquitous Web applications: The WUML approach. *Proceedings of the International Workshop on Data Semantics in Web Information System*.

Kindberg, T., Barton, J., Morgan, J., Becker, G., Caswell, D., Debaty, P., et al. (2002). People, places, things: Web presence for the real world. *Mobile Networks and Applications, 7*(5), 365–376.

Mokhtar, B., Fournier, D., Georgantas, N., & Issarny, V. (2005). Context-aware service composition in pervasive computing environments. *Proceedings of the 2nd International Workshop on Rapid Integration of Software Engineering techniques* (pp. 129-144).

Satyanarayanan, M., & Ellis, C. (1996). Adaptation: The key to mobile I/O. *ACM Computing Surveys, 28*(4es), 211.

Schilit, B., Adams, N., & Want, R. (1994). Context-aware computing applications. *1st International Workshop on Mobile Computing systems and Applications* (pp. 85-90).

Schilit, B., & Theimer, M. (1994). Disseminating active map information to mobile hosts. *IEEE Network, 8*(5), 22-32.

Schilit, B., Theimer, M., & Welch, B. (1993). Customizing mobile applications. *Proceedings of the USENIX Mobile & Location-Independent Computing Symposium* (pp. 129-138).

Schmidt, A., & Laerhoven, K. (2001). How to build smart appliances. *IEEE Personal Communications, 8*(4), 66-71.

Vredenburg, K., Isensee, S., & Righi, C. (2001). *User centered design: An integrated approach.* Prentice Hall PTR, ISBN: 0130912956.

KEY TERMS

Adaptivity or Adaptability: It is the ability of a service/application to react to its environment and change its behavior according to the context.

Application-Aware Adaptation: It is the adaptation process to context, which takes place within the application as well as out of it.

Application Transparent Adaptation: It is the adaptation process to context, which takes place exclusively outside the application.

Context: Context is any information that can be used to characterize the situation of an entity. An entity is a person, or object that is considered relevant to the interaction between a user and an application, including the user and the application themselves.

Context-Aware: Context-aware is a system that uses context to provide relevant information and/or services to the user, where relevancy depends on the user's task.

Laisser-Faire Adaptation: It is the adaptation process to context, which takes place inside the application.

Service: It is piece of an information product that materializes a concrete functionality.

Exploiting the Strategic Potential of Data Mining

Chandra S. Amaravadi
Western Illinois University, USA

INTRODUCTION

Data mining is a new and exciting technology to emerge in the last decade. It uses techniques from artificial intelligence and statistics to detect interesting and useful patterns in historical data. Thus traditional techniques such as regression, Bayesian analysis, discriminant analysis, and clustering have been combined with newer techniques such as association, neural nets, machine learning, and classification (Jackson, 2002).

Applications of data mining have ranged from predicting ingredient usage in fast food restaurants (Liu, Bhattacharyya, Sclove, Chen, & Lattyak, 2001) to predicting the length of stay for hospital patients (Hogl, Muller, Stoyan & Stuhlinger 2001). See Table 1 for other representative examples. Some of the important findings are: (1) corporate bankruptcies can be predicted from variables such as the "ratio of cash flow to total assets" and "return on assets," (2) gas station transactions in the U.K. average £20 with a tendency for customers to round the purchase to the nearest £5 (Hand & Blunt, 2001), (3) 69% of dissatisfied airline customers did not contact the airline about their problem, (4) sales in fast food restaurants are seasonal and tend to peak during holidays and special events (Liu et al., 2001), (5) patients in the age group over 75 are 100% likely to exceed the standard upper limit for hospital stays (Hogl et al., 2001). In recent years, data mining has been extended to mine temporal, text, and video data as well. The study reported by Back, Toivonen, Vanharanta, and Visa (2001) analyzed the textual and quantitative content of annual reports and found that there is a difference between them and that poor organizational performance is often couched in positive terms such as "improving," "strong demand," and so forth. Both applications and algorithms are rapidly expanding. The technology is very promising for decision support in organizations. However, extracting knowledge from a warehouse is still considered somewhat of an art. This article is concerned with identifying issues relating to this problem.

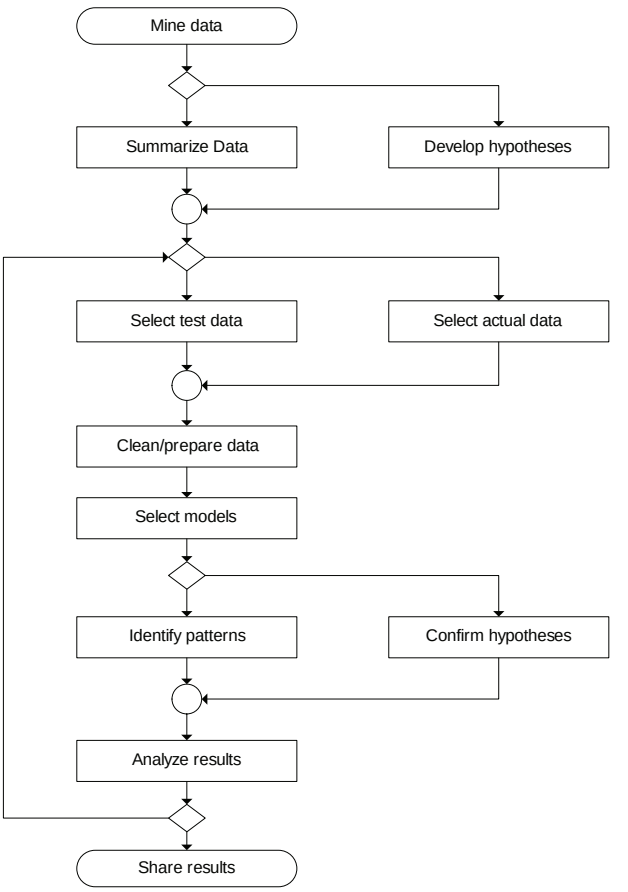
Table 1. Examples of data mining applications

• Predicting supplies in fast food restaurants (Liu, Bhattacharyya, Sclove, Chen & Lattyak 2001).
• Quality of health care (Hogl, Muller, Stoyan & Stuhlinger 2001).
• Analyzing Franchisee sales (Chen, Justis & Chong 2003).
• Predicting customer loyalty (Ng & Liu 2001).
• Mining credit card data (Hand & Blunt 2001).
• Analyzing annual reports (Back et al. 2001).

BACKGROUND

The mining process is often labeled as knowledge discovery in databases (KDD). An extended KDD model is presented in Figure 1. As illustrated in the figure, mining is carried out in two modes: "data-driven" or "hypothesis-driven" ("question-driven"). The *data-driven* approach is also referred

Figure 1. The extended KDD process



to as exploratory data mining and is often carried out to develop a preliminary understanding of the data. As a first step, data are summarized to identify their characteristics. Typical measures such as frequency distribution, mean, variance, and so forth are computed. For credit card data, what is the average monthly spending for customers? What credit card products are most frequently purchased? Based on results from this step, models are selected and the mining is performed.

Hypothesis-driven methods attempt to verify whether or not a particular pattern exists (Hogl et al., 2001). In this mode, the process starts with the identification of hypotheses that are motivated by business concerns. Each hypothesis can be thought of as a micro theory (MT) about the domain or an assumption to be verified.

The space of patterns that can be explored in data mining is very large and is restricted by computational constraints. Large data sets with high dimensionality will preclude data driven methods. Hypothesis driven approaches are more tractable and therefore more appealing. But even with these approaches, organizations still lack the resources to mine all available data. Formulation of good hypotheses is therefore important and will influence data selection.

The mining data are partitioned into a test set and the evaluation set (Jackson, 2002). The test set is generally 10-20% of the actual data. Neural nets and machine learning algorithms typically require more test data. Models developed with the test set are validated with the actual data. In the question driven approach, the required columns are selected depending on the hypothesis to be tested. This is a difficult problem ("curse of dimensionality") in data mining and sometimes referred to as *feature selection*. The large number of attributes such as demographic variables in a warehouse makes data selection a challenging process because the variables contributing to a pattern cannot be known a priori. Selection of irrelevant attributes adds to computational complexity and perhaps even misleading results. At the present time, the best method is trial and error.

Mining starts with data that are often integrated from several sources. Integration can present challenges due to differences in formats or attribute definitions. The data are cleaned and transformed by (Peacock, 1998): (a) checking for file transfer errors, where some of the records are not properly loaded or some of the columns are missing, (b) standardizing formats or codes such as converting from text to numeric codes or vice versa (1-married, 2-single, etc.), (c) dealing with sparse records; missing values are either filled or the record is deleted, (d) performing required calculations such as debt/equity, number of faculty/student, student credit hour per faculty, and so forth, (e) identifying and deleting outliers. Outliers are detected during the process of data summarization.

Data selection is followed by *model selection*. As shown in Table 2, mining techniques are broadly classified into summarization/visualization, clustering, classification, association, and sequence with a choice of algorithms in each. The limitations and conditions of each algorithm have to be borne in mind when selecting a suitable algorithm. For example, the decision tree approach may produce erroneous results for data with small training sets. Similarly, time series analyses are computationally expensive and cannot be effectively carried out on large data sets (Kumar, 2002).

Testing with the model will result in initial results. The test data are iteratively used to refine the model, and the analysis is run on the evaluation set. Results can take the form of cluster plots, dependency rules, regression coefficients, bar graphs, and so forth. These are analyzed to identify patterns and discussed with functional area specialists to ensure that they are meaningful in the organizational context. Findings are then shared.

THE KDD PROCESS AND DOMAIN UNDERSTANDING

Data mining results in the identification of patterns. Identifying those that are relevant and interesting to the organization is dependent on the analyst's skill, experience, and understanding of the *organizational context* of the data. To develop a general understanding of the organization, the analyst can look through the company's annual reports, Web site, and news articles. The analyst can also meet with managers working in the area to become familiar with its goals, objectives, and critical issues. For example, a manager may be concerned about how to improve catalog mailings. Another may be concerned about increasing catalog sales. While this general understanding is critical in KDD activities, the analyst must also develop a detailed understanding of the data, that is, the relationships among variables. We will discuss the relationship between important KDD activities and this understanding.

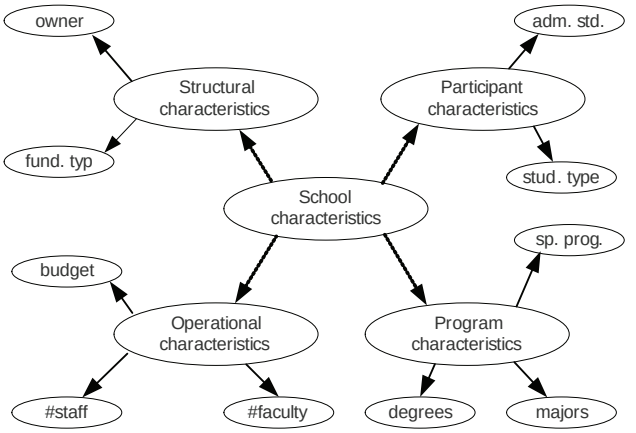
Data Selection/Hypothesis Formation

Knowledge of the domain is essential in pre-mining as well as post-mining activities. Hypothesis formation, data selection, and transformation require conceptualizing relationships among attributes (dimensions) and their impact. For example, converting numerical income data in a bank warehouse to categorical attributes ("hi," "mid," "lo") requires knowledge of expected income levels. Generally miners anticipate two types of patterns: (a) attributes concerning an entity or issue of interest such as computer defects, user complaints, employees/suppliers, and so forth; (b) attributes influencing organizationally relevant behavior such as hiring/firing, bankruptcy, and hospital stay. Some questions that might be asked are:

- What are the most common customer complaints?
- Which service engineers handled the most complex machine problems?
- What types of manufacturing defects are most likely in a computer? When are they most likely to occur?
- How long is an 80-year-old patient likely to stay in a hospital?
- What are the hiring/firing patterns in a company?
- What factors influence personal bankruptcies?

Data selection and transformation are dependent on issue of interest in the relevant functional area. For instance, an insurance organization with a 10^2 x 10^7 warehouse may be interested in demographic factors affecting claims. Dimensionality is somewhat high, but there are no guidelines for the analyst regarding selection of attributes. Random trial and error techniques will be limited by available resources. Statistical methods to support feature extraction have been found only marginally better than trial and error techniques (Lin & McClean, 2001). A model of the domain will aid in extracting relevant data. Both formal and informal techniques will suffice. Causal mapping has been proposed as one of the methods to understand interdependencies in high dimensional data (Fayyad & Stolorz, 1997). A *causal map* or influence diagram depicts cause-effect relationships between decision variables (Eden & Ackermann, 2004). The technique can be made more informal by simply omitting the signs ("+"/"-") on the cause maps. An example domain model is shown in Figure 2.

Figure 2. A domain model for understanding school similarities



Model Selection

Mining models are classified into descriptive or predictive. *Descriptive* techniques describe the data, while *predictive* techniques attempt to predict behavior or variable of interest. Examples of the former include data summarization and clustering, while examples of the latter include regression, association, and dependency analysis (Jackson, 2002). Model selection is governed by a number of factors including the limitations of the technique as well as the mining objectives. As evidenced from Table 2, different techniques are applicable in different situations. For instance, factors influencing catalog sales can be identified from a regression or dependency analysis. The information regarding the service engineers handling the most complex machine problems is easily obtained from a bar graph of service engineers and machines serviced. This is a type of problem that is best visualized. In general, descriptive techniques are useful in identifying information about an entity of interest, while predictive techniques are useful for understanding behavior.

There has been a tendency to treat models as black boxes, but models are sophisticated and require grappling with them while analyzing data. Model parameters also require domain understanding. For instance, regression and dependency analysis require understanding dependent and

Table 2. Popular mining techniques and representative modeling issues (Amaravadi & Daneshgar, 2003)

Technique	Popular algorithms	Application
Summarization/ Visualization	Descriptive statistics, OLAP, graph, 3D plots	Data exploration, average values, outliers
Association	A priori, hash tree, partitioning, and sampling	Relationships among attributes as pertaining to target variable
Classification	CART, ID 3, $C_{4.5}$, K-NN, neural nets, discriminant analysis, and Bayesian classification	To predict class membership
Clustering	K-means, K-Mediod, hierarchical methods.	Determine naturally occurring groups
Prediction	Regression, Bayesian analysis	Predict value of target variable.
Time series analysis	Simple trend analysis, exponential, Box-Jenkins, seasonal ARIMA models	Trends that vary with time, longitudinal variation

Table 3. Part of the AACSB sample data (from www.aacsb.edu)

Name	Cl	Mission	Research	Commun	OpBudget	OpBud/ FTFac
University1	I	T,I,S	A,I,B	Rural	$9,019,497	$111,352
University2	I	T,I,S	Equal	Rural	$10,763,523	$114,506
University3	I	T&I,S	B&A,I	Rural	$6,914,404	$132,969
University4	I	T,I,S	A,I,B	Urban	$3,136,878	$142,585
University5	Doc	T,I,S	A,B,I	Suburban	$12,589,979	$148,117

Name	UGrad	FTEQ	FT	FTPHD	FTSt/FTFac	UGrad
University1	$4,997	83	81	58	17	Res
University2	$3,038	97	94	73	23	Res
University3	$3,792	56	52	40	24	Commut
University4	$2,592	23	22	15	27	Commut
University5	$5,984	92	85	73	30	Res

Name	UGrad	Grad	FT	PT	TDegrees	AcctDegs
University1	Res	Res	1,274	174	300	56
University2	Res	Res	2,086	48	565	55
University3	Commut	Commut	1,170	378	447	83
University4	Commut	Commut	557	287	115	27
University5	Res	Commut	2,502	205	431	71

independent variables. Similarly, cluster analysis requires understanding similarity characteristics and the number of expected clusters. Also of great importance to the data miner is the efficacy of models in different situations. See the work by Lin and McClean (2001) for a comparative discussion of the effectiveness of classification models.

Testing

Once micro theories are identified and data sets are selected/transformed, the next step in the KDD process is testing. As already noted, testing proceeds in two stages, first with "test data" to develop the model and then with the evaluation set to guard against overfitting. As mentioned, the DM techniques include clustering, association, classification, and dependency analysis. The hypothesis test list is used by the analyst as a guide in selecting a suitable technique. For instance, an assumption about the reliability of a supplier could be confirmed by an association analysis between suppliers, delivery times, and the number of times the specifications were met 100%. It should be noted that the raw data may not be available in this form, and therefore may require tabulating and aggregation especially with respect to the variable, "specifications being met 100%." If the association analysis confirms some vendors meeting these criteria, this is again tested on the remainder of the data in the second stage. A number of situations may arise with tested hypotheses: (a) the hypothesis is supported in its entirety at the 90% confidence level or higher, (b) the hypothesis is not supported at the 90% confidence level, but at a lower level of confidence, (c) the hypothesis is not supported at any confidence level. Situations "a" and "c" are clear-cut, resulting in confirmation or disconfirmation of the MT, but "b" will probably be more common. A finding such as "70% of the time patients who are 65 are likely to take one week to recover from surgery" is only partially useable. In such cases, analysts can change the attributes and run the analysis again. Thus the process is carried out iteratively with the strategy being modified.

Case Study of AACSB Schools

The author was part of a committee that was involved in a data mining study involving AACSB schools. The objective of the study was to identify schools that were similar to the author's university. This mini-case illustrates some of the ideas discussed in this article. Part of the sample data is illustrated in Table 3 and is available from the AACSB

Web site (www.aacsb.edu). The names of the schools are disguised for confidentiality. The columns are explained in the appendix.

Collecting the data was labor intensive, as it was listed as a Web page profile for each individual school. The data were printed, and the information was manually entered into a database. Leaving out the international schools, there were a total of 423 colleges of business in the U.S. Eliminating the top tier and private schools resulted in approximately 63 records (of universities), each with 26 attributes. The labels for the data are duplicated, and this needs to be understood. It should be noted that the data are also sparse with values missing from a number of records.

The raw data are overwhelming to the novice. A cluster analysis with all the variables yielded no clustering. This is predictable both because of the size and because of extraneous columns. It is clear that an understanding of university characteristics is required to reduce the number of columns. Discussions with committee members resulted in the model illustrated in Figure 2. As evidenced from the diagram, a school's character is composed of four sets of factors, structural, operational, geographical, and participant. There are sub-factors for each of these. Thus operational characteristics include the "budget," "#faculty" and "#staff." This model is invaluable in reducing dimensionality. For example, it is clear that columns such as "tuition," "OpBudget/FTFac," "FT students/FT faculty," and "number of accounting degrees" are not relevant and can be safely dropped. Also instead of having three separate variables for "FT," "FTEQ," and "FTPHD," only one, which is "FTEQ," can be chosen. Other variables can similarly be collapsed, leading to a dimensionality reduction of 50%. Records for which values were missing were dropped. Since cluster analysis can be carried out with only non-numerical data, columns with text are either dropped or converted to values assigned on a numerical scale. A decision was made to take the latter course of action. The analysis showed good results with *six* clusters. There were 11 universities in the peer group of Western, including University#2 in Table 3. For brevity the cluster results are not reproduced here. The case highlights the importance of understanding the data context as well as the subjectivity involved in the mining process.

FUTURE TRENDS

It is expected that new technologies such as mobile commerce, RFID, digital videography, and smart dust are going to multiply the data that organizations collect. Warehouses with 100 TB are expected to become common in 10 years (Kimball, 2003). Competition will force organizations to delve deeply into these data resources. In some cases, mining will be extended to real time data.

While hardware will undoubtedly evolve, mining techniques that are already tried by large data sets will be further tested with massive data sets. Efficient algorithms will be critical. Unfortunately, algorithm complexity tends to be a function of the number of records and number of dimensions, both of which will continue to plague miners. Algorithms with a complexity of $O(N^2)$ where N is the number of records will not be suitable (Fayyad & Stolorz, 1997). The problem seems formidable without change in paradigms. Dealing with large data sets will always present challenges and paradigms such as parallel computing may need to be introduced (ibid). Mining image data such as customer shopping patterns will present even greater challenges, one of which will be the visualization of data. Trends and patterns need to be presented in their natural context, which will also increase demands on the mining hardware and software. Mining is of utmost importance in a number of scientific and industrial applications, all of which will have the effect of driving advances in technology.

CONCLUSIONS AND IMPLICATIONS

The mining literature has focused mostly on applications, algorithms, and techniques. There has been comparatively less emphasis on pre- and post-mining activities. It is important for IS researchers to address this lacunae. We have discussed some of these activities as part of the KDD process. Pre-mining activities such as data selection and transformation are influenced by the questions of interest. We have suggested domain modeling augmented by traditional analysis techniques as a way to understand the issues at hand. The AACSB case study clearly illustrates the importance of such models.

The potential of mining has been barely scratched. Organizations have sizeable and expanding repositories of data. Competition will always force organizations to tweak every aspect of its operations for the utmost efficiency. To exploit mining, a theoretical framework is required for understanding the relationship between the domain and KDD activities. Also of strategic importance is an understanding of the applicability of models to different data and organizational contexts.

REFERENCES

Amaravadi, C., & Daneshgar, F. (2003). The role of data mining in organizational cognition. In H. Nemati & C. Barko (Eds.), *Organizational data mining* (pp. 46-60). Hershey, PA: Idea Group Publishing.

Back, B., Toivonen, Vanharanta, H., & Visa, A. (2001). Comparing numerical data and text information from annual

reports using self-organizing maps. *International Journal of Accounting Information Systems, 2*(4), 249-269.

Chen, Y-S., Justis, R., & Chong, P. P. (2003). Data mining in franchise organizations. In H. Nemati & C. Barko (Eds.), *Organizational data mining*. Hershey, PA: Idea Group Publishing.

Eden, C., & Ackermann, F. (2004). Cognitive mapping expert views for policy analysis in the public sector. *European Journal of Operational Research, 152*(3), 615-630.

Fayyad, U., & Stolorz, P. (1997). Data mining and KDD: Promise and challenges. *Future Generation Computer Systems, 13*(2-3), 99-115.

Hand, D. J., & Blunt, G. (2001). Prospecting for gems in credit card data. *IMA Journal of Management Mathematics, 12*(2), 173-200.

Hogl, O. J., Muller, M., Stoyan, H., & Stuhlinger, W. (2001). Using questions & interests to guide data mining for medical quality management. *Topics in Health Information Management, 22*(1), 36-50.

Hui, S. C., & Jha, G. (2000). Data mining for customer service support. *Information & Management, 38*(1), 1-13.

Jackson, J. (2002). Data mining: A conceptual overview. *Communications of the Association for Information Systems, 8*, 267-296.

Kimball, R. (2003). RFID tags and smart dust. *Intelligent enterprise,* July 18. Retrieved from http://www.intelligententerprise.com/030718/612warehouse1_1.jhtml

Kumar, V. (2002, January). *Data mining algorithms*. Tutorial presented at IPAM 2002 Workshop on Mathematical Challenges in Scientific Data Mining.

Lin, F. Y., & McClean, S. (2001). A data mining approach to the prediction of corporate failure. *Knowledge-Based Systems, 14*(3), 189-195.

Liu, L. M., Bhattacharyya, S., Sclove, S. L., Chen, R., & Lattyak, W. J. (2001). Data mining on time series: An illustration using fast-food restaurant franchise data. *Computational Statistics and Data Analysis, 37*, 455-476.

Ng, K., & Liu, H. (2000) Customer retention via data mining. *Artificial Intelligence Review, 14*(6), 569-590.

Peacock, P. R. (1998). Data mining in marketing: Part 2. *Marketing Management, 7*(1), 8-18.

Thearling, K., Becker, B., DeCoste, D., Mawby, B., Pilote, M., & Sommerfield, D. (2001). Visualizing data mining models. In U. Fayyad, G. Grinstein, & A. Wierse (Eds.), *Information visualization in data mining and knowledge discovery*. San Mateo, CA: Morgan Kaufman.

KEY TERMS

Association: A technique in data mining that attempts to identify similarities across a set of records, such as purchases that occur together across a number of transactions.

Classification: A technique in data mining that attempts to group data according to pre-specified categories such as "loyal customers" vs. "customers likely to switch."

Clustering: A technique in data mining that attempts to identify the natural groupings of data, such as income groups to which customers belong.

Data Driven: If the data drive the analysis without any prior expectations, the mining process is referred to as a data driven approach.

Dimensionality: Dimensionality refers to the number of attributes. For instance, in sales data, "product," "sales location," and "time" are attributes.

Micro Theories: Beliefs that need to be tested during the data mining process.

Overfitting: A condition that occurs when there are too many parameters in a model. In such cases, the model learns the idiosyncrasies of the test data set. This can happen in models such as regression, time series analysis, and neural networks.

Question Driven: In question-driven (hypothesis-driven) approaches, the analysis is preceded by an identification of questions of interest.

APPENDIX: EXPLANATION OF VARIABLES

ATTRIBUTE NAME	**DESCRIPTION**
Name	Name of institution
Cl	Carnegie classification
Mission	Teaching, intellectual contribution, service
Research	Applied, instructional, or basic
Commun	Whether urban or rural
OpBudget	Operating budget
OpBud/FTFac	Operating budget/Full time faculty member
UGrad	Undergraduate tuition
MBA	MBA tuition
FTEQ	Full time equivalent faculty
FT	Full time faculty
FTPHD	Full time PhD
FTSt/FTFac	Full time students per full time faculty
UGrad	Residential or commuter students
Grad	Residential or commuter students
FT	Number of full time undergraduate students
PT	Number of part time undergraduate students
TDegrees	Total number of undergraduate degrees awarded
AcctDegs	Undergraduate accounting degrees awarded
FT	Full time graduate enrollment
PT	Part time graduate enrollment
Tdegrees	Total number of graduate degrees awarded
FT	Full time degrees in special programs
PT	Part time degrees in special programs
Tdegrees	Total degrees in special programs

Extensions to UML Using Stereotypes

Daniel Riesco
Universidad Nacional de San Luis and Universidad Nacional de Rio Cuarto, Argentina

Marcela Daniele
Universidad Nacional de Rio Cuarto, Argentina

Daniel Romero
Universidad Nacional de Rio Cuarto, Argentina

German Montejano
Universidad Nacional de San Luis, Argentina

INTRODUCTION

The Unified Modeling Language (UML) allows to visualize, to specify, to build and to document the devices of a system that involves a great quantity of software. It provides a standard form for writing the models of a system, covering so much of the conceptual aspects (such as processes of the business and functions of the system) as the concrete ones (such as the classes written in a specific programming language, schemas of databases and software components).

In 1997, UML 1.1 was approved by the OMG becoming the standard notation for the analysis and the design oriented to objects. UML is the first language of modelling in which a metamodel in its own notation has been published. It is a strict subset called Core. It is a self-referential metamodel.

It is a very expressive language that covers all of the necessary views to develop and to deploy systems. UML is a language that provides three extension mechanisms (Booch, Rumbaugh, & Jacobson, 1999): stereotypes, tag values, and constrains. The stereotypes allow to create new types of elements of model based on the elements that form the metamodel UML extending the semantics of the same one, the tag values are an extension of the properties of an element of UML, allowing to add new information to the specification of the same one, and the constrains are an extension of the semantics of UML that allow to add new rules or to modify the existent ones.

The organization of this overview is given in the following way: first, we present the stereotypes according to the standard of OMG; second, we expose the analysis of works that extend UML using stereotypes in diverse real domains; third, we make an analysis of the stereotypes of UML; and we finish giving a general conclusion where we focus ourselves in the distinction of the works according to their inclusion or not of the created stereotypes in the metamodel of UML.

STEREOTYPE ACCORDING TO THE STANDARD OF OMG

A stereotype provides a form of classifying elements in such a way that they work in some aspects as if they were instances of a new constructor of the "virtual"

Attributes	
BaseClass	This specifies the names of one or more elements from UML modeling to which the stereotype is applied, such as classes, associations.
Icon	This is the geometric description of the icon that will be used to present an image of the element of the marked model with the stereotype.

metamodel. A stereotype could also be used to indicate a meaning or different use between two elements with identical structure. A stereotype can also specify a geometric icon to be used to present elements with the stereotype.

USING STEREOTYPES IN DIVERSE REAL DOMAINS

UML adapts to any technique, because it has extension mechanisms that don't need to redefine the nucleus UML, allowing to obtain a modeling more appropriate to the different particular domains. All of the extensions should follow the standard proposed by the OMG (2001).

Modeling of Business with UML

UML was initially designed to describe aspects of a software system. For the modeling of business, UML needed to be extended to identify and to visualize resources, processes, objectives and rules more clearly. These are the primary concepts that define a business system. The Eriksson-Penker Business Extensions (Eriksson & Penker, 1999) provide

new stereotypes for their model. In a diagram of class of UML, they represent a process through a specific symbol that corresponds to an activity stereotyped in an activity diagram. The resources used by the process are modeled with a stereotyped dependence *«supply»* and the resources controlled by the process are modeled with a stereotyped dependence *«control»*.

Modeling of Web Applications with UML

In Baresi, Garzotto, and Paoloni (2001), they propose a framework denominated W2000, for the design of Web applications. They combine the use of UML and HDM (Hypermedia Model). The Web applications require the integration of two different, but interrelated, activities: the design hypermedia that is focused in the navigation way and the structure of the information, and the functional design that is focused in the operations. Among their main purposes is the extension to the standard of UML of the dynamic diagrams.

It uses *«node type»* like stereotype of UML to define a node type that allows to reach different structures of information and defines the symbol "@" to indicate that a node stereotyped with *«node type»* will be the node for defect where all the users begin to navigate.

They propose a symbol called "index" that allows the users of a navigation to select one of the elements from a list of indexes.

The "collection links" define how the users can navigate between the core and the members of a collection, and they add a symbol to graphically represent the pattern "Index + Guided Tour".

For the functional design, they define the diagrams of scenarios, and these are represented through an extended sequence of diagrams of UML. The extensions refer as much to objects as to the step of messages. The objects are organized in entities (components and nodes), semantic associations and collect. The "free navigation" is represented with dotted lines and the "constrained navigation" is represented with a line with a diamond in it.

In 1999, Jim Conallen, Principal Consultant of Conallen Inc., Object Oriented Application Development in Conallen Inc., presented in their paper an extension to UML, in a formal way, to model applications Web.

The extension was presented at several other conferences in 1999, including the Rational Users Conference in Seattle (July 1999), and two Wrox Press ASP conferences in Washington, DC (September 1999) and in London (November 1999).

Various summaries and introductions to the extension have or will appear in the *Communications of the ACM* (ASPToday, *http://www.asptoday.com/articles/19990517.htm*), and in the UML Resources Web site at Rational Software. A full explanation of this work is currently being prepared for the book, *Building Web Applications with UML* (Conallen, 2002), published in the Object Technology Series of Addison-Wesley Longman.

This article presents an extension of UML for Web application designs. Part of the extension mechanism of UML is the ability to assign different icons to stereotyped classes. A list of prototype icons for the most common class stereotypes can be found as an appendix. It defines two new stereotypes to model the difference between the executed methods in the server and the executed functions in the client. In a page, a method that executes on the server will be stereotyped as *«server method»* and functions that run on the client *«client method»*. This solves the problem of distinguishing attributes and methods of a page object. It proposes the modeling of a page with two stereotyped classes, *«server page»* and *«client page»*. They define several stereotypes to represent the associations, such as: *«builds»* that is modeled with an unidirectional association from the server page to client page, *«redirects»* to model the redirection to other *«server page»*, *«links»* for defined associations between pages clients and other pages (client or server). Also, they define stereotypes to model components, *«server component»* and *«client component»*, for Forms *«form»*, for Framesets *«frameset»*. Other defined stereotypes, *«scriplet»* for cached client page, and *«xml»* for a hierarchical data object that can be passed to and from a Web server and client browser.

In Gorshkova and Novikov (2001), a UML extension capable to refine the design of the client part of Web application is defined. Several new diagrams are specified which provide a precise definition of the content of Web pages and navigation between them. The composition diagram is a special case of class diagram. We use it to express the structure of the Web pages and identify their content: how they are connected together and what data is carried from one page to another. The main notion of the composition diagram is the page, defined as an autonomous block of screen. Each screen in the navigation diagram is mapped into several pages in the composition diagram. The tool may provide links from pages to screens and vice versa to show their relationship. A page is modeled in the composition diagram as a class stereotyped *«page»*. A page may play the role of container for other pages. Nested pages are modeled as aggregated classes. The page has elements like buttons, links and input fields. They are modeled as attributes of the corresponding page. The *«form»* stereotype is a child of *«page»*. It is used to model HTML forms. The navigable association between source and target pages is stereotyped *«link»*. Each *«link»* has a tag context with expression as value.

In Koch, Baumeister, and Mandel (2000), the authors define a set of stereotypes that are used in the construction of intuitive analysis and design models in the development of Web applications. These models are the navigation space model, the navigation structure model, and the static presentation models.

The basis of the navigation design is the conceptual model, and the outcome is a navigational model, which can be seen as a view over the conceptual model. The navigational model is defined in a two-step process. In the first step, the navigational space model is defined, and in the second, the navigational structure model is built. The navigation space model defines a view on the conceptual model showing which classes of the conceptual model can be visited through navigation in the Web application. The navigational structure model defines the navigation of the application. It is based on the navigation space model, but also additional model elements are included in the class diagram to perform the navigation between navigational objects: menus, indexes, guided tours, queries, external nodes, and navigational contexts.

The static presentational model is represented by UML composition diagrams that describe how the user interfaces are built. They define stereotypes to be able to build these diagrams.

The set of defined stereotypes is the *«Navigational Class»*. It represents a class whose instances are visited during navigation. The *«Direct Navigability»* are associations in the navigation model. These associations are interpreted as direct navigability from the source navigation class to the target navigation class. An *«Index»* is modeled by a composite object which contains an arbitrary number of index items. A *«Guided Tour»* is an object which provides sequential access to the instances of a navigational class. A *«Query»* is represented by an object which has a query string as an attribute. This string may be given, for instance, by an OCL select operation. A *«Menu»* is a composite object which contains a fixed number of menu items. A *«Presentational class»* models the presentation of a navigational class or an access primitive, such as an index, a guided tour, query, or menu. A *«Frameset»* is a top-level element which is modeled by a composite that contains (lower level) presentational objects but may also contain an arbitrary number of nested framesets. An area of the frameset is assigned to each lower level element, so called *«frame»*, the same stereotype is also used in Eriksson and Penker (1999). A *«window»* is the area of the user interface where framesets or presentational objects are displayed, and also defined are *«text»*, *«anchor»*, *«button»*, *«image»*, *«audio»*, *«video»* and *«form»*. A *«collection»* is a list of text elements that is introduced as a stereotype to provide a convenient representation of composites. An *«anchored collection»* a list of anchors.

Real-Time Systems Modeling with UML

In Selic and Rumbaugh (1998), a set of constructs that facilitate the design of software architectures in the domain of real-time software systems is described. The constructs, derived from field-proven concepts originally defined in the ROOM methodology (http://wwweb.org/smo/bmc/mb/mb35.html) are specified using the UML standard. In particular, it showed how these architectural constructs can be derived from more general UML modeling concepts by using the powerful extensibility mechanisms of UML. The following stereotypes are defined as UML extension: *«protocol»*, *«protocolRole»*, *«port»*, *«capsule»,* and *«chainState»*.

A protocol role is modeled in UML by the *«protocolRole»* stereotype of Metamodel Class ClassifierRole. A protocol is modeled in UML by the *«protocol»* stereotype of Metamodel Class Collaboration with a composition relationship to each of its protocol roles representing the standard relationship that a collaboration has with its "owned elements". A port object is modeled by the *«port»* stereotype, which is a stereotype of the UML Class concept. A capsule is represented by the *«capsule»* stereotype of Class. The capsule is in a composition relationship with its ports, sub-capsules (except for plug-in sub-capsules), and internal connectors. A state whose only purpose is to "chain" further automatic (triggerless) transitions onto an input transition is defined as a stereotype *«chainState»* of the UML State concept.

In Toetenel, Roubtsova, and Katwijk (2001), the paper shows an extension UML with mechanisms for specifying temporal constrains and properties. It defines schemes for the translation of UML specifications into semantically equivalent XTG-based (eXtended Timed Graphs) specifications such as the properties given on UML specification can be proved on the XTG specification. The realization of the approach uses the extensibility interface of the Rational Rose UML Tool (Rational Rose 98i, 2000).

ANALYSIS OF UML STEREOTYPES

In Gogolla and Henderson-Sellers (2002), the paper takes up ideas from Gogolla (2001) where stereotypes have been introduced for relational database design. The expressiveness of UML stereotypes has been analyzed, and some concrete suggestions have been made for the improvement of the UML metamodel. Use OCL to define precise stereotypes.

In Riesco, Martellottto, and Montejano (2003) and Riesco, Grumelli, Maccio, and Martellotto (2002), proposals of "evolutionary stereotypes" are presented. These are incorporated into the modeling tool in such a way that they can extend the UML metamodel, including the new elements with their corresponding semantics. In this way, the environment of a tool can dynamically change its appearance and functionality to allow software engineers to use the stereotypes previously defined in the diagrams.

The Clark, Evans, Kent, Brodsky, and Cook study[1] proposes a new metamodelling facility (MMF) containing: Metamodelling Language (MML); Metamodelling Tools (MMT); a satisfaction checker (for instance, does X satisfy constraint C from model M?); to check that a model satisfies

its metamodel; to check that a metamodel satisfies the MML rules; to check that MML satisfies the MML rules.

CONCLUSION

UML is a universal language adopted for the modeling of applications in a wide range of domains. It is an open language, which provides extension mechanisms in order to extend the metamodel. The UML extension mechanisms include: Tag values, Constrains and Stereotypes.

In particular, the stereotypes should be carefully declared and should only be used when the message to be communicated could not be expressed in any other UML terms. In order to achieve a better use of stereotypes, there are certain necessary conditions. The UML metamodel should be adjusted and support tools should be provided for the complete part of the UML metamodel dealing with stereotypes.

The papers analyzed in this work propose a varied number of stereotypes to extend UML, with the purpose of using this language to model particular domains, such as Web applications, real-time systems, business modeling, XML, and so forth.

After the reading and the analysis of these papers, it has been detected that only in three of them (Gogolla, 2001; Riesco, Grumelli, Maccio, & Martellotto, 2002; Selic & Rumbaugh, 1998) the addition of stereotypes is defined as a UML complete extension. This occurs through the incorporation of new elements to the metamodel with new semantic that assures the consistency of the metamodel UML.

In the other analyzed papers, the stereotypes are presented being used in their specific context, but without extending the metamodel UML. This is a clear deficiency in their definitions, since the maintenance of the metamodel UML in a consistent way is very difficult, or even impossible, as their stereotypes are not explicitly added to the metamodel UML. Besides, this means that the stereotypes won't be available to be used in generic solutions to problems belonging to the selected particular domain.

REFERENCES

ASPToday. Available at *http://www.asptoday.com/articles/19990517.htm*

Baresi, L., Garzotto, F., & Paoloni, P. (2001). *Extending UML for modeling Web applications.*

Booch, G., Rumbaugh, J., & Jacobson I. (1999). *The Unified Modeling Language user guide.* Addison-Wesley.

Conallen, J. (1999). *Modeling Web applications with UML.*

Conallen, J. (2002). *Building Web applications with UML* (2nd ed.). Addison-Wesley.

Eriksson, H.-E., & Penker, M. (1999). *Business modeling with UML: Business patterns at work*, Wiley & Sons.

Gogolla, M. (2001). Using OCL for defining precise, domain-specific UML stereotypes. In A. Aurum & R. Jeery (Eds.), *Proc. 6th Australian Workshop on Requirements Engineering (AWRE'2001)* (pp. 51-60). Centre for Advanced Software Engineering Research (CAESER), University of New South Wales, Sydney, 2001.

Gogolla, M., & Henderson-Sellers, B. (2002). Analysis of UML stereotypes within the UML metamodel. In J.-M. Jezequel, H. Hussmann, & S. Cook (Eds.), *Proc. 5th Conf. Unified Modeling Language (UML'2002)*, Springer, Berlin, LNCS.

Gorshkova, E., & Novikov, B. (2001). UML extensibility in the design of Web applications. *Exploiting*.

Koch, N., Baumeister, H., & Mandel, L. (2000). Extending UML to model navigation and presentation in Web applications. In G. Winters & J. Winters (Eds.), *Modeling Web applications, Workshop of the UML'2000*. York, England, October.

OMG (2001). OMG Unified Modeling Language specification. Retrieved from the World Wide Web at *http://www.omg.org*

Rational Rose 98i. (2000). Rose Extensibility Reference. Rational Software Corporation, *http://www.rational.comwww.se.fhheilbromm.de/usefulstuff/Rational Rose 98i Documentation*

Rational Software and Miller Freeman, Inc, a United Newa & Media Company. (1999). Business modeling with UML. Retrieved from the World Wide Web at *http://www.therationaledge.com/rosearchitect/mag/archives/fall99/f5.html*

Riesco, D., Grumelli, A., Maccio, A., & Martellotto, P. (2002). Extensions to UML metamodel: Evolutionary stereotypes. *3rd ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel / Distributed Computing*, Madrid, Spain.

Riesco, D., Martellottto, P., & Montejano, G. (2003). Extension to UML using stereotypes. In L. Favre (Ed.), *UML and the Unified Process* (Chapter XIV, p.40), Hershey, PA: IRM Press.

ROOM Methodology. *http://wwweb.org/smo/bmc/mb/mb35.html*

Selic, B., & Rumbaugh, J. (1998). *Using UML for modeling complex real-time systems.*

Toetenel, H., Roubtsova, E., & Katwijk, J. (2001). A timed automata semantics for real-time UML specifications. *Proceedings of the IEEE Symposia on Human-Centric Computing Languages and Environments (HCC'01).*

KEY TERMS

Class Diagram: Show the classes of the system, their interrelationships, and the collaboration between those classes.

Extension Mechanisms: Specify how model elements are customized and extended with new semantics.

Metamodel: An abstraction which defines the structure for a UML model. A model is an instance of a metamodel. Defines a language to describe an information domain.

Object Constraint Language (OCL): A notational language for analysis and design of software systems. It is a subset of the industry standard UML that allows software developers to write constraints and queries over object models.

OMG: Has been "Setting The Standards For Distributed Computing™ through its mission to promote the theory and practice of object technology for the development of distributed computing systems. The goal is to provide a common architectural framework for object-oriented applications based on widely available interface specifications (OMG, 2001).

Real Domains (Particular or Specific): The different application areas that can require to be modeled with UML. For example: Web applications, real-time system, XML, business modeling, frameworks, communication protocols, workflows, geographical information systems, and so forth.

Stereotype: Allows to create new types of elements of modeling, based on the elements that form the goal-pattern UML, extending its semantics.

ENDNOTE

[1] See www.puml.org for the document "A Feasibility Study in Rearchitecting UML as a Family of Languages using a Precise OO Meta- Modeling Approach" (Clark, Evans, Kent, Brodsky, Cook) and associated tools.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1169-1173, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Software & Systems Design

Extreme Programming for Web Applications

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

The engineering environment of Web Applications is in a constant state of technological and social flux. These applications face challenges posed by new implementation languages, variations in user agents, demands for new services, and user classes from different cultural backgrounds, age groups, and capabilities.

We require a methodical approach towards the development *life cycle* and maintenance of Web Applications that can adequately respond to this constantly changing environment. In this article, we propose the use of an *agile methodology* (Highsmith, 2002), namely Extreme Programming (XP) (Beck & Andres, 2005), for a systematic development of Web Applications. In general, agile methodologies have show to be cost-effective for projects with certain types of uncertainties (Liu, Kong, & Chen, 2006) and, according to surveys (Khan & Balbo, 2005), been successfully applied to Web Applications.

The organization of the article is as follows. We first outline the background necessary for the discussion that pursues and state our position. This is followed by a discussion of the applicability and feasibility of XP practices as they pertain to Web Applications. Then the shortcomings of XP towards Web Applications are highlighted, and suggestions for improvement are presented. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

Over the last decade, Web Applications have become increasingly large and complex as they respond to the expectation of sophisticated services. The need for managing increasing size and complexity of Web Applications and the necessity of a planned development was realized in the late 1990s (Coda, Ghezzi, Vigna, & Garzotto, 1998; Powell, Jones, & Cutts, 1998). This led to the introduction of the notion of Web Engineering (Ginige & Murugesan, 2001), which subsequently has been treated comprehensively (Kappel, Pröll, Reich, & Retschitzegger, 2006).

In the last few years, a number of methods for realizing Web Applications have been proposed including, but not limited to, Web Site Design Method (WSDM) (De Troyer & Leune, 1998), Object-Oriented Hypermedia Design Method (OOHDM) (Schwabe & Rossi, 1998), and Web Object-Oriented Model (WOOM) (Coda, et al., 1998). However, the focus in these approaches has been on specific aspects of Web Applications (like modeling, designing, or implementing) rather than on the *process*.

We adopt the most broadly-used and well-tested *agile methodology*, namely XP, for the development of Web Applications. There are several differences between traditional software and Web Applications (including aspects of delivery, legality, privacy, security, and *usability*) that makes the realization of XP challenging.

XP is a test-driven, "lightweight" methodology designed for small teams that emphasizes customer satisfaction and promotes teamwork. XP was created to tackle uncertainties in development environment, and in doing so, put more emphasis on the social (people) component (engineer, customer, and end-user). The XP practices are set up to mitigate project risks (dynamically changing requirements, new system due by a specific time line, and so on) and increase the likelihood of success. The use of XP has been suggested for a "rapid application development" of Web Applications (Maurer & Martel, 2002; Wallace, Raggett, & Aufgang, 2002), however, a detailed analysis has not been carried out.

It is not the purpose of this article to evaluate the merit of XP on its own or with respect to other *agile methodologies*; such assessments have been carried out elsewhere (Mnkandla & Dwolatzky, 2004).

ENGINEERING WEB APPLICATIONS USING EXTREME PROGRAMMING

In this section, we discuss in detail how the *twelve practices* put forth by XP manifest themselves in the development of Web Applications (Table 1).

We note that some of these practices such as *Testing*, *Refactoring*, or *Pair Programming* are not native to XP and were discovered in other contexts previously. In this sense, by aggregating them in a coherent manner, XP bases itself on "best" practices. These practices are also not necessarily mutually exclusive and we point out the relationships among them where necessary. We also draw attention to the obstacles in the realization of these practices that pose challenges to the deployment of XP for Web Applications.

Table 1. XP practices corresponding to process workflows in a Web application

Process Workflow	XP Practices
Planning	40-Hour Week, The Planning Game (Project Velocity)
Analysis (Domain Modeling, Requirements)	On-Site Customer, The Planning Game (User Stories)
Design	Metaphor Guide (Natural Naming), Simple Design, Refactoring
Implementation	Collective Ownership, On-Site Customer, Metaphor Guide, Coding Standards, Pair Programming, Continuous Integration
Verification and Validation	On-Site Customer, Testing (Unit Tests, Acceptance Tests)
Delivery	Small Releases

1. *The Planning Game.* The purpose of *The Planning Game* is to determine the scope of the project and future releases by combining business priorities and technical estimates. For that, it first solicits input from the "customer" to define the business value of desired features and uses cost estimates provided by the programmers. This input documented in form of *user stories* (Alexander & Maiden, 2004). A *user story* is a user experience informally expressed in a few lines with a Web Application such as navigating or using a search engine. The estimation is limited to the assessment of *project velocity*, a tangible metric that determines the pace at which the team can produce deliverables. The plan is iterative: it is prone to modifications based on the current reality. For example, a new user story may lead to revisitation and evolution of the current plan.
2. *Small Releases.* The idea behind *Small Releases* is to have a simple system (an evolutionary prototype) into production early, and then via short cycles, iteratively, and/or incrementally, reach the final system. To have a concrete proof-of-concept up-and-running can be used to solicit feedback for future versions and can help convince customers and managers of the viability of the project. This is useful for Web Applications that are highly interactive such as those making broad use of fill-out-forms. However, there is cost associated with prototypes and, therefore, their number should be kept under control.
3. *Metaphor Guide.* The use of metaphors (Boyd, 1999) is prevalent in all aspects of software development. A *Metaphor Guide* is an effort to streamline and standardize efforts for naming software objects and is available for team-wide use. The main concerns in naming are of sensitivity to the domain under consideration (that is, use of terminology of the application area) and familiarity (as user background is assumed to be non-technical). *Natural naming* (Keller, 1990) is a technique initially used in source code contexts that encourages the use of names that consist of one or more full words of the natural language for program elements in preference to acronyms or abbreviations. Indeed, natural naming strengthens the link between the underlying conceptual entity and its given name. For example, DeviceProfile is a combination of two real-world metaphors placed into a natural naming scheme.
4. *Simple Design.* The motivation behind a *Simple Design* is that in XP's view, requirements are *not* complete when the design commences. This is inline with the reality of Web Applications which have to respond to the market pressures and the competition that are beyond their control, or other unavoidable circumstances such as variations in implementation technology. Therefore, the design is minimal based on *current* (not future) requirements. It aims for simplicity, and to ensure "good" design, its quality (specifically, structural complexity) is improved by frequent revisitations; that is, *Refactoring*. The interfaces of Web Applications are particularly amenable for simple design as they are likely to change often during the process. Design patterns (Van Duyne, Landay, & Hong, 2003) are forms of reusable knowledge based on past experience and expertise that can lead to simplified design.
5. *Testing.* There is a strong emphasis in XP on validation and verification of the software at all times. By being test-driven, there is transition from one phase to another only if the tests succeed. The tests range from *unit tests* (using tools such as HTMLUnit, HTTPUnit, XMLUnit, XSLTUnit, and JUnit) written by programmers to acceptance tests involving customers

(to satisfy customer requirements). There are variations in user agents (browsers) with respect to their support for information representation (markup or style sheet) languages. Therefore, interface testing is uniquely critical to Web Applications. As part of that activity, syntactical validation of documents being served is critical. A detailed treatment of tools, techniques, and methods for testing Web Applications as well as for test plans is given in (Nguyen, Johnson, & Hackett, 2003).

6. *Refactoring.* The artifacts created during analysis or design may need to evolve for reasons such as discovery of "impurities" (or code "smells") or obsolescence. The refactoring (Fowler, Beck, Brant, Opdyke, Roberts, 1999) methods are structural transformations that help eradicating the undesirables without changing the functionality of the application. Examples of such smells include inconsistent names of classes, operations, or attributes that hinder communication, redundancy (duplication), classes with unnecessary responsibility (non-cohesivity), and so on. The goal of *Refactoring* is to improve the design of the system throughout the entire development.
7. *Pair Programming.* This is one of the practices of XP that highlights the social aspects of engineering. The idea behind *Pair Programming* is to encourage collaborative work. In some controlled experiments (Williams & Kessler, 2003), Pair Programming has been shown to produce better code at similar or lower cost than programmers working alone. Empirical studies (Katira, 2004) have shown that some level of compatibility among partners in Pair Programming is necessary for it to be effective. The notion of Pair Programming can be extended to artifacts created during early stages (namely, analysis and design phases) of the development process that focus on modeling (Kamthan, 2005). For example, the use of the Unified Modeling Language (*UML*) (Booch, Jacobson, & Rumbaugh, 2005) for visual modeling of Web Applications has been suggested (Conallen, 2003). A pair can be responsible for several other practices such as using *Refactoring* to obtain a *Simple Design*, *Continuous Integration*, and *Testing*. The *On-Site Customer* can be a partner in a pair but only as a co-pilot.
8. *Collective Ownership.* According to the XP philosophy, one of reasons for inertia in modifications to software is that when change is warranted, the team has to wait for specific personnel to carry it out. Therefore, in XP, all the code belongs to all the programmers and anyone can change code anywhere in the system at any time. However, for such an arrangement to be effective, configuration management that provides trace of person, date/time, and of nature and location of the change carried out is needed.
9. *Continuous Integration.* In an incremental and iterative approach of XP, standalone units (such as a corporate logo, navigation icons, and so on) are created and then integrated. However, it is not automatic that if the individual parts work, then their sum would also work. For example, a graphical navigation bar may work well individually, but not when included in an Extensible HyperText Markup Language (XHTML) document due to, say, incorrect encoding or link syntax. By "continuous," XP means integrating and building the software system multiple times a day. The advantage of *Continuous Integration* is minimal propagation of errors (limited to the last addition).
10. *40-Hour Week.* The term *40-Hour Week* is to be taken figuratively rather than literally. It simply implies that, due to the emphasis on the social aspect in XP, "overwork" is not recommended. XP believes that excessive overtime leads to low productivity in the long-term. For example, tired programmers are prone to more mistakes, which in turn may slow down progress of the project.
11. *On-Site Customer.* The availability of a full-time *On-Site Customer* helps in understanding the application domain, determining requirements, setting priorities, and answering questions as the programmers have them. In XP, every contributor to the project, including the customer, is an integral part of the *entire* team. This has two major implications for the team: its structure is not hierarchical and it requires physical proximity of the participants to function.
12. *Coding Standards.* There are a variety of languages for expressing information in Web Applications, which can be served using any general or special-purpose programming language. It is critical that instances based on these languages be communicable to stakeholders and interoperable with respect to user agents. For example, for *Pair Programming* and for *Collective Ownership* to be effective, there needs to be a common understanding. *Coding Standards* provide means for doing that. It is known (Schneidewind & Fenton, 1996) that, when applied judiciously, standards can contribute towards quality improvement.

CHALLENGES TO THE DEPLOYMENT OF EXTREME PROGRAMMING FOR WEB APPLICATIONS

In this section, we highlight certain caveats of applying XP practices as-is as well as certain aspects that are essential to Web Applications but are not covered by these practices *per se*.

- XP does not mandate a rigorous feasibility study, including a formal means for cost estimation, as part of *The Planning Game*. A feasibility study could, for instance, determine if one could take advantage of reuse. For example, the functionality of a Web Application for different classes of desktop computers should not be all that different.
- The notions of *Pair Programming*, *40-Hour Week*, and *On-Site Customer* make sense for a development in a non-distributed environment only. This would present a coordination obstacle if a Web Application were being developed in different natural languages, each in different parts of the world.
- Testing for accessibility or *usability* of Web Applications can be prohibitive for small-to-medium size enterprises, particularly if it involves specialized rooms, dedicated infrastructure with video monitoring and recordings for subsequent analysis. Also, testing cannot always detect all errors in systems. For example, that user supplied correct address, or that internal documentation corresponds to source code, are beyond the scope of testing. Furthermore, testing is only one approach to verification and defect removal. XP does not include any support for formal inspections (Wiegers, 2002), although that is somewhat ameliorated by support for *Pair Programming*, which can be viewed as "informal" inspections.

Finally, we point out that agility is not a panacea (Boehm & Turner, 2004) and there are several issues associated with *agile methodologies* in general, and XP in particular. For example, XP is not applicable to large (greater than 15) team sizes, distributed development, or for very large projects. There is no explicit support for metrics or rigorous measurement in XP. XP does not explicitly take into account the licensing conditions under which the software is developed. For example, Web Applications that are Open Source or outsourced will not be able to comply with some of the practices mandated by XP. However, XP provides a feasible first step from an ad-hoc approach to an organized view of developing Web Applications.

FUTURE TRENDS

Some degree of cost estimation is relevant to any software development. COCOMO II (Boehm, Abts, Brown, Chulani, Clark, Horowitz, et al., 2001) provides a rigorous approach to cost estimation, and could assist in *The Planning Game*. However, that would require some adjustments in measures and calibrations of data, as Web Applications are different from traditional applications for which the COCOMO II cost estimation model is defined.

Since XP is driven by testing, there is an urgent need for unit testing frameworks (Hamill, 2004) for Web Applications beyond those that are currently available for testing documents of markup languages.

For a widespread acceptance of the proposed use of XP in Web Applications, its adoption in initiatives for *standard*-izing the development of Web Applications (IEEE, 2003) will be crucial.

An extension of XP to Web Applications developed within the Semantic Web architecture (Hendler, Lassila, & Berners-Lee, 2001), such as interfaces to traditional reasoners for ontological inferencing, would be of interest.

Finally, for large-scale Web Applications, a "heterogeneous" process environment approach that mixes agility with discipline (Boehm & Turner, 2004) could be useful. A natural extension of the previous discussion would be to deploy a simplified version of the Unified Process (UP) (Jacobson, Booch, & Rumbaugh, 1999), which is a process *framework* that can be tailored to produce a process model, to Web Applications. Indeed, such a "WebUP" would, on one hand, be model-driven, iterative, customer-centric, and would, on the other hand, still emphasize a top-down team hierarchy and document-based communication.

CONCLUSION

Web Applications continue to increase in size and complexity, and to sustain and manage this growth, require a systematic approach towards their development. For that, there is a need to move away from thinking at the implementation language level and focus on abstractions created *earlier* in the process. At the same time, we wish to avoid the bureaucracy in development processes that have plagued Software Engineering in the past.

XP provide one such viable option for development of small-to-medium size Web Applications for new or not-well-understood domains, and where close collaboration within the team and with the customer is encouraged. The aforementioned shortcomings inherent to XP are largely resolvable, and pave the way towards improvements as well as considerations for other process models tailored to Web Applications.

REFERENCES

Alexander, I., & Maiden, N. (2004). *Scenarios, Stories, Use Cases Through the Systems Development Life-Cycle*. John Wiley and Sons, Inc.

Beck, K., & Andres, C. (2005). *Extreme Programming Explained: Embrace Change* (2 Ed.). Addison-Wesley.

Boehm, B. W., Abts, C., Brown, A. W., Chulani, S., Clark, B. K., Horowitz, E., Madachy, R., Reifer, D., & Steece, B. (2001). *Software Cost Estimation with COCOMO II*. Prentice Hall.

Boehm, B., & Turner, R. (2004). *Balancing Agility and Discipline: A Guide for the Perplexed*. Addison-Wesley.

Booch, G., Jacobson, I., & Rumbaugh, J. (2005*). The Unified Modeling Language Reference Manual (Second Edition)*. Addison-Wesley.

Boyd, N. S. (1999). Using Natural Language in Software Development. *Journal of Object-Oriented Programming*, 11(9).

Coda, F., Ghezzi, C., Vigna, G., & Garzotto, F. (1998). Towards a Software Engineering Approach to Web Site Development. *The Ninth International Workshop on Software Specification and Design (IWSSD-9)*. Ise-shima, Japan. April 16-18, 1998.

Conallen, J. (2003). *Building Web Applications with UML* (2 Ed.). Addison-Wesley.

De Troyer, O., & Leune, C. (1998). WSDM: A User-Centered Design Method for Web Sites. *The Seventh International World Wide Web Conference (WWW7)*. Brisbane, Australia, April 14-18, 1998.

Fowler, M., Beck, K., Brant, J., Opdyke, W., & Roberts, D. (1999). *Refactoring: Improving the Design of Existing Code*. Addison-Wesley.

Ginige, A., & Murugesan, S. (2001). Web Engineering: An Introduction. *IEEE Multimedia*, 8(1), 14-18.

Hamill, P. (2004). *Unit Test Frameworks: Tools for High-Quality Software Development*. O'Reilly Media, Inc.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The Semantic Web. *Scientific American*, 284(5), 34-43.

Highsmith, J. (2002). *Agile Software Development Ecosystems*. Addison-Wesley.

IEEE. (2003). *IEEE Standard 2001-2002. IEEE Recommended Practice for the Internet - Web Site Engineering, Web Site Management, and Web Site Life Cycle*. Internet Best Practices Working Group, IEEE Computer Society.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The Unified Software Development Process*. Addison-Wesley.

Kamthan, P. (2005). Pair Modeling. *The 2005 Canadian University Software Engineering Conference (CUSEC 2005)*. Ottawa, Canada. January 14-16, 2005.

Kappel, G., Pröll, B., Reich, S., & Retschitzegger, W. (2006). *Web Engineering*. John Wiley and Sons, Inc.

Katira, N. (2004). *Understanding the Compatibility of Pair Programmers*. M.Sc. Thesis. North Carolina State University.

Keller, D. (1990). A Guide to Natural Naming. *ACM SIGPLAN Notices*, 25(5), 95-102.

Khan, A., & Balbo, S. (2005). Agile versus Heavyweight Web Development: An Australian Survey. The Eleventh Australian World Wide Web Conference (AusWeb 2005), Gold Coast, Australia, July 2-6, 2005.

Liu, L., Kong, X., & Chen, J. (2006). An Economic Model of Software Development Approaches. The Twelfth Australian World Wide Web Conference (AusWeb 2006), Australis Noosa Lakes, Australia, July 1-5, 2006.

Maurer, F., & Martel, S. (2002). Extreme Programming: Rapid Development for Web-Based Applications. *IEEE Internet Computing*, 6(1), 86-90.

Mnkandla, E., & Dwolatzky, B. (2004). A Survey of Agile Methodologies. *Transactions of the South African Institute of Electrical Engineers*, 95(4), 236-247.

Nguyen, H. Q., Johnson, R., & Hackett, M. (2003). *Testing Applications on the Web: Test Planning for Mobile and Internet-Based Systems* (2 Ed.). John Wiley and Sons, Inc.

Powell, T. A., Jones, D. L., & Cutts, D. C. (1998). *Web Site Engineering*. Prentice-Hall.

Schneidewind, N. F., & Fenton, N. E. (1996). Do Standards Improve Product Quality? *IEEE Software*, 13(1), 22-24.

Schwabe, D., & Rossi, G. (1998). *An Object Oriented Approach to Web-Based Application Design*. John Wiley and Sons, Inc.

Van Duyne, D. K., Landay, J., & Hong, J. I. (2003). *The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience*. Addison-Wesley.

Wallace, D., Raggett, I., & Aufgang, J. (2002). *Extreme Programming for Web Projects*. Addison-Wesley.

Wiegers, K. (2002). *Peer Reviews in Software: A Practical Guide*. Addison-Wesley.

Williams, L., & Kessler, R. (2003). *Pair Programming Illuminated*. Addison-Wesley.

KEY TERMS

Agile Development: A philosophy that embraces uncertainty, encourages team communication, values customer

satisfaction, vies for early delivery, and promotes sustainable development.

Coding Standard: A documented agreement that addresses the use of a formal (such as markup or programming) language.

Pair Programming: A practice that involves two people such that one person (the primary person or the pilot) works on the artifact while the other (the secondary person or the co-pilot) provides support in decision making and provides input and critical feedback on all aspects of the artifact as it evolves.

Refactoring: A structural transformation that provides a systematic way of eradicating the undesirable(s) from an artifact while preserving its behavioral semantics.

Semantic Web: An extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning.

Web Application: A Web site specific to a domain that behaves more like an interactive software system and will, in general, require programmatic ability on the server-side and may integrate/deploy additional software (such as application servers, media servers, or database servers) for some purpose (such as dynamic delivery of resources).

Web Engineering: A discipline concerned with the establishment and use of sound scientific, engineering, and management principles and disciplined and systematic approaches to the successful development, deployment, and maintenance of high quality Web Applications.

Category: IT Education

Facilitating Roles an E-Instructor Undertakes

Ni Chang
Indiana University South Bend, USA

INTRODUCTION

A discussion of roles that an instructor plays in the traditional classroom does not seem to be an innovative focus in the educational field. Yet, such discussions continue because of the topic's paramount impact on student learning. Discussions regarding the roles that an online instructor plays in a virtual learning environment are essential because teaching and learning via course management systems are completely different from that in the face-to-face setting and are still in their infancy, thereby requiring a great deal of exploration.

BACKGROUND

Roles of an Online Instructor: The Paradigm Shift

Traditional face-to-face meetings differ distinctively from online teaching (Coppola, Hiltz, & Rotter, 2001; Lim & Cheah, 2003) largely due to the following reasons. The former relies heavily on a specific location and time, whereas the latter is independent of time and location. The former mostly constitutes speaking and listening, while the latter is exercised primarily by reading and writing. The former makes an instructor and learners easily visible to one another, while the latter leaves the instructor and all learners in individual and invisible locations (Pelz, 2004; Sloan Consortium, 2006). The former expects learners to have a moderate level of self-regulation, whereas the latter requires learners to have a higher level of self-regulation (Pelz, 2004; Sloan Consortium, 2006). All the changes from the familiar to the unfamiliar explicitly generate a sizable barrier for online teaching and learning. Removing barriers to student success necessitates the online instructor to undertake a variety of responsibilities (Lim & Cheah, 2003; Morris, Xu, & Finnegan, 2005). Discovering what roles an instructor ought to play in a virtual learning environment is conducive to and vital in the successful facilitation of student learning.

MAIN FOCUS: ROLES UNDERTAKEN BY AN E-INSTRUCTOR

In the following text, the roles of an e-instructor are characterized horizontally by two categories on the basis of a study conducted by the author (Chang, 2007): Pedagogical Efficacy (8 roles) and Affective Promotion (19 roles). These roles are also set apart vertically by three distinct stages: Course Development (7 roles), Course Delivery (18 roles), and Course Completion (2 roles). To address these roles one by one, the author will present them in the form of stages, namely, Course Development, Course Delivery, and Course Completion. The first two stages contain both categories of Pedagogical Efficacy and Affective Promotion with the last stage showing only Pedagogical Efficacy.

COURSE DEVELOPMENT

During the Course Development stage, under the category of Pedagogical Efficacy, the instructor assumes four roles from those of gaining technological skills to those of getting the course ready for teaching. The instructor is responsible for acquiring necessary and useful technological skills and becomes familiar with the learned skills through practice. In the same stage, the instructor engages in research to decide the content of a course plan, which is in line with what Wilson, Varnhagen, Krupa, Kasprzak, Hunting, and Taylor (2003) found from their interviews of eight e-instructors. The researchers noticed that the information covered in virtual learning environments was not equivalent in amount to that in face-to-face meetings. One of the interviewees in the Wilson et al. (2003) study noted, "I went from about 13 individual classes or modules to about six modules." Followed by the decision-making, the instructor lays out a course plan appropriate for the students' learning needs. The reduction in content should not only be measured in quantity, but also be in quality. The environment set for learning should be responsive to students' needs and their learning levels (Berge, 1995). Topics for discussions need to be meaningful and related to students' experiences and interests to attract and maintain students' learning and desire for an in-depth study of concepts and tasks (Lim & Cheah, 2003).

During the stage of Course Development, student affective learning should be seriously attended to. Three roles are involved in this teaching process. An e-course instructor

ought to keep in mind that designing an online course is by no means "curriculum conversion" (Palloff & Pratt, 1999), because this practice is insufficient to guide students in their acquisition of knowledge in a self-controlled manner (Chang, 2007). Learning through a virtual classroom seems intimidating to some and confusing to others. To minimize the degree of apprehension and anxiety, much work is necessary, such as an analysis of the learning environment (Tessmer, 1990) and offering details of requirements and expectations (Lim & Cheah, 2003). The purpose is to avoid the phenomenon that "an instructional design project may produce a theoretically sound but practically unable product" (Tessmer, 1990, p. 56). Additionally, if the same course is repeated in a following semester, redesigning/revising should be something that an online instructor needs to exercise based on the instructor's self-reflections and voices from students taught in the previous semester.

COURSE DELIVERY

It is through the stage of Course Delivery that interaction between students and the instructor transpires. During the interaction with students, an instructor, undertaking four roles, works as an academic guide instead of attempting, as much as possible, to remove the instructor from actual course delivery (Lim & Cheah, 2003; Morris et al., 2005). Over the course of this process, the instructor not only is a learner acquiring knowledge both in content areas and technological skills, but also needs to lecture students. Lecturing still is one of the appropriate instructional methods employed via the Internet or face-to-face meetings. Lectures would address both technology and content in order to prepare students for online learning and to avoid unnecessary confusion later in the learning process. During the lecture, the instructor should exercise caution as not to drive students further away by making a poor presentation, as some students are already intimidated by the idea of learning with computers. Lectures can also be viewed as the time when the instructor responds to students' emails and when he or she interacts with students about their assignments via course management systems. The way that the instructor offers assistance to a learner works as a form of scaffolding or individualized instruction.

Individualized instruction transpires as each communication between the instructor and a student is tailored to specific needs or misconceptions that a student expresses through emails or a submitted assignment. Lim and Cheah (2003) supported this notion with the analysis of a questionnaire and two focus-group interviews. The researchers found that feedback provoked students' thinking and enhanced reflection. In this sense, reviewing the student's submitted work should not be restricted by a traditional responsibility: granting a grade. Reading student submitted assignments works as an informative process that helps the instructor be aware of the status of the student's learning and the quality of the instructor's facilitation. In this sense, awarding a grade itself is not an end to the cycle of communication between the student and instructor. The circular communication continues if a student is willing to continue working on the assignment based on the instructor's feedback. This exercise corresponds to Pelz's (2004) argument that the instructor's thinking is stimulated by the student's work, with an emphasis on helping the student to reach a high level of reflective expression. Moreover, this assessment process provides opportunities for the instructor to improve the course design and teaching strategies. To an e-instructor, this practice is part of ongoing assessment. Ongoing assessment is essential in the process of teaching and learning, as it allows an instructor to continue to improve the work by constant reflections on students' posted work. It is a way to guide students on the side as an effort to keep online learning on the right track (Berge, 1995). Teaching is intellectual work; examining one's teaching and asking questions about ongoing teaching are part of it (Anthony, 1999). Analyzing the process of teaching in a detailed and organized way is scholarly (Shulman, 1993).

Although pedagogical efficacy is significant in student learning, student emotional involvement in learning is equally essential to the success of learning and should be taken into serious account. There are 14 roles regarding affective promotion during this stage, three times more roles than those in the Course Development stage. To ease the transition from the mode of speaking and listening to that of frequently reading and writing and to help students establish self-responsibility, self-time management, self-motivation, and a sense of autonomy, the instructor's appropriate and incremental support in these aspects is pivotal (ESRC Economic and Social Research Council, 2002). An e-instructor should be purposefully committed to employing various strategies to encourage learners to become owners of their own learning (Morris et al., 2005; Pelz, 2004) in hopes that students would actively participate in online discussions and activities. In learning with computers, some students jump right in, whereas others resist the situation. These students usually struggled, to a certain degree, due to the dramatic shift of the learning paradigm (Wilson et al., 2003). Monitoring students' learning can help students ease into the transition and boost their self-confidence in online learning when the instructor provides just-in-time needed support and assistance. Furthermore, to gradually help students transit from the familiar to the unfamiliar learning environment, the instructor frequently sent out email, reminding students of matters requiring their attention at an appropriate time. In the reciprocal interaction with students, it is fundamental for the instructor to understand that one student's question may be representative of others' and that emerging problems in the process of teaching and learning may become a potent opportunity for the instructor to reexamine the course design and instructional strategies. These can be the basis for the

instructor to clarify topics under discussion and to modify course content and ongoing approaches to teaching (Anthony, 1999; Chang, 2001b). These endeavors may lessen the learners' anxiety level to a certain degree.

To further set up an emotionally supportive learning environment, assessments of the effect of course design and delivery are continuous throughout each semester. The instructor purposefully organizes a course by evaluating it based on unexpected e-mail messages and communications with students. For example, in receiving email messages, the instructor may learn which assignments are most helpful to students in their acquisition of knowledge. Students may also consciously and unconsciously share what they like or dislike about an assignment. E-mail messages also are an avenue for students to reveal their thoughts and feelings concerning an ongoing course. In the meanwhile, to avoid commotion and chaos, an e-instructor ought to make known to the students that assistance is readily available to them via e-mail. Students are not left completely alone in a novel learning environment.

Although many researchers have paid much attention to an issue of course management (see Berge, 1995, Coppola et al., 2001; Morris et al., 2005), this effort is actually intended to motivate students and maintain their desire to learn. It is of primary importance that an instructor asks questions while engaging in instruction throughout a semester. Asking questions and searching for answers are actions that can lead to quality learning (Chang, 2001b). It is beneficial also for the instructor to take advantage of every available opportunity to solicit feelings and reactions with regard to online course delivery for supporting and strengthening student learning. In a geographically isolated learning setting, it would make e-students feel comfortable and encourage their conscious learning efforts if an instructor cares about their well-being and sincerely listens to their feelings and problems that they have about online learning. This is a way to support and sustain students' affective learning. To this end, the instructor must maintain an open mind to learn about and understand students in order to provide supportive assistance to their learning at a level compatible to their needs and interests.

Because online teaching and learning is still in its infancy, many unexpected events may occur during the Course Delivery stage. It requires an instructor to remain flexible and creative. For instance, even if an instructor introduces students to the "Comment" feature of Microsoft Word in corresponding feedback, some students still are unable to adopt this feature for various reasons. Some may lack the updated software, may have different platforms, or may feel ill at ease using the technical skill. Under these circumstances, the instructor's flexibility and creativity should come into play by developing other useful methods to benefit student learning. For instance, instead of using the Microsoft Word feature to submit their work as an attachment, an e-instructor, when reading student posted homework, may place a question in {{{{}}}}, for example, {{{{How does this statement relate to the criterion?}}}}} or a comment in [[[[[]]]]], for example, question may be made at the end of the triple dashes inside the created symbols, such as this {{{{How does this statement relate to the criterion? --- Since the teacher has detailed knowledge of the students, the curriculum is designed to suit their learning level. Therefore, I think that this statement supports the criterion.}}}}.

During the Course Delivery, the instructor functions as a traditional instructor. With the utilization of Announcement on a course management system and e-mail, students can be informed of class-related information as well as business unrelated to the course content. Due to different platforms or operation systems that the students and the instructor use, files that students send as attachments sometimes cannot be read. For troubleshooting the problem, the instructor can write the student an e-mail note, providing step-by-step instruction to save time that the student may otherwise have to spend on communicating with personnel at HelpDesk. This endeavor and immediate assistance also strengthen students' emotional involvement in learning, which, in turn, is beneficial to effective inquiry.

Motivation may become extrinsic if students are unduly influenced by words of praise from their teachers (Starko, Sparks-Langer, Pasch, Frankes, Gardner & Moody, 2003). However, according to Chang (2007), acknowledging students' submissions and effort allow the instructor to recognize students' work without excessive praise, to encourage their continuous endeavors, and to promote their intrinsic motivation. Before electronic feedback is provided, the first sentence is often, "Thank you for your submission." The sentence at the end of feedback usually could be, "Keep up with the good work!" and/or "Thank you for your time and work." If a revision is needed, the message may be, "Thank you for your time and cooperation." If a response to the instructor's question is prompt, the student receives a reply, such as "Thank you for your quick response. I appreciate your effort." Sometimes, the instructor recognizes students' hard work by highlighting, "Keep up your good work."

Patience plays an important role in online teaching. Sometimes, an explanation of an assignment needs to be repeated and/or is given in various ways. Sometimes, one concept may require several elaborations. If some students are unable to follow given guidelines, the instructor needs to provide directions for needed materials. If students fail to understand or misunderstand underscored points embedded in readings, they are asked to reread required class materials and/or offered guidance, if needed. If a student has difficulty locating a document online, the instructor helps provide the direction and alternative assistance. If a student fails to follow the direction, the instructor provides further detailed instruction and also may attach the article that the student intends to obtain.

COURSE COMPLETION

During the third stage of Course Completion, an instructor primarily assumes two roles under the category of Pedagogical Efficacy by focusing attention on decision making and reflection for the next semester's instructive preparation. An e-instructor needs to discern as well as to reflect upon the data collected through the entire semester in order to perform general analytical work. The collected data may include students' e-mail, feedback from students concerning their submissions, student surveys, and records of conversations between students and the instructor at different occasions and in different formats. From these resources, answers to the instructor's questions may be obtained that may further improve cyber instruction. Some of these questions may be: What has been achieved in the past semester? What has been viewed as a failure? What lessons should be drawn to improve future online instruction? Anthony (1999) agreed that making inquiries and seeking solutions is a reflective practice; "Teaching is reflective and informed" (p. 3). Targeted curriculum and a suitable learning condition chiefly stem from reflection and the use of pedagogical and content knowledge.

The results of reflective thinking, in turn, enable the instructor to make informed decisions for future course design and instructional strategies in the virtual learning environment. The process is cyclical (Chang & Petersen, 2006; Chang, 2001a).

FUTURE TRENDS

Redden (2005) underscored the significance of students' affective involvement in student learning: "We ignore emotions in the cyber classroom at our own risk. We may want to focus on learning outcomes, but we cannot ignore the process that facilitates or hampers those outcomes" (n.p). Hall (2002) further noted that affective learning had been recognized as more potent and influential than academic efforts in student learning. How to promote students' emotional involvement in online learning requires more research efforts. Online learning isolates learners and the instructor geographically, leaving the community members feeling lonely and, sometimes frustrated. As learners move from the familiar to the unfamiliar learning environment, they need emotional support and promotion in order to be successful learners. Redden (2005) noted the notion that emotions affect learning is indeed in line with brain research results. To build an emotional foundation for successful learning and to reduce the level of discomfort, Redden suggested that the instructor must be "visible" in order to render emotional support. Despite the fact that much of the literature has underlined the importance of establishing a friendly and healthy learning community and has recognized that students are able to lend helping hands to one another for successful learning (Pelz, 2004), socialization would be short-lived or a forced participation without an instructor's commitment to it. Therefore, future efforts should be exerted to explore and discover mechanisms to promote students' affective learning, namely, how and when an e-instructor enters into the interaction in order to heighten the intructor's visibility to students. Additional research may also focus on the promotion of students' affective involvement in learning through an e-instructor's appropriate communication compatible to each student's individual level of cognition revealed via the student's submitted homework. That is, what and how much an e-instructor needs to say so that students would be likely to read constructive feedback to better their learning. Research efforts may be extended to ways to encourage and strengthen learners' self-regulation and self-motivation and to the continuous discovery of roles that an e-instructor plays in a virtual learning classroom.

CONCLUSION

The chapter classifies the 27 roles into two distinctive categories: Pedagogical Efficacy and Affective Promotion. The instructor assumes eight roles in the category of Pedagogical Efficacy. These roles are intended to promote student learning, ranging from assisting students in gaining technological skills to teaching them academic content. This category also demonstrates that the instructor is responsible for acquiring technological skills and becomes familiar with newly learned skills through practice. Throughout the teaching and learning process, the instructor develops meticulous plans and designs an e-course appropriate for the students' learning needs. Reflection enables the instructor to reexamine the whole process of teaching and learning and to make useful informed decisions in order to develop an improved course in the future.

In comparison with Pedagogical Efficacy, there are 19 roles that the instructor assumes in the category of Affective Promotion—twice as many roles as in Pedagogical Efficacy. Ways to help promote affective learning include interactive communication about an issue or an assignment, encouragement, acknowledgement, technical assistance, troubleshooting, and other logistics.

These 27 roles identified through the aforementioned discussion are seemingly familiar to some educators. However, despite the superficial overlapping nature, the roles addressed here are not completely the same as those carried out in the traditional classroom. This is largely due to the paradigm shift that has entirely altered how teaching and learning is exercised. It is inappropriate for an online instructor to blindly adopt what the instructor has utilized in the traditional classroom to interact with students in a virtual learning environment.

On the whole, an e-instructor not only has to know content knowledge and pedagogical skills, but also has to be equipped with technological know-how. Apart from these, the instructor should possess a set of strategies in order to appropriately facilitate learning while maintaining and increasing learners' emotional involvement in a novel learning environment. The ideology of consistent efforts made to explore the responsibilities of an e-instructor is likely to be advantageous to students' successful learning.

REFERENCES

Anthony, A. C. (1999, November). Introductory remarks. In *Proceedings of the 6th Annual Preparing Future Faculty Conference: The Scholarship of Teaching*, Jacksonville, FL.

Berge, Z. L. (1995). Facilitating computer conferencing: Recommendations from the field. *Educational Technology, 35*(1), 22–30.

Chang, N. (2007). Responsibilities and accountabilities of an early childhood e-instructor. *Journal of Early Childhood Teacher Education, 28*(4), 315-331.

Chang, N. & Pertersen, N. J. (2006). Cybercoaching: An emerging model of personalized online assessment. In D. D. Williams, S. L. Howell, & M. Hricko (Eds.), *Online assessment, measurement, and evaluation: Emerging practices* (pp. 110–130). Hershey, PA: Idea Group.

Chang, N. (2001a). It is developmentally inappropriate to have young children work alone at the computer. *Information Technology in Childhood Education (ITCE) Annual.* VA: Association for the Advancement of Computers in Education (AACE).

Chang, N. (2001b). The role of reflective practice in teaching: An evolution from a lecture format to service and online learning. *Online Journal of Teaching Excellence,* 2. Retrieved June 17, 2008, from http://www.uwplatt.edu/~journal/

Coppola, W., Hilz, R., & Rotter, N. (2001, January). Becoming a virtual professor: Pedagogical roles and ALN. In *Proceedings of the 34the Hawaii International Conference on System Sciences*, Maui, Hawaii.

ESRC Economic and Social Research Council (2002). *ESRC research seminar series: Understanding the implications of networked learning for higher education*.

Retrieved June 17, 2008, from http://csalt.lancs.ac.uk/esrc/manifesto.pdf

Hall, R. (2002). Aligning learning, teaching and assessment using the web: An evaluation of pedagogic approaches. *British Journal of Educational Technology, 33*(2), 149–158.

Lim, C. P. & Cheah, P. T. (2003). The role of the tutor in asynchronous discussion boards: A case study of a preservice teacher course. *Education Media International.* Retrieved June 17, 2008, from www.tandf.co.uk/journals/routledge/09523987.html

Morris, L. V., Xu, H., & Finnegan, G. L. (2005). Roles of faculty in teaching asynchronous undergraduate courses. *Journal of Asynchronous Learning Networks, 9*(1), 65–82.

Palloff, P. M. & Pratt, K. (1999). *Building learning communities in cyberspace: Effective strategies for the online classroom.* San Francisco: Jossey-Bass.

Pelz, B. (2004). Three principles of effective online pedagory. *Journal of Asynchronous Learning Networks, 8*(3), 33–46.

Redden, C. A. (2005, October). *Emotions in the cyber classroom. Educator's voice – archive.* Retrieved June 17, 2008, from http://www.ecollege.com/news/EdVoice_arch_10_12_05.learn

Shulman, L. S. (1993). Teaching as community property: Putting an end to pedagogical solitude. *Change, 25*(6), 6–7.

Sloan Consortium (2006). *Growing by degrees: Online education in the United States, 2005*. Retrieved June 17, 2008, from http://www.sloan-c.org/publications/survey/survey05.asp

Starko, A. J., Sparks-Langer, G. M., Pasch, M., Frankes, L., Gardner, T.G., & Moody, C. D. (2003). *Teaching as decision making: Successful practices for the elementary teachers* (3rd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

Tessmer, M. (1990). *Environment* analysis: A neglected stage of instructional design. *Educational Technology Research and Development, 38*(1), 55–64.

Wilson, D., Varnhagen, S., Krupa, E., Kasprzak, S., Hunting, V., & Taylor, A. (2003). Instructors' adaptation to online graduate education in health promotion: A qualitative study. *Journal of Distance Education, 18*(2), 1–15.

KEY TERMS

Affective Promotion: encompasses endeavors and strategies made by an e-instructor in fostering students' emotional involvement in e-learning and in setting up an emotionally supportive learning environment to facilitate student learning.

Knowledge Building (I): Refers to an e-instructor, who keeps professionally up-to-date through self-development and learning alongside students and who attains technological knowledge and skills by attending relevant workshops and by frequent interaction with a computer.

Knowledge Building (S): Refers to the e-students' enhancement of content-bound knowledge and skills as well as their understanding of computer technology via various mechanisms that an e-instructor employs to guide students in an effort to achieve their mutual academic goals.

Instructive Preparation: Is related to avenues in which an e-instructor is engaged to make decisions based on information at hand as well as collected through previous experiences of working with students in order to help plan instruction suited to learners' needs.

Meaningful Management: Refers to an e-instructor who manages a course in ways that may help ease students' unnecessary frustration resulting from their being situated in a novel learning environment. This type of course management aims to promote students' affective learning in the virtual classroom.

Pedagogical Efficacy: Refers to the growth and development of both faculty and students concerning academics-oriented knowledge and skills ranging from content-specific areas to technological skills through efforts exerted by an e-instructor.

Purposeful Commitment: Refers to an e-instructor who is committed to helping students become owners of their own learning by the instructor becoming visible through various means in the shared virtual classroom in order to support learning.

Purposeful Organization: Refers to an e-instructor who is committed to helping students become owners of their own learning, achieved when the instructor becomes visible through various means in the virtual classroom.

Reflective Practice: Refers to an e-instructor's consistent behaviors in assessing the course by an ongoing, even daily, basis as well as at the end of a semester in order to motivate learners to succeed in learning.

Factors for Global Diffusion of the Internet

Ravi Nath
Creighton University, USA

Vasudeva N.R. Murthy
Creighton University, USA

INTRODUCTION

There is overwhelming evidence that the use of the Internet-enabled applications and solutions provide unprecedented economic growth opportunities. However, the Internet diffusion rates remain low in many countries. According to the International Telecommunications Union (ITU), in 2004, less than 3% of the Africans used the Internet, whereas the average Internet subscription rate for G8 countries (Canada, France, Germany, Italy, Japan, Russia, the UK, and the US) is about 50%. Also, in nearly 30 countries the Internet penetration rates still remain below 1% (ITU, 2006). So, what are the key factors that explain this wide variation in Internet subscription rates in countries around the world? An understanding of these factors will be highly useful for policy makers, economic developmental agencies and political leaders in establishing and implementing suitable national developmental strategies and policies.

BACKGROUND

The Internet is playing a pivotal role in the economic development of nations. The adoption of the Internet and related business applications such as e-business, voice over IP (VoIP), mobile commerce, and integrated supply chains have become the primary drivers of the growth of economic activities in many countries (Dedrick, Gurbaxani, & Kraemer, 2003; Kenny, 2003, Koh & Chong, 2002). For example, it is estimated that during the 1990's, investments in information and communication technologies contributed around 10 to 20 percent to the output growth of the economies of the countries such as Canada, Finland, and the United States (Lawrence, 2002). In fact , the recently published *Global Information Technology Report 2005-2006*, and Global *Competitiveness Report 2006-2007* attribute the enhanced degree of competitiveness of such high ranking economies as the United States, Singapore, Switzerland, and Nordic countries to their high levels of networked readiness and technological readiness. An important factor of technological readiness is the high level usage of the internet. *Global Technology Report 2006-2007* (p. 10) observes a positive correlation between the global competitiveness index for 2006-2007 and technological readiness index for 2005-2006 in a large number of countries emphasizing the key role played by the usage of information and communication technology (ICT). Accelerated economic growth rates in India and China are also prime example of how ICT in concert with appropriate economic, intellectual property protection, and infrastructure improvement policies promote rapid economic development.

GLOBAL INTERNET DIFFUSION FACTORS

Several factors have been observed to determine the penetration rates of the internet in various countries. These factors include the availability of reasonably-priced telecommunications infrastructure, access to personal computers, educational and training opportunities for individuals, income levels, and innovative capability of the country (Beilock & Dimitrova, 2003; Chinn & Fairlie, 2007; Dewan, Ganley, & Kraemer, 2006; Dholakia, Dholakia, & Kshetri, 2003; Kiiski & Pohjola, 2002; Oyelaran-Oyeyinka & Lal, 2005; Meijers, 2006; Murthy, 2004; Nath & Murthy, 2003, 2004). Also, the rule of law (e.g., property rights, strong legal system) governing the country's trading system, government regulations and market liberalization policies, and credible payment systems (e.g., credit cards, digital wallet, and cash) are necessary for migrating to digital commerce.

Human Capital Development Factors

It is difficult to realize the full potential of the Internet if people cannot read or write or have a basic understanding of the computer and its functions. Therefore, an individual needs a minimum level of computer and language literacy to use the Internet and accrue its benefits. One way to assess this is to consider measures such as:

a. Adult literacy rate
b. Percent of school age children enrolled in schools
c. Per capita spending on education

Many studies including those by Baliamoune-Lutz (2003) and Nath and Murthy (2003) have concluded that literacy rates and tertiary enrollment are strong predictors of Internet diffusion. Also, literate population is more accepting of

information and communication technology (ICT) innovations which subsequently lead to increased acceptance of advanced technologies such as the Internet.

Technological Factors

Availability and reliability of the telecommunications infrastructure is clearly important for people to use the Internet. This includes the bandwidth, the number of Internet hosts, the reliability of electric power, and the percent of the population of a nation that have access the Internet. In addition, people who are familiar and comfortable with using other technologies such as a phone, a mobile phone and a personal computer are more likely to adopt and use the Internet. Specific variables that one may consider are:

a. Number of personal computers per 100 inhabitants.
b. Telephone lines per 100 inhabitants.
c. Cell phone subscribers per 100 inhabitants
d. Number bandwidth (bits) per capita
e. Reliability of electrical power

One key aspect of the new economy is the availability of cost-effective information and communication technologies (ICT) and the above listed items represent leading indicators of ICT (Baliamoune-Lutz, 2003). Several studies have established that countries with higher penetration rates of personal computers, telephones, and mobile phones, sufficient bandwidth and reliable supply of electricity that powers most ICT devices, also tend to have more Internet users (Nath & Murthy, 2003, 2004).

Economic Factors

How expensive is it to get an Internet connection? If the cost is prohibitively high relative to the income, then the Internet penetration rate is likely to be low. The two variables that are relevant here are:

a. Real gross domestic product per capita (in US purchasing parity $)
b. Average monthly cost of 20 hours of Internet access.

Kiiski and Pohjola (2002) found that, across countries, GDP per capita and Internet access costs were the best predictors of growth in Internet hosts. Further, more recently Nath and Murthy (2004) using data from 62 countries, demonstrated that higher "average monthly cost of 20 hours of Internet access" had a significant negative impact on the Internet diffusion rates.

Political Factors

A nation's political and economic policies and environment are important in determining the diffusion rate of the Internet. Some nations restrict the use of the Internet (e.g., Iran and North Korea). Also, the economic policies of the government affect the extent to which citizens use the Internet. Variables that measure "political and economic" policies include:

a. **Economic Freedom Index (EFI):** Beach and O'Driscoll (2003) define this index as the "…. absence of government coercion or constraint on the production, distribution, or consumption of goods and services beyond the extent necessary for citizens to protect and maintain liberty itself." This index aggregates several factors covering broad issues such as corruption, non-tariff barriers to trade, the fiscal burden of government, the rule of law and efficiency of the judiciary, regulatory hurdles for businesses, labor market restriction, and black market activities. Complete details regarding the development and description of this index can be found in Beach and O'Driscoll (2003). The values of EFI can vary from 1 to 5. A value of 1 indicates set of national policies that promote economic freedom and a value of 5 signifies policies that are least conducive to economic freedom.
b. **Innovation Capability of the Country:** This variable is calculated as the product of the number of patents granted per million inhabitants and gross tertiary enrollment rate. Note that the number of patents reflects the nation's innovation intensity and the enrollment rates denote the degree of investment in human capital. Thus, this measure reflects a country's capability, ability to provide a conducive environment for augmenting technological advance and its capacity for innovation in technologies and products (McArthur & Sachs, 2000).

In a study of the diffusion of the Internet across countries, Nath and Murthy (2003) have shown that both the economic freedom index and the innovation capability of a country play a positive role towards the diffusion of the Internet. These findings have been further supported by Baliamoune-Lutz (2003) and Nath and Murthy (2004).

Cultural Factors

Cultural factors do impact the rate of diffusion of the internet. Culture is defined as: "the collective programming of the mind which distinguishes the members of one human group from another (Hofstede, 1991, p.5)." In his book titled *Culture's Consequences,* Hofstede (1980) suggested four dimensions of culture.

a. **Power Distance (PD):** Power distance is defined as "the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally" (Hofstede, 1991, p.27). In cultures with high power distance, decisions are centralized and subordinates are often fearful of disagreeing with their superiors. On the other hand, cultures with low power distance are more participative and have less tolerance for the lack of autonomy. In a society with low power distance, we expect people to be more innovative and willing to try new things.
b. **Individualism versus Collectivism (IND):** This dimension relates to the way people live together. Individualism "pertains to societies in which the ties between individuals are loose: everyone is expected to look after himself or herself and his or her immediate family" (Hofstede, 1991, p.51). On the other hand collectivism "pertains to societies in which people from birth onwards are integrated into strong, cohesive groups, which throughout people's lifetime continue to protect them in exchange for unquestioning loyalty" (Hofstede, 1991). Cultures with high individualism value personal time and achievement. Such societies are expected to be more innovative and open to new ideas.
c. **Uncertainty Avoidance (UA):** Uncertainty avoidance refers to "the extent to which the members of a culture feel threatened by uncertain or unknown situations" (Hofstede, 1991, p. 113). A culture high in uncertainty avoidance is rule oriented, has less tolerance for opinions and behaviors different from its own, and avoids taking risks. There is also resistance to change. Cultures with high uncertainty avoidance are expected to be less innovative and less accepting of new things.
d. **Masculinity versus Femininity (MAS):** This is possibly the most controversial dimension of culture advocated by Hofstede. In highly masculine cultures "men are supposed to be assertive, tough and focused on material success; women are supposed to be more modest, tender and concerned with the quality of life" (Hofstede et al., 1998). The more modern and popular perspective on this dimension is to view the masculine and feminine culture in terms of competitiveness and material success versus nurturing behavior and quality of life, as opposed to gender roles for the sexes.

There is considerable evidence supporting that cultural factors influence the use of various technological innovations. Yaveroglu and Donthu (2002) have shown that the use of cell phones, home computers, and microwave ovens is high in countries with low power distance, low uncertainty avoidance, and high individualism. Yeniyurt and Townsend (2003) investigated the role that culture plays in the acceptance of new products. Their findings indicate that lower acceptance of new products is related to power distance and uncertainty avoidance. With respect to the adoption of enterprise resource planning (ERP) software, national culture is shown to have a significant influence on its adoption rate (van Everdingen & Waarts, 2003). Cultural considerations even play a role in how information systems, in general, are designed, implemented, and used (Gallupe & Tan, 1999; Jarvenpaa & Leidner, 1998; Montealegre, 1997; Nelson & Clark, 1994; Straub, 1994; Watson, Ho, & Raman, 1994). For instance, Straub (1994) showed that in Japan, workers prefer fax over e-mail because of the intricacies of the Japanese language and other cultural factors. With respect to the role of culture on Internet diffusion, Nath and Murthy (2004) have shown that "uncertainty avoidance" and "masculinity" are positively associated with the diffusion of the Internet.

FUTURE TRENDS

The outlook for the Internet and its usage looks very promising. First, the broadband access to the Internet will increase rapidly allowing people very high-speed access to the Internet. Consequently it will be possible for the users to download rich content with little delay. New technological innovations such as WiMax will also accelerate the diffusion of broadband Internet services in a cost effective manner. Second, Internet access costs will decline due to global competition and economic reforms taking places in many countries around the world. Third, as many nations become receptive to economic reforms and provide additional economic freedom opportunities to its citizens, entrepreneurship will thrive, incomes will rise, information economies will increase and more people will use the Internet and become fully vested in the information economy. Overall, the existing digital divide will narrow considerably. In addition, mobile phone access to the Internet will yield unprecedented economic benefits to citizens at the "bottom of the pyramid" around the world.

CONCLUSION

A clear and vivid understanding of the factors that determine the Internet usage is crucial for researchers, practitioners, and policy makers around the world. Researchers who wish to study international information technology issues and their economic implications must be cognizant of these factors in assessing computer user behaviors and technology adoption patterns. Also, an understanding of inter-national differences with respect to these factors can improve IT/IS research design strategies and enhance the implementation of research studies in different national settings.

At the policy level, national policy makers and political leadership can benefit by incorporating the insights gleaned from this research into their deliberations resulting in more informed macro-level strategic policies. Without a cohesive and carefully crafted national development strategies and policies, nations are likely to miss out on the benefits of the Internet and thus leave a large portion of the population way behind on the economic ladder.

REFERENCES

Baliamoune-Lutz, M. (2003). An analysis of the determinants and effects of ICT diffusion in developing countries. *Information Technology for Development, 10*, 151-169

Beach, W. W., & O'Driscoll, G. P. (2003). *Explaining the factors of the index of economic freedom.* Retrieved from www.heritage.org

Beilock, R., & Dimitrova, D. V. (2003). An exploratory model of inter-country Internet diffusion. *Telecommunications Policy, 27,* 237-252.

Chinn, M. D., & Fairlie, R. W. (2007).The determinants of the global digital divide: A cross-country analysis of computer and internet penetration. *Oxford Economic Papers, 59*(1), 16-44.

Dedrick, J., Gurbaxani, V., & Kraemer, K. L. (2003). Information technology and economic performance: A critical review of the empirical evidence. *ACM Computing Surveys, 35*(1), 1-29.

Dewan, S., Ganley, D., & Kraemer, K. L. (2006). Across the digital divide: A cross-country multi-technology analysis of the determinants of IT penetration. *Journal of the Association for Information Systems, 6*(12), 409-424.

Dholakia, N., Dholakia R. R., & Kshetri, N. (2003). *Internet diffusion. The Internet encyclopedia.* New York: Wiley.

Gallupe, R. B., & Tan, F. (1999). A research manifesto for global information management. *Journal of Global Information Management, 7*(3), 5-18.

Hofstede, G. H. (1980). *Culture's consequences: International differences in work-related values.* Beverly Hills: Sage Publications.

Hofstede, G. H. (1991). *Cultures and organizations: Software of the mind*. New York: McGraw-Hill.

Hofstede, G. H. et al. (1998). Masculinity and femininity: The taboo dimension of national cultures. In W. Lonner, & J. Berry (Eds.), *Cross-cultural psychology series*. Newbury Park, CA: Sage Publications.

International Telecommunication Union. (2006). Retrieved November 30, 2006, from www.itu.int/itu-d/ict/statistics/ict/index.html

Jarvenpaa, S., & Leidner, D. (1998). An information company in Mexico: Extending the resource-based view of the firm to a developing country context. *Information Systems Research, 9*(4), 342-361.

Kenny, C. (2003). The Internet and economic growth in less-developed countries: A case of managing expectations. *Oxford Development Studies, 21*(1), 99–114.

Kiiski, S., & Pohjola, M. (2002). Cross-country diffusion of the Internet. *Information Economics and Policy, 14*, 297-310.

Koh, C. E., & Chong, H. (2002). Does the Internet improve business? An empirical inquiry into the perceived strategic value and contribution of the Internet. *Journal of International Technology & Information Management, 11*(1), 81-97.

Lawrence, S. (2002, February). Technology and the global economy. *Red Herring*, 28-29.

McArthur, J. W., & Sachs, J. D. (2000). The growth of competitiveness index: Measuring technological advancement and the stages of development. In *The Global Competitiveness Report 2000*. New York: Oxford University Press.

Meijers, H. (2006). Diffusion of the Internet and low inflation in the information economy. *Information Economics and Policy, 18*(1), 1-23.

Montealegre, R. (1997). The interplay of information technology and the social milieu. *Information Technology and People, 10*(2), 106-131.

Murthy, N. R. (2004). Internet diffusion: An econometric analysis. *Asian-African Journal of Economics and Econometrics, 4*(1), 45-54.

Nath, R., & Murthy, N. R. (2003). An examination of the relationship between digital divide and economic freedom: An international perspective. *Journal of International Technology & Information Management, 12*(1), 15-23.

Nath, R. & Murthy, N. R. (2004). A study of the relationship between Internet diffusion and culture. *Journal of International Technology & Information Management, 13*(2), 123-132.

Nelson, K., & Clark, T. (1994). Cross-cultural issues in information systems research: A research program. *Journal of Global Information Management, 2*(4), 19-29.

Oyelaran-Oyeyinka, B., & Lal, K. (2005). Internet diffusion in sub-Saharan Africa: A cross-country analysis. *Telecommunications Policy, 29*(7), 507-527.

Straub, D. (1994). The effect of culture on IT diffusion: E-mail and FAX in Japan and the U.S. *Information Systems Research, 5*(1), 23-47.

Van Everdingen, Y. M., & Waarts, E. (2003). The effect of national culture on the adoption of innovations. *Marketing Letters, 14*(3), 217-232.

Watson, R. T., Ho, T. H., & Raman, K. S. (1994). Culture: A fourth dimension of group support systems. *Communications of the ACM, 37*(10), 44-55.

World Economic Forum. (2006). *Global Competitiveness Report 2006-2007*. Palgrave-Macmillan.

World Economic Forum. (2006). *Global Information Technology Report 2005-2006*.Palgrave-Macmillan.

Yaveroglu, I.S. & Donthu, N. (2002). Cultural influences on the diffusion of new products. *Journal of International Consumer Marketing, 14*(4), 49-63.

Yeniyurt, S., & Townsend, J. D. (2003). Does culture explain acceptance of new products in a country? *International Marketing Review, 20*(4), 377-396.

KEY TERMS

Economic Freedom Index: A measure of the absence of government constraints on the economic activities of a nation.

Hofstede's Cultural Dimensions: Four factors proposed by Hofstede along which cultures might differ.

Internet Diffusion Rate: Percent of population of a nation having access to the Internet.

Power Distance (PD): A measure of the unequal distribution of power in a society. Cultures with high power distance are centralized whereas cultures with low power distance are more participative and individuals like more autonomy.

Uncertainty Avoidance (UA): A measure of how well a culture accepts uncertainty. A culture with high uncertainty avoidance is mostly rule oriented and is less tolerant of divergent opinions and behaviors.

Faculty Competencies and Incentives for Teaching in E-Learning Environments

Kim E. Dooley
Texas A&M University, USA

Theresa Pesl Murphrey
Texas A&M University, USA

James R. Lindner
Texas A&M University, USA

Timothy H. Murphy
Texas A&M University, USA

INTRODUCTION

In 2001, Michele Bunn offered her readers *timeless* and *timely* issues in distance education. There were predictions that virtual universities would shift from a teacher-centered to a student-centered learning environment. Although emerging technologies have allowed for more student-centered approaches, university instructors remain a key factor in the success or failure of distance education efforts in university settings. Shifts in technological advances over the past five years have made the term "distance education" less accurate. Learners are not necessarily located away from campuses, but instead choose the flexibility to learn asynchronously. Therefore, we will use the term e-learning, rather than distance education.

Timeless issues impacting e-learning typically include the core values of universities in regard to strategic planning, faculty competence, and incentives. Administrative decision-making determines relevant *timely* issues that tend to fall into three categories: (1) student-related issues, (2) instructional issues, and (3) organizational issues (Bunn, 2001). These three areas will frame our discussion.

BACKGROUND

The research on distance education faculty participation in the 1990s focused on the need to provide appropriate technological infrastructure, faculty training, and support for course development efforts. Enhancing faculty participation required that resources be directed to adequate levels of support and training so that instructional technologies were used for the benefit of students (Howard, Schenk, & Discenza, 2004).

While faculty recognized the potential of technology-enhanced instruction, intervention strategies were necessary to alter how people perceived and reacted to distance education technologies. Incentives such as release time, mini-grants, continuing education stipends, and recognition in the promotion and tenure process, were recommended to encourage faculty participation (Dooley & Murphrey, 2000; Murphrey & Dooley, 2000).

Today, technological infrastructure at universities is often in place. Many universities now include wireless technology and require students to come to the university armed with their own laptops and technological skills. Similarly, many faculty enter the profession equipped with technological competence. The need for training and support, while still important to some, has become replaced by the need to locate the appropriate resource personnel to create advanced multimedia. University instructors continue to want incentives to participate in e-learning due to the time and effort required to participate effectively. Most importantly, time and effort spent on e-learning activities is time diverted from scholarly work in research and the development of new knowledge. In addition, many faculty often lack the mechanics of how online teaching differs from lecture in terms of instructional material development, communication channels, interactions, and more authentic forms of assessment.

STUDENT-RELATED ISSUES

According to Bunn (2001), timeless student-related issues for e-learning incorporate the "policies and practices that define how students are treated in the administrative system. These timeless issues include the basic approaches to recruitment, enrollment, retention, and graduation" (p. 58).

Students are viewed as clients or customers with choices for their courses and programs. The view of a student as a customer poses a timely issue in regard to student-centered instructional design and delivery.

E-learning draws on the ability of learners to be self-directed, thus, incorporating adult learning principles (andragogy) in the design and delivery of content (Richards, Dooley, & Lindner, 2004). Andragogy is based on the following six assumptions about the learner: (1) learner's need to know; (2) self-concept of the learner; (3) prior experience of the learner; (4) readiness to learn; (5) orientation to learning; and (6) motivation to learn (Knowles, Holton, & Swanson, 1998).

Educators, who put their interests and needs (intentional or unintentional) over those of the learners, restrict meaningful learning. The ultimate goal of an educator should be to facilitate learning (Leamnson, 1999). This will require the educator to be a teacher, coach, mentor, facilitator, motivator, and/or authoritarian depending on the learners' personal characteristics.

Developing learning activities that require learners to draw on and share their prior experiences facilitates deeper and more meaningful learning. Noted philosopher, educator, and author, John Dewey (1938) stated that the education process begins with experience and that "all genuine education come about through experience" (p. 13). For example, learner-led threaded discussion groups can be used to help learners think about how course materials can be used in various contexts.

It is necessary for the educator to maintain a sense of community regardless of where the learning takes place. While this is readily accomplished in a classroom setting, it requires more planning and effort for e-learning courses (Brown, 2001). Grow (1991) theorized that in asynchronously delivered courses, an educator's traditional role of providing feedback is less important than the role of motivator, coach, or delegator, implying that the instructor must establish a learning climate. Effective learning seems to require student engagement (Kearsley & Shneiderman, 1999). However, instructor behaviors alone cannot determine student success rate. Success is at least partially controlled by student behavior. Previous research has shown, for example, that length of engagement in an asynchronously delivered course was positively related to a student's perception of learning (Lindner, Hynes, Murphy, Dooley, & Buford, 2003).

Learners can be successful in a variety of settings, and few, if any, differences in performance will be identified based on delivery method. While the delivery method may not impact success, delivery strategies do. Delivery strategies can be defined as those methods used to engage students in the instructional materials. It can be argued that through various interactions, engagement results in learning—not grades (Merrow, 2003). Moore (1989) developed a model for describing interactions used to engage learners in a distance education environment. His model included three types of interaction: (1) learner to learner; (2) learner to content; and (3) learner to instructor. Hillman, Willis, and Gunawardena (1994) expanded this model to include an additional type of interaction: learner to technology (interface). Zhang and Fulford (1994) expanded the model yet again to include vicarious interaction. Vicarious interaction captures the value to a learner of the interactions between others in the learning environment—the learning benefits of watching, reading, (or listening to) others interact. Discussion boards, for instance, are said to encourage vicarious interaction. Vicarious interaction contributes to overall perceived interaction, learner satisfaction, and quite possibly learning (Kawachi, 2003; Swan, 2004).

Figure 1. Vicarious interaction (Based on Hillman, Willis, & Gunawardena, 1994; Moore, 1989)

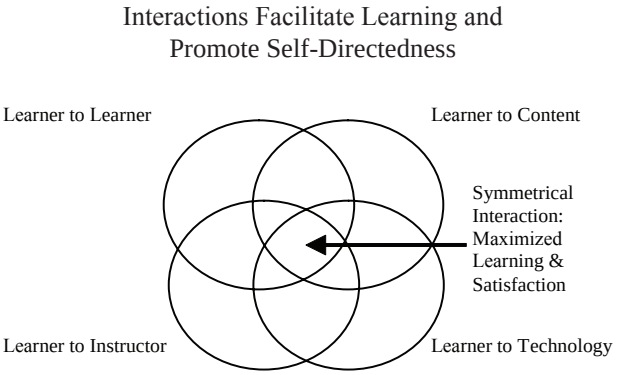


Figure 1 illustrates the interconnectedness among the types of interactions. To maximize learning and increase satisfaction in e-learning environments, opportunities for learner to learner, learner to content, learner to instructor, and learner to technology interactions should be included, and vicarious interaction should be encouraged. The authors believe that students will self-select the amount of each type of interaction necessary to maximize their learning and satisfaction. This individually balanced or "symmetrical" selection of interaction opportunities should facilitate deeper and more meaningful learning, increase overall learning satisfaction, and promote self-directedness among learners. The challenge for faculty members is to remain up-to-date with the ever-increasing range of technologies developed to support the various types of interaction. Acquiring the knowledge needed for the systematic selection of these tools into an instructional design, as well as the skills needed for their day-to-day management in instructional settings, presents tremendous challenges for faculty members.

INSTRUCTIONAL ISSUES

The focus of this article is on university faculty competencies and incentives for participation in e-learning. The student-related issues directly impact the philosophical beliefs and values of the faculty as they create instructional content and develop delivery strategies. If the learner is the customer, then timeless instructional issues will focus on the creation and delivery of courses for the learner (Bunn, 2001). Learner-centered instructional methods based on emerging technologies require new perspectives on developing materials, delivering courses, and creating faculty incentives for innovation and experimentation (Bunn, 2001).

Although university faculty have general technical competence, they usually do not have experience in the design and delivery of online courses and programs (andragogy, instructional design, etc.). Academic institutions are generally concerned that faculty are both *willing* to create content for online delivery and *able* to create effective instructional materials (Furnell, Evans, & Bailey, 2001).

As we reflect on instructional issues related to the delivery of courses online or at a distance, it is important to recognize the competencies required of faculty to teach in these environments. Murphrey and Dooley (2006) reported seven core areas as necessary for working in the e-learning arena: (1) proficiency with computers and programs, including interface design, (2) organizational skills, (3) instructional design, (4) evaluation and assessment strategies, (5) adult learning theory, (6) written communication skills, and (7) student/teacher relationships to build a sense of community. Each of these areas is critical in creating engaging and effective instruction for delivery online. In many ways these competencies are not very different from those required for traditional face-to-face instruction. The differences lie in the way in which each area can be accomplished and the magnitude by which the area impacts overall effectiveness. For example, organizational skills are important in both traditional and e-learning settings. However, when a faculty member lacks this skill and is operating in an e-learning setting, the overall negative impact can be much more due to the nature of the environment.

Magjuka, Shi, and Bonk (2005) share 10 critical design and administrative issues related to online education. These issues include which student group to serve, how the online program will be treated relative to other programs, the extent of blended versus stand-alone delivery, faculty load assignments, use of funds for course development versus faculty training, source of interactivity, course design versus faculty led interactions, content management system selection, and corporate partnerships/alliances. Faculty competencies and incentives will impact how an institution addresses each of these areas. Faculty must be competent in the delivery of course material in an online environment while at the same time recognize and embrace the incentives offered by administration to engage in these activities.

ORGANIZATIONAL ISSUES

Timeless organizational issues are directly related to student and instructional issues. They are inseparable from "technology, infrastructure, human resources, and administration" (Bunn, 2001, p. 62). Timely issues related to the organizational system include a reorganization of thinking and operation, seeking "cost-efficient ways to 'deliver' knowledge" (Bunn, 2001, p. 65).

In addition to concerns about faculty e-learning competence, many institutions struggle with providing appropriate incentives for online course development. Many institutions encourage faculty participation by paying additional stipends or salary for the development and delivery of online instruction. This is sometimes distributed in the form of mini-grants or additional summer salary for faculty not on 12-month contracts. In some cases faculty members receive "overload" funding for teaching additional sections in addition to their faculty teaching load. Additionally, many institutions have "distance education" affiliated with continuing education and pay faculty a fee per student for delivery of online instruction. These entrepreneurial models work well to reward faculty for their time and expertise.

FUTURE TRENDS

Instructors who have e-learning competence, well-developed instructional materials, and sought-after content are quite marketable. There is a trend for faculty to teach for other institutions using a distributed virtual faculty model. The University of Phoenix has been quite successful in the recruitment of distributed faculty to teach their online courses. This is cost effective for the institution and allows faculty to earn extra dollars, much like with consulting services. Furnell et al. (2001, p. 285) purport that "… [A]cademics could take encouragement from the observation that, in a virtual market, they (as subject experts) are no longer geographically constrained to providing their services to a single [institution], and could well find wider employment in the ODL [online distance learning] domain.

CONCLUSIONS

It must be recognized that teaching online requires a unique set of competencies (Richards et al., 2004) and attention to faculty incentives (Furnell et al., 2001). Bunn (2001) emphasized that timeless issues involve the fundamental core

values of the institution. Faculty participation in distance education, e-learning, or whatever the name of the future may be must consider both the willingness (incentives) and ability (e-learning competencies) of the faculty to be successful. Fundamentally, attention to vicarious interaction and learner-centered approaches is key. Faculty participation is therefore both a timeless and timely issue impacting the design and delivery of instruction for e-learning environments.

REFERENCES

American Psychological Association (APA). (1997). *Learner-centered psychological principles: A framework for school redesign and reform.* Retrieved October 9, 2003, from http://www.apa.org/ed/lcp.html

Brown, R. E. (2001). The process of community building in distance learning classes. *Journal of Asynchronous Learning Environments, 5*(2). Retrieved August 9, 2002, from http://www.aln.org/alnweb/journal/vol5_issue2/brown/brown.htm

Bunn, M. D. (2001). Timeless and timely issues in distance education planning. *The American Journal of Distance Education, 15*(1), 55-68.

Dewey, J. (1938). *Experience and education.* New York: Collier Books.

Dooley, K. E., & Murphrey, T. P. (2000). How the perspectives of administrators, faculty, and support units impact the rate of distance education adoption. *The Journal of Distance Learning Administration*, *3*(4). Retrieved November 2, 2006, from http://www.westga.edu/~distance/ojdIa/winter34/dooley34.html

Furnell, S., Evans, M., & Bailey, P. (2001). The promise of online distance learning: Addressing academic and institutional concerns. *The Quarterly Review of Distance Education, 1*(4), 281-291.

Grow, G. O. (1991). Teaching learners to be self-directed. *Adult Education Quarterly, 41*(3), 125-149. Retrieved November 2, 2006, from http://www.longleaf.net/ggrow/SSDL/SSDLIndex.html

Henry, P. (2001). E-learning technology, content and services. *Education & Training, 43*(4/5), 249-255.

Hillman, D. C., Willis, D. J., & Gunawardena, C. N. (1994). Learner-interface interaction in distance education: An extension of contemporary models and strategies for practitioners. *The American Journal of Distance Education, 8*(2), 30-42.

Howard, C., Schenk, K., & Discenza, R. (2004). *Distance learning and university effectiveness: Changing educational paradigms for online learning.* Hershey, PA: Information Science Publishing.

Jones, E. T., Lindner, J. R., Murphy, T. H., & Dooley, K. E. (2002). Faculty philosophical position toward distance education: Competency, value, and educational technology support. *The Journal of Distance Learning Administration*, *5*(1). Retrieved November 2, 2006, from http://www.westga.edu/~distance/ojdla/spring51/jones51.html

Kawachi, P. (2003). Vicarious interaction and the achieved quality of learning. *International Journal on E-Learning, 2*(4), 39-45.

Kearsley, G., & Shneiderman, B. (1999). Engagement theory: A framework for technology-based teaching and learning. Retrieved August 8, 2002, from http:://homesprynet.com/~gkearsley/engage.htm

Knowles, M. S., Holton, E. F., III., & Swanson, R. A. (1998). *The adult learner: The definitive classic in adult education and human resource development*. Woburn, MA: Butterworth-Heinemann.

Leamnson, R. (1999). *Thinking about teaching and learning: Developing habits of learning with first year college and university students.* Sterling, VA: Stylus.

Lindner, J. R., Hynes, J. W., Murphy, T. H., Dooley, K. E., & Buford, J. A., Jr. (2003). A comparison of on-campus and distance student's progress through an asynchronously delivered Web-based course. *Journal of Southern Agricultural Education Research, 53*(1), 80-92. Retrieved February 9, 2004, from, http://pubs.aged.tamu.edu/jsaer/pdf/vol53/jsaer-53-080.pdf

Magjuka, R. J., Shi, M., & Bonk, C. J. (2005). Critical design and administrative issues in online education. *Online Journal of Distance Learning Administration, 8*(4). Retrieved November 2, 2006, from http://www.westga.edu/~distance/ojdla/winter84/magjuka84.htm

Merrow, J. (2003, February 5). Easy grading makes 'deep learning' more important. *USA Today* (pp. 12A).

Moore, M. G. (1989). Three types of interaction. *The American Journal of Distance Education, 3*(2), 1-7.

Murphrey, T. P., & Dooley, K. E. (2000). Perceived strengths, weaknesses, opportunities, and threats impacting the diffusion of distance education technologies for colleges of agriculture in land grant institutions. *Journal of Agricultural Education, 41*(4), 39-50.

Murphrey, T. P., & Dooley, K. E. (2006). Determining e-learning competencies using Centra™ to collect focus group data. *The Quarterly Review of Distance Education, 7*(1), 78-82.

Richards, L. J., Dooley, K. E., & Lindner, J. R. (2004). Online course design principles. In C. Howard, K. Schenk, & R. Discenza (Eds.), *Distance learning and university effectiveness: Changing education paradigms for online learnging* (pp. 99-118). Hershey, PA: Information Science Publishing.

Swan, K. (2004). Relationships between interactions and learning in online environments. *The Sloan Consortium*. Retrieved November 2, 2006, from www.sloan-c.org/publications/books/interactions.pdf

Zhang, S., & Fulford, C. (1994). Are interaction time and psychological interactivity the same thing in the distance learning television classroom? *Educational Technology, 34*(4), 58-64.

KEY TERMS

Andragogy: Design and instructional philosophy based on the following six assumptions about the adult learner: (1) learner's need to know; (2) self-concept of the learner; (3) prior experience of the learner; (4) readiness to learn; (5) orientation to learning; and (6) motivation to learn (Knowles et al., 1998).

Competence: A measure of perceived level of ability by faculty in the use of electronic technologies often associated with distance education (Jones, Lindner, Murphy, & Dooley, 2002).

E-Learning: "The appropriate application of the Internet to support the delivery of learning, skills, and knowledge in a holistic approach not limited to any particular course, technologies, or infrastructures" (Henry, 2001, p. 249).

Incentive: Intrinsic or extrinsic motivational factors that impact faculty decisions to participate in distance education.

Instructional Issues: Determining the instructional needs, resource availability, curriculum design, course development, and faculty capacity and incentives for effective delivery of distance education (Bunn, 2001).

Learner-Centered Instruction: Any formal or nonformal education that accounts for a learner's cognitive and metacognitive factors, motivational and affective factors, developmental and social factors, and individual differences (APA, 1997).

Organizational Issues: Developing infrastructure, technical systems, resource allocations, professional development, and organizational restructuring necessary for the efficient delivery of distance education (Bunn, 2001).

Student-Related Issues: Policies and practices that impact the needs of the learner (client or customer) in terms of recruitment, enrollment, retention, and graduation (Bunn, 2001).

Vicarious Interaction: A student's perception of the interactions between others in the learning environment (Zhang & Fulford, 1994).

Financial Trading Systems Using Artificial Neural Networks

Bruce Vanstone
Bond University, Australia

Gavin Finnie
Bond University, Australia

INTRODUCTION

Soft computing represents that area of computing adapted from the physical sciences. Artificial intelligence techniques within this realm attempt to solve problems by applying physical laws and processes. This style of computing is particularly tolerant of imprecision and uncertainty, making the approach attractive to those researching within "noisy" realms, where the signal-to-noise ratio is quite low. Soft computing is normally accepted to include the three key areas of fuzzy logic, artificial neural networks, and probabilistic reasoning (which include genetic algorithms, chaos theory, etc.).

The arena of investment trading is one such field where there is an abundance of noisy data. It is in this area that traditional computing typically gives way to soft computing as the rigid conditions applied by traditional computing cannot be met. This is particularly evident where the same sets of input conditions may appear to invoke different outcomes, or there is an abundance of missing or poor quality data.

Artificial neural networks (henceforth ANNs) are a particularly promising branch on the tree of soft computing, as they possess the ability to determine non-linear relationships, and are particularly adept at dealing with noisy datasets.

From an investment point of view, ANNs are particularly attractive as they offer the possibility of achieving higher investment returns for two distinct reasons. Firstly, with the advent of cheaper computing power, many mathematical techniques have come to be in common use, effectively minimizing any advantage they had introduced (see Samuel & Malakkal, 1990). Secondly, in order to attempt to address the first issue, many techniques have become more complex. There is a real risk that the signal-to-noise ratio associated with such techniques may be becoming lower, particularly in the area of pattern recognition, as discussed by Blakey (2002).

Investment and financial trading is normally divided into two major disciplines: fundamental analysis and technical analysis. Articles concerned with applying ANNs to these two disciplines are reviewed.

BACKGROUND

There are a number of approaches within the literatures, which deal with applying ANN techniques to investment and trading. Although there appears to be no formal segmentation of these different approaches, this review classifies the literature into the topics proposed by Tan (2001), and augments these classifications with one more category, namely, hybrid. These categories of ANN, then, are:

- **Time series:** Forecasting future data points using historical data sets. Research reviewed in this area generally attempts to predict the future values of some time series. Possible time series include Base time series data (e.g., closing prices), or time series derived from base data, (e.g., indicators--frequently used in technical analysis).
- **Pattern recognition and classification:** Attempts to classify observations into categories, generally by learning patterns in the data. Research reviewed in this area involved the detection of patterns, and segregation of base data into "winner" and "loser" categories as well as in financial distress and bankruptcy prediction.
- **Optimization:** Involves solving problems where patterns in the data are not known, often non-polynomial (NP)-complete problems. Research reviewed in this area covered the optimal selection of parameters, and determining the optimal point at which to enter transactions.
- **Hybrid:** This category was used to distinguish research, which attempted to exploit the synergy effect by combining more than one of the previous styles.

There appears to be a wide acceptance of the benefit of the synergy effect, whereby the whole is seen as being greater than the sum of the individual parts.

Further, the bias in this style of research toward technical analysis techniques is also evident from the table, with one-third of the research pursuing the area of pattern recognition and classification. Technical analysis particularly lends itself to this style of research, as a large focus of technical analysis concerns the detection of patterns in data, and the

examination of the behavior of market participants when these patterns are manifest.

USING NEURAL NETWORKS TO DEVELOP TRADING SYSTEMS

This section briefly considers the characteristics of each of the four main categories previously described. The selected articles were chosen as they are either representative of current research directions, represent an important change in direction for this style of research, or represent a novel approach.

Research into Time Series Prediction

The area of time series predictions is normally focused on attempting to predict the future values of a time series in one of two primary ways, either:

- Predicting future values of a series from the past values of that same series
- Predicting future values of a series using data from different series

Typically, current research in this area focuses on predicting returns, or some variable thought to correlate with returns (e.g., earnings). Some researchers focus on attempting to predict future direction of a series (e.g., increasing from last known value, decreasing from last known value, no change). Research of this nature is essentially a classification problem, and is discussed in that section.

The following articles were selected and reviewed as they are representative of the current research in Time Series Prediction (Austin et al., 1997; Chan & Foo, 1995; Falas et al., 1994; Hobbs & Bourbakis, 1995; Quah & Srinivasan, 2000, Wang et al., 2003; Yao & Poh, 1995). The articles reviewed consider both fundamental and technical data. For example, Falas et al. (1994) used ANNs to attempt to predict future earnings based on reported accounting variables. They found no significant benefit using ANNs compared to the logit model and concluded that the accounting variables chosen were not appropriate earnings predictors. This conclusion represents one of the major problems encountered when working with ANNs, namely, their non-existent explanatory capability. It is not unusual to find conclusions of this type when reviewing ANN research with non-correlation often being reported as wrongly chosen input variables. Quah et al. (2000) use mainly accounting variables to predict excess returns (with limited success). Chan et al. (1995) use ANNs to predict future time series values of stock prices, and use these "future" values to compute a variety of technical indicators. The ANN produced showed particularly promising results, the authors conclude that the networks ability to predict allows the trader to enter a trade a day or two before it is signalled by regular technical indicators, and that this accounts for the substantially increased profit potential of the network.

In many ways, these two primary prediction methodologies relate quite closely to technical analysis strategies. For example, the use (and projection) of a moving average over a series of stock prices could be regarded as predicting future values of a series (the moving average) from past values of the same series. Indicators in technical analysis are often composed of a number of constituent data items, like price, volume, open-interest, etc. These indicators are commonly used to give indications of future direction of price.

Research into Pattern Recognition and Classification

Pattern recognition techniques and classification techniques have been grouped together, as their goal is normally not to predict future values of a time series, but to predict future direction of a time series. For example, the primary goal of chartists (a style of technical analyst) is to attempt to predict trend turning points by studying chart price action, looking for certain patterns. Chartists have noticed that these patterns tend to re-occur, and are reasonably reliable indicators of the future direction of price trends. There are a great deal of these chart patterns, and different analysts attach different weightings to the predictive power of any given pattern. Also, these patterns normally need to be confirmed by values from another time series (such as volume) to be considered "reliable." For more detail on this area, the reader is encouraged to refer to Pring (1999). Non-pattern matching techniques, which also attempt to predict future direction of a time series are also classification problems. Quite often, in addition to predicting future direction of a time series, classification research attempts to classify stocks into two main groups, namely "winners" and "losers" as in bankruptcy and financial distress predictions.

The following articles were selected and reviewed as they are representative of the current research in pattern recognition and classification (Baba & Handa, 1995; Baba et al., 2004; Baba & Nomura, 2005; Baba et al., 2001; Baek & Cho, 2000; Enke & Thawornwong, 2005; Fu et al., 2001; Kamijo & Tanigawa, 1990; Michalak & Lipinski, 2005; Mizuno et al., 1998; Skabar & Cloete, 2001; Suh & LaBarre, 1995; Tan & Quek, 2005). As previously described, the research can generally be classified as "winner" and "loser" detection or pattern matching. The work of Tan et al. (2005), and later, Tan and Dihardjo uses the concept of "winner" and "loser" classification, as does Longo et al. (1995) and Skabar et al. (2001). Specifically, Skabar et al. (2001) do not predict "winners" and "losers," but predict two disparate categories,

namely "up" and "down" (direction of returns). The work of Kamijo et al. (1990) provides an excellent example of pattern matching with the authors building ANNs to identify "triangle" patterns in stock market data (the "triangle" is a specific pattern used in technical analysis).

Classification involving pattern matching could also be validly discussed under the previous section on time series prediction, due to the fact that pattern constructs must occur in specific time order, and the majority of patterns are not time invariant. This leads to the desire of researchers to identify time invariant patterns, or attempt to determine a fixed period of time in which a pattern should occur. The work of Fu et al. (2001) provides examples of using genetic algorithms to "fix" the length of patterns, making them suitable for study using ANNs.

Research into Optimization

The focus of optimization is directed toward research that uses soft-computing specifically to attempt to optimize an otherwise accepted achievable result. Typical of this style of research article, an already accepted result is discussed, then considered for optimization. The optimization is characteristically proven by excess returns compared to the un-optimized case.

For an example of this style of optimization using ANNs, an index arbitrage timing has been proposed by Chen et al. (2001) and Zimmerman and Grothmann (2005). Their model attempts to optimise the correct entry point timing for index arbitrage positions. Current arbitrage models propose establishing an arbitrage position immediately an opportunity arises; the neural network approach is to attempt to locate the timing when there will be a maximum basis spread for the arbitrage, thereby increasing profit potential. Their research concludes that the neural model significantly outperforms the traditional approach.

Research into Ensemble Approaches

Research is classified as an ensemble approach if it combines work from more than one of the areas described next, effectively attempting to leverage the synergy effect by achieving an end result greater than that expected from each of the individual constituents. Among soft-computing research, there is a growing trend towards using the ensemble approach to analysis.

The following articles were selected and reviewed as they are representative of the current research in ensembles (Abdullah & Ganpathy, 2000; Baba et al., 2002; Chenoweth et al., 1995; Doeksen et al., 2005; Jang et al., 1991; Leigh et al., 2002; Liu & Lee, 1997; Wong & Lee, 1993). The majority of the ensembles draw their components from a variety of soft computing methods. The use of ANNs and genetic algorithms (GAs) together is quite popular, and is used by Leigh et al. (2002) to combine pattern recognition techniques with price forecasting. Another approach combining ANNs and GAs is provided by Baba et al. (2002) using ANNs for their predictive ability, and GAs to determine the best way to react to that information. Some ensembles combine multiple ANNs, for example, Jang et al. (1991) combine two ANNs, one that takes a short-term view of market movement, with one that takes a longer-term view. They then build a model, which reacts to a weighted output sum of the outputs of both models. Both Liu and Lee (1997) and Abdullah et al. (2000) also used ensembles of ANNs and concluded that the predictive ability of the ensemble approaches exceeded that of the individual ANNs. Other research reviewed combined ANNs with fuzzy logic and expert systems.

FUTURE TRENDS

Essentially, the field of financial trading is in a state of transition between traditional pricing models, the efficient market hypothesis, and ideas about behavioral finance. The challenge that presents itself is how best to unify financial trading pricing models. There is much debate about the validity of the efficient market hypothesis, which effectively contends that prices cannot be predicted using methods such as technical analysis. There is a large body of evidence that appears to contradict the efficient market hypothesis, and there seems little chance of academically moving forward en-masse unless an alternative valid pricing model exists. This offers substantial opportunity for soft computing research techniques, particularly neural models. These models are capable of acting as universal approximators, and determining complex non-linear relationships. The goal with these methods is to attempt to mine deep relationships, which can shed new light about the behaviour of markets and prices. These new relationships would inherently provide more scope for developing feasible and effective pricing models.

CONCLUSION

This article has surveyed recent and key literature in the domain of applying artificial neural networks to investment and financial trading. Within the context of investment discipline, this survey shows that the majority of this type of research is being conducted in the field of technical analysis. As discussed earlier, soft computing is particularly data intensive, and it is suggested that this observation goes someway to explaining this obvious bias in research.

Within the area of soft computing styles, the survey finds that the majority of research is within the area of both hybrid systems and pattern recognition and classification. It is suggested the reason for this is that the technical analysis approach lends itself towards the pattern recognition and classification

areas. Also, many hybrid systems include pattern recognition and classification as one of their constituents.

REFERENCES

Abdullah, M. H. L. B., & Ganapathy, V. (2000). Neural network ensemble for financial trend prediction. *Tencon 2000 Theme: Intelligent Systems and Technologies for the New Millennium* (pp. 157-161). Kuala Lumpur, Malaysia.

Austin, M., C. Looney, & Zhuo, J. (1997). Security market timing using neural network models. *New Review of Applied Expert Systems, 3,* 3-14.

Baba, N., & Handa, H. (1995). Utilization of neural network for constructing a user friendly decision support system to deal stocks. *IEEE International Conference on Neural Networks* (pp. 818-823). Australia.

Baba, N., N. Inoue, & Yanjun, Y. (2002). Utilization of soft computing techniques for contructing reliable decision support systems for dealing stocks. *IJCNN'02: 2002 International Joint Conference on Neural Networks* (pp. 2150-2155). Honolulu, Hawaii.

Baba, N., T. Kawachi, Nomura, T., & Sakatani, Y. (2004). Utilization of NNs & GAs for improving the traditional technical analysis in the financial market. *SICE 2004 Annual Conference* (pp. 1409-1412). Sapporo, Japan.

Baba, N., & Nomura, T. (2005). An intelligent utilization of neural networks for improving the traditional technical analysis in the stock markets. *KES 2005* (pp. 8-14).

Baba, N., Y. Yanjun, Naoyuki, I., Lina, X., & Zhenglong, D. (2001). Knowledge-based decision support systems for dealing Nikkei-225 by soft computing techniques. *Knowledge-Based Intelligent Information Engineering Systems and Allied Technologies KES 2001* (pp. 728-732). Netherlands: IOS Press.

Baek, J., & Cho, S. (2000). "Left shoulder" detection in Korea composite stock price index using an auto-associative neural network. *Intelligent Data Engineering and Automated Learning—IDEAL 2000: Data Mining, Financial Engineering and Intelligent Agents* (pp. 286-291). Hong Kong.

Blakey, P. (2002). Pattern recognition techniques. *IEEE Microwave Magazine, 3,* 28-33.

Chan, K. C. C., & Foo, K. T. (1995). Enhancing technical analysis in the Forex market using neural networks. *IEEE International Conference on Neural Networks* (pp. 1023-1027). Australia.

Chen, A., Chianglin, C., & Chung, H. (2001). Establishing an index arbitrage model by applying neural networks method—A case study of Nikkei 225 index. *International Journal of Neural Systems, 11*(5), 489-496.

Chenoweth, T., Obradovic, Z., & Lee, S. (1995). Technical trading rules as a prior knowledge to a neural networks prediction system for the S&P 500 index. *IEEE Technical Applications Conference and Workshops* (pp. 111-115). Portland, Oregon.

Doeksen, B., Abraham, A., Thomas, J., & Paprzycki, M. (2005). Real stock trading using soft computing models. *ITCC 2005: International Conference on Information Technology* (pp. 162-167).

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications, 29*(4), 927-940.

Falas, T., Charitou, A., & Charalambous, C. (1994). *The application of artificial neural networks in the prediction of earnings*. Orlando: IEEE Press.

Fu, T. C., Chung, F. L., Ng, V., & Luk, R . (2001). Evolutionary segmentation of financial time series into subsequences. *2001 Congress on Evolutionary Computation* (pp. 426-430). Seoul, Korea.

Hobbs, A., & Bourbakis, N. G. (1995). A neurofuzzy arbitrage simulator for stock investing. *International Conference on Computational Intelligence for Financial Engineering (CIFER)* (pp. 160-177). New York.

Jang, G., Lai, F., Jiang, B., & Chien, L. (1991). An intelligent trend prediction and reversal recognition system using dual-module neural networks. The *1st International Conference on Artificial Intelligence Applications on Wall Street* (pp. 42-51). New York.

Kamijo, K., & Tanigawa, T. (1990). Stock price pattern recognition: A recurrent neural network approach. In *Proceedings of the International Joint Conference on Neural Networks* (pp. 215-221). San Diego.

Leigh, W., Purvis, R., & Ragusa, J. M. (2002). Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: A case study in romantic decision support. *Decision Support Systems, 32*(4), 361-377.

Liu, N. K., & Lee, K. K. (1997). An intelligent business advisor system for stock investment. *Expert Systems, 14*(3), 129-139.

Michalak, K., & Lipinski, P. (2005). Prediction of high increases in stock prices using neural networks. *Neural Network World, 15*(4), 359-66.

Mizuno, H., Kosaka, M., Yajima, H., & Komoda, N. (1998). Application of neural network to technical analysis of stock

market prediction. *Studies in Informatics and Control, 7*(2), 111-120.

Pring, M. J. (1999). *Martin Pring's introduction to technical analysis*. Singapore: McGraw-Hill.

Quah, T. S., & Srinivasan, B. (2000). Utilizing neural networks in stock pickings. *International Conference on Artificial Intelligence* (pp. 941-6).

Samuel, C., & Malakkal, I. (1990). Leading-edge investors downplay debate on fundamental vs. technical analysis. *Wall Street Computer Review, 7*, 22-28,53.

Skabar, A., & Cloete, I. (2001). Discovery of financial trading rules. *IASTED International Conference on Artificial Intelligence and Applications (AIA 2001)* (pp. 121-125). Marbella, Spain.

Suh, Y. H., & LaBarre, J. E. (1995). An application of artificial neural network models to portfolio selection. *Journal of Computer Information Systems, 36*(1), 65-73.

Tan, A., & Quek, C. (2005). Maximizing winning trades using a rough set based pther-product (RSPOP) fuzzy neural network intelligent stock trading system. *IEEE CEC2005: IEEE Congress on Evolutionary Computation* (pp. 2076-83).

Tan, C. N. W. (2001). *Artificial neural networks: Applications in financial distress prediction and foreign exchange trading*. Gold Coast, QLD: Wilberto Press.

Wang, X., P. K. H. Phua, et al. (2003). Stock market prediction using neural networks: does trading volume help in short-term prediction. *International Joint Conference on Neural Networks 2003: IJCNN 03* (pp. 2438-2442).

Wong, F., & Lee, D. (1993). Hybrid neural network for stock selection. *The 2nd Annual International Conference on Artificial Intelligence Applications on Wall Street*, New York (pp. 294-301).

Yao, J., & Poh, H. L. (1995). Forecasting the KLSE index using neural networks. *IEEE International Conference on Neural Networks (ICNN'95)* (pp. 1012-1017). Perth, Western Australia.

Zimmermann, H. G., & Grothmann, R. (2005). Optimal asset allocation for a large number of investment opportunities. *International Journal of Intelligent Systems in Accounting, Finance, and Management, 13*(1), 33-40.

KEY TERMS

Continuation Pattern: A pattern in technical analysis, which suggests, on the balance of probabilities, that price trend will continue in its current direction.

Fundamental Analysis: The use of company reported financial data to determine an intrinsic (or fair value) for a security. Used to identify cases where companies are undervalued, with a view to profiting from the future price movements. This style of analysis is generally long-term.

Noisy Data: This term is generally used to describe data and datasets where there is a low signal-to-noise ratio. Any algorithm attempting to filter out the signal has to be capable of identifying and dealing appropriately with noise. In this sense, noise is that element of the data, which obscures the true signal.

Reversal Pattern: A pattern in technical analysis, which suggests, on the balance of probabilities, that price trend will change direction.

Technical Analysis: The study of the behavior of market participants, as reflected in the technical data. Used to identify early stages in trend developments, with a view to profiting from price movements. This style of analysis is generally short term.

Technical Data: Technical data is the term used to describe the components of price history for a security. These components are open price, low price, high price, close price, volume traded, and open interest.

Technical Indicators: Technical indicators are produced as results of various computations on technical data. They are primarily designed to confirm price action.

Triangle Pattern: A triangle is a particular pattern observed using technical analysis. There are a variety of circumstances under which a triangle can occur, and dependant on the circumstances, the triangle can be either a reversal or continuation pattern.

Focused Requirements Engineering Method for Web Application Development

Ala M. Abu-Samaha
Amman University, Jordan

Lana S. Al-Salem
SpecTec Ltd & MEP, Greece

INTRODUCTION

The requirements phase of the system/application development process typically involves the activities of requirements elicitation, analysis, validation, and specification. The main goal of such a process is "to develop a requirements specification document which defines the system to be procured and which can act as a basis for the system design" (Sawyer, Sommerville, & Viller, 1996). Hence the underpinning assumption of the requirements engineering (RE) process is to transform the operational needs of an organisation into complete, consistent, and unambiguous system/application specifications through an iterative process of definition and validation (Pohl ,1994).

The Web engineering (WE) literature provides a limited number of methods and techniques that can be used to manage the RE process in a Web development context [*e³-value* framework (Gordijn, Akkermans, & van Vliet, 2000), SOARE approach (Bleistein, Aurum, Cox, & Ray, 2004), e-prototyping (Bleek, Jeenicke, & Klischewski, 2002), AWARE (Bolchini & Paolini, 2004), and SSM/ICDT (Meldrum & Rose, 2004)]. Despite the availability of such a limited number of Web requirements engineering (WRE) methods, many researchers criticised such methods for their failure to address the necessity to align the Web application' requirements to the organisation's business strategy. Hence, the recommendation of many researchers (Al-Salem & Abu-Samaha, 2005a; Bleistein 2005; Bleistein, Cox, & Verner, 2004; Vidgen, Avison, Wood, & Wood-Harper, 2002) is to utilise a general WRE framework for the development of Web applications that can align the application's requirements to the organisation's business needs and its future vision. The objective of such a WRE framework is to incorporate the elicitation/analysis of business strategy as part of the application's RE process.

This chapter presents a WRE method that extends Sommerville and Kotonya's viewpoint-oriented requirements definition (VORD) and Kaplan and Norton's balanced scorecard (BSC) to elicit the Web application' requirements and to plan/analyze the business strategy, respectively. In addition, eWARE (extended Web application requirements engineering) deploys the concept of "requirements alignment" to attain business objectives during the requirements discovery, elicitation, and formalisation process to identify the services of the Web application that will achieve the business objectives in order to improve the organisation's profitability and competitiveness. The chapter is organised into a number of sections. The second section of this chapter provides a background to Web applications in terms of definition and differentiating characteristics. The third section provides a discussion of eWARE method in terms of phases and activities. This section is divided into two subsections to cover the activities of the two prominent phases of the eWARE process in more detail. The fourth and fifth sections provide a discussion of possible future trends in WRE and a number of concluding remarks.

BACKGROUND

Web applications provide organisations an unprecedented chance to stretch their existence beyond the typical boundaries of an organisation to include customers, trading partners, and suppliers.Little attention has been paid to the process of RE for Web application development, in comparison to other areas of the development process [modelling, design, and coding] (Ginige & Murugesan, 2001). Hence, Web applications can be defined as "applications that tend to be used to integrate and streamline an organisation's business processes beyond organisational (customers, agents, suppliers, others) and geographical borders; provide an organisation with competitive products and services that give it a strategic advantage over its competitors in the marketplace; promote business innovation; and/or improve operational efficiency" (Al-Salem & Abu-Samaha, 2005a).

There is a pressing need in the WE discipline for RE approaches and techniques that (a) take into account the multiplicity of user profiles and the various stakeholders involved [a stakeholder is defined as "anyone who can share information about the system, its implementation constraints or the problem domain" (Potts, Takahashi, & Anton, 1994)]; (b) eliciting overall functionality and the business environ-

ment of the Web application; (c) specifying technical and nontechnical requirements of the Web application, and (d) aligning the Web application' requirements to the overall business strategy (Bleistein et al., 2005; Ceri, Fraternali, Bongio, Brambilla, Comai, & Matera, 2003; Ginige & Murugesan, 2001; Kautz & Madsen, 2003; Lowe, 2003; Meldrum & Rose, 2004; Nuseibeh & Easterbrook, 2000; Vidgen et al., 2002). More importantly, a Web application must be developed with an emphasis on how the services of such an application can achieve the business vision and strategy and fulfil the business processes (Haire, Henderson-Sellers, & Lowe, 2001).

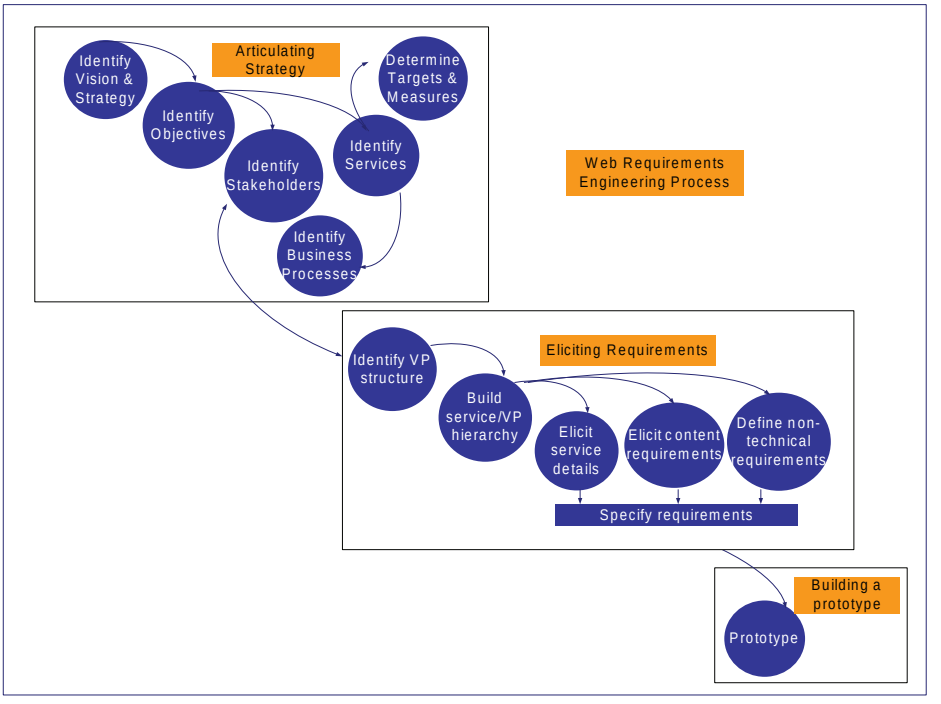
eWARE Process

eWARE process can be best perceived as a series of activities grouped into three phases; strategy articulation via BSC, Web application' requirements elicitation via VORD, and prototype building; Figure (1) presents the phases and activities of eWARE. Such a process aims to develop a Web application requirements specification (WRS) document that is aligned with business strategy and detailed enough to be used for contractual purposes.

The strategy articulation phase of eWARE can be best thought of as a structure of many layers. The vision of the organisation is at the top of the structure, while the strategic objectives are presented in the next layer of the structure followed by the Web application services. The next layer of the structure contains the measurements and targets for meeting the strategic objectives. The level of detail tends to increase as we move down the structure of the strategy. This articulation of the organisation's vision, objectives, and measurements aims to translate the future vision of the organisation into detailed and prioritised Web application requirements (this will be fully covered in the coming subsections).

The requirements elicitation phase of the eWARE process is used to produce a WRS document based on requirements collected during the strategy articulation phase. Requirements elicitation relies on the identification of the relevant viewpoints (VPs), their sub-VPs, and requirements for each viewpoint (VP). Kotonya and Sommerivlle (1996) define a VP as anyone who may have some direct or indirect influence on the system/application requirements. Goals and objectives of the different stakeholder groups need to be identified to define success or failure measures for each stakeholder. Moreover, nonfunctional requirements {NFR) need to be

Figure 1. The eWARE process



identified; these include some of the "-ilities" of the Web application, such as reliability, supportability, maintainability, affordability, and so forth. Finally, in the prototyping phase of the eWARE process, the system/application stakeholders consider the unclear set of requirements, and agree on prototyping the ambiguous requirements to verify them. Moreover, the user interface (UI) and Web site structure are presented in a throw away prototype.

The mentioned WRE phases and activities are perceived to be iterative and incremental in nature where unmet targets are questioned. This cyclic view of the WRE process will trigger the strategy articulation phase to enter in a feedback loop in order to refine services and change requirements of the Web application in order to enhance the organisation's chance to achieve its set vision and strategy.

Strategy Articulation Phase

As mentioned earlier, the strategy articulation phase of eWARE process aims to align the Web application' requirements to the organisation's business strategy through a process of strategy analysis. eWARE delivers such alignment via eBSC (extended Balanced Score Card). According to eBSC, aligning the Web application' requirements to the business objectives yields four perspectives to focus upon (stakeholders, strategic objectives, internal processes, and Web application services). Figure 2 provides a diagrammatic representation of the perceived relationship between the four perspectives.

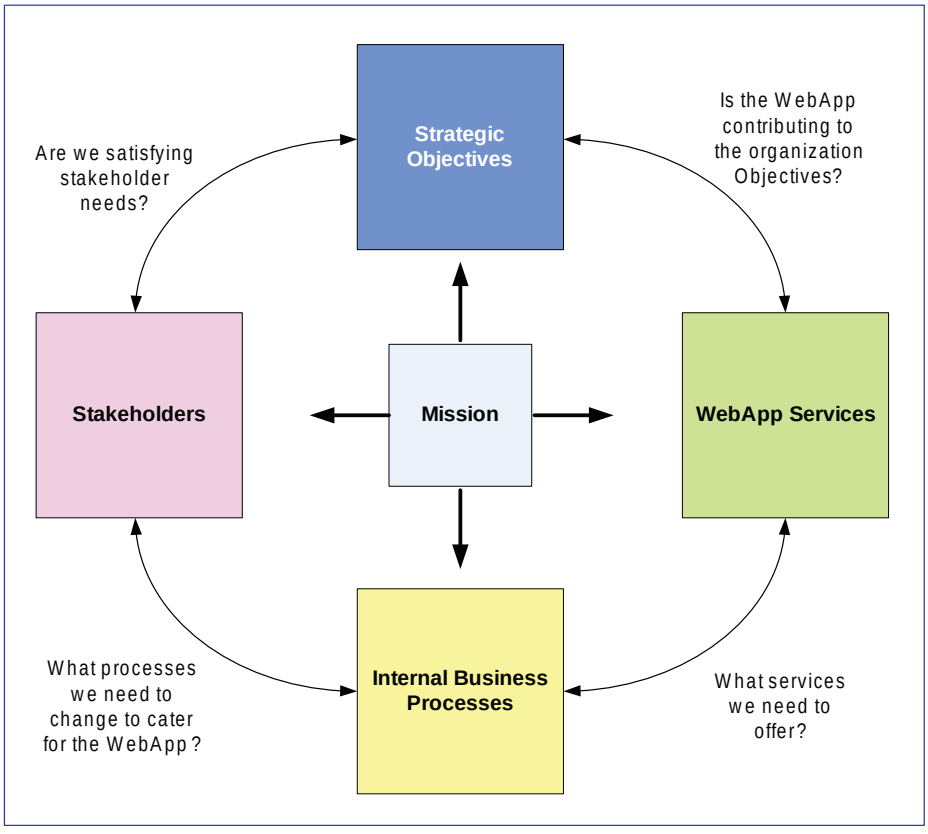
Strategic Objectives Perspective

Objectives are statements that clarify what the strategy aims to achieve. Hence, organisations pursue strategies in the belief that, when implemented, they will enable the organisation to better achieve its strategic objectives.

Web Application Services Perspective

The Web application services are the collection of functionality, quality, content, and all what the Web application must provide in order to achieve the strategic objectives of the

Figure 2. eBSC perspectives



organisation and to meet the expectations of its stakeholders. This perspective is considered as a precondition to process improvement, stakeholder satisfaction, and business objectives' realisation. Hence, the development team, comprising of requirement engineers and end user representatives, must be empowered to select which services (functionalities) of the Web application will be included, in order to deliver the optimal business solution. Decisions and trade-offs need to be made between new services elicited through the eBSC process, and less important original functionalities that may have to be excluded, being less strategically important.

Stakeholders[1] Perspective

The stakeholder perspective is arguably the most important perspective in the Web application development process. As mentioned earlier, a stakeholder is defined by Potts et al. (1994) as "anyone who can share information about the system, its implementation constraints or the problem domain." According to Potts et al. (1994), the list of stakeholders will include end users, indirect users, other customer representatives, and developers. Since a stakeholder is, in essence, a requirements source; it can be an application user, a competitor, or even a third party. Hence, if the stakeholder represents a customer, then the requirements elicitation effort will focus on new customer acquisition, customer retention, and customer profitability. The goal of using eBSC is to classify the organisation's strategy by stakeholder, which will lead to introducing Web application services for each group to meet their strategy or objective. The organisation must determine whom it serves and how their requirements can be met.

Internal Process Perspective

Organisations' activities can be grouped into "business processes" that describe the way work is to be implemented. Some of the organisation's activities will be affected by the introduction of a Web application, since such applications have the potential to significantly change an organisation's work practices and procedures (Ginige & Murugesan, 2001; Pressman, 2004). The internal process perspective focuses on key processes at which the organisation must excel in order to add value to its stakeholders through the Web application.

Figure 3. Web application cause and effect relationships

When designing a score card, the starting point will be asking "what strategies do we have to put in place to satisfy the wants and needs of the key stakeholders?" The inclusion of BSC within the WRE process aims to help clarify, consolidate, and gain consensus around the business–Web application strategy of the organisation, and to translate the business strategy into Web application services (requirements) to ensure that the elicited requirements are strategy focused. Hence, a strategy can be best described as a series of cause and effect relationships that provide a translation from future vision to Web requirements, as shown in Figure (3).

eVORD for Eliciting Web Requirements

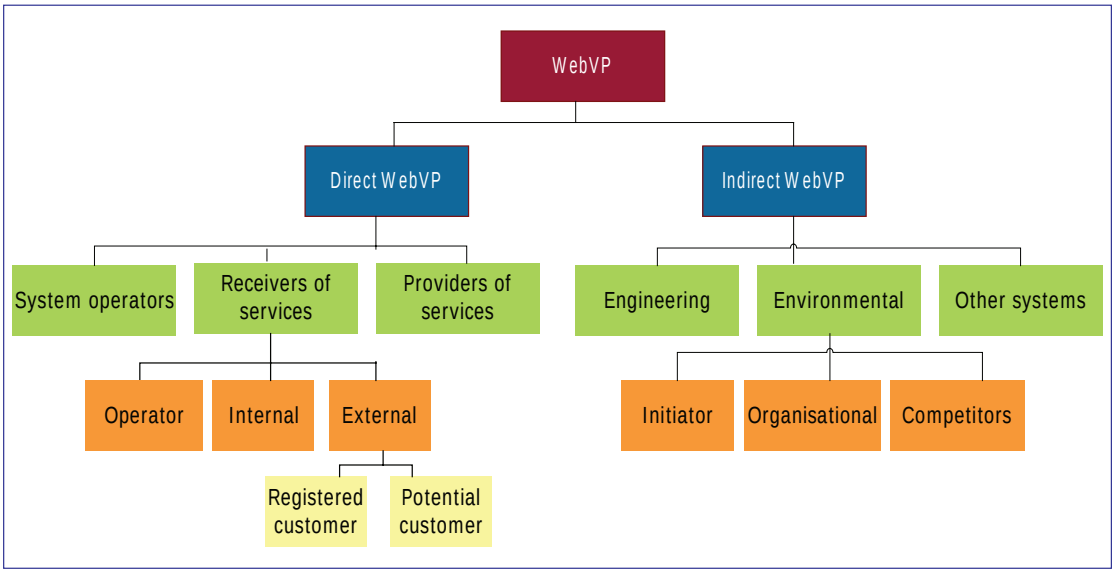
The second phase of the proposed WRE process presents the requirements elicitation phase of the Web application development process. eWARE delivers such elicitation via eVORD (extended viewpoint-oriented requirements definition). According to eVORD, templates are created to describe each viewpoint (VP), service, nonfunctional requirement (NFR), and content. As mentioned earlier, Kotonya and Sommerville (1996) define a VP as anyone who may have some direct or indirect influence on the system/application requirements. The VORD requirements engineering process includes activities concerned with VP identification, VP service description, cross-viewpoint analysis to discover inconsistencies, omissions, and conflicts, and developing an object-oriented model of the system/application from the VP analysis. In addition, a VP diagram is used to show the relationships among VPs, while sequence diagrams illustrate the interactions among VPs.

Web Application Viewpoints (WebVPs) Identification and Structuring

The construction of Web applications involves a great number of stakeholders who have different views of the Web application. These perspectives are partial or incomplete descriptions of the system/application, and reflect the environment in which the system/application will operate in. The integration of multiple views can contribute to augment the overall understanding of the Web application. The authors recommend the use of a new and amended list of VPs for developing Web applications. These WebVPs' abstracts have been identified as a starting point that acts as a template for WebVPs classes and hierarchy. Figure (4) shows a high-level abstract of WebVPs structure.

WebVPs can be classified into *direct VPs* that interact directly with the Web application and fall into two subclasses (receivers and providers of services), and *indirect VPs* that have "interest" in some or all of the services that are delivered by the system but do not interact directly with it; hence, they provide high-level organisational requirements and constraints. Indirect VPs fall into a number of subclasses: *environmental WebVPs,* which reflect the requirements of the business domain, that is, legalisation, localisation, taxa-

Figure 4. Abstract of WebVPs structure



tion, and competitors; *engineering WebVPs,* which reflect the requirements of the development team, that is, software engineers, team leaders, and creative designers; and *system WebVPs,* which include all existing information systems that the application being analysed needs to interface to, that is, payment systems and supplier systems.

Eliciting Requirements Details

The second step of this phase is concerned with documenting the details of each WebVP (identified in the previous step) and its associated services. For every identified WebVP, a number of templates are used to elicit and specify its requirements. The authors have extended and adapted VORD templates to cater to the particularities of Web applications development (please refer to Al-Salem & Abu-Samaha, 2005a and Al-Salem & Abu-Samaha, 2005b for more details).

Documenting Requirements

The objective of the requirements process is to deliver a requirements specification document that defines the system/application to be developed (Sawyer, Viller, & Sommerville, 1997). This document is used for contractual purposes, and can be used as a basis for facilitating a competitive tendering for the system/application design and implementation (IEEE, 2004). The authors have enhanced the typical software requirements specification (SRS) document to reflect the changes introduced to eVORD and eBSC, as depicted in Figure (5).

FUTURE WORK

The domain of RE, in general, and WRE, in particular, is evolving with the ever-changing contexts of software en-

Figure 5. Software requirements specification (SRS) document template

Table of Contents

gineering and information systems development projects. Despite the availability of many requirements, engineering methods, processes, techniques, and tools, such artefacts are in need of constant extensions and enhancements to bring such artefacts to the changing contexts of the development projects. The presented eWARE method is an extension of two existent methods used to align requirements of a Web application to the organisation's business strategy. eWARE needs to be tested further in different industrial settings to validate the applicability of the method. As well, eWARE, like many other RE methods, needs to be supported through the development of a computer aided software engineering (CASE) tool to facilitate the generation of the WRS document. Compared to similar WRE methods (Bleek et al., 2002; Bleistein et al., 2004; Bolchini & Paolini, 2004; Gordijn et al., 2000; Meldrum & Rose, 2004), eWARE provides its users with a number of benefits: it extends the BSC approach to help with the formulation of Web application strategy as a stage in the WRE process in advance to requirements elicitation; it provides a prioritisation framework synchronised with business strategy; it extends VORD to elicit requirements for Web applications; and it creates a WRS document that specifies both the business strategy and requirements for the Web application. Such advantages proved to be valuable in Web application development projects, as Web applications are different in many aspects when compared to other applications. The differing characteristics of Web applications can be summarized as diverse and volatile requirements, vast and unknown end users, multiple stakeholders, adaptable architecture, short development life cycle, high visibility, heavy content, integration with backend databases and third party applications, Web applications' relevance and direct effect on business, and multidisciplinary development team.

CONCLUSIONS

There are two general opinions on whether current software engineering methods and techniques are applicable to face the challenges of developing a Web application. One opinion advocates the need for a new "software engineering" discipline that handles the Web application particularities (Ginige & Murugesan, 2001, Murugesan, 1999; Murugesan, Deshpande, Hansen, & Ginige, 1999). In contrast, the other opinion believes that current software engineering methods, tools, and techniques are applicable to Web application development. For example, McDonald and Welland (2001) recommended developing new methods, techniques, and approaches to address the challenges of developing Web applications in order to increase the possibility of their success. This implies the need for a new breed of WRE methods, tools,and working practices. In contrast, Pressman (2004) argues that Web applications are a natural evolution of existing applications/systems offering a solution to classical problems exhibited by previous information systems (IS). The authors of this chapter hold a middle position in between these two opinions. The authors perceive Web applications as a natural evolution of "traditional" information systems, yet they believe that such applications possess special characteristics that need to be provided for by the "traditional" RE method(s). Hence, existing RE methods are considered valid for WRE, though they need to be enhanced and extended to cater to the distinguishing features of Web applications.

The chapter has presented a WRE approach that enables businesses to develop/procure Web application(s) capable to achieve the organisations' business strategic objectives, and to effectively harness their business processes. The combination of BSC (Kaplan & Norton 1993, 1996a & b) and VORD (Kotonya & Sommerville 1996) within eWARE process provides the development team with the ability to translate strategy into Web application requirements and to incorporate views from different complementary perspectives. Hence, eWARE is a Web-specific RE process to cope with the complex aspects of requirements elicitation, alignment, and specification for Web applications development/procurement. eWARE views Web applications as organisational initiatives and as such, it takes into account the need to address strategic objectives, business processes issues, requirements details of services, NFR, and integration with other systems.

REFERENCES

Al-Salem, L. S., & Abu-Samaha, A. (2005a). *Assessing the usability of VORD for Web applications requirements engineering - An industrial case study* -. 1st International Workshop on Requirements Engineering for Business Need and IT Alignment (REBNITA 2005), Paris.

Al-Salem, L. S., & Abu-Samaha, A. (2005b). *Assessing the usability of VORD method for Web applications requirements elicitation.* International Conference on Internet Technologies and Applications (ITA 05), Wrexham, North Wales, UK.

Bleek, W.-G., Jeenicke, M., & Klischewski, R. (2002). *Developing Web-based applications through e-prototyping*. 26th Annual International Computer Software and Applications Conference (COMPSAC 2002), Oxford England.

Bleistein, S. J. (2005). *Requirements analysis framework for alignment of IT with competitive strategy of business organizations*. IEEE International Conference on Requirements Engineering - Doctoral Consortium, Paris.

Bleistein, S. J., Aurum, A., Cox, K., & Ray, P. (2004). Strategy-oriented alignment in requirements engineering: Linking

business strategy to tequirements of e-business systems using the SOARE approach. *Journal of Research and Practice in Information Technology, 36*, 259-276.

Bleistein, S. J., Cox, K., & Verner, J. (2005). Validating strategic alignment of organizational IT requirements using goal modelling and problem diagrams. *Journal of Software and Systems, 79*, 362 - 378.

Bolchini, D., & Paolini, P. (2004). Goal-driven requirements analysis for hypermedia-intensive Web applications. *Requirements Engineering Journal Special Issue.*

Ceri, S. Fraternali, P., Bongio, A., Brambilla M., Comai S., & Matera M. (2003). *Designing data-intensive Web applications*. Morgan Kaufman.

Ginige, A. (2002). Web engineering: Managing the complexity of Web systems development. In *14th International Conference on Software Engineering and Knowledge Engineering (SEKE '02),* Ischia, Italy: ACM Press.

Ginige, A., & Murugesan, S. (2001). The essence of Web engineering: Managing the diversity and complexity of Web application development. *IEEE Multimedia 8*(2), 22-25.

Gordijn, J., Akkermans, H. & van Vliet, H. (2000). Value based requirements creation for electronic commerce applications. In *33rd Hawaii International Conference on System Sciences,* IEEE, Hawaii, USA.

Haire, B., Henderson-Sellers, B., & Lowe, D. (2001). Supporting Web development in the open process: Additional tasks. *COMPSAC'2001: International Computer Software and Applications Conference,* Chicago, Illinois, USA, IEEE Computer Society.

IEEE. (2004). *Guide to the software engineering body of knowledge*. IEEE.

Kaplan, R., & Norton, D. (1993). Putting the balanced scorecard to work. *Harvard Business Review,* (September/October), 134-147.

Kaplan, R., & Norton, D. (1996a). *The balanced scorecard: Translating strategy into action*. Boston: Harvard Business School Press.

Kaplan, R., & Norton, D. (1996b). Using the balanced scorecard as a strategic management system. *Harvard Business Review,* (January/February), 75-85.

Kautz, K., & Madsen, S. (2003). *Web development - The differences, similarities and in-betweens.* Melbourne, Australia: Information Systems Development Conference.

Kotonya, G., & Sommerville, I. (1996). Requirements engineering with viewpoints. *BCS/IEE Software Engineering Journal, 11*(1), 5-18.

Lowe, D. (2003). Web system requirements: An overview. *Requirements Engineering Journal, 8,* 102–113.

McDonald, A., & Welland, R. (2001). *Agile Web engineering (AWE) process.* Scotland: Department of Computing Science, University of Glasgow.

Meldrum, M., & Rose, J. (2004). *Activity based generation of requirements For Web-based information systems: The SSM/ICDT approach*. The 12th European Conference on Information Systems (ECIS 2004), Turku Finland.

Murugesan, S. (1999). Web engineering. *ACM SIGWEB Newsletter, 8*(3), 28 - 32.

Murugesan, S., Deshpande, Y., Hansen, S., & Ginige, A. (1999). Web engineering: A new discipline for development of Web-based systems. In *Proceeding of the Int'l Conf. Software Engineering (ICSE) Workshop on Web Engineering* (pp: 1-9). Sydney, Australia: IEEE Multimedia.

Nuseibeh, B., & Easterbrook, S. (2000). Requirements engineering: A roadmap. In *International Conference on Software Engineering (ICSE-2000),* Limerick, Ireland, ACM Press.

Pohl, K. (1994). The three dimensions of requirements engineering. *Information Systems, 19*(3), 243-258.

Potts, C., Takahashi, K., & Anton, A. (1994). Inquiry-based requirements analysis. *IEEE Software, 11*(2), 21-40.

Pressman, R. S. (2004). *Software engineering: A practitioner's approach*. McGraw-Hill.

Sawyer, P., Sommerville, I., & Viller, S. (1996). *PREview: Tackling the real concerns of requirements engineering*. Lancaster: Computing Department, Lancaster University.

Sawyer, P., Viller, S., & Sommerville, I. (1997). Requirements process improvement through the phased introduction of good practice. *Software Process - Improvement and Practice, 1*, 19-34.

Sommerville, I. (1995). *Software engineering*. Addison Wesley.

Sommerville, I., & Sawyer, P. (1997). *Requirements engineering: A good practice guide*. John Wiley & Sons Lts.

Vidgen, R., Avison, D., Wood, J. R. G., & Wood-Harper, A.T. (2002). *Developing Web information systems*. Butterworth Heinemann.

KEY TERMS

Balanced Score Card (BSC): "A multidimensional framework for describing, implementing and managing

strategy at all levels of an enterprise by linking objectives, initiatives and measures to an organisation's strategy" (Kaplan & Norton, 1993, 1996 a & b).

eWARE (extended Web application requirements engineering): "A strategy-focused requirements engineering method used to align Web application requirements to business strategy and to elicit legal, technological, business, marketing and content requirements".

A Requirement: "A condition or capability that must be met or fulfilled by a system to satisfy a contract, standard, specification, or other formally imposed documents." (*IEEE Standard, 610*,12-1990).

Requirements Engineering (RE): "The process of discovering that 'purpose' by identifying stakeholders and their needs, and documenting them in a form that is amenable to analysis, communication, and subsequent implementation" (Nuseibeh & Easterbrook, 2000).

Stakeholder: "Anyone who can share information about the system, its implementation constraints or the problem domain, including end users, indirect users, other customer representatives and developers" (Potts et al., 1994).

Viewpoint (VP): "Any one who may have some direct or indirect influence on the system requirements" (Kotonya & Sommerville, 1996).

Viewpoint Oriented Requirements Definition (VORD): "A software requirements engineering approach used to organise both the elicitation process and the requirements themselves into viewpoints" (Sommerville, 1995).

Web Business Application (WebApp): "An application that tends to be used to integrate and streamline an organisation's business processes beyond organisational (customers, agents, suppliers, others) and geographical borders; to provide an organisation with competitive products and services that give it a strategic advantage over its competitors in the marketplace; to promote business innovation and/or to improve operational efficiency" (Al-Salem & Abu-Samaha, 2005a).

ENDNOTE

[1] Throughout eWARE, a stakeholder /viewpoint is used interchangeably.

A Formal Definition of Information Systems

Manuel Mora
Autonomous University of Aguascalientes, Mexico

Ovsei Gelman
Universidad Nacional Autónoma de México, Mexico

Francisco Cervantes
Universidad Nacional Autónoma de México, Mexico

Guisseppi Forgionne
University of Maryland, Baltimore County, USA

INTRODUCTION

Since its conceptualization in the 1960s (Adam & Fitzgerald, 2000), information systems (IS) has undertaken a hard effort to be recognized as a scientific discipline. Nowadays, indicators such as the existence of undergraduate, master, and doctoral programs; research centers focused on IS topics; specialized conferences and journals; and professional and academic associations suggest that the IS discipline is a scientific field that is independent from its root disciplines (e.g., computer science, management science, accounting, and behavioral sciences).

On the other hand, during this 50-year path, the discipline of information systems can be critiqued for the multiple self-identities perceived by the different stakeholders (e.g., IS researchers, IS practitioners, and IS users). Gelman, Mora, Forgionne, and Cervantes (2005) point out the following weaknesses IS exhibits, making it a still immature field:

i. the scarce utilization of deductive and formal (e.g., logical-mathematical) research models and methods (Farhoomand, 1987, p. 55);
ii. the lack of a formal and standard set of fundamental well-defined concepts used in the discipline (Banville & Landry, 1989, p. 56; Alter, 2001, p. 3; Wand & Weber; 1990, p. 1282); and
iii. the excessive number of availiable micro-theories (Barkhi & Sheetz, 2001, p. 11).

Additionally, the partial, disparate, and not consensual conceptualizations of what is the focus of study in IS is(Alter 2003; Benbazat & Zmud, 2003), along with the lack of integration of multiple research methodologies to cope with the complexity of the phenomena of study (Mingers, 2001), also suggest that the maturity-development process for the IS discipline still is an ongoing process.

Gelman et al. (2005), based on a profound study of the term *information system* (Mora, Cervantes, Mejia, & Weitzenfeld, 2002), confirmed that the fundamental concepts used in most IS research are based on few and misused core concepts from what is the Theory of Systems (Ackoff, 1960, 1971), and that the few proposals for formalization (Wand & Weber, 1990; Mentzas, 1994; Alter, 2001, 2003) are still incomplete. Furthermore, although Systems Science concepts were used in the two most comprehensive IS research frameworks reported in the IS literature (Ives, Hamilton, & Davis, 1980; Nolan & Wetherbe, 1980), a recent study also identified conceptual inconsistency and incompleteness in both frameworks from a formal systemic view (Mora, Gelman, Cano, Cervantes, & Forgionne, 2006). Hence, it can be inferred that the utilization of an informal, conflicting, and ambiguous communicational system in the IS discipline (Banville & Landry, 1989) and the lack of a comprehensive IS research framework have hindered the development of a cumulative research tradition and delayed the maturation of the field (Wand & Weber, 1990; Farhoomand, 1987).

As reported in Mora et al. (2002) and extended in Gelman et al. (2005), the formalization of the core concepts used in the IS discipline becomes a relevant and mandatory, as well as urgent, research purpose. This article furthers this purpose by utilizing the core principles from the Theory of Systems and a recent IS research framework (Mora et al., 2006) to extend and update the conceptualizations reported in previous studies. Formal definitions are updated and built upon the terms *system* (Ackoff, 1971; Gelman & Garcia, 1989), *organization, business process,* and *information system* (Mora et al., 2002; Gelman et al., 2005). Finally, this article examines the implications for IS research and practice.

BACKGROUND

The term *information system* has been defined in textbooks and research papers usually in non-formal terms. Table 1 shows a sample of the main definitions posed in the literature. An examination of these definitions suggests that the IS no-

Table 1. A sample of informal definitions of what an information system is

Definition	Reference
An IS is a system composed of subsystems of hardware, programs, files and procedures to get a shared goal.	Senn (1989, p. 23)
An IS is a system composed of application software, support software, hardware, documents and training materials, controls, job roles and people that uses the software application.	Hoffer, George, and Valachi (1996, p. 8)
An IS is a system composed of inputs, models, outputs, technology, data bases and controls.	Burch and Grudnitski (1989, p. 58)
A complete information system is a collection of subsystems defined by functional or organizational boundaries.	Ives et al. (1980, p. 910)
MIS is an integrated man-machine system for providing information to support the operation, management, and decision-making functions in an organization. The system utilizes computer software and software, manual procedures, management and decision models and a database.	Nolan and Wetherbe (1980, p. 3, quoting Davis, 1974)
We conceptualize the IT artifact as the application of IT to enable or support some task(s) embedded within a structure(s) that itself is embedded within a context(s).	Benbazat and Zmud (2003, p. 186)

tion: (i) lacks fundamental standardized and formal concepts (Alter, 2001); (ii) lacks competitive formal macro-structures to cumulate theories (Farohoomand & Drury, 2001, p. 14), and (iii) has an excessive variety of micro-theories (Barkhi & Sheetz, 2001).

There have been few, if any, efforts to formalize the the core concepts of IS. Despite attempts to reduce ambiguity, which have increased consequently the quality of the definitions, these proposals (Wand & Weber, 1990; Alter, 2001) have relied on partial views—for example, syntactical and structural perspectives that hide core semantic information—of the concept *system* formulated in the Systems Science literature (Sachs, 1976; Mora et al., 2002). Another study (Mentzas, 1994) offers a more articulated definition than exhibited in Table 1, by the identification of five subsystems and their functional properties. Nevertheless the resulting definition still lacks formalization and is based on a common-sense language that has been critiqued in the IS literature (Banville & Landry, 1989). Therefore, the concept *information system* still has multiple meanings. A systems-based research stream (Paton, 1997; Alter, 2001; Mora et al., 2002) combined with an ontological perspective (Wand & Weber, 1990) suggest that formal foundations from the Theory of Systems (Xu, 2000, p. 113) can reduce this ambiguity and strengthen the rigor that a scientific discipline requires to mature and simultaneously be relevant and useful for practitioners.

THE FORMALIZATION OF THE CONSTRUCT *INFORMATION SYSTEMS*

Formalization reported in this article is adapted and extended from previous definitions of the formal concepts of *system-I, system-II,* and *general system* (Gelman & Garcia, 1989). In turn, the concepts *organization* ***O(X)****, Information System* ***IS(X),*** and *envelop* ***EE(X)*** are updated from Mora et al. (2002), and the original concept of *environment* ***W(X)*** is replaced by the French term *entourage* ***ENT(X).*** To complete this set of formal definitions, Mora et al. (2006) also introduce the following concepts: *high-level business process* ***HLBP(X)****, low-level business process* ***LLBP(X)****, socio-political business process* ***SSBP(X)****, supra-suprasystem* ***SSS(X)****, non-entourage* ***NENT(X),*** and *world* ***W(X).*** Updates are mainly based on ideas reported by Oliva and Lane (1998) on soft systems and originally developed by Checkland (2000). As in similar works from the authors and related literature (Wand & Weber, 1990; Wand & Woo, 1991), we follow a conceptual development based on an ontological path to define primitive concepts and postulates to derive the set of updated and new definitions.

Formal Definition 1

System-I: An object of study **X,** formalized as ***system-I*** and denoted as $\boldsymbol{S_I(X)}$**=<B(X),RB(X),E(X)>,** is a whole **X** that fulfills the following conditions: (I.1) it has a *conceptual structure* **§(X)** that defines its set of *attributes* **B(X**), its set of *events* **E(X),** and its set of *range of attributes* **RB(X)**; (I.2) for any subset **B'(X)** of *attributes* of **B(X),** the set of *events* **E(X)** associated with **B(X)** differs in at least one element from the set of *events* **E'(X)** associated with **B'(X).**

Therefore, to define a situation of study as a *system-I* implies to specify $\boldsymbol{S_I(X)}$ **= <§(X)> =<B(X), E(X), RB(X)>** and to fulfill condition I.2.

Formal Definition 2

System-II: An object of study **X,** formalized as ***system-II*** and denoted as $\boldsymbol{S_{II}(X)}$**=<**$\mathbf{C_X}$**,** $\boldsymbol{\Re_s(C_X')}$**>** is a whole **X** that fulfills

the following conditions: (II.1) the whole **X** is a set $\mathbf{C_X}$ of elements $\mathbf{X_1, X_2, ..., X_k}$, called *subsystems*, where each $\mathbf{X_i}$ for i=1,2,..., k can be formalized as $\boldsymbol{S_I(X_i)}$ or $\boldsymbol{S_{II}(X_i)}$; (II.2) there is a collection finite $\mathbf{\Re_S(C_X')}$ of *set-relations* where $\mathbf{\Re_S(C_X')=\{\Re_1(C_X'), \Re_2(C_X'),...\}}$ on the set $\mathbf{C_X'=\{C, S_I(X)\}}$ and where each *set-relation* $\mathbf{\Re_p(C_X')=\{ {}_p\Re_1 , {}_p\Re_2 , ... \mid {}_p\Re_n = < X_i, {}_ia_j , X_j > \text{ or } {}_p\Re_n = < X_i, {}_ia_X , S_I(X)> \text{ or } {}_p\Re_n = < S_I(X), {}_Xa_j , S_j>}$ and $\mathbf{{}_ia_j}$ stands by the output-input parameters or acts between the two elements}; and (II.3) exists at least a *non-directed-path* among two any items X_i and X_j in the *set-relation* $\mathbf{\Re_S(C_X')}$.[1]

Formal Definition 3

General System: An object of study **X,** formalized as ***general system*** and denoted as $\boldsymbol{S_G(X)}$, is a whole **X** that can be defined simultaneously as a *system-I* $\boldsymbol{S_I(X)}$ and as a *system-II* $\boldsymbol{S_{II}(X)}$.

Postulate 1: Any *general system* $\boldsymbol{S_G(X)}$ defined as *system-I* $\boldsymbol{S_I(X)}$ can be mapped onto a *system-II* $\boldsymbol{S_{II}(X)}$ and vice versa.

Auxiliary Formal Definition 1

Universe: In a general sense a whole **UX** is called the ***universe of a system*** **X** and is denoted as ***U(X)*** if (III.1) **UX** can be formalized as $\boldsymbol{S_G(UX)}$ and (III.2) the whole **X** is a *subsystem* of **UX.**[2]

Auxiliary Formal Definition 2

World: In a general sense a whole **WX** is called the ***world of a system X*** and is denoted as ***W(X)***, if (IV.1) **WX** can be formalized as $\boldsymbol{S_G(WX)}$ and (IV.2) given the $\mathbf{S_{II}(U(X)) = <C_{U(X)}, \Re_S(C_{U(X)}')>}$ then $\mathbf{C_{U(X)} = \{X, W(X)\}}$.

Auxiliary Formal Definition 3

Suprasystem: In a particular sense a whole **SX** is called the ***suprasystem of a system*** **X** and is denoted as ***SS(X),*** if (V.1) **SX** can be formalized as $\boldsymbol{S_G(SX)}$ and (V.2) the whole **X** is a *subsystem* of **SX.**[3]

Auxiliary Formal Definition 4

Entourage: In a particular sense a whole **EX** is called the ***entourage of a system X*** and is denoted as ***ENT(X)***, if (VI.1) **EX** can be formalized as $\boldsymbol{S_G(EX)}$ and (VI.2) given the $\mathbf{S_{II}(SS(X)) = <C_{SS(X)}, \Re_S(C_{SS(X)}')>}$ then $\mathbf{C_{SS(X)} = \{X, ENT(X)\}}$.[4]

Auxiliary Formal Definition 5a

Non-Entourage: In a particular sense a whole **NEX** is called the ***non-entourage of a system X*** and is denoted as ***NENT(X)***, if (VIIa.1) **NEX** can be formalized as $\boldsymbol{S_G(NEX)}$ and (VIIa.2) **NEX** is the *world* of ***SS(X)***.[5]

Auxiliary Formal Definition 5b

Environment[6]*:* In a general sense a whole **ENVX** is called the ***environment of a system X*** and is denoted as ***ENV(X)***, if (VIIb.1) **ENVX** can be formalized as $\boldsymbol{S_G(ENVX)}$ and (VIIb.2) **ENVX** is the *world* of ***SS(X)***.[7]

Auxiliary Formal Definition 6

Supra-Suprasystem: In a particular sense a whole **SSX** is called the ***supra-suprasystem of a system*** **X** and is denoted as ***SSS(X),*** if (VIII.1) **SSX** can be formalized as $\boldsymbol{S_G(SSX)}$ and (VIII.2) there exists a whole S**X** that is a *subsystem* of ***SSS(X)*** and X is in turn is a *subsystem* of SX.

Auxiliary Formal Definition 7

Envelope: A whole **EEX** is called the ***envelope of a system X*** and is denoted as ***EE(X)***, if (IX.1) **EEX** can be formalized as $\boldsymbol{S_G(EEX)}$ and (IX.2) the whole **EEX** is the *entourage* of the *suprasystem* of **X.**

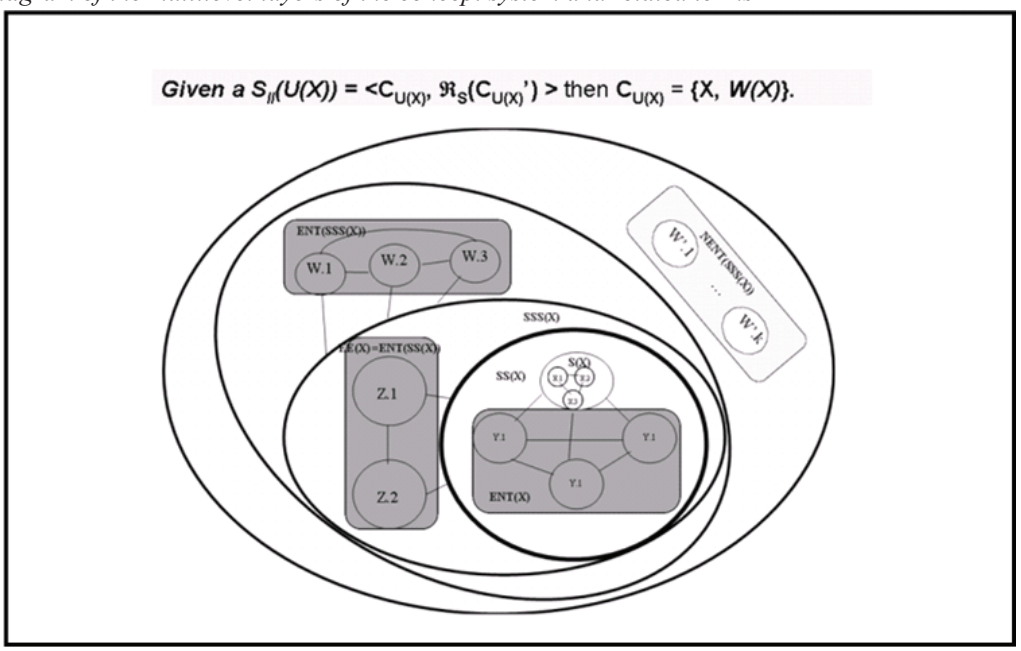
Postulate 2: Any *general-system* $\boldsymbol{S_G(X)}$ has a *world* **W(X)** and a *universe* ***U(X)***.

Hence, the first formal definition of the concept *system, system-I,* forms an external view that sees the *system* as a single unit with special characteristics called *attributes* and potential acts to execute called *events*. In turn, the second formal definition, *system-II,* represents the internal view that sees the system as a graph. Furthermore, the definitions of the *set-relations* $\Re_{S1}(C_X')$, $\Re_{S2}(C_X')$, ..., $\Re_{Sm}(C_X')$ consider the *system* as a multi-digraph, instead of a digraph, and therefore eliminates some limitations of classic digraph-like definitions critiqued in the Systems Theory literature (Sachs, 1976). Auxiliary definitions and the second postulate help to support the expansionist systemic perspective that indicates every system always belongs to another larger system (Ackoff, 1971). Figure 1 exhibits a general graphic interpretation of such formal definitions.

Formal Definition 4

Socio-Political Business Process as *System:* An object of study **X** is called a *socio-political* ***business process*** and denoted as **SSBP(X)**, if it satisfies the following conditions: (X.1) **X** can be defined as a *system-II* $\mathbf{S_{II}(X)=<C_X, \Re_S(C_X')>}$, where $\mathbf{C_X =\{S_G(Soc\text{-}SS), S_G(Pol\text{-}SS)\}}$ and $\mathbf{S_G(Soc\text{-}SS), S_G(Pol\text{-}SS)}$ are respectively called the social and the political subsystems; and $\mathbf{\Re_S(C_X')=\Re_S(X, S_I(X))=\{\Re_1(C_X'), \Re_2(C_X'), ..., \Re_6(C_X') \mid \Re_1(C_X')=\{ {}_1\Re_1 = < X_{Soc\text{-}SS}, {}_{Soc\text{-}SS}a_{Pol\text{-}SS}, X_{Pol\text{-}SS}>\}, ..., \Re_6(C_X')=\{ {}_6\Re_1 = <X_{Pol\text{-}SS}, {}_{Pol\text{-}SS}a_X, S_I(X)> \} \}}$; and (X.2) $\mathbf{\Re_S(X, S_I(X))}$ satisfies condition (II.3).

Figure 1. A diagram of the multilevel layers of the concept system and related terms



Formal Definition 5

Low-Level Business Process as *System:* An object of study **X** is called a ***low-level business process*** and denoted as **LLBP(X),** if it satisfies the following conditions: (XI.1) **X** can be defined as a *system-II* $\mathbf{S_{II}(X)}=<\mathbf{C_X}, \Re_S(\mathbf{C_X'})>$ where $\mathbf{C_X}$ **=**$\{\mathbf{S_G(T\text{-}SS)}, \mathbf{S_G(P\text{-}SS)}, \mathbf{S_G(T\&I\text{-}SS)}, \mathbf{S_G(M\&P\text{-}SS)}, \mathbf{SSBP(LSP\text{-}SS)}\}$ and $\mathbf{S_G(T\text{-}SS)}, \mathbf{S_G(P\text{-}SS)}, \mathbf{S_G(T\&I\text{-}SS)}, \mathbf{S_G(M\&P\text{-}SS)}, \mathbf{SSBP(LSP\text{-}SS)}$ are respectively called the ***task***, ***people***, ***tools&infrastructrure***, ***methods&procedures,*** and ***socio-political subsystems*****;** and $\Re_S(\mathbf{C_X'})= \Re_S(\mathbf{X, S_I(X)})$ $=\{\Re_1(\mathbf{C_X'}), \Re_2(\mathbf{C_X'}), \ldots, \Re_{24}(\mathbf{C_X'}) \mid \Re_1(\mathbf{C_X'})=\{{}_1\Re_1 = <\mathbf{X_{T\text{-}SS}}, {}_{\mathbf{T\text{-}SS}}\mathbf{a_{P\text{-}SS}}, \mathbf{X_{P\text{-}SS}}>\}, \ldots, \Re_{24}(\mathbf{C_X'})=\{{}_{24}\Re_1=<\mathbf{X_{P\text{-}SS}}, {}_{\mathbf{P\text{-}SS}}\mathbf{a_X}, \mathbf{S_I(X)}>\}\}$**;** and (XI.2) $\Re_S(\mathbf{X, S_I(X)})$ satisfy condition (II.3).

Formal Definition 6

High-Level Business Process as *System:* An object of study **X** is called a ***high-level business process*** and denoted as **HLBP(X)**, if it satisfies the following conditions: (XII.1) **X** can be defined as a *system-II* $\mathbf{S_{II}(X)}= <\mathbf{C_X}, \Re_S(\mathbf{C_X'})>$ where $\mathbf{C_X}$ **=**$\{\mathbf{LLBP(C\text{-}SS)}, \mathbf{LLBP(O\text{-}SS)}, \mathbf{LLBP(I\text{-}SS)}, \mathbf{SSBP(HSP\text{-}SS)}\}$ and **LLBP(C-SS), LLBP(O-SS), LLBP(I-SS), SSBP(HSP-SS)** are respectively called the ***control***, ***operational***, ***informational,*** and ***socio-political subsystems*****;** and $\Re_S(\mathbf{C_X'})= \Re_S(\mathbf{X, S_I(X)}) =\{\Re_1(\mathbf{C_X'}), \Re_2(\mathbf{C_X'}), \ldots, \Re_{20}(\mathbf{C_X'}) \mid \Re_1(\mathbf{C_X'})=\{{}_1\Re_1 = <\mathbf{X_{C\text{-}SS}}, {}_{\mathbf{C\text{-}SS}}\mathbf{a_{O\text{-}SS}}, \mathbf{X_{O\text{-}SS}}>\}, \ldots, \Re_1(\mathbf{C_X'})=\{{}_{20}\Re_1 = <\mathbf{X_{C\text{-}SS}}, {}_{\mathbf{C\text{-}SS}}\mathbf{a_X}, \mathbf{S_I(X)}>\}\}$**;** (XII.2) and $\Re_S(\mathbf{X, S_I(X)})$ satisfy condition (II.3).

Formal Definition 7

Organization as a *System:* An object of study **X** is called an ***organization*** and denoted as **O(X)**, if it satisfies the following conditions: (XIII.1) **X** can be defined as a *system-II* $\mathbf{S_{II}(X)}= <\mathbf{C_X}, \Re_S(\mathbf{C_X'})>$ where $\mathbf{C_X}$ **=**$\{\mathbf{S_G(X.1)}, \mathbf{S_G(X.2)}, \ldots \mathbf{S_G(X.k)}\}$ and $\Re_S(\mathbf{C_X'})= \Re_S(\mathbf{X, S_I(X)})= \{\Re_1(\mathbf{C_X'}), \Re_2(\mathbf{C_X'}), \ldots, \Re_p(\mathbf{C_X'}) \mid$ for N=1,2,...p, each set-relation $\Re_N(\mathbf{C_X'})$ has item-relations ${}_N\Re_{1,N}\Re_{2,N}\Re_{3,\ldots}$ of the format < X.i, ${}_i\mathbf{a}_j$, X.j> or < X.i, ${}_i\mathbf{a}_X$, $S_I(X)$> or < $S_I(X)$, ${}_X\mathbf{a}_j$, X_j> and ${}_\mathbf{i}\mathbf{a}_\mathbf{j}$ stands by the output-input parameters or acts between the two elements X.i and X.j}; (XIII.2) $\Re_S(\mathbf{X, S_I(X)})$ satisfies condition (II.3); and (XIII.3) for j=1,2, … , k either $\mathbf{S_G(X.j)= HLBP(X.j)}$**, or** $\mathbf{S_G(X.j)= SSBP(X.j)}$ or $\mathbf{S_G(X.j)= S_{II}(X.j)}=<\mathbf{C_{X.j}}, \Re_S(\mathbf{C_{X.j}'})>$ where $\mathbf{C_{X.j}}$ **=**$\{\mathbf{BP(X.j.1)}, \mathbf{BP(X.j.2)}, \ldots, \mathbf{BP(X.j.n)}\}$ and $\Re_S(\mathbf{C_{X.j}'})$ exists.

Several features should be noted from previous definitions: (a) the *item-relations* $\Re_1$, $\Re_{2,\ldots,a}$ and so on take into account at least all possible interrelationships between any two subsystems and the whole system; (b) a *high-level business process* can be considered a *general system* with three low-level business processes: control, operational, informational, and a socio-political business process; and (c) an *organization* is a *general system* composed of at least two *general systems* that in turn can be either a *high-level business process* or a *system-II* composed of *high-level business processes.* Figure 2, adapted from Mora et al. (2006), exhibits a diagram of an *organization* ***O(X)***.[8]

An extended cybernetic-based paradigm (Gelman & Negroe, 1981, 1982) is used in Figure 2, where ***S_{II}(X.1)*** and

$S_{II}(X.2)$ are conceptualized as a driving-org-subsystem and a driven-org-subsystem respectively, $S_{II}(X.3) = HLBP(X.3)$ for an *information-org-subsystem,* and $S_{II}(X.4) = SSBP(X.4)$ for a *socio-political-org-subsystem.* Interactions between subsystems are not shown. However, it must be noted that general definitions of what is an *organization* can adopt other management perspectives such as Porter's added-value chain (Porter & Millan, 1985). In that case, it would be necessary to consider three *subsystems* of the *primary activities,* the support activities, and the top management activities. With these previous antecedents, the formalization of the term *information system* is straightforward.

Formal Definition 8

Information System as *System:* An object of study **X** is called an ***Information System*** and is denoted as ***IS(X)*** if it fulfills either one of the following conditions: (XIV.1) ***IS(X)*** is the ***HLBP(X.3)*** of an ***O(X)*** or (XIV.2) ***IS(X)*** is the ***LLBP(I-SS)*** of a ***HLBP(X.i)*** and that belongs to a ***O(X).***

Figure 2 also presents a visualization of the formal concept of *information system.* Due to the diagram's style, all notation related to the *item-relations* of all *systems-II* are hidden as are *item-relations* into the *subsystems* of the *information system.* The posed formal definition indicates that an *information system,* as a *system,* can correspond either to the general organizational function or area of *information systems* (e.g., ***IS(X)*** is the ***HLBP(X.3)***) or to the specific view of an *information system* as part of any business process (e.g., ***IS(X)*** is the ***LLBP(I-SS)*** of a ***HLBP(X.i)***). It is congruent with the systemic perspective that any *information system* is a *system* that is part of a larger system—that is, a *high-level business process* that in turn is the *suprasystem* of the *information system,* denoted as ***SS(IS)***.

FUTURE TRENDS

From a theoretical perspective, the development and utilization of formal definitions help to advance standardization, theoretical soundness and completeness, and academic maturation. Ambiguous, informal, and non-standard concepts of the main objects of study or the main phenomena studied by the discipline obfuscate its development and diminish its theoretical accumulation. Previous frameworks developed by the authors and this update constitute a conceptual tool to model and specify an *information system* at the level of detail demanded by researchers and practitioners. In Gelman et al. (2005), four possible applications of the utilization of the formal definitions were presented under a theoretical and practice perspective: (i) formal integration of the common-sense concepts used in the informal definitions of what is an *information system*; (ii) elimination of the ambiguity caused by the "Siamese Twin Problem" (Alter, 2001, pp. 30-31); (iii) analysis of failures on information systems through the systemic classification of problems; and (iv) generation of conceptual and dynamic simulation models of information systems. In this updated article, based on recent research frameworks for the IS discipline (Mora et al., 2006), we have presented the usefulness of the concept. Figure 3 compares the previous systemic IS research frameworks with the proposal offered here. The four past issues are reviewed under this updated framework:

Figure 2. Diagram of the overall and systemic perspective of what are an organization and an information system

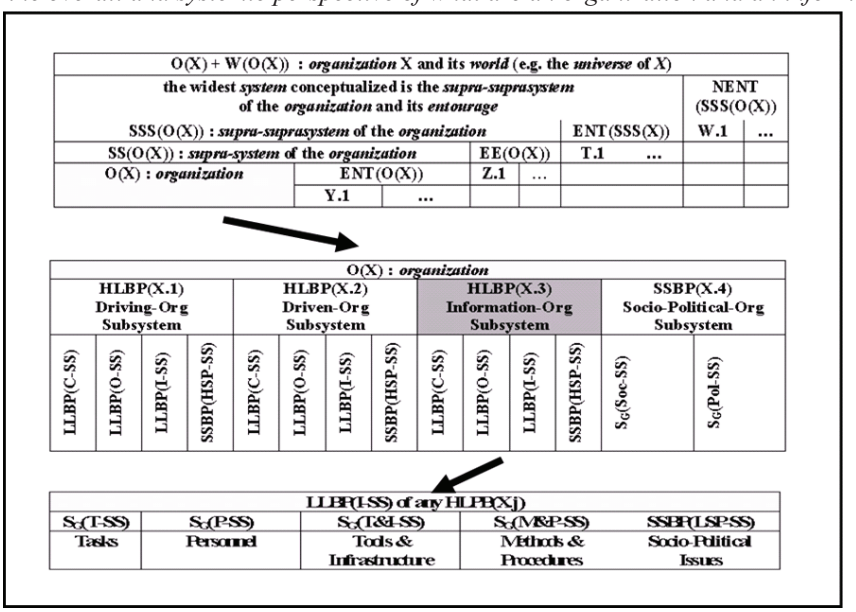


Figure 3. Comparison map of the systemic concepts for IS research frameworks

Mapping of Concepts		
Ives' et al Framework	**Nolan and Wetherbe's Framework**	**Mora et al's Framework**
<external environment> : legal social, cultural, economic, educational, resource and industry/trade systems	**Not explicitly considered**	**EE(O(X))** (the systems to be modeled at this level of analysis are determined by the researcher)
Not explicitly considered	**<environment of the organization>** : competitors, government, suppliers, customers, etc.	**ENT(O(X))** (the systems to be modeled at this level of analysis are determined by the researcher)
The concept of **<organizational environment>** used by Ives et al, really does not considers the environment of an organization, but organization's attributes per se such as: goals, tasks, structure, volatility and management philosophy/style	**<organization>** The 5 subsystems of <organization>: goals and value SS, psychosocial SS, managerial SS, structural SS and technical SS	**O(X)** **Goals & Value SS, psychosocial SS, and structural SS,** are considered in the **SSBP(X.4)** and the **SSBP(HSP-SS)** for any **HLBP(X.j).** The same applies for the organization's attributes posited by Ives' et al such as: goals, tasks, volatility and management philosophy/style.
Not explicitly considered		The **managerial** and **technical SS** respectively corresponds to the **HLBP(X.1) (e.g. the driving-org subsystem)** and to **HLBP(X.2) (e.g. the driven-org subsystem)**
<user environment> + <use process>	**Not explicitly considered but implicit in the 5 subsystems above reported**	Any **HLBP(X.j)** and **SSBP(X.4)** of the **O(X)**
<IS development environment> + <IS development process>	**<MIS Technology> : (hardware, software, data base, procedures and personnel)** (Development, operations and maintenance aspects are considered in this concept)	**HLBP(X.3)** and **SSBP(X.4)** (e.g. **the high-level business process** that corresponds to the **information-org subsystem** and the **socio-political general subsystem** of the ***organization*** **O(X)**)
<IS operations environment> + <IS operations process>		
<Information Subsystem> (attributes of content, presentation, time, etc)	**Not explicitly considered as system but as the output of the MIS technology-based system**	**LLBP(I-SS)** of any **HLBP(X.j)** and that is conformed by the **SG(T-SS) +SG(P-SS) + SG(T&I-SS) + SG(M&P-SS) + SSBP(LSP-SS)** (e.g. **the low-level business process** that corresponds to the **informational subsystem** of any **high-level business process** in an **O(X)**)

(1) The informal definition from Hoffer et al. (1996) establishes that "an IS is a system composed of application software, support software, hardware, documents and training materials, controls, job roles and people that use the software application." This notion can be conceptually mapped to the proposed framework (and consequently to the formal definitions) under the second interpretation of what is an information system (e.g., condition (XIV.2) of the formal definition 8: ***IS(X)*** is the ***LLBP(I-SS)*** of a ***HLBP(X.i)*** and that belongs to a ***O(X)***). Given this condition and condition (XI.1) of the formal definition 5 (e.g., **X** can be defined as a *system-II* $\mathbf{S_{II}(X)}$= $\langle \mathbf{C_X}, \Re_S(\mathbf{C_X}')\rangle$ where $\mathbf{S_{II}(X)}$= $\langle \mathbf{C_X}, \Re_S(\mathbf{C_X}')\rangle$ where $\mathbf{C_X}$ ={$\mathbf{S_G(T\text{-}SS)}$, $\mathbf{S_G(P\text{-}SS)}$, $\mathbf{S_G(T\&I\text{-}SS)}$, $\mathbf{S_G(M\&P\text{-}SS)}$, **SSBP(LSP-SS)**} and $\mathbf{S_G(T\text{-}SS)}$, $\mathbf{S_G(P\text{-}SS)}$, $\mathbf{S_G(T\&I\text{-}SS)}$, $\mathbf{S_G(M\&P\text{-}SS)}$, **SSBP(LSP-SS)** are the ***task***, ***people***, ***tools&infrastructrure***, ***methods&procedures,*** and ***socio-political subsystems, respectively***), then the elements <application software, support software, hardware> correspond to the $\mathbf{S_G(T\&I\text{-}SS)}$, and <documents and training materials > and <people> correspond to $\mathbf{S_G(M\&P\text{-}SS)}$ and $\mathbf{S_G(P\text{-}SS)}$ respectively. The remainder <job roles> element can be modeled as the system's attribute $\boldsymbol{B(P\text{-}SS)}$=$\{_{IW}\beta_1$=<job roles>} and specifies all possible <job roles> by the definition of the range of attributes $\boldsymbol{RB(P\text{-}SS)}$=$\{_{IW}\mathbf{R}\beta_1$={operational-user; staff-user; executive-user; etc.}}, if $\mathbf{S_G(P\text{-}SS)}$ is further formalized as a $\mathbf{S_I(X)}$.

(2) As reported in Gelman et al. (2005), the systemic properties of *organization* and *hierarchy* mapped onto the concept of *information system* are useful to resolve the "Siamese Twin Problem" (Alter, 2001) that indicates the

fallacy of studying an *information system* without studying its w*ork system.* The *hierarchy* property of *systems* and the *Expansionism Approach* suggest that any *system* should not be studied in isolation. A framework and formal definitions support scientific study and indicate that any *information system* should be viewed as part of a larger *system*—that is, the *high-level business process* (e.g., Alter's w*ork system* concept). A Synthetic Thinking view (Ackoff, 1960), that the behavior of every system can be best understood when studied in relationship to its *suprasystem* and wider systems, is also achieved.

(3) Formal definitions and the updated framework exhibited in Figures 1, 2, and 3 are simple but powerful conceptual lenses to study holistically complex problems as the failures of *information systems* implementations. According to the classification of problems based on the Theory of Systems by Ackoff (1973, p. 666), the failures of *systems* are caused by the conflicts between the system's purpose and the environment's purpose, the system's purpose and the subsystems' purposes, and the system with itself—for example, the environmentalization, humanization, and self-control problems. Thereby, the systemic formalisms posed can be used and extended through the addition of parameters in the set of *a-actions* in the *item-relations*, to model social, technical, and political features through the specification of *attributes* and *events* assigned to *subsystems* in the *system's entourage* ***ENT(IS(X)).*** These features either can be the *driving, driven,* and *socio-political subsystems* of the *organization* ***O(X)*** or the *control, operational,* and *socio-political* subsystems of any ***HLBP(X.j).*** For example, soft features, such as <top management support>, <environmental hostility>, <environmental dynamism>, and <organizational climate>, once specified and measured, can be easily assigned as subsystems *attributes* of this ***ENT(IS(X))***.

(4) Complex conceptual and dynamic-based simulations models of *information systems* can be specified with the integration of the hard (e.g., System Dynamics) and soft systems approach. A case in point is the recent development of a conceptual and system dynamic-based simulation model of the implementation process of decision-making support systems (Mora, 2003; Mora, Cervantes, Gelman, & Forgionne, 2004) that generated satisfactory predictive results for several well-known statistical and case-study reports in the literature.

(5) Finally, according to Mora et al. (2006), the framework exhibited in Figure 3 and rooted in the formal definitions of the concepts *system-I, system-II,* and *general system,* and the updated or new concepts (*suprasystem, supra-suprasystem, envelop, entourage, world, universe, low-level business process, high-level business process, organization,* and *information systems*) offer evidence that previous IS research frameworks were insufficiently comprehensive. The new framework:

(i) is congruent with formal definitions of system; (ii) permits the modeling of all variables reported as sub-systems or attributes of sub-systems; (iii) includes the time variable if required through the consideration of the state of the system; (iv) and integrates both technical and socio-political perspectives. (Mora et al., 2006, p. 8)

In summary, this article updates the formal definitions of the concepts of *systems* developed by some of authors in 1989, and offers adjustments to previous definitions elaborated in 2002 and 2005 and applied to the concept of *information system.* All definitions are based on, and congruent with, formal principles from the Theory of Systems. As reported in Gelman et al. (2005), these findings enable researchers and practitioners to:

(i) avoid ambiguity from informal definitions; (ii) account for practically all informal definitions; (iii) specify and customize a structure of the concept IS with the level of detail demanded by the modelers; and (iv) build complex systemic models of organizations that use IS.

CONCLUSION

This updated article supports the thesis that formalisms—from the Theory of Systems—are required in the field of information systems to reach the rigor and maturation needed for a respectable scientific discipline. Given that posed formalisms offer a non-trivial way to understand and model an information system, their utilization will contribute toward the maturation of our field. However, gains will not be free since researchers and practitioners could be required to add to their conceptual tools the utilization of these logical-mathematical models. This article also can be considered as a research basis for the emergent topic of formal ontologies for information systems. Nevertheless, we believe that future contributions based on computerized ontologies should be theoretically deepened on Theory of Systems.

REFERENCES

Ackoff, R. (1960). Systems, organizations and interdisciplinary research. *General System Yearbook, 5,* 1-8.

Ackoff, R. (1971). Towards a system of systems concepts. *Management Science, 17*(11), 661-671.

Ackoff, R. (1973). Science in the systems age: Beyond IE, OR and MS. *Operations Research, 21*(3), 661-671.

Adam, F., & Fitzgerald, B. (2000). The status of the information systems field: Historical perspective and practical orientation. *Information Research, 5*(4), 1-16.

Alter, S. (2001). Are the fundamental concepts of information systems mostly about work systems? *Communication of AIS, 5*(11), 1-67.

Alter, S. (2003). 18 Reasons why IT-reliant work systems should replace "the IT artifact" as the core subject matter of the IS field. *Communications of AIS, 12*(23), 366-395.

Banville, C., & Landry, M. (1989). Can the field of MIS be disciplined? *Communications of the ACM, 32*(1), 48-60.

Barkhi, R., & Sheetz, S. (2001). The state of theoretical diversity of information systems. *Communication of AIS, 7*(6), 1-19.

Benbazat, I., & Zmud, R. (2003). The crisis identity within the IS discipline: Defining and communicating the discipline's core properties. *MIS Quarterly, 27*(2), 187-194.

Burch, J.G., & Grudnitski, G. (1989). *Design of information systems.* New York: John Wiley & Sons.

Checkland, P. (2000). Soft systems methodology: A thirty year retrospective. *Systems Research and Behavioral Science, 17,* S11-S58.

Davis, G. (1974). *Management information systems: Conceptual foundations, structure and development.* New York: McGraw-Hill.

Farhoomand, A. (1987). Scientific progress of management information systems. *Database,* (Summer), 48-57.

Farhoomand, A., & Drury, D. (2001). Diversity and scientific progress in the information systems discipline. *Communication of AIS, 5*(12), 1-22.

Gelman, O., & Garcia, J. (1989). Formulation and axiomatization of the concept of general system. *Outlet IMPOS* (Mexican Institute of Planning and Systems Operation), *19*(92), 1-81.

Gelman, O., & Negroe, G. (1981). Role of the planning function in the organizational conduction process. *Outlet IMPOS* (Mexican Institute of Planning and Systems Operation), *11*(61), 1-17.

Gelman, O., & Negroe, G. (1982). Planning as organizational conduction process. *Journal of the Mexican National Academy of Engineering, 1*(4), 235-270.

Gelman, O., Mora, M., Forgionne, G., & Cervantes, F. (2005). Information systems and systems theory. In *Encyclopedia of Information Science and Technology* (pp. 1491-1496). Hershey, PA: Idea Group.

Hoffer, J., George, J., & Valachi, J. (1996). *Modern systems analysis and design.* Menlo Park, CA: Benjamin/Cummings.

Ives, B., Hamilton, S., & Davis, G. (1980). A framework for research in computer-based management information systems. *Management Science, 26*(9), 910-934.

Mentzas, G. (1994). Towards intelligent organizational information systems. *International Journal of Information Management, 14*(6), 397-410.

Mingers, J. (2001). Combining IS research methods: Towards a pluralist methodology. *Information Systems Research, 12*(3), 240-253.

Mora, M. (2003). *Theoretical foundations of the systems approach and its application to the study of the dynamic of the implementation process of decision making support systems.* Doctoral Dissertation, School of Engineering, National Autonomous University of Mexico, Mexico.

Mora, M., Cervantes, F., Gelman, O., & Forgionne, G. (2004, July 1-3). Understanding the strategic process of implementing decision making support systems (DMSS): A systems approach. *Proceedings of the 2004 IFIP International Conference on Decision Support Systems* (DSS2004), Prato, Italy (pp. 557-567).

Mora, M., Gelman, O., Cano, J., Cervantes, F., & Forgionne, G. (2006, July 9-14). Theory of Systems and information systems research frameworks. *Proceedings of the International Society for the Systems Sciences 50th Annual Conference,* Sonoma State University, CA (pp 282-1, 282-7).

Mora, M., Gelman, O., Cervantes, F., Mejia, M., & Weitzenfeld, A. (2002). A systemic approach for the formalization of the information system concept: Why information systems are systems. In J. Cano (Ed.), *Critical reflections of information systems: A systemic approach* (pp. 1-29). Hershey, PA: Idea Group.

Nolan, R., & Wetherbe, J. (1980). Toward a comprehensive framework for MIS research. *MIS Quarterly,* (June), 1-20.

Oliva, R., & Lane, D. (1998). The greater whole: Towards a synthesis of systems dynamics and soft systems methodology. *European Journal of Operational Research, 107,* 214-235.

Paton, G. (1997). Information system as intellectual construct—its only valid form. *Systems Research and Behavioral Science, 14*(1), 67-72.

Porter, M., & Millan, V. (1985). How information gives you competitive advantage. *Harvard Business Review,* (July-August), 149-174.

Sachs, W. (1976). Toward formal foundations of teleological systems science. *General Systems, XXI,* 145-154.

Senn, J. (1989). *Analysis and design of information systems.* New York: McGraw-Hill.

Xu, L. (2000). The contributions to systems science to information systems research. *Systems Research and Behavioral Science, 17,* 105-116.

Wand, Y., & Weber, R. (1990). An ontological model of an information system. *IEEE Transactions on Software Engineering, 16*(11), 1282-1292.

Wand, Y., & Woo, C. (1991). *An approach to formalizing organizational open systems concepts*, 141-146. Retrieved March 5, 2002, from the ACM Library Digital.

KEY TERMS

Attribute (Informal Definition): A substantial feature of a whole that is perceived by an observer with the potential to produce or cause a product or effect.

Event (Informal Definition): An act performed by a whole or to the whole that is perceived by an observer directly or through its consequences on other(s) whole(s).

Subsystem: Any immediate inner system that is subsumed to the system.

Suprasystem: The immediate outer system that subsumes any system.

System (Informal Definition): A whole composed of subsystems, and at the same time included in a suprasystem in such way that some particular properties of the whole and of the subsystems are lost when they are considered analytically—for example, by separation of the parts of the whole.

Theory of Systems: A research paradigm based on an expansionist world view, synthetic and holistic thought, and teleological principles that is suitable and effective for studying complex phenomena.

ENDNOTES

1 It must be noted that: (i) condition II.3 assures that for any two elements X_i and X_j in the multi-digraph of **X**, X_i is reachable from X_j and vice versa that implies the nonexistence of an isolated element X_i; (ii) it is a recursive definition that permits a *subsystem* to have *subsystems;* and (iii) this definition supports the output/input relationships between any *subsystem* and the whole *system.* Therefore, to define a situation of study as a *system-II* implies to specify: $\boldsymbol{S_{II}(X)=<C_X, \Re_S(C_X')>}$ where $\boldsymbol{C_X=\{S_I(X_I)}$ **or** $\boldsymbol{S_{II}(X_I)\}}$ for i = 1,2, ..., k; $\boldsymbol{\Re_S(C_X')=\{\Re_{S1}(C_X'), \Re_{S2}(C_X'), \ldots\}}$ and the fulfillment of the condition II.3.

2 Then, the universe of **X** includes the element.

3 Note that **U(X)** is the broadest system considered as a *suprasystem* of **X.**

4 **W(X)** is the broadest system considered as an entourage of **X.**

5 It must be noted that **NENT(X)** is the set of all elements that are not part of the system **X** nor its entourage **ENT(X)**.

6 Readers could note that *non-entourage* and *environment* represent the same concepts. However, both terms are used to enable the researcher or analyst to use the term better suited to the level of details used in the analysis of a system.

7 It must be noted that **ENV((X)** is the set of all elements that are not part of the suprasystem **X**.

8 Figure 2 also presents in advance the graphical interpretation of an *information system.* Both are reported in the same figure due to space limitations.

Formal Development of Reactive Agent-Based Systems

P. Kefalas
CITY College, Greece

M. Holcombe
University of Sheffield, UK

G. Eleftherakis
CITY College, Greece

M. Gheorghe
University of Sheffield, UK

INTRODUCTION

Recent advances in both the testing and verification of software based on formal specifications have reached a point where the ideas can be applied in a powerful way in the design of agent-based systems. The software engineering research has highlighted a number of important issues: the importance of the type of modelling technique used; the careful design of the model to enable powerful testing techniques to be used; the automated verification of the behavioural properties of the system; and the need to provide a mechanism for translating the formal models into executable software in a simple and transparent way.

An agent is an encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its design objectives (Jennings, 2000). There are two fundamental concepts associated with any dynamic or reactive system (Holcombe & Ipate, 1998): the environment, which could be precisely or ill-specified or even completely unknown and the agent that will be responding to environmental changes by changing its basic parameters and possibly affecting the environment as well. Agents, as highly dynamic systems, are concerned with three essential factors: a set of appropriate environmental stimuli or inputs, a set of internal states of the agent, and a rule that relates the two above and determines what the agent state will change to if a particular input arrives while the agent is in a particular state.

One of the challenges that emerges in intelligent agent engineering is to develop agent models and agent implementations that are "correct." The criteria for "correctness" are (Ipate & Holcombe, 1998): the initial agent model should match the requirements, the agent model should satisfy any necessary properties in order to meet its design objectives, and the implementation should pass all tests constructed using a complete functional test-generation method. All the above criteria are closely related to stages of agent system development, i.e., modelling, validation, verification, and testing.

BACKGROUND: FORMAL METHODS AND AGENT-BASED SYSTEMS

Although agent-oriented software engineering aims to manage the inherent complexity of software systems (Wooldridge & Ciancarini, 2001; Jennings, 2001), there is still no evidence to suggest that any methodology proposed leads toward "correct" systems. In the last few decades, there has been strong debate on whether formal methods can achieve this goal. Software system specification has centred on the use of models of data types, either functional or relational models, such as Z (Spivey, 1989) or VDM (Jones, 1990), or axiomatic ones, such as OBJ (Futatsugi et al., 1985). Although these have led to some considerable advances in software design, they lack the ability to express the dynamics of the system. Also, transforming an implicit formal description into an effective working system is not straightforward. Other formal methods, such as finite state machines (Wulf et al., 1981) or Petri Nets (Reisig, 1985) capture the essential feature, which is "change," but fail to describe the system completely, because there is little or no reference to the internal data and how these data are affected by each operation in the state transition diagram. Other methods, like statecharts (Harel 1987), capture the requirements of dynamic behaviour and modelling of data but are informal with respect to clarity and semantics. So far, little attention has been paid in formal methods that could facilitate all crucial stages of "correct" system development, modelling, verification, and testing.

In agent-oriented engineering, there have been several attempts to use formal methods, each one focusing on different aspects of agent systems development. One was to

formalise the PRS (procedural reasoning system), a variant of the BDI architecture (Rao & Georgeff, 1995), with the use of Z, in order to understand the architecture in a better way, to be able to move to the implementation through refinement of the specification, and to be able to develop proof theories for the architecture (D'Inverno et al., 1998). Trying to capture the dynamics of an agent system, Rosenschein and Kaebling (1995) viewed an agent as a situated automaton that generates a mapping from inputs to outputs, mediated by its internal state. Brazier et al. (1995) developed the DESIRE framework, which focuses on the specification of the dynamics of the reasoning and acting behaviour of multiagent systems. In an attempt to verify whether properties of agent models are true, work has been done on model checking of multiagent systems with reuse of existing technology and tools (Benerecetti et al., 1999, Rao & Georgeff, 1993). Toward implementation of agent systems, Attoui and Hasbani (1997) focused on program generation of reactive systems through a formal transformation process. A wider approach is taken by Fisher and Wooldridge (1997), who utilised Concurrent METATEM in order to formally specify multiagent systems and then directly execute the specification while verifying important temporal properties of the system. Finally, in a less formal approach, extensions to Unified Modelling Language (UML) to accommodate the distinctive requirements of agents (AUML) were proposed (Odell et al., 2000).

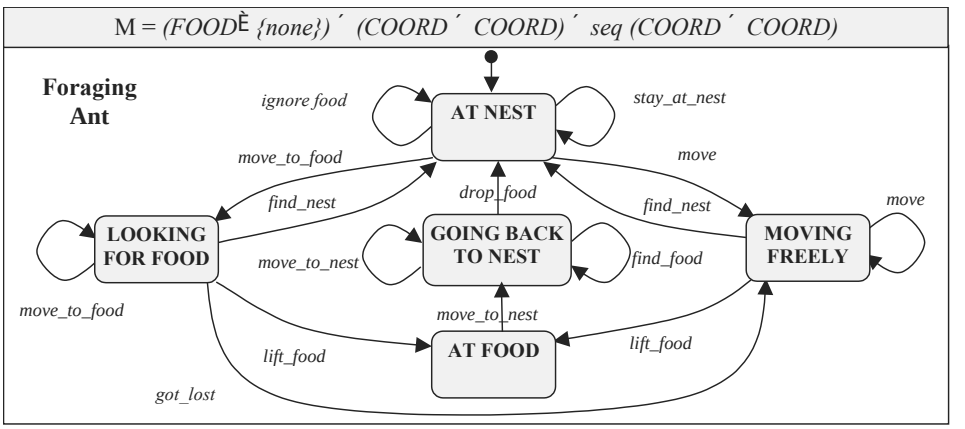
X-MACHINES FOR AGENT-BASED SYSTEM DEVELOPMENT

An X-machine is a general computational machine (Eilenberg, 1974) that resembles a finite state machine but with two significant differences: there is memory attached to the machine, and the transitions are labeled with functions that operate on inputs and memory values. The X-machine formal method forms the basis for a specification language with great potential value to software engineers, because they can facilitate modelling of agents that demand remembering as well as reactiveness. Figure 1 shows the model of an ant-like agent that searches for food but also remembers food positions in order to set up its next goals. Many other biological processes seem to behave like agents, as, for example, a colony of foraging bees, tissue cells, etc. (Kefalas et al., 2003a; Gheorghe et al., 2001; Kefalas et al., 2003b). Formally, the definition of the X-machine requires the complete description of a set of inputs, outputs, and states; a memory tuple with typed elements; a set of functions and transitions; and finally, an initial state and a memory value (Holcombe, 1988).

Having constructed a model of an agent as an X-machine, it is possible to apply existing model-checking techniques to verify its properties. *CTL** is extended with memory quantifier operators: M_x (for all memory instances) and m_x (there exist memory instances) (Eleftherakis & Kefalas, 2001). For example, in the ant-like agent, model checking can verify whether food will eventually be dropped in the nest by the formula: $AG[\neg M_x(m_1 \neq \text{none}) \vee EFM_x(m_1 = \text{none})]$, where m_1 indicates the first element of the memory tuple.

Having ensured that the model is "correct," we need to also ensure that the implementation is "correct," this time with respect to the model. Holcombe and Ipate (1998) presented a testing method that under certain design-for-test conditions can provide a complete test-case set for the implementation. The testing process can be performed automatically by checking whether the output sequences produced by the implementation are identical to the ones expected from the agent model through this test-case set.

Figure 1. An X-machine that models an ant.



A methodology for building complex agent systems by aggregating a set of behaviors of individual agents is available, namely, communicating X-machines. It is demonstrated that they are a powerful extension to the X-machines that also facilitate modelling of multiagent systems (Kefalas et al., 2003a).

FUTURE TRENDS

There are currently three directions to future work. First, there is a need to investigate potential applications. We already identified the area of biology-inspired agent-based systems, such as modelling of biological cells and tissues, as an area that can largely benefit from formal modelling, verification, and testing. Second, current research is underway to extend the testing and verification methods in order to be applicable for communicating asynchronous and possibly nondeterministic systems. Last, there has been an attempt to build tools around the X-machine formal method in order to facilitate the actual agent development process. A markup language, namely X-Machine Definition Language (XMDL), has been defined, and around it, a number of prototype tools, such as a modeller, an automatic translator to Prolog, an animator, and a model checker, have been constructed.

CONCLUSION

Because the X-machine method is fully grounded in the theory of computation, it is fully general and will be applicable to any type of computational task. The paradigm of the X-machine is also convenient when it comes to implementing the models in an imperative programming language. In fact, the translation is more or less automatic. The existence of the powerful testing method described lays the foundation for the method to be used in potentially critical applications. Finally, the model-checking developments will lead to a situation in which one of the key issues in agent software engineering can be solved, namely, how can we guarantee that the agent system constructed will exhibit the desired emergent behavior, or at least substantial progress toward this goal will be achieved.

REFERENCES

Attoui, A., & Hasbani, A. (1997). Reactive systems developing by formal specification transformations. In *Proceedings of the Eighth International Workshop on Database and Expert Systems Applications (DEXA '97)*, (pp.339-344). Washington, DC: IEEE Computer Society.

Benerecetti, M., Giunchiglia, F., & Serafini, L. (1999). A model checking algorithm for multiagent systems. In J. P. Muller, M. P. Singh, & A. S. Rao (Eds.), *Intelligent Agents V* (LNAI Vol. 1555) (pp.163-176). Heidelberg: Springer-Verlag.

Brazier, F., Dunin-Keplicz, B., Jennings, N., & Treur, J. (1995). Formal specification of multi-agent systems: A real-world case. In V. Lesser (Ed.), *Proceedings of International Conference on Multi-Agent Systems (ICMAS'95)* (pp. 25-32), San Francisco, CA, June 12-14. Cambridge, MA: MIT Press.

Eilenberg, S. (1974). *Automata, machines and languages* (Vol. A). New York: Academic Press.

Eleftherakis, G., & Kefalas, P. (2001). Towards model checking of finite state machines extended with memory through refinement. In G. Antoniou, N. Mastorakis, & O. Panfilov (Eds.), *Advances in signal processing and computer technologies* (pp. 321-326). Singapore: World Scientific and Engineering Society Press.

Fisher, M., & Wooldridge, M. (1997). On the formal specification and verification of multi-agent systems. *International Journal of Cooperating Information Systems, 6*(1), 37-65.

Futatsugi, K., Goguen, J., Jouannaud, J. -P., & Meseguer, J. (1985). Principles of OBJ2. In B. Reid (Ed.), *Proceedings, Twelfth ACM Symposium on Principles of Programming Languages* (pp. 52-66). Association for Computing Machinery.

Georghe, M., Holcombe, M., & Kefalas, P. (2001). Computational models for collective foraging. *Biosystems, 61*, 133-141.

Harel, D. (1987). Statecharts: A visual approach to complex systems. *Science of Computer Programming, 8*(3).

Holcombe, M. (1988). X-machines as a basis for dynamic system specification. *Software Engineering Journal, 3*(2), 69-76.

Holcombe, M., & Ipate, F. (1998). *Correct systems: Building a business process solution*. Heidelberg: Springer-Verlag.

Inverno, d'M., Kinny, D., Luck, M., & Wooldridge, M. (1998). A formal specification of dMARS. In M. P. Singh, A. Rao, & M. J. Wooldridge (Eds.), *Intelligent Agents IV* (LNAI Vol. 1365) (pp. 155-176). Heidelberg: Springer-Verlag.

Ipate, F., & Holcombe, M. (1998). Specification and testing using generalised machines: A presentation and a case study. *Software Testing, Verification and Reliability, 8*, 61-81.

Jennings, N. R. (2000). On agent-based software engineering. *Artificial Intelligence, 117*, 277-296.

Jennings, N. R. (2001). An agent-based approach for building complex software systems. *Communications of the ACM, 44*(4), 35-41.

Jones, C. B. (1990). Systematic software development using VDM (2nd ed.). New York: Prentice Hall.

Kefalas, P., Eleftherakis, G., & Kehris, E. (2003a). Communicating X-Machines: From theory to practice. In Y. Manolopoulos, S. Evripidou, & A. Kakas (Eds.), *Lecture notes in computer science* (Vol. 2563) (pp. 316-335). Heidelberg: Springer-Verlag.

Kefalas, P., Eleftherakis, G., Holcombe, M., & Gheorghe, M. (2003b). Simulation and verification of P systems through communicating X-machines. *BioSystems, 70*(2), 135-148.

Odell, J., Parunak, H.V.D., & Bauer, B. (2000). Extending UML for agents. In G. Wagner, Y. Lesperance, & E. Yu (Eds.), *Proceedings of the Agent-Oriented Information Systems Workshop at the 17th National Conference on Artificial Intelligence*, (pp. 3-17), Austin, TX.

Rao, A.S., & Georgeff, M. P. (1993). A model-theoretic approach to the verification of situated reasoning systems. In R. Bajcsy (Ed.), *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI'93)* (pp. 318–324). San Francisco, CA: Morgan Kaufmann.

Rao, A.S., & Georgeff, M. (1995). BDI agents: From theory to practice. In V. Lesser & L. Gasser (Eds.), *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS'95)* (pp. 312-319), MIT Press.

Reisig, W. (1985). Petri nets—An introduction. In *EATCS Monographs on Theoretical Computer Science, Vol. 4*. Heidelberg: Springer-Verlag.

Rosenschein, S. R., & Kaebling, L. P. (1995). A situated view of representation and control. *Artificial Intelligence, 73*(1-2), 149-173.

Spivey, M. (1989). *The Z notation: A reference manual*. New York: Prentice Hall.

Wooldridge, M., & Ciancarini, P. (2001). Agent-oriented software engineering: The state of the art. To appear in the S.K. Chang, *Handbook of Software Engineering and Knowledge Engineering*. Singapore: World Scientific Publishing.

Wulf, W. A., Shaw, M., Hilfinger, P. N., & Flon, L. (1981). *Fundamental structures of computer science*. Reading, MA: Addison-Wesley.

KEY TERMS

Agent: An agent is an encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its design objectives. Agents normally exhibit autonomous, reactive, proactive, and social behaviors.

Communicating X-Machine: A communicating X-machine is a set of stream X-machine components that are able to communicate with each other by exchanging messages.

CTL*: A temporal logic formalism used in model checking. *CTL** employs operators, such as *A* (for all paths), *E* (there exists a path), *X* (next time), *F* (eventually), *G* (always), *U* (until), and *R* (release), that facilitate the construction of temporal logic formulas that correspond to desirable properties of a model.

Formal Methods: Formal methods are rigorous techniques based on mathematical notation that can be used to specify and verify software models.

Model Checking: Model checking is a formal verification technique that determines whether given properties of a system are satisfied by a model. A model checker takes a model and a property as inputs, and outputs either a claim that the property is true or a counterexample falsifying the property.

X-Machine: A deterministic stream X-machine is an 8-tuple $(\Sigma, \Gamma, Q, M, \Phi, F, q_0, m_0)$, where Σ, Γ is the input and output finite alphabet, respectively; Q is the finite set of states; M is the (possibly) infinite set called memory; Φ is the type of the machine, that is, a finite set of partial functions φ that map an input and a memory state to an output and a new memory state, $\varphi: \Sigma \times M \rightarrow \Gamma \times M$; F is the next state partial function that, given a state and a function from the type Φ, denotes the next state—*F* is often described as a state transition diagram, $F: Q \times \Phi \rightarrow Q$; q_0 and m_0 are the initial state and memory, respectively.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1201-1204, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Formalization Process in Software Development

Aristides Dasso
Universidad Nacional de San Luis, Argentina

Ana Funes
Universidad Nacional de San Luis, Argentina

INTRODUCTION

Nowadays, software engineering (SE) is considered more frequently an engineering discipline. Several definitions have been proposed by different authors, and many of them agree in affirming that SE is the application of principles and systematic practices for the development of software. That is—as it was established by the IEEE (1990)—SE is the application of engineering to the software.

As a general rule all engineering applications use mathematics or mathematical tools as a basis for their development. However, software engineering is an exception to this rule. Not all the techniques and software development methods have a formal basis. Formal methods[1] (FM) rely on mathematical foundations.

FM are a collection of methodologies and related tools, geared to the production of software employing a mathematical basis. There are a number of different formal methods each having its own methodology and tools, specially a specification language.

As it is expressed in Wikipedia (2006) and Foldoc (2006), we can say that FM are "mathematically based techniques for the specification, development and verification of software and hardware systems."

FM are based on the production of formal specifications—for which they have a formal language to express it. Sometimes there is also a method to use the language in the software development process.

The aims of FM can vary according to the different methodologies, but they all shared a common goal: the production of software with the utmost quality mainly based on the production of software that is error free. To achieve this, the different FM have developed not only a theory, but also different tools to support the formal process.

FM can cover all the steps of the life cycle of a software system development from requirement specification to deployment and maintenance. However, not all FM have that capacity, and not always it is convenient to apply them. It is necessary to make an evaluation between pros and cons before applying FM in the software development process.

FORMAL METHODS

It is important to make a distinction between a formal notation or language and a formal system. A formal notation is used to produce a formal specification, and it has a formal syntax and semantics. A formal system, besides these two components, includes a proof system—the deductive mechanism of the formal system. The syntax is described by a grammar and defines the set of well-formed formulas of the language. The meaning of these formulas is given by the semantics of the language. Finally, the syntactic manipulations of these formulas are achieved by using the inference mechanism of the formal system, which allows the derivation of new well-formed formulas from those present in the language.

Alagar and Peruyasamy (1998) establish that the difference between a FM and a formal system is the automatic support for specification and the availability of mechanized proofs.

All FM are based on mathematical formalisms, but we can distinguish two different development approaches. On the one hand, we can find the transformational methods based in transformations and on the other hand, those based in the "invent and verify" principle. The former rely for development on a calculus or transformation, where the engineer starts with an expression and then following predefined rules applies them to obtain an equivalent expression. Successive calculations lead to implementation. In FM relying on "invent and verify" technique, the engineer starts by inventing a new design, which afterward needs to be verified as correct. From this verified design implementation follows.

There are several styles of formal specification. Some are mutually compatible, while others are not. Table 1 shows a possible classification of the different styles.

Formal languages have formal definitions, not only of their syntax, but also of their semantics. Table 2 shows a possible classification for the different formal semantic definitions styles.

In NASA's Langley Research Center site for formal methods there is a nice definition and also an explanation of different degrees of rigor in FM (http://shemesh.larc.nasa.gov/fm/fm-what.html):

Table 1. Summary of specification language characteristics

Model-oriented. Based on mathematical domains. For example numbers, functions, sets, and so forth. Concrete.	**Property-oriented**. Based on axiomatic definitions. Abstract.
Applicative. Does not allow the use of variables.	**Imperative or State-oriented**. Allows the use of variables.
Static. Does not include provisions for handling time.	**Action**. Time can be considered in the specification. There are several ways of doing this: considering time as linear or branching, synchronous, asynchronous, and so forth.

Table 2. Summary of semantic definitions styles of specification languages

Operational. Concrete, not well suited for proofs. The meaning of a system is expressed as a sequence of actions of a simpler computational model.
Denotational. Abstract, well suited for proofs. The meaning of a system is expressed in the mathematical theory of domains.
Axiomatic. Very abstract, normally only limited to conditional equations. The meaning of the system is expressed in terms of preconditions and postconditions.

"Traditional engineering disciplines rely heavily on mathematical models and calculation to make judgments about designs. For example, aeronautical engineers make extensive use of computational fluid dynamics (CFD) to calculate and predict how particular airframe designs will behave in flight. We use the term 'formal methods' to refer to the variety of mathematical modelling techniques that are applicable to computer system (software and hardware) design. That is, formal methods is the applied mathematics of computer system engineering, and, when properly applied, can serve a role in computer system design analogous to the role CFD serves in aeronautical design.

Formal methods may be used to specify and model the behavior of a system and to mathematically verify that the system design and implementation satisfy system functional and safety properties. These specifications, models, and verifications may be done using a variety of techniques and with various degrees of rigour. The following is an imperfect, but useful, taxonomy of the degrees of rigour in formal methods:

Level-1:
Formal specification of all or part of the system.

Level-2:
Formal specification at two or more levels of abstraction and paper and pencil proofs that the detailed specification implies the more abstract specification.

Level-3:
Formal proofs checked by a mechanical theorem prover.

Level 1 represents the use of mathematical logic or a specification language that has a formal semantics to specify the system. This can be done at several levels of abstraction. For example, one level might enumerate the required abstract properties of the system, while another level describes an implementation that is algorithmic in style.

Level 2 formal methods goes beyond Level 1 by developing pencil-and-paper proofs that the more concrete levels logically imply the more abstract-property oriented levels. This is usually done in the manner illustrated below.

Level 3 is the most rigourous application of formal methods. Here one uses a semi-automatic theorem prover to make sure that all of the proofs are valid. The Level 3 process of convincing a mechanical prover is really a process of developing an argument for an ultimate skeptic who must be shown every detail.

Formal methods is not an all-or-nothing approach. The application of formal methods to only the most critical portions of a system is a pragmatic and useful strategy. Although a complete formal verification of a large complex system is impractical at this time, a great increase in confidence in the system can be obtained by the use of formal methods at key locations in the system."

As it is said in NASA's definition of the levels of degree of rigor, they are imperfect; others exist. Most of the FM included next have as an integral part not only a language but also a methodology included, and most of the time this methodology implies different levels of rigor in its use. For

example, RAISE—that has its own method—presents three degrees of formality (The RAISE Method Group, 1995):

- **Formal Specification Only:** Where formality is only applied to the specification procedure.
- **Formal Specification and Rigorous Development:** Where formality is applied to the specification procedure as previously, and rigor to the development process. This means that the developer starts writing abstract specifications, goes on developing more concrete ones and recording the development relations between them. These relations are then examined; however, they are not justified.
- **Formal Specification and Formal Development:** Where formality is applied not only to the specification but also to the development process including justification.

FM finds more adepts in critical software development, where the presence of errors can have grave human, economic, and social consequences. NASA, as well as other government bodies in the USA, Europe, and elsewhere, is using FM especially in avionics and systems where the utmost reliability is needed. Some examples from NASA are Small Aircraft Transportation System (SATS) (http://shemesh.larc.nasa.gov/fm/fm-now-sats.html) and Formal Analysis of Airborne Information for Lateral Spacing (AILS). Also, NASA's contractors use FM (http://shemesh.larc.nasa.gov/fm/fm-now-ails.html). For more information on these and other projects see http://shemesh.larc.nasa.gov/fm/fm-main-research.html.

Next we present a not all-inclusive list of FM. For more information on these and other FM see http://vl.fmnet.info. This site has links to the sites of many other FM.

- **Abstract State Machines (ASM):** "[M]ethodology for describing simple abstract machines which correspond to algorithms" (http://www.eecs.umich.edu/gasm/).
- **B-Method:** "B is a formal method for the development of program code from a specification in the Abstract Machine Notation" (http://www.afm.lsbu.ac.uk/b/).
- **Communicating Sequential Processes (CSP):** Process algebra originated by C. A. R. Hoare (http://www.afm.lsbu.ac.uk/csp/).
- **Duration Calculus:** "Modal logic for describing and reasoning about the real-time behavior of dynamic systems" (http://www.iist.unu.edu/newrh/II/2/1/2/page.html).
- **Extended ML:** "Framework for specification and formal development of Standard ML (SML) programs" (http://www.dcs.ed.ac.uk/home/dts/eml).
- **HOL:** Automatic theorem proving system based on Higher Order Logic (http://www.afm.lsbu.ac.uk/hol/)
- **Larch:** Specification language based on logic (http://nms.lcs.mit.edu/spd/larch/).
- **Model Checking:** "Method for formally verifying finite-state concurrent systems" (http://www-2.cs.cmu.edu/~modelcheck/).
- **OBJ3:** Specification language based on algebra.
- **Petri Nets:** "Formal, graphical, executable technique for the specification and analysis of concurrent, discrete-event dynamic systems" (http://www.petrinets.org/).
- **Prototype Verification System (PVS):** "PVS consists of a specification language, a number of predefined theories, a theorem prover, various utilities, documentation, and examples that illustrate different methods of using the system in several application areas" (http://pvs.csl.sri.com/introduction.shtml).
- **Rigourous Approach to Industrial Software Engineering (RAISE):** "[C]onsists of the RAISE development method and RSL, the RAISE Specification Language" (http://spd-web.terma.com/Projects/RAISE. See also http://www.iist.unu.edu/raise/).
- **Vienna Development Method (VDM):** Based on sets and relations; "VDM (The Vienna Development Method) is a set of techniques for modelling computing systems analyzing those models and progressing to detailed design and coding" (http://www.csr.ncl.ac.uk/vdm).
- **Z Notation:** Based on Zermelo-Fraenkel set theory and first order predicate logic (http://www.zuser.org/z).

Of course a question to ask oneself is why there are so many flavors of FM. There are undoubtedly several possible answers to this question. One is that since there are many ways to describe a system, not everyone agrees on a particular style. Why are there so many different programming languages? The answer to this question could be also the answer to our question. We can also say not all the FM addressed the same problem. Some of them are geared to system design, others to domain description; some deal with time, while others do not consider it, and so forth.

PROS AND CONS

The issue of whether FM are useful or not was discussed in the literature around 10 years ago—see for example work by Bowen and Hinchey (1995a, 1995b) and Luqi and Goguen (1997).

The use of formal specifications has benefits ranging from the possibility of building unambiguous specifications, to the possibility of proving system properties, to automatic code generation. Formal specifications provide a deep understanding of software requirements that reduce the possibility of making mistakes and omissions.

By using formal specifications it may be possible to prove system consistency; however, verification is an expensive and time-consuming task, and it requires a high level of expertise in algebra and mathematical logic.

Formal specifications can be used as the basis for the derivation of test cases. This has been treated by Aichernig (1999, 2001a, 2001b, 2001c)) among others.

Most opponents give as a main reason the fact that FM make difficult the communication with the end users when requirements must be validated. However, lately there is an ongoing interest in the combined use of semi-formal notations and formal specifications to overcome this problem and the ambiguity and inconsistency inherent in semi-formal notations. This kind of integration is intended to improve understandability—given by graphical notations—and to gain in unambiguity—offered by formal specifications.

According to Pons and Baum (2000), we can carry out integration in four different ways:

1. **Supplemental:** Where the informal notations are enriched with formal concepts.
2. **Extension:** Where the formal notations are extended with concepts from others paradigms, for example, from the object-oriented paradigm.
3. **Interface:** Where formal notations are provided with graphic interfaces to help in the development of models.
4. **Semantics:** Where the semantics of an accepted semi-formal modeling language is given by a formal language.

There is an important number of theoretical works that deal with the integration of graphical notations and mathematically precise formalisms. Proof of this are the growing interest in providing a more traditional methodology such as UML with formal basis, either through the OCL language or by giving to UML graphic language a formal semantics—see the works of the 2U Consortium (2003) and The Precise UML Group (2003). The extensive work done on formalizing Java and Java Machine can be seen in work by Hartel and Moreau (2001) and Bertelsen (2003). Other good examples can be found in work by Amálio, Stepney, and Polack (2003), DeLoach and Hartrum (2000), France (1999), Funes and George (2003), Moreira and Clark (1996), Kim and Carrington (2000), Lano (1991), Meyer and Souquieres (1999), Reggio and Larosa (1997), and Weber (1996), among others.

Finally, we should not forget the important role FM have in the development of systems where security and reliability are crucial. Hung et al. (2002) cite three factors that were the traditional arguments in favor of FM: first the growing number of applications with minimal or zero tolerance for errors, second social, environmental, and economic consequences of design errors in hardware and software products, raising social awareness and concern, and third there are a number of standards and regulatory organizations enforcing quality standards for electronic products.

EXAMPLES

Just so the reader can get the flavor of FM we present in this section an example of specifications for a stack of natural numbers using the RAISE Specification Language (RSL) (The RAISE Language Group, 1992).

The first specification in Example 1 illustrates the use of the property-oriented style, which is abstract and whose definitions are given by axioms. The second specification given for the stack is model-oriented, based in the use of

Example 1. A stack of natural numbers in RSL (using static applicative praoperty-oriented style)

```
scheme STACK_0 =
class
  type
    Stack
  value
    empty: Stack,
    push: Nat × Stack → Stack,
    pop: Stack ~→ Stack ,
    top: Stack ~→ Nat
  axiom
    ∀e: Nat, s: Stack • pop(push(e, s)) ≡ s,
    ∀e: Nat, s: Stack • top(push(e, s)) ≡ e,
end
```

Example 2. A stack of natural numbers in RSL (using static applicative model-oriented style)

```
scheme STACK_1 =
class
  type
    Stack = Nat *
  value
    empty: Stack,
    push: Nat × Stack → Stack,
    pop: Stack ~→ Stack ,
    top: Stack ~→ Nat
  axiom
    empty ≡ <>,
    ∀e: Nat, s: Stack • push(e, s) ≡ <e> ^ s,
    ∀s: Stack • pop(s) ≡ tl s
          pre s ≠empty,
    ∀s: Stack • top(s) ≡ hd s
          pre s ≠empty
end
```

concrete types (see Example 2). Both specifications use RSL static and applicative style.

The *Stack* type in Example 1 is given by an abstract type since we do not say explicitly how a stack is going to be implemented or of what type its elements are. The stack is defined by the main operations needed to manage it. To formally specify its behavior, we give the signature of three functions: *pop*, *push*, and *top*, the constant *empty* that represents the empty stack, and a set of axioms to express the intended meaning of these operations. The axioms define the essential properties that must always be true when the operations are applied.

The operations *pop* and *top* are partial functions that are not defined for the empty stack. This fact is reflected by the absence of axioms about *pop(empty)* and *top(empty)*. The operation *push* is a total function.

The first axiom states that pushing an element *e* into a stack *s* and then doing a *pop* returns the former stack *s*. The second axiom states that every time we push an element *e* into a stack and then apply *top* to the resulting stack, we get the last pushed element *e*.

But, where is the formalism in this specification? We must not forget that as any FM, RSL has defined formally not only its syntax but also its semantic. This allows, for instance, the use of axioms, which are truths that must be preserved all along the specification and that in turn can be proved—or not, if the specification has errors or inconsistencies. So the formalism can be found in the fact that using the underlying semantic—in this case logic and algebra—the specification can be subjected to proof, and therefore any errors can be found.

In this example, the axioms specify the behavior of the stack—remember that axioms must always be true; if they are not, then there is an error in the specification.

So, besides the abstractness and precision given by a formal specification language, one of the major reasons for expressing specifications in a formal language is the possibility of proving properties of the specifications. Formulating properties and then trying to prove them is a way of ensuring correctness and detecting errors.

The second example—see Example 2—is a refinement of the first one. Here, we define the *Stack* type not as a sort but as a list of natural numbers—the * next to **Nat** indicates list of **Nat** in RSL.

The axioms that define the operations on the stack are defined using the primitive operations for lists in RSL—remember that everything in RSL has its semantics formally defined, and that includes the type list as well as the elementary operations on the type. The operation *empty* is defined as the empty list. To *push* an element into the stack corresponds to adding the element at the head of the list, returning the resulting augmented stack. To *pop* an element from the stack corresponds to remove the head of the list and return the resulting reduced stack; and to get the *top* of the stack corresponds to get the head of the list and return it. Note that since *pop* and *top* have a pre-condition establishing that they cannot be applied to the empty stack, they are partial functions.

In RAISE—as in many other FM—besides being formal when specifying, we can also be formal in the development process. Software can be developed in a sequence of steps. We can start writing a suitably abstract specification and proceed developing more concrete and detailed ones until the final specification can be automatically translated to a programming language. Each step in the development must conform to the previous and must be proved as a correct development step. For example, we should prove that the specification STACK_1 is a correct development for STACK_0, that is, to prove that STACK_1 implements STACK_0, that is, STACK_0 and STACK_1 are in the implementation relation. In RAISE this means that the specifications have to meet two requirements:

- **Property of Preservation:** All properties that can be proved about STACK_0 can also be proved for STACK_1.
- **Property of Substitutivity:** In any specification an instance of STACK_0 can be replaced by an instance of STACK_1, and the resulting specification should implement the former specification.

FUTURE TRENDS

Although traditional methods seem to be the most popular nowadays, the use of FM has grown to occupy more and more a place in software engineering for development of systems where security and reliability are important. However, its use continues being expensive and limited.

Besides its leading role in the development of critical systems, there is also a growing interest in proposing methods and techniques to use formal specifications combined with semi-formal notations to give more rigor to the first phases of the development process.

If formal techniques become in the future the foundations of a new generation of CASE tools, is possible that a bigger portion of software practitioners will adopt them, making formal methods a truly practical tool for all kinds of systems.

CONCLUSIONS

FM are methods based primarily in formal, mathematical notation and principles. They are ideal for defining unambiguously not only the requirements, but also every stage in system development including implementation since there are FM that have translators to programming languages.

Because of their profound roots in logic and algebra, they seem harder to understand and learn. However the FM community aware of this view has been making efforts in producing tools that make their use easier.

REFERENCES

2U Consortium (2003). *Unambiguous UML*. Retrieved November 1, 2003, from http://www.2uworks.org

Aichernig, B. K. (1999). Automated black-box testing with abstract VDM oracles. In M. Felici, K. Kanoun, & A. Pasquini (Eds.), *Computer safety, reliability and security: Proceedings of the 18th International Conference, SAFECOMP'99* (pp. 250-259). Toulouse, France: Springer.

Aichernig, B. K. (2001a). Test-case calculation through abstraction. In J. N. Oliveira & P. Zave (Eds.), *Proceedings of Formal Methods Europe 2001, FME 2001: Formal Methods for Increasing Software Productivity* (pp. 571-589). Berlin, Germany: Springer.

Aichernig, B. K. (2001b). Test-design through abstraction—A systematic approach based on the refinement calculus. *Journal of Universal Computer Science, 7*(8), 710-735.

Aichernig, B. K. (2001c). *Systematic black-box testing of computer–based systems through formal abstraction techniques.* Unpublished doctoral dissertation, Technischen Universität Graz, Graz.

Alagar, V. S., & Peruyasamy, K. (1998). *Specification of software systems*. New York: Springer.

Amálio N., Stepney S., & Polack F. (2003). Modular UML semantics: Interpretations in Z based on templates and generics. In *FACS'03: Formal Aspects of Component Software, International Workshop*, Pisa, Italy (Tech. Rep. No. UNU/IIST 284).

Bertelsen, P. (2003). *Semantics of Java Byte Code*. Retrieved November 1, 2003, from ftp://ftp.dina.kvl.dk/pub/Staff/Peter.Bertelsen/jvm-semantics.ps.gz

Bowen, J., & Hinchey, M. (1995a). Seven more myths of formal methods. *IEEE Software, 12*(3), 34-40.

Bowen, J., & Hinchey, M. (1995b). Ten commandments of formal methods. *IEEE Computer, 28*(4), 56-63.

Butler, R. (2006). *Singular vs. plural*. Retrieved July 7, 2006, from http://shemesh.larc.nasa.gov/fm/fm-is-vs-are.html

DeLoach, S., & Hartrum, T. (2000). A theory-based representation for object-oriented domain models. *IEEE Transactions on Software Engineering, 6*(6), 500-517.

Foldoc. (2006). *Foldoc, the free online dictionary of computing*. Retrieved July 7, 2006, from http://foldoc.doc.ic.ac.uk/foldoc

France, R. (1999). A problem-oriented analysis of basic UML static requirements modeling concepts. In *Proceedings of OOPSLA '99* (pp. 57-69). Denver, CO. ACM Press.

A. Funes and C. George (2003). Formalizing UML class diagrams. In L. Favre (Ed.) *UML and the unified process* (pp. 129-198). Hershey, PA: Idea Group Publishing.

Hartel, P., & Moreau, L. A. V. (2001). Formalizing the safety of Java, the Java virtual machine and Java card. *ACM Computing Surveys, 33*(4), 517-558.

Hinchey, M., & Bowen, J. (Eds.) (1995). *Applications of formal methods*. Prentice Hall International in Computer Science..

Hung, D. V., Janowski, G. T., & Moore, R. (Eds.) (2002). *Specification case studies in RAISE*. Springer.

IEEE. (1990). Standards collection: Software engineering. *IEEE Standard 610.12-1990*. Retrieved from http://standards.ieee.org/reading/ieee/std/se/610.12-1990.pdf

Kim, S-K., & Carrington, D. (2000). *A formal specification mapping between UML models and object-Z specifications*. (LNCS 1878, pp. 2-21). London: Springer.

Lano, K. (1991). Z++, an object-oriented extension to Z. In J. Nicholls (Ed.), *Z user workshop, workshops in computing* (pp. 151-172). Oxford: Springer-Verlag.

Luqi, & Goguen, J. (1997). Formal methods: Promises and problems. *IEEE Software, 14*(1), 73-85.

Meyer, E., & Souquieres, J. (1999). A systematic approach to transform OMT diagrams to a B specification. In *Proceedings of FM '99* (vol. I) (pp. 875-895).

Moreira, A.M.D., & Clark, R.G. (1996), LOTOS in the object-oriented analysis process. In S. Goldsack & S. Kent (Eds.), *Formal methods and object technology* (pp. 33-46), Springer-Verlag.

Pons, C., & Baum, G. (2000). Formal foundations of object-oriented modeling notations. In *Proceedings of the Third IEEE International Conference on Formal Engineering Methods (ICFEM'00)* (pp. 101-110).

Regio, G., & Larosa, M. (1997). A graphic notation for formal specification of dynamic systems. In J. S. Fitzgerald, C. B. Jones, & P. Lucas (Eds.), *Proceedings of FME'97* (Vol. 1313, pp. 40-61). London: Springer-Verlag.

The Precise UML Group. (2003). Retrieved November 1, 2003, from http://www.puml.org/

The RAISE Language Group. (1992). *The RAISE specification language*. Prentice Hall International.

The RAISE Method Group. (1995). *The RAISE development method*. Prentice Hall International.

Weber, M. (1996). Combining statecharts and Z for the design of safety-critical control systems. In M.-C. Guade & J. Woodcock (Eds.), *Proceedings of the Third International Symposium of FME'96* (LNCS, pp. 307-326). London: Springer-Verlag.

Wikipedia. (2006). *Wikipedia, the free encyclopedia*. Retrieved July 7, 2006, from http://en.wikipedia.org/wiki/Formal_methods

KEY TERMS

Action-Oriented Formal Specification Language: Time can be considered in the specification. There are several ways of doing this: considering time as linear or branching, synchronous, asynchronous, and so forth.

Applicative-Oriented Formal Specification Language: Does not allow the use of variables.

Axiomatic Semantics: The meaning is given in terms of conditions, pre and post.

Denotational Semantics: The meaning is given in terms of mathematical functions.

Imperative or State-Oriented Formal Specification Language: Allows the use of variables.

Model-Oriented Formal Specification Language: Based on mathematical domains. For example, numbers, functions, sets, and so forth. Concrete.

Operational Semantics: The meaning is given in terms of rules that specify how the state of a computer—real or formal—changes while executing a program.

Property-Oriented Formal Specification Language: Based on axiomatic definitions. Abstract.

Semantics: In a language, it is the meaning of a string, as opposed to syntax, that describes how the symbols of the language are combined. Most programming languages have their syntax defined formally (traditionally in BNF), while formal specification languages have also their semantics defined formally.

Specification: A document describing what a system should do, what a problem is, or what a domain is all about. In formal methods this document is written in a formal language.

Static-Oriented Formal Specification Language: Do not include provisions for handling time.

Verification: The process of determining whether or not the products of a specification phase fulfill a set of established requirements. Sometimes this is also used to indicate the process of proving that a more concrete specification preserves the properties of a more abstract specification.

ENDNOTE

[1] Some authors use the singular while others employ the plural in referring to FM. We can refer to Ricky W. Butler (2006):

"Some of you are saying, 'Formal Methods is ...? What's wrong with these people, ain't nobody learned them no grammar!' In an age in which few people care about the proper use of language, your concern is commendable; however, in this instance, your concern is also unwarranted.

In these pages, we are using the term formal methods to refer to a particular collection of knowledge. Just as the plural-sounding term fluid dynamics is treated as singular, so too may the term formal methods be treated as singular. A legitimate argument can be made as to the acceptability of treating the term as plural, but no legitimate argument can be made as to the necessity of doing so."

Category: Software & Systems Design

Foundations for MDA Case Tools

Liliana María Favre
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

Claudia Teresa Pereira
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

Liliana Inés Martinez
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

INTRODUCTION

The model driven architecture (MDA) is an initiative proposed by the object management group (OMG), which is emerging as a technical framework to improve productivity, portability, interoperability, and maintenance (MDA, 2003).

MDA promotes the use of models and model-to-model transformations for developing software systems. All artifacts, such as requirement specifications, architecture descriptions, design descriptions, and code are regarded as models. MDA distinguishes four main kinds of models: computation independent model (CIM), platform independent model (PIM), platform specific models (PSM), and implementation specific model (ISM).

A CIM describes a system from the computation independent viewpoint that focuses on the environment of and the requirements for the system. In general, it is called domain model. A PIM is a model that contains no reference to the platforms that are used to realize it. A PSM describes a system with full knowledge of the final implementation platform. In this context, a platform is "a set of subsystems and technologies that provide a coherent set of functionality which any application supported by that platform can use without concern for the details of how the functionality is implemented" (MDA, 2003, p. 2-3). PIMs and PSMs are expressed using the unified modeling language (UML) combined with the object constraint language (OCL) (Favre, 2003; OCL, 2004; UML, 2004).

The idea behind MDA is to manage the evolution from CIMs to PIMs and PSMs that can be used to generate executable components and applications. In MDA is crucial to define, manage, and maintain traces and relationships between different models and automatically transform them and produce code that is complete and executable.

Metamodeling has become an essential technique in model-centric software development. The metamodeling framework for the UML itself is based on architecture with four layers: meta-metamodel, metamodel, model, and user objects. A metamodel is an explicit model of the constructs and rules needed to build specific models, its instances. A meta-metamodel defines a language to write metamodels. OCL can be used to attach consistency rules to models and metamodels. Related OMG standard metamodels and meta-metamodels such as meta object facility (MOF), software process engineering metamodel (SPEM) and common warehouse model (CWM) share a common design philosophy (CWM, 2001; MOF, 2005; SPEM, 2005).

MOF defines a common way for capturing all the diversity of modeling standards and interchange constructs. MOF uses an object modeling framework that is essentially a subset of the UML core. The four main modeling concepts are "classes, which model MOF metaobjects; associations, which model binary relationships between metaobjects; data types, which model other data; and packages, which modularize the models" (MOF, 2005, p. 2-6). The query, view, transformation (QVT) standard depends on MOF and OCL for specifying queries, views, and transformations. A query selects specific elements of a model, a view is a model derived from other model, and a model transformation is a specification of a mechanism to convert the elements of a model, into elements of another model, which can be instances of the same or different metamodels (QVT, 2003).

The success of MDA depends on the existence of CASE (computer-aided software engineering) tools that make a significant impact on software processes such as forward engineering and reverse engineering processes (CASE, 2006). This article explains the most important challenges to automate the processes that should be supported by MDA tools. We propose an integration of knowledge developed by the community of formal methods with MDA. We describe a rigorous framework that comprises the metamodeling notation NEREUS and bridges between MOF-metamodels and NEREUS, and between NEREUS and formal languages. NEREUS can be viewed as an intermediate notation open to many other formal specifications. We analyze metamodeling techniques for expressing model transformations such as refinements and refactorings. Our approach focuses on interoperability of formal languages in model driven development (MDD).

This article is organized as follow. We first analyze the limitations of the existing MDA-based CASE tools. Then, we describe the bases of rigorous MDA-based processes. Next, we show how the formalization of MOF metamodels and metamodel-based model transformations allows us automatic software generation. Finally, we highlight the key directions in which MDA is moving forward.

BACKGROUND

To date, there are about 120 UML CASE tools that vary widely in functionality, usability, performance, and platforms (CASE, 2006). Some of them can only help with the mechanics of drawing and exporting UML diagrams. The mainstream object-oriented CASE tools support forward engineering and reverse engineering processes and can help with the analysis of consistency between diagrams. Only a few UML tools include extension for real time modeling. The tool market around MDA tools is still in flux and only about 10% of them provide some support for MDA. Table 1 exemplifies a taxonomy of the UML CASE tools (CASE, 2006).

The current techniques available in the commercial tools do not allow generating complete and executable code and after generation, the code needs additions. A source of problems in the code generation processes is that, on the one hand, the UML models contain information that cannot be expressed in object-oriented languages while, on the other hand, the object-oriented languages express implementation characteristics that have no counterpart in the UML models.

Moreover, the existing CASE tools do not exploit all the information contained in the UML models. For instance, cardinality and constraints of associations and preconditions, postconditions, and class invariants in OCL are only translated as annotations. It is the designer's responsibility to make good use of this information either selecting an appropriate implementation from a limited repertoire or implementing the association by himself.

On the other hand, many CASE tools support reverse engineering, however, they only use more basic notational features with a direct code representation and produce very large diagrams. Reverse engineering processes are facilitated by inserting annotations in the generated code. These annotations are the link between the model elements and the language. As such, they should be kept intact and not be changed. It is the programmer's responsibility to know what he or she can modify and what he or she cannot modify.

UML CASE tools provide limited facilities for refactoring on source code through an explicit selection made for the designer. However, it will be worth thinking about refactoring at the design level. The advantage of refactoring at UML level is that the transformations do not have to be tied to the syntax of a programming language. This is relevant since UML is designed to serve as a basis for code generation with MDA (Sunye et al., 2001).

Techniques that currently exist in UML CASE tools provide little support for validating models in the design stages. Reasoning about models of systems is well supported by automated theorem provers and model checkers, however, these tools are not integrated into CASE tools environments. Another problem is that as soon as the requirements specifications are handed down, the system architecture begins to deviate from specifications (Kollmann & Gogolla, 2002). Only research tools provide support for formal specification and deductive verification.

All of the MDA CASE tools are partially compliant to MDA features. They provide good support for modeling and limited support for automated transformation. In general, they support MDD from the PIM level and use UML class diagrams for designing PIMs. Some of them provide only one level of transformation from PIM to code (Codagen, Ameos, Arcstyler) and, in general, there is no relation between QVT and the current existing MDA tools. As an example, OptimalJ from Compuware supports MDD from PIM level. It allows generating PSMs from a PIM and a partial code generation. It distinguishes three kinds of models: a domain model that correspond to a PIM model, an application model that includes PSMs linked to different platforms (Relational-PSM, EJB-PSM and Web-PSM), and an implementation model.

Table 1. UML CASE tools

Basic drawing tools	Visio
Main stream object oriented case tools	Rational Rose, Argo/UML, Together, UModel, MagicDraw, MetaEdit+, Poseidon
Real time/embedded tools	Rapsody, Rational Rose Real Time, RapidRMA
MDA-based tools	OptimalJ, AndroMDA, Ameos, Together Architect, Codagen, ArcStyler, MDE Studio, Objecteering

The transformation process is supported by transformation and functional patterns.

The MDA-based tools use MOF to support OMG standards such as UML and XMI (XML metadata interchange). MOF has a central role in MDA as a common standard to integrate all different kinds of models and metadata and to exchange these models among tools; however, MOF does not allow capturing semantic properties in a platform independent way and there is no rigorous foundations for specifying transformations among different kinds of models.

A lot of research work has been carried out dealing with the advanced metamodeling techniques and formalization of different kinds of transformations. For instance, the main task of USE tool (Gogolla, Bohling, & Ritchers, 2005) is to validate and verify specifications consisting of UML/OCL class diagrams. Key (Ahrendt et al., 2002) is a tool based on together (CASE, 2006) enhanced with functionality for formal specification and deductive verification.

Akehurst and Kent (2002) propose an approach that uses metamodeling patterns that capture the essence of mathematical relations. The proposed technique is to adopt a pattern that models a transformation relationship as a relation or collections of relations, and encode this as an object model. Hausmann (2003) defined an extension of a metamodeling language to specify mappings between metamodels based on concepts presented in Akehurst et al. (2002). Kuster, Sendall, and Wahler (2004) compare and contrast two approaches to model transformations: one is graph transformation and the other is a relational approach. Czarnecki and Helsen (2003) describe a taxonomy with a feature model to compare several existing and proposed model-to-model transformation approaches.

RIGOROUS MODEL-DRIVEN DEVELOPMENT

Developing or reengineering a system in an MDA perspective should be done through automated transformation with the help of tools. Figure 1 illustrates the different processes and artifacts beyond this idea. Forward engineering and reverse engineering processes should be supported in MDA tools. Forward engineering is the process of transforming higher-level or abstract models into concrete ones. Reverse engineering reconstructs higher-level models from low ones. Reengineering is the process that transforms one concrete representation to another, while reconstituting the higher-level models along the way. We describe a rigorous framework compliant to MDA forward engineering processes. A model-driven development is carried out as a sequence of model transformations that includes, at least, the following steps: construct a CIM, transform the CIM into a PIM that provides a computing architecture independent of specific platforms, transform the PIM into one or more PSMs, each one suited for specific platforms, and derive code directly from the PSMs.

A model transformation is the process of converting one model into another model preserving some kind of equivalence relation between them. We can distinguish two types of transformations to support model evolution from CIMs to ISMs: refinements and refactorings. A refinement is the process of building a more detailed specification that conforms to another that is more abstract. On the other hand, a refactoring means changing a model leaving its behavior unchanged, but enhancing some non-functionality quality factors such as simplicity, flexibility, understandability, and performance.

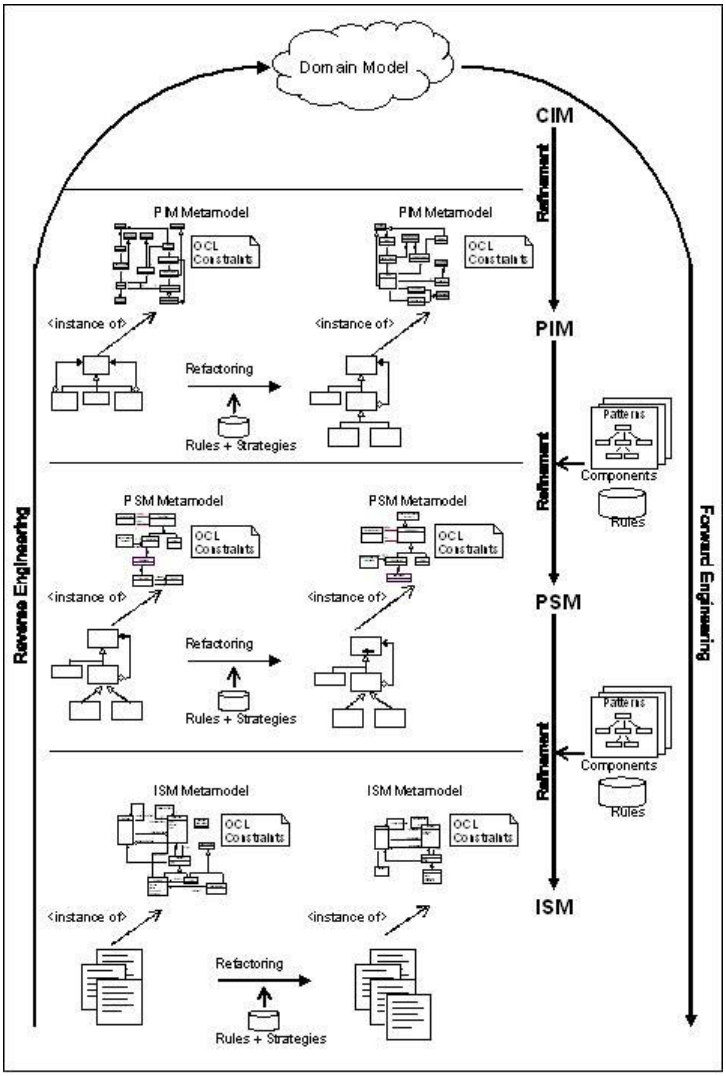
Metamodeling is a powerful technique to specify families of models and model transformations. Figure 1 shows the different correspondences that may be held between several models and metamodels and their interrelations. A CIM is related to one or more PIM-metamodels. A PIM-metamodel is related to more than one PSM-metamodels, each one suited for different platforms (e.g., .NET, J2EE, or relational). The PSM-metamodels correspond to ISM-metamodels. A metamodel is a description of all the concepts that can be used in the respective level. For instance, a metamodel linked to a relational platform refers to concepts of table, foreign key and column. An ISM-metamodel includes concepts of programming languages such as constructor and method.

The following types of model transformations can be distinguished:

- **CIM to PIM refinement:** It describes how a CIM that is an instance of a MOF-metamodel is transformed into a PIM that is an instance of a specialized metamodel for a specific computation dependent model.
- **PIM to PSM refinement:** It describes how a PIM that is an instance of a MOF-Metamodel is transformed into a PSM that is an instance of a specialized MOF-metamodel for a specific platform.
- **PSM to ISM refinement:** It describes how a PSM is transformed into code (which is an instance of MOF-metamodel for a platform and specific language technologies).
- **Refactoring:** It specifies how a model in a given level is transformed into a new restructured model in the same level (for instance, PIM to PIM, PSM to PSM, ISM to ISM). The source and target models are instances of the same MOF-metamodel.

Metamodel transformations are a specific type of model transformations that impose relations between pairs of metamodels. A metamodel-based transformation is a specification of a mechanism to convert the elements of a model, that are instances of a particular metamodel, into elements of another model, which can be instances of the same or different metamodels. We specify metamodel-based model

Figure 1. Rigorous model-driven development



transformations as OCL contracts that are described by means of a transformation name, parameters, preconditions, postconditions, and additional operations.

The MDA-based processes are based on the adaptation of reusable components and systems of transformations rules. We analyzed basic techniques for MDA-based processes such as refactoring (Kerievsky, 2004; Long, Jifeng & Liu, 2005; Mens, Demeyer, Du Bois, Stenten, & Van Gorp, 2004) and design pattern (France, Kim, Ghosh, & Song, 2004; Gamma, Helm, Johnson, & Vlissides, 1995).

Pereira and Favre (2006) propose a metamodeling technique to define refactorings at different abstraction levels (e.g., PIM, PSM, and ISM). A transformational system based on behaviour-preserving model-to-model transformations was defined. To reason about correctness and robustness we propose to specify refactorings as OCL contracts that are based on metamodels capturing common properties to a family of refactorings.

Martinez and Favre (2006) describe a metamodeling technique to define design pattern components from an MDA perspective. In this context, we propose a "megamodel" for defining reusable components that integrates different kinds of models with their respective metamodels. We analyze metamodel-based model transformations among levels of PIMs, PSMs and ISMs. We illustrate the approach to define reusable design pattern components using the popular Gamma patterns (Gamma et al., 1995).

FORMALIZATION OF MDA-BASED PROCESSES

UML and OCL are too imprecise and ambiguous when it comes to simulation, verification, validation, and forecasting of system properties and even when it comes to generating

models/implementations through transformations. Although OCL is a textual language, OCL expressions rely on UML class diagrams (i.e., the syntax context is determined graphically). OCL does also not have the solid background of a classical formal language. In the context of MDA, model transformations should preserve correctness. To achieve this, the different modeling and programming languages involved in an MDD must be defined in a consistent and precise way. Then, the combination of UML/OCL specifications and formal languages offers the best of both worlds to software developer. In this direction, we define NEREUS to take advantage of all the existing theoretical background on formal methods, using different tools such as theorem provers, model checkers, or rewrite engines in different stages of MDD.

Favre (2006) proposes a rigorous framework to model driven developments. The bases of this approach are the metamodeling notation NEREUS and, bridges between UML/OCL and NEREUS and between NEREUS and algebraic languages.

NEREUS can be viewed as an intermediate notation open to many other formal specifications, such as algebraic, functional or logic ones. NEREUS is suited for specifying MOF. Most of the MOF concepts for metamodels (entity, associations, and packages) can be mapped to NEREUS in a straightforward manner. This language is relation-centric which means that it expresses different kinds of UML relations (dependency, association, aggregation, and composition) as primitives to develop specifications. In Favre (2006), we show how to integrate NEREUS with algebraic languages using the common algebraic specification language (CASL) (Bidoit & Mosses, 2004).

The formalization of MDA-based processes implies to specify metamodels and metamodel-based transformations.

On the one hand, we define a bridge between MOF-metamodels and NEREUS that is based on a system of transformation rules to convert automatically UML/OCL into NEREUS specifications. Starting from UML class diagrams, an incomplete algebraic specification can be built by instantiating reusable schemes and components, which already exist in the NEREUS predefined library. Analyzing OCL specifications, it is possible to derive axioms that will be included in the NEREUS specification. Preconditions written in OCL are used to generate preconditions in NEREUS. Postconditions and invariants allow us to generate axioms in NEREUS. Thus, an incomplete specification can be built semi-automatically (Favre, 2005; Favre, Martinez, & Pereira, 2003).

On the other hand, we formalize transformations (refinements and refactorings) as OCL contracts that are translated into NEREUS specifications by instantiating reusable schemes.

We have applied the approach to transform UML/OCL class diagrams into NEREUS specifications, which in turn, are used to generate object-oriented code (Favre, 2005; Favre et al., 2005). The process is based on the adaptation of MDA-based reusable components. NEREUS allows us to keep a trace of the structure of UML models in the specification structure that will make easier to maintain consistency between the various levels when the system evolves. All the UML model information (classes, associations, and OCL specifications) is overturned into specifications having implementation implications. The transformation of different kinds of UML associations into object-oriented code was analyzed, as well as, the construction of assertions and code from algebraic specifications. The proposed transformations preserve the integrity between specification and code. The transformation process is based on reusable components.

In Favre and Martinez (2006) we describe how formalize MOF- metamodels and metamodel-based transformations exemplifying with MDA design pattern components.

In contrast to other works, our approach is the only one focusing on interoperability of formal languages in model-driven software development. There are UML formalizations based on different languages that do not use an intermediate language such as NEREUS. However, this extra step provides some advantages. NEREUS would eliminate the need to define formalizations and specific transformations for each different formal language. The metamodel specifications and transformations can be reused at many levels in MDA. Languages that are defined in terms of NEREUS metamodels can be related to each other because they are defined in the same way through a textual syntax.

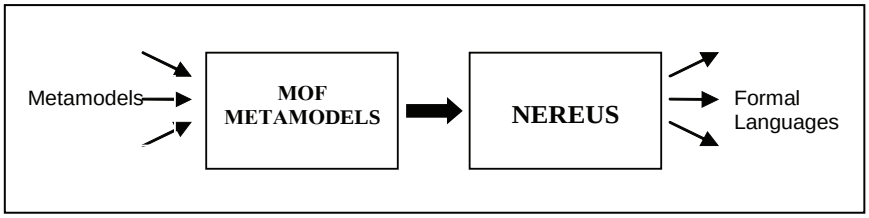
Any number of source languages (modeling language) and target languages (formal language) could be connected without having to define explicit model/metamodel transformations for each language pair. NEREUS embraces changes at different levels of abstraction (Figure 2).

FUTURE TRENDS

Nowadays, there exists an increased demand of reengineering of legacy systems towards new technologies. Advanced MDA tools should reverse existing code to abstract models to facilitate platform migration. It will probably take several years before a full round trip engineering based on standards occurs (many authors are skeptical about this). The existing MDA-based tools do not provide sophisticated transformation from PIM to PSM and from PSM to code.

To solve problems basic research on formalisms and theories will have to be carried out dealing with software evolution in MDA. If MDA becomes a commonplace, adapting it to formal development will become crucial. Formal and semi-formal techniques can play complementary roles in software development processes. This integration is benefi-

Figure 2. Interoperability of formal languages



cial for both graphical and formal specification techniques. On the one hand, semi-formal techniques have the ability to visualize language constructs allowing a great difference in the productivity of the specification process, especially when the graphical view is supported by means of good tools. On the other hand, formal specifications allow us to produce a precise and analyzable software specification before implementation and to define semi-automatic forward engineering processes.

The integration between ontology (that are essentially CIMs) and MDA will occupy a central place in MDD (Djuric, Gasevic, & Devedzic, 2006). The use of formal specification will make it possible to perform automated reasoning about ontology. A new type of MDA tools that do a more intelligent job might emerge. Probably, the next generation of tools might be able to describe the behavior of software systems in terms of domain models and translate it into executable programs on distributed environment.

CONCLUSION

There is a great number of UML CASE tools in existence that facilitates code generation and limited support for reverse engineering. Unfortunately, the current techniques available in these tools provide little automation for MDD. The formalization of metamodels and metamodel-based model transformations can help to overcome these problems. We propose to integrate knowledge developed by the community of formal methods with MDA. A rigorous framework for MDD was defined. It is comprised of a metamodeling notation NEREUS, a "megamodel" for defining MDA components and the definition of metamodeling/model transformations based on MOF and NEREUS. We define basic techniques for forward engineering and reverse engineering.

We define systems of transformation rules that allow translating MOF-metamodels to formal specifications and implementations. A bridge between NEREUS and algebraic languages was defined by using CASL. Our approach focuses on interoperability of formal languages.

We want to define foundations for MDA tools that permit designers to directly manipulate the visual models they have created. However, meta-designers need to understand metamodels and metamodel transformations.

This research is still evolving and additional issues will have to be tackled in order to fit advances in MDD.

ACKNOWLEDGMENT

This work is partially supported by the Comisión de Investigaciones Científicas de la Provincia de Buenos Aires (CIC).

REFERENCES

Ahrendt, W., Baar, T., Beckert, B., Giese, M., Hähnle, R., Menzel, W., Mostowski, W., & Schmitt, P. (2002). The Key system: Integrating object-oriented design and formal methods. In R. Kutsche, & H. Weber (Eds.), *Lecture notes in computer science 2306* (pp. 327-330). Berlin: Springer-Verlag.

Akehurst, D., & Kent, S. (2002). A relational approach to defining transformations in a metamodel. In J. M. Jezequel, H. Hussmann, & S. Cook (Eds.), *Lecture notes in computer science 2460* (pp. 243-258). Berlin: Springer-Verlag.

Bidoit, M., & Mosses, P. (2004). CASL User manual--Introduction to using the common algebraic specification language. *Lecture Notes in Computer Science 2900*. Berlin: Springer-Verlag,

CASE. (2006). *CASE TOOLS*. Retrieved December 2006, from www.objectsbydesign.com

CWM. (2001). *Common warehouse metamodel (CWM) Specification, Version 1.1*. Retrieved December 2006, from www.omg.org/cgi-bin/doc?ad/2001-02-01

Czarnecki, K., & Helsen, S. (2003). Classification of model transformation approaches. In J. Bettin, G. Van Emde, A. Agrawal, E. Willink, & J. Bezivin (Eds.), *Proceedings of OOPSLA'03 Workshop on Generative Techniques in the Context of Model-Driven Architecture*. Retrieved December 2006, from www.oopsla.org/oopsla2003

Djuric, D., Gasevic, D., & Devedzic, V. (2006). *Model driven architecture and ontology development*. Berlin: Springer.

Favre, L. (2003). *UML and the unified process*. USA: IRM Press.

Favre, L. (2006). A rigorous framework for model driven development. In K. Siau (Ed.), *Advanced topics in database research* (Vol. 5, Chapter I, pp. 1-27). Hershey, PA: Idea Group Publishing.

Favre, L. (2005). Foundations for MDA-based forward engineering. *Journal of Object Technology (JOT), 4*(1), 129-153.

Favre, L., & Martinez, L. (2006). Formalizing MDA components. In M. Morisio (Ed.), *Proceedings of the 9th International Conference on Software Reuse. Lecture Notes in Computer Science 4039* (pp. 326-339). Berlin: Springer-Verlag.

Favre, L., Martinez, L., & Pereira, C. (2005). Forward engineering of UML static models. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology* (pp. 1212-1217). Hershey, PA: Idea Group Publishing.

France, R., Kim, D., Ghosh, S., & Song, E. (2004). A UML-based pattern specification technique. *IEEE Transactions on Software Engineering, 30*(3), 193-206.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns. Elements of reusable object-oriented software*. USA: Addison-Wesley.

Gogolla, M., Bohling, J., & Richters, M. (2005). Validating UML and OCL models in USE by automatic snapshot generation. *Journal on Software and System Modeling*. Retrieved December 2006, from http://db.informatik.uni-bremen.de/publications

Hausmann, J. (2003). Relations-relating metamodels. In A. Evans, P. Sammut, & J. Williams (Eds.), *Proceedings of Metamodeling for MDA. The 1st International Workshop*. Retrieved December 2006, from http://wwwcs.uni-paderborn.de/cs/ag-engels/Papers/2004/MM4MDAhausmann.pdf

Kerievsky, J. (2004). *Refactoring to patterns*. USA: Addison-Wesley.

Kim, S., & Carrington, D. (2002). A formal model of the UML metamodel: The UML state machine and its integrity constraints. *Lecture Notes in Computer Science 2272* (pp. 477-496). Berlin: Springer-Verlag.

Kuster, J., Sendall S., & Wahler M. (2004). Comparing two model transformation approaches. In J. Bezivin et al. (Eds.), *Proceedings of OCL and Model Driven Engineering Workshop*. Lisboa, Portugal. Retrieved December 2006, from http://www.cs.kent.ac.uk/projects/ocl/oclmdewsuml04

Kollmann, R., & Gogolla, M. (2002). Metric-Based Selective Representation of UML Diagrams. In T. Gyimóthy & F. Brito e Abreu (Eds.), *Proceedings of 6th European Conf. Software Maintenance and Reengineering (CSMR'02)*. Los Alamitos: IEEE.

Long, Q., Jifeng, H., & Liu, Z. (2005). *Refactoring and pattern-directed refactoring: A formal perspective*. Technical Report 318, UNU-IIST, P.O. Box 3058, Macau. Retrieved December 2006, from www.iist.unu.edu/home/Unuiist/newrh/I/3/14/docs/report_14.html

Martinez, L., & Favre, L. (2006). MDA-based design pattern components. In M. Khosrow-Pour (Ed.), *Proceedings of the 17th IRMA International Conference. Emerging Trends and Challenges in Information Technology Management* (pp. 259-263). Hershey, PA: Idea Group Publishing.

MDA. (2003). *MDA Guide V1.0.1*. Retrieved December 2006, from omg/03-06-01 http://www.omg.org/cgi-bin/doc?omg/03-06-01

Mens, T., Demeyer, S., Du Bois, B., Stenten, H., & Van Gorp, P. (2003). Refactoring: Current research and future trends. *Electronic Notes in Computer Science, 82*(3).

MOF. (2005). *Meta object facility (MOF ™) 1.4*. Document formal/2002-04-03. Retrieved December 2006, from www.omg-org/mof

OCL. (2006). *OCL specification. Version 2.0*. Document: formal/2006-05-01. Retrieved December 2006, www.omg.org

Pereira, C., & Favre, L. (2006). Specifying refactorings as metamodel-based transformation. In Mehdi Khosrow-Pour (Ed.), *Proceedings of the 17th IRMA International Conference. Emerging Trends and Challenges in Information Technology Management* (pp. 264-268). Hershey, PA: Idea Group Publishing.

QVT. (2003). *Revised submission for MOF 2.0 Query/Views/Transformations RFP. Version 1.1*. OMG Adopted Specification. ptc/05-11-01. Retrieved December 2006, from www.omg.org

SPEM. (2005). *Software process engineering metamodel, version 1.1*. Retrieved December 2006, from www.omg.org/cgi-bin/doc?formal/2005-01-06

Sunyé, G., Pollet, D., Y. Le Traon, Y., & Jezequel, J.M. (2001). Refactoring UML Models. In M. Gogolla, & C. Kobryn (Eds.) *Lecture Notes in Computer Science 2185* (pp. 134-148). Berlin: Springer-Verlag.

Szyperski, C., Gruntz, D., & Murer, S. (2002). *Component software. Beyond object-oriented programming* (2nd ed.). USA: Addison-Wesley.

Thomas, D. (2005). Refactoring as meta programming? *Journal of Object Technology, 4*(1), 7-11. Retrieved December 2006, from www.jot.fm/issues/issue_2005_01/column1

UML. (2005). *UML 2.0 Superstructure Specification*. Document formal/2005-07-04. Retrieved December 2006, from www.omg.org/cgi-bin/doc?formal/05-07-04

KEY TERMS

CASE Tool: Computer aided software engineering (CASE); a tool to aid in the analysis and design of software systems.

Forward Engineering: The process of transforming a model into code through a mapping to a specific implementation language.

MDA (Model Driven Architecture): A framework based on UML and other industry standards for visualizing, storing, and exchanging software design and models. It separates the specification of functionality from the specification of the implementation of that functionality on a specific technology platform.

Metamodel: A model that defines the language for expressing a model.

Meta-Metamodel: A model that defines the language for expressing a metamodel.

Model Transformation: The process of converting one model into another model preserving some kind of equivalence relation between them.

OCL (Object Constraint Language): A notational language for analysis and design of software systems that allows software developers to write constraints and queries over object models such as UML models.

Refactoring: A change to a system that leaves its behavior unchanged but enhances some nonfunctional quality factors such as simplicity, flexibility, understanding and performance.

Reverse Engineering: The process of transforming code into a model through a mapping from a specific implementation language.

UML (Unified Modeling Language): An OMG standard language for visualizing, specifying, constructing, and documenting the artifacts of a software-intensive system.

A Framework for Communicability of Software Documentation

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

The role of communication is central to any software development. The documentation forms the *message carrier* within the communication infrastructure of a software project.

As software development processes shift from predictive to adaptive environments and serve an ever more hardware diverse demographic, new communication challenges arise. For example, an engineer may want to be able to remotely author a document in a shell environment without the need of any special purpose software, port it to different computer architectures, and provide different views of it to users without making modifications to the original. However, the current state of affairs of software documentation is inadequate to respond to such expectations.

In this article, we take the position that the ability of documents to be able to communicate at all levels intrinsically depends upon their *representation*. The rest of the article proceeds as follows. We first outline the background necessary for later discussion. This is followed by a proposal for a quality-based framework for representing software documentation in descriptive markup and application to agile software documentation. Next, challenges and avenues for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

Since the origins of software, and subsequently the recognition of software engineering as a discipline, documentation has had an important role to play. The use of documentation in software has a long and rich history (Furuta, Scofield, & Shaw, 1982; Goldfarb, 1981; Knuth, 1992).

There are various means that have been used for expressing software documentation. The documents could for example be expressed in structured natural language text (mimicking typewriting); Rich Text Format (RTF) and its implementations such as Microsoft Word; Hypertext Markup Language (HTML), which supports multiple language characters and symbols that can reach a broad demographic worldwide, incorporates features of print publishing, and supports hyperlinking; the Portable Document Format (PDF); and TEX/LATEX and their variations that are oriented to mathematical typesetting. However, these traditional means suffer from one or more of the following limitations: they are proprietary and can only be authored or rendered by a proprietary software; the focus is not on software engineering but other disciplines such as generic office or scientific use; and the focus is mainly on the presentation or processing rather than on representation.

Descriptive markup (Goldfarb, 1981) is based on a rich model of text known as the ordered hierarchy of content objects (OHCO) (Coombs, Renear, & DeRose, 1987; DeRose, Durand, Mylonas, & Renear, 1997) that lends a hierarchical structure to documents. The Standard Generalized Markup Language (SGML) and its simplification the Extensible Markup Language (XML) are exemplary of descriptive markup. SGML/XML are both *meta-markup* mechanisms that lend a suitable basis for a concrete serialization syntax for expressing information in a software document. They define a document in terms of its OHCO structure with mnemonic names, usually inspired by the domain being addressed, for the content objects of the data. There is a large and increasing base of markup languages based on SGML/XML.

The focus in this article is primarily on XML. Indeed, the use of XML for software process documents has been proposed (Clements et al., 2002; Mundle, 2001). The DocBook/XML markup language has been deployed for software user documentation. These efforts, however, are oriented towards technology rather than descriptive markup or communicability (or quality in general); they do not provide comparisons with other means of representations, and they do not include details of challenges posed during document engineering.

DESCRIPTIVE MARKUP AND REPRESENTATION OF SOFTWARE DOCUMENTATION

We look at a software document from two viewpoints, namely that of a *producer* and that of a *consumer*. Based on that, the representation requirements that we consider pertinent for software documentation are the following:

- **Communicability Concerns for a Document Producer:** A provider, who is responsible for both internal

and external documentation, could be interested in any of the following aspects: be able to express the domain under consideration well; have the flexibility of authoring and serving documents in different modalities; readily move documents between computing environments and over networks; easily manage the collection of documents, particularly as they scale (grow in number).

- **Communicability Concerns for a Document Consumer.** A consumer, whose main concern is external documentation, could be interested in any of the following aspects: access the documents on the device he/she is using that may be stand alone or connected to the Internet, and in the natural language/characters of choice; read or listen to the documents in the way he/she prefers that they should be presented; may want to simply look at the table of contents before reading further, or look up the definition of a term used in the main document. In some sense, a consumer would like a document to be "personalized."

These requirements motivated the proposal of a communicability framework, which we now discuss in detail.

A Framework for Communicability of Descriptive Markup-Based Software Documentation

The discussion of software documents and their representations that follows is based on the framework given in Table 1.

Semiotics (Nöth, 1990) is concerned with the use of symbols to convey knowledge. From a semiotics' perspective, a representation can be viewed on three interrelated levels: *syntactic, semantic,* and *pragmatic*. Our concern here is the pragmatic level, which is the practical knowledge needed to use a language for communicative purposes. Indeed, the goal of pragmatic quality is comprehension (Lindland, Sindre, & Sølvberg, 1994), and communicability is a prerequisite to that.

We acknowledge that there are time, effort, and budgetary constraints on producing a software document. We therefore include *feasibility*, a part of decision theory, as an all-encompassing factor to make the representation framework practical.

We now consider other aspects of the framework in more detail, starting with the principles underlying documents and their descriptive markup realizations.

Document Engineering Principles with a Descriptive Markup Perspective

The principles presented here are inspired by the established software engineering principles (Ghezzi, Jazayeri, & Mandrioli, 2003).

- **[DEP1] Separation of Concerns:** There are various concerns in document production and delivery, and to manage them effectively, each of these semantically different concerns need to be addressed separately. In particular, separation by parts, time, quality, and views are of special interest. SGML/XML provide means for separating structure, presentation, and logic in a document.
- **[DEP2] Abstraction:** This principle is based on the idea that it is often necessary in documents to highlight only the essentials while suppressing the details. The provision for table of contents and index of terms in a document are examples of abstraction. In descriptive markup, this could be realized by using mnemonic labels `<section>` and `<term>` to encapsulate the name of a section and a special term, respectively and

Table 1. A high level view of the elements of a framework for communicability of software documentation

Software Document		
Pragmatic goal	Communicability	Feasibility
Quality attributes of concern	Evolvability, Heterogeneity, Interoperability, Processability, Renderability, Traceability, Universality	
Engineering principles	Abstraction, Anticipation of Change, Formality, Generality, Incrementality, Modularity, Separation of Concerns	
Representation model	Descriptive Markup	

collating such appearances to obtain the desired result. Other examples in SGML/XML include the use of comments (`<!-- ... -->`) and in general metadata, the use of entities, or the use of a style sheet to hide content that one typically does not want rendered.

- **[DEP3] Modularity:** This principle is one way to realize separation by parts and therefore a special case of [DEP1] but is significant enough to be included separately here. [DEP2] is a prerequisite for modularity, which is essential for flexibility of future documents (Malloy, 2005). SGML and particularly XML provide various opportunities for modular documents, for example, by basing OHCO structure on a parent-child hierarchy, by including the concept of entities, or by "hiding" attributes inside the definition of an element in a document.
- **[DEP4] Anticipation of Change:** This principle is based on the premise that change in a (nontrivial) document is inevitable and to accommodate change we must be adequately prepared. By being text based (rather than binary) and being vendor and device independent, SGML/XML support this principle.
- **[DEP5] Incrementally:** This principle is related to [DEP1] and [DEP4] and is based on the notion that creating a representation of a document in increments should be encouraged. SGML/XML are in agreement with this. Indeed, it is a common markup practice to develop documents iteratively (in manageable steps) and check for conformance after each iteration.
- **[DEP6] Generality:** This principle is based on the assertion that the more general a document representation is, the broader the audience it can reach and the more reusable in different situations it is. SGML/XML are neutral to user context, vendor, domain, device, network, or programming language. They also support the largest character set currently available, namely the Universal Character Set (UCS)/Unicode.
- **[DEP7] Formality:** This principle, by requiring a formal (logical) syntax and semantics, aims to reduce the potential of ambiguity, contradictions, or misinterpretations and enables a more precise description of documents. XML has a well-defined syntax and semantics (Renear, Dubin, Sperberg-McQueen, & Huitfeldt, 2002). The Document Type Definition (DTD) provides a grammar for structural and data type constraints on the syntax and content of the elements and attributes in SGML/XML documents. In the case of XML, the capabilities of DTD have been strengthened in other grammar languages such as XML Schema and RELAX NG.

Attributes Impacting Communicability of a Software Document

In the following discussion we consider communicability as a "meta-concern" and discuss the low level list of attributes to address it. We also highlight relevant relationships between these attributes and with the abovementioned principles where needed.

- **[CA1] Evolvability:** This attribute especially depends on [DEP1-5]. Once a software document is created, it is highly likely that it will be maintained for corrective, adaptive, or perfective purposes. For example, a user manual may need to be transformed on computers with different capabilities for users with individual needs. The Resource Description Framework (RDF) and Dublin Core Metadata Element Set (DCMES) provide support for metadata (such as author information, date/time, history, or versioning) that can help track modifications.
- **[CA2] Heterogeneity:** This attribute especially depends on [DEP4-6] and is related to [CA3]. The elements of a document may need to be represented in a variety of different forms such as text, graphic, mathematical symbols, and so forth, and subsequently aggregated. Therefore, a representation must accommodate the possible compound or heterogeneous nature of a document. SGML/XML documents can be heterogeneous, where fragments of different markup (or even nonmarkup as long as the characters in it do not violate the document's character encoding) could be placed in a single document. XML Inclusions (XInclude) allows multiple possibilities of reuse: using a fragment of another document, a fragment of the current document, or an entire document (Figure 1).
- **[CA3] Interoperability:** This attribute is related to [CA2]. The heterogeneous forms in a document not only need to coexist in the same information container but also need to "talk" (interface) with each other and with the parent document. Namespaces in XML is a mechanism for uniquely identifying XML elements and attributes of a markup language, thus making it possible to create heterogeneous documents that unambiguously mix elements and attributes from multiple different XML documents (Figure 2).
- **[CA4] Processability:** This attribute depends on all [DEP1-7] and supports [CA5]. At times, we may need to manipulate (recast/transform, extract/filter, query, and so on) software documents to suit different circumstances. Support for querying XML documents is provided by XQuery and client- or server-side tree-based processing of XML documents is enabled by the Document Object Model (DOM) for which stable implementations are available. Extensible

Figure 1. Opportunities of reuse for a software document in an XML environment

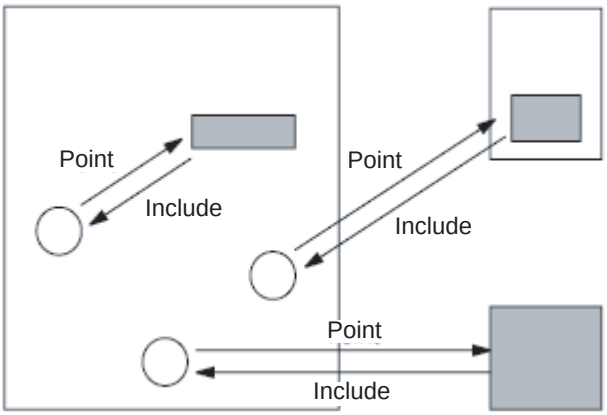
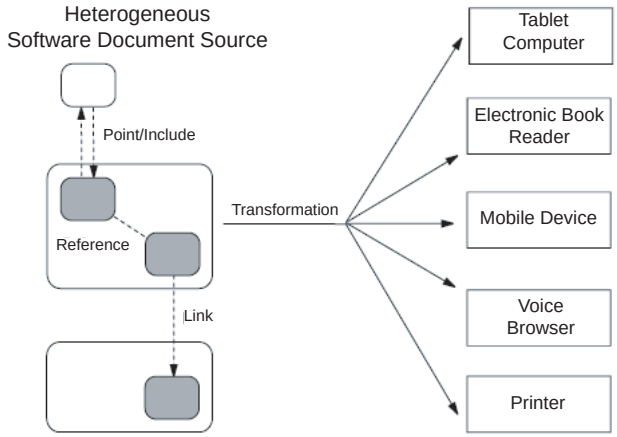


Figure 2. A heterogeneous software document in a descriptive markup transformation environment



Stylesheet Language (XSL) is a style sheet language for associating presentation semantics with arbitrarily complex XML documents, while its companion XSL Transformations (XSLT) is a style sheet language for transforming XML documents into other documents, including nonXML, documents (Figure 2).

- **[CA5] Renderability:** This attribute especially depends on [DEP1-2]. A software document must ultimately be rendered to a user's environment, and therefore presentation semantics (such as fonts, horizontal and vertical layout, pagination, and so on). The Document Style Semantics and Specification Language (DSSSL) is a style sheet language for SGML documents (that inspired XSL/XSLT). The Cascading Style Sheets (CSS) is a style sheet language for presenting simple XML documents on devices and agents with a variety of different configurations such as in Figure 2.
- **[CA6] Traceability:** This attribute depends on [DEP1,3]. A trace from `A` to `B` requires identification and location of, and means to reach, `B` from `A`. XML provides an `id` attribute for local identification of an element that can be used in conjunction with the Uniform Resource Identifier (URI) for global identification. HyTime, an ISO Standard, provides sophisticated linking functionalities for SGML documents including multi-directional links and linking to nonSGML data such as video clips. XML Linking Language (XLink) extends the unidirectional linking support in HTML and provides powerful bidirectional linking capabilities necessary for hypertext.
- **[CA7] Universality:** This attribute depends on [DEP6], and its underlying premise is that software documents should be created for all, irrespective of their individual context. XML is in agreement with the standards for accessibility and internationalization.

Application of the Framework: The Case of Agile Documentation

The bureaucracy inherent to traditional predictive software development processes and their perhaps over emphasis on documentation has in the last decade led to a shift to adaptive environments such as Extreme Programming (XP) and the Unified Process (UP). Although they are not document driven, documents continue to play an important role in stakeholder communication.

The term *agile document* was introduced in (Ambler, 2002) to imply customer-oriented, lightweight documents that could serve XP and UP. A collection of patterns for writing effective agile software documentation is given in (Rüping, 2003). Many of the patterns in it that deal with presentation and representation of documents can be concretely dealt with in our framework. For example, communicability aspects of structuring individual documents patterns are given by [CA1-7], layout and typography patterns are subsumed by [CA4, 5], whereas infrastructure and technical organization patterns are covered by [DEP1] and [CA1-4, 6].

Challenges to the Descriptive Markup Approach to Software Documentation

There are a few technical challenges to representations in descriptive markup. For example, descriptive markup in its source form, particularly when there are complex (usually

nonlinear) structural relationships involved, can be error prone for direct authoring and is not considered very readable. There are openly available and mature SGML/XML authoring tools that ameliorate this to a certain extent. The processing based on DOM or transformations based on XSLT lead to in-memory tree, which is not always efficient for large documents for demanding situations (say, being transformed over a low bandwidth network in real time). One solution to this approach has been to use XSLT compilers instead of interpreters. Similarly, event-based processing such as Simple API for XML (SAX) could complement the DOM.

FUTURE TRENDS

As software documents become increasingly large in size and number and aim to serve diverse users, a systematic document engineering process is highly desirable. There are currently limited efforts in that direction (Glushko & McGrath, 2005).

XML, like any other technology, has its own set of issues if not used appropriately while authoring or processing (Harold, 2003) or if used beyond its scope (Megginson, 2005). In fact, addressing the issue of quality in all early software representations is crucial. To do that, an evaluation of XML using the Cognitive Dimensions of Notations (CDs) (Green, 1989), a generic framework for describing the utility of information artifacts by taking the system environment and the user characteristics into consideration, would be of interest.

A natural amplification of the previous discussion is a closer synergy with the current knowledge representation initiatives that are based on descriptive markup. The Semantic Web has recently emerged as an extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning (Hendler, Lassila, & Berners-Lee, 2001) and could provide much more powerful representations of knowledge in a software document.

Finally, in spite of its broad use, documentation is not automatically useful by itself: Its value is realized only if it is well done (Weinberg, 1998), which could be viewed as one of or a combination of more than one dimensions of communicability. As documents become more like interactive applications, where they act as *user interfaces* to services provided by the software, the significance of assuring and evaluating their communicability will only increase.

CONCLUSION

Representation of documents so as to assure their communicability to the participants at all levels is critical. One way to foster that is by *partitioning* communicability into a manageable number of dimensions that could be dealt with directly and by choosing a suitable means of representation of software documents such as descriptive markup. The principles and attributes presented in our framework provide a basis for communicability dimensions and the SGML/XML *family* provides established technologies for descriptive markup.

The most important requirement for communicability from both the producer's and the consumer's standpoint is that they should have *options*. Indeed, if software documents are seen as carriers of commonsense knowledge (Minsky, 2000), then we should not seek one uniform way to present or represent them. This is echoed in the Redundant Recoding principle (Green, 1989), which is the ability to express information in a representation in more than one way, each of which simplifies different cognitive tasks. For that, documents need to be systematically created and strive for high quality. They also need to become *intelligent*, which in computational context usually implies that they carry *knowledge* amenable for automated processing. The Semantic Web provides a vehicle for representing knowledgeable software documents in descriptive markup.

REFERENCES

Ambler, S.W. (2002). *Agile modeling: Effective practices for extreme programming and the unified process*. John Wiley & Sons.

Clements, P., Bachmann, F., Bass, L., Garlan, D., Ivers, J., Little, R., et al. (2002). *Documenting software architectures: Views and beyond*. Addison-Wesley.

Coombs, J. H., Renear, A. H., & DeRose, S. J. (1987). Markup systems and the future of scholarly text processing. *Communications of the ACM, 30*(11), 933-947.

DeRose, S. J., Durand, D. G., Mylonas, E., & Renear, A. H. (1997). What is text, really? *Journal of Computer Documentation, 21*(3), 1-24.

Furuta, R., Scofield, J., & Shaw, A. (1982). Document formatting systems: Survey, concepts, and issues. *ACM Computing Surveys, 14*(3), 417-472.

Ghezzi, C., Jazayeri, M., & Mandrioli, D. (2003). *Fundamentals of software engineering* (2nd ed.). Prentice Hall.

Glushko, R. J., & McGrath, T. (2005). *Document engineering*. Cambridge, MA: MIT Press.

Goldfarb, C. F. (1981, June 8-10). A generalized approach to document markup. *Proceedings of the ACM SIGPLAN SIGOA Symposium on Text Manipulation*, Portland, (pp. 68-73).

Green, T. R. G. (1989). Cognitive dimensions of notations. In V. A. Sutcliffe & L. Macaulay (Eds.), *People and computers* (pp. 443-460). UK: Cambridge University Press.

Harold, E. R. (2003). *Effective XML*. Addison-Wesley.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The Semantic Web. *Scientific American, 284*(5), 34-43.

Knuth, D. E. (1992). *Literate programming* (CSLI Lecture Notes, Number 27). USA: Stanford University, Center for the Study of Language and Information.

Lindland, O. I., Sindre, G., & Sølvberg, A. (1994). Understanding quality in conceptual modeling. *IEEE Software, 11*(2), 42-49.

Malloy, T. (2005, November 2-4). The future of documents. *Proceedings of the 2005 ACM Symposium on Document Engineering (DocEng 2005)*, Bristol, UK, (pp.1-1).

Megginson, D. (2005). *Imperfect XML*. Addison-Wesley.

Minsky, M. (2000). Commonsense-based interfaces. *Communications of the ACM, 43*(8), 66-73.

Mundle, D. (2001, May 15). Using XML for software process documents. *Proceedings of the XML Technologies and Software Engineering (XSE 2001)*, Toronto, Canada.

Nöth, W. (1990). *Handbook of semiotics*. Bloomington: Indiana University Press.

Renear, A., Dubin, D., Sperberg-McQueen, C. M., & Huitfeldt, C. (2002, November 8-9). Towards a semantics for XML markup. *Proceedings of 2002 ACM Symposium on Document Engineering (DocEng 2002)*, McLean, USA (pp. 119-126).

Rüping, A. (2003). *Agile documentation: A pattern guide to producing lightweight documents for software projects*. John Wiley & Sons.

Spiel, C. (2002). Writing documentation, Part III: DocBook/XML. *Linux Gazette, 75*.

Weinberg, G. M. (1998). *The psychology of computer programming (silver anniversary edition)*. New York: Dorset House.

KEY TERMS

Agile Document: A customer-oriented lightweight document that need not be perfect but just good enough.

Descriptive Markup: A model of text that focuses on the description of information using markup delimiters for consumption by both humans and machines.

Document Engineering: A discipline that is concerned with principles, tools, and processes that improve the ability to create, manage, and maintain documents.

Knowledge Representation: The study of how knowledge about the world can be represented and the kinds of reasoning that can be carried out with that knowledge.

Ontology: An explicit, formal specification of a conceptualization that consists of a set of terms in a domain and the relations among them.

Semiotics: The field of study of signs and their representations.

Single Source Approach: A technique that encourages a once-only creation of a resource, such as a document, in a manner so that it could be reused or repurposed for different contexts.

Framing Political, Personal Expression on the Web

Matthew W. Wilson
University of Washington, USA

INTRODUCTION

The World Wide Web, as a collection of Web sites, Web services, and Web-enabled technologies, is a space of expression and contestation—a social construction of sorts. Additionally, the Web, as a locus of investigation, is gaining attention from scholars in the social sciences, feminist and critical theorists, as well as more recent poststructural reconceptualizations across many disciplines. One unifying interest is precisely the topic of this article: How might we recognize what is considered *political* and *personal* in a virtual space? To what sense can we distinguish political and personal expression online? This article frames the diverse perspectives for interrogating political and personal expression on the Web, while offering considerations for why these sorts of projects are at all necessary or useful.

The determinacy of virtual, Web-based locations as political and/or personal is a complex endeavor. Does a pro-choice posting to an anti-abortion online discussion group constitute a political act? What is potentially meant by "political"? Several discussion forums or news groups contain categories like "politics" or "government and politics" (see *Yahoo! Groups* for example); and yet, such groups may or may not be perceived as "political". This perception of "being political" is dependent on certain philosophical tensions about what can be considered political in certain spaces and times. Other Web sites seek to build politics through the Web, via such movements as e-democracy, online deliberation, or public participation geographic information systems (Davies & Novack, forthcoming; Dragicevic & Balram, 2006). However, while building politics is certainly political, surficial analysis of such online-coalition building endeavors may resist or gloss the multiple political implications for constructing a politics. Therefore this entry contains a discussion of politics and "the political"; each as a perspective has certain methodological and empirical contingencies. Namely, how do we study online interactions? What sorts of data might we collect? Furthermore, how are we, as researchers, already implicated in our studies of online interactions?

This entry proposes a diversity of approaches in studying interactions within the Web as informed by both the information sciences and the humanities and is organized into four sections: first, a background section which contemplates more traditional debate in political theory made relevant to studies of the Web; a second section which proposes (post)modernist and poststructuralist framings for researching personal and political expression; third, a section offering future research questions in this research area; and finally, conclusions that reflect upon research on the Web.

BACKGROUND

While certain academic traditions analyze political encounters as separate from those situations that are supposedly personal, critical interruptions in these traditions have shown that the personal and political are quite interrelated and inseparable, if not fictitious designations of expression. Most obvious of these critical interruptions are the women's and GLBT (gay, lesbian, bisexual, and transgender) movements, with the blurring of these boundaries as central to an active social movement. These interruptions highlight the co-constitutiveness of the political and the personal, and render suspect analyses that enforce a strict dualism. Therefore, it may be as no surprise that analyses of interactions on the Web are similarly complex—given that the Web is simultaneously coded as personal and political, public and private. Identity and embodiment—or what constitutes the "self"—are intriguingly negotiated in virtuality, making a mess of any rigid enforcement of public/private, political/personal evaluation of online culture. This section explores these public/private, political/personal frameworks as a background for studies of interaction on the Web.

Publicness and Privateness

Social interaction over the Web, or what Steven Johnson (1997) argues is an "interface culture", is acutely part of everyday life, although admittedly this is not the case for everyone. Out-of-the-box high-speed, wireless networking allows computers and mobile-computing devices to find use across the space of the home, from the kitchen refrigerator to the bedroom Web-enabled television, and throughout the local neighborhood, within grocery stores, cafés, and

while walking the sidewalks of busy metropolitan streets. Johnson's central thesis is that our contemporary culture is one of *interfaces*—a rearranging/repackaging/refiltering of digital data to serve various information consumers. Web sites like *WifiMug.org* point computer-savvy-sans-caffeine users to local coffeeshops, which have free wireless Internet access, and handy power outlets. Others, like *BlogHer.org*, collect women-authored Web-logs (or blogs), to build a sense of "women-friendly" online political community. What about contributing to online discussions on transportation planning while sitting in bed could be experienced or reinforced as a public, political act within a place coded as private? Furthermore, how might we theorize the privateness of certain discussions of personal violence, in a supposedly public, online forum? The Web is certainly a site where that which is coded public and private meets in virtual contradiction. Analyses of online interaction that seek to code separately what is considered public from private, are engaged in projects that frame their observations by a certain transformation of expression, *pace* Habermas (1979), from the private to the public individual. The Internet, some argue, enables this transformation of expression, from private mobile devices into actively participating citizens of a public (Stromer-Galley, 2003a, 2003b). However, these sorts of framings are a slippery slope, susceptible to a problematic dualism of the political and personal.

Political and Personal

Central to the popular critique that "the personal *is* political", feminists have challenged the public-private dualism as misrecognition of the political status of supposedly private, personal acts. As Nancy Fraser (1992) argues, that which is "private" is conceptualized within public-private frameworks as some "prepolitical starting point" (p. 130). The moment requiring critical intervention occurs when the public-private dualism discriminates personal claims as separate from and irreconcilable with political claims. Feminist political geographers Michael Brown and Lynn Staeheli (2003) offer a framework to organize these diverse critical interventions into three moments of "the political": the distributive, the antagonistic, and the constitutive. These three moments are essentially research perspectives, and can be appropriately translated to online interaction research. The *distributive* is an interrogation of which individuals and groups have unequal access to certain resources; studies of Internet and interface accessibility are good examples of the distributive approach (e.g., Servon, 2002). The *antagonistic*, Brown and Staeheli continues, is a recognition of the oppositional conflict which structures the political, similar to studies of partisan-politic Web forums (e.g., Stromer-Galley, 2003a). Finally, the *constitutive* research perspective looks to analyze the productive dimension of the political; research which analyzes the assertion of identity by appropriating online discussion forums might take a more constitutive approach (e.g., Wincapaw, 2000). Each of these perspectives on political research is also interestingly contingent on the collections of interfaces, protocols, software, and hardware within our constructed and engineered (Web-based) spaces of interaction. The following section discusses these emerging spaces of interaction, and then proposes two ways of framing expression within these spaces.

METHODOLOGICAL-THEORETICAL FRAMINGS FOR INTERNET RESEARCH

Methods of analysis of political and personal expression over the Internet are especially complex due to the dynamic and differential development of Web-based systems. Those interested in social studies of the Internet should consider new developments in interaction design, particularly the advancement of *Web 2.0*, but must remain conscious of the differential development of Web-based systems, from the simple-yet-elegant, HTML Web sites (such as *Craigslist.org*) to the more interactive, AJAX Web sites (such as *Flickr.com*). The Web, with its various interfaces of carefully designed interaction, privileges, and constrains various methods of online research (Cherny & Weise, 1996; Crampton, 2003; Davis, 1999; Graham, 1998; Kitchen, 1998; White, 2006). Web developers explicitly construct certain operations and logs of data collection within their Web sites; researchers may or may not have access to these datasets, and furthermore, may or may not be able to analyze the data even as it becomes accessible. However, beyond the data acquisition dilemma for online interaction research, researchers should additionally consider the nuanced implications of interface and interaction design.

Web sites have become more interactive, with recent developments such as *Flickr*, *Blogger*, and the suite of *Google* Web-applications. Many of these more recent developments fall under the paradigm of *Web 2.0*. This ethos of Web development focuses intensely on the design of interactions—leading toward not only technical or programmatic changes in Web software development, but also (more importantly) advances new concepts in user interaction (Garrett, 2002, 2005). Web sites architected from a notion of *Web 2.0*, follow design protocols that create a seamless experience for users; Web pages have the look and feel of a computer desktop application, with interactive menus and nearly instantaneous feedback. Clicking controls on the Web page no longer require jarring page reloads; instead, components within a Web page update and change, never requiring the user to adjust to a new page or layout. Beyond aesthetic and cinematic changes to the interaction design, *Web 2.0* also introduces such technologies as RSS (Rich Site Summary) feeds, which employ XML (Extensible Markup Language) standards, allowing users to be served with vari-

ously tailored information content, instead of the "hide and seek" paradigm of more traditional Web design.

Given these various nuances in Web construction (albeit differentially), with new interfaces and new ways of interacting, the opportunities for studying expression are diverse. With better systems of interaction come enhanced user experiences (Garrett, 2002). Although this equation of enhanced experience with more application-like Web environments might be problematic (or indicative of more *distributive*, political research), researchers have the opportunity to explore how these enhanced online interactions shift or enable identity formation in different ways (Balsamo, 1996; Foster, 2005; Gray, 1995; Halberstam & Livingston, 1995; Haraway, 1997; Schuurman, 2002; Turkle, 1995). As these virtual spaces become impacted by new concepts in Web design, political inquiry becomes impacted by new ways of "being political" online. Unfortunately, research to date has not explored these social and political implications for the *Web 2.0* design/development movement and the sorts of expression that are accordingly enabled or disabled (however, scholars are beginning to recognize this; see Erickson, 2007). The following sub-sections propose two broad methodological-theoretical approaches for investigating these expressions.

(Post)modernist Perspectives

The interpretation of human behavior on the Web is subject to certain philosophical traditions, whether explicit or implicit. In other words, we must ask an obvious, yet revealing, question of our observations of online behavior: "how can we know this?" If our online selves are fully knowable to ourselves, and thereby to others, then certain methodologies are applicable for interrogating these senses of virtual self. In other words, epistemologies of behavior and the "self" inform the data we collect and interpret. Particularly *modern* of these epistemologies of behavior and self is the observational behavior, or behavior that can be directly observed (or indirectly observed through proxy or mediator). These sorts of observations could be characterized as modern, as part of the progression of modernity, by continually adding to our knowledge of our increasingly comprehendible self, through actual, observed behavior.

> *I am looking for numbers. Scientific surveys, focus group, usability studies, Internet traffic statistics, log file analysis etc.—data that helps give a quantifiable sense of* what people say they want *with the Internet and politics/government/media as well as* what people actually do *online in those areas.* (Clift, 2001, emphasis mine)

Steven Clift, one leading figure in the e-democracy movement, posted the previous statement to his online democracy listserve. This is an appropriate example of observational behavior, or "what people say they want" and "what people actually do". Clift is calling for "numbers"—for quantified, observed behavior. It is important to note here that "numbers" and statistics obtained about online behavior are not necessarily indicative of a *modernist* approach; instead, it is how these data are employed— "how we say we know this to be what it is" —that makes these data subject and suspect to foundational critique (Lawson, 1995).

What then of *postmodern* approaches to research? The *post* indicates an important nuance in the interpretation of these "numbers" —namely, that any observation of behavior (online or otherwise) is always-already within a "hall of mirrors". This "hall of mirrors" metaphor underlines the partiality of all observation, interpretation, and representation—that all observation is confined, in some sense, by an amusement park "filled with hucksters and con artists" (Shaviro, 1995, p. 44). Data are always framed and relational; there is no observational, transferable, or generalizable truth to human behavior. Regardless of foundational perspective or belief, there is always data. Data collection that interrogates expression on the Web might include methods of collecting user data (profiles, geographic-demographic positioning, etc.), as well as collecting "user-using-Web site" data (logs of system use, including operations and queries performed, travel to/from hyperlinks, etc.) (e.g., Golder & Huberman, 2006). Other more "interpretive" data collection includes user interviews (focus groups, surveys, etc.) and user ethnographies (e.g., Stromer-Galley, 2003a; Wincapaw, 2000). System developers also look for broader Web site data, including performance statistics and traffic-time profiles. Again, the power in representing this data relies upon certain philosophical underpinnings. A post-structural approach works to reveal these underpinnings, through continual reflection on these projects of inquiry.

Post-Structuralist Perspectives

What might a post-structural perspective bring to projects around political and personal expression on the Web? *Post-structuralism* refers specifically to a theoretical approach informed by French theorists (notably, Michel Foucault), and more broadly serves to identify approaches which challenge and complicate structuralist and foundationalist accounts of ideology, hegemony, subjectivity, resistance, and dominance. The Web, under this perspective, is an immense discursive and material formulation, constituting new identities and subjects through various operations of power (Foucault, 1990 [1978]). Furthermore, the Web is a "space" in precisely the sense of the poststructural, feminist geographer Doreen Massey (2005): as "constituted through interactions" and thereby reliant upon the "existence of multiplicity" (p. 9). Web developers and designers, users, lurkers, and researchers negotiate this (virtual) space of multiplicities, each of which have partial and hybrid identities/subject positions.

Power operates, from a Foucaultian perspective, upon these identities or subject positions to constitute a sense of self and other—even as they are already mediated by their status as a technician, designer, user, and/or researcher.

Specific to studies of online interaction, one thread of research (possibly described as posthuman or cyberpunk or technoculture studies[1]) takes up this topic of the fractured online identity, and employs post-structural methods (genealogies, power-geometries, discourse analyses) to interrogate the "cyborg" as a particular politicized figure of postmodern, technicized society. The cyborg, as theorized by Donna Haraway (1991) and further extended by feminists and others in science (and science fiction) studies, is the figure of fragmented embodiment—a blurring of human and technology, where "mind, body, and tool are on very intimate terms" (p. 165), in some cases a literal merger of the body with various prostheses and other techno(bio)logical enhancements. Tom Foster (2005) describes this blurring as indicative of a posthuman world, where:

technology does in fact "become me" not by being incorporated into my organic unity and integrity, but instead by interrupting that unity *and opening the boundary between self and world. The point ... is not to reject the value of embodiment, however, but to make it possible to rearticulate the relationship between mind and body, "inner being" and external form, in more complex and diverse ways, by revealing the ways in which that relationship was always mediated socially.* (p. 10, emphasis mine)

Foster's argument extends Haraway's post-gender cyborg as a subject position that interfaces through communication prostheses—especially that found through the Internet. Interfaces support these moments of "interrupted unity", and thereby beg to be problematized to uncover the power operations that constitute certain identities in particular (virtual) spaces. The following section, proposes future research questions which are informed by both post-structural and (post)modern perspectives.

(Post-)modern and post-structural perspectives provide the theoretical and methodological framings for studying political and personal expression on the Web. These perspectives are not necessarily oppositional, nor incongruent. Instead, continued research on the Web will draw from a diversity of perspectives; this article has hopefully served to articulate two such framings.

FUTURE TRENDS

As disciplines find research about the Web to be relevant to studies of human behavior and interaction, there shall be growing interest in how to frame these explorations. Interactions on the Web are insistently political even as they are coded personal. This political-ness has been categorized as being distributive, antagonistic, and constitutive (Brown & Staeheli, 2003). As such, future research questions are offered below, categorized by these three moments of the political:

1. **Distributive.**
   - How does *Web 2.0* Internet design influence accessibility for underrepresented groups?
   - Which online community designs promote inclusive versus exclusive membership guidelines? How are these norms and regulations enforced?
   - How might collaborative tagging systems enable or constrain online interaction?
   - How are coalition-building Web sites able to affect policy formation?
2. **Antagonistic.**
   - How do "government and politics" discussion forums construct an immediate other? For what purpose?
   - To what extent are discussion forums largely homogenous in affiliation or commentary?
   - Is 'community' really the appropriate theoretical metaphor for online discussion and coalition-building Web sites?
3. **Constitutive.**
   - How might the purpose of an identity politics discussion forum be multiple? How do members of this discussion forum negotiate the political potential of such a Web site?
   - How might collaborative tagging systems or folksonomies contribute to the shared ontologies of Web site content?
   - How do users perceive their interactions within discussion forums as (un)useful to other users?
   - What emerges from online interaction studies? How do these studies in turn enact new forms of online interaction or interfaces?

CONCLUSION

The study of political and personal expression is dependent on how one understands "the political". This entry has reviewed how "the political" has been problematized, and has proposed two framings for expression on the Web: (post-)modern and post-structural approaches. Students and researchers should make use of this entry as a re-mapping of the topic—to push the boundaries of what is potentially meant by expression over the Internet. Indeed, there are new technological concepts, such as *Web 2.0*, that structure the online interaction between users, and certainly "change" is the status quo for

the Web. However, such a dynamic environment should simply enrich the various studies of interaction from diverse disciplinary and theoretical perspectives—some of which touched on previously (if a bit prescriptive). To conclude, I present some final thoughts which expand upon the topics and problematics from the previous sections.

Why is it interesting to study online interaction? The answer to this question is (of course) multiple. Depending on your perspective (a few of which are discussed earlier), online expression is interesting because it tells us something general about human behavior, as well as something about our specific, online selves; or maybe, online expression is interesting precisely because it has grown as a site of inquiry across various disciplines, drawing attention from the sciences and the humanities, not to mention private industry. Online expression can be approached as a culture, a technical achievement, a public sphere, a site of resistance or discrimination, or a virtual community. Due to the various positionings of this topical area from multiple disciplines, we must be more explicit about our politics in conducting this research—as the subjects of our research continue to virtually express themselves politically and personally.

REFERENCES

Balsamo, A. M. (1996). *Technologies of the gendered body: Reading cyborg women*. Durham: Duke University Press.

Brown, M., & Staeheli, L. A. (2003). "Are we there yet?" Feminist political geographies. *Gender, Place and Culture, 10*(3), 247-255.

Cherny, L., & Weise, E. R. (1996). *Wired women: Gender and new realities in cyberspace*. Seattle, Wash. [Emeryville, CA]: Seal Press; Distributed to the trade by Publishers Group West.

Clift, S. (2001). *Surveys on Internet and Elections/Governance—"Numbers" research request*. Retrieved August 31, 2006, from http://www.mail-archive.com/do-wire@tc.umn.edu/msg00230.html

Crampton, J. W. (2003). *The political mapping of cyberspace*. Chicago: University of Chicago Press.

Davies, T. (Ed.). (forthcoming). *Online deliberation*.

Davis, R. (1999). *The Web of politics: The Internet's impact on the American political system*. New York: Oxford University Press.

Dragicevic, S., & Balram, S. (Eds.). (2006). *Collaborative geographic information systems*. Hershey, PA: Idea Group, Inc.

Erickson, K. (2007, 1 July). *Web 2.0 is broken and here's why*. Retrieved July 14, 2007, from http://www.profy.com/2007/07/01/Web-20-is-broken-and-here-s-why/

Foster, T. (2005). *The souls of cyberfolk: Posthumanism as vernacular theory*. Minneapolis: University of Minnesota Press.

Foucault, M. (1990 [1978]). *The history of sexuality, vol. I: An introduction* (R. Hurley, Trans.). New York: Vintage Books.

Fraser, N. (1992). Rethinking the public sphere: A contribution to the critique of actually existing democracy. In C. Calhoun (Ed.), *Habermas and the public sphere* (pp. 109-142). Cambridge, MA: The MIT Press.

Garrett, J. J. (2002). *The elements of user experience: User-centered design for the Web* (1st ed.). Indianapolis, IN: New Riders.

Garrett, J. J. (2005). *Ajax: A new approach to Web applications*. Retrieved July 26, 2006, from http://www.adaptivepath.com/publications/essays/archives/000385print.php

Golder, S., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science, 32*(2), 198-208.

Graham, S. (1998). The end of geography or the explosion of place? Conceptualizing space, place and information technology. *Progress in Human Geography, 22*(2), 165-185.

Gray, C. H. (1995). *The cyborg handbook*. New York: Routledge.

Habermas, J. (1979). What is universal pragmatics? (T. McCarthy, Trans.). In *Communication and the evolution of society* (pp. 1-68). Boston: Beacon Press.

Halberstam, J., & Livingston, I. (Eds.). (1995). *Posthuman bodies*. Bloomington: Indiana University Press.

Haraway, D. J. (1991). *Simians, cyborgs, and women: The reinvention of nature*. New York: Routledge.

Haraway, D. J. (1997). *Modest_Witness@Second_Millennium. FemaleMan©_Meets_OncoMouse™: feminism and technoscience*. New York: Routledge.

Johnson, S. (1997). *Interface culture: How new technology transforms the way we create and communicate* (1st ed.). San Francisco: HarperEdge.

Kitchen, R. M. (1998). Towards geographies of cyberspace. *Progress in Human Geography, 22*(3), 385-406.

Lawson, V. (1995). The politics of difference: Examining the quantitative/qualitative dualism in post-structuralist feminist research. *The Professional Geographer, 47*(4), 449-457.

Massey, D. (2005). *For space*. London; Thousand Oaks, CA: SAGE.

Schuurman, N. (2002). Women and technology in geography: A cyborg manifesto for GIS. *The Canadian Geographer, 46*(3), 258-265.

Servon, L. J. (2002). *Bridging the digital divide: Technology, community, and public policy*. Malden, MA: Blackwell Pub.

Shaviro, S. (1995). Two lessons from Burroughs. In J. Halberstam, & I. Livingston (Eds.), *Posthuman bodies* (pp. 38-54). Bloomington: Indiana University Press.

Stromer-Galley, J. (2003a). Diversity of political conversation on the Internet: Users' perspectives. *Journal of Computer-Mediated Communication, 8*(3).

Stromer-Galley, J. (2003b). Voting and the public sphere: Conversations on Internet voting. *PS: Political Science and Politics, 36*(4), 727-731.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon & Schuster.

White, M. (2006). *The body and the screen: Theories of Internet spectatorship*. Cambridge, MA: MIT Press.

Wincapaw, C. (2000). The virtual spaces of lesbian and bisexual women's electronic mailing lists. In G. Valentine (Ed.), *From nowhere to everywhere: Lesbian geographies* (pp. 45-60). New York: Harrington Park Press.

KEY TERMS

Cyborg: The cyborg is a figure deployed most popularly by Donna Haraway to depict the partial subjects of post-modern technoculture. The cyborg is a method of critique, allowing theorists to challenge nature/human, mind/body, virtual/material dualisms.

Epistemology: Epistemology is the study of how we are able to know what we know. As a way to study knowledge, epistemological critique challenges assumptions about truth, causal relationships, and evidence.

Foucaultian Perspective: Foucaultian perspectives are post-structuralist critique specific to Michel Foucault, a French political theorist. Foucault's writings have been used in various academic disciplines and political projects. His interest in power is well cited as a methodology exploring processes of subject formation through normalization.

Postmodernism: Postmodernism is a category of thought within literature, architecture, and the arts, emphasizing complexity, chaos, and indeterminacy as a response to the modern tradition which emphasizes progress, determinism, and singular narratives. Within investigations of interaction on the Web, postmodernism informs the notion that data are always framed and relational. Postmodern perspectives advocate that there is no observational, transferable, or generalizable truth inherent to human behavior.

Post-structuralism: Post-structuralism is a response to foundationalism, concerned particularly with the work that language performs. Similar to postmodernism, post-structuralism challenges singular narratives and objectivity, but focuses on the processes of subjectification and identity formation through language.

Social Construction: Social construction is a perspective in research that argues all experience (including language, information, and knowledge) is premised on social relationships.

Web 2.0: A movement in Internet development and design, focusing on interaction and complementarity. Jesse James Garrett is a well-known advocate/critic of Web 2.0 approaches. Web sites that deploy Web 2.0 concepts usually build interactions with the "look and feel" of a computer application, using XML and Javascript (AJAX) to allow for greater responsiveness.

ENDNOTE

[1] However, note that cyberpunk, cyborg, posthuman, and technoculture studies connote different audiences, empirics, disciplinary contexts, and theoretical positionings.

Free and Open Source Software

Mohammad AlMarzouq
Clemson University, USA

Guang Rong
Clemson University, USA

Varun Grover
Clemson University, USA

INTRODUCTION

Free and open source software (F/OSS) is emerging as a promising alternative to proprietary software. The interest in F/OSS solutions is growing as firms realize that it could help reduce IT expenditures. Unfortunately, despite the heightened interest, F/OSS solutions remain misunderstood, and a number of myths regarding this approach still prevail.

F/OSS has been described as simply software, or even software specific for the Linux operating system, with hardly any reliable support available. Some have considered it a silver bullet solution that will always create superior quality software at lower or no cost (Wheatley, 2004).

The purpose of this article is to demystify these misconceptions surrounding F/OSS and provide an understanding of its basic concepts. Then, based on these concepts, we try to illustrate how companies can benefit from F/OSS. Our hope is that this article would assist interested observers to better understand F/OSS and help managers make more informed decisions regarding F/OSS solutions.

BACKGROUND

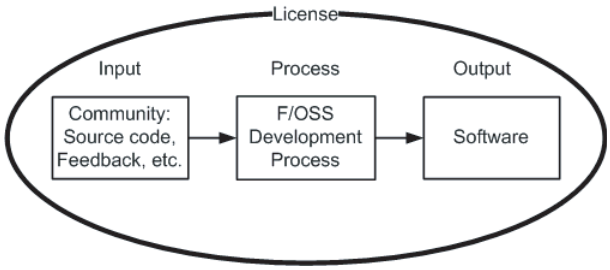
When computers first originated, the norm in most academic and corporate labs was to freely exchange programs and ideas between programmers. This was the early form of F/OSS. This norm changed when IBM unbundled its software and hardware in the 1970s. This move created a market and value for software. In order to preserve this newly found software value, software producers restricted user access to human readable source code in order to protect software secrets (Glass, 2004). This meant that users could not modify the software that they owned and more importantly, restricted the free flow of ideas. This did not fit well with many programmers, most notable was Richard Stallman from MIT's Artificial Intelligence Lab. Stallman believed strongly in the freedom of the user to use his software and created the Free Software Foundation (FSF) in 1985 to promote the development and use of free software. The following success of projects using the free software development model, such as Apache and Linux, inspired Eric Raymond to write his seminal piece *The Cathedral and the Bazaar* in 1997 that brought the attention of the corporate world to the free software development model (Raymond, 1999a). In a brain storming session on February 3, 1998 Todd Anderson, Christine Peterson (of the Foresight Institute), John "maddog" Hall and Larry Augustin (both of Linux International), Sam Ockman (of the Silicon Valley Linux User's Group), and Eric Raymond, agreed on the need for a marketing campaign to win the support of fortune 500 companies to ensure the long-term survival of the movement. The participants saw the need to use a term other than "free" which they figured would hurt the movement's chances of gaining support from the corporate world because of its ambiguous meaning. As a result of this session, the term *open-source* was coined by Christine Peterson and the Open Source Initiative Organization (OSI) was established ("History of the OSI," n.d.).

Richard Stallman of the FSF opposed this movement because in his opinion the term open-source was not pure enough. So the FSF and OSI remained separate movements promoting similar practical principles, but with different philosophies and beliefs. FSF decided to be vocal about its beliefs and maintained the *free* software label. They believed that the user's "freedom" is a priority, an end of itself, and they should be able to do whatever they want with their software. OSI on the other hand did not explicitly express the user's right for software freedom, but promoted it as a means of producing better software. The end is to get the corporate world to buy into this concept. Because of the coexistence of both philosophies, we refer to the group of software that adheres to the OSI or FSF principles as free and open source software (F/OSS). [1]

F/OSS IPO MODEL

The F/OSS development system can be conceptualized as an input-process-output (IPO) system, with the license as the boundary, the community as the input provider, the F/OSS

Figure 1. F/OSS IPO system



development methodologies as the process, and the software as the output (see Figure 1). So we would expect that all four components to have implications on the quality of the final product (software).

The License

The F/OSS license acts as the boundary that identifies the system as an F/OSS system. It specifies the terms by which the software is to be used and distributed. It serves as a governing mechanism that enforces the norms of the F/OSS community and provides motivation for programmers by protecting their efforts from appropriation (Bonaccorsi & Rossi, 2003).

There are numerous F/OSS licenses available that all maintain the openness and free redistribution of the source code.[2] The difference between the licenses reflects the philosophical differences between the FSF and OSI on how to advance the F/OSS projects. There are two principles that observers should be aware of (Lee, 1999):

- **The copyleft principle:** Software derived/revised from original F/OSS source code must remain F/OSS, and privatization of any part or whole of the program is prohibited.
- **The GPL compatibility principle:**[3] Licensed F/OSS cannot be mixed with proprietary source code.

Both the FSF and OSI have very similar criteria in qualifying licenses and there is a great deal of overlap between the approved licenses of both organizations. The fundamental difference between the two lies in the underlying philosophy. FSF recommends more open licenses that enforce both the GPL compatibility and copyleft principles, while the OSI is less stringent on this matter. Generally, an F/OSS license must allow programmers to access, modify, and redistribute the source code. F/OSS licenses do not prevent a software producer from demanding a distribution fee for his product. F/OSS licenses however prevent the software producer from placing restrictions on how the software should be used or redistributed after it is in the hands of the users ("*Selling Free Software,*" n.d.).

The Community

The community consists of all the developers and users of the F/OSS. The community and all their contributions are conceptualized as the input to the IPO system, which includes source code, documentation, and feedback (i.e., bug reports, support requests, and feature requests) (Raymond, 1999b).

The growth of the community ensures the ongoing survival of the F/OSS project and further improvement of the product. The advancement of the projects and community is dependent on the members who have the motivation and the ability to contribute. The most active of the community contributors are known as the *core*. The core is responsible for the majority of source code development. They also have the most control over the features and design of the software product. Occasional source code contributors are known as *co-developers*. They contribute by modifying or reviewing code or submitting bug fixes in addition to feedback. But the majority of the community members are the *users* who do not contribute with code submissions. Depending on the level of feedback, users can be active by providing some feedback, or passive, by providing none (Crowston & Howison, 2005).

THE DEVELOPMENT PROCESS

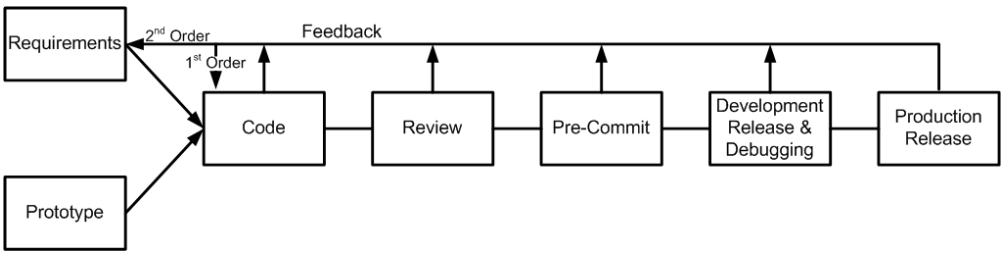
F/OSS development communities do not seem to adopt or practice traditional software development processes (e.g., the waterfall model). Many F/OSS projects begin with a prototype with predefined requirements developed from scratch or based on existent older product (Scacchi, 2002). Then, this early version incrementally evolves through rapid development iterations from the community, while concurrently managing as many designing, building, and testing activities as possible. Five main steps were identified for this approach: (1) code submission, (2) peer review, (3) precommit test, (4) development release and parallel debugging, and (5) production release (Jorgensen, 2001).

The F/OSS development process is an iterative process with feedback loops in every stage. The source code is constantly updated to meet the dynamic requirements that change along with the needs of the users (see Figure 2). These requirements are updated based on user feedback.

The Software

A continuous output of the F/OSS system is the software, which demonstrates some unique benefits when compared to proprietary software, such as:

Figure 2. The F/OSS development process



- **Reliability:** The reliability of F/OSS can be best reflected in Linus' Law, "Given enough eyeballs, all bugs are shallow." (Raymond, 1999b, p. 41). Due to a virtually unlimited number of contributors, software bugs are more likely to get uncovered. Short feedback and revision cycles also mean that improvements are incorporated quickly (Williams, Clegg, & Dulaney, 2005).
- **Security*:*** In a similar vein, community access to the source code makes it more likely to uncover and solve security problems. In contrast, securing the software by restricting access to the source code will only create a false sense of security, as it might take extended periods of time to locate and fix security holes (Raymond, 1999b).
- **Low Cost*:*** Software operation costs are drastically reduced due to the elimination of multiple licensing fees and the availability of support and updates from the community. However, when considering the total cost of ownership (TCO), training and customizing cost should also be accounted, which may vary by company (Stevenson, 2005).

All four components of F/OSS will have different implications on the quality of the software. The development process and community have a direct impact on the quality of the software. A large community will be worthless if the members are not motivated to contribute source code and feedback to the development process. The restrictions of the license will mean that only those who can accept these restrictions can contribute—essentially affecting the level and quality of contributions. Even the software design will have an effect on not only the quality of the software, but also how the community will organize around it (Crowston, Annabi, & Heckman, 2004).

SUCCESS WITH F/OSS

With a better understanding of F/OSS, we now discuss how companies can maximize the benefits gained from F/OSS depending on their specific needs. We have identified three distinct approaches centered on three components: software, community, and license. With these approaches on a rough continuum, companies can choose to have no involvement in the community or to be part of the F/OSS community core. Regardless of approach chosen, all four components of the F/OSS system should be evaluated (AlMarzouq, Zheng, Rong, & Grover, 2005).

Software-Centered Approach

The software-centered approach is about the autonomous use of the final product of the system (software). The software could be directly obtained from the community or provided by a vendor (who is usually part of the F/OSS community). It is most appropriate for companies with static and clear requirements that need well-developed software with a strict budget or time line. It focuses on utilizing the product and follows the same selection criteria as proprietary software.

In addition to reduction in time and costs in obtaining the desired software, companies can enjoy the improved reliability and security of F/OSS. However, less involvement and less input mean that the companies would have no influence on the features of the software. As passive users, companies have to accept the restrictions set up by the software and license.

Technical support or minor customization requirements also posit potential concern. Since technical support for F/OSS is dependent on the community, the evaluation should include an assessment of how responsive and supportive the community is or the availability of support vendors.

Community-Centered Approach

For the community-centered approach, the companies become more involved into the F/OSS community, and therefore are entitled to some influence on the features of the software. It is more suited when the desired software product or the requirements of the company are evolving.

The companies could get involved through this approach in two ways depending on the capabilities of the company.

They can either be active users providing feedback and feature requests on the product. Or they can, as co-developer, further contribute by providing source code contributions that would help enhance the software and make it more suitable for their use.

Becoming a co-developer has its benefits over being an active user. By participating in the development process, companies can improve the software development skills of their employees as the training process is embedded into the participation process (Lussier, 2004). Such valuable learning experience could enhance a companies' technical capability and facilitate the in-house application or customization of the software or even build competencies associated with the software that can be leveraged so the company could act as a support vendor for the F/OSS. The challenge with this approach is that companies have to devote more time and effort and possess certain technical capabilities to be able to act as an active player or co-developer in the community. Meanwhile, as players but not initiators, they still have to follow the restrictions of the licenses. Furthermore, the desired level of control over the software is not guaranteed since the company is part of a community. A potential challenge specific to becoming a co-developer with this approach is that in some cases companies have to adjust their own routines and structures to fit in the F/OSS development process. Such adaptation might introduce further costs and conflict within the company.

License-Centered Approach

The idea from this approach is to initiate an F/OSS community. As a result, the initiator has control over the choice of the terms of the license and becomes part of the core of the community, hence the term *license-centered*. An initiator can either release existing code, or embark on creating the initial prototype of the F/OSS.

Since users of the software can potentially be active members and developers within the F/OSS community, the initiating company can tighten the relationship with its customers should they get involved with the community. The customers can then be leveraged as a resource for innovative ideas, feedback, and developmental assistance to improve user satisfaction and prolong the product life cycle (Scacchi, 2004). This approach is most suited for companies seeking to improve their competitive position, especially with the existence of network effects working against the company. But given the long-term nature of the community building process, the benefits from this approach will be realized in the long run (West, 2003).

The biggest issue with having a company release software as F/OSS is that it might relinquish any competitive advantage associated with the software as the secrets of the software can be easily obtained by competitors. One solution is to "open parts," which is to release commoditized layers that are not a source of competitive advantage and retain full control of the layers that can be a source of competitive advantage. Another solution is to "partly open" the source code through using licenses to put legal restrictions that can provide value for the customer but prevent competitors from using the technology (West, 2003).

The biggest challenge that remains after releasing the software is getting the community to participate. Having the project develop into a self-sustaining one will require foresight from the company that initiated the effort in setting up the control structures and a willingness to hand off that control to the community (West & O'Mahony, 2005). The initiating company has to make a trade-off between its own interest and the growth of the community.

Another challenge is to align the company's internal development process with an F/OSS development process that enables community contributions. This may require some change management on the part of the initiating company.

Overall, the benefits and challenges of the three approaches are summarized in Table 1.

FUTURE TRENDS

As chief information officers (CIOs) get pressured to further reduce IT expenditures, we expect to see increased adoption and improved understanding of the benefits and utilization of the F/OSS system as a whole. We will not only see the success and influence of F/OSS on infrastructure-type software such as operating systems and database system (e.g., Linux and MySQL), but more complex applications. This move has already started with systems such as Compiere which is an F/OSS ERP system.

While research will continue to advance our understanding of the structure, efficacy, and strategic implications of F/OSS, there will clearly be regulatory issues (e.g., software patents) that will need resolution. It will be interesting to see how the change in legal environment will affect the F/OSS movement and whether F/OSS licenses can be effectively enforced.

CONCLUSION

Given the increasing pervasiveness of F/OSS, it is important that business managers evaluate alternatives to proprietary software and explore potential opportunities F/OSS may bring to their business. In doing this, they should understand and evaluate all the components of F/OSS that will impact the quality of the software. F/OSS is not software that is specific to an operating system, nor is it simply just software, but it is a complete system that includes the license, the community, the development process, in addition to the software.

Table 1. Summary of F/OSS benefiting approaches

Approach	Level of involvement in the community	Potential benefits	Potential challenges
Software-centered approach	None	• Reliable and secure F/OSS • Customizable software depending on the degree of competence of the user	• No influence on features developed by the community • Subject to license restrictions
Community-centered approach	Low to medium	• Reliable and secure F/OSS • Some influence on features • Learning/training through participating	• More time and effort required • Subject to license restrictions • Organizational changes required
License-centered approach	High	• Reliable and secure F/OSS • Influence on features • Learning/training through participating • Control over licenses • Improved customer relationship management	• Lose of competitive advantage • Motivate participation • Organizational changes required

Cost saving has been the main driver for F/OSS adoption, but even these savings might not be realized when the total cost of ownership is factored in. However, managers can maximize the benefits they could gain from F/OSS should they understand the whole system and learn to leverage all of its components. We have identified three different approaches companies could use to benefit from F/OSS. In addition to software cost saving, these different approaches (summarized in Table 1) could reduce development costs, increase innovation, improve technical training, and improve competitive position. Each approach provides a different set of benefits and challenges.

REFERENCES

AlMarzouq, M., Zheng, L., Rong, G., & Grover, V. (2005). A tutorial on open source: Concepts, benefits and challenges. *Communications of the AIS, 16*(37), 756-784.

Bonaccorsi, A., & Rossi, C. (2003). Why open source software can succeed. *Research Policy, 32*(7), 1243-1258.

Crowston, K., Annabi, H., & Heckman, R. (2004, Dec. 12). *A structurational model of the dynamics of free/libre.* Paper presented at the International Federation for Informational Processing WG 8.2 Organizations and Society in Information Systems Workshop, Washington, DC.

Crowston, K., & Howison, J. (2005). The social structure of free and open source software development. *First Monday, 10*(2). Retrieved February 20, 2006, from http://www.firstmonday.org/issues/issue10_2/crowston/index.html

The Free Software Definition. (n.d.). Retrieved February 12, 2006 from, http://www.fsf.org/licensing/essays/selling.html

Glass, R. L. (2004). A look at the economics of open source. *Communications of the ACM, 47*(2), 25-27.

History of the OSI. (n.d.). Retrieved February 12, 2006, from http://www.opensource.org/docs/history.html

Jorgensen, N. (2001). Putting it all in the trunk, incremental software development in the freeBSD open source project. *Information Systems Journal, 11*, 321-336.

Lee, S. H. (1999). *Open source software licensing*. Retrieved May 15, 2005, from http://cyber.law.harvard.edu/openlaw/gpl.pdf

Lussier, S. (2004). New tricks: How open source changed the way my team works. *IEEE Software, 21*(1), 68-72.

O'Mahony, S. (2003). Guarding the commons: How community managed software projects protect their work. *Research Policy, 32*(7), 1179-1198.

Raymond, E. S. (1999a). A brief history of hackerdom. In C. DiBona, S. Ockman, & M. Stone (Eds.), *Open sources* (1st ed.). Cambridge, MA: O'Reilly.

Raymond, E. S. (1999b). *The cathedral & the bazaar musings on linux and open source by an accidental revolutionary* (1st ed.). Cambridge, MA: O'Reilly.

Scacchi, W. (2002). Understanding the requirements for developing open source software systems. *IEEE Software, 149*(1), 24-39.

Scacchi, W. (2004). Free and open source development practices in the game community. *IEEE Software, 21*(1), 59- 66.

Selling Free Software. (n.d.). Retrieved Feb 12, 2006, from http://www.fsf.org/licensing/essays/selling.html

Stevenson, R. (2005, April 4). Study shows Microsoft, Linux costs neck-and-neck. *Reuters*.

West, J. (2003). How open is open enough? Melding proprietary and open source platform strategies. *Research Policy, 32*(7), 1259-1285.

West, J., & O'Mahony, S. (2005, January 3-6). *Contrasting community building in sponsored and community founded open source projects.* Paper presented at the Annual Hawaii International Conference on System Sciences, Waikoloa.

Wheatley, M. (2004, March 1). The myths of open source. *CIO Magazine.*

Williams, J., Clegg, P., & Dulaney, E. (2005). The advantages of adopting open source software. In *Expanding choice: Moving to Linux and open source with Novell open enterprise server*. Indianapolis, IN: Novell Press.

KEY TERMS

Brook's Law: Adding manpower to a late software project makes it later.

Copyleft: The copyleft principle prevents the privatization of the whole or parts of the software that has a license that implements this principle.

Free Software: Software is considered free software if the user has the freedom to run, copy, distribute, study, change and improve the software ("The Free Software Definition," n.d.).

General Public License (GPL): GNU General Public License is a license developed by Richard Stallman to ensure that the GNU system remained free. It employs the copyleft principle and prevents mixing with proprietary source code (GPL Compatibility) (O'Mahony, 2003).

GNU: A recursive acronym that stands for GNU is Not Unix. It is a UNIX compatible operating system.

Open Source Software: Software that complies with the criteria set forth by the OSI which can be found at http://www.opensource.org/docs/definition.php

ENDNOTES

1 For more information on understanding the difference between OSI and FSF please refer to: OSI perspective: http://opensource.org/advocacy/free-notfree.php FSF perspective: http://www.fsf.org/licensing/essays/free-software-for-freedom.html

2 For a comprehensive list, please visit: OSI approved licenses: http://www.opensource.org/licenses/ FSF approved licenses: http://www.fsf.org/licensing/licenses/index_html

3 The "GPL compatibility" name is used by the FSF, please see http://www.fsf.org/licensing/licenses

Functional and Object–Oriented Methodology for Analysis and Design

Peretz Shoval
Ben-Gurion University, Israel

Judith Kabeli
Ben-Gurion University, Israel

INTRODUCTION

The chapter provides an overview of FOOM–Functional and Object-oriented Methodology–for analysis and design of information systems. FOOM integrates the functional and object-oriented approaches. In the analysis phase, two main models are created: a) a conceptual data model, in the form of an initial class diagram; and b) a functional model, in the form of OO-DFDs (object-oriented data-flow diagram). In the design phase, the above models are used to design the following products: a) a complete class diagram, including Data, Menus, Forms, Reports and Transactions classes, including their attributes, relationships and methods; b) the user interface–a menus tree; c) the input and output form and report; and d) detailed descriptions of the class methods, expressed in pseudo-code or message charts.

BACKGROUND

Background on Traditional Approach to Information System Development

Many paradigms for system analysis and design have been proposed over the years. Early approaches have advocated the functional approach. Common methodologies that support this approach are SSA and SSD (DeMarco, 1978; Yourdon & Constantine, 1979). SSA is based on the use of data flow diagrams (DFDs), which define the functions of the system, the data stores within the system, the external entities, and the data flows among the above components. Early SSA and similar methodologies emphasized the functional aspects of system analysis, neglecting somehow the structural aspects, namely the data model. This was remedied by enhancing those methodologies with a conceptual data model, usually the entity-relationship (ER) model (Chen, 1976), that is used to create a diagram of the data model, which is later mapped to a relational database schema.

SSD is based on the use of structure charts, which describe the division of the system to program modules as well as the hierarchy of the different modules and their interfaces. Certain techniques have been proposed to create structure charts from DFDs (Yourdon & Constantine, 1979). The main difficulty of an approach where functional analysis is followed by structured design lies in the transition from DFDs to structure charts. In spite of various guidelines and rules for conversion from one structure to the other, the problem has not been resolved by those methodologies (Coad & Yourdon, 1990).

Shoval (1988, 1991) developed the ADISSA methodology that solved this problem. It uses hierarchical DFDs during the analysis stage (similar to other functional analysis methodologies), but the design centers on ***transactions***. A transaction is a process that supports a user who performs a business function, and is triggered as a result of an event. Transactions will eventually become the application programs. Transactions are identified and derived from DFDs: A transaction consists of elementary functions (namely, functions which are not decomposed into subfunctions) that are chained through data flows, and of data-stores and external-entities that are connected to those functions. A transaction includes at least one external-entity, which serve as its trigger. The process logic of each transaction is defined by means of structured programming techniques, for example, pseudo-code.

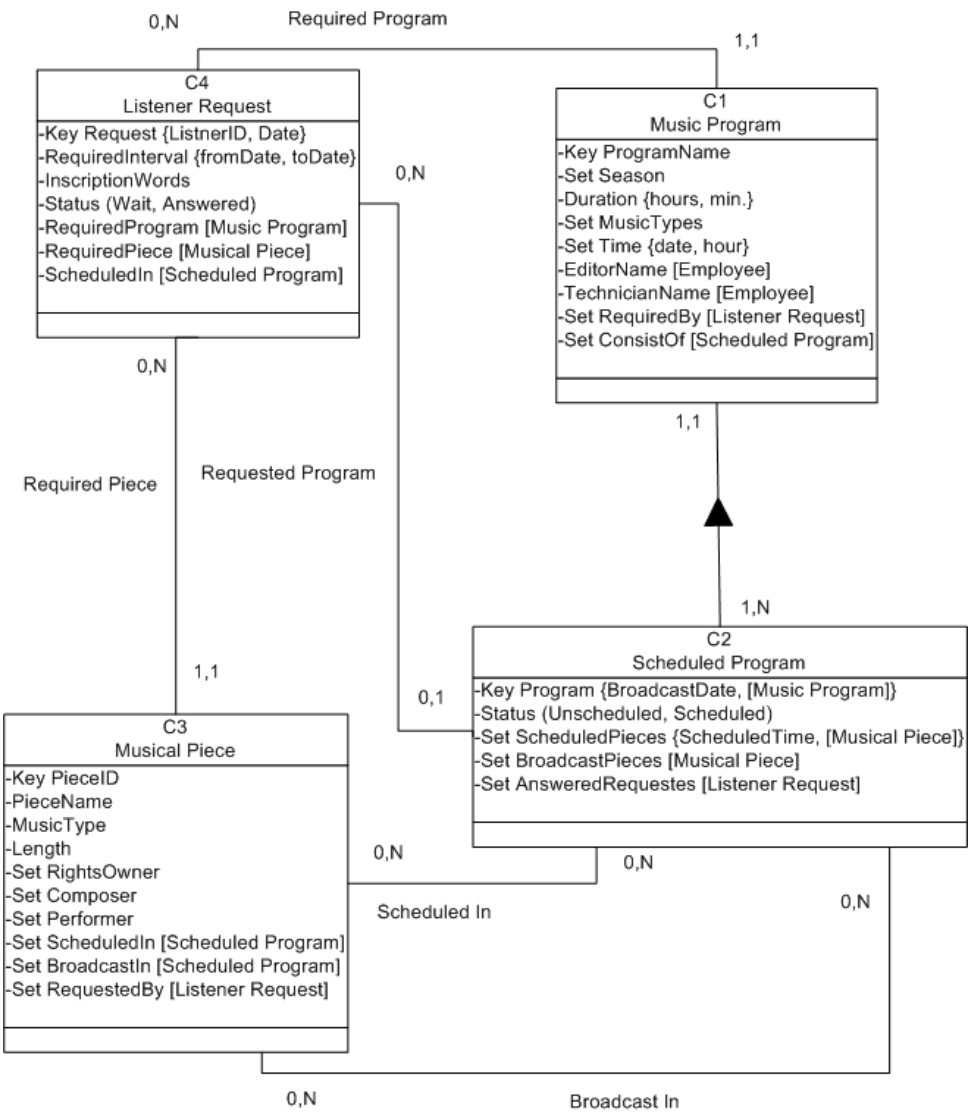
Based on the hierarchical DFDs and the transactions, ADISSA provides structured techniques to design the user-system interface (a menus-tree), the inputs and outputs (forms and reports), the database schema, and detailed descriptions of the transactions, which will eventually become the application programs. The menus-tree is derived from the hierarchy of DFDs in a semi-algorithmic fashion, based on functions that are connected to user-entities. The design of the forms and reports is based on data flows from user-entities to elementary functions and from elementary functions to user-entities. The design of the relational database schema is based on the analysis of dependencies among the data elements within the data-stores. The data flows from elementary functions to data-stores and from data-stores to elementary functions serve as a basis for defining access-steps, namely update and retrieval operations on the relations. Access-steps are expressed as SQL statements that will be embedded in the program code of the respective transactions. The products of the design stages can be easily implemented using various programming environments.

Background on the Object-Oriented Approach to Information System Development

The development of object-oriented (OO) programming languages gave rise to the OO approach and its penetration into system analysis and design. Many OO methodologies have been developed in the early 1990s (e.g., Booch, 1991; Coad & Yourdon, 1990, 1991; Jacobson, 1992; Rumbaugh, Blaha, Premerlani, Eddy, & Lorensen, 1991; Shlaer & Mellor, 1992; Wirfs-Brock, Wilkerson, & Wiener, 1990). In the OO approach, the world is composed of objects with attributes (defining its state) and behavior (methods), which constitute the only way by which the data included in the object can be accessed. When using the OO approach, a model of the system is usually created in the form of a class diagram consisting of data classes with structural relationships between them (e.g., generalization-specialization), and each class having its attributes and methods.

The early OO methodologies tended to neglect the functionality aspect of system analysis, and did not show clearly how to integrate the application functions (transactions) with the class diagram. Another difficulty with those methodologies was that they involved many types of non-standard diagrams and notations. The multiplicity of diagram types in the OO approach has been a major motivation for developing the UML (Booch, Rumbaugh, Jacobson, 1999;

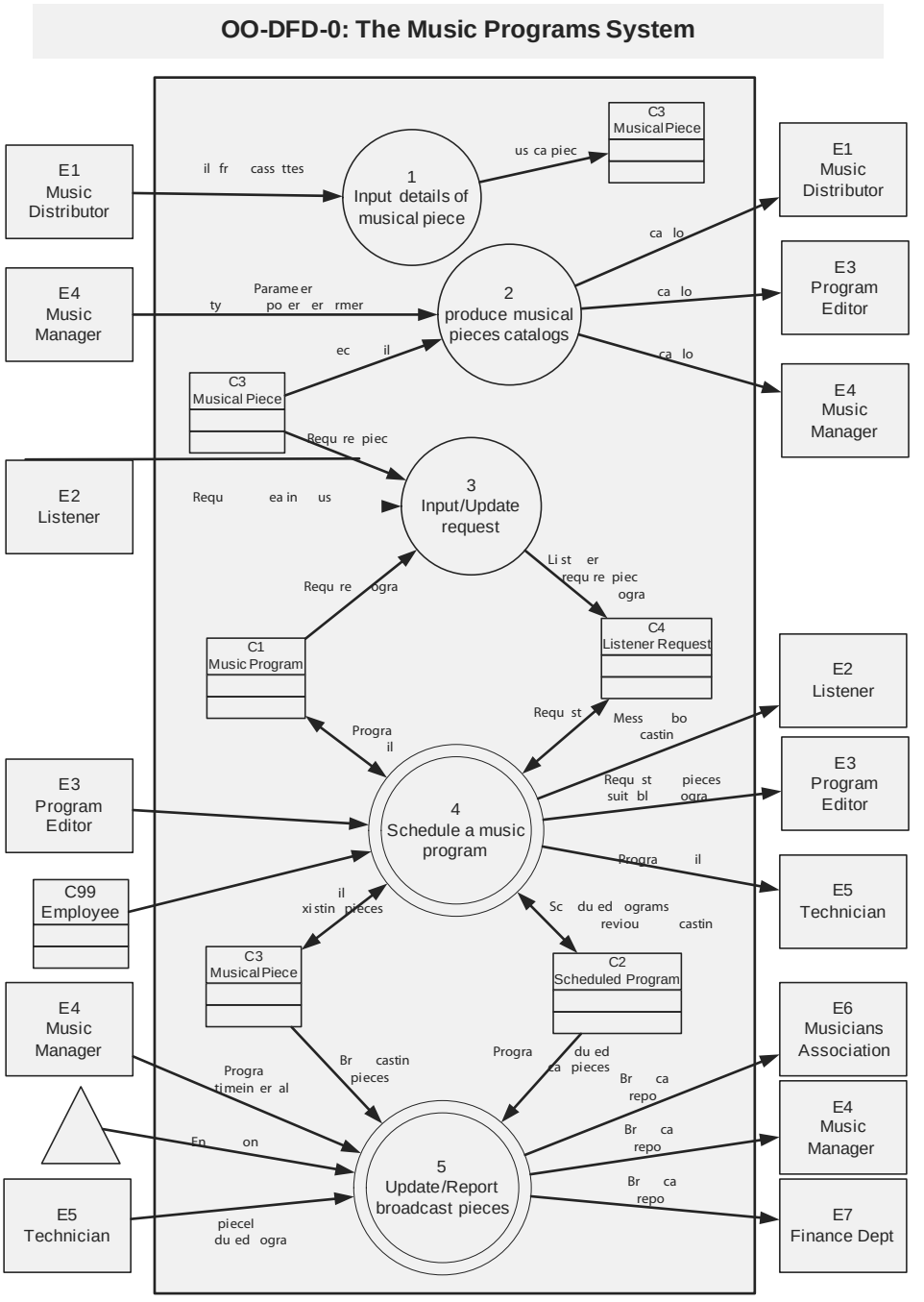
Figure 1. The initial class diagram of Music Program system



Fowler, 2004; Larman, 1998; UML, 2007). UML provides a standard ("unified") modeling language. It consists of several types of diagrams with well-defined semantics and syntax, which enable a system to be presented from different points of views. But UML in itself is not a development methodology that guides the developer how to do and which techniques to use. Moreover, some of the diagram types are redundant, and it is not always clear if and when they should be used in system development (e.g., sequence diagrams vs. collaboration diagrams).

Our approach is to integrate the two paradigms of system development. In our view, functions are as fundamental

Figure 2. The top-level OO-DFD of Music Programs system



as objects, complementing each other, and a development methodology should combine the functional and the object-oriented approaches. FOOM is a system analysis and design methodology that presents this combination.

ESSENTIALS OF FOOM METHODOLOGY

The essential steps of FOOM are described below. For more details see the initial paper, Shoval and Kabeli (2001), and the textbook, Shoval (2007).

The Analysis Phase

The analysis phase consists of two main activities: data analysis and functional analysis. The products of this stage are a conceptual data model, in the form of an **initial class diagram**, and a functional model, in the form of hierarchical **OO-DFDs**.

The **initial class diagram** consists of data (entity) classes, namely classes that are derived from the application/user requirements and contain real-world data. Each class includes attributes of various types (e.g., atomic, multivalued, sets, and reference attributes). Association types between classes include "ordinary" (1:1, 1:n and n:n) relationships, generalization-specialization (inheritance) links between super and subclasses, and aggregation-participation (part-of) links. Relationships are signified by links between respective classes, and by reference attributes to those classes. The initial class diagram does **not** include methods; these will be added at the design phase. An example of an initial class diagram is shown in Figure 1.

The **OO-DFDs** specify the functional requirements of the system. Each OO-DFD consists of general or elementary functions, external entities—mostly user-entities—but also time and real-time entities, data-classes (instead of the traditional data-stores), and data flows among them. An example of an OO-DFD is shown in Figure 2; it is the top-level (root) OO-DFD. For each general function, signified by a double circle, there is a separate OO-DFD which details its subfunctions and related entities and classes (not shown).

Data analysis and functional analysis may be performed in any order. When starting with functional analysis, the analyst elicits the user requirements and based on that creates the OO-DFDs, which include (besides other components) data-class rather than data-stores. Then, the analyst can create an initial class diagram, using the classes already appearing in the OO-DFDs. This means mainly defining proper class associations and attributes. Alternatively, the analyst can first create an initial class diagram (based on user requirements), and then create OO-DFDs, using the already defined classes. We investigated the pros and cons of the two alternative orders in an experimental setting, and found out that an analysis process which starts with data analysis provides better products and is preferred by analysts (Shoval & Kabeli, 2005). At any rate, the methodology provides rules for synchronizing the two analysis products, so that all classes appearing in the class diagram appear also in the OO-DFDs, and vise versa. This is done with the assistance of a data dictionary which is also created as part of the analysis stage and then keeps being updated in the following stages.

THE DESIGN PHASE

Top-Level Design of Transactions

This stage is performed according to ADISSA methodology, where the application transactions are derived from DFDs. The products of this stage include transactions diagrams, as extracted from the OO-DFDs, and top-level descriptions of the transactions. In addition, a new class—Transactions class—is added to the class diagram. This class will contain transaction methods that will be designed from the transactions' descriptions, as will be described later. Figure 3 shows an example of a "simple" transaction, consisting of an elementary-function, a class and a user-entity; this transaction is derived from OO-DFD (Figure 2).

Figure 3. A "simple" transaction diagram

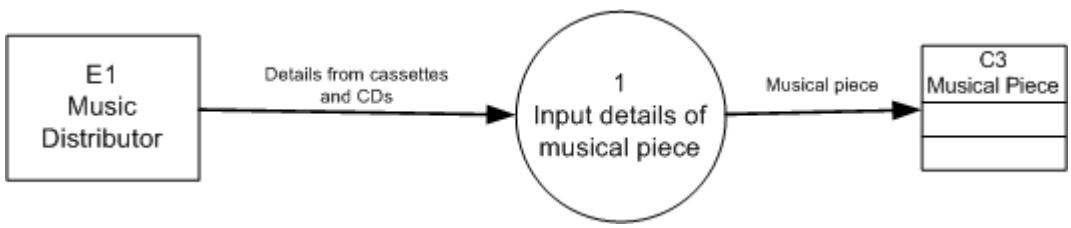


Figure 4. Top-level description of the transaction

```
Begin Transaction_1

  Repeat

    Piece=Forms.Input_Musical_Piece.Display (cassettes and CDs)

      /* Input_Musical_Piece is a form, which enable the user to fills in details of a musical piece. When
                                        the user completes the form it returns the filled piece object */

    If Piece <> NULL then        /* if the user selects to add a piece*/
            Musical_Piece.Construct(Piece)    /*Construct an object of Musical Piece*/
```

A top-level transaction description is provided in Structured-English (pseudo-code), and it refers to all components of the transaction: every data-flow from or to an external entity is translated to an "Input from entity..." or "Output to entity..." command; every data-flow from or to a class is translated to a "Read from class..." or "Write to class..." command; every data flow between two functions is translated to a "Move from... to..." command; and every function in the transaction is translated into an "Execute function…" command. The process logic of the transaction is expressed by standard structured-logic patters (e.g., if... then... else...; do-while...). The analyst and the user, who presents the requirements, determine the process logic of each transaction. This cannot be deducted automatically from the transaction diagram alone, because a given diagram can be interpreted in different ways, and it is up to the user to determine its proper semantics. The top-level description of the above transaction is shown in Figure 4. This description will be used in further stages of design, namely input/output design, and behavior (methods) design, to provide detailed descriptions of the application-specific class methods.

Design of the Interface: The Menus Class

As in ADISSA methodology, the menus-tree is derived in a semi algorithmic way from the hierarchy of OO-DFDs (Shoval, 1990). An example of a menus-tree that is derived from our example is shown in Figure 5. The main menu contains five lines/items; three marked by "T" (for "trigger"), indicating lines that trigger transactions; and two marked by "S" (for "selection") indicating lines that call other menus. (The numbers next to each line indicate the functions included in the transaction being triggered by the respective menu line). The menus-tree is translated into a new class, Menus; the instances (objects) of this class are the individual menus. Note that at run time, a user who interacts with the menu of the system actually works with a certain menu object. He/she may select a menu line that will cause the presentation of another menu object, or invoke a transaction, which is a method of the Transactions class.

Design of the Inputs and Outputs: The Forms and Reports Classes

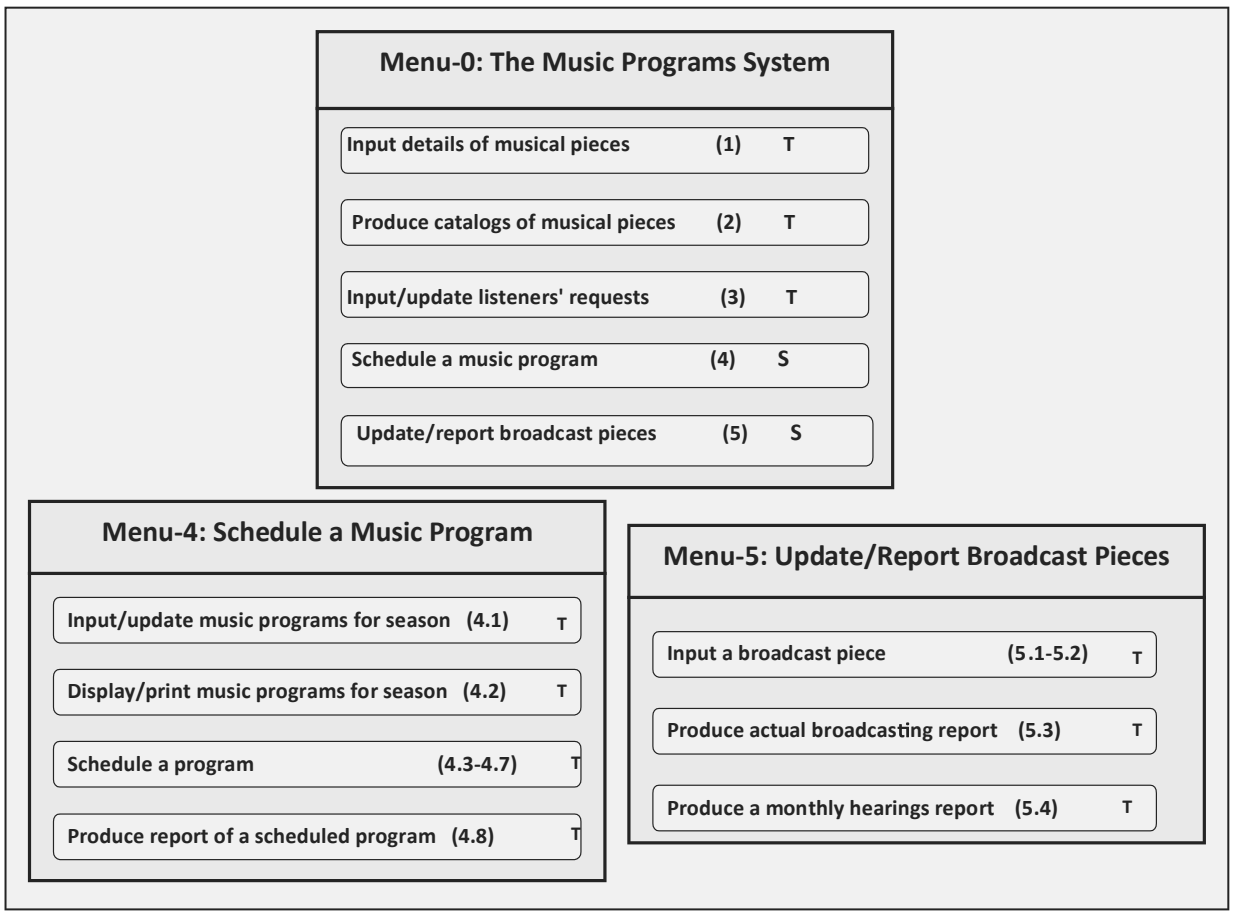
This stage is also performed according to ADISSA and is based on the input and output commands appearing in each of the transaction descriptions. For each "Input from..." command, an input screen/form will be designed, and for each "Output to..." command an output screen/report will be designed. Depending on the process logic of each transaction, some input or output screens may be combined. Eventually, two new classes are added to the class diagram: Forms, for the inputs; and Reports, for the outputs. Obviously, the instances (objects) of each of these classes are the input screens and output screens/reports, respectively.

Design of the System Behavior (Class Methods)

In this stage, the top-level descriptions of the transactions are converted into detailed descriptions of application programs and then into class methods. The transition from a top-level description of a transaction to detailed descriptions of the class methods is done as follows: Every "Input from entity..." and "Output to entity..." command in the top-level description is translated to a message calling an appropriate Display method of the Forms or Reports class, which will display the proper input or output screen or report. Every "Read from class..." or "Write to class..." command is translated to a message calling a "basic" method of the appropriate class. Basic methods include Create, Read, Update and Delete (CRUD methods), which are assumed to exist for every data class.

Every "Execute-Function..." command can be translated to a "basic" method of certain classes, or to an "application-specific" method, which will be attached to a proper class. An application-specific method may be one that performs a procedure/function that is specific to the application/transaction, beyond what is done by any "basic" method; for example, a procedure that performs some computations on values of attributes, and so forth. Each procedure/function

Figure 5. Interface design: The Menus tree



within the transaction description that is identified as a "basic" method or defined as an "application-specific" method of some class, is removed from the transaction's description and replaced by a message to that class-method. But a procedure/function that involves several classes, or performs some general computations, is not defined as an "ordinary" class method and is not attached to a specific class; rather, it remains within the transaction. The remaining parts of the transaction's description are defined as "transaction method." This method is actually the "main" part of the transaction's program, and belongs to the Transactions class. Hence, when users ask the system (via proper menu selections) to perform some activity/transaction, they actually trigger the "main" method of the transaction; that method executes, and while executing it may call (send messages to) other methods ("basic" or "application-specific") of respective classes, according to the process logic of the transaction.

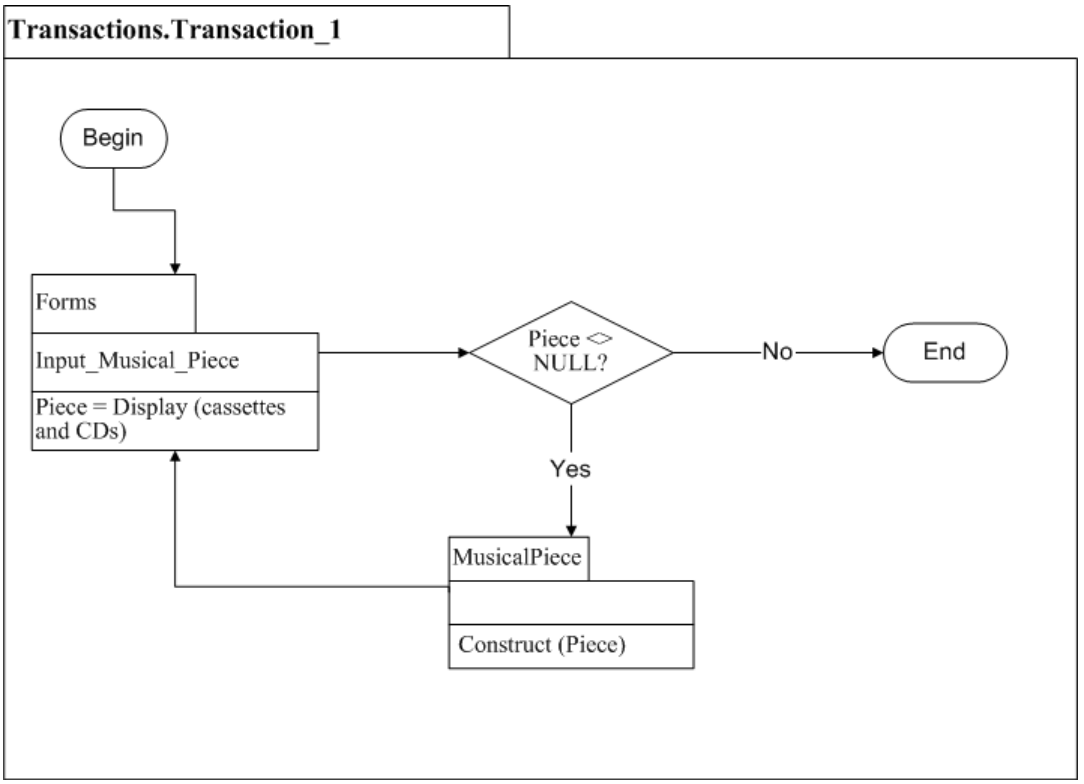
Each "transaction method" or "application-specific" method can be described in two complementing forms: pseudo-code and message chart. Figure 6 is an example of pseudo-code of the transaction. This description must not be too "formal;" it should include comments and explanations that will clarify the meaning of the transaction to the programmer who will implement it in the proper programming language. An equivalent way to describe a method is message chart (see example in Figure 7) which shows the classes, methods and messages included in a method, and the order/process logic of their execution. A message chart is actually an enhanced program flowchart which includes also the classes involved in the method and the messages to proper methods of those classes. As said, message charts are equivalent to pseudo-code and can be used interchangeably.

To summarize, the products of the design phase are: a) a complete class diagram, including Data, Menus, Forms, Reports and Transactions classes, each with various attribute types and method names, and various associations among the classes; b) detailed menu objects of the Menus class; c) detailed form and report objects of the Forms and Reports classes; d) detailed descriptions of the "transaction methods" and the "application-specific" methods, expressed in pseudo-code or in message charts. At the implementation

Figure 6. Pseudo-code of the transaction

```
Begin Transaction_1
   Repeat
      Piece=Forms.Input_Musical_Piece.Display (cassettes and CDs)
         /* Input_Musical_Piece is a form, which enable the user to fills in details of a musical piece. When the
                                    user completes the form it returns the filled piece object */
      If Piece <> NULL then        /* if the user selects to add a piece*/
               Musical_Piece.Construct(Piece)    /*Construct an object of Musical Piece*/
      End if
   Until Piece = NULL                          /* the user selects to stop adding pieces
End
```

Figure 7. Message chart of the transaction



stage, the programmers will use the above design products to create the software with any common OO programming language.

FUTURE DEVELOPMENTS

FOOM methodology is currently not supported by specific CASE tools. Any drawing tool, for example, MS-Visio, can be used to draw the class diagram, the OO-DFDs and the message charts. This has advantages and disadvantages: among the advantages are simplicity (of using any general-purpose tool) and flexibility (the possibility to adopt changes, if needed). The disadvantages are that a simple drawing tool does not provide for checking the syntactic correctness of the diagrams and for transformations from one product to the other.

In the future, we plan to develop a set of specialized CASE tools that will overcome the above limitations. More specifically, the set will include: a tool for drawing a

class diagram and hierarchical OO-DFDs, which will also enable checking the synchronization between them; a tool for deriving transactions diagrams from the OO-DFDs and specifying their process logic; a tool for designing the user interface (i.e., menus); a tool for designing the input and output forms and reports; a tool for mapping the class diagram to a relational database schema; and a tool for creating pseudo-code and message charts of the methods, based on the transactions.

CONCLUSION

The advantages of the FOOM methodology are: (a) in the analysis phase two complementary and synchronized models which represent the users' requirements are created: a conceptual data model, in the form of an initial class diagram, and a functional model, in the form of hierarchical OO-DFDs; (b) the design phase uses these products to create the various design artifacts: The class diagram is augmented with a Menus class which is derived from the menus designed earlier from the OO-DFDs. Inputs and Outputs classes are also derived from the input forms and the outputs of the system (earlier products of the design stage). The transactions are defined as "basic" methods, "application-specific" methods and "transaction methods;" and (c) the end products of the design phase can be easily implemented with any OO programming environment.

REFERENCES

Booch, G. (1991). Object-oriented design with applications. Benjamin/Cummings.

Booch, G., Rumbaugh, J., & Jacobson, I. (1999). The unified modeling language user guide. Addison-Wesley.

Chen, P. (1976). The entity-relationship model: Toward a unified view of data. ACM Transactions on Database Systems, 1(1), 9-36.

Coad, P., & Yourdon, E. (1990). Object-oriented analysis. Englewood Cliffs, NJ: Prentice Hall.

Coad, P., & Yourdon, E. (1991). Object-oriented Design. Englewood Cliffs, NJ: Prentice Hall.

DeMarco, T. (1978). Structured analysis and system specification. New York: Yourdon Press.

Fowler, M. (2004). UML distilled. Addison-Wesley.

Jacobson, I. (1992). Object-oriented software engineering: A use case driven approach. ACM Press.

Larman, C. (1998). Applying UML and patterns—an introduction to object oriented analysis and design. Englewood Cliffs, NJ: Prentice Hall.

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., & Lorensen, W. (1991). Object-oriented modeling and design. Englewood Cliffs, NJ: Prentice Hall.

Shlaer, S., & Mellor, S. (1992). Object life cycles: Modeling the world in states. Englewood Cliffs, NJ: Yourdon Press.

Shoval, P. (1988). ADISSA: Architectural design of information systems based on structured analysis. Information Systems, 13(2), 193-210.

Shoval, P. (1990). Functional design of a menu-tree interface within structured system development. International Journal of Man-Machine Studies, 33, 537-556.

Shoval, P. (1991). An integrated methodology for functional analysis, process design and database design. Information Systems, 16(1), 49-64.

Shoval, P. (2007). Functional and object oriented analysis and design: An integrated methodology. Hershey, PA: IGI Global.

Shoval, P., & Kabeli, J. (2001). FOOM: Functional- and object-oriented analysis & design of information systems—an integrated methodology. Journal of Database Management, 12(1), 15-25.

Shoval, P., & Kabeli, J. (2005). Data modeling or functional modeling—which comes first? An experimental comparison. Communications of the Association for Information Systems, 16, 831-847.

UML. (2007). Unified modeling language. Retrieved May 27, 2008, from http://www.uml.org/

Wirfs-Brock, R., Wilkerson, B., & Wiener, L. (1990). Designing object-oriented software. Englewood Cliffs, NJ: Prentice Hall.

Yourdon, Y., & Constantine, L. L. (1979). Structured design. Englewood Cliffs, NJ: Prentice Hall.

KEY TERMS

ADISSA – Architectural Design of Information Systems based on Structured Analysis: A systems analysis and design methodology which follows the functional- (process) oriented approach (Shoval, 1988). In the analysis phase of development it utilizes hierarchical DFDs. In the design phase the DFDs are used to design the various components of the system. These include: a) top-level descriptions of the transactions, which eventually become detailed descriptions

of the applications programs; b) the user interfaces (menus); c) the input and output screens and reports; and d) the database schema, or normalized relations, and SQL commands for retrieving and updating the database.

DFD - Data Flow Diagram: A diagram used in functional analysis which specifies the functions of the system, the inputs/outputs from/to external (user) entities, and the data being retrieved from or updating data stores. There are well-defined rules for specifying correct DFDs, as well as for creating hierarchies of interrelated DFDs.

ER - Entity-Relationship: A conceptual data model which defines the domain in terms of entities, attributes and relationships (Chen, 1976). ERD is an ER diagram, in which entities are represented as rectangles, attributes as ellipses and relationships between entities as diamonds.

FOOM – Functional and Object-Oriented Methodology: A systems analysis and design methodology which integrates these two approaches, as described in this chapter (Shoval, 2007; Shoval & Kabeli, 2001).

OO-DFD – Object-Oriented DFD: A variant of DFDs introduced in FOOM methodology, which include object (data) classes rather than data stores.

SSA - Structured System Analysis: A traditional, functional-oriented methodology for analyzing information systems, which utilized data flow diagrams.

SSD - Structured System Design: A traditional methodology for designing information systems, which utilized structure charts.

SQL – Structured Query Language: A standard language for querying and updating a relational database.

UML – Unified Modeling Language: An "industrial standard" notation for object-oriented development. It consists of a several types of (mostly) diagrams that enable describing systems from different perspectives, including structural, behavioral (functional) and managerial.

Fundamentals of Multirate Systems

Gordana Jovanovic Dolecek
INSTITUTE INAOE, Puebla, Mexico

INTRODUCTION

Digital signal processing (DSP) is an area of science and engineering that has been rapidly developed over the past years. This rapid development is a result of significant advances in digital computers technology and integrated circuits fabrication (Mitra, 2005; Smith, 2002).

Classical digital signal processing structures belong to the class of single-rate systems since the sampling rates at all points of the system are the same.

The process of converting a signal from a given rate to a different rate is called sampling rate conversion. Systems that employ multiple sampling rates in the processing of digital signals are called multirate digital signal processing systems. Sample rate conversion is one of the main operations in a multirate system (Harris, 2004; Stearns, 2002).

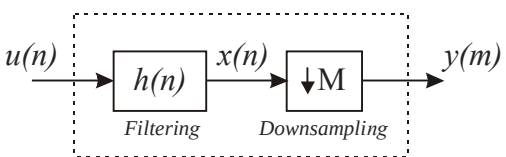
BACKGROUND

Decimation

The reduction of a sampling rate is called decimation, because the original sample set is reduced (decimated). Decimation consists of two stages: filtering and downsampling, as shown in Figure 1. The discrete input signal is $u(n)$ and the signal after filtering is $x(n)$. Both signals have the same input sampling rate f_i.

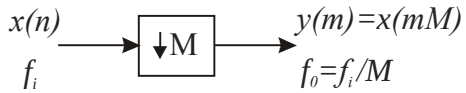
Downsampling reduces the input sampling rate f_i by an integer factor M, which is known as a downsampling factor. Thus, the output discrete signal $y(m)$ has the sampling rate f_i/M. It is customary to use a box with a down-pointing arrow, followed by a downsampling factor as a symbol to represent downsampling, as shown in Figure 2.

Figure 1. Decimation



DECIMATION

Figure 2. Downsampling



The output signal $y(m)$ is called a downsampled signal and is obtained by taking only every M-th sample of the input signal and discarding all others,

$$y(m) = x(mM). \tag{1}$$

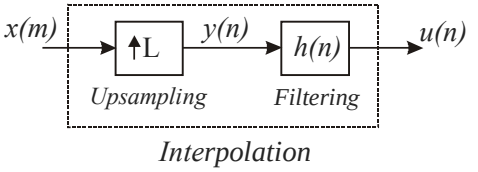
The operation of downsampling is not invertible because it requires setting some of the samples to zero. In other words, we can not recover $x(n)$ from $y(m)$ exactly, but can only compute an approximate value.

In spectral domain downsampling introduces the repeated replicas of the original spectrum at every $2\pi/M$. If the original signal is not bandlimited to π/M, the replicas will overlap. This overlapping effect is called aliasing. In order to avoid aliasing, it is necessary to limit the spectrum of the signal before downsampling to below π/M. This is why a lowpass digital filter (from Figure 1) precedes the downsampler. This filter is called a decimation or antialiasing filter.

Three useful identities summarize the important properties associated with downsampling (Jovanovic Dolecek, 2002). The First identity states that the sum of the scaled, individually downsampled signals is the same as the downsampled sum of these signals. This property follows directly from the principle of the superposition (linearity of operation). The Second identity establishes that a delay of M samples before the downsampler is equivalent to a delay of one sample after the downsampler, where M is the downsampling factor. The Third identity states that the filtering by the expanded filter followed by downsampling, is equivalent to having downsampling first, followed by the filtering with the original filter, where the expanded filter is obtained by replacing each delay of the original filter with M delays. In the time domain this is equivalent to inserting M-1 zeros between the consecutive samples of the impulse response.

The polyphase decimation, which utilizes polyphase components of a decimation filter, is a preferred structure for decimation, because it enables filtering to be performed at a lower sampling rate (Diniz, da Silva & Netto, 2002).

Figure 3. Interpolation



Interpolation

The procedure of increasing the sampling rate is called interpolation, and it consists of two stages: upsampling and filtering (shown in Figure 3).

The upsampler increases the sampling rate by an integer factor *L*, by inserting *L*-1 equally spaced zeros between each pair of samples of the input signal *x(n)* as shown by

$$y(n) = \begin{cases} x(n/L) & for\ \ n = mL \\ 0 & otherwise \end{cases} \qquad (2)$$

where *L* is called interpolation factor.

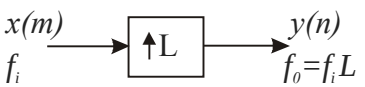
As Figure 4 illustrates, the symbol for this operation is a box with an upward-pointing arrow, followed by the interpolation factor. We can notice that the input sampling rate f_i is increased *L* times.

The process of upsampling does not change the content of the input signal, and it only introduces the scaling of the time axis by a factor *L*. Consequently, the operation of upsampling (unlike downsampling) is invertible, or in other words, it is possible to recover the input signal *x(m)* from samples of *y(n)* exactly.

The process of upsampling introduces the replicas of the main spectra at every $2\pi/L$. This is called imaging, since there are *L*-1 replicas (images) in 2π. In order to remove the unwanted image spectra, a lowpass filter must be placed immediately after upsampling (Figure 3). This filter is called an anti-imaging filter. In the time domain, the effect is that the zero-valued samples introduced by upsampler are filled with "interpolated" values. Because of this property, the filter is also called an interpolation filter.

We have already seen three useful identities of the downsampled signals, and now we will state the identities associated with upsampling. The Fourth identity asserts that the output signal obtained by upsampling followed by scaling of the input signal will give the same result as if the signal is first scaled and then upsampled. The Fifth identity states that a delay of one sample before upsampling is equivalent to the delay of *L* samples after upsampling. The Sixth identity, which is a more general version of the Fifth identity, states that filtering followed by upsampling is equivalent to having upsampling first followed by expanded filtering (Jovanovic Dolecek, 2002; Mitra, 2005; Diniz, da Silva & Netto, 2002).

Figure 4. Upsampling



Cascade of Sampling Converters

An interchange of cascaded sampling converters can often lead to a computationally more efficient realization (Fliedge, 2000; Vaidyanathan, 1993). If upsampling precedes downsampling, where both operations have the same factor, the signal is not changed. However, if downsampling is performed before upsampling, and both operations have the same factor, the resulting signal will be different from the input signal. Rational sampling conversion, that is, changing the sampling rate by a ratio of two integers, *L*/*M* can be efficiently performed as a cascade of upsampling and downsampling, where the interpolation and decimation filters are combined into one filter.

FUTURE TRENDS

Multirate systems have applications in digital radio, speech processing, telecommunications, wavelet transform, digital filtering, A/D converters, spectrum estimation, and so forth.

There are many applications where the signal of a given sampling rate needs to be converted into an equivalent signal with a different sampling rate. For example, in digital radio, three different sampling rates are used: 32 kHz in broadcasting, 44.1 kHz in digital compact disc (CD), and 48 kHz in digital audiotape (DAT), (Fliedge, 2000; Mitra, 2005). Conversion of the sampling rate of audio signals between these three different rates is often necessary. For example, if we wish to play CD music which has a rate of 44.1 kHz in a studio which operates at a 48 kHz rate, then the CD data rate must be increased to 48 kHz using a multirate technique.

In speech processing, multirate techniques are used to reduce the storage space or the transmission rate of speech data (Damiani, Dipanda, Yetongnon, Legrand, Schelkens & Chbeir, 2007; Meana, 2007). In the past years, multirate speech and audio signal processing has been a research topic that has produced several efficient algorithms for echo and noise cancellation, active noise cancellation, speech enhancement, and so forth. (Diniz, da Silva & Netto, 2002; Jovanovic Dolecek, 2002; Meana, 2007).

An example of an application of multirate signal processing in telecommunications is the translation between two multiplexing formats, time division multiplexing (TDM)

and frequency division multiplexing (FDM), (Fliedge 2000; Harris, 2004; Smith, 2001). The filter banks are used to separate a signal into two or more signals (Analysis filter bank) or to compose two or more signals into a single signal (Synthesis filter bank). Multirate systems and filter banks have found many applications in source and channel coding, thereby providing a bridge between communication system design/analysis and signal processing tools (Vaidyanathan, 1993).

The application of multirate techniques to a software radio (SWR) considerably increases the efficiency of the design, (Arslan, 2007; Harada & Prasad, 2002; Reed, 2002). It enhances flexibility and lowers the costs of constructing and operating wireless infrastructure (Bard & Kovarik, 2007). Software radio is one of the key enabling technologies for the wireless revolution and is considered as one of the more important emerging technologies for the future wireless communications (Burachini, 2000; Johnson & Sethares; Kenington, 2005.) Nowadays the term "software radio" generally refers to a radio that derives its flexibility through software while using the same piece of hardware, that is, the same piece of hardware can perform different functions at different times. In that way a completely reconfigurable physical hardware is obtained by programming in software. As a difference, the traditional radio architecture is primarily determined by hardware and so any upgrading means completely abandoning the old design and start with a new design. Software radio terminals must be able to process many various communication standards. These standards are generally based on different master clock rates and thus employ different bit/chip rates. The most obvious solution to cope with the diversity of master clock rates in one terminal is to provide a dedicated master clock for each standard of operation (Hentschel & Fettweis, 2000). However, this approach is too costly and limits the applicability of a realized terminal. Hence, it is much more elegant to run the terminal on a fixed clock rate, and perform digital sample rate conversion controlled by software (Babic & Renfors, 2005; Johansson & Gustafsson, 2005; Saud and Stuber, 2006; Tkacenko, 2007).

One of the most fascinating developments in the field of multirate signal processing has been the establishment of its link to the discrete wavelet transform (Diniz, da Silva & Netto, 2002; Rao, 2002). This link has been responsible for the rapid application of wavelets in fields such as image compression.

Multirate processing has found important application in the efficient implementation of DSP functions (Sheng, Chen & Shan, 2005). The multirate approach increases the computational speed, decreases the overall filter order, reduces word-length effects, and decreases power consumption, making it vital for efficient filtering. Basic multirate techniques used to satisfy the filter requirements are the polyphase decomposition, multistage filtering and frequency masking approach (Mitra, 2005; White, 2000).

The need for inexpensive, high resolution analog/digital (A/D) converters has led to the use of oversampling techniques in the design of such converters, that is, to sample the analog signal at a rate much higher than the Nyquist rate, one uses a fast low-resolution A/D converter, and then decimates the digital output of the converter to the Nyquist rate. Such A/D converter relaxes the sharp cutoff requirements of the analog antialiasing filter, resulting in a more efficient structure (Harris, 2004; Jovanovic Dolecek, 2002; Mitra, 2005).

Another application of multirate signal processing is in the area of spectrum estimation. By using the sampling rate conversion, the computational requirements for narrowband spectrum estimation based on discrete Fourier transform can be significantly reduced (Vaidyanathan, 1993).

Multirate systems also become an active research area in systems and control (Ding & Chen, 2005).

It is expected that in future the applications of multirate systems will play a very important role for resolving problems in communications, control, wavelet analysis, analog/digital converters etc. From the other side, the different applications will give an impulse to develop new algorithms and design methods for multirate systems.

CONCLUSION

There are many applications where the signal of a given sampling rate needs to be converted into an equivalent signal with a different sampling rate. The main reasons could be to increase efficiency or simply to match digital signals that have different rates. During the past several years the multirate processing of digital signals has been attracted by many researchers and the utilization of multirate techniques is becoming an indispensable tool of the electrical engineering profession.

Changing the sampling rate can reduce the computation necessary to complete some DSP operations, and thus reduce the overall system cost. Consequently, the main advantage of multirate systems lies in their high computational efficiency.

REFERENCES

Arslan, H. (2007). *Cognitive radio, software defined radio, and adaptive wireless systems* (Signals and Communication Technology). Springer.

Babic Dj & Renfors, M. (2005). Power efficient structure for conversion between arbitrary sampling rate. *IEEE Signal Processing Letters 12*(1), 1-4.

Bard, J. & Kovarik, V. (2007). *Software defined radio: The software communications architecture*. John Wiley.

Buracchini, E. (2000). The software radio concept. *IEEE Communications Magazine*, September, 138-143.

Ding, F. & Chen, T. (2005). Modeling and identification of multirate systems. *Acta automatica sinica, 31*(1), 105-122.

Damiani, E., Dipanda, A., Yetongnon, K., Legrand, L., Schelkens, P., & Chbeir, R. (Eds.) (2007). *Signal processing for image enhancement and multimedia processing* (Multimedia systems and applications). Springer.

Diniz, P. S. R., da Silva, E. A .B., & Netto, S. L. (2002). *Digital signal processing, system analysis and design*. Cambridge: Cambridge University Press.

Fliedge, N. J. (2000). *Multirate digital signal processing.* New York: John Willey & Sons.

Harada, H. & Prasad, R. (2002). *Simulation and software radio for mobile communications*. Artech House.

Harris, F. (2004). *Multirate signal processing for communication systems*. Prentice Hall PTR.

Hentschel, T. & Fettweis, G. (2000). Sample rate conversion for software radio. *IEEE Communications Magazine*, August, 142-150.

Johansson, H. & Gustafsson, O. (2005). Linear-phase FIR interpolation, decimation, and M th band filters utilizing the farrow structure. *IEEE Transactions on Circuits and Systems, I Regular Papers, 52*(10), 2197-2207.

Johnson, R. & Sethares, W. (2003). *Telecommunications breakdown: Concepts of communication transmitted via software-defined radio*. Prentice Hall.

Jovanovic Dolecek, G. (Ed.) (2002). *Multirate systems: Design & applications.* Hershey, PA: Idea Group Publishing.

Kenington, P. (2005). *R.F and baseband technique*. Artech House Publishers.

Meana, H. P. (2007). *Advances in audio and speech signal processing: Technologies and applications.* Hershey, PA: IGI Global.

Mitra, S. K. (2005). *Digital signal processing: A computer-based approach*. New York: The McGraw-Hill Companies, Inc.

Rao, R. (2002). Wavelet transforms and multirate filtering. In G. Jovanovic-Dolecek (Ed.), *Multirate systems: Design & applications.* Hershey, PA: Idea Group Publishing. 86-104.

Reed, J. H. (2002). *Software radio*: *A modern approach to radio engineering*. Prentice Hall.

Saud, A. & Stuber, W. G. (2006). Efficient sample rate conversion for software radio systems. *IEEE Transactions on Signal Processing, 54*(3), 932-939.

Sheng, J., Chen, T., & Shan, S. L. (2005). Optimal filtering for multirate systems. *IEEE Transactions on Circuits and Systems II: Express Briefs, April*(4), 228-232.

Smith, D. (2001). *Digital signal processing technology: Essentials of the communications revolution*. Amer Radio Relay League.

Smith, S. (2002). *Digital signal processing: A practical guide for engineers and scientists*. Newnes.

Stearns, S. D. (2002). *Digital signal processing with examples in MATLAB*. CRC Press.

Stein, J. (2000). *Digital signal processing: A computer science perspective*. New York: Wiley- Interscience.

Tkacenko, A. (2007). Variable sample rate conversion techniques for the advanced receiver. *IPN progress report* (pp. 42-168).

Vaidyanathan, P. P. (1993). *Multirate systems and filter banks*. New Jersey: Prentice Hall, Inc.

White, S. (2000). *Digital signal processing: A filtering approach*. Delmar Learning.

KEY TERMS

Analysis Filter Bank: Decomposes the input signal into a set of subband signals with each subband signal occupying a portion of the original frequency band.

Continuous-Time Signals: Continuous signals are defined along a continuum of time *t* and thus are represented by continuous independent variables, for example $x_c(t)$. Continuous-time signals are often referred to as analog signals.

Decimation: The process of decreasing the sampling rate. It consists of filtering and downsampling.

Decimation Filter: The filter used in decimation to avoid aliasing caused by downsampling.

Digital Filter: The filter is a discrete-time system, which changes the characteristics of the input discrete signal in a desired manner to obtain the discrete output signal.

Digital Filter Bank: Set of digital bandpass filters with the common input or a common output.

Discrete-Time Signals: Discrete-time signals are defined at discrete time values and thus the independent variable has discrete values *n*, as for example *x(n).*

Interpolation: The process of increasing the sampling rate. It consists of upsampling and filtering.

Interpolation Filter: The filter used in interpolation to remove the unwanted images in the spectra of the upsampled signal.

Multirate System: Discrete–time systems with unequal sampling rates at various parts of the system.

Sampling: The generation of a discrete-time signal $x(n)$ from a continuous signal $x_c(t)$ is called sampling, where $x(n) = x_c(nT)$. T is called the sampling period and its inverse $1/T$ is the sampling frequency or the sampling rate.

Software Radio: The term software radio generally refers to a radio that derives its flexibility through software while using the same piece of hardware, that is, the same piece of hardware can perform different functions at different times. In that way a completely reconfigurable physical hardware is obtained by programming in software.

Subband Coding (SBC) Filter Bank: Consists of an analysis filter bank followed by synthesis filter bank. This type of filter bank is used for partitioning signals into subbands for coding purposes, and vice versa.

Synthesis Filter Bank: Combines subband signals into one signal.

Transmultiplexer Filter Bank (TMUX): Consists of a synthesis filter bank followed by an analysis filter bank. This type of filter bank is used for converting time-multiplexed signals (TDM) into frequency-multiplexed signals (FDM), and vice versa.

Fuzzy and Probabilistic Object-Oriented Databases

Tru H. Cao
Ho Chi Minh City University of Technology, Vietnam

INTRODUCTION

For modeling real-world problems and constructing intelligent systems, integration of different methodologies and techniques has been the quest and focus of significant interdisciplinary research effort. The advantages of such a hybrid system are that the strengths of its partners are combined and complementary to each other's weakness.

In particular, object orientation provides a hierarchical data abstraction scheme and a mechanism for information hiding and inheritance. However, the classical object-oriented data model cannot deal with uncertainty and imprecision pervasive in real world problems. Meanwhile, probability theory and fuzzy logic provide measures and rules for representing and reasoning with uncertainty and imprecision. That has led to intensive research and development of fuzzy and probabilistic object-oriented databases, as collectively reported in De Caluwe (1997), Ma (2005), and Marín & Vila (2007).

BACKGROUND

The key issues in research on extending the classical object-oriented data models to deal with uncertainty and imprecision are:

1. Modeling partial subclass relationship.
2. Definition of partial class membership.
3. Representation of uncertain and/or imprecise attribute values.
4. Representation and execution of class methods.
5. Expression of partial applicability of class properties.
6. Mechanism for inheritance under uncertainty and imprecision.

In the classical object-oriented data model, a class hierarchy defines the subclass/super-class relation on classes. A class *A* is derived as a subclass of a class *B*, which is then called *A*'s super-class, either by narrowing the crisp value ranges of *B*'s attributes or by adding new properties to *B*'s ones. In the probabilistic and fuzzy case, due to the uncertain applicability of class properties or the imprecision of attribute value ranges, the inclusion between classes naturally becomes graded, which could be computed on the basis of the value ranges of their common attributes (George & Buckles & Petry, 1993, Rossazza & Dubois & Prade, 1997).

As discussed in Baldwin, Cao, Martin, and Rossiter (2000), a set of classes with a graded inclusion or inheritance relation actually forms a network rather than a hierarchy, because if a class *A* has some inclusion degree into a class *B* based on a fuzzy matching of their descriptions, then *B* usually also has some inclusion degree into *A*. Moreover, naturally, a concept is usually classified into sub-concepts that are totally subsumed by it, though the sub-concepts can overlap each other, as assumed in Dubitzky, Büchner, Hughes, and Bell (1999) for instance.

Uncertain and imprecise attribute values lead to partial membership of an object into a class, and there are different measures proposed. Yazici and George (1999), for instance, defined for each class a membership function on a set of objects. Bordogna, Pasi, and Lucarella (1999) used linguistic labels to express the strength of the link of an object to a class. Dubitzky et al. (1999) defined membership as similarity degrees between objects and classes. Blanco, Marín, Pons, and Vila (2001) mentioned different measures, including probabilistic one, to be used for membership degrees. Nevertheless, it is to be answered how measures of different meanings, such as possibility and probability, on various levels of a model are integrated coherently.

Most of the works on fuzzy object-oriented data models, which are referred in this paper, were mainly based on fuzzy set and possibility theories, and used fuzzy sets or possibility distributions to represent imprecise attribute values. Bordogna, Pasi, and Lucarella (1999) and Blanco et al. (2001) also modeled uncertainty about an attribute having a particular value. However, much less concern was given for uncertainty over a set of values of an attribute and a foundation to combine probability degrees and fuzzy sets in the same model.

While class attributes were paid much attention and treatment, class methods, as common in object-oriented systems for modeling object behaviors and parameterized properties, were often neglected. In Dubitzky et al. (1999) and Blanco et al. (2001) methods were not considered.

Bordogna, Pasi, and Lucarella (1999) mentioned about methods but did not provide formal representation and explicit manipulation in their model. In Yazici (1999) and Cao and Rossiter (2003) methods were formally defined as Horn clauses and executed as a reasoning process, which were thus for declarative and deductive in contrast to imperative and procedural models.

In the classical object-oriented data model, the properties that represent a class are necessary and sufficient to define the class. However, there is no commonly agreed set of defining properties for many natural, scientific, artificial, and ontological concepts. Arguing for flexible modeling, Van Gyseghem and De Caluwe (1997) introduced the notion of *fuzzy property* as an intermediate between the two extreme notions of required property and optional property, each of which was associated with a possibility degree of applicability of the property to the class. Meanwhile, Dubitzky et al. (1999) addressed the issue by contrasting the prototype concept model with the classical one, assuming each property of a concept to have a probability degree for it occurring in exemplars of that concept.

We note the distinction between the notion of uncertain property values and that of uncertain property applicability. In the former case, an object surely has a particular property but it is not sure which one among a given set of values the property takes. Meanwhile, in the latter, it is even not sure if the object has that property. For example, "John owns a car whose brand is probably BMW" and "It is likely that John owns a car" express different levels of uncertainty. In Bordogna, Pasi, and Lucarella (1999), Blanco et al. (2001), and Cao and Rossiter (2003), the two levels were mixed.

Uncertain class membership and uncertain property applicability naturally result in *uncertain inheritance* of class properties. This was not considered in Bordogna, Pasi, and Lucarella (1999), Dubitzky et al. (1999), and Yazici and George (1999). In Blanco et al. (2001), class membership degrees were used as thresholds to determine what part of the properties in a class would be inherited. In Cao and Rossiter (2003), both membership of an object into a class and applicability of a property to the class were represented by support pairs (Baldwin, Lawry & Martin, 1996) and combined into the support pair for the object to inherit the property.

Recently, Cross (2003) reviewed existing proposals and presented recommendations for the application of fuzzy set theory in a flexible generalized object model. Furthermore, De Tré and De Caluwe (2005) focused on representing data as constraints on object attributes and query answering as constraint satisfaction. For realization of fuzzy object-oriented data models, Berzal et al. (2005) were concerned with implementation of their model on an existing platform. Meanwhile, Fril++, a fuzzy object-oriented logic programming language, was also developed in Rossiter and Cao (2005).

While the fuzzy object-oriented data models referred in this paper were mainly based on fuzzy set and possibility theories, Eiter, Lu, Lukasiewicz, and Subrahmanian (2001) introduced a probabilistic model to handle object bases with uncertainty, called POB (Probabilistic Object Base). For a POB class hierarchy, although a class was assumed to be fully included in its super-classes, the model specified the conditional probability for an object of a class belonging to each of its subclasses. Intuitively, it specified how likely an object of a class belonged to a subclass of that class. Accordingly, the partial class membership was measured by probability degrees. For each attribute of an object, uncertainty about its value was represented by lower bound and upper bound probability distributions on a set of values. The authors also developed a full-fledged algebra to query and operate on object bases.

However, the two major shortcomings of the POB model are: (1) it does not allow imprecise attribute values; and (2) it does not consider class methods. For instance, in the Plant example therein, the values of the attribute *sun light* are chosen to be only enumerated symbols such as *mild*, *medium*, and *heavy* without any interpretation. Meanwhile, in practice, those linguistic values are inherently vague and imprecise over degrees of sun light. Moreover, without an interpretation, they cannot be measured and their probability distributions cannot be calculated.

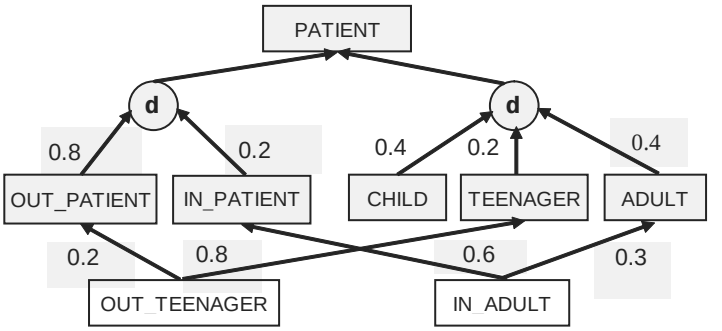
A HYBRID MODEL

In Cao and Nguyen (2007) and Nguyen and Cao (2007), POB is extended with fuzzy attribute values, class methods, and uncertain applicability of class properties. Here, the term *property* is used to subsume both the terms *attribute* and *method*. This hybrid model is called FPOB (Fuzzy-Probabilistic Object Base).

Figure 1 is an FPOB hierarchy of patients, who are classified as being children, teenagers or adults and, alternatively, as being out-patients, or in-patients. Those subclasses of a class that are connected to a **d** node are mutually disjoint, and they form a cluster of that class. The value in [0, 1] associated with the link between a class and one of its immediate subclasses represents the probability for an object of the class belonging to that subclass. For instance, the hierarchy says 80% of patients are non-resident, while the rest 20% are resident. As such, each object could be a member of a class with some probability.

Basically, imprecise and uncertain values of an attribute is expressed by a *fuzzy-probabilistic triple* of the form $\langle V, \alpha, \beta \rangle$, where V is a set of *fuzzy values*, that is,, those defined by fuzzy sets, and α and β are lower and upper bound probability distributions on V. For example, $\langle \{young, middle\_aged\}, .8u, 1.2u \rangle$, represents that the probability for the age of a patient is *young* or *middle-aged* is between 0.4 and 0.6, where *young* and *middle-aged* are linguistic labels of fuzzy sets, and u denotes the uniform distribution.

Figure 1. An example FPOB class hierarchy



For a unified treatment of class attributes and methods, an attribute could be considered as a special method with a fixed output, having no input argument. Alternatively, a method could be considered as a parameterized attribute, whose value depends on its input arguments. In other words, a method is a function from products of fuzzy-probabilistic triples to fuzzy-probabilistic triples. For example, a method could be define for the class PATIENT to compute the total treatment cost of a patient, depending on the daily treatment cost, the patient's treatment duration, and insurance cover.

Uncertain applicability of a class property in FPOB is expressed by a probability interval. For example, since not all patients may have a medical history recorded, that property might be defined with the interval [.8, 1] saying that at least 80% of patients have medical histories. For computation, a property value $\langle V, \alpha, \beta\rangle$ associated with a probability interval $[l, u]$ is assumed to be equivalent to the fuzzy-probabilistic triple $\langle V, \alpha\otimes l, \beta\otimes u\rangle$, where $\otimes$ denotes a probabilistic conjunction operator (Lakshmanan et al., 1997, Ross & Subrahmanian, 2005).

Let $[l, u]$ be the applicability probability interval of a property P to a class c, and $[x, y]$ be the membership probability interval of an object o into c. Then the applicability probability interval of P to o is defined to be $[l, u] \otimes [x, y]$. For multiple uncertain inheritance, in the logic-based fuzzy and probabilistic object-oriented model in Cao (2001), each uncertainly applicable property was considered as a defeasible probabilistic logic rule, and then uncertain inheritance was resolved by probabilistic default reasoning.

In FPOB, complex object structures, or types, can be recursively defined to be of the form $\tau = [P_1(\tau_{11},\ldots, \tau_{1n_1}): \tau_1[l_1, u_1],\ldots, P_k(\tau_{k1},\ldots, \tau_{kn_k}): \tau_k[l_k, u_k]]$. For each property P_i, τ_i and τ_{ij}'s (j from 1 to n_i) are respectively the types of its output and input parameters, and $[l_i, u_i]$ is the applicability probability interval of the property to the class in which it is defined. Simple types are atomic types like real numbers and strings, or fuzzy sets on atomic types.

Since complex object types can be so nested, one has the notion of *path expressions*. In particular, P_i is a path expression for the type τ aforementioned, and $[l_i, u_i]$ is its associated probability interval. $P_i[l_i, u_i]$ is called an *uncertain path expression*. Inductively, if λ_i is a path expression for τ_i with associated probability interval $[l_{\lambda_i}, u_{\lambda_i}]$, then $P_i.\lambda_i$ is a path expression for τ, and $[l_i, u_i] \otimes [l_{\lambda_i}, u_{\lambda_i}]$ is its associated probability interval.

For a particular object, a *fuzzy-probabilistic tuple value* of the type τ is $[P_1(\langle V_{11}, \alpha_{11}, \beta_{11}\rangle, \ldots, \langle V_{1n_1}, \alpha_{1n_1}, \beta_{1n_1}\rangle)$: $\langle V_1, \alpha_1, \beta_1\rangle[l'_1, u'_1], \ldots, P_k(\langle V_{k1}, \alpha_{k1}, \beta_{k1}\rangle, \ldots, \langle V_{kn_k}, \alpha_{kn_k}, \beta_{kn_k}\rangle)$: $\langle V_k, \alpha_k, \beta_k\rangle[l'_k, u'_k]]$, where $\langle V_i, \alpha_i, \beta_i\rangle$ and $\langle V_{ij}, \alpha_{ij}, \beta_{ij}\rangle$ are fuzzy-probabilistic triples of types τ_i and τ_{ij}, (i from 1 to k and j from 1 to n_i). Here, $[l'_i, u'_i]$ specifies the uncertain applicability of P_i to that particular object, which is not necessarily the same as the default value $[l_i, u_i]$ for a generic object of the type τ.

The most important database operation is selection. For FPOB, a *fuzzy-probabilistic selection expression* is inductively defined to be in one of the following forms:

1. $x \in c$, where x is an object variable and c is a class.
   This is to select those objects that belong to the class c.
2. $x.\lambda$, where x is an object variable and λ is a path expression.
   This is to select those objects to which the path expression λ is applicable.
3. $x.\lambda\ \theta\ v$, where x is an object variable, λ is a path expression, θ is a binary relation from $\{=, \neq, \leq, <, \subseteq, \in, \rightarrow\}$, and v is a value.
   This is to select those objects for which the values of the property defined by the path expression λ satisfy the relation θ with the value v.
4. $x.\lambda_1 =_{\otimes} x.\lambda_2$, where x is an object variable, λ_1 and λ_2 are path expressions, and $\otimes$ is a probabilistic conjunction strategy of combining the probabilities for $x.\lambda_1 = v_1$, $x.\lambda_2 = v_2$, and $v_1 = v_2$.

This is to select those objects for which the values of the properties defined by the path expressions λ_1 and λ_2 are equal.

5. $E_1 \otimes E_2$, where E_1 and E_2 are selection expressions over the same object variable and $\otimes$ is a probabilistic conjunction strategy of combining the probabilities for E_1 and E_2 being true.
   This is to select those objects for which both the selection expressions E_1 and E_2 are satisfied.
6. $E_1 \oplus E_2$, where E_1 and E_2 are selection expressions over the same object variable and $\oplus$ is a probabilistic disjunction strategy of combining the probabilities for E_1 and E_2 being true.

This is to select those objects for which either the selection expression E_1 or the selection expression E_2 is satisfied.

The *probabilistic interpretation* of a selection expression for an object gives the probabilistic interval for that object to satisfy the expression.

A *fuzzy-probabilistic selection condition* is then inductively defined as a selection expression to be satisfied with a probability in an interval as follows:

1. If E is a selection expression and $[l, u]$ is a subinterval of [0, 1], then $(E)[l, u]$ is a selection condition.
2. If ϕ and ψ are selection conditions, then $\neg\phi$, $(\phi \wedge \psi)$ and $(\phi \vee \psi)$ are selection conditions.

For example, selection of those patients who is young and likely have medical history records with a probability of at least .8 can be posed by the condition:

$(x.\text{age} \rightarrow \textit{young})[1, 1] \wedge (x.\text{medical\_history})[.8, 1]$

where *young* is a linguistic label of a fuzzy set.

An algebra has been developed to compute with fuzzy-probabilistic triples. Meanwhile, the basis for matching a selection expression to an object is the definition of probabilistic interpretation of binary relations on fuzzy sets. The probabilistic interpretation $prob(A\ \theta\ B)$ of a relation $A\ \theta\ B$, where A and B are respectively fuzzy sets on domains U and V, $\theta \in \{=, \neq, \leq, <, \subseteq, \in\}$, is a value in [0, 1] defined by:

$$\sum_{S\subseteq U,\ T\subseteq V} \Pr(u\ \theta\ v \mid u\in S,\ v\in T).m_A(S).m_B(T)$$

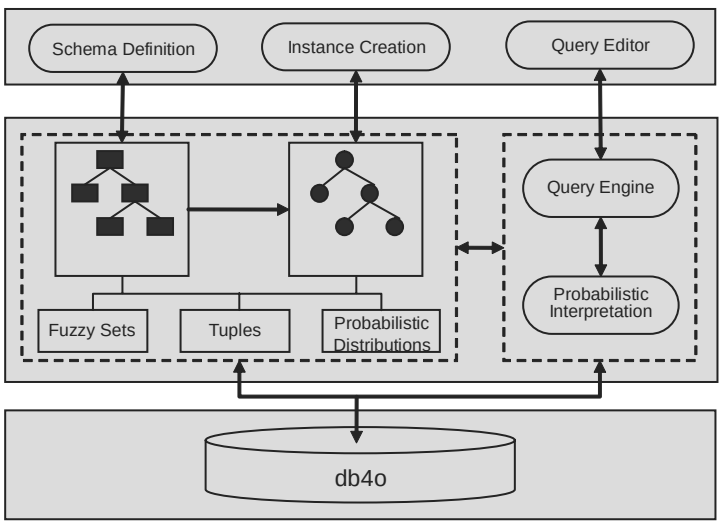
where $m_A(S)$ and $m_B(T)$ are the mass assignments of S and T with respect to A and B (Baldwin, Lawry & Martin, 1996). Intuitively, given fuzzy propositions $x\in A$ and $y\in B$, $prob(A\ \theta\ B)$ is the probability for $x\ \theta\ y$ being true. Especially, the probabilistic interpretation $prob(A \rightarrow B)$ of the relation $A \rightarrow B$, where A and B are fuzzy sets on the same domain U, is a value in [0, 1] defined by:

$$\sum_{S,T\subseteq U} \Pr(u\in T \mid u\in S).m_A(S).m_B(T).$$

The intuitive meaning of $prob(A \rightarrow B)$ is that it is the probability for $x\in B$ being true given $x\in A$ being true. These definitions are the foundation to combine probabilities and fuzzy sets into the coherent framework of FPOB.

Other basic algebraic operations on databases are Cartesian product, join, intersection, union, difference have also been defined for FPOB. For applications of FPOB, Nam et al. (2007) implement an FPOB management system called

Figure 2. FPDB4O architecture



FPDB4O, whose architecture comprises three layers as illustrated in Figure 2:

1. GUI Layer: this is for user interface, allowing users to draw an FPOB class hierarchy and define attributes for classes (Schema Definition), to create a new FPOB instance by creating and modifying data objects (Instance Creation), and to query on an FPOB (Query Editor).
2. Core Layer: this is the main component of the architecture. It includes two main blocks, one for representing the FPOB model, and one for executing queries. In more detail, the former determines the structure for representing schemas, data objects, probabilistic distributions of attribute values, and fuzzy sets. The latter parses queries, performs probabilistic interpretations, and gives answers.
3. Data Layer: this provides functions for storing and retrieving FPOB data objects, employing related functions of DB4O (Grehan, 2005).

FUTURE TRENDS

Research on fuzzy and probabilistic object-oriented databases has been extensive and mature. Due to complexities of theoretical foundation and practical implementation, no model would be so universal that could include all measures and tackle all aspects of uncertainty and imprecision. Different models are thus to complement each other to deal with certain facets of the complex real world. In the coming years, besides continuing work on development of new hybrid models, application tools, languages, and systems are expected to emerge.

CONCLUSION

Various fuzzy and probabilistic object-oriented database models have been developed during the last 15 years, introducing basic concepts and proposing solutions for key issues of fuzzy and probabilistic object-oriented modeling. They are truly extension of the classical object-oriented models with imprecise and/or uncertain class properties. FPOB, a state-of-the-art model, combines fuzzy set and probability theories into a coherent framework for modeling classes and objects and computing with their attributes and methods.

Currently, only natural join operation is defined in FPOB. To support general join operation, probabilistic interpretation of relations on fuzzy-probabilistic triples has to be defined. Also, developed fuzzy and probabilistic object-oriented database languages and systems are to be applied to real world problems.

REFERENCES

Baldwin, J.F., Cao, T.H., Martin, T.P., & Rossiter, J.M. (2000). Toward soft computing object-oriented logic programming. In *Proceedings of the 9th IEEE International Conference on Fuzzy Systems* (pp. 768-773).

Baldwin, J. F., Lawry, J. M., & Martin, T. P. (1996). A mass assignment theory of the probability of fuzzy events. *International Journal of Fuzzy Sets and Systems, 10*, 353-367.

Berzal, F., Marín, N., Pons, O., & Vila, M. A. (2005). A framework to build fuzzy object-oriented capabilities over an existing database system. Z. Ma, (Ed.), *Advances in fuzzy object-oriented database: Modeling and applications* (pp. 177-205). Hershey, PA: Idea Group Publishing.

Blanco, I., Marín, N., Pons, O., & Vila, M. A. (2001). Softening the object-oriented database model: Imprecision, uncertainty and fuzzy types. In *Proceedings of the 1st International Joint Conference of the International Fuzzy Systems Association and the North American Fuzzy Information Processing Society* (pp. 2323-2328).

Bordogna, G., Pasi, G., & Lucarella, D. (1999). A fuzzy object-oriented data model managing vague and uncertain information. *International Journal of Intelligent Systems, 14*, 623-651.

Cao, T. H. (2001). Uncertain inheritance and recognition as probabilistic default reasoning. *International Journal of Intelligent Systems, 16*, 781-803.

Cao, T. H. & Nguyen, H. (in press). Modelling and computing with imprecise and uncertain properties in object bases. *International Journal of Intelligent Information and Database Systems*.

Cao, T. H. & Rossiter, J. M. (2003). A deductive probabilistic and fuzzy object-oriented database language. *International Journal of Fuzzy Sets and Systems, 140*, 129-150.

Cross, V. V. (2003). Defining fuzzy relationships in object models: abstraction and interpretation. *International Journal of Fuzzy Sets and Systems, 140*, 5-27.

De Caluwe, R. (Ed.) (1997). *Fuzzy and uncertain object-oriented databases: Concepts and models*. World Scientific.

De Tré, G. & De Caluwe, R. (2005). A constraint based fuzzy object-oriented database model. In Z. Ma (Ed.), *Advances in fuzzy object-oriented databases: Modelling and applications* (pp. 1-45). Hershey, PA: Idea Group Publisher.

Dubitzky, W., Büchner, A. G., Hughes, J. G., & Bell, D. A. (1999). Towards concept-oriented databases. *Data & Knowledge Engineering, 30*, 23-55.

Eiter, T., Lu, J. J. & Lukasiewicz, T. & Subrahmanian, V. S. (2001). Probabilistic object bases. *ACM Transactions on Database Systems, 26*, 264-312.

George, R., Buckles, B. P., & Petry, F.E. (1993). Modelling class hierarchies in the fuzzy object-oriented data model. *International Journal of Fuzzy Sets and Systems, 60*, 259-272.

Grehan, R. (2005). Complex object structures, persistence, and DB4O. Series of db4o whitepaper, db4objects Inc.

Lakshmanan, L. V. S., Leone, N., Ross, R., & Subrahmanian, V. S. (1997). ProbView: A flexible probabilistic database system. *ACM Transactions on Database Systems, 22*, 419-469.

Ma, Z. (Ed.) (2005). *Advances in fuzzy object-oriented database: Modeling and applications*. Hershey, PA: Idea Group Publishing.

Marín, N. & Vila, M. A. (2007). Special issue on intelligent fuzzy information systems: Beyond the relational data model. *International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems, 15*.

Nam, M., Ngoc, N. T. B., Nguyen, H., & Cao, T. H. (2007). FPDB4O: A fuzzy and probabilistic object base management system. In *Proceedings of the 16th IEEE International Conference on Fuzzy Systems* (pp. 676-681).

Nguyen, H. & Cao, T. H. (2007). Extending probabilistic object bases with uncertain applicability and imprecise values of class properties. In *Proceedings of the 16th IEEE International Conference on Fuzzy Systems* (pp. 487-492).

Ross, R. & Subrahmanian, V. S. (2005). Aggregate operators in probabilistic databases. *Journal of the ACM, 52*, 54-101.

Rossazza, J-P., Dubois, D. & Prade, H. (1997). A hierarchical model of fuzzy classes. In R. De Caluwe (Ed.), Fuzzy and uncertain object-oriented databases: Concepts and models (pp. 21-61). World Scientific.

Rossiter, J. & Cao, T. H. (2005). Fril++ and its applications. Z. Ma (Ed.), *Advances in fuzzy object-oriented databases:* Modelling and Applications (pp. 113-152).Hershey, PA: Idea Group Publisher.

Van Gyseghem, N. & De Caluwe, R. (1997). The UFO database model: Dealing with imperfect information. R. De Caluwe (Ed.), *Fuzzy and uncertain object-oriented databases: Concepts and models* (pp. 13-185). World Scientific.

Yazici, A. & George, R. (1999). Fuzzy database modelling. *Studies in Fuzziness and Soft Computing, 26*, Physica-Verlag.

F

KEY TERMS

Fuzzy Class Hierarchy: An extended conventional class hierarchy where each link between a class and one of its subclasses is associated with an inclusion degree in [0, 1].

Fuzzy Class Membership: A class is considered as a fuzzy set on a set of objects, for which each object is a member of a class to a certain degree.

Fuzzy Property: A property that is applicable to a class with a certain possibility degree.

Fuzzy-Probabilistic Selection Condition: A condition of a selection operation on a fuzzy-probabilistic object-oriented database, which is a selection expression associated with a probability interval.

Fuzzy-Probabilistic Selection Expression: An expression of a selection operation on a fuzzy-probabilistic object-oriented database, which includes constraints on uncertain path expressions and fuzzy-probabilistic triple values.

Fuzzy-Probabilistic Triple Value: An imprecise and uncertain value expressed by lower and upper bound probability distributions on a set of fuzzy set values.

Probabilistic Class Hierarchy: An extended conventional class hierarchy where each link between a class and one of its subclasses is associated with a conditional probability for an object of the class belonging to that subclass.

Probabilistic Class Membership: There is a probability for an object belonging to a class.

Probabilistic Interpretation of Selection Expression: A probability interval for an object satisfying a selection expression.

Probabilistic Property: A property that is applicable to a class with a certain probability.

Uncertain Inheritance: An object may inherit a property of a class to a certain degree only.

Uncertain Path Expression: A sequence of nested properties associated with a probability interval expressing the uncertainty of its applicability to a class or object.

Gender and Computer Anxiety

Sue E. Kase
The Pennsylvania State University, USA

Frank E. Ritter
The Pennsylvania State University, USA

INTRODUCTION

Because of their ability to enhance productivity, computers have become ubiquitous in the workplace. By the early 1990s the use of computers in the workplace reached a per capita penetration that the telephone took 75 years to achieve (Webster & Martocchio, 1992). During the past several decades, there has been both speculation and hard research related to the psychological effects of computer technology. More recently the role of attitudes towards computers in influencing the acceptance and use of computer-based management information systems (MIS) has been highlighted by a growing number of MIS researchers. Generally, these studies focus on the negative attitudes towards computers and concerns about the impact of MIS on individual performance in the workplace.

Computer anxiety has been reported to be associated with negative attitudes towards computers. As computers play a pervasive role in MIS and decision support systems, these findings emphasize the need for additional empirical research on the determinants of computer anxiety and attitudes towards computers. Furthermore, with the increasing participation of women in information technology professions, important questions are whether men and women differ with regard to computer anxiety and attitudes towards computers, and what factors explain such differences where they exist, and how to ameliorate anxiety where it occurs.

The Concept and Correlates of Computer Anxiety

Much has been speculated about computer anxiety, both what it is and what to do about it. Computer anxiety is context specific and covers a wide variety of situations in which people interact with computers. Context-specific anxiety tests ask the question: "How do you feel when a specific type of situation occurs?" Commonly, the relationship between a measure of computer anxiety and other variables is examined. For example, the relationship of computer anxiety to computer-related experience has historically been a hotly contested question in MIS research, human-computer interaction (HCI), and educational psychology. Demographic variables posited or found to be related to computer anxiety include gender, age, organizational level, and academic major (Dambrot, Watkins-Malek, Silling, Marshall & Garver, 1985; Gutek & Bikson, 1985; Zmud, 1979). Personality variables examined as potential determinants of computer anxiety include trait anxiety, math anxiety, cognitive style, and locus of control (Howard & Smith, 1986; Igbaria & Parasuraman, 1989; Morrow, Prell & McElroy, 1986). Additionally, several studies have examined the relationship between computer anxiety and academic achievement. For example, Hayek and Stephens (1989) and Marcoulides (1988) reported significantly lower computer anxiety being associated with higher academic achievement.

BACKGROUND

Initially, computer anxiety became of interest during the technological revolution. In 1963 a social psychologist at IBM completed a nationwide study to examine popular beliefs and attitudes about one of the prime symbols of our rapidly changing technology—the electronic computer. Lee's (1970) findings concluded that the American public viewed computers on two independent dimensions. The first dimension, the "Beneficial Tool of Mankind Perspective," described a positively toned set of beliefs that computers are beneficial in science, industry, and business. The second dimension, the "Awesome Thinking Machine Perspective," connoted fear of an incomprehensibly complex machine with capabilities far exceeding those of a human. This perspective, which reflects ignorance about the capabilities and limitations of computers, is one of the generic origins of computer anxiety.

Later, during the 1980s, much of the writing about computer anxiety and attitudes towards computers was concentrated in trade and business publications (e.g., Howard, 1986; Igbaria & Parasuraman, 1989). During this time period uncertainty was often considered the primary predictor of computer anxiety. This uncertainty referred to an individual's ability to learn to use the computer or to the potential the machine had to rearrange traditional office functions and power structures. Sabotage and hostility were sometimes responses to these uncertainties, especially when they were

accompanied by fear of replacement by the machine. This particular concern was often voiced by middle managers who viewed their jobs as information conduits or as a mosaic of clerical tasks, all of which could be performed more efficiently by a computer. Managers with longer tenure with a company and those who felt they were currently utilizing their time quite effectively were likely to resist computer adoption and use. Additionally, computer usage required typing skills. Those persons who did not know how to type or considered typing a low-status skill were reluctant to adapt to the new technology.

Collectively two groups displayed the most susceptibility to computer anxiety: individuals without computer experience overestimated the difficulties involved in learning and interacting with computers; and individuals whose jobs appeared threatened resisted adaptation to technological improvements (Gilroy & Desai, 1986). It has been well documented that among individuals demonstrating computer anxiety are significant numbers of women, as examined in the next section.

The Role of Gender in Computer Anxiety

According to feminist technology studies, computers are widely perceived as belonging to the "male domain" of mathematics, science, electronics, and machinery (Beyer, 1999; Cockburn & Ormrod, 1993; Faulkner, 2001). This, coupled with reports of greater prevalence of math anxiety among women than men (e.g., Brown & Josephs, 1999; Chipman, Krantz & Silver, 1992), suggests that women are likely to have a more negative view of computer use than men. It is not surprising that men have been found to display lower computer anxiety, higher computer aptitude, and more positive attitudes towards computers in general than women (Chua, Chen & Wong, 1999; Coffin & Machintyre, 2000; Colley, Gale & Harris, 1994; Whitely, 1997).

The limited empirical research on gender differences in computer anxiety, attitudes towards computers, and computer experiences among working adults reveals conflicting results, however. By the early 1990s, only 25 studies presented sufficient statistical information that could be converted to correlations, and an additional 13 qualitative research reports supported only slight differences between men and women in computer anxiety (Rosen & Maguire, 1990). In contrast, other studies reported no gender differences associated with computer use in the workplace.

More specific studies found stronger correlations and gender differences. One of the most frequently cited studies on computer anxiety is by Rosen, Sears, and Weil (1987). They examined the relationship between computer anxiety and gender role as measured by the Bem Sex Role Inventory (Bem, 1974). This instrument identified individuals as belonging to one of four identity groups: masculine, feminine, androgynous, or undifferentiated. They found that feminine-identity individuals had more computer anxiety and more negative computer attitudes than did masculine-identity individuals, regardless of gender. Another influential instrument employed extensively in early studies on microcomputer anxiety in management is the Computer Anxiety Rating Scale (CARS) developed by Raub (1981). Raub investigated math anxiety, gender, age, trait anxiety, and knowledge of computers as possible correlates of computer anxiety. She suspected a gender effect based on the negative socialization of women toward mathematics, science, and technology and on the resulting production of anxieties. Raub found the relationship between computer anxiety and gender so strong that she ran separate regressions for males and females.

CARS has been used in more recent research as well. Anderson (1996) utilized CARS to determine whether or not perceived knowledge of software, computer experience, overall knowledge of computers, programming experience, and gender were predictors of computer anxiety. Table 1 displays the CARS portion of the Anderson questionnaire. Collaborating Raub's results, Anderson's study showed higher computer anxiety is accompanied by less experience and less perceived knowledge of computers, and that at higher levels of computer anxiety, women are over-represented.

Indirectly, the use of CARS led to the introduction of statistical modeling as a more formal investigation of the psychological mechanisms that trigger computer anxiety and the remedies for it. For example, the work of Howard (1986) is distinctive during the 1980s in its similarity to more recent research on computer anxiety. Howard developed a sequence of models addressing the predictors of computer anxiety and the use of computers in management. In schematic form, Figure 1 shows the possible relationships between psychological variables and the attitudes of managers toward the usefulness of microcomputers as management tools. Howard's study confirmed that computer anxiety is a significant inverse correlate of managers' positive attitudes toward microcomputers.

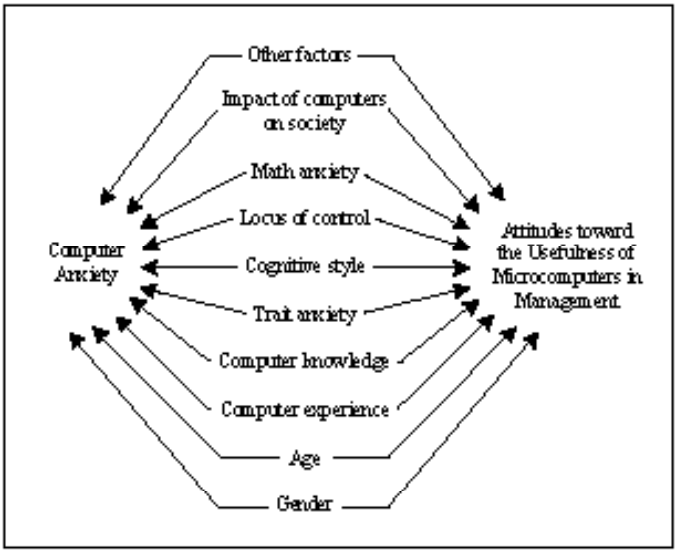
Similar to Raub, Howard speculated that gender may correlate with math anxiety and possibly with computer anxiety. Gender as a math anxiety correlate reflects psychological differences between men and women with regard to mathematics that may result from early socialization of females away from scientific and technical endeavors. If math anxiety and computer anxiety are similar phenomena, then they are likely to have common psychological roots. Thus, for both Raub and Howard, the prime psychological root of computer anxiety appeared to be that certain people simply did not see themselves as technological types. Math anxious types see mathematics and computers and all the paraphernalia of technology as for someone else and when required to use it experience stress.

A current perspective of Howard's usefulness of microcomputers research is several recent studies examining the

Table 1. The Computer Anxiety Rating Scale (CARS) is based on research by Raub (1981)

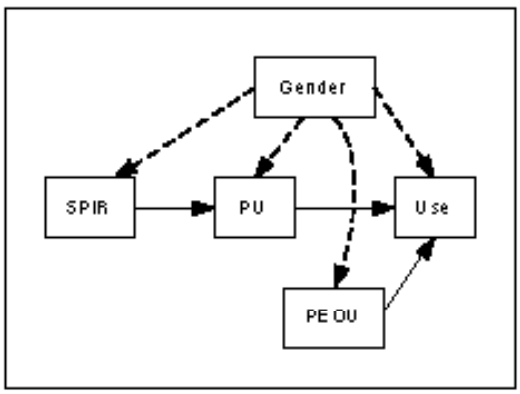
The responses are scaled as follows: 1 = Strongly Agree 2 = Agree 3 = Unsure 4 = Disagree 5 = Strongly Disagree
Items 2 through 10 are reverse scored so that high scores indicate high levels of computer anxiety. 1. I am confident that I could learn computer skills. 2. I am unsure of my ability to learn a computer programming language. 3. I will be able to keep up with the important technological advances of computers. 4. I feel apprehensive about using the computer. 5. If given the opportunity to use a computer, I'm afraid that I might damage it in some way. 6. I have avoided computers because they are unfamiliar to me. 7. I hesitate to use the computer for fear of making mistakes that I cannot correct. 8. I am unsure of my ability to interpret a computer printout. 9. I have difficulty understanding most technological matters. 10. Computer terminology sounds like confusing jargon to me.

Figure 1. Computer anxiety and other possible correlates of managers' attitudes towards microcomputers adapted from Howard (1986)



diffusion of information technology (IT) in the workplace. Many of these studies have demonstrated a strong link between self-efficacy and individual reactions to computing technology, both in terms of adoption and use of computers, and in terms of learning to use computer systems and applications. Beliefs about our capabilities to use technology successfully are related to our decisions about whether and how much to use technology, and the degree to which we are able to learn from training (Compeau, Higgins & Huff, 1999). While inconsistencies in group differences exist, overall findings of IT diffusion research suggest that women and men exhibit different perceptual tendencies and usage patterns in computer-related circumstances (Chou, 2001; Gefen & Straub, 1997; Hackbarth, Grover & Yi, 2003; Venkatesh, Morris & Ackerman, 2000).

Figure 2. Gender effects on TAM variables (Gefen & Straub, 1997)



Gefen and Straub (1997), and Venkatesh, Morris, and Ackerman (2000) are two prevalent examples of IT diffusion research predicated on influential theories evaluating information technology perception and use: the Technology Acceptance Model (TAM) and the Theory of Planned Behavior (TPB). Gefen and Straub (1997) modeled IT diffusion by extending TAM and the Social Presence/Information Richness (SPIR) factor addendum to include gender in the context of e-mail system usage. Figure 2 shows the effects of gender on TAM and cultural extensions, specifically the perceived attributes of SPIR, perceived ease of use (PEOU), and perceived usefulness (PU). Gefen and Straub sampled 392 female and male knowledge workers using e-mail systems in the airline industry in North America, Asia, and Europe. Study findings indicated that women's perceptions of e-mail are different from male co-workers. Covariates gender and culture accounted for 37% of the variance in SPIR, 53% of the variance in PEOU, and 59% of the variance in SPIR, culture, and PU combined. In this study gender definitely had an impact on the IT diffusion process.

Using TPB as a framework, Venkatesh, Morris, and Ackerman (2000) investigated gender differences in the context of adoption and sustained usage of technology in the workplace. User reactions and technology usage behavior were studied over a five-month period among 355 workers being introduced to a new software application. Men's attitude toward using the new technology was strongly motivated by achievement needs and task-oriented or instrumental behavior. In contrast, women's attitude toward using the new technology was more strongly influenced by subjective norm and perceived behavioral control. Sustained technology usage behavior in both men and women was driven by early usage behavior, emphasizing the importance of gender-based early evaluations of a new technology.

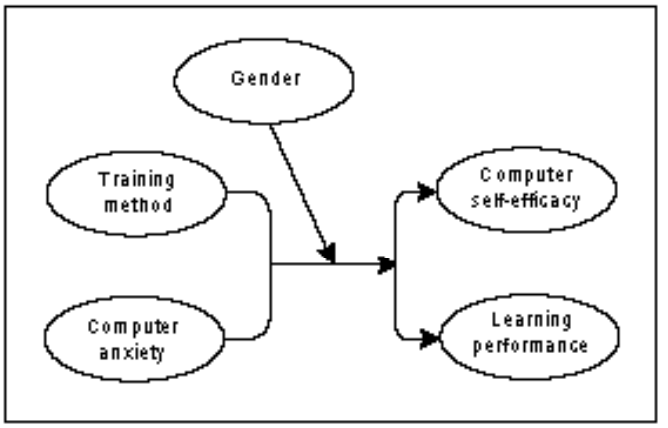
The above studies suggest several implications: when marketing IT and considering its effects, the gender of the users should be considered; and when training users on particular information technology such as e-mail systems, groups composed primarily of women should be addressed in a different manner than mixed or mainly masculine groups. When training mostly female groups, user friendliness and the ability of the system to convey the presence of the communicator should be emphasized.

Computer training has been widely researched and considered an essential contributor to the success of organizational computing. Computer anxiety and attitudes toward computers have often been identified as the critical factors influencing computer learning performance and training methodology. From the late 1980s up to today, the relationship of training techniques and personal characteristics has played a key role in training end users of information systems.

Characteristic of such training studies, Chou (2001) developed a conceptual model to evaluate how training method, an individual's gender, and computer anxiety level affect learning performance and computer self-efficacy. Chou used two types of training methods: an instruction-based method and a behavior-modeling method. The instruction-based method is a traditional approach that teaches primarily by lecture and follows a deductive way to learning, where learners proceed from general rules to specific examples (Davis & Davis, 1990; Simon, Grover, Teng & Witcomb, 1996). On the other hand, the behavior-modeling approach is a task-focused method involving a visual observation of the behaviors of a user performing a task. Learners then imitate and extend the demonstrated behavior in practice and experimentation to master the task. The behavior-modeling method employs an inductive approach that teaches by hands-on demonstrations first, followed by complimentary lectures (Compeau & Higgins, 1995; Gist, Schwoerer & Rosen, 1989). Figure 3 shows Chou's research model. In this model, gender was proposed as a moderating variable that moderates the effects of training method and computer anxiety on both learning performance and computer self-efficacy. When the training methods were tested on students, gender effects in general were found to be significant: male subjects performed better then female subjects. Male students had better learning performance, higher computer self-efficacy, and lower computer anxiety. Female students had significantly lower self-efficacy and a lower self-image about their computer learning capabilities. The behavior-modeling method appeared to enhance the computer self-efficacy of male students, whereas the instruction-based method benefited female students, suggesting that instruction technique preference varies with gender.

Studies on the relationships between computer anxiety, learning performance, and gender contribute solid knowledge about end-user training potential. This knowledge aids educators and trainers in the development of more personalized

Figure 3. Moderating effects of gender on computer self-efficacy and learning performance (Chou, 2001)



and therefore effective training programs for groups and individuals (Bostrom, 1998; Chou, 2001; Davis & Davis, 1990; Santhanam & Sein, 1994).

FUTURE TRENDS

In the developing research area of profiling Internet users, extrapolation of the computer anxiety literature offers parallels between gender differences in Internet use and differences in expertise and attitudes towards computers. It is widely assumed that women's participation in the Internet is hampered by their attitudes towards computers, which in turn is reflective of their attitudes towards new technology (Durndell & Haag, 2002; Gackenback, 1998; Jackson, Ervin, Gardner & Schmitt, 2001; Kraut et al., 1998; Schumacher & Morahan-Martin, 2001; Weiser, 2000).

Supporting this claim, Durndell and Haag (2002) utilized a Computer Self-Efficacy Scale, a Computer Anxiety Scale, and an Attitude to the Internet Scale in obtaining information from 74 female and 76 male university students on their Internet usage. Durndell and Haag found significant gender effects throughout, with males tending to report greater computer self-efficacy, lower computer anxiety, more positive attitudes towards the Internet, and longer use of the Internet than females. Similarly, Schumacher and Morahan-Martin (2001) reported males feeling more comfortable and competent with computers and the Internet. One explanation is that males, from childhood on, have more experience with computers than females, especially with games and programming that enhance technological sophistication and increase overall levels of competence and comfort with computers. Sussman and Tyson (2000) suggested the nature of communication on the Internet may vary by gender, and Balka and Smith (2000) proposed gender differences in Web navigation strategies.

Employing psychological methodology, DeYoung and Spence (2004) designed an instrument, the Technology Profile Inventory (TPI), to profile information technology users for dynamic personalization of software interfaces. Their approach was to generate a broad range of items for assessing responses to information technology, to examine their factor structure in a normal population, and to investigate potential associations between the emergent factors and variables that have, in the past, been associated with responses to computers, including gender, age, experience with information technology, and use of information technology. In application, a Web page could conform itself to suit the technology profile of each user who encounters it, for instance, a program could display all of its options for someone high in "computer interest" while displaying only the most functional options for someone who just wants to accomplish a task as simply as possible.

CONCLUSION

Computers are a vital asset in today's business and education world. The emergence of computers and information systems has been perhaps the single largest factor influencing organizations during the past three decades. Despite the increasing dispersal of computers, there is significant evidence that individual computer usage is affected by the computer anxiety or fear of computers that is widespread, and negative attitudes towards computers in general. This suggests that the potential benefits of computers as aids to professionals may not be fully realized, and the success of using a computer is dependent on the user's acceptance and commitment. The presence of computer anxious individuals in the workplace can lead to performance problems, decline in motivation, work quality, and moral, and can increase errors,

absenteeism, interpersonal conflicts, and turnover (Brosnan, 1998; Mikkelsen, Ogaard, Lindoe & Olsen, 2002).

Historically, a commonly held stereotype of computer anxiety is that of a frightened female secretary, struggling to learn the new word processor that her (male) boss is making her use as a replacement for her old, trusty IBM Selectric typewriter. The majority of gender and computer anxiety research does not support this description—both the secretary and the boss in this story may be anxious, but for different reasons and with different implications. In this particular case role has a greater effect than gender. However, the confusing continuum of disagreement characterizing the role of gender as it relates to attitudes toward computers, and by extension, IT diffusion and learning performance in the workplace, carries on today. While men may still represent a majority of the IT workforce, the number of women in technology-oriented areas continues to rise. As a result, the implementation of new technology requires an understanding of the factors that are likely to lead to user acceptance and sustained usage across gender and experience levels.

Does gender matter when examining attitudes and anxiety toward computers? The only valid conclusion that can be drawn from the existing body of literature is: it is important, but we do not fully understand gender as a moderating variable in the context of computer anxiety. One standard recommendation is to do more research and bring to the forefront the need to be cognizant of sex differences. The hypothesis that some users will have less overall experience with computers and are therefore more likely to have negative attitudes towards computers should be kept in mind when creating, adopting, and using information technology systems. Also, if cognitive style is distributed differently across genders, which it currently appears to be, it too may be a relevant factor in IT diffusion. One way to ameliorate these differences is for new technology introductions to be accompanied by user involvement, training, and active practical use. Special attention should be paid to the user's sex and their experience level. Because of the ramifications in society, education, and the workplace, educators and managers need to know how to recognize computer anxiety and the strategies to help alleviate or eliminate it. It is important that future research has focus and direction, and answers larger questions, such as: understanding how high levels of computer anxiety develop; what role such anxiety plays in career choices; and constructing methods to reduce computer anxiety within technological environments.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Anderson, A.A. (1996). Predictors of computer anxiety and performance in information systems. *Computers in Human Behavior, 12*(1), 61-77.

Balka, E. & Smith, R. (Eds.). (2000). *Women work and computerization*. Boston: Kluwer.

Bem, S.L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*, 155-162.

Beyer, S. (1999). The accuracy of academic gender stereotypes. *Sex Roles, 40*, 787-813.

Brosnan, M.J. (1998). The impact of computer anxiety and self-efficacy upon performance. *Journal of Computer Assisted Learning, 14*, 223-234.

Brown, R.P. & Josephs, R.A. (1999). A burden of proof: Stereotype relevance and gender differences in math performance. *Journal of Personality and Social Psychology, 76*(2), 246-257.

Chipman, S.F., Krantz, D.H. & Silver, R. (1992). Mathematics anxiety and science careers among able college women. *Psychological Science, 3*(5), 292-295.

Chou, H.W. (2001). Effects of training method and computer anxiety on learning performance and self-efficacy. *Computers in Human Behavior, 17*, 51-69.

Chua, S., Chen, D. & Wong, P. (1999). Computer anxiety and its correlates: A meta analysis. *Computers in Human Behavior, 15*(5), 609-623.

Cockburn, C. & Ormrod, S. (1993). *Gender and technology in the making*. London: Sage Publications.

Coffin, R. & Machintyre, P. (2000). Cognitive motivation and affective processes associated with computer-related performance: A path analysis. *Computers in Human Behavior, 16*(2), 199-222.

Colley, A.M., Gale, M.T. & Harris, T.A. (1994). Effects of gender role identity and experience on computer attitude components. *Journal of Educational Computing Research, 10*(2), 129-137.

Compeau, D.R. & Higgins, C.A. (1995). Application of social cognitive theory to training for computer skills. *Information Systems Research, 6*(2), 118-143.

Compeau, D., Higgins, C.A. & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. *MIS Quarterly, 23*(2), 145-158.

Dambrot, F.H., Watkins-Malek, M.A., Silling, M.S., Marshall, R.S. & Garver, J.A. (1985). Correlates of sex differences in attitudes toward and involvement with computers. *Journal of Vocational Behavior, 27*, 71-86.

Davis, F. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340.

Davis, D.L. & Davis, D.F. (1990). The effect of training techniques and personal characteristics on training end users of information systems. *Journal of Management Information System*, *7*(2), 93-110.

DeYoung, C.G. & Spence, I. (2004). Profiling information technology users: En route to dynamic personalization. *Computers in Human Behavior*, *20*, 55-65.

Durndell, A. & Haag, Z. (2002). Computer self-efficacy, computer anxiety, attitudes towards the Internet and reported experience with the Internet, by gender, in an East European sample. *Computers in Human Behavior*, *18*, 521-535.

Faulkner, W. (2001). The technology question in feminism: A view from feminist technology studies. *Women's Studies International Forum*, *24*(1), 79-95.

Gachenback, J. (Ed.). (1998). *Psychology and the Internet: Intrapersonal, interpersonal and transpersonal implications*. New York: Academic Press.

Gefen, D. & Straub, D.W. (1997). Gender differences in the perception and use of e-mail: An extension of the technology acceptance model. *MIS Quarterly*, *21*(4), 389-400.

Gilroy, F.D. & Desai, H.B. (1986). Computer anxiety: Sex, race and age. *International Journal of Man- Machine Studies*, *25*, 711-719.

Gist, M.E., Schwoerer, C. & Rosen, B. (1989). Effects of alternative training methods on self-efficacy and performance in computer software training. *Journal of Applied Psychology*, *74*, 884-891.

Gutek, B.A. & Bikson, T.K. (1985). Differential experience of men and women in computerized offices. *Sex Roles*, *13*(3/4), 123-136.

Hackbarth, G., Grover, V. & Yi, M.Y. (2003). Computer playfulness and anxiety: Positive and negative mediators of the system experience effect on perceived ease of use. *Information & Management*, *40*, 221-232.

Hayek, L.M. & Stephens, L. (1989). Factors affecting computer anxiety in high school computer science students. *Journal of Computers in Mathematics and Science Teaching*, *8*(4), 73-76.

Hofstede, G. (1980). *Culture's consequences: International differences in work related values*. London: Sage Publications.

Howard, G.S. (1986). *Computer anxiety and the use of microcomputers in management*. Ann Arbor, MI: UMI Research Press.

Howard, G.S. & Smith, R. (1986). Computer anxiety in management: Myth or reality? *Communications of the ACM*, *29*(7), 611-615.

Igbaria, M. & Parasuraman, S. (1989). A path analytic study of individual characteristics, computer anxiety and attitudes toward microcomputers. *Journal of Management*, *15*, 373-388.

Jackson, L., Ervin, K., Gardner, P. & Schmitt, N. (2001). Gender and the Internet: Women communicating and men searching. *Sex Roles*, *44*(5/6), 363-379.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukopadhyay, T. & Scherlis, W. (1989). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, *53*(9), 1017-1031.

Lee, R.S. (1970). Social attitudes and the computer revolution. *Public Opinion Quarterly*, *34*, 53-59.

Marcoulides, G.A. (1988). The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research*, *4*, 151-158.

Mikkelsen, A., Ogaard, T., Lindoe, P. & Olsen, O. (2002). Job characteristics and computer anxiety in the production industry. *Computers in Human Behavior*, *18*, 223-239.

Morrow, P.C., Prell, E.R. & McElroy, J.C. (1986). Attitudinal and behavioral correlates of computer anxiety. *Psychological Reports*, *59*, 1199-1204.

Raub, A.C. (1981). *Correlates of computer anxiety in college students.* Unpublished doctoral dissertation, University of Pennsylvania, USA.

Riding, R.J. & Rayner, S. (1998). *Cognitive styles and learning strategies: Understanding style differences in learning and behavior*. London: D. Fulton Publishers.

Rosen, L.D. & Maguire, P. (1990). Myths and realities of computer phobia: A meta-analysis. *Anxiety Research*, *3*, 175-191.

Rosen, L.D., Sears, D.C. & Weil, M.M. (1987). Computerphobia. *Behavior Research Methods, Instruments & Computers*, *19*, 167-179.

Santhanam, R. & Sein, M.K. (1994). Improving end user proficiency effects of conceptual training and nature of interaction. *Information Systems Research*, *5*, 378-399.

Schumacher, P. & Morahan-Martin, J. (2001). Gender, Internet, and computer attitudes and experiences. *Computers in Human Behavior*, *17*(1), 95-110.

Simon, S.J., Grover, V., Teng, J.T. & Whitcomb, K. (1996). The relationship of information system training methods and cognitive ability to end-user satisfaction, comprehension,

and skill transfer: A longitudinal field study. *Information Systems Research*, *7*(4), 466-490.

Straub, D.W. (1994). The effect of culture on IT diffusion: E-mail and FAX in Japan and the U.S. *Information Systems Research*, *5*(1), 23-47.

Sussman, N. & Tyson, D. (2000). Sex and power: Gender differences in computer mediated interactions. *Computers in Human Behavior*, *16*(4), 381-394.

Venkatesh, V., Morris, M.G. & Ackerman, P.L. (2000). A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. *Organizational Behavior and Human Decision Processes*, *83*(1), 33-60.

Webster, J. & Martocchio, J.J. (1992). Microcomputer playfulness: Development of a measure with workplace implications. *MIS Quarterly*, *16*, 201-226.

Weiser, E. (2000). Gender differences in Internet use patterns and Internet application preferences: A two-sample comparison. *CyberPsychology and Behavior*, *3*(2), 167-178.

Whitely, B. (1997). Gender differences in computer related attitudes and behavior: A meta analysis. *Computers in Human Behavior*, *13*(1), 1-22.

Witkin, H., Moore, C., Goodenough, C. & Cox, P. (1977). Field dependent and field cognitive styles and their educational implications. *Review of Educational Research*, *47*, 1-64.

Zmud, R.W. (1979). Individual differences and MIS success: A review of the empirical literature. *Management Science*, *25*(10), 966-979.

KEY TERMS

Cognitive Style: Information processing habits that represent an individual's typical modes of perceiving, thinking, remembering, and problem solving. Various cognitive styles have been identified, measured, and shown to affect the manner in which individuals perceive their environments. As just one example, two such styles are field-independence and field-dependence. Field-independent individuals perceive objects as separate from the field, impose personal structures on the environment, set self-defined goals, work alone, choose to deal with abstract subject matter, are socially detached and rely on their own values, and are self-reinforcing. In contrast, field-dependent individuals tend to rely on the environment for clues about an object, prefer a structure provided by the environment, experience the environment more globally, are interested in people, use externally defined goals, receive reinforcement from others, focus on socially oriented subject matter, and prefer to work with others (Riding & Rayner, 1998; Witkin, Moore, Goodenough & Cox, 1977).

Computer Anxiety: The tendency of a particular individual to experience a level of uneasiness over his or her impending use of a computer, which is disproportionate to the actual threat presented by the computer. Computer anxiety, defined by Raub (1981), is "the complex emotional reactions that are evoked in individuals who interpret computers as personally threatening."

Computer Anxiety Rating Scale (CARS): A self-report inventory consisting of 10 statements designed to measure computer anxiety. The scale comprises a mix of anxiety-specific statements (e.g., "I feel apprehensive about using the computer") and positive statements (e.g., "I am confident that I could learn computer skills") (Raub, 1981).

Computer Self-Efficacy: Computer self-confidence or perceptions of ability. Beliefs about one's ability to perform a specific behavior or task on a computer.

Locus of Control: Individuals' perceptions of whether they themselves influence events and outcomes in their lives (internal control), or that events and outcomes are influenced by factors such as luck, fate, chance, or powerful others (external control). Locus of control is considered a trait characteristic that is unlikely to change significantly in an individual's lifetime.

Math Anxiety: The psychological fear or anxiety associated with engaging in mathematical activity. Characteristics of math anxiety are an above-average number of negative attitudes (e.g., nervousness, solitude, uneasiness, and low confidence) and/or intense emotional reactions to math based on past experiences. Math anxiety and test anxiety are generally significant correlates and somewhat resemble computer anxiety as a situational manifestation of a general anxiety construct.

Perceived Ease of Use (PEOU): The degree to which an individual believes that using a particular information technology system would be free of effort. An application perceived to be easier to use than another is more likely to be accepted by users (Davis, 1989).

Perceived Usefulness (PU): The degree to which an individual believes that using a particular information technology system would enhance his or her job performance. A system high in perceived usefulness is one that a user believes has a positive usage to performance relationship (Davis, 1989).

Social Presence/Information Richness Factor (SPIR): A factor appended to TAM derived from Hofstede's (1980) work on dimensions of cultural differences among countries

that include a disposition toward masculine attitudes and other behavioral indexes. The extension combines perceived social presence and the sense of human contact embodied in a medium with the information richness of the medium (Straub**,** 1994).

Technology Acceptance Model (TAM): A causal model hypothesizing that actual information technology system use is affected by behavioral intentions that themselves are affected by attitudes toward use. Beliefs about the system, perceived usefulness, and perceived ease of use in TAM directly affect attitudes toward use (Davis, 1989).

Technology Profile Inventory (TPI): A psychological instrument that generates technology profiles to predict how individuals are likely to respond to various aspects of information technology. The ability to profile information technology users facilitates the design of software capable of dynamic personalization (DeYoung & Spence, 2004).

Theory of Planned Behavior (TPB): Defines relationships among beliefs, attitude toward a behavior, subjective norm, perceived behavioral control, behavioral intention, and behavior. The theory has been widely applied across a range of disciplines such as marketing, consumer and leisure behavior, medicine, and information technology. When applied in technology adoption and usage contexts, TPB explains an individual's adoption of new technologies (Ajzen**,** 1991).

Trait Anxiety: Traits are properties of individuals that dispose them to react in certain ways in given classes of situations. Trait anxiety is a chronic predisposition to be anxious and nervous that may be based on feelings of inadequacy, usually due to poor past performances, low-self image, or low-self esteem.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1257-1265, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Genetic Algorithms in Multimodal Search Space

Marcos Gestal
University of A Coruña, Spain

Julián Dorado
University of A Coruña, Spain

INTRODUCTION

Genetic algorithms (GAs) (Holland, 1975; Goldberg, 1989) try to find the solution for a problem using an initial group of individuals—the population—where each one represents a potential solution.

Actually they are successfully applied in very different and actual fields (Yang, Shan, & Bui, 2008; Yu, Davis, Baydar, & Roy, 2008); nevertheless, GAs have some restrictions on a search space with more than a global solution or a unique global solution, together with multiple local optima. A classical GA faced with such a situation tends to focus the search on the surroundings of the global solution; however, it would be interesting to know a higher number of possible solutions for several reasons: precise information about the search space, easy implementation of the local solutions compared with the global one, simple interpretation of certain solutions compared with others, and so forth. To achieve that knowledge, an iterative process will be executed until reaching the desired goals. Such process will start with the grouping of the individuals into species that will independently search a solution in their environments; following, the crossover operation will involve individuals from different species in order not to leave unexplored any search space area. The process will be repeated according to the goals achieved.

BACKGROUND

Multimodal Problems

There are problems that do not exclusively have a unique global solution, but they have multiple optima, either global or local: the multimodal problems (Ehrgott, 2005).

For dealing with such type of problems, it is interesting to know the higher possible number of solutions. On one hand, the knowledge about the problem might not be complete; this fact leads to the uncertainty about the goodness of the obtained solution, as it cannot be guaranteed that no better solutions might be found at the unexplored search space. On the other hand, and even achieving the best solution, there might be other possible solutions that, due to different reasons (economy, simplicity), might be preferable.

BRIEF INTRODUCTION TO GENETIC ALGORITHMS

GAs are adaptive methods, generally used in problems of search and of parameter optimization, based on sexual reproduction and on the "survival of the fittest" theory (Tomassini, 1995; Beasley, Bull, & Martin, 1993; De Jong, 2002).

A population, a group of individuals where each one represents a potential solution that will evolve through different generations, is initially created. The best individual would tend to be kept after several evolutions, but other less-fitted individuals will be also kept in order to keep diversity. The diversity will enable that there might be individuals with different characteristics that could, some of them, be suitably adapted to the eventual changes on the environment.

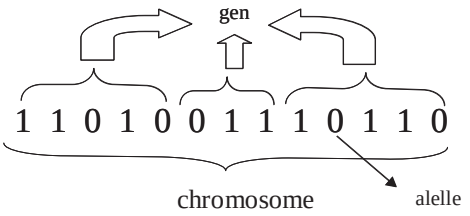
The best-fitted natural individuals are those that have more possibilities of having descendants, following the natural selection principles proposed of Darwin (1859).

In nature, individuals usually establish different groups, each of them specialized in different tasks: hunting, harvesting, and so forth. But the traditional GAs do not envisage this possibility for reaching several solutions; the present work will study an extension that will bear this in mind.

Problems Encoding

Any potential solution to a problem can be represented by providing values for a series of parameters. The whole of these parameters (or genes) is codified by a strand of values: the chromosome. The encoding is usually done by means of binary values, although other representations can also be used (see Figure 1).

Figure 1. Genetic individual



Main Algorithm

The generic functioning of a classic GA can be observed in Figure 2. A generation is obtained from a previous one by means of the reproduction operators. There are two types: the crossover and the copy. The crossover is a sexual reproduction that originates new descendants after the exchange of the genetic information of the parents. In the second case, a given number of individuals pass, with no variation, to the following population. Once the new individuals have been generated, the mutation is carried out with a P_m probability in such way that the transcription failures, which occurred in the copy of the genetic material during the sexual reproduction, are mimicked.

The GA run finishes when there are solutions good enough shaped as best individuals, when they all concur on a similar value, or when a prefixed generation's maximum number is reached.

For the GA to work correctly, a method should also exist that might indicate whether the population individuals represent or not, and to what extent, good solutions for the problem put forward. The later task would be carried out by the evaluation function, which establishes a numerical measurement (fitness) of the solution goodness (Koza, 1992).

EVOLUTIONARY APPROACHES TO MULTIMODAL PROBLEMS

Several approaches related to evolutionary techniques have been tried. A brief summary is shown in this section.

The modification proposed here is based on niching techniques. These are techniques that try to make and maintain stable subpopulations in GAs. This idea comes from nature, where individuals have different roles that allow them to survive in their natural ecosystems. These roles are called "ecological niches." Given some maximums, and given a limited capacity to locate them, the best niching algorithm will choose the global maximum. In addition, since they are not selective, they will keep both global and local maximums.

Different niching techniques have emerged throughout time. Some of the most important are:

- *Fitness Sharing:* This was first implemented by Goldberg and Richardson (1987) for its use with multimodal functions. This technique uses the concept of similarity

Figure 2: GA pseudocode

```
initialise current population randomly
WHILE the termination criterion is not fulfilled
 create temporal empty population
 WHILE temporal population don not fill
    select parents
    cross parents with Pc probability
   IF the crossover has occurred
               mutate one of the descendants with Pm probability
         assess descendants
        add descendants to the temporal population
   ELSE
      add parents to the temporal population
   END IF
END WHILE
increase generation counter
establish the temporal population as current new population
END WHILE
```

to determine the level of sharing between individuals of the same population. The fitness is scaled according to the similarity level with other individuals of the same population. It has two components: distance function (which measures the overlapping between individuals) and sharing function (returns "1" if the individuals are identical, or values closer to "0" as the difference between individuals increases). Based on this technique, Miller and Shawn (1995) propose a variant using a dynamic variation of the fitness.

- *Cavicchio's Pre-Selection:* This pre-selection is applied in Cavicchio (1970) so the good descendants replace one of their parents to maintain diversity. Therefore, the good individuals are favored respect by their neighbors, being later affected by crossovers and mutations. This is required by the complexity of the used chromosomes.
- *De Jong's Crowding:* The replacement of the existing individuals with the descendants is based on their similarity. The new individual is compared with a random subset of the population. The individual of this subset that presents more similarity with the new one is replaced. This technique is based on the natural effect produced when two similar individuals compete for the same resources in natural populations(De Jong, 1975). There are other variants of crowding. Among them, one of the most important is Mahfoud's (1992). An example of a more recent approach is found in Thomsem (2004).

These techniques have several limitations, some of them very common, like the necessity of calibrating the functions that control the niching process. This requires an a priori knowledge of the problem environment. A more important problem is that it is necessary to have a definition of similarity between individuals, but if the problem is not well understood, it will be complicated to specify a good function. As will be shown below, a grouping algorithm has been adopted here, independently of the problem domain.

Another type of technique used for the resolution of multimodal problems is *particle swarm optimization* (PSO) (Kennedy & Eberhart, 1995, 2001; Clerc, 2004; Iwamatsu, 2005). This technique is based in the social behavior of several organisms, like bird flocking or fish schooling. The search procedure is based in the changes of the positions of the individuals, named particles, through the multidimensional search space. During its movement, each particle updates its position according to its experience and to the experience of a neighbor particle, making use of the best position found by it and its neighbor. Therefore, it combines local and global search methods, trying to find a balance between exploration and exploitation.

GENETIC ALGORITHMS WITH DIVISION INTO SPECIES

During the course of evolution, the individuals tend to gather themselves into different species that will be adapted to a given environment and will evolve differently. It would be as much optimal individuals as species. Besides, sometimes the individuals, due to different reasons, will split from the initial group for them to experience crossover with individuals belonging to other groups from different places.

When a multimodal problem is intended to be solved, it is interesting to find several solutions. The use of traditional GA is not optimum for that scenario because, as generations advance, they tend to focus the search near the best solution. However, there are several options for at least minimizing this problem.

The simpler and most immediate (but inefficient) option would be the execution of the GA several times. This alternative randomizes the management of the problem to a great extent; other disadvantages are the repetition of solutions or the time-consuming execution until reaching the goal.

A more efficient technique, from the computational and solution quality viewpoint, implies the grouping of individuals of the genetic population into species; in such way, each of these species will carry out the search of a different solution.

Broadly speaking, the grouping technique into species entails gathering the initial population into groups of similar individuals. It is hoped that every group could specialize in a given search space area. Each species will search a solution existing in its surroundings that will be different from the ones provided by the rest of species.

The natural distribution of individuals into species and their separate evolution is tried to be mimicked. In nature, there are individuals adapted to cold, dry, or hot environments; each group keeps its life in a given environment by means of adapting specific characteristics that distinguish it from other groups.

Nevertheless, this technique is not free of disadvantages; certain requirements are needed for a good functioning, and they are not directly achieved during the initial arrangement of the problem. For instance, it should be desirable that the population could be perfectly distributed throughout the whole search space, as well as that the groups were fairly distributed along that space and in a quantity in accordance with the number of total solutions of the problem. Unexplored areas might exist if these characteristics were not present; in contrast, there might be another area highly explored where, depending on the grouping procedure, several species might coexist. Most of these problems can be avoided by means of an automated mechanism for the number of existing species. It seems clear that if the number of solutions is different for

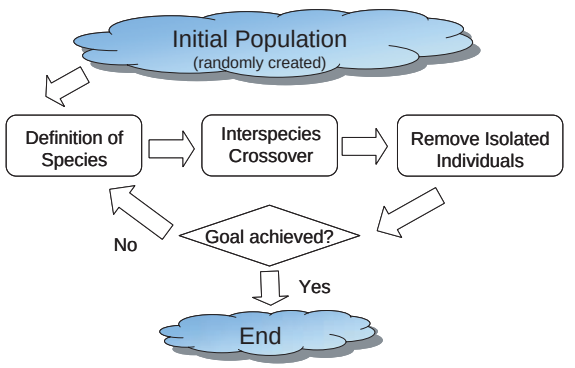
different problems, the GA should be the one who manages the number of evolved species in each generation.

In order to do this, the GA will be allowed to expand the starting number of species as generations advance. The increase of the species number will be achieved by performing crossover on individuals from different species throughout several generations. As the resulting descendants mix the knowledge from the species of their ancestors, a new species can be created in a different location from the ones of their progenitors. In this way, the species stagnation can be avoided and the exploration of new areas, together with the appearance of new knowledge, may be obtained; in short, environment diversity is achieved. As individuals might migrate or be expelled and afterwards create new, the performance of individuals is again modeled in their natural environment.

From an initial population generated for these techniques to be implemented, some following steps take place (see Figure 3):

1. random creation of the genetic population,
2. organization into species with similar characteristics,
3. application of the GA on every species,
4. introduction of new individuals coming from the crossover on different species (These individuals are located in another area of the search space. A variant of the functioning implies the elimination of the individuals that, after several generations, do not create a species large enough.); and
5. verifying whether the number of evolutions reaches the maximum allowed or if the population reaches a top level of individuals (defined at the beginning of the execution).

Figure 3. Overview of the proposed system



If some of these conditions are not fulfilled, the algorithm execution finishes; if they are, the individuals of the existing population will again be arranged into species (step 2).

As Figure 3 shows, the process is quite similar to standard parallel GAs. The main difference resides in points 3 and 4. Usually parallel GA involves different populations in isolated ways. After a predefined iteration count, an interchange of individuals is produced between populations. Here, there is only one population where GA differences between species instead of multiple populations are usual. So, crossover operations are able to produce individuals that will be inserted in the according population (different form original) when the grouping algorithm will be re-executed. So, species are really a mechanism to allow the execution of a set of iterations of an internal GA which involves only the individuals of that species. It allows improving the task of reaching the local optima.

Furthermore, as grouping algorithm is executed before each iteration, it allows that population size not be fixed; even the individuals within each population vary between iterations.

Grouping Algorithms

The grouping process of the genetic individuals into species is a key value in the proposed solution. Non-supervised grouping techniques have been chosen for carrying this grouping out. The classification is done according to concrete parameters that are specific for each type of grouping algorithm and that will later be discussed.

The most relevant of these heuristic methods are the *adaptive method* and the *algorithm of Batchelor and Wilkins,* also known as the *maximum distance algorithm* (Batchelor & Wilkings, 1969).

The adaptive method is a heuristic method that uses two intimately related parameters: τ, acting as the distance threshold for creating the groups, and θ, τ fraction that indicates up to what extent two individuals should have similarities for belonging to the same species. The basic functioning of the algorithm involves the incremental creation of groups; every one of the samples and a portion of the whole of patterns to be grouped will be individually processed. According to the distance threshold, it can be established whether they belong to an existing group or if a new group should be created.

The main advantage to be highlighted is its simplicity, as it is an algorithm that bases the grouping on simple comparisons and it does not additionally need the predetermination of the number of groupings. However, it has some disadvantages, the main one being that the grouping is biased by the patterns first used for the learning, meaning that it depends on the sequence of presentation.

The algorithm of Batchelor and Wilkins is also an incremental heuristic method that only uses a unique parameter, θ, which determines the mean distance among the existing

groups. It is used for calculating a distance threshold useful for deciding whether a new grouping should be created. As in the adaptive algorithm, θ value ranges as follows: $0 \leq \theta \leq 1$.

Basically, the algorithm creates a group if the distance from a given pattern to the nearest group exceeds the threshold value. Different from the adaptive algorithm, the distance threshold is not fixed and it is calculated according to the θ parameter and the mean distance among the existing groups.

Figures 4 and 5 show grouping examples of the use of grouping algorithms on the same initial sample set. It can be observed that, although both groups are not identical, they are quite similar.

The yellow area of Figure 4 is the acceptance area. The intermediate area is the uncertainty area. Despite that the figure represents the final stage, these mentioned areas are used as acceptance criteria throughout. In this way, it is considered that the individuals within the yellow area belong to that class and they have influence on its center. On the contrary, the individuals within the uncertainty area will induce the execution of the algorithm until the stabilization, until they cannot belong to any other group.

Tests

Once the approach for obtaining multiple solutions has been defined, it should be tested on different examples. The first

Figure 4. Adaptive method example

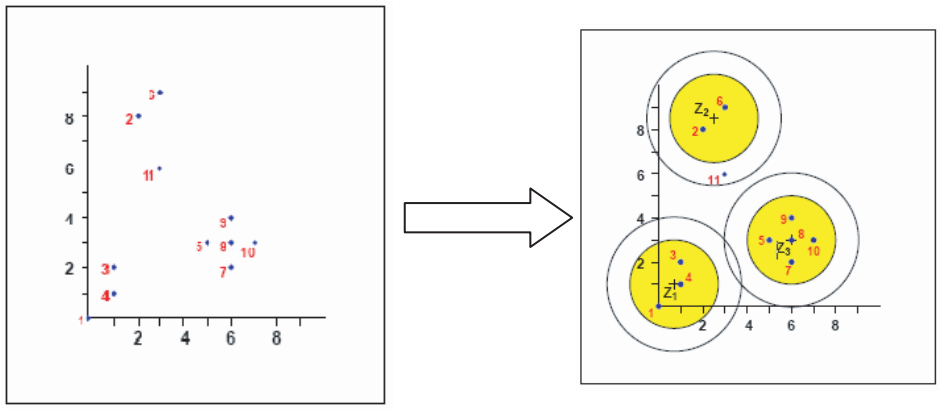
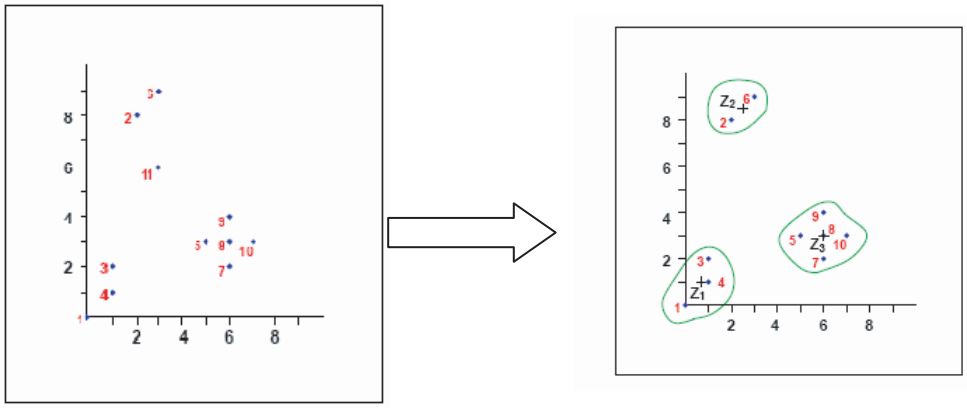


Figure 5. Batchelor & Wilkins algorithm example



of them will include an undetermined number of equations that, therefore, would have infinite possible solutions. This simpler example will facilitate the adjustment of the configuration to the different parameters that are involved in the process.

Resolution of an Undetermined Equation System

As has been mentioned, the proposed approach will be applied on an undetermined equation system (see Figure 6). The solution to such a system is the straight line that crosses the two planes, so the solutions will be infinite; it might be then considered a multimodal system.

The individuals ought to have (m-1) genes. Each one will represent the value designated for each unknown factor in order to solve the equation system.

The fitness value will be the addition of the absolute values of the differences between each of the independent terms specified in the system and the independent terms obtained after replacing the genotype values in the subsequent equation.

Several steps of the evolution are graphically represented in Figure 7. Each subfigure represents up until what extent the solutions provided by the GA are close to the straight line that represents the solution of the undetermined equation system. Every one of the red spots represents the place where the best individual is after all the foreseen generations. On the other hand, the appearance of a line indicates the movement of a species along that generation. If no line is present, it means that the change of location of the best individual along evolution is small. Figure 7 seems to show that the proposed solution is valid for providing multiple solutions to a problem.

The same example has been used for approximately determining the optimum GA execution parameters. Obviously, the final configuration will depend on the specific problem, but the values in Figure 8 could be perfectly used as a starting point. The validity of a given parameter-value partnership is tested by executing the application and checking not only the number of found solutions, but also the additive error of all the solutions.

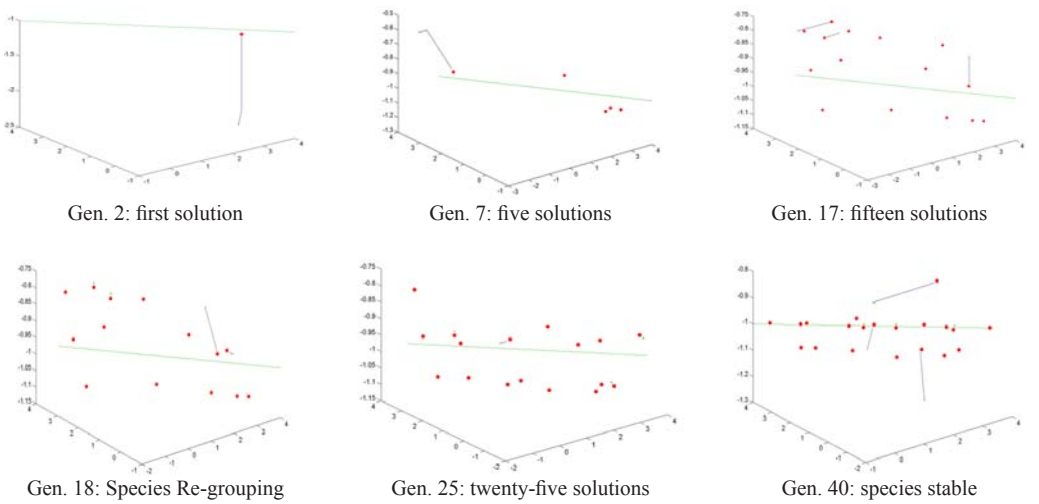
Figure 6. System representation

$$\begin{aligned} 2*x + 2*y + 6*z &= 0 \\ x - y + 25*z &= -22 \\ 2*x + 2*y + 6*z &= 0 \end{aligned}$$

Rastrigin Function

The Rastrigin function is a good example of a multimodal problem. The GA has been applied with the proposed modifications on a Rastrigin function with two variables, and it

Figure 7. Example of execution



Gen. 2: first solution | Gen. 7: five solutions | Gen. 17: fifteen solutions

Gen. 18: Species Re-grouping | Gen. 25: twenty-five solutions | Gen. 40: species stable

Figure 8. Optimum parameters

Selection Algorithm 1	**Roulette**	**Initial population size**	**200**
Selection Algorithm 2	**Random**	**Mutation probability**	**2%**
Crossover Algorithm	**1 point**	**Crossover probability**	**90%**
Replacement Algorithm	**Worst**	**Grouping algorithm**	**Adaptive (τ: 50.000; θ: 0.66)**

Figure 9. Rastrigin function: Example of execution

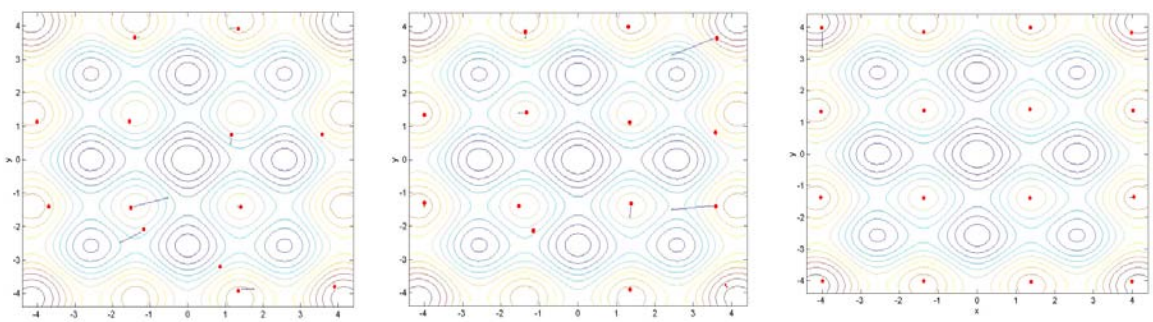
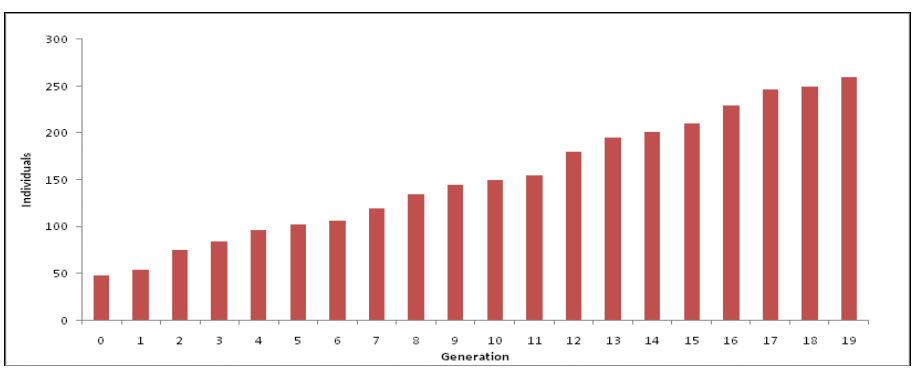


Figure 10. Rastrigin function: Evolution of the number of individuals



can be observed in the following example, where the results are clearly and legibly represented with graphs. In this case, the explored space is between X = [-4..4] and Y = [-4..4], where there are 16 maximums clearly identifiable.

Figure 9 shows that almost all the existing optimums of the specified search space are obtained after less than 20 generations. In this example is also interesting to see that the number of individuals increase as generations advance (see Figure 10). This increment is due to the fact that the inter-species crossovers create new individuals in unexplored areas of the search space and therefore originate new species as generations advance.

FUTURE TRENDS

The mutation operator is currently applied on the crossover-generated individuals. However, it should be desirable that it could be applicable on any randomly chosen individual in order to increase the diversity of the global population. This variant, and some others, should be implemented for providing the user with the option of choosing among them according to a specific problem.

Another aspect to be improved is the response time. In problems where the evolution of the individuals has high computational cost, it should be desirable to distribute the application execution among different machines in order to achieve a considerable reduction of time.

CONCLUSION

Due to the Gas' traditional trend of returning a unique solution for a given problem, they are not an optimal option for the search of solutions to multimodal problems. The present work has proposed the division of the population into species for them to individually search a solution in their own development areas. In this way, the solutions are found in a more organized way, not having to execute the GA with the whole of the population, but within each species.

Nevertheless, a problem arises when dividing the problem into species and applying the GA to them. If the species finds a solution in its area, it tends to remain there, as the exploration of its boundaries does not achieve anything better. After several evolutions, there would be species fixed on a solution, but others would not find any because there are no solutions in their surroundings. The crossover between different species avoids the stagnation. The resulting individuals will not strictly belong to the species of their progenitors, but they will belong to a different species. Therefore, new species will appear in order to increasingly cover, as much as possible, the complete search space. The validity of such approaches can be observed in the examples.

REFERENCES

Batchelor, B.G., & Wilkins, B.R. (1969, October 2). Method for location of clusters of patterns to initialise a learning machine. *Electronics Letters, 5*(20), 481-483.

Beasley, D., Bull, D.R., & Martin, R.R. (1993). An overview of genetic algorithms: Part 1, fundamentals. *University Computing, 15*(2), 58-69.

Cavicchio, D.J. (1970). *Adaptative search using simulated evolution.* PhD Thesis, University of Michigan, USA.

Clerc, M. (2004). Discrete particle swarm optimization. In B.V. Babu & G.C. Onwubolu (Eds.), *New optimization techniques in engineering* (pp. 219-239). Berlin: Springer-Verlag.

Darwin, C. (1959). *On the origin of species by means of natural selection.* London: John Murray.

De Jong, K.A. (1975). *An analysis of the behavior of a class of genetic adaptative systems.* PhD Thesis, University of Michigan, USA.

De Jong, K.A. (2002). *Evolutionary computation.* Cambridge, MA: MIT Press.

Ehrgott, M. (2005). *Multricriteria optimization* (2nd ed.). Berlin: Springer-Verlag.

Goldberg, D.E. (1989). *Genetic algorithms in search, optimization and machine learning.* Reading, MA: Addison-Wesley.

Goldberg, D.E., & Richardson, J. (1987). Genetic algorithms with sharing for multimodal function optimization. *Proceedings of the 2nd International Conference on Genetic Algorithms* (pp. 41-49). Hillsdale, NJ: Lawrence Erlbaum.

Holland, J.H. (1975). *Adaptation in natural and artificial systems.* Ann Arbor: University of Michigan Press.

Iwamatsu, M. (2005). Comparison of particle swarm and evolutionary programming as the global conformation optimizer of clusters. *International Journal of Modern Physics, 16*(4), 591-606.

Kennedy, J., & Eberhart, R.C. (1995). Particle swarm optimization. *Proceedings of the IEEE International Conference on Neural Networks* (pp. 1942-1945). Piscataway, NJ: IEEE Press.

Kennedy, J., & Eberhart, R.C. (2001). *Swarm intelligence.* San Francisco: Morgan Kaufmann.

Koza, J.R. (1992). *Genetic programming: On the programming of computers by means of natural selection.* Cambridge, MA: MIT Press.

Mahfoud, S.W. (1992). Crowding and preselection revisited. *Proceedings of the Conference on Parallel Problem Solving from Nature II* (pp. 27-36). New York : Elsevier Science.

Miller, B.L., & Shaw, M.J. (1995). *Genetic algorithms with dynamic niche sharing for multimodal function optimization*. IlliGAL Report No. 95010.

Thomsem, R. (2004). Multimodal optimization using crowding-based differential evolution. *Proceedings of the Congress on Evolutionary Computation.* New York: ACM.

Tomassini, M. (1995). A survey of genetic algorithms. *Annual Reviews of Computational Physics, 3,* 87-117.

Yang, A., Shan, Y., & Bui, L.T. (Eds.). (2008). *Success in evolutionary computation*. Berlin: Springer-Verlag.

Yu, T., Davis, L., Baydar, C., & Roy, R. (Eds.). (2008). *Evolutionary computation in practice.* Berlin: Springer-Verlag.

KEY TERMS

Diversity: Measure of the genotypic difference between different individuals. It is necessary to keep a high ratio of diversity to explore in depth the search space and avoid a premature gene convergence due to the genetic drift.

Evolutionary Technique: Technique that tries to provide valid solutions for a specific problem using concepts taken from nature or biology, such as survival of fittest.

Fitness: Value derived from the evaluation of an individual with respect to its reproduction capability. This measure determines the goodness of the solution encoded by an individual. Usually, selection in evolutionary algorithms depends on the fitness.

Genetic Algorithm: A special type of evolutionary technique where the potential solutions are represented by means of chromosomes (usually either binary or real sets of values). Each gene (or set of genes) represents a variable or parameter within the global solution.

Genotype: Set of values (real, binary, …) that codifies the internal solution representation to which the crossover and mutation operators are applied.

Multimodal Problem: Special kind of problem with several global solutions or one global solution with several local peaks instead of a unique global optimum.

Niching: Separation of individuals according to their states in the search space or maintenance of diversity by appropriate techniques, for example, local population models, fitness sharing, or distributed evolutionary algorithms.

Phenotype: Expression of the properties coded by the individual's genotype. The precise definition of phenotypes is mostly problem dependent. For parameter optimization the phenotype is usually identical with the object parameters, whereas for structure optimization (e.g., of neural networks) the phenotype represents a specific structure (in this case, the genotype represents the value parameters of the structure).

Search Space: Set of all possible situations of the problem that want to be solved. Combination of all the possible values for all the variables involved with the problem.

Species: Within the context of genetic algorithms, subset of genetic individuals with similar genotypes (genetic values) that explores a similar area in the search space.

Geographic Information Systems as Decision Tools

Martin D. Crossland
Oklahoma State University, USA

INTRODUCTION

Geographic information systems (GISs) as a technology have been studied and reported extensively and, not unexpectedly, in the field of geography. The various ways of capturing spatial data, arranging attribute data into appropriate database structures, and making the resulting large data sets efficient to store and query have been extensively researched and reported (Densham, 1991). However, the geographic research community has only recently noted the need to study how GISs are used as decision tools, especially with regard to how such decision making might be related to a decision maker's cognitive style (Mennecke, Crossland, et al., 2000). As an example, the University Consortium for Geographic Information Science called for research examining how geographic knowledge is acquired through different media and by users with different levels of experience and training (University Consortium for Geographic Information Science, 1996).

Researchers in the fields of decision sciences and information systems have more recently begun to make contributions in the area of decision making with GISs. When a GIS is employed as a decision support system, in these studies the resultant system is often referred to as a *spatial decision support system,* or *SDSS* (see Crossland, 1992; Crossland, Perkins, et al., 1995; Mennecke et al., 2000).

A *geographic information system* in its simplest form is a marriage of accurately scaled digital maps with a database. The digital maps comprise spatially referenced details such as natural elements (lakes, rivers, topographic elevation contours, etc.), manmade objects (buildings, roads, pipelines, etc.), and political boundaries (city limits, state and county lines, international boundaries, etc.). These natural elements are typically referenced, with varying degrees of precision, to latitude/longitude coordinates on the earth's surface. It must be noted here that the degree of precision and, more importantly, differences in degrees of precision for the various elements are the subjects of much research and user consternation in applications of GISs to solving problems. The database, in turn, catalogs information about the various spatial elements (e.g., the names of rivers, names of buildings, building owner, operator of a pipeline, etc.). These descriptive entries in the database are often referred to as *attributes* of the various spatial elements.

A GIS may be paired with the *global positioning system* (*GPS*), from which real-time, satellite-derived location information may be derived, as provided by an appropriate GPS receiver.

BACKGROUND

With regard to the effectiveness of decision making when using information tools, there is a relatively long history of researchers emphasizing that tools which provide graphical presentations and graphical representations of information are deserving of special note and study. For example, Ives (1982) discussed at great length the role of graphics in business information systems. He even went so far as to state, "The map, perhaps more than any other chart form, gains the most from the availability of computer graphics" (p. 16).

Several more recent studies have drawn from Image theory (Bertin, 1983) to help explain why decision makers using GISs may experience greater effectiveness in decision making. Image theory states that one graphical representation of information may be considered more efficient than another for a particular question, if that question can be answered in the mind of the decision maker in a lesser amount of time. In his *Semiology of Graphics*, Bertin defined image theory and put forth the constructs of images and figurations. An *image* is a meaningful visual form, perceptible in a minimum instant of vision. A *figuration* is a more complex construction comprising multiple images. Figurations are inherently less efficient than images, according to image theory. This is because the viewer is able to grasp the full informational content of an image in a brief moment of viewing it. Figurations, on the other hand, comprise multiple images which must be mentally extracted, processed, and related in the viewer's perception. Although the informational content may be richer in a figuration, it is inherently less efficient for quick extraction of specific information.

The more recent studies propose that one role of GISs is to collapse more complex figurations into simpler figurations or even to simple images. This has the net effect of increasing a decision maker's efficiency in extracting relevant information for the purpose of evaluating and making a decision. For examples the reader is encouraged to review Crossland

(1992), Crossland, Herschel, et al. (2000), Crossland et al. (1995), and Mennecke et al. (2000).

Although there seems to be a common assumption that GISs improve decision making (Morrison, 1994), only a few studies to date have performed controlled experiments to actually test this assumption. Those that have been accomplished typically used dependent variables of decision time and decision accuracy to measure decision-making effectiveness. These include Crossland (1992), Dennis and Carte (1998), Mennecke et al. (2000), Smelcer and Carmel (1997), and Swink and Speier (1999). All of these studies found that the addition of a GIS to a spatially referenced decision-making task had a positive effect on decision outcomes.

THE ROLE OF COGNITIVE STYLE IN DECISION MAKING WITH GISS

With respect to decision making, the term *cognitive style* has been used to refer to enduring patterns of an individual's cognitive functioning that remain stable across varied situations. Various elements of cognitive style have been speculated upon and studied in various disciplines. With respect to decision making using GISs, two elements have been studied in some depth, *field dependence* and *need for cognition*.

Field dependence (FD) measures a person's ability to separate an item from an organized field or to overcome an embedded context (Witkin, Lewis, et al., 1954). Zmud and Moffie (1983) proposed that people with lower field dependence tend to outperform those with higher field dependence in structured decision tasks and that they tend to make more effective use of transformed information (e.g., aggregated values and graphical formats, such as are typically found in a GIS). FD can be measured using commercially available testing instruments. Because making decisions using a GIS, by its nature, involves mentally extracting relevant information from a potentially complex field of information, studies have hypothesized that low field dependence should predict better decision making with a GIS or other spatially referenced tool. In particular, field dependence is seen as an inverse proxy for an individual's level of spatial cognition—the ability of an individual to grasp and analyze information within a spatial context.

Need for cognition (NFC) was proposed by Caccioppo and Petty (1982) as a measure of a person's internal motivation to pursue and enjoy cognitive tasks and activities. They developed a questionnaire which can be used to measure this cognitive-style attribute. People who score high on the need for cognition scale tend to enjoy the engagement of thought activity in a task as much or more than even the result of a task. The studies named below hypothesized that this tendency to engage more fully in a task should lead to more effective decision making, as measured by the dependent variables of decision time and decision accuracy.

Studies that looked at FD, NFC, or both as independent variables of decision-making performance using GISs include Crossland (1992), Crossland et al. (1995), and Mennecke et al. (2000). In general, the findings may be summarized as follows:

- Field dependence exhibits an inverse main effect on decision time, but not on decision accuracy. That is, subjects with lower field dependence tend to solve spatially referenced problems more quickly, but not more accurately. It may be that the efficiency predicted by image theory does contribute to faster decision making, but not to more accurate decisions.
- Need for cognition exhibits a positive main effect on decision accuracy, but not on decision time. That is, higher-NFC subjects tend to solve spatially referenced problems more accurately, but not more quickly. This last finding was noted as unexpected by Crossland (1992). He speculated that perhaps an individual with a high NFC might tend to spend longer in thinking about the problem and its solution, thereby extending the decision time. It would seem, however, that this extra thinking effort may have contributed to a more accurate solution.

FUTURE TRENDS

Some questions and issues in this area of research that remain to be addressed include:

- How do other important measures of cognitive style affect a decision maker's ability to solve spatially referenced problems accurately and quickly?
- How does problem complexity factor in or even interact with the decision maker's task? Several studies also examined problem complexity as an independent variable (Crossland, 1992; Crossland et al., 1995; Mennecke et al., 2000). Crossland et al. (1995) reported an observed interaction of field dependence with problem complexity that would be interesting to explore further.
- To what extent are SDSSs/GISs effective in collapsing figurations (as defined by image theory) into images or into simpler figurations? Are there certain levels of complexity beyond which it becomes impractical or ineffective to combine or collapse displays into simpler decision tools? How does the cognitive style of the decision maker factor into this consideration?
- Are SDSSs/GISs even necessary for certain types of problems? Perhaps a series of static, hard-copy outputs are sufficient for some decisions by some decision makers, and the combined or flattened displays are not

necessary. The cognitive style of the decision maker may be an important factor in this.

- Studies in this area may be useful in understanding more generally how technology is useful in supporting decision makers in other contexts. For example, Vessey (1991) and Vessey and Galletta (1991) suggested that three variables would influence the mental representation that the decision maker develops: (1) the problem representation, (2) the problem-solving task, and (3) the decision maker's problem-solving skills.
- It may be useful to apply cognitive fit theory to these types of problem-solving tasks, as proposed by Mennecke et al. (2000), to better understand the factors that appear to be important in influencing a user's formation of a mental representation of spatial tasks.

CONCLUSION

The roles of components of cognitive style of a decision maker are important factors in how well he can solve problems and carry out spatially referenced tasks when using a GIS. Although current research has shown that simply using a GIS can enhance decision-making effectiveness, the cognitive style of the subject is a less well understood element of the process. More research is needed in this area to better define and understand it.

REFERENCES

Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps*. University of Wisconsin Press.

Caccioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *4*(1), 116-131.

Crossland, M. D. (1992). *Individual decision-maker performance with and without a geographic information system: An empirical investigation*. Unpublished doctoral dissertation, , Indiana University, Bloomington.

Crossland, M. D., Herschel, R. T., et al. (2000). The impact of task and cognitive style on decision-making effectiveness using a geographic information system. *Journal of End User Computing*, *12*(1), 14-23.

Crossland, M. D., Perkins, W. C., et al. (1995). Spatial decision support systems: An overview of technology and a test of efficacy. *Decision Support Systems*, *14*, 219-235.

Dennis, A. R., & Carte, T. (1998). Using geographical information systems for decision making: Extending cognitive fit theory to map-based presentations. *Information Systems Research*, *9*(2), 194-203.

Densham, P.J. (1991). Spatial decision support systems. In P. J. Densham, M. F. Goodchild, & D. W. Rhind (Eds.), *Geographical information systems: Principles and applications* (Vol. 2, pp. 403-412). London: Longman Scientific & Technical,

Ives, B. (1982). Graphical user interfaces for business information systems. *MIS Quarterly*, 15-42.

Mennecke, B. E., Crossland, M. D., et al. (2000). Is a map more than a picture? The role of SDSS technology, subject characteristics, and problem complexity on map reading, and problem solving. *MIS Quarterly*, *24*(4), 601-629.

Morrison, J. L. (1994). The paradigm shift in cartography: The use of electronic technology, digital spatial data, and future needs. In T. C. Waugh & R. G. Healey (Eds.), *Advances in GIS research* (pp. 1-15). London: Taylor and Francis.

Smelcer, J. B., & Carmel, E. (1997). The effectiveness of difference representations for managerial problem solving: Comparing tables and maps. *Decision Sciences*, *28*, 391-420.

Swink, M., & Speier, C. (1999). Presenting geographic information: Effects of data aggregation, dispersion, and users' spatial orientation. *Decision Sciences*, *30*(1), 169-195.

University Consortium for Geographic Information Science. (1996). Research priorities for geographic information science. *Cartography and Geographic Information Systems*, *23*(3), 1-18.

Vessey, I. (1991). Cognitive fit: Theory-based analysis of the graphs vs. tables literature. *Decision Sciences*, *22*(1), 219-241.

Vessey, I., & Galletta, D. (1991). Cognitive Fit: An empirical study of information acquisition. *Information Systems Research*, *2*(1), 63-84.

Witkin, H. A., Lewis, H. B., et al. (1954). *Personality through perception*. New York: Harper.

Zmud, R. W., & Moffie, R. P. (1983). The impact of color graphic report formats on decision performance and learning. *International Conference on Information Systems*.

KEY TERMS

Attributes: are the pieces of information contained in a GIS database that describe or detail a spatially referenced element.

Cognitive Style: refers to enduring patterns of an individual's cognitive functioning that remain stable across varied situations.

Digital Map: any form of geographic boundaries or spatially referenced drawings that have been captured, or "digitized," into an electronic form. Each element of the map is or may be linked to various descriptive or identifying types of information in a database.

Field Dependence (FD): measures a person's ability to separate an item from an organized field or to overcome an embedded context.

Figuration: as defined in image theory, is a complex construction comprising multiple images. Figurations are inherently less efficient for extracting information than images, according to image theory.

Geographic Information System (GIS): a marriage of accurately scaled digital maps with a database. The digital maps comprise spatially referenced details such as natural elements (lakes, rivers, topographic elevation contours, etc.), manmade objects (buildings, roads, pipelines, etc.), and political boundaries (city limits, state and county lines, international boundaries, etc.). These natural elements are typically referenced, with varying degrees of precision, to latitude/longitude coordinates on the earth's surface.

Global Positioning System (GPS): provides real-time, satellite-derived location information based on information received by an appropriate GPS receiver. GPS is funded by and controlled by the U.S. Department of Defense (DOD). While there are many thousands of civil users of GPS worldwide, the system was designed for and is operated by the U.S. military. A GPS may be employed in the original construction of the digital map information to be stored in a GIS. Or, if the GIS is already constructed, the GPS may be employed to accurately render the position of new elements to be added to the GIS or the current position of a mobile element to be referenced against the information stored in the GIS. A good example might be a freight truck moving on a highway. The GPS receiver on the truck can derive its current latitude and longitude and then send that information to the GIS system in the truck cab, to a GIS in a central control center via radio, or to both for subsequent reporting and analysis.

Image: as defined in image theory, is a meaningful visual form, perceptible in a minimum instant of vision.

Need for Cognition (NFC): a measure of a person's internal motivation to pursue and enjoy cognitive tasks and activities.

Spatial Decision Support System (SDSS): typically, a geographic information system (GIS) that has been extended to provide knowledge workers with decision-making tools and support data.

Spatially Referenced Data: entities, typically recorded as records in a database, which have some notion of a definite location in space.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1274-1277, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Geography and Public Health

Robert Lipton
Prevention Research Center, USA

D. M. Gorman
Texas A&M University, USA

William F. Wieczorek
Center for Health and Social Research, Buffalo State College-State University of New York, USA

Aniruddha Banerjee
Prevention Research Center, USA

Paul Gruenewald
Prevention Research Center, USA

INTRODUCTION

From John Snow's pioneering work on cholera in the 19th century until the present day, placing illness and disease within the context of a geographic framework has been an integral, if understated, part of the practice of public health. Indeed, geographical/spatial methods are an increasingly important tool in understanding public health issues. Spatial analysis addresses a seemingly obvious yet relatively misunderstood aspect of public health, namely, studying the dynamics of people in places. As advances in computer technology increase almost exponentially, computer intensive spatial methods (including mapping) have become an appealing way to understand the manner in which the individual relates to larger frameworks that compose the human community and the physical nature of human environments (streets with intersections, dense vs. sparse neighborhoods, high or low densities of liquor stores or restaurants, etc.). Spatial methods are extremely data intensive, often pulling together information from disparate sources that have been collected for other purposes such as research, business practice, governmental policy, and law enforcement. Although initially more demanding in regard to data manipulation compared to typical population level methods, the ability to compile and compare data in a spatial framework provides information about human populations that lies beyond typical survey or census research. We will discuss general methods of spatial analysis and mapping that will help to elucidate when and how spatial analysis might be used in a public health setting. This discussion will include a method for transforming arbitrary administrative units, such as zip codes, into a more useable uniform grid structure. In addition, a practical research example will be discussed focusing on the relationship between alcohol and violence. A relatively new Bayesian spatial method will be part of this example.

BACKGROUND: GIS CAPABILITIES AND PREVENTION

A basic understanding of the capabilities of geographic information systems (GISs) is critical to the development of prevention activities because alcohol-related problems are not evenly distributed across space. GIS can be defined as a combination of computer hardware, software, spatial data (digital maps), and data with a geographic reference (e.g., alcohol outlets or crime locations) that facilitates spatial analysis. The key functions of GIS provide access to the broad spectrum of potential spatial analyses that can support the simple targeting of resources as well as the development of more complex models of spatial interactions. Both simple maps of problem rates or clusters and spatial interaction models may be useful for targeting traditional individual-based prevention programs or environmental interventions. Spatial interaction models, however, may be more appropriate for identifying the locations of events (e.g., assaults or crashes) that may be most amenable to environmental or regulatory prevention. In addition, GIS capabilities promote the development of a basic spatial/geographic epidemiology of alcohol use and related consequences, which is critical to the development of prevention programs (see Wieczorek, 2000, and Wieczorek and Hanson, 1997, for more details).

The key functions of GIS include geocoding, data overlays, reclassification functions, and distance/adjacency measures. Geocoding is a generic term used to describe the GIS function of providing a specific location to descriptive data. Geocoding applies to point data (e.g., alcohol outlet)

as well as to area data (e.g., number of assaults in a census tract). Sometimes geocoding is known as address matching because the process of matching points to addresses is very common. The advent of the Census Bureau's TIGER system has made geocoding a relatively low cost and widely available GIS function. However, professional geocoding services have developed to assist persons who are not comfortable in geocoding their own data or because of the high cost of updating digital maps based on TIGER in areas of changing population. Geocoding is the most basic of GIS functions because it transforms descriptive information into a format suitable for spatial analysis.

A GIS-based map may consist of multiple sources of data. The ability to combine multiple layers of information is known as the overlay function. An example of an overlay function is to place geographic boundaries (such as the outline of a town) on top of individual points (such as residences of DWI offenders). The points within each area can then be automatically counted to create rate-based maps such as those shown in Figure 1. To create rate-based maps from relevant point information, at least three layers of data are necessary (i.e., map of the points, a map with relevant boundaries, and Census data on population). The ability to perform an intersection between separate maps, to aggregate data into meaningful geographic areas, and to link data to standard sources, such as Census data, highlights some of the processing capabilities of GIS overlay functions, processes nearly impossible to accomplish by non-automated methods (see Wieczorek & Hanson, 2000, for an example using regions and mortality data).

One major contribution of GIS to prevention is its ability to provide useful visualizations of spatial data. The reclassification function of GIS allows the user to easily manipulate the number of categories or select specific information (e.g., crashes by time of day or day of the week) for display. Figure 1 shows how the reclassification function can assist in the targeting of prevention by reclassifying the same data to emphasize highest rate areas.

A second major contribution of GIS to prevention is that the technology enables the development of models of spatial patterns and interactions within and between populations and environments. These models require accurate information on the distance between individual objects (e.g., bars and traffic crashes) and their spatial relationships to one another. Distance and adjacency functions of GIS allow assessments of these relationships. Data generated from these assessments of spatial relationships can be exported from the GIS and used in spatial modeling software. Information about adjacencies of different geographic objects can be used to assess contributions of environmental features (e.g., bars) to problematic public health outcomes in surrounding areas (e.g., assaults) (Gruenewald et al., 1996; Lipton & Gruenewald, 2002). Other important GIS functions are based upon the assessment of distance relationships: neighborhood functions calculate the number of a specific characteristic (e.g., assaults) within a specific radial distance (e.g., 300 yards) of

Figure 1. Reclassification and targeting prevention

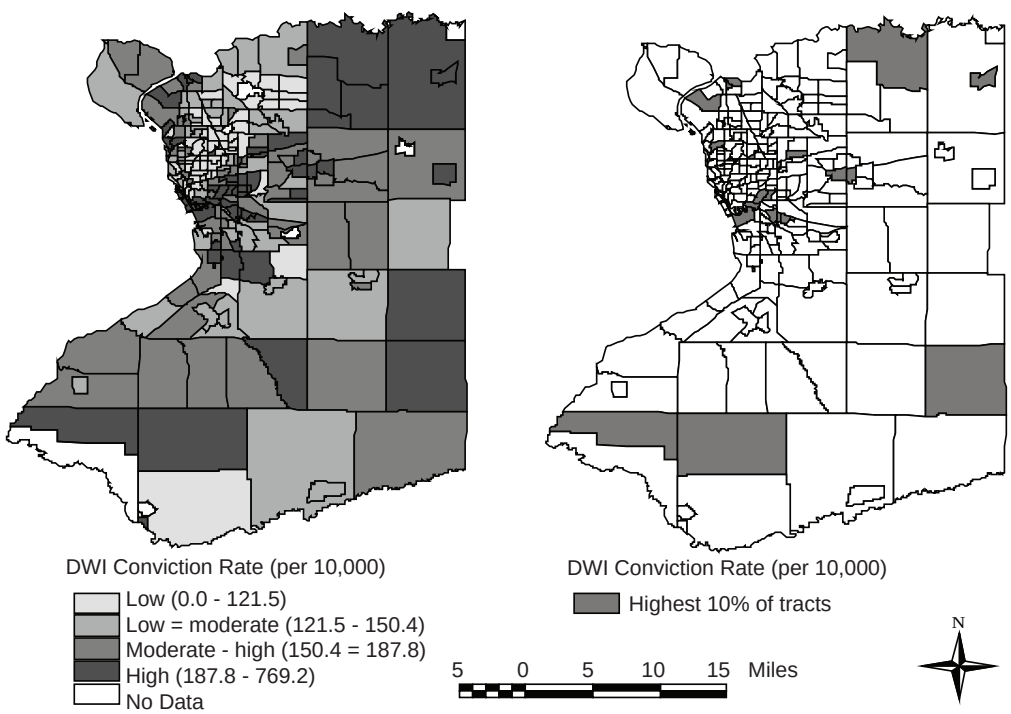
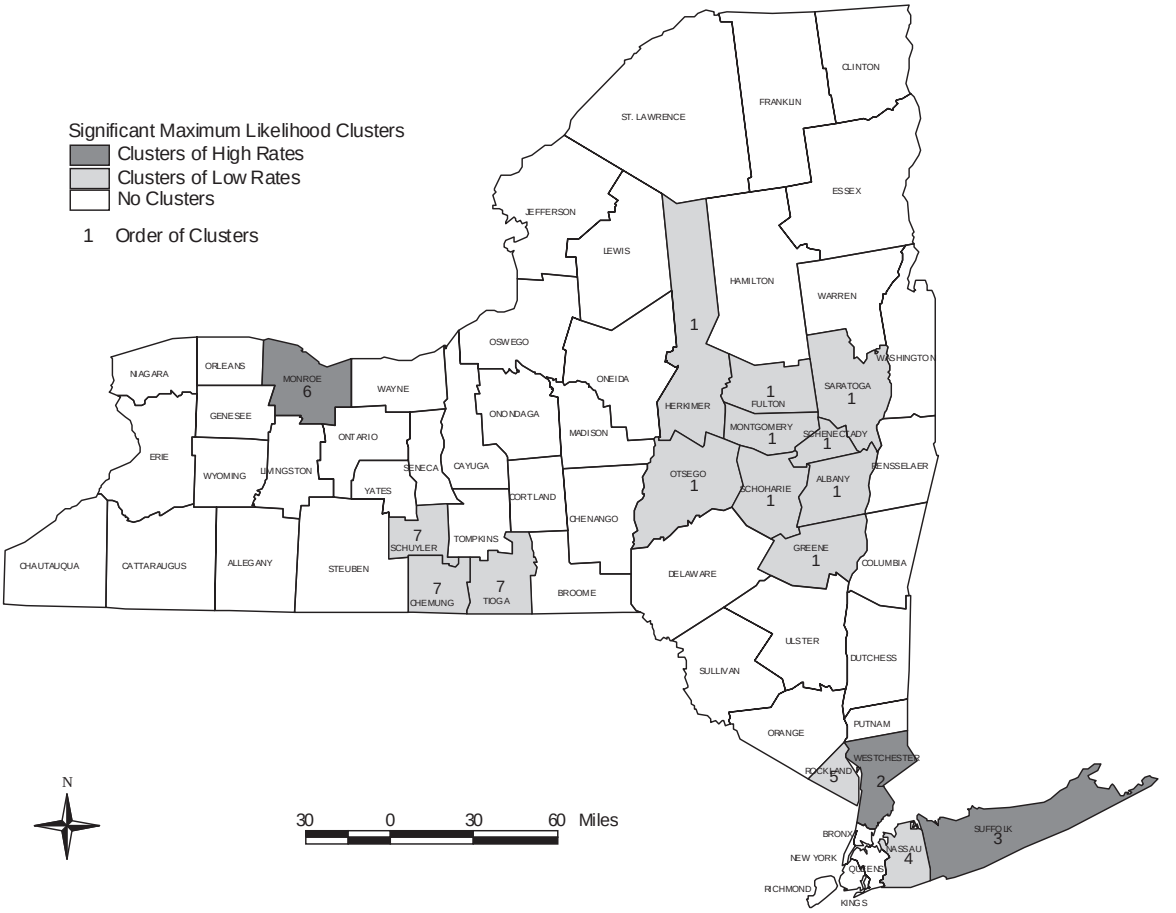


Figure 2. Significant spatial scan clusters of alcohol-explicit mortality



point features (e.g., bars). Buffer functions use the distance function on a complex feature, such as the road network, to identify points within a set distance of the feature (e.g., homes of DWI offenders within 400 yards of a bus line). These GIS functions can also be combined in complex ways to provide new insights for targeting prevention activities to areas with the greatest need (see Harding and Wittman, 1995, for additional applications in support of prevention).

Spatial Clusters

Spatial clusters are a greater than expected geographically close group of occurrences or events (e.g., deaths, crashes, alcohol outlets). Spatial clusters are a natural result of spatial dependencies in the data; by definition, spatially dependent data will have an uneven geographic distribution. The use of spatial cluster analysis was pioneered for finding cancer clusters, especially for rare cancers (Aldrich, 1990; Ricketts, Savitz, Gesler, & Osborne, 1994). Specific spatial clustering techniques can be used with point or geographic area data and may also be used for space-time cluster analysis to examine temporal trends (Jacquez, 1994). Spatial cluster analysis is useful for identifying areas with significantly high or low rates of alcohol problems where services can be targeted, to identify new research questions (e.g, why are rates highest in certain areas), to empirically identify the appropriate scale of analysis in small area studies, and to examine the impact of interventions on communities over time (e.g., do the clusters change or disappear in response to interventions).

It is important to note that low rate clusters are as important to identify for prevention purposes as high rate clusters. High rate clusters clearly have problems that require prevention/intervention; however, low rate clusters may be areas at high risk of developing problems, especially if the low rate area is embedded in surrounding high rate areas. Characteristics of low rate areas may also provide important insights into factors important for prevention application in high problem locations. Figure 2 shows a map of spatial scan clusters of alcohol-related mortality in New York. Note that both high and low rate clusters are identified. The analysis of county-level data also shows the potential for regional level prevention approaches.

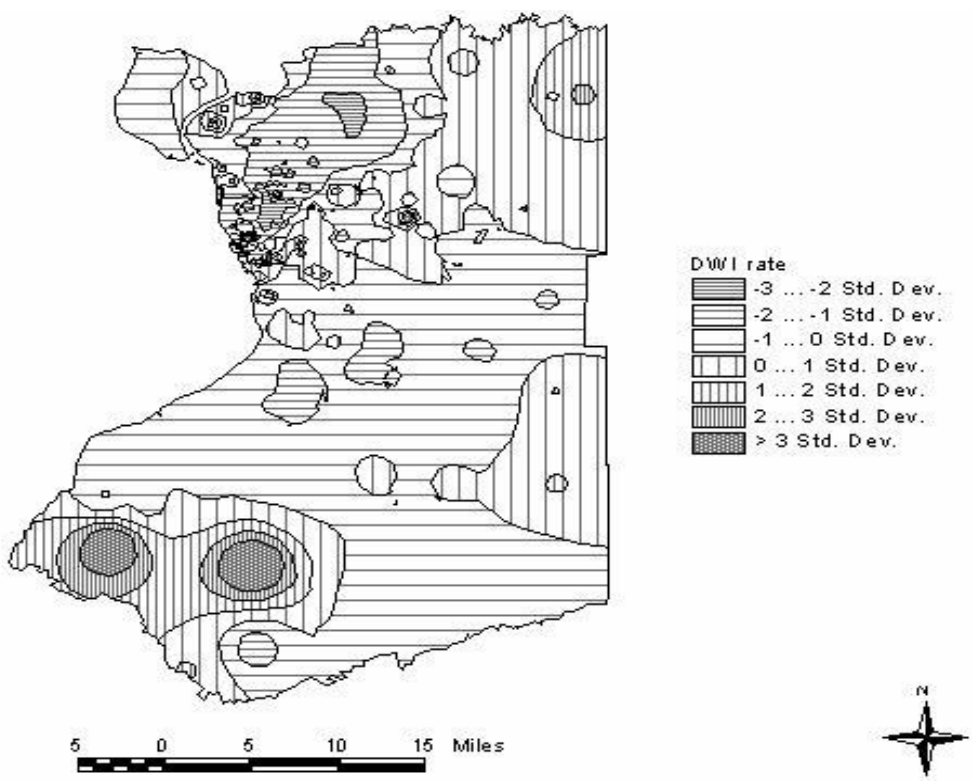
Other Spatial Analytic Techniques

Three additional approaches deserve mention in the context of spatial analysis for prevention. The first technique is a relatively simple method to control spatial autocorrelation in multiple regression analysis. Spatial autocorrelation, correlated measurement error between spatially adjacent units, is a substantial source of statistical bias in these analyses. The method is to use a GIS to calculate a generalized spatial potential for the dependent variable used in multiple regression of geographic area data (e.g., Census tracts or zip code areas). Wieczorek and Coyle (1998) provide an example of this technique in the context of targeting the neighborhoods of DWI offenders. A generalized spatial potential (GSP) for the DWI rate was calculated for each tract by summing the ratio of DWI rates (V) and distances (D) to every other tract ($GSP_i = V_1/D_1 + V_2/D_2 + \ldots V_n/D_n$). By including the GSP as an independent variable in multiple regressions, some aspects of biases due to spatial autocorrelation can be controlled (allowing more appropriate interpretations of model coefficients and statistical tests). This approach is not as statistically complete as direct methods for assessing and controlling spatial autocorrelation (see Gruenewald et al., 1996), but it is a substantial improvement that may be implemented relatively easily.

The second technique is the development of continuous surface models, or "kriging." Kriging is a modeling technique for spatial data that can be used to develop contour maps (e.g., maps that show lines of equal value such as DWI rates) from a limited number of assessment points or areas (Isaaks & Srivastava, 1989). These continuous surface models overcome a central limitation of area data: that actual rates within the geographic areas are unlikely to be as uniform as suggested in area maps. Kriging creates a continuous surface model by overlaying a grid of cells over the entire areas and calculating a weighted value for each cell based on the distance to surrounding centroids. The values calculated for the grid are then used to create a contour map. An example of kriging is provided in Wieczorek and Hanson (1997). Figure 3 shows a continuous surface model created by applying kriging to the specific tract rates used to generate Figure 1. A continuous surface model may provide a more realistic version of geographic variation that can be used to target prevention and assist in the overall planning of alcohol-related services.

The third technique allows for the transforming of zip code level data into a uniform geo-spatial Grid. Zip code information, due to its primarily administrative and political nature, is quite difficult to use for panel data analysis and public health purposes. Using irregular area units (like zip

Figure 3. DWI conviction rate continuous surface model

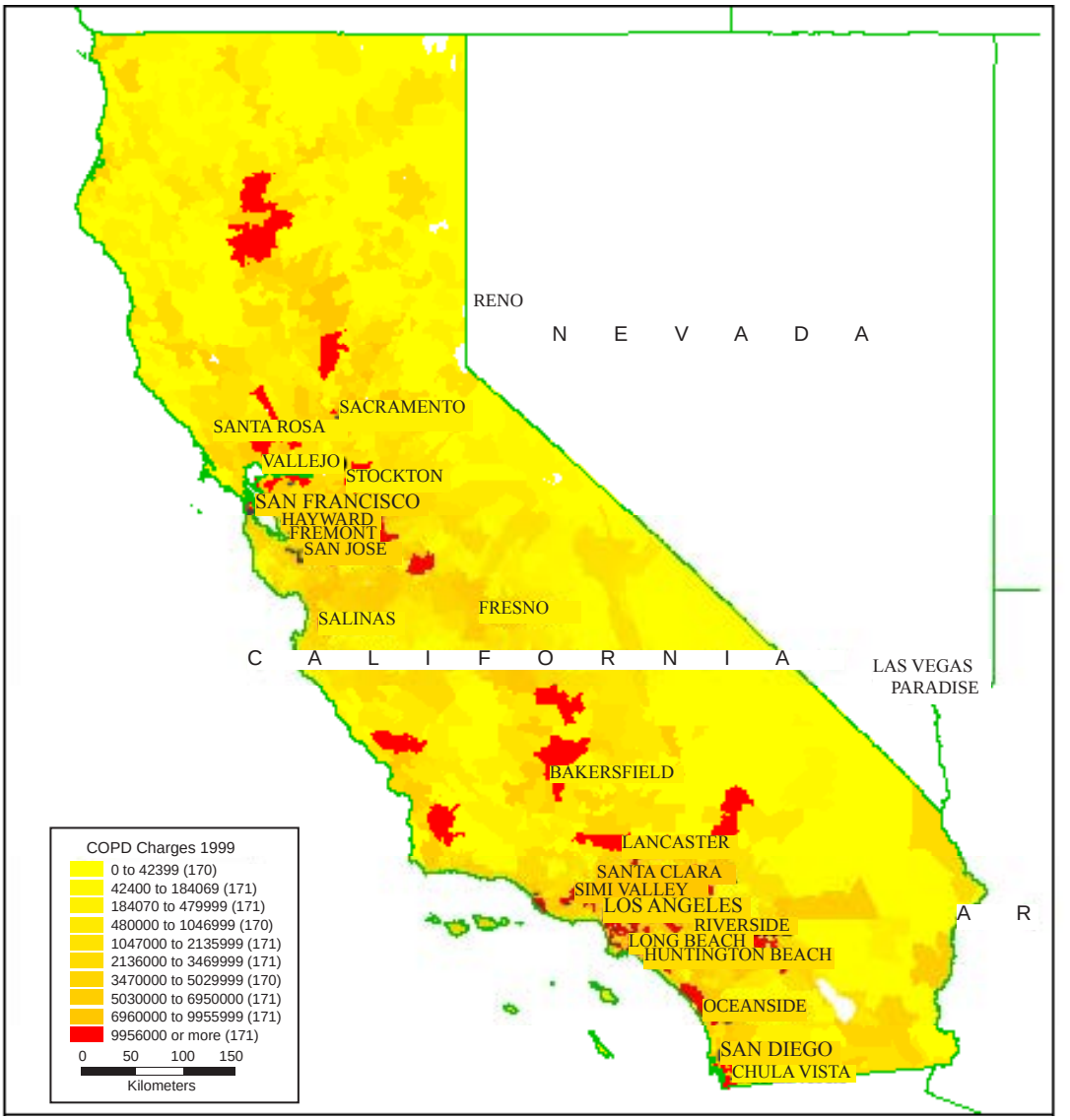


codes) for calculating disease risks poses problems of geo-statistical consistency. Changing the boundaries of collection units or grouping them differently produces different spatial patterns and gives rise to the Modifiable Areal Unit Problem or MAUP (Openshaw & Taylor, 1979). The ecological inference problem (or ecological fallacy) (Robinson, 1950), which refers to the failure to incorporate relevant, spatial information about individuals that changes the summary statistics, is a more generalized form of the MAUP.

According to Gotway and Young (2002), the MAUP and ecological fallacy are special cases of a mathematically well-defined problem known as the change of support problem (or COSP). COSP addresses the "specification bias" that can violate the properties of statistical inference and underpins the basis of probability theory (Hogg & Craig, 1995). Gotway and Young (2002) outline a combination of spatial smoothing and geo-statistical upscaling or aggregation of data with point support to avoid statistical pitfalls associated with the COSP. One way to minimize the effects of the COSP is to collect point addresses of health events so that they are not affected by scale changes. Flexible aggregation of these points with the help of a grid (as opposed to ZCTAs or census tracts) eliminates the effect of COSP. Although simple comparisons across time (panel data) are almost impossible with zip code analysis, they can be rendered in a straightforward fashion with this grid approach.

In one example examining chronic obstructive pulmonary disease (COPD) we employ a spatial overlay that applies a

Figure 4(a). COPD charges 1999 (ZCTA deciles)

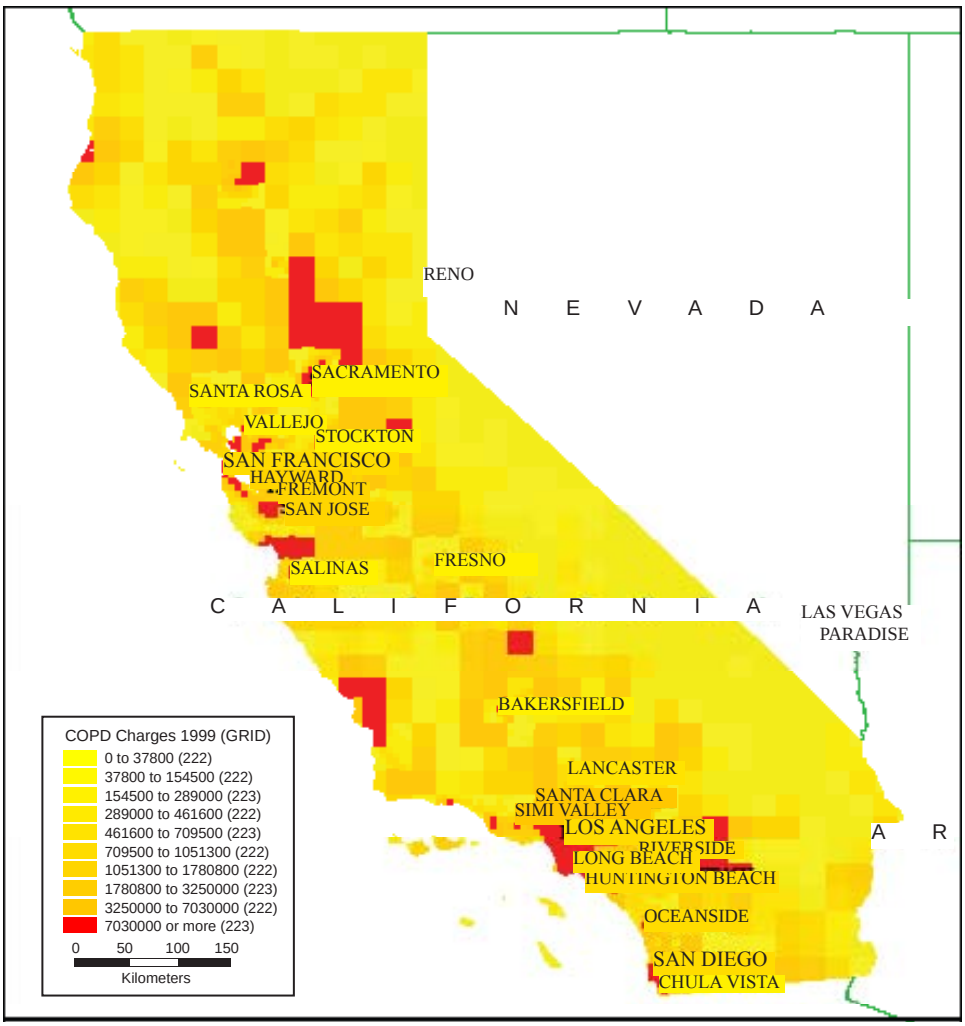


linear transformation of the zip code data to the grid, employing a "4 x 4" mile square grid for urban areas and a "16 x 16" mile grid for rural areas. This overlay procedure estimates the attributes of one or more features by superimposing them over other features and determining the extent to which there was overlap between the grid and a spatial unit—in this instance, the degree of overlap between a spatial unit and a zip code. Information for each zip code was then proportionally divided into their share of the grid by estimating the ratio of the area overlaid. Statistically, this equates to a transformation using a uniform probability density function from one area to another area of support (Goodchild & Lam, 1980). There were 1,527 zip code areas in 1993 and 1,707 zip code areas in 1999. After the spatial overlay procedure, both years had 2,224 grid units with exactly the same shape and size. In Figure 4a we show COPD hospital charges in 1999 by zip code, and in Figure 4b we show COPD charges using a uniform grid structure. While the two maps are similar in terms of showing differences across California for COPD charges, the uniform grid structure provides a much more consistent and manageable data source than the zip code map. Indeed, using a uniform grid structure for a temporal analysis eliminates differential statistical support thereby minimizing COSP (Gotway & Young, 2002). A possible disadvantage associated with this procedure is that some information will be lost when converting zip code areas into grid areas; however, the stability of the new units over time more than compensates for this by improving statistical support and minimizing statistical misspecification.

SPATIAL DECISION SUPPORT SYSTEMS (SDSS)

GIS can help us link traditional data analysis with good geo-spatial data. Traditional data analysis that deals with optimization, as in graph theory or network theory, can be

Figure 4(b). COPD charges 1999 (4&16 mile grid)



implemented in a geo-spatial context. For example, locating a new hospital where a certain outcome (like distance to patient population) can be minimized can be very helpful from a public health perspective.

Some of the applications of optimization (TransCAD, 2005) in a spatial and public health context are as follows:

- **Vehicle Routing:** Describes procedures used to provide efficient routes for making ambulance pick-up and deliveries, including restrictions on times at which stops are made and multiple vehicle dispatches during an emergency.
- **Partitioning:** Describes how districts (patient districts) can be created that optimize the hospital loads based on numbers served and minimized travel times for patients.
- **Facility Location Models:** Describes models that find efficient location for facilities (like new health centers in addition to old ones) that can optimize some objective (like improving level of service, cost of service, etc.).

These types of sophisticated analyses, can be viewed as a move (for spatial analysis) from mere data modeling to applying techniques of operations research—a highly developed field in applied and theoretical mathematics.

SPATIAL ANALYSIS: AN EXAMPLE FROM VIOLENCE PREVENTION

Knowing where problem events take place is different from knowing why they take place where they do. And understanding the etiology of public health problems in different geographic areas requires knowing all the spatial technologies discussed to this point and their suitable application in spatial statistical analysis, the use of spatial data for the purpose of explicating the etiological dynamics of public health problems. This section will provide an example of one such application, a first approach to understanding the environmental correlates of violence.

Much of the criminology and public health literature is concerned with the determinants of violence in different community areas. A particular focus has been upon the role of alcohol outlets in violence. Similar to the work of Morenoff, Sampson, and Raudenbush (2001) and Baller, Anselin, Messner, Deane, and Hawkins (2001), our analysis includes demographic and socio-economic data so as to capture violence related to population characteristics (i.e., high unemployment, low rates of high school graduates, etc.). These population characteristics are analyzed in relation to the moderating effects of alcohol outlets on the production of violence. Moderation can simply be thought of as interaction of outlets with people (with a mix of characteristics). This outlet interaction could serve to increase or decrease violence, depending on the composition of population characteristics and outlet presence and type. Further, we examine spatial components of these moderating effects. Rates of violence may be affected by characteristics of populations living in adjacent areas (Gorman, Speer, Gruenewald, & Labouvie, 2001).

Our analysis assesses whether such spatial relationships exist and controls for spatial autocorrelations that may obscure the relationship between population characteristics and the production of violence. The sample comprised 766 zip codes from four selected areas of California: Los Angeles, the Bay Area, Sacramento, and the northern section of the state. The first three areas are heavily urbanized, while the last is quite rural. The three urban areas are heterogeneous with regard to ethnic, age, and socio-economic composition, particularly in relation to the rural area that is more homogeneous in most population level measures. In this research, data are taken from three different sources: Census data (1990), using a three-item scale representing concentrated disadvantage, immigrant concentration, and residential stability, based on the work of Sampson, Raudenbush, and Earls (1997). Hospital discharge data for 1991, using patient home address, contained information on assaults. California state data on alcohol outlets (1991) gave type and address of outlet.

The graphic presented in Figure 5 outlines the general conceptual framework that guides the analysis. We hypothesize that populations of people produce assaults at a given rate that is moderated by population characteristics (e.g., greater in places where there are more young people) and environmental characteristics (e.g., where there are more alcohol outlets). This adjusted rate may be further modified by the numbers and characteristics of nearby populations (not shown). It is assumed that errors in estimation are not independent, but rather are spatially autocorrelated, r_s.

In Figure 6 assaults per roadway mile are presented for each of the four regions. The difference in concentrations of assaults is apparent in this map, with greater densities

Figure 5. Conceptual model and statistics for actual spatial regression

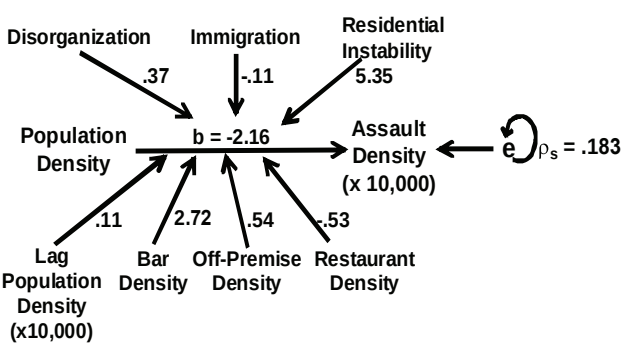
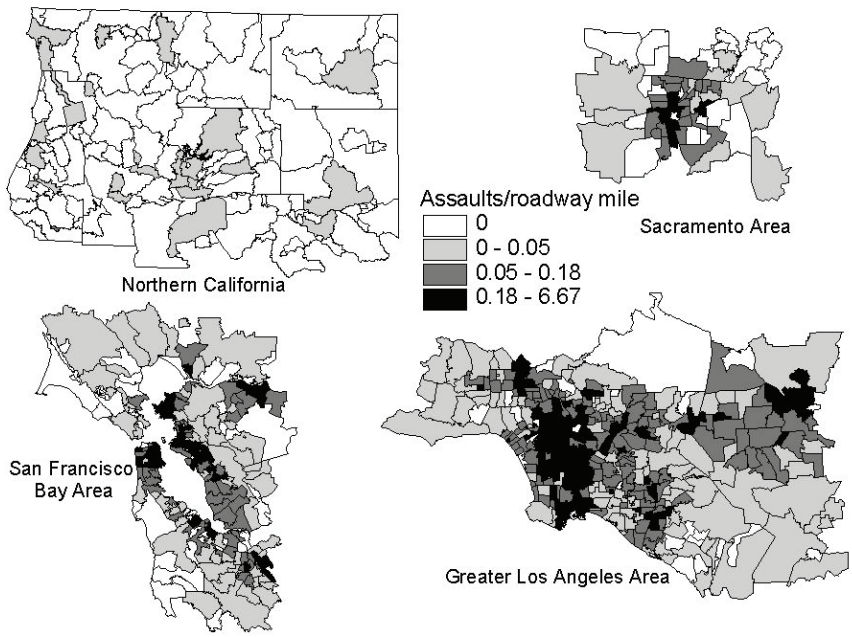


Figure 6. Assaults per mile of roadway



occurring where there are greater densities of population. This is not, however, universally the case. For example, the western region of the Los Angeles basin appears to exhibit relatively greater assaults than expected from the population distribution observed. Overlays of zip codes are represented on this particular map.

Table 1 and Figure 5 (note numbers in the figure) present the results of an analysis that includes the direct effects of population variables, alcohol outlets, and adjacent population density. Many, but not all, of the estimated coefficients from the model are significant. A direct interpretation of the coefficients of the model suggests that as population density increases, there is a reduction of approximately 2.16 assaults for every 10,000 persons in each zip code area. This rate, however, is that expected for an isolated population living in an area with no bars, restaurants, or off-premise alcohol establishments (and rather unrealistically, with values of zero for the measures of population characteristics). This rate is moderated, however, by the presence of large populations in adjacent areas, the densities of restaurants, bars, and off-premise establishments, and local population characteristics. This rate is greater in areas where bar densities are greater and restaurant densities are less. This rate is further greater in areas where social disadvantage is greater, immigrant presence less, and residential stability greater. Finally, the rate at which local population density produces assaults is greater in areas surrounded by larger populations.

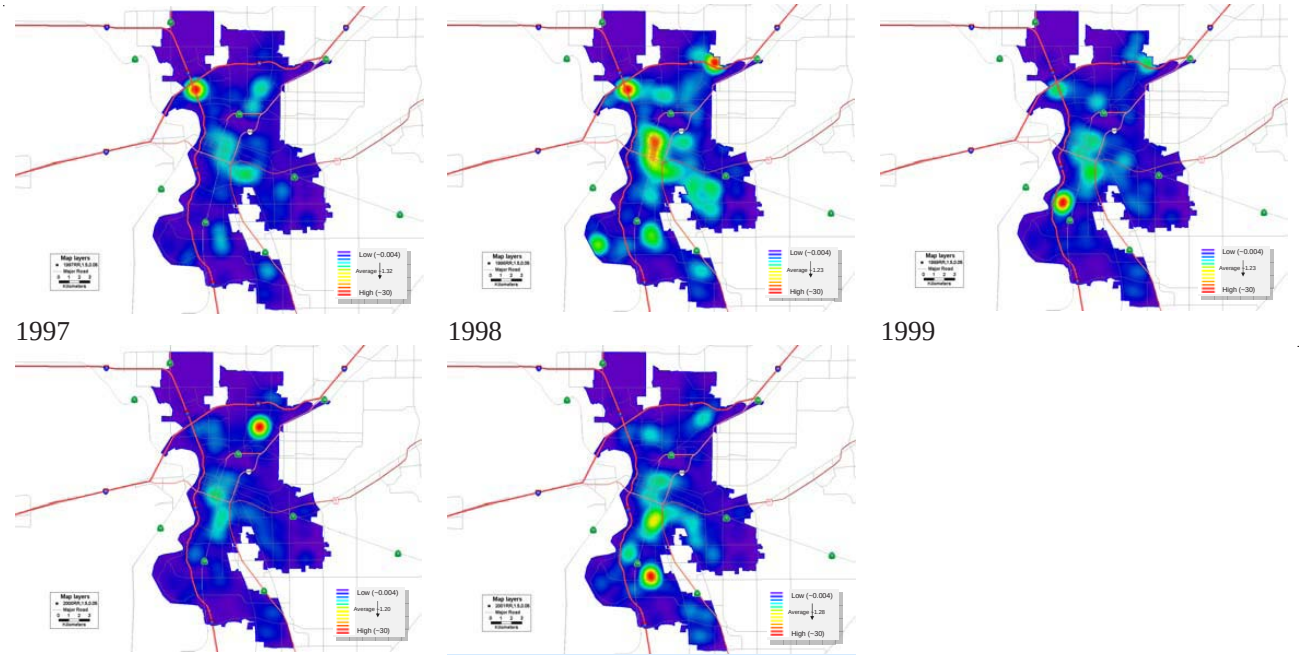
Table 1. Associations of outlet densities with rates of assault hospitalizations (x 10,000)

Variable Name	b:	t:	p:*
Population Density	-2.16	02.62	.009
Outlet Densities			
Bars	2.72	4.36	<.001
Off-Premise	.54	1.64	.101
Restaurants	-.53	-3.97	<.001
Population Characteristics			
Social Disadvantage	.37	32.74	<.001
Immigrant presence	-.11	-15.17	<.001
Resident Stability	5.35	2.41	.016
Adjacent (Lag) Population Density (x 10,000)	.11	2.21	.027
Model based estimate for spatial autocorrelation: $\rho_{s=}$ .183 ** Z = 3.51 P < .001			

**two-tailed significance*

*** a measure of spatial autocorrelation for the model*

Figure 7. Predicted changes in relative rates of assault using a Bayesian hierarchal modeling framework



The results of this analysis indicate that it is possible to construct a conceptually well-framed spatial analysis of assault rates that explains to a substantial degree variation in rates of assault between places in California. Shifting from a representation that suggests that outlets on their own create violence to one that presents outlets as providing contexts for violence, the usual pattern of relationships of violence to environmental densities of alcohol outlets continues to be observed. The current analysis suggests, however, that alcohol outlets moderate rates at which violence is produced within areas and that these effects persist when controlling for spatial effects and other covariates related to the production of violence in local populations (Table 1). Notably, positive relationships continue to exist between bar and off-premise outlet densities and assaults, with no relationship to densities found for restaurants (Table 1 and Figure 5). At the geographic scale of the current study (zip codes), greater rates of violence are observed in stable, non-immigrant areas with greater concentrated disadvantage. Bar densities, when controlling for other environmental or socio-demographic measures, are clearly connected to an increase in assaults. Thus, in this spatial example, beyond the obvious finding that denser populations have more assaults, we are able to observe important environmental effects that may be actionable in terms of prevention policy.

BAYESIAN MODELS

Before the advent of the Markov Chain Monte Carlo (MCMC) revolution in 1990 a small group of researchers, consisting of mathematical statisticians and theoretical probabilists, were involved with hierarchical Bayesian methods (Banerjee, Carlin, & Gelfand, 2004). MCMC made Bayesian methods popular due to its ability to translate the complex combination of theoretical random variables and random coefficients into numerical outputs for real world problems. Before MCMC, hierarchical Bayesian models were simple to construct but difficult to implement due to the lack of computing power to integrate hundreds and thousands of unknown parameters. Hierarchical Bayesian methods are suitable for spatial statistical applications. Modeling coefficients as random effects allows us to induce a specific correlation structure for the coefficients. Thus, a specified correlation structure such as spatial autocorrelation can be modeled within any multivariate study. For example, the relationship between violence and alcohol problems can be modeled at the local level using hierarchical Bayesian modeling. Gruenewald and Banerjee (2005) hypothesized that:

- **Hypothesis #1:** Drug activity will generally be related to levels of violence across neighborhood areas ("systemic" violence).

- **Hypothesis #2:** Drug activity will lead to explosive growth in related violence in areas where "core groups" form (core group subcultures).
- **Hypothesis #3:** Built environments in public areas (e.g., alcohol outlets) will tend to suppress violence associated with drug activities (surveillance, guardianship).

Analysis features:

- **City of Sacramento: Population 403,796 (1997) Years 1997-2001, 304 Census Block Groups (n * t = 1,520)**
- **Dependent Variable:** Number of Emergency Medical System (EMS) Assaults. At-site treatment for physical assaults by EMS.
- **Independent Variables:** Population size, percent population in poverty, number of Hispanics, number of African Americans, drug activities (police calls for service), number of bars, number of off-premise establishments.
- **Analysis Model:** Poisson space-time model with unit and time specific random effects and statistical controls for spatial heterogeneity (spatial autocorrelation) using Conditional Autoregressive Models implemented in a Bayesian framework. Asymmetric bootstrapped standard errors (Bernardinelli et al., 1995).

In Figure 7, we can see that when modeling the relative risk of assaults in Sacramento using Bayesian methods, over time, the loci of assaults shift. Such a result is only available using conditional hierarchal multivariate models. Although theoretically possible using traditional techniques, in reality, only Bayesian methods are currently practical. Further, the results described are in accord with the previous hypotheses, which provides a coherent dynamic spatial explanation for the production of violence.

FUTURE TRENDS

Spatial analysis in public health is in its infancy. This is due to the rapid increase in computing power that has only in the last few years allowed for a more comprehensive development of mapping and spatial statistical techniques. There are two general areas of development that are interrelated: (1) GIS will become more intuitive and popular with more data becoming accessible (e.g., through geo-coding techniques that will become more consistent and "industry standard" through time), and (2) spatial statistical methods will become more available and more user friendly with statistical software companies offering spatial statistical solutions. As the comfort level grows with both mapping and spatial statistics, there will be an increasing demand for data and software that more capably handle research needs. This trend will only accelerate. The ability to add a spatial component to typical population level data opens up an entirely new approach toward thinking about public health issues. The individual will no longer be considered the sole unit of analysis; place will start to constitute its own unit of analysis or as complexly interacting with the individual. Further, as spatial methods improve we will be able to model the dynamic relationship between a target area and areas around the target. The previous example is one of the first public health examples of such an approach. In the future this dynamic modeling will become much more routine and essential to characterizing people in places. In addition, with the introduction of Bayesian hierarchal spatial approaches, we have a readily available ability to see how such dynamic spatial systems change over time. There is no methodological or theoretical obstacle to being able to, for example, analyze how characteristics in one area, such as density of liquor stores, affects violence in an adjacent area and to see how this relationship changes through time.

SUMMARY AND CONCLUSIONS

In this article we have described how spatial methods may be applied in specific public health areas, namely alcohol and alcohol related problems and COPD. Although important areas of research, they are by no means unique in being amenable to spatial/ecological analysis. There are several important factors that should be considered when contemplating public health spatial analysis: (1) Are there specific environmental features (such as the presence of bars or liquor stores) that might help explain an outcome (such as violence)? (2) Is there a dynamic relationship between individual behavior and environmental (area) setting? (3) Do environmental factors, such as alcohol outlets, modify the relationship between socio-demographic factors and the outcome of interest (e.g., violence)? (4) Is there data available to support a spatial analysis? (5) Is the effect of adjacent areas likely to obscure relationships between exposures and outcomes (spatial autocorrelation)?

When studying alcohol related problems, spatial analysis allows for the integration of disparate types of information into a meaningful story from both a public health and criminological/sociological point of view. Further, being able to use a uniform grid structure as seen in the earlier COPD example, notwithstanding certain strong assumptions, is a method that can be used to overcome limitations of using typical administrative geographical units such as zip codes. Given that zip codes are perhaps the most common basic areal unit of analysis, novel approaches toward dealing with such data are necessary. Similarly, the Bayesian spatial method described allows for a more powerful and

efficient ability to handle spatial statistical models compared to other methods.

The ability to put people in places in more than a purely descriptive framework signals a new generation in research that transcends traditional proscriptions against the use of ecological data. Further, measures of community health such as social cohesion take on a more fully realized form in a spatial analytical context. Indeed, given that most public health and criminological data are collected at a population level, spatial analysis allows researchers to more clearly observe population level effects for whatever measures chosen. Ultimately, GIS and spatial methods will become an integral part of public health research and practice as will the concept of place. In the future, measures of place will be as essential to public health research as race and class are now.

REFERENCES

Aldrich, T. E. (1990). *CLUSTER: User's manual for software to assist with investigations of rare health events*. Atlanta, GA: Agency for Toxic Substances and Disease Registries.

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical Analysis, 27*(2), 93-115.

Baller, R. D., Anselin, L., Messner, S. F., Deane, G., & Hawkins, D. F. (2001). Structural covariates of U.S. county homicide rates: Incorporating spatial effects. *Criminology, 39*(3), 561-590.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: Chapman & Hall/CRC.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995). Bayesian-analysis of space-time variation in disease risk. *Statistics Medicine, 14*(21-22), 2433-2443.

Goodchild, M. F., & Lam, N. S. N. (1980). Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing, 1*(3), 297-312.

Gorman, D. M., Speer, P. W., Gruenewald, P. J., & Labouvie, E. W. (2001). Spatial dynamics of alcohol availability, neighborhood structure and violent crime. *Journal of Studies on Alcohol, 6*(5), 628-636.

Gotway, C. A., & Young L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association, 97*(*458*), 632-648.

Gruenewald, P. J., & Banerjee, A. (2006). The structure of the built environment and space-time models of drugs and violence. In *NIDA/Association of American Geographers Symposium on Geography and Drug Addiction*. Chicago.

Gruenewald, P. J., Millar, A., Treno, A. J., Ponicki, W. R., Yang, Z., & Roeper, P. (1996). The geography of availability and driving after drinking. *Addiction, 91*(7), 967-983.

Harding, J. R., & Wittman, F. D. (1995). GIS enhances alcohol/drug prevention planning. *GIS World, 8*(6), 80-83.

Hogg, R. V., & Craig, A. L. (1995). *Introduction to mathematical statistics*. Englewood Cliffs, NJ: Prentice-Hall.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics*. New York: Oxford University Press.

Jacquez, G. M. (1994). *Stat! Statistical software for the clustering of health events*. Ann Arbor, MI: BioMedware.

Lipton, R. I., & Gruenewald, P. J. (2002). The spatial dynamics of violence and alcohol outlets. *Journal of the Study of Alcohol, 63*(2), 187-195.

Morenoff, J. D., Sampson, R. J., & Raudenbush, S. W. (2001). Neighborhood inequality collective efficacy, and the spatial dynamics of urban violence. *Criminology, 39*(3), 517-560.

Openshaw, S., & Taylor, P. J. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In N. Wrigley (Ed.), *Statistical applications in the spatial sciences* (pp. 127-144). London: Pion.

Ricketts, T. C., Savitz, L. A., Gesler, W. M., & Osborne, D. N., (Eds.). (1994). *Geographic methods for health services research*. New York: University Press of America.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(3), 351-357.

Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science, 277*(5328), 918-924.

TransCAD™ (2005). *Caliper Corporation*. Newton, MA. Retrieved from http://www.caliper.com/TransCAD/ApplicationModules.htm#Network%20Analysis

Wieczorek, W. F. (2000). Using geographic information systems for small area analysis. In R. E. Wilson & M. C. Dufour (Eds.), *The epidemiology of alcohol problems in small geographic areas* (pp.137-162). Bethesda, MD: NIH National Institute on Alcohol Abuse and Alcoholism.

Wieczorek, W. F., & Coyle, J. J. (1998). Targeting DWI prevention. *Journal of Prevention and Intervention in the Community, 17*(1), 15-30.

Wieczorek, W. F., & Hanson, C. E. (1997). New modeling methods: Geographic information systems and spatial analysis. *Alcohol Health and Research World, 21*(4), 331-339.

Wieczorek, W. F., & Hanson, C. E. (2000). Regional patterns of alcohol-specific mortality in the United States. In R. C. Williams, M. M. Howie, C. V. Lee, & W. D. Henriques (Eds.), *Geographic information systems in public health: Proceedings of the third national conference* (pp. 669-676). Atlanta, GA: Centers for Disease Control and Prevention.

KEY TERMS

Bayesian Spatial Analysis: Spatial probabilities are often conditional and hierarchical. Bayes's theorem provides a convenient way to compute such conditional probabilities. Because the computational impediments to computing posterior probabilities in a Bayesian setting are eliminated after the advent of Markov Chain Monte Carlo algorithms, this method is an efficient and flexible method for conducting spatial analysis.

Geocoding: A generic term used to describe the GIS function of providing a specific location to descriptive data. Geocoding applies to point data (e.g., alcohol outlet) as well as to areal data (e.g., assaults in a census tract).

Geographical Information Systems (GIS): The geographic uses of data to develop maps and statistical relationships that help describe processes like the relationship between alcohol outlets and violence or vehicle crashes and alcohol outlets.

GIS Distance and Adjacency Function: The distance between individual objects (e.g., bars and crashes) and whether areas are adjacent to one another.

Kriging: A technique that can be used to develop contour maps (e.g., maps that show lines of equal value such as DWI rates) from a limited number of points or areas (which can be given a value at the centroid).

Overlay Function: The ability to combine multiple layers of information.

Spatial Analysis: Using geographic data to mathematically model the relationship between measures such those mentioned previously, that is, alcohol outlets and violence.

Spatial Autocorrelation: The measure of similarity between values (for a given variable, for example, income) located in space. Similarity of values in spatial proximity may indicate some underlying mechanism that is spatial in nature and contributes to the spatial pattern of the predictor variable. Controlling for spatial autocorrelation reduces statistical bias in parametric modeling.

Spatial Clusters: A greater than expected geographically close group of occurrences or events (e.g., deaths, crashes, or alcohol outlets).

Spatial Decision Support Systems (SDSS): Decision algorithms and software developed for particular types of decision support that deal with geographic space. Developed with GIS technology, but without requiring programming skills or knowledge of GIS software, SDSS uses digital maps, tables, and charts that are intuitive and help reduce the amount of information processing needed to make complex decisions.

Uniform Grid Unit Transformation: A method for taking information from arbitrary administrative spatial units like zip codes and transforming the data into a uniform grid structure.

Geospatial Information Systems and Enterprise Collaboration

Donald R. Morris-Jones
SRA, USA

Dedric A. Carter
Nova Southeastern University, USA

INTRODUCTION

Organizations and teams are becoming increasingly more distributed as groups work to expand their global presence while rationalizing team members across skill sets and areas of expertise instead of geographies. With this expansion comes the need for a robust and comprehensive language for pinpointing locations of globally distributed information systems and knowledge workers. Geospatial information systems (GISs) provide a common framework for jointly visualizing the world. This shared understanding of the world provides a powerful mechanism for collaborative dialogue in describing an environment, its assets, and procedures. The collaborative framework that GIS provides can help facilitate productive dialogue while constraining impulses of extreme positions. Collaboration and GIS intersections take many forms. Under a collaborative work-flow model, individuals use GIS to perform their job and post data back to the central database (e.g., engineering designs and as-built construction).

This article addresses the increasing role of GIS in emerging architectures and information systems in a number of applications (e.g., land planning, military command and control, homeland security, utility-facilities management, etc.). Real-time applications, mobile access to data, GPS (global positioning satellite) tracking of assets, and other recent developments all play a role in extending the scope and utility of the GIS-enabled enterprise. The impact of new GIS Web services standards and open geospatial-data archives are also addressed as areas of increased potential for remote GIS collaboration in global organizations. The expansion of enterprise GIS within organizations increases the opportunity and necessity of using GIS collaboratively to improve business processes and efficiency, make better decisions, respond more quickly to customers and events, and so forth.

BACKGROUND

The term *geospatial* is increasingly used to describe digital data about the earth in GIS, image, or GPS formats. The related technologies of GIS, remote sensing image-processing systems, and GPS data collection are all components of geospatial information systems. Geospatial-technologies use continues to expand in a great variety of applications ranging from land planning to utility-engineering design and military command and control. Those applications, which were once relegated to discrete groups of specialists, have now begun to take a more prominent role in the enterprise. Duffy (2002) describes the transition of GIS from a specialist technology to a more mainstream environment in the industry information-systems department from the end of the last decade into 2002.

The essence of collaboration is people and organizations working together to accomplish a common goal. Information–technology- (IT) enabled collaboration has improved business processes in many organizations and contributed to more functional and profitable operations. Collaboration technologies are characterized by three major generic attributes: communication, information sharing, and coordination (Munkvold, 2003). These characteristics can be further refined into available channels such as synchronous or asynchronous, the medium of sharing information through repositories or real-time interaction, and work-flow management to coordinate steps in a decision process or protocol. Geospatial technologies and systems extend collaboration in unique ways for problems that are related to location.

GIS provides a geographic dimension to enterprise collaboration, which helps solve a variety of problems that are difficult to address by any other means. For example, vehicle-routing and dispatching applications make it possible for Sears to deliver goods to customers more efficiently within tighter time windows. As a result, Sears is more profitable and customers are more satisfied. This example of distributed-network optimization using efficient queuing mechanisms based on location information is a simple illustration of the impact that GIS data may have on existing

business processes. In fact, most aspects of business-process automation initiatives at present require some element of collaboration either between networked systems or dispersed individuals.

Collaboration utilizing GIS and geospatial frameworks continues to be a focus of research both in the United States and abroad (Boettcher, 2000; Songnian, 2004; Stasik, 2000).

EMERGING GEOSPATIAL INFORMATION-SYSTEMS ARCHITECTURES AND COLLABORATIVE ENTERPRISE APPLICATIONS

As organizations become more dispersed in an effort to rationalize across areas of expertise in lieu of geographies, complex infrastructures for location analysis and coordination may emerge (Munkvold, 2003). In recent years, GIS software companies have developed an expanding and increasingly capable enterprise suite of tools. Early generations of GIS were used by GIS specialists only; these systems were available in stand-alone or project-systems configurations. GIS product options have improved and now provide a sound basis for supporting casual users as well as specialists with desktop, distributed client-server, and Internet solutions.

Geospatial data standards and interoperability have greatly improved the ease of using data in different formats or geographic projections. Geospatial Web-services standards provide Internet access to geospatial data stored in geospatial-data archives. Federal-government initiatives (e.g., Geospatial One Stop, the National Map, Homeland Infrastructure Foundation Level Database, etc.) will increase data standardization and access, and reduce expensive, redundant data collection.

GIS-enabled collaboration can now involve a broad range of different types of users within and outside of a particular organization. These users can be expert or casual as well as stationary or mobile. Medeiros, de Souza, Strauch, and Pinto (2001) present an analysis of aspects of coordination in a collaborative system for spatial group-decision support that resulted in a prototype system for a distributed GIS.

Geospatial Information-Systems Products and Architectures

GIS-product vendors continue to innovate and expand the solution set available to the user (Atkinson & Martin, 2000).

Enterprise suites of GIS include the following different types of products.

- Desktop GIS with varying levels of functionality
- Spatial analysis extensions
- Internet GIS with limited functionality or full functionality
- Mobile GIS
- Geospatial-data middleware
- Software to embed geospatial functionality in business applications
- 3D GIS
- Geospatial-data visualization software
- GPS tracking software
- Remote sensing image-processing software
- Geospatial Web-services software
- Location services

GIS products are available to support stand-alone users, and distributed client-server and centralized Internet architectures. GPS tracking units and mobile GIS on Personal Digital Assistants (PDAs) and pocket PCs extend the range of the technology into the field. Wireless communication of data is improved through the use of data compression and area-of-interest extraction techniques.

Geospatial-data management functionality is improving but is less capable than business-data management functionality. While Oracle states that their products now provide equivalent data-management functionality for spatial and business data, experience is limited for enterprise replication of geospatial data. ESRI, the GIS-software market leader, promises to add geospatial-data replication to its ArcSDE product with the release of ArcSDE 9.1, which is projected for 2005. The large size of geospatial-data files means that substantial bandwidth is needed to move data through a communications network.

Location services refer to mobile geospatial services that will primarily be delivered to location-aware smart phones. The E-911 legislation mandates that cell phones must become location aware so emergency vehicles can locate 911 callers who use cell phones. Cell-phone operators and partners are and will offer an increasing array of location services to provide users with directions for driving, the nearest services of different types, and the location of buddies (i.e., those who have authorized sharing this information, etc.). An example system is presented in the location-based tourist-guide application of Simcock, Hillenbrand, and Thomas (2003). This tool combines a mobile PDA device and GPS technology to provide the user with location tours that are self-guided.

Collaborative Land-Use Planning

Geospatial data provides a common view or abstraction of an area. GIS has been used extensively in land-use planning. While GIS initially served and continues to serve as a tool for planners, it is increasingly used to facilitate collaboration with the public. The ability to show land and its characteristics to groups of people with divergent views provides a common

frame of reference, which can make it easier to develop consensus or agreement on difficult issues. The common GIS data framework serves to constrain more extreme positions, which are more likely to be presented without geospatial data. The implementation of the collaborative system in Medeiros et al. (2001) described earlier was targeted as a land-use and planning application for a distributed GIS.

The Urban and Regional Information Systems Association (URISA) is in the third year of hosting an annual conference on the topic of public participation in GIS (URISA, 2004b; Voss et. al., in press). A discussion of presentations at the second conference in the series mentioned that Internet GIS technology was used to present and solicit public comment regarding alternative proposed designs for the World Trade Center buildings and parks (URISA, 2004a). From the number of questions and issues that were identified in the conference, it is clear that there is not yet a commonly accepted model for how GIS should be used to facilitate public participation in planning. As the role and power of planners, politicians, GIS specialists, developers, and the public could potentially shift with more publicly available geospatial data and analytic capabilities, there is no simple answer to the question of how collaboration between government and the public should be enhanced using GIS. Clearly, the technology increases options for public participation in land planning.

The Orton Family Foundation, a Vermont-based nonprofit organization dedicated to better decision making by communities, has developed 2D and 3D GIS software to help people develop, visualize, and analyze the implications of alternative approaches to land- and growth-management planning (Orton Family Foundation, 2004). Their CommunityViz GIS software has been used extensively by smaller communities to help involve the public in land planning. In Eureka Township, Minnesota, GIS was used to help the Eureka Township Envisioning Project understand the impacts of alternative growth-management scenarios on septic-system placement and water quality (Orton Family Foundation).

Utilities-Engineering Design and Construction

Utilities use geospatial-information systems to plan and manage their assets and design new components of their outside plant infrastructure (e.g., poles, wires, transformers, pipes, etc.). While most of the collaboration occurs within the utility, collaboration occurs to some degree with developers and government officials. For example, a developer will submit their design for a new subdivision, usually in CAD (computer-aided design) format. Field surveys are often conducted at the site and these surveys are increasingly conducted using GPS technology. The utility designer will import this CAD into either a CAD or GIS tool, which will be used to design the utility infrastructure that is needed to support the proposed development. Typically, the proposed design is analyzed as if it were connected to the network, so a proposed version of the design is developed. Alternative designs may be developed and costs of the alternatives may be determined so that the most cost-effective option can be chosen. Designs may be reviewed by other designers or a design supervisor to select the preferred option, which may also be with the developer and government officials. When approved for construction, the design will be attached to a work order and assigned to a crew which will be dispatched to the construction site. The design may be modified in the field if logistical difficulties are encountered. Construction crews will record as-built conditions for the project, and this data will be reviewed and posted to the database, which represents the utility assets that are constructed.

This utility design and construction process occurs over an extended period of time, which makes it a long transaction. Engineers will typically extract a version of the database to use to develop their designs. Usually, a utility will use optimistic logic and not formally lock the portion of the database that is included in the versioned database. Conflicts between alternative designs in the same general area are flagged for resolution by an engineering supervisor.

This CAD- and GIS-based process represents a major increase in productivity in design and maintenance of utility maps and records. Utility GIS practitioners often network through an organization called the Geospatial Information & Technology Association (GITA, 2004).

C4ISR and Defense Geospatial Applications

Command, control, communications, computers, intelligence, surveillance, and reconnaissance (C4ISR) is a military application that makes extensive use of geospatial technologies. The National Geospatial-Intelligence Agency, the National Reconnaissance Office, and other agencies collect imagery and produce different types of geospatial data for military and intelligence purposes. For example, cruise missiles use topographic data as a model of the terrain. The terrain is sensed by the cruise missile and pattern-matching techniques are used to follow the landscape to reach a target. High-resolution imagery is interpreted to help select targets.

C4ISR provides the military forces with geospatially enabled command and control capabilities. Trends are to make C4ISR more widely available to help cut through the fog of war and provide a better coordinated understanding of the battle space. The Geo-Intel 2003 conference included discussions of the importance of geospatial data and technologies and the role played by them in the rapid and successful movement of U.S. troops in the Iraqi War.

In the future, geospatial intelligence and C4ISR will be more ubiquitous. Currently, the military is still in the process of adopting geospatial data for battle-space operations. The Commercial Joint Mapping Toolkit initiative could place GIS-enabled C4ISR in the hands of up to one million U.S. military forces.

Homeland-Security Command Centers

Homeland–security command centers are making effective use of GIS to provide planning and response capabilities. The emphasis on situational awareness and command and control is similar to DOD C4ISR applications. 3D GIS applications developed by IT Spatial are playing a major role in command centers in Washington, DC, and elsewhere. Plume modeling software combined with demographic data, 3D GIS data, and real-time weather data provide a capability to predict terrorist acts, like a dirty nuclear explosion, and their consequences. There is no shortage of investigation of 3D GIS technologies (Manoharan, Taylor, & Gardiner, 2002). The better the consequences are understood, the more effective the response can be from emergency first-responder forces. First responders in the field can access data and also feed back information regarding emergencies and events to provide command and control personnel and other responders with better understanding. This impact of geospatial information in emergency situations has been the subject of research by several in the wake of the attacks of September 11, 2001 (Kevany, 2003; Kwan & Lee, in press; Rauschert, Agrawal, Sharma, Fuhrmann, Brewer, & MacEachren, 2002), in addition to some early research examining this area prior to the attacks (Kumar, Bugacov, Coutinho, & Neches, 1999). GPS tracking can be used to understand the location of police cars, ambulances, and other emergency vehicles. Integration of GIS with live video data feeds also provides a useful real-time picture of ground conditions and events. There are some, however, that continue to point out the numerous impediments to effectively using GIS in disaster-control decision situations (Zerger & Smith, 2003).

GIS additionally continues to play an important role in the containment of medical outbreaks and the analysis of systems mapped to locations, such as in the work of Cockings, Dunn, Bhopal, and Walker (2004).

Presence Awareness in Instant Messaging and RFID-Based Asset Tracking

In an effort to better display the impact of emerging technologies such as radio-frequency identification (RFID), Web services on GIS, and collaboration for real-time presence management, AMS, through its Center for Advanced Technologies, customized the BuddySpace instant messaging (IM) product to allow IM users the ability to display the location of other users. This information was automatically captured through the implementation of a GIS Web service and an RFID network (Del Vecchio & Carter, 2004). This location or presence awareness adds an additional useful dimension to dialogue between IM users. There are alternative methods of collecting the location data and these options include

- entering and geocoding addresses,
- GPS data collection, and
- RFID-tag readings.

The RFID-tag approach is interesting due to future expectations for expansion in the use of this technology. RFID tags must be read by a stationary reading device, and the location of this reader can be determined by address geocoding or GPS. RFID tags that are read can be related to their location, and this could prove useful for supply chain applications. This experiment builds on the assertions of Mitchell (2003), which state that Web services will further propel GIS into the spotlight.

FUTURE TRENDS

Geospatial technologies have begun to play mission-critical roles on a large scale in a variety of different types of organizations. Early adopters of geospatial technologies included many organizations that traditionally have maintained maps and records to manage their assets and operations. Geospatial functionality will increasingly be embedded in various business applications (e.g., SAP ERP) that might support almost any organization with simple geospatial data visualization, queries, and analysis. In addition, spatial extensions to commercial RDBMS products will permit organizations to visualize the geospatial dimensions of their business data without a major investment.

The ability to deliver data in real time to mobile users extends GIS to the field. The number of mobile GIS users and the wireless communication of data to and from these mobile users will continue to expand. Location services will deliver focused geospatial-information services (e.g., directions, nearest services, etc.) to cell-phone users willing to pay for these supplemental options. According to Nellis (2004), these innovations are "…now at the heart of a vast array of real-time interactive mobile computing, geo-location applications and asset management, along with wireless geographic services that are revolutionizing the role of geography and geospatial information analysis in meeting the needs of everyday society."

The high cost of geospatial data has been a major impediment to its adoption. As geospatial data are increasingly available from geospatial Web-services-compliant sites, costs to access data will decline and users will be able to develop

and use applications more rapidly. There will continue to be some significant policy concerns and debate with regard to the privacy implications of GIS technology as outlined in Balough (2001).

CONCLUSION

Geospatial information systems provide unique capabilities for geospatial-data visualization and analysis. While the large data sizes associated with this technology make real-time and mobile uses more challenging, it is now possible, and increasingly practical and necessary, to use GIS across an enterprise for real-time, mobile geospatial applications. Military and homeland-security markets are currently driving the market for collaborative GIS enterprise applications. Other markets have also emerged and more will surely follow as the advantages of location awareness and geospatial analysis are better appreciated.

REFERENCES

Atkinson, P., & Martin, D. (2000). Innovation in GIS application. *Computers, Environment, and Urban Systems, 24*, 61-64.

Balough, R. C. (2001). *Global Positioning System and the Internet: A combination with privacy risks* (Excerpt from the Chicago Bar Association's CBA record). Retrieved January 20, 2004, from *http://www.isoc.org/internet/issues/privacy/balough.shtml*

Boettcher, R. L. (2000). Collaborative GIS in a distributed work environment. *Master's Abstracts International, 38*(4), 1097-1190. (UMI No. AAT MQ46234).

Cockings, S., Dunn, C. E., Bhopal, R. S., & Walker, D. R. (2004). Users' perspectives on epidemiological, GIS, and point pattern approaches to analyzing environment and health data. *Health & Place, 10*, 169-182.

Del Vecchio & Carter, D. A. (2004). Enhanced presence management in real-time instant messaging systems. *Proceedings of the 14th Information Resource Management International Conference.*

Duffy, D. (2002, August 1). GIS goes worldwide. *CIO Magazine*. Retrieved August 20, 2002, from *http://www.cio.com*

Geospatial Information & Technology Association (GITA). (2004). GITA Web site. Retrieved March 15, 2004, from *http://www.gita.org*

Kevany, M. J. (2003). GIS in the World Trade Center attack: Trial by fire. *Computers, Environment, and Urban Systems, 27*, 571-583.

Kumar, V., Bugacov, A., Coutinho, M., & Neches, R. (1999). Integrating geographic information systems, spatial digital libraries and information spaces for conducting humanitarian assistance and disaster relief operations in urban environments. *Proceedings of the Seventh ACM International Symposium on Advanced Geographic Information Systems*, 146-151.

Kwan, M. P., & Lee, J. (in press). Emergency response after 9/11: The potential of real-time 3D GIS for quick emergency response in micro-spatial environments. *Computers, Environment, and Urban Systems*.

Manoharan, T., Taylor, H., & Gardiner, P. (2002). A collaborative analysis tool for visualization and interaction with spatial data. *Proceedings of the Seventh International Conference on 3D Web Technology*, 75-83.

Medeiros, S. P. J., de Souza, J. M, Strauch, J. C. M., & Pinto, G. R. B. (2001). Coordination aspects in a spatial group decision support collaborative system. *Proceedings of the 2001 ACM Symposium on Applied Computing*, 182-186.

Mitchell, R. L. (2003, December 15). Web services put GIS on the map. *Computerworld*. Retrieved December 15, 2003, from *http://www.computerworld.com*

Munkvold, B. E. (2003). *Implementing collaboration technologies in industry: Case examples and lessons learned*. London: Springer-Verlag.

Nellis, D. M. (2004, February 20). Geospatial information, cybergeography, and future worlds. *Directions Magazine*. Retrieved March 15, 2004, from *http://www.directionsmag.com*

Orton Family Foundation. (2004). CommunityViz Web site. Retrieved March 15, 2004, from *http://www.communityviz.org*

Rauschert, I., Agrawal, P., Sharma, R., Fuhrmann, S., Brewer, I., & MacEachren, A. (2002). Designing a human-centered, multimodal GIS interface to support emergency management. *Proceedings of the 10th ACM International Symposium on Advanced Geographic Information Systems*, 119-124.

Simcock, T., Hillenbrand, S. P., & Thomas, B. H. (2003). Developing a location based tourist guide application. *Proceedings of the Australian Information Security Workshop Conference on ACSW Frontiers 2003, 21*, 177-183.

Songnian, L. (2004). Design and development of an Internet collaboration system to support GIS data production management. *Dissertation Abstracts International, 64*(7), 3147-3487. (UMI No. AAT NQ82574)

Stasik, M. I. (2000). Collaborative planning and decision-making under distributed space and time conditions. *Dissertation Abstracts International, 60*(7), 2629-2758. (UMI No. AAT 9938995)

Urban and Regional Information Systems Association. (2004a). *Hotbutton questions/issues from 2002 conference*. URISA Public Participation GIS Web site. Retrieved March 15, 2004, from *http://www.urisa.org/ppgis.html*

Urban and Regional Information Systems Association. (2004b). URISA Public Participation GIS Web site. Retrieved March 15, 2004, from *http://www.urisa.org/ppgis.html*

Voss, A., Denisovich, I., Gatalsky, P., Gavouchidis, K., Klotz, A., Roeder, S., et al. (in press). Evolution of a participatory GIS. *Computers, Environment, and Urban Systems*.

Zerger, A., & Smith, D. I. (2003). Impediments to using GIS for real-time disaster decision support. *Computers, Environment, and Urban Systems, 27*, 123-141.

KEY TERMS

C4ISR: Command, control, communications, computers, intelligence, surveillance, and reconnaissance, a military application framework that makes extensive use of GIS technologies.

CAD: Computer-aided design.

Enterprise Collaboration: Application of systems and communications technologies at the enterprise level to foster the collaboration of people and organizations to overcome varying levels of dispersion to accomplish a common goal. This term, when applied to the use of technologies, is also known as e-collaboration or distributed collaboration.

Geospatial Information Systems (GISs): Systems that provide a common framework for jointly visualizing the world.

Global Positioning Satellite (GPS): A format of presenting geospatial data for tracking purposes used often in location-based services.

RFID: Radio-frequency identification. This technology uses the electromagnetic spectrum radio signals to transmit information from a transponder (tag) to a receiver for purposes of identifying items. This technology has been in development for a standard to replace the Universal Product Code (UPC) symbol with the Electronic Product Code (ePC) symbol through the Auto ID Center, formerly of MIT.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1278-1283, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Geospatial Interoperability

Manoj Paul
Indian Institute of Technology, India

S.K. Ghosh
Indian Institute of Technology, India

INTRODUCTION

Spatial information is an essential component in almost all decision support system due to the capability it provides for analyzing anything that has reference to the location on earth. Spatial data generally provides thematic information of different aspects over a region. Geospatial information, a variant of spatial information, is generally collected on thematic basis, where individual organizations are involved on any particular theme. Geospatial thematic data is being collected from decades and huge amount of data is available in different organizations (Stoimenov, Dordevi´c, & Stojanovi´c 2000). Information communities find it difficult to locate and retrieve required geospatial information from other geospatial sources in reliable and acceptable form. The problem that has been incurred is the lack of standards in geospatial data formats and storage/access mechanism (Devogele, Parent, Spaccapietra, 1998). Heterogeneity in geospatial data formats and access methods poses a major challenge for geospatial information sharing among a larger user community.

With the growing need of geospatial information and widespread use of Internet has fostered the requirement of geospatial information sharing over the Web. The *Geo-Web* (Lake, Burggraf, Trninic, & Rae, 2005) is being envisioned to be a distributed network of interconnected geographic information sources and processing services that are:

- Globally accessible, that is, they live on the internet and are accessed through standard Open Geospatial Consortium (OGC) and W3C interfaces,
- Globally integrated data sources that make use of standard data representation for sharing and transporting geospatial data.

Unless a standard means for geospatial information sharing is developed, interoperability cannot be realized. Without successful interoperability approaches, the realization of Geo-Web is not possible. Geo-Web is being developed to address the need for access to current and accurate geospatial information from diverse geospatial sources around the world. The *National Spatial Data Infrastructure* (NSDI) initiative has been taken by many nations for providing integrated access of geospatial information (Budak, Sheth, & Ramakrishnan, 2004). Actual data will be kept under the jurisdiction of the organization producing that data. A user will be interested in availing geospatial services through well-defined interface. Without some internationally agreed upon standards for geospatial data and computational methodology, this cannot be made into existence. This chapter discusses several issues towards geospatial interoperability and adoption of *geography markup language* (GML) (Cox, Cuthbert, Lake, & Martell, 2001; Lake et al., 2005) as a common geospatial data format. The associated technologies that can be used for realizing geospatial interoperability have also been discussed.

BACKGROUND

The need for integrated and interoperable geospatial system has been felt for long time and several methods for information integration have been adopted into geospatial domain as well (Devogele et al., 1998; Guan et al., 2003). NSDI (Shekhar, Vatsavai, Sahay et al., 2001) attempts to bring the single point accessibility of geospatial information. But the heterogeneity in geospatial data formats and access mechanism immediately puts into concern about some standard way of sharing data.

The *Open Geospatial Consortium* (OGC) (2000) is an international voluntary consensus standards organization defining standards to bring geospatial computing into mainstream computing. In OGC, several organizations worldwide collaborate in an open consensus process encouraging development and implementation of standards for geospatial content and services, GIS data processing and exchange. The standards defined by OGC are popularly known as *open GIS standards*. These are being adopted by GIS vendors, which in turn increase the possibility of geospatial data sharing. The proposition of geography markup language (GML) (Cox et al., 2001; Lake et al., 2005) as standard data transformation format and service based sharing of geospatial interoperability is going to add new dimension in geospatial interoperability.

The proposition of GML has changed the concept of sharing geospatial information among large-scale users.

Being an XML-based encoding method, GML allows users to share the geospatial information irrespective of the platform or the system in which the data repositories are residing (Badard & Richard, 2001; Zaslavsky, Marciano, Gupta, & Baru, 2000). The advancement of Web service technology and service oriented architectures (Erl, 2004) has led the further progress in geospatial information domain. Geospatial data are now increasingly becoming available on the Web (Kim & Kim, 2002) as geospatial services. The advantage is that they are capable of providing geospatial data in GML-encoded format. Thus, anybody having the knowledge about the geospatial service interface can utilize the information in their application.

GEOSPATIAL INTEROPERABILITY

The main objective of the chapter is to discuss the support that GML and its associated technologies provide towards interoperability.

Geography Markup Language

XML technology today is extremely widespread. It is the "lingua franca" of emerging e-business frameworks, and it powers the generation of thousands of Web sites[1]. Due to its text format it is easily processable across different computing platform. XML has changed the way people thought about interoperability among large-scale disparate systems. GML is based on XML with added support for spatial geometries. In GML, geospatial entities are treated as features. Each feature should have its own properties, both spatial and non-spatial. It is proposed that data has to be shared to the user in GML encoded format irrespective of the heterogeneity of their data formats (Cox et al., 2001; Lake et al., 2005).

There are several base structures defined for modelling a geospatial domain in GML like geometry, features, and so forth. A sample GML instance depicting a geospatial feature *School* is shown in Figure 1. It assigns a feature-id (say, *fid*) for the feature. The polygon geometric element describes the spatial expansion of the feature. It is also geo-referenced by spatial reference system by the element *srsName*.

GML Geometry

Several basic primitive geometry elements have been described in GML for describing the geometric properties of real world objects like rivers, roads, states and so forth. Table 1 provides detailed description of different GML geometry elements. The base schema *geometry.xsd* provides the constructs and structure for geometries. There are a number of homogeneous geometry collections that are predefined in the geometry schema[2]. The fundamental geometry element is "co-ord." All other geometry elements are derived from this basic geometric element. Some of the composite geometry structures in the geometry core schema are as follows: a *multipoint* is a collection of *points*; a *MultiLineString* is a collection of *LineStrings*; and a *MultiPolygon* is a collection of *polygons*. Many new geometric elements have been added to the geometry schemas in GML 3.0, including *curve*, *surface*, *solid* and so forth.

GML Feature

GML defines features, which are different from the concept of geometry objects. A feature is an application object that represents a physical entity, for example, a road, a river, or a forest. A feature may or may not have geometric aspects. The distinction between features and geometry objects in GML contrasts with models used in other geographic information systems.

In GML, a feature can have various geometric properties that describe aspects or characteristics of the feature (e.g., the feature's *point* or *extent* properties). The feature structure

Figure 1. A sample GML instance for a feature school

```
<Feature fid="142" featureType="school" Description="A middle school">
  <Polygon name="extent" srsName="epsg: 27354">
    <LineString name="extent" srsName="epsg: 27354">
      <CData>
        491888.99,5458045.99,491904.99,5458044.99 491908.99,
        5458064.99, 491924.99, 5458064.99 491925.99, 5458079.99,
        491977.99, 5458120.99 491953.999999466, 5458017.99963357
      </CData>
    </LineString>
  </Polygon>
</Feature>
```

Table 1. Geometric elements of GML base schemas with example

Point: Used to encode instances of the point geometry class.	<Point gid="P1" srsName="http://www. opengeospatial.org/gml/srs/epsg.xml#4326"> <coord> <X>56.1</X> <Y>0.45</Y> </coord> </Point>
LineString: A linear path defined by a list of coordinates that are assumed to be connected by straight line segments	<LineString srsName="http://www.opengeospatial.org/gml/srs/epsg.xml#4326"> <coordinates> 0.0,0.0 20.0,35.0 100.0,100.0 </coordinates> </LineString>
LinearRing: Closed, simple linear path which is defined by a list of coordinates that are assumed to be connected by straight line segments.	<LinearRing srsName="http://www.opengeospatial.org/gml/srs/epsg.xml#4326"> <coordinates> 0.0,0.0 100.0,0.0 100.0,100.0 0.0,100.0 0.0,0.0 </coordinates> </LinearRing>

of geospatial objects is defined in *feature.xsd*. GML also provides the ability for features to share a geometry property with one another by using a remote property reference on the shared geometry property. For example, a *building* feature in a particular GML application schema might have a position given by the primitive GML geometry object type *point*. However, the *building* is a separate entity from the *point* that defines its position. Any such property may share its geometry object with properties of other features. An *xlink: href* attribute on a GML geometry property means that the value of the property is the resource referenced in the link.

GML Application Schema

For modeling any geospatial domain, we need to identify the different features of the domain and their properties, relationships. A subset of the base schema, proposed by OGC, can be used for this purpose. A GML application schema is an XML schema written according to the GML rules for application schemas and defines a vocabulary of geographic objects for a particular domain of discourse (Cox et al., 2001; Lake et al., 2005).

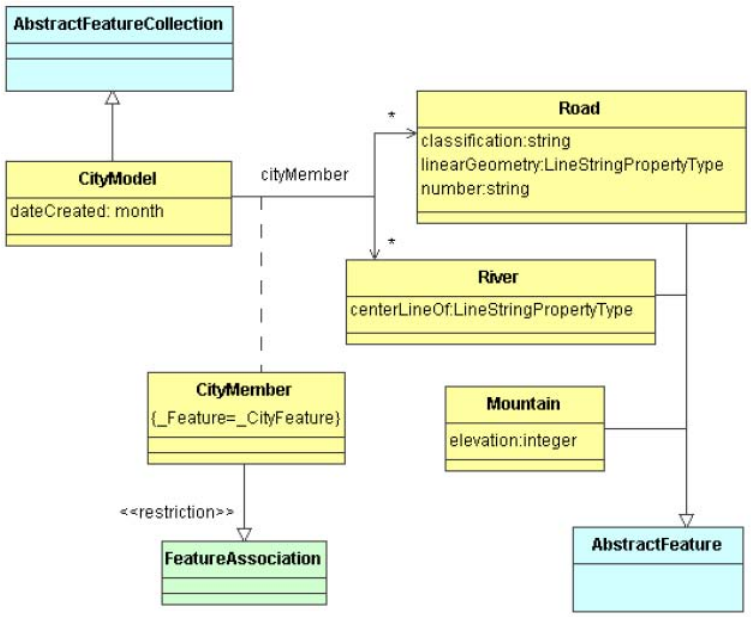
Application schemas are generally designed using ISO 19103 conformant unified modeling language (UML), and then the GML application schema created by following the rules given in ISO DIS 19136. Let us give an example for describing the overall process of application schema generation. For modelling a *city*, we need to incorporate features like *road, river,* and so forth. Each of these features will have several properties and relationships with other features. A UML model incorporating the base schemas can be generated to capture the *City* model precisely as shown in figure 2 (OGC, 2000). It shows different features, which are included for modeling the *city* feature, the relationships among them. Some rules can be applied to generate the application schema from the UML model. For ensuring interoperability, data providers have to publish their application schema. The data user can easily import them to their data repository if the schema is provided along with the actual data.

Advanced Features in GML

Several advanced concepts have been incorporated in GML 3.1.3[3]. New additions in GML 3.0 support complex geometries, spatial and temporal reference systems, topology, units of measure, metadata, gridded data, and default styles for feature and coverage visualization. The most important of these is the support for 3D geometry and topology definition. Topology is the branch of mathematics describing the properties of objects which are invariant under continuous deformation. For example, a circle is topologically equivalent to an ellipse because one can be transformed into the other by stretching. In geographic modelling, the foremost use of topology is in accelerating computational geometry. Topology, realized by the appropriate geometry, also allows a compact and unambiguous mechanism for expressing shared geometry among geographic features.

The GML temporal schemas extend the core elements of GML to include elements for describing the temporal characteristics of geographic data; their purpose is to provide a means of describing the history of a dynamic feature basic temporal schema. The underlying spatio-temporal model strives to accommodate both feature-level and attribute-level time stamping; basic support for tracking moving objects is also included. *gml:_TimeObject, gml:_TimePrimitive, gml: TimeInstant; gml:timePosition, gml:TimePeriod, gml:_dura-*

Figure 2. Sample application schema for a city

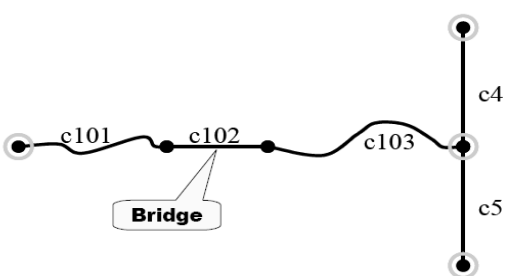


tion have been defined in GML 3 to capture the temporal behavior of dynamic objects.

GML 3.1.1 also supports the creation of complex geometry using simple geometry primitives. Figure 3 is an example of instances of two different feature types, a *road* and a *bridge*, sharing a *gml:LineString* geometry using an *XLink*. It also demonstrates a *gml:Composite* curve composed of *gml:curveMember* elements that are geometries with different interpolation.

GML supports geo-spatial interoperability in a number of ways. It provides a common schema framework for defining geo-spatial features. GML further supports interoperability by providing a common set of GML geometry types. While two different schema authors might for example model a road in different ways they can share the same mechanisms for geometry description and it is then very likely that one can interpret the correspondence between the two schemas. The overall scheme for interoperability can be as shown in Figure 4, which resemblence the service-based integration of geospatial repositories.

Figure 3. Composite geometry support in GML



Supporting Technology

GML alone cannot be sufficient towards achieving interoperable *Geo-Web*. It is just a data description language and there is the need for other associated technologies for its full-fledged use. This section provides some of these related technologies to make GML useful.

Visualization

GML represents the content of data without providing any information of how to visualize the data. Since GML is an XML standard, XSLT can be used to convert it into vector graphics format like SVG, VML, and so forth. SVG is ideal for displaying GML-encoded geospatial information in a Web browser. SVG provides support for 2D graphics including rectangles, circles, ellipses, lines, polylines, polygons, symbols and path elements. In order to visualize GML data, XSLT stylesheets are required. Stylesheets are used to convert and format GML into SVG and different style and symbols can be applied. The overall methodology fro GML visualization is shown in figure 5. Some works on GML visualization have been proposed by (Guo, Zhou, Zu, & Zhou, 2003; Shekhar et al., 2001).

Figure 4. GML and application schema for ensuring interoperability

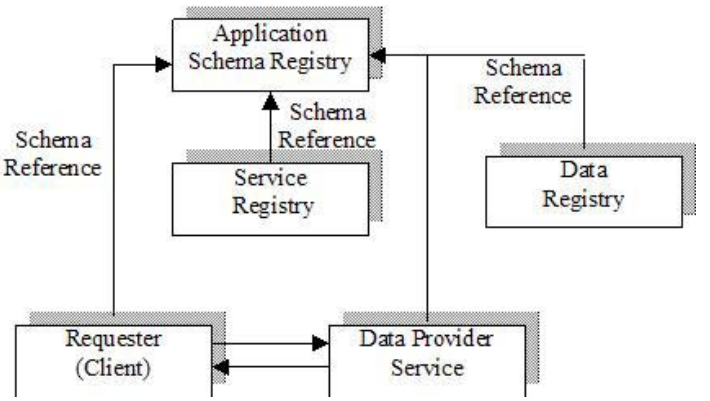
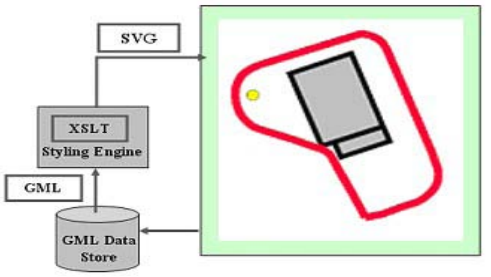


Figure 5. Relationship between GML and SVG



Geospatial Services

With Web service technology providing interoperability among disparate systems, more and more geospatial data are being provided as geospatial services. Similar to general Web service computing, where XML has been the backbone for data sharing and communication, geospatial services make use of GML for data and message transportation. OGC has come up with several specifications for service-based computations in geospatial domain (Doyle & Reed, 2001). These are mainly the following:

- **Web feature service:** Allows feature level access of geospatial data in GML format along with the feature schema
- **Web map service:** Provides geospatial features in the form of map/images
- **Web coverage service:** Allows access of data at the coverage level
- **Catalog service:** Standard registry for services for service discovery purpose

The service based geospatial computation methodology with the help of GML is perceived to be successful in interoperable geospatial data sharing. A framework for service-based geospatial information integration for achieving interoperability has been described in (Paul & Ghosh, 2006b).

GML APPLICATION

Some of recent works, primarily on geospatial infrastructure development, are making use of GML as the standard geospatial data formats. The integration framework proposed in (Boucelma, Garinet, & Lacroix, 2003; Paul & Ghosh, 2006a,2006b; Stoimenov et al., 2000) uses GML for achieving interoperability.

Boucelma et al. (2003) proposes an approach, which uses a WFS-based mediation with the help of *derived wrappers* for geospatial information integration. It proposes a multi-tier, client-server architecture and uses standard wrappers to access data, extended by derived wrappers that capture additional query capabilities. It uses GML as the geospatial data sharing format. The mediation architecture is composed of three main layers: a GIS mediator, *Web feature service (WFS)* and data sources. The GIS mediator is in charge of query processing. The WFS layer receives the query, executes them in the data sources and returns the result in GML format to the mediator. Service-based integration methods for geospatial information using GML have been proposed in (Erl, 2004; Guan et al., 2003; Paul & Ghosh, 2006b).

With increasing adoption of GML for geospatial information sharing, some query language for GML has also been

proposed. Corcoles and Gonzalez (2001) propose a spatial query language for GML. The data model and the algebra underlying the query language has been designed that support spatial features.

With increasing availability of geospatial information in GML-encoded format, the need has arisen for the efficient storage of it in standard DBMSs. Li, Li, and Zhou (2004) propose an approach in which a schema tree is generated from the GML application schema. The tree is subsequently mapped into the relational schema. Paul and Ghosh (2006a) proposed methodology uses a combination of *Element Level* mapping and *Feature Level* mapping for GML storage and utilizes the semantic concepts. Corcoles and Gonzalez(2002) analyse different methodologies for GML storage. They compare three approaches for GML storage: *LegoDB*, a structure-mapping approach, and two simple model-mapping approaches, *Monet* over Relational database and *XParent*. A performance study is conducted using different data sets.

FUTURE TRENDS

The requirement of using XML-based data exchange format for geospatial domain (Badard & Richard, 2001) has given rise to the concept of GML. The usability of GML is increasing day-by-day and thus providing several research issues on effective use of it in different application. There are several future scopes involving GML—GML querying, GML mining, and so forth. The research work on GML querying focuses on query optimization and query expansion involving composite geospatial features. GML mining, on the other hand, involves finding some pattern form a set of GML data. This has direct correspondence with information retrieval form GML document. Advancement in research with XML is being used for GML also.

CONCLUSION

With the increasing need of geospatial information among organizations and the heterogeneity in GIS has led to formalizing a standard way for geospatial data sharing. OGC proposed GML towards the realization of interoperable access of geospatial information. Application schema is aimed at providing interoperability to a significant extent. The adoption of Web service technology for providing geospatial services has brought a new era in geospatial integration paradigm. GML has found extensive use for service description, data transportation, and registry description. The NSDI approaches are increasingly focusing on the realization of service based geospatial information access portal with extensive use of GML.

REFERENCES

Badard, T., & Richard, D. (2001). Using XML for the exchange of updating information between geographical information system. Computers, Environment and Urban Systems, 2001

Boucelma, O., Garinet, J., & Lacroix, Z. (2003). The VirGIS WFS-Based Spatial Mediation System. *Proceedings of ACM CIKM,* New Orleans.

Budak, A., Sheth, A., & Ramakrishnan, C. (2004). *Geospatial ontology development and semantic analytics. Handbook of geographic information science.* Blackwell Publishing.

Corcoles, J.E., & Gonzalez, P. (2001). A Specification of a Spatial Query Language over GML. *Proceedings of the 9th ACM international symposium on Advances in geographic information systems,* (pp. 112-117).

Corcoles, J.E., & Gonzalez, P. (2002). Analysis of Different Approaches for Storing GML Documents. *Proceedings of ACM GIS*, (pp 11-16). ACM Press.

Cox, S., Cuthbert, A., Lake, R., & Martell, R. (2001). Geography Markup Language (GML) 2.0 - URL: http://www.opengis.net/gml/01-029/GML2.html

Devogele, T., Parent, C., & Spaccapietra, S. (1998). On spatial database integration. *International Journal of Geographic Information Science, 4*, 335–352.

Doyle, A., & Reed, C. (2001). Introduction to OGC web services. OGC Interoperability Program White Paper.

Erl, T. (2004). *Service-oriented architecture: A field guide to integrating XML and web services.* Upper Saddle River, NJ: Prentice Hall PTR,

Guo, Z., Zhou, S., Xu, Z., & Zhou, A. (2003). G2ST: A Novel Method to Transform GML to SVG. *Proceedings of ACM-GIS 2003*, ACM Press.

Guan, J. Zhou, S. Chen, J. et al. (2003). Ontology-based GML schema matching for information integration. *Proceedings of 2nd IEEE International Conference on Machine Learning and Cybernetics,* Vol.4, (pp. 2240-2245). Xi'an, China: IEEE CS Press.

Kim, D.H., & Kim, M.S. (2002). Web GIS service component based on open environment. *Geoscience and Remote Sensing Symposium*.

Lake, R., Burggraf, D.S., Trninic, M., & Rae, L. (2005). *Geography markup language.* John Wiley and Sons Ltd.

Li, Y., Li, J., & Zhou, S. (2004). GML Storage: A Spatial Database Approach. P*roceedings of ACM ER Workshops,* Shanghai, China.

OGC, Ed. (2000). *The OpenGIS® guide—introduction to interoperable geoprocessing and the OpenGIS specification*. MA: Open GIS Consortium, Inc.

Paul, M., & Ghosh, S.K. (2006a). An Approach for Geospatial Data Management for Efficient Web Retrieval. *IEEE International Conference on Computer and Information Technology (CIT)*. Seoul, Korea: IEEE Computer Society Press.

Paul, M., & Ghosh, S.K. (2006b). An Approach for Service Oriented Discovery and Retrieval of Spatial Data. *Proceedings of ICSE International Workshop on Service Oriented Software Engineering (IW-SOSE '06),* (pp. 88-94). Shanghai: ACM Press.

Shekhar, S., Vatsavai, R.R., Sahay, N., Burk, T.E., & Lime, S. (2001). WMS and GML based interoperable web mapping system. *In proceedings of the 9th ACM International Symposium on Advances in Geographic Information Systems*, ACM Press.

Stoimenov, L., Dordevi´c–Kajan, S., & Stojanovi´c, D. (2000). Integration of GIS Data Sources over the Internet Using Mediator and Wrapper Technology. *In Proceedings of MELECON 2000*, Cyprys.

Zaslavsky, I., Marciano, R., Gupta, A., & Baru, C. (2000). XML-based Spatial Data Mediation Infrastructure for Global Interoperability. *In proceedings of the 4th Global Spatial Data Infrastructure Conference,* Cape Town

KEY TERMS

Application Schema: A metadata structure describing the geospatial feature for the domain of interest. GML adheres to XML-based schema language for application schema design.

Geo-Web: A globally integrated and accessible spatial infrastructure that comprises a number of interconnected geospatial datasets and Web services.

Geospatial Data: Geographically referenced spatial data that provides some thematic information over a region.

Geospatial Web Services: A Web service that provides access to, or data processing on, geographic information. The OGC Web Feature Service (WFS), Web Map Service (WMS) are examples of geospatial Web service.

Geospatial Interoperability: Ability to access, share and manipulate of geospatial data stored in heterogeneous distributed repositories**.**

GIS: A system of computer hardware, software and data for collecting, storing, analyzing and disseminating information about areas of the earth.

Spatial Reference System: A co-ordinate system that defines the maximum possible extent of space that is referenced by a given range of coordinates with respect to a point on earth.

Spatial Geometry: Describes the structure of spatial objects in terms of points, lines, polygons, polylines, and so forth.

UML: UML is a language for specifying, constructing, visualizing, and documenting the artifacts of a software-intensive system.

XML: An open standard for exchanging structured documents and data over the Internet that was introduced by the World Wide Web Consortium (W3C)

ENDNOTES

1 http://gislounge.com/ucon/ucgml2.shtml
2 http://cite.opengeospatial.org/test_engine/GML_2.1.2/files/gml_spec_2_1_2/
3 http://www.opengis.net/gml/

GIS and Remote Sensing in Environmental Risk Assessment

X. Mara Chen
Salisbury University, USA

INTRODUCTION

The existence, well-being, and sustainable development of the global economy hinges upon the state of the earth's environment. Effective environmental risk assessment and management issues have become increasingly important. With the ever-growing global population and expanding economic development, we consume more natural resources, produce more waste, and develop more areas into the regions that are prone to environmental risks. Although humans have interacted with the environment for thousands of years, environmental risk assessment and management is only a recent research undertaking. As the industrialization has made the human-environment interactions more dynamic and complex, the increased environmental risks have propelled and compelled people to use technologies for identifying and solving problems. The earliest global environmental applications of remote sensing and GIS technologies began in the 1960s, particularly marked by the successful launch of the TIROS-1, the first meteorological satellite, and the development of computer-based geographic information systems (GIS). The story *Silent Spring* (Carson, 1962) awoke the public's environmental consciousness and promoted the public demands for governments to set up environmental protection policies and research priorities. The birth of the U.S. Environmental Protection Agency (EPA) in 1970 set the stage for modern environment risk assessment. The launch of the LANDSAT program in 1972 created a new way for monitoring global land use and land cover changes (Foley, 1999; Goward, Masek, Williams, Irons, & Thompson, 2001).

BACKGROUND

Environmental risks ranging from natural to human-induced hazards present growing threats to communities at local, national, regional, and global scales. Effective and timely environmental risk assessment and management has become a forefront issue in ensuring the health and functions of modern civilization. Information technologies offer a promising approach of integrating and processing information from various sources and formulating comprehensive solutions to complex environmental problems. In particular, GIS and remote sensing technologies together offer the abilities of rapidly collecting data, processing and integrating data and information, and displaying results in geographic-referenced maps and reports. Environmental professionals have increasingly utilized remote sensing and GIS to study human activities and the environment (Chen, Blong, & Jacobson, 2003; Turner, 2003). Multi-spectral and multi-resolution sensors mounted on different platforms (aircrafts or spacecrafts) have become our "eyes" in space, providing constant and consistent environmental surveillance. In the mean time, GIS has provided us with the extended brain-power to store, process, analyze, and display unprecedented vast amounts of complex data. The technological marriage of remote sensing and GIS created a powerhouse that allows remotely sensed data to be directly fed into GIS for integrated analysis and visualization. Satellite remote sensing provides a systematic and synoptic knowledge base about the earth's complex geophysical phenomena (Tralli, Blom, Zlotnicki, Donnellan, & Evans, 2005). A GIS-based integrated approach can be used for the risk management of natural hazards (Chen et al., 2003).

CURRENT STATE

Effective environmental risk assessment and management is a complex process (Figure 1). The success of the process depends upon the prerequisite steps of comprehensive data collection, data integration, and analysis. Remote sensing is very critical in capturing the dynamic and vicissitudinary nature of hazards. The essential environmental risk assessment database must encompass the measurements and information on hazard types, occurrence probability and frequency, intensity and magnitude, and their proximity to the human environment. Remote sensing technology offers spatial, spectral, and temporal monitoring functionalities to fully measure these environmental variables.

Multi-platform remote sensing allows the earth's environment to be monitored at different spatial scales (local, regional, and global), spatial resolution (fine, medium, and coarse), and from a variety of spectral ranges beyond human's visible spectral vision (see Table 1). High resolution imagery (of 5m or less pixel size) is used for precise topography mapping and complex ecosystem change detection in densely populated regions (Ehlers, Gähler, & Janowsky, 2002; Ellis et al., 2006; Morsdorf, Meier, Kotz, Itten, Dobbertin, & Allgower, 2004). Medium-resolution satellite imagery (5-100m) is utilized for diverse global environment monitoring and assessment. For example, data from many

Figure 1. Flowchart of GIS and remote sensing in environmental risk assessment and management

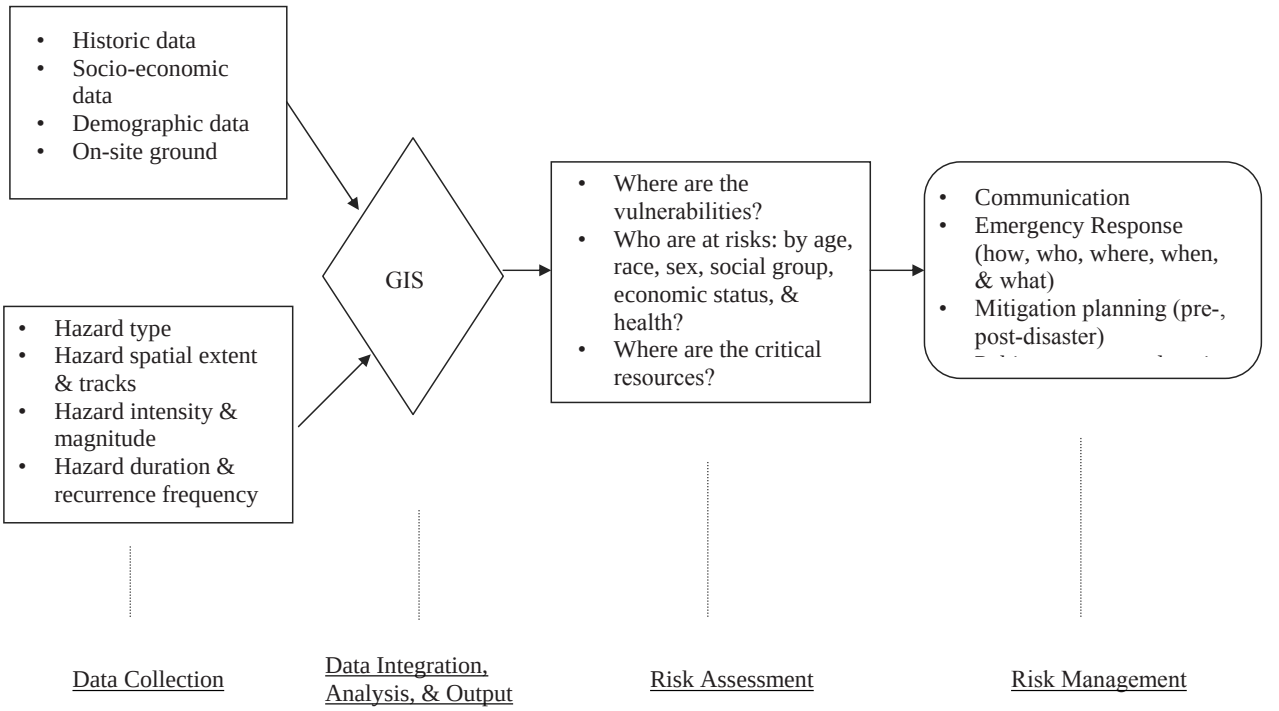


Table 1. Selected remote sensing application in environmental risk assessment

Remote Sensing Type		**Characteristics**	Environmental Risk **Assessment**
Platform	Aircraft	Local & regional coverage at certain time interval	A variety of environmental monitoring and assessment at relatively local levels
		low, medium, and high altitudes	
	Spacecraft (Shuttle & Satellite)	regional and global coverage	Monitoring the environment at regional or global levels in a long-term repetitive manner
		long term and repetitive surveillance	
		high & very high altitudes	
Sensor Spectral Range	Ultra violet (UV)	0.3–0.4 μm	Oil spills, wildlife inventory
	Visible	0.4–0.7 μm	Various land use and land cover assessment
	Near & mid infrared (NIR, MIR)	0.7–3 μm	Water and land boundary, vegetation differentiation
	Thermal infrared (TIR)	3–14 μm	Fire, volcanic activities, thermal pollution
	Microwave/RADAR	1 mm–1 m	Oil spill, deforestation, polar ice study
Sensor Spectral Resolution	Panchromatic mode	One broad band	Preliminary assessment
	Multispectral mode	Several to tens of broad bands with spectral range in μm	Comprehensive comparative study, feature discrimination
	Hyperspectral mode	Hundreds of narrow bands with spectral range in nm	Possible specific feature identification
Image Spatial Resolution	Fine resolution	Less than 5m	Precise and detailed study
	Medium resolution	5–100m	Local & regional study
	Coarse resolution	Greater than 100m, often in km	Sub-continental & global environment assessment

satellite imaging systems (e.g., Landsat, IRS, SPOT, MODIS) have created new ways of monitoring fundamental land use and land cover changes and a variety of natural hazardous phenomena. Landsat data provide systematic observation data for assessing human risks to earthquakes, volcanoes, floods, landslides, coastal inundation, and forest fires (Mbow, Goïta, & Bénié, 2004; Tralli et al., 2005). Coarse resolution satellite imagery is often used for tracking highly dynamic hazardous weather phenomena such as hurricanes/typhoons, tornadoes, and other types of severe storms.

In addition to precise and accurate observations and measurements of hazards, integrated data processing, and analysis is crucial for a complete and comprehensive risk assessment and management. Many hazards, however powerful, only become harmful when their dynamic spheres intersect with the spheres where people live and develop (Chen, Parrott, & Johnson, 2006). The analytical capacity of GIS ensconces its increasingly important role in environmental risk assessment and management. GIS assists environmental analysts and decision makers to better understand the spatial, socio-economic, and historic aspects of hazards and their associated risks to humans and the human environment.

Spatially, over half of the global population are concentrated along the coastal zones (Finkl, 2000), about half of the population are clustered in cities (Thouret, 1999), especially in coastal cities. Eleven of the world's 15 largest cities are on the coast (Cohen & Small, 1998). About 500 million people live close to active volcanoes (Thouret, 1999). GIS is used for delineating the spatial distribution of hazards, assisting community risk assessment, and emergency response planning (Chen et al., 2006; Mondschein, 1994; Wood & Good, 2004), documenting global offshore hazardous materials sites (Lindsay & Aguirre, 2004), supporting large area land use and land cover studies and sustainable land management (Foley, 1999; Groot, 1997; Lidrah, 2000). Socio-economically, people who are at the same level of exposure to hazards do not necessarily face the same degree of risk. Young children and old people who are of poor health and scarce resources and limited social supporting networks are the most vulnerable. GIS not only helps map exposure risks, but also highlights the importance of understanding environmental injustice and inequality problems (Dolinoy & Miranda, 2004; Fielding & Burningham, 2005). Historically, human activities have significant accumulative effects on transforming the environment. Historic data are often combined with the data of human settlement patterns, historic fire occurrences, and economic developments in the GIS for monitoring coastal evolution, land use changes, and wildfire recurrences (Foley, 1999; Mountford & Sheppard, 2000; Riva, Pérez-Cabello, Lana-Renault, & Koutsias, 2004). GIS is capable of processing data at a full gamut of temporal cycles, ranging from century, to decadal, to annual update cycles, and to even real time observations (Chen, Yang, & Chen, 2005). Historic data analyses help people reflect upon what has happened and what could have been done better in the past, observe what is happening in the present, and predict what potential risks would be and what can be done to better mitigate the risks in the future.

FUTURE TRENDS

With the continuous evolution of computing and information technologies, GIS and remote sensing will facilitate more effective and efficient environmental risk assessment and management. Remote sensing will continue to improve its spatial resolution, radiometric precision, spectral differentiation and coverage, and timely or real time imaging capability. It will become an indispensable tool in providing data for climate change detection and sustainable development study (Cihlar, 2000). In the mean time, GIS will provide more powerful analytical and visualization capabilities to process the increasingly complicated and comprehensive environmental databases. Together, GIS and remote sensing will not only able to estimate post-disaster impact risks, but also project and predict the potential risks and optimum management strategies. Hybrid knowledge-based GIS systems are explored for simulating exposure risks (Fedra & Winkelbauer, 2002; Swartz, Rudel, Kachajian, & Brody, 2003). Potential real time 3D GIS is examined for rapid urban emergency response and flood management (Al-Sabhan, Mulligan, & Blackburn, 2003; Kwan & Lee, 2005).

Despite the genuine utilities and potential promises, many challenges still exist in using GIS and remote sensing for environmental risk assessment and management. Major aspects include:

- No systematic methodology in using the technologies for different types of environmental risk assessment and management applications at different spatial scales (local, national, regional, and global context);
- Inadequate algorithms and procedures developed for integrating heterogeneous data that are acquired at different scales and resolutions, during different times, from different methods, and by different agencies;
- Large gaps between high-level research agendas and local-level practical actions due to a lack of financial resources and technical know-how in using GIS and remote sensing in many poor hazard-prone communities and regions;
- Growing needs for portable and mobile GIS. They can be used in a distributed computing environment simultaneously by multi-agency first responders and decision makers to facilitate optimal procedures for pre-, during-, and post-disaster risk assessment, response, and management. Finding suitable info in the distributed geographic information services is crucial in disaster management (Klien, Lutz, & Kuhn, 2006).

CONCLUSION

As the rapid growing global population and expanding industrialization place more pressure on the earth's environment, GIS and remote sensing will become increasingly critical for environment risk assessment and management. Remote sensing imagery of multi-spectral bands, multi-spatial resolutions, and multi-temporal cycles will provide accurate and timely fundamental data sources for measuring ever-changing environmental variables. GIS will help combine geo-spatial data, from remote sensing and other data sources, socio-economic data, and demographic data to formulate integrated solutions and strategies. More advanced image-processing algorithms and methods will streamline real time data update and dissemination, which will help develop a comprehensive global environment database. New developments in computing and information technologies will propel GIS to overcome current technical challenges for carrying out real time three-dimensional simulations in a distributed computing environment. Together, GIS and remote sensing will help environmental scientists and decision makers to optimize resources allocation, enable global data and information sharing, and promote effective and cost-efficient environmental risk assessment and management practices.

REFERENCES

Al-Sabhan, W., Mulligan, M., & Blackburn, G. A. (2003). A real-time hydrological model for flood prediction using GIS and the WWW. *Computers, Environment and Urban Systems, 27*(1), 9-32

Carson, R. (1962). *Silent Spring (2002 ed.)*. Boston: Houghton Mifflin Company.

Chen, K., Blong, R., & Jacobson, C. (2003). Towards an integrated approach to natural hazards risk assessment using GIS: with reference to bushfires**.** *Environmental Management, 31*(4), 546-560.

Chen, X. M., Parrott, C., & Johnson, K. E. (2006). Key aspects in community-based coastal emergency response GIS. In *Proceedings of 2006 IRMA International Conference: Emerging Trends and Challenges in IT Management* (pp. 114-117).

Chen, X. M., Yang, C., & Chen, S. (2005). Evolution and computing challenges of distributed GIS. *Journal of Geographic Information Sciences, 11*(1), 61-70.

Cihlar, J. (2000). Land cover mapping of large areas from satellites: Status and research priorities**.** *International Journal of Remote Sensing, 21*(6&7), 1093-1114.

Cohen, J. E., & Small, C. (1998). Hypsographic demography: The distribution of human population by altitude. In the *Proceedings of National Academy Science USA, 95*(24)**:** 14009–14014.

Dolinoy, D. C., & Miranda, M. L. (2004). Environmental health perspectives. *112*(17), 1717-1724.

Ehlers M., Gähler, M., & Janowsky, R. (2003). Automated analysis of ultra high resolution remote sensing data for biotope type mapping: New possibilities and challenges. *ISPRS Journal of Photogrammetry & Remote Sensing, 57*, 315-326.

Ellis, E. C., Wang, H., Xiao, H. S., Peng, K., Liu, X. P., Li, S. C., Ouyang, H., Cheng, X., & Yang, L. Z. (2006). Measuring long-term ecological changes in densely populated landscapes using current and historical high resolution imagery**.** *Remote Sensing of Environment, 100*(2006), 457-473.

Fedra, K., & Windelbauer, L. (2002). A hybrid expert system, GIS, and simulation modeling for environmental and technological risk management. *Computer-Aided Civil and Infrastructure Engineering, 17*(2002), 131-146.

Fielding, J., & Burningham, K. (2005). Environmental inequality and flood hazard. *Local Environment, 10*(4), 379-395.

Foley, J. A. (1999). Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Global Biogeochemical Cycles, 13*(4), 997-1027.

Goward, S. N., Masek, J. G., Williams, D. L., Irons, J. R., & Thompson, R. J. (2001). The Landsat 7 mission terrestrial research and applications for the 21st century**.** *Remote Sensing of Environment, 78*(2001), 3-12.

Klien, E., Lutz, M., & Kuhn, W. (2006). Ontology-based discovery of geographic information services: An application in disaster management**.** *Computers, Environment and Urban Systems, 30,* 102-123.

Kwan, M., & Lee, J. (2005). Emergency response after 9/11: The potential of real-time 3D GIS for quick emergency response in micro-spatial environments**.** *Computers, Environment and Urban Systems, 29*(2005), 93-113.

Lindsay, J. A., & Aguirre, R. A. (2004). Global offshore hazardous materials sites GIS. *Marine Technology Society Journal, 38*(3), 36-43.

Mbow, C., Goïta, K., & Bénié, G. B. (2004). Spectral indices and fire behavior simulation for fire risk assessment in savanna ecosystems. *Remote Sensing of Environment, 91*(2004), 1-13.

Mondschein, L. G. (1994). The role of spatial information systems in environmental emergency management. *Journal*

of the American Society for Information Science, 45(9), 678-685.

Morsdorf, F., Meier, E., Kotz, B., Itten, K. I., Dobbertin, M., & Allgower, B. (2004). LIDAR-based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management. *Remote Sensing of Environment, 92*(2004), 353-362.

Mountford, K., & Sheppard, C. R. C. (2000). Chesapeake Bay: The United States' largest estuarine system. *Seas at the Millennium-An Environmental Evolution, 1*(2000), 335-349.

Riva, J., Pérez-Cabello, F., Lana-Renault, N., & Koutsias, N. (2004). Mapping wildfire occurrence at regional scale. *Remote Sensing of Environment, 92*(2004), 363-369.

Swartz, C. H., Rudel, R. A., Kachajian, J. R., & Brody, J. G. (2003). Historical reconstruction of wastewater and land use impacts to groundwater used for public drinking water: Exposure assessment using chemical data and GIS. *Journal of Exposure Analysis and Environmental Epidemiology, 13*(5), 403-416.

Thouret, J. (1999). Urban hazards and risks; consequences of earthquakes and volcanic eruptions: an introduction. *GeoJournal, 49*, 131-135.

Tralli, D. M., Blom, R. G., Zlotnicki, V., Donnellan, A., & Evans, D. L. (2005). Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. *ISPRS Journal of Photogrammetry & Remote Sensing, 59*(2005), 185-198.

Turner, M. D. (2003). Methodological reflections on the use of remote sensing and geographic information science in human ecological research. *Human Ecology, 31*(2), 255-280.

Wood, N. J., & Good, J. W. (2004). Vulnerability of port and harbor communities to earthquake and tsunami hazards: The use of GIS in community hazard planning. *Coastal Management, 32*, 243-269.

KEY TERMS

GIS: A digital-based data storage, data manipulation and analysis, and visualization system and science, consisting of hardware, software, and organization structure. The modern GIS began in the 1960s, and it has evolved into a maturing science discipline. GIS was an acronym for geographic information systems, but was first used to represent geographic information science in 1992 by Michael F. Goodchild.

Human Environment: The complete continuum of matter and conditions that surround human and human society. The state of the earth's environment dictates the existence, safety, and health conditions of human and is affected by the interactions between human and the natural environment.

Image Resolution: The quality of remote sensing imagery is controlled by its resolution, which includes spatial resolution, spectral differentiation, radiometric calibration, and temporal revisit cycle. Spatial resolution (image pixel size) is the minimum ground separations between adjacent objects that appear distinct on the image. Spectral resolution is the sensor's overall spectral range and the width of individual spectral band, with higher resolution having narrower bands and/or overall broader spectral range. Radiometric resolution is the sensor's calibration capability to detect the minimum spectral response differences of imaged objects. Temporal resolution is the time interval between two repetitive sensing cycles for a given region. A smaller time interval gives a better coverage to monitor changes or dynamic phenomena.

Remote Sensing: The technology of obtaining, analyzing, and displaying the information about an object or a phenomenon through remote detection of its reflected and emitted electromagnetic energy. The term "remote sensing" was first used in the United States in the 1950s by Ms. Evelyn Pruitt of the U.S. Office of Naval Research.

Risk & Vulnerability: Risk is the possibility for human to suffer the loss and/or injury (physical, social, economic, property) from hazards. Vulnerability is an area's susceptibility to damage and degradation. Risk and vulnerability often occur side by side.

Risk Assessment: A process to identify hazard types, their intensities/magnitudes, occurring probabilities, and the associated consequences of impact.

Risk Management: The responding practices to better mitigate hazards or recover from any associated adversary impact. Typically, effective risk management includes pre-disaster mitigation planning by means of insurance, land use regulation, and technology; rapid and effective emergency responses during the emergency; post-disaster recovery and rebuilding management; and increased public awareness education.

Global Digital Divide

Nir Kshetri
University of North Carolina at Greensboro, USA

Nikhilesh Dholakia
University of Rhode Island, USA

INTRODUCTION

Despite rapidly falling costs of hardware, software and telecommunications services, a wide gap persists between rich and poor nations in terms of their capabilities of accessing, delivering, and exchanging information in digital forms (Carter and Grieco, 2000). According to a report published by the United Nations Conference on Trade and Development in 2006, a person in a high-income country was more than 22 times likely to use the Internet than someone in a low-income country (UNCTAD, 2006). The ratios were 29 times for mobile phones and 21 times for fixed phones.

An estimate suggested that more than 95% of e-commerce transactions in 2003 were industrialized countries (Tedeschi, 2003). Another estimate suggested that 99.9% of business-to-consumer e-commerce in 2003 took place in the developed regions of North America, Europe, and Asia Pacific (Computer Economics, 2000). This is a form of commercial divide (UN Chronicle, 2003). Another estimate suggests that 80 percent of the global trade in high technology products originates from Europe, the U.S., and Japan (Bowonder, 2001) and 92 % of the patents granted in the world are owned by the members of Organisation for Economic Co-operation and Development (Archibugi and Iammarino, 2000).

Whereas high-income countries have income 63 times that of low-income countries, the respective ratios are 97 for PCs, 133 for mobile phones, and over 2,100 for Internet hosts (Dholakia and Kshetri, 2003). While reliable data on e-commerce transactions are not available, the ratio is likely to be even higher for e-commerce transactions since e-commerce is virtually non-existent in many developing countries. The pattern indicates that the gap between developed and developing countries is wider for more recent technologies such as PC, mobile phone, and the Internet than for technologies which were introduced earlier.

This article provides an assessment of three computer networks that redefine the conventional definition of market value by allowing developing nations and communities (Brooks, 2001) reap the benefits of modern ICTs: Global Trade Point Network (GTPNet) and Little Intelligent Communities (LINCOS).

BACKGROUND

The "global digital divide" is the outcome of the complex interactions between information and communication technologies (ICT) and various economic, political, and social factors in the environment. The global digital divide arguably is one of the strongest non-tariff barriers to the world trade with potentially adverse social, economic and other consequences influencing a developing country's ability to take advantage of opportunities provided by modern ICTs (UN Chronicle, 2003). First, a large majority of potential users in developing countries are unable to afford a telephone line, a PC, and the telephone and Internet services provider (ISP) access charges. Whereas the cost of a PC is 5% of per capita GDP in high-income countries, it is as high as 289% in low-income countries (ITU, 2001). Furthermore, monthly Internet access charge as a proportion of per capita GDP in the world varies from 1.2% in the U.S. to 614% in Madagascar (UNDP, 2001).

Second, even if consumers are willing to pay for the connection of a telephone line, there is a big gap between demand and supply in developing countries. For instance, in 2001, 33 million people in the developing world were on the registered waiting lists for telephone connections, the average waiting periods being over 10 years in some countries.

A third problem is related to the lack of skills. A majority of potential users in developing countries lack English language and computer skills, prerequisites to the use of Internet. For instance, in 1998 about 85% of the text on the Internet was in English (Nunberg, 2000). This proportion was estimated at 80% in 1999. Another estimate suggested that about 70% of the world's Web sites were in English in 2003 (UN Chronicle (2003). Although a shift of Internet content to non-English languages is under way, some knowledge of English is still necessary to use the Internet as the bulk of software used in the Internet is in English (Hedley, 1999) and most of the human-computer interfaces favor English language users (Goodman, 1994).

A fourth problem is related to the lack of relevant content or the content divide (UN Chronicle, 2003). Although there are over 17 billion Web pages[1] in existence, the content remains largely geared to the needs of advanced nations. Most of the information available on the WWW is not relevant to

the needs of people in the developing world (UN Chronicl, 2003). Edejer (2000) observes the difficulty of finding reliable health related information relevant to developing countries online:

Few reports of health research from developing countries are published in journals indexed by Western services such as Medline. Western indexing services cover some 3,000 journals, of which 98% are from the developed world. The whole of Latin America accounted for 0.39% of the total number of articles referenced by Medline in 1996... Because only a small number of journals from developing countries are indexed by Medline, research from these countries is almost invisible.

CREATIVE WAYS TO BRIDGE THE DIGITAL DIVIDE: SOME EXAMPLES

The effectiveness of a network in bridging the global digital divide is thus a function of (1) the network's ability to identify priorities of digitally excluded populations, and (2) the network's ability to attack the major barriers to Internet and e-commerce adoption. In the following section, we examine three projects aimed to enable e-business systems for the global poor: DDD, GTPNet (Figure 1) and LINCOS (Figure 2).

Digital Divide Data (http://www.digitaldividedata.com/)

Digital Divide Data (DDD), a nonprofit organization, was started in 2001 by a group of North Americans. Canada's Jeremy Hockenstein and Jaeson Rosenfeld of the U.S. came up with the idea for DDD. When Hockenstein visited Cambodia in November, 2000, many Cambodians were in Internet cafes, but he realized that they were not using the technology to increase economic productivity (The Associated Press, 2001). Hockenstein felt that Cambodia should focus on attracting IT-related businesses, rather than factories that produce clothing or other manufacturing goods (Reed, 2001).

The two North Americans invested $25,000 of their own money and received a $25,000 grant. An Indian firm donated technical advice and software (Shih, 2003). In 2002, DDD received another $45,000 grant to support DDD's expansion[2].Its main office is located in Phnom Penh, Cambodia. Subsequently, DDD expanded its operations in Vietnam and Laos also (Fast Company, 2005).

DDD was started with 20 employees and the number grew to about 115 by 2003 (Tedeschi, 2003). By the spring of 2004, DDD had 140 people employed in the three countries (St. John, 2004). The employees digitize data from maps or documents and send them back to the country of origin (e.g., the U.S.). DDD's first project was a $50,000 contract to digitize 100 years of archives of the Harvard Crimson, Harvard University's student newspaper (Shih, 2003). Digital Divide Data generated $178,974 in 2003, a 67% increase over 2002, which covered the operating expenses, with some money left for further business development (St. John, 2004). In 2003, DDD workers earned $1,200 a year, which is four times the average Cambodian's income (Tedeschi, 2003).

Most of DDD's employees are women, polio and landmine victims, orphans and other categories of internally displaced people (Anderton, 2003). Whereas the wage offered at Phnom Penh garment factories exporting clothing to the U.S. was $11.25 for a 48-hour work week in 2003, DDD typists earned $16.25 a week working a 36-hour week (Shih 2003). DDD workers do data entry jobs for six hours a day and get English and computer training for another six hours (Helm and Kripalani, 2006). Thus, the work makes them employable in more challenging and better paying jobs. In addition, if employees set aside $20 per month for educational purposes, DDD matched with a $20 scholarship (St. John, 2004). The company awarded 80 such scholarships by the spring of 2004 (St. John, 2004).

In the beginning, DDD faced a number of problems. For instance, typists did not save their documents, which resulted in every new day's typing erasing the previous day's work, and workers were shy to ask questions and embarrassed to point out errors in their colleagues' works (Shih, 2003). Efficiency and quality were problems attributed to DDD's hiring practices which employed people with little IT experience and very limited English language skills. DDD utilizes a rigorous proofreading system and other quality-control processes such as specialized "double entry" software to minimize typographical errors (St.John, 2004). Accuracy rates gradually improved. Estimates suggest that U.S. companies can save 50-60% by outsourcing works to DDD (Anderton, 2003).

Global Trade Point Network (GTPNet) of the World Trade Point Federation (http://www.wtpfed.org/)

The United Nations Conference on Trade and Development (UNCTAD) launched the Global Trade Point Program in 1992 to facilitate the access to international markets for small and medium-sized enterprises (SMEs). The program was taken over by the World Trade Point Federation (WTPF) in November 2002. In mid-2003, GTPNet had a human network of 121 Trade Points in over 80 countries on the 5 continents[3].

In a trade point, participants in foreign trade transactions (e.g., customs authorities, foreign trade institutes, banks, chambers of commerce, freight forwarders, transport and

Figure 1. UNCTAD GTPN

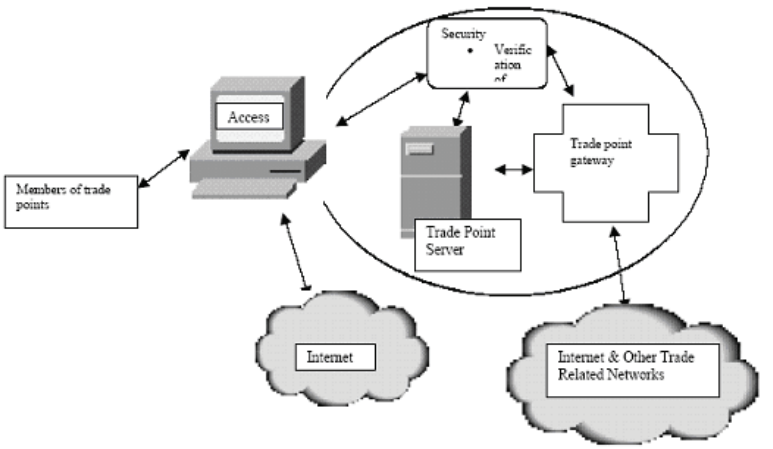
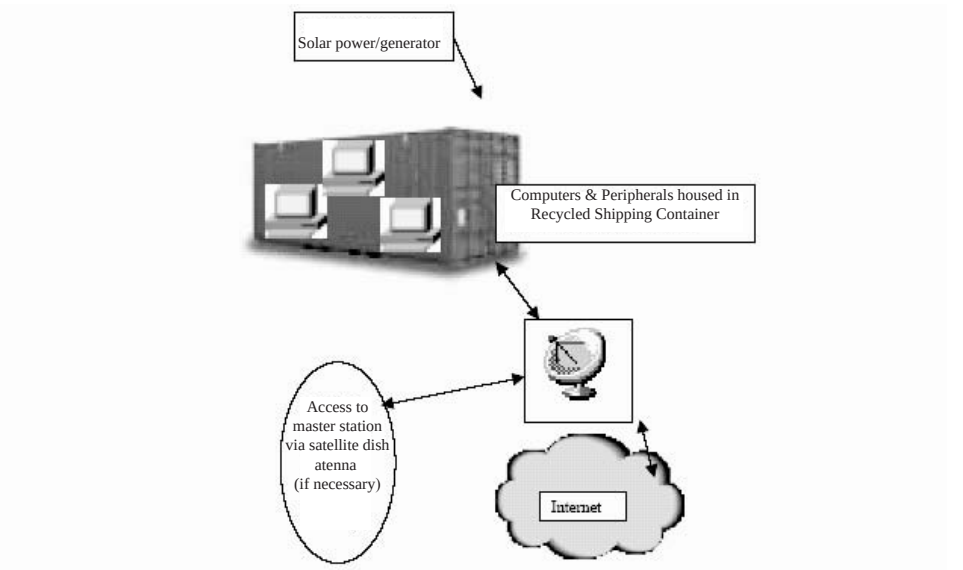


Figure 2. LINCOS network



insurance companies) are grouped together under a single physical or virtual roof to provide all required services at a reasonable cost. It is a source of trade-related information providing actual and potential traders with data about business and market opportunities, potential clients and suppliers, trade regulations and requirements, etc. A survey found that 85.7% of trade point customers are SMEs and micro-enterprises (UNCTAD, 1997).

The ETO System is probably one of the most important aspects of the GTPNet. It was started by the UN Trade Point Development Center (TPDC) in June 1993 and is the world's largest Internet-based business opportunities system. ETOs are offers and demands for products, services, and investment and are distributed point-to-point and company-to-company. They are forwarded to the GTPNet system by Trade Points and third-party information providers. A random survey of ETO users conducted in 1998 revealed that 48% of the ETO users received 1-10 responses per posted ETO, an additional 14% received 10-30 responses, about 7% received over 100 reactions. About a third of respondents made business deals on the basis of ETOs.

Developing countries also have a much higher share in the ETOs than in the overall global e-commerce. A United Nations Trade Point Development Center's (UNTPDC) analysis of ETOs posted on the GTPNet during March 1- July 15, 1998 indicated that 20% of them were posted by U.S. based companies followed by companies in China (19%), South Korea (11%), and India (7%) (UNCTAD, 1998). It is interesting to note that the U.S. accounted for 74% of the global Internet commerce market in 1998 (Wang, 1999).

Password-restricted areas have been added to the GTPNet site which uses state-of-the-art tools for uploading, downloading, automatic updating, and searching for information. Only Trade Points and members of Trade Points can send electronic trading opportunities (ETOs) and see hot ETOs[4]. Java is used to control access, certify trading partners, and handle payments. The Java-based secure infrastructure ensures integrity and confidentiality of all trade information. Certification is the first step in secure trading. Prospective traders download the UNTPDC's 100% Pure Java-based applet and use it to provide the UN with reference data about their banking, trading, and services. After the UN certifies it, the company uses the Java applet residing in their standard Internet browser to access the network. Similarly, the smart card project of the UNCTAD is facilitating the payment flow in international trade. As discussed in the previous section, the first and second level smart cards allow secure ETO, confidentiality, payment information integrity, authentication, etc.

The Secure Electronic Authentication Link (SEAL) project and concept were developed by the United Nations Trade Point Development Center. Its smart card project facilitates payment flows in international trade. The first level smart card allows users to automatically authenticate their user profile to the SEAL and secure electronic trading opportunity (ETO) on the GTPN. The second level smart card allows confidentiality of information, payment information integrity, cardholder account authentication, merchant authentication, and interoperability with the ETO system on the Internet and the GTPN.

Most of the trade points are, however, located in big cities and many less developed countries are still deprived of the services of trade points. GTPNet sites are also vast and not well organized (Lehrer, 2003). A recent international forum on ways to improve the Trade Point programs for SMEs came up with many suggestions[5] including the necessity of encouraging and assisting non-exporting companies possibly by even creating special e-commerce programs open to "new exporters" only; providing guidance for the management; providing translation services so that import-export business can be conducted in multiple languages; and providing convenient online payment mechanisms.

LINCOS: Little Intelligent Communities (http://www.lincos.net/)

The LINCOS initiative was developed jointly by the Fundación Costa Rica para el Desarrollo Sostenible, the Media Lab at the MIT, and the Instituto Tecnológico de Costa Rica in 1998 (Saxe et al., 2000). LINCOS selected Hewlett Packard's E-inclusion as the model. It has alliances with over ten academic institutions, and at least ten technology companies (United Nations, 2000). Initially, LINCOS used recycled shipping containers to house computers, peripherals and generators. Each unit comprised of five computers and other facilities to provide a broad range of services including Internet access, health, education, banking, government services, electronic trading, technical support for SMEs, telecom and Information center, video conference and entertainment, forest, soil and water analysis, etc. The ultimate goal of the project is to achieve sustainability by allowing each community within the network to make decisions related to the technologies and advance them independently[6].

Each unit is satellite operated and solar power enabled and can operate independently of traditional infrastructures. The satellite dish antennae link them to any telecom network or master station as needed. Eggers and Siefken (2000) comment on LINCOS' effctiveness:

LINCOS units can be taken anywhere—mountain, jungle or village—and make Web-browsing, telephony and e-mail available even in the most remote spot. Equipped with their own generators, these units need neither external energy nor communication cables; they connect directly via satellite.

The units are installed in a container equipped with five computers. Additional LINCOS units are to be provided if more computers need to be added. The investment for each

unit was as high as $85,000 U.S. in the Dominican Republic and $50,000 U.S. in San Marcos de Terazu, Costa Rica (Proenza, 2001), but is expected to decrease with the increase in production (United Nations, 2000). In 2000, LINCOS won ALCATEL's first place prize for Technological Innovation in Latin America. Latin America participated in the competition[7] with over 30 projects.

The prototype LINCOS sites, which have already been deployed in Costa Rica, are providing several benefits. For instance, coffee growers use LINCOS sites to find the best prices in the world as well as next week's weather. Thanks to LINCOS' assistance, Costa Rican coffee farmers have also created Web pages and learned how to request budgets for buying equipment and register their trademarks (Amighetti and Nicholas, 2003).

While a container has had obvious advantages for carrying the equipments to the remote communities, the "container concept" was later abandoned because it created the concept of "temporariness"; the target users did not regard it as "rooted in the community" and tended to develop the feeling that the project was brought to the community in a "top-down" manner as a wrapped-up "development package"[8]. A second drawback of LINCOS is that it has attracted relatively rich people instead of helping the poor (Amighetti and Nicholas, 2003). According to the World Bank's Charles Kenny, LINCOS' have had mixed results because of the lack of sufficient interest of the target audience. He argues: "Poor people don't seem to think that the Internet is the answer to all their problems" (Rich, 2003, p. 93).

FUTURE TRENDS

Since the global digital divide is the result of fundamental economic, political and social gaps between the developed and the developing world, such a divide is likely to remain in place (Guillén & Suárez, 2005). In the absence of appropriate policy measures, it is likely that the global digital divide may become even wider (Dholakia and Kshetri, 2003; Economist.com, 2000). Policy measures directed at making appropriate networks available to the digitally excluded populations at reasonable costs, however, could bridge the gap or at least decrease the rate at which it widens.

The nature of the 'divide', which entails several gaps (UN Chronicle, 2003) appears to be changing rapidly. While some poor-friendly technologies such as cellular phones (Kshetri, Dholakia and Schiopu, 2006) and open source software (Kshetri, 2004) are bridging the basic form of digital divide in terms of access to technology, other forms of digital divide are widening and new forms are rapidly emerging. Evidence indicates that the digital divide has shifted from basic to advanced communications and more generally from quantity to quality (World Telecommunication Development Report, 2002).

CONCLUSION

Rapidly dropping costs of ICTs, developments of user-friendly software and interfaces, and versatility of the Internet offer the potential for leapfrogging many of the development obstacles. Civil society, governments, and entrepreneurs of developing and industrialized countries can take actions to bridge the digital divide by targeting highly excluded communities, by designing appropriate combination of new and old technologies, and by setting projects in the context of a longer term plan to extend the benefits more widely.

Some companies have already taken exemplary measures against the global digital divide. For instance, in May 2006, Intel announced its plan to invest a billion dollars to help provide access to technology and educational resources in developing countries (Clark, 2006).

Although, the networks discussed in this article are helping to bridge the digital divide between the developing and developed countries to some extent, most of them are not yet able to reach really excluded populations in developing countries. For instance, mainly rich people are benefiting from LINCOS. Similarly, there are no trade point programs in small villages of developing countries.

There is a huge untapped market for modern ICTs in developing countries if the services are *affordable* and *appropriate* to the target population and e-business companies need not provide their services in philanthropic ways. Lyle Hurst, director of HP e-inclusion, a partner of LINCOS, says that the mobile digital community centers will be a significant market opportunity for all involved. To exploit the potential, comprehensive research on the needs of the digitally excluded population and on the most appropriate networks to satisfy such needs by using locally available expertise and resources is needed.

The experiences of LINCOS and GTPNet indicate that lack of awareness and interest with the target audience are the major drawbacks of the networks discussed in this chapter. National governments, international agencies and technology marketers are required to work together to educate the target users about the potential benefits of such networks.

REFERENCES

Amighetti, A., & Reader, N.(2003). Internet project for poor attracts rich. *The Christian Science Monitor*.

Anderton, A. (2003). Business of ope: A charity-funded project is training young Cambodians to capitalise on new technology. *Asia Inc.*, January, pp. 42-43.

Archibugi, D., & S. Iammarino. (2000). Innovation and globalisation: Evidence and implications. In F. L. Chesnair & G. S. Roberto (Eds.), *European integration and multinational corporations strategies* (pp. 95-120). Routledge: London.

Brooks, K. (2000). Pas de deux—the dance of digital design. *Design Management Journal, 12*(2), 10-15.

Carter, C., & Grieco, M. (2000). New deals, no wheels: Social exclusion, tele-options and electronic ontology. *Urban Studies, 37*(10), 1735-1748.

Clark, D. (2006). Intel Aims to Bridge Digital Divide, *Wall Street Journal—Eastern Edition*, May 2, *247*(102), p. B2.

Computer Economics. (2000). The global economy is not so global. *Internet & E-Business Strategies, 4*(4), 1-3.

Dholakia, N., & Kshetri, N. (2003). Electronic Architectures for Bridging the Global Digital Divide: A Comparative Assessment of E-Business Systems Designed to Reach the Global Poor. In Shi Nansi (Ed.), *Architectural issues of Web-enabled electronic business* (pp. 23-40). Hershey, PA: Idea Group Publishing.

Economist.com. *Falling through the net*. Retrieved September 23, 2000, from http://economist.com/surveys/displaystory.cfm?story_id=E1_PSVGQV

Edejer, T. (2000). Disseminating health information in developing countries: The role of the Internet. *British Medical Journal, 321,* 797-800.

Eggers, I., & Siefken,S. (2000). Four Countries Connect. *UN Chronicle XXXVII* (2). Retrieved from http://www.un.org/Pubs/chronicle/2000/issue2/0200p32.htm

Fast Company. (2005). *The Rising Stars,* January, p. 58.

Goodman, S. E., Press, L. I., Ruth, S. R., & Ruthowski, A. M. (1994). The global diffusion of the Internet: Patterns and problems. *Communications of the ACM, 37*(8), 27-31.

Guillén, M. F., & Suárez, S. L. (2005). Explaining the global digital divide: Economic, political and sociological drivers of cross-national Internet use. *Social Forces, 84*(2), 681-708.

Hedley, R. A. (1999). The information age: Apartheid, cultural imperialism, or global village? *Social Science Computer Review, 17*(1), 78-87.

Helm, B., & Manjeet K. (2006). Life on the Web's factory floor, *Business Week*, May 22, 3985, p. 70.

ITU. (2001).The Internet: Challenges, Opportunities and Prospects. *World Telecommunication Day*. Retrieved on from http://www.itu.int/newsroom/wtd/2001/ExecutiveSummary.html

Kshetri , N. (2004). Economics of Linux adoption in developing countries. *IEEE Software, 21*(1), 74-81.

Kshetri, N., Dholakia, N., & Schiopu, A. (2006). *Is the cellular technology bridging the global digital divide?* Working Paper, Department of Business Administration, The University of North Carolina at Greensboro.

Lehrer, B. (2003). *Finding business opportunities on the Internet*. Retrieved from http://www.fita.org/aotm/tbird.html

Nunberg, G. (2000). Will the Internet always speak English? *The American Prospect*, March 27-April 10, 40-43.

Reed, M. (2001). Project Outlines Different Path to Development, *Cambodia Daily*. Retrieved from http://www.digitaldividedata.com/Press/cambodia_daily.htm

Rich, J. L . (2003). Not-so-simple solution. *Foreign Policy*. Nov/Dec, p. 93.

Saxe, E. B. (2000). *Taskforce on bridging the digital divide through education*. Retrieved from http://www.worldbank.org/edinvest/lincos.htm

Shih, J. (2003). At foreign firm, only Cambodia's abject need apply. *The Christian Science Monitor*, January 8, p. 1.

St. John, C. (2004). The Humanitarian Divide. *Stanford social innovation review, 1*(4), 52-53.

Tedeschi, B. (2003). Sensing economic opportunities, many developing nations are laying the groundwork for online commerce, *New York Times*. November 24, p. C7.

The Associated Press (2001). Cambodia Eyes Global Tech Industry. Retrieved from http://www.digitaldividedata.com/Press/associated_press.htm

UN Chronicle (2003). *We are embarked on an endeavour that transcends technology*. Dec 2003-Feb2004, *40*(4), p. 4.

UNCTAD (1998c). *Trade Point Review*. Geneva: United Nations Conference on Trade and Development. Retrieved from http://www.sdnp.undp.org/mirrors/lc/pan/untpdc/gtpnet/tpreview/%20The%20current%20status%20of%20Trade%20Points

UNDP. Human Development Report (2000). *United Nations Development Program*, New York, 2001. Retrieved from http://www.undp.org/hdr2001/completenew.pdf

United Nations. (2000). *Report of the high-level panel of experts on information and communication technology, general assembly, economic and social council*. Retrieved May 22, from http://www.un.org/documents/ga/docs/55/a5575.pdf.

UNCTAD (2006). The Digital Divide Report: ICT Diffusion Index. *2005 United Nations Conference on Trade and Development*, Retrieved from http://www.unctad.org/Templates/webflyer.asp?docid=6994&intItemID=2068&lang=1&mode=highlights

Wang, A. (1999). Verio Expands Global Reach of E-Commerce, *E-Commerce Times.* Retrieved from http://www.ecommercetimes.com/perl/story/143.html

World Telecommunication Development Report (2002). Reinventing Telecoms' & Trends in Telecommunication Reform 2002, *Effective Regulation*. Retrieved from http://www.itu.int/newsarchive/wtdc2002/backgrounder.html

KEY TERMS

Digital Divide: Refers to the fact that people in developing countries use modern ICTs less than those in developed countries.

Gross Domestic Product (GDP): Is the sum of the total value of consumption expenditure, total value of investment expenditure, and government purchases of goods and services.

High-Income Countries: Have per capita income $9,266 U.S. or more (in 2001).

Information and Communications Technologies (ICTs): Are technologies that facilitate the capturing, processing, storage, and transfer of information.

Internet: Is the "global information system that (1) is logically linked together by a globally unique address space based on the Internet Protocol (IP) or its subsequent extensions/follow-ons; (2) is able to support communications using the Transmission Control Protocol/Internet Protocol (TCP/IP) suite or its subsequent extensions/follow-ons, and/or other IP-compatible protocols; and (3) provides, uses or makes accessible, either publicly or privately, high level services layered on the communications and related infrastructure described herein" (The Federal Networking Council definition).

Low-Income Countries: Have per capita income less than $875 U.S. (in 2001).

PPP (Purchasing Power Parity): Is a rate of exchange that accounts for price differences across countries, allowing international comparisons of real output and incomes.

ENDNOTES

1 See How Big Is The Internet?, How Fast Is The Internet Growing?, http://www.metamend.com/internet-growth.html

2 See http://www.digitaldividedata.com/Press/MIT_Sloan_Cambodians.htm

3 See Introductory word by the WTPF President (May 23, 2003) at http://www.wtpfed.com/newsite/index1.php#.

4 ETOs that are less than eight days old.

5 These suggestions are condensed from UNCTAD/WTO email discussions on SMEs, reported at http://www.intracen.org/e_discuss/sme/welcome.htm (accessed on August 27, 2003).

6 See the details of the project at: http://projects.takingitglobal.org/lincos

7 See http://www.lincos.net/webpages/english/acerca/historia.html.

8 See "Assessing ICT efforts in marginalized regions... LINCOS-Dominican Republic", http://www.developmentgateway.org/node/603248/browser/?&sort_by=title

Global Software Team and Inexperienced Software Team

Kim Man Lui
The Hong Kong Polytechnic University, Hong Kong

Keith C. C. Chan
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Given that the number of qualified programmers cannot be increased drastically and rapidly, software managers in most parts of the world will likely have to live with a human resources shortage in this area for some time. One way of dealing with this shortage is to form global software teams in which members are recruited from all over the world and software is developed in a distributed manner. Forming such a global software teams can have many advantages. In addition to alleviating the problems caused by scarcity of human resources, programmers on a global team would be free to work without being confined by physical location.

Although forming global software teams may increase the size of the pool of programmers that can be recruited, both team quality and software quality are issues of great concern. Some software companies would prefer to establish a global software team with software programmers in developing countries, such as China, Poland, and South Africa (Sanford, 2003). Given the tremendous salary gap between skilled and unskilled developers or between developed and developing countries, it is not difficult to see that maintaining a team with a proportion of less experienced members significantly reduces running expenses (Figure 1). On the other hand, however, it would present the problem of managing inexperienced programmers.

This chatper shares our experience of managing inexperienced software teams in China. To simplify our discussion, we deal separately with the two topics of inexperienced software teams and global software teams. However, it should be noted that a global software team can be composed of both inexperienced and experienced software subteams. We categorize the problems in these two types of software teams which will help software managers learn more how to manage the two types of software teams.

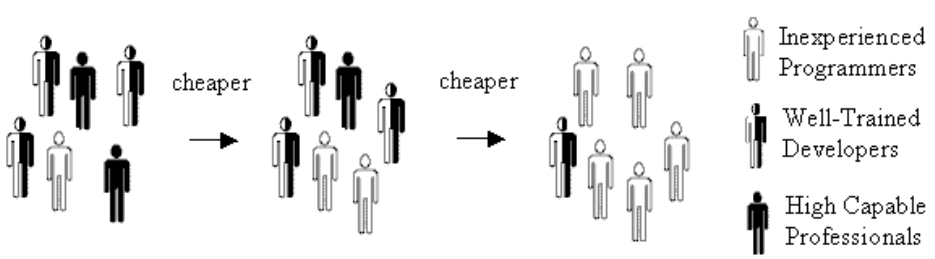
BACKGROUND

This section reviews real cases that have driven the formation of an inexperienced software team and a global software team. The motivation behind managerial decisions to build such teams is both financial and environmental.

Discovering Developing Areas: Inexperienced Software Team

Active rural industrialization involves manufacturing plants moving from more developed regions to less developed ones so as to exploit the lower costs of land, labor, and distribution channels (Otsuka, 2001). In less developed regions it is an

Figure 1. In the software world, the proportion of highly professional to less experienced teams may fall as companies operate under the constraints of tighter cash flows and for reasons of cost replace more experienced programmers with junior programmers and seek to avoid the costs associated with the professional development of senior programmers.



easy matter to recruit labor for manufacturing, but these plants also require management information systems (MIS) and it is not at all easy to find and recruit the IT professionals that are required to develop an integrated, customized MIS.

In developing countries, the demand for IT professionals in larger cities is currently so high that it is almost impossible for any manufacturing plant in a rural area to recruit people. As a result, programmers in poor rural areas are usually inexperienced. Even though the alternative of employing software expatriates might sound reasonable, it is feasible to recruit one or two highly qualified programmers from developed regions but it is not practical to recruit a team of them. Instead of in-house development, we might evaluate a third-party solution. The additional expenses incurred in purchasing vendor products, in consultancy services, annual maintenance, version upgrading, training, traveling, and so forth, outweigh the savings that are sought by setting up in the country in the first place.

In less developed areas, many programmers do not receive proper training in computing. In addition, the turnover rate is typically high. As soon as they have received even a little training, many workers will seek a job with better career prospects in a more developed city. The result is that the project manager always has to work with inexperienced programmers and in a constant mood of crisis management as the high turnover rate is aggravated by resignations without notice as people tender their resignation and leave on the same day. Clearly, the process of handing over work is unsatisfactory, teams are constantly under-staffed and the working environment suffers, making work elsewhere an even more attractive option.

One may suggest that educating inexperienced people or allocating suitable jobs according to an individual's ability should fulfil the same purposes. However, when the knowledge and experience of staff members is not aligned with the tasks assigned, the learning curve can be steep and long (Amrine, Ritchey, Moodie & Kmec, 1993). Nevertheless, when a staff member becomes well trained in some less developed regions in China, or in a small company in Denmark, for example, the determination of the staff member to look for better job prospects elsewhere will become stronger. Training, therefore, does not provide a promising solution in this case. In contrast with what we might find in well-developed regions, in China, staff who are provided with certified professional programs will leave a company or a less developed region even sooner. Senior managers are disturbed by this phenomenon and say that they are always training another company's staff. The idea of allocating developers according to their skill set is not feasible when all team members are inexperienced. Human resource allocation can therefore be implemented only to a limited extent. Better knowledge management, rather than adopting conventional principles, is required.

Around the Clock: Global Team

A small but ambitious company selling weight-loss and nutritional products, which was headquartered in New York, had a number of small offices of 40 employees in different parts of the world. Being close to customers is always a key to business success.

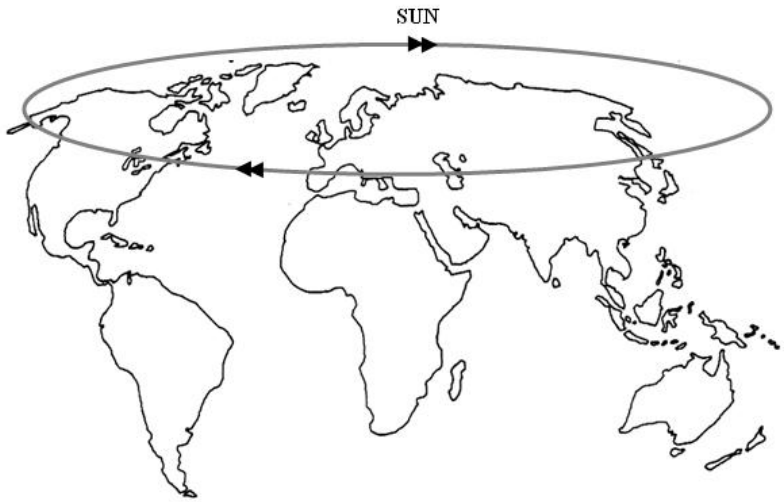
One to two staff members in each office had the task of providing IT support. When the MIS system needed to be modified to meet requirements for local processing, requests for modifications would be sent to the head office. The result was that more resources were required at the head office to provide ongoing support to the branches. Although a larger software team was thus required at the head office, IT staff at the branches might have time to spare.

The load-balancing problem got worse when the number of branches increased. The question, naturally, was to decide if it was possible to link people to establish a global software team. The team in each site then plays a role more or less as distributed agents following a communication scheme from a coordination agent.

A global software team can even be formed locally, if the team is set up in different locations within the same country or in nearby countries or regions. This means that there may not be much difference in time zones and cultures. In such situations, the term "multisite" or "distributed" software team can be used more generally to describe a software system developed by teams that are physically separate from each other in different cities of a country or in different countries. A multisite team in nearby time zones can be managed with less complexity and fewer challenges than a global team and but has more limited service hours. Clients on the other side of the world want a reply inside their own local business hours but if all teams are in the same time zone, it becomes hard to respond promptly outside office hours. In any case, the management framework required for a global or multisite team should be very similar. To further explore around-the-clock development (see Figure 2) and global development, we realize the intrinsic difference is how synchronization of work-in-progress proceeds. We concluded that challenges of managing around-the-clock tasking widely cover managerial and technical problems of non-around-the-clock global software development.

Around-the-clock development exploits time zone differences as a way to improve time-to-market . But, there has not been a model for this kind of global software development (Carmel, 1999; Karolak, 1998). It is easiest to manage a global software team with less strict synchronization among different sites. In this case, a team that is waiting for a result from another site could work on other tasks for the same project. However, in around-the-clock development, work-in-progress and communications are rigidly synchronized and the progress of a team is tied to the progress of

Figure 2. Around-the-clock development (also called around-the-sun development)—A global team with one site in Asia, one in Europe, and one in North America maximizes time use by working around the clock



another team that posts deliverables to another team at the end of the day.

A traditional framework does not provide a sound solution (McMahon, 2001). A less strict model is available if there are recognized steps towards a standard solution because then the work of one team can be continued by another team at the point where they stopped. Work is thus forwarded from team to team and time zone to time zone until it has been finished.

The success of around-the-clock development thus greatly depends on what type of application we build and what methodology we use to manage a global team. It requires standard approaches to standard problems, for example, generic systems for a particular application such as a database or a Web system. The bad news is that this approach is unable to cope with the totality of any new type of project. The good news is that a high portion of information projects nowadays are related to familiar commercial database applications and Web applications.

HOW TWO TEAMS ARE RELATED

At a glance, an inexperienced development team and a global software team conjure up two totally different pictures. However, some problems can be dealt with by the same common solution. We start with the examination of commonalities between them, shown in Table 1.

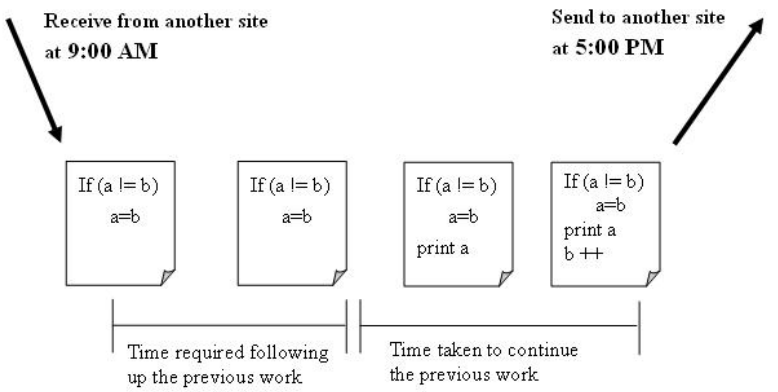
Turnover

In less developed regions in a developing country, programmers are usually inexperienced. As soon as they gain experience and receive some training, many of them will look for a new job in the more developed cities either in the same country or in other countries (Morris, 1995). As a result, the rate of staff turnover is high. This situation is aggravated by the fact that many people who tender their resignations prefer to leave immediately, making work effective work

Table 1. Characteristics of software development teams: An inexperienced team and a global team

	An Inexperienced Software Team in a Less Developed Region	A Global Software Team using Around-the-clock Development
1. Turnover	High Personnel Turnover	High Task Turnover
2. Knowledge Management	Weak IT knowledge	Varied knowledge and skills at each site
3. Communication	Lack of IT project experience for effective team communication	Varied culture at each site

Figure 3. Anatomy of the process of around-the-clock development



handovers either difficult or impossible. It can be seen from this how personnel turnover can severely affect the rate of task turnover.

In some respects, the use of global teams for around-the-clock software development can produce similar problems as a team following up the work done by another team cannnot communicate with the predecessor team in real time as that team is asleep on the other side of the globe.

The basic challenge of around-the-clock work here is the time needed to follow up in order to continue the task delivered electronically from another site, and the remaining time for working and then relaying to another site at sunset, depicted in Figure 3. There are two unconventional problems. The subprocess in a site must not be interrupted, as this will also suspend the whole process. In addition, the sum of the working hours to follow up and to continue is just one day. Suppose each site works for eight hours a day. The efficiency of around-the-clock development at one site will be as follows:

$$\text{Efficiency} = \frac{8\,\textit{hours} - \text{Time used to follow up previous works}}{8\,\text{hours}} \times 100\%$$

Around-the-clock software development has problems similar to those associated with a high personnel turnover rate. Neither guarantees that the expected outcome is achievable. And both strongly require a very quick job handover without face-to-face, lengthy explanations from previous developers.

Personnel turnover and task turnover can unexpectedly interrupt or sometimes even halt our project activities. Many software teams seem to be taking a very ad hoc approach but their project activities are always progressing and these teams normally complete projects on time and on schedule. In these cases, the rhythms of running software projects (referred to as software development rhythms), rather than software methodologies alone, provide a deeper understanding of what and how a software team should plan and execute their projects (Lui & Chan, 2008).

Knowledge Management

In developing a database application, a software team may encounter many kinds of technical problems that require different skill sets, such as inserting records into a database, deleting records from a database, updating those records, controlling data integrity, controlling transactions (Taylor, 2003). In order to do programming, a software team should be equipped with the minimum expertise that allows the team to complete part, if not all, programming jobs. Developers below that level could do nothing by themselves. Figure 4 illustrates this idea.

Our goal is to lower the line of minimum expertise—but how? Let us look at an example: suppose you have a group of people, say those with a learning disability, who are able to count numbers but do not understand addition. If we want them to do the addition without a calculator, the best way would be to teach them the calculation. This would be the minimum expertise for this problem. Still, the learning curve may be long. *(If they are your employees, your boss cannot help wondering why you would hire these people and make the office a learning center. Subsequently, you cannot help worrying about the full support that your boss previously committed to you.)* Another approach to getting the same work done is a mechanical method that asks the workers to

Figure 4. Minimum expertise for programmers

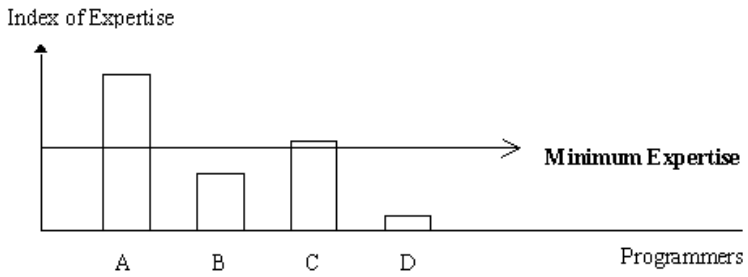


Figure 5. Lowering minimum expertise

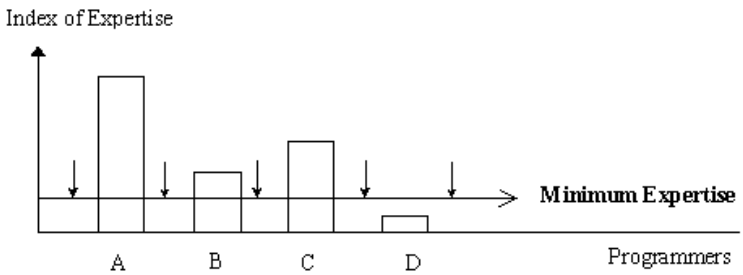
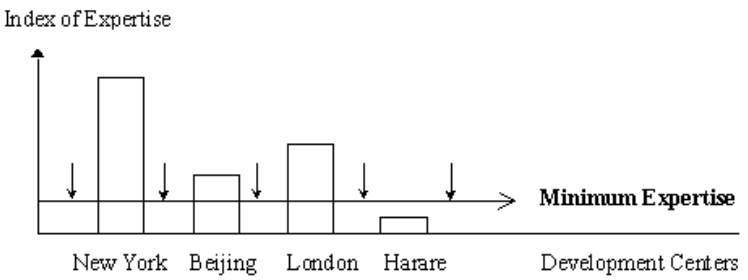


Figure 6. Minimum expertise for development center



follow a predefined mechanism for counting marbles. As for 4+3, the rule might be as follows:

(*Step 1*) Count four marbles and put them aside.
(*Step 2*) Count three marbles and put them aside as in Step 1.
(*Step 3*) Count all the marbles together.

The minimum expertise is now counting marbles and putting counted marbles aside. Obviously, learning to add numbers first requires understanding how to count numbers. Putting counted marbles together means the workers follow the steps. In short, the minimum expertise is lowered to some degree, as shown in Figure 5, although the approach might not appear to be intelligent, management would be happy with it as employees are able to start working and get the work done in a predictable way and time.

The problem of different technical skills among team members parallels the situation of different knowledge levels among many sites in the global software team, as shown in Figure 6. For example, in Figure 6, programmers in Beijing may use techniques in a program that developers in Harare would spend a whole day figuring out, whereas developers in London may follow up the work quickly and can continue to work.

This may happen that all of the teams spend more time on some of project tasks tremendously owing to technical skills unmatched to a point that requires completing the tasks or technical skills among members being diversified to a point that members do not understand (or even complain) the ways other members have done their jobs.

COMMUNICATION

Communication does not only require the application of knowledge of surface features such as words; it also requires knowledge of the cognitive and cultural practices of different groups and disciplines. This means that on the one hand, global communications between software programmers is aided by the fact that they are all trained in the same field, even if some of the programmers are not as well trained as others. On the other hand, regional cultural differences can also mean there are many very different assumptions about matters of communication, such as who is responsible for clarity in a communication (is it my job to explain or yours to understand?), what can be assumed to be already understood, in what order ideas should be presented, whether authorities can be questioned by subordinates. Clearly, communication can become a central issue in the effectiveness of global teams and so it is important that global teams work within a clear corporate communicative culture. This can be especially critical in software projects, where project progress is intangible and must be constantly assessed and reported in ways that rely very much on communication.

FUTURE TRENDS

Globalism and the expansion of the web will continue to underwrite the global demand for IT professionals (Sangwan et al., 2007). There is no doubt that we will have to either recruit and train less experienced programmers for our software teams or build a global software team so as to acquire human assets not limited to a particular job market. Yet other innovative approaches will also be adopted.

Recently, agile software development (Cockburn, 2007) has provided an alternative way to alleviate some of the software development problems addressed previously. For example, pair programming helps to minimize the impact on personal turnover as every piece of production code is now written by two programmers (Lui & Chan, 2006). Putting the focus on compiling working software daily can also facilitate effective communications not only between customers and programmers but between different programmers as well. Although many agile practices (e.g., daily stand-up meetings and pair programming) appears to be easily adopted for colocated small teams, software practitioners are trying to take the agile philosophy into distributed environments by extending agile practices (Finden, 2006). Unfortunately, some colocated agile software teams still fail to manage their projects (Stephens & Rosenberg, 2003), needless to say distributed agile teams in distributed envrionments; such failures are not due to any difficulty of learning and adopting a single agile practice. The real challenging issue is to combine a number of agile practices (with existing software practices already adopted by a software team) in such a way that the team is able to integration with software practices. This has been referred to as *software development rhythms*. For example, software team members who have been used to writing programs with their private instant massagers on will take more time to get used to pair programming. This however does not mean that it takes time to go agile. It only means that we have to understand the relationships between our team and newly adopted practices. Going back to our example, the team may try to adopt pair-solo programming (Lui & Chan, 2008) to see if it provides an effective rhythm for the team.

CONCLUSION

We have addressed some interesting problems relating to inexperienced software teams and global software teams. Both require new managerial, technical, and social approaches and underline the fact that software projects fail not because of the failure of software methodologies or software teams but because of the failure of software development rhythms.

REFERENCES

Amrine, H. T., Ritchey, J. A., Moodie, C. L., & Kmec, J. F. (1993). *Manufacturing organization and management*. Englewood Cliffs, NJ: Prentice Hall.

Carmel, E. (1999). *Global software teams: Collaborating across borders and time zones*. Upper Saddle River, NJ: Prentice Hall.

Cockburn, A. (2007). *Agile software development: The cooperative game* (2nd ed.). Upper Saddle River, NJ: Addison-Wesley.

Finden, A. (2006). Achieving agility in globally distributed software development. *Agile Journal*. Retrieved June 16, 2008, from http://www.agilejournal.com/articles/articles/achieving-agility-in-globally-distributed-software-development.html

Karolak, D. W. (1998). *Global software development: Managing virtual teams and environments*. **Los Alamitos, CA:** IEEE Computer Society.

Lui, K. M. & Chan, K. C. C. (2006). Pair programming productivity: Novice-novice vs. expert-expert. *International Journal of Human Computer Studies, 64*, 915-925.

Lui, K. M. & Chan, K. C. C. (2008). *Software development rhythms: Harmonizing agile practices for synergy*. John Wiley and Sons.

McMahon, P. E. (2001). *Virtual project management: Software solutions for today and the future*. Boca Raton, FL: St. Lucie Press.

Morris, S. G. (1995). *Turnover among professionals: The role of person-culture fit and mentoring*. **Boulder, CO:** University of Denver.

Otsuka, K. (2001). Book reviews: Growth and development from an evolutionary perspective. *Journal of Development Economics, 65*, 237–241.

Rischpater, R. (2001). *Palm enterprise applications: A Wiley tech brief*. New York: John Wiley and Sons.

Sangwan, R. et al. (2007). *Global software development handbook*. Boca Raton, FL: Auerbach Publications.

Sanford, J. E. (2003). *Developing countries: Definitions, concepts and comparisons*. New York: Nova Science.

Sapaty, P. (1999). *Mobile processing in distributed and open environments*. **New York:** John Wiley & Sons.

Stephens, M. & Rosenberg, D. (2003). *Extreme programming refactored: The case against XP*. Berlin: Springer, Apress.

Taylor, A. **(2003).** *JDBC: Database programming with J2EE*. Upper Saddle River, NJ: Prentice Hall.

KEY TERMS

Agile Software Development: Agile software development is a conceptual framework for software engineering that promotes development iterations throughout the life-cycle of the project.

Around-the-Clock Development: A software development style in which software teams that are geographically distributed make use of time zones to develop software.

Around-the-Sun Development: See Around-the-Clock-Development.

Efficiency of Around-the-Clock Development: This is an index used to indicate a ratio between time required to follow up the previous work and time spent advancing the work.

Global Software Team: Software teams located in different countries collaborate as a single team for a clear project objective.

Inexperienced Software Team: Most members of a software team are graduates, or inexperienced in disciplined software development.

Multisite Software Team: Software teams located in different cities and/or in different countries collaborate as a single team for a clear project objective.

Time-to-Market: A work product to the market as quickly as possible, before competitors, in order to get a larger market share or to begin earlier cost recovery.

Software Development Rhythms: Software development rhythms (SDR) respects and builds upon the inherent flexibility of agile practices (e.g., pair programming, refactoring, test-driven development, stand-up meeting, plagiarism programming, etc.), focusing on understanding the "why and when" of the effective application of practice-move-practice or activity-move-activity

Globalization of Consumer E-Commerce

Daniel Brandon, Jr.
Christian Brothers University, USA

INTRODUCTION

This article reviews globalization aspects of "business to consumer" (B2C) electronic commerce. According to *Computerworld*, "Globalization is the marketing and selling of a product outside a company's home country. To successfully do that on the Internet, a company needs to *localize* – make its Web site linguistically, culturally, and in all other ways accessible to customers outside its home territory" (Brandon, 2001). This overview describes the key issues in the globalization of electronic commerce; for more detail, see the full book chapter (Brandon, 2002).

BACKGROUND

"Ever since the end of the Cold War, the world has been rushing toward ever-higher levels of national convergence, with capital markets, business regulation, trade policies, and the like becoming similar" (Moschella, 2000). The value of cross-border mergers grew six-fold from 1991 to 1998 from U.S. $85 billion to $558 billion. The world has not witnessed such a dramatic change in business since the Industrial Revolution (Korper & Ellis, 2000). More than 95% of the world population lives outside of the U.S., and for most countries, the majority of their potential market for goods and services is outside of their borders. Over 60% of the world's *online* population resides outside of the United States (IW, 2000).

Today, the majority of Fortune's 100's Web sites are available only in English (Betts, 2000). In our rush to get on the WWW, we sometimes forget that WW is for "World Wide" (Giebel, 1999). Today's average Web site gets 30% of its traffic from foreign visitors, yet only 1% of small and mid-size American businesses export overseas (Grossman, 2000b).

KEY ISSUES

"Localization" (shortened to L12N in Internet terms) considers five global dimensions: geographic, functional, regulatory, cultural, and economic (Bean, 2000). We shall overview each of these somewhat overlapping and interrelated issues in these groupings: language, cultural, legal, payment/currency, dates/units, and logistics.

Language

According to IDC, by 2005, more than 70% of the one billion Web users around the world will be non-English speakers (Wonnacott, 2001). For the immediate future, most of the Internet community will still understand English, but overall English is the native language to only 8% of the world. Most users in foreign countries prefer content in their own language; for example, 75% of users in China and Korea have such a preference (Ferranti, 1999). It was found that visitors spend twice as long, and are three times more likely to buy from a site presented in their native language (Schwartz, 2000). We also have to take into account differing dialects that are used across various countries speaking a specific language. The combination of language and dialect is called a "locale".

One can convert Web pages by hiring a translator or using a computer-based translation product or service. Hiring a translator will provide the best localization but is more costly than the automatic methods. Translators can easily be found in the Aquarius directory (http://aquarius.net) or Glen's Guide (www.gleensguide.com). It is best to use a translator that "lives" in the local region; if a translator has not lived in a region for a decade, he has missed 10 years of the local culture. There are also many companies that provide translation services such as: Aradco, VSI, eTranslate, Idiom, iLanguage, WorldPoint, and others. The cost of these services is about 25 cents per word per language (Brandon, 2002). Automatic translation software is another option, but it is still in its infancy (Reed, 2000). Some popular software products for translation are: www.e-ling.com, www.lhs.com, and www.systransoft.com. The automatically-translated text typically does not convey the meaning of the original text.

There are several Web sites which provide free translation services such as: http://babelfish.altavista.com, *http://translator.go.com*, and www.freetranslation.com. For example, Figure 1 shows the "BabelFish" Web site where we are requesting a translation of an English sentence into Spanish. Figure 2 shows the translation results.

Another alternative, although certainly not optimal, is to provide a link on your English Web page for these free services so that visitors can translate your content themselves. Figure 3 shows a portion of the CBU School of Business English version Web site.

The automatic Spanish translated version (using BabelFish) is shown in Figure 4. Note that automatic version, while syntactically and grammatically correct, does not

Figure 1.

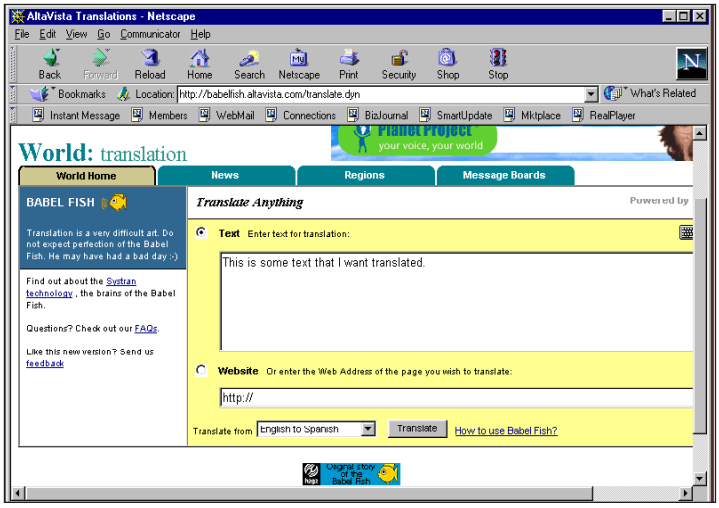
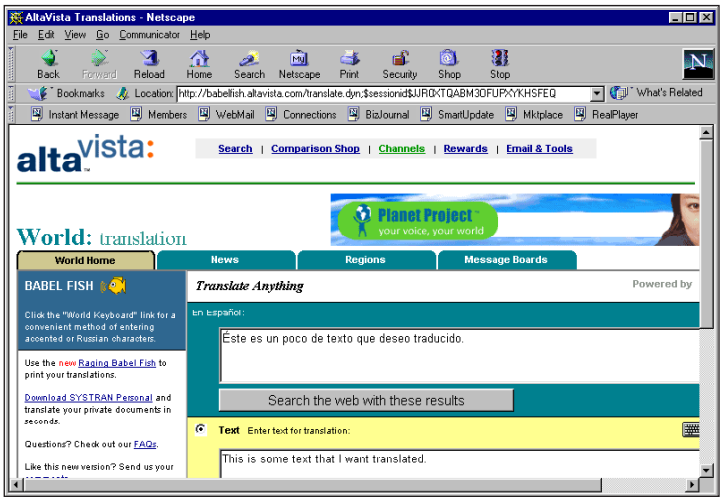


Figure 2.



convey the exact intended meaning to most of the titles and phrases.

Figure 5 is the version converted by a translator manually, and even though you may not speak Spanish, you can see the extent of the differences (Brandon, 2000). Shown in Figure 6 is the home page for FedEx (www.fedex.com). One can select from over 200 countries for specific language and content.

Cultural

Creating an effective foreign Web site involves much more than just a good language translation. Not only do languages differ in other countries but semantics (the meaning of words and phrases) and cultural persuasions in a number of key areas are different. "Sensitivity to culture and national distinction will separate success from failure" (Sawhney & Mandai, 2000). To be effective, a Web site has not only to be understandable and efficient, but has to be culturally pleasing and inoffensive. To accomplish that, it may be necessary

Figure 3.

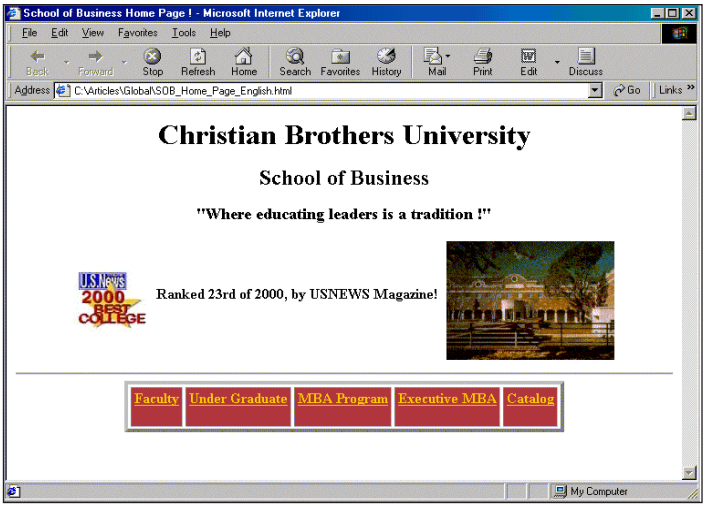
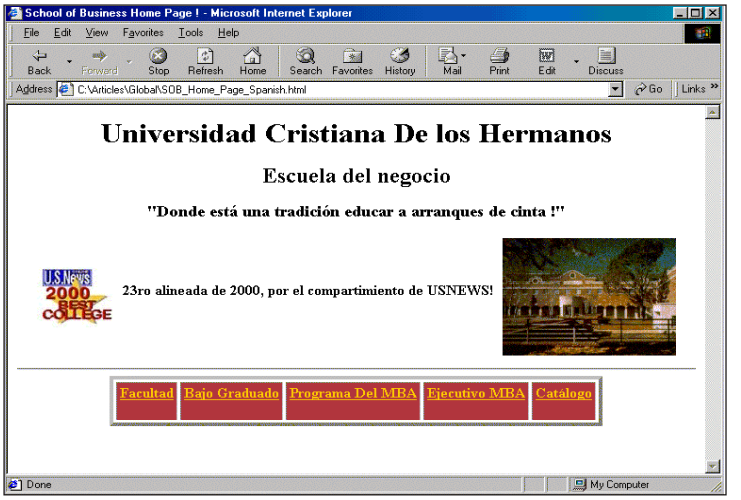


Figure 4.



that not only language be localized, but that content, layout, navigation, color, graphics, text/symbol size, and style may be different. Many companies have put forth global Web sites simply by translating the English into the targeted language, but then had to pull back and redesign the localized site due to cultural offenses.

A country's humor, symbols, idioms, and marketing concepts may not send the same messages to other countries in the world. Some areas of global disagreement to avoid are: equality of the sexes or races, body parts and sexuality, abortion, child labor and majority age, animal rights, nudity, guns, work hours and ethic, capital punishment, scientific theories, and religious particulars (Brandon, 2002).

Colors have symbolic and special meaning in most locals. Purple is a problem in many places; it symbolizes death in catholic Europe and prostitution in the Middle East. Euro Disney had to rework its European sites after the first version used too much purple. Overall blue is the most culturally accepted color (Brandon, 2001). It is also very important to respect other cultures "symbols" (heroes, icons, etc.) both positive and negative (swastika). One guide site is *Merriam Webster's Guide to International Business* (*www.bspage.com/address.html*).

Figure 5.

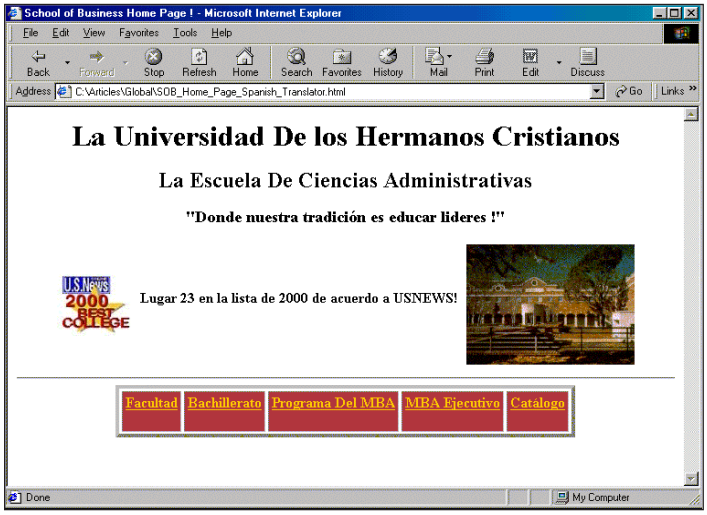
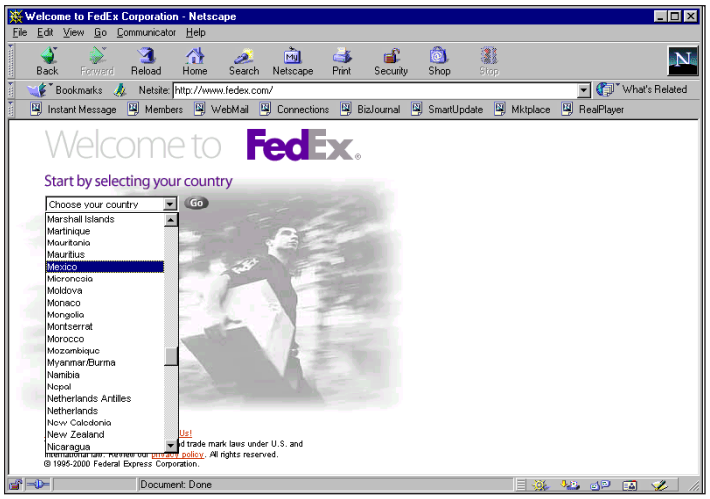


Figure 6.

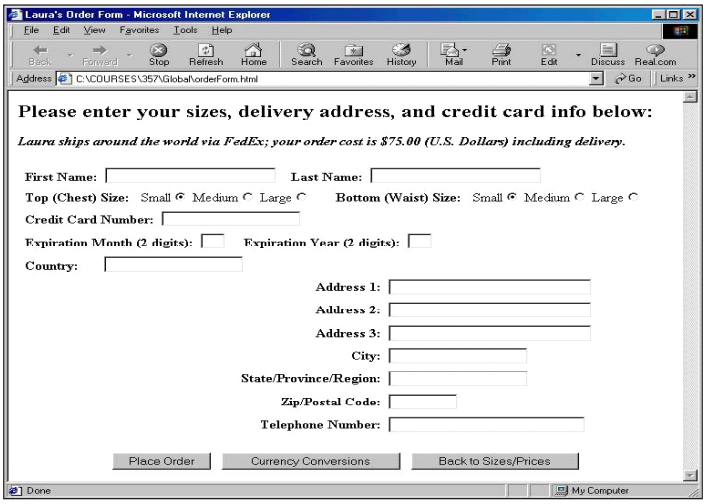


Legal

Recently French court's ruling that Yahoo must make auctions of Nazi memorabilia unavailable in France indicates how uncertain and risky international e-business can be. "The troubling aspect of this case is that different countries can say that content not even targeted at their population breaks the law" (Perrotta, 2000). With the Internet, it is not possible to know for sure where a user is logged in due to "IP tunneling" possibilities.

"Freedom" laws (such as the U.S. First Amendment) are not universal, and saying/printing some things can be illegal in some parts of the world. In the U.S., you can say what you like about "public figures" but not so in most of the rest of the world. Another legal issue concerns the privacy of personal data collected online. Many parts of the world have stricter laws than does the U.S., and U.S. companies have had judgments rendered against them in foreign courts. There are other areas that could cause legal problems, too. One is foreign advertising restrictions; for example, in Germany, one cannot directly compare your product with that of a competitor. In some other countries this comparison may not be illegal but may leave a bad taste.

Figure 7.



Payment and Currency

Nearly half of the U.S. Web sites refuse international orders because they are unable to process them (Grossman, 2000a). Foreign exchange rates vary daily so indicating that your prices are in your country's funds (exclusive of local taxes and custom duties) and using credit cards (so the credit card company does the conversion) is one way to deal with that issue. One can also link to a converter site (www.xe.net/ucc, www.oanda.com) or place a calculator on your page (www.xe.net/currency) as a utility for your customers or do your own conversions (see Figure 8, later).

However, credit cards are rare in Japan as is the use of checks. There, postal workers collect cash on delivery (CODs), and some companies send goods to brick and mortar places for consumers to pick up. In Germany, only 5% of Web users (second to U.S. in overall net usage) use credit cards. Eighty-eight percent of European merchants use invoice billing (with a long net payment due time). So while credit cards are a convenient and popular mechanism in the U.S., it is not so in the rest of the world. To complicate matters even further, there are many (and always changing) international sales taxes, value added taxes (VAT) in Europe, with different exempt items in each country. One approach to avoid all these problems is to use an escrow service such as Paymentech (www.paymentech.com) which now handles about three billion transactions a year.

Time-Date and Units of Measure

Dates are very important in e-commerce when being used for events such as: delivery dates, credit card expiration dates, product expire dates, etc. There is an international standard on dates (ISO 8601 Date Format), and even though you may not use it internally in your programs (for database operations and calculations), your Web display should be in the localized format. For example, the common U.S. format of 10/6/2000 is not uniformly understood; instead use Oct-6-2000. Major databases (i.e., Oracle) allow you to switch date formats per session or connection so the way a date is input (inserted into a table) or output (selected from a table) is automatically converted to the internal table representation of the date.

In the U.S., a 12-hour clock is common, except in U.S. military establishments. The rest of the world uses mostly a 24-hour clock, so it is best to display time in the 24-hour format. Of course, time zones will be different, so include your time zone along with the phone numbers for personal customer support. It is best to spell out the time zone in the native language. You could instead give your support time in GMT (Greenwich 2000 Standard) and use or link to www.timeanddate.com for a customizable world clock and calendar. In addition to dates and times, other units of measure will be different also. Only the U.S. and Canada still use the "English System"; the rest of the world is on the metric system now, even Britain.

"Addressing" a customer may be more involved; some foreign addresses may have longer and more address fields. There is a universal standard of sorts here called the "UPU" (universal address formats). Generally, it is of good advice including a country code (for validation of remaining fields), at least three address lines (40 characters each), city field (30 characters), a "state/province/region" field (20 characters), a postal code/zip field (10 characters), and a contact phone

number (20 characters). Figure 7 shows an order form using these specifications.

Logistics

Logistics involve both getting your products to the customer, as well as allowing the customer to return unwanted goods. Some parts of the world have relatively primitive transportation networks. In China, villages do not have postal service. Also, each locale typically has a set of customs and tariffs that you may need to add to the price of your goods. This "landed cost" of an order is the sum of the price of goods, shipping charges, insurance, duties/customs, VAT, and any import or export fees. You may need a "Shippers Export Declaration" depending on value and mode of transportation (www.census.gov/foreign-trade/www/correct.way.html) or other documents depending on countries and goods. As well as normal shipping insurances, you may need to consider export insurance (www.exim.gov). Of course, the language as well as logistic terminology varies; however, there is a standard set of international logistic acronyms ("incoterms" - *www.schenkerusa.com/incoterms.html*).

Many countries have *foreign import restrictions* and/or quotas. In addition, many countries have certain *export restrictions*. Japan has more than 200 trade laws and 17,000 regulations on imports (Pfenning, 2001). Today, 85% of U.S. companies do not ship to customers seeking delivery abroad, and the 15% that do ship ignore these compliance issues and push the responsibility of customs, restrictions, and payment onto their customers (Shen, 2000).

There are several ways to handle all these logistics issues. One is to use shipping companies that handle all these problems for you (at a nominal charge) such as FedEx (www.fedex.com) or UPS (www.ups.com). Another alternative is to use software or services that handle all these payment, custom, and restrictions issues by preparing the paperwork and calculating "landed costs"; one example can be found at www.mycustoms.com.

Still another alternative is to use a centralized distribution center in foreign regions to reduce shipping costs and eliminate some import taxes and tariffs (Tapper, 2000), either directly or with a partner. There are also total fulfillment providers such as: National Fulfillment Services, DupliSoft, Fill It, SubmitOrder, Equire, FedexLogistics, and so forth. These organizations not only handle delivery but also inventory, returns, customer service, and in some cases, Web ordering and payment.

FUTURE TRENDS

In the not too distant future, the Web will be everywhere, and by "everywhere", we mean not only in all of our electronic devices but everywhere in the world. Future trends in B2C e-commerce will involve this convergence of Web connectivity both in geographics, demographics, and electronics. For example in the future, automobiles are likely to have satellite and wi-fi Internet connections not just for navigation and information, but for marketing and sales also (in several languages). As one drives down a street, they will be able to receive live commercial messages from businesses as they physically approach the retail outlets, and these retail outlets could be outlets for international products as well.

CONCLUSION

Is globalization right for an organization? It can be very costly to build and maintain a foreign presence. We have identified and briefly discussed the key issues in this article. But there are many other issues that may affect your global e-commerce. "Building a global e-business calls for hosts of strategies that include partnering with or acquiring foreign companies, assembling sales and support operations, understanding new laws, languages, cultures, and implementing technology that can sustain a global endeavor" (Bacheldor, 2000). Many organizations are successful by using foreign partners such as: E-Steel, GlobalFoodExchange, and Office Depot. It has been said that the "Net brutally punishes latecomers" (Sawhney & Mandai, 2000), so it is essential to start planning the internationalization and localization of e-commerce now. Also remember the Web is a two-way street; foreign corporations will be coming after your customers soon!

REFERENCES

Bacheldor, B. (2000, May). Worldwide e-commerce: It's more than a Web site. *Information Week*.

Bean, J. (2000, March). A framework for globalization. *Enterprise Development*.

Betts, M. (2000, August). Global Web sites prove challenging. *Computerworld*.

Brandon, D. (2001). Localization of Web content. *15th Southeastern Small College, Computing Conference, 17*(1), Nashville, TN, November.

Brandon, D. (2002). Issues in the globalization of electronic commerce, In *Architectural issues of Web-enabled electronic business*. Hershey, PA: Idea Group Publishing.

Ferranti, M. (1999, October). From global to local. *Infoworld*.

Ferranti, M. (2000, November). Globalization tidal wave. *Infoworld*.

Giebel, T. (1999, November). Globalize your Web site. *PC Magazine*.

Grossman, W. (2000a, July). The outsiders. *Smart Business*.

Grossman, W. (2000b, October). Go global. *Smart Business*.

IW (staff). (2000, November 20). Weekly stats. *InternetWeek*.

Kiplinger, K. (2000, November). Globalization – alive & well. *Fidelity Outlook*.

Klee, K. (2001, March). Going global: Out ten tests can help you get started. *Forbes Small Business*.

Korper, S., & Ellis, J. (2000). *The e-commerce book, building the e-empire*. Academic Press

Moschella, D. (2000, December). Ten key IT challenges for the next 20 years. *Computerworld*.

Perrotta, T. (2000, July). Yahoo ruling exposes risks of being global. *InternetWorld*.

Pfenning, A. (2001, March 19). E-biz must chart international path. *InternetWeek*.

Reed, S. (2000, August). Want to limit the audience for your Web site? Keep it English only. *Infoworld*.

Sawhney, M., & Mandai, S. (2000, May). Go global. *Business*.

Schwartz, H. (2000, September). Going global. *WebTechniques*.

Shen, J. (2000, November). The commerce diplomats. *WebTechniques*.

Tapper, S. (2000, September). Is globalization right for you. *WebTechniques*.

Wonnacott, L. (2001, April). Going global may bring new opportunities for existing customers. *InfoWorld*.

KEY TERMS

Character Set: The set of symbols used to represent a language (alphabet, numerals, special symbols).

Encoding: The bit pattern to use for each symbol in a character set.

Export Restrictions: Restrictions on the type, quantity, or destination of goods that can be exported out of a country.

Globalization: The marketing and selling of a product outside a company's home country.

Import Restrictions: Restrictions on the type, quantity, or origin of goods that can be imported into a country.

Incoterms: A standard set of international logistic acronyms.

Landed Cost: The cost of an order including the price of goods, shipping charges, insurance, duties/customs, value added tax (VAT), and any import or export fees.

Locale: The combination of language and dialect.

Localize: Make a Web site linguistically, culturally, and in all other ways accessible to customers outside ones home territory.

Shippers Export Declaration: Documentation necessary to export goods outside one's home country.

UPU: Universal address formats.

VAT: Value added tax.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1293-1298, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Governance Structures for IT in the Health Care Industry

Reima Suomi
Turku School of Economics and Business Administration, Finland

INTRODUCTION

The pressures for the health care industry are well known and very similar in all developed countries (i.e., altering population, shortage of resources for staff and from taxpayers, higher sensitivity of the population for health issues, new and emerging diseases, etc.). Underdeveloped countries experience different problems, but they have the advantage of learning from the lessons and actions that developed countries underwent perhaps decades ago. On the other hand, many solutions also exist, but they all make the environment even more difficult to manage (i.e., possibilities of networking, booming medical and health-related research and knowledge produced by it, alternative caretaking solutions, new and expensive treatments and medicines, promises of biotechnology, etc.).

From the public authorities' points of view, the solution might be easy—outsource as much as you can out of this mess. Usually, the first services to go are marginal operational activities, such as laundry, cleaning, and catering services. It is easy to add information systems to this list, but we believe this is often done without a careful enough consideration. Outsourcing is often seen as a trendy, obvious, and easy solution, which has been supported by financial facts on the short run. Many examples show that even in the case of operational information systems, outsourcing can become a costly option, not to mention lost possibilities for organizational learning and competitive positioning through mastering of information technology.

BACKGROUND

We have found the following reasons for the late adoption of modern information technology in the health care sector (Suomi, 2000):

- Fragmented industry structure
- Considerable national differences in processes
- Strong professional culture of medical care personnel
- One-sided education
- Handcrafting traditions
- Weak customers
- Hierarchical organization structures

ICT and governance structures meet in two ways. On one side, ICT enables new governance structures for the health care industry. On the other, it is an object in need of governing. As both sectors offer a multitude of new possibilities, innovations are called for in the industry (Christensen, Bohmer, & Kenagy, 2000).

IT governance thinking matures in organizations as any other discipline. Van Grembergen, De Haes, and Guldentops (2003) have defined the following stages in their IT Governance Maturity Model:

- Non-existent
- Initial/ad-hoc
- Repeatable but intuitive
- Defined process
- Managed and measurable
- Optimized

Needless to say, in the health care industry, IT Governance thinking is non-existent or initial/ad hoc in the best situation.

THE MEANING OF ICT GOVERNANCE STRUCTURE IN HEALTH CARE

IT is an old acronym for information technology. Nowadays, it is replaced often with the term ICT, referring to information and communication technology. This emphasizes the communication services that are developing very quickly, such as the Internet and mobile services. The letter *C* is often upgraded to the second dimension: alongside communication it can refer to contents. IT or ICT governance is defined (IT Governance Institute, 2001) as follows:

IT governance is the responsibility of the board of directors and executive management. It is an integral part of enterprise governance and consists of the leadership and organizational structures and processes that ensure that the organization's IT sustains and extends the organization's strategies and objectives.

For many, there is a temptation to understand governance as just a synonym for management. This is an oversimplification. Management is a goal-oriented activity, whereas governance is often given from the outside, and organizations just have to live with it. This is not to say that all governance structures would be beyond management control; management can influence most governance structures, at least in the long run. The long run is a key term in many aspects. We talk about structures that are semi-permanent and not changed very frequently. Structure is a term closer to architecture than to infrastructure; governance structures are architectural terms and are then implemented into infrastructures through different organizational forms. The terms *organization form* and *governance structure* are not synonyms. Organizational forms are more formal and touch upon one organization, whereas governance structures are found in a richer selection of forms and organize themselves over a number of organizations. Table 1 summarizes our discussion.

Governance structures are present in almost any human decision-making situation. In Table 2, we have a collection of key aspects of governance structure issues in health care.

FUTURE TRENDS

One of the biggest changes in the industry is that information related to health, sickness and medicines is not scarce. Internet is a rich source of such information, at different levels of expertise and at different languages. The gap between what information is available and what a health-care professional should know is growing very fast (Weaver 2002). This will shift the power balance between health professionals and patients: increasingly often the patients are the best experts on their disease. Different electronic forums or Virtual Communities (Rheingold 1993) related to health are born on the Internet. They have different services and values to offer to the healthy ones and to the chronic and acute sick (Utbult, 2000). Similarly, the interaction between the patients and health care professionals is going to change: electronic means are going to take share from face-to-face meetings (Cain, Sarasohn-Kahn & Wayne 2000; Gibson 2003).

For organizing patient flows through the health care system modern ICT offers many possibilities. Should patient data be all the time available anywhere though electronic means, would the Healthcare Supply Chains be much more effective (More & McGrath 2001). Effectiveness means that patients are taken care of in the best and most effective places, be they public or private, and of right level of expertise. As patient data can be electronically cumulated into huge databases, these databases can be used for different statistical, research and other purposes. This calls for care and proper legislation giving the principles.

Managing and building governance structures for ICT in health care organizations is not that much different from other organizations. Even in health care organizations, the scope and status of information resource management has to be decided. Issues such as sourcing decisions, charging arrangements, data privacy, and security issues all deserve their attention. There are certain problems that need to be solved in this area:

- Data privacy and security needs are extremely important and might sometimes conflict with optimal care.
- As the area is new, legislation is often lagging behind.
- The field is a meeting place for two strong professional cultures—medical doctors and ICT-professionals—that might bring along difficulties.

Table 1. *Comparison of terms management, organizational form, and governance structure*

	Management	Organizational Form	Governance Structure
Time perspective	Short	Medium	Long
Focus	Action	Internal organization	Inter-organizational structures
Management Control	In action	Easy	Difficult
Metaphor	Communication channels	Infrastructure	Architecture
Character	Concrete	Formal	Abstract

Table 2. *ICT governance structure issues in health care*

ICT as an enabler
- Health-related information on the web
- Private-public sector co-operation
- Allocation of patients to different levels of care
- Customer contacts distribution between electronic and classical means
- Ownership, structure and allocation of patient, population-level and other critical data
- Electronic forums for patients to interact
- Electronic prescription systems

ICT as an object to be governed
- New legislation needs because of the new data processing possibilities
- Data privacy and security
- Structure and status of the information resource management in health care units
- ICT-general management partnership
- Sourcing decisions of ICT
- Charging arrangements on ICT-services

CONCLUSION

Health is undoubtedly among the most important issues for all of us. In a modern society, the threats towards health are changing all the time, but at the same time, the possibilities to maintain health and cure illnesses grow exponentially. The task is to make needs and solutions meet in an effective way. This is about information and communication technologies and governance structures.

Modern ICT allows health care organizations to structure themselves in new, innovative ways, and simultaneously to empower the customers to interact with the organizations, with fellow patients, and with information sources in revolutionary new ways. In this environment, health care professionals also have to adjust their roles.

REFERENCES

Cain, Mary M., Sarasohn-Kahn, Jane, & Wayne, Jennifer C. (2000). *Health e-people: The online consumer experience*. California Health Care Foundation, Oakland, CA.

Christensen, Clayton M., Bohmer, Richard, & Kenagy, John (2000, September-October). Will distruptive innovations cure health care? *Harvard Business Review*, 102-112.

Gilson, L. (2003). Trustnext term and the development of health care as a social institution. *Social Science & Medicine, 56*(7), 1453-1468.

IT Governance Institute (2001). IT governance executive summary. Retrieved January 10, 2004, from www.itgi.org

More, Elizabeth, & McGrath, G. Mike (2001). Reengineering the healthcare supply chain in Australia: The PeCC initiative. In Robert Stegwee & Ton Spil (Eds.), *Strategies for healthcare information systems* (pp. 114-125). Hershey, PA: Idea Group Publishing.

Rheingold, Howard (1993). *The virtual community – Homesteading on the electronic frontier.* Addison-Wesley, New York.

Suomi, Reima (2000). Leapfrogging for modern ICT usage in the health care sector. *Proceedings of the 8th ECIS conference*, Vienna, Australia.

Utbult, Mats (2000). Näthälsä. *Internetpatienter möter surfande doktorer – uppstår confrontation eller samarbete*. TELDOK Rapport 138. Stockholm.

Van Grenbergen, Wim, De Haes, Steven, & Guldentops, Erik (2003). Structures, processes and relational mechanisms for IT governance. In Wim Van Grembergen (Ed.), *Strategies for information technology governance* (pp. 1-36). Hershey, PA: Idea Group Publishing.

Weawer, Robert R. (2002, March). Resistance to computer innovation: Knowledge coupling in clinical practice. *Computers and Society*, 16-21.

KEY TERMS

Electronic Patient Record: All health-related information related to a patient in electronic form, assembled as a single entity.

Electronic Prescription: Prescriptions created and handled in electronic form in an integrated information system.

Healthcare Supply Chain: A managed set of activities related to the health care activity of a patient, organized so that all necessary information is available all the time and the participants in the chain have a complete picture of the total process.

Sourcing Decision: Decision whether to buy goods/services from the market or to make them self. Sourcing decision can be made as an independent decision unit or as a part of a bigger group.

Virtual Community: A social aggregation on the Internet when people interact long enough to form personal relationships.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1305-1308, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Government Intervention in SMEs E-Commerce Adoption

Ada Scupola
Roskilde University, Denmark

INTRODUCTION

Innovation and technological change has been considered an important factor for economic development. Information technology has been among the fastest growing innovations in both production and use in the second half of the last century. In the last decade, a particular type of information technology, the Internet, has been changing business processes, organizational and industrial structures and given form to new communication and business forms as for example e-commerce.

The institutional environment created by governments in the form of policies and interventions is very important for the economic development of developed as well as developing nations (e.g., North, 1990). The external environment, and especially the role of government, has been very important in the adoption and diffusion of technological innovations such as telecommunications and more recently e-commerce (e.g., Tornatzky & Fleischer, 1990). Government intervention is and has been especially important at sustaining technological development in SMEs (Rothwell, 1994). Recently, many governments and international organizations are taking initiatives to foster the adoption of electronic commerce in small and medium size enterprises (OECD, 1999). For example the American government has set up a set of guidelines to foster the diffusion of electronic commerce in SMEs and the European Union has approved a series of "Directives" aiming at guaranteeing free availability of products and services for electronic signatures, copyright protection, taxation policy, and so forth (http://europa.eu.int/).

This study provides insights into small and medium size enterprises' perception of government intervention in e-commerce adoption in Southern Italy. The research question addressed is: "How do SMEs perceive government intervention in adoption and diffusion of e-commerce and what do they believe government intervention should focus on?" This study does not however differentiate between different types of governments, such as local, regional and national governments. The research was designed as a case study (Yin, 1994) and was conducted in Southern Italy.

The chapter is structured as follows. The next section provides a background of the institutional roles in adoption and diffusion of IT. The following section presents the research methodology. This is followed by the main thrust of the chapter that presents the major findings. Finally the last two sections discuss future trends and give some concluding remarks and suggestions for further research respectively.

BACKGROUND

The literature on adoption and diffusion of innovations, especially that focusing on information technology, has mostly focused on the factors affecting adoption and diffusion. These factors have been classified into three main groups or other categories that can be reconnected to these three groups: technological context, organizational context, and environmental context (e.g., Scupola, 2003a; Tornatzky & Fleischer, 1990).

Within the environmental context, the institutional research has focused on the influence of institutions on adoption and diffusion of technological innovations. Institutions have been historically important in the shaping of organizational and economic life and their importance is always increasing. King, Gurbaxani, Kraemer, McFarlan, Raman, and Yap (1994) identifies a series of institutions that influence IT adoption among which government authorities, international agencies and trade and industry associations.

Many studies have used institutional perspectives to study implications of information technology for organization and economic development (e.g., Corbitt & Al-Quirim, 2004; Gengatharen & Standing, 2005; Gibbs, Kraemer & Dedrick, 2002; Kraemer, Gibbs & Dedrick, 2002; Wong, 2003). Recent literature is addressing the disparity in local government readiness and community demand for e-commerce by investigating the tripartite relationship between State, local government and community in e-commerce adoption and diffusion. For example, Howell and Terziovski (2005) investigated adoption and diffusion of e-commerce in 12 local government councils in Australia funded by the Victorian e-Commerce Early Movers Assistance Scheme (VEEM). They found that the VEEM scheme was successful in raising awareness of e-commerce within the community; however there is a wide disparity in local government readiness for e-commerce and community demand for e-com-

merce emphasizing the importance of tripartite relationship between State, local government, and the community in e-commerce diffusion.

The research on the role of institutions in the adoption and diffusion of information technology is summarized here in three main frameworks: Andersen, Bjørn-Andersen, and Dedrick (2003) model of environmental drivers, Lal (2001) analytical framework encompassing the interactions among different factors and King et al. (1994) model of institutional actions.

Andersen et al. (2003) model for analyzing environmental factors mainly focuses on the demand drivers. Such drivers include industry structure (e.g., concentration, sectoral distribution, vertical integration, size of firms, etc.), information infrastructure (telecommunication, wireless and Internet infrastructure, technology acceptance, etc.), financial and human resources (e.g., venture capital, population, IT skills, education) and social and cultural factors (consumption patterns, consumer preferences, language, business culture, etc.). The second group of factors of the model includes initiatives taken by the government and private sector institutions to promote e-commerce. The model identifies four main initiatives: knowledge diffusion, economic incentives, regulation and legislation and electronic government.

Lal (2001) proposes an analytical framework that encompasses the interactions among government policies, information infrastructure, the IT industry and the markets. The framework shows that governments can influence the growth of an industry (in Lal's study the IT industry in India) by embarking on economic policies that affect supply-side and demand-side factors. Supply-side factors include telecommunications networks, power, transport and human resources development, while demand-side factors include encouragement of the use of IT in domestic markets.

King et al. (1994) classify the nature of institutional intervention in IT innovation on whether the desired changes are in production or use. Production concerns the actors that make innovative products, while use concerns the actors and ways in which innovations are used in the society. Institutions can affect IT adoption in several ways, for example

Table 1. Dimensions of institutional intervention (King et al., 1994)

	SUPPLY-PUSH	**DEMAND-PULL**
INFLUENCE	(I) KNOWLEDGE BUILDING Funding of research projects KNOWLEDGE DEPLOYMENT Provision of Educational Services SUBSIDY Funding Development of Prototypes Encouragement of capital markets to support R&D activity Provision of tax benefits for investment in R&D (e.g., investment tax credits, rapid depreciation) INNOVATION DIRECTIVE Direct institutional operation of production facilities for innovation	(II) KNOWLEDGE DEPLOYMENT Training programs for individuals and organizations to provide base of skilled talent for use SUBSIDY Procurement of innovative products and services Direct or indirect provision of complementarities required for use Direct or indirect suppression of substitute products or services MOBILIZATION Programs for Awareness and promotion
REGULATION	(III) KNOWLEDGE DEPLOYMENT Require education and training to the citizens SUBSIDY Reduction in general liabilities for organizations engaging in innovative activity Modification of legal, administrative or competitive barriers to innovation and trade STANDARDS Establishment of standards under which innovative activity might be encouraged INNOVATION DIRECTIVE Establishment of requirements for investment in R&D by organizations	(IV) SUBSIDY Procurement Support for products and processes that facilitate adoption and use STANDARDS Require particular products or processes to be used in any work for the institution Require conformance with other standards that essentially mandate use of particular products or processes INNOVATION DIRECTIVE Require that specific innovative products or Processes be used at all times

by using legal forces or by stimulating demand through the creation of a need for innovative products and processes (Montealegre, 1999).

King et al. (1994) classify the forms of institutional actions into influence and regulation. Influential initiatives have the purpose of changing behavior of those under the institution's way, without direct use of force or exercise of command. Regulatory actions have the purpose of affecting the behavior of entities under formal institutional jurisdiction such as directives. Furthermore, influence and regulation can play different roles depending on whether the innovation is driven by demand-pull or by supply-push. Supply-push forces for innovation come from the production of the innovative product or process. Demand-pull forces arise from the willingness of potential users to use the innovation (King et al., 1994). Based on the categories of influence, regulation, demand-pull and supply-push, King et al. (1994) identify six types of institutional actions that can stimulate or retard IT adoption, summarized in Table 1. These six categories are knowledge building, knowledge deployment, subsidy, mobilization, and standard setting and innovation directives.

Knowledge Building. Knowledge building consists of the institutional actions undertaken with the purpose of providing the base of scientific and technical knowledge necessary to produce and exploit innovations. The most obvious form of knowledge building is sponsored research that can be either basic or applied for which the government is the most common supporter.

Knowledge deployment. Knowledge deployment involves institutional actions aimed at disseminating new knowledge, either in form of knowledgeable individuals and organizations, or in the form of repositories of knowledge as archives and libraries of scientific and technical facts. The most important form of knowledge deployment is the general provision of education to the population. Finally, knowledge deployment can also be achieved by stimulating the use of innovations through the training of a group of potential users.

Subsidy. Subsidies have the purpose of defraying the otherwise unavoidable costs or risks to innovators and users in the process of innovation adoption and diffusion. They take different forms: funding of prototyping, institutional procurement of innovations, and support for provision of necessary complements to be used with innovative products or processes.

Mobilization. Mobilization refers to institutional actions aimed at encouraging decentralized actors to think in a positive or negative way about an innovation as for example through promotional and awareness campaigns.

Standard setting. Standard setting comprises actions that regulate the operation of decentralized actors and institutions to bring them into line with larger social or institutional objectives.

Innovation Directives. Innovation directives are institutional actions aiming at producing innovations, using them, or engaging in some activities facilitating their production or use.

King et al. (1994) model has been the basis for the following analysis (Scupola, 2003).

METHODOLOGY

The data used in this article are part of a larger research project on adoption and diffusion of electronic commerce in Southern Italian SMEs (e.g., Scupola 2002; 2003a; 2003b). The research was designed as a case study (Yin, 1994) to understand issues in adoption and diffusion of e-commerce in small and medium size enterprises, including actual, and desirable government intervention.

The Sample and the Sample Selection Process

Six interviews in six different companies were conducted. The companies have been chosen on the basis of representativeness and accessibility according to the following criteria:

- They should be a registered company and could be classified as SME according to the number of employees that should not exceed 500 according to the OECD (1999) definition.
- They should have been early adopters of Internet. Having had an Internet connection for at least 3 years was chosen as criterion of early adoption. This is based on the consideration that these companies with their experience are in a better position to identify and evaluate issues related to e-commerce adoption, including possible government intervention.
- The companies should be located in the same geographical region. This criteria should ensure that external factors such as government influence, average level of education of the population, availability of qualified labor force, and so forth are the same for all the sample companies.

The sample includes two IT consulting companies, two distributors, one producer of textiles, and an intermediary in the textile business. They are all located in the Southern Italian region called Puglia according to the last selection criterion (Table 2).

The local yellow page directory and a directory of companies distributed by the Chamber of Commerce of the city of Lecce were the primary sources in the selection of the cases. One consulting company and the local chamber of commerce were very helpful to narrow down the sample to

Table 2. Companies description

Company	Type of Business	No. Of Employees	Year of E-commerce Adoption
F1	IT Consultants	80	1996
F2	IT Consultants	1 (family driven)	1996
F3	Distributor of Watches	15	1996
F4	Intermediary in the Textile Business	2 (Family Driven)	1998
F5	Production and Commercialization of Textiles	300	1996
F6	Distributor of Car Parts	19 (family driven)	1998

Table 3. Desired intervention by different companies

Company/ Desired Intervention	F1	F2	F3	F4	F5	F6
Knowledge Building						
Knowledge Deployment	X	X	X	X	X	X
Subsidy	X	X	X	X		X
Innovation Directive						
Mobilization	X		X			
Standards						

those companies that satisfied the criteria mentioned previously. Many more companies were contacted by telephone, but only those in Table 2 were willing to participate to the study.

RESEARCH PROCESS AND DATA ANALYSIS

One person was interviewed in each company. The company has suggested the person to be interviewed on the basis of the author's requirement to talk with the employee responsible for and most knowledgeable about e-commerce. In all the cases this person has been the owner, often functioning as CEO. Semistructured interviews were the main data collection method. The interview questions were formulated with the intention of understanding issues of adoption and diffusion of e-commerce in SMEs in this region, including government intervention. Specifically, the questions around government intervention were aiming at understanding what kind of intervention (if any) the companies had been getting from the government and what they believed future intervention should address. The questions did not address whether the intervention was coming or should be come from local, regional or national government. It actually turned out that even though government intervention initiatives had been present in the region, small companies were not aware of their existence. Interviews also covered demographic data on each firm and informant.

The interviews have been conducted by the author at the company's site and have lasted between 1.5 and 3 hours each. Each interview was tape-recorded and the contents of the tape fully transcribed. Notes were also taken during the interview. By following Iacovou, Benbasat, and Dexter (1995) to enhance validity summaries of the major findings of each interview were verified by the participants after the end of each interview session. To increase reliability an interview protocol was used and a case study database was developed (Yin, 2003). The interview protocol was first tested with one of the companies and successively adjusted

to make it more clear and comprehensive. By following Yin (2003) the data were analyzed by following the "general strategy of relying on theoretical orientation" of the case study. By following Miles and Huberman (1994, p. 58) a provisional "start list" of codes was created prior to the field work to guide the analysis. The coding was manual. This "start list" came from the conceptual framework and the research question.

MAIN FOCUS OF THE CHAPTER

Being the focus of this chapter, government intervention in SMEs adoption and diffusion of e-commerce, it is natural to position the analysis in a demand-pull rather then a supply-push perspective (see King et al., 1994). The demand-pull perspective deals, in fact, with influence and intervention regarding the use of technological innovations. The government actions that either are perceived to be taking place or are desired by small enterprises to foster e-commerce adoption are described in the rest of this paragraph.

Knowledge Deployment. Southern Italian SMEs believe that State intervention to deploy knowledge is very essential for their uptake of e-commerce. This could take especially the form of e-commerce training programs and more widespread knowledge of English. These training programs could be used for example to spread knowledge about potential benefits of e-commerce and increase first hand-on experience of the uses of e-commerce as similar studies also have found out (e.g., Poon & Swatman, 1999). Some support for this purpose was available at the time of the study, but small companies were not taking much advantage of it. The companies dealing with suppliers and buyers located in foreign countries also expressed the wish to increase the English language skills among the local population and SMEs employees in particular, similarly to what found by Madon (2000) in Latin America. It seems therefore that intervention aimed at increasing the knowledge of English among the nonEnglish speaking population would contribute to adoption and diffusion of e-commerce.

Subsidies. Direct or indirect subsidies were important both as influence and regulation mechanisms. Small businesses wished generally an increase in indirect subsidies such as procurement of e-commerce technologies and services by governmental institutions and an increase in direct subsidies such as procurement support to small companies for products and processes that facilitate adoption and use. Some companies expressed especially the wish for government institutions and the public administration to procure and use Internet technologies and services both to the public and to the average citizen. By doing so, they believed that government institutions would contribute to decrease uncertainty about e-commerce, serve as role model for corporations and create the need for its use (Rogers, 1995). Another type of subsidy mentioned by SMEs in this study is direct or indirect procurement support for e-commerce systems to small businesses such financial aid for the acquisition of the system and relative training, tax deduction or financial aid that totally or partially covers acquisition and installation costs. This kind of support existed at the time of the study; however some small and medium size enterprises were not aware of it.

Mobilization. Some companies also expressed the wish for more programs aiming at increasing awareness of Internet technologies and e-commerce. This can be achieved through educational and informational campaigns aimed both to the larger population and to small businesses (Poon & Swatman, 1999). However, these campaigns should also have the objective of informing small and medium size enterprises of the existence of government subsidies and other forms of intervention, as they often are not aware of them.

FUTURE TRENDS

As the analysis has showed, small businesses wish government intervention, both in terms of influence and regulation to foster adoption and diffusion of electronic commerce. Such intervention should concentrate on three different areas: knowledge deployment, subsidies, and mobilization. Mobilization should aim at increasing awareness of the technology, related benefits, and ways of use (Poon & Swatman, 1999). In addition, mobilization initiatives should also aim at informing the companies of state and other institutions' e-commerce support programs and initiatives.

Knowledge deployment should aim at increasing knowledge of e-commerce (e.g., through targeted training programs), but also and especially at increasing knowledge of English among the population in general and small businesses employees in particular. Subsidies have emerged important both as influence and regulation mechanisms. The most important desired form of subsidy is indirect subsidies aiming at improving e-government. Direct subsidies such as financial support, tax deductions, and e-commerce pilot programs are also considered important and desirable. The study has not found evidence for standard setting, innovation directives and knowledge creation. Furthermore the study has found significant evidence for Conjuncture 2 and 4 in King et al. (1994) framework stating:

Conjecture 2: "Significant (production or) use of IT innovation requires serious and sustained institutional interventions for knowledge deployment,"

Conjecture 4. "Mobilization efforts are important but not essential in stimulating (production and) use of IT innovation, and are useful mainly in conjunction with other institutional interventions."

The study has, in fact, found that knowledge deployment is considered by SMEs a very important type of intervention

to increase adoption of e-commerce and that mobilization is also important, but not really essential and it would lead to some results mainly in conjunction with knowledge deployment and subsidy.

However the study has found only partial support for Conjecture 3 stating:

Conjecture 3: "Subsidies are often crucial but not always essential instruments of institutional intervention in both the production and use of IT innovation."

In fact, if it is true that financial subsidies such as tax breaks, and so forth are crucial, but not essential, SMEs believe that indirect subsidies in the form of government procurement of e-commerce and e-services are very essential to the adoption and diffusion of e-commerce among SMEs.

CONCLUSION

The main contribution of this study consists in illustrating what are the types of government intervention presently offered to SMEs in Southern Italy, how much these companies know about them and what they desire government intervention should address. This study has showed that SMEs desire institutional intervention, both in term of influence and regulation. Such intervention should concentrate on three main different areas: knowledge deployment, subsidies, and mobilization.

The study has also found that there is starting to be a convergence between what companies want and what the government does. This is especially happening regarding indirect subsidies such as deployment of e-services by the government and public administration.

The study has also not taken into consideration whether the intervention was undertaken by the local, the regional or the national governments. It could be interesting to take this into consideration in future studies as different initiatives have been taken place at different levels within the same nation-state and at a pan European level (e.g., Scupola, 2003c).

To conclude, more attention should be given to understand e-commerce related intervention initiatives at national, regional and local government levels in different geographical regions and what SMEs believe that their needs for intervention are. An analysis of the convergence and divergence of such results could also be done to highlight the successes and failures of current government intervention initiatives.

REFERENCES

Andersen, K. V., Bjørn-Andersen, N., & Dedrick, J. (2003). Governance initiatives creating a demand-driven e-commerce approach: The Case of denmark. The Information Society.

Corbitt, B. & Al-Quirim, N. (Eds.) (2004). *E-business, e-government & small and medium-size enterprises: Opportunities and challenges*. Hershey, PA: IGI Publishing.

Gengatharen, D. & Standing, C. (2005). A framework to assess the factors affecting success or failure of the implementation of government-supported regional e-marketplaces for SMEs. *European Journal of Information Systems, 14*(4), 417-433.

Gibbs, J., Kraemer, K. L., & Dedrick, J. (2003). Environment and policy factors shaping e-commerce diffusion: A cross-country comparison. *The Information Society, 19*(1), 5-18.

Howell, A. & Terziovski, M. (2005). E-commerce, communities and government - A snapshot of the Australian experience. In P. Van Den Besselaar, G. De Michelis, J. Preece & C. Simone (Eds.), In *Proceedings of the Second Communities and Technologies Conference* (pp. 341-357). Milano: Springer-Verlag.

Iacovou, C. L., Benbasat, I., & Dexter, A. S. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. *MIS Quarterly, 19*(4), 465-485.

King, J. L., Gurbaxani, V., Kraemer, K. L., McFarlan, F. W., Raman, K. S., & Yap, C. S. (1994). Institutional factors in information technology innovation. *Information Systems Research, 5*(2), 139-169.

Kraemer, K. L., Gibbs, J. & Dedrick, J. (2002). Impacts of globalization on e-commerce use and firm performance: A cross-country investigation. *The Information Society, 21*(5).

Madon, S. (2000). The internet and socio-economic development: Exploring the interaction. *Information Technology and People, 13*(2), 85-101.

Miles, M. B. & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Sage Publications

Montealegre, R. A. (1999). Temporal model of institutional interventions for information technology adoption in less-developed countries. *Journal of Management Information Systems, 16*(1), 207-232.

North, D. C. (1990). *Institutional structure and institutional change*. New York: Cambridge University Press.

OECD (1999). Business-to-business e-commerce: Status, economic impact and policy implications. *OECD Working Paper, NO. 77*.

Poon, S. & Swatman, P. (1999). An exploratory study of small business internet commerce issues. *Information and Management, 35*, 9-18.

Rogers, E. M. (1995). The diffusion of innovations (4th ed.). New York: Free Press.

Rothwell, R. (1994). The changing nature of the innovation process: implications for SMEs. In R. Oakey (Ed.), *New technology based firms in the 1990s*. London: Paul Chapman Publishing.

Scupola, A. (2002). Adoption issues of business-to-business internet commerce in European SMEs. In R. Sprague, Jr. (Ed.), In *Proceedings of the 35th Hawaii International Conference on System Sciences*. IEEE Computer Society.

Scupola, A. (2003a). The adoption of internet commerce by SMEs in the south of Italy: An environmental, technological and organizational perspective. *Journal of Global Information Technology Management, 6*(1), 52-71.

Scupola, A. (2003b). Adoption of e-commerce in SMEs: Lessons from stage models. In K. V. Andersen, S. Elliot, P. Swatman, E. Trauth & N. Bjørn-Andersen (Eds.), *Seeking success in e-business: A multidisciplinary approach*. Amsterdam: Kluwer Academic Publishers.

Scupola, A. (2003c). The critical role of SMEs and e-commerce in the formation of a European information society: An assessment and policy considerations. In *Proceedings of the International Conference Innovation In Europe: Dynamics, Institutions and Values*, Roskilde University, Denmark.

Tornatzky, L. G. & Fleischer, M. (1990). *The process of technological innovations*. Lexington Books.

Wong, P-K. (2003). Global and national factors affecting e-commerce diffusion in Singapore. The Information Society, *19*(1), 19-32.

Yin, R. K. (1994). Case study research - Design and methods (2nd ed.). Sage Publication. Retrieved June 16, 2008, from http://europa.eu.int/

KEY TERMS

Adoption: E-commerce adoption is here defined as the decision to use Internet technologies and the Web to share business information, maintain business relationships, and conduct business transactions.

Diffusion: It is the spread of the capacity to produce and/or use an innovation, and its use in practice.

E-Commerce: Is here defined as "the sharing of business information, maintaining business relationships, and conducting business transactions by means of telecommunications networks. Here e-commerce is equivalent to Internet commerce.

Government Intervention: Here it includes the set of policy measures and other actions and initiatives that governments and international organizations are taking to foster the adoption of electronic commerce in SMEs.

Innovation: It is characterized by three stages: invention, innovation and diffusion. An invention is a new idea or product, which becomes an innovation when it starts diffusing in the society or move into a usable form.

Institutions: It is any standing, social entity that exerts influence and regulation over other social entities by outlasting the social entities it influences and regulates.

Small and Medium Size Enterprises (SMEs): There are many definitions of SMEs. Here they are defined as companies with up to 500 employees according to the OECD (1999) definition.

Category: Software & Systems Design

Graph Encoding and Transitive Closure Representation

Yangjun Chen
University of Winnipeg, Canada

INTRODUCTION

Composite objects represented as directed graphs are an important data structure that require efficient support Web and document databases (Abiteboul, Cluet, Christophides, Milo, Moerkotte, & Simon, 1997; Chen & Aberer, 1998, 1999; Mendelzon, Mihaila, & Milo, 1997; Zhang, Naughton, Dewitt, Luo, & Lohman, 2001), CAD/ CAM, CASE, office systems, and software management. It is cumbersome to handle such objects in relational database systems when they involve ancestor-descendant relations (or say, reachability relations). In this article, we present a new graph encoding based on a tree labeling method and the concept of branchings that are used in the graph theory for finding the shortest connection networks. A branching is a subgraph of a given digraph that is in fact a forest, but covers all the nodes of the graph. Concretely, for a DAG G (directed acyclic graph) of n nodes, the space needed for storing its transitive closure can be reduced to $O(b \cdot n)$, where b is the number of the leaf nodes of G's branching. Such a compression is, however, at the expense of querying time. Theoretically, it takes $O(\log b)$ time to check whether a node is reachable from another. The method can also be extended to digraphs containing cycles.

BACKGROUND

A composite object can be generally represented as a directed graph (digraph for short). For example, in a CAD database, a composite object corresponds to a complex design, which is composed of several subdesigns. Often, subdesigns are shared by more than one higher-level design, and a set of design hierarchies thus forms a directed acyclic graph (DAG). As another example, the citation index of scientific literature, recording reference relationships between authors, constructs a directed cyclic graph. As a third example, we consider the traditional organization of a company, with a variable number of manager-subordinate levels, which can be represented as a tree hierarchy.

In a relational system, composite objects must be fragmented across many relations, requiring joins to gather all the parts. A typical approach to improving join efficiency is to equip relations with hidden pointer fields for coupling the tuples to be joined. The so-called *join index* is another auxiliary access path to mitigate this difficulty. Also, several advanced join algorithms have been suggested, based on hashing and a large main memory. In addition, a different kind of attempts to attain a compromise solution is to extend relational databases with new features, such as *clustering* of composite objects, by which the concatenated foreign keys of ancestor paths are stored in a primary key. Another extension to relational system is *nested relations* (or NF^2 relations). Although it can be used to represent composite objects without sacrificing the relational theory, it suffers from the problem that subrelations cannot be shared. Moreover, recursive relationships cannot be represented by simple nesting because the depth is not fixed. Finally, *deductive database*s and *object-relational databases* can be considered as two quite different extensions to handle this problem (Chen, 2003; Ramakrishnan & Ullman, 1995).

In the past decade, another kind of research has been done to avoid *join* operation based on *graph encoding*. In this article, we provide an overview on most important techniques in this area and discuss a new encoding approach to pack "ancestor paths" in a relational environment (Chen, 2004; Chen & Cooke, 2006). It needs only $O(e \cdot b)$ time and $O(n \cdot b)$ space, where b is the number of the leaf nodes of the graph's branching. This computational complexity is better than any existing method for this problem, including the graph-based algorithms (Bender, Farach-Colton, Pemmasani, Skiena, & Sumazin, 2004; Schmitz, 1983), the graph encoding (Abdeddaim, 1997; Bommel & Beck, 2000; Zibin & Gil, 2001), and the matrix-based algorithms (La Poutre & Leeuwen, 1988).

TREE LABELING

In this section, we mainly discuss the concept of tree labeling, based on which our algorithm is designed. For any directed tree T, we can label it as follows. By traversing T in *preorder*, each node v will obtain a number $pre(v)$ to record the order in which the nodes of the tree are visited. In a similar way, by traversing T in *postorder*, each node v will get another number $post(v)$. These two numbers can be used to characterize the reachabilities of nodes as follows.

- **Proposition 1:** Let v and v' be two nodes of a tree T. Then, v' is a descendant of v iff $pre(v') > pre(v)$ and $post(v') < post(v)$.

- **Proof:** See Exercise 2.3.2-20 in Knuth (1969).

The following example helps for illustration.

- **Example 1:** See the pairs associated with the nodes of the directed tree shown in Figure 1. The first element of each pair is the preorder number of the corresponding node and the second is its postorder number. Using such labels, the reachabilities of nodes can be easily checked. For instance, by checking the label associated with b against the label for f, we know that b is an ancestor of f in terms of Proposition 1. We can also see that since the pairs associated with g and c do not satisfy the condition given in Proposition 1, g must not be an ancestor of c and *vice versa*.

Let (p, q) and (p', q') be two pairs associated with nodes u and v. We say that (p, q) is subsumed by (p', q'), denoted $(p, q) \prec (p', q')$, if $p > p'$ and $q < q'$. Then, u is a descendant of v if (p, q) is subsumed by (p', q').

GRAPH DECOMPOSITION AND COMPUTATION OF TRANSITIVE CLOSURES

Now we discuss how to recognize the ancestor-descendant relationships w.r.t. a general structure: a DAG or a graph containing no cycles. First, we discuss the concept of branching in 4.1. Then, we address the problem of DAGs in 4.2. Next, cyclic graphs are discussed in 4.3.

Branchings of DAGs

What we want is to apply the technique discussed above to a DAG. For this purpose, the concept of *branchings* needs to be first specified.

- **Definition 1:** (*branching* (Tarjan, 1977)) A subgraph $B = (V, E')$ of a digraph $G = (V, E)$ is called a branching if it is cycle-free and $d_{\text{indegree}}(v) \leq 1$ for every $v \in V$.

Clearly, if for only one node r, $d_{\text{indegree}}(r) = 0$, and for all the rest of the nodes, v, $d_{\text{indegree}}(v) = 1$, then the branching is a directed tree with root r. Normally, a branching is a set of directed trees. Now, we assign each edge e a same cost (e.g.,

Figure 1. Tree labeling

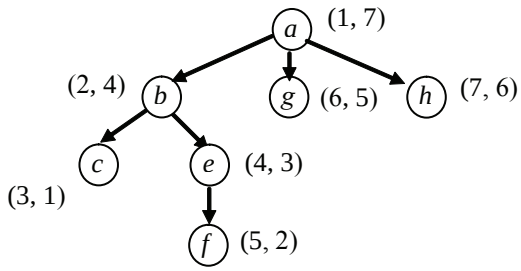
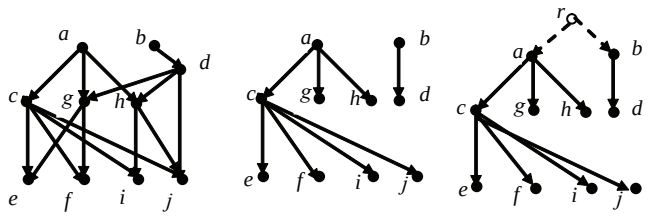


Figure 2. A DAG and its branching



(a) (b (c)

let cost $c(e) = 1$ for every edge). We will find a branching for which the sum of the edge costs:

$$\sum_{e \in E'} c(e),$$

is maximum.

For example, the trees shown in Figure 2(b) are a maximal branching of the graph shown in Figure 2(a) if each edge has a same cost.

Assume that the maximal branching for $G = (V, E)$ is a set of trees T_i with root r_i ($i = 1, ..., m$). We introduce a *virtual root* r for the branching and an edge $r \rightarrow r_i$ for each T_i, obtaining a tree G_r, called the core of G. For instance, the tree shown in Figure 2(c) is the core of the graph shown in Figure 2(a). Using Tarjan's algorithm for finding optimum branchings (Tarjan, 1977), we can always find a maximal branching for a directed graph in $O(|E|)$ time if the cost for every edge is equal to each other. Therefore, the core tree for a DAG can be constructed in linear time.

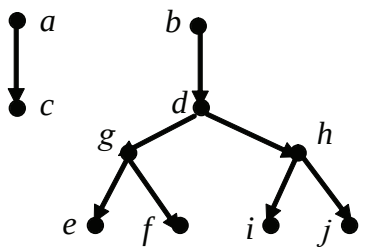
However, we may have more than one branchings for a given DAG. For instance, for the graph shown in Figure 2(a), we can find another branching as shown in Figure 3.

The number of the leaf nodes of this branching is 5 while the number of the leaf nodes of the branching shown in Figure 2(b) is 7.

An interesting question is: can we always find a branching with the least number of leaf nodes in an efficient way? To answer this question, we give the following algorithm, which works in a recursive way and searches G bottom-up as follows. Let S be the set of all nodes in G that have no descendants. We will first determine, in some way, the parents of the nodes (in S) in the branching to be generated. After that, we remove all the nodes in S from G, obtaining another DAG G'. Applying the above procedure to G' recursively, we will eventually get a branching. Whether it has the least number of leaf nodes depends on how we choose the parents of the nodes in S in each recursive step.

In the following algorithm, we use T to represent the set of the nodes, which have only the children in S. Associated with each $v \in T$ is an integer, denoted by $\alpha(v)$, equal to the

Figure 3. A branching



Box 1. Algorithm branching-generation (G,S)

Algorithm *branching-generation*(G, S)
begin
1. find a set T such that each $v \in T$ has all its children in S;
2. $T' \leftarrow \Phi$; $S' \leftarrow S$
3. **while** T is not empty **do**
4. {**find** v in T with the least α-value
5. let $v_1, v_2, ..., v_k$ be the children of v
6. add edges (v, v_i) $(1 \le i \le k)$ into *branching*
7. remove v from T; $T' \leftarrow T' \cup \{v\}$
8. remove $\{v_1, v_2, ..., v_k\}$ from S
9. **for** any singular v' **do**
10. {remove v' from T; $T' \leftarrow T' \cup \{v'\}$
11. Calculate new α-values for any of the remaining nodes in T if some of its children are removed
12. }
13. $G' \leftarrow G/S'$; (* remove all nodes in S from G *)
14. call *branching-generation*(G', T')
end

number of the nodes (in T), whose children all appear in the set of v's children. For instance, for the DAG shown in Figure 2(a), $S = \{e, f, i, j\}$ and $T = \{c, g, h\}$. In addition, $\alpha(c) = 2$, $\alpha(g) = 0$ and $\alpha(h) = 0$. When we generate part of the branching including only S and T, such α-values are used as follows:

1. Choose $v \in T$ such that $\alpha(v)$ is the smallest. (Here, the tie is resolved arbitrarily.) Put all the edges from v to its children into the branching. Remove v from T and its children from S. Remove any node in T, which becomes singular due to the removing of the nodes from S (by a singular, we mean a node not connecting any other node in the graph.) Change the α-values for some of the remaining nodes in T, whose children are partly removed.
2. Repeat step (1) until T is exhausted.

Box 1 shows the formal description of the algorithm.

The algorithm searches G bottom-up. First, we determine the parents of all the nodes in S for the branching to be created. This is done by executing lines 3-12. During this process, we choose a node v with the least α-value from T one by one, and add the edges connecting it to its children into the branching (see line 4-6). After this, both v and its children will be removed from T and S, respectively (see lines 7-8). Also, all those node in T, which become singular, will also be eliminated from T (see lines 9-10). In addition, for some of the remaining nodes in T, their α-values will be accordingly changed (see line 11). This arrangement enables us to find a branching with the least number of leaf nodes. Otherwise, choosing a node in T with a larger α-value earlier as a parent will make more nodes in T become leaf nodes. In a next step, we remove S' (which contains all the nodes in the initial S) from G, getting another DAG G', on which a recursive call *branching-generation*(G', T') will be carried out (see line 14). Here, T' is a copy of T, containing all the nodes that do not have descendants in G'.

The following example helps for illustration.

- **Example 2:** Consider the graph shown in Figure 2(a) once again. When we apply the previous algorithm to it, we will have altogether three recursive calls, which are shown below in great detail.
  - First recursive call - *branching-generation*(G, S_1):
    $S_1 = \{e, f, i, j\}$, $T_1 = \{c, g, h\}$.
    $\alpha(c) = 2$, $\alpha(g) = 0$, $\alpha(h) = 0$.
    Part of the branching generated is shown in Figure 4(a).
    $G_1 = G/S_1$ is shown in Figure 4(b).
  - Second recursive call - *branching-generation* (G_1,S_2):
    $S_2 = \{c, g, h\}$, $T_2 = \{a, d\}$.
    $\alpha(a) = 1$, $\alpha(d) = 0$.
    Part of the branching generated is shown in Figure 5(a).
    $G_2 = G_1/S_2$ is shown in Figure 5(b).
  - Third recursive call - *branching-generation*(G_2, S_3):
    $S_3 = \{a, d\}$, $T_3 = \{b\}$.
    $\alpha(b) = 0$.
    The branching generated is shown in Figure 3.
    $G_3 = G_2/S_3$ contains only node b.
    The process terminates.

The dominant cost of the previous algorithm is the execution of line 4, by which we will search the whole T to find a node with the least α-value. But it is bounded by O(b). Since such an operation can be done at most n times, the total cost of the previous algorithm is on the order of O($b \cdot n$).

Figure 4. Results of the first recursive call

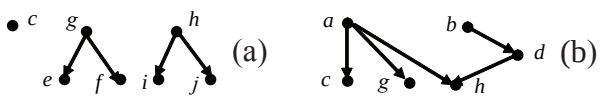


Figure 5. Results of the second recursive call

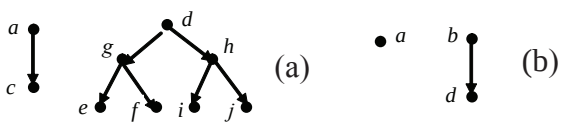
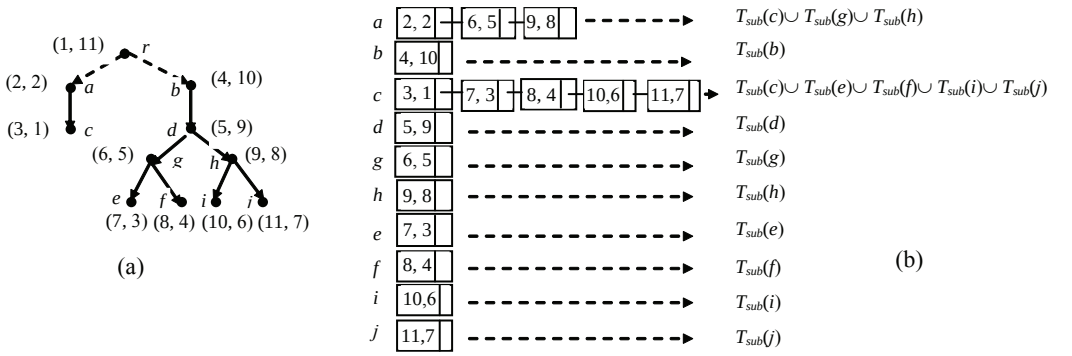


Figure 6. Tree labeling and illustration for transitive closure representation



According to the previous discussion, we have the following proposition.

- **Proposition 2:** Algorithm *branching-generation*( ) can always find a branching for any DAG with the least number of leaf nodes.

Transitive Closures of DAGs

Obviously, we can always label G_r (for some G) as discussed in Section 2.

In a G_r, a node v can be considered as a representation of the subtree rooted at v, denoted $T_{sub}(v)$; and the pair (*pre*, *post*) associated with v can be considered as a pointer to v, and thus to $T_{sub}(v)$. (In practice, we can associate a pointer with such a pair to point to the corresponding node in G_r.) In the following, what we want is to construct a pair sequence: $(pre_1, post_1), ..., (pre_k, post_k)$ for each node v in G, representing the union of the subtrees (in G_r) rooted respectively at $(pre_j, post_j)$ ($j = 1, ..., k$), which contains all the descendants of v. In this way, the space overhead for storing the descendants of a node is dramatically reduced. Later we will shown that a pair sequence contains at most O(b) pairs, where b is the number of the leaf nodes of G's branching.

- **Example 3:** The core tree G_r of the DAG G shown in Figure 2(a) can be labeled as shown in Figure 6(a). Then, each of the generated pairs can be considered as a representation of some subtree in G_r. For instance, pair (5, 9) represents the subtree rooted at *d* in Figure 6(a).

If we can construct, for each node v, a pair sequence as shown in Figure 6(b), where it is stored as a linked list, the descendants of the nodes can be represented in an economical way. Let $L = (pre_1, post_1), ..., (pre_k, post_k)$ be a pair sequence and each $(pre_i, post_i)$ be a pair labeling v_i ($i = 1, ..., k$). Then, L corresponds to the union of the subtrees $T_{sub}(v_1)$,..., and $T_{sub}(v_k)$. For example, the pair sequence (2, 2)(6, 5)(9, 8) associated with *a* in Figure 7(b) represents a union of 3 subtrees: $T_{sub}(a)$, $T_{sub}(g)$, and $T_{sub}(h)$, which contains all the descendants of *a* in G.

The question is how to construct such a pair sequence for each node v so that it corresponds to a union of subtrees in G_r, which contains all the descendants of v in G.

First, we notice that by labeling G_r, each node in $G = (V, E)$ will be initially associated with a pair as illustrated in Figure 7. That is, if a node v is labeled with (*pre*, *post*) in G_r, it will be initially labeled with the same pair (*pre*, *post*) in G.

To compute the pair sequence for each node, we sort the nodes of G topologically, that is, $(v_i, v_j) \in E$ implies that v_j appears before v_i in the sequence of the nodes. The pairs to be generated for a node v are simply stored in a linked list A_v. Initially, each A_v contains only one pair produced by labeling G_r.

We scan the topological sequence of the nodes from the beginning to the end and at each step we do the following:

Let v be the node being considered. Let $v_1, ..., v_k$ be the children of v. Merge A_v with A_{vi} each for the child node v_i ($i = 1, ..., k$) as follows. Assume $A_v = p_1 \rightarrow p_2 \rightarrow ... \rightarrow p_g$ and $Av_i = q_1 \rightarrow q_2 \rightarrow ... \rightarrow q_h$, as illustrated in Figure 8. Assume that both A_v and Av_i are increasingly ordered. (We say a pair p is larger than another pair p', denoted $p > p'$ if $p.pre > p'.pre$ and $p.post > p'.post$.)

We step through both A_v and from left to right. Let p_i and q_j be the pairs encountered. We will perform the following checkings to merge into A_v.

1. If $p_i.pre > q_j.pre$ and $p_i.post > q_j.post$, insert q_j into A_v after p_{i-1} and before p_i and move to q_{j+1}.
2. If $p_i.pre > q_j.pre$ and $p_i.post < q_j.post$, remove p_i from A_v and move to p_{i+1}. (*p_i is subsumed by q_j.*)

Figure 7. Graph labeling

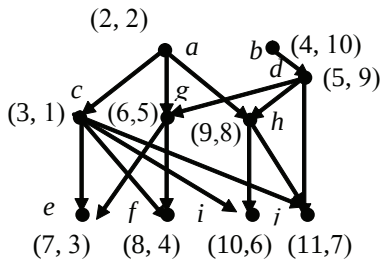
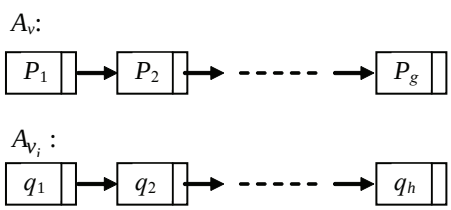


Figure 8. Linked lists associated with nodes in G



Box 2. Algorithm pair-sequence-merge (A1, A2)

Input: A_1 and A_2 - two linked lists associated with v_1 and v_2.
Output: A - modified A_1, containing all the pairs in A_1 and A_2 with all the subsumed pairs removed.

```
begin
1    p ←first-element(A1)
2    q ←first-element(A2)
3    while p ≠ nil do{
4            while q ≠ nil do{
5            if (p.pre > q.pre ∧ p.post > q.post) then
6            {insert q into A1 before p
7            q ←nextq);} (*nextq) represents the pair next to q in A2.*)
8            else if (p.pre > q.pre ∧ p.post < q.post) then
9                    {p' ←p; (*p is subsumed by q; remove p from A1.*)
10                   remove p from A1
11                   p ←next(p');} (*next(p') represents the pair next to p' in A1.*)
12                   else if (p.pre < q.pre ∧ p.post > q.post) then
13                           {q ←nextq);} (*q is subsumed by p; move to the next element of q.*)
14                   else if (p.pre < q.pre ∧ p.post < q.post) then
15                                   {p ←nextp);}
16                           else if (p.pre = q.pre ∧ p.post = q.post)
17                                   then {p ←nextp); q ←nextq);}
18   if p = nil ∧q ≠nil then
     {attach the rest of A2 to the end of A1;}
end
```

Figure 9. An entire merging process

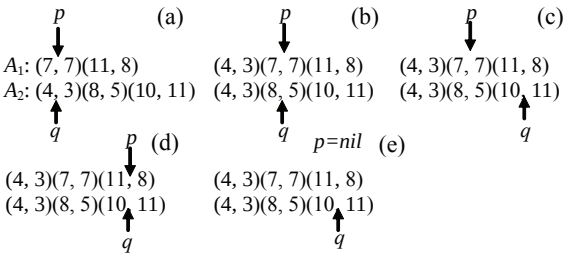
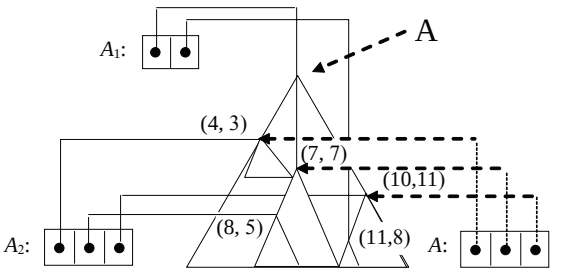


Figure 10. Illustration of merging two pair sequences



3. If $p_i.pre < q_j.pre$ and $p_i.post > q_j.post$, ignore q_j and move to q_{j+1}. (*q_j is subsumed by p_i; but it should not be removed from .*)
4. If $p_i.pre < q_j.pre$ and $p_i.post < q_j.post$, ignore p_i and move to p_{i+1}.
5. If $p_i = p_j$' and $q_i = q_j$', ignore both (p_i, q_i) and (p_j', q_j'), and move to (p_{i+1}, q_{i+1}) and (p_{j+1}', q_{j+1}'), respectively.

We notice that initially each A_v contains only one pair and is trivially sorted. Then, when we merge a sorted pair sequence into another sorted pair sequence as above, the result pair sequence must also be sorted.

In terms of the previous discussion, we have the algorithm (see Box 2) to merge two pair sequences together.

The following example helps for illustration.

- **Example 4:** Assume that $A_1 = (7, 7)(11, 8)$ and $A_2 = (4, 3)(8, 5)(10, 11)$. Then, A = *pair-sequence-merge*(A_1, A_2) = (4, 3)(7, 7)(10, 11). Figure 9 shows the entire merging process.

In each step, the A_1-pair pointed to by p and the A_2-pair pointed to by q are compared. In the first step, (7, 7) in A_1 will be checked against (4, 3) in A_2 (see Figure 9(a)). Since (4, 3) is smaller than (7, 7), it will be inserted into A_1 before (7, 7) (see Figure 9(b)). In the second step, (7, 7) in A_1 will be checked against (8, 5) in A_2. Since (8, 5) is subsumed by (7, 7), we move to (10, 11) in A_2 (see Figure 9(c)). In the third step, (7, 7) is smaller than (10, 11) and we move to (11, 8) in A_1 (see Figure 9(d)). In the fourth step, (11, 8) in A_1 is checked against (10, 11) in A_2. Since (11, 8) is subsumed by (10, 11), it will be removed from A_1 and p becomes *nil* (see Figure 9(e)). In this case, (10, 11) will be appended to A_1 (see line 18 of Algorithm *pair-sequence-merge*( )), forming the result A = (4, 3)(7, 7)(10, 11) (see Figure 6(e)). Figure 10 is a pictorial illustration of the result of merging A_2 into A_1 (note that A_2 itself is not changed).

In the following, we establish several propositions to clarify the properties of the previous algorithm.

- **Proposition 3:** Let A_1 and A_2 be two pair sequences sorted in ascending order. Let A be the result obtained by merging A_2 into A_1 using Algorithm *pair-sequence-merge*( ). Then, A is also sorted increasingly.
- **Proof:** During the execution of the algorithm, some pairs may be removed from A_1 and some pair of A_2 may be inserted into A_1. Obviously, removing a pair from A_1 will not change the ordering of A_1. Let q be a pair of A_2 inserted into A_1. It may be done at line 6 or at line 18. If it is done at line 6, there must be a pair p in A_1 such that $p > q$. Consider the pair p' before p. We have $p' < q$; otherwise, q will be inserted before p' or will not be inserted into A_1 at all. In this case, the proposition holds. If q is inserted into A_1 by executing line 18, all the pairs in A_1 must be used up before the line 18 is carried out. We notice that at this moment,

Box 3. Algorithm all-sequence-generation

```
begin
1     Let v_n, v_{n-1}, ..., v_1 be the topological sequence of the nodes of G;
2     for i from n downto 1 do
3           {let v_1,...,v_k be the child nodes of v_i;
4                 for j from 1 to k do
5                       call pair-sequence-merge(A_i, Av_i );
6           }
end
```

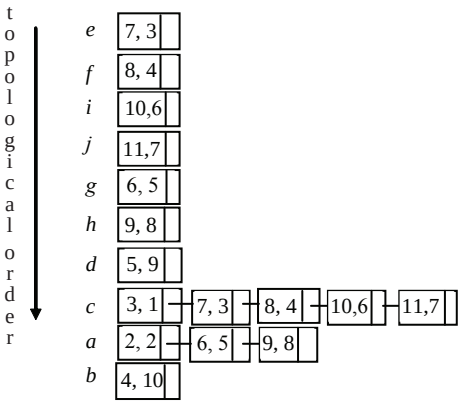
all the pairs in A_1 are increasingly ordered and smaller than all the remaining pairs in A_2, which are originally in ascending order. Therefore, in this case, the proposition holds, too.

- **Proposition 4:** Let A_1 and A_2 be two pair sequences sorted in ascending order. Let A be the result obtained by merging A_2 into A_1 using Algorithm *pair-sequence-merge*( ). If v is a node in a subtree of G_r, which is rooted at some node labeled with a pair in A_2, then there must be a pair in A such that the subtree rooted at it contains v.
- **Proof:** Assume that v is in a subtree rooted at u labeled with (*pre*, *post*) that appears in A_2. If (*pre*, *post*) appears in A, the proposition holds. Suppose that (*pre*, *post*) does not appear in A. In this case, there must be a pair (*pre*', *post*') in A_1, which subsumes (*pre*, *post*). Notice that (*pre*', *post*') cannot be subsumed by any pair in A_2 since it subsumes (*pre*, *post*). Otherwise, we will have a pair (*pre*'', *post*'') in A_2 such that *pre*'' < *pre*' and *post*'' > *post*'. But we have (*pre*, *post*) < (*pre*'', *post*'') or (*pre*, *post*) > (*pre*'', *post*''). In the former case, we have *pre* < *pre*'' < *pre*'. It contradicts the fact that (*pre*', *post*') subsumes (*pre*, *post*). In the latter case, we have *post* > *post*'' > *post*'. It also contradicts the fact that (*pre*', *post*') subsumes (*pre*, *post*). Therefore, (*pre*', *post*') will appear in A. Since v is in the subtree rooted at (*pre*, *post*), it must be in the subtree rooted at (*pre*', *post*'). Thus, the proposition holds.
- **Proposition 5:** Let A_1 and A_2 be two pair sequences sorted in increasing order. Let A be the result obtained by merging A_2 into A_1 using Algorithm *pair-sequence-merge*( ). If v is a node in a subtree of G_r, which is rooted at some node labeled with a pair in A_1, then there must be a pair in A such that the subtree rooted at it contains v.
- **Proof:** Similar to Proposition 4
- **Proposition 6:** The time complexity of Algorithm *pair-sequence-merge* is bounded by $O(\max\{|A_1|, |A_2|\})$.
- **Proof:** During the execution of the algorithm, each pair in A_1 and A_2 is visited at most once.

Based on the merging operation previously discussed, the pair sequences for all the nodes in a DAG can be computed as shown in Box 3.

- **Proposition 7:** The space complexity of Algorithm *all-sequence-generation* is bounded by $O(n \cdot b)$, where b is the number of the leaf nodes of G's branching.
- **Proof:** In the algorithm, each node v is associated with a linked list A_v. We claim that the size of A_v is bounded by b. Assume that A_v contains $b + 1$ pairs that are different from each other. Then, there must exists two pairs p and q so that p subsumes q or *vice versa*. Therefore, the space needed for Algorithm *all-sequence-generation* is bounded by $O(n \cdot b)$.
- **Proposition 8:** The time complexity of Algorithm *all-sequence-generation* is bounded by $O(e \cdot b)$, where b is the number of the leaf nodes of G's branching.
- **Proof:** In the out for-loop of the algorithm, n steps are performed. In each step, d_i merge operations are made for each v_i, where d_i represents the outdegree of v_i. Therefore, the time spent for each step is $O(d_i \cdot b)$. The whole time complexity is thus $O() = O(e \cdot b)$.
- **Proposition 9:** Let v be a node in G. Any descendant u of v must be in a subtree of G_r rooted at a node labeled with a pair in A_v constructed by Algorithm *all-sequence-generation*.
- **Proof:** Assume that $v_n \rightarrow v_{n-1} \rightarrow ... \rightarrow v_1$ *is a topological sequence of* G. We prove the proposition by induction on the ordinal number m in the topological sequence.
  - **Basis:** When $m = 1$, v_n is a leaf node in G and its linked list contains only one label associated with v_n. The proposition holds.
  - **Hypothesis:** Suppose that when $m \leq k$ the proposition holds. That is, each linked list A_i associated with v_i $(i = n, ..., n - k)$ contains all the pairs covering all the descendants of v_i.
  - **Consider m = k + 1:** According to the property of the topological sequence, all the child nodes of v_{n-k-1} must appear in $\{v_n, v_{n-1}, ..., v_{n-k}\}$. Then, from lines 3 - 6 of Algorithm *all-sequence-gen-*

Figure 11. Linked lists representing pair sequences



eration, as well as Proposition 2, 3 and 4, we can see that the linked list A_{n-k-1} associated with v_{n-k-1} must contain all the pairs covering all the descendants of v_{n-k-1}. It completes the proof.

- **Example 5:** A possible topological sequence for the graph shown in Figure 2(a) is: $e \rightarrow f \rightarrow i \rightarrow j \rightarrow g \rightarrow h \rightarrow d \rightarrow c \rightarrow a \rightarrow b$. The initial pairs associated with them are: (4, 1), (5, 2), (6, 3), (7, 4), (8, 6), (9, 7), (11, 9), (3, 5), (2, 8) and (10, 10), respectively. They are obtained by labeling the tree shown in Figure 3(a). Applying Algorithm *all-sequence-generation* to this sequence, we will produce a linked list for each of them as shown in Figure 11.

We notice that each pair sequence associated with a node is increasingly ordered. Therefore, we can store a pair sequence in two integer sequences: one for preorder numbers and the other for postorder numbers. For example, the pair sequence associated with *c* in Figure 11 is (3, 1)(7, 3)(8, 4)(10, 6)(11, 7). It can be stored in two sequences of integers: 3, 7, 8, 10, 11, and 1, 3, 3, 4, 6, 7. To check whether *v* is a descendant of *u*, we do the following. Let *p* be the pair of *v* and *s* the pair sequence of *u*. Assume that *s* is stored in two integer sequences: s_1 (for all the preorder numbers in *s*) and s_2 (for all the postorder numbers in *s*). We first make a binary search of s_1 to find the largest *k* such that $s_1[k] < p{\cdot}pre$. Then, we search s_2 from the 1st to the *k*th element to find whether there exists *l* such that $s_2[l] > p{\cdot}post$. If it is the case, *p* is subsumed by some pair in *s*; otherwise not. Obviously, the time complexity of this process is $O(\log_2|s|)$. Since $|s|$ is bounded by the breadth *b* of the graph, this method requires $O(\log_2 b)$ time for reachability checking.

Transitive Closures of Cyclic Graphs

Based on the method discussed in the previous subsection, we can easily develop an algorithm to compute transitive closures for cyclic graphs. First, we use Tarjan's algorithm for identifying *strongly connected components* (*SCCs*) to find all the cycles in a cyclic graph (Tarjan, 1972) (which needs only $O(n + e)$ time). Then, we take each SCC as a single node (i.e., collapse each SCC to a node) and transform a cyclic graph into a DAG while maintaining all the reachability information. This is because all nodes on a cycle are mutually reachable, and hence have identical reachability properties. Next, we handle the DAG as discussed in 3.2. In this way, however, all nodes in an SCC will be assigned

Figure 12. A cyclic graph and its reduction with SCC collapsed

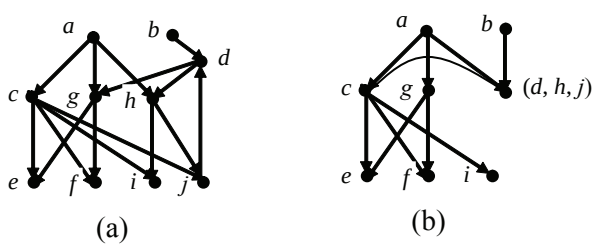
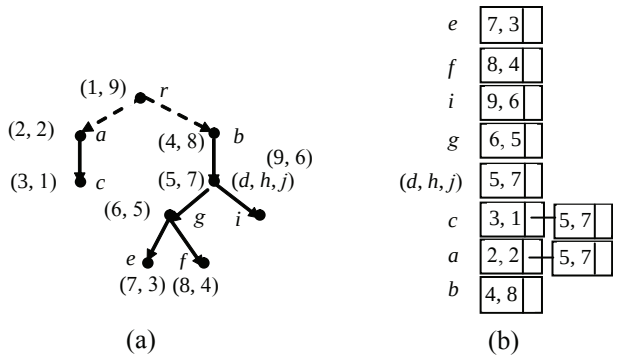


Figure 13. Linked lists representing pair sequences



the same pair (and the same pair sequence). For a better understanding, see the following example.

- **Example 6:** Consider the graph shown in Figure 12(a), which is the graph shown 2(a) with one of the edges reversed (i.e., edge (*d*, *j*) is changed to (*j*, *d*)). Then, the nodes *d*, *h* and *j* form an SCC, which can collapse to a single node, denoted by (*d*, *h*, *j*) in the graph shown in Figure 12(b).

The branching of this graph is shown in Figure 13(a) and can be labeled as discussed in Section 2. Along the reverse topological order of the graph shown in Figure 12(a), we will generate the pair sequences for all the nodes as shown in Figure 13(b), which is the representation of the graph's transitive closure.

Finally, we point out that the time complexity of computing transitive closure of a cyclic graph is still O(*e*·*b*) since Tarjan's algorithm runs in O(*n* + *e*) time (Tarjan, 1972). The idea of collapsing an SCC into a single node was first proposed in Munro (1971).

FUTURE TREND

The computation of transitive closures and recursive relationships is a classic problem in the graph theory and has a variety of applications in data engineering, such as CAD/CAM, office systems, databases, programming languages and so on. For all these applications, the problems can be represented as a directed graph with the edges being not labelled, and can be solved using the techniques described in this article. In practice, however, there exits another kind of problems, which can be represented only by using the so-called weighted directed graphs. For them, the edges are associated with labels or distances and the shortest (or longest) paths between two given nodes are often asked. Obviously, the above techniques are not able to solve such problems. They have to be extended to encode path information in the data structure to speed up query evaluation. For this, an interesting issue is how to maintain minimum information but get high efficiency, which is more challenging than transitive closures and provides an important research topic in the near future.

CONCLUSION

In this article, we provide an overview on the recursion computation in a relational environment and present a new encoding method to label a digraph, which is compared with a variety of traditional strategies as well as the methods proposed in the database community. Our method is based on a tree labeling method and the concept of branchings that are used in graph theory for finding the shortest connection networks. A branching is a subgraph of a given digraph that is in fact a forest, but covers all the nodes of the graph. On the one hand, the proposed encoding scheme achieves the smallest space requirements among all previously published strategies for recognizing recursive relationships. On the other hand, it leads to a new algorithm for computing transitive closures for DAGs in O(*e*·*b*) time and O(*n*·*b*) space, where *n* represents the number of the nodes of a DAG, *e* the numbers of the edges, and *b* is the number of the leaf nodes of the DAG's branching. In addition, this method can be extended to cyclic digraphs and is especially suitable for a relational environment.

REFERENCES

Abdeddaim, S. (1997). On incremental computation of transitive closure and greedy alignment. In A. Apostolico & J. Hein (Ed.), *Proceedings of the 8th Symp. Combinatorial Pattern Matching* (pp. 167-179).

Abiteboul, S., Cluet, S., Christophides, V., Milo, T., Moerkotte, G., & Simon, J. (1997). Quering documents in object databases. *International Journal of Digital Libraries, 1*(1), 5-19.

Agrawal, R., Borgida, A., Jagadish, J. V. (1989). Efficient management of transitive relationships in large data and knowledge bases. In *Proceedings of the ACM SIGMOD International Conference on the Management of Data* (pp. 253-262).

Bender, M., Farach-Colton, M., Pemmasani, G., Skiena, S., & Sumazin, P. (2004). Lowest common ancestors in trees and directed acyclic graphs. *Journal of Algorithms, 57*(2005), 75-94.

Booth, K. S., & Leuker, G. S. (1976). Testing for the consecutive ones property, interval graphs, and graph palanity using PQ-tree algorithms, *Journal of Computer System Science, 13*(3), 335-379.

Chen, Y. (2004). A new algorithm for computing transitive closures. In *Proceedings of the ACM SAC 2004* (pp. 1091-1092).

Chen, Y. (2003). On the graph traversal and linear binary-chain programs. *IEEE Transactions on Knowledge and Data Engineering, 15*(3), 573-596.

Chen, Y., & Aberer, K. (1999). Combining pat-trees and signature files for query evaluation in document databases. In *Proceedings of the 10th International DEXA Conference on Database and Expert Systems Application* (pp. 473-484), Florence, Italy: Springer Verlag.

Chen, Y., & Aberer, K. (1998). Layered index structures in document database systems. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM)* (pp. 406-413). Bethesda, MD: ACM.

Chen, Y., & Cooke, D. (2006). On the transitive closure representation and adjustable compression. In *Proceedings of the 21st ACM Symposium on Applied Computing* (pp. 450-455), Dijon, France.

Cohen, N. H. (1991). Type-extension tests can be performed in constant time. *ACM Transactions on Programming Languages and Systems, 13*, 626-629.

Cattell, R.G.G., & Skeen, J. (1992). Object operations benchmark. *ACM Trans. Database Systems, 17*(1), 1 -31.

Fall, A. (1995). Sparse term encoding for dynamical taxonomies. In *Proceedings of the 4th International Conference on Conceptual Structures (ICCS-96): Knowledge Representation as Interlingua* (pp. 277-292). Berlin.

Knuth, D. E. (1969). *The art of computer programming* (Vol. 1). Reading: Addison-Wesley.

Kuno, H.A., & Rundensteiner, E. A. (1998). Incremental maintenance of materialized object-oriented views in multiview: Strategies and performance evaluation. *IEEE Transactions on Knowledge and Data Engineering, 10*(5), 768-792.

Krall, A., Vitek, J., & Horspool, R. N. (1997). Near optimal hierarchical encoding of types. In M. Aksit & S. Matsuoka (Eds.), *Proceedings of the 11th European Conference on Object-Oriented Programming* (pp. 128-145). Jyvaskyla, Finland.

La Poutre, J.A., & van Leeuwen, J. (1988). Maintenance of transitive closure and transitive reduction of graphs. *Proceedings of Workshop on Graph-Theoretic Concepts in Computer Science, Lecture Notes in Computer Science 314* (pp. 106-120). Springer-Verlag.

Lee, W. C., & Lee, D. L. (1998). Path dictionary: A new access method for query processing in object-oriented databases. *IEEE Transactions on Knowledge and Data Engineering, 10*(3), 371-388.

Mendelzon, A. O., Mihaila, G. A., & Milo, T. (1997). Querying the World Wide Web. *International Journal of Digital Libraries, 1*(1), 54-67.

Munro, I. (1971). Efficient determination of the transitive closure of directed graphs. *Information Processing Letters, 1*(2), 54-58.

Ramakrishnan, R., & Ullman, J. D. (1995). A survey of research in deductive database systems. *Journal of Logic Programming*, 125-149.

Schmitz, L. (1983). An improved transitive closure algorithm. *Computing, 30*, 359-371.

Stonebraker, M., Rowe, L., & Hirohama, M. (1990). The implementation of POSTGRES. *IEEE Transaction Knowledge and Data Engineering, 2*(1), 125-142.

Tarjan, R. (1977). Finding optimum branching. *Networks, 7,* 25-35.

Tarjan, R. (1972). Depth-first search and linear graph algorithms. *SIAM Journal of Computing, 1*(2), 146-140.

van Bommel, M. F., & Beck, T. J. (2000). *Incremental encoding of multiple inheritance hierarchies supporting lattice operations*. Linkoping Electronic Articles in Computer and Information Science. Retrieved from http://www.ep.liu.se/ea/cis/2000/001

Zibin, Y., & Gil, J. (2001, October 14-18). Efficient subtyping tests with pq-encoding. *Proceedings of the 2001 ACM SIGPLAN Conference on Object-Oriented Programming Systems, Languages, and Application* (pp. 96-107). Florida.

Zhang, C., Naughton, J., DeWitt, D., Luo, Q., & Lohman, G. (2001). On supporting containment queries in relational database management systems. *Proceedings of ACM SIGMOD International Conference on Management of Data*, California.

KEY TERMS

Branching: A branching is a subgraph of a directed graph, in which there is no cycles and the indegree of each node is 1 or 0.

Cyclic Graph: A cyclic graph is a directed graph that contains at least one cycle.

DAG: A DAG is a directed graph that does not contain a cycle.

Graph Encoding: Graph encoding is a method to assign the nodes of a directed graph a number or a bit string, which reflects some properties of that graph and can be used to facilitate computation.

Strongly Connected Component (SCC): A SCC is a subgraph of a directed graph, in which between each pair of nodes there exists a path.

Topological Order: A sequence S of nodes of a DAG $G = (V, E)$: $v_1,\ldots, v_n$ such that $(v_i, v_j) \in E$ implies that v_j appears before v_i in S.

Transitive Closure: The transitive closure of a directed graph G is a graph G^*, in which there is an edge from node a to node b if there exists a path from a to b in G.

Tree: A tree is a graph with a root, in which the indegree of each node is equal to 1.

Handheld Programming Languages and Environments

Wen-Chen Hu
University of North Dakota, USA

Yanjun Zuo
University of North Dakota, USA

Chyuan-Huei Thomas Yang
Hsuan Chuang University, Taiwan

Yapin Zhong
Shandong Sport University, China

INTRODUCTION

Mobile commerce is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile, handheld devices such as smart cellular phones and PDAs (personal digital assistants). It is widely acknowledged that mobile commerce is a field of enormous potential. However, it is also commonly admitted that the development in this field is constrained. There are considerable barriers waiting to be overcome. One of the barriers is most software engineers are not familiar with the design and development of mobile applications (Kiely, 2001). This chapter gives a study of handheld computing and programming to help software engineers better understanding this subject. Handheld computing is to use handheld devices to perform wireless, mobile, handheld operations such as personal data management and making phone calls. They can be achieved by using server or client- side handheld computing and programming:

- **Server-side handheld computing and programming:** Server-side handheld computing is to use handheld devices to perform wireless, mobile, handheld operations, which require the supports of server-side computing. The most common applications of server-side handheld programming are the mobile Web contents.
- **Client-side handheld computing and programming:** Client-side handheld computing is to use handheld devices to perform handheld operations, which do not need the supports of server-side computing. Most client-side handheld programming languages are a version of either C/C++ or Java. Examples of the application development of Java ME, a version of Java, and Palm OS, using a version of C, will be given.

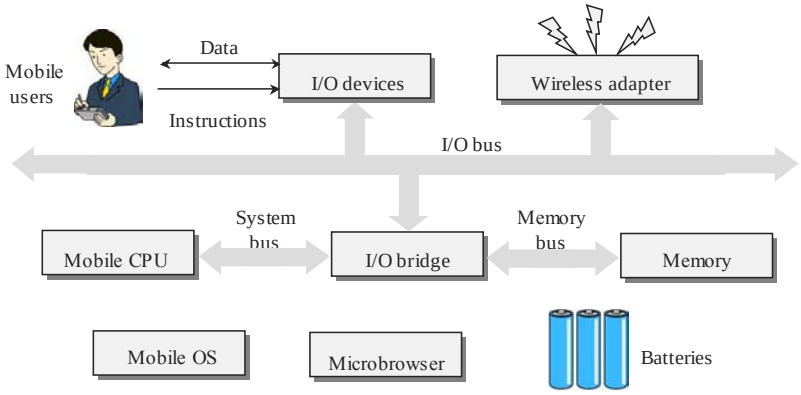
BACKGROUND

Mobile users interact with mobile commerce applications by using small wireless Internet-enabled devices, which come with several aliases such as handhelds, palms, PDAs (personal digital assistants), pocket PCs, and smart phones. To avoid any ambiguity, a general term, mobile handheld devices, is used in this article. Mobile handheld devices are small general-purpose, programmable, battery-powered computers, but they are different from desktop PCs or notebooks due to the following special features:

- Limited network bandwidth,
- Small screen/body size, and
- High mobility.

Short battery life and limited memory, processing power, and functionality are additional features, but these problems are gradually being solved as the technologies improve and new methods are constantly being introduced. The limited network bandwidth prevents the display of most multimedia on a microbrowser. Though the Wi-Fi and 3G networks go

Figure 1. A system structure of mobile handheld devices



some way toward addressing this problem, the wireless bandwidth is always far below the bandwidth of wired networks. The small screen/body size restricts most handheld devices to using a stylus for input.

Figure 1 shows a typical system structure for handheld devices, which includes the following six major components, (i) a mobile operating system, (ii) a mobile central processor unit, (iii) a microbrowser, (iv) input/output devices, (v) a memory, and (vi) batteries (Hu, Yeh, Chu et al, 2005). Synchronization connects handheld devices to desktop computers, notebooks, or peripherals to transfer or synchronize data. Without needing serial cables, many handheld devices now use either an infrared (IR) port or Bluetooth technology to send information to other devices.

MAIN FOCUS OF THE CHAPTER

Handheld computing is a fairly new computing area and a formal definition of it is not found yet. Nevertheless, the author defines it as follows:

Handheld computing is to use handheld devices such as smart cellular phones and PDAs to perform wireless, mobile, handheld operations such as personal data management and making phone calls.

Again, handheld computing includes two kinds of computing: server and client- side handheld computing, which are defined as follows:

- **Server-side handheld computing:** It is to use handheld devices to perform wireless, mobile, handheld operations, which require the supports of server-side computing.
- **Client-side handheld computing:** It is to use handheld devices to perform handheld operations, which do not need the supports of server-side computing.

The terms of computing and programming are sometimes confusing and misused. The handheld programming, defined as programming for handheld devices, is different from handheld computing and includes two kinds of programming too:

- **Server-side handheld programming:** It is design and development of handheld software such as CGI programs that reside on the servers.
- **Client-side handheld programming:** It is design and development of handheld software such as Java ME programs that reside on the handheld devices.

Server-Side Handheld Computing and Programming

Server-side handheld computing and programming usually involve complicated procedures and advanced programming such as TCP/IP network programming. This chapter will focus on the most popular server-side handheld computing and programming, mobile Web contents design and development. For other kinds of server-side handheld computing and programming such as instant messaging and telephony, readers may refer to other technical reports or articles. A database-driven mobile Web site is often implemented by using a three-tiered client/server architecture consisting of three layers as shown in Figure 2: (i) user interface, (ii)

Figure 2. A generalized system structure of a database-driven mobile Web site

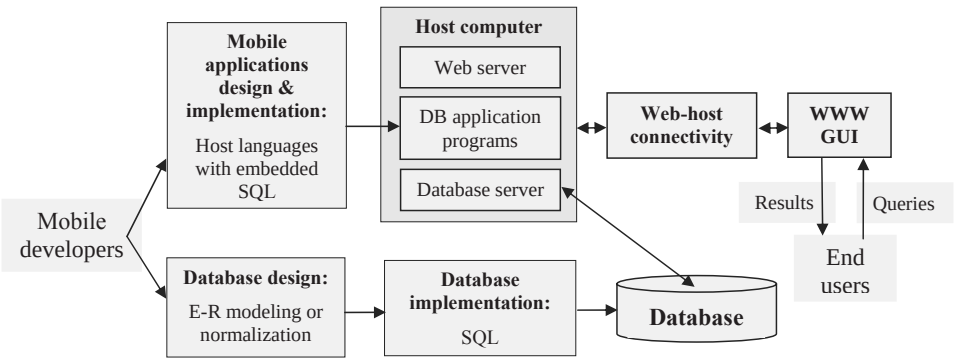
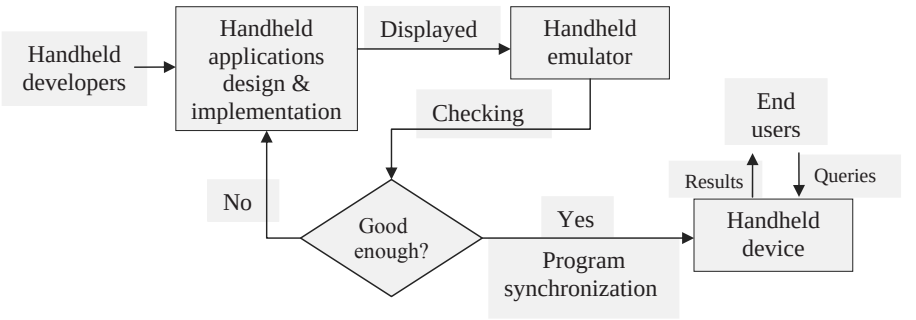


Figure 3. A generalized client-side handheld computing development cycle



functional modules, and (iii) database management systems. The three-tier design has many advantages over traditional two-tier or single-tier designs, the chief one being: The added modularity makes it easier to modify or replace one tier without affecting the other tiers.

Client-Side Handheld Computing and Programming

Various environments/languages are available for client-side handheld computing and programming. Five of the most popular are (i) BREW, (ii) Java ME, (iii) Palm OS, (iv) Symbian OS, and (v) Windows Mobile. They apply different approaches to accomplishing the development of mobile applications. Figure 3 shows a generalized development cycle applied by them and Table 1 gives a comparison among the five languages/environments.

BREW: Binary Runtime Environment for Wireless

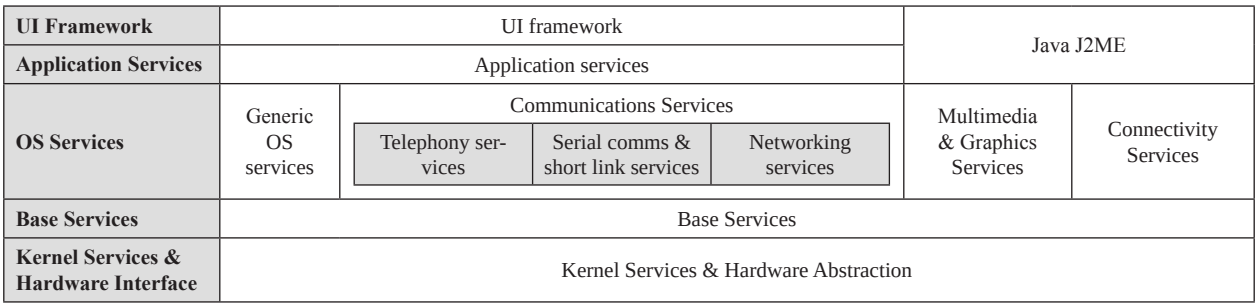
BREW is an application development platform created by Qualcomm for CDMA (code division multiple access)-based mobile phones (Qualcomm Inc., 2003). The CDMA is a digital wireless telephony transmission technique and it has two major features:

- The CDMA allows multiple frequencies to be used simultaneously (Spread Spectrum).

Table 1. A comparison among five handheld-computing languages/environments

	BREW	Java ME	Palm OS	Symbian OS	Windows Mobile
Creator	Qualcomm	Sun Microsystems	PalmSource	Symbian	Microsoft
Language/Environment	Environment	Language	Environment	Environment	Environment
Market Share (PDA) as of 2005	N/A	N/A	3rd	4th	1st
Market Share (Smartphone) as of 2006	?	N/A	4th	1st	5th
Primary Host Language	C/C++	Java	C/C++	C++	C/C++
Target Devices	Phones	PDAs & phones	PDAs	Phones	PDAs & phones

Figure 4. Symbian OS architecture



- The CDMA standards used for second-generation mobile telephony are the IS-95 standards.

BREW is a complete, end-to-end solution for wireless applications development, device configuration, application distribution, and billing and payment. It includes three major components:

- BREW SDK (software development kit) for application developers,
- BREW client software and porting tools for device manufacturers, and
- BREW distribution system (BDS) that is controlled and managed by operators—enabling them to easily get applications from developers to market and coordinate the billing and payment process.

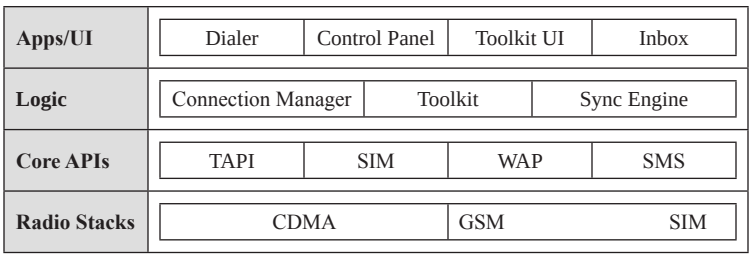
Symbian OS

Symbian is a software licensing company that develops and supplies the open operating system—Symbian OS—for data-enabled mobile phones (Symbian Ltd, 2005). Symbian was established as a private independent company in June 1998. It is an independent, for-profit company whose mission is to establish Symbian OS, whose architecture is given in Figure 4, as the world standard for mobile digital data systems, primarily for use in cellular telecoms. It is owned by Ericsson (15.6 percent), Nokia (47.9 percent), Panasonic (10.5 percent), Samsung (4.5 percent), Siemens (8.4 percent) and Sony Ericsson (13.1 percent). Headquartered in the UK, it has more than 1,300 staff with offices in Japan, Sweden, UK and the USA and a development centre in Bangalore in 2006. It is the most popular mobile operating system for smartphones. Cumulative shipments of Symbian OS phones since Symbian's formation reached 70.5 million phones in 2006.

Windows Mobile

Windows Mobile is a compact operating system for mobile devices based on the Microsoft Win32 API (Microsoft Corp., 2005). It is designed to be similar to desktop versions of Windows. In 1996, Microsoft launched Windows CE, a version of the Microsoft Windows operating system designed specially for a variety of embedded products, including handheld devices. However, it was not well received

Figure 5. Windows Mobile-based Smartphone architecture

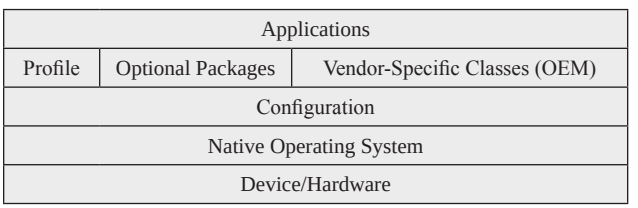


primarily because of battery-hungry hardware and limited functionality, possibly due to the way that Windows CE was adapted for handheld devices from other Microsoft 32-bit desktop operating systems. Windows Mobile includes three major kinds of software: (i) Pocket PC, (ii) Smartphones, and (iii) Portable Media Centers. Figure 5 shows the Smartphone architecture, which provides a core set of services that will abstract a variety of underlying links for both voice and data services (Finan, 2002). The primary Smartphone architecture consists of four layers: (i) applications/UI, (ii) logic, (iii) core APIs, and (iv) radio stack.

Java ME (Java Platform, Micro Edition)

Java ME provides an environment for applications running on consumer devices, such as mobile phones, PDAs, and TV set-top boxes, as well as a broad range of embedded devices (Sun Microsystem Inc., 2002a). Like its counterparts for the enterprise (Java EE), desktop (Java SE) and smart card (Java Card) environments, Java ME includes Java virtual machines and a set of standard Java APIs defined through the Java community process. The Java ME architecture, as shown in Figure 6, comprises a variety of configurations, profiles, and optional packages that implementers and developers can choose from, and combine to construct a complete Java runtime environment that closely fits the requirements of a particular range of devices and a target market. There are two sets of Java ME packages, which target different devices:

Figure 6. Java ME architecture



- **High-end devices:** They include connected device configuration (CDC), foundation, and personal profile.
- **Entry-level devices and smart phones:** They include connected limited device configuration (CLDC) and mobile information device profile (MIDP).

The following description gives an example of Java ME programming (Sun Microsystem Inc., 2004). Sun Java Wireless Toolkit is a toolbox for developing wireless applications that are based on Java ME's CLDC and MIDP. The toolkit includes the emulation environments, performance optimization and tuning features, documentation, and examples that developers need to bring efficient and successful wireless applications to market quickly. Program 1 gives a Java ME example, which displays the text "Hello, World!" and a ticker with a message "Greeting, world." Figure 7 shows an emulator displaying the execution results. For further references, the packages provided by the MIDP are given in Mobile Information Device Profile Specification 2.0 (Sun Microsystem Inc., 2002b). The packages `javax.*` are the extensions to standard Java packages. They are not included in the JDK or JRE. They must be downloaded separately.

PalmOS

Palm OS is a fully ARM-native, 32-bit operating system designed for used on Palm handhelds and other third-party devices. Its popularity can be attributed to its many advantages, such as its long battery life, support for a wide variety of wireless standards, and the abundant software available. The plain design of the Palm OS has resulted in a long battery life, approximately twice that of its rivals (PalmSource Inc., 2002). Two major versions of Palm OS are currently under development:

- **Palm OS Garnet:** It is an enhanced version of Palm OS 5 and provides features such as dynamic input area,

Program 1. An MIDlet program displaying the text "Hello, World!"

```
// This package defines MIDP applications and the interactions between
// the application and the environment in which the application runs.
import  javax.microedition.midlet.*;

// This package provides a set of features for user interfaces.
import  javax.microedition.lcdui.*;

public class HelloMIDlet  extends MIDlet  implements CommandListener {

  public void  startApp( ) {
    Display  display     = Display.getDisplay( this );
    Form     mainForm    = new  Form   ( "HelloMIDlet" );
    Ticker   ticker      = new  Ticker ( "Greeting, World" );
    Command  exitCommand = new  Command( "Exit", Command.EXIT, 0 );

    mainForm.append           ( "\n\n         Hello, World!" );
    mainForm.setTicker        ( ticker );
    mainForm.addCommand       ( exitCommand );
    mainForm.setCommandListener( this );
    display.setCurrent        ( mainForm );
  }

  public void  pauseApp ( ) { }

  public void  destroyApp( boolean unconditional ) {
    notifyDestroyed( );
  }

  public void  commandAction( Command c, Displayable s ) {
    if ( c.getCommandType( ) == Command.EXIT )
      notifyDestroyed( );
  }
}
```

Figure 7. A screenshot of an emulator displaying the execution results of Java ME program in Program 1

Figure 8. Palm OS 5 block diagram

Palm Applications	
PACE: Palm Application Compatibility Environment	
Core Palm OS	Licensee libraries
DAL: Device Abstraction Layer	
HAL: Hardware Abstraction Layer	

improved network communication, and support for a broad range of screen resolutions including QVGA.

- **Palm OS Cobalt:** It is Palm OS 6, which focuses on enabling faster and more efficient development of smartphones and integrated wireless (WiFi/Bluetooth) handhelds.

Figure 8 shows the structure of Palm OS 5, which consists of five layers: (i) applications, (ii) PACE, (iii) core Palm OS and licensee libraries, (iv) DAL, and (v) HAL.

The *Palm OS Developer Suite*, which is the official development environment and tool chain from PalmSource, is intended for software developers at all levels. It is a complete IDE (Integrated Development Environment) for:

- Protein applications (all ARM-native code) for Palm OS Cobalt and
- 68K applications for all shipping versions of the Palm OS.

Program 2 gives a Palm example, which displays the text "Hello, Mobile world!," an image, and a button "OK" on a Palm device. For how to create Palm OS applications, check Palm OS Developer Documentation at http://www.palmos.com/dev/support/docs/ . Figure 9 shows an emulator

Program 2. A Palm OS program displaying the text "Hello, Mobile world!"

```
// This header is from the Palm SDK and contains the needed refer-
ence
// materials for the use of Palm API and its defined constants.
#include  <PalmOS.h>

// The following IDs are from using Palm Resource Editor.
#define   Form1  1000
#define   OK     1003

// ----------------------------------------------------------------------
// PilotMain is called by the startup code and implements a simple
// event handling loop.
// ----------------------------------------------------------------------
UInt32  PilotMain( UInt16 cmd, void *cmdPBP, UInt16 launchFlags ) {
  short      err;
  EventType  e;
  FormType   *pfrm;

  if ( cmd == sysAppLaunchCmdNormalLaunch ) {
    // Displays the Form with an ID 1000.
    FrmGotoForm( Form1 );

    //  Main event loop
    while( 1 ) {
      //  Doze until an event arrives or 100 ticks are reached.
      EvtGetEvent( &e, 100 );
      //  System gets first chance to handle the event.
      if ( SysHandleEvent( &e ) )  continue;
      if ( MenuHandleEvent( (void *) 0, &e, &err ) )   continue;

      switch ( e.eType ) {
        case ctlSelectEvent:
          if ( e.data.ctlSelect.controlID == OK )
            goto  _quit;
          break;
        case frmLoadEvent:
          FrmSetActiveForm( FrmInitForm( e.data.frmLoad.formID ) );
          break;
        case frmOpenEvent:
          pfrm = FrmGetActiveForm( );
          FrmDrawForm( pfrm );
          break;
        case menuEvent:
          break;
        case appStopEvent:
          goto  _quit;
          break;
        default:
          if ( FrmGetActiveForm( ) )
            FrmHandleEvent( FrmGetActiveForm( ), &e );
          break;
      }
    }
_quit:
    FrmCloseAllForms( );
  }
  return 0;
}
```

Figure 9. A screenshot of the execution results of the Palm program in Program 2

displaying the execution results. Since this article is never intended to be a comprehensive Palm programming guide, further references for Palm OS SDK (Software Development Kit) can be found from the Internet (PalmSource Inc., 2004a, 2004b, 2004c).

FUTURE TRENDS

A number of mobile operating systems, as the previous list, with small footprints and reduced storage capacity have emerged to support the computing-related functions of mobile devices. For example, Research In Motion Ltd's BlackBerry 8700 smartphone uses RIM OS and provides Web access, as well as wireless voice, address book, and appointment applications (Research In Motion Ltd., 2005). Because the handheld device is small and has limited power and memory, the mobile OSes' requirements are significantly less than those of desk or laptop OSes. Although a wide range of mobile handheld devices are available in the market, the operating systems, the hub of the devices, are dominated by just few major organizations. The following two lists show the operating systems used in the top brands of smart cellular phones and PDAs in descending order of market share:

- **Smart cellular phones:** Symbian OS, Linux, RIM OS, Palm OS, Windows Mobile-based Smartphone, and others (Symbian Ltd., 2006).
- **PDAs:** Microsoft Pocket PC, RIM OS, Palm OS, Symbian OS, Linux, and others (Gartner Inc., 2005).

The market share is changing frequently and claims concerning the share vary enormously. It is almost impossible to predict which will be the ultimate winner in the battle of mobile operating systems. Because each mobile OS has its own unique development tools and programming languages, handheld computing and programming, especially the client-side ones, become difficult. A dominant mobile OS in the future may not be all that bad as the Microsoft's Windows do to the PCs.

CONCLUSION

Mobile commerce is a coming milestone after electronic commerce blossoming in the late 1990s. However, it is also commonly admitted that the development in this field is constrained. There are some considerable barriers waiting to be overcome. One of the barriers is most software engineers are not familiar with handheld programming, which is the programming for handheld devices such as smart cellular phones and PDAs (personal digital assistants). This chapter attempts to give a study of handheld computing to help software engineers better understand this subject. Client-side handheld computing is to use handheld devices to perform handheld operations, which do not need the supports of server-side computing. Various environments/languages are available for client-side handheld computing and programming. Five of the most popular are:

- **BREW:** It is created by Qualcomm Inc. for CDMA-based smartphones.
- **Java ME:** Java ME is an edition of the Java platform that is targeted at small, standalone or connectable consumer and embedded devices.
- **Palm OS:** It is a fully ARM-native, 32-bit operating system running on handheld devices.
- **Symbian OS:** Symbian OS is an industry standard operating system for smartphones, a joint venture originally set up by Ericsson, Nokia, and Psion.
- **Windows Mobile:** Windows Mobile is a compact operating system for handheld devices based on the Microsoft Win32 API. It is a small version of Windows, and features many "pocket" versions of popular Microsoft applications, such as Pocket Word, Excel, Access, PowerPoint, and Internet Explorer.

They apply different approaches to accomplishing the development of handheld applications and it is almost impossible to predict which approaches will dominate the client-side handheld computing in the future, as the Windows to desktop PCs. Most client-side handheld programming languages are a version of either C/C++ or Java. This chapter also shows application examples of Java ME, a version of Java, and Palm OS, using a version of C.

REFERENCES

Finan, T. (2002). *Developing applications for Windows Mobile-based Smartphones*. Retrieved September 12, 2006, from http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnppcgen/html/devappsp.asp

Gartner Inc. (2005). *Gartner Says Worldwide PDA Shipments Increased 32 Percent in the Second Quarter of 2005*. Retrieved January 13, 2006, from http://www.gartner.com/press_releases/asset_133230_11.html

Hu, W.-C., Yeh, J.-h., Chu, H.-J. Chu, & Lee, C.-w. Lee. (2005). Internet-enabled mobile handheld devices for mobile commerce. *Contemporary Management Research*, *1*(1), 13-34.

Kiely, D. (2001). Wanted: programmers for handheld devices. *IEEE Computer*, *34*(5),12-14.

Microsoft Corp. (2005). *What's New for Developers in Windows Mobile 5.0*? Retrieved June 21, 2006, from http://msdn.microsoft.com/mobility/windowsmobile/howto/documentation/default.aspx?pull=/library/en-us/dnppcgen/html/whatsnew_wm5.asp

PalmSource Inc. (2002). *Why PalmOS*? Retrieved June 23, 2006, from http://www.palmsource.com/palmos/Advantage/index_files/v3_document.htm

PalmSource Inc. (2004a). *Palm OS programmer's API reference*. Retrieved August 15, 2006, from http://www.palmos.com/dev/support/docs/palmos/PalmOSReference/ReferenceTOC.html

PalmSource Inc. (2004b). *Palm OS Programmer's Companion, Vol. I*. Retrieved February 21, 2006, from http://www.palmos.com/dev/support/docs/palmos/PalmOSCompanion/CompanionTOC.html

PalmSource Inc. (2004c). *Palm OS Programmer's Companion, Vol. II*. Retrieved February 21, 2006, from http://www.palmos.com/dev/support/docs/palmos/PalmOSCompanion2/Companion2TOC.html

Qualcomm Inc. (2003). *BREW and J2ME—A Complete Wireless Solution for Operators Committed to Java*. Retrieved February 12, 2006, from http://brew.qualcomm.com/brew/en/img/about/pdf/brew_j2me.pdf

Research In Motion Ltd. (2005). *BlackBerry Application Control—An Overview for Application Developers*. Retrieved January 05, 2006, from http://www.blackberry.com/knowledgecenterpublic/livelink.exe/fetch/2000/7979/1181821/832210/BlackBerry_Application_Control_Overview_for_Developers.pdf?nodeid=1106734&vernum=0

Sun Microsystem Inc. (2002a). *Java 2 Platform, Micro Edition*. Retrieved January 12, 2006, from http://java.sun.com/j2me/docs/j2me-ds.pdf

Sun Microsystem Inc. (2002b). *Mobile Information Device Profile Specification 2.0*. Retrieved March 25, 2006, from http://jcp.org/aboutJava/communityprocess/final/jsr118/

Sun Microsystem Inc. (2004). *J2ME Wireless Toolkit 2.2—User's Guide*. Retrieved April 21, 2006, from http://java.sun.com/j2me/docs/wtk2.2/docs/UserGuide.pdf

Symbian Ltd. (2005). *Symbain OS Version 9.2*. Retrieved May 20, 2006, from http://www.symbian.com/technology/symbianOSv9.2_ds_0905.pdf

Symbian Ltd. (2006). *Fast Facts*. Retrieved June 12, 2006, from http://www.symbian.com/about/fastfacts/fastfacts.html

KEY TERMS

BREW (binary runtime environment for wireless): BREW is an application development platform created by Qualcomm for CDMA (code division multiple access)-based mobile phones.

Client-Side Handheld Programming: It is design and development of handheld software such as Java ME programs that reside on the handheld devices.

Handheld Computing: It is to use handheld devices such as smart cellular phones and PDAs (personal digital assistants) to perform wireless, mobile, handheld operations such as personal data management and making phone calls.

Java ME (Java Platform, Micro Edition): Java ME provides an environment for applications running on consumer devices, such as mobile phones, PDAs, and TV set-top boxes, as well as a broad range of embedded devices.

Mobile Handheld Devices: They are small general-purpose, programmable, battery-powered computers, but they are different from desk or laptop computers mainly due to the following special features: (i) limited network bandwidth, (ii) small screen/body size, and (iii) mobility.

Palm OS: Palm OS, developed by PalmSource Inc., is a fully ARM-native, 32-bit operating system running on handheld devices. Two major versions of Palm OS are currently under development: Palm OS Garnet and Palm OS Cobalt.

Server-Side Handheld Programming: It is design and development of handheld software such as CGI programs that reside on the servers.

Symbian OS: Symbian is a software licensing company that develops and supplies the open operating system—Symbian OS—for data-enabled mobile phones

Windows Mobile: It is a compact operating system for mobile devices based on the Microsoft Win32 API and is designed to be similar to desktop versions of Windows.

Category: Software & Systems Design

Handling Extemporaneous Information in Requirements Engineering

Gladys N. Kaplan
Universidad Nacional de La Matanza, Argentina & Universidad Nacional de La Plata, Argentina

Jorge H. Doorn
Universidad Nacional de La Matanza, Argentina & Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina

Graciela D. S. Hadad
Universidad Nacional de La Matanza, Argentina & Universidad Nacional de La Plata, Argentina

INTRODUCTION

The key of the success or failure of a software project depends on solving the right problem (Rumbaugh, 1994; Sawyer, 2005). Thus, software requirements should be correct, unambiguous, consistent, and as complete as possible (Institute of Electrical and Electronics Engineers [IEEE], 1998). Errors in requirements raise software development and maintenance costs notoriously (Katasonov & Sakkinen, 2006). The later the requirement error is detected, the higher the correction cost turns out to be. Error correction costs have been widely studied by many researchers (Bell & Thayer, 1976; Davis, 1993). Errors in requirements may be due to several reasons such as poor communication among requirements engineers, clients, and users; poor or nonexistent requirements validation; and the level of sternness of the models being used, especially to model relevant information captured from the universe of discourse (UofD).

Requirements engineering is the area of software engineering responsible for proposing and developing solutions to elicit, model, and analyze requirements by means of heuristics, guidelines, models, and processes which tend to requirements' completeness, quality, correctness, and consistency. Many proposals have been put forward by many researchers (Bubenko & Wrangler, 1993; Jacobson, Christerson, Jonsson, & Overgaard, 1992; Leite & Oliveira, 1995; Macaulay, 1993; Reubenstein & Waters, 1991).

BACKGROUND

Eliciting and modeling software requirements or related information are two different but highly related activities (Hull, Jackson, & Dick, 2005; Zowghi & Coulin, 2005). They may be coupled in several ways, being canonicals, such as in model-driven elicitation and elicitation-driven modeling. In the former, the requirements engineer tries to capture only the information that he or she needs for the model under construction. In the latter, the requirements engineer creates all models at the same time recording every piece of information gathered in the model it belongs to. Each of these approaches has advantages and drawbacks.

If the information is elicited for a given model, the requirements engineer pays attention only to some part of the things he or she is seeing, reading, or hearing. Then, he or she will discard any information that is not focus oriented. When the requirements engineer starts the creation of another model, he or she will change the focus and perhaps will now pay attention to information previously disregarded, provided that he or she comes across the same information. Unfortunately, this does not always happen, especially when the source of information is people. In other words, model-driven elicitation tends to make completeness difficult.

If all information obtained is registered at the same time, every model of the process is opened at the same time, and also, none are finished during an important period. A lack of coherence among models, poor understanding of the information gathered, and misplaced information are the main drawbacks of this approach. However, the loss of information is minimized.

Most researchers explicitly or implicitly prefer model-driven elicitation over elicitation-driven modeling (Leite, Hadad, Doorn, & Kaplan, 2005; Loucopoulos & Karakostas, 1995; Potts, Takahashi, & Antón, 1994). This means that model-driven elicitation has to deal with the risk of information loss. Looking closely into such risk, it can be seen that regardless of the elicitation technique, whether involving document reading, interviews, observation, or any other method (Goguen & Linde, 1993), the requirements engineer does not get exactly whatever he or she is looking for at any time in the information gathering activity (Faulk, 1996). He or she will need some of the information captured in the later stages of the process; however, he or she also needs some of that information in advance. In other words,

some of the information gathered is out of time (earlier or delayed). Dealing with the risk of loss of information means dealing with extemporaneous information (EI).

The research on which this chapter is founded was collected using a given process (Doorn, Hadad, & Kaplan, 2002; Leite et al., 1997; Young, 2004). However, the conclusions obtained can be applied to any requirements engineering process provided that it builds more than one model, having some degree of precedence among them.

In this article, the existence of extemporaneous information is studied and a proposal for its appropriated handling is given.

ADVANCED INFORMATION

Every phase of the requirements engineering process has its own specific objective. Although the requirements engineer tries to adhere to this objective, usually he or she comes across pieces of unexpected information that will be later necessary in the process. This advanced information (AI) may come from any source of information. However, most likely, AI comes from people.

Two main factors influence the quality and the quantity of Advanced Information in interviews.

Expectations on the new software: When the interviewed person is waiting for the new software to fix some organizational problem, he or she will be biased to describe his or her expectations even when the requirements engineer is trying to elicit knowledge about current business practices or even trying to build a glossary of the macrosystem. On the other hand, when the interviewed person does not have any expectation on the software system, he or she will not be an important source of AI.

Relative position in the organization: When interviewing a person in a relatively high position in the organization (directors, managers), the requirements engineer should be prepared to deal with abundant AI usually expressed as objectives and goals with an important degree of abstraction. On the other hand, operative people in the organization tend to provide less AI.

It should also be taken into account that AI may increase accordingly with the amount of changes planned for the business process after software installation. To estimate such amount of changes, the following factors should be carefully watched, among others.

Internal factors:

- Quality improvement projects
- Outsourcing projects
- Opening of new business lines
- Production technology changes
- Production volume increases
- Products or services changes
- Global organization objective changes

External factors:

- Sociocultural changes affecting the organization
- Economic changes affecting business
- Customer preferences changes

When accessing written sources of information, the risk to lose AI reaches a peak when the document has a moderate amount of AI. This becomes obvious considering that every document with much AI will be tagged to be read later in the process.

The requirements engineer cannot be in any way considered as a passive actor in the process of knowledge elicitation. He or she may usually have useful ideas about the context in which the software system will be involved in the future. On the other hand, it is very well known that short memory might be unfair to creative people in all activities. Sometimes, after having a good idea, the creator recalls just that: He or she has had a good idea. This falls into the paradigm of AI, although it was not originated in clients or users.

It is hard to determine at that early stage whether the elicited information or the requirements engineer's ideas are valuable or not. Perhaps they are the stub of what will be later an important requirement or perhaps they will become unplayable in the future context of the software system.

While model-driven elicitation is applied, the requirements engineer is trying to focus on his or her main current objective, and AI is actually felt as a disturbance; the risk is either to fail to collect the desired information or to lose the main track. The problem, visualized in this way, is very similar to visiting every node in a graph with bifurcations. Thus, the solution should also be very alike. The requirements engineer must keep attached to the main activity as much as possible, but at the same time he or she should find a way to register the minimum information needed to be able to follow the secondary track later.

When registering the minimum information needed to follow the secondary track, an ad hoc notepad could be used or a more structured document might be also chosen.

EXTEMPORANEOUS INFORMATION

In order to choose between a preplanned and an ad hoc registering strategy, it is valuable to quote Faulk (1996) who stated that the use of regular and predictable structures with limited information eases the comprehension and visualization of large quantities of information. This article summarizes the advantages of the use of a minimum card that allows registering the basic information required to

recover the EI when it can be appropriately understood and processed. This card introduces a very small disturbance in the requirements engineering activity when filled in, and it is very valuable when recovered.

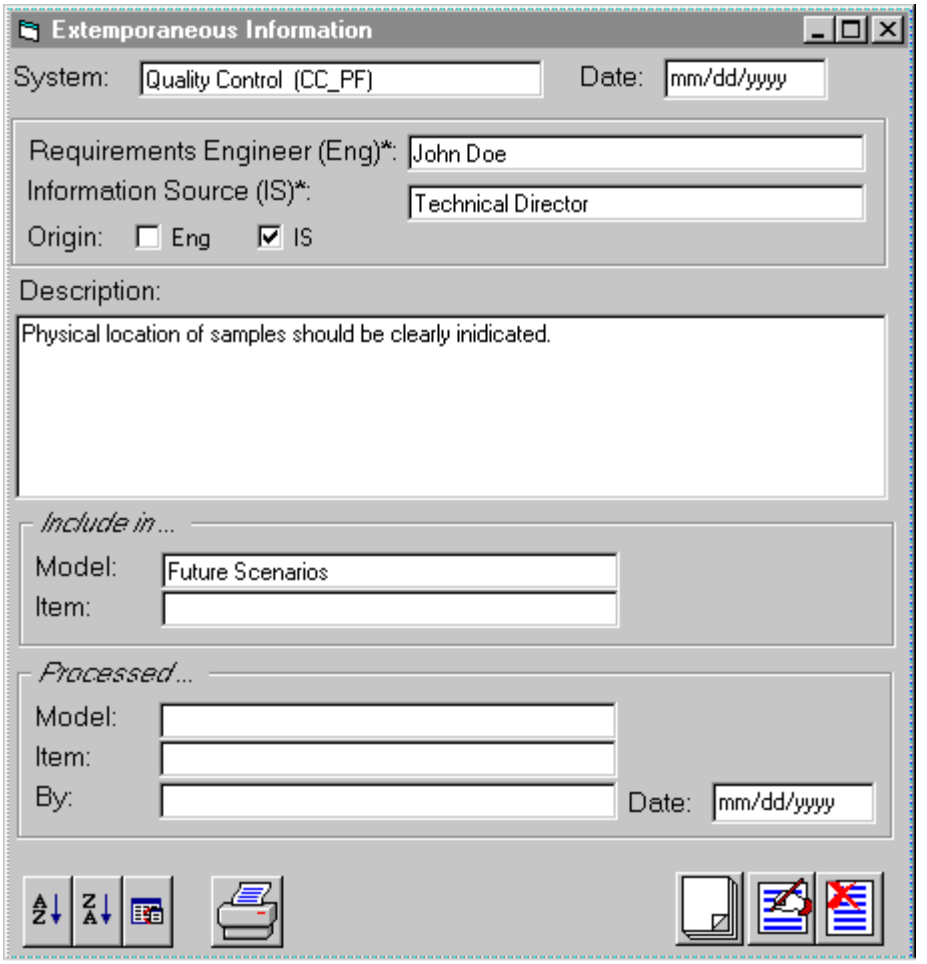
Every card should contain a brief description of the EI, specifying the source of information. If the requirements engineer can figure out how to use EI, the *include in* field could also be filled out (see Figure 1). The *system*, *date*, and *requirements engineer* fields may be automatically completed from previous cards, and other fields should be left blank until the card is used.

Although this study was motivated by the existence of advanced information during the requirements engineering process, the application of this approach in actual cases has shown that retarded information appears almost as much as advanced information. Then, it becomes more appropriate to use extemporaneous information when referring to both.

SOME PRACTICAL EXPERIENCES

During the testing of the registering of extemporaneous information in actual cases, the initial hypothesis was validated in the sense that for the chosen cases, EI could be adequately managed. However, a new and very important unplanned advantage showed up. The existence of an ordered registering of EI introduces an important increment in the points of view from which EI originates, such as security, availability, cost, and customer satisfaction, among many others. In the mainstream of the requirements engineering process, some of these points of view are disregarded. This is not a minor issue; on the contrary, when the requirements engineer begins to evaluate the collected EI cards, he or she puts himself or herself very close to the point of view that originated the EI card. This opens many new lines of thought about unconsidered aspects, becoming an important

Figure 1. EI card



contribution to the completeness of the comprehension and modeling of the problem.

Quantitative studies in two different cases reported 6.7% and 5.2% of requirements elicited by EI cards. These requirements would not have been picked by other means during the requirement phase. There is no way to know how soon such information will show up to the development team.

FUTURE TRENDS

Current experience should be extended to more cases by different requirements engineers. It is expected that some process or personal-flavor changes on the EI card will be introduced. However, it seems that the basic idea of the use of an EI card during any requirements engineering process is useful. A few project keywords may be added to later allow queries on them.

CONCLUSION

The problem of the appearance of EI has been studied in this article, which paid more attention to AI since it is hard to manage. The cases studied showed that EI is an important aspect in the process of UofD modeling, especially future UofD. Most EI deals with the context in which the software system will run. Retarded information will introduce a perturbation factor as it forces the requirements engineer to update already finished models, but it does not carry the risk of losses.

Another important conclusion is that some of the AI surpassed the requirements engineer's expectations in relation to the model in which it was finally included. Consequently, the inclusion point had to be, from a new point of view, deeply analyzed to understand the relationship with previously registered knowledge about the UofD. This means that the insertion of EI in any model is a far more complex task than merely inserting a line of text in a model. In other words, qualitative evaluation shows that EI has more importance than quantitative figures may indicate.

REFERENCES

Bell, T. E., & Thayer, T. A. (1976). Software requirements: Are they really a problem? In *Second International Conference on Software Engineering* (pp. 61-68).

Bubenko, J. A., & Wrangler, B. (1993). Objective driven capture of business rules and information systems requirements. In *Proceedings of IEEE Conference on Systems, Man and Cybernetics* (pp. 670-677).

Davis, A. M. (1993). *Software requirements: Objects, functions and states.* Englewood Cliffs, NJ: Prentice Hall.

Doorn, J. H., Hadad, G. D. S., & Kaplan, G. N. (2002). Comprendiendo el universo de discurso futuro. In *Anales del Workshop en Ingeniería de Requisitos*, Valencia, Spain.

Faulk, S. (1996). Software requirements: A tutorial. *Software Engineering*, pp. 82-103.

Goguen, J. A., & Linde. (1993). Techniques for requirements elicitation. In *Proceedings of the International Symposium on Requirements Engineering* (pp. 152-164). IEEE Computer Society.

Hull, E., Jackson, K., & Dick, J. (2005). *Requirements engineering.* Springer.

Institute of Electrical and Electronics Engineers (IEEE). (1998). *IEEE recommended practice for software requirements specifications* (Std 830-1998). Author.

Jackson, M. (1995). *Software requirements & specifications: A lexicon of practice, principles and prejudices.* Addison Wesley, ACM Press.

Jacobson, Y., Christerson, M., Jonsson, P., & Overgaard, G. (1992). *Object-oriented software engineering: A use case driven approach.* New York: Addison Wesley, ACM Press.

Katasonov, A., & Sakkinen, M. (2006). Requirements quality control: A unifying framework. *Requirements Engineering Journal, 11*(1), 42-57.

Leite, J. C. S. P., Hadad, G. D. S., Doorn, J. H., & Kaplan, G. N. (2005). Scenario inspections. *Requirement Engineering Journal, 10*(4), 1-21.

Leite, J. C. S. P., & Oliveira, A. P. A. (1995). A client oriented requirements baseline. In *Proceedings of the Second IEEE International Symposium on Requirements Engineering* (pp. 108-115). IEEE Computer Society Press.

Leite, J. C. S. P., Rossi, G., Balaguer, F., Maiorana, V., Kaplan, G. N., Hadad, G. D. S., et al. (1997). Enhancing a requirements baseline with scenarios. *Requirements Engineering Journal, 2*(4), 184-198.

Loucopoulos, P., & Karakostas, V. (1995). *System requirements engineering.* London: McGraw-Hill.

Macaulay, L. (1993). Requirements capture as a cooperative activity. In *IEEE International Symposium on Requirement Engineering* (pp. 174-181). San Diego, CA: IEEE Computer Society Press.

Potts, C., Takahashi, K., & Antón, A. I. (1994). Inquiry-based requirements analysis. *IEEE Software, 11*(2), 21-32.

Reubenstein, H. B., & Waters, R. C. (1991). The requirements apprentice: Automated assistance for requirements acquisition. *IEEE Transaction on Software Engineering, 17*(3), 226-240.

Rumbaugh, J. (1994). Getting started: Using use case to capture requirements. *Journal on Object-Oriented Programming*, pp. 8-12.

Sawyer, P. (2005). Maturing requirement engineering process maturity model. In J. L. Maté & A. Silva (Eds.), *Requirement engineering for sociotechnical systems* (pp. 84-99). Information Science Publishing.

Young, R. R. (2004). *The requirements engineering handbook.* Artech House.

Zowghi, D., & Coulin, C. (2005). Requirement elicitation: A survey of techniques, approaches and tools. In A. Aurum & C. Wohlin (Eds.), *Engineering and managing software requirements* (pp. 19-46). Springer.

KEY TERMS

Advanced Information: It is information acquired when it is not yet needed.

Elicitation: Elicitation is the activity of acquiring knowledge of a given kind during the requirements engineering process. There exist several techniques for elicitation such as interviews, observation, and document reading, among others.

Extemporaneous Information: It is information acquired before or after the moment it is needed.

Requirements Engineering: It is an area of software engineering that is responsible for acquiring and defining the software system's needs. The aim of requirements engineering is to improve the way in which the services should behave in the future. It covers all activities involved in discovering, understanding, modeling, analyzing, and maintaining the set of requirements for a software system.

Requirements Modeling: It is the activity that represents, organizes, and registers the information gathered during elicitation. The model itself may be composed by more than one representation.

Retarded Information: It is information acquired after the moment it is needed.

Sources of Information: These include documents, key people, books, and so forth that can provide useful information about the subject matter under study.

Universe of Discourse: It is the environment in which the software artifact will be used. It includes the macrosystem and any other source of knowledge.

Heuristics in Medical Data Mining

Susan E. George
University of South Australia, Australia

HISTORICAL PERSPECTIVE

Deriving—or discovering—information from data has come to be known as data mining. Within health care, the knowledge from medical mining has been used in tasks as diverse as patient diagnosis (Brameier et al., 2000; Mani et al., 1999; Cao et al., 1998; Henson et al., 1996), inventory stock control (Bansal et al., 2000), and intelligent interfaces for patient record systems (George at al., 2000). It has also been a tool of medical discovery itself (Steven et al., 1996). Yet, it remains true that medicine is one of the last areas of society to be "automated," with a relatively recent increase in the volume of electronic data, many paper-based clinical record systems in use, a lack of standardisation (for example, among coding schemes), and still some reluctance among health-care providers to use computer technology. Nevertheless, the rapidly increasing volume of electronic medical data is perhaps one of the domain's current distinguishing characteristics, as one of the last components of society to be "automated."

Data mining presents many challenges, as "knowledge" is automatically extracted from data sets, especially when data are complex in nature, with many hundreds of variables and relationships among those variables that vary in time, space, or both, often with a measure of uncertainty, as is common within medicine. Cios and Moore (2001) identified a number of unique features of medical data mining, including the use of imaging and need for visualisation techniques, the large amounts of unstructured nature of free text within records, data ownership and the distributed nature of data, the legal implications for medical providers, the privacy and security concerns of patients requiring anonymous data used, where possible, together with the difficulty in making a mathematical characterisation of the domain.

Strictly speaking, many ventures within medical data mining are better described as exercises in "machine learning," where the main issues are, for example, discovering the complexity of relationships among data items, or making predictions in light of uncertainty, rather than "data mining," in large, possibly distributed, volumes of data that are also highly complex. Large data sets mean not only increased algorithmic complexity but also often the need to employ special-purpose methods to isolate trends and extract "knowledge" from data. However, medical data frequently provide just such a combination of vast (often distributed) complex data sets.

Heuristic methods are one way in which the vastness, complexity, and uncertainty of data may be addressed in the mining process. A heuristic is something that aids discovery of a solution. Artificial intelligence (AI) popularised the heuristic as something that captures, in a computational way, the knowledge that people use to solve everyday problems. AI has a classic graph search algorithm known as A* (Hart et al., 1968), which is a heuristic search (under the right conditions). Increasingly, heuristics refer to techniques that are inspired by nature, biology, and physics. The genetic search algorithm (Holland, 1975) may be regarded as a heuristic technique. More recent population-based approaches have been demonstrated in the Memetic Algorithm (Moscato, 1989), and specific modifications of such heuristic methods in a medical mining context can be noted (Brameier et al., 2000).

Aside from the complexity of data with which the medical domain is faced, there are some additional challenges. Data security, accuracy, and privacy are issues within many domains, not just the medical (Walhstrom et al., 2000). Also, while ethical responsibility is an issue in other contexts, it is faced by the medical world in a unique way, especially when heuristic methods are employed. One of the biggest ethical issues concerns what is done with the knowledge derived combined with a "forward-looking responsibility" (Johnson et al., 1995). Forward-looking responsibility is accountable for high-quality products and methods and requires appropriate evaluation of results and justification of conclusions.

George (2002) first identified and proposed a set of guidelines for heuristic data mining within medical domains. The proposed guidelines relate to the evaluation and justification of data-mining results (so important when heuristic "aids to discovery" are utilised that "may" benefit a solution) and extend to both where and how the conclusions may be utilised and where heuristic techniques are relevant in this field. The remainder of this article summarises some heuristic data-mining applications in medicine and clarifies those proposed guidelines.

BACKGROUND

First, we will explain some of the heuristic methods that have been employed in medical data mining, examining a range of application areas. We broadly categorise applications as clinical, administrative, and research, according to whether they are used (or potentially used) in a clinical context, are infrastructure related, or are exploratory, in essence. We also note that with the exception of some medical imaging applications and mining of electronic medical records, the databases are small.

There is a wide variety of automated systems that have been designed for diagnosis—systems that detect a problem, classify it, and monitor change. Brameier and Banzhauf (2000) described the application of linear genetic programming to several diagnosis problems in medicine, including tests for cancer, diabetes, heart conditions, and thyroid conditions. Their focus was upon an efficient algorithm that operates with a range of complex data sets, providing a population-based heuristic method that is based upon biological principles. Their heuristic method is based on an inspiration from nature about how "introns" (denoting DNA segments with information removed before proteins are synthesised) are used in generating new strings. They suggest that introns may help to reduce the number of destructive recombinations between chromosomes by protecting the advantageous building blocks from being destroyed by crossover. Massive efficiency improvements in the algorithm are reported.

An interesting administrative application of data mining in a medical context comes in the area of interfaces for electronic medical records systems that are appropriate for speedy, accurate, complete entry of clinical data. At the University of South Australia, George et al. (2000) reported on the use of a data-mining model underlying an adaptive interface for clinical data entry. As records are entered, a database is established from which predictive Bayesian models are derived from the diagnosis and treatment patterns. This heuristic is used to predict the treatment from new diagnoses that are entered, producing intelligent anticipation. The predictive model is also potentially incremental and may be re-derived according to physician practice. This application addresses issues in incremental mining, temporal data, and highly complex data with duplication, error, and nonstandard nomenclatures.

One interesting ongoing database mining project at Rutgers is the development of efficient algorithms for query-based rule induction, where users have tools to query, store, and manage rules generated from data. An important component of the research is a facility to remember past mining sessions, producing an incremental approach. They are using heuristics for efficiently "re-mining" the same or similar data in the face of updates and modifications. In their trials, a major insurance company was trying to explore anomalies in their medical claims database. The new data-mining techniques aided the isolation of high-cost claims and scenarios in each disease group that would lead to high-cost claims. They also identified characteristics of people who were likely to drop out of their health plans and locations where there were higher dropout rates. This is a general approach to mining, where information from prior mining is utilised in new mining to prevent the need to compute relationships from scratch every time data is added to the database. This is, naturally, a general approach to mining large-scale changing databases that may be considered in a variety of fields.

Medical data mining is a natural method of performing medical research, where new relationships and insights are discovered in human health. The University of Aberdeen address the problem of mammographic image analysis using neural nets together with conventional image analysis techniques to assist in the automated recognition of pathology in mammograms (Undrill, 1996). The group also addresses the use of genetic algorithms for image analysis, applying this powerful general optimisation technique to a variety of problems in texture segmentation and shape analysis in two-dimensional and three-dimensional images (Delibassis, 1996). Mining information from the data in these tasks must address many of the problems of finding patterns within large volumes of highly complex data.

Banerjee et al. (1998) described the use of data mining in medical discovery. They reported on a data-mining tool that uncovered some important connections between diseases from mining medical literature. The data-mining tool compared the article titles in various medical journals. Medical discoveries were made, such as the connection between estrogen and Alzheimer's disease, and the relationship between migraine headaches and magnesium deficiency. Ngan et al. (1999) reported on medical discovery using data mining based upon an evolutionary computation search for learning Bayesian networks and rules. They were able to discover new information regarding the classification and treatment of scoliosis as well as knowledge about the effect of age on fracture, diagnoses, and operations and length of hospital stays.

Kargupta and colleagues (1999) were interested in an epidemiological study that involved combining data from distributed sources. Their study investigated what affects the incidence of disease in a population, focusing upon hepatitis and weather. They illustrated the collective data-mining approach, emphasising the importance within medicine of merging data from heterogeneous sites. Their solution minimises data communication using decision-tree learning and polynomial regression. As more hospitals and general practitioners, pharmacists, and other health-care-related professions utilise electronic media, mining ventures are going to have to cope with mining across data sources. They will have to address issues such as those addressed by this study, such as minimising data exchange and adopting suitable heuristic approaches.

GUIDELINES FOR HEURISTIC MEDICAL DATA MINING

Responsibility is clearly an issue in medical data mining given the unique human arena in which the conclusions are outworked. If medical data-mining products are ever produced by "professionals" or are ever exploited "commercially," there may be serious legal consequences for their creators in the wake of harmful consequences from information produced. In the context of software engineering, the computer field seeks to promote high-quality software products, so too,

should data miners seek to guarantee high-quality data-mining techniques.

Johnson and Nissenbaum (1995) distinguished "backward-looking" responsibility from "forward-looking" responsibility. A "backward-looking" responsibility asks questions in the wake of a harmful event and seeks to discover who is to blame for the harm and who should be punished. It is often conducted in the context of discovering legal liability. The Therac-25 computer-controlled radiation treatment is a shocking example of a malfunction disaster that resulted in loss of life for people who were receiving computer-controlled radiation treatments for cancer. In contrast a "forward-looking" responsibility addresses the particular responsibilities in advance. It defines guidelines for creating quality products; measures the quality of the product; defines the method of evaluation, and the limitations and scope of the operation in advance of harmful incidents.

One of the biggest ethical issues in medical mining concerns what is done with the knowledge derived. There is tremendous potential for good in improving quality of human life, managing disease effectively, efficiently administering programs, and preserving life, but the same knowledge can also be put to less-constructive ends, or benefit only a few, or conform to the contemporary political agendas influenced by the philosophy of the age. Forward-looking responsibility requires not only making ethical uses of data but also ensuring the quality of automated techniques and knowing the limitations and scope of methods in advance of harmful consequences.

Crucial to forward-looking responsibility is a way to evaluate products. This is not as pertinent as when heuristic methods are utilised to derive that knowledge. Whatever is ultimately done with the conclusions, we know that heuristics do not guarantee "optimality" or even the "accuracy" or "validity" of the conclusion. Forward-looking responsibility within medical data mining will address, among other things, how knowledge is evaluated, how conclusions are justified, what is the scope of validity, and the limitations of "products." One of the best forms of evaluation for clinical data-mining solutions is a clinical trial. Another approach to evaluation makes use of benchmark data sets, where various techniques (heuristic and other) could be compared to assess quality and efficiency of solutions. Additionally, some types of specialist data may be invaluable resources for data-mining researchers. It is also important to define the scope, including justification of explanations, and limitations of systems, from technical to clinical applicability of algorithms (in terms of patient populations and other), especially under heuristic conditions.

CONCLUSION

This article has reviewed heuristic medical data mining and some of the applications of medical mining, identifying administrative, clinical, and medical areas of applicability, focusing on the guidelines for use of heuristics in such a field. Forward-looking responsibility is vital, focusing upon appropriate use of knowledge, a means to evaluate the heuristic solutions and assess the scope and limitations of the system, including explanations of the behaviour.

REFERENCES

Bansal, K., Vadhavkar, S., & Gupta, A. (2000). Neural networks based data mining applications for medical inventory problems. Retrieved September 21, 2000, from *http://scanner-group.mit.edu/htdocs/DATAMINING/Papers/paper.html*

Brameier, M., & Banzhaf, W. (2001). A comparison of linear genetic programming and neural networks in medical data mining. *IEEE Transactions on Evolutionary Computation*, 5(1), 17-26. Retrieved September 22, 2000, from *http://ls11-www.cs.uni-dortmund.de/people/banzhaf/ieee_taec.pdf*

Cao, C., Leong, T. Y., Leong, A. P. K., & Seow, F. C. (1998). Dynamic decision analysis in medicine: A data driven approach. *International Journal of Medical Informatics, 51*(1), 13–28.

Cios, K., & Moore. (2001). Medical data mining and knowledge discovery: Overview of key issues. In K. Cios (Ed.), *Medical data mining and knowledge discovery*. Heidelberg: Springer-Verlag.

Delibassis, K., & Undrill, P. E. (1996). Genetic algorithm implementation of stack filter design for image restoration. *IEE Proc. Vision, Image & Signal Processing, 143*(3), 177-183.

George, S. E. (2002). Heuristics and medical datamining (Chap. 13). In H. A. Abbass, R. A. Sarker, & C. S. Newton (Eds.), *Heuristics and optimisation for knowledge discovery* (pp. 226-240). Hershey, PA: Idea Group Publishing.

George, S. E., & Warren, J. R. (2000). Statistical modelling of general practice medicine for computer assisted data entry in electronic medical record systems. *International Journal of Medical Informatics, 57*(2–3), 77-89.

Hart, P. E., Nilsson, N. & Raphael, B.(1968). A formal basis for the heuristic determination of minimum cost paths. *IEE Transactions on SSC, 4*, 100-107.

Holland, J. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control and artificial intelligence*. Ann Arbor, MI: University of Michigan Press.

Johnson, D., & Nissenbaum, H. (1995). *Computers, ethics and social values*. Englewood Cliffs, NJ: Prentice Hall.

Kargupta, H., Park, B., Hershberger, D., & Johnson, E. (1999). Collective data mining: A new perspective toward distributed

data mining. In H. Kargupta & P. Chan (Eds.), *Advances in distributed data mining*. Cambridge, MA: AAAI/MIT Press.

Mani, S., Shankle, W., Dick, M., & Pazzani, M. (1999). Two-stage machine learning model for guideline development. *Artificial Intelligence in Medicine, 16*, 51-71.

Moscato, P. (1989). On evolution, search, optimization, genetic algorithms and martial arts: Towards Memetic algorithms. Caltech Concurrent Computation Program, C3P Report 826.

Ngan, P. S., Wong, M. L., Lam, W., Leung, K. S., & Cheng, J. C. Y. (1999). Medical data mining using evolutionary computation. *Artificial Intelligence in Medicine, 16*, 73-96.

Undrill, P. E., Gupta, R., Henry, S., Downing, M., & Cameron, G. G. (1996). Outlining suspicious lesions in mammography by texture focussing and boundary refinement. In M.H. Loew & K.M. Hanson (Eds.) *Proceedings SPIE Medical Imaging: Image Processing*, 2710 (pp. 301-310).SPIE, Newport Beach, CA.

Vajdic, S. M., Brooks, M. J., Downing, A., & Katz, H. E. (1996). AI and medical imagery: Strategy and evaluation of inexact relational matching. Retrieved September 21, 2000, from *http://www.csu.edu.au/ci/vol2/vajdic/vajdic.html*

Walhstrom, K., Roddick, J. F., & Sarre, R. (2000). *On the ethics of data mining*. Research Report ACRC-00-003, January 2000, School of Computer and Information Science, University of South Australia.

KEY TERMS

Backward-Looking Responsibility: When backward-looking, we seek to discover who is to blame in wake of a harmful event. There are frequently connotations of punishment, legal intervention, and determination of guilt.

Data Mining: Analysis of data using methods that look for patterns in the data, frequently operating without knowledge of the meaning of the data. Typically, the term is applied to exploration of large-scale databases in contrast to machine-learning methods that are applied to smaller data sets.

Data-Mining Guidelines: A set of standards by which medical data mining, in particular, might be conducted. This is a framework that adopts a forward-looking responsibility in the evaluation of methods and explanation of conclusions, especially in the context of heuristic methods (with outcomes that may be ill-defined). This extends not only to the methods of the data-mining procedure, the security and privacy aspects of data, but also to where and how the results of data mining are utilised, requiring that an ethical reference be made to the final purpose of the mining.

Forward-Looking Responsibility: Addresses the particular responsibilities of individuals, groups, and partners in advance of a product's use or a system's implementation; it defines guidelines for creating quality products, measures the quality of the products, and defines the method of evaluation, the limitations, and the scope of the operation in advance of harmful incidents.

Heuristic: From the Greek "heuriskein," meaning "to discover." A heuristic aids discovery, particularly the search for solutions in domains that are difficult and poorly understood. It is commonly known as a "rule of thumb." Unlike algorithms, heuristics do not guarantee optimal or even feasible solutions and frequently do not have a theoretical guarantee.

Medical data: Frequently demonstrate increased complexity (e.g., uncertainty) and occurrence in large volumes, possibly distributed over many sources. There may be images involved and a high frequency of nonstandardisation (especially in the context of electronic medical records) in coding schemes and other medical concepts utilised. Data security, accuracy, and privacy are particular issues with medical data, as are the ethical issues involved in what is done with the data.

Medical Data Mining: This is the application of data-mining methods to medical data, typically for clinical or administrative and medical research investigation use, particularly in epidemiological studies.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1322-1326, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

High-Performance Virtual Teams

Ian K. Wong
Queen's University, Canada

D. Sandy Staples
Queen's University, Canada

INTRODUCTION

In the past several decades, we have seen tremendous advancements in the development of communication technology. Since the invention of the Internet in 1969, there has been rapid development of Internet-based communication tools and technologies. This technology has revolutionized business practices by offering another important and effective channel for communication (Foo & Lim, 1997) and has allowed people to work on projects irrespective of their physical location. One resulting business practice that has been adopted in recent years is virtual teamwork. Virtual teams are groups of individuals who work together in different locations (i.e., are geographically dispersed), work at interdependent tasks, share responsibilities for outcomes, and rely on technology for much of their communication (Cohen & Gibson, 2003). The use of virtual teams has become widespread in organizations, and its use is expected to grow (Martins, Gilson, & Maynard, 2004; Powell, Piccoli, & Ives, 2004).

BACKGROUND

In addition to the basic definition of a virtual team, all virtual teams have important characteristics that contribute to their overall success. To analyze the characteristics of the team's situation, Cohen's (1994) model of team effectiveness can be used as an organizing framework. The model identifies strengths and weaknesses that readers can use to inform their own design and operations of effective virtual teams. According to Cohen, there are several broad characteristics that all potentially effect how successful the team will be at meeting its task, and are therefore worthy of examination. These characteristics are listed in Table 1 and will be examined in detail in the following paragraphs. Although Cohen's team effectiveness model is based on traditional teams (i.e., collocated), these characteristics have been found to be very important in empirical research on virtual teams (Pinsonneault & Caya, 2005; Staples & Cameron, 2004; Wong & Staples, 2004).

Table 1. Characteristics of virtual teams affecting team effectiveness

• Design of the team's task
• The characteristics of the members of the team
• The processes used by the team
• The organizational context of the team

TASK DESIGN

Appropriate task design can be a powerful motivator (Cohen, 1994). Both job characteristics theory (e.g., Hackman & Oldman, 1976, 1980) and sociotechnical theory (e.g., Cummings, 1978) suggest that group task design is critical for employee motivation, satisfaction, and performance. Both theories suggest that to positively impact performance and attitudes, the task should be designed according to the criteria specified in Table 2. The design of the virtual team and the structuring of its interactions in the early stages of team development have been found to help team members develop a shared language and shared understanding (Powell et al., 2004).

Job characteristics theory, which has fairly strong empirical support, suggests that task attributes influence effectiveness through their impact on critical psychological states such

Table 2. Task design criteria necessary to positively impact performance and attitudes

The task should be designed such that:
• A variety of skills are required (leadership, communication, different technical skills, etc.) such that a team of people are needed to work together to complete the overall task
• A whole and identifiable piece of work exists so that members can see the outcome of their efforts
• It is perceived to have significant impact on the lives of other people so that team members feel their work is important and are motivated to complete the task
• The team has considerable autonomy and independence in determining how the work will be done so that team members feel empowered and responsible for their actions
• The team is provided with regular and accurate feedback such that the team can understand how it is performing and make adjustments as needed

Table 3. Characteristics of the team members that affect team effectiveness

• The size of the team
• The stability of the team, in terms of turnover
• The skills of the members of the team
• The relative locations of the team members (i.e., their virtualness)
• The team members' beliefs about their team's capabilities
• The diversity of the team

as motivation and satisfaction with the work. For example, in a case study of one particular business development virtual team, team members commented that high satisfaction and motivation levels reflected the high perceived significance of the project (Wong & Staples, 2004). Positive motivation and satisfaction levels have a positive effect on the quality of the work and overall productivity of the team (i.e., an indirect effect exists between task design and productivity and quality) (Cohen, 1994). Also, the team must have autonomy in determining how their work will be done, because autonomy enhances worker attitudes, behaviors, and performance (Cohen & Bailey, 1997). Finally, when a remote worker receives managerial feedback in the form of advice and help, the worker's effectiveness increases (Staples, 2001). This would result in an increase in virtual team performance.

CHARACTERISTICS OF THE TEAM AND ITS MEMBERS

Team member characteristics that influence the success of a virtual team are listed in Table 3 and are described in more detail next (Cohen, 1994).

The size of the team can affect the ability of the team to do its task (Cohen, 1994). If the team is too big, higher coordination costs result (Goodbody, 2005). If the team size is too small, it will not have the resources needed to complete its work, and team members will be less likely to be committed to the team. The size of the team should also correspond to the stage of the project. For example, a virtual team developing a new product may need more human resources as the product moves from the design stage into the manufacturing stage.

Stability of team membership is necessary for team effectiveness. If turnover is high, time and effort will be spent orientating new members, performance norms will not develop, and performance will suffer. However, some turnover can be beneficial, in that it could revitalize a stagnant team and enhance creativity (Cohen, 1994).

The collective knowledge and skills of a team will greatly impact the team's ability to carry out its task. Such skills include technical skills, information systems (IS) skills, and interpersonal skills. Information systems skills are needed to use the information technology tools and systems that are available to communicate virtually and share information virtually, which is the norm given the lack of face-to-face interaction in virtual teams. Effective communication skills among team members are vital to the effectiveness of a virtual team (Jones, Oyund, & Pace, 2005).

The degree of virtuality (degree of team geographic distribution) could contribute to team effectiveness. Most research on virtual teams suggests that geographic distance among team members is detrimental to team performance (Lu, Watson-Mannheim, Chudoba, & Wynn, 2006). This is presumably due to reduced face-to-face contact, reduced opportunities to build social relationships, and the difficulties of communicating and coordinating work using communication technology rather than communicating face-to-face. A recent meta-analysis (Ortiz de Guinea, Webster, & Staples, 2005) did find evidence that virtualness negatively impacts team processes (such as communication), although the total effect of virtualness on team effectiveness (quality, productivity, and satisfaction) was positive. Therefore, it is especially important to design team processes well in highly virtual teams. Other research has also pointed to the benefits of virtual teaming, such as increased rigor in processes and formal documents (Delone, Espinosa, Lee, & Carmel, 2005).

Team performance beliefs have been found to be a strong predictor of group effectiveness in previous research (Cohen, 1994). For example, team beliefs, assessed via a concept called group potency, were found to be positively related to the commitment to the team, satisfaction with being part of the team, and motivation with the team's tasks (Staples & Cameron, 2004). Therefore, it is important that potency beliefs are high within a team. Team members should know each other's strengths and abilities and celebrate achievements together.

Diversity is a characteristic of virtual teams that may positively influence team effectiveness if managed properly. Virtual teams are often composed of team members from various locations with diverse cultural backgrounds. Diversity may benefit a team because it increases the collective knowledge and expertise of the team. However, team diversity can also pose challenges, and thus it would benefit team members to openly discuss individual differences and expectations in order for new team-level expectations to be established (Gibson & Cohen, 2003).

TEAM PROCESSES

There are several behavioral characteristics pertaining to team process that positively affect team effectiveness. These are coordination, caring (i.e., team spirit), sharing of expertise, and effectiveness of communications. According to Cohen

Table 4. Organizational context factors that affect team effectiveness

• Human resource policies that emphasize and reward teamwork
• Human resource policies that do not penalize (i.e., overlook) virtual workers
• The availability of training to work virtually, work effectively in teams, and carry out task responsibilities
• The amount of support management provides to the team
• The degree of autonomy the team is given to make decisions and take responsibility for their own actions
• The ability of the organization to share information that is needed to carry out tasks and make effective decisions

and Bailey (1997), how team members coordinate is an important characteristic of a team. Since team members are interdependent on each other to get their work done, coordinating interdependent tasks is critical. Good coordination among team members leads to working together without duplication and wasted efforts, and has been shown to predict higher performance in virtual teams (Powell et al., 2004). Caring about each other implies working together cohesively with energy and team spirit. This can motivate team members and foster commitment to the team goals, resulting in higher performance. The development of team cohesion/team spirit has been associated with higher levels of performance and satisfaction in virtual teams (Pinsonneault & Caya, 2005; Powell et al., 2004). Sharing and benefiting from others' knowledge and expertise is also important because it supports effective cross training and decision-making to fulfill interdependencies. Empirical results from both virtual and traditional teams support a positive link between sharing of knowledge and team outcomes (e.g., Cohen, 1994; Cummings, 2004; Hong, Doll, Nah, & Li, 2004; Majchrzak, Rice, King, Malhotra, & Ba, 1995). Good communication processes are required in order to make this possible. Effective communication processes were important in building a successful virtual team according to 84% of team members interviewed in six case studies of virtual teams (Staples, Wong & Cameron, 2004). Furthermore, research has found that the effectiveness and frequency of team communication may also help build a more cohesive team (Pinsonneault & Caya, 2005). These team process variables are part of most models of team effectiveness and have been found to be associated with group effectiveness in previous traditional team research (Cohen, 1994) and virtual team research (Ortiz de Guinea et al., 2005).

ORGANIZATIONAL CONTEXT

Lastly, the organizational context that a team works in can create the conditions for a team to be successful or for it to fail (Cohen, 1994). The team with the best internal processes may still perform poorly if it lacks the resources or information needed to do its task. A team will not be able to make good decisions without proper information, without sufficient training, and without adequate resources. The key variables that potentially interact to create an environment where the employee wants to be involved and can participate to complete their tasks effectively are listed in Table 4 and explained next.

According to Duarte and Snyder (2001), human resource policies must be designed and integrated in such a way that virtual team members are recognized, supported, and rewarded for their work. Cohen, Ledford, and Spreitzer (1996) found that management recognition was positively associated with team ratings of performance, trust in management, organizational commitment, and satisfaction for both self-directed and traditionally managed groups in organizations. As such, it is important that an effective reward system with performance measures is in place to reward results. Lurey and Raisinghani (2001) and Hertel, Konradt, and Orlikowski (2004) also suggest that it is important for the organization to reward high levels of *team* performance. If rewards are solely at the individual level, this does not stimulate the completion of interdependent tasks and the sharing of information and expertise.

Next, it is important that team members have access to continual online training and technical support (Duarte & Snyder, 2001; O'Hara-Devereaux, & Johansen, 1994). Training and team building is important because it ensures that employees develop the knowledge required to contribute to organizational performance (Cohen, 1994). It should be available to enable employees to develop the necessary skills and knowledge required to complete their tasks at hand. In addition, providing training to team members shows management's commitment to continual growth and development for members. Research has also shown that training in early stages of team development in particular helps members not only focus on performance, but also to provide structure for their interactions (Jarvenpaa, Shaw, & Staples, 2004).

Management support and the power structure are also key variables in attaining virtual team effectiveness. Teams

not only need to be well budgeted and resourced, but team members also want encouragement and symbolic gestures of appreciation. Symbolic gestures such as a pat on the back or having management verbally communicate that the team is doing a good job demonstrates respect for the team members; this can positively motivate the members. With respect to power structure, it has been shown that team members must have independence and decision-making capability to be successful in virtual groups (Lipnack & Stamps, 1997). In addition, according to Cohen and Bailey (1997), an organization needs to provide team members autonomy in their work. Worker autonomy has been shown to have clear benefits because it enhances worker attitudes, behaviors, and performance (whether measured objectively or rated subjectively by team members).

Finally, access to information is necessary for team members to effectively complete their tasks. A practical study of a project-based virtual team suggested that most of the information team members needed was held either by themselves or by other team members (Wong & Staples, 2004). Therefore, good communication between team members is essential for information to be easily accessed. In addition, resources need to be available for virtual communication to be possible and effective. These resources include information technology infrastructure and information technology tools that are needed to communicate and share information electronically in the virtual setting.

FUTURE TRENDS

The need to compete in a rapidly changing, hypercompetitive, and global marketplace is prompting many organizations to transform their organizational structures from large, hierarchical structures to agile, flexible, new structures (Morris, Marshall, & Rainer, 2001). Consequently, we will see a continuing trend of virtual teams emerging that have team members from multiple countries with different national cultures. Cross-cultural issues are especially important for virtual work (Pinsonneault & Boisvert, 2001); however, recent research has suggested that it remains to be seen what aspects of national culture specifically affect virtual teams (Webster & Staples, 2006). Therefore, future research could focus on the impact of cultural differences of team members in virtual teams, because this information would be useful for designing training program for members and for selecting members.

Furthermore, management will be continually trying to adopt new communication tools and technologies that have improvements in quality and capabilities. As such, training to use new technology will become especially important; how well team members can communicate and interact with each other depends very much on their ability to properly use the communication tools. Future research may deal with the effectiveness of new, emerging technologies and how they may positively or negatively affect virtual team performance. In addition, we need a better understanding of how the effect of different media combinations affects virtual team processes (Pinsonneault & Caya, 2005).

CONCLUSIONS

The success of a virtual team depends heavily on its characteristics. Specifically, the main characteristics are the team's task design, the characteristics of the team members, the processes used by the team, and the organizational context. Analyzing these characteristics can reveal a great deal about how well a team is functioning, and what potential it has to improve. An organization's management also can use these characteristics to effectively design its virtual teams and thereby give it the best chance for success possible. Therefore, it is in management's best interests when putting together a virtual team to carefully consider the characteristics of the team. This will increase the team's chance of success if the variables that yield positive outcomes are maximized.

REFERENCES

Cohen, S. G. (1994). Designing effective self-managing work teams. In M. M. Beyerlein, D. A. Johnson, & S. T. Beyerlein (Eds.), *Advances in interdisciplinary studies of work teams, volume 1, series of self-managed work teams* (pp. 67-102). Greenwich, CT: JAI Press.

Cohen, S. G., & Bailey, D. E. (1997). What makes teams work: Group effectiveness research from the shop floor to the executive suite. *Journal of Management, 23*(3), 239-290.

Cohen, S. G., & Gibson, C. B. (2003, April). *Putting the team back in virtual teams*. Paper presented at the 18th Annual Conference of the Society for Industrial/Organizational Psychology, Orlando, FL.

Cohen, S. G., Ledford, G. E., & Spreitzer, G. M. (1996). A predictive model of self-managing work team effectiveness. *Human Relations, 49*(5), 643-676.

Cummings, J. N. (2004). Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science, 50*(3), 352-364.

Cummings, T. G. (1978). Self-regulating work groups: A sociotechnical synthesis. *Academy of Management Review, 3*(3), 625-634.

Delone, W., Espinosa, J. A., Lee, G., & Carmel, E. (2005). Bridging global boundaries for IS project success. In *Pro-*

ceedings of the 38th Hawaii International Conference on System Science (p. 48b).

Duarte, D. L., & Snyder, N. T. (2001). *Mastering virtual teams: Strategies, tools, and techniques that succeed*. San Francisco: Jossey-Bass.

Foo, S., & Lim, E. (1997). A hypermedia database to manage World Wide Web documents. *Information and Management, 31*(5), 235-249.

Gibson, C. B., & Cohen, S. G. (2003). The last word: Conclusions and implications. In D. Truxillo (Ed.), *Virtual teams that work: Creating conditions for virtual team effectiveness* (pp. 403-421). San Francisco: John Wiley & Sons, Inc.

Gibson, B. G., Randel, A. E., & Earley, P. C. (2000). Understanding group efficacy: An empirical test of multiple assessment methods. *Group & Organizational Management, 25*(1), 67-97.

Goodbody, J. (2005). Critical success factors for global virtual teams. *Strategic Communication Management, 9*(2), 18-21.

Hackman, J. R., & Oldman, G. R. (1976). Motivation through the design of work: Test of a theory. *Organizational Behavior and Human Performance, 16*, 250-279.

Hackman, J. R., & Oldman, G. R. (1980). *Work redesign*. Reading, MA: Addison-Wesley Publishing Company.

Hertel, G., Konradt, U., & Orlikowski, B. (2004). Managing distance by interdependence: Goal setting, task interdependence, and team-based rewards in virtual teams. *European Journal of Work and Organizational Psychology, 13*(1), 1-28.

Hong, P., Doll, W. J., Nah, A. Y., & Li, X. (2004). Knowledge sharing in integrated product development. *European Journal of Innovation Management, 7*(2), 102-112.

Jarvenpaa, S. L., Shaw, T. R., & Staples, D. S. (2004). Towards contextualized theories of trust: The role of trust in global virtual teams. *Information Systems Research, 15*(3), 250-267.

Jones, R., Oyund, R., & Pace, L. (2005). *Working virtually: Challenges of virtual teams*. Hershey, PA: Cybertech Publishing.

Lipnack, J., & Stamps, J. (1997). *Virtual teams: Reaching across space, time, and organizations with technology.* New York: John Wiley & Sons.

Lu, M., Watson-Mannheim, M. B., Chudoba, K. M., & Wynn, E. (2006). Virtuality and team performance: Understanding the impact of variety of practice. *Journal of Global Information Technology Management, 9*(1), 4-23.

Lurey, J., & Raisinghani, M. (2001). An empirical study of best practices in virtual teams. *Information & Management, 38*(8), 523-544.

Majchrzak, A., Rice, R., King, N., Malhotra, A., & Ba, S. (1995). Computer-mediated inter-organizational knowledge-sharing: Insights from a virtual team innovating using a collaborative tool. *Information Resources Management Journal, 24*(4), 44-53.

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A temporally based framework and taxonomy of team processes, *The Academy of Management Review, 26*(3), 356-376.

Martins, L. L., Gilson, L. L., & Maynard, M. T. (2004). Virtual teams: What do we know and where do we go from here. *Journal of Management, 30*(6), 805-835.

Morris, S., Marshall, T., & Rainer, R. (2001). Impact of user satisfaction and trust on virtual team members. *Information Resources Management Journal, 15*(2), 22-30.

O'Hara-Devereaux, M., & Johansen, R. (1994). *Global work: Bridging distance, culture & time*. San Francisco: Jossey-Bass.

Ortiz de Guinea, A., Webster, J., & Staples, D. S. (2005, oCTOBER 12). *A meta-analysis of the virtual teams literature.* Presented at the Symposium on High Performance Professional Teams, Industrial Relations Centre, School of Policy Studies, Queen's University, Kingston, Ontario, Canada.

Pinsonneault, A., & Boisvert, M. (2001). The impacts of telecommuting on organizations and individuals: A review of the literature. In N. J. Johnson (Ed.), *Telecommuting and virtual offices: Issues & opportunities* (pp. 163-185). Hershey, PA: Idea Group Publishing.

Pinsonneault, A., & Caya, O. (2005). Virtual team: What we know, what we don't know. *International Journal of e-Collaboration, 1*(3), 1-16.

Powell, A., Piccoli, G., & Ives, B. (2004). Virtual teams: A review of current literature and directions for future research. *The DATA BASE for Advances in Information Systems, 35*(1), 6-36.

Staples, D. S. (2001). Making remote workers effective. In N. J. Johnson (Ed.), *Telecommuting and virtual offices: Issues & opportunities* (pp. 186-212). Hershey, PA: Idea Group Publishing.

Staples, D. S., & Cameron, A. F. (2004). Creating positive attitudes in virtual team members. In S. Godar & P. Ferris (Eds.), *Virtual & collaborative teams: Process, technologies, & practice* (pp. 76-98). Hershey, PA: Idea Group Publishing.

Staples, D. S., Wong, I. K., & Cameron, A. F. (2004). Best practices for virtual team effectiveness. In D. Pauleen (Ed.), *Virtual teams: Projects, protocols and processes* (pp. 160-185). Hershey, PA: Idea Group Publishing.

Webster, J., & Staples, D. S. (2006). Comparing virtual teams to traditional teams: An identification of new research opportunities. In J. J. Martocchio (Ed.), *Research in personnel and human resources management* (Vol. 25, pp. 183-218). Greenwich, CT: JAI Press.

Wong, I., & Staples, D. S. (2004). A virtual team in action—An illustration of a business development virtual team. In D. Pauleen (Ed.), *Virtual teams: Projects, protocols and processes* (pp. 91-114). Hershey, PA: Idea Group Publishing.

KEY TERMS

Group Potency: A collective belief in the capability of the group to meet a task objective (Gibson, Randel, & Earley, 2000).

Job Characteristics Theory: Task attributes influence effectiveness through their impact on critical psychological states such as motivation and satisfaction with the work.

Organizational Context: The conditions within the organization that a team works in that influences the successfulness of the team's activities and the involvement of the team members. Reward systems, level of management support, resources provided, and organizational culture all are important organizational factors that potentially affect a team's ability to succeed.

Task Design: The way key attributes of the task are arranged, in terms of the influence of these attributes on the effectiveness of a team in performing the task. Research has found key design attributes include the need for a variety of skills, the perceived importance of the task, the independence and autonomy given to people to determine how the task will be done, and the way task feedback is provided.

Team Characteristics: The composition of the team and the shared beliefs held within the team about the team. Team composition includes such things as the number of team members, the skills the team members collectively possess, and the stability of team membership.

Team Effectiveness: The ability of a team to perform its tasks on time, on budget, and with acceptable quality, as well as the satisfaction, motivation, and commitment of the team members.

Team Processes: "Members' interdependent acts that convert inputs to outcomes through cognitive, verbal, and behavioral activities directed toward organizing taskwork to achieve collective goals" (Marks, Mathieu, & Zaccaro, 2001, p. 357).

Virtual Team: A group of individuals who work at interdependent tasks, who share responsibility for outcomes, and who work together from different locations.

Virtuality/Virtualness: The degree to which team members are geographically distributed such that opportunities for meeting informally and/or face-to-face are reduced.

Highly Available Database Management Systems

Wenbing Zhao
Cleveland State University, USA

INTRODUCTION

In the Internet age, real-time Web-based services are becoming more pervasive every day. They span virtually all business and government sectors, and typically have a large number of users. Many such services require continuous operation, 24 hours a day, seven days a week. Any extended disruption in services, including both planned and unplanned downtime, can result in significant financial loss and negative social effects. Consequently, the systems providing these services must be made highly available.

A Web-based service is typically powered by a multi-tier system, consisting of Web servers, application servers, and database management systems, running in a server farm environment. The Web servers handle direct Web traffic and pass requests that need further processing to the application servers. The application servers process the requests according to the predefined business logic. The database management systems store and manage all mission-critical data and application states so that the Web servers and application servers can be programmed as stateless servers. (Some application servers may cache information, or keep session state. However, the loss of such state may reduce performance temporarily or may be slightly annoying to the affected user, but not critical.) This design is driven by the demand for high scalability (to support a large number of users) and high availability (to provide services all the time). If the number of users has increased, more Web servers and application servers can be added dynamically. If a Web server or an application server fails, the next request can be routed to another server for processing.

Inevitably, this design increases the burden and importance of the database management systems. However, this is not done without good reason. Web applications often need to access and generate a huge amount of data on requests from a large number of users. A database management system can store and manage the data in a well-organized and structured way (often using the relational model). It also provides highly efficient concurrency control on accesses to shared data.

While it is relatively straightforward to ensure high availability for Web servers and application servers by simply running multiple copies in the stateless design, it is not so for a database management system, which in general has abundant state. The subject of highly available database systems has been studied for more than two decades, and there exist many alternative solutions (Agrawal, El Abbadi, & Steinke, 1997; Kemme, & Alonso, 2000; Patino-Martinez, Jimenez-Peris, Kemme, & Alonso, 2005). In this article, we provide an overview of two of the most popular database high availability strategies, namely database replication and database clustering. The emphasis is given to those that have been adopted and implemented by major database management systems (Davies & Fisk, 2006; Ault & Tumma, 2003).

BACKGROUND

A database management system consists of a set of data and a number of processes that manage the data. These processes are often collectively referred to as database servers. The core programming model used in database management systems is called transaction processing. In this programming model, a group of read and write operations on a data set are demarcated within a transaction. A transaction has the following ACID properties (Gray & Reuter, 1993):

- **Atomicity:** All operations on the data set agree on the same outcome. Either all the operations succeed (the transaction commits) or none of them do (the transaction aborts).
- **Consistency:** If the database is consistent at the beginning of a transaction, then the database remains consistent after the transaction commits.
- **Isolation:** A transaction does not read or overwrite a data item that has been accessed by another concurrent transaction.
- **Durability:** The update to the data set becomes permanent once the transaction is committed.

To support multiple concurrent users, a database management system uses sophisticated concurrency control algorithms to ensure the isolation of different transactions even if they access some shared data concurrently (Bernstein, Hadzilacos, & Goodman, 1987). The strongest isolation can be achieved by imposing a serializable order on all conflicting read and write operations of a set of transactions so that the transactions appear to be executed sequentially. Two operations are said to be *conflicting* if both operations access the same data item, at least one of them is a write operation, and they belong to different transactions. Another popular isolation model is snapshot isolation. Under the snapshot

isolation model, a transaction performs its operations against a snapshot of the database taken at the start of the transaction. The transaction will be committed if the write operations do not conflict with any other transaction that has committed since the snapshot was taken. The snapshot isolation model can provide better concurrent execution than the serializable isolation model.

A major challenge in database replication, the basic method to achieve high availability, is that it is not acceptable to reduce the concurrency levels. This is in sharp contrast to the replication requirement in some other field, which often assumes that the replicas are single-threaded and deterministic (Castro & Liskov, 2002).

DATABASE HIGH AVAILABILITY TECHNIQUES

To achieve high availability, a database system must try to maximize the time to operate correctly without a fault and minimize the time to recover from a fault. The transaction processing model used in database management systems has some degree of fault tolerance in that a fault normally cannot corrupt the integrity of the database. If a fault occurs, all ongoing transactions will be aborted on recovery. However, the recovery time would be too long to satisfy the high availability requirement. To effectively minimize the recovery time, redundant hardware and software must be used. Many types of hardware fault can in fact be masked. For example, power failures can be masked by using redundant power supplies, and local communication system failures can be masked by using redundant network interface cards, cables, and switches. Storage medium failures can be masked by using RAID (redundant array of inexpensive disks) or similar techniques.

To tolerate the failures of database servers, several server instances (instead of one) must be used so that if one fails, another instance can take over. The most common techniques are database replication and database clustering. These two techniques are not completely distinct from each other, however. Database replication is typically used to protect against total site failures. In database replication, two or more redundant database systems operate in different sites — ideally in different geographical regions — and communicate with each other using messages over a (possibly redundant) communication channel. Database clustering is used to provide high availability for a local site. There are two competing approaches in database clustering. One uses a shared-everything (also referred to as shared-disk) design, such as the Oracle Real Application Cluster (RAC) (Ault & Tumma, 2003). The other follows a shared-nothing strategy, such as the MySQL Cluster (Davies & Fisk, 2006) and most DB2 shared database systems. To achieve maximum fault tolerance and hence high availability, one can combine database replication with database clustering.

Database Replication

Database replication means that there are two or more instances of database management systems, including server processes, data files, and logs, running on different sites. Usually one of the replicas is designated as the primary, and the rest of the replicas are backups. The primary accepts users' requests and propagates the changes to the database to the backups. In some systems, the backups are allowed to accept read-only queries. It is also possible to configure all replicas to handle users' requests directly. But doing so increases the complexity of concurrency control and the risk of more frequent transaction aborts.

Depending on how and when changes to the database are propagated across the replicas, there are two different database replication styles, often referred to as *eager* replication and *lazy* replication (Gray & Reuter, 1993). In eager replication, the changes (i.e., the redo log) are transferred to the backups synchronously before the commit of a transaction. In lazy replication, the changes are transferred asynchronously from the primary to the backups after the transactions have been committed. Because of the high communication cost, eager replication is rarely used to protect site failures where the primary and the backups are usually far apart. (Eager replication has been used in some shared-nothing database clusters.)

Eager Replication

To ensure strong replica consistency, the primary must propagate the changes to the backups within the boundary of a transaction. For this, a distributed commit protocol is needed to coordinate the commitment of each transaction across all replicas. The benefit for doing eager replication is that if the primary fails, a backup can take over instantly as soon as it detects the primary failure.

The most popular distributed commit protocol is the two-phase commit (2PC) protocol (Gray & Reuter, 1993). The 2PC protocol guarantees the atomicity of a transaction across all replicas in two phases. In the first phase, the primary (which serves as the coordinator for the protocol) sends a *prepare* request to all backups. If a backup can successfully log the changes, so that it can perform the update even in the presence of a fault, it responds with a "Yes" vote. If the primary collects "Yes" votes from all backups, it decides to *commit* the transaction. If it receives even a single "No" vote or it times out a backup, the primary decides to *abort* the transaction. In the second phase, the primary *notifies* the backups of its decision. Each backup then either commits or aborts the transaction locally according to the primary's decision and sends an acknowledgment to the primary.

As can be seen, the 2PC protocol incurs significant communication overhead. There are also other problems such as the potential blocking if the primary fails after all backups

have voted to commit a transaction (Skeen, 1981). Consequently, there has been extensive research on alternative eager replication techniques, for example, the epidemic protocols (Agrawal et al., 1997; Stanoi, Agrawal, & El Abbadi, 1998), and multicast-based approaches (Kemme & Alonso, 2000; Patino-Martinez et al., 2005). However, they have not been adopted by any major commercial product due to their high overhead or complexities.

Lazy Replication

Most commercial database systems support lazy replication. In lazy replication, the primary commits a transaction immediately. The redo log, which reflects the changes made for the recently committed transactions, is transferred to backups asynchronously. Usually, the backup replicas lag behind the primary by a few transactions. This means that if the primary fails, the last several committed transactions might get lost.

Besides the primary/backup replication approach, some database management systems allow a multi-primary configuration where all replicas are allowed to accept update transactions. If this configuration is used with lazy replication, different replicas might make incompatible decisions, in which case manual reconciliation is required.

Database Clustering

In recent years, database clustering has evolved to be the most promising technique to achieve high availability as well as high scalability (Ault & Tumma, 2003; Davies & Fisk, 2006). Database clustering, as the name suggests, uses a group of computers interconnected by a high-speed network. In the cluster, multiple database server instances are deployed. If one instance fails, another instance takes over very quickly so high availability is ensured.

Database clustering not only brings high availability, but the *scaling-out* capability as well. Scaling-out means that the capacity of a database management system can be dynamically increased by adding more inexpensive nodes while keeping the old equipment.

There are two alternative approaches in database clustering. One approach pioneered in Oracle RAC adopts a shared-everything architecture. A number of other products choose to use the shared-nothing architecture. Both approaches have their challenges and advantages.

Shared-Everything Cluster

In a shared-everything database cluster, all server instances share the same storage device, such as a storage area network. The cluster nodes typically connect to the shared storage device via a fiber channel switch or shared SCSI for fast disk I/O. The shared storage device must also have built-in redundancy such as mirrored disks to mask disk failures. To minimize disk I/O, all server instances share a common virtual cache space. The virtual cache space consists of local cache buffers owned by individual server instances. A number of background processes are used to maintain the consistency of the data blocks in the cache space. These processes are also responsible to synchronize the access to the cached data blocks because only one server instance is allowed to modify a data block at a time.

Each server instance has its own transaction logs stored in the shared disk. If a server instance fails, another server instance takes over by performing a roll-forward recovery using the redo log of the failed server instance. This is to ensure that the changes made by committed transactions are recorded in the database and do not get lost. The recovery instance also rolls back the transactions that were active at the time of the failure and releases the locks on the resources used by those transactions.

The shared-everything design makes it unnecessary to repartition the data, and therefore eases the tasks of cluster maintenance and management. However, this benefit does not come for free. The most prominent concern is the cost of inter-node synchronization. Unless high-speed interconnect is used and the workload is properly distributed among the server instances, the inter-node synchronization might limit the scalability of the cluster. Also, the requirement for a high-speed shared disk system also imposes a higher financial cost than using conventional disks.

Shared-Nothing Cluster

In a shared-nothing database cluster, each node runs one or more server instances and has its own memory space and stable storage. Essential to the shared-nothing approach, the data must be partitioned either manually or automatically by the database system across different nodes. Each partition must be replicated in two or more nodes to keep the desired redundancy level. Concurrency control and caching are carried out in each local node, and therefore they are more efficient than those in shared-everything clusters. However, to ensure the consistency of replicated data and fast recovery, the two-phase commit protocol is often used to ensure atomic commitment of the transactions in the cluster. Comparing with the shared-everything approach, the cost of inter-node synchronization is essentially replaced by that of distributed commit.

The shared-nothing approach faces the additional challenge of *split-brain syndrome* prevention (Birman, 2005). The split-brain syndrome may happen if the network partitions, and if each partition makes incompatible decisions on the outcome of transactions or their relative orders. To prevent this problem, typically only the main partition is allowed to survive. The minor partition must stop accepting

new transaction and abort active transactions. Usually, the main partition is the one that consists of the majority of the replicas or the one that contains a special node designated as the arbitration node (Davies & Fisk, 2006).

FUTURE TRENDS

Existing database systems are designed to tolerate process crash fault and hardware fault. However, considering the increased pace of security breaches, future database management systems must be designed to be intrusion tolerant — that is, they should provide high availability against a variety of security threats, such as the unauthorized deletion and alteration of database records, the disruption of distributed commit (may cause replica inconsistency), and the exposure of confidential information.

To make a database system intrusion tolerant, many fundamental protocols such as the 2PC protocol must be enhanced. There may also be a need to design special tamper-proof storage devices to protect data integrity (Strunk, Goodson, Scheinholtz, Soules, & Ganger, 2000). Even though there has been intensive research in this area (Castro & Liskov, 2002; Malkhi & Reiter, 1997; Mohan, Strong, & Finkelstein, 1983; Deswarte, Blain, & Fabre, 1991), the results have rarely been incorporated into commercial products yet. The primary barrier is the high commutation and communication cost, the complexity, and the high degree of replication required to tolerate malicious faults.

CONCLUSION

Database systems are the cornerstones of today's information systems. The availability of database systems largely determines the quality of service provided by the information systems. In this article, we provided a brief overview of the state-of-the-art database replication and clustering techniques. For many, a low-cost shared-nothing database cluster that uses conventional hardware might be a good starting point towards high availability. We envisage that future generation of database management systems will be intrusion tolerant — that is, they are capable of continuous operation against not only hardware and process crash fault, but a variety of security threats as well.

REFERENCES

Agrawal, D., El Abbadi, A., & Steinke, R.C. (1997). Epidemic algorithms in replicated databases. *Proceedings of the ACM Symposium on Principles of Database Systems* (pp. 161-172), Tucson, AZ.

Ault, M., & Tumma, M. (2003). *Oracle9i RAC: Oracle real application clusters configuration and internals.* Kittrell, NC: Rampant TechPress.

Bernstein, P.A., Hadzilacos, V., & Goodman, N. (1987). *Concurrency control and recovery in database systems.* Reading, MA: Addison-Wesley.

Birman, K. (2005). *Reliable distributed systems: Technologies, Web services, and applications.* Berlin: Springer-Verlag.

Castro, M., & Liskov, B. (2002). Practical Byzantine fault tolerance and proactive recovery. *ACM Transactions on Computer Systems, 20*(4), 398-461.

Davies, A., & Fisk, H. (2006). *MySQL clustering.* MySQL Press.

Deswarte, Y., Blain, L., & Fabre, J.C. (1991). Intrusion tolerance in distributed computing systems. *Proceedings of the IEEE Symposium on Research in Security and Privacy* (pp. 110-121). Oakland, CA: IEEE Computer Society Press.

Gray, J., & Reuter, A. (1993). *Transaction processing: Concepts and techniques.* San Mateo, CA: Morgan Kaufmann.

Kemme, B., & Alonso, G. (2000). A new approach to developing and implementing eager database replication protocols. *ACM Transactions on Database Systems, 25*(3), 333-379.

Malkhi, D., & Reiter, M. (1997). Byzantine quorum systems. *Proceedings of the ACM Symposium on Theory of Computing* (pp. 569-578), El Paso, TX.

Mohan, C., Strong, R., & Finkelstein, S. (1983). Method for distributed transaction commit and recovery using Byzantine agreement within clusters of processors. *Proceedings of the ACM Symposium on Principles of Distributed Computing* (pp. 89-103), Montreal, Quebec.

Patino-Martinez, M., Jimenez-Peris, R., Kemme, B., & Alonso, G. (2005). Middle-R: Consistent database replication at the middleware level. *ACM Transactions on Computer Systems,* 375-423.

Skeen, D. (1981). Nonblocking commit protocols. *Proceedings of the ACM International Conference on Management of Data* (pp. 133-142), Ann Arbor, MI.

Stanoi, I., Agrawal, D., & El Abbadi, A. (1998). Using broadcast primitives in replicated databases. *Proceedings of the IEEE International Conference on Distributed Computing Systems* (pp. 148-155), Amsterdam, The Netherlands.

Strunk, D., Goodson, G., Scheinholtz, M., Soules, C., & Ganger, G. (2000). Self-securing storage: Protecting data in compromised systems. *Proceedings of the USENIX As-*

sociation Symposium on Operating Systems Design and Implementation (pp. 165-189), San Diego, CA.

KEY TERMS

Database Cluster (Shared-Everything, Shared-Nothing): A database management system runs on a group of computers interconnected by a high-speed network. In the cluster, multiple database server instances are deployed. If one instance fails, another instance takes over very quickly to ensure high availability. In the shared-everything design, all nodes can access a shared stable storage device. In the shared-nothing design, each node has its own cache buffer and stable storage.

Database Recovery (Roll-Backward, Roll-Forward): Recovery is needed when a database instance that has failed is restarted or a surviving database instance takes over a failed one. In roll-backward recovery, the active transactions at the time of failure are aborted and the resourced allocated for those transactions are released. In roll-forward recovery, the updates recorded in the redo log are transferred to the database so that they are not lost.

Database Replication (Eager, Lazy): Multiple instances of a database management system are deployed in different computers (often located in different sites). Their state is synchronized closely to ensure replica consistency. In eager replication, the updates are propagated and applied to all replicas within the transaction boundary. In lazy replication, the changes are propagated from one replica to others asynchronously.

High Availability (HA): The capability of a system to operate with long uptime and to recover quickly if a failure occurs. Typically, a highly available system implies that its measured uptime is five nines (99.999%) or better, which corresponds to 5.25 minutes of planned and unplanned downtime per year.

Split-Brain Syndrome: This problem may happen if the network partitions in a database cluster, and if each partition makes incompatible decisions on the outcome of transactions or their orders. To prevent this problem, typically only the main partition is allowed to survive.

Transaction: A group of read/write operations on the same data set that succeeds or fails atomically. More accurately, a transaction that has atomicity, consistency, isolation, and durability properties.

Two-Phase Commit Protocol (2PC): This protocol ensures atomic commitment of a transaction that spans multiple nodes in two phases. During the first phase, the coordinator (often the primary replica) queries the prepare status of a transaction. If all participants agree to commit, the coordinator decides to commit. Otherwise, the transaction is aborted. The second phase is needed to propagate the decision to all participants.

Histogram Generation from the HSV Color Space

Shamik Sural
Indian Institute of Technology, Kharagpur, India

A. Vadivel
Indian Institute of Technology, Kharagpur, India

A. K. Majumdar
Indian Institute of Technology, Kharagpur, India

INTRODUCTION

Digital image databases have seen an enormous growth over the last few years. However, since many image collections are poorly indexed or annotated, there is a great need for developing automated, content-based methods that would help users to retrieve images from these databases. In recent times, a lot of attention has been paid to the management of an overwhelming accumulation of rich digital images to support various search strategies. In order to improve the traditional text-based or SQL (Structured Query Language)-based database searches, research has been focused on efficient access to large image databases by the contents of images, such as color, shape, and texture. Content-based image retrieval (CBIR) has become an important research topic that covers a large number of domains like image processing, computer vision, very large databases, and human computer interaction (Smeulders, Worring, Santini, Gupta & Jain, 2000). Several content-based image retrieval systems and methods have recently been developed.

QBIC (Query By Image Content) is one of the first image retrieval systems developed at IBM (Niblack et al., 1993). Color, texture, and shape features are combined to represent each image in this system. The VisualSeek system, developed at the Columbia University, is an image retrieval system based on visual features (Chang, Smith, Mandis & Benitez, 1997). The NeTra system is a prototype image retrieval system, which uses color, texture, shape, and spatial location information as features to retrieve similar images (Ma & Manjunath, 1997). Some of the other popular CBIR systems are MARS (Ortega et al., 1998), Blobworld (Carson, Thomas, Belongie, Hellerstein & Malik, 1999), PicToSeek (Gevers & Smeulders, 2000), and SIMPLIcity (Wang, Li & Wiederhold, 2001).

An analysis of these systems reveals that all of them give a lot of importance on the image color for retrieval. In fact, color is always considered to be an important attribute, not only in content-based image retrieval systems, but also in a number of other applications like segmentation and video shot analysis. In color-based image retrieval, there are primarily two methods: one based on color layout (Smith & Chang, 1996) and the other based on color histogram (Swain & Ballard, 1991; Wang, 2001). In the color layout approach, two images are matched by their exact color distribution. This means that two images are considered close if they not only have similar color content, but also if they have similar color in approximately the same positions. In the second approach, each image is represented by its color histogram. A histogram is a vector whose components represent a count of the number of pixels having similar colors in the image. Thus, a color histogram may be considered to be a signature extracted from a complete image. Color histograms extracted from different images are indexed and stored in a database. During retrieval, the histogram of a query image is compared with the histogram of each database image using a standard distance metric like the Euclidean distance or the Manhattan distance. Since color histogram is a global feature of an image, the approaches based on color histogram are invariant to translation and rotation, and scale invariant with normalization.

Color histograms may be generated using properties of the different color spaces like RGB (Red, Green, and Blue), HSV (Hue, Saturation, and Intensity Value), and others. In this article, we give an overview of the different histogram generation methods using the HSV color space. We first present a brief background of the HSV color space and its characteristics, followed by the histogram generation techniques for various applications.

BACKGROUND

A color space or a color model is a specification of a coordinate system and a subspace within that system where a single point represents a distinct color value. There are several well-known color spaces that are used to represent the pixels of

an image. This representation is used for image analysis like extraction of color histograms. Each color space has its own merits and demerits depending on the application and hardware specification where it is going to be used. RGB, CMY, CMYK, and HSV are some of the popular color spaces. The RGB color space contains three color components, namely red, green and blue, each of which appears in its primary spectral components. Devices that deposit colored pigments on paper use CMY color space, and the representation of this color space is with the secondary colors of light, which are Cyan, Magenta, and Yellow. CMYK (Cyan, Magenta, Yellow, Black) color space is similar to CMY but is used to produce true black color, which is muddy-black in the CMY color space.

The HSV (Hue, Saturation, Value) color space, on the other hand, closely corresponds to the human visual perception of color. The HSV color space can be represented as a three-dimensional hexacone, where the central vertical axis represents intensity which takes a value between 0 and 255 (Shapiro & Stockman, 2001). Hue is defined as an angle in the range $\pi[0,2]$ relative to the red axis with red at angle 0, green π at 2/3, blue π at 4/3, and red again π at 2. Saturation is the depth or purity of color and is measured as a radial distance from the central axis to the outer surface. For zero saturation, as we move higher along the intensity axis, we go from black to white through various shades of gray. On the other hand, for a given intensity and hue, if the saturation is changed from zero to one, the perceived color changes from a shade of gray to the most pure form of the color represented by its hue. When saturation is near zero, all pixels, even with different hues, look alike and as we increase the saturation towards one, they tend to get separated out and are visually perceived as the true colors represented by their hues. Thus, the effect of saturation may be considered as that of introducing visual shadows on the image for any given value of hue and intensity.

The HSV model is an ideal tool for developing image and video processing algorithms based on color descriptions. A number of histogram generation methods from the HSV color space have recently been proposed for different applications. We next describe some of these approaches.

HISTOGRAM GENERATION FROM THE HSV COLOR SPACE

The HSV color space in general, and the HSV color histogram in particular, plays an important role in image analysis. A color histogram can be used in image retrieval, segmentation, video shot detection, color and intensity-based clustering, place recognition for topological localization, person identification and authentication using biometric techniques, as well as in many other applications.

For image retrieval applications, an HSV color histogram can be generated using an approach similar to the RGB color space. The hue scale is divided into eight groups, saturation scale is divided into two groups, and the intensity scale is divided into four groups. By combining each of these groups, we get a total of 64 cells to represent a 64-component HSV color histogram. The reason for having a different number of groups for the three scales is that, of the three axes, hue is considered to be the most important, followed by intensity, and finally, saturation. For the H, S, and V combination of values, the corresponding histogram component is determined. The respective histogram component is updated by one for each pixel having the corresponding color combination. An efficient indexing of the histograms can enhance the performance of a CBIR application to a great extent. Smith and Chang (1996) exploit this idea in their color set approach. This method extracts spatially localized color information and provides efficient indexing of the color regions. The large single color regions are extracted first, followed by multiple color regions. They utilize binary color sets to represent the color content as a color histogram. The H and S dimensions are divided into N and M bins, respectively, for a total of N×M bins (Ortega et al., 1998). Each bin contains the percentage of pixels in the image that have corresponding H and S colors for that bin. Intersection similarity is used as a measure to capture the amount of overlap between two histograms.

From the properties of the HSV color space, it is observed that for low values of saturation, a color is approximated by a gray value specified by the intensity level while for higher saturation, the color is approximated by its hue. This captures the human visual properties effectively and can be used to generate a histogram for image retrieval applications (Sural, 2003; Sural, Qian & Pramanik, 2002). The saturation threshold that determines this transition is once again dependent on the intensity. Thus, the value of saturation projected onto the hue and intensity plane is useful for the extraction of color information. A threshold function can be used to determine if a pixel should be represented by its hue or by its intensity in the color histogram. For an intensity value of zero, all the colors are considered as black, whatever their hue or saturation may be. On the other hand, with increasing values of intensity, the saturation threshold that separates hue dominance from intensity dominance goes down. This approach treats the pixels as a distribution of "colors" in an image where a pixel may be of a "gray color" or of a "true color." The histogram is a logical combination of two independent histograms—one for the true colors and one for the gray colors. One drawback of this approach is that for saturation values near the threshold, a pixel is neither a true color pixel nor a gray color pixel. In order to capture the fuzzy nature of human visual perception of color, there is a need for using a soft threshold to determine the dominant property of a pixel. In the soft threshold approach, two

components of the histogram are updated for each pixel in an image, namely, a gray color component and a true color component. The quantum of update is determined both by the saturation and the intensity of the pixel, and the sum of the weights of the two contributions equals unity. Also, for the same saturation, the weight varies with intensity. For a lower intensity value, the same saturation gives a lower weight on the true color component and vice versa. This histogram has a high recall and precision of retrieval, and is effectively used in content-based image retrieval systems (Vadivel, Majumdar & Sural, 2003). The soft threshold approach can also be used for video shot detection (Sural, Mohan & Majumdar, 2004).

An input image in a content-based retrieval system may be textured or non-textured. Similarly, the images stored in a database can also be classified as textured or non-textured. When the query image is compared with the database images, the search should be restricted to the relevant portion of the database. An HSV histogram can be effectively used to classify an image into a textured or a non-textured class (Li, Wang & Wiederhold, 2000). In this approach, an image is first segmented into 4X4 pixel regions. HSV color histogram is extracted for each such region. When an image is a color-rich image, hue plays an important role in its representation. Although, hue can represent millions of colors, the human visual system cannot distinctly recognize all of them. The visually similar colors can be combined together in the same color band, as suggested by Gong, Proietti, and Faloutsos (1998). After combining the similar colors, the image is segmented and indexed for content-based retrieval. It is observed that human visual system can perceive distinct colors depending on the NBS (National Bureau of Standards) color distance. Colors with NBS distance below 3.0 are indistinguishable to the human eye. Besides hue, if the intensity axis is also divided into a number of bands, then segmentation, clustering, indexing, and retrieval of images can be done even more effectively (Zhang & Wang, 2000). In this approach, the histograms are generated separately—one from the hue component and the other from the intensity component. K-means clustering algorithm is then applied on these two histograms to obtain the center for each class for indexing.

It should be noted that the histogram of an image does not keep track of the spatial information of the pixels in an image. Two images with the same number of color pixels but at different locations would have the same histogram. This results in higher false-positives. If the spatial relationship is also captured during histogram generation, then the retrieval performance can be enhanced. Color correlogram is a type of histogram generated from the HSV color space that retains spatial information (Ojala, Rautiainen, Matinmikko & Aittola, 2001).

In addition to content-based image retrieval, the HSV histogram is used for a large number of other applications. One important class of applications is the domain of topological localization in robotics. Ulrich and Nourbakhsh (2000) use the HSV histogram for appearance-based place recognition. They first determine the candidate locations based on the current belief of a robot's location and transform the input image into six one-dimensional histograms. Then for each candidate location and for each of the two color bands, RGB and HSV, the reference histogram is determined that matches the input histogram most closely. In order to reduce the resource requirements for storage and transmission of image and video, various data compression techniques are used in practice. A fundamental goal of data compression is to obtain the best possible fidelity for a given data rate or, equivalently, to minimize the rate required for a given fidelity. Matching Pursuit (MP) image representation has proven to give good compression results. It is found that the MP coefficients are interestingly distributed along the diagonal of the color cube. Coding of the MP coefficients is done in the HSV color space, where V becomes the projection of RGB coefficients on the diagonal of the cube, S is the distance of the coefficient to the diagonal, and H is the direction perpendicular to the diagonal where the RGB coefficient is placed. MP coefficients are quantized and a histogram of the quantized coefficients is generated in the HSV space for efficient data compression (Rosa, Ventura & Vandergheynst, 2003).

FUTURE TRENDS

The HSV color space provides a close representation of human visual perception of color. New and effective algorithms are being developed to extract more useful information from the HSV color space. Recent research shows that the HSV color space can even be used to represent both color and texture information in a single histogram. An extension of this work could be in the domain of fuzzy feature extraction for segmentation and retrieval. Further theoretical and experimental comparisons should be made between the HSV and other color spaces.

CONCLUSION

We have discussed a number of histogram generation techniques using the HSV color space and made a critical assessment of their merits and demerits. Color histograms generated from the HSV color space have been used in a variety of image and video processing applications. Content-based image retrieval is one of the most popular domains in which the HSV histogram has been used effectively for high recall and precision of retrieval. The histogram may also be combined with other features for representing semantic contents in image and video.

ACKNOWLEDGMENT

The work done by Shamik Sural is supported by research grants from the Department of Science and Technology, India, under Grant No. SR/FTP/ETA-20/2003 and by a grant from IIT Kharagpur under ISIRD scheme No. IIT/SRIC/ISIRD/2002-2003.

REFERENCES

Carson, C., Thomas, M., Belongie, S., Hellerstein, J.M., & Malik, J. (1999). Blobworld: A system for region-based image indexing and retrieval. *Proceedings of the Third International Conference on Visual Information Systems* (pp. 509-516).

Chang, S-F., Smith, J.R., Mandis, B., & Benitez, A. (1997). Visual information retrieval from large distributed online repositories. *Communications of the ACM, 40*, 63-71.

Gevers, T., & Smeulders, A.W.M. (2000). PicToSeek: Combining color and shape invariant features for image retrieval. *IEEE Transactions on Image Processing, 9,* 102-119.

Gong, Y., Proietti ,G., & Faloutsos, C. (1998). Image indexing and retrieval based on human perceptual color clustering. *Computer Vision and Patten Recognition,* 578-583.

Li, J., Wang, J.Z., & Wiederhold, G. (2000). Classification of textured and non-textured images using region segmentation. *Proceedings of the Seventh International Conference in Image Processing* (pp. 754-757).

Ma, W.Y., & Manjunath, B.S. (1997). NeTra: A toolbox for navigating large image databases. *Proceedings of the IEEE International Conference on Image Processing* (pp. 568-571).

Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E., Pektovic, D., Yanker, P., Faloutsos, C., & Taubin, G. (1993). The QBIC project: Querying images by content using color texture and shape. *Storage and Retrieval for Image and Video Databases, 1908,* 173-187.

Ojala, T., Rautiainen, M., Matinmikko, E., & Aittola, M. (2001). Semantic image retrieval with HSV correlograms. *Proceedings of the Scandinavian Conference on Image Analysis* (pp. 621-627).

Ortega, M., Rui, Y., Chakrabarti, K., Porkaew, K., Meharotra, S., & Huang, T.S. (1998). Supporting ranked Boolean similarity queries in MARS. *IEEE Transactions on Knowledge and Data Engineering, 10*(6), 905-925.

Rosa, M., Ventura, F., & Vandergheynst, P. (2003). *Scalable color image coding with Matching Pursuit.* Technical Report, ITS-TR-05.03. Signal Processing Institute.

Shapiro, L., & Stockman, G. (2001). *Computer vision.* Englewood Cliffs, NJ: Prentice-Hall.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349-1380.

Smith, J.R., & Chang, S-F. (1996). Tools and techniques for color image retrieval. *SPIE Storage and Retrieval for Image and Video Databases,* 426-437.

Sural, S. (2003). Histogram generation from the HSV color space using saturation projection. In S. Deb (Ed.), *Multimedia systems and content-based image retrieval.* Hershey, PA: Idea Group Publishing.

Sural, S., Mohan, M., & Majumdar, A.K. (2004). A soft-decision histogram from the HSV color space for video shot detection. In S. Deb (Ed.), *Video data management and information retrieval.* Hershey, PA: Idea Group Publishing.

Sural, S., Qian, G., & Pramanik, S. (2002). Segmentation and histogram generation using the HSV color space for content-based image retrieval. *Proceedings of the IEEE International Conference on Image Processing* (pp. 589-592).

Swain, M.J., & Ballard, D.H. (1991). Color indexing. *International Journal of Computer Vision, 7*(1), 11-32.

Ulrich, I., & Nourbakhsh, I. (2000). Appearance-based place recognition for topological localization. *IEEE International Conference on Robotics and Automation, 2,* 1023-1029.

Vadivel, A., Majumdar, A.K., & Sural, S. (2003). Perceptually smooth histogram generation from the HSV color space for content-based image retrieval. *Proceedings of the International Conference on Advances in Pattern Recognition* (pp. 248-251).

Wang, J.Z., Li, J., & Wiederhold, G. (2001). SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23,* 947-963.

Wang, S. (2001). *A robust CBIR approach using local color histograms.* Masters thesis. Department of Computing Science, University of Alberta, Canada.

Zhang, C., & Wang, P. (2000). A new method of color image segmentation based on intensity and hue clustering. *Proceedings of the International Conference on Pattern Recognition, 3,* 3-8.

KEY TERMS

Content-Based Image Retrieval: Retrieval of images similar to a given image based only on features present in the image and not any external information.

Histogram: A vector whose components represent similar colors in an image. The value of a component is the number of image pixels having that color.

HSV Color Space: A color space consisting of hue, saturation, and intensity value. It is a popular way of representing color content of an image.

Precision: The number of relevant images retrieved as a percentage of the total number of images retrieved.

Recall: The number of relevant images retrieved as a percentage of the total number of relevant images in the database.

Soft Threshold: A fuzzy approach to decide the importance of a feature. This is in contrast to a hard threshold where a yes/no decision is made.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1333-1337,

Histogram-Based Compression of Databases and Data Cubes

Alfredo Cuzzocrea
University of Calabria, Italy

INTRODUCTION

Histograms have been extensively studied and applied in the context of *Selectivity Estimation* (Kooi, 1980; Muralikrishna & DeWitt, 1998; Piatetsky-Shapiro et al., 1984; Poosala et al., 1996; Poosala, 1997), and are effectively implemented in commercial systems (e.g., Oracle Database, IBM DB2 Universal Database, Microsoft SQL Server) to **query optimization** purposes. In statistical databases (Malvestuto, 1993; Shoshani, 1997), histograms represent a method for approximating **probability distributions**. They have also been used in data mining activities, intrusion detection systems, scientific databases, that is, in all those applications which (*i*) operate on huge numbers of detailed records, (*ii*) extract useful knowledge only from condensed information consisting of summary data, (*iii*) but are not usually concerned with detailed information. Indeed, histograms can reach a surprising efficiency and effectiveness in approximating the actual distributions of data starting from **summarized information**. This has led the research community to investigate the use of histograms in the fields of database management systems (Acharya et al., 1999; Bruno et al., 2001; Gunopulos et al., 2000; Ioannidis & Poosala, 1999; Kooi, 1980; Muralikrishna & DeWitt, 1998; Piatetsky-Shapiro et al., 1984; Poosala, 1997; Poosala & Ioannidis, 1997), *online analytical processing* (OLAP) systems (Buccafurri et al., 2003; Cuzzocrea, 2005a; Cuzzocrea & Wang, 2007; Furfaro et al., 2005; Poosala & Ganti, 1999), and data stream management systems (Guha et al., 2001; Guha et al., 2002; Thaper et al., 2002), where, specifically, compressing data is mandatory in order to obtain fast answers and manage the endless arrival of new information, as no bound can be given to the amount of information which can be received.

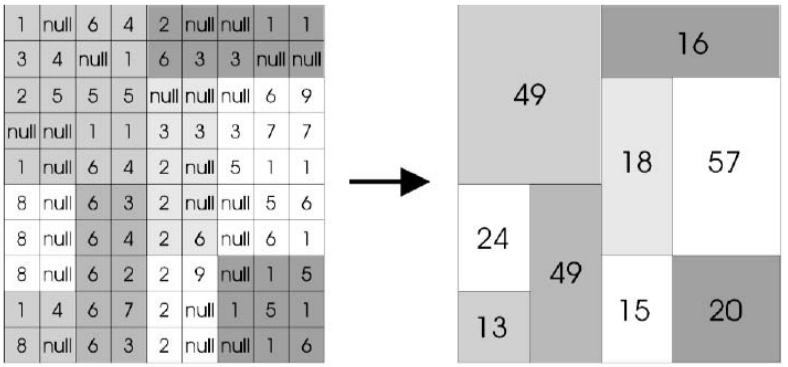
Histograms are data structures obtained by partitioning a **data distribution** (or, equally, a data domain) into a number of mutually disjoint blocks, called ***buckets***, and then storing, for each bucket, some **aggregate information** of the corresponding range of values, like the sum of values in that range (i.e., applying the SQL aggregate operator SUM), or the number of occurrences (i.e., applying the SQL aggregate operator COUNT), such that this information retains a certain "summarizing content."

Figure 1 shows an instance of a histogram built on a two-dimensional **data cube** (left-side of the figure), represented as a matrix. The corresponding (two-dimensional) histogram (right-side of the figure) is obtained by (*i*) partitioning the matrix into some rectangular buckets which do not overlap, and (*ii*) storing for each so-obtained bucket the sum of the measure attributes it contains.

Histograms are widely used to support two kinds of applications: (*i*) selectivity estimation inside *Query Optimizers* of DBMS, as highlighted before, and (*ii*) *approximate query answering* against databases and data cubes. In the former case, the data distribution to be compressed consists of the frequencies of values of attributes in a relation (it should be noted that, in this vest, histograms are mainly used within the core layer of DBMS, thus dealing with databases properly). In the latter case, the data distribution to be compressed consists of the data items of the target domain (i.e., a database or a data cube) directly, and the goal is to provide fast and approximate answers to resource-intensive queries instead of waiting-for time-consuming exact evaluations of queries. To this end, a widely-accepted idea is that of evaluating (with some approximation) queries against ***synopsis data structures*** (i.e., succinct, compressed representations of original data) computed over input data structures (i.e., a database or a data cube) instead of the same input data structures. Histograms are a very-popular class of synopsis data structures, so that they have been extensively used in the context of approximate query answering techniques. Some relevant experiences concerning this utilization of histograms are represented by the work of Ioannidis and Poosala (Ioannidis & Poosala, 1999), that propose using histograms to provide approximate answers to set-valued queries, and the work of Poosala and Ganti (Poosala & Ganti, 1999), that propose using histograms to provide approximate answers to *range-queries* (Ho et al., 1997) in OLAP.

In both utilizations, a relevant problem is how to reconstruct the original data distribution form the compressed one. In turn, this derives from the fact that the original data distribution summarized within a bucket cannot be reconstructed exactly, but can be approximated using some estimation strategies, like *continuous value assumption* (CVA) (Colliat, 1996) or *uniform spread assumption* (USA) (Poosala et al., 1996). For a given storage space reduction, the problem of determining the "best" histogram (i.e., the histogram which minimizes the approximation of reconstructing the original content of ranges corresponding to buckets) is crucial. Indeed,

Figure 1. A two-dimensional data cube and its corresponding two-dimensional histogram



different partitions lead to dramatically different errors in reconstructing the original data distribution, especially for *skewed* (i.e., asymmetric) data. This issue has been investigated for some decades, and a large number of techniques for arranging histograms have been proposed (Buccafurri et al., 2003; Christodoulakis, 1984; Donjerkovic et al., 1999; Ioannidis & Poosala, 1995; Jagadish et al., 2001). The aim of every partition technique is to build a histogram whose buckets contain values with "small" differences, so that one can estimate a range query inside a bucket assuming that data distribution is uniform, thus successfully exploiting linear interpolation. Indeed, finding the optimal solution to this problem in multiple dimensions is NP-hard (Muthukrishnan et al., 1999). Several techniques and heuristics have been proposed to find sub-optimal solutions with provable quality guarantees (Jagadish et al., 1998; Gilbert et al., 2001). These guarantees regard the "distance" of the provided solution from the optimal one, but do not provide any measure of the approximation of each estimated answer to a range-query.

Another important, more recent utilization of histograms concerns with the *data visualization problem*, where the data compression paradigm is intended as a solution to aid the visualization of complex and multidimensional domains. This is a quite-unexplored line of research: pioneeristic works can be found in the DIVE-ON (Ammoura et al., 2001) and *Polaris* (Stolte et al., 2002) projects, whereas recent works can be found in (Cuzzocrea et al., 2007).

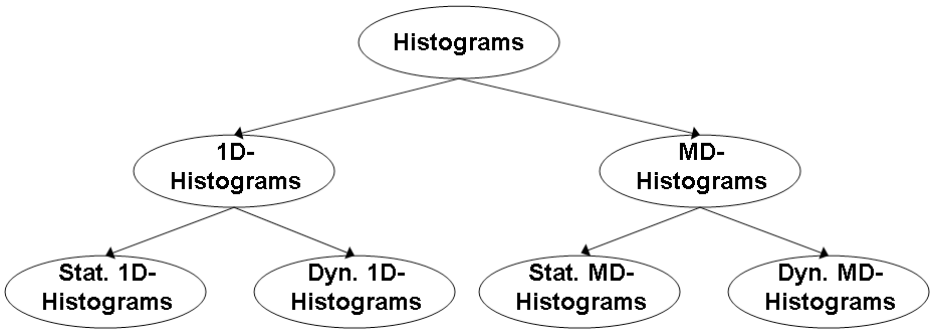
In this chapter, we survey several state-of-the-art histogram-based techniques for compressing databases and data cubes, ranging from one-dimensional to multidimensional data domains. Specifically, we highlight similarity and differences existing among the investigated techniques, and put-in-evidence how proposals have evolved over time towards more and more sophisticated and very-efficient solutions, beyond early experiences focused on selectivity estimation issues within the core layer of DBMS.

BACKGROUND

A *database* D is a tuple $D = \langle W,I,F \rangle$ such that (*i*) W is the *schema* of D; (*ii*) I is the *instance* of D, that is, its realization in terms of collections of tuples adhering to W; (*iii*) F is the collection of *functional dependencies* defined over W. In turn, W is a collection of *relation schemas* $W = \{T_0, T_1,\ldots, T_P\}$, with $P = |W| - 1$, such that T_i, with $0 \leq i \leq P$, is defined as a tuple $T_i = \langle K, A_{i,0}, A_{i,1},\ldots, A_{i,G} \rangle$, with $G = |T_i| - 1$, such that K is the *key* of T_i, and $A_{i,j}$ is the j^{th} *attribute* of T_i. A functional dependence is expressed as a *logical rule* over attributes of T_i. The instance of a relation scheme T_i is named as *relation*, and denoted by R_i.

A *data cube* L is a tuple $L = \langle C,J,H,M \rangle$, such that: (*i*) C is the data domain of L containing (OLAP) *data cells*, which are the basic SQL aggregations of L computed against the relational data source S alimenting L; (*ii*) J is the set of *dimensions* of L, that is, the *functional attributes* (of S) with respect to which the underlying OLAP analysis is defined (in other words, J is the set of attributes with respect to which relational tuples in S are aggregated); (*iii*) H is the set of *hierarchies* related to the dimensions of L, that is, hierarchical representations of the functional attributes shaped-in-the-form-of generic trees; (*iv*) M is the set of *measures* of L, that is, the *attributes of interest* (of S) for the underlying OLAP analysis (in other words, M is the set of attributes with respect to which SQL aggregations stored in data cells of L are computed).

Figure 2. A taxonomy of histograms



HISTOGRAM-BASED COMPRESSION TECHNIQUES FOR DATABASES AND DATA CUBES

Classes of Histograms

We distinguish between *one-dimensional histograms*, those devoted to compress one-dimensional data domains, and *multidimensional histograms*, those working on multidimensional data domains, which are more interesting than the former, and can be found in a wide range of modern, large-scale, *data-intensive* applications. Moreover, we can also further distinguish between *static histograms* and *dynamic histograms*. The first ones are statically computed against the target domain, and are not particularly suitable to efficiently accomplish data updates occurring on original data sources. The second ones are dynamically computed by taking into consideration, beyond the target domain, or, in some cases, a synopsis of it, other entities related to the dynamics of the target DBMS/OLAP server such as *query-workloads*, *query feedbacks*, *load balancing issues,* and so forth. Contrary to the previous histogram class, dynamic histograms efficiently support update management, being their partition dependent on a "parametric" configuration that can be easily (re-)computed at will. From this classification, a simple-yet-effective *taxonomy* of histograms can be derived (see Figure 2).

Static One-Dimensional Histograms

One-dimensional histograms deal with the problem of approximating a one-dimensional data domain or, equally, a one-dimensional data distribution. As an example, one-dimensional histograms can be used to approximate the domain/distribution of an attribute R.A of a given relation R of a RDBMS server to selectivity estimation purposes. Such histograms represent first research experiences in this field, and have been proposed in the context of query optimizers mainly (Kooi, 1980).

Let B be the whole number of buckets of the final histogram, obtaining an equal number of rows per bucket is the goal of (one-dimensional) *Equi-depth histogram*, introduced by Piatetsky-Shapiro and Connell (Piatetsky-Shapiro et al., 1984), that propose a simple construction procedure based on first sorting, and then taking B—1 equally-spaced splits of the target domain. This approach can be improved by first *sampling*, and then taking equally-spaced splits of the sampled domain instead of the original domain (Piatetsky-Shapiro et al., 1984), or applying *one-pass quantile algorithms* (Greenwald & Khanna, 2001). Maintaining such histogram is also very efficient as it can be maintained using one-pass algorithms, or a *backing sample* (Gibbons et al., 1997), which, essentially, consists in keeping bucket counts up-to-date.

Poosala et al. (1996) propose *compressing Equi-Depth* histograms by creating singleton buckets for largest values, and maintaining the equi-depth strategy over the rest of data. This allows the "original" version of the *Equi-Depth* histogram to be improved significantly, as exact information is kept for largest values, whereas less-detailed information is maintained for the rest of data. The construction of such compressed histogram can be implemented by (*i*) sorting buckets, (*ii*) scanning the buckets looking for the largest values (this is done in time $O(B \cdot logB)$), and, finally, (*iii*) performing a one-pass scan in order to process the remaining data. Alternatively, sampling can be used to improve performance, similarly to the "original" version of the histogram. Gibbons et al. (1997) propose *maintaining* compressed *Equi-Depth* histograms by using the *split-&-merge* approach as in the

uncompressed ones, but also taking in account the issue of deciding when to create and remove singleton buckets.

In a *V-Optimal histogram*, introduced by Ioannidis and Poosala (1995), buckets are selected in such a way as to minimize frequency variance within them. In Ioannidis and Poosala (1995), authors show that *V-Optimal* histogram minimizes the average selectivity estimation error for equality-joins and selections. Following the original idea, *V-Optimal* construction method has been further improved by Jagadish et al. (1998), that give-to-this-end an $O(B \cdot n^2)$ time dynamic programming algorithm, such that (*i*) B is (still) the overall number of buckets of the histogram, and (*ii*) n is the number of items of the input data domain D (i.e., the size of D).

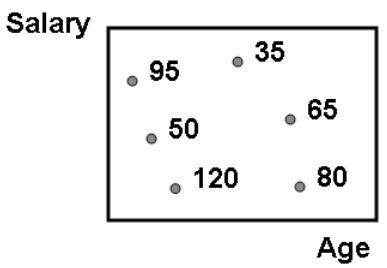
Dynamic One-Dimensional Histograms

In the case of dynamic one-dimensional histograms, the final histogram is computed by considering additional information apart from the target data domain (or a synopsis of it). A relevant instance of such class of histograms is represented by the so-called *Self-Tuning histogram* (*ST-histogram*), introduced by Aboulanaga and Chaudhuri (1999), that propose improving the original version of *Equi-Depth* histogram via tuning bucket frequencies by (*i*) comparing actual selectivity to histogram estimate, (*ii*) using this information to adjust bucket frequencies by dividing the values of each bucket by a quantity equal to $D_F \cdot S_E$, where D_F is the *dampening factor*, and S_E is the *actual* error, and, finally, (*iii*) restructuring the histogram by merging buckets of near-equal-frequencies and splitting large-frequency buckets.

Static Multidimensional Histograms

The basic problem addressed by techniques for compressing multidimensional domains is to approximate the *joint data distribution* of multiple attributes, obtained by jointly combining the distributions of singleton attributes. For instance, in Figure 3 a two-dimensional data domain on salary and age of employers is depicted; in each dimension, a distinctive data distribution related to the singleton attribute can be identified.

Figure 3. A two-dimensional joint data distribution



The goal of these techniques is to provide selectivity estimation for queries with multiple predicates, like in Muralikrishna and DeWitt (1998), and to approximate general relations, like in Poosala and Ioannidis (1997), or OLAP data cubes, like in Vitter et al. (1998). A popular and conventional approach is based on the well-known *attribute-value independence* (AVI) assumption (Selinger et al., 1979), according to which any query involving a set of attributes can be answered by applying it on each attribute singularly. This approach is theoretically reasonable, but it has been recognized as source of gross errors in practice (e.g., (Christodoulakis, 1984; Faloutsos & Kamel, 1997; Poosala, 1997; Poosala & Ioannidis, 1997). To cope with this problem, multidimensional histograms use a small number of multidimensional buckets to *directly* approximate the joint data distribution. The approximate d-dimensional distribution $\tilde{f}$ is obtained from the actual d-dimensional distribution f via (*i*) setting the total bucket frequency F, and (*ii*) approximating data points of f on a $w(0) \times w(1) \times \ldots \times w(|d|-2) \times w(|d|-1)$ uniform grid G, such that $w(k)$ is the number of distinctive values along the dimension d_k, and each cell of G has a bucket frequency equal to $F / (w(0) \cdot w(1) \cdot \ldots \cdot w(|d|-2) \cdot w(|d|-1))$.

However, even if reasonable, this approach introduces serious limitations when applied to real-life databases and data cubes. Indeed, constructing histograms is much harder for two dimensions than for the one-dimensional case, as recognized by Muthukrishnan et al. (1999). This problem gets worse in the case that the dimension number increases, and becomes a problematic bottleneck in real-life data-intensive systems where corporate databases and data cubes with more than 100 dimensions can be found. From this evidence, various and heterogeneous alternatives to the problem of computing multidimensional histograms have been proposed in literature, each of them based on one or more properties of data distributions characterizing the input data domain. Conventional approaches take into consideration statistical and error-metrics-based properties of data distributions. Other approaches expose *greedy solutions*, as traditional approaches introduce excessive computational overheads on highly-dimensional data domains.

Among all the alternatives, we focus our attention on the following (static) multidimensional histograms, mainly because they can be considered as representative and significant experiences in this context: *Equi-Depth* (Muralikrishna & DeWitt, 1998), *MHist* (Poosala & Ioannidis, 1997), *Min-Skew* (Acharya et al., 1999), *GenHist* (Gunopulos et al., 2000), and *GHBH* (Furfaro et al., 2005) histograms.

Given a d-dimensional data domain D (e.g., a data cube), (multidimensional) *Equi-Depth histogram* $H_{E\text{-}D}(D)$, proposed by Muralikrishna and DeWitt (1998), is built as follows: (*i*) fix an ordering of the d dimensions; (*ii*) set α

$\approx |d|^{th}$ root of desired number of buckets; (*iii*) initialize $H_{E\text{-}D}(D)$ to the input data distribution of D; (*iv*) for each k in $\{0, 1, \ldots, |d|\text{-}1\}$ split each bucket in $H_{E\text{-}D}(D)$ in α equi-depth partitions along d_k; finally, (*v*) return resulting buckets to $H_{E\text{-}D}(D)$. This technique presents some limitations: fixing α and the dimension ordering can result in poor partitions, and, consequently, there could be a limited level of *bucketization* that, in turn, involves low quality of $H_{E\text{-}D}(D)$ in its general goal of approximating D.

MHist histogram, proposed by Poosala and Ioannidis (1997), overcomes *Equi-Depth* performance. *MHist* build procedure depends on the parameter p (specifically, such histogram is denoted by *MHist-p*): contrarily to the previous technique, at each step, the bucket b_i in a *MHist* histogram $H_{MH}(D)$ containing the dimension d_k whose *marginal* is the most in need of partitioning is chosen, and it is split along d_k into p (e.g., $p = 2$) buckets. Experimental results shown in Poosala and Ioannidis (1997) state that *MHist* overcomes AVI and *Equi-Depth*.

Min-Skew histogram was originally designed by Acharya et al. (1999) to tackle the problem of selectivity estimation of *spatial data* in *geographical information systems* (GIS). Spatial data are referred to *spatial* (or *geographical*) *entities* such as points, lines, poly-lines, polygons and surfaces, and are very often treated by means of minimal rectangles containing them, namely *Minimum bounding rectangles* (MBR). *Min-Skew* is more sophisticated than *MHist*. The main idea behind a *Min-Skew* histogram $H_{M\text{-}S}(D)$ is to follow the criterion of minimizing the *spatial skew* of the histogram by performing a *binary space partitioning* (BSP) via recursively dividing the space along one of the dimensions each time. More formally, each point in the space of a given GIS instance is associated to a *spatial density*, defined as the number of MBR that contain such a point. When performing the partition, each bucket b_i is assigned the spatial skew s_i, defined as follows: $s_i = \sum_{j=0}^{n_i-1} (f_j - \bar{f})^2 / n_i$, where (*i*) n_i is the number of points contained within b_i, (*ii*) f_j is the *spatial frequency* of the j^{th} point within b_i, and (*iii*) $\bar{f}$ represents the *average frequency* of all the points within b_i. The total skew S of $H_{M\text{-}S}(D)$ is defined as follows: $S = \sum_{i=0}^{B-1} n_i \cdot s_i$, where (*i*) B is the total number of buckets, (*ii*) s_i is the spatial skew associated with bucket b_i, and (*iii*) n_i is the number of points of b_i. The construction technique of $H_{M\text{-}S}(D)$ tries, at each step, to minimize the overall spatial skew of the histogram by selecting (*i*) a bucket to be split, (*ii*) a dimension of the multidimensional space along which split, and (*iii*) a splitting point along that dimension such that the overall spatial skew computed after the split is smaller than the one computed at the previous step (i.e., before the current split). Finally, noticing that the spatial skew captures the variance of the spatial density of MBR within each bucket, we can say that *Min-Skew* follows, in some sense, the spirit of *V-Optimal*.

Gunopulos et al. propose *GenHist histogram* (2000), a new kind of multidimensional histogram that is different from the previous ones with respect to the build procedure. The key idea is the following: given an histogram H whit h_b buckets on an input data domain D, a *GenHist* histogram $H_{GH}(D)$ is built by finding n_b overlapping buckets on H, such that n_b is an input parameter. To this end, the technique individuates the number of distinct regions that is much larger than the original number of buckets h_b, thanks to a greedy algorithm that considers *increasingly-coarser grids*. At each step, such algorithm selects the set of cells J of highest density, and moves enough randomly-selected points from J into a bucket to make J and its neighbors "close-to-uniform." Therefore, the novelty of the proposal consists in defining a truly multidimensional splitting policy, based on the concept of *tuple density*. A drawback of the *GenHist* proposal is the difficulty of choosing the right values for setting the input parameters, which are: (*i*) the degree of the grid ξ used to obtain regular partitions of the data domain at each iteration; (*ii*) the number of buckets b created at each iteration; (*iii*) the value α, which controls the rate by which ξ decreases. Indeed, authors state that: (*i*) the optimal setting for the initial value of ξ, denoted by ξ_0, must be such that, at the first iteration, the percentage of points that are removed from the input data cube to provide b buckets is at least

$$\frac{1}{\log_{1/\alpha} \xi};$$

(*ii*) the optimal setting for α is $\alpha = (1/2)^{1/d}$, such that d is the number of dimensions of the input data cube.

More recently, Furfaro et al. (2005) propose a new kind of multidimensional histogram which can be considered as very innovative if compared with previous ones. They first propose the definition of a new *class* of histograms, called *flat binary histogram* (FBH), which are characterized by the property that buckets are represented independently from one another, without exploiting the hierarchical structure of the underlying partition. Then, they classify *MHist* and *Min-Skew* inside the class FBH histograms, and give theoretical motivations to this claim. Finally, they introduce the *grid hierarchical binary histogram* (GHBH), a new histogram belonging to the FBH class. The partition scheme used by a *GHBH* histogram $H_{GHBH}(D)$ is the same of that of *MHist* and *Min-Skew* histograms, thus based on (*i*) statistical properties of data, like the standard deviation, and (*ii*) greedy (generating) algorithms. The novelty of $H_{GHBH}(D)$ is that the hierarchy adopted to determine the structure of the histogram is also used to represent it, thus introducing surprising efficiency in terms of both space consumption and accuracy of query estimations. Furthermore, $H_{GHBH}(D)$ is based on a constrained partition scheme, where buckets of data cannot be split anywhere along one of their dimensions, but the

split must be laid onto a grid partitioning the bucket into a number of equally-sized sub-buckets. The adoption of this constrained partitioning enables a more efficient physical representation of the histogram with respect to other histograms using more traditional partition schemes. Thus, the saved space can be invested to obtain finer grain buckets, which approximate data in more detail. Indeed, the ability of creating a larger amount of buckets (in a given storage space) does not guarantee a better accuracy in estimating range-queries, as it could be the case that buckets created by adopting a constrained scheme contain very skewed (i.e., non-uniform) distributions. In Furfaro et al. (2005), authors show that given a *d*-dimensional data domain *D* and the space bound *M* (i.e., the storage space available for housing the output compressed data structure), the maximum number of buckets $\beta_{FBH}^{\max}$ within *M* of a FBH over *D*, denoted by $H_{FBH}(D)$, is

$$\beta_{FBH}^{\max} = \left\lfloor \frac{M}{32 \cdot (2 \cdot |d| + 1)} \right\rfloor$$

assuming that an integer value is represented using 32 bits. On the contrary, using $H_{GHBH}(D)$, the maximum number of buckets $\beta_{GHBH}^{\max}$ within *M* is

$$\beta_{FBH}^{\max} = \left\lfloor \frac{M + \log \delta + \lceil \log |d| \rceil - 30}{3 + \log \delta + \lceil \log |d| \rceil} \right\rfloor,$$

such that δ is the "degree" of the grid-based constrained partition scheme.

Dynamic Multidimensional Histograms

Dynamic multidimensional histograms extend capabilities of static multidimensional histograms by incorporating inside their generating algorithms the amenity of building/refining the underlying partition in dependence on non-conventional entities related to the dynamic behavior of the target DBMS/OLAP server, such as query-workloads. Among this class of histograms, relevant proposal are: *wavelet-based* (Matias et al., 1998) and *STHoles* (Bruno et al., 2001) histograms, and our innovative data structure **TP**-*Tree* (Cuzzocrea & Wang, 2007).

Wavelet-based histograms, proposed by Matias et al. (1998), aim at combining the benefits coming from the usage of *wavelet transformations* in the context of data-compression/approximate-query-answering (Vitter et al., 1998) (basically, these benefits can be synthesized in a greater flexibility in supporting different classes of queries rather than histograms) with histograms. The key idea of this approach is to use a compact subset of *Haar* (or linear) *wavelet coefficients* to approximate the data distribution. This approach has provided good results in range-query selectivity estimation issues, and has outperformed performance of original histograms. Subsequently, authors have then addressed the problem of the *dynamic maintenance of wavelet-based histograms* (Matias et al., 2000). Here, they observe that updates in singleton distribution value can affect the values of many coefficients through propagations like paths-to-the-root of the *decomposition tree* given by the wavelet transformation; therefore, they develop a dynamic technique in which as distribution changes, "most significant" (e.g., largest) wavelet coefficients are updated consequentially.

Bruno et al. (2001) propose a different kind of multidimensional histogram, based on the analysis of the query-workload on it: the *workload-aware histogram*, which they call *STHoles*. Rather than an arbitrary overlap, a *STHoles* histogram $H_{ST}(D)$ allows bucket nesting, thus achieving the definition of the so-called *bucket tree*. Query-workloads are handled as follows: the query result stream Q^R to be analyzed is intercepted and, for each query Q_j belonging to Q^R and for each bucket b_i belonging to the current bucket tree, the number $|Q_j \cap b_i|$ is counted. Then, "holes" in b_i for regions of different *tuple density* are "drilled" and "pulled out" as children of b_i. Finally, buckets of similar densities are merged in such a way as to keep the number of buckets constant. $H_{ST}(D)$ makes use of a tree-like in-memory-data-structure, since the parent-child relationship in a tree is comparable to the nest relationship, and the sibling-sibling relationship is represented by buckets nested within the same external bucket, (nested buckets look-like holes within external buckets—the name *STHoles* comes from this). The construction algorithm of $H_{ST}(D)$ does not take into account the original data set; indeed, the needed information is instead gathered by inspecting the target query-workload and query feedbacks. This amenity makes $H_{ST}(D)$ *self tunable*, that is, adaptable to updates and modifications in the original data cube. On the basis of this approach, a relevant amount of the total storage space available for housing the histogram is invested in representing "heavy-queried regions," thus providing a better approximation for such regions, whereas a fewer storage space is reserved to "lowly-queried regions," admitting some inaccuracy (however, tolerable in OLAP context (Cuzzocrea, 2005a)) for such regions. More specifically, $H_{ST}(D)$ construction algorithm is outlined in the following. Given an input data cube L and a query-workload $QWL = \{Q_0, Q_1, ..., Q_{|QWL|-1}\}$, for each query $Q_j \in QWL$ at iteration *j*, new buckets are generated by intersecting each bucket b_i of the current partition $P_j(L)$ with Q_j, thus determining the set of candidate buckets $U = \{b_c \mid b_c = Q_i \cap b_i, b_i \in P_j(L), Q_i \in QWL, Q_i \cap b_i \neq \emptyset\}$; if such buckets have densities notably different from their parents' ones, then these new buckets are definitively added to the current partition. This process is iterated until the histogram reaches the maximum number of allowed buckets (i.e., the whole available storage space

is consumed). At this point, in order not to exceed the given amount of storage space, after candidate buckets are added to the current partition, the algorithm finds couples of buckets linked together by either a parent-child relationship or a sibling-sibling relationship, and having the closest densities. The size reduction is accomplished by performing either a parent-child or a sibling-sibling merge between the components of these couples, until the histogram fits within the given amount of storage space. A thorough set of experimental results on both real-life and synthetic data sets demonstrates that *STHoles* overcomes performance of *Equi-Depth*, *MHist*, and *GenHist*; in addition to this, (Bruno et al., 2001) authors also show that, on the DBMS Microsoft SQL Server, query-workload analysis overheads introduced by *STHoles* are very low, less than 10 percent of the overall DBMS throughput.

T*unable*-**p***artition-tree* (**TP**-*Tree*), proposed by Cuzzocrea & Wang (2007), is a tree-like, highly-dynamic data structure that codifies a multidimensional histogram for massive (multidimensional) data cubes, denoted by $H_{TP}(D)$, whose partition *varies over time* according to the query-workload against the target OLAP server. For this reason, partitions of $H_{TP}(D)$ are named as *tunable partitions*, and $H_{TP}(D)$ is said to be a "workload-aware" synopsis data structure. This approach resembles that proposed by Bruno et al. (2001). The main contribution of the **TP**-*Tree* proposal with respect to previous techniques consists in introducing models and algorithms having low computational costs, whereas previous techniques are, usually, time-consuming and resource-intensive. Data stored inside buckets of **TP**-*Tree* partitions are obtained by (*i*) *sampling* the input data cube (from this solution, low computational costs required by the **TP**-*Tree* approach follow), and (*ii*) separately representing, storing, and indexing *outliers* via high performance *quad-tree* based (data) structures. Outlier management is another innovative and relevant contribution of the **TP**-*Tree* proposal, as it allows us to provide *probabilistic guarantees* over the degree of approximation of the answers, which is a leading research topic very often neglected by the research community (for instance, see (Cuzzocrea, 2005b)). Finally, the **TP**-*Tree* data organization is characterized by a hierarchical and multi-resolution nature, which allows us to efficiently answer relevant-in-practice classes of OLAP queries, like range-queries.

FUTURE TRENDS

For what regards the basic problem of computing histograms, in order to achieve more and more sophisticated representations beyond actual capabilities of state-of-the-art techniques, next years' challenges for histogram research are the following. (*i*) *Error guarantees*. No one of the surveyed state-of-the-art techniques consider the relevant issue of building histograms capable of ensuring a fixed threshold over the (query) error due to the approximate evaluation of queries against such data structures. Indeed, the next frontier of histogram research is represented by the challenge of ensuring error guarantees over the final data structures; some initiatives devoted to fill the actual gap in this context are (Cuzzocrea & Wang, 2007; Koudas et al., 2000), whereas similar efforts in wavelet-based database and data cube approximation, which is a research issue close to ours, are Garofalakis and Gibbons (2002); Garofalakis and Kumar (2004). (*ii*) *Flexibility for wide families of queries*. One of the most important limitations of histograms is the fact that they are "build-one-use-many-times" data structures, meaning that state-of-the-art techniques construct histograms for a *specific* family of queries, for example, those given by typical query-workloads of the target DBMS/OLAP server, so that it could happen to obtain high accuracy on certain queries, and, contrarily, low accuracy on different queries. This is a problematic issue to be handled, and can be reasonable considered as a leading open problem for histogram research.

Also in-consequence-of influencing insights given by Ioannidis (2003), for what regards the integration of histogram-based techniques with other popular data-intensive techniques, main future directions can be summarized by the following points. (*i*) *Integration of histogram-based and clustering techniques*. Computing *clusters* of a given item set defines the same problem of computing a histogram as a collection of buckets over that item set, given that (*i*) clustering is, similarly to histogram-based ones, a partition-based technique, and (*ii*) buckets are conceptually the same of clusters, since in clustering we want to generate clusters such that intra-cluster distance is small and extra-bucket distance is big, and, symmetrically, in buckets we aggregate data having, for instance, low variance (which, in turn, is a *distance*-based metrics), whereas among buckets we want to obtain big variance. In literature, there exists a wide number of clustering techniques, so that it is very interesting to study how these proposals can be applied to the problem of computing histograms, and, specifically, multidimensional histograms, given that research communities have devoted a great deal of attention to the problem of clustering high-dimensional domains. (*ii*) *Integration of histogram-based and pattern recognition techniques*. Choosing the parameters according to which computing buckets (e.g., variance, skewness, etc.) is similar to the problem of establishing which parameters are similar for items of a given item set in order to group items according to these parameters. The simplest case of such kind of problems is to choose a dimension among all the dimensions of a given domain to grouping/clustering purposes. Specifically, this general problem is recognized in literature as the *pattern recognition problem*. Just like the histograms/clustering symmetry, it is also very interesting to study how pattern recognition techniques can influence

histogram research. (*iii*) *Integration of histogram-based and tree-like indexing techniques*. A hierarchical indexing data structure is very similar to a *hierarchical histogram*, where each level is partitioned according to a criterion that is inspired to classical histogram-based techniques. This similarity leads to, from a side, (*i*) the definition of high-performance indexing data structures for massive data sets that exploit consolidate results of histogram research, and, from the other side, (*ii*) the idea of adopting successful methodologies of RDBMS indexes (e.g., B^+-trees, *R*-trees, etc.) to the problem of computing histograms for one-dimensional and multidimensional domains.

CONCLUSION

Due to an impressive proliferating of data-intensive systems in real-life applications, the problem of compressing massive databases and data cubes play a critical role in database and data warehouse research. On the other hand, this problem has significantly stimulated research communities during the last two decades, and, presently, it continues attracting attention and efforts of research communities, also thanks to innovative, previously-unrecognized knowledge production, processing, and fruition paradigms drawn by emerging technologies like those of data stream management systems and sensor network data management systems. In consequence of this, a plethora of database and data cube compression techniques have been proposed in literature, with different aims and goals (and fortune). Among these, histogram-based techniques are a very popular solution of tackling research challenges posed by compressing massive databases and data cubes, and have been extensively studied and investigated since early 1980 and before. The result we observe at now is a wide literature on histogram-based techniques, which, in our opinion, will continue to stimulate research communities towards investigating new and exciting problems that, basically, concern the core layer of DBMS and OLAP servers, with also important influences on commercial platforms. Given these considerations, in this article we surveyed state-of-the-art histogram-based techniques for compressing databases and data cubes, ranging from one-dimensional to multidimensional data domains. As an additional contribution, a simple-yet-effective taxonomy of histograms has been provided.

REFERENCES

Aboulnaga, A., & Chaudhuri, S. (1999). Self-Tuning Histograms: Building Histograms Without Looking at Data. *Proceedings of the 1999 ACM International Conference on Management of Data*, (pp. 181-192).

Acharya, S., Poosala, V., & Ramaswamy, S. (1999). Selectivity Estimation in Spatial Databases. *Proceedings of the 1999 ACM International Conference on Management of Data*, (pp. 13-24).

Ammoura, A., Zaiane, O.R., & Ji, Y. (2001) Immersed Visual Data Mining: Walking The Walk. *Proceedings of the 18th British National Conference on Databases*, (pp. 202-218).

Bruno, N., Chaudhuri, S., & Gravano, L. (2001). STHoles: A Multidimensional Workload-Aware Histogram. *Proceedings of the 2001 ACM International Conference on Management of Data*, (pp. 211-222).

Buccafurri, F., Furfaro, F., Saccà, D., & Sirangelo, C. (2003). A Quad-Tree based Multiresolution Approach for Two-Dimensional Summary Data. *Proceedings of the 15th International Conference on Scientific and Statistical Database Management*, (pp. 127-140).

Christodoulakis, S. (1984). Implications of certain assumptions in database performance evaluations. *ACM Transactions on Database Systems*, *9*(2), 163-186.

Colliat, G. (1996). OLAP, relational, and multidimensional database systems. *ACM SIGMOD Record*, *25*(3), 64-69.

Cuzzocrea, A. (2005a). Overcoming Limitations of Approximate Query Answering in OLAP. *Proceedings of the 9th IEEE International Database Engineering and Applications Symposium*, (pp. 200-209).

Cuzzocrea, A. (2005b). Providing Probabilistically-Bounded Approximate Answers to Non-Holistic Aggregate Range Queries in OLAP. *Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP*, (pp. 97-106).

Cuzzocrea, A., Saccà, D., & Serafino, P. (2007). Semantics-aware advanced OLAP visualization of multidimensional data cubes. *International Journal of Data Warehousing and Mining*, to appear.

Cuzzocrea, A., & Wang, W. (2007). Approximate range-sum query answering on data cubes with probabilistic guarantees. *Journal of Intelligent Information Systems*, *28*(2), 161-197.

Donjerkovic, D., Ioannidis, Y., & Ramakrishnan, R. (1999). Dynamic histograms: Capturing evolving data sets. *University of Wisconsin-Madison Technical Report CS-TR-99-1396*.

Faloutsos, C., & Kamel, I. (1997). Relaxing the uniformity and independence assumptions using the concept of fractal dimension. *Journal of Computer and System Sciences*, *55*(2), 229-240.

Furfaro, F., Mazzeo, G.M., Saccà, D., & Sirangelo, C. (2005). Hierarchical Binary Histograms for Summarizing Multi-

Dimensional Data. *Proceedings of the 20th Annual ACM Symposium on Applied Computing*, (pp. 598-603).

Garofalakis, M.N., & Gibbons, P.B. (2002). Wavelet Synopses with Error Guarantees. *Proceedings of the 2002 ACM Conference on Management of Data*, (pp. 476-487).

Garofalakis, M.N., & Kumar, A. (2004). Deterministic Wavelet Thresholding for Maximum-Error Metrics. *Proceedings of the 23rd ACM International Symposium on Principles of Database Systems*, (pp. 166-176).

Gibbons, P.B., Matias, Y., & Poosala, V. (1997). Fast Incremental Maintenance of Approximate Histograms. *Proceedings of the 23rd International Conference on Very Large Data Bases*, (pp. 466-475).

Gilbert, A.C., Kotidis, Y., Muthukrishnan, S., & Strauss, M. (2001). Optimal and Approximate Computation of Summary Statistics for Range Aggregates. *Proceedings of the 20th ACM International Symposium on Principles of Database Systems*, 227-236.

Greenwald, M., & Khanna, S. (2001). Space-Efficient Online Computation of Quantile Summaries. *Proceedings of the 2001 ACM Conference on Management of Data*, (pp. 58-66).

Guha, S., Indyk, P., Muthukrishnan, S., & Strauss, M. (2002). Histogramming Data Streams with Fast Per-Item Processing. *Proceedings of the 29th International Colloquium on Automata, Languages and Programming*, (pp. 681-692).

Guha, S., Koudas, N., & Shim, K. (2001). Data Streams and Histograms. *Proceedings of the 33th ACM Symposium on Theory of Computing*, (pp.471-475).

Gunopulos, D., Kollios, G., Tsotras, V.J., & Domeniconi, C. (2000). Approximating Multi-Dimensional Aggregate Range Queries over Real Attributes. *Proceedings of the 2000 ACM Conference on Management of Data*, (pp. 463-474).

Ho, C.-T., Agrawal, R., Megiddo, N., & Srikant, R. (1997). Range Queries in OLAP Data Cubes. *Proceedings of the 1997 ACM International Conference on Management of Data*, (pp. 73-88).

Ioannidis, Y. (2003). The History of Histograms (abridged). *Proceedings of the 29th International Conference on Very Large Data Bases*, (pp. 19-30).

Ioannidis, Y., & Poosala, V. (1995). Balancing Histogram Optimality and Practicality for Query Result Size Estimation. *Proceedings of the 1995 ACM International Conference on Management of Data*, (pp. 233-244).

Ioannidis, Y., & Poosala, V. (1999). Histogram-based Approximation of Set-Valued Query Answers. *Proceedings of the 25th International Conference on Very Large Data Bases*, (pp.174-185).

Jagadish, H.V., Jin, H., Ooi, B.C., & Tan, K.L. (2001). Global Optimization of Histograms. *Proceedings of the 2001 ACM International Conference on Management of Data*, (pp. 223-234).

Jagadish, H.V., Koudas, N., Muthukrishnan, S., Poosala, V., Sevcik, K.C., & Suel, T. (1998). Optimal Histograms with Quality Guarantees. *Proceedings of the 24th International Conference on Very Large Data Bases*, (pp. 275-286).

Kooi, R.P. (1980). The Optimization of Queries in Relational Databases. *PhD Thesis, Case Western Reserve University*.

Koudas N., Muthukrishnan S., & Srivastava D. (2000). Optimal Histograms for Hierarchical Range Queries. *Proceedings of the 19th ACM International Symposium on Principles of Database Systems*, (pp. 196-204).

Malvestuto, F. (1993). A Universal-Scheme Approach to Statistical Databases Containing Homogeneous Summary Tables. *ACM Transactions on Database Systems*, 18(4), 678-708.

Matias, Y., Vitter, J.S., & Wang, M. (1998). Wavelet-Based Histograms for Selectivity Estimation. *Proceedings of the 1998 ACM International Conference on Management of Data*, 448-459.

Matias, Y., Vitter, J.S., & Wang, M. (2000). Dynamic Maintenance of Wavelet-based Histograms. *Proceedings of the 26th International Conference on Very Large Data Bases*, (pp. 101-110).

Muralikrishna, M., & DeWitt, D.J. (1998). Equi-Depth Histograms for Estimating Selectivity Factors for Multi-Dimensional Queries. *Proceedings of the 1998 ACM International Conference on Management of Data*, (pp. 28-36).

Muthukrishnan, S., Poosala, V., & Suel, T. (1999). On Rectangular Partitionings in Two Dimensions: Algorithms, Complexity, and Applications. *Proceedings of the 7th IEEE International Conference on Database Theory*, (pp. 236-256).

Piatetsky-Shapiro, G., & Connell, C. (1984). Accurate Estimation of the Number of Tuples Satisfying a Condition. *Proceedings of the 1984 ACM International Conference on Management of Data*, (pp. 256-276).

Poosala, V. (1997). Histogram-based Estimation Techniques in Database Systems. *PhD Thesis, University of Wisconsin-Madison*.

Poosala, V., & Ganti, V. (1999). Fast Approximate Answers to Aggregate Queries on a Data Cube. *Proceedings of the 11th International Conference on Statistical and Scientific Database Management*, (pp. 24-33).

Poosala, V., & Ioannidis, Y. (1997). Selectivity Estimation Without the Attribute Value Independence Assumption. *Proceedings of the 23rd International Conference on Very Large Data Bases*, (pp. 486-495).

Poosala, V., Ioannidis, Y., Haas, P.J., & Shekita, E.J. (1996). Improved Histograms for Selectivity Estimation of Range Predicates. *Proceedings of the 1996 ACM International Conference on Management of Data*, (pp. 294-305).

Selinger, P., Astrahan, M., Chamberlin, D., Lorie, R., & Price, T. (1979). Access Path Selection in a Relational Database Management System. *Proceedings of the 1979 ACM International Conference on Management of Data*, (pp. 23-34).

Shoshani, A. (1997). OLAP and Statistical Databases: Similarities and Differences. *Proceedings of the 16th ACM International Symposium on Principles of Database Systems*, (pp. 185-196).

Stolte, C., Tang, D. & Hanrahan, P. (2002). Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, *8*(1), 52-65.

Thaper, N., Guha, S., Indyk, P., & Koudas, N. (2002). Dynamic Multidimensional Histograms. *Proceedings of the 2002 ACM International Conference on Management of Data*, (pp. 428-439).

Vitter, J.S., Wang, M., & Iyer, B. (1998). Data Cube Approximation and Histograms via Wavelets. *Proceeding of the 7th ACM International Conference on Information and Knowledge Management*, 96-104.

KEY TERMS

Multidimensional OLAP (MOLAP): An in-memory-storage model that represents a multi-dimensional data cube in form of a multi-dimensional array.

OLAP Query: A query defined against a data cube that introduces a multidimensional range (via specifying an interval for each dimension of the data cube) and a SQL aggregate operator, and returns as output the aggregate value computed over cells of the data cube contained in that range.

Online Transaction Processing (OLTP): A methodology for representing, managing and querying DB data generated by user/application transactions according to flat (e.g., relational) schemes.

Online Analytical Processing (OLAP): A methodology for representing, managing and querying massive DW data according to multidimensional and multi-resolution abstractions of them.

Relational Query: A query defined against a database that introduces some predicates (e.g., Boolean) over tuples stored in the database, and returns as output the collection of tuples satisfying those predicates.

Selectivity of an OLAP Query: A property of an OLAP query that estimates the cost required to evaluate that query. It is usually model in terms of the geometrical volume of the query range.

Selectivity of a Relational Query: A property of a relational query that estimates the cost required to evaluate that query. It is usually model in terms of the number of tuples involved by the query.

Historical Overview of Decision Support Systems (DSS)

Udo Richard Averweg
eThekwini Municipality and University of KwaZulu-Natal, South Africa

INTRODUCTION

During the late 1970s the term "decision support systems" was first coined by P. G. W. Keen, a British Academic then working in the United States of America. In 1978, Keen and Scott Morton published a book entitled, *Decision Support Systems: An Organizational Perspective* (Keen & Scott Morton, 1978), wherein they defined the subject title as computer systems having an impact on decisions where computer and analytical aids can be of value but where the manager's judgment is essential. Information systems (IS) researchers and technologists have developed and investigated decision support systems (DSS) for more than 35 years (Power, 2003b).

The structure of this article is as follows: The background to DSS will be given. Some DSS definitions, a discussion of DSS evolution, development of the DSS field and frameworks are then presented. Some future trends for DSS are then suggested.

BACKGROUND

Van Schaik (1988) refers to the early 1970s as the era of the DSS concept because during this period the concept of DSS was introduced. DSS was a new philosophy of how computers could be used to support managerial decision-making. This philosophy embodied unique and exciting ideas for the design and implementation of such systems. There has been confusion and controversy in respect of the interpretation of the decision support system notion and the origin of this notion originated in the following terms:

- **Decision** emphasises the primary focus on decision-making in a problem situation rather than the subordinate activities of simple information retrieval, processing or reporting.
- **Support** clarifies the computer's role in aiding rather than replacing the decision maker.
- **System** highlights the integrated nature of the overall approach, suggesting the wider context of machine, user and decision environment.

DSS deal with semi-structured and some unstructured problems.

DECISION SUPPORT SYSTEMS

With the ever-increasing advances in computer technology, new ways and means of computer-assisted decision-making was born. As a result hereof, over the passage of time, different DSS definitions arose:

- Little (1970) defines DSS as a "model-based set of procedures for processing data and judgments to assist a manager in his decision making (*sic*)."
- The classical definition of DSS, by Keen and Scott Morton (1978), states that "Decision Support Systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semi-structured problems."
- Mann and Watson (1984) state that "a decision support system is an interactive system that provides the user with easy access to decision models and data in order to support semi-structured and unstructured decision-making tasks."
- Bidgoli (1989) defines DSS as "a computer-based information system consisting of hardware/software and the human element designed to assist any decision-maker at any level. However, the emphasis is on semi-structured and unstructured tasks."
- Sprague and Watson (1996) define a DSS as computer-based systems that help decision makers confront ill-structured problems through direct interaction with data and analysis models.
- Sauter (1997) notes that DSS are computer-based systems that bring together information from a variety of sources, assist in the organisation and analysis of information and facilitate the evaluation of assumptions underlying the use of specific models.
- Turban, Rainer, and Potter (2005) broadly define a DSS as "a computer-based information system that combines models and data in an attempt to solve semi-structured and some unstructured problems with extensive user involvement."

From these definitions it seems that the basis for defining DSS has been developed from the perceptions of what a DSS

does (e.g., support decision-making in semi-structured or unstructured problems) and from ideas about how a DSS's objectives can be accomplished (e.g., the components required and the necessary development processes).

Bidgoli (1989) contends that as the DSS field is in a state of flux, an exact definition of DSS is elusive. Turban (1995) indicates that previous researchers have collectively ignored the central issue in DSS; that is, "support and improvement of decision-making". Bidgoli (1989) suggests that there are several requirements for a DSS which must embrace a definition of a DSS. These are that a DSS:

- requires hardware;
- requires software;
- requires human elements (designers and end-users);
- is designed to support decision-making;
- should help decision makers at all levels; and
- emphasises semi-structured and unstructured tasks.

Turban (1995) states that there is no consensus on what a DSS is and there is therefore no agreement on the characteristics and capabilities of DSS. As the definition by Turban et al. (2005) underscores Bidgoli's (1989) DSS requirements, for the purposes of this article, the DSS definition by Turban et al. (2005) will be used.

Evolution of DSS

During the 1970s and 1980s, the concept of DSS grew and evolved into a field of research, development and practice (Sprague & Watson, 1996). Clearly DSS was both an evolution and a departure from previous types of computer support for decision-making.

Currently DSS can be viewed as a third generation of computer-based applications. Sprague and Watson (1996) note that initially there were different conceptualisations about DSS. Some organisations and scholars began to develop and research DSS which became characterised as *interactive* computer based systems which *help* decision makers utilise *data* and *models* to solve *unstructured* problems. According to Sprague & Watson (1974), the unique contribution of DSS resulted from these key words. However, a serious definitional problem arose in that the words had certain "intuitive validity"—any system that supports a decision (in any way) is a "decision support system". This term had such an instant intuitive appeal that it quickly became a "buzz word" (Sprague & Watson, 1996). However, neither the restrictive nor the broad DSS definition provided guidance for understanding the value, the technical requirements or the approach for developing and implementing a DSS. For a discussion of DSS implementation, see, for example, Averweg (1998).

Development of the DSS Field

According to Sprague and Watson (1996), DSS evolved as a "field" of study and practice during the 1980s. During the early development of DSS, several principles evolved. Eventually, these principles became a widely accepted "structural theory" or framework—see Sprague and Carlson (1982). The four most important of these principles are summarised:

- **The DDM Paradigm:** The technology for DSS must consist of three sets of capabilities in the areas of **d**ialog, **d**ata and **m**odelling and what Sprague and Carlson call the DDM paradigm. The researchers make the point that a good DSS should have *balance* among the three capabilities. It should be *easy to use* to allow non-technical decision makers to interact fully with the system. It should have access to a *wide variety of data* and it should provide *analysis and modelling* in a variety of ways. Sprague and Watson (1996) suggest that many early systems adopted the name DSS when they were strong in only one area and weak in the other. Figure 1 shows the relationship between these components in more detail and it should be noted that the models in the model base are linked with the data in the database. Models can draw coefficients, parameters and variables from the database and enter results of the model's computation in the database. These results can then be used by other models later in the decision-making process.
  Figure 1 also shows the three components of the dialog function wherein the database management system (DBMS) and the model base management system (MBMS) contain the necessary functions to manage the data base and model base respectively. The dialog generation and management system (DGMS) manages the interface between the user and the rest of the system.
- **Levels of Technology:** Three levels of technology are useful in developing DSS and this concept illustrates the usefulness of configuring *DSS tools* into a *DSS generator* which can be used to develop a variety of *specific DSS* quickly and easily to aid decision makers—see Figure 2. The system which actually accomplishes the work is known as the *specific DSS*, shown as the circles at the top of the diagram. It is the software/hardware that allow a specific decision maker to deal with a set of related problems. The second level of technology is known as the *DSS generator*. This is a package of related hardware and software which provides a set of capabilities to quickly and easily build a specific DSS. The third level of technology is *DSS tools* which facilitate the development of either a DSS generator or a specific DSS.

Figure 1. The components of DSS (Source: Adapted from Sprague & Watson, 1996)

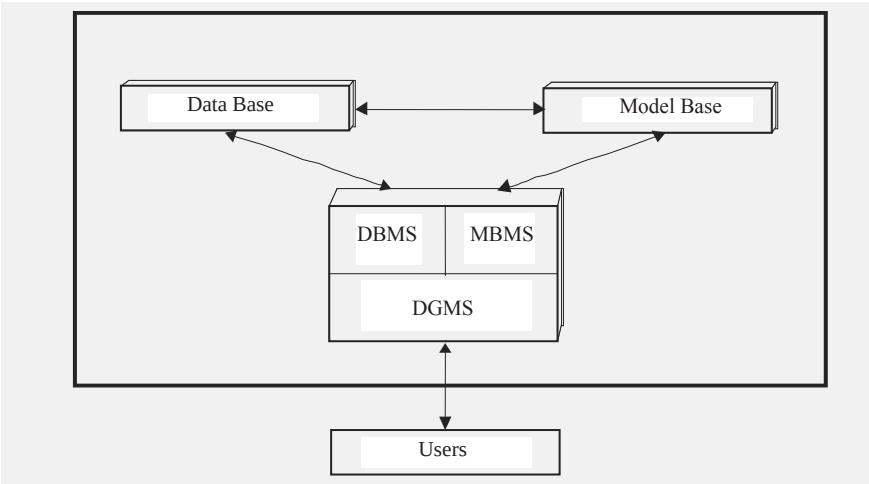
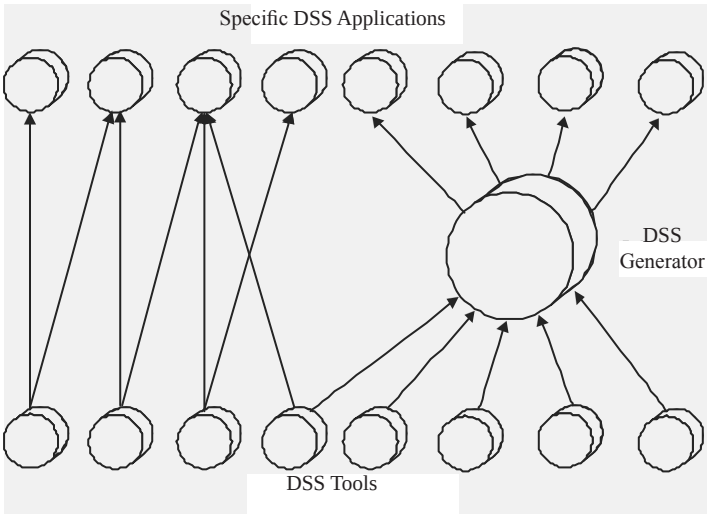


Figure 2. Three levels of DSS technology (Source: Adapted from Sprague & Watson, 1996)



While new technologies such as World Wide Web ("Web") browsers and data warehouses have emerged since Sprague and Watson's (1996) conceptual framework, nowadays the framework is still relevant.

- **Iterative Design:** Instead of the traditional development process, DSS require a form of iterative development which allows them to evolve and change as the problem or decision situation changes. They need to be built with short, rapid feedback from users thereby ensuring that development is proceeding correctly. In essence they must be developed to permit change quickly and easily.
- **Organisational Environment:** The effective development of DSS requires an organisational strategy to build an environment within which such systems can originate and evolve. The environment includes a group of people with interacting roles, a set of software and hardware technology, a set of data sources and a set of analysis models.

The IS called DSS are not all the same. DSS differ in terms of capabilities and targeted users of a specific system and how the DSS is implemented and what it is called (Power, 2003a). Some DSS focus on data, some on models and some on facilitating collaboration and communication. DSS can also differ in terms of targeted users, for example, a "primary" user or "generic" users.

Holsapple and Whinston (1996) identified five specialised types of DSS:

- Text-oriented;
- Database-oriented;
- Spreadsheet-oriented;
- Solver-oriented; and
- Rule-oriented.

Donovan and Madnick (1977) classified DSS as *ad hoc* DSS or institutional DSS. An *ad hoc* DSS supports problems that are not anticipated and which are not expected to reoccur. An institutional DSS supports decisions that reoccur. Hackathorn and Keen (cited in Power, 2003a) identified DSS into three interrelated categories:

- Personal DSS;
- Group DSS; and
- Organisational DSS.

DSS Frameworks

Power (2003a) suggests that the following DSS frameworks help categorise the most common DSS currently in use:

- **Communications-Driven DSS:** These systems are built using communication, collaboration and decision support technologies.
- **Data-Driven DSS:** These systems analyse large "pools of data" found in major organisational systems and they support decision-making by allowing users to extract useful information that was previously buried in large quantities of data. Often data from various transactional processing systems (TPS) are collected in data warehouses for this purpose. Online analytical processing (commonly known as OLAP) and data mining can then be used to analyse the data.
- **Document-Driven DSS:** These systems integrate a variety of storage and processing technologies to provide complete document retrieval and analysis.
- **Knowledge-Driven DSS:** These systems contain specialised problem-solving expertise wherein the "expertise" consists of knowledge about a particular domain (and understanding of problems within that domain) and "skill" at solving some of those problems.
- **Model-Driven DSS:** Early DSS developed in the late 1970s and 1980s were model driven as they were primarily standalone systems isolated from major organisational IS that used some type of model to perform "what if" and other kinds of analysis. Such systems were often developed by end-user groups or divisions not under central IS control (Laudon & Laudon, 1998). A DSS is not a black box—it should provide the end-user with control over the models and interface representations used (Barbosa & Hirko, 1980). Model-driven DSS emphasise access to and manipulation of a model.

Watson (2005) suggests that "I don't think that we need to find a single theory or framework. Furthermore, I don't think that we will see a single overarching theory emerge. Rather, there will be multiple theories, each one being appropriate for specific situations". Despite all the rapid developments of the late 1980s, 1990s, and early 2000s, DSS as a field is now at a crossroads. Some functions that were once considered part of DSS now appear to be migrating to other areas. For example, Watson (2005) suggests that there is an increasing trend to integrate and embed decision support applications into operational systems (e.g., fraud detection system embedded in credit card processing).

FUTURE TRENDS

In future, it is envisaged that traditional DSS applications will be extended to a larger number of potential applications where the data required is only an interim stage or a subset of the information required for the decision. This will require the construction of DSS where the end-user can concentrate on the variables of interest in their decision while "other" processing is performed without the need of extensive end-user interaction. Some future trends for DSS are suggested:

- Organisations that consolidate there is into a single environment reduce administration and license costs. By consolidating organisational data into a Web visualisation application, will facilitate better decision support.
- All organisations use metrics and key performance indictors to undertaken business and remain competitive. With the advent of Web-based technologies (e.g., portal technologies), a decision support portal will be able to present key information to the right audience.
- In the future all data collection and analysis will be automated. This will "free up" domain experts from verifying the validity of data from TPS and data warehouses allowing them to *act* on the information from DSS instead.
- There will be an increase in visualised information in context with user-centric displays. By having the most recent data correlated and aggregated, will allow for better decisions and which are more relevant to a user's current conditions.
- There will be a surge to use advanced display techniques to highlight key issues. Consequently the design of future DSS interfaces will receive greater prominence since the interface should bring attention to the most important areas almost immediately.
- Decision support technology will continue to broaden to include monitoring, tracking and communication tools to support the overall process of unstructured

problem solving. The broadening of this technology will be as a result of an increased availability of mobile computing and communication.

CONCLUSION

DSS continue to impact decision-making in organisations and this is largely dependent on the nature of the application. In order that optimal solutions may be identified, more alternatives may need to be explored and some decisions may need to be automated. The Internet and the Web have accelerated developments in decision support and decision-making and nowadays provide a new research focus area for DSS development and implementation.

REFERENCES

Averweg, U. R. F. (1998). *Decision support systems: Critical success factors for implementation.* Master of Technology: Information Technology dissertation, M L Sultan Technikon, Durban, South Africa.

Barbosa, L. C., & Hirko, R. G. (1980, March). Integration of algorithmic aids into decision support systems. *MIS Quarterly, 4*, 1-12.

Bidgoli, H. (1989). *Decision support systems: Principles and practice*. St Paul: West Publishing Company.

Donovan, J. J., & Madnick, S. E. (1977). Institutional and ad hoc DSS and their effective use. *Data Base, 8*(3).

Holsapple, C. W., & Whinston, A. B. (1996). *Decision support systems: A knowledge-based approach.* Minneapolis: West Publishing Co.

Keen, P. G. W., & Scott Morton, M. S. (1978). *Decision support systems: An organizational perspective.* Reading: Addison-Wesley.

Laudon, K. C. & Laudon, J. P. (1998). *Management information systems.* New Jersey: Prentice-Hall, Inc.

Little, J. D. C. (1970). Models and managers: The concept of a decision calculus. *Management Science, 16*(8).

Mann, R. I., & Watson, H. J. (1984). A contingency model for user involvement in DSS development. *MIS Quarterly, 8*(1), 27-38.

Power, D. J. (2003a). Categorizing decision support systems: A multidimensional approach (Chapter 2). In M. Mora, G. Forgionne, & J. N. D. Gupta (Eds.), *Decision making support systems: Achievements and challenges for the new decade* (pp. 20-27). Hershey: Idea Group Publishing.

Power, D. J. (2003b). *A brief history of decision support systems.* DSSResources.COM (Editor), version 2.8, 31 May. Retrieved from http://www.dssresources.com/history/dsshistory.html

Sauter, V. L. (1997). *Decision support systems: An applied managerial approach*. New York: John Wiley & Sons, Inc.

Sprague, R. H., & Carlson, E. D. (1982). *Building effective decision support systems.* Englewood Cliffs, NJ: Prentice-Hall.

Sprague, R. H., & Watson, M. J. (1974). Bit by bit: Toward decision support systems. *California Management Review, 22*(1), 60-67.

Sprague, R. H., & Watson, H. J. (1996). *Decision support for management.* Upper Saddle River: Prentice-Hall.

Turban, E. (1995). *Decision support and expert systems.* Englewood Cliffs: Prentice-Hall.

Turban, E., Rainer, R. K., & Potter, R. E. (2005). *Introduction to information technology*. Hoboken: John Wiley & Sons.

Van Schaik, F. D. J. (1988). *Effectiveness of decision support systems*. PhD dissertation, Technische Universiteit Delft, Holland.

Watson, H. (2005). *Hugh Watson: Understanding computerized decision support.* Thought Leader Interview by Dan Power, Editor DSSResources.com, October. Retrieved from http://www.dssresources.com/interviews/watson/watson11042005.html

KEY TERMS

Analytical Processing: Involves analysis of accumulated data, frequently by end-users in an organisation. Analytical processing activities include data mining, decision support and querying.

Communications-Driven DSS: Systems built using communication, collaboration and decision support technologies.

Data-Driven DSS: These systems analyse large "pools of data" found in major organisational systems and they support decision-making by allowing users to extract useful information that was previously buried in large quantities of data.

Data Warehouse: A repository of subject-oriented historical data that is organised to be accessible in a form readily acceptable for analytical processing activities.

Document-Driven DSS: These systems integrate a variety of storage and processing technologies to provide complete document retrieval and analysis.

Information Systems (IS): A combination of technology, people and process to capture, transmit, store, retrieve, manipulate and display information.

Knowledge-Driven DSS: These systems contain specialised problem-solving expertise wherein the "expertise" consists of knowledge about a particular domain.

Model-Driven DSS: Model-driven DSS emphasise access to and manipulation of a model.

History of Artificial Intelligence

Attila Benkő
University of Pannonia, Hungary

Cecília Sik Lányi
University of Pannonia, Hungary

INTRODUCTION

George Boole was the first to describe a formal language for logic reasoning in 1847. The next milestone in artificial intelligence history was in 1936, when Alan M. Turing described the Turing-machine. Warren McCulloch and Walter Pitts created the model of artificial neurons in 1943, and it was in 1944 when J. Neumann and O. Morgenstern determined the theory of decision, which provided a complete and formal frame for specifying the preferences of agents. In 1949 Donald Hebb presented a value changing rule for the connections of the artificial neurons that provide the chance of learning, and Marvin Minsky and Dean Edmonds created the first neural computer in 1951.

Artificial intelligence (AI) was born in the summer of 1956, when John McCarthy first defined the term. It was the first time the subject caught the attention of researchers, and it was discussed at a conference at Dartmouth. The next year, the first general problem solver was tested, and one year later, McCarty—regarded as the father of AI—announced the LISP language for creating AI software. Lisp, which stands for list processing, is still used regularly today.

Herbert Simon in 1965 stated: "Machines will be capable, within twenty years, of doing any work a man can do." However, years later scientists realized that creating an algorithm that can do anything a human can do is nearly impossible. Nowadays, AI has a new meaning: creating intelligent agents to help us do our work faster and easier (Russel & Norvig, 2005; McDaniel, 1994; Shirai & Tsujii, 1982; Mitchell, 1996; Schreiber, 1999).

Perceptrons was a demonstration of the limits of simple neural networks published by Marvin Minsky and Seymour Papert in 1968. In 1970, the first International Joint Conference on Artificial Intelligence was held in Washington, DC.

PROLOG, a new language for generating AI systems, was created by Alain Colmerauer in 1972. In 1983, Johnson Laird, Paul Rosenbloom, and Allen Newell completed CMU dissertations on SOAR.

BACKGROUND

In 1950 Alan Turing suggested a definition for deciding whether software is intelligent or not. In his theory the software's intelligent behavior can be measured like a human intellectual efficiency. The software is intelligent when a human being does not know if he or she is chatting with the software or with another human. That test was called the Turing test, and here is how it works: if the software passes the test, it is called intelligent software—also called intelligent agent—which percepts the environment with sensors and acting with effectors.

The term *embodied conversational agent* (ECA) (Cassel, 2007; Huget, 2003; Cassel, Sullivan, Prevost, & Churchill, 2000) is used for special software or hardware as an extension of an intelligent agent, not just because these are able to communicate with the user via natural language, but also for their emotion system (Benkő & Sik Lányi, 2007). There are many emotional models (Ruebenstrunk, 1998) for creating an embodied conversational agent, including:

- Theory of Ortony, Clore and Collins;
- Theory of Roseman;
- Theory of Scherer;
- Theory of Frijda; and
- Theory of Oatley and Johnson-Laird.

Virtual reality (VR) can be used for designing and testing an ECA because the developing process can be easier and cheaper with VR technology (Ortiz, Oyarzun, Carretero, & Nestor, 2006; Takacs & Kiss, 2003). An avatar is a spatial creature that usually symbolizes or simulates a human being in exterior and in behavior also. The next article describes the VTR, the modeled emotions, and the avatars that were created for a virtual therapy room.

INTELLIGENT AGENTS WITH EMOTIONS

The term *embodied conversational agent* (Cassel et al., 2000) stands for intelligent software with an emotion system

that simulates human emotions. Churchill, Cook, Hodgson, Prevost, and Sullivan (2000) described the method of designing ECA using scenarios and storyboards. The levels of description and the embodied agent issues were determined as personality, appearance, communication, and domain expertise. The dimensions of agent embodiment were also described by them.

The method of the communication via conversational dialogs and the aspects of human-computer interaction were discussed by Ball and Breese (2000). Computational models of emotions and the problem of recognizing human emotions were also discussed.

In an ambient intelligent environment, it is important for the user to have a good human-centric computer interface. It is possible to design a more natural interface using the techniques of natural language processing during the development. With such an interface the user is able to communicate with the intelligent environment via natural language. A virtual therapy room (Benkő & Sik Lányi, 2007) was created for aphasic patients. With the VTR the therapist can furnish the room, and the patient can practice at home also. During the therapy, patients can learn to express their emotions in a good way. Therapy is represented by special exercises created for practicing communication. Avatars in a virtual therapy room with the methods of artificial intelligence can answer the user's questions. Emotions can also be expressed by the avatars. The emotion system of an avatar was described as a deterministic finite automaton, and it uses the emotion model of Oatley.

An avatar is the virtual therapist or the virtual patient in the exercise. Usually, there are several virtual patients and only one therapist. They are sitting behind the table. The arrangement can be altered, but it is essential for the patient to think that the user interface is convenient and appropriate to solve exercises. Exercises are appearing on the virtual blackboard. Patients are capable of doing exercises like in a real therapy room.

Oatley assumed a hierarchy of parallel working processing instances, which work on asynchronously different tasks. A central control system—also called an operating system—manages the instances. The control system has a model from the entire system. Two kinds of communication—symbolical and emotional—exist between the modules. The name "Communicative Theory of Emotions" was chosen because it is the task of emotions to convey certain information to all modules of the overall system.

Introverted or extroverted personalities can be created with a push-down automaton-controlled emotion system. The methods of artificial intelligence can be used for creating more intelligent ECAs. The therapist decides the avatars' exterior and interior characteristics based on the type of the therapy and the abilities of the aphasic patients. Generating the ECA's knowledge base is also important in order to make progression in therapy and develop the communication abilities of the patients. VTR can also be used for language teaching because the exercises can be modified by the therapist. Integrating a text-to-speech engine and sound recognition engine provides user friendliness. The emotion model can be improved also.

El-Nasr, Ioerger, and Yen (2000) presented a fuzzy logic adaptive model of emotions. A new computational model of emotions was proposed that can be incorporated into intelligent agents and other complex, interactive software. It uses a fuzzy logic representation to map events and observations to emotional states. The model includes several inductive learning algorithms for learning patterns of events, associations among objects, and expectations. The adaptive components of the model are crucial to users' assessments of the believability of the agent's interactions.

FUTURE TRENDS

Dr. Rodney Brooks, the director of MIT's artificial intelligence laboratory, says that right now, AI is about at the same place as the PC industry was in 1978. Within 30 years, we will have an understanding of how the human brain works that will give us "templates of intelligence" for developing strong AI, and by 2050, our lives will be populated with all kinds of intelligent robots.

"Referring to Spielberg's movie AI in which a company creates a robot that bonds emotionally like a child, Dr. Brooks says: 'A scientist doesn't wake up one day and decide to make a robot with emotions.' Despite the rapid advance of technology, the advent of strong AI will be a gradual process, they say. 'The road from here to there is through thousands of these benign steps,' Mr. Ray Kurzweil says." (BBC News, 2001)

CONCLUSION

Because of the combinatorial explosion, there is no algorithm that can solve all types of problems. Every agent has its own knowledge base, each fit for only a specific field of science. In the early 1980s these types of software were called expert systems because they have special knowledge based on human experiences. Nowadays, expert systems are used in hospitals to help the doctor diagnose a disease or in geology to identify materials. There are many types of expert systems, and they can give very good advice to humans. In the future, expert systems are going to be necessary in our everyday life as well.

However, in addition to expert systems, another form of artificial intelligence aims to understand the operation of how a human brain functions and simulate its emotions. There are several emotion models already in place that can nearly

describe the emotion changes of a human, but there is no full representation of brain functions. It is still a philosophical question if there is any.

REFERENCES

AIBO. (2007). *Homepage.* Retrieved from: http://support.sony-europe.com/aibo/index.asp

Ball, G., & Breese, J. (2000). *Emotion and personality in a conversational agent, embodied conversational agents* (pp. 189-219). Cambridge, MA: MIT Press.

Benkő, A., & Sik Lányi, C. (2007, November 22-24). Developing intelligent emotional agent. *Proceedings of the Regional Conference on Embedded and Ambient Systems* (pp. 75-80), Budapest.

BBC News. (2001). *Predicting AI's future.* Retrieved from http://news.bbc.co.uk/1/hi/in_depth/sci_tech/2001/artificial_intelligence/1555742.stm

Cassel, J. (2007). *Research*. Retrieved from: http://web.media.mit.edu/~justine/research.html

Cassel, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied conversational agents* (pp. 212-214). Cambridge, MA: MIT Press.

Churchill, E., Cook, L., Hodgson, P., Prevost, S., & Sullivan, J. (2000). *Designing embodied conversational agent allies, embodied conversational agents* (pp. 64-94). Cambridge, MA: MIT Press.

Engadget. (2006). Retrieved from http://www.engadget.com/2006/07/21/hiroshi-ishiguro-builds-his-evil-android-twin-geminoid-hi-1/

El-Nasr, M., Yen, J., & Ioerger, T. (2000). FLAME—fuzzy logic adaptive. *Autonomous Agents and Multi-Agent Systems, 3,* 219-257.

Huget, M. (2003). *Communication in multiagent systems* (pp. 318-322). Berlin: Springer-Verlag.

McDaniel, G. (1994). *IBM dictionary of computing* (pp. 32-33, 586). New York: McGraw-Hill.

Mitchell, M. (1996). *An introduction to genetic algorithms* (pp. 35-81). Cambridge, MA: MIT Press.

Ortiz, A., Oyarzun, D., Carretero, M., & Nestor, V. (2006). *Virtual characters as emotional interaction element in the user interfaces* (p. 238).

RoboCup. (2007). *Homepage.* Retrieved from http://www.robocup.org/

Ruebenstrunk, G. (1998). *Emotional computers: Computer models of emotions and their meaning for emotion-psychological research.* Retrieved from http://www.ruebenstrunk.de/emeocomp/4e.HTM

Russel, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach* (p. 69). Panem-Prentice Hall.

Schreiber, G. (1999). *Knowledge engineering and management. The Commonkads methodology* (pp. 13-23). Cambridge, MA: MIT Press.

Shirai, Y., & Tsujii, J. (1982). *Artificial intelligence.* Tokyo: Iwanami Shoten.

Takacs, B., & Kiss, B. (2003). Virtual human interface. A photo-realistic digital human. *IEEE Computer Graphics and Applications, 23*(5), 38-45.

KEY TERMS

Artificial Intelligence: The capability of a device to perform functions that are normally associated with human intelligence, such as reasoning, learning, and self-improvement.

Backward Reasoning: Searching from the initial state to the final state in, for example, action planning.

Bayesian Network: A mathematic model in graphic form that represents a set of variables and their probabilistic independencies. It can be used, for example, to calculate the probability of a patient having a specific disease.

Expert System: A computer program that contains subject-specific knowledge of human experts. Such systems are used for giving advice. Also known as a knowledge-based system.

Forward Reasoning: Searching from the initial state to the final state in, for example, action planning.

Fuzzy Logic: Techniques for reasoning under uncertainty. It is capable of working with concepts such as 'thin', 'fat', 'long', and 'short', if there is no exact data for supporting the decision.

General Problem Solver: Uses means-ends-analysis heuristic for solving formalized symbolic problems. A GPS computer program solves simple problems that can be formalized such as the Towers of Hanoi.

Genetic Algorithm: A method of evolutionary computation for problem solving. There are states also called sequences and a set of possibility final states. Methods of mutation are used on genetic sequences to achieve better sequences.

Heuristic: General advice that is usually efficient but sometimes cannot be used; also it is a validate function that adds a number to the state of the problem.

Intelligent Agent: An agent is an entity with the capability to observe and act in an environment. It is intelligent if it interacts like a human being. It can be a robot or a software system, depending on the environment.

Knowledge Engineering: Related to mathematical logic, and building, maintaining, and developing knowledge-based systems.

Knowledge Representation: Describes how information can be stored efficiently. It uses state space and the information is represented with a graph. States are represented with nodes and actions with arcs.

Neural Network: A computation system containing a set of connected elements to solve arithmetic problems. The basis of the neural network computation is to analyze how the brain works and simulate it.

Robot: A device that performs programmed operations or that operates by remote control. A robot senses external feedback derived from ongoing operations and reacts to sensed data by modifying its actions accordingly.

Searching: A searching problem contains the initial state, the operators, the final state, and the cost of searching. The searching is efficient if it finds an optimal solution if a solution exists and it can be reached in a short time. Searching methods include: blind search (breadth-first, depth-first, depth-limited, iterative deepening, iterative broadening, uniform-cost) and heuristic search (hill-climbing, best-first, A, A*, IDA*, SMA*, simulated annealing).

Strong AI: The main goal of strong AI is to create an AI agent that can *think* and have a *mind.*

Turing Test: Test created by Alan M. Turing, who said that a machine can think. The test analyzes if a person in an isolated room can decide exactly who he or she is chatting with—a human being or an AI agent. The agent is in the other room and the chatting is in written form.

Weak AI: Refers to software to study or solve problems and reasoning tasks that do not need the full range of human cognitive abilities.

History of Artificial Intelligence Before Computers

Bruce MacLennan
University of Tennessee, USA

INTRODUCTION

The history of artificial intelligence (AI) is commonly supposed to begin with Turing's (1950) discussions of machine intelligence, and to have been defined as a field at the 1956 Dartmouth Summer Research Project on Artificial Intelligence. However, the ideas on which AI is based, and in particular those on which symbolic AI (see below) is based, have a very long history in the Western intellectual tradition, dating back to ancient Greece (see also McCorduck, 2004). It is important for modern researchers to understand this history for it reflects problematic assumptions about the nature of knowledge and cognition: assumptions that can impede the progress of AI if accepted uncritically.

BACKGROUND

Symbolic AI is the approach to artificial intelligence that has dominated the field throughout most of its history and remains important. It is based on the physical symbol system hypothesis, enunciated by Newell and Simon (1976), which asserts, "A physical symbol system has the necessary and sufficient means for general intelligent action." In effect, it implies that knowledge is represented in the brain by language-like structures, and that thinking is a computational process that rearranges these structures according to formal rules. This view has also dominated cognitive science, which applies computational concepts to understanding human cognition (Gardner, 1985).

Many symbolic AI systems are based on formal logic, which represents propositions by symbolic structures, in which all meaning is conveyed in the structure's form, and which implements inference by the mechanical manipulation of those structures. Therefore, we will discuss the origins of formal logic and of the idea that knowledge and inference can be represented in this way. We will also consider the combinatorial methods used before the invention of computers as well as in modern AI for generating possible solutions to a problem, which leads to combinatorial explosion, a fundamental limitation of symbolic AI. Then we describe early modern attempts to design comprehensive knowledge representation languages (predecessors of those used in symbolic AI) and mechanical inference machines. We conclude with a mention of alternative views of knowledge and cognition.

THE HISTORICAL ROOTS OF SYMBOLIC AI

Formal Logic

It is surprising, perhaps, that the original inspiration for symbolic knowledge representation can be found in ancient Greece, in particular in Pythagorean number theory (Burkert, 1972, Riedweg, 2005). In ancient Greece, as in many cultures, ancient and modern, pebbles were used for calculation by being moved in grooves in a similar way to the beads on an abacus. Indeed, the Latin word for *pebble* is *calculus*, and our word *calculate* comes from this manipulation of *calculi* (pebbles). In logic and mathematics, we use the word *calculus* for any system of notation in which we can accomplish some purpose by the manipulation of tokens according to formal, game-like, mechanical rules. (For example we have differential and integral calculi in mathematics and propositional and predicate calculi in logic.) To the extent that the rules are purely mechanical, they can, in principle, be carried out by a machine, which is why calculi are important in AI; if a process can be reduced to a calculus, it can be calculated by a machine.

The ancient Pythagoreans (Pythagoras, 572–497 B.C.E.) investigated number theory by means of arrangements of pebbles (Burkert, 1972; Riedweg, 2005). For example, they observed that certain numbers could be arranged into a square shape, and we still call these numbers *squares*. However, they also investigated triangular numbers as well as rectangles, pentagons, cubes, pyramids, and so forth. Although they did not prove theorems in the modern sense, they were able to demonstrate the truth of theorems in number theory by means of these arrangements. Thus, they discovered calculi could be used for reasoning as well as computation.

According to tradition, Pythagoras was the first to explain consonant musical intervals in terms of numerical ratios (Burkert, 1972). For example, a string one-half the length of another string sounds an octave higher; the shorter of two strings of lengths with the ratio 2:3 sounds a fifth higher, and so forth. Thus, a subtle perceptual distinction (the rela-

tive consonance of pitches) could be rendered logical and rational by reducing it to numerical ratios (Greek *logoi* and Latin *rationes*, terms that also refer to the articulation of thought in words or symbols; Maziarz & Greenwood, 1968). It is an example of the representation of expertise in terms of formal structures; judgments of consonance can be replaced by calculation.

The Pythagoreans believed that everything could be reduced to numbers and thus made intelligible, rational, and logical (Burkert, 1972; Burnet, 1930). Therefore, they were committed to the idea that all knowledge could be represented in terms of arrangements of otherwise meaningless tokens, that is, in formal structures (and hence, we may conclude, in computer data structures).

Aristotle (384-322 B.C.E.) is known as the originator of the science of logic, but two of his contributions in this area are especially relevant to AI. First, he began the development of formal logic by showing that valid inference could be distinguished from invalid inference on the basis of its form rather than on the meaning of its particular terms (words). In other words, Aristotle showed that valid inference is a matter of syntax (the grammatical form of an argument) rather than semantics (its meaning). This is important because it shows how inference can be carried out by the manipulation of symbols independently of their meaning, which means that, in principle, inference is a kind of computation. Stated differently, there is a calculus of valid inference.

Aristotle also began the study of modal logic, that is, logic in which propositions are not simply true or false, but in which the propositions may be possible, impossible, necessary, or contingent (Bocheński, 1970; Kneale & Kneale, 1962). Modal logic and its derivatives (such as tense logic, which deals with propositions whose truth values may change in time) are important in AI (Sowa, 1984).

Another contribution of Aristotle was the organization of knowledge into formal deductive structures, in which all the facts of a science were either stated as axioms or formally derivable from the axioms. The best-known example is Euclidean geometry, which was the exemplar of a systematic body of knowledge for over two millennia (Maziarz & Greenwood, 1968). Similar formal axiomatic structures are used in AI for representing a knowledge domain.

The investigation of logic continued over the following centuries. For example, the medieval scholastics (roughly 6th to15th centuries) refined logic into a very precise instrument, although it was still based on a natural language (Latin) in contrast to modern symbolic logic. As a consequence, they became conscious of the limitations of natural language for exact knowledge representation and strove to compensate for its deficiencies. For example, they knew that the word *dogs* is used differently in the propositions "dogs are mammals" and "*dogs* is a plural noun." AI knowledge representation languages have to deal with similar issues (Sowa, 1984). In the end, dissatisfaction with natural languages led to an interest in developing artificial languages that were intended to be more rational (logical and precise). Behind this was the assumption that there is a universal grammar underlying all natural languages, and that it corresponds to the "language of thought"; therefore an artificial language, as an ideal vehicle for thought, ought to reflect this deep structure. Similar motivations underlie the development of AI knowledge representation languages (see below).

Combinatorial Methods

The Middle Ages also saw the development of combinatorial approaches to solving problems (Bocheński, 1970). For example, the medieval scholastics used a combinatorial procedure to generate the 192 possible Aristotelian syllogisms, and then they crossed out the invalid ones. This is an example of a generate-and-test procedure, an approach still widely used in AI. The problem with generate-and-test procedures is combinatorial explosion: The number of combinations to be tested increases exponentially with their size.

These combinatorial procedures acquired an increased significance, which contributed to the eventual development of AI, from the kabbalah, a Jewish mystical tradition with Pythagorean affinities, which became popular in the Middle Ages (Eco, 1997; Scholem, 1960). According to this tradition, the text of the Torah reflects the logos (rational structure) of the universe. Therefore, since the Torah is written in the letters of the Hebrew alphabet, these letters correspond to the elementary categories and archetypal forms underlying the universe. As a consequence, the letters of the Hebrew words for things reveal their logical structure to one who knows how to interpret them. Combinatory processes figure prominently in kabbalah, and significant words, especially the names of God, were permuted in order to reveal hidden wisdom and discover new truths. For this purpose the kabbalists used rotating wheels and other devices to ensure that they did not omit any combinations of letters, an example of a mechanized generate-and-test procedure.

Similar in spirit to the kabbalah, and perhaps in part inspired by it, was the Great Art (Ars Magna) of Raymond Lull (also spelled Llull, 1232-1315; Bonner, 1985; Johnston, 1987; Yates, 1966). He intended it to be a "universal science of all sciences," a systematic method by which knowledge could be discovered and proved. There were several versions of his system, but the most common one made use of nine "divine dignities," or attributes of God, which took different forms in each domain of knowledge but provided the fundamental categories in each domain. These abstract qualities (Goodness, Magnitude, Duration, etc.) correspond closely to certain kabbalistic names of God. In Lull's Art, as in kabbalah, we see an attempt to isolate the most basic categories that constitute all knowledge and to discover, therefore, an alphabet of thought. This remains an important goal in contemporary symbolic AI.

A distinctive characteristic of Lull's Art was the extensive use of rotating wheels to generate combinations of these elementary categories in order to discover and to demonstrate philosophical truths. Thus, the Great Art combines an alphabet of elementary concepts with mechanical procedures for generating their combinations in order to produce an automated method of knowledge discovery and proof. Such, at least, was its goal. In fact, it did not work, and for the most part it could be used only for proving the theological propositions that the operator already believed. Nevertheless, it inspired many later thinkers to attempt to correct its deficiencies and to construct machines for knowledge discovery and inference, but first it had to be recast into a more logical form.

Knowledge Representation and Mechanized Inference

As knowledge and inference became more systematized, the idea developed that reasoning, when carefully and methodically executed, was a kind of calculation. One clear exponent of this view was Thomas Hobbes (1588-1679), who said, "By *ratiocination* I mean *computation*" (*Elem. Phil.*, 1.1.1.2). He explained, however, that the addition and subtraction of concepts was not the same as the addition and subtraction of numbers.

Hobbes also distinguished reasoning from causes to their effects (forward chaining in modern AI terminology) and reasoning from effects to their causes (backward chaining). In both cases, thought is a kind of mental discourse, which corresponds to a defining assumption of symbolic AI: that there is a language of thought (sometimes called "Mentalese"). Words, whether external or in the mind, are tokens, manipulated according to mechanical rules, and correct reasoning is analogized to balancing account books. That is, thought is calculation. Furthermore, since Hobbes was a complete materialist, he understood thought as a kind of matter in motion, a strictly mechanical process.

Over the centuries there have been many attempts to design ideal languages, that is, artificial languages without the perceived deficiencies of natural languages (Eco, 1997; Large, 1985). As modern science emerged in the 17th century, the goal was often to develop a philosophical language, that is, a language suitable for philosophical analysis and scientific discourse. One of the most famous of these projects was the Real Character of John Wilkins (1614-1672), the first president of the Royal Society (Large; Lewis, 2007; Rossi, 2000; Vickers, 1987). He began by isolating a universal grammar that he believed to underlie the particular grammars of all natural languages, and so it is in effect the grammar of Mentalese and hence reflects the laws of thought. Inspired by Chinese writing, Wilkins also concluded that the forms of words should reflect their logical analysis (based on a class hierarchy), and he designed a vocabulary and symbolic writing system based on a comprehensive conceptual taxonomy. His language had little direct impact beyond inspiring the conceptual taxonomy used by *Roget's Thesaurus*, but symbolic logic and AI knowledge representation languages have similar goals and approaches (and, arguably, similar failings).

Gottfried Wilhelm von Leibniz (1646-1716) investigated knowledge representation and mechanized reasoning, in which he was influenced by kabbalah, Lull, Wilkins' language, Hobbes, and Chinese writing and philosophy (Buchanan, 2005; Coudert, 1995; Kneale & Kneale, 1962; Perkins, 2004; Styazhkin, 1969). For example, although he had already invented the binary number system, he later found it in the Chinese *I Ching* (*Book of Changes*) and saw how it reduced all change in the universe to two opposites (yin and yang). This accorded with the kabbalistic and Lullian idea that the world was organized in terms of an alphabet of fundamental ideas and with his own rationalistic philosophy, which sought the true essences of concepts in a small number of atomic (indivisible) categories.

Leibniz was very impressed by Lull's Great Art and by Wilkins' Real Character, but concluded that they would not work, and so he constructed a number of knowledge representation schemes on a more logical plan. For example, any positive integer can be decomposed into a unique product of prime numbers, which is analogous to the rationalist idea that any concept can be reduced to a unique conjunction of atomic concepts. Therefore, if a prime number were assigned to each atomic concept, then every possible concept would have a unique numerical value. Conversely, if we looked up in a philosophical dictionary the number corresponding to any concept, we could discover its essence, or true definition, by reducing the number into its prime factors.

There are two ways that classes are treated in mathematics and logic: extensionally and intensionally. The extensional approach is to define a class in terms of its members, its extension, whereas the intensional approach defines it in terms of its intension, or essential attributes. Although modern logic and mathematics tend to treat classes extensionally, AI treats them intensionally (i.e., a concept is represented by a property list) for the simple reason that most concepts have small intensions but infinite extensions, so it is easier to compute with intensions. For the same reasons, Leibniz settled on an intensional representation.

Leibniz agreed with Hobbes' assertion that thought is computation, and worked on a calculus for logical inference. For example, he discovered that propositions of the form "all *S* are *P*" can be decided computationally if we know the numbers corresponding to *S* and *P*. For if all *S* are *P*, then the essential attributes of *P* are among the essential attributes of *S*; numerically, the prime factors of *P* are among the prime factors of *S*. Therefore, to decide if a proposition "all *S* are *P*" is true, all we need to do is to look up the numbers for *P* and *S* and see if the number for *P* evenly divides the number for *S*.

In summary, we can see that Leibniz had all the components of a system of knowledge representation and mechanical inference. In principle, all concepts could be analyzed into a relatively small number of elementary atomic concepts, and each concept could be assigned a unique number on the basis of this analysis. All philosophical questions, then, could be answered rationally and logically by calculation, literally by ratios (*rationes*, *logoi*). Indeed, Leibniz constructed one of the earliest digital calculating machines (1671), the first capable of multiplication and division, and so he had in principle (but not capacity) the means for actual mechanized reasoning.

George Boole (1815-1864) is well known to computer scientists and information technologists as the inventor of Boolean algebra, which is applied to digital circuit design and in many other ways to computer technology. However, his goals were much more ambitious, for in his *Investigation of the Laws of Thought*, he says he intends "to investigate the fundamental laws of those operations of the mind by which reasoning is performed" and to express them in a calculus (Boole, 1854, p. 1). In common with contemporary logicians, he expressed logical operations in an algebraic notation as opposed to a natural language, thus contributing to the development of symbolic logic. He developed an extensional class logic, in which operations on classes correspond to operations on their extensions, that is, on the sets of their members, and so he invented the algebra of sets. However, he also showed how the same algebraic operations could be interpreted as a propositional logic, which laid the foundation for Boolean circuit design, later developed by Claude Shannon (1938), the inventor of information theory. Boole stressed the formality of his logic, that is, that its rules of inference depended only on the algebraic properties of the operators (commutativity, associativity, etc.) and not on any interpretation of the terms. Therefore, these operation were not restricted to human thought, but could be implemented by machines, which was accomplished about a decade later by W. S. Jevons.

William Stanley Jevons (1835-1882) was a prolific mathematician, scientist, and philosopher who contributed to statistics, economics, meteorology, and the philosophy of science (Mays & Henry, 1953). However, he was also the first to construct fully functional logic machines capable of automated reasoning (Jevons, 1870, 1894, 1958), and thus predecessors of AI technology. His system is based, first, on the idea of a logical alphabet that lists all the possible conjunctions of a given set of terms and their complements. Second, he uses an indirect method of deduction, which is simply to eliminate from the logical alphabet those combinations that are inconsistent with the premises. Obviously, this is a generate-and-test procedure: List all the possible combinations and remove the impossible conclusions; the result is the broadest conclusion compatible with the premises.

Jevons' indirect method, like most generate-and-test procedures, is too tedious and error prone to perform manually, so he invented a succession of devices that increasingly automated the process. His most sophisticated was a completely mechanical device, called the logical piano. It had a keyboard marked with the terms (*A*, *B*, *C*, *D*, and their complements) and with various logical symbols, which was used for entering a series of logical equations representing the premises of a deduction. Above the keyboard was a (mechanical) display, a kind of spreadsheet, which represented all of the logical combinations consistent with the premises that had been entered so far. Thus, the operator could watch the developing mathematical analysis, and even try out hypothetical premises to see how they might affect the conclusion. In 1869, Jevons constructed and demonstrated a four-term machine and planned the development of a 10-term reasoning engine, which would have required an entire wall to display the 1,024 combinations of its logical alphabet. Although the machine performs relatively simple operations on bit strings, Jevons (1958, pp. 110-111) enthused that "after the Finis key has been used the machine represents a mind endowed with powers of thought," and that as each proposition is entered, "the machine analyses and digests the meaning of it and becomes charged with the knowledge embodied in that proposition." Thus, Jevons invented AI hype!

FUTURE TRENDS

In the light of this history, symbolic AI, which has dominated AI research, can be seen as the continuation of a centuries-old tradition concerning the nature of knowledge and inference. This, of course, does not imply that it is the best approach to AI, or conversely that it is not. Although some prominent researchers have declared that the symbolic approach to cognition is "the only game in town," there are alternatives, most notably connectionism (or parallel distributed processing), which is based on simplified models of neural networks in the brain (Garson, 2007; Rumelhart, McClelland, & PDP Research Group, 1986). This new approach promises to compensate for many of the limitations of the symbolic approach, and also to shed light on cognitive processes in the brains of humans and other animals. Connectionism, however, is beyond the scope of this article.

This article has focused on a few of the principal thinkers who contributed to AI before the era of the computer, but there were many others. Therefore, much work remains to be done in exploring and explaining the intellectual background of artificial intelligence. Through a deeper understanding of the assumptions we bring to knowledge and cognition, we may see new approaches to AI technology.

CONCLUSION

We may draw several conclusions from this historical survey. First, symbolic AI is built upon a foundation of philosophical and psychological premises that have been part of Western intellectual history since ancient Greece. Since these assumptions are so deeply embedded in our intellectual background, they easily may be taken for granted and escape adequate scrutiny. Nevertheless, alternatives, such as connectionism, are being explored. Second, although earlier philosophers discussed the idea that thought is a kind of computation, it was only with the advent of modern computers that there was sufficient computing power to test these theories empirically. As a consequence, experimental AI research in the late 20th century revealed both the capabilities and limitations of symbolic AI and motivated the search for alternatives. Finally, perhaps the most important conclusion is that AI is not an isolated technological discipline, nor simply the applied side of cognitive science, but it is intimately related to intellectual issues about the mind and thought that have occupied civilization for millennia.

REFERENCES

Barnes, J. (1982). *Aristotle.* Oxford, United Kingdom: Oxford University Press.

Bocheński, I. M. (1970). *A history of formal logic.* New York: Chelsea Publishing.

Bonner, A. (Ed. & Trans.). (1985). *Selected works of Ramon Llull (1232–1316)* (Vol. 1). Princeton, NJ: Princeton University Press.

Boole, G. (1854). *An investigation of the laws of thought.* New York: Macmillan.

Buchanan, B. G. (2005). A (very) brief history of artificial intelligence. *AI Magazine*, pp. 53-60.

Burkert, W. (1972). *Lore and science in ancient Pythagoreanism.* Cambridge, MA: Harvard University Press.

Burnet, J. (1930). *Greek philosophy part I: Thales to Plato.* London: Macmillan.

Coudert, A. P. (1995). *Leibniz and the kabbalah*. Dordrecht, The Netherlands: Kluwer.

Eco, U. (1997). *The search for the perfect language* (J. Fentress, Trans.). Cambridge, MA: Blackwell.

Gardner, H. (1985). *The mind's new science: A history of the cognitive revolution.* New York: Basic Books.

Garson, J. (2007). Connectionism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy.* Retrieved from http://plato.stanford.edu/archives/spr2007/entries/connectionism/

Jevons, W. S. (1870). On the mechanical performance of logical inference. *Philosophical Transactions, 160*, 497.

Jevons, W. S. (1894). *Elementary lessons in logic: Deductive and inductive.* New York: Macmillan.

Jevons, W. S. (1958). *The principles of science: A treatise on logic and scientific method* (2nd ed.). New York: Dover.

Johnston, M. D. (1987). *The spiritual logic of Ramon Llull.* Oxford, United Kingdom: Clarendon Press.

Kneale, W., & Kneale, M. (1962). *The development of logic.* Oxford, United Kingdom: Oxford University Press.

Large, A. (1985). *The artificial language movement.* New York: Basil Blackwell.

Lewis, R. (2007). *Language, mind and nature: Artificial languages in England from Bacon to Locke.* New York: Cambridge University Press.

Mays, M., & Henry, D. P. (1953). Jevons and logic. *Mind, 62*, 484.

Maziarz, E. A., & Greenwood, T. (1968). *Greek mathematical philosophy.* New York: Frederick Ungar.

McCorduck, P. (2004). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence.* New York: AK Peters.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery, 19*, 113-126.

Perkins, F. (2004). *Leibniz and China: A commerce of light.* New York: Cambridge University Press.

Riedweg, C. (2005). *Pythagoras: His life, teaching, and influence.* New York: Cornell University Press.

Rossi, P. (2000). *Logic and the art of memory: The quest for a universal language* (S. Clucas, Trans.). Chicago: University of Chicago Press.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition.* Cambridge, MA: MIT Press.

Scholem, G. (1960). *On the kabbalah and its symbolism.* New York: Schocken.

Shannon, C. E. (1938). A symbolic analysis of relay and switching circuits. *Transactions of the AIEE, 57*, 713-723.

Sowa, J. F. (1984). *Conceptual structures: Information processing in mind and machine.* Reading, MA: Addison-Wesley.

Styazhkin, N. I. (1969). *History of mathematical logic from Leibniz to Peano.* Cambridge, MA: MIT Press.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433-460.

Vickers, B. (1987). *English science: Bacon to Newton.* Cambridge, United Kingdom: Cambridge University Press.

Yates, F. A. (1966). *The art of memory.* Chicago: University of Chicago Press.

KEY TERMS

Calculus: A calculus is a system of physical symbols and mechanical rules for their manipulation intended to accomplish some purpose, such as calculation, differentiation, integration, or formal inference. In principle, any process that can be accomplished by a calculus can be programmed on a digital computer.

Generate-and-Test Procedure: Is a common method of search, used in AI and other applications, in which possible solutions are generated systematically and evaluated until a suitable solution is found. For example, a game-playing program might generate possible moves, which are evaluated in terms of their likelihood of leading to a win. The greatest weakness of generate-and-test procedures is combinatorial explosion, which refers to the exponential increase of the number of possible solutions of increasing complexity (e.g., the number of moves that a game-playing program looks forward).

Knowledge Representation Language: Is a formal language, implementable in the data structures of a digital computer, intended to be capable of representing all knowledge or at least all knowledge in some AI application domain. It is intended as a medium for storing knowledge and for mechanized inference in its domain. A knowledge representation language is the analogue in AI of the language of thought in cognitive science.

Language of Thought ("Mentalese"): Is a hypothesized language-like system in whose terms all human cognition is supposed to take place. Advocates of this hypothesis acknowledge that not all of our thinking is discursive (by means of an inner dialogue), but they argue that the systematic structure of ideas and thinking implies that there must be a language of thought, albeit below the level of conscious access. The language-of-thought hypothesis partly justifies symbolic AI as a sufficient basis for AI.

Semantics: Refers to the meanings of expressions in a natural or artificial language and to the study of these meanings and their relation to the expressions. It is often contrasted with syntax. Since formal systems, calculi, and symbolic AI systems deal only with the forms of expressions, they can be sensitive to semantics only to the extent that the semantics is encoded in the system's syntax.

Symbolic AI: Is an approach to AI based on the manipulation of knowledge represented in language-like (symbolic) structures in which all relevant semantics (meaning) is explicit in the syntax (formal structure). The language-of-thought hypothesis provides part of the justification of the sufficiency of the symbolic approach to AI.

Syntax: Refers primarily to the grammar rules of a language (natural or artificial), that is, to the allowable forms of expressions without reference to their meaning (semantics). In the context of AI, syntax refers to the rules of knowledge representation in terms of data structures and to the computational processes that operate on these structures.

History of Simulation

Evon M. O. Abu-Taieh
The Arab Academy for Banking and Financial Sciences, Jordan

Asim Abdel Rahman El Sheikh
The Arab Academy for Banking and Financial Sciences, Jordan

Jeihan M. O. Abu-Tayeh
Ministry of Planning, Jordan

Hussam Al Abdallat
The Arab Academy for Banking and Financial Sciences, Jordan

INTRODUCTION

The great philosopher Aristotle said, "If you would understand anything, observe its beginning and its development." Therefore, understanding simulation requires observing its history.

Accordingly, simulation can be understood in many ways: "Simulation is the use of a model to represent over time essential characteristics of a system under study" (El Sheikh, 1987). Another definition is "Simulation is the imitation of the operation of a real-world process or system over time" (Banks, 1999).

Simulation was known long before computers. According to Araten et al. (1992), "The first econometrics model of the United States economy was constructed by J. Tinbergen in 1939." Later, as computers developed in the late 1950s and early 1960s, a spawn of computer simulation methodologies and approaches came to life. Computer simulation, like any industry, both affected and was affected by the development of different programming languages and computer capabilities and advances.

This article will first give a background about simulation in general, then it will discuss the classical simulation methodologies. We will address the current trends in simulation by presenting currently used Java-based simulation languages. In this regard, the classical simulation methodologies discussed in this article include the three-phase approach, activity scan, process interaction, event scheduling, transaction flow approach, Petri nets, and Monte Carlo. The languages discussed are simjava, DEVSJAVA, JSIM, JavaSim (J-Sim), JavaGPSS, Silk, WSE (Web-enabled simulation environment), SLX, and SRML (simulation reference markup language). As such, this article will tackle the history of the approaches and methodologies while shedding light on the genealogy of the simulation languages.

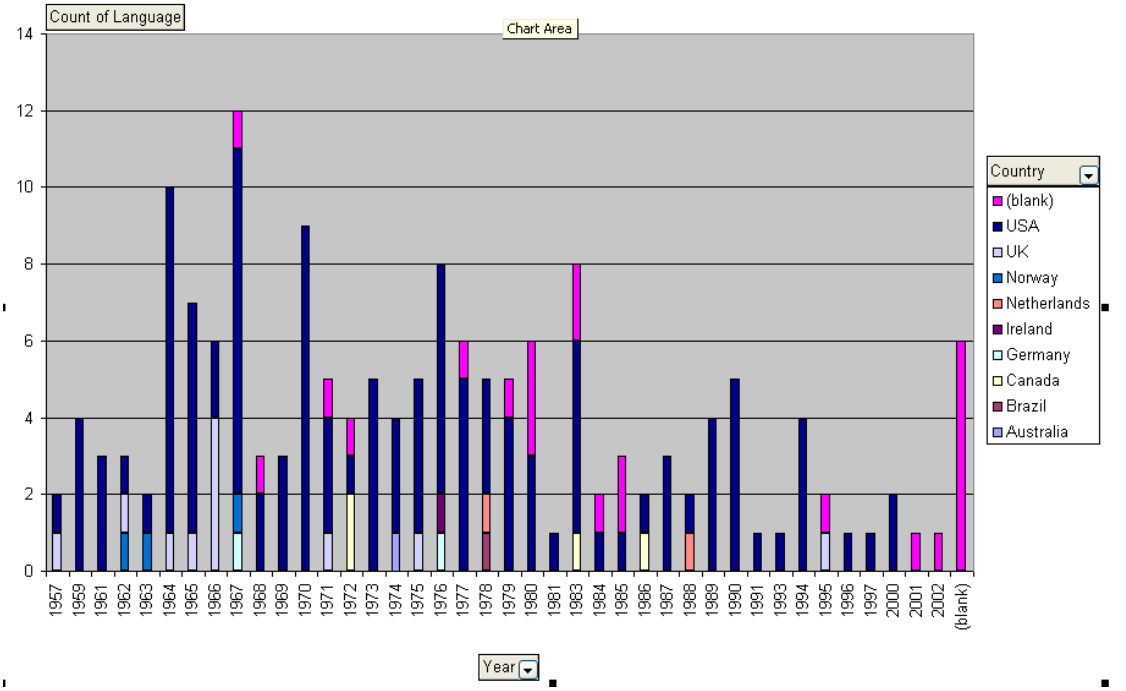
BACKGROUND

Defining simulation in its broadest aspect, we can say that it embodies a certain model to represent the behavior of a system, whether that may be an economic or an engineering one, with which conducting experiments is attainable. When studying models currently used, such a technique enables management and allows us to take appropriate measures and make fitting decisions that would further complement today's growth sustainability efforts, apart from cost decreases as well as service delivery assurance. As such, the computer simulation technique contributed to cost decline, depicted cause and effect, pinpointed task-oriented needs and service delivery assurance, explored possible alternatives, identified problems, proposed streamlined measurable and deliverable solutions, provided the platform for change-strategy introduction, introduced potential prudent investment opportunities, and finally provided a safety net for conducting training courses. Yet, simulation development is hindered due to many reasons. Like a rose, the computer simulation technique does not exist without thorns. Simulation reflects real-life problems; hence, it addresses numerous scenarios with a handful of variables. Not only is it costly and liable to human judgment, but also, the results are complicated and can be misinterpreted.

Within this context, there are four characteristics that distinguish simulation from any other software-intensive work according to Page and Nance (1997). First, simulation uses time as an index variable. Second, one of the objectives in simulation is to achieve correctness. Third, simulation software is involves computational intensiveness. Last but not least, the use of simulation is not typical; in fact, "no typical use for simulation can be described" (Page & Nance, 1997, p. 91).

Hence, there are many methodologies and approaches that simulation practitioners use when working on a simula-

Figure 1. Simulation languages in chronological order



tion project: the three-phase approach, activity scan, process interaction, event scheduling, transaction-flow approach, Petri nets, and Monte Carlo (Abu-Taieh & El Sheikh, 2007). Based on the previously mentioned methodologies and approaches, as well as other programming languages, more than 170 simulation programming languages and more than 60 simulation packages (Abu-Taieh & El Sheikh) were developed, for example, GPSS, GSP, GASP, SIMULA, and so forth. The simulation programming languages are reflected in *Figure **1*** in terms of the date and country of origin for each programming language.

CLASSICAL SIMULATION APPROACHES AND METHODOLOGIES

There are many simulation approaches and methodologies. For this purpose, the most familiar will be thoroughly discussed, namely, the three-phase approach, activity scan, process interaction, event scheduling, transaction flow approach, Petri nets, and Monte Carlo.

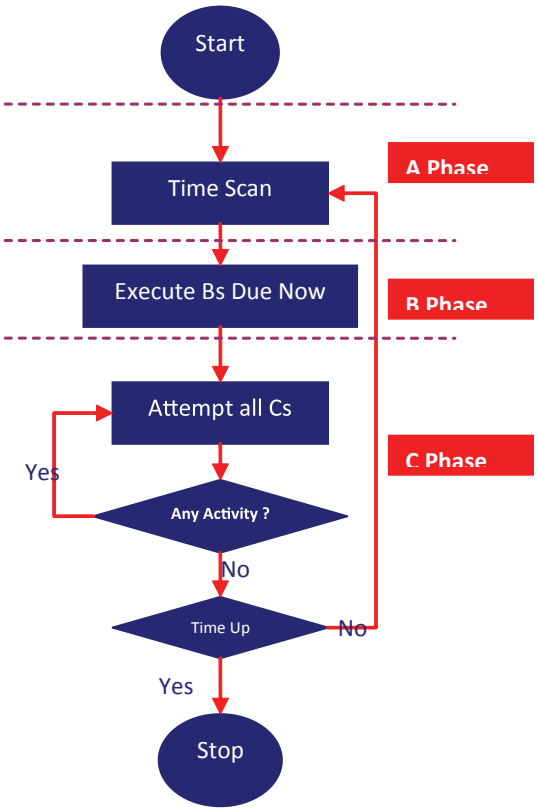
Three-Phase Approach

The first simulation modeling structure is known as the three-phase method. It was described by Keith Douglas Tocker in 1963 in his book *The Art of Simulation* (Odhabi, Paul, & Macredie, 1998), and then discussed in detail by Pidd and Cassel (1998).

Tocker introduced the three-phase approach through the General Simulation Program (GSP), which is considered as the first language effort (Nance, 1995; Pidd, 1998). In 1966, Tocker introduced wheel charts, which were later replaced by the activity cycle diagram and the language CSL (control and simulation language), which represented a simpler approach called activity scan (Nance). Tocker claimed that the structure would enable automatic programming simulation as Nance (1995) states was obvious in the development of OPS-1, OPS-2, OPS-3, and OPS-4, which are simulation languages intended for non-computer-programmers developed by students in MIT.

The three-phase approach has, as the name suggests, three phases—the A phase, B phase, and C phase—as seen in Figure 2. Each phase will be further discussed next.

Figure 2. A three-phase executive (Pidd, 1998)



In the A phase, time is advanced until there is a state change in the system or until something happens next. The system is examined to determine all of the events that will take place at this time (that is, all the activity completions that occur). The A phase is defined formally by Pidd and Cassel (1998): "the executive finds the next event, and advances the clock to the time in which this event is due."

The B phase is the release of those resources scheduled to end their activities at this time. According to Pidd and Cassel (1998), the B phase "executes all B activities (the direct consequence of the scheduled events) which are due at the time."

The C phase starts the activities given a global picture of resource availability. The C phase is defined formally by Pidd and Cassel (1998): "the executive tries to execute all of the C activities (any actions whose start depends on resources and entities whose states may have changed in the B phase)."

In this regard, Michael Pidd (1998) criticized the three-phase model and claimed that this methodology will be less popular as computer power continues to grow, indicating that the reason is "because discrete time slices must be specified. If the time interval is too wide, detail is lost." It is worthwhile to note that Pidd attributed the same criticism to activity scanning, which will be discussed next.

Activity-Scanning Approach

In 1963 the activity-scanning approach was introduced by John Buxton and John Laski (Nance, 1993, 1995). Activity scanning is also known as the two-phase approach. Activity scanning produces a simulation program composed of independent modules waiting to be executed. In the first phase, a fixed amount of time is advanced or scanned. In Phase 2, the system is updated (if an event occurs) as seen in Figure 3. Activity scanning is similar to rule-based programming (if the specified condition is met, then a rule is executed).

Many simulation packages adopted activity scanning, one of which is FirstSTEP.

The activity-scanning approach was represented by CSL, which in turn was influenced by the FORTAN programming language. Later in 1966, Clementson published the research paper *Extended Control and Simulation Language*, which introduced the ECSL simulation programming language. ECSL is based on C.S.L., yet one must note here that CSL is not C.S.L.

In 1978 Parkin and Coats published a research paper titled *EDSIM: Event Based Discrete Simulation using General Purpose Languages such as FORTRAN*. According to Nance (1993, 1995), EDSIM built on ECSL, but reintroduced the FORTRAN-like component.

Figure 3. An activity-scanning executive (Pidd, 1998)

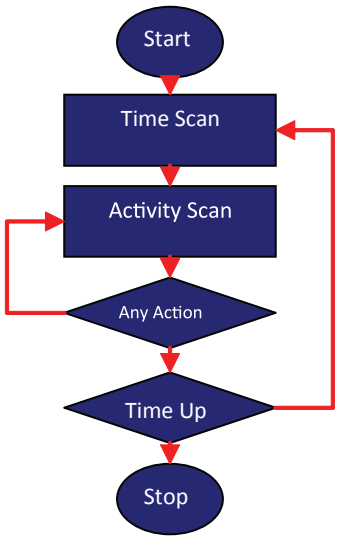
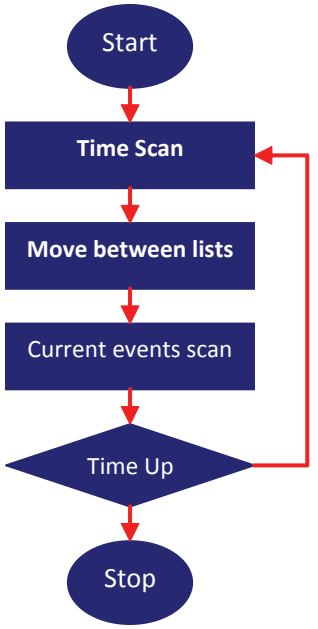


Figure 4. A process-interaction executive (Pidd, 1998)



Process-Interaction Approach

During the early 1960s, the process-interaction approach was introduced in the form of the simulation-oriented language (SOL) and SIMULA. SOL, created by Knuth and McNeley, was an extension of ALGOL, according to Nance (1993, 1995). However, SIMULA was created between 1962 and 1965, and later SIMULA was developed in 1967 by Ole-Johan Dahl and Kristen Nygaard. SIMULA produced CØNSIM (conflict simulator) and DEMOS in the late 1970s (Clema & Kirkham, 1971).

The simulation structure that has the greatest intuitive appeal is the process-interaction method. In this method, the computer program emulates the flow of an object (for example, a load) through the system. The load moves as far as possible in the system until it is delayed and either enters an activity or exits from the system. When the load's movement is halted, the clock advances to the time of the next movement of any load.

This flow, or movement, describes in sequence all of the states that the object can attain in the system as seen in Figure 4. The process-interaction approach was used by many commercial packages, among them, Automod.

Transaction-Flow Approach

The transaction-flow approach was first introduced by GPSS in 1962 as stated by Henriksen (1997). Transaction flow is a simpler version of the process-interaction approach as the following clearly states: "Gordon's transaction flow world-view was a cleverly disguised form of process interaction that put the process interaction approach within the grasp of ordinary users" (Schriber, Ståhl, Banks, Law, Seila, & Born, 2003).

The transaction-flow models consist of entities (units of traffic), resources (elements that service entities), and control elements (elements that determine the states of the entities and resources) as described by Henriksen (1997), Schriber and Brunner (1997), and *GoldSim Web* (n.d.). Discrete simulators that are generally designed for simulating detailed processes such as used in call centers, factory operations, and shipping facilities rely on such an approach, such as in "ProModel, Arena, Extend, and Witness" (GoldSim Web, n.d.). Some scholars (Schriber & Brunner, 1997) it as a "transaction-flow world view" and they add that it "often provides the basis for discrete-event simulation."

In view of the aforementioned, the same scholars continue to describe the best fitted applications to such an approach: "manufacturing, material handling, transportation, health care, civil, natural resource, communication, defense, and information processing systems, and queuing systems in general" (Schriber & Brunner, 1997).

Stock and Flow Approach

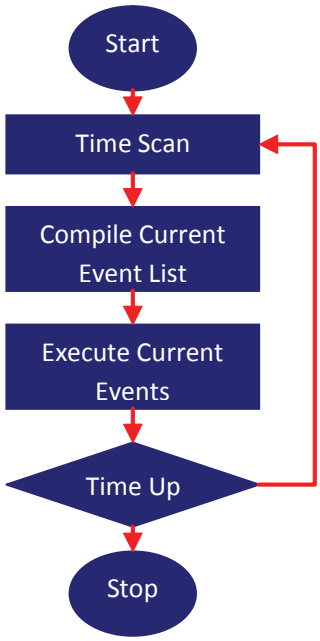
Another approach that is worth mentioning is the stock and flow approach, which is used in dynamic systems. This approach was created at MIT in the '60s by J. W. Forrester (GoldSim Web, n.d.). Stock and flow is based on system dynamics that are built using three principal element types: stocks, flows, and converters.

Event-Scheduling Approach

The event-scheduling approach came into existence through SIMCRIPT in the mid 1960s. SIMSCRIPT spawned from the programming language FORTRAN. The major designer of SIMSCRIPT is Harry Markowitz (Nobel prize winner) and the sole programmer is Bernard Hausner. SIMSCRIPT was fathered by SPS-1 and GEMS, and resembles SEAL (simulation, evaluation, and analysis language). SPS-1 was developed by Jack Little of RAND and Richard W. Conway of Cornell University, and the sole programmer was Bernard Hausner. Many versions came from SIMSCRIPT, like CLP, QUICKSCRIPT, SIMSCRIPTII, SIMSCRIPTII plus, SIMSCRIPTII.5, ESC II, and CSP II, according to Nance (1993, 1995) and Araten et al. (1992).

The basic concept of the event-scheduling method is to advance time to the moment when something happens next (that is, when one event ends, time is advanced to the time of the next scheduled event). An event usually releases a resource. The event then reallocates available objects or

Figure 5. An event-scheduling executive (Pidd, 1998)



entities by scheduling activities, in which they can now participate. Time is advanced to the next scheduled event (usually the end of an activity) and activities are examined to see whether any can now start as a consequence, as seen in Figure 5. The event-scheduling approach has one advantage and one disadvantage as Schriber et al. (2003) states, "The advantage was that it required no specialized language or operating system support. Event-based simulations could be implemented in procedural languages of even modest capabilities." On the other hand, the disadvantage "of the event-based approach was that describing a system as a collection of events obscured any sense of process flow." As such, "in complex systems, the number of events grew to a point that following the behavior of an element flowing through the system became very difficult." Many simulation packages adopted the event-based approach, of which are Supply Chain Builder, Factory Explorer, GoldSim, and ShowFlow.

Petri Nets

Petri nets have been under development since the 1960s, when Carl Adam Petri defined the language in *Kommunikation mit Automaten* (History, 2004): "It was the first time a general theory for discrete parallel systems was formulated."

Furthermore, the idea of Petri nets was developed to answer the question of concurrency, which naturally always arises when discussing simulation. As such, Petri nets handle concurrent discrete events in dynamic systems simulation. As an example, some simulation packages like Optsim (Artifex) use Petri nets.

In order to understand Petri nets, a comprehensive definition must be initially realized. Following are two formal

Table 1. Classical simulation approaches

Approach	Creator	Year	Tool	Language
Three-Phase	Keith Douglas Tocker	1963	wheel charts	General Simulation Program (GSP)
Activity-Scanning Approach	John Buxton and John Laski	1963	activity cycle diagram	CSL
Process-Interaction Approach	Knuth and McNeley Ole-Johan Dahl & Kristen Nygaard	1963		SOL, SIMULA
Event Scheduling	Major Designer: Harry Markowitz (Nobel prize winner) Sole programmer: Bernard Hausner	1963		SIMSCRIPT based on SPS-1 and GEMS
Stock and Flow Approach	Professor Jay W. Forrester at MIT	1960s		GoldSim
Petri Nets	Carl Adam Petri	1960s	tokens	graphical
Monte Carlo Methods (statistical sampling)	Enrico Fermi	1930s	many flavors	

definitions of Petri nets: one that calls it a technique, while the other calls it a language. The first definition follows: "A formal, graphical, executable technique for the specification and analysis of concurrent, discrete-event dynamic systems; a technique undergoing standardization" (PetriNets, 2004).

The second definition states, "Petri Nets is a formal and graphical appealing language, which is appropriate for modeling systems with concurrency….The language is a generalization of automata theory such that the concept of concurrently occurring events can be expressed" (History, 2004).

Petri nets were classified in different ways. The following classifications were listed on *Class Web* (n.d.), among others: Petri net systems of Level 1, Petri net systems of Level 2, and Petri net systems of Level 3. The Level 1 Petri nets systems are characterized by Boolean tokens; on the other hand, Level 2 uses integer tokens. Level 3 is characterized by high-level tokens.

Monte Carlo Methods

At the onset, the Monte Carlo methods were used in the 1930s and denoted generic names (statistical sampling). Enrico Fermi used Monte Carlo to calculate the properties of the neutron. Consequently, during the 1950s, the Monte Carlo methods were used at Los Alamos for the development of the hydrogen bomb. Monte Carlo was popularized and pioneered by Stanislaw Marcin Ulam, Enrico Fermi, John von Neumann, and Nicholas Metropolis (http://en.wikipedia.org/wiki/Monte_Carlo_methods). Another alternative to Monte Carlo is applied information economics (AIE), which is a decision analysis method.

In this context, the Monte Carlo method is formally defined: "Numerical methods that are known as Monte Carlo methods can be loosely described as statistical simulation methods" (*CSEP Web*, 1995). Note that three words stand out in the definition—*loosely, random,* and *statistical*—while statistical simulation is also defined as a "method that utilizes sequences of random numbers to perform the simulation" (*CSEP Web*), as can also be seen within the following definition: "Monte Carlo was coined by Metropolis (inspired by Ulam's interest in poker) during the Manhattan Project of World War II, because of the similarity of statistical simulation to games of chance" (CSEP Web, 1995).

Furthermore, Monte Carlo was referenced in 1987 when Metropolis wrote, "Known to the old guards as statistical sampling: in it new, surroundings and owing to its nature there was no denying its new name of the Monte Carlo Method."

Monte Carlo methods, known for its diversity of applications, are used in nuclear reactor simulation, quantum chromo dynamics, radiation cancer therapy, traffic flow, stellar evolution, econometrics, Dow Jones forecasting, oil well exploration, and VSLI design.

THE JAVA INCLINATION

Since Java came to life in the 1990s, many people from the simulation arena thrived for its utilization. This did not come as a surprise given the history of the object-oriented trail of thought. Additionally, upon examining the process interaction methodology, it was bound to be the basis for today's object-oriented mind-set.

In this context, currently there are several Java-based discrete simulation environments (Kuljis & Paul, 2000): simjava, DEVSJAVA, JSIM, JavaSim (J-Sim), JavaGPSS, Silk, WSE, SLX, and SRML. Following is an elaborated overview.

The first in this list of definitions is simjava, which is based on the process-interaction approach with the purpose of being able to build complex systems. Obviously, simjava is based on the Java programming language, which is inherently objecting oriented (Kuljis & Paul, 2000; Page, Moose, & Gri, 1997).

The second on the list is the DEVSJAVA environment, which was built using Java and is based on the discrete event system specification (Kuljis & Paul, 2000): "A user of DEVSJAVA is able to experiment with any DEVS model from any machine at any time and to interactively and visually control simulation execution."

Likewise, the third on the list is JSIM, which is described by Kuljis and Paul (2000) as a "Java based simulation and animation environment supporting web based simulation as well component-based technology." The component-based technology that JSIM utilizes in this case is Java Beans, as Kuljis and Paul claim in their paper. The idea is to build up the environment from reusable software components that "can be dynamically assembled using visual development tools" (Kuljis & Paul).

In this regard, it is worth noting that JavaSim was later changed to the name J-Sim, which is the fourth definition on the list. J-Sim is considered to be "a set of Java packages for building discrete event process-based simulation" (Kuljis & Paul, 2000). JavaSim was renamed because the word Java is a trade mark owned by SUN Microsystems. J-Sim is an implementation of a simulation tool kit named C^{++}SIM developed in the University of Newcastle (Kuljis & Paul). The official Web site of J-Sim describes it as a "component-based, compositional simulation environment" (*J-Sim Web*, n.d.). Yet, the Web site adds "unlike the other component-based software packages/standards, components in J-Sim are autonomous" (*J-Sim Web*).

The fifth on the list is the JavaGPSS compiler: "The *JavaGPSS* compiler is a simulation tool which was designed for the Internet" (Kuljis & Paul, 2000). JavaGPSS was built so that GPSS can be run on the Internet (Kuljis & Paul).

Within this context, Silk, being the sixth on the list, is defined as the "general-purpose simulation language based around a *process-interaction* approach and implemented

in Java" (Kuljis & Paul, 2000). The purpose of silk is "to encourage better discrete-event simulation through better programming by better programmers" (Kilgore, 2003). According to Healy and Kilgore (1997), "The *Silk* language is an opportunity to make simulation more accessible without sacrificing power and flexibility."

Furthermore, the Web-enabled simulation environment is the seventh item on the list. WSE "combines web technology with the use of Java and CORBA" (Kuljis & Paul, 2000). Also, according to Kuljis and Paul, the WSE environment provides location and distribution transparency, and platform independence.

The eighth on the list, SLX is considered a simulation language that is "C-like," as cited by Henriksen (1997), who first introduced the framework of SLX in 1995 and went on to describe it as a "wolverine software" for the next generation. R. C. Crain discussed SLX in 1997 in the paper entitled "Simulation Using GPSS/H" at the winter simulation conference, and later Henriksen also covered it in his paper entitled "Introduction to SLX" in 1998.

Finally, the simulation reference markup language and the simulation reference simulator were both developed by Boeing and used in many projects (Reichenthal, 2002). Like HTML (hypertext markup language), SRML represents simulation models and as a Web browser represents universal client application. SRML binds the declarative with procedures and like HTML contains both declarative and procedural definitions. According to Reichenthal, "*SRML* is an XML-based language that provides generic simulation markup for adding behavior to arbitrary XML documents."

FUTURE TRENDS

Simulation is not a stand-alone science; it interacts with technology simultaneously as technology advances. There are two major things driving simulation today: the Internet and object-oriented thinking. While the major driving force for simulation is still the need for simulation, the need for virtual reality and augmented reality inter alia demands the prompt advancement of many aspects of simulation. Nevertheless, the elimination of the usual inhibitors like computer hardware speed is yet another prerequisite for advancement in this regard.

CONCLUSION

This article tried to shed light on the numerous facets of simulation history, spanning over the approaches and methodologies of discrete event simulation. The article discussed the most famous discrete event simulation methodologies (three-phase approach, activity scan, process interaction, event scheduling, transaction-flow approach, Petri nets, and Monte Carlo) represented in the programming languages that stemmed from the methodologies. Then the article discussed the future trends through the current history by discussing nine Java-based languages.

REFERENCES

Abu-Taieh, E., & El Sheikh, A. (2007). Commercial simulation packages: A comparative study. *International Journal of Simulation, 8*(2), 1473-804x.

Araten, M., Hixson, H. G., Hoggatt, A. C., Kiviat, P. J., Morris, M. F., Ockene, A., et al. (1992). The Winter Simulation Conference: Perspective of the founding fathers. In J. J. Swain, D. Goldsman, R. C. Crain, & J. R. Wilson (Eds.), *Proceedings of the 1992 Winter Simulation Conference.*

Banks, J. (1999, December 5-8). Introduction to simulation. In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 Winter Simulation Conference*, Phoenix, AZ (pp. 7-13). New York: ACM Press.

Class Web. (n.d.). Retrieved April 1, 2004, from http://www.cse.fau.edu/~maria/COURSES/CEN4010-SE/C10/10-7.html

Clema, J., & Kirkham, J. (1971). CONSIM (conflict simulator): Risk, cost and benefit in political simulations. In *Proceedings of the 1971 26th Annual Conference* (pp. 226-235). New York: ACM Press.

El Sheikh, A. (1987). *Simulation modeling using a relational database package.* Unpublished doctoral dissertation, The London School of Economics, London.

GoldSim Web. (n.d.). Retrieved September 1, 2003, from http://www.goldsim.com

Healy, K. J., & Kilgore, R. A. (1997, December 7-10). Silk: A Java-based process simulation language. In S. Andradóttir, K. J. Healy, D. Withers, & B. Nelson (Eds.), *Proceedings of the 1997 Winter Simulation Conference*, Atlanta, GA (pp. 475-482). New York: ACM Press.

Henriksen, J. (1997, December 7-10). An introduction to SLX™. In S. Andradóttir, K. J. Healy, D. H. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 Winter Simulation Conference*, Atlanta, GA (pp. 559-566).

History. (2004). Retrieved April 18, 2004, from http://www.daimi.au.dk/PetriNets/faq/

J-Sim Web. (n.d.). Retrieved April 10, 2004, from http://www.j-sim.org

Kilgore, R. (2003, December 7-10). Object-oriented simulation with SML and silk in .Net and Java. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 Winter Simulation Conference*, New Orleans, LA (pp. 218-224).

Kuljis, J., & Paul, R. (2000, December 10-13). A review of Web based simulation: Whither we wander? In J. A. Joines, R. R. Barton, K. Kang, & P. A. Fishwick (Eds.), *Proceedings of the 2000 Winter Simulation Conference*, Orlando, FL (pp. 1872-1881). San Diego, CA: Society for Computer Simulation International.

Metropolis, N. (1987). The beginning of Monte Carlo method. *Los Alamos Science, 15*, 125-130. Retrieved April 18, 2004, from http://jackman.stanford.edu/mcmc/metropolis1.pdf

Nance, R. (1993). A history of discrete event simulation programming languages. *ACM SIGPLAN Notices, 28*(3), 149-175.

Nance, R. (1995). Simulation programming languages: An abridged history. In C. Alexopoulos, K. Kang, W. R. Lilegdon, & D. Goldsman (Eds.), *Proceedings of the 1995 Winter Simulation Conference.*

Nance R. (1996). A history of discrete event simulation programming languages. In *History of programming languages* (Vol. 2). New York: ACM Press.

Odhabi, H., Paul, R., & Macredie, R. (1998, December 13-16). Making simulation more accessible in manufacturing systems through a "four phase" approach. In D. J. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the 1998 Winter Simulation Conference*, Washington, DC (pp. 1069-1075). Los Alamitos, CA: IEEE Computer Society Press.

Page, E., Moose, R., & Gri, S. (1997, December 7-10). Web-based simulation in Simjava using remote method invocation. In S. Andradóttir, K. J. Healy, D. H. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 Winter Simulation Conference*, Atlanta, GA (pp. 468-474). New York: ACM Press.

Page, H., & Nance, R. (1997). Parallel discrete event simulation: A modeling methodological perspective. *ACM Transactions on Modeling and Computer Simulation, 7*(3), 88-93.

PetriNets. (2004). Retrieved April 18, 2004, from http://www.petrinets.info/graphical.php

Pidd, M. (1998). *Computer simulation in management science* (4th ed.). Chichester, England: John Wiley & Sons.

Pidd, M., & Cassel, R. (1998, December 13-16). Three phase simulation in Java. In D. J. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the 1998 Winter Simulation Conference*, Washington, DC (pp. 267-371). Los Alamitos: IEEE Computer Society Press.

Reichenthal, S. (2002, December 8-11). Re-introducing Web-based simulation. In E. Yücesan, C.-H. Chen, J. L. Snowdon, & J. M. Charnes (Eds.), *Proceedings of the 2002 Winter Simulation Conference*, San Diego, CA (pp. 847-852).

Schriber, T., & Brunner, D. (1997, December 7-10). Inside discrete-event simulation software: How it works and why it matters. In S. Andradóttir, K. J. Healy, D. H. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 Winter Simulation Conference*, Atlanta, GA (pp. 14-22).

Schriber, T. J., Ståhl, I., Banks, J., Law, A. M., Seila, A. F., & Born, R. G. (2003, December 7-10). Simulation text-books: Old and new (panel). In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 Winter Simulation Conference*, New Orleans, LA (pp. 238-245).

KEY TERMS

Continuous Simulation Systems: These are systems that deal with time as a continuous function, for example, when simulating the flow of water from a reservoir.

Discrete Simulation Systems: These are simulation systems that treat the time variable as a discrete variable. Usually any systems that deal with queues (supermarkets, banks) are of discrete nature.

Monte Carlo Methods: These are statistical simulation methods.

Petri Nets: Petri nets are "a formal, graphical, executable technique for the specification and analysis of concurrent, discrete-event dynamic systems; a technique undergoing standardization" (PetriNets, 2004).

Simulation: It "is the imitation of the operation of a real-world process or system over time" (Banks, 1999).

Simulation Methodology: It is a world view of abstracting the real world and mapping it into a computer program.

How Teachers Use Instructional Design in Real Classrooms

Patricia L. Rogers
Bemidji StateUniversity, USA

INTRODUCTION

"I've learned how to use the [insert new instructional technology here], so now how do I use it in class?"

From filmstrips and mimeographs, to computer-based simulations and virtual reality, technology seems to dominate teachers' lives as they master the new instructional media for use in their classrooms. Good teaching and learning practices tend to take a back seat while the focus on mastery of the technology reduces teaching into basic presentations and lectures, a format most easily controlled by the instructor. While most pre-K-12 and post-secondary instructors do develop effective courses in which students learn, many would be hard pressed to describe *how* they arrive at certain goals and teaching strategies.

BACKGROUND

The field of instructional design provides sound practices and models that, once modified for use by working teachers, can be used to design effective instruction in any content area (Rogers, 2002). The more difficult issue is helping teachers move beyond the tendency to focus on technology rather than instructional goals. Such focus occurs at lower levels of what can be described as a technology adoption hierarchy (summarized in Table 1): familiarization, utilization, integration, reorganization, and evolution (Hooper & Rieber, 1999).

Somewhere at the integration stage, a "magic line" is crossed and the focus is no longer on the technology but on the teaching and learning. A supporting practical design model can help teacher-designers cross this magic line more efficiently and with a high degree of success.

FUTURE TRENDS

A Modified Instructional Design Model

Prescriptive behavioral models in learning would seem, at first encounter, to be inappropriate in light of the more constructivist practices of current educators. However, most constructivists would concur that one must have solid building blocks or elements before construction of new knowledge can be achieved. Dick and Carey's (1990) original systems design model and subsequent modifications by Gagné, Briggs and Wager (1992) and others offer examples of all of the elements necessary for designing and evaluating effective instruction. What the models lacked, however, was a connection to real classroom teachers: those of us who are really teacher-designers and who must create and develop our courses without benefit of design teams and lengthy pilot tests with target audiences.

Figure 1 is a modification based on several interpretations of the most typical instructional design model (Dick & Carey, 1990). Notice that the five phases of design: analyze, design, develop, implement, and evaluate, are focused not

Table 1. A summary of the technology adoption hierarchy

EVOLUTION	Highest level: is most able to cope with change and has skills to adapt newer technologies as needed or desired in teaching and learning environment.
REORGANIZATION	Re-designs teaching strategies with focus on learning and goals of instruction. Students become more involved in the learning environment.
INTEGRATION	Beginning to accept the technology. Focus soon shifts from learning the technology (and fearing its breakdown) to effective use of the technology in teaching.
UTILIZATION	Basic trial of the new technology. Focus is on finding a use for the technology that may or may not continue, particularly if the technology breaks down.
FAMILIARIZATION	Lowest level of exposure to a technology.

on designing teacher-proof curricula but rather on teacher-designers staying focused on their own environment and learners.

The model helps teachers begin designing with the constraints, issues, community demands, and state and federal mandates in mind before thinking about instructional media or "activities". Once parameters are identified, teacher-designers move into the design phase as they document the overall goals of their course (or, in the case of primary teachers, their school year) while simultaneously considering their learners. What does it mean to be a 3rd grade person? What skills should learners have as they move into 4th grade? What new knowledge is gained in 4th grade to allow learners to become 5th grade students? And so on.

Within this phase, assessments are also considered. Effective design, as well as effective teaching, requires teacher-designers to carefully match goals and objectives to appropriate assessments. Desired types of learning, from basic verbal information to higher order thinking skills (Gagné, Briggs & Wager, 1992) must have matched assessments that allow learners to demonstrate their new skills and abilities. Mismatched goals and assessments are common errors in designing instruction.

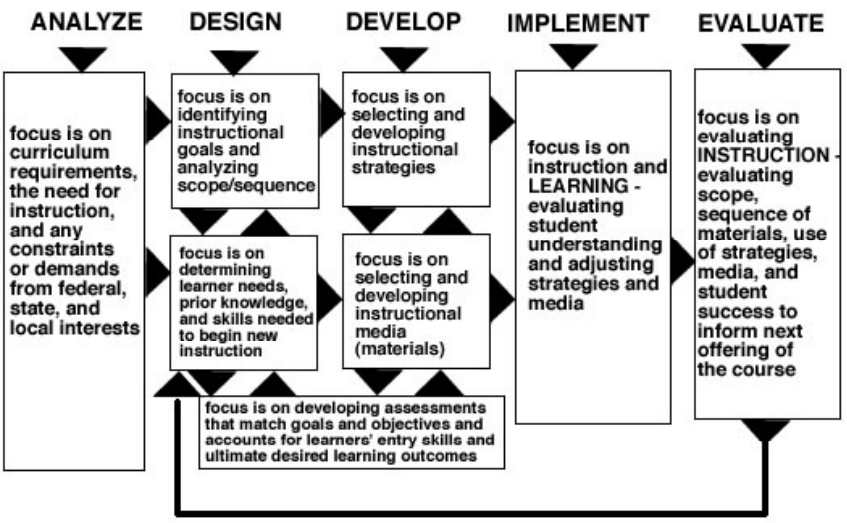
Using this model essentially forces us to wait until the development phase to select teaching strategies and instructional media. For those teachers who are struggling to leave the lower levels of the technology adoption hierarchy, this placement will seem uncomfortable. However, starting with the technology and trying to build an instructional environment is, as should be apparent, in essence turning the design process inside out! Once the focus is away from the goals and objectives and the *learners,* any further course development will likely result in a design that falls far short of the intended learning:

I am elated that I had the opportunity to work on curriculum design for the first time the right way and with a group of faculty members who supported my learning. I have watch[ed] part-time faculty members and even seasoned classroom teachers jump into material they are not familiar with, plan day by day, never really having clear objectives and methods of evaluation [in mind]. (A. Vidovic, personal communication, July 30, 2003)

Notice that the development of assessments also crosses this phase of the design. It is critical to select strategies and media that support the goals and objectives as well as allow students to demonstrate their understanding. Using strategies and media that are similar to the assessment situation strengthens the learning. For example, if students were learning to write poetry, a true-false test would be a very inadequate measure of their skills.

Implementation, *teaching*, is the phase of a teacher-designer's true test. It is here that this model is quite different from traditional instructional design models in that teacher-designers rarely have a chance to "try out" a course on a sample of students. Rather, they often have to simply try things and hope it all works well. However, by follow-

Figure 1. Modified instructional design model for teacher-designers. Modifications first introduced in Designing Instruction for Technology-Enhanced Learning*, Rogers, 2002, Idea Group Publishing. Further modifications by Patricia L. Rogers and Catherine E. McCartney, Bemidji State University, for the Online Graduate Program, 2002-2003).*



ing the model thus far, teacher-designers have an advantage over others who do not have clear goals and objectives in mind. During this phase, student achievement and perhaps student evaluations of the course should be examined as evidence that all elements of the design thus far actually form a cohesive course that meets the goals of the instruction. Teacher-designers should take notes on a daily basis regarding which strategies are working with learners, which activities supported new learning, and which instructional medium was appropriate for certain types of learning.

The evaluation phase in this model relies heavily on the evidence from the previous phase and includes a critical look at any notes from the teaching experience, comparison to a previous experience teaching the course, and so on:

In designing and developing this online class using the first couple assignments (objectives, goals, subgoals, etc.), I really feel like [my] course's material fits together much better than it has when I taught it in the past. Though this [instructional design] process took a fair amount of time, I know I would never tackle another class design without using this process first. It does seem to speed up the material/content piece considerably by doing this first. (N. Gregg, personal communication, July 28, 2003)

Barriers to Designing Effective Instruction in Distance Learning

By following a model that is based in practical, real-world experiences of teachers, teacher-designers are able to develop effective and well-documented instruction. However, we should note that there are many reasons good instructional design practices are not followed, and that most are out of the teacher-designer's control. Table 2 is a summary of some of the issues and barriers faced by teacher-designers.

CONCLUSION

A strong case can be made for working with teacher-designers at all levels of education on sound instructional design practices. "Winging it" when it comes to designing effective instruction is ill-advised in the rarified air of the 21st century knowledge and information age, with many demands from

Table 2. A summary of barriers to designing effective instruction

Fear of change
Changing teaching methods (strategies) to accommodate newer technologies, different modes of delivery, and the reality of managing a larger student market carries a certain amount of risk and challenge. The human tendency to want things to remain the same introduces a fear factor in designing and delivering instruction in the 21st century (Dublin, June 2003).

Unfamiliarity with newer technologies
The introduction of newer technologies in teaching usually results in teachers defaulting to presentations and lectures. Once the "magic line" is crossed, teaching and learning with technology refocuses from the technology to learning (Dublin, June 2003; Hooper & Rieber, 1999; Strauss, June 2003).

Correspondence, Lecture, and Interactive Learning
Real classrooms rely on interactions among students and the instructor. Some online courses are actually stand-alone correspondence courses that are self-paced and lack high interactivity levels. Lecture courses tend to be one-way communications while other strategies emphasize interactivity. There is a critical need to be clear about levels of interactivity in learning environments (Cavalier, June 2003).

Ill-defined goals and objectives
Defining goals and objectives is often a new experience for many faculty. Goals and objectives may not match teaching style or adequately address desired learner outcomes.

Unrealistic administrative, policy, or economic pressures
Some teachers have encountered serious constraints when designing instruction. A partial list includes: forced use of traditional "activities" that become the central focus of the instruction, district-wide adoption of specific texts or programs designed to be "teacher-proof" with little flexibility, limited development time for teachers, and a focus on state-wide test scores directly tied to school funding (Rogers, 2000).

Difficulty in translating from one environment to another, such as onground to online
Moving a course from onground delivery to the online environment sets up barriers for inexperienced teachers: some try to limit all transactions to real time and have a felt need to recreate their onground course exactly. Others err on the other side and resort to a type of glorified correspondence approach.

learners for high-quality educational experiences. Educational institutions, particularly colleges and universities, are faced with harsh competition for the teaching aspect of their institution from for-profit companies. Such companies outspend higher education in development, maintenance, and marketing of educational offerings, particularly in online learning (Rogers, 2001). Non-profit educational institutions can compete most effectively by providing (a) affordable pricing, (b) greater accessibility to education, and (c) high-quality, personalized educational experiences for their learners. A and B are usually easily attained. High quality education (c) begins with great teachers and support staff and is built and sustained with solid instructional design practices.

REFERENCES

Cavalier, R. (2003, June). Interactions in education: A conversation with Brenda Laurel. *Syllabus. http://www.syllabus.com/article.asp?id=7764*

Dick, W., & Carey, L. (1990). *The systematic design of instruction* (3rd ed.). Glenview, IL: Scott Foresman.

Dublin, L. (2003, June 24). If you only look under street lamps...Or nine e-learning myths. *E-learning Insider, 2003*(1). *http://www.elearningguild.com/pbuild/linkbuilder.cfm?selection=doc.421*

Gagné, R.M., Briggs, L.J., & Wager, W.W. (1992). *Principles of instructional design.* Orlando, FL: Harcourt, Brace, Jovanovich.

Hooper, S., & Rieber, L. (1999). Teaching, instruction, and technology. In A.C. Ornstein & L.S. Behar-Horenstein (Eds.), *Contemporary issues in curriculum* (2nd ed., pp. 252-264). Boston: Allyn and Bacon.

Rogers, P.L. (2000). Barriers to adopting emerging technologies in education. *Journal of Educational Computing Research, 22*(4), 455-472.

Rogers, P.L. (2001). Traditions to transformations: The forced evolution of higher education. *Educational Technology Review, 9*(1). *http://www.aace.org/pubs/etr/issue1/rogers.cfm*

Rogers, P.L. (Ed.). (2002). *Designing instruction for technology-enhanced learning.* Hershey, PA: Idea Group Publishing.

Seels, B.B., & Richey, R.C. (1994). *Instructional technology: The definitions and domains of the field.* Washington, DC: Association for Communications and Technology.

Strauss, H. (2003, June). My dog knows html—Should your faculty? *Syllabus. http://www.syllabus.com/article.asp?id=7774*

KEY TERMS

ADDIE: The five phases of most instructional design models: analyze, design, develop, implement, and evaluate. Some models follow the phases in a linear fashion, while others may approach the phases in a holistic or phenomenologic manner.

E-Learning: A term used to describe learning that takes place usually online, but includes all forms of electronically-enhanced and mediated learning. Computer-aided instruction, just-in-time learning, and intelligent systems can be included in the term "e-learning".

Instructional Design: The field of instructional design includes a range of professions from programmers and graphic artists, to the instructional designer. Designers are able to analyze instruction, learners, environments, strategies, and media to develop effective instruction of training. Designers may or may not be subject matter experts.

Instructional Design Models: Traditional design models are prescriptive step-by-step processes, usually associated with behaviorist instructional strategies. Phenomenological models incorporate constructivist philosophies and practices. In either aspect, design models guide the user in designing effective instruction that takes all aspects of design (see ADDIE) and reminds the user of critical elements and decisions in designing effective instruction.

Instructional (Educational) Technology: Instructional technology is the theory and practice of design, development, utilization, management and evaluation of processes and resources for learning (Seels & Richey, 1994).

Teacher-Designer: "…if you have any experience with instructional design you know that the field and the various models of design associated with it seem most appropriate for teams of people working on the course materials together. Once in a while, some of us are fortunate enough to have instructional designers, subject matter experts, graphic artists, programmers and so on available on our campus or in our school district to assist us with our technology-enhanced course. But most often, it the teacher alone who must rethink and redesign his or her course for technology-enhanced learning. And very often it is the teacher who must also prepare the materials for the Internet, interactive television, or some other delivery medium. They often do not have any background in instructional design theory or practices and have only just mastered the skills for using the delivery medium. These are the people I call 'teacher-designers'" (Rogers, 2002, p. 2).

Technology Adoption Hierarchy: "The model…has five steps or phases: familiarization, utilization, integration, reori-

entation, and evolution. The full potential of any educational technology can only be realized when educators progress through all five phases; otherwise, the technology will likely be misused or discarded…The *traditional* role of technology in education is necessarily limited to the first three phases, whereas contemporary views hold the promise to reach the evolution phase" (Hooper & Rieber, 1999, p. 253).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1344-1348, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Human-Centric E-Business

H.D. Richards
MAPS and Orion Logic Ltd, UK

Harris Charalampos Makatsoris
Brunel University, UK & Orion Logic Ltd, UK

Yoon Seok Chang
Korea Aerospace University School of Air Transport, Transportation and Logistics, Korea

INTRODUCTION

This article studies the transformation processes occurring in industry and business at large. It deals with the social and economic challenges, and explores the new concepts arising from an unprecedented technology revolution underpinned by advances and innovation in ICT. In addition it sets the scene for a new era of industrial capitalism.

Over the last decade of the twentieth century, a large number of companies faced the future with trepidation while others lacked a good strategy (Possl, 1991; Kidd, 1994; Ashkenas, 1997). Many changes had taken place including Just In Time (JIT) manufacturing and logistics, lean manufacturing (Womack, Jones, & Roos, 1990), shorter product lifecycles (Davenport, 1993), more intelligent approaches to IT (Drucker, 1992; MacIntosh, 1994; Nonaka, 1998), and costing (Wilson, 1995; Ansari, Bell, & the CAM-I Target Cost Core Group, 1997), but making money was becoming more and more difficult. It was a time and climate for dramatic new approaches (Warnecke, 1993; Drucker, 1994; Goldman, Nagel, & Preiss, 1995) with greater agility. New technologies were replacing old at a faster rate, and information technology provided better management and control vision, albeit on a limited local scale (Arguello, 1994; Leachman, Benson, Lui, & Raar, 1996; Makatsoris, Leach, & Richards, 1996). Also, push to pull manufacturing (Mertins, 1996) distinctly changed the approach to customers and service, which increased competitive and economic pressures resulted from the global reach of customers, manufacturers, and service providers keen to exploit the wealth of opportunities in both global markets and differences in worldwide regional markets (Bitran, Bassetti, & Romano. 2003). Even players only operating in local markets (Bologni, Gozzi, & Toschi, 1996; Zabel, Weber, & Steinlechner, 2000; Bonfatti & Monari, 2004) could not resist the tide of change. As a result many companies and economies (Hutton, 1995) were in a state of upheaval, and as a consequence some fell by the wayside. This was a climate in which there was an uncertain outcome, and it was into this melting pot that the Internet and the World Wide Web (WWW) were to produce an environment for a much-needed revolutionary change in the industrial approach. Later, broadband for landline and also wireless networking provided a much-needed speedier access.

Businesses looked to the wider horizons and the dynamics of their supply chains as well as their markets to discover new ways of working with both customers and suppliers, to grow and remain viable. The diverse industrial, commercial, and operational practices and processes needed to be remolded. And the collaborative aspects of external relationships to the advantage of company performance and the creation of new opportunities were the ones to be targeted. This resulted in increasing use of new forms of communication and innovation in multimedia technologies. In this unsettled environment, once fear of change had been forced into the background, chaos became the domain of creative experimentation (Weiland-Burston, 1992). It is during this period of confusion and anxiety that the process of metamorphosis started to take place.

A surge of new software tool ideas have helped, including Enterprise Resource Planning (ERP) Supply Chain Management (SCM) (Chang, McFarlane, & Shaw, 2001); Customer Relationship Management (CRM) (Greenberg, 2002); electronic commerce (e-commerce) and procurement (Chang, Makatsoris, & Richards, 2004); extensions in order management, fulfillment, and demand lifecycle control (Makatsoris, Chang, & Richards, 2004a, 2004b; Makatsoris & Chang, 2004); electronic business (e-business) (CEC, 2000); and new forms of conducting business, among many others. Further, mobile devices have enabled access to systems and software from any place in the world, and these technological improvements are transforming the way people work. All of these have stimulated the reformation of business attitudes to the flow of goods, services, information, and knowledge (Hardwick, Spooner, Rando, & Morris, 1996; Richards, Dudenhausen, Makatsoris, & de Ridder, 1997; Bouet & Martha, 2000; Johnston, 2001; Introna, 2001; Zobel & Filos, 2002).

BACKGROUND

Life has become more hectic: the hustle and bustle of global business, developing everyday situations, and worldwide

instant news coverage have intermingled business with leisure more than ever before. Collaboration, especially e-collaboration, is very important for today's business. It can take place at any time between enterprises and organizations, and moreover can be between people who are located in different places around the globe. But people are different and are motivated in many different ways, as well as having to work in multi-tasking environments within a variety of e-collaboration activities. Understanding how to work effectively with others in modern industrial and service-oriented society is key.

To bring about real benefits to society and business, modern communication means will need to be improved, extended, and seamlessly integrated with support services that can speedily call upon suitable tools, models, data, information and knowledge, and visualizations of entities from anywhere and at anytime to match priority and context. These will be the new e-collaboration environments, or electronic collaborative working environments, that are human centered and intuitive for the practical use of people, teams, and heterogeneous groups in an enriched virtual world serving them in their everyday tasks.

THE VISION

An enterprise network in an e-business context must be considered holistically with respect to its scale and scope to enable better e-collaboration across all its nodes and enable efficient and sustainable operations (Ballesteros & Richards, 2006). However, understanding and embracing people's needs is critical to support human interaction in such working environments (CEC, 2005, 2006). This calls upon the development of better and newer forms of approaches that are in stark contrast to existing approaches and practice. The transformation process involves bridging the gap between the way people think and work with others with the emergence of virtual service-oriented collaborative working environments. The challenge though is to have more than just an intuitive interface. To be human centered is more about having the right services with responsive personalized features in a fully immersive virtual environment that must be designed and implemented with full involvement of the real users.

THE CHALLENGES

ICT tools and systems are important enablers (CEC, 2000) in enterprise management and the transformation processes taking place. They have played and will continue to play a major role in the emergence of new ways of conducting business and ensuring their sustainable development. However, open global standards, protocols and interfaces, interoperable applications and platforms, trusted and sustainable infrastructure, and compatibility between business practices must be developed before interconnection for broader-based business is fully realized (Frick & Lill, 2000; Kidd, 2001).

However, innovative ICT alone is not enough. The necessary social and organizational changes to business (McCarthy, 1996) are at least as significant and are enabled by ICT. For instance, a Web-like organizational network has emerged from the more loosely coupled supply chains of the 1990s. The adaptive value network (Makatsoris, 2004) and virtual enterprise permit new forms of communication, participation, leadership, and decision making to develop. In turn these create new economic balances, shared learning, and new procedures embracing human involvement rather than strict structures dictated by inflexible ICT, which would just collapse space and time (Franke, 2001; Duttas, 2001) and increase resistance to change that must be overcome (Hunziker & Sieber, 1999; Deloitte & Touche, 2000).

Three basic aspects to change have emerged, before smarter business is accomplished, to drive the change process. These are developing in parallel to carry business forward to the future.

- **Organization:** How organization and inter-company relations are developed to ensure greater collaboration—that is, working jointly together, cooperating, and coordinating; trusting each other; sharing information and knowledge where appropriate and refining the skills in the organization to cope with the economics, strategic aims; and increasing the rate of innovation, day-to-day operations, and service excellence.
- **Information and Communication Technologies:** How tools and systems are created, developed, and introduced to ensure open, effective, and efficient dynamic engagement between companies, using all the appropriate communication channels open to them as necessary and of preferred choice. This applies in business networks, supply chains, and value networks, as well as at the retail end where new opportunities may be found. Such opportunities include competencies and innovative products and services that can enable the creation, enlargement, or optimization of adaptive value networks. For example such tools include, among others: distributed planning, distributed event-driven decision assistance and tracking, demand lifecycle control, collaborative management of uncertainty and risk, and collaborative design and provision of an environment for context-based e-collaboration services which can rapidly switch the context of working on demand.
- **Environment:** How exclusivity may be stripped away to provide global trade opportunities for all in the world of electronic trade, not only buying but also selling to any region or nation irrespective of differences in

language, culture, or local laws or regulation. Also importantly when designing, making, distributing, and selling products, discovering how to ensure a sustainable development of business that balances economic growth with energy consumption and environmental impact.

Some of these challenges are outlined in Tables 1 to 3.

FUTURE TRENDS

Consolidation of ideas and rethinking of strategy has been enabled not only by a political will (Timmers, 2002; CEC, 2002) and further research and development (CEC, 2003; CEC, 2005), but also by the current global socio-economic, environmental, and energy trends as well as by a greater respect for the social and economic changes among all the industrial players.

ICT providers and consultants to date have provided much of the impetus for industrial change, but industry has now to fully engage with strategic plans to complete the transformation process. One of the challenges highlighted in European statistics is that less than 20% of SMEs[1] had put in place any CRM or SCM software. Change through national incentives, joint projects, and education through regional clusters of companies in the same industry sector is being encouraged. However, it is now realized that technology cannot do it alone in the context of industry, business, and services. Both social and cognitive facets need to be fully understood and solutions found for deficiencies in order to create truly human-centered systems.

Europe, one of the world's largest trading blocks, has reconfirmed its Lisbon objective. The March 2005 European Council declared the aim of increasing the potential for economic growth and of strengthening European competitiveness by investing above all in knowledge, innovation, and human capital. This is reflected in other relevant and complementary work and objectives taking place in the Asian and American trading blocks. Current CEC research has provided many strategies for future manufacturing (Geyer, Scapolo, Boden, Dory, & Ducatel, 2003), e-collaboration (CEC, 2005), as well as delivering results for biometrics (IBG, 2003) of particular relevance to future authentication, and specific industrial supply chain innovations. Research initiatives, such as EC's Framework 7 with a focus on key information

Table 1. Some challenges to organization

- Lack of awareness by very large sections of the business community:
  - How best to educate and train staff
  - Provision of systems and interactive tools to encourage catalytic creative thinking
- Lack of trust for successful e-business
- Insufficient partners in an e-market — liquidity of an e-market
- Lack of perceived benefit
- Inequitable benefits:
  - How benefits and risk are fairly shared
- Lack of sustainable business models
- Limitation of e-collaboration to low-tier suppliers
- Needs to accelerate the business performance targets, for example:
  - Reduction of 'time to market'
  - Better precision for 'just-in-time production and logistics'
  - Provision of faster innovation cycles
  - Working capital efficiency
  - Increased resource utilization across the whole network
  - Improvements to distributed inventory and capacity management
  - Creation of new metrics for value network performance improvement
- Demand for specialty products in small batches through intelligent automated production planning and logistics systems
- Demand for astute fast mobile complex service engineering
- Inter-company e-collaborative design of high-quality complex products
- Lack of standard models and metrics for e-collaboration
- Lack of relevant support for people in multi-tasking jobs
- Lack of ability to manage and control large equipment/product integration produced in value networks to meet the necessary performance targets
- Needs to meet the demand of mass customization and personalization through detailed models and processes and smart, agile order processing
- Ability of companies to conceptualize a value network that is advantageous to them and to identify requirements for new skills, new metrics, and changes to be made
- Needs to have the ability to use shared knowledge and interpret shared information effectively among partners in a value network
- How much transparency should be between partners in a value network
- Belonging to many value networks at the same time
- Lack of standard business processes

Table 2. Some challenges for ICT

• Lack of a cost-effective and affordable, real-time worldwide communication of complex business information to serve the globally distributed nature of business
• Lack of interoperability between systems and applications
• Lack of appropriate multi-lingual facilities
• Lack of an affordable end-to-end high-quality seamless integration
• Lack of integrated workflow
• Free-flowing information flow in an end-to-end, timely, and secure manner
• Right economics for ICT suppliers
• Lack of smart applications and good decision assistance tools for collaborative business
• Need for data capture and filter systems that communicate in real time pertinent information to all partners in the value network
• Knowledge sharing and usage across the value network
• Cognitive decision making
• Tracking individual ordered products from source to consumer and providing a succinct record of such; the human food chain is a critical issue
• Development of intelligent agents as electronic proxies for individuals
• Serve the special needs of SMEs
• How to respond against unplanned/unexpected event with less cost
• Provision of an immersive environment that will automatically handle priorities by user choice, switch contexts of collaboration fast and in real time, and quickly provide all the necessary tools, data, information and knowledge, advanced visualizations to all involved in the collaboration process, and provide the means of management and administration for collaborations, which include new generation security measures and user choice IP protection
• Context orientation
• Intelligent lifecycle for both products and their demand support
• Integration methodologies and technology
• Provision of truly human-centered system
• Easy access to advanced simulators and computing power

Table 3. Some challenges for the business environment

• Too many standards developers and lack of coordination with sources of new work
• Slowness of standards take-up by commercial ICT developers
• Legal challenges:
○ Contract law
○ Cross-border provision of services
○ Protection of intellectual property rights
○ Privacy
○ Consumer protection
○ Electronic payments
• Regulation
• Security:
○ Trust
○ Cyber crime
○ Fraud
○ Unauthorized access
○ Computer attack
• Systematic framework to balance economic development with environmental and energy considerations

and communication technologies, will be a catalyst over the next few years for new working methods through the use of more innovative ICT and application. Other publications suggest that in the near future every single object will be connected to the Internet through a unique wireless address identifier allowing complete lifecycle tracking (Sarma, Brock, & Ashton, 2000; Murray, 2003; Datta, 2004). A roadmap for intelligent agent research is also available (SEEM, 2002).

Productivity is as much to do with people and organization than anything else, and a new way of thinking is highly desirable to deal with business survival and economics. New management styles that allow for greater flexibility and rapid smart responses and lifelong learning are essential to get right. Freeing up time is essential through proper prioritizations of all incoming messages, and events for pressures in the workplace are increasing, such as the overload from e-mail. Help from technology through personal electronic agents looking after both context and priority is one way of accomplishing this. Encouragement of effective and efficient collaboration too between people and mixed discipline teams within and without businesses will result in more creative thinking and help to increase the rate of innovation. Early studies and analysis of organizational webs (Tatnall & Gilding, 1999), value (Zobel & Filos, 2002), and inter-node relationships (Underwood, 2002) and new methodologies for design of networks (Soares, Sousa, & Barbedo, 2003), among many others, have helped to develop the social processes in complex organizations and their symbiosis with new business systems. And the Semantic Web will help to foster the much-needed greater understanding between organizations (Berners-Lee, 2001).

New people—known as the 'gaming generation'—are already arriving at the workplace and are more used to working in virtual spaces that supplement the already diverse backgrounds, characteristics, and behaviors found.

CONCLUSION

Metamorphosis will not be completed over night; the transformation process will take many years before completed. It is expected that a new end state for global business is characterized by an expanding global economy, greater collaboration in smarter value networks, versatile virtual organizations, better deployment of knowledge, greater dynamism, versatility and unhindered opportunities in markets, highly dynamic processes, new standards of customer service excellence, better use of both capital and human resource, and better decision making. This will be brought about through better designed and correct deployments of ICT, for structured, trusted, and secure electronic exchanges between companies. How we procure software and services will change dramatically. Technological advances will continue unabated to feed the desires for improvement. Also applications and tools

will continue to improve with aids for helping organization building, revising or creating new inter-business processes, as well as providing aids for flexible sizing for smart e-collaboration value network configurations. Universal standards will make integration between applications and distributed databases and knowledge bases both easier and cheaper. And new telecommunications technology will improve both the speed and the volume of information flow per unit time from anywhere at anytime. The workforce and end customers will experience a new way of living and fulfillment with the help of human-centered systems. The basic novelty of our age is the spirituality of consumerism and voracious appetite for information. How these may be blended with an increasing capacity for sympathy and mutual understanding will inspire both ways of life and collaborative e-business.

REFERENCES

Ansari, S.L., Bell, J.E., & the CAM-I Target Cost Core Group. (1997). *Target costing: The next frontier in strategic cost management.* New York: McGraw-Hill.

Arguello, M. (1994). *Review of scheduling software.* Technology Transfer 93091822A-XFER, Sematech, USA.

Ashkenas, R. (1995). Capability: Strategic tool for a competitive edge. *Journal of Business Strategy, 16*(6), 13-14.

Ballesteros, I.L., & Richards, H.D. (2006). *Workshop report on collaborative environments in manufacturing.* CEC D.G. Information Society and Media.

Bitran, G., Bassetti, P.F., & Romano, G.M. (2003). *Supply chains and value networks: The factors driving change and their implications to competition in the industrial sector.* MIT Center for E-Business Research Brief (vol. II, no. 3). Retrieved from *http://ebusiness.mit.edu*

Bologni, L., Gozzi, C., & Toschi, E. (1996). *Selecting software for small companies: The SPI 072/2 Experience.*

Bonfatti, F.,& Monari, P.D. (2004). Special needs of SMEs and micro-businesses. In *Evolution of supply chain management* (pp. 135-159). Boston: Kluwer Academic.

Bouet, D., & Martha, J. (2000). *Value nets: Breaking the supply change to unlock hidden profits.* New York: John Wiley & Sons.

CEC. (2000, January). *Directorate General Information Society, DGIS-C3-Electronic Commerce, Developing a coherent policy and regulatory framework for advancing electronic commerce in Europe.* Author.

CEC. (2002, May). Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of Regions. *Proceedings of eEurope 2005: An Information Society for All,* Brussels, Belgium.

CEC. (2003). *Sixth framework program.* Retrieved from *http://www.cordis.lu*

CEC. (2005). *Collaboration@Work: The 2005 report on new working environments and practices.* Author.

CEC. (2006). *Workshop report on collaborative environments in manufacturing.* Author.

Chang, Y., Makatsoris, C., & Richards, H.D. (2004). Design and implementation of an e-procurement system. *Production Planning and Control, 15*(7), 634-646.

Chang, Y., McFarlane, D.,& Shaw, A. (2001, August). *State-of-the-art system review.* Technical Report CUED/E-MANUF/TR.17, University of Cambridge, UK.

CSIRT. (2002). *Handbook of legislative procedures of computer and network misuse in EU countries for assisting computer security incident response teams.* European Commission, DG Information Society C4.

Datta, S. (2004). Adaptive value networks. In *Evolution of supply chain management* (pp. 3-67). Boston: Kluwer Academic.

Davenport, T.H. (1993). *Process innovation: Re-engineering work through information technology.* Boston: Harvard Business School Press.

Deloitte & Touche. (2000). *Manufacturing with a small e: An account of e-business in UK and US manufacturers.* Manufacturing Group Report, Deloitte & Touche, USA.

Drucker, P. (1992). *Managing for the future.* Oxford: Butterworth-Heinemann.

Drucker, P. (1994). The theory of business. *Harvard Business Review, 72,* 95-104.

Frick, V., & Lill, A. (2000, August 4). *Ten imperatives for e-business success.* Report, Gartner Group, USA.

Geyer, A., Scapolo, F., Boden, M., Dory, T., & Ducatel, K. (2003). *The future of manufacturing in Europe 2015 to 2020: The challenge of sustainability.* Scenario Report, Joint Research Center European Commission.

Greenberg, P. (2002). *CRM at the speed of light: Capturing and keeping customers in Internet real time* (2nd ed). New York: McGraw-Hill Osborne.

Goldman, S.L., Nagel, R.N., & Preiss, K. (1995). *Agile competitors and virtual organizations — strategies for enriching the customer.* Van Nostrand Reinhold.

Hardwick, M., Spooner, D.L., Rando, T., & Morris, K.C. (1996). Sharing information in virtual enterprises. *Communications of the ACM, 39*(2), 46-54.

Hunziker, D., & Sieber P. (1999). Turbulence and the dynamics of Internet diffusion. *Electronic Journal of Organizational Virtualness, 1*(1), 237-261.

Hutton, W. (1995). *The sate we're in.* Jonathan Cape.

Introna, L. (2001). Defining virtual organizations. In *E-commerce and v-business: Business models for global success.* Butterworth-Heinemann.

Kidd, P.T. (1994). *Agile manufacturing: Forging new frontiers.* San Francisco: Addison-Wesley.

Kidd, P.T. (2001). *E-business strategy: Case studies, benefits and implementations.*

Leachman, R.C., Benson, R.F., Lui, C., & Raar, D.J. (1996). IMPReSS—an automated production-planning and delivery-quotation system at Harris Corporation–Semiconductor Sector. *Interfaces, 26* (1), 6-37.

Macintosh, A. (1994). Corporate knowledge management state of the art review. *Proceedings of the SMICK (Management of Industrial and Corporate Knowledge) Conference,* Edinburgh, Scotland.

Makatsoris, C., & Chang, Y. (2004). Design of a demand-driven collaborative supply chain planning and fulfillment system for distributed enterprises. *Production Planning & Control, 15*(3), 256-269.

Makatsoris, C., Leach N.P., & Richards, H.D. (1996). Addressing the planning and control gaps in semiconductor virtual enterprises. In *IT and Manufacturing Partnerships: Delivering the promise.* Amsterdam: IOS Press.

Makatsoris, H., Chang, Y., & Richards, H. (2004a). Design of a distributed order promising system and environment for a globally dispersed supply chain. *International Journal of Computer Integrated Manufacturing, 17*(8), 679-691.

Makatsoris, H., Chang, Y., & Richards, H. (2004b). Collaborative sense-and-respond ICT for demand-driven value network management. In *Evolution of supply chain management: Symbiosis of adaptive value networks and ICT* (pp. 483-514). Boston: Kluwer Academic.

McCarthy, E. (1996). Culture, mind and technology: Making the difference. In *Human machine symbiosis* (pp. 143-176). London: Springer-Verlag.

Mertins, K. (1996). PULL-oriented synchronization of logistics and production flow in automobile industries. In *IT and manufacturing partnerships — delivering the promise.* Amsterdam: IOS Press.

Murray, C.J. (2003, September). *Network specifications released for every day products.* UK: E.E. Times.

Nonaka, I. (1998). The concept of "Ba." Building a foundation for knowledge creation. *California Management Review, Special Issue on Knowledge and the Firm, 40*(3).

Possl, G.W. (1991). *Managing in the new world of manufacturing.* Englewood Cliffs, NJ: Prentice Hall.

Richards, H.D., Dudenhausen, H.M., Makatsoris, C., & de Ridder, L. (1997). Flow of orders through a virtual enterprise—their proactive planning, scheduling and reactive control. *IEEE Computing & Control Engineering Journal,* (August), 173-179.

Sarma, S., Brock, D.L., & Ashton, K. (2000). *The networked physical world — proposals for engineering the next generation of computing, commerce & automatic identification.* Technical Report MIT-AUTOID-WH-001, MIT Auto-ID Center.

SEEM. (2002). *Workshop reports, October 1, 2002, and March 11, 2003.* Retrieved from *Europa.eu.net/information_society/topics/ebusiness/ecommerce/seem/index_en.htm*

Soares, A., Sousa, J., & Barbedo, F. (2003). Modeling the structure of collaborative networks: Some contributions. In *Processes and foundations for virtual organizations.* Boston: Kluwer Academic.

Tatnall, A., & Gilding, A. (1999). Actor-network theory and information systems research. *Proceedings of the 10th Australasian Conference on Information Systems* (pp. 955-966).

Underwood, J. (2002). Not another methodology: What ANT tells us about systems development. *Proceedings of the 6th International Conference on Information Systems Methodologies.* Retrieved from *http://www-staff.mcs.uts.edu.au/ ~jim/papers/ismeth.htm*

Warnecke, H.J. (1993). *The Fractal Company — a revolution in corporate culture.* Berlin: Springer-Verlag.

Wilson, R.M.S. (1995). *Strategic management accounting. Issues in management accounting* (pp. 159-190). Englewood Cliffs, NJ: Prentice Hall.

Womack, J.P., Jones, D.T., & Roos, D. (1990). *The machine that changed the world.* Rawson Associates.

Zabel, O., Weber, F., & Steinlechner, V. (2000). *Process reengineering and e-business models for efficient bidding and procurement in the tile supply chain.* Retrieved from *http://www.ebip.net*

Zobel, R., & Filos, E. (2002). *Work and business in the e-economy. Technology and policy issues in challenges and achievements in e-business and e-work.* Berlin: IOS Press.

KEY TERMS

Adaptive Value Network (AVN): An arrangement where companies form a web of close relationships and work together as a system that delivers the right customized product and expected service at the right quality in a coordinated manner and are responsive and adaptable to changes in the environment (Makatsoris et al., 2004b).

Collaborative Working Environment: A virtual environment that allows seamless access to all the necessary services for the context of effective and efficient collaboration, and permits all people and teams involved that may be located in any global site and belong to any organization to be fully engaged. The highly distributed and integrated and connected resources are managed to provide fast context switching, IP protection, intuitive user interfaces, multimedia and smart assistance tools, and so forth.

Electronic Business (E-Business): Any form of business or administrative transaction or information exchange that is executed using information and communications technology. This may be transaction performed in a peer–to-peer fashion between companies or organizations or with a customer. Electronic business impacts the way business is perceived.

Electronic Market (E-Market): A market free from inhibiting constraints and affordable for all businesses in any shape, form, or size, and to allow them to easily take part in e-business with beneficial returns. It is a market in which trust, security, and dependability apply and in which regulatory and legal issues are unified. It is a market where buyers and sellers ubiquitously execute business transactions online. These may include searching and identifying competence; ability to identify the right product or service together with quality, price, and quantity; and virtual auctions. It is also based on an open, secure, and reliable collaborative platform for knowledge exchange, joint product design, production planning, and logistics in stable customer-supplier relationships.

Intelligent Software Agent: Acts at speed over the electronic communication channel on behalf of human individuals or companies as their proxy; a program acting on behalf of another person, entity, or process. An intelligent software agent is an autonomous program that is capable of perceiving and interpreting data sensed from its environment, reflecting events in its environment, and taking actions to achieve given goals without permanent guidance from its user. Agents must have the intrinsic ability to communicate, cooperate, coordinate, negotiate, and learn, as well as have the capability to evolve through interactions with other agents. Agents can be standalone or part of a multi-agent system.

Smart Organization: A further evolution of value networks and virtual corporations through the use of more advanced business models taking account of human ICT symbiosis and utilizing more intelligent applications and tools for collaborative work and holistic development of both product and service engineering.

Supply Chain: In its basic form, a buyer-centric chain or network of independent companies that are loosely interlinked by activity along a manufacturing, servicing, and distribution channel of a product service specialty, from sources of raw material to delivery to an end customer. Supplementary to this supply chain management is a set of approaches utilized to integrate suppliers, manufacturers, warehouses, retail stores, and so on. Consequently, merchandise is produced and distributed in right quantities, to right locations at the right time, to minimize system-wide costs while at the same time satisfying service-level requirements.

Value Network: This term is ambiguous, as the analytical perspective colors its meaning. Nevertheless, the value network in general terms evolves from a supply chain through mutual use of ICT and more closely linked collaboration and mutual dependency between the partner organizations or independent companies. Collaboration means electronic communication via extranet, or Internet, co-operation and co-ordination of work flow, information and knowledge exchange, negotiation and dynamic trading, and joint decision making. Value is derived through the exchanges with partner organizations in the network and its shared knowledge. The value network also aims to deliver the highest value to the end consumer and to its stakeholders.

Virtual Enterprise/Virtual Corporation: A virtual corporation or enterprise is formed from a pool of competencies and capabilities resulting from a club of pre-qualified partners that may be expanded or contracted through the mutual desires of the club. The management body for a virtual enterprise selects partners from the pool of competence available to provide products or comprehensive services to any industry in direct competition to single companies or other virtual enterprises. It is necessary to have strong collaborative relationships between partners in the club. The virtual enterprise may exist only on a temporary basis to take market chances, for example tendering. It may also exist for a longer term for optimization of a value network to service a market need.

ENDNOTE

[1] SMEs together with micro-companies account for 70% of employment in Europe.

ICT and E-Democracy

Robert A. Cropf
Saint Louis University, USA

INTRODUCTION

The virtual public sphere does not exist and operate the same everywhere. Every virtual public sphere is different because each country's economic, social, political, and cultural characteristics and relations are varied. As a result, the impact of **information communication technology** (ICT) on political and social conditions will also differ from one country to another. According to the German philosopher, **Jürgen Habermas** (1989,1996), the public sphere is a domain existing outside of the private sphere of family relations, the economic sphere of business and commerce, and the governmental sphere dominated by the state. The public sphere contributes to democracy by serving as a forum for deliberation about politics and civic affairs. According to Habermas, the public sphere is marked by liberal core beliefs such as the freedoms of speech, press, assembly and communication, and "privacy rights, which are needed to ensure society's autonomy from the state" (Cohen & Arato, 1992, p. 211). Thus, the public sphere is defined as a domain of social relations that exist outside of the roles, duties and constraints established by government, the marketplace, and kinship ties.

Habermas' public sphere is both a historical description and an ideal type. Historically, what Habermas refers to as the bourgeois public sphere emerged from the 18th century Enlightenment in Europe and went into decline in the 19th century. As an ideal type, the public sphere represents an arena, absent class and other social distinctions, in which private citizens can engage in critical, reasoned discourse regarding politics and culture.

The remainder of this article is divided into three parts. In the first part, the background of virtual public spheres is discussed by presenting a broad overview of the major literature relating to ICT and democracy as well as distinguishing between virtual and public spheres and e-government. The second section deals with some significant current trends and developments in virtual public spheres. Finally, the third section discusses some future implications for off-line civil society of virtual public spheres.

BACKGROUND

ICT, in the eyes of some forward-thinking observers, (e.g., Abramson, Arterton, & Orrren, 1988; Barber, 1984; Becker & Slaton, 2000; Cleveland, 1985; Clift, 2004; Coleman & Gøtze, 2001; Cropf & Casaregola, 1996, 1998; Davis, Elin, & Reeher, 2002; Grossman, 1995; Negroponte, 1998; Rheingold, 1993; Saco, 2002) makes possible the type of public sphere envisioned by Habermas. According to these e-democracy advocates, ICT provides citizens with numerous opportunities to engage in the political process and take a more active role in the governance process. For example, a guide to effective public engagement notes: "A spectacular array of tools are emerging that give ordinary citizens a greater 'voice' in nearly every aspect of society today" (Lukensmeyer &Torres, 2006). Benkler (2006), for example, asserts ICT, in the form of the ubiquitous World Wide Web, encourages a more open, participatory, and activist approach to politics because it enables users to interact with other users in a way that the mass media does not and is therefore less susceptible to corruption than the mass media (p. 11). Nonetheless, the view that ICT can facilitate deliberative democracy is far from universal; a number of authors assert that technology creates its own set of problems with regard to democratic discourse practices (e.g., Margolis & Resnick, 2002; Sunstein, 2001; Taylor & Saarinen, 1995) Furthermore, these critics of technology argue correctly that advocates do not adequately account for the continued tenacity of mass media's stranglehold over the public discourse in liberal democracies well into the 21st century.

HOW ICT MAKES VIRTUAL PUBLIC SPHERES POSSIBLE

The idea that ICT serves as a catalyst for social and political change has been a motivating force behind netactivism since the dawn of the personal computing era. The ability of individuals to gain access to, store and manipulate vast amounts of information, which ICT makes possible, would lead to a situation where "vast numbers of people empowered by knowledge…assert the right or feel the obligation to make policy" (Cleveland, 1985). The potential of ICT, then and now, is that it enables a many-to-many, decentralized, and nonhierarchical flow of information. By contrast, mass media information flows in a top-down, one-to-many and centralized manner, which requires large amounts of capital investment, concentrating political and economic power in the hands of a small number of multinational or state-run corporations.

ICT consists of numerous tools that create linkages across space and time; allow people to build and participate in communities of their own choice; spread their message to diverse constituencies and collaborate over a networked environment. These tools include but are not limited to: 1) the World Wide Web (WWW), the graphical interface with the Internet; 2) **wikis**, a tool that enables individuals to collaborate on content creation using the WWW; 3) **blogs**, or Web logs, which allows individuals to upload personal content to the WWW; 4) social networking sites, which allow individuals with no pre-existing ties with each other to form online communities and engage in **peer-production**; and 5) handheld and other portable computing devices, which can be used by individuals and groups to communicate "from the field."

HOW VIRTUAL PUBLIC SPHERES DIFFER FROM E-GOVERNMENT

It is necessary here to distinguish between e-governance, which is the "product" of virtual public spheres, and **e-government**. E-government is the use of IT to provide governmental information to citizens and to assist in the delivery of public goods and services. E-government emerged as a phenomenon among Western governments during the mid-1990s. At that time, governments borrowed techniques and processes involving ICT already in use by businesses to facilitate consumer access to goods and services and to optimize management and organizational operations. The principal focus, then and now is on a "services first, democracy later approach" (Clift, 1998); in other words, using technology to provide government services more efficiently, that is, cut costs, rather than as a means to foster greater public engagement and civic deliberation about politics and government (Northrup, Kraemer, Dunkle, & King, 1990).

In this article, I define virtual public sphere as the use of ICT to achieve two chief ends of **e-democracy**: 1) to empower ordinary citizens to engage in effective public discourse regarding the proper ends of politics and the means to attain those ends and 2) to provide the technological means to effect public policy change. While building the necessary public infrastructure to support e-governance typically lags behind a service-based strategy in terms of public sector online efforts, more governments are collaborating with civil society to build online deliberation and policy-making spaces. As the global ICT revolution reshapes social and economic institutions, e-democracy, or "the use of information and communication technologies and strategies by democratic actors (government officials, the media, political organizations, citizens/voters)" (Clift, 2004, p. 38) will continue to make significant strides. Interestingly, the country with perhaps the greatest international reputation for the advanced use of ICT, the U.S., actually lags behind in the global e-government revolution of other countries such as Canada, South Korea and the UK. According to a report issued by the IBM Center for the Business of Government, the above-named countries "have taken giant strides toward modernizing citizen participation by creating policy frameworks and departments with mandates to coordinate citizen engagement online" (p. 34).

CURRENT TRENDS AND DEVELOPMENTS IN VIRTUAL PUBLIC SPHERES

As noted, a growing body of literature builds the case for ICT creating the necessary infrastructure for virtual public spheres, best exemplified currently by the WWW. According to empirical research, small-scale, virtual public spheres have emerged around the world, which closely resemble the ideas put forth by the visionary thinkers discussed earlier. This article, however, can provide only a surface treatment of the numerous experiments in deliberative e-democracy such as those below (year of implementation is in parentheses after project):

1. **The Hansard Society eDemocracy Programme, UK. (Ongoing). (http://www.hansardsociety.org.uk/programmes/e-democracy).** An effort to develop virtual public spheres around policy deliberations involving members of the UK Parliament, public officials and private citizens. The group's official Web site refers to these as "digital dialogues" or the use of ICT to enable and enhance civic engagement. A current initiative, as of October 2007, involves the Ministry of Justice and has three principal aims: 1) to promote knowledge of ICT-based engagement, 2) to cultivate online engagement skills in the central government, and 3) to analyze case studies to develop benchmarks for administrator and user demographics, attitudes and behaviors.
2. **Dialogue with the City, Perth, Australia (2003**). According to the Web site (http://www.wapc.wa.gov.au/Coast/Perth+coastal+planning+strategy/306.aspx), this process was implemented by the Western Australian government to engage the citizens of Perth in planning for their future in the face of some of the highest population and economic growth rates of any city in Australia, which places a significant demand on land, resources and environment. The stated aim of the project is to make Perth the most livable city by 2030.
3. **The International Centre of Excellence for Local eDemocracy (ICELE). UK. (2006).** The goal of

ICELE (http://www.icele.org/site/index.php), according to one of its founders, is "to harness new technologies to make it easy for people across the country to get involved in the democratic process." The Centre carries forward the work of the UK E-Democracy project, which commissioned the study of local best practices in e-democracy in the UK and ended in February 2005. The stated goal of the ICELE is to "provide best practice advice, support and practical solutions to help local authorities increase national eParticipation rates."

4. **North Jutland Democracy Project, Denmark. (2000).** The object of the Democracy Project was to create a virtual public sphere to facilitate deliberative e-democracy among the government and civil society just prior to Election Day, in response to the lowest voter turnout in Danish history just 4 years earlier. Citizens and public officials took part in the project, which resulted in "a very lively and well-visited Web site" (Coleman & Gøtze, 2001, p. 45).
5. **Kalix, Sweden. (2001)** An online "town hall" was created (http://www.stockholmchallenge.se/data/kalix_radslag) in September 2000 as a forum for citizens to engage directly with local politicians and public officials. Some limited electronic voting takes place and residents were given the opportunity to suggest city center redesigns. According to the Web site listed above, the project engaged 86% of its participants through the WWW.
6. **21st Century Town Meeting, USA (Ongoing)**. This is more of a series of ongoing deliberative democracy events rather than any one particular forum. The process was developed by AmericaSpeaks, a nonprofit organization for civic engagement (http://www.americaspeaks.org/). The objective of these town meetings is to bring together a diverse group of individuals to discuss issues and develop policy options. The importance of these deliberative forums from this article's standpoint stems from their innovative use of ICT: Networked laptop computers and wireless technology are used to share small group results with the larger body and decision makers. Examples of 21st Century Town Meetings include a national forum on social security reform, redevelopment of the World Trade Center site after the 9/11 terrorist attack, and as part of a process for biennial strategic planning in Washington, DC (Lukensmeyer & Torres, 2006, p. 25). More recently, 21st Century Town Meetings have been used to help New Orleans residents displaced by Hurricane Katrina develop a housing plan and to plan for pandemic flu outbreak in Maryland.
7. **Canadian International Polcy eDiscussions, Canada (Ongoing).** In 2003, Foreign Affairs Canada launched an effort to involve ordinary Canadians in online discussions on foreign policy issues affecting their country (See http://geo.international.gc.ca/cip-pic/participate/menu-en.aspx). Topics covered in the e-discussions have included "Canada's Role in North America" and "Canada's Approach to Democracy Promotion." E-discussion summaries are distributed to senior Canadian officials whose response is put up on the Web site.

These seven examples barely scratch the surface of the literally hundreds of other virtual public spheres which could have been chosen to illustrate the variety of deliberative e-democracy around the world. Clearly, the virtual public spheres described above indicate the promise inherent in the use of ICT in strengthening democracy. Benkler (2006), however, adds a note of caution: There is not much actual experience with public spheres "built on a platform that is widely distributed and independent of both governmental control and market demands" (p. 176). Thus, while citizens' ability to enter into direct contact with governments is greatly enhanced by ICT, it is still far more common for even democratic Western governments to downplay or ignore virtual public spheres in favor of those elements of the technology which largely improve operational efficiency.

FUTURE TRENDS

In general, while most of democratic governments' online efforts are well meaning, they often stop far short of enhancing civil society. However, as several of the earlier examples show, the truly outstanding efforts have been undertaken by grassroots, nongovernmental organizations that are actively engaged in virtual public sphere creation throughout the world. Indeed, the literature strongly suggests that ICT can help facilitate off-line civil society. Thus, the increase and augmentation of information sharing across individuals and groups in the future will be a major factor in social capital formation, which can be translated into collective action directed toward a common goal (Kavanaugh & Patterson, 2001). For example, in a study of a community technology center, the researchers found that the social relations engendered by the center were critical in promoting community change (Alkalimat & Williams, 2001). The effects on civil society of a virtual public sphere in Melbourne, Australia, found an increase in bonding capital (Meredyth, Hopkins, Ewing, & Thomas, 2002). Another study found that Internet usage bolsters social ties in a suburban community near Toronto, Canada (Hampton & Wellman, 1999). A study of over 20,000 Internet users found that ICT links geographically dispersed groups and individuals sharing common interests (Quan-Haase & Wellman, 2002). What all these studies have in common is the idea that online communities can help strengthen off-line communities. In the future, as

online communities grow and in many cases become intertwined, this will have a beneficial effect in bolstering off-line democracy, whether or not government takes an active role in promoting virtual public spheres.

CONCLUSION

Virtual public spheres, mostly on a small scale, are a reality in many parts of the world today. The networked and widely distributed nature of ICT renders virtual public spheres less vulnerable to the centralizing tendencies of the broadcast mass media, which has resulted in the concentration of economic and political power. ICT de-concentrates and diffuses power as exemplified by peer-production, the wide-scale cooperative behavior of the type that makes open-source software (e.g., **Linux**) and massive projects such as Wikipedia possible. Moreover, ICT makes possible the types of communications and linkages that enhances and expands social capital, thereby serving to strengthen civil society.

REFERENCES

Abramson, J.B., Arterton, F.C., & Orren, G.R. (1988). *The electronic commonwealth: The impact of new media technologies on democratic politics*. New York: Basic Books.

Alkalimat, A., & Williams, K. (2001). Social capital and cyberpower in the African American community: A case study of a community technology center in the dual city. In L. Keeble & B.D. Loader (Eds.), *Community informatics: Community development through the use of information and communications technologies.* London: Routledge.

Barber, B. (1984). *Strong democracy: Participatory politics for a new age*. Berkeley, CA: University of California Press.

Barber, B. (2003). Which democracy and which technology? In H. Jenkins & D. Thorburn (Eds.), *Democracy and new media* (pp. 33-48). Cambridge, MA: The MIT Press.

Becker, T., & Slaton, C. (2000). *The future of teledemocracy*. Westport, CT: Praeger.

Benkler, Y. (2006). *The wealth of networks*. New Haven, CT: Yale University Press.

Cleveland, H. (1985). Twilight of hierarchy. *Public Administration Review, 45*, 189-195.

Clift, S. (1998, March-April). Democracy is online. *OnTheInternet Magazine*. Retrieved May 28, 2008, from http://www.publicus.net/articles/democracyisonline.html

Clift, S. (2004). *E-government and democracy: Representation and citizen engagement in the information age.* Retrieved May 28, 2008, from http://www.publicus.net/articles/clift-egovdemocracy.pdf

Cohen, J., & Arato, A. (1992). *Civil society and political theory*. Cambridge, MA: The MIT Press.

Cropf, R., & Casaregola, V. (1998). Virtual town halls: Using computer networks to improve public discourse and facilitate service delivery. *Research and Reflection, 4*(1). Retrieved May 28, 2008, from http://www.iog.ca/policity/CP/Public%20Library/library_reference_virtual_town_halls.html

Cropf, R., & Casaregola, V. (2006). The virtual town hall. In A. Anttiroiko & M. Mälkiä (Eds.), *Encyclopedia of digital government.* Hershey, PA: Idea Group Reference.

Davis, S., Ellin, L., & Reeher, G. (2002). *Click on democracy: The Internet's power to change political apathy*. Boulder, CO: Westview Press.

Grossman, L.K. (1995). *The electronic republic*. New York: Viking.

Habermas, J. (1989). *The structural transformation of the public sphere*. Cambridge, MA: The MIT Press.

Habermas, J. (1996). *Between facts and norms: Contributions to discourse theory of law and democracy.* Cambridge, MA: The MIT Press.

Hampton, K.N., & Wellman, B. (1999). Netville online and off-line: Observing and surveying a wired world. *American Behavioral Scientist, 43*, 475-492.

Leadbeater. C. (2007). *Social software for social change: A discussion paper for the Office of the Third Sector.* Retrieved May 28, 2008, from http://www.publicus.net/articles/edempublicnetwork.html

Lukensmeyer, C., & Bringham, S. (2002). Taking democracy to scale: Creating a town hall meeting for the 21st Century. *National Civic Review, 91*, 351-366.

Margolis, M., & Resnick D. (2000). *Politics as usual: The cyber space "revolution."* Thousand Oaks, CA: Sage.

Meredyth, D., Hopkins, L., Ewing S., & Thomas, J. (2000). Measuring social capital in a networked housing estate. *First Monday.* Retrieved May 28, 2008, from http://www.firstmonday.org/issues/issue7_10/meredyth/index.html

Northrup, A., Kraemer, K., Dunkle, D., & King, J. (1990). Payoffs from computerization: Lessons over time. *Public Administration Review, 50*, 505-514.

OECD. (2003). *Promise and problems of e-democracy: Challenges of online citizen engagement.* Retrieved May 28, 2008, from: www.oecd.org/dataoecd/9/11/35176328.pdf

Putnam, R. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

Quan-Hasse, A., &Wellman, B. (2002). *How does the Internet affect social capital?* Retrieved May 28, 2008, from http://www.chass.utoronto.ca/~wellman/netlab/PUBLICATIONS/_frames.html

Rheingold, H. (1993). *The virtual community*. Reading, MA: Addison-Wesley.

Saco, D. (2002). *Cybering democracy: Public space and the Internet*. Minneapolis, MN: University of Minnesota Press.

Sunstein, C. (2001). *Republic.com*. Princeton, NJ: Princeton University Press.

Taylor, M., & Saarinen, E. (1995*). Imagologies: Media philosophy*. New York: Routledge.

KEY TERMS

Blog: Short for Web log. Generally, blogs are personal journals that are kept on the WWW and that can be updated frequently by a user, also known as a "blogger."

Civil Society: Refers to the sector that is different from business and government, which is constituted by voluntary associations including religious groups, labor organizations, citizens' groups and more.

E-Democracy: Using telecommunications technology by democratic actors, including governments, elected representatives, civic organizations, communities, political groups, and activists to improve the political process and political institutions. Examples of e-democracy include online discussion groups, blogs, government Web sites, and other forms of networked participation and civic engagement.

E-Government: Using telecommunications technology as a means to facilitate public administration and improve public access to government information and services.

Linux: An example of open-source software that is peer-produced. A free operating system for servers originally created by Linus Torvalds.

Peer-Production: Large-scale, often worldwide, collaborative efforts to create information, knowledge, and culture (see "Linux" and "Wikis."). Coined by Internet scholar, Yoshai Benkler.

Public Sphere: A concept that originates with the German social thinker, Jürgen Habermas, that refers to communications and relationships that are separate from the state, marketplace, and family structures. It serves to strengthen democratic institutions by serving as a space for deliberation regarding the means and ends of government and politics.

Social Networking: The use of ICT, particularly the WWW, to connect people who share common interests. Examples of successful social networking sites include MySpace, YouTube and Facebook.

Virtual Public Sphere: A group of people whose primary interaction is online but the felt experience of the individuals constituting the group is similar to an actual physical community.

Wiki: A software platform which enables any user to create, edit and otherwise add to Web page. An example of peer-production.

ICT Exacerbates the Human Side of the Digital Divide

Elspeth McKay
RMIT University, Australia

INTRODUCTION

Ethnic and racial tensions are aggravated by social inequities; perhaps it is the media that unwittingly feeds this dilemma. Look at how often we are directed to the Internet for further information. While exploring the Internet may be easy for some computer users, others demonstrate a complete avoidance for this type of knowledge exchange. Moreover, some misunderstandings that occur between cultural communities may be exacerbated by a phenomenon that has become known as the digital divide. There are various definitions of this term. One view takes a purely socio-political focus relating to a socio-economic gap between communities that have access to computer technologies and the Internet and those who do not. Another view covers a broader technological spectrum to include media of any sort, and the information and communications technologies (ICTs). No matter which definition is used, the lack of access to information, for whatever reason, may perpetuate a meaningful gap in cultural differences; the result of which may lead to a more serious communication breakdown throughout the community. This short article argues for more research on measuring the effectiveness of increased opportunities for Web-mediated cross-cultural/intergenerational knowledge sharing that is designed to overcome the ever widening digital divide.

BACKGROUND

It is useful to look at the ways in which a country such as Australia faces its significant socio-economic challenges. The literature reveals that this nation is one of the most multicultural countries in the world (Tsang, 1995), weaving cultural diversity and associated tensions into the social fabric. Furthermore, like many other nations, census statistics show us that Australia is fast becoming an aging nation. These two demographic features may give rise to communication problems associated with cultural and intergeneration discord. Unfortunately, current research appears to be ignoring the importance of the relationship between socio-cultural interaction and Web-mediated knowledge exchange. Moreover, there has been an unrealistic expectation that Web technologies would facilitate the engagement of people to share information through collaborative team work. Consequently, there were calls for researchers to become involved with ICTs to investigate these promises of collaborative Web-mediated information-sharing. Although this work has been taken up by the computer-supported collaborative learning (CSCL) protagonists, projects are still needed to correctly identify the complexity of the Web-mediated interactivity between humans and technology.

Even though the problem of rapidly evolving electronic multimedia was identified over a decade ago, the technologists are still excited today with our ability to create virtual information environments. However, for the more technologically challenged person, it would seem that we have become oblivious to how much we rely on ICTs that continue to change at an ever-increasing rate (Flicker, 2002). Using Australia again as the example, there is a distinct gap between theory and practice that exists within the population for opportunities to utilize ICTs to promote multicultural interaction and knowledge sharing. This disparity can be seen in terms of marked differences in access to the Internet for: enhancing multicultural sharing, promoting knowledge transfer between generations, and facilitating quality outcomes in special education (Stephanidis, 2001).

The aim of this short article is to suggest that ICTs can provide a useful set of easy tools to reduce some of the accessibility problems created by rapidly changing communications media. Issues that are causing concern among the communities who are cognizant of the harmful effects of the digital divide include: the forgotten human-dimension, cultural diversity, and the unequal accessibility to online information.

HUMAN ICT INTERFACE: IMPACTING FACTORS CAUSING CONCERN

To the discerning reader, the dualistic nature of human-computer interaction (HCI) is apparent. Some believe there is an intrinsic capability for HCI to span a socio-dimension in seeking solutions for people's problems; and moreover that ICTs provide effective conduits for producing appropriate outcomes. Yet as we get brighter, smarter machinery, the greater the perils of the digital divide become (McKay, 2005). If, in time, research can show this is the case, then, the predicted Big Brother phenomenon of the 1960s has

won. One only needs to look at the wireless technologies currently available. Consider the mobile phone technologies; the connectivity that is possible today is amazing. However, it is no secret that people do need to be in a position to purchase these smart-technology devices in order to join in this electro-knowledge revolution.

Inappropriate Web-Technologies can Exacerbate the Digital Divide

Deciding which technology to implement adds to the dilemma. While hardware and software standards are still evolving (Sonwalkar, 2005), the standards development's emanating from the Institute of Electrical and Electronics Engineers Inc. (IEEE) and the International Organization for Standardization (ISO) are largely mechanistic and concentrate on interoperability and integration. However, the human-dimension is taken up by the World Wide Web Consortium (W3C) and Web Access Initiative (WAI) accessibility standards (W3C, 2005). The W3C makes recommendations for Web-technologies; while the WAI deals with the issues of increasing accessibility for people with disabilities. The WAI Web-content accessibility guidelines are making progress to clarify the distinctions between technology and content. Unfortunately, Web-technologies can only be viewed as inert mechanisms that initiate nothing more than a means for online communication. Instead, our expectations of the social interface that implements ICT in a global context should involve two levels (Chan, 2003): secure ideological exchange between individuals, and clear representation of cultural perspectives. Therefore, Web-mediated knowledge sharing in this multi-dimensional environment will always be problematic.

Forgotten Human-Dimension

What is to happen with the people who cannot join in? This is where the human-dimension of HCI has an important role to play; to avoid or lessen the inevitable widening digital divide. Sadly, this role is not currently being fulfilled. Inequable access to information through ICT remains the status quo. The digital divide emphasizes the weaknesses of the human-dimensions of current HCI ideology. Moreover, there is a wrongly-focused assumption that all people can access information through a normal range of perceptual senses. To this end, we can see that the literature reveals there is a growing awareness of the belief that one-size-fits-all. This is more poignant through the business/government sectors where access to information affects the bottom line. An example of where ICT is not serving the wider community is to recall our recent graphic witnessing of naturally occurring disasters that affect whole nations; much of this global information sharing was only made possible through ICTs. These types of macro-events test out our ability to provide accessible information to all. Although the initial tsunami struck Indonesia 3½ hours before landing in Sri Lanka and India, our ICTs failed to convey any effective warnings! When it may be reasonable to think that equitable access to information is on the rise, the reverse may be true.

Why has this problem surrounding equal access to global information not been solved? In searching for an explanation of why the issue is now so acute; part of the answer may be that many of us with easy access to computers and the Internet appear to be blinded by technology. We have become accustomed to working quickly, many suffering from information overload. No wonder e-mail has become a common communications tool. However, in the rush to increase their coverage of the electronic information placed before them, tracking the snippets of misplaced knowledge scattered throughout our server mailboxes soon becomes unmanageable. Because of this, the human-dimension of HCI is now quite frail. To understand this, it is useful to reflect on how far we have come in a relatively short period of time.

It has taken less than 2-decades to see the information revolution unfolding. Our obsession on machinery was identified early by Dreyfus and Dreyfus (1986). Another example of where we commenced clinging to the notions that HCI could bring about the seamless ICTs we experience today can be seen in work on the Pask Conversation Theory and the subsequent computer language called *Protologic* (Lp). Pask concentrated on how to emulate the unique ways humans communicate. Although he was convinced that computer learning systems could teach by adapting to a learner's requirements (Pask, 1984), we are still unable to say this has been achieved (Izard & McKay, 2004).

Nonetheless, there is more compelling evidence of this ongoing tension between effective HCI and ICT continuing into the coming decade. Here is where the biggest problem exists. This is perhaps the real nexus of the HCI divide. Questions continue to arise about how best to represent human intelligence on a machine. Nobody really knows! The social sciences and artificial intelligence (AI) proponents cannot agree. On the one hand, the AI view of HCI continues to reflect their commitment towards a more mechanistic view of this phenomenon. While the softer sciences (philosophy, sociology, and anthropology) argue for understanding of the consequences for the human-dimension of technological developments (Preece, 2002). This vexing struggle is about making balanced and sensible choices for the type of HCI employed by our ICTs to advantage the human-dimension. No wonder that effective HCI remains unconquered. Even when there are recognized standards in place, where are the mechanisms in place to maintain compliance?

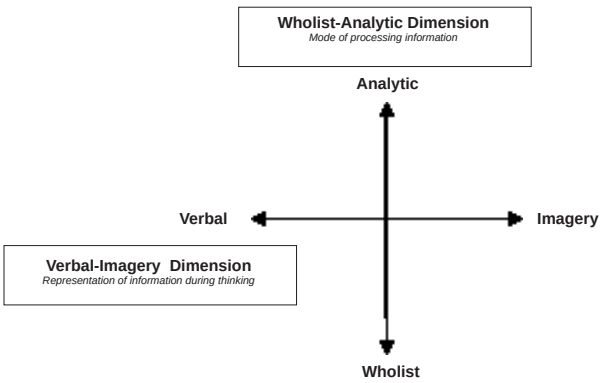
Cultural Sharing

With aging populations in the Western World and global migration in increasing numbers, intergenerational and cultural conflicts are major concerns. Web-mediated cross-cultural/intergenerational knowledge sharing can be designed to overcome facets of the ever-widening digital divide. According to Henry Tsang OAM, the success of cultural diversity depends on the spirit of sharing (Tsang, 1995). One common vehicle for cross-cultural sharing is through the abundant courses available for learning a second language (including ESL). Collecting experimental evidence is underway to provide the opportunity for dealing with the diverse nature of cultural and socio-contexts (McKay & Nishihori, 2004). This work adopts three distinct experiential environments. The first involves senior citizens in recalling traditional stories and games. The second engages youth (teenagers) to interact in multicultural settings, while the third relates to young children's delight in playful sharing of experiential learning. This notion of bringing about such interaction between the generations is not new. Nevertheless, much of this previous work is carried out within, not across, the disparate age groups. At the younger end of the age scale, new works are coming forward to describe early childhood development and ICT. More evidence is available for the youth with a rapidly growing interest in mobile technology for the school-aged group (Friedlander, 2004). While at the upper end of the age scale, there is a more limited number of successful projects that can describe intergenerational knowledge sharing (Kolodinsky, Cranwell, & Rowe, 2002). To bring about a positive change in this rather blinkered approach to reducing the negative effects of the so-called digital divide, this article suggests there is a real possibility of making positive inroads drawing on the strengths of HCI.

The Human Dimension of Accessibility

Perhaps all is not lost. There is a refreshed interest in the ways humans communicate online. The literature reveals research which distinguishes human ability to process information, as a combination of mode of processing information, and the way people represent information during thinking (see Figure 1). Riding and Rayner (1998) identified two fundamental cognitive dimensions (wholist-analytic and verbal-imagery) that affects performance in two ways. The first way is in how we perceive and interpret information we are given. While the second way is how we conceptualise related information already in our memory. Cognitive style is understood to be an individual's preferred and habitual approach to organizing and representing information. Measurement of an individual's relative right/left hemisphere performance and their cognitive style dominance has been a target of researchers from several disciplines over the last decade. Different theorists make their own distinctions on an individual's cognitive

Figure 1. Cognitive style construct (Riding & Rayner, 1998)



differences. However, little is known about the interactive effects of the cognitive style construct and multimedia delivery techniques on performance outcomes.

FUTURE TRENDS

Experimentation of the interactivity between humans and technology within a Web-mediated context has renewed the innovations that are emerging within the CSCL community (McKay & Nishihori, 2004). The expected collaborative interactions within such an online knowledge sharing entity are not just computerised collections of traditional games and stories but an opportunity for the capture and reuse of the interactive play with embedded instructional strategies.

It is proposed that Web-mediated knowledge sharing opportunities will nurture and cultivate individual creativity in both children and adult learners of English (Schank, 1988), by offering them an abundant supply of traditional games and stories. Children can be invited to a tightly-monitored, Web-mediated playing system, which includes specially designed playing environments where they can participate in a virtual exhibition room to enjoy collaborative play with other children. This type of research project pioneers Web-mediated processes for capturing and analysing this cross-cultural and inter-generational interaction (Nishihori, Okabe, & Yamamoto, 2002).

Spin-offs from this type of research project will provide an understanding of new ways to unite members of the community that are enjoyable, playful, and encourage lifelong learning. There is a real chance to discover ways to engage a diverse cross-cultural user network, including elderly citizens and younger family members to benefit from the collaborative story telling activities. Breaking down barriers to the digital divide will assist in strengthening family relationships and reduce cross-cultural tension in the Asia/Pacific region. Equitable access to information technol-

ogy will ensure each person has the choice to partake in the richness of 21st century communications technologies.

CONCLUSION

Seamless navigation tools are a move in the right direction. System's development teams must establish an understanding of their project's linguistics. Developing such a common nomenclature promotes synergies between ordinarily divergent disciplines. For example, product engineers are more familiar with technological issues than their end-user consulting counterparts who concentrate on the human factors relating to accessibility issues. More work is needed to bring about a paradigmatic knitting together of disparate professional communities of practice. For instance, in a system's development team, the knowledge engineer may not have sufficient technological skills to build the devices they perceive; while the software specialist may not be able to articulate the human-dimension.

Unless research can correctly evaluate the ingredients of the digital divide, the gap between theory and practice will continue to widen. Lost will be the understanding of the benefits that semiotic/discovery activities may elucidate for cross-cultural sharing. Lost too will be the knowledge and experiences of a dying generation who fought wars to allow our creativity to flourish.

REFERENCES

Chan, A. (2003). Communication technology and theory: Research into the interpersonal and social interface. *Gravity7*. Retrieved January 23, 2006, from http://www.gravity7.com/articles_investigation_toc.html

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York: Free Press.

Flicker, B. (2002). *Working at warp speed: The new rules for projectsuccess in a sped-up world.* San Francisco: Berrett-Koehler.

Friedlander, J. (2004). Cool to be wired for school. *Sydney Morning Herald,* April 16.

Izard, J., & McKay, E. (2004, November 28-December 2). *Automated educational/academic skills screening: Using technology to avoid or minimise effects of more formal assessment.* Paper presented at the Australian Association for Research Education (AARE 2004): Positioning education research, Melbourne. Retrieved January 23, 2006, from http://www.aare.edu.au/04pap/iza04951.pdf

Kolodinsky, J., Cranwell, M., & Rowe, E. (2002). Bridging the generation gap across the digital divide: Teens teaching Internet skills to senior citizens. *Journal of Extension, 40*(3). Retrieved January 23, 2006, from http://www.joe.org/joe/2002june/rb2.html

McKay, E. (1999). An investigation of text-based instructional materials enhanced with graphics. *Educational Psychology, 19*(3), 323-335.

McKay, E. (2005, July 10-13). *Human-computer interaction: Perils of ubiquitous information and communications technologies.* Paper presented at the 9th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2005), Orlando, FL. Retrieved January 23, 2006, from http://www.iiisci.org/sci2005/website/default.asp

McKay, E., & Nishihori, Y. (2004, December 4-6). Towards closing the digital divide: A multicultural, Intergenerational ICT case study. In E. McKay (Ed.), *International Conference on Computers in Education: Acquiring and Constructing Knowledge Through Human-Computer Interaction: Creating new visions for the future of learning,* Melbourne, Australia.

Nishihori, Y., Okabe, S., & Yamamoto, Y. (2002). *Creating cross-cultural learning communities on the Internet and the Japan gigabit network—Integration of media tools intocollaborative learning*. Paper presented at the International Conference on Computers in Education ICCE 2002, Auckland, NZ.

Pask, G. (1984, Spring). Review of conversation theory and a protologic (or protolanguage), Lp. *Educational Communications & Technology Journal, 32*(1), 3-40.

Preece, J. (2002). *Interaction design: Beyond human-computer interaction* (1st ed.). Harlow, UK: Addison-Wesley.

Riding, R. J., & Rayner, S. (1998). *Cognitive styles and learning strategies.* UK: Fulton.

Schank, R. C. (1988). Creativity as a mechanical process. In R. J. Sternberg (Ed.), *The nature of creativity.* New York: Cambridge University Press.

Sonwalkar, N. (2005). Demystifying learning technology standards part I: Development and evolution. *Campus Technology: 101 Communications.* Retrieved August 19, 2005, from http://www.campus-technology.com/article.asp?id=6134

Stephanidis, C. (2001). *Towards universal access in the information society: Achievements and challenges (Internal Report).* Crete: Paper presented at the Workshop on Universal Accessibility of Ubiquitous Computing: Providing for the Elderly. ICS-FORTH (HCI Laboratory) Internal

Report. Retrieved August 6, 2005, from http://www.ics.forth.gr/hci/files/PAPER_ON_UNIVERSAL_ DESIGN_stephanidis_2001.pdf

Thomas, T. (1998, October). *Intergenerational perspectives: Older persons through the eyes of the younger generations.* Paper presented at the 33rd Annual Conference of the Australian Association of Gerontology, Melbourne.

Tsang, H. (1995). *Designing for diversity: The multicultural city.* Paper presented at the 1995 Global Cultural Diversity Conference, Sydney.

W3C. (2005). *Leading the Web to its full potential...* Retrieved January 23, 2006, from http://www.w3.org/

WAI. (2002). *Web Access Initiative (WAI): Five primary areas of work.* Retrieved August 19, 2005, from http://www.w3.org/WAI

KEY TERMS

Cognitive Style Construct: This term represents Richard Riding's cognitive style dimensions. First, the wholist-analytic dimension (mode of processing information) defines that wholist learners are able to perceive the whole concept, but may find difficulty in disembedding its separate facts (McKay, 1999). Whereas Analytic learners analyse material into its parts but find difficulty in seeing the whole concept. Second, the verbal-imagery continuum (mode of representing information while thinking) measures whether an individual is inclined to represent information verbally, or in mental pictures, during thinking (Riding & Rayner, 1998).

Computer-Supported Collaborative Learning (CSCL): An emerging and important educational research paradigm that focuses on socially-oriented theories of learning using computer technologies to support collaborative methods of instruction.

Cross-Cultural/Intergeneration Knowledge Sharing: The issue of cross generational communications breakdown. Difficulties that arise through intergenerational perspectives (Thomas, 1998) can be witnessed in many parts of the community: at home with parents and siblings, in workplace reporting networks with age differences of employees, and at school between staff and students. Knowledge sharing across these boundaries will improve the disparity that currently exists.

Digital Divide: Understanding of this term is changing with time. Originally, it signified a socio-political environment and referred to the socio-economic gap between groups of people who had access to computers and the Internet. However, in 2005, this term reaches much further to bring about a more international context to highlight where disadvantaged groups and developing or poorer nations are pitted against the wealthy countries that have unlimited access to the plethora of electronic gadgetry.

Human-ComputerInteraction (HCI): The concept of HCI is not new. However, the proliferation of HCI has only occurred during the last decade. People building computerized systems have struggled for many years to define interactive relationships of HCI. The term should not be used just for the physical interaction between humans and computers; it includes all the cognitive processes required for humans to effectively utilize computerized medium.

Interoperability: Hardware and software devised to help people exchange information ensuring computer systems can to talk to each other, interpret data, and exchange information.

Smart-Technology Devices: Hardware/products that have AI enhanced capabilities and/or the ability to access the Internet. Some of these devices are designed to sense your actions or learn your patterns and alter their behaviour accordingly.

Virtual Information Environments: Internet-based information systems that provide a user with an interface that implements a multi-dimensional and real-world context through AI contrived metaphors that create highly visualized and often interactive experiential space.

Web Content Accessibility Guidelines (WCAG): The four design principles for Web accessibility include:

1. content must be perceivable;
2. interface elements in the content must be operable;
3. content and controls must be understandable; and
4. content must work with current and future Web technologies (WAI, 2002).

Web-Mediated Knowledge Exchange: The interactive use of online instruction as an effective tool in bringing about a knowledge-sharing culture, linking professional practice and education in life-long learning.

IDS and IPS Systems in Wireless Communication Scenarios

Adolfo Alan Sánchez Vázquez
University of Murcia, Spain

Gregorio Martínez Pérez
University of Murcia, Spain

INTRODUCTION

In principle, computers networks were conceived to share resources and certain computing devices among a select group of people working in academic institutions. In this context, the security did not have high importance. Today, through the network circulates a lot of valuable data (budgets, credit card numbers, marketing data, etc.), much of which can be considered confidential. Here is where security takes great importance—so that these data cannot be read or modified by any third party, and the services offered are always available and only to authorized people (confidentiality, integrity, and readiness).

When we refer to security, there are some terms of great importance. *Risk* is defined as any accidental or not prospective exhibition of information as consequence of the bad operation of hardware or the incorrect design of software. *Vulnerabilities* indicate when a failure in the operation of software and/or hardware elements exposes the system to penetrations. Starting from here we can define *attack* as an event against the good operation of a system, and it can be successful or not. If the attack is successful and access is obtained to the files and programs or control is obtained to the computers without being detected, then we are dealing with a *penetration.* This leads to an *intrusion,* which is a group of actions compromising the integrity, confidentiality, and readiness of computer resources (Sobh, 2006).

The main objective of this article is to explain to the reader the main concepts regarding intrusion detection systems (IDSs) and intrusion prevention systems (IPSs), and the particular issues that should be additionally considered when protecting wireless communication scenarios (in comparison with IDSs/IPSs in traditional wired networks). It also includes an extended view of the current state of the art of IDSs and IPSs in wireless networks, covering both research works done so far in this area, as well as an analysis of current open source IDSs and IPSs, and how they are dealing with the specific requirements of wireless communication networks.

This article is organized as follows: First, we start with a summary of the main related works in the *background* section; then we give a description of the important concepts of security, a classification of intrusion detection systems, and a brief comparative of the operation of IDSs in wired and wireless networks. Next, we highlight certain research works exemplifying efforts done so far in wireless scenarios. We present the main ideas behind our current research work to model intrusions in wireless scenarios, before offering future directions of work and a summary of the main ideas expressed in the article.

BACKGROUND

Many have been the efforts carried out to counteract the main weaknesses of IDSs and IPSs in wireless networks. In this sense, we can speak of different investigations directed to specific attacks in wireless scenarios by means of detection mechanisms based on artificial intelligence, design of monitoring IDSs, proactive IDSs, modeling of IDSs, and approaches based on system requirements and political issues.

Aime, Calandriello, and Lioy (2006) propose a mechanism of attack detection based on the shared monitoring of the networks by all the nodes, where one will be able to determine if the event is a bad behavior or an attack. The key idea is to install a monitor (ethereal) in each node of the network, and to produce evidence (information about the state of the network) and to share that among all the nodes. For each captured package, the system spreads a complete view of the packet headers, some general statistics are added such as timestamp, frame number, and longitude in bytes. The focus is 802.11 frames, although they are also considering source, destination, and BSSID addresses, sequence number, frame type, subtype, and retry flag. With this data, a list of events is built in each node.

We also find efforts based on anomaly detection by means of artificial neural networks, where the intrinsic characteristics and observables of the normal behavior are different from the abnormal behavior. A clear example is the proposal of Liu, Tian, and Li (2006), whose method of detection of intrusions is based on DGNN (dynamic grow-

ing neural network) and consists of Hebbian learning rules to which are added new neurons under certain conditions. Three of the more common attacks in wireless networks are: war driving, MAC spoofing, and WEP cracking, and we find investigation focused on preventing this type of attack. Hsieh, Lo, Lee, and Huang (2004) developed a proactive wireless hacker detection system (WIDS) basically based on a framework that proposes an answer plan designed to prevent the user in intranets of additional damages for each attack type. The structure of this proposal consists of five modules: packet capture, session analysis, hacker detection, honeypot, and alarm modulates. Tsakountakis, Kambourakis, and Gritzalis (2007) propose a design of WIDS modules to tackle 802.11i-specific attacks. Also, it evaluates the 802.11i-enabled WIDS modules, namely WIDZ. The tests were performed utilizing the majority of well-known open source attack tools and specific attack generators. Some types of attacks were detected and others were not detected.

WIRELESS INTRUSION DETECTION AND PREVENTION SYSTEMS

Wireless networks are particularly susceptible to attacks as interception and injection, to mention just one example. The problem is inherent to wireless protocols since they use the air as its primary means of transmission. Contrary to the conventional network where the location of a network is physically limited by the infrastructure of the network, the locus for a wireless device is not limited to a connection network backplane. This represents serious problems in the moment of deploying IDS. In wireless networks, IDSs receive packages of any closer networks or next antennas; this means that the IDS may process a great quantity of malicious packages and of unknown origin. The signs in wireless networks vanish in cloud form around the access point, and the signal radiated is attenuated, allowing that, in the periphery of the cloud, a corrupted signal and certain packets originate.

The IDS based on a network (NIDS) commonly listens to the network, captures and examines the packages flowing through the same network in contrast with the firewalls; NIDS can analyze the entire packet, not just the header. It is able to look at the payload within a packet to see which particular host application is being accessed, and raise alerts when attackers try to exploit a bug in such code. NIDSs are host independent, and can run like "black box" monitors to cover entire networks. In practice, active scanning slows down the network considerably and can effectively analyze a limited bandwidth network. NIDSs often required dedicated hosts/special equipment and can be prone to the network attack (Kachirski & Guha, 2003).

The IDS based on a host (HIDS) only cares what is happening in each individual host; monitors specific files, logs, and registries installed in a single computer; and they can alert any access or modification in the object monitored (Chari & Cheng, 2003).

They are able to detect actions such as repeated failed access attempts or changes to critical system files, and normally operate by accessing log files or monitoring real-time system usage. To ensure effective operation, host IDS clients must install every host of the network, tailored to specific host configuration. Host-based IDSs do not depend on network bandwidth and are used for smaller networks, where each host dedicates processing power towards the task of system monitoring (Kachirski & Guha, 2003).

IDSs, according to their functionality, are classified in anomaly detection and misuse detection systems (Tombini, Devar, Mé, & Duccassé, 2004). Anomaly detection consists of the detection of intrusions based on the non-habitual behavior of a system or the resources of the same one. The objective is to evaluate the correct use and acceptable behavior of a certain system, pointing out any activity that is outside of this behavior. Among the more highlighted efforts in this aspect, we find the use of statistical analysis, data mining, and limitation of flow or traffic.

Misuse detection is directed to the identification of intrusions in a system by means of the establishment of static information, providing a necessary base to determine a malicious activity according to how the static information is structured. A classification is obtained that is of vital importance; a classifier trains to discriminate, being based on the data of the network traffic.

The use of wireless networks has generated important changes in the implementation of the IDS, as wireless networks are conceptually and operationally different. For the wired system, the IDSs are distributed applications analyzing events in a network system to identify malicious behavior (Vigna, Valeur, & Kemmerer, 2003). And for wireless systems, intrusion detection systems analyze events in mobile and ad hoc networks. For this reason any discussion in relation to these architectures and their operational atmospheres will allow an in-depth exploration of the important aspect to consider in the moment of displays in an effective way wireless intrusion detection systems.

Stallings (2003) shows us in the Table 1 the clear limitation of IDS systems in wireless networks.

Already mentioned were the natural risks of a wireless network, helping us to identify a series of possible attacks that put in risk the security of the information and the stability of the computational systems. Table 2 summarizes these possible attacks, found in a recent investigation by Zhong, Khoshgoftaar, and Nath (2005).

IPS

An intrusion prevention system (IPS) is a mechanism that tries to provide an automatic, efficient, quick, and exact answer

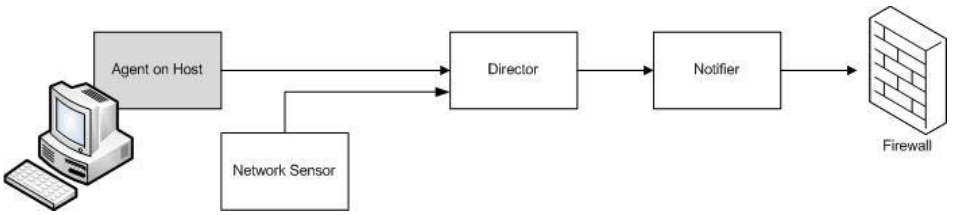
Table 1. A summary of the clear limitation of IDSs in wireless networks

• Network-based IDS sensors that have been placed on the wired network behind the wireless access point will not detect attacks directed from one wireless client to another wireless client (i.e., peer to peer) on the same subnet.
• Network-based IDS sensors on the wired network usually will not detect attempts to "dissociate" a legitimate client from the wireless network and will not detect the association of an unauthorized wireless client with the wireless network.
• A network-based IDS for a wired network will not be able to detect the physical location of the compromised host or rouge access point.
• A network-based IDS for a wired network will not be able to detect an authorized wireless device communicating peer to peer with an unauthorized wireless device.

Table 2. A summary of possible attacks to wireless networks

Attack Type	Attack Name
Passive	War Driving
	Man-in-the-Middle Attack
	High-Power Amplifiers
	Dictionary Attack–WPA
Masquerade	MAC Address Spoofing
	Bypassing Access Control List
	Authenticated User Impersonation
	Invalid State
	De-Authentication
	Disassociation
	ARP Poisoning
	MAC-Based Inference of ACL
	Virtual Carrier Sense Attack
Replay	Packet Re-Routing
Modify	Packet alteration
	Packet insertion
Denial of Service	Denial of Service
	RTS/CTS Flood
	Fragmentation Attacks
	Wormhole Attacks
	Network Injection Attacks
	Multiple Virtual Access point

Figure 1. The architecture of an IPS (Bishop, 2003)



to an intrusion attempt or attack that can cause damages in computational resources. The objective of the IPS is mainly to prevent attacks against the place being protected. Some of the main characteristics of such a system are that it should be easily adjustable and be integrated in the same way to the network of the organization. In this way, it can quickly answer, in real time, an attack (well known) in progress, although it is important to mention that these systems should have an index of false negatives. The main decisions of an IPS take in the expert motor of the system based on training; these periods of training should be minimal, thus allowing that details in a manual way are tuned. Figure 1 shows the architecture of an IPS.

MOTIVATION FOR AN INFORMATION MODEL IN WIRELESS INTRUSION DETECTION SYSTEMS

Network, service, and application management today face numerous challenges, ones that older ways of doing things cannot solve. The concept of policy-based management (PBM) addresses some of these problems and offers possible solutions. It provides a system-wide view of the network and its services and applications, and shifts the emphasis of network management and monitoring away from specific devices and interfaces towards users and applications. One of the main goals of policy-based management (Verma, 2002) is to enable network, service, and application control and management at a high abstraction layer.

The administrator specifies rules that describe domain-wide policies that are independent of the implementation of the particular network node, service, and/or application. It is then the policy management architecture that provides support to transform and distribute the policies to each node and thus enforce a consistent configuration in all the elements involved.

Policy rules are independent of a specific device and implementation, but they define in abstract terms a desired behavior. There are two main perspectives to define a policy (Westerinen, Schnizlein, & Strassner, 2001), and each represents a different level of abstraction: business level and technological level. A human policymaker initially provides the business-level policy using an informal language, then the task of a policy administrator is to transform the business policies into formal policies using a representation. For doing this, the administrator uses a policy language that assures that representation of policies will be unambiguous and verifiable.

As part of our investigation, and following this motivation, we are currently developing a common model based on the CIM (common information model) standard of the Distributed Management Task Force for the representation of intrusions in wireless scenarios. Since there are a lot of attacks that can be developed in these wireless environments, and authors do not come to an agreement on the exact meaning and scope of concepts like virus, worm, and so forth, this has led us to the idea of extracting the semantic content and providing the concept of patterning the power that allows the user to create a classification by him or herself.

FUTURE TRENDS

As statement of direction, we are now designing and testing an extensible information model where different concepts relevant to wireless IDSs can be expressed. We are also implementing this model as part of Snort and linking the information expressed here with the attack rules provided by the international community already working in wireless IDSs. Additionally, this model will be tested in our labs using complex attacks for wireless scenarios, as those explained as part of this article.

CONCLUSION

Intrusions in wireless networks represent a serious problem for economic, political, and educational organizations since many of them are supporting their main activities in wireless environments. Wireless intrusion detection/prevention systems are still needed tools for this monitoring and management of countermeasures against attacks.

In this article, we have provided a revision of the state of the art of IDSs in wireless environments and their possible attacks. This allows having a clear idea of the problem of representing attack information and its possible solutions.

Wireless devices and their corresponding protection have propitiated an increment in the use of intrusion detection systems which generates in many occasions the problem of representing the same information in diverse formats as a result of the generation of heterogeneous environments. It is necessary to highlight that there is a need to generate a model that represents in a uniform way the intrusions that can be generated in these environments.

REFERENCES

Aime, M.D., Calandriello, G., & Lioy, A. (2006). A wireless distributed intrusion detection system and a new attack model. *Proceedings of the 11th IEEE Symposium on Computers and Communications* (ISCC'06) (pp. 35-40).

Bishop, M. (2003). *Computer security: Art and science.* Reading, MA: Addison-Wesley Professional.

Chari, N.S., & Cheng, P. (2003). BlueBoX: A policy-driven, host-based intrusion detection system. *ACM Transactions on Information and System Security*, 6(2), 173-200.

Tombini, E., Debar, H., Mé, L., & Ducassé, M. (2004). A serial combination of anomaly and misuse IDSs applied to HTTP traffic. *Proceedings of the 20th Annual Computer Security Applications Conference* (ACSAC'04) (pp. 428-437).

Hsieh, W.-C., Lo, C.-C., Lee, J.-C., & Huang, L.-T. (2004). The implementation of a proactive wireless intrusion detection system. *Proceedings of the 4th International Conference on Computer and Information Technology* (CIT'04) (pp. 581-586).

Iheagwara, C., Blyht, A., & Bennet, M. (2005). Architectural and functional issues in systems requirements specifications for wireless intrusion detection system implementation. *Proceedings of the 2005 Conferences on Systems Communications* (ICW'05, ICHSN'05, ICMCS'05, SENET'05) (pp. 434-441).

Kachirski, O., & Guha, R. (2003). Effective intrusion detection using multiple sensors in wireless ad hoc networks. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences* (HICSS'03).

Karygiannis, T., & Owens, L. (2002). Wireless network security 802.11, Bluetooth and handheld devices. In *Recommendations of the national institute of standards and technology* (special publication, pp. 800-848). NIST.

Liu, Y., Tian, D., & Li, B. (2006). A wireless intrusion detection method based on dynamic growing neural network. *Proceedings of the 1st International Multi-Symposiums on Computer and Computational Sciences* (IMSCCS'06) (vol. 2, pp. 611-615).

Network Chemistry. (n.d.). *Security networks from rogue devices.* Retrieved from http://networkchemistry.com

Sobh, S.T. (2006). Wired and wireless intrusion detection system: Classifications, good characteristics and state-of-the-art. *Computer Standards & Interfaces, 28,* 670-694.

Stallings, W. (2003). *Cryptography and network security principles and practices* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Tsakountakis, A., Kambourakis, G., & Gritzalis, S. (2007). Towards effective wireless intrusion detection in IEEE 802.11i. *Proceedings of the 3rd International Workshop on Security, Privacy and Trust in Pervasive and Ubiquitous Computing* (SecPerU 2007) (pp. 37-42).

Verma, D.C. (2002). *Simplifying network administration using policy-based management.* (vol. 16, pp. 20-26). Yorktown Heights, NY: IBM Thomas J. Watson Research Center.

Vigna, G., Valeur, F., & Kemmerer, R.A. (2003). Designing and implementing a family of intrusion detection systems. *Proceedings of the 9th European Software Engineering Conference held jointly with the 10th ACM SIGSOFT International Symposium on Foundations of Software Engineering* (pp. 88-97).

Westerinen, A., Schnizlein, J., & Strassner, J. (2001). *Terminology for policy- based management.* Request for Comments 3198, IETF Network Working Group.

Zhong, S., Khoshgoftaar, T.M., & Nath, S.V. (2005 November 14-16). A clustering approach to wireless network intrusion detection. *Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence* (pp. 190-196). Washington, DC: IEEE Computer Society.

KEY TERMS

Common Information Model (CIM): An approach from the Distributed Management Task Force for the man-

agement of systems and networks; it applies to the basic techniques of the object-oriented paradigm. The approach uses a uniform modeling formalism that, together with the basic repertoire of object-oriented constructs, supports the development of an object-oriented schema across multiple organizations.

Host-based Intrusion Detection System (HIDS): Protects the machine in which it is installed. The data generated are used as a source of information, especially by the computer that operates at the operating system level: audit files of the system, files logs, or any file that the user wants to protect.

Intrusion Detection System (IDS): Hardware and/or software with certain intelligence monitoring able to analyze automatic events that happen in a computing system or network. The system's main objective is to identify possible threats and to carry out response actions.

Intrusion Prevention System (IPS): The set of mechanisms trying to provide an automatic, efficient, quick, and exact answer to intrusion intent or attack that can cause damages in a host or network. The objective of the IPS is mainly to prevent attacks against the entity (or entities) being protected.

Network Security: Those activities, techniques, or rules dedicated to prevent, protect, and preserve information and resources.

Network-based Intrusion Detection System (NIDS): Uses traffic networks and TCP/IP packages as sources of information. These systems revise the packages that circulate through the network searching for elements that denote an attack against some of the systems located in it. Of these packages they verify the validity of some parameters and the behavior of the protocols.

Policy Rules: Set of management rules independent of a specific device and implementation, and defining in abstract terms a desired behavior. These are stored and interpreted by the policy framework, which provide a heterogeneous set of components and mechanisms that are able to represent, distribute, and manage policies in an unambiguous, interoperable manner, thus providing a consistent behavior in all affected policy enforcement points

Wireless Attack: Malicious activities putting at risk the security of the information and of the computing resources in wireless scenarios.

Image Compression Concepts Overview

Alan Wee-Chung Liew
Griffith University, Australia

Ngai-Fong Law
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Image compression aims to produce a new image representation that can be stored and transmitted efficiently. It is a core technology for multimedia processing and has played a key enabling role in many commercial products, such as digital camera and camcorders. It facilitates visual data transmission through the Internet, contributes to the advent of digital broadcast system, and makes possible the storage on VCD and DVD. Despite a continuing increase in capacity, efficient transmission and storage of images still present the utmost challenge in all these systems. Consequently, fast and efficient compression algorithms are in great demand.

The basic principle for image compression is to remove any redundancy in image representation. For example, simple graphic images such as icons and line drawings can be represented more efficiently by considering differences among neighbor pixels, as the differences always have lower entropy value than the original images (Shannon, 1948). These kinds of techniques are often referred to as lossless compression. It tries to exploit statistical redundancy in an image so as to provide a concise representation in which the original image can be reconstructed perfectly.

However, statistical compression techniques alone cannot provide high compression ratio. To improve image compressibility, lossy compression is often used so that visually important image features are preserved while some fine details are removed or not represented perfectly. This type of compression is often used for natural images where the loss of some details is generally unnoticeable to viewers.

This articles deals with image compression. Specifically, it is concern with compression of natural color images because they constitute the most important class of digital image. First, the basic principle and methodology of natural image compression is described. Then, several major natural image compression standards, namely JPEG, JPEG-LS, and JPEG 2000 are discussed.

BACKGROUND

A common characteristic of most images is that the neighboring pixels are correlated and thus contain redundant information. The main goal of image compression is to reduce or remove this redundancy. In general, two types of redundancy can be identified (Gonzalez & Woods, 2002):

- **Spatial redundancy:** This refers to the correlation between neighboring pixels. This is the only redundancy for grayscale images.
- **Spectral redundancy:** This refers to the correlation between different color planes or spectral bands. This redundancy occurs in color images or multispectral images and exists together with the spatial redundancy.

Image compression techniques aim at reducing the number of bits needed to represent an image by removing the spatial and spectral redundancies as much as possible. The compression is lossless if the redundancy reduction does not result in any loss of information in the original image.

Besides redundancy, an image may also contain visually irrelevant information. The visually irrelevant information refers to information that is not perceived by human observers. Irrelevancy reduction thus aims at removing certain information in the image that is not noticeable by the Human Visual System (HVS). In general, some form of information loss is incurred when irrelevancy reduction is performed (Xiao, Wu, Wei, & Bao, 2005).

A number of standards have been established over the years for natural image compression. JPEG is the most common image file format that is found in existing Internet and multimedia systems (Pennebaker & Mitchell, 1993; Wallace, 1991). JPEG stands for Joint Photographic Experts Group. It is the name of the joint ISO/CCITT committee that created the image compression standard in 1992. There are two compression modes in the JPEG compression standard: lossless and lossy. However, the lossy mode dominates in almost all applications. The JPEG image compression codec has low complexity and is memory efficient. However, its

main criticism is the appearance of the blocking artifacts, especially at high compression ratios.

In 2001, the Joint Photographic Experts Group created another new image compression standard, called the JPEG 2000 (Taubman & Marcellin, 2002). This new standard provides an improved compression performance over JPEG and avoids the blocking artifacts completely. Besides the better compression performance, it also provides progressive capability in which the JPEG 2000 bitstream is organized in such a way that the image quality gets better progressively in terms of quality or resolution (Lee, 2005). Compared to the JPEG standard, JPEG 2000 is not widely supported at present. It is hindered by the fact that some of the algorithms are patented. As a result, it cannot be included in open-source Web browsers, which affects its popularity.

IMAGE COMPRESSION METHODOLOGY

Basic Compression Scheme

In general, most natural image compression schemes have a common structure, as shown in Figure 1. The first stage is usually a color space conversion module. Typically, a color image is stored in the RGB format. Because compression in the RGB domain is very inefficient, the image is converted into a luminance-chrominance color representation, that is, YCbCr, where the Y component represents the luminance information while the Cb and Cr components represent the color information. The image is then subjected to a transformation like the discrete cosine transform (DCT) (Ahmed, Natarajan, & Rao, 1974) or discrete wavelet transform (DWT) (Heil, Walnut, & Daubechies, 2006). The transformation decorrelates the image data, and thus reduces redundancy. The resulting coefficients are next quantized and entropy encoded. To quantize a signal means to describe it with less precision. Hence, some image information is inevitably discarded. Scalar quantization quantizes each coefficient separately using a predefined quantization table. It is the most common quantization scheme due to its simplicity. The rate-distortion (RD) unit controls the quantization step-size as a function of the bit-rate R and distortion D (Sarshar & Wu, 2007). Sometimes a RD unit is not explicitly defined, but is indirectly controlled by the nature of the quantizer.

Instead of quantizing each coefficient separately as in a scalar quantizer, vector quantization (VQ) can be used to represent a signal piecewisely by short vectors from a codebook. The codebook generally contains a limited number of entries that approximate the signal pieces. Compression is achieved in VQ because only the index of the best codebook entry needs to be encoded and transmitted.

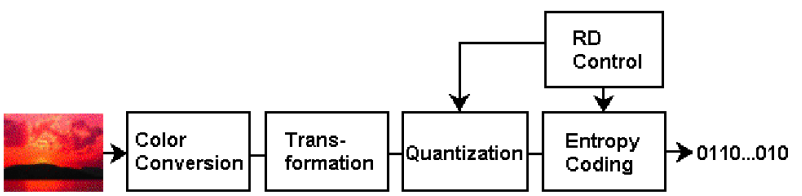
Exploiting the Limitations of the Human Visual System (HVS)

High compression ratio is usually achieved by aggressively exploiting the limitations of the HVS. Psycho-visual experiments have shown that the HVS has reduced sensitivity for patterns with high spatial frequencies. The phenomenon is parameterized by the contrast sensitivity function (CSF). Exploiting this behavior can significantly improve the compression ratio without incurring noticeable distortion. The quantization table in JPEG makes use of this phenomenon to some extent by using a large quantization step for high spatial frequency DCT transform coefficients in the luminance channel.

The sensitivity of the HVS for compression artifacts also varies with respect to the strength of local contrasts. Thus, an artifact might be hidden by the presence of strong contrasts or locally active image regions. This phenomenon, referred to as masking, is exploited in some sophisticated compression schemes (Gonzalez & Woods, 2002).

Subjective quality evaluations showed that the HVS is very sensitive to the loss of texture information. Blurred image with texture loss usually appear unnatural. However, the exact encoding of texture information is bit-rate intensive. To overcome this problem, a generative approach for texture region encoding is sometime employed in advance compression algorithms (Egger, Fleury, Ebrahimi, & Kunt, 1999; Ryan et. al. 1996). In this approach, the texture is characterized by only a few parameters that can be encoded for a modest increase in bit-rate. During decoding, the texture is synthe-

Figure 1. General structure of image compression algorithms



sized from these parameters. Even if the synthesized texture is pixel-wise different from the original texture, the HVS is nevertheless deceived due to their apparent similarity.

The HVS is unable to distinguish small differences in color as easily as it can to changes in brightness value. This phenomenon is exploited in chroma subsampling by dropping half or more of the chrominance information in the image (Poynton, 2003). At normal viewing distances, there is no perceptible loss incurred by sampling the color detail at a lower bit rate.

Compression Quality Evaluation

For lossy image compression, the image is not reproduced exactly. An approximation of the original image is enough for most purposes, as long as the error between the original and the compressed image is tolerable. Two type of quality evaluation are often used to assess the performance of different compression algorithms: (i) objective error metrics and (ii) subjective quality evaluation. Objective error metrics are quantitative and do not take into account the properties of the HVS, whereas subjective quality evaluation measures only the perceptually important distortion as determined by the HVS and is usually qualitative.

The Peak Signal to Noise Ratio (PSNR) is the most commonly used objective error metrics for measuring the quality of the reconstructed image. The PSNR is defined as:

$$PSNR = 10 * \log_{10}(MAX_I^2 / MSE),$$

where MAX_I is the maximum pixel value of the image (255 for 8 bit grayscale image). The MSE is the Mean Square Error between two images I and I', and is given by:

$$MSE = 1/MN \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i,j) - I'(i,j)]^2.$$

It is well known that PSNR does not correlate well with perceptual judgment of reconstruction quality.

The most common subjective quality evaluation is by visual inspection. Mean opinion score (MOS) can also be used to provide a numerical indication of the perceived quality of reconstruction quality (ITU-R Recommendation, 1992). The MOS averages the subjective evaluation of a group of viewers on different visual criteria based on a numerical score, for example, 1 = lowest perceived quality to 5 = highest perceived quality. Subjective quality evaluation is a difficult problem and currently there is no universally accepted method of performing subjective quality evaluation (Wei, Li, & Chen, 2006).

Post-Processing for Compression Artifacts Removal

Highly visible compression artifacts appear in the reconstructed image at high compression ratio. These artifacts are usually visually annoying to the viewers and a lot of research has been done to suppress them. Post-processing the reconstructed image for compression artifacts suppression has been a popular solution to this problem (Liew & Yan, 2004; Liew, Yan, & Law, 2005; Yarnatani & Saito, 2006).

For coding algorithms based on block-based DCT such as JPEG, the major artifacts are characterized by blockiness in flat areas and ringing along object edges. For wavelet-based coding techniques, ringing is the most visible artifact. Reduction of these artifacts can result in a significant improvement in the overall visual quality of the decoded images. The general approach is to reduce the block discontinuities along the 8×8 block boundaries by certain form of smoothing while preserving genuine edges in the image (Law & Siu, 2001). Alternatively, smoothing can be applied to region segments that exhibit slow intensity variation (Liew, Yan, & Law, 2005). Ringing suppression can be done by locally suppressing ringing ripples near the vicinity of strong edges (Liew & Yan, 2004).

Compression Choices

Some of the important choice/considerations for image compression are as follows:

- **Lossless compression:** For some applications, lossless compression is required. In this case the reconstructed image will be exactly equal to the original image. Lossless compression is usually used for medical applications. The best compression factor that can be achieved is typically around three.
- **Visually lossless compression:** If a human observer cannot see any difference between the original and the compressed image, the compression is visually lossless. Most compression algorithms exploit the properties of the human visual system (HVS) to achieve high compression factor, that is, up to 20 or 30, without noticeable degradation of the perceived image. This is done by discarding part of the image information that is not perceivable by the HVS.
- **Scalable or progressive coding:** A nonprogressively encoded bitstream must be received in its entirety before the inverse transform can be applied. However, for a progressively encoded bitstream, the inverse transform can be applied to a partially decoded bitstream. Progressive encoding arranges the bitstream in such a way that the most important information is near the

front end of the bitstream and decreasingly important information is toward the back of the bitstream. During decoding, the critical information at the front of the bitstream can be used to construct an approximate version of the image and the quality or resolution of the reconstructed image is progressively increased when further bitstream is received and decoded.

CURRENT IMAGE COMPRESSION STANDARDS

JPEG

The lossy mode in the JPEG image compression standard is widely used in WWW and multimedia systems. First, the image is converted from the RGB color space to the YCbCr color space. Because the human eye is relatively insensitive to the chrominance information, the Cb and Cr components are usually downsampled. Next, each of the Y, Cb and Cr components are grouped into 8×8 blocks. Discrete cosine transform (DCT) is then applied independently to each block to get the DCT coefficients. This is followed by the process of quantization in which each DCT coefficient is divided by a number and rounded to the nearest integer. Because humans are not good at seeing differences in high frequency components as compared to the low frequency components, the DCT coefficients are quantized differently: the divisor for the low frequency DCT coefficients is much smaller than that for the high frequency DCT coefficients. Many high frequency components are rounded to zero after quantization. "Zig-Zag" scanning then organized the two dimensional DCT coefficients block into a one dimensional coefficient stream so that similar frequency components are grouped together. Run-length or Huffman coding can then be applied to further compress the resulting bitstream.

The image quality depends greatly on the choice of the divisors in the quantization process. Good image quality is often obtained for a compression ratio of less than 10. For compression ratio larger than 100, images appear to be blocky due to the fact that quantization is applied independently to each of the 8×8 blocks.

The JPEG standard also has a lossless mode. Lossless JPEG was developed as a late addition to JPEG in 1993, using a completely different technique from the lossy JPEG standard. It uses a predictive scheme based on the three nearest neighbors (upper, left, and upper-left), and entropy coding is used on the prediction error. It was never widely adopted and its performance is not state-of-the-art.

JPEG-LS

JPEG-LS is the new lossless/near-lossless compression standard for continuous-tone images, ISO-14495-1/ITU-T.87. JPEG-LS was developed with the aim of providing a low complexity lossless image compression standard that could be able to offer better compression efficiency than lossless JPEG. Part 1 of this standard was finalized in 1999. The standard is based on the LOCO-I algorithm (LOw COmplexity LOssless COmpression for Images) developed at Hewlett-Packard Laboratories, that relies on prediction, residual modeling and context-based coding of the residuals (Weinberger, Seroussi, & Sapiro, 2000). Most of the low complexity of this technique comes from the assumption that prediction residuals follow a two-sided geometric distribution (also called a discrete Laplace distribution) and from the use of Golomb-like codes, which are known to be near-optimal for geometric distributions. Besides lossless compression, JPEG-LS also provides a lossy mode where the maximum absolute error can be controlled by the encoder. Compression for JPEG-LS is generally much faster than JPEG 2000 and much better than the original lossless JPEG standard.

JPEG 2000

JPEG 2000 is another newly developed image compression standard, which provides better image quality and stronger compression power than JPEG. Instead of using the block-based DCT, JPEG 2000 uses the wavelet transform so that the blocking artifacts can be avoided completely. Similar to the JPEG standard, an image is first converted from the RGB color space to YCbCr or the reversible component transform (RCT) color space. The image is then split into rectangular regions called tiles so that wavelet transform can be applied to each tile independently with different decomposition levels. The JPEG 2000 standard uses the floating-point biorthogonal 9/7 wavelet kernel for lossy compression and integer 3/5 kernel for lossless compression.

The wavelet coefficients are organized into a set of subbands that shows spatial details under a certain frequency range. These coefficients are then quantized to produce a set of integer numbers. The quantized subbands are split into rectangular regions in wavelet domain called precincts. Precincts are further split into codeblocks in which the Embedded Block Coding with Optimal Truncation scheme is employed to each code block in a biplane order. There are three coding passes: significant propagation, magnitude refinement and clean up passes. The significant propagation encodes bits of insignificant coefficients with significant neighbors while the magnitude refinement refines the significant coefficients.

A context-driven binary arithmetic coding is then used to code the bits after the three coding passes. The resultant

bit stream is organized into packets where a packet groups selected passes of all code blocks from a precinct into one unit. These packets can be organized in a flexible way depending on the target application. For example, these packets from all subbands can be arranged into layers in such a way that the image quality gets improved progressively from layer to layer.

Because of the inherent multiresolution nature of wavelets and the encoding schemes, the image compressed by the JPEG 2000 standard has bitstreams arranged progressively, from very low image quality (i.e., small number of bits and high compression ratio) to effectively lossless image compression. Often compression ratio of 20 can be achieved for natural images without any visible artifacts. In JPEG 2000, blocking artifacts are totally removed but smoothing and ringing artifacts are introduced at high compression ratio.

JPEG 2000 includes a lossless mode based on a special integer wavelet filter (biorthogonal 3/5). JPEG 2000's lossless mode runs more slowly and usually has worse compression ratios than JPEG-LS. However, it is scalable and progressive, and, because its algorithm is similar to JPEG 2000, it is more widely supported.

NEW FEATURES IN IMAGE COMPRESSION

The aim of the JPEG 2000 is not only to have an improved compression performance as compared to the JPEG standard, but also to introduce some added features which are valuable to various kinds of applications. New features introduced include region of interesting coding, random access and progressive transmission concept (Lee, 2005).

Region of interest coding implies that certain parts of the image can be encoded with higher (or lossless) quality than other regions. This feature is important for applications such as medical imaging where certain critical parts can be coded with very high quality while other less critical parts are coded with low resolution. JPEG 2000 also introduces the idea of random codestream access. This means that the compressed image bitstream support random spatial access at varying resolutions.

JPEG 2000 induces a scalable structure on image representation. This scalable structure enables the generation of a single bit-stream for different purposes without having to rerun the compression algorithm. For example, the quality or the image resolution/size can be progressively improved. Thus, the same bit-stream can be used in a High Definition Television with a 1280×720 display, as well as in a PDA with a 160×160 display.

FUTURE TRENDS

Next generation compression algorithms are characterized by the use of an object-oriented image representation and a single universal bit-stream for different transmission/display media. The object-oriented compression enables the manipulation of objects in the scene as individual items that can be processed separately. The single universal bit-stream allows the compressed data to be transmitted or displayed by different media without having to rerun the compression algorithm. The JPEG 2000 bitstream enables region of interest coding for objects and are scalable due to the use of wavelet transforms. Its popularity, however, is hindered by the software patents. Either the standards need to be modified to avoid those patented software or more researches have to be done to replace these software with other free license software. Otherwise, the open-source Web browsers cannot include the JPEG 2000 decoder.

CONCLUSION

Image compression is a core technology for multimedia systems. Efficient transmission and storage of images is often required despite the continuing increase in network bandwidth. Transform coding is currently the most popular class of compression method for natural images. To achieve high compression ratio, lossy compression that exploits the limitation of the HVS is often employed. There are a number of international standards in image compression. The JPEG standard was established in 1992 and is very popular in WWW and multimedia systems. However, it is well known that blocking artifacts would appear at low bit-rates, which could greatly affect the perceived image quality. For near lossless compression applications, the JPEG-LS standard was introduced in 1999. Recently, the JPEG 2000 standard was introduced. It not only removes the blocking artifacts completely and improves the compression performance over the JPEG standard, but also induces a scalable structure on image representation. This scalable structure fits well with the perception of human eyes. It enables the generation of a single bit-stream for different transmission/display media, and potentially induces an object-oriented image representation. However, because of the software patent issue, the JPEG 2000 has not been included in open-source Web browsers. This affects its popularity.

REFERENCES

Ahmed, N., Natarajan, T., & Rao, K.R. (1974). Discrete cosine transform. *IEEE Transactions on Computers, 23,* 90-93.

Egger, O., Fleury, P., Ebrahimi, T., & Kunt, M. (1999). High performance compression of visual information, a tutorial review part I: Still picture. In *Proceedings of the IEEE, 87*(6), 976-1013.

Gonzalez, R.C., & Woods, R.E. (2002). *Digital image processing*. Prentice Hall.

Hei. C., Walnut, D.F., & Daubechies, I. (2006). *Fundamental papers in wavelet theory*. USA: Princeton University Press.

ITU-R Recommendation. (1992). *Method for the subjective assessment of the quality of television pictures* (pp. 500-505).

Law, N.F., & Siu, W.C. (2001). Successive structural analysis using wavelet transform for blocking artifacts suppression. *Signal Processing, 81*(7), 1373-1387.

Lee, D.T. (2005). JPEG 2000: Retrospective and new developments. In *Proceedings of the IEEE, 93*(1), 32-41.

Liew, A.W.C., & Yan, H. (2004). Blocking artifacts suppression in block-coded images using overcomplete wavelet representation. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(4), 450-461.

Liew, A.W.C., Yan, H., & Law, N.F. (2005). POCS-based blocking artifacts suppression using a smoothness constraint set with explicit region modeling. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(6), 795-800.

Pennebaker W.B., & Mitchell, J.L. (1993). *JPEG: Still image data compression*. Van Nostrand Reinhold.

Poynton, C. (2003). *Digital video and HDTV: Algorithms and interfaces*. USA: Morgan Kaufmann.

Ryan, T.W., Sanders, D, Fishers, H.D., & Iverson, A.E. (1996). Image compression by texture modeling in the wavelet domain. *IEEE Transactions on Image Processing, 5*(1), 26-36.

Sarshar, N., & Wu, X. (2007). On rate-distortion models for natural images and wavelet coding performance. *IEEE Transactions on Image Processing, 16*(5), 1383-94.

Shannon, C.E. (1948). A mathematical theory of communication. Bell System. *The Bell System Technical Journal, 27*, 379-423.

Taubman, D.S., & Marcellin, M.W. (2002). *JPEG 2000: Image compression foundations, standards and practice*. Kluwer Academic.

Wallace, G. (1991). The JPEG still picture compression standard. *Communications of the ACM, 34*(4), 30-44.

Weinberger, M., Seroussi, G., & Sapiro, G. (2000). The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. *IEEE Transactions on Image Processing, 9*(8), 1309-1324.

Wei, X., Li, J., & Chen, G. (2006). An image quality estimation model based on HVS. In *Proceedings of the 2006 IEEE Region 10 Conference,* (pp. 1-4).

Xiao, L., Wu, H.Z., Wei, Z.H., & Bao, Y. (2005). Research and application of a new computational model of human vision system based on Ridgelet transform. In *Proceedings of the 2005 International Conference on Machine Learning and Cybernetics*, (Vol. 8, pp. 5170-5175).

Yarnatani, K., & Saito, N. (2006). Improvement of DCT-based compression algorithms using Poisson's equation. *IEEE Transactions on Image Processing, 15*(12), 3672-3689.

KEY TERMS

Blocking Artifacts: This is one of the artifacts often exhibited by JPEG standard at high compression ratios. Images appear to have regular block structures.

JPEG: An image compression standard proposed in 1992 by ISO/CCITT committee. It is one of the most common image file format that is found in Internet and consumer products.

JPEG 2000: An image compression standard that aims not only to provide an improved compression performance over JPEG, but also to provide new features such as region of interest coding, random access and progressive coding.

Progressive Transmission: This implies that the bitstream is arranged so that most important information is near the front end of the bitstream and the least important information is at the back of the bitstream. Thus, in decoding, the quality of the decoded image is progressively increased.

Region of Interest Coding: This means that certain parts of the image (i.e., the interested regions) are encoded with more bits and thus have better quality than other parts of the image.

Ringing Artifacts: This type of artifacts often appears near the edges of an image in which edges are blurred and have oscillating effect.

Scalability: It refers to a successive quality change by bitstream manipulation. For example, PSNR scalability means the PSNR improves as more bits in the bitstream are decoded.

Significant Wavelet Coefficients: This refers to wavelet coefficients that have large absolute magnitude. Usually, this implies important structural information such as edges in an image.

Transform Coding: This refers to a type of compression in which the image data is first transformed into another domain so that the data becomes uncorrelated in this new domain for further processing.

Wavelet Subband: This refers to a group of the wavelet coefficients at certain frequency ranges.

Category: Multimedia Technology

Image Segmentation Evaluation in this Century

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

Image segmentation consists of subdividing an image into its constituent parts and extracting those parts of interest (objects). Due to its importance in image analysis, many research works have been conducted for this process. After 40 years of development, a large number of image (and video) segmentation techniques have been proposed and utilized in various applications (Zhang, 2006). With many algorithms developed, some efforts have been spent also on their evaluation, and these efforts have resulted around 100 evaluation papers that can be found in literature for the last century. Several studies have been made in the past in attempt to characterize these existing evaluation methods (Zhang, 1993; Zhang, 1996; Zhang 2001).

Segmentation evaluation methods can be classified into analytical methods and empirical methods (Zhang, 1996). The analysis methods treat the algorithms for segmentation directly by examining the principle of algorithms while the empirical methods judge the segmented image (according to predefined criteria or comparing to reference image) so as to indirectly assess the performance of algorithms.

Empirical evaluation is practically more effective and usable than analysis evaluation (Zhang, 1996). Recent advancements for segmentation evaluation are mainly made by the development of empirical evaluation techniques. After providing a list of evaluation criteria and methods proposed in the last century as background, this article will provide a summary of the recent (in 21st century) research works for empirical evaluation of image segmentation. These new research works are classified into three groups: (1) those based on existing techniques, (2) those made with modifications of existing techniques, and (3) those that used dissimilar ideas than that of existing techniques. A comparison of these evaluation methods is made before going to the future trends and conclusion.

BACKGROUND

Empirical evaluation methods can be classified into *goodness method group* and *discrepancy method group* (Zhang, 1996). They use different empirical criteria for judging the performance of segmentation algorithms. The goodness method can perform the evaluation without the help of reference images while the discrepancy method needs some reference images to arbitrate the quality of segmentation. In Zhang (1996) the eight mostly used criteria (three goodness ones and five discrepancy ones) have been discussed in details. All these criteria have been grouped into a table in Zhang (2001, pp. 148-151), and the table is reproduced in Table 1.

There are other criteria discussed in Zhang (1996), though they were not very well liked in that time. One is the number of regions in a segmented image. In case no ground truth was available, it would be expected to get a modest result, so the *moderate number of regions* could be counted as a criterion. Some others come from the class D-5, such as *region consistency*, *grey level difference*, and *symmetric divergence (cross-entropy)*. Finally, several criteria used in special methods have attracted certain attention recently, such as *amount of editing operations, visual inspection,* and *correlation between original image and segmented bi-level image.* All these criteria are listed now in Table 2 as a complementary of Table 1.

MAIN FOCUS OF THE CHAPTER

Getting into the new century, the research on segmentation evaluation has attracted even more attention in the segmentation community. In the following section, some evaluation works published since 2002, that is, after the last review paper on evaluation (Zhang, 2001), are sketched and discussed. Among these new empirical evaluation works, some are based on existing techniques, some are made with modifications/improvements of existing techniques, and some have dissimilar ideas than that of existing techniques.

Evaluation Works Based on Existing Techniques

In Cavallaro, Gelasca, and Ebrahimi (2002), a single objective metric is formed by using both spatial and temporal consistency information. The metric was defined based on two types of errors. One is the number of (both positive and negative) false pixels. Another is the distance of false pixels to their correct places. The spatial context was introduced to weight the false pixels according to their distance to the

Table 1. A list of empirical criteria and their method groups

Class	Criterion name	Method group
G-1	Intra-region uniformity	Goodness
G-2	Inter-region contrast	Goodness
G-3	Region shape	Goodness
D-1	Number of mis-segmented pixels	Discrepancy
D-2	Position of mis-segmented pixels	Discrepancy
D-3	Number of objects in the image	Discrepancy
D-4	Feature values of segmented objects	Discrepancy
D-5	Miscellaneous quantities	Discrepancy

Table 2. A complementary list of empirical criteria and their method groups

Class	Criterion name	Method group
G-4	Moderate number of regions	Goodness
D-5a	Region consistency	Discrepancy
D-5b	Grey level difference	Discrepancy
D-5c	Symmetric divergence (cross-entropy)	Discrepancy
S1	Amount of editing operations	Special
S2	Visual inspection	Discrepancy like
S3	Correlation between original image and bi-level image	Goodness like

reference boundary. In addition, temporal context has been used to assign weight inversely proportional to the duration of an error for evaluating the quality variation over time. The overall metric was eventually formulated as nonlinear combination of the number of false pixels and the distances, weighted by the temporal context factor.

In Prati, Mikic, and Trivedi (2003), a comparative empirical evaluation of representative segmentation algorithms selected from four classes of techniques (two statistical ones and two deterministic ones) for detecting moving shadows has been made with a benchmark for indoor and outdoor video sequences. Two quantitative metrics: (1) good detection (low probability of misclassifying a shadow point) and (2) good discrimination (the low probability of classifying non-shadow points as shadow) are employed.

In Rosin and Ioannidis (2003), an evaluation of eight different threshold algorithms for shot change detection in a surveillance video has been made. Pixel-based evaluation is applied by using true positive (TP), true negative (TN), false positive (FP), and false negatives (FN).

In Lievers and Pilkey (2004), a comparison of 12 automatic global thresholding methods has been made. Among them, eight are point-dependent algorithms and four are region-dependent algorithms. Some multimodal images have been tested. Authors defined a cost function for selecting the appropriate thresholds. This cost function is based on intra-class variations, so it is not surprising that the best algorithm found by authors is a minimum cross-entropy method.

In Marcello, Marques, and Eugenio (2004), a survey of 36 image thresholding methods, with a view to assess their performance when applied to remote sensing images and especially in oceanographic applications, has been conducted. Those algorithms have been categorized into two groups: local and global thresholding techniques. For performance judgment, only visual inspection is carried out.

In Renno, Orwell, and Jones (2004), four different shadow suppression algorithms have been evaluated by using video from a nightly soccer match with quite some shadow because of the lighting used. All evaluation metrics are based on the number of correctly detected pixels. The metrics used are the detection rate, the false positive rate, the signal-to-noise ratio, and the tracking error (the average distance between ground truth boxes and tracked targets). Finally, using an average over time, the performances of shadow segmentation of the four shadow suppression algorithms are compared.

In Carleer, Debeir, and Wolff (2004), four algorithms were applied to high spatial resolution satellite images and their performances were compared. Two empirical discrepancy evaluation criteria are used: (1) the number of mis-segmented pixels in the segmented images compared with the visually segmented reference images, and (2) the ratio between the number of regions in the segmented image and the number of regions in the reference image.

In Ladak, Ding, and Wang (2004), a comparison of three kinds of segmentation algorithms for 3-D images: (1) segmenting parallel 2-D slice images, (2) segmenting rotated 2-D slice images, and (3) directly segmenting volume-based 3-D image, was carried out. The judging parameter used is the percent difference in volume (volume error) between automatically segmented objects and the manually determined (by a trained person) ground truth. The times needed for editing the segmented objects obtained by using the three kinds of algorithms to fit the ground truth are also compared.

Evaluation Works Made with Modifications/Improvements

In Oberti, Stringa, and Vernazza (2001), Receiver Operating Curve (ROC) is used to extract useful information about the segmentation performance when changing external parameters that describe the conditions of the scene. ROC has also been used in Niemeijer, Staal, and Ginneken (2004) for studying the performance of five vessel segmentation algorithms.

In Udupa, LeBlanc, and Schmidt (2002), three groups of factors: (1) precision, (2) accuracy, and (3) efficiency, are considered for evaluating segmentation methods in assessing

how good a segmentation method is and in comparing it with other methods. Precision factors assess the reliability of the method; accuracy factors describe the validity of the method; and efficiency factors determine the human operator time and computational time required to complete segmentation task. To characterize the range of behavior of a segmentation method from accuracy point of view, a Delineation Operating Characteristic (DOC) analysis is proposed, and the DOC curves have been used for comparing the accuracy of three segmentation methods: (1) thresholding, (2) fuzzy connectedness, and (3) fuzzy c-means (Udupa & Zhuge, 2004).

In Li, Li, and Chen (2003), four metrics have been used for evaluation of video segmentation. They are metrics for contour-based spatial matching, temporal consistency, user workload, and time consumption. The first two metrics can be considered as combined extensions of the number of mis-segmented pixels and the position of mis-segmented pixels. The last two metrics are somewhat closely interrelated.

In Kim, Chalidabhongse, and Harwood (2004), a Perturbation Detection Rate (PDR) analysis has been proposed. It measures the sensitivity of a background subtraction algorithm in detecting low contrast targets against background. Four background subtraction algorithms are evaluated for their segmentation performance.

In Erdem, Sankur, and Tekalp (2004), three goodness measures to evaluate quantitatively the performance of video object segmentation and tracking methods have been proposed. The spatial differences of color and motion along the boundary of the estimated video object plane and the temporal differences between the color histogram of the current object plane and its predecessors are used to localize (spatially and/or temporally) regions where the segmentation results are good or bad. They are further combined to yield a single numerical measure to indicate the goodness of the boundary segmentation and tracking results over a video sequence.

Evaluation Works Supplying New Inspiration

One of the major focuses in the evaluation of segmentation in this century is in combining different metrics. Though the earliest work of this kind was made a quarter of century ago (Zhang, 1996), new research in this direction still attracts attention, only they have used different strategies.

In Everingham, Muller, and Thomas (2002), an approach to formulate the evaluation problem as finding out the Pareto front in a multi-dimensional fitness space is proposed.

In Li et al. (2003), the aforementioned idea has been followed. One 4-D fitness space has been built for four different evaluation metrics. A search in multi-dimensional space is performed to find the best choice of a system with optimal parameters.

In Zhang, Fritts, and Goldman (2005), another cooperation framework is proposed, in which different effectiveness measures are combined by using a machine-learning approach, which combines the results obtained from different measures. Three strategies based on training for combinations are used: (1) weighted majority (WM), (2) Bayesian, and (3) support vector machine (SVM).

In Correia and Pereira (2003), they considered that the evaluation of video segmentation quality could have two targets: (1) individual object segmentation quality evaluation, in which single object identified by the segmentation algorithm is independently evaluated in terms of its segmentation quality, and (2) overall segmentation quality evaluation, in which the complete set of objects (for the whole scene) identified by the segmentation algorithm is globally evaluated in terms of its segmentation quality.

In Desurmont, Wijnhoven, and Jaspers (2005), an outlined framework for performance evaluation of video content analysis (VCA) system is proposed. Four main components of this framework are: (1) creation of ground-truth (GT) data, (2) available evaluation data sets, (3) performance metrics, and (4) presentation of the evaluation results.

Classes of Criteria

In this section, the criteria used in the aforementioned works are analyzed and the results are summarized for each group of methods by using existing techniques, with modified criteria and with novelties.

Table 3 gives the list of evaluation works using existing techniques. Most works are based on discrepancy criteria, in which the criteria belonging to class D-1 appears more frequently than others do. Still few works use goodness criteria, but most of them use discrepancy criteria.

Table 4 gives the list of evaluation works with some modifications to existing techniques. Most of these works are still based on discrepancy criteria, in particular on class D-1. Several works made the use of the combination of criteria from different classes. The newly defined class S-1 has also been used by two works.

Table 5 gives the list of evaluation works with some novelties. In general, these approaches are quite divergent, one can only reveal that all the first three take the procedure of combining different criteria for evaluation.

Comparison of Methods

To compare different methods for segmentation evaluation, the following four factors have been considered, taking into consideration of the techniques and criteria used in evaluation (Zhang, 1993; Zhang, 2001): (1) generality for evaluation; (2) subjective versus objective and qualitative versus quantitative; (3) complexity for evaluation; and (4) evaluation requirements for reference images.

Table 3. Evaluation methods using existing techniques

Method #	Source	Criteria used	Method #	Source	Criteria used
M-1	(Cavallaro et al., 2002)	D-1, D-2	M-5	(Marcello et al., 2004)	S-2
M-2	(Prati et al., 2003)	D-1	M-6	(Renno et al., 2004)	D-1, D-4
M-3	(Rosin & Ioannidis, 2003)	D-1	M-7	(Carleer et al., 2004)	D-1, D-3
M-4	(Lievers & Pilkey, 2004)	G-1	M-8	(Ladak et al., 2004)	D-1, S-1

Table 4. Evaluation methods with modified criteria

Method #	Source	Criteria used (modification)
M-9	(Oberti et al., 2001)	D-1 (ROC, curve of FP vs. FN)
M-10	(Udupa et al., 2002)	D-1, S-1 like (efficiency)
M-11	(Li et al., 2003)	D-1 and D-2, (contour matching, temporal consistency), S-1
M-12	(Erdem et al., 2004)	G-1, G-2 (with extension to color, motion, color histograms)
M-13	(Niemeijer et al., 2004)	D-1 (ROC, curve of TP vs. FP)
M-14	(Udupa & Zhuge, 2004)	D-1 (DOC, curve of TP vs. FP)
M-15	(Kim et al., 2004)	D-1 (PDR, modified detection rate)

Table 5. Some evaluation works with novelties

Work #	Source	Novelty
W-1	(Everingham et al., 2002)	Finding out the Pareto front in a multi-dimensional fitness space
W-2	(Li et al., 2003)	Finding out the Pareto front in a 4-D fitness space
W-3	(Zhang et al., 2005)	Using WM, Bayesian, and SVM
W-4	(Correia & Pereira, 2003)	Using contextual relevance metric to match human visual system (HVS)
W-5	(Desurmont et al., 2005)	Performing evaluation in different semantic levels

Table 6. A comparison of methods listed in Table 3 and Table 4

Method #	Generality	Complexity	Method #	Generality	Complexity
M-1	Video [1]	Medium/High	M-9	General	Medium
M-2	General	High	M-10	General	Medium
M-3	Video [1]	Medium	M-11	General	High (Human) [3]
M-4	Thresholding [2]	Medium	M-12	Video [1]	High
M-5	General	High (Human) [3]	M-13	General	Medium
M-6	General	Medium/High	M-14	General	Medium
M-7	Numerous objects [4]	Low/Medium	M-15	Video [1]	Medium
M-8	General	High (Human) [3]			

[1]*Only usable for video segmentation evaluation (mostly because temporal factor is critical).*
[2]*Only suitable for evaluating thresholding technique.*
[3]*Human visual factors are involved, so the complexity becomes high.*
[4]*More appropriate for treating images composed of numerous object regions.*

Among these four factors, some of them are related to the method groups. For example, most empirical criteria provide quantitative results. On the other side, the subjective versus objective and the consideration of segmentation application are closely related and can be determined according to the criteria which belong either to goodness group or discrepancy group.

As mentioned in the beginning, only empirical methods are compared here, so the focus will be put on the generality for evaluation and the complexity for evaluation. Thus,

obtained comparison results for the methods listed in Table 3 and Table 4 are given in Table 6.

FUTURE TRENDS

Following the recent evolution of segmentation evaluation as described previously and by analyzing and comparing the methods and criteria used in these works, it seems two further research directions should be considered:

1. **To make evaluation in reflecting the final goal of segmentation:**
   Image segmentation is the first step in image analysis; its performance will influence all the subsequent processes. Therefore, the purpose of segmentation, especially the goals of analysis task, should be considered in segmentation evaluation. From one side, the consideration of the final goal of segmentation would make the evaluation procedure more objective and the results more useful. From other side, the success of the following processes would be a suitable indication of the success of image segmentation.
2. **To efficiently combine multiple metrics in evaluation:**
   The results of segmentation and the performance of segmentation algorithms are related to many factors, so it is not surprising that the strategy of using several criteria to form multiple metrics for evaluation was considered a long time ago, and the recent works on evaluation also show this tendency as indicated before. The tactic for combining multiple metrics has evolved from simply making weighted sums to complicated techniques, such as matching learning approaches. Further works to combine multiple metrics to effectively cover different aspects of evaluation in wide range of applications are still needed.

CONCLUSION

According to the aforementioned investigation with more than 20 works for less than 4 years, it is clear that much more efforts have been put on segmentation evaluation in the beginning of 21st century than in the last century. We have seen a number of works based on previous proposed principles; a number of works made modifications and/or improvements on previous proposed techniques; and also several works with new ideas have appeared.

The generality and complexity of these newly proposed evaluation methods and performance criteria have been thoroughly compared. It seems that though much more efforts have been put in this subject, no (or very few) radical progress has been reported. Many evaluation works were just using previously proposed methods and criteria for particular application; some evaluation works made certain improvements on early works but still used similar principles. To probe further, continuing efforts for segmentation evaluation are still called for.

ACKNOWLEDGMENTS

This work has been supported by the Grants NNSF-60573148 and SRFDP-20050003013.

REFERENCES

Carleer, A. P., Debeir, O., & Wolff, E. (2004). Comparison of very high spatial resolution satellite image segmentations. *SPIE, 5238,* 532-542.

Cavallaro, A., Gelasca, E. D., & Ebrahimi, T. (2002). Objective evaluation of segmentation quality using spatio-temporal context. *Proceedings of the International Conference on Image Processing* (Vol. 3, pp. 301-304).

Correia, P. L., & Pereira, F. M. B. (2003). Objective evaluation of video segmentation quality. *IEEE Transaction on Image Processing, 12*(2), 186-200.

Desurmont, X., Wijnhoven, R., Jaspers, E., Olivier, C., Mike, B., Wouter, F., et al. (2005). Performance evaluation of real-time video content analysis systems in the CANDELA project. *SPIE, 5671,* 200-211.

Erdem, C. E., Sankur, B., & Tekalp, A. M. (2004). Performance measures for video object segmentation and tracking. *IEEE Transaction on Image Processing, 13*(7), 937-951.

Everingham, M., Muller, H., & Thomas, B. (2002). Evaluating image segmentation algorithms using the Pareto Front. *Proceedings of the European Conference on Computer Vision,* (Vol. 4, pp. 34-48).

Kim, K., Chalidabhongse, T. H., Harwood, D., & Davis, L. (2004). Background modeling andsubtraction by codebook construction. *Proceedings International Conference on Image Processing*, (Vol. 5, pp. 3061-3064).

Ladak, H. M., Ding, M., Wang, Y., Hu, N., Downey, D. B., & Fenster, A. (2004). Evaluation of algorithms for segmentation of the prostate boundary from 3D ultrasound images. *SPIE, 5370,* 1403-1410.

Li, N., Li, S., & Chen, C. (2003). Multimetric evaluation protocol for user-assisted video object extraction systems. *SPIE, 5150,* 20-28.

Lievers, W. B., & Pilkey, A. K. (2004). An evaluation of global thresholding techniques for the automatic image segmentation of automotive aluminum sheet alloys. *Materials Science and Engineering: A381*(1-2), 134-142.

Marcello, J., Marques, F., & Eugenio, F. (2004). Evaluation of thresholding techniques applied to oceanographic remote sensing imagery. *SPIE, 5573,* 96-103.

Niemeijer, M., Staal, J., Ginneken, B., Loog, M., & Abramoff, M. D.(2004). Comparative study of retinal vessel segmentation methods on a new publicly available database. *SPIE, 5370,* 648-656.

Oberti, F., Stringa, E., & Vernazza, G. (2001). Performance evaluation criterion for characterizing video-surveillance systems. *Real-Time Imaging, 7*(5), 457-471.

Prati, A., Mikic, I., Trivedi, M. M., & Cucchiara, R. (2003). Detecting moving shadows: Algorithms and evaluation. *IEEE Transaction on Pattern Aanalysis and Machine Intelligence, 25*(7), 918-923.

Renno, J. R., Orwell, J., & Jones, G. A. (2004). Evaluation of shadow classification techniques for object detection and tracking. In *Proceedings of International Conference on Image Processing* (pp. 143-146).

Rosin, P. L., & Ioannidis, E. (2003). Evaluation of global image thresholding for change detection. *Pattern Recognition Letters, 24*(14), 2345-2356.

Udupa, J. K., LeBlanc, V. R., Schmidt, H., Imielinska, C., Saha, P. K., Grevera, G. J., et al. (2002). A methodology for evaluating image segmentation algorithms. *SPIE, 4684,* 266-277.

Udupa, J., & Zhuge, Y. (2004). Delineation operating characteristic (DOC) curve for assessing the accuracy behavior of image segmentation algorithms. *SPIE, 5370,* 640-647.

Zhang, Y. J. (1993). Comparison of segmentation evaluation criteria. In *Proceedings of the 2nd International Conference on Signal Processing* (Vol. 1, pp. 870-873).

Zhang, Y. J. (1996). A survey on evaluation methods for image segmentation. *Pattern Recognition, 29*(8), 1335-1346.

Zhang, Y. J. (2001). A review of recent evaluation methods for image segmentation. *Proceedings of the 6th International Symposium on Signal Processing and Its Applications* (pp. 148-151).

Zhang, Y. J. (2006). *Advances in image and video segmentation*. Hershey, PA: IRM Press.

Zhang, H., Fritts, J. E., & Goldman, S. A. (2005). A co-evaluation framework for improving segmentation evaluation. *SPIE, 5809,* 420-430.

KEY TERMS

Composite Criteria: Criteria formed by combining several performance metrics in order to better cover the various aspects of the algorithms in segmentation. The combination can be made in different ways, such as by linear combination, by machine learning approach, etc. and so forth.

Evaluation Criteria: Criteria used in evaluation process to judge the performance of segmentation algorithms under consideration. They are also called performance metrics, performance measures, or performance indices.

Image Segmentation: A process consists of subdividing an image into its constituent parts and extracting these parts of interest (objects) from the image. It is a fundamental step and a critical task in image analysis.

Objective Criteria: Criteria based on objectively determined quantities or values, which indicate the difference between the segmented images and reference images. They are mostly used in empirical discrepancy methods for segmentation evaluation.

Segmentation Characterization: Segmentation characterization is an intra-algorithm process of segmentation evaluation. The purpose of evaluation for a specific algorithm is to quantitatively recognize its behavior in treating various images and/or to help appropriately setting its parameters regarding different applications to achieve the best performance of this algorithm.

Segmentation Comparison: Segmentation comparison is an inter-algorithm process of segmentation evaluation. The purpose of comparison for different algorithms is to rank their performance and to provide guidelines in choosing suitable algorithms according to applications as well as to promote new developments by effectively taking the strong points of several algorithms.

Segmentation Evaluation: Segmentation evaluation is a process used to judge the performance of segmentation algorithms based on some defined quality criteria and/or ground truth in view to assess or reveal the property of algorithms in use.

Subjective Criteria: Criteria based on human judgment or perception, which reflect some desirable properties of segmented images. They are used in empirical goodness methods for segmentation evaluation.

Category: Multimedia Technology

Image Segmentation in the Last 40 Years

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

Image segmentation is an important image technique well known by its utility and complexity. To extract the useful information from images or groups of images, an inevitable step is to separate the objects from the background. Segmentation is just the right process and technique required for this task. Image segmentation is often described as the process that subdivides an image into its constituent parts and extracts those parts of interest (objects). It is one of the most critical tasks in automatic image analysis, which is at the middle layer of image engineering. Image engineering (which is composed of three layers from bottom to top: (1) image processing, (2) image analysis, and (3) image understanding) is a new discipline and a general framework for all image techniques (Zhang, forthcoming).

The history of segmentation of digital images using computers can be traced back to 40 years ago. In 1965, an operator for detecting the edges between different parts of an image, Roberts operator (also called Roberts edge detector), was introduced and used for partition of image components (Roberts, 1965). Since then, the field of image segmentation has evolved very quickly and has undergone great change (Zhang, 2001a). In this article, after an introduction and explanation of the formal definition of image segmentation as well as three levels of research on image segmentation, the statistics for the number of developed algorithms in these years are provided; the scheme for classifying different segmentation algorithms is discussed; and a summary of existing survey papers for image segmentation is presented. All these discussions provide a general picture of research and development of image segmentation in the last 40 years.

BACKGROUND

Formal Definition of Image Segmentation

A formal definition of image segmentation, supposing the whole image is represented by R and R_i, $i = 1, 2, \ldots, n$ are disjoint nonempty regions of R, consists of the following conditions (Fu & Mui, 1981):

$$\bigcup_{i=1}^{n} R_i = R; \quad (1)$$

For all i and j, $i \neq j$, there exits $R_i \cap R_j = \varnothing$; (2)

For $i = 1, 2, \ldots, n$, it must have $P(R_i) = TRUE$; (3)

For all $i \neq j$, there exits $P(R_i \cup R_j) = FALSE$; (4)

where $P(R_i)$ is a uniformity predicate for all elements in set R_i and $\varnothing$ represents an empty set.

The following condition is also important for segmentation and is often included in the conditions for the formal definition (Zhang 2001a):

For all $i = 1, 2, \ldots, n$, R_i is a connected component. (5)

In the aforementioned conditions, each of them has particular meanings. The condition (1) points out that the union of segmented regions could include all pixels in an image. The condition (2) points out that the different segmented regions could not overlap each other. The condition (3) points out that the pixels in the same regions should have some similar properties. The condition (4) points out that the pixel belonging to different regions should have some different properties. The condition (5) points out that the pixels in the same region resulted from segmentation are connected.

Three Levels of Research on Image Segmentation

Though many efforts have been devoted to the research of segmentation techniques, there is no general theory for image segmentation, yet. Therefore, the development of segmentation algorithms has traditionally been an ad hoc process. As a result, many research directions have been exploited, some very different principles have been adopted, and wide varieties of segmentation algorithms have appeared in the related literatures. It was noted by many people that none of the developed segmentation algorithms are generally applicable to all kinds of images and different algorithms are not equally suitable for a particular application (Zhang, 2006).

With the increase of the number of algorithms for image segmentation, how to evaluate the performance of these algorithms becomes indispensable in the study of segmentation. Considering the various modalities for acquiring different images and the large number of applications requiring image segmentation, how to select appropriate algorithms for segmentation turns into an important task. A number

of evaluation techniques have been proposed. For those published in the last century, see survey papers by Zhang (1996) and Zhang (2001b).

While the evaluation of segmentation techniques has gained more and more attention, with numerous evaluation methods frequently designed, how to characterize the different existing methods for evaluation has also attracted some interest in recent years (Zhang, 2001a). In fact, different evaluation criteria and procedures, their applicability, advantages, and limitations need to be studied carefully and systematically.

According to the previous discussion, the research for image segmentation is carried out in three levels (Zhang, 2006). The first one and the basic one is the level of algorithm development. The second one is the level of algorithm evaluation. The third one is the level of systematic study of evaluation methods. This present article will mainly concentrate on the first level.

MAIN THRUST

The current study focuses on three points:

1. A worldwide statistics about the number of segmentation algorithms already developed.
2. A method for classifying different segmentation techniques into groups.
3. A general overview of survey papers for segmentation, published in the last 40 years.

Amount of Developed Segmentation Algorithms

Over the last 40 years, the research and development of segmentation algorithms are going on and making very rapid progress. A great number of segmentation algorithms have been developed and this number continually increases each year. Table 1 gives a list of the numbers of records (for every 5 years) found in EI Compendex (the most comprehensive bibliographic database of engineering research available today, see http://www.ei.org) by using the term *image segmentation* to search only in the field of "Subject/Title/Abstract."

Figure 1 gives a plot of the number of records found, together with a tendency curve obtained by using the third order polynomial. It is interesting to note the very fast increasing rate (an exponential raise) for the number of papers published. It is also interesting to note there is not any sign for the slowdown of augmenting.

A Classification of Segmentation Algorithms

With so many publications appearing in the literature and so many segmentation algorithms being developed, the classification of various algorithms for image segmentation becomes an essential task in studying image segmentation.

A classification of algorithms into groups, in principle, is a problem of set partition into subsets. With reference to the conditions for the definition of segmentation (Fu & Mui, 1981), it was believed that the resulted groups after

Table 1. List of records found in EI compendex

1965-1969	1970-1974	1975-1979	1980-1984	1985-1989	1990-1994	1995-1999	2000-2004	Total
10	20	233	680	1499	3423	7665	12727	26257

Figure 1. Number of records and the tendency of development in the last 40 years

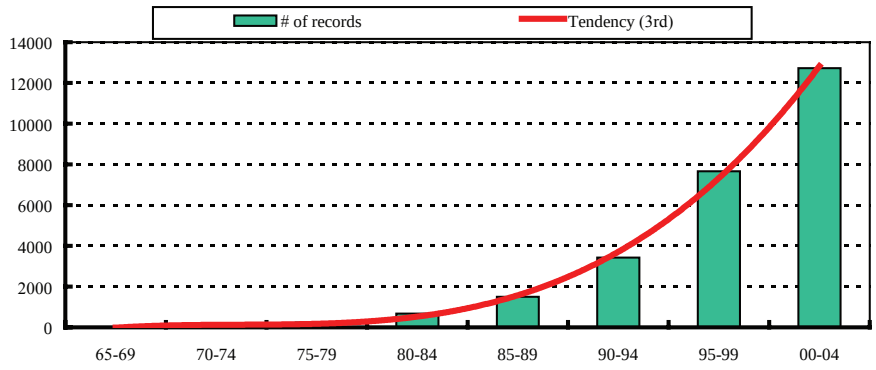


Table 2. General classification of segmentation algorithms

Classification	Edge-based (discontinuity)	Region-based (similarity)
Parallel process	G1: Edge-based parallel process	G3: Region-based parallel process
Sequential process	G2: Edge-based sequential process	G4: Region-based sequential process

an appropriate classification of segmentation algorithms, according to the process and objective, should satisfy the following four conditions (Zhang, 1997):

1. Every algorithm must be in a group,
2. all groups together can include all algorithms.
3. The algorithm in the same group should have some common properties and
4. the algorithm in different groups should have certain distinguishable properties.

Classifications of algorithms are performed always according to certain classification criteria. The first two conditions imply that the classification criteria should be suitable for classifying all different algorithms. The last two conditions imply that the classification criteria should determine the representative properties of each algorithm group.

Taking the aforementioned conditions in mind, the following two criteria turn up to be suitable for the classification of segmentation algorithms. The first is the discontinuity or similarity of pixel property; the second is the sequential or parallel of processing strategy. All segmentation algorithms can be classified into four groups, namely G1, G2, G3, and G4, according to these two criteria. The results are shown in Table 2.

Some simple and typical examples of each group are as follows (they can be easily found in the literature):

- **G1:** Edge detector, gradient operator, SUSAN operator, fuzzy edge detection.
- **G2:** Graph search, dynamical programming, active contour model, active shape model (snake).
- **G3:** Thresholding, K-means, feature space clustering, pixel classification.
- **G4:** Region growing, split, and merge, morphological watersheds.

Overview of Survey Papers on Image Segmentation

Along with the development of image segmentation algorithms, a number of survey papers for general image segmentation algorithms have been presented in the literature over the last 40 years (Borisenko, Zlatotol, & Muchnik, 1987; Buf & Campbell, 1990; Davis, 1975; Fu & Mui, 1981; Haralick & Shapiro, 1985; Nevatia, 1986; Pal & Pal, 1993; Pavlidis, 1986; Peli & Malah, 1982; Riseman & Arbib, 1977; Rosenfeld, 1981; Sahoo, Soltani, Wong, & Chen, 1988; Sarkar & Boyer, 1993; Weszka, 1978; Zucker, 1976; Zucker, 1977).

By partitioning the last 40 years into four decades, it is interesting to note that all these survey papers are dated in the second and third decades. The reason for no survey paper published in the first decade is that the research results were just cumulated in that period (some important works for this period have been indicated in the survey papers published in the second period). The reason for no survey in the last decade could be attributed to the fact that so many techniques have already been presented, thus a comprehensive survey becomes less feasible.

Though no general survey for the whole scope of image segmentation has been made in the last 10 years, some specialized surveys are nevertheless published in recent years. These survey papers can be classified into two subcategories:

1. Survey Papers Focused on Particular Group of Segmentation Algorithms

Many segmentation algorithms have been developed by using certain mathematical/theoretical tools, such as fuzzy logic, genetic algorithm, neural network (NN), pattern recognition, wavelet, and so forth; or based on some unique framework, such as active contour model (ACM), thresholding, watershed, and so forth. Some surveys for algorithms using the same tools or based on the same frameworks have been made. The following paragraphs present some examples.

Considering that the fully automatic methods sometimes would fail and produce incorrect results, the intervention of a human operator in practice is often necessary. To identify the patterns used in the interaction for the segmentation of medical images and to develop qualitative criteria for evaluating interactive segmentation method, a survey of computational techniques for human-computer interaction in image segmentation has been made (Olabarriaga & Smeulders, 2001). This survey has taken into account the type of information provided by the user, how this information affects the computational part, and the purpose of interaction in the segmentation process for the classification and comparison of a number of human-machine dialog methods.

Algorithms combining edge-based and region-based techniques will take the advantage of the complementary nature of edge and region information. A review of differ-

ent segmentation methods, which integrate edge and region information has been made (Freixenet, Muñoz, Raba, Martí, & Cufí, 2002). Seven different strategies to fuse such information have been highlighted.

Active shape model (ASM) is a particular structure for finding the object boundary in images. Under this framework, various image features and energy functions as well as different search strategies can be used, which makes a wide range of ASM algorithms. A number of these variations for segmentation of anatomical bone structures in radiographs have been reviewed in Behiels, Maes,Vandermeulen, and Suetens (2002).

Thresholding technique is a very popular, relatively simple and fast technique. A survey of thresholding methods with a view to assess their performance when applied to remote sensing images has been made recently (Marcello, Marques, & Eugenio, 2004). Some image examples are taken from oceanographic applications in this work.

2. Survey Papers Focused on a Particular Application of Image Segmentation

Image segmentation has many applications. For each application, a number of segmentation algorithms could be developed. Some surveys for certain particular application areas have been made.

In medical imaging applications, image segmentation is used for automating or facilitating the delineation of anatomical structures and other regions of interest. A survey considering both semi-automated and automated methods for the segmentation of anatomical medical images has been made (Pham, Xu, & Prince, 2000). The advantages and disadvantages of these methods for medical imaging applications are also discussed and compared.

While video could be considered as a particular type of general images, its segmentation is just an extension of image segmentation. For video data, the temporal segmentation is used for determining the boundary of shots. A survey has made for techniques that operate on both uncompressed and compressed video stream (Koprinska & Carrato, 2001). Both types of shot transitions: abrupt and gradual transitions are considered. The performance, relative merits, and limitations of each of the approaches are comprehensively discussed.

For temporal video segmentation, except the ability and correctness of shot detection, the computation complexity is also a criterion that should be considered, especially for real-time application. A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval has been made (Lefèvre, Holler, & Vincent, 2003). Depending on the information used to detect shot changes, algorithms based on pixel, histogram, block, feature, and motion have been selected.

Vessel extraction in bioengineering is essentially a segmentation process. A survey for related algorithms has been made (Kirbas & Quek, 2003). Six groups of techniques proposed for this particular application are involved: (1) pattern recognition techniques; (2) model-based approaches; (3) tracking-based approaches; (4) artificial intelligence-based approaches; (5) neural network-based approaches; and (6) miscellaneous tube-like object detection approaches.

In many vision applications, moving shadows must be detected. Moving shadows can be considered as object in video streams, and the detection of moving shadows is a video segmentation problem. A survey has been made for four classes of techniques (two statistical ones and two deterministic ones) designed specially for detecting moving shadows (Prati, Mikic, Trivedi, & Cucchiara, 2003).

FUTURE TRENDS

Though much progress has been made in the last 40 years, the subject of image segmentation still needs additional study efforts, a few further research directions would be:

- **Incorporating Human Factors:** Since image segmentation is a process at the middle layer of image engineering, it is influenced strongly by human factors. It seems that the assistance of humans, who are knowledgeable in the application domain, will remain essential in any practical image segmentation method. Incorporating high-level human knowledge algorithmically into the computer should be a challenge in the future.
- **Introducing More Mathematical Models and Theories:** The introduction of various mathematical models and theories into the research of image segmentation has proved to be quite effective. Since many novel models and theories have been invented and/or created in these years, introducing them into the research on image segmentation would be promising.
- **Enlarging the Scope of Applications:** Though no general theory for segmentation exists, researches on segmenting different particular images from numerous applications have made much progress. As a lot of new applications still call for segmentation, developing the suitable algorithms for those new areas would have great potential, either from the point of view technique developments or from the point of view of image applications.

CONCLUSION

An overview of the development of image and video segmentation in the last 40 years is provided with emphasis on

showing the number of segmentation algorithms already developed, on describing the techniques for classifying these algorithms and on analyzing the survey papers for image segmentation. With such an expansive overview, readers should perceive a general idea about the 40 years' progresses of research and application on image segmentation.

ACKNOWLEDGMENTS

This work has been supported by the National Natural Science Foundation under Grant NNSF-60573148 and the Ministry of Education under Grant SRFDP-20050003013.

REFERENCES

Behiels, G., Maes, F., Vandermeulen, D., & Suetens, P. (2002). Evaluation of image features and search strategies for segmentation of bone structures in radiographs using active shape models. *Medical Image Analysis, 6*(1), 47-62.

Borisenko, V. I., Zlatotol, A. A., & Muchnik, I. B. (1987). Image segmentation (state of the art survey). *Automatic Remote Control, 48,* 837-879.

Buf, J. M. H., & Campbell, T. G.. (1990). A quantitative comparison of edge-preserving smoothing techniques. *Signal Processing, 21,* 289-301.

Davis, L. S. (1975). A survey of edge detection techniques. *Computer Graphics and Image Processing, 4,* 248-270.

Freixenet, J., Muñoz, X., Raba, D., Martí, J., & Cufí, I. (2002). Yet another survey on image segmentation: Region and boundary information integration. In *Proceedings Proc. European Conference on Computer Vision 2002* (pp. 408-422). Copenhagen, Denmark: Springer.

Fu, K. S., & Mui, J. K. (1981). A survey on image segmentation. *Pattern Recognition, 13,* 3-16.

Haralick, R. M., & Shapiro, L. G. (1985). Image segmentation techniques. *Computer Vision, Graphics, and Image Processing, 29,* 100-132.

Kirbas, C., & Quek, F. K. H. (2003). Vessel extraction techniques and algorithms: A survey. *Proceedings of the 3rd Bioinformatics and Bioengineering Symposium* (pp. 238-245). Bethesda, MD: IEEE Computer Society Press.

Koprinska, I., & Carrato, S. (2001). Temporal video segmentation: A survey. *Signal Processing: Image Communication, 16*(5), 477-500.

Lefèvre, S., Holler, J., & Vincent, N. (2003). A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval. *Real-Time Imaging, 9*(1), 73-98.

Marcello, J., Marques, F., & Eugenio, F. (2004). Evaluation of thresholding techniques applied to oceanographic remote sensing imagery. *SPIE, 5573,* 96-103.

Nevatia, R. (1986). Image segmentation. In *Handbook of pattern recognition and image processing, 86*, 215-231.

Olabarriaga, S. D., & Smeulders, A. W. M. (2001). Interaction in the segmentation of medical images: A survey. *Medical Image Analysis, 5*(2), 127-142.

Pal, N. R., & Pal, S. K. (1993). A review on image segmentation techniques. *Pattern Recognition, 26,* 1277-1294.

Pavlidis, T. (1986). Critical survey of image analysis methods. *Proceedings 8th International Conference on Pattern Recognition,* Paris, France (pp. 502-511). New York: IEEE.

Peli, T., & Malah, D. (1982). A study of edge determination algorithms. *Computer Graphics and Image Processing, 20,* 1-20.

Pham, D., Xu, C., & Prince, J. (2000). Current methods in medical image segmentation. *Annual Review of Biomedical Engineering, 2,* 315-337.

Prati, A., Mikic, I., Trivedi, M., & Cucchiara, R. (2003). Detecting moving shadows: Algorithms and evaluation. *IEEE Pattern Aanalysis and Machine Intelligence, 25*(7), 918-923.

Riseman, E. M., & Arbib, M. A. (1977). Computational techniques in the visual segmentation of static scenes. *Computer Graphics and Image Processing, 6,* 221-276.

Roberts, L. G. (1965). Machine perception of three-dimensional solids. In J. T. Tippett, D. A. Berkowitz, L. C. Clapp, C. J. Koester, & A. Vanderburgh (Eds.), *Optical and electro-optical information processing* (pp. 159-157). Cambridge, MA: MIT Press.

Rosenfeld, A. (1981). Image pattern recognition. *Proceedings of IEEE, 69*(5), 596-605.

Sahoo, PK., Soltani, S., Wong, AKC., & Chen, YC., (1988). A survey of thresholding techniques. *Computer Vision, Graphics, and Image Processing, 41*, 233-260

Sarkar, S., & Boyer, K. L. (1993). Perceptual organization in computer vision: A review and a proposal for a classificatory structure. *IEEE Systems, Man and Cybenetics, 23,* 382-399.

Weszka, J. S. (1978). A survey of threshold selection techniques. *Computer Graphics and Image Processing, 7,*

259-265.

Zhang, Y. J. (1996). A survey on evaluation methods for image segmentation. *Pattern Recognition, 29*(8), 1335-1346.

Zhang, Y. J. (1997). Evaluation and comparison of different segmentation algorithms. *Pattern Recognition Letters, 18*(10), 963-974.

Zhang, Y. J. (2001a). *Image segmentation*. Beijing, China: Science Publisher.

Zhang, Y. J. (2001b). A review of recent evaluation methods for image segmentation. *Proceedings of the 6th International Symposium on Signal Processing and Its Applications* (pp. 148-151). Kuala Lumpur, Malaysia: IEEE Signal Processing Society Press.

Zhang, Y. J. (2006). *Advances in image and video segmentation*. Hershey, PA: IRM Press.

Zhang, Y. J. (forthcoming). Ten years' survey on image engineering. In M. Khosrow-Pour (Ed.) *Encyclopedia of information science and technology* (2nd ed.). Hershey, PA: Ide Group Reference.

Zucker, S. W. (1976). Region growing: Childhood and adolescence. *Computer Graphics and Image Processing, 5,* 382-399.

Zucker, S. W. (1977). Algorithms for image segmentation. In A. Rosenfeld & J. C. Simon (Eds.) *Digital image processing and analysis* (pp. 169-183). Nordhoff International Publisher.

KEY TERMS

Active Contour Model: Active contour model is a sequential technique for image segmentation. Given an approximation of the boundary of an object in an image, an active contour model can be used to find the *actual* boundary by deforming the initial boundary to lock onto features of interest within this image.

Clustering: Clustering is also called unsupervised learning and is a powerful technique for pattern classification. It is a process to group, based on some defined criteria, two or more terms together to form a large collection In the context of image segmentation, it is often considered as the multi-dimensional extension of the thresholding technique.

Edge Detection: Edge detection is the most common approach for detecting discontinuities in images, and is the fundamental step in edge-based parallel process for segmentation. An edge is a local concept. To form a complete boundary of an object, edge detection should be followed by edge linking or connection.

Gradient Operator: Gradient operator is the first type of operator used for edge detection. The gradient of an image is a vector consisting of the first order derivatives (including the magnitude and direction) of an image.

Graph Search: Graph search is a particular type of segmentation technique which combines edge detection and linking together. It represents edge segments in the form of a graph and searches the graph for low-cost paths that correspond to significant edges or boundaries of objects.

Image Engineering: Image engineering is an integrated discipline/subject comprising the study of all the different branches of image and video techniques. It mainly consists of three levels: image processing, image analysis, and image understanding.

Image Segmentation: A process consists of subdividing an image into its constituent parts and extracting these parts of interest (objects) from the image.

Region Growing: Region growing is a region-based sequential technique for image segmentation by assembling pixels into larger regions based on predefined seed pixels, growing criteria, and stop conditions.

Thresholding: Thresholding techniques are the most popularly used segmentation techniques. A set of suitable thresholds need to be first determined, and then the image can be segmented by comparing the pixel properties with these thresholds.

Watersheds: Watershed technique is inspired from the topographic interpretation of image segmentation by watersheds embodies many concepts of edge detection, thresholding and region processing techniques, and often produces stable and continuous results.

Imaging Advances of the Cardiopulmonary System

Holly Llobet
Cabrini Medical Center, USA

Paul Llobet
Cabrini Medical Center, USA

Michelle LaBrunda
Cabrini Medical Center, USA

INTRODUCTION

A technological explosion has been revolutionizing imaging technology of the heart and lungs over the last decade. These advances have been transforming the health care industry, both preventative and acute care medicine. Ultrasound, nuclear medicine, computed tomography (CT), and magnetic resonance imaging (MRI) are examples of radiological techniques which have allowed for more accurate diagnosis and staging (determination of severity of disease). The most notable advances have occurred in CT and MRI. Most medical subspecialties rely on CT and MRI as the dominant diagnostic tools an exception being cardiology. CT and MRI are able to provide a detailed image of any organ or tissue in the body without the necessity of invasive or painful procedures. Virtually any individual could be tested as long as they are able to remain immobile for the duration of the study.

Image generation traditionally has been limited by the perpetual motion of the human body. For example, the human heart is continually contracting and relaxing without a stationary moment during which an image could be obtained. Lung imaging has been more successful than cardiac imaging, but studies were limited to the length of time an ill person is able to hold his or her breath.

Historically, imaging technology was limited by inability to take a picture fast enough of a moving object while maintaining a clinically useful level of resolution. Recent technologic innovation, resulting in high speed electrocardiogram-gated CT and MRI imaging, now allows the use of these imaging modalities for evaluation of the heart and lungs. These novel innovations provide clinicians with new tools for diagnosis and treatment of disease, but there are still unresolved issues, most notably radiation exposure. Ultrasound and MRI studies are the safest of the imaging modalities and subjects receive no radiation exposure. Nuclear studies give an approximate radiation dose of 10mSv and as high as 27mSv (Conti, 2005). In CT imaging, radiation dose can vary depending on the organ system being imaged and the type of scanner being used. The average radiation dose for pulmonary studies is 4.2mSv (Conti, 2005). The use of multi-detector CT (MDCT) to evaluate the heart can range from 6.7—13mSv. To put it into perspective, according to the National Institute of Health, an average individual will receive a radiation dose of 360mSv per year from the ambient environment. It is unlikely that the radiation doses received in routine imaging techniques will lead to adverse reactions such as cancer, but patients should be informed of the risks and benefits of each procedure so that they can make informed decisions. It is especially important that patients be informed when radioactive material is to be injected into their bodies. The reasons for this will be discussed later on in the chapter.

BACKGROUND

Coronary artery disease (CAD) is the leading cause of death in the U.S.. It is also the most common cause of the most costly heart disease in the U.S.—heart failure (Thomas & Rich, 2007). With a prevalence of almost 13,000,000, it carries an estimated cost of US$130 billion per year (Sanz & Poon, 2004). The American Heart Association calculates the cost of caring for those with cardiovascular disease at $300 billion per year (Raggi, 2006). According to the American Lung Association, lung disease is the third most common cause of death in America, responsible for one in seven deaths every year. Infections, particularly those of the lungs, are the number one killer of infants. An estimated 35 million Americans are currently living with lung disease such as emphysema, asthma, or chronic bronchitis. Even small technological advancement can lead to dramatic improvements in health care**.** Raggi (2006), the American Heart Association/American College of Cardiology and the

National Cholesterol Education program III have describe simple screening tools to risk stratify patients (group patients into prognostic categories). These stratification tools involve assessment of lifestyle factors, genetic factors, and simple blood tests. For example, patients who smoke, are obese, inactive, have high cholesterol, and a family history of heart disease are more likely to have heart disease than people who lack these risk factors. Based on the results of the screening, recommendations can be made for further diagnostical procedures, treatments, and lifestyle interventions. Some of the more advanced diagnostic procedures include echocardiography, nuclear medicine (NM) scans, CT, and MRI.

ECHOCARDIOGRAPHY

Echocardiography uses the principle of ultrasound (sound wave) reflection off cardiac structures to generate images of the heart just as it does in producing images of an unborn baby in a pregnant woman. In the past three decades, echocardiography has rapidly become a fundamental component of the cardiac evaluation. Measuring the same feature from different angles (windows) with different types of sound wave detectors (transducer) is done to produce an image. The entire heart and its major blood vessels can be displayed in real time and in various 2-D planes. Transthoracic echocardiogram (TTE) imaging is performed with a handheld transducer placed directly on the chest wall. In select situations in which a more direct image needs to be taken without the interference of the chest wall, chest muscles, and chest fat, a transesophageal echocardiogram (TEE) may be performed. In TEE, an ultrasound transducer is mounted on the tip of an endoscope and placed into the esophagus of a patient. The transducer is directed toward the heart so that high-resolution images can be attained with minimal interference. TEE is a riskier and more invasive procedure but is able to provide details on the parts of the heart distant from the transducer on TEE.

Newer echocardiographic machines have the advantage of being lightweight and portable (Figure 1). Some machines weigh less than 6 pounds and can be carried like a briefcase. The benefit of these advances is the ability to immediately assess heart function without the risk of having to move unstable patients (Beaulieu, 2007). This will soon become an essential part of the physical examination both hospitals and the outpatient setting. The echocardiograph permits quantification of the overall ability of the heart to supply blood to the body, description of the appropriate opening and closing of the heart valves, determination of leaks within the heart valves, visualization of abnormal fluid around the heart, visualization fluid trapped in the lungs and detection of structural abnormalities that disturb the blood flow through the heart (Figure 2). Although one of the older modalities in cardiopulmonary imaging, it still is the test of choice for certain conditions. TTE is advantageous in emergency situations when time is limited and a patient may be too unstable to transport (Shiga, Wajima, Apfel, Inoue, & Ohe, 2006). The most significant limitation is in the quality of the image that TTE delivers. A large amount of information can be generated from TTE, but the images may be particularly limited in obese individuals or those severe lung disease. In addition echocardiography is highly operator dependant for generating and interpreting images.

Doppler echocardiography is a newer modality in heart imaging. It uses ultrasound reflecting off moving red blood

Figure 1. Portable echocardiogram machine (Sonosite MicroMaxx, product photographs reprinted with permission from SonoSite; Sonosite trademarks owned by SonoSite, Inc.)

Figure 2. Echocardiogram of the heart (product photographs reprinted with permission from SonoSite)

cells to measure the velocity of blood flow across several different parts of the heart. Normal and abnormal blood flow patterns can be assessed and by adding color. Blood flow can be evaluated for smoothness or turbulence. The addition of color to Doppler echoes is an important advancement yielding more accurate readings.

NUCLEAR MEDICINE

NM techniques consist of the injection of a radioactive isotope into the blood stream of an individual. This isotope emits a photon (usually a gamma ray) generated during radioactive decay as the nucleus of one isotope changes to a lower energy level. Special cameras are placed over the chest of the individual and capture the photons released.

Nuclear studies of the heart and lungs have numerous applications. One can assess blood flow to specific areas of the heart. Blood provides a fresh supply of oxygen and nutrients required for the proper cardiac function. NM scans occurs in two parts, one while the patient is at rest and the second after the heart has been exercised. While at rest, the isotope is injected into the blood stream while the individual lies flat. A specialized camera rests over the heart capturing the emitted photons. The second phase of the test occurs just after the heart rate has been increased, either by medication or exercise. The camera then captures the photons emitted during increased cardiac stress and a comparison is made between the two studies. In a healthy heart, the photons emitted during both tests should be equal creating identical pictures. If any region of the heart that has poor circulation an image produced from the nuclear scan shows a black area from which few or no photons are emitted. This signifies that there is a reduced blood flow to that area of the heart and allows treatment to be initiate before the decreased blood flow leads to a heart attack.

Similar techniques are used when examining the lungs. A lung scan is most commonly used to detect a blood clot obstructing normal blood flow to part of a lung (pulmonary embolism). The test is broken down into two parts—a ventilation scan (V) and a perfusion scan (Q)—which combined comprise the V/Q scan. The ventilation scan involves inhaling of a known amount of radioactive tracer by an individual. Cameras are then placed over the chest to detect the emitted photons. An image representative of the movement of air through the lung is generated. Areas with inadequate ventilation appear dark because few photons arrive to those areas. Conversely, areas of the lung containing too much air emit an increased number of photons. The perfusion scan works in a similar fashion to that of the NM heart scan. The radioactive tracer is injected into the bloodstream as the blood passes through to the lungs. The scan will show areas of the lungs that are not receiving enough blood. Normally the tracer should be evenly distributed throughout the lung. When both tests are completed, a comparison of the images is done and any mismatch may indicate a that not enough blood is arriving or not enough air is arriving to a specific region of the lung.

A newer modality in nuclear medicine is positron emission tomography (PET). A PET scan is a diagnostic examination that involves the acquisition of physiologic images based on the detection of radiation from the emission of positrons. Positrons are tiny particles emitted from a radioactive substance injected into a patient's blood stream. PET scans of the heart and lungs can be done to examine blood flow to the heart, viability of the heart muscle, detect cancer in the lungs, and determine the effectiveness of cancer therapy using chemotherapeutics. Unlike the previous tests, which are only able to describe structure, PET scans are able to relay information on the functionality of the tissues being studied. Current research is investigating the possibility of using PET scan to routinely measure cardiac blood flow (deKemp, Yoshinaga & Beanlands, 2007; Di Carli, Dorbala, Meserve, El Fakhri, Sitek, & Moore, 2007).

One major problem in generating images from this technique is that the photons are emitted in a random array of directions from the origin. False readings can occur if the cameras are not placed correctly. A second problem occurs when tissues other than those targeted absorb photons. When this occurs insufficient numbers of photos reach the camera to create an interpretable image. The use of higher energy isotopes, most commonly technetium 99m (99m Tc) and thallium 201 (201 Tl) helps avoid these difficulties (Hoheisel, 2006). Both isotopes are frequently used and many times choices of isotope are made solely on the cost or experience of the operator.

COMPUTED TOMOGRAPHY

CT has emerged as a cutting edge technique able to evaluate the structure and function of both the heart and lungs. Often, both lung and heart imaging can be done simultaneously with no added patient inconvenience. Advancements in imaging speed have allowed for more accurate evaluation of both relatively stationary structures, such as large arteries as well as rapidly moving structures, such as the heart. Traditional mechanical CT scanners produce images by rotating an x-ray tube around a circular gantry through which the patient advances on a moving bed. The basic principle of CT is that a fan-shaped, thin x-ray beam passes through the body at many angles providing cross-sectional images. The corresponding x-ray transmission measurements are collected by a detector array. Information entering the detector array and x-ray beam itself is collimated to produce thin sections and avoid unnecessary photon scatter. The transmission measurements recorded by the detector array are digitized into picture elements with known dimensions and are reconstructed. What

this provides is a 3-D 360 degree examination of not only the heart, lungs, and its arteries, but also the lumen and walls of the arteries and airways. These images can provide detailed information on the size of the blood vessels lumen (the hole through the blood vessel). If the lumen of the blood vessels of the heart are significantly reduced and the individual is at high risk for a heart attack and may already be having chest pain resulting from decreased blood flow to the heart. In some institutions, cardiac CT is replacing cardiac catheterization for the diagnosis of coronary artery disease.

Although x-ray angiography remains the "gold standard" in diagnosing changes in artery lumen size, it is an invasive procedure. Non-invasive modalities such as CT and MRI are becoming routine in the diagnostic examination of patients with vascular disease (Dowe, 2007; Raff & Goldstein, 2007). Conventional CT scanners have been modified and now use a different approach to gather data. Spiral (helical) CT and now MDCT are now replacing older scanners providing detailed information in one sweep. Since spiral CT became a part of the routine diagnostic imaging in the early 1990's, CT as a whole has matured. In spiral CT scanners, the x-ray tube rotates continuously around the patient as the bed moves through the scanner eliminating stops in between traditional shots. This prevents data misread and reduces the time needed to acquire images. The recent development and refinement of MDCT systems is a pivotal advance in CT technology (Chartland-Lefebvere et al., 2007). MDCT can take a complete 3-D picture of the heart within one heartbeat simultaneously imaging pulmonary structures.

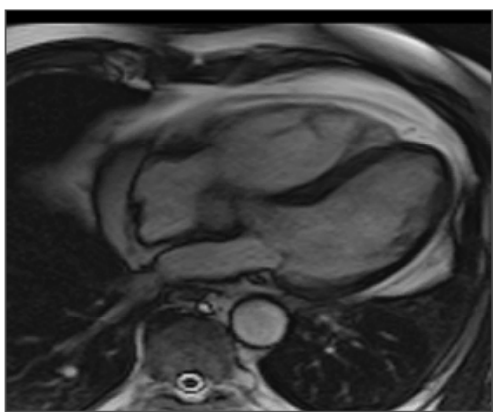
Many researchers now consider CT angiography/CT venography the test of choice when evaluating lungs for pulmonary embolism (blood clots) (Garcia, Lessick, & Hoffmann, 2006; Stein et al., 2006). As with cardiac CT imaging, the test involves placing a patient on the bed of the scanner. The bed moves through the doughnut-shaped scanner. In a cardiac CT, the patient is attached to a cardiac monitor so that it can follow the heart rate. An optimal study occurs if the heart rate remains regular and kept a maximal rate of 65 beats per minute (Poon, 2006). Many times a short acting medication is given to slow the heart rate. An iodinated contrast is injected into the blood stream of the patient, and the scanner is set on a timer. The timer predicts when the contrast will most likely reach the heart and images are recorded accordingly. Mis-timing can lead to poor images and an uninterpretable test. The patient is asked to hold his or her breath for 15 to 20 seconds as the scan is being done. The breath hold prevents the heart from moving as the chest rises and falls. Finally, the images are reconstructed able to be reviewed. Not only have these advancements changed how heart disease is measured, but have also substantially improved the quality of CT angiography of extracranial, thoracic, abdominal, pulmonary, and peripheral vasculature.

MAGNETIC RESONANCE IMAGING

MRI is a technique based on the magnetic properties of hydrogen protons present in water molecules. A large magnetic field can be used to induce, nuclear spin transitions from the ground state to excited states. As the nuclei lose energy, they return to their ground state, releasing energy in the form of electromagnetic radiation, which can be detected and processed into an image. Different tissues have a different concentration of water protons and molecular environments, therefore generating signals with different characteristics, which is the basis of the high quality tissue-contrast resolution of MRI (Sanz & Poon, 2004). The development of cardiac MRI has been particularly challenging because of motion of the heart and coronary arteries. Long scanning times needed in old MRI techniques made it technically difficult to get interpretable results.

Figure 3. MRI system (reprinted with permission from Siemens)

Figure 4. MRI of the heart (reprinted with permission from Siemens)



Much like the CT, an individual is placed on a sliding table and positioned in the MRI scanner. It also is doughnut-shaped with the table sliding back in forth through the hole (Figure 3). Depending on the number of images needed, the scan time can take 15 to 45 minutes. Contrast material can be injected into the blood stream but unlike CT contrast or nuclear medicine scans, there is no radiation or radioactive isotopes used.

The high-resolution 3-D images without the use of radiation of cardiac MRI sets it apart from other imaging modalities, but it also had its limitations (Figure 4). Disadvantages included long scanning times preventing acutely ill individuals and those who suffer from claustrophobia from tolerating the test. Also, anyone with metallic foreign bodies embedded in their bodies such as pacemakers, surgical clips, or bullet fragments are typically unable to undergo MRIs. Despite the advantages, MRI is rarely used as the initial imaging technique because of lack of availability, time delay, incompatibility with implanted metal devices, and monitoring difficulties during the study (Shiga et al., 2006). The most common patient complaint is remaining still during the study, which can last 30 minutes or more.

FUTURE TRENDS

Ultrasound and nuclear medicine have undergone technological breakthroughs, and newer devices provide more accurate and detailed results, which manufactures and researchers continue to focus on developing improved MDCT and MRI scanners. MDCT scanners will become faster with greater precision and detail allowing for improved diagnostic accuracy and reproducibility (Sanz & Poon, 2004). Similar advances are taking place with MRI. Scanners are being designed utilizing stronger magnetic fields, improving data acquisition technology, and improving contrast agents to produce higher quality images. The use of CT may decline as MRI becomes better, faster, and cheaper. The advancements of all these imaging modalities will change the field of medicine for the better.

CONCLUSION

Each imaging modality provides a specific type of information and when used appropriately can aid physicians in making rapid diagnoses, and each also has its limitations. At present it may appear as multi-slice CT imagers are taking center stage especially with the development of 16-slice, 32-slice, 64-slice, and 256-slice imagers now available. Presently, MR is the imaging modality of choice for peripheral vasculature and is more suited for elective diagnosis rather than clinical emergencies. CT plays a large role for pulmonary and cardiac imaging. Since technological advances and clinical research continue to improve all modalities, it is impossible to know which will be of most utility in the future. Whatever the future of cardiopulmonary imaging brings, there will always be a need to understand basic principles of ultrasound technology, nuclear medicine, CT, and MRI.

REFERENCES

Beaulieu, Y. (2007). Bedside echocardiography in the assessment of the critically ill. *Critical Care Medicine, 35*(5 Suppl), S235-249.

Chartrand-Lefebvre, C., Cadrin-Chênevert, A., Bordeleau, E., Ugolini, P., Ouellet, R., Sablayrolles, J., & Prenovault, J. (2007). Coronary computed tomography angiography: Overview of technical aspects, current concepts, and perspectives. *Canadian Association of Radiology Journal, 58*(2), 92-108.

Conti, C. (2005). One-stop cardiovascular diagnostic imaging (and radiation dose). *Clinical Cardiology, 28*(10), 450-453.

deKemp, R., Yoshinaga., K., & Beanlands, R. (2007). Will 3-dimensional PET-CT enable the routine quantification of myocardial blood flow? *Journal of Nuclear Cardiology, 14*(3), 380-397.

Di Carli, M., Dorbala, S., Meserve, J., El Fakhri, G., Sitek, A., & Moore, S. (2007). Clinical Myocardial Perfusion PET/CT. *Journal of Nuclear Medicine, 48*(5), 783-793.

Dowe, D. (2007). The case in favor of screening for coronary artery disease with coronary CT angiography. *Journal of the American College of Radiology, 4*(5), 295-299.

Garcia, M., Lessick, J., & Hoffmann, M. (2006). Accuracy of 16-row multidetector computed tomography for the assessment of coronary artery stenosis. *The Journal of the American Medical Association, 296*(4), 403-411.

Hoheisel, M. (2006). Review of medical imaging with emphasis on x-ray detectors. *Nuclear Instruments and methods in Physics Research, section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 563*(1), 215-224.

Poon, M. (2006). Technology insight: Cardiac CT angiography. *Nature Clinical Practice Cardiovascular Medicine, 3*(5), 265-275.

Raff, G., & Goldstein, J. (2007). Coronary angiography by computed tomography: Coronary imaging evolves. *Journal of the American College of Cardiology, 49*(18),1827-1829.

Raggi, P. (2006). Nonivasive imaging of atherosclerosis among asymptomatic individuals. *Archives of Internal Medicine, 166*(10), 1068-1070.

Sanz, J., & Poon, M. (2004). Evaluation of ischemic heart disease with cardiac magnetic resonance and computed tomography. *Expert Review of Cardiovascular Therapy, 2*(4), 601-615.

Shiga, T., Wajima, Z., Apfel, C., Inoue, T., & Ohe, Y. (2006). Diagnostic accuracy of transesophageal echocardiography, helical computed tomography, and magnetic resonance imaging for suspected thoracic aortic dissection: Systematic review and meta-analysis. *Archives of Internal Medicine, 166*(13), 1350-1356.

Stein, P., Woodard, P., Weg, J., Wakefield, T., Tapson, V., Sostman, H., Sos, T., Quinn, D., Leeper, Jr., K., Hull, R., Hales, C., Gottschalk, A., Goodman, L., Fowler, S., & Buckley, J. (2006). Diagnostic pathways in acute pulmonary embolism: Recommendations of the PIOPED II investigators. *The American Journal of Medicine, 119*(12), 1048-1055.

Thomas, S., & Rich, M. (2007). Epidemiology, pathophysiology and prognosis of heart failure in the elderly. *Clinical Geriatric Medicine, 23,* 1-10.

KEY TERMS

Atherosclerosis: A condition in which fatty material collects along the walls of arteries. This fatty material thickens, hardens, and may eventually block the arteries.

Computed Tomography: An imaging technology completed with the use of a 360-degree x-ray beam and computer production of images. These scans allow for cross-sectional views of body organs and tissues.

Coronary Heart Disease: A narrowing of the small blood vessels that supply blood and oxygen to the heart. CHD is also called coronary artery disease (CAD).

Echocardiography: A procedure that evaluates the structure and function of the heart by using sound waves recorded on an electronic sensor producing a moving image of the heart and heart valves.

Magnetic Resonance Imaging: The use of a nuclear magnetic resonance spectrometer to produce electronic images of specific atoms and molecular structures in solids, especially human cells, tissues, and organs.

Multidetector Computed Tomography: An imaging system which incorporates approximately 1,500 solid-state ceramic X-ray detector elements, each approximately 1 mm wide, arranged in rows. This detector array has been combined with improved X-ray tube design, allowing faster, multilevel scanning with high spatial resolution.

Nuclear Medicine Scan: The use of a camera to image certain tissue in the body after a radioactive tracer accumulates in the tissue.

Spiral (Helical) Compute Tomography: A newer version of CT scanning which is continuous in motion and allows for three-dimensional image generation.

The Impact of Network-Based Parameters on Gamer Experience

Dorel Picovici
Institute of Technology Carlow, Ireland

David Denieffe
Institute of Technology Carlow, Ireland

Brian Carrig
Institute of Technology Carlow, Ireland

INTRODUCTION

Most of the existent games consist of multiple players, connected over a network, collaborating and competing in a virtual world. In this world, each player typically controls a single virtual entity. Communication between players can be achieved by sharing entity state information, such as positioning using synchronisation messages. These messages are periodically transmitted across the connecting network, and update the remote state of the virtual entity, which is the state replicated on other players' computers. If the games are using the Internet as a connecting network, latency, jitter, and packet loss, can have a significant impact upon the service experienced by application user (end-user). More specifically, latency or delay can by introduced by various types of delays, such as propagation, serialisation, and queuing delays.

Jitter is the variation in latency experienced by consecutive packets. Not all of the packets in a given flow will take the same path through the network. The time taken to traverse different routes is likely to vary due to factors such as their different physical distance, the number of hops, or physical link properties. On lower bandwidth links, which have a greater serialization delay, variation in packet lengths can introduce jitter. The size distribution and arrival patterns of other traffic flows on shared links may influence the queuing delay experienced by packets in one particular flow and is, in itself, a source of jitter. There are a number of different points on the network where packets may be lost. At the physical layer, all links experience some rate of data corruption, known as the bit error rate (BER). This may be caused by a high signal-to-noise ratio (SNR) during digital to analog conversion processes, which causes erroneous encoding or decoding of data, or it may be caused by faulty hardware. Forward error correction (FEC) is sometimes used at the link layer to recover from one- or two-bit errors. A cyclic redundancy check (CRC) may be applied to detect whether or not errors are included in the frame. Occasionally, transient congestion, with the subsequent queuing of packets, is so severe that it causes the routing queue buffer to overflow. When this occurs, newly arriving packets will be dropped until there is sufficient space in the buffer to place new packets. On other occasions, dynamic routing changes or route flapping may result in a temporarily incomplete network path, which causes losses.

BACKGROUND

The current Internet model provides only a single level of service, known as best effort (BE) delivery of data. In this model, the network will attempt to route traffic as quickly as possible to its destination, but provides no guarantees that those packets will traverse the same path across the network, arrive in the same order, or even arrive at all. Streaming and interactive applications (often referred to as *nonelastic* applications) require upper bounds to be placed on delay and jitter to facilitate smooth delivery. Packet loss and corruption must be also bounded to ensure adequate subjective quality, as it may not be possible to retransmit packets within the required time frame. How the required level of service might best be provided to the multitude of applications in use on the Internet has been the subject of extensive study for over a decade. When discussing the motivation for quality of service (QoS) in communications networks, it is important to define the term "quality of service." The International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T) describes the term "quality of service" as "The collective effect of service performance which determines the degree of satisfaction of a user of the service" (ITU-T, 1994). When this definition is used, it can be clearly seen that QoS is something that can only be correctly determined by the user of a service, as it relates to the user's expectation of service quality.

The quality of a voice/video telephony call, or a multiplayer game can be assessed using objective and subjective

methods. For voice/video, objective methods allow subjective quality to be predicted on the basis of psychoacoustic modeling (Hollier & Cosier, 1996). Similar applications are very much in their infancy for gaming. Subjective methods of assessment involve playing sample stimuli to a target audience in order to gather opinion data. The primary difficulty with subjective assessment is that of eliciting useful objective information when a user expresses their satisfaction or otherwise with a service. Quantifying what is meant by "good" is difficult without contextual information about the user's previous experiences and context. This points to the inherent complexity of measuring subjective quality and the difficultly involved in trying to obtain uniform answers from a diverse population of users. On the other hand, objective methodology, though often preferred because it is easier to measure and gather data for, is not without problems.

Subjective Measures of Mean Opinion Score (MOS)

Assessment measures that are based on ratings by human listeners are called subjective measures. When used for telecommunications systems, these tests seek to quantify the range of opinions that listeners express when they hear speech transmission of systems that are under test. Properly designed subjective tests provide the most accurate way of assessing speech quality. However, the results of subjective tests are influenced by the conditions of the tests, and great care must be taken of a number of factors in order to obtain reliable and reproducible results.

Although subjective assessment of speech quality requires substantial efforts, it is indispensable as a reference for evaluating the performance of objective speech quality

Table 1. Listening-quality scale

Quality of speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 2. Listening-effort scale

Effort required to understand the meaning of sentence	Score
Complete relaxation possible; no effort required	5
Attention necessary; no appreciable effort required	4
Moderate effort required	3
Considerable effort required	2
No meaning understood with any feasible effort	1

Table 3. Listening-effort scale

Loudness preference	Score
Much louder than preferred	5
Louder than preferred	4
Preferred	3
Quieter than preferred	2
Much quieter than preferred	1

measures. All subjective methods involve the use of large numbers of human listeners to produce statistically valid subjective quality indicator. The indicator is usually expressed as a *mean opinion score (MOS)*, which is the average value of all the rating scores registered by the subjects. For rating, the following opinion scales are approved and recommended by ITU-T:

Listening-Quality Scale

Listening-quality scale is a five-point category-judgment scale representing perceptual impression to speech quality as shown in Table 1.

The arithmetic mean of the listening-quality scale accumulated from all the subjects is known as the mean listening-quality opinion score, or simply mean opinion score, and is represented by the symbol MOS.

Listening-Effort Scale

Typical layout and wording of the listening effort scale is given in Table 2. The quantity evaluated from the scores accumulated (mean listening-effort opinion score) is represented by the symbol MOS_{LE}.

Loudness-Preference Scale

The layout and wording of the loudness-preference scale is given in Table 3. The quantity evaluated from the scores (mean loudness-preference opinion score) is represented by the symbol MOS_{LP}.

Amongst all the scales presented, the most used one is "listening–quality scale," as shown in Table 1, where the arithmetic mean of the listening quality scale accumulated is represented by MOS.

Objective Measures of MOS

Objective methods attempt to extract and catalog the motivations behind a user's decisions during an interaction. They are particularly applicable in the early stages of research or experiment design to establish a framework in which to operate. However, such methods should always be used in conjunction with subjective data to enable users to properly evaluate their experiences.

The E-model is a mathematical model of objective measurements, detailed in ITU-T Recommendation G.107 (ITU-T, 1995). The basic principle of the model is that "Psychological factors on the psychological scale are additive." It estimates the user satisfaction of a narrowband handset conversation, as perceived by the listener. The transmission rating factor *R* is obtained using a series of network impairment factors. Once computed, this R-value can be converted to an MOS score. The basic equation for the model is given by

$$R = Ro - Is - Id - Ie + A$$

where:

Ro = *S/N* at 0 dBr point
Is = Impairments simultaneous to voice signal
Id = Impairments delayed after voice signal
Ie = Impairments of special equipment such as codecs
A = Advantage factor which takes account of user advantages such as mobility

Another technique for estimating the effect of network impairments on Quake 3 clients is proposed by Ubicom (2005). The Quake 3 G-model is based on the ITU-T recommended E-model (ITU-T, 1995), as described previously. The impairment factor R is determined by:

$$R = (WL * L + WJ * J)(1 + E)$$

where

- *WL* is the latency weighting factor
- *WJ* is the jitter weighting factor
- *E* is the packet loss ratio
- *L* is the latency (one-way) in ms
- *J* is the jitter in ms as defined in Request for Comment (RFC) 1889

Some metrics used are based on work by Zander and Armitage (2004), and are correlated to average player performance, as defined by average player frags per minute. Obviously such a performance measure will not be applicable to all gaming genres, hence, the use of the more generic OPScore value. The measure has received criticism in some quarters for failing to account for other causes of lag, such as server and client processing delays, and the accuracy of the measure has been questioned on some forums.

Work by Wattimena (2006) is successive and focused on developing a similar model for QuakeIV. Data is taken from subjective tests on user samples, where network impairments are introduced, before applying multidimensional linear regression analysis to determine the appropriate mapping function. This function is given by

$$MOS = -0.00000587X^3 + 0.00139X^2 - 0.114X + 4.37$$

where the network impairment *X* is given by function:

$$X = 0.104 * ping_{average} + jitter_{average}$$

This mapping function displays very high correlation (R = 0.98) with subjectively determined MOS scores. It also displays high correlation with user gaming performance.

Analysis of impairment factors across a broader range of applications can be found in Dick et al., (Dick, Wellnitz, & Wolf, 2005). Here, the influence of latency, jitter, and overall skill level on player perception and performance for four separate multiplayer games is evaluated. One of the interesting things about this study is that it did not concentrate solely on FPS games, though two of the games studied are this genre, namely Counter Strike and Unreal Tournament 2004. There is also a car racing game, Need for Speed Underground 2, and an RTS title, World of Warcraft III. Both a subjective MOS and an objective normalized game score metric is determined for all titles. Multidimensional linear regression analysis was used to show coherences between the results. The analysis determined that not only do different multiplayer games behave differently under the same networking conditions, but also that even titles in the same genre behaved differently. This difference likely arises from the varying degrees of success with which compensatory techniques are employed by games developers.

NEW OBJECTIVE GAME QUALITY ASSESSMENT

Most existent game quality assessment techniques described take into consideration only network impairments, therefore, the measured games quality is only correlated with the network impairments (delay, jitter, and packet loss). To estimate the player's overall perception of games quality, the proposed objective game quality assessment extends the traditional objective game quality methods by introducing the end-user experience/knowledge. As shown in Figure 1, the proposed objective game quality assessment takes into consideration the following parameters:

a. end-user experience
b. distortions introduced by game client equipment (memory, graphic card) and I/O devices (screen, keyboard, and joystick)
c. distortions introduced by the network (end-to-end delay, jitter, packet loss)
d. distortions introduced by game server (number of users, game type, game capability to adapt to network distortions)

Using these parameters, a "game rating factor" (*GRF*) is proposed. The *GRF* is inspired from an ITU-T recommended computational model (E-Model) successfully used to assess the combined effects of variation in several parameters that may affect end-user perception of speech quality. The computation of the *GRF* can be described as follows: a maximum value that reflects a high level of game quality will be reduced in proportion with the distortions caused by various impairment parameters. The following equation is proposed for *GRF* calculation:

$$GRF = GRFMAX - IGCD - IN - IGS + A$$

where:
GRFMAX is the maximum Game Rating Factor
IGCD: impairment factor representing all impairments due to Game Client and I/O device
IN: impairment factor representing all impairments due to network connection between the game server and game client

Figure 1. Proposed game quality assessment

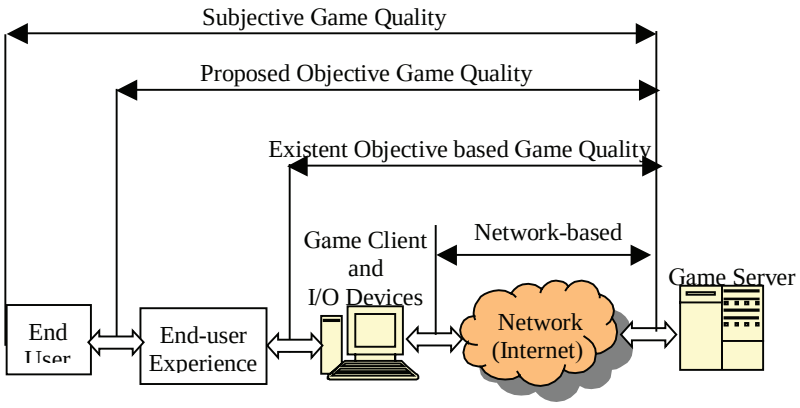


Table 4. Relation between the GRF, MOS_{GQE} and user satisfaction

GRF **(lower limit)**	$\mathbf{MOS_{GQE}}$ **(lower limit)**	**User satisfaction**
90	4.34	Very satisfied
80	4.03	Satisfied
70	3.60	Some users dissatisfied
60	3.10	Many users dissatisfied
50	2.58	Nearly all users dissatisfied

IGS: impairment factor representing all impairments due to Game Server
A: represents the end-user experience with online games.

The *GRF* can lie in the range from 0 to 100, where *GRF*=0 represents an extremely bad game quality and *GRG*=100 represents a very high game quality. The maximum value of 100 is in line with the ITU-T Recommendation G.107 (ITU-T, 1995). These parameters (excepting *GRFMAX*) will be individually calibrated and their effect on user perception will be measured directly using subjective assessment (MOS_{GPE}). An estimated MOS_{GQE} for an online game situation in the scale 1-5 can be obtained from the GRF using the following mapping function:

For $GRF < 0$: $MOS_{GQE}=1$
For $0 < GRF < 100$:
$MOS_{GQE}=1+0.035GRF+GRF(GRF-60)(100-GRF)7*10-6$
For $GRF > 100$: $MOS_{GQE}=5$

Table 4 shows the provisional guide for the relation between the *GRF*, $MOS_{GQE,}$ and user satisfaction.

MULTIPLAYER GAMING

Multiplayer gaming is becoming an increasingly important component of the Internet traffic mix, as evidenced by traffic measurements on Internet backbones (Joyce, 2000, McCreary, 2000). McCreary (2000) estimates that somewhere between 3% and 4% of traffic on one Internet backbone could be associated with just six popular gaming applications. Even this data might be drastically understated, as much like P2P traffic, multiplayer games traffic is difficult to identify based on simple port number mapping (Zander, Williams, & Armitage, 2006).

In an extensive survey of UK broadband subscribers, 48% of those surveyed listed "playing games" as a task for which they regularly use their Internet connection. This is only slightly less than the percentage who use their connections to download or listen to music, and greater than those who regularly use their connections for the purposes of online banking and downloading or watching videos. In a 10-month study of traffic at major Internet peering exchanges, games accounted for 14 of the 25 largest identifiable UDP applications (McCreary, 2000).

Wireless games are already considered to be a popular application globally, according to ITU-T (2002), with the wireless gaming market set to be worth in excess of $20 billion annually by 2010 (GIA, 2007). Europe is the world's largest wireless gaming market, and is estimated to account for approximately 43% of the global market in 2007. The remainder of the global market, comprising Canada, Asia-Pacific, Middle East, and Latin American regions, is projected to be the fastest growing wireless gaming market, with a compounded annual growth rate of circa 77% over a 10-year analysis period. The hosted server platform, offered by Vollee, allows for potentially any game to be played on a 3G handset by streaming a video of the game screen to the mobile. This vastly expands the range of available games without the difficulty of porting existing games or placing undue processing demands on the handset. It has been reported that the market for wireless gaming is developing in a manner similar to that of the games console market in the early nineties (Ward, 2005).

INTERNET AND ONLINE MULTIPLAYER GAMES

The most popular activities on the Internet today include e-mail, search, surfing, shopping, travel bookings, instant

messaging, listening to and downloading music, and playing games (Dutton, Gennaro, & Millwood 2005). Thus, it can be said that of the nine most popular things to do on the Internet, playing games is the only one where it is not possible to use a buffer to mitigate transient delay. Despite evidence that the demand for games grows alongside resource capacity and the high cost associated with providing sufficient bandwidth, many developers consider customer service to be the primary concern of games providers (Dutton et al., 2005). As defined by Veilleux (2002):

"The degree to which those companies need bandwidth is staggering. When a single server can host around 3,000 players at any time, and keeping in mind most of these online games often have many servers on which you can play, the cost must be sky high ... [but] the customer service issue is what the developers consider to be the most costly, and potentially the issue that could mean the difference between a loss and a profit ..."

The availability of Internet has provided an extensive infrastructure with global connectivity for the games industry to develop and deploy online games. Over the last 5 years, existent research reports revealed that online games have more severe requirements that are not fulfilled by the Internet's best effort model when compared with bi-dimensional Internet interfaces such as World Wide Web (WWW).

As indicated by Henderson (2003), multiplayer gaming on the Internet has proved enormously popular in spite of the lack of QoS guarantees. The many comments and discussions about "ping times" and "lag" on online gaming forums would suggest it is unlikely, but the hypothesis that users may actually be insensitive to network delays is subjectively evaluated by Henderson and Bhatti (2003). They monitored, over a 2-week period, user reactions to an advertised server latency of between 0 and 250ms, as well as introducing delays in this range, during the game, on a server with low advertised delay. The results revealed that delay had a significant impact upon a user's decision to join a server, but much less influence on a player's decision to leave a server. This occurred even though players indicated the delay was noticeable and their performance, as measured by frags per minute, decreased correspondingly.

Ward (2005) investigates the effects of both loss and delay on user performance in Unreal Tournament 2003 games. The main findings establish some bounds on typical values for loss and delay through analysis of public UT2003 servers. Loss was quite low, with a maximum reported loss of 3% and 80% of public servers reported no loss during the analysis period. Latency was less than 140ms for more than 80% of all servers measured. These values were then applied in a controlled experiment that demonstrated that packet losses in this range have no observable effect on player's performance. Latency however, of between 75ms and 100ms, resulted in significantly decreased player performance, as measured by observing shooting accuracy and the total number of frags. Users reported that once latency exceeded 75ms, the game developed a "sluggish" feel.

Evaluation of a car game by Pantel and Wolf (2002), finds that delays in excess of 100ms deteriorate gameplay and should be avoided. Results based on an evaluation of Counter Strike data (Farber, 2004) relating game lag (which includes nonnetworking delay) to subjective user data. These data are given in Table 5.

FUTURE TRENDS

The growth of gaming is a reflection of a worldwide trend, as many now predict that revenue from the games industry will soon exceed that of the US movie industry excluding DVD sales and foreign film rights. Combined video and computer games sales exceeded $7 billion in 2005. Research, conducted in 2006 by the US Entertainment Software Association (ESA), revealed that some 42% of all Americans had purchased or planned to purchase a game in 2006 (ESA, 2006). The same survey also demonstrated that 44% of frequent gamers play online, a figure which has risen from 19% in 2000. Another report, conducted by Joyce (2001), has shown that networked gaming, where multiple

Table 5. FPS game lag and its effect on user opinion

Lag	User opinion
<50 ms	Excellent gameplay
50ms-100ms	Good gameplay
100ms-150ms	Noticeably decreased gameplay
150ms-200ms	Significantly affected gameplay
>200ms	Intolerable gameplay

players play interactive games using network connections, are becoming an increasingly important component of the Internet traffic mix.

It is clear that, not only is the games industry undergoing substantial growth, but, within that industry, the importance of online gaming has been increasing. As broadband proliferation increases, and games development companies seek to extend the lifetime of their releases by encouraging online play, this trend seems likely to continue. This can be seen in the popularity of massively multiplayer online role playing games (MMORPG) like *World of Warcraft,* which has currently over 8.5 million subscribers. Worldwide software games sales are expected to reach $26 billion in 2010, with portable game sales due to hit $10 billion this year. The most popular first person shooter (FPS) game, *Counter Strike,* has, at any particular moment in time, an average of 85,000 concurrent players. In an extensive survey of UK broadband subscribers, 48% of those surveyed listed playing games as a task for which they regularly use their Internet connection (Hutton, 2005). This was only slightly less than the percentage using their connections to download or listen to music, and greater than those who used their connection for online banking and downloading or watching videos. It was nearly four times the percentage of those who used their Internet connections to make phone calls.

CONCLUSIONS

Interactive applications, such as online games, have certain requirements with respect to latency, jitter, and packet loss; usually referred to as quality of service (QoS) requirements. If the network cannot meet these service requirements, such impairments will impact negatively on the application user's experience. Over the last decade, game quality of service (QoS) has been, and still is, a challenging research area. Although there have been several standards and approaches published, the deployment of these technologies has been lacking.

This chapter detailed the network-based parameters mentioned (latency, jitter, and packet loss) and their impact on end user's experience. Measurement techniques for measuring these parameters were also given. A detailed review of existent standards and research advances into this research were presented. A new objective method for game quality assessment that could be used by both end-users and game providers was also presented.

REFERENCES

Dick, M., Wellnitz, O., & Wolf, L. (2005), Analysis of factors affecting players' performance and perception in multiplayer games. In *Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games, NetGames '05* (pp. 1-7).

Dutton, W. H., Gennaro, C., & Millwood A. (2005). *Oxford Internet survey 2005 report: The Internet in Britain*. Oxford: Oxford Internet Institute.

Entertainment Software Association (ESA). (2006). *Essential facts about the computer and video game industry.* Retrieved from http://www.theesa.com/archives/2006/05/2006_essential.php

Farber, J. (2002). Network game traffic modeling. In *Proceedings of the 1st workshop on Network and system support for games, NetGames '02* (pp. 53-57).

Farber, J. (2004). Traffic modeling for fast action network games. *Multimedia Tools Appl., 23*(1), 31-46.

GIA. (2007). *Wireless gaming: Global strategic business report*. Global Industry Analysts.

Henderson, T. (2003). *The effects of relative delay on networked games*. PhD thesis,

University of London, London.

Henderson, T., & Bhatti, S. (2003). Networked games: A QoS-sensitive application for QoS-insensitive users? In *Proceedings of the ACM SIGCOMM workshop on Revisiting IP QoS, RIPQoS '03* (pp. 141-147).

Hollier M. P., & Cosier G. (1996). Assessing human perception. *British Telecom technology, 14*(1), 206-215.

Hutton, 2005

International Telecommunication Union (ITU-T). (1994). *Terms and definitions related to quality of service and network performance including dependability*. *ITU-T Recommendation E.800*.

International Telecommunication Union (ITU-T). (1995). *The E-model; A computational model for use in transmission planning*. *ITU-T Recommendation G.107*.

International Telecommunication Union (ITU-T). (2002). *Internet for a mobile generation. ITU-T Internet Technical Report, 24*0.

Joyce, S. (2000). Traffic on the Internet; A study of Internet games. Retrieved from http://citeseer.ist.psu. edu/joyce-00traffic.html

Joyce, S. (2001). Traffic on the Internet – a study of Internet games. *Network Analysis Times, 2*(1), 6-8.

McCreary, S., & Claffy, K. (2000). Trends in wide area IP traffic patterns. In *ITC Specialist Seminar* (pp. 18-20).

Mullin, J., Henderson, A., Sasse, M. A. Jackson, M., Watson, A., Smallwood, L., & Wilson, G. (2001). *Assessment methods for assessing audio and video quality in real-time interactive communications*. Retrieved from http://www-mice.cs.ucl.ac.uk /multimedia/projects/etna/ assessment-methods.pdf

Pantel, L., & Wolf, L. C. (2002). On the impact of delay on real-time multiplayer games. In *Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video, NOSSDAV '02* (pp. 23-29).

Ubicom. (2005). Opscore, or online playability score: A metric for playability of online games with network impairments, In *Technical report, Ubicom Inc*.

Veilleux, M. E. (2002). *Ekim's gamer view: To pay, or not to pay*? Retrieved from http://www.mmorpgdot.com/index.php?hsaction=10053\&ID=400\&sid=1678270ec2b51c971a25ec09609f9f43

Ward, M. (2005). *Mobile games poised for take-off*. Retrieved from http://news.bbc.co. uk/2/hi/technology /4498433.stm

Wattimena, A. F. (2006). *Performance modeling of interactive gaming*. Master's thesis, Vrije Universiteit, Amsterdam, Netherlands.

Zander, Z., & Armitage, G. (2004). Empirically measuring the QoS sensitivity of interactive online game players. In *Australian Telecommunications Networks and Applications Conference, ATNAC'04.*

Zander, S., Williams, N., & Armitage, G. (2006). Internet archeology: Estimating Individual Application Trends in Incomplete Historic Traffic Traces. In *Passive and Active Measurement Workshop, PAM 2006* (pp. 30-31).

KEY TERMS

Best Effort (BE): BE denotes "best effort" service delivery, the prevalent level of service offered by the Internet. It is analogous to the postal system and does not provide guarantees for timeliness, order or delivery itself.

First Person Shooter (FPS): An FPS game is a type of action game where the player views everything from a first-person perspective. Such video games are typically characterized by a high level of violence, and a focus on the use of handheld ranged weapons. The objective is to shoot and destroy objects or other players. Examples include Quake, Counter Strike and Unreal Tournament.

ITU-T E-Model: A mathematical model recommended by ITU-T to be used in assessing the speech/voice quality.

Jitter: Jitter is an unwanted variation of one or more characteristics. Jitter may be seen in characteristics such as the interval between successive moments, or the amplitude, frequency, or phase of successive cycles.

Latency: A time delay between the moment something is initiated and the moment its first effect begins.

Mean Opinion Score (MOS): MOS provides a numerical indication of the perceived quality of received media after compression and/or transmission. The MOS is expressed as a single number in the range 1 to 5, where 1 is lowest perceived quality and 5 is the highest perceived quality.

Packet-Loss: Packet loss occurs when one or more packets of data travelling across a computer network fail to reach their destination.

Quality of Experience (QoE): QoE is a higher-layer more abstract term than QoS. It is used in reference to the user's perception of QoS, which may differ greatly depending on a vast number of subjective criteria.

Quality of Service (QoS): A term used to describe performance criteria within a network. The criteria is quite varied, ranging from a loss rate below a given percentage or a bounded delay. BE is generally the term used when there is no specified QoS.

The Impact of Risks and Challenges in E-Commerce Adoption Among SMEs

Pauline Ratnasingam
University of Central Missouri, USA

INTRODUCTION

E-commerce provides different opportunities to small businesses as it overcomes part of their technical, environmental, organizational, and managerial inadequacies (Bergeron, Raymond, & Rivard, 2001; Hussin, King, & Cragg, 2002). According to Forrester Research, e-commerce in the US will grow at 19% reaching $230 billion by 2008. Further, the Internal Revenue Service (IRS) estimated that in 2003, there were 27 million small business tax returns. Small businesses are an important and integral part of every nation's economy (Hambrick & Crozier, 1985). The US Small Business Administration (SBA) defines a small business as "an independent business having fewer than 500 employees or is independently owned and not dominant in its field of operation." Small firms play an increasingly crucial role in US economy. They employ more than one half of the US private sector work force, are responsible for about one-half of the GDP, and generate more than one half of all sales in the US create 60%-80% of net new jobs annually (Ibrahim, Angelidids, J. & Parsa, 2004).

Alternatively, small businesses are often more challenged than larger firms by resource constraints, such as lack of financial capital, and technical or managerial skills, knowledge and expertise that significantly reduce the number and types of options available to management (Hodgetts & Kuratko, 2001). Previous research suggests that although most small businesses were connected to the Internet, the potential use of the Internet in their business was rarely explored. Security concerns has a direct impact on every critical part of the small business including reputation, productivity, and business continuity, as they need to adhere to the legal requirement for information management. The research question thus designed for this study is what factors inhibit or pose challenges for e-commerce adoption among small businesses? We discuss the findings of an exploratory case study with four firms, across a section of different industries, on the risks and challenges they encountered when adopting e-commerce. The study contributes to managerial and theoretical implications by increasing the importance and awareness of small businesses in e-commerce adoption.

BACKGROUND

The rapid growth of the Web, and its importance to the US economy, make it imperative to develop a greater understanding about the challenges and barriers experienced by small businesses. Small businesses use the Internet mainly to send electronic mail messages, and to conduct e-commerce, including financial transactions. Other uses include communicating internally and externally, thereby sharing data; providing customer service and vendor support; purchasing and selling products and services; and collaborating with other businesses (Schneider & Perry, 2001; Turban, Lee, King, Chung, 2006).

Small businesses are typically characterized by a flat organizational hierarchy and close proximity to coworkers, thereby contributing to effective communication practices comprised of informal channels (Vinten, 1999; Wickert & Herschel, 2001). Their communications are typically carried out face-to-face as the need arises, rather than applying a formal standard operating procedure. Small businesses lack human resources and budgets allocated for information security management. Many small businesses think they are not at risk because of the size of their business information. They display the "if it's not broke, don't fix it attitude." Due to the lack of IT, staff leads to no one thinking about the IT security of small businesses, thereby leading to the lack of IT security policies. Further, they are often pressured by their large manufacturers to adopt e-commerce. Their systems are often vulnerable, and are related to unpatched systems that are improperly configured. The next section discusses factors that inhibit small business adoption.

Factors Inhibiting E-Commerce Adoption

The factors that inhibit e-commerce adoption stem from two perspectives, namely, technology-related factors and relationship-related factors. Risks refer to the possibility of an adverse outcome and uncertainty. Risks can be derived both internally and externally, thereby causing a concern to the smooth flow of e-commerce operations.

Technology-Related Factors

Technology related factors are derived from misuse of IT, viruses, and the lack of confidentiality, integrity, and availability mechanisms. Some of the risks include viruses, worms, spy ware, spam, phishing scams, hackers, and bot networks. They impact the compatibility, infrastructure, complexity, and uncertainties of e-commerce systems and operations.

Preadoption Negotiation and Integration Challenges

E-commerce adoption, unlike traditional information systems adoption, demands high levels of negotiation, cooperation, and commitment from participating organizations. Selecting transaction sets, negotiating legal matters, and defining performance expectations can burn up hours of employees' time and demand financial and technological resources (Senn, 2000). This becomes even more challenging for small businesses, as they lack the technical skills and knowledge to negotiate effectively. Previous studies suggest an internal fear of opening their organization's systems to suppliers, as implementing e-commerce could affect critical business processes including procurement, inventory management, manufacturing, order fulfillment, shipping, invoicing, payments, and accounting, as in difficult to measure success (Nath, Akmanligil, Hjelm, Sakaguch, & Schultz, 1998; Senn, 2000; Storresten, 1998).

High Implementation Costs

Startup costs for implementing e-commerce applications can be high. These include connection costs, hardware, software, set up, and maintenance (Iacovou, Benbasat, & Dexter, 1995; Nath et al., 1998). Implementation costs may also include conducting an initial search costs, costs of writing contracts, and paying staff to update and maintain electronic databases.

Lack of Standards and Policies

Due to the small size and physical environment, most small businesses operate with a lack of formal standards and best-known practices, which can lead to potential compromises in network controls, maintenance, data ownership, internal and external security, and permissions (Riggins & Rhee, 1998; Senn, 2000). For example, current methods of standardization for structuring data exchanged in extranet applications totally ignore how e-commerce applications were designed to operate. Most small businesses do not know what policies to set and many do not even have a complete security policy in place (Marcella, Stone, & Sampias, 1998). Lack of established standards, regulatory policies, and best business practices can impact effective business operations in e-commerce.

Technology Uncertainties and Security Concerns

The proliferation of e-commerce applications has left most small businesses uncertain of e-commerce operations and unaware of the full potential of e-commerce technology (Ghosh, 1998). Uncertainties arise when small businesses encounter barriers in communication (such as incompatible e-commerce systems, or lack of uniform standards) that may lead to conflicts.

Relationship-Related Factors

Relationship-related risks are derived from mistrust among small business partners. Trust is defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (Mayer, Davis, & Schoorman, 1995).They arise from a lack of experience, training, and a lack of technical knowledge about the security concerns, task uncertainties, environment uncertainties, false impressions of unreliability, and concerns about the enforceability of transaction records in the electronic trade area. They examine opportunistic behavior, conflicting attitudes, poor reputation, lack of training, and reluctance to change in business partners.

Competitive Market Pressure

Small businesses that form electronic partnerships between buyers and suppliers or manufacturers and distributors are subjected to competitive pressures in the global environment that demand quality. Iacovou et al. (1995) suggest that external pressures and organizational readiness may affect e-commerce adoption. For most organizations, the biggest challenge is not if or when to consider an Internet commerce solution, but rather how to select the best Internet commerce strategies to develop and sustain competitive advantage.

Lack of Trust

Security is one barrier, but the real underlying factor is insufficient trust in the reliability of e-commerce systems to absorb the rapid increase in use (Keen, 2000). Despite the opportunities of Internet commerce, many small businesses

are reluctant to go online because they perceive the Internet as an intrinsically uncertain insecure environment (Bhimani, 1996; Cavalli, 1995; Jarvenpaa et al 2002; Storrosten, 1998). Small businesses fear that their business transactions and operations may be hacked and disrupted from confidentiality, integrity, and availability security concerns.

Lack of Top Management Support

With poor internal management and the lack of top management commitment, small businesses, implementing e-commerce, experience challenges. If management is unwilling to provide adequate financial resources, poor business practices might follow. For example, without full support a small business might neglect the need for a paper audit trail that would ensure the reliability of electronic certification and business continuity. Further, the lack of top management involvement and commitment impacts the extent of participation, extent of coordination with employees, and their strategic rationale for e-commerce adoption (Alavi & Leidner, 2001; Chatterjee & Segars, 2001).

Lack of Technical Skills, Knowledge, and Expertise

E-commerce was in its formative stages in the mid 1990s (Norlan and Norlan, KPMG, 1999). Many small businesses lacked the necessary IT skills, resources, and technical know-how to implement policies and strategies for secure e-commerce. This is consistent with previous empirical research that suggests the lack of technical knowledge, expertise, and resources hindered e-commerce adoption among small businesses (Heck & Ribbers, 1998; Iacovou et al., 1995; Reekers & Smithson, 1996).

Findings

An exploratory multiple case study approach was applied to examine the factors in four small business firms. According to Yin (1994), a research design should incorporate research questions including "how" and "why" types of questions.

The interviews were conducted applying a semistructured questionnaire for implementing and maintaining the security of the e-commerce system. The interviews lasted for 90 minutes, and two visits per firm were made. Four small businesses from a cross section of different industries, namely, a retail import-export firm, transportation firm, manufacturer, and a real estate firm participated in this study. Brief background information of each firm is given followed by Table 1, which summarizes the demographic information of the firms.

Firm A

Firm A is an import/export trading firm that conducts its business on a Web site developed in Nov 2003. They sell unique accessories including jewelries, artwork, handbags, and unique children books in Spanish and English. They aim to cater the needs of the women, gay, and young children in the Spanish and African American communities.

There have two full-time employees and their main business transactions include e-mail, fax, and telephone. They use QuickBooks, which takes in the orders and manages the shopping cart outsourced and managed by an IT solutions provider. The firm selects their suppliers for the variety of products via word of mouth.

Firm B

Firm is a transportation broker located in the town of Warsaw in Missouri with five employees in their branch office, paid on commission basis. They match trucks with freight that load goods across the US, in particular, the southeastern states of Mississippi (head office), Alabama, and Georgia. The Gulf States Paper or Georgia Pacific papers are their main shippers. The firm applies a data-entry software, called "Expressions," that records each order and matches a truck for it. They use e-mail extensively and have six phone lines. They also use the Internet to choose their truckers and offer low freight rates in order to stay competitive. They have 6,000 trucking companies in their database and use between 200-400 on a regular basis. Further, they have 20 shippers who use extranets. Some of their truckers use a special system, called Qual.com, that helps them to track where the truck is and when the goods will be expected to be delivered. Firm B mainly communicates via e-mail, fax, and telephones. They also do have a few shippers that use extranets. Their revenue each year is between $1-$10 million annually, and they spend about $5,000 on their IT computer systems as they have brand new computers.

Firm C

Firm C is an equipment supply and plastic packaging manufacturer located in Harrisonville (Missiouri). They have 15 full-time employees and sell equipment to global customers via their B2C Web site implemented in 2001. They manufacture and distribute equipment and plastic bottles on the requested size, shape, and color for their customers. They have 10 customers for the plastic bottles business and 50 customers each year on their B2C Web site, which sells specialized equipment. Their main business is conducted via the phone, fax, and e-mail for the plastic business and

the Internet for the equipment business. Their business documents include purchase order, invoices, quotations, and e-mail acknowledgements.

Firm D

Firm D is a real estate broker who has been in business for 25 years and owned this company for 20 years. They sell residential houses, farms, and commercial property primarily to the local area and to national customers who are mostly in the military. They have nine full-time employees. Their business transactions include closing agreement and foreclosure agreement. Most of the other documents are between the buyer and the bank. They rely on the lending institutions to take care of the financial process. Once the loan is approved, they order the title insurance and conduct a termite inspection. Their e-commerce system includes the multiple listing service system, implemented since 1990, that has all the information about the property, and is used by all realtors to network. It is crucial to their business strategy. In addition, they use the e-mail, fax, and telephone extensively. Through their MLS system, they have the voyager processor, which is the electronic tool used to facilitate the preparation of electronic contracts.

Table 1 presents the demographic information of the four firms.

Technology-Related Factors

Download delays and search problems from incompatible systems or lack of system functionality was not an issue for most of the firms, as they had an IT person who would fix their system if they had problems. Further, they had other ways of communicating with their customers such as the telephone and fax to do business if their network was down. Firm A spends nearly $40,000 for the initial IT implementation costs. Their annual maintenance costs of the Web site vary from $2,000-$5,000.

The manager of Firm A stated that the functionality of the Web site nor the user friendliness was important as they were able to observe the Web traffic as to how many customers visited their Web site. Firm B's computer does crash and they will call their IT service provider to come and fix it. The manager of Firm B noted: "In this business the better the tools are they make you the money because the system is able to provide you with timely accurate information that in turn enables you to make profit. Further, knowing our providers was important as we know our shippers very well in terms of what products they want us to transport. We need to know what the product is as we need special licenses and the driver needs special licenses." Firm B indicated that it was difficult to measure the success of their IT implementation as they moved from the old to the new system. The old

Table 1. Demographic information of the four firms

Characteristics	Firm A	Firm B	Firm C	Firm D
Type of Industry	Retail Marketing	Transport Freight	Plastic Packaging and Equipment Manufacturer	Real Estate broker
Number of Employees	2	15	15	7
Years Using IT/E-Commerce solution	Since November 2003	2001	2001	1990
Type of IT system	Internet B2C shopping cart	Internet, E-mail	B2C, Phone, fax and e-mail	Internet E-mail
Types of business transactions	Purchase order, Invoices, shipping notices	Purchase order, Invoices Shipping Notices Bills of Lading Payment Info	Purchase order, Invoices Quotations E-mail acknowledgments	Closing agreement, Foreclosure
Product Focus	Accessories – jewelries, artwork, handbags		Plastic bottles & Specialized equipment	Residential houses, farms, and commercial properties
Organizational Reach	Global National	National	National Global	Regional National

system had the complete database and 6,000 filters not in the new system. They are working on it.

Firm D's business partners include four local lenders and bankers. Their Web site was shared by the MLS (Mortgage Listing Service) as all can see it. Further, security was not a concern because each user is required to use a log-on user id and a password to get into the MLS system. Firm D spends about $2,000 to $5,000 each year to upgrade and maintain our IT systems.

The manager of Firm D noted: "In the past we had an MLS but it had a lot of redundant information. For example 5 different realtors will list the same house in the listing. It was a nightmare for the seller to keep track of who had listed and who did not. Firm D indicated that everything is legal and looks the same no matter which realto the seller or the buyer goes to."

Competition from traditional brick and mortar firms going online was not a concern as most of the firms sold specialized goods and services that served their customers locally and regionally. Complexity costs, training, and implementation costs was a concern for Firm B. The manager stated: "The main challenges that the firms experienced included; was the lack of IT skills as they had to rely on an IT solution provider to customize their systems when it was down."

Firms C and D had customized software and solutions that were not connected to the Internet but rather to intranets that had various log-in authorizations for their different employees.

Relationship-Related Factors

The impact of top management support for all firms was low as most of the firms were family owned business; hence, this was not a concern. The manager of Firm A indicated: "When you open a Web site business it is difficult for suppliers to know the size of your firm. Our suppliers thought that our business was big. A large image was impressive although in reality there were only two partners."

Issues pertaining to trust and uncertainties were not a concern for Firms C and D, as they maintained a long-term relationship with their business partners. Firm B indicated that most of their business partners are derived from word of mouth and is based on past reputations. They adopted the e-mail system in 1995 and found that it was crucial to their business strategy.

In order for Firm B to stay competitive, they had to provide good freight rates, sign contracts, thereby, providing customer satisfaction. Demands from larger trading partners related to timely delivery pickups. Firm B manager indicated: "We do not charge them if the shipper delays on the product. If they cancel on us and do not pick up the load there is nothing we can do. A bad experience like that will affect

Table 2

Factors that Inhibit IT/EC Adoption ***Technology-related Factors***	**Firm A**	**Firm B**	**Firm C**	**Firm D**
Competitors have also adopted new IT	L-3	M-4	M-3	L-4
Limited flexibility due to resource constraints	L-2	M-4	M-3	L-3
Difficult to measure success of IT implementation	L-3	M-2	L-3	M-5
Inability to sustain e-initiatives	L-2	M-4	M-4	L-3
IT sophistication–resistance to change	L-2	M-2	M-4	L-3
Functionality of the Web site–difficult and not user friendly	L-2	M-3	L-2	L-3
Concerns over security and reliability of the system and the stakeholders	L-3	L-2	L-3	L-1
Difficult to integrate	L-3	L-3	L-3	L-3
Incompatible systems	L-2	L-3	L-2	L-2
Increased implementation costs	M-5	M-4	L-3	L-3
Relationship Factors				
Complexity costs, costs of training	L-2	L-2	L-3	L-2
Lack of financial resources	L-3	M-4	M-5	L-2
Lack of IT/IS skills	L-3	M-4	M-5	L-3
Lack of trust from external IT sources	L-3	L-2	L-3	L-2
Level of top management support	L-2	M-4	L-2	L-3
Market pressure	L-2	M-5	L-3	M-5

our business and usually give three chances. For example, when the trucker loads with the wrong product to the wrong place. We have to correct any mistake even if we have to spend more. It could be the drivers fault, shippers fault or the guy who loaded it. Papers can go spoilt if the van has a leakage. Usually the shipper will find a way to sell it as scrap so that the entire load is not wasted."

Market pressure was a concern for Firm C as the manager mentioned that although China and India could provide the specialized machinery items cheaper, it was seen as our weakness. Firm C has established a very strong relationship with their regular customers and hence, were able to compete effectively with the overseas market. Firm D encourages a team concept and so, hires only seven employees. We work really hard as a leader and discourage interoffice competition.

Impact of security risks and challenges experienced by the four firms were measured using a Likert scale where Low (L= 0-3), Medium (M=4-6), and High (H=7-10). Table 2 presents the summary findings based on the impact on factors that inhibit small businesses e-commerce adoption via a Likert scale.

FUTURE TRENDS

In this section, we discuss the observations of the findings as lessons learned and future trends of this study. Although most of the firms that participated in this study were not established and matured in the use of the e-commerce systems, as in an active shopping cart, it was found that most of the small businesses were operating using e-mails, extensively, to circulate documents, notes, attach invoices, financial reports, design documents, purchase order acknowledgments, and other business documents.

The firms did not undertake an evolutionary process when implementing their e-commerce systems but rather, implemented them as the need and demand arose, that is, they took a reactionary approach. This is because small businesses do not have the necessary financial, technical IT skills, knowledge, and strategic resources to build their system in an evolutionary fashion. Rather, most of them have outsourced their IT business operations to Web developers, IT consultants, or even subcontracted part of their IT operations to another firm. This indicates their resistance to change, to expand, educate, and increase their awareness of the full potential of the e-commerce technologies.

Further, small businesses were very cautious of sustaining their revenue in order to meet their ongoing overhead expenses such as utilities, wages, salaries for their current employees, cost of purchasing inventory, and other current operating expenses. In essence, they were not big risk takers. The Internet assists small businesses globally, but a physical store helps to build reputation and trust locally, and within suppliers and future trends of most small businesses is to maintain and upgrade their Web sites, which, in turn, will be reflected in their business operations.

CONCLUSIONS

In this study, we examined the factors that inhibit and pose risks and challenges to e-commerce adoption among small businesses. We tested an exploratory case study questionnaire with four small business firms from a cross section of different industries. The findings suggest the importance of adopting e-commerce to maintain a competitive and strategic advantage. It increased the awareness and importance of e-commerce adoption among small businesses. Future research should aim to examine these factors extensively. Small businesses need to have a brick and mortar office to do e-commerce because there are so many scamps going on.

REFERENCES

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, *25*(1), 107-136.

Bergeron, F., Raymond, L., & Rivard, S. (2001). Fit in strategic information technology management research: An empirical comparison of perspectives. *Omega*, *29*(2), 125-142.

Bharadwaj, P., Nagendra, S., & Ramesh G. (2007). E-commerce usage and perception of e-commerce issues among small firms: Results and implications from an empirical study. *Journal of Small Business Management*, *45*(4), 501-521.

Bhimani, A. (1996). Securing the commercial Internet. *Communications of the ACM, 39*(6), 29-35.

Brunetto, Y., & Farr-Wharton, R. (2007). The moderating role of trust in SME owner/managers' decision-making about collaboration. *Journal of Small Business Management*, *45*(3),p362-387.

Cavalli, A. (1995). Electronic commerce over the Internet and the increasing need for security. *TradeWave*.

Chatterjee, D., & Segars, A. H. (2001). *Transformation of the enterprise through e-business: An overview of contemporary practices and trends*. Report to the Advances Practices Council of the Society for Information Management, July.

Ghosh, S (1998) Making Business sense of the Internet. *Harvard Business Review, March-April*, 126-135

Heck and Ribbers, (1999) The adoption and impact of edi in dutch SMEs. The *Hawaii International Conference in Information Systems.*

Hodgetts, R. M.., & Kuratko, D. F. (2001). *Effective small business management*. Ft. Worth, TX: Harcourt College.

Hussin, H., King, M., & Cragg, P. (2002). IT alignment in small firms. *European Journal of Information Systems*, *11*(2), 108-127.

Iacovou, C. L., Benbasat, I., & Dexter, A. S. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. *MIS Quarterly, 19*(4), 465-485.

Ibrahim, N. A., Angelidids, J. P., & Parsa, F. (2004). The status of planning in small busineses. *American Business Review*, 52-60

Jarvenpaa, S. L., Tractinsky, N., & Vitale, M. (2000). Consumer trust in an Internet store: Information technology and management. *Information Technology and Management, 1*, 45-71.

Johnston, D. A., Wade, M., & McClean, R. (2007). Does e-business matter to SMEs? A comparison of the financial impacts of Internet business solutions on European and North American SMEs. *Journal of Small Business Management*, *45*(3), 354-361.

Keen, 2000

Marcella, A. J., Stone, L., & Sampias, W. J. (1998). *Electronic commerce: Control issues for securing virtual enterprises*. The Institute of Internal Auditors.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734.

Nath, R., Akmanligil, M., Hjelm, K., Sakaguch, T., & Schultz, M. (1998). Electronic commerce and the Internet: Issues, problems and perspectives. *International Journal of Information Management*, *18*(2), 91-101.

Norlan and Norton Institute, KPMG. (1999). *Electronic commerce – The future is here.*

Reekers, N and Smithson, S (1996) The role of EDI in inter-organizational coordination in the European automotive industry. *European Journal of Information Systems, 5*, 120-130.

Riggins, F. J., & Mukhopadhyay, T. (1999). Overcoming EDI adoption and implementation risks. *International Journal of Electronic Commerce*, *3*(4), 103-123.

Riggins, F.J., and Rhee, H.S (1998) Toward a unified view of electronic commerce. *Communications of the ACM, 41*, 10, 88-95.

Senn, J. A. (2000). Business to business e-commerce. *Information Systems Management*, 23-32.

Smith, C. W. (2007). On governance and agency issues in small firms. *Journal of Small Business Management*, *45*(1), 176-178.

Street, C. T., Cameron, A-F. (2007). External relationships and the small business: A review of small business alliance and network research. *Journal of Small Business Management*, *45*(2), 239-266.

Storrosten, M. (1998). Barriers to electronic commerce. *European Multimedia, Microprocessor Systems and Electronic Commerce Conference and Exhibition*, Bordeaux, France.

Turban, E., Lee, J., King, D., & Chung, H. M. (2006). *Electronic commerce: A managerial perspective*. Prentice Hall Inc.

Vinten, G. (1999). Corporate communications in small- and medium-sized enterprises. *Individual and Commercial Training*, *31*(3), 112-119.

Wickert, A., & Herschel, R. (2001). Knowledge management issues for smaller businesses. *Journal of Knowledge Management*, *5*(4), 329-337.

Yin, R. K. (1994). *Case study research: Design and methods*, 2nd ed. Thousand Oaks, CA: Sage Publications.

KEY TERMS

E-Commerce: Refers to buying and selling over the Internet

Relationship-Related Factors: Refer to risks derived from mistrust among business partners.

Risks: Refer to possibility of an adverse outcome and uncertainty.

Small Business: Refers to an independent business having fewer than 500 employees or is independently owned and not dominant in its field of operation and the lack of confidentiality, integrity, and availability mechanisms

Trust: The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party.

Impediments for Knowledge Sharing in Professional Service Firms

Georg Disterer
University of Applied Sciences and Arts, Germany

INTRODUCTION

Professional service firms (PSFs), where professionals (consultants, lawyers, accountants, tax advisors, etc.) work, are interested in knowledge management because their businesses are heavily dependent on the knowledge of their employees. A core asset is their ability to solve complex problems through creative and innovative solutions, and the basis for this is their employees' knowledge. The "product" that PSFs offer their clients is knowledge (Kay, 2002; Ofek & Sarvary, 2001; Chait, 1999).

Sharing knowledge between colleagues improves the economical benefits a firm can realize from the knowledge of employees. This is especially true for PSFs (Huang, 1998; Quinn, Anderson, & Finkelstein, 1996), where broad ranges of knowledge must be kept to provide intellectual services, and real-life experiences with certain questions and situations are an important asset. The organizations and its members are spread over various offices across the country or the world. The necessity for sharing grows because the network of professionals in most cases can offer significantly better professional advice than any individual. "We sell knowledge…the most valuable thing we can offer is the collective, institutional knowledge of our firm" (Roger Siboni, KPMG executive, in Alavi, 1997, p. 1). Working together openly without holding back or protecting vital pieces of knowledge will result in more productivity and innovation than could be reached individually.

BACKGROUND

No professional is denying the worth of using working documents and materials produced by others. All PSFs are trying to set up collections of knowledge acquired in projects in order to share it and conserve it for reuse. Knowledge databases can address what is sometimes called the traditional weakness of PSF: "…narrow specialists who see only their own solutions, self-centered egoists unwilling or unable to collaborate with colleagues" (Liedtka, Haskins, Rosenblum, & Weber, 1997, p. 58). Many authors signal that sharing knowledge seems to be "unnatural" (Quinn et al., 1996; Barua & Ravindran, 1996; Holloway, 2000).

However, attempts to use knowledge databases often fail. Only a few databases are accepted as up to date. The special fields of expertise are covered only in fragments. The access is laborious and uncomfortable. Heterogeneous sources (text, internal and external databases, journals, books, comments, codes of law, and so forth) cannot be integrated. The lack of actuality and completeness causes quality risks if dealt with thoughtlessly and if not reflected upon.

People issues are meant to be critical for successful knowledge sharing. According to Ruggles (1998), "In fact, if the people issues do not arise, the effort underway is probably not knowledge management. If technology solves the problem, yours was not a knowledge problem" (p. 88). Therefore, we analyze the reasons why knowledge sharing needs dedicated efforts and describe possible actions to foster knowledge sharing. Through our research (Disterer, 2000, 2001, 2002a) and analyses drawn from literature, we categorize and discuss the various impediments encountered by people sharing knowledge (see Figure 1). There are some empirical results that confirm these impediments (APQC, 1996; Ruggles, 1998; KPMG, 2003; Govindarajan & Gupta, 2001M; Hooff & Ridder, 2004; Lee, Kim, & Kim, 2006). Then we show various approaches to overcome these impediments.

IMPEDIMENTS TO KNOWLEDGE SHARING

Loss of Power

Knowledge can be used to take action and to enforce spheres of influence. Passing knowledge to colleagues might grant some of this potential. Those who do not have this knowledge are deprived of the capacity to act or to influence. That applies for knowledge about customers, competitors, suppliers, procedures, recipes, methods, formulas, and so forth. In this sense, someone who passes on knowledge to a colleague loses the exclusiveness of his or her influence, which might have suggested some professional respect and job security. "Knowledge is power" is the well-known citation to describe situations in which experts with rare knowledge have the highest reputation, and monopolies of knowledge causes knowledge hoarding instead of knowledge sharing (Reimus, 1997; Andrews, 2001; Kankanhalli, Tab, & Wei, 2005).

In industries like professional services, employees are competing directly with each other through their special

knowledge, gifts, and talents. It might be part of the individual culture of high-performing employees that they voluntarily enter into the competition for scarce seats on their career paths because they like to compete and excel (Quinn et al., 1996). But, the drawbacks of competition are obvious: knowledge workers would be cautious to share their knowledge, because they could possibly give up an individual lead.

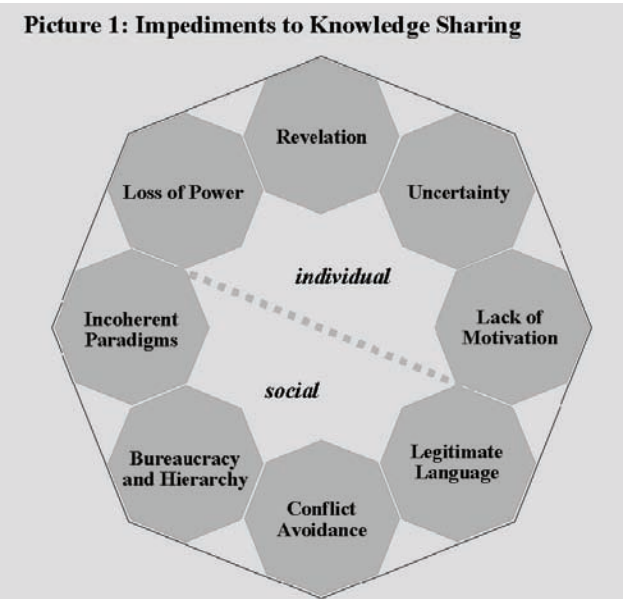
Revelation

Passing on knowledge to colleagues or entering working results into a knowledge database may be considered a revelation because it proclaims that this knowledge has a certain value and rarity. If this assessment is not shared by others, embarrassment may result (Rodwell & Humphries, 1998). Additionally, hasty colleagues rush to suggest "necessary" improvements to emphasize their expertise. For an individual, knowledge justified as "true belief" is not of particular concern. But in situations of knowledge sharing, more than one individual is involved. At this point, "…justification becomes public. Each individual is faced with the challenge of justifying his true beliefs in presence of [an]other" (Krogh, 1998, p. 35).

Uncertainty

Less-experienced colleagues may feel uncertain, because they cannot judge if their working results and experiences represent valuable knowledge for others. They cannot estimate if their knowledge is too general or too well known or, on the other side, that some results are too specific for a special situation and therefore useless for colleagues in other situations. Positioning on the scale of "general" to "specific" is not trivial and thus results in uncertainty.

Figure 1. Impediments to knowledge sharing



Lack of Motivation

Sharing knowledge is often seen as additional work because of the time necessary for reflection, documentation, communication, and so forth. Time is scarce, especially if the performance of an organization is measured by billable hours only. Reflection of work and sharing experiences are more an investment for future work than a billable action in the present. As stated in Dixon (2000), "In an organization with a bias for action, the time for reflection may be hard to come by" (p. 18; Hunter, Beaumont, & Lee, 2002).

Some employees do not expect reciprocal benefits from sharing, because they do not believe in these benefits or they did not experience it. Benefits of contributing to a knowledge database are gotten by a different stakeholder later on—the benefits will not be earned by the provider but by others (Nissen, Kamel, & Sengupta, 2000). Therefore, one precondition for contributing is the assumption of an equilibrium—a balanced give and take between colleagues. The insight that knowledge sharing can only be beneficial if everybody provides knowledge unselfishly may have charm only theoretically. In day-to-day practice, the benefit is too uncertain, and payback is not going to be immediate; therefore, the individual's commitment to share knowledge fails.

Legitimate Language

Some organizations lack a legitimate language (Krogh, 1998) that is known and accepted by all colleagues and can carry individual knowledge. This covers the need for a common language to communicate analogies and metaphors to externalize tacit knowledge hidden in individual mental models, viewpoints, working models, schemata, paradigms, and beliefs (Nelson & Cooprider, 1996; Nonaka, 1994; Haldin-Herrgard, 2000).

Conflict Avoidance

Attitudes of conflict avoidance and some conservative habits may prevent knowledge sharing, if the knowledge contains some new thoughts or innovative ideas. If most executives of an organization are not comfortable with change and are not willing to take risks, new ideas may be covered up easily. Different views and perspectives would be hidden, and knowledge not culturally legitimated may be suppressed ("do not rock the boat" attitude).

Bureaucracy and Hierarchy

Bureaucratic and hierarchical organizations show formal and administrative procedures that prevent the sharing of knowledge and new ideas. Strong hierarchical organizations prevent cross-functional communication, cooperation, and knowledge sharing.

Incoherent Paradigms

A lack of alignment between the personal intents of the individuals and the paradigms of the organization (strategic intent, vision, mission, strategies, values, etc.) can cause difficulties in articulating and justifying personal beliefs that do not fit with the ruling paradigms of the firm (Krogh, 1998).

ACTIONS TO FOSTER KNOWLEDGE SHARING

There is no complete methodology and no set of procedures and policies to address systematically all of the above impediments to knowledge sharing. Various approaches should be discussed further. Some of them sound like common sense but need to be emphasized because of their importance.

Concern and Trust

A precondition for knowledge sharing within organizations is an attitude of concern and trust among members of the organizations. Krogh (1998) called this "care" and defined it as serious attention, a feeling of concern and interest within an organization. His concept includes phenomena like trust among the people, interest in different viewpoints and experiences, access to help, lenience in judgment, courage to voice opinions, courage to allow experiments, and courage to take risks.

Organizations must strive for a culture of accepting mistakes (Soliman & Spooner, 2000). They should not penalize errors to foster a climate of constructive conflicts, giving members the chance to "fall forward." Organizational development processes should build and establish a common set of ethical standards and values for an organization and should achieve a consensus of accepted working practices and habits.

Leadership

Knowledge sharing is based on the consistent, reliable, and plausible behavior of management. Members of management must positively communicate that they are thoroughly convinced that knowledge needs to be "nurtured, supported, enhanced, and cared for" (Nonaka & Konno, 1998, p. 53) and that they financially support knowledge management initiatives. Management must afford time for communication and reflection. There must be organizational slack that permits time for employees to network (Krogh, 1998; Wiig, 1997).

To openly share knowledge, mutual trust is necessary among all organization members. Trust results in common expectations of reliability, consistency, and plausibility. Trust reduces the fear that others will act opportunistically. Likewise, management must act as examples for knowledge sharing. They have to walk-the-talk and give up knowledge hoarding first. Members of a profession or a community accept standards of behavior and working habits from their peers (Quinn et al., 1996); therefore, management must act as peers to be an example in knowledge sharing (McDermott & O'Dell, 2001).

Rewards and Incentives

Special rewards and incentive methods can act as extrinsic motivation for employees willing to share knowledge. Organizations are successful with the provision of personal recognition and reputation when people have contributed to knowledge databases or actively participated in knowledge sharing (Hunter et al., 2002).

Some examples for direct rewards and how to provide chances to build reputation and fame include the following: Texas Instruments created an annual award named "Not Invented Here, But I Did It Anyway Award" (Dixon, 2000, p. 57) to reward usage of other employees' knowledge. Buckman Labs rewards the top 150 "knowledge sharers" (judged by knowledge managers) with a laptop and an incentive trip to a resort (Davenport, Long, & Beers, 1998). AMS honors contributors to the knowledge center with a bronze plaque at the headquarters and regularly publishes a top 10 list of most frequently used contributions (King, 1998). Forum, a consultancy in Boston, Massachusetts, holds a "World Cup Capture" to encourage its consultants to make explicit and sharable what they have learned from their latest engagements (Botkin, 1999). An Australian law firm honors individuals who contributed the most by having a star named after them (Robertson, 1999).

Contrary to this, there might be professions with different views: "A major concern of software engineers…is the fear of being known as an expert" (Desouza, 2003, p. 100). They fear being staffed to projects based on their past experience instead of being allocated more challenging tasks with room for learning. In this situation, the brand-like identity of a software engineer works to his or her disadvantage and builds a barrier to individual professional development.

Many organizations incorporate issues of knowledge sharing into their compensation plans and promotion policies. The big consulting and accounting firms commonly base their personal evaluations partly on how many contributions are

made to knowledge databases, how many new employees people have tutored, and how many training courses have been designed (Quinn et al., 1996; Whiting, 1999).

Tutoring and Mentoring

Administrative actions may define responsibilities for tutoring and mentoring in an organization. Ongoing programs that systematically develop employees can foster common habits and attitudes and can support communication among the members of the organization.

Project Experiences

At the end of bigger projects and transactions, time and effort for explicit debriefing should be provided to learn systematically by experience. The lessons learned could be systematically analyzed and stored for access by other employees. In other actions, it can help to use knowledge and experiences gained in projects (Disterer, 2002b).

Communities of Practice

A popular approach for fostering knowledge sharing is to develop communities of practice. These groups of professionals enhance the ability of their members to think together, to stay in touch with each other, and to share ideas with each other. These informal networks, sometimes also called knowledge fairs or clubs, competence centers, or creativity centers, consist of groups of professionals, informally bound to one another through a common class of interests and problems and a common pursuit of solutions. People who are exposed to a common class of interests and problems often develop a common language with which to communicate and develop a sense of mutual obligation to help each other (Manville & Foote, 1996; McDermott, 1999). These phenomena can be used to overcome some of the individual and social barriers to knowledge sharing within communities of practice.

Focus on Codification or Personalization

In PSFs, the knowledge of experts is a core asset, and therefore, careful management of this asset is important. Management is responsible for ensuring that the organization is as independent as possible from individuals. At the same time, these companies are operating in a "people business," where the personal and individual link between clients and professionals is critical (Morris & Empson, 1998). This special situation requires special approaches to manage knowledge. Quite popular are two approaches that consulting firms apply that address cultural issues differently (Hansen, Nohria, & Tierney, 1999).

One strategy ("codification") centers on information technology (IT): the knowledge is carefully codified and stored in knowledge databases and can be accessed and used by others. With the other strategy ("personalization"), knowledge is tied to the person who developed it and is shared mainly through direct person-to-person contact (Hansen et al., 1999). With a codification strategy, knowledge is extracted from the person who developed it, is made independent from the individual, and is stored in the form of interview guides, work schedules, checklists, and so forth. Knowledge is then searched and retrieved and used by other employees. Personalization focuses on dialogue between individuals; knowledge is shared primarily in personal meetings and in one-on-one conversations.

Individual barriers are significantly lower with a personalization strategy, because professionals keep control through the whole knowledge management cycle. The individual is recognized as an expert and is cared for. In fact, focusing on personalization could be called a communication strategy, because the main objective is to foster personal communication between people. Core IT systems are yellow pages (directories of experts, who-knows-what systems, people finders) that show people with whom they should discuss special topics or problems.

Organizational Design

Some organizational designs can foster intra-organizational collaboration. Partnerships and other forms of ownership by employees can be utilized to produce involvement and commitment (Hildebrand, 1994; Miles, Miles, Perrone, & Edvinssen, 1998). Moreover, these organizational forms address the hesitation of professionals with specialized knowledge to work within strong hierarchies and in working environments with strong regulations (Quinn et al., 1996).

Office Design and Construction

To lower the disadvantages of bureaucracy and formal communications, modern office layouts reduce the distance between colleagues to foster ad hoc, informal, and face-to-face communication.

FUTURE TRENDS

Further research will be necessary to understand barriers to knowledge sharing, because cultural barriers are dominant over technical problems while implementing knowledge management initiatives. The connections between corporate culture and corporate climate and knowledge sharing within firms need more research for better understanding. Today, appropriate reward and incentive systems and well-established communities of practice are the most promising ways to overcome the barriers.

Inter-organizational knowledge sharing will become more important when larger and more complex projects will be distributed among several professional service firms. Sharing knowledge across organizational boundaries will raise questions of exchanging important knowledge while keeping business secrets.

CONCLUSION

Ways to support knowledge management with IT are manifold, but certain cultural aspects of knowledge sharing must be addressed. We describe some possible actions to overcome typical resistance often articulated with phrases like "this is client confidential," "only I know how to use it," "what's in it for me?" and "I have no time to document my experiences." The descriptions of the impediments and the possible actions make clear that knowledge management in professional service firms could not be seen as a technical field, as it is deeply social in nature. Corporate culture and corporate climate are important factors to foster—or hinder—knowledge sharing. Till now, there is no complete methodology and no set of procedures and policies to address all issues systematically, therefore managerial actions are described that can lower impediments of knowledge sharing.

REFERENCES

Alavi, M. (1997). *KPMG Peat Marwick U.S.: One great brain.* Case No. 9-397-108, Harvard Business School, USA.

Andrews, D. (2001). Knowledge management: Are we addressing the right issues? *Managing Partner, 4*(1), 23-25.

APQC (American Productivity and Quality Center). (1996). *Knowledge management—Consortium benchmarking study final report.* Houston, TX: Author.

Barua, A., & Ravindran, S. (1996). Reengineering information sharing behavior in organizations. *Journal of Information Technology, 11*(3), 261-272.

Botkin, J. (1999). *Smart business: How knowledge communities can revolutionize your company.* New York: The Free Press.

Chait, L.P. (1999). Creating a successful knowledge management system. *Journal of Business Strategy, 20*(2), 23-26.

Davenport, T.H., Long, D.W., & Beers, M.C. (1998). Successful knowledge management projects. *Sloan Management Review, 39*(4), 43-57.

Desouza, K.C. (2003). Barriers to effective use of knowledge management systems in software engineering. *Communications of the ACM, 46*(1), 99-101.

Disterer, G. (2000). Knowledge management—barriers to knowledge databases in law firms. *Managing Partner, 2*(3), 24-27.

Disterer, G. (2001). Individual and social barriers to knowledge transfer. In R.H. Sprague (Ed.), *Proceedings of the 34th Hawaii International Conference on System Sciences* (pp. 1-7). Los Alamitos, CA: IEEE Computer Society Press.

Disterer, G. (2002a). Social and cultural barriers for knowledge databases in professional service firms. In D. White (Ed.), *Knowledge mapping and management* (pp. 124-130). Hershey, PA: IRM Press.

Disterer, G. (2002b). Management of project knowledge and experiences. *Journal of Knowledge Management, 6*(5), 512-520.

Dixon, N.M. (2000). *Common knowledge: How companies thrive on sharing what they know.* Cambridge, MA: Harvard University Press.

Govindarajan, V., & Gupta, A.K. (2001). Building an effective global business team. *Sloan Management Review, 42*(4), 63-71 .

Haldin-Herrgard, T. (2000). Difficulties in diffusion of tacit knowledge in organizations. *Journal of Intellectual Capital, 1*(4), 357-365.

Hansen, M.T., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge. *Harvard Business Review,* (2), 106-116.

Hildebrand, C. (1994). The greater good. *CIO Magazine, 8*(4), 32-40.

Hooff, B., & Ridder, J.A. (2004). Knowledge sharing in context: The influence of organizational commitment and communication climate and CMC use on knowledge sharing. *Journal of Knowledge Management, 8*(6), 117-130.

Holloway, P. (2000). Sharing knowledge—and other unnatural acts. *Knowledge Management Magazine,* (1).

Huang, K.-T. (1998). Capitalizing on intellectual assets. *IBM Systems Journal, 37*(4), 570-583.

Hunter, L., Beaumont, P., & Lee, M. (2002). Knowledge management practice in Scottish law firms. *Human Resource Management Journal, 12*(2), 4-21.

Kankanhalli, A., Tan, B.C.Y., & Wei, K.-K. (2005). Contributing knowledge to electronic knowledge repositories: An empirical investigation. *MIS Quarterly, 29*(1), 113-143.

Kay, S. (2002). *Benchmarking knowledge management in U.S. and UK law firms.* Retrieved August 15, 2002, from *http://www.llrx.com/features/benchmarkingkm.htm*

King, J. (1998). Knowledge management promotes sharing. *Computerworld,* (June 15).

KPMG. (2003). *Insights from KPMG's European Knowledge Management Survey 2002/2003.* Author.

Krogh, G. (1998). Care in knowledge creation. *California Management Review, 40*(3), 133-153.

Lee, J.-H., Kim, Y.-G., & Kim, M.-Y (2006). Effects of managerial drivers and climate maturity on knowledge-management. *Information Resources Management Journal, 19*(3), 48-60.

Liedtka, J.M., Haskins, M.E., Rosenblum, J.W., & Weber, J. (1997). The generative cycle: Linking knowledge and relationships. *Sloan Management Review, 38*(1), 47-58.

Manville, B., & Foote, N. (1996). Harvest your workers' knowledge. *Datamation,* (7), 78-81.

McDermott, R. (1999). Why information technology inspired but cannot deliver knowledge management. *California Management Review, 41*(4), 103-117.

McDermott, R., & O'Dell, C. (2001). Overcoming cultural barriers to sharing knowledge. *Journal of Knowledge Management, 5*(1), 76-85.

Miles, G., Miles, R.E., Perrone, V., & Edvinssen, L. (1998). Some conceptual and research barriers to the utilization of knowledge. *California Management Review, 40*(3), 281-288.

Morris, T., & Empson, L. (1998). Organization and expertise: An exploration of knowledge bases and the management of accounting and consulting firms. *Accounting, Organizations and Society, 23*(5/6), 609-624.

Nelson, K.M., & Cooprider, J.G. (1996). The contribution of shared knowledge to IS group performance. *MIS Quarterly, 20*(4), 409-432.

Nissen, M., Kamel, M., & Sengupta, K. (2000). Integrated analysis and design of knowledge systems and processes. *Information Resources Management Journal, 13*(1), 24-43.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 2,* 14-37.

Nonaka, I., & Konno, N. (1998). The concept of "Ba": Building a foundation for knowledge creation. *California Management Review, 40*(3), 40-54.

Ofek, E., & Sarvary, M. (2001). Leveraging the customer base: Creating competitive advantage through knowledge management. *Management Science, 47*(11), 1441-1456.

Quinn, J.B., Anderson, P., & Finkelstein, S. (1996). Managing professional intellect: Making the most of the best. *Harvard Business Review, 74*(2), 71-80.

Reimus, B. (1997). *Knowledge sharing within management consulting firms.* Retrieved February 9, 1999, from *http://www.kennedyinfo.com/mc/gware.html*

Robertson, G. (1999). The impact of knowledge management on Australian law firms. In: Proceedings of Deciphering Knowledge Management KNOW99 (pp. 191-202): Sydney.

Rodwell, I., & Humphries, J. (1998). The legal face of knowledge management. *Managing Information, 5*(7), 31-32.

Ruggles, R. (1998). The state of the notion: Knowledge management in practice. *California Management Review, 40*(3), 80-89.

Soliman, F., & Spooner, K. (2000). Strategies for implementing knowledge management: Role of human resources management. *Journal of Knowledge Management, 4*(4), 337-345.

Whiting, R. (1999). Knowledge management: Myths and realities. *Informationweek Online,* (November 22), 1-5.

Wiig, K.M. (1997). Knowledge management: Where did it come from and where will it go? *Expert Systems with Applications, 13*(19), 1-14.

KEY TERMS

Community of Practice: A group of people in an organization who are (somehow) held together by common interest in their work topic, purpose, and activities.

Culture: Covers the pattern of basic assumptions accepted and used about behaviors, norms, and values within an organization.

Knowledge Management: The systematic, explicit, and deliberate approach to creating, sharing, and using knowledge in order to enhance organizational performance.

Knowledge Sharing: The processes of transforming and transferring knowledge through an organization are designated by knowledge sharing.

Implementation Management of an E-Commerce-Enabled Enterprise Information System

Joseph Sarkis
Clark University, USA

R.P. Sundarraj
University of Waterloo, USA

INTRODUCTION

The integration of enterprise systems and the supply chain to an organization is becoming more critical in an ever-changing, globally competitive environment. Quick response will require close relationships, especially communications and information sharing among integrated internal functional groups as well as the suppliers and customers of an organization. Texas Instruments (TI), headquartered in Dallas, Texas, has come to realize this requirement for building and maintaining its competitive edge. Thus, it sought to implement an enterprise resource planning (ERP) system with a focus on linking it with a global electronic commerce (e-commerce) setting, an innovative and current issue (Weston, 2003).

There were a number of major players, including project management direction from Andersen Consulting Services, software vendors such as SAP and i2 Technologies, hardware vendors such as Sun Microsystems, and various suppliers and customers of TI.

The purpose of this case is to provide some aspects of implementation of strategic systems that provide valuable lessons for success. We begin and rely on the foundation of a strategic systems implementation model, which is initially described. A description of the case follows, with the various stages as related to strategic systems implementation described. We complete our discussion with implications and conclusions.

BACKGROUND

A process-oriented framework for ERP management is presented to help guide the discussion of this case (see Cliffe, 1998; Davenport, 1999; Miranda, 2002; Sarkis & Sundarraj, 2000).

The elements include the following:

- Strategy formulation and integration—One of the results of this step in the process is determination of an organization's core competencies that need specific technology support.
- Process planning and systems design—Also known as the reengineering phase, three studies are usually undertaken at this stage, and they are named AS-IS, SHOULD-BE, and TO-BE.
- System evaluation and justification—Here, analysis focuses on the economic, technical, and operational feasibility and justification of the system.
- System configuration—As a packaged software system, there are likely to be discrepancies (at the detailed level) between the needs of an organization and the features of the software. Hence, a significant amount of effort can be expected to configure the system or the organizational processes in order to produce an alignment between them.
- System implementation—The implementation stage can be classified into startup, project management, and a migration handing the switch over from the old to the new system.
- Postimplementation audit—This last "feedback" stage, although very important from a continuous-improvement perspective, is one of the more neglected steps.

As can be seen, the process suggested above can be arduous, but this necessary effort must be anticipated for the successful integration of complex and strategic systems into an organization.

IMPLEMENTING A GLOBAL ERP SYSTEM AT TI

Company Background

Texas Instruments Incorporated (TI) is a global semiconductor company and the world's leading designer and supplier of digital signal processing (DSP) solutions and analog

technologies (semiconductors represent 84% of TI's revenue base). The company has manufacturing or sales operations in more than 25 countries and, in 1999, derived in excess of 67% of its revenues from sales to locations outside the United States. Prior to the implementation of ERP, TI had a complex suite of stand-alone nonintegrated marketing, sales, logistics, and planning systems consisting of thousands of programs that were based on many independent databases and were running on proprietary mainframe systems.

OVERVIEW

Since the 1980s, TI had used a highly centralized infrastructure utilizing proprietary mainframe computers for meeting its IT requirement. As the first step toward global business processes, certain planning processes and systems were standardized in 1989. Starting in 1996, TI underwent a company-wide reengineering effort that led to the implementation of a 4-year, $250 million ERP system using Sun Microsystems' hardware platform, SAP AG's ERP software, i2's advanced planning tools, and Andersen Consulting's implementation process. In 1998, Texas Instruments implemented the first release of the ERP system, which primarily consisted of a prototype implementation of the i2 system running on a Sun E10000 platform. In early 1999, TI began rolling out the second release. In the middle of 1999, TI completed the i2 Technologies software implementation as part of the third release. Finally, TI turned on the remaining financials, and new field sales, sales, and distribution modules. A high-level architecture of TI's pioneering ERP implementation consists of SAP and the i2 system for advanced planning and optimization. The system is a pioneering large-scale global single-instance implementation of seven modules (finance, procurement and materials management, logistics, planning, field sales, sales, and marketing) for all of TI's divisions, and it is in use by 10,000 TI employees to handle 45,000 semiconductor devices and 120,000 orders per month. This solution also enabled global Web access to information for TI's 3,000 external users at customer, distributor, and supplier sites.

STAGES IN MANAGING THE GLOBAL ERP SYSTEM IMPLEMENTATION

Strategy Formulation

Traditionally, TI was primarily running what was called a "commodity" business, wherein orders were received, manufactured, and shipped as a batch. Mass customization combined with the maturity of TI's business caused it to reexamine its goals and strategies. TI started its shift toward a more customized product environment.

Within this new customized product environment, TI had a number of customer needs that could not be met easily. Thus, the goal was to determine the appropriate processes and information systems that must be put in place in order to support such agile design and manufacturing strategies. Another goal was a move toward supplier-managed inventory and customer-managed orders. Finally, standardizing systems was another integrative corporate goal. TI made extensive use of metrics. Strategic goals are translated into tactical and operational quantifiable objectives.

Process Planning and Systems Design

TI conducted a massive reengineering effort for the whole organization, with the goal of setting standard processes globally. The major result of this effort was to declare that all inventory and manufacturing management be done globally.

TI decided to implement a single-instance ERP system so as to fully leverage the system's capabilities to support the flexibility and standardization demanded by global processes. After site visits by major ERP vendors, TI selected SAP mostly because of its scalability to handle voluminous amounts of data.

System Justification

A budget of approximately $250 million was set for the implementation. The justification of the system was done using a combination of tangible and intangible factors at both the enterprise and business-unit levels. Standard hard-justification measures such as ROI and IRR were used to ensure the financial viability of the project.

Through this business case justification, acceptable financial returns, along with strategic factors such as competing effectively within a given niche market, and operational factors, such as global inventory management, all played roles in ERP's justification at TI.

System Configuration

The goals and processes entailed numerous and significant changes to all aspects of the business process design of the system.

Implementation

In this phase, concepts and goals were translated into tangible action, and as a result, it is perhaps one of most difficult phases of the project. General principles such as global

processes and standard systems need to be backed up by convincing and deploying the right people to implement the processes.

Startup

A number of key personnel, along with their families, were expatriated to the United States and stationed in Dallas, Texas, for a few years. About 250 people were transitioned from TI to Andersen Consulting that became the main provisioner of services with respect to the ERP system. IT outsourcing in this case involved Andersen Consulting taking over the employment and management of former TI people.

Project Management

Change management played a large role in this stage. The roles of training, planning, and communicating were of equal importance. All management levels were involved in this process as well as various vendors and suppliers. Some of the practices included the following:

- On-site experts were made available to new users of the system.
- A help desk was set up to handle problems that could not be addressed by these experts.
- A ticketing system for managing and prioritizing problems was also established (e.g., a system stop was a high-priority ticket that would get round-the-clock attention).

Handling Go-Live

To get prepared for "go-live," the key managers who were stationed in Dallas were sent back to their territories for educating the next level of users. Using selected experts, user-acceptance scripts were defined and tested, with problems, if any, being resolved as per one of the schemes outlined above. Daily conference calls were set up for 30 days prior to go-live to obtain status checks on progress and on the tickets.

Based on the results of these checks, a risk analysis was conducted weekly to determine the effects of various potential failures. The implementation plan was to have a few go-live dates one after another, but in relatively quick succession. Except for the planning system, in all the other stages, in this case, a direct conversion was employed. That is, with a downtime of about 2 to 3 hours during a weekend, the old system was turned off and the new one turned on.

Postimplementation Status

The system met most of its goals 9 months after the complete implementation. There are around 13,000 users on the system, with concurrent users ranging from 300 to 1,700. Some of the key performance measures and parameters evaluated were as follows:

- Productivity dip
- Ontime delivery
- Single-instance, global system
- Better response
- Inventory reduction

FUTURE TRENDS—MANAGERIAL IMPLICATIONS

The following lessons are summarized:

- Conduct a thorough strategic plan—The case illustrated how market forces had compelled the company to make radical shifts in its organizational environment and culture.
- Align IT plans with business plans—Conduct reengineering studies and develop strategic IT plans to align key IT needs with those of the business (Barker & Frolick, 2003).
- Get top management support—The prescription of top management support has been made ever since early IT implementations to the present (Mabert et al., 2001).
- Change management—Set realistic user expectations, such as the initial productivity dips. User involvement is critical. Andersen Consulting's process helped to ensure that such was the case. Make sure that the user was supported to help improve user satisfaction (Lee et al., 2003; Legare, 2002).
- Strong champion characteristics (Dean, 1987)—In TI's situation, the manager of the ERP project had over two decades of experience in various levels of the organization. This manager had broad knowledge of corporate operations. Previously, he was a vice president of one of TI's divisions.
- Rationalize business models and processes—Make sure the business models and processes fit within the strategic direction and goals of the organization. Time, mass customization, and flexibility concerns led to a global model (Gardiner et al., 2002). Global cultural issues were also a concern and needed to be managed (Davison, 2002).
- Manage external enterprises—Appropriate and well-planned involvement of consultants is important for

keeping the project on a tight schedule. Further, with the advent of e-commerce, companies are more likely to ship and order goods on the basis of Web-based inputs.

- Manage using metrics (Skok et al., 2001; Hitt et al., 2002)—TI and Andersen Consulting have a corporate culture and policy that require the stringent and formal use of metrics in the management and evaluation of projects. They attribute this policy adherence as one of the key reasons for success of the ERP implementation. Key performance indicators included such issues as reduction in inventory, percentage of suppliers linked, and productivity of outputs.

CONCLUSION

TI's ERP implementation with an e-commerce perspective required a significant amount of features that added issues to its management:

- It is a single-instance system, providing access to the same data, irrespective of the geographic location of the user.
- It provides access to 3,000 external users (customers and suppliers), thereby enabling 70% of the transactions to be conducted electronically.

Management saw some problems in this implementation process and tried to address the issues. Some of the major problems included the following:

1. The software for supply chain management (Red Pepper) that was initially chosen did not meet expectations of TI. This system had to be scrapped, and this resulted in a multimillion dollar cost.
2. A productivity dip occurred, and the implementation had to address this issue for all managers throughout the organization who had some stake in the performance of the system. The expectations that this would occur were communicated through newsletters and messages. Consistent and continuous communication helped to mitigate a situation that could have caused a major project failure.
3. Getting buy-in from internal functions not directly associated with the implementation process was difficult. This occurred with the marketing function.

Future extension of developing appropriate infrastructure is another issue that needs to be faced (Kovacs & Paganelli, 2003). Lessons learned here may be appropriate for small or large organizations, but some differences appear in the practices of what is successful and not for ERP implementations in different size organizations. One of the major issues is that small companies get greater operational benefits from ERP, while larger companies get more financial benefits (Mabert et al., 2003).

REFERENCES

Barker, T., & Frolick, M. N. (2003). ERP implementation failure: A case study. *Information Systems Management, 20*(4), 43-49.

Cliffe, S. (1999). ERP implementation. *Harvard Business Review, 77*(1), 16.

Davenport, T. (1998). Putting the enterprise into the enterprise systems. *Harvard Business Review, 77*(4), 121-131.

Davison, R. (2002). Cultural complications of ERP. *Communications of the ACM, 45*(7), 109-111.

Dean, J. (1987). *Deciding to innovate: How firms justify advanced technology.* Cambridge, MA: Ballinger Publishing Company.

Gardiner, S. C., Hanna, J. B., & LaTour, M. S. (2002). ERP and the reengineering of industrial market processes. *Industrial Marketing Management, 31*(4), 357-365.

Hitt, L. M., Wu, D. J., & Zhou, X. (2002). Investment in enterprise resource planning: Business impact and productivity measures. *Journal of Management Information Systems, 19*(1), 71–98.

Kovacs, G. L., & Paganelli, P. (2003). A planning and management infrastructure for large, complex, distributed projects—Beyond ERP and SCM. *Computers in Industry, 51*(2), 165–183.

Lee, J., Siau, K., & Hong, S. (2003). Enterprise integration with ERP and EAI. *Communications of the ACM, 46*(2), 54-60.

Legare, T. L. (2002). The role of organizational factors in realizing ERP benefits. *Information Systems Management, 19*(4), 21-42.

Mabert, V. A., Soni, A., & Venkataramanan, M. A. (2001). Enterprise resource planning: Common myths versus evolving reality. *Business Horizons, 44*(3), 69-76.

Mabert, V. A., Soni, A., & Venkataramanan, M. A. (2003). The impact of organization size on enterprise resource planning (ERP) implementations in the U.S. manufacturing sector. *Omega, 31*(3), 235-236.

Miranda, R. (2002). Needs assessments and business case analysis for technology investment decisions. *Government Finance Review, 18*(5), 12-16.

Reich, B., & Benbasat, I. (2000). Factors that influence the social dimensions of alignment between business and

information technology objectives. *MIS Quarterly, 24*(1), 81-113.

Sarkis, J., & Sundarraj, R. P. (2000). Factors for strategic evaluation of enterprise information technologies. *International Journal of Physical Distribution and Logistics Management, 30*(3/4), 196-220.

Skok, W., Kophamel, A., & Richardson, I. (2001). Diagnosing information systems success: Importance-performance maps in the health club industry. *Information and Management, 38*(7), 409.

Weston, F. C. (2003). ERP II: The extended enterprise system. *Business Horizons, 46*(6), 49-55.

KEY TERMS

Audit: Reviewing and monitoring the performance of a system.

Enterprise Resource Planning (ERP) System: An information system that spans organizational boundaries with various organizational functional modules and systems integrated and managed by one system application.

Go-Live: The actual operation of an information system.

Mass Customization: Producing basically standardized goods but incorporating some degree of differentiation and customization.

Reengineering: Activities that seek to radically change business processes and support systems in an organization.

Single Instance: A one-time full-fledged company-wide initial operation of a system, as opposed to incremental (functionally or organizationally) or modular implementations.

Strategic Justification: The process of evaluating and selecting systems based on tangible (financial) and intangible factors that have implications for long-term and broad management of the organization.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1397-1401, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Implementation of ERP in Human Resource Management

Zhang Li
Harbin Institute of Technology, China

Wang Dan
Harbin Institute of Technology, China

Chang Lei
Harbin Institute of Technology, China

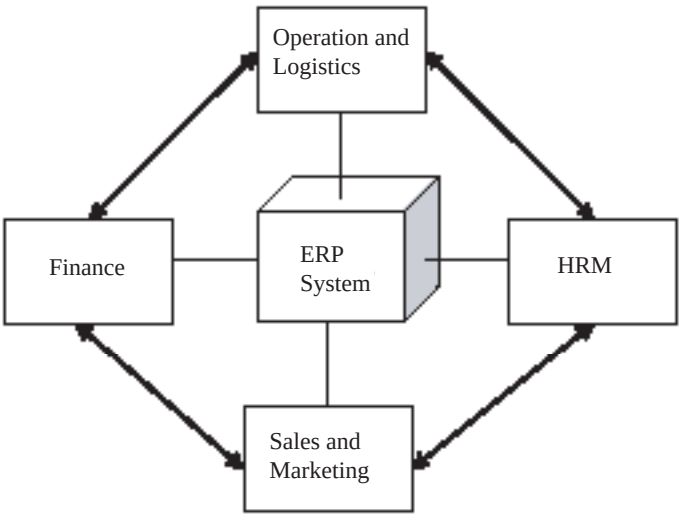
INTRODUCTION

In 1999, Peter Drucker said: "A new Information Revolution is well under way. It is not a revolution in technology, machinery, techniques, software or speed. It is a revolution in concepts." As a result of information technology (IT) innovation and reorganization, enterprise resource planning (ERP) was proposed by the Gartner Group in the early 1990s. It is a successor to manufacturing resource planning (MRP II) and attempts to unify all departmental systems together into a single, integrated software program that runs off a single database so that the various departments can more easily share information and communicate with each other (Koch, 2002). Over 60% of the U.S Fortune 500 had adopted ERP by 2000 (Kumar, & Hillegersberg, 2000; Siau, 2004), and it was projected that organizations' total spending on ERP adoptions was an estimated $72.63 billion in 2002 (Al-Marshari, 2002).

Many scholars have recognized the importance of people in organizations, and this viewpoint is the central focus of the human resource management (HRM) perspective (Pfeffer, 1995). In this perspective, HRM has the potential to be one of the key components of overall enterprise strategy. Additionally, HRM may provide significant competitive advantage opportunities when they are used to create a unique (i.e., difficult to imitate) organizational culture that institutionalizes organizational competencies throughout the organization (Bowen & Ostroff, 2004).

Typically, an ERP system supports HRM, operation and logistics, finance, and sales and marketing functions (Daven-

Figure 1. Function modules of an ERP system



port, 1998) (see Figure 1). But the early development stage of ERP in enterprises was all along with the center of production and sales course. Until recently, research has empirically supported the positive relationship between corporate financial performance and HRM function, and managers have also realized that HRM can deliver organizational excellence and competitive advantage for enterprises (Boudreau & Ramstad, 1997; Huselid, 1995; Wright, McMahan, Snell, & Gerhart, 2001). The HRM module was introduced into ERP, forming a highly integrated and efficient resource system with the other function modules of ERP. However, there are still many HRM-related problems that may result in the failure of ERP projects arising. So, there have been regular appeals to scholars for more research about the implementation of ERP systems in the HRM perspective in the last few years (Barrett & Mayson, 2006).

This article introduces the functions of an HRM module in ERP systems from the fields of human resource planning, recruitment management, training management, time management, performance management, compensation management, and business trip arrangement. Then it analyzes five HRM-related problems that may block the enterprises from implementing ERP successfully, and it provides reasonable recommendations. Finally, the article discusses future trends and suggests emerging research opportunities within the domain of the topic.

BACKGROUND

ERP, a term coined by the Gartner Group, is not simply a tool that provides singular outputs, but rather an infrastructure that supports the capabilities of all other IT-based tools and processes utilized by a firm (Enslow, 1996). Shang and Seddon (2000) classified the different types of ERP benefits as: IT infrastructure benefits, operational benefits, managerial benefits, strategic benefits, and organizational benefits. Palaniswamy (2002) pointed out that the failures of ERP projects were not because the software were coded incorrectly, rather the companies failed to understand the real organizational needs and systems required to solve their problems to improve performance. Lynne, Axline, Petrie, and Cornelis (2000) analyzed the adopters' problems with ERP including project phase problems, problems with product and implementation consultants, shakedown phase problems, underestimating data quality problems and reporting needs, and so on.

Within the managerial literatures, a coherent approach provides a conceptual basis for asserting that human resource is a key source of competitive advantages, since it offers a unique contribution to value creation, rarity, imperfect imitability, and non-substitutability of a firm's strategic resources (Bellini & Canonico, 2007). Stone (2007) considered the past, present, and future of HRM theory and research. He concluded that HRM theory and research has considerable potential to enhance organizational efficiency and effectiveness. Ashbaugh and Rowan (2002) summarized the technology features of a modern HRM system (see Table 1).

In addition, some scholars have already studied the relationship or connection of ERP implementation with HRM. For instance, Ashbaugh and Rowan (2002) argued that the major difference between ERP and its predecessors (e.g., MRP II) is the linkage of financial and HRM applications through a single database in a software application that is both rigid and flexible. Wright and Wright (2002) listed two of the most-cited HRM risks in an ERP system: lack of user involvement and inadequate training. Hsu, Sylvestre, and Sayed (2006) supplied another often-overlooked HRM factor when implementing an ERP system—that is, the result of high stress levels on the staff, particularly in the finance or accounting departments, which are already under stress from the heavy workload in a legacy system. Li (2001) studied the HRM function module in an ERP system. He insisted that the practical HRM system should be built up to improve incentive mechanism and to strengthen the training of employees while applying ERP.

Table 1. Technology features of modern HRM system

Integration	User friendliness
Common relational database	Enhanced reporting and analysis
Flexible and scalable technology	Process standardization and malleability
Audit trail and drill down capabilities	Internet and capabilities
Robust security	Document management and imaging
Workflow	

IMPLEMENTATION OF ERP RELATED TO HRM

Functions of the HRM Module in ERP System

We have studied the necessity and essentiality of HRM to the implementation of ERP in the preceding part of this article. The adoption of ERP also greatly impacts HRM by extending its functions to the all-direction management category (see Figure 2). The functions of HRM have developed from simple compensation calculating and personnel management to the fields of human resource planning, recruitment management, training management, time management, performance management, compensation management, and business trip arrangement (Ahmad & Schroeder, 2003; Li, 2001; Stone, 2007). Data from all function systems will be collected into a central database, and the database can further supply data needed for all function systems by integration.

Human Resource Planning

Based on the requirements of enterprises, managers can use the HRM module of an ERP system to establish human resource planning conveniently. The ERP system assists the decision making of managers by simulating the performance of human resource planning and comparing the data. Additionally, the ERP system is also able to analyze or forecast the human resource planning costs by integrating relevant information.

Recruitment Management

Recruitment should be taken as a significant investment because human resources are the foundational assets of an enterprise. To keep advantages in competition, the human resources department must have a reasonable recruitment system to select talents for enterprise. The ERP system can support recruitment management in three ways. First, it optimizes the recruitment process to reduce the workload. Second, it offers scientific management to recruitment costs. Third, it provides useful information for the decision making on recruitment management.

Training Management

Training in the use of multiple skills including process improvement skills, which can provide long-term work-life security rather than job security (Schonberger, 1994). The implementation of ERP can help train employees to acquire technical, interpersonal, and business skills required to become fully participating team members in the early stage of team development (Pasmore & Mlot, 1994). In other team development stages, by giving support to the human resource department to make an appropriate training plan, the ERP system can also help train team members to accept new skills, improved management regulations, and so on.

Time Management

Time management may support the planning, controlling, and management processes of HRM. It means to arrange the time table for the enterprises and staff flexibly according to the local calendar. The ERP system can record the attendance rate and other relevant information by using a Telematics Control Unit (TCU). For example, data related to the compensation will be further processed in the compensation management system.

Performance Management

Performance evaluation might consider the following issues: How are the facilitative and operational activities allocated to individuals in an organization, and how does the facilitative content of a task vary in an organization (Nilakant, 1994). The human resource department can establish an evaluation index system according to these issues. By integrating the performance management system with the time management system, the ERP system will record data in a central database and keep relevant data timely for each evaluation index. These data will be useful to the decision making of managers on corporate strategy, too.

Compensation Management

A reasonable compensation system should be able to apply proper calculation methods in terms of different regions, departments, positions, and so forth. The implementation of ERP will achieve this objective by integrating the compensation management system with other systems (e.g., timing management system, performance management system) so that it can update relevant data in a timely fashion so as to establish a dynamic compensation calculation system. The human resource department can simulate the performance of the calculation system to forecast compensation information needed and to adjust the structure of the compensation management system. This is an excellent improvement because it decreases management costs as well as problems caused by the intervention of manpower. Compensation management also includes other functions such as salary payments, loans for staff, and so forth.

Business Trip Arrangement

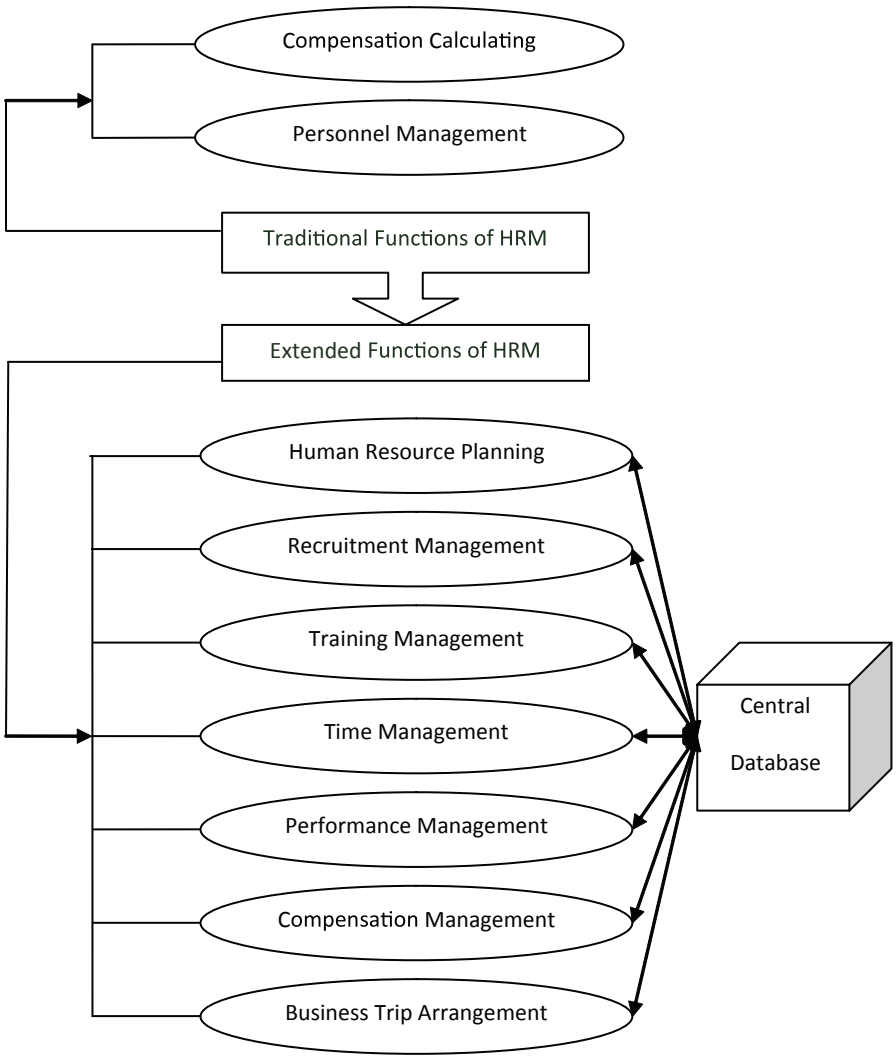
A business trip arrangement system can control the whole flow of a business trip from application to ratification and

reimbursement. These data will be further processed in other function modules of ERP (e.g., finance module) through systems integration.

Main HRM-Related Problems of ERP Implementation in Enterprises

ERP has been broadly applied in enterprises for nearly 20 years because of the enormous potential economic benefits. However, 53% of the ERP projects in U.S. firms were failed by 1996, and the success rate of ERP projects in Chinese enterprises was only less than 20% by 2002 (Edwards, 1999; Yang & Zhao, 2003). Hsu et al. (2006) pointed out that HRM factors played a significant role in almost all failed information systems. This article analyzes five main HRM-related problems that may block the success of ERP projects in enterprises (Lynne et al., 2000; Wright & Wright, 2002; Hsu et al., 2006; Sun, & Xu, 2006).

Figure 2. Traditional and extended functions of HRM



First is the shortage of professionals, especially the inter-disciplinary talents who are expert at both IT and management. ERP is not just an advanced technique, it is also an advanced management concept. Therefore, the inter-disciplinary professionals are crucial to the success of ERP. This problem is especially serious in the small and medium-sized enterprises because of the weakness of strength and management level.

Second is the deficient talent introduction mechanism. Employees who take charge of the ERP project of an enterprise are under high pressure because they take great responsibilities for the success of the implementation. If an enterprise does not have an effective talent introduction mechanism to attract talented people who are needed, it will be hard to start the ERP project at all, not to mention the successful application.

Third is the insufficient education and training for employees. The cultivation of talents is a process that needs much time investment as well as money, while some enterprises do not want to pay much investment to the education and training of employees due to lack of an in-depth understanding of ERP.

Fourth is the poor incentive mechanism for employees. Above all, the compensation mechanisms of enterprises may not be attractive enough. A competitive compensation mechanism cannot only attract applicants, but it also prevents the job hopping of employees. Still, the ERP projects have not been supported enough by superior administration departments. For example, the application of ERP needs the support of all involved departments, while managers hesitate to place departmental backbones on the ERP implementation. It may also influence the working enthusiasm of employees and kill more innovations if superiors interfere too much with the implementation.

Fifth is the lack of exterior consultation or a supervision system. An enterprise must pass the scientific verification of experts if it wants to adopt ERP. The implementation must be carried out under the direction of exterior professional organizations and the supervision of a special system. Enterprises may neglect the necessity and importance of these functions.

RECOMMENDATIONS

To implement ERP successfully, enterprises must pay attention to such tasks of HRM as follows (Lynne et al., 2000; Ashbaugh & Rowan, 2002; Sun & Xu, 2006):

- *Enterprises should realize the importance of human resources.* An ERP system is a production of IT, while human resources are part of the essential power of the invention or improvement of IT. The idea that HRM is just a secondary function must be updated.
- *Enterprises should redesign the functions of HRM.* ERP extends the traditional functions of HRM greatly. The simple compensation calculating or personnel management cannot meet the requirements of an ERP system. The extended functions of HRM are more favorable for inter-departmental cooperation, because they can conveniently provide the decision making with information needed.
- *Enterprises should establish an effective talent introduction mechanism and an effective incentive mechanism for employees.* An attractive compensation mechanism can provide useful market competitive advantage for enterprises to hold talents. On the other hand, enterprises should supply employees with reasonable freedom to make decisions within position responsibilities. This can enhance the working initiative as well as the sense of responsibility of employees.
- *Enterprises should place great emphasis on personnel education and training.* The human resource department can carry an ideological education that involves both senior managers and common staff to arouse their concerns about ERP implementation. The frequency and quality of personnel training should be strengthened as well.
- *Enterprises should enhance exterior consultation as well as the supervision function for the adoption of ERP.* It is a good idea to invite exterior experts to enterprises to make practical guidance. The human resource department can also organize staff to visit the enterprises that implement ERP successfully to draw on experience. Additionally, the human resource department should be in charge of the supervision function or help your enterprise establish a supervision department, made up of interior managers and exterior experts.

FUTURE TRENDS

Enterprise resource planning is a new concept introduced by the Gartner Group in 2000 to label the latest extensions to ERP (Classe, 2001). The new concept is that, having successfully integrated internal business applications such as finance, sales, and marketing to increase efficiency and create a total overview of the business, ERP II can be used to integrate external applications with collaborative commerce arrangements, e-business, and the supply chain (Payne, 2002). Traditional ERP is the main component in an ERP II system, but for the purposes of the collaboration, an ERP II system is opened to inflow and outflow of information (Moller, 2003). On the other hand, during the last decade, a new wave of human resource technology known as electronic human resource management (e-HRM) has emerged with the advent of intranet- and Internet-based technologies. E-HRM is mainly connecting staff and managers with the

human resource department electronically through the human resource portal (Lai, 2006). The basic expectations are that using *e-HRM* will decrease costs, will improve the human resource service level, and will give the human resource department space to become a strategic partner (Rual, Bondarouk, & Velde, 2007).

When we combine the ERP II concept with e-HRM technology, we find a valuable issue to study—that is, the research about the implementation of ERP II based on e-HRM. Additionally, we can also study other details of this issue according to different countries, cultures, industries, and so forth.

CONCLUSION

As IT continues to development, there will be more and more enterprises adopting ERP systems. ERP can extend the traditional functions of HRM greatly and also heighten the importance of HRM in enterprises. Enterprises that implement ERP must perfect the functions of HRM to raise the success rate, so to enhance the whole management level of enterprises. A poignant hope is that this article will be helpful to both scholars and practitioners who wish to improve the current situation of ERP implementation.

REFERENCES

Ahmad , S., & Schroeder, R.G. (2003). The impact of human resource management practices on operational performance: Recognizing country and industry differences. *Journal of Operations Management, 21,* 19-43.

Al-Marshari, M. (2002). Enterprise resource planning (ERP) systems: A research agenda. *Industrial Management and Data Systems, 102*(3), 165-170.

Ashbaugh, S., & Rowan, M. (2002). Technology for human resources management: Seven questions and answers. *Public Personnel Management, 31*(1), 7-20.

Barrett, R., & Mayson, S. (2006). Exploring the intersection of HRM and entrepreneurship: Guest editors' introduction to the special edition on HRM and entrepreneurship. *Human Resource Management Review, 16*(4), 443-446.

Bellini, E., & Canonico, P. (2007). Knowing communities in project driven organizations: Analysing the strategic impact of socially constructed HRM practices. *International Journal of Project Management,* (September), 29.

Boudreau, J.W., & Ramstad, P.M. (1997). Measuring intellectual capital: Learning from financial history. *Human Resource Management, 36,* 343-356.

Bowen, D.E., & Ostroff, C. (2004). Understanding HRM-firm performance linkages: The role of the strength of the HRM system. *Academy of Management Review, 29,* 203-221.

Classe, A. (2001). Business-collaborative commerce—the emperor's new package. *Accountancy,* (November).

Davenport, T.H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review,* (July-August), 121-131.

Drucker, F.P. (1999). *Management challenges for the 21st century* (p. 97). New York: HarperCollins.

Edwards, J. (1999). *Three-tier client—server at work.* New York: John Wiley & Sons.

Enslow, B. (1996). Which comes first: ERP or supply chain planning projects? *Gartner Group Best Practices and Case Studies.*

Hsu, K., Sylvestre, J., & Sayed, E.N. (2006). Avoiding ERP pitfalls. *Journal of Corporate Accounting & Finance,* (May-June), 67-74.

Huselid, M.A. (1995). The impact of human resource management practices on turnover, productivity, and corporate financial performance. *Academy of Management Journal, 38,* 635-672.

Koch, C. (2002). The ABCs of ERP. *CIO Magazine,* (February).

Kumar, K., & Hillegersberg, J.V. (2000). ERP experiences and evolution. *Communications of the ACM, 43*(4), 24-26.

Lai, W.H. (2006). Implementing e-HRM: The readiness of small and medium sized manufacturing companies in Malaysia. *Asia Pacific Business Review, 12*(4), 465-485.

Li, Y.F. (2001). Thoughts about the ERP human resource management. *Journal of Yunnan University of Finance and Economic, 17*(10), 12-16.

Lynne, M.M., Axline, S., Petrie, D., & Cornelis, T. (2000). Learning from adopters' experiences with ERP: Problems encountered and success achieved. *Journal of Information Technology, 15,* 245-265.

Moller, C. (2003). ERP II—next-generation extended enterprise resource planning. *Proceedings of the 7th World Multi-Conference on Systemics, Cybernetics and Informatics.*

Nilakant, V. (1994). Transdisciplinary approach to a theory of performance in organizations. *Human Systems Management, 13*(1), 41-48.

Palaniswamy, R. (2002). An innovation-diffusion view of implementation of enterprise resource planning (ERP)

systems and development of a research model. *Information & Management, 40,* 87-114.

Pasmore, W.A., & Mlot, S. (1994). Developing self-managing work teams: An approach to successful integration. *Compensation and Benefits Review, 26*(4), 15-23.

Payne, W. (2002). The time for ERP? *Work Study, 51*(2), 91-93.

Pfeffer, J. (1995). Producing sustainable competitive advantage through the effective management of people. *Academy of Management Executive, 9,* 55-69.

Rual, J.M.H., Bondarouk, V.T., & Velde, M. (2007). The contribution of e-HRM to HRM effectiveness: Results from a quantitative study in A Dutch ministry. *Employee Relations, 29*(3), 280-291.

Schonberger, R.J. (1994). Human resource management lessons from a decade of total quality management and re-engineering. *California Management Review, 36*(4), 109-123.

Shang, S., & Seddon, P.B. (2000). A comprehensive framework for classifying the benefits of ERP systems. *Proceedings of the 6th Americas Conference on Information Systems* (pp. 1005-1014).

Siau, K. (2004). Enterprise resource planning (ERP) implementation methodologies. *Journal of Database Management, 15*(1), 1-4.

Stone, D.L. (2007). The status of theory and research in human resource management: Where have we been and where should we go from here? *Human Resource Management Review, 17*(2), 93-95.

Sun, X., & Xu, W. (2006). Research on ERP and enterprise's human resources informatization. *Science Technology Information Development & Economy, 16*(7), 229-230.

Wright, P.M., McMahan, G.C., Snell, S.A., & Gerhart, B. (2001). Comparing line and HR executives' perceptions of HR effectiveness: Services, roles, and contributions. *Human Resource Management, 40,* 111-123.

Wright, S., & Wright, A. (2002). Information system assurance for enterprise resource planning systems: Unique risk considerations. *Journal of Information Systems, 16,* 99-113.

Yang, J.Y., & Zhao, X.W. (2003). Research on HRM in ERP system. *Journal of Beijing Institute of Technology,* (August), 73-77.

KEY TERMS

Electronic Human Resource Management (e-HRM): The planning, implementation, and application of information technology for both networking and supporting at least two individual or collective actors in their shared performing of HR activities.

Enterprise Resource Planning (ERP): An approach to the provision of business support software that enables companies to combine the computer systems of different areas of the business—production, sales, marketing, finance, human resources, and so forth—and run them off a single database. Also defined as an application and deployment strategy to integrate all things enterprise-centric.

Human Resource: The people that staff and operate an organization.

Human Resource Management (HRM): The function within an organization that focuses on recruitment of, management of, and providing direction for the people who work in the organization.

Information Technology (IT): The collection of technologies that deal specifically with processing, storing, and communicating information, including all types of computer and communications systems as well as reprographics methodologies.

Manufacturing Resource Planning (MRP II): A method for the effective planning of all resources of a manufacturing company, including functions of business planning, production planning and scheduling, capacity requirement planning, job costing, financial management, forecasting, and so forth.

Implementation of Programming Languages Syntax and Semantics

Xiaoqing Wu
The University of Alabama at Birmingham, USA

Marjan Mernik
University of Maribor, Slovenia

Barrett R. Bryant
The University of Alabama at Birmingham, USA

Jeff Gray
The University of Alabama at Birmingham, USA

INTRODUCTION

Unlike natural languages, programming languages are strictly stylized entities created to facilitate human communication with computers. In order to make programming languages recognizable by computers, one of the key challenges is to describe and implement language syntax and semantics such that the program can be translated into machine-readable code. This process is normally considered as the **front-end** of a compiler, which is mainly related to the programming language, but not the target machine.

This article will address the most important aspects in building a compiler front-end; that is, syntax and semantic analysis, including related theories, technologies and tools, as well as existing problems and future trends. As the main focus, formal syntax and semantic specifications will be discussed in detail. The article provides the reader with a high-level overview of the language implementation process, as well as some commonly used terms and development practices.

BACKGROUND

The task of describing the syntax and semantics of a programming language in a precise but comprehensible manner is critical to the language's success (Sebesta, 2008). The **syntax** of a programming language is the *representation* of its programmable entities, for example, expressions, declarations and commands. The **semantics** is the actual *meaning* of the syntax entities. Since the 1960s (Sebesta, 2008), intensive research efforts have been made to formalize the language implementation process. Great success has been made in the syntax analysis domain. Context-free grammars are widely used to describe the syntax of programming languages, as well as notations for automatic parser generation (Slonneger & Kurtz, 1995). Formal specifications are also useful in describing semantics in a precise and logical manner, which is helpful for compiler implementation and program correctness proofs (Slonneger & Kurtz, 1995). However, there is no universally accepted formal method for semantic description (Sebesta, 2008), due to the fact that the semantics of programming languages are quite diverse, and it is difficult to invent a simple notation to satisfy all the computation needs of various kinds of programming languages. Overall, compiler development is still generally considered as one of the most appropriate software applications that can be implemented systematically using formal specifications.

Context-free grammar, BNF and EBNF. In the 1950s, Noam Chomsky invented four levels of grammars to formally describe different kinds of languages (Chomsky, 1959). These grammars, from Type-0 to Type-3, are rated by their expressive power in decreasing order, which is known as the **Chomsky hierarchy.** The two weaker grammar types (i.e., regular grammars, Type-3; and **context-free grammars**, Type-2) are well-suited to describe the lexemes (i.e., the atomic-level syntactic units) and the syntactic grammar of programming languages, respectively. **Backus-Naur Form** (BNF) was introduced shortly after the Chomsky hierarchy. BNF has the same expressive power as context-free grammar and it was first used in describing ALGOL 60 (Naur, 1960). BNF has an extended version called **Extended BNF**, or simply EBNF, where a set of operators are added to facilitate the expression of production rules.

LL, LR and GLR parsing. Based on context-free grammars and BNF, a number of parsing algorithms have been developed. The two main categories among them are called **top-down parsing** and **bottom-up parsing**. Top-

down parsing recursively expands a nonterminal (initially the start symbol) according to its corresponding productions and matches the expanded sentences against the input program. Because it parses the input from **L**eft to right, and constructs a **L**eft parse (i.e., left-most derivation) of the program, a top-down parser is also called an LL parser. A typical implementation of an LL parser is to use recursive descent function calls for the expansion of each nonterminal, which are easy to develop by hand. Bottom-up parsing, on the other hand, identifies terminal symbols from the input stream first, and combines them successively to reduce to nonterminals. Bottom-up parsing also parses the input from **L**eft to right, but it constructs a **R**ight parse (i.e., reverse of a right-most derivation) of the program. Therefore, a bottom-up parser is also called an LR parser. LR parsers are typically implemented by a pushdown automaton with actions to shift (i.e., push an input token into the stack) or reduce (i.e., replace a production right-hand side at the top of the stack by the nonterminal which derives it), which are difficult to code by hand. The table size of a canonical LR parser is generally considered too large to use in practice. Consequently, an optimized form of it, the LALR (Look Ahead LR) parser is widely used instead, which significantly reduces the table size (Aho, Lam, Sethi, & Ullman, 2007).

The grammars recognized by LL and LR parsers are called LL and LR grammars, respectively. They are both subsets of context-free grammars. LL grammars cannot have left-recursive references (i.e., a nonterminal has a derivation with itself as the leftmost symbol) and LR grammars cannot create action conflicts (i.e., shift-reduce conflicts, reduce-reduce conflicts). Any LL grammar can be rewritten as an LR grammar, but not vice versa. Both LL and LR parsers can be extended by using *k* tokens of lookahead. The associated parsers are called LL(k) parsers and LR(k) parsers, respectively. Lookahead can eliminate most of the action conflicts existing in an LR grammar, unless the grammar contains ambiguity. To resolve action conflicts in a generic way, an extension of the LR parsing algorithm, called **GLR (Generalized LR) parsing** (Tomita, 1986), has been invented to handle any context-free grammar, including ambiguous ones. The basic strategy of the algorithm is, once a conflict occurs, the GLR parser will process all of the available actions in parallel. Hence, GLR parsers are also named parallel parsers. Due to its breadth-first search nature, the GLR parsing suffers from its time and space complexity. Various attempts have been made to optimize its performance (e.g., McPeak & Necula, 2004). Currently, GLR is still not widely used in programming language implementation, but its popularity is growing. There are a number of tools available to automatically generate LL, LR and GLR parsers from grammars[1]. These tools are generally referred to as parser generators or compiler-compilers (Aho, Lam, Sethi, & Ullman, 2007).

Figure 1. Attribute grammar of the Robot language for location calculation

Production	Semantic Rules
Program → begin Moves end	Program.out_x = Moves.out_x; Program.out_y = Moves.out_y; Moves.in_x = 0; Moves.in_y = 0;
$Moves_0$ → Move $Moves_1$	$Moves_0$.out_x = $Moves_1$.out_x; $Moves_0$.out_y = $Moves_1$.out_y; Move.in_x = $Moves_0$.in_x; Move.in_y = $Moves_0$.in_y; $Moves_1$.in_x = Move.out_x; $Moves_1$.in_y = Move.out_y;
Moves → ε	Moves.out_x = Moves.in_x; Moves.out_y = Moves.in_y;
Move → left	Move.out_x = Move.in_x - 1; Move.out_y = Move.in_y;
Move → right	Move.out_x = Move.in_x + 1; Move.out_y = Move.in_y;
Move → up	Move.out_x = Move.in_x; Move.out_y = Move.in_y + 1;
Move → down	Move.out_x = Move.in_x; Move.out_y = Move.in_y - 1;

Attribute Grammar: Semantic analysis is usually embedded into syntax analysis. **Attribute grammars** (Knuth, 1968) were invented to specify the static semantics of programming languages (those properties that can be acquired from the source program directly) as an extension of context-free grammars. Each symbol has an associated set of attributes that carry semantic information, and each production has a set of semantic rules associated with attribute computation. The attributes are divided into two groups: synthesized attributes and inherited attributes. The synthesized attributes are used to pass the semantic information up the parse tree and the inherited attributes are passed down from parent nodes. An evaluation scheme traverses the tree one or more times to compute all the attributes. Figure 1 shows an attribute grammar of a simple Robot language. In this language, a robot can move in four directions from the initial position (0, 0). The attribute grammar uses two synthesized attributes *out_x* and *out_y*, as well as two inherited attributes *in_x* and *in_y*, to compute the final result location.

Denotational Semantics: Although an attribute grammar can provide a formal description of the static semantics of a programming language, it is still not powerful enough to express dynamic semantics, which are those properties that "reflect the history of program execution or user interactions with the programming environment" (Kaiser, 1989, p. 169). A number of formal methods have been created to define the full semantics of programming languages such as operational semantics, axiomatic semantics and denotational semantics (Slonneger & Kurtz, 1995). Among these approaches, **denotational semantics** is the most researched. The denotational semantics of a programming language map a program in that language directly to its meaning represented as a mathematical value called its denotation. The primary components of a denotational semantics specification are the semantic algebras and evaluation functions. The semantic algebras define a collection of semantic domains and the valuation functions map syntax domains into functions over semantic domains. The denotational semantics for the simple Robot language is shown in Figure 2. Denotational semantics is very suitable for mathematical reasoning, but it is rarely used for practical implementation due to the fact that it cannot be used to derive an efficient compiler implementation. For more details about denotational semantics and other formal methods for semantic description, please refer to Slonneger and Kurtz (1995).

LANGUAGE IMPLEMENTATION IN PRACTICE

There are various approaches to develop a compiler front-end, such as handcrafting a compiler from scratch, fully implementing the language with formal specifications, or utilizing parser generators to describe syntax and using a programming language to code semantics. Depending on the nature of the language and available environment, different development strategies can be selected. In the following paragraphs, two of the most commonly used strategies are introduced.

Figure 2. Denotational semantics of the Robot language

Abstract Syntax
P ∈ Program
M ∈ Move
P ::= begin M end
M ::= left | right | down | up | M_1 M_2

Semantic algebra
Point = Integer × Integer

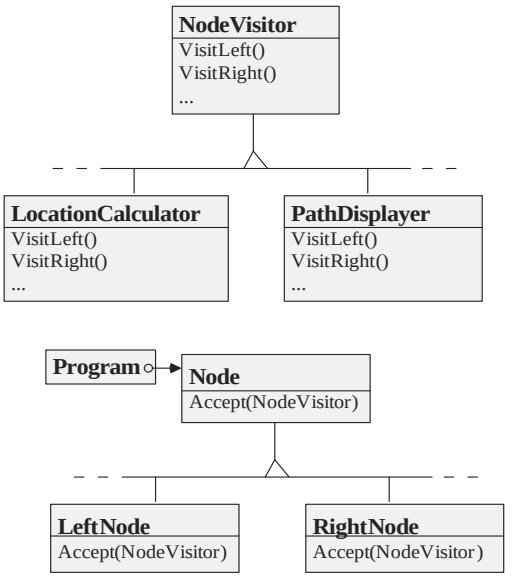
Semantic evaluation functions
P : Program → Point
M : Move → Point → Point
P ⟦begin M end⟧ = M ⟦M⟧(0,0)
M ⟦left ⟧ = λ(x,y). (x-1, y)
M ⟦right ⟧ = λ(x,y). (x+1, y)
M ⟦up⟧ = λ(x,y). (x, y+1)
M ⟦down ⟧ = λ(x,y). (x, y-1)
M ⟦M_1 M_2 ⟧ = λ(x,y). M ⟦M_2⟧(M ⟦M_1⟧(x,y))

Scenario I: Parser Generator + Visitor Pattern

In this scenario, syntax and semantics are implemented separately (i.e., language syntax is described by a formal specification and semantics is implemented by a general purpose programming language). For syntax analysis, the language is fully specified with a context-free grammar, which serves as input to a parser generator such as Yacc (Johnson, 1975). Most parser generators allow attaching semantic code with each grammar rule in the parser specification, but the mixed grammar and code fragments generally make the overall specification hard to read and debug (Wu, 2007). In reality, as long as a representation of the program (i.e., an Abstract Syntax Tree -AST) is built after parsing, the semantic code can be factored out to later phases as AST traversing logic. Therefore, ideally the attached code fragment should only contain code related to AST construction. To support polymorphism and inheritance, these tree nodes are generally implemented using object-orientation (i.e., each node is represented by a class).

After the tree is successfully built, the remaining phases of the front-end can be implemented by using the Visitor pattern (Gamma, Helm, Johnson, & Vlissides, 1995), which is an object-oriented pattern used to separate functional operations with object structure. In applying this pattern, all the methods pertaining to one semantic pass are encapsulated inside a visitor class. Each AST node class has a general *accept* method, which can redirect a semantic evaluation request to the appropriate method in the provided visitor class. The benefit of using this pattern is that each tree traversal phase is isolated as a class, which is independent of other node classes and can be freely modified or extended. Figure 3 provides an illustration of the Visitor pattern in implementing two analysis phases for the Robot language. The visitor *LocationCalculator* contains the same semantics as described in Figures 1 and 2. The second visitor *PathDisplayer* contains printing actions to display the path of the robot.

Figure 3. Illustration of the Visitor pattern



Scenario II: Attribute Grammar Implementation

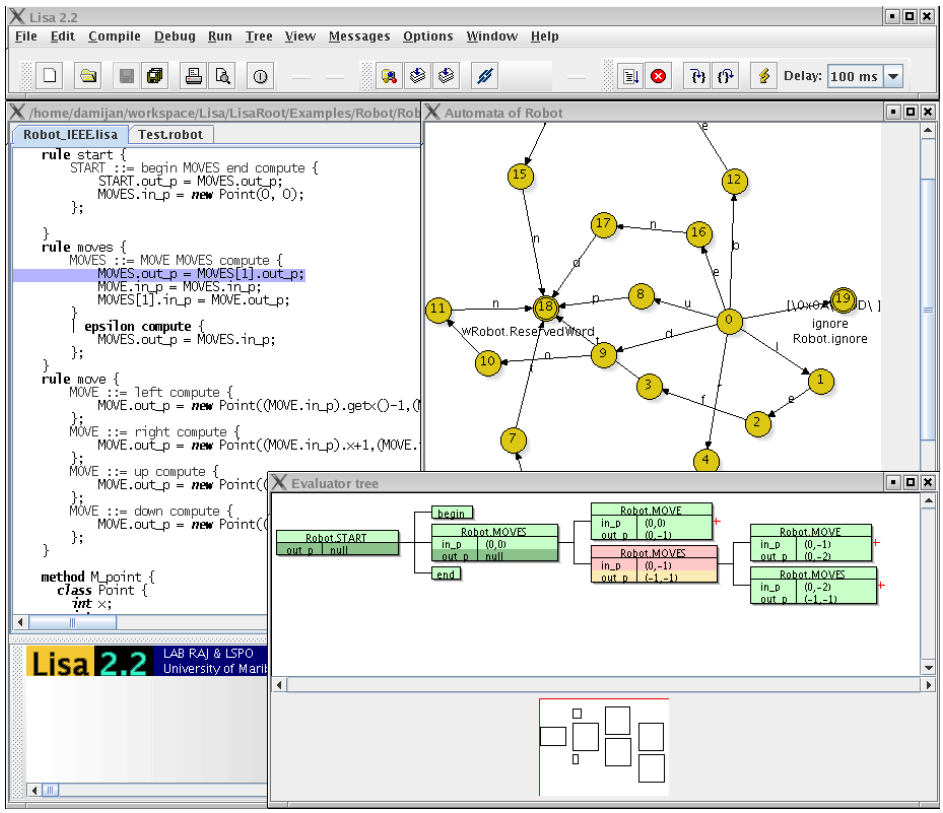
In this scenario, both syntax and semantics are described by formal specifications. Specifically, syntax is described by a variant of BNF, while semantics are described with attribute grammars where specialized languages can be used for semantic rules (e.g., FNC-2 [Jourdan, Parigot, Julie, Durin, & Le Bellec, 1990]) or assignment statements from general-purpose programming languages are employed (e.g., LISA-Language Implementation System based on Attribute grammars [Mernik & Žumer, 2005]). An important distinction for parser generators, where semantic actions are mixed with productions, is that attribute grammars are declarative. Hence, the semantic part can be separated from the syntax. Moreover, the order of evaluation of semantic rules does not need to be specified when describing the language. It is automatically obtained from an attribute dependency graph, which alleviates the burden from the language implementer.

Various compiler generators based on attribute grammars exist (e.g., FNC-2 and LISA) where systems are differentiated by their AST representation (e.g., structured, modular, or object-oriented), traversal logic (e.g., L-attributed or ordered attribute grammars) and the power of semantic rules (e.g., logic, functional, parallel, incremental). Overall, the whole front-end is automatically generated from such specifications and often used in implementation of domain-specific languages (Mernik, Heering, & Sloane, 2005). An informative description of attribute grammars as a programming language implementation methodology is given in Paakki (1995).

The development of the Robot language in LISA is shown in Figure 4. LISA moves toward modular, reusable and extensible language specifications. Different parts of language specifications (e.g., lexical definitions, generalized syntax rules that encapsulate semantic rules, and operations on semantic domains) can be inherited, specialized or overridden from ancestor specifications. Due to space limitations these features are not present in Figure 4. Interested readers can find more information in Mernik and Žumer (2005).

From the above two scenarios, it can be seen that language implementation has been greatly facilitated by automatic generation tools. However, being generative is not enough. The key challenge of language development resides in

Figure 4. LISA integrated development environment



providing high-quality modularity of the implementation (Mernik, Wu, & Bryant, 2004), which helps the developer to divide-and-conquer the complexity of constructing a complex language. Large grammars may well run into several hundred or even thousands of productions (e.g., the GOLD grammar of COBOL 85 is 2500 lines[2]). Developers sometimes are forced to put a fair amount of production rules inside one module and develop the parser as a whole, which results in a system that is hard to develop and maintain. This is an obstacle that neither of the two scenarios introduced above can overcome. Moreover, although the use of object-oriented techniques and concepts greatly improves language specification toward better modularity, there are still incidents where object-orientation is not a natural specification for semantic actions that crosscut AST node classes. For instance, since object-orientation describes a system by a collection of objects rather than a collection of operations, the complicated implementation of the Visitor pattern introduces a lot of side-effects (Wu et al., 2006), despite the fact that it brings advantages in separation of concerns.

FUTURE TRENDS

It can be foreseen that, for syntax description and implementation, context-free grammar and its variants will still be the mainstream specification for syntax analysis. Sophisticated parser generators will be used as the main method for developing a parser, while hand-coded parsers will only exist in legacy systems. GLR parsing and its equivalent technologies will gain more attention when the implementation becomes more mature.

Due to the diversity of language semantics, describing semantics using formal specifications will remain an intricate problem. The current available formal specification methods will be enhanced in terms of the comprehensibility and automatic generation ability. But the ideal solution, that a language designer can easily use a specification language to produce an efficient compiler implementation without using any general purpose programming language, is less likely to appear in forthcoming years (e.g., attribute grammar-based systems like LISA use general-purpose languages for semantic rules).

To achieve modularity, extensibility and reusability to the fullest extent, it can be expected that more state-of-the-art techniques will be brought into language development. For example, component-based development technology (Szyperski, 2002) could help decompose the workload of developing a large language into developing a number of smaller languages such as languages for expression and commands, which can be developed and tested individually (Wu, 2007). Instead of generating a single parser for a complex language, the future parser generator should be able to generate a set of parsers based on well-defined specification modules. These parser components should be composed in an organic way to deliver the parsing functionality. Moreover, due to the crosscutting nature of semantic phases, aspect-oriented programming (Kiczales et al., 1997) should be utilized for developing semantics, which is already being investigated, as described in de Moor, Peyton-Jones, and Van Wyk (2000), Hedin and Magnusson (2003) and Wu et al. (2006).

CONCLUSION

The article provides background knowledge to help the reader understand the current state and future direction in language development technologies. It first describes the roles of syntax and semantics in implementing a programming language, followed by the introduction of formal syntax and semantic specifications. Some theory of syntax analysis is highlighted by describing grammar classes and their associated parsing algorithms such as LR(k), LL(k) and GLR. Practical information on compiler construction is the main part of this article, where two most widely used development scenarios are presented. Some of the current difficulties in language implementation are also presented, with a special focus on the modularity problem of a compiler system. Various research directions are suggested at the end, including using the state-of-the-art programming technologies such as aspect-oriented programming and component-based development to help modularize language implementations. Due to space restrictions, some topics such as component-based compiler development and aspect-oriented semantic implementation are not covered in depth in the article. Interested readers may refer to the Web site (http://www.cis.uab.edu/wuxi/research/) for more details.

REFERENCES

Aho, A.V., Lam, M.S., Sethi, R., & Ullman, J. D. (2007). *Compilers: Principles, techniques, and tools* (2nd ed.). Boston, MA: Addison-Wesley.

Chomsky, N. (1959). On certain formal properties of grammars. *Information and Control, 2*(2), 137-167.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns, elements of reusable object-oriented software*. Reading, MA: Addison-Wesley.

Hedin, G., & Magnusson, E. (2003). Jast add-an aspect-oriented compiler construction system. *Science of Computer Programming, 47*(1), 37-58.

Johnson, S. C. (1975). *YACC: Yet another compiler compiler* (Tech. Rep. No. 32). Murray Hill, NJ: AT&T Bell Laboratories.

Jourdan, M., Parigot, D., Julie, C., Durin, O., & Le Bellec, C. (1990). Design, implementation and evaluation of FNC-2 attribute grammar system. In *Proceedings of the ACM Sigplan Conference on Programming Language Design and Implementation*, (pp. 209-222).

Kaiser, G.E. (1989). Incremental dynamic semantics for language-based programming environments. *ACM Transactions on Programming Languages and Systems, 11*, 169-193.

Kiczales, G., Lamping, J., Mendhekar, M., Maeda, C., Lopes, C., Loingtier, J-M, & Irwin, J. (1997). Aspect-oriented programming. In *Proceedings of the 11th European Conference on Object-Oriented Programming (ECOOP '97)*, (pp. 220-242).

Knuth, D. E. (1968). Semantics of context-free languages. *Mathematical Systems Theory, 2*(2), 127-145.

McPeak, S., & Necula, G. C. (2004). Elkhound: A fast, practical GLR parser generator. In *Proceedings of Conference on Compiler Construction (CC '04*), (Vol. 2985, pp. 73-88).

Mernik, M., Heering, J., & Sloane, A. (2005). When and how to develop domain-specific languages. *ACM Computing Surveys, 37*(4), 316-344.

Mernik, M., Wu, X., & Bryant, B. R. (2004, June). Object-oriented language specification: Current status and future trends. In *Proceedings of the ECOOP Workshop on Evolution and Reuse of Language Specifications for DSLs*.

Mernik, M., & Žumer, V. (2005). Incremental programming language development. *Computer Languages, Systems and Structures, 31*, 1-16.

de Moor, O., Peyton-Jones, S., & Van Wyk, E. (2000). Aspect-oriented compilers. In *Proceedings of Generative and Component-based Software Engineering (GCSE)*, (pp. 121-133).

Naur, P. (1960). Report on the algorithmic language ALGOL 60. *Communications of the ACM, 3,* 299-314.

Paakki, J. (1995). Attribute grammar paradigms—a high-level methodology in language implementation. *ACM Computing Surveys, 27*(2).

Sebesta, R.W. (2008). *Concepts of programming languages* (8th ed.). Boston, MA: Addison-Wesley.

Slonneger, K., & Kurtz, B.L. (1995). *Formal syntax and semantics of programming languages*. Addison-Wesley.

Szyperski, C. (2002). *Component software: Beyond object-oriented programming* (2nd ed.). New York: ACM Press, Addison-Wesley.

Tomita, M. (1986). *Efficient parsing for natural language*. Kluwer Academic.

Wu, X. (2007, May). *Component-based language implementation with object-oriented syntax and aspect-oriented semantics*. Docotral dissertation, University of Alabama at Birmingham, Retrieved December 14, 2007, from http://www.cis.uab.edu/wuxi/thesis/Thesis_Xiaoqing_Wu.pdf.

Wu, X., Bryant, B.R., Gray, J., Roychoudhury, S., & Mernik, M. (2006). Separation of concerns in compiler development using aspect-orientation. In *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC '06),* (pp.1585-1590).

KEY TERMS

Aspect-Oriented Programming: Aspect-oriented Programming (AOP) provides special language constructs called aspects that modularize crosscutting concerns in conventional program structures (e.g., a concern that is spread across class hierarchies of object-oriented programs). A translator called a weaver is responsible for merging the additional code specified in an aspect language into the base language at compile time.

Attribute Grammar: Attribute grammar is an extension of context-free grammars in which each symbol has an associated set of attributes that carry semantic information, and each production has a set of semantic rules associated with attribute computation.

Bottom-Up Parsing: Bottom-up parsing is a parsing strategy that identifies terminal symbols from the input stream first, and combines them successively in a rightmost way to reduce to nonterminals, until the start symbol is reduced.

Compiler-Compiler: Compiler-compilers are tools that can automatically generate compilers or interpreters from programming language descriptions.

Context-Free Grammar: A context-free grammar is a quadruple (N, T, P, S), where N is a set of nonterminal symbols; T is a set of terminal symbols with $T \cap N = \varnothing$; the relation $P \subseteq N \times (N \cup T)^*$ is a finite set of production rules and S is the start symbol with $S \in N$. The production of the form $A \rightarrow \alpha$ means A derives α, where $A \in N$ and $\alpha \in (N \cup T)^*$.

Denotational Semantics: The denotational semantics of a programming language map a program in that language directly to its meaning represented as a mathematical value called its denotation. Its primary components are the semantic algebras that define a collection of semantic domains and evaluation functions that map syntax domains into functions over semantic domains.

Front-End: Among all the compiler phases, those mainly related to the programming language but not the target machine are normally considered as the front-end of a compiler.

Semantics: The semantics of a programming language is the actual meaning of the syntax entities. It describes what the programs written in this language can achieve.

Syntax: The syntax of a programming language is the representation of its programmable entities, for example, expressions, declarations and commands. It describes the appearance of programs written in this language.

Top-Down Parsing: Top-down parsing is a parsing strategy that iteratively expands the leftmost nonterminal (initially, the start symbol) according to its corresponding productions and matches the expanded sentences against the input program from left to right.

ENDNOTES

1 Available at http://en.wikipedia.org/wiki/List_of_compiler-compilers.

2 Available at http://www.devincook.com/GOLDParser/grammars/index.htm

Implementation of Web Accessibility Related Laws

Holly Yu
California State University, Los Angeles, USA

INTRODUCTION

In the past three decades, the general method of delivering and receiving information has shifted from paper-based, typewriter-generated, hand-edited, and printing-press-produced publications to more technology-mediated, intelligent, WYSIWYG software-generated forms and interactive design. The Web has expanded its horizon as a gateway in carrying and delivering information to include audio and video formats. Further, the advent of the *Web 2.0*, or social networking/virtual community via the Web has changed the nature of the Web not only as an information carrier, but also a tool for all to use, share, and participate. Consequently, the concept of delivery of, access to, and interaction with information has changed to reflect this phenomenon. The new forms of utilizing the Web that have made it easier for non-disabled people have often created barriers for people with disabilities because, in a large part, the standard methods of access and delivery are inaccessible for people with disabilities.

The disability rights movement in the United States originated during the post World War II era when large numbers of veterans who were disabled in the war joined the efforts of parents seeking education and independent living options for their children with disabilities (Slatin & Rush, 2003). The notion of access to information involving the civil rights of people with or without disabilities arises from the fact that access to information through technology has increasingly become a necessary tool for success and a source of opportunity in education and employment. With the unprecedented opportunities they created for people with and without disabilities, it has become apparent that information technologies have a tremendous potential for allowing people with disabilities to participate in mainstream activities and to support their ability to live independently. The legal foundation for protecting the right to access for persons with disabilities has been established through a series of federal and state laws, and court decisions. These laws provide a legal ground on Web accessibility implementation.

BACKGROUND

A person with a disability is defined in the *Americans with Disabilities Act (ADA)* as "someone who has a physical or mental impairment that substantially limits one or more major life activities, a person who has a record of such impairment, or a person who is regarded as having such impairment" (ADA, 1990). It is estimated by the American Foundation for the Blind (AFB) that approximately 10 million blind and visually impaired people in the United States, and 1.3 million Americans, are legally blind (AFB, 2007). In defining Web accessibility, Hackett and Parmanto (2005) state that accessibility may be direct or through the use of *assistive technologies*, hardware or software that aids a person in accessing the information. Section 508 of the Rehabilitation Act of 1973, as amended in 1998, documents that "Web sites are accessible when individuals with disabilities can access and use them as effectively as people who don't have disabilities." (Section 508, 1998).

Disabled people who benefit the most from accessible design and assistive technology in interacting with the Web, are those who are blind, visually impaired, people with hearing impairment, physical impairment, and learning difficulty such as dyslexia. Recently, we have seen a growing body of significant laws, regulations, and standards concerning Web accessibility that impact people with disabilities and their ability to fully overcome digital barriers and participate in the Web environment.

LAWS, REGULATIONS, STANDARDS, AND GUIDELINES

Under the provisions of laws, some of the legal milestones that have direct impact on Web accessibility are Section 504 of the Rehabilitation Act of 1973, Americans with Disabilities Act (ADA) of 1990, and Section 508 of the Rehabilitation Act of 1973, as amended in 1998.

Section 504, Rehabilitation Act, 1973

Signed on October 1, 1973, Section 504 of the Rehabilitation Act is regarded as landmark legislation and the first civil rights law prohibiting recipients of federal funds from discriminatory practices on the basis of disability.

Core areas of the legislation consist of the prohibition of such activities as discriminatory employment practices,

and discrimination in the delivery of educational offerings, health, welfare and social services, or any other type of programs benefit, or service supported in whole or in part by federal funds.

Section 504 is currently applied to all entities that receive federal government funds, including states, counties, cities, towns, villages, and their political subdivisions, public and private institutions, public and private agencies, and other entities that receive federal money. Each Federal agency has its own set of Section 504 regulations that guide its own programs. Over the years, the Rehabilitation Act has been amended several times to address the constant changes in technology and its impact on society. The amendments most relevant to the access to information technology are those made to *Section 508*. The significance of Section 504 lies not only in that it was the first statute applying civil rights protections to people with disabilities, but that it also "furnished the model for major subsequent enactments, including the ADA" (NCD, 2001). Section 504 was legislated too early to specifically address the issue of access to services and programs provided over the Web.

Americans with Disabilities Act (ADA), 1990

Signed by President George H.W. Bush on July 26, 1990, the ADA establishes a clear and comprehensive prohibition of discrimination on the basis of disability. While Section 504 applies to federal government agencies and those that receive federal funds, the ADA extends the rights of equal treatment for people with disabilities to the private area, to all places of public accommodation, employers, and entities that deliver government services. The core sections of the law are found in the first three titles: Employment, State and Local Government Activities, and Public Accommodation. Title II of the ADA requires that state and local governments give people with disabilities an equal opportunity to benefit from all of their programs, services, and activities, such as public education, employment, transportation, recreation, health care, social services, courts, voting, and town meetings. Section 202, Title II indicates that "no qualified individual with a disability shall, by reason of such disability, be excluded from participation in or be denied the benefits of the services, programs, or activities of a public entity, or be subjected to discrimination by such entity" (ADA, 1990). Title II recognizes the special importance of communication, which includes access to information in its implementing regulation at 28 CFR Section 35.160(a). The regulation requires that a public entity must take appropriate steps to ensure that communications with persons with disabilities are as effective as communications with persons without disabilities. The ADA mandates for "effective communication, reasonable accommodations, and auxiliary aides and services" (ADA, 1990).

However, Web accessibility did not become prominent until 1996 when the Department of Justice (DOJ) responded to Senator Tom Harkin (D-Iowa), the author of the ADA, when he inquired on behalf of one of his constituents on Web page compatibility for the blind and other people with disabilities. In response, DOJ stated that ADA Title II and III do indeed require covered entities to provide "effective communication" regardless of the media used, and that information offered through digital media must be offered through "accessible means" as well. The Internet is an excellent source of information and, of course, people with disabilities should have access to it as effective as people without disabilities (DOJ, 1996). This response involves understanding to what extent the ADA requires Web pages to be accessible to people with disabilities. The DOJ's ruling explains how the mandate for "effective communication" in ADA should apply to Web pages and Web design.

Section 508, Rehabilitation Act, 1998

Signed into law on August 7, 1998 as part of the Workforce Investment Act, Congress revised Section 508, an amendment to the Rehabilitation Act of 1973. The core area of this amendment is to ensure that all Americans have access to information technology. The law applied specifically to all U.S. government agencies, but it also affects anyone who works with the U.S. government. The law requires that when federal departments or agencies develop, procure, maintain, or use Electronic and Information Technology (EIT), they should ensure that the EIT allows federal employees with disabilities to have access to and use of information and data that is comparable to the access to and use of information and data by other federal employees (Section 508, 1998).

Section 508 charges the Architectural and Transportation Barriers Compliance Board (Access Board) with the task of producing accessibility standards for all electronic and information technologies used, produced, purchased, or maintained by the federal government. On December 21, 2000, the Access Board finalized and issued the Standards. With its publication, the federal government for the first time set specific access standards for information presented over the Web. According to the law, federal agencies were permitted a six-month period after the Standards were issued to demonstrate that their information technology systems met the Standards. Federal agencies become subject to civil rights litigation for noncompliance under Section 508 after June 21, 2001. These Standards have been folded into the federal government's procurement regulations, namely the Federal Acquisition Regulations (FAR) to which most agencies are subject. The Standards define means of disseminating information, including computers, software, and electronic office equipment. They provide criteria that make these products accessible to people with disabilities, including those with vision, hearing, and mobility impairments. The

scope of Section 508 and the Access Board's Standards are limited to the federal sector. It does not explicitly apply to the private sector, or to state and local governments unless a site is provided under contract to a federal agency, in which case only that Web site or portion covered by the contract would have to comply (Section 508 Standards, 2001).

The statements issued by the US Department of Education and the Access Board support a concept of state government obligations wherein they abide by Section 508 by virtue of their linkage with funding supplied to state governments through the Assistive Technology Act (29 USC 3001)**.** There is a statement in the US Department of Education's enforcing regulations to the Assistive Technology Act at 34 CFR Part 345, requiring "compliance with Section 508 of the Rehabilitation Act of 1973" (34 CFR 345.31, 2002).

Web Content Accessibility Guidelines (WCAG)

Founded in 1994, the World Wide Web Consortium (W3C) was originally organized by the European Center for Nuclear Research (CERN) and by the Massachusetts Institute of Technology (MIT). The W3C launched the Web Accessibility Initiative (WAI) in April 1997 specifically to address the question of how to expand access to the Web for people with disabilities with a goal of providing a single shared standard for Web content accessibility that meets the needs of individuals, organizations, and governments worldwide. The WAI published the Web Content Accessibility Guidelines 1.0 (WCAG) in May 1999, and Authoring Tool Accessibility Guidelines 1.0 (ATAG) in February 2000, and the User Agent Accessibility Guideline 1.0 (UAAG) in December 2002.

The WCAG establishes three priority checkpoints for Web content developers to meet. The Guidelines includes 64 checkpoints arranged under 14 separate guidelines within the three priorities.

Priority I -- Developers MUST satisfy Priority I checkpoints in order to attain a minimum degree of accessibility. Otherwise, one or more groups will find it impossible to access some part of the information.

Priority II – Developers SHOULD satisfy Priority II checkpoints for a higher degree of accessibility.

Priority III – Developers MAY address Priority III checkpoints for maximum accessibility and usability (WAI, 2000).

Topics of the Guidelines include images, programming scripts, navigation, multimedia, forms, frames, and tables. The Guidelines are detailed and prioritized with an associated checklist. WCAG 1.0 has become the basis for accessibility standards adopted by the international community.

Currently, WCAG 2.0 is being developed to apply to more advanced Web technologies, and be more precisely testable with a combination of automated testing and human evaluation. Specifically, the WCAG2.0 is developed to address many unanswered questions in WCAG 1.0 on how to implement, how to evaluate and reasons behind its requirements to reflect the diverse needs of a broad community including industry, disability organizations, accessibility researchers, government, and others interested in Web accessibility (WCAG 2 FAQ, 2007).

Web standards developed by the Web Accessibility Initiative (WAI) of W3C are called W3C Recommendations, and they encompass guidelines and technical reports. The milestones that a technical report goes through on its way to becoming a W3C Recommendation are working draft, last call working draft, candidate recommendation, proposed recommendation and W3C Recommendation (Web Standard).

Different from WCAG 1.0, WCAG 2.0 is organized around four design principles of Web accessibility. Each principle has guidelines, and each guideline has success criteria at level A, AA, or AAA. The basis for determining conformance to the WCAG 2.0 Working Draft are the success criteria (Overview of WCAG 2.0 Documents, 2007).

The Four Principles are stated in the Working Draft of Web Content Accessibility Guidelines 2.0 issued in May 2007:

- Perceivable – Information and user interface components must be perceivable by users;
- Operable – User interface components must be operable by users;
- Understandable – Information and operation of user interface much be understandable by users;
- Robust – Content must be robust enough that it can be interpreted reliable by a wide variety of user agents, including assistive technologies (Web Content Accessibility Guidelines 2.0, 2007).

WCAG 2.0 ensures accessible Web design, development, evaluation, and implementation**.** Key differences between WCAG 1.0 and WCAG 2.0 lie in that WCAG 2.0 applies more broadly to different Web technologies and is designed to apply as technologies develop in the future (Overview of WCAG 2.0 Documents, 2007).

Laws, Regulations and Policies in Other Countries

World Wide Web has no boundaries, therefore, "nations everywhere are developing policies and standards to enhance the accessibility of electronics and information technology" (Paciello, 2000). These nations include European countries (Belgium, France, Greece, Norway, Portugal, Sweden, and United Kingdom), Canada, Australia, and Japan. Davies finds that the drivers that underpin appropriate and adequate provision to visually impaired people include legislation,

international conventions and codes of practice (Davies, 2007). In many countries and the international community, accommodating visually impaired people are made mandatory in legislations. For instance, there are Standard Rules on the Equalization of Opportunities for Persons with Disabilities issued by the United Nations in 1993, and in the United Kingdom, there is the Disability Discrimination Act (1995) and the Special Educational Needs and Disability Act (2001). Australia and Canada have also enacted legislations. Prominent Australia legal standards include the Disability Discrimination Act (DDA) of 1992, and the Anti-Discrimination Act of New South Wales of 1977. Australia's Human Rights and Equal Opportunity Commission, a government agency, is responsible for creating and establishing legal decrees mandating equality of access for the disabled. In Canada, the Equity and Diversity Directorate of the Public Service Commission of Canada (PSC), "was the first government institution in any country to create a series of Web accessibility guidelines used to evaluate Web pages" (Paciello, 2000). The Common Look and Feel Working Group created in 1998 by the Treasury Board Secretariat's Internet Advisory Committee is charged to establish standards and guidelines as a base for official government policy. There are practical methods based on research and evaluation to achieve desirable results for people with disabilities. As a founding contributor, the European Community provides financial support to the WAI. European Union (EU) Web Accessibility Benchmarking (WAB) cluster of three EU-funded projects --the European Internet Accessibility Observatory (EIAO), SupportEAM, and BenToWeb, working in liaison with the W3C/WAI to develop a harmonized European methodology for evaluation and benchmarking of Web sites, the Unified Web Accessibility Methodology, or UWEM (Brophy & Craven, 2007).

ISSUES RELATED TO WEB ACCESSIBILITY

The current concerns encountered in the implementation of accessible Web design do not only stem from design issues, nor are they because of a lack of laws, regulations and government-related standards, although there are unanswered questions in that laws create difficulty in the implementation process. And, among these concerns are that the current legal framework for electronic and information technology accessibility is actually a patchwork of laws covering certain categories of technology in some settings, other categories in other settings, but nowhere reflecting an overview or comprehensive assessment of either the issues or the solutions (NCD, 2001). We have not seen that this situation has been improved since the statement was made by the National Council on Disability (NCD) in 2001.

There is a tendency that most Web designers either unaware of the laws and regulations, or ignore the accessibility requirements, or only pay attention to them as part of legal requirements. This situation is resulted from the absence of an obligation to fulfill legal requirements and not being aware that ensuring resource accessibility is a legal mandate. In many cases, the absence of obligations is because of the unfamiliarity with the legal responsibility for creating accessible Web sites. This is the reason why so many institutions are still in the beginning stage of applying accessible Web design techniques.

Compounding these problems have been the conflicting views between minimum compliance and maximum accessibility.

Technically, stand-alone workstations utilizing assistive technology software solutions are no longer sufficient because in the Web environment, the linkage between an individual and the Internet community as a whole is addressed, and access to the Web cannot be handled on a case-by-case basis using workstations with assistive technology. When approaching the issue of accessibility, John Mueller, the author of *Accessibility for Everyone: Understanding the Section 508 Accessibility Requirements*, points out that the issue should be approached from the viewpoint of universal design. "The application you create should work equally well with any hardware the user might have attached to the machine" (2003). Assistive technology alone cannot overcome the barriers that are created at a more basic level, the format in which content is presented (Schmetzke, 2001). Access barriers created by inaccessible design cannot be overcome even with the most sophisticated assistive technology. In defining the concept of accessible design, Brophy and Craven state that "the information provided on screen must be presented in a way that can be interpreted by any kind of access technology, "which can be assistive, adaptive, or enabling technology that enables a visually impaired user to access on-screen information receiving output in a way that is appropriate to their needs" (Brophy & Craven, 2007). Therefore, good Web interface design and the use of assistive technology are two fundamental methods to improve the accessibility of Web-based information access, delivery and interaction.

Further, *assistive technology* may also give rise to issues of incompatibility, and different types of *assistive technologies* present different problems to accessing Web sites. Training of both the user and the service provider adds another layer to the complexity of the issues surrounding the successful use of assistive technologies.

Design issues revealed by usability studies are also primary factors to impacting people with disabilities interacting with Web sites. Such issues include navigation structure, and content organization, form design, etc. In their studies, Brophy and Craven (2007) and others (Hackett & Parmanto,

2005**,** Pilling, Barrett, & Floyd, 2004) indicate the following issues in addition to those major concerns stated above:

- Higher education Web sites become progressively inaccessible as complexity increases
- Lack of support and training in the use of assistive technologies
- Disabled people cannot afford, or are not motivated, to upgrade their assistive software to the latest version
- Web pages with cluttered contents, graphics and links
- Poor accessibility for voice recognition system users
- Lack of ALT text or poor use of ALT text

Technical issues are deeply rooted in the lack of legal obligations in accessible Web design, and the accessible design is still an afterthought. Compounding these are cost in time and resources and conflicts with aesthetic and other design considerations.

IMPLEMENTATION

The overwhelming benefits of creating universally accessible Web pages can hardly be disputed. However, *accessibility barriers* are usually systematic. There are practical and legal reasons for addressing accessibility issues in our policy, education and design. As the rapid development of new Web applications continues, particularly Web 2.0 or social networking applications, it is necessary to ensure that barriers themselves will not continue to expand. Awareness, policy, education, and assessment are primary aspects to ensure Web sites designed with maximum accessibility. We have seen tremendous efforts being made in these processes.

To increase awareness of Web accessibility issues, understanding the laws related to accessibility, and the consequences of major disparities in our society is the first step. Lacking of knowledge in legal obligations and applications in making Web sites accessible can be resolved through systematical training. It is crucial that people with disabilities be involved in the procurement and/or development of accessibility solutions. Thus, issues of compatibility and accessibility can be anticipated and addressed during strategic planning stage. Identifying problems and implementing repairs are two integral elements of Web site accessibility assessment. One way to ensure the assessment can be carried out is to recommend its inclusion into the institution's Web planning and design process. Another is for the institution to develop its own checklists or follow the most up-to-date guidelines published by the W3C. The assessment should indicate the areas of violations and content that is inaccessible. Based on the assessment, a plan should be made to integrate the findings and recommendations for change. Developing tools to facilitate accessible Web design is a key to the ultimate success of an accessible Web site. Brophy & Craven point out that, the three assessment methodologies include semi-automatic and automatic testing using validation tools, manual evaluation using relevant criteria for assessment such as WCAG 1.0/2.0 guidelines and success criteria, and user testing of specific features of a Web site. Using guidelines and automated testing tools are not enough to assess the accessibility of Web sites and that involving users—and in particular disabled people—in the design and testing process will help improve accessibility usability (2007). Principles of universal design should be achieved, and the view of depending heavily on the tool – the assistive devices to mediate and decode should be utilized as a last result. Conducting usability studies with the understanding that the accessibility is an integral part of usability will assist in promoting inclusion.

FUTURE TRENDS

Discussions about the "digital divide" problem, needs for accessible Web design, and practical tips for designing barrier-free Web sites found in literature in recent years have demonstrated awareness-raising efforts (Brophy & Craven, 2007; Casey, 1999; Clapper & Burke, 2005; Hackett & Parmanto, 2005; Jobe, 2000, & Valenza, 2000; Loiacono & McCay, 2004; Pilling, Bannett, & Floyd, 2004; Rouse, 1999).

The ultimate goal for Web sites is universal design, or pervasive accessibility-interfaces that adapt to the user, regardless of ability. It is so often that with each release of new technology, there will be fixes to accessibility problems. "Subsequently, that technology that increases access to people with disabilities is termed assistive and adaptive as opposed to being a technology that is in inherently accessible" (Paciello, 2000). As we observed, while much of today's technology designed lacking in accessibility, recent awareness of laws and guidelines, and advances in promoting accessibility give rise to hope for achieving the ultimate goal for accessibility. The "design for all" concept – a single version of the Web site which is accessible to everyone regardless of abilities, is changing the way of the provision of a parallel text-only version. The benefits resulting from the implementation of universal Web design principles, which allow pages to be accessible to the largest number of segments of the population, have been emerging.

Needless to say, the progress in designing more accessible Web pages has been made. However many problems cited in this article five years ago are still in existence.

CONCLUSION

The concept of accessible design or universal design is increasingly becoming an important component of Web design. Today, public and private entities are subject to the requirements of the ADA, Section 504, and Section 508. According to the definition of the law, denial of access resulting from the inaccessibility of mainstream information technology is considered to be discrimination. Barriers mostly resulting from the absence of obligations to fulfill legal requirements should be eliminated as the awareness of the importance of accessibility increases. Functionalities found in assistive technology can streamline our digital architecture, and the very functionality required in Web design by people with disabilities can meet dynamic requirements for Web-based transactions. A universal Web design will greatly reduce the cost for assistive technology geared specifically to individual computer workstations, and allow universal access for users from anywhere at anytime. A universal Web design allows the disability community to benefit as a whole rather than achieving accessibility through a segregated, compartmentalized, and ad hoc approach. The benefits and value for overcoming these barriers for the community of people with disabilities cannot be disputed, and maximum accessibility is a goal for us to achieve.

REFERENCES

Americans with Disabilities Act of 1990 (ADA), (1990). Pub.L.No. 101-336, §2,104 Stat. 328.

American Foundation for the Blind (AFB), (2007). Blindness Statistics. Retrieved October 15, 2007 from http://www.afb.org/.

Brophy, Peter, & Craven, Jenny. (2007). Web Accessibility. Library Trends, 55(4), 950-972.

Casey, C. A. (1999). Accessibility in the virtual library: creating equal opportunity Web Sites. *Information Technology and Libraries*, 18(1), 22-25.

Clapper, D.L., & Burke, D.D. (2005). Edu dilemma: the Web accessibility challenge facing public and private universities. *Journal of Strategic E-Commerce*, 3 (1/2), 71-98.

Davies, J. E. (2007). An overview of international research into the library and information needs of visually impaired people. *Library Trends*, 55(4), 785-795.

Department of Justice. (DOJ), (2000). Information Technology and People with Disabilities: The Current State of Federal Accessibility. Washington, D.C.: DOJ. Retrieved June 6, 2001 from http://www.usdoj.gov/crt/508/report/content.htm

Hackett, Stephanie, & Parmanto, Bambang. (2005). A longitudinal evaluation of accessibility: higher education Web sites. *Internet Research, 15(3), 281-294.*

Hricko, Many. (2003). *Design and Implementation of Web-Enabled Teaching Tools.* Hershey, PA: Information Science Publishing.

Jobe, M.M., (1999). Guidelines on Web accessibility for the disabled. *Colorado Libraries*, 25(3).

Loiacono, Eleanor, & McCoy, Scott. (2004). Web site accessibility: an online sector analysis. *Information Technology & People*, 17 (1), 87-95.

Mueller, John. (2003). *Accessibility for Everyone: Understanding the Section 508 Accessibility Requirements*. Berkeley, CA: Apress.

National Council on Disability (NCD), (2001). *The accessible Future*. Washington, D.C.: NCD. Retrieved July 10, 2001 from http://www.ncd.gov/newsroom/publications/accessiblefuture.html

Paciello, Michael G. (2000). *Web Accessibility for People with Disabilities*. Lawrence, Kansas: CMP Books.

Pilling, D., Barret, P., & Floyd, M. (2004). Disabled people and the Internet: Experiences, barriers and opportunities. Retrieved December 4, 2006 from http://www.jrf.org.uk/bookshop/details.asp?pubID=597.

Rouse, V. (1999). Making the WEB accessible. *Computers in Libraries*, 19(6).

Slatin, John M. & Rush, Sharon. (2003). *Maximum Accessibility: making your Web site more usable for everyone*. New York: Addison-Wesley.

Schmetzke, Axel. (2001). Distance Education, Web-resources design, and compliance with the Americans with Disabilities Act. *Proceedings of the Tenth National Conference of College and Research Libraries*, March 15-18, Denver, CO. Chicago: Association of College and Research Libraries, 137-142. Retrieved Oct 20, 2003 from http://www.ala.org/Content/NavigationMenu/ACRL/Events_and_Conferences/schmetzke.pdf.

Section 504. (1973). Pub.L. No. 93-112, § 504, 87 Stat. 355, 394 (codified as amended at 29 U.S.C. § 794 (1994).

Section 508. (1998). Pub. L. No. 105-220, §112, Stat. 936 (codified at 29 U.S. C. § 798).

State Grants Program for Technology-Related Assistance for Individuals With Disabilities, (2002). 34 C.F.R. § 345.

Summary of Section 508 Standards. (2007). Retrieved October 9, 2007 from http://www.section508.gov/.

Valenza, Joyce Kasman, (2000). Surfing Blind. *Library Journal*, Fall 2000 Supplement Net Connect, 125 (14), 34+.

Web Accessibility Initiative (WAI). (2000). Web Content Accessibility Guidelines (WCAG). Retrieved October, 20,2003 from http://www.w3.org/TR/WAI-WEBCONTENT/.

Web Accessibility Initiative (WAI). (2007). WCAG 2 FAQ. Retrieved October 19, 2007 from http://www.w3.org/WAI/WCAG20/wcag2faq.

Web Accessibility Initiative (WAI). (2007). Overview of WCAG 2.0 Documents. Retrieved October 19, 2007 from http://www.w3.org/WAI/intro/wcag20.php.

KEY TERMS

ALT-Text: Stands for Alternative Text, primarily used to render graphics when the image is not being displayed.

Assistive Technology (AT): A term includes assistive, adaptive, enabling and rehabilitative devices and the process used in selecting, locating, and using them to access information. AT promotes greater independence for people with disabilities by enabling them to perform tasks that they are normally unable to accomplish, or have great difficulty accomplishing.

Americans with Disabilities Act (ADA): U.S. public law enacted 1990 ensuring rights for people with disabilities. This legislation mandates reasonable accommodation and effective communication.

Section 508: Section 508 of the Rehabilitation Act is U.S. legislation that establishes requirements for electronic and information technology developed, maintained, procured or used by the federal government.

W3C: The World Wide Web Consortium (W3C) develops interoperable technologies such as: specifications, guidelines, software, and tools to lead the Web to its full potential. W3C is a forum for information, commerce, communication, and collective understanding.

Web Accessibility Initiative (WAI): Established by W3C and the Yuri Rubinsky Insight Foundation in 1997, the WAI works with organizations around the world to promote Web accessibility in five key areas: technology, guidelines, tools, education and outreach, and research and development.

Web Content Accessibility Guidelines 1.0 (WCAG 1.0): The official set of guidelines published by the W3C to assist Web content developers in the creation of accessible Web sites. The guidelines established three priority levels with 64 checkpoints for Web developers to meet.

Web Content Accessibility Guidelines 2.0 (WCAG 2.0): WCAG 2.0 is organized around four design principles of Web accessibility. Each principle has guidelines, and each guideline has success criteria at level A, AA, or AAA. The basis for determining conformance to the WCAG 2.0 are the success criteria. Currently the WCAG 2.0 is still at its Working Draft stage. It is expected to become W3C Recommendation (Web Standard) in early 2008.

Improving Data Quality in Health Care

Karolyn Kerr
Simpl, New Zealand

Tony Norris
Massey University, New Zealand

INTRODUCTION

The increasingly information intensive nature of health care demands a proactive and strategic approach to data quality to ensure the right information is available to the right person at the right time in the right format. The approach must also encompass the rights of the patient to have their health data protected and used in an ethical way. This article describes the principles to establish good practice and overcome practical barriers that define and control data quality in health data collections and the mechanisms and frameworks that can be developed to achieve and sustain quality. The experience of a national health data quality project in New Zealand is used to illustrate the issues.

BACKGROUND

Tayi and Ballou (1998) define data as "the raw material for the information age." English (1999) builds on the idea of information as being data in context, with knowledge being information in context, where you know the significance of the information. Translating information into knowledge requires experience and reflection.

Klein and Rossin (1999) note there is no *single* definition of data quality accepted by researchers and those working in the discipline. Data quality takes a consumer-focused view (consumers being people or groups who have experience in using organisational data to make business decisions) that quality data are "data that are fit for use" (Loshin, 2001; Redman, 2001; Wang, Strong, & Guarascio, 1996). Data quality is 'contextual'; the user defines what is good data quality for each proposed use of the data, within its context of use (Pringle, Wilson, & Grol, 2002; Strong, Lee, & Wang, 1997). Therefore:

Data are of high quality if they are fit for their intended uses in operations, decision-making, and planning. Data are fit for use if they are free of defects and possess desired features (Redman, 2001).

Data quality is now emerging as a discipline, with specific research programmes underway within universities, the most significant being that of the Sloan School of Management Information Quality Programme at the Massachusetts Institute of Technology (MIT)[1]. The field is based upon the well-established Quality Discipline, drawing on the work of Deming (1982) and the adaptation of the "*plan-do-check-ac*" cycle (the Deming Cycle). It also draws upon the "*quality is free*" concept of Crosby (1980) arising from the notion that doing things wrong is costly, and imports the ideas behind the Six Sigma approach and Total Quality Management (Juran & Godfrey, 1999) adapted to Total Data Quality Management (TDQM) and the management of information as a product (Wang, Lee, Pipino, & Strong, 1998).

The research programmes are developing ways to combine TDQM with the strategic direction of the organization, aligning the data quality requirements with overall goals. At present, there is little research published in this area, although some organizations do have data quality programmes with some strategic alignment to the business requirements.

Data quality is also becoming an increasingly important issue for health care providers, managers and government departments. The movement towards total quality management in health care to improve patient safety and health care efficiency is demanding high quality information. Further, evidenced based care requires the assimilation of large amounts of relevant and reliable research data available at the point of clinical decision making. Strategic prevention, national consistency of improvement practices, evolving data standards, and targeted improvements with increasing consumer involvement are all moving health care towards a TDQM model of data quality management.

Using the TDQM model, many of the methods developed in other industries can be useful for improving health care data. However, health data differ from data that arise in other fields in several ways. For example, the complex and multidisciplinary nature of medicine means that, unlike other sciences, health care data have no internationally standardized terminologies. This lack of standardization produces many homonyms, where the same term can mean different things depending on context, and synonyms, where there are several ways of expressing the same meaning. Added to these sources of ambiguity is the plethora of healthcare abbreviations which have no agreed format. The required longevity, privacy, and confidentiality concerns associated

with health data are also distinguishing characteristics. Combine these inherent features with the need to share and integrate the data across distributed and disparate entities such as the Ministry or Department of Health, regional health boards, hospitals, general practices, and individual specialists, and it is hardly surprising that the improvement of healthcare data quality is multifaceted and more involved than in many other domains.

THE DEVELOPMENT OF A DATA QUALITY EVALUATION FRAMEWORK

A common imperative in the health sector is the need to structure and improve the management of data quality within regional or national health data collections that aggregate patient health data and use the combined data for epidemiological and service planning purposes. Clinicians and managers recognize the need for assessment tools that indicate the level of data quality, identify where the problems are, and who should be accountable. This information can then be used to construct a strategy for quality improvement. This was recently the situation in New Zealand where previous work to develop the Ministry of Health information systems strategic plan, and a current state analysis of data quality, provided further support for the commencement of a programme of data quality improvement.

The starting point for any Data Quality Evaluation Framework (DQEF) in the health sector is the pioneering work of the Canadian Institute for Health Information (CIHI) (2003a, 2003b) The CIHI has developed a comprehensive data quality framework based on a hierarchy of quality attributes ranging from criteria at the base level that are grouped into characteristics, and then further reduced into a small number of high-level dimensions[2]. The CIHI framework is of course developed for Canadian conditions and must be adapted for local circumstances to ensure applicability. This can be done by engaging data custodians and stakeholders in an iterative sequence of focus group, interviews, and discussions that modify the data attributes at each level to reflect existing organizational structures and practice. The iterative process is most readily facilitated by action research using grounded theory (Strauss & Corbin, 1998) to structure the analysis of qualitative data through inductive coding and comparison.

In the New Zealand project, for example, analysis confirmed the utility of the Canadian quality dimensions of accuracy, relevancy, timeliness, usability, and comparability, but found it necessary to add a dimension of privacy and security to ensure that this aspect is explicitly managed to meet New Zealander's expectations for national data collections. The project also revealed that New Zealand's unique patient identifier, the National Health Index (NHI), is a powerful tool for the location of poor quality (e.g., redundant or incorrect) database entries.

The New Zealand project also revealed the advantages of the action research approach. Discussion and feedback, coupled with rigorous analysis, raises awareness of the importance of data quality, an issue which is frequently underrated at all levels by data users and managers. The approach similarly creates a common understanding of terminology and recognition of the importance that quality improvement is an ongoing venture needing continued support and clear lines of accountability. Thus, in developing and applying a data quality evaluation framework, it is important to:

- define the underpinning data quality criteria carefully involving all stakeholders to ensure common understanding and direction;
- consider the critical quality dimensions that reflect how the organization uses data and how data flow throughout the business processes;
- document business processes identifying data sources and their reliability;
- appreciate that the framework requires practical supporting tools, for example, documentation, to make it effective;
- customize the language of user guidelines or manuals with regard to the level and experience of the intended users;
- be aware of the importance of both education and training at all necessary stages and levels – training is essential to affect the culture change that must accompany the realization of the importance of data quality; and
- be aware that application of the framework is an iterative, on-going process – the required outcomes cannot be achieved in a single-pass.

THE DEVELOPMENT OF A DATA QUALITY IMPROVEMENT STRATEGY

Porter (1991) states that the function of a strategy is to integrate the activities of diverse functional departments to ensure consistency within an organization with explicit, reinforcing goals and policies when senior management cannot participate or monitor all decisions directly. It will be apparent from the above discussion that a Data Quality Evaluation Framework will not, by itself, automatically lead to required improvements and that a strategy, as defined by Porter, is critical to its successful implementation. Robson (1997) also notes that a successful strategy "exploits opportunities and fits the circumstances at the time" with a requirement, therefore, to undertake a systematic, skillful, accurate and realistic assessment of the opportunities and to re-evaluate them. The design and execution of a Data Quality Improvement Strategy (DQIS) in health care therefore begins with a current state analysis followed by an assess-

ment of the capability and maturity of the relevant health care organizations to effect the necessary improvements and understand the implications. Benchmarking against international organizations and other business sectors also helps to root the strategy so that it identifies local priorities and exploits current capabilities.

As is typical of many health systems at the present time, an extensive analysis of practices in New Zealand found that data quality management is entirely initiated through "bottom up" processes, in general through information services or information technology teams. Management has not yet taken responsibility, and indeed does not understand their role in the data quality process. Data quality improvement tends to occur when staff become frustrated with their inability to use the data for its intended purpose and there is no one else in the organization responsible for overall data quality management. This bottom up approach introduces silos of data quality improvements that can lead to even more quality issues by increasing discrepancies and producing conflicting reports from the same source. The New Zealand analysis shows that this situation is evident in all types of health sector organizations. It confirms that TDQM and the strategic management of data quality are relatively new phenomena, still rarely found across organizations in either New Zealand or overseas. New Zealand health care is certainly not lagging behind in the management of data quality but its political system, national public sector provision, structured assessment tool, and the NHI suggest that it has the infrastructure and capability to become a world leader in the implementation of a strategic data quality programme based on TDQM principles.

The New Zealand DQIS thus provides the Ministry of Health and the sector with detailed guidelines on how to develop and implement TDQM at all levels of the health sector. Roles and responsibilities are clearly defined, along with data ownership. A series of projects provides the required development for "business as usual" initiatives that institutionalize data quality into every day practice and make use of existing sector knowledge through the development and dissemination of best practice guidelines. For a data quality improvement strategy, it is important therefore to:

- derive and impose standards that facilitate data and information transfer whilst preserving quality;
- re-engineer the business processes to deliver the quality data needed for efficient service planning and the effective practice of integrated patient care;
- identify and disseminate best practice to reduce the development time needed to improve data quality;
- ensure data quality levels are not unnecessarily rigorous to maintain user ownership and workloads at reasonable levels;
- define user accountabilities for data quality and the mechanisms to enforce them; and
- seek to embed the search for data quality in normal working practices and recognize its achievement in appropriate ways such as accreditation.

As these requirements suggest, the strategic management of data quality requires wide ranging skills from practitioners with experience in many different areas. The challenge is, therefore, to encourage these practitioners to develop an interest and ownership in data quality and then to undertake further study to apply their skills to the development of the necessary strategic programmes.

There is considerable theory available in disciplines outside of data quality to support the New Zealand findings that organization wide teams enhance the learning and innovation of those who participate leading to improvements throughout the organization as a whole. Encouraging staff to develop solutions themselves helps to institutionalize data quality in the organization, through the development of emergent strategy. Data quality practitioners are empowered to implement "business as usual" initiatives that they themselves believe work. This emergent strategy is guided by the overarching organizational or national strategy that provides the simple rules that encourage the seeds of innovation and maintain momentum towards the defined vision.

Whilst comparison with data quality initiatives in other sectors is important, the development of a data quality framework and strategy for health care must always recognize the inherent complexity of the "business" of health. Plsek and Greenhalgh (2001) introduce the science of Complex Adaptive Systems (CAS) to help understand and bring about change in the health care environment. A complex system is defined as "a system with many independent agents, each of which can interact with others" (Penchas, 2003), that can behave very sensitively and be influenced by small initial differences (Champagne, 2002). Health care systems are just such systems creating complex adaptive interactions that contain emergent learning and change potential (Penchas, 2003).

In summary, the management and improvement of data quality in health care relies on an action research approach through collaborative groups to understand the relevant issues, an organization wide data quality team to develop the maturity capability of the organization, and an awareness of the inherent complexity of health care that demands multidisciplinary solutions to multidisciplinary problems.

CONCLUSION AND FUTURE DIRECTIONS

The development of a Data Quality Evaluation Framework and a Data Quality Improvement Strategy as described in this article provide clear direction for a holistic and "whole of health sector" way of viewing data quality, enabling

organizations to implement local innovations through locally developed strategies and data quality improvement programmes. Simple rules, such as the TDQM process and the data quality dimensions guide the change, leaving room for innovation. With these essential features in place, practitioners can use a range of other devices to embed data quality awareness and accountability in normal practice and facilitate further change.

For example, experience in using the DQEF and the benefits of better data quality information will help to generate appropriate data quality metrics. Suitable metrics are paramount to the success of the DQEF and to provide data suppliers with applicable key performance indicators for expected levels of data quality. The New Zealand research referred to provided an embryonic methodology for metrics development via its ethnicity data collections but the methodology needs further refinement and validation through empirical testing.

Organizationally, data quality practitioners should have a pivotal role in the development and implementation of new systems, including ongoing training of data collection staff on the impact of errors and the downstream uses of the data. Further, the importance of data quality, and the staff who work in this area, needs to be recognized by management teams. Recruitment of staff should include those with the skills of strategic thinking and analysis. Providing such a team with a structured programme of data quality management, with clear roles and responsibilities for the organization to manage their data from a whole of system view, will enable data quality practitioners to concentrate on improvement initiatives.

Other valuable approaches include change management theories such as appreciative enquiry (Fry, 2002), which can help to encourage the utilization of existing organizational knowledge and enact change with a minimum period of diminished performance (Elrod, 2002). The theory of "complex systems of adjustment" (Champagne, 2002; Stacey, 1993) can be instilled in the organization to encourage change through the constant interaction of people throughout the organization. Champagne (2002) notes the similarities between complexity and learning theories. Both see change in a global, integrated way as forming part of the routine life of organizations, with the change process a collective one.

This research that underpins this article was undertaken in New Zealand, but the many similarities between international health care systems and the correspondence in data quality capabilities suggest strongly that the results are generally applicable. This raises the possibility of external benchmarking (Hamel & Prahalad, 1994) between international healthcare organizations to compare performance, develop best practice, and identify key challenges. Even so, the health sector environment has some way to go before reaching second and third generation data quality management as discussed by Redman (2001). Continued research into the prevention of poor quality data through process management could lead to a health sector where all data quality is managed through prevention rather than "find and fix." This may be a difficult, even elusive, goal but its pursuit will bring benefits to all citizens.

REFERENCES

Canadian Institute for Health Information. (2003a). Data Quality Framework. Retrieved October 13, 2005 from http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=quality_e

Canadian Institute for Health Information. (2003b). Earning Trust. Key Findings and Proposed Action Plan from the Data Quality Strategies Study. Retrieved October 13, 2005, from http://secure.cihi.ca/cihiweb/en/downloads/DADEarningTrust2003_e.pdf

Champagne, F. (2002). *The Ability to Manage Change in Health Care Organisations*: Commission of the Future of Health Care in Canada.

Crosby, P. (1980). *Quality is Free*. New York: Penguin Group.

Deming, W. E. (1982). *Out of the Crisis*. Cambridge: Massachusetts Institute of Technology.

Elrod, P. D. (2002). The death valley of change. *Journal of Organisational Change, 15*(3), 273-291.

English, L. (1999). *Improving Data Warehouse and Business Information Quality*. New York: John Wiley & Sons.

Fry, R. (2002). Appreciating your past in order to transform the future. Retrieved September 26, 2005, from http://www.clevelandshrm.com/.

Hamel, G., & Prahalad, C. K. (1994). *Competing for the Future. Breakthrough Strategies for Seizing Control of Your Industry and Creating the Markets of Tomorrow*. Boston: Harvard Business School Press.

Juran, J. M., & Godfrey, A. B. (1999). *Juran's Quality Handbook* (5 ed.). New York: McGraw-Hill.

Klein, B., & Rossin, D. F. (1999). Data errors in neural network and linear regression models: An experimental comparison. *Data Quality, 5*(1), 25.

Loshin, D. (2001). *Enterprise Knowledge Management. The Data Quality Approach*. California: Academic Press.

Penchas, S. (2003). Complexity in health care systems. *Complexus, 1*, 149-156.

Plsek, P., & Greenhalgh, T. (2001). The challenge of complexity in health care. *British Medical Journal, 232*, 625-628.

Porter, M. (1991). Towards a dynamic theory of strategy. *Strategic Management Journal, 12*, 95-117.

Pringle, M., Wilson, T., & Grol, R. (2002). Measuring "goodness" in individuals and healthcare systems. *British Medical Journal, 325*, 704-707.

Redman, T. C. (2001). *Data Quality. The Field Guide.* Boston: Digital Press.

Robson, W. (1997). *Strategic Management and Information Systems* (Second ed.). London: Pitman Publishing.

Stacey, R. D. (1993). *Strategic Management and Organisational Dynamics*. London: Pitman.

Strauss, A.., & Corbin, J. (1998). *Basics of Qualitative Research*. Thousand Oaks: Sage Publications.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103-110.

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM, 41*(2), 54-57.

Wang, R. Y., Lee, Y. W., Pipino, L. L., & Strong, D. M. (1998). Manage your information as a product. *Sloan Management Review.* (Summer), 95-105.

Wang, R. Y., Strong, D. M., & Guarascio, L. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33.

KEY TERMS

Complex Adaptive System: A system with many independent agents, each of which can interact with others.

Data Quality Dimensions: Quality properties or attributes of data; a set of data quality attributes that most data consumers react to in a fairly consistent way.

Data Quality Framework: A tool for the assessment of data quality within an organization; a vehicle that an organization can use to define a model of its data environment, identify relevant data quality attributes, analyse data quality attributes in their current or future context, and provide guidance for data quality improvement.

Data Quality Improvement Strategy: A cluster of decisions centered on organizational data quality goals that determine the data processes to improve, solutions to implement, and people to engage.

Datum: A fact or value assigned to a variable; single observational point that characterises a relationship. Data is the plural noun of datum

Emergent Strategy: A series of actions converges into patterns that become deliberate when the pattern is recognized and legitimized by senior management.

Grounded Theory: A method of extracting meaning and theories from data by systematically and intensively analysing and coding the data, sentence-by-sentence, or phrase-by-phrase.

Total Data Quality Management: An approach that manages data proactively as the outcome of a process, a valuable asset rather than the traditional view of data as an incidental by-product.

ENDNOTES

1 http://mitiq.mit.edu

2 Each familiar, high-level, dimension is defined in context by appropriate characteristics that answer "what is/are?" questions such as "what is the level of error?". Underpinning these characteristics are "criteria" that define processes and metrics used to assess the presence of potential data quality issues, for example, "Is the error within acceptable limits?"

Improving Public Sector Service Delivery through Knowledge Sharing

Gillian H. Wright
Manchester Metropolitan University Business School, UK

W. Andrew Taylor
University of Bradford, UK

INTRODUCTION

Since the publication of the first knowledge management article in *Harvard Business Review* (Nonaka, 1991), the world has witnessed a revolution in management practice. While the origins of knowledge management extend further back in history (see Prusak, 2001; Wiig, 1997), it is certainly true that in the last decade the creation, sharing and application of knowledge are increasingly seen as a source of competitive advantage. However, knowledge management is largely a private sector innovation at the present time, although gradually moving towards the public service sector (Bate & Robert, 2002; Hartley & Allison, 2002). The implementation of knowledge management places an emphasis on organizational factors such as learning capability, culture and leadership as well as renewed focus on the importance of information quality (Alavi & Leidner, 2001). The ability to manage the sharing of information (and hence knowledge) effectively remains one of the most important but still least understood activities in modern organizations, no less so in public services.

BACKGROUND: KNOWLEDGE SHARING IN CONTEMPORARY PARTNERSHIP ORGANIZATIONS

Public services represent a significant economic sector in most countries and public demands on services are increasingly consumerist. This has led to escalating scrutiny of the performance of public services. Consequently, the strategic use of information and knowledge to improve service delivery and financial performance has become a key skill for managers in this sector. Partnership working represents a formal departure from the traditional compartmentalized approach to public service delivery. Often referred to as joined-up thinking, partnership working challenges existing hierarchies, encouraging the partner organizations to work together at all levels, including strategy, service planning and service delivery to enhance efficiency and improve user experience and satisfaction. For partners to work effectively together, knowledge of best practices must be shared and utilized towards the common goal of improving the overall quality of service delivery.

Our research has focused on health and social care as an area of public service in which organizations responsible for commissioning and delivering all aspects of care are increasingly expected to work together, to reduce fragmentation of access to the user. Management of the provision of high quality public services continues to be a major social and political issue in many countries. Our research was conducted in the context of UK national policies for performance management (DETR, 2001a), partnership working (DETR, 2000, 2001b; Fordham, 1998), the reduction of health inequalities (DoH, 1998a, 1999), and overall improvements in service quality (DoH, 1998b, 2000). We have concentrated particularly on the issue of making public service partnerships work effectively, to achieve strategic objectives, that is, to improve individual health and personal well being as well as to achieve gains in public health. Specifically, our research questions relate to assessing the readiness of the partners to work together, and to share knowledge that each possesses about their part in the overall service delivery process. By understanding the factors that influence effective knowledge sharing, managers can take practical steps toward improving these antecedent preconditions.

MANAGING THE ANTECEDENTS TO KNOWLEDGE SHARING

The key to partnerships is a focus on the creation of an explicit understanding of what needs to be done to meet strategic objectives –akin to Choo's concept of a "knowing organization" (Choo, 1988). We conceptualize the role of knowledge in the partnership process (Figure 1) in terms of two core aspects, viz:

- The effective management of information to support the vertical deployment of organizational strategy in terms of communication and development of meaningful performance measures, and
- The wider organizational culture to support attitudes conducive to new ways of working.

Figure 1. The public service partnership: The knowing organization

Figure 1.
The public service partnership:
The knowing organization

We have identified six key factors that are associated with successful knowledge sharing in public sector partnerships (Wright & Taylor, 2003), namely:

- Innovative culture
- Change readiness
- Information quality
- Clarity of responsibility
- Strategy formulation and deployment
- Accountability

Innovative Culture

An innovative culture is one where people are receptive, rather than resistant to, new ideas, and where they are motivated to embrace and develop these ideas and shape them into improved working practices. Such cultures provide people with time to reflect, to learn from both success and failure, providing supporting systems to facilitate reflection and capture lessons learned. Finally, innovation is focused on the user or customer, whereby people actively search for new ways of improving service delivery. The legal and political constraints on public service managers and persistent demands for strict oversight can lead to rigidity and bureaucracy in public sector organizations that counter the development of an open and inclusive culture (Scott & Falcone, 1998).

Change Readiness

Change and innovation are closely linked. An innovative culture needs to be able to implement changes to working practices and behaviors generated by innovation. This requires a positive attitude to doing things differently, rather than seeking to maintain the status quo. Change requires leadership, to proactively seek opinions and listen to views whilst engendering an atmosphere where ideas are freely expressed and there is no perception of a need for staff to cover their backs to protect themselves from criticism and retribution. Involvement and commitment will decline and the organization's innovative potential will be diminished if there is a culture of reluctance to challenge current ways of working. High levels of media scrutiny of public sector organizations (Perry & Rainey, 1988) and the top-down nature of government-imposed changes (Collier, Fishwick & Johnson, 2001) can reduce public sector employees' receptivity to change (Halachmi & Bovaird, 1997). Being ready to change implies a concomitant sensing of the need to change. Information about performance gaps, that highlights the need for change, must be communicated throughout an organization. If people feel that managers pay little attention to performance statistics, they too will ignore them and continue working in ways that maintain the status quo. The nature of the change and the benefits that it will bring need to be understood.

Quality of Information

Good quality information facilitates performance review, and reflection on service delivery. It supports people in their work tasks and it provides a medium for the capture and dissemination of lessons learned. If timely and meaningful information is not provided, people will find it difficult to know how well they are performing, and they will spend extra time searching for the information they really need. Public sector organizations often place less importance on the quality of information and perceive less need to invest in information systems (Rocheleau & Wu, 2002). Unless there is clarity about the basis of performance measurement, information systems will not be perceived as providing appropriate support.

Clarity of Responsibility

It is important for people to understand their specific roles and responsibilities and to know whom to contact elsewhere in the service value chain. People need to see clearly how their jobs fit into the bigger picture, to have their responsibilities delineated clearly in relation to the organizational strategy, and to see how their roles contribute to its achievement. Thus clarity of responsibility is concerned with the effectiveness of strategy delivery. Managers must ensure that people can grasp the significance of strategy in relation to their own responsibilities, and that the performance measures that derive from the strategic process (see Figure 1) are useable for managing service delivery. All too often, public sector performance measures are regarded by employees as vague and of limited utility (Townley, 2002). To achieve clarity of responsibility in partnerships there needs to be:

- A joint strategy developed, and owned by all partners
- Re-definition of the service value chain and its business processes
- Re-examination of roles and responsibilities within the value chain
- Development of partnership-based performance measures
- Logical and explicit derivation of performance measures from the joint strategy

Strategic Connections

This factor addresses communication gaps in the strategy formulation and deployment process. Public sector managers must ensure that people feel involved in the strategy formulation process and do not feel that strategy is imposed from on high, as is often the case in public services. Staff need to understand the meaning of strategy in their own situations. Strategic plans need to be living documents, owned by all, rather than uninspiring rhetoric that gathers dust on a shelf in the strategist's office. In public service partnerships, strategy must be communicated in terms of improved relationships and outcomes for service users, rather than a structural end in itself.

Accountability

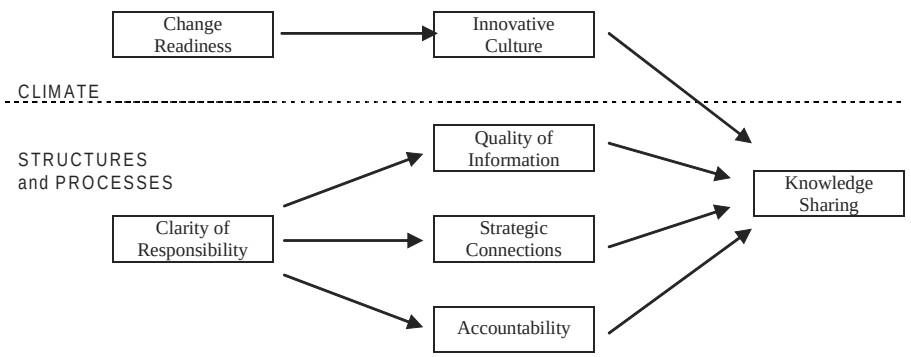
People need to know where the buck stops. In an integrated service delivery chain that crosses organizational boundaries, it must be made clear who is ultimately accountable for performance. Poor redefinition of accountability is symptomatic of, and consistent with, top-down imposition of strategy with little staff involvement, and poor deployment of strategies into meaningful processes and activities.

FUTURE TRENDS: MANAGING KNOWLEDGE SHARING

Six antecedents to successful knowledge sharing are illustrated in Figure 2, and these form a basis for the management of knowledge sharing. Increasingly, organizations will move from simply practicing knowledge management and knowledge sharing, toward an emphasis on evaluating its effectiveness. Using this model as a basis for diagnosing an organizational climate that is supportive of knowledge sharing, a targeted, proactive approach can be developed for implementing and maintaining effective knowledge sharing.

Figure 2. Predictors of knowledge sharing

Figure 2. The predictors of knowledge sharing



This empirically validated model indicates that these six factors are important precursors in the development of an effective knowledge sharing process and can give guidance to managers about where they should focus their efforts. We have found that an innovative culture is the strongest factor in explaining knowledge sharing, with acceptance of new ideas being the strongest element within this factor.

Our research identified the issues of trust and power as being underpinning and tacit impacts on knowledge sharing and our factors include items that embody trust and power (Dirks & Ferrin, 2001). For example, trust can result in higher levels of co-operation, increased willingness to take risks (Mayer, Davis & Schoorman, 1995), increased involvement of employees in decision making (Spreitzer & Mishra, 1999), increased sharing of information and greater satisfaction with organizational change programs (Rousseau & Tijoriwala, 1999). Thus this model incorporates the effects of trust and power in the ways in which they are experienced in a knowledge sharing environment.

CONCLUSION: A MANAGEMENT AGENDA FOR KNOWLEDGE SHARING AND INNOVATION

There are some clear antecedents to the management of effective knowledge sharing in partnerships developed to enhance public services. The change implications of knowledge sharing are very important, and are represented strongly in the innovative culture and change readiness factors. These factors are well established as key strategic management challenges and underline that knowledge sharing is a key strategic issue, rather than an IT-centred initiative, as it is often portrayed. While public policy usually addresses structural and process issues, our research suggests that this may inadvertently be misguided and that the key to improved service delivery lies in changes to the underlying culture, particularly with regard to an innovative orientation and change readiness. Applying structural solutions to behavioral problems is not recommended.

This model of knowledge sharing suggests a management agenda to improve service delivery through knowledge sharing. To develop an innovative culture, there must be a climate where both motivation to innovate and acceptance of new ideas are equally strong, and further, there must be appropriate systems in place to facilitate learning and reflection. To generate a shared understanding of best practices and a sharing of knowledge about lessons learned, staff in such partnerships need to be given time and opportunities to engage in socialisation processes that facilitate learning (Nonaka & Takeuchi, 1995). Organizations need to take on board Garvin's assertion that while much knowledge can be generated from reflecting upon success, there is even more to be gained from reflection on failure (Garvin, 1993). The changes needed to move from a departmental focus to an inter-organizational one point very forcibly to the organizational culture and especially the role of senior management.

Public services are embedded in an environment wherein government requirements for performance reporting and accountability often seem to have little relevance for informing staff about what needs to be improved. The intangibility of public services means that public service managers rely very much on knowledge – insight, understanding and empathy. Extant studies of knowledge have generally focussed on the private sector. However, the prominence of a knowledge sharing culture is clearly a key element in achieving a partnership-based, user-focused public service.

REFERENCES

Alavi, M., & Leidner, D.E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107.

Bate, S.P., & Robert, G. (2002). Knowledge management and communities of practice in the private sector: Lessons for modernizing the National Health Service in England and Wales. *Public Administration, 80*(4), 643-684.

Choo, C.W. (1988). *The knowing organisation*. New York: Oxford University Press.

Collier, N., Fishwick, F., & Johnson, G. (2001). The processes of strategy development in the public sector. In K. Scholes (Ed.), *Exploring public sector strategy* (pp. 17-32). London: Pearson Education.

DETR. (2000). *Local government act*. London: Department of the Environment Transport and the Regions, the Stationary Office.

DETR. (2001a). *Best value and audit commission indicators for 2001/2002: Consultation*. London: Department of the Environment Transport and the Regions, the Stationary Office.

DETR. (2001b). *Local strategic partnerships: Government guidance*. London: Department of the Environment Transport and the Regions, the Stationary Office.

Dirks, K.T., & Ferrin, D.L. (2001). The role of trust on organizational settings. *Organization Science, 12*, 450-467.

DoH. (1998a). *Independent inquiry into inequalities in health report (the Acheson report)*. London: Department of Health, The Stationary Office.

DoH. (1998b). *The new NHS: A national framework for assessing performance: Consultation document*. London: Department of Health, NHS Executive.

DoH. (1999). *Reducing health inequalities: An action report*. London: The Stationary Office.

DoH. (2000). *NHS plan: A plan for investment, a plan for reform*. London: Department of Health, the Stationary Office: Cm 4818.

Fordham, G. (1998). *Building partnerships in the English regions: A guide to good practice*. DETR.

Garvin, D.A. (1993). Building a learning organization. *Harvard Business Review, 71*(4), 78-92.

Halachmi, A., & Bovaird, T. (1997). Process reengineering in the public sector: Learning some private sector lessons. *Technovation, 17*(5), 227-235.

Hartley, J., & Allison, M. (2002). Good, better, best? Inter-organizational learning in a network of local authorities. *Public Management Review, 4*(1), 101-118.

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*, 709-734.

Nonaka, I. (1991). The knowledge-creating company. *Harvard Business Review, 69*, 96-104.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company: How Japanese companies create the dynamics of innovation*. New York: Oxford University Press.

Perry, J.L., & Rainey, H.G. (1988). The public-private distinction in organization theory: A critique and research strategy. *Academy of Management Review, 13*(2), 182-201.

Prusak, L. (2001). Where did knowledge management come from? *IBM Systems Journal, 40*(4), 1002-1007.

Rocheleau, B., & Wu, L. (2002). Public versus private information systems: Do they differ in important ways? A review and empirical test. *American Review of Public Administration, 32*(4), 379-397.

Rousseau, D., & Tijoriwala, S. (1999). What's a good reason to change? Motivated reasoning and social accounts in promoting organizational change. *Journal of Applied Psychology, 84*, 514-528.

Scott, P.G., & Falcone, S. (1998). Comparing public and private organizations: An exploratory analysis of three frameworks. *American Review of Public Administration, 28*(2), 126-145.

Spreitzer, G., & Mishra, A. (1999). Giving up control without losing control. *Group and Organization Management, 24*, 155-187.

Townley, B. (2002). The role of competing rationalities in institutional change. *Academy of Management Journal, 45*(1), 163-179.

Wiig, K.M. (1997). Knowledge management: Where did it come from and where will it go? *Expert Systems with Applications, 13*(1), 1-14.

Wright, G.H., & Taylor, W. A. (2003). Strategic knowledge sharing for improved public service delivery: Managing an innovative culture for effective partnerships. In E. Coakes (Ed.), *Knowledge management: Current issues and challenges* (ch. XV, pp. 187-211). Hershey, PA: IRM Press.

KEY TERMS

Accountability: Transparency of responsibility for performance, the management of performance and resulting implications for the deployment of future resources.

Change Readiness: An organizational mindset that welcomes challenges to established structures and processes and administrative orthodoxies.

Clarity of Responsibility: An understanding of the roles and responsibilities of individuals and business units working together to deliver a holistic service proposition.

Information Quality: The accuracy, completeness, timeliness and utility of performance related information that is used as the basis of management decision making.

Innovative Culture: An organizational climate that fosters new ideas and reflection on learning from experiences.

Knowledge Sharing: Formal, deliberate and systematic activities of transferring or disseminating knowledge from one person, group or organization to another.

Partnership Working: A network of organizations working together to provide a total service offering to a targeted group of users.

Public Services: Social infrastructure services, delivered wholly or partly with the benefit of public funds and strategically driven through national or regional administrations.

Strategic Connections: The relationships between the strategy formulation process and the deployment of resources to achieve it.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1414-1418, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Improving the Usability in Learning and Course Materials

Maria Elizabeth Sucupira Furtado
University of Fortaleza and Estadual of Ceara, Brazil

INTRODUCTION

Human-computer interaction (HCI) is a discipline concerned with the study, design, and development of high-usability interactive systems (ISs) focusing on users' needs and their experiences with technologies, among others. In a simplified way, the usability of an IS refers to how easy it is to use and to learn. HCI is a very broad discipline that encompasses different specialties with different concerns regarding computer development: software engineering (SE) is concerned with the design and development of high-quality ISs focusing on schedule, budget, communication, and productivity. The quality of an IS refers to how satisfied the system clients and/or users are, verifying whether the system is performing exactly what was requested.

In order to achieve both IS usability and quality, it is necessary to go beyond designing user interfaces (UIs) and that they are easier to use and learn. It is important to define methods and use techniques (as ethnographic, semiotic, prototypes), which help designers to understand HCI concepts and build better interactive artifacts (as widgets) and to understand the effects that systems will have on humans (Cooper & Reimann, 2003). Some HCI concepts are characteristics of users (such as their preferences, language, culture, and system experience) and their contexts of use (such as great familiarity with a device, easy accessibility, and good luminosity of the environment).

In the interactive learning context, it is necessary to consider HCI concepts into an interactive learning system development method. The pedagogic usability of an interactive learning system is related to how easy and effective it is for a student to learn something using multiple devices (such as palm, camera, cell phone) to interact with the system. For these reasons, it is important not only to think about the IS quality, but about its usability as well. In this text, an interactive learning system is composed of a virtual learning environment (VLE), with tools to support a collaborative learning and interactive course materials available for the users through this environment. So, it is important not only to think about the VLE usability, but also about the interactive course material usability.

We have identified some problems to achieve a successful deployment of interactive learning systems (Furtado, Mattos, Furtado, & Vanderdonckt, 2003):

- **Lack of learning quality:** Many academic staffs are worried about the learning process quality through the course materials available in VLE. However, the material of a face-to-face course is hardly ever adapted to the students' needs and experiences. This way, it is expected that a VLE allow students exploring possibilities brought by new technologies in order to participate in the elaboration of this material.
- **Lack of adaptive tools:** Learning systems are very useful, but most of them are not adaptive and neither consider the user experiences with technology. Interactivity and personalization are factors that help for allowing a user participating in the community, which he or she makes part of (McCarthy & Wright, 2004).
- **Lack of training in modern and collaborative technologies:** Any academic staff (such as a teacher), as part of his/her professional development, needs continuous and sophisticated training. Such training should help overcoming the limits found by this community in accessing to digital technologies for the creation of interactive information and multimedia content, in a collaborative way with their own students. It is necessary to fulfill these needs by adopting an integrated pedagogical-technological content (Perrenoud, 2001).

All of these issues have a critical impact on the usability and quality of interactive learning systems. Thus, we developed a general architecture for such systems, which aims to show the concepts that must be considered to increase the quality of the learning process and to increase their UIs usability.

The remainder of this article is structured as follows: in the next section, we explain the main concepts that helped us to develop such general architecture. Then, we provide the best practices used in a development cycle of an IS, focusing on the usability issue. Finally, we summarize the main points of this text.

BACKGROUND

There is a trend about technological convergence, which companies are thinking in providing users common access to content by using any device in any place. It involves dif-

ferent technologies such as mobile phones interconnected with other surrounding interfaces (e.g., i-TV, PCs, PDAs, in-car-navigators, smart-house appliances, etc.) (Roibás, Geerts, Furtado, & Calvi, 2006). This technological convergence will be decisive in the creation of pervasive virtual learning environments. Many HCI researchers focus on the interaction design process in order to build UIs to these environments with which the users (students) can interact with no usability problem. In this process, several design decisions are made concerning the system navigation, the feedback mechanism, and the information organization and by taking into account the users' device and their context of use. However, it is important to point out that the user controls a VLE for the purpose of learning the content (course material). Then improving the usability of these systems is also to improve the usability of the multimedia content. A system UI should ideally be designed to be an integral part of the content (Garrett, 2003). It means that the UI design paradigm involving the task efficiency concepts (such as response time, errors control, task completion) must be extended to involve the content efficiency concepts (such as to allow users to read a content again, to stop of seeing it and to return when they want, to show someone an interesting content, to add a comment, etc). By combining these efficiency concepts, users can have more control of the systems and consequently, they can change their posture of passive learner to a more active one. In Chorianopoulos and Spinellis (2006), a UI evaluation approach was described in which the content quality is a relevant part to the quality of interactive television systems. Mattos (2005) described an approach, which can be used by teachers to persuade and motivate their learners to define their own contents.

VLE AND INTERACTIVE COURSE MATERIAL BACKGROUND

As we have mentioned before, an interactive learning system is composed of a VLE and tools for creation in a collaborative way of interactive course instructional materials.

A VLE has to provide students with spatial freedom and time flexibility. It has to be flexible enough so that every student may profit from his or her own skills and abilities, use his or her previously developed idiosyncratic characteristics (cognitive, social, or emotional), and apply his or her previously gained experience and expertise (Karoulis & Pombortsis, 2003). Some tools available in a VLE are links to tutorials and course materials, collaborative tools (as blog, skype), evaluation tools, and administrative tools.

There are some authoring tools, which users (students and teachers) can use to develop their own contents and make them available and accessible in various devices (such as palmtop, digital television, kiosks, and mobile phones) (Maia & Furtado, 2006). The user of this system follows a flow of activities to edit and update the Web content and to design the UIs of these contents. During UI design, HCI patterns are made available in order to assure the consistence between user interface objects of different devices. We can define HCI patterns as a tested solution for a usability problem (such as lack of orientation, difficulty in finding information) that happens in a certain context (search, visualization, etc.). Other tools, such as those for specific programming languages (HTML, FLASH, SVG), are only used by specialized teams.

BASIC CONCEPTS RELATED TO USABILITY IN INTERACTIVE LEARNING SYSTEMS

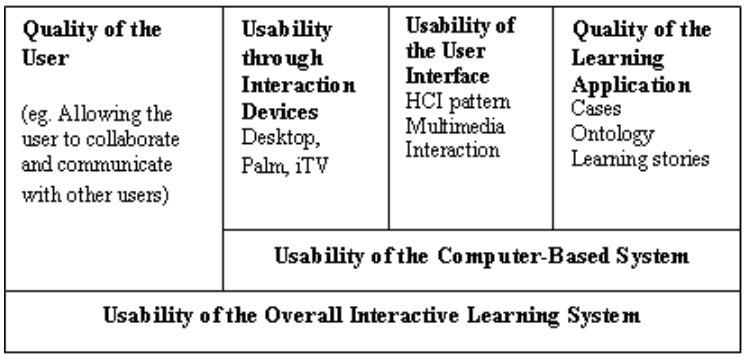
The general architecture initially proposed in Furtado et al. (2003) but updated here in Figure 1 aims at the development of VLE and interactive materials, taking into account some concepts studied in different areas (human-computer interaction, cognitive sciences, ergonomic, artificial intelligence, and pedagogy).

According to Figure 1, an interactive learning system's usability can be assured when its components have been built with quality and when users' needs have been taken into account. Quality of a component means: (1) quality in the application that corresponds to content, which refers to the information and knowledge involved in the system. Information (such as learning stories and objects) is related to the development of instructional materials, and knowledge (such as cases) is especially related to the collaborative practices; (2) usability in the UI, which refers to a good specification of the interactive information of the system (its windows, its buttons, etc.); and (3) usability through interaction devices, which makes the interaction with different media (sound, text, image) possible through different interaction resources as pen for palmtops, cameras and microphones, remote control for iTV, and so on. The quality of the user refers to his or her ability to use new interaction devices and technologies, experience with technologies, and acquaintance of the domain in question.

The concepts related to usability in an interactive learning system are:

- Utilization of ontology to assure the flexibility in modeling learning applications. The ontology notion comes from the artificial intelligence area where it is identified as the set of formal terms with one knowledge representation, since the representation completely determines what "exists" in the system (Guarino, 1995). During an application modeling, models (such as the user model), knowledge (such as cases studies), and

Figure 1. The general architecture proposed



learning stories and their learning objects associated to an instructional material can be represented using ontology. The advantage of using this representation is that the ontology can be defined once and used as many times as necessary (evolutionary approach). In addition, the ontology is useful to create learning objects and reuse them when a new course is initialized.

- Utilization of human factors and HCI patterns to assure effective interaction, maximum performance, and flexibility (multiplicity of ways the user and the system exchange information). Human factors such as the teachers' beliefs and guidelines related to graphic aspects and characteristics of the users and their context of use must be considered. Guidelines are suggestions about the ergonomic aspects of the interfaces such as showing only the necessary information or letting the user control the system dialog. Many HCI patterns are associated to guidelines, in order to help the designer to determine the best way in which the information is to be provided to users and to ensure optimal accessibility of the system (Appleton, 2000).
- Convergence of various devices and resources to improve interaction. The use of different interaction devices makes it possible to explore the possibilities brought by mobile and iTV technologies, in order to bring more facilities of access to information. As consequence, special attention should be given to issues such as sociability, creativity, and context awareness. It implies in developing systems that employ appropriate mechanisms of feedback, and content sharing and creation.
- Utilization of collaboration and communication mechanisms to assure the quality of the user and the continuous usability of the learning system. Some VLEs implement the collaborative aspect by allowing users to share an application (such as TELE (Neto, Raimir, Bezerra, & Sarquis, 2001)) and/or by using forums and chats. It allows users to share knowledge, when they are motivated to collaborate through principles of participation, for instance, in problem definition practical situations. There is a need to continuously assure usability of a system due to technological changes and to the evolution of users' needs. We believe instructional materials of a course should evolve from human interactions between students and teachers occurring within an interactive discussion forum (Mattos, Maia, & Furtado, 2003). Hence, we believe in a collaborative process between users and designers to adjust accessibility, acceptability, and usability criteria of a system.

USABILITY OF INTERACTIVE LEARNING SYSTEMS: THE BEST PRACTICES OF REQUIREMENTS ENGINEERING

In this section, the concept of usability (learnability, flexibility, and robustness) is related to some best practices of the requirements modeling and validation of an IS (see Table 1).

An IS must be developed with the participation of users throughout the development process because it is easier for developers to define and evaluate the functional and nonfunctional (usability) requirements. The functional requirements of a VLE are related to the tasks that the user wants to perform (user tasks), for instance, to interact with other students and tutors, and to access the rules and regulations of the course. Usability requirements are related to users' satisfaction and the performance of the system. These requirements directly influence aspects of the system quality of use (e.g., never lose sight of navigational functions).

Sutcliffe (2002) gives practical guidance for requirements modeling and validation based on scenarios. A scenario represents a story or example of events taken from real-world

experience. These stories are close to the common sense use of the word and may include details of the context of use for a system. These representations help users think about how the system would support their tasks.

During the material design, users must be able to inform their learning requirements, which are related to content (students must perform the study tasks for a particular unit). These requirements are usually represented in storyboard sketch or animated sequences. A storyboard represents a future vision of a designed material with sequences of behavior and possibly contextual description. These representations serve as a starting point in determining the information that should have a specific material.

However, a single scenario or storyboard shows only some possible sequences of events among many possible sequences permitted during the interaction of a user with an IS. Interactive prototype is an interactive medium that allows the users to explore all alternative paths, and it gives a look and feel overview of the IS. To do a prototype, the designers should take into account HCI patterns. During a session of requirements validation, it is necessary to actively engage users in checking that the designed IS actually does what they want. If the IS developed is accepted by the user, we can say that its development process was user centered, according to their needs and suggestions. As we mentioned, the usability of an interactive learning system is not just the UI. The focus must also be on the support that the VLE or instructional material provides for its interactive students.

As requirements are so volatile, it is important to examine how they can be specified so software can evolve and adapt to the users´ needs. The evolutionary process involves continuous requirement adjustments in two points of view: (1) of the user—when using the system, his or her practices and working methods can be adapted to satisfy the evolving needs individually and/or collectively; and (2) of the system—the system's behavior can be adapted. The evolution of users' needs can involve modifying a system's characteristics related to its design options, for instance. In order for an IS to be considered evolutionary, its quality must be continuously verified. In VLEs, this means that its functionalities must be changed (for instance, to realize that students would like to publish their work to be viewed by any of their tutors other than themselves). In course materials, the changes can require a definition of a new learning object from an existing one.

To develop an adaptive IS, the UI designer must consider guidelines that must conform to usability requirements defined previously and users' characteristics and their context of use. It is usual to gather users sharing the same value for a given set of characteristics into stereotype. The problem is there is no predefined information on the users to ensure that an IS has high quality of interaction to a stereotype of all users (Furtado et al., 2001). In case of difficulties, it is better to design the best interaction possible for the most representative users, who are the intermediates (Cooper et al., 2003). They establish the functions that they use with regularity and those that they only use rarely.

Ontology can be used to represent a variety of parameters that are not necessarily identified nor truly considered in the requirements analysis. The notion of ontology allows the definition of the meta-models, which define the specification language with which any model can be specified. This resource makes it easier to consider new information in models (Vanderdonckt et al., 2004). In interactive adaptive learning systems, a flexible user modeling approach is very important. The ontology of a user model can be updated accordingly to consider more information.

So the success of an IS will depend on a complex trade-off between the classic view of requirements being satisfied by a design, but evolving, and the desired degree of satisfaction in acquiring the desired product (Sutcliffe, 2002).

These practices and the HCI concepts described here are more detailed in Sousa, Mendonça, & Furtado, 2005). It provides the developers and users a framework, including the activities that must be performed using these HCI concepts and artifacts.

FUTURE TRENDS

The multidisciplinary dimension of this work is characterized by the studies done in diverse areas of knowledge,

Table 1. A summary of requirements engineering practices to have usability

Usability of VLEs	Usability of course instructional materials
Participatory design of users	Participatory design of students and teachers
User-centered design	Student-centered design
Focuses on the UI and pedagogical functions of the VLE	Focuses on the UI and pedagogical content of the material
Requirements are represented in scenarios and prototypes	Requirements are represented in storyboards and prototypes
Convergence of various devices and resources of interaction	Authoring tool assuring the content sharing and creation using various media (photo, video, and text).

from human-computer interaction and software engineering to pedagogy. Distance learning adheres to a vision where interactive learning applications are developed for the widest population of users in the most varied contexts of use by taking into account individual differences. This population of users is being called "interactive community." To create an interactive community-centered design will be the best practice of the requirements engineering for developing the applications, which will be able to better support the collaborative learning. In addition, it will be possible to develop applications with adaptive user interfaces. These interfaces must be able to adapt themselves to the community and contexts of use characteristics.

CONCLUSION

From our experience with the community of students and teachers in the interactive learning systems and with the development of interactive systems, it was possible to establish the following conclusion about how to obtain more usable ISs: the main question for the success of learning processes for users through the different technologies does not lie exclusively on the choice of pedagogical methodology and techniques. It lies, fundamentally, on three factors: (1) the understanding of the needs of both teachers and students though a participatory design, (2) the transformation of such needs in a consistent UI and adaptive functions of a VLE and its available course materials; (3) the need to continuously assure usability of a system through an extensible representation of requirements, and (4) the technological convergence.

ACKNOWLEDGMENT

I thank the CAPES for the financial support given to this research.

REFERENCES

Appleton, B. (2000) *Patterns and software: Essential concepts and terminology*. Retrieved from http://www.cmcrossroads.com/bradapp/docs/patterns-intro.html

Chorianopoulos, K., & Spinellis D. (2006). User interface evaluation of interactive TV: A media studies perspective. *Univ. Access Inf Soc. 5*, 209-218.

Constantine, L., Windls, H., Nolbe, J., & Lockwood, L. (2003). *From abstraction to realization in user interface designs: Abstract prototype based on canonical abstract components*. Working Paper on Tutorial Usage-Centered Software Engineering. ICSE'03, Portland, Oregon.

Cooper A., Reimann, R. (2003). *About Face 2.0. The essentials of interaction design*. John Wiley & Sons.

Furtado, E., Furtado, V., Bezerra, W., William, D., Taddeo, L., Limbourg, Q., & Vanderdonckt, J. (2001). An ontology-based method for universal design of user interfaces. *Proceedings of the Workshop on Multiple User Interfaces over the Internet: Engineering and Applications Trends*, Lille. France. Retrieved July, 10, 2001, from www.cs.concordia.ca/%7Efaculty/seffah/ihm2001/program.html

Furtado, E., Mattos, F. L., Furtado, J. J. V., & Vanderdonckt, J. (2003). Improving usability of an interactive learning system. In C. Ghaoui (Ed.), *Usability evaluation of interactive learning programs* (pp. 69-86).

Guarino, N. (1995). Formal ontology, conceptual analysis, and knowledge representation: The role of formal ontology in information technology. *International Journal of Human-Computer Studies, 43*(5-6), 623-640.

Garrett J. (2003). The elements of user experience. Ed. AIGA.

Karoulis, A., & Pombortsis, A. (2003). Heuristic evaluation of Web-based ODL programs. In C. Ghaoui (Ed.), *Usability evaluation of interactive learning programs* (pp. 89-109).

Maia M., & Furtado, E. (2006). System to support publishing, editing, and creating Web content for various devices. In *Proceedings of the 6th International Conference on Computer-Aided Design of User Interfaces—CADUI'2006*. Bucharest, Romania, June 5-8, 2006.

McCarthy, J., & Wright, P. (2004). *Technology as experience.* MIT Press.

Mattos, F. L., (2005). Concepção e desenvolvimento de uma abordagem pedagógica para processos colaborativos a distancia utilizando a internet. PhD. UFC.

Mattos, F. L., Maia, M., &, Furtado, E. S. (2003). Formação docente em processos colaborativos interactive: Em direção a novos "círculos de cultura"? In *Proceedings of the Workshop em Informática na Educação (WIE)*.

Neto, H., Raimir H., Bezerra W., & Sarquis O. (2000). Especificando o tele-ambiente no contexto da educação a distância. In *Proceedings of the Simpósio Brasileiro de Informática Educativa (SBIE'2000)* (pp. 120-132). Alagoas. Universidade Federal de Alagoas Editora.

Perrenoud, P. (2001). Formando professores profissionais: Quais estratégias? Quais competências? Porto Alegre: Artmed.

Roibás, A. C., Geerts, D., Furtado, E., & Calvi, L. (2006). Investigating new user experience challenges in iTV: Mobility & sociability. In *Proceedings of Conference in Human*

Factors in Computing Systems CHI'2006)pp. 1659-1662). (Montreal, Ca, 22-27 April 2006). ACM 1-59593-298-4/06/0004. 2006.

Sousa, K., Mendonça, H., & Furtado, E. (2005) UPi--A software development process aiming at usability, productivity and integration. *CLIHC '2005- II Congresso Latino Americano de IHC*, Cuernavaca. México.

Sutcliffe, A. (2002). *User-centered requirements engineering. Theory and practice*. London: Springer-Verlag.

Vanderdonckt, J., Furtado, E., Furtado, V., Limbourg, Q., Bezerra, W., William, D., & Taddeo, L. (2004). *Multi-model and multi-layer development of user interfaces in multiple user interfaces*. London: Ahmed Seffah and Homa Javahery, John Wiley & Sons.

KEY TERMS

Evolutionary System: Involves continuous adjustments of its functionalities and UI according to the user and/or technological changes.

Extensible Representations of Requirements: Ways to represent easy requirements that were not necessarily identified nor truly considered in the requirements analysis.

HCI Pattern: A representation (graphical, textual and so on) about a tested solution for a usability problem (such as lack of orientation, difficulty in finding information) that happens in a certain context (search, visualization, etc.).

Pedagogic Usability of an Interactive Learning System: Related to how easy and effective it is for a student to learn something using the system.

Requirements Engineering: The human acts of identifying and understanding what people want from an IS.

Usability of an IS: Refers to how easy it is to use and learn the system.

Usability Requirements: Related to users' satisfaction and the performance of the system.

Improving Virtual Teams through Creativity

Teresa Torres-Coronas
Universitat Rovira i Virgili, Spain

Mila Gascó-Hernández
Open University of Catalonia, Spain

INTRODUCTION

Many studies have already shown how a team can become more creative, and therefore more efficient, but only a few researchers have focused on how a virtual team can use creativity techniques to perform better. In this article, we study what differences there are (both in terms of processes and in terms of results) when creativity techniques are used in the management of traditional and virtual teams. To do this, we discuss three main elements: the definition of creativity and its relationships with team performance, the variables that enhance creativity in a virtual team, and the most suitable creativity techniques for a virtual environment.

BACKGROUND

Most researchers and practitioners believe that the key to organizational success lies in developing intellectual capital and acquiring a new set of thinking: the creativity to produce an idea and the innovation to translate the idea into a novel result (Roffe, 1999). Explaining the meaning of creativity is not straightforward; there are thousands of definitions of the term. So, for the purpose of this article, we will understand creativity as the shortest way to search for unconventional wisdom and to produce paradigm-breaking ideas and innovation. This unconventional wisdom through the generation and use of creative knowledge is the key to building sustainable competitive advantages (Carr, 1994).

In order to develop more innovative products, services, or processes, organizations must encourage their employees to become more creative. During the last few decades, several researchers (Andriopoulous, 2001; Nemiro & Runco, 2001; McFadzean, 1998; Amabile, Conti, Coon, Lazenby & Herron, 1996) tried to describe contextual factors largely under the control of managers that influence creativity, though as creativity is a multidimensional concept, there is not a universal theory yet (Walton, 2003). This section focus on how managers and/or team leaders can improve creative climate within virtual structures.

The literature review conducted by Andriopoulous (2001) highlights five major organizational factors that enhance creativity in a traditional work environment: 1) organizational climate, or designing a working atmosphere that fosters participation and freedom of expression; 2) a democratic and participative leadership style; 3) an organizational culture that nourishes innovative ways of solving problems; 4) new resources and skills through the development of human resources creative talent; and 5) a structure and systems that include building flat structures, and rewards, recognition, and career systems that emphasize people creative thinking. Scholars argue that these factors create conditions that enhance creativity both at the team and individual levels.

From a study of the social psychology of creativity, Amabile (1996) cites the three main origins of creative performance as: task motivation, domain-relevant skills, and creativity-relevant skills. She differentiates between intrinsic and extrinsic motivation, proposing that the intrinsic motivation enhances creativity. In Amabile's research, the work team environment is also considered to exert a powerful impact on creativity by influencing the employee's intrinsic motivation. Management practices indicate that performance can be fostered by allowing freedom and autonomy to conduct one's work, matching individuals to work assignments, and building effective work teams that represent a diversity of skills and are made up of individuals who trust and communicate well with each other, challenge each other's ideas, are mutually supportive, and are committed to the work they are doing (Amabile & Gryskiewicz, 1987). Creativity is best achieved in open climates (Feurer, Chaharbaghi & Wargin, 1996).

These studies have not specifically addressed dimensions that may be necessary when groups no longer interact in traditional structures (Nemiro, 2001). In fact, so far, the only research that has been seriously conducted about this issue is that by Nemiro (2001), who identifies several key elements that influence creativity in virtual teams and therefore result in effectiveness and high levels of performance. Table 1 summarizes some of these factors as described by Nemiro (2001, p. 94).

A creativity-based management aimed at fostering virtual team creativity and performance must manage the above environmental variables in order to enhance employees' internal drive to perceive every project as a new creative challenge (Andriopoulos & Lowe, 2000).

A quick analysis of the variables shown in Table 1 gives rise to the conclusion that there are no meaningful differences

between the factors that affect creativity in traditional environments and those that affect creativity in virtual contexts. On the other hand, most of the factors that influence creativity (such as work characteristics and situational constraints) are also considered as factors that impact team performance, as the conceptual model of Prasad and Akhilesh (2002) shows. Nevertheless, due to the particular way virtual teams work, there is a need to consider some elements related to the previous variables. Thus, communication and trust become very relevant issues.

In this sense, Henry and Hartzler (1998) find that keeping the synergy and creativity flowing, without frequent face-to-face interaction, is the greatest challenge a virtual team has. Virtual teams lose non-verbal communication and, as has been argued, electronic communication increases the level of social isolation. Schein (1993) points out that most communication workshops emphasize active listening, which means paying attention to the spoken words, the body language, the tone of voice, or the emotional content. Virtual teams that want to communicate successfully cannot actively listen in this sense. Other tools must therefore be explored¾for example, the use of multiple media or several communication technologies (Bal & Teo, 2001). However, as Van der Smagt (2000) showed, it is crucial to ensure that dialogue is the primary form of interaction between team members and that two-way monologues are avoided. Rich media—those that transmit nonverbal cues—are not the solution.

"In a dialogue, the difficult part is to make one's own assumptions manifest, not the exchange of insights with others. The attitude in relation to other actors is one of openness, which makes it relatively easy to get behind the position and possibilities of actors." (Van der Smagt, 2000, p. 155)

Collaborative work also requires a level of personal familiarity and trust. Without trust, building a true team is almost impossible (Duarte & Snyder, 1999). For most newly forming virtual teams, achieving an effective level of trust is not an easy task. Increasingly, virtual teams will form without the advantage of prior face-to-face team building opportunities, but with the added challenges of geographic isolation, time zone differentials, and cultural diversity (Holton, 2001). With virtual team heterogeneity there is a high probability that team members are confronted with mistrust (Prasad & Akhilesh, 2002), though such diversity within a team has the potential to increase opportunities to be innovative and creative (Lipnack & Stamps, 1997), if trust can be established (Dyer, 1995). But how can trust be built? The qualitative research project of Holton (2001) concludes that standard team-building tools can be used to enhance collaboration and trust in a virtual team. The book of Simon Priest (2001) is full of examples for virtual team building. But, as with all team building, there is no quick fix for virtual teams.

These difficulties related to communication and trust are only an example that illustrates the need to conduct in-

Table 1. A summary of factors that can foster creativity in a team context

Autonomy and Freedom. Allowing individuals responsibility for initiating new ideas and making decisions; a sense of control over one's work.
Challenge. Work that is stimulating, engaging, and meaningful; a sense of having to work hard on challenging and important tasks.
Clear Direction. Goals that facilitate creativity are clear, negotiated, attainable, shared, and valued.
Diversity/Flexibility/Tension. Diversity, both in terms of the work assignments offered and the people one interacts with, and a tolerance of differences. In order to be tolerant of differences, flexibility is needed. Both diversity and flexibility can lead to creative tension.
Support for Creativity. An organizational focus on support for or encouragement of creativity.
Trust and Participative Safety. Especially crucial for group creativity is trust and participative safety. The emphasis is on encouraging participation in a non-threatening, non-evaluative environment.

depth studies on the rest of environmental factors that, in non-virtual contexts, have been proven to directly impact teamwork creativity.

TOOLS AND TECHNIQUES TO IMPROVE CREATIVE PERFORMANCE

How can team creativity be encouraged? Until now no serious research has been conducted into which creativity techniques are the most suitable in a virtual environment. In traditional environments, one method of achieving this is to encourage teams to utilize creative problem-solving (CPS) techniques such as synectics, brainwriting, or wishful thinking.

In this context, McFadzean (2000, 1999, 1998) explores creative problem solving and presents a model that helps facilitators and team members choose appropriate techniques. McFadzean (1996a, 1996b) classifies creative problem solving (CPS) techniques into three categories: paradigm preserving, paradigm stretching, and paradigm breaking. Paradigm preserving techniques do not tend to change a participant's perspective. Paradigm stretching techniques encourage users to stretch the boundaries of the problem space. Paradigm breaking techniques allow participants to completely break down the boundaries of the problem space and to look at something entirely new.

In a virtual environment, three variables must also be considered when selecting a technique (Gascó-Hernández & Torres-Coronas, 2004): 1) the effectiveness of the method in finding innovative solutions, considering that quality solutions require the right balance between knowledge of the business issue and novelty (Kim, 1990); 2) the technological context or support system through which the technique can be implemented; and 3) the level of interaction that the technique requires. It is also important that the virtual facilitator has experience in running virtual creative sessions. Team members must also be taught about the dynamics of virtual interactions and about the use of technological tools, such as chats, e-mail, video conferencing, or interactive whiteboards.

Next, McFadzean's creativity continuum (1996a, 2000) will be used to summarize how virtual teams can choose among different techniques, which will be briefly described, to generate ideas. To make valuation accessible, the techniques will be classified according to the above three criteria. They will be rated as well (low, medium, and high). Finally, common technological tools by which each technique can be used will be shown.

It is a fact that there is a remaining need for in-depth research studies that help clarify which is the best creative problem-solving tool in terms of virtual team creative performance. Table 2 intends to be the basis to start evaluating the creativity continuum in a virtual context.

The use of these techniques will only be effective if the organization has a creative culture (McFadzean, 1998). Environmental factors that help managers to build a creative climate within virtual communities and creative problem-solving tools are two sides of the same coin.

FUTURE TRENDS

Virtual teams can use creativity in order to perform better. Nevertheless, there is a need to adapt those tools and techniques to a virtual environment. In this article, we have approached some of the issues that must be considered when studying the relationship between teams that work online and creativity. Nevertheless, several questions still remain unanswered and, in this sense, further research is required. In particular, three important issues need further development. First, the relationship between creativity and virtual team performance needs to be thoroughly explored. Successful studies will determine how structural and environmental factors influence team creativity. Second, although virtual teams are already using idea-generation techniques, their strengths and weaknesses need to be carefully and academically explored. Finally, it is also important to consider the effects of technology on both individual and team creativity. Technology has risks that can sometimes outweigh its benefits. When applying creativity techniques, people need to focus on the creative process, not on the technology being used. Technology must be easy to use, it must be effortless and unsophisticated—the simpler the technology, the better.

CONCLUSION

The emergence of tools based on the new information and communication technologies is currently affecting team creative processes. Nowadays, achieving high levels of creative performance is still an unresolved problem within virtual teams. Only a few researchers have focused on how a virtual team can use creativity techniques to perform better or how to build a creative virtual environment to foster creativity. The critical issues discussed in this article summarize many challenges to implement creative management within both virtual teams and organizations. With greater emphasis being placed on creative thinking and processes, team creative performance will increase day by day, allowing organizations to succeed and to become more innovative and adaptable.

Table 2. Valuing the creativity continuum in a virtual context

PRESERVING PARADIGM TECHNIQUES	**STRETCHING PARADIGM TECHNIQUES**	**BREAKING PARADIGM TECHNIQUES**
• Problem boundaries: unchanged • Creative stimulation: low • Stimuli: related • Expression: verbal/written • Can be used by experienced and inexperienced groups	• Problem boundaries: stretched • Creative stimulation: medium • Stimuli: unrelated • Expression: verbal/written	• Problem boundaries: broken • Creative stimulation: high • Stimuli: unrelated • Expression: unlimited • Should only be used by experienced groups
TECHNIQUES WITHIN EACH GROUP		
BRAINSTORMING	**OBJECT STIMULATION**	**WISHFUL THINKING**
Generation of ideas without criticism	Group members generate ideas using objects unrelated to the problem.	Participants are asked to look at a perfect future, examine each fantasy statement, and look for ideas on how these ideas can be achieved.
Effectiveness: Medium	Effectiveness: Medium	Effectiveness: Medium
Technological context: Online chat rooms Electronic mail Video conferencing	Technological context: Online chat rooms Electronic mail Video conferencing	Technological context: Online chat rooms Electronic mail Video conferencing
Level of interaction: High	Level of interaction: Medium	Level of interaction: Medium
BRAINWRITING	**METAPHORS**	**RICH PICTURES**
Group members write down their ideas on different sheets of paper. They are encouraged to build on others' ideas.	Group members use a metaphor to generate ideas to solve a problem.	Participants draw pictures which can be a metaphor of the problem. The descriptions of the picture help participants to generate ideas.
Effectiveness: High	Effectiveness: Medium	Effectiveness: Medium
Technological context: Online chat rooms Electronic mail Video conferencing	Technological context: Online chat rooms Electronic mail Video conferencing	Technological context: Whiteboard software
Level of interaction: Medium	Level of interaction: Medium	Level of interaction: Low

REFERENCES

Amabile, T.M., Conti, R., Coon, H., Lazenby, J. & Herron, M. (1996). Assessing the work environment for creativity. *Academy of Management Journal, 39*(5), 1154-1184.

Andriopoulos, C. (2001). Determinants of organizational creativity: A literature review. *Management Decision, 39*(10), 834-840.

Andriopoulos, C. & Lowe, A. (2000). Enhancing organizational creativity: The process of perpetual challenging. *Management Decision, 38*(10), 834-840.

Bal, J. & Teo, P.K. (2001). Implementing virtual teamworking: Part 2—a literature review. *Logistics Information Management, 14*(3), 208-222.

Carr, C. (1994). *The competitive power of constant creativity.* New York: AMACOM.

Duarte, D.L. & Snyder, N.T. (1999). *Mastering virtual teams.* San Francisco, CA: Jossey-Bass.

Dyer, W.G. (1995). *Team building: Current issues and new alternatives* (3rd Edition). Reading, MA: Addison-Wesley.

Feurer, R., Chaharbaghi, K. & Wargin, J. (1996). Developing creative teams for operational excellence. *International Journal of Operations & Production Management, 16*(1), 5-18.

Gasco-Hernández, M. & Torres-Coronas, T. (2004). Virtual teams and their search for creativity. In F. Pixis & S. Godar (Eds.), *Virtual and collaborative teams: Process, technologies, and practices* (pp. 213-231). Hershey, PA: Idea Group Publishing.

Henry, J.E. & Hartzler, M. (1998). *Tools for virtual teams.* Milwaukee, WI: ASQC Quality Press.

Holton, J.A. (2001). Building trust and collaboration in a virtual team. *Team Performance Management: An International Journal, 7*(3/4), 36-47.

Kim, S.H. (1990). *Essence of creativity—a guide to tackling difficult problems.* New York: Oxford University Press.

Lipnack, J. &, Stamps, J. (1997). *Virtual teams: Reaching across space, time and organizations with technology.* New York: John Wiley & Sons.

McFadzean, E.S. (1998). Enhancing creative thinking within organizations. *Management Decision, 36*(5), 309-315.

McFadzean, E.S. (1996a). *The classification of creative problem-solving techniques.* Working Paper No. 9632, Henley Management College, Henley-on-Thames, Oxon, UK.

McFadzean, E.S. (1996b). *New ways of thinking: An evaluation of K-Groupware and creative problem solving.* Doctoral Dissertation, Henley Management College/Brunel University, Henley-on-Thames, Oxon, UK.

McFadzean, E.S. (1999). Encouraging creative thinking. *Leadership & Organization Development Journal, 20*(7), 374-383.

McFadzean, E.S. (2000). Techniques to enhance creative thinking. *Team Performance Management: An International Journal, 6*(3/4), 62-72.

Nemiro, J.E. (2001). Connection in creative virtual teams. *The Journal of Behavioral and Applied Management, 2*(2), 92-112.

Prasad, K. & Akhilesh, G.B. (2002). Global virtual teams: What impacts their design and performance? *Team Performance Management: An International Journal, 8*(5/6), 102-112.

Priest, S. (2001). *100 of the best virtual team-building events.* Tarrack Publication.

Roffe, I. (1999). Innovation and creativity in organizations: A review of the implications for training and development. *Journal of European Industrial Training, 23*(4/5), 224-237.

Schein, E.H. (1993). On dialogue, culture and organizational learning. *Organizational Dynamics, 2*2(2), 40-51.

Van der Smagt, T. (2000). Enhancing virtual teams: Social relations v. communication technology. *Industrial Management & Data Systems, 100*(4), 148-156.

Walton, A.P. (2003). The impact of interpersonal factors on creativity. *International Journal of Entrepreneurial Behaviour & Research, 9*(4), 146-162.

Williams, S. (2001). Increasing employees' creativity by training their managers. *Industrial and Commercial Training, 33*(2), 63-68.

KEY TERMS

Autonomy and Freedom: Allowing individuals responsibility for initiating new ideas and making decisions; a sense of control over one's work.

Challenge: Work that is stimulating, engaging, and meaningful; a sense of having to work hard on challenging and important tasks.

Clear Direction: Goals that facilitate creativity are clear, negotiated, attainable, shared, and valued.

Converging Thinking Techniques: Tools used during the convergent phases of the CPS to improve the evaluation and selection of the most relevant ideas, thoughts, or data. Pluses, potentials, and concerns (PPC); highlighting; and the evaluation matrix are some of the most common converging thinking techniques.

Creative Performance: High level of capability in an idea or solution, applied to solve a problem in an imaginative way, resulting in effective action. Environmental factors such as autonomy and freedom, challenge, clear direction, diversity/flexibility/tension, support for creativity, trust, and

participative safety directly affect the creative performance within work teams.

Creative Problem Solving (CPS): A systematic process model to solve problems and to harness creativity. Its six steps include objective-finding, data-finding, problem-finding, idea-finding, solution-finding, and acceptance-finding. Each step has a divergent and convergent phase. During the divergent phase, a free flow of ideas is elicited. Convergent phases involve the evaluation and selection of the ideas with the greatest potential or relevancy. The defer-judgment rule separates idea generation from idea evaluation.

Creativity: The production of something new or original that is useful; the act of creating recombining ideas or seeing new relationships among them. Creativity is usually defined in terms of either a process or a product and at times has also been defined in terms of a kind of personality or environmental press. These are four Ps of creativity: process, product, person, and press.

Divergent Thinking Techniques: Tools used during the divergent phases of the CPS to improve the generation of ideas, thoughts, or data without evaluation. These tools are classified according to their primary use of related or unrelated problem stimuli. Brainstorming, brainwriting, forced connections, analogies, and metaphors are some of the most used divergent thinking techniques.

Diversity/Flexibility/Tension: Diversity, both in terms of the work assignments offered and the people one interacts with, and a tolerance of differences. In order to be tolerant of differences, flexibility is needed. Both diversity and flexibility can lead to creative tension.

Support for Creativity: An organizational focus on support for or encouragement of creativity.

Trust and Participative Safety: Especially crucial for group creativity is trust and participative safety. The emphasis is on encouraging participation in a non-threatening, non-evaluative environment.

Virtual Team: A group of people who are geographically separated and who work across boundaries of space and time by utilizing computer-driven technologies such as desktop video conferencing, collaborative software, and Internet/intranet systems. How these teams interact defines them as "virtual."

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1419-1424, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

An Inclusive IS&T Work Climate

Debra A. Major
Old Dominion University, USA

Valerie L. Morganson
Old Dominion University, USA

INTRODUCTION

Employees develop perceptions regarding which behaviors are expected, supported, and rewarded in their organization through a series of workplace events, practices, and procedures; these beliefs comprise a workplace climate (Schneider, Wheeler, & Cox, 1992). An inclusive workplace climate is one in which everyone has a sense of belonging, is invited to participate in decisions, and feels that their input matters (Hayes, Bartle, & Major, 2002; Major, Davis, Sanchez-Hucles, Germano, & Mann, 2006). For an inclusive climate to exist, all organizational members should feel equally welcome in the IT work environment and feel free to make suggestions regardless of their gender or ethnicity. Moreover, all organizational members should feel that their contributions have an impact (Major et al., 2006). Rather than simply tolerating diversity, organizations with an inclusive climate embrace it and capitalize upon it. In an IT sample, inclusive climate was positively associated with job satisfaction, organizational and career commitment, and intentions to remain with one's employer (Major et al., 2003). In contrast, exclusion is associated with turnover, reduced organizational commitment and decreased job satisfaction (Greenhaus, Parasuraman, & Wormley, 1990). Research has highlighted the role of three contributors to inclusive climate: (1) strong supervisor/subordinate relationships, (2) supportive coworkers, and (3) a supportive culture (Margolis & Fisher, 2003; Major, Davis, Sanchez-Hucles, & Mann, 2003). The current article briefly reviews social factors that have hindered the realization of a gender and minority inclusive IT climate and draws upon these three contributors to identify strategic levers to guide managers and researchers toward fostering inclusion in the IT workforce.

BACKGROUND

Women's underrepresentation in IT education and careers has been recognized as a global problem (Huyer, 2005; Rosser, 2005). In addition, certain U.S. ethnic minority groups, including African Americans, Hispanic Americans, and Native Americans, are underrepresented in programs preparing people for IT careers (Tapia, Kvasny, & Trauth, 2004). Researchers refer metaphorically to a "leaky pipeline" to describe the attrition of women and minorities from pathways leading to participation and success in IT education and careers. From childhood to adulthood, a variety of experiences, such as limited access and exposure to computers, and the unappealing portrayal of the IT industry (e.g., depictions of individuals working in solitude and the stereotypical IT "geek image"), discourage female and minority interest in IT (see Rosenblum, Ash, Coder, & Dupont, 2006, Splender, 1997; Tapia et al., 2004).

Barriers persist in the IT workplace. Women encounter a "glass ceiling," a situation in which men hold top-level positions and women are limited in their ability to move up due to barriers that are not readily obvious or overt (Martin, 2005). In fact, U.S. women hold fewer than 5% of IT executive positions, such as chief information officer (Gingras, 1999). Instead, they are more likely to hold support positions such as help desk operator or support center staff (Belt, 2002). The nature of IT work can be a barrier, requiring long and irregular hours and, in some positions, the flexibility to travel. These job demands infringe upon family life and women are frequently less willing or able to sacrifice their home and childcare duties (Panteli, Stack, & Ramsay, 1999; Roldan, Soe, & Yakura, 2004).

Populated primarily by white males, the IT work climate can be "chilly" because of characteristics that make it unfriendly and inhospitable to women and minorities (Roldan et al., 2004). An inclusive climate is

warm and inviting to all regardless of their demographic characteristics. This review considers key factors that research has shown to contribute to an inclusive climate and to the success of a diverse workforce. We begin with a discussion of self-fulfilling prophecy as a mechanism through which workforce inclusion disparities function. Mentoring, leadership and coworker relationships are each discussed as levers for social change.

SELF-FULFILLING PROPHECY

Based on experience and perceptions, people develop expectations about what behaviors to anticipate from other individuals. Expectations, in turn, can realize themselves through the actions of the perceiver and subsequent reactions of the perceptual target. Social scientists refer to the tendency of others' assumptions to evoke behavior as *self-fulfilling prophecy* and have highlighted its influence in the professional development and success of women and minorities. Expectations are particularly influential when they are held by powerful individuals. Through the process of self-fulfilling prophecy, managers form expectations of employees and behave according to their expectations by either providing or withholding emotional and social support. In turn, subordinates exposed to high expectations and greater support gain experience, confidence, and ultimately perform better, whereas those confronted with low expectations and a lack of support are less likely to be high performers, given a relative lack of opportunity to bolster confidence and experience (Eden, 1997).

Self-fulfilling prophecy can be either negative or positive and can operate on an individual level or pertain to entire social groups. In IT, women often feel that they need to outperform men to be viewed equally. When they act upon this expectation, the quality of their work may be adversely affected, and they appear less competent as a result (Valian, 1998). Along the same lines, members of underrepresented groups feel extra pressure because they fear confirming and reinforcing negative stereotypes about their group. *Stereotype threat* is the circumstance when this extra pressure reduces performance and persistence (Steele, 1997).

Organizational development experts have recognized the power of positive expectations as an instrument to change work climates for the better. They advocate *appreciative inquiry* or emphasis upon optimism, positive expectations, and challenge (Srivasta & Cooperrider, 1990). By focusing on what the organization does well, rather than what it does poorly, and acknowledging individual contribution, organizations become better aligned and equipped to achieve inclusive climate. Holding positive expectations can be especially transforming in situations where low expectations are usually held (McNatt, 2000) and when gender and racial stereotypes tend to influence expectations (Jussim & Eccles, 1992).

MENTORING

Mentoring is a process by which a more experienced employee (e.g., a supervisor) guides, advises, counsels and otherwise enhances the professional development of another employee. Mentors act as role models, provide social support, and serve as sources of information, especially early in an employee's organizational tenure. Mentors can be instrumental in the new employee's career development (Scandura, 1992). For example, mentoring is associated with higher performance ratings, more frequent promotions, greater job satisfaction and higher income (e.g. Allen, Eby, Poteet, Lentz, & Lima, 2004; Ragins & Cotton, 1999).

Developing an effective mentor-protégé relationship can be challenging for minority and female employees. Although women and minorities may prefer a mentor who is demographically similar to themselves (Murrell, Crosby, & Ely, 1999), they are in short supply given the composition of the IT workforce. Compared to white males, incumbent women and minorities in the IT workforce are more likely to be in the junior stages of their careers. At early career stages, professionals have more limited social networks, and less access to information, and therefore may have limited effectiveness as mentors (Benishek, Bieschke, Park, & Slattery, 2004). White males may be better mentoring resources to the extent that they have greater social status and power in the organization (Dreher & Cox, 1996; Ragins & Cotton, 1999).

However, mixed-gender mentoring relationships also have obstacles. In the case of a male mentor and a female protégé, concern that others will make sexual attributions about the nature of their relationship can make both parties reluctant to pursue a mentoring relationship (Ragins & Cotton, 1999). When mixed-gender mentoring relationships do develop, they may offer

less support than same-sex mentoring relationships. Female protégés are sometimes viewed by the mentor as not taking their professional goals seriously, and may be mentored in a qualitatively different way than male protégés (Schwiebert, Deck, Bradshaw, Scott, & Harper, 1999). Moreover, in mixed-gender mentorships, the female protégé is less likely to be included in informal networking activities than her male counterpart (Ragins & McFarlin, 1990).

To equitably disperse the many benefits of mentoring relationships, organizations should be active in fostering collaboration among junior and senior employees. Organizations profit from formal mentorship programs, although informal mentorships are associated with even greater benefit, especially for women (Ragins & Cotton, 1999). As an effective strategy, organizations may identify a pool of potential mentors and protégés, provide training, and allow dyads to form according to individual preference (Forret, Turban, & Dourghty, 1996).

LEADERSHIP

IS&T work requires effective leaders to maintain a balance of technical/task-oriented and interpersonal skills (Major et al., 2007). Yet, the technical aspects of the IS&T profession have historically been emphasized while interpersonal requisites have been downplayed. This emphasis may disadvantage females whose strengths are often relationships, understanding, communication and other interpersonal aspects of IT (Nielsen, von Hellens, Greenhill, & Pringle, 1997). This is not to infer that females are weak at technical aspects of IT work. Indeed, there is no evidence to support such a conjecture. Instead, our intent is to recognize the contribution that research shows interpersonal skills make to IT leadership effectiveness.

Major et al. (2007) found that IT professionals appreciate supervisors with effective interpersonal and relationship building skills. Some research suggests that female IS&T supervisors are better able to develop high-quality working relationships with their subordinates than male supervisors (Major et al., 2006). Mayo (1982) offers a plausible explanation for this finding. She described a condition of positive marginality, whereby minorities and women develop strong leadership skills because they have observed effective and ineffective practices from the sidelines. Promoting women and minorities to positions of leadership may act as a catalyst for inclusive climate by synergistically changing norms to welcome diversity and by providing female and minority subordinates with role models.

In addition to recognizing the value of a relational approach to IT leadership, IT professionals should be aware of how gender stereotypes can hinder inclusive climate. Women attempting to lead, especially in male-dominated environments, may face a double-standard (Gherardi, 1995). Women who lead with an instrumental and assertive style (i.e., a stereotypically masculine style) may be rejected and labeled as overly "pushy" and are likely to be viewed as behaving inconsistently with female gender role expectations. IT professionals should be mindful of the influence of their own gender socialization and expectations in order to prevent the manifestation of such double standards.

COWORKER RELATIONSHIPS

The expression "the people make the place" is sometimes used to emphasize human capital as an organizational asset (see Schneider, 1987). This holds true for creating inclusive climate. Through effective coworker relationships, individuals are better able to adjust and are more likely to feel anchored in their environment. Indeed, coworker support is associated with a variety of positive work outcomes, including reduced stress, greater organizational commitment, higher job satisfaction and reduced intentions to turnover (Baruch-Feldman, Brondolo, Ben-Dayan, & Schwartz, 2002; Ducharme & Martin, 2000; Lee, 2004). Among IT professionals, men and women reported equal satisfaction with the IT social environment and both felt equal emotional support from their coworkers (Major et al., 2006). In addition to affective ties, social networks contribute to well-being and career progress (Podolny & Baron, 1997; Seibert, Kraimer, & Liden, 2001) and provide women in IT with a means of coping (Pegher, Quesenberry, & Trauth, 2006). However, compared to men, women report being less able to employ social networks to make career moves (Sumner & Werner, 2001). This finding illustrates a distinction between participation and influence. To truly realize an inclusive climate, individuals should be equally able to realize results and benefits from their participation (Major et al., 2006).

FUTURE TRENDS

Research has identified several levers to increase inclusive climate. To warm the "chilly" climate, policies and norms should be set in place to welcome diverse lifestyles. For example, women value working for organizations that promote child care, maternity leave, equal opportunity, sexual harassment laws, and training (Trauth, 2002). In addition, efforts should be made to promote minorities and women and to develop their careers so that they are present to mentor subordinates. Inclusive relationships enhance mentoring and advising, and increase retention of women and minority IT professionals (Gürer & Camp, 2002). Finally, it is essential that IT professionals experience a sense of influence rather than simply being allowed to participate. In one study of IT professionals, women, particularly minority women, reported feeling included in a relational sense only; they felt their contributions had less of an impact than those of their male counterparts (Major et al., 2006). It is important to note that organizational policies and practices aimed at increasing the inclusion of women and minorities do not exclude majority group members (e.g., white males). Instead, the levers we have discussed here promote the inclusion of all.

Minority status can be especially stressful due to reduced access to mentors and to information (Jackson, Stone, & Alvarez, 1992). Just as organizational polices and practices can promote inclusion, there are also steps that individuals can take to improve inclusion. Active coping is defined as taking steps to overcome these stressors and mitigate their effects. Individuals practicing active coping (1) *plan* or think about how to cope with a stressor, (2) *seek social support* for instrumental and emotional reasons, (3) *suppress competing activities* to allow focus on the stressor at hand, and (4) *restrain action* until the appropriate time (Carver, Scheier, & Weintraub, 1989). Through active coping individuals can capitalize on a constellation of positive outcomes, including increased optimism, control, self-esteem, and resilience and decreased anxiety. Teaching active coping has great potential as an intervention to increase inclusive climate in the field of IT.

CONCLUSION

Inclusive climate is essential for recruiting and retaining a diverse workforce in IT organizations. Inclusion offers the benefit of diverse viewpoints and promotes race and gender equity. Effective IT management practices include embracing participation, providing support for individual lifestyle accommodations, and taking an individual interest in each employee (Major et al., 2007). Inclusive practices, such as sourcing from hiring agencies that specialize in providing female and minority applicants and promoting family-friendly polices and procedures (e.g., relaxed environment, flexible work arrangements, child care services and telecommuting) are associated with success in the IT industry (Agarwal & Ferratt, 2002). Individuals and organizations should be proactive in creating an inclusive environment because failing to actively encourage women and minorities may have the same effect as actively discouraging them (Leggon, 2003).

REFERENCES

Agarwal, R., & Ferratt, T. W. (2002). Enduring practices for managing IT professionals. *Communications of the ACM, 45*(9), 73-79.

Allen T. D., Eby, L. T., Poteet, M. L., Lentz, E., & Lima, L. (2004). Career benefits associated with mentoring for protégés: A meta-analysis. *Journal of Applied Psychology, 89*, 127-136.

Baruch-Feldman, C., Brondolo, E., Ben-Dayan, D., & Schwartz, J. (2002). Sources of social support and burnout, job satisfaction, and productivity. *Journal of Occupational Health Psychology, 7*, 84-93.

Belt, V.A. (2002). A female ghetto? Women's careers in call centres. *Human Resource Management Journal, 12*(4), 51-67.

Benishek, L.A., Bieschke, K.J., Park, J., & Slattery, S.M. (2004). A multicultural feminist model of mentoring. *Journal of Multicultural Counseling and Development, 32*, 428-442.

Carver, C.S., Scheier, M.F., & Weintraub, J.K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology, 56*, 267-283.

Dreher, G. F., & Cox, Jr., T. H. (1996). Race, gender, and opportunity: A study of compensation attainment and the establishment of mentoring relationships. *Journal of Applied Psychology, 81*, 297-308.

Ducharme, L. J., & Martin, J. K. (2000). Unrewarding work, coworker support, and job satisfaction: A test of the buffering hypothesis. *Work and Occupations, 27,* 223-243.

Eden, D. (1997). Leadership and expectations: Pygmalion effects and other self-fulfilling prophecies in organizations. In R.P. Vecchio (Ed.), *Leadership: Understanding the dynamics of power and influence in organizations* (pp. 177-193). Notre Dame, IN: University of Notre Dame Press.

Forret, M.L., Turban, D.B., & Dougherty, T.W. (1996). Issues facing organizations when implementing formal mentoring programmes. *Leadership and Organization Development Journal, 17,* 27-30.

Gherardi, S. (1995). *Gender, symbolism, and organizational cultures.* Thousand Oaks, CA: Sage.

Gingras, A. (1999, January 18). Cherchez la femme. *Computerworld*, 49.

Greenhaus, J.H., Parasuraman, S., & Wormley, W.M. (1990). Effects of race on organizational experiences, job performance evaluations and career outcomes. *Academy of Management Journal, 33,* 64-86.

Gürer, D., & Camp, T. (2002). An ACM-W literature review of women in computing. *SIGCSE Bulletin, 34,* 121-127.

Hayes, B.C., Bartle, S.A., & Major, D.A. (2002). Climate for opportunity: A conceptual model. *Human Resource Management Review, 12,* 445-468.

Huyer, S. (2005). Women, ICT and the information society: Global perspectives and initiatives. In *Proceedings from the International Symposium of Women and ICT: Creating Global Transformation,* Baltimore, MD, (pp. 1-6).

Jackson, S.E., Stone, V.K., & Alvarez, E.B. (1992). Socialization amidst diversity: The impact of demographics on work team oldtimers and newcomers. In L.L. Cummings & B.M. Staw (Eds.), *Research in organizational behavior* (Vol. 15, pp.45-109). Greenwich, CT: JAI Press.

Jussim, L., & Eccles, J.S. (1992). Teacher expectations: II. Construction and reflection of student achievement. *Journal of Personality and Social Psychology, 63,* 947-961.

Lee, P. C. B. (2004). Social support and leaving intention among computer professionals. *Information and Management, 41,* 323-334.

Leggon, C.B. (2003). Women of color in IT: Degree trends and policy implications. *Technology and Society Magazine, IEEE, 22*(3), 36-42.

Major, D.A., Davis, D.D., Germano, L.M., Fletcher, T.D., Sanchez-Hucles, J., & Mann, J. (2007). Managing human resources in information technology: Best practices of high performing supervisors. *Human Resource Management, 46,* 411-427.

Major, D. A., Davis, D. D., Sanchez-Hucles, J., Germano, L. M., & Mann, J. (2006). IT workplace climate for opportunity and inclusion. In E. M. Trauth (Ed.), *Encyclopedia of gender and information technology* (Vol. 2, pp. 856-862). Hershey, PA: Idea Group Reference.

Major, D.A., Davis, D.D., Sanchez-Hucles, J., & Mann, J. (2003, October). Climate for opportunity and inclusion: Improving the recruitment, retention, and advancement of women and minorities in IT. In *Proceedings of the National Science Foundation's ITWF & ITR/EWF Principal Investigator Conference* (pp. 167-171). Albuquerque, NM: The University of New Mexico.

Margolis, J., & Fisher, A. (2003). *Unlocking the clubhouse: Women in computing*. Cambridge, MA: MIT Press.

Martin, U. (2005). *New group aims to get women into top IT research posts by the British Computer Society.* Retrieved December 11, 2007, from http://www.egovmonitor.com/node1178

Mayo, C. (1982). Training for positive marginality. In C. L. Bickman (Ed.), *Applied social psychology annual* (Vol. 3, pp. 57-73). Beverly Hills, CA: Sage.

McNatt, D.B. (2000). Ancient Pygmalion joins contemporary management: A meta-analysis of the result. *Journal of Applied Psychology, 85,* 314-322.

Murrell, A. J., Crosby, F. J., & Ely, R J. (1999). *Mentoring dilemmas: Developmental relationships within multicultural organizations.* Mahwah, NJ: Lawrence Erlbaum.

Nielsen, S.H., von Hellens, L.A., Greenhill, A., & Pringle, R. (1997). Collectivism and connectivity: Culture

and gender in information technology. In *Proceedings of the 1997 ACM SIGCPR Conference on Computer Personnel Research,* San Francisco, CA, (pp. 9-13).

Panteli, A., Stack, J., & Ramsay, H. (1999). Gender and professional ethics in the IT industry. *Journal of Business Ethics, 22*(1), 51-61.

Pegher, V., Quesenberry, J. L., & Trauth, E. M. (2006). A reflexive analysis of questions for women entering the IT workforce. In E. M. Trauth (Ed.), *Encyclopedia of gender and information technology* (Vol. 2, pp. 1075-1080). Hershey, PA: Idea Group Reference.

Podolny, J.M., & Baron, J.N. (1997). Resources and relationships: Social networks and mobility in the workplace. *American Sociological Review, 62,* 673-693.

Ragins, B. R., & Cotton, J. L. (1999). Mentor functions and outcomes: A comparison of men and women in formal and informal mentoring relationships. *Journal of Applied Psychology, 84,* 529-550.

Ragins, B.R., & McFarlin, D.B. (1990). Perceptions of mentor roles in cross-gender mentoring relationships. *Journal of Vocational Behavior, 37,* 321-339.

Roldan, M., Soe, L., & Yakura, E.K. (2004). Perceptions of chilly IT organizational contexts and their effect on the retention and promotion of women in IT. In *Proceedings of the 2004 SIGMIS Conference on Computer Personnel Research: Careers, Culture and Ethics in a Networked Environment,* Tucson, AZ, (pp. 108-133).

Rosenblum, J. L., Ash, R. A., Coder, L., & Dupont, B. (2006). IT workforce composition and characteristics. In E. Trauth (Ed.), *Encyclopedia of gender and information technology* (Vol. 2, pp. 850-855). Hershey, PA: Idea Group Reference.

Rosser, S.V. (2005). Women in ICT: Global issues and actions. In *Proceedings of the International Symposium of Women and ICT: Creating Global Transformation,* Baltimore, MD, (pp. 7-12).

Scandura, T. A. (1992). Mentorship and career mobility: An empirical investigation. *Journal of Organizational Behavior, 13,* 169-174.

Schneider, B. (1987). The people make the place. *Personnel Psychology, 40,* 437-454.

Schneider, B., Wheeler, J.K., & Cox, J.F. (1992). A passion for service: Using content analysis to explicate service climate themes. *Journal of Applied Psychology, 77,* 705-716.

Schwiebert, V. L., Deck, M. D., Bradshaw, M. L., Scott, P., & Harper, M. (1999). Women as mentors. *Journal of Humanistic Counseling, Education, and Development, 37,* 241-253.

Seibert, S., Kraimer, M.L., & Liden, R.C. (2001). A social capital theory of career success. *Academy of Management Journal, 44,* 219-237.

Splender, D. (1997). The position of women in information technology–or who got there first and with what consequences, *Current Sociology, 45*(2), 135-147.

Srivasta, S., & Cooperrider, D.L. (1990). *Appreciative management and leadership: The power of positive thought and action in organizations.* San Francisco: Jossey-Bass.

Steele, C.M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52,* 613-629.

Sumner, M., & Werner, K. (2001). The impact of gender differences on the career experiences of information systems professionals. In *Proceedings of the ACM SIGCPR Conference on Computer Personnel Research,* San Diego, CA, (pp. 125-131).

Tapia, A., Kvasny, L., & Trauth, E. (2004). Is there a retention gap for women and minorities? The case for moving in vs. moving up. In M. Igbaria & S. Conrad (Eds.), *Strategies for managing IS/IT personnel* (pp. 228-244). Hershey, PA: Idea Group.

Trauth, E. M. (2002). Odd girl out: An individual differences perspective on women in the IT profession. *Information Technology and People, 15*(2), 98-118.

Valian, V. (1998). Why so slow? The advancement of women. Cambridge, MA: MIT Press.

KEY TERMS

Active Coping: Directly seeking to overcome stressors and reduce their impact by (a) planning, (b) suppressing competing activities (i.e., focusing on

the stressor at hand), (c) restraint coping (i.e., waiting for an appropriate time to act), and (d) seeking social support (Carver et al., 1989).

Appreciative Inquiry: An intervention that focuses on an organization's positive aspects and uniting individuals under a single positive view.

Glass Ceiling: The condition in which men hold top-level positions and women are limited in their ability to move up due to barriers that are not readily apparent.

Inclusive Climate: Workers' perception of a workplace atmosphere where everyone has a sense of belonging, is invited to participate in decisions, and feels that their input matters (Major et al., 2006).

Mentoring: A process by which a more experienced employee (a mentor) guides, advises, counsels and otherwise enhances the professional development of another employee (a protégé).

Positive Marginality: The constructive and positive use of one's minority experiences of exclusion.

Self-Fulfilling Prophecy: A phenomenon that begins with an assumption about a person or group, and evokes behavior that makes the original conception become reality.

Stereotype Threat: The fear that an underrepresented group member experiences about confirming or reinforcing a negative stereotype about their group; it leads to impaired performance.

Increasing the Accuracy of Predictive Algorithms: A Review of Ensembles of Classifiers

Sotiris Kotsiantis
University of Patras, Greece & University of Peloponnese, Greece

Dimitris Kanellopoulos
University of Patras, Greece

Panayotis Pintelas
University of Patras, Greece & University of Peloponnese, Greece

INTRODUCTION

In classification learning, the learning scheme is presented with a set of classified examples from which it is expected tone can learn a way of classifying unseen examples (see Table 1).

Formally, the problem can be stated as follows: Given training data $\{(x_1, y_1)\ldots(x_n, y_n)\}$, produce a classifier h: X->Y that maps an object x ϵ X to its classification label y ϵ Y. A large number of classification techniques have been developed based on artificial intelligence (logic-based techniques, perception-based techniques) and statistics (Bayesian networks, instance-based techniques). No single learning algorithm can uniformly outperform other algorithms over all data sets.

The concept of combining classifiers is proposed as a new direction for the improvement of the performance of individual machine learning algorithms. Numerous methods have been suggested for the creation of ensembles of classifiers (Dietterich, 2000). Although, or perhaps because, many methods of ensemble creation have been proposed, there is as yet no clear picture of which method is best.

Table 1. Instances with known labels (the corresponding correct outputs)

Data in standard format					
case	Feature 1	Feature 2	...	Feature n	Class
1	xxx	x		xx	good
2	xxx	x		xx	good
3	xxx	x		xx	bad
...					...

BACKGROUND

Generally, support vector machines (SVMs; Scholkopf, Burges, & Smola, 1999) and artificial neural networks (ANNs; Mitchell, 1997) tend to perform much better when dealing with multidimensions and continuous features. In contrast, logic-based systems (e.g., decision trees [Murthy, 1998] and rule learners [Furnkranz, 1999]) tend to perform better when dealing with discrete or categorical features. For neural-network models and SVMs, a large sample size is required in order to achieve the maximum prediction accuracy whereas the naive Bayes model (Jensen, 1996) may need a relatively small data set. Most decision-tree algorithms cannot perform well with problems that require diagonal partitioning. The division of the instance space is orthogonal to the axis of one variable and parallel to all other axes. Therefore, the resulting regions after partitioning are all hyperrectangles. The ANNs and the SVMs perform well when multicolinearity is present and a nonlinear relationship exists between the input and output features.

Although training time varies according to the nature of the application task and data set, specialists generally agree on a partial ordering of the major classes of learning algorithms. For instance, lazy learning methods require zero training time because the training instance is simply stored (Aha, 1997). Naive Bayes methods also train very quickly since they require only a single pass on the data either to count frequencies (for discrete variables) or to compute the normal probability density function (for continuous variables under normality assumptions). Univariate decision trees are also reputed to be quite fast—at any rate, several orders of magnitude faster than neural networks and SVMs.

Naive Bayes methods require little storage space during both the training and classification stages: The strict minimum is the memory needed to store the prior and conditional probabilities. The basic *k*-nearest-neighbor (k-NN) algorithm

(Aha, 1997) uses a great deal of storage space for the training phase, and its execution space is at least as big as its training space. On the contrary, for all nonlazy learners, the execution space is usually much smaller than the training space since the resulting classifier is usually a highly condensed summary of the data.

There is general agreement that k-NN is very sensitive to irrelevant features: This characteristic can be explained by the way the algorithm works. In addition, the presence of irrelevant features can make neural-network training very inefficient and even impractical. Logic-based algorithms are all considered very easy to interpret, whereas neural networks and SVMs have notoriously poor interpretability. k-NN is also considered to have very poor interpretability because an unstructured collection of training instances is far from readable, especially if there are many of them.

While interpretability concerns the typical classifier generated by a learning algorithm, transparency refers to whether the principle of the method is easily understood. A particularly eloquent case is that of k-NN; while the resulting classifier is not quite interpretable, the method itself is very transparent because it appeals to the intuition of human users, who spontaneously reason in a similar manner. Similarly, naive Bayes methods are very transparent as they are easily grasped by users, like physicians, who find that probabilistic explanations replicate their way of diagnosing. Moreover, decision trees and rules are credited with high transparency.

MAIN FOCUS OF THE ARTICLE

Mechanisms that are used to build ensembles of classifiers include (a) using different subsets of training data with a single learning method, (b) using different training parameters with a single training method (e.g., using different initial weights for each neural network in an ensemble), and (c) using different learning methods.

Bagging is a method for building ensembles that uses different subsets of training data with a single learning method (Breiman, 1996). Given a training set of size *t*, bagging draws *t* random instances from the data set with replacement (i.e., using a uniform distribution). These *t* instances are learned, and this process is repeated several times. Since the draw is with replacement, usually the instances drawn will contain some duplicates and some omissions as compared to the original training set. Each cycle through the process results in one classifier. After the construction of several classifiers, taking a vote of the predictions of each classifier produces the final prediction. Another method that uses different subsets of training data with a single learning method is the boosting approach (Freund & Schapire, 1997). Boosting is similar in overall structure to bagging except that it keeps track of the performance of the learning algorithm and concentrates on instances that have not been correctly learned. Instead of choosing the *t* training instances randomly using a uniform distribution, it chooses the training instances in such a manner as to favor the instances that have not been accurately learned. After several cycles, the prediction is performed by taking a weighted vote of the predictions of each classifier, with the weights being proportional to each classifier's accuracy on the training set. AdaBoost is a practical version of the boosting approach (Freund & Schapire). A number of studies that compare AdaBoost and bagging suggest that AdaBoost and bagging have quite different operational profiles (Bauer & Kohavi, 1999; Opitz & Maclin, 1999). In general, it appears that bagging is more consistent, increasing the error of the base learner less frequently than does AdaBoost. However, AdaBoost appears to have greater average effect, leading to substantially larger error reductions than bagging on average. A number of recent studies have shown that the decomposition of a classifier's error into bias and variance terms can provide considerable insight into the prediction performance of the classifier (Bauer & Kohavi). Bias measures the contribution to error of the central tendency of the classifier when trained on different data. Variance is a measure of the contribution to error of deviations from the central tendency. Generally, bagging tends to decrease variance without unduly affecting bias (Breiman; Bauer & Kohavi). On the contrary, in empirical studies, AdaBoost appears to reduce both bias and variance (Breiman; Bauer & Kohavi). Thus, AdaBoost is more effective at reducing bias than bagging, but bagging is more effective than AdaBoost at reducing variance. The decision on limiting the number of subclassifiers is important for practical applications. To be competitive, it is important that the algorithms run in reasonable time. Quinlan (1996) used only 10 replications, while Bauer and Kohavi used 25 replications, Breiman used 50, and Freund and Schapire used 100. For both bagging and boosting, much of the reduction in error appears to have occurred after 10 to 15 classifiers. However, AdaBoost continues to measurably improve test-set error until around 25 classifiers for decision trees (Opitz & Maclin, 1999). As mentioned in Bauer and Kohavi, the main problem with boosting seems to be robustness to noise. On the contrary, they pointed out that bagging improves the accuracy in all data sets used in the experimental evaluation. MultiBoosting (Webb, 2000) is another method of the same category. It can be conceptualized as wagging committees formed by AdaBoost. Wagging is a variant of bagging: Bagging uses resampling to get the data sets for training and producing a weak hypothesis, whereas wagging uses reweighting for each training instance, pursuing the effect of bagging in a different way. Webb, in a number of experiments, showed that MultiBoost achieved greater mean error reductions than AdaBoost or bagging decision trees in both committee sizes

that were investigated (10 and 100). Another metalearner, DECORATE (diverse ensemble creation by oppositional relabeling of artificial training examples), was presented by Melville and Mooney (2003). This method uses a learner (one that provides high accuracy on the training data) to build a diverse committee. This is accomplished by adding different randomly constructed examples to the training set when building new committee members. These artificially constructed examples are given category labels that disagree with the current decision of the committee, thereby directly increasing diversity when a new classifier is trained on the augmented data and added to the committee.

There are also methods for creating ensembles that produce classifiers that disagree on their predictions. Generally, these methods focus on altering the training process in the hope that the resulting classifiers will produce different predictions. For example, neural-network techniques that have been employed include methods for training with different topologies, different initial weights, and different parameters (Mitchell, 1997). Another effective approach for the generation of a set of base classifiers is ensemble feature selection. Ensemble feature selection involves finding a set of feature subsets for the generation of the base classifiers for an ensemble with one learning algorithm. Ho (1998) has shown that the simple random selection of feature subsets may be an effective technique for ensemble feature selection. This technique is called the random subspace method (RSM). In RSM, one randomly selects $N^*<N$ features from the N-dimensional data set. By this, one obtains the N^*-dimensional random subspace of the original N-dimensional feature space. This is repeated S times so as to get S feature subsets for constructing the base classifiers. Then, one constructs classifiers in the random subspaces and aggregates them in the final integration procedure. An experiment with a systematic partition of the feature space, using nine different combination schemes, was performed by Kuncheva and Whitaker (2001), showing that there is no best combination for all situations and that there is no assurance that in all cases a classifier team will outperform the single best individual.

Voting denotes the simplest method of combining predictions from multiple classifiers (Roli, Giacinto, & Vernazza, 2001). In its simplest form, called plurality or majority voting, each classification model contributes a single vote. The collective prediction is decided by the majority of the votes; that is, the class with the most votes is the final prediction. In weighted voting, on the other hand, the classifiers have varying degrees of influence on the collective prediction, which is relative to their predictive accuracy. Each classifier is associated with a specific weight determined by its performance (e.g., accuracy, cost model) on a validation set. The final prediction is decided by summing up all weighted votes and by choosing the class with the highest aggregate. Kotsiantis and Pintelas (2004) combined the advantages of classifier fusion and dynamic selection. The algorithms that are initially used to build the ensemble are tested on a small subset of the training set and, if they have statistically worse accuracy than the most accurate algorithm, they do not participate in the final voting. Stacking (Ting & Witten, 1999) aims to improve efficiency and scalability by executing a number of learning processes and combining the collective results. The main difference between voting and stacking is that the latter combines base classifiers in a nonlinear fashion. The combining task, called a metalearner, integrates the independently computed base classifiers into a higher level classifier, a metaclassifier, by relearning the metalevel training set. This metalevel training set is created by using the base classifiers' predictions on the validation set as attribute values and the true class as the target. Ting and Witten have shown that successful stacked generalization requires the use of output class distributions rather than class predictions. In their experiments, only the MLR algorithm (a linear discriminant) was suitable for use as a Level 1 classifier. Cascade generalization (Gama & Brazdil, 2000) is another algorithm that belongs to the family of stacking algorithms. Cascade generalization uses the set of classifiers sequentially, at each step performing an extension of the original data by the insertion of new attributes. The new attributes are derived from the probability class distribution given by a base classifier. This constructive step extends the representational language for the high-level classifiers, reducing their bias. Todorovski and Dzeroski (2003) introduced meta-decision-trees (MDTs). Instead of giving a prediction, MDT leaves specify which classifier should be used to obtain a prediction. Each leaf of the MDT represents a part of the data set, which is a relative area of expertise of the base-level classifier in that leaf. MDTs can use the diversity of the base-level classifiers better than voting, thus outperforming voting schemes in terms of accuracy, especially in domains with a high diversity of errors made by base-level classifiers.

Another attempt to improve classification accuracy is the use of hybrid techniques. Lazkano and Sierra (2003) presented a hybrid classifier that combines the Bayesian network algorithm with the nearest-neighbor distance-based algorithm. The Bayesian network structure is obtained from the data and the nearest-neighbor algorithm is used in combination with the Bayesian network in the deduction phase. Wang, Yuan, Li, and Li (2004) presented flexible NBTree: a decision-tree learning algorithm in which nodes contain univariate splits as do regular decision trees, but the leaf nodes contain general naive Bayes classifiers, which is a variant of the standard naive Bayesian classifier. Zhou and Chen (2002) generated a binary hybrid decision tree according to the binary information gain ratio criterion. If attributes cannot further distinguish training examples fall-

ing into a leaf node whose diversity is beyond the diversity threshold, then the node is marked as a dummy node and a feed-forward neural network named FANNC is then trained in the instance space defined by the used attributes. Zheng and Webb (2000) proposed the application of lazy learning techniques to Bayesian induction and presented the resulting lazy Bayesian rule learning algorithm, called LBR. This algorithm can be justified by a variant of the Bayes model, which supports a weaker conditional attribute independence assumption than is required by naive Bayes. For each test example, it builds the most appropriate rule with a local naive Bayesian classifier as its consequent. Xie, Hsu, Liu, and Lee (2002) proposed a similar lazy learning algorithm: selective neighborhood-based naive Bayes (SNNB). SNNB computes different distance neighborhoods of the new input object, lazily learns multiple naive Bayes classifiers, and uses the classifier with the highest estimated accuracy to make the final decision. Domeniconi and Gunopulos (2001) combined local learning with SVMs. In this approach, an SVM is used to determine the weights of the local neighborhood instances.

FUTURE TRENDS

The key question when dealing with classification is not whether a learning algorithm is superior to others, but under which conditions a particular method can significantly outperform others on a given application problem. Metalearning is moving in this direction, trying to find functions that map data sets to algorithm performance (Kalousis & Gama, 2004). To this end, metalearning uses a set of attributes, called metaattributes, to represent the characteristics of learning tasks, and searches for the correlations between these attributes and the performance of learning algorithms. Some characteristics of learning tasks are the number of instances, the proportion of categorical attributes, the proportion of missing values, the entropy of classes, and so forth. Brazdil, Soares, and Da Costa (2003) provided an extensive list of information and statistical measures for a data set.

CONCLUSION

After a better understanding of the strengths and limitations of each method, the possibility of integrating two or more algorithms together to solve a problem should be investigated. The objective is to utilize the strengths of one method to complement the weaknesses of another. If we are only interested in the best possible classification accuracy, it might be difficult or impossible to find a single classifier that performs as well as a good ensemble of classifiers.

REFERENCES

Aha, D. (1997). *Lazy learning.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning, 36*, 105-139.

Brazdil, P., Soares, C., & Da Costa, J. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning, 50*, 251-277.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*, 123-140.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning, 40*, 139-157.

Domeniconi, C., & Gunopulos, D. (2001). Adaptive nearest neighbor classification using support vector machines. *Advances in Neural Information Processing Systems, 14*, 665-672.

Freund, Y., & Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *JCSS, 55*(1), 119-139.

Furnkranz, J. (1999). Separate-and-conquer rule learning. *Artificial Intelligence Review, 13*, 3-54.

Gama, J., & Brazdil, P. (2000). Cascade generalization. *Machine Learning, 41*, 315-343.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 832-844.

Jensen, F. (1996). *An introduction to Bayesian networks.* Springer.

Kalousis, A., & Gama, G. (2004). On data and algorithms: Understanding inductive performance. *Machine Learning, 54*, 275-312.

Kotsiantis, S., & Pintelas, P. (2004, August 26-28). Selective voting. In *Proceedings of the Fourth International Conference on Intelligent Systems Design and Applications (ISDA 2004)*, Budapest, Hungary (pp. 397-402).

Kuncheva, L., & Whitaker, C. (2001). Feature subsets for classifier combination: An enumerative experiment. In *Lecture notes in computer science* (Vol. 2096, pp. 228-237). Springer-Verlag.

Lazkano, E., & Sierra, B. (2003). BAYES-NEAREST: A new hybrid classifier combining Bayesian network and distance

based algorithms. In *Lecture notes in computer science* (Vol. 2902, pp. 171-183).

Melville, P., & Mooney, R. (2003). Constructing diverse classifier ensembles using artificial training examples. In *Proceedings of the IJCAI-2003*, Acapulco, Mexico (pp. 505-510).

Mitchell, T. (1997). *Machine learning.* McGraw Hill.

Murthy. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery, 2*, 345-389.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Artificial Intelligence Research, 11*, 169-198.

Quinlan, J. R. (1996). Bagging, boosting, and C4.5. In *Proceedings of the 13th National Conference on Artificial Intelligence* (pp. 725-730). AAAI/MIT Press.

Roli, F., Giacinto, G., & Vernazza, G. (2001). Methods for designing multiple classifier systems. In *Lecture notes in computer science* (Vol. 2096, pp. 78-87). Springer-Verlag.

Scholkopf, C., Burges, J. C., & Smola, A. J. (1999). *Advances in kernel methods.* MIT Press.

Ting, K., & Witten, I. (1999). Issues in stacked generalization. *Artificial Intelligence Research, 10*, 271-289.

Todorovski, L., & Dzeroski, S. (2003). Combining classifiers with meta decision trees. *Machine Learning, 50*, 223-249.

Villada, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Artificial Intelligence Review, 18*, 77-95.

Wang, L., Yuan, S., Li, L., & Li, H. (2004). Improving the performance of decision tree: A hybrid approach. In *Lecture notes in computer science* (Vol. 3288, pp. 327-335).

Webb, G. I. (2000). MultiBoosting: A technique for combining boosting and wagging. *Machine Learning, 40*, 159-196.

Xie, Z., Hsu, W., Liu, Z., & Lee, M. L. (2002). SNNB: A selective neighborhood based naive Bayes for lazy learning. In *Lecture notes in computer science* (Vol. 2336, pp. 104-115).

Zheng, Z., & Webb, G. (2000). Lazy learning of Bayesian rules. *Machine Learning, 41*(1), 53-84.

Zhou, Z., & Chen, Z. (2002). Hybrid decision tree. *Knowledge-Based Systems, 15*(8), 515-528.

KEY TERMS

Artificial Neural Networks: They are nonlinear predictive models that learn through training and resemble biological neural networks in structure.

Bagging: Bagging uses different subsets of training data with a single learning method. After the construction of several classifiers, taking a vote of the predictions of each classifier produces the final prediction.

Boosting: It is similar in overall structure to bagging, except that it keeps track of the performance of the learning algorithm and concentrates on instances that have not been correctly learned.

Cascade Generalization: It uses the set of classifiers sequentially, at each step performing an extension of the original data by the insertion of new attributes.

Majority Voting: The collective prediction is decided by the majority of the votes of the classifiers; that is, the class with the most votes is the final prediction.

Rule Induction: It is the extraction of useful if-then rules from data based on statistical significance.

Stacking: Stacking integrates the independently computed base classifiers into a higher level classifier: a metaclassifier.

Indexing Techniques for Spatiotemporal Databases

George Lagogiannis
University of Patras, Greece

Christos Makris
University of Patras, Greece

Yannis Panagis
University of Patras, Greece

Spyros Sioutas
University of Patras, Greece

Evangelos Theodoridis
University of Patras, Greece

Athanasios Tsakalidis
University of Patras, Greece

INTRODUCTION

We can define as spatiotemporal any database that maintains objects with geometric properties that change over time, where usual geometric properties are the spatial position and spatial extent of an object in a specific *d*-dimensional space. The need to use spatiotemporal databases appears in a variety of applications such as intelligent transportation systems, cellular communications, and meteorology monitoring. This field of database research collaborates tightly with other research areas such as mobile telecommunications, and is harmonically integrated with other disciplines such as CAD/CAM, GIS, environmental science, and bioinformatics.

Spatiotemporal databases stand at the crossroad of two other database research areas: spatial databases (Güting, 1994; Gaede & Gunther, 1998) and temporal databases (Salzberg & Tsotras, 1999). The efficient implementation of spatiotemporal databases needs new data models and query languages and novel access structures for storing and accessing information. In Güting, Bohlen, Erwig, Jensen, Lorentzos, Schneider, and Vazirgiannis (2000) a data model and a query language capable of handling such time-dependent geometries, including those changing continuously that describe moving objects, were proposed; the basic idea was to represent time-dependent geometries as attribute data types and to provide an abstract data type extension to the traditional database data models and query languages. In that paper, it was also discussed how various temporal and spatial models could possibly be extended to be spatiotemporal models.

BASIC ALGORITHMIC TOOLS

The problem of spatiotemporal indexing is considered in the standard external memory model. In this model each disk access transfers a contiguous block of B data items in a single input/output operation (I/O). The space complexity of a data structure is measured in terms of the amount of disk blocks it uses, while the time complexity of its various operations is expressed with the number of needed I/Os. In the sequel, we let N denote the number of stored objects and let $n=N/B$ and $k=K/B$, where K denotes the size of the desired output.

R-Tree and Variants

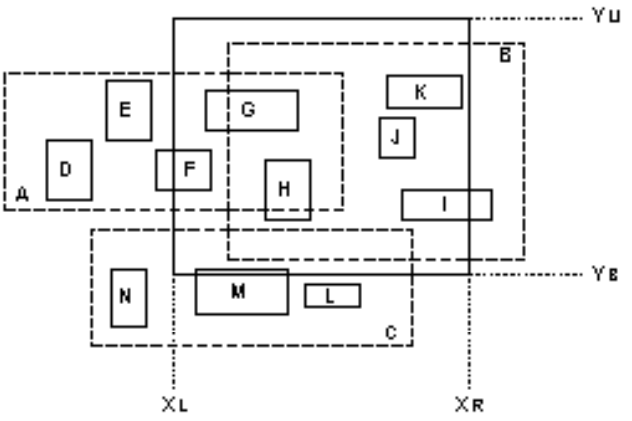
The R-tree (Guttman, 1984) is a spatial access method; it stores objects in R^d with a spatial position and extent and is able to answer various geometric queries. In a nutshell, it is an hierarchical structure, resembling a classical B-tree, where every node has size equal to the disk block size, all objects are stored at the leaves of the structure, and all leaves are at the same distance from the root. Each node v of the R-tree corresponds to a d-dimensional rectangle Rect(v); the rectangle corresponding to a leaf is the minimum rectangle that encloses the objects stored at the leaf, while the rectangle

corresponding to an internal node encloses all the rectangles corresponding to its children. Moreover, every node contains between m and M children where m, M are parameters whose value depends on the block size; this means that the height of an R-tree is $O(\log_m n)$. Searching in the R-tree is done in a similar way as in the B-tree, where for both point and region queries, the paths where rectangles intersect with the query object are followed. In contrast to the B-tree, the R-tree does not guarantee that traversing a path of the tree is enough when searching for an object, as the bounding rectangles of entries in the same nodes may overlap one another. In the worst case, the search algorithm may have to visit all nodes, in order to answer a query. See, for example, the 2-dimensional R-tree of Figure 1 that stores 11 rectangles (n=11) with block size B=4. For the given axis-parallel range query (x_L, x_R, y_B, y_U) of the figure below, the search algorithm may have to visit all nodes in order to determine the K=7 rectangles that are enclosed or intersected by the query rectangle.

The search efficiency of the R-tree could be improved if the spatial overlap between sibling nodes could be minimized. There exists a set of heuristics to achieve that, leading to two known variants of the R-tree: the R+-trees (Sellis, Roussopoulos, & Faloutsos, 1987) and the R*-trees (Beckmann, Kriegel, Schneider, & Seeger, 1990). R+-trees do not allow the subspaces of each internal node to overlap with the subspace of its sibling nodes. However, this clipping of the subspaces has, as an immediate consequence, the creation of much more subspaces and thus much more leaves. Also an object is assigned to more than one leaf causing a much larger tree structure. On the other hand, R*-trees embed maintenance algorithms that aim at minimizing the following penalty metrics: (i) the area and the perimeter of each bounding rectangles, (ii) the overlap between two sibling bounding rectangles, and (iii) the distance between the centroid of a bounding rectangle and that of the node containing it. As discussed in the original publication, the minimization of these metrics decreases the probability that a node is accessed by a range query. Experimental studies show that R*-trees can achieve 50% better query times from the common R-trees.

The interested reader should consult Gaede and Gunther (1998) and Manolopoulos, Theodoridis, and Tsotras (2000) for more information concerning the aforementioned structures.

Figure 1. An example of an R-tree



Partition Trees

Partition trees are data structures that can handle simplex range searching queries and are based on the idea of simplicial partitions. A simplicial partition for a set S of N points in R^d is a collection of pairs $P(S) = \{(S_1, s_1\}, (S_2, s_2), \ldots, (S_m, s_m)\}$ where the S_i's are disjoint subsets of S whose union is S and s_i is a simplex containing S_i. The crossing number of $P(S)$ is the maximum number of simplices in $\{s_1, \ldots, s_m\}$ that can be crossed by an arbitrary hyperplane. Matousek (1992) has proved that for any set S and given parameter r, $r \leq N$, it is possible to construct a simplicial partition of size r for S, whose crossing number is $O(r^{1-1/d})$, in $O(N^{1+\varepsilon})$ time, for any $\varepsilon>0$. Using this theorem recursively, it is possible to come up with a partition tree T for the set S. The root node has $O(r)$ children, which correspond to the simplices of the initial partition and are the roots of recursively defined partition trees for the simplicial partitions. A query with a given query simplex can be answered by starting at the root of T and descending towards the leaves, based on the relation between the query simplex and the simplices in the partitions of the various nodes. In Agarwal, Arge, Erickson, Franciosa, and Vitter (2000), they described an external memory version of static partition trees that needed $O(n)$ disk blocks, so that d-dimensional simplex queries could be answered in $O(n^{1-1/d+\varepsilon}+k)$ I/O s.

INDEXING TECHNIQUES

Research on spatiotemporal access methods has mainly focused on two aspects: (i) storage and retrieval of historical information concerning the positions of the moving points, and (ii) prediction of future positions. There are two kinds of spatiotemporal databases: those that deal with *discrete* and those that deal with *continuous* movements. In the sequel, we will briefly refer to discrete spatiotemporal databases, and we will mainly focus our presentation on continuous spatiotemporal databases.

A sequence of spatiotemporal movements in a discrete environment can be considered to be an ordered sequence of database snapshots of the object positions/extents taken at time instants $t_1<t_2<\ldots$, with each time instant denoting the moment where a change took place. By taking this point of view then it could be possible to handle the index-

ing problems in such environments by suitably extending indexing techniques from the area of temporal (Salzberg & Tsotras, 1999) and of spatial databases (Gaede & Gunther, 1998); in Manolopoulos, Theodoridis, and Tsotras (2000), it is elegantly exposed how these indexing techniques can be generalized in order to handle efficiently queries in a discrete spatiotemporal environment.

In the continuous spatiotemporal environment there exists a plethora of efficient data structures (Agarwal, Arge, & Erickson, 2003; Kollios, Gunopulos, & Tsotras, 1999; Patel, Chen, & Chakka, 2004; Saltenis, & Jensen, 2002; Saltenis, Jensen, Leutenegger, & Lopez , 2000; Tao, Papadias, & Sun, 2003). The common thrust behind these structures lies in the idea of abstracting each object's position as a continuous function *f(t)* of time and updating the database whenever the parameters of the function change; accordingly an object is modeled as a pair consisted of its extent at a reference time (design parameter) and of its motion vector.

The basic queries supported by these structures are: (Q_1) given an orthogonal range and a time stamp *t*, report the objects that lie in the query range at time *t*, (Q_2) given an orthogonal range and two time stamps t_1 , t_2 report all the objects that lie in the query range at any time instant between t_1, t_2, (Q_3) given a query point *q*, a constant δ>0 and a time stamp *t*, report a δ-approximate neighbor of *p* at this timestamp (δ designates the degree of the approximation).

We can partition the most prominent of the aforementioned indexing structures into two broad categories; those that are based on geometric duality and represent the stored objects in the dual space (Agarwal, Arge, & Erickson, 2003; Kollios, Gunopulos, & Tsotras, 1999; Patel, Chen, & Chakka, 2004) and those that leave the original representation intact indexing data in their native *d*-dimensional space (Saltenis, Jensen, Leutenegger, & Lopez, 2000; Saltenis, & Jensen, 2002; Tao, Papadias, & Sun, 2003). The geometric duality transformation is a tool heavily used in the computational geometry literature that maps hyper-planes in R^d to points and vice-versa. In general, the straightforward approach of representing an object moving on an 1-dimensional line is by plotting the trajectories as lines in the time-location (*t*,*y*) plane (same for (*t*,*x*) plane). The equation describing each line is $y(t)=ut+a$ where *u* is the slope (velocity vector in this case) and *a* is the intercept (initial position vector in this case), which is computed using the motion information (Figure 2). Based on this setting, the query is expressed as the 2-dimensional interval $[(y_{1q},y_{2q}),(t_{1q},t_{2q})]$, and it reports the objects that correspond to the lines intersecting the query rectangle.

For example, one dual transform for mapping the line with equation $y(t)=ut+a$ to a point in R^2 is by using the dual plane where one axis represents the slope of an object's trajectory (i.e. velocity) and the other axis its intercept. Thus we have the dual point *(u,a)*. Accordingly, the 1-d query $[(y_{1q},y_{2q}),(t_{1q},t_{2q})]$ becomes a polygon in the dual space (see Figure 3).

In Kollios, Gunopulos, and Tsotras (1999) a set of indexing techniques were presented supporting operations Q_1, Q_2 for objects in the one- and two-dimensional Euclidean space. For the one-dimensional case the application of duality transformation makes the problem of indexing moving objects equivalent to the simplex range searching problem in two dimensions which is solved by applying the external memory partition tree of Agarwal, Arge, Erickson, Franciosa, and Vitter (2000); this tree uses O(*n*) space, answers queries in $O(n^{1/2+\varepsilon}+k)$ I/Os and handles updates in O(log*N*) I/Os. Partition trees, though having a guaranteed worst case performance are generally considered non-practical since they entail large hidden factors. Hence, in (Kollios, Gunopulos, & Tsotras, 1999) two more structures are presented: one based on k-d-trees and a more complex one based on B+-trees. Both these structures use linear space and work well in the average case. Moreover they extend their results in the two-dimensional case for two distinct versions of the problem; in the first version the objects are allowed

Figure 2. Trajectories and query in (t,y) plane

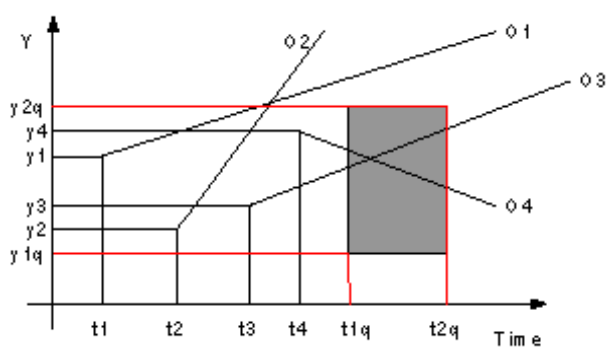
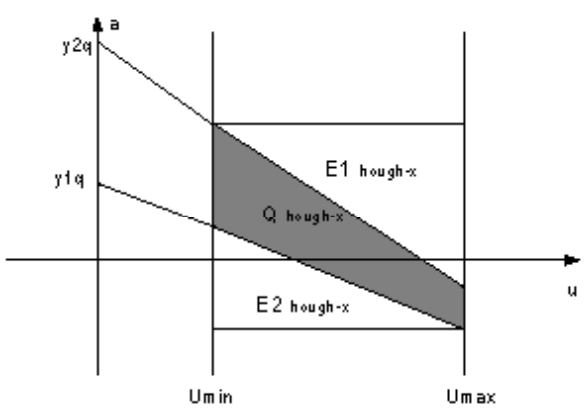


Figure 3. Query on the dual plane



to move on a network of one-dimensional routes while the second version allows arbitrary movements. The first version, reduces to a number of one-dimensional subproblems that use the previously described structures, while the second is equivalent (through geometric duality) to simplex range queries in R^3, which can be solved in $O(n^{2/3}+k)$ I/Os with the use of external memory partition trees.

In Agarwal, Arge, Erickson, Franciosa, and Vitter (2000), results were refined and improved. In particular, a new version of partition tree was introduced that could handle questions Q_1, Q_2 for objects in the plane in $O(n)$ space, $O(n^{1/2+e}+k)$ query time, and $O(\log_B n)$ expected amortized update cost; the results apply also for higher dimensional spaces slowing down only the update time (it becomes $O(\log^2_B n)$ I/Os). If it is assumed that the queries arrive in chronological order then the query time for problem Q_1 can be further reduced to $O(\log^2_B n/\log_B \log_B n)$I/Os; this is achieved by employing the kinetic data structures framework (Basch, Guibas, & Hershberger, 1997) at external range trees. Moreover by combining multiversion kinetic data structures with partition trees, they develop an indexing scheme having a query time that is small for near-future queries and is increasing for queries that are far-away, without exceeding the bound of $O(n^{1/2+\varepsilon}+k)$ I/Os. Finally an indexing technique is described for handling δ-approximate queries; the query time is $Ï(n^{1/2+\varepsilon}/\sqrt{\delta})$, the expected update time is $O(\log^2_B n/\sqrt{\delta})$ and the space is $O(n/\sqrt{\delta})$ disk blocks.

The TPR tree was introduced in Saltenis, Jensen, Leutenegger, and Lopez (2000) and is basically a generalization of the R*-tree data structure in order to store and access linearly moving objects. The leaves of the structure store pairs with the position of the moving point and the moving point, while internal nodes store pointers to subtrees with associated rectangles that minimally bound all moving points or other rectangles in the subtree. The difference with the classical R*-tree lies in the fact that the bounding rectangles are time parameterized (their coordinates are functions of time), and it is considered that a time parameterized rectangle bounds all enclosed points or rectangles at all times not earlier than current time. The algorithms for search and update operations in the TPR tree are straightforward generalizations of the respective algorithms in the R*-tree, moreover the various kinds of spatiotemporal queries can be handled uniformly in one-, two-, and three-dimensional spaces.

The TPR-tree constituted the base structure for further developments in the area. Saltenis and Jensen in (Saltenis, & Jensen 2001) presented the R^{EXP}-tree, an R*-based access method that building on the TPR –tree handles efficiently moving objects that have the extra property that their positions may expire after specific time periods; the problem is adequately motivated from Internet-service scenarios, employed by standards such as WAP or Bluetooth, where objects that have not reported their position for a specific time period are considered to have been expired.

In Tao, Papadias, and Sun (2003), an extension to the TPR-tree was proposed, the so called TPR*-tree, that for some scientific publications (Tao, Faloutsos, Papadias, & Liu, 2004) is considered as the most efficient spatiotemporal structure. The crucial improvement (in comparison to the TPR-tree) is in the update operations, where it is shown that local optimization criteria (at each tree node) may degrade seriously the performance of the structure and more particularly in the use of update rules that are based on global optimization criteria. Moreover the authors propose a novel probabilistic cost model for validating the performance of a spatiotemporal index and analyze with this model the optimal performance for any data-partition index.

Finally, in Patel, Chen, and Chakka (2004), the STRIPES index is proposed. STRIPES is based on the application of the duality transformation and it employs disjoint regular space partitions (disk based quadtrees [Gaede, & Gunther, 1998]); the authors claim, though the use of a series of implementations, that STRIPES outperforms TPR*-trees for both update and query operations.

EXTENSIONS

The previously described methods have as common characteristic that the movement of the objects follows a linear function. While this assumption does not hold (obviously) for every movement that needs to be practically handled, this linear model is often justified in two ways: (i) its adoption avoids the complications arising from arbitrary motion patterns and permits the handling of several problems that otherwise could be intractable, and (ii) it is possible, through the effective use of piecewise linear segments to approximate (virtually, to arbitrary precision) any curve. These two arguments are however only partially true since it is possible to construct examples where stored objects follow motion patterns that cannot possibly be represented with the use of linear motion patterns. In order to overcome these problems, in Tao, Faloutsos, Papadias, and Liu (2004), a general framework was introduced, that can handle effectively arbitrary motion types. This general framework contains three important contributions to handle the specific problems: (i) a general client-server architecture for objects with unknown and possibly variable movement types, where a filter refinement mechanism is employed so that candidate objects are contacted for more refined information, (ii) an elegant technique, for expressing in a concise format, the so-called recursive motion function, a large number of movement types, and (iii) an access method, the STP-tree (spatiotemporal prediction tree), for indexing the expected trajectories (at the server). The STP-tree constitutes a generalization of the previously mentioned data structures in order to handle arbitrary polynomial functions; both construction and update algorithms follow closely their counterparts in

TPR and TPR* trees. Finally, the contributions of the presented framework are validated through an extensive set of experimental results.

FUTURE TRENDS

Concerning future trends, since partition trees are considered impractical (due to the large constant factors involved) then various research teams focus on presenting improved variants of the R-tree structure. However, R-trees, due to their overlapping nature, can cause backtracking in search, and they may not be best suited for every application so there is need for more efficient data structures. Finally, another interesting problem (Roddick, Egenhofer, Hoel, Papadias, & Salzberg, 2004) lies in the fact that in quite a few of the latest spatiotemporal applications (i.e., sensor networks) data arrive in the form of data streams; therefore in such applications the indexing techniques should aim at maintaining approximate summarized information, and handle queries effectively using as guide such approximations.

CONCLUSION

In this article, we have presented a set of indexing techniques for spatiotemporal databases; the various presented structures have various advantages and disadvantages and the choice of a particular structure depends on the application, the hardware configuration, and the range of required operations. The use of these structures is motivated by real-life applications such as intelligent transportation systems, meteorology monitoring, and mobile computing.

REFERENCES

Agarwal, P. K., Arge, L., & Erickson, J. (2003). Indexing moving points. *Journal of Computer and System Sciences, 66,* 207-243.

Agarwal, P. K., Arge, L., Erickson, J., Franciosa, P. G., & Vitter, J. S. (2000). Efficient searching with linear constraints. *Journal of Computer and System Sciences, 61,* 194-216.

Basch, J., Guibas, L. J., & Hershberger J. (1997). Data structures for mobile data. *Proceedings of the Eightth ACM SIAM Symposium on Discrete Algorithms* (pp. 747-756).

Beckmann, N., Kriegel, H. Schneider, R., & Seeger, B. (1990). The R*-tree: An efficient and robust access Method for points and rectangles. In *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data* (pp. 322-331).

Gaede, V., & Gunther, O. (1998). Multidimensional access methods. *ACM Computing Surveys, 30*(2), 170-231.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD*, Boston (pp. 47-57).

Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal, 3*(4), 357-399.

Güting, R., Bohlen, M., Erwig, M., Jensen, C., Lorentzos, N., Schneider, M., & Vazirgiannis, M. (2000). A foundation for representing and querying moving objects. *ACM Transactions on Database Systems, 25*(1), 1-42.

Kollios, G., Gunopulos, D., & Tsotras, V. (1999). On indexing mobile objects. In *Proceedings of the 18th ACM Symposeum on Principles of Database Systems (PODS)* (pp. 261-272).

Manolopoulos, Y., Theodoridis, Y., & Tsotras, V. (2000). *Advanced database indexing.* Kluwer Academic Publishers.

Matousek, J. (1992). Efficient partition trees. *Efficient Partition Trees, Discrete and Computational Geometry, 8,* 432-448.

Patel, J., Chen, Y., & Chakka, V. (2004). STRIPES: An efficient index for predictedtrajectories. In *Proceedings of the ACM SIGMOD* (pp. 637-646).

Roddick, J., Egenhofer, M., Hoel, E., Papadias, D., & Salzberg, B. (2004). Spatial, temporal and spatiotemporal databases: Hot issues and directions for PhD research. *SIGMOD Record, 33*(2), 126-131.

Saltenis, S., & Jensen, C. S. (2002). Indexing of moving objects forlocation-based services. In *Proceedings of the 18th International Conference on Data Engineering,* San Jose, CA (pp. 507-518).

Saltenis, S., Jensen, C., Leutenegger, S., & Lopez, M. A. (2000). Indexing the positions of continuously moving objects. In *Proceedings of the ACM SIGMOD* (pp. 331-342).

Salzberg, B., & Tsotras, V. J. (1999). A comparison of access methods for time-evolving data. *ACM Computing Surveys, 31*(2), 158-221.

Sellis, T., Roussopoulos, N., & Faloutsos, C. (1987). The R+-tree : A dynamic index for multidimensional objects. In *Proceedings of the 13th International Conference on Very Large Data Bases* (pp. 507-518).

Tao, Y., Faloutsos, C., Papadias, D., & Liu, B. (2004, June 13-18). Prediction and indexing of moving objects with unknown motion patterns. In *Proceedings of the ACM Conference on the Management of Data (SIGMOD),* Paris (pp. 611-622).

Tao, Y., Papadias, D., & Sun, J. (2003). The TPR*-tree: An optimized spatio-temporal access method for predictive queries. *VLDB* (pp. 790-801).

KEY TERMS

Bitemporal Databases: Temporal databases that support both the valid and the transaction time.

Database: A collection of interrelated persistent information stored and organized as a unit in order to serve a specific purpose and satisfy the demands of a set of users. A database can be considered to be an electronic filling system stored on mass-storage systems such as magnetic tape or disk. A database is one component of a database management system.

Data Structures: A collection of methods for storing and organizing sets of data in order to facilitate access to them. More formally data structures are concise implementations of abstract data types, where an abstract data type is a set of objects together with a collection of operations on the elements of the set.

Secondary Memory Algorithms: Algorithms for handling information on secondary media, like hard disks, CD-ROMs, and so forth, and trying to minimize the number of accesses on them.

Spatial Database: A database storing and handling objects that are equipped with spatial information (a position and an extent).

Spatiotemporal Database: A database that maintains objects with geometric properties that change over time, where usual geometric properties are the spatial position and spatial extent of the object in a specific d-dimensional space.

Temporal Database: A database that stores and handles objects that are equipped with built-in time information. These databases are components of database management systems that should offer a suitable temporal data model and a temporal version of a structured query language.

Transaction Time Databases: Temporal databases, the supported time dimension of which is the transaction time dimension that is the time when a fact is stored in the database. Transaction time is consistent with the transaction serialization order.

Valid Time Databases: Temporal databases, the supported time dimension of which is the valid time dimension that is the time when a fact becomes valid (effective) with respect to the real world.

Indexing Textual Information

Ioannis N. Kouris
University of Patras, Greece

Christos Makris
University of Patras, Greece

Evangelos Theodoridis
University of Patras, Greece

Athanasios Tsakalidis
University of Patras, Greece

INTRODUCTION

Information retrieval is the computational discipline that deals with the efficient representation, organization, and access to information objects that represent natural language texts (Baeza-Yates, & Ribeiro-Neto, 1999; Salton & McGill, 1983; Witten, Moûat, & Bell, 1999). A crucial subproblem in the information retrieval area is the design and implementation of efficient data structures and algorithms for indexing and searching information objects that are vaguely described. In this article, we are going to present the latest developments in the indexing area by giving special emphasis to: data structures and algorithmic techniques for string manipulation, space efficient implementations, and compression techniques for efficient storage of information objects.

The aforementioned problems appear in a series of applications as digital libraries, molecular sequence databases (DNA sequences, protein databases [Gusûeld, 1997)], implementation of Web search engines, web mining and information filtering.

BACKGROUND

Dictionary Data Structures

The dictionary data structure stores a set S of n elements in order to support the operations of insertion, deletion, and the test of membership. A basic criterion for categorizing dictionary data structures is whether only comparisons are used, or the representation of elements for guiding the search is also employed. Typical representatives of the former group are *search trees* and of the latter *tries* and *hashing*. Search trees need O(log$n$) update/search time and O(n) space and the most prominent examples of them are: AVL-trees, red-black trees, (α,b)-trees, BB[α]-trees and Weight Balanced B-trees (Arge, & Vitter, 1996; Cormen, Leiserson, & Rivest, 1990; Mehlhorn, 1984). On the other hand, *tries* and *hashing* structures (Cormen, Leiserson, & Rivest, 1990; Czegh, Havas, & Majewski, 1997; Pagh, 2002) try to use the representation (for example, the value of the element written as a string of digits or the value itself), to compute directly the element's position in system's memory. The time and space complexities of these structures generally vary; however, it should be mentioned that a lately developed structure (Anderson & Thorup, 2001) answers both search and update operations in $O(\sqrt{\log n / \log\log n})$ time. This structure is also able to retrieve the largest element in the stored set smaller than a *query* element (*predecessor* query).

Finding Occurrences of Patterns

The string matching (or pattern matching) problem is one of the most frequently encountered and studied problems in the area of system/algorithm design. In this problem, we are searching for the occurrences of a pattern P in a sequence of symbols T. The naive $O(|P||T|)$ algorithm aligns the pattern at each one of the O($|T|$) possible positions of the sequence and executes O($|P|$) comparisons; however, there exist elegant, though complex, algorithms whose overall time complexity is linear, that is, $O(|P|+ |T|)$. The most known linear time algorithms that achieve that are Knuth-Morris-Pratt (Knuth, Morris, & Pratt, 1977) and variants of the Boyer-Moore (Boyer & Moore, 1977) algorithm. For the case that we are searching for a set of patterns in the sequence, the Aho-Corasick automaton (Aho & Corasick, 1975) can be used. This automaton accepts all the patterns of the set and can be constructed in time linear to the sum of the lengths of them. Running the automaton with the characters of T, all the occurrences of the patterns are reported in $O(|T|)$ time.

On most of the modern applications, the patterns arrive in an on line manner and the $O(|T|+ |P|)$ computational

time is prohibitive; there is need for indexing structures (indices) that can perform the queries as closer as possible to $O(|P|)$ computational time, assuming that the text has been preprocessed *once*. The indices that try to satisfy this demand are divided in two categories: the *word-based* (or *keyword-based*) indices, which have been designed for sequences of symbols that can be divided in tokens/words, and the *full-text* (or *sequential scan*) indices, where the previous feature does not hold and the strings involved are non-tokenizable.

TEXT AND STRING DATA STRUCTURES

Word Based Indices

The most commonly used indexing structures in this category are *inverted files, signature files* and *bitmaps*. An inverted file consists of two parts: a structure for storing the set of all different words in the text and, for each such word, a list of the text positions where the word appearances are stored. Signature files are term-oriented structured based on hashing while bitmaps represent each document as a bit vector having length equal to the size of the lexicon. In typical applications compressed inverted files are considered to be superior to both signature files and bitmaps (Faloutsos, 1985; Zobel, Moffat, & Ramamohanarao, 1998).

More analytically, consider a document collection and a lexicon containing the terms that appear in the documents of the collection. An inverted file consists of a search structure containing all the distinct terms that appear in the lexicon and, for each distinct term, a list termed inverted (or postings) list storing the identifiers (usually integer numbers starting from 1) of the documents containing the term. The search structure can be implemented as a search tree or a hashing structure, and queries are evaluated by using the search structure to find the relevant terms, fetching the inverted lists for the corresponding terms, and then intersecting them for conjunctive queries or merging them for disjunctive queries. There are many algorithms for constructing inverted files while their performance can be significantly improved by employing compression techniques for representing the postings lists. The interested reader should find relevant material in Witten, Moffat, and Bell (1999) and Zobel, Moffat, and Ramamohanarao (1998).

On the other hand, in the signature file method, all the documents in the collection are stored sequentially; each document is hashed (mapped to) a distinct signature and all the signatures are stored in a signature file.

Finally, in bitmaps for every term in the lexicon a bit-vector is stored, each bit corresponding to a document, a bit with value 1 means that the term appears somewhere in the document; otherwise it has value 0.

Full Text Indices

In general, these indices are more powerful from word-based indices since they answer a wider range of queries like arbitrary substring queries, motifs, and repetitions selection queries, and they have the ability to index non-tokenizable strings like biological sequences. The price for these extra capabilities is larger space consumption. The common feature of full-text indices is that they organize in a proper order (frequently lexicographic), all the suffixes of the sequence.

Static Indices: For the setting that texts do not underlie on updates (insertions/deletions of characters), there are many classical confrontations coming quite a few years ago. The most famous data structure of this type is the *suffix tree*. A suffix tree of a string T is a compacted trie of all suffixes of T. The suffix tree has linear, to the length $|T|$ of the string, construction time. For bounded alphabets, there exist the algorithms of Weiner (1973), McCreight (1976) and Ukkonen (1995) and, for arbitrary large alphabets, the algorithm of Farach (1997). The main counterparts of suffix trees are *suffix arrays*. A suffix array of a string T is an array that indicates the lexicographic order of all suffixes of T. Suffix arrays were proposed in Manber, and Myers (1993), where they described an $O(|T|log|T|)$ construction time algorithm and how pattern matching queries can be performed in $O(|P|+log|T|)$ time, assuming that the longest common prefixes among each pair of adjacent suffixes in the array have been precomputed. Later on, there were proposed three linear time construction algorithms in Karkkainen and Sanders (2002); Ko, and Aluru (2003); Kim, Sim, Park, and Park (2003). Suffix arrays, though having a slowdown term, are usually preferable from suffix trees due to their smaller space requirements. The interested reader should refer to Grossi, and Italiano (1993) and Gusfield (1997) to familiarize with the numerous applications of the aforementioned structures.

Dynamic Indices: In many applications, for example, text editors, log files and so forth, the sequences that underlie pattern matching queries change dynamically by insertions or deletions of symbols in arbitrary places. For this setting, we would like to answer the queries as closer as possible to the desirable $O(|P|)$ time without preprocessing the whole sequence from the start and paying $O(|T|)$ computational time. The static structures from the previous subsection cannot achieve that since even a small change can modify all the suffixes and enforce a full reorganization of them. In Gu, Farach, and Pagli (1994), Ferragina (1994), Ferragina, and Grossi (1995a), Sahinalp, and Vishkin (1996), Ferragina and Grossi (1995b), Alstrup, Brodal, and Rauhe (2000), the researchers developed various indices, trying to keep the

update computation cost as closer to the size of the changed (in number of characters) text and the query time closer to the linear bound to the length of pattern. More specifically, the more efficient of these structures are described in Ferragina and Grossi (1995b); Alstrup, Brodal, and Rauhe (2000). Both solutions are quite complex and use a large number of algorithmic techniques from the areas of data structures and stringology. In Ferragina and Grossi, (1995b), an algorithm was presented than needed $O(x+\sqrt{|T|})$ time for inserting/deleting a string of length x in T and $O(|P|+\alpha)$ computational time for retrieving $O(a)$ occurrences. They propose a novel technique called *balanced partition* that keeps the indexed text in equal chunks. Also, they make heavy use of properties of the periods of a string to detect the occurrences of the pattern in the above bounds. In Alstrup, Brodal, and Rauhe (2000), another algorithm was proposed with $O(\log^2|T|\text{loglog}|T|\log^*|T|)$ time for insertions/deletions of characters and $O(|P|+\log|T|\text{loglog}|T|+\alpha)$ time for retrieving $O(a)$ occurrences. They combine a deterministic signature encoding technique originally described in Mehlhorn, K., Sundar R., and Uhrig Christian (1997), and a transformation of the pattern matching problem to the range searching problem to achieve the above bounds.

Another interesting problem in the dynamic setting is when we have a set of patterns that changes dynamically, when inserting or deleting a pattern. The solution of Aho-Corasik automaton is not efficient for this setting because it has to be reconstructed from the start. In Sahinalp and Vishkin (1996), an optimal solution was presented that needed $O(|P|)$ time for inserting/deleting a pattern P from the set and $O(|T|+\alpha)$ time for finding all $O(a)$ occurrences of the patterns in T.

SPACE EFFICIENT IMPLEMENTATIONS COMPRESSION

The design of efficient text compression algorithms is already a very rich area, and most text compression methods can be placed in one of two classes: *symbolwise* methods and *dictionary* methods. Symbolwise methods work by first estimating the probability of appearance of each symbol (character) in the text and then coding each symbol so that frequent characters take codewords with smaller length and rare characters take codewords with larger length. The prominent symbolwise compression algorithms are *Huffman coding* and *arithmetic coding*. On the other hand, dictionary methods achieve compression by replacing groups of consecutive characters by an index in a dictionary. The main such technique is Ziv Lempel coding and its various variants (LZ77, LZ78, LZW). For more details concerning text compression techniques, the interested reader should consult Bell, Cleary, and Witten (1990), Salomon (1992), and Witten, Moffat, and Bell (1999).

COMPRESSED STRING STRUCTURES

Full-text indices, although having a very wide repertoire of applications, fall short in space occupation issues. They are many times larger than inverted files that have sub-linear size. Towards improving the space consumption, there have been proposed structures with succinct representations and self-indexing structures. Self-indexing structures are indices that encapsulate in them the information so there is no need to have aside a copy of data. The first step towards compressed string structures was made by Grossi and Vitter in (2000), where they introduced the *compressed suffix arrays*. Compressed suffix arrays use $\Theta(|T|)$ bits and answer the pattern matching query in $O(|P|/\log|T| + \alpha\log^c|T|)$ time where α is the size of the answer and c is a constant. Sadakane (2003) made compressed suffix arrays self-indexed.

Known compressed string indices, besides the one reported in Grossi and Vitter (2000), are the *opportunistic data structures* and the *LZ-index*. The notion of an opportunistic data structure was introduced in Ferragina and Manzini (2000) (see also Ferragina & Manzini, 2005), it is based on the exploitation of an elegant transform, the Burrows Wheeler transform (Burrows & Wheeler, 1994) and the size of the proposed index depends on the entropy of the text. The LZ-index was proposed in Navarro (2004) and is based on a combination of Ziv-Lempel encoding techniques and tries and occupies $4|T|(1+o(1))$ bits, while answering pattern matching queries in $O(|P|^2\log(|P||\Sigma|)+(|P|+\alpha)\log|T|)$ time, where α is the size of the answer and Σ is the alphabet. Finally, more recent developments can be looked at in Cormode and Muthukrishnan (2005), Ferragina, Giancarlo, Manzini, and Sciortino (2005), and Ferragina, Manzini, Mäkinen, and Navarro (2004).

DATA STRUCTURES IN SECONDARY MEMORY

The most common data structure for handling the dictionary problem on disk is the B-tree (Bayer & McCreight, 1977) and its variants. The B-tree is a tree structure where the degree of each node is required to be $\Theta(B)$ with B being the size of the disk block. A quite powerful variant is the weight balanced B-tree (Arge & Vitter, 1996) that has the property that the weight (the number of nodes in its subtree) of any node at level h is $\Theta(c^h)$, where c denotes a fixed parameter. In a wide range of applications, the weight balanced B-tree can schedule the rebalancings in an amortized way with only

O(1) disk I/Os. The interested reader should refer to Vitter (2001) for a complete presentation of several kinds of data structures on secondary memory including spatial, graph, and hashing data structures.

Suffix trees and suffix arrays, when placed on secondary storage devices, perform very poorly. They need a large number of disk I/Os for their construction and a worst case number of $O(|P|)$ I/Os for answering the pattern matching query. This happens due to the unbalanced nature of the suffix tree, the possibility of large nodes when the alphabet is bigger than the disk block size *B,* and the need to make more access to disk when we cross an edge. The suffix arrays, which perform a binary search upon the array, perform poorly too because of the need for non-local movements.

There are two general methodologies to overcome these difficulties. The first one tries to stuff as much as possible information in main memory in order to perform a draft search in main memory and a more coarse in a small range of information in the secondary memory. This methodology has been followed in Baeza-Yates, Barbosa, and Ziviani (1996). The other one combines the classic B-tree with the plain full-text structures and every internal node is organized like a Patricia tree or a suffix array. Following this approach, Ferragina and Grossi (1999) managed to answer the pattern matching queries in $O(|P|/B+\log_B|T|)$ I/Os with a novel data structure the String *B*-tree.

FUTURE TRENDS

As the handling of large amounts of information come along with all modern applications, the algorithmic area exhibited in this article is going to be stimulated more and grow constructively in the coming years. There is a number of very interesting problems under theoretic and practical perspective to be solved yet. A listing of prospective research problems is the following: (1) more efficient processing of compressed sequences and of compressed text indexing and adaptation of the methods to biological sequences as well as semi structured data and general (hyper) linked data, (2) design of new succinct data structures and implicit data structures, and (3) provision of software for new types of compressed matching and corresponding methods of compression.

CONCLUSION

In this article, we presented the latest developments in the text indexing area by giving special emphasis to string algorithms, secondary memory indices, and space efficient implementations. The areas covered were: data structures and algorithmic techniques for string manipulation, space efficient implementations, and compression techniques. Especially the area of compressed indexing structures is expected to be the central focus for future research as the amount of data to be handled continuously increases, and it is related to the areas of large-scale processing, network communication, and multimedia. This article can hold as a survey for indexing techniques used in text manipulation that can be useful for both academic researchers and practitioners (software developers) that implement indexing structures for various kinds of information systems.

REFERENCES

Aho, A.V., & Corasick, M.J. (1975). Efficient string matching: an aid to bibliographic search. *Communications of the ACM, 18*(6), 333-340.

Alstrup, S., Brodal, G.S., & Rauhe, T. (2000). Pattern matching in dynamic texts. *In Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms* (pp. 819-828).

Anderson, A., & Thorup, M. (2001). Tight(er) worst-case bounds on dynamic searching and priority queues. *Proceedings of the 32nd ACM Symposium on Theory of Computing* (pp. 335-342).

Arge, L., & Vitter, J.S. (1996). Optimal dynamic interval management in external memory. In *Proceedings of the 37th Annual Symposium on Foundations of Computer Science* (pp. 560-569).

Baeza-Yates, R., Barbosa, E., & Ziviani, N., (1996). Hierarchies of indices for text searching. *Information Systems, 21*(6), 497-514.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. ACM Press Addison-Wesley.

Bayer, R., & McCreight, E. (1977). Organization and maintenance of large ordered indices. *Acta Informatica , 1*(3), 173-189.

Bell, T.C., Cleary, J.C., & Witten, I.H. (1990). *Text compression* (pp. 99, 102, 388). Englewood, NJ: Prentice Hall.

Boyer, R.S., & Moore, J.S. (1977). A fast string-searching algorithm. *Communications of the ACM, 20*(10), 62-72.

Burrows, M., & Wheeler, D. J. (1994). *A block sorting lossless data compression algorithm*. Tech. Rep. 124. Digital Equipment Corporation, Palo Alto, CA.

Cormen, T., Leiserson, C., & Rivest, R. (1990). *Introduction to algorithms*. Cambridge, MA: MIT Press.

Cormode, G. & Muthukrishnan, S. (2005). *Substring compression problems*. In ACM-SIAM Symposium on Discrete Algorithms.

Czegh, Z., Havas, G., & Majewski, M. (1997). Fundamental study: Perfect hashing. *Theoretical Computer Science*, 182, 1-143.

Faloutsos, C. (1985). Access methods for text. *ACM Computing Surveys, 17*(1), 49-74.

Farach, M., (1997). Optimal suffix tree construction with large alphabets. In *38th Annual Symposium on the Foundations of Computer Science*, New York (pp. 137-143).

Ferragina, P. (1994). Incremental text editing: A new data structure. In *Proceedings of the 2nd European Symposium on Algorithms* (pp. 495-507).

Ferragina, P., Giancarlo, R., Manzini, G., & Sciortino, M. (2005). Boosting Textual compression in optimal linear time. *Journal of the ACM, 52,* 688-713.

Ferragina, P., & Grossi, R. (1995a). Fast incremental text editing. In *Proceedings of the 6th ACM-SIAM Symposium on Discrete Algorithms* (pp. 531-540).

Ferragina, P., & Grossi, R. (1995b). Optimal online search and sublinear time update in string matching. In *Proceedings of the 36th Annual Symposium on Foundations of Computer Science* (pp. 604-612).

Ferragina, P., & Grossi, R. (1999). The string B-tree: A new data structure for string search in external memory and its applications. *Journal of the ACM, 46*(2), 236-280.

Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science* (pp. 590-598).

Ferragina, P., & Manzini, G. (2005). Indexing compressed text. *Journal of the ACM, 52,* 552-581.

Ferragina, P., Manzini, G., Mäkinen, V., & Navarro, G. (2004). An alphabet friendly FM-index. *Proc. 11th Symposium on String Processing and Information Retrieval (SPIRE '04),* Padova, Italy, Lecture Notes in Computer Science n. 3246 (pp. 150-160). Springer Verlag.

Grossi, R., & Italiano, G. (1993). Suffix trees and their applications in string algorithms. In *1st South American Workshop on String Processing* (pp. 57-76).

Grossi, R., & Vitter, J. (2000). Compressed suffix arrays and suffix trees with applications to text indexing and string matching (extended abstract). In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing* (pp. 397-406).

Gu, M., Farach, M., & Pagli, L. (1994). An efficient algorithm for dynamic text indexing. In *Proceedings of the 5th ACM-SIAM Symposium on Discrete Algorithms* (pp. 697-704).

Gusfield, D. (1997). *Algorithms on strings, trees and sequences*. Cambridge, UK: Cambridge University Press.

Hon, W. K., Sadakane, K., & Sung, W. K. (2003). Breaking a time-and-space barrier in constructing full-text indices. *IEEE FOCS 2003*.

Karkkainen, J., & Sanders, P. (2002). Simple linear work suffix array construction. In *Proceedings of 30th International Colloquium on Automata, Languages and Programming* (pp. 943-955).

Kim, D.K., Sim, J., Park, H., & Park, K. (2003). Linear time construction of suffix arrays. In *Proceedings of the 14th Annual Symposium on Combinatorial Pattern Matching* (pp. 186-199).

Knuth, D.E., Morris, J.H., & Pratt, V.R. (1977). *Fast pattern matching in strings. SIAM Journal of Computing, 6*(2), 323-350.

Ko, P., & Aluru, S. (2003). Space efficient linear time construction of suffix arrays. In *14th Annual Symposium on Combinatorial Pattern Matching* (pp. 200-210).

Manber, U., & Myers, G. (1993). Suffix arrays: A new method for online string searches. *SIAM Journal of Computing, 25*(5), 935-948.

McCreight, E.M. (1976). A space-economical suffix tree construction algorithm. *Journal of the Association for Computing Machinery, 23*(2), 262-272.

Mehlhorn, K. (1984). Data structures and algorithms 1: Sorting and searching. *EATCS Monographs on Theoretical Computer Science*. Springer-Verlang.

Mehlhorn, K., Sundar, R., & Christian, U. (1997). Maintaining dynamic sequences under equality tests in polylogarithmic time. *Algorithmica, 17*(2), 183-198.

Navarro, G. (2004). Indexing text using the ziv-lempel trie. *Journal of Discrete Algorithms, 2*(1), 87-114.

Pagh, P. (2002). *Hashing, randomness and dictionaries*. PhD thesis, University of Aarhus.

Sadakane, K. (2003). New text indexing functionalities of the compressed suffix arrays. *Journal of Algorithms, 48*(2), 294-313.

Sahinalp, S.C., & Vishkin, U. (1996). Efficient approximate results and dynamic matching of patterns using a label paradigm. In *Proceedings of the 37th Annual Symposium on Foundations of Computer Science* (pp. 320-328).

Salomon, D. (1992). *The data compression book.* M&T Books.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. McGraw Hill.

Ukkonen, E. (1995). On-line construction of suffix trees. *Algorithmica, 14*(3), 249-260.

Vitter, J.S. (2001). External memory algorithms and data structures: Dealing with massive data. *ACM Computing Surveys, 33*(2), 209-271.

Weiner, P. (1973). Linear pattern matching algorithms. In *14th IEEE Annual Symposium on Switching and Automata Theory* (pp. 1-11).

Witten, I., Moffat, A., & Bell, T. (1999). *Managing Gigabytes: Compressing and indexing documents and images*. San Francisco: Morgan Kaufmann Publishers, Inc.

Zobel, J., Moffat, A., & Ramamohanarao, K. (1998). Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems, 23*(4), 453-490.

KEY TERMS

BioInformatics: A scientific field that stands at the crossroads of Biology and Informatics, and uses techniques from informatics, computer science, applied mathematics, and statistics to solve biological problems. A synonym term is computational biology, but whereas computational biology deals mainly with the scientific filled bioinformatics is usually used to indicate the infrastructure part.

Compression: The process of encoding information Maintain using fewer information units than a more obvious representation would use, by employing specific encoding schemes that try to exploit the inherent entropy and/or redundancy of the input.

Database: A collection of interrelated persistent information stored and organized as a unit in order to serve a specific purpose and satisfy the demands of a set of users. A database can be considered to be an electronic filling system stored on mass-storage systems such as magnetic tape or disk. A database is one component of a database management system.

Data Structures: A collection of methods for storing and organizing sets of data in order to facilitate access to them. More formally, data structures are concise implementations of abstract data types, where an abstract data type is a set of objects together with a collection of operations on the elements of the set.

Information Retrieval: The scientific discipline that deals with the representation, organization, storage, and maintenance of information objects and in particular textual objects. The representation and organization of the information items should provide the user with easy access to the relevant information and satisfy the user's various information needs.

Secondary Memory Algorithms: Algorithms for handling information on secondary media, like hard disks, CD-ROMs, etc., and try to minimize the number of page accesses in them.

String: A sequence of symbols drawn from an alphabet; differently, it is a series of alphanumeric characters or a series of keywords used to characterize an information object.

World Wide Web: A distributed hypertext-based information system that operates over the Internet and permits the easy access to available information by employing a special software called a "Web browser".

Indicators and Measures of E-Government

Francesco Amoretti
University of Salerno, Italy

Fortunato Musella
University of Naples Federico II, Italy

INTRODUCTION

Although the question of measurement is crucial when defining any concept, little attention has been devoted to a comprehensive view of information and communication technologies (ICTs) applications, spanning qualitative and quantitative assessments.

Due to the lack of a clear definition of e-government, many differences can be noted in the way in which digital policies have been interpreted by academics and practitioners. Coined by the U.S. programme for reinventing government under the Clinton administration (*National Performance Review)*, the term e-government refers to a public sector reorganisation which aims at increasing the efficiency of the public administration and reducing its budget through the use of new technologies. In the words of Douglas Holmes (2001), e-government is "the use of information technology, in particular the Internet, to deliver public services in a much more convenient, customer oriented, cost effective and altogether different and better way. It affects an agency's dealing with citizens, business and other public agencies as well as its internal business processes and employees" (p. 2).

Yet many definitions go beyond the role of e-government in improving the provision of public services. Indeed, the label e-government supports other definitions, not necessarily limited to the computerisation of the public administration (Osborne & Gaebler, 1992). The concept of e-government seems to contain both the redesigning of public services system and a wider transformation of the relationship between private and public actors, so that the restructuring of public administration–influenced by the ideal of a new public management–is combined with the renewal of the democratic decision-making process. Digital policies are presumed to be a key element in improving online service quality and other factors, casting a new role for the citizen-costumer.

At the same time, although e-government is becoming a catch-all concept, from an analytical point of view, official reports produced by international actors show a significant convergence in the way in which this is evaluated and measured. Diffusion of e-government practices are often closely related, and limited, to features of public administration Web sites, with reference to dimensions of openness and interactivity (La Porte, Demchak, & De Jong, 2002). Other studies focus exclusively on how citizens and businesses *perceive* the quality of public e-service, with reference to customer satisfaction, benefits conceived in terms of value and utility of services offered and opportunity of use as strategic factors for performance efficacy and efficiency (Graafland-Essers & Ettedgui, 2003; Stowers, 2004). Only recently a new approach has taken shape, which concentrates more attention on socio-political aspects of the intensive use of new technologies.

BACKGROUND

It can be assumed that measurement is a relevant component of any form of rational decision-making, acting as a mechanism for improving allocative decisions and technical efficiency (Townley, 2005). However, only recently has performance evaluation become a central part of the activity of the public administration.

The increasing interest in the measurement of government activities goes back to the second half of the 20th century, and is strongly related to the growth of public sector expenditure. It was a product of the more general turn to "planning," a key element during the period in question (Boivard, 2005). In the 1980s, the introduction of market principles in public bureaucracies, the process of privatisation, the contracting out of public services and their management though performance contracts gave significant momentum to evaluation.

This move toward performance measurement represented one of the most important issues within the dominant management philosophy, which has sought to modernise the public administration over the last two decades in line with the so-called "new public management" (Schedler & Proeller, 2000). In its attempts to render public services more efficient and to downsize government activities, the resulting wave of public sector reforms has focused greater attention on the outcomes of public activity, emphasising output indicators in reporting administrative performance. Although technical devices have often been presented as a "transparent snapshot of activities," it is clear that they reflect a broader range of political considerations. The importance given to outcomes of government activity has accompanied state restructuring in favour of an increasing market component to public sector delivery.

As part of the "reinventing government" programme, strategic applications of new technologies for the renewal and innovation of public administration has produced several approaches to the evaluation of digital policies. As Jane Fountain (2001) puts it, "applying performance measures is essential to evaluate whether e-government is cost efficient, is serving stakeholders, and is being used effectively by government agencies, staff, citizens, and businesses" (p. 41).

This heightened attention to the question of measurement has led to the conclusion that e-government represents a very complex and multidimensional issue. At a first stage, e-government evaluation concerns the value of ICT investment, that is, the relationship between the costs incurred for the acquisition and introduction of new technologies and the gains made in terms of better management. Indeed, e-government assessments may consider a wide range of variables, including the type of technology, management, as well as organisational and legal issues (Eddowes, 2004; Gupta & Jana, 2003).

Studies on e-government evaluation may present different perspectives: For instance, a clear difference can be noted by comparing *Strategic Value Analysis* method, focusing on organisational issues, and *Cost-Benefit Analysis*, evaluating the impact of public programmes in terms of community well-being. A more articulated idea of performance evaluation is tied to diffusion of the category of governance, which leads to interpret public activity as a negotiated process among private and public actors, rather than a managerial question. Such developments have conducted to look to the social and environmental effects of public action, because it has attempted to measure the consequences of public activity on the whole community of stakeholders rather than on a restricted group of direct service users.

Despite the number of variables effectively involved (Gil-Garcıa & Pardo, 2005), official reports, especially at the international level, seem to encounter considerable difficulties in acknowledging this complex framework of e-government evaluation.

UNPACKING E-GOVERNMENT IN AMERICAN AND EUROPEAN REPORTS

Many efforts have been made to measure the progress and impact of e-government in several countries. The reports produced by international and supranational organisations seem to converge on the following three points:

- The Web site is adopted as the unit of analysis for evaluation of e-government experiences. The way in which information is organised and provided to the citizen-user is considered a key element (Wang, Bretschneider, & Gant, 2005), with specific reference to the possibility of creating interactive systems.
- The path of e-government development seems to follow pre-established stages. Most e-government evaluation is based on a "stages model," a metaphor based on organic growth (Lee, 2007, p. 33). In this deterministic view, countries belonging to different geo-political areas are presumed to be similar in the process of e-government implementation.
- Electronic democracy is not distinct from electronic government, because forms of citizenship involvement are fostered as a direct consequence of the process of administrative restructuring (Mayer-Schönberger & Lazer, 2007). It is often argued that participation will occur when a two-way flow of communication between citizens and political institutions is created. Information provision and transaction processing are used to measure governmental responsiveness to citizens' demands.

Taking such three observations as our point of departure, common elements can be identified in a range of different official documents. A report published by the American Society for Public Administration on behalf of the United Nations, the *Global E-Government Readiness Report* (2005), singles out four goals for electronic government. As outlined in Table 1, the evolutionary relationship between electronic government and electronic democracy is presented through strictly interrelated developmental stages.

The evaluation of digital policies attempts to unify measurement of online services quality and other factors, enabling a participative relationship between citizens and political institutions. An *index of e-readiness* for the performance of public agencies is combined with an *index of e-participation* focusing on citizens' involvement. The participation framework includes several important dimensions: the quantity and quality of information on programmes, budgets, laws and regulation, and other materials dealing with key issues of public interest; e-consultation tools and procedures that encourage citizen participation in public debates; a decision-making e-process to increase citizen input and feedback about specific government decisions.

The findings of a number of European research projects also support a similar approach, stressing the centrality of e-government within the broader family of digital policies, the weak autonomy of e-democracy and the identification of a developmental pattern. Electronic government is considered an essential premise for future experiments in electronic democracy, as shown particularly at the national level. In a study produced by the European University Institute and the University of Geneva (Trechsel, Kies, Mendez, & Schmitter, 2003), focused on the qualitative analysis of 26 recent national reports, e-government infrastructures are viewed as the foundation for future experiments in e-democracy. These reports aim at evaluating the use of ICTs within European

Table 1. E-government development stages according to United Nations' report (2005)

- *Efficient government information management*
- *Better service delivery to citizens*
- *Improved information access and outreach*
- *People empowerment through a participatory decision-making process*

Table 2. E-government developmental goals according to Trechsel et al. (2004)

- *information*
- *bilateral interactivity*
- *multilateral interactivity*
- *user friendliness*

Table 3. E-government development stages according to European Commission (Capgemini, 2006)

- *information* – data necessary to launch the procedure aimed to make public services available online;
- *one-way interaction* – a publicly-accessible Web site providing access (by download) to the paper forms needed to activate the service procedure;
- *two-way interaction* – a publicly-accessible Web site providing access to official electronic forms to obtain certain services to be compiled directly online; and
- *full electronic-case handling* – a publicly-accessible Web site providing full service management within a single Web site, including processing and delivery. No additional paperwork or procedures are necessary.

parliaments and parties, addressing four areas (Trechsel et al., 2003, p. 14). Table 2 presents the e-government goals in such report.

The study concentrates on parliament and party Web sites and on the impact of ICTs on representative institutions, mostly considered in terms of the features of Web sites. Options such as usability and communication approach are considered the main components of a working definition of e-democracy, without addressing the effects of citizens' participation on the decision-making process.

Another authoritative study of electronic government, funded by the European Commission and edited by Capgemini, *Online availability of public services: How is Europe progressing?* (Capgemini, 2006), takes into account a selection of Web sites and provides case studies from all EU member states, evaluating the online services available and their overall structure. There are few references to the political implications of the introduction of new media. The application of new technologies is appraised through a measurement table identifying four developmental stages (Capgemini, 2006, p. 7). Table 3 presents four levels of e-government implementation:

From this point of view, the percentage of public services fully available online is considered a central point, as a consequence of the assumption of the public agencies' Web site as the only–or at least the most important–unit of analysis. Moreover, the meagre attention paid to political variables in assessments of e-government may be the result of the narrow focus of most e-government initiatives on service delivery: "Projects seem to focus on the role of citizens as consumers of public services, and less on the possibilities of ICT to improve the interaction with citizens to develop new policies" (Snijkers, Rotthier, & Janssen, 2007, p. 76).

Such consumer-centricity is also evident in those studies dedicated to understanding the perceptions, attitudes and intentions of citizens in their use of digital services. Despite official reports concentrating on the *provision* rather than *usage* of public services, an important field of the literature on technology adoption regards how digital policies are perceived by users, a question not less important than economic investments in determining the success or failure of e-government initiatives (Gilbert, Balestrini, Kolsaker, & Littleboy, 2007).

FUTURE TRENDS

Future trends are likely to reveal further convergence of e-government methodologies and indicators. A new articulation of evaluation techniques regarding digital policies can also be expected, so that other analytical dimensions may be included in action plans and other official reports. As recent developments suggest, more attention will be devoted to social and political variables, especially those relating to participation.

In the first stage, the overlap between e-government and e-democracy puts emphasis on technological dimensions, frequently using the Web sites of public agencies as a measure of the progress produced by ICT applications (La Porte et al., 2002). More recently, there has been a re-evaluation of political components: Terms like e-rulemaking, e-consultation and e-participation express new conceptions and directions. The potential of new technologies is being investigated not only for boosting the efficiency of the public administration, but also in terms of their effect on democratic regimes. The category of e-governance may be considered an idea that expresses the aim of stimulating the involvement of public and private actors in a networked and horizontal political system, as well as the related and widespread concept of "good governance." This expression has begun to refer to new and more participative processes of coordination made possible or even necessary by the diffusion of online activities.

A report by the OECD (2003a, 2003b) argues that the dilemma of early thinking about e-government was that most Internet enthusiasts did not understand or care very much about political democracy. This is a position echoed too by Chadwick (2007), according to whom democratisation has represented the forgotten promise of e-government. Although online consultations integrating groups based within civil society with bureaucracies and legislatures, internal democratisation of the public sector itself, involvement of users in the design and delivery of public services and diffusion of open-source collaboration in public organisations constitute the main areas of convergence between technological revolution and the objective of improving democratic mechanisms.

During the period of 2000-2003, front office presence of e-government in the form of online service delivery was considered the most important variable for the evaluation of ICT applications. By contrast, more attention is now being paid to back-office strategies and to wider impact of digital policies on society (Albright, 2005; Kunstelij & Vintar, 2004). The "second generation e-government paradigm" furthers the thin objective of achieving improvements in service delivery, increasingly looking toward e-government-as-a-whole concept. In this way, the focus has shifted from the provision of services to the use of ICTs for increasing the value of services, as such a new paradigm "maintains that genuine cost savings and quality improvements will occur only if there is a re-engineering of the internal structures and processes of the administration" (United Nations, 2008, p. 5). As a strategic tool and as an enabler for public service innovation, technology is presumed to lead to a connected or networked governance, which engages the creative efforts of all of society. Indeed, recent trends have moved toward considering the most sophisticated level of online e-govern-

Table 4. Families of e-government indicators

FAMILY 1. *Transparency*

- SUB-FAMILY 1.1. ACCESSIBILITY
- SUB-FAMILY 1.2. PROACTIVITY
- SUB-FAMILY 1.3. ORGANISATION

FAMILY 2. *Citizen Participation*

- SUB-FAMILY 2.1. SERVICES PROMOTING EFFECTIVE CITIZEN PARTICIPATION
- SUB-FAMILY 2.2. CITIZEN SATISFACTION

FAMILY 3. *Culture of Participation*

- SUB-FAMILY 3.1. TRAINING/EDUCATION
- SUB-FAMILY 3.2. PROMOTION

ment initiatives characterized by an integrated back-office infrastructure and more citizen engagement supported and encouraged by governments in the decision-making process (United Nations, 2008, p. 16).

An articulated set of indicators have been used in a significant report produced by *The Conference of European Regional Legislative Assemblies* (2005), composed by the chairman of regional and local parliaments, whose e-democracy working group is in charge of evaluating the experiences of European local and regional institutions in applying new ICTs. The mission of this project is to identify e-democracy practices based on two categories: transparency and participation. Some indicators are intended to capture the quality and nature of service accessibility, along with changes in the internal organisation and administration of regional or local authorities.

Other indicators are intended to highlight the kinds of activities promoting effective citizen participation through digital communication and to measure the level of citizen satisfaction when using services specifically designed and implemented to promote participation. At the same time, as the following table shows, some indicators locate information on education and training activities offered to citizens, as well as issues related to democratic values and ICTs.

A greater awareness of the political implications of ICTs has been achieved as a result of these recent experiences. E-government initiatives preserve a strong link to democratic values such as transparency, inclusion, participation. Although digital experiments have often been accompanied by the rhetoric of democratic participation–depicted as their only source of justification–many difficulties in e-government implementation can be identified. E-government initiatives continue to appear rather immature, especially those regarding provision of services for citizens. Indeed, the most recent benchmarking report on online services provided in Europe (Capgemini, 2006) reveals that the degree of advancement of business-oriented services vastly exceeds those directed toward citizens. The former are included in the "two-way interaction" category in many countries, while the latter remain confined within the "one-way interaction" mode (Capgemini, 2006, p. 9).

Leaving aside these considerations of quality, other differences are noticeable in relation to delivery: business-oriented and citizen-oriented services cover, respectively, two thirds and one third of the overall distribution (Capgemini, 2006, p. 10). Moreover, a recent report by Darrel West (2007) shows that countries vary enormously in their overall e-government performance. Only 28% of the 1,687 Web sites examined around the world present services that are fully executable online. If the political potential of new technologies is often taken for granted, empirical analyses call into doubt the effectiveness of ICT applications for democratic purposes.

CONCLUSION

The analysis of evaluation systems for e-government policy has revealed an overlapping set of indicators for the development of new technologies. American and European reports seem to make reference to a common interpretative platform, intended to blur the borders between e-government and e-democracy. Although high aspirations are tied to new technologies—a remedy for the evils of "poor communication between general public and decision-makers in the political system; a lack of political participation, either caused by structural or functional deficits in the political system; and a negative effect of mass media both on the political system in general and on political participation in particular" (Hagen, 2003)—reports tend to concentrate attention on the modernisation of the administrative machinery through new tools, and on the delivery of public services in a more efficient and customer-oriented approach. In addition to this, they mostly focus on the technical features of Web sites, limiting themselves to the electronic façade of online services, with little or no reference to redesigning back-office administrative procedures or wider political implications. The evaluation of e-government mainly refers to output indicators, as initiatives are typically ranked on the basis of the amount of services that are brought online.

The same vision is embraced by a report produced by the World Bank for the Information for Development Program (infoDev), under a very meaningful title, *The Handbook of E-government* (World Bank, 2002). This document is an implementation guide, a practical tool underlining key resources and best practices when establishing e-government plans. The democratisation process is presumed to conform to a similar path in countries with very different traditions and history. The first stage involves the use of new technologies in order "to expand government information access" (World Bank, 2002, p. 3). If the public administration of modern states produce a huge amount of documents, data management possibilities enable the distribution of "government information to an audience as wide as possible" (World Bank, 2002, p. 3). The second stage involves a larger impact of technologies on the quality of democracy, by activating greater citizen involvement and interaction with political institutions. A two-way information flow enables citizens to receive answers to their questions directly from policy makers. Finally, in the third phase of digital policy development, known as the "transaction stage," citizens can access public services directly online. Once again, it is difficult here to differentiate between e-government and e-democracy, because the first is defined as a tool to implement better government.

More recently, the association between objectives of participative governance and the restructuring of administrative design has been confirmed. As the last United Nations *Global*

E-Government Survey (2008) put it, in the movement from e-government to connected governance, "efforts are aimed at an improved cooperation between governmental agencies, allowing for an enhanced, active and effective consultation and engagement with citizens, and a greater involvement with multi-stakeholders regionally and internationally." Thus, new developments of e-government paradigm show more consciousness on socio-political aspects of the introduction of information technologies.

However, the central role of technical variables in methodologies used to measure e-government can be underlined, as well as the consequent fading of the distinction between e-government and e-democracy. The World Wide Web is becoming a vehicle for the evaluation of the openness of public organisations, so that measuring the diffusion and quality of modern networked information technologies is often considered a way to measure the spread of administrative arrangements that are vital to emerging forms of democratic governance.

REFERENCES

Albright, K. S. (2005). Global measures of development and the information society. *New Library World*, *106*(7-8), 320-331.

Bovaird, T. (2005). Public sector performance. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. III). Oxford: Elsevier.

Capgemini. (2006). *Online availability of public services: How is Europe progressing? Web-based survey on electronic public services.* European Commission. Bruxelles.

Chadwick, A. (2003). Bringing e-democracy back in. What it matters for future research on e-governance. *Social Science Computer Review*, *21*(4), 443-455.

Chadwick, A. (2007). Digital network repertoires and organizational hybridity. *Political Communication*, *24*, 283-301.

Conference of European regional legislative assemblies. (2005). *Catalogue of indicators*. In e-democracy project for Carle, version 9.2. Brussels.

Eddowes, L. A. (2004). The application of methodologies in e-government. *Electronic Journal of e-Government*, *2*(2), 115-126.

Fountain, J. E. (2001). *Building the virtual state: Information technology and institutional change*. Washington, DC: Brookings Institution.

Graafland-Essers, I., & Ettedgui, E. (2003). *Benchmarking e-government in Europe and the U.S.* Rand Europe.

Gil-Garcıa, J. R., & Pardo, T.A. (2005). E-government success factors: Mapping practical tools to theoretical foundations. *Government Information Quarterly*, *22*, 187-216.

Gilbert, D., Balestrini, P., Kolsaker, A., & Littleboy, D. (2007). *Citizen adoption of e-government in the UK: Perceived benefits and barriers*. In D. Griffin, P. Trevorrow, & E. Halpin (Eds.), *Developments in e-government: A critical analysis.* Amsterdam: IOS Press.

Gupta, M. P., & Jana, D. (2003). E-government evaluation: A framework and case study. *Government Information Quarterly*, *20*, 365-387.

Hagen, M. (2003). *A typology of electronic democracy*. Retrieved May 31, 2008, from

http://www.uni-giessen.de/fb03/vinci/labore/netz/hag_en.htm

Holmes, D. (2001). Egov: *Ebusiness strategies for government*. London: Nicholas Brealey.

Kunstelij, M., & Vintar, M. (2004). Evaluating the progress of e-government development: Critical analysis of current approaches. In *Paper presented at the EGPA Annual Conference*, Ljubljana.

La Porte, T. M., Demchak, C., & De Jong, M. (2002). Democracy and bureaucracy in the age of the Web: Empirical findings and theoretical speculations. *Administration & Society*, *34*(4), 411-446.

Lee, J. (2007). Search for stage theory in e-government development. In D. Griffin, P. Trevorrow, & E. Halpin (Eds.), *Developments in e-government: A critical analysis.* Amsterdam: IOS Press.

Macintosh, A. (2006). eParticipation in policy-making: The research and the challenges. *Exploiting the knowledge economy: Issues, applications, case studies*. Amsterdam: IOS Press.

Mayer-Schönberger, V., & Lazer, D. (2007). *Governance and information technology. From electronic government to information government*. Cambridge, MA: MIT Press.

OECD. (2003a). *The e-government imperative.* Paris.

OECD. (2003b). *Promises and problems of e-democracy. Challenges of online citizen engagement.* Paris.

Osborne, D., & Gaebler, T. (1992). *Reinventing government—how the entrepreneurial spirit is transforming the public sector.* Reading, MA: Addison-Wesley.

Sanders, L. (1997). Against deliberation. *Political Theory*, *5*(25), 347-377.

Schedler, K., & Proeller, I. (2000). *New public management*. Bern, Stuttgart, Wien: Haupt.

Snijkers, K., Rotthier, S., & Janssen, D. (2007). Critical review of e-government benchmarking studies. In D. Griffin, P. Trevorrow, & E. Halpin (Eds.), *Developments in e-government: A critical analysis.* Amsterdam: IOS Press.

Stowers, G. N. (2004). *Measuring the performance of e-government*. Washington, DC: IBM Center for The Business of Government, E-government series.

Trechsel, A., Kies, R., Mendez, F., & Schmitter, P. (2004). *Evaluation of the use of new technologies in order to facilitate democracy in Europe*: *E-democratizing the parliaments and parties of Europe.* European University Institute and University of Geneva.

Townley, B. (2005). Critical views of performance measurement. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. I). Oxford: Elsevier.

United Nations. (2005). *Global e-government readiness report 2005: From e-government to e-inclusion*. New York.

United Nations. (2008). *Global e-government survey: From e-government to connected governance*. New York.

Wang, L., Bretschneider, S., & Gant, J. (2005). Evaluating Web-based e-government services with a citizen-centric approach. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, (Vol. 5, p. 129).

West, D. M. (2000). *Assessing e-government: The Internet, democracy, and service delivery by state and federal government*. The Genesis Institute. Retrieved May 31, 2008, from http://www.insidepolitics.org/egovtreport00.html

West, D. (2007). *Global e-government*. Providence, RI: Brown University.

World Bank. (2002). *The handbook of e-government: The information for development program*. infoDev.

KEY TERMS

Benchmarking: An interactive process by which the activities of an actor are evaluated in relation to best practices.

E-Consultation: The use of the Internet to disseminate the developments in a policy field to the wider public, experts and interest groups, and to invite them to respond.

E-Democracy: The use of the ICTs for increasing the transparency of the political process, for enhancing the direct involvement and participation of citizens and for improving the quality of opinion formation by opening new spaces of information and deliberation.

E-Rulemaking: A "notice and comment" method, following three steps: announcement, comment and publication. The agency publishes a notice containing a proposed law and the interested public is invited to send comments and proposals via e-mail during a fixed time period, so that the agency can analyse and consider these comments in its final version of the law.

E-Governance: A concept and emerging practice, seeking to realise processes and structures for harnessing the potentialities of information and communication technologies at various levels of government and the public sector and beyond, for the purposes of enhancing good governance.

E-Participation: The use of information and communication technologies to broaden and deepen political participation by enabling citizens to connect with one another and with their elected representatives.

Good Governance: The processes and structures that guide political and socio-economic relationships, with particular reference to a set of eight major characteristics: participation, consensus, accountability, transparency, responsiveness, effectiveness and efficiency, inclusiveness and the rule of law.

Governmental Openness: A governmental strategy leading citizens to play a stronger role in interacting with government and in controlling its activities, making decision-making more transparent. It may also be considered a measure of government accountability, in that government agencies can be continuously assessed by citizens through everyday interactions.

Monitoring: The systematic collection of information to provide indications on how a programme or service is performing.

ICTs: Acronym for information and communications technologies. It is a general term that describes any technology that helps to produce, manipulate, store, communicate, or disseminate information.

Performance Measures: Benchmarks which indicate the economy, efficiency and effectiveness of a current or past activity, unit or organisation.

Category: Bioinformatics

Individual-Based Modeling of Bacterial Genetic Elements

Venetia A. Saunders
Liverpool John Moores University, UK

Richard Gregory
University of Liverpool, UK

Jon R. Saunders
University of Liverpool, UK

INTRODUCTION

Individual-based computational modeling of biological systems is an important complement to experimental research. The individual-based model (IbM) is a bottom-up approach that considers the fate of individuals, their properties and interactions, and the influence of these interactions, holistically, on properties of the system. This contrasts with population-based models dependent on averaged behaviour of the whole system (DeAngelis & Gross, 1992; Huston, DeAngelis, & Post, 1988). IbMs can track individuals in time so that unusual events can be captured. They are particularly suited to biological simulations, where individuals might represent virtual plants, animals, or microorganisms in differing ecosystems. Lower complexity, coupled with the wealth of genetic knowledge about bacteria, allow for more realistic simulations compared with higher organisms. Accordingly, a lineage of IbMs, including Bacteria Simulator (BacSim) (Kreft, Booth, & Wimpenny, 1998; Kreft, Picioreanu, Wimpenny, & van Loosdrecht, 2001), INDividual DIScrete SIMulation (INDISIM) (Ginovart, Lopez, & Gras, 2005; Ginovart, Lopez, & Valls, 2002; Prats, Lopez, Giro, Ferrer, & Valls, 2006), COmputing Systems of Microbial Interactions and Communications (COSMIC) (Gregory, Paton, Saunders, & Wu, 2004; Paton, Gregory, Vlachos, Saunders, & Wu, 2004), RUle-based BActerial Modeling (RUBAM) (Paton, Vlachos, Wu, & Saunders, 2006; Vlachos, Paton, Saunders, & Wu, 2006) and COSMIC-Rules (Gregory, Saunders, & Saunders, 2006, 2008b), based on COSMIC and RUBAM, has been developed for bacterial simulations.

Although all these models are individual-based, underlying simulation mechanisms and aims vary. BacSim was the first to use IbM in a recognizable biological context (Kreft *et al*., 1998, 2001) aiming to model growth and cell division, quantitatively, at the population level, using a pseudocontinuous 2-dimensional world with restricted nutrients. INDISIM is based on stronger mathematical foundations, and is a discrete space and time stochastic simulation of colony growth, largely based on random variables (Ginovart *et al*., 2002). Each cell is a set of parameters existing at a discrete location. COSMIC uses pseudocontinuous space and discrete time to model evolution of cells (Gregory *et al*., 2004). Each cell contains a bit string genome that interacts with itself and the environment. This model is largely deterministic, although random events do have a role. It can run in a parallel machine, though any random effects this creates have been removed. RUBAM is a simplification of COSMIC, with pseudocontinuous space, discrete time, and a much more simplified genome. It aims to model adaptation (Vlachos *et al*., 2006). The simplified genome allows for comparatively rapid simulations that show adaptation and acquired resistance to antibiotics. COSMIC-Rules is a culmination of IbM modeling design, having an effective balance of modeling detail while being computationally tractable (Gregory *et al*., 2006, 2008b). Like COSMIC, it is a parallel simulation with pseudocontinuous space and discrete time. It uses a genome abstraction to represent the conditions and outputs of complex biochemical pathways, while incorporating an element of specificity and means of simulating evolution. Like the other IbMs considered here, each individual has its own parameters and state. Unlike the other IbMs, the scope of COSMIC-Rules covers vertical and horizontal gene transfer using populations of millions of cells.

BACKGROUND

IbMs describe behaviour in a system, acknowledging the uniqueness of the individual, its characteristics and interactions with other individuals. Individuals are only considered together as a population or community when analysed. The majority of IbM approaches to bacterial simulations have focussed on growth and metabolism from ecological and evolutionary perspectives. However, COSMIC-Rules (Gregory *et al*., 2006, 2008b), the IbM described here, has been designed to simulate genetic interactions in bacteria

within a framework that allows adaptation and evolutionary processes to be observed. An advantage of the IbM in these simulations is that bacterial evolution becomes open-ended: emergence, growth, and death of individual bacteria, their interactions with other bacteria, and any infection events can be monitored over time. Such an approach permits questions about impact of individual variability on adaptive evolution to be addressed. Genetic elements, such as plasmids and viruses, can spread within bacterial populations mediating genetic exchange (Sørensen, Bailey, Hansen, Kroer, & Wuertz, 2005) and, coupled with mutation, provide raw materials for adaptive evolution (Marri, Hao, & Golding, 2007). By acquiring new or mutated genes, bacteria can adapt and survive in changing environments. Some plasmids are conjugative (self-transmissible), transferring by the horizontal gene transfer process of conjugation. This requires cell-to-cell contact involving a conjugation ligand, encoded by the donor, and a receptor on the recipient (Manning & Achtman, 1979).

Bacteriophages (phages) are viruses that infect bacteria (Carter & Saunders, 2007). The infective cycle is initiated by attachment of phages to susceptible bacteria through specificity of phage ligand-host cell receptor interactions. Phages may be temperate or virulent. A temperate phage is capable of operating in lytic (host killing) or lysogenic (without harming the host) modes. Phage replication occurs in the lytic cycle, culminating in cell lysis and release of progeny phages. For lysogeny, an inactive phage genome is stably inherited; there is neither bacterial lysis nor production of progeny phages. Phages that lysogenize the host (lysogen) confer immunity to superinfecting, homologous phages, and may effect changes to host properties by lysogenic conversion (Brussow, Canchaya, & Hardt, 2004). Furthermore, lysogeny promotes adaptation to survival in poor/unstable environments where resources are limited, as well as providing a potential reservoir of progeny phages.

MODELING GENETIC ELEMENTS IN BACTERIA

The Model: COSMIC-Rules

COSMIC-Rules (Gregory *et al*., 2006, 2008b) incorporates three levels using IbM philosophy: the genome, the cell, and an environment populated by such cells. Organisms possess individually defined physical locations, size, cell division status, and genomes including extrachromosomal elements (*e.g.*, plasmids and phages). The virtual environment consists of multiple, individual substances (substrates and toxic agents *e.g.*, antibiotics) whose relative nutrient status and/or toxicity is specified by the make-up of particular bacterial genomes. Individuals have specific, mutable genotypes and phenotypes evolving in a medium of initially defined, though changeable, composition. The environment is a 3-dimensional space, with the third dimension being of one cell diameter, so that cells effectively move in two dimensions.

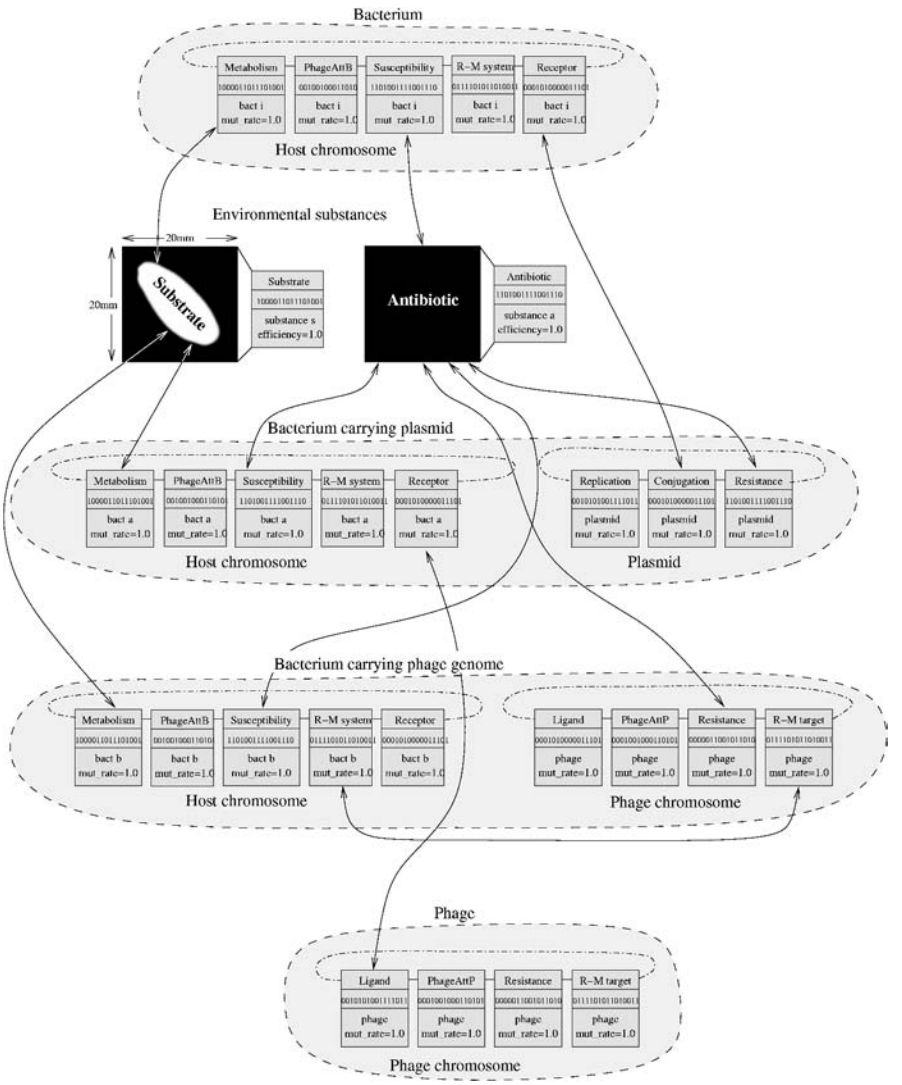
Central to the model are novel features for representing genotypes and phenotypes in a compressed manner (genome compression). Each gene/gene set is represented by a unique tagged bit string that defines coding capacity and mediates specific genotypic and phenotypic interactions through bit string matching (Figure 1). This allows modeling of ligand-receptor interactions required, for example, for bacterial cell contact and conjugation or phage infection, metabolism-substance (substrate) interactions for cell growth, and susceptibility-substance (antibiotic/toxic agent) interactions for antibiotic action and cell death. If there is interaction with a resistance tag, probability of death is reduced to zero. A successful outcome demands that tagged bit strings match, and matching depends on them being no more than two bits different. The degree of similarity for matching varies with events under consideration. There are multiple metabolism, susceptibility, and resistance tags for each individual, together with multiple substances, and the best match is used in each case. "Best" refers to an outcome that produces either highest growth rate or highest probability of cell death. Collectively, bit strings form the genome and only genes directly involved in a particular simulation are considered; other required functions are assumed to be provided by covert gene sets to reduce computational load. However, flexibility of the model allows additional bit strings to be incorporated to increase genome complexity as necessary.

COSMIC-Rules models simplifications of real-world situations, aiming to reproduce biologically realistic conditions, by applying a series of rules, informed by physical laws and principles of bacterial genetics. Simulation parameters reflect biological values, as determined experimentally. Rules govern behavior of individuals in simulations and are varied for different scenarios.

The model is built for parallel execution utilizing a development cluster and scales to large HPC systems. To achieve parallelism, the environment is partitioned into demes, each containing individuals that move and interact with other individuals, and for one time step, the deme is isolated from other demes. Individuals then move and may migrate to another deme. Isolating demes in this way facilitates computability, since cell-to-cell synchronization between demes is not required.

The simulation treats each bacterium or free phage as an object instance, with its own associated parameters and genome. Genomes are also handled as objects. Within each genome are tagged bit strings representing compressed genes that make this approach tractable. Individual bacteria or free phages are subject to mutation, and soon each individual has its own unique bit strings associated with each tag. Once parallel synchronization has been achieved, a typical cycle

Figure 1. Genomes of individuals and their interactions. Compressed genomes comprise keywords plus bit string pairs that specify overall function (consequences of a successful interaction). Examples of interactions are shown by connecting arrows. mut_rate, mutation rate associated with bit strings; R-M system, restriction-modification system; R-M target, site for action of restriction-modification system; PhageAttB, phage attachment site on bacterial genome; PhageAttP, attachment site on phage genome.



of the simulation involves randomly iterating over all tags in all individual bacteria and free phages. For each matching tag, a rule is "triggered" (see Figure 1 for instances of matching).

Triggered rules invoke some combination of time delay and probabilistic or deterministic action, the exact outcome depending on the tag involved, the bit string, and possibly position of individuals in the environment. These rules implement, for example, substrate uptake, antibiotic-induced death, plasmid transfer, phage infection, and cellular decay, some involving several rules triggered in succession.

Case Studies

COSMIC-Rules has been validated by modeling case studies including antibiotic action and resistance in bacterial populations (Gregory *et al*., 2008b), conjugative plasmid transfer (Gregory *et al*., 2008a), and phage infection. The simulations illustrate adaptation and survival of bacteria in changing environments through genetic events (mutation and gene transfer) that alter phenotype and can confer selection benefits on the host.

A. Emergence of Antibiotic Resistance in Bacteria

The simulation in Figure 2 demonstrates the action of a bactericidal antibiotic that kills sensitive bacteria. A few resistant mutants survive after exposure to the antibiotic. Growth of mutants is promoted by nutrients released from dead cells. Thus, the model incorporates the influence of the organism on the environment and the reverse.

B. Plasmid Transfer in Bacteria

Plasmids are responsible for a large proportion of resistance to antibiotics, and their transmission through bacterial populations renders individual bacteria resistant. Figure 3 simulates spread of a virtual antibiotic resistance (R) plasmid through an antibiotic-sensitive bacterial population.

A substrate is provided for growth and an antibiotic that overlaps part of the substrate zone is present. Bacteria grow around the antibiotic zone, but do not survive in it, so the initial level of substrate is maintained here. When R plasmid-

Figure 2. Simulation of antibiotic action and emergence of antibiotic resistance in a bacterial population. (a) Snapshots of key moments in the simulation. Panels 1-7 show growth of the antibiotic-sensitive bacteria. A substrate (dashed oval area, panel 1) supports the population. Addition of antibiotic kills most cells. Panel 8 shows growth of the few antibiotic-resistant mutants. (b) Graphical representation of simulation in (a). Antibiotic-resistant mutants (whose "susceptibility" bit string has mutated sufficiently to avoid death) survive, supported by nutrients released from dead cells.

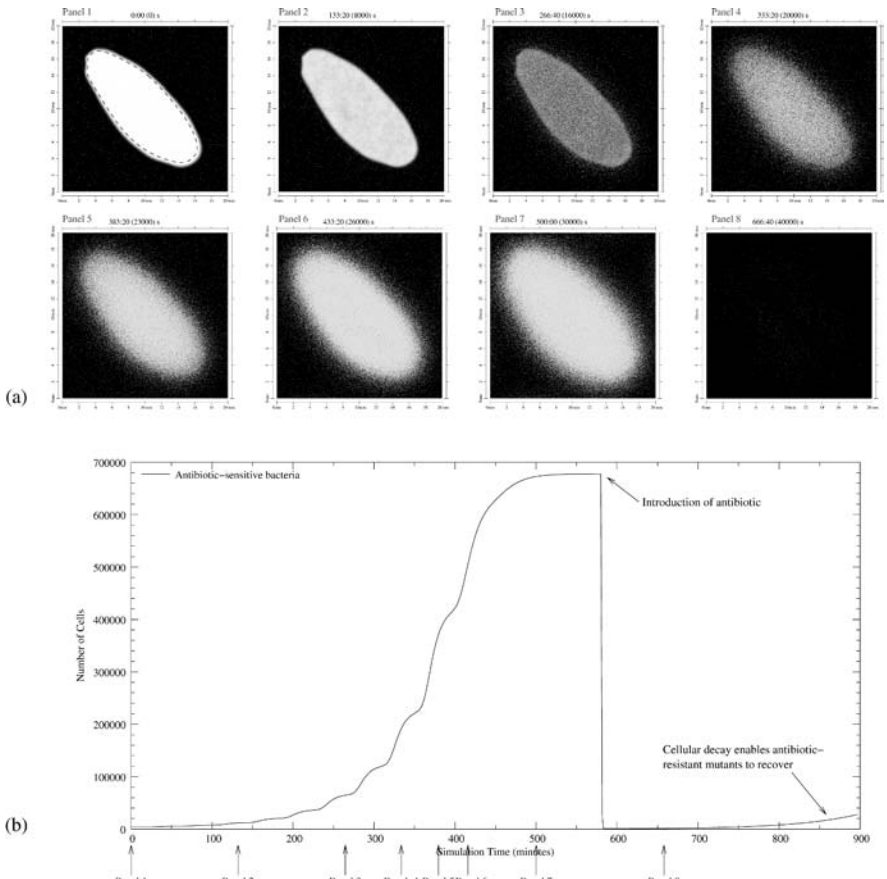
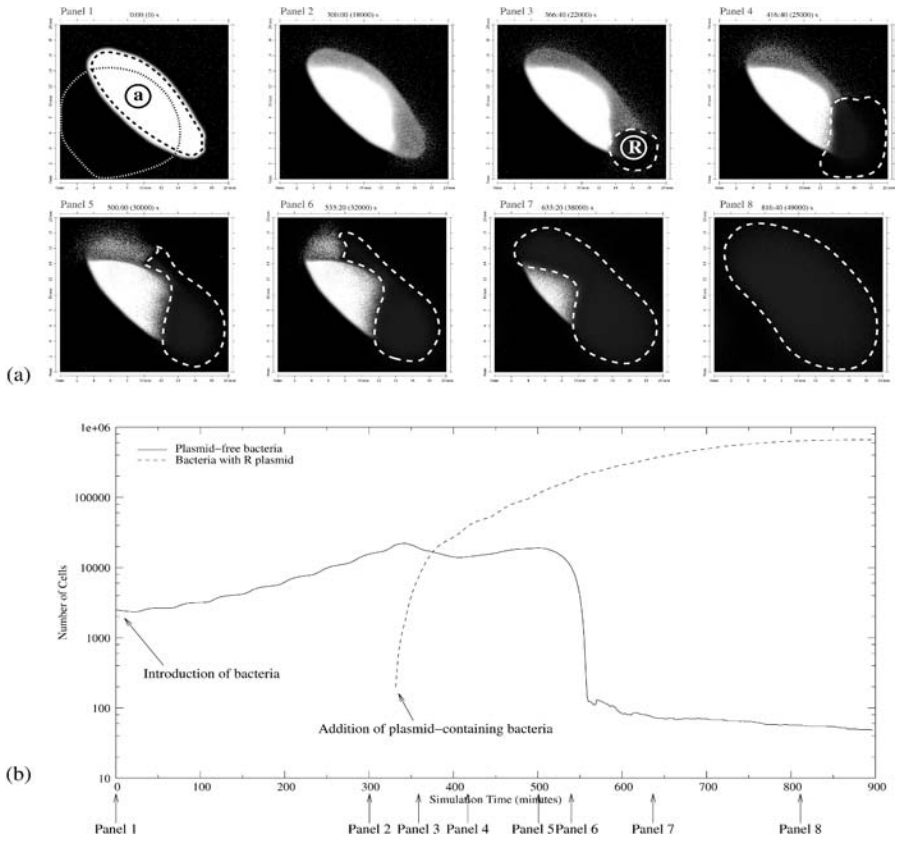


Figure 3. Simulation of plasmid spread in a bacterial population. (a) Snapshots of key moments in simulation using an antibiotic to monitor plasmid transfer. Panel 1 shows an oval area of substrate (dashed outline) and a circular area of antibiotic (dotted outline); (a) is the region of overlap of substrate and antibiotic. Sensitive bacteria only grow in the area with substrate and no antibiotic (panel 2). Plasmid-containing bacteria are added at "R" (white dashed outline, panel 3). Panels 4-8 show spread of the resistance plasmid. (b) Graphical representation of simulation in (a).



containing bacteria, resistant to the antibiotic, are added, the plasmid transfers to plasmid-free recipients, rendering them resistant. Such recipients are now able to colonize the zone of overlap of substrate and antibiotic. With further plasmid transfer there is rapid spread of antibiotic-resistance throughout the population. This mimics the emergence and spread of antibiotic-resistance, for example, amongst clinically important bacteria.

C. Phage Infection

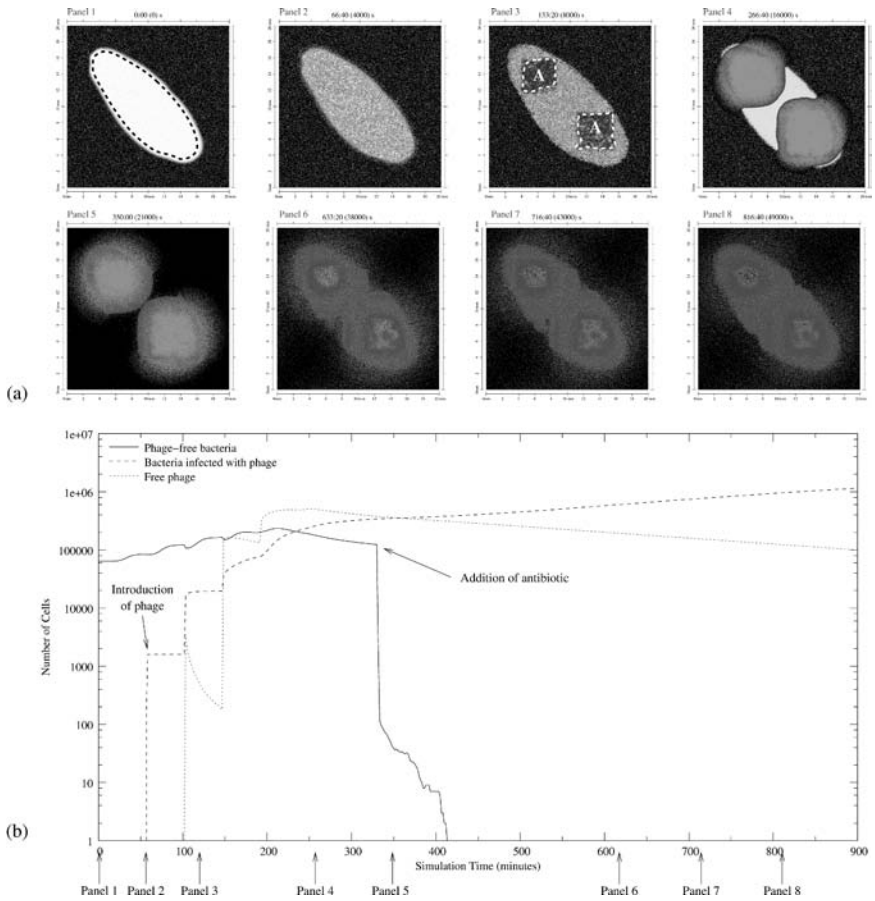
Phages, which are ubiquitous in nature, have a crucial role in controlling abundance of bacteria by infecting and killing susceptible cells. Figure 4 describes the spread of a virtual temperate phage through a susceptible population. The phage genome carries a gene, encoding resistance to an antibiotic to facilitate selection of lysogens, upon exposure to antibiotic. Addition of phages results in infection of phage-free cells. Phages either enter the lytic cycle, killing host cells and releasing progeny phages, or establish lysogeny and confer an antibiotic-resistance phenotype on resultant lysogens. Such lysogens, which exhibit superinfection immunity, continue to grow, despite the presence of free phages. Addition of antibiotic kills uninfected bacteria, with the exception of a few resistant mutants, but lysogens survive due to the antibiotic-resistance gene. Growth of surviving cells is encouraged by nutrients released from dead bacteria. Free phages, having a finite half-life, gradually decline through lack of susceptible host cells. This is reflected in natural populations that fre-

quently comprise bacteria that are multiply lysogenized by different phages. Moreover, the genomes of extant bacteria contain evidence of extensive lysogenization during bacterial evolution (Brussow *et al*., 2004; Williamson, Radosevich, Smith, & Wommack, 2007).

FUTURE TRENDS

COSMIC-Rules has been used here to examine some simple case studies involving virtual transmission of plasmids and phages, and demonstrates the applicability of the model to studying the role of genetic elements in bacterial populations. It is a highly flexible model, and provides a springboard for generating more complex simulations involving multiple genetic transactions by, for example, combining plasmids and phages in bacteria subject to natural and anthropogenic environmental change. Extending the model, using larger scale Grid technology, should allow simulations more representative of mixed populations in natural environments. Accordingly, the model should inform studies on bacterial adaptation and evolution, including gene transfer and spread of antibiotic-resistance in bacteria, and behaviour of pathogenic bacteria and their viruses. It is also applicable

Figure 4. Simulation of phage infection of a susceptible bacterial population. (a) Snapshots of key moments in simulation of phage infecting bacteria. Panel 1 shows an oval area of substrate (dashed outline) supporting growth of the bacterial population. The population increases as substrate is consumed. Phages are introduced (at "A," panel 3). Panel 4 shows phages spreading through uninfected bacteria: the lighter areas represent phage-free bacteria, the darker areas phage-infected bacteria/free phages. Addition of antibiotic kills uninfected bacteria (antibiotic-sensitive), except for a few resistant mutants. Panels 6-8 show phage-infected bacteria that express antibiotic resistance and free phages. Surviving bacteria grow on nutrients released from dead bacteria. (b) Graphical representation of simulation in (a).



to modeling the epidemiology of infectious diseases. More widely, the approach would have applications in other areas of information science and technology, where an evolutionary dimension is involved.

CONCLUSION

COSMIC-Rules has been developed and validated using scenarios involving virtual bacteria, their plasmids, and phages in environments varying between supportive (with added nutrients) and inhibitory (with antimicrobial agents). The simulations reinforce our basic notion (Gregory *et al*., 2008b) that compressing the representation of the genome is justified and retains biological realism. The model has demonstrated the utility of IbM philosophy to simulating plasmid transfer and phage infection in bacteria. A rule-based approach thus provides a valuable tool for predicting behaviour of bacterial populations in response to past and future episodes of environmental change.

REFERENCES

Brussow, H., Canchaya, C., & Hardt, W-D. (2004). Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews*, *68*(3), 1092-2172.

Carter, J. B., & Saunders, V.A. (2007). *Virology. Principles and applications*. UK: John Wiley and Sons Ltd.

DeAngelis, D. L., & Gross, L. J. (1992). *Individual-based models and approaches in Ecology: Populations, communities and ecosystems*. New York: Chapman and Hall.

Ginovart, M., Lopez, D., & Gras, A. (2005). Individual-based modelling of microbial activity to study mineralization of C and N and nitrification process in soil. *Nonlinear Analysis: Real World Applications*, *6*(4), 773-795.

Ginovart, M., Lopez, D., & Valls, J. (2002). INDISIM, an individual-based discrete simulation model to study bacterial cultures. *Journal of Theoretical Biology*, *214*(2), 305-319.

Gregory, R., Paton, R. C., Saunders J. R., & Wu, Q. H. (2004). A model of bacterial adaptability based on multiple scales of interaction. In R. Paton, H. Bolouri, M. Holcombe, J. H. Parish, & R. Tateson (Eds.), *Computation in cells and tissues: Perspectives and tools of thought, Series in Natural Computing* (pp. 131-158). Heidelberg: Springer.

Gregory, R., Saunders, J. R., & Saunders V. A. (2006). The Paton individual-based model legacy. *Biosystems*, *85*(1), 46-54.

Gregory, R., Saunders, J. R., & Saunders, V. A. (2008a). Rule-based modelling of conjugative plasmid transfer and incompatibility. *Biosystems*, *91*(1), 201-215.

Gregory, R., Saunders, V. A., & Saunders, J. R. (2008b). Rule-based system for microbial interactions and communications: Evolution in virtual bacterial populations. *Biosystems*, *91*(1), 216-230.

Huston, M. A., DeAngelis, D. L., & Post, W. (1988). New computer models unify ecological theory. *BioScience*, *38*(1), 682-691.

Kreft, J. U., Booth G., & Wimpenny, J. W. T. (1998). BacSim, a simulator for individual-based modelling of bacterial colony growth. *Microbiology*, *144*, 3275-3287.

Kreft, J. U., Picioreanu, C., Wimpenny, J. W. T., & van Loosdrecht, M. C. M. (2001). Individual-based modelling of biofilms. *Microbiology*, *147*, 2897-2912.

Manning, P. A., & Achtman, M. (1979). Cell-to-cell interactions in conjugating Escherichia coli: The involvement of the cell envelope. In M. Inouye (Ed.), *Bacterial outer membranes: Biogenesis and functions* (pp. 409-447). New York: John Wiley and Sons Inc.

Marri, P. R., Hao, W., & Golding, G. B. (2007). The role of laterally transferred genes in adaptive evolution. *BMC Evolutionary Biology*, *7* (suppl.1) S8.

Paton, R., Gregory, R., Vlachos, C., Saunders, J., & Wu, H. (2004). Evolvable social agents for bacterial systems modelling. *IEEE Transactions on Nanobioscience*, *3*(3), 208–216.

Paton, R. C., Vlachos, C., Wu, Q. H., & Saunders, J. R. (2006). Simulated bacterially inspired problem solving—the behavioural domain. *Natural Computing*, *5*(1), 43–65.

Prats, C., Lopez, D., Giro, A., Ferrer, J., & Valls, J. (2006). Individual-based modelling of bacterial cultures to study the microscopic causes of the lag phase. *Journal of Theoretical Biology*, *241*(4), 939-953.

Sørensen, S. J., Bailey, M., Hansen, L. H., Kroer, N., & Wuertz, S. (2005). Studying plasmid horizontal transfer in situ: A critical review. *Nature Reviews*. *Microbiology*, *3*(9), 700-710.

Vlachos, C., Paton, R. C., Saunders J. R., & Wu, Q. H. (2006). A rule-based approach to the modelling of bacterial ecosystems. *BioSystems*, *84*(1), 49-72.

Williamson, K. E., Radosevich, M., Smith, D. W., & Wommack, K. E. (2007). Incidence of lysogeny within temperate and extreme soil environments. *Environmental Microbiology*, *9*(10), 2563-2574.

KEY TERMS

Antibiotic: A natural or artificial substance that can kill or inhibit growth of microorganisms.

Bacterial Conjugation: A naturally occurring horizontal gene transfer process in which donor and recipient cells come into direct contact for exchange of genetic material in bacteria.

Bacterial Plasmid: A self-replicating, extrachromosomal genetic element found in bacteria. Plasmids may carry genes for various functions, including antibiotic resistance and virulence.

Bacteriophage (Phage): A virus that specifically infects bacteria. Broadly, there are two types: virulent (or lytic) and temperate (or lysogenic). Upon infection, a virulent phage replicates and releases progeny phages. A temperate phage may enter the lytic cycle or lysogenize the host, with the phage genome remaining dormant. The genome may later become active, directing the synthesis of progeny phages and destruction of the host.

Bit String Matching: Method by which bits in a tagged bit string are compared with bits in another tagged bit string to determine if they are acceptably similar. COSMIC-Rules uses the fast exclusive-or operation followed by counting the number of set bits.

COSMIC-Rules: Model and modeling framework simulating bacterial adaptation using IbM philosophy and genome compression to achieve realistic and qualitatively accurate simulations of, for example, substance affinity, plasmid transfer, phage spread, and cellular decay.

Genome Compression: An abstraction that reduces the complexity of pathways into single components. Each "gene" can represent an otherwise intractable pathway if its external overall effect, behaviour, or input/output relationship can be characterised by probability, mathematical formula, lookup table, logical expression, or a combination of any of these methods.

Individual-Based Model (IbM): Modeling philosophy in which numerical quantities, representing the size of some population, are replaced by individuals that make up the population. Each individual would have its own state, allowing analysis to include both population level crowd effects and individualism. The form of the individual is inherently open-ended and need not be tied to mathematical expressions. The IbM philosophy also allows use of nested levels of individuality.

Tagged Bit String: Consists of both a type and a bit string. The type specifies what pathway it abstracts, and what other tags must be present for this pathway abstraction to be active. A bit string provides specificity by adding another (mutable) condition to activation of a pathway. Both tag and bit string exist as an inseparable pair and are passed through both vertical and horizontal inheritance.

Influential Agile Software Parameters

Subhas C. Misra
Carleton University, Canada

Vinod Kumar
Carleton University, Canada

Uma Kumar
Carleton University, Canada

INTRODUCTION

Successful software systems development is a delicate balance among several distinct factors (Jalote, 2002) such as enabling people to grow professionally; documenting processes representing the gained experiences and knowledge of the organization members; using *know how* to apply the suitable processes to similar circumstances; and refining processes based on achieved experience.

Software projects have two main dimensions: engineering and project management. The engineering dimension concerns the construction of a system, and focuses mainly on issues such as *how to* build a system. The project management dimension is in charge with properly planning and controlling the engineering activities to meet project goals for optimal cost, schedule, and quality.

For a project, the engineering processes specify how to perform activities such as requirement specification, design, testing, and so on. The project management processes, on the other hand, specify how to set milestones, organize personnel, manage risks, monitor progress, and so on (Jalote, 2002).

A software process may be defined as "a set of activities, methods, practices, and transformations that people use to develop and maintain software, and the associated products and artifacts."[1] This is pictorially depicted in Figure 1 (Donaldson & Siegel, 2000).

BACKGROUND

Premise of Agile Software Development

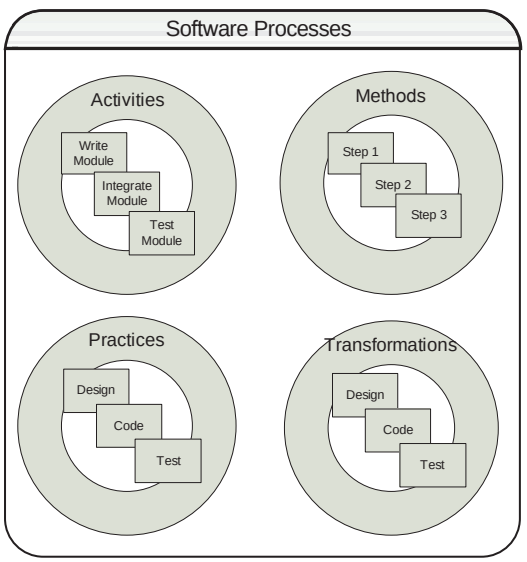
The professional goal of every development team is to deliver the highest possible value to the project and customers. Yet, projects fail, or fail to deliver value, at a frustrating rate due to an increase in process inflation. Plan-driven methods are those in which work begins with the elicitation and documentation of a *complete* set of requirements, followed by architectural and high-level design development and inspection. In this context, the concept of *agile* appeared where principles and values were shaped as a way to help teams avoid the cycle of process inflation and to focus on simple techniques for reaching their goals.

Agile processes allow adjustments of requirements during all phases of the development cycle and stress collaboration between software developers and customers and early product delivery (Donaldson & Siegel, 2000).

Key motivations of *agile methods* apparition are

- Iterative development is of lower risk than waterfall development (Larman, 2004).
- Early risk discovery and improvement.
- Promotes early change: consistent with new product development.
- Early partial product apparition.
- Satisfaction through early and repeated successes.
- Continuous testing activity.
- Final product matches client's desires better.

Figure 1. Software processes



Features Of Agile Software Development

The core of the "Manifesto for Agile Software Development" (n.d.) is as follows (Fowler, 2002; Fowler & Highsmith, 2001; Martin, 2001):

- individuals and interactions over processes and tools;
- working software over comprehensive documentation;
- customer collaboration over contract negotiation; and
- responding to change over following a plan.

The central aspects of agile methods are simplicity and speed. These goals can be achieved by software whose development is incremental, cooperative, straightforward, and adaptive.

The agile software development approach considers people the main resource of the development. In this context, its approach is collaborative considering that software development is, in fact, a collaborative team activity. In this way it steps away from the individualistic software engineering paradigm, and instead considers software development as a collaborative team activity. This enables agile software development teams to learn how to work together and thereby provides a mechanism for resolving the inevitable misunderstandings that occur during a project.

The set of consistent approaches that arise from agile software development processes are (Abrahamson, Salo, Ronkainen, & Warsta, 2002; Fowler & Highsmith, 2001; Larman 2004)

- human resource issues,
- amount of documentation to be as reduced as possible,
- communication is a critical issue, and
- modeling tools are not as useful as in other development processes.

Agile processes characteristics are stated as follows (Fowler & Highsmith, 2001):

- modularity is used on development process level;
- iterative activities with short cycles enable fast verifications and corrections;
- time spent with iteration cycles takes from 1 to 6 weeks;
- temperance in development process that removes all useless activities;
- adaptive;
- incremental; and
- collaborative and communicative working style.

There are many agile methods sharing common characteristics and starting from the same approach. Some of them are: *extreme programming, crystal family of methodologies, rational unified process, dynamic systems development method,* and *adaptive software development.* Each of them has its own processes, principles, practices, roles, and responsibilities, but they all have in common the agile approach.

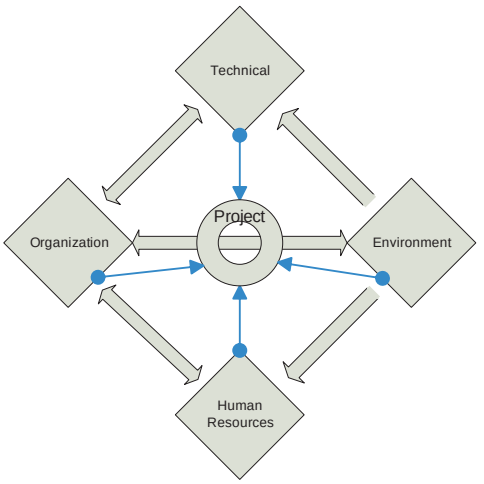
INFLUENCE PARAMETERS OF AGILE SOFTWARE DEVELOPMENT PROJECTS

In agile approach it is quite hard to distinguish between success and failure factors of projects. What was considered decisive for a project echoing success may be considered as a limit in another one. At the same time, a certain factor may influence not only a project, but may also persuade the rest of agile factors. In this context, it is difficult to draw a fixed and clear line between influence factors and how they act in agile software projects. Different factors[2] that influence the success or failure of software projects are shown in Figure 2 and are described hereafter.

Organizational Factors and Their Influences

- **Organization Culture:**
  1. A dynamic, agile organization will find agile methods extremely suitable for it (Abrahamson et al., 2002). Importance of customer feedback and control on an agile project requires an adaptive and collaborative working environment. In this working environment, which is, in fact, the organization, all its members, for example, management, developers, and testers must be in total agreement to use agile processes, because without organization commitment to being agile, failure may develop into a strong possibility (Smith & Pichler, 2005).
  2. In a bureaucratic organization where respecting plans, rules, and directives are a way of work, agile is inappropriate as a new path in creating value. A stagnant organization with a culture believing in efficiency, control, and process rigor tends to neglect the fact that success evolves from new successes, failures, and different alternatives used. The deeper these factors are grounded in the organizational culture, the more difficult they are to revolutionize, and the more easily they can became obstructions in adopting new approaches (Highsmith, 2000).

Figure 2. Main influence factors of agile projects



- **Team Dimension:**
  1. Team dimension is directly dependent on project size and influences communication between team members. The agile approach with face-to-face communication suits teams with less than 20-40 people.
  2. As team size grows, coordinating the interfaces becomes a dominant issue. The communication required in the agile approach breaks down and becomes increasingly difficult and complex with more than 20-40 people (Dyba, 2000). If more than 20-40 people are involved in an agile project, scaling strategies must be adopted, for example, breaking a large team into few smaller teams and applying agile for team coordination.
- **Team Distribution:**
  1. One of the factors that are likely to positively influence the success of an agile software development project is the centralized organization of the teams. Companies involved in distributed international projects will be affected by the cultural and political situations in those regions. For instance, if there are teams that are located in regions where the political and cultural environment supports collaborative work advocated by the agile approach, those teams are more likely to be successful over the ones that are located in regions where this is not the case.
  2. Directly dependant on project type and environment (especially for international projects), geographic distribution of teams may be an inconvenience for an agile project. In such a case, face-to-face communication is extremely difficult and documents apparitions are highly desirable.
- **Decision Time:**
  Teams using an agile approach are more likely to make decisions more quickly than other teams, relying on frequent informal communication.
- **Customer Commitment:**
  1. Close collaboration with customers provides a team following an agile approach the chance to react fast to adherence (or any change) in customer requirement. Customers have the opportunity of driving development along with the project managers (Lindvall et al., 2002). A project following the agile approach represents a partnership between customers and development teams, where each member has a specific role, responsibility, and authority. During the development effort, relationships between customers and developers must be collaborative.
  2. Software is not an ordinary commodity. Unlike other products, it cannot be ordered and received as a final product according to the specified requirements. A customer must describe its needs, and more than anything else time, budget, and resource constraints must be part of a contract. Such needs may change over time following interactions between the customers and the project team. Also, customer requirements may become a reason for failure of a system, if they are not "filtered" by the developers. The reason is that some of the requirements of the customers might be infeasible. Therefore, customer commitment becomes an important factor, the absence of which might negatively impact the project.

Human Resource Factors and Their Influences

The success of a software development project is often related to people factors (Turner & Boehm, 2003). Human resource factors are also hypothesized as important factors for the success of agile software development projects.

- **Competency:**
  One may be said to be competent for a software development project if one has real-world experience in the technology domain, has built similar systems in the past, and possesses good interpersonal and communication skills.
  1. One of Boehm's principles of top talent: "use better, and fewer people" is central to an agile process (Keith, 2002). Having high competency

members in a team is one of the most important requirements of the agile approach.

2. Even if a final product looks and works in the same manner, differences between competencies of the members influence all development phases. Too many slow workers either slow the pace of development of an entire team or end up being left behind by their faster colleagues (Cohn & Ford, 2003).

- **Personal quality:**
  1. Team members must not necessarily be extremely skilled and experienced people, but honest, collaborative, responsible, ready to learn, and work well with others. A team, where there are integrated distinct types of IT specialists (programmers, testers, analysts) will work well with such members.
  2. Just professional competencies are not enough anymore for a programmer. Agile's core principle: "Communication is critical" has secondary effects on the requirements of team members' qualities—without communication abilities and the desire to work in an agile team, even an expert may become a hindrance to the success of the process.
- **Percentage of experts in project:**
  1. A well-fabricated team, even without having various experiences in a specific field, will work efficiently, if for any success-critical issue at least one person has the time and expertise to resolve the issue. Davies's 131st principle "People are the key to success," combined with close team communication, allows to any team members to learn from his/her peer.
  2. For highly agile, tactical environments, as also for extremely specialized domains, external experts may be of limited help. If there is no such expert in an organization, it will take a long time to train one.

Technical Factors and Their Influences

- **Project Size:**
  1. A large project may influence other agile factors such as: team dimension, team distribution, and communication with a large number of customer representatives. Each of them has its own influence on the project.
- **Documentation:**
  1. An organization's goal should be to communicate effectively. As per the agile approach requirements, documentation should be given secondary importance. On the other hand, since documentation has important functions (Highsmith, 2004) in supporting project phases such as sustaining team collaboration and communication, preserving historical information and information evolution, and fulfilling requirements, documentation in agile projects must be concise and have the ability to respond to project specific needs, as per necessity. In other words, documentation should be strictly limited to what is optimal.
  2. "Good" products can be produced with or without documents; nevertheless, without a documented process it is difficult, for an organization to create products on time and within budget. Code is not the ideal medium for communicating the rationale and structure of a system. In this context, the team needs to produce human readable documents that describe the system and the rationale for their design decisions.

Environmental Factors and Their Influences

- **Culture:**
  1. As any other human activity, software development is highly influenced by regional (local) culture. A communicative, dynamic, progressive environment has a major influence on the members of a society. With such social values recognized, any individual may fit well in an agile team.
  2. On the other hand, an oppressive society, full of constraints has a bad influence on human behavior (considering agile set of values).
- **Cultural differences between team members:**
  1. As any new experience brings new values, encountering people with similar culture may have a positive influence on team communication and foster a willingness to work together.
  2. On the other hand, if there are extremely high differences between team members, it may take a long time until they are in fact able to work together efficiently.

FUTURE TRENDS

In addition to identifying the important factors influencing success, it is important to obtain empirical evidences from the industry. Similarly, most of the changes and challenges identified in the existing literature that are required for adopting agile processes in a traditional development organization are either based on experience of the developers, or are based on the opinions of agile experts without much empirical support. Additionally, different previous articles report long lists of such changes and challenges. However,

for an organization considering adopting agile processes, such long lists are of limited help. It is important to identify the critical factors responsible for the transition of traditional software development organizations into agile methods based on empirical survey-based evidences.

CONCLUSION

Agile software development methodologies have recently gained widespread popularity. The "Manifesto for Agile Software Development" (n.d.) states valuing "individuals, and interactions over processes, and tools, working software over comprehensive documentation, customer collaboration over contract negotiation, and responding to change over following a plan" (Fowler, 2002). However, little is known about how effective and efficient agile practices are over the traditional methodologies, and what their success factors are. There have been several disparate anecdotal evidences about the success of software development projects using agile methodologies.

In this article we introduced the concepts surrounding agile software development. We have also mentioned some of the important factors influencing the success or failure of agile software development projects. These factors are based on previously published literature. This article should help software development teams willing to adopt agile practices to be aware of the "what works and what does not" in agile projects.

We reviewed the parameters that affect the success of projects adopting agile software development methodologies based on previous anecdotal and practical experience stories. We also presented a conceptual framework showing the relationships between the success of agile software development and its predictors.

REFERENCES

Abrahamson, P., Salo, O., Ronkainen, J., & Warsta, J. (2002). Agile software development methods—Review and analysis. (VTT Publications No. 478). Retrieved from http://www.inf.vtt.fi/pdf/publications/2002/P478.pdf

Cohn, M., & Ford, D. (2003). Introducing an agile process to an organization. *IEEE Computer, 36*(6), 74-78.

Donaldson, S., & Siegel, S. (2000). *Successful software development.* Prentice Hall.

Dyba, T. (2000). Improvisation in small software organization. *IEEE Software, 17*(5), 82-87.

Fowler, M. (2002). The agile manifesto: Where it came from and where it may go. Retrieved from http://martinfowler.com/articles/agileStory.html

Fowler, M., & Highsmith, J. (2001, July 16). The agile manifesto. *Software Development Magazine.* Retrieved from http://www.sdmagazine.com/documents/s=844/sdm0108a/0108a.htm

Highsmith, J. (2000, July/August). Retiring lifecycle dinosaurs. *Software Testing and Quality Engineering Magazine.*

Highsmith, J. (2004). Agile revolution. In *Agile project management: Creating innovative products.* Addison-Wesley.

Jalote, P. (2002). *Software project management in practice.* Addison-Wesley.

Keith, E. R. (2002). Agile software development processes: A different approach to software design. Retrieved from http://www.agilealliance.com/articles/keitheveretteragileso/file

Larman, C. (2004). *Agile & iterative development: A manager's guide.* Addison-Wesley.

Lindvall, M., Basili, V., Boehm, B., Costa, P., Dangle, K., Shull, F., et al. (2002). Empirical findings in agile methods. *Proceedings of the 2nd XP Universe and First Agile Universe Conference on Extreme Programming and Agile Methods* (pp. 197-207).

Manifesto for agile software development. (n.d.). Retrieved from http://agilemanifesto.org

Martin, R. C. (2001). *Agile processes.* Prentice Hall.

Smith, P. G., & Pichler, R. (2005, April). Agile risks, agile rewards. *Software Development Magazine,* 50-53.

Turner, R., & Boehm, B. (2003, December). People factors in software management: Lessons from comparing agile and plan driven methods. *CrossTalk: The Journal of Defense Software Engineering,* 4-8.

KEY TERMS

Agile: The quality of being quick.

Agile Manifesto: A public declaration of principles and intentions shared by a group of software developers who advocated the philosophy of software development.

Agile Methodologies: Methodologies that follow the manifesto, principles, and values of agile software development.

Agile Software Development: An approach for software development advocated by a group of software developers believing in developing software in shorter time boxes, faster communication, working closely with customers, and allowing changes to requirements even late in the software development process, among others.

Iterative Development: The art of developing software incrementally in versions by taking advantage of lessons learned from producing previous deliverables of software.

Process: A sequence of steps following which can lead to an outcome.

Project: A plan or proposal, which is undertaken, usually over a fixed time duration.

ENDNOTES

1 http://www.sei.cmu.edu/iso-15504/resources/glossary.PDF

2 These factors are selected based on previous literature. Many of these pieces of literature came up with these factors based on the authors' experiences, or empirical studies involving cases and surveys.

Information and Communication Technology for E-Regions

Koray Velibeyoglu
Izmir Institute of Technology, Turkey

Tan Yigitcanlar
Queensland University of Technology, Australia

INTRODUCTION

Information and communication technologies (ICTs) are essential components of the knowledge economy, and have an immense complementary role in innovation, education, knowledge creation, and relations with government, civil society, and business within city regions. The ability to create, distribute, and exploit knowledge has become a major source of competitive advantage, wealth creation, and improvements in the new regional policies. Growing impact of ICTs on the economy and society, rapid application of recent scientific advances in new products and processes, shifting to more knowledge-intensive industry and services, and rising skill requirements have become crucial concepts for urban and regional competitiveness. Therefore, harnessing ICTs for knowledge-based urban development (KBUD) has a significant impact on urban and regional growth (Yigitcanlar, 2005). In this sense, e-region is a novel concept utilizing ICTs for regional development.

Since the Helsinki European Council announced Turkey as a candidate for European Union (EU) membership in 1999, the candidacy has accelerated the speed of regional policy enhancements and adoption of the European regional policy standards. These enhancements and adoption include the generation of a new regional spatial division, NUTS-II statistical regions; a new legislation on the establishment of regional development agencies (RDAs); and new orientations in the field of high education, science, and technology within the framework of the EU's Lisbon Strategy and the Bologna Process. The European standards posed an ambitious new agenda in the development and application of contemporary regional policy in Turkey (Bilen, 2005). In this sense, novel regional policies in Turkey necessarily endeavor to include information society objectives through efficient use of new technologies such as ICTs. Such a development seeks to be based on tangible assets of the region (Friedmann, 2006) as well as the best practices deriving from grounding initiatives on urban and local levels. These assets provide the foundation of an e-region that harnesses regional development in an information society context.

With successful implementations, the Marmara region's local governments in Turkey are setting the benchmark for the country in the implementation of spatial information systems and e-governance, and moving toward an e-region. Therefore, this article aims to shed light on organizational and regional realities of recent practices of ICT applications and their supply instruments based on evidence from selected local government organizations in the Marmara region. This article also exemplifies challenges and opportunities of the region in moving toward an e-region and provides a concise review of different ICT applications and strategies in a broader urban and regional context.

The article is organized in three parts. The following section scrutinizes the e-region framework and the role of ICTs in regional development. Then, Marmara's opportunities and challenges in moving toward an e-region are discussed in the context of ICT applications and their supply instruments based on public-sector projects, policies, and initiatives. Subsequently, the last section discusses conclusions and prospective research.

BACKGROUND

New Regionalism and Information Society in Turkey

In the 1950s, Turkey was divided into seven geographic regions based on topographic and climatic conditions without paying attention to administrative aspects. In terms of territorial division, Turkey has a national administration, and 81 province, 873 district, and 3,227 local administrations. Provinces and districts are both administrative units of the national government and territorial units of local government. Representatives of the national government, governors of provinces, and heads of districts on the one hand, and local government bodies and provincial local governments on the other hand work in the same areas but carry out different duties (Sagbas, 2003). The administrative reorganization and spatial division of regions have been under review posed by

the new regulations covering the establishment of RDAs and NUTS-II statistical regions.

Adoption of EU's regional standards and information society objectives are challenging ambitions for Turkey. The EU's Lisbon agenda (2000) has also come up with a similar ambitious plan with a strategic vision to become the most competitive and dynamic knowledge economy in the world that is capable of sustainable economic growth with more jobs and greater social cohesion (Campano et al., 2004). To address these objectives, a comprehensive e-transformation program, e-Turkey, was prepared rapidly after Turkey participated in the e-Europe initiative in 2001. The main goals of this initiative include cheaper, faster, and secure Internet; investing in people and skills; stimulation of Internet use in European regions; and acceleration in forming information society foundations (Tuzun & Sezer, 2002). In conjunction with this initiative, the e-transformation Turkey project was launched in 2002. The information society department of the State Planning Organization (SPO) was assigned for the coordination of this project. The prime ministry, nongovernment organizations, and all public institutions are identified as affiliated organizations for the project (SPO, 2004).

Local governments play a key role in developing local ICT policies in order to coordinate with national-level policy implications, which are indexed to the e-Europe initiative (Akin, 2005). The e-Turkey initiatives, however, are undertaken by both national and local governments, where no coordinating regional authority is allotted with these initiatives. These regional policy initiatives focus on the promotion of wealth, welfare, and sustainability in a broader information society context. Therefore, focusing on the regional level would likely increase the success chance in the implementation of ICT applications within the e-Turkey initiatives.

E-Region and ICT

E-region can be considered as the set of innovative actions to achieve economic and social cohesion and to raise the technological level of regions through the use of information ICTs.

ICTs are the backbones of the knowledge economy and in recent years have been recognized as effective tools for promoting economic growth and sustainable development (Chen & Dahlman, 2005). According to Millard (2002), the five *E*s (entity, economy, equity, environment, e-technology) provide basic conditions to achieve sustainable regional development. For example, entity promotes territorial identity and integration; similarly, economy is the engine for growth and efficiency, equity resembles cohesion and inclusion in encountering the spatial digital divide and promoting welfare, and environment is an important tangible asset for regions inducing sustainability. E-technology or ICTs complement the other four dimensions and can widen the spectrum of innovativeness and creativity of a region.

Another emphasis for the development of the knowledge economy is to enhance regional governance. It is widely accepted that good governance and effective institutional structure are important sources of regional competitiveness. This requires the partnership of private-, public-, and voluntary-sector bodies aimed at driving forward a region's e-agenda.

The EU's regional approach and projects for the information society and urban technologies have a good framework toward understanding e-regions. These policies have been discussed under three major objectives: Support the provision of ICT infrastructure to reduce the digital divide and regional disparities, stimulate new electronic services and innovative ICT applications ranging from e-commerce to e-governance, and invest in people to ensure necessary skills and capabilities via distance learning and digital literacy (European Commission [EC], 2006). In this context, a variety of e-region initiatives in the European region has been under way: Kaunas E-Region (Latvia), E-Region Blagoevgrad (Bulgaria), E-Bourgogne Programme (France), Kuyavia and Pomerania E-Region (Poland), and E-Region Schleswig-Holstein (Germany). These initiatives are parts of the e-Europe region, which represents the information society at the service of regional development.

MARMARA: TOWARD AN E-REGION?

Marmara is the most developed region in Turkey, covering approximately 60% of the output of the Turkish manufacturing industry, 37% percent of the gross domestic product (GDP), and 26% of the total population. The region's dominant position also reflects the share of public investment (28.7%) and private investment incentives (46.3%; Karadag & Deliktas, 2004). On the other hand, Marmara has also been highly innovative in the implementation of cutting-edge ICT applications and associated public-sector (national and local governments) supply instruments, which is explored based on the categories posed by Heeks (2005).

Supply Environment (Policies, Strategies, and Legislations)

Grounding the national-level ICT policies such as e-Turkey to the urban and regional level is a major challenge that needs to be tackled. Within the frame of e-Turkey, Yalova province in Marmara is selected as a pilot city for the initiative. ICT projects of Yalova were presented as best practices in various national and international conferences, meetings, and platforms. In this context, various local ICT policies have been deployed in order to enhance public Internet access (public Internet kiosks), economic development (call centers), e-literacy (adult IT certification programs), and online

public services (local e-government). Macro factors such as economic instability and change, and the ever-changing local political context, however, lead to the sustainability failure of so-called IT City Yalova projects (Velibeyoglu, 2006).

On the strategic level, two recent developments are important in the implementation of ICT applications internal to local governments' structure. First, the strategic decision of Turkey to join the EU and the need to adopt EU principles in the field of local government have constituted a powerful driving force and accelerated local government reform in Turkey (Kosecik & Sagbas, 2004; Ozkaynak, 2005). In the new Local Government Act (2005) establishment of geographic and urban information systems (UIS) for intermunicipality tasks has become obligatory for all metropolitan municipalities. Second, through total quality management (TQM) strategies, some local governments (Bursa, Kocaeli, Yalova) in Marmara have reorganized their departmental structures and processes that allowed them to accommodate rational technical systems like GIS (geographic information system). In Yalova, for example, online local-government services are measured and evaluated through TQM principles (Velibeyoglu, 2006).

Supply of Resources (Infrastructure, Skills, and Other Resources)

Local governments in Turkey have experienced serious policy bottlenecks in governing and investing in their ICT infrastructures. Infrastructure provision is left for the national government and private sector. The detrimental effects of the Marmara earthquake (1999) had important influence on local governments in Marmara as well as Turkey in terms of recognition of the importance of telecommunications infrastructure and ICT-based services. For example, Yalova is selected as Turkish Telecom's pilot city for the provision of a natural-disaster-resistant Internet infrastructure. Another impact of the earthquake was the recognition of the vital importance of an information infrastructure that accelerated the development of spatial information systems in post-earthquake cities (Sakarya, Duzce, Yalova, Kocaeli, Bursa) in the region (Velibeyoglu & Saygin, 2005).

The availability of skilled personnel for information systems use is one of the biggest problems in public-sector organizations in Turkey including local governments. This was largely because of the problem of the public-sector employment policy that neither computer skills nor individual productivity was encouraged and rewarded by the administrative system (Tecim, 2004). However, some innovative and careful attempts can be observed in the local governments of Marmara. In the Bursa metropolitan municipality, for example, the UIS Division was founded to support the functioning of municipal services, in-house production and maintenance of information, and training of the staff in IS applications (Velibeyoglu, 2005).

As hubs of the knowledge economy, universities and research centers play a critical role in the creation of e-regions (Marceau & Martinez, 2005). In this sense, in Marmara there are a considerable number of universities and research centers (e.g., TUBITAK Marmara Research Center) that help to facilitate human resources and adoption capability in information systems and technologies. Bilisim Vadisi (Informatics Valley), for example, was designed in the Marmara region Istanbul to ensure development of Turkey among the regional countries as a center of production and operation for international IT corporations as well as to attract foreign direct investment in the domestic IT sector (SPO, 2006). In addition, new technology-park developments in Marmara have been planned to focus on R&D (research and development) firms in the automotive and telecommunication sectors in order to support the Specialization in Technology Development Zones objectives of the 2006-2010 Information Society Strategy Action Plan of Turkey (SPO).

Supply Mechanisms (Local and Global Organizations, Initiatives, and Networks)

Building strong partnerships in disseminating and sharing knowledge between institutions such as academia, the public sector, and the private sector has become a vital issue for KBUD and the transition to a knowledge economy (Yigitcanlar, 2005).

The information need of postdisaster management and recovery have provided some international and national donor aid in the establishment of UIS in post-earthquake cities in Marmara. The Sakarya governorship GIS center, for example, has developed several applications for emergency situations (e.g., to provide tents, prefabricated houses, food, and social activities) with the help of UNICEF-Turkey and the Ministry of Internal Affairs (Tecim, 2004).

On the European level, local governments in Marmara have affiliated regional information society initiatives in Europe. The Yalova municipality currently has the only Turkish member of the Telecities network, which is a regional institute that aims to bring together towns and cities for the development of urban ICT applications.

ICT Applications (Local E-Government and Urban Information Systems)

As aforementioned, new EU integration initiatives including e-Turkey have visible impacts on the implementation of ICT applications such as local e-government and UIS in local governments. The national and local e-government initiative lies at the heart of the e-Turkey project. It is seen as an essential part of government reform and restructuring (SPO, 2004). Within the framework of the e-government project, restructuring the state, raising the level of educa-

tion and health of the society, strengthening scientific and technological capability, developing new technologies, and improving physical infrastructure have been determined as critical. About 200 large, medium, and small e-government projects are now under way. Among them, local e-government projects in the provinces of Marmara (e.g., Istanbul, Bursa, Yalova) have dominated e-governance awards that promote the best projects and initiatives in ICT applications in Turkey (Velibeyoglu, 2006).

Initiatives supporting e-municipality and e-government, and transition from government to e-governance and then to e-democracy raised the importance of transparency, communication, public accountability, and participation issues (Yigitcanlar & Baum, 2006). In this sense, the concept of UIS began to be popular among the local governments. The UIS concept is used as an umbrella term encapsulating all the efforts for an information system, whether GIS or MIS (management information system), or information technologies like the Internet within an integrated system that is supposed to be performed in local-government operations in order to support organizational rationality (Velibeyoglu, 2005).

Although so-called UISs were being marketed by vendors as the panacea for all problems, the implementation of large-scale information systems generally ended in failure because ISs require large changes in an organization's existing structure. Implementation of UIS generally incorporates a significant set of rational structures, processes, culture, professional strategies, and involvement (Saygin, 2003).

In the Turkish case, although no local government has been able to complete the establishment of a citywide UIS so far, the most promising applications have come from the local governments in Marmara (e.g., Bursa UIS) because of several reasons. First, local governments in Marmara have relatively more experience with UIS (e.g., Istanbul since 1989, and Bursa since 1996). Second, there has been positive reception from local governments in the implementation of various technological systems and therefore some of them rearranged their organizational structure for better utilization of IS (e.g., TQM or semiautonomous UIS departments or centers). Third, financial resources have been available particularly as a part of large-scale infrastructure projects. In the Bursa UIS case, for example, funding for UIS was obtained through an international donation within the framework of the fresh-water infrastructure project (Velibeyoglu, 2005). Finally, the specific supply instruments outlined above have been influential in the development of UIS in the region.

FURTHER TRENDS

For future studies and research, regional-level indicators should be developed in relation to both supply and demand from businesses, government, and households. Similarly, the tangible assets of the region should be taken into account. The best-practice applications obtained from the Marmara region should be extended, updated, and shared with other public-sector organizations, yet only a few studies focus on the institutional dimension of the ICT applications and the context of supply instruments. The important and challenging task of researchers, then, is to demonstrate the multilevel evidence in recorded case studies and research.

CONCLUSION

The EU harmonization process has set the challenge of developing information society objectives in Turkey. In this context, there is an urgent need to find implementation paths to realize ICT policies not only at local and national levels, but also on the regional level as well. This concise review revealed that local governments in Marmara have already taken part in the European urban and ICT networks. With best-practice implementations, local governments in Marmara are accelerating the efforts of the country in terms of spatial IS and e-governance applications. Although there are some indicators of e-region and some short-term individual best practices from Marmara, the efforts moving toward an e-region is still embryonic and largely suffers from the uncoordinated nature of supply instruments.

As international relations have been intensified, and as common concerns have been shared, local governments in Marmara have required new networking mechanisms for interorganizational as well as international cooperation. Therefore, new ICT supply mechanisms should be introduced toward an e-region to support KBUD, allowing public and private partnership and community participation in decision-making processes, and also encouraging local economic development and social cohesion. This is to say that ICT applications need to be smoothly adopted for unstable, rapidly changing sociospatial circumstances and considered soft organizational realities of local organizations.

Supply mechanisms and ICT applications should be fully accommodated in the future developments of the region. For example, a coordinating autonomous or semiautonomous regional body for ICT policy and implementation needs to be established. Such a mechanism is less vulnerable to environmental changes and political instability, which are very important in a developing-country context.

REFERENCES

Akin, U. (2005). *Strategic urban ICT management in metropolitan governance.* Paper presented at the 45th European Regional Science Congress, The Netherlands.

Bilen, G. (2005). *Novel regional policy of Turkey inline with EU standards.* Paper presented at the Regional Science Conference, Denmark.

Chen, D., & Dahlman, C. (2005). *The knowledge economy.* Washington, DC: The World Bank.

European Commission (EC). (2006). *Europe's regions & the information society.* Retrieved from http://europa.eu.int/information_society/regwor/reg/index_en.htm

Friedmann, J. (2006). *The wealth of cities.* Canada: UN-Habitat, University of British Columbia.

Heeks, R. (2005). *Foundations of ICTs in development* (eDevelopment briefing). Manchester, United Kingdom: Development Informatics Group, University of Manchester.

Karadag, M., & Deliktas, E. (2004). The effects of public infrastructure on private sector performances in the Turkish regional manufacturing industries. *European Planning Studies, 12*(8), 1145-1156.

Kosecik, M., & Sagbas, I. (2004). Public attitudes to local government in Turkey. *Local Government Studies, 30*(3), 360-383.

Marceau, J., & Martinez, C. (2005). *Stocktake of NSW as a potential knowledge hub.* Sydney, Australia: AEGIS, University of Western Sydney.

Millard, J. (2002). *Regional development and cohesion in the European information society* (working paper). Denmark: Danish Technological Institute.

Ozkaynak, B. (2005). *Indicators and scenarios for urban development and sustainability.* Unpublished doctoral dissertation, Universitat Autonoma de Barcelona, Spain.

Sagbas, I. (2003). *Financing local government in Turkey.* Istanbul, Turkey: Istanbul Municipality.

Saygin, O. (2003). *GIS based urban policy development.* Unpublished doctoral dissertation, Dokuz Eylul University, Turkey.

State Planning Organization (SPO). (2004). *State Planning Organization contribution of Turkey to eEurope+* (progress report). Turkey: Information Society Department.

State Planning Organization (SPO). (2006). *Information society strategy report 2006-2010.* Turkey: State Planning Organization.

Tecim, V. (2004). *Disaster management system with GIS in Sakarya.* Paper presented at the 24th Urban Data Management Symposium, Venice, Italy.

Tuzun, G., & Sezer, S. (2002). *National report on Turkey.* Johannesburg, South Africa: UNDP.

Velibeyoglu, K. (2005). Urban information systems in Turkish local governments. In S. Marshall, W. Taylor, & X. Yu (Eds.), *Encyclopedia of developing regional communities with ICT.* Hershey, PA: Idea Group.

Velibeyoglu, K. (2006). *Urban ICT policies for Turkish local governments.* Paper presented at the Electronic City Workshop, Bratislava, Slovakia.

Velibeyoglu, K., & Saygin, O. (2005). *Spatial information systems in Turkish local governments.* Paper presented at the Ninth International Conference on Computers in Urban Planning and Urban Management, CASA, London.

Yigitcanlar, T. (2005). *The making of knowledge cities.* Paper presented at the International Symposium on Knowledge Cities, Saudi Arabia.

Yigitcanlar, T., & Baum, S. (2006). Benchmarking local e-government. In M. Khosrow-Pour (Ed.), *Encyclopedia of e-commerce, e-government and mobile commerce* (pp. 37-42). Hershey, PA: Idea Group.

KEY TERMS

E-Region: It is a set of policy actions to achieve economic and social cohesion and to raise the technological level of regions through the use of ICTs. It covers a wide range of ICT initiatives from the provision of ICT infrastructure to the promotion of new electronic services and innovative ICT applications.

Knowledge-Based Urban Development: This is a knowledge-intensive urban planning and development approach to nourish the transformation and renewal of cities into knowledge cities.

Local E-Government: Local e-government refers to information, services, or transactions that local governments provide online to citizens using the Internet and Web sites.

Nomenclature of Territorial Statistical Units: It is a system of the classification of regions across the EU used by the European Commission.

Regional Development Agency: This is a semiautonomous, regionally based body operating at arms length vis-à-vis its sponsoring political authority. It is a multifunctional and integrated agency, the level of which may be determined by the range of policy instruments it uses.

Total Quality Management: TQM is an organization-wide effort to improve quality and make it the responsibility of all employees.

Urban Information Systems: UISs are powerful means for governments in meeting long-term strategic planning and management challenges. They provide a heightened awareness of the interdependency among environmental, social, and economic health and the impact of decisions made by neighboring jurisdictions, government agencies, and private business.

Category: Multimedia Technology

Information Fusion of Multi-Sensor Images

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

The human perception to the outside world is the results of action among brain and many organs. For example, the intelligent robots that people currently investigate can have many sensors for sense of vision, sense of hearing, sense of taste, sense of smell, sense of touch, sense of pain, sense of heat, sense of force, sense of slide, sense of approach (Luo, 2002). All these sensors provide different profile information of scene in same environment. To use suitable techniques for assorting with various sensors and combining their obtained information, the theories and methods of multi-sensor fusion are required.

Multi-sensor information fusion is a basic ability of human beings. Single sensor can only provide incomplete, un-accurate, vague, uncertainty information. Sometimes, information obtained by different sensors can even be contradictory. Human beings have the ability to combine the information obtained by different organs and then make estimation and decision for environment and events. Using computer to perform multi-sensor information fusion can be considered as a simulation of the function of human brain for treating complex problems.

Multi-sensor information fusion consists of operating on the information data come from various sensors and obtaining more comprehensive, accurate, and robust results than that obtained from single sensor. Fusion can be defined as the process of combined treating of data acquired from multiple sensors, as well as assorting, optimizing and conforming of these data to increase the ability of extracting information and improving the decision capability. Fusion can extend the coverage for space and time information, reducing the fuzziness, increasing the reliability of making decision, and the robustness of systems.

Image fusion is a particular type of multi-sensor fusion, which takes images as operating objects. In a more general sense of image engineering (Zhang, 2006), the combination of multi-resolution images also can be counted as a fusion process. In this article, however, the emphasis is put on the information fusion of multi-sensor images.

BACKGROUND

There are many modalities for capture image and video, which use various sensors and techniques (Brakenhoff, 1979; Committee, 1996; Bertero, 1998), such as visible light sensor (CCD, CMOS), infrared sensor, depth sensor, con-focal scanning light microscopy (CSLM), a variety of computer tomography techniques (CT, ECT, SPECT), magnetic resonance imaging (SAR), synthesis aperture radar, millimeter wave radar (MMWR), etc.

Main Steps of Image Fusion

For image fusion, many image techniques can be used in three steps (Zhang, 2007).

Image Pre-Processing

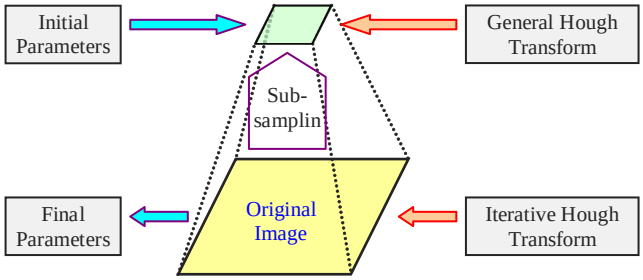
It includes image normalization (gray level equipoise, re-sampling, and interpolation), image filtering, color enhancement, edge sharpening, etc. Image fusion is carried out among images of different sizes, different resolutions, and different dynamic ranges of gray levels or colors. Image normalization is to normalize these parameters. Image filtering is to high pass filter the higher resolution image to obtain high frequency texture information, to keep it in fusion with lower resolution image. Image color enhancement is to increase the color contrast in lower resolution image, to reflect the spectrum information into the fused image. Edge sharpening is performed on high-resolution image for making the boundary clear and reducing noise, thus it fuses the space information from high-resolution image to low resolution image.

Image Registration

It is to align different images in space. In a more general sense, it is a special case of image matching, which has many existing techniques (Kropatsch, 2001; Shapiro, 2001; Buckley, 2003; Zhang, 2007). Image fusion has high requirement for accurate registration. If the registration error is higher than one pixel, then the fused results will show superposition effect and the visual quality of image will be greatly reduced.

Image registration can be classified as relative registration and absolute registration. The relative registration takes one image from many images of the same category as a reference image; other images will be aligned relatively to this reference image. Absolute registration takes the space

Figure 1. Framework for image registration using control point-based multi-scale Hough transform



coordinate system as the reference system; images to be fused will be aligned relative to this system.

Image registration can be classified also as region-based registration and feature-based registration. Control point (corresponding points in both images to be registrated) is a typical feature used in feature-based registration. Once the correspondence between control points was determined, the registration process can be carried out with determined parameters. The general Hough transform (GHT) is a commonly used technique. It can be considered as the evidence accumulation method. The global search space depends on the scale and rotation parameters, and can be very huge. To reduce the complexity of GHT, iterative Hough transform (IHT) can be used. However, IHT is influenced by the initial parameters and the range of parameter values, and often converged to local maximum. By using Hough transform in a multi-resolution decomposition environment, as shown in Figure 1, the robustness of GHT and computation efficiency of MIHT can be combined (Li, 2005).

In multi-resolution decomposition-based techniques, few control points are used in low-resolution layer in which GHT is used to obtain accurately the initial values of transform parameters. While in high resolution, IHT is used to accelerate the process.

Image Fusion

It is performed after image pre-processing and registration. The quantitative fusion is to fusion a group of data to obtain a consistent data, which is a conversion from data to data. The qualitative fusion is to fusion many single decisions to form a combined decision, which is a conversion from several uncertainty representations to a relative coherent representation. The quantitative fusion often treats information represented by numeric value while the qualitative fusion mainly treats information represented by non-numeric values.

Three Layers of Image Fusion

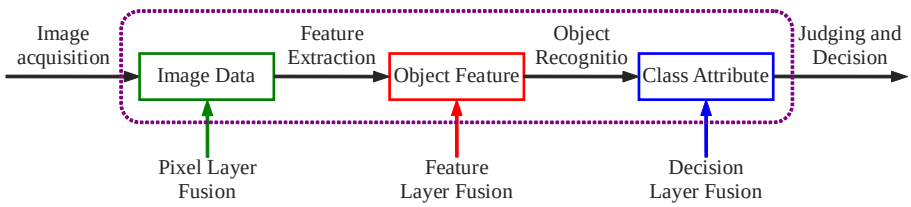
The multi-sensor image fusion can be split into three layers. They are, from low to high, pixel-based fusion layer, feature-based fusion layer, and decision-based fusion layer (Polhl, 1998). In recent years, the development tendency of fusion is going from pixel to region (Piella, 2003).

The flowchart of multi-sensor image fusion with three layers is illustrated in Figure 2. There are three steps from capturing scene image to making judgment and decision: feature extraction, object recognition, and decision creation. Three layers of image fusion are just corresponding to these three steps. The pixel-based fusion is made before the feature extraction step, the feature-based fusion is made before the object recognition step, and the decision-based fusion is made before the decision creation.

Pixel Layer Fusion

Pixel layer fusion is conducted in low layer, data layer. It operates directly on captured images and produces a single

Figure 2. Flowchart of multi-sensor image fusion



fused image. Pixel layer fusion provides the basis for high layer fusion. The advantages of pixel layer fusion is that it will keep as more as possible original information, so the precision obtained would be higher than that of other two fusions. The main disadvantages are the huge information to be treated and high computation cost. Moreover, pixel layer fusion often requires the data to be fused are captured by the same type or similar type sensors.

Feature Layer Fusion

Feature layer fusion is conducted in middle layer. It needs to extract features, obtain scene information, and integrate them to provide higher believable decision. Feature layer fusion not only keeps the important information, but also compresses the data volume. It is suitable for sensors of different types. The advantage of feature layer fusion is that it deals with less data than the pixel layer fusion, so it is more suitable for real time process. The main disadvantage of feature layer fusion is that its precision is worse than pixel layer fusion.

Decision Layer Fusion

Decision layer fusion is conducted in the highest layer, often performed with the help of symbolic computation. It makes directly the optimal decision according to the reliability of each decision, as each process unit has already finished the tasks of object classification and recognition. The advantage of decision layer fusion is that it has the properties of high tolerance, opening and real time. The main disadvantage of decision layer fusion is that the information has already had a lot of loss before fusion, so the precision, either in time of in space, would be inferior to other two fusions.

Principal properties of the three fusion modes are listed in Table 1.

Some typical techniques used in three fusion modes can be found in (Zhang 2007).

MAIN FOCUS OF THE CHAPTER

Techniques in pixel layer fusion will be illustrated with real examples. In pixel layer fusion, the original images to be fused have some different but complementary properties.

Basic Fusion Methods

In the following, by taking the fusion of TM (thematic map) multi-spectrum images captured by Landsat earth resource satellite and of complete spectrum images captured by SPOT remote sensing satellite as example, several basic methods are introduced. Currently, TM multi-spectrum images covers seven bands ranged from blue to infrared (with wave length 0.45 ~ 12.5 μm), and SPOT whole spectrum images cover five bands ranged from visible light to near infrared (with wave length 0.5 ~ 1.75 μm). The space resolution of SPOT image is higher than that of TM image, but the spectrum coverage of TM image is wider than that of SPOT image. Figures 3(a) and 3(b) show a TM image $f_t(x, y)$ in Band-5 (with wavelength 1.55 ~ 1.75 μm) and a SPOT image $f_s(x, y)$ with wavelength 0.5 ~ 0.73 μm, taken from same place.

Weighted Average Fusion

It is an intuitionistic method with the following steps:

(1) Select the region of interest in $f_t(x, y)$.
(2) Re-sample different wave band images in this region to extend $f_t(x, y)$ to a high-resolution image.
(3) Select the corresponding region of interest in $f_s(x, y)$, and match it with $f_t(x, y)$.
(4) Perform the following algebra operation to obtain weighted average fused image.

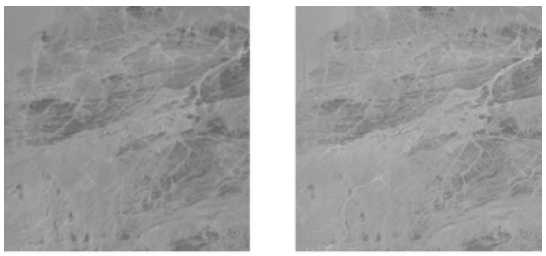
$$g(x, y) = w_s f_s(x, y) + w_t f_t(x, y) \quad (1)$$

where w_s and w_t are the weighting values for $f_s(x, y)$ and $f_t(x, y)$, respectively.

Table 1. Principal properties of three fusion modes

Fusion Layer	Fusion Level	Information Loss	Tolerance	Anti-disturb	Precision	Real Time	Computation Complexity
Pixel Layer	Low	Small	Bad	Bad	High	Poor	Big
Feature Layer	Middle	Moderate	Middle	Middle	Middle	Moderate	Middle
Decision Layer	High	Big	Good	Good	Low	Good	Small

Figure 3. Examples of TM image and SPOT image



(a) (b)

Pyramid Fusion

Pyramid is a common data structure to represent images in multi-scale. Pyramid fusion can be carried out by:

(1) Select the region of interest in $f_t(x, y)$.
(2) Re-sample different wave band images in this region to extend $f_t(x, y)$ to a high-resolution image.
(3) Decompose all $f_t(x, y)$ and $f_s(x, y)$ to be fused according to pyramid structure.
(4) Fuse the corresponding decomposition results of $f_t(x, y)$ and $f_s(x, y)$ in every layer of pyramid.
(5) Reconstruct the fused image from the fused pyramid by using inverse process for generating pyramid.

HSI Transform Fusion

HSI (hue, saturation, and intensity) transform converts the color image from RGB space to HSI space. HSI transform fusion performs fusion operation with the following steps:

(1) Select three bands of images from $f_t(x, y)$, take them as *R*, *G*, and *B* images, and transform them into *H*, *S*, and *I* images.
(2) Substitute *I* image after HSI transform (this image determine the details) by $f_s(x, y)$.
(3) Perform inverse HSI transform; take thus obtained RGB image as the fused image.

PCA-Based Fusion

PCA (principal component analysis) relies on KL transform. The main steps for PCA-based fusion are:

(1) Select three or more bands of images from $f_t(x, y)$ to perform PCA.
(2) Take the first principal component obtained by PCA operation, match it with $f_s(x, y)$ by using their histograms and make them having comparable mean and variance values.
(3) Substitute the first principal component by the above matched $f_s(x, y)$, perform inverse PCA, and take thus obtained image as the fused image.

Wavelet Transform Fusion

Wavelet transform decomposes an image into sub-images corresponding to different structures of the image. The main steps for wavelet transform fusion are (Pajares, 2004):

(1) Perform wavelet transform for both $f_t(x, y)$ and $f_s(x, y)$, obtain low frequency and high frequency sub-images for each image.
(2) Substitute the low frequency sub-images of $f_s(x, y)$ by those of $f_t(x, y)$.
(3) Combine substituted low frequency sub-images of $f_t(x, y)$ with the high frequency sub-images of $f_s(x, y)$; perform the inverse transform to obtain the fused image.

Combination of Fusion Methods

The above introduced fusion methods have their own particularity.

Weighted average fusion is simple and fast, but ineffective for anti-jamming and the quality of fused image is questionable. One typical problem is the blur caused by averaging.

Pyramid fusion is also simple to implement and can provide clearly fused image. However, the different layers in pyramid have correlation, so the images in different layers have redundancy. Besides, the reconstruction of pyramid has some instability, especially for distinct images.

HSI transform fusion, when used to fuse TM multi-spectrum image and SPOT whole spectrum image, can make the fused image having high definition to enhance the spatial detail information in image. However, if substitute all *I* component of TM image by SPOT image, the spectrum information will have a big loss and the resulted fusion image will have a large distortion.

PCA-based fusion produces fused image with high spatial definition and high spectrum definition, and the details for objects will be even clear. However, if substitute the first principal component of TM image by SPOT image, some useful information in the first principal component of TM image that related to the spectrum property will be lost. In this case, the spectrum definition of fused image will be affected.

Wavelet-transform fusion can effectively keep the spectrum information from multi-spectrum image and the detailed information from whole spectrum image, so the fused image would be better both in visual appearance and in statistics. However, the standard wavelet transform has two problems.

One is that it is equivalent to filter image with high-pass and low-pass filters, this filtering process will cause the loss of some original information. Another is that the gray levels of TM and SPOT images are quite different, the fusion would cause the change of TM image's spectrum information and induce the generation of noise.

To overcome the problems in using only one type of fusion methods, different fusion methods have been combined in practice.

Fusion by Combining HSI Transform and Wavelet Transform

It has the following steps:

(1) Select three bands from $f_t(x, y)$; perform the transform from RGB space to HSI space.
(2) Perform wavelet-transform for both $f_s(x, y)$ and I component.
(3) Substitute the high frequency coefficients obtained from the decomposition of I component by the high frequency coefficients obtained from the wavelet transform of $f_s(x, y)$.
(4) Perform inverse wavelet transform for all wavelet coefficients after substitution to obtain a new intensity component I'.
(5) Perform the transform from H, S and I' to R, G, B and obtain the fused image.

Figures 4(a) and (b) show the fused images obtained from fusing Figures 3(a) and (b) with HSI transform fusion and wavelet transform fusion, respectively (Bian, 2005). Figure 4(c) shows the fused images obtained by using the combined method. Since the new fused result not only keeps the high frequency information from SPOT image but also keeps a lot of texture information from TM image, it has clear details and provides better visual impression than either Figure 4(a) and Figure 4(b).

Fusion by Combining PCA and Wavelet Transform

It has the following steps:

(1) Perform PCA operation for all bands of $f_t(x, y)$.
(2) Take the first principal component obtained by PCA operation, match it with $f_s(x, y)$ by using their histograms and make them having comparable mean and variance values.
(3) Perform wavelet transform for matched two images.
(4) Substitute the high frequency coefficients obtained from the decomposition of $f_s(x, y)$ by the high frequency coefficients obtained from the first principal component of $f_s(x, y)$.
(5) Perform inverse wavelet transforms for all wavelet coefficients after substitution to obtain a new first principal component of $f_t(x, y)$.
(6) Perform the inverse PCA operation for the new first principal component and other components of $f_t(x, y)$ to obtain the fused image.

Figures 5(a) and (b) show the fused images obtained from fusing Figures 3(a) and (b) with PCA-based fusion and wavelet transform fusion, respectively (Bian, 2005). Figure 5(c) shows the fused images obtained by using the combined method. It is seen that the new fused result enhances the texture property, enriches the spectrum information, as well as makes the object contour more visible.

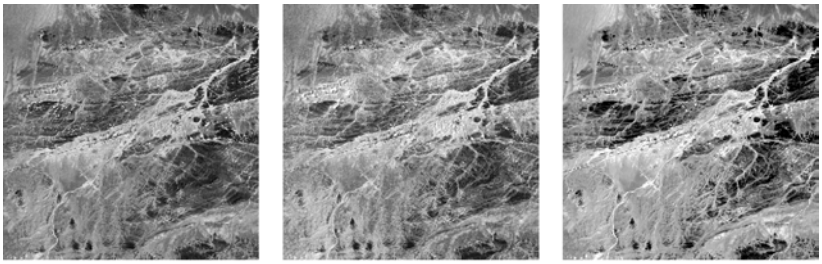
FUTURE TRENDS

The basic fusion techniques use only static images; such an analysis constitutes an ill-posed, under-determined problem. A new paradigm of "active, qualitative, purposive" vision has been introduced (Andreu, 2001). With the goal of effective fusion in mind, multiple moving images and suitable analysis

Figure 4. Fused image with the combination of HSI transform and wavelet trans

(a) (b) (c)

Figure 5. Fused image with the combination of PCA and wavelet transform



(a) (b) (c)

are employed. The uncertainty problem will be well defined. Further researches in this direction should be continued.

The sensor model plays an important role in multi-sensor image fusion. It should not only describe sensors' own properties, but also describe the influence of exterior conditions to sensor and the ability of cross actions among different sensors.

Finally, the scope of fusion can be further extended from the three layers described. For example, to the low end by considering the detection of signals, and to the high end by considering the situation state and tendency.

CONCLUSION

In the above, the principles of multi-sensor information fusion, especially of multi-sensor image fusion, are introduced. Some basic techniques used in multi-sensor image fusion are presented with real applications.

Multi-sensor information fusion is a basic ability of human beings. In mimicking such ability by computer, the fuzziness in perceiving environment could be reduced, the reliability of making decision could be increased, and much more completed problems could be solved. Except the examples for the fusion of remote sensing images here, the fusions of visible-light image and infrared image, CT image and PET image also have provided to be successful. It is expected that this technology will be applied in more application areas.

ACKNOWLEDGMENTS

This work has been supported by the National Natural Science Foundation under Grants NNSF-60573148.

REFERENCES

Andreu J. P., Borotsching H., Ganster H., etc. (2001). Information Fusion in Image Understanding. Digital Image Analysis – Selected Techniques and Applications. Springer.

Bertero M., Boccacci P. (1998). Introduction to Inverse Problems in Imaging. IOP Publishing Ltd.

Bian H., Zhang Y. J., Yan W. D. (2005). The study on wavelet transform for remote sensing image fusion. Proc. FJCSIP, 109-1139.

Brakenhoff G. J., Blom P., Barends P. (1979). Confocal scanning light microscopy with high aperture immersion lenses. Journal of Microscopy, 117: 219-232

Buckley F., Lewinter M. (2003). A Friendly Introduction to Graph Theory. Pearson Education, Inc.

Committee on the Mathematics and Physics of Emerging Dynamic Biomedical Imaging. (1996). Mathematics and Physics of Emerging Biomedical Imaging. National Academic Press.

Kropatsch W. G., Bischof H., eds. (2001). Digital Image Analysis – Selected Techniques and Applications. Springer.

Li R., Zhang Y. J. (2005). Automated image registration using multi-resolution based Hough transform. SPIE, 5960: 1363-1370

Luo Z. Z., Jiang J. P. (2002). Matching Vision and Multi-information Fusion. China Machine Press

Pajares G. (2004). A wavelet-based image fusion tutorial. Pattern Recognition 37: 1855~1872

Piella G. (2003). A general framework for multiresolution image fusion: from pixels to regions. Information Fusion, 4: 259-280

Polhl C., Genderen J. L. (1998). Multisensor image fusion in remote sensing: Concepts, methods and applications. International Journal of Remote Sensing, 19(5): 823-854

Shapiro L., Stockman G. (2001). Computer Vision. Prentice Hall.

Zhang Y. J. (2006). Image Engineering (I): Image Processing. 2nd ed. Tsinghua University Press.

Zhang Y. J. (2007). Image Engineering (III): Image Understanding. 2nd ed. Tsinghua University Press.

KEY TERMS

Bayesian fusion: A probabilistic method for fusing information from different sensors. It can be used for feature level fusion and decision level fusion.

Evidence reasoning fusion: A new method for fusing information from different sensors. It also called D-S theory. It can be used for feature level fusion and decision level fusion.

Fusion based on rough set theory: A fusion method for decision level. Instead of exact set, it uses rough set to manipulate sensor data. It can compress redundant information so avoid the composition exploitation problem during fusion procedure.

Image Engineering: An integrated discipline/subject comprising the study of all the different branches of image and video techniques. It mainly consists of three levels: Image Processing, Image analysis, Image understanding.

Information fusion: Combined process of information from the source of same object or scene to obtain more complex, reliable and accurate information.

Objective evaluation of image fusion results: To judge the quality of image fusion with some computable metrics based on fusion results.

Sensor model: An abstract representation of the physical sensors and its information manipulation process.

Subjective evaluation of image fusion results: To judge the quality of image fusion with subjects' perception on fusion results.

From Information Management to Knowledge Management

Călin Gurău
GSCM – Montpellier Business School, France

INTRODUCTION

The continuous evolution of theory and practice has modified the existing organizational paradigms and has introduced new models which attempt to explain how information is created, transmitted, used, and managed within various organizations. Many authors have outlined the fact that information no longer represents the most important asset of a firm. In the present competitive conditions, the managers must also consider knowledge and its relationship with enterprise information systems (Gray & Densten, 2005; Jorna, 2002; Nonaka & Takeuki, 1995).

Using both a theoretical and empirical approach, this study attempts to investigate the implication of a new paradigm of *knowledge management* on an organization's structure and functioning, considering knowledge management in direct relation with data management and information systems. This article shows, using two organizational examples, that the development of effective *knowledge management* systems requires a well-organized information system, as well as the clear identification of the main knowledge and decision-making centers within the business organization.

After briefly defining the concepts of *information management* and *knowledge management,* the article presents a comprehensive literature review of the academic and professional publications that investigate the inter-relationship between these two organizational functions. Based on this secondary information, we propose a model that integrates both information and knowledge management systems, and provides an analysis of two UK business firms in order to illustrate the integration between these elements.

BACKGROUND

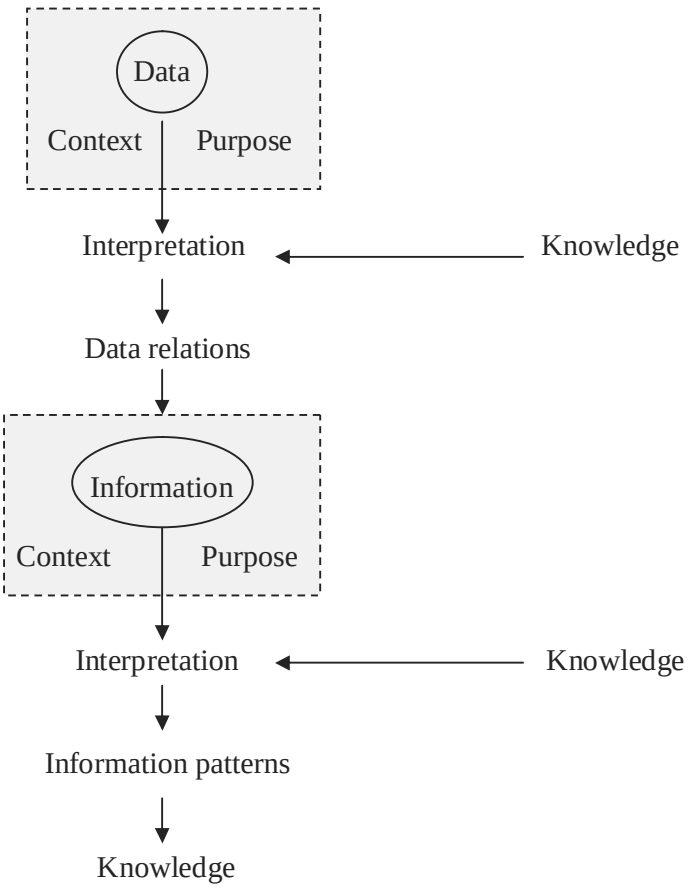
Before considering the research made on the relationship between *information management* and *knowledge management,* it is important to understand clearly the meaning of concepts such as data, information, and knowledge, and the progression from one to another within an organization.

A simple collection of data does not represent information, and equally, a simple collection of information cannot be considered as knowledge. An isolated datum has no meaning, and a collection of randomly combined isolated data is even more confusing (Schreiber et al., 2000). In order to transform a data collection into information, a person or a system must order the data, applying a specific interpretative pattern, which is determined by the context and the objectives of data analysis. Through the application of this interpretative pattern, specific relations among the collected data are discovered and defined, which transforms data in information, but only for a specific context and purpose (Bellinger, 2004). When the resulting information is ordered and interpreted in a specific context and with a specific purpose, patterns can be identified and defined as knowledge (Bellinger, 2004). Considering this transformation of data in information and then in knowledge, it is possible to draw a descriptive model (see Figure 1). It is interesting to note that in order to properly interpret the data and then the information, certain information patterns (knowledge) must be applied which create a dynamic cycle of knowledge creation and application within organizational systems.

However, this model is still too simplistic for several reasons. First of all, the knowledge used to define interpretation rules might not be created inside the organization, but rather acquired and transferred from outside (e.g., from a consulting firm), and it might be completely different from the knowledge resulting as an output of the entire process of interpretation.

Secondly, knowledge can be of different types (Wilson, 2002). Nonaka and Takeuki (1995) identify two types of knowledge—tacit and explicit knowledge—the first being derived from the second. On the other hand, Jorna (2002) defines three types of knowledge that are integrated into a dynamic model (van Heusden, & Jorna, 2001): (a) tacit or perceptual knowledge, (b) coded knowledge, and (c) theoretical knowledge. Perceptual knowledge is based on the perception of a specific difference in the environment, which allows one to identify and become aware of a specific situation or context (perceived as a pattern). Jorna (2002) considers this type of knowledge as uni-dimensional. The step towards coded knowledge is realized when the perceiver identifies a specific relation between recognized events or processes. This type of knowledge is defined as bi-dimensional. Coded knowledge is easier to communicate, because it can be represented and reproduced using specific signs (e.g., letters, mathematical operators, symbols, etc.). Finally, knowledge becomes theoretical when coded signs relate to the events represented not on a basis of a convention, but on the basis

Figure 1. The progressive transformation of data in information and information in knowledge



of formal or structural qualities (Jorna, 2002). At this level, abstract signs can be used as knowledge operators, in order to predict the development and evolution of real events or processes (e.g., scientific formulas).

Depending on the purpose and utility of knowledge, Zack (1999) classifies knowledge as:

a. *declarative knowledge* knowledge about something;
b. *procedural knowledge* knowledge of how something occurs or is performed; and
c. *causal knowledge* knowledge of why something occurs.

On the other hand, considering the specific subject/object or the form of knowledge, Lemken, Kaler, and Rittenbruch (2000) identify:

a. *tacit knowledge* you know it but you can't say it;
b. *experience-based knowledge* physical experience;
c. *coded knowledge* still available when people leave;
d. *conceptual knowledge* cognitive ability, abstraction;
e. *social knowledge* shared knowledge, culture, groups;
f. *event knowledge* events and trends; and
g. *process knowledge* operations and context.

Thirdly, the knowledge is mainly connected with people, and therefore human resource management is considered one of the main tracks of *knowledge management* (Parise, 2007; Sveiby, 2001), together with information systems. Choi and Lee (2003) make a clear distinction between a system-oriented and a human-oriented approach in knowledge management. System orientation emphasizes codified knowledge, focuses on codifying and storing knowledge via information technology, and attempts are made to share knowledge formally. On the other hand, human orientation emphasizes dialogue through social networks and person-

to-person contacts, focuses on acquiring knowledge via experienced and skilled people, and attempts are made to share knowledge informally (Sorge & Warner, 2001).

The management of various forms of knowledge within organizations is the subject of many recent studies (Gray & Densten, 2005; Jorna, 2002). One main area of research is the way in which knowledge can be created, managed, and distributed in complex business organizations.

Zack (1999) indicates four primary resources that are used for the management of explicit knowledge: (1) repositories of explicit knowledge; (2) refineries for accumulating, refining, managing, and distributing knowledge; (3) organization roles to execute and manage the refining process; and (4) information technologies to support these repositories and processes.

Although this research framework has its merits, allowing an analytical investigation of the various components of knowledge management systems, the approach neglects the role of the basic building blocks of knowledge: data and information. Considering the three main elements presented in Figure 1—data, information, and knowledge—it becomes obvious that a clear vision of the organizational knowledge system must also take into account the relationship among data, information, and knowledge, and consider the role of various employees in these three inter-related systems (Jorna, 2002).

Investigating Data Systems Inter-Relations within Business Organizations

In order to understand the relationship between data, information, and knowledge, two case studies of active UK-based companies have been investigated. The firms have been contacted as a result of previous collaborative work with the researcher, and invited to participate in this study. The first company (X) is a small business consulting firm, and the second one is a medium-sized biotechnology company (Y). Both firms are located in London and have more than 20 years of working experience. For confidentiality purposes, the names of the analyzed firms were not disclosed in this study.

The investigation of these organizational systems has been realized using both secondary and primary data. First of all, the public presentation of these companies, their annual reports and press releases, have been analyzed in order to understand organizational profiles of the firms, their portfolios of activities, and their main marketing strategies. Secondly, a series of face-to-face interviews, lasting between 45 and 90 minutes, have been conduced with the managers and the employees of these firms, regarding the organization and management of knowledge within the firm, as well as its connections with data and information systems. Finally, this information has been compared with the data collected during a direct observation of the internal functioning of these firms. The data collected was analyzed using simple qualitative techniques, such as the identification of the main representations and narratives of respondents about the organizational knowledge system.

The Relation Between Data, Information, and Knowledge Management

The analysis of these companies indicated both a number of common and specific elements concerning the relationship between data, information, and knowledge systems. In terms of common findings, both organizations used data, information, and knowledge systems, which they recognized as essential for the good functioning and the performance of the firm. The firms used a combination of internal and external data and information. Two types of knowledge have been identified as essential: routine organizational knowledge and creative organizational knowledge. *Routine knowledge* can be defined as knowledge with a high degree of stability which is applied repetitively for the good day-to-day functioning of the company. On the other hand, *creative knowledge* is applied to solve singular and/or unexpected situations that cannot be fully predicted and defined in advance.

There is no direct correspondence between routine knowledge and the coded knowledge defined by Jorna (2002), because in both companies the respondents recognized that some of the routine knowledge is implicit. Similarly, the creative knowledge is both codified and implicit, although the most substantial part is implicit, based on the personal professional experience and talent of employees.

Despite these common elements, the investigated companies present specific profiles in what concerns the predominance and the areas of application of routine and creative knowledge. Firms X collects or acquires both general market data and individual customer data, while the biotechnology firm focuses its data collection activities only on general market data. This approach is logical considering that the relation of the first firm with customers is direct and personal. On the other hand, company Y develops health therapies for a global consumer segment, and the clinical trials of its products are realized by a partner organization.

The routine knowledge is used by these organizations to process general market data and to interpret information in relation to a specific context and purpose. The resulting information is then transmitted to the centers of creative knowledge which develop and define the general strategy and the specific tactics of the firm. Sometimes, a drastic change in the market conditions determines a change of strategic orientation, which consequently has an impact on the routine knowledge (interpretation rules and objectives) used to process and interpret general market data and information.

Firm X

Figure 2 presents the relationship between data, information, and knowledge within company X. As can be seen, the main components of the company data system are the processes of data collection/acquisition, data archiving, and data processing—this last process being realized mainly using routine knowledge procedures.

The direct interaction between customers and consultants results also in data collection, which is processed in real time by a company's employees and transformed in information which is then interpreted using creative knowledge procedures, in order to provide customized and effective consulting solutions. On the other hand, the customer-consultant interaction enriches the professional experience of the company's personnel, enhancing the available creative knowledge.

At an organizational level, the information interpreted though routine and creative knowledge represents an input for the *strategic decision-making* center, which is using the received information to define the strategic orientation of the firm and the tactical day-to-day operations. These decisions are consequently enriching/modifying both the routine and the creative knowledge of the firm. The presence of various *double-loop connections* among data, information, and knowledge systems indicates the main *organizational learning* processes, which result in a dynamic evolution of company's knowledge (Antonacopoulou & Chiva, 2007; Engeström, 2007).

Firm Y

The connection among the three internal organizational systems is different in the case of firm Y (see Figure 3). The various information resulting from data processing represents an input for the operational block of the company, which is composed of three main departments: Research and Development (R&D), Production, and Commercialization. On the other hand, internal or external (outsourced) creative knowledge is also involved in the functioning of these departments. An exact representation of data, information, and knowledge systems for each of these three departments would present, in fact, a similar structure to the general model, with department-specific operations of data collection, archiving, and processing; with information processing mechanisms that apply department-specific routine and creative knowl-

Figure 2. The inter-relations among data system, information system, and knowledge system in company X

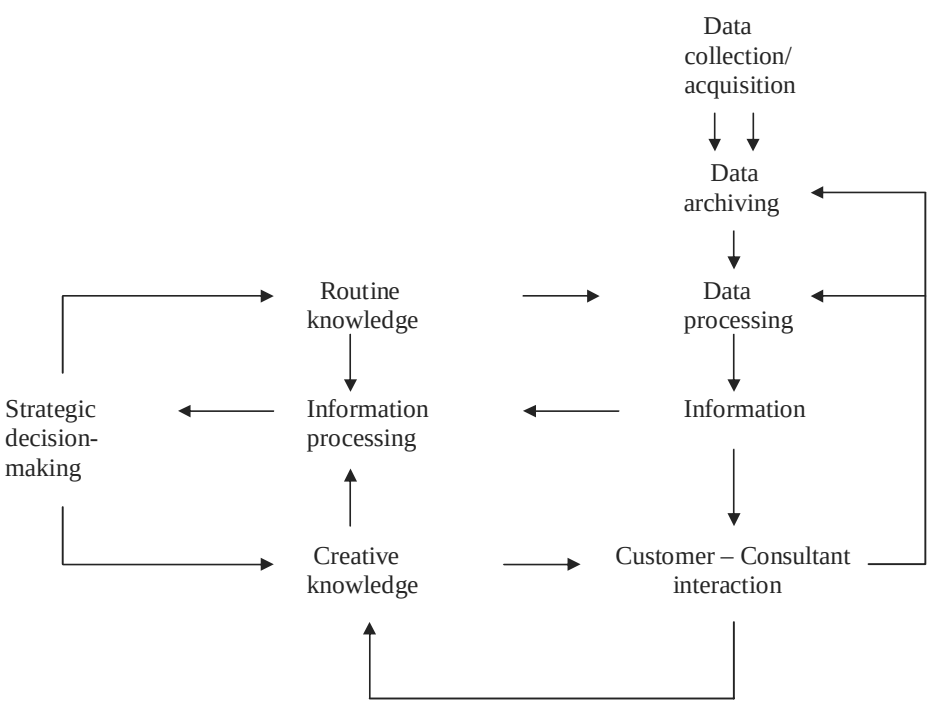
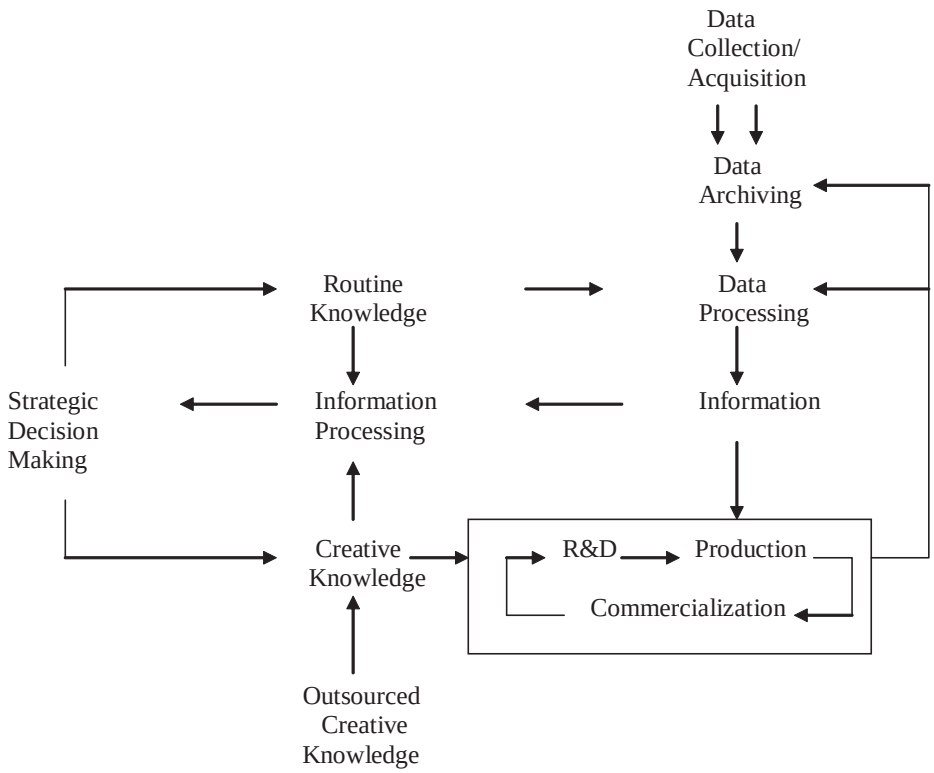


Figure 3. The inter-relations among data system, information system, and knowledge system in company Y



edge; and with strategic decision making, which influences the strategic orientation and the operational tactics of each department.

The complexity of this model is further enhanced by the fact that each department has connections with the other two departments in terms of exchange of data, information, and knowledge, as well as with the general structure and functioning of the organization. As in the previous case, the closed loops indicate an organizational learning process, functioning at a specific level within the company.

FUTURE TRENDS

The 'knowledge management' science and the connection of organizational knowledge with enterprise data and information systems represent very recent areas of research and analysis. The increased awareness that knowledge is the main asset of the company has forced management to transform the classical organizational paradigm of information systems/information technology in a complex model centered on various types of knowledge.

However, despite the increased professional and academic interest in this area, much progress is still to be done. The business modeling approach used in this study provides a clear representation of the inter-relationship among various organizational systems, but needs to be complemented by a clear identification of the main sources of routine and creative knowledge, and with an analysis of various roles and competencies that must be fulfilled by a company's personnel in the functional architecture of the company. On the other hand, researchers should explore the possibility of developing quantitative models of the relationship between data, information, and knowledge within the organizational system.

The connection between data, information, and knowledge flows should also be analyzed, in order to identify the possible blockages in the system, by applying simulation and

modeling procedures. The efficient transformation of data in information and then in knowledge, as well the effective application of knowledge in all organizational departments and processes, can significantly enhance the organizational learning process, which can provide an invaluable competitive advantage in a dynamic market environment (Cader, 2007; Reich, 2007).

CONCLUSION

Starting from a general discussion about the importance of data, information, and knowledge systems for the modern business organization, this study has developed and presented a basic graphical representation of data, information, and knowledge flows within two real companies that provided empirical data for this project.

The modeling approach that has been used has a number of advantages, permitting a clear representation of various organizational processes, and in the case of inter-organizational analysis, an easy comparison of various company structures. However, it also has a number of important limitations, offering only a general representation of the company's functional architecture.

The analysis of two different enterprises indicates that although the general representation of data, information, and knowledge flows is relatively similar, the level of complexity might vary from one enterprise to another. On the other hand, previous studies have indicated that the structure and the functioning of the organizational knowledge system might also be influenced by the specific corporate culture: adhocracy, clan, hierarchy, and market (Gray & Desten, 2005; Jorna, 2002; Lemken et al., 2000; Stoica, Liao, & Welsch, 2004). Future projects can identify additional characteristics that might determine the shape and functioning of the internal knowledge system, such as the company' stage of development, its size, or the level of technology/innovativeness that is integrated in its products and services.

The findings of this study can represent the basis for future research projects. The analysis of various data, information, and knowledge flows must be complemented by an investigation of the human resource management and technology used by the organization. For each specific process identified in the model, the enterprise should identify the main personnel capacities and the technological platform required to ensure an effective functioning of the system. Ultimately, these requirements should be compared with the real situation of the enterprise in order to map the main organizational gaps between the necessary and the existing level of resources.

REFERENCES

Antonacopoulou, E., & Chiva, R. (2007). The social complexity of organization learning: The dynamics of learning and organizing. *Management Learning, 38*(3), 277-295.

Bellinger, G. (2004). *Knowledge management emerging perspectives.* Retrieved November 2006 from http://www.systems-thinking.org/kmgmt/kmgmt.htm

Cader, Y. (2007). Knowledge management and knowledge-based marketing. *Journal of Business Chemistry, 4*(2), 46-58.

Choi, B., & Lee, H. (2003). An empirical investigation of KM styles and their effect on corporate performance. *Information & Management, 40*(5), 403-417.

Engeström, Y. (2007). From stabilization knowledge to possibility knowledge in organizational management. *Management Learning, 38*(3), 271-275.

Gray, J.H., & Densten, I.L. (2005). Towards an integrative model of organizational culture and knowledge management. *International Journal of Organizational Behaviour, 9*(2), 594-603.

Jorna, R. (2002). Organizational forms and knowledge types: A cognitive multi-actor approach. *Australasian Journal of Information Systems, 10*(1), 29-40.

Lemken, B., Kahler, H., & Rittenbruch, M. (2000, January 4-7). Sustained knowledge management by organizational culture. *Proceedings of the Hawaii International Conference on System Sciences,* Maui, HI.

Nonaka, I., & Takeuki, H. (1995). *The knowledge creating company: How Japanese companies create the dynasties of innovation.* Oxford: Oxford University Press.

Parise, S. (2007). Knowledge management and human resource development: An application in social network analysis methods. *Advances in Developing Human Resources, 9*(3), 359-383.

Reich, B.H. (2007). Managing knowledge and learning in IT projects: A conceptual framework and guidelines for practice. *Project Management Journal, 38*(2), 5-17.

Schreiber, G., Akkermans, H., Anjewierden, A., de Hoog, R., Shadbolt, N., van de Velde, W., & Wielinga, B. (2000). *Knowledge engineering and management: The CommonKADS methodology.* Cambridge: MIT Press.

Sorge, A., & Warner, M. (Eds.). (2001). *The IEBM handbook of organizational behaviour.* London: Thomson Learning.

Sveiby, K.E. (2001). *What is knowledge management.* Retrieved September 2006 from http://www.sveiby.com/faq.html#Whatis

Van Heusden, B., & Jorna, R. (2001). Toward a semiotic theory of cognitive dynamics in organizations. In K. Liu (Ed.), *Organizational semiotics.* Amsterdam: Kluwer.

Stoica, M., Liao, J., & Welsch, H. (2004). Organizational culture and patterns of information processing: The case of small and medium-sized enterprises. *Journal of Developmental Entrepreneurship, 9*(3), 251-266.

Wilson, T.D. (2002). The nonsense of 'knowledge management'. *Information Research, 8*(1). Retrieved February 2007 from http://informationr.net/ir/8-1/paper144.html

Zack, M.H. (1999). Managing codified knowledge. *Sloan Management Review, 40*(4), 45-58.

KEY TERMS

Adhocracy Culture: An organizational culture in which various groups of individuals reach consensus by responding in an ad hoc fashion to frequently changing priorities.

Clan Culture: An organizational culture that emphasizes the internal maintenance organizational structures and processes, using flexibility, concern for people, and sensitivity toward customers.

Explicit Knowledge: Knowledge that is coded and cataloged in order to facilitate its use by people.

Hierarchy Culture: An organizational culture that focuses on the development and maintenance of stable organizational rules, structures, and processes, by implementing a hierarchical system of power and management.

Implicit Knowledge: Knowledge that is contained in people's 'modes', and based on their personal experience and insight.

Knowledge Management: A central function of an organization that contains a number of rules and processes applied so to comprehensively collect, organize, share, analyze, and distribute knowledge in order to maximize the organizational performance.

Market Culture: An organizational culture that attempts to achieve market performance using internal processes that emphasize stability and control.

Information Project Assessment by the ANDA Method

Alexandru Tugui
"Alexandru Ioan Cuza" University, Romania

INTRODUCTION

It is well known that the decision to invest is preceded by (pre-)feasibility analysis and studies, which should show that the investment is necessary, opportune, and efficient. As concerns the information field, we consider this feasibility study practice as partially adequate, as there is no full financial assessment of the *necessity* and *opportunity* of the information project by taking into account its advantages/disadvantages, the studies presenting only a list of these advantages/disadvantages. Moreover, given the impossibility of determining the contribution of each information function to the economic efficiency indicators of the organization (turnover, profit, etc.), the final result is also a partial assessment of the efficiency of the *information project*, which may lead to bad influences on the managerial team when making the decision of investing or not in that project. This is why we proposed in this study *a new method of financial assessment of the necessity and opportunity of the information project*, namely the annual net discounted advantages method or, more simply, the **ANDA method**). This chapter also includes, besides theoretical aspects, a case study with a concrete application of the ANDA method to an information project.

We mention that the ANDA method may be applied to the assessment of the efficiency and of the opportunity of any investment project of modernization in any business area.

BACKGROUND

Modern management lays increasing emphasis on the organization on *projects* of investment activities and their funding, by means of the *budget technique*.

Briefly, a project is the first decomposition of a program, with a beginning and an ending, including a series of logically chained and efficiently planned actions/activities, which are supposed to be carried out in a certain period of time, for which a financial support is associated/assigned by budget, aimed at achieving one or a set of objectives (Belanger, 1995; Devaraj & Kohli, 2002; Hayes, 1989; Lewis, 2000; Lientz & Rea, 1999; Mantel, Meredith, Shafer, & Sutton, 2001; Oprea, 2001; Project Management Institute [PMI], 1996; Sages Group [SG], 1997; Țugui & Fătu, 2004).

The previous definition leads to the idea that a project should include:

1. A (logic) schemata for running the established activities depending on the objectives undertaken;
2. An adequate financial support, that would enable the mobilization of the resources necessary for objective achievement;
3. A project administration team with a project manager that should coordinate the necessary procedures like a traditional manager.

The three items required for objective/objectives achievement within a project, support the project management concept, which has been thoroughly developed and widely used lately.

Basically, project management is defined as being the application of the knowledge and techniques specific to the project-oriented activities, so that the stakeholders' expectations and requirements should be attained or even overcome (Beise, Neiderman, & Mattord, 2006; Kern, Galup & Nemiro, 2000; Laudon & Laudon, 2000; PMI, 1996; Sisco, 2001; Walker, 2001).

Information project is a particular case of the project concept. As to the information-related activities in an organization, we can state that the project which reassembles them is an information project within a program run by such organization.

For a project to succeed, it is necessary to perform all the tasks assigned to the project management team resulting from the objective set, so the project management work becomes of primary importance.

Irrespective of the type of project considered for financial efficiency and opportunity analysis, the following information is necessary (Țugui & Fătu, 2004, p. 139):

- The initial investment value;
- The future cash flows;
- The analysis period; and
- The advantages and disadvantages as to the scenario "without project".

Please note that decisions related to information projects rely on feasibility studies stating their necessity, opportunity, and efficiency.

ANDA METHOD: THE METHOD OF ANNUAL NET DISCOUNTED ADVANTAGES

Motivation of ANDA Method

The assessment of the *necessity*, *opportunity*, and *efficiency* of an investment project is a stage of the utmost importance in any economic entity.

Usually, the *necessity* and *opportunity* of an investment project is justified by a description, under the form of a list, of the technical-economic advantages or disadvantages, without an *integrated financial value estimation*, under the form of impact studies, market surveys, and so forth, while for economic *efficiency*, the information included in the previous studies under the form of business plans, cost-profit analysis and, finally, feasibility studies, is molded.

As for the assessment of the efficiency of investment projects, there are specific indicators, such as: internal rate of return (IRR or ROI), payback time, net present value (NPV), and so forth. All these indicators apply to future cash flows for each project. In Table 1, we present the work schemata to obtain the analysis indicators of the efficiency of an investment project.

There is also the practice of justifying the *necessity* and *opportunity* of investment projects through sensitivity analysis of *efficiency indicators* between the scenario "with project" and the scenario "without project". The disadvantage of this comparison consists of the fact that it does not provide actual indicators to estimate the necessity and opportunity of the project, but only some differences that should be interpretation and supported.

***Note*:** The scenario "*without project*" is the variant in which the analyzed investment project is not applied, while the scenario "*with project*" is meant to commission the analyzed project.

Thus, the building of a set of economic-financial assessment indicators is justified for the assessment of the necessity and opportunity of an investment project.

The Essence of ANDA Method

As seen earlier, there is no set of economic-financial assessment indicators to assess the necessity and opportunity of an investment project, should the organization management request such information.

Under such circumstances, in order to assess the *necessity* and *opportunity* of an investment project we propose the employment of the **method of annual net discounted advantages (ANDA)**.

The essence of the *ANDA method* consists in substituting the cash flows with net advantages within the method of discounted cash flows (DCF) in order to calculate the indicators

Table 1. Model of data organization for the assessment of the efficiency of an information project. Note: IRR: Internal rate of return; NPV: Net present value; Vr: Residual value; ra: Discounted rate

Explanations	**Years**					
	0	**1**	**2**	**3**	**4**	**5**
Investment value (I_1)	I_1					
Net cash flows (F)		F1	F2	F3	F4	F5+Vr
Discount factor		$(1+ra)^{-1}$	$(1+ra)^{-2}$	$(1+ra)^{-3}$	$(1+ra)^{-4}$	$(1+ra)^{-5}$
Net discounted flows		$F_1*(1+ra)^{-1}$	$F_2*(1+ra)^{-2}$	$F_3*(1+ra)^{-3}$	$F_4*(1+ra)^{-4}$	$(F_5+Vr)*(1+ra)^{-5}$
Cumulated net discounted flows		$F_1*(1+ra)^{-1}$	$F_1*(1+ra)^{-1}+F_2*(1+ra)^{-2}$	$F_1*(1+ra)^{-1}+F_2*(1+ra)^{-2}+F_3*(1+ra)^{-3}$	$F_1*(1+ra)^{-1}+F_2*(1+ra)^{-2}+F_3*(1+ra)^{-3}+F_4*(1+ra)^{-4}$	$F_1*(1+ra)^{-1}+F_2*(1+ra)^{-2}+F_3*(1+ra)^{-3}+F_4*(1+ra)^{-4}+(F_5+Vr)*(1+ra)^{-5}$
Payback period	The investment value is compared to the cumulated net discounted flow. The year that overruns the investment value is the payback period.					
NPV	$-I+\sum F_i*(1+ra)^{-i}+Vr*(1+ra)^{-i}$ i takes values from 1 to 5					
IRR	For NPV = 0, ra is determined $-I+\sum F_i*(1+ra)^{-i}+Vr*(1+ra)^{-i}=0$					

Table 2. Data organization model for an information project opportunity assessment. Note: IRR: Internal rate of return; NPV: Net present value; Vr: Residual value; ra: Discounted rate

Explanations	**Years**					
	0	**1**	**2**	**3**	**4**	**5**
New Investment Value (I_1)	I_1					
Residual value of Old Project (Vro)	Vro					
Net New Investment Value (I_{1net})	$(I_1 – Vro)$					
Residual Value of New Project (Vr)						Vr
Net advantage /disadvantage (A)		A1	A2	A3	A4	A5+Vr
Discount factor		$(1+ra)^{-1}$	$(1+ra)^{-2}$	$(1+ra)^{-3}$	$(1+ra)^{-4}$	$(1+ra)^{-5}$
Net discounted advantage/disadvantage		$A_1*(1+ra)^{-1}$	$A_2*(1+ra)^{-2}$	$A_3*(1+ra)^{-3}$	$A_4*(1+ra)^{-4}$	$(A_5+Vr)*(1+ra)^{-5}$
Cumulated net discounted advantage/disadvantage		$A_1*(1+ra)^{-1}$	$A_1*(1+ra)^{-1}+$ $A_2*(1+ra)^{-2}$	$A_1*(1+ra)^{-1}+$ $A_2*(1+ra)^{-2}+$ $A_3*(1+ra)^{-3}$	$A_1*(1+ra)^{-1}+$ $A_2*(1+ra)^{-2}+$ $A_3*(1+ra)^{-3}+$ $A_4*(1+ra)^{-4}$	$A_1*(1+ra)^{-1}+$ $A_2*(1+ra)^{-2}+$ $A_3*(1+ra)^{-3}+$ $A_4*(1+ra)^{-4}+$ $(A_5+Vr)*(1+ra)^{-5}$
Payback period	The investment value is compared with the cumulated net discounted advantage. The year that overruns the investment value is the payback period.					
NPV	$- I_{1net} + \sum A_i*(1+ ra)^{-i} + Vr*(1+ ra)^{-i}$ i assumes values from 1 to 5					
IRR	For NPV = 0, ra is determined $- I_{1net} + \sum A_i*(1+ ra)^{-i} + Vr*(1+ ra)^{-i} = 0$					

IRR, NPV, and so forth. Hence, this method requires having run the following stages:

1. **Value assessment of annual net advantages/disadvantages**, that an information resource investment brings, as they are presented in Table 1;
2. **Discounting such advantages/disadvantages** upon the completion of the investment project;
3. **Calculation of the efficiency assessment indicators** based on the annual net discounted advantages/disadvantages.

In Table 2, we present the work scheme for obtaining the analysis and opportunity assessment indicators of an information project.

As shown in Tables 1 and 2, the difference between the DCF method and the ANDA method are:

1. the replacement of cash flows (Fi) by annual net advantages (Ai);
2. the replacement of investment value (I_1) by net investment value (I_{1net}), obtained by deducting from (I_1) the residual value of the old project (Vro), if it exists.

Figure 1 shows clearly that from the old project to the new one there are financial advantages/disadvantages, which we presented separately. It is obvious that these financial advantages/disadvantages are present in all the years of our analysis. Please note that most of the times, in the first year there may be greater advantages as compared to the old project, after which these advantages may diminish and/or become uniform in the following years.

Figure 2 shows the financial advantages/disadvantages between the two projects for a period of five years. Depending on the nature of the field we analyze, on the type of project funding, one may choose a shorter or longer period of analysis. However, in case of information projects, we advise that this period should not exceed five years, given the high obsolescence of information technologies.

Figure 3 presents the graph of the ANDA method, as described in Table 3.

The main strengths of the ANDA method for any type of project, including information projects, are:

1. The necessity and opportunity of a new project are also supported by figures;
2. There is a direct connection between the “old project” and the “new project”, or between the scenario “with project” and the one “without project”;

Figure 1. Comparison between the financial advantages/disadvantages of the new project and of the old project

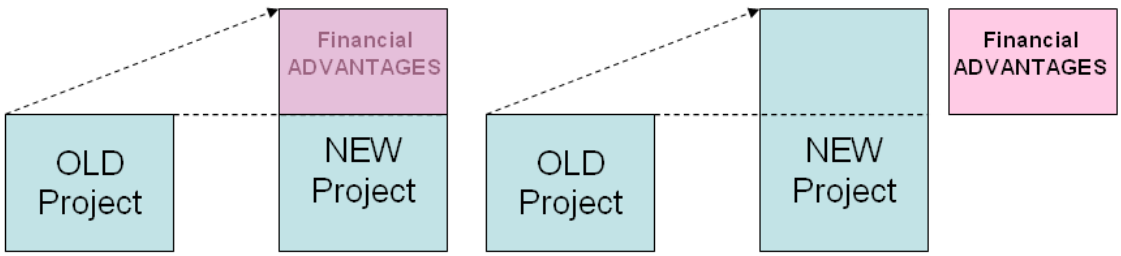


Figure 2. Five-year presentation of advantages/disadvantages Note: A1: 1st year's Financial Advantages

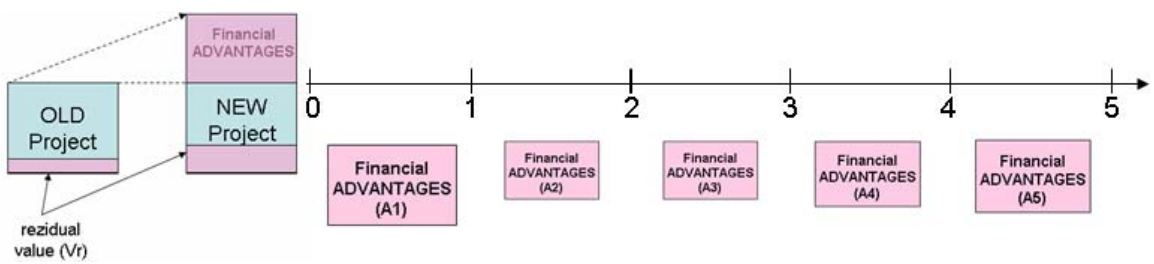


Figure 3. Graphic presentation of the ANDA method

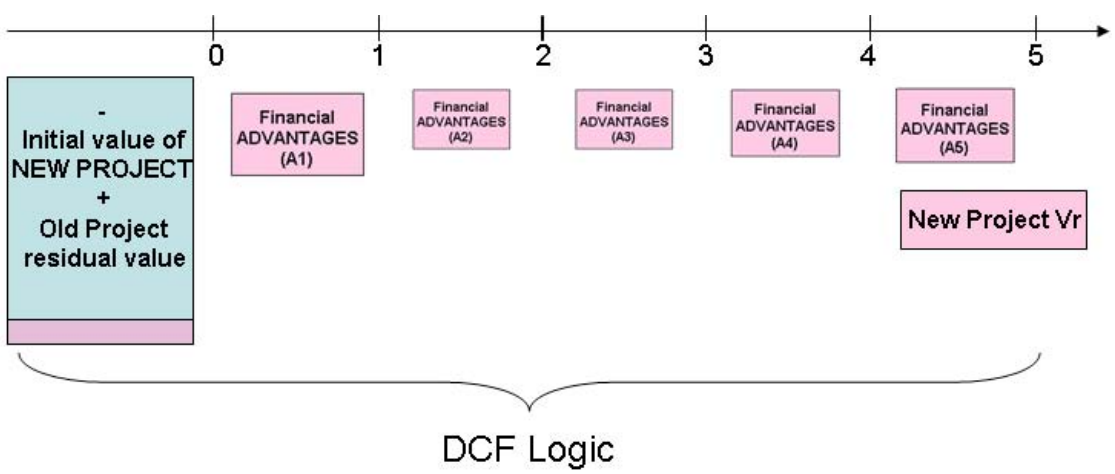


Table 3. Recommendations for investment project assessment

Indicators		Efficiency	
		Favorable	**Unfavorable**
Necessity Opportunity	**Favorable**	Investment project accepted.	The investment project may be accepted, but we advise the use of a more efficient technology.
	Unfavorable	The investment project should be rejected. We advise its later resuming.	Investment project rejected.

3. The same set of indicators as the ones used for assessing the financial efficiency of a project are employed;
4. It is an efficient tool for those who want to convince the organization management, including the financial and accounting department, of the necessity and opportunity of an investment project.

Here are some of the weaknesses of the ANDA method:

1. The indicators calculated by means of the ANDA method may have unfavorable values. If so, we advise its later resuming, maybe even finding new variants for the project, if the project efficiency indicators are favorable;
2. The difficulty of determining and quantifying advantages for all the years of analysis;
3. The residual value of the replaced project may be null or even negative;
4. The values of the indicators are lower if an old project is modernized.

Thus, the forecast economic data for the two variants, scenario "with project" and scenario "without project" are combined, which provides the grounds in figures to decision-makers. Moreover, we can see that the ANDA method achieves a hybrid between efficiency and opportunity in supporting the *financial opportunity*.

Table 3 contains a few recommendations depending on the favorable or unfavorable values pertaining to the efficiency assessment indicators and to the indicators used to assess the necessity and opportunity of an investment project.

A few aspects that we should consider in assessing a project opportunity are presented in the following:

1. If calculated efficiency indicators are positive, then the opportunity of the scenario "with project" may be supported, otherwise it is not appropriate, as the scenario "without project" has more advantages.
2. In case we have several new projects, the net advantages of each of them compared to the scenario "without project" will be assessed for the envisaged analysis periods, concomitantly to their discounting and the calculation of these indicators. Having compared the results, the project with the best indicators will be chosen, that is the project with the greater internal rate of return, with the shortest payback period and the greatest net present value.
3. If the indicators calculated on the basis of the new advantages presents lower values than those calculated on the basis of the cash flows afferent to the scenario "with project". This is the very fact that justifies the opportunity of resorting to the scenario "with project".
4. It is difficult to assess the values of certain qualitative advantages and disadvantages. In this case, we shall tackle categories of projects and types of decisions, the sources of advantages and disadvantages between the scenario "with project" and the one "without project". For information projects, we present in Table 2 a centralizing example regarding the main decisions on categories of resources.

INFORMATION-RELATED DECISION-MAKING

An *information project* may be regarded from the perspective of the time it is carried out as plain investment, but also from the perspective of the time it is carried out as continuous activity going on after the investment.

In all cases the information project may regard the automation of a function inside an organization or the re-engineering of an older computer investment, as well as all the activities that are meant to reorganize information-related processes.

As for information-related decisions, we are presenting an analysis starting from the information resources which are subject to these decisions. So we present in Table 5 the main decisions on categories of information resources.

In other words, an information project may be an investment (marked with +) or an disinvestment (market with -),

Table 4. Main information decisions on categories of information resources and their advantages/disadvantages

Explanations	Information resources			
	Human Resources	**Equipment**	**Software**	**Documentations/ Information**
Investment (+)	Training Continuous training Employment Motivation	Integral purchases Leasing Modernization Extension Maintenance	Purchase Own production Production in END USER logic Update Subscriptions	Purchase Internal development Update
Advantages/ Disadvantages	Training expenses Quality enhancement Continuous training expenses	Efficiency enhancement Information expense lowering Initial investment Operation indicators improvement Maintenance expenses	Initial investment Information expense cut Efficiency enhancement Update or subscription expense	Initial investment Time saving in documentation Update expenses
Disinvestments (-)	Externalizing Discharge Transfer	Externalizing Replacement Annulment Sale	Externalizing Replacement Sale	Replacement Sale
Advantages/ Disadvantages	Expense reduction Income reception Compensatory payments	Expense reduction Income from sale and annulment assignment	Expense reduction Income from license sale	Income from sale Expense reduction

under the conditions in which one may employ one, two or several types of information resources. For instance, the procurement of an information product also assumes the involvement of the human resources of the information department or of other departments, as the training and the licensing to use and maintain the information product is required.

When presenting the ANDA method, the determination of the financial implications of the advantages/disadvantages supporting the necessity and opportunity of a project is an extremely complex stage. Therefore, we present in Table 4 the main decisions to make by the information function of an organization and the related financial advantages and disadvantages.

Similarly to the processing in Table 4, we may analyze any investment field by its specialists.

Based on the analysis of Table 4, we may establish the value advantages and disadvantages attached to any computer-related decision that is meant for any resource.

CASE STUDY

We are providing the example of how to calculate the estimation indicators of the financial opportunity for an information project.

At a trade company, the computer system is replaced with a new one. This replacement is the subject of an information project that will be conducted over a month. To approve the investment project, a feasibility study is requested, by which its efficiency and opportunity are to be analyzed. From the data provided, we realize that the project is efficient because these indicators are positive. As to project opportunity, the company management requests its demonstration. To this purpose, we use the ANDA method. Hence, in order to assess the advantages that result by applying this project, the following financial information is available:

- Increase of customer handling rate by 5%, under the conditions of a profitability of 12%;
- Reduction of information expenses with turnover by 1.25%;
- Before-project turnover is of $400,000;
- Investment value is $42,000;
- The residual value of the replaced project is $5,500;
- The discounted rate afferent to the period is of 8%;
- Project assessment period is of 5 years;
- The residual value of the new project is of $7,500;
- The turnover increase annual rate is of 7%.

In Tables 5 and 6 we present how data are organized in order to appreciate the investment project opportunity as a worksheet is being used.

Table 5. Work hypotheses % or $

Explanations	Values
Prior turnover	400000
Increase of customer handling rate	5%
Information expense cut rate (% of turnover)	1.25%
Rate of return/general profitability (%)	12%
Recovery from the old project	5500
New project investment	42000
Discount rate	8%
Project assessment duration (years)	5
Turnover increase rate	7%
Residual value	7500

Table 6. Efficiency indicators % or $

Explanations	Values	Years				
		1	2	3	4	5
Investment value	42000					
Recovery from the old project	5500					
Net investment value	36500					
Residual value						7500
Project advantage calculation						
Additional profit	2400	2400	2568	2748	2940	3146
Expense reduction	5250	5250	5618	6011	6431	6882
Total advantages	7650	7650	8186	8758	9372	17528
Discount rate		10%	10%	10%	10%	10%
Discount factor		91%	83%	75%	68%	62%
Net discounted advantages		6955	6765	6580	6401	10883
Cumulated net discounted advantages		6955	13719	20300	26701	37584
Investment payback period	5 years					x
Net present value	1084					
Internal rate of return (IRR)						
Net investment value	-36500					
Year 1	7650					
Year 2	8186					
Year 3	8758					
Year 4	9372					
Year 5	17528					
IRR	11,01%					

Table 6 shows the financial opportunity of this project, because positive indicators are obtained only from the advantages of commissioning it.

FUTURE TRENDS

We consider that the ANDA we proposed may be extended for the project management of any area, being necessary to list the specific advantages and disadvantages. The advantages/disadvantages assessment requires a theoretical approach similar to that of Table 4 and a practical approach like in the previous case study.

The ANDA method enables the building of new indicators for the assessment of the necessity and appropriateness of an investment project, irrespective of the business area.

CONCLUSION

The assessment of an information project efficiency is a primary issue for the future information society. The speed of information equipment renewal is increasing sensibly, fact which requires that computer system managers should resort to the scenario and option theories.

The method we propose hands to the decision-maker a valuable tool in making information-related decisions.

At the same time, we are aware that ANDA is a radical method of both opportunity and efficiency assessment because if its indicators are positive, it results that the set of indicators calculated based on the discounted cash flows are positive. Hence, the ANDA method succeeds in harmoniously blending aspects that regard the efficiency and the opportunity of an investment project of any kind.

We believe that this method stands for further improvement especially in the field of advantages/disadvantages determination and assessment. This will be the subject of our further research.

REFERENCES

Beise, M. C., Neiderman, F., & Mattord. (2006). IT project managers' perceptions and use of virtual team technologies. In M. Khosrow-Pour, & N. Herman (Eds.), *Advanced topics in information resources management, vol. 5* (pp. 25-43). Hershey, PA: Idea Group Publishing.

Belanger, T. C. (1995). *Successful project management.* USA: American Management Association.

Devaraj, S., & Kohli, R. (2002). IT payoff, the measuring the business value of information technology investments. *Financial Times.* Prentice Hall. Retrieved September 20, 2004, from http://safari.informit.com

Hayes, E. M. (1989). *Project management.* California: CRISP Publication, Inc.

Kern, H., Galup, S., & Nemiro, G. (2000). *IT organisation*. Prentice Hall.

Laudon, K., & Laudon, J. (2000). *Management information system. Organisation and technology in the networked enterprise* (6th ed.). New Jersey: Prentice Hall.

Lewis, J. P. (2000). *The project manager's desk reference*. New York: McGraw-Hill.

Lientz, B. P., & Rea, K. P. (1999). *Guide to successful project management*. San Diego, CA, USA: Harcourt Brace Professional Publishing.

Mantel, S. J., Meredith, J. R., Shafer, S. M., & Sutton, M. M. (2001). *Project management in practice*. New York: John Wiley & Sons, Inc.

Oprea, D. (2001). *Project management* (pp. 10-35). Sedcom Libris, Iaşi.

Project Management Institute. (1996). *A guide to the project management body of knowledge.* USA. Retrieved September 30, 2006, from http://www.pmi.org/info/PP_OPM3ExecGuide.pdf

Sages Group. (1997). *Project management manual* (pp. 7-23), Government of Romania, DEI.

Sisco, M. (2001). IT manager development series. MDE Enterprises. Retrieved September 30, 2006, from www.mde.net

Ţugui, A., & Fătu, T. (2004). *Information resources management.* Sedcom Libris, Iaşi, Romania.

Walker, G. (2001). *IT problem management*. Prentice Hall. Retrieved September 20, 2004, from http://safari.informit.com

KEY TERMS

ANDA Method: This consists in substituting the cash flows with net advantages within the method of present value cash flows in order to calculate the indicators IRR, NPV, and so forth.

Financial Opportunity: In practical life, it is much simpler to convince a manager or a managing team of the opportunity of an investment project if the indicators calculated

by the ANDA method are positive because the project value can be redeemed only from the highlighted advantages.

Information Project: This is a particular case of the project concept. Information project may regard the computerizing of a function inside an organization or the modernizing of an older computer investment, as well as all the activities that are meant to reorganize information-related processes.

Opportunity/Efficiency Assessment: The opportunity and efficiency assessment of an investment project is a very important stage in any economic entity. As for the assessment of the efficiency of investment projects, there are specific indicators, such as: internal rate of return (IRR), payback time, net present value (NPV), and so forth.

Project: First decomposition of a program, with a beginning and an end, made up of a set of interlinked actions planned to be run over a certain time period, with its own budget established in order to reach a well-defined objective/objectives set.

Project Management: The application of the knowledge and techniques specific to the project-oriented activities, so that the stakeholders' expectations and requirements should be attained or surpassed.

Scenario Project: The scenario "without project" is the variant in which the analyzed investment project is not applied, while the scenario "with project" is meant to commission the analyzed project.

Information Resources Development in China

Maosheng Lai
Peking University, China

Xin Fu
University of North Carolina at Chapel Hill, USA

Liyang Zhang
Baidu.Com Co., Ltd., China

Lin Wang
Peking University, China

INTRODUCTION

In its several thousand years of social progress, China has put continuous effort into cultural development, which to a certain extent contributed to the exploitation and utilization of information resources. This article reviews the history and present situation of China's information resources development (IRD), with the focus on some IRD projects launched since the mid-1990s.

The specific projects that will be introduced include the China Academic Library and Information System, the China Digital Library Project, the construction of the China National Science and Technology Library, the China Online Government Project, and the construction of the National Institute for Information Resources Management. The goal of each project is described and its initial impact is discussed.

BACKGROUND

Since the founding of the People's Republic of China in 1949, the government has been attaching great importance to information resources development. In 1956, the government set "Marching Towards Science" as the directing principle for the course of information resources management, and made a conscientious plan in information resources development with emphasis on collecting, rearranging, analyzing, indexing, and reporting scientific and technical documents from home and abroad to serve the needs of professionals in various disciplines. By 1987, the scientific and technical information sector alone had already possessed 26,000 foreign periodicals, 6,000 domestic periodicals, 120 million patent manuals, and more than 32 million books. There were 236 abstracting and indexing journals published annually, covering more than 1.2 million documents and articles. Also, there were 2,038 public libraries at the county level and higher, collecting more than 200 million books. There were 745 academic libraries, collecting 250 million books. There were also more than 4,000 libraries at research institutes (Guan, 1988).

In the late 1980s and early 1990s, however, information resources development was affected by the readjusting of China's economy. Non-profitable libraries and information service institutions suffered from a severe shortage of money for collection development. As a result, information resources development was captured in a severe logjam or even retrogresses. Types of document collections in some libraries dropped by half or even two-thirds (Fu, 1996). Many abstracting and indexing journals stopped publication. But on the other hand, some new abstracting and indexing journals emerged, as did bibliographical databases that catered to market demand.

Under the promotion of the international information technology revolution, China has been experiencing an upsurge in information development since the last decade of the 20th century. Information infrastructure construction keeps a rapid pace in development. The ownership of telephones, cell phones, and computers has been increasing steadily. The overall scale of China's information infrastructure in terms of network capacity ranks first in the world (China Telecommunications, 2003; He, 2004), and the number of users ranks second in the world (CNNIC, 2006a). However, information resources development is lagging far behind. The lack of information, especially Chinese information, in networks and information systems influences the benefit of investment in information technology, which has become a major obstacle not only to China's informationalization drive, but also to the competitiveness of the Chinese economy.

Since the mid-1990s, under the promotion of the tide of information superhighway construction in many countries, information resources development in China entered a new phase. In 1997, the Chinese government constituted the "Draft on China's Informationalization," drawing the outline of China's information infrastructure (Zou, 1997), which includes six elements: information resources, national information network, information technology (IT) applica-

tion, information industry, information professional, and information policy code and standard.

Information resources was set as the primary element among the six, showing the state's emphasis on its development. This also indicated that people once again realized the importance of information resources development. Several years later, the proposal was accepted as a part of China's tenth "five-year plan," which marked that information resources development became the central task of China's informationalization drive. In China's eleventh "five-year plan," several parts mentioned the issue of information resources development and regarded it indispensable. For example, in Chapter 15, developing information resources is listed as a central task of advancing Chinese informationalization, which includes accelerating the construction of national fundamental databases; adjusting information resources structure; and strengthening the development, disposal, dissemination, and utilization of information resources. Information resources development is also mentioned in Chapter 16. In that chapter, several areas of information resources development, such as governmental and geographic domains, are emphasized.

At the forth meeting of General Office of State Council Informationalization Leading Group in 2004, the main agenda was information resources development. After the meeting, the General Office of the State Council together with the General Office of the CPC Central Committee issued the document "Some Suggestions on Strengthening the Work of Information Resources Development and Utilization." Until now, this file is still one of the most authoritative guiding policies of Chinese information resources development. It outlined the guidelines, basic principles, and general tasks of information resources development in the country. It also attached great importance to fostering the information resources market and industry.

The concept "Information Resources Development" used in this item refers to collection, processing, organization, and dissemination of document resources, as well as their digitalization and networking. Factual and data resources ought to be included in the concept. However, China's progress in these aspects is relatively slow. In recent years, people started to realize the importance of factual and data resources development. The departments concerned have started to work out a plan for constructing the National Data Center.

MAJOR INITIATIVES IN CHINA'S INFORMATION RESOURCES DEVELOPMENT

Under the guidance of the policies introduced in the last section, the Chinese government initiated several major information resources development projects to change the current situation of inconsistency between information resources development and information network construction, as well as to lessen the discrepancy between information resources available and that required by the public.

CALIS (China Academic Library and Information System)

CALIS is an initiative under China's plan to build 100 key universities in the 21st century (named "211 Project" by the Ministry of Education). It aims at constructing a networked information resources sharing system based on the China Education and Research Network (CERNET) so as to parallel the development of a communication network and an information resources network, and provide university faculty and students as well as professionals in research institutions with easy access to a national information service system that is characterized by abundant information resources, advanced technologies, and a convenient service system.

The service system consists of a CALIS national management center, four CALIS national information centers (covering sciences and social sciences, engineering, agricultural science, and medical science respectively), and seven CALIS regional information centers — in Beijing, Shanghai, Nanjing, Guangzhou, Xi'an, Chengdu, Wuhan, and Changchun. The system will be also linked to major information service systems outside China to form China's Academic Library and Information System. The construction of CALIS will greatly increase the amount of information available to academic libraries and also improve their capability in information services (data from *www.calis.edu.cn*).

Digitalization Projects

The China Digital Library Project was carried out under the coordination of the Ministry of Culture. In July 1997, the National Library of China (then Beijing Library), together with the Shanghai Library and a few other institutions, started the Chinese Pilot Digital Library Project (CPDLP). Later in 1998, the Ministry of Culture formally put forward the proposal of constructing the China Digital Library. A variety of enterprises and organizations — such as China Telecom, the National Library of China, the Chinese Academy of Sciences (CAS), the China Aerospace Industrial Corporation, Peking University, and Tsinghua University — participated in the project.

Together with the second phase of the Chinese National Library construction, the Chinese Digital Library Project was formally initiated at the end of 2004. The investment of the project was about 400 million RMB (US$40 million). It has 27 sub-projects belonging to four categories: the technical support environment construction, information resources construction, service system construction, and standards

construction. The whole project is set for completion before the 2008 Beijing Olympic Games, at which time, the Chinese National Digital Library will become the largest Chinese digital information preservation and service facility in the world (Fu, 2005a).

According to the plan, when the project is completed, the National Digital Library will have the information processing capability of:

- Digitalizing 300,000 paper-based documents per year,
- processing 90,000 bibliographic records per year,
- indexing 6,000 hours of video and audio resources,
- storing 200 million metadata records, and
- providing 100 million pages of full-text online.

It will also provide information resources with ancient Chinese characters, such as the bones-tortoise inscription database and Dun Huang database (Fu, 2005b).

Besides the China Digital Library Project, various other digital library projects were also carried out. The construction of Chinese National Science Digital Library (CSDL) started in late 2001. The project, as part of the Knowledge Innovation Project of the Chinese Academy of Sciences, aims to build a digital information service system that meets the international developing trends of digital libraries and caters to the development of the Chinese Academy of Sciences (Zhang, 2002). It should be able to serve the needs of researchers and professionals in information accessing and knowledge innovation.

The total investment is 140 million RMB (US$17.5 million). Until June 2005, CSDL has provided access to 128 scientific documents databases. It collected 13,000 e-books. CSDL service includes online cooperative cataloging, integrated cross-database searching and browsing, online interlibrary loan, and virtual reference. Two databases, the Science China Database and the subject portal, are only available through CSDL. The scope of the former covers almost 3,000 types of Chinese science and technological journals since 1986. The number of abstracts adds up to 1.2 million, still increasing at a rate of 200,000 per year. The number of citation data is 5 million, with a growth rate of 1.2 million per year (Zhang, 2005).

In China's Taiwan Province, eight digital library initiatives are currently underway, including the construction of a Digital Library and Information Center and the building of the Haoran Digital Library in Jiaotong University. The objects of the initiatives are to promote information exchanges among learning and research institutions in Taiwan, and coordinate their purchase of information resources such as databases from foreign countries. Another object is to promote the research on Chinese culture, especially on Chinese history (Lv, 1999).

There are also digitalization projects other than construction of digital libraries. In January 1999, the Geology Department of the Chinese Academy of Sciences raised to the State Council a proposal on strategies of China's "Digital Globe" development, indicating the importance of building a national Global Information Infrastructure and establishing a digital global spatial information sharing system (*Information Industry Newspaper*, November 22, 1999). In November 1999, the first "Digital Globe" International Conference was held in Beijing, showing that the Chinese government attached great importance to international cooperation in this area (*China Computer World*, December 6, 1999). Currently, the prototype of China's "Digital Globe," called the Digital Earth Prototype System (DEPS CAS1.0), has been built. It has been applied in several domains, such as digital tourism and digital Olympic.

Construction of the China National Science and Technology Library

In June 2000, the China National Science and Technology Library (NSTL, *www.nstl.gov.cn*) was formally established through the cooperation of China's Ministry of Science and Technology, the State Committee of Economics and Trade, the Ministry of Agriculture, the Ministry of Health, and the Chinese Academy of Sciences. As a virtual scientific and technical resource center, it consists of nine library and information institutions such as the Library of the Chinese Academy of Sciences, the National Engineering Library of China, the Library of the Chinese Academy of Agricultural Science, and the Institute of Scientific and Technical Information of China. The center utilizes advanced technologies and methods to collect information from domestic and foreign sources. Moreover, the center serves as a bridge of cooperation between Chinese information resources management professionals and their foreign counterparts (Yuan & Meng, 2001).

Special-Topic Information Resources Development

Government Information Resources Development—China Online Government Project

On January 22, 1999, the China Online Government Project Start-Up Conference was held in Beijing, sponsored by China Telecom and the State Economic and Trade Commission, together with the information sectors of more than 40 ministries. At the conference, the China Online Government Project was started.

According to the "White Paper on China Online Government Project," the project refers to the practice by the

government at all levels to establish formal Web sites to promote Office Automation in government work, offer public services via the Internet, and fulfill the roles of management and service in the fields of society, economy, and social life (*www.gov.cn*).

On January 1, 2006, the Web site *www.gov.cn* was formally launched. It has three versions: simplified Chinese, traditional Chinese, and English. Its main columns include China today, national institutions, rules and laws, and news release network service. It received 40.48 million hits from 337,000 users on its debut day. Through the Web site, the public is better informed and can monitor government operation more actively than before. It is believed that the Web site will make a positive impact on building a transparent, efficient, and clean government (BIAN, 2006).

Patent Information Resources Development

Patent information is an essential part of a country's technical information resources. To meet the users' requirement of searching and utilizing patent information, China Patent InfoNet was established by the Retrieving and Consulting Center under the State Intellectual Property Office in May 1998. In January, 2002, its new version (*www.patent.com.cn*) was published online and started to offer all-around services — such as patent information retrieval, introduction of patent laws and regulations, and guidelines for patent application — to patent users and researchers.

Construction of the National Institute for Information Resources Management

Three national research centers for information resources management have recently been set up in Beijing, Nanjing, and Wuhan to promote research on theories, applications, policies, and technologies in IRM; they are affiliated with the Department of Information Management of Peking University, the Department of Information Management of Nanjing University, and the School of Information Management of Wuhan University.

INITIAL IMPACT OF SOME INITIATIVES

CALIS

Started in November 1998, CALIS completed its first phase of construction by the end of 2001. Currently, the system can provide an online public access catalog, interlibrary loan (ILL), Internet navigation, online cataloging, cooperative literature purchasing, and various other services through digitalization of information resources, networking of information services, and cooperation among participating academic libraries. As a result of the first phase of construction, universities and colleges in China now possess greater information resources than ever before; for example, the variety of foreign periodicals increased by one-third, 95% of Chinese literature and 80% of foreign literature are now available, and more than 100 academic libraries offer 24-hour online information services. In addition, 25 distinctive databases and 194 disciplinary navigation databases are built.

In its second phase of construction starting in 2002, CALIS aimed to further strengthen the document supporting ability of academic libraries. Until now the membership of CALIS totaled up to 700. Thirty academic libraries have been developed into digital library bases, acting as the kernels of information service systems and distributing centers of information resources. Besides, digitalized information resources imported from foreign countries are expected to cover all subject, areas while domestic information resources will be as much as several terabytes (Zhu, 2001). The second phase of CALIS construction was combined with the CADAL project in 2004, adopting a new name CADLIS (Chinese Academic Digital Library & Information System). It has strengthened the construction of the full-text database, whose goal is to offer 30,000 e-journals (including 20,000 international journals), 300,000 theses, and 30,000 education and reference books (Xiao & Chen, 2005).

China Digital Library Project and NSTL Construction

Construction of the China Digital Library and the National Science and Technology Library has been advancing smoothly. In April 1999, China Cultural Information Net started operation as the top level of the China Digital Library. In November 1999 and February 2000 respectively, the Capital Library and China Radio International (CRI) became experimental units of the China Digital Library Project. It should include information resources not only from libraries, but also from the government, even from international channels. The ultimate goal of the project is to build a "Digital China." The National Culture Information Resources Sharing Project has been carried out. All resources have been published through Web pages to allow access for people in rural and remote areas. By digitalizing the information resources maintained in public libraries, museums, and art galleries in China, the project aimed to provide cultural services to people in rural areas. The project investment totals 400 million RMB (US$50 million). Until now, 4,226 local centers and backbone service sites have been established (China Radio International, 2006).

The initiative of building the National Science and Technology Library is near conclusion. Through two years of construction, participating libraries now collect more than 16,000 types of foreign scientific and technical literatures (including periodicals, conference proceedings, technical reports, etc.), as compared to no more than 4,000 types in 1996.

Some 6.5 million bibliographical records were put online by the end of March 2002, and this number is expected to increase at a rate of 2 million per year. The network service system provides 24-hour free secondary literature retrieval service to Internet users. In March 2002 alone, 1.37 million users visited the system, as compared to 150,000 when the system was started in January 2001. More than 60,000 users have received full-text document service.

Now it can provide a variety of services such as bibliography searching, full-text browsing, virtual reference, and Chinese and foreign language preprint acquisition through its network service system. By the end of 2005, 35 million bibliographical records had been put online, which is 20 times as many as in 2000 when it opened. The system responds to more than 10,000 requests for full-text service each month, compared to 60 at the beginning of 2000. In 2005 alone, nearly 28 million searches took place with the system. The principle of NSTL is "unified acquisition, standardized processing, data integration, and resources sharing." The main purpose of establishing NSTL was to satisfy the new science and technology information needs, and revitalize the industry-oriented science and technology information service system during the transition from the planned economy to the market economy in China (Zhang, 2006).

Development of Commercial Information Products

Information resource development in a market-oriented approach achieves great effect. Many database and information service providers (such as ICP, ISP) have come into operation, among which the following enjoy a nationwide reputation: China Academic Journals CD-ROM database, ChinaInfo Group, Chongqing Weipu Information Consulting Corporation Ltd. (*www.vipinfo.com.cn*), Beijing Scholar Sci-Tech Co., Ltd., and China Infobank. The Chinese Journal full-text Database includes about 18 million articles from 7,626 major periodicals published in Mainland China since 1976. The database is available both online and in CD-ROM form.

In a broader context, Internet-based information resources have also undergone rapid development. According to statistics from the "Survey on Information Resources in China," which was released by the China Internet Network Information Center in March 2006 (CNNIC, 2006b), there were 2.6 million registered domain names, 694,000 Web sites, 2.4 billion Web pages, and 295,000 online databases within China.

FUTURE TRENDS

The projects we have introduced above lay a solid foundation for the further development of information resources in China. Recently, the Ministry of Science and Technology started up the Science and Technology Documents Resource and Service Platform. Founding of the National Informationalization Directing Committee also boosted the development and utilization of information resources.

In the "2006-2020 National Informationalization Development Strategy" issued by the General Office of the CPC Central Committee and the General Office of State Council, information resources development was positioned in a significant place. It was listed as one of the national informationalization strategic goals. It was also one of the nine strategic tasks of national informationalization. Network information resources development was one of the six strategic plans of national informationalization (*Tianjin Daily*, May 9, 2006).

According to the "2004-2010 National Science and Technology Fundamental Platform Construction Blueprint," the Science and Technology Documents Sharing Platform is one of the four fundamental platforms. The platform construction includes two projects: one is science and technology document resources construction; the other is documents service ability construction. The platform includes a network service system, an integrated resources navigation system, and an individualized knowledge service system. The investment of the system is estimated at 150 million RMB (US$12.5 million).

Looking into the future, we can feel the long road ahead for China to improve its information resources. Efforts need to be made in a number of areas: first, the digital library projects need to be further expanded; second, the government should continue its support for information resources development projects; and third, the issue of nationwide cooperation across industries needs to be addressed properly as one of the most important problems in information resources development.

CONCLUSION

This article introduces the history and present situation of information resources development in China. Some major information resources development projects at the national level are discussed. Several of the projects have already generated positive impact on the society and laid a solid foundation for the further development of information resources in China.

REFERENCES

Bian, Z. (2006). A milestone of Chinese e-government informationalization construction. *Informationalization Construction*, (1/2), 6-9.

China Telecommunications. (2003). *China telecommunications yearbook.* Beijing: Posts and Telecom Press.

CNNIC. (2006a). *Statistical reports on the Internet development in China.* Retrieved from http://www.cnnic.net.cn/download/2006/18threport-en.pdf

CNNIC. (2006b). *Survey reports on quantity of Internet information in China.* Retrieved from *http://www.cnnic.net.cn/download/2006/20060516.pdf*

CRI Editor. (2006). *Stage achievement of Chinese cultural information resources sharing projects.* Retrieved from http://gb.chinabroadcast.cn/1321/2006/05/25/1569@1059967.htm

Fu, L. (1996). Some thinking on the strategies of China's sci-tech information resources construction. *Journal of the China Society for Scientific and Technical Information, 15*(5), 374-377.

Fu, P. (2005a). The standards construction of Chinese National Digital Library national library. *Journal of the National Library of China*, (4), 13-16.

Fu, P. (2005b). Construction and development perspectives of National Digital Library. *Library and Information Service, 11*, 5-8.

Guan, J. (1988). *Information work and information science development strategy.* Beijing: Sci-Tech Document Publishing House.

He, W. (2003). *Rapid development of communication industry in China.* Retrieved from http://www.cnii.com.cn/20030915/ca226082.htm

Lv, Y. (1999). Digital libraries in Taiwan. *China Computer World,* (May 31).

Xiao, L., & Chen, L. (2005). CALIS and development of academic digital libraries in China. *Library and Information Service, 11*, 9-14.

Yuan, H., & Meng, L. (2001). A practice on information resource sharing in Web-based environment — construction and development of National Research Centers for Information Resources Management. *Proceedings of the Academic Seminar Commemorating the 45th Anniversary of China's Scientific and Technical Information Cause,* Beijing, China.

Zhang, X. (2002). China Scientific Digital Library: User-oriented digital information service system. *Sci-Tech International,* (4), 21-23.

Zhang, X. (2005). CSDL and its advances. *Bulletin of the Chinese Academy of Sciences, 20*(4), 344-346.

Zhang, Z. (2006). The network service system of NSTL. *China Information Review, 4*, 48-51.

Zhu, Q. (2001). A rewarding practice on information resource sharing oriented towards the 21st Century — advances in the construction of China Academic Library & Information System. *Proceedings of the 2001 Annual Meeting of the China Society for Library Science*, Chengdu, China.

Zou, J. (1997). Promoting national informationalization. *China Electronics Daily,* (September 16).

KEY TERMS

Digital Library (DL): A cultural infrastructure that collects and stores information in electronic format and supports its users in accessing a large collection of information effectively through digital means.

Information Resource: A collection of valuable information generated by human activities. In a broader sense, it also includes related equipment, personnel, and capital.

Information Resources Development: The process of collecting, processing, storing, disseminating, and utilizing information resources according to the need of the society.

Information Resources Management (IRM): Refers to the planning, organization, allocation, and utilization of information and related resources through legal, technological, and other methods to support institutional goals and missions.

Information Service: The activity of providing information products and related services according to users' need. In a broader sense, it refers to providing users with information through any form of product or service.

Informationalization: The process of social advances in which human society transforms from industrial society to information society.

Information Sharing in Innovation Networks

Jennifer Lewis Priestley
Kennesaw State University, USA

Subhashish Samaddar
Georgia State University, USA

BACKGROUND

Innovation networks help members develop new products at a faster rate with lower investment commitments. The R&D consortium named Semiconductor Manufacturing Technology (SEMATECH), with member firms such as Motorola, Texas Instruments, and others, is an example of such a network. In a study of this network, Lim (2000) found that the network members were able to develop an innovative copper-based semiconductor that rivaled a similar product developed by (at the time) an independently operating IBM. The SEMATECH consortium experienced a significantly abbreviated time line and collectively invested significantly less money than did IBM with almost identical results. Lim attributed the innovative success of SEMATECH to the "connectedness" of the firms.

Researchers engaged in studies examining interorganizational alliances generally agree with the findings of Lim and others that innovation network alliances represent a potential solution to mitigate environmental uncertainty, in part through the sharing of information (e.g., Gulati & Gargiulo, 1999). Van de Ven (2005) refers to this strategy for dealing with environmental uncertainty as "Running In Packs." The basic logic is that as a network grows in membership, the amount of information any individual firm can access grows, and the value of membership in that network grows. Consequently, firms engaged in networks typically realize superior economic gains from their increased access to information relative to independent or nonaligned firms (e.g., Carlsson, 2002; Van de Ven, 2005).

Since organizations join networks to mitigate costs and uncertainties, the question of how network characteristics affect (or not) the transfer of information is relevant to both practitioners as well as researchers in knowledge management and/or organizational learning. For instance, some innovation networks are composed of members engaged in similar activities while other networks are composed of members engaged in very different activities. Some networks tolerate more competition among their members than others. Finally, some networks are more centrally governed than others. These differences in how an innovation network is formed and governed raises an important question—Given that firms embedded within organizational networks experience greater exchange of information relative to firms operating outside of a network, how do the different characteristics of these networks impact the movement of that information?

In this chapter, we will first review the two primary factors that have been demonstrated to influence the transfer of information—*absorptive capacity* and *causal ambiguity*. We then review three characteristics of *multi-organizational networks—governance structure, scope of operations*, and *intensity of competition*—with particular attention to the issue of information transfer. We develop six testable propositions regarding how these network characteristics would be expected to affect absorptive capacity and causal ambiguity among networked firms. Finally, we discuss future and emerging trends related to the transfer of information among networked firms.

INFORMATION SHARING

Economic theories such as the *knowledge-based view of the firm,* view information as an asset that will move unencumbered and without cost within and among organizations; although information is recognized as an asset, unlike other assets, its transferability has no associated costs. However, some authors have suggested that this may not be the case (e.g., von Hippel, 1994). In fact, the transfer of information is not necessarily frictionless and has even been described as "sticky"and the organizational implications associated with transfer "stickiness" can reach beyond issues of cost and simple inefficiencies (Szulanski, 1996). Information is increasingly recognized as the engine of economic growth and a source of competitive advantage, and where its transfer is difficult, the implications are more strategic and may threaten a firm's long-term competitiveness, including, new enterprise formation; the exploitation of technological know-how; and the successful development and commercialization of new products and services (Teece, 2001).

Absorptive Capacity

An organization's absorptive capacity has been described as the organization's ability to first recognize and then realize any value from the external information to which it is ex-

posed (Cohen & Levinthal, 1990). Such exposition arguably increases when a firm becomes a member of a network. As a result, if absorptive capacity is low, the transfer of information is less likely to occur. In a networked context, the absorptive capacity of the recipient organization is integral to the success of *information sharing*. The ability to identify new, relevant information and have the processes in place to then bring it internal to the organization quickly becomes a competitive advantage when translated into economic rents. However the paradox of absorptive capacity is that an organization that does not have it may not understand that they need it; organizations with low absorptive capacity will be less likely to value external information (Mosakowski, 1997).

Causal Ambiguity

Unlike absorptive capacity, which is considered to be an enabler of information sharing, the presence of causal ambiguity has been identified as an isolating mechanism of information, impeding its movement within and among organizations (Knott, 2003). The concept of causal ambiguity is centered around the organizational inputs and the causal factors used in combination to generate known outcomes. Organizational inputs can be the raw materials used to manufacture a product, and the causal factors can be viewed as the processes used. When an organization is successful in benefiting from an innovative process but does not know what combination of inputs and process factors generated the final outcome, their knowledge is, at best, causally ambiguous.

Causal ambiguity as an inhibitor of information transfer has been recognized across much of the research in organizational learning. Mosakowski (1997) developed a useful typology through which to examine the effects of causal ambiguity on decision making. Extending the work of Lippman and Rumelt (1982), Mosakowski determined that although increased causal ambiguity has the potential to increase competitive advantage by increasing the difficulties associated with imitation by competitors, increased causal ambiguity has the impact of decreasing information transferability by associating its application.

CHARACTERISTICS OF ORGANIZATIONAL NETWORKS

In this chapter, we approached the examination of networks using two established perspectives. The first, *transaction cost economics*, recognizes that exchange agreements between and among firms must be governed and contingent on the transactions to be organized; some forms of governance are better than others (Williamson, 1973, 1975). Specifically, this includes examination of centralized and decentralized governance. The second perspective, *social network theory*, examines the individual *nodes* and *linkages* within a network to explain how organizations (or individuals) will interact (e.g., Westlund, 1999). Using these well-established perspectives as a basis, we discuss the three primary characteristics of an interorganizational network that would be expected to have particular influence on the transfer of information—governance structure, scope of operations, and intensity of competition.

Governance Structure

Networks of organizations represent an organizing principle residing between pure market-based transactions and complete organizational self-sufficiency (Thorelli, 1986). However, once "within" the network, the question of governance structure remains to be determined. In his work on transaction cost economics, Williamson (1973, 1975) identifies the preferred governance structure for providing authority to address issues related to opportunistic behavior, information impactedness and bounded rationality as centralized or hierarchical. This governance structure would also be expected to have the ability to mandate standardization of operations, language, policies, and so forth. Conversely, a decentralized governance structure is described as one of peer group associations, without subordination, involving collective and usually cooperative activities, but deficient in its ability to address opportunism and free-rider abuses. A decentralized governance structure has been suggested as preferable to facilitate innovation and new knowledge creation, where the former structure has been suggested to better facilitate the transference of existing information (e.g., Adler, 2001; Chen & Edgington, 2005).

Scope

Researchers engaged in social network theory and organizational alliances have stated that the degree to which the members of a multi-organizational network or of a dyadic alliance demonstrate operational homogeneity affects the likelihood of information transfer (e.g., Westlund, 1999). Specifically, operational similarity has been used to explain, in part, when information does or does not transfer between or among alliance partners (e.g., Simonin, 1999). For the purposes of this chapter, we will refer to this network characteristic of member similarity as the *scope of operations*, where a high scope network will have operationally dissimilar members while a low scope network will have operationally similar members.

Intensity of Competition

The concept of linkages among the nodes or members in a network has been identified to have significant impact on

how well information does or does not transfer (e.g., Uzzi & Lancaster, 2003). The linkages that exist among network entities have been described as being either *integrated* or at *arm's length* (Dacin, Ventresca, & Beal, 1999). Integrated linkages are defined as shifting the expectations among the network members away from a behavioral basis of opportunism to a behavioral basis of cooperative behavior, where these expectations then facilitate the transfer of information (Uzzi & Lancaster, 2003). Alternatively, linkages at arm's length are described as creating impersonal, atomistic behavior, characterized by instrumental profit-seeking on an individual member basis (Uzzi & Lancaster, 2003).

Why would an organization voluntarily join a network, where the linkages among the members would be characterized as being at arm's length? Consider the VISA network. Individual banks are fierce competitors, yet collectively benefit from the functionality of the global card acceptance afforded by the VISA network—relationships among the members would be described as impersonal and profit seeking, with linkages created for the purposes of decreased transaction costs. This seemly paradoxical concept of arm's-length linkages among networked entities has been found to be particularly evident in industries characterized by rapid change. Powell, Koput, and Smith-Doerr (1996) found that as the technological sophistication of an industry increases, the intensity and number of competitive alliances also increases.

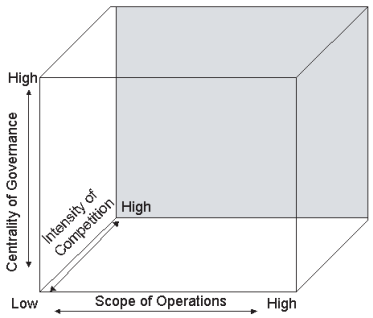
When there is a regime of rapid technological development, research breakthroughs are so broadly distributed that no single firm has all the internal capabilities necessary for success ... Firms thus turn to collaboration to acquire resources and skills they cannot produce internally, when the hazards of cooperation can be held to a tolerable level. (Powell et al., 1996, p. 117)

In this article, we will refer to this characteristic of a network as the *intensity of competition*, where low intensity of competition will equate to integrated linkages and a high intensity of competition will equate to arm's-length linkages.

Using the three characteristics of a network described in this section, different network forms can be examined

Consider Figure 1. Although a conceptually infinite number of networks could be described using the three characteristic continuums represented, consider a typical innovation network that has been formed with the objective of mitigating costs, risks, and environmental uncertainty. Innovation networks are typically characterized by a decentralized governance structure, with limited ability to regulate or punish, comprised of nonsubordinated peers. These networks also typically exhibit a high scope of operations—where member firms may include public companies, universities, nonprofit firms, governmental and quasi-governmental firms.

Figure 1. Characteristics of multi-organizational networks



Finally, innovation networks may exhibit a range of intensity of competition—where otherwise competitive firms agree to collaborate for specific purposes—such as Motorola and Texas Instruments did as members of the R&D consortium SEMATECH, referenced previously.

An assessment of how each characteristic of this network would be expected to affect absorptive capacity (an enabler of information transfer) and causal ambiguity (an isolating mechanism of information transfer) is developed next.

FUTURE TRENDS

Cohen and Levinthal (1990) highlight the importance of possessing certain commonalities to facilitate absorptive capacity and the effective sharing of information among organizations. Cohen and Levinthal and others (e.g., Lane and Lubatkin, 1998) have identified several factors, which contribute to an organization's absorptive capacity. Two of these factors include a commonality of language and a common base knowledge—where common knowledge translates to an intersection, not an overlap of knowledge. A complete overlap of knowledge is inefficient and represents limited opportunity for transfer.

Given that organizations generally join innovation networks for the purposes of gaining access to new information, they would be expected to come to the network with some common or *base* knowledge. This is true because if organizations did not have a base knowledge of the topic in question, transfer would be almost impossible. For example, if a network was developed to research a particular form of cancer, and a strong base knowledge of biochemistry were required, an organization with no previous experience in the area of biochemistry would have a limited ability to absorb or contribute to the exchange of information. In ad-

dition, within the context of this base knowledge, it would be logical to conclude that a common language is used. For example, the Human Genome Project, incorporates government (e.g., U.S. Department of Energy), quasi-government (e.g., National Institutes of Health), private (e.g., Wellcome, IBM), and research (e.g., MIT, Baylor College of Medicine, Washington University) institutions. Each organizational entity approaches the project with some common working knowledge of, for example, genetics, which is then used to develop new knowledge related to gene sequencing, bioinformatics, and other topics. Although these respective organizations may use completely different vernaculars within their respective operating environments, within the context of the Human Genome Project, it would be expected that in the pursuit of new knowledge, the organizations would engage in a common language.

Proposition 1. *Low centrality of governance in the innovation network contributes to high organizational absorptive capacity.*

Proposition 2. *High scope of operations in the innovation network contributes to low organizational absorptive capacity.*

Proposition 3. *Low to medium intensity of competition in the innovation network contributes to high organizational absorptive capacity.*

The innovation network is most common in environments characterized by rapid and turbulent change, where inputs and/or causal factors may not be understood prior to an outcome (Mosakowski, 1997). Scenarios characterized by complexity and ill-structured problems are considered to have high causal ambiguity. This is logical—if a particular process or product has many interdependent components, identifying or isolating the impact of each one on the eventual outcome would be difficult. Simon (1962) suggested that more hierarchical structure helps to mitigate this complexity. However, this structure is generally not present in an innovation network, in part because it has been shown to decrease new knowledge creation and innovation (e.g., Adler, 2001).

Proposition 4. *Low centrality of governance in the innovation network contributes to high organizational causal ambiguity.*

Proposition 5. *High scope of operations in the innovation network contributes to high organizational causal ambiguity.*

Proposition 6. *Low to medium intensity of competition in the innovation network contributes to high organizational causal ambiguity.*

CONCLUSION

In this chapter we examined three characteristics of the innovation network. The developed propositions could indicate that this network would experience problems with information sharing. Initially, this expectation may violate conventional wisdom—we stated earlier that the primary motivation to participate in this network form was access to new information. However, a further consideration of knowledge creation versus established information transfer may explain the (possibly) perceived violation. This network type is not based on the need to transfer an existing information asset within the network, but rather on the need to develop new, unproven knowledge. Therefore, it is not surprising that these characteristics—some competition, low centralization of governance, and high scope of operations—could be expected to restrict the flow of information, but facilitate new knowledge creation.

It is our intention that this chapter will provide researchers in the areas of organizational learning and knowledge management with rich concepts for further examination. From our perspective, a rich opportunity for further examination is embedded in questions regarding how differently organized networks, which reside in different positions within the cube identified in Figure 1, would experience information transfer differently. Would a supply chain network experience absorptive capacity and causal ambiguity differently than would an innovation network? How would a co-op network of members simultaneously competing and collaborating, experience information transfer? Although much research has been done to establish that networks are superior to dyads and dyads are superior to independently operating firms regarding the transfer of and access to information, a large theoretical gap exists regarding how differently organized networks experience information transfer differently.

REFERENCES

Adler, P. (2001). Market, hierarchy and trust: The knowledge economy and the future of capitalism. *Organization Science, 12*(2), 215-234.

Baum, J. A. C., & Ingram, P. (1998). Survival-enhancing learning in the Manhattan hotel industry 1898-1980. *Management Science, 44*(7), 996-1016.

Carlsson, S. A. (2002). Strategic knowledge managing within the context of networks. In C. W. Holsapple (Ed.),

The handbook on knowledge management (pp. 623-650). Berlin, Germany: Springer Verlag.

Chen, A., & Edgington, T. (2005). Assessing value in organizational knowledge creation: Consideration for knowledge workers. *MIS Quarterly, 29*(2), 279- 309.

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly, 35*(1), 128-153.

Dacin, M., Ventresca, M., & Beal, B. (1999). The embeddedness of organizations: Dialogue and directions. *Journal of Management, 25*(3), 317-356.

Darr, E. P., Argote, L., & Epple, D. (1995). The acquisition, transfer and depreciation of knowledge in service organizations: Productivity in franchises. *Management Science, 41*(11), 1750-1762.

Dyer, J. H. (1997). Effective interfirm collaboration: How firms minimize transaction costs and maximize transaction value. *Strategic Management Journal, 18*(7), 535-556.

Gulati, R., & Gargiulo, M. (1999). Where do interorganizational networks come from? *American Journal of Sociology, 104*(5), 1439-1493.

Knott, A. (2003). The organizational routines factor market paradox. *Strategic Management Journal, 24,* 929-943.

Lane, P. J., & Lubatkin, M. (1998). Relative absorptive capacity and inter-organizational learning. *Strategic Management Journal, 19*(5), 461-477.

Lim, K. (2000). *The many faces of absorptive capacity: Spillovers of copper interconnect technology for semiconductor chips*. Paper presented at the 2000 Academy of Management Conference.

Lippman, S. A., & Rumelt, R. P. (1982). Uncertain imitability: An analysis of interfirm differences in efficiency under competition. *Bell Journal of Economics, 13*(2), 418-439.

Madhavan, R., Koka, B. R., & Prescott, J. E. (1998). Networks in transition: How industry events (re)shape interfirm relationships. *Strategic Management Journal, 19*(5), 439-459.

Mosakowski, E. (1997). Strategy making under causal ambiguity: Conceptual issues and empirical evidence. *Organization Science, 8*(4), 414-442.

Powell, W., Koput, K., & Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly, 41*(1), 116-146.

Simon, H. (1962). New developments in the theory of the firm. *American Economic Review, 52*(2), 1-16.

Szulanski, G. (1996, Winter). Exploring internal stickiness: Impediments to the transfer of best practice within the firm [Special Issue]. *Strategic Management Journal, 17,* 27-43.

Teece, D. J. (2001). Strategies for managing knowledge assets: The role of firm structure and industrial context. In I. Nonaka & D. J. Teece (Eds.), *Managing industrial knowledge: Creation, transfer and utilization* (pp. 125-144). London: Sage.

Thorelli, H. B. (1986). Networks: Between markets and hierarchies. *Strategic Management Journal, 7*(1), 37-52.

Uzzi, B., & Lancaster, R. (2003). Relational embededdness and learning: The case of bank loan managers and their clients. *Management Science, 49*(4), 383-400.

Van de Ven, A. (2005). Running in packs to develop knowledge-intensive technologies. *MIS Quarterly, 29*(2), 365-378.

von Hippel, E. (1994). "Sticky information" and the locus of problem solving: Implications for innovation. *Management Science, 40*(4), 429-438.

Westlund, H. (1999). An interaction-cost perspective on networks and territory. *The Annals of Regional Science, 33,* 93-121.

Williamson, O. E. (1973). Markets and hierarchies: Some elementary considerations. *American Economic Association, 63*(2), 316-325.

Williamson, O. E. (1975). *Markets and hierarchies: Analysis and antitrust implications*. New York: Free Press.

KEY TERMS

Absorptive Capacity: The ability of a firm to recognize the value of new, external information and assimilate it and apply it to commercial ends.

Causal Ambiguity: The "knowability" (the extent to which something *can* be known) and "knowness" (the extent to which something *is* known) of two sets of elements—(1) the organizational inputs and (2) the causal factors that are used in combination to generate outcomes.

Innovation Network: A structured network of N organizations sharing common goals related to research and/or development of new products/technologies (e.g. The Human Genome Project). This network type is characterized by a decentralized structure, low-medium competition and uncommon scope of operations among members.

Inter-Organizational Network: A community of practice of N organizations, where N is more than two. In this chapter, network types are defined through three primary characteristics—the degree of centralization of authority, competition, and commonality of operations.

Knowledge Management: Knowledge management (KM) is the organization, creation, sharing, and flow of knowledge within and among organizations.

Social Network Theory: A theory, which explains how individuals or organizations will interact based upon the nodes and linkages within the network.

Transaction Cost Economics: A theory most typically associated with Williamson, which explains that firms will organize in a manner to minimize the costs of production, including the costs of transaction and exchange.

Information Society Discourse

Lech W. Zacher
Leon Kozminski Academy of Entrepreneurship and Management, Poland

INTRODUCTION

Information society (IS) has a short history as a form of human organization and social context. However, information (signals, communications, various data, etc.) and use thereof have always been fundamental to people's existence, survival, and development. Some important milestones included the Gutenberg printing press, telephone, radio, TV, computer, and all electronic devices and systems related to ICTs. In fact, the progress of technology, especially of electronics and telecommunications, marked out the directions and potentialities of social change.

Coined as a term in the 1960s, information society is just emerging nowadays mostly in developed countries. As a result of the effect of present technological, economic, and political globalization processes, the whole world is being impacted and transformed by ICTs. IS can be *per se* perceived as the intellectual (scientific) model or ideal type having a set of specific characteristics and assigned interpretations.

Needless to say, in the real world there are only *concrete individual different* information societies. Their difference concerns mostly: geographic, historical, educational, technological, cultural, political, and economic aspects and advancements already achieved in IS development (i.e., its stage, directions, pace, and so on) and their multifaceted impacts on societies, organizations, and individuals. In the social sciences—especially in sociology and political science—there are some indicators enabling measurement of these advancements and their consequences.

The aforementioned societal advancements, initially always pre-informational or not yet informational, are constantly emerging from some "embryos"—often scientific and technological—and are progressing via multidimensional processes of organizational, social, economic, political, cultural innovations, and by their diffusion. In fact, all segments and features of society are heavily affected by them. These impacts are rather difficult to measure and evaluate. Quite often, they are treated generally as ICTs' impact on a society. Certain analytical methods and procedures connected with *technology assessment* or—more comprehensive—*impact assessment* can be applied to this end. Since IS is still emerging, or in other words *in* the *statu nascendi* stage, it is reasonable and necessary to apply a prospective approach to its investigations and evaluations.

Therefore, the future of ISs should be of interest not only to researchers, but also governments, business, and the public—referred to as *civil society* in democratic countries. Increasing use of the word "future" in its plural form, "futures," has been accepted for a long time. In English this form has already functioned for decades, while in other languages "future" is used only in singular. The other reason is that people (and scientists) often perceived the future as non-optional (a rather fatalistic approach). By using the plural form, we emphasize the conviction and hopes that the future will be multi-optional, thus very differentiated for regions, states, societies, communities, and individuals. Therefore, differentiated ISs will not have the same futures. As such, the future of the whole world will be extremely complex. It does not seem probable that there will be one future for all.

Historically, various societies have had divergent take-off points, possibilities, development opportunities, trajectories, as well as performance, behavior, policies, cultural heritage, social capital, and so forth. In spite of some universalistic tendencies in production and consumption patterns, many diverse *gaps* currently exist in political systems, media performance, and so forth. Some time ago, it was fashionable to refer to them as technological, organizational, or managerial, information. There are other forms and names, for example, presently we talk about the digital divide and knowledge gap. Technological developments, their diffusion and transfer all over the world do not make the world equal regarding the stage and impacts of IS progress (understood in the abstract).

The irregular development of economies and societies throughout the world seems to be a historical regularity. The same applies to the present stage of development connected with ICTs. A historical perspective of IS development in particular countries requires grouping such into classes:

- *pioneering countries*—in ICTs production, use and wide diffusion in all sectors of economy and social life;
- *imitators*—taking advantage of technology transfer, FDIs, and global networking, however the diffusion of ICTs may be limited to selected sectors; and
- *lagging behind*—for a variety of reasons, for example, educational, technological, economic, political, cultural, and so forth, such countries may have trouble introducing and effectively utilizing ICTs.

A similar division is possible made within particular countries. The developmental dualism seems to be common in many parts of the world, especially in the less advanced

states. However, some regrouping is occurring. Until recently the only pioneers were the United States, Japan, and Western Europe. Due to the global reach of transnational corporations, FDIs, international trade, and global networking, as well as national strategies and efforts, some countries have become increasingly competitive (e.g., China, India). Moreover, the internationalization of ICT production is rapidly growing. In addition, some countries (including the entire European Union) declare they are building an information society.

Nevertheless, particular ISs will not have the same faces throughout the world despite some strong similarities, universalistic trends, similar strategies and policies of governments and business, and certain parallel human activities. The chaos of developing a diversified and turbulently changing environment may outweigh some, mostly technological, deterministic tendencies. Technological determinism that assumes "one way for all" seems to be merely an intellectual idea or simplistic concept rarely functioning in reality (if so, with some time and space limits).

Summing up: various emerging information societies are highly differentiated and will probably also have differentiated info-futures. Apart from certain similarities and some evident universalization, the growing info-diversity may occur and greatly shape the world's societies. Therefore, even a general abstractive model or pattern of an information society may need reinterpretation based on actual experience. So far, the known prophecies and visions of IS development, elaborated in Japan, the United States, and Western Europe, will probably not match the real course of events, the real potentials, needs, and aspirations of billions of people. All the aforementioned differences were reflected in the IS discourse.

BACKGROUND: INFORMATION SOCIETY - DEFINITIONS AND DISCUSSIONS

ICTs, their multifaceted impacts, the change of sociocultural context, and the global dimensions of all emerging transformations need permanent investigation, analysis, interpretation, and forecasting, required not only for research, theorizing, or education.

All theorists, futurists, and analysts dealing with the IS problem express a conviction that there is some possibility and social ability to steer and control occurring changes and transformations.

For all these reasons, there have been numerous efforts to define problems, recognize and evaluate processes, and predict the possible future course of IS around the world since the 1960s.

Providing one commonly accepted definition of IS seems to be very difficult. There are many terms or qualifications directly or indirectly connected with broadly understood IS. To name several examples: *information society, information rich society, cyber-society, computer society, telematic society, network society, virtual society, e-society,* and the like. These terms underline the role of various characteristics and symbols like, for example, access to information, cyberspace as a new social space, use of computers, telecommunications networking, virtualization, and electronization. All are relevant and in fact complementary. However, various authors tend to support their own interpretation concerning the most important features. The long list includes examples such as Masuda (1981a, 1981b), Negroponte (1996), Dertouzos (1998), Castells (2000, 2004), Wellman (1999), and Virilio (1998). Many authors add such qualifications as *digitalization*, referring to the advances of info-technology (e.g., Tapscott, 1998), and *mediatization*, referring to the overwhelming role of mass media (e.g., Lievrouw & Livingstone, 2002; Downing, 2000).

It is quite difficult to find truly precise definitions in the very extensive literature on this subject. Quite often, there are descriptions, characteristics, and qualifications that are rather general and vague (i.e., not comprehensive). In many cases, though the term "information society" appears in the title and in the text of a book, article, or document, it is not explicitly defined, but used as a kind of label, taking for granted that the content would sufficiently explain all terms.

However, there have been many efforts in the past to describe, analyze, and evaluate ongoing technological, socioeconomic, and cultural changes connected with new ICTs. For example, Bezold and Olson (1986) reviewed first the specific *societal impacts*—past, present, and probable—of the information revolution. Subsequently, they discussed different *whole-system* images of how an information society may evolve. They contrasted images of the civilizational and societal transformations of leading future-oriented thinkers, such as Bell (1973), Toffler (1980, 1990; Toffler & Toffler, 1995), Naisbitt (1982), Harman and Markley (1985), and Masuda (1981a, 1981b), who believed that a new stage of civilization is emerging, with information and ICT playing a pivotal role in the social transformation. However, they differ on such matters as the *key driving forces* for societal change, the *main features,* and the *overall pattern* of change. Bell (1973) announced the emergence of post-industrial society in which the critical driving force for change is the codification of theoretical knowledge generating the exponential growth of science, systematic R&D, and new intellectual technologies. Toffler (1980, 1990; Toffler & Toffler, 1995) developed the theory of a third wave driven by growing socio-economic complexity, diversity, heterogeneity, and connected with demassification of production, media, lifestyles, and so forth. The newly emerging social order demanded higher levels of information flow.

Naisbitt (1982) believed social development rather than technological change leads to information society, although

technology greatly stimulates the transformations. He stressed the changing structure of the economy (rising dominance of the information sector) and the role of information (including new information jobs) as the key strategic resources.

Harman and Markley (1985) noted that current problems require something other than industrial technology and a worldview hailing from the industrial era. They argued that new images of the future and new values were required, along with a shift in the hierarchy of information types (data, information, knowledge) with emphasis on wisdom. In their view, this may lead to a "learning society," but "failure futures" are also possible in spite of information richness.

Masuda (1981a, 1981b) elaborated "The Plan for an Information Society: Japan's National Goal Toward the Year 2000," in which he set out the most radical vision of development based on three stages—automation, knowledge creation, and system innovation—with the computer as a key driving force.

All authors mentioned above appreciated the great transformational role of information and information technologies for social change. All except Bell (1973) envisioned a *multi-optional future*—not necessarily positive and democratic. However, all authors assumed some *possibility and social ability* to steer change by decisions, policies, business behavior, education, citizen participation, and so forth.

Based on the *alternative futures approach,* Bezold and Olson (1986) worked out four scenarios:

- the High-Tech Information Society,
- the Creative Society,
- Things Bog Down, and
- 1984 and Beyond.

The first scenario was similar to Bell's (1973) image, the second to the images forecasted by Toffler (1980, 1990; Toffler & Toffler, 1995), Naisbitt (1982), Harman and Markley (1985), and Masuda (1981a, 1981b) (see Bezold & Olson, 1986). The last two presented a less optimistic future, even with the possibility of authoritarianism.

Bezold and Olson (1986) in their report used the following definition of IS:

A society that reflects the growing importance of information in the shift from manufacturing jobs, in the growth of information purchases, in the importance of information in business and personal life, and in the enhanced level of information built into products. (Bezold & Olson, 1986, p. A-24)

This definition is economically (business) oriented.

A more recent approach was presented within the ESRC Research Program, "Virtual Society?: The Social Science of Electronic Technologies." The program "was set up to research the implications of the continued massive growth in new electronic technologies" (ESRC, 2000, p. 3). It posed two questions: "Are fundamental shifts taking place in how people behave, organize themselves and interact as a result of the new technologies? Are electronic technologies bringing about significant changes in the nature and experience of interpersonal relations, in communications, social control, participation, inclusion and exclusion, social cohesion, trust and identity?" (ESRC, 2000, p. 3).

The answers had to serve the *policy agenda* and contribute to *business success,* to the *quality of life,* and to the *better future of society* (see also Woolgar & Ingram, 2000). The program formulated the rules of virtuality and especially emphasized the role of the local social context—finding expression in the cultural capacity to make the best use of new technologies. It further considered relations between technologies and social theory (Woolgar, 2000).

Interestingly, computer specialists such as Bill Gates (Microsoft), Nicholas Negroponte, and Michael Dertouzos (the latter two from MIT) preferred to use terms such as *information revolution, information age, digital life, world,* and *future.* Their books (Gates, 1996; Negroponte, 1996; Dertouzos, 1998) are the visionary guides to the future. They focused on how ICTs transform human life and the world. As active participants in the creation of the information age, they present deep insights into the technological problem. They do not theorize on society, dealing rather with the lives of individuals. They are optimistic, though they see a dark side of new technologies. For Negroponte (1996), the coming digital future is an "age of optimism" (p. 227).

Mattelart (2001) presents a very broad historical and cross-disciplinary perspective. Instead of using the term IS, he prefers "society of information." The central focus of Mattelart's investigations are information and its history. He is skeptical about the idea of commonly accessible information, pointing out the conflicting interests of governments and societies, even stimulated by the international context (e.g., terrorism) and the continuing role of development (Mattelart, 2001). Interestingly, Mattelart refers to the historical anticipations of the ideas of globalism and world networks, which are very fashionable at the present time.

The apologetic discourse on IS is also criticized by May (2002). In fact, his criticism is directed against the overoptimistic and exaggerated beliefs and convictions concerning the stage, development, span, and transformational potential of the "information age," "new economy," and "information society." He rightly points out that all these new ideas do not necessarily invalidate what we already know about society, the economy, and the world.

There have been many changes we might link to the development of new information and communication technologies (ICTs). However, there is also much about the global information society which is similar to previous modes of

social interaction: economics is still recognizable as modern (or perhaps late) capitalism; despite forecasts of increased 'virtualization', politics, communities and other aspects of social existence remain located in the material world; states continue to play an active and important role in our lives. Thus, there is no need to dispense with our previous 'ideological baggage'. Indeed, the claim that we should is ideological in itself. It represents a dismissed of well-developed arguments regarding the contested and contingent character of capitalism, while also presenting a specific set of contemporary social relations as natural and outside history. (May, 2002, p. 149; see also Schiller, 1999, who used the term "digital capitalism")

Castells (2000, 2004) is a very influential author who argues that the widespread deployment of ICTs produces a networked society in which new communication capacities can be beneficial both for organizations (like companies) and individuals. Further, he finds that electronically mediated networks greatly support development and diffusion of information and knowledge. Physical resources are increasingly substituted for the mobilization and coordination of knowledge and information resulting in information capitalism and the network society. In spite of some controversies and criticism (also by May, 2002), Castells' (2000, 2004) approach seems to be useful. Needless to add that some other researchers made use of the network approach as well (e.g., van Dijk, 2006; Wellman, 1999).

It seems to be quite justified to review the IS problem in the EU documents and publications since the European Union is a historically unprecedented grand socio-political project. IS building is part of that project's agenda. Evidently, this agenda is more policy oriented.

FAST, the early program of the Commission of the European Communities, subtitled "Sub-Program Information Society—Research Activities," focused on "the societal changes towards which Europe and the world are heading, based on the enormously enhanced capabilities in information handling which microelectronics and associated technologies provide" (FAST, 1980, p. 1). The program emphasized computer technology, microelectronics, and automation. Subsequently, innovation, job creation, representation, and sharing of power, impacts on the way of life (transport, communication, work, leisure), along with distribution of risk and benefits, were the subjects of investigations and debate.

The famous Bangemann Report on "Europe and the Global Information Society—Recommendations to the European Council" stated:

The widespread availability of new information tools and services will present fresh opportunities to build a more equal and balanced society and to foster individual accomplishment. The information society has the potential to improve the quality of life of Europe's citizens, the efficiency of our social and economic organization and to reinforce cohesion. (Bangemann Report, 1994, p. 5)

However, the report concentrates not on technology, but on market-driven changes. It considers issues such as protection of intellectual property rights, privacy, electronic protection, legal protection and security, media ownership, competition policy, and financing.

Another EU document, referred to as the "green paper" on "Living and Working in the Information Society: People First," stresses fundamental social problems, for example, work (skills revolution, job insecurity, new forms of work organization), employment (management of the job transformation process, education and training to match the ICT revolution), social cohesion, public policy, new regulatory framework, human resources, empowerment, and integration (Green Paper, 1996). The EU document, subtitled "A European Way for the Information Society," points out:

[ICTs] are rapidly becoming central to economic, cultural and civic life in the industrial countries. Access to ICT will increasingly be essential to full participation and citizenship…Internet access will become a fundamental right…ICT abolishes distance and ignores borders…The information society is necessarily a global society. (A European Way, 2000, p. 5)

Moreover, it proposes a distinctive European Way based on liberty, equality, fraternity, solidarity, and sustainability. It also addresses other issues, including the rights of the individual, lifelong learning, global communication, new culture of government and of public service oriented towards a "network mentality," global governance, and dialogue (A European Way, 2000). This document of the Information Society Forum encourages joint reflection on the future of the information society through global dialogue based on the necessary core values. This approach is not only normative, but to some extent also policy oriented, because it also provides a set of more concrete recommendations.

The global dimension of the information society was discussed in EU documents even earlier (e.g., I&T Magazine, 1995).

F.R. de Bruïne, of the European Commission DG XIII, explains:

Information superhighway, Infobahn, global networked economy, global information infrastructure, global village…these and other popular labels abound. In Europe, the preferred term 'information society' reflects concern with the wider social and organizational changes that will result from the information and communications revolution. This revolution is driving the transformation from a society based

on physical goods to one increasingly based on knowledge and information. (Bruïne, 1995, p. 10)

In the EU, some other terms like "information society based on knowledge" or "e-Europe" have been recently used (see, e.g., the Lisbon Strategy or the E-Europe Web site). These are extensions of previous definitions and concepts, and serve the debates and policies well.

Information society as a social and political project has always been the subject of interest for government, politicians, international organizations, as well as NGOs and various social movements. Examples follow.

During the 1992 U.S. presidential campaign, Al Gore introduced the issue of national (then global) information infrastructure. A special advisory council at the White House was appointed after the election. The Project of National Infrastructure promised many good jobs for highly skilled persons (e.g., for "symbolic analysts"), mass development of telemedicine, stimulation of democracy ("interactive citizen"), and modernization of education. The actual goals, however, were U.S. competitive advantage in world trade, and technology transfer and research.

G7 and G8 meetings have also dealt with the "new world information order" (in 1995 and 2000) and adopted the notion "global information society." Information superhighways were left to private sector initiatives and market forces in accordance with the recommendation of neo-liberal economics (e.g., presented in the form of "new economy"). E-commerce, digital market, global competition, deregulation, piracy, intellectual property rights, and so forth gained more interest than social visions, even in the EU.

The United Nations—via its agency, UNESCO—organized workshops and conferences (e.g., the World Summit on Information Society in 2003 and 2006) in order to debate the conditions for common access to cyberspace, limiting the digital divide, and introducing NGOs to the discourse, The international discourse on global IS is continued both in global competition—technological and economic, piracy protection, rights of intellectual property, standards, telecom deregulation—and also under the banner of e-governance, strengthening democracy, and global civil society.

BETWEEN THE IS DISCOURSE AND THE REAL PROCESSES AND CONTEXTS

The IS discourse will certainly continue. The above presented definitions and concepts are in fact complementary in spite of their diverse orientations (cognitive, ideological, practical). Therefore what really matters for the future are the real processes and the contexts of the emerging information societies. So not to discuss more on the terms and their interpretations, it seems reasonable to adopt the following general definition: *information societies are societies in which info-activities are prevailing,* which implies sufficient infrastructure, education, access, and efficiency. However, the predominance of info-activities may differ substantially. There are societies able to create ICTs and use them effectively on a large scale like the most advanced countries, sometimes referred to as *high-tech societies.* There are also societies that simply participate, via technology transfer, international technological cooperation, adaptation of info-oriented patterns of education, mass culture, consumption, and so forth, in the civilizational mega trend related to the mass production, wide diffusion, and mass application of information. The leading or pioneering countries (societies) have the potential (educational, scientific, financial, managerial), not to mention the proper infrastructure and government policies, to create the new ICTs and exploit them everywhere. Government orders that stimulate ICTs are often in the field of military systems and equipment connected with automated battlefields, military robots, intelligent ammunition, telecommunications systems supporting info-war, and the like. In the beginning of the information revolution (approximately a half a century ago), the governments of technologically leading countries usually supported intensive research in the R+D sector and also the processes of harnessing new technologies into industry—for example, acting in various ways to diminish risk for companies. This support is important nowadays too. It can be implemented in the form of public-private partnership, favorable legal regulations, and policy measures.

Therefore, the supply side of ICTs and their multifaceted uses and impacts were ensured by the R+D sector–public and private: public university labs, labs of big corporations (like Bell, IBM, Microsoft, Philips), small innovative firms (like Intel or, at one time, Apple), high-tech centers (like Silicon Valley or similar centers, e.g., Bangalore in India or science parks in Europe), and entrepreneurial companies that immediately commercialized emerging innovations and placed them in markets locally and globally. Scientists, governments, and businesses created new knowledge and new products and services based on this knowledge, as well as transformational patterns and potentials for change of all leading societies and—by transfer, exchange, impacts—other societies all over the world. Great TNCs play a tremendous role in diffusion of IC technologies, systems, and products through FDIs with new technology and organization, international trade, joint ventures, and so forth, not to mention the influence of their immense marketing and advertisement efforts.

There is of course the demand side of ICTs and their uses. This side is predominantly stimulated through education, national and global media, imitation of consumption patterns of technologically leading countries, import of

technologically advanced goods, as well as advertisement and marketing of novel technologies and products. Here, psychological and cultural (also religious) openness and capacities are also crucial. Readiness to use ICTs, to adapt to them, to modify our own behavior, and finally to be a part of a global network seem to be the *conditio sine qua non* of the information civilization's success and of social change (both in a collective and an individual sense).

To make potential demand truly effective, some technical and financial barriers should be overcome, for example wide access to infrastructure, inexpensive computers and programs, and cheap Internet access are necessary. Countries with inferior education, low technology, inadequate technical infrastructure, and low income levels have serious problems in joining the "ICT world club." Thus, they should be the object of effective assistance and cooperation from international organizations and businesses as well. Only their inclusion in the processes of ICT globalization will allow achievement of global rationality and sustainability, not only of the technological kind.

Information society (IS) is currently the emerging form of human organization. General IS discourse tends to concern an abstract model (i.e., ideal construct) with its characteristics and interpretations. In the real world, there are only concrete individual information societies emerging mostly in advanced countries. They differ significantly from each other, because of the experience and potentials (educational, technological, economic, cultural, etc.) of their achieved stage of IS development, along with the diversity of their effects and impacts. Furthermore, their strategies, policies, and performance thereof also play extensive roles. Finally, the external environment is undoubtedly crucial in the current highly interdependent and networked world.

As result of all these conditions, potentials, strategies and actions, diversity of impacts, and unequal external dependence, the future of the world's information societies will be quite diversified and heterogeneous. The info-futures will be full of gaps, divides, collisions, and conflicts. Therefore, the idea of a uniform global information society, in spite of the globalization of ICTs and their prevailing impacts, also seems rather utopian due to its incompatibility with the existing and predictable diversity of the world.

FUTURE TRENDS

There are at least two emerging trends in IS research. One is the general, global, and civilizational approach, the other is represented by various disciplines and studies in particular areas. It is also possible to separately categorize investigations focusing strongly on technology from non-technological ones, which reflects "two cultures" in the sciences.

Studies of particular areas are more empirical and should form a cognitive base for more general studies. A cross-disciplinary approach should be recommended, because such is the actual nature of the problem.

Moreover, two other (not only) methodological—problems emerge:

1. whether and how much the general and domain studies reflect the actual transformations in real world (also in the cyber-world), and
2. what kind of feedback dominates between general and particular studies.

Of course, a kind of cognitive time-lag is always possible. Additionally, the ever-present propensity toward utopian (or dystopian) thinking can serve ideological, political, or even religious goals. Increasingly complex reality and progressively narrowing empirical investigations are immense challenges for any reasonable generalization. Thus, the problem at hand and its research should be future oriented and open ended. But how does one avoid excessive value judgments, overstatements, absolutist approach, one-sided opinions, unjustified simplifications, and optimistic or pessimistic emotions notoriously present in theories and their criticisms?

It is interesting to consider which approach connected with understanding of man-computer-cyberspace relations will win. For the time being, there are two presumed attitudes. One treats a computer as a machine to be used as other machines in the past; the same refers to the Net, regarded as an additional social space in which human life existence will carry on as usual. The other emphasizes a strong transforming force embedded in computers and networks which can radically change our lives and ourselves. Therefore, the difference is between *using* the Net and *being* (*living*) in the Net. The increasing convergence of media and VR technologies and their massive diffusion will probably act more for the latter option. The problem is, however, the low predictability of the course of events, behavior, and choices of the next generations brought up in the digital culture (i.e., e-generations).

The discourse on IS is rather controversial regarding civilizational, technological, and especially social and cultural issues. Some authors tend to deal mostly with the technology-man relation, others concentrate on social and cultural change; there are many examples of both orientations.

Some works may already be referred to as classics. Their ideas, theories, and empirical analyses are fundamental (additionally, one can find the comprehensive literature reviews in these works). Their general observations and views were influential and often survived. More contrasts and controversies emerge when considering and debating specific problems and issues. Moreover, the time factor is verifying the past intuitions, hypotheses, and criticisms.

Besides general views concerning IS as a result of

civilizational and technological change modified by various cultures, there are several more particular orientations and fields of IS discourse. They present interesting current and future research opportunities. The following exemplification is limited to keywords reflecting their possible profiles:

- *future or futures:* info-futures;
- *globalization:* global information society;
- *capitalism:* information or digital capitalism;
- *LDCs:* pre-information societies, info, late-comers;
- *culture:* cyber-culture, cyber-art;
- *values:* Internet ethics, multiculturalism, cosmopolitanism;
- *knowledge:* network production of knowledge, knowledge society, new episteme, knowledge property;
- *education:* e-learning, connected intelligence;
- *democracy:* teledemocracy, netocracy, democracy.com, info-civil society, e-democracy;
- *governance:* e-government, e-governance;
- *business:* e-business, e-commerce, e-banking;
- *Internet:* cyber-space, telework, net games, VR, crimes, terrorism;
- *communications:* mobile phone, global media, massive surveillance;
- *society, community, individual:* IS/knowledge-based society, telecommunities, virtual societies, netizens, global netizens, multiidentity, networked invidualism, monadization in the Net; and
- *posthuman era:* post-info-society, bio-info-society, posthuman society, e-herd, e-swarm, human-machine aggregates, intelligent mobs, cyborgization.

There are numerous additional research interests (e.g., power, state, war, work, terrorism, environment, feminism) representing areas also radically transformed by ICTs.

Area studies can generate some important warnings as well as some policy-oriented directives and recommendations. They can stimulate thinking, decision making, and multi-optional acting, along with the greatest possible exploitation of multi-criterial rationality, also in the long term.

CONCLUSION

The problem discussed above has three dimensions. The first refers to *real phenomena and processes* driven by or connected with ICTs. They should be observed, identified, recognized, measured, and analyzed. The second concerns their *interpretations and evaluations.* This is more subjective, normative, often biased, ideologized, and controversial. Controversies even concern fundamental problems like continuity vs. discontinuity in development, transformations of forms vs. substance, meaning and importance of ICT implications, and so on. Thus, the ICT discourse tends to be vivid and, ultimately, inconclusive. The third dimension of the problem, the *future*, significantly reinforces this situation. This dimension is inherently burdened with risks and uncertainties connected with the future course of real processes and human behavior, values, and choices. In spite of all these difficulties and limitations, there are efforts to make the ICT discourse policy-oriented and to use it for elaboration of strategies and policies of governments, businesses, and civil societies.

The discussion presented here is extremely broad, complex, and further complicated because it is relatively new and interdisciplinary. Moreover, entities such as information civilization, information economy, and information society have only emerged in recent decades, while also being highly diversified in time and space. As such, it is very difficult and demanding to investigate processes, which are often radical and hardly predictable especially in their long-term effects and impacts.

As a result, theories, recommendations, and conclusions have rather limited validity (cognitive, scientific, intellectual, practical, political). Apart from being to some extent general and perhaps universal, they ought to relate to real particular societies or other human groupings (entities) that can emerge in the future.

Nevertheless, some concrete practical recommendations seem to be possible for governmental and corporate strategies, for example concerning procedures of e-government or e-business strategies. Some may also be useful for people in their various roles—as citizens, netizens, consumers, electorate, and media auditorium. At present, it is not reasonable enough to elaborate recommendations based on past experience for would-be societal forms such as e-herd, e-swarm, intelligent mobs, for possible posthuman creatures. It can only be advised to stimulate the efforts of our imagination and our alternative thinking abilities, which can be instrumental for elaborating multi-optional scenarios of info-futures.

REFERENCES

Bard, A., & Söderqvist, J. (2000). *Nätokraterna—boken om det elektroniska klassamhället.* Stockholm: BookHouse.

Baudrillard, J. (1981). *Simulacres et simulation.* Paris: Éditions Galilée.

Bell, D. (1973). *The coming of post-industrial society.* New York: Basic Books.

Bezold, C., & Olson, R.L. (1986). *The information millennium: Alternative futures.* Washington, DC: Information Industry Association.

Bugliarello, G. (1997). Telecommunities: The next civilization. *The Futurist, 31.*

Castells, M. (2000). *The rise of the network society* (vol. 1, 2nd ed.). Oxford: Blackwell.

Castells, M. (Ed.). (2004). *The network society. A cross cultural perspective.* Cheltenhem: Elgar.

Currie, W. (2000). *The global information society.* Chichester/New York: John Wiley & Sons.

Dertouzos, M.L. (1998). *What will be—how the new world of information will change our lives*. New York: HarperCollins.

Downing, J.D.H. (2000). *Radical media: Rebellious communication and social movements.* Thousand Oaks, CA: Sage.

Ester, P., & Vinken, H. (2003). Debating civil society: On the fear for civic decline and hope for the Internet alternative. *International Sociology, 18*(4).

Everard, J. (2000). *Virtual states—the Internet and the boundaries of the nation–state*. London/New York: Routledge.

Featherstone, M. (2000). Technologies of post-human development and the potential for global citizenship. In J.N. Pieterse (Ed.), *Global futures—shaping globalization.* London/New York: Zed Books.

Fisher, D., & Wright, L. (2001). On utopias and dystopias: Towards an understanding of the discourse surrounding the Internet. *Journal of Computer Mediated Communication, 6*(2).

Florida, R. (2002). *The rise of creative class.* New York: Basic Books.

Gates, B. (1996). *The road ahead* (2nd ed.). London: Penguin Books.

Gray, C.H. (2001). *Cyborg citizen.* New York: Routledge.

Harman, W., & Markley, O.W. (1985). *Changing images of man.* New York: Pergamon.

Heeks, R.B. (2001). *Reinventing government in the information age.* London: Routledge.

Kamarck, E.C., & Nye, J.S. (Eds.). (1999). *Democracy.com? Governance in a networked world*. Hollis, NH: Hollis.

Kerckhove, D. (1997). *Connected intelligence—the arrival of the Web society.* Toronto: Somerville House.

Lievrouw, L.A., & Livingstone S. (Eds.). (2002). *The handbook of new media.* London: Sage.

Masuda, Y. (1981a). *The information society.* Bethesda, MD: World Future Society.

Mattelart, A. (2001). *Histoire de la societé d'information.* Paris: Editions La Découverte.

May, C. (2002). *The information society—a skeptical view.* Cambridge: Polity Press.

Naisbitt, J. (1982). *Megatrends*. New York: Warner Books.

Negroponte, N. (1996). *Being digital.* New York: Vintage Books.

Rheingold, H. (1993). *The virtual community: Homestanding on the electronic frontier.* Reading, MA: Addison-Wesley.

Rheingold, H. (2002). *Smart mobs—the next social revolution— transforming culture and communities in the age of instant access.* Cambridge, MA: Basic Books.

Schiller, D. (1999). *Digital capitalism.* Cambridge, MA: MIT Press.

Tapscott, D. (Ed.). (1998). *Blueprint to the digital economy: Wealth creation in the era of e-business.* New York: McGraw-Hill.

Toffler A. (1980). *The third wave.* New York: William Morrow.

Toffler, A. (1990). *Power shift—knowledge, wealth, and violence at the edge of the 21st century.* New York/Auckland: Bantam Books.

Toffler, A., & Toffler, H. (1995). *Creating a new civilization—new directions. The politics of the third wave.* Atlanta: Turner.

Van Dijk, J. (2006). *The network society: Social aspects of new media* (2nd ed.). London/Thousand Oaks/New Delhi: Sage.

Virilio, P. (1998). *La bombe informatique.* Paris: Éditions Galilée.

Webster, F. (1995a). *Images of information society.* London: Routledge.

Webster, F. (1995b). *Theories of the information society.* London: Routledge.

Wellman, B. (1999). *Networks in the global village.* Boulder, CO: Westview Press.

Woolgar, S. (2000). *Virtual technologies and social theory.* In R. Rogers (Ed.), *Preferred placement: Knowledge politics on the Web*. Maastricht: Jan van Eyck Editions.

Woolgar, S., & Ingram, C.S. (2000). Virtual society? The social science of new technologies. *Assignation, 17,* 2.

Zacher, L.W. (2006). E-transformations of societies. In *Encyclopedia of digital government* (vol. 2). Hershey, PA: Idea Group.

Zubov, S. (1988). *In the age of the smart machine—the future of work and power.* New York: Basic Books.

KEY TERMS

Future: An image of what may be, composed of a set of elements in which interactions are shaped largely by a set of driving forces; also called a macroscenario (according to Bezold & Olson, 1986, p. A-23).

Future Scenarios: Comprehensive, internally consistent descriptive images of the future based on assumptions about relevant forces of all kinds (technological, economic, political, social, educational, environmental, etc.) together with their multiple interactions; external conditions and influences should be also considered.

Future Types: Possible (open to all possibilities); plausible (an image of the future based on forecasts that might reasonably occur); preferred (the future desirable and sought after).

Info-Futures: Future alternatives of development of societies, organizations, and institutions related to information (production and applications) and ICTs.

Information Era: An era where information is the main strategic resource upon which individuals, organizations, and societies rely for their growth and development. Also called *information millennium.*

Information Society (IS): A society predominantly dealing with production and applications of information in all spheres of life (i.e., economic, social, political, cultural, etc.). It is also assumed that the mass info-activities are based on sufficient infrastructure, access, education, cultural capacity, efficiency, and so forth. Synonyms and closely related terms include *information-rich society, network (or networked) society, e-society, virtual society, information society based on knowledge* (or *knowledge society*).

IS Determinants of Development: There are many—mostly complex, interrelated, and often fuzzy—conditions, factors, and mechanisms (including historical, geographic, economic, technological, educational, social, cultural, political, psychological, etc.) that determine the character, structure, pace, and effectiveness of IS emergence and its progress. They are connected with certain existing potentials, situations, and external context, as well as with deliberate efforts (like business strategies, government policies, activities of international organizations, people's attitudes, education systems, and media presentations of the advanced patterns of IS development).

IS Diversity: The real world's societies are much diversified in terms of the advancement of IS characteristics and indicators. The most developed create ICTs and use them widely and effectively. Most world societies are merely users and imitators (via technology transfer). There are also many limited to being merely impacted by ICTs (via the global Net, FDIs, international trade, etc.).

IS Futures: The future of particular information societies should be presented and debated as multi-optional, not as universal/uniform—that is, the same for all societies. Such is implied by their diversity, which determines their differentiated progress now and in the future. Therefore, any conclusions and recommendations ought to consider their specificity, especially if they have practically oriented ambitions. It is important to remember that all long-term visions, predictions, and forecasts of IS are uncertain, even if we hardly strive for them.

Information Systems and Small Business

M. Gordon Hunter
University of Lethbridge, Canada

INTRODUCTION

The subject area of the application of information systems to small business is a thoroughly interesting, yet relatively under-researched topic. Small business is an important part of any economy. In the United Kingdom, 25% of the gross domestic product is produced by small business, which employs 65% of the nation's workers (Ballantine et al., 1998). In Canada, 43% of economic output is accounted for by small business, employing 50% of private sector employees (Industry Canada, 1997). Further, governments view the small business sector as that component of the economy that can best contribute to economic growth (Balderson, 2000). Given the importance of this sector of the economy, it is incumbent upon researchers and managers of small business to develop a better understanding of how information systems may contribute to the operation and growth of individual businesses as well as the overall sector.

The objective of this article is to provide an overview of information systems used by small business. Research projects are presented that describe the current situation. Recommendations are then proffered for various stakeholders who should contribute to a more effective use of information systems by small business.

BACKGROUND

There does not seem to be a commonly accepted definition of a small business. Thus, individual researchers have adopted a definition for their specific projects. Some definitions include annual revenue, amount of investment, or number of employees. The definition mostly used is number of employees (Longnecker et al., 1997). The European Parliament (2002) has also adopted number of employees as a definition and has further refined the category. Thus, 0 to 10 employees represent micro businesses, small businesses include 10 to 50 employees, and medium businesses have 50 to 250 employees.

Beyond the size aspect of small business, there are others that differentiate them from large businesses. Stevenson (1999) has determined that from a strategic perspective, managers of small businesses tend to respond to opportunities presented by their environment in a multi-staged approach by committing a minimum of resources. Another differentiating factor is "resource poverty" (Thong et al., 1994). This term refers to the lack of time, finances, and human resources.

Laudon and Laudon (2001) suggest that an information system is "interrelated components working together to collect, process, store, and disseminate information to support decision making, coordination, control, analysis, and visualization in an organization" (Laudon & Laudon, 2001, p. 7). As indicated previously, managers of small businesses emphasize short-term decisions in their allocation of scarce resources. However, most information systems require a long-term plan with a significant one-time initial financial commitment. This conflict may result in inefficient investment in information systems, which in turn may negatively impact the financial situation of the small business.

Recent research has supported the contention that the use of information systems by small business represents a unique approach. For instance, Belich and Dubinsky (1999) and Pollard and Hayne (1998) determined that the issues being faced by small business managers (lack of time, skills, and financial resources) are different than those faced by large business managers. Further, Taylor (1999) investigated the implementation of enterprise software in small businesses and found that neither the businesses themselves, nor the software vendors were fully cognizant of the unique problems (matching system capability to functional requirements) encountered by small business managers. Finally, Hunter et al. (2002) identified two major themes regarding small business use of information systems. These themes are "dependency" and "efficiency". The authors suggest that the adoption of information systems increased the small business' dependency on an internal champion, and a series of external stakeholders, including consultants and suppliers. Hunter et al. (2002) suggest this increased dependency results from the approaches to business (Stevenson, 1999) taken by the manager and the concept of resource poverty (Thong et al., 1994). The efficiency theme suggests that small business managers primarily use information systems as an operational tool to help complete daily activities.

Earlier research (Nickell & Seado, 1986) determined that small business was mainly using information systems for accounting and administrative purposes. Research conducted in the 1990s (Berman, 1997; Canadian Federation of Independent Business, 1999; Fuller, 1996; Lin et al, 1993; Timmons, 1999) noted a growing interest by small business in employing information systems for daily operations. While small business has been more than prepared to exploit

the use of information systems to support daily operations (El Louadi, 1998), there exists little evidence that they are prepared to employ the technology in a strategic manner (Berman, 1997). Bridge and Peel (1999) determined that small businesses employed computers mainly to support daily operations and tended not to use them to support decision-making or long-term planning. Current research suggests this situation has not changed. For instance, Dandridge and Levenburg (2000) found that information systems were being employed for daily operations and there was little use of computerization for competitiveness aspects such as accessing the Internet.

A number of research projects have identified that small businesses have not adopted Internet use because of lack of knowledge and experience (Damsgaard & Lyytinen, 1998; Iacovou et al., 1995; Kuan & Chau, 2001). Another set of contributing factors relates to the lack of personnel and time (Bennett et al., 1999). Even when time and personnel are available, there seems to be reluctance by small businesses to investigate the use of the Internet (Chapman et al., 2000).

Burgess and Trethowan (2002) examined the use of Web sites by small businesses, represented by general practitioners, in Australia. They found that while there was reasonably high use of computers to improve efficiency and lower costs, there was not much use of computers for Web sites. Those who had Web sites mainly employed them to provide basic information and contact details.

FUTURE TRENDS

This section presents a number of recommendations for various stakeholders intent on increasing the use of information systems within the small business sector. These recommendations represent suggestions for the future and reflect a synthesis of available literature, presented in the previous section, in the area in conjunction with, and in the specific context of, previous research (Hunter, 2002; Hunter & Long, 2003;Hunter et al., 2002).

Small Business Manager

To overcome the limitations of being dependent upon others' expertise, managers need to gain an understanding of the capabilities of information systems. While managers do not need to know how to design or develop information systems, they do need to understand how technology might be used as a key resource in adding value to the firm's core business products or services.

Further, the small business manager should establish a relationship with a specific individual regarding a source for advice. The recommendation is for the manager to establish a relationship with someone who is independent of a specific solution and who will be prepared to play a strategic role, taking a long-term perspective. It is incumbent upon the manager to review the relationship to ensure that the recommendations being proffered are appropriately contributing to the long-term success of the firm.

Also, managers should take a proactive approach toward the adoption of information systems. This would involve actively seeking out ways to leverage information systems to create or improve products or services offered to customers. However, the manager should avoid being an early adopter of new hardware and/or software applications.

Consultant

It is important for consultants to recognize that in regard to the nature, timing and acquisition of resources, the small business manager generally aims to minimize the amount of resources used at each stage of the firm's growth. Generally, consultants need to be able to provide opportunities for small businesses to "phase in" information systems in stages. Doing so will accommodate small business practice and form the foundation for a mutually beneficial ongoing relationship.

Vendor

Vendors should make a visible commitment to small business through the establishment of an entity specifically directed at small business. The small business sector is a large and important one; thus target marketing this sector makes good business sense. Also, software vendors must ensure that an application performs the necessary functions for small business. It is incumbent upon the vendor to ensure the hardware or software addresses the appropriate functionality of the small business.

Government

Government can help overcome resource poverty by providing advice and financial incentives to small business. By initiating relationships with small business managers, individuals representing government services can more effectively support these managers by tapping into their informal networks to exchange required information. The role of tax and other financial incentives may be employed to encourage the expanded use of information systems.

CONCLUSION

Information systems have increased the efficiency of daily operations for small business.

Generally, evidence suggests that information systems expenditures are being made in reaction to needs or problems, rather than as a result of a long-term coordinated plan. The suggestions included here should contribute to more effective use of information systems by small business.

REFERENCES

Balderson, D.W. (2000). *Canadian entrepreneurship and small business management.* Toronto: McGraw-Hill Ryerson.

Ballantine, J., Levy, M., & Powell, P. (1998). Evaluating information systems in small and medium-sized enterprises: Issues and evidence. *European Journal of Information Systems, 7,* 241-251.

Belich, T.J., & Dubinsky, A.J. (1999, Fall). Information processing among exporters: An empirical examination of small firms. *Journal of Marketing Theory and Practice, 7*(4), 45-58.

Bennett, J., Polkinghorne, M., Pearce, J., & Hudson, M. (1999, April). Technology transfer for SMEs. *Engineering Management Journal,* 75-80.

Berman, P. (1997). *Small business and entrepreneurship.* Scarborough, Ontario: Prentice Hall.

Bridge, J., & Peel, M.J. (1999, July-September). A study of computer usage and strategic planning in the SME sector. *International Small Business Journal, 17*(4), 82-87.

Burgess, S., & Trethowan, P. (2002). GP's and their Web sites in Australia: Doctors as small businesses. *Proceedings of ISOneWorld Conference,* Las Vegas, NV.

Canadian Federation of Independent Business.(1999). Results of members' opinion surveys #37-42. Retrieved August 29, 2000, from *http://www.cfib.ca/research/98internet.asp*

Chapman, P., James-Moore, M., Szczygiel, M., & Thompson, D. (2000). Building Internet capabilities in SMEs. *Logistics Information Management, 13*(6).

Damsgaard, J., & Lyytinen, K. (1998). Contours of diffusion of electronic data interchange in Finland: Overcoming technological barriers and collaborating to make it happen. *Journal of Strategic Information Systems, 7,* 275-297.

Dandrige, T., & Levenburg, N.M. (2000, January-March). High-tech potential? An exploratory study of very small firms' usage of the Internet. *International Small Business Journal, 18*(2), 81-91.

El Louadi, M. (1998). The relationship among organizational structure, information technology and information processing in small Canadian firms. *Canadian Journal of Administrative Sciences, 15*(2), 180-199.

European Parliament. Retrieved July 6, 2002, from *www.europarl.eu.int/dg4/factsheets/en/4_14_0.htm*

Fuller, T. (1996). Fulfilling IT needs in small businesses: A recursive learning model. *International Journal of Small Business, 14*(4), 25-44.

Hunter, M.G. (2002). Information systems development outcomes: The case of song book music. In S. Burgess (Ed.), *Managing information technology in small business: Challenges and solutions* (ch. 3). Hershey, PA: Idea Group Publishing.

Hunter, M.G., Diochon, M., Pugsley, D., & Wright, B. (2002). Unique challenges for small business adoption of information technology: The case of the Nova Scotia Ten. In S. Burgess (Ed.), *Managing information technology in small business: Challenges and solutions* (ch. 6). Hershey, PA: Idea Group Publishing.

Hunter, M.G., & Long, W.A. (2003). Adopting the entrepreneurial process in the study of information systems and small business. In G. Gingrich (Ed.), *Managing information technology in government, business, and communities* (ch. 1). Hershey, PA: IRM Press.

Iacovou, C., Benbasat, I., & Dexter, A. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. *MIS Quarterly, 19*(4), 465-485.

Industry Canada. (1997). *Your guide to government of Canada services and support for small business: Trends and statistics* (Catalogue No. C1-10/1997E). Ottawa: Canadian Government Publishing Centre.

Kuan, K., & Chau, P. (2001). A perception-based model of EDI adoption in small businesses using a technology-organized environment framework. *Information and Management, 38,* 507-521.

Laudon, K.C., & Laudon, J.P. (2001). *Essentials of management information systems – organization and technology in the networked enterprise* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Lin, B., Vassar, J., & Clack, L. (1993). Information technology strategies for small business. *Journal of Applied Business Research, 9*(2), 25-29.

Longnecker, J., Moore, C., & Petty, J. (1997). *Small business management.* Cincinnati: South-Western College Printing.

Nickell, G., & Seado, P. (1986). The impact of attitudes and experiences on small business computer use. *American Journal of Small Business, 101,* 37-48.

Pollard, C., & Hayne, S. (1998). The changing faces of information systems issues in small firms. *International Small Business Journal,16*(3), 70-87.

Stevenson, H.H. (1999). A perspective of entrepreneurship. In H.H. Stevenson, H.I. Grousebeck, M.J. Roberts & A. Bhide (Eds.), *New business ventures and the entrepreneur* (pp. 3-17). Boston: Irwin McGraw-Hill.

Taylor, J. (1999). Fitting enterprise software in smaller companies. *Management Accounting, 80*(8), 36-39.

Thong, J., Yap, C., & Raman, K. (1994). Engagement of external expertise in information systems implementation. *Journal of Management Information Systems, 11*(2), 209-223.

Timmons, J.A. (1999). *New venture creation* (5th ed.). Boston: Irwin McGraw-Hill.

KEY TERMS

Dependency: The requirement to rely upon another individual or organization.

Effectiveness: The ability to accomplish a task with fewer errors.

Efficiency: The ability to accomplish a task with few resources.

Information Systems: Interrelated components working together to collect, process, store, and disseminate information to support decision-making, coordination, control, analysis, and visualization in an organization.

Internal Champion: Highly respected individual within the organization who possesses expertise in a specific area, specifically information systems.

Resource Poverty: The lack of time, finances, and human resources.

Sales/Revenue: Receipt of income for the exchange of goods or services.

Small Business: Various categories may be employed to define this term. Some examples are as follows:

Employees:

- Micro business: 0 – 10 employees
- Small business: 10 – 50 employees
- Medium business: 50 – 250 employees

Stakeholder: An independent party who may have the ability to impact another party.

Strategy: A business plan of action.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1487-1490, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: IT Education

Information Systems Curriculum Using an Ecological Model

Arthur Tatnall
Victoria University, Australia

Bill Davey
RMIT University, Australia

INTRODUCTION

To those of us involved in research and teaching in information systems (IS), it is clear that curriculum innovation and change is complex, and anything but straightforward. The amount of control that individual IS academics have over the curriculum varies between universities. In some cases there is complete control over curriculum content whereas in others just control over delivery with content determined externally. This article concentrates on the former situation but still has some relevance to the later. All curriculum innovation is complex (Fullan, 1993) due to the involvement of a large number of human actors, but in information systems curriculum change this is particularly so due to the need to consider the part played by such non-human actors (Latour, 1996) as the technology itself.

We will argue that if you want to understand *how* IS curriculum is built, you need to use models and metaphors that relate to how people interact with each other, with the environment, and with non-human artifacts. One such approach is provided by the ecological metaphor described in this article in which we argue that systems of education may be seen as ecosystems containing interacting individuals and groups. The interactions between these will sometimes involve co-operation and sometimes competition, and may be interpreted in terms of these forces along with mechanisms for minimizing energy expenditure. In this article we will examine the application of this metaphor to curriculum change in information systems.

BACKGROUND

Nordvall (1982) identifies several models for curriculum change that he suggests all have relevance, in the higher education context, at the subject, course, and institutional levels. These are:

- Research, development, and dissemination models
- Problem solving models; social interaction models
- Political and conflict models
- Diffusion, linkage, or adaptive development models

Models of change based upon a process of research, development, and dissemination (RDD) are probably the most common way of attempting an explanation of the process of curriculum development (Nordvall, 1982). In models like this, relying on logical and rational decisions, curriculum change depends on the use of convincing arguments based on programs of research. A rational and orderly transition is then posited from research to development to dissemination to adoption (Kaplan, 1991). These could then be considered as "manufacturing models" as they follow a fairly logical and straightforward mechanical approach with one thing leading directly to another and do not allow for or consider other influences such as those due to human interactions. If we were to accept a manufacturing model like this, then we might expect some curriculum outcomes to be consistently apparent across the world:

- As research would have shown that several specific programming languages were much more widely used and better to teach than others, all courses requiring programming would use just these few languages, and there would be no arguments regarding the best language to teach.
- As research would show the advantages of object-oriented methodologies all computing courses would teach only these and ignore other approaches.
- The content of courses around the world would be designed to achieve similar goals and outcomes, and contain similar content.
- Research would show the ideal method of teaching computing concepts and issues and classroom delivery of content would be moving toward this researched ideal. Everyone would then use these ideal delivery methods.

It is easy to illustrate that these predictions are not borne out in fact, as programs of study show wide variance within any given country and around the world. Many different programming languages and development methodologies are used, and a wide variety of techniques is adopted for classroom delivery. Some innovations seem to be accepted world wide, but many are accepted only locally. We will here

provide an alternative model that we believe better explains how IS curriculum is really developed.

Metaphors and Models

Before proceeding, however, we need to caution the reader on the limitations of models and metaphors. The dictionary describes a metaphor as a term "applied to something to which it is not literally applicable, in order to suggest a resemblance" (Macquarie Library, 1981, p. 1096). Metaphors are useful, not in giving a literal interpretation, but in providing viewpoints that allows us to relate to certain aspects of complex systems.

We contend that most curriculum models and metaphors are too simplistic to allow a useful view of a curriculum development as a complex system involving human and non-human interactions. In this regard, ecological model offers two main advantages:

- A way of allowing for the inclusion of complexity
- A new language and set of analytical and descriptive tools from the ecological sciences

AN ECOLOGICAL MODEL OF CURRICULUM CHANGE

In ecology organisms are seen to operate within a competitive environment, which ensures that only the most efficient of them will survive. In order to survive, they behave in ways that optimize the balance between their energy expenditure and the satisfaction they obtain from this effort. These two key principles underlie the discipline of ecology, which is concerned with the relationship of one organism to another and to their common physical environment (Case, 2000; Townsend, Harper, & Begon, 2000). Habitat, ecological niches, and the exploitation of resources in predator-prey interactions, competition, and multi-species communities (Case, 2000) are all important considerations in ecology.

We have argued (Tatnall & Davey, 2002, 2003, 2004) that these ideas correspond to the process of curriculum development in that an educational system may be seen as an ecosystem and that the interactions within this can then be analysed in terms of ecological concepts such as competition, co-operative behavior, and niche-development. Curriculum change can be interpreted in terms of mechanisms for minimizing energy expenditure and decisions that individuals make about whether to cooperate or to compete.

In information systems curriculum development we should thus look at all the factors, both human and artifact, to see which could be expected to compete, and which to cooperate to become part of the surviving outcome. A non-human stakeholder such as a development tool or methodology must cooperate with the environment, compete successfully, or die out. This may mean a new curriculum element becomes incompatible with an old element and so replaces it. Alternatively it may mean that two new design tools can be used together, or that a particular curriculum element is compatible, or perhaps incompatible, with the desires and interests of a particular faculty member.

Ecological metaphors have been used in areas other than biology and IS curriculum change. An ecological framework has been used quite successfully in other areas including mathematics curriculum (Truran, 1997) and a study of the effects of violence on children (Mohr & Tulman, 2000). Ecology as a framework tells us to expect progress of a task through cooperative or competitive behaviors of the animate and inanimate factors in the environment. A factor that cannot compete or cooperate is inevitably discarded.

Ecosystems and Complexity

An ecosystem contains a high degree of complexity due to the large number of creatures and species living in it, as well as the variety of interactions possible between each of these. The "ecosystem" represented by the curriculum in a university information systems department contains (at least) the following "species": lecturers, researchers, students, professional bodies, university administrators, and representatives of the computer industry. The "environment" also contains many inanimate objects relevant to the formation of the curriculum, including computers, programming languages, textbooks, lecture rooms, analysis and design methodologies, networks, laboratories, programming manuals, and so on.

Curriculum development can be seen as attempting to introduce change within an ecosystem. The problem, of course, is the large number of interested parties that must be contended with before change can be implemented. Curriculum development is more complex than resolving the conflicting needs of students, employers, academics, and the academy. There is ongoing conflict between many things such as educational philosophies, pedagogical preferences, perceived resource constraints, and personal issues. To investigate the interrelationships between these entities we will look now at competition, co-operation, niche formation, and energy expenditure.

Competition

Competition in nature can occur both within and between species. In many species the males compete with each other for mates, while different species of fish compete for the best feeding areas. In IS curriculum we see many examples of competition, some of which are useful in determining the "fittest" topics and techniques best suited for survival (Darwin, 1859) in the curriculum, while others involve time-wasting clashes of personality between academics.

One example of competition seen in recent years in many IS departments is in programming between .Net and Java. The advocates of Java will contend that its use in producing Web-based applications and its non-proprietary nature mean that it is the best language to teach. .Net advocates, on the other hand, argue that while this may be so .Net is easier to use, and being backed by Microsoft has a considerable advantage in its use by industry. This, they contend, makes it the best vehicle to introduce students to programming. The result of this competition is, most likely, that one language will survive in the curriculum and the other die out. Similar examples can often be seen in competition between different methodologies and between software products. Most university courses now make use of Microsoft Office rather than Lotus, Word Perfect and the like, as Microsoft has clearly won the competition and become dominant in this area.

Co-Operation

There are many examples of unexpected co-operation between organisms in nature: the oxpecker bird that lives with a rhinoceros, sharks, and suckerfish, barnacles that attach themselves to whales, and dogs and cats living in close proximity with people. It is also possible to think of an organism living in co-operation with its environment: something the native peoples of many countries speak about.

In an educational program such as an IS degree some courses rely on earlier courses, that is, they have prerequisites. This can be seen as a form of co-operation in which each course benefits from the existence of the other. Another similar example is in software and programming languages where, for instance, the use of VB in a computer laboratory requires the presence (and co-operation) of Microsoft Windows. Likewise subject material that relies on the use of a specific textbook could also be seen as an example of co-operation.

Ecological Niches

An ecological niche is a place where a particular species that is well suited to this environment is able to thrive, where other species may not. A curriculum example of this is in the teaching of the PICK operating system by a university in Australia. Some years ago PICK was a serious challenger to UNIX for the "universal operating system" in business, but PICK has now decreased in importance. Despite the fact that no other university in the region now teaches it, and its place being challenged by more recent operating systems, PICK has remained in the curriculum of this university. It has remained largely because an academic involved in its teaching was able to argue convincingly (Tatnall & Davey, 2001) that learning PICK allowed students to take up jobs in the small number of prominent local industries using this system: that it filled an important ecological niche.

Energy Expenditure

It is easy to find examples of minimization of energy expenditure in curriculum development in the use of curriculum templates and the copying of curriculum from other institutions. Perhaps the greatest reduction in energy expenditure can be gained by using, without change, a model curriculum or the curriculum from another university. A related example is seen in choosing curriculum elements so that they fit in with existing university resources.

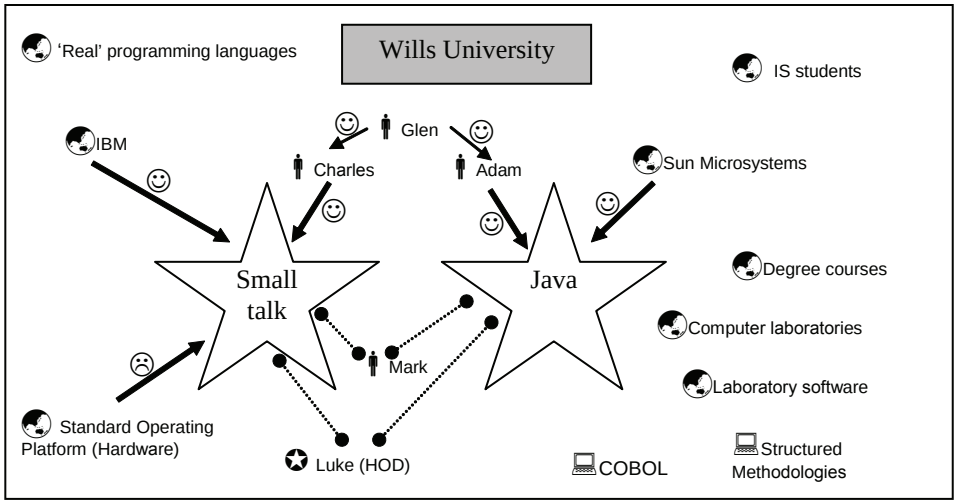
AN EXAMPLE: CHOOSING BETWEEN OBJECT TECHNOLOGIES

This example involves introduction of object-oriented (OO) technology into an undergraduate degree curriculum at Wills University in 2003 (although the case is real, the name of the university is fictitious). The non-human organisms in this case include existing structured methodologies and technologies such as structured analysis and design, and structured programming languages such as COBOL (Tatnall & Davey, 2004). New contenders with some support amongst the humans included an IBM version of Smalltalk and surrounding object-oriented methods in direct opposition to the Java tool set from Sun Microsystems.

In retrospect it can be seen from study of the minutes of meetings and discussion papers that a small group of human "organisms" played significant parts in the ecosystem during the introduction of the technology. These humans included an academic with ultimate responsibility for signing off any change, the existing academics in charge of target subjects, and the co-ordinator of teaching into the degree. The issue of changing technology had been raised several times over a number years, and the University was clearly several years behind in making decisions on this technology. During this period both the head of the department and the controller of the degree had been co-operating to minimize energy expenditure so that the output of the degree was constant (and a zero input for some output can be seen as very efficient). This co-operation between organisms can be seen to explain conservative behavior in organizations, but such a view overlooks the amount of energy required to overcome the desire of other organisms to make change. In this case a predator arrived in the environment in the form of a visiting professor who had been funded by the department to come to Australia to contribute to their research efforts. As an "outsider" the professor could be expected to have little impact as many factors existed in the environment strongly promoting the technology, and these had resulted in no change.

The effect of the visiting professor, however, was marked, and one of the academics involved with the target subjects used the forums provided to mount a campaign for change.

Figure 1. Wills University environmental interactions diagram (Adapted from Tatnall & Davey, 2004)



This co-operation between predators was aimed at the grazing organisms in that the visiting professor was motivated to have *some* effect occur as a result of his funded trip, and the interested academic could establish a reputation by being associated with such a change. At this stage the exact nature of the technology could be seen as irrelevant, as almost none of the relevant decision makers had any knowledge of the technology. As pressure mounted to make change the two other main organisms (academics) arose as interacting entities. Whereas one was championing the IBM Smalltalk technology, the other had been researching Java. In each case the co-operation between the technology and the human stemmed from invested time in gaining knowledge of the technology. A significant problem, however, was that the university's standard hardware operating platform did not contain sufficient memory to run Smalltalk satisfactorily.

Eventually the decision makers were spending so much energy in containing the move to new technologies that a decision was forced upon them. At this stage a coalition of predators achieved dominance in the environment as shown in Figure 1. Faced with this array of predators, the Java technology achieved dominance.

FUTURE TRENDS: APPLICATION OF THE ECOLOGICAL MODEL

No one individual can be identified as the main element in seeking to introduce the object-oriented technologies into the degree course. Let us suppose that an academic with a strong interest in OO had desperately wanted to introduce Java into the curriculum. He or she would have quickly seen the visiting professor as an ally with whom he or she could co-operate and then gone in a search for other allies, and any potential competitors. When it became clear that the main problem was one of energy expenditure—"why change when all is going well?"—he or she could have specifically addressed this by attempting to show how the change was inevitable and how inaction would just result in more energy expenditure at a later time.

CONCLUSION

Researchers investigating curriculum development, or any other field, must use language in framing their research questions. The language used often reflects a general viewpoint of the field being investigated and will always embody some metaphor for the principle components of the field. The metaphor is not useful in *proving* relationships but can be used to convey meaning once relationships are discovered, and an appropriate metaphor can lead the researcher toward or away from useful possible conclusions. Many of the metaphors for curriculum development are simple ones from areas such as the manufacturing-type research, development, and dissemination models described earlier. Any investigation of development processes in rapidly changing areas such as information systems shows that a common factor is complexity. This leads the search for a suitable metaphor to those disciplines that have accommodated complexity. One such area is ecology, and we have shown how ecological principles appear to provide good descriptions of common curriculum development activities. The ease with which the metaphor can be used to describe

Table 1. Application of the ecological model

Step	Ecological model	Example
1.	Examine the environment in which the IT curriculum change occurs.	This might be a university department of information systems or computer science, a departmental sub-unit, or perhaps an entire university faculty.
2.	Look for all relevant entities that might constitute this ecosystem.	Academics, students, university administrators, course advisory committees, local industry representatives, computer networks, programming languages, development methodologies, text books, courses of study, university handbooks, and so on.
3.	Look at all interactions between entities and classify these as co-operative, competitive, or niche forming.	Some interactions may not easily fit a single category.
4.	Look for examples of potential co-operation and for co-operative entities.	One academic course can be seen to co-operate with another by acting as a pre-requisite. Visual Basic requires the co-operation of Microsoft Windows in laboratory computers in order to operate.
5.	Look for examples of potential competition and for competitive entities.	The Java and .Net programming systems can be seen to be in competition with each other. Another example is OO development methodologies and conventional structured methodologies.
6.	Look for potential niche applications	In one particular university the teaching of the PICK operating system and PICK basic constitutes a niche application, as it prepared this institution's students for work in specific local companies.
7.	Look at the level of energy expenditure (both in keeping the current curriculum in place and in introducing change).	How much energy does a reactive head of a department have to expend to prevent change? Alternatively, how much energy does an enthusiastic faculty member have to expend to bring about drastic change? (There are, of course, many other possibilities regarding energy expenditure between these extremes.)

actions within IS curriculum development shows that it can be useful as a set of language elements that might lead the researcher to framing useful questions that do not trivialize the complexity of the field.

IS curriculum development involves a complex process of negotiation between actors, and one that cannot be simply explained by reference to a set process of referring new ideas to a series of university committees. The choices of individual academics, or groups of academics, to adopt or ignore a new concept or technology, and to compete or co-operate, must also be considered. This inevitably involves a negotiation process between many different actors. We have argued that this negotiation process can be analyzed in terms of ecological behavior and have utilized an ecological metaphor to assist in visualizing the curriculum development process.

REFERENCES

Case, T. J. (2000). *An illustrated guide to theoretical ecology*. New York: Oxford University Press.

Darwin, C. (1859). *On the origin of species by means of natural selection* (Folio Society ed., 2006). London: Folio Society.

Fullan, M. (1993). *Change forces: Probing the depths of educational reform*. London: The Falmer Press.

Kaplan, B. (1991). Models of change and information systems research. In H.-E. Nissen, H. K. Klein, & R. Hirschheim (Eds.), *Information systems research: Contemporary approaches and emergent traditions* (pp. 593-611). Amsterdam: Elsevier Science Publishers.

Latour, B. (1996). *Aramis or the love of technology*. Cambridge, MA: Harvard University Press.

Macquarie Library. (1981). *The Macquarie dictionary*. Sydney: Macquarie Library.

Mohr, W. K., & Tulman, L. J. (2000). Children exposed to violence: Measurement considerations within an ecological framework. *Advances in Nursing Science, 23*(1), 59-67.

Nordvall, R. C. (1982). *The process of change in higher education institutions*. Washington, DC: American Association for Higher Education.

Tatnall, A., & Davey, B. (2001). How visual basic entered the curriculum at an Australian university: An account informed by innovation translation. In E. Cohen (Ed.), *Challenges to informing clients: A transdisciplinary approach (Informing Science 2001)* (pp. 510-517). Krakow, Poland.

Tatnall, A., & Davey, B. (2002). Information systems curriculum development as an ecological process. In E. Cohen (Ed.), *IT education: Challenges for the 21st century* (pp. 206-221). Hershey, PA: Idea Group Publishing.

Tatnall, A., & Davey, B. (2003). ICT and training: A proposal for an ecological model of innovation. *Educational Technology & Society, 6*(1), 14-17.

Tatnall, A., & Davey, B. (2004). Improving the chances of getting your IT curriculum innovation successfully adopted by the application of an ecological approach to innovation. *Informing Science, 7*(1), 87-103.

Townsend, C. R., Harper, J. L., & Begon, M. (2000). *Essentials of ecology*. Boston: Blackwell Science.

Truran, J. M. (1997). Reinterpreting Australian mathematics curriculum development using a broad-spectrum ecological model. In *Old boundaries and new frontiers in histories of education: Australian and New Zealand History of Education Society Conference* (pp. 241-262). Newcastle, Australia: The University of Newcastle.

KEY TERMS

Competition: When two individuals or species are in competition with each other, they are each striving for the same thing. In biological systems it is typically food, space, or some other physical need, but in the model described in this article, it can be any matter relating to IS curriculum. When the thing they are striving for is not in adequate supply for both of them, the result is that both are hampered, or adversely affected, in some manner.

Co-Operation: Occurs when one species works with another in order to achieve an outcome beneficial to one or both. Proto co-operation is the situation in which both benefit by the co-operation, but can survive without it. Mutualism occurs when each benefits and cannot otherwise survive. Commensalism occurs when two species habitually live together; one species being benefited by this arrangement and the other unharmed by it.

Ecological Metaphor: A way of describing a complex situation, such as IS curriculum development, by providing a way of allowing for the inclusion of complexity, and a language and set of analytical and descriptive tools from the ecological sciences.

Ecological Niche: A place where a particular species that is well suited to this environment is able to thrive, where other species may not.

Ecosystem: In the context of this article, the ecosystem represented by the curriculum in a university information systems department contains (at least) the following "species": lecturers, researchers, students, professional bodies, university administrators, and representatives of the computer industry.

Metaphor: A term applied to something to which it is not literally applicable, in order to suggest a resemblance.

Minimization of Energy Expenditure: A principle of ecology in which a species uses the least possible amount of energy to achieve its purpose.

Information Systems Research Relevance

Shirish C. Srivastava
National University of Singapore, Singapore

Thompson S. H. Teo
National University of Singapore, Singapore

INTRODUCTION

Though there have been extended deliberations for making information systems (IS) research more relevant[1] and useful for IS executives, to our knowledge, there has been no empirical study which examines the *extent of relevance* in the current IS research. In this chapter, we analyze the *topical relevance* of 388 published academic articles in the three top IS journals: *MIS Quarterly (MISQ), Information Systems Research (ISR),* and *Journal of Management Information Systems (JMIS),* for a 5 year period from 2000-2004. We do this by examining their *fit* with the key issues for information technology (IT) executives identified by the latest Society for Information Management (SIM) survey. Based on our results, we make recommendations for making IS research more meaningful for practitioners.

BACKGROUND

The importance of relevance of IS research has been highlighted by a number of scholars. For example, Benbasat and Zmud (1999) quoting from the title of a 1990 Business Week article highlighted that useful research should not be "in the ivory tower, fuzzy, irrelevant and pretentious" (p. 3) rather it should be relevant for practitioners. In a similar vein, there has been a growing debate about crisis in the IS discipline. Scholars have identified various ways of resolving this crisis (Agarwal & Lucas, 2005; Benbasat & Zmud, 2003; Hirschheim & Klein, 2003; Lucas, 1999; Markus, 1999). Some of them have recommended the need for according greater sociopolitical and cognitive legitimacy by addressing the needs of all stakeholders of IS research (Aldrich, 1999). As IS professionals form a major part of the IS discipline stakeholders, an important way in which the IS discipline can address their needs is by making research more useful and relevant for them (Agarwal & Lucas, 2005; Benbasat & Zmud, 2003).

Scholars like Davenport and Markus (1999) and Benbasat and Zmud (1999) view the goal of "research relevance" as critical to the long term survival and success of the field. They suggested that IS researchers can make their studies more relevant by choosing research topics which are considered to be important by practitioners. Three of the four dimensions of relevance identified by Benbasat and Zmud (1999) pertain to the content of articles (interesting, applicable, and current) as shown in Table 1.

IS research has often been criticized for its failure to address the issues relevant for business practitioners (Saunders, 1998; Zmud, 1996a, 1996b). Researchers have been exhorted to incorporate greater relevance in their research, in addition to instilling research rigor, to make it more useful and applicable for practitioners (Benbasat & Zmud, 1999; Davenport & Markus, 1999). However to date, there have been no studies which assess the "extent of relevance" of current IS research. Over half a decade after the deliberations of senior IS researchers on the issue of instilling greater relevance in IS research, it is an opportune time to take stock of the practitioners' concerns in subsequent IS research. In this study, we address this vital issue about relevance in IS research in three ways. First, we investigate how relevant for practitioners is the current IS research? For doing this we analyze the topical relevance of the published IS research, which encompasses the dimensions of articles' content (interesting, applicable, and current). Second, we develop a measure called the journal relevance coefficient (JRC)

Table 1. Dimensions of research relevance (Benbasat & Zmud, 1999)

Category	Dimensions of Relevance	Description
Content	Interesting	Does IS research address the problems or challenges that are of concern to IS professionals?
	Applicable	Does IS research produce the knowledge and offer prescriptions that can be utilized by practitioners?
	Current	Does IS research focus on the current technologies and business issues?
Style	Accessible	Are IS research articles able to be understood (in terms of tone, style, structure, and semantics) by IS professionals?

which is used to assess the relevance of IS journals. Third, we provide specific recommendations for addressing the concerns of IS practitioners.

METHODOLOGY

For seeking an answer to our research question about the relevance of current IS research, we examine the *fit* of the relevant topics identified by the IS professionals, with the topics of research in the top three IS journals namely, *MISQ, ISR, and JMIS* for 5 years (from 2000-2004). We posit that the top three IS journals are a reflection of IS academic research.

The key issues identified in the sixth[2] formal survey by the SIM have been used as current concerns of IS professionals (Table 2) for conducting our analysis in this study (Luftman & McLean, 2004). The sixth formal survey was authorized by the SIM executive board, 9 years after the last formal survey was conducted in 1994 (Brancheau et al., 1996).

We analyzed 388 articles published during last 5 years (2000-2004) in the three top IS journals: 104 in *MISQ*, 111 in *ISR*, and 173 in *JMIS*. Based on the topics and the abstracts of the articles, we identified the dominant concern being addressed in each article. If the content matched with one of the top 20 concerns identified in the latest SIM survey (Luftman & McLean, 2004), it was grouped there, otherwise it was grouped into a 21st category: others. The *others* category indicates that the dominant theme of the article does not address any of the concerns mentioned in the top 20 concerns.

Table 2. IT management concern: Ranking of importance (Luftman & McLean, 2004)

Rank	IT Management Concern
1	IT and business alignment
2	IT strategic planning
3	Security and privacy
4	Attracting, developing, and retaining IT professionals
5	Measuring the value of IT investments
6	Measuring the performance of the IT organization
7	Creating an information infrastructure
8	Complexity reduction
9	Speed and agility
10	IT governance
11	Business process reengineering
12	Introducing rapid business solutions
13	Evolving CIO leadership role
14	IT asset management
15	Managing outsourcing relationships
16	Leveraging the legacy investment
17	Sarbanes-Oxley Act of 2002
18	Globalization
19	Offshore outsourcing impacts on IT careers
20	Societal implications of IT

DATA ANALYSIS AND DISCUSSION

To our knowledge, there are no known measures for analyzing the relevance of academic journals. For this study, to understand the relevance of journals, we followed a two pronged approach. First, we developed a measure called JRC to analyze the aggregate relevance trends of journals across the years. Second, we analyzed the data to understand the trends in terms of topics in IS research.

Journal Relevance Coefficient

The aim of the JRC is to understand the relevance aspect of the published articles in the three journals, in an aggregate way. For calculating JRC, we use the following methodology. First, we assign a weight to each of the 21 issues identified in the IT executive survey (Table 2). For simplicity we assign equal interval weights in the reverse direction of ranks, for example, the top ranked topic *IT and business alignment* is assigned a weight of 21, whereas the last ranked item *others* is assigned a weight of 1. Next, we multiply these weights to the corresponding values (or frequencies of articles). This gives us a rank weighted table across the years for each of the journals. Next we sum up the weighted value for each year for each journal. To calculate the JRC, we divide this value in each year for each journal by the maximum possible value that can be attained (i.e., assuming all the articles in that year address the top concern for IT professionals, a weight of 21). The resulting value gives the JRC for that journal, for that particular year. JRC expresses in a uniform way the extent of relevance being addressed by the journals in a particular year. In notational terms, this can be expressed as follows:

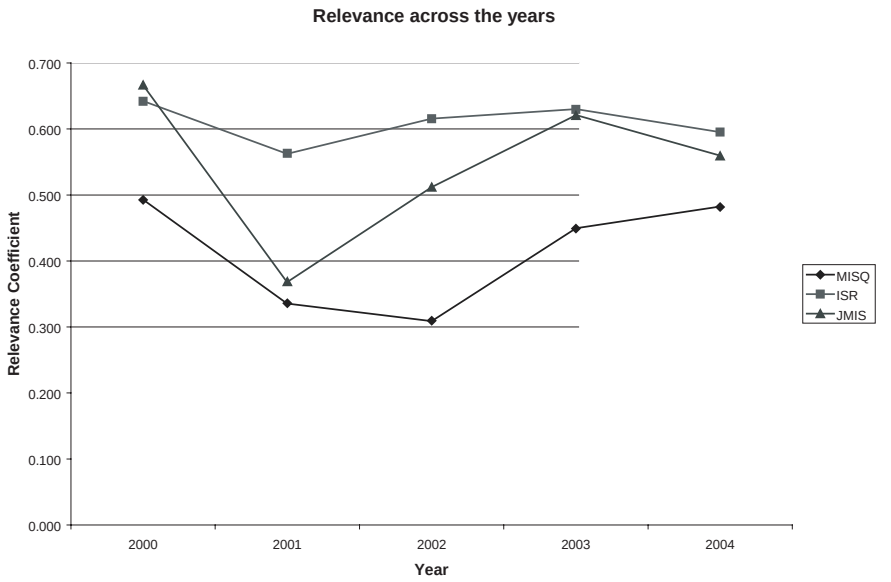
$$(JRC)_{j,y} = \frac{\sum_{i=1}^{n} x_i w_i}{N \sum_{i=1}^{n} x_i}$$

,where $(JRC)_{j,y}$ is the journal relevance coefficient for journal j (MISQ, ISR, or JMIS) for year y (2000-2004), n is the rank of topics identified as relevant for practitioners x_i is the number of articles for the i^{th} rank and w_i is the weight assigned for i^{th} rank article and N is the total number of articles analyzed.

The JRC is a fair indicator of the relevance of published research in the journals and can be used to compare the *relevance* of the IS journals across the years from 2000-2004. Figure 1 shows the movement of relevance coefficients across the years for the three journals in this study: *MISQ*, *ISR*, and *JMIS*.

From the chart in Figure 1, we observe that *ISR* is cur-

Figure 1. Journal relevance coefficients over the years



rently the *most* relevant journal for IS professionals *in terms of topics studied* and *MISQ* is the *least*. The interpretation of these results comes with some caveats (1) the analysis takes into consideration only the "topical relevance," it is possible that the writing style is more practitioner friendly in *MISQ*; (2) *MISQ* also publishes another journal especially for practitioners namely, *MISQ Executive*, where the issues discussed tend to be more relevant for practitioners; and (3) we have assumed that the relevant topics identified in the survey of 2003 do not change remarkably in the period from 2000-2004, but the dynamism of IT may inhibit this stability of relevant topics. Even with these limitations, the point to be observed is that the current trend (2003-2004) for *ISR* and *JMIS* is a falling one in terms of relevance, whereas for *MISQ* it is an increasing trend, which is encouraging for *MISQ*.

Though there is an increasing trend in the relevance of IS articles for the top three journals, the absolute values of the JRC still range between 0.48 and 0.60 for the three journals in the year 2004 as shown in Table 3. This broadly signifies that we are addressing about 60% of the concerns of the IS executives. There is a greater need and scope for addressing the needs of IS practitioners in terms of selection of research topics.

From the detailed account of topical analysis of journals, we observe a perceptible gap in the requirements of the IS executives and the actual research by IS academics. Combined topic-wise academic research published in all the three journals is indicated in Table 4.

We observe that comparatively less research (< 14%) has been done on the top *four* issues identified by the SIM survey (Luftman & McLean, 2004). Hence this study identifies four areas where research should be taken up by academics, namely: (1) IT and business alignment; (2) IT strategic planning; (3) security, and privacy; and (4) attracting, developing, and retaining IT professionals.

Apart from the under-researched areas, we observe that there are some areas which have been the topic for most of the research in the IS field. For example, measuring value of IT investment and measuring the performance of IT organization form about 19% of the total IS research. Similarly IT governance is also an area which has been paid a lot of attention by IS researchers and forms over 13% of the total IS research.

Table 3. Journal relevance coefficient (JRC) across the years for the three journals

	2000	2001	2002	2003	2004	Total
MISQ	0.49	0.34	0.31	0.45	0.48	0.43
ISR	0.64	0.56	0.62	0.63	0.60	0.61
JMIS	0.67	0.37	0.51	0.62	0.56	0.54

RECOMMENDATIONS

Based on this study, we offer a set of recommendations for IS researchers and academics.

Table 4. Aggregate topic wise published research in the three journals (2000-2004)

Rank	Issues	MISQ	ISR	JMIS	Total	Percent
1	IT and business alignment	3	2	1	**6**	**1.55**
2	IT strategic planning	2	3	8	**13**	**3.35**
3	Security and privacy	0	7	9	**16**	**4.12**
4	Attracting, developing, and retaining IT professionals	5	6	7	**18**	**4.64**
5	Measuring the value of IT investments	7	18	24	**49**	**12.63**
6	Measuring the performance of the IT organization	3	11	10	**24**	**6.19**
7	Creating an information architecture	7	13	9	**29**	**7.47**
8	Complexity reduction	2	5	5	**12**	**3.09**
9	Speed and agility	3	4	5	**12**	**3.09**
10	IT governance	16	12	24	**52**	**13.40**
11	Business process reengineering	2	1	4	**7**	**1.80**
12	Introducing rapid business solutions	11	4	18	**33**	**8.51**
13	Evolving CIO leadership role	1	1	3	**5**	**1.29**
14	IT asset management	1	2	8	**11**	**2.84**
15	Managing outsourcing relationships	1	4	1	**6**	**1.55**
16	Leveraging the legacy investment	1	1	0	**2**	**0.52**
17	Sarbanes-Oxley Act of 2002	0	0	0	**0**	**0.00**
18	Globalization	0	0	1	**1**	**0.26**
19	Offshore outsourcing impact on IT careers	1	0	0	**1**	**0.26**
20	Societal implications of IT	5	6	8	**19**	**4.90**
21	Others	33	11	28	**72**	**18.56**
	Total	**104**	**111**	**173**	**388**	**100.00**

- **Recommendation #1. Researchers, while choosing research topics should consider not only the theoretical significance of the topics but also their relevance for practitioners:** As highlighted by earlier researchers (Benbasat & Zmud, 1999; Davenport & Markus, 1999), business is an applied field and the needs of the IS executives should be addressed more closely. Our study indicates that only about 48-60% of the concerns of the IS professionals are being addressed by the current research, hence there is a need for instilling greater relevance in IS research.
- **Recommendation #2. Regular surveys to feel the pulse of IT executives should be conducted to provide guidelines for practical research to academics:** For conducting research on topics relevant for IS executives, it is important to know the topics which are interesting, applicable, and current for them. This can be known only if we have regular surveys like the one conducted by SIM in 2003. This particular survey was conducted by SIM after a gap of 9 years (the last one was in 1994). The conduct of surveys should be institutionalized on a regular basis and the results should be proactively propagated to IS academics.
- **Recommendation #3. Professional bodies and conferences should proactively assist in disseminating information and details about conducting relevant research:** Professional bodies like Association of Information Systems (AIS), IS conferences [like International Conference on Information Systems (ICIS), International Federation for Information Systems (IFIP), AMCIS (Americas' Conference on Information Systems) , etc.] can be used as a platform for disseminating information about relevant topics. This will increase the awareness of IS academics about these topics; also the institutional endorsement of these topics will add to the value of doing research on the identified topics by the IS researchers.
- **Recommendation #4. There is a greater need to study the topics of IT and business alignment; IT strategic planning; security and privacy; and attracting, developing, and retaining IT professionals:** Our study indicates that the aforementioned top four key issues for IS executives are highly under researched in terms of published papers in the top three IS journals. There appears to be an imperative need to incorporate these as important topics of research by IS academics.

FUTURE TRENDS

Results from our research reveal that the issue of relevance for IS executives is not being adequately addressed by current IS research. Based on the results, we make recommendations for increasing the relevance of future IS research for practitioners. We also identify four top concerns of the IS professionals, which have not been adequately researched, namely (1) IT and business alignment; (2) IT strategic planning; (3) security, and privacy; and (4) attracting, developing, and retaining IT professionals. Future IS researchers must

focus on these topics to address the requirements of the IS practitioners thereby making research more meaningful for them.

Though in our study we restricted our analysis to the topical relevance, it is equally important to make the article's style amenable for the IS practitioners, which implies that the tone and language should be direct, simple, and easily comprehensible by practitioners (Benbasat & Zmud, 1999). This important aspect of relevance can be analyzed in detail by future research.

CONCLUSION

This article revisits the often debated question about the relevance of current IS research. Though there has been a lot of discussion and deliberation on the issue, to our knowledge, there has been no empirical investigation about the extent of relevance of current IS research. Our empirical study, which investigates the extent of relevance of published research for the last 5 years (2000-2004) in the three top journals in the field of IS: *MISQ, ISR*, and *JMIS* provides a broad overview of the entire discipline and its focus on the requirements of IS practitioners. We develop a measure for estimating the relevance of IS journals (JRC), which can be used by future studies for estimating the extent of relevance exhibited by various journals. A periodical evaluation of the JRC can help us assess if we are adequately addressing the needs of the IS practitioners through our research.

The results in this study indicate that the current level of relevance of top three IS journals is grossly inadequate (the maximum value of JRC is 0.60, for the year 2004). This implies that for ensuring long term survival and growth of the IS discipline, the topics relevant for IS executives have to be studied in a more organized fashion, so that this important stakeholder group understands the value added by research for them.

REFERENCES

Agarwal, R., & Lucas, H. C., Jr. (2005). The information systems identity crisis: Focusing on high-visibility and high-impact research. *MIS Quarterly, 29*(3), 381-398.

Aldrich, H. E. (1999). *Organizations evolving.* Thousand Oaks, CA: Sage.

Ball, L., & Harris, R. (1982). SIM members: A membership analysis. *MIS Quarterly, 6*(1), 19-38.

Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. *MIS Quarterly, 23*(1), 3-16.

Benbasat, I., & Zmud, R. W. (2003). The identity crisis within the IS discipline: Defining and communicating the discipline's core properties. *MIS Quarterly, 27*(2), 183-194.

Brancheau, J. C., Janz, B. D., & Wetherbe, J. C. (1996). Key issues in information systems management: 1994-95 SIM Delphi results. *MIS Quarterly, 20*(2), 225-242.

Brancheau, J. C., & Wetherbe, J. C. (1987). Key issues in information systems management. *MIS Quarterly, 11*(1), 23-45.

Davenport, T. H., & Markus, M. L. (1999). Rigor and relevance revisited: Response to Benbasat and Zmud. *MIS Quarterly, 2*(1), 19-23.

Dickson, G. W., Leitheiser, R. L., Wetherbe, J. C., & Nechis, M. (1984). Key information systems issues for the 1980's. *MIS Quarterly, 8*(3), 135-159.

Hirschheim, R., & Klein, H. K. (2003). Crisis in the IS field? A critical reflection on the state of discipline. *Journal of the AIS, 4*(5), 237-293.

Lucas, H. (1999). The state of the information systems field. *Communications of the AIS Systems, 5*(1), 1-6.

Luftman, J., & McLean, E. R. (2004). Key issues for IT executives. *MIS Quarterly Executive, 3*(2), 89-103.

Markus, M. L. (1999). Thinking the unthinkable: What happens if the IS Field as we know it goes away? In W. Currie & R. Galliers (Eds.), *Rethinking management information systems* (pp. 175-203). Oxford: Oxford University Press.

Niederman, F., Brancheau, J. C., & Wetherbe, J. C. (1991). Information systems management issues in the 1990's. *MIS Quarterly, 15*(4), 474-499.

Sarbanes Oxley Act (2002). *Sarbanes Oxley Act 2002.* Retrieved June 25, 2006, from http://www.sarbanes-oxley.com/

Saunders, C. (1998, Winter). The role of business in IS research. *Information Resources Management Journal,* 4-6.

Zmud, R. (1996a). Editor's comments. *MIS Quarterly, 20*(2), 21-23.

Zmud, R. (1996b). Editor's comments. *MIS Quarterly, 20*(3), 37-38.

KEY TERMS

Discipline Stakeholders: Discipline stakeholders are all the logical segments of population who are impacted by and also impact the research in a particular discipline, for example, the stakeholders in IS discipline are IS and non IS academics, students, and industry and business practitioners.

Information Systems (IS) Practitioners: IS practitioners are the professionals involved with planning and implementing IT resources for their organizations which includes chief information officers (CIOs), IT managers, and other professionals with similar job descriptions.

Journal Relevance Coefficient (JRC): JRC is a measure developed in the current study which can be used to ascertain the extent of *topical relevance* addressed to by the journals.

Research Relevance: Research in applied fields like IS has to be responsive to the needs of business and industry to make it useful and practicable for them. There are four dimensions of relevance in research which deal with the content and style of research articles. For an article to be relevant it must not only be interesting, applicable and current to the needs of the practitioners but should also be written in an accessible and simple style.

Research Rigor: There are two dimensions of rigor in research. *Methodological rigor* implies following established methodological procedures and *philosophical rigor* signifies incorporating theoretical basis into research. For a piece of research to be justified as an academic piece it must have substantial rigor on both these dimensions.

SIM Survey: Society for Information Management (SIM) conducts surveys to assess the needs of the IT practitioners and identifies current areas and topics of concern for the IT executives. Last such survey was done in 2003, the findings of which are available in Luftman and McLean (2004).

Topical Relevance: Topical relevance means that the research topic is useful for the practitioners. This implies that the topic is interesting, applicable, and current to the needs of business and industry.

ENDNOTES

1 Useful and applicable in practice.

2 Ball and Harris (1982) conducted the first survey and produced a list of 18 issues. Subsequently, surveys were held to identify the top IT management concerns by Dickson, Leitheiser, Wetherbe, and Nechis (1984), (1984), Brancheau and Wetherbe (1987), Niederman, Brancheau, and Wetherbe (1991), and Brancheau, Janz, and Wetherbe (1996).

Category: Business Information Systems

Information Technology Business Continuity

Vincenzo Morabito
Bocconi University, Italy

Gianluigi Viscusi
University of Milano, Italy

INTRODUCTION

Continuity could be and should be strategic for the business competitive advantage. Besides natural disaster, from blackout to tsunami, businesses face in daily activities critical challenges in IT management for assuring business continuity; for example, business continuity management results must be strategic, because of the infrastructural, organizational, and information systems changes that are required to assure compliance with regulatory norms (see, e.g., the impact of Basel II norms in financial sector), or must have and maintain a time-to-market advantage (disasters can facilitate competitors in a first mover perspective). Nevertheless, business continuity is at present often synonymous with risk management at the IT level, disaster recovery at the hardware level, or in the best case—at the data management level—with data quality management. These perspectives fail to unveil the strategic value of IT business continuity as a framework assuring alignment of strategy, organization, and systems, allowing a competitive advantage in a dynamic competitive environment. Moreover, even when business continuity, under these perspectives, has become one of the most important issues in IT management, there still appears to be some discrepancy as to the formal definitions of what precisely constitutes a disaster, and there are difficulties in assessing the size of claims in the crises and disaster areas. Taking these issues into account, we propose: (a) an analysis of the different facets of the concept of business continuity, and (b) an integrated framework for strategic management of IT business continuity. To these ends, we move from the finance sector—a sector in which the development of information technology (IT) and information systems (IS) have had a key impact upon competitiveness. Indeed, banking industry IT and IS are considered "production," not "support" technologies. The evolution of IT and IS has challenged the traditional ways of conducting business within the finance sector. These changes have largely represented improvements to business processes and efficiency but are not without their flaws, in as much as business disruption can occur due to IT and IS sources. The greater complexity of new IT and IS operating environments requires that organizations continually reassess how best they may face changes and exploit these later for organizational advantage. As such, IT and IS have supported massive changes in the ways in which business is conducted with consumers at the retail level. Innovations in direct banking would have been unthinkable without appropriate IS, and merger and acquisition (M&A) initiatives represent the ideal domain to show what value can lead strategic management of IT business continuity. Taking these issues into account, we point out the relevance of continuity for maintaining customers, and time-to-market in complex and evolutionary competitive environments. Due the relevance of IT to maintain a value-added continuity, our contribution aims to clarify the concept of IT business continuity, providing a framework, exploiting the different facets that it encompasses, and showing the strategic implications to the field of IS&T.

BACKGROUND

The evolution of IT and IS has challenged the traditional ways of conducting business within the finance sector, as a consequence of new business models emerging, for example, from e-business (Müller, Viering, Ahlemann, & Riempp, 2007; Pennings & Harianto, 1992; Ross, Vitale, & Weill, 2001). These changes have largely represented improvements to business processes and efficiency, introducing new challenges and critical issues as much as business interruptions can occur due to IT and IS sources. The greater complexity of new IT and IS operating environments requires that organizations continually reassess how challenges change and exploit those for organizational competitive advantage. In particular, this article seeks to investigate how companies in the financial sector understand and manage their business continuity problems. In fact, business continuity has become one of the most important issues in the banking industry (Lam, 2002), but its relationship with disaster recovery still causes some discrepancy in providing a formal definition on the one hand of what precisely constitutes a disaster, and on the other hand of what is business continuity beyond disaster recovery. Taking into account the different typologies of disaster that can occur in particular in the financial sector (Lam, 2002; Nemzow, 1997), we can define a disaster as an incident that leads to the formal invocation of contingency/continuity plans or any incident that leads to a loss of

revenue; indeed, we can consider a disaster any accidental, natural, or malicious event that threatens or interrupts normal operations or services, causing the failure of the enterprise. In the area of organizational crises and disasters, the degree to which a company has been affected by one or more of such interruptions is the defining factor.

These preliminary definitions are relevant because as estimated by the Business Continuity Institute (2007), most organizations facing a significant crisis, without either a contingency/recovery or a business continuity plan, fail to maintain market position competitively or even to survive a year further. Moreover, state-of-the-art analyses (Bank of Japan, 2003; Barnes, 2001; Elliott & Swartz, 1999; Lam, 2002; Nemzow, 1997; Zambon, Bolzoni, Etalle, & Salvato, 2007) point out that only a small number of organizations have disaster and recovery plans, and among those, few have been renewed to reflect the changing nature of the organization.

In this article we consider in particular our experience in studying practices of the Italian banking industry, where major differences emerge in preparing and implementing strategies that enhance business process security. Comparing them with state-of-the-art literature, we notice two prevalent approaches. On the one hand, there are disaster recovery (DR) strategies that are internally and hardware focused (Lewis, Watson, & Pickren, 2003); on the other hand, there are strategies that treat the issues of IT and IS security within a wider internal-external, hardware-software framework. The latter deals with IS as an integrating business function rather than as a standalone operation. We consider this second type of strategy as part of what in literature (Barnes, 2001; British Standard Institute, 2006; Cerullo & Cerullo, 2004; Nemzow, 1997) is defined as business continuity planning (BCP). Taking these issues into account, we point out the need for a comprehensive IT business continuity approach because of the relevance of the IT for the business continuity of the organizations in the finance sector, encompassing technological, organizational, and strategic facets of the whole system carrying out the business activities. We define the IT business continuity approach (IT-BCA) as a framework of disciplines, processes, and techniques aiming to provide continuous operation for essential business functions under all circumstances. IT-BCA considers business continuity planning as a core element of business continuity initiatives.

More specifically, business continuity planning can be defined as "a collection of procedures and information [that have been] developed, compiled and maintained [and are] ready to use—in the event of an emergency or disaster"(Elliott & Swartz, 1999). BCP has been addressed by different contributions to the literature, such as Allen's (2001) studies on Cert's Octave method, the activities of the Business Continuity Institute (2007) and of the British Standard Institute (2006) in defining certification standards and practice guidelines, the EDS white paper on Business Continuity Management (Decker, 2004), and the activity of financial institutions such as the study carried out by the Bank of Japan (2003). This last study illustrates the process and activities for successful business continuity planning in three steps:

1. formulating a framework for robust project management,
2. identifying assumptions and conditions for business continuity planning, and
3. introducing action plans.

Considering the first step above, banks should (i) develop basic policy and guidelines for business continuity planning (*basic policy*); (ii) develop a study of firm-wide aspects (*firm-wide control section*); and (iii) implement appropriate progress control (*project management procedures*). In the second step, banks should (i) recognize and identify the potential threats, analyze the frequency of potential threats, and identify the specific risk scenarios (*disaster scenarios*); (ii) focus on continuing prioritized critical operations (*critical operations*); and (iii) target times for the resumption of operations (*recovery time objectives*).

Finally, in the third step where actions plans must be introduced, the banks should (i) study specific measures for business continuity planning (*business continuity measures*); (ii) acquire and maintain backup data (*robust backup data*); (iii) determine the managerial resources and infrastructure availability capacity required (*procurement of managerial resources*); (iv) determine strong time constraints, a contact list, and a means of communication on emergency decisions (*decision-making procedures and communication arrangements*); (v) realize practical operational procedures for each department and level (*practical manual*); and (vi) implement a test/training program (*testing and reviewing*).

IT business continuity, indeed, is not only an IT/IS issue, but involves organizational facets, having a strategic impact on banks' competitive advantage and value.

IT BUSINESS CONTINUITY AS A STRATEGIC PATH TO VALUE

In this section we discuss the IT business continuity approach as a path to value for banks, encompassing three fundamental facets that can be viewed in a systemic way: *technology, people,* and *process.*

Technology refers to the recovery of mission-critical data and applications contained in the Disaster Recovery Plan (DRP). It establishes technical and organizational measures in order to face events or incidents with potentially huge impact that could even lead to the unavailability of data centers. The DRP development defines and ensures IT emergency procedures that intervene and protect the data relevant for

company activities and services (Wang, Yin, Yuan, & Zhou, 2005). DRP is usually considered the only part of the BCP in banking business continuity initiatives.

Further, *people* refers to the recovery of the employees and physical workspace. In particular, BCP teams should be drawn from a variety of company departments, including those from personnel, marketing, and internal consultants. The managers of these teams should possess "general" skill and be partially drawn from business areas besides the ones coming from the IT departments. Nowadays, this is essential to emphasize the role of human assets and value rather than the traditional focus on hardware and software resources that in most cases are probably protected by backup systems. Finally, the term *process* refers to the development of a strategy for the deployment, testing, and maintenance of the plan. All BCPs should be regularly updated and modified in order to take into consideration the latest types of threats, both at physical and technological levels.

Whereas a simple disaster recovery (DR) approach aims to rescue those facilities and services that can be recovered, a BC approach should treat IT and IS security with a wider internal-external, hardware-software framework where all processes are neither in-house nor subcontracted out, but are a mix of the two, integrating business functions' standalone operations. Thus, the BC approach is a dual approach providing a unified perspective on management and technology function. The BC approach as a global approach also considers all existing relationships in total value chain perspective for business, giving value to clients and suppliers and to protect business both in-house and out. The BC approach incorporates the disaster recovery approach but rejects its exclusive focus upon facilities. The BC approach defines the process as business wide, enabling competitive and/or organizational advantages.

The starting point for planning processes that an organization will use as its BCP must include an assessment of the likely impact of different types of 'incidents' on the business. As mentioned above, the adoption of new technologies continues to become more and more integral and critical for the financial activities. In addition, in order to assess the likely impact on the entire organization, banks must consider the effects on their different business areas. To these ends, we introduce a *vulnerability and business impact matrix,* which is a tool that can be used to summarize the inter-linkages between the various information system services, their vulnerability, and the impact on business processes. For the strategic focus of BCP, an understanding of the relationships between value-creating activities is a key determinant of the effectiveness of any process, both at IT and organizational levels. This way, we can define the correct business continuity perimeter by trying to extract the maximum value from the BCP within a context of bounded and limited resources. Indeed, the BCP teams in the organizations focus on how resources were utilized and how they were added to value-creation, rather than merely being "support activities," consuming financial resources with no productivity outcome. Furthermore, the growing relevance of customer-oriented technologies demands that those managing the BCP process are aware of the need to expand the role of unforeseen events and contingencies, not merely looking inward but actually looking out of the IT and back-office operations. Such a dual focus improves the linkages between customer and client which create competitive advantage. Indeed, in cases where client-oriented business fundamentally depends upon information exchange (e.g., the online equity brokerage services provision), it might be argued that there is a *virtual value chain*, which the BCP team protects, providing the 'market-space' for value creation to take place. Finally, another benefit is that vulnerability and business impact can improve the prioritization of particular key areas.

In fact, today's approach to BCP (Cerullo & Cerullo, 2004; Lam, 2002; Nemzow, 1997; Zambon et al., 2007) is focused on process management and business-driven paradigms; nevertheless, when considering large institutions with systemic impact—not only on their own but on clients businesses as well—two key objectives need to be considered when facing a disrupting event. These have been named *recovery point objective* (RPO) and *recovery time objective* (RTO). The former deals with how far in the past you have to go to resume a consistent situation; the latter considers how long it takes to resume a standard or regular situation. The definitions of RPO and RTO can change according to data center organization and how high a level a company wants its own security and continuity to be (Disaster Recovery Journal & DRI, 2007). For instance, a dual site recovery system organization must consider and evaluate three points of view, namely *application's availability, business continuity process*, and *data perspective.*

Focusing on RTO, data are first impacted before the crisis event due to the closest *consistent point* from which to restart. The crisis opening or declaration occurs after the crisis event. We now define a relevant RTO's event indicators:

- the *computing environment restored point* considers the length of time the computing environment needs in order to be restored (i.e., when servers, network, etc. are once again available);
- the *mission critical application restarted point* indicates the critical applications (in rank order) that are working once again;
- the *applications and data restored point* is the point from which all applications and data are restored; but (and it is a big but)
- the *previous environment restored point* is the true end point when the previous environment is fully restored (this later pointing out the all business continuity solutions are properly working).

Taking these issues into account, natural risks are also increasing in scope and frequency, thus extending actual geographical recovery distance, and forcing businesses and institutions to consider a new technological approach. This later must be introduced to undertake synchronous-asynchronous data replication, focusing on intervals and quality. Therefore, more complex analyses on RPO and RTO are required.

However, from a business point of view, when faced with an imminent and unforeseen disaster, the most important issue is how to reduce the restore or restart time, trying to minimize this window to few seconds or less. New technologies, such as the serial ATA (SATA) and the massive arrays inexpensive disk (MAID), seem to make some progress in reducing the time-related problem in disaster recovery and business continuity initiatives.

The *vulnerability and business impact analysis matrix* discussed above considers these issues and treats each selected business in accordance with the value chain and value system. In fact, in addition to assessing the likely disaster impact upon IT departments, organizations should consider disaster impacts over all company departments and their likely effects upon customers. Organizations should avoid the so-called 'Soccer Star Syndrome' (Elliott & Swartz, 1999); drawing an analogy with the football industry, this syndrome recognizes that management attention is often focused on the playing field rather than on the unglamorous, but very necessary, locker room and stadium management support activities. Defenders and goalkeepers, let alone the stadium manager, do not get paid at the same level as the "star" player, yet their functions are just as vital to achieving the overall objectives of the football team. The value chain provides an opportunity to examine the connection between the whole linkages that deliver customer value. The evolution of crisis preparations from the IT-focused disaster recovery solutions towards the business continuity approach reflects a growing understanding that business continuity depends upon the maintenance of all elements that provide organizational efficiency-effectiveness and customer value, whether directly or indirectly.

A final key characteristic of the proposed business continuity approach concerns the primary role of prevention. In the literature, a number of authors have argued that the potential for crises is a "normal" issue for organizations (Greiner, 1989; Pauchant & Mitroff, 1998). Nowadays, crisis avoidance requires a strategic approach and a good understanding of both the organization's operating processes and systems, and the environment where the business operates. To these ends, in the IT business continuity approach, an organization must develop a business continuity planning culture in order to remove the barriers to the development of crisis prevention strategies. In particular, these organizations should recognize that incidents and disasters are more and more triggered by external factors and not only by technical causes, but their effects are largely determined by internal factors that are within the control of their organizations. To face these issues from a BCP-enabled business continuity approach, a *cluster of crises* must be identified and categorized along the axis of internal-external and human/social-technical/economic causes and effects. By adopting a strategic approach, decisions could be made about the extent of exposure in particular product markets or geographical sites. An ongoing change management program could contribute to real commitment from middle managers. Here the focus is not on change management as per se topic, but as a way to the commitment from middle managers who, from our first investigation of the Italian banking sector, emerged as key determinants of the success of the IT-business continuity approach.

Nevertheless, BCP is the core of a business continuity approach, and its success requires the commitment of middle managers. Hence managers need to avoid considering BCP as a costly, administrative inconvenience that diverts time away from primary business activities. As a consequence, strategic business units should own BCP plans, and CEO involvement is a key factor supporting the BCP process. Thus, the recognition that responsibility for the process rests with business managers must be improved and reinforced through a formal appraisal and other reward systems; these initiatives must be integrated by others leveraging peer pressure as relevant to assume responsibility and affect a more receptive culture. Finally, BCP teams need to regard BCP as a process rather than as a specific end point, exploiting business continuity not as a "one facet" event-based recovery activity, but as an organizational change factor, improving IT/IS-oriented innovation and value for the whole business through a risk-aware management culture.

FUTURE TRENDS

Future research should test the IT business continuity approach on strategic activities, such as cross-border M&A initiatives. These analyses are relevant to exploit the whole meaning of the business continuity concept, integrating the still prevailing one based on technological issues in disaster recovery initiatives. Furthermore, future works should investigate alternative measures involving organizational and strategic issues, together with the technological ones. These measures must be introduced to enhance the strategic value of business continuity instruments, such as the above discussed *vulnerability and business impact analysis matrix.*

CONCLUSION

This article argues that to adopt an IT business continuity approach involves a strategic perspective on the different

facets leading to a sustainable competitive advantage. Indeed, IT business continuity is not only a technological issue; it is a proxy for organizational change in terms of culture, structure, and communications. The IT business continuity approach is more and more a driver to generate competitive advantage from efficient and effective flexible information systems, allowing the organization to attract and maintain customers by assuring trust on the capabilities of the organization in facing disaster and a better perceived quality of services in terms of timeliness and currency. Nowadays, referring to organizational change and culture, the IT business continuity approach must be considered a business-wide approach, not an IT-focused one. Supportive measures must be introduced to encourage managers to adhere to the IT business continuity strategic vision. In particular, management as a whole should be committed to the IT business continuity ongoing process. Continuity requires changes of key assumptions and values within the organizational structure and culture, having implications for the role that the IT business continuity approach must play within the strategic management processes of the organization, as well as within the levels of strategic risk that an organization may wish to undertake to secure a sustainable competitive advantage.

REFERENCES

Allen, J.H. (2001). *CERT® guide to system and network security practices.* Reading, MA: Addison Wesley Professional.

Bank of Japan. (2003). *Business continuity planning at financial institutions.* Retrieved July 2003 from http://www.boj.or.jp/en/type/release/zuiji/kako03/fsk0307a.htm

Barnes, J.C. (2001). *A guide to business continuity planning B2.* New York: John Wiley & Sons.

British Standard Institute. (2006). *Business continuity management—part1: Code of practice*. Technical Report 25999-1, British Standard Institute, UK.

Business Continuity Institute. (2007). *About the BCI.* Retrieved from http://www.thebci.org/about.htm

Cerullo, V., & Cerullo, J. (2004). Business continuity planning: A comprehensive approach. *Information Systems Management Journal*, (Summer).

Decker, A. (2004). *Business continuity management: A model for survival.* EDS White Paper.

Disaster Recovery Journal and DRI. (2007). *Business continuity acronym glossary.* Retrieved from http://www.drj.com/glossary/drjglossary.html

Elliott, D., & Swartz, E. (1999). Just waiting for the next big bang: Business continuity planning in the UK finance sector. *Journal of Applied Management Studies, 8*(1), 45-60.

Greiner, L. (1989). Evolution and revolution as organisations grow In D. Asch & C. Bowman (Eds.), *Readings in strategic management* (pp. 373-387). London: Macmillan.

Lam, W. (2002). Ensuring business continuity. *IT Professional, 4*(3), 19-25.

Lewis, W., Watson, R.T., & Pickren, A. (2003). Virtual extension: An empirical assessment of IT disaster risk. *Communications of the ACM, 46*(9), 201-206.

Müller, B., Viering, G., Ahlemann, F., & Riempp, G. (2007, June 7-9). Towards understanding the sources of the economic potential of service-oriented architecture: Findings from the automotive and banking industry. *Proceedings of the 15th European Conference on Information Systems* (ECIS2007), St. Gallen, Switzerland.

Nemzow, M. (1997). Business continuity planning. *International Journal of Network Management, 7,* 127-136.

Pauchant, T.C., & Mitroff, I. (1998). Crisis prone versus crisis avoiding organisations: Is your company's culture its own worst enemy in creating crises? *Industrial Crisis Quarterly, 2*(4), 53-63.

Pennings, J.M., & Harianto, F. (1992). The diffusion of technological innovation in the commercial banking industry. *Strategic Management Journal, 13,* 29-46.

Ross, J., Vitale, M., & Weill, P. (2001). *From place to space: Migrating to profitable electronic commerce business models.* Cambridge, MA: MIT Sloan School of Management.

Wang, K., Yin, Z., Yuan, F., & Zhou, L. (2005). A mathematical approach to disaster recovery planning. *Proceedings of the 1st International Conference on Semantics, Knowledge and Grid* (SKG'05).

Zambon, E., Bolzoni, D., Etalle, S., & Salvato, M. (2007, July 1-5). A model supporting business continuity auditing & planning in information systems. *Proceedings of the 2nd International Conference on Internet Monitoring and Protection* (ICIMP), San Jose, CA.

KEY TERMS

Business Continuity Planning (BCP): "A collection of procedures and information [that have been] developed, compiled and maintained [and are] ready to use in the event of an emergency or disaster" (Elliott & Swartz, 1999).

Crisis: A critical event, which, if not handled in an appropriate manner, may dramatically impact an organization's profitability, reputation, or ability to operate. Or, an occurrence and/or perception that threatens the operations, staff, shareholder value, stakeholders, brand, reputation, trust, and/or strategic/business goals of an organization (Disaster Recovery Journal & DRI, 2007).

Disaster: An incident that leads to the formal invocation of contingency/continuity plans, or any incident that leads to a loss of revenue; indeed, a disaster is any accidental, natural, or malicious event that threatens or interrupts normal operations or services, causing the failure of the enterprise (Disaster Recovery Journal & DRI, 2007).

Disaster Recovery: The ability of an organization to respond to a disaster or an interruption in services by implementing a disaster recovery plan to stabilize and restore the organization's critical functions (Disaster Recovery Journal & DRI, 2007).

Disaster Recovery Plan (DRP): A plan that establishes technical and organizational measures in order to face events or incidents with potentially huge impact that could even lead to the unavailability of data centers. The DRP development defines and ensures IT emergency procedures that intervene and protect the data relevant for the company activities and services. DRP is usually considered as the only part of the BCP in banking business continuity initiatives.

IT Business Continuity: The ability of an organization to provide service and support for its customers, and maintain its viability before, during, and after a business continuity event (Disaster Recovery Journal & DRI, 2007) through the leverage of IT resources.

IT-Business Continuity Approach: A framework of disciplines, processes, and techniques aiming to provide continuous operation for "essential business functions" under all circumstances.

Recovery Point Objective (RPO): The maximum amount of data loss an organization can sustain during an event (Disaster Recovery Journal & DRI, 2007). RPO deals with how far in the past you have to go to resume a consistent situation.

Recovery Time Objective (RTO): The period of time within which systems, applications, or functions must be recovered after an outage (e.g., one business day). RTOs are often used as the basis for the development of recovery strategies, and as a determinant as to whether or not to implement the recovery strategies during a disaster situation (Disaster Recovery Journal & DRI, 2007). RTO considers how long it takes to resume a standard or regular situation.

Category: Business Information Systems

Information Technology in Franchising

Ye-Sho Chen
Louisiana State University, USA

Grace Hua
Louisiana State University, USA

Bob Justis
Louisiana State University, USA

INTRODUCTION

Franchising is "a business opportunity by which the owner (producer or distributor) of a service or a trademarked product grants exclusive rights to an individual for the local distribution and/or sale of the service or product, and in return receives a payment or royalty and conformance to quality standards. The individual or business granting the business rights is called the *franchisor*, and the individual or business granted the right to operate in accordance with the chosen method to produce or sell the product or service is called the *franchisee*." (Justis & Judd, 2002)

Information technology (IT) has been widely used in today's businesses. In his best seller, *Business @ the Speed of Thought*, Bill Gates (1999) wrote: "Information Technology and business are becoming inextricably interwoven. I don't think anybody can talk meaningfully about one without talking about the other." Thus, to see how IT is used in franchising, one needs to know how franchising really works. The objective of this paper is to propose an attention-based IT infrastructure that can cultivate the relationship building between the franchisors and their franchisees which will ultimately lead to the success of the franchise organizations.

BACKGROUND

In addition to the popular growth strategy for many businesses, franchising has emerged over the years as a pathway to wealth creation for entrepreneurs (Justis & Vincent, 2001). This article first discusses the operations at both the franchisor headquarters and the franchisee outlets. Second, it reviews the franchisor/franchisee relationship and points out the essential indicators needed to pertain and flourish the good relationship. Third, it shows the inevitability of collaborative learning and innovation, which leads us to the discussion of the working knowledge development among the franchisor and the fellow franchisees. Fourth, we discuss that the proposed attention-based IT infrastructure will enable the knowledge sharing and dissemination between the franchisor and the franchisee; and suggest outsourcing the initial architectural stages of the IT infrastructure to trusted applications service providers.

Understanding the Franchisor

In this section we examine the operational activities at the franchisor headquarters. Figure 1 illustrates the interactions of the franchisor with all four of its entities: business units, prospective franchisees, suppliers, and government; as well as performing relevant activities (represented by rectangles): marketing its products and services, assisting in creating distinguished brand names indispensable in attracting new customers, selling to the franchisees, and handling the diversified financing quandaries.

The franchisor headquarters is required to provide both initial and ongoing support/service to all business units. Business units here include company units, all of the start-up, established and mastered franchisees, and the co-branded units. Among the five different types of business units, the franchisor needs to have intense concentration on supporting the start-up franchisees, since a good start is as efficient as the half way completion of any task. On the other hand, established and mastered franchisees are the ones in need of appealing incentives (e.g., having cobranded units) in order to encourage growth and expansion. Company units are typically used as role models for the franchisees.

To expand the business, the franchisor ought to select and contact the prospective partners (franchisees). The partner selection process is crucial to the success of franchising and requires exceptional attention. Prospective franchisees can be contacted through: (1) leads from marketing channels; (2) referrals, such as satisfied customers; (3) consumers who feel affection to the product/service and would like to be in possession of the business; (4) community and media relationships; (5) public services, like recruiting veterans; and (6) international contacts generated from master franchisees.

Franchise suppliers can be anywhere from products and goods distributors up to business service providers, such as real estate agents, human resources providers, uniform

vendors, marketing and advertising agents, trade show and exposition organizers, accountants, information systems vendors, insurance providers, attorneys, translators, and many others.

Franchisors also need to comply with regulations that govern the sales of the franchises and business transactions in the places where the business located. The overall legal landscape of franchising is complex which includes: (1) federal, state, and international taxes; (2) local, regional, and global laws; (3) insurances, such as workers compensation; (4) possibilities of litigations from government, customers, and franchisees; and (5) supports for international expansion.

As is seen in Figure 1, the franchisor interacts with various business entities and conducts various deeds to fulfil his/her obligations. Moreover, the franchisor goes through a learning process, composed of five stages (Justis & Judd, 2002): (1) beginner—learning how to do it; (2) novice—practicing doing it; (3) advanced—doing it; (4) master—teaching others to do it; and (5) Professional—becoming the best that you can be. Once attaining the advanced stages of development, most preceding struggles have been overcome. However, further convoluted and challenging enquiries will arise as the franchise continues the expansion. This is especially factual once the system reaches the professional stage, where various unpredicted and intricate problems could arise. Bud Hadfield (1995), the founder of Kwik Kopy franchise and the International Center of Entrepreneurial Development, explained it the finest: "The more the company grows, the more it will be tested."

To capture the learning process in Figure 1, a counter-clockwise round arrow is used in each of the five categories. It depicts the increasing intensity of learning in every area of the sub-activities (represented by rectangles) as the franchise struggles to survive and thrive. For example, as the system expands, the real estate sub-activity will become much more complex, since the quandary of territory encroachment becomes more significant and harder to manage.

Understanding the Franchisee

As illustrated in Figure 2, the franchisee sells to customers, perform marketing and advertising activities, handle financial/accounting issues, and manage sales people. Akin to the franchisor, the franchisee outlets (shortened form, franchisee) work together with four entities: customers, franchisor headquarters, suppliers and government. The franchisee customers are divided into five categories and consist of potential, infrequent, frequent, online, and co-branded. Supports from the franchisor headquarter may include demonstrations from field representatives, training and continued education from the management groups, in addition to discussion forums and distance learning.

Suppliers of the franchisee are similar to those of the franchisor. They include both products and goods distributors, business service providers such as real estate agents, human resources providers, uniform vendors, marketing and advertising agents, trade shows and exposition organizers, accountants, information systems vendors, insurance providers, attorneys, translators, and others. The franchisee is regulated by the franchising laws at local, state, and federal level. The regulatory framework contains of: (1) federal, state, and international taxes; (2) local, regional, and global laws; (3)

Figure 1. Understanding how the franchisor works

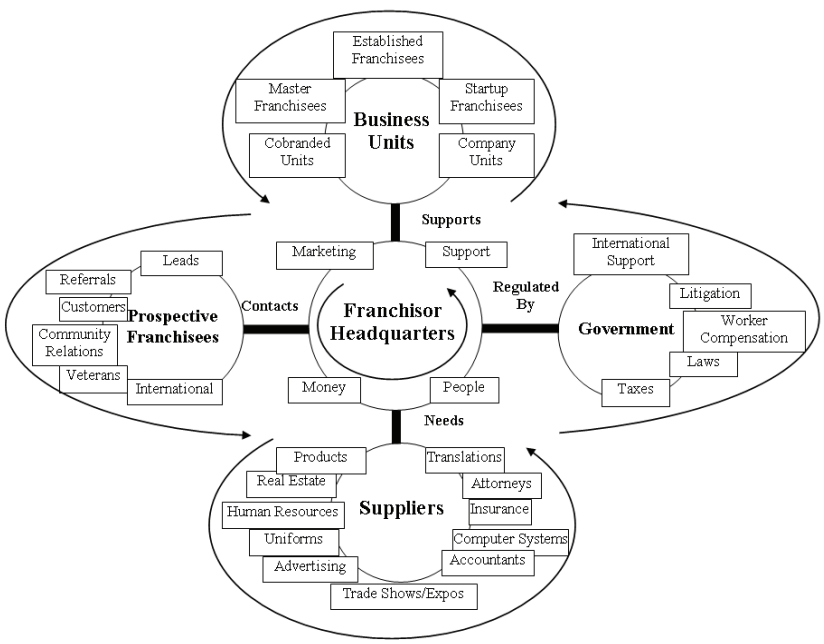
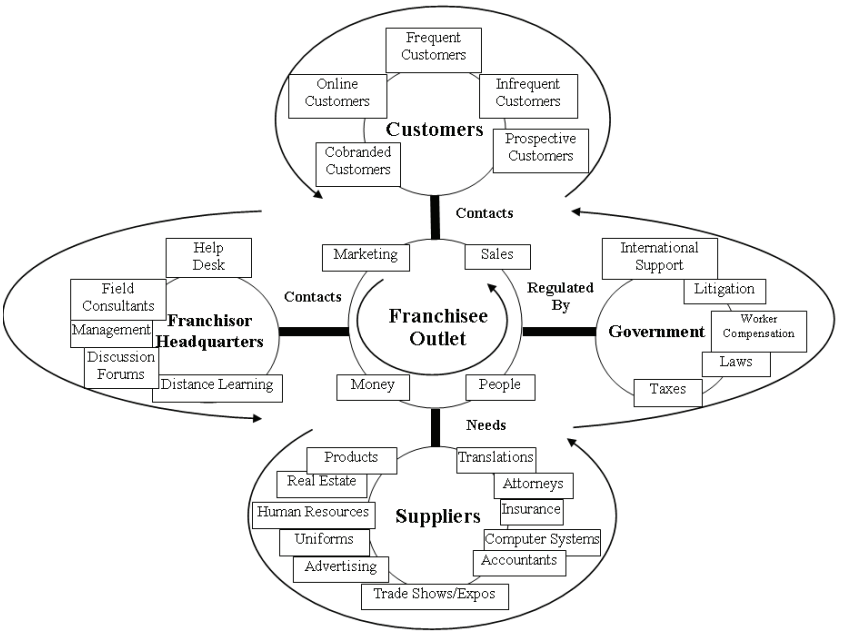


Figure 1. Understanding how the franchisee works



insurances such as workers compensation; (4) possibilities of litigations from government, customers, and franchisees; and (5) support for international expansions.

The franchisee also goes through the above mentioned five stages of learning (i.e., beginner, novice, advanced, master, and professional) represented by a counter-clockwise round arrow used in each of the five categories in Figure 2. Once again, the arrow depicts the increasing intensity of learning as the franchisee continues growing the business and many unforseen problems/issues may rise up to challenge the practices.

Understanding the Franchisor/ Franchisee Relationship

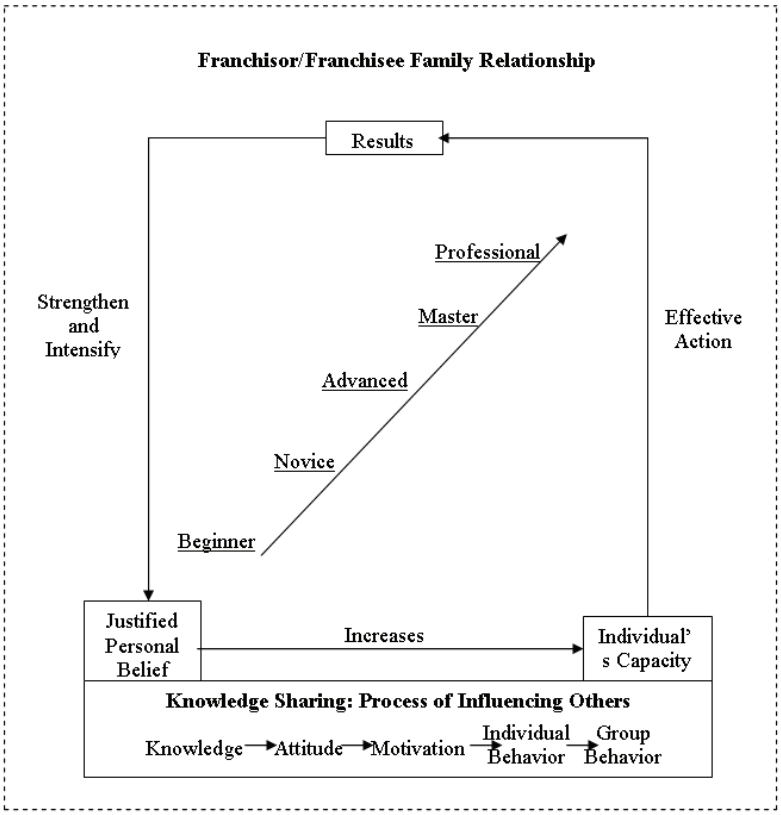
Developing a high-quality "family" relationship between the franchisor and the franchisee is believed to be the most significant factor for the success of a franchise (Justis & Judd, 2002). To understand how the relationship is developed, one needs to know the franchisee life cycle (Schreuder, Krige, & Parker, 2000):

- **The Courting Phase:** Both the franchisee and the franchisor are eager with the relationship. This typically corresponds to the beginner stage of the franchisee.
- **The "We" Phase:** The relationship starts to deteriorate, but the franchisee still values the relationship. This typically corresponds to the novice stage of the franchisee.
- **The "Me" Phase:** The franchisee starts to question the reasons for payments to their franchisors. They may start to think that the success so far is purely of his/her own work. This typically corresponds to the advanced stage of the franchisee.
- **The Rebel Phase:** The franchisee starts to challenge the restrictions being placed upon. This typically corresponds to the master stage of the franchisee. The rebel ones tend to be those who know the system very well and are capable of influencing others to follow them.
- **The Renewal Phase:** The franchisee realizes that the "win-win" solution is to continue teaming up with the franchisor to develop the system. This typically corresponds to the professional stage of the franchisee.

Thus, the major challenge for the franchisor is to turn a rebel franchisee into the renewal one. A viable solution for the franchisor is to provide a learning and innovative environment where the franchisee can continue contributing to the growth of the system. Successful collaborative learning and innovations will provide a strong incentive for the professional franchisees to continue their renewal relationship with the firm, which in turn will have positive impact on maintaining the good relationship with other franchisees and recruiting new ones. On the other hand, constant failures in this stage will intensify the rebel franchisees to dessert the firm, which in some cases lead to the demise of the franchise.

Lying behind the successful collaboration is the working knowledge of the franchise firm. Knowledge is defended

Figure 3. Understanding the franchisor/franchisee family relationship



as "a justified personal belief that increases an individual's capacity to take effective action" (Alavi & Leidner, 1999). Knowledge becomes "working" when the action produces results. When knowledge produces results, the personal confidence strengthens, intensifies, and justifies. The more superior the individual's capacity becomes, the better results are attained. This spiral-up cycle of working knowledge development is very important in the context of franchising. Figure 3 shows that the development process is incrementally developed through the five stages of the spiral-up cycle defined earlier: beginner, novice, advanced, master, and professional.

The foundation of the learning cycle is the capability of sharing and coaching the working knowledge throughout the franchise system. The process of influencing others for knowledge dissemination consists of five steps (Justis & Vincent, 2001): (1) knowledge, the proven abilities to solve problems in the franchise environment; (2) attitude, constructive ways of presenting and sharing the working knowledge; (3) motivation, incentives for learning or teaching the working knowledge; (4) individual behavior, the strengths of the participants to learn and enhance the working knowledge; and (5) group behavior, collaborative ways to create, dissimilate, and manage the hard-earned working knowledge.

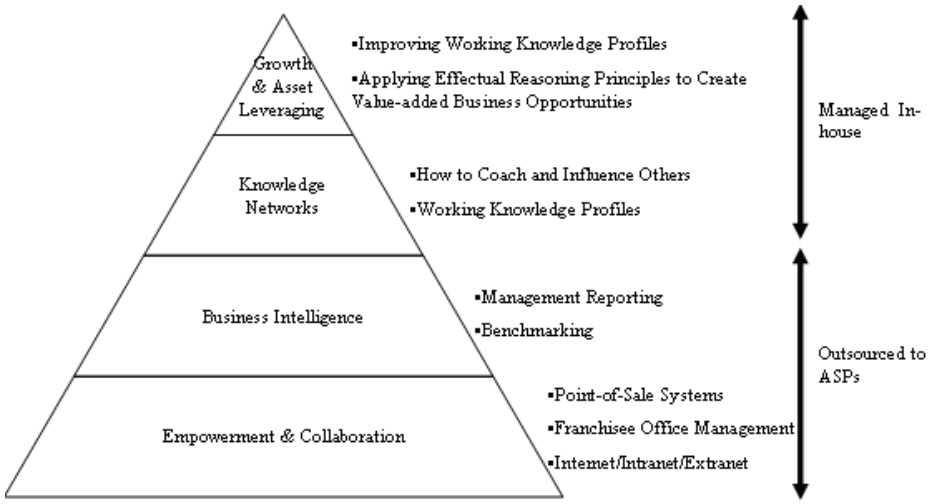
The franchisor/franchisee "family" relationship building in Figure 3 is also surrounded with dashed line, denoting that the relationship is enlarged and expanded without limits as the franchisee incrementally learns the working knowledge through the influencing of the franchisor and the fellow franchisees. By going through the processes of learning and influencing, both the franchisor and the franchisee progressively gain the working knowledge and mange the franchisee life cycle effectively.

FUTURE TRENDS

An Attention-Based IT Infrastructure in Franchising

In an information-rich world, Herbert Simon (Nobel laureate in Economics in 1978) wrote: "a wealth of information creates a poverty of attention" (Simon, 1971). As such, the "proper aim of a management information system is not to bring the manager all the information he needs, but to recognize the manager's environment of information so as to reduce the amount of time he must devote to receiving it" (Simon,

Figure 4. An attention-based IT infrastrucutre in franchising



1971). From the discussions in the last section, it is obvious that an attention-based IT infrastructure in franchising shall be one devoted to enabling the building of a good "family" relationship between the franchisor and the franchisee. In Figure 4, we propose such IT architecture (Chen, Chen, & Wu, 2005, 2006) for franchise organizations to manage the immense information produced by the firms.

The architecture, adapted from Gates' Digital Nervous Systems (1999), consists of the following four layers:

- **Empowerment and Collaboration:** Point-of-sale and office management systems are used to empower the franchisees (Chen, Justis, & Chong, 2002). Internets, Extranets, and Intranets networking the franchisor, the franchisees, customers and suppliers are deployed to improve collaboration (Chen, Chong, & Justis, 2002).
- **Business Intelligence:** Data warehousing and data mining techniques are used to convert volumes of data into reports and benchmarks for management to glean business intelligence (Chen, Justis, & Watson, 2002; Chen, Justis, & Chong, 2004; Chen, Zhang, & Justis, 2005).
- **Knowledge Networks:** Intranet-based systems consisting of the skills of coaching/influencing others (Chen, Seidman, & Justis, 2005) and working knowledge profiles (Chen, Hammerstein, & Justis, 2002) are implemented for knowledge sharing and learning within the franchise business. A distance-learning curriculum (Chen, Chong, & Justis, 2000) of working knowledge modules can also be deployed.
- **Growth and Asset Leveraging:** Value networks (Chen, Justis, & Wu, 2006) are developed with the goal of improving working knowledge profiles and applying effectual reasoning principles (Chen, Yuan, & Dai, 2004) to create value-added business opportunities (Chen, Justis, & Yang, 2004).

To demonstrate how the attention-based IT infrastructure is related to the activities of the franchisor, the franchisee (shown in Figures 1 and 2), and the "family" relationship building (shown in Figure 3), we show in Figure 5 the character of the business process, the information flow in the franchise business, and clarify the rationale behind the necessity of the attention-based IT infrastructure. The foundation of the architecture, adapted from Inmon (1996), consists of four levels: (1) data collected from the empowerment and collaboration activities of the franchisor (Figure 1) and the franchisee (Figure 2); e-business strategy shall be one empowering the franchisor and the franchisees to do their activities (Chen, Chong, & Justis, 2002); (2) reconciled data in the franchise data warehouse (Chen, Justis, & Watson, 2002); (3) derived data residing in data marts based on various franchisee-centered segmentations such as franchise development and support; and (4) business intelligence reports generated from analytical data analysis; for example, management reporting and benchmarking could be produced by online analytical processing (OLAP) periodically which periodically may lead to more proactive analysis of top/low performer attributes using the data mining techniques (Chen, Justis, & Chong, 2004).

The business intelligence reports will help the franchise system to identify key top performers and their success attributes. With their involvement, working knowledge profiles can be developed and dissimilated throughout the knowledge networks of the business using coaching/influencing techniques (Chen, Hammerstein, & Justis, 2002). The top of Figure 5 depicts such a knowledge sharing and

Figure 5. Implementing the attention-based IT infrastructure in franchising

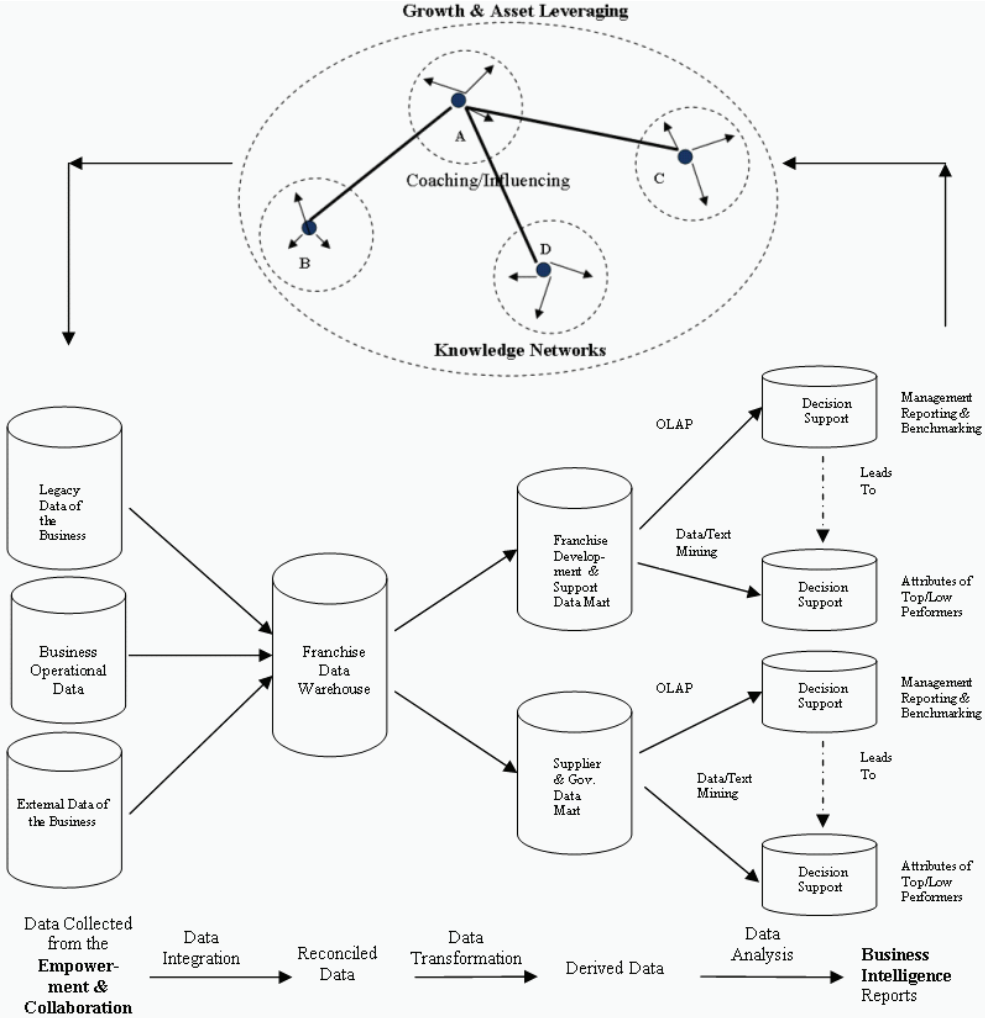


Figure 5. Implementing the Attention-based IT Infrastructure in Franchising

dissemination idea described in Figure 3 in detail. There are four franchisees (A - D) in the figure. Each franchisee (a dot) has his/her personal knowledge network (arrows pointing out of the dot). The knowledge network may include the customers' likes and dislikes, the kind of employees to hire, the competitors' and suppliers' pricing strategies, the social needs in the local community. Each franchisee is surrounded with a circle with dashed lines, meaning there is no limit to the personal knowledge network. Supposing franchisee A is the top performer and is charged and rewarded with coaching/influencing other franchisees to survive and thrive. Thus, clusters (connected dots) of knowledge network are formed and surrounded with a circle with dashed lines, meaning there is no limit for improving and leveraging the assets of the business (Chen, Justis, & Yang, 2004).

Using the attention-based IT architecture as a guide, we recommend franchise organizations to outsource IT in the first two layers to application service providers (ASPs) (Chen, Ford, Justis, & Chong, 2001). The concept of subscribing IT services through ASPs has special appeal in the franchising industry because an ASP can duplicate success for former similar franchises quickly and economically.

CONCLUSION

Franchising has been popular as a growth strategy for small businesses; it is even more so in today's global and e-commerce world (Chen & Wu, 2006). Although IT is quite important in franchising, IT researchers have largely ignored

this arena. One major reason is that few IT researchers are vested in the knowledge of how franchising functions, and without this intimate knowledge it is difficult to implement effective IT systems. Based on years of academic and consulting experiences in franchising and IT, an effective framework of IT in franchising is proposed. At the heart of the framework is to manage working knowledge and use IT to build up the "family" relationship between the franchisor and the franchisee.

REFERENCES

Alavi, M., & Leidner, D. E. (1999, February). Knowledge management systems: Issues, challenges, and benefits. *Communications of the Association for Information Systems,* (Vol. 1, Article 7).

Chen, Y., Chen, G., & Wu, S. (2005). Issues and opportunities in e-business research: A Simonian perspective. *International Journal of E-Business Research, 1*(1), 37-53.

Chen, Y., Chen, G., & Wu, S. (2006). A Simonian approach to e-business research: A study in netchising. In *Advanced topics in e-business research: E-business innovation and process management* (Vol. 1, pp. 133-161). Hershey, PA: Idea Group Publishing.

Chen, Y., Chong, P., & Justis, R.T. (2000, February 19-20). Franchising Knowledge Repository: A Structure for Learning Organizations. In *Proceedings of the 14th Annual International Society of Franchising Conference,* San Diego, California.

Chen, Y., Chong, P. P., & Justis, R. T. (2002, February 8-10). E-business strategy in franchising: A customer-service-life-cycle approach. In *Proceedings of the 16th Annual International Society of Franchising Conference,* Orlando, Florida.

Chen, Y., Ford, C., Justis, R. T., & Chong, P. (2001, February 24-25). Application service providers (ASP) in franchising: Opportunities and issues. In *Proceedings of the 15th Annual International Society of Franchising Conference,* Las Vegas, Nevada.

Chen, Y., Hammerstein, S., & Justis, R. T. (2002, April 5-6). Knowledge, learning, and capabilities in franchise organizations. In *Proceedings of the 3rd European Conference on Organizational Knowledge, Learning, and Capabilities,* Athens, Greece.

Chen, Y., Justis, R. T., & Chong, P. P. (2002). Franchising and information technology: A framework. In S. Burgess (Ed.), *Managing Information Technology in Small Business: Challenges and Solutions* (pp. 118-139). Hershey, PA: Idea Group Publishing.

Chen, Y., Justis, R. T., & Chong, P. P. (2004). Data mining in franchise organizations. In H. R. Nemati & C. D. Barko (Eds.), *Organizational data mining: Leveraging enterprise data resources for optimal performance* (pp. 217-229). Hershey, PA: Idea Group Publishing.

Chen, Y., Justis, R., & Watson, E. (2000). Web-enabled Data Warehousing. In M. Shaw, R. Blanning, T. Strader, & A. Whinston (Eds.), *Handbook of electronic commerce* (pp. 501-520). New York: Springer-Verlag.

Chen, Y., Justis, R., & Wu, S. (2006, February 24-26). Value Networks in Franchise Organizations: A Study in the Senior Care Industry. In *Proceedings of the 20th Annual International Society of Franchising Conference,* Palm Springs, California.

Chen, Y., Justis, R. T., & Yang, H. L. (2004, March 5-7). Global e-business, international franchising, and theory of netchising: A research alliance of east and west. In *Proceedings of the 18th Annual International Society of Franchising Conference,* Las Vegas, Nevada.

Chen, Y., Seidman, W., & Justis, R. (2005, May 20-22). Strategy and docility in franchise organizations. In *Proceedings of the 19th Annual International Society of Franchising Conference,* London.

Chen, Y., & Wu, S. (2006). E-business research in franchising (invited editorial preface). *International Journal of E-Business Research, 2*(4), i-ix.

Chen, Y., Yuan, W., & Dai, W. (2004, December 12-15). *Strategy and nearly decomposable systems: A study in franchise organizations.* International Symposium on "IT/IS Issues in Asia-Pacific Region, Co-sponsored by ICIS-2004, Washington, DC.

Chen, Y., Zhang, B., & Justis, R. T. (2005). Data mining in franchise organizations. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 714-722). Hershey, PA: Idea Group Reference.

Gates, W. (1999). *Business @ the speed of thought.* New York: Warner Books.

Hadfield, B. (1995). *Wealth within reach.* Houston, TX: Cypress Publishing.

Inmon, W. H. (1996). *Building the data warehouse.* New York: John Wiley & Sons.

Justis, R. T., & Judd, R. J. (2003). *Franchising* (3rd ed.). Houston, TX: DAME Publishing.

Justis, R. T. & Vincent, W. S. (2001). *Achieving wealth through franchising.* Holbrook, MA: Adams Media Corporation.

Schreuder, A.N., Krige, L., & Parker, E. (2000, February 19-20). The franchisee lifecycle concept—A new paradigm in managing the franchisee-franchisor relationship. In *Proceedings of the 14th annual International Society of Franchising Conference*, San Diego, California.

Simon, H. A. (1971). Designing organizations for an information rich world. In M. Greeberger (Ed.), *Computers, communications, and the public interest* (pp. 38-52). Baltimore: The Johns Hopkins Press.

KEY TERMS

Attention-Based IT Infrastructure: An IT infrastructure which is able to sort through volumes of data and produce right information at the right time for the right persons to consume.

Franchisee: The individual or business who receives the business rights and pay the royalties for using the rights.

Franchisee Life Cycle: The stages a franchisee goes through in the franchise system: Courting, "We", "Me", Rebel, Renewal.

Franchising: A business opportunity based on granting the business rights and collecting royalties in return.

Franchisor: The individual or business who grants the business rights.

Franchisor/Franchisee Learning Process: The stages of learning, including Beginner, Novice, Advanced, Master, and Professional.

Franchisor/Franchisee Relationship Management: The vital factor for the success of a franchise, including: Knowledge, Attitude, Motivation, Individual Behavior, and Group Behavior.

Information Technology in Survey Research

Jernej Berzelak
University of Ljubljana, Slovenia

Vasja Vehovar
University of Ljubljana, Slovenia

INTRODUCTION

Data collection based on standardized questionnaires represents one of the central tools in many research areas. Early surveys date back to the 18th century (de Leeuw, 2005), while a major breakthrough came in the 1930s with the application of probability samples. By using surveys, today governments monitor conditions in the country, social scientists obtain data on social phenomena and managers direct their business by studying the characteristics of their target customers.

The importance of survey research stimulates ongoing efforts to achieve higher data quality and optimized costs. Early on researchers recognized the potential of technological advances for the achievement of these goals. In the early 1970s telephone surveys started replacing expensive face-to-face interviews. Computer technology developments soon enabled computer-assisted telephone interviewing ("CATI"). The 1980s brought new approaches based on personal computers. Interviewers started to use laptops and respondents sometimes completed questionnaires on their own computers. Another revolution occurred with the Internet in the subsequent decade. The pervasive availability of Internet access, and the growing number of Internet-supported devices, coupled with the advance of interactive Web technologies (like Ajax) are facilitating developments in contemporary survey research.

Internet surveys show the potential to become the leading survey approach in the future. According to the Council of American Survey Research Organizations ("CASRO"), the Internet already represents the primary data collection mode for 39% of research companies in the USA (DeAngelis, 2006). The rate of adoption is slower in academic and official research but it is far from negligible. These technological innovations have, however, created several new methodological challenges.

BACKGROUND

Survey research can be performed using different modes like paper-and-pencil, mail or Internet surveys. Two characteristics of these modes (Groves et al., 2004) highlight the impact of modern technology on survey research: The presence of information technology during data collection and the degree of the interviewer's involvement.

Computer-assisted survey information collection ("CASIC") is a term embracing various modes that rely on computer technology for data collection (Couper & Nicholls, 1998). Computerized questionnaires offer numerous advantages. Answers are entered directly into a computer, which eliminates transcription-related errors. Enhanced possibilities of standardization ensure higher data quality and a lower burden on respondents. For example, answers can be limited to predefined options, irrelevant questions are automatically skipped over, answers can be subjected to real-time control and so forth.

Some computerized modes require an interviewer to be present either physically or remotely (e.g., CATI). In others, respondents complete a questionnaire by themselves. Self-administration offers benefits to respondents and researchers. Respondents can complete a questionnaire at the time and place of their preference and might have an increased sense of privacy. Researchers benefit especially from the absence of interviewers which prevents interviewer-related biases and significantly lowers research costs.

The widespread computer technology allowed both aspects to merge into computerized self-administered questionnaires ("CSAQ"). The trend toward paperless (computerized) and interviewer-less (self-administered) surveying corresponds with the general cost-optimization efforts of the research industry. Table 1 summarizes the most important survey modes based on CSAQ. They rely on electronic networks which make data instantly available to a research organization (Nathan, 2001).

Information technology has introduced new input and output technologies. Questions can be presented not only textually but also as audio or video clips. Answers, on the other hand, can be provided and recorded either manually (e.g., using a keyboard) or automatically with speech recognition. Even more advanced technologies that recognize gestures, writing and touch are establishing new potential for surveying respondents with certain disabilities (Johnston, 2007).

Internet surveys are probably the most revolutionary mode. They were enabled by progress in transmission procedures, standardized Web browsers, e-mail clients and integrated technologies (Lozar Manfreda, 2001). E-mail

Table 1. Common survey modes based on computerized self-administered questionnaires

CSAQ survey mode	
Touch-tone data entry ('TDE')	A telephone survey where respondents answer prerecorded questions by pressing appropriate numerical keys on a handset.
Interactive voice response ('IVR')	A telephone-based approach supported by the computerized voice recognition of answers. Modern systems utilize advanced speech recognition to automatically record complex answers.
Internet surveys	A range of modes based on one or more Internet services (like e-mail or Web). Respondents access and answer a questionnaire using an Internet-enabled device, usually a personal computer.
Virtual interviewer	Largely an evolving mode that can integrate various technologies (Web, speech recognition etc.). A virtual interviewer, usually a video clip of a real person, poses questions to respondents. In the future, completely virtualized characters might be used.

surveys already offered faster and cost-effective data collection (Bachmann, Elfrink, & Vazzana, 1996; Sheehan & Hoy, 1999). However, they were soon replaced by more powerful Web surveys. Interactive Web surveys are based on a continuous interaction between the system and the respondents (Conrad, Couper, & Tourangeau, 2003). The advance of modern Web technologies, like Java, JavaScript and ActiveX, fostered the implementation of feature-rich and flexible Web questionnaires. The Internet has thus become a medium for survey research, enabling different combinations of input and output technologies. Internet surveys currently remain prevailingly based on textual questionnaires. Yet, some visual or audio elements, including multimedia, are already used as enhancements of textual contents.

The implementation of Internet surveys is simplified by dedicated software tools. They provide features of questionnaire design, respondent recruitment, survey administration and data analysis. According to the WebSM portal (2007), in 2007 there were more than 300 of such tools available. They range from simple tools for single-question daily polls to advanced integrated solutions for complex data collection (Berzelak, Lozar Manfreda, & Vehovar, 2006). Especially promising is the development of some open-source solutions that allow a high level of flexibility for deployment according to a researcher's specific needs.

In the subsequent part, we focus on information technology's impact on different steps of the survey process. The emphasis is on Internet surveys which are the most promising approach and integrate several IT-based modes.

THE INTEGRATION OF INFORMATION TECHNOLOGY IN THE SURVEY PROCESS

Questionnaire Design

The appropriate implementation of a survey questionnaire is critical to the quality of obtained data (Dillman, 2007). There are important differences between perceptions of CSAQ and paper questionnaires. Graphical elements of Web survey questionnaires, for example, attract more attention than text (e.g., Tourangeau, Couper, & Galesic, 2005) and navigation with a keyboard and mouse causes a loss of eye-hand centralization (Bowker & Dillman, 2000).

Computerized questionnaires, including those of interactive Web surveys, offer a lot of design flexibility. Questions can be presented using different visual elements like check-boxes, radio buttons, sliders and so forth. Images, video and audio clips can be easily included and without extra cost. While multimedia elements might increase the motivation of respondents, the effects can also be negative or unpredictable (Dillman, 2007; Lozar Manfreda, Batagelj, & Vehovar, 2002). There is also strong evidence that different visual presentations of questions impact the answers provided by respondents (Couper, Tourangeau, & Conrad, 2004; Smyth, Dillman, Christian, & Stern, 2006). The employment of design features should be thus subjected to careful methodological examination.

Technological limitations should also be considered. The extensive use of images and multimedia is likely to exclude respondents with slow Internet connections. Further, questionnaires based on modern Web technologies can cause incompatibilities between technological platforms. The research by Buchanan and Reips (2002), for example, found personality differences between PC and Mac users. Incompatibilities can thus prevent the surveying of specific groups and significantly distort the results.

Web survey questionnaires can be simultaneously deployed on personal computers, mobile phones, handhelds, interactive TV and other Internet-enabled devices. However, some important barriers remain. The screens of mobile devices have low resolution which aggravates the deployment of complex questionnaires. Mobile phones also lack technological standardization (Tjøstheim, Thalberg, Nordlund, & Vestgården, 2005) and usually only offer limited bandwidth. Interactive TV devices (e.g., WebTV) could bring surveying closer to the everyday activities of respondents, but they remain underdeveloped. These technologies, however, might become an important vehicle of future survey research.

Sampling and Recruiting

The basic principles of sampling remain largely the same as in traditional survey modes. Yet technology has aggravated the problem of noncoverage. The employment of IT-based surveys is usually limited to segments of the population that are adequately covered with necessary technology. For example, those without Internet access cannot be reached using Web surveys. As they often significantly differ from those who do have access, the results can be highly distorted. While the coverage of organizations in the USA and the EU is almost complete, it is far lower when it comes to households. Data for 2006 show that Internet access exceeds 70% of the adult population in only a few EU member states and amounts to 73% in the USA (Eurostat, 2006; Madden, 2006). The integration of Web surveys into probability samples thus represents one of the most challenging problems of contemporary research methodology.

Information technology has enabled new approaches to the recruiting of respondents. Random visitors to Web pages are often invited to a survey using intercept sampling, usually implemented with pop-up windows or banners. In addition, invitations can be conveniently distributed by e-mail. Complete and consistent lists of e-mail addresses are, however, rarely available. The exceptions only include specific groups like the employees of an organization.

These problems can be partly overcome by mixed-mode surveys (de Leeuw, 2005). The preselection of respondents for a Web survey can, for example, be made by postal mail. Those without adequate technology are then offered to participate using a traditional mode (e.g., a mail survey). Mixed modes are, however, expensive to realize. In addition, various modes can differently impact answers and cause a mode effect (Dillman, 2007).

Data Collection

The logic of the data collection process is similar in all CSAQ modes. Respondents answer a computerized questionnaire remotely. The data are then transferred through a corresponding network (like the Internet or GSM) and automatically stored in a database.

IT-based data collection allows the recording of paradata, or data about the process of data collection (Couper, 2005). In Web surveys, they can provide information on a respondent's IP address, the duration of surveying, navigation throughout the questionnaire and so forth. This helps to effectively monitor, understand and manage the survey process.

An adequate response rate is necessary to obtain sufficiently reliable and valid data. Recent research shows that Web surveys yield, on average, 11% lower response rates than traditional modes (Lozar Manfreda, Bosnjak, Berzelak, Haas, & Vehovar, 2008). Information technol-

ogy, however, provides new opportunities for improving response rates. Innovative features can help to create user-friendly questionnaires that might heighten the motivation of respondents. Some new forms of incentives (e.g., PayPal) are available as well.

Survey Management

The effective management of a survey project ensures the high quality of obtained data at optimized costs. It is thus important to adequately balance errors and costs. In CSAQ, errors usually arise due to noncoverage and nonresponse problems. Especially Web surveys are often regarded as being an inherently cost-effective alternative. However, this becomes questionable as errors are taken into account (Vehovar, Lozar Manfreda, & Batagelj, 2001).

The survey process should adhere to relevant ethical standards and guidelines. These are well-established in general survey research, but are relatively incomplete for Internet surveys. Some of them (e.g., ESOMAR, 2005; MRA, 2000; MRS, 2006) already address important issues caused by the incorporation of information technology. These include the problems of unsolicited e-mail invitations, privacy and security threats, online informed consent, combining data from different sources, surveying children and minorities and several others. However, in practice it is often difficult to satisfy these principles. The confidence of respondents is also frequently shaken by online frauds and the large number of unsolicited e-mails. Such issues will need to be resolved by future standards and guidelines.

FUTURE TRENDS

The strong impact of information technology has importantly determined the future of survey research. It is likely that further development will be characterized by the integration of technologies, devices, data collection methods and different data sources.

CSAQ modes will be additionally fostered by the growing importance of related fields like e-learning, human-computer interaction, usability studies and online research. The wider availability of Internet access among the general population will help overcome the noncoverage problem. Text-to-speech ("TTS") and speech recognition are likely to further impact the technological foundations of surveying. The integration of these approaches will enable synergetic effects, leading to new important advances. A nearly science-fictional example might be completely virtualized interviewers that would automatically adapt to respondents' specific characteristics.

An important trend involves the distribution of questionnaires among a variety of devices. As smart-phones (like iPhone) are gaining in popularity, more powerful mobile devices might become prevalent. Advances in mobile communications, including the broad growth of 3G and WLAN, will encourage the implementation of mobile computerized self-administered questionnaires ("MCSAQ"). This will greatly overtake the current use of mobile phones, which remains largely restricted to telephone interviewing (Kuusela, Vehovar, & Callegaro, 2006) and very limited SMS surveys (Townsend, 2005). Respondents could then be reached virtually anywhere and anytime, facilitating the continuous measurement of target groups (Couper, 2005).

Future survey research tends to be increasingly based on mixed-mode surveys. Different modes will be employed in different steps, compensating for the weaknesses of individual modes (de Leeuw, 2005; Dillman, 2007). The whole project will be managed through integrated centralized data management. Information technology might also encourage combinations of online quantitative (e.g., Web surveys) and qualitative methods, including online focus groups and in-depth interviews (Lobe, 2006).

Technologically, it is already easy to combine survey data with data about phone calls, TV watching, shopping and so forth. Integrations with the Global Positioning System ("GPS") and Geographic Information Systems ("GIS") provide new opportunities for spatial and location-based research. However, such combinations raise serious ethical and methodological questions. As Couper (2005) states: "The more we can potentially learn about people, the fewer people there may be willing to give us such access into their everyday activities" (p. 493). These challenges will therefore have to be appropriately addressed by professionals and thrown open to public debate.

CONCLUSION

Recent innovations in survey data collection have been greatly influenced by information technologies. Cost-reduction trends continuously demand a move toward interviewer-less and paperless surveying, an area in which technology offers enormous potential. It provides innovative possibilities for questionnaire design, communication with respondents and data collection. These factors can significantly contribute to higher data quality, which is inevitably related to research costs.

Internet surveys are already widespread in the business sector and are being increasingly adopted in academic and official research. The existing limitations, like inadequate Internet coverage, methodological dilemmas and ethical issues, will almost certainly be adequately resolved. This will reduce the barriers for the Internet becoming a central surveying technology, at least in more economically developed countries. Further innovations and integrations will provide fresh opportunities for survey measurement, but will also raise new issues that will require appropriate solutions.

Information technology will definitely underlie the future of survey research. This, of course, does not mean that the demise of traditional survey modes is soon likely. However, they will be increasingly complemented and substituted by new, IT-based survey data collection.

REFERENCES

Bachmann, D. P., Elfrink, J., & Vazzana, G. (1996). Tracking the progress of e-mail versus snail-mail. *Marketing Research, 8*(2), 31-35.

Berzelak, J., Lozar Manfreda, K., & Vehovar, V. (2006). Software tools for Web surveys. In *Paper presented at Applied Statistics, 2006.*

Bowker, D., & Dillman, D. A. (2000). An experimental evaluation of left and right oriented screens for Web questionnaires. In *Paper presented at the Annual Meeting of the American Association for Public Opinion Research (AAPOR).*

Buchanan, T., & Reips, U.-D. (2002). Platform-dependent biases in online research: Do Mac users really think different? In *Paper presented at the German Online Research (GOR) Conference, 2002.*

Conrad, F. G., Couper, M. P., & Tourangeau, R. (2003). Interactive features in Web surveys. In *Paper presented at the Joint Meetings of the American Statistical Association.*

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review, 23*(4), 486-501.

Couper, M. P., & Nicholls, W. L., II. (1998). The history and development of computer assisted survey information collection methods. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls, II, & J. M. O'Reilly (Eds.), *Computer assisted survey information collection* (pp. 1-21). New York: John Wiley & Sons.

Couper, M. P., Tourangeau, R., & Conrad, F. G. (2004). What they see is what we get: Response options for Web surveys. *Social Science Computer Review, 22*(1), 111-127.

de Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics, 21*(2), 233-255.

DeAngelis, C. (2006). 2006 CASRO Data Trends Survey. In *Paper presented at the 2006 Data Collection Conference*. Retrieved May 28, 2008, from http://www.casro.org/techform/2006-datatrends.cfm

Dillman, D. A. (2007). *Mail and Internet surveys: The tailored design method* (2007 update, 2nd ed.). Hoboken, NJ: John Wiley & Sons.

ESOMAR. (2005). *ESOMAR guideline on conducting market and opinion research using the Internet*. Retrieved May 28, 2008, from http://www.esomar.org/uploads/pdf/ESOMAR_Codes&Guideline-Conducting_research_using_Internet.pdf

Eurostat. (2006). *Internet usage in the EU25*. Luxembourg: Eurostat.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons.

Johnston, M. (2007). Automating the survey interview with dynamic multimodal interfaces. In *Paper presented at The American Association for Public Opinion Research (AAPOR) 62nd Annual Conference.*

Kuusela, V., Vehovar, V., & Callegaro, M. (2006). Mobile phones—influence on telephone surveys. In *Paper presented at the 2nd International Conference on Telephone Survey Methodology.*

Lobe, B. (2006). *Mixing qualitative and quantitative methods in the environment of new information-communication technologies.* Unpublished doctoral dissertation, University of Ljubljana, Ljubljana.

Lozar Manfreda, K. (2001). *Web survey errors.* Unpublished doctoral dissertation, University of Ljubljana, Ljubljana.

Lozar Manfreda, K., Batagelj, Z., & Vehovar, V. (2002). Design of Web survey questionnaires: Three basic experiments. *Journal of Computer Mediated Communication, 7*(3).

Lozar Manfreda, K., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes—a meta-analysis comparing response rates. *International Journal of Market Research, 50*(1), 79-104.

Madden, M. (2006). *Internet penetration and impact: April 2006*. Washington: Pew Internet.

MRA. (2000). *Use of the Internet for conducting opinion and marketing research: Ethical guidelines*. Retrieved May 28, 2008, from http://www.mra-net.org/pdf/internet_ethics_guidelines.PDF

MRS. (2006). *Internet research guidelines*. Retrieved May 28, 2008, from http://www.mrs.org.uk/standards/downloads/revised/active/internet_mar06.pdf

Nathan, G. (2001). Telesurvey methodologies for household surveys—a review and some thoughts for the future? *Survey Methodology, 27*, 7-31.

Sheehan, K. B., & Hoy, M. G. (1999). Using e-mail to survey Internet users in the United States: Methodology

and assessment. *Journal of Computer Mediated Communication, 4*(3).

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in Web surveys. *Public Opinion Quarterly, 70*(1), 66-77.

Tjøstheim, I., Thalberg, S., Nordlund, B., & Vestgården, J. I. (2005). Are mobile phone users ready for MCASI? In ESOMAR (Ed.), *Excellence in international research* (pp. 465-488). Amsterdam: ESOMAR.

Tourangeau, R., Couper, M. P., & Galesic, M. (2005). Use of eye-tracking for studying survey response processes. In *Paper presented at the ESF SCSS Exploratory Workshop: Internet Survey Methodology: Toward Concerted European Research Efforts*.

Townsend, L. (2005). The status of wireless survey solutions: The emerging "Power of the Thumb." *Journal of Interactive Advertising, 6*(1), 52-58.

Vehovar, V., Lozar Manfreda, K., & Batagelj, Z. (2001). Sensitivity of e-commerce measurement to the survey instrument. *International Journal of Electronic Commerce, 6*(1), 31-52.

WebSM. (2007). Web survey methodology portal. Retrieved May 28, 2008, from www.websm.org

KEY TERMS

Computerized Self-Administered Questionnaires (CSAQ): Survey modes implemented with computerized questionnaires that are completed by respondents themselves (without an interviewer). CSAQ modes include Internet surveys, interactive voice response, touch-tone data entry and others.

Interactive Voice Response (IVR): A telephone survey mode based on prerecorded questions or text-to-speech technology and technology for voice recognition. Answers provided by respondents are automatically recognized and stored in a database. A modern IVR system can incorporate advanced speech recognition, enabling the automatic textual recording of complex answers.

Internet Surveys: computerized self-administered survey modes with questionnaires distributed and answered using one or more Internet services. The prevailing type is Web surveys, which has almost completely replaced e-mail surveys. Internet surveys can be distributed across various devices, including personal computers, mobile devices and interactive TV.

Mixed-Mode Survey: A survey based on a combination of different modes at various stages of a survey project. This helps to overcome the limitations of an individual mode. For example, respondents without Internet access can be offered the opportunity to complete a paper questionnaire instead of a Web one. Modern technology enables the advanced centralized data management of mixed-mode surveys.

Mobile Computerized Self-Administered Questionnaires (MCSAQ): Computerized questionnaires completed by respondents using mobile devices, usually mobile phones. Common examples are the very limited SMS surveys and the more powerful mobile Internet surveys.

Virtual Interviewer: A computerized survey mode in which questions are presented to respondents by a virtual interviewer. This is usually a prerecorded video clip of a real person asking questions. It can be provided via different media, including the Internet. The future development and integration of information technologies may enable completely virtualized characters with an adaptable appearance for different surveying contexts.

Web Survey: An Internet survey mode with questionnaires administered on the World Wide Web. Respondents access and answer the questionnaire using a Web browser. Modern Web technologies enable the client-side execution of advanced questionnaire features, including real-time skips over questions, the control of answers and others. Images and multimedia elements can also be included to enhance the contents of a questionnaire.

Information Technology Outsourcing

Anne C. Rouse
Deakin University, Australia

INTRODUCTION

Organizations have used external vendors to supply information technology (IT) functions since the first commercial implementations. In the sixties, the use of facilities management, contract programmers, and contract project management were common, but during the 70s, many organizations relied increasingly on internal delivery of IT services. The term "outsourcing" arose in the late 80s. Since that time industry has seen a fundamental change in the way information technology (IT) services are organized and delivered, with increasing reliance on external, outsourced providers. Managing outsourced IT service delivery has now become a core competence for organizations

BACKGROUND

According to Willcocks and Lacity (1998, p. 3), outsourcing involves "handing over to a third party [the] management of IS/IT assets, resources and/or activities for required results". There is general consensus that outsourcing involves delegating the responsibility for "how" to produce definable outcomes to an external party, while retaining responsibility for specifying "what" is to be delivered. Instead of controlling the behavior of service staff directly, the purchaser controls performance through a contract or service agreement, which articulates the services required, and the performance criteria.

The rise of the term "outsourcing" occurred when, in the 80s, several large U.S. corporations announced they were handing over control of their IT function to one or more vendors. The most prominent of these was Eastman Kodak. At that time, it was common in the trade literature (and some academic literature) to argue that IT had become a commodity. By outsourcing IT, it was asserted that organizations could more easily concentrate on core business. In the early 90s, announcements like Eastman Kodak's tended to produce a rise in share price, as the market anticipated consequent cost savings or improved organizational performance. Thus, the stock market response was an important outsourcing driver. Another significant driver was the growth of communications technologies, which enabled vendors to provide services remotely.

A less-frequently acknowledged reason for the rise in outsourcing was IBM's entry into the IT services arena, joining Electronic Data Systems (EDS), which had been spun off as a separate IT service vendor from its parent, General Motors. The dwindling profitability of hardware and software sales acted as an impetus, as outsourcing provided a relatively stable and long-term source of income and profits. Thus, in many ways, outsourcing is a vendor-driven phenomenon.

In the fifteen years since IT outsourcing emerged as an academic topic, the phenomenon has grown and adapted, and now embraces a range of variants. These include "business process outsourcing" (BPO), "off shore outsourcing" or "offshoring", and "application service provider" services (ASPs). Gartner (2005) reports that outsourcing is the prime driver for the IT services market, estimated to be around $US600 billion in 2004. While growth in the IT outsourcing market has slowed, the growth in new outsourcing forms (offshoring, and BPO) is reportedly strong, so the topic is likely to remain important in the IS discipline for some time.

THEORETICAL UNDERPINNINGS

There is no generally agreed "theory" of outsourcing but a range of theories drawn from economics, strategy, marketing, and public policy have been used to understand the phenomenon.

Economic Theory

Economic theories tend to view outsourcing as a variation on the "make or buy" decision that organizations must take and to view sourcing decisions as being based on relative costs. Outsourcing is seen to lead to lower cost of delivery under certain circumstances.

The most influential economic theory, and probably the most widely used in outsourcing research, is *transaction cost theory (TCT)* (Williamson & Masten, 1999). This theory predicts when decision makers will choose the *market* to deliver services and when they will choose in-house delivery through the organizational *hierarchy*. These are seen as polar forms of service provision, although a range of hybrid forms are possible.

According to TCT, the relative costs for these two strategies depend on two types of costs: production costs—usually reduced in markets because of competition—and transaction costs; and the costs of finding, contracting with and dealing

with vendors in the market. Transaction costs are difficult to measure and can be so high as to outweigh the outsourcing savings associated with reduced production costs. TCT predicts that several factors will influence whether outsourcing leads to cost savings, including level of uncertainty, frequency of transactions, and extent to which the services are "asset specific", that is, tailored to a specific vendor or purchaser and so not easily deployed elsewhere. Some TCT propositions have been confirmed for IT outsourcing (Aubert, Rivard, & Patry, 2004) though it is also argued that many of the TCT constructs are difficult to operationalize and so the theory is difficult to disprove (Ghoshal & Moran, 2005).

Resource dependency theory, and the *Resource-based view (RBV)* of the firm are two economic theories that underlie the "core competency" argument discussed next. Resource dependency theory argues that organizations will seek to reduce dependency on external providers for key resources and that factors increasing dependency include a small number of potential suppliers, and switching costs. Resource-based theory (Barney, 2002) argues that it is differences (heterogeneity) in resources between firms that allow some to sustain competitive advantage and that outsourcing allows firms to access resources (usually intangible) they do not currently have. They can then devote attention to resources and capabilities they do possess that are likely to lead to greater profitability. These include resources/capabilities that are rare, valuable, difficult to imitate, and not easily substituted—such services should not be outsourced.

Another economic theory with implications for outsourcing is *agency theory* (Laffont & Martimort, 2002). This recognizes that the provider has different motives from the purchaser and that there is often information asymmetry between the two. This theory proposes that the purchaser needs different forms of control for different types of services. Where it is difficult to measure the effort involved in service delivery, agency theory predicts that an outcomes-based, contractual form of control (like that associated with outsourcing) will lead to lower costs.

Strategic Management Theories

Much of the impetus to outsource IT has come from strategic theories related to the idea of *core competency* and the notion that managerial attention is a limited resource. The underlying proposition is that modern organizations cannot concentrate on all business functions and still achieve sustainable advantage so they must focus on those key processes or capabilities where they have unique advantages. Organizations should delegate to external providers as many non-core processes as possible, in the expectation that vendors will, through specialization and economies of scale, be able to provide higher quality services at lower cost.

As a result of this theory, there appears to be general consensus (e.g., Lacity & Willcocks, 2001) that core, or "strategic" IT services, should be kept in-house, while non-core services and commodities should be outsourced. However, this proposition has proved difficult to test as few IT services are commodities, and it is not easy to operationalize "core" services. An assumption underlying the core-competency view is that managing the relationship with a vendor will drain less attention than providing services in-house and that economies of scale and specialization will reduce vendor production costs low enough to outweigh increased transaction costs. These propositions have not been empirically verified.

Marketing Theories

While economic theories concentrate on the relative costs of outsourcing, marketing theories concentrate on the way quality, success, and value are judged by purchasers of outsourced services, and the way relationship elements, like trust, affect these judgments. An important notion that is used in studying outsourcing is that of *service quality* (Parasuraman & Zeithaml, 2002), which predicts long-term satisfaction with a vendor and intention to re-purchase.

There has been increasing attention devoted in the IT outsourcing literature to the effects that the quality of the outsourcing relationship and notions of trust have on outsourcing satisfaction and other success measures (Kern & Willcocks, 2002; Lee, Huynh, Kwok, & Pi, 2003). Theories underlying this research include service quality theory, theories related to trust, and exchange theory. Good overviews of service marketing and relationship theories can be found in White and Schneider (2003).

Public Policy Theories

The public sector is a major user of IT outsourcing, and in some regions (particularly, the UK and Australasia) has pioneered large-scale outsourcing arrangements. In the public policy literature, outsourcing is often couched as part of the "steer rather than row" philosophy (Osborne & Gaebler, 1993). This philosophy, in turn, has been influenced by strategic management theories discussed previously.

OUTSOURCING RESEARCH

A brief check of electronic resources (like ABI Inform) reveals thousands of articles on outsourcing and hundreds of papers labeled "peer reviewed" or "academic". An examination of these, though, will show that most of them are practitioner (or academic) opinion. Much of the IT outsourcing literature is written either directly, or indirectly by staff employed by outsourcing vendors or by specialist outsourcing advisory services.

A detailed review of the mainstream academic literature to 2001 is found in Dibbern, Goles, Hirschheim, and Jayatilaka (2004). This reveals that systematic empirical research into outsourcing has been limited and that much of this has concentrated on the reasons purchasers choose to outsource (or not outsource). The predominant research methodology has been "case studies" but the depth to which case studies are reported, and the extent to which the researchers adopted a critical, theory-testing approach, has varied. There have been few theory-testing studies in the literature and very few studies that have statistically tested propositions related to outsourcing practices.

A widely cited and influential paper was published by Lacity and Willcocks (1998). This summarized 61 sourcing cases the authors had undertaken (of which around half involved outsourcing). The authors compared these cases to explore various outsourcing propositions but recognized that opportunistic cases are often atypical and that their research might not generalize to wider populations. More recent survey research (Lee & Kim, 1999; Rouse & Corbitt, 2003) has not been able to confirm one of their most widely-recognized assertions—that "selective" outsourcing is generally more successful than "total" outsourcing. It is possible that other findings Lacity and Willcocks reported were unique to the cases they studied.

Drawing on either their own outsourcing experience, or on case studies, a number of writers have produced practical guidebooks for managing outsourcing arrangements. The prescriptions supplied in these guides provide rich source material for researchers, though few recommended practices have been confirmed empirically. Instead, much of the empirical research has investigated motivations for IT outsourcing. The most common goals include: to save costs, to concentrate more on core business, to gain access to skills and technologies not provided in-house, and to get better quality service or advice. Outsourcing is often employed to gain greater long-term flexibility by exchanging fixed costs or capital for variable costs. Public sector agencies also seek to outsource to increase accountability. Surprisingly, little evidence exists to confirm that these benefits (other than access to new skills and technologies) are widely obtained from outsourcing.

More recent qualitative research has been concerned with the quality of the outsourcing relationship (e.g., Kern & Blois, 2002; Barthelemy, 2003; Levina & Ross, 2003; Lacity, Feeny, & Willcocks, 2004). Confirmatory research has also investigated the role of relationship elements in outsourcing success (e.g., Lee, 2001). However, confirmatory research is currently held back by the lack of agreement about what the dependent variable should be and how it should be measured. It is not yet clear whether judgments about the relationship are the result of vendor performance, and consequent purchaser satisfaction, or whether a high quality relationship leads to satisfaction.

There is growing interest in the academic literature on the reconciliation of competing goals, and the management of the downsides or risks of outsourcing. Hirschheim and Lacity (2000) warned that in many cases there is a trade off between different goals. So, for example, flexibility often involves higher short-term costs while exchanging capital for variable costs is often more expensive in the long run. A common theme in the literature is that outsourcing tends to bring both benefits and downsides and that even after 15 years of research still involves substantial risks. These may include failure to obtain expected cost savings, reduced business flexibility, loss of control, increased dependence on the provider (even "lock in"), threats to privacy/confidentiality, and intellectual property issues (Aubert et al., 2004).

Key researchers currently conducting empirical research into IT outsourcing are included in Table 1. Excluded from this table are researchers who have reported one-off case studies or studies into the motivations/drivers of outsourcing.

In summary, a review of the literature reveals that, despite a large body of academic literature, the empirical evidence base for making outsourcing decisions is surprisingly thin. The research available to guide decision makers consists mainly of singular case studies, argumentation, and opinion, and there have been few attempts to validate claims.

FUTURE TRENDS

A growing trend in the trade literature has been the reporting of "consultant" surveys claiming IT outsourcing has generally failed to meet expectations, particularly for cost savings. This may be self-serving, though the limited academic data that has been gathered to date tends to confirm it (see Rouse & Corbitt, 2003). It is expected that more researchers will undertake hypothesis testing studies to determine whether the theoretical benefits of outsourcing are widely encountered and what the boundary conditions are.

The growth in new variations of outsourcing (BPO, offshoring, and ASPs) provides additional research focuses, and proponents argue that these may lead to greater economic benefits than IT outsourcing has been able to provide. However, as with IT outsourcing, there has been little systematic research to date, and the little research that has been done has generally involved isolated, often anecdotal case studies.

An emerging topic is how the risks and downsides of IT outsourcing (and later variants like BPO and offshoring) can better be identified and managed (e.g., Bahli & Rivard, 2005; Rouse & Corbitt, 2003b) . There is growing interest in the quality of the vendor-client relationship. With increasing proportions of IT services now outsourced, this is likely to be an area of growing interest in the future.

Table 1. Key IT outsourcing researchers and their research focus

Researcher(s)	Nature of research	Key findings
Ang; Ang and Straub	Large quantitative studies	Outsourcing is related to skills shortage; outsourcing success is predicted by fulfilled obligations; outsourcing adopters tend to concentrate on production cost savings but non-adopters on transaction costs.
Aubert, Rivard and Patry	Quantitative studies, theoretical studies, case studies	Uncertainty is the major deterrent to outsourcing, access to technical skills is the major driver; outsourcing is risky; risk management model proposed.
Barthelemy	Case studies	Outsourcing needs both hard (contractual) and soft management strategies; outsourcing has hidden costs, though some can be mitigated.
Kern; Kern and Willcocks; Willcocks and colleagues	Case studies focusing on vendor-client relationship	A range of practices recommended as likely to lead to success; outsourcing can result in negative outcomes where both vendor and client suffer from over-promising; the vendor-client relationship evolves over time and involves contractual and non contractual elements; post contract management is important to success.
Lacity, Hirschheim	In depth case studies of outsourced and insourced IT services	Outsourcing motives are frequently political; the strategy has substantial risks; outsourcing benefits can often be obtained through in-house delivery; outsourcing usually involves trade-offs.
Lacity, Willcocks, Feeny and colleagues	Case studies; cross case comparisons; small-n surveys, global outsourcing	The search for cost savings is the major driver for outsourcing; "selective outsourcing" is generally successful; a range of practices are recommended as likely to be successful; models of the outsourcing process proposed; key criteria for benchmarking suppliers and determining "knowledge potential" proposed.
Lee, Leet et al., Lee and Kim	Large quantitative studies	Relationship quality and trust predict outsourcing success; outsourcing falls into distinct configurations.
Rouse; Rouse and Corbitt	Large quantitative study, longitudinal case study, focus groups	Outsourcing is riskier than recognized; outsourcing leads to satisfaction, cost savings, in only a minority of organizations; some practices lead to discernible benefits, while others (e.g. "selective outsourcing") do not; selective outsourcing not more successful than total outsourcing.

CONCLUSION

Despite more than 15 years of study, there are large gaps in academic knowledge related to IT outsourcing. The emphasis on case study research, and the scarcity of statistically-reliable studies have resulted in an abundance of conflicting findings and claims, with limited hypothesis-testing research to reconcile these. Consequently, a number of unchallenged recommendations regarding when, what, and how to outsource have arisen. There is a need for researchers to shift emphasis from theory-generation to theory testing so as to discover which plausible recommendations are supported by evidence, and to determine the boundaries for current theory. This in turn requires that attention be paid to the dependent variable(s) associated with outsourcing success and to the various competing expectations and consequent trade-offs that outsourcing involves.

REFERENCES AND KEY READINGS

Ang, S., & Straub, D. (1998). Production and transaction economies and IS outsourcing. A study of the U.S. banking industry. *MIS Quarterly, 22*(4), 535-552.

Aubert, B. A., Rivard, S., & Patry, M. (2004). A transaction cost model of IT outsourcing. *Information and Management, 41*(7), 921-933.

Bahli, B., & Rivard, S. (2005). Validating measures of information technology outsourcing risk factors. *Omega, 33*(2), 175-187.

Barney, J. (2002). *Gaining and sustaining competitive advantage* (2nd ed.). NJ: Prentice-Hall.

Barthelemy, J. (2003). The hard and soft sides of IT outsourcing management. *European Management Journal, 21*(5), 539-549.

Dibbern , J., Goles, T., Hirschheim, R., & Jayatilaka, B. (2004, Fall). Information systems outsourcing: A survey and analysis of the literature. *ACM SIGMIS Database, 35*(4), 6-102.

Gartner (2005). *Outsourcing drives IT services growth*. Retrieved May 2005, from www.gartner.com/5_about/press_releases/pr2004.jsp

Ghoshal, S., & Moran, P. (2005). Bad for practice: A critique of the transaction cost theory. In J. Birkinshaw & G. Piramal (Eds.), *Sumantra Ghoshal on management: A force for good*. NJ: Prentice-Hall.

Hirschheim, R., & Lacity, M. (2000). the myths and realities of information technology insourcing, *Communications of the ACM, 43*(2), 99-107.

Kern, T., & Willcocks, L. P. (2002) *The relationship advantage: Information technologies, sourcing, and management*. Oxford: Oxford University Press.

Lacity, M., Feeny, D., & Willcocks, L. (2004). Commercializing the back office at Lloyd's of London: Outsourcing and strategic partnerships revisited. *European Management Journal, 22*(2), 127-140.

Lacity, M. C., & Willcocks, L. (1998). An empirical investigation of information technology sourcing practices: Lessons from experience. *MIS Quarterly, 22*(3), 363-408.

Lacity, M. C., & Willcocks, L. (2001).*Global IT outsourcing: In search of business advantage*. Chichester, UK: Wiley.

Laffont, J. J., & Martimort, D. (2002). *The theory of incentives: The principal-agent model*. Princeton NJ: Princeton University Press.

Lee, J. N. (2001) The impact of knowledge sharing, organizational capability and partnership quality on IS outsourcing success. *Information and Management*, 323-335.

Lee, J.-N., & Kim, Y.-G. (1999, Spring). Effect of partnership quality on IS outsourcing: Conceptual framework and empirical validation. *Journal of Management Information Systems, 15*(4), 29-61.

Lee, J. N., Huynh, M. A., Kwok, R. C., & Pi, S. M. (2003). IT outsourcing evolution: Past, present, and future. *Communications of the ACM, 46*(5), 84-89.

Levina, N., & Ross, J. W. (2003). From the vendor's perspective: Exploring the value proposition in IT outsourcing. *MIS Quarterly, 27*(3), 331-364.

Osborne, D., & Gaebler, T. (1993). *Reinventing government: How the entrepreneurial spirit is transforming the public sector*. Reading, MA: Addison-Wesley.

Parasuraman, A., & Zeithaml, V. A. (2002). Measuring and improving service quality: A literature review and research agenda. In B. Weitz (Ed.), *Handbook of marketing*. CA: Sage Publications.

Rouse, A. C., & Corbitt, B. (2003). Revisiting IT outsourcing risks: Analysis of a survey of Australia's Top 1000 organizations. *The 14th Australasian Conference on Information Systems*, Perth.

Rouse, A., & Corbitt, B. J. (2003b). Minimising risks in IT outsourcing: Choosing target services. *The 7th Pacific Asia Conference on Information Systems*, Adelaide, South Australia.

White, S. S., & Schneider, B. (2003). *Service quality: Research perspectives*. CA: Sage Publications.

Williamson, O. E., & Masten, S. E. (1999). *The economics of transaction costs: Elgar critical writings reader*. London: Edward Elgar Publishing.

Willcocks, L., & Lacity, M. C. (1998). *Strategic sourcing of information systems: Perspectives and practices*. Chichester, UK: Wiley.

KEY TERMS

Application Service Provider (ASPs): Standardized IT applications (such as enterprise processing, office systems, and e-mail) or software that is hosted by a provider, and accessed over the Internet by purchasers, who are charged on a transaction basis.

Business Process Outsourcing (BPO): The outsourcing of relatively complex IT-supported business functions or processes. In practice, this often also involves "offshoring".

IT Insourcing: This is a term with multiple meanings and so often unclear. It can mean services delivered in-house; services put to tender and then awarded to an in-house team in competition with the market; or services originally outsourced then brought back in house (this latter is sometimes labelled "backsourcing").

IT Outsourcing: The provision, at an agreed price, of specified IT services by an external vendor that is contracted to manage the day-to-day activities (and IS/IT assets and resources) so as to meet agreed performance and quality standards. Outsourcing can involve either the once-off development of a new information system or the ongoing provision of IT services (such as mainframe operations, PC support, software maintenance, etc.) over a specific period.

Off-Shore Outsourcing or "Offshoring": Outsourcing that is provided across national borders. In most cases this involves sourcing services from a low-salary nation like India, China, or parts of the former Soviet Union, where "salary arbitrage" (definition follows) leads to reduced costs.

Salary Arbitrage: Differences in average salary rates between developing nations (such as India, China, various South American countries) and developed nations in Europe, North America, etc.

Transaction Costs: The costs of contracting with a vendor through the marketplace, in contrast to coordinating and managing service provision "in-house" (i.e., through the hierarchy). Key costs include finding, choosing, contracting with, monitoring, and controlling the work of the vendor, as well as coordinating the vendor's activities with others being carried out by the purchaser.

Information Technology Strategy in Knowledge Diffusion Lifecycle

Zhang Li
Harbin Institute of Technology, China

Jia Qiong
Harbin Institute of Technology, China

Yao Xiao
Harbin Institute of Technology, China

INTRODUCTION

A progressive liberalization and deregulation of international trade, and the rapid development and diffusion of information and communication technology (IT) have fundamentally changed the global competitive dynamic environment (Ernst & Kim, 2002). Growing around these is a new information age economy whose fundamental sources of wealth are knowledge and communication rather than natural resources and physical labor (Kanter, 1994). The simultaneous development of the knowledge economy (Dunning, 2000) and the information technology economy (Varian, Farrell, & Shapiro, 2004) provides both opportunity and challenge for the organizations, and also requires us to develop from a comprehensive perspective by combining knowledge management with the information technology strategy.

In the knowledge economy, the importance of knowledge diffusion dynamics has been increasingly recognized in development economics over the last decade (World Bank, 1999). Knowledge diffusion can be defined as the adaptations and applications of knowledge documented in scientific publications and patents (Crane, 1972). Knowledge diffusion is part of the knowledge management process, realizing the proliferation of knowledge and information among different individuals across time and space (Chen & Hicks, 2004). According to the extent of knowledge diffusion, the knowledge diffusion lifecycle can be divided into four stages, including incubation, nurture, promotion, and popularization (Lang & Yuan, 2004). In this lifecycle, knowledge diffusion refers to promoting the innovation and core competence formation, so how to accelerate the knowledge diffusion has become an important issue for organizations.

The development of information technology establishes a solid base to accelerate knowledge diffusion. IT and related organizational innovations provide effective mechanisms for constructing flexible infrastructures that can link together and coordinate economic transactions at distant locations (Broadbent, Weill, & St. Clair, 1999). In essence, IT fosters the development of leaner, meaner, and more agile production systems that cut across firm boundaries and national borders. The underlying vision is that accelerating knowledge diffusion can speed up the dissemination of information technology. Knowledge diffusion is an essential content of the business strategy (Borghoff & Pareschi, 2003).

However, existing theories of both information technology and knowledge have not specified the information technology strategy in the knowledge diffusion. This article introduces the information technology strategy in knowledge diffusion based on the knowledge cycle theory. The article describes how to advance knowledge diffusion by using the matched information technology strategy in a different knowledge diffusion lifecycle. The article shows how firms innovate and research to imitate knowledge and improve the diffusion of knowledge.

BACKGROUND

The relationship between knowledge and information is essential. Knowledge may be defined as information whose validity has been established through a test of proof and can therefore be distinguished from opinion, speculation, beliefs, or other types of unproven information (Liebeskind, Oliver, Zucker, & Brewer, 1996). This definition of knowledge consists of two primary classifications: information (explicit knowledge) and know-how (tacit knowledge) (Nonaka, 1991). Knowledge in this article refers to explicit knowledge. Information is knowledge that can be transmitted without loss of integrity once the syntactical rules required for deciphering it are known. Thus, knowledge as information implies knowing what something means and that it can be written down (Nonaka, 1994).

Throughout the 1990s and early 2000s, both researchers and practitioners (e.g., Cowan & Jonard, 1999; Morone & Taylor, 2004) have discussed the model of knowledge diffusion within organizations. They develop a model in the framework of graph theory. The aim of their model is to capture effects of incremental innovation and their diffu-

sion over a network of heterogeneous agents. The idea that knowledge diffusion is necessary to an organization's success has become the focal point of strategy and the strategic planning process (Liebeskind et al., 1996). Knowledge has emerged as the most strategically significant resource of the firm (Grant, 1996b).

"Lifecycle" within knowledge management exists because it is evident that organizational knowledge does indeed have a lifecycle. The knowledge is discovered, captured, utilized, and eventually retired rather than killed. Siemieniuch et al. (1999) refer to the knowledge lifecycle starting point—that knowledge is not a unitary thing and it has a lifecycle in a competitive environment. In other words, if a company is to keep competitive, it must address the issues of new knowledge generation, propagation across the organization, and the subsequent knowledge retirement. They indicate that: (1) knowledge will age as the context changes; (2) humans will be intrinsic components in all processes involving the creation, utilization, and retirement of knowledge; and (3) the management of knowledge is a critical, core competence of the organization.

Siemieniuch and Sinclair (2004) also introduced the Cross Sectoral Learning in the Virtual Enterprise (CLEVER) process framework for knowledge lifecycle management (KLM). The model was developed to help organizations in the manufacturing and construction domains to tackle ill-defined knowledge management problems. Focusing on organizational and cultural issues, rather than technological ones, the framework aids the user organization in translating vague KLM problems into a set of specific knowledge management issues.

Sakol (2002) introduces the knowledge lifecycle interplay between the user study and product development phases, and introduces a method, concept, and model for the entire design process. The three proposed solutions are: object-mediated user knowledge elicitation–OMUKE, pattern of user knowledge–PUK, and use process-based product architecture–UPBPA.

OMUKE is a method proposed for capturing user knowledge. The method is built from empirical research of existing methods (convergent perspective approach) and an experimental study with the OMUKE software agents. The method can be effectively used to capture user knowledge and use it to form the product architecture in knowledge lifecycle processes.

According to the extent of knowledge diffusion, the lifecycle of knowledge diffusion will be divided into four stages, including incubation, nurture, promotion, and popularization. The knowledge lifecycle can be represented as the curve in Figure 1. The development of a core IT and communications infrastructure supports knowledge lifecycle management.

INFORMATION TECHNOLOGY STRATEGY

I

Information technology strategy refers to the IT applications used to help the organization gain a competitive advantage, reduce competitive disadvantage, or meet other strategic enterprise objectives (Bergeron, Bateau, & Raymond, 1991). Clearly, this is a critical resource, as discussed earlier. It is therefore vital that a suitable IT infrastructure is in place, with the right applications implemented. There must also be the right information- and knowledge-sharing policies in place. While scholars have explained the knowledge diffusion lifecycle and information technology in knowledge management (Siemieniuch, 1999), combining the information technology strategies with knowledge diffusion will be a potential power to accelerate the knowledge use.

Information Technology Strategy for Knowledge Diffusion

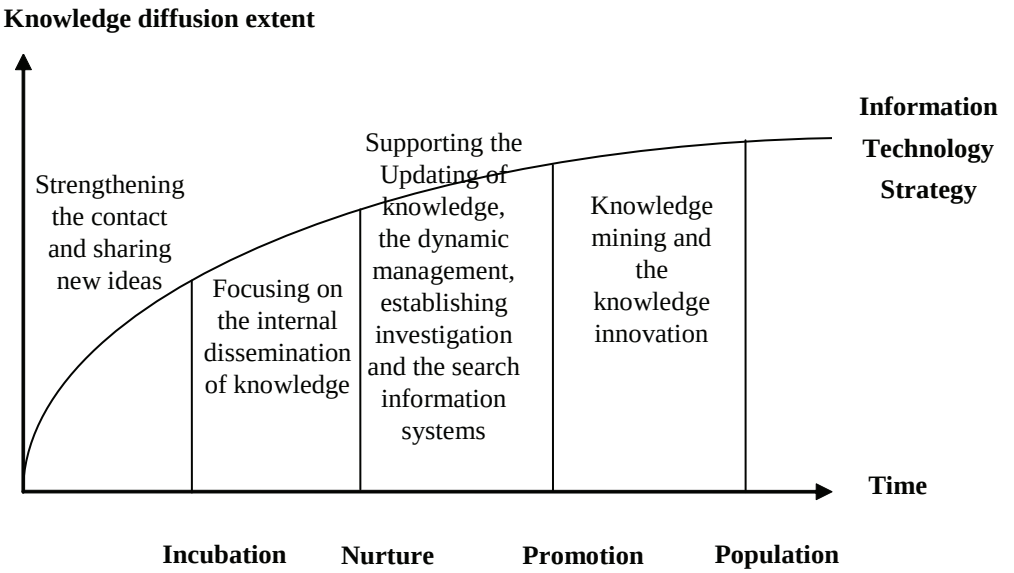
According to the extent of knowledge diffusion, the lifecycle of knowledge diffusion will be divided into four stages, including incubation, nurture, promotion, and popularization. Based on the literature, this article mainly explains how to integrate information technology with knowledge diffusion effectively in the four stages of the knowledge diffusion lifecycle.

The article provides a framework in which the knowledge diffusion lifecycle can be represented from an IT perspective, as Figure 1 shows (Lang & Yuan, 2004). Different IT strategies can be matched with the different sections of knowledge development.

Firstly, in the knowledge incubation stage, the information technology strategy emphasizes the strengthening of the contact and sharing new ideas. The knowledge incubation stage is the process of the generation of knowledge. The knowledge transforms into structural knowledge and was integrated into the knowledge resources of the company.

As stated above, knowledge capture and formalization is a critical process. This knowledge becomes an intellectual asset for the organization, owned by the shareholders in theory. It also becomes a tool for others to use, with a shorter learning time than would otherwise be the case. Perhaps more importantly, it can encapsulate best practice as a standard process for the organization. In this phase, new knowledge is still in a state of non-clarity: knowledge encoding and storage systems do not have much value. So strengthening the contact and the sharing of these new viewpoints and ideas, information systems play a very important role. The Linux system succeeds in providing a platform for exchange for everyone talking about the system, so that the system's various ideas and viewpoints on the platform by in-depth discussions can be further discussed and improved.

Figure 1. Information technology strategy in the knowledge diffusion lifecycle



Current practice for knowledge incubation seems to vary, depending on the circumstances. For problem solving, it has been suggested that discussion-capturing software agents is an appropriate methodology (Conklin, 1996). Software agents for assessing, selecting, and accumulating knowledge have appeared as good tools.

Secondly, in the knowledge nurture stage, the organization tries to gain the interests from the knowledge, so the information technology strategies focus on the internal dissemination of knowledge.

As in the knowledge incubation stage, information technology systems equally focused on the internal knowledge and information dissemination. To achieve this goal, establishing a detailed database is very important. Organizations also need to establish an online area in the existing internal network to exchange ideas and expertise. What is clear form for those companies who are regarded as leaders in this field, to incubate the knowledge is a non-trivial process. The mere provision of access to formal knowledge by wide distribution of books, software agent applications, and so on does not ensure that the knowledge will be understood, absorbed, and utilized. Each of these steps requires a process and maintenance of the process. Two highly regarded texts are those by Eason (1988) on the implementation of change and Humphrey (1989) on capability maturity models of processes, which includes the management of their lifecycles.

To ensure that new knowledge is error free, that includes a context to provide the knowledge with meaning and the conditions under which it can be used. Moreover, creating an organizational environment where accumulation and use of knowledge is in people's best interest. The use of performance assessment and the redesign of jobs are very important. Structuring workgroups is also conducive to innovative performance. This implies the autonomy for workgroups, with a clear target, sufficient resources, and the development of a commitment for the organization's goals. None of these is easy to achieve, but they are essential for motivating the new knowledge and new style of working.

Thirdly, in the knowledge promotion stage, organizations will grab interests by the spreading of knowledge. Therefore, information technology systems must support the updating of knowledge, the dynamic management, the establishment of an effective investigation, and the search information systems. Knowledge database plays an important role in the promotion of a universal stage. Knowledge coding and standardization in the organization enable it to open to the public, and it is easy for competitors to copy. The competitive advantage will come from the extent to which the information is easy to access and the extent of the quality previously made, rather than hide the knowledge.

Finally, in the knowledge popularization stage, the effective utility of the information technology in the knowledge

mining and the knowledge innovation are the most essential strategies. To popularize the knowledge that has already been popularized, it is very important to update the knowledge. Organizations must maintain the dynamic management on the original content of the document, increase new information in accordance with the applicable principles, and make the document files so they no longer change. With the passage of time, there will be a lot of documents on this knowledge published, thus the establishment of an effective investigation and search system in the literacy stage is also essential.

In this stage, an alternative approach is to provide the knowledge within a context, so that potential users can gauge for themselves the usefulness or otherwise of the knowledge (for example, when created and for what purpose, when used and for what purpose, amendments and why, and so on). More than this might be required, for example, many design ideas that are discarded for one reason or another may well contain knowledge about components that are worth keeping for future use. A particular instance of this concerns a revolutionary design in the late 1980s for fuel-injection systems, wherein an obsolete design from another company in the 1930s proved to have within it a key chunk of knowledge that overcame a severe design limitation and enabled the new design to work. In this particular instance, the retrieval from the archives happened because the designer was old enough to have been in the company when the obsolete design was still in use, and an evaluative analysis had been performed on it.

The lifecycle of knowledge diffusion can help enterprises develop each stage of knowledge diffusion and the strategic management of knowledge. From the perspective of the lifecycle, the enterprise managers could understand clearly which stage the enterprises are in. Then, the enterprises could find a balance between the storing of the knowledge and the sharing of knowledge, but they should also realize that in any case the efforts cannot avoid the possibility that the knowledge created is eroded with the passage of time. Finally, enterprises should focus on the characteristics of the knowledge in the different stages of the lifecycle, making good use of information technologies to transform the new ideas into concrete products and services, and acquire more benefits from the current stage of development of their own technology and knowledge management strategy.

FUTURE TRENDS

As stated above, the combination between the knowledge diffusion lifecycle and information technology provides us with an innovative way to implement the matched information technology strategies in the four different stages of the knowledge diffusion lifecycle. The information technology strategy supports the process of knowledge diffusion simultaneously. Especially in the knowledge economy, more organizations will become knowledge based. Knowledge diffusion may become an important process of knowledge management. Combining the information technology strategies with the corporate strategies may enhance competitiveness of enterprises greatly. Thus, the integration of the information technology strategy and the other strategy of the enterprises, such as the research and development strategy, will establish a more effective knowledge management model.

Moreover, more information technologies that have been expressly designed with knowledge management will be available in the future. The design and application of knowledge-oriented information technology provided the focus for the conference on Practical Applications of Knowledge Management held in October 1996 in Basel, Switzerland. Borghoff and Pareschi (1997) selected several contributions related to technologies supporting various types of organizational knowledge during different phases of its lifecycle—for example, a workgroup system, an experimental workflow management system in the process management of knowledge, a full-fledged knowledge engineering approach suitable for building corporate memories, the intelligent filtering system, and so on. Those information technologies were developed for process management, cooperate memories, and information filtering in knowledge management. All of these information technologies provide the guidance for the application of them into the knowledge diffusion process. More specific information technologies should be researched and developed in the future.

CONCLUSION

Based on the knowledge lifecycle theory, this article introduces effective information technology management strategies in the four stages of the knowledge diffusion lifecycle. Fundamentally, knowledge diffusion is part of the knowledge sharing process, realizing the proliferation of knowledge and information among different individuals across time and space. According to the extent of knowledge diffusion, the lifecycle of knowledge will be divided into four stages, including incubation, nurture, promotion, and popularization. Combing with the literature, this article mainly explains how to integrate the information technology with the knowledge diffusion effectively in the four stages of the lifecycle. Firstly, in the knowledge incubator stage, the information technology strategy focuses on strengthening the contact and sharing new ideas. Secondly, in the knowledge development stage, the organization tries to gain the interests from the knowledge, so the information technology strategies focus on the internal dissemination of knowledge. Thirdly, in the knowledge promotion stage, organizations will grab interests by the spreading of knowledge. Therefore, information technology systems must support the updating of knowledge, dynamic management, the establishment of

an effective investigation, and the searching of information systems. Finally, in the knowledge popularization stage, the effective utility of information technology in knowledge mining and knowledge innovation are the most essential strategies.

In conclusion, based on the knowledge diffusion lifecycle, this article makes detailed explanations on the information technology strategy of knowledge diffusion at different stages. This research will enlighten the implementation of an information technology strategy in knowledge diffusion and promote the development of the society in fierce knowledge competition.

REFERENCES

Bergeron, F., Bateau, C., & Raymond, L. (1991). Identification of strategic information systems opportunities: Applying and comparing two methodologies. *MIS Quarterly, 15*(1), 89-103.

Borghoff, U.M., & Pareschi, R. (1997). Information technology for knowledge management. *Journal of Universal Computer Science, 3*(8), 835-842.

Broadbent, M., Weill, P., & St. Clair, D. (1999). The implications of information technology infrastructure for business process redesign. *MIS Quarterly, 23*(2), 159-182.

Conklin, E.J. (1996). *Designing organizational memory: Preserving intellectual assets in a knowledge economy.* Retrieved from http://www.gdss.com/DoM.htm

Chen, C.M., & Hicks, D. (2004). Tracing knowledge diffusion. *Scientometrics, 59*(2), 199-211.

Cowan, R., & Jonard, N. (1999). *Network structure and the diffusion of knowledge.* MERIT Working Papers (pp. 99-128).

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities.* Chicago: University of Chicago Press.

Dunning, J.H. (2000). *Regions, globalization, and the knowledge-based economy.* Oxford: Oxford University Press.

Ernst, D., & Kim, L. (2002). Global production networks, knowledge diffusion. *Research Policy, 31,* 1417-1429.

Eason, K.D. (1988). *Information technology and organizational change.* London: Taylor & Francis.

Grant, R.M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal, 17*(Winter), 109-122.

Humphrey, W.S. (1989). *Managing the software process.* Reading, MA: Addison-Wesley.

Kanter, R.M. (1994). Do cultural differences make a business difference? Contextual factors affecting cross-cultural relationship success. *Journal of Management Development, 13*(2), 5-24.

Lang, J.B., & Yuan, A.F. (2004). The knowledge management in the diffusion lifecycle. *Journal of Information, 7,* 64-76.

Liebeskind, J.P., Oliver, A.L., Zucker, L., & Brewer, M. (1996). Social networks, learning, and flexibility: Sourcing scientific knowledge in new biotechnology firms. *Organization Science, 7*(4), 428-443.

Morone, P., & Taylor, R. (2004). Small world dynamics and the process of knowledge diffusion: The case of the metropolitan area of Greater Santiago de Chile. *Journal of Artificial Societies and Social Simulation, 7*(2), 327-351.

Nonaka, I. (1991). The knowledge-creating company. *Harvard Business Review,* (November-December), 96-104.

Nonaka I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

Sakol, T. (2002). *An approach to user knowledge and product architecture for knowledge lifecycle.* Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Design in the Graduate College of the Illinois Institute of Technology, USA.

Siemieniuch, C.E., & Sinclair, M.A. (1999). Organizational aspects of knowledge lifecycle management in manufacturing. *International Journal of Human Computer Studies, 51,* 517-547.

Siemieniuch, C.E., & Sinclair, M.A. (2004). CLEVER: A process framework for knowledge lifecycle management. *International Journal of Operations & Production Management, 24*(11/12), 1104.

Varian, H.R., Farrell, J., & Shapiro, C. (2004). The economics of information technology: An introduction. Cambridge: Cambridge University Press.

World Bank. (1999). *World development report 1998/1999.* New York: Oxford University Press.

KEY TERMS

Information: Knowledge that can be transmitted without loss of integrity once the syntactical rules required for deciphering it are known (Nonaka, 1994).

Information Technology Economy: The economy based on the production, assignment, and application of information technology (Varian et al., 2004).

Information Technology Strategy: The information technology applications used to help the organization gain a competitive advantage, reduce competitive disadvantage, or meet other strategic enterprise objectives (Bergeron et al., 1991).

Knowledge: Information whose validity has been established through a test of proof and can therefore be distinguished from opinion, speculation, beliefs, or other types of unproven information (Liebeskind et al., 1996). This definition of knowledge consists of two primary classifications: information (explicit knowledge) and know-how (tacit knowledge) (Nonaka, 1991). Knowledge in this article refers to explicit knowledge.

Knowledge Diffusion: The adaptations and applications of knowledge documented in scientific publications and patents (Crane, 1972).

Knowledge Diffusion Lifecycle: According to the extent of knowledge diffusion, the knowledge diffusion lifecycle can be divided into four stages, including incubation, nurture, promotion, and popularization (Lang & Yuan, 2004).

Knowledge Economy: The economy based on the production, assignment, and application of knowledge. Knowledge is the key economic asset that drives long-run economic performance (Dunning, 2000).

Knowledge Lifecycle: Knowledge is not a unitary thing, and in a competitive environment it has a lifecycle. There are four different phases of the knowledge lifecycle—socialization, internalization, externalization, and combination (Nonaka & Takeuchi, 1995).

Knowledge Lifecycle Management: The management of the environment that makes knowledge flow through all the different phases of its lifecycle (Borghoff & Pareschi, 1997).

Knowledge Management: The management activities of creating, acquiring, interpreting, retaining, and transferring knowledge to improve performance by purposefully modifying behavior based on new knowledge (Borghoff & Pareschi, 1997).

Category: Software & Systems Design

Inheritance in Programming Languages

Krishnaprasad Thirunarayan
Wright State University, USA

INTRODUCTION

Inheritance is a powerful concept employed in computer science, especially in artificial intelligence (AI), object-oriented programming (OOP), and object-oriented databases (OODB). In the field of AI, inheritance has been primarily used as a concise and effective means of representing and reasoning with common-sense knowledge (Thirunarayan, 1995). In programming languages and databases, inheritance has been used for the purpose of sharing data and methods, and for enabling modularity of software (re)use and maintenance (Lakshmanan & Thirunarayan, 1998). In this chapter, we present various design choices for incorporating inheritance into programming languages from an application programmer's perspective. In contrast with the language of mathematics, which is mature and well-understood, the embodiment of object-oriented concepts and constructs in a concrete programming language is neither fixed nor universally accepted. We exhibit programs with similar syntax in different languages that have very different semantics, and different looking programs that are equivalent. We compare and contrast method inheritance, interaction of type system with method binding, constructs for method redefinition, and their implementation in widely used languages such as C++ (Stroustrup, 1997), Java (Arnold, Gosling, & Holmes, 2005), and C# (Hejlsberg, Wiltamuth, & Golde, 2006), to illustrate subtle issues of interest to programmers. Finally, we discuss multiple inheritance briefly.

BACKGROUND

SIMULA introduced the concepts of *object*, *class*, *inheritance*, and *polymorphism* for describing discrete event simulations (Meyer, 1999). Subsequently, object-oriented programming languages such as Smalltalk, C++, Object Pascal., and so forth used these concepts for general purpose programming (Budd, 2002).

Object/Instance is a run-time structure with state and behavior. Object state is stored in its fields (variables) and behavior as its methods (functions). *Class* is a static description of objects. (In practice, a class itself can appear as a run-time structure manipulated using reflection APIs such as in Java, C#, CLOS, and so forth.) A class defines type of each field and code for each method, which inspects and/or transforms field values. (By field we mean *instance* field, and by method we mean *instance* method. The discussion of static fields and static methods, and access control primitives such as private, protected, and private, are beyond the scope of this chapter.) *Inheritance* is a binary relation between classes (say P and Q) that enables one to define a class (Q) in terms of another class (P) *incrementally,* by adding new fields, adding new methods, or modifying existing methods through overriding. A class Q is a *subclass* of class P if class Q inherits from class P.

```
class P {
      int i;
      int f() { return 2;}
      int f1() { return f() + i;}
      void g() {}
}
class Q extends P {
      int j;
      int f() { return 4;}
      int h() { return i + j;}
}
class Main {
      public static void main(String [] args) {
            Q q = new Q();
            P p = q;
            p.f1();
      }
}
```

Every P-object has an int field i, and methods f(), f1() and g() defined on it. Every Q-object has int fields i and j, and methods f(), f1(), g() and h() defined on it. Q inherits i, f1(), and g() from P, and overrides f() from P.

The variable q of type Q holds a reference to a Q-instance (an object of class Q) created in response to the constructor invocation 'new Q()' (Gosling, Joy, Steele, & Bracha, 2002). The variable p holds a reference to the same Q-instance as variable q (*dynamic aliasing*) through the *polymorphic assignment* 'p = q.' In general., *polymorphism* is the ability of a variable of type T to hold a reference to an instance of class T and its subclasses. The method f1() can be invoked on the variable p because it is of type P and f1() is defined in class P. The method f1() can be successfully invoked on a Q-instance due to method inheritance. *For a subclass to be able to reuse separately compiled method binaries in the ancestor class, the layout of the subclass instances should coincide with the layout of the ancestor instances on com-*

mon fields. The body of f1() invokes f() defined in class Q on a Q-instance referred to by variable p through *dynamic binding.* In other words, it runs the code associated with the object's class Q rather than the variable's type P. *For the method calls compiled into the ancestor's method binaries to be dynamically bound, the index of the method pointers in the ancestor method table and the descendent method table should coincide on the common methods.*

The implementation technique for reusing parent method binaries in a straightforward way works for languages with only single inheritance of classes. A language supports *multiple inheritance* if a class can have multiple parents. The implementation of multiple inheritance that can reuse separately compiled parent method binaries requires sophisticated manipulation of object reference (self/this pointer adjustment) (Stroustrup, 2002, Chapter 15), or hash table based approach (Appel, 2002, Chapter 14), in general.

Object-oriented paradigm and imperative/procedural paradigm can be viewed as two orthogonal ways of organizing heterogeneous data types (data and functions) sharing common behavior. The relative benefits and short comings of the two paradigms can be understood by considering the impact of adding new functionality and new data type. Procedural paradigm incorporates new functions incrementally while it requires major recompilation to accommodate new data types. In contrast, the object-oriented paradigm assimilates new data types smoothly but requires special *Visitor* design pattern to deal with procedure updates. Another significant advantage of object-oriented paradigm is its use of *interface* to decouple clients and servers. Wirth (1988) elucidates type extension that bridges procedural languages such as Pascal to object-oriented languages such as Modula-3 via the intermediate languages such as Modula and Oberon.

COMPARISON OF METHOD INHERITANCE IN C++, JAVA, AND C#

In this section, we discuss subtle issues associated with method inheritance in programming languages using examples from C++, Java and C#.

Single Inheritance and Method Binding in C++ vs Java

Consider a simple class hierarchy consisting of Rectangle, ColoredRectangle, and Square, coded in Java. The state of the instance is formalized in terms of its width and its height, and stored in int fields w and h. The behavior is formalized using perimeter() method which is *defined in* Rectangle, *inherited by* ColoredRectangle, and *redefined/overridden in* Square (let us say for efficiency!).

```
class Rectangle {
    int w, h;
    int perimeter() { return (2*(w+h)); }
}
class ColoredRectangle extends Rectangle {
    int c;
}
class Square extends Rectangle {
    int perimeter() { return (4*w); }
}
class OOPEg {
    public static void main (String[] args) {
        Rectangle [] rs = { new Rectangle(),
            new ColoredRectangle(), new Square()};
        for (int i = 0 ; i < rs.length ; i++ )
            System.out.println( rs[i].perimeter() );
    }
}
```

The array of rectangles is a polymorphic data structure that holds instances that are at least a Rectangle. The for-loop invokes the "correct" perimeter-method on the instance referred to by the polymorphic reference rs[i] through run-time binding of the call rs[i].perimeter() to the method code based on the class of the instance referred to by rs[i] (dynamic type of rs[i]) rather than the declared type of rs[i] (static type of rs[i]).

This code can be minimally massaged into a legal C++ program. (Note that perimeter is explicitly prefixed with keyword *virtual* in C++. #include's have been omitted.)

```
class Rectangle {
    protected int w, h;
    public virtual int perimeter() { return (2*(w+h)); }
}
class ColoredRectangle : public Rectangle {
    private int c;
}
class Square extends Rectangle {
    public int perimeter() { return (4*w); }
}

void main (char* argv, int argc) {
        Rectangle rs [3] = { Rectangle(),
            ColoredRectangle(),  Square()} ;
        for (int i = 0 ; i < RSLEN ; i++ )
            cout << rs[i].perimeter() << endl;
}
```

The main(…)-procedure in C++ resembles the corresponding main()-method in Java syntactically, but they are very different semantically. The array of Rectangle is a homogeneous structure with each element naming a Rectangle instance. The initialization assignments cause the common fields to be copied and the additional subclass instance fields to be ignored (projection). In other words, there is no polymorphism involved. Similarly, the call rs[i].

perimeter is *statically* bound and invokes the code in class Rectangle on a "direct" instance of Rectangle.

The driver code can be modified into another C++ program that is *equivalent* to the Java program given earlier using pointers and dereferencing operator.

```
void main (char* argv, int argc) {
      Rectangle* rs [3] = { new Rectangle(),
            new ColoredRectangle(), new Square()} ;
      for (int i = 0 ; i < RSLEN ; i++ )
            cout << rs[i]->perimeter() << endl;
}
```

The array of Rectangle is a homogeneous structure of polymorphic references with each element holding a reference to a (possibly indirect) Rectangle instance. The initialization assignments cause the array to hold references to three instances : a Rectangle, a ColoredRectangle and a Square. Similarly, the call rs[i]->perimeter is dynamically bound. In other words, for the ColoredRectangle instance, it runs the inherited code from class Rectangle, while for the Square instance, it runs the redefined/overriding code from class Square.

Java uses dynamic binding of methods as the default, while C++ uses static binding of methods as the default. To specify dynamic binding in C++, an explicit keyword *virtual* is necessary in front of the original overridden method definition.

Method Redefinition in Java and C#

C# resembles C++ and differs from Java in that dynamic binding of methods is not the default. To override a method, the overriding method signature must not only match the overridden method's signature, but it must also contain the keyword *override*. It is also possible to support a new subclass method that matches the parent method's signature but is logically distinct from it using the keyword *new*, instead of *override*. The rationale for this design decision is to achieve robustness in the context of code evolution in the face of changes to the parent class methods.

Consider the following C# class definitions.

```
class Parent {…}
class Child : Parent {   public virtual void m(){…} }
```

A subsequent update to the parent class with a new definition of the method m() as shown below goes undetected in Java, while C# throws an error.

```
class Parent {  public virtual void m(){…}  }
```

In C#, the programmer can state that the method m() in the subclass child is related to the method m() in the class parent by changing the former using the keyword *override* or proclaiming their independence using the keyword *new*.

```
class Child : Parent {public virtual override void m(){…}}

class Child : Parent {public virtual new void m(){…}}
```

Method Overriding in C++ and Java

A method defined in a class is guaranteed to be available in all descendant subclasses in Java (signature-based subtyping). A subclass may either inherit the method code from the ancestor or override it. In contrast, this claim does not hold in C++.

The following Java code compiles without any error. Both child and grandchild instances have two methods defined on them: one with signature m(int) and another with signature m(int, boolean).

```
class Parent { void m(int i) {  } }
class Child extends Parent { void m(int i, boolean b) {  } }
class GrandChild extends Child { void m(int i, boolean b)
{  } }
class Overload {
      public static void main(String[] args) {
            Child c = new Child();
            GrandChild gc = new GrandChild();
            c.m(5,true);                  c.m(6));
            gc.m(1,false);                gc.m(2));
}}
```

On the contrary, the following C++ code compiles with two errors. Both child and grandchild instances have only one method defined on them, and its signature is m(int, boolean). The method with signature m(int) is defined only for direct instances of Parent, and is not defined for child and grandchild instances.

```
class Parent {
      public:     void m(int i) {  }
};
class Child : public Parent {
      public:     void m(int i, bool b) {   }
};
class GrandChild : public Child {
      public:     void m(int i, bool b) {   }
};

int main() {
      Child* c = new Child();
      GrandChild* gc = new GrandChild();
      c->method(5,true);      c->method(6);      // ErRoR
      gc->method(1,false);    gc->method(2);     // ErRoR
}
```

The reason for this discrepancy can be traced to the way the method calls are resolved in C++ and Java. In C++, the

method name is searched starting from the class in which the method call appears, up the class hierarchy. The search stops at the first class that contains a definition of the method name m. Subsequently, C++ tries to match the entire signature of the method call m(…) using all the explicitly given overloaded definitions of m(…) in the "matched" class. If a method matching the method call signature is found, the search stops with success. Otherwise, C++ gives a compile-time error. In contrast, Java continues its search all the way to the root of the tree-structured class hierarchy to find a method that matches the method call signature. Effectively, in Java, if a method m(…) is defined in a class, it is also defined on all instances of its descendant subclasses.

INTERACTION OF TYPE SYSTEM WITH METHOD INHERITANCE IN JAVA AND C#

A method call in Java and C# is processed in two steps: The signature of the method to be called is fixed at compile-time, while the method definition to be invoked for a method call is determined at run-time. The static determination of method signature uses type coercion rules.

Consider the following Java class definitions and the driver program illustrating a variety of method calls. There is only one method in class P, while there are two methods in class C. The method f(P) defined in class P has been redefined in class C.

```
class P {
   public void f(P p) {   }
}
class C extends P {
   public void f(P p) {   }
   public void f(C cp) {   }
}

class Calls {
   public static void
      main(String[] args) {
         P pp = new P();   C cc = new C();   P pc = cc;
         pp.f(pp); pp.f(cc); pc.f(pp); pc.f(cc)  cc.f(pp); cc.f(cc);
   }
}
```

To process the call objExp.meth(arg), the compiler determines the static type of objExp, say T, and searches for a definition of meth compatible with the argument arg in the associated class T. The compiler freezes the signature of the method call at this stage. At run-time, the method code is chosen using the frozen signature in the class associated with the run-time object that objExpr evaluates to. For the call pp.f(pp), the frozen signature is f(P) and method run is the one defined in class P. For the call pp.f(cc), the frozen signature remains f(P), because the compiler searches for the definition of f(…) in class P, yielding the unique signature f(P), which is admissible as cc's type C is compatible with class P (coercion). The method run is the one defined in class P because pp refers to a P-instance. For the calls pc.f(pp) and pc.f(cc), the frozen signature remains f(P), but the method run is the one defined in class C because pc refers to a C-instance (dynamic binding). For the calls cc.f(pp) and cc.f(cc), the frozen signature is f(P) and f(C) respectively determined by searching class C, the declared type of cc, for a method signature match. (See Table 1)

Table 1.

Call	Compile-time Signature	Run-time Code
pp.f(pp)	f(P) in P	f(P) in P
pp.f(cc)	f(P) in P	f(P) in P
pc.f(pp)	f(P) in P	f(P) in C
pc.f(cc)	f(P) in P	f(P) in C
cc.f(pp)	f(P) in C	f(P) in C
cc.f(cc)	f(C) in C	f(C) in C

Compilation and Interpretation of C++, Java and C#

C++ programs (*.h and *.cc files) are separately compiled into assembly language to yield object code (*.o files). These are statically or dynamically linked with library routines to obtain the executables (*.exe or a.out files) that are ready to run.

Java compiler translates source programs into machine independent Java byte code. Java 1.0 run-time for a specific platform (hardware/operating system combination) interprets Java byte code by repeatedly converting each instruction into machine code for the platform and then running it. Execution of loops could be improved by factoring out the translation of byte code into machine code and caching it first-time around the loop. This led to Java 1.1 that improves the execution efficiency by *just-in-time* compilation of the entire Java byte code program into platform-specific (native) code before running it. Subsequently, this approach was observed to cause slow start-up because of wasteful dynamic compilation of rarely used byte code segments. The Hotspot virtual machine introduced with Java 2.0 begins as an interpreter, and through profiling determines frequently executed code segments (hotspots). Subsequently, it selectively just-in-time compiles only the "hotspots." This reduces the start-up time and improves the behavior of long-running server programs.

C# sources are compiled into Microsoft intermediate language (MSIL). At run-time, the MSIL code is just-in-

time compiled and executed using Common Language Runtime (CLR).

In relation to object-oriented programming, it is possible to replace dynamically bound calls to the *final* methods (methods that cannot be overridden), or to methods in the terminal classes, by static binding. Hotspot virtual machine goes further in aggressively inlining small method bodies, and statically bound calls. Due to potential dynamic class loading, Hotspot virtual machine can reverse its inlining decisions if newly loaded classes extend the existing class hierarchy.

Another major aspect of modern object-oriented language run-time is the garbage collection. Refer to (Venners, 2000) for lucid details on Java virtual machine in particular and garbage collection in general.

MULTIPLE INHERITANCE

Even though the software engineering benefits of multiple inheritance are abundantly clear, the incorporation of multiple inheritance into concrete programming language is fraught with significant complexity. Essentially, undesirable problems sneak in when accommodating good examples. Thus, there is no consensus among researchers on the semantics of multiple inheritance in the presence of method overriding and potential conflicts due to multiple definitions (Meyer, 1997).

Thirunarayan et al. (2001) reviews the approach taken in C++, Java, and Eiffel, and explores the patterns and the idioms used by the Java designers and programmers to redeem the advantages of multiple inheritance. The chapter also discusses an alternative to multiple inheritance using constructs for type-safe automatic forwarding.

FUTURE TRENDS

The systems programming languages such as SmallTalk, Common LISP, C++, Java, C#, and so forth provide the necessary infrastructure for building large applications, with efficiency of execution as an important goal. With the advent of the Internet and the WWW, there has been a sudden surge of scripting languages such as TCL/TK, PERL, Python, Ruby, Visual Basic, JavaScript, PHP, and so forth that are primarily designed for gluing applications quickly, with programming flexibility as an important goal (Ousterhout, 1998). Several scripting languages support advanced object-oriented programming features including multiple inheritance, prototypes and delegation, run-time addition of members to instances, meta-programming and reflection, and so froth. We expect future programming languages to support features that enhance programmer productivity through code reuse, program reliability through strong typing and exception facility, rich functionality through domain-specific APIs, program efficiency through dynamic optimization, and ease of use through GUI.

CONCLUSION

We selectively reviewed the basics of object-oriented programming language features and illustrated the subtleties associated with the inheritance of instance methods in widely used systems programming languages such as C++, Java and C#. For example, we explained the superiority of Java's signature-based method inheritance over C++'s use of method name-based search, C# rationale of versioning robustness for deviating from Java's instance method inheritance, and clarified Java and C#'s implementation of instance method binding using statically computed method signature and dynamically determined method definition. Overall, C++ comes across as feature-rich with lot of legacy applications, while Java and C# come across as cleaner designs for developing new applications.

In future, we expect the scripting languages popularized by improved hardware resources, and demanded by rapidly evolving, distributed and heterogeneous WWW, to support object-oriented programming features to improve programmer productivity, program reliability, functionality and ease of use.

REFERENCES

Appel, A. W. (2002). *Modern compiler implementation in Java* (2nd ed.). Cambridge: Cambridge University Press.

Arnold, K., Gosling, J., & Holmes, D. (2005). *The Java programming language*,(4th ed.). Addison-Wesley.

Budd, T. (2002). *Introduction to object-oriented programming*, (3rd ed.). Addison-Wesley.

Gosling, J., Joy, B., Steele, G., & Bracha, G. (2002). *The Java language specification* (2nd ed.). Addison-Wesley.

Hejlsberg, A., Wiltamuth, S., & Golde, P. (2006). *The C# programming Language*, (2nd ed.). Addison Wesley.

Lakshmanan, L. V. S., & Thirunarayan, K. (1998). Declarative Frameworks for Inheritance. In J. Chomicki & G. Saake (Eds.), *Logics for databases and information systems* (pp. 357-388). Kluwer Academic Publishers.

Meyer, B. (1999). *Object-oriented software construction*, (2nd ed.). Prentice Hall.

Ousterhout, J. K. (1998). Scripting: Higher-level programming for the 21st century. *IEEE Computer, 31*(3), 23-30.

Stroustrup, B. (1997). *The C++ programming language*, (3rd ed.). Addison Wesley.

Thirunarayan, K., Kniesel, G., & Hampapuram, H. (2001). Simulating multiple inheritance and generics in Java. *Computer Languages, 25*(4), 189-210.

Thirunarayan, K. (1995). Local theories of inheritance. *International Journal of Intelligent Systems, 10*(7), 617-645.

Venners, B. (2000). *Inside Java 2 Virtual Machine*, (2nd ed.). McGraw-Hill.

Wirth, N. (1988). Type extensions. *ACM Transactions on Programming Languages and Systems (TOPLAS), 10*(2), 204-214.

KEY TERMS

Class: A static description of objects. It defines types for fields and code for methods.

Multiple Inheritance: A language supports multiple inheritance if a class can have multiple parents.

Object: A run-time structure with state and behavior. Object state is stored in its fields (variables) and behavior as its methods (functions).

Static/Dynamic Binding: Binding is the association of method code to a method call. Binding carried out at compile-time is called static binding, while the binding carried out at run-time is called dynamic binding.

Strong/Static/Dynamic Typing: A strongly typed language guarantees that in a program all the operations are applied to operands of compatible type, and any type violation is flagged by the language implementation. A statically typed language makes such guarantees by compile-time analysis of a program. A dynamically typed language makes such guarantees using run-time checks. Modern object-oriented languages such as Java and C# are strongly typed and straddle the two extremes, by checking type constraints statically as much as possible, and generating code for performing additional type constraints at run-time in situations where those checks are data dependent.

Subclass (Inheritance): Inheritance is a binary relation between classes that enables one to define a class in terms of another class incrementally, by adding new fields, adding new methods, or modifying existing methods through overriding. A class Q is a subclass of class P if class Q inherits from class P.

Subtype (Polymorphism): A class Q is a subtype of a class P if an instance of Q can be used where an instance of P is required. A variable is said to be polymorphic if it can hold a reference to objects of different forms. Typically, a variable of type P can hold a reference to an instance of a subclass of class P.

Innovation Generation and Innovation Adoption

Davood Askarany
The University of Auckland, New Zealand

INTRODUCTION

The growing level of global competition is forcing organizations to make dramatic change and improvements in order to compete, prosper, and survive (Kotter, 1996). During the past three decades, the world has witnessed some spectacular changes that have provided a totally new environment for organizations. These changes include technological and administrative innovations that organizations are dealing with in different areas of their operations such as manufacturing process, operation technologies, and information systems (Shields, 1997). It can be argued that the growing level of global competition has lead to the adoption of technological evolution, which may also require the adoption of complementary administrative innovation (e.g., Baines & Langfield-Smith 2003). Given this information, to keep pace with other competitors in the global market, the adoption and the diffusion of latest ideas, techniques, practices, and processes has become a key factor for success in organizations. Therefore, this places a greater emphasis on the innovation generation and innovation adoption in organizations in order to compete, prosper, and survive.

BACKGROUND

According to diffusion theory (Rogers, 2003), we may expect that innovation generation and innovation adoption do not simply emerge and develop full-blown. Some groups of people, some places, or some organizations may have immediate access to the innovation; some may access it later, and some may never access it. As Schumpeter (1934) remarked, an innovation and its diffusion are part of a larger pattern of social, political, and economic activity. Therefore, it is expected that innovation generation and innovation adoption be influenced by a variety of factors.

Mansfield (1961) suggested that innovation adoption is a function of the degree of uncertainty associated with the innovation, the amount of investment required to adopt the innovation, and the extent of economic advantage of the innovation. Other researchers, including Brown (1981) and Robinson and Lakhani (1975), proposed a supply and demand rationale for the explanation of innovation generation and innovation adoption. Obviously, Mansfield's (1961) suggested factors for innovation adoption could be included in such a supply and demand concept. Brown (1981) explained that the market and infrastructure factors provide the supply side of innovation adoption and shape its course. He further emphasized that the central element of a supply framework is the diffusion agency.

Clark (1984) believes that the demand approach in the process of innovation generation and innovation adoption is more diverse and more extensive. It focuses on the adoption of innovations, which are available to everyone. He thinks the supply approach is dealing with cases where the innovation is not universally available due to the fact that the supply is under control. In other words, when every potential adopter of an innovation does not have equal access to an innovation, the supply factor might be considered as an important influencing factor in the diffusion process of that innovation.

The learning perspective has been introduced as another factor affecting the innovation generation and innovation adoption (Sahal, 1981). Expanding the scope of influencing factors, Hagerstrand (1967) proposed an information transfer explanation as another factor influencing the innovation generation and innovation adoption. Other researchers like Sharif and Kabir (1976) considered the diffusion of an innovation as a replacement process and claimed that the dynamics of this replacement process account for the diffusion rate during the diffusion period. However, in some cases, an innovation might be an addition to an employed technique or an ongoing capacity and not a complete replacement.

Surprisingly, continuity of the innovation progress (as an influential factor) is suggested as having a negative impact on the diffusion of an innovation. Brown (1981, p. 158) confirms that "deliberation and slowness in the adoption decision is encouraged by the continuity of the innovation process which results in many improvements during the course of diffusion." Rosenberg (1976) also confirms that there is often a delay in adoption because of the expectation of future improvements in the innovation. He emphasizes that the expectation of continued improvement might lead to a slowing down in the rate of diffusion of an innovation.

Røvik (1996) introduces "fashion" as an influencing factor, which could play an import role in the diffusion process of an innovation. He argues that the process of diffusion of innovation follows selective perception, which adjusts to the social environment and copes with what is in fashion

and what is out of fashion and usually certain innovations are chosen that seem to be more fashionable. According to Røvik (1996), "fashion" is a human-made and dynamic phenomenon that spreads by drawing attention to it. Fashion can present itself in many ways: as ideas, social organizing, specific structures and processes in organizations, and so forth. Røvik refers to fashion as an institutionalized standard for implementing new ideas, change/innovation in order to organize successfully, be up-to-date, and efficient. According to Røvik (1996, p. 159), fashion also refers to the notion that organizations are torn between "signalling a common identity and belonging to a group of organisations" and "the motive of distinguishing themselves from the other organisations and attracting attention." From this perspective, fashionable idea and innovations/changes spread by imitation, but, after a while they will be so common that some organizations may wish to demonstrate their uniqueness by developing new ideas (innovation generation) or implementing new innovations (innovation adoption), which in turn become fashionable, and so the process starts all over again.

Kanter, Barry, and Todd (1992) stress that executive sponsorship, participation, coalition building, and change agents are critical to the success of change initiatives in the process of implementing organizational change. Change agents should identify and involve opinion leaders, decision-makers on resources, functional experts, and other important persons as early as possible in the project-planning phase. They further emphasize the importance of the involvement of the people in successful implementation of changes in organizations. They suggest that all members of the change team and other employees affected by the change must not feel like as if they are just the tools for change or the subject of change; rather they should be given the chance to become actively involved, to contribute their own experiences. Every employee should feel that his/her contribution to the change process in organizations is important and valued. Thus, people will develop a sense of responsibility and ownership regarding the process of changes, which, in turn may serve as a major source of motivation to facilitate the process of such change(s) in organizations.

Most of contextual factors addressed in the diffusion literature are consistent with other theories in relation to innovation/change such as institutional, contingency, cognition, and expectancy theories. For instance, the influence of institution on diffusion of innovation addressed in diffusion literature (Rogers, 2003) supports DiMaggio and Powell's (1991) suggestion that the power of institutions may play an important role in the diffusion process of innovation. DiMaggio and Powell (1991) see the influence of institutions in almost every aspect of human life, from the way people eat to the way they shake hands and engage in conversation. According to DiMaggio and Powell (1991), institutions can be habits and social protocols right through to cultural templates and frames of meaning that define what is expected and what is regarded as "rational" or appropriate in a given situation.

Resistance to change has been mentioned by Brown (1981) as another factor, which might result in lags in the use of innovation or a slow rate of diffusion. He also refers to the development of technical skills among users as another influencing factor, which is expected to facilitate the diffusion of an innovation. Another factor influencing the rate of diffusion of an innovation is said to be profitability. Linstone and Sahal (1976) propose that the more profitable the innovation and the smaller the required investment, the greater the rate of diffusion. Profitability of an innovation can be interpreted as cost saving, relative advantage, or cost effectiveness of that innovation. Competition is said to be another factor influencing the diffusion rate of an innovation. Parker (1974) states that some of the early adoption takes place because certain firms wish to gain an advantage over their competitors. Then later adopters follow the adoption either to remain competitive or to take advantage of the innovation.

Innovation generation or innovation adoption is also expected to be affected by the characteristics of the innovation. Rogers (2003) has identified five aspects of an innovation, which affect its rate of diffusion in a population to whom the innovation is relevant. He argues that the high rate of diffusion of an innovation would be a feature of its "relative advantage" over the current practice, its "compatibility" with other aspects of the culture, its "complexity" of understanding, its "trialability" to experience, and its "observability" to see the results.

Despite Rogers (2003), Goss (1979) and Gotsch (1972) argued that there should be less stress on the role of innovation itself (characteristics of innovation) and more on the spatially variable character of society into which innovation is introduced. They believed that the diffusion rate of an innovation depends less on the nature of innovation than on the type of society existing before the innovation, while Rogers (2003) believes that the level of diffusion of an innovation depends more on its characteristics than any other influencing factors.

Yapa and Mayfield (1976) indicates that the availability and distribution of resources or individual access to the means of production and public goods affect the innovation generation and innovation adoption. Examples of resources in this context would include capital, information, public goods or services such as electricity, transportation, water systems, network communications, and education.

Innovation generation or innovation adoption is expected to be influenced by the size of firms, too. In general, large firms have several advantages over smaller firms in terms of innovation generation and innovation adoption. Brown (1981) argued that one of the advantages of large firms is their greater ability to afford capital, to put up with the costs of innovation, and bear the risk of failure. Larger firms are

also capable of better affording managerial and technical specialists. However, he further confirmed that while this might be true for higher-cost demanding innovations, there is evidence that for lower-cost innovations, size might not be important. Also, it might be argued that diffusion of some kinds of innovations may be much easier in small firms than in large firms, as small firms are expected to have less bureaucratic systems.

With regard to the previous discussion, there are, however, some significant differences in the emphasis on the factors, which are introduced as influential factors on innovation generation and innovation adoption. Some suggested factors tend to give more attention to the characteristics of the innovation itself and of the adopting firms; whereas, some other suggested factors have tended to give relatively more emphasis to the society, economy, and communication or information flow process. According to contingency theory, it is difficult to generalize such preferences; depending on the type of societies and innovations, the importance and the influence of factors responsible for innovation generation and innovation adoption might change. Social concerns in some societies might give added importance to some innovations that might have otherwise not generated of diffused, and reduce the importance of other innovations. Furthermore, an innovation generation or innovation adoption might be associated with a specific social, economic, geographical, and institutional situation within which its generation and its diffusion are more likely.

Given the variation of influencing factors addressed in the literature, Askarany (2005) develops a diffusion model classifying all influencing factors into three main categories—characteristics of innovation, characteristics of adopters, and other social and environmental factors—and divides all innovations into two main groups: innovation generation and innovation adoption.

INNOVATION GENERATION AND INNOVATION ADOPTION MODEL

Rogers (2003, p. 12) defines an innovation as "an idea, practice, or object that is perceived as new by an individual or other unit of adoption." Further, he suggests that if the individual has no perceived knowledge about an idea and sees it as new, it is an innovation. Likewise, Damanpour and Gopalakrishnan (1998, p. 3) define innovation as "the adoption of an idea or behaviour new to the organisation." The common criterion in any definition of innovation is newness. According to Rogers (2003), newness in an innovation might be expressed not only in terms of new knowledge, but also in terms of first persuasion, or a decision to adopt. The second element that needs some clarification is diffusion. Wolfe (1994) explains diffusion of an innovation as a way the new ideas are accepted (or not) by those to whom they are relevant. Rogers (2003) extends this definition to consider diffusion as a process by which an innovation is communicated through certain channels over time among the members of a social system.

A clear understanding of the complexities of the innovation process and of alternative diffusion methods is central to any innovation diffusion study. Depending on the source of innovation, the diffusion of the innovation might follow different stages, so that alternative approaches and perspectives might be applicable under different innovation diffusion processes.

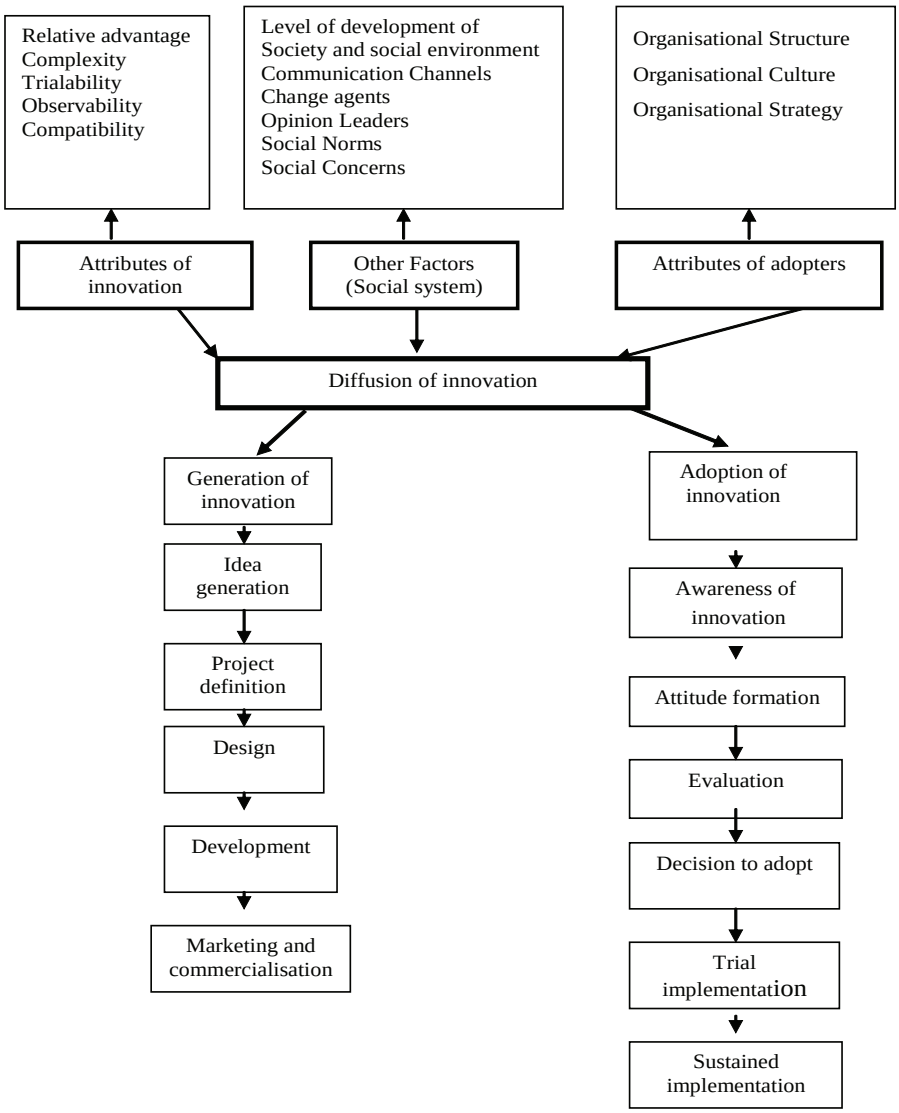
According to Damanpour and Gopalakrishnan (1998), diffusion of innovations in organizations takes place in two ways: generation and adoption. In the case of generation, innovations are generated by organizations for their own use or for export to other organizations. In the case of adoption, innovations are imported into the organization for adoption. The process of adoption of an innovation is a very long and difficult process, especially because many innovations need a long period of time to become widely adopted (Rogers, 2003). Rogers further emphasises that increasing the diffusion rate of an innovation is a common problem for potential adopters of that innovation.

The process of innovation diffusion is different when the innovation is generated by the organization; in this case, the main stages include idea generation, project definition, design, development, and marketing and commercialization (Cooper & Kleinchmidt, 1990). In the case of adoption of an innovation, which has been developed outside the organization, the stages will be awareness of innovation, attitude formation, evaluation, decision to adopt, trial implementation, and sustained implementation (Zaltman, Duncan, & Holbek, 1973). Furthermore, Wolfe (1994, p. 411) added that when innovation is generated in the organization, the stages "tend to be mulled and overlapping," while in the case of adoption the stages "tend to occur in the expected order." Depending on whether the innovation is generated within or adopted by an organization, two alternative general models can be formulated to describe the diffusion process.

Contributing to the diffusion of innovation literature, Rogers (2003) suggests that there are six phases for the diffusion of an innovation: recognition of a problem or need, basic and applied research, development, commercialization, diffusion and adoption, and consequences. Given this explanation, Rogers emphasises that these six phases are somehow arbitrary as they might not always occur in order, and some of them might be skipped in the case of particular innovations. An innovation development consists of all decisions and activities and their impacts, which occur during these phases. These suggested stages for innovation development are largely consistent with the generation approach of Damanpour and Gopalakrishnan (1998).

However, Zahra and Covin (1994) adopt a different perspective suggesting that there are three major sources

Figure 1. A general diffusion model



of innovation: imitative, acquisitive, and incubative. They define these three major sources of innovation as follows. Imitative sources are those innovations that are first introduced by other firms and then copied by organizations. Acquisitive sources also include those innovations that have been developed by other firms but are acquired through purchase, licensing, acquisition, or merger. Finally, incubative sources are those innovations that have been developed in organizations for their own use. This categorization is compatible with the generation and adoption approach of Damanpour and Gopalakrishnan (1998) in that "imitative innovations" and "acquisitive innovations" can be classified as "adopted" innovations and "incubative innovations" as "generated" ones.

From a process point of view, Rogers (2003) divides the innovation process in organizations into two sub-processes: an initiation process and an implementation process. The initiation process itself includes two stages: agenda setting and matching. These two stages involve all activities such as information gathering, conceptualizing, and planning for the adoption of an innovation. The implementation process includes three stages: redefining/restructuring, clarifying, and routinizing. These three stages contain all of the actions, events, and decisions involved in implementing an innovation. This classification is again consistent with the adoption method explained by Damanpour and Gopalakrishnan (1998).

Given the previous classification, Askarany (2005) suggests that the following general diffusion model (Figure 1) can be developed. This diffusion model is highly likely to be applicable to any diffusion study with minor modifications. Under this model, in general, what makes a diffusion research different from other diffusion research is the type or group of influencing factors and the nature of the diffusion process (innovation generation and/or innovation adoption). Therefore, depending on the type of influencing factors and the nature of diffusion process, a more detailed model can be adopted from the suggested general model and modified to tailor a specific diffusion research.

One of the main advantages of the presented model is that it is not only a comprehensive model that distinguishes between innovation generation and innovation adoption/diffusion, but it also is a simplified model that recognizes most of the contextual factors addressed in the literature and classifies them into three main categories: characteristics of innovation, characteristics of adopters, and other social and environmental factors.

Under this model, characteristics of innovations are one of three main categories of influencing factors. Moore and Benbasat (1991) classify characteristics of innovation into eight categories: "voluntariness," "relative advantage," "compatibility," "image," "ease of use," "result demonstrability," "visibility," and "trialability." However, Rogers (2003) after reviewing several thousands of instances of diffusion research, concluded that the main characteristics of innovation can be explained just by five factors: relative advantage, compatibility, complexity, trialability, and observability. Consistent with Rogers, it can be argued that "voluntariness" is more related to attributes of adopters rather than attributes of innovation, and "image" and "visibility" can be explained by the "observability" category. Nevertheless, five categories of characteristics of innovations addressed by Rogers include a variety of influencing factors such as the degree of uncertainty associated with the innovation, the amount of investment required to adopt the innovation, the extent of economic advantage of an innovation, continuity of the innovation progress, overall benefit of an innovation (including economic and non-economic advantages of an innovation), reinventing and dynamics aspects of innovations, profitability, flexibility and capability of modification of an innovation, availability of an innovation and the information about it for potential adopters, and the type of innovation.

Factors related to the adopters of innovations included three categories: organizational strategy, organizational structure, and organizational culture. In other words, most characteristics of organizations can be explained by these three categories. These categories might include factors such as size of organizations, the aggressiveness and innovativeness of their managers, level of information of organizations about the innovation, the learning perspective of organizations, resistance to change, technical skills of the users of an innovation in organizations, competition, and awareness of an innovation as a possible solution or as an available technique for progress.

Factors related to social system include the level of development of a society, communication channels in a society, social concerns, change agents, opinion leaders, and social norms. It might also be possible to include all of the influential factors that could not be related either to the innovation category or to the adopter's category under a social system category.

The previous diffusion model does not give any priority to any influencing groups. However, according to Rogers (2003), the accumulated body of literature on diffusion indicates that much effort has been devoted to studying the innovativeness and determining the characteristics of adopters, while relatively little effort has been spent in investigating how the "properties" of innovations affect their rate of adoption. He suggests that between 51 and 87 percent of variance in the adoption rate of innovations can be explained by the *characteristics of innovations*, suggesting that it is these that have the most significant influence on their diffusion. Highlighting this view, recent studies investigating the impact of other influencing factors (e.g., characteristics of innovators and social systems) on the diffusion of administrative innovations have failed to explain the majority of variances in the diffusion of such advanced techniques (e.g., Chenhall, 2003; Rogers, 2003).

FUTURE TRENDS

Given the speed and the scope of recent technological changes and innovations, it is highly likely that the diffusion of innovation will remain an import subject for future research. An overview of innovation generation and innovation adoption along with the recognition of factors influencing the diffusion of such innovations (presented in this article) is expected to provide valuable guidelines for future diffusion research. Such research is expected to facilitate the innovation generation and innovation adoption, which are seen as key factors for organizations in order to compete, prosper, and survive. Given this, the diffusion model presented in this article could be applicable to studies investigating the innovation generation and innovation adoption of both technological and administrative innovations.

CONCLUSION

The technological and administrative changes and innovations of the last three decades have placed a greater emphasis on facilitating the diffusion of such innovations in organizations in order to provide them with some advantages over

competitors to prosper, compete, and survive. To facilitate the diffusion of innovation in organization and provide a guideline for future diffusion research, the current article explains innovation generation and innovation adoption processes and addresses a variety of contextual factors influencing the diffusion processes of such innovations.

REFERENCES

Askarany, D. (2005). Diffusion of innovations in organizations. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (p. 5). Hershey, PA: Idea Group Reference.

Baines, A., & Langfield-Smith, K. (2003). Antecedents to management accounting change: A structural equation approach. *Accounting Organizations and Society, 28*(7-8), 657-698.

Brown, L. A. (1981). *Innovation diffusion: A new perspective*. New York: Methuen.

Chenhall, R. H. (2003). Management control systems design within its organisational context: Findings from contingency-based research and direction for the future. *Accounting, Organizations and Society, 28*(2-3), 127-168.

Clark, G. (1984). *Innovation diffusion: Contemporary geographical approaches*. Norwich: Geo Books.

Cooper, R. G., & Kleinchmidt, E. J. (1990). New product success factors: A comparison of "Kills" versus success and failures. *R & D Management, 20*(10), 47-63.

Damanpour, F., & Gopalakrishnan, S. (1998). Theories of organizational structure and innovation adoption: The role of environmental change. *Journal of Engineering and Technology Management, 15*(1), 1-24.

DiMaggio, P. J., & Powell, W. W. (1991). *The new institutionalism in organisational analysis*. Chicago: The University of Chicago Press.

Goss, K. F. (1979). Consequences of diffusion innovations. *Rural Sociology, 44*(4), 754-772.

Gotsch, C. H. (1972). Technical change and the distribution of income in rural areas. *American Journal of Agricultural Economics, 54*(2), 326-341.

Hagerstrand, T. (1967). *Innovation diffusion as a spatial process*. Chicago: University of Chicago Press.

Kanter, R. M., Barry A. S., & Todd D. J. (1992). *The challenge of organizational change*. New York: The Free Press.

Kotter, J. P. (1996). *Leading change*. Boston: Harvard Business School Press.

Linstone, H. A., & Sahal, D. (1976). *Technological substitution: Forecasting techniques and applications*. New York: Elsevier.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica, 29*(4), 741-766.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Parker, J. E. S. (1974). *The economics of innovation*. London: Longman.

Robinson, B., & Lakhani, C. (1975). Dynamic price models for new product planning. *Management Science, 21*(10), 1113-1122.

Rogers, E. M. (2003). *Diffusion of innovation* (5th ed.). New York: Free Press.

Rosenberg, N. (1976). On technological expectations. *Economic Journal, 86*(343), 523-535.

Røvik, K.-A. (1996). Deinstitutionalization and the logic of fashion. In B. Czarniawska & G. Sevón (Eds.), *Translating organizational change (de Gruyter Studies in Organization 56)* (pp. 139-172). Berlin: de Gruyter.

Sahal, D. (1981). *Patterns of technological innovation*. Reading MA: Addison weekly.

Schumpeter, J. A. (1934). *The theory of economic development*. Cambridge, MA: Harvard University Press.

Sharif, M. N., & Kabir, C. (1976). A generalized model for forecasting technological substitution. *Technological Forecasting and Social Change, 21*(8), 301-323.

Shields, M. D. (1997). Research in management accounting by North Americans in the 1990s. *Journal of Management Accounting Research, 9*, 3-61.

Wolfe, R. A. (1994). Organizational innovation: Review, critique and suggested research directions. *Journal of Management Studies, 31*(3), 405-431.

Yapa, L. S., & Mayfield, R. C. (1976). Non-adoption of innovations: Evidence from discriminant analysis. *Economic geography, 54*(2), 145-156.

Zahra, S. A., & Covin, J. G. (1994). The financial implications of fit between competitive strategy and innovation types and sources. *Journal of High Technology and Management, 5*(2), 183-211.

Zaltman, G., Duncan, R., & Holbek, J. (1973). *Innovations and organizations*. New York: Wiley.

KEY TERMS

Compatibility of Innovation: The degree of consistency of an innovation with the needs, expected values, and the norms of potential adopters and their social systems.

Complexity of Innovation: The degree to which an innovation seems difficult to understand and use.

Diffusion: A process by which an idea, product, practice, behavior, or object is communicated/circulated to those to whom it is relevant.

Innovation: Any idea, product, practice, behavior, or object that is apparent as new.

Observability of Innovation: The degree to which the results of an innovation can be observed or demonstrated.

Relative Advantage of Innovation: The degree to which an innovation seems to be better than the idea, object, practice, or process that it is replacing.

Trialability of Innovation: The degree to which an innovation can be tried on a limited basis before full implementation.

Innovations for Online Collaborative Learning in Mathematics

Rodney Nason
Queensland University of Technology, Australia

Earl Woodruff
OISE - University of Toronto, Canada

INTRODUCTION

The field of computer-supported collaborative learning (CSCL) has been growing in a number of areas and across a number of subjects (Koschmann, 1996; Koschmann, Hall, & Miyake, 2002; Wasson, Baggetun, Hoppe, & Ludvigsen, 2003). One of the most promising pedagogical advances, however, for online collaborative learning that has emerged in recent years is Scardamalia and Bereiter's (1996) notion of knowledge-building communities. Unfortunately, establishing and maintaining knowledge-building communities in CSCL environments such as Knowledge Forum® in the domain of mathematics has been found to be a rather intractable problem (Bereiter, 2002b; Nason, Brett, & Woodruff, 1996). In this chapter, we begin by identifying two major reasons why computer-supported knowledge-building communities in mathematics have been difficult to establish and maintain.

1. The inability of most "textbook" math problems to elicit ongoing discourse and other knowledge-building activity
2. Limitations inherent in most CSCL environments' math representational tools

Therefore, in this chapter, we argue that if mathematics education is to exploit the potentially powerful new ways of learning mathematics being provided by online knowledge-building communities, then the following innovations need to be designed and integrated into CSCL environments:

1. authentic mathematical problems that involve students in the production of mathematical models that can be discussed, critiqued, and improved, and
2. comprehension-modeling tools that (a) enable students to adequately represent mathematical problems and to translate within and across representation modes during problem solving, and (b) facilitate online student-student and teacher-student hypermedia-mediated discourse.

Both of the above innovations are directed at promoting and sustaining mathematical discourse. The requirement that the mathematical problems need to be authentic ensures that the students will have the contextual understanding necessary to promote a discussion about the mathematical models. Comprehension-modeling (Woodruff & Nason, 2003) further promotes the discourse by making student understanding yet an additional object for discussion.

Most textbook math problems do not require multiple cycles of designing, testing, and refining (Lesh & Doerr, in press), and therefore do not elicit the collaboration between people with special abilities that most authentic math problems elicit (Nason & Woodruff, 2004). Another factor that limits the potential of most textbook math problems for eliciting knowledge-building discourse is that the answers generated from textbook math problems do not provide students with much worth discussing (Bereiter, 2002b).

Another factor that has prevented most students from engaging in ongoing discourse and other mathematical knowledge-building activity within CSCL environments is the limitations inherent in their mathematical representational tools (Nason et al., 1996). Most of these tools are unable to carry out the crucial knowledge-building functions of (a) generating multiple representations of mathematical concepts, (b) linking the different representations, and (c) transmitting meaning, sense, and understanding.

Two clear implications can be derived from this review of the previous research. First is that different types of mathematical problems that have more in common with the authentic types of mathematical problems investigated by mathematics practitioners than most existing types of textbook math problems need to be designed and integrated into CSCL environments. Second, a new generation of iconic mathematical representation tools also needs to be designed and integrated into CSCL environments. In order to differentiate these tools from previous iconic math representation tools, we have labeled our new generation of tools as comprehension-modeling tools. Each of these two issues will be discussed in the next two sections of this chapter.

AUTHENTIC MATH PROBLEMS

Credence for the viewpoint that the integration of more authentic types of mathematical problems into CSCL environments may lead to conditions necessary for the establishment and maintenance of knowledge-building activity is provided by the findings from two recent research studies conducted by the coauthors. Although both of these studies were situated within elementary schools, it should be noted that the same math problems used in these research studies could also be used within online CSCL environments to facilitate the development of mathematical subject-matter knowledge in high school students and preservice teacher-education students. Therefore, we believe that the findings from these two studies have much relevance for the establishment and maintenance of math knowledge-building communities not only in elementary schools, but also in secondary school and higher education institutions, too.

In a series of research studies, Nason, Woodruff, and Lesh have been investigating whether having students engage in model-eliciting mathematical problems with collective discourse mediated by Knowledge Forum would achieve authentic, sustained, and progressive online knowledge-building activity. In this section, we focus on two of these research studies.

In the first of the research studies (Nason & Woodruff, 2004), 21 students in a Grade-6 class at a private urban Canadian school for girls were asked to devise an alternative model that could be used for ranking nations' performance at the Olympic games that de-emphasized the mind-set of "gold or nothing." In the second research study (Nason, Woodruff, & Lesh, 2002), 22 students in another Grade-6 class at the same school were asked to build a model that could help rank Canadian cities in terms of quality of life.

In both studies, the students were initially presented with an article setting the scene for the model-eliciting activity and a set of focus questions based on the article. After this 45-minute warm-up activity, the students went through the phases of (a) initial model building (Phase 1, one session of 45 minutes), (b) sharing of initial models (Phase 2, one session of 45 minutes), and (c) iterative online critiquing and revision of models within Knowledge Forum (Phase 3, four sessions of 45 minutes). The sharing of the initial models in Phase 2 was done face to face within the classroom. After the face-to-face sharing of the initial models had been completed, each group attached their math model to a Knowledge Forum note where it could be viewed and evaluated by other participants within the online CSCL community. During the online critiquing and revision of models in Phase 3, Knowledge Forum provided the contexts and scaffolds for intergroup online discourse.

Five important elements of activity consistent with Scardamalia's (2002) principles of knowledge building were observed during the course of these two studies.

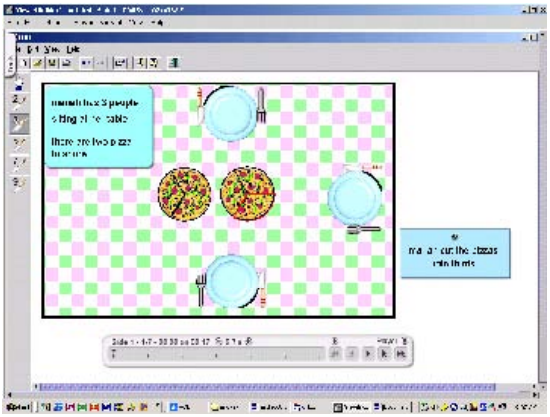
1. Redefinition of the problems, which highlights Scardamalia's principles of improvable ideas and rising above
2. Inventive use of mathematical tools, which highlights Scardamalia's principle of improvable ideas
3. Posing and exploration of conjectures, which highlights Scardamalia's principles of idea diversity and knowledge-building discourse
4. Collective pursuit of the understanding of key mathematical concepts, highlighting Scardamalia's principles of community knowledge and collective responsibility
5. Incremental improvement of mathematical models, which highlights Scardamalia's principle of improvable ideas

Much of the success in establishing and maintaining the online mathematics knowledge-building communities in these two studies can be attributed to the rich context for mathematical knowledge-building discourse provided by the model-eliciting problems. In both problems, students were required to produce a mathematical model for issues that the students found meaningful and relevant. Therefore, they were willing to proceed through multiple cycles of developing, evaluating, and revising their models. This process of proceeding through multiple cycles encouraged much online discourse between the groups in each classroom. The model-eliciting problems also had many different possible solutions. Because of this, there was much heterogeneity in the initial models produced by the groups of students. In order to understand other groups' models and also to explain their own model to other groups, each group had to engage in much iterative online discourse with other groups. During this discourse, they had to ask good questions, propose how other groups' models could be improved, and elaborate on and/or modify their explanations. Finally, the models themselves provided students with artifacts that could be discussed, evaluated, compared, and improved (just like the artifacts built by mathematics practitioners). Unlike the answers produced in most textbook problems that tend to only enable discourse about correctness (or incorrectness), the models produced from the model-eliciting problems were artifacts that could be evaluated and discussed in terms of not only correct usage of mathematical concepts and processes, but also in terms of subjective, nonmathematical factors.

COMPREHENSION MODELING

Evidence to support the notion that the inclusion within a CSCL learning environment of comprehension-modeling tools can do much to facilitate knowledge-building discourse

Figure 1. Screen shot of the problem students are attempting to solve



has been provided by research during the development of CHiLE (constructivist hypermedia interactive learning environment; Charles & Nason, 2000). CHiLE situates the learning of fractions in the context of a restaurant in which the children play the role of a waiter and are asked to partition and share out equal objects such as pizza and apple pies to customers sitting at the restaurant table (see Figure 1). The number of customers sitting at the table and the number of objects to be partitioned and shared can be varied. CHiLE provides the children with five different slicers (a knife-like tool) that enable objects to be cut in halves, thirds, fifths, sevenths, and ninths. With these slicers, the children can also create other fractions such as quarters (by halving the halves) and sixths (by halving the thirds).

CHiLE enables children to generate multiple representations of fraction problems and provides the iconic tools for facilitating synchronous hypermedia-mediated, knowledge-building, child-child and teacher-child discourse. CHiLE, however, has an added facility that enables teachers and children to also engage in online asynchronous knowledge-building discourse. With CHiLE, children are able to make an animated sequence of slides with accompanying text that not only enables them to communicate the solution to a fraction problem, but also the process (or model) that was used to generate the solution. CHiLE thus uses hypermedia as a way to animate and promote mathematical discourse: The strategy (or model) is reified on the screen via the iconic representation, the animation shows the "story," and everything is recorded, thus promoting reflection and revisitation. This is illustrated below in a series of figures (see Figures 1 to 4) generated by two 8-year-old children who had been asked to share one pizza fairly between three people.

Nason and Woodruff (2004) have found that online knowledge-building discourse facilitated by CHiLE operates on two different levels. First, the discourse can occur at the global level and focus on the overall strategy (or model). For example, another group of students, when given the same problem as in Figure 1, decided to slice the pizza into sixths and give two sixths to each person. After looking at one another's models, the two groups of students engaged in robust online debate about which was the "better" strategy (and solution). During this debate, they were able to identify similarities and differences between the strategies, but more importantly, build conceptual links between thirds and sixths. Second, the discourse can occur at the language level. For example, there can be discourse about the best language to insert in a sequence of slides. This discourse often provides the contexts for the introduction of formal mathematical language as a more precise way of communicating meaning within mathematical contexts than natural language.

The hypermedia facilities provided by CHiLE enabled children to engage in online knowledge-building discourse synchronously and asynchronously via iconic, natural language, and/or mathematical language representations.

Figure 2. Early screen shot of students' initial steps toward a solution

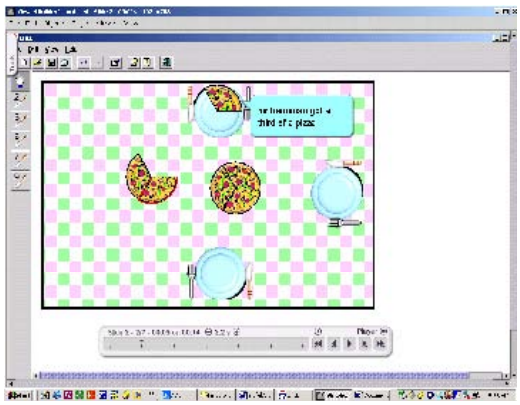


Figure 3. Screen shot midway toward a solution

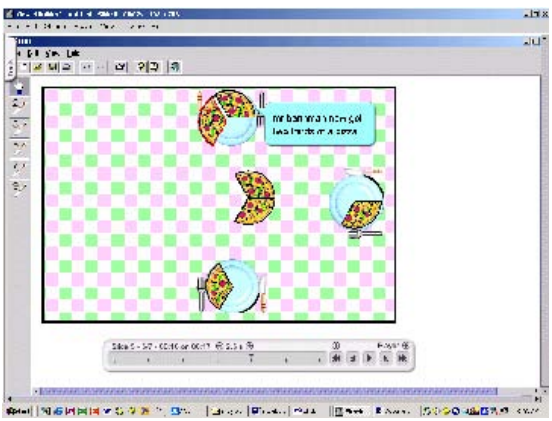


Figure 4. Screen shot of students' proposed solution

CHiLE thus provided one of the most important dynamics that Scardamalia (2002) identified as being a technical determinant of knowledge building and knowledge advancement within online CSCL environments.

CHiLE also provided two other dynamics that Scardamalia indicated were technological determinants of knowledge building and knowledge advancement. First, there is her notion that computer technology should include facilities for bringing together different ideas in such a way that productive use can be made of diversity. The iconic tools provided by CHiLE met this criterion by enabling children to readily

1. generate diverse solutions and solution processes to the same mathematics problem, and
2. communicate both synchronously and asynchronously via the iconic models, natural and mathematical language, and mathematical symbols their diverse solutions and solution processes to others within the online learning community.

Scardamalia also indicated that the computer technology also should provide children with the opportunity and the means to make revisions. Without this, she claimed that children will not be able to work continuously to improve the quality, coherence, and the utility of their ideas. One of the major qualities of CHiLE is the ease with which children can revisit and revise the sequences of slides and their accompanying text. The comprehension-modeling tools provided by CHiLE thus promoted idea diversity, improvable ideas, and knowledge-building discourse, three of the sociological and technological determinants of knowledge building identified in Scardamalia (2002).

FUTURE TRENDS

The research in progress reported in the previous two sections indicates that the inclusion of model-eliciting problems and of comprehension-modeling tools (such as CHiLE) into online collaborative learning environments both have the potential to facilitate the establishment and maintenance of online collaborative mathematics knowledge-building communities in schools and higher education institutes.

However, two important issues still need to be addressed before this potential can be realized. First, the set of principles for informing the design of model-eliciting problems developed by Lesh and Doerr (2003) need to be modified to take cognizance of the differences between online collaborative and traditional classroom environments. Second, the theoretical framework informing the design of comprehension-modeling tools needs to be modified to include not just ideas from research into external mathematical representations that were used to inform the design of CHiLE (e.g., Kaput, 1992; Olive, 2000), but also ideas from research conducted in other areas such as online collaboration (e.g., Klopfer & Woodruff, 2002), cognitive science, and multimedia learning (e.g., Mayer, 2001; Sweller, 1999). Both these issues are the foci of a series of design experiments (Bereiter, 2002a) currently being conducted by Nason and Woodruff.

CONCLUSION

In this chapter, we identified two major reasons why mathematics educators have had limited success in establishing and maintaining online knowledge-building communities.

1. The inability of most textbook math problems to elicit ongoing discourse and other knowledge-building activity
2. Limitations inherent in most CSCL environments' math representational tools

We then proposed how these two problems could be overcome, namely, by the inclusion of mathematical problems that children can analyze and describe through a mathematical model (such as the steps necessary to divide two pizzas among three people) and comprehension-modeling tools (that allow observers to later see how the students have solved the problem) within CSCL environments.

We have targeted our discussion within one CSCL groupware product called Knowledge Forum, but we believe the same principles will apply to any online computer-supported collaborative learning system. To that end, we argued that the development of model-eliciting problems suitable for use in online CSCL environments and of comprehension-modeling

tools is being restricted by the lack of adequate theoretical frameworks to inform the research and development of these two types of artifacts. Therefore, we have proposed that the development of adequate theoretical frameworks to inform the design of these two types of artifacts should be a major research priority in this field.

REFERENCES

Bereiter, C. (2002a). Design research for sustained innovation. *Cognitive Studies, Bulletin of the Japanese Cognitive Science Society, 9*(3), 321-327.

Bereiter, C. (2002b). *Education and mind in the knowledge age*. Mahwah, NJ: Erlbaum.

Charles, K., & Nason, R. A. (2000). Towards the specification of a multimedia environment to facilitate the learning of fractions. *Themes in Education, 1*(3), 263-288.

Kaput, J. J. (1992). Technology and mathematics education. In D. A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 515-556). New York: Macmillan.

Klopfer, E., & Woodruff, E. (2002). *The impact of distributed and ubiquitous computational devices on the collaborative learning environment.* Proceedings from the Annual CSCL Conference, Boulder, CO.

Koschmann, T. (Ed.). (1996). *CSCL, theory and practice of an emerging paradigm*. Mahwah, NJ: L. Erlbaum.

Koschmann, T., Hall, R., & Miyake, N. (Eds.). (2002). *CSCL2: Carrying forward the conversation.* Mahwah, NJ: L. Erlbaum.

Lesh, R., & Doerr, H. (2003). Foundations of a models and modelling perspective on mathematics teaching, learning and problem solving. In H. Doerr & R. Lesh (Eds.), *Beyond constructivism: A models and modelling perspective on mathematics learning, problem solving and teaching*. Mahwah, NJ: Erlbaum.

Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.

Nason, R. A., Brett, C., & Woodruff, E. (1996). Creating and maintaining knowledge-building communities of practice during mathematical investigations. In P. Clarkson (Ed.), *Technology in mathematics education* (pp. 20-29). Melbourne: Mathematics Education Research Group of Australasia.

Nason, R. A., & Woodruff, E. (2004). Online collaborative learning in mathematics: Some necessary innovations. In T. Roberts (Ed.), *Online learning: Practical and theoretical considerations* (pp. 103-131). Hershey, PA: Idea Group Inc.

Nason, R. A., Woodruff, E., & Lesh, R. (2002). Fostering authentic, sustained and progressive mathematical knowledge-building activity in CSCL communities. In B. Barton, C. Irwin, M. Pfannkuch, & M. O. J. Thomas (Eds.), *Mathematics education in the South Pacific* (Proceedings of the Annual Conference of the Mathematics Education Research Group of Australasia, Auckland, pp. 504-511). Sydney, Australia: MERGA.

Olive, J. (2000). Computer tools for interactive mathematical activity in the elementary school. *International Journal of Computers for Mathematical Learning, 5*(3), 241-62.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67-98). Chicago: Open Court.

Scardamalia, M., & Bereiter, C. (1996). Adaptation and understanding: A case for new cultures of schooling. In S. Vosniadou, E. De Corte, R. Glaser, & H. Mandel (Eds.), *International perspectives on the psychological foundations of technology-based learning environments* (pp. 149-165). Mahwah, NJ: Lawrence Erlbaum Associates.

Sweller, J. (1999). *Instructional design in technical areas.* Melbourne: ACER.

Wasson, B., Baggetun, R., Hoppe, U., & Ludvigsen, S. (Eds.). (2003). *CSCL2003: Community events, communication and interaction.* Bergen, Norway: University of Bergen.

Woodruff, E., & Nason, R. (2003). Math tools for knowledge-building and comprehension modeling in CSCL. In B. Wasson, R. Baggetun, U. Hoppe, & S. Ludvigsen (Eds.), *International Conference on Computer Support for Collaborative Learning, CSCL 2003: Community Events, Communication and Interaction* (pp. 31-34). Bergen, Norway: University of Bergen.

KEY TERMS

Comprehension-Modeling Tools: Math representation tools that enable users to (a) generate multiple representations of mathematical concepts and processes, (b) dynamically link the different representations, (c) communicate the mathematical ideas they have constructed, and (d) make movie-like sequences of animation slides that enable others to replay the process used to generate the solution.

Computer-Supported Collaborative Learning: collaborative learning mediated by computers.

CSCL: Acronym for computer-supported collaborative learning.

Knowledge Building: Production and improvement of knowledge objects that can be discussed, tested, compared, hypothetically modified, and so forth, and not simply the completion of school tasks.

Knowledge Forum ®: A single, communal multimedia database designed to facilitate computer-supported collaborative learning.

Mathematical Representations: concrete, pictorial, and symbolic models used to represent mathematical ideas.

Model-Eliciting Problems: Mathematical problems that involve producing models for constructing, describing, explaining, manipulating, predicting, and controlling complex systems (Lesh & Doerr, 2003).

Problem Solving: Situation involving an initial state, a goal (or solution) state, and a blockage between the initial and goal states that requires the construction of new knowledge to proceed from the initial to the goal state.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1529-1534, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Innovative Thinking in Software Development

Aybüke Aurum
University of New South Wales, Australia

INTRODUCTION

As we enter the third millennium, organizations have to cope with accelerating rates of change in technology and increased levels of competition on a global scale more than ever before. In order to stay competitive within this changing business environment, organizations are forced to constantly pursue new strategies to differentiate themselves from their competition, such as offering a stream of new products and services (Satzinger et al., 1999). Furthermore, there is growing recognition that an organization's capability to deal with change, improve services and quality, cut costs, develop new products, and compete in a global market will depend upon the level of creative and innovative thinking of its workforce (Covey, 1989). In short, in order to remain competitive in an era of increasing uncertainty and market globalization, organizations must constantly be creative and innovative with their products and services.

Software has been widely considered as central to all sophisticated innovations. In the age of the Internet the challenge is to identify and evaluate new ideas, processes and applications. In many of the fastest growing industries, including computer, entertainment, communications, advertising, logistics and finance, software has been the end product itself, or the highest value component in the end product. In other cases, software has been used to support value creation and innovation processes. The growing importance of software-based innovations suggests the need to improve the creative skills of IT professionals. This need, in turn, requires an appropriate response from the IT education and training sector. Moreover, IT education and training should better nurture students' creativity, so that they can be successful in their future roles as innovative professionals and business people. It is particularly important that IT students be given an opportunity to develop and apply creative and innovative skills to software processes and products.

Given the crucial importance of creativity and innovativeness for success in a knowledge economy, the main purpose of this article is to explore concepts about creativity and how they relate to software development by providing empirical research examples in IT education.

CONCEPT OF CREATIVITY

The literature offers diverse conceptual definitions of creativity. Glass (2001) argues that creativity is hard to define, hard to judge and hard to quantize. Kappel and Rubenstein (1999) reason that this is due to fact that creativity is used to describe a variety of things; that is, supporting the creativity process, the creative person or the creative product present different requirements for the definition of the creativity. Tomas (1999), for example, defines creativity in terms of an original idea. Shalley and Perry-Smith (2001) point out that it is not enough to only be original; also, appropriateness is vital in order to distinguish creative ideas from surreal ideas that may be unique, but have unlawful or highly unrealistic implications.

Central to creativity is the ability to generate ideas. Some psychologists and philosophers have argued that idea formation can be explained by way of association (Mednick & Mednick, 1964). This theory suggests that association occurs when two stimuli take place together (contiguity), when two stimuli are similar to each other (similarity), or when two stimuli are different from each other (contrast). Associations may be stimulated by environmental factors, by previous associations, or may be mediated by ideas related to other associates. Therefore, it is possible to have many combinations and permutations. Associations can vary in strength, depending on how often associated ideas occur together or separately.

Lateral thinking is an aid to creativity when one needs to have diverse ideas. It is a function of knowledge and imagination that may bring out discovery, innovation, imagination, and exploration. Lateral thinking consists of seeking as many alternative options as possible to the extent of one's adventurousness. In other words, it is a mental activity involving making connections between knowledge and ideas that were previously unrelated. The basis of lateral thinking is that since many problems require a different perspective to be solved successfully, individuals should suspend their judgment about what is relevant to a course of action.

CREATIVITY TECHNIQUES

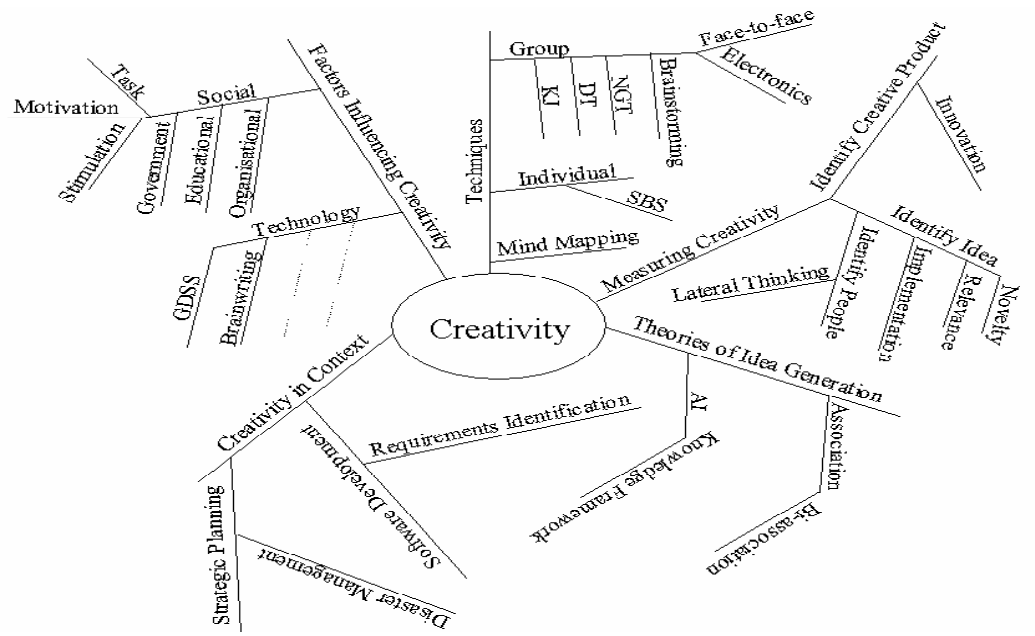
Consistent with the view that creative thinking can be learnt by appropriate stimulation and instruction, a variety of formal techniques have been developed to assist the production of novel ideas including brainstorming, mind mapping or solo brainstorming. Brainstorming and similar idea generation techniques aim to increase the production of novel ideas. The objective is to promote creativity by appropriately managing interaction within group as well as enhancing the creative environment. The procedures involved in the following examples are not difficult and may involve "lateral thinking," where ideas are stimulated by members of the group.

Brainstorming is an idea generation technique that was conceptualized by Walt Disney in the late 1920s and then expanded by Alex Osborn (1957). The objective of brainstorming is to encourage associations. The basic assumption is that it is possible for an individual(s) to generate many ideas, provided that he or she is exposed to stimuli and has experience, knowledge, and the personal flexibility to develop various permutations and combinations, and the capacity to make correct selections. This method initially emphasizes the quantity of ideas generated, leaving the assessment of quality to a later stage. Brainstorming sessions can be conducted electronically or verbally. In electronic brainstorming systems (EBS), group members share their knowledge and ideas by sending their ideas to each other, and by viewing the ideas of other members. Ideas generated from a brainstorming session can be recorded and stored in electronic files, making them easily accessible for printing or later reference (Nunamaker et al., 1991).

Another free association technique is mind mapping. This method begins with writing down a main idea in the centre of the page, and then working outward in all directions, producing a growing and organized structure composed of key words and key images, as illustrated in Figure 1. Mind mapping therefore relies on association (and clustering) of concepts/issues. The association process underling construction of the mind map actually facilitates making connections between concepts, and hence tends to generate new ideas and associations that have not been thought of before.

An example of an individual creativity technique is solo brainstorming (SBS), originally proposed by Aurum (1997). This technique is especially suited to environments where sentential analysis is appropriate, or information sources are document-based (e.g., reports, abstracts). SBS requires the individuals to adhere to a formal protocol, where a series of documents are examined and then edited. The ultimate aim in an SBS session is to determine a sufficient set of issues. As applications of the SBS protocol have been computer-based, all issues are automatically available in electronic form for further analysis. The SBS protocol touches upon an important research issue in the area of knowledge management: whether an increase in an individual's level of domain knowledge will necessarily increase their capacity to be creative within that domain. Central to the SBS protocol is the encouragement

Figure 1. Mind mapping (Aurum & Gardiner, 2003)



of participants to use their cognitive abilities by asking them to make "lateral comments".

CREATIVITY IN SOFTWARE DEVELOPMENT

Software engineering is another domain in which creativity plays an important role. The value of creativity is also well recognised in the field of system requirements determination. Robertson (2001) addresses requirements determination as "requirements discovery," which suggests that many users may not even be aware of their true requirements (e.g., unconscious requirements) without application of techniques for reflection and creativity.

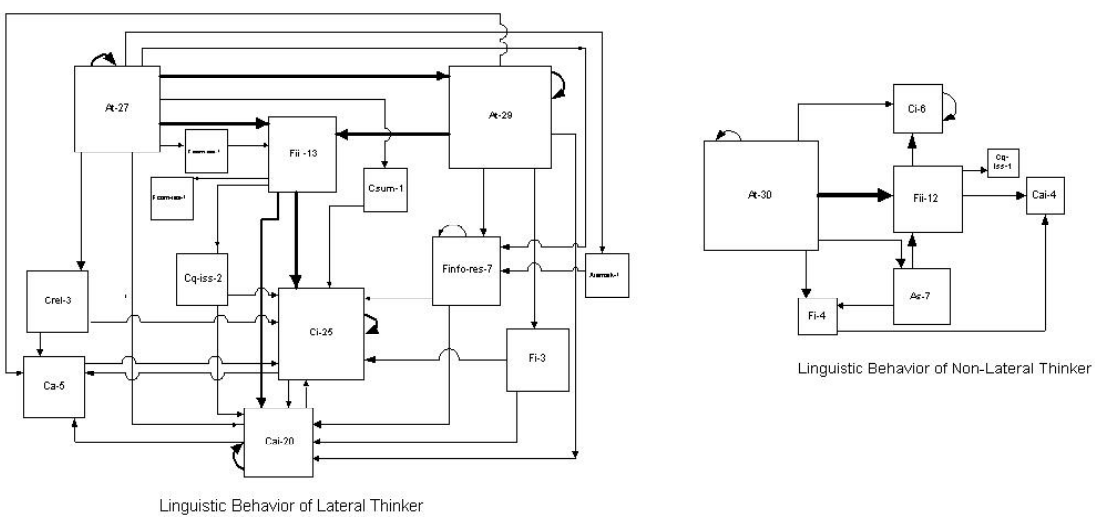
In an experiment focusing upon requirements elicitation, Aurum and Martin (1999) applied the SBS protocol to determine whether application of the protocol would deliver a richer set of requirement statements and insights. An experiment was conducted in which participants were told to adopt the role of a systems analyst retained by a fictitious organization, The Cultural Heritage Authority (CHA), to write requirements specification for their main information systems. The types of documents used as input (external information) to this study included fictitious interviews with users and abstracts from published articles addressing either heritage or marketing issues. Participants' task was to generate ideas and identify issues to be included in CHA's software requirements specification (SRS).

The objective of the research was to investigate whether the application of the SBS protocol had indeed led to a richer level of requirement specifications. The following two research questions are investigated (a) whether application of the protocol would result in identification of more relevant, workable, and original requirements issues, (b) to measure the lateral thinking by examining the linguistic behavior of participants.

The main findings of this study indicated that an SBS-based learning tool had a positive effect on participants' creative performances in development of SRS. This outcome was evident in the originality aspect of task performance, but not relevance and workability. Users were found to generate significantly more original ideas as the result of their interaction with the tool, while maintaining similar levels of relevance and workability (Aurum et al., 2003).

Aurum (1997) suggests that the level of "laterality" for any thought for a given problem can only be assessed with respect to the thoughts generated by others for the same problem. Aurum also found that documents generated from SBS session exhibited some unique characteristics. From linguistic analysis, it was possible to identify those users who were able to think laterally in the SBS session. Furthermore, lateral thinkers displayed a more complex linguistic pattern than non-lateral thinkers, as illustrated in Figure 2. Participants who generated many ideas and identified many issues were

Figure 2. Linguistic behavior of SBS participants



also found to be the lateral thinkers. The findings showed that lateral thinkers wrote many "issue loaded" or "idea loaded" sentences, whereas non-lateral thinkers produced fewer ideas. However, the distinction between these two groups was not clear-cut, but rather a continuum.

The results of this study indicate that the SBS is a promising method for stimulating creative thinking and idea generation in a software development task. Essentially, the brainstorming session helped students uncover ideas without being constrained, stimulate their own thinking by external influences, and capture their thoughts.

These findings also have some important implications for software development and IT education. They demonstrate that creativity can be improved, leading to higher quality software designs. The findings also suggest that the type of tool tested here may be a useful teaching tool in a variety of IT courses involving creative thinking and problem solving. Furthermore, the tool is likely to be most valuable in situations where the problem is unstructured, goals indistinct, and where the outcome of an action cannot always be clearly identified. The tool is a relatively generic one, since it uses a technique that can be applied to a variety of scenarios and can help people process relevant documents whilst identifying issues. These documents act like a 'trigger to stimulate domain-specific ideas from users.

CONCLUSION

Many organizations have come to realize that the creativity of their management and employees is an important source for competitive advantage. However, arguably more can be done within these organizations to promote a creative culture – for example, more organizations should seek to reward management and employees for creative (or divergent) displays, and make creativity supporting technologies more readily available to them.

A number of techniques have been developed to facilitate creativity, with many techniques based upon some form of brainstorming. One theme common to some of the more recent studies on creativity is the importance of a rich source of stimuli to support the creative process, whether the stimuli are: documents, as in Aurum and Martin (1999); group memory, as in Satzinger et al. (1999); or models, as in Shalley and Perry-Smith (2001). Other forms of stimuli include: text, audio, graphics, simulations, video, and so forth (Kletke et al., 2001). Indeed, the effectiveness of the brainstorming technique relies on participants being stimulated by the ideas contributed by others. The potential to cascade ideas is referred to as synergy (Dennis & Valacich, 1993) – that is, the ability of an idea from one participant to trigger in another participant a new idea that would otherwise not have been produced. Another technique is formalizing the creative process through some protocol that can be an effective strategy in terms of supporting the level of intrinsic motivation and mental effort required by participants undertaking a creativity task (Aurum 1999; Paulus & Yang, 2000). Application of a formal protocol is usually at the heart of a creativity technique, and ensures a more systematic and thorough approach to information analysis, which is essential for many creativity tasks.

REFERENCES

Aurum, A. (1997). *Solo brainstorming: Behavioral analysis of decision-makers.* PhD thesis. University of New South Wales, Australia.

Aurum, A., & Gardiner, A. (2003). Creative idea generation. In H. Hasan & M. Handzic (Eds.), *Studies on knowledge management* (pp. 57-91). University of Wollongong Press.

Aurum, A., Handzic, M., & Gardiner, A. (2003). Preparing IT professionals for creative development. In T. McGill (Ed.), *Supporting creativity in requirements engineering: An application in IT education.* Hershey, PA: Idea Group Publishing.

Aurum, A., & Martin, E. (1999). Managing both individual and collective participation in software requirements elicitation process. *14th International Symposium on Computer & Information Sciences,* Kusadasi, Turkey (pp. 124-131).

Covey, S.R. (1989). *The 7 habits of highly effective people.* New York: Rockfeller Center.

Dennis, A.R., & Valacich, J.S. (1993). Computer brainstorms: More heads are better than one. *Journal of Applied Psychology, 78*(4), 531-537.

Glass, R.L. (2001, September/October). A story about the creativity involved in software work. *IEEE Software,* 96-97.

Kappel, T.A., & Rubenstein, A.H. (1999). Creativity in design: The contribution of information technology. *IEEE Transaction on Engineering Management, 46*(2), 132-143.

Kletke, M.G., Mackay, J.M., Barr, S.H., & Jones, B. (2001). Creativity in the organization: The role of individual creative problem solving and computer support. *International Journal of Human-Computer Studies, 55,* 217-237.

Mednick, S.A., & Mednick, M.T. (1964). An associative interpretation of the creative process. In C.W. Taylor (Ed.), *Widening horizons in creativity.* New York: John Wiley & Sons.

Nunamaker, J.F., Dennis, A.R., Valacich, J.S., Vogel, D.R., & George, J.F. (1991). Electronic meeting systems to support group work. *Communications of the ACM, 34*(7), 40-61.

Osborne, A. (1957). *Applied imagination: Principles and procedures of creative thinking.* New York: Charles Scribner's Sons.

Paulus, P.B., & Yang, H. (2000). Idea generation in groups: A basis for creativity in organizations. *Organizational Behavior and Human Decision Processes, 82*(1), 76-87.

Robertson, S. (2001). Requirements trawling: Techniques for discovering requirements. *International Journal of Human-Computer Studies, 55,* 405-421.

Satzinger, J.W., Garfield, J.M., & Nagasundaram, M. (1999). The creative process: The effects of group memory on individual idea generation. *Journal of Management Information Systems, 14*(4), 143-160.

Shalley, C.E., & Perry-Smith, J.E. (2001). Effects of social-psychological factors on creative performance: The role of informational and controlling expected evaluation and modelling experience. *Organizational Behavior and Human Decision Processes, 84*(1), 1-22.

Tomas, S. (1999). Creative problem-solving: An approach to generating ideas. *Hospital Material Management Quarterly, 20*(4), 33-45.

KEY TERMS

Creativity: There are many views about the definition of creativity. In the context of discovery, creativity is the ability to generate or recognize ideas, alternatives that might be useful solving problems. There are several aspects of creativity, including creative product or value, creative person/people, creative environment, creative symbols and creative process.

Electronic Brainstorming Systems (EBS): A computer-based system that facilitates brainstorming between group members.

Information System (IS): A system that uses IT to capture, transmit, store, retrieve, manipulate or display data for business processes in an organization.

Information Technology (IT): Computer hardware and software, as well as the peripheral devices closely associated with computer-based systems that facilitate data processing tasks, such as capturing, transmitting, storing, retrieving, manipulating or displaying data. IT includes matters concerned with design, development, and implementation of information systems and applications.

Internet: A worldwide network of computer networks that use the TCP/IP network protocols to facilitate data transmission. It provides access to a vast amount of information resources including multimedia (movies, sound, and images), software, text documents, news articles, electronic journal, travel information and so forth. It also provides an environment for buying and selling products and services over a network.

Knowledge Economy: Economic growth is driven by the accumulation of knowledge, which is the basic form of capital. A knowledge driven economy is one in which the generation and exploitation of knowledge plays the predominant part in the creation of wealth.

Knowledge Management (KM): The collection of processes that manage the creation, dissemination, and utilization of knowledge for learning, problem solving, and decision-making. KM often encompasses identifying intellectual assets within organizations. The management of knowledge is regarded as a main source of competitive advantage for organizations. KM brings together three organizational resources: people, process and technologies, and enables the organization to use and share information more effectively

Software Requirements Specification: A document that contains all requirements, for example functional and non-functional requirements and project issues, of the system as agreed upon by customers and software developers.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1535-1539, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Institutional Isomorphism and New Technologies

Francesco Amoretti
University of Salerno, Italy

Fortunato Musella
University of Naples Federico II, Italy

INTRODUCTION

Technological factor is mainly underestimated in the literature on institutions and organizations. Although organizational studies and information technology are disciplines dedicated respectively to studying socio-political and technical aspects of organizing, cross-fertilization among such fields has remained quite limited. Only rarely the variable of technology has been interpreted as a crucial element for explaining institutional uniformity. From a more general point of view, changing technical factors have been considered "relatively unimportant sources of organizational change in a mature organizational field" (Yang, 2003, p. 433).

Only after the spread of the information and communication technologies (ICTs), a good number of studies has started to consider the relationships among information technology and organizational structure (Guthrie, 1999). Neo-institutional analysis on the use of information technology was mostly directed at showing how the embeddedness of organizational actors "in cognitive, cultural, social, and institutional structures influences the design, perceptions, and uses of the Internet and related [information technology]" (Fountain, 2001, p. 88). Therefore, it can been argued that most of the literature on this field concerns the way in which technology represents a social construct, because it shows that any technological application is strongly influenced by social aspects, such as cognitive frames, political culture, local traditions and so forth.

Yet, a few contributions have been dedicated until now to investigate how institutions change through the introduction of new technologies. Although technological innovation is said to be the source of variation in a given institutional context, as "new technology offers new possibilities for solving problems [and] new practices arise when innovative organizations take advantage of its novel benefits" (Leblebici, 1991, p. 335), little attention is focused on technological variables. Despite such disregard, in the following article some examples of the strategic use of information and communication technologies will be included, with specific reference to pressures exerted by ICTs for producing "institutional isomorphism."

BACKGROUND

Institutional isomorphism represents a central issue in the neo-institutional approach. Such concepts refer to the way organizations in a population are forced to "resemble" other organizations that "face the same set of environmental conditions." (DiMaggio & Powell, 1983, p. 66). It deals in part with the organizational process of homogenization.

The main contributions on institutional isomorphism can be divided in two different fields of research, only rarely put together. The first one can be identified with the analysis of organization, which looks to pressures leading to conformity among organizational actors. The second one has found new areas of application after the process of globalization: it is the study of policy transfers, which aims at underlining the policy convergence in different political contexts. Both the approaches show many similarities in the discourse on institutional change and can be also associated for the lack of consideration for technological variables. How the adoption of a new technology may influence and being influenced by pre-existing institutional setting has represented a crucial and underestimated issue. Yet, after the spread of ICTs, the relevance of the strategic use of technology for producing institutional effects is becoming more evident.

The analysis of organizational isomorphism refers to a notable source of inspiration. In *The Protestant Ethic and the Spirit of Capitalism*, Max Weber (1905/1958) introduced the imagery of the iron cage to catch the process of bureaucratic homogenization in which the humanity was imprisoned. Organizations were deemed to a destiny of increased rationalization that would make them more similar to each other in structure, culture and outputs. The same image of imprisonment was used—and revised—by DiMaggio and Powell (1983) in their note studies on the mechanisms of institutional isomorphic change. As a constraining pressure which forces one unit in a population to resemble other units that face the same sets of environmental conditions, the concept of isomorphism aims at explaining why there is homogeneity of forms and practices in a given organizational field.

Three type of isomorphism can be identified. Coercive forces occur when an actor influences other actor's behaviour through formal and informal instruments. Generally, in this circumstance an organism legally imposes rules that are to be followed by a specific set of organizations. The most classical and remarkable example is represented by the state in its capacity of acting as a normative power which defines standards and rules of action. Mimetic isomorphism does not derive from a coercive authority, mainly responding to the search for efficiency and the need of institutional legitimacy. It can be argued that an organization that builds a reputation for excellence will attract other organizations trying to imitate its practices. Finally, the third source, the normative one, for producing isomorphism is constituted by the relevant trend of *professionalization*, interpreted as a "collective struggle of members of an occupation to define the conditions and methods of their works, to control 'the production of producers'" (DiMaggio & Powell, 1983, p. 152).

The processes by which isomorphism mechanisms intervene show a mix of different institutional logics. Action is driven by the principle of rationality as well as by rules of appropriate or exemplary behaviour. According to March and Olsen (1995) social actors follow prescriptions of what is socially defined as normal, true, right or good, without, or in spite of, calculation of consequences and expected utility. Isomorphism derives from institutional arrangements that link roles/identities, accounts of situations, resources and prescriptive rules and practices. Internalized principles and prescriptions have to be balanced by the calculated expected utility and constrains that a specific institutional frame determines.

The category of institutional isomorphism has been usually applied to the analysis of organizational uniformity in specific organizational fields, which are, in the words of March and Olsen, recognised areas of institutional life. Other contributions have tried to enlarge their perspective. With the increase of interdependence due to globalization, the idea of isomorphism has been used to underline a growing phenomenon of policy convergence from one political setting to another. Dolowitz and Marsh note that a significant body of literature in political science and in international studies refers to terms as lesson-drawing, policy convergence, policy diffusion and policy transfer in order to indicate "the process by which knowledge about policies, administrative arrangements, institutions and ideas in one political system (past or present) is used in the development of policies, administrative arrangements, institutions and ideas in another political system" (Dolowitz & Marsh, 2000, p. 5). The reasons why the policy transfer occurs run, in a simplified and heuristic model, from voluntary adoption—a rational response starting from the "dissatisfaction with the status quo"—to direct imposition. In addition to this, what is transferred is also variable, and Rose (1993, p. 30) set at the one extreme "direct coping" which regards full "lesson-drawing" of a programme or a policy from one jurisdiction to another, and, at the other extreme "inspiration," with reference to mimetic mechanisms. Although many factors influence the process of institutional adaptation, a crucial point in the literature on isomorphism is that the institutional arena contains a number of exogenous pressures that influence the structure and behaviour of organizations, based on socio-cultural norms or interdependency (Dacin, 1997; Tucker, 1983).

It can be confirmed that both classical studies on organizational isomorphism and analyses on policy transfer underestimate the role of technological variables. Although, as La Porte et al. (2002) put it, "growing homogeneity among modern organizations emerges as a function of unprecedented amounts of newly available information about other organizations in the environment" (p. 435), the role of information and communication technologies remains quite implicit.

The reasons behind such a scarce attention to the potentials of technology in terms of creating institutional uniformity probably rests on a fundamental distinction between institutional and technical sources of organizational practices: technology is said to be the source of variation while institutions are considered responsible for conformity and of predictability of behaviour (Meyer & Rowan, 1977, in Leblebici et al., 1991). Yet it is difficult to defend this statement after the spread of new technologies, when the definition of a computer-based infrastructure was used to establish a true code able to regulate the various actors' behaviour.

For instance, Jane Fountain, by showing how the intensive use of new technologies is embedded in an institutional context with its social, cultural and legal features, observes the collision between traditional practices and traditions with technological innovation. In her contribution *Building the Virtual State: Information Technology and Institutional Change* (2001) she focuses on the role of digital policies, and in particular of e-government, in offering new incentives and constrains to change governments. According to Fountain, it is to be assumed that within a government agency there is a dominant homogenous culture, and the strategic use of new technologies can constitute an instrument for constructing such discourse (Yang, 2003). If her main finding is that institutions are the most important factors in explaining how information technology is being adopted and used in government, at the same time it can be argued that new technologies themselves are a relevant element for the shaping of institutions.

Also, Lawrence Lessig (1999) considers the effects of the introduction of new technologies in influencing institutional settings. What he calls "the code," the complex system of software and hardware instructions defining the Internet rules, seems to show a constitutive potential. Indeed, the idea of the code tends to consider the effects of new technologies not only for creating convergence in the public administrations'

behavior, but also for a more incisive cultural homogenization. A computer code may regulate conducts more than a legal corpus does, as it intervenes at a deeper level, the level of ideas and their diffusion (Lessig, 2001).

THE SEARCH FOR UNIFORMITY: THE CASE OF EUROPEAN UNION

The European Union can be considered one of the most relevant examples of the strategic use of new technologies for institution building (Baptista, 2005; Overeem, Witters, & Peristeras, 2007).

An important body of literature has already put attention on the process of institutional isomorphism developed in Europe. The homogenization pressures in the Old Continent seem to produce "a process of (a) construction, (b) diffusion and (c) institutionalization of formal and informal rules, procedures, policy paradigms, styles, ways of doing things and shared beliefs and norms which are first defined and consolidated in the making of EU decisions and then incorporated in the logic of domestic discourses, identities, political structures and public policies" (Radaelli, 1997, p. 4). Such phenomenon has justified the introduction of categories such as that of "Europeanization" in regard to the transformation of structures and practices of MS's institutions, as activated by the process of European integration. Even if this process is producing different influences in the different contexts, there is no doubt that the various member states' administrations are currently undergoing a transformation stage where they share some common directions (Börzel & Risse, 1999; Heritier, 2001; Riekmann, Puntsher, & Latzer, 2006). For instance, according to Siedentopf and Speer (2003), in Europe an evolving process of increasing convergence between national administrative legal orders and administrative practices of member States is realizing "Convergence is influenced by several driving forces, such as economic pressures from individuals and firms, regular and continuous contacts between public officials of member States, and finally and especially, the jurisprudence of the European Court of Justice" (p. 13).

Literature on institutional isomorphism in Europe seems to underestimate technological variables. Yet digital policies held a central role for the EU development due to its meaningful potential for transformation. It has to be also underlined the priority of the strategic use of new technology for economic purposes. Despite of a poor definition for a comprehensive project, the heavy use of new technologies and the development of a supra-national computer-based architecture have become, indeed, much valued tools aiming at EU major targets, particularly in the area of economic development. For example, the recent *i2010 E-government Action Plan* (European Commission, 2006a) clearly states that those countries with a higher e-government development degree are also at the top level in the main economy indicators: "This strong link between national competitiveness, innovation strength and the quality of public administrations means that in the global economy a better government is a competitive must" (ibidem, p. 3). There are in fact several evidences supporting the positive impact of new technologies, in both the short and long term, on innovation and growth within the public sector. The crucial point here is that digital policies create not only more opportunity of economic development, but also more integrity and standardization in the European administration. Securing competition in the European market required the definition of an integrated service system able to overcome EU's fragmentation and differences, in regard to both the increasing number of MSs and local governments' accountability following the introduction of the subsidiarity principle. With the EU undergoing an expansion process, thus becoming more and more diversified, it was necessary to implement effective public services systems covering the entire European territory and ensuring the full mobility of citizen and goods (p. 3).

Through new ICTs standardization in organisations, infrastructure, administrative procedures, and also in semantic codes, legislation is encouraged in order to produce interaction and integration between European public administrations. If establishing behaviour standards could be considered an important part of the current process of globalization, in different areas, from environmental issues to financial matters (Cassese, 2006), now the computer code seems in charge of deciding the norms to be applied at a global level to cooperation, harmonization and standardization procedures.

A recent report presented by the European Commission to the European Parliament (European Commission, 2006b) also focuses on interoperability as the main goal of the administrative structure. The modernization of public administration is thus finalized to the establishment of a common market and an effective interaction between citizens and companies: "The single market relies on modern and efficient public administrations which facilitate the mobility and seamless interaction of citizens and businesses" (European Commission, 2006b, p. 2). The same approach applies to administrative structures integration in various government levels. The document states that "to be affordable and effective, implementation of the infrastructure required for the delivery of pan-European eGovernment services will have to be guided by an overall conceptual architecture, based on standards" (European Commission, 2006b, p. 9). Therefore, interoperability goal is described as intertwined with the strengthening of economic competition and the overcoming of any obstacle to the establishment of a common market (European Commission, 2006a). Defined as "the key enabler for the delivery of e-government services across national and organizational boundaries," such interoperability systems are regarded as the right tools to ensure the mobility of businesses and individuals, a larger interaction among the

stake-holders, and an effective administrative cooperation (European Commission, 2006b).

Such interoperability can only be actualized by intervening on three different elements, respectively related to the organizational, technical, and semantic dimensions:

- Organizational interoperability is about being able to identify those players and organizational processes involved in the delivery of a specific e-government service and achieving agreement among them on how to structure their interactions, that is, defining their "business interfaces."
- Technical interoperability is about knitting together IT systems and software, defining and using open interfaces, standards and protocols in order to build reliable, effective and efficient information systems.
- Semantic interoperability is about ensuring that the meaning of the information exchanged is not lost in the process, that it is retained and understood by the people, applications and institutions involved (European Commission, 2006b, p. 6).

Finally, the realization of interoperable systems leads to a "one-stop government:" a single point of access to electronic services and information offered by different public authorities. Online one-stop government requires that all public authorities will be interconnected and that the customer (citizen, private enterprise or other public administration) will be able to access public services via a single point, even if these services are provided by different public authorities or private service providers.

The definition of a technological architecture constitutes the premise for further developing of the European Union. If the lack of a cultural integration is among the main causes of the delay of the European experiment—of which a clear example is represented by the failure of the process of treaty ratification—ICTs face the challenge to keep together organizations at different levels of government and to offer sharing cognitive tools for the citizen-user as well. Therefore, the strategic use of new technology can be easily interpreted as a process of institutional building.

FUTURE TRENDS

Due to its more defined characteristics, ICTs policies are doubtless one of the most effective tools for establishing common institutional standards.

The existence of a wider process leading toward increased convergence between national administrative practices can be questioned (Harlow, 2005). In an era in which globalization and international homogenization of regulatory regimes are gradually undermining national institutional distinctions, public officials are increasingly inclined to borrow policy solutions and adopt organizational structures experimented in other countries. Moreover, the need—and imperative—of cooperation in a global market has led to the adoption of standardization procedures in the administrative process (La Porte et al., 2002). Although it remains uncertain that pressures are strong enough to create a uniform model of public administration, ICTs will become a crucial lever toward a greater integration within administrative agencies by providing standards to public actors' behaviour. So, while digital architecture encourages the creation of new networks and executive layers, it seems also to introduce new forms of homogenization of political and administrative action. An example could be taken from the modalities of creation of public Web sites, as governments move toward standardization among various agency portals (West, 2007). By allowing easy Web orientation, the standardization of online environments may increase administrative interoperability.

This discourse can be also enlarged from administrations to other socio-political actors. For instance, parties, interest groups and social movements' organizational features and policy impacts will converge after the impact of the Internet (Chadwick, 2007). Briefly, we can formulate the hypothesis that new technologies encourage organizational hybridity, by noting that a process of transplantation and adaptation of digital networks repertoires previously considered typical of social movements is occurring. At the same time, due to this process, new forms of organization are emerging. By providing more tools of political action, the Internet probably will lead to a convergence in contemporary political mobilization.

CONCLUSION

New technologies are a relevant and underestimated instrument for producing institutional convergence. The concept of institutional isomorphism, as it emerges from the literature on institutional studies, has to be rearticulated in order to consider the role of technological tools, starting from the assumption that ICTs may in several ways conduct to standardization within and between institutions. Indeed, especially supranational governments are betting on their transformative potential for creating integration and interoperability between different levels of government.

Is the homogenization a specific European product or a further puzzle piece of the broader globalization process? The idea of knitting together the world's economies into a single unit seems to suggest the necessity for nation states to adopt a shared administrative model. In addition to this, studies on what it is happening in other nonwestern countries seem to support the thesis of uniformity. For instance, if the European Union tries to reach a difficult equilibrium between the aspiration of administrative decentralization and the need for uniformity, e-government initiatives have

been recently interpreted in China as "vehicles intended to support economic development through an increasingly transparent and decentralized administration while at the same time providing the central government the information and ability to efficiently monitor" (Ma, Chung, & Thorson, 2005, p. 20). Further research may be dedicated to evaluate differences and similarities of various geopolitical areas. Anyway, the strategic use of new technology can begin to be intended as one of the most important variables of institutional isomorphism.

REFERENCES

Baptista, M. (2005). E-government and state reform: Policy dilemmas for Europe. *Electronic Journal of e-Government*, *3*(4), 167-174.

Börzel, T., & Risse, T. (2000). When Europe hits home: Europeanization and domestic change. *European Integration Online Papers*, *4*(15). Retrieved May 29, 2008, from http://eiop.or.at/

Cassese, S. (2006). *Oltre lo Stato.* Roma-Bari: Laterza.

Chadwick, A. (2007). Digital network repertoires and organizational hybridity. *Political Communication, 24,* 283-301.

Dacin, M. T. (1997). Isomorphism in context: The power and prescription of institutional norms. *The Academy of Management Journal*, *40*(1), 46-81.

DiMaggio, J. P., & Powell, P. P. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*, *48*(2), 147-160.

Dolowitz, D. P., & Marsh, D. (2000). Learning from abroad: The role of policy transfer in contemporary policy-making. *Governance: An International Journal of Policy and Administration*, *13*(1), 5-24.

European Commission. (2006a, April 25). *i2010 e-government action plan: Accelerating e-government in Europe for the benefit of all*. Brussels.

European Commission. (2006b, February 13). *Interoperability for Pan-European e-government services.* Brussels.

Fountain, J. E. (2001). *Building the virtual state: Information technology and institutional change*. Washington, DC: Brookings Institution.

Guthrie, D. (1999). A sociological perspective on the use of technology: The adoption of Internet technology in U.S. organizations. *Sociological Perspectives, 42*(4), 583-603.

Harlow, C. (2005). Law and public administration: Convergence and symbiosis. *International Review of Administrative Sciences, 7*(2), 279-294.

Heritier, A. (2001). Differential Europe: National administrative responses to community policy. In M. Green, M. Cowles, J. Caporaso, & T. Risse (Eds.), *Transforming Europe. Europeanization and domestic change*. Ithaca, NY: Cornell University Press.

La Porte, et al. (2002). Democracy and bureaucracy in the age of the Web: Empirical findings and theoretical speculations. *Administration & Society*, *34*(4), 411-446.

Leblebici, H. (1991). Institutional change and the transformation of interorganizational fields: An organizational history of the U.S. radio broadcasting industry. *Administrative Science Quarterly*, *36*(3), 333-363.

Lessig, L. (1999). *Code and other laws of cyberspace.* New York: Basic Books.

Lessig, L. (2001). *The future of Ideas*. New York: Vintage.

Ma, L., Chung, J., & Thorson, S. (2005). E-government in China: Bringing economic development through administrative reform. *Government Information Quarterly*, *22*(1), 20-37.

March, J. G., & Olsen, J. P. (1995). *Democratic governance*. New York: Free Press.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*, 340-363.

Mignerat, M., & Rivard, S. (2005). Positioning the institutional perspective in information technology research. In *Proceedings of the Asac Congress*, Toronto.

Olsen, J. (2002). Toward an administrative European space?. *Arena Working Papers*, (26).

Orlikowski, W. J., & Barley, S. R. (2001). Technology and institutions: What can research on information technology and research on organizations learn from each other? *MIS Quarterly*, *25*(2), 145-165.

Overeem, A., Witters, J., & Peristeras, V. (2007). An interoperability framework for Pan-European e-government services (PEGS). In *Proceedings of the 40th Hawaii International Conference on System Sciences*.

Radaelli, C. (1997). Policy transfer in the European Union: Institutional isomorphism as a source of legitimacy. *Jean Monnet Working Papers in Comparative and International*

Politics, 10. Retrieved May 29, 3008, from http://www.fscpo.unict.it/EuroMed/jmwp10.htm

Riekmann, S., Puntsher, M. M., & Latzer, M. (2006). *The state of Europe: Transformation of statehood from a European perspective*. Chicago University Press.

Rose, R. (1993). *Lesson drawing in public policy*. Chatam: Chatam House.

Siedentopf, H., & Speer, B. (2003). The European administrative space from a German administrative science perspective. *International Review of Administrative Science, 69*(1), 9-28.

Tolbert, P. S., & Zucker, L. G. (1983). Institutional sources of change in the formal structure of organizations: The diffusion of civil service reforms, 1880-1935. *Administrative Science Quarterly, 23*, 22-39.

Weber, M. (1905). *Die Protestantische Ethik und der Geist des Kapitalismus*. Tubinga: Mohr Verlag. American translation: (1958). *The protestant ethic and the spirit of capitalism*. New York: Scribner.

West, D. (2007). *Global e-government*. Providence, RI: Brown University.

Yang, K. (2003). Neoinstitutionalism and e-government. *Social Science Computer Review, 21*(4), 432-442.

KEY TERMS

Code: The digital architecture which regulates the cyberspace, the complex system of software and hardware instructions defining Internet rules

E-Government: The application of new information and communication technology for the restructuring of public administration and the renewal of the relationship between public institutions and citizen-users

European Administrative Convergence: The process for which administrations become more similar and close to a common European model

ICTs: Acronym for information and communications technologies. It is a general term that describes any technology that helps to produce, manipulate, store, communicate, or disseminate information

Interoperability: Ability of ICT systems and business processes to exchange data and enable the sharing of information and knowledge

Isomorphism: A constraining process that forces one unit in a population to resemble other units facing the same set of environmental conditions. It stimulates an evolutionary path from diversity to homogeneity, as the result of imitation among organizations or independent development under similar constraints

Organizational Hybridity: (a) a process of hybridization based on the selective transplantation and adaptation of digital network repertoires previously considered typical of social movements (b) the emerging of new organizational forms that exist only in hybrid form and that could not function in the ways that they do without the Internet and the complex spatial and temporal interactions it facilitates.

Standardization: The process of creating uniformity through established standards

Category: IT Education

Instructional Support for Distance Education

Bernhard Ertl
Universität der Bundeswehr München, Germany

INTRODUCTION

During the late '90s, distance education and e-learning were believed to be able to solve almost every problem associated with the further qualification of employees in organizations. Distance education was credited with saving costs for companies, by reducing time and expenses for traveling and with flexible time management. Consequently, many companies started programs for distance education. However, after this initial euphoria, several organizations experienced problems with their programs (e.g., Haben, 2002). The costs for distance education courses exploded, employees refused the new style of learning, and the general question arose as to the effectiveness of distance education (see, e.g., Bernard et al., 2004). Looking at the range of distance education courses at this time, one could see that they used a broad variety of technologies to deliver learning contents to the learners, for example, videos, Web pages, dedicate software for learning, Weblogs, wikis, collaboration tools, videoconferencing, chat, and discussion boards. However, in contrast to the variety of technologies available, the instructional design of these courses was elementary and traditional (see Ertl, Winkler, & Mandl, 2007). Many courses offered recorded classroom lectures and streamed them to participants, or they just presented texts or slides in the style of a book. Such courses experienced a lack in acceptance and thus several efforts of distance education failed because the instructional design of these courses was not able to take advantage of the innovative technologies.

BACKGROUND

To take advantage of the emerging technologies, a new philosophy of learning and teaching is necessary. Moderate constructivist approaches focus on several activities of learners that are necessary for the successful implementation of distance education courses. They build on learners' active knowledge construction and postulate that learning requires learners' active participation. This is in contrast to traditional approaches, which set learners in a receptive role. According to constructivist approaches, learning is mediated by learners' individual prior knowledge, their motivation, and other individual learning prerequisites. Reinmann-Rothmeier and Mandl (2001) describe several key-elements for construction of knowledge according to this philosophy (see also Ertl, Winkler, & Mandl, 2007). They state that a learning process is:

- **Active**, because only active involvement enables learning.
- **Self-directed** and learners have to take active control and responsibility for their learning activities.
- **Constructivist**, which means that learners have to embed new knowledge in their existing knowledge structures.
- **Social** and knowledge acquisition requires a social context.
- **Situated** because knowledge acquisition happens in a specific context and is linked to this context.
- **Emotional**; the emotional component is particularly important for the motivation of the learners.

Besides these constructivist aspects, learning environments require a certain amount of instruction (Ertl et al., 2007; Kirschner, Sweller, & Clark, 2006; Reinmann-Rothmeier & Mandl, 2001). Consequently, learning environments need to find a balance between construction and instruction. This balance can be realized by the design of problem-oriented learning environments (see Mandl, Gräsel, & Fischer, 1998) and case-based learning scenarios (Kolodner et al., 2003). Such learning environments can benefit from new technologies; they can provide tools for supporting the active construction of knowledge (Roschelle & Teasley, 1995), provide an authentic situational context by the display of video cases (CTGV, 1997), enable the social context for spatially-divided learners (Mandl, Ertl, & Kopp, 2006), and motivate learners by the provision of gimmicks and animations (Mayer, Hegarty, & Mayer, 2005). However, none of these benefits are caused by the technology itself—they are introduced by the instructional design of the learning environment including the use of the new technologies.

This chapter focuses on two particular aspects how the instructional design can apply new technologies for the improvement of learning environments: on collaboration-specific methods structuring learners' collaboration, and on content-specific methods that are supporting learners' active construction of knowledge.

COLLABORATION-SPECIFIC METHODS

Methods for facilitating learners' collaboration may be associated with several tools, particularly software products that aim at enabling collaborative work or at supporting particular collaborative tasks (e.g., collaborative drawing or text editing). These tools can support collaboration between learning partners, yet the fact remains that collaborative skills often do not come naturally to the individual learner and must therefore be acquired (see Salomon & Globerson, 1989). Instructional approaches focusing on the improvement of collaboration often refer to methods such as *scripted cooperation* (O'Donnell & King, 1999). Such scripts sequence learners' work on the task. Furthermore, they may provide roles for the learners and encourage them to apply beneficial strategies for solving a task.

As an example, the MURDER-script (Dansereau et al., 1979; O'Donnell & Dansereau, 1992) is comprised of several different aspects, and will therefore demonstrate the potential elements of scripts and their combination. This script relates to a collaboration process in which learners work collaboratively on text comprehension. It divides the collaborative learning process into six phases that focus on individual as well as on collaborative activities. The first phase relates to learners personal motivation for the task ahead (*M*ood). The second phase focuses on individual text comprehension (*U*nderstand). In the third phase, one partner repeats contents of the text from his memory (*R*epeat) while the other partners try to find difficulties and give feedback (4th phase; *D*etect). In the following, learners reflect and elaborate about the content to link the learning material with their prior knowledge (5th phase; *E*laborate). Finally, they check their work against the original text material (*R*eview, 6th phase). Learners may repeat these six phases for several text paragraphs and for each cycle, a different learning partner takes the role to repeat the text contents.

Technologies can integrate such scripts into collaborative learning environments. They may structure the collaboration process or the proceeding in the work on the task. Baker and Lund (1997), for example, report a script, which specifically directed the collaboration process. Their learning environment provided a shared graphics editor for working on a collaborative product and the instructional design added several *speech act buttons* to this editor. Each time a learner had made changes to the collaborative product, the learning environment required both partners to agree on these changes before continuing. They were required to demonstrate this by pressing the respective speech act buttons. The intention of this mechanism was that both learning partners increased their grounding (Dillenbourg & Traum, 2006) and their collaborative commitment to the joint product (Baker & Lund, 1997).

Ertl, Reiserer, and Mandl (2005) showed a different example for scripting in distance education using a videoconferencing scenario. The aim of this script was to facilitate learners during the task of collaborative teaching. This script structured the collaborative proceeding on the task, the roles of the learners, and the application of beneficial strategies regarding the collaborative negotiation. Therefore, the script assigned two different roles to the learners, the role of a teacher, and the role of a learner. Furthermore, it divided the collaboration process into four different phases. Learners worked with a shared application in this scenario, and this application offered learners a space for written externalizations. Furthermore, the application was pre-structured with instructional elements that guided learners according to the script. In the first phase, the participant in the teacher role explained the text material while the partner in the learner role asked comprehension questions. In the second phase, the learner rehearsed the concepts acquired and fixed important aspects in the shared application. The teacher supported the partner and clarified misinterpretations. In the third phase, both partners reflected individually, and in the fourth phase, they discussed the learning material. In this phase, the learner also noted important aspects in the shared application. After these four phases, learners switched their roles and continued with another text.

Results of the study showed that the learning environment with the script was able to facilitate learners' negotiation with theoretical concepts during collaboration. With respect to the individuals' learning outcomes, the script particularly facilitated learners in the learner role. They acquired more knowledge during collaboration than learners without a script (see Ertl, Reiserer, & Mandl, 2005). Other studies also report beneficial effects of scripts in distance education courses. These were related to the learning processes (Baker & Lund,

Table 1. Taxonomy of support methods with different goals

Goal of support	Collaboration-specific methods	Content-specific methods
Improving collaborative processes	Scripts	
Understanding impact factors		Simulations
Understanding structures		Templates
Understanding relations		Conceptualization tools

1997; Weinberger, Ertl, Fischer & Mandl, 2005) as well as to the individuals' outcomes (Rummel & Spada, 2005). Scripts may improve general processes of collaboration (Baker & Lund, 1997), lead to a more homogeneous work on the task (see Weinberger, 2003) and to the acquisition of beneficial collaboration strategies (Rummel & Spada, 2005).

CONTENT-SPECIFIC METHODS

Content-specific methods rely on particular affordances of the course's content domain. They may provide domain categories or ontologies for the learners (see, e.g., Ertl, Fischer, & Mandl, 2006), facilitate the visualization of conceptual relations (see Fischer, Bruhn, Gräsel & Mandl, 2002), or provide simulations or visualizations which help learners to understand particular mechanisms (see Roschelle & Teasley, 1995). Content-specific methods aim for support at a conceptual level and try to facilitate learners' understanding of particular conceptual aspects, relations, or mechanisms of the content domain (see Table 1).

Content-specific methods often rely on a particular representation of important content structures. Zhang and Norman (1994) postulate that this representation of content has an influence on learners' ability to deal with the content. If a method changes the representation of the content then it might be that learners perceive this content in a different manner. This may facilitate as well as impede learning—depending on the match between the representation and the learners' cognitive structure (see Zhang & Norman, 1994). This means that the content structure remains the same (it is isomorph) but the way in which it is presented changes. A rather simple example for this mechanism would be to provide a diagrammatic representation instead of a textual description (see, e.g., Larkin & Simon, 1987). The representation can make important task characteristics salient and function as a representational guide for the learners (see Suthers & Hundhausen, 2003). There is a broad variety of methods and tools for this kind of facilitation (see Löhner & van Joolingen, 2001). They offer different amounts of facilitation to the learners, and they vary with respect to the degrees of freedom the learners have when working with them.

In distance education, one has to distinguish between tools, which enable content-specific facilitation, and the instructional design, which applies the tools to a particular context and provides the facilitation. Powerful tools may offer many possibilities and much freedom to the learners. However, this may be too complex for the learners, who may not have the cognitive ability to apply it correctly and thereby suffer from cognitive overload (see Sweller, van Merrienboer, & Paas, 1998). Consequently, it may be too complex for beneficial activities (see Dobson, 1999) and negate the potential facilitation effect. The instructional design of a distance education course should therefore consider the skills and the prior knowledge of the learners (see Mandl, Ertl, & Kopp, 2006; Shapiro, 2004) and aim for a balance between learners' experiences and the demands of the tools.

In the following, we will describe briefly different forms of content-specific support:

- Tools for simulations (see Roschelle & Teasley, 1995) allow learners to *simulate* scientific processes. The instructional design of these tools is such that the learner can simulate a process according to various parameters. The particular tools for simulations are modeled specifically for this one purpose and might also include visualizations or animations of these processes. Learners can modify the parameters of the simulation and observe the results of this change in the simulation. Thus, simulations aim at understanding the influence of particular factors on a whole (simulated) system.
- Templates are different from simulations in that they *pre-structure* a content domain (see Brooks & Dansereau, 1983; Ertl, Fischer, & Mandl, 2006; Suthers & Hundhausen, 2003). In this case, the tool provides the features to create templates, and the instructional design specifies the contents of these templates. It provides categories that are particularly important for content-specific negotiation and often uses tables for their representation. These tables provide empty spaces for the learners which help them to focus on the important categories. However, learners cannot change the structure of the tables and model new relations. Consequently, templates aim at internalizing the structure of a content area.
- Conceptualization tools allow the visualization of connections between different concepts within a subject matter. They enable learners to illustrate connections between concepts and theories by creating a mind map or a similar diagram. The tool provides the concepts and various types of connecting lines that are then sorted and put together to demonstrate the connections. Learners may thereby create their own representation, but the process is supported by the pre-existing elements used (see Fischer et al., 2002; Suthers & Hundhausen, 2003). Consequently, conceptualization tools are intended to facilitate a deeper understanding of the relationships within a particular content area.

Ertl, Reiserer, and Mandl (2005) present an example for a content-specific method in a distance education course in the style of a template. This template aimed at facilitating learners' learning of text material. It focused learners on important aspects of theories, particularly on the categories of theory concepts, evidence, and personal elaborations with respect to consequences and learners' individual opinion. They used a shared application for providing the template to the learners. The instructional design provided a table

with four cells headed by the respective category names. Furthermore, it anchored the rather broad categories by different prompts in each table cell.

Results of the study show that this template provided several benefits for the learners. They reached a higher score in the category of evidence, and they provided more personal elaborations (see Ertl, Reiserer, & Mandl, 2005). Thus, the template was able to direct the learning process not only to the memorization of facts, but also on the personal contextualization of these facts. Moreover, several other studies have shown beneficial effects of content-specific methods in the context of distance education (see Ertl et al., 2006; Fischer et al., 2002; Roschelle & Teasley, 1995; Suthers & Hundhausen, 2003). Roschelle and Teasley (1995) report beneficial effects of simulations for transactive discourse and knowledge co-construction. Ertl et al. (2006) present a template, which provided benefits for learners' collaborative learning process as well as for their individual knowledge acquisition. Suthers and Hundhausen (2003) reported that a template had facilitated the learners to draw relations between theoretical concepts and evidence. Furthermore, Fischer, Bruhn, Gräsel, and Mandl (2000) found that conceptualization tools homogenized collaborative learning processes.

FUTURE TRENDS

Studies which compared learning environments with a sound instructional design and traditional courses report an increased quality of education, a more active role of the learners, and more motivated learners if they were working in the well-designed learning environment (see, e.g., Hiltz, 1997; Lehtinen, 2003). In contrast, studies which just compared different technologies were hardly able to find any beneficial effects of the technologies for learning (e.g., Clark, 1994; Salomon, 1984; Schweizer, Pächter, & Weidenmann, 2001; Storck & Sproull, 1995). This means that distance education courses can provide "powerful learning environments". However, this power comes from the collaborative setting and from their instructional design rather than from technology. The future of e-learning will evoke some kind of consolidation in the field. Distance education courses will be more and more subject to evaluation. This will disclose how far a particular course or technology can provide benefits for the learners.

CONCLUSION

This chapter dealt with instructional support for distance education courses. This is of particular importance for distance education because many distance education courses have a fairly simple instructional design. They provide either lectures without any opportunity for learners' individual knowledge construction or merely offer resources without any guidance for the learners. Courses for distance education should use well balanced aspects of construction and instruction to provide benefits for the learners. The instructional design of courses may be featured by several methods which apply information technology. Collaboration-specific methods structure collaboration tools to optimize collaborative learning processes. Content-specific methods use tools to facilitate learners' collaborative knowledge construction on a conceptual level. Both can enhance the instructional design and the outcomes of distance education courses.

ACKNOWLEDGMENT

This work was funded by Deutsche Forschungsgemeinschaft (German science foundation, DFG), Grant Nos. MA 978/13-1, MA 978/13-3 and MA 978/13-4.

REFERENCES

Baker, M., & Lund, K. (1997). Promoting reflective interactions in a CSCL environment. *Journal of Computer Assisted Learning, 13*(3), 175-193.

Bernard, R. M., Abrami, P. C., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Wallet, P. A., Fiset, M., & Huang, B. (2004). How does distance education compare with classroom instruction? A meta-analysis of the empirical literature. *Review of Educational Research, 74*(3), 379-439.

Brooks, L. W., & Dansereau, D. F. (1983). Effects of structural schema training and text organization on expository prose processing. *Journal of Educational Psychology, 75*(6), 811-820.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development, 42*(2), 21-29.

Cognition and Technology Group at Vanderbilt. (1997). *The Jasper Project: Lessons in curriculum, instruction, assessment, and professional development*. Mahwah, NJ: Erlbaum.

Dansereau, D. F., Collins, K. W., McDonald, B. A., Holley, C. D., Garland, J. C., Diekhoff, G., et al. (1979). Development and evaluation of a learning strategy training program. *Journal of Educational Psychology, 71*(1), 64-73.

Dillenbourg, P., & Traum, D. (2006). Sharing solutions: persistence and grounding in multimodal collaborative problem solving. *Journal of the Learning Sciences, 15*(1), 121-151.

Dobson, M. (1999). Information enforcement and learning with interactive graphical systems. *Learning and Instruction, 9*(4), 365-390.

Ertl, B., Fischer, F., & Mandl, H. (2006). Conceptual and socio-cognitive support for collaborative learning in videoconferencing environments. *Computers & Education, 47*(3), 298-315

Ertl, B., Reiserer, M., & Mandl, H. (2005). Fostering collaborative learning in videoconferencing: the influence of content schemes and collaboration scripts on collaboration outcomes and individual learning outcomes. *Education, Communication & Information, 5*(2), 147-166.

Ertl, B., Winkler, K., & Mandl, H. (2007). E-learning: Trends and future development. In F. M. M. Neto, & F. V. Brasileiro (Eds.), *Advances in computer-supported learning* (pp. 122-144). Hershey, PA: Information Science Publishing.

Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2000). Kooperatives Lernen mit Videokonferenzen: Gemeinsame Wissenskonstruktion und individueller Lernerfolg [Cooperative learning in videoconferencing. Collaborative knowledge construction and individual learning outcomes]. *Kognitionswissenschaft, 9*, 5-16.

Fischer, F., Bruhn, J., Gräsel, C., & Mandl, H. (2002). Fostering collaborative knowledge construction with visualization tools. *Learning and Instruction, 12*, 213-232.

Haben, M. (2002). E-Learning in large German companies: Most of the concepts are not effective. *Computerwoche, 30*(22), 12-16.

Hiltz, S. R. (1997). Impacts of college-level courses via asynchronous learning networks: Some preliminary results. *Journal of Asynchronous Learning Networks, 1*(2), 1-19.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., et al. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting learning by Design™ into practice. *The Journal of the Learning Sciences, 12*(4), 495-547.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65-99.

Lehtinen, E. (2003). Computer-supported collaborative learning: An approach to powerful learning environments. In E. De Corte, L. Verschaffel, N. Entwistle, & J. J. G. v. Merrienboer (Eds.), *Powerful learning environments: Unravelling basic components and dimensions* (pp. 35-53). Amsterdam: Elsevier.

Löhner, S., & van Joolingen, W. (2001). Representations for model construction in collaborative inquiry environments. In P. Dillenbourg, A. Eurelings, & K. Hakkarainen (Eds.), *Proceedings of the First European Conference on Computer-Supported Collaborative Learning (euroCSCL)* (pp. 577-584). Maastricht: McLuhan Institute.

Mandl, H., Ertl, B., & Kopp, B. (2006). Computer support for collaborative learning environments. In L. Verschaffel, F. Dochy, M. Boekaerts, & S. Vosniadou (Eds.), *Instructional psychology: Past, present and future trends. Sixteen essays in honor of Erik De Corte* (pp. 223-237). Amsterdam: Elsevier.

Mandl, H., Gräsel, C., & Fischer, F. (1998). Facilitating problem-orientated learning: The role of strategy modeling by experts. In W. Perring, & A. Grob (Eds.), *Control of human behavior, mental processes and awareness. Essays in honor of the 60th birthday of August Flammer* (pp. 165-182). Mahwah, NY: Erlbaum.

Mayer, R. E., Hegarty, M., & Mayer, S. (2005, April). *Does animation improve learning?* Paper presented at the 86th Conference of the American Educational Research Association (AERA), Montréal.

O'Donnell, A. M., & Dansereau, D. F. (1992). Scripted cooperation in student dyads: A method for analyzing and enhancing academic learning and performance. In R. Hertz-Lazarowitz, & N. Miller (Eds.), *Interactions in cooperative groups. The theoretical anatomy of group learning* (pp. 120-141). New York, NY: Cambridge University Press.

O'Donnell, A. M., & King, A. (Eds.). (1999). *Cognitive perspectives on peer learning*. Mahwah, NJ: Erlbaum.

Reinmann-Rothmeier, G., & Mandl, H. (2001). Unterrichten und Lernumgebungen gestalten [Teaching and the instructional design of learning environments]. In A. Krapp, & B. Weidenmann (Eds.), *Pädagogische Psychologie* (pp. 601-646). Weinheim: Beltz.

Roschelle, J., & Teasley, S. D. (1995). The construction of shared knowledge in collaborative problem solving. In C. O'Malley (Ed.), *Computer supported collaborative learning* (pp. 69-97). Berlin: Springer.

Rummel, N., & Spada, H. (2005). Learning to collaborate: An instructional approach to promoting collaborative problem solving in computer-mediated settings. *The Journal of the Learning Sciences, 14*(2), 201-241.

Salomon, G. (1984). Television is "easy" and print is "tough": The differential investment of mental effort in learning as a

function of perceptions and attributions. *Journal of Educational Psychology, 76*(4), 647-658.

Salomon, G., & Globerson, T. (1989). When teams do not function the way they ought to. *International Journal of Educational Research, 13*(1), 89-99.

Schweizer, K., Pächter, M., & Weidenmann, B. (2001). A field study on distance education and communication: Experiences of a virtual tutor. *Journal of Computer Mediated Communication, 6*(2).

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*(1), 159-189.

Storck, J., & Sproull, L. (1995). Through a glass darkly: What do people learn in videoconferences? *Human Communication Research, 22*(2), 197-219.

Suthers, D. D., & Hundhausen, C. D. (2003). An experimental study of the effects of representational guidance on collaborative learning processes. *The Journal of the Learning Sciences, 12*(2), 183-218.

Sweller, J., van Merrienboër, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251-296.

Weinberger, A. (2003). *Scripts for computer-supported collaborative learning. [Dissertation, Ludwig-Maximilian-University Munich]*. Retrieved from http://edoc.ub.uni-muenchen.de/archive/00001120/01/Weinberger_Armin.pdf

Weinberger, A., Ertl, B., Fischer, F., & Mandl, H. (2005). Epistemic and social scripts in computer-supported collaborative learning. *Instructional Science, 33*(1), 1-30.

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*(1), 87-122.

KEY TERMS

Collaboration: Tightly working together with a strong commitment of collaboration partners.

Collaborative Knowledge Construction: Learners' joint activities to acquire or create new knowledge.

Content Scheme: Tabular representation of domain-specific structure to facilitate learners.

Instructional Design: The didactical rationale for a learning scenario which includes instructional elements as well as the application of tools.

Knowledge Construction: Learners' work with their knowledge in a way that they link their new knowledge to their existing knowledge base in stead of memorizing facts.

Learning Environment: Learners' context in distance education courses that is comprised of instructional, social, and technical aspects.

Powerful Learning Environment: A learning environment which includes instructional elements that evoke learners' active construction of knowledge.

Script: Specification of learning processes which contains procedural aspects, the assignment of roles, and the evocation of beneficial cognitive activities.

Category: Software & Systems Design

Integrating Domain Analysis into Formal Specifications

Laura Felice
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

Daniel Riesco
Universidad Nacional de San Luis, Argentina

INTRODUCTION

Reusability is widely suggested as a key to improve software development productivity. It has been further argued that reuse at domain level can significantly increase reuse at later stages of development. The development of a particular system that exploits previously accumulated domain knowledge can be the source for new insights about the domain that adds to or refines. The classification of similar problems grouped by domains is the key to find reusable solutions that belong to similar problems. This work deals with the integration of a reusability model into a formal method in order to enhance the benefits of reusability at the domain and design levels. Working with formal methods, software reuse problems such as the detection of inconsistencies in component integration can be revealed in early stages of development.

The rigorous approach to industrial software engineering (RAISE) (George et al., 1995) formal method was originally designed to be applied at different levels of abstraction as well as stages of software development. It includes several definition styles of specifications such as model-based or property-based, applicative or imperative, sequential or concurrent. However, it is not easy to identify reusable specifications, due to the fact that objects identified from this method tend not to be reusable in other applications as they are defined without a proper domain perspective. Even though RAISE supports object-orientation, the major approaches to object identification (keyword analysis, structured analysis, scenario-based analysis) can not provide support for reusability and adaptability of applications without analyzing the commonality and variability among a family of applications in a domain (Lee, Kang, Chae, & Choi, 2000).

To address this problem, there have been methods for reuse whose development has been based on the notion of "domain orientation" (Barstow, 1985; Prieto-Diaz, 1987), which emphasizes a group of closely related applications in a domain rather than a single application. The exploitation of commonality across related software systems is a fundamental technical requirement to achieve successful software reuse. Software product lines (PL) (Kang, Kim, Lee, & Kim, 2003) present a solid approach in large scale reuse. Due to the PL's inherit complexity, many PL methods use the notion of "features" to support domain modeling (e.g., FODA) (Kang et al., 1990), FORM (Kang, Kim, Lee, & Kim, 1998), FeatuRSEB (Griss, Allen, & d´Alessandro, 1998). These methods identify common abstractions across the applications of a domain in order to engineer reusable domain components. Feature modeling mainly focuses on identifying commonalities and variabilities among products of a PL, and organizing them in terms of structural relationships (e.g., aggregation and generalization) and configuration dependencies (e.g., required and excluded).

This work is related to the introduction of a feature-oriented reuse method to the RAISE components specifications and to give a solution to "bridge" the gap between the domain analysis and the specifications in RAISE (Riesco, Felice, Debnath, & Montejano, 2005). In particular, FORM method (feature-oriented reuse method) is introduced in order to gradually improve the reuse of RSL (George et al., 2002) components specifications, incorporating the effectiveness of software reusability as an integral part of the software specification process, considering the following:

- To create a plan of reuse as part of the project plan
- To add steps to look for and evaluate reusable component candidates to use in the project
- To add guidelines to create a future component for reuse
- To evaluate the benefits and costs associated with the practice of reuse in the project
- To finally identify created components of the project which have the potential for reuse in the future

BACKGROUND

PL engineering is an emerging software engineering paradigm, which guides organizations toward the development of products from core assets rather than the development of products one by one from scratch. Two major activities of PL are core asset development (i.e., product line engineering) and product development (i.e., product engineering) using the core assets. The paradigm of developing core assets for application development has been called domain engineer-

ing (DE), in which emphasis is given to the identification and development of reusable assets from an application "domain" perspective (Lee, Kwanwoo, Kang, & Lee, 2002). The purpose of domain engineering is to develop domain artifacts that may be used and reused in development applications for a given domain. DE consists of activities for gathering and representing information on systems that share a common set of capabilities and data. In usual approaches to software reuse, the product of such domain engineering might include only the reusable components and their parametric representations applicable to that domain.

DE, the key to systematic software reuse, has two phases: domain analysis (DA) and domain implementation.

DA is the process of discovering and recording the commonalities and variabilities in a set of software systems, while domain implementation is the use of the information from DA to create reusable assets and new systems within a domain (Frakes, 1994).

The view of DA follows the line of thought pioneered in Neighbors' DRACO system (Neighbors, 1980), where the importance of DA in reusability was pointed out. This is summarized as "to identify objects and operations for a class of similar systems."

DA is a process that affects the maintainability, usability, and reusability of a system and includes:

- Domain definition
- Domain analysis
- The development of a domain architecture
- The construction of components (specifications, design, documentation)

In the field of software engineering, DA is seen as a prerequisite for successful reuse not only by the researchers in the reuse community but also by the methodologists who have introduced component-based development methods (D'Souza & Wills, 1999; Gamma, Helm, Johnsson, & Wlissides, 1995; Jacobson, Booch, & Rumbaugh, 1999). A review of some domain analysis methods is presented in Kang (1990) and an extensive domain analysis bibliography can be found in Hess et. al. (1990).

The result of the analysis, usually known as the domain model, is retrieved for reuse in future developments or similar systems and also, for maintainability of legacy systems.

In a reuse strategy, DA must be maintained over many systems, and the repository should contain domain models that form the basis of subsequent development activities.

DA is essential to formalize reuse. However, it is missing from most software development methods. Reuse engineering extends information engineering by adding the new stage "domain analysis" to provide a place in the life cycle where the most valuable reusable components for the domains of the enterprise can be identified and a library containing these components can be created. At this stage of the software development, working with formal methods (or formal specification languages, specifically) implies to provide a means of unambiguously stating the requirements of a system, or of a system component. In this way, formally specified system components that meet the requirements of components of the new system can be easily identified. Thus, components that have been formally specified and sufficiently well documented can be identified, reused, and combined to form components of the new system.

Nevertheless, the main problem is that the requirements may not be clear. Specially, when the requirements are written in a natural language the result is likely to be ambiguous. The aim of the initial specification is to capture the requirements in a precise way applying a reusability model.

In particular, there are two main activities in the RAISE method: writing an initial specification, and developing it towards something that can be implemented in a programming language (George et al., 2002). Writing the initial specification is the most critical task in software development. If it is wrong (i.e., if it fails to meet the requirements) the following work will be largely wasted. It is well known that mistakes made in the life cycle are considerably more expensive to fix than those made later. So, the introduction of a DA method is a crucial task considering the possibility of reusing the specifications in the future.

Examples of more relevant DA methods include FODA, FORM, and FeatuRSEB. They support the notion of feature-oriented. This is a concept based on the emphasis this method places on finding the features or functionalities usually expected in applications for a given domain. The FORM engineering processes are illustrated in Figure 1.

THE FORM METHOD

FORM product line engineering consists of two major processes: asset development and product development, as it can be seen in Figure 1. Asset development consists of analyzing a product line (domain analysis, feature modeling) and developing architectures and reusable components based on analysis results. Product development includes analyzing requirements, selecting features, selecting and adopting an architecture, and adapting components and generating code for the product.

This method has been applied to several industry application domains including elevator control systems, electronic bulletin board systems yard automation systems and PBX, to create product line software engineering environments and software packages (Kang et al., 2003).

CONCEPTS

Feature-Oriented

The use of "features" is motivated by the fact that users and developers often speak of product characteristics in terms of "features the product has and/or delivers." In a DA, primary inputs are documents of applications (users manual, design documents, etc.). The volume of the documents to be analyzed tends to be enormous in a domain of any reasonable size. The experience has shown that the DA can be performed efficiently by instead analyzing the domain language. That is, services provided, and techniques used in applications are abstracted as "features," and they are used by domain experts to communicate their ideas, needs, and problems. Feature models have been used to support different engineering techniques, re-engineering legacy systems (Kang, Kim, Lee, & Kim, 2005), and other purposes (Czarnecki, & Antkiewicz, 2005). An informal definition of feature is stated as essential characteristics of applications in a domain. Also, rigorous conceptual foundations for feature modeling have been introduced (Asikainen, Mannisto, Soininen, 2006).

Summarizing, features may serve as a means of:

- Modeling large domains
- Managing the variability
- Future planning
- Communication between system stakeholders
- Guiding the PL development

To create coherent models, feature analysis involves tasks for identifying, classifying, and organizing product features as models.

Feature Classification

As potential features are identified, they are classified according to the types of information they represent. For example, users are concerned with functions provided by the systems (i.e., service features), analysts and designers are concerned with domain technologies, and developers are concerned about implementation techniques.

Thus, four categories are distinguished: those about application capabilities, operating environments, domain technologies, and implementation techniques.

A feature model should cover all four categories of features for a domain. To make it possible, FORM uses the following constructs:

- A feature diagram, a graphical AND/OR hierarchy of features, capturing the logical relationships (composition/generalization) among features. Three types of relationships are represented in this diagram: "composed-of," "generalization/specialization," and "implemented-by." Also, features may be "mandatory" (unless specified otherwise), "optional" (denoted with a circle), or "alternative" (denoted with an arc) (see the example in Figure 2).
- Composition rules that supplement the feature diagram with mutual dependency and mutual exclusion relationships.

Depending on the domain, it is possible that AND/OR diagram tends to become complex.

A feature model with AND-nodes at an upper level and OR-nodes at a lower level indicates a high level of reuse opportunity. Alternatives (i.e., OR-nodes) at the upper level

Figure 1. FORM engineering processes

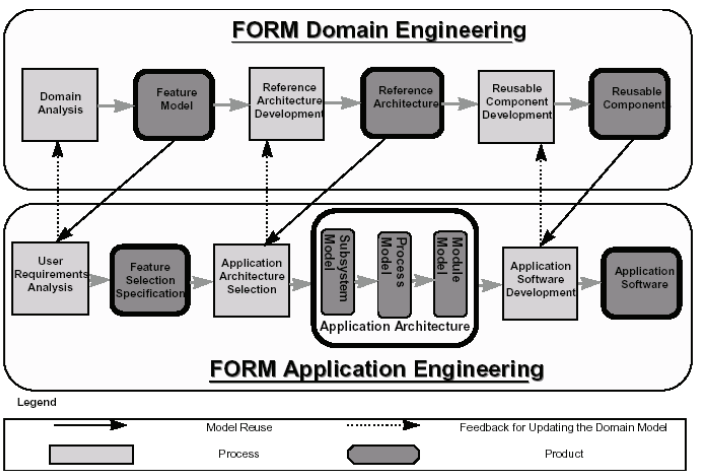
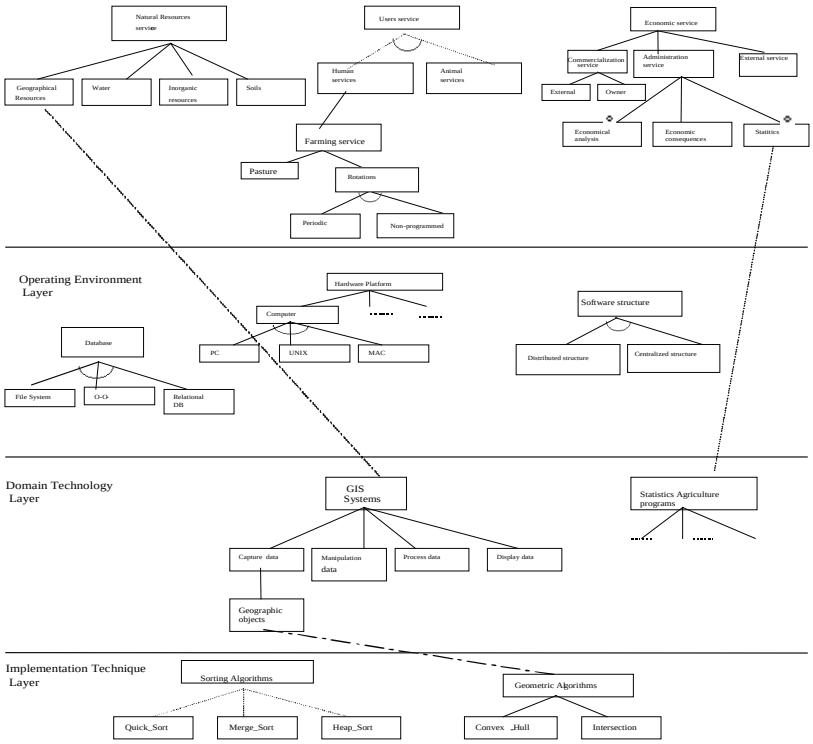


Figure 2. Feature Model of Agriculture system domain



Relationship

Composed of

Generalization/ Specialization

Implemented by

usually mean that applications in the domain do not share much commonalities in terms of services and functions provided by them. This indicates that the domain might not have much reuse opportunity at the application level, although there might still be opportunities for reuse at low-level generic components such as math libraries and abstract data types. Alternatives (OR-nodes) at a lower level indicate different ways of implementing certain components and, with an appropriate application of information hiding and encapsulation techniques, reusable components can be developed.

Features are largely classified into functional and non-functional features. Functional features include services and operations that are needed to provide services. Figure 2 shows the feature model of agriculture system domain (Riesco et al., 2005).

Features are considered following the four level feature hierarchy since the hierarchy reflects step-wise refinement in the reference architecture. These concepts are strongly connected with the style of development in RAISE, where the separate and step-wise are the basis to build a solid specification of an infrastructure.

Figure 2 shows some features (represented as tree nodes) exhibited in the Agriculture system domain. A capability feature characterizes a distinct service, operation, function or performance that a given domain may possess. For instance, “natural resources service,” “users service,” and “economic service” are features that characterize the functional aspects of the agriculture systems. It is assumed that natural resources are geographical resources, water, inorganic resources and soils.

On the other hand, users may be animal or human services. Humans have motivations related to the services that farming groups can do. With respect to economic service, it includes people and other services linked to a different domain. The agriculture system is an information system whose objective is the model, which will help deciders to identify problems with the management and the access to resources for several purposes. Thus, an operating environment feature represents attributes of environment in which this system is operated. In this case, this is not a crucial objective; the hardware platform will be related to the equipment (e.g. “PC”), operating system, database systems,

and software structures. Domain technology feature is more specific to a given domain. For instance, GIS systems are closely related with geographical features and the way to capture process and display data. It works with geographic objects linked directly with geometric algorithms allocated to Implementation technology feature that are more generic and may be used in another domain.

The domain model products should represent the relevant information about the objects and the functionality of a family of a similar system in a domain. The validation of the model is obtained by reproducing known applications through the selection of specific features and building a prototype system.

RAISE SPECIFICATION ARCHITECTURE

To be able to incorporate the idea of reusability in different forms of development, it is necessary to take into consideration many activities that have to be applied to the RAISE method. It provides guidelines to hierarchically structure a specification, aiming the encouragement of separate development and step-wise development.

A development in RAISE begins with an abstract specification and gradually evolves to concrete implementations. Figure 3 shows these phases. RAISE proposes to structure modules hierarchically in order to get a particular component over by reference only to it and its suppliers, to limit the effects of changes of a module to it and its clients, and to limit the properties of a module to it and its suppliers.

It is an object-oriented method and covers a large portion of systems development phases. Nevertheless, in these phases, DA is absent before the first specifications. Our work is focused on the incorporation of DA upon the idea of specifying and designing a family of systems to produce qualitative applications in a domain, promoting early reuse and reducing development costs.

This architecture is the basis to start applying the steps of the RAISE method. Also, in Riesco, Felice, Debnath, & Montejano, 2004) concepts and assumptions of component reuse were analyzed based on the class generality assessment and relations among classes.

Our aim is to concentrate on an earlier analysis according to the "features" for a domain and then, to improve the specifications development by building a feature model.

The Architectural Model

FORM gives a set of guidelines that can be used to derive domain products from the feature model. Each feature somehow constrains the selection of the final reference model considering differences in types of features following the four level hierarchy (Figure 4).

Following, it is discussed how the feature model serves as a guideline to identify RSL reusable modules. In the example illustrated in Figure 3, model service features such as natural resources service, users service, and economic service play different user roles for the purpose of the agriculture systems. Thus, they can be mapped respectively into natural resources, users and economic RSL modules, each of which performs a set of operations with its specific role. Once a feature is mapped into a module, the sub-features of the feature such us geographical resources, water, inorganic resources, and soils can be modules that are part of natural resources following the same type of relationships in the feature model (e.g., generalization, aggregation).

Besides, operations can be mapped as internal functions to provide services, and they are a collection of types and values without type of interest.

On the other hand, non-functional features include end-user visible application characteristics that cannot be identified in terms of services or operations like quality attribute, cost, etc, so, they can not be mapped into any RSL constructions.

With respect to the model operating environment features, the RSL specifications are independent from the operating environment. These features can be mapped into the subsidiary RSL modules, which are less important from the point of view of development.

Figure 3. Phases of the RAISE method

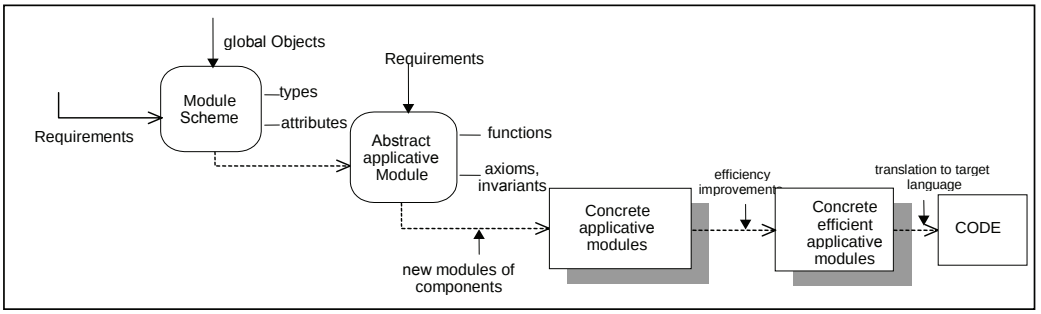
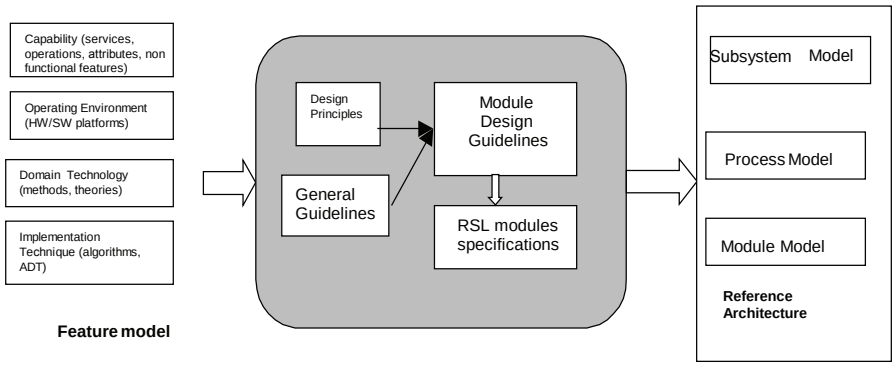


Figure 4. Mapping from feature categories to RSL specification



Model domain technology features such as GIS systems will be considered by the RSL system modules and they will be expected to be finally implemented as software modules. In object-oriented terms, they will form the objects of the software system.

Modules derived from implementation technique features are generally used to implement or to derive concrete applicative specifications derived from capability or domain technology features. Both sorting algorithms and geometric algorithms would be part of a module called "algorithms." This module will be defined as a generic module (i.e., a module we expect to instantiate more than once with different parameter), being the sub-features (QuickSort, Merge_Sort, Intersection) the possible instantiations.

Each RSL module derived would be later refined and completed with the definition of functions. In Mauco, Riesco, and George (2001) some heuristics to derive functions in a RSL specification are defined.

Once the RSL modules are derived from the feature model, they need to be organized into a model in order to represent how they are related to each other and what the contexts for their use are.

To finally get the reference architecture model it is necessary to make a sequence of refinements: subsystem architecture, process architecture and the module architecture.

CONCLUSION

Domain analysis is seen as a prerequisite in successful reuse not only by the researches but also by the methodologists who have introduced component-based development methods. FORM is one of the promising methods for the identification of the desired system functionalities. The contribution of this work consists of the integration of this methodology into a formal method. It is a useful tool that allows to work with the identification of commonalties and variabilities among related applications creating a feature model of a specific domain. More precisely, a feature model has been developed for the Agriculture system being the basis to the specifications of the RAISE reusable component.

The use of formal methods in system development can help to overcome inconsistencies, and should aid the promotion of software reuse in early stages of software development.

REFERENCES

Asikainen, T., Mannisto, T., & Soininen, T. (2006). A unified conceptual foundation for feature modelling. In *Proceedings of the 10th International Software Product Line Conference (SPLC 06)* (pp. 31-40).

Barstow, D.R. (1985). Domain-specific automatic programming. *IEEE Transaction on Software Engineering: SE11* (Vol. 11, pp 1321-1366).

Czarnecki, K., & Antkiewicz, M. (2005). Mapping features to models: A template approach based on superimposed variants. In *Proceedings of GPCE '05*.

D'Souza, D., & Wills, A. (1999). *Objects, components, and frameworks with UML: The catalysis approach.* Addison-Wesley.

Frakes, W. B. (1994). Software reuse: Advances in software reusability. In *Proceedings of the 3rd International Conference on Software Reuse*, Los Alamitos, CA. IEEE CS Press.

Gamma, E., Helm, R., Johnsson, R., & Wlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software*. Reading, MA: Addison-Wesley Publishing Company.

George, C., Haff, P., Havelund, K., Haxthausen, A., Milne, R., Nielsen, C., Prehn, S., & Ritter, K. (2002). *The RAISE specification language*. UK: Prentice Hall.

George, C., Haxthausen, A., Hughes, S., Milne, R., Prehn, S., & Pedersen, J. (1995). *The RAISE Development Method*. BCS Practitioner Series. Denmark: Prentice Hall.

Griss, D., Allen, R., & d´Alessandro, M. (1998). Integrating feature modelling with the RSEB. In *Proceedings of the 5th International Conference of Software Reuse* (*ICSR*-5).

Hess, J., Novack, W., Carrol, P., Cohen, S., Holibaugh, R., Kang, K., & Peterson, A. (1990). *A domain analysis bibliography*. SEI-90-SR-3 Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The unified software development process*. Addison Wesley.

Kang, K., Cohen, S., Hess, J., Novak, W., Peterson, A. (1990). *Feature-oriented domain analysis (FODA) feasibility study (CMU/SEI-90-TR-21, ADA235785)*. Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

Kang, K., Kim, S., Lee, J., & Kim, K. (2003). Feature-oriented product line software engineering: Principles and guidelines. In *Domain Oriented Systems Development: Practices and Perspectives* (pp. 19-36). New York: Taylor & Francis.

Kang, K., Kim, S., Lee, J., & Kim, K. (1998). FORM: A feature-oriented reuse method with domain-specific reference architectures. *Annals of Software Engineering* (Vol. 5, pp. 143-268).

Kang, K. C., Kim, M., Lee, J., & Kim, B. (2005). Feature-oriented re-engineering of legacy systems into product line assets -a case study. In *Proceedings of SPLC 2005*.

Lee, K., Kang, K., Chae, W., & Choi, B.(2000). Feature-based approach to object-oriented engineering of applications for reuse. In *Software practice and experience*, *30*(9), pp. 1025-1046.

Lee, K., Kang, K., & Lee, J. (2002). Concepts and guidelines of feature modeling for product line software engineering. *Software Reuse: Methods, techniques, and tools, the 7th International Conference, ICSR-7*, Austin, TX, Proceedings. Lecture Notes in Computer Science 2319 (pp: 62-77).

Mauco, V., Riesco, D., & George, C. (2001). Using a scenario model to derive the functions of a formal specification. *The 8th Asia-Pacific Software Engineering Conference (APSEC 2001)* China. (pp. 329-332). IEEE Computer Society Press.

Neighbors, J. (1980). *Software construction using components*. Ph.D. dissertation, Department of Information and Computer Science, University of California, Irvine.

Prieto-Diaz, R. (1987). Domain analysis for reusability. In *Proceedings of COMPSAC 87: The 11th Annual International Computer Software and Applications Conference* (pp. 23-29). Washington DC. IEEE Computer Society.

Riesco, D., Felice, L., Debnath, N., & Montejano, G. (2005). Using a feature model for RAISE specification reusability. In *Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration IRI-2005*. Las Vegas, NV (pp. 306-311).

Riesco, D., Felice, L., Debnath, N., & Montejano, G. (2004). Incorporating a reuse model to the RAISE formal method. In *Proceedings of the 2004 IEEE International Conference on Information Reuse and Integration IRI-2004*. Las Vegas, NV (pp. 133-138).

KEY TERMS

Domain Analysis: DA is a process that affects the maintainability, usability, and reusability of a system and includes:

- Domain definition
- Domain analysis
- The development of a domain architecture
- The construction of components (specifications, design, documentation)

Domain Engineering: DE consists of activities for gathering and representing information on systems that share a common set of capabilities and data. In usual approaches to software reuse, the product of such domain engineering might include only the reusable components and their parametric representations applicable to that domain. DE, the key to systematic software reuse, has two phases: domain analysis (DA) and domain implementation.

Feature:

- Is anything users or client programs might want to control about a concept?
- Is a coherent and identifiable bundle of system functionality that helps characterize the system from the user perspective?
- Is a prominent or distinctive user-visible aspect, quality, or characteristic of a software system or systems?

Feature Modeling: The process where it is documented only functional features but also implementation features, various optimizations, alternative implementation techniques, and so on.

RAISE Method: RAISE is a formal method. It is an acronym for “rigorous approach to industrial software

engineering" It provides facilities for the industrial use of formal methods in the development of software systems. RAISE consists of the RAISE specification language (RSL), which is a powerful specification and design language and a comprehensive development method.

Software Product Line: A software product line is a set of software-intensive systems that share a common, managed set of features satisfying the specific needs of a particular market segment or mission and that are developed from a common set of core assets in a prescribed way.

Integrating Enterprise Systems

Mark I. Hwang
Central Michigan University, USA

INTRODUCTION

In the last two decades many organizations installed enterprise resource planning (ERP) systems as a means to integrate their back-office operations. The need for integration, however, actually amplified with the advent of ERP. In addition to integrating ERP with legacy systems, consolidating multiple copies of ERP running in different business units posed major challenges. Moreover, recent strategic initiatives such as customer relationships management (CRM), supply chain management (SCM), business to consumer (B2C), and business to business (B2B) all require a free flow of information between ERP and other enterprise systems to be successful. It is, therefore, more critical than ever to plan for and implement integration projects involving ERP properly. Hwang (2005) describes the need for integrating enterprise systems in detail. He also discusses several success factors cited in practitioner journals. Since then a handful of empirical studies have been published in the scholarly literature. This article provides a review of those studies with a special focus on the success factors. A consolidated list of success factors is developed for practitioners. Promising research directions are discussed.

BACKGROUND

While researchers have examined integration issues for some time, it was not until the early 2000s that empirical studies involving ERP began to appear in the literature. Table 1 summarizes the characteristics of the five empirical studies reviewed, and Table 2 summarizes the critical success factors (CSF) discussed. Alshawi, Themistocleous, and Almadani (2004) investigated the feasibility of minimizing ERP customization through integrating two ERP packages. They found that an enterprise application integration (EAI) tool was useful in integrating SAP R/3 with an Oracle H/R module at a telecommunication company. Sharif, Irani, and Love (2005) studied the integration project of a global industrial company involving ERP and legacy systems. The integration effort was deemed unsuccessful based on a post hoc evaluation model that they developed. Lam (2005) proposed a CSF model for EAI projects. He termed this the BOTP model, after the categories into which the success factors fall: business, organization, technology, and project. A case study involving a large financial services provider integrating its consumer banking systems revealed three broad groups of success factors: top management support, integration strategy, and project planning and execution. Mendoza, Perez, and Griman (2006) developed a set of 20 CSFs and tested them in two case studies, one in a B2B and the other in an ERP setting. Many but not all of the success factors were present in the two companies. Finally, Stefanou and Revanoglou (2006) examined a successful ERP implementation at a hospital.

Alshawi et al. (2004) and Stefanou and Revanoglou (2006) did not discuss their findings in the context of some type of success models. The three studies that did classified various CSFs by their types (e.g., organization vs. technology) or the type of integration involved (e.g., intra- vs. inter-organization). One group of variables discussed by Sharif et al. (2005), for instance, deals with ERP II tailorability, the ability to integrate ERP with customers via CRM and B2C and with business partners via SCM. Building on the maturity model of Schmidt (2000), Mendoza et al. (2006) developed their list of CSFs based on different levels of integration, from level one point-to-point integration to level

Table 1. Study characteristics

Study	Case Study	CSF Model
Alshawi et al. (2004)	Integrating two ERP systems	No
Sharif et al. (2004)	Integrating ERP with legacy systems	Yes
Lam (2005)	Integrating ERP with legacy systems	Yes
Mendoza et al. (2006)	Integrating ERP with legacy systems; B2B integration	Yes
Stefanou and Revanoglou (2006)	Integrating ERP with legacy systems	No

Table 2. Critical success factors in the literature

Study-	Business	Organization	Technology	Project
Sharif et al. (2004)	1. optimization of business models 2. acceptability of success	1. effect of influencers	**1. vertical specialization** **2. horizontal specialization** **3. extended ERP functionality**	1. scope of technical effort involved
Lam (2005)	1. strong business case 2. overall integration strategy **3. process interoperability with business partners**	1. top management support 2. business process change and overcoming resistance to change 3. good organizational and cultural fit	1. handling legacy systems 2. technology planning 3. common data standards 4. use of right tools 5. use of mature technology	1. realistic project plans and schedule 2. client involvement, communication, consultation, and training 3. required skills and expertise on-board, vendor competence 4. monitoring and feedback 5. proper migration approach 6. adequate testing plans
Mendoza et al. (2006)	1. careful strategy of implementation	**1. valuable support by senior management** **2. change determined and justified at a productivity level** 3. effective organizational change management **4. appropriate strategy of security** 5. known organizational structure	1. standard data model documentation, unification, and updating 2. appropriate configuration of communication software 3. helpful technical support 4. complete technological infrastructure	**1. effective outgoing and incoming communication** **2. adequate management of project scope** 3. appropriate outsourcing management 4. high expertise project team 5. low impact of IS on the organization 6. effective internal and external training plan 7. relevant user involvement 8. valuable project management 9. effective project leadership 10. significant administrative support for the project consultant

four external integration. Lam (2005) does not distinguish internal from external integration projects but acknowledges that some factors such as "process interoperability with business partners" are more important in inter-organization settings than in intra-organization settings.

Table 2 organizes the CSFs into four groups of Lam (2005): business, organization, technology, and project. This is a general classification scheme into which any success factor can be classified. At the same time, it makes sense to differentiate factors that are oriented toward more external integration or ERP II tailorability (Sharif et al., 2005) than internal integration. Those external-oriented factors are boldfaced in Table 2. It is, however, important to note that the internal/external dimension should be treated as a continuum rather than a dichotomy because some factors apply to both intra- and inter-organizational settings (Mendoza et al., 2006)

CRITICAL SUCCESS FACTORS

As can be seen in Table 2, the success factors discussed by different researchers share a number of commonalities. A consolidated list of success factors is presented in Table 3, with factors that are either common across different studies or fit closely with the theme of each category. As shown in Table 3, for instance, the theme of the business factors category is to define the value and strategy of integration. The list in Table 3 is concise and presented in an easy-to-use format for practitioners. It can be expanded or modified as more studies appear in the literature in the future. The next paragraphs discuss all the factors listed in Table 2.

Factors dealing with business aspects are related to the value and strategy of integration. In today's business environment it is critical to demonstrate the return on investment (ROI) of any major endeavors, especially for expensive and complicated integration projects (Lam, 2005). It is also important to develop an integration strategy (Lam, 2005; Mendoza et al., 2006) at the outset including key performance indicators (Mendoza et al., 2006). Sherif et al. (2004) similarly discuss the need for defining the success for integration projects. They also suggest "optimization of business models" as a success factors, because organizations with flexible and adaptive business models are likely to value integration efforts. Finally, Lam (2005) describes "process interoperability with business partners," a factor admittedly more important to external than internal integration projects.

Organizational factors deal with the acceptance of the mission by all the constituencies. Top management support

Table 3. Critical success factors for practitioners

Business Factors: Define the value and strategy of integration • Strong business case; clear key performance indicators • Adequate integration strategy • Proper business models/processes
Organization Factors: Secure acceptance of mission by constituencies • Strong top management support • Strong support/commitment from all levels • Good culture/organizational fit • Proper change management
Technology Factors: Mull over the technical aspects of projects • Good extensibility of ERP software • Appropriate technology planning • Complete technological infrastructure • Standard data model • Good use of right tools
Project Factors: Deal with the management and execution of projects • Proper definition of scope and schedule • Strong user involvement • Adequate skills and expertise • Strong project leadership • Appropriate project plan including testing, training, and conversion procedures

consistently ranks high among success factors for major information systems projects including integration (Lam, 2005; Mendoza et al., 2006). Sharif et al. (2005) discuss the effect of all influencers, including stakeholders both inside and outside of the organization. It is conceivable that commitment from not only the senior level but also every level of the organization is required. Mendoza et al. (2006), for instance, recommend that the implementation team study the organization structure and determine its support for the integration process. The full cooperation from all personnel is more attainable if the integration project fits well with the culture and organizational environment (Lam, 2005; Stefanou & Revanoglou, 2006). Similarly, effective change management is a precondition for organizational acceptance (Lam, 2005; Mendoza et al., 2006; Stefanou & Revanoglou, 2006). Mendoza et al. (2006) further identify appropriate security strategy as a success factor for external integration projects due to the need for safeguarding data and applications that span organizational boundaries.

Technology factors deal with the technical aspects of integration efforts. Sharif et al. (2005) focus on the attributes of ERP packages including vertical specialization, horizontal specialization, and extended functionality. These are all factors affecting how well an ERP package integrates with external systems. Vertical specialization pertains to integration within an industry, whereas horizontal specialization is related to integration with other best-of-breed solutions to create an end-to-end e-business offering. Extended functionality affects how easy it is to integrate an ERP package via componentization or modularization. Mendoza et al. (2006) describe the importance of having a complete technological infrastructure, including an internal network, operating systems, and software tools. They also discuss the need for proper communications software, which applies to point-to-point integrations primarily, and a standard data model to ensure consistent data and secure transactions in structural integrations. Another technology factor is helpful technical support provided by software vendors. Lam (2005) similarly suggests common data standards and the ability to integrate with legacy systems. He also discusses the use of right tools and mature technology, which was echoed by other researchers (e.g., Alshawi et al., 2004; *Themistocleous, 2004)*. All of these require proper technology planning.

Project factors are related to the management and execution of integration projects. An important factor emphasized by all researchers is the proper definition of scope. User involvement is another oft-cited success factor for IS projects and is recommended by both Lam (2005) and Mendoza et al. (2006). Having the requisite skills and expertise, which may come from a team of employees, consultants, and vendors working together, is another common success factor (Lam, 2005, Mendoza et al., 2006). Mendoza et al. (2006) also include effective project leadership, appropriate outsourcing management, and significant support to consultants as additional human capital factors. The project plan should include a proper migration approach (Lam, 2005) and other traditional project management variables such as testing (Lam, 2005), training, and conversion procedures to ensure low impact on the business process (Mendoza et al., 2006). Lastly, effective outgoing and incoming communication is needed to properly determine requirements and needs for external integration projects (Mendoza et al., 2006).

FUTURE TRENDS

Table 2 presents a summary of the critical success factors discussed in the academic literature. One technology factor that is receiving a lot of attention in the practitioner literature is service-oriented architecture (SOA). In a recent survey of IT executives, more than half have invested or are considering implementing SOA to integrate their applications (CIO research reports, 2006). The majority of those respondents also agree that SOA tools and protocols are still evolving. SOA has received major endorsement from key software development tools vendors including IBM, Oracle, and BEA Systems (Morejon, 2005). The potential benefits of SOA include software reuse, low cost of integration, business agility, and risk reduction (Lager, 2006). Similar to many emerging technologies, no universally accepted definition for SOA exists. One view is that it is a paradigm for developing loosely coupled software components or services that

encourages software reuse, integration, and interoperability (He, 2003; Kobielus, 2005). SOA can be implemented under various platform and middleware environments, including Web services (Khanna, 2005; Kobielus, 2005). It is likely that both SOA and Web services will play an important role in integrating enterprise systems in the near future.

The three empirical studies provided support for several but not all of the success factors. Other researchers are encouraged to test these factors in additional case studies or cross-section surveys. Another promising research stream is to further consolidate and refine the critical success factors list. One approach is to combine the most promising factors from the three empirical studies. Another approach is to compare and augment the list with factors found in the practitioner literature. Several factors discussed by Hwang (2005), for instance, robust infrastructure and human capital, have similar counterparts in Table 3. Incorporating practitioners' definitions of these variables into the operationalization of factors in future study will increase the relevance of the research, as will inclusion of new factors such as prioritization of projects. Another useful research area is the dependent variable, the integration success. Sharif et al. (2005) suggest that success be explored from strategic, tactical, and operational aspects. Mendoza et al. (2006) similarly discuss the importance of defining the key performance indicators. Without an agreed-upon definition of success, it will not be meaningful to assess the effect of success factors. Finally, the specific effect of individual factors on success is significant. Some factors are likely more important than others, but which ones? To what degree are they more important and under what circumstances? These are all critical research questions.

CONCLUSIONS

Systems integration is an ongoing issue in the use of information technology by businesses. The introduction of ERP is a step in the right direction toward integration, but it also brings new challenges to the creation of a truly integrated enterprise. Integrating enterprise systems can be expensive and risky; at the same time the financial savings can be substantial (Lam, 2005; Puschmann & Alt, 2004). Many factors come into play in determining if an integration project will be successful. More research is warranted, but the factors that are highlighted in this article should help organizations increase their odds of success.

REFERENCES

Alshawi, S., Themistocleous, M., & Almadani, R. (2004). Integrating diverse ERP systems: A case study. *Journal of Enterprise Information Management, 17*(6), 454-462.

CIO research reports. (2006). *CIO and Computerworld research: The forecast for SOA*, March 06. Retrieved July 24, 2006, from http://www2.cio.com/research/surveyreport.cfm?id=106

He, H. (2003). *What is service-oriented architecture?* Retrieved July 24, 2006, from http://www.xml.com/pub/a/ws/2003/09/30/soa.html

Hwang, M. I. (2005). Enterprise resource planning and systems integration. In Mehdi Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 1083-1088). Hershey, PA: Idea Group.

Khanna, P. (2005). SOA plants the seeds of true system integration. *Computing Canada, 31*(10), 18.

Kobielus, J. (2005). Three steps to SOA nirvana. *NetworkWorld*, (October 10). Retrieved July 24, 2006, from http://www.networkworld.com/techinsider/2005/101005-soa-steps.html

Lager, M. (2006). SOP simple. *Customer Relationship Management, 10*(2), 20-24.

Lam, W. (2005). Investigating success factors in enterprise application integration: A case-driven analysis. *European Journal of Information Systems, 14*(2), 175-187.

Mendoza, L. E., Perez, M., & Griman, A. (2006) Critical success factors for managing systems integration. *Information Systems Management, 23*(2), 56-75.

Morejon, M. (2005). Making a connection with SOAs. *CRN, 1143*, 89-90.

Puschmann, T., & Alt, R. (2004). Enterprise application integration systems and architecture: The case of the Robert Bosch Group. *The Journal of Enterprise Information Management, 17*(2), *105-116.*

Schmidt, J. (2000). *Enabling next-generation enterprises. EAI Journal, 2*(7), 74-80.

Sharif, A. M., Irani, Z., & Love, P. (2005). Integrating ERP using EAI: A model for post hoc evaluation. *European Journal of Information Systems, 14*, 162-174.

Stefanou, C. J., & Revanoglou, A. (2006). ERP integration in a healthcare environment: A case study. *The Journal of Enterprise Information Management, 19*(1), *115-130.*

Themistocleous, M. (2004). Justifying the decisions for EAI implementations: A validated proposition of influential factors. *The Journal of Enterprise Information Management, 17*(2), 85-104.

KEY TERMS

Critical Success Factors: Those things that must go right in order for an organization to achieve its mission.

Enterprise Application Integration: Comprehensive middleware software suits that allow connection to an array of applications including enterprise resource planning, customer relationship management, and to various databases.

Enterprise Resource Planning: Configurable enterprise software that integrates business processes across functions.

Enterprise Systems: Information systems that allow companies to integrate information across operations on an enterprise-wide basis.

Service-Oriented Architecture: A paradigm for developing loosely coupled software components or services, which encourages software reuse, integration, and interoperability.

Systems Integration: The process of tying together two or more computer systems for sharing data and functionality.

Web Services: Technologies that allow easy integration of applications over the Internet or Internet protocol-based networks.

Integrating Natural Language Requirements Models with MDA

María Carmen Leonardi
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

María Virginia Mauco
Universidad Nacional del Centro de la Pcia. de Buenos Aires, Argentina

INTRODUCTION

The model driven architecture (MDA) is a framework for software development defined by the OMG (Object Management Group, 2006). The MDA initiative shifts the focus of software development from writing code to building models, separating the specification of functionality from the specification of the specific implementation of that functionality (Miller & Mukerji, 2003). Key to MDA is the importance of models and transformations between them in the software development process.

The first model of MDA is the computation independent model (CIM), which describes the business model independently of the software system to be implemented. The CIM is described with a vocabulary that is familiar to stakeholders. As it captures the domain without reference to a particular system implementation or technology, the CIM would remain the same even if the system were implemented mechanically, rather than in computer software (Meservy & Fenstermacher, 2005). CIM reduces the gap between stakeholders and software engineers (Miller et al., 2003).

Recently, some proposals related to business modeling and MDA have appeared, for example, the use of activity diagrams (Mersevy et al., 2005), BPMN (White, 2004), or goal-oriented strategies (Birol, 2006). Several authors agree in the importance of using the "language of the business experts" (Francis, 2006) during the first stages of development, enhancing communication between the domain experts domain and software engineers. From the requirements engineering area, we have been working with natural language requirements models to describe the universe of discourse (Leite, Hadad, Doorn, & Kaplan, 2000). But, to fit in MDA framework, we have to map them into object-oriented models, defining a CIM that will be the basis for a MDA software development. To do this, we have proposed a transformation strategy and developed its associated tool, CIMTool, thus allowing the integration of the natural language models into the MDA framework. In this article, which is an integration of Leonardi (2003), Leonardi (2005), Leonardi and Mauco (2004), and Leonardi, Mauco, and Leoni (2005), we present the overall strategy defining OCL based transformation rules to derive a CIM from the language extended lexicon (LEL) and the scenario model (Leite et al., 2000).

BACKGROUND

MDA is an approach that makes modeling the primary focus of the software development (Miller et al., 2003). It is based on modeling different aspects and levels of abstraction of a system, and exploiting interrelationships between these models. The MDA starts with the idea of separating the specification of the operation of the system from the details of its implementations. In MDA, all artifacts such as requirements specification, architecture descriptions, design descriptions, and code are regarded as models. One of the key features of this framework is the notion of automatic transformations that are used to modify one model in order to obtain another one. MDA defines how models expressed in one language can be transformed into models in other languages. The MDA standard proposes UML as the specification language, and is divided into the following main steps:

- Construct a model describing the business system: Computation independent model (CIM)
- Construct a model with a high level of abstraction: Platform independent model (PIM)
- Transform the PIM into one or more platform specific models (PSMs)
- Transform the PSMs to code

There are some works that propose UML extensions to represent business models (Johnston, 2006; Vasconcelos et al., 2001). These extensions, though not conceived in MDA context, allow the construction of UML models that can be considered as CIMs as they model the business knowledge without considering any software system. However, stakeholders think in terms of processes, goals, resources, actors, among others rather than objects. Object orientation is an abstraction not easily understandable by stakeholders because objects encapsulate data and behavior in the same level (Jackson, 1995). Then, the introduction of techniques

and models closer to stakeholders' way of thinking would be really useful.

During the early stages of development, when the interaction with the stakeholders is crucial, the use of natural language oriented requirements models seems necessary in order to enhance communication. Everyone can read natural language, so it is still used to define the requirements documents (Sommerville, 2005). However, natural language can be ambiguous, surprisingly opaque, and is often misunderstood. This kind of requirements models has to be reinterpreted by software engineers into a more formal design on the way to a complete implementation. In particular, in MDA context it is necessary to transform them to UML models representing the CIM. We can mention some strategies that, though not developed in MDA context, derive UML models from different business models. For example, Díaz, Pastor, Moreno, and Matteo (2004) obtain sequence diagrams from use cases and in Alencar, Pedroza, Castro, Silva, and Ramos (2006), a strategy is proposed to derive class diagrams starting from i* model. In order to improve CIM construction, we describe in the next section a transformation process to automatically derive a class diagram representing a CIM, starting from two natural language requirements models.

A STRATEGY TO DERIVE A CIM FROM REQUIREMENTS MODELS

In this section, we describe the transformation strategy to map the natural language models into a CIM. The section is organized in three subsections: one to present the requirements models, the other presents the derivation strategy, and finally CIMTool, the tool implementing the strategy.

Natural Language-Oriented Requirements Models

The models presented in this section are well known, used, and accepted by the requirements engineering community (Leite et al., 2000). The models are: language extended lexicon model and scenario model.

- **Language extended lexicon:** The language extended lexicon (LEL) is a structure that allows the representation of significant terms of the Universe of Discourse. The purpose of the lexicon is to help to understand the vocabulary and its semantics. It unifies the language allowing communication with the stakeholder. LEL is composed by a set of symbols with the following structure: symbol name: word or phrase and set of synonyms, notions defining the denotation of the symbol and behavioral responses describing the symbol connotation. In the description of the symbols, two rules must be followed simultaneously: the "closure principle" that encourages the use of LEL symbols in other LEL symbols, and the "minimum vocabulary principle" where the use of symbols external to the application language is minimized. LEL terms define objects, subjects, verbal phrase, and states. Figure 1 shows the heuristics to define each type of symbol.
- **Scenario model:** A scenario describes situations of universe of discourse. A scenario uses natural language as its basic representation and it is connected to LEL. Figure 2 describes the components. In Leite et al. (2000) the scenario construction process is described.

Figure 1. Heuristics to represent LEL terms

Subject	**Notions:** Who the subject is.
	Behavioral responses: Register actions executed by the subject.
Object	**Notions:** Define the object and identify relationship with other objects.
	Behavioral responses: Describe the actions that may be applied to this object.
Verb	**Notions:** Describe who executes the action, when it happens, and procedures involved in it.
	Behavioral responses: Describe the constraints on the happening of an action, which are the actions triggered in the environment and new situations that appear as consequence.
State	**Notions:** What it means and the actions, which triggered the state.
	Behavioral responses: Describe situations and actions related to it.

Figure 2. Components of scenario

Title: identifies a sce nario.
Goal: describes the purpose of the scenario.
Context: defines g eographical and te mporal l ocations and preconditi ons.
Resources: identif y passive entities with which actors wor k.
Actors: detail entities actively involved in the scenari o
Episodes: each episode represents an action pe rformed by actors using resources.

The Derivation Strategy

Once the natural language requirements models are defined and validated with the stakeholders, they have to be manipulated in order to be mapped to class diagram representing the structural aspects of a CIM. We propose a rules-based strategy to do this that fit in MDA framework. We illustrate the application of the strategy with examples taken from a milk production system (Mauco, 2004).

As we shown in Figure 3, the process takes as the source model a LEL and a scenario model from a concrete case study and produces a class diagram described following UML 1.5 syntax (Unified Modeling Language specification V.1.5, 2003).

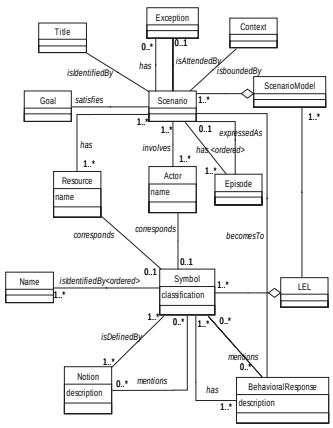
The steps described next organize the application of the transformation rules:

- **Identification of classes:** Taking as input LEL symbols classified as subjects and objects, transformation rules named TRC1 and TRC2 propose the definition of one class for each symbol. TRC2 also defines the methods for the classes coming from object LEL symbols.
- **Identification of methods:** Considering behavioral responses of subject LEL symbols, transformation rule named TRM1 defines the methods for the classes coming from subject LEL symbols (obtained after applying TRC1). Then, transformation rule TRM2 completes the corresponding parameters.
- **Identification of relationships:** The object diagram is completed with the definition of inheritance, aggregation and association relationships through transformation rule TRR by analyzing notions of LEL symbols defined as classes.

The Transformation Rules

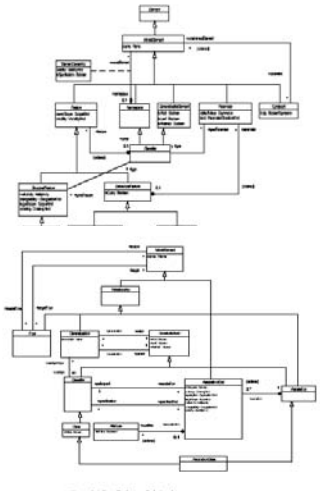
This section describes the transformation rules that allow the derivation of CIM from the requirement models. The transformation language we use is based on the transformation language proposed in Kleppe, Warmer, and Bast (2003), which is an OCL extension. Each transformation rule specification contains a name, the signature, a brief natural language description, and the OCL specification. Parameters may be any of the components of Figure 3, that

Figure 3. The source and the target models of the derivation process



UML class diagram of LEL and Scenario

Transformation rules



Backbone and Relationships Core Packages from UML V.1.5 Metamodel

is, any component of the Requirements Models or any of the components of the target model, referenced in each transformation rule as RM and UML respectively. The parameter TP represents a language dictionary with the language used in the construction of the requirements models (an English dictionary in this case).

TRC1_Transformation SubjectToClass (RM, UML, TP)

DESCRIPTION: Each subject LEL symbol becomes a UML class. The attributes are defined as follows: for each notion that does not contain a LEL symbol, the transformation identifies nouns and defines them as attributes.

Figure 4 shows a subject LEL symbol defining a Dairy Farmer and the class obtained by applying the transformation rule TRC1.

TRC2_ Transformation ObjectToClass (RM, UML, TP)

DESCRIPTION: Each object LEL symbol becomes a UML class. The attributes are defined as follows: for each notion that does not contain a LEL symbol, the transformation identifies nouns and defines them as attributes. Methods are defined adding SET and GET prefixes for each attribute.

Figure 5 shows the notion of the object LEL symbol Plot and the class defined by applying the transformation rule TRC2 to it.

TRM1_TransformationSubjectBehavioral ResponsesToMethods(RM,UML,TP)

DESCRIPTION: Each behavioral response of a subject LEL symbol modeled as a class by TRC1 becomes a method.

Box 1. TRC1 OCL specification

```
OCL Specification
SOURCE: S1: RM:: Symbol
            D: TP:: Dictionary
 TARGET: C1: UML :: Class
 SOURCE CONDITION
        S1.classification :: subject
 TARGET CONDITION
        C1.name = S1.isIdentifiedBy → first()
        let
            plainNotions :Set =
                    S1.isDefinedBy → excludes (n/ n.mentions -> notempty())
            nounofNotions: Set =
                    plainNotions → collect(n/ D.returnNouns(n.description)) asSet
               at: OrderedSet=
                    C1.features → collect (f/ f.oclIsTypeOf (Attribute))
         in
                    at → forAll (a/ nounofNotions → one (n: String / n = a.name))
```

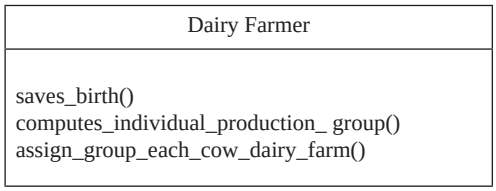
Figure 4. Dairy farmer subject LEL symbol and its corresponding class

DAIRY FARMER
NOTION
Person in charge of all the activities in a dairy farm
He has a name
He has a salary
He may have one or more employees
BEHAVIOURAL RESPONSE
He assigns to a group each cow of the dairy farm
He saves birth
He computes individual production of a group
He computes birth date for each dairy cow or heifer

Box 3. TRM1 OCL specification

```
OCL Specification
SOURCE: S1: RM:: Symbol
            D: TP :: Dictionary
-- D. ProcessString deletes spaces between strings, and deletes articles, prepositions and conjunctions,
returning nouns and verbs concatenated by _
TARGET: C1: UML :: Class
SOURCE CONDITION
        S1.classification:: subject
                C1.name = S1.isIdentifiedBy → first ()
TARGET CONDITION
        let
                behavioralNames : Sequence =
                        S1.has → (collect (br/ D.processString (br))) → AsSequence
        methods : Sequence =
                        C1.features → collect (f/ f.oclIsTypeOf(Operation))
        in
methods → forAll ( m/ behavioralNames → one (n: String /
n= m.name))
```

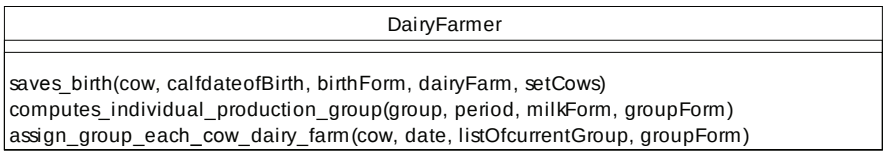
Figure 6. Methods of Dairy Farmer class



Box 4. TRM2 OCL specification

```
OCL Specification
SOURCE: S1: RM:: Symbol
TARGET: C1: UML :: Class
SOURCE CONDITION
                        S1.classification :: subject
                C1.name= S1. isIdentifiedBy → first ()
TARGET CONDITION
                    let
                      opers: OrderedSet =
                        C1.features → select (f/ f.oclIsTypeOf(Operation)
                    in
                        opers → forAll( o/ o.parameter =
                        (S1.has → select (description=o.name).becomesTo.has → collect (name)) union
                        (S1.has → select (description=o.name).becomesTo.involves → excludes (S1) →
                        collect (name)))
```

Figure 7. Defining parameters to methods of Dairy Farmer class



already defined (parameter listOfCurrentGroup, Figure 7, corresponds to a set of group).

TRR_Transformation LELRelationshipsToClassRelationships(RM,UML,TP)

DESCRIPTION: This transformation applies to subject as well as object LEL symbols. Notions of a LEL symbol, called L1, modeled as a class are analyzed in order to detect other LEL symbols also defined as classes. For each LEL symbol detected, named L2, the definition of an association relationship between the corresponding classes is considered, taking into account the kind of verbs involved in the descriptions (Juristo, Moreno, & López, 2000):

- **Inheritance relationships:** L1 and L2 have the same classification (object or subject). Besides, L1 appears in one of the notions of L2. The involved notions of L1 and L2 contain, in a complementary way, two kinds of verbs : bottom-up verbs (is a, is a type of, is a class

Box 5. TRR OCL specification

```
OCL Specification
SOURCE: S1: RM:: Symbol
            D1: TP:: Dictionary
TARGET: C: UML :: Class
SOURCE CONDITION
                C.name = S1.isIdentifiedBy(first)
TARGET CONDITION
let
cadidateInheritanceNotions: Set=
        S1.isDefinedBy → select (D1.BottonUpVerbsIncludes(n.description))
CandidateAggregationNotions: Set=
        S1.isDefinedBy → select(D1.Component_Composition.Includes(n.description))
CandidateAssociationNotions: Set=
        S1.isDefinedBy → excludes(cadidateInheritanceNotions union candidateAggregationNotions)
in
cadidateInheritanceNotions → forAll (n.mentions → exists (s: Symbol / s.classification = S1.classification and
Class.allInstances → exists (c1 / c1.name = s.name) and s.isDefinedBy → exists(n1/ n1.mentions-> includes(S1) and
D1.TopdownVerbsIncludes (n1.description)))
        implies  G.oclIsTypeOf(Generalization) and G.child = c1 and G.parent = C
        and c1.generalization = G and C.especialization = G) (*)
candidateAggregationNotions → forAll (and n.mentions → exists (s: Symbol / s.classification = S1.classification and
  (Class.allInstances → exists (c1 / c1.name = s.name)) and s.isDefinedBy → exists (n:notion / D1.
  Content_Composition_VerbIncludes (n.description) and n.mentions → includes (S1)))
        implies A.oclIsTypeOf (Association) and A.connection → at(1).participant = C
        and A.connection → at (1).aggregation = aggregate and  A.connection →
          at(2). participant= c1 and A.connection → at (2).aggregation=none)
candidateAssociationNotions → forAll( n.mentions → exists(s: Symbol/ class.AllInstances → exists(c/c.name=s.name))
          implies A.oclIsTypeOf(Association) and A.connection → at(1).participant= C
          and A.connection → at (2).participant= c1)
```

(*) To simplify the OCL expression we have omitted the expression to define c1 in the right side of each implies expression.

of) or top-down verbs (is, may be, may be classified as, classifies as).

- **Aggregation relationships:** in the notions of the LEL symbol considered as container, verbs of the type "component_composition_verb" must appear "to consist / to contain / to include / to form, to compose, to divide" (these three last in passive voice[1]). In the notions of the "component" symbol, verbs of the type content_composition_verb must appear "it is part, it belongs, it is a component, it is included," among others. As it is not possible to automatically distinguish between an aggregation or a composition relationship, the transformation rule defines the relationship as an aggregation.
- **Association relationships:** any relationship between LEL symbols that does not represent a relationship of the previous types, represents an association. The general verb that appears in the notion is taken as the name of the association.

We illustrate this transformation rule in next subsection.

CIMTool

CIMTool implements the derivation strategy allowing the integration of natural language requirements models with MDA framework (Figure 8).

CIMTool was developed with ´Oracle JDeveloper 10g´ in ´Java 2 Runtime Environment Standard Edition v1.4.2´. As we can see in Figure 9, CIMTool has the following input and output files:

- **XRD file:** An input file with the LEL and scenarios models for a given case study following an XRD specification. The XRD specification is a specialization of XML that integrates the LEL and Scenarios models (Leonardi et al, 2005).
- **XML dictionary file:** Is an input file with a dictionary of the language used to describe the requirements models.
- **XMI file:** An output file with a XMI specification of the diagram class representing the CIM. The file has the XMIv1.2 (OMG XML Metadata Interchange) format; therefore it is possible to integrate the output of CIMTool with other CASE tools, for example, Poseidon

Figure 8. The derivation strategy in the context of MDA

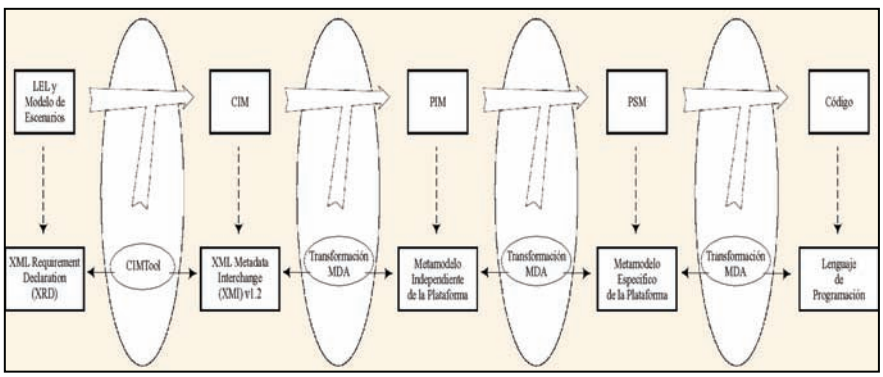


Figure 9. CIMTool inputs and outputs

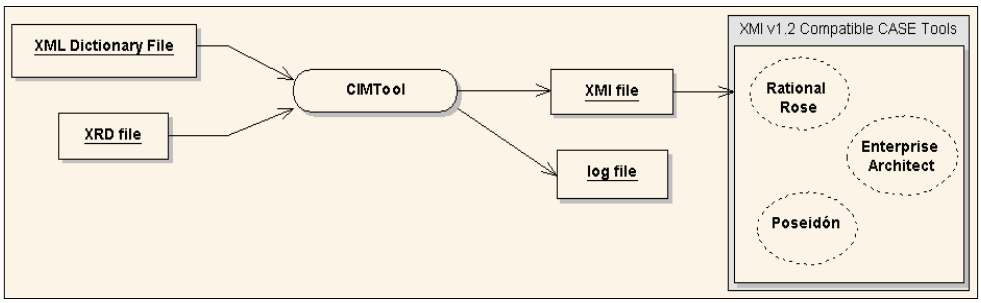
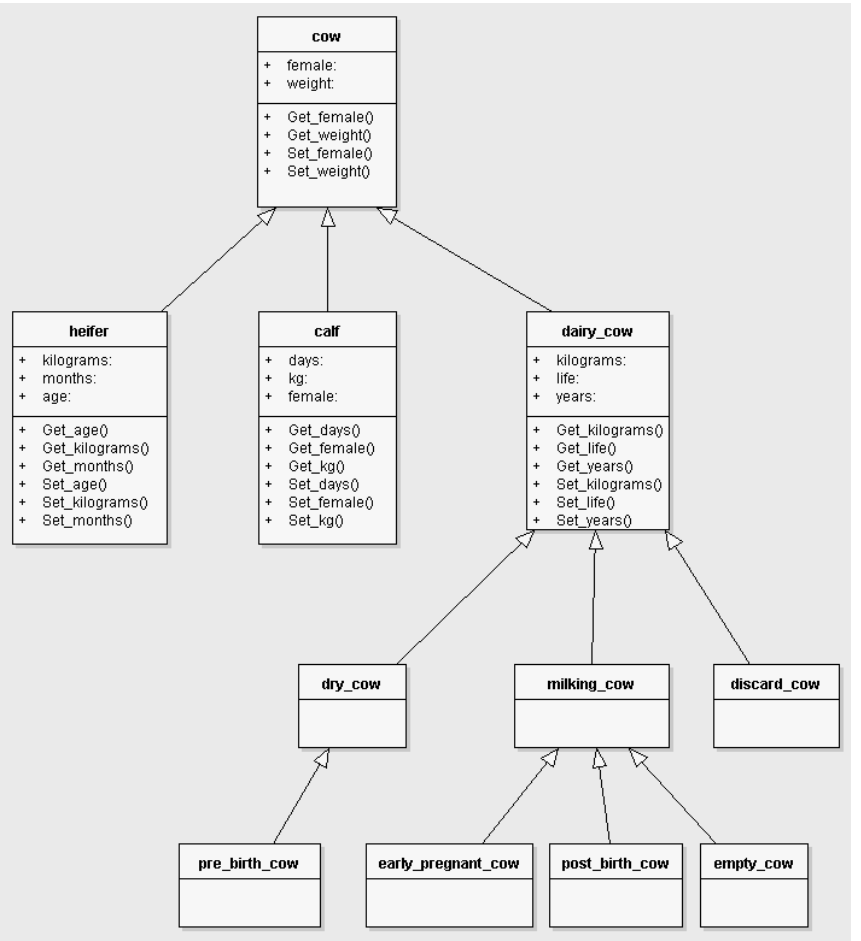


Figure 10. XRD description of LEL symbols

```
<symbol classification="object" id="cow">
  ...
  <notion>
    <text>It may be a calf, a heifer, or a dairy cow.</text>
    <symbol_ref symbol_id="calf" verb="may be"/>
    <symbol_ref symbol_id="heifer" verb="may be"/>
    <symbol_ref symbol_id="dairy_cow" verb="may be"/>
  </notion>
  ...
</symbol>
...
<symbol classification="object" id="calf">
  <notion>
    <text>It is a cow of less than 12 months age.</text>
    <symbol_ref symbol_id="cow" verb="is a"/>
  </notion>
  ...
</symbol>
```

Figure 11. Inheritance relationship defined by TRR



Box 2. TRC2 OCL specification

```
OCL Specification
SOURCE: S1: RM:: Symbol
            D: TP :: Dictionary
 TARGET: C1: UML :: Class
 SOURCE CONDITION
                          S1.classification :: object
 TARGET CONDITION
                          C1.name = S1.isIdentifiedBy → first()
                          let
                              plainNotions: Set =
                           S1.isIdentifiedBy → excludes (n/ n.mentions → notempty())
                          nounOfNotions: Set =
        plainNotions → collect(n/ D.returnNouns(n.desription)) asSet
                          at: OrderedSet=
                           C1.features → collect (f/ f.oclIsTypeOf (Attribute))
                         oper: OrderedSet =
                               C1.features → select (f/ f.oclIsTypeOf(operation))
                          in
                            at → forAll (a/nounsOfNotions →
                                  one (n: String /n= a.name))
                            oper → forAll(o/ at → one(a / o.name = "set" concat (a.name) or o.name =
                             "get" concat (a.name)))
```

Figure 5. Plot object LEL symbol and Plot class

PLOT
NOTION
It is a part of a field.
It has identification.
It has a location inside the field.
It has a size.
It has a starting date.
It has an approximated period of duration in days.
In any time it is occupied by one group.

Plot
identification size date period duration days
setIdentification() getIdentification() ...

Applying the transformation rule TRM1 to the LEL symbol shown in Figure 4, the methods described in Figure 6 are obtained.

TRM2_Transformation SubjectInformationToMethodParameter (RM, UML)

DESCRIPTION: Each behavioral response of a subject LEL symbol originates a scenario (Leite et al., 2000). The rule models actors and resources of each scenario as parameters of the method obtained by TRM1 from the behavioral response that originated the scenario. The actor referring to the subject LEL symbol in consideration is excluded.

For example, for each method previously defined by TRM1 (Figure 6), parameters are identified considering the scenarios involved (Mauco, 2004). As parameters come from resources and actors, they are modeled as classes when the corresponding resource and actor is a subject or object LEL symbol (TRC1 and TRC2); for example, parameters cow and groupForm in the method assign_group_each_cow_dairy_farm (Figure 7). When the resource or the actor does not belong to the LEL, two things may happen. It may be a word that does not need a LEL entry because it belongs to the minimum vocabulary or it may represent a set. In the first case, it is modeled with a primitive class or type (parameter date, Figure 7), and in the second one no new classes are needed because the parameter is a set of a class

Community Edition v2.6 (Poseidon, Gentleware) and Enterprise Architect v4.0 (Enterprise Architect, Sparx Systems).

- **Log file:** an output file with information of the process.

Figure 10 shows a portion of the XRD input file with a partial description of LEL symbols corresponding to Cow, Dairy Cow, Heifer, and Calf. By applying TRR rule, CIMTool produces an XMI based inheritance relationship. Figure 11 shows the inheritance relationship visualized by Enterprise Architect v4.0.

It is important to remark that the output XMI file representing the CIM must be analyzed by software engineers who will adjust the results obtained after the application of the transformation rules.

FUTURE TRENDS

In order to complete the transformation process, we must define transformation rules for the business rule model based on the manual heuristics proposed in Leonardi (2003). We must also define transformation rules to include the dynamic aspects of the business models. As the strategy deals with natural language oriented models, it would be necessary to incorporate linguistic approaches to achieve a better processing of the information (Diaz et al., 2004; Juristo et al., 2000).

Traceability plays a crucial role. In an MDA specification, CIM requirements should be traceable to the PIM and PSM constructs that implement them and vice versa (Miller et al., 2003). The transformation process we have proposed allows the trace between the source and the target. However, we want to enhance this mechanism by defining another complementary and independent model to capture and represent the relationships created by the application of the transformation rules, as the one proposed in Leonardi (2003).

The success of MDA depends on the definition of transformation languages and tools that make a significant impact on full forward engineering processes. MDA is still evolving and many products claim to be complaint with it. The new results will impact both the process and the tool presented in this article, therefore, they will be adapted to incorporate these results.

CONCLUSION

We present a strategy and its corresponding tool to derive a CIM starting from natural language oriented requirements models. The application of the transformation rules by CIMTool provides a systematic and consistent way of defining a CIM in MDA framework. Though a manual derivation (Leonardi, 2003), generally produces a better and more accurate model definition, transformation rules are a starting point to deal with the great amount of requirements information. In addition, we also take advantage of all the time and effort the definition of requirements and business models consumes, thus reducing the gap between requirements and other development models.

Though the requirements models presented in this article have a precise structure, the use of natural language allows the same semantics to be usually expressed with many different natural language sentences. CIMTool takes always the same decision about certain modeling issues, loosing, in some cases, the real meaning of the essential concept. As a consequence, this strategy unavoidably needs software engineer's participation in order to adjust the CIM obtained after the application of the transformation rules.

Natural language oriented models are widely used in requirements modeling due to their well-known advantages (Francis, 2006; Leite et al., 2000). This kind of requirements models have to be reinterpreted by software engineers into a more formal design on the way to a complete implementation. Therefore, a semiautomatic transformation to map their knowledge into conceptual object models would be really useful. Our proposal is a first step into this direction, aligning with the MDA framework.

REFERENCES

Alencar, F., Pedroza, F., Castro, J., Silva, C., & Ramos, R. (2006). XGOOD: A tool to automatize the mapping rules between I* framework and UML. *IX Workshop Iberoamericano de Ingeniería de Requisitos y Ambientes de Software*. Argentina.

Birol, B. (2006). How to align IT with the changes using UML and according to BMM. *Journal of Object Technology*, *5*(2), 85-102.

Díaz, I., Pastor, O., Moreno, L., & Matteo, A. (2004, May). Una Aproximación Lingüística de Ingeniería de Requisitos para OO-Method. In *Proceedings of IIV Workshop Iberoamericano de Ingeniería de Requisitos y Desarrollo de Ambientes de Software* (pp. 270-281). Perú.

Sparx Systems. (2006). *Enterprise architect.* Retrieved July, 2006, from http://www.sparxsystems.com.au/ea.htm

Francis, J. (2006). *Managing BPM. The normal modeler*. Retrieved July 2006, from http://www.bprtrends.com

Gentleware (2006). *Poseidon.* Retrieved July 14, 2006, from http://www.gentleware.com

Jackson, M. (1995). *Software requirements & specifications*. ACM Press, Addison Wesley.

Johnston, S. (2006). *Rational UML Profile for business modeling. IBM Rational Developer Works*. Retrieved November 2006, from http://www-128.ibm.com/developerworks/rational/library/5167.html

Juristo, N., Moreno, A., & López, M. (2000). How to use linguistic instruments for object-oriented analysis. *IEEE Software, 17*(3), 80-89.

Kleppe, A., Warmer, J., & Bast, W. (2003). *MDA explained: The model driven architecture™: Practice and promise*. Addison Wesley.

Leonardi, M. C. (2005). Business modeling with client-oriented requirements strategy. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology I-V* (pp. 339-344). Hershey, PA: IRM Press.

Leonardi, M. C. (2003). Enhancing RUP business model with client-oriented requirements models. In L. Favre (Ed.), *UML and the Unified Process* (pp. 80-115). Hershey, PA: IRM Press.

Leonardi, M. C., & Mauco, M. V. (2004). Integrating natural language oriented requirements models into MDA. In *Proceedings of the VII Workshop on Requirements Engineering* (pp. 65-76). Argentina.

Leonardi, M. C., Mauco, M. V., & Leoni, H. (2005). CIM-Tool: Una herramienta para la definición de un diagrama de clases UML. *CACIC 2005 - XI Congreso Argentino de Ciencias de la Computación*, Argentina.

Leite, J.C. S. P., Hadad, G., Doorn, J., & Kaplan, G. (2000). *A Scenario Construction Process Requirements Engineering Journal*, *5*(1), 38-61.

Mauco, M. V. (2004). *A technique for an initial specification in RSL*. Master thesis, Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina.

Meservy, T., & Fenstermacher, K. (2005). Transforming software development: An MDA road map. *IEEE Computer* (pp. 38-44), September.

Miller, J., & Mukerji, J. (2003). *MDA Guide Version 1.0.1*. Retrieved July 2006, from http://www.omg.org/docs/omg/03-06-01.pdf

Object Management Group (2003). *Unified modeling language specification. V.1.5*. Retrieved December 2005, from http://www.omg.org/cgi-bin/doc?formal/03-03-01

Object Management Group (2006). *OMG model driven architecture.* Retrieved July 2006, from http://www.omg.org

Object Management Group (2006). *OMG XML metadata interchange: (XMI) specification*. Retrieved July 2006, from http://www.xmi.org

Sommerville, I. (2005). Integrated requirements engineering: A tutorial (2005). IEEE Software. *IEEE Computer Society Press, 22*(1), 16-23.

Vasconcelos, A., Caetano, A., Neves, J., Sinogas, P., Mendes, R., & Tribolet, J. (2001). A framework for modeling strategy, business processes, and information systems. In *Proceedings of EDOC '01: The 5th IEEE International Enterprise Distributed Object Computing Conference* (pp. 69-80).

White, S. A. (2004). *Introduction to BPMN*. Retrieved July 2006, from http://www.bpmn.org/Documents/Introduction%20to%20BPMN.pdf

KEY TERMS

Business Model: An abstraction of how a business works. It provides a simplified view of the business structure that will act as the basis for communication or innovations and define the information systems requirements that are necessary to support the business.

CIM: MDA computation independent model that describes the business model of the organization. This model is independent of how the system is implemented.

MDA: An approach to IT system specification that separates the specification of functionality from the specification of the implementation of that functionality on a specific technology platform.

Requirements Engineering: It comprehends all the activities involved in eliciting, modeling, documenting, and maintaining a set of requirements for a computer-based system. The term "engineering" implies that systematic and repeatable techniques should be used to ensure that system requirements are consistent, complete, and relevant.

Stakeholder: People or organizations who will be affected by the system and who have a direct or indirect influence on the system requirements. They include end-users of the system, managers and others involved in the organizational processes.

Universe of Discourse: The overall context in which the software will be developed and operated.

ENDNOTE

[1] We decided to eliminate the verbs to have and to posses as indicators of aggregation relationships since, from our experience, they are commonly used by stakeholders to describe properties of concepts.

Integration of MES and ERP

Vladimír Modrák
Technical University of Košice, Slovakia

INTRODUCTION

In the present manufacturing paradigm, manufacturing execution systems (MESs) play a significant role in effective manufacturing management. Offered software solutions simultaneously close the gap between Enterprise Resource Planning (ERP) systems and production equipment control or SCADA (Supervisory Control and Data Acquisition) applications. Current ERP systems usually contain modules for material management, accounting, human resource management and all other functions that support business operations. In the past years, the role of ERP has been extended to cross-organizational coordination. Nowadays, as optimization of production activities is increasingly topical, a cooperation of ERP and MES becomes a serious concern of manufacturing managers.

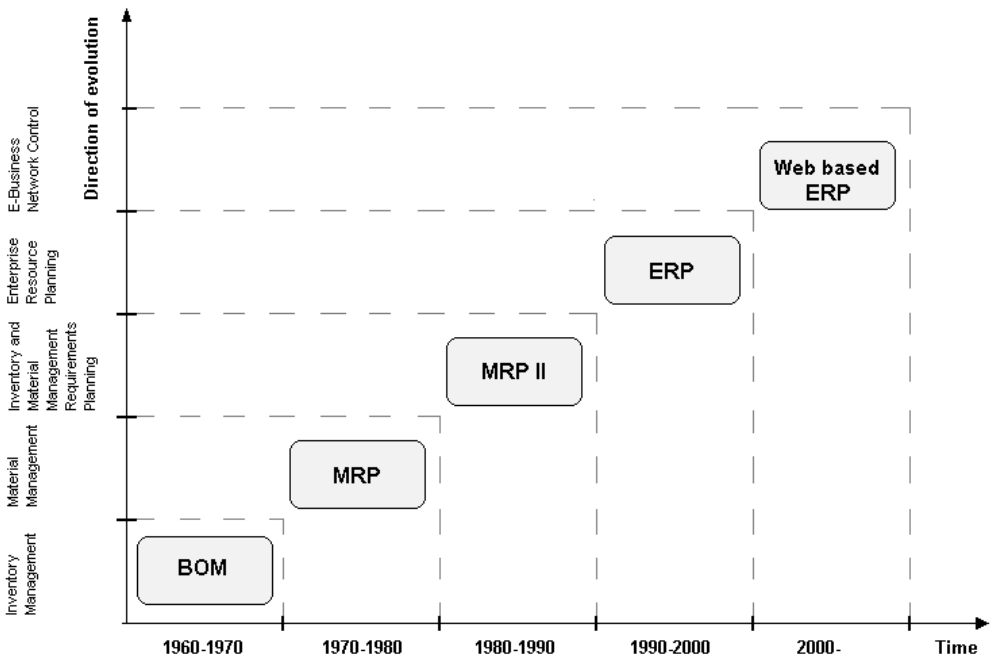
BACKGROUND OF ERP AND MES EVOLUTION

From a historical perspective, the infiltration of information technology into manufacturing technology was conditioned by the development and advancement of host mainframe computing in the 1950s and 1960s. It gave manufacturers the ability to capture, manipulate, and share information and automate calculation and analysis in order to support design of increasingly complex and capable products. Simultaneously, in the framework of manufacturing management, an inventory control took on great importance and most of the software in the 1960s was developed for this purpose. Typically, inventory control was handled by a tool called BOM (bill of materials) processors, which were used as a means to represent process plans. The focus shifted in the 1970s to Material Requirement Planning (MRP) as the complexity of manufacturing operations increased. This managerial instrument enabled financial managers both to view and control their business processes much more closely. The tools to automate business processes were enhanced by adding further functionalities to meet the increased requirements. Subsequently, in the 1980s the term Manufacturing Resources Planning (MRP II) became popular. An MRP II presented extension of MRP functions to achieve integration of all aspects of the planning and control of the personnel, materials and machines (Kimble & McLoughlin, 1995). Following, solutions that are marked by acronym ERP were performed in the early 1990s. An ERP system can be defined as an integrated information processing system supporting various business processes such as finance, distribution, human resources and manufacturing (Choi & Kim, 2002). The newest version ERP II has been much publicized by the Gartner group (Mohamed & Fadlalla, 2005). Fundamentally, ERP II signals a shift in traditional ERP applications from focusing on internal data gathering and management process information to partners, vendors and customers externally via the Web (Farver, 2002). The overall view on evolution of ERP system is shown in Figure 1. Initially, this concept attained a huge popularity among manufacturers, but as the scope of managed systems increased, the ERP system was not suitable for controlling activities on the shop floor level. For this purpose, a new tool of manufacturing management called Manufacturing Executive System was evolved and utilized during the 1990s. There is more interpretation of MES depending on different manufacturing conditions, but the common characteristic to all is that an MES aims to provide an interface between an ERP system and shop floor controllers by supporting various "execution" activities such as scheduling, order release, quality control, and data acquisition (MESA #6, 1997). In a context of the MES development and deployment, it is important to point out that Manufacturing Execution Systems were originally designed to provide first-line supervision management with a visibility tool to manage work orders and workstation assignments. Consecutively, MES expanded into the indispensable link between the full range of enterprise stakeholders and the real-time events occurring in production and logistics processes across the extended value chain (McClellan, 2004).

INTEGRATION OF ERP AND MES

Manufacturing execution systems besides their typical functions were developed and used also as the interface between ERP and process control, because it was generally recognized that ERP systems weren't scalable. The seamless connections often required skilled coding to connect to ERP and process control systems (Siemens Energy & Automation, Inc., 2006). Today, the availability of Web-based XML communications successfully bridges the gaps between MESs and ERP systems. Built on XML, the B2MML (business-to-manufacturing markup language) standard specifies accepted definitions and data formats for informa-

Figure 1. The evolution of ERP systems

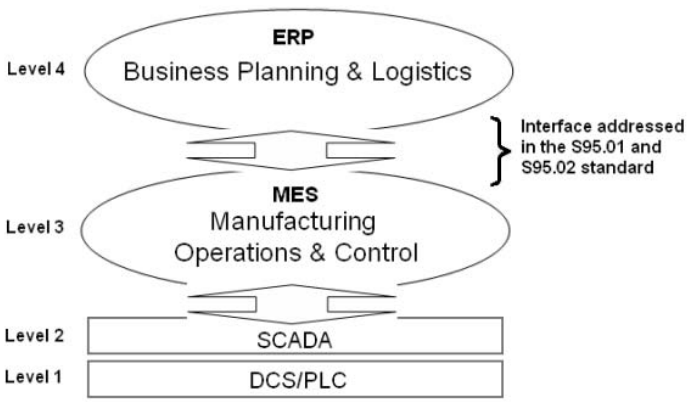


tion exchange between systems, and facilitates information flow and updates between ERP and manufacturing execution systems. It also instigated redefinition of the role of the MES. The ISA SP-95 model (see Figure 2) breaks down business to plant floor operations into four levels.

Levels 1 and 2 include process control zone. The MES layer consists of manufacturing management, dispatching production, detailed production scheduling, reliability assurance, and so forth. A point of debate about MES functionalities is connected with more aspects like different types of manufacturing and others (Modrák, 2005). Level 4 corresponds to the business planning and logistics.

The goal of ISA-95 standard was to reduce the risk, cost and errors associated with implementing interface between ERP and MES. The ISA-95 "Enterprise - Control System Integration" is a multipart series of ANSI/ISA standards that define the activity models and interfaces between manufacturing functions and other enterprise functions. Parts 1 (Models and Terminology), parts 2 (Objects Attributes) and part 5 (Business to Manufacturing Transactions) define the exchange of production data between business and plant systems. B2MML provides a schema implementation of the ANSI/ISA-95 and represents an independent technology implementation of this standard. B2MML has been developed

Figure 2. Position of MES in the hierarchy of IT systems



by The World Batch Forum (WBF) and adopted by players such as SAP and Wonderware. Coupled together, B2MML and ISA-95 permit designers to define the data mapping using a standardized, common terminology and models that can be carried over to the B2MML XML vocabulary (Zurawski, 2007).

The mentioned standards, and other ISA standards, significantly facilitate the implementation of integrated manufacturing systems. It is aimed to integrate ERP systems with control systems like DCS and SCADA. To support batch control level optimization, the standard S88.01 (ANSI/ISA, 1995) has been developed. It provides standard models and terminology for the design and operation of batch control systems. At the control level the key attribute is integration of all process information into one place. For this purpose, both programmable logic controllers and SCADA software are ordinarily used.

An important function of MES is to provide feedback to ERP with the aim to adjust their scheduling data and algorithms in a more realistic manner. In this connection, it is necessary to take in consideration certain aspects of incompatibility. Originally, ERP systems have been developed more for financial managers than manufacturing managers. Moreover, the data flow frequency in ERP system is lower then at the MES level. Accordingly, it would not make sense to feed a quantity of output data available at shop floor level into the ERP system and a reversal. An effort to do so often involved large amounts of manual data entry. Therefore, one important role of manufacturing management optimization by MES may be seen in the reduction of manual data entry on the boundaries of information system layers. Basically, it's relevant in this case when enterprise information and control systems are designing as one complex. Then, expected reductions of manual activities may be compromised by integration of system levels.

Supplementary integration between MES and ERP is seldom uncomplicated. One manner to get around presumable complications is by execution of whole business process modeling. The sense of business process modeling of whole enterprise is to make new organizations of work more feasible than in the past. That is why business process modeling and improvement becomes a key issue of successful implementation of any ERP or MES strategy.

FUTURE TRENDS

As it is observed, production planning activities have become more complex and therefore need to be in principle optimized. Manufacturing Execution Systems, which are positioned between the Enterprise Resource Planning and control systems levels, have significant potential to be effectively used to optimize business processes on the shop floor. Besides that fact, MES are being viewed as critical in getting the most value out of existing investments in automation. Speaking about MES's role, a frequent interest of manufacturers concerns a balanced scale of MES functionalities. As mentioned earlier, it depends on more factors. For instance, when an existing ERP system contains factory floor control functionality, then the functionality model of MES has only supplement character. The scope of functionality is influenced also by changes in using automated identification (AID) technologies, which can have positive impact on the plant floor optimization. To this category of progressive AID technologies that have the potential to change the future in manufacturing, undoubtedly belongs Radio-Frequency Identification (RFID). Mass use of this technology can bring significant rationalizations in the manufacturing automation in the future. This tendency indirectly confirms such IT players as Oracle, SAP, Microsoft and IBM, which are all accelerating efforts to meet the RFID challenge (Rockwell Automation, 2004). The rules concerning manufacturing execution such as control, scheduling, routing, tracking, and monitoring must all be modified to collect as well as respond to new RFID-information. It is predicted that RFID will complement existing MES efforts in genealogy tracking. In this connection, according to Rockwell Automation, RFID could be used in varying scales, either locally or across the entire facility to provide visibility into incoming raw materials, work in process, production sequencing, packaging, palletizing, and warehousing operations as well as in the supply chain management.

CONCLUSION

In reality, the majority of enterprise information and control systems on different levels are developed and operated on incompatible technologies and based on heterogeneous architectures. On the other hand, today's ample standardized communication tools achieve the successful integration of MES and ERP. From "Siemens Energy & Automation, Inc.' view point of view" an integrated system will show real returns: from the ability to monitor—in real time—key performance indicators on productivity, quality, yields, and throughput; to managing inventory locations and raw materials; through remediation processes to isolate and or rework nonconforming products." One of expected directions of MES and ERP integration is synthesis of MES and ERP systems. It is hopefully in the future to be able to simplify implementation issues of information system integration.

REFERENCES

ANSI/ISA S88.01. (1995). *Batch control part 1: Models and terminology*. International Society for Measurement and Control (ISA).

Barkmeyer, E., Denno, P., Feng, S., Jones, A., & Wallace, E. (1999). *NIST response to MES request for information* (pp. 1-24). NISTIR 6397, National Institute of Standards and Technology.

Bradley, J. (2005). Balancing risks and rewards of ERP. *Encyclopedia of information science and technology (I)* (pp. 205-210). Hershey, PA: Idea Group Reference.

Choi, B.K., & Kim, B.H. (2002). MES (manufacturing execution system) architecture for FMS compatible to ERP (enterprise planning system). *International Journal of Computer Integrated Manufacturing, 15*(3), 274-284.

Daneels, A., & Salter, W. (1999). What is SCADA? In *Proceedings of the International Conference on Accelerator and Large Experimental Physic Control systems*, Trieste, Italy, (pp. 339-343).

Falco, J., Stouffer, K., Wavering, A., & Proctor, F. (2004). *IT security for industrial control systems* (pp. 1-16). National Institute of Standards and Technology, Gaithersburg, USA. Retrieved May 27, 2008, from http://www.isd.mel.nist.gov/documents/falco/ITSecurityProcess.pdf

Farver, D. (2002). 2 ERP or ERP 2. *Journal of Business Innovation*. Retrieved May 27, 2008, from http://www.agilebrain.com

Hwang, M.I. (2005). Enterprise resource planning and systems integration. *Encyclopedia of information science and technology (II)* (pp. 1083-1088). Hershey, PA: Idea Group Reference.

Kimble, C., & McLoughlin, K. (1995, March). Computer-based information systems and managers' work. *New Technology, Work and Employment, 10*(1), 56-67.

McClellan, M. (2004, June 12). *Execution systems: The heart of intelligent manufacturing*. Intelligent Enterprise.

MESA #6. (1997). *MES explained: A high level vision* (White Paper 6). Pittsburgh, PA: Manufacturing Execution Systems Association.

Modrák, V. (2005). Functionalities and position of manufacturing execution systems. *Encyclopedia of information science and technology (II)* (pp. 1243-1248). Hershey, PA: Idea Group Reference.

Mohamed, M., & Fadlalla, A. (2005). ERP II: Harnessing ERP systems with knowledge management capabilities. *Journal of Knowledge Management Practice, 6*, 1-13.

Rockwell Automation. (2004). *RFID in manufacturing* (White paper). Retrieved May 27, 2008, from www.rockwellautomation.com/solutions/rfid/get/rfidwhite.pdf

Siemens Energy & Automation, Inc. (2006). Why integrate MES and ERP? Because you can't afford not to. *Siemens white paper* (pp. 1-8).

Zurawski, R. (2006). *Integration technologies for industrial automated systems*. CRC Press, Taylor & Francis Group.

KEY TERMS

Distributed Control System (DCS): A supervisory control system typically controls and monitors set points to subcontrollers distributed geographically throughout a factory (Falco, Stouffer, Wavering, & Proctor, 2004).

Enterprise Resources Planning: Configurable enterprise software that integrates business process across functions (Hwang, 2005).

Manufacturing Execution System: A collection of hardware/software components that enables the management to control production activities from order launch to finished goods. While maintaining current and accurate data, a MES guides, initiates, responds to and reports on plant activities as they occur. MES provides mission-critical information about production activities to decision support processes across the shop floor level of manufacturing management (Barkmeyer, Denno, Feng, Jones, & Wallace, 1999).

Manufacturing Resources Planning II: Extends MRP by addressing all resources, in addition to inventory. MRPII links material requirements planning with capacity requirements planning, avoiding over- and under-shop loading typical with MRP (Bradley, 2005).

Programmable Logic Controller (PLC): A small industrial computer used in factories originally designed to replace relay logic of a process control system and has evolved into a controller having the functionality of a process controller (Falco et al., 2004).

Scalability: Understood as the ability to incorporate into the existing system additional resources or meet diverse quality requirements.

Supervisory Control and Data Acquisition System (SCADA): It is a part of the control system, but focused on the supervisory level. As such, it is a purely software package that is positioned on top of hardware to which it is interfaced, in general via programmable logic controllers (PLCs) or other commercial hardware modules (Daneels & Salter,1999).

Integrative Document and Content Management Solutions

Len Asprey
Practical Information Management Solutions Pty Ltd., Australia

Michael Middleton
Queensland University of Technology, Australia

INTRODUCTION

Developments in office automation, which provided multiple end-user authoring applications at the computer desktop, heralded a rapid growth in the production of digital documents and introduced the requirement to manage capture and organization of digital documents, including images. The process of capturing digital documents in managed repositories included metadata to support access and retrieval subsequent to document production (D'Alleyrand, 1989; Ricks, Swafford & Gow, 1992).

The imperatives of documentary support for workflow in enterprises, along with widespread adoption of Web-oriented software on intranets and the Internet World Wide Web (WWW), has given rise to systems that manage the creation, access, routing, and storage of documents, in a more seamless manner for Web presentation. These content management systems are progressively employing document management features such as metadata creation, version control, and renditions (Megill & Schantz, 1999; Wiggins, 2000), along with features for management of content production such as authoring and authorization for internal distribution and publishing (Addey et al., 2002; Boiko, 2002; Hackos, 2002; Nakano, 2002).

If business applications are designed taking into account document and Web content management as integral constructs of enterprise information architecture, then the context of these solutions may be an integrative document and content management (IDCM) model (Asprey & Middleton, 2003). As the name implies, the IDCM model aspires to combine the features of a document management system with the functionality of Web content management. An integrative business and technology framework manages designated documents and their content throughout the continuum of their existence and supports record-keeping requirements.

The IDCM model supports system capabilities for managing digital and physical documents, e-mail, engineering and technical drawings, document images, multimedia, and Web content. These systems may be deployed individually to address a specific requirement. However, due to the volume and varied formats of important documents held in digital format, these systems are often deployed collectively based on a strategic IDCM approach for better managing information assets. An organizational approach to IDCM supports enterprise knowledge strategies by providing the capability to capture, search, and retrieve documented information.

SCOPE

IDCM depends upon effective integration of organizational systems that together are used for managing both digital and physical document types. The scope of this management is across all stages of document lifecycles. It includes provision for distribution of the document content over intranets and the Internet.

Features of enabling IDCM technologies are described in the following section. The technologies may be differentiated into those with core capabilities and supporting technologies.

Core capabilities are: document management; e-mail management; drawing management; document imaging; Web content management; enterprise report management; and workflow. Supporting technologies include: Web services; database management systems; digital signatures; portals; universal interfaces; and network management.

Significant issues that need to be addressed with respect to IDCM solutions include the provision of seamless functionality that may be employed across different capabilities so that currency, integrity, and authority are managed effectively. These in turn must be complemented by user interfaces that provide stylistic consistency and that are augmented by metadata that enhances retrieval capabilities through the supporting technologies.

The following section itemizes the types of features that are required.

SYSTEM FEATURES: CORE TECHNOLOGIES

Document Management

An encompassing approach to document management sees documents within a framework that supports integrity, security, authority, and audit, and that are being managed so that effective descriptions of them are used to support access, presentation, and disposal (Bielawski & Boyle, 1997; Wilkinson et al., 1998). In this context document management applications implement management controls over digital and physical documents. The general capabilities of a document management application are:

- *Document production and capture*—interface with common office productivity software.
- *Classification*—support business classification schemes (e.g., folder structures, document properties).
- *Metadata*—capture of properties that describe document.
- *Check-in/checkout*—maintain document integrity during editing.
- *Version control*—increment versions of document to support integrity.
- *Complex relationships*—manage links and embedded content within digital documents.
- *Security*—implement user/group access permission rights over documents.
- *Document lifecycles*—manage the transition of document states through pre-defined lifecycles.
- *Integrated workflow* to automate review and approvals; controlled distribution of documents.
- *Search and information retrieval*—search metadata or text within documents, or both.
- *Viewing*—view documents in native application or using integrated viewer.

These should be associated with recordkeeping features such as disposal scheduling and archiving.

E-MAIL MANAGEMENT

The growth in e-mail has brought a high demand for solutions that allow enterprises to manage e-mails that have value to the business. The IDCM model offers two types of capabilities:

- *Direct capture*—These applications are often referred to as e-mail archiving applications.
- *End-user capture*—These capabilities are typically offered as a module within document management systems.

Direct capture or archiving facilities intercept incoming and outgoing e-mail. They operate by taking a copy of incoming and outgoing messages that are managed by the e-mail messaging system, and use customized business rules to extract e-mail that may not have a business context. Unwanted e-mail such as spam, or that received from news lists or information bulletins, can be eliminated.

These systems may feature auto-categorization based on metadata such as that contained in e-mail message headers, and possibly also within attachments. Categorization can also occur using the content and context of e-mail by applying techniques such as learning by example from previously processed e-mail. These types of solutions might be valuable for capturing statistical information differentiated by the types of requests made by customers. For example, statistics can aid call centers to monitor turnaround timeframes for responding to e-mail requests, or undertake trend analysis.

Search options include the capabilities to search messages and text attachments. Depending on the capabilities of the system, searches might be invoked from an e-mail client, desktop client application, or Web browser.

Some systems are able to apply rules defined in disposal authorities so that e-mails are purged from the system within a legal framework. In some cases, different retention schedules can be applied to specific categories of e-mail.

End-user capture facilities are adopted by some enterprises to save relevant sent and received e-mails that evidence business transactions into an e-mail management repository, such as a document management system, leaving it up to the user to identify e-mails that need to be saved according to organizational guidelines.

The document management system would need then to integrate effectively with the existing enterprise e-mail client software. This capability would enable end-users to save e-mails and/or attachments to the managed repository, automatically derive metadata from the header of e-mail messages, add custom metadata, and store the e-mail and attachment/s (where appropriate) as a digital record.

DRAWING MANAGEMENT

Many systems for registering or managing drawings have been developed independently of more generic approaches to document management. They may include information systems that enable users to register or index physical drawings in a database, along with generation of transmittals for issue of new documents, and management of the distribution of revisions to drawings and technical documents.

A drawing management system may be differentiated from a registry system in that the software implements automated management controls over the digital drawing objects maintained within a vault-like repository. This capability evolved to support the capture and management of drawings created by Computer Aided Design (CAD) packages. Functionality should include base capabilities such as:

- Integration with CAD tools for capturing electronic drawings.
- Automated features for drawing revision control and revision numbering.
- Management of electronic and hardcopy drawings, technical specifications, and manuals.
- Management of parent-child relationships between multiple drawings.
- Registration and tracking of physical copies of controlled drawings.
- Management of incoming and outgoing transmittals.
- Electronic document review and authorization using integrated workflow.
- Provision of viewing, red line, markup, and annotation functions.
- Maintenance of history logs and audit trails.

Extended capabilities may include automation of drawing numbering, synchronization of digital and physical drawing objects, synchronization of title block and metadata registration and updates, and management of drawing status during engineering change lifecycle transitions.

Drawing management capabilities may be provided by a dedicated drawing management system; as unified functionality within a document management system; as an inbuilt module of an Enterprise Resource Planning (ERP) system, maintenance management system, or similar; or an integral component of a document management application.

DOCUMENT IMAGING

Imaging systems have evolved from the principles of film-based imaging and may now be characterized in two groups for document imaging, these being (a) film-based imaging (micrographics) and (b) digital imaging systems.

In film-based imaging, micrographics technology is used to capture images of physical documents on microfilm, so that the images may subsequently be viewed using a reader, and printed if required. In digital imaging, images of physical documents are captured in a digital file format, with subsequent viewing or printing from the image format.

Digital imaging systems may be differentiated as desktop (ad hoc scanning), workgroup (shared tools in network), or production (high volume, diverse type). IDCM normally implies a workgroup or production environment. Capabilities offered include image manipulation functions such as:

- Capture of hardcopy documents into digital format.
- Capability for scanning and conversion of different sizes, sides (duplex scanning), physical orientation, and physical structure of documents.
- Managing multi-page images as a single entity (e.g., multi-page TIF file).
- Images may be saved to specified file formats. For example, a document might be saved in PDF or JPG for publishing on a Web server, or as a multi-page TIF for viewing/transmission.
- Support for a range of resolution, contrast, threshold, and size settings to meet the diverse requirements of document capture.
- Capture of color and/or grayscale images to suit forms processing and other applications (e.g., colored contour maps).
- Despeckling/deskewing and border removal.
- Multi-level registration capabilities, including batch-, folder-, envelope-, and document-level indexing.

Imaging systems are often integrated with recognition systems to facilitate capture and retrieval. These include technologies for automatically capturing data encoded in barcodes, integration with optical character recognition (OCR and ICR) technologies to enable text information to be extracted from scanned images, and integration with optical mark recognition (OMR).

WEB CONTENT MANAGEMENT

IDCM has the capability to provide a managed environment for the processes associated with publishing Web content. It has been said that a content management system is a concept rather than a product (Browning & Lowndes, 2001). This adds weight to analysis of it within the context of an IDCM model, where documents and their content may be considered more broadly than in terms of Web presentation. Document creation, management, and utilization can thus be undertaken with reference to business requirements and workflow of business processes.

Typically, functionality is characterized in terms of content creation, presentation, and management (Arnold, 2003; Robertson, 2003). Increasingly this functionality is seen to be employed within a unified content strategy for an enterprise (Rockley, Kostur & Manning, 2003).

Content creation functionality includes separation of presentation and content, utilization of elements of documents such as illustrations in different contexts, and continuation of associations between pages after restructuring. Metadata

support should also be available, and markup should be transparent to the content creator.

Presentation elements include multiple formats for distribution of internal material such as manuals and business forms over intranets, and for external material such as marketing information and application forms. Other features expected are template availability through style sheets, integration of multiple formats as compound documents, provision of alternative renditions, and personalization of display according to user profiles.

Management features include version control and integrity maintenance among multiple users, and associated security procedures and audit trails. Managed interfaces to other subsystems should provide for dynamic provision of content to pages so that current data can be presented in validated form within compound documents. There should also be utilization of workflow for accommodating distributed users, content review, and approval processes.

These capabilities are shared at least in part with other IDCM systems' functionality. The IDCM environment has the capability to manage Web content within a continuum that includes initial document creation processes, potentially in a distributed environment, through to managed archiving of content.

ENTERPRISE REPORT MANAGEMENT (COLD)

As digital media have been developed, businesses with high volumes of management information reporting have made increasing use of Enterprise Report Management (ERM). These capabilities enable organizations to capture reports from business application databases and store them in a managed repository, to reduce printing, improve information accessibility, and maintain records.

Technologies that provide support for ERM include output reports in a range of formats. Examples of these include text-based digital format (e.g., XML) that is stored and searched via a document management application, and image format such as TIF or JPG that can be captured and accessed via a document management or imaging application. Reports may be captured on optical disk, using a capability known as Computer Output Laser Disk (COLD), which stores digital reports and enables data to be represented with graphical overlays to facilitate interactive communications.

General ERM capabilities are defined as follows:

- Capture of digital report objects to managed repository (document management, imaging, or COLD application).
- Utilization of indexing capabilities for capturing metadata relevant to the report.
- Support for inquiry and retrieval of metadata or report contents (where applicable).
- Management of database growth to support performance.
- Support for repository that can include different data objects.
- Support for document integrity.
- Control of processes through workflow.
- Provision of extraction and use of parts of reports.
- Support for high-volume printing.
- Management of security—user authentication, group and user levels.

WORKFLOW

Workflow management systems are designed to automate and implement controls over a diverse range of business processes, from the initiation of a process through to execution of all tasks, and process closure. The need for transparent interfaces between the workflow management system and IDCM is vital to maintain the integrity of documents or Web content files during their transition through a workflow process.

There are a number of technology options for enterprises that are seeking a workflow management capability. The most suitable workflow engine will depend on the nature and complexity of the requirement and the functionality supported by the workflow technology options. Options for workflow within the context of IDCM are:

- *Messaging/collaboration systems/workflow:* IDCM should support ad hoc and cooperative review and production of documents and reports, and it may be desirable to support integration with electronic forms for recordkeeping purposes.
- *Embedded workflow:* This capability is offered in systems such as document and Web content applications, or in application suites such as ERP systems. The host application provides inbuilt workflow for facilitating document-centric or process-centric modules.
- *Autonomous workflow:* These types are functional without any additional application software, with the exception of database and message queuing. When used in the context of IDCM, the functionality may support automation of document or Web content review and approval processes.

IDCM ARCHITECTURE

The IDCM system model should feature scalable, flexible, and extensible applications that integrate with the enterprise information architecture, enabling the system to grow with

the organization and facilitate knowledge sharing. Some architectural scenarios include:

- Scalability, flexibility, extensibility.
- Intuitive interface, preferably Web based, to facilitate usability, software upgrades, and support.
- Integrate with heterogeneous operating environments (where required).
- Implement three-tier (or "n-tier") client server architecture to facilitate Web client functionality and usability.
- Support distributed computing environment (databases, document/Web content repositories, replication services).
- Support mobile workers—access from remote sites, limited bandwidth.
- Integration with enterprise backup/recovery and business continuity regimes.

REASONS FOR UTILIZING IDCM SOLUTIONS

The business justification for implementing IDCM solutions ranges widely with respect to policy, compliance, and economics. For example, policy initiatives may include support for customer service initiatives, knowledge management (Laugero & Globe 2002), or risk reduction in relation to brand damage. Compliance with legislative requirements may include administrative requirements that support privacy and freedom of information legislation. Economic justification may include support for timely delivery of product to market, reduction in operational costs, continuous process improvement initiatives, and profit maximization strategies.

CRITICAL ISSUES OF DOCUMENT AND CONTENT TECHNOLOGIES

There are critical issues that must be managed, and it is imperative that organizations undertake risk analysis in order to identify risks and develop strategies to mitigate them. Table 1 summarizes some of the critical issues.

CONCLUSION

The IDCM model supports a business value proposition that aligns enabling systems and technology with an enterprise's strategic, tactical, and operational planning imperatives. The IDCM system architecture provides a range of enabling applications and technologies that support end-to-end business process improvement initiatives and provide a key foundation for knowledge management strategies.

Table 1. A summary of critical issues of IDCM technologies

Business Issues	Technology Issues
Executive management lacks resolve Lack of executive management engagement, both at the start and during the project, may impair outcomes.	**Inadequate infrastructure** Client, server, and network architecture needs to be adequate to optimize performance.
Inadequate planning Poor definition of scope and inadequate product and project lifecycle management may impair outcomes.	**Incorrect technology application** Inadequate business definition and failure to examine solution options results in implementation of inappropriate technology solution.
Inadequate specifications Lack of analysis and determination of requirements leads to project complications and inhibits extensibility of applications across enterprise.	**Integrity of metadata** Metadata can be abused in non-validated fields, which may create significant retrievability issues.
Mandated use may cause rejection Document management is often mandated, without appropriate consultation and analysis.	**Security** IDCM solutions contain vital documents and content files, and security should reflect importance.
Lack of process integration Mandated use is often accompanied by failure to integrate capabilities with existing processes, often meaning duplication of effort.	**System incompatibilities** Lack of proven integration capabilities may impact delivery document/content management enabled end-to-end business solutions.

REFERENCES

Addey, D. et al. (2002). *Content management systems*. Birmingham, UK: glasshaus.

Arnold, S.E. (2003). Content management's new realities. *Online*, *27*(1), 36-40.

Asprey, L. & Middleton, M. (2003). *Integrative document and content management: Strategies for exploiting enterprise knowledge*. Hershey, PA: Idea Group Publishing.

Bielawski, L. & Boyle, J. (1997). *Electronic document management systems: A user-centered approach for creating, distributing and managing online publications*. Upper Saddle River, NJ: Prentice-Hall.

Boiko, B. (2002). *Content management bible*. New York: Hungry Minds.

Browning, P. & Lowndes, M. (2001). JISC TechWatch report: Content management systems. Retrieved October 10, 2003, from *www.jisc.ac.uk/uploaded_documents/tsw_01-02.pdf.*

D'Alleyrand, M.R. (1989). *Image storage and retrieval systems: A new approach to records management*. New York: Intertext Publications.

Hackos, J. (2002). *Content management for dynamic Web delivery*. New York: John Wiley & Sons.

Laugero, G. & Globe, A. (2002). *Enterprise content services: A practical approach to connecting content management to business strategy*. Boston: Addison-Wesley.

Megill, K.A. & Schantz, H.F. (1999). *Document management: New technologies for the information services manager*. East Grinstead, UK: Bowker-Saur

Nakano, R. (Ed.). (2002). *Web content management: A collaborative approach*. Boston: Addison-Wesley.

Ricks, B.R., Swafford, A.J. & Gow, K.F. (1992). *Information and image management: A records systems approach* (3rd ed.). Cincinnati, OH: South-Western Publishing.

Robertson, J. (2003). So, what is a content management system? Retrieved March 18, 2004, from *www.steptwo.com.au/papers/kmc_what/index.html.*

Rockley, A., Kostur, P. & Manning, S. (2003). *Managing enterprise content: A unified content strategy*. Indianapolis: New Riders.

Wiggins, B. (2000). *Effective document management: Unlocking corporate knowledge*. Aldershot, UK: Gower.

Wilkinson, R. et al. (1998). *Document computing: Technologies for managing electronic document collections*. Boston: Kluwer Academic Publishers.

KEY TERMS

Content Management: Implementation of a managed repository for digital assets such as documents, fragments of documents, images, and multimedia that are published to intranet and Internet WWW sites.

Document Capture: Registration of an object into a document, image, or content repository.

Document Imaging: Scanning and conversion of hard-copy documents to either analogue (film) or digital image format.

Document Management: Implements repository management controls over digital documents via integration with standard desktop authoring tools (word processing, spreadsheets, and other tools) and document library functionality. Registers and tracks physical documents.

Drawing Management System: Implements repository management controls over digital drawings by integration with CAD authoring tools and using document library functionality. Registers and tracks physical drawings.

E-Mail Management: Implements management controls over e-mail and attachments. These controls may be implemented by direct capture (e-mail archiving software) or invoked by the end-user in a document management application.

Recognition Technologies: Technologies such as barcode recognition, optical character recognition (OCR), intelligent character recognition (ICR), and optical mark recognition (OMR) that facilitate document registration and retrieval.

Workflow Software: Tools that deal with the automation of business processes in a managed environment.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1573-1578, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Intellectual Property Protection on Multimedia Digital Library

Hideyasu Sasaki
Ritsumeikan University, Japan

INTRODUCTION

The principal concern of this article is to provide researchers and practitioners in information science and technology with legal references on the concepts, issues, trends, and frameworks of intellectual property protection regarding *multimedia digital library* in engineering manner.

Digital library is the global information infrastructure in the networked society (Borgman, 2003). A digital library, as an information system, consists of digital contents in databases and retrieval mechanisms. The right protection of digital library is a critical issue in the digital library community that demands frameworks for recouping their investment in database design and system implementation. Intellectual property law gives incentive to advance appropriate investment in database design and implementation with two types of intellectual property protection: copyright and patent (Jakes & Yoches, 1989; Junghans & Levy, 2006).

Multimedia digital contents take a variety of forms including text, images, photos, and video streams, which often commingle in multimedia digital libraries. Nevertheless, present legal studies are not satisfactory as the source of technical interpretation of the intellectual properties regarding multimedia digital libraries. The intellectual property protection of the multimedia digital libraries demands clear and concise frameworks.

BACKGROUND

In this section, we discuss two main issues on the intellectual property protection regarding multimedia digital libraries. The first issue is the copyright protection of databases to which the multimedia digital contents are stored in multimedia digital libraries. The second issue is the patent protection of the retrieval mechanisms of multimedia digital libraries.

Copyright on Multimedia Digital Libraries

U.S. Copyright Act (2005) defines that a compilation or assembling of individual contents, that is, preexisting materials or data, is a copyrightable entity as an original work of authorship. Gorman and Ginsburg (2002) and Nimmer, Marcus, Myers, and Nimmer (2006) state that a compilation is copyrightable as far as it is an "original work of authorship that is fixed in tangible form."

Multimedia digital libraries consist of multimedia digital contents which are indexed and stored in databases for appropriate retrieval operations and the retrieval mechanisms which are optimized and applied to object domains of those databases. A database of a multimedia digital library is copyrightable in the form of a component of *contents-plus-indexes* while static indexes or metadata are fixed to multimedia digital contents in a tangible medium of repository, that is, database. Static indexes or metadata represent a certain kind of categorization of the entire content of each database in a multimedia digital library (see Figure 1).

The originality on the categorization makes each database copyrightable as is different from the mere collection of its individual contents. What kind of categorization should be *original* to constitute a copyrightable compilation on the database in a multimedia digital library? The court of *American Dental Ass'n vs. Delta Dental Plan Ass'n* (1997) determined that minimal creativity in compilation sufficed this requirement of originality on databases. Any standard or framework on the requirement is not clear in the technical or engineering meanings. A uniform framework on the categorization regarding indexes or metadata of databases in multimedia digital libraries must be formulated in engineering manner.

The European Union has legislated and executed a scheme for protecting a database including its content per se, known as the sui generis right of database protection (Aplin, 2005; Reinbothe, 1999; Samuelson, 1996). That European scheme shares the same issue on the originality regarding the categorization of multimedia digital contents in databases of multimedia digital libraries.

Patent on Multimedia Digital Libraries

U.S. Patent Act (2005) defines that a data-processing process or method is patentable subject matter in the form of a computer-related invention, that is, a computer program. The computer program is patentable as far as the "specific machine … produce(s) a useful, concrete, and tangible result … for transforming … " physical data ("*physical transformation*") (*In re* Alappat, 1994).

The computer-related inventions often combine means for data processing, some of which are prior disclosed inventions. A retrieval mechanism in a multimedia digital library consists of a number of *processes*, that is, methods or means for data processing in the form of combination of computer programs. A set of programs focuses on image processing, while another set of programs operates text mining, for example.

Meanwhile, the processes in a retrieval mechanism of a multimedia digital library comprise means or components for parameter setting which is adjusted to retrieve specific kinds of multimedia digital contents, for example, images in certain domains. The problem is that which process is to realize technical advancement (nonobviousness) on its combination of the prior arts and is to be specific/enable on its parameter setting. These two issues are emerging problems in the advent of multimedia digital libraries. Uniform frameworks on the novel combination and the specific parameter setting must be formulated in engineering manner, respectively.

FRAMEWORKS FOR RIGHT PROTECTION

In this section, we outline the frameworks for intellectual property protection regarding multimedia digital library: copyrightable database and patentable retrieval mechanism.

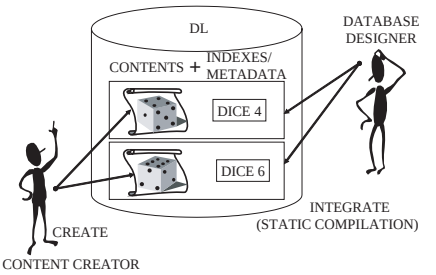
Multimedia Digital Library as Copyrightable Database

Our framework for copyrighting the databases of multimedia digital libraries determines which type of database should be copyrightable in the form of a component of contents-plus-indexes (Sasaki & Kiyoki, 2003, 2004, 2005b). The collection of static indexes and individual contents forms a component of contents-plus-indexes. That component identifies the entire content of each database, as is a static and copyrightable compilation. Copyrightable compilation is to be of sufficient creativity, that is, originality in the form of a component of contents-plus-indexes.

The set of conditions on the original categorization regarding indexes or metadata is formulated as follows (Sasaki & Kiyoki, 2004, 2005b). A categorization regarding indexes or metadata is original only when:

1. the type of indexes or metadata accepts discretionary selection in the domain of a problem database; otherwise, and
2. the type of taxonomy regarding indexes or metadata accepts discretionary selection in the domain of a problem database.

Figure 1. Copyrightable digital library as contents-plus-indexes



A typical case of nonoriginal categorization is a photo film album database which has indexes of consecutive numbers. That case does not accept any discretion in the selection of the type of indexes or metadata, or the type of taxonomy. The photo film album database uses its respective film numbers as indexes for its retrieval operations. The taxonomy of the indexes is only based on the consecutive numbering without any discretion in its selection of the type of indexes or taxonomy regarding a database in a multimedia digital library.

Meanwhile, the discretionary selection of the type of indexes or metadata, or taxonomy constitutes copyrightable compilation of minimal creativity, that is, originality on the categorization regarding indexes or metadata. A typical case of discretionary selection of the type of indexes or metadata is the Web document encyclopedia as a multimedia digital library. Suppose that a database restores pictures of starfish which are manually and numerically numbered by day/hour-chronicle interval that is based on their significant life stages from birth to death. That database is to be an original work of authorship as a copyrightable compilation in the form of a component of contents-plus-indexes. That database of discretionary type of numbering or indexing is an original, that is, copyrightable database in a multimedia digital library.

Multimedia Digital Library as Patentable Mechanism

Our framework for patenting the retrieval mechanisms of multimedia digital libraries determines which type of retrieval mechanism should be patentable in the form of a component of novel combination of prior disclosed processes and/or a component of specific parameter setting (Sasaki & Kiyoki, 2002a, 2002b, 2005a, 2005b). The frameworks focus on the

following three requirements for patentability: *patentable subject matter* (entrance to patent protection), *nonobviousness* (technical advancement) and *enablement* (specification) (Merges & Duffy, 2002).

The requirement for nonobviousness on the combination of the processes for data processing as the retrieval mechanism in a multimedia digital library is listed as follows (Sasaki & Kiyoki, 2005a):

1. The processes for performing a retrieval mechanism must comprise the combination of prior disclosed means to perform a certain mechanism which is not predicated from any combination of the prior arts and
2. the processes for performing a retrieval mechanism must realize quantitative and/or qualitative advancement.

Otherwise, the discussed processes are obvious so that they are not patentable as the processes for performing a retrieval mechanism.

First, a combination of prior disclosed means should not be "suggested" from any disclosed means "with the reasonable expectation of success" (*In re* Dow Chemical Co., 1988). Second, its asserted function on the discussed mechanism must be superior to the conventional functions which are realized in the prior disclosed or patented means in the field of the retrieval mechanism of multimedia digital libraries. On the latter issue, several solutions for performance evaluation are proposed and applicable. Another general strategy is restriction of the scope of problem claims into a certain narrow field to which no prior arts have been applied. This claiming strategy is known as the local optimization of application scope.

A component for parameter setting realizes thresholding operations in the form of a computer program with a set of ranges of parametric values. In retrieval mechanisms, parametric values determine, as thresholds, which candidate image is similar to an exemplary requested image by computation of similarity of visual features (Deb, 2004; Rui, Huang, & Chang, 1999; Smeulders, Worring, Santini, Gupta, & Jain, 2000; Yoshitaka & Ichikawa, 1999). That parameter setting component is to be a computer-related invention in the form of computer program as far as that parameter setting is sufficiently specified to enable a claimed invention or retrieval mechanism (U.S. Patent and Trademark Office, 1996a)

The requirement for enablement on the parameter setting component of the retrieval mechanism in a multimedia digital library is listed as follows (Sasaki & Kiyoki, 2005a):

1(a). The descriptions of the processes for performing a retrieval mechanism must specify the formulas for parameter setting.

1(b). Otherwise, the disclosed invention of the processes should have its co-pending application that describes the formulas in detail.

2(a). the processes must perform a new mechanism by a combination of the prior disclosed means; otherwise,

2(b). the processes should have improved formulas for parameter setting which is based on the prior disclosed means for performing a retrieval mechanism and should also give examples of parametric values on parameter setting in descriptions.

For 2(b), the processes must specify the means for parameter setting by "giving a specific example of preparing an" application to enable those skilled in the arts to implement their best mode of the processes without undue experiment (*Autogiro Co. of America vs. United States,* 1967; *Unique Concepts, Inc. v. Brown,* 1991). U.S. Patent and Trademark Office (1996a, 1996b) suggested that the processes comprising the means, that is, the components for parameter setting must disclose at least one of the following examples of parametric values on parameter setting:

1. working or prophetic examples of initial values or weights on parameter setting; or
2. working examples of the ranges of parametric values on parameter setting.

The "working examples" are parametric values that are confirmed to work at an actual laboratory or as prototype testing results. The "prophetic examples" are given without actual work by one skilled in the art.

The retrieval mechanisms of multimedia digital libraries are patentable in the form of components of novel combinations of prior disclosed processes and/or components of specific parameter settings while they are to satisfy the aforementioned conditions.

FUTURE TRENDS

In the field of visual information retrieval, the digital library community faces a variety of problems. Especially, two problems demand urgent solutions for the future progress of digital library: (1) a framework for protecting a database as a whole and (2) a scheme for protecting parameter settings in the form of trade secret.

Databases contain multimedia information including images and videos. Portable information devices allow people to easily access a large amount of downloadable multimedia files stored in distributed databases around the world. Network technology for efficient data transactions often triggers unauthorized misappropriation of those multimedia files that are important intellectual assets. Even the sui generis right of database protection discussed in Europe

is not to protect any database as a whole in the present legal system. It is necessary to prepare a framework for protecting entire databases including their contents. That framework should determine how and which type of database is to be protected as a whole.

Another emerging problem is discussed on the parameter settings of retrieval mechanisms. Patent application on the parameter setting components demands applicants as developers to make public the detailed know-how on the best range of parametric values in practice. The discovery of those parametric values needs considerable pecuniary investment in research and development. That kind of knowledge should be kept covered in the form of trade secret but not be open in public via patent application. The digital library community demands a scheme that determines which parameter setting component should be patentable or kept secret regarding multimedia digital libraries. It is necessary to prepare a scheme that determines how and which part of parameter setting components should take the form of trade secret.

CONCLUSION

In this article, we have discussed issues on intellectual property protection regarding multimedia digital libraries which consist of indexed multimedia digital contents in databases and retrieval mechanisms. We have presented the frameworks for copyrighting the database of multimedia digital library in the form of a component of contents-plus-indexes, and for patenting the retrieval mechanism of multimedia digital library in the form of a combination of processes and/or a component of parameter settings. We have also pointed out an emerging problem on the trade secret of parameter settings.

REFERENCES

American Dental Ass'n v. Delta Dental Plan Ass'n, 126 F.3d 977 (7th Cir. 1997).

Aplin, T. (2005). *Copyright law in the digital society: The challenges of multimedia*. Oxford, UK: Hart.

Autogiro Co. of America vs. United States, 384 F.2d 391, 155 U.S.P.Q. 697 (Ct. Cl. 1967).

Borgman, C. L. (2003). *From Gutenberg to the global information infrastructure: Access to information in the networked world (Digital Libraries and Electronic Publishing).* Cambridge, MA: MIT Press.

Deb, S. (2004). *Multimedia systems and content-based retrieval.* Hershey, PA: Idea Group Publishing.

Gorman, R. A., & Ginsburg, J. C. (2002). *Copyright: Cases and materials* (6th ed.). University casebook series. Charlottesville, NC: The Michie Company.

In re Alappat, 33 F.3d 1526, 31 U.S.P.Q.2d 1545 (Fed. Cir. 1994) (en banc).

In re Dow Chemical Co., 837 F.2d 469, 473, 5 U.S.P.Q.2d 1529, 1531 (Fed. Cir. 1988).

Jakes, J. M., & Yoches, E. R. (1989). Legally speaking: Basic principles of patent protection for computer science. *Communications of the ACM, 32*(8), 922-924.

Junghans, C., & Levy, A. (2006). *Intellectual property management: A guide for scientists, engineers, financiers, and managers*. Hoboken, NJ: John Wiley & Sons.

Merges, R. P., & Duffy, J. F. (2002). *Patent law and policy: Cases and materials* (3rd ed.). Dayton, OH: LexisNexis.

Nimmer, M. B., Marcus, P., Myers, D. A., & Nimmer, D. (2006). *Cases and materials on copyright & other aspects of entertainment litigation including unfair competition* (7th ed.). Dayton, OH: LexisNexis.

Reinbothe, J. (1999, September 14-16). The legal protection of non-creative databases. In *Proceedings of the Database Workshop of the International Conference of Electronic Commerce and Intellectual Property,* (*WIPO/EC/CONF/99/SPK/22-A*). Geneva, Switzerland: WIPO.

Rui, Y., Huang, T. S., & Chang, S.-F. (1999). Image retrieval: Current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation, 10*(4), 39-62.

Samuelson, P. (1996). Legally speaking: Legal protection for database content. *Communications of the ACM, 39*(12), 17-23.

Sasaki, H., & Kiyoki, Y. (2002a, November 6-7). Patenting advanced search engines of multimedia databases. In S. Lesavich (Ed.), *Proceedings of the 3rd International Conference on Law and Technology* Cambridge, MA (pp. 34-39). International Society of Law and Technology (ISLAT). Calgary: Acta Press.

Sasaki, H., & Kiyoki, Y. (2002b, December 11-14). Patenting the processes for content-based retrieval in digital libraries. In E.-P. Lim, S. Foo, C. Khoo, H. Chen, E. Fox, S. Urs, & T. Costantino (Eds.), *Proceedings of the 5th International Conference on Asian Digital Libraries (ICADL) — Digital Libraries: People, Knowledge, and Technology, (LNCS 2555* pp. 471-482). Singapore. Berlin: Springer-Verlag.

Sasaki, H., & Kiyoki, Y. (2003, May 27-31). A proposal for digital library protection. In *Proceedings of the 3rd ACM/ IEEE-CS Joint Conference on Digital Libraries*, Houston, TX, (p. 392). . Los Alamitos, CA: IEEE Computer Society Press.

Sasaki, H., & Kiyoki, Y. (2004, December 11-14). Copyrighting digital libraries from database designer perspective. In *Proceedings of the 7th International Conference on Asian Digital Libraries (ICADL)*, Shanghai, China *(LNCS 3334*, pp. 626-629). Berlin, Germany: Springer-Verlag.

Sasaki, H., & Kiyoki, Y. (2005a). A formulation for patenting content-based retrieval processes in digital libraries. *Journal of Information Processing and Management, 41*(1), 57-74.

Sasaki, H., & Kiyoki, Y. (2005b). Multimedia digital library as intellectual property. In Y. L. Theng, & S. Foo (Eds.) *Design and usability of digital libraries: Case studies in the Asia Pacific* (pp. 238-253). Hershey, PA: Information Science Publishing.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349-1380.

Unique Concepts, Inc. vs. Brown, 939 F.2d 1558, 19 U.S.P.Q.2d 1500 (Fed. Cir. 1991).

U.S. Copyright Act, 17 U.S.C. Sec. 101, 103 (2005).

U.S. Patent Act, Title 35 U.S.C. Sec. 101, 103, 112 (2005).

U.S. Patent and Trademark Office. (1996a). *Examination guidelines for computer-related invention*s, 61 Fed. Reg. 7478 (Feb. 28, 1996) ("*Guidelines*").

U.S. Patent and Trademark Office. (1996b). *Examination guidelines for computer-related inventions training materials directed to business, artificial intelligence, and mathematical processing applications ("Training Materials"*). Retrieved July 1, 2006 from http://www.uspto.gov/web/offices/pac/compexam/examcomp.htm

Yoshitaka, A., & Ichikawa, T. (1999). A survey on content-based retrieval for multimedia databases. *IEEE Transactions on Knowledge and Data Engineering, 11*(1), 81-93.

KEY TERMS

Combination of Processes: Combination of processes is a patentable computer program or programs as a number of "processes," that is, methods or means for data processing, some of which are prior disclosed inventions.

Contents-Plus-Indexes: Contents-plus-indexes is a component that comprises static indexes or metadata fixed in a database of multimedia digital library.

Copyrightable Database: Copyrightable database is a compilation of individual contents, that is, preexisting data of originality, as fixed in tangible form.

Digital Library: Digital library is a system as an infrastructure for global information, which consists of digital contents in databases and retrieval mechanisms.

Multimedia Digital Contents: Multimedia digital contents are data entities which are stored in multimedia digital libraries in a variety of forms of text, images, photos, or video streams, which often commingle therein.

Multimedia Digital Library: Multimedia digital library consists of multimedia digital contents which are indexed and stored in databases for appropriate retrieval operations and the retrieval mechanisms which are optimized and applied to object domains of those databases.

Parameter Setting: Parameter setting is a patentable computer program or programs which realize thresholding operations with a set of ranges of parametric values.

Patentable Computer Program: Patentable computer program is a computer-related invention in the form of a data-processing process or method to produce a useful, concrete, and tangible result for transforming physical data.

Sui Generis Right of Database Protection: Sui generis right of database protection means a kind of right of database protection which the European Union has legislated as a scheme for protecting databases including contents per se.

Trade Secret on Parameter Setting: Trade secret on parameter setting is a legal framework to keep secret the range of parametric values in practice rather than just patenting parameter setting components.

Intelligent Information Systems

John Fulcher
University of Wollongong, Australia

INTRODUCTION (INFORMATION SYSTEM TYPES & FUNCTIONS)

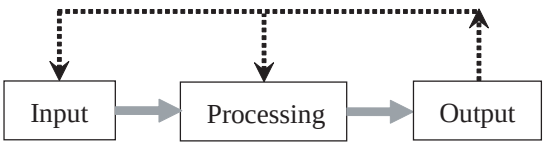
Information Systems (IS), not surprisingly, process information (data + meaning) on behalf of and for the benefit of human users. Information Systems comprise the basic building blocks shown in Figure 1, and as such can be likened to the familiar Von Neumann computer architecture model that has dominated computing since the mid 20th Century. In practice, IS encompass not just computer system hardware (including networking) and software (including DataBases), but also the *people* within an organization (Stair & Reynolds, 1999).

Information Systems are ubiquitous in today's world–the so-called "Digital Age"–and are tailor-made to suit the needs of many different industries. The following are some representative application domains:

- Management Information Systems (MIS)
- Business IS
- Transaction processing systems (& by extension, eCommerce)
- Marketing/Sales/Inventory IS (especially via the Internet)
- Postal/courier/transport/fleet/logistics IS
- Geographical Information System (GIS)/Global Positioning Satellite (GPS) systems
- Health/Medical/Nursing IS

The roles performed by IS have changed over the past few decades. More specifically, whereas IS focussed on data processing during the 1950s and 1960s, management reporting in the 1960s and 1970s, decision support during the 1970s and 1980s, strategies and end user support during the 1980s and 1990s, these days (the early years of the 21st Century) they focus more on global Internetworking (O'Brien, 1997). Accordingly, we nowadays find extensive use of IS in e-business, decision support, and business integration (Malaga, 2005). Let us take a closer look at one of these–Decision Support Systems. A DSS consists of (i) a (Graphical) User Interface, (ii) a Model Management System, and (iii) a Data Management System (comprising not only Data/Knowledge Bases but also Data Warehouses, as well as perhaps incorporating some Data Mining functionality). The DSS GUI typically displays output by way of text, graphs, charts and the like, enabling users to visualize recommendations/advice produced by the DSS. The Model Management System enables users to conduct simulations, perform sensitivity analysis, explore "what-if" scenarios (in a more extensive manner than what we are familiar with in spreadsheets), and so forth.

Figure 1. Generic information system



BACKGROUND

Whenever we encounter "intelligence" in relation to IS, it is usually in the context of (i) the decision making process itself, (ii) intelligent organizations, (iii) software agents, or (iv) the incorporation of Artificial Intelligence (AI) techniques.

For instance, Filos (2006) characterizes a "smart/intelligent" organization as being networked in the following three dimensions: (a) Information & Communications Technology (ICT), (b) organizational, and (c) *knowledge* (it is interesting to note the link here between (c) and the discussion which follows in this article). Furthermore, in the "Digital Age," the latter necessarily incorporates uncertainty and unpredictability. In this regard, Kelly and Allison (1999) demonstrate how the following concepts from Complexity Theory can be applied to improve business:

1. Nonlinear dynamics
2. Open systems
3. Feedback loops
4. Fractal structures
5. Evolutionary theory
6. Group self-organization

Filos (2006) further observes that rather than attempt to control their environments, organizations in the Digital Age will *adapt* to them (lest if they don't, they run the risk of stifling creativity, imagination, and innovation). Again, this notion of "adaptability" will resurface in our impending discussion of "intelligence."

According to Simon and Newell (1961), the human decision-making process comprises three phases, namely: (i) "intelligence," (ii) design, and (iii) choice. The use of the term "intelligence" here has a different meaning from that used in this article. More specifically, it is used by Simon and Newell in the sense of backgrounding a specific topic, as one performs in undertaking a literature review prior to commencing a new research project; in other words, to learn what has gone before (in order to "stand on the shoulders of giants," to paraphrase Einstein), and to avoid "re-inventing the wheel."

"Intelligence" can take yet another meaning within the context of IS, more specifically in relation to system security. Intelligence gathering is a prime concern of the latter, with various IS developed to support these endeavours. There are inherent dangers in such systems however, as flagged by the American Civil Liberties Union (http://www.aclu.org) in relation to the use of biometrics generally, and more especially to the 2008 RFID legislation; stated simply, whereas RFID is capable of providing audit trails, their indiscriminate use *could* encroach on ones privacy). The interests of "Big Brother" (i.e., Government) need to be balanced against citizens' rights in this regard, especially because identity theft has become such a major concern since the turn of the century. Space does not allow us to pursue these issues further, as this really warrants another article (book?) entirely.

INTELLIGENT INFORMATION SYSTEMS

Before we proceed further with our discussion of Intelligent IS, we need to define what *we* mean by the term "intelligence." We can regard "information" as being data + meaning. By extension, we can characterize "knowledge" as information + experience. Defining "intelligence," however, is a rather more challenging task. In the AI world, intelligence is generally viewed as encompassing:

- Awareness of (knowledge about and the ability to interact with) the surrounding environment, and
- An ability to learn from experience and adapt accordingly.

The first of these criteria presupposes an efficient method of encoding, storing and retrieving knowledge. Several different methods exist for doing so, including if…then (or fuzzy) production rules, frames (schema), semantic networks, propositional/predicate logic, or Artificial Neural Networks. The second criteria raises the issue of "learning" from experience, incorporating such knowledge into a "Knowledge Base" (KB), and consulting this knowledge (wisdom?) when encountering new situations and circumstances, whether consciously or unconsciously (i.e., relying not just on reasoning but also on intuition). It also implies some pattern recognition ability, in order to extrapolate from known situations, to apply heuristics (rules-of-thumb), and to build upon existing knowledge.

The claim that IBM's Deep Blue is intelligent because it defeated a human (World) Chess Champion is somewhat missing the point. The ability of the former to look a long way ahead with sequences of next moves–up to 100 billion–is indicative of little more than a brute force approach, after all (i.e., *un*intelligent).

In this article, we will restrict our focus to natural/biological systems that appear to exhibit "intelligence" (irregardless of our specific definition of the latter term). Our premise is that by mimicking (or alternatively taking inspiration from) nature, we stand to develop systems which *naturally* exhibit intelligence." Table 1 compares and contrasts the attributes of such biologically-inspired ("soft computing") approaches with that of logic and reasoning–the underpinning of conventional (algorithmic-based) computing.

REPRESENTATIVE INTELLIGENT SYSTEMS

After briefly describing the underlying principles of each approach, we proceed to cite representative examples where researchers have applied "intelligent" techniques to solve real-world problems. We restrict ourselves to six biologically

Table 1. Classical vs. soft computing

Classical computing	Soft computing
2-valued (Boolean/crisp) logic	many-valued (Fuzzy) logic
precise	approximate
deterministic	stochastic (i.e., incorporates some randomness/unpredictability)
exact/precise data	ambiguous/approximate/inconsistent data
sequential processing	parallel processing

inspired soft computing methods here, these being Artificial Neural Networks, Genetic Algorithms, swarms, DNA immune- and membrane-based computing. We basically don't consider Fuzzy Systems, reasoning systems or rule-based Expert Systems as such. However, we *do* mention such systems in the context of combinations/hybrids of such soft computing techniques, which has become the province of the field of Computational Intelligence–CI–in recent times (Fulcher & Jain, 2008).

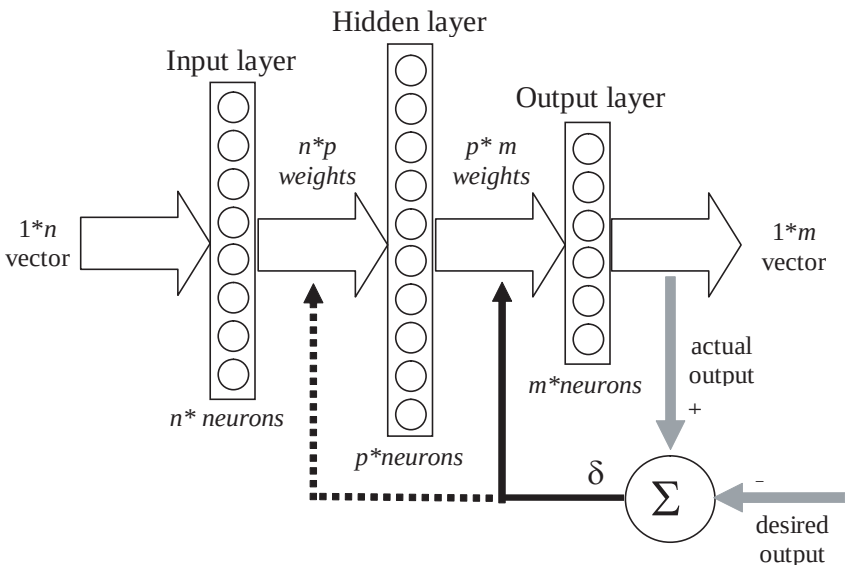
Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) are simplistic models of biological neural networks (brains), typically comprising dozens (but not billions) of neurons and hundreds (not tens of billions) of synapses (connections between neurons). The output (axon) of a biological neuron "fires" (produces a pulse train signal output) whenever the weighted sum of the signals from the inputs (synapses) exceeds some preset threshold. Now excitation (inhibition) of individual neurons is essentially an electrochemical process, involving different concentrations of potassium and sodium ions within and outside of the cell body; moreover, this is inherently an analog (linear) process. In the simplified neuron model commonly used in ANNs, neuron "firing" corresponds to a simple output level shift (0→1; 1→0). One characteristic of biological networks is their localized behaviour; in other words, certain areas of the brain are responsible for processing information incoming from our senses (although substantial preprocessing often takes place in the cerebral cortex prior to arriving in the brain proper), or for producing the necessary outputs (motor movement, speech, and so forth). Some ANN models reflect this localized behaviour, while others employ a more uniform, holistic architecture.

Another characteristic of biological brains is their massively parallel (analog) processing capability, such that despite the relative slow processing capability of individual neurons (milliseconds), their collective processing power far exceeds that of the fastest supercomputers, at least for some tasks. Realization of parallel, analog, neural network hardware in practice is by way of (sequential) digital computer software simulation. The ANN models in common usage are very much simplified versions of the biological networks from which they derive their inspiration. The most popular ANN model (Wong, Lai, & Lam, 2000) is the Multilayer Perceptron (MLP) of Figure 2.

ANNs are not programmed in the traditional algorithmic sense, but rather learn by example, at least in the *supervised* kind. Accordingly, supervised networks require numerous input-output training data pairs in order to learn the underlying "intelligence" of the system under study. Once trained, an ANN is capable of correctly recognizing input patterns not previously met during the training process; in other words, it exhibits generalization ability. Note that such a training process is an inherently data-driven, bottom-up approach (in contrast to conventional model-driven, top-down, algorithmic approaches). Furthermore, the training process can be quite time consuming; however, once trained, an ANN can respond almost instantaneously to new inputs.

Figure 2. A (fully-connected) 3-layer MLP/BP



The Multilayer Perceptron (MLP) of Figure 2 is a fully connected, 3-layer, supervised, feedforward ANN, comprising input, hidden, and output layers, each of which contain *n*, *p* and *m* neurons, respectively. By "feedforward," we mean that connections (weights) only exist in a forward direction, that is, from one of the *n* neurons in the Input Layer to one of the *p* neurons in the Hidden Layer (or from one of the *p* neurons in the Hidden Layer to one of the *m* neurons in the Output Layer). By contrast, no such restrictions apply in brains. The MLP employs the so-called BackPropagation learning rule, which simply stated says that upon presentation of an input-output training exemplar pair, the *actual* output produced by the network is compared with the *desired* output. During each successive training iteration, the weights are adjusted in proportion to this error (Δ or difference) signal: firstly, to adjust the weights connecting the Hidden Layer to the Output Layer, then those between the Input Layer and the Hidden Layer. In this manner, the error signal "propagates" backward from the ANN output to its input, adjusting its weights in the process; hence its name (BP).

Presentation of all input-output training pairs (exemplars)–one "epoch"–will see the weights change in many different (and incompatible/conflicting) directions. In practice *many* epochs will be necessary in order for the network to converge to an acceptable solution (which corresponds to the network having learned *all* I/O pattern associations).

It has been proven mathematically that the BP algorithm will *eventually* converge to an acceptable solution, although this might not be within a convenient timeframe from a user's perspective! In practice, training of ANNs can take several hours, perhaps even overnight, even on top-of-the-range computers. Training ceases when the error (difference) signal falls below a certain level (say 0.1%), or alternatively after a certain predetermined number of training epochs.

Common practice is to divide the available training data in two, then use one half for training and the other half to test (verify) the network once trained. Now in practice such labelled I/O training data (exemplars) may not always be available, and hence some people prefer to use *un*supervised neural networks. One has to exercise caution with the latter, however, because the resulting classes/clusters the network produces are often suspect. We restrict our current discussion here to supervised ANNs, and indeed to only *one* type (MLP/BP).

There is also the issue of how many I-O training exemplar pairs constitute a "minimum-yet-sufficient" set: too few will not lead to network convergence, whereas too many could result in "overtraining" (akin to "overfitting" in mathematical function approximation/curve fitting).

ANNs are especially good at pattern recognition or pattern classification, irrespective of what the pattern actually represents. This means that in practice we need to be able to encode the pattern of interest (be this vision, speech, time series, or whatever) into an appropriate form. Indeed, preprocessing is often the most challenging aspect of applying ANNs to real-world problems. Typical preprocessing tasks include the handling of missing, incomplete or noisy data, and most especially dimensionality reduction (because from what we have already seen, ANN training times are quite long; in fact, they increase exponentially as a function of the number of network weights; accordingly, any reduction in the dimensionality of the training data will have a dramatic effect on network convergence times).

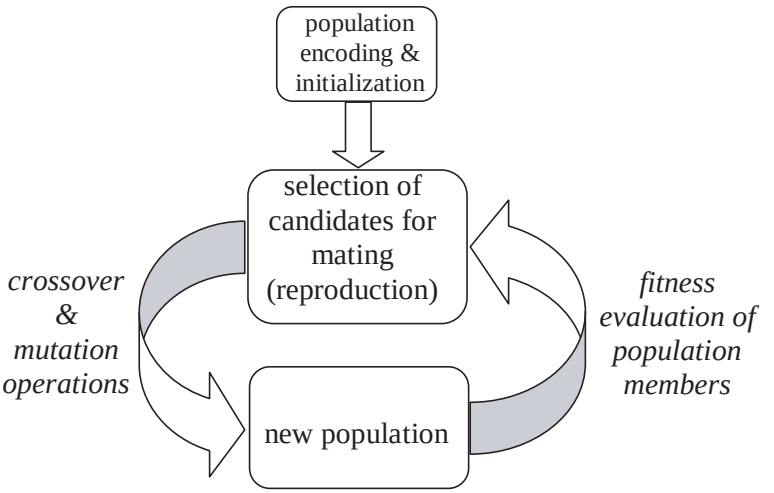
Verma & Panchal (2006) used ANNs (supervised MLP/BP) in a standard pattern classification task, that of discriminating between malignant (cancerous) and benign pap smears. Fyfe (2008) applied an unsupervised ANN–the self-organizing map–to data clustering and visualization. Likewise, Yin (2008) showed how the SOM–and variants thereof–could also be applied to vector quantization, image segmentation, density modeling, gene expression analysis, and text mining. More sophisticated (Higher-Order, supervised) ANN models have been used for both satellite weather prediction (Zhang & Fulcher, 2004) and financial time series prediction (Fulcher, Zhang & Xu, 2006). Zeleznikow (2004) combined ANNs and rule-based reasoning in the development of an Intelligent Legal Decision Support System. By contrast, Fu, Li, Wang, Ong, and Turner (2008) combined ANNs and multi-agents in order to predict network traffic over media grids.

Evolutionary Algorithms

As is the case with ANNs, evolutionary methods take their inspiration from Nature, in this case Darwinism and "survival-of-the-fittest." Although there are other variants–most notably evolutionary programming and genetic programming–we will restrict our discussion here to that of Genetic Algorithms (GAs). We assume the simplified evolutionary model of Figure 3.

Prior to evolving a solution to the problem of interest, we must first be able to encode potential (candidate) solutions into (fixed-length) genetic string form. As with ANNs, in practice this preprocessing stage can often prove the most difficult part of the exercise. Commencing with random bit strings, we first select two "parent" strings from the available population on the basis of an objective (cost or fitness) function, and proceed to "mate" them. As in biological evolution, a "child" will inherit half of its genetic code (attributes, characteristics) from either parent. The aim is that over time stronger members will "evolve" more appropriate solutions to the problem at hand, while at the same time maintaining sufficient diversity among the population as a whole to ensure healthy future generations. As in nature, a certain degree of randomness (mutation) needs to be injected into this process, in order to prevent "inbreeding" and proceeding too far down evolutionary "blind alleys" (dead ends).

Figure 3. The steps in evolution



Not surprisingly, evolution of an acceptable solution can take a very long time, typically even longer than is the case with ANN training.

Mumford (2008) showed how GAs could be applied to set partitioning problems (such as graph colouring, bin packing and timetabling). Ishibuchi, Nojima and Kowajima (2008) evolved Fuzzy Classifiers using evolutionary techniques. Beale and Pryke (2006) combined GAs with interactive 3D dynamic visualization techniques in the realization of their Haiku Knowledge Discovery system. Tran, Abraham and Jain (2006) combined ANNs, EAs and Fuzzy inference methods in the development of intelligent Decision Support Systems (DSS).

Swarm Intelligence

The inspiration for this approach stems from the collective behaviour of bee/ant colonies, bird flocks, animal herds, and other social insects/animals. What we are attempting to exploit using such techniques is a system in which the whole is greater than the sum of the parts, in the sense that whereas individual members are relatively unintelligent, the collective behaviour of the colony/flock/herd exhibits "intelligent" behaviour.

Swarms differ from GAs in that there is no direct influence from one generation to the next; the focus is rather on how *present*-generation members affect the behaviour of others; "peer pressure" in a sense. Such influence is indirect, and often takes the form of general, "broadcast" messages, rather than "peer-to-peer" communication, as it were. A good example of this is the depositing of chemical (pheromone) trails which mark the path from a hive to a food source, say.

Apart from such reliance on indirect rather than direct communication between population members, a couple of other constraints apply to swarms, these being:

- *Intra*generational learning only (i.e., no *inter*generational learning), and
- Individual swarm members are assumed to have identical form, function, and status, and hence are interchangeable.

Actually there are several variants of swarms in common usage, including Swarm Intelligence (SI) (Bonabeau, Dorigo, & Theraulaz, 1999), Particle Swarm Optimization (PSO) (Kennedy & Eberhardt, 1995), Ant Colony Optimization (ACO) (Dorigo, Maniezzo, & Colorni, 1996), and Autonomous Nanotechnology Swarms (ANTS), the latter having been proposed by NASA for future space exploration projects (Hinchey, Sterritt, & Rouff, 2007).

Hendtlass (2004) applied both Ant Colony and Particle Swarm Optimization (PSO) to the Travelling Salesman Problem (TSP). His ACO algorithm can be paraphrased as follows:

- Initialise pheromone levels on each path segment and randomly distribute *N* ants among *C* cities;
- **Repeat**
  - Repeat

    - Each ant decides which city to move to next (provided it does not *re*visit a city)
  - **Until** it returns to its starting city
- Each ant calculates the length of this most recent tour and updates information about the shortest tour found to date;
- The pheromone levels on each path segment are updated (refreshed);
- All ants having completed a predefined maximum number of tours "die off" & are replaced by new ants at randomly selected cities;
- **Until** termination condition met (e.g., shortest path < predefined threshold, or maximum number of tours made).

Khosla, Kumar and Aggrawal (2006) combined PSO and the Taguchi Method to derive optimal Fuzzy Models of a Ni-Cd battery charger. Sharkey and Sharkey (2006) combined swarms and software agents in their work with collective/swarm robotics.

DNA, Immune-Based, and Membrane-based Computing

The three approaches considered so far, while being inspired by nature, are realized in practice by way of software simulations on (silicon-based) digital computers. With the emerging field of DNA computing, we turn our attention to carbon-based computing, or so-called "wetware," a "computer-in-a-test tube," as it were. Classical algorithms are employed, rather than the data-driven approaches characteristic of ANNs, GAs and swarms.

The potential we are attempting to exploit using such an approach is the massive parallelism which results from the *simultaneous* reactions of large numbers of DNA molecules within a single test tube, despite the computation times of individual reactions being quite slow (a similar phenomenon to that previously encountered in relation to biological neural networks; in other words, fast overall behaviour results from the relatively *slow* computations that take place within individual neurons).

Not surprisingly, one of the biggest challenges with DNA computing–just as with Quantum Computing, as it happens–is Input/Output. More specifically, how does one first encode the problem of interest into DNA strand form? Next, having done so, how does one decode the result of the chemical reaction(s) into an intelligible form (and moreover, one that relates back to the problem at hand)?

Watada (2008) demonstrated how DNA computing can be applied to scheduling problems, in particular synchronizing the movements of multiple elevators in a multistorey (high-rise) building.

Immune-based computing (IBC) utilizes "antibodies" to discriminate between "self" (good cells) and "nonself" (bad/cancerous cells) and to affect self-repair. Ishida (2008) applied IBC to the so-called "stable marriage problem," and further showed how IBC could be extended to a general problem solver (the latter, by way of interest, was one of the traditional goals of AI).

The allied field of membrane-based computing is inspired by the so-called "reaction rules" between objects located within the compartments defined by a membrane structure. Not only are objects able to react with each other, they also on occasion pass through the membrane; also, the membrane itself can change shape, divide, dissolve or alter its permeability. Membranes are thus in a constant state of (nondeterministic) transition. Sequences of such transitions are the mechanism whereby we are able to realize parallel computations (Paun, 2002).

FUTURE TRENDS

Now, while much stands to be gained by employing the (largely biology-inspired) "intelligent" techniques discussed above, currently many advances emanate from *combinations* (hybrids) of these, perhaps also incorporating statistical or Machine Learning methods more commonly encountered in Data Mining. Indeed, such hybrid approaches have become a growing concern within the discipline of Computational Intelligence, as previously mentioned (Fulcher & Jain, 2008).

A few representative examples have been mentioned. Considerable activity is currently underway in the research community in developing such hybrid systems, which no doubt will set the research agenda for the foreseeable future. Duch (2007) even suggests that Computational Intelligence *might* be capable of realizing a truly "intelligent" machine where AI has failed to do so during the past 50 years.

CONCLUSION

In this article, we have focussed our attention on Intelligent IS, emphasizing the incorporation of principles imitated from/inspired by nature, in attempts to realize more efficient and better performing systems. The primary areas in nature which have served as inspiration to date include (i) Artificial Neural Networks (biological brains), (ii) Evolutionary Algorithms (Darwin's theories of evolution and survival-of-the-fittest), (iii) swarms (i.e., ants, bees, and flocks of birds), and (iv) DNA, immune-based and membrane-based computing. Several different examples of IS which have employed such biologically-inspired approaches (and most especially *hybrid* techniques) were briefly described. It is the contention of the present author that there is yet more to be gained by adopting such approaches, and no doubt we

will witness the development of many more Intelligent IS in the years (decades) to come.

REFERENCES

Beale, R., & Pyrke, A. (2006). Knowledge through evolution. In J.A. Fulcher (Ed.), *Advances in applied artificial intelligence* (pp. 234-250). Hershey, PA: Idea Group.

Bonabeau, E., Dorigo, M., & Theraulaz, G. (1999). *Swarm intelligence: From natural to artificial systems.* Oxford, UK: Oxford University Press.

Dorigo, M., Maniezzo, V., & Colorni, A. (1996). The ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man and Cubernetics–Part B*, *26*, 29-41.

Duch, W. (2007). What is computational intelligence and where is it going? In W. Duch, & J. Mandziuk (Eds.), *Challenges for computational intelligence*. Berlin: Springer-Verlag.

Filos, E. (2006). Smart organizations in the digital age. In I. Mezgar (Ed.), *Integration of ICT in smart organizations* (pp. 1-37). Hershey, PA: Idea Group.

Fu, X., Li, X., Wang, L., Ong, D., & Turner, S.J. (2008). Data mining in QoS-aware media grids. In J.A. Fulcher & L.C. Jain (Eds.), *Computational Intelligence: A Compendium* (pp.689-714). Berlin: Springer-Verlag.

Fulcher, J.A., & Jain, L.C. (Eds.). (2004). *Applied intelligent systems: New directions*. Berlin: Springer-Verlag.

Fulcher, J.A., & Jain, L.C. (Eds.). (2008). *Computational Intelligence: A Compendium.* Berlin: Springer-Verlag.

Fulcher, J.A., Zhang, M., & Xu, S. (2006). Higher order neural networks for financial prediction. In J. Kamaruzzaman (Ed.), *Artificial neural networks in finance & manufacturing* (pp. 80-108). Hershey, PA: Idea Group.

Fyfe, C. (2008). Topographic maps for clustering and data visualization. In J.A. Fulcher & L.C. Jain (Eds.), *Computational Intelligence: A Compendium* (pp. 111-153). Berlin: Springer-Verlag.

Hendtlass, T. (2004). An introduction to collective intelligence. In J. A. Fulcher & L.C. Jain (Eds.), *Applied intelligent systems: New directions* (pp. 133-178). Berlin: Springer-Verlag.

Hinchey, M.G., Sterritt, R., & Rouff, C. (2007). Swarms and swarm intelligence. *IEEE Computer*, *40*(4), 111-113.

Ishibuchi, H., Nojima, Y., & Kuwajima, I. (2008). Evolutionary multi-objective design of fuzzy rule-based classifiers. In J.A. Fulcher & L.C. Jain (Eds.), *Computational Intelligence: A Compendium* (pp.641-685). Berlin: Springer-Verlag.

Ishida, Y. (2008). The next generation of immunity-based systems. In J.A. Fulcher & L.C. Jain (Eds.), *Computational Intelligence: A Compendium* (pp. 1093-1121). Berlin: Springer-Verlag.

Kelly, S., & Allison, M.A. (1999). *The complexity advantage: How the science of complexity can help your business achieve peak performance.* New York: McGraw-Hill.

Kennedy, J., & Eberhardt, R.C. (1995). Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, Perth, Western Australia, (Vol. IV, pp. 1942-1948).

Koshla, A., Kumar, S., & Aggrawal, K.K. (2006). Swarm intelligence and the Taguchi method for identification of fuzzy models. In J.A. Fulcher (Ed.), *Advances in applied artificial intelligence* (pp. 273-295). Hershey, PA: Idea Group.

Malaga, R.A. (2005). *Information systems technology.* Upper Saddle River, NJ: Pearson/Prentice Hall.

Mumford, C. (2008). An order-based memetic evolutionary algorithm for set partitioning problems. In J.A. Fulcher & L.C. Jain (Eds.), *Computational Intelligence: A Compendium* (pp. 881-925). Berlin: Springer-Verlag.

O'Brien, J.A. (1997). *Introduction to information systems* (8th ed.). Chicago, IL: Irwin.

Paun, G. (2002). *Membrane computing: An introduction.* Berlin: Springer-Verlag.

Sharkey, A.J.C., & Sharkey, N. (2006). The application of swarm intelligence to collective robots. In J.A. Fulcher (Ed.), *Advances in applied artificial intelligence* (pp. 157-185). Hershey, PA: Idea Group.

Simon, H.A., & Newell, A. (1961). Computer simulation of human thinking and problem solving. *Datamation*, 35-37.

Stair, R.M., & Reynolds, G.W. (1999). *Principles of information systems* (4th ed.). Cambridge, MA: Thomson.

Tran C., Abraham, A., & Jain, L.C. (2006). Soft computing paradigms and regression trees in decision support systems. In J.A. Fulcher (Ed.), *Advances in applied artificial intelligence* (pp. 1-28). Hershey, PA: Idea Group.

Verma, B., & Panchal, R. (2006). Neural networks for the classification of benign and malignant patterns in digital mammograms. In J.A. Fulcher (Ed.), *Advances in applied artificial intelligence* (pp. 251-272). Hershey, PA: Idea Group.

Watada, J. (2008). DNA computing and its application. In J.A. Fulcher & L.C. Jain (Eds.), *Computational Intelligence:*

A Compendium (pp. 1067-1091). Berlin: Springer-Verlag.

Wong, B., Lai, V., & Lam, J. (2000). A bibliography of neural network business applications research: 1994-1998. *Computer and Operations Research*, *23*, 1045-1076.

Yin, H. (2008). The self-organizing map: Background, theory, extension, and applications. In J.A. Fulcher & L.C. Jain (Eds.), *Computational Intelligence: A Compendium* (pp. 715-762). Berlin: Springer-Verlag.

Zeleznikow, J. (2004). Building intelligent legal decision support systems: Past practice and future challenges. In J. A. Fulcher & L.C. Jain (Eds.). *Applied intelligent systems: New directions* (pp. 201-254). Berlin: Springer-Verlag.

Zhang, M., & Fulcher, J.A. (2004). Higher-order neural networks for satellite weather prediction. In J. A. Fulcher & L.C. Jain (Eds.), *Applied intelligent systems: New directions* (pp. 17-57). Berlin: Springer-Verlag.

KEY TERMS

Artificial Intelligence (AI): The field of study devoted to building machines which exhibit "intelligence," as commonly understood in relation to humans.

Artificial Neural Networks (ANNs): Simplified models of the human brain (biological neural network) which are particularly adept at pattern recognition or classification.

Backpropagation (BP) Algorithm: Used to train Multilayer Perceptrons (supervised, feedforward neural networks); BP adjusts the weights connecting neurons in the various layers according to the error (or difference) between actual and desired outputs generated in response to presentation of input-output training pattern pairs (exemplars).

Computational Intelligence (CI): Incorporates ANN, Fuzzy and Evolutionary approaches, and more especially *hybrids* of these (some authors extend this definition to include intelligent agents, stochastic reasoning and other techniques).

DNA Computing: The implementation of classical computing algorithms by way of chemical reactions within a test tube (so called "wet computing").

Evolutionary Algorithms (EAs): An iterative procedure which involve the "mating" of suitable parents from a population of solutions to a problem of interest, in the hope that more suitable "offspring" (i.e., solutions) will evolve over time.

Expert (or Knowledge-Based) System (ES/KBS): Comprise a (Graphical) User Interface, an Inference Engine and a Knowledge Base. The GUI accepts user queries (inputs), and presents the results to these queries (outputs) in a comprehensible manner, usually together with some justification (rules) or confidence level.

Immune-Based Computing: Uses "antigens" to discriminate between "self" and "nonself," and to affect self-repair within a system.

I**ntelligence:** Difficult to define *exactly*, but incorporates awareness of (and ability to interact with and adapt to) ones environment, as well as an ability to learn from experience (thereby increasing our knowledge).

Intelligent System: Used in this article primarily to mean (a) biologically-inspired "soft computing" techniques which can be incorporated into an information system (IS) in order to improve performance, and secondarily (b) in the sense of intelligence gathering, in order to render an IS more secure.

Soft Computing: An older term for Computational Intelligence (see above).

Swarm Intelligence (SI): Refers to a class of algorithms inspired by the collective behaviour of insect swarms, ant colonies, the flocking behaviour of some bird species, or the herding behaviour of some mammals, such that the behaviour of the whole can be considered as exhibiting a rudimentary form of "intelligence."

Category: Artificial Intelligence

Intelligent Software Agents and Multi-Agent Systems

Milan Stankovic
University of Belgrade, Serbia

Uros Krcadinac
University of Belgrade, Serbia

Vitomir Kovanovic
University of Belgrade, Serbia

Jelena Jovanovic
University of Belgrade, Serbia

INTRODUCTION

Agent-based systems are one of the most important and exciting areas of research and development that emerged in information technology (IT) in the past two decades. In a nutshell, an agent is a computer program that is capable of performing a flexible, autonomous action in typically dynamic and unpredictable domains (Luck, McBurney, Shehory, & Willmott, 2005). Agents emerged as a response of the IT research community to the new data-processing requirements that traditional computing models and paradigms were increasingly incapable to deal with (e.g., the huge and ever-increasing quantities of available data). Many IT researchers believe that agents represent one of the most important software paradigms that have emerged since the object orientation.

From the historic point of view, the agent-oriented research and development (R&D) originates from different disciplines. Undoubtedly, the main contribution to the field of autonomous agents came from artificial intelligence (AI). Ultimately, AI is all about building intelligent artifacts and if these artifacts sense and act in some environment, then they can be considered agents (Russell & Norvig, 1995). Also, object-oriented programming (Booch, 2004), concurrent object-based systems (Agha, Wegner, & Yonezawa, 1993), and human-computer interaction (Maes, 1994) are fields that constantly drive forward the agent R&D in the last few decades.

In addition, the concept of an agent has become important in a diverse range of sub-disciplines of IT, including software engineering, computer networks, mobile systems, control systems, decision support, information retrieval and management, electronic commerce, and many others. Agents are being used in an increasingly wide variety of applications—ranging from comparatively small systems such as personalized email filters to large, complex, mission critical systems such as air-traffic control.

BACKGROUND

Even though there is no universal consensus over some key definitions in the field, it is intuitively clear what an "agent" is. One of the most widely used definitions states that "*an agent is a computer system, situated in some environment, that is capable of flexible autonomous action in order to meet its design objectives"* (Jennings, Sycara, & Wooldridge, 1998, p. 8).

There are three key concepts in this definition: *situatedness*, *autonomy*, and *flexibility*. *Situatedness* means that an agent receives sensory input from its environment and that it can perform actions which change the environment in some way. *Autonomy* is seen as the ability of an agent to act without the direct intervention of humans and that it has control over its own actions and internal state. In addition, the autonomy implies the capability of learning from experience. By *flexibility*, we mean the agent's ability to perceive its environment and respond to changes in a timely fashion; it should be able to exhibit opportunistic, goal-directed behaviour and take the initiative whenever appropriate. Also, an agent should be able to interact with other agents and humans, thus to be *social*. Some authors emphasize

the importance of the concept of *rationality*, which will be discussed in the next section.

Moreover, agent technologies can be considered from three perspectives (Luck, McBurney, Shehory, & Willmott, 2005):

- As a *design metaphor*, agents offer designers a way of structuring an application around autonomous, communicative elements;
- As a *source of technologies*, agent-based computing spans a range of specific techniques aimed at supporting interactions among entities in dynamic and open environments; and
- As a *simulation tool*, multi-agent systems offer robust models for representing complex and dynamic real-world environments, such as economies, societies and bio-systems.

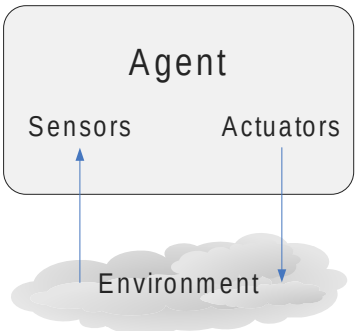
INTELLIGENT SOFTWARE AGENTS

Agents and Environments

Agents can be viewed as software entities that perceive their *environment* through *sensors* and act upon that environment through *actuators* (Russell & Norvig, 1995). There is an obvious analogy with a human agent who has ears, eyes, and other organs as sensors, and arms, legs, and other organs as actuators.

When we refer to an agent's perceptual inputs, we refer to the *agent's percepts.* An agent typically collects its percepts during the time, so its action in any moment generally depends on the whole sequence of percepts up to that moment. If we could generate a decision tree for every possible percept sequence of an agent, we could completely define the agent's behavior. Strictly speaking, we would say that we have defined the *agent function* that maps any sequence of percepts to the concrete action. The program that defines the agent function is called the *agent program*. These two concepts are different; the agent function is a formal description of the agent's behavior. The agent program is a concrete implementation of that formalism.

Figure 1. Agent and environment



As Russell and Norvig (1995) stipulate, one of the most desirable properties of an agent is its *rationality.* We say that an agent is rational if it always does the action that will cause the agent to be the most successful.

The rationality of an agent depends on:

- The performance measure that defines what is a good action and what is a bad action;
- The agent's knowledge about the environment;
- The agent's available actions;
- The agent's percept history.

One of the most cited definitions of a rational agent is:

for each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure, given the evidence provided by the percept sequence and whatever built-in knowledge the agent has. (Russell & Norvig, 1995, p. 36)

The main challenge in the field of intelligent software agents is to develop an agent program that implements the desired functionalities. Since it is a computer program, we need to have some computing device with appropriate sensors and actuators on which the agent program will run. We call this *agent architecture*. So an agent is essentially made of two components: the agent architecture and the agent program.

The Types of Agents

When we deal with the structure of an agent, we consider various implementation models for agent development. There are several basic types of agents with respect to their structure (Russell & Norvig, 1995).

The simplest kind of agents are the *simple reflex agents*. Such an agent only reacts to its current percept, totally ignoring its percept history. When a new percept is received, a rule that maps that percept to an action is fired. Such rules are known as *condition-action rules*.

Model-based reflex agents are more powerful agents, because they maintain some sort of internal state of the environment that depends on the percept history. For maintaining this sort of information, an agent must have two types of knowledge: (1) how the environment evolves, and (2) how its actions affect the environment.

Goal-based agents have some sort of goal information that describes desirable states of the world. Such an agent's decision making process is fundamentally different, because when a goal-based agent is considering performing an action it is asking itself "would this action make me happy?" along

with the standard "what this action will have as a result?".

Utility-based agents use a utility function that maps each state to a number that represents the degree of happiness. They are able to perform rationally even in the situations when there are conflicting goals, as well as when there are several goals that can be achieved, but none with certainty.

Learning agents do not have a priori knowledge of the environment, but *learn* about it. This is advantageous because these agents can operate in unknown environments and to a certain degree facilitates the job of developers because they do not need to specify their whole knowledge base.

Multi-Agent Systems and Environments

With an agent-oriented view of the world, it soon becomes clear that a single agent is insufficient. Most real-world problems require or involve multiple agents: to represent the decentralized nature of the problem, multiple perspectives, or competing interests.

Systems composed of multiple autonomous components (agents) are considered *multi-agent systems* (MAS), and historically belong to *distributed artificial intelligence* (Bond & Gasser, 1998). MAS can be defined as a loosely coupled network of problem solvers that work together to solve problems that are beyond the individual capabilities or knowledge of each problem solver (Durfee & Lesser, 1989). In an MAS, each agent has incomplete information or capabilities for solving the problem and thus has a limited viewpoint. There is no global system control, the data is decentralized, and the computation is asynchronous.

In addition to MAS, there is also the concept of a *multi-agent environment*, which can be seen as an environment that includes more than one agent. Thus, it can be cooperative, or competitive, or a combined one.

In an MAS and a multi-agent environment, the individual agents need to interact with one another (socialization) either to achieve their individual objectives, or to manage the dependencies that ensue from being situated in a common environment. These interactions range from simple semantic interoperation (exchanging comprehensible communications), through client-server interactions (the ability to request that a particular action is performed), to rich social interactions (the ability to cooperate, coordinate, and negotiate about a course of action).

Agent communication is achieved by exchanging messages represented by mutually understandable syntax and containing mutually understandable semantics. In order to find a common ground for communication, an *agent communication language* (*ACL*) should be used to provide mechanisms for agents to negotiate, query, and inform each other. The most important such languages today are *KQML* (Knowledge Query and Manipulation Language) (ARPA Knowledge Sharing Initiative, 1993) and *FIPA ACL* (FIPA, 1997) proposed by FIPA (IEEE Foundation for Intelligent Physical Agents, http://www.fipa.org/).

Because of the obstacles stemming from heterogeneous nature of agents involved in communication (e.g., finding one another), there is a need for *middle-agents*, which facilitate cooperation among agents and connect service providers with service requesters in the agent world. These agents are useful in various roles, such as *matchmakers* or *yellow page agents* that collect and process service offers ("advertisements"), *blackboard* agents that collect requests, and *brokers* that process both (Sycara, Decker, & Williamson, 1997).

Usability Aspects

The obvious advantage of agents is the clarity that the sole definition brings with regards to the representation of functionality. When an end user is presented with certain software as an agent (for example, an agent for filtering spam email), its functionalities are naturally obvious and the user experiences the system as a personal assistant that will work for him without much need for supervision.

Concerning the agents that communicate directly with end-users, the concept of agents shifts the paradigm of user interaction from simple user computer manipulation to assigning tasks to the computer. Within this scenario, the user assigns the tasks and acquires perception about an agent as an anthropomorphic software entity.

Controversies and Pitfalls of Agent Development

It is not uncommon that some users, not having enough knowledge about intelligent agents express concern (even fear) about their usage. They tend to confuse them with software daemons that send spam email messages, or with software viruses that can damage their systems.

Others perceive danger in the amount of independency that agents possess. Lead by skepticism, they are likely to discard the concept of autonomous software agents in favor of software that allows complete control over every operation it executes. This point of view goes along with arguments that giving too much decision making autonomy to a software entity would lead to dependency on such an entity's "will". Dangers arising from such autonomy are subject of various science fiction works (e.g., 2001: A Space Odyssey).

In the field of software engineering, agents are also criticized. The critiques are primarily concerned with traps in which a developer may fall when developing agents. Those pitfalls include applying agent technology in areas where other software engineering concepts are more suitable, the inappropriate use of other AI techniques, the inadequate number of agents in an agent-system, and so forth (Wooldridge & Jennings, 1998).

APPLICATIONS

Intelligent software agents are a suitable software engineering concept in a wide variety of application domains. Their autonomy qualifies them for successful application in the fields such as: traffic and transportation control, computer networks and telecommunication control, healthcare, and so forth.

Process control software systems require an entity that can supervise a process (e.g., production process) and react when needed. Reactive and responsive, agents perfectly fit the needs of such a task. Example domains in this field include: production process control, climate monitoring, spacecraft control, and monitoring nuclear power plants.

An important application domain is information gathering, where agents are used to search trough heterogeneous information sources (e.g., World Wide Web) and acquire relevant information for their users. One of the most common domains is Web browsing and search, where agents are used to adapt the content (e.g., search results) to the users' preferences and offer relevant help in browsing.

Agent cooperation opens possibilities for applying multi-agent systems to solving constraint satisfaction problems. Auction negotiation model, as a form of communication, enables a group of agents to find good solutions by achieving agreement and making mutual compromises in case of conflicting goals. Such an approach is applicable to trading systems, where agents act on behalf of buyers and sellers. Financial markets, as well as meeting scheduling, travel arrangement composing, and fault diagnosing also represent prominent fields for agent application.

Intelligent tutoring systems often include pedagogical agents, which represent software entities constructed to present the learning content in a user-friendly fashion and monitor the user's progress through the learning process. These agents are responsible for guiding the user and suggesting additional learning topics related to the user's needs (Devedzic, 2006).

Mobility Agents for People with Cognitive Disabilities

Mobility agents is an agent-based architecture that helps a person with cognitive disabilities to travel using public transportation. Agents are used to represent transportation participants (buses and travelers) and to enable notification of bus approaching and arrival. Information is passed to the traveler using a rich multimedia interface, via a handheld device. Customizable user profiles determine the most appropriate modality of interaction (voice, text, and pictures) based on the user's abilities (Repenning & Sullivan, 2003).

This architecture actually imposes a personal agent to guide users with cognitive disabilities and take care that abstract goals as "go home" are translated into concrete directions. To achieve this, an agent needs to collect information about user-specific locations and must be able to suggest the right bus for the particular user's current location and destination.

BluScreen

BluScreen is an intelligent public display, developed at the University of Southampton, in order to adapt the selection of adverts for display to the present audience detected by Bluetooth technology. Customization of content is achieved using history information of past users' exposure to certain sets of adverts, in order to predict which advert is likely to gain the highest attention.

The system is implemented as a multi-agent auction-based mechanism. Each agent represents a stakeholder wishing to advertise, and it is provided with a bidding strategy that utilizes heuristics to predict future advert exposure, based on the expected audience composition. These agents compete in an auction to gain advertising space, ensuring that the most suitable advertising content is selected (Payne, David, Jennings, & Sharifi, 2006).

BluScreen employs the concept of agents' socialization to achieve context aware, intelligent behavior of the system as a whole, relying on particular agents' autonomy to act on behalf of stakeholders and maximize their satisfaction.

Talaria

Talaria System (The autonomous lookup and report internet agent system), named after the Greek Messenger God Hermes's winged sandals, is a multi-agent system, developed for academic purposes at the University of Belgrade, Serbia, School of Business Administration (http://goodoldai.org.yu/talaria/). It was built as a solution to the common problem of gathering information from diverse Web sites that do not provide RSS feeds for news tracking. The system was implemented using the JADE modeling framework (http://jade.tilab.com/).

Talaria integrates information gathering and filtering in the context of supporting a user to manage her/his Web interests. The system provides each user with a personal agent, which periodically monitors the Web sites that the user expressed interest in. The agent informs its user about relevant changes, filtered by assumed user preferences and default relevance factors. Human-agent communication is implemented via email, so that a user can converse with her/his agent in natural language, whereas the agent heuristically interprets concrete instructions from the mail text (e.g., "change site list" or "kill yourself").

Human-like interaction, autonomy-related aspects of this system, and acting on behalf of the user emphasize the usability advantages of this agent-based software.

FUTURE TRENDS

The development of agent technologies is taking place within a context of broader visions and trends in IT, which are about to drive forward the whole field of intelligent agents. We especially emphasize *the semantic Web, Web services, ambient intelligence* and *peer-to-peer computing*.

The semantic Web is the vision of the future Web based on the idea that the data on the Web can be defined and linked in such a way that it can be used by machines for the automatic processing and integration (Berners-Lee, Hendler, & Lassila, 2001). The key to achieving this is by augmenting Web pages with descriptions of their content in such a way that it is possible for machines to reason automatically about that content. The semantic Web offers a solid ground for further development of the agent technologies as well as successful deployment of a variety of agent-based applications. Actually, we share the opinion that the semantic Web itself will be a form of intelligent infrastructure for agents, allowing them to "understand" the meaning of the data on the Web.

The other important drivers of agent development are the *Web services* and *service-oriented computing*, which are likely to become the dominant base technology in the foreseeable future. The Web service technology provides standard means for establishing interoperability between heterogeneous software applications that run on a variety of different platforms. Accordingly, this technology is almost ideal for use in supporting agent interactions in a multi-agent system. Moreover, an agent-oriented view of Web services is gaining an increasing attention, since agent-based systems are naturally seen as "provider and consumer" Web services environments (Booth et al., 2004).

The concept of *ambient intelligence*, being a vision that describes a shift away from PCs to a variety of devices which are unobtrusively embedded in our environment and which are accessed via intelligent interfaces, with no doubt require agent-like technologies in order to achieve autonomy, distribution, adaptation, and responsiveness (Booth et al., 2004). *Peer-to-peer (P2P) computing*, presenting networked applications in which every node is in some sense equivalent to all others, tends to become more complex in the future. Auction mechanism design, agent negotiation techniques, increasingly advanced approaches to trust and reputation, and the application of social norms, rules and structures—presents some of the agent technologies that are about to become relevant in the context of P2P computing (Booth et al., 2004).

Almost each of today's compelling IT visions and trends such as the abovementioned, but also grid computing, complex systems, and many more, will require agent technologies (or something similar to them), before being fully implemented. Agent technologies are upstream of these visions and mission-critical to them. However, to be able to support these visions, the agent-based computing needs further development and strengthening. Some considerable challenges have remained in the agent-based world, among which we emphasize the lack of sophisticated software tools, techniques and methodologies that would support the specification, development, integration and management of agent systems.

CONCLUSION

Research and development in the field of intelligent agents and multi-agent systems is rapidly expanding. It can be viewed as a melting pot of different ideas originating from diverse areas such as artificial intelligence, object-oriented systems, software engineering, distributed computing, economics, and so forth. At its core is the concept of autonomous agents interacting with one another for their individual and/or collective benefit.

A number of significant advances have been made over the past two decades in design and implementation of individual autonomous agents, and in the way in which they interact with one another. These concepts and technologies are now finding their way into commercial products and real-world software solutions. Future IT visions share the common need for agent technologies and prove that agent technologies will continue to be of vital importance.

REFERENCES

Agha, G., Wegner, P., & Yonezawa, A. (Eds.). (1993). *Research directions in concurrent object-oriented programming*. Cambridge, MA: The MIT Press.

ARPA Knowledge Sharing Initiative. (1993). *Specification of the KQML agent-communication language – plus example agent policies and architectures*. Retrieved January 30, 2007, from http://www.csee.umbc.edu/kqml/papers/kqml-spec.pdf

Berners-Lee, T., Hendler, J., & Lassila, O. (2001, May). The semantic Web. *Scientific American*, 35-43.

Bond, A. H., & Gasser, L. (Eds.). (1998). *Readings in distributed artificial intelligence*. San Mateo, CA: Morgan Kaufmann Publishers.

Booch, G. (2004). *Object-oriented analysis and design (2nd ed.)*. MA: Addison-Wesley.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., & Orchard, D. (2004, February). *Web services architecture*. W3C working group note 11. Retrieved January 30, 2007, from http://www.w3.org/TR/ws-arch/

Devedzic, V. (2006). *Semantic Web and education*. Berlin, Heidelberg, New York: Springer.

Durfee, E. H., & Lesser, V. (1989). Negotiating task decomposition and allocation using partial global planning. In L. Gasser, & M. Huhns (Eds.), *Distributed artificial intelligence: Volume II* (pp. 229-244). London: Pitman Publishing and San Mateo, CA: Morgan Kaufmann.

FIPA. (1997). *Part 2 of the FIPA 97 specifications: Agent communication language*. Retrieved January 30, 2007, from http://www.fipa.org/specs/fipa00003/OC00003A.html

Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems, 1*(1), 7-38.

Luck, M., McBurney, P., Shehory, O., & Willmott, S. (2005). *Agent technology: Computing as interaction*. Retrieved January 30, 2007, from http://www.agentlink.org/roadmap/al3rm.pdf

Maes, P. (1994) Agents that reduce work and information overload. *Communications of the ACM, 37*(7), 31-40.

Payne, T. R., David, E., Jennings, N. R., & Sharifi, M. (2006). Auction mechanisms for efficient advertisement selection on public displays. In B. Dunin-Keplicz, A. Omicini, & J. Padget (Eds.), *Proceedings of the Fourth European Workshop on Multi-Agent Systems* (in press).

Repenning, A., & Sullivan, J. (2003). The pragmatic Web: Agent-based multimodal Web interaction with no browser in sight. In G. W. M. Rauterberg, M. Menozzi, & J. Wesson (Eds.), *Proceedings of the Ninth International Conference on Human-Computer Interaction* (pp. 212-219). Amsterdam, The Netherlands: IOS Press.

Russel, S. J., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. New Jersey: Prentice-Hall.

Sycara, K., Decker, K., & Williamson, M. (1997). Middle-agents for the Internet. In M. E. Pollack (Ed.), *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (pp. 578-584). Morgan Kaufmann Publishers.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115–152.

Wooldridge, M., & Jennings, N. R. (1998). Pitfalls of agent-oriented development. In K. P. Sycara, & M. Wooldridge (Eds.), *Proceedings of the Second International Conference on Autonomous Agents* (pp. 385-391). Minneapolis: ACM Press.

KEY TERMS

Actuators: Software component and part of the agent used as a mean of performing actions in the agent environment.

Agent Communication Language (ACL): Language used by agents in exchange of messages, defining common syntax for cooperation between heterogeneous agents.

Agent Percepts: Every information that an agent receives through its sensors, about the state of the environment or any part of the environment.

Intelligent Software Agent: An encapsulated computer system that is situated in some environment and that is capable of flexible, autonomous action in that environment in order to meet its design objectives (Wooldridge & Jennings, 1995).

Middle-Agents: Agents that facilitate cooperation among other agents and typically connect service providers with service requesters.

Multi-Agent System (MAS): A software system composed of several agents that interact in order to find solutions of complex problems.

Sensors: Software component and part of the agent used as a mean of acquiring information about current state of the agent environment (i.e., agent percepts).

Intelligent Agents and Their Applications

Alexa Heucke
Munita E.V., Germany

Georg Peters
Munich University of Applied Sciences, Germany

Roger Tagg
University of South Australia, Australia

INTRODUCTION

An agent, in the traditional use of the word, is a person that acts on behalf of another person or group of persons. In information technology, the term *agent* is broadly used to describe software that carries out a special range of tasks on behalf of either a human user or other pieces of software. Such a concept is not new in computing. Similar things have been said about subroutines, reusable objects, components, and Web services. So what makes agents more than just another computer technology buzzword and research fashion?

BACKGROUND

The idea of intelligent agents in computing goes back several decades. Foner (1993, p. 1) dates the first research on software agents to the late 1950s and early 1960s. However, with the breakthrough of the Internet, intelligent agents have become more intensively researched since the early 1990s. In spite of this long heritage, the uptake of these ideas in practice has been patchy, although the perceived situation may be partly clouded by commercial secrecy considerations. Even today, the many different notions of the term *software agent* suggest that the computing profession has not yet reached a generally accepted understanding of exactly what an agent is.

DEFINITIONS AND CLASSIFICATIONS

According to Jennings, Sycara, and Wooldridge (1998, p. 8), "An agent is a computer system, situated in some environment that is capable of flexible autonomous action in order to meet its design objectives." Thus, the determining characteristics of an software agent are:

- **Reactivity:** An agent has profound knowledge of its environment and has the ability to interact directly with it. It can receive input from the outside and can perform reactions with external effects.
- **Autonomy:** An agent is in charge of its own internal status and actions. It can perform independently without the explicit interference of any user or other agents.
- **Proactivity:** An agent has the ability to interpret even minor changes in its environment and can take the initiative to act upon them. It can communicate and interact with entities and can delegate tasks to other agents.
- **Intelligence:** An agent's degree of intelligence is determined by its capability to apply methods of AI in order optimize its action (Meier, 2006, pp. 20-320).

The research literature discusses many different types of agents, carrying out all sorts of functions with what can be termed primary and secondary characteristics. Primary characteristics include autonomy, cooperation, and learning, while secondary characteristics include aspects like multi-functionality, goodwill, or trustworthiness.

A typology of software agents was proposed by Nwana (1996, pp. 7-38):

- **Collaborative agents** feature a high degree of cooperation and autonomy. They are determined by the idea of distributed artificial intelligence and by the concept of task sharing, cooperation, and negotiation between agents.
- **Interface agents** focus on the characteristics of learning and autonomy. By collaborating with the user and by sharing knowledge with other agents, they learn a user's behavior and are trained to take the initiative to act appropriately.
- **Mobile agents** are not static but have the ability to travel. This entails non-functional benefits such as freeing local resources, showing more flexibility, and enabling an asynchronous work scenario.
- **Information or Internet agents** emphasize managing enormous amounts of information. Their main task is to know where to search for information, how to retrieve it, and how to aggregate it.
- **Reactive agents** show a stimulus-response manner as opposed to acting deliberatively. Since they are

based in the physical world and only react to present changes, their behavior is not predetermined.
- **Hybrid agents** comprise more than one agent philosophy and benefit from the combination of different architectures.

Wooldridge and Jennings (1995, pp. 24-30) offer a two-way classification, based on contrasting approaches to building agents. They distinguish the following representative architectures:

- **Deliberative agent architecture:** This classical agent architecture consists of one definite, symbolic world model with all decisions being made on the basis of logical reasoning. Challenges of this approach are the translation of the real world into an accurate model and the establishment of an efficient reasoning.
- **Reactive agent architecture:** In contrast to the deliberative agent architecture, this alternative approach is lacking an explicit and symbolic model of the world as well as extensive reasoning.

Wooldridge and Jennings (1995) also allow for hybrid agent architectures that are built as a hierarchy of deliberative and reactive agent architecture layers.

DISCUSSION

Four aspects are of particular interest when trying to understand how agents work and could be successfully employed in applications and environments: agent knowledge, agent applications, agent standards, and multi-agent systems.

Agent Knowledge

To operate autonomously, any software agent must build up a collection of knowledge, typically data and rules that enable it to serve the principal it is acting for. According to Maes (1994, pp. 2f), an agent's knowledge base should be built up gradually by learning from users and other agents. The key issues are competence and trust. To be competent, the agent must have a knowledge base that is comprehensive and flexible enough to adapt to the user's profile. For an agent to be trusted, a human user must feel comfortable when accepting help from the agent or when delegating tasks to it. Generally, an agent can only learn from its user and other agents if their actions show an iterative pattern. Maes (1994) suggests four different ways of training an agent to build up competence: observation and imitation of the user's habits, user feedback, training by example, and training by other agents.

However, Nwana and Ndumu (1999, p. 10) have criticized Maes' approach, claiming that an agent would not only need to know all peculiarities of the deployed operating system, but also must understand all tasks its user is engaged in. Furthermore, the agent would need to be capable of gathering the user's intent at any time, thus continuously modeling its user. Nwana and Ndumu (1999) identify four main competences for an agent: domain knowledge about the application, a model of its user, strategies for assistance, and a catalog of typical problems that users face in the environment.

Agent Applications

Software agents can be employed in many fields of information technology. One role for agents is to act as an assistant or helper to an individual user who is working with a complex computer system or physical equipment. Examples are:

- **Information agents** (Davies, Weeks, & Revett, 1996, pp. 105-108) that help a human researcher in finding the most relevant material — for example by additionally taking browsing information into consideration (Sharon, Lieberman, & Selker, 2002).
- **Decision support agents** that help a user assess alternative courses of action; functions include filtering and summarization of data, optimizing algorithms, heuristics, and so forth.
- **E-mail agents** (Maes, 1994, p. 5f), which filter spam, allocate incoming mail to folders, and work out addresses to which outgoing mail should be sent.
- **Buying and selling agents**, which assist a user in finding good deals in Internet marketplaces, or *bidding agents* (Morris, Ree, & Maes, 2000), which assist participants in auctions (He, Jennings, & Prügel-Bennett, 2006). These agents have characteristics of information agents as well as of decision support agents.

A second group of applications is where the agent acts as a coordinator of activities, or "virtual manager." Any workflow management system could qualify for this category. Other examples include meeting scheduling agents (Kozierok & Maes, 1993, p. 5), and dynamic scheduling agents that are able to reallocate resources to meet the goals of a business process (Lander, Corkill, & Rubinstein, 1999, p. 1ff). Delegation agents are another example in this category, although they could also be regarded as individual support.

A third group of applications is where the agent continually monitors data and rules in an organization, and on that organization's behalf alerts or sends messages to human recipients. Examples are advertising agents, notification agents, recommendation agents, and selling agents. Such agents are at work when you receive an e-mail from an Internet bookstore about a book that might interest you.

Other agents act as a third party between two humans or pieces of software that need to cooperate. Examples include brokering agents, negotiation agents, mediation

agents, and ontology agents (Helal, Wang, & Jagatheesan, 2001; Pivk & Gams, 2000). An area of application is an electronic marketplace. For example, He, Jennings, and Leung (2003) discussed agent-mediated e-commerce, and Loutchko and Teuteberg (2005) suggested an agent-based electronic marketplace.

Many humans, computer systems, and even other agents depend on one particular specialized task, which is a common agent or subagent, especially useful in an era of information overload. This is a categorization agent (Segal & Kephart, 2000, pp. 2f). Such an agent has the task of applying, and where necessary building up, a classification structure for incoming data. This structure may be particular to an individual (e.g., for e-mail filtering and filing) or it may be for an organizational unit.

Agent Standards

Intelligent agents are intended to function in heterogeneous system environments. To interact smoothly and efficiently in such environments, standardization is essential.

Although agent technology is relatively immature and many researchers still have their own definition of agents, professional bodies have been developing standards for agents since the late 1990s.

These organizations include (Dickinson, 1997):

- ARPA knowledge sharing effort (KSE)
- Agent Society
- OMG Mobile Agent System Interoperability Facility (MASIF)
- The Foundation for Intelligent Physical Agents (FIPA)

The FIPA and MASIF standards are regarded as of special importance for intelligent agents. While the FIPA standard has its origins in the intelligent agent community and has been influenced by the KQML (Knowledge Query and Manipulation Language), MASIF deals primarily with agent mobility.

The FIPA 2000 standard (*www.fipa.org*) specification deals with mobility and tries to integrate MASIF (Milojicic et al., 1998, pp. 50-67). Therefore this specification bridges the gap between the intelligent and mobile agent communities. FIPA's specification is divided into five main categories: applications, abstract architecture, agent communication, agent management, and agent message transport.

Multi-Agent Systems

Much of the recent literature on agents envisages a system with a community of agents that cooperate in some way to achieve an overall set of goals. According to Jennings et al. (1998, pp. 9, 17f), "Multi-agent systems are ideally suited to representing problems that have multiple problem-solving methods, multiple perspectives and/or multiple problem solving entities." Since each agent has a restricted view of any problem and only limited information, multi-agent systems feature a flexible and advanced infrastructure to solve issues beyond individual capabilities. Thus, the system can benefit from every agent's expert knowledge. Other characteristics of multi-agent systems are decentralized data, asynchronous computation, and the lack of a central control system.

A major challenge of multi-agent systems is clearly the means of coordination between agents. Nwana, Lee, and Jennings (1997, pp. 33-55) have identified the following key components in such coordination: foreseeable structures of agent interaction, defined agent behavior and social structures, flexibility and dynamics, and the knowledge and reasoning to utilize the above.

Possible coordination techniques are:

- **Organizational structuring:** The agents' roles, responsibilities, and their communication chains and paths of authority are defined beforehand.
- **Contracting:** All tasks and resources distributed among agents are controlled by a contract net protocol.
- **Multi-agent planning approach:** Agents decide on a detailed and interlaced plan of all activities and goals aimed at. Multi-agent planning can be centralized with one agent reviewing all individual plans and coordinating them into a single — or a distributed — multi-agent plan.

When agents are interacting in a multi-agent system, they may have to negotiate in order to fulfill their interests. Nwana et al. (1997) suggest two different negotiation theories:

- In the **Game Theory-based negotiation** approach, each agent holds a utility matrix that lists how much a certain interaction or goal is worth. During the negotiation process, which is defined by a negotiation protocol, the parties exchange bids and counteroffers following their strategies.
- In the **Plan-based Negotiation Theory**, each agent schedules its actions individually before all plans are coordinated. This is similar to the multi-agent planning coordination approach, but any agent can play the role of central coordinator.

For recent progress on negotiating agents, the reader is referred to Luo, Jennings, and Shadbolt (2006) or Fatima, Wooldridge, and Jennings (2006).

FUTURE TRENDS

We believe the area of support for human users in the carrying out of highly heterogeneous workloads represents a

promising area for the development of agent applications (see e.g. Padgham & Winikoff, 2004, for a practical guide to developing intelligent agent systems). The current support for users who work with a mixture of word processing, e-mail, spreadsheets, databases, digital libraries, and Web search tools is very primitive. The user has to do most of the work in correlating the different sources, and current tools are poor at learning the user's commonly repeated work patterns. The authors of this article feel that agents are the most promising technology to redress this shortcoming, and we have worked on architecture for linking agents with tools such as Groupware and Workflow.

CONCLUSION

In spite of a considerable amount of research, the killer application for intelligent agents is still somewhat elusive. The IT industry still has not reached a consensus about the use of agents now and in the future. Nwana and Ndumu (1999, p. 14) are even more critical, claiming that "not much discernible progress has been made post 1994." The main reason might be that intelligent agent theory integrates some of the most challenging concepts in science, including artificial intelligence (AI), data mining, or contract theory. Agent technology can be considered another new demanding application of these concepts and will succeed or fail depending on any progress in these areas. The take-up of agent technology is therefore likely to suffer the same ups and downs that AI has experienced in recent decades. In the longer term, however, there is a large area of opportunity for agents supporting human users of computer systems which has yet to be fully developed.

REFERENCES

Davies, J., Weeks, R., & Revett, M. (1996). Information agents for the World Wide Web. *BT Technology Journal, 14*(4), 105-114.

Dickinson, I.J. (1997). *Agents standards.* HP Labs Technical Report HPL-97-156, Hewlett-Packard, USA. Retrieved January 20, 2007, from http://www.hpl.hp.com/techreports/97/HPL-97-156.html

Fatima, S.S., Wooldridge, M., & Jennings, N.R. (2006). Multi-issue negotiation with deadlines. *Journal of Artificial Intelligence Research, 27*, 381-417.

Foner, L. (1993). *What's an agent, anyway? A sociological case study.* Agents Memo 93-01, Agents Group, MIT Media Lab, USA. Retrieved January 20, 2007, from http://foner.www.media.mit.edu/people/foner/Reports/Julia/Agents--Julia.pdf

He, M., Jennings, N.R., & Leung, H.F. (2003). On agent-mediated electronic commerce. *IEEE Transactions on Knowledge and Data Engineering, 15*(4), 985-1003.

He, M., Jennings, N.R., & Prügel-Bennett, A. (2006). A heuristic bidding strategy for buying multiple goods in multiple English auctions. *ACM Transactions on Internet Technology, 6*(4), 465-496.

Helal, S., Wang, M., & Jagatheesan, A. (2001). Service-centric brokering in dynamic e-business agent communities. *Journal of Electronic Commerce Research.* Retrieved January 10, 2007, from *http://www.icta.ufl.edu/projects/publications/BPtemp148.pdf*

Jennings, N.R., Sycara, K., & Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous Agents and Multi-Agent Systems, 1*, 7-38.

Kozierok, R., & Maes, P. (1993). A learning interface agent for scheduling meetings. *Proceeding of the 1st International Conference on Intelligent User Interfaces* (pp. 81-88), Orlando, FL.

Lander, S., Corkill, D., & Rubinstein, Z. (1999). KPM: A tool for intelligent project management and execution. *Proceedings of the 16th International Joint Conference on Artificial Intelligence, Workshop on Intelligent Workflow and Process Management,* Stockholm, Sweden.

Loutchko, I., & Teuteberg, F. (2005). An agent-based electronic job marketplace: Conceptual foundations and fuzzyMAN prototype. *Computer Systems Science and Engineering, 20*(4), 295-309.

Luo, X., Jennings, N.R., & Shadbolt, N. (2006). Acquiring user strategies and preferences for negotiating agents: A default then adjust method. *International Journal of Human Computer Studies, 64*(4), 304-321.

Maes, P. (1994). Agents that reduce work and information overflow. *Communications of the ACM, 37*(7), 31-40.

Meier, P. (2006). *Agentenkomponenten.* Munich: Martin Meidenbauer.

Milojicic, D., Breugst, M., Busse, I., Campell, J., Covaci, S., Friedman, B., Kosaka, K., Lange, D., Ono, K., Oshima, M., Tham, C., Virdhagris, S., & White, J. (1998). MASIF: The OMG Mobile Agent System Interoperability Facility. *Mobile Agents,* 50-67.

Morris, J., Ree, P., & Maes, P. (2000). Sardine: Dynamic seller strategies in an auction marketplace. *Proceedings of the ACM Conference on Electronic Commerce* (pp. 128-134), Minneapolis, MN.

Nwana, H. (1996). Software agents: An overview. *The Knowledge Engineering Review, 11*(3), 205-244.

Nwana, H., Lee, L., & Jennings, N. (1997). Coordination in multi-agent systems. In H. Nwana & N. Azarmi (Eds.), *Software agents and soft computing. Towards enhancing machine intelligence* (pp. 42-58). Berlin: Springer-Verlag.

Nwana, H., & Ndumu, D. (1999). A perspective on software agents research. *The Knowledge Engineering Review, 14*(2), 125-142.

Padgham, L., & Winikoff, M. (2004). *Developing intelligent agent systems: A practical guide.* Chichester: John Wiley & Sons.

Pivk, A., & Gams, M. (2000). E-commerce intelligent agents. *Proceedings of the 3rd International Conference on Telecommunications and Electronic Commerce* (pp. 418-429), Dallas, TX.

Segal, R., & Kephart, J. (2000). Incremental learning in SwiftFile. *Proceedings of the 17th International Conference on Machine Learning* (pp. 863-870), Stanford, CA.

Sharon, T., Lieberman, H., & Selker, T. (2002). Searching the Web with a little help from your friends. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work*, New Orleans, LA.

Wooldridge, M. (2002). *An introduction to multi agent systems.* Chichester: John Wiley & Sons.

Wooldridge, M., & Jennings, N. (1995). Intelligent agents: Theory and practice. *The Knowledge Engineering Review, 10*(2), 115-152.

KEY TERMS

Artificial Intelligence: Computer systems that feature automated human-intelligent, rational behavior and employ knowledge representation and reasoning methods.

Business Process: A process at the business layer of an organization. Since the 1990s, the focus of any business reengineering project and one of the central inputs for IT design. Also called *workflow.*

Categorization: The process of deducing, from the content of an artifact, the potentially multiple ways in which the artifact can be classified for the purpose of later retrieval from a database, library, collection, or physical storage system.

Contract Theory: Theory dealing with aspects of negotiation and contracting between two or more parties.

Data Mining: Integrating statistics, database technology, pattern recognition, and machine learning to generate additional value and strategic advantages.

Game Theory: Mathematical theory of rational behavior for situations involving conflicts of interest.

Workflow: The automation of a business process, in whole or part, during which documents, information, or tasks are passed from one participant to another for action, according to a set of procedural rules. Also called *business process.*

Intelligent Software Agents in E-Commerce

Mahesh S. Raisinghani
TWU School of Management, USA

Christopher Klassen
University of Dallas, USA

Lawrence L. Schkade
University of Texas at Arlington, USA

INTRODUCTION

Agent technology is one of the most widely discussed topics in information systems and computer science literature. New software products are being introduced each day. A growing number of computer information professionals recognize that there are definite issues surrounding intelligent agent terminology. These must be resolved if agent technology is to continue to develop and establish.

Current research on intelligent agent software technology can be categorized as two main areas: technological and social. In the excitement of emergent technology, people often forget to scrutinize how new technology may impact their lives. The social dimension of technological progress is the driving force and most central concern of technology. Technology is not created for its own sake as a technological imperative. This article critiques the current state of software intelligent agents by examining technological issues and the social implications of intelligent agent software technology.

TECHNOLOGICAL ISSUES

An attempt to arrive at a generally accepted definition is the first hurdle. In order for this term to have any effectiveness, there must first be a universal definition that can be agreed upon and used consistently. Unfortunately, there is none. Many proposals for defining an "intelligent agent" have been put forth, but none has received wide acceptance. Some of these proposals are the following:

- "An agent is anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors." Russell and Norvig (1995)
- "Let us define an agent as a persistent software entity dedicated to a specific purpose. 'Persistent' distinguishes agents from subroutines; agents have their own ideas about how to accomplish tasks, their own agendas. 'Special purpose' distinguishes them from other entire multifunction applications; agents are typically much smaller."
- "An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future." Franklin and Graesser (1996)

While these terms attempt to describe characteristics of intelligent agents, no comprehensive and generic description for these agents has gained wide recognition as the definitive description of a software agent. A consensus definition has not yet been achieved. As Franklin and Graesser (1996) indicate, most of the definitions proposed are derived from conceptualizations peculiar to the subjective views of the individuals. It is important to note that it is this intuitive aspect of an "intelligent agent" which makes it difficult to establish a broadly accepted formal definition. Ironically, it facilitates marketing of intelligent agent software technology.

A second reason for a lack of a consensus definition is that much of the agent research is proprietary. Companies that make investments to sponsor such research do not wish to reveal their competitive edge nor give away the value of their work. Standardization of new technology is difficult. Uncertainty will continue until the companies and individuals with the proprietary information recognize that sharing knowledge benefits everyone.

A third reason for the difficulty for the lack of a generally approved definition of an intelligent software agent, and probably the most important reason of the three outlined in this article, is that intelligent agent software does not seem to be qualitatively different from other software. "Is it an agent, or just a program?" Franklin and Graesser (1996) ask and observe, correctly, that all software agents are programs. The authors also state that not all programs are agents. The implication is that some programs are, in fact, agents. If an "intelligent agent" were just an added complex program, the term "intelligent agent" would simply mean that a software program was simply extended, made more composite

and possibly more useful than other typical programs. An intelligent agent differs from a procedural program in two ways. First, it is an agent and broadly speaking, it is defined as someone or something that acts. To be able to act, the entity must have a purpose or a goal. A computer program can only perform a prescribed set of instructions. An intelligent software agent has the same capability and is similar to a computer program in this respect.

Computer programs act utilizing a relatively low level of logic. These programs cannot act autonomously. For any entity to act with autonomy there must be concomitant independence and freedom. Procedural computer programs do not have volition, because whatever is written into the program is executed. The key factor is logic bound and a closed program. The term "react" is an inherent limitation of closed computer programs. An agent, in the true sense of the word, initiates action. The several reasons illustrate why some time is required for an acceptable definition of software agent. This process is likely to be somewhat similar to the emergence of the distinction between artificial intelligence, expert systems and decision support systems, which became clear gradually with more widespread usage of this distinctive software.

We do not demean the effort that has been invested into these products. Systems that are based on the detection of patterns in conjunction with explicit user commands and preferences are based on straightforward computational mathematics and logic. Technical challenges exist in the areas of security, connectivity, storage, peer group collaboration, network-based services, user interface, stability and standards (Bantz et al., 2003). Park and Park (2003) propose an agent-based system for merchandise management and verify its application in a duty-free shop, which performs evaluating and selecting merchandise and predicting seasons and building purchase schedules autonomously in place of human merchandise managers under a business-to-business (B2B) electronic commerce (EC) environment. In order to facilitate the agent's intelligent behavior, several analysis tools such as data envelopment analysis (DEA), genetic algorithm (GA), linear regression and rule induction algorithm are incorporated into the system.

E-mail and filters reject messages that do not comply to the user's defined preferences. Help engines and data warehousing tools search for built-in patterns. Patterns are pre-built into the engines, which are limited by the closed logic bounds specified by the designs. News and searching tools have a great potential, albeit they pose a concern. The dilemma is if many users have news searching intelligent agent tools constantly searching for information on the Internet, the Internet may possibly be clogged up by too many of these searching tools. Imagine if one of these intelligent agents had a built-in error (bug) that caused the program to continuously spawn even more agents to search the Internet. Moreover, some intelligent agents searching the Internet for information could get lost and not return with the requested information. Thus, one can see the latent technical threat in employing such ill-designed intelligent agents. These "lost" agents may create severe bottlenecks on the Internet.

Although intelligence means people thinking, it may be possible to replicate the same set of behaviors using computation. This idea was discussed by Turing in the 1940s. In 1950 he proposed a test, now called the Turing Test (TT), for computational intelligence. In the test, a human judgment must be made concerning whether a set of observed behaviors is sufficiently similar to human behaviors that the same word - intelligent - can justifiably be used. Feigenbaum (2003) discuss the challenges for computational intelligence, including: 1. an alternative to TT that tests the facet of quality (the complexity, the depth) of reasoning, 2. building a large knowledge base by reading text, thus reducing knowledge engineering effort, and 3. distilling from the WWW a huge knowledge base, thus reducing the cost of knowledge engineering.

Olin et al. (2001) suggest that although tools in the shape of distributed artificial intelligence will be available and be particularly applicable to the complexities of decision-making in the typical global enterprise, a number of issues will arise in the next few years as intelligent agents become the mainstream enablers of "real-time" enterprise process, such as the need to ensure continuity in the transition phase by careful integration of intelligent agent systems in to a legacy systems.

Van Den Heuvel and Maamar (2003) propose a framework for contract-based support to establish virtual collaboration using loosely coupled and heterogeneous intelligent Web services (IWS) in which contracts encapsulate the control information for IWSs engaged in e-business transactions. Since IWS technology is still in its infancy, several important issues must be addressed before agentified Web services can be successfully deployed at e-marketplaces, such as the integration of IWSs with wrapped legacy systems, the semantic integration of Web services, and the integration of Web services into e-commerce transactions (Van Den Heuvel & Maamar, 2003).

While search engines have become the major decision support tools for the Internet, there is a growing disparity between the image of the World Wide Web stored in search engine repositories and the actual dynamic, distributed nature of Web data. The traditional static methods in which search and retrieval are disjoint seem limited. Menczer (2003) proposes using an adaptive population of intelligent agents mining the Web online at query time and presents a public Web intelligence tool called MySpiders, a threaded multi-agent system designed for information discovery and augmenting search engines with adaptive populations of intelligent search agents for significant competitive advantage. Weippl et al. (2003) describe the manifold security threats that occur in mobile agent systems. For instance, when masquerading, an

agent claims the identity of another agent and tries to gain unauthorized access to resources. A possible solution to this security flaw is signing agent code with digital signatures such as those provided by the Java cryptographic extension (JCE 1.2.1). Mobile agents could launch denial of service attacks since too many agents could flood the Java runtime environment. Another relevant issue that needs to be addressed in the future is database intelligent agents' independence of the Java agent platform used for transport agents. Due to the lack of support offered by major Java agent platforms, the unconfined employment of standardized interfaces is not yet possible. Finally the issue of multiple, heterogeneous data warehouse islands that need to be made interoperable is being investigated by using intelligent agents to realize the federation. For instance, Xu et al. (2003) propose a two-layer approach for the formal modeling of logical agent mobility (LAM) using predicate/transition (PrT) nets in which a mobile agent system is viewed as a set of agent spaces and agents could migrate from one space to another.

SOCIAL IMPLICATIONS

The social implications, as with any new technology, include both positive and negative aspects. We are aware that the current literature debate and research on intelligent agent software technology deals in moderation with the topic of the social and ethical implications. Technologies are human artifacts, developed, presumably, for the benefit of humankind and the improvement of the quality of life. New technologies must, however, be tested to determine if the product meets these. Most important is to check if the new technology is significantly different from the existing ones.

One of the major benefits from using intelligent agents is the potential for liberating humans from the tedious task of searching for information on the Internet. The intelligent agent can aid in the search by filtering information that has little or no value. Unfortunately, intelligent agents also have the potential for damage. First, with excessive reliance on intelligent agents, humans can risk an excessive loss of choice. According to Lanier (1999), confining to an artificial world created by a programmer(s) can limit human potential for innovation. Another objection Lanier raises is that human beings degrade themselves by using intelligent agents. When individuals begin to think of computers anthropomorphically, as actually possessing intelligence and autonomy, people will tend to relate to computers as if they were human. The opposite is true too. This is a serious dilemma that must be avoided. An additional technical problem that Lanier (1999) raises is that info-consumers see the world through agents' eyes. The point is that if intelligent agents are used to find useful information, the agents themselves may be manipulated.

FUTURE TRENDS

The technological, social, and ethical issues notwithstanding, the new business environment is characterized not only by a rapid pace of change but also the dynamically discontinuous nature of such change. The issues such as lack of standardization in mobile agents may cause lack of identity traceability due to multiple transfers among networks. The security concerns relate to machine protection without artificially limiting agent access rights. Finally, there are issues surrounding performance and scalability, such as the performance effects that high levels of security would have on the network, as well as the effects of having multiple mobile agents in the same system. The emergence of intelligent mobile/software agents not only will change the way that we communicate across networks, but also have a profound impact on the way that we accomplish many tasks.

CONCLUSION

In conclusion, while intelligent agent technology has the potential for being useful to humankind, many fundamental issues remain unsolved. These problems are technical, social and/or ethical and require careful thought and consideration by developers. This discussion is critical of the current state of intelligent agent software technology and research, in the hope of encouraging developers to be aware of spillover effects. Unfortunately, issues raised by authors such as Lanier (1999) are often thrust aside as being extremist. Intelligent agent software technology has advanced, but greater progress must be made socially and ethically before these agents can be accepted as tools for the improvement of the quality of human life.

REFERENCES

Bantz, D.F., Bisdikian, C., Challener, D., & Karidis, J.P. (2003). Autonomic personal computing. *IBM Systems Journal, Armonk: 42*(1), 165-177.

Feigenbaum, E.A. (2003, January). Some challenges and grand challenges for computational intelligence. *Journal of the Association for Computing Machinery, 50*(1), 32-42. New York: Association for Computing Machinery.

Franklin, S., & Graesser, A. (1996). Is it an agent, or just a program?: A taxonomy for autonomous agents. *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages.* Springer-Verlag.

Lanier, J. (1999). Agents of alienation. Retrieved May 18, 2000, from *http://www.well.com/ user/ jaron/ agent alien. html*

Menczer, F. (2003, May). Complementing search engines with online Web mining agents. *Decision Support Systems, 35*(2), 195-206.

Olin, J., Greis, N., & Morgan, L. (2001). The enterprise at the edge: Agents to the rescue. *European Management Journal, 19*(5), 489-501.

Park, J.H., & Park, S.C. (2003, June). Agent-based merchandise management in business-to-business electronic commerce. *Decision Support Systems, 35*(3), 311-324.

Russell, S.J., & Norvig, P. (1995). *Artificial intelligence: A modern approach.* Englewood Cliffs, NJ: Prentice Hall.

Smith, D.C., Cypher, A., & Spohrer, J. (1994). KidSim: Programming agents without a programming language. *Communications of the ACM, 37*(7), 55-67.

The IBM Agent. *http://activist.gpl.ibm.com: 81/WhitePaper/ptc2.htm*

Van Den Heuvel, W.-J., & Zakaria, M. (2003, October). Moving toward a framework to compose intelligent Web services. *Communications of the ACM, 46*(10), 103-114. New York: Association for Computing Machinery.

Vinaja, R., & Sircar, S. (1999). Agents delivering business intelligence. *Handbook of information technology* (pp. 477-490). CRC Press.

Weippl, E., Klug, L., & Essmayr, W. (2003, January-March). A new approach to secure federated information bases using agent technology. *Journal of Database Management, 14*(1), 48-69.

Xu Dianxiang, Y., Jianwen, D.Y., & Ding, J. (2003, January). A formal architectural model for logical agent mobility. *IEEE Transactions on Software Engineering, 29*(1), 31-42. New York.

KEY TERMS

Agent: A program designed to provide specialized and well defined services. Agent can be *static* – executing on the computer where it was installed, or *mobile* – executing on computer nodes in a network.

Autonomous Agent: A system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future.

Bottlenecks: A stage in a process that causes the entire process to slow down or stop.

Data Warehousing: A form of data storage geared towards business intelligence. It integrates data from various parts of the company. The data in a data warehouse are read-only and tend to include historical as well as current data so that users can perform trend analysis.

Decision Support Systems: A specific class of computerized information system that supports business and organizational decision-making activities. DSS is an interactive software-based system that compiles useful information from raw data, documents, personal knowledge, and/or business models to identify and solve problems and make decisions

Expert Systems: A computer system that facilitates solving problems in a given field or application by drawing inference from a knowledge base developed from human expertise. Some expert systems are able to improve their knowledge base and develop new inference rules based on their experience with previous problems.

Intelligent Agent: A program that gathers information or performs some other service without your immediate presence and on some regular schedule.

Turing Test (TT), for computational intelligence: In the test, a human judgment must be made concerning whether a set of observed behaviors is sufficiently similar to human behaviors that the same word - intelligent - can justifiably be used.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1603-1606, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Intelligent Technologies for Tourism

Dimitris Kanellopoulos
University of Patras, Greece

INTRODUCTION

Nowadays, the tourism industry is a consumer of a diverse range of information (Buhalis & O'Connor, 2005). Information communication technologies (ICTs) play a critical role for the competitiveness of tourism organizations and destinations. According to Staab and Werthner (2002), ICTs are having the effect of changing:

- The ways in which tourism companies contact their business; reservations and information management systems;
- The ways tourism companies communicate; how customers look for information on, and purchase travel goods and services.

In the tourism industry, the supply and demand sides form a worldwide network in which tourism product's generation and distribution are closely worked together. Most tourism products (e.g., hotel rooms or flight tickets) are time constrained and nonstockable. Generally, the tourism product is both "perishable" and "complex," and itself is a bundle of basic products aggregated by intermediaries. Consequently, basic products must have well-defined interfaces with respect to consumer needs, prices, or distribution channels. In addition, a tourism product cannot be tested and controlled in advance. During decision-making, only an abstract model of the product (e.g., its description) is available. Besides, the tourism industry has a heterogeneous nature, and a strong small and medium-sized enterprises (SMEs) base. Undoubtedly, intelligent technologies are increasingly changing the nature of, and processes in, the tourism industry. This chapter reviews, in brief, such technologies applied to the e-tourism domain.

BACKGROUND

E-tourism is defined as the use of ICTs in the tourism industry. It involves the buying and selling of tourism products and services via electronic channels, such as the Internet, cable TV, and so forth. E-tourism includes all intranet, extranet, and Internet applications, as well as all the strategic management and marketing issues related to the use of technology. ICTs include the entire range of electronic tools that facilitate the operational and strategic management of organizations by enabling them to manage their information, functions, and processes, as well as to communicate interactively with their stakeholders for achieving their mission and objectives. Currently, e-tourism makes use of (syntactic) Web technology for tours, infrastructure, related interesting information, such as public transport, timetables, weather, online reservation, and so forth. However, the major barriers using the syntactic Web are:

- Creating complex queries involving background knowledge on tourism issues.
- Solving ambiguities and synonyms.
- Finding and using Web services for tourism

From another perspective, the characteristics of the tourism product require information on the consumers' and suppliers' sides, involving high information search costs and causing informational market imperfections. These outcomes sequentially lead to the establishment of specific product distribution information and value-adding chains. Given such a framework, Staab and Werthner (2002) state that intelligent Information Systems (ISs) should:

- Be heterogeneous, distributed, and cooperative.
- Enable full autonomy of the respective participants.
- Support the entire consumer life cycle and all business phases.
- Allow dynamic network configurations.
- Provide intelligence for customers (tourists) and suppliers as well as in the network.
- Be scalable and open.
- Focus on mobile communication enabling multichannel distribution.

Hereafter, we present intelligent technologies for tourism.

INTELLIGENT TECHNOLOGIES FOR TOURISM

Web intelligence combines two topics: (1) *Web analytics*, which examines how Web site visitors view and interact with a Web site's pages and features; and (2) *business intelligence*, which allows a corporation's management to use data on customer purchasing patterns, demographics, and

demand trends to make effective strategic decisions (Zhong, 2003). In tourism, the developments of artificial intelligence (AI) are at the cutting edge. Many applications are provided to the users, such as individualized pricing (http://www.priceline.com), reversed multiattribute auctioning (http://www.mytraveldream.com), recommendations in bundling products, and semantic Web, as well as mobile applications (Kanellopoulos, 2006; Kanellopoulos & Kotsiantis, 2006). Using the Web, travelers can get information on routes, timetables, seat availabilities, accommodations, rental cars, and restaurants to help them plan their travels. Remarkable progress has been made in the automation of travel planning with the help of the easily accessible information. There are also many semiautomated commercial service Web sites like travelocity.com, expedia.com, and orbitz.com (Paprzychi, Gilbert, & Gordon, 2002).

WEB SERVICES

The Web services technology is a set of standards that could allow Web applications for the tourism domain to "talk" to each other over the Internet. These standards are:

- XML (eXtensible Markup Language: http://www.w3.org/XML/) for driving applications services.
- SOAP (Simple Object Access Protocol: http://www.w3.org/TR/soap) for communication.
- WSDL (Web Services Description Language: http://www.w3.org/TR/wsdl/) as the service description language.
- UDDI (Universal Description, Discovery and Integration: http://www.uddi.org/) as the service discovery protocol.

The Web services technology offers distributed tourism services capability over a network (Ouzzani & Bouguettaya, 2004). The platform- and language-independent interfaces of Web services allow the easy integration of heterogeneous tourism ISs. Web services offer mechanisms for describing tourism-related Web documents, methods for accessing them, and discovery methods that enable the identification of relevant Web service providers. Recently, the OTA (Open Travel Alliance) has developed open data transmission specifications for the electronic exchange of business information for the travel industry, including but not limited to the use of XML.

SEMANTIC WEB

The *Semantic Web* is an extension of the current Web in which information is given well-defined meaning. It enables computers and users to work in cooperation, and allows the tourism content to become semantic annotated (Kanellopoulos, 2006). This characteristic allows users and software agents to query and infer knowledge from Web tourism information quickly and automatically. The semantic Web is based on formal domain models (ontologies) that define domain specific conceptualization and impose description on the domain knowledge structure and content. An ontology comprises the classes of entities, relations between entities, and the axioms that apply to the entities of the domain. Ontologies can provide a shared understanding of the tourism domain to sustain communication among users and software agents typically being represented in a machine-processable representation language like OWL (Web Ontology Language, http://www.w3.org/2004/OWL/). Through the use of metadata organized in several interrelated ontologies, information concerning tourism objects (e.g., hotels, attractions) can be tagged with descriptors that facilitate its retrieval, analysis, processing, and reconfiguration. Introducing semantics to Web services for tourism brings the following advantages:

- Ontologies offer a promising infrastructure to cope with heterogeneous representations of Web documents (Chandrasekaran, Josephson, & Benjamins, 1999). Semantically enriched Web services can handle the interoperability at the technical level, that is, they make Web applications "talk" to each other independent of their hardware and software platforms (Dell' Erba, 2004).
- Semantics can be used for the discovery and composition of Web services.
- The main mechanism for service discovery is *service registries,* and semantics can be used for the discovery of registries of Web services.

ONTOLOGIES FOR TOURISM

Tourism ontologies allow machine-supported tourism data interpretation and integration. The *e-tourism* ontology (http://e-tourism.deri.at/ont/) was deployed in the OnTour project, and describes the domain of tourism using OWL. It focuses on accommodation and activities, and it is based on an international standard: the "*Thesaurus on Tourism & Leisure Activities*" of the World Tourism Organization (WTO). This thesaurus is an extensive collection of terms related to tourism. The ISO 18513 standard ("tourism services–hotel and other types of tourism accommodation–terminology") defines terms used in tourism in relation to the various types of tourism accommodation and other related services. MONDECA's tourism ontology (http://www.mondeca.com) defines tourism concepts based on the WTO thesaurus. These concepts include terms for tourism object profiling, tourism and cultural objects, tourism packages, and tourism multimedia content. A reference ontology, named

COTRIN (Comprehensive Ontology for the Travel Industry) is presented in Cardoso and Lang (2007). The objective of COTRIN ontology is the implementation of the semantic XML-based OTA specifications. Major airlines, hoteliers, car rental companies, leisure suppliers, travel agencies, and others may use COTRIN to bring together autonomous and heterogeneous tourism Web services, Web processes, applications, data, and components residing in distributed environments.

From another perspective, a destination management system (DMS) provides complete and up-to-date information on a particular tourist destination. DMSs is a perfect application area for semantic Web and P2P (peer-to-peer) technologies. Kanellopoulos and Panagopoulos (2008) developed an ontology for tourist destinations in the LA_DMS (layered adaptive semantic-based DMS based on P2P technologies) project. The aim of LA_DMS project was to enable DMSs adaptive to tourists' needs concerning destination information. The LA_DMS system incorporates a metadata model to encode semantic tourist destination information in an RDF-based P2P network architecture. This metadata model combines ontological structures with information for tourist destinations and peers. Kanellopoulos and Kotsiantis (2007) proposed a semantic-based architecture in which semantic Web ontology is used to model tourist destinations, user profiles, and contexts. The semantic Web service ontology (OWL-S) is extended for matching user requirements with tourist destination specifications at the semantic level, with context information taken into account. Semantic Web rule language (SWRL) is used for inferencing with context and user profile descriptions. Their architecture enables DMSs to become fully adaptable to user's requirements concerning tourist destinations.

In the group package tour domain, an intelligent Web portal was proposed that helps people living in Europe to find package tours that match their personal traveling preferences (Kanellopoulos, 2008). For this purpose, the knowledge of the package tour domain has been represented by means of ontology.

The HARMONISE (http://www.harmonise.org) is an EU tourism harmonisation network (THN) established by the ECommerce and Tourism Research Laboratory, IFITT (International Federation for IT and Travel & Tourism), and others. The HARMONISE project allows participating tourism organizations to keep their proprietary data format and use ontology mediation while exchanging information (Fodor & Werthner, 2005). HARMONISE is an ontology-based mediation and harmonization tool that establishes the bridges between existing and emerging online marketplaces.

In the SATINE project (semantic-based interoperability infrastructure for integrating Web service platforms to P2P networks), a secure semantics-based interoperability framework was developed for exploiting Web service platforms in conjunction with P2P networks in the tourist industry (Dogac, Kabak, Laleci, Sinir, Yildiz, Kirbas, & Gurcan, 2004).

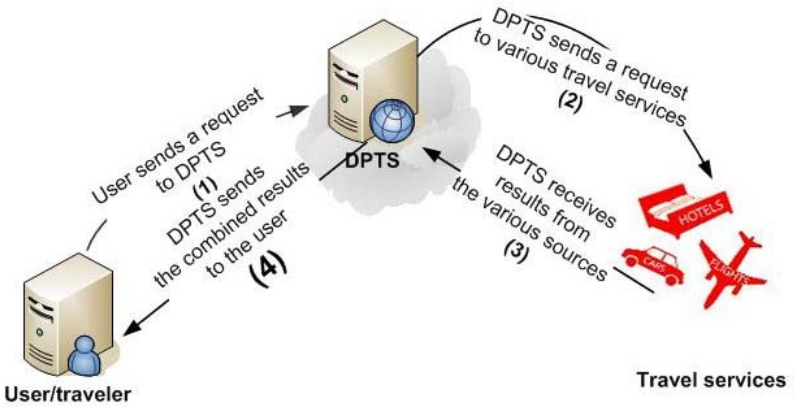
INTELLIGENT SOFTWARE AGENTS

The semantic Web includes intelligent software agents that "understand" semantic relationships between Web resources, and seek relevant information as well as perform transactions for users. Intelligent agents can provide various tourism products and services into an integrated tourism package that can be personalized to a tourist's needs. A variety of traveler, hotel, museum, and other agents can enhance the tourism marketing and management reservation processes (Kanellopoulos, 2006). There are many research prototypes of intelligent travel support systems based on software agent technology (Camacho, Borrajo, & Molina, 2001). Traveler software agents can assist travelers in finding sources of tourism products and services, and in documenting and archiving them. A set of agents can be deployed for various tasks including tracking visitor schedules, monitoring meeting schedules, and monitoring user's travel plans. For example, if the user specifies the travel itinerary and his/her required services, then a set of information agents can be spawned to perform the requested monitoring activities (Camacho et al., 2001). An additional capacity of the semantic Web is realized when intelligent agents extract information from one application and subsequently, utilize the data as input for further applications (Kanellopoulos, 2006). Therefore, software agents can create greater capacity for large-scale automated collection, processing, and selective dissemination of tourism data.

DYNAMIC PACKAGING SYSTEMS

A package tour consists of transport and accommodation advertised and sold together by a vendor known as a tour operator. Tour operators provide various services like a rental car, activities, or outings during the holiday. Consumers can acquire packages from a diversity of Web sites including online agencies and airlines. The objective of *dynamic packaging* is to pack all the components chosen by a traveler to create one reservation. Regardless of where the inventory originates, the package that is created is handled seamlessly as one transaction, and requires *only one* payment from the consumer. Cardoso and Lang (2007) proposed a framework and a platform to enable dynamic packaging using semantic Web technologies. A dynamic packaging application allows consumers and/or travel agents to bundle trip components. The range of products and services to be bundled is too large: guider tour, entertainment, event/festival, shopping, activity, accommodation, transportation, food and beverage,

Figure 1. The operation of DPTS



and so forth. Figure 1 depicts the operation of a dynamic packaging tour system (DPTS).

TRAVEL RECOMMENDER SYSTEMS

Recommender systems can predict what the user needs based on the information provided by the user (Ricci, 2004). A Web-based recommendation system was developed in the intelligent Recommendation for Tourist Destination Decision Making project (DieToRecs) (http://etd.ec3.at). This system helps the tourist destination selection process and accommodates individual traveler's preferences. Based on the user profiles, personalized recommendations are created to support potential tourists to choose their ideal destination. The Travel Recommender System (Trip@dvice) (http://tripadvise.itc.it) assists travelers in their search for tourism products and services. A prototype called NutKing (http://itr.itc.it) is available. Using the WAP (wireless application protocol), the Mobile Tourism Recommender System (mITR) (http://mobile.itc.it) implements mobile tourism services such as airlines (reservations, check-in, flight status, etc.), hotels and restaurants (reservations), maps, transportation (schedules, connections etc.), traffic, and weather conditions. Context-aware applications, such as mobile tourism guides, utilize contextual information, such as location, display medium, and user profile, in order to provide tailored functionality to the end-user.

MOBILE TOURISM GUIDES

A user interacts with a mobile tourism guide using a "map." Many mobile tourism guides have been deployed. The COMPASS (COntext-aware Mobile Personal ASSistant) guide provides tourists with context-aware recommendations and services (Van Setten, Pokraev, & Koolwaaij, 2004). The GUIDE system (Schmidt-Belz, Polsad, Nick, & Zipf, 2002) provides tourists with context-aware information about a city via a PDA (personal digital assistant). It is based on a client/server architecture and utilizes software agent technology, with a Fujitsu TeamPad 7600 used as a terminal. CRUMPET (CReation of User-friendly Mobile services Personalized for Tourism) is an EU project that developed a guide system that provides mobile services personalized for tourism. In the CRUMPET framework, tourism-related value-added services for nomadic users (across mobile and fixed networks) are provided. In most mobile guides, the characteristics of "context" are categorized into: (1) scope of context, (2) its representation and acquisition, as well as (3) the access mechanism used.

UBIQUITOUS COMPUTING AND AMBIENT INTELLIGENCE FOR TOURISM

Ubiquitous computing is a model of human-computer interaction in which information processing has been thoroughly integrated into everyday objects and activities. Recently, several projects have focused on ubiquitous computing. Such projects include Xeroz PARC's ubiquitous computing, IBM's pervasive computing, and MIT's Oxygen initiative. An interesting ubiquitous travel service delivery system is presented in O'Brien and Burmeister (2003).

Ambient intelligence is the convergence of ubiquitous computing and communication, and intelligent user-friendly interfaces. Ambient intelligent systems are embedded,

personalized, adaptive, and anticipatory, as well as they provide access for tourists, anywhere, at any time (Manes, 2003). In an ambient intelligent environment, tourists are surrounded by intelligent interfaces supported by computing and networking technologies that are embedded in everyday objects such as clothes, vehicles, and smart materials.

FUTURE TRENDS

In the near future, the mode of users' interaction will become laid-back (relaxed and enjoyable) in e-tourism. Users will enjoy computer interaction for travel planning, and technology will move to the background. In addition, users will become an integral part of tourism product creation. Therefore, personalized services and complex market mechanisms should be deployed and provided to the users. For that reason, researchers must consider nontechnical issues related to markets and users, such as dynamic market and network structures; pricing and market design; design and experimenting business models; user decision modeling; and usage analysis.

CONCLUSION

E-tourism is a decent area for Web and semantic Web technologies by assisting users and agencies with quick information searching, integrating, recommending, and various intelligent services. Semantic Web technology has an enormous potential for e-tourism by providing: (1) integration and interoperability, (2) personalized and context-aware recommendations, (3) semantically enriched information searching, and (4) internationalization. Staab and Werthner (2002) state that the requirements of intelligent tourism ISs will raise a number of important technical research issues such as (1) semantic interoperability and mediated architectures; (2) e-business frameworks supporting processes across organizations; (3) mobility and embedded intelligence; (4) natural multilingual interfaces and novel interface technologies; (5) personalization and context-based tourism services; (6) information-to-knowledge transformations, data mining and knowledge management. To these directions, much work has to be done.

REFERENCES

Buhalis, D., & O'Connor, P. (2005). Information communication technology—Revolutionising tourism. *Tourism Recreation Research, 30*(3), 7–16.

Camacho, D., Borrajo, D., & Molina, J. (2001). Intelligent travel planning: A multiagent planning system to solve Web problems in the e-tourism domain. *Autonomous Agents and Mmultiagent Systems, 4*(4), 387-392.

Cardoso, J., & Lange, C. (2007). A framework for assessing strategies and technologies for dynamic packaging applications in e-Tourism. *Information Technology & Tourism, 9*(1), 27-44.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *Intelligent Systems and Their Applications, 14*(1), 20–26.

Dell' Erba, M. (2004). Exploiting semantic Web technologies for data interoperability. *AIS SIGSEMIS Bulletin, 1*(3), 48-52.

Dogac, A., Kabak, Y., Laleci, G., Sinir, S., Yildiz, A., Kirbas, S., & Gurcan, Y. (2004). Semantically enriched Web services for the travel industry. *ACM Sigmod Record, 33*(3), 21-27.

Fodor, O., & Werthner, H. (2005). Harmonise – a step towards an interoperable e-tourism marketplace. *International Journal on Electronic Business, 9*(2), 11-39.

Kanellopoulos, D. (2006). The advent of semantic Web in tourism information systems. *Tourismos: An International Multidisciplinary Journal of Tourism, 1*(2), 75-91.

Kanellopoulos, D. (2008). An ontology-based system for intelligent matching of travelers' needs for group package tours. *International Journal of Digital Culture and Electronic Tourism, 1*(2).

Kanellopoulos, D., & Kotsiantis, S. (2006). Towards intelligent wireless Web services for tourism. *International Journal of Computer Science and Network Security, 6*(7), 83-90.

Kanellopoulos D., & Kotsiantis S. (2007). A semantic-based architecture for intelligent destination management systems. *International Journal of Soft Computing, 2*(1), 61-68.

Kanellopoulos, D., & Panagopoulos A. (2008). Exploiting tourism destinations' knowledge in an RDF-based P2P network. *Network and Computer Applications, 31*,179-200.

Manes, G. (2003). The tetherless tourist: Ambient intelligence in travel & tourism. *Information Technology & Tourism, 5*(4), 211-220.

O'Brien, P., & Burmeister, J. (2003). Ubiquitous travel service delivery. *Information Technology & Tourism, 5*(4), 221-233.

Ouzzani, M., & Bouguettaya, A. (2004). Efficient access to Web services. *IEEE Internet Computing, 8*(2), 34-44.

Paprzychi, M., Gilbert, A., & Gordon, M. (2002). Knowledge representation in the agent-based travel support system.

Lecture Notes on Computer Science, 2457, 232-241. Berlin: Springer.

Ricci, F. (2002). Travel recommender systems. *IEEE Intelligent Systems, 17*(6), 55-57.

Schmidt-Belz, B., Polsad, S., Nick, A., & Zipf, A. (2002). Personalized and location-based mobile tourism services. In *Workshop on Mobile Tourism Support Systems, in conjunction with Mobile HCI 2002*. Pisa.

Staab, S., & Werthner, H. (2002). Intelligent systems for tourism. *IEEE Intelligent Systems, 17*(6), 53-55.

Van Setten, M., Pokraev, S., & Koolwaaij, J. (2004). Context-aware recommendations in the mobile tourist application COMPASS. In W. Nejdl & P. De Bra (Eds.), *Adaptive Hypermedia 2004*, LNCS 3137. Springer-Verlag.

Zhong, N. (2003). Toward Web intelligence. *Lecture Notes in Artificial Intelligence (LNAI), 2663*, 1–14. Springer, Berlin.

KEY TERMS

CRS (Computerized Reservation System): A CRS enables travel agencies to find what a customer is looking for and makes customer data storage and retrieval relatively simple.

DMO (Destination Management Organization): It is an entity or a company that promotes a tourist destination such as to increase the amount of visitors to this destination. It uses a DMS to distribute its properties and to present the tourist destination as a holistic entity.

DMS (Destination Management System): A DMS provides complete and up-to-date information on a particular tourist destination. It handles both the pre-trip and post-arrival information, as well as integrates availability and booking service too. It is used for the collection, storage, manipulation, and distribution of tourism information, as well as for the transaction of reservations and other commercial activities. Well-known DMSs are TISCover, VisitScotland, and Gulliver.

GDS (Global Distribution System): A GDS provides travel information services, such as real-time availability and price information for flights, hotels, and car rental companies. Dominant GDSs are Sabre and Galileo.

IFITT (International Federation for IT and Travel & Tourism): The IFITT (http://www.ifitt.org/) is a not-for-profit organization aiming to promote international discussion about ICTs and tourism.

LA_DMS (Layered Adaptive Semantic-Based DMS Based on P2P Technologies): It is a DMS that is adaptive to tourists' needs for tourist destination information. It uses a metadata model to encode semantic destination information in an RDF-based P2P network architecture. Its metadata model combines ontological structures with information for tourism destinations and peers.

OTA (Open Travel Alliance): The OTA (http://www.opentravel.org/) is an organization that develops open data transmission specifications for the electronic exchange of business information for the travel industry, including, but not limited to the use of XML.

WTO (World Tourism Organization): The WTO (http://www.world-tourism.org/) is a global body concerned with the collection and collation of statistical information on international tourism. It represents public sector tourism bodies from most countries, and the publication of its data makes possible comparisons of the flow and growth of tourism on a global scale.

Interactive Television Context and Advertising Recall

Verolien Cauberghe
University of Antwerp, Belgium

Patrick De Pelsmacker
University of Antwerp, Belgium

INTRODUCTION

Interactive digital television (IDTV), the merging of the Internet and television, has the potential of reaching many consumers. Introducing interactivity in television content will replace lean-backward viewing with a more active lean-forward viewing (Van den Broeck, Pierson, & Pauwels, 2004). This new way of watching TV can have implications for the way people process the advertisements embedded in programmes. We examine the impact of two dimensions of interactivity induced by a TV quiz show, that is, user control and two-way communication (McMillan & Hwang, 2002) on the ad and brand recall of an embedded commercial. User control means the possibility of accessing extra information about the quiz show, the host, and the candidates with the remote control. Two-way communication allows the viewer to play along with the quiz using the remote control.

BACKGROUND

Advertising Context Effects

The impact of responses to programme context (e.g., mood, excitement, involvement) on embedded advertisements have been studied extensively (e.g., De Pelsmacker, Geuens, & Anckaert, 2002). This context can have either a stimulating or an inhibiting effect on the processing of an embedded advertisement. Mental processes evoked by the programme have an influence on the processing of the advertisement embedded in the programme. Positive or congruent context effects are caused by the "carry-over" principle of the programme-induced attention, liking, interest, or arousal toward the advertisement that follows (Moorman, Neijens, & Smit, 2005). Negative or contrasting relationships have been explained by the cognitive absorption of the programme, leaving less cognitive abilities for processing the advertisement (Lang, 2000).

Interactivity and Information Processing

Interactivity can be defined in different ways. Wu (2005) distinguishes between actual or feature-based interactivity and perceived interactivity. McMillan and Hwang (2002) define three underlying dimensions of interactivity: two-way communication, user control, and time delay.

A considerable amount of empirical studies have investigated the effects of the interactivity of a message vehicle on an individual's cognitive information processing in an Internet context. Some studies found a positive impact of interactivity on memory (Chung & Xinshu, 2004; Macias, 2003), while others found no impact or even a negative impact (Bezijan-Avery, Calder, & Iacobucci, 1998). The cognitive load theory (CLT; Van Merriënboer & Sweller, 2005) can be used to explain both. The CLT assumes that the human working memory is limited in processing novel information. There are broadly two types of cognitive load that can affect the working memory (Van Merriënboer & Sweller): intrinsic cognitive load, which is related to the intrinsic nature of the information (in this study the questions and answers of the quiz programme, and the content of the additional programme information), and extraneous cognitive load, which corresponds to the mental effort imposed by the way the information is presented (for instance, in this study, programme interactivity). According to the elaboration likelihood model (Petty & Cacioppo, 1981), extensive information processing only occurs when the consumer is motivated to process the information. When this prerequisite is accomplished, he or she must also have the ability (cognitive capacity) to do so. The amount of interactivity can have an influence on both mechanisms. Interactivity can increase involvement with the content (Fortin & Dholakia, 2005), which increases the motivation to process the information. However, following the CLT, interactivity can also increase the total cognitive load, and thus diminishes the ability to process information. Therefore, depending on the strength of the intrinsic cognitive load, the individual will or will not have enough ability to process the information (the interac-

tive programme), which will further influence the processing of the embedded advertisement.

In this study, we investigate the context effects of two dimensions of programme interactivity representing a low level of intrinsic load (user control regarding information about the programme, candidates, host) and a high level of intrinsic load (two-way communication involving the questions and answers in the quiz) on ad and brand recall.

User Control

User control is "the range of ways to manipulate the content" (Coyle & Thorson, 2002) and refers to the amount of possible interactions the user has to get the information in the order and pace he prefers (in this study the amount of hyperlinks in the additional programme information). Different levels of this user-control dimension could influence the motivation to process the programme. Because the intrinsic load (extra programme information) is low, we do not expect that the additional load imposed by the interactivity (user control) will lead to limited information processing capacity. Although the respondents have the ability to process the programme and the embedded advertisement when the programme has no user control, the involvement with the programme and thus the motivation to process it will be relatively low. This low processing motivation is expected to be transferred to the advertisement, leading to a superficial processing of the advertisement. At a moderate level of user control and no two way, the motivation to process the information and programme involvement increase, thus facilitating the processing of the programme. This attentive state is expected to be transferred to the subsequent advertisement. A high level of user control will increase the motivation to process the programme but, given the relatively low intrinsic cognitive load of the user control process, this motivation to process information may be higher than is required. This may lead to the development of negative thoughts, which may inhibit advertising processing. Also, more clicks lead to less information per click. This decrease in information complexity may also lead to the development of feelings of boredom and irritation. We expect the following.

H1: A moderate level of user control will lead to a higher ad and brand recall than a low or a high level of user control.

Two-Way Communication

This dimension of interactivity can be characterized as a mutual discourse or the capability of providing feedback (Ha & James, 1998). In this study, two-way communication is manipulated through the possibility of playing along with a quiz show. This implies that the interactivity leads to a high intrinsic load as a result of answering the quiz questions (multiple choices) on screen using the remote control. Although the individual may have a high motivation to process the programme when playing along, he or she may lack the ability to do so given the linear time flow of the programme, which demands the working memory to process the information very fast. This limited cognitive capacity to process the programme may lead to a cognitive capacity problem when the individual is exposed to the embedded advertisement. We expect the following.

H2: Programme embedded two-way communication (playing along) will lead to a lower ad and brand recall than no two-way communication (not playing along).

Interaction Effect between User Control and Two-Way Communication

It is unclear what the combined effect of two-way communication and user control in the program context will be on the recall of the embedded ad and brand. On the one hand, a combination of playing along and the availability of more user control could result in a higher cognitive load and less recall of the embedded ad. Following this argument, a combination of low user control and no two-way communication should lead to the highest recall, and high user control combined with two-way communication (playing along) to the lowest. On the other hand, when the viewer cannot play along with the quiz and has no user control, his or her motivation to process the programme will be low, and consequently also the motivation to process the ad. Earlier we also hypothesized that a moderate level of user control leads to the most optimal cognitive activation state to process the programme. This activation state will be transferred to the ad when the consumer has the ability to process the ad in depth. When the consumer does not play along with the game, he or she will have sufficient cognitive resources to process the programme, and thus we could expect that this combination of a moderate level of user control and two-way communication will lead to a positive effect on ad and brand recall. Given that playing along with the quiz absorbs a lot of cognitive resources, the additional cognitive load induced by the moderate level of user control, might lead to a limited capacity problem. Since the combined effect of user control and two-way communication is not clear, we formulate the following open research question.

Q1: What is the interaction effect of user control and two-way communication on ad and brand recall?

EMPIRICAL STUDY

Research Method

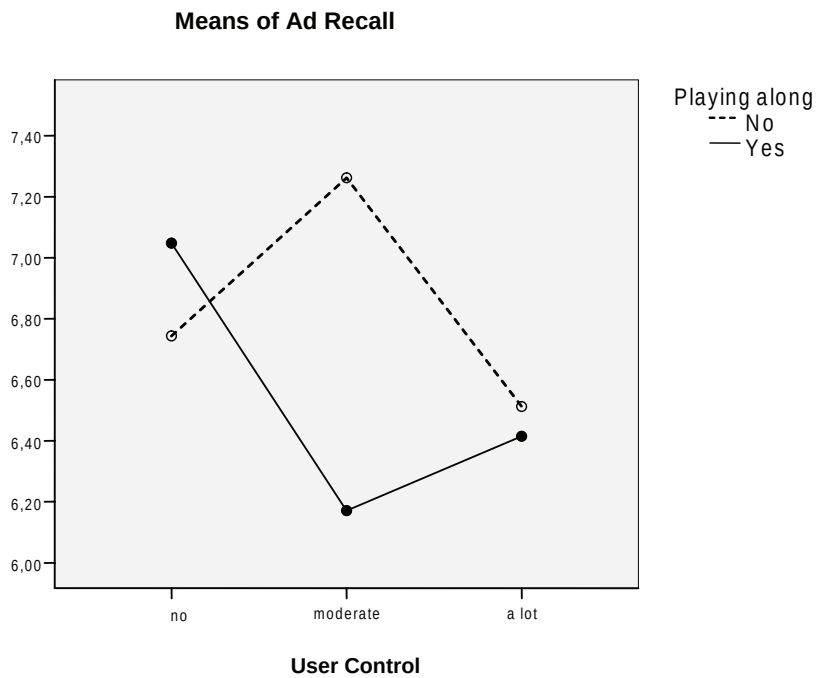
The hypotheses and research question were tested by means of a 3x2 (level of user control x level of two-way communication) between-participant experimental design. The programme context was an old episode of a well-known Belgian quiz show. User control was manipulated in the programme through the amount of clicks in the additional transparent information overlays that appeared automatically on screen during the sequence of the programme and that could be accessed by means of the remote control. Additional information about the host, candidates, quiz rules, and prizes was provided. The first level of user control did not enable any interactivity. Instead, the viewer got the extra programme information on paper. At the second level, a moderate amount of clicks (five clicks) was available during two moments in the programme. In the highest user-control condition, 26 clicks during four moments in the programme were available. The amount of information was kept constant across conditions. Two-way communication was varied on two levels. In the playing-along condition, for each question in the quiz, the viewer had the opportunity to make a choice out of three multiple choice answers using the remote control. In the not-playing condition, no interactivity was made possible. The advertisement and brand embedded in the program were unknown in Belgium at the time of the study to avoid confounding effects of existing knowledge and/or experience. We used a toothpaste advertisement from The Netherlands.

A sampling frame of 521 Flemish individuals was randomly selected by an Internet research company, taking age, gender, and education quota into account to be representative of the Belgian population. A net sample of 246 persons participated in the study. The average age of the participants was 38 years (range 21-56 years, 50% between 21-40 years old), 61.8% were males, 44.7% held a high school diploma, and 55.3% finished a higher education level.

The participants were individually invited to an experimental living-room setting. The participants were told that they were participating in a test viewing of an IDTV application. They were randomly assigned to one of the six experimental conditions. The participants viewed 15 minutes of the TV quiz show with an advertising break in the middle of the programme, after which they entered a computer-assisted questionnaire containing, amongst others, the recall measures. The experiment lasted 30 minutes in total. Each respondent received an incentive of €25

Advertisement recall was measured using 10 multiple-choice statements about content aspects of the advertisement (true vs. false). The scores were calculated by counting the

Figure 1. Interaction effect of user control and two-way communication on advertisement recall



number of correct answers (score between 0 and 10). Unaided brand recall was measured at two moments: immediately after the experiment and 10 days after the experiment. Of the 246 respondents, 168 cooperated in the delayed brand recall test (response rate of 68%).

Results

Advertising recall was measured as a score on a 10-point scale, and thus it was a ratio-scaled variable. The data met the conditions for parametric analysis. The ANOVA (analysis of variance) results indicate that user control ($p=.046$) has a significant main effect on ad recall. However, unexpectedly, the significance of this factor was driven by the difference between no user control ($M=6.90$) and high user control ($M=6.46$; $t=2.661$; $p=.009$), and not between the moderate level ($M=6.71$) vs. the high ($t=1.390$; $p=.167$) and low level ($t=.957$; $p=.341$), as expected. User control had no significant effect on brand recall measured immediately after the experiment (no=34.6%, moderate=36.1%, high=32.9%; $Chi^2=.189$; $p=.910$) or on delayed brand recall (no=28.8%, moderate=25.0%, high=27.7%; $Chi^2=.971$; $p=.896$). H1 is not supported.

Two-way communication also has a significant negative effect on ad recall ($p=.038$). Not playing along with the quiz resulted in a higher ad recall ($M=6.84$) than playing along ($M=6.55$, $t=2.011$; $p=.045$). As expected, two-way communication also had a significant negative effect on brand recall measured immediately after the experiment (40.2% vs. 29%; $Chi^2=3.370$; $p=.066$). This difference became even more significant after a delay of 10 days (36.7% vs. 18.1%; $Chi^2=7.992$; $p=.008$). H2 is accepted.

In Figure 1, the significant interaction effect between user control and two-way communication ($p<.001$) is shown. A moderate level of user control results in the highest ad recall rate when there is no play-along possibility ($M=7.26$) than when the consumer could play along. The lowest ad recall rate across all conditions results from the combination of a moderate level of user control and playing along ($M= 6.17$). This difference was significant ($t= 4.085$; $p<.001$). When there was no user control embedded in the programme, playing along had no effect on ad recall ($M= 6.74$ vs. $M= 7.05$; $t= -1.315$; $p=.192$). Also, when the level of user control was high, there appeared to be no difference in ad recall as a result of playing along or not ($M=6.51$ vs. $M=6.42$; $t= .416$; $p=.679$). The same interaction effect of user control and two-way communication was noticeable on immediate and delayed brand recall.

FUTURE TRENDS

Interactivity is assumed to be the major difference between new and traditional media. With IDTV, the viewer or user can interact with the TV content using the remote control, providing him or her with more control over the viewing experience (Van den Broeck, 2004). Requesting additional information about actors in a TV soap, participating in TV talk shows through on-screen chatting, and requesting a coupon in an advertisement are a few of the new possibilities made possible by this medium. This hypermultimedia content changes the way people watch TV, and thus can have implications on the way advertisements are perceived and processed.

The results of this study have implications for broadcasters, media planners, and advertisers. For broadcasters, the IDTV technology may increase the effectiveness of advertising in terms of ad and brand recall compared to traditional TV. The challenge for broadcasters is to reach enough viewers for their live interactive programming. Once this is achieved, our results demonstrate that IDTV compared to traditional TV can increase the brand recall up to 5.6% (difference between brand recall in the condition of no user control and no play-along possibility, and the condition with a moderate level of user control with no play-along possibility; 41% vs. 46.6%). For media planners and advertisers, the results may be of interest when the advertising goal of the campaign is to increase the awareness of the ad message or the brand. Placing an advertisement in an interactive programme might enhance ad and brand recall if the programme has a moderate level of interactivity. Too little programme interactivity might lead to boredom and to negative advertising results, whereas too much interactivity might load the respondent's cognitive abilities so much that he or she will no longer pay attention to the advertisements that appear during the commercial break. To achieve the most optimal recall results, programme interactivity should increase the motivation of the viewers to watch the content more in depth without overloading them. Therefore, a moderate level of user control without two-way communication would lead to the best results.

Further research should investigate the effects of less-cognitive-demanding two-way communication applications than the one used in this study. Perhaps imposing fewer questions on the viewers and giving them more time to respond might lead to this optimal IDTV experience for broadcasters and advertisers. In this study, we investigated memory-related aspects of processing the advertisement. Investigating the impact of programme interactivity on ad and brand attitudes would also be interesting. Further research is also warranted to study more in depth the effects of interactivity on memory and attitudes after a time delay. The results of the current study should be tested in a real-time (live) setting to corroborate the findings of this experimental context. Finally, research in larger samples is needed to reliably assess the differences in effects between different groups of viewers (e.g., by gender, age, education, etc.).

CONCLUSION

Both programme-induced user control and two-way communication had an influence on recall. However, the effect of user control was less outspoken than that of two-way communication. User control did not have a significant main effect on brand recall, and only had a weak negative effect on ad recall. Two-way communication (playing along with the TV quiz show) did have a strong and significantly negative impact on both immediate and delayed brand recall and on ad recall. The high motivation to process the quiz when playing along in combination with the high intrinsic load of the questions and answers apparently left little cognitive resources to process the embedded advertisement. The highest recall is obtained as a result of a moderate level of user control and playing along. In this condition apparently the viewer has both the ability and the motivation to process the programme, a processing state which is transferred to the embedded ad.

Besides the difference in cognitive intrinsic load, the fact that two-way communication is more likely to imply a limited-capacity problem compared to user control may also be explained by the fact that user control is a more voluntary interaction in terms of the time spent in the interactive information and the order in which viewers access the information. On the contrary, when playing along, the viewer has to follow the sequence and the order of the questions presented in the quiz programme. Past studies also found that time pressure increases the likelihood for cognitive overload to occur.

REFERENCES

Bezjian-Avery, A., Calder, B., & Iacobucci, D. (1998). New media interactive versus traditional advertising. *Journal of Advertising Research, 38*(4), 23-32.

Chung, H., & Xinshu, Z. (2004). Effects of perceived interactivity on Web site preference and memory: Role of personal motivation. *Journal of Computer-Mediated Communication, 10*(1). Retrieved from http://jcmc.indiana.edu/

Coyle, J. R., & Thorson, E. (2002). The effects of progressive levels of interactivity and vividness in Web marketing sites. *Journal of Advertising*, *30*(3), 65-77.

De Pelsmacker, P., Geuens, M., & Anckaert, P. (2002). Media context and advertising effectiveness: The role of context appreciation and context/ad similarity. *Journal of Advertising, 31*(2), 49-61.

Fortin, D. R., & Dholakia, R. R. (2005). Interactivity and vividness effects on social presence and involvement with Website-based advertisement. *Journal of Business Research, 58*(3), 387-396.

Ha, L., & James, L. (1998). Interactivity reexamined: A baseline analysis of early business Web sites. *Journal of Broadcasting and Electronic Media, 42*(4), 457-474.

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication, 50*(3), 46-67.

Macias, W. (2003). A beginning look at the effects of interactivity, product involvement and Web experience on comprehension: Brand Websites as interactive advertising. *Journal of Current Issues and Research in Advertising, 25*(2), 31-44.

McMillan, S. J., & Hwang, J. (2002). Measures of perceived interactivity: An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *Journal of Advertising, 31*(3), 29-42.

Moorman, M., Neijens, P. C., & Smit, E. G. (2005). The effects of program responses on the processing of commercials placed at various positions in the program and the block. *Journal of Advertising Research*, pp. 49-59.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitudes & persuasion: Classic & contemporary approaches.* Dubuque, IA: William C. Brown.

Van den Broeck, W., Pierson, J., & Pauwels, C. (2004). *Does interactive TV imply new uses?* Paper presented at the IDTV Conference, Brighton, United Kingdom.

Van Merriënboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17*(2), 147-177.

Wu, G. (2005). Mediating role of perceived interactivity in the effect of actual interactivity on attitude toward the Website. *Journal of Interactive Advertising, 5*(2). Retrieved from http://www.jiad.com

KEY TERMS

Carry-Over Advertising Context Effect: It is when psychological reactions evoked by a programme do not immediately disappear when the programme is interrupted by a commercial break, but have an influence on the processing of the advertisement embedded in the programme

Cognitive Overload: This occurs when the volume of information supply exceeds the information processing capacity of the individual.

Extrinsic Load: It is the cognitive load that is related to the representation of the information (form, style, etc.).

Interactive Digital Television (IDTV): IDTV is the merging of the Internet and television.

Intrinsic Load: It is the cognitive load that is related to the information content itself.

Two-Way Communication: It refers to mutual discourse, the capability of providing feedback, or the exchange of roles.

User Control: It is the range of ways to manipulate the content.

Interface Design Issues for Mobile Commerce

Susy S. Chan
DePaul University, USA

Xiaowen Fang
DePaul University, USA

INTRODUCTION

Effective interface design for mobile handheld devices facilitates user adoption of mobile commerce (m-commerce). Current wireless technology poses many constraints for effective interface design. These constraints include limited connectivity and bandwidth, diverse yet simplistic devices, the dominance of proprietary tools and languages, and the absence of common standards for application development.

The convergence of mobile Internet and wireless communications has not yet resulted in major growth in mobile commerce. Consumer adoption of m-commerce has been slow even in countries such as Finland, which have broadly adopted wireless technology (Anckar & D'Incau, 2002). An international study of mobile handheld devices and services suggests that mobile commerce is at a crossroads (Jarvenpaa, Lang, Takeda & Tuunainen, 2003). The enterprise and business use of wireless technology holds greater promise, but it demands the transformation of business processes and infrastructure. Poor usability of mobile Internet sites and wireless applications for commerce activities stands out as a major obstacle for the adoption of mobile solutions. For example, even with the latest 3G phones in Japan, consumers still find the small screen display and small buttons on these devices difficult to use (Belson, 2002).

BACKGROUND

Mobile Commerce

Mobile commerce broadly refers to the use of wireless technology, particularly handheld mobile devices and mobile Internet, to facilitate transaction, information search, and user task performance in business-to-consumer, business-to-business, and intra-enterprise communications (Chan & Fang, 2003). Researchers have proposed several frameworks for the study of m-commerce. Varshney and Vetter's framework (2001) presents 12 classes of m-commerce applications, ranging from retail and online shopping, auction, mobile office, and entertainment to mobile inventory emphasizing the potential of mobile B2B and intra-enterprise applications. The framework by Kannan, Chang, and Whinston (2001) groups mobile services into goods, services, content for consumer e-commerce, and activities among trading partners.

Waters (2000) proposes two visions for the potential and opportunities of m-commerce. One perspective argues that the mobile, wireless channel should be viewed as an extension of the current e-commerce channel or as part of a company's multi-channel strategies for reaching customers, employees, and partners. The second, more radical view suggests that m-commerce can create markets and business models.

Recent development in m-commerce has substantiated the first perspective. Major e-commerce sites have implemented their mobile Internet sites as an extension of wired e-commerce to support existing customers (Chan & Lam, 2004; Chan et al., 2002). Consumers have shown relatively low willingness to use m-commerce, but adopters of e-commerce are more likely to embrace this new technology (Anckar & D'Incau, 2002). Furthermore, perceived difficulty of use can affect consumers' choice of m-commerce as a distribution channel (Shim, Bekkering & Hall, 2002). These findings suggest that in a multi-channel environment, m-commerce *supplements* e-commerce instead of becoming a *substitute* for e-commerce.

Enterprise and business applications of m-commerce technologies seem to hold greater promise, because it is easier for companies to standardize and customize applications and devices to enhance current work processes. An Ernst & Young study (2001) of the largest companies in Sweden shows that, except for the retail industry sector, most industries have viewed m-commerce as being vital for growth and efficiency strategies, but not necessarily for generating new revenue. However, integrating the wireless platform in an enterprise requires significant structural transformation and process redesign.

Research on Wireless Interface Design

Several recent studies have examined interface design for mobile applications using handheld devices. Researchers have found that direct access methods were more effective for retrieval tasks with small displays (Jones, Marsden, Mohd-Nasir, Boone & Buchanan, 1999). Novice WAP phone users perform better when using links instead of action screens for navigation among cards, and when using lists of links instead of selection screens for single-choice lists (Chittaro & Cin,

2001). Ramsay and Nielsen (2000) note that many WAP usability problems echo issues identified during the early stage of Web site development for desktop computers, and could be alleviated by applying good user interface design. Such design guidelines for WAP applications include: (1) short links and direct access to content, (2) backward navigation on every card, (3) minimal level of menu hierarchy, (4) reduced vertical scrolling, (5) reduced keystrokes, and (6) headlines for each card (Colafigi, Inverard & Martriccian, 2001; Buchanan et al., 2001). Buyukkokten, Garcia-Molina, and Paepcke (2001) have found that a combination of keyword and summary was the best method for Web browsing on PDA-like handheld devices.

Diverse form factors have different interface requirements. The study by Chan et al. (2002) of 10 wireless Web sites across multiple form factors reveals that user tasks for the wireless sites were designed with steps similar to the wired e-commerce sites, and were primarily geared towards experienced users. Many usability problems, such as long download and broken connections, information overload, and excessive horizontal and vertical scrolling, are common to three form factors—WAP phone, wireless PDA, and Pocket PC. Interface design flaws are platform independent, but the more limitations imposed on the form factors, the more acute the design problems become.

Mobile users access information from different sources and often experience a wide range of network connectivity. Context factors have a particular impact on the usability of mobile applications. Based on a usability study conducted in Korea, three use context factors—hand (one or two hands), leg (walking or stopping), and co-location (alone or with others)—may result in different usability problems (Kim, Kim, Lee, Chae & Choi, 2002). Therefore, the user interface design has to consider various use contexts. Researchers also suggest a systems-level usability approach to incorporating hardware, software, "netware," and "bizware" in the design of user-friendly wireless applications (Palen & Salzman, 2002). Perry, O'Hara, Sellen, Brown, and Harper (2001) have identified four factors in "anytime anywhere" information access for mobile work: the role of planning, working in "dead time," accessing remote technological and informational resources, and monitoring the activities of remote colleagues.

Multimodal interfaces are gaining importance. The MobileGuiding project developed in Spain is aimed at building a European interactive guide network on a common, multimodal, and multilingual platform in which contributors will provide leisure information and cultural events in their locations (Aliprandi et al., 2003). Furthermore, there has been a study conducted in Finland that addresses the design and evaluation of a speech-operated calendar application in a mobile usage context (Ronkainen, Kela & Marila, 2003).

MAIN THRUST OF THIS ARTICLE

Five issues are essential to the interface design for mobile commerce applications, including: (a) technology issues, (b) user goals and tasks, (c) content preparation, (d) application development, and (e) the relationship between m- and e-commerce.

Technology Issues

Limitation of Bandwidth

Most mobile communication standards only support data rates that are less than 28.8 kbps. Connections to the wireless service base stations are unstable because signal strength changes from place to place, especially on the move. These constraints limit the amount of information exchanged between device and base station. Indication of the download progress and friendly recovery from broken connections are necessary to help users gain a better sense of control.

Form Factor

Mobile commerce services are accessible through four common platforms: wireless PDA devices using Palm OS, Pocket PCs running Microsoft Windows CE/Pocket PC OS, WAP phone, and two-way pagers. Within the same platform, different form factors may offer different functionalities. A developer should consider the form factor's unique characteristics when developing m-commerce applications.

User Goals and Tasks

Mobile users can spare only limited time and cognitive resources in performing a task. Services that emphasize mobile values, and time-critical and spontaneous needs, add more value for m-commerce users. These mobile services may include the ability to check flight schedules, check stock prices, and submit bids for auction (Anckar & D'Incau, 2002). In addition, mobile tasks that demonstrate a high level of perceived usefulness, playfulness, and security are the ones most likely to be adopted by users (Fang, Chan, Brzezinski & Xu, 2003).

Content Preparation

Constraints in bandwidth and small screen size demand different design guidelines. Most design guidelines for e-commerce (e.g., Nielsen, Farrell, Snyder & Molich, 2000) support the development of rich product information sets and a complete shopping process. In contrast, wireless Web sites have to simplify their content presentation.

Amount of Information

Content adaptation is necessary to convert information for the mobile Web (Zhou & Chan, 2003). However, users should have sufficient, if not rich, information to accomplish the goals for the application.

Navigation

Navigation systems vary from one form factor to another because the design of handheld devices differs. Currently, there is no consensus on which functions or features should be provided by the application, or built into the device itself.

Depth of Site Structure

Since mobile users have limited time for browsing wireless applications, the organization of information is critical. A flatter structure with fewer steps for wireless applications would allow users to review more options in the same step, and to locate the desired information more quickly.

Graphics or Text

Text is a better choice for displaying information on small screen browsers. However, better technology may improve the screen quality of handheld devices to display more complicated graphics. When determining the format of information to present, it is important to consider the form factor, because it may pose additional constraints on the format.

Development Environment

Mobile computing alters the assumption of "fixed" context of use for interface design and usability testing (Johnson, 1998). Traditional means of user interviews or usability testing in a laboratory environment cannot reveal insights into users' activities and mobility in real life. Contextual consideration is critical for gathering information about user requirements. For example, when developing and testing a mobile application for grocery shoppers, user requirement gathering and prototype evaluation should be conducted in a grocery store (Newcomb, Pashley & Stasko, 2003). The method of contextual inquiry can augment user interface design by exploring the versatility of usage patterns and usage context (Väänänen-Vainio-Mattila & Ruuska, 1998). While contextual inquiry may help developers gain a realistic understanding of contextual factors affecting user behaviors in motion, it is difficult to conduct non-obtrusive observations and inquiries. Developers for mobile applications need to consider the application context surrounding the relationship between the mobile device and user goals and tasks.

Relationship Between M-Commerce and E-Commerce

The wireless channel for e-commerce has raised many new questions regarding coordination between interactions with users across multiple channels. Some researchers suggest that because of the "transaction aware" and "location aware" characteristics of the wireless technology, mobile consumers may increase impulse purchases, especially in low-value, low-involvement product categories, such as books and CDs (Kannan et al., 2001). At present, many Web sites have extended the wireless channel to leverage relationships with exiting customers (Chan et al., 2002). The current state of technology and poor usability of mobile Web sites makes it difficult to expand m-commerce as an independent channel. Many analysts believe that the wireless channel is promising for customer relationship management (CRM) because of its ability to: (1) personalize content and services; (2) track consumers or users across media and over time; (3) provide content and service at the point of need; and (4) provide content with highly engaging characteristics (Kannan et al., 2001). The challenge is how to coordinate interface and content across multiple channels so that experienced users and repeat customers can handle multiple media and platforms with satisfaction.

FUTURE TRENDS

Technology Trends

User interface design for mobile commerce will likely be influenced by four trends. First, multiple standards for wireless communication will not be resolved quickly, especially in North America. Second, the high cost of third-generation (3G) technology may delay the availability of broadband technology for complex functionality and content distribution for mobile applications. Third, instead of the convergence of functionalities into a universal mobile handheld device, there may be a variety of communication devices operating in harmony to support users in their everyday lives. Fourth, input and output format may expand to incorporate voice and other formats, as well as expandable keyboards. The introduction of the voice-based interfaces may complement the text-based interface and remedy some of the information input/display problems of the handheld devices. These trends suggest opportunities to conceptualize wireless user interface beyond text-based interaction. The new challenges are to design better multimodal interfaces for inter-device communication in order to simplify tasks for mobile users.

Development Trends

Alternative methods for interface design and evaluation will be necessary to support m-commerce applications development. First, requirement analysis should focus on the context of mobile users' behaviors and tasks. Contextual inquiry and other methods may be developed to facilitate the understanding of interactions between mobility and usability. Second, usability testing should be conducted with an understanding of contextual variables beyond user behavior. Third, mapping form factors, user tasks, data needs, and content across multiple channels and platforms is necessary to synchronize content and coordinate functionality in a distributed system. Fourth, user-centered design guidelines for mobile applications will be important. These trends require a fresh look at current methodology and will help determine new ways of incorporating user interface design and usability testing for distributed wireless application development. The reference framework proposed by Lee and Benbasat (2003) may be useful in this regard. Their framework incorporates seven design elements for m-commerce interface: context, content, community, customization, communication, connection, and commerce.

M-Commerce Business Models

Wireless technology for m-commerce is likely to evolve in two areas. For intra-enterprise and business-to-business uses, wireless technology provides location-aware and mobility-aware solutions for mobile workers. There is a broad range of possibilities for B2B applications because such deployment can be controlled more easily. Content distribution may be integrated with the enterprise systems. Context-based applications, interfaces, functionality, and even devices can be customized according to the mobile tasks and user groups in the B2B context. This approach makes application development, deployment, and integration easier to manage. In contrast, it is far more challenging to manage the design, development, and deployment of wireless applications for customers. Wireless technology's capability for personalization seems to be the strongest argument for m-CRM services to enhance customer retention (Chan & Lam, 2004). A careful mapping of tasks, data, form factors, and the CRM process is essential for user interface design.

CONCLUSION

Wireless technology and the mobile Internet continues to evolve. Until the technology matures and bandwidth improves, wireless applications will be geared toward users requiring limited bandwidth, short exchange of data and text, and simple functionality. Two areas of wireless applications, CRM and enterprise efficiency, may reap greater success. Consumer e-commerce Web sites should focus on the selection of tasks that are most suitable for the wireless channel and demonstrate mobile values, especially for experienced users. Such mapping process requires a solid understanding of the CRM strategy, user preferences, and the constraints imposed by a mobile environment. For enterprise adoption, consolidating the wireless platforms and form factors will facilitate interface design. In either case, additional research to improve usability for mobile commerce is essential.

REFERENCES

Aliprandi, C., Athenour, M., Martinez, S.C., & Patsis, N. (2003). MobileGuiding: A European multimodal and multilingual system for ubiquitous access to leisure and cultural contents. In C. Stephanidis & J. Jacko (Eds.), *Proceedings of the 10th International Conference on Human-Computer Interaction* (vol. 2, pp. 3-7). Mahwah, NJ: Lawrence Erlbaum.

Anckar, B., & D'Incau, D. (2002). Value creation in mobile commerce: Findings from a consumer survey. *Journal of Information Technology Theory & Application, 4*(1), 43-64.

Belson, K. (2002, April 22). Japan is slow to accept the latest phones. *The New York Times*, C4.

Buchanan, G., Farrant, S., Jones, M., Thimbleby, H., Marsden, G., & Pazzani, M. (2001). Improving mobile Internet usability. *Proceedings of the 10th International World Wide Web Conference* (pp. 673-680). New York: ACM Press.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Seeing the whole in parts: Text summarization for Web browsing on handheld devices. *Proceedings of the 10th International World Wide Web Conference*. New York: ACM Press.

Chan, S., & Fang, X. (2003). Mobile commerce and usability. In K. Siau & E. Lim (Eds.), *Advances in mobile commerce technologies* (pp. 235-257). Hershey, PA: Idea Group Publishing.

Chan, S., Fang, X., Brzezinski, J., Zhou, Y., Xu, S., & Lam, J. (2002). Usability for mobile commerce across multiple form factors. *Journal of Electronic Commerce Research, 3*(3), 187-199.

Chan, S., & Lam, J. (2004). Customer relationship management on Internet and mobile channels: A framework and research direction. In C. Deans (Ed.), *E-commerce and m-commerce technologies*. Hershey, PA: Idea Group Publishing.

Chittaro, L., & Cin, P.D. (2001). Evaluating interface design choices on WAP phones: Single-choice list selection and navigation among cards. In M.D. Dunlop & S.A. Brewster (Eds.), *Proceedings of Mobile HCI 2001: Third International Workshop on Human Computer Interaction with Mobile Devices.*

Colafigli, C., Inverard, P., & Martriccian, R. (2001). Infoparco: An experience in designing an information system accessible through WEB and WAP interfaces. *Proceedings of the 34th Hawaii International Conference on System Science.* Los Alamitos, CA: IEEE Computer Society Press.

Ernst & Young. (2001). *Global online retailing: An Ernst & Young special report.* Gemini Ernst & Young.

Fang, X., Chan, S., Brzezinski, J., & Xu, S. (2003). A study of task characteristics and user intention to use handheld devices for mobile commerce. *Proceedings of the 2nd Annual Workshop on HCI Research in MIS* (pp. 90-94).

Jarvenpaa, S., Lang, K., Takeda, Y., & Tuunainen, V. (2003). Mobile commerce at crossroads. *Communications of the ACM, 46*(12), 41-44.

Johnson, P. (1998). Usability and mobility: Interactions on the move. In C. Johnson (Ed.), *Proceedings of the 1st Workshop on Human Computer Interaction with Mobile Devices.*

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., & Buchanan, G. (1999). Improving Web interaction on small displays. *Computer Networks: The International Journal of Distributed Informatique, 31,* 1129-1137.

Kannan, P., Chang, A., & Whinston, A. (2001). Wireless commerce: Marketing issues and possibilities. *Proceedings of the 34th Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press.

Kim, K., Kim, J., Lee, Y., Chae, M., & Choi, Y. (2002). An empirical study of the use contexts and usability problems in mobile Internet. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press.

Lee, Y. ,& Benbasat, I. (2003). Interface design for mobile commerce. *Communications of the ACM, 46*(12), 49-52.

Newcomb, E., Pashley, T., & Stasko, J. (2003). Mobile computing in the retail arena. *Proceedings of the Conference on Human Factors in Computing Systems, 5*(1), 337-344.

Nielsen, J., Farrell, S., Snyder, C., & Molich, R. (2000). *E-commerce user experience: Category pages.* Nielsen Norman Group.

Palen, L. & Salzman, M. (2002). Beyond the handset: Designing for wireless communications usability. *ACM Transactions on Computer-Human Interaction, 9*(2), 125-151.

Perry, M., O'Hara, K., Sellen, A., Brown, B., & Harper, R. (2001). Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction, 8*(4), 323-347.

Ramsey, M., & Nielsen, J. (2000). *WAP usability: Déjà vu: 1994 all over again.* Nielsen Norman Group.

Ronkainen, S., Kela, J., & Marila, J. (2003). Designing a speech operated calendar application for mobile users. In C. Stephanidis & J. Jacko (Eds.), *Proceedings of the 10th International Conference on Human-Computer Interaction* (vol. 2, pp. 258-262). Mahwah, NJ: Lawrence Erlbaum.

Shim, J.P., Bekkering, E., & Hall, L. (2002). Empirical findings on perceived value of mobile commerce as a distributed channel. *Proceedings of the 8th Americas Conference on Information Systems* (pp. 1835-1837).

Varshney, U., & Vetter, R. (2001). A framework for the emerging mobile commerce applications. *Proceedings of the 34th Hawaii International Conference on System Sciences.* Los Alamitos, CA: IEEE Computer Society Press.

Väänänen-Vainio-Mattila, K., & Ruuska, S. (1998). User needs for mobile communication devices: Requirements gathering and analysis through contextual inquiry. In C. Johnson (Ed.), *Proceedings of the 1st Workshop on Human Computer Interaction with Mobile Devices.*

Waters, R. (2000, March 1). Rival views emerge of wireless Internet. *Financial Times FT-IT Review,* 1.

Zhou, Y., & Chan, S. (2003). Adaptive content delivery over the mobile Web. *Proceedings of the 9th Americas Conference on Information Systems* (pp. 2009-2019).

KEY TERMS

Contextual Inquiry: This interface design method employs an ethnographic approach such as observing user activities in a realistic context.

Fixed Context of Use: Traditional user interface design and testing assumes a single domain, with the users always using the same computer to undertake tasks alone or in collaboration with others.

Form Factor: This platform or operating system runs on a handheld device. Major form factors include Palm, Pocket PC, and WAP.

Interface Design: Design of the interactions between humans and computers.

Location-Aware Service: Mobile services that provide information based on a user's location through the support

of a global positioning system. Such services include mobile maps, weather, restaurants, and movie directories.

M-CRM: Interactions between a company and its customers for marketing, sales, and support services through the mobile Web and wireless channel.

Multimodal Interface: An interface that communicates with users through multiple modes.

Usability: Usability refers to how well an application is designed for users to perform desired tasks easily and effectively.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1612-1617, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

International Digital Studies Approach for Examining International Online Interactions

Kirk St.Amant
Texas Tech University, USA

INTRODUCTION

As global access to the Internet increases, so does the potential for miscommunication in international online interactions (IOIs). Unfortunately, many models for examining cross-cultural communication focus on conventional (offline) interactions or settings. As a result, researchers lack a mechanism for examining how cultural factors could affect online discourse.

This article presents an approach—international digital studies—for examining how cultural factors could affect IOIs. The purpose of this approach is to identify points of contention or areas where online media can create conflicts in cultural expectations associated with credibility. Once identified, these points of contention can serve as the subject of future research related to culture and communication.

BACKGROUND

Creating credibility, or ethos, is not a random process. Rather, audiences use certain factors, or *ethos conditions*, to develop a checklist for determining if a presentation is credible (worthy of attention). That is, audiences come to a particular presentation situation thinking, "This individual must do x, y, and z if I am to consider him or her credible/worth listening to." If all of these expectations are met (can be "checked off"), then the presenter and his or her ideas will be considered credible. If one or more ethos conditions are not met, then audiences will be less likely to view a presenter as credible (see St.Amant, 2002a, for a more in-depth discussion of this concept).

The ethos conditions one expects to encounter can vary from culture to culture, and such differences have been noted at a variety of levels (Campbell, 1998; Tebeaux, 1999; Lewis, 2001). Persons from different cultures, for example, often use different organizational structures (e.g., stated vs. implied conclusions) and different methods of citing sources to establish the credibility of a presentation (Woolever, 2001; Lewis, 2001; Hofstede, 1997). Cultures can also associate different credibility expectations with sentence length. Southern Europeans, for example, associate longer sentences with credible presentations, while Americans view shorter and more direct sentences as being more credible (Ulijn & Strother, 1995). Additionally, the kind of relationship associated with the use of a particular word can cause cross-cultural credibility problems (Li & Koole, 1998; Li, 1999).

Online media complicate cross-cultural interactions by creating conditions that affect credibility expectations. In many cases, online media reduce human interaction to typed words. Typed online messages, however, tend to follow patterns related to spoken discourse. This mix of written and spoken communication creates a new and interesting situation, for recipients of online text messages do not obtain nonverbal identity cues key to communicating in spoken exchanges. The sender of an online message therefore seems faceless and anonymous (Gauntlett, 2000; St.Amant, 2002b).

As a result, notions of authority, identity, and credibility take on new forms in cyberspace. As Fernback (1999) notes, in online exchanges, the markers of credibility—marks that draw others to listen to you—are not, "brawn, money, or political clout," but are rather "wit, and tenacity, and intelligence" (p. 213). Thus, wit, tenacity, and intelligence become ethos conditions individuals can use to appear more authoritative or more credible than other participants in an online exchange. These factors therefore become *digital ethos conditions*, for individuals come to expect them when assessing the credibility of online presentations. These digital ethos conditions, however, can conflict with the communication expectations of different cultural groups.

Understanding how cultural factors can affect online exchanges can be a complicated and seemingly overwhelming process. Yet, now that more of the world is getting online, it is becoming increasingly important to understand IOI situations so that miscommunications and mistakes can be avoided. (Such culture-related mistakes, moreover, could affect everything from online social exchanges to international outsourcing and international e-commerce activities.) For this reason, researchers can benefit from an approach that helps them focus their analysis of IOIs on a more manageable set of topics. The international digital studies approach is designed to establish such a focus.

MAIN THRUST OF THE ARTICLE

International digital studies is a research approach used to examine how cultural groups differ in their responses to digital ethos conditions. According to this perspective, the objective of the researcher is two-fold:

- First, the researcher must identify *actual* digital ethos conditions—presentation factors that actually contribute to a presenter's credibility in online exchanges. Once isolated, these digital ethos conditions can become variables used to evaluate how different cultures communicate online.
- Second, the researcher must determine if a digital ethos condition is also a factor that varies in relation to cultural communication expectations. That is, researchers need to determine if cultures would differ observably in how they responded to a particular digital ethos condition.

The key to this line of research becomes identifying variables that could affect communication in IOIs. To achieve this objective, researchers must use a two-part literature review involving the fields of Internet studies and intercultural communication.

The purpose of the dual-field literature review is to determine how factors of medium and of culture might create conflicting expectations of ethos conditions in IOIs. To identify these situations, individuals must first survey the research literature in Internet studies in order to identify digital ethos conditions in direct, two-way interactions online. The focus of this review is to isolate behavior resulting from online communication conditions vs. the transfer of communication patterns from more traditional media to an online setting. Name-dropping, for example, can be used to create credibility in both print and online media; the use of emoticons to create credibility, however, is more restricted to online communication.

After researchers identify digital ethos conditions, they must determine if these factors could cause confusion or conflict in cross-cultural exchanges. The goal then becomes evaluating if a particular digital ethos condition is also a *point of contention*—or a situation in which the communication patterns documented in the literature of one field (Internet studies) conflict with patterns noted in the literature of another field (intercultural communication).

To identify points of contention, the second part of the international digital studies process involves a review of the research literature in intercultural communication. In this second review, the researcher would look specifically for indications that digital ethos conditions identified in the initial (Internet studies) literature review relate to findings reported in the intercultural communication literature. If little or no mention of this factor is made, or if this variable appears to cause no real conflict in cross-cultural exchanges, then that variable would be a *weak* point of contention. If, for example, different cultural groups did not react differently to uses of wit (a key ethos condition noted in Internet studies), then uses of wit would be a weak point of contention, for there is little evidence of different cultural behavior related to this digital ethos condition. If, however, the intercultural literature review reports that the ethos condition noted in the Internet studies literature can cause problems in cross-cultural interactions, then that factor would be a *strong* point of contention that could affect IOIs.

Researchers must next determine if a strong point of contention could actually cause problems in IOIs. That is, just because the two-part literature review indicates a particular ethos condition could be a point of contention.

- Does that ethos condition actually affect discourse in IOIs?
- Do reactions to that ethos condition vary along cultural lines in IOI? (e.g., Do some cultures use it more than others? Are some cultures more confused by its use than others?)
- Can researchers develop a ranking system to compare how specific cultural groups vary in relation to uses of and responses to a particular ethos condition?

To answer these questions, researchers must use strong points of contention as the foundation for experiments that test if and how a strong point of contention can affect IOIs. In this way, the international digital studies approach helps researchers identify suitable topics for conducting further research into IOIs.

FUTURE TRENDS

An application of international digital studies indicates that the concept of identity could be a key problem area in future IOIs. For this reason, it is important that researchers understand how aspects and perceptions of identity could cause problems in online exchanges involving individuals from different cultures.

A review of the Internet studies literature reveals identity is a factor that affects discourse in online forums. Many researchers note that, in cyberspace exchanges, identity is neither fixed nor stable; rather, it can easily change because of online media that reduce interactions to typing words (Gauntlett, 2000; St.Amant, 2002b). By reducing identity to texts, online media allow individuals to create their online identity on their own terms (Arnold & Plymire, 2000).

Other researchers note that by limiting identity to texts, online media allow other users to co-opt or to change someone else's online identity by cutting and pasting another person's words into a different message (Warnick, 1998). Cutting and

pasting parts of a message allows individuals to separate what was said from who said it (Warnick, 1998). The ability to separate authors from their original online message gives presenters a great deal of liberty, for they can:

- attribute segments of forwarded messages to whomever they wish, thus altering the online identity of someone else;
- create new identities for themselves by co-opting the words or ideas of another;
- alter what another said, creating a new online identity for that author by "putting words in that person's mouth."

In all three cases, altering textual factors allows one individual to change the online identity of another person. The question then becomes: Could this plasticity of online identity result in reactions that vary along cultural lines? To answer this question, the researcher must review the intercultural communication literature in search of information relating to identity.

A review of the intercultural communication literature reveals that a fixed or verifiable identity is often essential to creating credible messages (Weiss, 1998; Ferraro, 2002; Ng & Van Dyne, 2001). In some cultures, one's identity is not based on the claims or the proof that the individual himself or herself presents, but is rather based on the claims of others. For example, in certain cultures, people interact within relatively large and complex social networks. These networks are often formed from long-term relationships developed between individuals over time, or from strong familial ties based on trust and senses of family duty and family honor (Hofstede, 1997; Weiss, 1998; Richmond, 1995).

In social network cultures, the identity of the presenter often determines if others will listen to or ignore information presented by that individual (Hofstede, 1997; Weiss, 1998; Richmond, 1995). The distinction is essentially one of in-group vs. out-group. If one can be identified as a member of the in-group as confirmed by someone else in the social network, then certain behaviors are expected and a certain level of credibility is awarded. Additionally, a special authority—or credibility—is given to the information that "identified" individual presents. If one is viewed as a member of the out-group (his or her identity cannot be confirmed by someone else in the social network), different behaviors are expected and different, more restricted levels of trust and disclosure are granted.

Also, in social network cultures, outsiders (members of a different culture) tend to be viewed with suspicion, and being "heard" or "listened to" often becomes a matter of having the proper introduction (Richmond, 1995; Hofstede, 1997; Scharf & MacMathuna, 1998). In such systems, if a person who is part of the network (has a known identity) says an outsider should be listened to, then that outsider is identified as "credible" by other members of the network. In this case, the identity of the outsider gains credibility by being associated with a particular individual who is known and is trusted by members of that system (Hofstede, 1997; Scharf & MacMathuna, 1998; Weiss, 1998). Success in such a cultural communication system becomes a matter of identity. That is, does the recipient of a message know the identity of the presenter and how do factors of identity affect perceptions of credibility?

Social network perspectives on identity might affect how individuals from such cultures use online communication technologies. For example, when an individual from a social network culture receives an e-mail message that lacks identity cues, will that individual trust the identity of the sender? Moreover, if individuals from such cultures (cultures in which a person's word is often that person's bond) see how easily online messages can be altered and reposted, would those individuals be willing to use online media to introduce new people or present new ideas? And how would resistance to disclosing information in cyberspace be perceived by individuals from cultures in which a fixed identity is not as important?

According to such cultural expectations, the identity of the presenter affects his or her credibility. This perceived credibility affects the degree of access an individual has to certain kinds of information in social network cultures. Persons with a solid and an easy-to-confirm identity are considered credible and gain access to information. Individuals with a limited/text-based identity, however, might be viewed as non-credible and would be limited in the information they could obtain. In such cases, identity = trust/credibility, and lack of identity = doubt.

The limited identity resulting from IOIs could therefore undermine the credibility of some participants in IOIs. This situation could lead to "unexpected" behaviors or reactions from members of social network cultures. (These behaviors might include ignoring important communiqués because the identity, and thus the credibility, of the sender is "suspect.") Such unexpected behavior could, in turn, lead to confusion or even offense.

This two-part literature review reveals instances where cultural perceptions of identity could cause problems in IOIs. As a result, identity appears to be a strong point of contention that merits further examination. From this point, researchers can use different methods to examine how this strong point of contention might affect discourse patterns in IOIs. For example, some researchers might conduct online case studies to observe how individuals from different cultural groups act and interact in online environments where identity remains restricted to texts. Other researchers might conduct controlled experiments in which different cultural groups interact in situations where identity is known and stable, and in situations where identity is limited and text based.

The purpose of these approaches would be to determine if identity actually affects how cultural groups interact in IOIs. Such research could also be used to establish a comparative system for determining how specific cultures differ according to this particular point of contention. The overall objective would therefore be to determine if a strong point of contention could allow for a relativistic comparison of cultural behavior according to a similar concept. The results of such research could provide important insights that could be used to shape everything from international e-commerce strategies to communication protocols used in international outsourcing projects.

CONCLUSION

The rapid evolution of online communication technologies is constantly changing how people think about space and time. Now, international communication often transpires in seconds or minutes, not days or weeks. This new degree of proximity, however, could lead to an increase in cross-cultural misunderstanding as many aspects of online interactions contradict the communication expectations of certain cultures. This essay has overviewed how the international digital studies approach can serve as a foundation for examining international online interactions. It now becomes the task of researchers to examine culture, technology, and communication in order to explore the true nature of IOIs.

REFERENCES

Arnold, E.A. & Plymire, D.C. (2000). The Cherokee Indians and the Internet. In D. Gauntlett (Ed.), *Web.Studies* (pp. 186-193). New York: Oxford University Press.

Campbell, C.P. (1998). Rhetorical ethos: A bridge between high-context and low-context cultures? In S. Niemeier, C.P. Campbell & R. Dirven (Eds.), *The cultural context in business communication* (pp. 31-47). Philadelphia: John Benjamins.

Fernback, J. (1999). There is a there there: Notes toward a definition of cybercommunity. In S. Jones (Ed.), *Doing Internet research* (pp. 203-220). Thousand Oaks, CA: Sage Publications.

Ferraro, G. (2002). *Global brains: Knowledge and competencies for the 21st century.* Charlotte, NC: Intercultural Associates.

Gauntlett, D. (2000). Web studies: A user's guide. In D. Gauntlett (Ed.), *Web.Studies* (pp. 2-18). New York: Oxford University Press.

Hofstede, G. (1997). *Culture and organizations: Software of the mind.* New York: McGraw-Hill.

Lewis, R. (2003). *The cultural imperative: Global trends in the 21st century.* Yarmouth, ME: Intercultural Press.

Li, X. (1999). *Chinese-Dutch business negotiations: Insights from discourse.* Atlanta, GA: Rodopi.

Li, X. & Koole, T. (1998). Cultural keywords in Chinese-Dutch business negotiations. In S. Niemeier, C.P. Campbell & R. Dirven (Eds.), *The cultural context in business communication* (pp. 185-213). Philadelphia: John Benjamins.

Ng, K.Y. & Van Dyne, L. (2001). Culture and minority influence: Effects on persuasion and originality. In C.K.W. De Dreu & M.K. De Vries (Eds.), *Group consensus and minority influence: Implications for innovation* (pp. 284-306). Malden, MA: Blackwell Publishers.

Richmond, Y. (1995). *From da to yes: Understanding East Europeans.* Yarmouth, ME: Intercultural Press.

St.Amant, K. (2002a). International digital studies: A research approach for examining international online interactions. In E. Buchanan (Ed.), *Virtual research ethics: Issues and controversies* (pp. 317-337). Hershey, PA: Idea Group Publishing.

St.Amant, K. (2002b). When cultures and computers collide. *Journal of Business and Technical Communication, 16*(2) 196-214.

Scharf, W.F. & MacMathuna, S. (1998). Cultural values and Irish economic performance. In S. Niemeier, C.P. Campbell & R. Dirven (Eds.), *The cultural context in business communication* (pp. 145-164). Philadelphia: John Benjamins.

Tebeaux, E. (1999). Designing written business communication along the shifting cultural continuum: The new face of Mexico. *Journal of Business and Technical Communication, 13,* 49-85.

Ulijn, J.M. & Strother, J.B. (1995). *Communicating in business and technology: From psycholinguistic theory to international practice.* Frankfurt, Germany: Peter Lang.

Ulijn, J.M. & Campbell, C.P. (1999). Technical innovations in communication: How to relate technology to business by a culturally reliable human interface. *Proceedings of the 1999 IEEE International Professional Communication Conference* (pp. 109-120). Piscataway, NJ: IEEE Professional Communication Society.

Warnick, B. (1998). Rhetorical criticism of public discourse on the Internet: Theoretical implications. *Rhetoric Society Quarterly, 28,* 73-84.

Weiss, S.E. (1998). Negotiating with foreign business persons: An introduction for Americans with propositions on six cultures. In S. Niemeier, C.P. Campbell & R. Dirven (Eds.), *The cultural context in business communication* (pp. 51-118). Philadelphia: John Benjamins.

Woolever, K.R. (2001). Doing global business in the information age: Rhetorical contrasts in the business and technical professions. In C.G. Paneta (Ed.), *Contrastive rhetoric revisited and redefined* (pp. 47-64). Mahwah, NJ: Lawrence Erlbaum Associates.

KEY TERMS

Cross-Cultural: Situations where individuals from different cultures interact with or exchange information with one another; interchangeable with the term "intercultural."

Digital Ethos Conditions: Factors individuals use to assess the credibility or the worth of an online presentation of information.

Ethos Conditions: Factors individuals use to asses the credibility or the worth of a presentation.

Intercultural: Situations where individuals from different cultures interact with or exchange information with one another; interchangeable with the term "cross-cultural."

International Online Interaction (IOI): Situation in which individuals from two or more cultures use an online medium to interact directly with one another.

Literature Review: Reviewing the research findings reported by a certain field of research; the process usually involves the examination of articles published in the research journals of a particular field.

Point of Contention: A communication factor that is found to be positive (contribute to one's credibility) in the research of one field, but is found to be negative (detract from one's credibility) in the research of another field.

Presentation: The sharing of information with other via spoken, written, or online media.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1618-1622, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Multimedia Technology

International Standards for Image Compression

Jose Oliver Gil
Universidad Politécnica de Valencia, Spain

Otoniel Mario López Granado
Miguel Hernandez University, Spain

Miguel Onofre Martínez Rach
Miguel Hernandez University, Spain

Pablo Piñol Peral
Miguel Hernandez University, Spain

Carlos Tavares Calafate
Universidad Politécnica de Valencia, Spain

Manuel Perez Malumbres
Miguel Hernandez University, Spain

INTRODUCTION

Only a few decades ago, the human-computer interaction was based on a rudimentary text user interface, a convenient method compared to the punch card era, but too tedious and not very appealing for the nonspecialist, and thereby, not suitable for the mass market. Later on, the multimedia era arrived, with personal computers and other devices having powerful graphic capabilities, plenty of full-coloured pictures shown to the user. Although images made more pleasant the interaction with computers, their use represented a new challenge for electronic engineers; while only a few bytes are needed to represent a text (typically one byte per character in extended ASCII), lots of data must be employed for images that, in a "raw" representation (i.e., uncompressed) for colour images, need as many as three bytes per single *pixel* (picture element, each dot forming an image). Thus, there was a clear urge to reduce the amount of bytes required to encode an image, mainly so as to avoid an excessive increase in both memory consumption and network bandwidth required to store and transmit images, which would limit or prevent their use in practice.

In general, in order to exploit the information system resources in an efficient way when dealing with images, compression is almost mandatory. Fortunately, most images are characterized by highly redundant signals (especially natural and synthetic images), since pixels composing an image present high homogeneity, and this redundancy, often called *spatial redundancy*, can be reduced through a compression process, achieving a more compact representation.

BACKGROUND

The main contribution of this chapter is a brief survey on ISO standards for image coding. This chapter is organised as follows. For continuous-tune still-image lossy compression, the chapter reviews the classic ISO JPEG standard and the newer ISO JPEG 2000, while for lossless compression, the ISO JPEG-LS is presented. In addition, the authors review the JBIG standard, which aims at binary image coding, being widely used for fax transmission.

For continuous tone images (e.g., those from a digital camera), each pixel takes a value in a (nonbinary) range; typically any value in the range $[0..2^8\text{-}1]$ for a greyscale image. Actually, for colour images, three bytes per pixel are commonly used (one byte per colour component, red, green, and blue, which is known as RGB colour space), where each pixel represents a colour from up to 2^{24} possibilities. The well-known standard, JPEG (JPEG, 1992), focuses on this type of image. Its sequential mode, widely used throughout the entertainment industry, is based on removing information that is hardly perceived by a human viewer (in particular, high-frequency components are less accurately encoded). Although the decompressed image is not equal to the original one if compared pixel by pixel, the perceptual quality could be nearly the same, provided that no heavy compression is applied. An evolution from this standard is the new JPEG 2000 standard (JPEG 2000, 2000), based on the same ideas, which is more efficient and flexible. However, the higher complexity of JPEG 2000 and the current widespread use

of JPEG make the success of this new version uncertain in the mass market.

This type of image coding process, where recovered data is not the encoded one, is called *lossy compression*. It is important to emphasise that the comparison of lossy encoders cannot be performed based on the final image size alone, but also on its visual quality, in a sort of cost/benefit ratio (usually represented as a rate/distortion curve).

If it is important to recover exactly the original image, for example, for legal reasons in medical imaging or in image editing, *lossless compression* can be done at the cost of lower compression performance (but no quality loss). Although JPEG 2000 has a lossless mode, specific standards, like JPEG-LS (JPEG-LS, 1997), offer better performance. Another famous lossless image encoder is GIF (CompuserveIncorporated, 1987). Frequently used on the Internet, it is intended for 256-colour images that are previously selected in a palette.

Finally, the authors deal with binary images where each pixel can take two different values. These images are typically employed in fax transmission, and are compressed using the JBIG standard (JBIG, 1993).

CONTINUOUS-TONE STILL-IMAGE COMPRESSION

The JPEG Standard

The most widely used algorithm for image coding is defined in the JPEG standard (JPEG, 1992). Introduced in the 1980s, and developed by the Joint Photographic Experts Group (from which the standard is named after), nowadays, it is the most common way of encoding pictures in a wide range of important applications, such as image transmission on the Internet and image storage for digital cameras. This algorithm is able to encode colour images with an average compression rate of 15:1 with good visual quality (Furth, 1995).

The JPEG standard provides four working modes, three of them are lossy and the other one is lossless (Pennebaker & Mitchell, 1992). All the lossy modes employ the two-dimensional discrete cosine transform (2D-DCT) to analyse the spatial-frequency features of the images so as to store with less precision (or even remove) those frequency components least important for a human observer (according to the human visual system). Another important role of the DCT is to achieve high compactness of the information: after the DCT is applied, a substantial part of the image information is concentrated in only a few transform coefficients, mainly the low frequency ones, and thus, it can be represented in an efficient way.

Among its four compression modes, the sequential mode is extensively used, being the simplest and most well known. In this mode, the input image can be a greyscale image or a colour image. Although a digital colour image is commonly represented in the RGB space by an image capture device, it can be stored more efficiently in YCbCr space, in which the luminance component (*Y*) is represented separately from the blue and red chrominance components (*Cb* and *Cr* respectively) (note that the remaining green component, *Cr*, is not needed because it can be inferred from the *Cb*, *Cr* and *Y* components). Since the human eye is more sensitive to brightness information than to colour, the YCbCr colour representation allows us to reduce the size of the chrominance components without a significant degradation of the perceptual image quality. Thereby, the first step of the JPEG algorithm for colour image coding is a colour space transform from the input colour space to YCbCr, followed by a chrominance downsampling, usually a 4:2:0 subsampling, where the chrominance components are reduced by two in both horizontal and vertical directions (other possible downsampling is 4:2:2, in which only the horizontal direction is reduced by two, or 4:4:4, where there is no subsampling at all).

In the sequential mode, each image component is divided into 8×8 nonoverlapping blocks, and they are compressed and transmitted (or stored) in scan order, from left to right, and from top to bottom, so that the decoder can recover the image sequentially, in the same order as it was encoded. Each block is then processed as follows:

1. The two-dimensional DCT is applied to the entire block (it can be separately applied by using a 1D-DCT, first on the rows and then on the columns). Details on how to compute the DCT can be readily found in the literature (Ghanbari, 2003; Pennebaker & Mitchell, 1992).
2. The transform coefficients are then quantised to reduce information, most of all in high frequency components. This step is responsible for introducing information loss in the encoding process. The quantization process is done by dividing each coefficient by an associated constant value from a quantization matrix, rounding the result obtained to the nearest integer. This matrix is defined in such a manner that the higher frequency a coefficient represents, the higher the denominator (quantization value) becomes, and thereby more information reduction takes place.
3. The DC component of the current block is differentially encoded by using, as a reference, the DC component of the previous block.
4. The rest of the components are scanned in zig-zag order, from lower to higher frequencies, and joint run-length and entropy coding is done. With the run-length coding, a count of zero-values is performed, while entropy coding is a statistical compression method that encodes symbols by using an amount of bits inversely

proportional to the probability of its appearance (i.e., more likely symbols are encoded with fewer bits, and vice versa).

In addition, the JPEG algorithm allows varying the compression ratio by increasing the value of elements in the quantization matrix, at the expense of reducing image quality. This process is called rate control.

The other two lossy modes are the progressive and hierarchical modes, and both are based on the sequential one, adding more features to it. In particular:

a. In the progressive mode, data from all the blocks is interleaved (by interleaving bit-planes, or coefficients, or both of them (Ghanbari, 2003)) so that a blurry full-resolution image can be displayed when the first bytes are received, and more bytes from the bitstream can be used to obtain more sharpened and defined versions of the image (this feature is known as quality scalability or SNR scalability).
b. In the hierarchical mode, the original image is first downsampled by two several times, and then encoded and transmitted as in the sequential mode. Then, the encoder transmits the error committed by upsampling by two the low-resolution version of the image, and it repeats this process successively until the original resolution is achieved. By this process, the decoder is able to reconstruct larger images as more bytes are received, approaching the original resolution (this feature is called resolution scalability).

The last operation mode in JPEG is the lossless mode. In this coding method, no block-division and transform is done, being completely different in concept and usage to the lossy modes. In fact, it was introduced as a late addition a few years after the JPEG standard was released.

The JPEG 2000 Standard

The JPEG 2000 standard (JPEG 2000, 2000) was proposed at the beginning of the new millennium, two decades after the JPEG standard was released, aiming at replacing it. Much research had been done in the field of image coding, and this new standard was intended to collect and apply much of it. During these years, a new mathematical tool, called discrete wavelet transform (DWT), had aroused great interest in the field of image coding, mainly because it achieves high compactness of energy in the lower frequency sub-bands for natural images, which is extremely useful in image compression. Moreover, it allows resolution scalability in a natural way, since more wavelet sub-bands are used to progressively enlarge the low frequency sub-bands (Mallat, 1989). For these reasons, the JPEG 2000 standard replaced the use of the DCT by the DWT.

Another advantage of the DWT compared to the DCT is that there are computationally efficient algorithms to apply the two-dimensional transform to the entire image as a whole, with no need to partition it into smaller blocks. One of the side effects of the block processing in the DCT, and perhaps the main problem of JPEG, is that blocking artifacts appear with moderate to high compression ratios. According to the human visual system, block edges are easily identified and, hence, the visual image quality is highly degraded. Moreover, when small blocks are independently encoded, spatial redundancy is not optimally removed from the image. The use of the wavelet transform also solves these problems.

As a result of the replacement of the DCT by the DWT, the JPEG 2000 standard offers superior coding performance, especially at high compression rates, where blocking artifacts are more noticeable. Despite the bit rate saving of the new standard, maybe its main feature is a different one: it produces an extremely flexible and versatile bitstream. First, the user can indicate exactly the desired file size of the compressed image, while in the original JPEG, only a quality parameter can be used (and the final file size is unknown a priori). Besides, after encoding the wavelet coefficients, the bitstream is reorganised to achieve different types of resolution and quality scalability, with almost no loss of coding performance (observe that the progressive and hierarchical modes are less efficient than the sequential one in JPEG).

Let us now briefly describe the image coding process of JPEG 2000. After a colour space transform (if needed), each image component is transformed by the DWT, which is applied to the entire component as a whole (unless there are memory limitations), and the resulting transform coefficients are quantized to obtain integer values. Then, the embedded block coding with optimized truncation (EBCOT) algorithm (Taubman, 2000) is used to encode the coefficients in blocks (typically of 64×64 coefficients each block). Although block coding is also used in JPEG 2000, blocking artifacts do not appear because the image transform and quantization is applied to the entire image, and the block size is considerably bigger than in JPEG. The EBCOT algorithm consists in a bit-plane encoder (with three passes per bit-plane), followed by an optimization algorithm based on the Lagrange multiplier method (Everett, 1963), which is used to achieve the target file size in an optimal way, and a bitstream reorganization to fulfil the desired scalability.

Moreover, the original JPEG 2000 algorithm is able to work in lossless mode, as described previously, simply by utilizing a reversible integer-to-integer wavelet transform that ensures the exact reversibility of the transform coefficient (avoiding possible rounding errors), and by skipping the quantization stages.

Specific Lossless Image Encoders

Most lossless coders are based on entropy coding and predictive techniques. Predictive schemes try to calculate every sample from the previously encoded samples that are available to the encoder and the decoder. In lossless image coding, prediction is usually done from nearby pixels. After computing a prediction, the residual pixel is entropy encoded as the difference between the prediction and the original value. Basically, this is how the lossless mode of the JPEG standard works, a process that differs significantly from the lossy modes defined by that same standard.

Clearly, the better a prediction is, the lower the residual error becomes. In order to improve the prediction, the CALIC scheme (Wu & Memon, 1996) uses a very large amount of contexts (i.e., different statistics for the entropy coding depending on the neighbour pixels), achieving one of the best coding performances among the continuous-tone still-image lossless encoders. The high number of contexts makes CALIC quite complex and therefore, a simplification of CALIC was adopted as the JPEG-LS standard (JPEG-LS, 1997), which aims at improving and replacing the lossless mode of the original JPEG standard. The core algorithm of JPEG-LS is a simplified version of CALIC called LOCO-I (Weinberger, Seroussi, & Sapiro, 2000), whose performance is close to CALIC with lower complexity. It is able to provide both lossless and "near lossless" image compression. Moreover, it is more efficient and much faster than JPEG 2000 working in lossless mode.

Lossless image coding is also used to transmit bitmaps on the Internet in well-known services, such as the World Wide Web. Thus, the graphics interchange format (GIF) (CompuserveIncorporated, 1987) was introduced to losslessly encode images formed by 256 colours that are previously selected and stored in a colour table (also known as colour palette) from 2^{24} different possibilities (all the colours generated from a 24-bit RGB colour space). GIF images are compressed with the Lempel-Ziv-Welch (LZW) lossless encoder (Welch, 1984). The fact that LZW was a patented algorithm caused the PNG format to be proposed as a royalty-free alternative to GIF, with an additional support for palettes of 24-bit RGB colours (PNG, 2003).

BINARY IMAGE COMPRESSION

The JBIG standard (JBIG, 1993) was released by the Joint Bi-level Image Experts Group as an efficient solution to encode images with only two possible values for each pixel (known as binary or bi-level images). One of the main applications of this standard is fax transmission. JBIG uses a special entropy coder, the Q-coder (Pennebaker, Mitchell, Langdon, & Arps, 1988), as a basis to transmit the bi-level pixels in a lossless way.

An advanced version of this standard, called JBIG 2 (JBIG 2, 2001), was released in 2001. This new version is suitable for lossless, lossy, and perceptually lossless compression, and provides large increases in coding performance. This new standard is able to take advantage of image segmentation by separating an image into text, picture, graphical regions, and so forth, and encoding each different segment with the most appropriate algorithm.

FUTURE TRENDS

One of the greatest doubts about the future of image compression is if JPEG 2000 will be able to become a widespread standard, as occurred with the previous JPEG standard. At the time of writing this chapter, in early 2008, the support of images encoded with JPEG 2000 was not generalised in the World Wide Web, mainly because the most popular browsers (such as Internet Explorer and Firefox) do not support them yet. In order to improve the use of JPEG 2000 in interactive networks with client/server protocols, a new part of this standard has been recently released. This is Part 9, and is called JPIP. In this extension, a client can control and optimise the data flow downloaded from the server to meet the needed requirements.

Other new parts of the JPEG 2000 standard include Part 8 (JPSEC), which addresses security and content protection issues, Part 10 (JP3D), dealing with volumetric and 3-D image coding, and Part 11 (JPWL), for image transmission on wireless networks, including (a) protection of the bitstream against transmission errors (Part 1 only conceals errors), (b) detection of sensible areas of the bitstream, which will be protected more intensively, and (c) detection of errors that could not be corrected by the error protection decoder.

As for digital cameras, an increase of the native support of JPEG 2000 is expected because the specific hardware is becoming cheaper (e.g., Analog Devices, Inc. has recently lowered the price of their JPEG 2000 chip from USD 30 to USD 8). On the other hand, JPEG 2000 is the standard adopted by the digital cinema industry for media content distribution (each movie frame is encoded with a JPEG-2000-based encoder).

The main drawback of JPEG 2000 implementations is that they are computationally intensive and need a lot of memory to work, especially if compared to the sequential mode of JPEG. Thereby, alternative methods have recently been proposed (Cho & Pearlman, 2007; Oliver & Malumbres, 2006) to reduce the computational requirements while trying to preserve coding efficiency, many times at the cost of a less flexible bitstream than in JPEG 2000.

CONCLUSION

The main contribution of this chapter is a survey of the main existing standards for digital image compression. First, the authors have focused on current standard solutions for continuous-tone image coding, JPEG being the most commonly employed compression method, widely used in digital cameras and on the web. Then, the JPEG 2000 standard has been presented. The JPEG 2000 standard aims at replacing JPEG, although its use in common fields is still limited, and its success in the mass market is an open question. On the other hand, lossless coding is required at times (e.g., in medical imaging for legal issues, or in image editing, to avoid the accumulative errors from successive editions progressively damaging the image quality). Hence, both JPEG and JPEG 2000 have lossless modes, though it has been shown that JPEG-LS, the specific standard for lossless image coding, is a better option in most cases because it is faster and more efficient. Finally, if the image to encode is not continuous toned but binary instead, the standard solution is JBIG, which has been employed in several important applications, like fax transmission.

REFERENCES

Cho, Y., & Pearlman, W. (2007). Hierarchical dynamic range coding of wavelet sub-bands for fast and efficient image decompression. *IEEE Transactions on Image Processing*, *16*(8), 2005-2015.

CompuserveIncorporated. (1987). *Graphics interchange format specification version 89a.* Retrieved from http://www.w3.org/Graphics/GIF/spec-gif89a.txt

Everett, H. (1963). Generalized Lagrange multiplier method for solving problems of optimum allocation of resources.= *Operations Research*, *11*(3), 399-417.

Furht, B. (1995). A survey of multimedia compression techniques and standards. Part I: JPEG standard. *Real-Time Imaging*, 1-49.

Ghanbari, M. (2003). *Standard codecs: Image compression to advanced video coding*. The Institution of Electrical Engineers.

JBIG Standard. (1993). *ISO/IEC 11544 (ITU-T Recommendation T.82), 1993.*

JBIG 2 Standard. (2001). *ISO/IEC 14492 (ITU-T Recommendation T.88).*

JPEG-LS Standard. (1997). *ISO/IEC 14495-1 (ITU-T Recommendation T.87).*

JPEG Standard. (1992). *ISO/IEC 10918-1 (ITU-T Recommendation T.81).*

JPEG 2000 Standard. (2000). *ISO/IEC 15444-1, (ITU-T Recommendation T.800).*

Mallat, S. (1989). A theory for multiresolution signal decomposition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674-693.

Oliver, J., & Malumbres, M. P. (2006). Low-complexity multiresolution image compression using wavelet lower trees. *IEEE Transactions on Circuits and Systems for Video Technology*, *17*(11), 1437-1444.

Pennebaker, W. B., & Mitchell, J. L. (1992). *JPEG Still Image Data Compression Standard*. Van Nostrand Reinhold, International Thomson Publishing.

Pennebaker, W. B., Mitchell, J. L., Langdon, G. G., & Arps Jr., R. B. (1988). An overview of the basic principles of the Q-Coder adaptive binary arithmetic coder. *IBM Journal of Research and Development*, *32*(6), 717.

Portable Network Graphics (PNG) Specification. (2003). *Information technology- Computer graphics and image processing- Portable Network Graphics (PNG): Functional specification* (2nd. ed.). *ISO/IEC 15948.*

Taubman, D. (2000). High performance scalable image compression with EBCOT. *IEEE Transactions on Image Processing*, *9*(7), 1158-1170.

Weinberger, M., Seroussi, G., & Sapiro, G. (2000). The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. *IEEE Transactions on Image Processing*, *9*(8), 1309-1324.

Welch, T. A. (1984). A technique for high-performance data compression. *Computer*, *17*(6), 8-19.

Wu, X., & Memon, N. D. (1996). Context-based, adaptive, lossless image coding, *IEEE Transactions on Communications*, *45*(5), 437-444.

KEY TERMS

Chrominance: The difference between a certain colour component and the luminance component is called chrominance. The *Cb* and *Cr* components refer to the blue and red chrominances, increased by 0.5 and rescaled by 2 and 1.6, respectively (Pennebaker & Mitchell, 1992)

DC Component: When the two-dimensional discrete cosine transform is computed, the DC coefficient refers to the mean value of the pixels in the block, usually scaled according to the normalization used in the transform (in the

DCT transform applied in JPEG, the DC is the mean value multiplied by 8).

Entropy Coding: An entropy coder is a general lossless data compression method that encodes symbols by using an amount of bits inversely proportional to the probability of the symbols.

Human Visual System (HVS): In order to improve coding performance, a lossy encoder removes the information that is not seen by the human eye. The human visual system is the part of the nervous system that allows us to see. It has been modelled to determine its behaviour, and hence, to identify the information that can be removed without noticeable artifacts.

Lossless Coding: In lossless image compression, the decoded pixels are exactly the same as those that were encoded. For this reason, it is considered that there is no loss of data.

Lossy Coding: An image encoder can modify a source image in order to achieve higher compression ratios, while trying to keep the perceived quality unaltered. This is the case of lossy image compression, in which the decoded pixels do not maintain the same values as when they were encoded.

Luminance: The brightness of an image is determined by the luminance component (usually referred to as *Y*). It is the main component in a YCbCr colour space because the HVS is more sensitive to this component than to chrominance components. It can be computed from the RGB components as $Y = 0.2126\ R + 0.7152\ G + 0.0722\ B$.

Internet Abuse and Addiction in the Workplace

Mark Griffiths
Nottingham Trent University, UK

INTRODUCTION

As with the introduction of other mass communication technologies, issues surrounding Internet use, abuse and addiction have surfaced. This article has a number of objectives. It will first introduce readers to the concept of Internet addiction before going on to look at the wider issue of Internet abuse in the workplace. In this section, generic types of Internet abuse will be described, in addition to further examination of the reasons why Internet abuse occurs. The chapter ends with some guidelines and recommendations for employers and human resources departments.

BACKGROUND: INTERNET ADDICTION

There have been a growing number of academic papers about excessive use of the Internet. These can roughly be divided into four categories:

- Studies that compare excessive Internet users with non-excessive users (e.g., Brenner, 1997; Young, 1998)
- Studies that have examined vulnerable groups of excessive Internet use; for example, students (e.g., Nalwa & Anand, 2003; Scherer & Bost, 1997)
- Case studies of excessive Internet users (Catalano, Catalano, Embi & Frankel, 1999; Griffiths, 2000a; Tsai & Lin, 2003; Young, 1996)
- Studies that examine the psychometric properties of excessive Internet use (e.g., Armstrong, Phillips & Saling, 2000; Charlton, 2002; Pratarelli et al., 1999).
- Studies examining the relationship of excessive Internet use with other behaviors; for example, psychiatric problems, depression, loneliness, academic performance and so forth (e.g., Kubey, Lavin & Barrows, 2001; Nie & Ebring, 2000; Shapira, Goldsmith, Keck, Khosla & McElroy, 2000)

Despite the predominance of drug-based definitions of addiction, there is now a growing movement that views a number of behaviors as potentially addictive, including those which do not involve the ingestion of a psychoactive drug (e.g., gambling, computer game playing, exercise, sex, and now the Internet) (Orford, 2001). Research has suggested that social pathologies are beginning to surface in cyberspace. These have been termed "technological addictions" (Griffiths, 1996a) and have been operationally defined as non-chemical (behavioral) addictions that involve excessive human-machine interaction. They can thus be viewed as a subset of behavioral addictions (Marks, 1990) and feature core components of addiction (Brown, 1993; Griffiths, 1996a); that is, salience, mood modification, tolerance, withdrawal, conflict and relapse. Young (1999) claims Internet addiction is a broad term that covers a wide variety of behaviors and impulse control problems. This is categorized by five specific subtypes:

- Cybersexual addiction: compulsive use of adult Web sites for cybersex and cyberporn
- Cyber-relationship addiction: over-involvement in online relationships
- Net compulsions: obsessive online gambling, shopping or day-trading
- Information overload: compulsive Web surfing or database searches.
- Computer addiction: obsessive computer game playing (e.g., Doom, Myst, Solitaire, etc.)

In reply to Young, Griffiths (2000a) has argued that many of these excessive users are not "Internet addicts" but just use the Internet excessively as a medium to fuel other addictions. Put very simply, a gambling addict or a computer game addict who engages in their chosen behavior online is not addicted to the Internet. The Internet is just the place where they engage in the behavior. However, in contrast to this, there are case study reports of individuals who appear to be addicted to the Internet itself (e.g., Young, 1996; 2000b). These are usually people who use Internet chat rooms or play fantasy role playing games - activities that they would not engage in except on the Internet itself. These individuals to some extent are engaged in text-based virtual realities and take on other social personas and social identities as a way of feeling good about themselves.

In these cases, the Internet may provide an alternative reality to the user and allow them feelings of immersion and anonymity that may lead to an altered state of consciousness. This in itself may be highly psychologically and/or physiologically rewarding. There are many factors that make the Internet seductive. It is clear from research in the area of computer-mediated communication that virtual environments

have the potential to provide short-term comfort, excitement, and/or distraction (Griffiths, 2000a). These reasons alone provide compelling reasons alone why employees may engage in non-work related Internet use. There are also other reasons that are outlined in more detail in the next section on Internet abuse.

Case study accounts (e.g., Griffiths, 2000b; Tsai & Lin, 2003; Young, 1996) have shown that the Internet can be used to counteract other deficiencies in the person's life (e.g., relationships, lack of friends, physical appearance, disability, coping, etc.). Internet addiction appears to be a bona fide problem to a small minority of people but evidence suggests the problem is so small that few employers take it seriously. It may be that Internet abuse (rather than Internet addiction) is the issue that employers should be more concerned about.

TYPES OF WORKPLACE INTERNET ABUSE

It is clear that the issue of Internet abuse and Internet addiction are related but they are not the same thing. Furthermore, the long-term effects of Internet abuse may have more far-reaching effects for the company that the Internet abuser works for than the individual himself or herself. Abuse also suggests that there may not necessarily be any negative effects for the user other than a decrease in work productivity.

As seen in the previous section, Young (1999) claims Internet addiction is a broad term that covers a wide variety of behaviors and impulse control problems categorized by five specific subtypes. These can be adapted and refined to produce a typology of Internet abuse within the workplace. These are cybersexual Internet abuse, online friendship/relationship abuse, Internet activity abuse, online information abuse, criminal Internet abuse, and miscellaneous Internet abuse. These are examined in more detail below.

- Cybersexual Internet abuse: this involves the abuse of adult Web sites for cybersex and cyberporn during work hours. Such online sexual services include the conventional (e.g., Internet versions of widely available pornographic magazines like Playboy), the not so conventional (Internet versions of very hardcore pornographic magazines) and what can only be described as the bizarre (various discussion groups). There are also pornographic picture libraries (commercial and free-access), videos and video clips, live strip shows, live sex shows and voyeuristic Web-cam sites (Cooper, 2000; Griffiths, 2001).
- Online friendship/relationship abuse: this involves the conducting of an online friendship and/or relationship during work hours. Such a category could also include the use of e-mailing friends and/or engaging in discussion groups, as well as maintenance of online emotional relationships. Such people may also abuse the Internet by using it to explore gender and identity roles by swapping gender or creating other personas and forming online relationships or engaging in cybersex (see above) (Griffiths, 2001; Whitty, 2003).
- Internet activity abuse: this involves the use of the Internet during work hours in which other non-work related activities are done (e.g., online gambling, online shopping, online travel booking, online computer gaming, online day-trading, etc.). This may be one of the most common forms of Internet abuse in the workplace.
- Online information abuse: this involves the abuse of Internet search engines and databases. Typically, this involves individuals who search for work-related information on databases and so forth but who end up wasting hours of time with little relevant information gathered. This may be deliberate work-avoidance but may also be accidental and/or non-intentional. It may also involve people who seek out general educational information, information for self-help/diagnosis (including online therapy) and/or scientific research for non-work purposes.
- Criminal Internet abuse: this involves seeking out individuals who then become victims of sexually-related Internet crime (e.g., online sexual harassment, cyberstalking, paedophilic "grooming" of children). The fact that these types of abuse involve criminal acts may have severe implications for employers.
- Miscellaneous Internet abuse: this involves any activity not found in the above categories such as the digital manipulation of images on the Internet for entertainment and/or masturbatory purposes (e.g., creating celebrity fake photographs where heads of famous people are superimposed onto someone else's naked body) (Griffiths, 2001).

WHY DOES INTERNET ABUSE OCCUR?

There are many factors that make Internet abuse in the workplace seductive. It is clear from research in the area of computer-mediated communication that virtual environments have the potential to provide short-term comfort, excitement, and/or distraction (Griffiths, 2000a). These reasons alone provide compelling reasons why employees may engage in non-work related Internet use. There are also other reasons (opportunity, access, affordability, anonymity, convenience, escape, disinhibition, social acceptance, and longer working hours), which are briefly examined next:

- Opportunity and access – Obvious pre-cursors to potential Internet abuse include both opportunity and access to the Internet. Clearly, the Internet is now commonplace and widespread, and is almost integral to most workplace environments. Given that prevalence of undesirable behaviors is strongly correlated with increased access to the activity, it is not surprising that the development of Internet abuse appears to be increasing across the population.
- Affordability - Given the wide accessibility of the Internet, it is now becoming cheaper and cheaper to use the online services on offer. Furthermore, for almost all employees, Internet access is totally free of charge and the only costs will be time and the financial costs of some particular activities (e.g., online sexual services, online gambling, etc.).
- Anonymity - The anonymity of the Internet allows users to privately engage in their behaviors of choice in the belief that the fear of being caught by their employer is minimal. This anonymity may also provide the user with a greater sense of perceived control over the content, tone, and nature of their online experiences. The anonymity of the Internet often facilitates more honest and open communication with other users and can be an important factor in the development of online relationships that may begin in the workplace. Anonymity may also increase feelings of comfort since there is a decreased ability to look for, and thus detect, signs of insincerity, disapproval, or judgment in facial expression, as would be typical in face-to-face interactions.
- Convenience - Interactive online applications such as e-mail, chat rooms, newsgroups, or role-playing games provide convenient mediums to meet others without having to leave one's work desk. Online abuse will usually occur in the familiar and comfortable environment of home or workplace, thus reducing the feeling of risk and allowing even more adventurous behaviors.
- Escape - For some, the primary reinforcement of particular kinds of Internet abuse (e.g., to engage in an online affair and/or cybersex) is the sexual gratification they experience online. In the case of behaviors like cybersex and online gambling, the experiences online may be reinforced through a subjectively and/or objectively experienced "high". The pursuit of mood-modifying experiences is characteristic of addictions. The mood-modifying experience has the potential to provide an emotional or mental escape and further serves to reinforce the behavior. Abusive and/or excessive involvement in this escapist activity may lead to problems (e.g., online addictions). Online behavior can provide a potent escape from the stresses and strains of real life.
- Disinhibition – Disinhibition is clearly one of the Internet's key appeals, as there is little doubt that the Internet makes people less inhibited (Joinson, 1998). Online users appear to open up more quickly online and reveal themselves emotionally much faster than in the offline world. What might take months or years in an offline relationship may only takes days or weeks online. As some have pointed out (e.g., Cooper & Sportolari, 1997), the perception of trust, intimacy and acceptance has the potential to encourage online users to use these relationships as a primary source of companionship and comfort.
- Social acceptability – The social acceptability and perception of being online has changed over the last 10 years (e.g., the "nerdish" image of the Internet is almost obsolete). It may also be a sign of increased acceptance as young children are exposed to technology earlier and so become used to socializing using computers as tools. For instance, laying the foundations for an online relationship in this way has become far more socially acceptable and will continue to be so. Internet interaction takes away the social isolation that we can all sometimes feel and there are no boundaries of geography, class or nationality.

FUTURE TRENDS: GUIDELINES FOR MANAGERS AND HUMAN RESOURCES DEPARTMENTS

As has been demonstrated, being able to spot someone who is an Internet addict or an Internet abuser can be very difficult. However, there are some practical steps that can be taken to help minimize the potential problem.

- Develop an "Internet Abuse At Work" policy. Many organizations have policies for behaviors such as smoking or drinking alcohol. Employers should develop their own Internet abuse policies by liaison between personnel services and local technology councils and/or health and safety executives.
- Take the issue of Internet abuse/addiction seriously. Internet abuse and addiction in all their varieties are only just being considered as potentially serious occupational issues. Managers, in conjunction with personnel departments, need to ensure they are aware of the issues involved and the potential risks it can bring to both their employees and the whole organization. They also need to be aware that for employees who deal with finances, the consequences of some forms of Internet abuse/addiction (e.g., Internet gambling) for the company can be very great.

- Raise awareness of Internet abuse/addiction issues at work. This can be done through e-mail circulation, leaflets, and posters on general notice boards. Some countries will have national and /or local agencies (e.g., technology councils, health and safety organizations, etc.) that can supply useful educational literature (including posters). Telephone numbers for these organizations can usually be found in most telephone directories.
- Ask employees to be vigilant. Internet abuse/addiction at work can have serious repercussions not only for the individual but also for those employees who befriend Internet abusers and addicts, and the organization itself. Fellow staff need to know the basic signs and symptoms of Internet abuse and addiction. Employee behaviors such as continual use the Internet for non-work purposes might be indicative of an Internet abuse problem.
- Give employees access to diagnostic checklists. Make sure that any literature or poster within the workplace includes a self-diagnostic checklist so that employees can check themselves to see if they might have (or be developing) an Internet problem.
- Monitor Internet use of your staff that you suspect may have problems. Those staff with an Internet-related problem are likely to spend great amounts of time engaged in non-work activities on the Internet. Should an employer suspect such a person, they should get the company's IT specialists to look at their Internet surfing history, as the computer's hard disc will have information about everything they have ever accessed. The fact that specific individuals may be monitored should be outlined in the organization's "Internet Abuse At Work" policy so that employees are aware they may be monitored.
- Check Internet "bookmarks" of your staff. In some jurisdictions across the world, employers can legally access the e-mails and Internet content of their employees. One simple check is to simply look at an employee's list of "bookmarked" Web sites. If they are spending a lot of employment time engaged in non-work activities, many bookmarks will be completely non-work related (e.g., online dating agencies, gambling sites).
- Give support to identified problem users. Most large organizations have counseling services and other forms of support for employees who find themselves in difficulties. In some (but not all) situations, problems associated with Internet use need to be treated sympathetically (and like other more bona fide addictions such as alcoholism). Employee support services must also be educated about the potential problems of Internet abuse and addiction in the workplace.

CONCLUSION

In this chapter, major issues that surround Internet abuse/addiction issues in the workplace have been highlighted. Internet abuse/addiction can clearly be a hidden activity and the growing availability of Internet facilities in the workplace is making it easier for abuse to occur in lots of different forms. Thankfully, it would appear that for most people Internet abuse is not a serious individual problem, although for large companies, small levels of Internet abuse multiplied across the workforce raises serious issues about work productivity. For those whose Internet abuse starts to become more of a problem, it can affect many levels including the individual, their work colleagues and the organization itself.

Managers clearly need to have their awareness of this issue raised, and once this has happened, they need to raise awareness of the issue among the work force. Knowledge of such issues can then be applied individually to organizations in the hope that they can develop an Internet abuse policy in the same way that many organizations have introduced smoking and alcohol policies. Furthermore, employers need to let employees know exactly which behaviors on the Internet are reasonable (e.g., the occasional e-mail to a friend) and those that are unacceptable (e.g., online gaming, cybersex, etc.). Internet abuse has the potential to be a social issue, a health issue *and* an occupational issue and needs to be taken seriously by all those employers who utilize the Internet in their day-to-day business.

REFERENCES

Armstrong, L., Phillips, J.G., & Saling, L. (2000). Potential determinants of heavier Internet usage. *International Journal of Human-Computer Studies, 53*, 537-550.

Brenner, V. (1997). Psychology of computer use: XLVII. Parameters of Internet use, abuse and addiction: The first 90 days of the Internet usage survey. *Psychological Reports, 80*, 879-882.

Brown, R.I.F. (1993). Some contributions of the study of gambling to the study of other addictions. In W.R. Eadington & J.A. Cornelius (Eds.), *Gambling behavior and problem gambling* (pp. 241-272). Reno: University of Nevada Press.

Catalano, G., Catalano, M.C., Embi, C.S., & Frankel, R.L. (1999). Delusions about the Internet. *Southern Medical Journal, 92*, 609-610.

Charlton, J.P. (2002). A factor analytic investigation of computer 'addiction' and engagement. *British Journal of Psychology, 93*, 329-344.

Cooper, A. (Ed.). *Cybersex: The dark side of the force* (pp. 5-29). Philadelphia: Brunner Routledge.

Cooper, A., & Sportolari, L. (1997). Romance in cyberspace: Understanding online attraction. *Journal of Sex Education and Therapy, 22,* 7-14.

Griffiths, M.D. (1996a). Behavioural addictions: An issue for everybody? *Journal of Workplace Learning, 8*(3), 19-25.

Griffiths, M.D. (1996b). Internet "addiction": An issue for clinical psychology? *Clinical Psychology Forum, 97,* 32-36.

Griffiths, M.D. (2000a). Internet addiction: Time to be taken seriously? *Addiction Research, 8,* 413-418.

Griffiths, M.D. (2000b). Does Internet and computer "addiction" exist? Some case study evidence. *CyberPsychology and Behavior, 3,* 211-218.

Griffiths, M.D. (2001). Sex on the Internet: Observations and implications for Internet sex addiction. *Journal of Sex Research, 38,* 333-342.

Joinson, A. (1998). Causes and implications of disinhibited behavior on the Internet. In J. Gackenback (Ed.), *Psychology and the Internet: Intrapersonal, interpersonal, and transpersonal implications* (pp. 43-60). New York: Academic Press.

Kubey, R.W., Lavin, M.J., & Barrows, J.R (2001, June). Internet use and collegiate academic performance decrements: Early findings. *Journal of Communication,* 366-382.

Marks, I. (1990). Non-chemical (behavioural) addictions. *British Journal of Addiction, 85,* 1389-1394.

Nalwa, K., & Anand, A.P. (2003). Internet Addiction in students: A cause of concern. *CyberPsychology and Behavior, 6,* 653-656.

Nie, N.H., & Ebring, L. (2000). *Internet and society: A preliminary report.* Stanford, CA: The Institute for the Quantitative Study of Society.

Orford, J. (2001). *Excessive appetites: A psychological view of the addictions* (2nd ed.). Chichester: Wiley.

Pratarelli, M.E., Browne, B.L., & Johnson, K. (1999). The bits and bytes of computer/Internet addiction: A factor analytic approach. *Behavior Research Methods, Instruments and Computers, 31,* 305-314.

Scherer, K., & Bost, J. (1997, August). *Internet use patterns: Is there Internet dependency on campus?* Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, IL.

Shapira, N.A., Goldsmith, T.D., Keck, P.E., Khosla, U.M., & McElroy, S.L. (2000). Psychiatric features of individuals with problematic Internet use. *Journal of Affective Disorders, 57,* 267-272.

Tsai, C.-C., & Lin, S.S.J. (2003). Internet addiction of adolescents in Taiwan: An interview study. *CyberPsychology and Behavior, 6,* 649-652.

Whitty, M.T. (2003). Pushing the wrong buttons: Men's and women's attitudes toward online and offline infidelity. *CyberPsychology and Behavior, 6,* 569-579.

Young, K. (1996). Psychology of computer use: XL. Addictive use of the Internet: A case that breaks the stereotype. *Psychological Reports, 79,* 899-902.

Young, K. (1998). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology and Behavior, 1,* 237-244.

Young, K. (1999). Internet addiction: Evaluation and treatment. *Student British Medical Journal, 7,* 351-352.

KEY TERMS

Conflict: This refers to the conflicts between the addict and those around them (interpersonal conflict), conflicts with other activities (job, social life, hobbies and interests) or from within the individual themselves (intrapsychic conflict) that are concerned with the particular activity.

Cybersex: The act of computer-mediated sex either in an online or virtual environment. Examples include two consenting adults engaging in an e-mail or real-time chat sex session. The advantages to this are that two people who are at opposite ends of the globe can maintain a relationship.

Internet Addiction: This is a term used to describe excessive Internet use and has been also been referred to as Internet addiction disorder (IAD), Internet addiction syndrome (IAD) and pathological Internet use. As with other addictions, Internet addiction features the core components of other addictive behaviors (salience, mood modification, tolerance, withdrawal, conflict and relapse) and can be defined as a repetitive habit pattern that increases the risk of disease and/or associated personal and social problems. It is often experienced subjectively as "loss of control" and these habit patterns are typically characterized by immediate gratification (short-term rewards), often coupled with delayed, deleterious effects (long-term costs). Attempts to change an addictive behavior (via treatment or by self-initiation) are typically marked by high relapse rates (see also technological addictions).

Mood Modification: This refers to the subjective experiences that people report as a consequence of engaging in the particular activity and can be seen as a coping strategy (i.e., they experience an arousing "buzz" or a "high" or paradoxically, tranquilizing feel of "escape" or "numbing").

Relapse: This is the tendency for repeated reversions to earlier patterns of the particular activity to recur and for even the most extreme patterns typical of the height of the addiction to be quickly restored after many years of abstinence or control.

Salience: This occurs when the particular activity becomes the most important activity in the person's life and dominates their thinking (preoccupations and cognitive distortions), feelings (cravings) and behavior (deterioration of socialized behavior). For instance, even if the person is not actually engaged in the behavior they will be thinking about the next time they will be.

Technological Addictions: These addictions are operationally defined as non-chemical (behavioral) addictions that involve human-machine interaction. They can either be passive (e.g., television) or active (e.g., computer games, Internet), and usually contain inducing and reinforcing features which may contribute to the promotion of addictive tendencies. Technological addictions can be viewed as a subset of behavioral addictions and feature core components of addiction, that is, salience, mood modification, tolerance, withdrawal, conflict and relapse.

Tolerance: This is the process whereby increasing amounts of the particular activity are required to achieve the former effects. For instance, a gambler may have to gradually have to increase the size of the bet to experience a euphoric effect that was initially obtained by a much smaller bet.

Withdrawal Symptoms: These are the unpleasant feeling states and/or physical effects that occur when the particular activity is discontinued or suddenly reduced, for example, the shakes, moodiness, irritability and so forth.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1623-1628, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Internet and Multimedia Communications

Dimitris Kanellopoulos
University of Patras, Greece

Sotiris Kotsiantis
University of Patras, Greece

Panayotis Pintelas
University of Patras, Greece

INTRODUCTION

Multimedia communications involve digital audio and video and impose new quality of service (QoS) requirements on the Internet (Lu, 2000). Different multimedia applications have different QoS requirements. For example, continuous media types such as audio and video require hard or soft bounds on the end-to-end delay, while discrete media such as text and images do not have any strict delay constrains. In addition, video applications require more bandwidth than audio applications. QoS requirements are specified by the following four closely related parameters: (1) bandwidth on demand; (2) low end-to-end delay; (3) low delay variation (or delay jitter); and (4) acceptable error or loss rate without retransmission, as the delay would be intolerable with retransmission. Multimedia applications are classified into the following three categories:

- *Two-way conversational applications*, which are characterized by their stringent requirement on end-to-end delay that includes total time taken to capture, digitize, encode/compress audio/video data, transport them from the source to the destination, and decode and display them to the user.
- *Broadcasting services* where the source is live. The main dissimilarity from the conversational applications is that it is one-way communication and it can stand more delay.
- *On-demand applications* (e.g., video on demand) where the user requests some stored items and the server delivers them to the user.

In designing and implementing multimedia applications, the characteristics of these application types should be used to provide required QoS, but using network and system resources efficiently. Even though we say that QoS should be guaranteed, the user states the degree of guarantees. Usually, there are three levels of guarantees:

- *Hard guarantee*, where user-specified QoS should be met absolutely. Reserving network and system resources based on the peak-bit rate of a stream achieves hard guarantees.
- *Soft guarantee*, where user-specified QoS is supposed to be met to a certain precise percentage. This is suitable for continuous media, as they usually do not need 100% accuracy in playback. This type of guarantee uses system resources more efficiently.
- *Best effort*, where no guarantee is given and the multimedia application is executed with whatever resources are available. More networks function in this mode.

These different types of guarantees may all be needed in a multimedia session established using proper association control protocols such as C_MACSE (Kanellopoulos & Kotsiantis, 2006). Different levels of guarantee are used for different types of traffic and the user determines which type of guarantee to use. Besides, the charging policy is related to the level of guarantee and the most expensive is the hard guarantee, while the best effort is the cheapest. At the source, multimedia data are either captured live or retrieved from storage devices. The transport module accepts these data, packetizes and passed them on to the Internet. At the destination (sink), multimedia data are reassembled and passed to the application for playback of audio/video. Packet processing time differences, network access time differences, and queuing delay difference can cause delay jitter, which has to be removed at the destination before data being played out.

BACKGROUND

Through a number of subsystems multimedia data flows. For example, an audio segment is encoded at the audio server program, sent through the underlying transport network, and decoded at the receiving application. In a multimedia multiparty call, the end-user issues diverse QoS parameters values,

which are interpreted onto specific performance parameter values in the communication subsystem and the operating system frameworks. To provide end-to-end QoS guarantees, an intensive effort is necessary from all subsystems, including end-subsystems, network hardware, and communication protocols, of a multimedia system.

Bandwidth On Demand

The speed of network links and routers in the next generation Internet will be improved radically so that network congestion will be uncertain and QoS guarantees will be provided by design. This endeavor will include optical wavelength-division multiplexing (WDM) technologies, being considered by Next Generation Internet (NGI) initiative (http://www.ccic.gov/ngi/).

Multicast Support

It is a common requirement of multimedia communication to send data from one source to multiple destinations. Efficient multicasting protocols can reduce bandwidth requirements (Paul, 1998). Given the multireceiver nature of video programs, real-time video distribution has emerged as one of the most important IP multicast applications and it requires bandwidth adaptability. Real-time video multicast applications have to adapt to the dynamic network conditions, but still offer reasonable playback quality to the receivers. Liu and Zhang (2003) present a survey on adaptive video multicast solutions. Because video and shared data are essential to many distributed tasks, audio of sufficient quality is a necessary condition for almost any successful real-time interaction.

Synchronization

Continuous media are characterized by well-defined temporal relationship between subsequent presentation units to be played. A presentation unit is a logical data unit that is perceivable by the user. *Multimedia synchronization* is the process of preserving the temporal order of one or more media streams. The problem of maintaining continuity within a single stream is referred as *intra-stream synchronization*; where as the problem of maintaining continuity among the streams is called *inter-stream synchronization*. These two types of synchronization are necessary for both live streams and for stored media streams presentations. Manvi and Venkataram (2006) proposed an agent-based synchronization framework to handle three synchronization mechanisms (point, real-time, and adaptive) at application service level depending on the life/run-time presentation requirements of the multimedia applications.

Adaptive Media Coding

Multimedia data should be coded in a way such that acceptable audio/video playback quality is still achieved, when some data packets are delayed extremely or lost. Coding multimedia data into multiple layers is the basic suggestion. Some layers are assigned high priority and they contain essential data to generate basic acceptable basic play out audio/video quality. Extra layers contain data that add additional details (or quality) to the basic quality and are assigned low priority. In the case of system overloading, low priority data are dropped first, leading to little effect to play out quality. This effect is named *graceful quality degradation*, and it can be obtained by the use of error control techniques such as forward error correction (FEC).

End System Support

End systems must offer mechanisms to handle multimedia data efficiently and effectively such as to provide end-to-end QoS guarantees (Lu, 1996). The end-system support for multimedia communications is required for two reasons. Firstly, the communications protocol stack is implemented mainly in software and it has thousands of instructions executed by the end system. If, for these instructions, the end system cannot guarantee the execution time, there will be no real-time communications system regardless of how well networking support is offered. Secondly, if the media data need to be compressed and decompressed before presentation, the processing time should be predictable. If not, a meaningful presentation is not obtained.

Hardware: Multimedia applications impose the following requirements on the hardware architecture:

- Digital audio and video are very data intensive, and therefore the hardware must have high data transfer throughput and high processing power.
- Parallel hardware architectures are preferred, as a lot of multimedia applications have to access several input/output devices simultaneously. In addition, multimedia host computers have usually I/O buses, which support lower transfer rates than of those of high-speed networks. This situation leads to the problem called "*mismatch in bandwidth.*" For the solution of this problem various network interface units must be implemented.
- The hardware architecture must be scalable to accommodate new input/output devices and applications.
- To support different types of data and applications, the architecture should be versatile and programmable.

Multimedia operating systems: They should meet the following requirements (Steinmetz, 1995).

- Multimedia operating systems should use the hardware resources efficiently so that use of these resources is maximized.
- QoS requirements of multimedia applications should be guaranteed by using proper resource management and process scheduling. At the operating system level, one of the main QoS requirements is the guaranteed processing time for each task.
- The multimedia operating system should execute multimedia applications as well as conventional applications. This has two implications. Firstly, the conventional application-programming interface (API) should be maintained. Secondly, the conventional applications should not be starved of resources, while QoS requirements of multimedia applications are guaranteed.

Mobility Support

The advent of wireless and cellular networks has improved multimedia applications with mobility. Mobility aspect has added another dimension of complexity to multimedia networks. It opens up questions on a host of complex issues like routing to mobile terminals, maintaining the QoS when the host is in motion, interworking between wireless and wired networks. Kanellopoulos, Pintelas, and Giannoulis (2006) provide an overview of key QoS issues required to support end-to-end QoS guarantees for wireless multimedia communications.

PROPOSED APPROACHES FOR QoS GUARANTEES OVER THE INTERNET

Supporting multimedia communications over the Internet introduces two major problems: *fairness* and *useless packet transmission* (UPT). In current Internet, a single FIFO-based queue is used to multiplex both adaptive and nonadaptive traffic at routers. When multimedia applications cross the same router, this may cause fairness problems. UTP is based on the fact that for packetised audio and video, packet loss rate must be preserved under a given threshold for any meaningful communication. Wu and Hassan (2004) have proposed two different management policies for UPT.

ITU-T has developed a series of recommendations together comprising the *H.323 system* (Toga & Ott, 1999) that provides for multimedia communications in packet-based (inter)networks. The H.323 series of recommendations explain systems, logical components, messages and procedures that enable real-time, multimedia calls to be established between two or more parties on a packet network.

Giordano, Salsano, Van den Berghe, Ventre, and Giannakopoulos (2003) describe the current evolution of QoS architectures, mechanisms, and protocols in the Internet, as it is ongoing in the framework of the European Union funded research projects (AQUILA, CADENUS, TEQUILA) on premium IP networks.

The Internet provides the best-effort service to applications and thus cannot meet the QoS requirements of multimedia communications. However, many research efforts have been made toward providing QoS guarantees. Communication architecture of the Internet is altered or improved to provide QoS guarantees. Three main architectures have been proposed: (1) integrated service model (IntSer); (2) differentiated service model (DiffServ); and (3) multiprotocol label switching (MPLS). Bhargava, Wang, Khana, and Habib (2005) developed an adaptable network architecture (ADNET) which allows different QoS provision schemes such as active networks, integrated services and differentiated services to co-exist.

Integrated Service Model

The integrated service model (IntServ) is based on resource reservation (Braden et al., 1994). An appropriate amount of resources is reserved for each flow to meet its requirements. Resources including bandwidth and memory are reserved to meet QoS requirements of an application or communication session (Braden, Zhang, Berson, Herzog, & Habib, 1997). The following fundamentals are needed to provide QoS guarantees:

- A QoS specification mechanism for applications to specify their requirements;
- Admission control to determine whether the new application should be admitted without affecting the QoS of other ongoing applications;
- A QoS negotiation process so that as many applications as possible can be served. During the QoS negotiation phase (Figure 1), each peer computer must determine whether it can support the desired QoS. If so, certain resources are reserved for this session. If the user is satisfied with the suggested QoS, the session is established. Otherwise, the session is rejected.
- Resource allocation and scheduling to meet the QoS requirement of accepted applications; and
- Traffic policing to make sure that applications generate the correct amount of data within the agreed specification.

A QoS renegotiation mechanism is required so that applications can request changes in their initial QoS specifications. The actual QoS provided to the ongoing sessions should be monitored (Figure 2) so that suitable actions can be taken in case of any problem in providing specified QoS guarantees. Media scalability and graceful quality degra-

Figure 1. QoS negotiation

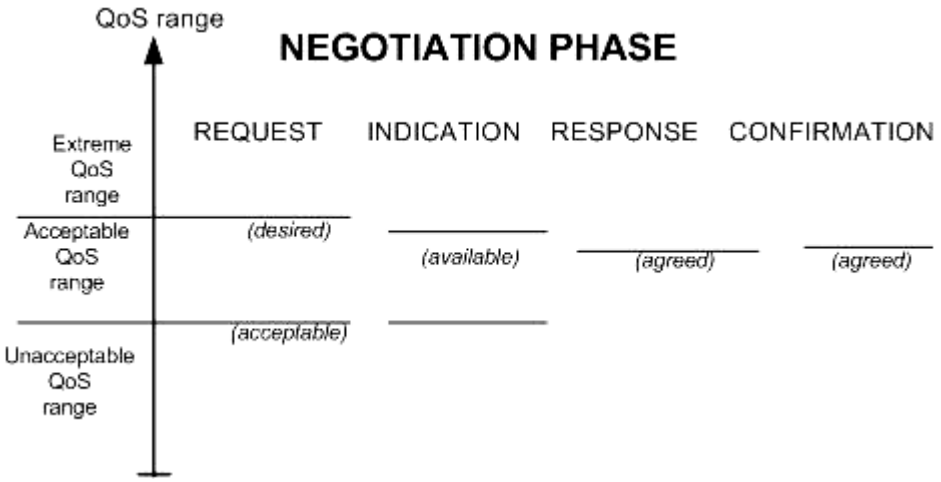
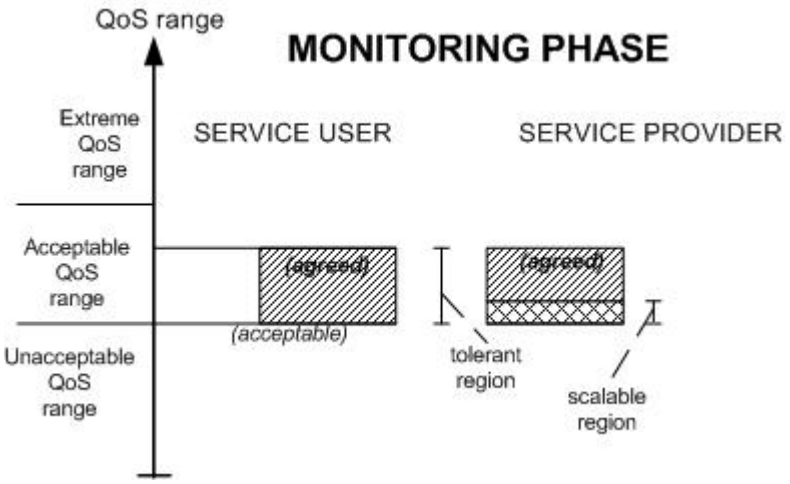


Figure 2. QoS monitoring



dation techniques should be used together with the above mechanisms to provide satisfactory services to multimedia applications.

Traffic-shaping schemes. Traffic shaping regulates a stream's traffic so that it is simple to describe and police. Traffic-shaping schemes are used when the traffic pattern is too complicated to describe directly or the traffic is not appropriate for networks to support directly. For example, when video is variable bit rate coded, it may be hard to characterize the coded bit stream.

Admission control, QoS negotiation and renegotiation. When a connection with specified QoS is established, QoS parameters are translated and negotiated among all relevant subsystems. Only when all subsystems are in agreement with and guarantee the specified QoS parameters, the end-to-end QoS requirements can be met.

Resource reservation protocols. A vital part of the IntServ model is the resource reservation protocol at the network layer. A resource reservation protocol transfers information about resource requirements and negotiates the QoS values that users desire for their end-to-end applications. In the current IntServ model, only RSVP (Braden et al., 1997) is used as the reservation protocol.

IntServ uses RSVP to make per-flow reservations at routers along a network path. While this allows the network to provide service guarantees at the flow level, it suffers from scalability issues. RSVP is a soft-state protocol, which means that the router's state has to be refreshed at regular intervals,

and this adds to traffic overhead. The service classes offered by IntServ are: (1) guaranteed service class, (2) controlled load service, and (3) best-effort service.

Differentiated Services

The main problem of the currently best effort model on the Internet is that all packets are treated the same, although different types of service can be determined in the IPv4 header. The main problem of the IntServ is that there is potentially an infinite number of different types of traffic so each router has to store essential information to provide QoS guarantees to each type of traffic. Differentiated services (Black, 1998) take a middle ground between the best-effort service and IntServ. It defines a *fixed* number of packet classes. All traffic types/packets are aggregated into these classes and the network/routers provide different services to different packet classes.

Service classification. IPv4 has an underused type-of-service byte in its header. The newer IPv6 has a header byte called *traffic class*. In DiffServ, the type-of-service (traffic class) byte is redefined as a *differentiated service* (DS) field. The first six bits of the DS field is called *DS CodePoint*, which indicates the behavior each router is required to apply to the individual packet. Packets with the DS CodePoint set to zero receive the same service as they get in the best effort service. Values between one and seven are defined to be backward compatible with the original IP precedence mechanism, to ensure that DiffServ technology can be deployed in the operational Internet progressively. The DS filed can be assigned by the customer (the transmitter process) to indicate the desired service. Alternatively, the ingress router marks the DS field based on *multifield* (MF) *classification*. MF classification classifies packets based on the contents of multiple fields such as source address, destination address, type-of-service byte, protocols ID, source port number and destination port number. As a packet moves from one Internet Service Provider (domain) to another, it may be reclassified. Many service classes can be defined. The Internet Engineering Task Force (IETF) working group has defined the following two new services.

- *Expedited* or *premium service*: to provide virtual leased line service to applications requiring low-delay and low delay jitter.
- *Assured service*: to provide better-than-best-effort services to applications.

In the past decade, Hou et al. (2000) proposed a Diffserv architecture for multimedia streaming applications in next generation Internet.

MULTIPROTOCOL LABEL SWITCHING (MPLS)

In an IP network, each router analyzes the packet's header and runs a network layer routing algorithm. Each router separately chooses a next hop for the packet, based on its analysis of the packet's header and the results of running the routing algorithm. This process introduces some latency, as the routing tables are very big and table lookups take time. Choosing the next hop consists of two steps. The first step classifies the entire set of possible packets into a set of *forwarding equivalence classes* (FECs). The second step maps each FEC to a next hop. As far as the forwarding decision is concerned, different packets, which are mapped into the same FEC, are indistinguishable and travel in the same path. In MPLS, the assignment of a particular FEC is done just once at the ingress router (Armitage, 2000). The FEC to which the packet is assigned is encoded as a short fixed length value known as a *label*. This label is inserted into the packet by the ingress router. At subsequent hops, there is no further analysis of the packet's network layer header. Instead, the label is used as an index into a table that specifies the next hop and a new label. The old label is replaced with the new label and the packet is forwarded to its next hop. The path taken by the packet is specified by a sequence of labels and is called a *label switched path* (LSP). The routers that support MPLS are called *label-switching routers* (LSRs). MPSL is a forwarding scheme and is much faster than IP routing. The label can represent a combination of a FEC and a precedence or class of service. Routers to provide differentiated QoS to different types of traffic can treat packets with different labels differently.

FUTURE TRENDS

In the future, high-speed multimedia networks will be mainly wireless at the edges, with access to a high-speed optical backbone infrastructure, including optical switches. One can imagine a network that consists only of wireless access to an optical backbone. Fourth-generation mobile wireless networks will emerge to offer capacities up to 150 Mb/s to fully mobile users in various environments. In such environments, resource management will remain an important issue. Efficient resource management is a hot issue due to: the rapid increase in size of the wireless mobile community; its demand for high-speed multimedia communications; and the limited resources.

CONCLUSION

This chapter provides an overview of technologies required to support end-to-end QoS guarantees over the Internet. Besides the network layer architectures, proper transport protocols such as RTP (Schulzrinne, Cassner, Frederick, & Jacobson, 1996) and application dependent protocols such as Session Initiation Protocol (SIP) (Johnston, 2000) are required for multimedia communications. To achieve end-to-end QoS guarantees an effort from all subsystems including end systems and all networking components is required. The communication architectures IntServ, DiffServ and MPLS will inter-operate with each other in the Internet environment. IntServ will be implemented in edge networks where the number of flows is small, while DiffServ will be putted into practice in the core of the Internet where the number of flow is high (Lu, 2000). MPLS together with DiffServ will provide differentiated services and QoS guarantees.

REFERENCES

Armitage, G. (2000). MPLS: The magic behind the myths. *IEEE Communications Magazine, 38*(1), 124-131.

Bhargava, B., Wang, S-Y., Khana, M., & Habib, A. (2005). Multimedia data transmission and control using active networks. *Computer Communications, 28*(6), 623-639.

Black, D. (1998, May). *An architecture for differentiated services*. Internet Draft.

Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997, September). *Resource ReSerVation Protocol (RSVP), version 1 functional specification*. Internet RFC2205.

Braden, R., et al. (1994, July). *Integrated services in the Internet architecture: An overview*. Internet RFC1633.

Giordano, S., Salsano, S., Van den Berghe, S., Ventre, G., & Giannakopoulos, D. (2003). Advanced QoS provisioning in IP networks: The European premium IP projects. *IEEE Communications Magazine, 41*(1), 30-36.

Hou, Y.-T., Wu, D., Li, B., Hamada, T., Ahmad, I., & Jonathan Chao, H. (2000). A differentiated services architecture for multimedia streaming in next generation Internet. *Computer Networks, 32*(2), 185-209.

Johnston, A.B. (2000). *Understanding the session initiation protocol*. Artech House.

Kanellopoulos, D., & Kotsiantis, S. (2006). C_MACSE: A novel ACSE protocol for hard real-time multimedia communications. *International Journal of Computer Science and Network Security, 6*(3), 57-72.

Kanellopoulos, D., Pintelas, P., & Giannoulis, S. (2006). QoS in wireless multimedia networks. *Annals of Mathematics, Computing & TeleInformatics, 1*(4), 66-75.

Liu, J., & Zhang, Y.-Q. (2003). Adaptive video multicast over the Internet. *IEEE Multimedia, 10*(1), 22-31.

Lu, G. (1996). *Communication and computing for distributed multimedia systems*. Artech House.

Lu, G. (2000). Issues and technologies for supporting multimedia communications over the Internet. *Computer Communications, 23*(14/15), 1323-1335.

Manvi, S., & Venkataram, P. (2006). An agent based synchronization scheme for multimedia applications. *The Journal of Systems and Software, 79*(5), 701-713.

Paul, S. (1998). *Multicasting on the Internet and its applications*. Dordrecht: Kluwer.

Schulzrinne, H., Cassner, S., Frederick, R., & Jacobson, V. (1996, January). *RTP: A transport protocol for real-time applications*. RFC1889.

Steinmetz, R. (1995). Analyzing the multimedia operating system. *IEEE Multimedia, 2*(1), 68-84.

Toga, J., & Ott, J. (1999). ITU-T standardization activities for interactive multimedia communications on packet-based networks: H.323 and related recommendations. *Computer Networks, 31*(3), 205-223.

Wu, J., & Hassan, M. (2004). Avoiding useless packet transmission for multimedia over IP networks: The case of multiple multimedia flows. *Computer Communications, 27*(7), 651-663.

KEY TERMS

ACSE (Association Control Service Element): In the Open Systems Interconnection Reference Model (OSI-RM), an element of the application layer, which is responsible for the establishment, termination and control of associations between two or more communication parties (programs).

Bandwidth on Demand: It refers to data rate measured in bit/s (channel capacity or throughput-bandwidth consumption), which is required in order to transfer continuous media data (e.g., video).

Delay Variation (or delay jitter): It is a loose term for the variation of end-to-end delay from one packet to the next packet within the same packet stream (connection/flow).

End-to-end Delay: It is a significant parameter affecting the user's satisfaction with the application. It includes capturing, digitizing, encoding/compressing media data,

transporting them from the source to the destination, and decoding and displaying them to the user.

Multicast: The capability of a network to transmit data simultaneously to many receivers with no need to replicate the data.

Multimedia Data: It refers to data that contain various media types such as text, graphics, animation, audio, and video.

QoS (Quality of Service): It refers to the capability of a networked system to provide scalable service to selected network traffic.

The Internet and SMEs in Sub-Saharan African Countries: An Analysis in Nigeria

Princely Ifinedo
University of Jyväskylä, Finland

INTRODUCTION

The Internet is a global network of interconnected computers using multiple Internet protocols (IP). Increasingly, it is being used to enhance business operation by both small and medium-sized enterprises (SMEs) and large organizations around the world (Bunker and MacGregor, 2002; Turban, Lee, King, & Chung, Lee, J., King, D. & Chung 2004). One reason is that the Internet, when used to facilitate e-commerce and e-business, offers several benefits for the adopting organizations (Walczuch, den Braven, & Lundgren, 2000; Turban, et al, 2004). Such benefits include the following: 1) reducing distance barrier, 2) the development of new products and services, 3) opening direct links between customer and suppliers, and 4) enhancing communication efficiency. Our study of the relevant literature reveals that the diffusion of the Internet among businesses in Sub-Sahara Africa (SSA), including SMEs, is the lowest in the world at around 2% (ITU, 2005). A recent report shows that the whole of Africa has only 1% of the total international Internet bandwidth (UNCTAD, 2005). Thus, it is to be expected that businesses in the region with such poor connectivity and use will be unable to fully reap the benefits of the technology. Against such unfavorable situations, it would seem reasonable for research efforts to uncover why such unfavorable conditions prevail in the region. Sadly, very few studies exist that have investigated such issues. Little is known about the perceptions of the Internet or the factors inhibiting its spread among SMEs in SSA. To fill this gap in research, this article aims at adding to knowledge by presenting a summary of the findings of a preliminary study designed to investigate the perceptions of the Internet and the sorts of barriers facing SMEs in SSA desiring to adopt Internet in their operations or for commerce. The study used SMEs in Nigeria, a Sub-Saharan African country. The country was chosen for illustration proposes as it is the most populous country in Africa and has favorable indicators for the use of information and communication technologies (ICT) compared to other SSA countries (Ifinedo, 2005). Importantly, researchers, for example, Ojukwu (2006) have discussed use of ICT among Nigerian SMEs and it is hoped that this present effort will complement similar research efforts.

BACKGROUND

SMEs and Economies

SMEs can be described in several ways, for example, the European Parliament's definition of SMEs refers to a business with up to 250 employees. de Klerk and Kroon (2005), writing from the perspectives of the Republic of South Africa, divided SMEs into three main subcategories: micro (< 5 people), small (between 5-50 people) and medium-sized (51-200 people). Nonetheless, we accept SMEs as businesses characterized by informal planning, strong owner's influence, lack of specialists, small management teams, heavy reliance on few customers, and limited knowledge, amongst others (Bunker and MacGregor, 2002; Ifinedo, 2006). It is generally accepted that SMEs are the engine of growth of all economies (Bunker and MacGregor, 2002), including those in Africa. According to Ojukwu (2006), 97% of all businesses in Nigeria employ less than 100 employees, and the same is true in many African countries (Ifinedo, 2006). That said, SMEs in developed nations have been able to use ICT products such as the Internet in establishing e-commerce and e-business initiatives (Bunker and MacGregor, 2002; Turban, et al., 2004) and have subsequently benefited from such exercises. On the contrary, little or no progress has been made on such fronts in many developing countries, including SSA ones due to a variety of reasons, including inadequate know-how and a lack of resources (Ifinedo, 2005, 2006; Ojo, 1996; Ojukwu, 2006; Okoli, 2003).

SSA and Internet Commerce

Africa, with its population of about 1 billion people, is the poorest continent in the world (World Bank, 2005). In terms of geography, Africa tends to be described as being comprised of two regions - North Africa and SSA. The Northern part is comparable to the Middle East economically and culturally (Ifinedo, 2005). Further, South Africa (also known as the Republic of South Africa (RSA)) tends to be excluded from the rest of SSA because of its relatively high socio-economic indicators. The conditions in SSA are different from those in the excluded regions, and the region of SSA

Figure 1. Nigeria on the map of Africa

Source: Worldcountries.info (2006).

typifies perceptions of Africa more than do the excluded regions. The map Africa highlighting Nigeria, the chosen SSA country for this study is shown in Figure 1.

According to the latest World Bank (2005) reports Africa continues to be the only continent with worsening socio-economic indicators. In particular, SSA lags behind on the adoption and use of ICT products such as the Internet. Africa has the lowest diffusion rates for ICT products (for example, computers and telephones) (UNCTAD, 2005). Further, ITU (2005) shows that there were only 4.9 Internet hosts per 10, 0000 Africans in 2004 and the statistics are unchanged two years. When ICT products are lacking, it is not surprising that Internet commerce (also known as e-commerce) is relatively low on the African continent compared to the rest of the world (Economist Intelligence Unit, 2003; Gateway, 2003; UNCTAD, 2005; Ifinedo, 2005, 2006). Reports by Gateway (2003) state that "E-commerce is concentrated in South Africa and Egypt (regions not considered in this article), while B2B (business-to-business) outside South Africa remains negligible." As noted above, few have investigated why the penetration rates of ICT products such as the Internet and other ICT products are relatively low in Africa in general and SSA in particular. The available few studies suggest that the slow diffusion can be attributed to the unavailability of technological/technical expertise and other know-how in the region (de Klerk and Kroon, 2005; Ojukwu, 2006; Ifinedo, 2006) while others attribute it to unfavorable cultural influences (Ojo, 1996; Okoli, 2003). Yet others trace the poor penetration rates to a lack of financial resources in the SSA region (Duncombe and Heeks, 1999; Beliamourne-Lutz, 2003). In particular, Duncombe and Heeks (1999) highlighted the factors below as the main reasons why the adoption of the Internet among SMEs in Botswana - an SSA country – is very low.

- Lack of money.
- Lack of skills or knowledge.
- Lack of technological infrastructure, for example, telecommunications, electricity supply.
- Lack of awareness of the business environment and how to take advantage of it.
- Lack of "critical mass": Few local people/organizations use computers/email; thus most are unable to engage in any meaningful Internet commerce.

As mentioned above, the focus of this article is to present the perception of the Internet and barriers for not using it by Nigerian SMEs. The issues relating to the perceptions of the Internet were adapted from the work of Duncombe and Heeks (1999), and with regard to the factors militating against the use of the Internet in Nigerian SMEs, we adapted the variables from Walczuch, den Braven, & Lundgren (2000) to suit the SSA context. That said, we conducted a survey in Nigeria using the judgmental or purposive sampling approach in selecting our respondents (Neuman, 1997). With the approach, the researcher selects his or her respondents given their suitability for the study's theme. We collected data from SMEs in three Nigerian cities, namely, Lagos, Ibadan, and Port-Harcourt. The chosen cities are among the largest commercial cities in the country. (There is an "urban-rural divide" with regard to the use of ICT facilities in Nigeria which suggests that SMEs in larger cities are more likely to use ICT products in their operations than those in other parts of the country (Ifinedo, 2006). The SMEs selected follow the definitions provided above. Overall, 63 usable questionnaires were obtained for analysis. The views of 22 (35%) proprietors and 41(65%) supervisory/managerial employees are used. Our respondents came from wide ranging industries, including Agro-based business, light engineering (metal works), constructions, retail, and IT consultancy. The majority of the respondents are aged between 26 and 55 years and more than half have university degrees. Our respondents classified according to the sub-classifications offered by de Klerk and Kroon (2005) are as follows: 32 (50.8%), 19 (30.2%), and 12 (19%) respectively for micro, small, and medium-sized organizations. Participation was voluntary and respondents were motivated with a promise that their responses will be kept anonymous. We used a five-point Likert scale to represent the issues, please see Table 1.

A summary of the results is shown in Table 1. Only 5 (7.9%) of the SMEs in our sample have their own Web pages. Many of the SMEs - 81% - plan to have their own Web pages at some future time. E-mail is the commonest Internet facility used by the Nigerian SMEs. The Internet (e-mail) is accessed more from Cybercafés than anywhere else which suggests that the SMEs do not have their own extranets and intranets. 35% of the sampled SMEs agree to using e-mails regularly. About 90% of the respondents "surely" accept that the Internet could be a source of business and business information. Finally, while all of the respondents noted that they prefer conducting business transactions face-to-face, only 22% indicated that the Internet could serve as an alternative transaction platform.

Table 2 shows the barriers or reasons for not using the Internet by SMEs in Nigeria. We briefly discuss the top ten issues due to space limitations. The "lack of IT skills for e-commerce development" placed as the topmost barrier. This harks back to the general limitations confronting the SSA region with respect to the availability of technological/technical expertise and know-how (Ifinedo, 2005; Ojukwu, 2006). The next barrier relates to the "general lack of finance", which researchers such as Beliamourne-Lutz (2003) and Duncombe and Heeks (1999) contend is the most prominent inhibitor to ICT (including the Internet) diffusion in developing countries. The third highly ranked barrier relates to the costs of buying computers. Feedback from one manager in an Agro-based business reads: "To do buying and selling online, our company needs facilities (computers, telephones, and so forth); we're a small business, we can't afford to invest such huge sums when apparently our business stands to benefit from any spare cash." The majority of the feedback received echo similar sentiments.

The next barrier concerns "Technical complexity of using IT in business". This, in some respects, is related to the topmost barrier on the list. The lack of technical know-how continues to be among the most cited inhibitors facing SMEs both in developed and developing countries in their use of ICT-enabled services (Ifinedo, 2006). High costs of using and subscribing to telephone services ranked fifth. This is

Table 1. Perception of the Internet among Nigerian SMEs

	Number (N)	Percent (%)		Number (N)	Percent (%)
Preference for conducting business:			***Are you planning having a Web page in the near future?***		
Face-to-face	63	100	Yes		
Telephone / fax	29	43	No	51	81
The Internet	14	22.2	Undecided	4	6.3
Others	11	17.5	Missing data	4	6.3
				4	6.3
Does your business have a Web page?			***Does your business have Internet access?***		
Yes	5	7.9	Yes	57	90.5
No	58	92.1	No	4	6.3
			Missing data	2	3.2
Which one in particular?			***How often do you use the Internet (e-mail, WWW etc.)?***		
E-mail	53	84.1	Very often	35	56.5
WWW (Web page)	5	7.9	Often	23	37.1
Others	4	6.4	Sometimes	3	4.8
Missing data	1	1.6	Never used it	1	1.6
Where do you (and your business) access the Internet?			***Accepting the Internet as a source of business and business information***		
Own business premises	8	12.9			
Cyber cafés	49	79	Surely	56	88.9
Public places (library etc.)	1	1.6	Not too sure	5	7.9
Friend's	2	3.2	It cannot be a source…	1	1.6
Other places	2	3.2	No comment	1	1.6

Table 2. Reasons for **NOT** *having the Internet used by SMEs in Nigeria*

Reasons or barriers	N	Min	Mean	Std dev.	Max
Lack of IT skills for e-commerce development	63	2	4.84	0.54	5
General lack of finance	63	3	4.43	0.78	5
Too expensive to buy computers	63	3	4.25	0.74	5
Technical complexity of using IT in business	63	3	4.11	0.79	5
Telephone costs is too high	63	3	4.10	0.89	5
Internet subscription fee is too high	63	2	3.97	1.08	5
Not clear government policy and support	63	2	3.97	1.15	5
Poor energy (power) supply	63	2	3.87	1.11	5
Not many of many our customers use the Internet	63	1	3.60	1.01	5
Not suitable for our business	63	1	3.49	1.03	5
The Internet is not safe for our business	63	1	3.32	0.76	5
Does not lead more efficiency or lower costs	63	1	3.14	1.12	5
Lack of time	60	2	3.08	0.79	5
The Internet is too slow	63	1	2.89	1.32	5
The Internet is too difficult to use	63	1	2.75	1.29	5
Does not lead to more sales in our business	63	1	2.62	0.73	5
Never thought about it	57	1	2.25	0.93	5

Likert scale ranging from strongly agree (5) to strongly disagree (1)

not surprising given the poor statistics for Africa on ICT use and diffusion (ITU, 2005). Apparently, the sixth placed barrier is related to the preceding one. The lack of clear government policy and support for e-commerce (and the use of the Internet for business in Nigerian SMEs) ranked seventh. Poor energy supply came in eighth in the ranking order of barriers. Feedback from a manager/co-owner of a Business Centre reads "If you do business with the Internet, would NEPA [National Electric Power Authority] understand and improve power supply?" She emphasized that due to the poor power supply, her business has had to procure a generating plant to supplement the services provided by the state through NEPA. Indeed, the EIU (2003) noted poor power generation in Nigeria as one of the impediments to e-business growth in the country.

The lack of critical mass among SMEs with the technology which is captured by the barrier of "not many of our customers use the Internet" came in ninth place. Finally, the last of the top-ten reasons that the Internet is not used by SMEs in Nigeria is related to the suitability of the technology for local business initiatives. To illustrate this point, a proprietress of a small confectionary business responded that "I make wedding cakes and also participate in the [wedding] ceremonies; my presence is a must for my business…. The Internet can't help my business", when asked why she has not considered the use of the Internet for business. Some scholars have suggested that the socio-cultural contexts in developing countries, especially those in SSA, might not be conducive for Internet commerce. Okoli (2003, p.16) comments that "Africans do not have the culture of buying a product without [having] tactile contacts.

CONCLUSIONS AND FUTURE STUDY

The findings of this preliminary study have added to our understanding of the perceptions and use of the Internet among SMEs in SSA. It is easy to notice that a lack of resources (that is, technological, financial, and human) explains the poor state of affairs with regard to the use of the Internet among SMEs in SSA. The general lack of IT skills among SMEs in Nigeria can be attenuated through well-planned schemes and programs. For instance, governments in SSA could follow cues from other developing countries in Asia where cooperation between states and SMEs has started yielding encouraging results with regard to how SMEs in those parts of the world perceive and use of the Internet in their business operations. For example, in Thailand, the Ministry of Commerce arranged free homepages for SMEs. In addition to paying for the homepages, the Thai govern-

ment helped to provide relevant technical skills and training for some of these SMEs. Our study shows that SMEs in Nigeria may be able to benefit from the Internet as well as if their infrastructural, human, and financial capabilities are improved. To assist SMEs in SSA make progress with the use of the Internet for their operations and commerce, SSA governments could consider the types of assistance being offered in Thailand or grant soft loans and subsidies to those showing interest in adopting the Internet in their businesses. The availability of funding (finance) is vitally important in redressing the poor use of ICT among SMEs in SSA (Beliamourne-Lutz, 2003; Duncombe and Heeks, 1999; Ifinedo, 2006), and this should be accorded great attention. Efforts need to be directed toward sensitizing SMEs in SSA about modern business practices and technologies. Norms, traditions, or culture must not be allowed to stand in the way of progressive development. This study provides a basis for further investigations. For example, studies could be commissioned to investigate the perceptions of the Internet and the sorts of factors that might be limiting its use among large organizations in SSA. Future studies could consider the perceptions of the Internet among SMEs located in rural communities in SSA; a comparative study between the two parts (rural vs. urban) could enrich the discourse of Internet use among SMEs in SSA.

REFERENCES

Beliamourne-Lutz, M. (2003). An analysis of the determinant and effects of ICT diffusion in developing countries. *Information Technology for Development, 10*, 151-169.

Bunker, D.J., & MacGregor, R.C. (2002). *The context of Information Technology and electronic commerce adoption in Small/Medium Enterprises: A global perspectives*. In Proceedings of the 8th. AMCIS 2002 conference, August 9 – 11, Dallas, Texas.

de Klerk, S. & Kroon, J. (2005). E-commerce adoption in South African businesses. *South African Journal of Business Management, 36(1)*, 33 - 40.

Duncombe, R. & Heeks, R. (1999). Informatics, ICT and small enterprises: Findings from Botswana Development. *Informatics, Working Paper Series, No. 7*, The University of Manchester, the UK. Retrieved September 13, 2006, from http: www.sed.manchester.ac.uk/idpm/publications/wp/di/di_wp07.pdf.

Economist Intelligence Unit (EIU) (2003). The 2003 E-readiness Rankings: A White Paper. Retrieved September 13, 2006, from http://unpan1.un.org/intradoc/groups/public/documents/APCITY/UNPAN010006.

Gateway (2003). Good prospects for IT industry in 2003. Retrieved August 11, 2005, from, http://www.gateway.hr/index.php?folder=148&article=43.

Ifinedo, P. (2005). Measuring Africa's e-readiness in the global networked economy: A nine-country data analysis. *The International Journal of Education and Development using Information and Communication Technology*, 1 (1), 53-71.

Ifinedo, P. (2006). Factors affecting e-business adoption by SMEs in Sub-Saharan Africa: An exploratory study from Nigeria. In N. Al-Qirim, (Ed.), *Global Electronic Business Research: Opportunities and Directions*. (pp. 319-346). Hershey, PA: Idea Group Publishing.

ITU (2005). ICT statistics . Retrieved September 1, 2006, from http://www.itu.int/ITU-D/ict/statistics/.

Neuman, W. L. (1997). *Social Research Method*, (2nd.). London: Allyn and Bacon.

Ojo, S.O. (1996). Socio-cultural and organizational issues in IT application in Nigeria. In M. Odedra-Straub (Ed.), *Global Information Technology and Socio-Economic Development*. (pp 99-109). Marietta, GA: Ivy League Publishing.

Ojukwu, D. (2006). Achieving sustainable growth through the adoption of integrated business and information solutions: A case study of Nigerian small & medium-sized enterprises. *Journal of Information Technology Impact, 6(1)*, 47- 60.

Okoli, C. (2003). *Expert assessments of e-commerce in Sub-Saharan Africa: A theoretical model of infrastructure and culture for doing business using the Internet.* Unpublished PhD thesis, Louisiana State University, USA.

Turban, E., Lee, J., King, D. & Chung, H.M. (2004). *Electronic Commerce: A Managerial Perspective*. New Jersey: Prentice-Hall.

UNCTAD (2005). United Nations Conference on Trade and Development. *The digital divide report: ICT diffusion index2005*. Retrieved September 7, 2006, from http://www.unctad.org/templates/webflyer.asp?docid=6994&intItemID=2068&lang=1&mode=downloads

Walczuch, R., den Braven, G., & Lundgren, H. (2000). *Internet adoption barriers for small firms in the Netherlands*. In Proceedings of the 6th. AMCIS conference, August 10 - 13 : Long Beach, California.

Worldcountries.info (2006). Map of Nigeria. Retrieved September 16, 2006, from http://www.worldcountries.info/Maps/Region/Africa-450-Nigeria.jpg

World Bank. (2005). World Development Reports. Retrieved September 10, 2006, from http://www.worldbank.org/.

KEY TERMS

Cybercafé: This a place where one can use a computer with access to the Internet for a fee; it may or may not serve as a regular café serving food and drinks.

Business-2-Business (B2B): This refers to the commercial activities between firms online.

Business-2-Customer (B2C): This refers to the commercial activities between firms and customers online.

E-Business: This refers to the sharing of business information, collaborating with business partners as well as conducting business transactions by means of the telecommunication networks.

E-Commerce: This term is used to describe the buying and selling of goods and services online.

Information and Communications Technology (ICT): These include technological products such as telephones, computers, the Internet, and so forth which are used to convert, store, protect, process, transmit, and retrieve information.

The Internet: This is a global network of interconnected computers using multiple Internet protocols.

The Internet Protocol (IP): This is a unique number that devices use in order to identify and communicate with each other in a computer network.

Small and Medium Sized Enterprises (SMEs): These are businesses characterized by informal planning, strong owner's influence, lack of specialists, small management teams, heavy reliance on a few customers, limited knowledge, and often employ less than 250 workers.

Sub-Saharan Africa (SSA): This is the region of Africa excluding the Northern part of the continent and the Republic of South Africa.

The Internet and Tertiary Education

Paul Darbyshire
Victoria University, Australia

Stephen Burgess
Victoria University, Australia

INTRODUCTION

For many years, information technology (IT) has been used to find ways to "add value" for customers to entice them to purchase the products and services of a business. This article examines the possibility of translating the benefits of "added value" to the use of the Internet by tertiary educators for subject and course delivery. Many educators use the Internet to supplement existing modes of delivery. Importantly, the Internet is providing a number of "added value" supplemental benefits for subjects and courses delivered using this new, hybrid teaching mode. There are two aspects to subject delivery to where "added value" benefits may be applied, and that is in the *administrative tasks* associated with a subject and the *educational tasks*. In both instances, IT solutions can be employed to either fully or partially process some of these tasks. Given the complex and often fluid nature of the education process, it is rare that a fully integrated solution can be found to adequately service both aspects of subject delivery. Most solutions are partial in that key components are targeted by IT solutions to assist the subject coordinator in the process. If we examine closely the underlying benefits gained in the application of IT to these tasks, there is a strong parallel to the benefits to be gained by business organizations with similar applications of IT. While the actual benefits sought by academics depend on the motivation for the IT solution, the perceived benefits can be classified using standard categories used to gauge similar commercial applications. However, from an *educational* viewpoint online technologies provide educators with new challenges, especially in relation to dealing with issues related to plagiarism and class attendance. These need to be considered by educators when deciding how, and if, to incorporate the Internet into their curriculum.

BACKGROUND

In order to investigate the benefits of using Web-based techniques to supplement traditional teaching in terms of business efficiencies, the reasons that commercial organizations use IT are examined. The different aspects of subject delivery also need to be considered in order to determine the ultimate benefits to be gained.

Information Technology: Efficiency and Added Value

There are a number of reasons for using IT in organizations (O'Brien, 1999):

- **For the support of business operations:** This is usually to make the business operation more efficient (by making it faster, cheaper and more accurate).
- **For the support of managerial decision making:** By allowing more sophisticated cost benefit analyses, providing decision support tools and so forth.
- **For the support of strategic advantage:** This refers to the use of Porter and Millar's (1985) three generic strategies as a means of using information technology to improve competitiveness by adding value to products and services.

It has been recognized for a number of decades that the use of computers can provide cost savings and improvements in efficiencies in many organizations. Porter et al. (1985) have generally been credited with recognising that the capabilities of information technology can extend further to providing organizations with the opportunity to add value to their goods. Value is measured by the amount that buyers are willing to pay for a product or service. Porter et al. (1985) identify three ways that organizations can add value to their commodities or services (known as *generic strategies for improving competitiveness*):

- Be the lowest cost producer
- Produce a unique or differentiated good (providing value in a product or service that a competitor cannot provide or match, at least for a period of time). If an organization is the first to introduce a particular feature, it may gain a competitive advantage over its rivals for a period. Some ways in which information technology can be used to differentiate between products and/or services are (Sandy & Burgess, 1999):

  - Improved quality
  - Superior product support
  - Time (delivering products or services faster)

- Provide a good that meets the requirements of a specialised market

The next sections examine the possibility of translating the benefits of "added value" to a particular application of IT, the use of the Internet by tertiary educators to assist with subject and course delivery.

Aspects of Course and Subject Delivery

There are two overall aspects to course and subject delivery, the educational and administrative components (Darbyshire & Wenn, 2000). Delivery of the educational component of a subject to students is the primary responsibility of the subject coordinator, and this task is the most visible from a student's perspective. However, the administration tasks associated with a subject form a major component of subject coordination, but these responsibilities are not immediately obvious or visible to the students.

It is essential that all aspects of subject delivery be carried out as efficiently as possible. To this end, IT, and in particular, Web-based solutions can be applied to both aspects of subject delivery. That Web-based solutions are a suitable vehicle to use has been almost universally accepted by students, teachers, and academic administrators (Scott Tillett, 2000). Other advantages are the ease with which information can be disseminated, its interactivity, its use as a real-time communication medium and the ability to use text, graphics, audio, and video (Kaynama & Keesling, 2000).

There are a number of administrative tasks associated with subject coordination for which IT solutions can be applied in the application. These include (Byrnes & Lo, 1996; Darbyshire et al., 2000):

- **Student enrollment:** While most universities have a student enrolment system administered at the institute level, there are often local tasks associated with enrolment such as user account creation and compilation of mail lists etc. Some of these tasks can be automated (Darbyshire et al., 2000).
- **Assignment distribution, collection, and grading:** The written assignment remains the basic unit of assessment for the vast majority of educators, and there have been many initiatives to computerize aspects of this task. Some of these include *Submit* (Hassan, 1991), *NetFace* (Thompson, 1988), *ClassNet* (Boysen & Van Gorp, 1997) and *TRIX* (Byrnes et al., 1996).
- **Grades distribution and reporting:** Techniques for this range from e-mail to password protected Web-based database lookup.
- **Informing all students of important notices:** Notice boards and sophisticated managed discussion facilities can be found in many systems. Examples include products such as *TopClass*, *Learning Space*, *Virtual-U*, *WebCT*, and *First Class* (Landon, 1998)

Many of the tasks viewed as educational can also employ IT solutions in order to gain perceived benefits. Some of these include *online class discussions, learning, course outline distribution, seminar notes distribution, and answering student queries*. Just how many of these are actually implemented will relate to a number of factors such as the amount of face-to-face contact between lecturers and students. However, using the Internet for many of these can address the traditional problems of students misplacing handouts, and staff running out of available copies.

Discussion management systems are being integrated into many Web-based solutions. These are usually implemented as threaded discussions, which are easily implemented as a series of Web pages. Other tools can include chat rooms or listserv facilities. Answering student queries can take place in two forums, either as part of a class discussion or privately. Private discussions online are usually best handled via an e-mail facility, or in some instances, *store and forward messaging systems* may replace e-mail.

Implementing IT solutions to aid in the actual learning process is difficult. These can range from intelligent tutoring systems (Cheikes, 1995; Ritter & Koedinger, 1995), to facilitated online learning (Bedore, Bedore, & Bedore, 1998). However, the major use of IT solutions in the learning process is usually a simple and straight forward use of the Web to present hypertext based structured material as a supplement to traditional learning.

Using Internet Technologies to Improve Efficiency and Add Value

With the recent explosion in Internet usage, educators have been turning to the Internet in attempts to gain benefits by the introduction of IT into the educational process. In this article, subject delivery at the university level is only considered. The benefits sought from such activity depend on the driving motivation of the IT solution being implemented. While many may not perceive a university as a business (and it is not advocated here), it is nonetheless possible to match the current uses of the Internet in tertiary education with traditional theory related to the reasons why firms use IT.

Internet technologies in education, which are used for the learning process itself, target the student as the main stakeholder. While the motivation may be the enhancement of the learning process to achieve a higher quality outcome, we can loosely map this to the "*support of managerial decision making*" concept identified earlier. Such technologies allow educators to obtain a far more sophisticated analysis

of individual student's learning progress, and thus provide them with decision support tools on courses of action to take to influence this process.

Technology solutions, which target the academic as the stakeholder (Darbyshire et al., 2000), implement improvements or efficiencies that can be mapped to the *support of the business operation* previously identified. Improvements or efficiencies gained from such implementations are usually in the form of automated record keeping and faster processing time, ultimately resulting in lower costs in terms of academic time, and added value to the students.

By default, the university also becomes a stakeholder in the implementation of either of the previous types of technology enhancements. Benefits gained by students and staff by such uses of technology translates ultimately to lower costs for the institution or the provision of more and/or better quality information. The benefits of such systems can be mapped onto the *support of strategic advantage* concept (as Porter et al's low cost and differentiation strategies), previously identified as a reason for using technology in business. If these institutions are to regard themselves as a business, then the successful use of IT in subject delivery could give the university a strategic advantage over other universities, which it would regard as its business competitors. Most of the reported advantages gained from online supplementation of teaching relate to cost savings in terms of efficiency, flexibility, and/or convenience. These represent the traditional added value benefits of lower cost and faster access to goods in the commercial world. Thus, we can use the measures of *money savings*, *time savings*, *improved quality*, and better *product information* as categories to measure the benefits gained from the introduction of IT to supplement teaching.

VALUE ADDED ACHIEVED BY ACADEMICS

The authors were interested to investigate the extent of appreciation of the "value added" benefits that the Internet can offer to tertiary educators, institutions, and their students. In the first instance, a simple survey was conducted in 2001 through the IS world discussion list to gain an initial idea of the level of appreciation that existed (Darbyshire & Burgess, 2005a).

The most common benefit for administrative uses was to *save time* for the institution and for students. Most administrative benefits were similar for both groups except for *save money* (where more than twice the respondents felt that the institution saved money than students). The *information provision* administrative usages were the most commonly used (important notices, schedules/timetables, assignment, and grade distribution). Less common were the more *interactive* options, assignment collection, and student enrolment. Educational uses of the Internet were seen as providing slightly more benefits for students than institutions. Their uses were seen as providing more information and improving quality more on average than the administrative uses. As with administrative usages, the easiest educational features to set up were the most commonly used (distribute course/subject notes, provide external links). Less common were the more *interactive* options, discussion lists, and online chat groups. About three quarters of respondents used the Internet to answer student queries (probably by e-mail). As with administrative uses, most of the benefits are similar for students and the institution, with (again) some differences for instances where the benefits save money more for the institution than students. More respondents saw the differences in the benefits of educational uses flowing to students than to institutions than with administrative uses. In three of the uses, saving time *was not* the most common benefit identified. These were the provision of external links to additional resources, discussion lists, and online chats, where improved quality of information and more information were more commonly identified. More recently, there have successful trials of automatically generated online examinations. For instance, Patterson (2006) reports on the conversion of a four-hour examination delivered online to almost 200 students. The examination consisted of 100 randomly generated questions—so the test was different for each student. The vast majority of students found the examination easy to access and complete online.

ISSUES FOR ONLINE TEACHING

Having decided to adopt the Internet as part of their curriculum, academics are faced with a number of issues in dealing with the new paradigm. Two of these are *plagiarism* and *class attendance*.

Plagiarism has long been a problem for educators, but recent discussions seem to indicate that incidents of plagiarism are increasing. The easy access of information via the Internet has been blamed for this perceived rise in plagiarism. However, while there is no direct evidence that plagiarism is occurring more often since the introduction of the Web into classrooms, anecdotal evidence suggests this is so. But is the Internet to blame? According to Tribe & Rendell (2003), there were very few publications dealing with plagiarism prior to 1995, although by the year 2000 it seems to have become a very serious problem. Tribe et al. provide a number of explanations for this including student commitment, lifestyle, organizational skills, and confusion. What the Internet offers is the ability to find vast amounts of information very quickly with basic mastery of a search engine and some carefully selected key words. There are a number of strategies available to combat plagiarism. McLafferty and Faust (2004) suggest that the best strategy

to combat plagiarism is to prevent its occurrence and that when students are given appropriate instructions and/or particular types of assignments, plagiarism is minimized or even eliminated completely. Online plagiarism detection systems can be used as a deterrent beforehand, as well as for detecting plagiarism of online materials (Martin, 2005). In Darbyshire and Burgess (2005b), we outlined two vastly different techniques to combat plagiarism in two subjects offered by the School of Information Systems at Victoria University, Melbourne Australia. Each take a different approach by concentrating on the learning required of the assessment tasks and combating plagiarism at a different phase of the task. In one case students are encouraged to embrace the diversity offered by the Internet and shown how to find materials online and how to reference materials found on the Internet. In the other case, students were presented with an ungraded assignment, for which they could research information on the Internet, with a test on that assignment then being set. Each technique was effective in its own way in reducing the level of plagiarism by students. The eventual conclusion was that an evolution in procedures for assessment tasks is underway, and as academics we should view the Internet as a tool to be used rather than an adversary aiding and abetting plagiarists. Plagiarism occurs for many reasons, and there are many things we can do combat this. We are operating in a new landscape and traditional assessment methods are open to abuse in this environment.

Another problem to consider in the teaching environment is that of class attendance. Over the last few years, we seem to be experiencing a phenomenon of dropping attendance rates in classes. It is unclear whether this phenomenon is linked to increasing student employment rates or some other societal direction not yet evident, but the trend is a matter of concern. There is some doubt as to whether there is a direct correlation between attendance rates and success rates in university courses. However, students run the risk of significantly increasing their chances of poor performance and (or) failure by missing vital information due to non-attendance. Policy makers are putting pressure on universities to be more accountable in their management of public funds. One of the methods universities can adopt to do this is to more closely monitor student progression rates (Dancer & Fiebig, 2004). Trying to reverse this trend is difficult, and it seems that at best we need to closely monitor attendance and warn those at risk as early as possible. Alternatively, we can build attendance into the grading scheme, a practice which is common in online courses due to the nature of the paradigms utilized, and the importance of attendance in this environment. In Burgess and Darbyshire (2006), we examined two separate cases related to tertiary student attendance—one in a traditional face-to-face environment and one in an online environment. In the *face to face* example, student attendance in lectures had been quite poor for a few semesters, with students apparently relying on online lecture notes and copying noted from friends to survive. Unfortunately, this reflected in some poor student final results. The lecturer introduced a number of strategies, including weekly polls in lectures related to the lecture topic in which students were identified in their responses, and a regular reinforcement via e-mail to students with a poor attendance rate. Eventually, the student attendance rates and overall performance improved. In the second case, one of the advantages of the online paradigm became apparent. For students to succeed in a solely online course offering they are virtually **required** to "attend" class (online) on a regular basis. In this instance, the case showed that students that attended poorly tended to be identified early on and often did not remain enrolled in the course.

FUTURE TRENDS

The use of the Internet has become ubiquitous in tertiary education. It has now become commonplace as a tool for administration and learning both to supplement traditional techniques and in the form of new online paradigms. The lessons are still being learned as to how to best employ online technologies to improve the overall tertiary experience. New initiatives, such as randomly generated online tests delivered to large numbers of students in different locations, will continue to be trialled. It is also a time to look for future developments in the manner in which plagiarism, and the strategies used to deal with it, are developed. We anticipate this battle to continue for a while yet.

CONCLUSION

The majority of tertiary educators use the Internet to supplement existing modes of delivery. Importantly, the Internet is providing a number of *added value* supplemental benefits for subjects and courses delivered. There are two aspects to subject delivery to where *added value* benefits may be applied, and that is in the *administrative tasks* associated with a subject and the *educational tasks*. Most of the reported advantages gained from online supplementation of teaching relate to cost savings in terms of efficiency, flexibility, and/or convenience. These represent the traditional added value benefits of lower cost and faster access to goods in the commercial world. The measures of *money savings*, *time savings*, *improved quality*, and better *product information* can be used as categories to measure the benefits gained from the introduction of IT to supplement teaching.

A 2001 survey of tertiary educators revealed similar usage levels of administrative and educational features to aid tertiary education on the Internet. The administrative uses showed slightly more benefits for the institution than for students and vice-versa for educational uses. In both types of uses, their adoption seemed to be based upon how

difficult the feature was to set up as well as the added value benefits it provided. More recently, the use of the online paradigm for educational delivery has introduced new issues in relation plagiarism and online attendance, which need to be confronted by academics intending to head down the online or hybrid delivery paths.

REFERENCES

Bedore, G. L., Bedore, M. R., & Bedore, G. L. Jr., (1998). *Online education: The future is now* (2nd ed.). Socrates Distance Learning Technologies Group, Phoenix, AZ.

Boysen, P., & Van Gorp, M. J. (1997). ClassNet: Automated support of Web classes. The *25th ACM SIGUCCS Conference for University and College Computing Services*, Monterey, California USA.

Burgess, S., & Darbyshire, P. (2006). Postgraduate student attendance: Face-to-face versus online. *2006 International Resources Management Association International Conference*, CD ROM proceedings, Washington DC, May.

Byrnes, R., & Lo, B. (1996). *A computer-aided assignment management system: Improving the teaching-learning feedback cycle*. Retrieved February 12, 1999, from http://www.opennet.net.au/cmluga/byrnesw2.htm

Cheikes, B. A. (1995). GIA: An agent-based architecture for intelligent tutoring systems.In *Proceedings of the CIKM'95 Workshop on Intelligent Information Agents*.

Dancer, D., & Fiebig, D. (2004). Modelling students at risk. *Australian Economic Papers*, *43*(2), 158-173.

Darbyshire, P. (1999). Distributed Web based assignment submission and access. In *Proceedings of the International Resource Management Association, IRMA '99*, Hershey, USA.

Darbyshire, P., & Burgess, S. (2005a). Tertiary education and the Internet. In M. Khosrow-Pour (Ed), *Encyclopedia of information science and technology I-V* (Vol V, pp. 2788-2792). Hershey, PA: Idea Group Publishing.

Darbyshire, P., & Burgess, S. (2005b). New landscapes: Teaching to avoid plagiarism in the Web environment. *2005 International Resources Management Association International Conference* (pp. 224-227), CD ROM proceedings, San Diego, May.

Darbyshire, P., & Lowry, G. (2000). An overview of agent technology and its application to subject management. In *Proceedings International Resource Management Association, IRMA '2000*, Alaska, USA.

Darbyshire, P., & Wenn, A. (2000). A matter of necessity: Implementing Web-based subject administration. *Managing Web Enabled Technologies in Organizations* (pp. 162-190). Hershey, Idea Group Publishing.

Hassan, H. (1991). *The paperless classroom*. Paper presented at ASCILITE '91, University of Tasmania, Launceston, Australia.

IS World. (2001). *Mission and objectives*. Retrieved March 10, 2001, from http://www.isworld.org/isworld/mission.html

Kaynama, S. A., & Keesling, G. (2000). Development of a Web-based Internet marketing course. *Journal of Marketing Education*, *22*(2), 84-89, August.

Landon, B. (1998). *Online educational delivery applications: A Web tool for comparative analysis*. Centre for Curriculum, Transfer and Technology, Canada. Retrieved October 10, 1998, from http://www.ctt.bc.ca/landonline/

McLafferty, C. L., & Foust, K. (2004). Electronic plagiarism as a college instructor's nightmare--Prevention and detection. *Journal of Education for Business*, *79*(3), 186, Jan/Feb.

Martin, D. F. (2005). Plagiarism and technology: A tool for coping with plagiarism. *Journal of Education for Business*, *80*(3), 149-152, Jan/Feb.

O'Brien J. A. (1999). *Management information systems: Managing information technology in the internetworked enterprise* (4th ed.). Irwin MaGraw Hill

Patterson, D. A. (2006). A large-scale, asynchronous, Web-based MSW comprehensive examination administration: Outcomes and lessons learned. *Journal of Social Work Education*, *42*(3), 655-668.

Porter, M. E., & Millar, V. E. (1985). How information gives you competitive advantage. *Harvard Business Review*, *63*(4), 149-160, July-August.

Ritter, S., & Koedinger, K. R. (1995). *Towards lightweight tutoring agents*. Paper presented at the AI-ED 95--World Conference on Artificial Intelligence in Education, Washington, D.C.

Sandy, G., & Burgess, S. (1999). Adding value to consumer goods via marketing channels through the use of the Internet. *CollECTeR'99: 3rd Annual CollECTeR Conference on Electronic Commerce*, Wellington, New Zealand, November.

Scott Tillett, L. (2000). Educators begin to reach out—The net cuts costs, simplifies management, and could make distance learning a winner. *InternetWeek*, (835), 49-56, Manhasset, October 30.

Thompson, D. (1988, March 14). *WebFace overview and history*. Monash University. Retrieved January 2, 1999, from http://mugca.cc.monash.edu.au/~webface/history.html

Tribe, D., & Rendell, C. (2003). Meeting the plagiarism challenge. Complexity, creativity, and the curriculum. *5th Annual LILI Conferenc*e, Jan 2003, University of Warwick, UK. Retrieved November 14, 2004, from http://www.ukcle.ac.uk/lili/2003/papers/tribe.html,14/11/04

KEY TERMS

Administrative Tasks: The tasks that support educational tasks (such as enrolment, recording results, and so forth).

Class Attendance: In the case of "face to face" classes this refers to the physical presence of students in class. In the case of "online" classes this refers to student participation occurring with the timelines as imposed by the course lecturer.

Educational Tasks: Those tasks directly associated with the delivery of the educational component to students (e.g., lecturers, tutorials, assessment, and so forth).

Efficiency: From an IT viewpoint this usually relates to improvements within the business, so for a business it may mean IT systems that reduce costs or perform tasks more reliably or faster.

Internet Technologies: That group of technologies that allow users to access information and communication over the World Wide Web (Web browsers, ftp, e-mail, associated hardware, Internet service providers, and so forth).

Plagiarism: Using the words or ideas of others and presenting them as your own without acknowledgment.

Value: The amount a "buyer" is willing to "pay" for a product or service. A business can "add" value by being a low cost providing, providing a unique or differentiated product or service, or filling a niche market.

Internet Auctions

Kevin K.W. Ho
The Hong Kong University of Science and Technology, Hong Kong

INTRODUCTION

Year 2005 marks the 10th anniversary of eBay (http://www.ebay.com), the most successful online marketplace for Business-to-Consumer (B2C) and Consumer-to-Consumer (C2C) Internet auctions in this decade. As of December 2005, eBay has a major auction Web site and 26 sister Web sites operating all over the world and is enjoying a 37% quarter-to-quarter growth in revenue (eBay, 2005).

Before eBay existed, Internet auctions were mainly held in Internet forums and newsgroups (Lucking-Reiley, 1999, 2000). Nowadays, people can auction goods in cyberspace through Web sites such as eBay, Yahoo! auction (http://auctions.yahoo.com), and Amazon auction(http://s1.amazon.com). Designated Web sites for niche markets, such as antiques or electronic products are also established.

With the rapid growth of Internet auctions, economists and information systems (IS) researchers are using these Web sites to conduct field experiments and to collect transaction records to support their research more and more frequently. They are also interested in analyzing new features developed by these Web sites, such as Peer Review System and Buy-It-Now (BIN) auction, and bidders' behaviors, such as snipping. This article aims to provide a brief review of the Internet auction research conducted in the past few years and to explore the future research trends in this area.

BACKGROUND

Auction is one of the most popular business activities in the world. Each year, goods worth hundreds of billions of dollars are auctioned online and off-line. The types of goods auctioned range from very expensive items such as land and properties, government permits/licenses, and antiques to relatively minor items such as used stuffs, toys, and food (McAfee & McMillian, 1987; Lucking-Reiley, 2000). Recently, even job openings are being auctioned via the Internet (see, Job Dumping, http://www.jobdumping.de.)

Before the Internet boom, people already auctioned goods on the Internet via forums and newsgroups (Lucking-Reiley, 1999, 2000), as well as in private networks (for example, AUCNET, http://www.aucnet.com) (Lee, 1997, 1998). On the contrary to the argument that the "frictionless" nature of electronic commerce would reduce the profit of sellers (Bakos, 1997), AUCNET shows that electronic marketplaces can bring a win-win situation to both sellers and buyers. AUCNET has, on the one hand, increased sellers' profit, and on the other hand, provided an efficient market for auction participants with the help of information technology and other supporting features (Lee, 1998).

In response to customers' suggestions and to improve its usability, new features have been developed in Internet auction Web sites in the past few years. While some of them are brand new auction mechanisms, others are new features which aim to build up the trust between auction participants. Among them, Buy-It-Now auction and Peer Evaluation System are the most important ones.

Before we start our discussion, readers are reminded that the terminologies used by auction Web sites are sometimes different from auction literature. The most significant difference is the definition of a "Dutch auction." Dutch auction is a descending auction commonly used for the auction of perishable goods such as fish and cut flowers (McAfee & McMillian, 1987). However, auction Web sites use the same term to describe multiple-item English auction (Bapna, Goes, & Gupta, 2001). Hence, readers are reminded to check with individual Web sites on the terminologies if they are in doubt.

CURRENT TOPICS IN INTERNET AUCTION RESEARCH

In this section, we will review the latest research findings in Internet auctions, including research on the Peer Evaluation System and Buy-It-Now auction, as well as how economists and IS researchers employ Internet auction Web sites to conduct field experiments.

Peer Evaluation System

Peer Evaluation Profile is a reputation profile of auction participants based on his/her behaviors in his/her previous transactions on the Web site. After each transaction, the auction Web site will invite auction participants to evaluate each other. Nowadays, nearly all Internet auction Web sites have developed their own feedback system to allow their users to evaluate each other. Most of them use a 3-point scale (e.g., eBay and Yahoo!) and the others use a 5-point scale (e.g., Amazon) to record participants' performance, supplemented with notepads to provide further comments

Figure 1. A sample of user profile of Internet auction

Bidder Profile: Auctioner (112)				
Feedback Score: 112 **Positive Feedback: 93.8%**	Recent Rating:			
		Past Month	Past 6 Months	Past 12 Months
	Positive	20	120	140
	Neutral	2	5	5
	Negative	0	8	8

on the Web. Figure 1 is a hypothetical user profile of an Internet auction Web site. A lot of research studies have been conducted to examine whether the feedback system can bring benefit to auction participants.

One of the earlier studies on the Peer Evaluation System was conducted by Lee, Im, and Lee (2000). Through regression analysis using auction data of monitors and printers collected from eBay, they observe that the negative feedback score of a seller will have negative impact on the final auction price. This effect is more significant to used and refurbished products than to brand new items. Standifird (2001) used eBay auction data of 3Com Palm Pilot V to show that while both positive and negative feedback scores affect the final price, negative feedbacks are more influential than the positive ones. McDonald and Slawson's study (2002) suggests that the peer evaluation profile could be used to measure the expected performance of the seller whereas Ba and Pavlou (2002) suggest that the profile can induce and build up trust between bidders and sellers.

A comprehensive review has been conducted by Dellarocas (2003). He suggests that Peer Evaluation System provides an opportunity for sellers and bidders to build up their word-of-mouth profiles in electronic marketplaces. Yoo, Ho, and Tam (2006) propose that as there is a large number of positive feedback when compared with the number of negative feedbacks, the per-unit effect of negative feedback becomes more influential when bidders evaluate the reputation of sellers. This may explain why some Web sites highlight the net feedback (i.e., the number of positive feedback minus negative feedback) of auction participants as this can minimize the adverse effect of negative feedbacks.

In conclusion, as shown in Table 1, the Peer Evaluation Profile can induce and build up trusts between buyers and sellers. For bidders, they can use the profile to evaluate the trustworthiness of sellers before they decide whether they should join the bidding competitions. On the other hand, sellers can use this system to boost up the auction price by building up their word-of-mouth profile.

Buy-It-Now and Buy-Price Auctions

BIN auction and Buy-Price (BP) auction are two similar variations of English auction developed by eBay and Yahoo! respectively, in the late 1990s. In BIN auction, the seller posts a BIN price and indicates that he/she is willing to close the auction at this preset price. This BIN price is valid until a bidder submits a bid which is higher than the reserve price. After that, it will disappear and the auction will convert to an English auction. BP auction operates in a similar way except it will not convert to English auction in any circumstances. This counter-intuitive arrangement, that is, capping the maximum profit of an auction by sellers themselves, creates an interesting research topic. Table 2 summarizes the result of several representative studies on BIN and BP auctions.

Most of the analytical studies are developed based on the assumption that bidders are risk averse. For example, Budish and Takeyama (2001) show that with risk averse bidders, BP auction can generate more revenue than First-Price Sealed Bid and Dutch auctions. Mathews (2004), on the other hand, shows that BIN auction is useful only when either bidders, sellers, or both of them are time impatient. Onur and Tomak (2003) also analyze whether bidders would prefer to use the BIN option or to snipe in an Internet auction. Their model, supported with empirical data, shows that the low value bidders prefer to snipe, and the high value bidders prefer to exercise the BIN option in a BIN auction.

To compare the difference between BIN and BP auctions, Reynolds and Wooders (2003) develop an analytical model to examine their differences. Similar to Budish and Takeyama's (2001) model, their model also predicts that both BIN and BP auctions will not bring any additional benefit to sellers if bidders are risk neutral. However, if bidders are risk averse, BIN and BP auctions will raise the revenue when compared with English auction. Furthermore, BP auction will have a better performance, that is, generate more revenue, than BIN auction. More recently, Hidvegi et al. (2006) have further extended the discussion and show that the social welfare and utilities of all participants can be improved if BP is properly set.

Empirical and experimental studies have also been conducted to examine the property of these auctions. Standifird et al. (2005) conducted a field experiment on eBay using American Eagle silver dollars. They observe that some eBay users do not exercise the BIN options even when they are set at a price lower than the market price. Hence,

Table 1. Result of studies on peer evaluation profile

Authors	Web site studied	Item studied	Result
Lee et al.(2000)	eBay	Printer Monitor	Negative feedback will decrease the final auction price.
Standifird (2001)	eBay	Palm Pilot V	Positive feedback will increase the final auction price. Negative feedback will decrease the final auction price. Negative feedback has a stronger influence than positive feedback.
McDonald and Slawson (2002)	eBay	Doll	The higher the net feedback score, the higher the final auction price.
Ba and Pavlou (2002)	eBay	Digital products, CD and software	Positive feedback has a stronger positive impact on the final auction price.
Yoo et al.(2006)	eBay Yahoo!	Digital Camera	Positive feedback will increase the final auction price. Negative feedback will decrease the final auction price. Negative feedback has a stronger influence than positive feedback.

Table 2. Result of studies on Buy-It-Now and Buy-Price auctions

Authors	Auction studied	Result
Budish and Takeyama (2001)	BP	If bidders are risk averse, revenue of BP > D and FP.
Onur and Tomak (2003)	BIN	Low value bidders prefer to snipe and high value bidders prefer to use BIN.
Reynolds and Wooders (2003)	BIN and BP	If bidders are risk averse, revenue of BP > BIN > E.
Mathews (2004)	BIN	BIN is useful when either bidders, sellers, or both of them are impatient.
Standifird, Roelofs, and Durham (2004-5)	BIN	Bidders are interested in the hedonic effect of Internet auctions. Hence, they will not exercise BIN option even if it is set below the market price.
Yoo et al. (2006)	BIN and BP	The revenue of BP > BIN > E. Besides, even if BP and BIN are not exercised, their performances are still better than E.
Hidvegi, Wang, and Whinston (2006)	BP	The revenue of BP > E. Social welfare of all participants can be improved if BP is properly set.

Key: BIN: Buy-It-Now auction BP: Buy-Price auction D: Dutch auction E: English auction FP: First-Price Sealed Bid auction

they propose that bidders may only be interested in taking part in the bidding competition. Another empirical study conducted by Yoo et al. (2006) using digital camera data obtained from eBay and Yahoo! auction Web sites shows that BIN and BP posted by sellers can provide additional information to bidders. This effect can probably help to improve the auction performance even when the BIN or BP option is not being exercised.

Field Experiments on the Internet

Experimental economics also benefits from the boom of the Internet. Before the Internet Age, most economic experiments were conducted in laboratories using student subjects. It is always arguable that whether the results obtained from student subjects can be generalized to the real world setting. Thus Internet auction provides an opportunity for researchers to conduct field experiments more easily. One of the early experiments using the Internet was conducted by Lucking-Reiley (1999) in an Internet newsgroup. In his study, he auctioned game cards in the newsgroup and examined whether the *revenue equivalence theory* proposed by Vickrey (1961) is held in cyberspace. He observed that Dutch auction outperforms first-price sealed bid auction and he suggests that the Dutch clock may have an anchoring effect to bidders' valuation on the product.

Hossain and Morgan (2006) also conducted a field experiment in eBay. They auctioned matched pairs of CDs and Xbox games with different combinations of reservation price and shipping and handling costs. Even though the opening offers, that is, reservation price plus shipping and handling costs, are the same for both settings, they observe that a combination of high shipping and handling costs and low opening price attracts more bidders than a combination of low shipping and handling costs and high opening price. They propose that mental accounting maybe one of the reasons to explain this phenomenon. Besides, Standifird et al. (2004-5) conducted another field experiment to examine the impact of BIN option on the auction price.

Other Studies on Internet Auctions

Other studies on Internet auctions include the study on the reputation of auction Web sites. Standifird (2002) conducted

Table 3. Result of experimental studies on Internet auctions

Authors	Web site studied	Item studied	Result
Lucking-Reiley (1999)	Internet newsgroup	Game Cards	Dutch auction outperforms First-price sealed bid auction.
Standifird et al. (2004-5)	eBay	Silver Coins	Bidders are interested in the hedonic effect of Internet auctions.
Hossain and Morgan (2006)	eBay	CDs Xbox Games	A combination of high shipping and handling cost and a low opening price will attract more bidders than a combination of low shipping and handling cost and a high opening price.

an empirical study to examine whether the brand name of the Web site will have any effect on the auction final price. He shows that using the auction of Iomega Zip Disks, eBay generates more revenue than CNET (a niche auction Web site for digital products) and Amazon auction as well. Moreover, the revenue generated from CNET is higher than Amazon. Onur and Tomak (2003) and Roth and Ockenfel (2002) both studied the sniping behavior of bidders, that is, last minute bidding for an auction with hard close. While the former study proposes that the type of bidders will determine the selection between sniping and using BIN option, the latter study concentrates on the effect of the closing method of the auctions on the sniping behavior.

FUTURE TRENDS

From our review of the literature we notice that Internet auction study is a very interesting research topic in economic and IS. It is because Internet auction research is one of the few research areas which can allow researchers to construct analytical models which are ready for validation using both empirical and experimental methods. These analytical models developed can be implemented in the real business world, and hence, practitioners can also benefit from the result of this area of research.

Concerning future research areas, it is expected that more research will be conducted on the development of new auction mechanisms which can either improve the trust between bidders and sellers or resolve the information asymmetry problem. Trust building is important as this can reduce the perceived risk associated in Internet auctions. This can in turn increase the number of participants in the electronic marketplace and increase the revenue. New mechanisms targeted to resolve information asymmetry will also help to reduce the perceived risk and improve the auction performance. The building of these new mechanisms will definitely bring benefits to the sellers (by increasing the revenue) and bidders (by reducing the perceived risk in participating in Internet auctions).

Moreover, it is also expected that more field experiments will be conducted using the Internet as it can provide a test bed for auction research. It can also provide a platform for practitioners to discover the pricing of new products.

CONCLUSION

As one of the most successful e-businesses in the world, eBay tells us that Internet auction can be one of the "killer" applications in the real world. Apart from financial reward for stockholders as well as sellers using their service, it also provides opportunities for researchers to conduct new economic and IS research. It is foreseeable that more and more e-business research will be conducted in this area in the near future.

REFERENCES

Ba, S., & Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly, 26(3)*, 243-268.

Bakos, J. Y. (1997). Reducing buyer search costs: Implications for electronic marketplaces. *Management Science, 43(12)*, 1676-1692.

Bapna, R., Goes, P., & Gupta, A. (2001). Insights and analyses of online auctions. *Communications of the ACM, 44(11)*, 42-50.

Budish, E. B., & Takeyama, L. N. (2001). Buy prices in online auctions: Irrationality on the Internet? *Economics Letters, 72(3)*, 325-333.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49(10)*, 1407-1424.

eBay. (2005, October 19). *eBay Inc. announces third quarter 2005 financial results.* Retrieved December 3, 2005, from

http://investor.ebay.com/news/Q305/EBAY1019-148259.pdf

Hossain, T., & Morgan, J. (2006). … Plus shipping and handling: Revenue (non)equivalence in field experiments on eBay. *Advances in Economic Analysis & Policy, 6(2),* Article 3.

Hidvegi, Z., Wang, W., & Whinston, A. B. (2006). Buy-price English auction. *Journal of Economic Theory, 129*(1), 31-56.

Lee, H. G. (1997). AUCNET: Electronic intermediary for used-car transactions. *Electronic Markets, 7(4),* 24-28.

Lee, H. G. (1998). Do electronic marketplaces lower the price of goods? *Communications of the ACM, 41(1),* 73-80.

Lee, Z., Im, I., & Lee, S. (2000). The effect of negative feedback on prices in Internet auction markets. In W.J. Orlikowski, S. Ang, P. Weill, H. C. Krcmar, & J. I. DeGross (Eds.), *Proceedings of the 21st International Conference on Information Systems,* Brisbane, Australia (pp. 286-287).

Lucking-Reiley, D. (1999). Using field experiments to test equivalence between auction formats: Magic on the Internet. *American Economic Review, 89(5),* 1063-1080.

Lucking-Reiley, D. (2000). Auctions on the Internet: What's being auctioned, and how? *Journal of Industrial Economics, 48(3),* 227-252.

Mathews, T. (2004). The impact of discounting on an auction with a buyout option: A theoretical analysis motivated by eBay's Buy-It-Now feature. *Journal of Economics, 81*(1), 25-52.

McAfee, R.P., & McMillian, J. (1987). Auctions and bidding. *Journal of Economic Literature, 25(2),* 699-738.

McDonald, C. G., & Slawson, V. C., Jr. (2002). Reputation in an Internet auction market. *Economic Inquiry, 40*(4), 633-650.

Onur, I., & Tomak, K. (2003). Buy-it-now or snipe on eBay. In L. Applegate, R. Galliers, & J. I. DeGross (Eds.), *Proceedings of the 24th International Conference on Information Systems* Seattle, WA (pp. 841-846).

Reynolds, S. S., & Wooders, J. (2003). *Auctions with a buy price.* [Working Paper]. Tucson: University of Arizona.

Roth, A. E., & Ockenfel, A. (2002). Last-minute bidding and the rules of ending second-price auctions: Evidence from eBay and Amazon auctions on the Internet. *American Economic Review, 92(4),* 1093-1103.

Standifird, S. S. (2001). Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management, 27(3),* 279-295.

Standifird, S. S. (2002). Online auctions and the importance of reputation type. *Electronic Markets,* 12*(1),* 58-62.

Standifird, S. S., Roelofs, M. R., & Durham, Y. (2005). The impact of eBay's buy-it-now function on bidder behavior. *International Journal of Electronic Commerce, 9(2),* 167-176.

Vickrey, W. (1961). Counterspeculation, auctions, and competitive sealed tenders. *Journal of Finance, 16*(1), 8-37.

Yoo, B., Ho, K., & Tam, K. Y. (2006). The impact of information in electronic auctions: An analysis of buy-it-now auction. In R.H. Sprague, Jr. (Ed.) *Proceedings of the 39th Hawaii International Conference on System Sciences,* Kauai, HI.

KEY TERMS

Auction With Hard Close: It is an Internet auction with a fixed end time. This closing method provides an opportunity for bidders to snipe in an auction.

Auction With Soft Close: It is an Internet auction with an extendable end time. The end time of auction will be extended if a new bid submitted within the last few minutes of the auction.

Buy-it-Now (BIN) Auction: It is a variation of English auction developed by eBay. The seller posts a BIN price and agrees to close the auction at this preset price. The BIN option will disappear and the auction will convert to an English auction when a bidder submits a bid which is higher than the reserve price of the auction.

Buy Price (BP) Auction: It is a variation of English auction developed by Yahoo! auction. Its setting is similar to BIN auction. Unlike BIN auction, BP auction will not covert to English auction in any circumstances.

Dutch Auction: It is a common descending auction for perishable goods. However, some Internet auction Web sites used this term to describe a multiple-item English auction.

Peer Evaluation Profile: It is a reputation profile of the auction participants based on her/his behaviors in her/his previous transactions on the Web site.

Snipping: It is a jargon for "last-minute bidding." It is only possible if the auction Web site is having a fixed end time (or "hard close").

Internet Diffusion in the Hospitality Industry

Luiz Augusto Machado Mendes-Filho
Faculdade Natalense para o Desenvolvimento do Rio Grande do Norte, Brazil

Anatália Saraiva Martins Ramos
Universidade Federal do Rio Grande do Norte, Brazil

INTRODUCTION

Tourism is the most important industry in the world in terms of the numbers of employees and its effect on the social and economical development of a region or country. Holjevac (2003) believes that, by the year 2050, tourism will by far be the largest industry worldwide, with 2 billion tourists and US$24 billion in domestic and international receipts. Moreover, the major tourist destinations will be India, China, Indonesia, and Brazil.

The use of information technologies for basic functions—conferences, business meetings in distant places, training, designed routes and airlines, reservations and tickets purchased through computer systems, tourist shops, restaurants—is becoming usual in tourism. All these services have led tourist companies to adopt more updated methods in order to increase competition. Consumers, who are already becoming familiar with new technologies, demand more flexible, interactive, and specialized products and services, bringing new management techniques from the intelligent use of IT used to accomplish tour company business processes (Buhalis, 2000).

The hotels depend progressively on the resources of new information technology to follow and update the tools which allow an efficient development of activities in each section of the company, leading to better results for its management (Mendes-Filho & Ramos, 2003a). To Phillips and Moutinho (1998), information technology (IT) is one of the critical factors of success in the hotel industry.

According to studies and data, the use of technological tools will allow a bigger competitiveness for hotels (Cline, 1999). Technology will be the catalyst of change, a source of growing connectivity and one of the most important factors in distinguishing success among hotel companies. Few issues have greater importance to the business of hospitality than the technological decisions that will be made in the coming years (Buhalis, 2000; Mendes-Filho & Ramos, 2004; Olsen & Connoly, 2000).

The hotel industry is one of the most important kinds of Web commerce. The data shows that all major companies linked to the tourism industry (hotels, agencies, air companies, and rentals) possess some kind of e-commerce activity through the Web (O'Connor, 1999; Scottish Executive, 2000; Werthner & Klein, 1999).

BACKGROUND

Decades ago, before the use of the computer in the accommodation sector, those charged with making reservations performed their service by checking availability tables exposed on the wall or in large, updated, hand-written lists (O'Connor, 1999). The hotels received innumerous telephone calls, letters, and telex from potential clients, sometimes larger than that of the hotel's reception, and worked to select correspondence, type letters, send telegrams, and deal with other demands. The delays were frequent, the cost of correspondence writing went sky-high, and specialized typists were in demand (World Tourism Organization, 2003).

A way found by the American hotel chains to streamline the reservation services was to centralize this function in a main office, serving the consumer better and offering a valuable service to the hotels belonging to that chain. O'Connor (1999) states that the reservation process in hotels in the USA was made even easier with the introduction of free telephone services in the mid-'60s, which permitted potential clients to perform an only call to obtain information or make reservations in any of the hotels of that chain in the world.

Although the reservation area became faster and more efficient, two large costs remained, those of telecommunications (free telephone service payment) and labor costs of the reservation agents necessary to answer the phones. With the increase in trips during the 1960s, the airline companies developed the computer reservation system (CRS), which pressured the hotel sector to develop its own (O'Connor, 1999).

The main focus in hotel and restaurant management has always been the maximization of consumer satisfaction and personalized attention. The use if IT has, at times, seemed incompatible with this objective, and the hotel sector has, in a way, delayed the application of IT in its operations. The technology has been viewed as a hindrance to personalized service because it creates an impersonal, mechanical, and cold environment with the clients.

However, the change of this belief is being changed within the hotel sector. Nowadays, according to Sheldon (1997), the establishments are noticing that IT can bring efficiency to the hotel, besides reducing costs and offering a great potential to increase the levels of personalized service to the clients.

In a survey performed by financial managers of American hotels, all stated that IT increased the hotel's productivity (David, Grabski, & Kasavana, 1996). The motives used to justify this statement were the following: Technology reduces the administration costs, decreases the amount of paperwork between sectors, minimizes operational errors, increases the earnings/profits of the hotel, and makes the reservation management more efficient. This same survey proved that IT is not only used to increase the hotel's productivity but also to improve the service, as well as to offer new services to the guests.

According to Namasivayam, Enz, and Siguaw (2000), IT can also be employed to reach business objectives. American and European hotel executives have plans to use the technology to reduce operational costs, increase sales, improve the service to the client, increase employees' productivity, and increase hotel earnings.

In research performed by Van Hoof et al. (1995), 550 American hotel managers answered questions about their perceptions of the use and implementation of technology in their establishments. Those responding identified the front office (reception and reservations) of the hotel as the sector that can benefit the most from the use of the technology, followed by sales and marketing, accounting, and the food and drink sector. According to Van Hoof et al. , having a quality service is a challenge to the hotel industry, which has high employee turnover indexes, employee salary increase, and low age of the most qualified people. Consequently, technological applications have been developed in hotels to increase this quality in the services and improve the interaction of the hotel employees with the guests.

IMPACTS OF THE INTERNET IN HOTEL INDUSTRY

During the '80s and '90s several authors from companies and universities had already foreseen that as new technologies were increasingly used, hotels could benefit from that in a great range of situations, for example: better qualified services for customers, increased sales and profits, efficiency in operation and integration of hotel sectors, rapid communication, and cost reduction (Laudon & Laudon, 1999).

Technological applications enable information and knowledge to bring a competitive advantage to the future profile of the hotel. The "Information Age" idea is that the most modern companies will build their success upon the amount of knowledge they have about their clients as well as information on their products and services and how they will make a profit in this new environment (Olsen & Connoly, 2000).

With the Internet being used as a means of communication, this brings several advantages or benefits compared to other vehicles. Flecha and Damiani (2000) state that when it comes to the tourist area, the main points are: the new relationship between consumers and companies, marketing for actively participating consumers, the importance of detailed information, self-service application, credibility, and agility of communication.

The use of the Internet and World Wide Web is spreading quickly in most consumer access areas to travel database developments. There are hundreds of thousands of suppliers' homepages, associations, e-news, newsgroups, and chats for the travel and tourism community. This group of technologies provides many opportunities for the industry to interact with its consumers and suppliers. It is also possible that, through information technology, products and services may be personalized according to the tourist's needs and thus may become a differential feature for those who adopt it (Buhalis, 2000; Sheldon, 1997).

The purchase of products and services through the Internet is revolutionizing the world of business and people's lives as well. For some clients it is more comfortable to book an e-ticket through the company home page rather than going to the travel agency (Franco, 2001).

As the Internet began and grew, the use of such technologies at home or work and also the new opportunities that arose from the lower costs in telecommunication equipment made it possible for suppliers to distribute information to their clients and process reservations directly with the clients (O'Connor, 1999).

According to Jeong and Lambert (2001), the Internet has already modified the competitive strategy of some hotels. It is through the Internet that the client can have a "self-understanding" in a service that is being offered to him in a more efficient way. In hotels, check-in processes can already be totally automatic, from the Internet booking until the moment the client takes his keys in an automatic dispenser. The result is that clients can become more informed and willing to have quick answers from the orders online. Though many experts and businessmen agree that the Internet is probably the most important technological tool, it is still relatively new and misused in the hotel industry (Van Hoof & Verbeeten, 1997).

Several authors have identified impediments to the growth of the Internet in the industry and, hence, have reservations about the willingness of hotel operators to adopt the Internet wholeheartedly (Wei et al., 2001). These problems include user-friendliness, the quality and accuracy of information obtained from the Web, and the issue of data security (Wei et al.). Here are other difficulties found by Lituchy and Rail (2000) in their research: problems in updating new

information in hotel Web pages, annoyance expressed at inaccessible Webmasters by managers, hard to find hotel Web sites, lack of knowledge on the employees' part of how to use the technology, and the impersonality of the medium. Because of these, consumers have been slow to adopt the Internet as a means of making hotel reservations. Only 4% of reservations are made online (Maselli, 2002).

Namasivayam et al. (2000) summarize that almost 60% of the hotels in their study had few technologies. To Feiertag (2000), a lack of proper training, high turnover rates, and limited financial resources were major barriers to the successful use and implementation of new technologies.

In addition to this, many hotels still believe that conventional means of advertising, such as radio, television, and printed material, are the most effective way of promoting their properties. The share of reservations received through the Internet remains minute compared to reservations received through conventional means, such as phone, fax, or mail (Van Hoof & Combrink, 1998). However, these problems are diminishing with the increase of the number of Internet users. And if customers become accustomed to browsing for rooms and making reservations through the Internet, more and more properties will be forced to get on the Internet as well (Mendes-Filho & Ramos, 2002).

In a specific way, the Internet provides an expansion of hotel services, changing this industry and giving new opportunities to clients, thus being a new channel to be developed. Besides online reservation services, the Internet allows hotels to sell their services and charge them electronically as well as offer new products through the World Wide Web (Blank, 2000; Laudon & Laudon, 1999).

Through the Web the customer can check hotel location, compare rates, see pictures and watch videos, get information about tourist destinations and other facilities, check room availability, and book and confirm reservations for the amount of time he wants to stay, among other services. Hence, the interactivity of the Web provides an ideal medium for distributing accommodations online, consolidating itself as a very adequate platform for bringing information and services to the client in a very straightforward, efficient, and quick way (Flecha & Damiani, 2000; Hotels, 2001).

Marriot, Hospitality Services of America, and Hilton are some of the hospitality industry's members that have successfully used marketing on the Internet to reach new markets, track customers, take online reservations, and offer information about their products and services (Lituchy & Rail, 2000).

FUTURE TRENDS

A hotel chain's success has always depended on excellent services performed by operation, marketing, and human resources sectors. For Withiam (2000) in the 21st century an essential factor will be technological support, making it possible for computers to process information of reservation systems, affinity programs, and marketing data banks.

Improvements in integration, centralized data banks, and the use of Web sites are some of the tendencies in the development of software for hotels (Adams, 2001). Therefore, the connection of a hotel system to the Internet will integrate information of the internal system with the Web site, and this will make a lot of information available to managers. The new systems are being developed with this integration with the Web site.

With the increasing demand of information in the tourist sector, the importance of IT use in this industry will only tend to increase in the future. Therefore, the tourist businesses must understand, incorporate, and use IT strategically to serve the target markets, improve their efficiency, maximize profitability, perfect services, and maintain the profitability in the long term (Buhalis, 2000).

To Olsen and Connoly (2000), the volume of information about the guests collected electronically is too large for the directors to be able to manage without the help of technology. Data warehousing and data mining are technologies that are gaining popularity to analyze information about clients. These technologies may be used to help hotel keepers construct good relationships with their guests, increasing their loyalty to them.

Using the Internet in the hotel industry has good prospects of growth, though in many hotels the use of such technology is still moving slowly. On the other hand, there are some hotels using and steadily setting the trend. It will be an important and strategic issue for businessmen to stimulate such Internet use policies inside the tourist trade so that they become wired to this new reality and can work on even terms with their competitors.

CONCLUSION

The Internet has decreased expenses and enabled small businesses to conduct international business from home (Lituchy & Rail, 2000). Small inns and bed and breakfasts are advertising on the Web and are therefore becoming a presence in the global market. So, they face the likelihood of serving foreign customers that may have different hospitality expectations.

Despite the fact that Internet use is very common, the proportion of reservations received from the Internet is small. The public could still be concerned about issues of

security for financial transactions or could not be satisfied with its inability to synchronize inquiries. A low reservation rate from the Internet may also be partly attributed to the lack of certain relevant information, such as room availability and virtual tours of the property not commonly included in the homepage (Wei et al., 2001).

In general, the Internet does enable tourist companies to increase their competitiveness. IT can improve the efficiency of suppliers and provide tools for the development and delivery of different tourist products (Mendes-Filho & Ramos, 2003b). One of the benefits reached is the reduction of the dependence on the middlemen in the distribution of tourist products. Hotel owners should invest more money in technology besides concentrating more time and attention to subjects in that area. IT affects all aspects of a hotel chain's value, going far beyond sectors and departments. As technology will be intrinsically linked to hotel business, its executives will insert technology in all their strategic decisions for the facility. That implies all the employees (including managers and directors) need to have enough knowledge to extract the potential the technology provides.

REFERENCES

Adams, B. (2001). The PMS picture. *Hotel and Motel Management*, *216*(2), 36-37.

Blank, D. (2000). Internet will shape revenue-management role. *Hotel and Motel Management*, *215*(11), 54-55.

Buhalis, D. (2000). Marketing the competitive destination of the future. *Tourism Management*, *21*(1), 97-116.

Cline, R. (1999). Hospitality 2000—The technology. *Lodging Hospitality*, *55*(7), 18-26.

David, J. S., Grabski, S., & Kasavana, M. (1996). The productivity paradox of hotel-industry technology. *Cornell Hotel and Restaurant Administration Quarterly*, *37*(2), 64-70.

Feiertag, H. (2000). Technology can help salespeople, but it can't replace them. *Hotel and Motel Management*, *215*(14), 22.

Flecha, A. C., & Damiani, W. B. (2000). Avanços da tecnologia da informação: Resultados comparados de sites da indústria hoteleria. *Proceedings of the 20th Production Engineering National Meeting*, (Vol. 1, pp. 153-161).

Franco, C. F., Jr. (2001). *E-business: Tecnologia de informação e negócios na Internet*. São Paulo, Brazil: Atlas.

Holjevac, I. A. (2003). A vision of tourism and the hotel industry in the 21st century. *Hospitality Management*, *22*, 129-134.

Hotels. (2001). Hotels' 2001 worldwide technology survey—Part 1. *Hotels*, *35*(2), 75-85.

Jeong, M., & Lambert, C. (2001). Adaptation of an information quality framework to measure customers' behavioral intentions to use lodging Web sites. *International Journal of Hospitality Management*, *20*(2), 129-146.

Laudon, K. C., & Laudon, J. P. (1999). *Sistemas de informação com Internet* (4th ed.). Rio de Janeiro, Brazil: LTC.

Lituchy, T. R., & Rail, A. (2000). Bed and breakfasts, small inns, and the Internet: The impact of technology on the globalization of small businesses. *Journal of International Marketing*, *8*(2), 86-97.

Maselli, J. (2002, April 22). Hotels take to the Web to battle discounters. *InformationWeek*.

Mendes-Filho, L. A. M., & Ramos, A. S. M. (2002). The Internet adoption in the hotel industry: A multiple cases study in Brazilian hotels. *Proceedings of the 13th Information Resources Management Association International Conference*, (Vol. 1, pp. 209-211).

Mendes-Filho, L. A. M., & Ramos, A. S. M. (2003a). The benefits and difficulties of the Internet use in hotels: The effect of hotel rate on the managers' perception. *Proceedings of the 14th Information Resources Management Association International Conference*, (Vol. 1, pp. 328-330).

Mendes-Filho, L. A. M., & Ramos, A. S. M. (2003b). The perception of managers on the impacts of Internet in Brazilian hotels: An exploratory study. In S. Kamel (Ed.), *Managing globally with information technology* (pp. 244-259). Hershey, PA: Idea Group Publishing.

Mendes-Filho, L. A. M., & Ramos, A. S. M. (2004). The benefits and difficulties of Internet use in hotels and its effects according to the facilities' rank, property size, manager's age and experience. In C. Deans (Ed.), *E-commerce and m-commerce technologies* (pp. 217-239). Hershey, PA: Idea Group Publishing.

Namasivayam, K., Enz, C. A., & Siguaw, J. A. (2000). How wired are we? The selection and use of new technology in U.S. hotels. *Cornell Hotel and Restaurant Administration Quarterly*, *41*(6), 40-48.

O'Connor, P. (1999). *Electronic information distribution in tourism and hospitality*. Wallingford, UK: CAB International.

Olsen, M. D., & Connoly, D. J. (2000). Experience-based travel. *Cornell Hotel and Restaurant Administration Quarterly*, *41*(1), 30-40.

Phillips, P. A., & Moutinho, L. (1998). *Strategic planning systems in hospitality and tourism*.Wallingford, UK: CAB International.

Scottish Executive. (2000). A new strategy for Scottish tourism. Edinburg. Retrieved July 21, 2002, from *http://www.scotland.gov.uk/library2/doc11/sfst.pdf*

Sheldon, P. (1997). *Tourism information technology*. Wall-

ingford, UK: CAB International.

Van Hoof, H. B., & Combrink, T. E. (1998). U.S. lodging managers and the Internet: Perceptions from the industry. *Cornell Hotel and Restaurant Administration Quarterly*, *39*(2), 46-54.

Van Hoof, H. B., & Verbeeten, M. J. (1997). Vendors receive mixed reviews. *Hotel and Motel Management*, *212*(11), 42.

Van Hoof, H. B., Collins, G. R., Combrink, T. E. & Verbeeten, M. J.(1995). Technology needs and perceptions: An assessment of the U.S.lodging industry. *Cornell Hotel and Restaurant Administration Quarterly,36*(5), 64-69.

Wei, S., Ruys, H. F., Van Hoof, H. B. & Combrink, T. E. (2001). Uses of the Internet in the global hotel industry. *Journal of Business Research,54*, 235-241.

Werthner, H., & Klein, S. (1999). *Information technology and tourism*: *A challenging relationship*. New York: Springer-Verlag.

Withiam, G. (2000). Carlson's "24K" consumer-centric computer. *Cornell Hotel and Restaurant Administration Quarterly, 41*(3), 13.

World Tourism Organization (2003). *E-Business para turismo: Guia prático para destinos e empresas turísticas*. Porto Alegre, Brazil: Bookman.

KEY TERMS

Bed and Breakfast: An establishment (as an inn) offering lodging and breakfast.

Computer Reservation System: A computer system that manages the distribution of the tourist products to transportation, lodging, and entertainment companies.

Data Mining: The process of analyzing data to determine relationships undiscovered by previous analyses.

Data Warehouse: A data warehouse is a central repository for all or significant parts of the data that an enterprise's various business systems collect.

E-Ticket: An e-ticket (electronic ticket) is a paperless electronic document used for ticketing passengers, particularly in the commercial airline industry. Virtually all major airlines now use this method of ticketing.

Front Office: The department of the hotel that deals directly with clients. Normally, it involves the reception and the reservation sector of the hotel.

Middleman: A dealer or agent intermediate between the producer of goods and the consumer or retailer.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1635-1639,

Internet Work/Play Balance

Pruthikrai Mahatanankoon
Illinois State University, USA

INTRODUCTION

Productivity gain can be achieved through utilitarian use of the Internet. Networked organizations foster intra- and inter-organizational communications, which amplify team collaborations, information sharing, and relationship building. The Internet also provides linkage to external global information sources, allowing organizations to analyze market trends, predict competitors' movements, and search for competitive advantages. However, Internet usage in the workplace is also a double-edged sword that can bring liabilities to modern workplaces. Employees can utilize their Internet connectivity and e-mail accounts for a variety of purposes. Publicized cases of Internet abuse in the workplace (i.e., pornography, employee harassment, information leakage, software piracy, etc.) have generated different ethical and legal concerns for many organizations.

To prevent such occurrences, practitioners utilize several strategies to deter Internet abuses (e.g., training of proper Internet usage, communicating Internet usage policy, installing Internet monitoring, and filtering software, etc.). These strategies have been effective against such behaviors, but they often decrease employees' job satisfaction and motivation. Understanding the underlying determinants of workplace Internet usage can bring balance to organizational work and play, and allow practitioners to apply the most effective Internet usage policies to increase job satisfaction.

A feasible balance between proper behavioral controls and employee motivation is attainable through the equilibrium of organizational and individual psycho-socio-technical factors. The ultimate goal of this balance is to maintain and improve employee satisfaction and organizational well-being. To identify the appropriate balance, this article examines different perceptions of Internet usage activities and suggests three Internet management strategies.

ASPECTS OF INTERNET USAGE

Any Internet behavior can be classified as a productive, personal, or pathological Internet usage. This article defines these perceptions as the 3Ps of Internet usage. These perceptions have a direct influence on social/technological/psychological situations of individuals and organizations. Studies of Internet usage behaviors are interdisciplinary in nature; therefore, the stimuli behind each perception still require extensive research that involves a different set of determinants and outcomes.

Productive Internet usage provides a utilitarian view of Internet technologies. Research on information technology adoption, information systems success factors, and technology-task fit all contribute to increase productive Internet use. Studies in this area include creating positive employees' attitude, establishing organizational/social usage norms, and lowering the psychological barriers of Internet usage. Ideally, productive Internet usage occurs at work. With today's networked organizations, productive work also can be performed at various virtual offices.

Personal Internet usage involves a recreational use of the Internet. This type of usage occurs privately at home and occasionally at work. Anandarajan and Simmers (2002) define these behaviors as voluntary online Web behaviors during working time in which employees use any of the organization's Internet resources for activities outside current customary job/work requirements. These activities include any personal use of the Internet at work such as searching for news and entertainment information, conducting electronic commerce, booking a vacation, and using personal e-mail (Mahatanankoon, Anandarajan, & Igbaria, 2004). Most people who engage in such behaviors are aware of their environment, social norms, and organizational policies. The consequences of personal Internet usage also have contradictory ramifications. Some studies suggest that these behaviors lead to productive work life (Stanton, 2002) and organizations should encourage a balance between and work and play (Belanger & Van Slyke, 2002; Oravec, 2002). Other studies find that these behaviors can lead to cyberloafing (Lim, 2002). Table 1 summarizes the benefits and potential risks of personal Internet usage.

The norms of Internet usage in the workplace are "co-evolving" (Kraut & Kiesler, 2003) and it is difficult for organizations to associate the relationships between personal Internet activities with any individual and organizational outcomes. However, excessive personal Internet usage is counterproductive. Internet abuse is a general term that often refers to any wrongful or improper use of the Internet in the workplace. Behaviors related to Internet abuse often are more severe in nature such as viewing pornography, harassing other employees, downloading illegal software, excessive gaming, etc. Extreme cases of Internet abuse habitually result in many negative, uncontrollable psychological consequences.

Table 1. Benefits and potential risks of personal Internet usage

Benefits	Potential Risks
Information sharing	Cyberloafing
Team collaboration	Productivity loss
Job satisfaction/performance	Social disintegration
Stress reduction	Wasted network bandwidth
Work-life balance	Possible legal liability
Empowerment and motivation	Information leakage
Social networking	Cyber-workplace deviance
Relationship building	
Organizational learning	

Pathological Internet usage or Internet addiction involves excessive Internet usage as a way to cope with personal problems or difficulties (David, 2001; Greenfield, 1999; Young, 1998). Pathological Internet users have low self-esteem and are socially suppressed (Niemz, Griffiths, & Banyard, 2005). Mood-altering, denial of responsibilities, guilt, and craving are the common symptoms of Internet addicts (Morahan-Martin & Schumacher, 2000). Other symptoms include loneliness, boredom, salience, and lack of control (Nichols & Nicki, 2004; Widyanto & McMurran, 2004). They have a higher tolerance level, withdrawal symptoms, and a craving for the Internet as compared to normal Internet users (Brenner, 1997). Occurrences of Internet addiction are rare in the workplace, and most organizations have taken harsh measures to punish Internet addicts including workplace reprimand, employee termination, and pressing criminal charges (Warden, Phillips, & Ogloff, 2004).

The next section explains how a proper balance of Internet usage can occur, and suggests potential strategies for effective Internet management in the workplace.

BALANCING OF WORK-PLAY

Work/play balance is based on the equilibrium between employee extrinsic and intrinsic motivations. Extrinsic motivation is related to productive Internet use. That is, employees utilize the Internet to achieve work-related rewards such as monetary, promotion, and/or recognition. Intrinsic motivation is related to personal Internet usage at work. Occasionally, such activities can lead to both positive and negative outcomes, as indicated earlier. Self-determination theory suggests that people seek out stimulating and challenging activities in order to fulfill their interests and enjoyment. When they do, these activities generate the sense of competence and self-determination (Deci & Ryan, 1985), which can facilitate positive emotional experience (Matsumoto & Sanders, 1988) and individual learning (Pintrich & Schrauben, 1992).

As personal Internet usage influences the well-being of employees, Internet usage policy (IUP) must be examined carefully so it will not decrease any positive motivation or work morale. The policy should include training and education, which can be used as a tool for effective communication between management and employees. When Internet monitoring is necessary to enforce appropriate behavioral norms, the policy should be based on maintaining employee job performance. To achieve work/play balance, the article recommends three Internet management strategies.

- **Individual-organizational Internet behavioral alignment:** Employees' perceptions of proper Internet activities need to be aligned with organizational policy. The act of alignment is not as straightforward as it seems, because the 3Ps of Internet usage also asserts its "weight" on individual and organizational psycho-socio-technological situations. In other words, increasing an individual's "productive" Internet usage will reduce the occurrence of personal/pathological Internet usage. However, the increase in productive Internet usage also is related to the ways in which a person behaves socially, technically, and psychologically.

  A proper balance occurs when there is a fit among psycho-socio-technological factors of an individual and those of an organization. Figure 1 shows the intricacies of various concepts described earlier. To create a balance, (1) organizations must have sufficient technological support and infrastructure that match the knowledge and efficacy of their employees; (2) the need for social relationships of an individual must equal the general social norms and culture in the workplace; and (3) organizational psychology (e.g., the organizational analyses of job, career paths, personality, ethics, morale, and attitudes its employees) must adhere to individuals' psyche and motivation. Once there is a mutual understanding of psycho-socio-technological factors, an adaptive internet monitoring and filtering policy (AIMF) can be implemented.

Figure 1. Internet work/play balance framework

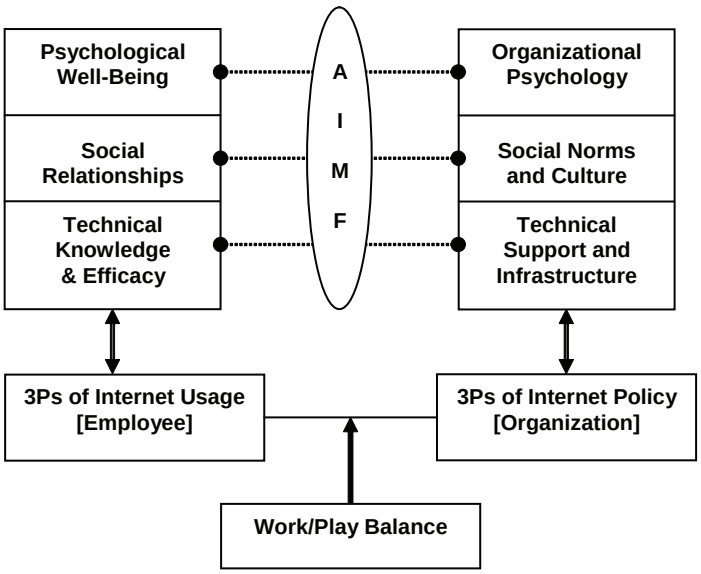


Table 2. Workplace Internet monitoring and filtering policy

Well-being	Performance	Types of Internet Activities	Restriction
+	+	• Learning about educational training classes related to work. • Visiting professional Web sites. • Reading current news/events to fulfill a career or promotional goal. • Searching for new business tools.	LOW/ MODERATE
-	-	• Visiting pornography sites. • Downloading music or illegal software. • Escaping from job responsibilities. • Excessive gaming and chatting. • Any pathological activities.	HIGH
+/-	+	• Accessing organizational intranet/extranet sites. • Sending and sharing information with co-workers. • Visiting customers, suppliers and/or competitors Web sites, etc.	LOW
+	-	• Researching personal hobbies. • Booking travel vacation. • Chatting with co-workers. • Engaging in minor leisure activities. • Surfing without any purpose or to reduce stress. • Reading general news and sport scores.	MODERATE/ HIGH

- **Adaptive Internet monitoring and filtering policy (AIMF):** This strategy requires a reciprocal sense of respect and fulfillment of an organizational—employee psychological contract (Mahatanankoon, 2005). Organizations may allow non-work-related Internet usage as long as employees satisfy their job requirements. Nevertheless, organizations need to provide adequate ethical education for their employees and establish Internet usage norms through informal peer influence. The psycho-socio-technical needs and job characteristics dictate the amount of Internet usage activities performed by the employees (see Figure 1); therefore, organizations must take these factors into consideration before implementing this strategy.
- **Workplace Internet monitoring and filtering policy:** This strategy examines Internet activities based on job satisfaction and job performance dimensions (Mahatanankoon & Igbaria, 2004). These dimensions are related to employee well-being and job performance, which can be used to develop a monitoring/filtering policy for personal Internet usage.

Employee well-being = *f* (job satisfaction, intrinsic motivation, quality of work life)
Job performance = *f* (productivity, extrinsic motivation)

The quality of any personal internet usage is based on its effects on well-being and job performance. Again, it is equally important to acquire the proper understanding of the 3Ps of Internet usage and the impact on individual and organizational psycho-socio-technological situations. In other words, an Internet monitoring/filtering policy should rely on general norms, cultural fits, and the objectives of each organization. Table 2 defines the type of Internet activities and the potential organizational policies.

FUTURE TRENDS

Management must decide on the tradeoffs between employee work/play balance and organizational liability. Some Internet activities can increase an employee's job satisfaction by amplifying his or her intrinsic motivations while reducing stress; however, such behaviors also can lead to higher organizational risks and monitoring costs. The new challenge for Internet-enabled workplaces is to find the right balance between work and play. Future studies need to examine:

- Relationships among the 3Ps of Internet usage and job satisfaction/motivation
- Suitable education and training to enhance work and IT ethics
- Filtering strategies to identify problem employees
- Strategies to reprimand valued employees
- Moderating effects of demographic and cultural differences on the 3Ps of Internet usage
- Possible psychological, social, and technical factors that could lead to the 3Ps of Internet usage

CONCLUSION

The Internet has become one of the most essential technological tools in today's workplace. The broad scope of its usefulness and ease of use makes the Internet beneficial to work and play activities. This article identifies the impact of the 3Ps of Internet usage on individual and organizational psycho-socio-technical factors, and recommends several management strategies to create a balance between work and play. These strategies will help researchers and practitioners better understand the Internet usage patterns of employees and assist them in implementing better Internet usage policies that fit the workplace environment, employees' personal agendas, and management goals.

REFERENCES

Anandarjan, M., Devine, P., & Simmers, C. A. (2004). A multidimensional scaling approach to personal Web usage in the workplace. In M. Anandarajan & C. A. Simmers (Eds.), *Personal Web usage in the workplace: A guide to effective human resources management* (pp. 61-78). Hershey, PA: Idea Group Publishing.

Anandarajan, M., & Simmers, C. (2002). Factors Influencing Web access behavior in the workplace: a structural equation approach. In M. Anandarajan (Ed.), *Internet usage in the workplace: A social, ethical, and legal perspective* (pp. 44-66). Hershey, PA: Idea Group Publishing.

Belanger, F., & Van Slyke, C. (2002). Abuse or learning? *Communications of the ACM, 45*(1), 64-65.

Brenner, V. (1997). Psychology of computer use: XLVII. Parameters of Internet use, abuse, and addiction: The first 90 days of the Internet Usage Survey. *Psychological Reports, 80*(3), 879-882.

David, R. A. (2001). A cognitive-behavioral model of pathological Internet use. *Computers in Human Behavior, 17*(2), 187-195.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.

Greenfield, D. N. (1999). *Virtual addiction: Help for Netheads, Cyberfreaks, and those who love them*. Oakland, CA: New Harbinger Publications, Inc.

Kraut, R., & Kiesler, S. (2003). The social impact of Internet use. *Psychological Science Agenda, 16*(3), 8-10.

Lim, K. G. (2002). The IT Way of loafing on the job: Cyberloafing, neutralizing, and organizational justice. *Journal of Organizational Behavior, 23*(5), 675-694.

Mahatanankoon, P. (2005). Personal Internet usage and quality of worklife. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology: Volume 4* (pp. 2277-2281). Hershey, PA: Idea Group Publishing.

Mahatanankoon, P., & Igbaria, M. (2004). Impact of personal Internet usage on employee's well being. In M. Anandarajan & C. A. Simmers (Eds.), *Personal Web usage in the workplace: A guide to effective human resources management* (pp. 246-263). Hershey, PA: Idea Group Publishing.

Mahatanankoon, P., Anandarajan, & Igbaria, M. (2004). Development of a measure of personal Web usage in the workplace. *CyberPsychology & Behavior, 7*(1), 93-104.

Matsumoto, D., & Sanders, M. (1988). Emotional experiences during engagement in intrinsically and extrinsically motivated tasks. *Motivation and Emotion, 12*(4), 353-369.

Morahan-Martin, J., & Schumacher, P. (2000). Incidence and correlates of pathological Internet use among college students. *Computers in Human Behavior, 16*(1), 13-29.

Nichols, L. A., & Nicki, R. (2004). Development of a psychometrically sound Internet addiction scale: A preliminary step. *Psychology of Addictive Behaviors, 18*(4), 381-384.

Niemz, K., Griffiths, M., & Banyard, P. (2005). Prevalence of pathological Internet use among university students and correlations with self-esteem: The general health questionnaire (GHQ) and Disinhibition. *CyberPsychology & Behavior, 8*(6), 562-570.

Oravec, J. A. (2002). Contructive approach to Internet recreation in the workplace. *Communications of the ACM, 45*(1), 60-63.

Pintrich, P. R., & Schrauben, B. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic task. In D. H. Schunk & J. L. Meece (Eds.), *Student perceptions in the classroom* (pp. 149-183). Hillsdale, NJ: Erlbaum.

Stanton, J. M. (2002). Company profile of the frequent Internet user. *Communications of the ACM, 45*(1), 55-59.

Warden, N. L., Phillips, J. G., & Ogloff, J. R. P. (2004). Internet addiction. *Psychiatry, Psychology, and Law, 11*(2), 280-295.

Widyanto, L., & McMurran, M. (2004). The psychometric properties of the Internet addiction test. *CyberPsychology & Behavior, 7*(4), 443-450.

Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior, 1*(3), 237-244.

KEY TERMS

Cyberloafing: Any production-deviant act in which of employees use their organization's Internet access and e-mail accounts during work hours for non-work-related purposes.

Internet Abuse: Any wrongful or improper use of the Internet in the workplace.

Internet Filtering and Monitoring Software: Software tools used to reduce occurrences of Internet abuse by blocking inappropriate Web sites and identifying frequently visited Web sites.

Internet Usage Policy (IUP): An organizational policy given to employees that governs the use of the Internet in a specific workplace. The goals of an IUP, if properly written and implemented, are to help organizations communicate proper Internet usage behaviors lessen employees' perceived expectation of privacy, and reduce costly litigation that may occur from the use of Internet monitoring and filtering software.

Pathological Internet Use: Excessive Internet usage by people as a means of coping with their personal problems or current personal difficulties.

Personal Internet Usage: Voluntary online Web behavior during working time in which employees utilize any of the organizations' resources for activities outside current customary job/work requirements.

Psycho-Socio-Technical: Factors that contribute to the impact of psychological, social, and technological aspects.

Quality of Work Life: Workplace factors that support the well-being and job satisfaction of employees.

Work/Play Balance: The balance between work and play that increases the quality of work life.

Interoperability between Distributed Systems and Web-Services Composition

Christophe Nicolle
Université de Bourgogne, France

INTRODUCTION

An information system is a multi-axis system characterized by a "data" axis, a "behavioral" axis, and a "communication" axis. The data axis corresponds to the structural and schematic technologies used to store data into the system. The behavioral axis represents management and production processes carried out by the system and corresponding technologies. The processes can interact with the data to extract, generate, and store data. The communication axis relates to the network used to exchange data and activate processes between geographically distant users or machines. Nowadays, technologies required for interoperability are extended to deal with the semantic aspect of the information systems. The aim of the semantic axis is to take into account new aspects of the sharing of the data and the processes, such as the understanding of the data and the processes, the access security, and owner rights (OWL Services Coalition, 2006).

Information system interoperation has emerged as a central design issue in Web-based information systems to allow data and service sharing among heterogeneous systems. Data heterogeneity stemming from the diversity of data formats or models used to represent and store information in the Web is a major obstacle to information systems interoperability. These data models range from the structured data models (network, relational, OO) found in traditional databases to flat files and emerging Web oriented semi-structured models. Information system interoperability aims at supporting the amalgamation autonomous heterogeneous systems to create integrated virtual environments or architectures in which information from multiple disparate sources can be accessed in a transparent and efficient manner. As an example of such integrated virtual systems, consider an airline reservation system based on the integration of a group of airlines reservation and ticket sale information systems. The specific airline systems provide various types of fares and

Figure 1. Axis for interoperability

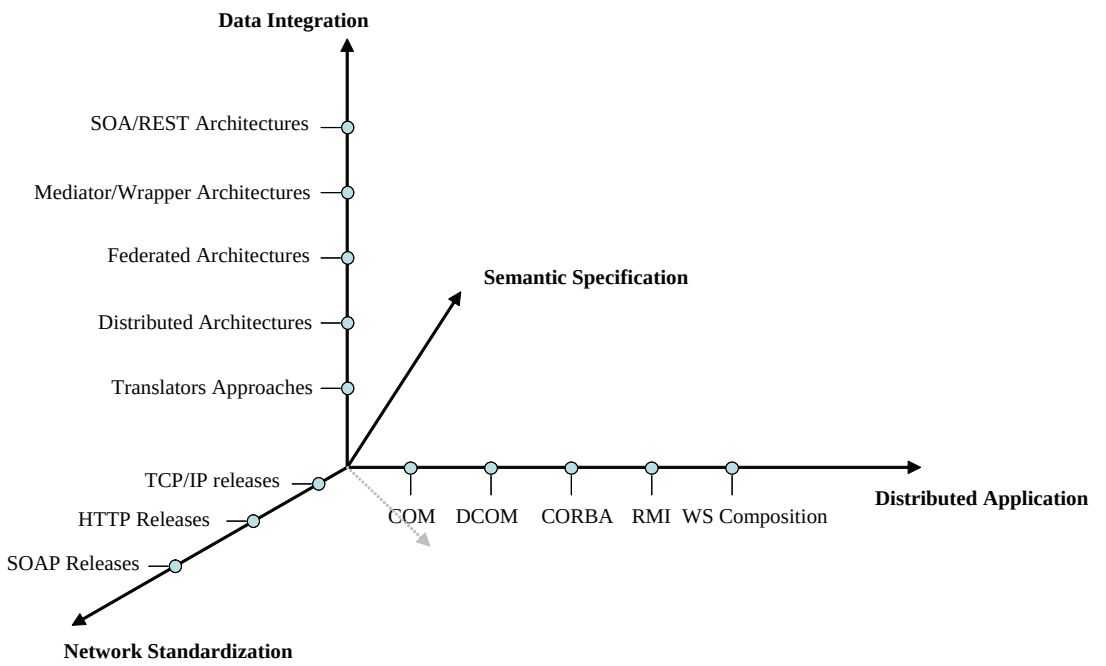


Figure 2. Database translation approach

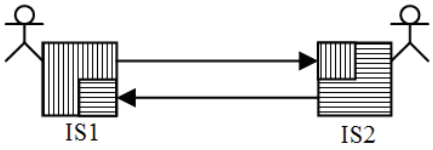


Figure 3. Standardization approach

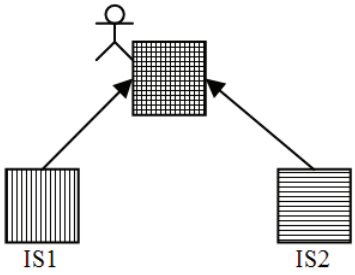


Figure 4. Federated systems

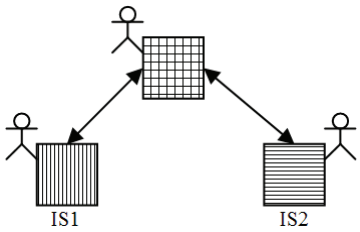
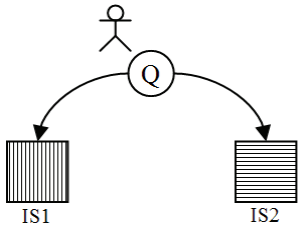


Figure 5. Multi-base systems



special discount trips. That can be searched and compared to respond to user queries for finding the best available prices for specified flights.

BACKGROUND

Database interoperability issues have been extensively studied in the past. Several approaches, including database translation, distributed systems, federations, language based multidatabase, ontology, and mediation, have been proposed to bridge the semantic gaps among heterogeneous information systems.

The *database translation* approach is a point-to-point solution based on direct data mappings between pairs of information systems. The mappings are used to resolve data discrepancies among the systems (Yan & Ling, 1992). The database translation approach is most appropriate for small-scale information processing environments with a reduced number of participants. The number of translators grows with the square of the number of components in the integrated system. For example, consider two information systems, IS1 and IS2 in the earlier travel agency example. The corresponding translators must be placed between the information systems as shown in Figures 2 and 3. Information in IS1 is represented by vertical lines, while the information in IS2 is shown as horizontal lines.

In the *standardization* approach, the information sources use the same model or standard for data representation and communication. The standard model can be a comprehensive metamodel capable of integrating the requirements of the models of the different components (Atzeni & Torlone, 1997). The use of a standard metamodel reduces the number of translators (this number grows linearly with the number of components) to resolve semantic differences. However, the construction of a comprehensive metamodel is difficult; the manipulation of high-level languages is complex; and there are no unified database interfaces. In our example, the travel agencies must define a common model to export their data. A centralized information system can be built to replace the original information systems (IS1, IS2). The global centralized schema is a combination of the entire data (horizontal and vertical lines) contained in IS1 and IS2.

Federated systems consist of a set of heterogeneous databases in which federation users can access and manipulate data transparently without knowledge of the data location (Sheth & Larson, 1990). Each federation database includes a federated schema that incorporates the data exported by one or more remote information systems. There are two types of federations. A tightly coupled federation is based on a global federated schema that combines all participant schemas. The federated schema is constructed and maintained by the federation administrator. A loosely coupled federation includes one or more federated schemas that are created by users or the local database administrator. The federated schema incorporates a subset of the schema available in the federation. This approach becomes rapidly complex when the number of translators required becomes large. In our example, the existing information systems are completely operational for local users. Only the shared data

Figure 6. Ontology approach

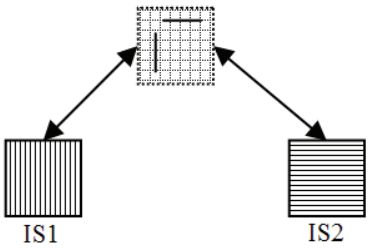
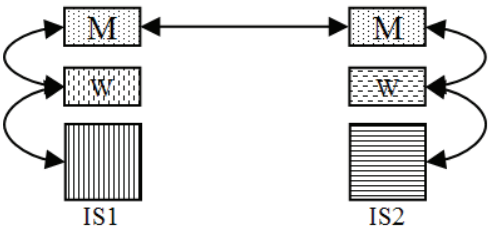


Figure 7. Mediation approach



are integrated in the federated schema. The federated system is only made of horizontal and vertical lines that IS1 and IS2 want to exchange.

Language-based *multi-base systems* consist of a loosely connected collection of databases in which a common query language is used to access the contents of the local and remote databases (Keim, Kriegel, & Miethsam, 1994). In this approach, in contrast to the distributed and federated systems, the burden of creating the federated schema is placed on the users, who must discover and understand the semantics of the remote databases. In our example, the various companies have to define a global common language (Q) to query their information systems (IS1, IS2). This solution is well adapted for information systems that are based on the same family of data models and does not require complex query translators.

The *ontology*-based interoperability approach uses ontology to provide an explicit conceptualization of the common domain of a collection of information systems (Benslimane, Leclercq, Savonnet, Terrasse, & Yétongnon, 2000). An ontology defines a common vocabulary that can be used by users from different systems. The construction of an ontology for a domain is a difficult task and often requires merging existing overlapping ontologies. The interoperability solutions based on ontology describe the semantics of information rather than their organization or their format. In our example, the companies have to define ontology to capture the semantics of their domain of activity.

The *mediation* approach is based on two main components: mediator and wrapper. The mediator is used to create and support an integrated view of data over multiple sources. It provides various services to support query processing. For instance, a mediator can cooperate with other mediators to decompose a query into sub queries and generates an execution plan based on the resources of the cooperating sites. The wrapper is used to map the local databases into a common federation data model. The wrapper component provides the basic data access functions (Hammer et al., 1995). In our example, a translator, which acts as a wrapper, is placed between the conceptual representation of the mediator and the local description of each information source.

As new data models are developed for Web-based information systems, there is a need to extend interoperability solutions to take into account requirements and specifications of the new models. For instance, *XML* (XML, 2004) has emerged recently as an important model for describing and sharing Web-based data. This importance stems from two major factors. First, XML is becoming a de facto data standard supported by many software vendors and applications developers. Second, XML is based on a relatively simple structure that is both user- and machine-readable and that can be used by non-expert database administrators. The existing Web technologies are not initially intended to address some of the issues involved in database integration. For instance the Web-browsing paradigm is efficient for data look up in large environments, but is inadequate for database integration support. To use this paradigm to locate and merge data requires costly applications that are often tailored to specific integration needs. New challenges have arisen from the development of Web-based information systems. One of the challenges is the need to develop Web-oriented tools to support information integration and allow access to local as well as remote information sources.

Since 2000, *Web services* have been proposed as a method to address some of the challenges of Web-based integrated systems. A Web service can be viewed as a set of layers contained in a stack. The layers are dynamically defined

Figure 8. WS approach

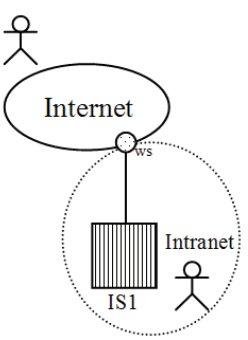


Table 1. Overview of architectures for interoperable information systems

Systems	Advantage	Limits	Tools or methods used	Level
Translation	• Better control of point-to-point translation.	• Requires a large number of translators in open environments. • Adding a new information system requires 2(n-1) translators.	Required n*(n-1) translators	☑ D ☐ B ☐ C
Standardization	• Use of pivot, canonical model or metamodel. • Reduce the number of translators.	• Definition of a common standard accepted by all ISs. • The construction of a comprehensive metamodel is difficult.	Required 2n translators	☑ D ☐ B ☐ C
Federation	• Derived from standardization. • Local IS are autonomous.	• Use of a global, static federated schema. • The construction of an integrated federal schema is difficult. • New addition requires redesign of federated schema.	Required 2n translators	☑ D ☐ B ☐ C
Multi-base	• Used of a single language for many IS.	• The common interoperate language does not export local system semantics. • Users need to discover and understand the semantic of remote IS.	Query Based	☑ D ☑ B ☐ C
Ontology	• Semantic-oriented solution.	• Extensive ontologies are voluminous. • Requires meta-level translation.	Semantic	☑ D ☐ B ☐ C
Mediation	• Combine translation and semantic. • Local ISs are autonomous.	• Difficult to construct automatic mediator process.	Required 2n semantic translators	☑ D ☑ B ☐ C
Web services	• Resolved format level. • Resolved process translation. • All levels of IS are managed. • Normalized solution. • Developed by both industrials and researchers.	• Security mechanism not finalized. • Combination of Web services not resolved.	Protocol SOAP, WSDL, UDDI	☑ D ☑ B ☑ C

Note: D = Data level, B = Behavioral level, C = Communication level.

following user needs and are called through a set of Internet protocols. The protocols are different; those propose for various network architectures. However, in all Web service architecture, a base set of protocols is always used (W3C, 2002). This base set is composed of SOAP (SOAP, 2003), WSDL (WSDL, 2004), and UDDI (UDDI, 2003). They allow the discovery, description and information exchanges between Web services.

SOAP is a mechanism that uses XML for the exchange of structured and typed information between several actors in a decentralized and distributed environment. SOAP does not define the semantics of the application, but provides a mechanism for expressing semantics by proposing a modular template and mechanisms for data coding.

WSDL uses XML syntax to describe the methods and parameters of Web services. These parameters include protocols, servers, ports, input and output messages format, and exceptions format. With WSDL, an application using SOAP can auto-configure the Web services exchanges, masking the majority of the low-level technical details.

UDDI is a Web-based company world directory combining "white pages" (information such as name, address, telephone number, and other contact information of a given business), "yellow pages" (information that categorizes businesses), and "green pages" (technical information about the Web services provided by a given business.). UDDI allows Web service references by automating all search procedures. Table 2 presents the advantages and limits of Web services.

In our example, a set of Web services can be built from each information system independently from the other information systems. The Web services become a standard interface to access the local information system. These Web services can be used by customers and partners via the Internet and by local users via an intranet. This solution is very flexible and reduces the complexity of the heterogeneity problem.

Industrial Development Platforms

To achieve Web service architecture, several industrial tools have been developed. Four main actors in the industrial world share the market. The solutions proposed by Microsoft and SUN are language oriented, while the solutions proposed by IBM and BEA are platform oriented.

Microsoft.NET proposes a software platform on which companies can exchange data and services on the Internet based on an ASP model (Application Provider Service). Most Microsoft products can be extended to use Web services developed with the *.NET*. The philosophy of this solution can be resumed by "one OS, many languages."

The *SUN J2EE* was developed by the Java Community Process. It is a set of services and specification containing JDBC (Java Data Base Connector), JMS (Java Message Services), JSP (Java Server Pages), EJB (Enterprise Java Beans), and so forth. J2EE 1.4 includes Web service specifications using an open source framework called AXIS (used by *IBM WebSphere*). In response to the *Microsoft .NET* solution, SUN proposes *ONE*, which groups the set of SUN Web services propositions. The philosophy of this solution can be resumed by "many OS, one language."

IBM WebSphere is a set of components allowing the creation of interoperable information systems based on Web services. These components include Interchange Server, which allows process integration; MQ Integrator Broker, which allows data integration; MQ Workflow, which allows processes management; and so forth. *IBM WebSphere* uses the SUN JAVA language for the development of its Web services.

The *BEA WebLogic Server* is based on the Java Connectors Architecture. This tool uses the notion of components and connectors that can be integrated between them. The integration of the connectors is carried out by the Application Integration framework and the Adapter Development Kit. To manage the resulting architecture, a Business Process Management is used in coordination with the B2B integration tool. This tool exploits standards such as XML, HTTP, or SSL and semantic solutions such as RosettaNet, cXML, ebXML, or EDI.

Table 1 summarizes the various architectures for the interoperation of information systems. In this table, a brief presentation of the advantages and limits of each approach is given.

FUTURE TRENDS

The next major challenge in the Web service world is to construct adaptive distributed applications that only exist for the time of their use and that are constructed according to the end user's waiting. This is possible by merging researches from the interoperability area and from the adaptive hypermedia systems.

CONCLUSION

For the past 20 years or so, the need to exchange information between various partners pushed researchers to develop architectures for the interoperability of information systems. The proposed architectures have addressed several key interoperability issues, ranging from the resolution of data format heterogeneity using translations-based architecture and the reduction of the number of required translators in standardization-based architecture to the resolution of semantic heterogeneity based on ontology and the resolution of process heterogeneity with mediation-based architecture.

Nowadays, information systems can be integrated or disassociated depending on the market trends of enterprise mergers. The Web service-based architecture allows the development of this type of interoperability by proposing a standard data format with XML, standard communication architecture based on the SOAP protocol, and a standard description of processes using WSDL and UDDI.

Nevertheless, the creation of distributed applications made of combined Web services is not really resolved. Since three years many researcher focus their works in the definition of a global framework to combined Web services. These frameworks are based on semantics characterization of services, orchestration languages, workflows, and so on (Agarwal et al., 2005).

REFERENCES

Agarwal, V., Dasgupta, K., Karnik, N., Kumar, A., Kundu, A., Mittal, S., et al. (2005, May 10-14). A service creation environment based on end to end composition of Web services. In A. Ellis & T. Hagino (Eds.), *Proceedings of the 14th International Conference on World Wide Web (WWW 2005)*, Chiba, Japan (pp. 128-137). ACM.

Atzeni, P., & Torlone, R. (1997, May 13-15). MDM: A multiple-data-model tool for the management of heterogeneous database schemes. In J. Peckham (Ed.), *Proceedings of the SIGMOD International Conference on Management Data* (pp. 528-553).

Benslimane, D., Leclercq, E., Savonnet, M., Terrasse, M. N., & Yétongnon, K. (2000). On the definition of generic multi-layered ontologies for urban applications. *International Journal of Computers, Environment and Urban Systems, 24*(2000), 191-214.

Hammer, J., Garcia-Molina, H., Ireland, K., Papakonstantinou, Y., Ullman, J. D., & Widom, J. (1995). Information

translation, mediation, and mosaic-based browsing in the TSIMMIS system. In M. J. Carey, D. A. Schneider (Eds.), *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data,* San Jose, CA (p. 483). ACM Press.

Keim, D. A., Kriegel, H. P., & Miethsam, A. (1994, May 17-20). Query translation supporting the migration of legacy databases into cooperative information systems. In M. L. Brodie, M. Jarke, M. P. Papazoglou (Eds.), *Proceedings of the Second International Conference on Cooperative Information Systems,* Toronto (pp. 203-214).

OWL Services Coalition. (2003). OWL-S 1.2 Pre-Release: Semantic markup for Web services. Retrieved March 2006, from http://www.ai.sri.com/daml/services/owl-2/1.2

Sheth, A. P., & Larson, J. A. (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM computing surveys, 22*(3), 183-236.

SOAP. (2003). *SOAP Version 1.2 Part 0: Primer, W3C recommendation.* Retrieved June 24, 2003, from http://www.w3.org/TR/soap12-part0/

UDDI. (2003). *UDDI Version 3.0.1, UDDI Spec. Technical Committee Specification.* Retrieved October 14, 203, from http://uddi.org/pubs/uddi-v3.0.1-20031014.htm

W3C. (2002). *Web services architecture requirements,* W3C, Working Draft. Retrieved November 14, 2002, from http://www.w3.org/TR/2002/WD-wsa-reqs-20021114

WSDL. (2004). *Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language,* W3C, Working Draft. Retrieved August 3, 2004, from http://www.w3.org/TR/wsdl20/

XML. (2004). *Extensible Markup Language 1.0* (3rd ed.), W3C Recommendation. Retrieved February 4, 2004, from http://www.w3.org/TR/REC-xml

Yan, L. L., & Ling, T. W. (1992). Translating relational schema with constraints into OODB schema. In D. K. Hsiao, E. J. Neuhold, R. Sacks-Davis (Eds.), *Proceedings of the IFIP WG2.6 Database Semantics Conference on Interoperable Database Systems (DS-5),* Lorne, Victoria, Australia (pp. 69-85). North-Holland.

KEY TERMS

eXtensible Makrup Language (XML): A language for creating markup languages. There are two kinds of XML documents: well-formed and valid. The first respects the XML standard for the inclusion and the names of the tags. The second must well-be formed and uses a grammar to define the structure and the types of the data described by the document.

Interoperability: The ability of heterogeneous software and hardware to communicate and share information.

Ontology: An explicit formal specification of how to represent the objects, concepts, and entities existing in some area of interest and the relationships among them.

Simple Object Access Protocol (SOAP): An XML-based message protocol used to encode information in Web service requests and response messages before sending them over a network. SOAP messages are independent of any operating system or protocol and may be transported using Internet protocols (SMTP, MIME, and HTTP).

Universal Description, Discovery, and Integration (UDDI): A Web-based distributed directory for discovery of Web services offered by companies. It is similar to a traditional phone book's yellow and white pages.

Web Service: A software system designed to support interoperable machine-to-machine interaction over a network. It has an interface described in a machine-readable format (specifically WSDL). Other systems interact with the Web service in a manner prescribed by its description using SOAP-messages, typically conveyed using HTTP with an XML serialization in conjunction with other Web-related standards.

Web Services Language Description (WSDL): An XML-formatted language used to describe a Web service's capabilities as collections of communication endpoints capable of exchanging messages.

Interventions and Solutions in Gender and IT

Amy B. Woszczynski
Kennesaw State University, USA

Janette Moody
The Citadel, USA

INTRODUCTION

The role of women in technology-related fields began with promising contributions from pioneers like Grace Hopper. In recent years, women have moved away from information technology (IT) fields, and the number of women selecting IT majors in universities continues to decline. Likewise, the number of women employed in the IT workforce remains low and declining.

Researchers have recognized the problem and have investigated the many reasons for low participation of women in IT-related fields. Researchers have proposed various interventions to fill the pipeline and retain women in computing.

In this chapter, we provide an overview of the current state of women in IT. We focus on girls and women at various life stages, from early education to the IT workplace. We also provide a discussion of the various methods and appropriate interventions that may be employed to encourage women to become empowered users of technology worldwide.

We use a broad definition of IT, which includes computer science (CS), computer engineering, information systems (IS), information technology (IT), and related professional fields. By examining research from multiple technology-related fields, we gain a clearer picture of the many ways that women may participate in IT.

Recent research on gender and IT has used an interdisciplinary approach, which has greatly expanded our potential for understanding why women decide not to pursue IT-related fields and how to implement appropriate interventions. Researchers from topics as diverse as IS, psychology, social sciences, education, and feminism, have taken a distinctive approach to understanding why women are not better represented in the IT workplace. We believe this broad, interdisciplinary approach has great potential to understand motivations for women pursuing IT-related careers. As Trauth & Niederman (2006, p. 8) said, "…the IT profession is challenged with meeting the demand to enlarge the IT workforce by recruiting and retaining personnel from historically underrepresented groups." This chapter looks at women in IT, shedding light on one historically underrepresented group.

BACKGROUND

Previous literature on women in IT has focused on education and the IT workforce. More recent research pursuits have focused on feminism as a lens through which to view gender and IT. The following sections discuss these areas.

Primary and Secondary Education

To increase the pipeline of women pursuing IT-related majors in universities, we must reach girls at a young age. Many factors, both structural and social, influence career choices of both genders, as seen in Adya & Kaiser's (2005) model. Generally speaking, social influences come from role models and influences by family members, peers, and the media; whereas structural influences are found in the support provided by educational institutions. One ubiquitous and early influence on a young girl's perceptions of computers comes through the mass media and its gendered implications, as reported by Gannon (2007). For example, magazines that appeal to teenage girls, including those with global editions for other cultures, consistently fail to portray women in professional careers using technology (Adya & Kaiser, 2005), but, rather, focus on beauty, fashion, and relationship items. Even as young women expand their readings and increase their exposure to home computing magazines, they will find images of women as novices when dealing with technology, in contrast to technologically competent and powerful males (Johnson & Lynch, 2006).

Although there has been little advice regarding how to counteract the media influences, researchers have made multiple suggestions on how to modify structural influences, one of which is the use of single-sex schools. However, a recent study showed that girls in single-sex schools did no better than their counterparts in coeducation schools in deciding to major in CS (Olivieri, 2005). Olivieri proposes that the lack of computer knowledge and understanding are more common reasons that girls do not choose to major in CS rather than the presence of mostly men in IT courses.

Some researchers have advocated exposing girls to programming as early as possible to increase their comfort and skill in developing simple programs on their own, be-

lieving this exposure will give girls an edge when they take the first programming course in college. However, Katz, Allbritton, Aronis, Wilson & Soffa (2006) note that girls who develop programming skills in high school may do so at the expense of advanced math skills. Clearly, that outcome is not satisfactory since math skills are highly correlated with success in CS.

Post-Secondary Education

Student perceptions of IT are "moderately gendered with a greater emphasis on masculine traits and abilities" (Joshi & Schmidt, 2006, p. 38). Students pursuing university degrees often do not understand IT-related fields and the diversity of career opportunities available. In fact, when college students are asked to draw a computer scientist, they usually draw a geeky, smart person, often with glasses, eating junk food and invariably male (Martin, 2004). Moreover, stereotypes associated with CS tend to be associated with IT in general, at least until undergraduate students are exposed to IT careers in introductory college courses. At that time, students begin to see the diversity in IT and how CS is uniquely different from IS (Joshi & Schmidt, 2006).

Most studies show that there are few significant differences for the variables of gender, persistence in IT-related courses, and achievement (Ilias & Kordaki, 2006). However, Katz, et al. (2006) showed that men who earned a grade below "B" in an early CS course were more likely to persist in CS than were their female counterparts. Perhaps average performing males have higher levels of self-confidence and believe they can succeed in subsequent courses as compared to their female colleagues. Interestingly, Ilias & Kordaki (2006) found that female graduate computer engineering students in Greece completed their studies faster than their male counterparts.

Researchers have recommended that university professors use care when integrating group projects and online learning tools into the classroom. Wolfe & Alexander (2005) showed that when coeducational project teams form, a single male often emerges as the resident computer expert. Then women do not learn how to effectively use technology tools, and men receive great credit for being the technical expert, thus perpetuating the myth of the masculine-dominated IT field. Moreover, when professors use technology tools in male-female classes, they should ensure that online tools allow equal participation by all students (Huynh & Schuldt, 2005).

Research has also shown that another factor affecting the retention of IT students is the type of assignments given in CS classes. For example, female students prefer to work on real-world applications, while males prefer to work on game problems. Yet, current textbooks continue to provide a large percentage of math problems (Wilson, 2006).

The IT Workforce

For those females who successfully persist to complete an IT degree, other challenges await them in the professional world. One of the major issues affecting women's job satisfaction is work-family life balance (Gallivan, 2004). Most women realize that IT careers require them to constantly re-train and learn new technologies. Much of the dissatisfaction with work-family life balance may be explained in how companies approach professional development opportunities. For example, in a company that paid for employees to undertake training on company time, women had lower levels of stress, as compared to women at a company that expected employees to undertake professional development on their own time (Gallivan, 2004). Since women are typically the primary caregivers for children and older parents, the work-family life balance issue may be more important to them than their male counterparts.

Another important issue related to work-family life balance is flexibility in scheduling. Armstrong, Riemenschneider, Allen & Reid (2007) found that women were concerned about the lack of consistency and equality in corporate policies regarding flexible work schedules. As the primary caregivers, women may need more flexible scheduling options than their male counterparts. As Tapia (2006, p. 94) notes, companies should develop organizational policies that are "sensitive, especially with regard to training and professional career development of a diverse IT staff."

Ironically, while medium and large corporations may not have been able to accommodate the work-family life balance requirements of women, women entrepreneurs have excelled in using IT to set up their own companies, thereby customizing how technology can help balance home and work needs. Computers and the Internet have enabled female small business owners to become independent and highly effective in developing new ways to meet market demands (Martin & Wright, 2005). Clearly the passive and novice role of women with respect to computers as portrayed in computer advertisements (Johnson, Rowna & Lynch, 2006) is not indicative of reality.

Global Issues

For issues related to IT education and the IT workforce, research needs to include a global perspective. There may be differences in groups not because of gender, but because of culture (Sagi, Carayannis, Dasgupta, & Thomas, 2004). When attempting to fill the pipeline and ultimately populate the IT workforce with women and other underrepresented groups, careful consideration of cultural and local contexts must be understood for successful development of interventions.

For example, the NetCorps Jordan Project, an IT training project in Jordan, found that culture, context, and gender

all play a significant role in who gets to do what with IT (Wheeler, 2005). These types of considerations may be particularly relevant in Middle Eastern and other countries where women do not typically work outside the home and usually have very traditional roles based on cultural contexts. In this case, offering IT training in primary and secondary schools has to be considered in the context of relevant social norms for the country in question. Much of the research on women in IT, however, has focused on a narrow view that considers mostly United States women. We cannot design interventions to help women break the IT glass ceiling without understanding the global environment.

We may need to develop training programs based on the particular culture and country situation. Understanding issues such as country infrastructure, availability of the Internet, and women's roles in society, is important to make an impact on the number of women who enter the IT workforce. Wheeler (2005) suggests using local people for IT training as a way to overcome cultural/country issues.

Unlike the United States, some countries may not have a gendered impression of IT. For example, women in Malaysia, who often outnumber men in undergraduate and graduate programs in CS, did not perceive IT as a masculine field (Othman & Latih, 2006). Of particular note, the universities studied employed a substantial number of women faculty in CS, with women in prominent positions such as Dean or Department Chair. Clearly, the success of women in IT-related programs in Malaysia demonstrates the importance of female mentors.

Feminist Research on Women in IT

Of interest to researchers is the recent emergence of feminist study as it relates to gender and IT. This research stream shifts the lens through which the topic of women and IT is viewed. As recounted by Trauth & Howcroft (2006), it is important to add the critical research approach to the well-entrenched positivist and interpretive research approaches of the past. Positivist research looks at the gender issue in order to quantify the dichotomy of the genders so that management can harness all of the firm's resources with the goal of "optimizing efficiency and enhancing corporate effectiveness." (p. 274). Interpretive research on gender issues looks at the why of these differences, investigating social influences and national cultures as they relate to gender identities. Such research has been seen as simply documenting the influences rather than confronting them for the purpose of change. In contrast to these two approaches, critical feminist research looks at and challenges the social and political power relations that create gender inequality.

Liberal feminist study looks to IT to put women on equal footing with their male counterparts (Rosser, 2005). Cyberfeminists note women's approximately equal use of the Internet as compared to their male counterparts, which coupled with women's lower interests in and development of hardware, illustrate a female's desire to remain connected (Rosser, 2005). Therefore, women may choose different types of career paths in IT, but they need to know about the diversity of options available to them in IT. As Miller (2005, p. 164) notes, "Integrating feminist models with traditional models may produce a computer science or cognitive science that reflects not only experiences associated with males but all the experiences associated with females." We believe this recent and emerging approach to studying women in IT shows promise for improving the understanding of women's motivations for pursuing IT-related degrees and the design of possible interventions to increase representation of women in IT.

INTERVENTIONS AND SOLUTIONS

One basic impediment to increasing the number of persons pursuing an IT career is the pernicious stereotype of the work itself and those entering it. Thus, we characterize the field of IT itself as having a diversity problem. Many women believe that IT is associated with masculine traits, epitomized by the geeky male without a social life who works 24 hours a day while sitting in front of a computer. That homogenous view of IT does not allow women to see the various career options for IT fields. We must design interventions to encourage girls to enter the IT pipeline, mentor women in universities to persist in IT, and support women as they move through life stages in the IT workplace. We also must consider context, culture, and global issues when trying to better understand women and IT. Specifically, the following interventions can increase the number of women in IT globally:

- Educate people about the nature of IT and the variety of career choices available. Start these programs early, reaching girls in primary and secondary school programs.
- Teach teamwork explicitly in the university classroom environment (Wolfe & Alexander, 2005) to avoid a division of labor that allows men to dominate the technology portion of projects.
- Allow time for professional development to better balance personal and career-related needs of women in IT.
- Encourage the use of flexible scheduling in the IT workplace to balance work-family life.
- Develop organizational policies and professional codes of ethics in organizations that encourage professional career development of a diverse IT staff. Doing so should help prevent the hostile work environments that emerged from the dot-com era when young white men

developed new companies consisting of a homogenous staff (Tapia, 2006).
- Increase the use of mentoring to encourage women to enter and persist in the IT pipeline. These mentoring endeavors should also extend into the IT workplace (Othman & Latih, 2006).
- On a global scale, educate men on how women can use IT to develop culturally appropriate businesses or to manage the home. The use of same-sex trainers may make it easier for women to participate in training programs in other countries.
- Consider the many faces of women in the IT profession (Trauth, 2002) and provide research that avoids characterizing all women as the same. Research needs to be expanded beyond its traditional focus on white, middle class, college educated women who are very different from other women.

FUTURE TRENDS

We see the future of research on women in IT as having an interdisciplinary approach, including perspectives from CS, information sciences, social sciences, women's studies, feminism, psychology, and education. We believe this interdisciplinary approach will allow us to better understand women and IT. We see future research as considering multiple interacting variables in addition to gender. Adya & Kaiser (2005) have provided a comprehensive and testable model of factors that influence female career choices as an excellent starting point for this research. By focusing on the many faces of women at various life stages, we will be able to provide a richer discussion of how to increase the number of women in the pipeline and who persist in IT careers.

CONCLUSION

The decreasing number of girls and women participating in IT education and employment continues to be a concern. Today's global economy presents opportunities and challenges for women in their choices of careers. This chapter has presented highlights of recent research into the causes of this decline and suggested interventions to reverse it. Clearly, the research and interventions must continue to adapt to the changing landscape of the IT field, and cross-cultural knowledge must be shared to provide a complete understanding. The future for women in IT looks brighter, broader, and more global.

REFERENCES

Adya, M., & Kaiser, K. (2005). Early determinants of women in the IT workforce: a model of girls' career choices. *Information Technology & People*, 18(3), 230-259.

Armstrong, D., Riemenschneider, C., Allen, M., & Reid, M. (2007). Advancement, voluntary turnover and women in IT: a cognitive study of work-family conflict. *Information & Management*, 44, 142-153.

Gallivan, M. J. (2004). Examining IT professionals' adaptation to technological change: The influence of gender and personal attributes. *The DATA BASE for Advances in Information Systems*, 35 (3), 28-49.

Gannon, S. (2007). Laptops and lipsticks: feminizing technology. *Learning, Media and Technology*, 32(1), 53-67.

Huynh, M. Q., Lee, J. N., & Schuldt, B. A. (2005). The insiders' perspectives: A focus group study on gender issues in a computer-supported collaborative learning environment. *Journal of Information Technology Education*, 4, 237-255.

Ilias, A., & Kordaki, M. (2006). Undergraduate studies in computer science and engineering: Gender issues. *Inroads – The SIGCSE Bulletin*, 38 (2), 81-85.

Johnson, N., Rowna, L., & Lynch, J. (2006). Constructions of gender in computer magazine advertisements: confronting the literature. *Studies in Media & Information Literacy Education*. 6(1). 1-11.

Joshi, K. D., & Schmidt, N. L. (2006). Is the information systems profession gendered? Characterization of IS professional and IS career. *The DATA BASE for Advances in Information Systems*, 37 (4), 26-41.

Katz, S., Allbritton, D., Aronis, J., Wilson, C., & Soffa, M. L. (2006). Gender, achievement, and persistence in an undergraduate computer science program. *The DATA BASE for Advances in Information Systems*, 37 (4), 42-57.

Martin, C. D. (2004). Draw a computer scientist. *Inroads, The SIGCSE Bulletin*, 36 (4), 11-12.

Martin, L., & Wright, L.T. (2005). No gender in cyberspace? Empowering entrepreneurship and innovation in female-run ICT small firms. *International Journal of Entrepreneurial Behaviour & Research*. 11(2), 162-178.

Miller, P. H. (2005). Gender and information technology: perspectives from human cognitive development. *Frontiers*, 26 (1), 148-167.

Olivieri, L. M. (2005). High school environments and girls' interest in computer science. *Inroads, the SIGCSE Bulletin*, 37 (2), 85-88.

Othman, M., & Latih, R. (2006). Women in computer science: No shortage here! *Communications of the ACM*, 49 (3), 111-114.

Rosser, S. V. (2005). Through the lenses of feminist theory: Focus on women and information technology. *Frontiers*, 26 (1), 1-23.

Sagi, J., Carayannis, E., Dasgupta, S., & Thomas, G. (2004). ICT and business in the New Economy: Globalization and attitudes towards eCommerce. *Journal of Global Information Management*, 12 (3), 44-64.

Tapia, A. H. (2006). Hostile work environment.com: Increasing participation of underrepresented groups, lessons learned from the dot-com era. *The DATA BASE for Advances in Information Systems*, 37 (4), 79-98.

Trauth, E. M. (2002). Odd girl out: The individual differences perspective on women in the IT profession. *Information Technology & People*, 15 (2), 98-117.

Trauth, E. M., & Howcroft, D. (2006). Critical empirical research in IS: an example of gender and the IT workforce. *Information Technology & People*, 19(3), 272-292.

Trauth, E. M., & Niederman, F. (2006). Special issue on achieving diversity in the IT workforce: Issues & interventions. *The DATA BASE for Advances in Information Systems*, 37 (4), 8.

Wheeler, D. L. (2005). Gender sensitivity and the drive for IT: Lessons from the NetCorps Jordan Project. *The Economist*, 8, 131-142.

Wilson, B. C. (2006). Gender differences in types of assignments preferred: implications for computer science instruction. *Journal of Educational Computing Research*, 34(3), 245-255.

Wolfe, J., & Alexander, K. P. (2005). The computer expert in mixed-gender collaborative writing groups. *Journal of Business and Technical Communication*, 19 (2), 135-170.

KEY TERMS

Computer Science: More traditional IT curriculum whose focus is technical and theoretical concepts rather than applied.

Critical Research: Methodological research perspective that approaches a research question with a stated point of view.

Feminist Research: Large body of research that seeks the "ideal of emancipation" of women (Trauth & Howcroft, 2006, p.273), including their relationship to technology.

Interpretive Research: Methodological research perspective that attempts to describe, from a neutral perspective, the factors observed in exploring the research question under consideration.

Information Systems: Curriculum which integrates technical skills and knowledge with applied business and organizational knowledge.

Information Technology: Term that encompasses a range of professional positions requiring at least a baccalaureate degree in CS, IS, or closely related majors.

Positivist Research: Methodological perspective extensively used by physical scientists and some social scientists that considers the researcher separate from the research object(s); best documented through quantitative analysis.

An Intranet within a Knowledge Management Strategy

Udo Richard Averweg
eThekwini Municipality and University of KwaZulu-Natal, South Africa

INTRODUCTION

An Intranet (or internal Web) is a network designed to serve the internal informational needs of an organisation (e.g., a municipality) using Internet concepts and tools (Averweg, 2007; Turban, McLean & Wetherbe, 2004). The cost efficiency of utilizing Internet technology has opened the door for organizations to use this same technology to share information within the organization (Botha, 2004). Information technology (IT) thus plays an important role in organizations. Given that advances in IT have made it easier to acquire, store and disseminate knowledge than ever before, many organizations are employing IT to facilitate sharing and integration of knowledge (Kankanhalli, Tanudidjaja, Sutanto & Tan, 2003). An Intranet is an application of technology within an organization for the purpose of information dissemination, communication, integration, and collaboration (Telleen, 1997).

Knowledge Management (KM) describes "the primary focus of these efforts has been developing new applications of information technology to support digital capture, storage, retrieval and distribution of an organization's explicitly documented knowledge" (Zack, 1999). In this chapter it is argued that, when aligned, organizational strategy and technical resources (e.g., IT) provide a sound framework to support KM within an organization. However, the question arises as to whether an organization is making the best investment in its IT resources and whether it is managing knowledge in the right way. One technical IT resource in an organization is an Intranet.

Every major organizational process should be regularly evaluated and the evaluation should be *purposeful* and *completed* (Debowski, 2006). One method of evaluation is a survey. Debowski (2006) suggests that survey "evaluations take a number of forms … and may be conducted via telephone, e-mail or mailouts". In this study the evaluation selected by the author is e-mail since the purpose and benefits of an e-mail survey justify the cost.

BACKGROUND

There is a need for KM practices in the workplace to enable managers to promote the sharing of knowledge and allow the organization to acquire and retain intellectual capital. For example, eThekwini Municipality in South Africa is "committed to using Information Technology to make a real difference … municipal decisions have to be based on sound research and information management in order to ensure [service] delivery" (eThekwini Municipality, 2006). KM initiatives in organizations are increasingly becoming important as organizations are making significant IT investments in deploying KM systems (Hahn & Subramani, 2000).

INTRANET AND INTRANET TECHNOLOGY

Tiwana and Ramesh (2001) contend that the Intranet is well suited for use as a strategic tool within the domain of KM owing to its ability to support distribution, connectivity and publishing. According to these authors, the Intranet should be seen as integral to an organization's KM system and should therefore be designed and tailored to enhance an organization's knowledge-sharing activities. This rationale raises the question whether an organization's existing Intranet facilitates knowledge-sharing and KM processes. The exploration of this question creates an opportunity for research within a field of application that seems particularly appealing: a metropolitan municipality – eThekwini Municipality in Durban, South Africa. The appropriate context and appeal arose from the fact that the author is situated within the organization's Information Services Department. Furthermore, given eThekwini Municipality's Integrated Development Plan (IDP), this study was considered pertinent and relevant. In surveying the parameters of the question, the overriding premise was established as follows: If knowledge is used effectively, it may well provide meaningful utility to the organization. Clark (2001) notes that "knowledge management initiatives are unlikely to be successful unless they are integrated with business strategy."

Intranets create a common communications and information-sharing system. Brelade and Harman (2003) suggest Intranets can be used on a "push" basis, where information is presented to employees, and on a "pull" basis, where employees may seek out and retrieve information for themselves. These mechanisms are described more fully as follows:

- "Push" technology is used when it is important that certain material is presented to employees at their

workstation. It ensures that no other function takes place until all the information is correctly accessed; and
- "Pull" technology allows employees to decide when to pull down information from the Intranet that they wish to view. The "views of the end users are more important than in most other studies" (Skok & Kalmanovitch, 2005).

To provide a seamless experience between viewing pages on the Web and viewing information on an Intranet, access is usually via a standard Internet browser. The commonly used Internet browser in eThekwini Municipality is Microsoft Internet Explorer.

WHAT IS KNOWLEDGE?

The question of defining knowledge has occupied the minds of philosophers since the classical Greek era and has led to many epistemological debates (Alavi & Leidner, 2001). Given the differing views of knowledge (e.g., a state of mind, an object, a process, a condition of having access to information or a capability), Carlsson, El Sawy, Eriksson, and Raven (1998) suggest that this leads to different perceptions of KM.

Many current theories and practices indicate that knowledge (and the management thereof) may prove useful if the scope and utility of knowledge is aligned with an organization's strategy. For this reason KM must have a business focus. It is therefore critical that KM aligns with the organization's business strategy and that it is structured in such a way that it articulates with the organization's purpose and goals. It may be further argued that knowledge should be viewed as a resource in the business, and that it should therefore tie in with the resource-based approach to strategy.

Although this chapter seeks to review the role of the Intranet and its contribution to a KM strategy, it also proposes that KM should be set on a broader scale than merely IT, that is, it is argued that the management of knowledge should go beyond a narrow technical focus and encompass other less tangible themes within an organization. Zack (1999) clarifies the intangible "as the knowledge existing within people's heads, augmented or shared via interaction and social relationships". This chapter draws together the technology, the notion of shared interaction and the creation of an opportunity for knowledge transfer.

Knowledge Management (KM)

Precisely what is KM? Kwalek (2004) suggests that "the literature on knowledge management is disjointed and disconnected". Pfeffer and Sutton (2000) indicate that KM "tends to treat knowledge as a tangible thing, as a stock or quantity, and therefore separates knowledge as some *thing* from the use of that thing". While there are different views on what KM is, Nomura (2002) suggests that the "objective of KM is to directly enhance corporate value according to business strategy." From a review of the literature and for the purposes of this chapter, the following definition of KM will be adopted: "KM is the organizational process for acquiring, organising and communicating both tacit and explicit knowledge (so that people may use it to be more effective)" (Gray, 2006). The argument for this selection is based on the recognition that the combined knowledge and expertise of people within an organization is what makes an organization unique.

The basic role of technology in KM can be briefly summarised in functional terms into the areas of:

- Facilitating communication;
- Enabling collaboration
- Collecting information;
- Storing information;
- Analysing information;
- Disseminating information; and
- Updating information (Brelade & Harman, 2003).

KM is not a centralized database that contains all the information known by an organization's workers. It is the idea of gaining business insights from a variety of sources – including databases, Websites, employees, and business partners – and cultivating that information wherever it resides into corporate value. Business insight emanates from capturing information and giving it greater meaning via its relationship to other information in the organization. It should be stated that KM is not about making plug-and-play workers dispensable because all they know is recorded for the next person who fills their shoes – it is about delivering information to knowledge workers, business processes and technology to make organizations and people successful and effective. The Intranet, the in-house version of the World Wide Web (the Web) browser based on Internet technology, creates a common corporate communications and information-sharing system (Brelade & Harman, 2003).

eThekwini Municipality in South Africa

eThekwini Municipality comprises six clusters/service units (Office of the City Manager, Treasury, Governance, Sustainable Development and City Enterprises, Corporate and Human Resources and Health, Safety and Social Services) and employs approximately 20,000 employees. The Information Services Department is located in the Office of the City Manager. eThekwini Municipality has some 6,000 networked desktops (personal computers, thin clients and laptops) and electronic communication (i.e., e-mail) takes place via Novell's GroupWise (Client version 6.5). A total of

6 654 GroupWise accounts exist in eThekwini Municipality. There are approximately 1,500 Internet accounts utilising either Microsoft Internet Explorer or Netscape Navigator Web browsers.

Research Methodology and Survey

Since research is varied, disparate approaches are taken. For this research, a mixed-methods research approach is adopted:

- Knowledge claim – pragmatism;
- Strategy of inquiry – transformative procedures; and
- Methods of data collection and analysis – secondary data and analysis are used. The data for eThekwini Municipality's Intranet have recently been collected by an independent research company, Ask Africa. The rationale for using secondary data is that (1) they are considered relevant to the study; and (2) there are savings of time and money by using available data rather than collecting original data.

On June 13, 2006, eThekwini Municipality employees were invited – by e-mail invitation from the Communications Department – to participate in an online Intranet survey. The aim of the survey was "to identify areas where the Intranet may need improvements" and "to allow positive user experiences to be obtained". eThekwini Municipality employees who expressed an interest in participating in this survey received an online questionnaire, which was e-mailed to them by Ask Africa's research partner, MicroIces. Data collation was handled by Ask Africa. The data used in this research are sourced from the eThekwini Municipality Intranet Research Report (July 2006), which was compiled by Ask Africa. The reported findings inform this study.

From the 150 e-mails sent to eThekwini Municipality employees, 39 responses were received. This represents 26 per cent of the total number of employees who originally expressed interest in participating in the survey. Debowski (2006) suggests that response "rates as low as 20% may still provide some sense of the issues". The author did not participate in this online survey.

Extracted from the eThekwini Municipality Intranet Research Report (Ask Africa, 2006), the results are now presented. The ranking in ascending order of Agree/Strongly Agree responses to benefits the Intranet holds is reflected in Table 1.

From Table 1, the greatest perceived benefit that the Intranet holds for employees using it is as a platform to share and access interdepartmental (i.e., clusters/service units) information. The second highest reported benefit was as "an effective way to conduct organizational interaction". Van der Walt, van Brakel, and Kok (2004) emphasized the importance of evaluating an organization's Intranet to ascertain its contribution to potential knowledge-sharing in an organization. The third highest reported benefit was as the quickest "focal point to disseminate and get organizational communication". The lowest reported benefit was for employees to use the Intranet for their daily work functions.

The ranking in ascending order of Agree/Strongly Agree responses to the design of the Intranet is reflected in Table 2.

From Table 2, it appears that most respondents surveyed (86 per cent) were satisfied with the text, font and colours used, but there was some disagreement on the images, pictures and overall design of the Intranet Website. For respondents surveyed, this suggests that images and pictures used on the Website require improvement for eThekwini Municipality employees to obtain user satisfaction (Ask Africa, 2006).

Table 1. Ranking in ascending order of Agree/Strongly Agree responses to benefits the Intranet holds

	Percentage (%) of Respondents (N=19)		
Statement	**Agree/ Strongly Agree**	**Neutral**	**Disagree**
Useful platform to share and access inter-departmental information	87%	9%	4%
The Intranet is an effective way to conduct organizational interaction	81%	14%	5%
Quickest focal point to disseminate and get organizational communication	77%	14%	9%
Enhances departmental communication 7	2% 5	%	24%
Helps the organization improve its service to customers	65%	15%	20%
Helps with productivity	63%	14%	23%
Using the Intranet is necessary for employees to perform daily work functions	50%	5%	45%

(Adapted from eThekwini Municipality Intranet Research Report compiled by Ask Africa (2006: 26))

Table 2. Ranking in ascending order of Agree/Strongly Agree responses to design of Intranet

Statement	Percentage (%) of Respondents (N=21)		
	Agree/ Strongly Agree	Neutral	Disagree
I am happy with the text and font used on the site	86%	5%	10%
I am happy with the colours used on the site	81%	10%	10%
I am happy with the layout and organization of the site	67%	19%	14%
I am happy with the images and pictures used on the site	62%	19%	19%
Overall I am happy with the design of the Intranet Website	57%	33%	10%

(Adapted from eThekwini Municipality Intranet Research Report compiled by Ask Africa (2006: 34))

Table 3. Ranking in ascending order of Agree/Strongly Agree responses to the usability of the Intranet

Statement	Percentage (%) of Respondents (N=20)		
	Agree/ Strongly Agree	Neutral	Disagree
The drop down menus are easy to use	70%	20%	10%
Overall I am happy with the functionality/usability of the site	67%	10%	24%
I am happy with the site labeling	62%	19%	19%
I am happy with the speed of the site	62%	14%	24%
I am able to navigate quickly and easily	50%	20%	30%
The site is self-explanatory – it indicates where I need to go to find the information I am looking for	43%	29%	29%

(Adapted from eThekwini Municipality Intranet Research Report compiled by Ask Africa (2006: 39))

The ranking in ascending order of Agree/Strongly Agree responses to the usability of the Intranet is reflected in Table 3.

From Table 3, it appears that navigation improvements are required. Furthermore, while respondents surveyed agreed that they are able to navigate the Intranet Website quickly and easily, they felt that there was no clear direction provided. This suggests the navigation needs to be improved for beginner users so that they have a better indication of where to go to find the information they are seeking (Ask Africa, 2006).

The ranking in ascending order of Agree/Strongly Agree responses to the content of the Intranet is reflected in Table 4.

From Table 4, it appears that respondents surveyed felt that the information on the Intranet is relevant and reliable. However, improvements in the updating of information and the quality of information-seeking are required. This suggests that while the information on the Intranet Website is generally seen to be reliable, the regular updating of content and finding information that an employee is looking for needs to be improved (Ask Africa, 2006). An important use of most

Table 4. Ranking in ascending order of Agree/Strongly Agree responses to the content of the Intranet

Statement	Percentage (%) of Respondents (N=18)		
	Agree/ Strongly Agree	Neutral	Disagree
The information and content on the Website is relevant	63%	11%	26%
The information on the Website is reliable	61%	17%	22%
Overall I am happy with the quality of content on the Website	57%	14%	29%
I am happy with the quality of the search process	57%	14%	33%
The content on the site is regularly updated	53%	11%	38%
There is a high likelihood of finding information I am looking for even though I do not know where to find it	52%	10%	38%

(Adapted from eThekwini Municipality Intranet Research Report compiled by Ask Africa (2006: 44))

Intranets is to find documents that "point" to employees who have knowledge and expertise. Wells, Sheina, and Harris-Jones (2000) indicate that less than 5 per cent of employee knowledge is actually captured and accessible across the organization. Intranet satisfaction is directly influenced by having the right content, features, and design factors (Kaplan, 2001).

From the survey results, there appear to be areas for improvement in the Intranet design, usability and content areas. The author's evaluation therefore appears to be both purposeful and completed.

FUTURE TRENDS

An Intranet may be classified as a KM application since it is capable of distributing knowledge. An Intranet is seen as a tool for the more efficient sharing and creation of knowledge within organizations, using both "push" and "pull" technologies. However, in the case of eThekwini Municipality's Intranet, it appears that far greater use is made of the "pull" technology (as opposed to "push" technology). This current trend will need to be addressed so that the "pull" technology is also facilitated.

The reported results tend to suggest that there is limited knowledge-sharing and/or KM in eThekwini Municipality. The trend will also need to be addressed so that "knowledge management [is] a planned structured approach to manage the creation, sharing, harvesting and leveraging of knowledge as an organizational asset … [and] should be in line with its business strategy" (Du Plessis & Boon, 2004).

CONCLUSION

Organizations generally make use of one of a variety of methodologies, or a combination of these, for strategy formulation when planning their longer-term interaction with the environment. Knowledge-based strategy places the organization's primary intangible asset, namely the competence of its people, at the centre. Given eThekwini Municipality's IDP and its overall intent to respond to social and economic needs of citizenry, the value of knowledge to organizational effectiveness is crucial at this point. IT, with the enabling role of Intranet technology, should be seen as significantly important to enhance the management of knowledge within eThekwini Municipality. By being aligned to the organizational strategy, the Intranet will provide a sound framework to support KM within the organization and become a strategic tool within a KM strategy.

REFERENCES

Alavi, M. & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Ask Africa (2006). *eThekwini municipality intranet research report*. Unpublished report 1-72, July.

Averweg, U. (2007). Impact of organisational intranets on profitability in organisations. In S. Lubbe (Ed), *Managing information communication technology investments in successful enterprises* (pp. 44-78). Hershey, PA: Idea Group Publishing.

Brelade, S. & Harman, C. (2003). *Knowledge management – The systems dimension*. London, United Kingdom: Thorogood.

Carlsson, S. A., El Sawy, O. A., Eriksson, I., & Raven, A. (1998). Gaining competitive advantage through shared knowledge creation: Search of a New Design Theory for Strategic Information Systems. In *Proceedings of the Fourth European Conference on Information Systems*. Lisbon, Portugal.

Cavaleri, S. & Reed, F. (2000). Designing knowledge generating processes, knowledge and innovation. *Journal of the KMCI, 1*(3), 27-54.

Clark, T. (2001). The knowledge economy. *Education & Training, 43*(4/5), 189-196.

Debowski, S. (2006). *Knowledge management*. Milton, Queensland, Australia: John Wiley & Sons.

Du Plessis, M. & Boon, J. A. (2004). Knowledge management in e-business and customer relationship management: South African case study findings. *International Journal of Information Management, 24*, 73-86.

eThekwini Municipality (2006). *Innovations – Good practice from the eThekwini Municipality, Durban, South Africa.* Corporate Policy Unit (CPU), eThekwini Municipality, Durban.

Gray, P. (2006). *Manager's guide to making decisions about information systems.* Hoboken, NJ: John Wiley & Sons, Inc.

Hahn, J. & Subramani, M. R. (2000). A framework of knowledge management systems: issues and challenges for theory and practice. In *Proceedings of the twenty-first international conference on Information Systems* (pp. 302-312).

Kankanhalli, A., Tanudidjaja, F., Sutanto, J., & Tan, B. C. Y. (2003). The role of IT in successful knowledge management initiatives. *Communications of the ACM, 46*(9), 69-73.

Kaplan, M. (2001). *Intranets and corporate portals: User study.* Agency.com Report. Retrieved June 18, 2008, from http://knowledgemanagement.ittoolbox.com/documents/document.asp?i=1557

Kwalek, J. P. (2004). Systems thinking and knowledge management: Positional assertions and preliminary observations. *Systems Research and Behavioral Science, 21*, 17-36.

Knowledge Workspace. *IBM Systems Journal, 40*(4), 925-941.

Nomura, T. (2002). Design of "Ba" for successful knowledge management: how enterprises should design the places of interaction to gain competitive advantage. *Journal of Network and Computer Applications, 25*(4), 263-278.

Pfeffer, J. & Sutton, R. (2000). *The knowing-doing gap: How smart companies turn knowledge into action.* Boston: Harvard Business School Press.

Skok, W. & Kalmanovitch, C. (2005). Evaluating the role and effectiveness of an intranet in facilitating knowledge management: A case study at Surrey County Council. *Information & Management, 42*, 731-744.

Telleen, S. L. (1997). *Intranet organization*. New York: Wiley Publishers.

Tiwana, A. & Ramesh, B. (2001). Integrating knowledge on the web. *IEEE Internet Computing, 5*(3), 32-39.

Turban, E., McLean, E., & Wetherbe, J. (2004). Information technology for management. *Transforming organizations in the digital economy* (4th ed.) Hoboken: John Wiley & Sons.

Van der Walt, C., Van Brakel, P. A., & Kok, J. A. (2004). Knowledge sharing via enterprise intranets – asking the right questions. *South African Journal of Information Management, 6*(2), 1-12.

Wells, D., Sheina, M., & Harris-Jones, C. (2000). Enterprise portals: New strategies for information delivery, **13**(8), London, England: Ovum.

Zack, M. H. (1999). Developing a knowledge strategy. *California Management Review, **41***(3), 125-146.

KEY TERMS

Information Systems (IS): A combination of technology, people and processes to capture, transmit, store, retrieve, manipulate and display information.

Information Technology (IT): The technology component of an IS.

Intranet: A network designed to serve the internal informational needs of an organisation using Internet concepts and tools.

Knowledge: The understanding, awareness or familiarity acquired through education or experience.

Knowledge Management (KM): KM is the organisational process for acquiring, organising and communicating both tacit and explicit knowledge.

Pull technology: An IT network communication where the initial request for data originates from the client and then is responded to by the server.

Push technology: An IT network communication where the request for data originates with the server.

World Wide Web (the Web): An information space consisting of hyperlinked documents published on the Internet.

Introduction to Basic Concepts and Considerations of Wireless Networking Security

Carlos F. Lerma
Universidad Autónoma de Tamaulipas, Mexico

Armando Vega
Universidad Autónoma de Tamaulipas, Mexico

INTRODUCTION

Local networks have been, from the beginning, a controversial topic. The organizations that have implemented these types of networks have shown their concern about their levels of security. Ever since the discovery of vulnerabilities among first-generation wireless networks (Borisov, Goldberg, & Wagner, 2001), analysts and security companies have tried to understand and mitigate those risks. Some of those efforts have contributed towards the study of wireless security. Other efforts have failed, presented a different group of vulnerabilities, or require expensive proprietary software and hardware. Finally, other efforts try to mitigate the problem piling up a complex group of security technologies, like virtual private networks.

Despite the benefits they bring, a great number of concerns related to security have limited the massive adoption of wireless networks, particularly in sectors that are highly aware of the existing security risks such as the financial and government sectors. Even though there are a significant number of risks inherent to the mass transmission of data to any individual within the boundaries of a wireless network, a good amount of these are installed without any security measure at all. However, the majority of businesses that have implemented some sort of wireless security measures have done so in the most rudimentary way, bringing a false sense of security to users.

When the first IEEE 802.11 wireless standards were in the phase of development, security was not as important as it is today. The level of complexity of network threats was much lower and the adoption of wireless technologies was still in an introductory phase. It was under these circumstances that the first standard for wireless network security, known as wired equivalent privacy (WEP), was originated. WEP underestimated the necessary means to turn air security into an element equivalent to the security provided by a cable. In contrast, the security methods of modern wireless networks are designed to work in hostile environments where there is a lack of well-defined physical network perimeters.

BACKGROUND

Every network environment is susceptible to risks, and wireless networks are not the exception. According to a survey by the Federal Bureau of Investigation of the United States, the only category of threats that shows a significant increase in number of attacks and/or possibility of misuse in the last few years is "wireless network abuse." The broadcasting nature of these networks has turned them into perfect targets for nonauthorized users.

According to Arbaugh (2001), these problems are exacerbated by the myriad of free security-threatening tools widely available for download on the Internet and because of the inherent vulnerabilities of wireless networks themselves. One of the most exploited vulnerabilities is the WEP protocol (Fluhrer, Mantin, & Shamir, 2002; Peikari & Forgie, 2002), which is such a severe problem that many companies have decided to abandon the wireless business.

On the other hand, a good amount of the deployment strategies of wireless networks lack a cohesive and effective integration with the authentication services infrastructure of the organization in which they are implemented (Arbaugh & Shankar, 2002). This common mistake is easy to mitigate, and its correction is evident almost immediately by closing the gap between the number of authorized and unauthorized users. This is evident because authorized users are checked against a database with secure access methods inside the wired network.

In other cases, security problems go beyond the merely technological element (National Institute of Standards and Technology, 2007). Commonly, the lack of planning of the wireless network is a decisive coverage and placement factor. Other elements, such as security policies, access procedures, internal policies governing the use of and access to resources and guidelines governing confidentiality and protection of information serve as a complementary regulatory framework that provides support to the technological infrastructure, establishing limitations related to the way in which information is and/or should be used.

TECHNOLOGICAL ANALYSIS

Wireless networks have experienced a rising trajectory in the last 8 years. Basically, access to wireless technology (access points, wireless network cards) has become easier due to relatively low equipment prices and easiness to set up equipment. Many pieces of network equipment are advertised under the commercial designation SOHO (small office home office), whose installation in inherently simple to carry out due to the fact that the people who purchase those pieces of equipment are relatively new to network equipment installations or users with basic computer skills.

Current advantages of implementing a wireless network include (Planet3 Wireless, 2005):

- Availability: Members of an organization can have access to information resources anywhere without depending on a wired infrastructure.
- Mobility: A user can go from one place to another inside a building or between buildings without leaving the network's coverage area while still having access to network resources. This characteristic trait poses one of the main advantages and threats to wireless networks.
- Productivity: Due to the fact that wireless networks can provide a connection virtually anywhere, this feature enables users to keep working no matter where they are. This feature gains a significant value when wireless networks are implemented in the business sector, due to factors like access to business information and management information systems.
- Ease of installation: A wireless network can be deployed in a matter of minute,s and can be transferred from one place to another as fast as it was set up in the first place. Basically, this advantage tends to be more noticeable in networks with a steady level of permanence and low level of complexity. However, ease of installation does not mean that important security and planning aspects should be omitted, such as carrying out a site survey, and the configuration of security protocols and authentication methods.
- Scalability: Wireless networks can rapidly adapt to an increasing population of users. To achieve this, a network administrator needs less pieces of wireless equipment than wired equipment. Scalability should also be determined before installation of a network, due to the fact that an administrator needs to know an estimated number of the users that will be connecting to the network and the number of pieces of equipment that will support those users, as well as the applications that will be supported by the network.
- Cost: Even though wireless network equipment is more expensive than their wired counterparts, their prices are still reasonable for the home user. Enterprise-scale equipment tends to be extremely expensive, but it comes preloaded with advanced security and management features, is built to be used outdoors, and can support a bigger amount of users.

On the other hand, wireless networks have clear disadvantages that are a result of their own nature (LaRocca & LaRocca, 2002). Generally, they can be:

- Security: Aided with proper equipment and pertinent knowledge concerning the basic operation of a wireless network, any person can capture information that travels through the air, product of a wireless transmission. For an intruder, it is only needed to position oneself within a relative close distance to the network from which one can have an acceptable level of signal reception and possess the tools needed to perform the decryption of the information.
- Distance: The coverage area of wireless networks used today is considered in tens of meters (taking into account that most networks used in infrastructure mode are part of the 802.11 standard). It is necessary to acquire and install accessories and equipment, such as antennas and repeaters, in order to further enhance the coverage area of the network.
- Reliability: Wireless technology is subject to the effects of interference, the environment, and terrain features. Most of the times, an administrator can deal with these phenomena, but it is an undeniable fact that their effects are strongly felt, producing unpleasant results and undermining network performance (in different degrees).
- Speed: Wireless network speed is low by nature, and it is not comparable to those of wired networks, which offer speeds and transfer rates much more advanced than wireless networks. Even though special components can be used to increase the speed and performance of a wireless network, this cannot be translated to higher transfer rates and increased speed, which is still a major disadvantage.

By analyzing disadvantages, it can be concluded that security is, without a doubt, the most important of them all when it comes to wireless networks, and is the one factor that originates the higher amount of challenges. Even though reliability is an area where antenna design and other important areas are put to better use, and are of extreme importance to the development of new generations of networks, security takes a more dynamic posture. A solution has not fully matured when there is a new problem to solve and a new countermeasure to develop. Wireless security is seen as a modular problem, that is, or can be solved by integrating several technologies.

This "integral solution" must be designed according to the security needs of the organization and the information and services that it handles or needs.

The search for solid security mechanisms that allow networks to be deployed in a safe manner is, and will always be, a priority, both for the industry and for researchers around the globe. Currently, the emphasis of wireless security has focused its efforts towards a framework that includes the following key areas (Federal Trade Commission, 2002):

1. General security: The wireless world differs greatly from its wired counterpart. Because wireless devices tend to be shared, and interact with exterior networks and users, passwords have ceased to be an effective method of authentication because of the fact that they are constantly in contact with dangerous and insecure resources. In this sense, wireless networks can benefit from a bipartite method of authentication in which the user makes use of an element he/she has and other element he/she knows. In this case, a USB device, an authentication token or its own fingerprint (read through a biometric device) are elements he/she possesses, while a password is an element he/she knows.
2. Encryption: Wireless networks must be able to support encryption schemes that provide that service in a solid manner, and along the entire communication path, in order to get to its destination. Previously, solutions such as WEP and WPA (Wi-Fi protected access) were proposed to alleviate the security problems that plagued wireless networks, but eventually, they turned out to be vulnerable when the attackers analyzed their mechanisms and published tools that crippled them (Peikari & Forgie, 2002). Currently, WPA2 seems like a viable solution, but it is evident that, with time, a new solution (or solutions) will be necessary (Mehta, 2001).
3. Digital signatures: From the beginning, wireless networks have been (and are already still, up to a certain extent) vulnerable to attacks that replicate or insert forged and/or altered packets in transmission sessions in order to lure users into traps to obtain their confidential information. It is evident that schemes like public key infrastructure (PKI), where integrity and reliability of data can be assured, have been deemed necessary to avoid "man-in-the-middle" type of attacks, where a third party hijacks a communication session and obtains data packets that traverse that communication's channel, forging or altering the information contained in the packets and then fooling the user who was intended to be the final destination of those packets.
4. Interoperability: Due to the fact that options like the exclusive use of WEP or WPA leaves important security holes open to a possible attack, it is evident that security solutions are built by integrating many products (which use, in most cases, proprietary technologies) that must interact with one another to alleviate the different and possible security threats that a network might encounter (Bhagyavati Summers, & DeJoie, 2004). Frequently, there is a possibility that, due to the proprietary nature of the aforementioned technologies, they might not be able to interact one with another in a correct way. This fact has focused the efforts of regulatory bodies and standards all over the world in order to guide the wireless networking industry towards the development of applications and solutions that can operate between them in a native way, and in any IP network platform.

Technological Integration

The implementation of a secure wireless network is founded on the integration of elements that can be classified inside two groups:

- Native 802.11 Security Solutions: Comprised by those solutions that conform, in a natural way, the standards established by IEEE's 802.11 workgroup: WEP, TKIP, 802.1x/EAP, WPA and WPA2 (Giller & Bulliard, 2004).
- Independent 802.11 Solutions: Comprised by all those technologies that are implemented "on top" of natural 802.11 solutions. Generally, they are proprietary, third-party solutions with different degrees of interoperability: VPNs, LDAP-based directory services (Microsoft, 2007), captive portals, physical configurations, and physical network design considerations (site surveys).

Once again, final wireless network implementations depend heavily on the privacy and confidentiality requirements of the organization, product of a previous network analysis. The possibility of mitigating multiple problems with a single tool is very distant from becoming a reality.

FUTURE TRENDS

Even though the future trends, in regards of security for the next generation of wireless networks, is very promising, companies need to continue the implementation of security countermeasures in order to guarantee the efficacy of their security programs aimed at wireless networks. Future work should be aimed towards standardization of security solutions in such a way that, any security solution can operate in an integral way with other solutions without altering the basic essence of any standard or protocol.

The basic security component inside SOHO equipment is also a security niche that should be addressed in the future. Since most security breaches happen in SOHO networks, manufacturers should aim their efforts into security solutions that are easy to implement, and whose complexity level is low. A solution to this issue is an area where the industry should focus its research, since it represents a good market opportunity and eventual future strategic advantage, as we see that enterprise equipment is hard to provide with quite hardened security countermeasures. However, since the industry is shifting towards more affordable types of equipment, this should be a priority in terms of security design and features in the near future.

CONCLUSION

The security-related concerns that have plagued wireless networks in both the enterprise and SOHO sectors finally seem to dissipate with security schemes that range from encryption methods using static keys, all the way through several improvements in dynamic encryption keys, and authentication solutions carried out in free software platforms (open source) and extensive solutions like VPN or captive portals. The security capabilities and levels of wireless networks are finally matching those provided by wired networks, due to the adoption and use of current available standards, certifying organisms like IEEE and the Wi-Fi Alliance, and integrated solutions, whether they might be open or proprietary, providing access control and encryption of the information that travels through the air.

Even though the first versions of wireless networks were not designed with security in mind (IEEE, 1999), the amount of security methods and solutions is growing extremely fast. With newly introduced solutions like 802.1x and 802.11i, there are now readily available and secure solutions in terms of encryption and authentication standards. These security features must be implemented in order to assure the integrity of the information inside a wireless network. With careful planning and analysis phases carried out well in advance of its implementation, a wireless network can be as secure as one of its wired alternatives. It is here where the human factor is as important as the technological.

REFERENCES

Arbaugh, W. (2001). *An inductive chosen plaintext attack against WEP/WEP2*". IEEE Document 802.11-02/230, May 2001; grouper.ieee.org/groups/802/11.

Arbaugh, W., & Housley, R. (2003). Security problems in 802.11-based networks. *Commun. ACM, 46*, 5.

Arbaugh, W.A., Shankar, N., & Wan, Y. C. (2002). Your 80211 wireless network has no clothes. IEEE Wireless Communications, 9, 44 – 51.

Bhagyavati, Summers, W. C., & DeJoie, A. (2004). Wireless security techniques: An overview. In *Proceedings of InfoSecCD '04: 1st Annual Conference on information Security Curriculum Development* (pp. 82-87). New York, NY: ACM Press.

Borisov, N., Goldberg, I., & Wagner, D. (July 2001). Intercepting mobile communications: The insecurity of 802.11. In *Proceedings of the International Conference on Mobile Computing and Networking* (pp. 180–189).

Federal Trade Commission. (2002). 16 CFR Part 314; Standards forsSafeguarding customer information. Final Rule, 2002, May 23. *Federal Register, 67*(100), 36484-36494. Retrieved on 25 May 2007, from http://www.ftc.gov/os/2002/05/67fr36585.pdf

Fluhrer, S., Mantin, I., & Shamir, A. (2002). Attacks on RC4 and WEP. RSA Laboratories, Cryptobytes, 5(2).

Fluhrer, S., Mantin, A., & Shamir,A. (2001). Weaknesses in the key scheduling algorithm of RC4. In Proceedings of the 8th Workshop on Selected Areas in Cryptography, LNCS 2259. Springer-Verlag.

Giller, R., & Bulliard, A. (2004). Security protocols and applications 2004: Wired equivalent privacy. Swiss Institute of Technology, Lausanne, March. 3.

IEEE. (1999). Std 802.11, *Standards for local and metropolitan area networks: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications.*

LaRocca, J., & LaRocca, R. (2002). *802.11: Demystified*. New York: McGraw-Hill.

Mehta, P. C. (2001). Wired equivalent privacy vulnerability. LevelOne Security Essentials Track.

Microsoft. (2007). *Choosing a strategy for wireless LAN Security*. Retrieved on 12 June 2007, from http://www.microsoft.com/technet/security/guidance/cryptographyetc/peap_int.mspx

National Institute of Standards and Technology. (n.d.). Wireless network security. 802.11, bluetooth and handheld devices. Retrieved on 21 December 2007, from http://csrc.nist.gov/publications/nistpubs/800-48/NIST_SP_800-48.pdf

Peikari, C., & Forgie, S. 2003). Cracking WEP. (Retrieved on 26 July 2007, from http://www.airscanner.com/publications.html#articles

Planet3 Wireless. (2005). CWNA certified wireless network administrator official study guide (3rd ed.). McGraw-Hill/Osborne.

KEY TERMS

802.11: Set of standards, established by the IEEE (Institute of Electrical and Electronics Engineers, a worldwide standardization body), that govern the functioning and design of wireless networks communicating in the 5GHz and 2.4 GHz unlicensed bands. It has evolved into many other specifications, but 802.11 still dictates the basic standards for what should be considered as a wireless network.

Access Point: A network device that serves as a central point of connection for devices with wireless networking capabilities. Similar to a network hub, it performs similar functions, being the difference that a wired hub uses a medium access control based in collision detection and an access point uses a collision avoidance method.

LDAP: Acronym for lightweight directory access protocol. A protocol used to send queries to user databases organized through directory services. Wireless networks use this protocol to communicate with servers housing user databases in a secure way, thus providing a consistent method of user authentication on wireless networks whose access is restricted.

SOHO: Acronym for small office home office. Types of network equipment intended for home use or offices with small amounts of employees. Generally, they offer most of the features and functionalities found in enterprise-class hardware, but are limited in certain aspects, and their price is much cheaper and competitive.

VPN: Acronym for virtual private network. Secure communication channels established from one network to another or inside one network. VPNs provide dedicated communications channels that offer higher levels of security and encryption. They can be established using a VPN server (software) or through a VPN concentrator (hardware). They assure that communications are somewhat free from eavesdropping.

WEP: Acronym for wired equivalent privacy. An encryption algorithm intended to secure the first generation of wireless networks. It uses a 40-bit key alongside a 24-bit initialization vector, originating an RC4 key. It proved to be useless in 2001, when it was cracked, and the cracking tools were made available on the Internet.

WPA: Acronym for Wi-Fi protected access. An encryption algorithm created by the Wi-Fi Alliance. Like WEP, it uses an RC4 encryption key, but WPA's key is longer: a 128-bit key alongside a 48-bit initialization vector. WPA also comes in two variations: enterprise and personal. The difference relies on the fact that enterprise is a method that relies in an 802.1x server to provide authentication services, while personal relies on a preshared key scheme.

WPA2: Acronym for Wi-Fi orotected access 2. It is also known as 802.11i. It is basically the same as the WPA standard, improved with a new algorithm to distribute encryption keys.

Category: IT Security & Ethics

Intrusion Detection Based on P2P Software

Zoltán Czirkos
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The importance of the network security problems come into prominence by the growth of the Internet. The article presents a new kind of software, which uses just the network, to protect the hosts and increase their security. The hosts running this software create an *Application Level Network* (ALN) over the Internet. Nodes connected to this ALN check their operating systems' log files to detect intrusion attempts. Information collected is then shared over the ALN to increase the security of all peers, which can then make the necessary protection steps by oneself.

The developed software is named *Komondor* (Czirkos, 2006), which is a famous Hungarian guard dog. The novelty of the system Komondor is that Komondor nodes of each host create a *Peer-To-Peer (P2P) overlay network*. Organization is automatic; it requires no user interaction. This network model ensures stability, which is important for quick and reliable communication between nodes. By this build-up, the system remains useful over the unstable network.

The use of the peer-to-peer network model for this purpose is new in principle. Test results proved its usefulness. With its aid, real intrusion attempts were blocked. This software is intended to mask the security holes of services provided by the host, not to repair them. For this it does not need to know about the security hole in detail. It can provide some protection in advance, but only if somewhere on the network an intrusion was already detected. It does not fix the security hole, but keeps the particular attacker from further activity.

BACKGROUND

The P2P networks comprise hundreds of thousands or millions of peers. That is why they are characterized by large dynamism, with a continuous process of nodes joining or leaving the P2P overlay.

Such large scale dynamism introduces several development problems. Neither a central authority nor a fixed communication topology can be employed to control the different components. Instead, a dynamically changing overlay topology is maintained and the maintenance is completely decentralized. The overlay is defined by links among nodes that are created and deleted based on the requirements of the particular application (Montresor, 2004).

Variability of P2P networks can be leveraged by implementing virtual networks based on super-peers. In the meantime, widely-used file-sharing systems such as Kazaa have applied the use of super-peers to enhance their search performance. In the field of the super-peer networks, the main focus is on centralized design of such networks (Yang & Garcia-Molina, 2003).

Until recently, most of the P2P applications deployed on the Internet had not any sophisticated mechanism for enforcing a particular overlay topology. The consequence of this was the adoption of simple communication models, such as flooding. Currently the situation has changed; many research projects have proved the importance of selecting, and proposed constructions and maintenance of appropriate topologies for robust P2P systems (Rowstron, & Druschel, 2001). Even popular file-sharing applications have started to consider more structured topologies (Kan, 2001). By introducing the concept of super-peer, their topologies are now organized through a two-level hierarchy. Nodes that are faster and/or more reliable than the ordinary nodes take on server-like responsibilities and provide services to a set of clients. A good example for this is the case of file sharing, where a super-peer builds an index of the files shared by its clients and participates in the search protocol on their behalf, leveraging them from participating in complicated protocols and reducing the overall traffic by forwarding queries only among super-peers.

The super-peer concept allows decentralized networks to run more efficiently by exploiting heterogeneity and distributing load to machines that can handle the burden. Also, it does not inherit the flaws of the client-server model, as it allows multiple, separate points of failure, increasing the robustness of the P2P network.

The applicability of the super-peer model is not limited to file-sharing, that is, it is possible to envisage distributed game systems (Smed, Kaukoranta, T., & Hakonen, 2003). In this case, multiple locations of a simulated virtual environment can be maintained by a distributed set of super-peers that control the virtual environment on behalf of their clients. Grid management systems and distributed storages are other

good possibilities for the usages of this architecture (Foster & Kesselman, 1999).

The construction and maintenance of a super-peer topology is, however, a difficult task. The extreme scale and dynamism call for robust and efficient protocols, capable to self-organize and self-repair a super-peer overlay in spite of both voluntary and unexpected events like joins, leaves and crashes. Another problem arises that in a P2P gaming, nodes must be clients of a single super-peer, and the sub-topology of super-peers must reflect the characteristics of the particular virtual environment that is simulated.

In order to avoid some problems that arise according to the super-peer model, Montresor (2004) proposed an enhanced protocol for the construction and management of super-peer-based overlay topologies. This protocol is based on the so-called gossip paradigm (Eugster, Guerraoui, Kermarrec, & Massoulié, 2003). In this case, every node periodically initiates an information exchange with a peer node selected randomly. The nodes involved in the exchange send each other information about their current status; whether they are super-peers or simply clients, the number of clients they are serving, and so on. Based on this information, role changes and/or client transfers can occur. A client may decide to become a super-peer and take responsibility for some of the clients of the other node to alleviate its load. Alternatively, a super-peer may decide to move all its clients to the other node and become a client by itself, to reduce the number of super-peers and, thus, the traffic generated by communication between super-peers.

THE SECURITY PROBLEMS OF P2P COMMUNICATION

The popular software, the *Instant Messaging* (IM) is the fastest increasing communications medium with an estimated 300 million consumer and enterprise IM users in 2005 (IMlogic, 2005). Global services such as AOL Instant Messenger, MSN Messenger, and Yahoo! Messenger each report over 1 billion messages sent per day, and IM traffic is expected to exceed email traffic by the end of 2006. As one of the most successful and widely-deployed applications on the Internet, IM has increasingly become the target for attackers to distribute IM-borne viruses, spyware software, worms, *SPam over IM* (SPIM), and malware attacks. Though widespread in adoption, IM is usually not protected and in user and corporate environments, leaving it vulnerable to attacks and exploits. These threats have grown exponentially over the past few years, increasing the need for real-time threat response for IM and P2P systems. As use of instant messaging clients and P2P networking increases, new viruses and other malware software are increasingly applying these mechanisms to disseminate.

Recently, the IM worms are more and more sophisticated and cross over from one network to another. In 2005, worms have been detected that are propagating over the Microsoft MSN public network and are crossing over into internal enterprise IM deployments, including Microsoft Live Communications Server environments (IMlogic, 2005). The growing prevalence of public to private network hopping by IM malware and worms will most likely increase in the near future, especially as IT organizations leverage public IM or connectivity to the public IM networks. The threats of multilingual worms are also growing, indicating a large sophistication for disseminating IM worms across geographic areas.

In order to monitor the network and to recognize the activity of the malware applications, a lot of different monitoring systems have been developed. One of them is the Netmon, which is a comprehensive network monitoring appliance that gives a complete perspective of the user's network (Netmon, 2005). Using Netmon, many kinds of malware can be identified, including worms, browser toolbars and plug-ins, and so forth. The Netmon can identify the activity of other types of network software, for example, P2P file sharing applications like KaZaA, E-Donkey and BitTorrent.

The Netmon apply the *Simple Network Management Protocol* (SNMP) to monitor the usage of many types of core network devices, including routers, gateway, firewalls, and switches. The implementation of the SNMP has two parts: one is the SNMP manager, and the other is the SNMP agent. While the SNMP manager is used by the network administrator, the SNMP agents are resided in the network devices to be monitored. The main task of the SNMP is extracting information from the *Management Information Bases* (MIB), which are maintained by the SNMP agents.

SNMP can also be applied for monitoring a lot of other kinds of devices, including network printers, hubs, and more. From the security viewpoint, the Netmon has an important feature, namely, its flexible port scanning tools. It can determine which ports are open for requests. The user is able to compare earlier scan results with the latest ones to spot newly opened ports.

TYPES OF PROTECTION

Security of a system can be increased with strict rules of usage. Applications doing this locally are Bastille-Linux (2006) and SeLinux (2006). These applications provide security mechanisms built in to the operating system or the kernel. Network security is also enhanced with these, but that is not the main purpose of them.

Protection of Host

No system can be completely secure. The literature defines the term of a properly skilled attacker (Toxen, 2001), a theoretical person, who by his infinite skills, can explore any existent security hole. We know that a completely secure system cannot be built. Every hidden bug of a system can be found, either systematically, or accidentally.

An average user has not much to do about the security of his system. Storing sensitive and valuable data (for example a bank company) demands dealing with security problems. Several companies offer *security management* to customers (Bauer, 2003). It is important to know that the more secure a system is, the more difficult the use it, and the less usability it has. A trade-off between security and usability has to be made (Bauer, 2005). Before initiating medium and large sized systems it is worth making up a so-called *security policy*.

Protection of Network

The simplest style of network protection is a firewall**.** This is a host which provides a strict gateway to the Internet for a subnetwork, checking traffic and maybe dropping some network packets. The three main types of firewalls are the following:

1. **Packet level firewalls**
   In this one, filtering rules are based on packet hearers, for example, the address of the source or the destination.
2. **Application level firewalls**
   These examine not only the header, but also the content of the network packets, to be able to identify unwanted input. They can also be used for an adaptive supervision of an application program.
3. **Personal firewalls**
   It is used usually for workstations and home computers. With these, the user can define access to the network for which running applications should be granted.

Intrusion Detection

Computer intrusion detection has three main types, which are the following:

- Traffic signatures (data samples) implying an intrusion,
- Understanding and examining application level *network protocols, and*
- Recognizing signs of *anomalies* (non-usual functioning).

Unfortunately, not every attack comes along with easily, automatically detectable signs. For example, the abusing of a system by an assigned user is hard to notice.

Data Acquisition

For accurate intrusion detection, authoritative and complete information about the system in question is needed. Authoritative data acquisition is a complex task on its own. Most of the operating systems provide records of different users' actions for review and verification. These records can be limited to certain security events, or can provide a list of all system calls of every process. Similarly, gateways and firewalls have event logs of network traffic. These logs may contain simple information like opening and closing network sockets, or may be the contents of every network packet recorded, which appeared on the wire.

The quantity of information collected has to be a trade-off between expense and efficiency. Collecting information is expensive, but collecting the right information is important, so the question is which types of data should be recorded.

Detection Methods

Supervising a system is only worth this expense if the intrusion detection system also analyzes the collected information. This technology has two main types: *anomaly detection* and *misuse detection*.

Anomaly detection has a model of a properly functioning system and well behaving users. Any deviation it finds is considered a problem. The main benefit of anomaly detection is that it can detect attacks in advance. By defining what is normal, every break of the rules can be identified whether it is part of the *threat model* or not.

The disadvantages of this method are frequent false alerts and difficult adaptability to fast-changing systems.

Misuse detection systems, practically speaking, define what is wrong. They contain intrusion definitions and alias *signatures*, which are compared with the collected supervisory information, searching for the signs of the known threats.

The advantage of these systems is that investigation of already known patterns rarely leads to false alerts. At the same time, it can only detect known attack methods, which have a defined signature. If a new kind of attack is found, the developers have to model it and add to the database of signatures.

ALGORITHMS USED IN THE DEVELOPED SYSTEM

This section describes basic security concepts, dangers threatening user data, and computers. We describe different means of attacks and their common features one by one, and also show the common protection methods against them. For the demonstration of the overlay network used in the Komondor, we also review the theory of P2P networks and the newest results of its research. Computers connected to networks are to be protected by different means (Kemmerer & Vigna, 2002), described in detail as follows.

Information stored on a computer can be personal or business character, private or confidential. An unauthorized person can, therefore, steal it. Its possible cases are shown in the *Table 1*.

We have to protect not only our data, but also our resources. Resources are not only hardware. A typical type of attack is to gain access to a computer to initiate other attacks from it. This is to make the identification of the attacker more difficult, because this way the next intruded host in this chain sees the IP address of the previous one as its attacker.

Intrusion attempts, based on their purpose, can be of different methods. But these methods share things in common, scanning networks ports or subnetworks for services, and making several attempts in a short time. This can be used to detect these attempts and to prepare for protection.

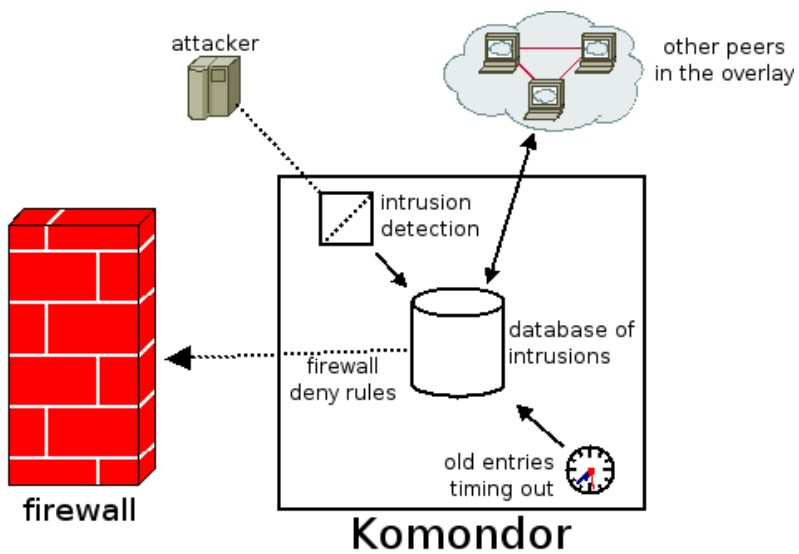
Through security holes, the attacker can also be trying to find resources. With this type of action, whole ranges of network addresses are scanned for a particular service having a bug or just being badly configured. The port number is fixed here. An example for this is scanning for an open e-mail (SMTP) relay to send junk mail anonymously.

The inner architecture of Komondor is presented on Figure 1. Different hosts run the uniform copies of this program, monitoring the occurring network intrusion attempts.

Table 1. The types of the information stealth

• An unauthorized person gains access to a host. • Monitoring or intercepting network traffic by someone. • Abuse of an authorized user.

Figure 1. Architecture of the Komondor system



Komondor nodes protect each other this way. If an intrusion attempt was recorded by a node, the other ones can prepare for the attack in advance.

Information about intrusion attempts is collected by two means: intrusion detected by this node, or intrusion detected by another node. The first working version of Komondor monitors system log files to collect information. These log files can contain various error messages, which may refer to an intrusion attempt. Possible examples are login attempt with an inexistent user name, or several attempts to download an inexistent file through a HTTP server.

Figure 2. Algorithm of the Komondor entity

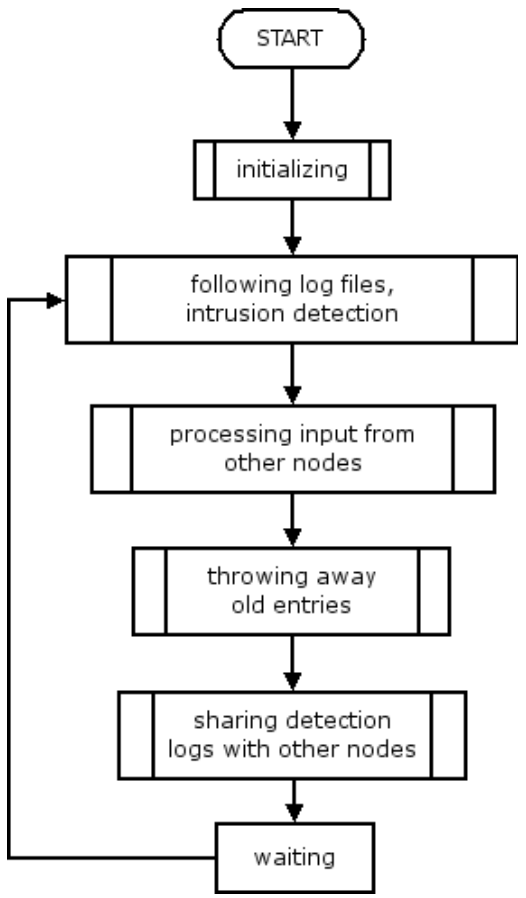


Figure 2 presents the algorithm of the software Komondor. It checks the log files every second, while the database should be purged only on an hourly or daily basis (Czirkos, 2005).

If a node has an empty connection slot, it tries to find other connection points (Hosszú, 2005). A list of active nodes can be downloaded from an anchor point. From this list, a connection is chosen randomly and the ones already tried to connect are deleted from the list. If it becomes blank, the node requests more addresses from the anchor server.

Each peer connects to *at least three; at most five* other nodes. These parameters can be easily adjusted in the program, to set the density of the overlay mesh. The denser the overlay is, the faster the database share and the more stable the system is.

Running of this module is continuous. After starting a Komondor process, it does not wait to connect to the overlay. Following system log files should be begun as soon as possible to be able to detect intrusions. If the node cannot connect to the overlay before the first intrusion is detected, the entry in the database is labeled as "unpublished." When at least one connection is up, then it is possible to send data to the overlay.

RATING OF THE NEW SYSTEM

The first version of Komondor network aided security enhancement system was under extensive testing for months. Parts of these examinations were simulated, and others were real intrusion attempts. This chapter is about some real intrusion attempts, which were averted by the Komondor system. *Table 2* summarizes some of the results achieved.

Effectiveness of the Komondor system is determined by the diversity of peers. Intrusion attempts attacking security holes are software and version specific. Daemons providing services (SSH, Web) could be of different versions on hosts running Komondor. An attempt detected on any (either buggy or not) system can also protect vulnerable ones. Three possible cases in which this can occur are listed in the *Table 3*.

Table 2. Results

Time of attack	Service under attack	Answer
2005-10-17	Secure shell	blocked IP
2005-06-17	Apache httpd	blocked IP
2005-09-20	Monkey httpd	blocked IP

Table 3. Examples of the heterogeneity of the Komondor's environment

• Hosts running the same software, but different version (Apache 2.0.54 and Apache 2.0.30). • Hosts providing the same service, but with means of different software (web server Apache and Zeus). • Hosts are based on different operating systems (Linux and Windows for example).

FUTURE TRENDS

In the second half of 2005, 54 day elapsed between the disclosure of a vulnerability and the release of a patch by the vendor. Security threats tend to emerge, as *exploits* are published in a much shorter time (Symantec, 2005). Development of firewall tools like Komondor is therefore important, while through blocking initiated attacks, they can even enable a vulnerable application to run safely.

We plan enhancing efficiency of the Komondor overlay by merging it with Snort. Also algorithms of present *P2P networks* mainly aim distributing resources and finding specific peers in the overlay. In contrast to this, our application demands a fast broadcast of information: data of detected intrusions. Therefore, we conduct research on selection of a particular P2P topology for this task.

CONCLUSION

This novel application of P2P theory helps the users of this system to increase security of their hosts. The system is easy to use; the nodes organize the P2P overlay automatically, and do not need any user interaction. The intrusion protection of the networked computers has obvious importance and introducing the cooperation among the hosts to be protected is a promising approach.

After reviewing the most important features of the P2P networks and their security related problems, the article presented the most common intrusion detection methods. Finally, a novel approach and the results of an ongoing research were introduced to utilize together the advantages of the P2P paradigm and the enhancements in the security methods of the operating systems. The presented experimental results proved the usefulness of the proposed method.

REFERENCES

Bastille-Linux (2006). Retrieved January 4, 2006, from http://www.bastille-linux.org/.

Bauer, M.D. (2003). *Building Secure Servers with Linux*. Hungarian Edition, Budapest, Hungary: Kossuth Publisher.

Bauer, M.D. (2005). *Linux Server security*. Second Edition, O'Reilly, 542 pages.

Czirkos, Z. (2005). Development of P2P Based Security Software, *Proceedings of the Conference of Scientific Circle of Students*, (Second Award). Budapest: November 11, 2005 (in Hungarian).

Czirkos, Z. (2006). *Komondor homepage*. Retrieved May 3, 2006, from http://jutas.eet.bme.hu/

Eugster, P.T., Guerraoui, R., Kermarrec, A.M., & Massoulié, L. (2003). From Epidemics to Distributed Computing. *IEEE Computer*, 21(4), 341-374.

Foster, I. & Kesselman, C. (1999). *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan Kaufmann.

Hosszú, G. (2005). Mediacommunication Based on Application-Layer Multicast. In Dasgupta, S. (Ed.), *Encyclopedia of Virtual Communities and Technologies* (pp. 302-307). Hershey, PA: Idea Group Reference.

IMlogic. (2005). Q3 2005 IM Security Threat Report. *IMlogic Threat Center. IMlogic, Inc*. Retrieved February 2, 2006, from http://www.imlogic.com/.

Kan, G. (2001). Gnutella. In A. Oram (Ed.), *Peer-to-Peer: Harnessing the Benefits of a Disruptive Technology* (Chapter 8). O'Reilly & Associates.

Kemmerer, R.A., & Vigna, G. (2002). Intrusion Detection: A Brief History and Overview. Security & Privacy-2002. *Supplement to Computer Magazine*. IEEE Computer Society, pp. 27-30.

Montresor, A. (2004). A Robust Protocol for Building Superpeer Overlay Topologies. *Proceedings of the Fourth International Conference on Peer-to-Peer Computing (P2P'04)*. Retrieved January 26, 2006, from http://www.cs.unibo.it/.

Netmon (2005). *Introducing a comprehensive network monitoring solution for today's enterprise*. Retrieved January 26, 2006 from www.netmon.ca.

Rowstron, A. & Druschel, P. (2001). Pastry: Scalable, Decentralized Object Location and Routing for Large-Scale Peer-to-Peer Systems. *In Proceedings of the 18th International Conference on Distributed Systems Platforms*. Heidelberg, Germany.

SeLinux (2006). *Security-Enhanced Linux*. Retrieved January 5, 2006, from http://www.nsa.gov/selinux/.

Smed, J., Kaukoranta, T., & Hakonen, H. (2003). Networking and Multiplayer Computer Games–The Story So Far. *International Journal of Intelligent Games & Simulation, 2(2)*.

Symantec Corporation (2005). *Symantec Internet Security Threat Report*. Trends for January 05 – June 05. Volume VIII, Published 2005.

Toxen, B. (2001). *Real World Linux Security*. Prentice Hall PTR.

Yang, B. & Garcia-Molina, H. (2003). Designing a Super-peer Network. *In Proceedings of the 19th International Conference on Data Engineering (ICDE)*. Bangalore, India.

KEY TERMS

Application Level Network (ALN): The applications, which are running in the hosts, can create a virtual network from their logical connections. This is also called *overlay network* (*see below*). The operations of such software entities are not able to understand without knowing their logical relations. The most cases this ALN software entities use the *P2P model (see below)*, not the *client/server* (*see below*) one for the communication.

Client/Server Model: A communicating way, where one host has more functionality than the other. It differs from the *P2P model* (*see below*).

Exploit: A small program which is designed specifically to attack a certain vulnerability in a system. These are dangerous, while their use requires no skills, and they are usually published shortly after a disclosure of a vulnerability.

Firewall: This is a host or router which provides a strict gateway to the Internet for a subnetwork, checking traffic and maybe dropping some network packets.

Overlay Network: The applications, which create an *ALN (see above)* work together and usually follow the *P2P communication model (see below)*.

Peer-to-Peer (P2P) Model: A communication way where each node has the same authority and communication capability. They create a virtual network, overlaid on the Internet. Its members organize themselves into a topology for data transmission.

Security Management: It means the calculation of the damage caused by a certain attack in advance so one can decide if a particular security investment as buying new devices or training employees is worth it or not.

Security Policy: It means a set of rules in which the expectations and provisions of usage for the users, and the administrators also, is made up. It is worth making up before initiating medium or large sized systems.

Simple Network Management Protocol (SNMP): It is a network management protocol used primarily for IP networks. Using it, the networked devices as routers and host, can be monitored and managed.

Intrusion Tolerance in Information Systems

Wenbing Zhao
Cleveland State University, USA

INTRODUCTION

Today's information systems are expected to be highly available and trustworthy — that is, they are accessible at any time a user wants to, they always provide correct services, and they never reveal confidential information to an unauthorized party. To meet such high expectations, the system must be carefully designed and implemented, and rigorously tested (for intrusion prevention). However, considering the intense pressure for short development cycles and the widespread use of commercial off-the-shelf software components, it is not surprising that software systems are notoriously imperfect. The vulnerabilities due to insufficient design and poor implementation are often exploited by adversaries to cause a variety of damages, for example, crashing of the system, leaking of confidential information, modifying or deleting of critical data, or injecting of erroneous information into a system.

This observation prompted the research on intrusion tolerance techniques (Castro & Liskov, 2002; Deswarte, Blain, & Fabre, 1991; Verissimo, Neves, & Correia, 2003; Yin, Martin, Venkataramani, Alvisi, & Dahlin, 2003). Such techniques can tolerate intrusion attacks in two respects: (1) a system continues providing correct services (may be with reduced performance), and (2) no confidential information is revealed to an adversary. The former can be achieved by using the replication techniques, as long as the adversary can only compromise a small number of replicas. The latter is often built on top of secrete sharing and threshold cryptography techniques. Plain replication is often perceived to reduce the confidentiality of a system, because there are more identical copies available for penetration. However, if replication is integrated properly with secrete sharing and threshold cryptography, both availability and confidentiality can be enhanced.

BACKGROUND

In this section, we introduce some basic security and dependability concepts and techniques related to intrusion tolerance. A secure information system is one that exhibits the following properties (Pfleeger & Pfleeger, 2002):

- **Confidentiality:** Only authorized users have access to the information.
- **Integrity:** The information can be modified only by authenticated users in authorized ways. Any unauthorized modification can be detected.
- **Availability:** The information is available whenever a legitimate user wants to access it.

Confidentiality is often ensured by using encryption, authentication, and access control. Encryption is a reversible process that scrambles a piece of plaintext into something uninterpretable. Encryption is often parameterized with a security key. To decrypt, the same or a different security key is needed. Authentication is the procedure to verify the identity of a user that wants to access confidential data. Access control is used to restrict what an authenticated user can access.

Information integrity can be protected by using secure hash functions, message authentication code (MAC), and digital signatures. For data stored locally, including the application binary files, a checksum is often used as a way to check data integrity. The checksum can be generated by applying a one-way secure hash transformation on the data. Before the data is accessed, one can verify its integrity by recomputing the checksum and comparing it with the original one. The integrity of a message transmitted over the network can be guarded by a MAC. A MAC is generated by hashing on both the original message and a shared secret key. If it is tampered with, the message can be detected in a way similar to that for the checksum. For stronger protection, a message can be signed by the sender. A digital signature is produced by first hashing the message using a secure hash function, and then encrypting the hash using the sender's private key.

High availability is achieved by using replication, checkpointing, and recovery techniques. Replication is a technique that relies on running redundant copies of an application so that if one copy fails, the services can be provided by the remaining copies. Checkpointing means to take a snapshot of the state of a replica. The saved state can be used to bring a new or a restarted replica up to date. Checkpointing is also useful to avoid log buildup (when a checkpoint is taken, all previous logs can be garbage collected). Recovery techniques concern the tasks of removing faulty replicas, repairing them, and reintegrating them back to the system.

INTRUSION TOLERANCE TECHNIQUES

Intrusion tolerance is built on two fundamental techniques: replication and secret sharing/threshold cryptography (Deswarte et al., 1991). In the context of intrusion tolerance, a very general fault model must be used because a compromised replica might exhibit arbitrary faulty behaviors. Such a fault model is often termed as Byzantine fault (Lamport, Shostak, & Pease, 1982).

Byzantine Fault Tolerance

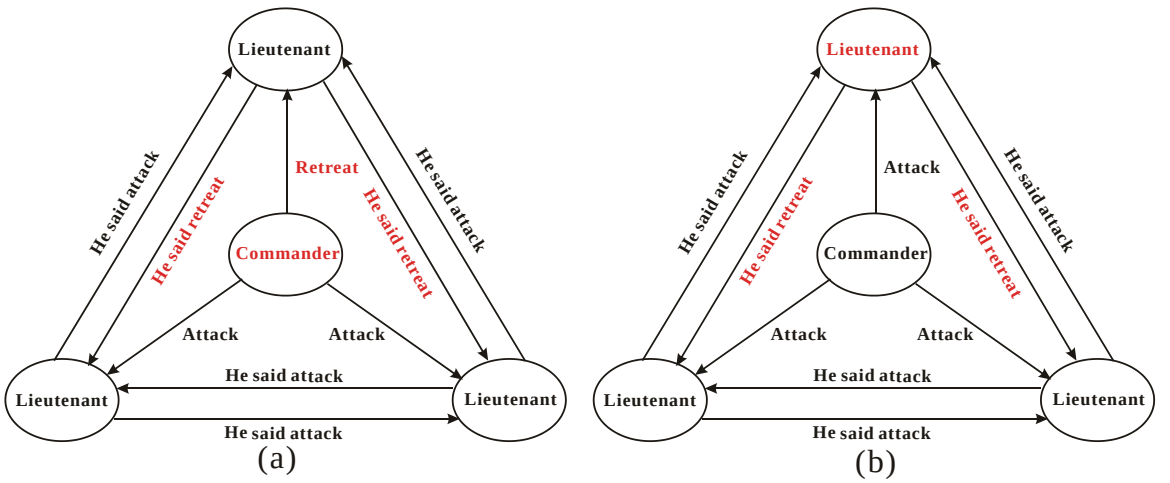
An intrusion attack might bring a service down or compromise the integrity of a service. An effective defense is to introduce redundancy into the system — that is, to replicate critical components in the system. Assuming that an intrusion attack can only penetrate a small fraction of the replicas, the service availability and integrity can be preserved by the remaining correct replicas. However, achieving this goal is not trivial — we must ensure consistent execution of all correct replicas despite the attacks launched by faulty replicas.

A Byzantine faulty replica may use all kinds of strategies to prevent the normal operations of the replicated service, in particular, it might propagate conflicting information to other replicas or components that it interacts with. To tolerate f Byzantine faulty replicas in an asynchronous environment, we need to have at least $3f+1$ number of replicas (Castro & Liskov, 2002). An asynchronous environment is one that has no bound on processing times, communication delays, and clock skews. Internet applications are often modeled as asynchronous systems. Usually, one replica is designated as the primary and the rest are backups.

There are two different approaches to Byzantine fault tolerance. In a Byzantine quorum system (Malkhi & Reiter, 1997), read and write operations issued by some clients are applied on a set of data items (which consists of the state of a service). It is assumed that the read and write operations are synchronized. A read operation retrieves information from a quorum of correct replicas, and a write operation applies the update to a quorum of correct replicas. In a system with $3f+1$ replicas, a quorum can be formed by $2f+1$ replicas so that any two quorums overlap by at least $f+1$ replicas, among which at least one is not faulty. This guarantees the correct operations of the quorum-based system.

A more general method is the state-machine-based approach (Schneider, 1990), in which a replica is modeled as a state machine. The state change is triggered by remote invocations on the methods offered by the replica. This approach is applicable to a much wider range of applications. Consider a client server application where the server is replicated using the state-machine-based approach (Castro & Liskov, 2002). The client first sends its request to the primary replica. The primary then broadcasts the request message to the backups and also determines the execution order of the message. To prevent a faulty primary from intentionally delaying a message, the client starts a timer after it sends out a request. It waits for $f+1$ identical replies from different replicas. Because at most f replicas are faulty, at least one reply must come from a correct replica. If the timer expires before it receives a correct reply, the client broadcasts the

Figure 1. The Byzantine agreement problem: To tolerate a single Byzantine fault, four replicas are needed. (a) If the commander (i.e., primary replica) is faulty, he may send conflicting information to its lieutenants (i.e., backup replicas). However, the lieutenants can exchange information regarding what they heard from the commander and reach the correct decision (attack) based on majority voting. (b) On the other hand, if a lieutenant is faulty, he may lie to other lieutenants regarding the information he has heard from the commander. Other lieutenants can still reach a correct decision based on majority voting. Reducing the number of replicas to three cannot guarantee an agreement among the correct replicas.



request to all server replicas. This enables the correct replicas to detect the primary failure so that a new primary can be elected (this is often called a view change).

All correct replicas must agree on the same set of input messages with the same execution order. In other words, the request messages must be delivered to the replicas reliably in the same total order. To understand this better, consider two replicas R1 and R2 and two requests M1 and M2 from two different clients. Let M1 be an update request and M2 be a read-only request. If R1 processes M1 and then M2, and R2 processes M2 and then M1, the reply messages for M2 from R1 and R2 would be different because one reflects the latest update while the other does not.

Furthermore, replicas may not process the same request deterministically — that is, given the same request (delivered in the same order to all replicas), they may not produce identical reply. There are many factors that can cause non-deterministic execution of a request, for example, differences in local clocks, process identifiers, and many other local resources that might be referenced by the replicas, and multi-threading. These factors must be controlled properly so that the replicas appear to execute deterministically. This is a rather difficult task in the face of Byzantine fault. So far, this issue has been addressed in a very limited fashion. Current solutions often assume that the replicas are single-threaded, and all non-deterministic operations are known *a priori*. The primary replica determines the values to be used for all replicas and disseminates them to the backups. The backups subsequently verify the proposed values. If the primary is detected to be faulty, a view change is initiated.

Byzantine fault tolerance mechanisms tend to suffer from scalability problems. Amir et al. (2006) have tackled the size scalability problem by using a hierarchical replication architecture so that Byzantine fault tolerance can be achieved in large-scale application running over wide area networks. Aiyer et al. (2005) considered the challenges faced by ensuring Byzantine fault tolerance over multi-administrative domains (administration scalability problems) and invented mechanisms to cope with selfish members, in addition to arbitrary faulty members.

Secrete Sharing and Threshold Cryptography

Another aspect of intrusion tolerance is to protect confidential information (e.g., security keys) even if some replicas have been penetrated (Deswarte et al., 1991). If all the secrets are maintained by a single process, an adversary can obtain these secrets by compromising the process. To defend against such an attack, each secret is divided into multiple shares, and each share is stored in a separate process. To obtain the secret, an adversary must now break into a significant number of processes. This is the basic objective of *secret sharing.*

A popular secret sharing scheme is the *(k, n) threshold scheme*, where *n* is the total number of shares and *k* is the minimum number of shares needed to reconstruct the secret. No useful information can be obtained as long as the number of shares collected is less than *k*, the threshold. This scheme was first proposed by Shamir (1979) and implemented using polynomial interpolation. In this scheme, each share is of similar size to the original secret, and shares can be dynamically added or deleted.

The (k, n) threshold scheme is quite expensive computationally and space-wise. Therefore, it is used only to protect the most crucial secret such as security keys. It might not be practical to apply the scheme to file systems or databases. Consequently, a more cost-effective scheme, called Fragmentation-Replication-Scattering (FRS), is proposed (Deswarte et al., 1991). It was initially designed to provide intrusion tolerance for file systems and was later extended to object-based systems (Fabre & Randell, 1992). The FRS scheme involves three steps. First, a file is partitioned into many smaller pieces. (To enhance the confidentiality, the file can be encrypted prior to the fragmentation step.) Second, each piece is replicated. Finally, the pieces are distributed pseudo-randomly according to some algorithm to the storage sites. The fragmentation and scattering protect data confidentiality against intrusion attacks, because to obtain the file, an adversary must first find out the locations of the fragments belonging to the file and then penetrate all the sites that store the fragments. The replication protects the availability of the file so that even if some pieces are destroyed by an adversary, there are enough copies left to reconstruct the original file.

Despite the elegancy of the secret sharing schemes, the secret must be reconstructed before it can be used. This poses a security threat because if the process that performs this task is compromised, the secret may be exposed. This prompted the development of threshold cryptography (Desmedt & Frankel, 1989). In threshold cryptography, security operations such as encryption, decryption, signature generation, and verification can be performed by a group of processes without reconstructing the shared secret. Threshold cryptography has been applied primarily to public key-based security services by sharing the private key among a group of processes (Zhou, Schneider, & van Renesse, 2002). The shares can be proactively updated to further enhance the security. By sharing the private key, each process can produce a partial signature. If a client obtains enough partial signatures, it can compute the complete signature. During this process, the private key is never reconstructed.

PROTECTING BOTH AVAILABILITY AND CONFIDENTIALITY

Byzantine fault tolerance ensures service high availability, but it does not protect data confidentiality against intrusions. Threshold cryptography guarantees both confidentiality and

high availability of some security operations, but not general services one might use. All secret sharing schemes require the reconstruction of the secret at a trusted site, which may be vulnerable to intrusion attacks.

COCA (Zhou et al., 2002) is the first attempt to integrate a Byzantine quorum system with threshold cryptography in the context of a certificate authority (CA) service. In COCA, the most critical state — that is, the service private key, used to generate signatures for the certificates issued to clients — is shared among the replicas to prevent an adversary from stealing it. There are other states in a CA service, such as the certificate issued to the clients. A client can query and update the certificate information through the CA services. The high availability of the CA services is provided by the group of CA replicas. To enhance failure resiliency, the replicas are periodically restarted proactively with a correct binary image and the latest state collected from other correct replicas. The server keys (used to communicate securely among the replicas) are also periodically refreshed.

FUTURE TRENDS

Even though there is moderate success in applying intrusion tolerance techniques to a few specific applications, it remains to be seen how to introduce intrusion tolerance to other types of information systems, such as credit card processing systems, online banking systems, and e-commerce systems. The lack of effective solutions has left many mission-critical systems vulnerable. This is evidenced by the increasingly frequent reports of high-impact security breaches that resulted in massive disclosure of confidential information, such as the break-in of CardSystems Solutions, which exposed the confidential data of more than 40 million credit card holders (Dash & Zelle, 2005), and the Veteran Affairs incident, which has caused the potential loss of personal data of more than 26.5 million veterans to the hands of adversaries (Dunhm, 2006).

There are many open issues to be resolved before we see widespread adoption of intrusion tolerance techniques. The most interesting research problem seems to be the design of a systematic methodology that allows direct operation on the shared secret that is dispersed among a number of processes. This will eliminate the vulnerability introduced by relying on a trusted process to reconstruct the secret. Another urgent issue is how to address the replica non-determinism problem, especially that caused by multi-threading. Practical information systems are very complicated and contain extensive non-deterministic operations. Unfortunately, all Byzantine fault tolerance strategies require deterministic execution of replicas. Even though the non-determinism resulting from a replica's accessing of some local resources (such as clock values and file descriptors) can be easily controlled if the access pattern is known in advance (Yin et al., 2003), there is no straightforward solution to render a multi-threaded application deterministic without significantly impacting the application's performance and correctness (Zhao, Moser, & Melliar-Smith, 2005). Furthermore, intrusion tolerance favors design diversity to avoid common vulnerabilities among the replicas. N-version programming was proposed as a potential solution by Chen and Avizienis (1978). However, it might not be an economically viable solution due to its heavy software development cost. Alternative, less costly solutions are needed to diversify the replica implementations. Code randomization appears to be a good approach to achieve replica diversity (Forrest, Somayaji, & Ackley, 1997).

Intrusion tolerance design does not conflict with the effort of intrusion prevention and intrusion detection (Verissimo et al., 2003). In fact, they are all indispensable techniques for building secure and dependable systems. Without applying good intrusion prevention techniques, the application would contain too many vulnerabilities for the replication strategy to take effect. Similarly, intrusion detection plays an essential role in intrusion tolerant systems. To remove a faulty replica and to recover it subsequently, the fault must be detected as quickly as possible. Intrusion detection also helps deter future intrusion attacks if there is enough information logged that can be used to prosecute the intruders (Dunlap, King, Cinar, Barsai, & Chen, 2002).

CONCLUSION

We believe that the research and development of intrusion tolerant systems will gain more momentum as more and more services are offered online. The expectation of such services is high, considering their essential roles in everyday operations of businesses and individuals as well. The impact of service unavailability and security breaches will only be more serious. In this chapter, we have surveyed the state-of-the-art techniques for building intrusion tolerant systems. We also illustrated a few most urgent open issues for future research. Finally, we pointed out that to build secure and dependable systems, we need a concerted effort in intrusion prevention, intrusion detection, and intrusion tolerance.

REFERENCES

Aiyer, A., Alvisi, L., Clement, A., Dahlin, M., Martin, J., & Parth, C. (2005). BAR fault tolerance for cooperative services. *Proceedings of the ACM Symposium on Operating Systems Principles* (pp. 45-58), Brighton, UK.

Amir, A., Danilov, C., Dolev, D., Kirsch, J., Lane, J., Nita-Rotaru, C., Olsen, J., & Zage, D. (2006). Scaling Byzantine fault-tolerant replication to wide area networks. *Proceedings of the International Conference on Dependable Systems and Networks* (pp. 105-114), Philadelphia, PA.

Castro, M., & Liskov, B. (2002). Practical Byzantine fault tolerance and proactive recovery. *ACM Transactions on Computer Systems, 20*(4), 398-461.

Chen, L., & Avizienis, A. (1978). N-version programming: A fault-tolerance approach to reliability of software operation. *Proceedings of the International Symposium on Fault Tolerant Computing* (pp. 3-9), Toulouse, France.

Dash, E., & Zeller, T. (2005). MasterCard says 40 million files are put at risk. *New York Times,* (June 18).

Desmedt, Y., & Frankel, Y. (1990). Threshold cryptosystems. *Lecture Notes in Computer Science, 435,* 307-315.

Dunhm, W. (2006). Personal data on millions of U.S. veterans stolen. *Computerworld,* (May 22).

Deswarte, Y., Blain, L., & Fabre, J. (1991). Intrusion tolerance in distributed computing systems. *Proceedings of the IEEE Symposium on Research in Security and Privacy* (pp. 110-121), Oakland, CA.

Dunlap, W., King, S., Cinar, S., Barsai, M., & Chen, P. (2002). ReVirt: Enabling intrusion analysis through virtual-machine logging and replay. *Proceedings of the Symposium on Operating Systems Design and Implementation* (pp. 211-224), Berkeley, CA.

Fabre, J., & Randell, B. (1992). An object-oriented view of fragmented data processing for fault and intrusion tolerance in distributed systems. *Lecture Notes in Computer Science, 648,* 193-208.

Forrest, S., Somayaji, A., & Ackley, D. (1997). Building diverse computer systems. *Proceedings of the Workshop on Hot Topics in Operating Systems* (pp. 67-72), Cape Cod, MA.

Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems, 4*(3), 382-401.

Malkhi, D., & Reiter, M. (1997). Byzantine quorum systems. *Proceedings of the ACM Symposium on Theory of Computing* (pp. 569-578), El Paso, TX.

Pfleeger, C., & Pfleeger, S. (2002). *Security in computing* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Schneider, F. (1990). Implementing fault-tolerant services using the state machine approach: A tutorial. *ACM Computer Survey, 22*(4), 299-319.

Shamir, A. (1979). How to share a secret. *Communications of the ACM, 22*(11), 612-613.

Verissimo, P., Neves, N., & Correia, M. (2003). Intrusion-tolerant architectures: Concepts and design. *Lecture Notes in Computer Science, 2677,* 90-109.

Yin, J., Martin, J., Venkataramani, A., Alvisi, L., & Dahlin, M. (2003). Separating agreement from execution for Byzantine fault tolerant services. *Proceedings of the ACM Symposium on Operating Systems Principles* (pp. 253-267), Bolton Landing, NY.

Zhao, W., Moser, L., & Melliar-Smith, P. (2005). Deterministic scheduling for multi-threaded replicas. *Proceedings of the IEEE International Workshop on Object-Oriented Real-time Dependable Systems* (pp. 74-81), Sedona, AZ.

Zhou, L., Schneider, F., & van Renesse, R. (2002). COCA: A secure distributed online certification authority. *ACM Transactions on Computer Systems, 20*(4), 329-368.

KEY TERMS

(k, n) Thread Scheme: A secret is divided into *n* shares. To reconstruct the secret, at least *k* shares are needed. No useful information can be obtained from *k-1* shares.

Byzantine Fault: Used to model arbitrary fault. A Byzantine faulty process might send conflicting information to other processes to prevent them from reaching an agreement.

Byzantine Fault Tolerance: A replication-based technique used to ensure high availability of an application subject to Byzantine fault.

Byzantine Quorum System: A system offering read and write services to its clients on a set of replicated data items. A read operation retrieves data from a quorum of correct replicas, and a write operation applies the update to a quorum of correct replicas. Any two quorums must overlap by at least one correct replica.

Fragmentation Redundancy Scattering: A secret sharing scheme that involves the following three steps: fragmenting a file, replicating each fragment, and distributing the replicated fragments to different storage sites.

Replica Consistency: The states of the replicas of an application should remain identical at the end of the processing of each request. Replica consistency is necessary to mask a fault in some replicas.

Threshold Cryptography: Security operations such as encryption, decryption, signature generation, and verification can be performed by a group of processes without reconstructing the shared secret. Threshold cryptography utilizes (k, n) threshold schemes internally.

Inventing the Future of E-Health

José Aurelio Medina-Garrido
Cadiz University, Spain

María José Crisóstomo-Acevedo
Jerez Hospital, Spain

INTRODUCTION

E-health involves the use of information and communications technologies to improve health in general and the healthcare system in particular (Alvarez, 2002; Chau & Hu, 2004; Roger & Pendharkar, 2000).

Healthcare, one of the largest industries in the world, suffers from some inefficiencies and inequities in both service provision and quality. Some of these problems are due to the poor management of the information flows (Kirsch, 2002). In this respect, there are business opportunities for e-health. But to understand what the future holds for e-health, we need to find a precise definition of the concept and identify the possible sources of business.

This article is structured as follows. The second section, the background, defines the concept of e-health. The third section outlines some of the business opportunities in the area of e-health based on the communications platform that is the Internet, and discusses some practical guidelines for e-health businesses to create value. The fourth section discusses the low level of adoption of e-health at present, as well as the future trends, in which e-health will presumably grow. e-health is also expected to be used to reduce the disparities in the population in access to healthcare, and for the treatment of the chronically ill. The fifth section is dedicated to the final conclusions.

BACKGROUND

The term *e-health* is relatively recent and refers to healthcare practice that is supported by electronic processes and communications. The term has many definitions, depending on the functions, stakeholders, context, or the theoretical framework referred to. It includes a wide range of medical informatics applications, both specific (for example, decision support systems, citizen health information) and general (for example, management systems, healthcare services provision, etc.). But the increased importance of the communication function in e-health, and the use of electronic networks (particularly the Internet), differentiate e-health from traditional medical informatics (Pagliari, 2005).

Thus, e-health goes beyond healthcare informatics and incorporates the most advanced information technologies to medicine and healthcare. Among the most significant applications of the technologies to healthcare are the following:

- *Electronic Medical Records*, which allow different healthcare professionals to share information about a particular patient.
- *Telemedicine*, which uses information and communications technology (ICT) to enable physician-patient contact at a distance.
- *Evidence-based Medicine*, in which a system updates information about the most appropriate treatments for each patient, thereby enhancing physicians' treatment possibilities.
- *Citizen-oriented Information*, through which citizens are provided with information about health topics.
- *Specialist-oriented Information*, whereby a system distributes information to specialists about medical journal articles, practices and protocols in the area of health, new medical advances, epidemiological alerts, etc.
- *Virtual healthcare teams*, made up of healthcare professionals sharing information about patients electronically to improve their knowledge and decision-making.
- *Health e-commerce*, which involves providing value-added electronic services to both professionals and citizens, and economically exploiting some or all of the services. In this respect, e-health is supported by the Internet and related technologies and combines medical informatics, public health, and business. This type of e-health does not exclude the previous ones. To the contrary, it includes them or complements them. The following section discusses the concept of Health e-commerce, indicating what types there are, what they consist of, and how they obtain their revenues.

Some authors go further than the concepts explained in this section and predict a change of mentality and culture among both citizens and practitioners. One author goes so far as to argue that e-health "..characterizes not only a technical development, but also a state-of-mind, a way of

thinking, an attitude, and a commitment for networked, global thinking, to improve healthcare locally, regionally, and worldwide by using information and communication technology" (Eysenbach, 2001).

BUSINESS OPPORTUNITIES BASED ON E-HEALTH

As we mentioned above, the inadequate management that a large part of the healthcare sector makes of its information flows (Kirsch, 2002) and processes, as well as the new advantages offered by present-day ICT, mean that e-health opens up significant business opportunities.

One of the most notable business opportunities offered by e-health is e-commerce. The most important forms that *e-commerce* can adopt on the Internet include (Parente, 2000): portals, connectivity sites, business-to-business applications, and business-to-consumer applications.

Portals dedicated to health tend to provide all types of information, guidance, and medical advice to consumers and professionals. Portals generally represent starting points for consumers, offering them various online activities as well as diverse information. Their general objective is to be the first place that customers go to when they are looking for something on the Internet. For this, they need to establish a brand that attracts visits and creates loyal customers. Their main sources of income come from the advertising they contain and occasionally from users' subscriptions.

Sites dedicated to facilitating connectivity in the healthcare sector have the objective of linking and integrating the various information systems seamlessly. The income of this business model comes from the company's external users, who pay fees to obtain information. Health e-commerce connectivity initiatives involve accessing electronic medical records on the Internet, evaluating the quality of providers according to their clinical results, and using quality information in the selection of physicians. For example, some hospitals provide their patients with directories of their physicians on the Internet, which are searchable by zip code and clinical specialty (Coile, 2000). Because these sites obtain their revenues from the fees generated by each information transaction, their objective is to maximize the number of transactions. These companies obtain transaction fees from health plan providers, physicians, hospitals, clinical laboratories, pharmacies, consumers, and companies offering financial, marketing, or delivery services in the healthcare sector.

Business-to-business (B2B) e-commerce involves selling products and services to other firms on the Internet. The income from this business model comes from the sale of the product or service itself. *B2B Health e-commerce* includes businesses dedicated to selling refurbished medical equipment or pharmaceutical refills on the Internet. Indeed, pharmaceutical refills are a large market with a high turnover, and are ideally suited to be traded on the Internet at competitive prices offering next-day home delivery (Coile, 2000). Apart from the products sold in this way, some companies are now beginning to offer services such as online management consultancy.

Business-to-consumer (B2C) e-commerce sells products and services directly to the consumer via the Internet. As in the previous model, the income comes from the sales themselves. The B2C business model in the healthcare area allows consumers to acquire products and services such as health insurance, prescription drug refills, over-the-counter drugs, medical supplies for the chronically ill, vitamins, homeopathic medicines, and home fitness equipment (Coile, 2000).

Table 1 shows examples of some of the most important firms in the e-health sector. The firms are classified under the business model that most closely matches their main activities or sources of income. But the limits are often hazy, and the firms can often be classified in more than one category.

After the bursting of the Internet bubble it became clear that an e-business, such as one based on e-health, should in the first place be a business venture, and not just a technological one. e-health firms are businesses, and so they must seek to create value. In this respect, some authors offer *practical guidelines to e-businesses* about how to generate value (Earle

Table 1. Examples of business models in e-health

Portal	Connectivity	B2B	B2C
Medscape	Healtheon/WebMD	Neoforma.com	drugstore.com
drkoop.com	TriZetto	Medical Manager	PlanetRx
OnHealth	XCare.net	Allscripts	HealthExtras
HealthGrades.com	Claimsnet.com	eBenx	

Source: Adapted from Parente (2000).

and Keen, 2000; Shapiro and Varian, 1999), and these are also valid for the particular case of e-health:

- Cultivate stable relationships with customers. Building a critical mass of loyal customers allows firms to avoid customer acquisition costs for each transaction. The idea is to build solid relationships with strong ties. For this reason, some firms offer some services for free on their Web sites.
- Build a powerful brand. The concept of brand is redefined on the Internet. It is a relationship brand. Customers cannot see the product or service until they pay for it. Thus, e-health firms need to have a good reputation. This reputation, which takes substantial time and money to build, can be quickly acquired working together or allying with another firm that already has a good reputation.•
- Improve the logistics. This is important for e-health firms that distribute physical products such as medical equipment or drugs. Logistics capabilities are critical for the generation of value. But some firms have opted to focus on those core activities they know how to do well, in order not to spread efforts or resources too thinly, and have allied with top logistics firms that can undertake this function.
- Harmonize the channels in the name of the customer. Customers choose the communication channel that offers them most advantages. Firms need to provide the option that best helps to build and maintain the relationship, and that choice is for the customer to make, not the firm. In this respect, customers need to be offered a number of communications channels (Web forms, e-mail, telephone, fax, post, cell phone SMS, personal digital assistants (PDAs), a combination of physical and virtual branches, etc.).
- Become an intermediary that provides value, or use one that does. Business on the Internet is dominated by nerve centers such as portals with powerful brands and other intermediaries that bring the supply and the demand together. These will control the interaction between providers and customers, and will advise customers about the Web sites they should visit when looking for a particular product or service. Only the intermediaries that offer consumers or firms value will survive and prosper. Intermediaries that do not provide value must use one that does if they wish to be profitable.
- Analyze how much the firm invests in producing and selling information. Information is expensive to produce, but extremely cheap to copy. In this respect, trading data electronically allows firms to distribute information enjoying economies of scale, which means they can cut unit costs and consequently the price of their product or service.
- When firms compete in commodity markets they need to create economies of scale that cut costs (and prices), be flexible to adapt to any change and quick, both to enter a market and exploit the business opportunity and to exit when this is no longer profitable.

FUTURE TRENDS

E-health tools show plenty of potential, but they are relatively undeveloped and have not yet been adopted to a great extent (Wilson, 2005). Some authors have predicted that some of the new technologies applicable to e-health will be adopted very quickly and massively (for example, telemedicine or PDAs), but there remain problems in this respect. Occasionally, the main *source of income of the business model* has seen undermining. This has been the case, for example, of the supply of information to the consumer, as they are not used to paying for this. On the other hand, some technologies do not have a large enough potential market to grow very quickly, for example, applications for supply-chain management (SCM) or procurement. Consequently, firms do not find them so attractive to invest in (Kirsch, 2002).

Electronic medical records are also expected to take off in the future. No paperless hospitals can be found just yet, but some authors predict that some hospitals will be completely paperless in the not-too-distant future (Coile, 2000). If these predictions prove right, extreme care will be needed to protect patients' privacy and interests.

A proper implementation of *e-health will require political commitment*, an adequate legal framework, and R&D and Innovation (Wilson, 2005). The political commitment should be reflected in the development of electronic health cards and health information networks and online health services. The legal framework must offer adequate coverage in terms of data protection, digital signatures, e-commerce regulations, and the professional qualifications required to use telemedical applications. The R&D and Innovation needs to promote the development of new technological tools and help to spread best practices in this new field.

Another potential future trend for *e-health is to apply information and communications technologies to reduce the disparities in the population* in access to the healthcare system (Ahern, Kreslake, & Phalen, 2006; Cashen, Dykes, & Gerber, 2004; Gibbons, 2005). Demographic and socioeconomic factors are arguably behind these disparities, such as: ethnic origin (whites vs. minorities), geographic (urban vs. rural) (Galea & Vlahov, 2005), gender (masculine vs. feminine) (Mcgrath & Puzan, 2004; Quinn & Overbaugh, 2005), income level (poor vs. non-poor) (Federico & Liu, 2003), and age (elderly vs. non-elderly) (Pyle & Stoller, 2003).

In the future, *e-health is also expected to increasingly facilitate the treatment and monitoring of patients with chronic illnesses* (Ahern, Kreslake & Phalen, 2006). This

will reduce the constant and inconvenient traveling to and from medical centers that these patients are subjected to, not to mention the healthcare system's work overload in this respect.

CONCLUSION

E-health will lead to a behavior change in the area of healthcare, through which the use of information technologies, the Internet and communications technologies enable improved and more effective healthcare (Eng, 2002).

One of the main areas of interest in e-health is to improve health communications by using technologies. In this respect, both healthcare organizations and public healthcare agencies are increasingly using the Internet in their communications and to transfer information.

These efforts are generating various socio-economic phenomena, in particular e-health as a business opportunity – the chance to create an e-business. Four models familiar from traditional e-businesses have also been adopted by e-health businesses: portals, connectivity sites, business-to-business applications, and business-to-consumer applications.

Although more attention is commonly paid to technological questions, we must also remember that an e-business is first of all a business. E-health businesses, as businesses, need to create value for their customers. In this respect, they are advised to cultivate stable relationships with their customers, build a powerful brand, consider the logistics carefully, offer a number of channels to their customers, be an intermediary that provides value, seek economies of scale when they produce and sell information, and be quick and flexible and create economies of scale when they compete in commodity markets.

At present, e-health activities are defusing only gradually, and have been slow to be accepted. But a greater development is forecast for the future; in particular, we will conceivably see phenomena such as paperless hospitals and a more complete coverage of the population's healthcare needs (education about health, treatment of chronically ill patients, and reduction in the disparities in access to healthcare).

REFERENCES

Ahern, D.K., Kreslake, J.M., & Phalen, J.M. (2006). What Is E-health: Perspectives on the Evolution of E-health Research. *Journal of Medical Internet Research*, 8(1), e4.

Alvarez, R.C. (2002). The promise of e-Health - a Canadian perspective. *E-health International*, 1(1), 4.

Cashen, M.S., Dykes, P., & Gerber B. (2004). E-health technology and Internet resources: barriers for vulnerable populations. *Journal of Cardiovascular Nursing*, 19(3), 209-222.

Chau, P.Y.K., & Hu, P. J. (2004). Technology Implementation for Telemedicine Programs. *Communications of the ACM*, 47(2), 87-92.

Coile, R.C. (2000). E-health: Reinventing healthcare in the information age. *Journal of Healthcare Management*, 45(3), 206-210.

Earle, N., & Keen, P. (2000). From .com to .profit. *Inventing Business Models that Deliver Value and Profit*. San Francisco, California: Jossey-Bass Inc.

Eng, T.R. (2002). E-health research and evaluation: challenges and opportunities. *Journal of Health Communication*, 7(4), 267-272.

Eysenbach, G. (2001). What is e-health? *Journal of Medical Internet Research*, 3(2), e20.

Federico, M.J., & Liu, A.H. (2003). Overcoming childhood asthma disparities of the inner-city poor. *Pediatric Clinics of North America*, 50(3), 655-75, vii.

Gibbons, M.C. (2005). A Historical Overview of Health Disparities and the Potential of E-health Solutions. *Journal of Medical Internet Research*, 7(5), Article e50.

Galea, S., & Vlahov, D. (2005). *Handbook of Urban Health: Populations, Methods and Practice*. New York: New York Academy of Medicine.

Kirsch, G. (2002). The business of e-health. International Journal of Medical Marketing, 2(2), 106-110.

Mcgrath, B.B., & Puzan, E. (2004). Gender disparities in health: attending to the particulars. *Nursing Clinics of North America*, 39(1), 37-51.

Pagliari, C., Sloan, D., Gregor, P., Sullivan, F., Detmer, D., Kahan, J.P., Oortwijn, W., & MacGillivray, S. (2005). What Is E-health (4): A Scoping Exercise to Map the Field. *Journal of Medical Internet Research*, 7(1), e9.

Parente, S.T. (2000). Beyond the hype: A taxonomy of e-health business models. *Health Affairs. Chevy Chase*, 19(6), 89-102.

Pyle, M.A., & Stoller, E.P. (2003). Oral health disparities among the elderly: interdisciplinary challenges for the future. *Journal of Dental Education*, 67(12), 1327-1336.

Quinn, T.C., & Overbaugh, J. (2005). HIV/AIDS in women: an expanding epidemic. *Science*, 308(5728), 1582-1583.

Rodger, J.A., & Pendharkar, P.C. (2000). Using telemedicine in the Department of Defense. *Communications of the ACM*, 43(3), 19-20.

Shapiro, C., & Varian, H.R. (1999). *Information Rules. A Strategic Guide to the Network Economy*. Boston, Massachusetts: Harvard Business School Press.

Wilson, P. (2005). *My Health / My E-health. Meeting the challenges of making e-health personal*. Presented at ICLM9. Brazil, September.

KEY TERMS

Business-to-Business (B2B) E-Commerce: Economic transactions between firms using information systems and technologies.

Business-to-Consumer (B2C) E-Commerce: Commercial transactions and activities between firms and the end-consumer using information systems and technologies.

Commodities: Raw materials, unfinished products, or products sold loose, or any other product characterized by being undifferentiated. Such products cannot be differentiated from other products in function of the producer that manufactures them or the supplier that sells them.

E-Health: The provision of any healthcare service that is supported by electronic processes and communications.

Electronic Medical Records: Computer-based patient medical records. Patient medical records are a systematic documentation of a patient's medical history and care.

Evidence-Based Medicine: Medical practice involving the sharing, updating and consultation of a system containing information about the most appropriate treatments for each patient. This helps to improve the treatments chosen by the physicians who use this system.

Health E-Commerce: E-business based on the economic exploitation of health-related information and services.

Health Plan: An individual or group plan that provides, or pays the cost of, medical care.

Over-the-Counter (OTC) Medicine: A medicine that can be bought without a doctor's prescription, such as some analgesics.

Telemedicine: The use of information and communications technologies to exchange information between practitioners, or to deliver medical services to a patient remotely.

Virtual Healthcare Teams: Teams made up of healthcare professionals that share information about patients electronically in order to improve their knowledge and decision-making.

Investigating Internet Relationships

Monica T. Whitty
Queen's University Belfast, UK

INTRODUCTION

The focus on Internet relationships has escalated in recent times, with researchers investigating such areas as the development of online relationships (e.g., McCown, Fischer, Page, & Homant, 2001; Parks & Roberts, 1998; Whitty & Gavin, 2001), the formation of friends online (Parks & Floyd, 1996), representation (Bargh, McKenna, & Fitzsimons 2002), and misrepresentation of self online (Whitty, 2002). Researchers have also attempted to identify those addicted to accessing online sexual material (Cooper, Putnam, Planchon, & Boies, 1999). Moreover, others have been interested in Internet infidelity (Whitty, 2003a, 2005) and cybersex addiction (Griffiths, 2001, Young, Griffin-Shelley, Cooper, O'Mara, & Buchanan, 2000). Notwithstanding this continued growth of research in this field, few researchers have considered the new ethical implications of studying this topic area.

While it is acknowledged here that some of the discussions in this article might be equally applied to the study of other Internet texts, such as religious or racial opinions, the focus in this article is on the concomitant ethical concerns of ongoing research into Internet relationships. Given that the development and maintenance of online relationships can be perceived as private and very personal (possibly more personal than other sensitive areas), there are potential ethical concerns that are unique to the study of such a topic area (Whitty, 2004; Whitty & Carr, 2006). For a broader discussion of virtual research ethics in general, refer to Ess and Jones (2004) and Whitty and Carr (2006).

BACKGROUND

Early research into this area has mostly focused on the similarities and differences between online and off-line relationships. Researchers have been divided over the importance of available social cues in the creation and maintenance of online relationships. Some have argued that online relationships are shallow and impersonal (e.g., Slouka, 1995). In contrast, others contend that Internet relationships are just as emotionally fulfilling as face-to-face relationships, and that any lack of social cues can be overcome (Lea & Spears, 1995; Walther, 1996). In addition, researchers have purported that the ideals that are important in traditional relationships, such as trust, honesty, and commitment, are equally important online, but the cues that signify these ideals are different (Whitty & Gavin, 2001). Current research is also beginning to recognize that online relating is just another form of communicating with friends and lovers, and that we need to move away from considering these forms of communication as totally separate and distinct entities (e.g., Wellman, 2004). Moreover, McKenna, Green, and Gleason (2002) have found that when people convey their "true" self online they develop strong Internet relationships and bring these relationships into their "real" lives.

Internet friendships developed in chat rooms, newsgroups, and MUDs or MOOs have been examined by a number of researchers. For example, Parks and Floyd (1996) used e-mail surveys to investigate how common personal relationships are in newsgroups. After finding that these relationships were regularly formed in newsgroups, Parks and Roberts (1998) turned to examine relationships developed in MOOs. These researchers found that most (93.6%) of their participants had reported having formed some type of personal relationship online, the most common type being a close friendship.

Researchers have also been interested in how the playful arena of the Internet impacts on the types of relationships formed in these places (e.g., Whitty, 2003b; Whitty & Carr, 2003, 2006). Turkle's (1995) well-known research on her observations while interacting in MUDs found that the role-playing aspect of MUDs actually creates opportunities for individuals to reveal a deeper truth about themselves. Whitty and Gavin (2001) have also contended that although people do lie about themselves online, this paradoxically can open up a space for a deeper level of engagement with others.

Importantly, some researchers are now starting to realize that cyberspace is not a generic space that everyone experiences in the same way. New theories are currently being developed to explain how individuals present themselves in different spaces online. For instance, Whitty (in press) devised the BAR theory to explain presentation of self on online dating sites, which she believes is different to other spaces within cyberspace. The BAR theory purports that most online daters find the best strategy for developing a "successful profile" is to create a balance between an "attractive self" and a "real self." The online daters Whitty and her research assistants interviewed (see Whitty, in press; Whitty & Carr, 2006) talked about the need to re-write their profiles if they were attracting either people they did not desire, or if they were attracting no one, or if their date appeared disappointed with them when they met face-to-face (given that they did not live out to their profile). Therefore, it would seem that

a successful profile has to appear attractive enough to stand out and be chosen, but also one that individuals could live up to in their first face-to-face date (which often took place within a couple of weeks of meeting online).

Cybersex addiction and the available treatment for these cybersex addicts and their partners has been an area of research and concern for psychologists (e.g., Schneider, 2000; Young, Pistner, O'Mara, & Buchanan, 1999). Research has also focused on what online acts might be considered as an act of infidelity. For example, Whitty (2003a) found that acts such as cybersex and hot-chatting were perceived as almost as threatening to the off-line relationship as sexual intercourse. In addition to these concerns, Cooper et al. (1999) identified three categories of individuals who access Internet erotic material, including recreational users, sexual compulsive users (these individuals are addicted to sex *per se*, and the Internet is but one mode where they can access sexual material), and at-risk users (these individuals would never have developed a sexual addiction if it were not for the Internet).

ETHICAL ISSUES PERTINENT TO THE STUDY OF INTERNET RELATIONSHIPS

Much of the research, to date, on Internet relationships and sexuality has been conducted online—either through interviews, surveys, or by carrying out analysis on text that is readily available online. There are many advantages to conducting research online as well as collecting text or data available online for analysis in one's research (see Table 1).

In spite of the numerous advantages to conducting research online, investigators also need to be aware of the disadvantages (see Table 2).

What all studies that research Internet relationships have in common is that they are researching a sensitive topic, which requires individuals to reveal personal and often very private aspects of themselves and their lives. Given the sensitive nature of this topic area, it is crucial that researchers give some serious thought to whether they are truly conducting research in an ethical manner.

Table 1. Practical benefits of conducting research online

- Easy access to a population of individuals who form relationships online and who access sexual material
- Internet provides researchers with a population that is sometimes difficult to research (e.g., people with disabilities, agoraphobia)
- Contact people in locations that have closed or limited access (e.g., prisons, hospitals)
- Requires relatively limited resources
- Ease of implementation

Photographs, video, sound bites, and text produced by individuals online are sometimes examined by researchers. The text can be produced in a number of different forums, including chat rooms, MUDs, newsgroups, MySpace, Bebo, and online dating sites. One way researchers collect data is by lurking in these different spaces in cyberspace. The development of online relationships (both friendships and romantic) and engaging in online sexual activities, such as cybersex, could easily be perceived by those engaging in such activities as a private discourse. Given the nature of these interactions, social researchers need to seriously consider if they have the right to lurk in online settings in order to learn more about these activities—despite the benefits of obtaining this knowledge.

There are fuzzy boundaries between what constitutes public and private spaces online, and researchers need to acknowledge that there are different places within cyberspace. For example, a chat room might be deemed a more public space than e-mail. It is contended here that lurking in some spaces online might be ethically questionable. We must, as researchers, debate how intrusive a method lurking potentially is. As Ferri (1999, cited in Mann & Stewart, 2000) contends, "who is the intended audience of an electronic communication—and does it include you as a researcher?" (p. 46).

Researchers also need to consider how the participant perceives the various online spaces. As Ferri suggests, private interactions can and do indeed occur in public places. It has been theorized that the Internet can give an individual a sense of privacy and anonymity (e.g., Rice & Love, 1987; Whitty & Carr, 2006). The "*social presence theory*" contends that "social presence" is the feeling one has that other persons are involved in a communication exchange (Rice & Love, 1987). Since computer-mediated-relating (CMR) involves less non-verbal cues (such as facial expression, posture, and dress) and auditory cues in comparison to face-to-face communication, it is said to be extremely low in social presence. Hence, while many others might occupy the space online, it is not necessarily perceived in that way. As researchers we need to ask some questions: Can researchers ethically take advantage of these people's false sense of privacy and security? Is it ethically justifiable to lurk in these sites and download material without the knowledge or consent of the individuals who inhabit these sites? This is especially relevant to questions of relationship development and sexuality, which are generally understood to be private

Table 2. Disadvantages of conducting research online

- Security issues
- Possible duplication of participants completing surveys
- Difficult to ascertain how the topic area examined impacts on the participant
- Restricted to a certain sample

matters. Therefore, good ethical practice needs to consider the psychology of cyberspace and the false sense of security the Internet affords.

It is suggested here that researchers need to maintain personal integrity as well as be aware of how their online investigations can impact the Internet relationships they study. For example, given researchers' knowledge of online relationships, interacting on online dating sites, chat rooms, and so forth could potentially alter the dynamics of these communities.

While it might be unclear as to how ethical it is for lurkers to collect data on the Internet, there is less doubt as to whether it is acceptable to deceive others online in order to conduct social research, especially with respect to online relationships and sexuality. Ethical guidelines generally state that deception is unethical because the participant is unable to give free and fully informed consent. For example, according to the Australian National Health and Medical Research Council (NHMRC), which set the ethical guidelines for Australian research:

As a general principle, deception of, concealment of the purposes of a study from, or covert observation of, identifiable participants are not considered ethical because they are contrary to the principle of respect for persons in that free and fully informed consent cannot be given. (NHMRC, 1999)

Generally, ethical guidelines will point out that only under certain unusual circumstances deception is unavoidable when there is no alternative method to conduct one's research. However, in these circumstances individuals must be given the opportunity to withdraw data obtained from them during the research that they did not originally give consent to.

FUTURE TRENDS

As with any other research conducted within the social sciences, some important ethical practices need to be adhered to when we conduct research on Internet relationships and sexuality (see Table 3).

Informed consent requires researchers to be up front from the beginning about the aims of their research and how they are going to be utilizing the data they collect. In off-line research individuals often sign a form to give their consent; however, this is not always achievable online. One way around this is to direct participants to a Web site that contains information about the project. This Web site could inform the participants about the purpose of the study, what the study entails, as well as contact details of the researcher, and the university Human Ethics Committee.

In some cases, spaces on the Web are moderated. In these instances, it is probably also appropriate to contact the moderators of the site prior to contacting the participants. This is analogous to contacting an organization prior to targeting individuals within that organization. Wysocki (1998), for instance, asked permission from the moderator of a sadomasochist bulletin board called the "pleasure pit."

Researchers also need to be aware that some European countries require written consent. If written consent is required, then the participant could download a form and sign it off-line and then return it by fax or postal mail (Mann & Stewart, 2000).

In research about relationships and sexuality, in particular, there is the risk that the interview or survey will stress the participant too much for them to continue with the study. As with off-line research, researchers need to consider up until what point a participant can withdraw consent. The end point of withdrawal of consent might be, for instance, after the submitting of the survey, or at the conclusion of the interview the interviewer might find confirmation that the participant is happy to allow the researcher to include the transcript in the study. Social scientists should also be aware that the lack of social cues available online makes it more difficult for them to ascertain if the participant is uncomfortable. Thus one should tread carefully and possibly make an effort to check at different points in the interview if the individual is still comfortable with proceeding.

There are other issues unique to Internet research in respect to withdrawal of consent. For example, the computer could crash mid-way through an interview or survey. Mechanisms need to be put into place to allow that participant to re-join the research if desired, and consent should not be assumed (Buchanan & Smith, 1999). In circumstances such as the computer or server crashing, we might need to have a system to enable debriefing, especially if the research is asking questions of a personal nature. Nosek, Banaji, and Greenwald (2002) suggest that debriefing can be made available by providing a contact e-mail address at the beginning of the study. They also suggest providing "a 'leave the study' button, made available on every study page, [which] would allow participants to leave the study early and still direct them to a debriefing page" (p. 163). In addition, they state that participants be given a list of FAQs, since they argue that there is less opportunity to ask the sorts of questions participants typically ask in face-to-face interviews.

There are various ways we might deal with the issue of confidentiality. As with off-line research we could elect to

Table 3. Ethical practices

• Informed consent
• Withdrawal of consent
• Confidentiality
• Psychological safeguards

use pseudonyms to represent our participants or even request preferred pseudonyms from them. However, a unique aspect of the Internet is that people typically inhabit the Web using a screen name, rather than a real name. Can we use a screen name given that these are not real names? While they may not be people's off-line identities, individuals could still be identified by their screen names if we publish them—even if it is only recognition by other online inhabitants.

As mentioned earlier in this article, research into the areas of relationships and sexuality is likely to cause psychological distress for some. It is perhaps much more difficult to deal with psychological distress online and with individuals in other countries. Nevertheless, it is imperative that we ensure that the participant does have counseling available to them if the research has caused them distress—which sometimes might be delayed distress. This could mean that there are limits to the kinds of topics about which we interview participants online or that we restrict our sample to a particular country or region where we know of psychological services that can be available to our participants if required.

Given that research into Internet relationships and sexuality is a relatively new area, future research might also focus on how to improve ethical practices. For instance, future studies might interview potential participants about how they would prefer social scientists to conduct research. Moreover, gaining a greater understanding of how individuals perceive private and public space could also influence how we conduct future studies in this topic area.

CONCLUSIONS

In concluding, while this article has provided examples of ways forward in our thinking about virtual ethics in respect to the study of online relationships, it is by no means prescriptive or exhaustive. Rather, it is suggested here that debate over such issues should be encouraged, and we should avoid setting standards for how we conduct our Internet research without also considering the ethical implications of our work. The way forward is to not restrict the debate amongst social scientists, but to also consult the individuals we would like to and are privileged to study.

REFERENCES

Bargh, J. A., McKenna, K. Y. A., & Fitzsimons, G. M. (2002). Can you see the real me? Activation and expression of the "true self" on the Internet. *Journal of Social Issues, 58*(1), 33-48.

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World-Wide Web. *British Journal of Psychology, 90*(1), 125-144.

Cooper, A., Putnam, D. E., Planchon, L. A., & Boies, S. C. (1999). Online sexual compulsivity: Getting tangled in the net. *Sexual Addiction & Compulsivity, 6*(2), 79-104.

Ess, C., & Jones, S. (2004). Ethical decision-making and Internet research: Recommendations from the AoIR Ethics Working Committee. In E. Buchanan (Ed.), *Readings in virtual research ethics: Issues and controversies* (pp. 27-44). Hershey, PA: Information Science Publishing.

Griffiths, M. (2001). Sex on the Internet: Observations and implications for Internet sex addiction. *Journal of Sex Research, 38*(4), 333-342.

Lea, M., & Spears, R. (1995). Love at first byte? Building personal relationships over computer networks. In J. T. Wood & S. W. Duck (Eds.), *Understudied relationships: Off the beaten track* (pp. 197-233). Newbury Park, CA: Sage.

Mann, C., & Stewart, F. (2000). *Internet communication and qualitative research: A handbook for researching online.* London: Sage Publications.

McCown, J. A., Fischer, D., Page, R., & Homant, M. (2001). Internet relationships: People who meet people. *CyberPsychology and Behavior, 4*(5), 593-596.

McKenna, K. Y. A., Green, A. S., & Gleason, M. E. J. (2002). Relationship formation on the Internet: What's the big attraction? *Journal of Social Issues, 58*(1), 9-31.

NHMRC. (1999). *National statement on ethical conduct in research involving humans.* Retrieved September 25, 2002, from http://www.health.gov.au/nhmrc/publications/humans/part17.htm

Nosek, B. A., Banaji, M. R., & Greenwald, A. G. (2002). E-research: Ethics, security, design, and control in psychological research on the Internet. *Journal of Social Issues, 58*(1), 161-176.

Parks, M. R., & Floyd, K. (1996). Making friends in cyberspace. *Journal of Communication, 46*(1), 80-97.

Parks, M. R., & Roberts, L. D. (1998). 'Making MOOsic': The development of personal relationships online and a comparison to their off-line counterparts. *Journal of Social and Personal Relationships, 15*(4), 517-537.

Rice, R. E., & Love, G. (1987). Electronic emotion: Socio-emotional content in a computer mediated communication network. *Communication Research, 14*(1), 85-108.

Schneider, J. P. (2000). Effects of cybersex addiction on the family: Results of a survey. *Sexual Addiction & Compulsivity, 7*, 31-58.

Slouka, M. (1995). *War of the worlds: Cyberspace and the high-tech assault on reality.* New York: Basic Books.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet.* London: Weidenfeld & Nicolson.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal and hyperpersonal interaction. *Communication Research, 23*(1), 3-43.

Wellman, B. (2004). Connecting communities: On and off line. *Contexts, 3*(4), 22-28.

Whitty, M. T. (2002). Liar, liar! An examination of how open, supportive and honest people are in Chat Rooms. *Computers in Human Behavior, 18*(4), 343-352.

Whitty, M. T. (2003a). Pushing the wrong buttons: Men's and women's attitudes towards online and offline infidelity. *CyberPsychology & Behavior, 6*(6), 569-579.

Whitty, M. T. (2003b). Cyber-flirting: Playing at love on the Internet. *Theory and Psychology, 13*(3), 339-357.

Whitty, M. T. (2004). Peering into online bedroom windows: Considering the ethical implications of investigating Internet relationships and sexuality. In E. Buchanan (Ed.), *Readings in virtual research ethics: Issues and controversies* (pp. 203-218). Hershey, PA: Idea Group Inc.

Whitty, M. T. (2005). The 'realness' of cyber-cheating: Men and women's representations of unfaithful Internet relationships. *Social Science Computer Review, 23*(1), 57-67.

Whitty, M. T. (in press). The art of selling one's self on an online dating site: The BAR approach. In M. T. Whitty, A. J. Baker, & J. A. Inman (Eds.), *Online matchmaking.* Palgrave Macmillan.

Whitty, M. T., & Carr, A. N. (2003). Cyberspace as potential space: Considering the Web as a playground to cyber-flirt. *Human Relations, 56*(7), 861-891.

Whitty, M. T., & Carr, A. N. (2006). *Cyberspace romance: The psychology of online relationships*. Hampshire, UK: Palgrave Macmillan.

Whitty, M., & Gavin, J. (2001). Age/sex/location: Uncovering the social cues in the development of online relationships. *CyberPsychology and Behavior, 4*(5), 623-630.

Wysocki, D. K. (1998). Let your fingers to do the talking: Sex on an adult chat-line. *Sexualities, 1*(4), 425-452.

Young, K. S., Griffin-Shelley, E., Cooper, A., O'Mara, J., & Buchanan, J. (2000). Online infidelity: A new dimension in couple relationships with implications for evaluation and treatment. *Sexual Addiction & Compulsivity, 7*(5), 59-74.

Young, K. S., Pistner, M., O'Mara, J., & Buchanan, J. (1999). Cyber disorders: The mental health concern for the new millennium. *CyberPsychology & Behavior, 2*(5), 475-479.

KEY TERMS

Bebo: A social networking site where members can communicate with school and university friends, connect with friends, share photos, comment on others sites and photos, and write a blog.

Blog: Online diaries on a Web page, where the blogger updates entries, typically fairly regularly, in reverse chronological sequence.

Chat Room: A Web site, or part of a Web site, that allows individuals to communicate in real time.

Cybersex: Two or more individuals using the Internet as a medium to engage in discourses about sexual fantasies. The dialogue is typically accompanied by sexual self-stimulation.

Hot-Chatting: Two or more individuals engaging in discourses that move beyond light-hearted flirting.

Lurker: A participant in a chat room or a subscriber to a discussion group, listserv, or mailing list who passively observes. These individuals typically do not actively partake in the discussions that befall in these forums.

MUDs and MOOs: Multiple-user dungeons, or more commonly understood these days to mean multi-user dimension or domains. These were originally a space where interactive role-playing games could be played, very similar to Dungeons and Dragons.

MySpace: A social networking site where members can communicate with school and university friends, connect with friends, share photos, comment on others' sites and photos, and write a blog.

Online Sexual Activity: Using the Internet for any sexual activity (e.g., recreation, entertainment, exploitation, education).

Screen Name: A screen name can be an individual's real name, a variation of an individuals' name, or a totally made-up pseudonym. Screen names are especially required on the Internet for applications such as instant messaging.

IS Project Management Contemporary Research Challenges

Maggie McPherson
University of Sheffield, UK

INTRODUCTION

Although project management is often said to have its roots in other traditional fields, such as construction, Morris (2002) asserts that modern project management practices have their origins in the 1950s US aerospace agencies. Much has been written about Information System (IS) / Information Technology (IT) project initiatives in both the public and private sectors. In fact, many information systems frequently fall short of their requirements, and are, more often than not, costlier and arrive later than anticipated, if indeed they are completed at all. For instance, according to a report for the Organization for Economic Co-operation and Development (2001), failures of major IT investments and key systems development projects have raised concerns for the achievement of service improvement through information technology. Additionally, it has been argued that failures in IT projects are more common than failures in any other aspect of modern business (Nulden, 1996). The widely-cited Standish Group (1994) study, carried out in the US, classified IT projects as follows:

- **Resolution Type 1 (Project Success)**: The project is completed on-time and on-budget, with all features and functions as initially specified.
- **Resolution Type 2 (Project Challenged)**: The project is completed and operational but over-budget, over the time estimate, and offers fewer features and functions than originally specified.
- **Resolution Type 3 (Project Impaired)**: The project is cancelled at some point during the development cycle.

The report estimated the success rate was only 16.2%, while challenged projects accounted for 52.7%, and impaired projects (cancelled) amounted to 31.1%. Since large complex projects in any area are difficult to organize, it could be said that the level of abstraction required often leads to a lack of understanding between all stakeholders involved with the project. Callahan and Moretton (2001) describe software design as being "in the code". They assert that since it is not visible, it makes it hard to use software design as a focal point for development project coordination and integration, unlike many physical designs which can be made visible to all project participants. As a result of this "invisibility", managing the development of an IS project is arguably more problematic than project management within the manufacturing sector because software development is often a highly conceptual and complex process.

Indeed, a lack of adequate project management knowledge could be said to be a major contributing factor to unsuccessful IS projects. For instance, as project managers should be aware, unless specific objectives and clear-cut end points have been set, it can be difficult to know if a milestone has been reached and indeed if the required end-product has been produced. However, making use of proprietary tools such as Microsoft™ Project is sometimes mistakenly thought of as project management, whereas real project management expertise goes beyond the mere production of Gantt or Pert (Program Evaluation Review Technique) charts, which simply represent project activities in the form of bar charts or flow diagrams. As Mandl-Striegnitz et al. (1998) point out, important project management techniques include estimation of costs and explicit identification of risks. Clearly, there is a need for more in-depth research to gain a better understanding relating to the complex role of project management within the whole IS design and development process. This discussion considers how these problems affect contemporary IS project management research and explores the methodological approaches open to researchers carrying out investigations in this area.

BACKGROUND

In order to better understand the challenges facing researchers of Information Systems Project Management (ISPM), it is necessary to explore what is meant by some of these terms. As stated by the American National Standard for Telecommunications (2000), an IS is "an organized assembly of resources and procedures united and regulated by interaction or interdependence to accomplish a set of specific functions, whether automated or manual, that comprises people, machines, and/or methods organized to collect, process, transmit, and disseminate data that represent user information". In its simplest terms, an IS can be described as a human activity or social system, which may or may not involve the use of computer systems; although, these days the former is more likely. According to Stoner et al. (1994), management can be regarded as a process of planning, organizing, leading and controlling the efforts of staff and other resources in order to

achieve organizational goals, and the Association for Project Management (2000) describes a project as a distinct set of coordinated activities "… with definite starting and finishing points, undertaken by an individual or organization to meet specific objectives within defined time, cost and performance parameters". By combining these terms, a definition for ISPM could be said to be the process of managing the creation of an IS through the establishment of project goals; organizing, leading, co-coordinating the efforts of staff processes and tasks; and controlling other resources to achieve a set of agreed objectives.

Since IS projects are frequently comprised of multi-disciplinary teams of people, a definition of what is meant by a team in this particular context is called for. Geddes et al. (1993), regard a team as comprising those individuals who have a significant contribution to make to the successful achievement of the project, whether this is through technical or specialist expertise; sponsorship, political support or sponsorship; or expectation of, and interest in, outcomes. Programmers and associated staff are often selected according to their ability to demonstrate the appropriate technical knowledge, which does not guarantee proficiency in managing successful projects. Despite the emphasis on team leadership ability, senior developers/project managers are often promoted from the programming team, with a continued emphasis on technical expertise (Mandl-Striegnitz et al., 1998).

In reality, IS project managers must not only be able to plan and break activities down into components that can be understood and to control tasks and monitor risks, but must additionally be able to consider people and process issues requiring significant team-building skills. Although IS may be implemented by staff with technical competence, they may well lack the necessary abilities to evaluate organizational contexts and analyze corresponding behaviors.

Nevertheless, since 1994 there has been an improvement in project management outcomes. By 2001, the Standish Group published another report stating that project time and cost overruns had reduced significantly. Although this improvement in project results was confirmed by a UK-based survey (Saur & Cuthbertson, 2004), the authors acknowledged that their sample could have been unrepresentatively experienced, signifying a continued need for further research.

CRITERIA FOR ISPM SUCCESS

Referring to the Standish Group report "Extreme Chaos" (2001), it seems that lessons can be learned from the successes and failures of past projects which warrant further study. From extensive research, the Standish Group identified ten criteria for project success:

1. Executive support
2. User involvement
3. Experienced project managers
4. Clear business objectives
5. Minimize scope
6. Standard software infrastructure
7. Firm basic requirements
8. Formal methodology
9. Reliable estimates
10. Other criteria such as small milestones, proper planning, competent staff and ownership

In the UK based study, Sauer and Cuthbertson (2004) reported a higher project success rate than the US Standish Report (1994). Nevertheless, Sauer and Cuthbertson suggested that in order to continue this general improvement, the following recommendations ought to be adhered to:

- Project managers should:
  - Structure projects into smaller units
  - Select the right team and involving them in decision making
  - Invest time and effort in self-development
- Senior IT managers should:
  - Establish a project management focus in the organization
  - Identify the right person for project management role
  - Create appropriate career paths
  - Be accountable through more effective performance management
- Senior business managers/sponsors should:
  - Develop client understanding of project management
  - Engage more actively with projects for which they have responsibility

Some reasons for the improvements described above were costs being cut, better tools being created to monitor and control processes and, not least, project managers becoming better skilled with better management processes being used, giving rise to optimism for the future of project management. Despite the change for the better as highlighted above, the Standish Group (2001) considered "Nirvana" still to be a long way off, indicating a need for continued research. In order to select more appropriate research approaches to investigate ISPM, it is necessary to explore some of the issues specifically related to this particular field.

INFORMATION SYSTEMS PROJECT MANAGEMENT RESEARCH ISSUES

Prior to the 1950s, computing was primarily associated with scientific applications. Even after computer installation began in business environments for data-processing tasks, associated research had a tendency to be dominated by scientific approaches. Thus, despite the increasing importance of IS within modern businesses, for historical reasons the close association of IS development with IT induced many researchers to consider IS problems using methods from the natural sciences (Baskerville & Wood-Harper, 1996; Garcia & Quek, 1997) that are more suited to science laboratories. As a consequence of this scientific tradition, investigations within the field of IS appear to have had a predominance of both positivist and technical points of view, despite major criticisms (op.cit.) that may these research methods might be not always be wholly appropriate.

Keen (1984) stressed that IS is not simply the installation of a technical system in an organization. He suggested that successful implementation needs to consider institutional issues affecting its use in the ongoing context of jobs, formal and informal structures, as well as personal and group processes. Hughes and Wood-Harper (2000) concurred, rejecting the notion that the process is exclusively technical and rational. In their view, IS research should not merely pay attention to the technology and called for IS development to be understood in its situated context, i.e. to consider the domain, the organizational constraints, the social actors and the politics in situ.

ISPM RESEARCH METHODS

The debate as to whether to adopt quantitative or qualitative research methods is well-documented and will only be touched on lightly here. However, as described by Wilson (2002), one of the most contentious debates is whether to adopt the positivist view of the nature of social reality, in which social facts can be known with certainty and in which laws of cause and effect can be discovered, or whether to apply humanistic approaches which generally see social reality as constructed through social action on the part of people who undertake those acts because they have meaning for them. Conventional "scientific" research can run the risk of being reductionist (Lincoln & Guba, 1985), since complex problems are condensed in order to produce models that can provide a simplified simulation of reality. Bryman (1988) suggests that the basic choice of methodological approach is largely influenced by the type of research question being asked and according to Yin (1994) a researcher's choice of methodological approach depends on the problem at hand and the control that the researcher has over the behavioral events. Given that many researchers now believe IS to be socially constructed in particular contexts, it is thought to be important to extend research methodologies for ISPM problems beyond the positivist paradigm in order to uncover rich qualitative data.

FUTURE TRENDS

Mumford et al. (1985) argued for a methodological pluralism within IS research domains, asserting that scientific proof was being regarded as the only valid method despite the fact that many IS problems were not susceptible to these systematic methods. A decade later, Allen (1995) agreed that research methods from different paradigms can be used simultaneously or consecutively and are equally valid. Despite the historical emphasis on positivist methods in IS research, there is increasing support for developing this type of methodological mix and this is equally applicable to project management research. This is demonstrated by the fact that although the data for the Standish Group (1994) project management research was primarily collected through a survey, focus groups were conducted to augment the survey results.

In fact, Myers (1997) argues that all studies are based on some underlying assumptions about what constitutes "valid" research and it is this which should dictate which research methods are appropriate. Avegerou (2000), writing about alternative reasoning for IS, notes that the development literature argues that developing societies need to recognize the limitations of the validity of techno-economic rationality and that they ought to pursue rationalities stemming from their own value systems. IS professional roles have been based and legitimated mostly on technocratic logic, without an obligation to consider the validity of the requirements, which are normally based on the social context (Avegerou, 2000). Morris (2002) concurs, stating that project management is not a science in the full or proper sense of the word. This is contrary to the view of Khazanchi and Munkvold (2000) who believe that methodological and philosophical diversity do not preclude researchers from making scientific inquiries into the fundamental nature of IS phenomena.

Nevertheless, Morris believes that it is a discipline worthy of theoretical study, with various questions susceptible to the methods of scientific enquiry, whilst other areas will always have a large element of unpredictability. He therefore considers that some knowledge of this field will always be personal and experiential. Following this line of reasoning, prominent researchers such as Galliers (1992) have proposed that an interpretative stance for IS research would be wholly applicable. Correspondingly, other researchers have explored and recommended various interpretivist approaches as being particularly suitable. These methodologies include case study research (Yin, 1994), where the researcher interprets data

without direct involvement, and action research (Baskerville, 1999), where the researcher is actively involved.

However, Garcia and Quek (1997) stress the need for further critical awareness, stating that importing methods into the IS field is not a simple task. They warn that without critical awareness, there is a danger of methods becoming stereotyped or distorted. Nonetheless, they supported the use of multiple methods to correspond with the complexity of research investigation which will allow a better understanding of the different aspects involved in the constitution of the object under investigation.

CONCLUSION

If one accepts that IS are socially constructed, as many researchers now appear to do, then it follows that evaluative research of IS projects needs to be situated in contextualized and authentic settings. It would therefore seem highly appropriate to set aside the positivist versus interpretivist debate, since Marcella and Knox (2004) suggest that "…it is only based upon a much fuller and more precise understanding of the complex and multifaceted needs of all users, internal and external, in all functional areas of the institution, that systems will be developed which are truly responsive and which function to meet overall … objectives". With this in mind, it seems reasonable to suggest that it is worthwhile combining diverse research methods, as endorsed by Fitzgerald and Howcroft (1998), with a view to maximizing their complementary strengths. This would seem to be particularly appropriate to address many of the concerns and issues relating to ISPM as highlighted in this discussion.

REFERENCES

Allen, D. (1995). Information systems strategy formation in higher education institutions. *Information Research*, *1*(1). Retreived March 14, 2004, from *http://InformationR.net/ir/1-1/paper3.html*

American National Standard for Telecommunications. (2000). *Telecom glossary*. Retreived March 14, 2004, from *http://www.its.bldrdoc.gov/projects/t1glossary2000/*

Association for Project Management. (2000). *Glossary of project management terms*. Retreived March 14, 2004, from *http://www.apm.org.uk/resources/p.htm*

Avegerou, C. (2000). Recognizing alternative rationalities in the deployment of information systems. *The Electronic Journal on Information Systems in Developing Countries*. Retrieved March 10, 2004, from *http://www.ejisdc.org*

Baskerville, R.L. (1999). Investigating information systems with action research., *Communications of the Association for Information Systems*, *2*(19). Retreived March 19, 2004, from *http://www.cis.gsu.edu/~rbaskerv/CAIS_2_19/CAIS_2_19.html*

Baskerville, R.L. & Wood-Harper, A.T. (1996). A critical perspective on action research as a method for information systems research. *Journal of Information Technology*, *11*, 235-246. Retreived March 14, 2004, from *http://taylorandfrancis.metapress.com/media/fcjlvprqvam991h5tw5w/Contributions/5/N/G/E/5NGER4X63FYQE50N.pdf*

Bryman, A. (1988). *Quantity and quality in social research*. London: Unwin Hyman.

Callahan, J. & Moretton, B. (2001). Reducing software product development time. *International Journal of Project Management*, *19*(1) 59-70.

Fitzgerald, B. & Howcroft, D. (1998). Competing dichotomies in IS research and possible strategies for resolution. In R. Hirschheim, M. Newman, & J.I. DeGross, (Eds.). *International Conference on Information Systems* (pp. 155-164). Helsinki, Finland: Association for Information Systems.

Fitzgerald, G., Hirscheim, R., Mumford, E. & Wood-Harper, A. (1985). Information systems research methodology: an introduction to the debate. In E. Mumford, R. Hirscheim, G. Fitzgerald, & A. Wood-Harper, (Eds.). *IS - a doubtful science? Research methods in information systems* (pp. 3-9). North Holland: Elsevier Publishers.

Galliers, R. (1992). Choosing information systems research approaches. In R. Galliers (Ed.), *Information systems research: Issues, methods and practical guidelines* (pp. 144-162). Oxford: Blackwell Scientific.

Garcia, L. & Quek, F. (1997) Qualitative research in information systems: Time to be subjective? In *Proceedings of IFIP WG8.2 Working Conference on 'Information Systems & Qualitative Research' Philadelphia, USA; 31 May-03 June 97*. Chapman and Hall, London. Available online at: *http://is.lse.ac.uk/iswnet/pub/ifip8297.htm* [last accessed 12 August 2004].

Geddes, M., Hastings, C. & Briner, W. (1993). *Project leadership*. Gower.

Hughes, J. & Wood-Harper, T. (2000). An empirical model of the information systems development process: A case study of an automotive manufacturer. *Accounting Forum. 24*(4), 391-406.

Keen, P. (1984). VDT's as agents of change. In J. Bennett, D. Case, J. Sandelin, & M. Smith (Eds.), *Visual display terminals*. London: Prentice Hall.

Khazanchi, D. & Munkvold, B. (2000). Is information sys-

tems a science? An inquiry into the nature of the information systems discipline. *The DATA BASE for Advances in Information Systems, 31*(3) 24-42.

Lincoln, Y. & Guba, E. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.

Mandl-Striegnitz, P., Drappa, A. & Lichter, H. (1998). Simulating software projects - An approach for teaching project management. In C. Hawkins, M. Ross, G. Staples, & J.B. Tompson (Eds.), *Proceedings of INSPIRE '98 (International conference on Software Process Improvement - Research into Education and training)*,(pp. 87–98). London: University of Sunderland.

Marcella, R. & Knox, K. (2004). Systems for the management of information in a university context: an investigation of user need. *Information Research, 9*(2). Paper 172. Retreaived March 27, 2004, from *http://InformationR.net/ir/9-2/paper172.html*

Morris, P.W.G. (2002). ICE James Forest lecture: science, objective knowledge, and the theory of project managemet. *ICE Civil Engineering, 150*(May), 82-90.

Myers, M. D. (1997). Qualitative research in information systems. *MIS Quarterly,* (21:2) 241-242. MISQ Discovery. Archival version, June 1997, Retrieved March 12, 2004, from *http://www.misq.org/discovery/MISQD_isworld/*. MISQ Discovery. Updated version, last modified: *www.qual.auckland.ac.nz*

Nuldén, U. (1996). Escalation in IT projects: Can we afford to quit or do we have to continue.*Information Systems Conference of New Zealand* (pp.136-142). Palmerston North, New Zealand: IEEE Computer Society Press.

OECD. (2001). Management of large public IT projects: Case studies. *Public management service, public management committee report*. In J. Kristensen (Ed.), Organization for Economic Co-operation and Development. Retreived March 12, 2004, from *http://www.olis.oecd.org/olis/2001doc.nsf/LinkTo/PUMA-SBO-RD (2001)1*

Sauer, C. & Cuthbertson, C. (2004). The state of IT project management in the UK 2002 - 2003. *ComputerWeekly.com Ltd*. Retreived March 15, 2004, from *http://www.cw360ms.com/pmsurveyresults/index.asp*

Standish Group. (1994). *The CHAOS Report*. Retreived March 15, 2004, from *http://standishgroup.com/sample_research/chaos_1994_1.php*

Standish Group. (2001). *Extreme CHAOS*. Retreived from *http://www.standishgroup.com/sample_research/PDFpages/extreme_chaos.pdf*

Stoner, J.A.F., Yetton, P.W., Craig, J.F. & Johnston, K.D. (1994). *Management*. Sydney, Australia: Prentice Hall.

Wilson, T. (2002). Information science and research methods. In J. Steinerová, & S. Kimlika (Eds.), Knižnicná a informacná veda (*Slovak* Library and Information Science), (pp. 63-71). Bratislava, Slovak Republic: Department of Library and Information Science, Comenius University.

Yin, R. (1994). *Case study research, design and methods* (2nd ed.). Newbury Park: Sage Publications.

KEY TERMS

Case Study Research: An in-depth investigation that attempts to captures lessons learned through studying the environment, procedures, results, achievements, and failures of a particular project or set of circumstances.

Development Literature: Literature about impoverished countries of the world that are trying to modernize or to find different ways of supporting their populations.

Focus Group: A small group interview, conducted by a moderator, which is used to discuss one or more issues.

Humanism: A philosophical approach that focuses on human value, thought, and actions.

Information Systems Project Management: The process of managing the creation of an IS through the establishment of project goals; organizing, leading, co-coordinating the efforts of staff processes and tasks; and controlling other resources to achieve a set of agreed objectives.

Interpretivism: A research approach that attempts to reach an understanding of social action in order to arrive at a causal explanation of its course and effects.

Positivism: A belief that natural science, based on observation, comprises the whole of human knowledge.

Project Team: All individuals who have made a significant contribution to make to the successful achievement of the project.

Project: A distinct set of coordinated activities with definite starting and finishing points, undertaken by an individual or organization to meet specific objectives within defined time, cost and performance parameters.

Techno-Economic Rationality: Logical justification for making a connection between technical advances and economic growth.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1673-1678,

Isochronous Distributed Multimedia Synchronization

Zhonghua Yang
Nanyang Technological University, Singapore

Yanyan Yang
University of California, Davis, USA

Yaolin Gu
Southern Yangtze University, China

Robert Gay
Nanyang Technological University, Singapore

INTRODUCTION

A multimedia system is characterized by the integrated computer-controlled generation, manipulation, presentation, storage, and communication of independent discrete and continuous media data. The presentation of any data and the synchronization between various kinds of media data are the key issues for this integration (Georganas, Steinmetz, & Nakagawa, 1996). Clearly, multimedia systems have to precisely coordinate the relationships among all media that include temporal and spatial relationships. Temporal relationships are the presentation schedule of media, and spatial relationships are the location arrangements of media. Multimedia synchronization is a process of maintaining these relationships by employing appropriate synchronization mechanisms and algorithms. Multimedia synchronization is traditionally challenging, especially in distributed environments.

Three types of multimedia synchronization can be distinguished: intrastream synchronization, interstream synchronization, and intermedia synchronization (Crowcroft, Handley, & Wakeman, 1999). The approaches used for interstream synchronization can also be used for intermedia synchronization.

The word *synchronization* refers to time. The easiest way of synchronizing between streams at different sites is to use a single time reference. There are several ways to provide this time reference.

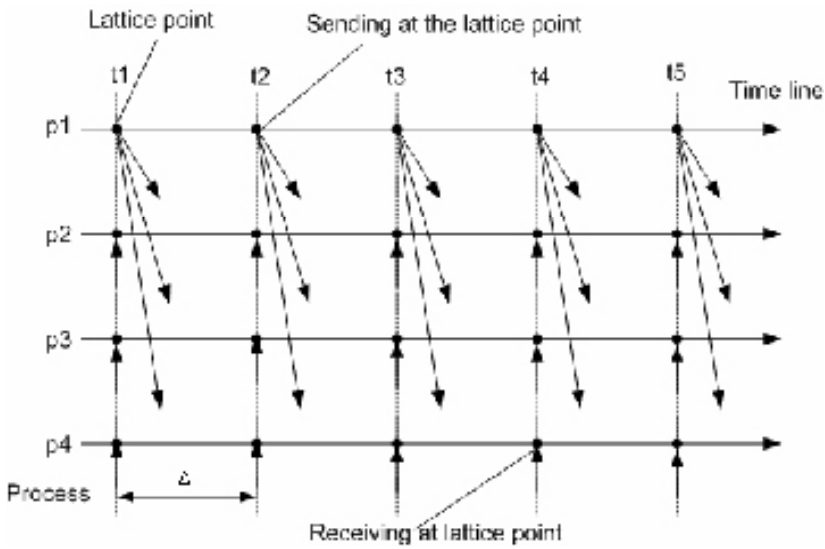
- The network will have a clock serve as a single reference. This approach is used in H.261/ISDN- (integrated services digital network) based systems. A single clock time is propagated around a set of codecs and multipoint control units (MCSs).
- The network deploys a clock-synchronization protocol, such as NTP (the network time protocol; Mills, 1993). The time stamps of media packets will be derived from the globally synchronized clocks. The isochronous synchronization approach as described in this article heavily relies on this time reference.

AN ISOCHRONOUS SYNCHRONIZATION APPROACH

The isochronous synchronization approach employs a clock-driven protocol for achieving multimedia synchronization (any one of three types of synchronization; Yang, Gay, Sun, Siew, & Sattar, 2002). This approach is particularly suitable for distributed collaborative multimedia environments where many-to-many multimedia communication is the basic interaction pattern. In this approach, multimedia synchronization is based on the use of synchronized physical clock time instead of any form of logical clock or sequence numbers, and thus clock synchronization across the distributed system is assumed. A real-time (synchronized) clock is incorporated in the system as a mechanism used for initiating significant events (actions) as a function of real time.

With globally synchronized clocks that satisfy the granularity condition, we can construct an *action lattice* (or *event lattice*; Kopetz, 1992). One dimension of this lattice represents the progression of time, the other dimension is the processes in the system (Figure 1). Processes in the system are designed to execute a simple *clock-driven protocol*, which requires that the events of sending and receiving messages are restricted to only occur at the lattice point of the globally synchronized space-time lattice (Figure 1). Thus, whenever an action has to be taken, it has to be delayed until the next lattice point of the event lattice.

Figure 1. Lattice structure



This lattice structure greatly simplifies multimedia synchronization and readily maintains the temporal and causal relationship among the media.

The idea behind the clock-driven, isochronous synchronization is very simple and intuitive in that the easiest way to synchronize processes is to get them all to do the same thing at the same time. Using the simple mechanism based on the synchronized clock without requiring complex algorithms, the approach can equally well be applied to various multimedia applications in distributed environments, including live multimedia applications (live teleconferencing and CSCW) and stored media applications.

THE ORDERING PROPERTIES OF THE SYNCHRONIZATION PROTOCOL

In essence, what is really required for distributed multimedia synchronization is *order*; that is, a synchronization protocol must ensure that multimedia messages or streams are sent, delivered, and presented in an order that is consistent with the expected behavior of the distributed multimedia system as a whole. Clearly, multimedia systems have to precisely coordinate the relationships among all media. These relationships include temporal and spatial relationships. Temporal relationships are the presentation schedule of media, and spatial relationships are the location arrangements of media.

There are two specific cases concerning temporal order: *causal order* and Δ-*causal order*. These ordering concepts are derived from a happens-before relation, which is a more fundamental notion in distributed computing. The expression $a \rightarrow b$ is read as "*a* happens before *b*" and means that all processes agree that first, event *a* occurs, then afterward, event b occurs. The happens-before relation can be observed directly in two situations in a distributed environment: (a) If *a* and *b* are events in the same process and *a* occurs before *b*, then $a \rightarrow b$ is true, and (b) if *a* is the event of a message being sent by one process and *b* is the event of the message being received by another process, then $a \rightarrow b$ is also true. Obviously, a message cannot be received before it is sent or even at the same time it is sent since it takes a finite, nonzero amount of time to arrive. Note that happens-before is a transitive relation, so if $a \rightarrow b$ and $b \rightarrow c$, then $a \rightarrow c$.

The notion of causal order, as introduced by Birman and Joseph (1994), states that for any process, the order in which it delivers messages must respect the happens-before relation of the corresponding sending of the messages. More formally, a distributed computation *E*, all of whose messages is denoted as a set $M(E)$, respects causal order if for any two messages $m1$ and $m2$ and corresponding message-sending events $send(m1)$ and $send(m\_2)$, $send(m1) \rightarrow send(m2)$. If $m1$ and $m2$ have the same destination process, then for the message delivering events, $deliver(m1) \rightarrow deliver(m2)$.

In the definition of causal order, nothing is mentioned about the time at which the messages are delivered; also, it does not prescribe what to do with the cases of message loss and late arrival. However, in a distributed multimedia system, messages have limited *validity time*, after which the

messages become useless and are allowed to be discarded. Messages that arrive at its destination within its validity time must be delivered within the expiration of its validity time and in its causal order. This motivates the notion of Δ-causal order, introduced in Yavatkar (1992) and formalized in Baldoni, Mostefaoui, and Raynal (1996). The Δ-causal order is defined as follows. A distributed computation respects Δ-causal order if (a) all messages in $M(E)$ that arrive within a time interval Δ are delivered within Δ and all the others are never delivered (they are lost or discarded), and (b) all delivery events respect causal order. That is, for any two messages $m1$ and $m2$ in $M(E)$ that arrive within Δ we have, if $send(m1) \rightarrow send(m2)$ and $m1$ and $m2$ have the same destination process, then $deliver(m1) \rightarrow deliver(m2)$.

The clock-based isochronous synchronization protocol requires the following ordering property be respected.

- **Same order:** The multimedia messages are delivered to the destinations in the same order.
- **Temporal order:** Different destinations see the different messages in the temporal order.
- **Simultaneity:** Different destinations see the same messages at about the same time.

Note that the temporal order is a prerequisite for the causal order. If and only if the occurrence of an event e_1 has preceded the occurrence of an event e_2 in the domain of real time, it is possible that e_1 has an effect on e_2. On the other hand, if it can be established that e_2 has occurred after e_1, e_2 cannot be the cause of e_1.

The basic mechanism is to use clock values as event time stamps to preserve temporal ordering of events and to achieve synchronization, and clocks in the system must have sufficient granularity or resolution. The granularity g of a synchronized clock is defined as the real-time duration between two consecutive global ticks. Obviously, the temporal order of two or more events, which occur between any two consecutive ticks of the synchronized clock, cannot be reestablished from their time stamps. This is a fundamental limit when using clock time for temporal ordering.

With globally synchronized clocks having sufficient granularity, we can construct an action lattice (or event lattice) as described above; that is, one dimension of this lattice represents the progression of time, and the other dimension is the processes in the system (Figure 1). Processes in the system are designed to execute a simple clock-driven protocol (an isochronous protocol below), which requires that the events of sending and receiving messages are restricted to only occur at the lattice point of the globally synchronized space-time lattice. Thus, whenever an action has to be taken, it has to be delayed until the next lattice point of the event lattice. This delay is the price we have to pay for the simple and intuitive synchronization protocols.

This lattice is a basic mechanism for the isochronous approach to multimedia synchronization. The lattice interval, Δ, is an important design parameter for the synchronization protocols. The following factors will affect how to choose Δ.

- The bounded end-to-end communication delay
- The validity time of multimedia, beyond which the multimedia objects become useless
- The granularity g and precision p of clock synchronization

As a general guideline, Δ must be large enough to accommodate all these factors; it must also be small enough not to unduly delay events.

AN ISOCHRONOUS SYNCHRONIZATION PROTOCOL

We now describe a general clock-driven, isochronous protocol to achieve desired multimedia synchronization. The protocol seeks to guarantee that in a time period, a set of processes will deliver the same messages at the same time and in the same temporal order. Here, "same time" must be understood to be limited by the clock skew (clock-synchronization precision) as much as π, meaning that two processes undertaking to perform the same action at the same time may in fact do so as much as π time units apart.

The protocol executes its events on every clock tick (i.e., at the lattice point), and executes a *No-Op* event by doing nothing if there is no communication event to take place on a clock tick. In practice, operations on a clock tick can be implemented by an interrupt-driven program; for example, the receipt of a message time-stamped T causes the setting of a clock interruption for $T+\Delta$, which in turn will cause the message to be processed at that time. When a process disseminates a message, it will not do so immediately; rather, it will wait until the next tick (i.e., next lattice point) on its clock and then time-stamp the message using its clock reading and send out the message. When the message arrives at its destination, it is not sufficient for the destination process to handle messages that have been received in ascending order by time-stamp. We must ensure that a process delivers messages only if no message with a smaller time stamp can be subsequently received. We say that a message is *stable for p* once no message with a lower time stamp can be delivered to the process p. Clearly, a message should be processed only after it becomes stable.

The lattice structure for the protocol operation provides a convenient way of establishing message stability. Here, testing the stability of a message can be accomplished by exploiting the bounds on delivery delays (i.e., the next lattice point) and process clocks. A message time-stamped T by

process p will be received by $T+\Delta$ at every other process in the system according to each process' local clock, which is synchronized with the others. The message that arrives later than $T+\Delta$ is considered useless and discarded (because the multimedia message is beyond its validity time).

The isochronous synchronization protocol follows the following rules.

- **Sending Rule:** A process sends out the message at every clock tick; if there is no message to send, the process will do nothing (or you can think of it as sending out a null message) at the clock tick.
- **Stability Rule:** A message is stable at process p_i if the time stamp on the message is T and the clock at p has a value equal or greater than $T+\Delta$.
- **Delivery Rule:** Stable messages are delivered to the application in ascending order by time stamp.
- **Tie-Breaking Rule:** Two messages with the same time stamp from different processes are ordered according to the process ID, which is assumed to be unique.

Note that the ordering properties as required by the synchronization are guaranteed by executing this protocol. Noticeably, the temporal order, causal order, and Δ-causal order are all respected without requiring additional sophisticated algorithms. In executing this protocol, all processes have a consistent behavior toward messages.

The isochronous synchronization approach assumes globally synchronized clocks in distributed systems. The improved NTP (version 3) enjoys synchronization to within a few tens of milliseconds in the global Internet of today; the clock synchronization for LANs (local area networks) can obtain accuracy as high as a few microseconds. The introduction of the Global Positioning System (GPS) in the 1990s has further advanced the clock-synchronization technique (Herring, 1996). GPS has introduced an inexpensive way to obtain accurate information (including time information) using a radio receiver, which consists of nothing more than a GPS receiver and a network interface. Time obtained in this manner is accurate to a few tens of microseconds. Accuracy such as this is adequate for even the most demanding real-time applications. With this development and such an accurate timing source in place, we believe that the clock-based isochronous distributed synchronization, as advocated by Lamport (1984), provides a promising yet simple and intuitive alternative.

ACHIEVING ISOCHRONOUS SYNCHRONIZATION USING RTP/RTCP

RTP (a transport protocol for real-time applications) is the real-time transport protocol within an Internet integrated-service architecture, which is designed to provide a quality-of-service guarantee beyond the current TCP/IP (transmission-control protocol/Internet protocol) best-effort service model. RTP provides end-to-end network transport functions suitable for applications transmitting real-time data (e.g., audio, video, or simulation data) over multicast or unicast network services. The data transport is augmented by a control protocol (RTCP) to allow monitoring of the data delivery in a manner scaleable to large multicast networks, and to provide minimal control and identification functionality. RTP and RTCP are designed to be independent of the underlying transport and network layers.

The noticeable feature associated with media synchronization is the 32-bit time stamp field in RTP data packets, the 64-bit NTP time stamp field, and the 32-bit RTP timestamp field in RTCP control packets. Although RTP does not mandate running the NTP to provide clock synchronization, running NTP is very useful for synchronizing streams transmitted from separate hosts. The mechanisms incorporated in RTP/RTCP enable the isochronous synchronization protocol.

One of primary functions of RTCP is to provide feedback in RTCP control packets (sender report, SR, and receiver report, RR) on the quality of the data distribution and information for intermedia synchronization. The RTP standard requires that the NTP time stamp (based on synchronized clocks) and corresponding RTP time stamp (based on data-packet sampling) are included in RTCP packets by data senders. This correspondence between the RTP time stamp and NTP time stamp may be used for intra- and intermedia synchronization for sources whose NTP time stamps are synchronized. Using the time-stamp mechanisms in RTP/RTCP and the lattice structure described in this chapter, our isochronous approach can be readily applied to multimedia systems, particularly in distributed many-to-many conferencing environments.

CONCLUSION

In distributed multimedia systems, there exist two approaches to protocol design, event driven and clock driven, and most protocols have taken an event-driven approach. While using physical time based on globally synchronized clocks for obtaining synchronization was advocated a long time ago, the clock-driven approach has not been popular in the distributed-system research community. In conjunction with the lattice structure, the isochronous protocol achieves the required synchronization, which guarantees the temporal order, including causal order and Δ-causal order, without additional sophisticated algorithms for respecting causality. The only assumption of the isochronous protocol is the synchronized clocks in a distributed system and the known bounds of the network communication delays. These assumptions can readily be satisfied with the deployment

of modern networks. The mechanisms incorporated in the Internet standard real-time protocols such as RTP/RTCP make the isochronous synchronization approach readily applicable.

REFERENCES

Baldoni, R., Mostefaoui, A., & Raynal, M. (1996). Causal delivery of messages with real-time data in unreliable networks. *Journal of Real-Time Systems.*

Birman, K., & Joseph, T. (1994). Reliable communication in the presence of failure. In K. P. Birman & R. van Renesse (Eds.), *Reliable distributed computing with the Isis toolkit* (pp. 176-200). IEEE CS Press. (Reprinted from *ACM Transactions on Computer Systems, 5*(1), 47-76, February 1987)

Crowcroft, J., Handley, M., & Wakeman, I. (1999). *Internetworking multimedia.* Morgan Kaufmann Publishers.

Georganas, N., Steinmetz, R., & Nakagawa, N. (Eds.). (1996). Synchronization issues in multimedia communications. *IEEE Journal on Selected Areas in Communications, 14*(1).

Herring, T. A. (1996). The Global Positioning System. *Scientific American, 274*(2), 32-38.

Kopetz, H. (1992). Sparse time versus dense time in distributed real-time systems. *Proceedings of the 12th International Conference on Distributed Computing Systems,* 460-467.

Lamport, L. (1984). Using time instead of timeout for fault-tolerant distributed systems. *ACM Transactions on Programming Languages and Systems, 6*(2), 254-280.

Mills, D. L. (1993). Precision synchronization of computer network clocks. *ACM Computer Communications Review, 24*(2), 28-43.

Yang, Z., Gay, R., Sun, C., Siew, C. K., & Sattar, A. (2002). An isochronous approach to multimedia synchronization in distributed environments. In S. M. Rahman (Ed.), *Multimedia networking: Technology, management, and applications* (chap. 16). Hershey, PA: Idea Group Publishing.

Yavatkar, R. (1992). MCP: A protocol for coordination and temporal synchronization in multimedia collaborative applications. *Proceedings of International Conference on Distributed Computing Systems*, 606-613.

KEY TERMS

Clock Synchronization: Physical clocks in a network are synchronized to within certain precision and accuracy. The precision refers to the difference between readings of clocks, and the accuracy refers to the difference between the clock reading and the universal standard time.

Intrastream Synchronization: This, also called play-out synchronization, ensures that the receiver plays out the medium a fixed time after it was generated at the source and it experienced variable end-to-end delay. In other words, intrastream synchronization assures that a constant-rate source at the sender again becomes a constant-rate source at the receiver despite delay jitter in the network.

Intermedia Synchronization: This is concerned with maintaining the requirements of the temporal relationships between two or more media. Lip synchronization between video and audio is an example of interstream synchronization where the display of video must synchronize with audio.

Interstream Synchronization: This ensures that all receivers play the same segment of a medium at the same time. Interstream synchronization may be needed in collaborative environments. For example, in a collaborative session, the same media information may be reacted upon by several participants.

Isochronous: The term refers to time-dependent processes where data must be delivered within certain time constraints. For example, multimedia streams require an isochronous transport mechanism to ensure that data is delivered as fast as it is displayed, and to ensure that the audio is synchronized with the video. Isochronous processes can be contrasted with asynchronous processes, which refers to processes in which data streams can be broken by random intervals, and synchronous processes, in which data streams can be delivered only at specific intervals. Isochronous service is not as rigid as synchronous service, but not as lenient as asynchronous service.

Multimedia: This term is used to indicate that the information and data being transferred over the network may be composed of one or more of the following media types: text, images, audio, and video.

NTP: NTP stands for network time protocol, and it is a standard Internet protocol used to synchronize the clocks of computers to some time reference.

Stream: This technique is for transferring data such that it can be processed as a steady and continuous stream. Streaming technologies are becoming increasingly important with the growth of the Internet because most users do not have fast-enough access to download large multimedia files quickly. If the stream is for transferring multimedia data, it is called a *multimedia stream*.

Validity Time: This is a time interval within which the message remains valid, available, and useful to its recipients. After the validity time of a message, the message becomes

useless and may be discarded. The notion of validity time is important in multimedia communication.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1679-1684, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: IT Education

Issues in Using Web-Based Course Resources

Karen S. Nantz
Eastern Illinois University, USA

Norman A. Garrett
Eastern Illinois University, USA

INTRODUCTION

Education over the Internet is going to be so big it is going to make e-mail usage look like a rounding error.

John Chambers, Cisco Systems, New York Times, November 17, 1990

Web-based courses (Mesher, 1999) are defined as those where the entire course is taken on the Internet. In some courses, there may be an initial meeting for orientation. Proctored exams may also be given, either from the source of the Web-based course or off-site at a testing facility. The Internet-based course becomes a virtual classroom with a syllabus, course materials, chat space, discussion list, and e-mail services (Resmer, 1999). Navarro (2000) provides a further definition: a fully interactive, multimedia approach. Current figures indicate that 12% of Internet users in the United States use the Internet to take an online course for credit toward a degree of some kind (Horrigan, 2006). That number is indicative of the rapid proliferation of online courses over the past several years.

The Web-enhanced course is a blend with the components of the traditional class while making some course materials available on a Web site, such as course syllabi, assignments, data files, and test reviews. Additional elements of a Web-enhanced course can include online testing, a course listserver, instructor-student e-mail, collaborative activities using RSS feeds and related technologies, and other activities on the Internet.

One of the biggest concerns about Web-based courses is that users will become socially isolated. The Pew Internet and America Life Project found that online communities provide a vibrant social community (Horrigan, Rainie, & Fox, 2001). Clearly, students are not concerned or feel that other benefits outweigh the potential drawbacks. According to government research (Waits and Lewis, 2003), during the 2000-2001 academic year alone, an estimated 118,100 different credit courses were offered via distance education (with the bulk of that using Internet-based methods) by 2- and 4-year institutions in the United States. Over 3 million students were registered in these courses.

Navarro (2000) suggests that faculty members are far more likely to start by incorporating Internet components into a traditional course rather than directly offering Web-based courses. These Web-enhanced courses might be considered the transition phase to the new paradigm of Internet-based courses. Rich learning environments are being created, with a shift from single tools to the use of multiple online tools, both to enhance traditional courses and to better facilitate online courses (Teles, 2002).

BACKGROUND

A 1999 research study showed that 27.3% of the faculty members thought they used the Internet for the delivery of course materials, but only 15.6% actually did so. Of this group, the major use was simply the substitution of a Web page for the printed page. Most faculty members (73.8%) updated their sites so infrequently that the sites only served to replicate printed handouts. In a follow-up study at the same university, the number of faculty who used Web pages to enhance their courses showed a decrease from the previous year (Garrett, Lundgren, & Nantz, 2000). In the same study, 22% of the faculty were never planning to use a Web site for delivery of any portion of their courses. Less than 5% were truly incorporating Web technology into their courses in a meaningful way. Lee Raines, Director of the Pew Internet and American Life Project notes that the role of experts, such as teachers, has changed. The Internet has empowered amateurs. New teaching models and methods have developed as educators try to adjust to changing student attitudes (Rainie, 2006). The new educational model becomes "the net-savvy, well-connected, teacher-independent end-user" (Castells, p. 20).

Overall, Internet penetration for U.S. adults is up to 73% as of April 2006, up 9% in just one year. In addition, "… the 40% in home broadband adoption from March 2005 to March 2006 is double the 20% rate of increase that occurred from March 2004 to March 2005" (Horrigan, 2006). For college age degreed adults, 91% go online regularly (Rainie, 2006). Researchers at Ball State University found that 30% of a waking day is spent with media as the sole activity with an additional 39% spent with media combined with some other

activity ("Average…", 2005). Fully one third of all Internet users in the U.S. say that the Internet has greatly improved the way they pursue hobbies and interests (Madden, 2006) and each day 44% of all Americans are online at some point, up from 36% in 2002 (Horrigan & Rainie, 2006).

Part of the expectation of the current college population is that two-way technologies are the norm (instant messaging, Weblogs, and online journaling, for example) and that online communities provide a rich environment for information sharing. According to Pew data, almost half of Internet users access listservs, RSS feeds, and bulletin boards to stay engaged. This shift to more collaborative tools provides new opportunities but creates numerous challenges. Learning management systems (LMS) are adding collaborative tools to reflect the changing habits of Internet users. All of the popular LMS tools, such as WebCT, Blackboard, and Moodle provide for online discussions, information posting, group assignments, synchronous chats, interactive quizzes, and a closed e-mail system. Students perceive collaborative activities, both synchronous and asynchronous, as cutting edge. Castell and Wellman refer to this synchronous and asynchronous environment as "networked individualism" (Castells, p. 20). In Figure 1, Garrett (2006) presents a breakdown of the myriad tools available in various combinations of synchronous/asynchronous and interactive/non-interactive. With these tools available in an almost endless variety of combinations, classroom experiences can be tailored to suit the content as well as the student learning styles (see Baggaley, 2003, for some examples).

Clearly, there are many compelling reasons to use Web-based resources in a course including greater efficiency in the delivery of materials, providing up-to-the-minute content, enhanced status for the course and faculty, fostering student-to-student collaboration, and the use of technologies with which the students are increasingly familiar and comfortable.

Despite the quantum leap in Internet technology adoption, some of the familiar problems still exist. Faculty still must adapt to a looser teaching environment. No longer are lectures delivered from a raised lectern, enough. The expectation by students is that the classroom paradigm has shifted, and faculty must adapt to a looser, more flexible teaching environment. Some of the issues inhibiting the use of Web-based resources include: lack of faculty knowledge of Web page design, html, server sites, and file transfer protocols (Nantz & Lundgren, 1998); perceived need for Web glitz to provide entertainment along with content such as highly interactivity, animation, audio, and video streaming; lack of accessibility to Web resources for both faculty and students (Rao & Rao, 1999); sufficient training for faculty (Rups, 1999); and compensation for cyberprofs who typically spend twice as much time developing and teaching Web-based courses for no extra pay (Navarro, 2000).

Carr notes that the high drop rates in online courses may result from faculty inexperience with the new classroom paradigm (Carr, 2000). Also, the need to continually retool to stay even with the student use of technology is daunting.

Illinois State University identified five major issues driving Web-delivered courses:

- Technology needs to be driven by sound pedagogical goals.
- Technology tools need to address a specific pedagogical task with technical expertise available.
- Faculty want and need to interact with peers who are doing similar tasks.
- Hardware must support teaching without frustrating students and faculty.
- Faculty need recognition for technology adoption ("Average…". 2005).

A Course Web Site Classification

Courses using Web-based resources can be classified in six different levels. At the top levels are the Internet-based classes (i.e., the course was created and organized to be Web delivered). The middle levels involve a Web class that uses the Internet for delivery of content and communication among the course registrants, but also uses face-to-face meetings for some classes, orientation, and testing. At the lowest level,

Figure 1. Instructional Communication (Adapted from Garrett, 2006)

	Non-interactive	***Interactive***
Synchronous	Lecture Web casts Videos	Discussion Managed Meetings IRC Chat Internet Messaging (IM) Webinars
Asynchronous	Podcasts / Vodcasts Webcasts Wikis	Discussion Boards Weblogs RSS Feeds / Syndication Cellular Text Messaging (SMS)

Table 1. Classification of academic Web pages

Level	Description	Typical Content	Maintenance Level Required
1	Traditional course presentation, basic-level course materials on Web—internal links	Instructor data (name, phone, office hours, e-mail address) course materials (syllabus, generic schedule, assignments); non-interactive	Low—static pages after initial upload. Low-volume e-mail correspondence.
2	Traditional course presentation—intermediate-level course materials on Web—external links	All Level 1 Some external links, such as textbook and reference sites; non-interactive.	Low—mostly static pages with occasional updates and checking of external links. Low-volume e-mail correspondence.
3	Traditional enhanced course presentation—intermediate-level course materials on Web and Web content delivery	All Level 2 All traditional course materials posted. Web access in class used for delivery of some course content. Some assignments/requirements involve interaction, e.g., e-mail submissions, listserv postings.	Weekly updates to schedule, FAQ, course materials, notes to students. Medium-volume e-mail correspondence.
4	Traditional enhanced course presentation—complete Web content and materials	All Level 3 Course Presentations and lectures dynamically available on Web. Data files, links, programs on Web for students. Forms for student "reply" assignments, course evaluations, etc. Link to course grades.	2-3 times per week. Regular updating of grades. Medium-volume e-mail correspondence.
5	Web-delivered course with orientation and testing meetings	All Level 4 plus any additional materials to allow for full Web delivery of course including audio and video augmentation; multimedia CD's. Few or no regular classes—orientation meeting may be necessary. Testing may be proctored off-site or unproctored on the Web.	Daily maintenance and access by instructor. High-level of e-mail correspondence. Regular updating of grades and course materials.
6	Virtual class	All Level 5 plus online testing and orientation. Discussion, chat groups, list serve, e-mail, and other interactive tools;. Teleconferencing. No class meetings.	Substantial daily maintenance (average 1-3 hours) by instructor including all course aspects. High-level of e-mail correspondence.

some course materials are simply presented in a hypertext format that replace traditional printed handouts. Table 1 shows the classification levels of academic Web pages by typical content and maintenance levels.

The six levels previously presented indicate progression from the most basic Web-enhanced course to a course delivered fully on the Internet. Faculty would likely proceed through the levels to reach level 4 for traditional classes unless limited by resources, expertise, and administrative factors. Levels 5 and 6 require significant changes in the academic structure and considerable support of the academic computing environment. The following table summarizes the resources that would be involved in the process of moving courses to the Web.

Although Table 2 shows a summary of the typical resources faculty need to develop Web course materials at varying levels, there are other elements that will be just as important in achieving a specific level of Web course expertise. The following list illustrates some of the issues involved. For a more comprehensive discussion, see Nantz and Lundgren (2003).

Issues Inhibiting Web-Enhanced Courses and Recommendations

- Be realistic about your own level of expertise and the instructional support you have available. Convert print-based materials to html using Word or some other

Table 2. Resources involved in moving to the Web

Level	Description	Resources Needed
1	Traditional course presentation, basic-level course materials on Web—internal links	Basic computer literacy, Web browsing experience. Course site can be created by the faculty member, professional designers, by use of Web course applications such as BlackBoard, WebCT, or the open-source Moodle.
2	Traditional course presentation—intermediate-level course materials on Web—external links	Experience with preceding level. Web application packages can be extended or with additional training, a general Web development package like MS FrontPage or DreamWeaver can be used.
3	Traditional enhanced course presentation—intermediate-level course materials on Web and Web content delivery	Experience with preceding level. Commitment to regular maintenance. Knowledge of e-mail attachments, listserv maintenance, or other interactive Web applications. Both Web application and general Web development packages can be extended for this level.
4	Traditional enhanced course presentation—complete Web content and materials	Experience with preceding level. Professional Web applications may not be able to accommodate this level without considerable difficulty. Usually requires considerable expertise with general Web development packages and some knowledge of HTML and programming concepts include Javascript, ASP, and XML. In addition, a working knowledge of social networking tools such as blogs and wikis, RSS feeds, and XML might be helpful.
5	Web-delivered course with orientation and testing meetings	All of the above. No additional faculty resources required; academic structural change to allow for registration and other student activities online.
6	Virtual class	Use of a sophisticated commercial Web course package that allows for secure online testing; considerable administrative support and faculty expertise in the selected package.

familiar software. Once a comfort level is achieved, incorporate other html code using simple programs like Netscape Composer. Cut and paste code from sample Web pages. Extend knowledge to knowledge of common gateway interface (CGI) scripts, Java, or XML (Extensible Markup Language). Use university's instructional support personnel to help you set up simple Web pages that are at your comfort level for maintenance.

- Be realistic about the cost in time and money to maintain course Web sites. The development of a full Web-delivered course may be as high as $115,000 (Navarro, 2000). Marchese (1998) suggests a range of $12,000 to $90,000 per credit hour. Any Web platform provider will charge a licensing fee, which may be based on the number of students. The equivalent of a 1-hour lecture may require 24 hours for writing, recording, and editing and up to 162 hours for full multi-media support.
- Be realistic about your access and the students' access to technology. Any administration who wants Web-delivered coursework must provide adequate technology to support it, either through on-campus servers or an off-campus Web host. Do not assume that all students have broadband connections at home, or their own computers on which to install specialized software.
- Be realistic about converting paper-based content to Web content. A direct conversion usually doesn't work well. The visual indicators on printed materials (headers, footers, page numbering) don't convert well to html. Content and access must be re-evaluated. Powerpoint slides can be posted, but good slide design means key points only. Substantial notes must be provided to expand on the slides. Some students see the notes as ancillary and don't get the depth of content. Text-based sites are seen as boring. An academic course site that simply creates text on a Web page defeats the purpose of using Web pages—of having the ability to create links to interesting sites, to provide graphics, to provide sound and video. Providing hundreds of pages of text is the death knell for a Web class.
- Be realistic about the use of Web-based multimedia materials. Unless these are professionally developed by staff that are familiar with the software, they are often viewed by students as low-quality even through they can be extremely time-consuming to produce. Overuse of multimedia can be counterproductive and can also become a storage and transmission burden, as multimedia can consume large amounts of storage and bandwidth. A good balance of quality multimedia

with other materials can significantly enhance learning in a class, but great care must be taken in the selection and application of the media.

- Be realistic about your expectations of a Web-enhanced course. Our research shows that when course content is placed on the Web, student attendance will drop by as much as 50%. Students see printing Web pages as a substitute for class attendance (Lundgren & Lundgren, 1996).
- Keep in mind that students may be more familiar with the technologies than you are. Many of them have grown up with technology and are more used to it than their older instructors. Recent statistics (Hitlin & Rainie, 2005) show that 87% of people in the U.S. aged 12 to 17 use the Internet and 78% of those have used it at school. Students, for example, are used to social networking and peer-to-peer collaboration, and studies show that they prefer synchronous interaction (such as instant messaging) to asynchronous interaction, such as e-mail (Lenhart, Madden, & Hitlin, 2005). You can use that to your advantage in a course if you use such tools as blogs, wikis, and RSS feeds (Harrsch, 2003), which are highly collaborative tools. These allow a higher degree of asynchronous collaboration than was previously possible and the technologies can be implemented at low cost. Similarly, instant messaging and chats provide the ability for synchronous interaction that is very much a part of the lives of many college students. This familiarity can be leveraged for classroom use with appropriate, and often free and readily available, software tools.
- Be realistic about the stability of the technology. Over 90% of instructors report frequent problems (Navarro, 2000). You can't expect to use transparencies as a backup to an interactive lecture with dynamic linking. Especially with the vulnerability of the Internet and servers to virus and worm threats, there will be times when the Internet is down or so slow that interactivity isn't possible. Plan for technical problems and have a contingency plan that is communicated to students.
- Be realistic about the reward system for incorporating technology. If you are at an institution where research is valued more than teaching, then you may need to forego creating Web content. Navarro (2000) also notes that "cyberprofs" are reporting strong negative reactions from their colleagues. If you are sitting in your office answering student e-mail or creating course content, you don't appear to be teaching.
- Make sure intellectual property and royalty procedures are clearly spelled out. Many faculty do not consider the issue of copyright and intellectual property when course materials are developed (Rueter, 2001). Most faculty believe they own their own course materials. This is often not the case. The issue becomes even muddier when entire courses are delivered on the Web (levels 5 and 6). A course that you developed could be offered by the university with someone else teaching it, without your consent or knowledge. Earnings from distance learning are viewed quite differently by faculty and administration (Guernsey & Young, 2001).
- Make sure you and your administration agree what "quality" teaching is. Age-old tenets of quality teaching include meaningful discussion, question and answer discourse, and significant teacher-student interaction. If faculty develop a Web-enhanced course following the myth of preserving student interaction, they will be quickly mired in Web activities that consume the majority of their time with no observable educational payoff. The less subtle problems that stem from a lack of administrative understanding include difficulties in obtaining resources, especially released time for the initial development of a Web class, lack of understanding about how many hours are needed to run the course when you aren't standing in front of a classroom, e-mail overhead, managing list serves, problems in obtaining reasonable hardware and software to develop and maintain a Web course, and the number of students that should be placed in the section.

FUTURE TRENDS

As younger, more computer-literate faculty emerges, there will be a slow and steady move toward the incorporation of Internet components into university courses. Many educational pundits believe that we are moving into a new learning paradigm with the integration of technology into our schools (Von Holzen, 2000). This new educational model envisions a complete shift in course delivery from the traditional lecture classroom to on-demand, flexible learning through the use of telecommunications technology or "just-in-time" learning. In this paradigm, the faculty will become the designers of interactive course materials.

CONCLUSION

With this new paradigm, many of the issues discussed in this chapter may take care of themselves. In the meantime, faculty who are considering Web-enhanced or Web-delivered courses need to be aware of the issues. But, most of all, we believe that learning is more than just content delivery; we need to create learning environments whether they are in the classroom or in cyberspace.*

*This article is based on work originally published by the authors and Terry D. Lundgren.

REFERENCES

Average Person Spends More Time Using Media Than Anything Else. (2005). Ball State University News Center. Retrieved July 10, 2006, from http://www.bsu.edu/news/article/0,1370,7273-850-36658,00.htm

Baggaley, J. (2003). *Blogging as a course management tool*. The Technology Source Archives. Retrieved July 20, 2006 from, http://technologysource.org/article/blogging_as_a_course_management_tool/

Carr, S. (2000). *As distance education comes of age, the challenge is keeping the students*. The Chronicle of Higher Education, 23, A1. Retrieved July 15, 2006, from http://www.chronicle.com/free/v46/i23/23a00101.htm

Castells, M. (2000). *The rise of the network society.* Malden, MA: Blackwell Publishers.

Garrett, N. A. (2006, May, 2006). *Setting up and using collaborative learning communities using RSS technologies.* Paper presented at the Faculty Summer Institute, 2006, University of Illinois at Urbana-Champaign.

Garrett, N. A., Lundgren, T. D., & Nantz, K. S. (2000). Faculty course use of the Internet. *Journal of Computer Information Systems* (Fall, 2000).

Guernsey, L., & Young, J. R. (1998). *Professors and universities Anticipate disputes over the earnings from distance learning*. Chronicle of Higher Education. Retrieved July 21, 2006, from http://www.chronicle.com/colloquy/98/ownership/background.shtml

Harrsch, M. (2003). *RSS: The next killer app for education*. The Technology Source Archives. Retrieved July 20, 2006, from http://technologysource.org/article/rss/

Hitlin, P., & Rainie, L. (2005). *Data memo: Teens, technology, and school* (Memo): Pew Internet and American Life Project.

Horrigan, J. B. (2006). *Home broadband adoption 2006.* Pew Internet and American Life Project.

Horrigan, J., & Rainie, L. (2006). *The Internet's growing role in life's major moments*. The Pew Internet and American Life Project.

Horrigan, J., Rainie, L., & Fox, S. (2001). *Online communities: Networks that nurture long-distance relationships and local ties.* The Pew Internet and American Life Project.

Lenhart, A., Madden, M., & Hitlin, P. (2005). *Teens and technology: Youth are leading the transition to a fully wired and mobile nation.* The Pew Internet and American Life Project.

Lundgren, T. D., & Lundgren, C. (1996). *College student absenteeism.* Paper presented at the 1996 Delta Pi Epsilon National Research Conference, Little Rock, Arkansas.

Madden, M. (2006). *Data memo: Internet penetration and impact.* Pew Internet and American Life Project.

Marchese, T. (1998). Not-so-distant competitors: How new providers are remaking the postsecondary marketplace. *AAHE Bulletin* (May, 1998).

Mesher, D. (1999). Designing interactivities for Internet learning. *Syllabus, 12*(7), 16-120.

Nantz, K., & Lundgren, T. (2003). Student attitudes towards Internet courses: A longitudinal study. *Journal of Computer Information Systems*, Spring.

Nantz, K. S., & Lundgren, T. D. (1998). Lecturing with technology. *College Teaching, 46*(2), 53-56.

Navarro, P. (2000). Economics in the cyberclassroom. *Journal of Economic Perspectives, 14*(2), 119-132.

Pew / Internet. (2006). *Internet activities*. Retrieved July 20, 2006, from http://www.pewinternet.org/trends/Internet_Activities_7.19.06.htm

Rainie, L. (2006). *How the Internet is changing consumer behavior and expectations.* Retrieved July 15, 2006, from http://www.pewinternet.org/PPF/164/presentation-display.asp

Rao, P. V., & Rao, L. M. (1999). Strategies that support instructional technology. *Syllabus, 12*(7), 22-24.

Resmer, M. (1999). IMS: Setting the course for distributed learning. *Syllabus, 12*(7), 10-14.

Rueter, J. (2001). *Modular courses and intellectual property rights*. Retrieved July 21, 2006, from http://Web.pdx.edu/~rueterj/rlw/modular.htm

Rups, P. (1999). Training instructors in new technologies. *T.H.E. Journal, 26*(8), 67-69.

Teles, L. (2002). *The use of Web instructional tools by online instructors*. The Technology Source Archives. Retrieved July 20, 2006, from http://technologysource.org/article/use_of_Web_instructional_tools_by_online_instructors/

Von Holzen, R. (2000). A look at the future of higher education. *Syllabus, 14*(4), 54-57, 65.

Waits, T., & Lewis, L. (2003). *Distance education at degree-granting postsecondary institutions: 2000-2001* (No. NCES 2003-017): National Center for Education Statistics (NCES).

KEY TERMS

HTML (Hypertext Markup Language): This is the foundation protocol for the world wide Web (WWW) that allows text, images, links, and other materials to be combined together into a single presentation.

LMS (Learning Management Systems): Also known as *content management systems*, these systems combine a variety of collaborative features into a single user interface, making it easier to administer and design content (faculty) and access and use (students). There are a number of commercial and open-source systems available with the most well known being WebCT and Blackboard (commercial), and Moodle (open-source).

RSS (Really Simple Syndication or, alternatively, Rich Site Summary (the former is the preferred term and is in wider use): This technology is based upon XML and is designed to facilitate the syndication, aggregation, and consumption of Web-based content.

Web-Based Course: A course, which is delivered entirely by electronic methods such as the Internet.

Web-Enhanced Course: A traditional course with some electronic enhancements, such as Web pages for course syllabi, data files, and test reviews.

XML (Extensible Markup Language): This markup language, which is much more robust than html, is used for numerous different applications but is primarily known as a container for networked database information and the foundation of RSS.

Issues of E-Learning in Third World Countries

Shantha Fernando
University of Moratuwa, Sri Lanka

INTRODUCTION

Around the world, e-learning is becoming popular, especially among higher education institutes (universities). Many highly ranked universities have either already deployed an e-learning system and are fully operational, or they are in a process of deployment where e-learning-based and non e-learning-based educational environments co-exist. It is also possible to find a few virtual universities. The amount of money and effort that has to be spent on e-learning is high. In addition to the initial e-learning system installation costs, there are ongoing maintenance, management and content development costs. Due to the rapid growth in the field of e-learning and the role it plays in today's education systems, those working in the field have begun to introduce standards for different aspects of e-learning. The Open Knowledge Initiative (OKI) which is described as "a collaboration among leading universities and specification and standards organizations to support innovative learning technology in higher education" is an example (OKI, 2003).

Many highly ranked universities use commercial e-learning systems such as BlackBoard, WebCT, e-college, Netschool, etc. Several open source products are available though their usage is not wide spread, although it is expected that collaborative projects such as Sakai will enable large-scale open source products to be introduced to the market. This effort is described on the Sakai website as, "The University of Michigan, Indiana University, MIT, Stanford, the uPortal Consortium, and the Open Knowledge Initiative (OKI) are joining forces to integrate and synchronize their considerable educational software into a modular, pre-integrated collection of open source tools" (OKI, 2003).

BACKGROUND

Many third world countries have become "Transitional Countries". The term "transitional country" has been used in different ways in different times and different contexts. However, today's meaning of a "transitional country" is a country that lies between a developed and a developing country, and has an evolving market economy. Dung (2003) states:

Generally speaking, the expression 'transition' is used, mainly by political scientists, in the context of changes that have followed the fall of regimes, usually when dictatorial regimes have given way to more democratic ones, but this usage has been extended to contexts where previously rigid structures, such as those governing the economy, are giving way to more liberal, market-friendly structures and associated features of liberal democracy.

Third world or transitional countries require sustainable development. Sustainable development of a country is very much dependent on industry, higher education and research, hence university education is vital. The importance of the higher education is stressed in the United Nations Resolution on the Decade of Education For Sustainable Development January 2005 – December 2014 (UN Report, 2002). For a third world country, as De Rebello (2003) puts it, "The university system was seen as being uniquely equipped to lead the way by their special mission in teaching and training the leaders of tomorrow, their experience in transdisciplinary research and by their fundamental nature as engines of knowledge."

CURRENT TRENDS IN INFORMATION TECHNOLOGY IN THIRD WORLD COUNTRIES

IT is becoming a driving force of economy. Realizing its potential, many transitional countries have embarked on projects in collaboration with funding agencies to improve IT services, though their IT infrastructure facilities are not adequate. Many foreign investors start IT based companies in transitional countries. The products are aimed at the US or European market, where the parent companies are based. India, in particular, exemplifies this for the IT sector, and many major IT companies have branches in India. In Sri Lanka, due to the limited market, poor infrastructure and slightly higher labor costs, such foreign investments are limited. However, the level of IT expertise is at a competitive level. Many local IT companies carryout sub-contracts for foreign IT companies. A few companies directly interact with the global market. Realizing the potential, the Sri Lankan government embarked on "e-Sri Lanka move" project to introduce e-governance and to improve e-services within the country, and formed the ICT Agency using World Bank

funds (Development Gateway, 2003). Motivated by these initiatives and realizing the importance of e-learning for today's form of higher education, some Sri Lankan universities have deployed e-learning systems as pilot projects and a few others have started exploring the possibility of using e-learning for their university education.

Due to the employment opportunities offered for IT professionals of transitional countries by developed countries, many professional IT programs have been initiated in transitional countries. In Sri Lanka, income generated by foreign employment has now become considerable compared to its other income sources such as garment, tea, rubber, minerals, spices, etc. Though most employment opportunities are labor-oriented, many professional opportunities are in the IT sector. However, this causes "brain drain".

IMPORTANCE OF E-LEARNING FOR HIGHER EDUCATION IN THIRD WORLD COUNTRIES

In order to understand the importance of e-learning, it is important to consider what we mean by e-learning. According to the definition of NCSA's e-learning group (Wentling, T.L. et al., 2000):

E-learning is the acquisition and use of knowledge distributed and facilitated primarily by electronic means. This form of learning currently depends on networks and computers but will likely evolve into systems consisting of a variety of channels (e.g., wireless, satellite), and technologies (e.g., cellular phones, PDA's) as they are developed and adopted. E-learning can take the form of courses as well as modules and smaller learning objects. E-learning may incorporate synchronous or asynchronous access and may be distributed geographically with varied limits of time.

In an abstract form, I would define it as "electronically facilitated, enhanced and managed learning". It can consist of many components or elements of a learning environment of a university system if they can be electronically facilitated, enhanced and managed. Some aspects that could be integrated into an e-learning system to make an impact in a university system, especially in the context of a third world country, are given below.

- Curriculum related aspects – courses and course contents, discussions, library catalogues, etc.
- Academic administration related aspects – registrations, student information, grading, etc.
- Technology infrastructure related aspects – alternative technologies, lab facilities, home use, etc.
- Societal context related aspects – cultural events, forums, activities, etc.
- Industrial collaboration related aspects – industrial expertise and contents, know-how dissemination, guidance to/from industry, etc.

These aspects, when incorporated in an e-learning system, will improve the quality of the higher education, if implemented using strategies and technologies suitable for constrained environments in third world countries. However, deployment of a suitable e-learning system requires a particular educational, administrative and technological environment, and the university educational system will also need to undergo changes. This is where the issues are faced in third world countries. One should not think that the deployment of e-learning is an adaptation to the required educational change. Contrarily, an ability to adapt is a must for the deployment of e-learning.

Bates (2000) states that higher education institutes consider technology-based learning for the following reasons:

- the need to do more with less
- the changing learning needs of society
- the impact of new technologies on teaching and learning (Bates, 2000, p. 8).

Although we observe that mainly the universities in developed countries tend to consider the above reasons, they are applicable to any university. It is in this context that e-learning is becoming attractive. However, when universities in third world countries embark on e-learning-based educational transformations, they face many barriers. In many cases, e-learning cannot be implemented in the way it is done at US or European universities. The approach has to be tailored to the environment, if it is to be a success.

COMMON ISSUES TO BE ADDRESSED

Administrative Issues

Most of the universities in third world countries are traditional universities. Gunn (2000) in his keynote paper states the following:

Perhaps the most critical challenge to traditional universities is develop capacity to change. This calls for major restructuring, removal of unnecessary processes and streamlined administration procedures. Motivation to progress, change and develop is hard found in the current insecure climate. . . The challenge this raises is being able to exploit the resources of commercial interests while maintaining quality and standards of service as a priority area. Ability to achieve

the right balance between opposing forces of cost and quality without reducing education to the lowest common factor will be a powerful survival strategy.

In many third world countries university academic administration is stream-lined and rigid. Changes are usually not welcomed. Many fear loosing the value of their jobs if IT strategies are introduced. Many administrative officers have the mentality that the others should come to them to get the work done. While this shows an attitude problem or an inferiority complex, it affects many productive plans.

However, rigid administrative procedures are sometimes required to prevent exploitation and use of facilities for personal advantage.

Some administrative functions can be handled efficiently through e-learning. Typical examples would be student semester and exam registrations, yearly progress archiving, student information management, etc. However, administrative officers such as registrars, examination branch officers, etc, are not comfortable when it is handled entirely by the e-learning system. There is the fear they might loose their job. Another fear is whether they will have any value for the university. A valid concern that is raised is whether the e-learning system is secure enough to protect confidential data and prevent students tampering with data.

IT Infrastructure Issues

IT infrastructure facilities in third world countries are often primitive. While IT infrastructure needs improvement for better interconnectivity of academic institutes, a countryman's concern is food, water supply, clothing, roads and transportation, housing, primary schools, and other essential items for their living. Governments in these countries have to allocate the majority of their funds for the latter and a low priority is given for IT infrastructure. It is not justifiable to allocate huge funds for the improvement of IT infrastructure when the basic needs of the people are not met. The good news is that some form of infrastructure is already available. The solution we propose for the improvement of higher education using e-learning has to consider alternative techniques given this serious limitation. This is not to say that mobile communications and other new inventions are not penetrating the market.

Consider Sri Lanka as a case, every university is interconnected by a university network called LEARN (Lanka Academic and Research Network). Some universities have 2 Mbps E1 links, while the rest have only 128 or 64 kbps links. Very soon the latter will be upgraded, but the maximum would be 2 Mbps in the foreseeable future. The current international bandwidth allocated for the whole university network is below 2 Mbps. This will gradually increase on demand, but on-demand increase implies the presence of congestion. The universities also experience disruption of the telecom services, either due to faults or non-payment of bills. However, within these infrastructure constraints, the majority of universities are able to have an acceptable level of communication for the current IT operations within the country. Web servers are acceptably fast and e-mail is heavily used for communication and collaboration among academics. A few e-learning systems are also operational.

Any e-learning-based solution has to work within these IT infrastructure constraints. Within a university it will work acceptably since many universities have local area networks with either gigabit fibre optics, or fast Ethernet or at least 10 Mbps links. Between universities it will work as long as it does not have heavy content delivery, congesting the links. However, international collaborations through e-learning will not be at the levels required by many e-learning systems in the near future.

Limitation of Equipment

In third world countries, equipment such as servers, routers, cabling, laboratory computers, etc. are usually procured under special university budgets, or grants and loans from funding agencies. It is not possible to expect frequent upgrades to equipment. It is very unlikely that high-end servers with redundant power supplies and disk arrays will be always available for the deployment of an e-learning system and redundancy and backup systems are not a priority. Sometimes valuable information stored in the system may be at stake. However alternative approaches such as weekly or critical time-based backups may be carried out.

Thus, any approach to introducing e-learning has to start with a low-end solution. Once the importance is recognized by the authorities, some form of ongoing support is feasible. Strategic planning is required to get the funds for improving the performance and reliability of the systems gradually.

It is not possible to assume that students will always have access to computers. While a few have their own computers, the majority of the students in transitional countries use common lab facilities to access computers. Labs are open only during working hours and usually scheduled for different groups of students based on assignments and workloads. In most cases, e-learning-based learning activities also need to be planned accordingly. For an example, if an assignment is given with a deadline for the submission through the e-learning system, this deadline has to be flexible in situations such as insufficient computers, labs being not open on demand, workers' strikes which are frequent in many third world countries, long electricity power cuts, etc.

Cost Factors of E-Learning Systems

Most of the commercially developed e-learning systems such as BlackBoard, WebCT, etc, used by US and European

universities are extremely expensive for the third world countries to purchase. A monetary grant may be a possibility, but then the question would be maintenance costs, purchase of additional modules to suite the changes as time passes, costs of customizations, etc., if these costs have to be born by the university, which are very high given the limited budgets. Therefore, any grant must include these costs. Otherwise, it will be a waste of funds.

An alternative is to select an open source solution. However, currently it is difficult to find the exact match of an open source solution, or to customize it to a particular university's environment. Projects such as Sakai may help solve this situation in the future, but we have to wait until their collaborative environment is functional. However, there is the concern whether it also will assume state-of-the-art technology infrastructure.

Another alternative is in-house development, however, for this to be a success, continuous employment of developers and good software development approaches with research input from an e-learning perspective is required. Finding developers is not a difficulty in most transitional countries, and the costs for this will be far below the purchase of a commercial e-learning system. To succeed, however, a vision to continue the project, and institutionalized incentives to the people involved, should be in place. While the result of this approach may not be as sophisticated as, or as reliable as, available commercial systems, it is possible to come up with an acceptable solution at a very low cost. In the author's environment, it was possible to get a group of students to start on the development of an e-learning system using research findings. Later, an expert was used to further improve it to be used as a production system. It needs further development, but the advantage is, while the required institutional changes for an e-learning-based education are conveyed to the rest of the faculty, the changes can also be synchronized with the development cycle, as illustrated by Collis and Moonen (2001). Even if a fully fledged e-learning system had been purchased, it would have been a failure due to the faculty being not ready to adapt immediately.

Reliability Issues

Reliability issues have already been mentioned under IT infrastructure issues and limitation of equipment. The following summary is provided to emphasize the issue of reliability.

- Frequent electricity power failures.
- Data communication connectivity failures or disruptions due to non payment of bills.
- Congested links.
- Less emphasis on backup and redundant systems.

Socio-Cultural Issues

In most of the third world countries, especially in South Asian and African continents, socio-cultural setting is very prominent. It affects how people engage in learning activities. Verbal and physical interactions are important and hence total virtual learning environments may not produce good results. This situation may change in the years to come, especially among the urban population. However, socio-cultural aspects cannot be neglected when dealing with education, and it is true also for technology-based education, as described by Gunawardena (1998).

Most of the e-learning systems and available contents are based on popular languages. However, this is not to say that they do not support other languages, but it will require an additional effort to prepare contents in native languages. In many third world countries primary education is done in native languages, although at university level popular languages like English or Spanish may be the medium. This situation can create communication barriers in e-learning-based learning processes.

Many people in third world countries believe that developments in IT will cause many people to lose their jobs. This is a serious social issue. However, there are situations where it is thought to be the other way round. For an example, in e-Sri Lanka move, the government expects that there will be an increase in job opportunities if IT is promoted. For an example, to deploy e-learning in a university environment, additional support staff is required for facilitation, content creation, maintenance, etc.

FUTURE TRENDS AND CONCLUSION

E-learning can play a major role in higher education in third world and transitional countries. It will help improve the higher education, thereby contributing to sustainable development. Using e-learning it is possible to improve curriculum, academic administration, industry collaboration, etc.

Emerging related standards such as Sharable Content Object Reference Model (SCORM, 2003), IEEE Learning Technology Standards Committee (LTSC, 2002) and collaborative work currently being carried out such as OKI (OKI, 2003) will make e-learning more widespread.

However it may not be possible to deploy it in third world countries in the way it is done in the highly ranked universities in the US and European countries. First, the related issues have to be addressed and alternative solutions should be explored. Given suitable alternative solutions, or desirable approaches, e-learning can be a success in many third world and transitional countries.

REFERENCES

Bates, A. W. (2000). *Managing technological change: Strategies for college and university leaders*. San Francisco: Jossey-Bass Publishers.

Collis, B., & Moonen, J. (2001). *Flexible learning in a digital world: Experiences and expectations*. UK: Kogan Page.

De Rebello, D. (2003). What is the role for higher education institutions in the UN decade of education for sustainable development?, *Theme IV, International Conference on Education for a Sustainable Future* (pp. 10-11). Prague, Czech Republic: Charles University, Karolinum.

Development Gateway. (2004). *e-Sri Lanka: Transforming government, business and society* (December 29, 2003). Retrieved March 01, 2004, from *http://www.developmentgateway.com/node/133831/sdm/docview?docid=841120*

Dung, L. T. (2003). Judicial independence in transitional countries, *The Democratic Governance Fellowship Program, United Nations Development Program, Oslo Governance Centre, January 2003* (page 5). [Electronic version]retrieved March 02, 2004, from *http://www.undp.org/oslocentre/docsjuly03/DungTienLuu-v2.pdf*

Gunawardena, C. (1998). Designing collaborative learning environments mediated by computer conferencing: Issues and challenges in the Asian socio-cultural context. *Indian Journal of Open Learning*, *7*(1), 101-119.

Gunn, C. (2000, Decmber). *Identity, control and changing reality*. Keynote paper at ASCILTE Conference, Coffs Harbour. [Electronic version] retrieved June 25, 2003, from *http://www.ascilite.org.au/conferences/coffs00/papers/cathy_gunn_keynote.pdf*

LTSC. (2002). *Learning object metadata*. Learning Object Metadata Working Group, Learning Technology Standards Committee (LTSC), IEEE. Retrieved March 24, 2004, from *http://ltsc.ieee.org/wg12/index.html*

OKI Project. (2003). *Open knowledge initiative*. Retrieved March 02, 2004, from OKI Project web site *http://web.mit.edu/oki/*

Sakai Project. (2003). Retrieved March 02, 2004, from Sakai Project Web site *http://www.sakaiproject.org/*

SCORM. (2003). SCORM overview. Advanced Distributed Learning. Retrieved May 05, 2004, from *http://www.adlnet.org/index.cfm?fuseaction=scormabt*

UN Report. (2002). *World summit on sustainable development: Plan of implementation* (para 117d). Retrieved March 01, 2004, from *http://www.johannesburg summit.org/html/documents/summit_docs/2309_planfinal.htm*

Wentling, T.L., Waight, C., Gallaher, J., La Fleur, J., Wang, C., & Kanfer, A. (2000, September), *E-learning – A review of literature*. Knowledge and Learning Systems Group, University of Illinois at Urbana-Champaign, NCSA.

KEY TERMS

Academic Administration: Administration procedures or formalities linked with university education, such as registrations for semesters or examinations, progress reviews and monitoring, eligibility formalities, student history records or progress archiving, promotions to levels or years, academic timetables, etc.

E-learning: Electronically facilitated, enhanced and managed learning.

IT Infrastructure: Technological infrastructure that enables the transfer of information.

Learning Environment: Overall university setting in which many educational and administrative processes interact.

Open Source E-learning Systems: E-learning systems developed by the Open Source Community and freely distributed with their own license or a GPL (General Purpose License) to use, modify and distribute together with the source code.

Third World Countries: Countries that are not yet developed.

Transitional Countries: A third world country that is in a transition process based on more liberal, market-friendly structures and associated features of liberal democracy.

Virtual Universities: All the learning and administration activities are done through e-learning and very minimum physical interactions, or no physical interactions at all.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1702-1707,

IT Application Development with Web Services

Christos Makris
University of Patras, Greece

Yannis Panagis
University of Patras, Greece

Evangelos Sakkopoulos
University of Patras, Greece

Athanasios Tsakalidis
University of Patras, Greece

INTRODUCTION

The advent of Web Services (WS) has signaled a true revolution in the way service-oriented computing and remote procedure invocation over the Web are conducted. Web Services comprise of a set of loosely coupled specifications to coordinate process execution from distance, based on common and widely accepted Web protocols such as HTTP, FTP, and XML, and therefore, providing increased development flexibility. Since the WS Framework was built on top of those protocols, Web Services have been widely acclaimed by the Web development community and paradoxically; they have marked one of the few examples in the history of computer protocols where a global consensus has been reached.

The Web Service framework consists of essentially three basic components:

1. The *Web Service Description Language* (WSDL), a language that allows formal functional characterization of the provided functionalities;
2. The *Simple Object Access Protocol* (simply SOAP from its version 1.2), a protocol that defines the format of the information interchange; and
3. The *UDDI* (Universal Description, Discovery and Integration) is a catalog of Web Service descriptions.

All three of these components are specified using XML markup. The elegance of the WS architecture lies in the fact that every WS transaction is taking place over established Web protocols such as HTTP and FTP. As remarked in Ballinger (2003, p. 5): "A Web Service is an application logic that is accessible using Internet standards." This very fact has accounted for the rapid and universal adoption of Web Services.

This work is organized as follows: First, a review of underlying technologies and tools is presented. Consequently, existing techniques for design methodologies are described. Next, an overview of storage and retrieval techniques for Web Services is given followed by real-world applications of Web Services. We conclude with open issues and discussion.

BACKGROUND

The need for executing process from remote computers seems to have emerged right after the first networking efforts. To put it simply, people need to share their data or access other peoples' data over the Internet, in the easiest possible way (Ballinger, 2003, p. 2). These needs are formally described under the term *Service Oriented Architecture* (SOA). Typically, SOAs are distributed system architectures focusing on network-centric, message-based and platform-independent communication (World-Wide Web Consortium [W3C], 2004a, 3.1).

Web Services constitute a brilliant example of an SOA implementation, albeit not the only one. Some of its precursors include: UNIX RPC, Microsoft's COM/DCOM, CORBA, and Java RMI. All of the latter have failed however due to their complex architecture.

REVIEW OF TECHNOLOGIES AND TOOLS

This section deals with the state-of-the-art in the technologies supporting the development of Web Services.

The Web Service Framework

The base protocols for the Web Services architecture are HTTP, XML, SOAP, WSDL, and UDDI. The role of HTTP and XML is more or less self-explanatory; they provide the wrapper protocols for every kind of data communication. In

the sequel, we shortly describe the remaining protocols.

- *SOAP (W3C, 2003) defines the format of message interchange. This interchange takes place when* discovering, binding, consuming a Web Service.
- WSDL (W3C, 2004c), is a language to describe the functionalities of a Web Service and provide additional details about the ways it can be accessed, the points it can be reached, and so forth.
- UDDI (UDDI) defines a separate entity (registry) that mediates in the development process by hosting descriptions of Web Services.

First of all, SOAP (W3C, 2003) is a protocol that requires the creation of XML-like documents defining the format of every communication taking place during a Web Service deployment. SOAP messages are created each time a computer seeks or runs a WS and each time it sends messages, that is, query results, to a remote computer. A typical SOAP message contains data format, required WSDL messages for the procedure and, if it calls a remote procedure, the WSDL-coded names of functions that will be called.

WSDL (W3C, 2004c) is another WS standard built around XML specifications. While SOAP defines the format of messages that will be exchanged, WSDL is the language that totally describes Web Services. First of all, a WSDL document contains descriptions of data types, XSD or custom, used in the procedure invocation. The messages that must be exchanged used during the execution, that is conveying I/O information, are also defined. Names of available remote procedures are also registered. Furthermore, apart from the names, *bindings* are also defined; bindings refer to the message protocol (most commonly, SOAP and HTTP) in use and that is, to a URL that provides access to the service. An important feature of WSDL is the possibility to insert in arbitrary elements inside a WSDL document, meta-information to be utilized for documentation purposes.

Once having joined the Web Service game, one needs to find out which function to call to carry out a specific task, by which parameter, which protocol to use, and so forth. It was agreed during the early days of WS architecture to use centralized registries for this purpose. These registries would gather all the information about available Web Services they host and they would provide technical details, WSDL descriptions, on how to use them. These registries were termed UDDIs. The globally available UDDIs are few, including Microsoft's[1] and IBM's[2], however UDDIs also exist inside large corporations. The UDDI protocol requires the registry to provide the corresponding APIs for service registration and querying.

Corporate records inside a UDDI registry are implemented via *businessEntities.* Each party that wishes to publish a set of Web Services registers a new *businessEntity.* A *businessEntity* area contains business related information about the entity, which publishes the Web Services, such as contact information, e-mail, business categorization, and textual descriptions. A further sub-entity inside each *businessEntity* is the *businessService.* Each *businessService* contains a record for a conceptually or otherwise related subset of the provided Web Services. Web Services in the same *businessService* can be, for example, geographically related or performing different functionalities of the same category, e.g. functions

Figure 1. The hierarchical structure in UDDI documents

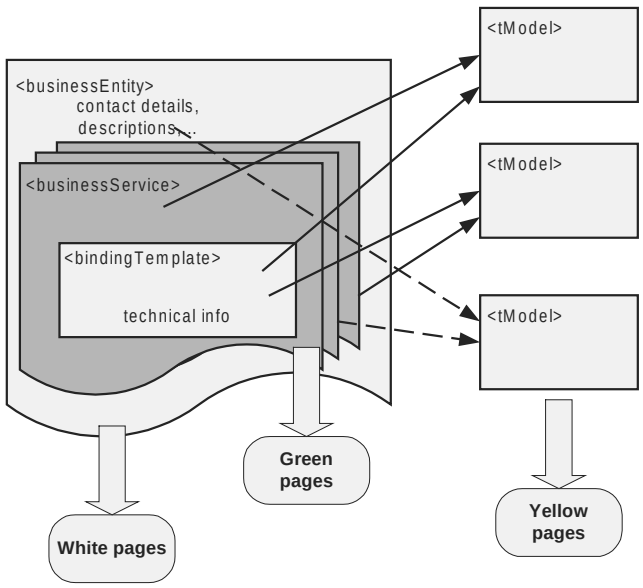
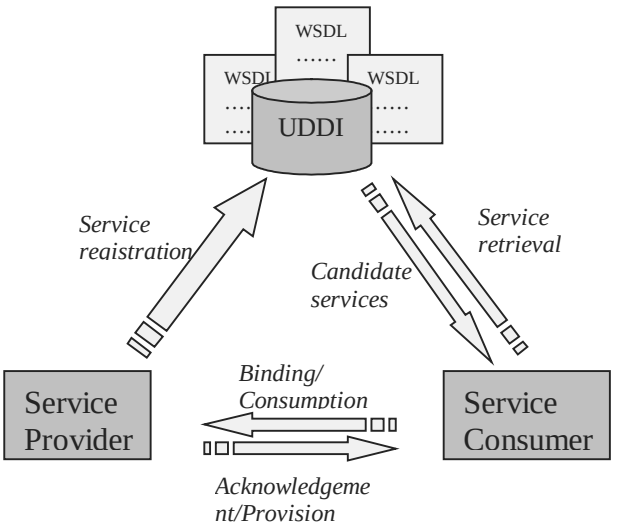


Figure 2. The Web Services Model (Arrows indicate communication with SOAP messages)



querying available rooms or booking a hotel room. Textual descriptions can also be part of the *businessService* structures. Each service in such a structure can contain one or more technical sections or *bindingTemplates.* A *businessService* may contain information on the service access point, but this information is not enough for the service consumption. The gap is filled by the *bindingTemplate*, which provides the required technical details, the service fingerprint, for service consumption.

Often Web Service fundamental functionalities are standardized. Many *bindingTemplates* provide pointers to external taxonomy elements that, in turn, specify the abstract functionalities of certain kinds of Web Services. Such an element is called a *tModel* and it is generally not part of the UDDI itself. A *tModel*, also provides URL references to the source of taxonomy and also textual descriptions.

According to this structuring of the UDDI entities, the contained structures can be informally characterized as follows (see also Figure 1):

- "white pages": information contained in *businessEntity*.
- "green pages": information required for the successful invocation of the Web Service, found in *businessService* and *bindingTemplate*.
- "yellow pages": general categorization of the provided functionality and conformance to prespecified industrial or commercial standards. This information is recorded in *tModel*.

Web Service Release and Consumption

Release and consumption is a three-step process:

1. The *provider* advertises its services in WSDL and registers in an available UDDI node.
2. The *consumer* searches the UDDI catalog, finds the required services, and retrieves the deployment details.
3. The consumer requests the service from the provider, and the provider responds with data.

All the necessary communication steps are conducted with SOAP messages. This process is best illustrated in Figure 2.

Development and Operation Platforms

One reason for the success of Web Services is the adoption of all leading providers of application development platforms with Java and .NET being the major delegates.

The .NET platform is supplemented by other products and services like Visual Studio .NET/.NET Framework for the development of applications, the .NET Passport service and .NET My Services. The .NET Framework is the programming model of the .NET platform for the development and operation of XML Web Services and applications. Visual Studio .NET is based on the previous editions of Visual Studio, so that it can offer flexibility and facilitate the development of integrated services. Finally, ASP.NET is a collection of objects

and classes, which allow the development of an application providing facile application development, since all the settings are kept in XML files while the components and their changes are automatically updated by the system.

Another possibility for WS development is to use Java-based APIs. The UDDI specifications do not directly define a Java-based API for accessing a UDDI registry. The Programmer's API specification only defines a series of SOAP messages that a UDDI registry can accept. Thus, a Java developer who wishes to access a UDDI registry can do so in a number of ways:

1. Using Java-based SOAP API;
2. Using a custom Java-based UDDI client API; or
3. Using JAXR.

Business Workflows

The demand expressing and orchestrating business workflows with the aid of Web Services was recognized early by the WS community. Therefore, languages to control business processes have been developed including *Business Process Execution Language for WS* (BPEL4WS) (Curbera et al., 2002) and *WS-Business Process Execution Language* version 2.0 (WS-BPEL)(OASIS, 2004). A very important recent development in the choreography field is the announcement of the *Web Services Choreography Description Language* Version 1.0 (W3C, 2004b). This initiative aims to serve as an additional layer on top of the corporate workflows in order to orchestrate their interactions.

Drawbacks of the Core Specification

The WS Architecture is not free of drawbacks and limitations. We briefly outline the most significant of them without going into much detail since this goes beyond the scope of this article.

- **Not flexible WS Discovery:** Discovery is based on matching keywords against functional descriptions. This will not adequately cover most of the service invocations needs
- **Quality of Service (QoS):** No method to deliver WSs under certain quality considerations is provided by the core specification.
- **Centralized Registries:** UDDIs may suffer from single point failure and bottlenecks.

RETRIEVAL OF WEB SERVICES

The UDDI data structures do not allow for advanced retrieval but only for strict matching of prespecified technical descriptions. We shortly present in this section some recent efforts to provide WS Information Retrieval.

Information Retrieval Based Methods

The simplest method is the *Keyword Based*. This model is followed by the legacy UDDI Standard and the discovery mechanism it supports. The retrieval corpus is based on textual descriptions found inside UDDI tModels. Queries are formed by keywords and are subsequently matched against keyword descriptions. This approach is similar to the Boolean IR model (Baeza-Yates & Ribeiro-Neto, 1999, Chap. 2).

A more elaborate approach (Sajjanhar, Hou, & Zhang, 2004) is to model WS descriptions as term vectors spanning a corpus vector space. WS vectors are collectively represented in a term-document matrix. Latent Semantic Analysis is used in the sequel to reduce the dimensionality of the vector space and map query vectors more closely to WS-vectors.

Distributed Retrieval

In the system of Schmidt and Parashar (2004), a unique ID was generated for each Web Service through the Hilbert curve mapping. The IDs are then stored in a Chord (Stoica, Morris, Karger, Kaashoek, & Balakrishnan, 2003) of Web-Service-storing peers. Thus, the storage and retrieval of WS inherits the load balancing capability and the dynamic nature of Chord.

A peer-to-peer solution (P2P) is also proposed in Li, Zou, Wu, and Ma, (2004). Their approach is to aggregate WS descriptions, hash them, and distribute them over a P2P overlay. Agent based solutions include Montebello and Abella (2003). This approach aims to describe an environment called DASD (DAML Agents for Service Discovery) where WS requesters and providers can discover each other with the intermediary action of a Matchmaking service.

On the other hand, a novel approach adopting main-memory dynamic interpolation searching to P2P overlays is NIPPERS (Makris, Sakkopoulos, Sioutas, Triantafillou, Tsakalidis, & Vassiliadis, 2005). NIPPERS is proved to display improved performance in searching and managing WS in overlay networks and performs asymptotically better than the popular current DHT-based overlay network, Chord.

APPLICATIONS

The WS development paradigm has been adopted in quite a few application domains, inside and outside IT. In the next paragraphs, we mention some representative examples.

IT-Related Applications

One of the standard WS applications would be incorporation to a help desk application. In this sort of application, the operator needs to retrieve and modify customer data from the central (and remote) company database. Even small Web-based tasks inside a corporation can be built with WSs; logging-in and out of an intranet, Web-based registrations are two more examples.

E-business is another field where WSs have met wide acceptance. Garofalakis, Sakkopoulos, Sirmakessis, and Tsakalidis (2003) give an insight on methods to use Web Services for a hypermedia environment offering various e-services for both customers and company employees in the context of a telecommunication carrier. Medjahed, Rezgui, Bouguettaya, and Ouzzani (2003), have built an integrated e-government system for welfare services, which is based on Web Services. Their system uses the WS framework not only for simple Web-based tasks but also to allow composition of government-to-citizen provisions. Notably, the WS framework seems to be ideal for implementing e-government applications, since it enables e-government applications to outsource from existing e-government services and handles privacy issues uniformly (Medjahed et al., 2003). Sakkopoulos, Kanellopoulos, and Tsakalidis (in press) build a Web-based vocational monograph system. Their work leverages the WS framework, which is enriched with semantics of a specialized ontology and with Web mining techniques.

Other Applications

Hu (2002) presents a methodology to combine Web services with an ontology in order to unify the underlying data subcomponents of an industrial control system. Liopa-Tsakalidis, Sakkopoulos, Savvas, Sideridis, and Tzimas (2005) have described HydroWeb, a system to facilitate hydroponic cultivation via Web-based training, communication with experts and online monitoring. Web Services were chosen as a platform to unify communication and application development from various sources of information.

FUTURE TRENDS

Web Services have been an active research area during the last years. Nevertheless, several steps have to be taken to elaborate on certain aspects of Web Service framework. We see as important further steps the provision of: semantic WS matching, efficient methods to deliver Quality of WS (QoWS), online QoWS-aware selections, and effective methods for approximate WS matching. Furthermore, it seems quite intriguing to use data mining and clustering techniques to improve retrieval of Web Service data. All the previously mentioned directions are especially challenging in the presence of orchestrated Web Services and Web Service-based workflows.

CONCLUSION

This article has only scratched the surface of a vast emerging architecture, the *Web Services framework*. We have presented up-to-date descriptions of the basic underlying concepts, the WS proponents, and we have given an essence of the potential applications of WS. Due to their simplicity, their flexibility, and the platform independency, Web Services have become the leading development paradigm in the most major IT projects. For the purposes of this article, we have presented some of these applications ranging from e-business, e-government to industrial applications. As the WS penetration becomes larger, some of the weaknesses of the initial specifications have been taken into serious consideration, including flexible service discovery (Garofalakis, Panagis, Sakkopoulos, & Tsakalidis, 2004), security, assurance of QoS characteristics, workflow support, orchestration, and so forth. Still, there is a large margin for improvement in QoS support and efficient discovery. All in all, everything points out that we are in front of a software development methodology that will mark the years to come.

REFERENCES

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: Addison-Wesley.

Ballinger, K. (2003). *.NET Web services architecture and implementation*. Boston: Addison-Wesley.

Curbera, F. et al. (Eds.). (2002). *Business process execution language for Web Services (BPEL4WS), Version 1.0*. BEA, IBM, Microsoft. Retrieved January 5, 2005, from http://www-106.ibm.com/developerworks/webservices/library/wsbpel/

Garofalakis, J., Panagis, Y., Sakkopoulos, E., & Tsakalidis, A. (2004). *Web Service discovery mechanisms: Looking for a needle in a haystack?* [Electronic version]. International Workshop on Web Engineering, "Hypermedia Development & Web Engineering Principles and Techniques: Put them in use." In conjunction with ACM Hypertext 2004, Santa Cruz, CA. Retrieved from http://www.ht04.org/workshops/WebEngineering/HT04WE_Garofalakis.pdf

Garofalakis, J., Sakkopoulos, E., Sirmakessis, S., & Tsakalidis, A. (2003). Integrating adaptive techniques with web-services. *Proceedings of the 2003 IEEE International Conference on Information Technology: Coding & Com-*

puting (IEEE ITCC 2003) (pp. 415-419). Las Vegas: IEEE Computer Society.

Hu, Z. (2002). Using ontology to bind web services to the data model of automation systems. In A.B. Chaudhri, M. Jeckle, Rahm, E., & U. Unland (Eds.), *Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web, Web-Services, and Database Systems, Erfurt, Germany, Lecture Notes in Computer Science* (Vol. 2593, pp. 154-168). Springer Verlag.

Li, Y., Zou, F., Wu, Z., & Ma, F. (2004). PWSD: A scalable Web service discovery architecturebased on peer-to-peer overlay network. In J. Xu Yu, X. Lin, X. Lu., & Y. Zhang (Eds.), *Proceedings of the Asian-Pacific Web Conference 2004, Lecture Notes in Computer Science* (Vol. 3007, pp. 291-300). Springer Verlag.

Liopa-Tsakalidis, A., Sakkopoulos, E., Savvas, D., Sideridis, A.B. & Tzimas, J. (2005). HydroNet: An Intelligent Hydroponics Web Service Environment. *Journal of Neural Parallel and Scientific Computations, 13*, 15-36.

Makris, Ch., Sakkopoulos, E., Sioutas, S., Triantafillou, P., Tsakalidis, A., & Vassiliadis, B. (2005). NIPPERS: Network of Interpolated PeERS for Web service discovery. *Proceedings of the 2005 IEEE International Conference on Information Technology: Coding & Computing (IEEE ITCC 2005)*. Las Vegas: IEEE Computer Society Press.

Medjahed, B., Rezgui, A., Bouguettaya, A., & Ouzzani, M. (2003). Infrastructure for e-government Web services. *IEEE Internet Computing, 7*(1), 58-65.

Montebello, M.C., & Abela, C. (2003). DAML enabled Web services and agents in the semantic Web. In A.B. Chaudhri, M. Jeckle, E. Rahm, & U. Unland (Eds.). *Revised Papers from the NODe 2002 Web and Database-Related Workshops on Web, Web-Services and Database Systems, Lecture Notes in Computer Science* (Vol. 2593, pp. 46-58). Springer Verlag.

OASIS (2004). *Web Services Business Process Execution Language Version 2.0.* Arkin, A. et al. (Eds.), OASIS Open Retrieved January 5, 2005, from OASIS Web site http://www.oasis-open.org/committees/download.php/10347wsbpel-specification-draft-120204.htm

Sajjanhar, A., Hou, J., & Zhang, Y. (2004). Algorithm for Web Services Matching. In J. Xu Yu, X. Lin, X. Lu, & Y. Zhang (Eds.), *Proceedings of the Asian-Pacific Web Conference 2004, Lecture Notes in Computer Science* (Vol. 3007, pp. 665-670). Springer Verlag.

Sakkopoulos, E., Kanellopoulos, D., & Tsakalidis, A. (in press). Semantic mining and Web service discovery techniques for media resources management. *International Journal of Metadata, Semantics and Ontologies*. Inderscience Publishers.

Schmidt, C., & Parashar, M. A. (2004). Peer-to-peer approach to Web service discovery. *World Wide Web: Internet and Web Information Systems, 7*, 211-229.

Stoica, I., Morris, R., Karger, D., Kaashoek M.F., & Balakrishnan, H. (2003). Chord: A scalable peer-to-peer lookup protocol for Internet applications. *IEEE/ACM Trans. Netw., 1*(11), 17-32.

W3C (2003). *SOAP Version 1.2 Part 1: Messaging Framework* (W3C). In M. Gudgin, M. Hadley, N. Mendelsohn, J. J. Moreau, & H. Frystyk Nielsen (Eds.). Retrieved May 3, 2005, from W3C Web site http://www.w3.org/TR/2003/REC-soap12-part1-20030624/

W3C (2004a). *Web services architecture* (W3C). In D. Booth et al. (Eds.). Retrieved May 3, 2005, from http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/

W3C (2004b). *Web Services Choreography Description Language Version 1.0*. (W3C). In N. Kavantzas, D. Burdett, G. Ritzinger, T. Fletcher, & Y. Lafon(Eds.). Retrieved May 5, 2005, from http://www.w3.org/TR/ws-cdl-10/

W3C (2004c). *Web Services Description Language (WSDL) Version 2.0 Part 1: Core Language* (W3C). In R. Chinnici, M. Gudgin, J.J. Moreau, J. Schlimmer, & S. Weerawarana (Eds.). Retrieved May 3, 2005, from http://www.w3.org/TR/wsdl20

KEY TERMS

Distributed Application Development: A programming paradigm focusing on designing distributed, open, transparent, scalable, fault tolerant systems. This paradigm is a natural result of computer internetworking.

Information Retrieval: Information Retrieval (IR) deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested.

Information Technology: Information technology (IT) or information and communication technology (ICT) is the technology required for information processing. In particular the use of electronic computers and computer software to convert, store, protect, process, transmit, and retrieve information from anywhere, anytime.

Web Engineering: A branch of software engineering, addressing the specific issues relating to design and development of large-scale Web applications. It focuses on the methodologies, techniques, and tools that are the foundation

of complex Web application development and which support their design, development, evolution, and evaluation.

Web Services: A family of standards promoted by the W3C for working with other businesses, developers, and programs, through open protocols, languages, and APIs, including XML, SOAP, WSDL, and UDDI.

World Wide Web: Computer network consisting of a collection of Internet sites that offer text and sound and animation resources through the hypertext transfer protocol.

XML: Short for Extensible Markup Language, a specification developed by the W3C. XML is a pared-down version of SGML, designed especially for Web documents. It allows designers to create their own customized tags, enabling the definition, transmission, validation, and interpretation of data between applications and between organizations.

ENDNOTES

1 http://uddi.microsoft.com
2 http://www.ibm.com/services/uddi

IT Evaluation Practices in Electronic Customer Relationship Management (eCRM)

Chad Lin
Curtin University of Technology, Australia

INTRODUCTION

Organizations are becoming increasingly aware of the need to scrutinize their bottom-line financial returns of business automation initiatives. To achieve this, organizations have to become more customer-centric. According to Karakostas, Karadaras and Papathanassiou (2005), a 5% increase in customer retention can result in an 18% reduction in operating costs. Therefore, the need to build and maintain customer relationship has become a priority for organizations. However, according to a KPMG survey, only a small percentage of companies were able to obtain even basic customer information despite the fact that 89% of companies consider customer information to be extremely important to the success of their business (McKeen and Smith, 2003). As a result, many organizations are adopting electronic customer relationship management (eCRM) applications in order to gather, organize, understand, anticipate, and respond to the constant evolution of customers' requirements and demands.

Indeed, eCRM is forecasted to become increasingly important as businesses seek to deliver their services and information as well as to provide transactional facilities via online and wireless platforms, in additional to the more traditional means of communication channels (e.g., call centers and customer service) (Tan, Yen and Fang, 2002). The market worldwide for eCRM applications is predicted to grow from US $3.4 billion in 2000 to US $10.5 billion in 2005 (EPS, 2001).

Yet, despite the huge investment and widespread agreement that eCRM has direct and indirect impact on customer satisfaction, loyalty, sales, and profit, it has been found that 70% of eCRM solutions that have been implemented by businesses fail (Feinberg, Kadam, Hokama and Kim, 2002). Moreover, studies carried out by Gartner, Forrester, AMR Research, and the Yankee Group claim that most of CRM implementations did not return the expected ROI (Foley, 2002). This is because management tends to be myopic when considering their IT (information technology) decisions, primarily because they are unable to evaluate (specifically the indirect benefits and costs) eCRM applications (Ernst and Young, 1999).

To address this issue, this paper sets out to investigate the current evaluation practices by Australian organizations implementing eCRM. The other objective is to identify the key issues faced by managers to justify and measure their eCRM. Hopefully, the finding can help business organizations to better manage their eCRM investment and its contribution to improving their long term profitability.

BACKGROUND

eCRM Characteristics and Elements

Advances in IT have provided businesses with an opportunity to deliver CRM functions more effectively. The use of IT to deliver CRM has lead to the emergence of electronic customer relationship management (eCRM) and specialist software vendors in the marketplace. This new generation of customer relationship management products is called eCRM because it supports the multiple electronic channels that are now available to customers (Bernett and Kuhn, 2002). The "e" is usually dropped when speaking about eCRM when it refers to CRM that has technology-facilitated interfaces with customers in a broad electronic commerce context which goes beyond the web (Chen and Chen, 2004). The followings are some of the definitions of eCRM found in the literature (Table 1).

Many researchers consider eCRM to be a subset of CRM, meaning that eCRM is one more channel through which an organization can deploy its customer relationship management strategy. eCRM differs from CRM is three important ways (EPS, 2001):

- It includes email, wireless channels, and Web;
- It is enterprise-ready rather than focused on departments or call centers; and
- It extends to cover partner channels such as extranets.

eCRM falls into three main types: operational, analytical, and collaborative (Fjermestad and Romano, 2003). Operational eCRM is concerned with the customer touch points such as automating sales force while the analytical eCRM utilizes technology to process and analyze large amounts of customer data (Sigala, 2004). Collaborative eCRM, on the other hand, focuses on creating a real-time eCRM infra-

Table 1. Various definitions of eCRM

Citations	eCRM Definitions
Steinmueller (2002)	eCRM is the collection of techniques that is employed, or that might be employed, to capture, retain, analyze, and productively utilize information about customers (or potential customers) for the purposes of pre-sales support, making sales and arranging delivery, and providing post-sales support.
Fjermestad and Romano (2003)	It is a combination of hardware, software, processes, applications, and management commitment.
Karakostas et al. (2005)	eCRM means that the sources of customer-related data are collected from the customer interactions with the Web and Internet-based systems.

structure for enterprise sales, service, marketing, and product development to better support customer requirements.

Evaluation of eCRM

As mentioned earlier, organizations invested substantial financial and organizational resources in eCRM annually, but had encountered extremely high failure rates, unhappy customers, and wasted money. While most eCRM vendors promised lots of benefits and dramatic return on investment results, it is difficult to substantiate their claims without undergoing proper evaluation and benefits realization processes (Lin, Pervan, McDermid, 2005; Love, Irani, Standing, Lin, and Burn, 2005). For example, a research conducted by Capgemini indicated that 52% of organizations surveyed could not measure their eCRM investments (Capgemini, 2004). Although there is now a well established field of research concerned with IT evaluation, there has been limited academic and practice based work undertaken on in the domain of eCRM applications (Kim, Suh and Hwang, 2003). The difficulties associated with determining the benefits and costs of IT are the major constraint to investment justification for eCRM. Consequently, many service organizations are faced with a dilemma, that is, how to manage the performance of an enterprise system that has both an internal and external focus and, thus, adds value for stakeholders (Dibb, 2001).

The difficulty in evaluation centers on the fact that both benefits and costs are difficult to quantify (Sugumaran and Arogyaswamy, 2004). In particular, the less precisely bounded environment of electronic commerce technology such as eCRM adds more complexity to the measurement problem as this type of investment is physically distributed between suppliers and customers, making the evaluation process even more difficult (Straub, Hoffman, Weber and Steinfield 2002). Indeed, many organizations have found that these IT project costs and benefits can be difficult to estimate and control. Some new and old measures need to be differentially applied for evaluating phenomena such as electronic commerce and the Internet (Straub, et al., 2002).

RESEARCH METHODOLOGY

As mentioned earlier, the benefits and added value that can be obtained from implementing eCRM have not materialized for many businesses (Ernst and Young, 1999). This is because there is currently a lack of clearly defined and measurable benefits as well as systematic approach to evaluate the eCRM systems (Auer and Petrovic, 2003). Effective evaluation of IT investments is critical to its successful implementation (Lin and Pervan, 2003; Tsao, Lin and Lin, 2004; Ward and Daniel, 2006). To address this issue, specific objectives of the research are to:

1. identify the key factors and issues faced by organizations to justify the implementation of eCRM projects; and
2. determine the current evaluation practices by Australian organizations implementing eCRM projects.

Case study utilizing semi-structured interviews (tape-recorded), observation, and document review were employed for this research, since the need for using multiple sources of data arises from the ethical need to increase the reliability and validity of the research processes (Mingers, 2001). According to Remenyi and Williams (1996), case study is one of the most frequently used research methods in information systems research.

A series of exploratory in-depth formal and informal interviews were conducted in Australia with senior managers and key personnel from several organizations to gain an overview of the business processes and the evaluation practices of their eCRM investments. Interviews were

carried out within 16 organizations in Australia that were involved in eCRM projects. The industries represented in the following cases: hospitality industry (8 organizations), education (1 organization), service industry (2 organizations), IT/Computer industry (2 organizations), and housing industry (3 organizations). Some of these organizations' customers were also contacted. At least two interviews were conducted for each organization. More than 35 interviews were conducted.

In addition to the use of the semi-structured interviews and observation data collection techniques, the researcher examined more than 1000 pages of relevant documents (e.g., annual reports, project reports) that were collected from the participating organizations. These documents provided some useful means of corroborating data from the other sources (e.g., observation and interview data) and expanded on details in order to eliminate or minimize the weakness of human memory when dealing with history.

DISCUSSION

A number of interesting and important issues have come from the analysis of the data gathered and some key issues are presented below in some detail.

1. Lack of proper assessment of business needs - Pre-project planning and justification processes were not properly carried out to assess the needs and feasibility of the eCRM system. These systems were implemented largely based on the "gut feeling" or intuition of the senior executives, the adoption of similar systems by their competitors, or persuasion by the eCRM vendors. Almost all organizations interviewed admitted that no proper pre-project planning, assessment, and justification was carried out before the implementation of eCRM. They were mostly done either through intuition or they believed in the eCRM vendors' words. For example, one project manager who was responsible for implementing an eCRM project said: *"This is the company we trust. This is the company that has just implemented a huge project for us ... this company has operated in couple of other companies in Australia."* This is consistent with research findings where the difficulties and uncertainties associated with IT investment evaluation forced senior executives to rely on gut feeling or intuition when making IT investment decisions (Lin and Pervan, 2003).
2. Different industry requires different implementation and use of eCRM systems - The extent to which the eCRM system was used was largely depending on the type of the industry (Rigby, Reicheld, and Schefter, 2002). Chen and Chen (2004) identified industries such as retail management, office supplies and equipment, hospitality, computer hardware/software, and entertainment as some of the industries that were mostly likely to employ eCRM. From the interview transcripts and other data collected, those industries which provided mainly face-to-face (personalized) services and where the employees were in competition for commissions (e.g., housing industry) tended to use either a standalone eCRM system or less sophisticated version of eCRM system. For example, it was difficult to have a centralized eCRM system within the housing industry as all sales representatives and consultants were in competition with each other in attracting new customers and retaining existing customers. When asked about the eCRM usage within his organization, a real estate agency managing director said: *"I am the only sales representative who is using it. The other sales representative use their own systems and they have to use other means of keeping in contact with their customers."* This is consistent with finding by Melville, Kraemer and Gurbaxani (2004) in which industry characteristics moderate the ability of firms to apply IT (such as eCRM) for improved organizational performance and to capture the resulting benefits. There was simply little incentive for them to share their customers' information via eCRM.
3. Lack of formal IT investment evaluation methodology – According to the interview data, less than one-third of the organizations interviewed had evaluation process. Only five out of 16 organizations interviewed had carried out some sort of evaluation processes (i.e., Scorecard, KPI analysis, benefits/costs and quantitative analysis). The rest were simply relied on their senior management's impressions or gut feeling/intuition. When asked about the evaluation process, one participant said: *"I have said to myself how much time it takes and what is the efficiency? If it can give me nil gain or plus gain that's good. If it gives me negative gain then I am not interested."* Most organizations indicated that they did not have the capability and resources to do so, or they did not know they had no evaluation process. One project manager even did not know about the evaluation process and suggested the executive director might be responsible for doing the evaluation. While almost all of them thought it would be worthwhile to do it, most of them simply did not do it, or relied on their intuition. This is consistent with finding by Karakostas, et al., (2005) where most of the respondents did not have a universal acceptance of metrics and failed to evaluate the performance of their eCRM.

Some of the other issues arising from this interpretive analysis are listed below but are not discussed due to space limitations. Details are available from the author.

4. Lack of user involvement.
5. Lack of benefits realization process.
6. Lack of integration with other systems.
7. Difficulties in identifying indirect costs (or intangible costs).
8. A gap in theory and practice in risk assessment by most organizations.
9. Lack of obvious linkage between the expected outcomes of the eCRM implementation and organizational objectives.
10. Lack of proper change management by many organizations.
11. Lack of incentives to use the eCRM systems.
12. Business processes versus software driven.

CONCLUSION

The results show that most organizations interviewed appeared to fail in some ways to conduct a proper assessment of business needs before implementing eCRM. Pre-project planning and justification processes were not properly carried out to assess the needs and feasibility of the eCRM projects.

In addition, the extent to which the eCRM system was used was largely depending on the type of the industry, size of the organizations, and type of job responsibilities. Large organizations and organizations in certain industries such as hospitality, and computer hardware and software were most likely to adopt eCRM. Not only did they have higher usage of eCRM, but also they were more likely to implement more sophisticated eCRM systems. eCRM usage had also something to do with the type of job responsibilities. The top management, for example, were more likely to use eCRM more often than their subordinates who did not always perceive eCRM systems as useful and necessary.

Most organizations did not carry out pre-project justification processes. Only half of the organizations interviewed had some sort of justification process. Those which did carry out the processes had very basic form of justification processes such as assessment of the vendor's demo or simple cost/benefit analysis. Furthermore, most organizations claimed to use a variety of criteria to evaluate their IT investments. However, only less than one-third of the organizations interviewed had carried out some sort of evaluation processes (i.e., Scorecard, KPI analysis, qualitative and quantitative analysis).

Finally, no formal IT benefits realization methodology (such as the Cranfield Process Model of Benefit Management (Ward and Daniel, 2006)) or process was specified by any of the participants. This is really a cause for concern as successful eCRM requires that organizations allocate sufficient resources for building customer relationships and continuously evaluating eCRM initiatives. The evaluation and benefits realization mechanisms can expedite the organizational learning process and help make eCRM work to the benefits of all customers and external partners.

FUTURE TRENDS

An article titled "IT Doesn't Matter" has argued that IT has become a commodity because it has become widespread, as happened to other innovations such as engines and telephones (Carr, 2003). However, Carr's (2003) views on IT are not shared by many IT practitioners and academics who argue that IT still has a lot to offer in the future and can deliver competitive advantages to organizations.

In addition, more recent evidence suggests that many organizations simply got carried away with IT and spent money unwisely in late 1990s and early 2000s. It is inevitable that more successful organizations will analyze their economics carefully, spend on only those IT applications that would deliver productivity gains, and evaluate their investments carefully through a disciplined approach with innovative management practices.

REFERENCES

Auer, C. and Petrovic, O. (2003). *Evaluation of CRM-system Success*. 5th Undergraduate and Graduate Students eCommerce Conference (Merkur Day, 2003), Naklo, Slovenia, October 17, 2003.

Bernett, H. G. and Kuhn, M. D. (2002). The Emergence of Electronic Customer Relationship Management, *The Telecommunications Review*, 91-96.

Capgemini. (2004). *Realizing Return on Investment from CRM*. Capgemini Consulting Technology Outsourcing, Source: [On-Line]: http://www.capgemini.com.

Carr, N. G. (2003). IT Doesn't Matter. *Harvard Business Review*, 8(1), 4-50.

Chen, Q. and Chen, H. (2004). Exploring the Success Factors of eCRM Strategies in Practice. *Journal of Database Marketing and Customer Strategy Management*, 11(4), 333-343.

Dibb, S. (2001). Customer Relationship Management and Barriers to the One Segment. *Journal of Financial Services Marketing*, 6(1), 10-23.

EPS, (2001). *eCRM: Putting the Customers First*, EPS Monthly Briefing Paper, December, Source: [On-Line] http://www.epsltd.com.

Ernst and Young. (1999). *Customer Relationship Management*. Special Report: Technology in Financial Services, New York, NY.

Feinberg, R., Kadam, R., Hokama, L., and Kim, I. (2002). The State of Electronic Customer Relationship Management in Retailing. *International Journal of Physical Distribution and Logistics Management*, 30(10), 470-481.

Fjermestad, J. and Romano, N.C., Jr. (2003). Electronic Customer Relationship Management: Revisiting the General Principles of Usability and Resistance. *An Integrative Implementation Framework, Business Process Management Journal*, 9(5), 572-591.

Foley, T. (2002). *Critical Success Factors for a Winning CRM Program*. Inforte Corporation. Source: [On-Line] http://www.realmarket.com/required/inforte2.pdf.

Karakostas, B., Kardaras, D., and Papathanassiou, E. (2005). The State of CRM Adoption by the Financial Services in the UK: An Empirical Investigation. *Information and Management*, 42(6), 853-863.

Kim, J., Suh, E., and Hwang, H. (2003). A Model for Evaluating the Effectiveness of CRM Using the Balanced Scorecard. *Journal of Interactive Marketing*, 17(2), 5-19.

Lin, C. and Pervan, G. (2003). The Practice of IS/IT Benefits Management in Large Australian Organizations. *Information and Management*, 41(1), 13-24.

Lin, C., Pervan, G., and McDermid, D. (2005). IS/IT Investment Evaluation and Benefits Realization Issues in Australia. *Journal of Research and Practices in Information Technology*, 37(3), 235-251.

Love, P., Irani, Z., Standing, C., Lin, C., and Burn, J. (2005). The Enigma of Evaluation: Benefits, Costs and Risks of IT in Small-Medium Sized Enterprises. *Information and Management*, 42(7), 947-964.

McKeen, J.D. and Smith, H.A. (2003). *Making IT Happen: Critical Issues in IT Management*, John Wiley and Sons, Ltd., Chichester, UK.

Melville, N., Kraemer, K., and Gurbaxani, V. (2004). Review: Information Technology and Organizational Performance: An integrative Model of IT Business Value, *MIS Quarterly*, 28(2), 283-322.

Mingers, J. (2001). Combining IS Research Methods: Towards a Pluralist Methodology. *Information Systems Research*, 12(3), 240-259.

Remenyi, D. and Williams, B. (1996). The Nature of Research: Qualitative or Quantitative, Narrative or Paradigmatic? *Information Systems Journal*, 6, 131-146.

Rigby, D.K., Reicheld, F.F., and Schefter, P. (2002). Avoid the Four Perils of CRM. *Harvard Business Review*, 80(2), 101-109.

Sigala, M. (2004). *Customer Relationship Management (CRM) Evaluation: Diffusing CRM Benefits into Business Processes*. The 12th European Conference on Information Systems (ECIS, 2004) , Turku, Finland, June 14-16, 2004.

Steinmueller, W.E. (2002). *Settling the e-CRM Frontier: The Experience of Innovating European Firms, Socio-Economic Trends Assessment for the Digital Resolution (STAR)*. Issue Report No. 23, September.

Straub, D.W., Hoffman, D.L., Weber, B.W., and Steinfield, C. (2002). Measuring e-Commerce in Net-Enabled Organizations: An Introduction to the Special Issue. *Information Systems Research*, 13(2), 115-124.

Sugumaran, V. and Arogyaswamy, B. (2004). Measuring IT Performance: "Contingency" Variables and Value Modes. *The Journal of Computer Information Systems*, 44(2), 79-86.

Tan, X., Yen, D.C., and Fang, X. (2002). Internet Integrated Customer Relationship Management: A Key Success Factor for Companies in the E-commerce Arena. *The Journal of Computer Information Systems*, 42(3), 77-86.

Tsao, H., Lin, K. H., and Lin, C. (2004). An Investigation of Critical Success Factors in the Adoption of B2BEC by Taiwanese Companies, *The Journal of American Academy of Business*, Cambridge, 5(1/2), 198-202.

Ward, J. and Daniel, E. (2006). *Benefits Management: Delivering Value from IS & IT Investments*, John Wiley & Sons, Ltd., Chichester, UK.

KEY TERMS

Analytical eCRM: It is concerned with the technology to process and analyze large amounts of customer data.

Benefits Realization: It is a managed and controlled process of making sure that expected business changes and benefits have been clearly defined, are measurable, and ultimately to ensure that the changes and benefits are actually achieved.

Collaborative eCRM: It is a business model that focuses on creating a real-time eCRM infrastructure to better support customer requirements.

CRM: Any initiative or process designed to assist an organization in optimizing interactions with customers via one or more touch points.

eCRM: It is the element of CRM that uses the Web to create a holistic approach to internal and external communication.

IT Investment Evaluation: This is the weighing up process to rationally assess the value of any acquisition of software or hardware which is expected to improve business value of an organization's information systems.

Operational eCRM: It is concerned with the customer touch points such as automating sales force.

Productivity Paradox: Despite large investments in IT over many years, there has been conflicting reports as to whether or not the IT benefits have actually occurred.

IT Outsourcing Practices in Australia and Taiwan

Chad Lin
Curtin University of Technology, Australia

Koong Lin
National University of Tainan, Taiwan

INTRODUCTION

Globally, information technology (IT) outsourcing has spread quickly in many countries and spending by organizations in IT outsourcing is increasing rapidly each year. According to Gartner (Blackmore, De Souza, Young, Goodness, and Silliman, 2005), total spending on IT outsourcing worldwide is likely to rise from US $184 billion in 2003 to US $256 billion in 2008. However, defining IT outsourcing is not an easy task as it can mean different things to different organizations. Hirschheim and Lacity (2000) define IT outsourcing as the "practice of transferring IT assets, leases, staff, and management responsibility for delivery of services from internal IT functions to third-party vendors." Willcocks and Lester (1997) define outsourcing as the "commissioning of third-party management of IT assets or activities to deliver required results." The scope and range of outsourcing services have also increased as well, as evidenced by the promotion of BPO (business process outsourcing), ASP (applications service providers), global outsourcing, R&D (research and development) outsourcing, and web and e-business outsourcing (Gonzales Gascon and Llopis, 2005; Huang, Lin, and Lin, 2005).

While there is already much research on the economics of IT outsourcing, critical success factors for IT outsourcing decision-making and for outsourcing vendor management (Barthelemy and Geyer, 2004; Hirschheim and Lacity, 2000), there is very little literature on the actual linkage between IT outsourcing and the use of evaluation methodologies in organizations, especially in how these organizations evaluate their IT outsourcing contracts and ensure that the benefits expected from these contracts are delivered eventually.

The aim of this paper is to examine issues surrounding the evaluation and benefits realization processes in Australian and Taiwanese organizations undertaking IT outsourcing. The paper first reviews relevant literature with respect to IT outsourcing, the evaluation of IT outsourcing, and IT benefits realization. Key findings from a survey of the top 2000 Australian organizations, as well as a survey to top 3000 Taiwanese organizations, will then be presented. The paper examines these findings and issues in light of these large organizations' evaluation practices.

BACKGROUND

IT Outsourcing

Whatever the objective, the possibility of IT outsourcing tends to generate strong emotions among the senior executives and external contractors. There are many reasons contributing to the growth of the outsourcing. A review of relevant IT outsourcing literature reveals the following organizational goals for their IT outsourcing projects: lower costs, access to world class expertise, economies of scale, risk sharing, increased efficiency/service level, elimination of internal irritants, higher quality of goods and services, greater focus on core functions, increased flexibility, and reduction in technological obsolescence risk (Aubert, Rivard, and Patry, 2003; Barthelemy, 2003; Kakabadse and Kakabadse, 2001).

There are several important factors that govern successful and less successful outsourcing decisions. These include: differentiation of the business from the competitors, strategic direction of the business, degree of uncertainty of the business environment, scope of outsourcing services, quality of outsourcing contract, technology maturity, level of IT integration, in-house capabilities, and trust (Barthelemy, 2003; Hormozi, Hostetler, and Middleton, 2003). In addition, there are other factors that are more critical for offshore outsourcing than for domestic outsourcing. According to Adelakun (2004), the following critical success factors are very important for offshore outsourcing: people factors (e.g., language skill and project management skill), technical factors (e.g., workers technical skill), business infrastructure factors (e.g., service level agreement details), regulatory factors (e.g., travel and visa restrictions), and client interface factors (e.g., security and trusting relationship). In particular, the traditional approaches to security are failing as we move to open networks and business models due to IT outsourcing (Grimshaw, Vincent, and Willmott, 2002; Wright, 2001). In

addition, IT outsourcing also forces organizations to extend the boundaries of trust outside of their former closed spheres (Wright, 2001). According to Khalfan (2004), these two issues are the most prominent risk factors that would affect the attitudes of organizations to IT outsourcing.

Furthermore, despite the promised savings from the IT outsourcing contracts, there have been problems. These include constant budget blowouts, dubious savings claims, deep dissatisfaction, and non-delivery of service levels (Aubert, et al., 2003; Sullivan and Ngwenyama, 2005). Reasons for this include failing to properly monitor and evaluate IT outsourcing contracts and projects, especially the performance of contractors (Lin, Pervan, and McDermid, 2005; Perrin and Pervan, 2004).

IT Investment Evaluation in IT Outsourcing

Complexity and scope are often the major constraints and difficulties in IT investment evaluation and benefits realization processes (Tallon, Kraemer, and Gurbaxani, 2000; Ward and Daniel, 2006). Many IT projects fail to deliver what is expected of them because organizations focus on implementing the technology rather than tracking and measuring the performance of IT projects (Lin and Pervan, 2003). One reason for this is that most organizations fail to properly monitor and evaluate their IT outsourcing projects (Perrin and Pervan, 2004; Willcocks and Lester, 1997). According to Kakabadse and Kakabadse (2001), the development of suitable methodologies for IT outsourcing has been very slow. For example, McIvor (2000) found that most organizations had no formal process to evaluate their IT outsourcing decision and, instead, relied on limited cost analysis associated with the outsourcing decision. Beaumont and Costa (2002) found that evaluating all costs relevant to outsourcing was a very difficult task. According to Hsu, Wu, and Hsu (2005), most large organizations (52.4%) in Taiwan do not perform evaluation on a regular basis and those organizations which do evaluate tend to do so irregularly. In fact, 15.1% of organizations surveyed did not evaluate at all (Hsu, et al., 2005).

Organizations that make extensive use of IT evaluation methodologies or measures have higher perceived payoffs from IT (Tallon, et al., 2000). Misra (2004) found that outsourcing organizations need to choose the evaluation methodologies which: (a) lead to the desired behavior by both outsourcers and outsourcing contractors; (b) are within the outsourcing contractors' control; (c) can be easily measured by both the outsourcers and outsourcing contractors; (d) can be evaluated by objective criteria rather than subjective criteria; and (e) can be aligned with business objectives.

IT Benefits Realization

While IT investment evaluation is important, it does not guarantee that the benefits identified and expected by organizations are realized (Lin, et al., 2005). This is because IT is just one enabler of process change and it only enables or creates a capability to derive benefits. The essence of benefits realization is to organize and manage so that the potential benefits arising from the use of IT can actually be realized (Changchit, Joshi, and Lederer, 1998).

The identification of expected benefits of a proposed IT outsourcing project is a challenging task. According to Lin and Pervan (2003), very few organizations have a benefits realization approach. Ironically, much attention is paid to ways of justifying investments with little effort being expended in ensuring that the benefits expected are realized. As benefits are frequently long term, uncertain and intangible future benefits are too wide-ranging to be estimated with any accuracy. After all, the critical role of benefits realization depends on external IT outsourcing contractors' ability to not just deliver excellent service, but also to turn this service into organizational consequences such as control of costs, meeting organizational goals, flexibility, and focusing on core functions (Rouse, Corbitt and Aubert, 2001).

While there is a clear indication in the literature of a greater reliance on IT outsourcing by organizations, the importance of outsourcing evaluation and benefits processes has received limited attention, as has the linkage between IT outsourcing and the use of IT investment evaluation and benefits realization methodologies.

RESEARCH METHODOLOGY AND FINDINGS

Research Objectives and Methodology

Corporate spending on IT outsourcing is increasing at a rapid rate. In Australia, there is an increasing push by businesses for offshore IT outsourcing (to India, in particular), although many industry executives believe that Australia can also become an offshore destination as it is at least 25% cheaper to run a commercial undertaking in Australia than in the US or Western Europe (Hollands, 2004). In Taiwan, foreign companies spent a total of US $66 billion on IT outsourcing to Taiwan in 2005 and over 70% of Taiwan's IT output was actually outsourced to China (Burns, 2006). However, no research has been carried out to obtain an overview of IT investments and benefits management processes and practices in these two economies. The research aims to provide new empirical evidence comparing Australia (a developed

economy) and Taiwan (a newly industrialized economy) on their IT outsourcing investment evaluation and benefits realization practices.

The survey approach was chosen as it has the advantage of being able to focus on problem solving and pursue a step-by-step logical, organized, and rigorous method to identify problems, gather data, analyze the data, and draw valid conclusions (Sekaran, 1984).

Specifically, the survey sought to:

1. establish current practices and norms in managing IT outsourcing benefits and evaluation by organizations in Australia and Taiwan; and
2. investigate the usage of the IT outsourcing investment evaluation and benefits realization methodologies or approaches by organizations in Australia and Taiwan.

The sample for the Australian study was obtained by mailing questionnaires in 2005 to the IT managers and CIOs of 900 Australian organizations randomly selected from the top 2000 Australian organizations (Dun and Bradstreet mailing list). Prior to determining the sample size for the survey, a pilot survey of IT managers/CIOs of ten companies in Australia and Taiwan was conducted. Comments about the pilot questionnaire were all positive and so no significant changes were made to the questionnaire. The survey elicited a total of 176 responses and a response rate of 19.6%.

The sample in Taiwan was selected from a list published by a semi-governmental organization, the Institute for Information Industry (III, 2005). Questionnaires were sent to top 3000 organizations in Taiwan in 2005 and 889 questionnaires were returned (a response rate of 29.6%). In the absence of objective data on the organizations' evaluation practices, the IT executives' perceptions were used. Although there has been some debate regarding the legitimacy of perceptual measures as a proxy for objective measures, research has succeeded in alleviating some of the concerns by showing that perceptual measures of organizational performance has a strong positive relationship with more traditional objective measures (Tallon, et al., 2000). For example, a study by Venkatraman and Ramanujam (1987) showed that there was a high degree of correlation between perceptual and objective performance measures in the process of measuring performance of several competing organizations.

Chi-squared Goodness of Fit tests, on industry sector, net revenue, and total number of employees, showed that the sample respondents were statistically similar (at the 1% significance level) to the target population. Late returns were compared with other response received earlier in order to check for non-response bias. No significant differences were detected between two samples (Armstrong and Overton, 1977). Therefore, the respondents can be considered representative of the population as a whole.

Most of the information presented below is based on descriptive statistics (i.e., frequencies) but some comparisons between groups were made using crosstabs, ANOVA, and correlation statistics. In the following discussion of results the percentages referred to normally represented the proportion of valid (answered) cases only and did not indicate missing values. A statistical software package, SPSS (v11.05), was deployed to analyze the quantitative data collected through the survey.

Survey Findings and Discussion

In the following discussion of results the percentages referred to normally represented the proportion of valid (answered) cases only and did not indicate missing values. Additionally, most of the information presented below was based on descriptive statistics.

Overall, the responding Australian organizations were large in revenue and number of employees, typical of the large corporate sector with large numbers from wholesale and retail (18.2%), government and utilities (15.3%), construction, mining, and engineering (11.9%), and health and pharmaceutical services (11.4%) (Table 1).

On the other hand, most Taiwanese organizations were from manufacturing (31.7%), wholesale and retail (25.3%), and information technology and communication industries (Table 1). More than half of the responding organizations had more than US $50 million in total revenue and 250 employees (Table 2).

IT Investment Evaluation and Benefits Realization for IT Outsourcing

Respondents were asked about adoption, usage and success with formal methodologies or processes for various IT outsourcing activities. The Australian results revealed a reasonably high adoption of methodology for IT investment evaluation (67.6%), but less for IT benefits realization (41.5%). However, the results also showed that 15.3% had failed to adopt an IT investment evaluation methodology while 32.4% of responding organizations failed to adopt an IT benefits realization methodology. Therefore, overall, their use was found to be commonplace, but by no means universal. In particular, a significant majority had a formal methodology or process for their IT investment appraisal. On the other hand, the survey results in Taiwan also revealed lower adoption rates for IT investment evaluation (43.9%) and IT benefits realization (42.1%). The ANOVA revealed that both Australian and Taiwanese organizations tend to adopt either both methodologies or none at all.

Table 1. Background information of the respondent organizations – industries & IT budget

	Australia	Taiwan
Range	**Percent (%)**	**Percent (%)**
(a) Industries		
Wholesale and retail	18.2	25.3
Government and utilities	15.3	2.1
Construction, mining & engineering	11.9	0.7
Health and pharmaceutical services	11.4	1.2
Manufacturing	9.7	31.7
Financial and Insurance Services	6.9	3.6
IT and communication	6.8	20.2
Education	5.1	6.6
Transportation	4.5	7.1
Other	10.2	1.5
Total	**100**	**100**
(b) IT budget (A$m)		
<1	18.8	9.5
1-5	25.0	20.9
6-10	13.6	23.1
11-20	18.2	10.7
21 and above	19.9	12.7
Unsure/do not know	4.5	23.1
Total	**100**	**100**

Table 2. Background information of the respondent organizations – IT budget & employees

	Australia	Taiwan
Range	**Percent (%)**	**Percent (%)**
(c) Total revenue (A$m)		
<6	4.0	1.4
6-10	5.1	10.5
11-50	7.4	13.1
51-100	16.5	16.5
101 and above	63.1	57.9
Unsure/do not know	3.9	0.6
Total	**100**	**100**
(d) Total number of employees		
<51	7.4	16.1
51-250	7.4	11.9
250-500	6.8	26.0
501-1000	21.6	6.6
1001-5000	46.0	16.3
5000 and above	10.8	23.1
Total	**100**	**100**

In addition, Australian respondents indicated that IT investment evaluation methodology was widely used in 50.6% of cases. However, only 29.0% of the respondents had pointed out that an IT benefits realization methodology was widely used in their organizations. On the other hand, the Taiwanese respondents reported that both methodologies were widely used in only 38.9% and 39.9% of cases, respectively.

In terms of *effectiveness* of those methodologies in ensuring successful IT outsourcing, 46.1% and 32.4%, respectively, of the Australian organizations pointed out that they were effective. Overall, the IT investment evaluation methodology was not effective in ensuring successful IT outsourcing. Furthermore, IT benefits realization methodology was neither widely used nor effective in ensuring successful IT outsourcing. The figures for the effective utilization of these methodologies by the Taiwanese organizations were 40.6%, and 38.7%, respectively.

Motivation for IT Outsourcing

Almost all organizations surveyed had outsourced at least part of their IT functions. Cost reduction (70.0%) due to lower salaries was considered an important factor for IT outsourcing by Australian organizations. However, this was given a very low priority by the Taiwanese organizations. Instead, time saving (68.7%) and ability to focus on core activities (68.4%) were the most often mentioned benefits for IT outsourcing in Taiwan. Furthermore, most Australian and Taiwanese organizations preferred those external IT outsourcing contractors which had a good track record and extensively skills and experience in outsourcing. The size of external IT outsourcing firms was not an important selection factor for both Australian and Taiwanese organizations.

FUTURE TRENDS

IT outsourcing spending will continue to rise in the future. The rising price of oil will put increasing pressure on organizations to both utilize technology and outsource to remain competitive. Despite the recent debates in the US and other western countries about outsourcing of skilled IT jobs to other low-cost countries such as India and China, and about organizations' obligations to the broader stakeholder community, offshore IT outsourcing has often been employed by most large organizations to reduce the cost of future IT investments and to improve the cash flow of the organizations (Burns, 2006; Hollands, 2004; Rottman and Lacity, 2004).

CONCLUSION

This paper seeks to provide new empirical evidence of IT investment evaluation and benefits realization practices in large Australian and Taiwanese organizations undertaking IT outsourcing projects. The results indicate that while the usage of IT investment evaluation methodologies by Taiwanese organizations is lower than organizations in Australia, the usage of IT benefits realization methodologies is about the same for both groups of organizations. In addition, large Australian organizations are more able to conduct IT investment evaluation methodologies more widely and effectively, whereas large Taiwanese organizations are more likely to deploy IT benefits realization methodologies widely and effectively.

The surveys also show that the extent of outsourcing is quite high in large Australian and Taiwanese organizations. Furthermore, Australian and Taiwanese organizations differ on their motivation for IT outsourcing. Cost reduction is the number one reason for IT outsourcing for large Australian organizations while time saving and ability to concentrate on core activities are the most often mentioned IT outsourcing reasons for Taiwanese organizations.

A key contribution of this exploratory comparative study is to provide new empirical evidence on IT outsourcing practices as well as IT investment evaluation and benefits realization processes in large Australian and Taiwanese organizations. The findings here can assist senior executives in making their IT outsourcing decisions. Finally, our study took place at a particular point in time. This research has relied on the information provided at a particular point in time. Further research could take a longitudinal approach as the perception and evaluation of benefits are likely to change over time. Alternatively, our study could be replicated in a few years time in other countries.

REFERENCES

Adelakun, O. (2004). *IT Outsourcing Maturity Model*. The 12th European Conference on Information Systems (ECIS2004). Turku, Finland, June 14-16.

Armstrong, J. S. & Overton, T. (1977). Estimating Nonresponse Bias in Mail Surveys. *Journal of Marketing Research*, 14 August, 396-402.

Aubert, B.A., Rivard, S., and Patry, M. (2003). A Transaction Cost Model of IT Outsourcing. *Information and Management*, 41, 921-932.

Barthelemy, J. (2003). The Hard and Soft Sides of IT Outsourcing Management. *European Management Journal*, 21(5), 539-548.

Barthelemy, J. and Geyer, D. (2004). The Determinants of Total IT Outsourcing: An Empirical Investigation of French and German Firms. *The Journal of Computer Information Systems*, 44(3), 91-97.

Beaumont, N. and Costa, C. (2002) Information Technology Outsourcing in Australia. *Information Resources Management Journal*, 15(3), 14-31.

Blackmore, D., De Souza, R., Young, A., Goodness, E., and Silliman, R. (2005). Forecast: IT Outsourcing. *Worldwide, 2002-2008* (Update), Gartner, 8 March.

Burns, S. (2006). IT Industry Spends $66bn in Taiwan. *VNU Business Publications*, VNUNet.com.

Changchit, C., Joshi, K.D., and Lederer, A.L. (1998). Process and Reality in Information Systems Benefit Analysis. *Information Systems Journal*, 8, 145-162.

Grimshaw, D., Vincent, S. and Willmott, H. (2002). Going privately: partnership and outsourcing in UK public services. *Public Administration* 80(3), 475-502.

Gonzalez, R., Gasco, J. and Llopis, J. (2005). Information Systems Outsourcing Reasons in the Largest Spanish Firms. *International Journal of Information Management*, 25(2), 117-136.

Hirschheim, R. and Lacity, M. (2000). The Myths and Realities of Information Technology Insourcing. *Communications of the ACM*, 43(2), February, 99-107.

Hollands, M. (2004). Status of Offshore Outsourcing in Australia: A Qualitative Study, AIIA Report. *Australian Information Industry Association*.

Hormozi, A., Hostetler, E., and Middleton, C. (2003). Outsourcing Information Technology: Assessing Your Options. *SAM Advanced Management Journal*, 68(4), 18-23.

Hsu, C., Wu, C., and Hsu, J. C. (2005). Performance Evaluation of Information System Outsourcing in Taiwan's Large Enterprises. *Journal of American Academy of Business*, 6(1), 255-259.

Huang, Y., Lin, C., and Lin, H. (2005). Techno-economic Effect of R&D Outsourcing Strategy for Small and Medium-sized Enterprises: A Resource-Based Viewpoint. *International Journal of Innovation and Incubation*, 2(1), 1-22.

III (2005). *Institute for Information Industry* (III), Source: [On-Line] http://www.iii.org.tw.

Kakabadse, A. and Kakabadse, N. (2001) Outsourcing in the Public Services: A Comparative Analysis of Practice, Capability, and Impact. *Public Administration and Development*, 21, 401-413.

Khalfan, A. M. (2004). Information Security Considerations in IS/IT Outsourcing Projects: A Descriptive Case Study of Two Sectors. *International Journal of Information Management*, 24, pp. 29-42.

Lin, C. and Pervan, G. (2003). The Practice of IS/IT Benefits Management in Large Australian Organizations. *Information and Management*, 41(1), 13-24.

Lin, C., Pervan, G., and McDermid, D. (2005). IS/IT Investment Evaluation and Benefits Realization Issues in Australia. *Journal of Research and Practices in Information Technology*, 37(3), 235-251.

McIvor, R. (2000). A Practical Framework for Understanding the Outsourcing Process. *Supply Chain Management*, 5(1), 22.

Misra, R. B. (2004). Global IT Outsourcing: Metrics for Success of All Parties. *Journal of Information Technology Cases and Applications*, 6(3), 21-34.

Perrin, B. and Pervan, G. (2004). *Performance Monitoring Systems for Public Sector IT Outsourcing Contracts*. Proceedings of the 15th International Conference of the Information Resources Management Association (IRMA 2004), New Orleans, LA, USA. 23-26 May.

Rottman, J. W. and Lacity, M. C. (2004). Twenty Practices for Offshore Sourcing. *MIS Quarterly Executive*, 3(3).

Rouse, A. C., Corbitt, B. and Aubert, B. A. (2001). *Perspectives on IT Outsourcing Success: Covariance Structure Modelling of a Survey of Outsourcing in Australia*. Cahier du GreSI, no 01-03, HEC, Montreal, Canada, 1-18.

Sekaran, U. (1984). *Research Methods for Managers: A Skill-building Approach*. John Wiley & Sons, New York, USA.

Sullivan, E. W. and Ngwenyama, O. K. (2005). How Are Public Sector Organizations Managing IS Outsourcing Risks? An Analysis of Outsourcing Guidelines From Three Jurisdictions. *Journal of Computer Information Systems*, 45(3), pp73-87.

Tallon, P. P., Kraemer, K. L., and Gurbaxani, V., (2000). Executives' Perception of the Business Value of IT: A Process-Oriented Approach., *Journal of Management Information Systems*, 16(4), 145-173.

Tallon, P. P., Kraemer, K. L. and Gurbaxani, V. (2000). Executives' Perceptions of the Business Value of Information Technology: A Process-Oriented Approach. *Journal of Management Information Systems*, 16(4), 145-173.

Venkatraman, N. and Ramanujam, V. (1987). Measurement of Business Economic Performance: An Examination of Method Convergence. *Journal of Management*, 13(1), 109-122.

Ward, J. and Daniel, E. (2006). *Benefits Management: Delivering Value from IS & IT Investments*. John Wiley & Sons Ltd., Chichester, UK.

Willcocks, L. and Lester, S. (1997). Assessing IT Productivity: Any Way Out of the Labyrinth?, In Willcocks, L., Feeny, D.F. & Islei, G. (Eds.). *Managing IT as a Strategic Resource*, Ch4, The McGraw-Hill Company, London, 64-93.

Wright, A. (2001). Controlling Risks of E-commerce Content. *Computers & Security*, 20(2), 147-154.

KEY TERMS

Benefits Realization: It is a managed and controlled process of checking, implementing, and adjusting expected results and continuous adjusting the path leading from investments to expected business benefits.

IT Benefits Realization Methodologies: Approaches that are used to ensure that benefits expected in the IT investments by organizations are eventually delivered.

IT Investment Evaluation: This is the weighing up process to rationally assess the value of any in-house IT assets and acquisition of software or hardware which are expected to improve business value of an organization's information systems.

IT Investment Evaluation Methodologies: Approaches that are used to evaluate and monitor organizations' IT investments.

IT Outsourcing: The practice of transferring IT assets, leases, staff, and management responsibility for delivery of services from internal IT functions to external contractors.

SPSS: It is a statistical and data management software package for analyzing collected questionnaire data.

Survey Research: It is a research method using questionnaires to obtain the required information.

IT Supporting Strategy Formulation

Jan Achterbergh
Radboud University of Nijmegen, The Netherlands

INTRODUCTION

This overview approaches information and communication technology (ICT) for competitive intelligence from the perspective of strategy formulation. It provides an ICT architecture for supporting the knowledge processes producing relevant knowledge for strategy formulation.

To determine what this architecture looks like, we first examine the process of strategy formulation and determine the knowledge required in the process of strategy formulation. To this purpose, we use Beer's viable system model (VSM). Second, we model the knowledge processes in which the intelligence relevant for the process of strategy formulation is produced and processed. Given these two elements, we describe an ICT architecture supporting the knowledge processes producing the knowledge needed for the strategic process.

BACKGROUND: STRATEGY FORMULATION, A VIABLE SYSTEM PERSPECTIVE

Strategy formulation aims at developing and selecting goals and plans securing the adaptation of the organization to its environment. These goals and plans may refer to specific product-market-technology combinations (PMCs) for which the organization hypothesizes that they ensure a stable relation with its environment. The process of strategy formulation needs to generate such goals and plans, needs to reflect upon their appropriateness, and needs to select certain goals and plans to guide the behavior of the organization. This is a continuous process. Goals and plans can be seen as hypotheses about what will work as a means to adapt and survive. Therefore, they should be monitored constantly and revised if necessary. In short, strategy formulation is a continuous contribution to maintaining organizational viability.

Although many authors deal with the process of strategy formulation, we choose the viable system model of Beer (1979, 1981, 1985) to define this process more closely. We select the VSM because Beer explicitly unfolds the functions required for the viable realization and adaptation of an organization's strategy.

To explain what these functions entail, it is useful to divide them into two groups: functions contributing to the *realization* of the organization's strategy and functions contributing to its *adaptation*.

The first group deals with the realization of the organization's strategy. It consists of three functions. Function 1 comprises the organization's primary activities constituting its "raison d'être" (Espejo, Schumann, Schwaninger, & Billello, 1996, p. 110). Function 2 (coordination) coordinates interdependencies between these primary activities. The third function is called the control function. It ensures the synergy of and cohesion between the primary activities by specifying their goals and controlling their performance.

To illustrate these functions, consider Energeco, a company servicing its environment with eco-energy. Function 1 of Energeco consists of three primary activities: supplying solar, tidal, and wind energy. To give an example of the coordination function, suppose that specialists in high-voltage energy are a shared resource between Energeco's business units. Also suppose that there is no coordination between these business units. In this case, the allocation of high-voltage specialists to a project in the business unit Solar Energy may require a revision of the allocation of these same specialists to a project in the business unit Wind Energy. Without a function supporting the coordination of these interdependencies, the business units Solar Energy and Wind Energy may become entangled in a process that oscillates between allocating and revising the allocation of these specialists to projects. It is the task of Function 2 to coordinate these interdependencies. The control function's task is to translate the identity and mission of the viable system (for Energeco, supplying eco-energy) into goals for the primary activities (in this example, supplying wind, solar, and tidal energy) and to control the realization of these goals.

The second group deals with the adaptation of the organization's strategy. It consists of control (Function 3), intelligence (Function 4), and policy (Function 5). Intelligence scans the organization's relevant environment and generates and proposes plans for adaptation. In the example of Energeco, developments in production technology may introduce the possibility of cost-effective, large-scale production of eco-energy from biomass. Intelligence should pick up these developments, assess them, and if relevant, translate them into proposals for innovation. Because of its knowledge of the potentials for change of the primary activities, control (Function 3) reviews the feasibility of the plans proposed by intelligence. For instance, it may object to the plans proposed

by intelligence because they require a change posing a risk to the performance of the primary activities.

Discussion about the relevance and feasibility of the proposals for adaptation between intelligence and control should produce finalized plans for adaptation. It is the task of the policy function to balance the discussion between intelligence and control and to consolidate the finalized proposal in the organization's strategy. For instance, in the discussion between intelligence and control about the feasibility of the adoption of large-scale production of eco-energy from biomass, the policy function should ensure that control and intelligence are equally represented in the discussion. By opting for the production of energy from biomass, the policy function consolidates producing eco-energy from biomass as a new goal for Energeco. Figure 1 depicts the process of strategy formulation in terms of the VSM functions and activities.

To contribute to the strategy-formulation process, control, intelligence, and policy require knowledge about particular domains. Table 1 provides an overview of the knowledge required by each function to contribute to the process of strategy formulation.

Given the overview of functions involved in the strategy-formulation process, their relations, and the knowledge required by these functions to contribute to the process of strategy formulation, it is now possible to look into the knowledge processes needed to produce this knowledge and the ICT architecture supporting these knowledge processes.

KNOWLEDGE PROCESSES CONTRIBUTING TO STRATEGY FORMULATION

The question for this section is by means of what processes knowledge in the knowledge domains should be produced and processed so that the process of strategy formulation can take place. To answer this question, we first need to specify what these knowledge processes are. Then we need to link these processes to the knowledge required by control, intelligence, and policy to contribute to the strategy-formulation process.

We distinguish four relevant processes for producing and processing knowledge: generating (G), sharing (S), retaining (R), and applying (A) knowledge (cf. Achterbergh & Vriens, 2002; Bukowitz & Williams, 1999; Davenport & Prusak, 1998).

These four knowledge processes can now be linked to the process of strategy formulation, as formulated according to the VSM. According to the VSM, the functions intelligence, control, and policy contribute to strategy formulation. This contribution involves the *application* of knowledge in the knowledge domains to arrive at the four core products of strategy formulation: proposals for innovation, their reviews, the finalized plans for innovation, and their consolidation. The knowledge applied by each function is *generated* either by that function or by one of the other functions of the VSM. In the latter case, knowledge must be *shared* between functions. Applying, generating, and sharing knowledge requires the *retention* or *storage* of knowledge.

Figure 1. The process of strategy formulation according to the VSM

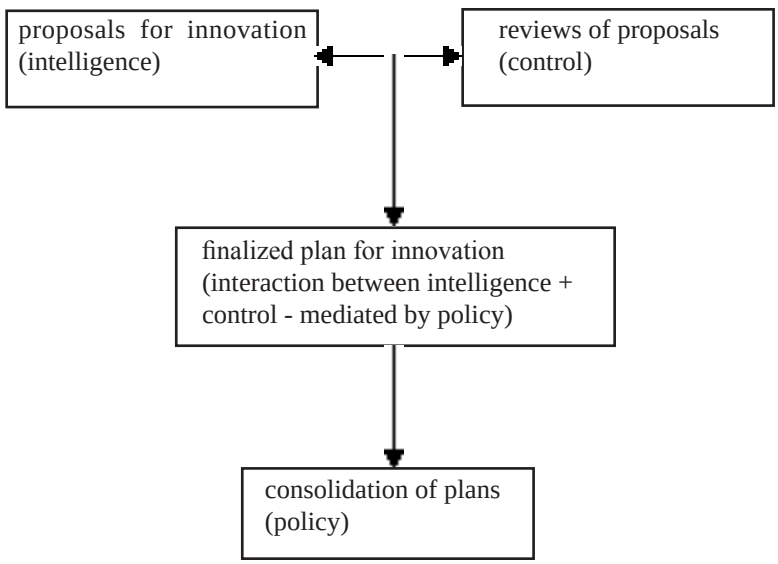


Table 1. Knowledge required by each function to contribute to the strategy formulation process

Function	Related domains of knowledge
F3: Function 3 (control)	For reviewing F4 proposals Organizational goals Proposals for innovation made by F4 Desired goals for F1 based on proposals for innovation Expected performance of the primary activities (goals for F1 activities) Gap between desired and current goals for F1 Required capacity for reorganization of F1 activities Modus operandi of F1 activities Actual capacity for reorganization of F1 activities Gap between required and actual capacity for reorganization Review of proposals for innovation Finalized plans for adaptation of organizational goals (a joint F3 and F4 product) Regulatory measures to counter the imbalance between F3 and F4 (see Function 5)
F4: Function 4 (intelligence)	Organizational goals Goals set by performance and modus operandi of F1 activities Developments in the relevant environment of the organization Reviews by F3 of proposals for innovation Regulatory measures to counter the imbalance between F3 and F4 (see Function 5) Finalized plans for adaptation of organizational goals (a joint F3 and F4 product)
F5: Function 5 (policy)	For balancing purposes Norms for balance between F3 and F4 Proposals by F4 and their reviews by F3 (relative contribution of F3 and F4 to the discussion on adaptation) Actual (im)balance between F3 and F4 Causes of imbalance between F3 and F4 Experiences with regulatory measures to counter the imbalance between F3 and F4 Regulatory measures to counter the imbalance between F3 and F4 For consolidation purposes Finalized plans for adaptation of organizational goals (a joint F3 and F4 product) Organizational goals

Table 2 provides an overview of the relation between the five functions in the VSM, the knowledge domains, and the application and generation of knowledge in these domains. Based on this table, it is possible to draw conclusions about sharing and retaining knowledge. In the table we only included the relevant knowledge for strategy formulation. However, some of this knowledge is generated by Function 1; this is the reason of its inclusion in the table.

The first column of Table 2 summarizes the knowledge domains listed in Table 1. In this column, we eliminated all redundant entries. Columns 2 to 5 indicate whether knowledge in a specific knowledge domain is generated and/or applied by a specific function.

Given the link between the knowledge processes, the functions contributing to the strategy-formulation process, and the knowledge required by them, it is now possible to outline an ICT architecture that can support the generating, retaining, sharing, and applying of this knowledge by these functions.

AN ICT ARCHITECTURE SUPPORTING KNOWLEDGE PROCESSES NEEDED FOR STRATEGY FORMULATION

Knowledge from several knowledge domains specified in Table 2 should be generated, stored, shared, and applied to take the steps in the process of strategy formulation: formulating proposals for innovation, reviewing them, making finalized plans for innovation, and consolidating them. We use these steps in the process of strategy formulation as a point of departure for outlining an ICT architecture (cf. Laudon & Laudon, 1997; Tan, 2003; Turban, McLean, & Wetherbe, 2002) for an information system supporting this process. In the literature, ICT architectures are presented as conceptual models, specifying (at a general level) the parts of an ICT infrastructure (applications, databases, technological ICT elements) and their relations. In this chapter we focus on the application and databases parts. An outline of the architecture is presented in Figure 2.

This architecture consists, ideally, of five modules and knowledge and/or databases. The modules (at the right in

Table 2. Functions, knowledge domains, and knowledge processes for strategy formulation

Knowledge domains	F1	F3	F4	F5
Goals set by performance and modus operandi of the primary activities in F1	G,A	A	A	
Organizational goals	A	A	A	G,A
Proposals for innovation made by F4		A	G,A	A
Desired goals for F1 based on proposals for innovation		G,A		
Gap between desired and current goals of F1		G,A		
Required capacity for reorganization of F1 activities		G,A		
Actual capacity for reorganization of F1 activities		G,A		
Gap between required and actual capacity for reorganization of F1 activities		G,A		
Reviews by F3 of proposals for innovation		G,A	A	A
Finalized plans for adaptation of organizational goals (a joint F3 and F4 product)		G,A	G,A	A
Regulatory measures to counter the imbalance between F3 and F4		A	A	G,A
Developments in the relevant environment of the organization			G,A	
Norms for balance between F3 and F4				G,A
Actual imbalance between F3 and F4				G,A
Causes of imbalance between F3 and F4				G,A
Experiences with regulatory measures to counter the imbalance between F3 and F4				G,A

Figure 2. Outline of an architecture of an information system supporting strategy formulation

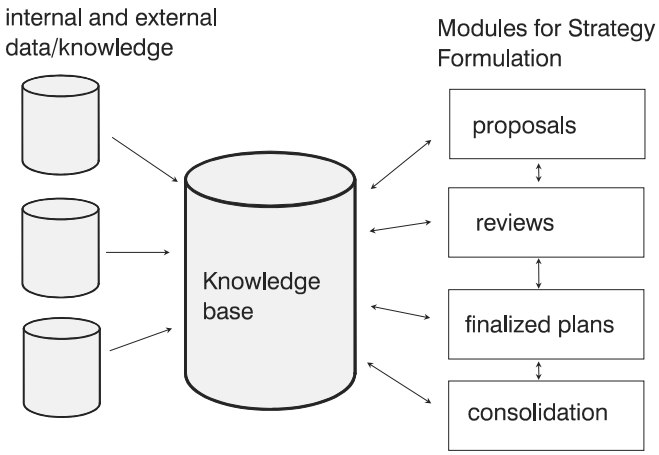


Figure 2) are applications helping to generate the products of the process of strategy formulation. With the help of these modules, the knowledge from the knowledge domains is applied to produce the proposals, reviews, and (consolidated) plans. The architecture further consists of a central knowledge base in which the knowledge in the knowledge domains necessary for strategy formulation (see Table 2) is stored. This central knowledge base in turn may receive knowledge from other internal and external knowledge and/or databases. Below, we discuss the modules and knowledge bases and their relation to relevant knowledge processes in the course of strategy formulation in more detail.

1. The proposal module
The main product of this module is a list of innovation proposals and their justification. To produce this list, one should have access to the knowledge in the relevant knowledge domains. To generate this knowledge, the module should have access to external and internal information. For instance, it may have access to a data warehouse by means of a front-end tool, or it may have access to external online databases. Furthermore, the module may have access to a database consisting of previously rejected or accepted proposals. The proposals for innovation produced with this module are stored in the central knowledge base.

2. The review module
   The input for this module consists of the proposals for innovation. The output is a list of accepted and rejected proposals and the reasons for their acceptance or rejection. To make this list, the module should apply the knowledge in the central knowledge base. This knowledge may be available or may have to be generated. To generate the knowledge, access to several internal and external databases may be required. Also, (external) data on the results of the current PMCs may be input for rejecting or accepting innovations. The review module may benefit from a database with (a classification of) reasons for acceptance or rejection.

3. The finalized-plans module
   This module is mainly a means for sharing proposals for innovation (and their reviews) in order to arrive at a finalized plan. It overarches the proposal and review module. By means of this module, results of the review module are shared and applied to revise the proposals (with the aid of the proposal module). The revised proposals are, in turn, used to produce new reviews (with the aid of the review module) and so forth. This module should (a) facilitate sharing proposals and (b) ensure the finalization of an innovation plan. To these ends, this module should support sharing knowledge about

   • the rules for interaction (such as discussion format and deadlines),
   • criteria for imbalance in the discussion,
   • a monitoring function regarding the imbalance,
   • rules and incentives for countering this imbalance, and
   • an overview of the history of the discussion (as well as an overview of previous discussion).

   Implementation could be by means of intranet applications (e.g., an internal discussion site).

4. The consolidation module
   This module has as its output the consolidation of (a specific selection of) the innovations on the finalized list of innovations. To make this selection, the arguments used in the previous modules should be scanned and valued. Its main goal is to share the results of the strategy-formulation process with relevant parties in the organization. It should enable sharing knowledge about (a) the selected innovations, (b) the reasons for their selection, and (c) their consequences for the current way of doing business. The process of sharing may benefit from a database with (previously successful) communication formats that can be a part of the consolidation module.

5. The central knowledge base
   The central knowledge base consists of all the knowledge in the knowledge domains relevant for strategy formulation. The knowledge base stores the knowledge produced in the modules and supports these modules by servicing them with knowledge relevant to their processes.

Above, an ICT architecture is outlined for an information system supporting strategy formulation. It shows how support should be focused on the products of strategy formulation. Moreover, the focus of the support is on the four knowledge processes involved in the production of proposals, reviews, plans, and consolidations. That is, the application of knowledge leads to proposals for innovation, reviews of these proposals, finalized plans, and consolidation of selected finalized plans. For these products, knowledge from the knowledge domains should be generated, stored, and/or shared. This knowledge is (partly) stored in the knowledge base. The knowledge may be generated by using the four modules and/or by using internal or external databases. Furthermore, knowledge from the knowledge domains may be shared by using connections between the modules.

The description of the architecture specifies the functionality of the different modules in it and how they should be connected. These specifications can be used to select or build the ICT tools to realize the architecture and the knowledge processes it supports.

FUTURE TRENDS

Given the particular outline of the proposed ICT architecture supporting the strategy formulation process, it is possible to link up with current trends that may enhance its performance.

- Developments in the technology for integrating databases (e.g., data warehouse technology) may support the intelligence function in the proposal phase by facilitating the integration and analysis of internal and external knowledge required for strategy formulation.
- Currently, data warehouses are often organized to fit the format of the Balanced Business Scorecard (Kaplan & Norton, 2001). This scorecard is primarily geared to strategy implementation. The format of the VSM, its related knowledge domains, and steps for formulating strategy may be used to organize data warehouses to fit the requirements for strategy adaptation (Achterbergh, Beeres, & Vriens, 2003).

- Proposing and reviewing proposals for adaptation may be enhanced by the application of computer-aided techniques such as gaming, system dynamics, scenario analysis, and group model building.
- By systematically linking strategy formulation to knowledge management, it becomes possible to enhance the quality of the knowledge processes related to strategy formulation by using acquired insights on improving infrastructures for knowledge management.

CONCLUSION

In this overview, we design an ICT architecture supporting strategy formulation on the basis of the viable system model. By applying the viable system model to the strategy-formulation process, it becomes possible to identify the functions required for strategy formulation, the relations between these functions, and the knowledge required by them.

By identifying the knowledge processes producing and processing this knowledge, and by linking these processes to the functions and the knowledge they require to contribute to the strategy-formulation process, it becomes possible to outline an ICT architecture supporting the processes of generating, retaining, sharing, and applying the knowledge needed for strategy formulation. This architecture consists of five modules dedicated to proposing, reviewing, finalizing, and consolidating strategy changes and related knowledge databases containing knowledge in the knowledge domains required for strategy formulation.

REFERENCES

Achterbergh, J. M. I. M., Beeres, R., & Vriens, D. (2003). Does the balanced scorecard support organizational viability? *Kybernetes, 32*(9/10), 1387-1404.

Achterbergh, J. M. I. M., & Vriens, D. (2002). Managing viable knowledge. *Systems Research and Behavioral Science, 19*, 223-241.

Beer, S. (1979). *The heart of enterprise*. Chichester, England: Wiley.

Beer, S. (1981). *Brain of the firm*. Chichester, England: Wiley.

Beer, S. (1985). *Diagnosing the system*. Chichester, England: Wiley.

Bukowitz, W. R., & Williams, R. L. (1999). *The knowledge management fieldbook*. Edinburgh, Scotland: Pearson.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge*. Boston: Harvard Business School Press.

Espejo, R., Schumann, W., Schwaninger, M., & Billello, U. (1996). *Organizational transformation and learning.* New York: Wiley.

Kaplan, R., & Norton, D. (2001). *The strategy-focused organization: How balanced scorecard companies thrive in the new business environment.* Boston: Harvard Business School Press.

Laudon, K. C., & Laudon, J. P. (1997). *Management information systems* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Tan, D. S. (2003). *Van informatiemanagement naar informatie-infrastructuurmanagement*. Leiderdorp: Lansa.

Turban, E., McLean, E., & Wetherbe, J. C. (2002). *Information technology for management* (3rd ed.). New York: Wiley.

KEY TERMS

ICT: Information and communication technology. ICT can be used to indicate the organization's technological infrastructure (comprising of all hardware, software, and telecommunications technology) and to indicate one or more specific collections of hardware, software, and telecommunications technology (i.e., one or more ICT applications).

ICT Architecture: The ICT architecture provides a conceptual model, specifying (at a general level) the parts of an ICT infrastructure (applications, databases, technological ICT elements) and their relations. In this chapter we concentrate on the application and databases parts.

Knowledge Domain: the knowledge related to defining, recognizing, and solving a specific problem.

Knowledge Processes: In the literature, one often finds four knowledge processes: (a) generating knowledge, (b) sharing knowledge, (c) storing knowledge, and (d) applying knowledge.

Strategy: In the literature, many definitions are given. A possible definition is the desired portfolio of product-market-technology combinations of an organization.

Strategy Formulation: The process by means of which the desired portfolio of product-market-technology combinations is defined and updated. This process can be modeled using the viable system model consisting of four steps: defining proposals for innovation, reviewing these proposals, finalizing proposals, and consolidating finalized proposals.

Viable System Model: This model is developed by Beer (1979, 1981) and specifies the necessary and sufficient functions organizations should possess to maintain a separate existence in their environment.

Viability: Viability is the ability of a system "to maintain a separate existence." Most organizations are continuously trying to maintain their viability.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1728-1734, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Key Factors and Implications for E-Government Diffusion in Developed Economies

Mahesh S. Raisinghani
TWU School of Management, USA

INTRODUCTION

E-government has grown in significance with the growth of the digital age and the global economy; however, at a slower pace. Its impact is pervasive and is evident in the availability and distribution of products and services within agencies, to business, and to citizens in western countries. There are many aspects of e-government; researchers have written extensively on the subject and have reached conclusions that will continue to evolve as new discoveries are made.

The business community, and society at-large, have been challenged by the complexities of e-government in meeting the needs in developing countries as well. The relationship between developed and underdeveloped countries is interdependent due to natural resources that may be available in underdeveloped countries and products that may be modified in their packaging and price for sale to the people in underdeveloped countries (e.g., toothpaste packaged in small, disposable packets, or shampoo in small vials, or second/third generation mobile phones to keep it affordable for the people in the underdeveloped countries). The polarization between e-government and society is due to conflicting financial, geopolitical/ethical and societal goals (Webber, 2006); this issue is evident in the adoption rates and usage of government Web sites. Although progress has been made in identifying e-government opportunities, the juxtaposition of government infrastructure, technology, and societal needs often conflict and, as a result, have adversely impacted the products and services offered by e-governments throughout the western world and ultimately, the adoption rates.

BACKGROUND

A recent visit to some state and local government Web sites, including New York City (http://www.nyc.gov/portal/site/nycgov/) and North Carolina (http://www.ncgov.com/), revealed that some local and state government Web sites in the United States offer static information. Unfortunately, governments have failed to take advantage of the full functionality of the Internet; moreover, they have ignored the upsurge in Internet usage in the commercial sector. Webber (2006) notes that while Canada outpaces the United States in nearly all Internet activities, both countries fail to capitalize on its potential to increase citizen interaction and to reduce costs associated with providing goods, services and information (Webber, 2006, 1). Most sites are based on a brochure-ware format; citizens can download forms, complete them and return the information by traditional postal mail; Webber notes that form downloading is one of the most common activities on US and Canadian sites. Further, he notes the growth and adoption of e-government may be hampered by an antiquated form-based approach to requesting and providing services (Webber, 2006, 1).

E-GOVERNMENT USAGE

An examination of e-government adoption must begin with an examination of Web site usage and an identification or profile of the typical Internet user. Figure 1 compares the profiles of the typical Australian, Canadian, and US online citizens. For those who do go online, many do not take advantage of general portal information.

Australian E-Government User Profile

In March 2006, the results of a recent government prospectus on Australia's e-government, which was, for the first time, endorsed by the prime minister and cabinet were announced,. The prospectus highlighted the government's flawed approach of layering Internet functionality over antiquated policies as opposed to first determining who is using the Internet, that is, user profile, and the purposes for which they use it. Today, only one in three online consumers uses the Internet to access government services. The profile of the typical online Australian government Web site users mirrors that of its American and Canadian counterparts (i.e., government Web site users) in many ways, however, there are some noteworthy differences; these differences are summarized as follows.

- The typical Australian user is a baby boomer in her mid-40s; it should be noted that females dominate online government site use in Australia with 62%. This contrasts with the US and Canada, where usage

Figure 1. Comparison of Australian, US, and Canadian Online Citizen Profiles (Webber, 2006, p. 3)

	Australia	US	Canada
Age	44	46	43
Average household income	A$50,303	US$65,942	C$53,096
Female	62%	52%	51%
Married or living with a partner	57%	70%	69%
Has a college degree	35%	37%	28%
Has broadband at home	40%	42%	57%
Technology optimist	46%	58%	56%
Owns one or two computers	69%	77%	81%
Uses the Internet more than 3 hours per week	79%	70%	61%
Goes online daily	69%	62%	56%

Base: Australian and North American online consumers

Source: Forrester's Consumer Technographics Q1® 2005 North American Retail, Automotive, A Online Study and Forrester's APCTAS Q1 2006 Survey

is equally spread between men and women. Fifty seven percent are married, however, 68% have no children. The average income of the e-government Web site user is 8% higher than their nonuser counterparts (Webber, 2006, p. 3).

- Online activities vary by federal, state, and local Web sites. Activities on a federal Web site include downloading or printing out a government form, filling in tax details or completing tax returns online; accessing information regarding benefit eligibility; accessing employment information; accessing tourism information, and applying online for benefits (Webber, 2006, p. 4). Activities on a state Web site include downloading or printing out government forms, accessing tourism information, license renewal or vehicle registration, accessing employment information and accessing motor vehicle related information (Webber, 2006, p. 5). On the local level, activities include accessing tourism information, accessing employment information, accessing real-estate information, downloading or printing out a government form, and ordering consumer publications (Webber, 2006, p. 5).

Canadian E-Government User Profile

Taken as a whole, Canadians prefer to use other channels to communicate with the government; these include the phone channel, mail, or in-person contact (Cardin, 2006, p. 1). The driver for this behavior is based on the type of inquiry, which, in turn, determines the type of interaction with the government (Cardin, 2006, p. 3). Citizens use the Internet to perform general research, for example, government statistics, laws and regulations, and possible employment opportunities within the government; in contrast, they prefer phones for personal research, including eligibility for health and government services, checking the status of tax returns, or applying for health benefits (Cardin, 2006, p. 2). In-person contact is the channel of choice for passport transactions, social security numbers, and government benefits; mail remains the channel of choice for filing taxes (Cardin, 2006, p. 3).

United States E-Government User Profile

While more than 38 million US citizens applied for Medicare benefits, only 5% applied online. This gap is evident in the security arena as well; although there has been a marked increase in attention to security and safety since 9/11, less than 10% of North American citizens visited government Web sites for information on security, health or safety issues, including the Avian flu, terrorism, or natural disasters (Webber, 2006). The most prolific e-government users are baby boomers; Young boomers range from age 41–50 and older boomers range from age 51-61. The most popular activity is downloading forms and research. Seventeen percent visit state government sites while 15% and 14% visit federal and local sites, respectively (Webber, Holmes, & Hanson, 2006, p. 1). One common thread among US e-government users is they are technology enthusiasts. Not surprisingly, Gen X and Gen Y are the most optimistic about technology, however, older boomers outpace younger boomers (65% to 61%); seniors are the most reticent at 58% (Webber et al., 2006, p. 3).

Although discussion, to this point, has been limited to Australia, Canada, and the US, an examination of the EU is warranted as well. The United Kingdom readily recognizes that numerous benefits could be achieved from increased Internet usage by its citizenry. These include wider participation/reduced social exclusion; improvements in information sharing between services and agencies; greater variety and choice and convenience for access; and improved speed and efficiency (Phippen & Lacohee, 2006). Unfortunately, citizen engagement is poor and lags significantly behind the US and Canada; only 1 in 10 citizens use online government services; this fact is particularly alarming, based on recent expenditures, and begs the question – Do they know their audience and what they want (Phippen & Lacohee, 2006, p. 205)?

Causes of Citizen Apathy Toward E-Government

The causes of apathy or reluctance to use e-government services are both qualitative and quantitative. Qualitative causes include appearance and performance, while quantitative causes include public safety, fraud services, and the digital divide (Webber, 2006).

Appearance and performance includes site responsiveness and readability; factors include speed, functionality, security, and reliable performance. Site trust is critical to older boomers and seniors; appearance and fun factor are less important to all except for Gen Yers. Public safety and fraud services include consumer and safety information leads. E-government users want information on public safety, including emergency services, courts, and health advisories; fraud warnings (scams), product recalls, and general consumer protection information.

Main Management Concerns/Issues in Regard to E-Government

The primary management concerns and issues with regard to e-government are listed as follows; they, too, are exacerbated by security, privacy issues, and financial/cost constraints. These same issues, when viewed from the perspective of

Figure 2. Citizens Look for an Increase in Online Government Services across a Spectrum of Topics (Webber, 2006)

"In which government service and/or department areas would you like to see an increase in eGovernment initiatives (i.e., an increase in information, transactions, and communication methods)?"

	Gen Yers (18 to 26)	Gen Xers (27 to 40)	Younger Boomers (41 to 50)	Older Boomers (51 to 61)	Seniors (62+)
Employment opportunities in government	62%	68%	67%	72%	58%
Worker benefits (e.g., unemployment and disability)	60%	64%	69%	72%	59%
ID cards (e.g., licenses, passports, and green cards)	54%	56%	51%	52%	42%
Parks and recreation opportunities	56%	61%	60%	61%	53%
Human rights, health, and welfare benefits	59%	60%	62%	59%	50%
Consumer information (e.g., fraud services)	59%	71%	75%	76%	68%
Environment and natural resources	55%	61%	63%	61%	54%
Public safety (e.g., emergency services, courts, and health advisories)	63%	68%	72%	74%	63%
Grants available from the government	72%	69%	74%	69%	58%
Housing, planning, and zoning	58%	66%	64%	61%	52%

Base: US online consumers who visit government Web sites twice per year or more (shading indicates the three highest percentages per generation)

Source: Forrester's NACTAS 2006 Benchmark Survey

40044 Source: Forrester Research, Inc.

the citizen, provide additional context on their concerns and issues.

- *Client definition* –. Agencies are set up to work in silos, however, consumers may be daunted by Web sites that are agency-focused when the information need is problem oriented, that is, "I know what I need I just don't know where to get assistance." Having to know the specific address of a government agency in order get services is not a "delighter" and may ultimately drive consumers either to more expensive channels such as phone or in-person or if like services are available, the private sector (Pavlichev & Garson, 2004).
- *Political implications* - The issue of using state Web portals for political purposes is an ethical issue. Sites with political messages should be segregated from sites focused on providing services (Pavlichev & Garson, 2004).
- *Centralization vs. decentralization* – From the consumer's perspective, information should be centralized to ease consumer access; however, decentralization may allow the agency to act as an entrepreneur in integrating horizontally, vertically, and externally to provide the best environment for the consumer (Weber, 2005).
- *Commercialization* – standardization of Web site naming conventions in an ongoing discussion and would provide ease of use for the consumer. Additionally, using the site for advertising space may set expectations for the consumer that the state is endorsing the product (Pavlichev & Garson, 2004).
- *Outsourcing* - While in-house expertise may be a consideration, the public may take a dim view of this practice, as there are issues of public information access and privacy. Additionally, fees may be assessed to access government information would have negative impact on usage (Pavlichev & Garson, 2004).

Figure 2 consolidates responses of survey participants to the question, "In which government Service and/or department areas would you like to see an increase of e-government initiatives?"

THE DIGITAL DIVIDE

The digital divide is defined as patterns of unequal access to information and communications technologies (ICTs); the divide was originally identified in the 1990s, and specifically referenced computers and the Internet (Khosrow-Pour, 2006). It is an accepted fact in academia and in political circles that the lack of computer and Internet access has created a divide between people of color, specifically, African-Americans and Latinos, and whites (Pavlichev & Garson, 2004). In a study conducted in September, 2000, researchers found that 50% of whites had Internet access; conversely, only 36% of African-Americans and 44% of Hispanics have Internet access (Pew Internet & American Life Project, 2000). The study also showed that the divisions were influenced by economic factors; of families with incomes greater than $75,000, 78% of whites, 79% of Hispanics, and 69% of African-Americans were online. The contrast was stark for those with income of less than $30,000; 68% of whites, 75% of African-Americans, and 74% of Hispanics were not online. By 2005, the number of people online line had increased,

Table 1. Digital divide demographics

CATEGORY	STATISTICS
Age – Online access increased as the age of the population decreased	18 – 19: 83%; 20 – 24: 81%; 25 – 29: 78%; 30 – 34: 76%; 35 – 39: 73%; 40 – 44: 69%; 45 – 49: 65%; 50 – 54: 61%; 55 – 59: 55% 60 – 64: 41%; 65 – 69: 27%; 70 – 74: 15%
Employment Status – Online access increased with employment or college attendance	Full time student: 94%; part-time student: 84%; employed: 76%; not employed: 51%; retired: 33%
Household Income – Online access increased in direct correlation to income	=> $75k: 89%; $50k - $75k: 81%; $30k – 50k: 69%; < $30k: 44%
Education attainment – Online access increased in direct correlation to education level	College degree +: 88%; some college: 75%; HS grad: 52%; less than HS: 32%
Race and ethnicity (English-speaking) – Online access varied by ethnic population	Asian-American: 82%; White: 67%; Hispanic: 59%; 43% African American
Community type - Online access varied by community	Suburban: 68%; Urban: 62%; Rural: 56%
Disability: Online access varied by disability	Not disabled: 67%; disabled: 38%

however, ethnic disparity still existed. When viewed by race, 70% of whites go online as opposed to 57% of African-Americans. Additionally, Americans age 65 and older, as well as those with less education, were less connected (Pew Internet & American Life Project, 2005).

Analysis of US Digital Divide Demographics

Table 1 summarizes the findings from the 2005 study on data gathered for the Congressional Internet Caucus (Pew Internet American Life Project 2005):

The digital divide determines which citizen can or will avail themselves of government services. Table 2 documents the number of citizens who used government online services. While these numbers may seem significant, they pale in comparison to the total US 2005 population of more than 296 million. Simply put, more than half of the American population is not participating in e-government. Those citizens who cannot or choose not to participate in government activities are at a disadvantage. One major area of concern is e-democracy, which is defined as all forms of electronic communications between government and its citizens (Pavlichev & Garson, 2004).

Many governments are taking steps to leverage the benefits of ICTs to increase accountability, transparency, and the quality of services to its citizens (Khosrow-Pour, 2006). E-democracy is based on five principals: access, convenience, awareness, communications, and involvement in the political process; the goal is to encourage active rather than passive participation (Pavlichev & Garson, 2004). E-democracy efforts are the focus of governments throughout the world; many governments are creating programs with the sole purpose of engaging citizens in the political and governmental process. While those efforts are admirable and are part of the equation necessary to close the digital divide, the issue of access is paramount and must be addressed.

Recent data shows that progress is being made in the African-American and Hispanic communities. As a result of significant reduction in laptop costs, more computers in public community sites, such as schools and libraries, and availability of mobile devices, such as cellphones, more minorities have online access (Marriott, 2006). A 2006 Pew study noted that 74% of whites go online compared to 61% of African-Americans and 80% of English-speaking Hispanics; these figures represent a significant increase from the 2005 statistics (Marriott, 2006).

One untapped market, from the perspective of government usage, is the mobile market. While the private sector has numerous efforts underway to use the tool for marketing and as an access point for the Internet, the government has yet to follow. In a recent Pew study, the benefits of cellphone usage were documented, and are indicative of strong user engagement. Additionally, they also indicate that users are willing to reallocate their personal and business time, and to intersperse both with mobile activity; moreover, they provide 24x7 availability. Governments should consider these benefits and develop plans to make information available via mobile access, thereby, bringing e-democracy into the 21st century.

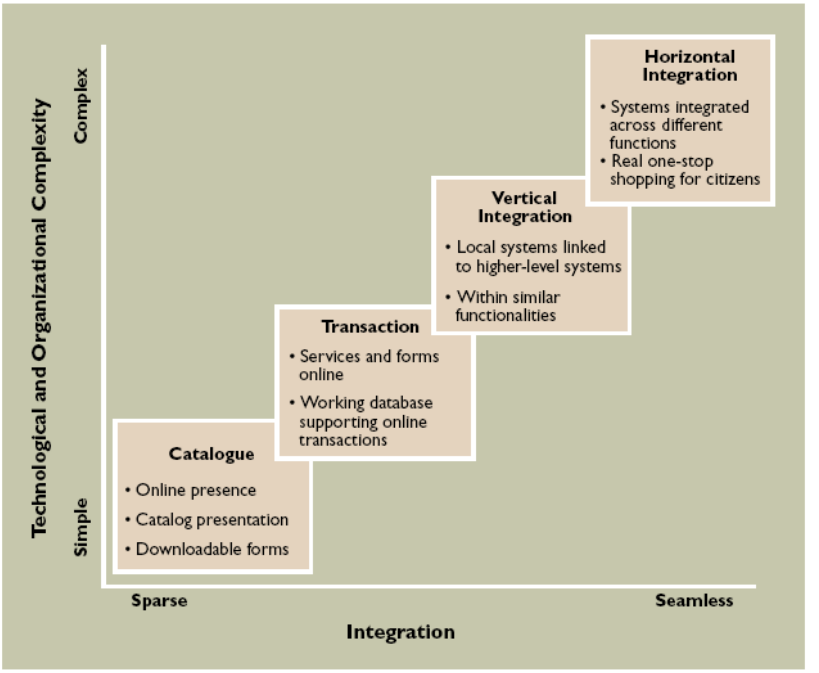
INTERNATIONAL DIGITAL DIVIDE

Efforts are underway to address issues in South American countries, specifically Brazil, Argentina, and Peru. The One Laptop per Child Program (OLPC) is a UN-sanctioned effort to close the digital gap on an international level. Libya, Liberia, Rwanda, and Thailand are included. The program distributes low-cost PCs with simple hardware and software configurations to children in these countries, thus, providing the tools needed for self-directed education. OLPC is not specifically an e-government initiative, but has the goal of giving children in these countries the tools needed to teach

Table 2. eCitizen activities

Activity	**Number of Users Millions)**
Government Web sites	83
2004 political campaign	63
Research on policy issues	52
E-mails regarding 2004 campaign	43
E-mails to government officials to affect changes in policy	38
Access to health and safety information	36
Research/application for government benefits	29
Participation in organized lobbying campaigns	24
Participation in 2004 campaign through donations and/or volunteer activities	11

Figure 3. Evolutionary progression of e-government



themselves (OLPC, 2007). The potential benefits of this effort are many; in addition to the immediate benefit of teaching children in these countries, the greater benefit is that the UN, in conjunction with its partners, Intel, and OLPC, is preparing a generation of computer literate citizens. Moreover, adoption rates may be higher, since the computers and associated Internet access may be the only means of accessing information. This effort may have positive implication for western countries, and should be considered, and potentially included in, future e-government strategies.

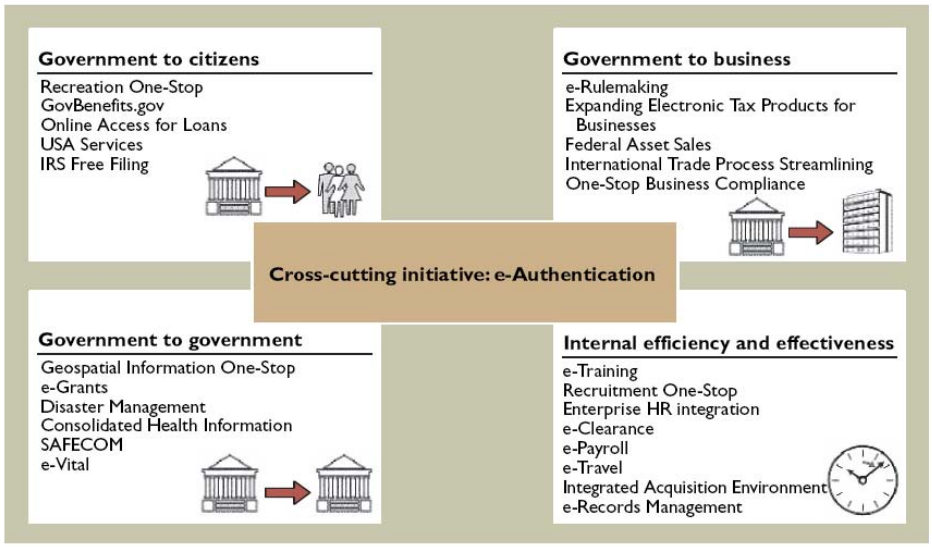
FUTURE TRENDS

The evolution of e-government can be charted as a comparison of technological and organizational complexity vs. the degree of integration. The progression is depicted as a linear relationship, however, government may choose to skip stages or combine them. Figure 3 depicts the dimension and stages of e-government development (Lee, Tan, & Trimi, 2005, p. 100).

Although the US, Canada, and the European Union each recognizes that improvements can be made to increase citizen engagement, each country is recognized as a leader in e-government development. The US effort includes the Office of Management and Budget's e-government strategy, an action plan with 25 initiatives (Lee et al., 2005, p. 100). The strategy, depicted in Figure 3, addresses government to business, government to citizen, government to government, and internal efficiencies and effectiveness.

Efforts in the European Union have been noteworthy as well. In April 2006, the European Commission issued the e-government action plan that provides specific actions needed to support e-government. These include horizontal government integration policies, funding, and coordination activities. The plan outlines high-level goals that will be addressed between 2006 and 2010. The goals include leaving no citizen behind, making efficiency and effectiveness a reality, providing high-impact key services for citizens and businesses, putting enablers in place, and strengthening participation and democratic decision making (Di Maio, 2006). Under this approach, e-government will be available to all citizens, regardless of socioeconomic status, and is expected to increase participation in e-democracy. Additionally, citizen's concerns regarding privacy and security are expected to be addressed as well through the deployment of identity management and document authentication (Di Maio, 2006).

Figure 4. U.S. Office of Management and Budget 25 e-government initiatives



CONCLUSION

E-government has had a pervasive impact on society (e.g., Government to Business (G2B), Government to Consumer (G2C), Government to Government (G2G) e-commerce). Millions of citizens around the world access government Web sites daily; unfortunately, this number is only a fraction of the total population. User profiles of citizens in the United States, Canada, and Australia are similar, and reflect similar usage patterns. Citizen apathy can be categorized as both qualitative, for example, appearance and performance, and quantitative, for example, public safety and fraud services. Within the government itself, issues have been identified that impact the adoption rate as well; these include client definition, political usage, centralization vs. decentralization of information, out-sourcing, and standardization of site naming conventions. One of the most highly publicized reasons identified as an impact to e-government adoption has been the digital divide. Citizens, specifically those of color and of lower socioeconomic status, have had less access to technology and as such, have not been able to avail themselves of the products and services offered via e-government sites. Recent studies show that while the gap is lessening, it still exists. The United Nations has efforts underway in developing countries, such as the One Laptop per Child program in South America and Africa; this program should be adopted in western countries as well.

Governments in western countries continue to look for ways to improve citizen adoption rates; both the United States and the European Union have developed strategic plans that target the issues listed here. In order to continue to improve adoption rates, governments must keep abreast of the technological advances that the Internet and associated technology offers; however, more importantly, they must listen to their customers and build a mousetrap that they want and need.

REFERENCES

Cardin, L. (2006). *Canadians prefer phone and in-person channels for critical government interactions* (09262006). Cambridge, MA: Forrester Research, Inc.

Di Maio, A. (2006). *European commission's e-government plan shows modest ambition* (G00139971). Stamford, CT: Gartner, Inc.

Khosrow-Pour, M. (Ed.). (2006). *Encyclopedia of e-commerce, e-government, and mobile commerce* (Vols. I, A-J (Volume Set)) [Electronic version]. Hershey, PA: IGI Publishing.

Lee, S., Tan, Z., & Trimi, S. (2005). Current practices of leading e-government countries. *Communications of the Acm, 48*(10), 99-104.

Marriott, M. (2006). Digital divide closing as blacks turn to Internet [Electronic version]. *The New York Times*, March 31, 2006, p. 1.

OLPC. (2007). *One laptop per child* [Brochure]. Retrieved August 3, 2007, from http://wiki.laptop.org/go/One_Laptop_per_Child

Pew Internet & American Life Project. (2000). *Who's not online: 57% of those without Internet access say they don't plan to log on* (09212000). Washington, D.C: Lenhart, A.

Pew Internet & American Life Project. (2005). *What people do online* (02092005). Washington, D.C: Rainie, L.

Pew Internet & American Life Project. (2005). *Digital divisions* (10052005). Washington, D.C: Fox, S.

Phippen, A., & Lacohee, H. (2006). E-government - issues in citizen engagement. *Bt Journal, 24*(2), 205.

Weber, A. (2005). *The future of e-government* (05122005). Cambridge, MA: Forrester Research, Inc.

Webber, A. (2006a). *Australia's apathetic e-government users* (06292006). Cambridge, MA: Forrester Research, Inc.

Webber, A. (2006b). *E-government adoption levels: 2006* (09062006). Cambridge, MA: Forrester Research, Inc.

Webber, A. (2006c). *E-government adoption levels: 2006* (09062006). Cambridge, MA: Forrester Research, Inc.

Webber, A., Holmes, B., & Hanson. (2006). *Boomers are the first to tap into online government* (10032006). Cambridge, MA: Forrester Research, Inc.

KEY TERMS

Electronic Government (E-Government): The use of digital technologies to transform government operations in order to improve efficiency, effectiveness, and service delivery.

Government to Business (G2B) E-Commerce: When a government entity sells products and services to businesses.

Government to Consumer (G2C) E-Commerce: The electronic commerce activities performed between the government and its citizens or consumers, including paying taxes, registering vehicles, and providing information and services.

Government to Government (G2G) E-Commerce: Either (1) the electronic commerce activities performed within a single nation's government or (2) the electronic commerce activities performed between two or more nations' governments including providing foreign aid.

Horizontal Government Integration: The electronic integration of agencies, activities, and processes across a special level of government.

Identity Theft: The forging of someone's identity for the purpose of fraud.

Internet: A vast network of computers that connects millions of people all over the world.

World Wide Web (Web): A multimedia-based collection of information, services, and Web sites supported by the Internet.

Keystroke Dynamics and Graphical Authentication Systems

Sérgio Tenreiro de Magalhães
University of Minho, Portugal

Henrique M. D. Santos
University of Minho, Portugal

Leonel Duarte dos Santos
University of Minho, Portugal

Kenneth Revett
University of Westminster, UK

INTRODUCTION

In information systems, authentication involves, traditionally, sharing a secret with the authenticating entity and presenting it whenever a confirmation of the user's identity is needed. In the digital era, that secret is commonly a user name and password pair and/or, sometimes, a biometric feature. Both present difficulties of different kinds once the traditional user name and password are no longer enough to protect these infrastructures, the privacy of those who use it, and the confidentiality of the information, having known vulnerabilities, and the second has many issues related to ethical and social implications of its use (Magalhães & Santos, 2005).

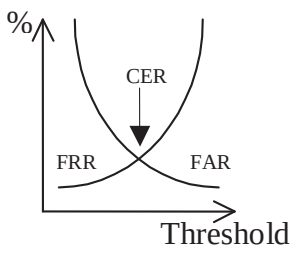
Password vulnerabilities come from their misuse that, in turn, results from the fact that they need to be both easy to remember, therefore simple, and secure, therefore complex. Consequently, it is virtually impossible to come up with a good password (Wiedenbeck, Waters, Birget, Brodskiy, & Memon, 2005). On the other hand, once users realize the need for securing their authentication secrets, even fairly good passwords become a threat when the security policies (if at all existing) fail to be implemented. The results of an inquiry made by the authors in 2004 to 60 IT professionals show that, even among those that have technical knowledge, the need for password security is underestimated (Magalhães, Revett, & Santos, 2006). This is probably one of the reasons why the governments increased their investment in biometric technologies after the terrorist attack of 9/11 (International Biometric Group [IBG], 2003).

The use of biometric technologies to increase the security of a system has become a widely discussed subject, but while governments and corporations are pressing for a wider integration of these technologies with common security systems (like passports or identity cards), human rights associations are concerned with the ethical and social implications of their use. This situation creates a challenge to find biometric algorithms that are less intrusive, easier to use, and more accurate.

The precision of a biometric technology is measured by its false-acceptance rate (FAR), which measures the permeability of the algorithm to attacks; its false-rejection rate (FRR), which measures the resistance of the algorithm to accept a legitimate user; and its crossover error rate (CER), the point of intersection of the FAR curve with the FRR curve that indicates the level of usability of the technology (Figure 1). For a biometric technology to be usable on a stand-alone base, its CER must be under 1%. As an algorithm becomes more demanding, its FAR is lower and its FRR is higher. Usually the administrator of the system can define a threshold and decide what the average FAR and FRR of the applied algorithm will be according to the need for security, which depends on the risk evaluation and the value of what is protected; also, the threshold can be, in theory, defined by an intrusion detection system (software designed to identify situations of attack to the system).

Establishing the error rates of a biometric technology is a complex problem. Studies have been made to normalize

Figure 1. Crossover error rate



their evaluation, but the fact is that the results are strongly dependent on the number of individuals involved in the process and, what is worst, on who is chosen. This means that, even with a large amount of data collected, the results can be very different if we change the evaluated group. The lack of trust in the precision evaluation methodologies and values is one of the reasons why the human rights associations are opposing the generalization of use of biometric technologies and their acceptance as standards for authentication procedures (Privacy International, Statewatch, & European Digital Rights, 2004). Even so, in an inquiry made by Epaynews (http://www.epaynews.com), 36% of users stated that they would prefer to use biometric authentication when using credit cards, a value only comparable to the use of personal identification numbers (PINs) and much higher than the 9% of authentication obtained by signature.

Considering all the advantages and disadvantages of biometric procedures, it seems that the only way is to allow the user a choice. Being so, the traditional password systems must be enhanced both in the biometrical way and in another completely different way. On the biometric component we propose keystroke dynamics, a biometrical authentication algorithm that tries do define a user's typing pattern and then verifies in each log-in attempt if the pattern existing in the way the password was typed matches the user's known pattern; it is the only biometric technology that can be used with the existing log-in and password systems without requiring any extra hardware. On the nonbiometric component, we propose the use of a graphical authentication system, a log-in system that verifies the user's knowledge of specific images or parts of images to grant or deny successful log-in, because it has been proven that it provides a wider key space and because it can be used to generate complex secret strings from simple passgraphs (the user's secret code to access a system protected by a graphical authentication system, constituted by a sequence of points where the user must click in order to obtain a successful log-in).

BACKGROUND

Keystroke Dynamics

As in many other problems, there have been two different approaches to the challenge of finding an algorithm for keystroke dynamics that minimizes the CER: machine learning and deterministic algorithms.

Among the solutions based on machine learning, we can find the work presented by Ord and Furnell (2000) that tested this technology with a 14-person group to study the viability of applying it to the simple use of PINs typed on a numeric pad. Unfortunately, the results suggest that, for large-scale use, the technology is not feasible. Deterministic algorithms have been applied to keystroke dynamics since the late '70s. In 1980, Gaines et al. (1980) presented a report on the study of the typing patterns of seven professional typists. The small number of volunteers and the fact that the algorithm is deducted from their data and not tested for other people later results in lower confidence in the FAR and FRR values presented. However, the method used to establish a pattern was a breakthrough: the study of the time spent to type the same two letters (digraph) when together in the text. Since then, many algorithms based on algebra and on probability and statistics have been presented. Joyce and Gupta presented in 1990 an algorithm to calculate a value that represents the distance between acquired keystroke latency times and correspondent times previously stored. In 1997, Monrose and Rubin used the Euclidean distance and probabilistic calculations based on the assumption that the latency times for one digraph exhibits a normal distribution. Later in 2000, they also presented an algorithm for identification based on the similarity models of Bayes, and in 2001 they presented an algorithm that uses polynomials and vector spaces to generate complex passwords from a simple one using the keystroke pattern (Monrose et al., 2001).

In 2005, Magalhães, Revett, and Santos presented an improvement of the Joyce and Gupta algorithm and tested it with 170.391 attacks to 143 patterns, obtaining a 0% FAR with an FRR of 26%, and an estimated CER below 5%.

Graphical Authentication Systems

A graphical authentication system is a log-in system that verifies the user's knowledge of specific images or parts of images to grant or deny successful log-in. Greg Blonder (1996) was the first to describe graphical passwords, presenting in a United States patent a system that would allow users to choose a picture, the number of regions to be clicked, and their sizes and positions. Since then, many variations of this system were presented and images have gained their way into the authentication processes.

Among the most popular graphical authentication systems, we find Passfaces™ from the Passfaces Corporation (2005), a commercial system where the user chooses a previously selected face from a set of faces and repeats this process for different faces in different sets for a defined number of times. However, being popular does not imply being secure, and a study of the users' choices demonstrated that they are, in some cases, similar for all users. For instance, 10% of the passwords of males could have been guessed with only two attempts (Davies, Monrose, & Reiter, 2004).

The déjà vu scheme involves a matrix of *m* images in a set, where *n* images are part of the user's portfolio, previously chosen from a set of proposed images. The user must identify those *n* images to log in.

The draw-a-secret (DAS) scheme is a graphical authentication system with an approach completely different. In DAS, the user draws something over a grid that becomes the

authentication secret. This system has been implemented with success on PDAs (personal digital assistants) and further studies will be made to analyse the users' choices and acceptance (Jermyn, Mayer, Monrose, Reiter, & Rubin, 1999).

In the visual identification protocol (VIP) several possibilities were created. From a set of 10 predefined images the user chooses 4 placed on the same position and typed in the same order (VIP1), or placed in random positions (VIP2). VIP3 is a process where four of the eight images existing in the user's portfolio are displayed along with 12 distractors, and the user must identify them in no particular order. The studies showed that the most common errors associated with VIP1 and VIP2 were related to bad sequences, where the identified images are correct but selected in the wrong order, and in VIP3 most of the errors were due to the wrong identification of the images, for instance, any flower being considered as the chosen flower (de Angeli, Coventry, Johnson, & Coutts, 2003).

In 2006, Magalhães et al. presented a graphical authentication system that included letters and numbers in images with the objective of allowing PDA users to click their user names in an easy way. From that system they discovered that the selection of the image and the rules that control the choice of passgraphs are critical factors in the success of the implementation of this kind of system. In particular, they found that users have a common tendency to choose the first available images, and that the use of images with corners and the existence of letters placed in a row create serious vulnerabilities to the system. Eyes are also a common choice and should be avoided. Therefore, the results suggest the use of images without corners, like nature images, cut in a round form. If the choice of keeping the letters is made, they must be placed in a random way throughout the images. Another dangerous tendency is the use of passgraphs constituted by regions placed in the same row or in the same column, therefore the system must reject the choice of passgraphs that meet this criteria, forcing the users to navigate inside the image by demanding the use of at least two different rows and two different columns.

ENHANCEMENT OF LOG-IN AND PASSWORD SYSTEMS

Since most of the existing systems trust passwords to provide access control and considering that passwords are not enough, we propose the enhancement of this process by adding a new module to the authentication system. This module gives two options to the user: a password with biometric control (keystroke dynamics) or a passgraph. If the user chooses the password system, he or she will be prompted (Figure 2) to enter the password several times in order to establish a pattern (this is called the enrollment process), and the everyday authentication process is, from the user's point of view, exactly the same as it was before the introduction of our module.

Each time that a user enters a password for authentication, the window captures both the characters stroked and the times between successive actions (pressing a key, releasing the key, pressing another key, and so on). The module verifies if the sequence of times matches the stored pattern (locally or in a portable device, like a smart card) and if (and only if) it does, the sequence of characters is sent to the original password authentication system that will verify correctness and allow or deny access to the user. Therefore, we have introduced another layer of security (biometrics) without any extra effort or equipment.

If the user chooses to use a passgraph, avoiding the biometrics component, he or she will have to choose several positions in an image. These positions, clicked in the same sequence, will be the secret access code of that user. Figure 3 shows a possible authentication window with a place to choose one of several possible images (in this case, the choice Mozart is the one that is active) and a place to enter the user name (in this case, the student's identification number). Nevertheless, this image would not be a good choice since it is not compliant with the best procedures in image selection for authentication, as described before.

Each time a user enters a passgraph, the sequence of clicks is transformed by a unidirectional function into a complex string that is passed into the original password field of the hosting system. In this way, we have obtained a simple and easy-to-memorize way of having extremely complex passwords.

As a last remark, one should notice that passgraph-area technology is very vulnerable to eavesdropping and, therefore, are more suitable for access made in private environ-

Figure 2. Keystroke-dynamics enrollment window

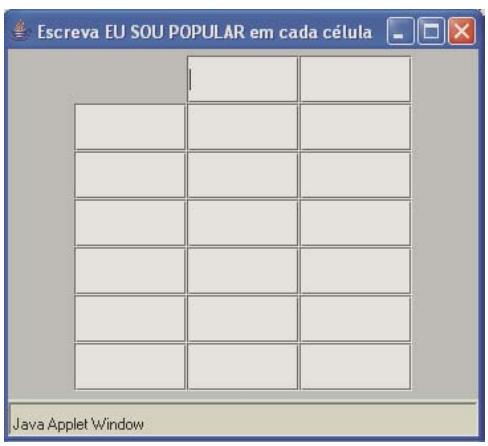
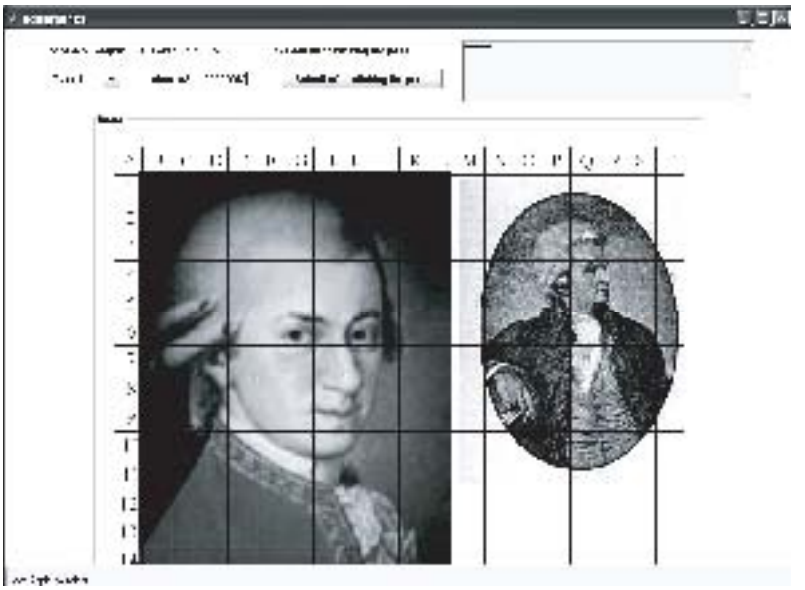


Figure 3. Passgraph authentication window



ments or on small portable devices, like smart phones or PDAs.

FUTURE TRENDS

Future work in this field will focus on improving the algorithms for keystroke dynamics, namely by combining the recent results provided by the artificial intelligence systems with the existing statistical algorithms.

Concerning passgraphs, studies are needed to improve the algorithms that convert the sequence of clicks into a sequence of characters and to improve the quality of the guidelines for image selection. This technology can also be integrated with other information security technologies in order to maximize their potential. On the other hand, artificial intelligence techniques can also be used to understand further more the use of secret codes in order to improve the quality of the proposed systems.

CONCLUSION

In conclusion, we can say that the technology has achieved a way to overcome, at least to a certain point, the entropy generated by users that continues to proceed in a way that is not the most efficient concerning security. Assuming human behaviour as a fact, ways were found to achieve best practices in security (like complex passwords) from the normal and traditional practices of users.

We have verified that keystroke dynamics and graphical authentication systems can, when used together, improve the security of the traditional log-in and password systems without adding significant complexity to their use and avoiding the ethical problems generated by biometrics when they are presented not as a choice but as an imposition. In fact, not only do these systems not present any ethical problems (when used together and leaving to the user the choice of which one to use), they can even provide a good use of the digital authentication processes by allowing those that cannot read or write (and therefore cannot use a password) to use the system, a matter especially relevant in the third-world countries that are now embracing new technologies, for instance, in electoral processes.

REFERENCES

Blonder, G. E. (1996). *Graphical password.*

Davies, D., Monrose, F., & Reiter, M. K. (2004). *On user choice in graphical password schemes.* Paper presented at the 13th USENIX Security Symposium.

de Angeli, A., Coventry, L., Johnson, G. I., & Coutts, M. (2003). Usability and user authentication: Pictorial passwords vs. PIN. In P. T. McCabe (Ed.), *Contemporary ergonomics 2003* (pp. 253-258). London: Taylor & Francis.

Gaines, R., et al. (1980). *Authentication by keystroke timing: Some preliminary results* (Rand Report No. R-256-NSF). Rand Corp.

International Biometric Group (IBG). (2003). *The biometric industry: One year after 9/11*. Retrieved November 2004 from http://www.biometricgroup.com/reports/public/reports/9-11.html

Jermyn, I., Mayer, A., Monrose, F., Reiter, M. K., & Rubin, A. (1999). *The design and analysis of graphical passwords.* Paper presented at the Eighth USENIX Security Symposium, Washington.

Joyce, R., & Gupta, G. (1990). Identity authorization based on keystroke latencies. *Communications of the ACM*, 33(2), 168-176.

Magalhães, S. T., Revett, K., & Santos, H. D. (2005). *Password secured sites: Stepping forward with keystroke dynamics.* Paper presented at the IEEE International Conference on Next Generation Web Services Practices (NweSP'05), Los Alamitos, CA.

Magalhães, S. T., Revett, K., & Santos, H. D. (2006). *Critical aspects in authentication graphic keys.* Paper presented at the International Conference on I-Warfare and Security, MD.

Magalhães, S. T., & Santos, H. D. (2005). An improved statistical keystroke dynamics algorithm. In *Proceedings of the IADIS MCCSIS 2005.*

Monrose, F., & Rubin, A. D. (1997). Authentication via keystroke dynamics. In *Proceedings of the Fourth ACM Conference on Computer and Communication Security*, Zurich, Switzerland.

Monrose, F., & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. *Future Generation Computing Systems (FGCS) Journal: Security on the Web.*

Monrose, F., et al. (2001). Password hardening based on keystroke dynamics. *International Journal of Information Security*.

Ord, T., & Furnell, S. M. (2000). User authentication for keypad-based devices using keystroke analysis. In *Proceedings of the Second International Network Conference: INC 2000*, Plymouth, United Kingdom.

Passfaces Corporation. (2005). *The science behind Passfaces.* Retrieved September 2005 from http://www.passfaces.com

Privacy International, Statewatch, & European Digital Rights. (2004). *An open letter to the ICAO: A second report on "Towards an International Infrastructure for Surveillance of Movement."* Retrieved from http://www.privacyinternational.org

Wiedenbeck, S., Waters, J., Birget, J. C., Brodskiy, A., & Memon, N. (2005, July 25-27). *Authentication using graphical passwords: Basic results.* Paper presented at Human-Computer Interaction International (HCII 2005), Las Vegas, NV.

KEY TERMS

Authentication: It is the process of verifying the identity alleged by a user who tries to gain access to a system.

Collaborative Biometric Technology: It is a biometric authentication technology that requires the user's voluntary and intended participation in the process. It opposes the stealth biometric technologies that can be used without the user's consent.

Crossover Error Rate (CER): Authentication algorithms need to simultaneously minimize permeability to intruders and maximize the comfort level, therefore they have to be both demanding and permissive. This contradiction is the base for the optimisation problem in authentication algorithms, and the measure of success for the overall precision of an algorithm and its usability is the CER, the value obtained at the threshold that provides the same false-acceptance rate and false-rejection rate.

False-Acceptance Rate (FAR): This rate is a measure of the permeability of an authentication algorithm. It is calculated by dividing the number of the intruder's successful log-in attempts by the total number of the intruder's log-in attempts.

False-Rejection Rate (FRR): This rate is a measure of the comfort level of an authentication algorithm. It is calculated by dividing the number of unsuccessful attempts made by legitimate users by the total number of legitimate log-in attempts.

Graphical Authentication System: It is a log-in system that verifies the user's knowledge of specific images or parts of images to grant or deny successful log-in.

Identification: It is the process of discovering the identity of a user who tries to gain access to a system. It differs from authentication because in the identification process, no identity is proposed to the system, while in authentication, an identity is proposed and the system will only verify if that identity is plausible.

Keystroke Dynamics: It is a biometrical authentication algorithm that tries do define a user's typing pattern and then verifies in each log-in attempt if the pattern existing in the way the password was typed matches the user's known pattern. Another application of keystroke dynamics, at least in theory, is the permanent monitoring of the user's typing pattern in order to permanently verify if the user that is typing is the legitimate owner of the system's account being used.

Passgraph: It is the user's secret code to access a system protected by a graphical authentication system. It is constituted by a sequence of points the user must click in order to obtain a successful log-in.

Stealth Biometric Technology: It is a biometric authentication technology that can be used without the user's consent. It opposes the collaborative biometric technologies that require the user's voluntary and intended participation in the process.

Threshold: It is the variable that defines the level of tolerance of an algorithm. It can be set to a more demanding value, raising the false-rejection rate and lowering the false-acceptance rate, or it can be set to a less demanding value, lowering the false-rejection rate and raising the false-acceptance rate.

Knowledge Architecture and Knowledge Flows

Piergiuseppe Morone
University of Foggia, Italy

Richard Taylor
Stockholm Environment Institute, UK

INTRODUCTION

Modern society is increasingly seen as a knowledge economy; institutions, firms and individuals progressively rely on knowledge as a key component for individual and collective growth. This calls for a clear understanding of knowledge and its sharing patterns. This article has a two-fold aim: on the one hand, it aims at reviewing some of the most common *definitions of knowledge* provided in the economic and science and technology literature; on the other hand, it aims at providing a taxonomy of *knowledge flows* which should help scholars in distinguishing among various forms of knowledge sharing. Subsequently, we shall present a description of future trends and put forward some possible extensions of knowledge literature. Finally, our concluding remarks will be presented in the last section of the article.

BACKGROUND

The growing information flow which characterises the so-called "information society" has made organisations increasingly concerned with the problem of selecting and organising information in a cost-efficient manner. However, it would be incorrect to refer to the learning activity simply as the accumulation of information. In fact, firms are increasingly concerned with the acquisition of knowledge which, as recognised by many scholars (see among many others: Foray, 2004; Steinmueller, 2002), differs substantially from information.

Knowledge and Information

This leads us to the core distinction between information and knowledge. Ancori, Bureth, and Chohendet observed how the classical approach of economics adopts a vision that "allows the reduction of knowledge to information, or more precisely allows knowledge to be considered a stock accumulated from interaction with an information flux" (2000, p. 259). However, this view has recently come under criticism as knowledge and information should be considered as two distinct concepts: the latter taking the form of structured data which can be easily transferred through physical supports, and the former involving cognition (see e.g., Tsoukas, 2005; Steinmueller, 2002). To clarify this distinction, we could analyse the differences between the reproduction processes of knowledge and information: While cost of reproducing information amounts solely to the physical cost of making a copy (e.g., the cost of a photocopy, the cost of duplicating an electronic file), the cost of reproducing knowledge is much higher as it involves a cognitive process required to disarticulate knowledge, transfer it to someone else, and rearticulate it for further use (Foray, 2004). Hence, reproducing knowledge involves an intellectual activity, whereas reproducing information simply involves duplication.

Tacit and Codified Knowledge

After having assessed the existence of a clear distinction between information and knowledge, we shall now turn our attention to the definition of knowledge itself. As mentioned above, knowledge has to be articulated in order to be transferred. This is because knowledge is, in its original form, completely embedded in the mind of the person who first developed it. In other words, we could say that knowledge is originally created as tacit and subsequently codified by means of a cognitive process which involves its articulation.

Before reasoning on the codification process, we need to better clarify what is tacit knowledge. The tacit dimension of knowledge corresponds, in the view of Polanyi (1967), to the form or component of human knowledge distinct from, but complementary to, the knowledge explicit in conscious cognitive processes. In the Hungarian polymath view, we know more than we can tell, where the portion of knowledge possessed and not communicable is the essence of tacitness.

In different moments in time and across different individuals, a different proportion of knowledge will be tacit and a different proportion will be codified. Hence, tacitness is a contextual rather than an absolute situation, this depending explicitly on the process of codification, which should be seen as a convergence process of tacit to codified knowledge. Cowan and Foray noted how "as the new knowledge ages, it goes through a process whereby it becomes more codified. As it is explored, used and better understood […] more of it is transformed into some systematic form that can be communicated at low cost" (1997, p. 595).

The relevance of codification for economic purposes has been largely debated. The core argument put forward is that codified knowledge, when compared to tacit, can be transferred more easily, more quickly, and at lower costs. Cowan, David, and Foray (2000) argued in favour of codification stating that an uncodifiable (unarticulable) knowledge is not very interesting for social science. This stance is criticised by Johnson, Lundvall, and Lorenz (2002) who contest the view that codification always represents progress. According to these authors, tacit knowledge is a relevant component in human training, including the kind of training provided in institutions such as schools, universities and research institutes.

Knowledge Flows: Tacit vs. Codified

This argument (Johnson et al., 2002) introduces a key point for us in the debate: Tacit and codified knowledge flow in very different ways. Specifically, once codified, knowledge can be stored in a mechanical or technological way, like in manuals, textbooks or digital supports; it can be transferred from one person to another relatively easily, incurring the effort of getting access to the source of codified knowledge and decoding it for further use. In this respect, as observed by Steinmueller (2000), the context and intended recipient of the decodified knowledge makes a great deal of difference to the costs and feasibility of the initial codification. However, if appropriately codified (i.e., codified keeping in mind the intended recipient), knowledge can be easily transferred, taking also great advantage of modern information and communication technologies.

On the contrary, "[d]ifferent methods like apprenticeship, direct interaction, networking and action learning that include face-to-face social interaction and practical experiences are more suitable for supporting the sharing of tacit knowledge" (Haldin-Herrgard, 2000). Haldin-Herrgard identifies five main difficulties associated with tacit knowledge flows, related to perception, language, time, value, and distance. Perception refers to the characteristic of unconsciousness which entails a problem of people not being aware of the full range of their knowledge; difficulties with language lie in the fact that tacit knowledge is held in a nonverbal form and hence involves extra efforts to be shared; the time issue refers to the fact that the internalization of tacit knowledge takes a long time as it involves direct experience and reflection on these experiences; value is a problem as many forms of tacit knowledge, like intuition and rule-of-thumb, have not been considered valuable, lacking the status of "indisputable methods;" finally, the issue of distance relates to the need for face-to-face interaction for the diffusion of tacit knowledge.

This last point brings us back to the tacit/codified distinction: As already observed, modern information technology can play a major role in diffusing codified knowledge, but tacitness is hard to diffuse technologically. Perhaps, as observed by Haldin-Herrgard (2000), today and in the future high technology will facilitate this diffusion in artificial face-to-face interaction, through different forms of meetings in real-time, using, for instance, audio and video conferences. This perspective is shared by other scholars; in a recent paper Brökel and Binder stated, for instance, that "[n]ew information technologies, for example, video conferences, cast doubt on the advantages of face-to-face contacts" (2007, p. 154).

PROPOSING A TAXONOMY OF KNOWLEDGE FLOWS

The discussed distinction between tacit and codified knowledge is at the heart of the problem of understanding knowledge flows. However, in our view, the existing literature has neglected to classify the different ways in which knowledge can flow among agents. This has created some confusion and has generated a misuse of specific concepts. In this section, we propose a taxonomy of knowledge flows which should help in clarifying the different forms of flow patterns.

Knowledge Gain vs. Knowledge Diffusion

We start our analysis distinguishing between the two broad concepts of *knowledge gain* and *knowledge diffusion*. The first relates, in our view, solely to those processes of knowledge flows which deliberately involve a barter among subjects: A portion of subject's A knowledge flows to subject B, who pays subject A back either with a portion of his or her knowledge or with a different coin.

We shall refer to the first of these two options (i.e., knowledge is paid back with other knowledge) as *knowledge exchange*, and to the second option (i.e., knowledge is paid back with a different coin) as *knowledge trade*. An example of knowledge exchange has been used by Cowan and Jonard who define a model in which knowledge flows "through barter exchange among pairs of agents" (2004, p. 1558). Patterns of knowledge trade, on the other hand, relate, for instance, to those cases where disembodied knowledge flows through technology and patent trade (Arora, Fosfuri, & Gambardella, 2002).

Note that knowledge gain relates to both tacit and codified knowledge. Codified knowledge can flow among distant agents, whereas tacit knowledge gains require always a direct interaction (i.e., face-to-face) among agents.

Substantially different is the concept of knowledge diffusion. Here, knowledge is no longer traded on a voluntary basis (*quid pro quo*), but freely flows while agents interact. Several scholars have referred to this process as knowledge

spillover (Jaffe, 1986), or knowledge percolation (Antonelli, 1996). The common idea behind these definitions is that knowledge flows freely, within a specific space, and can be economically exploited by the recipient agent. The kind of knowledge being spilled-over is tacit in nature, and requires some "absorptive capacity" to be effectively recombined in the cognitive framework of the recipient agent.

Decomposing Knowledge Diffusion

We shall now look more carefully into knowledge diffusion, decomposing it into knowledge spillover, knowledge transfer and knowledge integration. The latter of these three concepts refers to a process which combines dispersed bits of knowledge held by individuals to be applied in a coordinated way, and only on a temporary base. On the contrary, knowledge spillover and knowledge transfer denote two similar processes in which bits of knowledge convey from one agent to another such that the recipient can absorb it into her/his already existent personal knowledge (i.e., some previously acquired related knowledge is required); the only difference between these two processes being that spillover are unintended processes of knowledge diffusion (e.g., while chatting with colleagues), whereas knowledge transfer requires a defined will (e.g., while jointly working on a project).

Knowledge transfers and knowledge spillovers are the most cited typologies of knowledge diffusion patterns (see, for instance, Cabrera & Cabrera, 2002; Morone & Taylor, 2004; van der Bij, Song, & Weggeman, 2003). However, these mechanisms present some disadvantages: They are expensive and often time-consuming and they off-set the specialisation of employees needed for innovation, as it assumes that individuals absorb diverse specialised knowledge by means of face-to-face encounters. In fact, here we are posing a question of depth of knowledge vs. breadth of knowledge. As suggested by Grant, "[d]ue to cognitive limits of human brain, [tacit] knowledge is acquired in a highly specialised form […]. However, production […] requires a wide array of knowledge, usually through combining the specialised knowledge of a number of individuals" (1996, p. 377). The possibility to integrate knowledge without having to acquire it might provide a solution to these drawbacks. In light of these arguments, Grant (1996) asserts that integration of specialist knowledge is at the heart of production in a knowledge-based society.

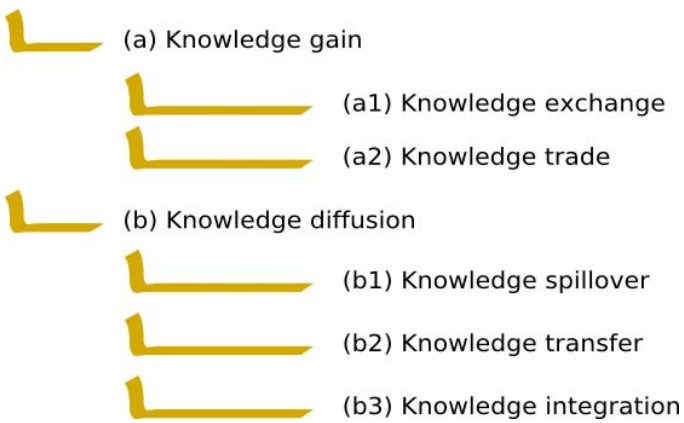
But how does integration occur? In a recent paper Berends, Debackere, Garud, and Weggeman (2004) examined knowledge integration in an industrial context. They defined knowledge integration as dominated by *thinking along*, that is, a mechanism through which an agent applies knowledge temporarily to a problem of somebody else and communicates the generated ideas to that other person. Hence, it involves temporary cognitive work with regard to a problem of someone else.

Interestingly, the concept of knowledge integration does not involve any permanent flow of knowledge from subject A to subject B in the conventional sense. We consider it as an "atypical" form of knowledge diffusion.

Now, recombining the analysis developed in this section, we shall propose a taxonomy of knowledge flows.

Figure 1 shows a taxonomy of concepts emerging from analysis of the knowledge flows literature. At the top level in the hierarchy are knowledge gain and knowledge diffusion, which we classify as distinct phenomena of flows. Knowledge exchange and trade are subclasses of knowledge gain, whereas knowledge spillover, transfer and integration are derived from a decomposition of knowledge diffusion.

Figure 1. A proposed taxonomy of knowledge flows



Assessing Knowledge Flows Taxonomy

It is important to assess the goodness and limitations of the proposed scheme. The following figure introduces a check-list evaluation of the taxonomy as a general framework for understanding knowledge flows.

The first four points should be evident from earlier sections of this article: it is grounded in previous studies and tries to integrate them; it can be used for comparing previous studies on knowledge flows and, specifically, for understanding knowledge flows in information systems. The model has been developed in a flexible (open) way and further insights could always be included. Our analysis also highlights different assumptions about governance and control of knowledge. In the case of knowledge gain, one assumes the functionality, as well as the ability, of locking flows in a rigidly controlled domain of knowledge. The strategy is to maximise the payoff of current knowledge assets and obtain a fair value in exchange. The drawback of this approach is that over the long term it tends to stifle creativity and diminish diversity in production of new knowledge and recombination of existent knowledge. The opposite strategy is the promotion of largely uncontrolled diffusion, where value is often derived from the outcomes on a larger scale: the generation and exploitation of whole new economic areas, and the impact this has on the opportunities and constraints for the organisation.

In spite of the discussion of such stylised facts, the taxonomy does not fully consider the implications of knowledge flows either for economic or innovation systems (points 7 and 8). With respect to empirical validation of the taxonomy, this point is addressed in the following section concerning measurement of knowledge flows.

FUTURE TRENDS

A Challenge for Future Research: Modeling Knowledge Flows

In the area of knowledge flows several different types of modeling have been used. Conceptual modeling ranging from organisational models to taxonomic models (such as the one presented above) are found. Mathematical modeling can be used to determine solution states and optimization behaviours. On the other hand, simulations are promising tools with which to investigate knowledge flows because they can express the dynamics in a model.

The role of formal modeling and simulation is to allow exploration of the hypotheses embodied in the program over a range of different conditions. Models can be, to a greater or lesser degree, based on empirical data on knowledge flows. Although measurement is often problematic (see the next section), efforts to improve the empirical basis of modeling are key to the increasing sophistication of recent knowledge diffusion models, to improving the clarity of conceptual models and to helping the modeler to arrive at a more rigorous conceptualisation.

Figure 2. A check-list for assessing the taxonomy

CHECK-LIST	"CASTING OUT NINES"	Yes	No
	1. integrates different models or studies	✓	
	2. allows to compare and contrast different models or studies	✓	
	3. has a clear structure	✓	
	4. is suitable for understanding knowledge flows in information systems	✓	
	5. could be modified or extended as new information emerges	✓	
	6. is suitable for evaluating knowledge governance and control	✓	
	7. fully covers the implications for innovation systems		✓
	8. fully covers the economics of knowledge		✓
	9. is empirically valid		✓

Measuring Knowledge Flows and Validating the Proposed Taxonomy

In order to validate and develop the proposed taxonomy, it is necessary to select suitable methods to observe and measure knowledge flows. Measurement is a key issue in assessing theoretical models, and in doing so we are concerned with distinguishing between measures of knowledge gains and measures of knowledge diffusion.

Measurement can be categorised according to the context in which an investigation is carried out. We shall define three possible investigation contexts within which data on knowledge flows can be gathered: (a) literature-based, (b) field-based, and (c) laboratory-based.

Since the seminal contribution of Jaffe (1986), many economists attempted to measure knowledge flows tracing the citation flows across patents (see, among others, Alcácer & Gittelman, 2006). Specifically, this strand of empirical literature aims at measuring "the benefits that one inventor receives from the innovations of others" (Fung & Chow, 2002, p. 353). Other researchers have investigated international flows of knowledge as measured through papers' citation (on bibliometric studies see, for instance, Sivadas & Johnson, 2005) or as measured in scientific meetings, using proceedings citations (see, for instance, Godin, 1998). These are all measures of knowledge diffusion; specifically, the first two refer to knowledge transfer (in fact, scientific papers and patents reflect the intention of transferring knowledge), whereas the last measure could incorporate pure knowledge spillovers as well.

The common ground of these studies is that they concentrate their attention on citations as a key empirical indicator of knowledge flows. However, not all knowledge flows through citations. For instance, citations refer solely to codified knowledge, hence dismissing all sources of tacit knowledge flows. A possible solution to this problem is offered by field-based research which, through field work and case studies, could better capture other sources of flows and would allow to clearly distinguish among knowledge gain, knowledge diffusion and their subcategories (see, for instance, Morone, Sisto, & Taylor, 2006).

A further step in this direction was made by Berends et al. (2004). The authors chose an ethnographic research strategy which combines interviews with community members with close observation of their work practices in their natural context. Hence, by means of such studies, researchers can analyse knowledge processes, in their natural context, as they are actively realised. Note that it was exactly through this empirical investigation that it was possible to define the abovementioned concept of knowledge integration.

The third approach involves laboratory experiments which might provide further insight on the actual processes of knowledge flows. This is a rather new research strategy, based on the analysis of agents' behaviour in a laboratory artificial environment. The main advantage of this approach is that it allows to track exactly who is interacting with whom and how knowledge flows from one agent to another (see Morone, Morone, & Taylor, 2007). One shortcoming of this methodology relates to the fact that knowledge structure is *ex-ante* determined and remains static through the experiment; moreover, knowledge flows fall in a predefined category (e.g., knowledge transfer or knowledge integration) which is imposed upon players. This results in a predetermined space of knowledge which bounds possible flows and does not allow classifying different categories of flows.

CONCLUSION

This article presented a review of recent studies of knowledge flows with relevance to economic and science and technology literature, arriving at some definitions of terms in the knowledge economy field. The main finding was that much of this discussion was based on the distinction between knowledge gain and knowledge diffusion, resulting from different assumptions about governance and control of knowledge.

This finding leads directly to the main contribution of the article presented in section three: A new taxonomy of knowledge flows, in which we expand on the concept of knowledge diffusion and highlight a further decomposition which, we hope, should help in distinguishing among various forms of knowledge sharing.

REFERENCES

Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *Review of Economic and Statistics, 88*(4), 774-779.

Ancori, B., Bureth, A., & Chohendet, P. (2000). The economics of knowledge: The debate about codification and tacit knowledge. *Industrial and Corporate Change, 9*(2), 255-287.

Antonelli, C. (1996). Localized knowledge percolation processes and information networks. *Journal of Evolutionary Economics, 6*(3), 281-295.

Arora, A., Fosfuri, A., & Gambardella, A. (2002). *Markets for technology: The economics of innovation and corporate strategy*. Cambridge, MA: MIT Press.

Berends, J.J., Debackere, K., Garud, R., & Weggeman, M. (2004). *Knowledge integration by thinking along*. Eindhoven Centre for Innovation Studies (Working paper 04.05).

Brenner, T. (2007). Local knowledge resources and knowledge flows. *Industry and Innovation, 14*(2), 121-128.

Brökel, T., & Binder, M. (2007). The regional dimension of knowledge transfers —a behavioral approach. *Industry and Innovation, 14*(2), 151-175.

Cabrera, A., & Cabrera, E.F. (2002). Knowledge-sharing dilemmas. *Organization Studies, 23*, 687-710.

Cowan, R., David, P.A., & Foray, D. (2000). The explicit economics of knowledge codification and tacitness. *Industrial and Corporate Change, 9*, 211-253.

Cowan, R., & Foray, D. (1997). The economics of codification and the diffusion of knowledge. *Industrial and Corporate Change, 6*, 595-622.

Cowan, R., & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics & Control, 28*, 1557-1575.

Foray, D. (2004). *Economics of knowledge*. Cambridge, MA: MIT Press.

Fung, M.K., & Chow, W.W. (2002). Measuring the intensity of knowledge flow with patent statistics. *Economics Letters, 74*, 353-358.

Godin, B. (1998). Measuring knowledge flows between countries: The use of scientific meeting data. *Scientometrics, 42*(3), 313-323.

Grant, R.M. (1996). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science, 7*, 375-387.

Haldin-Herrgard, T. (2000). Difficulties in diffusion of tacit knowledge in organisations. *Journal of Intellectual Capital, 1*(4), 357-365.

Jaffe, A. (1986). Technological opportunity and spillovers of R&D: Evidence from firms patents, profits, and market value. *American Economic Review, 76*, 984-1001.

Johnson, B.H., Lundvall, B., & Lorenz, E. (2002). Why all this fuss about codified and tacit knowledge? *Industrial and Corporate Change, 11*(2), 245-262.

Lundvall, B.Å., & Johnson, B. (1994). The learning economy. *Journal of Industry Studies, 1*(2), 23-42.

Morone, A., Morone, P., & Taylor, R. (2007). A laboratory experiment of knowledge diffusion dynamics. In U. Cantner & F. Malerba (Eds.), *Innovation, industrial dynamics and structural transformation.* Berlin: Springer.

Morone, P., Sisto, R., & Taylor, R. (2006). Knowledge diffusion and networking in the organic production sector: A case study. *EuroChoices, 5*(3), 40-46.

Morone, P., & Taylor, R. (2004). Knowledge diffusion dynamics of face-to-face interactions. *Journal of Evolutionary Economics, 14*, 327-351.

Polanyi, M. (1967). *The tacit dimension*. London: Routledge.

Sivadas, E., & Johnson, M.S. (2005). Knowledge flows in marketing: An analysis of journal article references and citations. *Marketing Theory, 5*(4), 339-361.

Steinmueller, E.W. (2000). Will new information and communication technologies improve the "codification" of knowledge?. *Industrial and Corporate Change, 9*(2), 361-376.

Steinmueller, E.W. (2002). Knowledge-based economies and information and communication technologies. *International Social Science Journal, 54*(171), 141-153.

Tsoukas, H. (2005). *Complex knowledge: Studies in organizational epistemology*. Oxford: Oxford University Press.

van der Bij, H., Song, X.M., & Weggeman, M. (2003). An empirical investigation into the antecedents of knowledge dissemination at the strategic business unit level. *Journal of Product Innovation Management, 20*, 163-179.

KEY TERMS

Codified Knowledge: Knowledge that has converged upon common concepts and usages such that it can be transferred more easily.

Knowledge Diffusion: A situation where knowledge flows freely, within a specific space, and can be economically exploited by the recipient agent.

Knowledge Gain: A process of knowledge flow which involves deliberate barter among subjects.

Knowledge Integration: A process which combines dispersed bits of knowledge held by individuals to be applied in a coordinated way.

Tacit Knowledge: Knowledge that is embedded in the mind of the person who has acquired it.

Taxonomy of Knowledge Flows: A conceptual model which attempts to distinguish among various forms of knowledge sharing.

Knowledge Combination vs. Meta-Learning

Ivan Bruha
McMaster University, Canada

INTRODUCTION

Research in intelligent information systems investigates the possibilities of enhancing their over-all performance, particularly their prediction accuracy and time complexity. One such discipline, data mining (DM), processes usually very large databases in a profound and robust way (Fayyad et al., 1996). DM points to the overall process of determining a useful knowledge from databases, that is, extracting high-level knowledge from low-level data in the context of large databases. This article discusses two newer directions in this field, namely knowledge combination and meta-learning (Vilalta & Drissi, 2002).

There exist approaches to combine various paradigms into one robust (hybrid, multistrategy) system which utilizes the advantages of each subsystem and tries to eliminate their drawbacks. There is a general belief that integrating results obtained from multiple lower-level decision-making systems, each usually (but not required) based on a different paradigm, produce better performance. Such multi-level knowledge-based systems are usually referred to as *knowledge integration* systems. One subset of these systems is called *knowledge combination* (Fan et al., 1996). We focus on a common topology of the knowledge combination strategy with base learners and base classifiers (Bruha, 2004).

Meta-learning investigates how learning systems may improve their performance through experience in order to become flexible. Its goal is to search dynamically for the best learning strategy. We define the fundamental characteristics of the meta-learning such as bias, and hypothesis space.

Section 2 surveys the various directions in algorithms and topologies utilized in knowledge combination and meta-learning. Section 3 represents the main focus of this article: description of knowledge combination techniques, meta-learning, and a particular application including the corresponding flow charts. The last section presents the future trends in these topics.

BACKGROUND

So far, commonly utilized decision-making systems have been exploiting a single technique, strategy, or topology. Consequently, their accuracy and overall performance have not been so high (Pratt & Thrun, 1997). New data mining (DM) systems utilize results obtained from several lower-level systems, each usually (but not required) based on different paradigm, or combine or refine them within a dynamic process. Thus, such a multi-strategy (hybrid) system consists of two or more individual 'agents' that interchange information and cooperate together.

It should be noted that there are in fact two fundamental approaches for combining the information from multi-data tasks:

1. In *data combination*, the data sets are merged into a single set before the actual knowledge acquisition.
2. In *knowledge* (*theory*) *combination*, or *sensor fusion*, several agents (base classifiers, sensors) process each input data set separately, and the induced models (knowledge bases) are then combined at the higher-level.

When we look at the issue of the multi-strategy systems from the other side, we come to the *meta-learning*. Generally speaking, meta-learning investigates the way the learning systems can increase their performance and efficiency over experience.

The base learners, the ones with a simple inductive paradigm, such as algorithms inducing decision trees or decision sets of rules, or neural nets, generate a hypothesis (concept description) by applying a fixed bias that is implanted in the knowledge base of the learner. The performance usually increases by larger training sets and losing the restrictions on the hypotheses (concept descriptions).

Using other words, a meta-learner searches dynamically for the best learning strategy and consequently, its performance is flexible. There are a few strategies of the meta-learning, however, various researches recognize it in various ways so that one cannot specify exactly which strategy belongs to meta-learning and which not (Vilalta & Drissi, 2002). Also, there is no sharp boundary between knowledge combination and meta-learning; some researches on machine learning (ML) and DM claim that the first is the subset of the latter, some not. Therefore, this article introduces the most common sights to this issue.

Another taxonomy of these systems distinguishes the way of arrangement of datasets and learning paradigms. We thus differentiate:

1. **Different subsets of training data with a single learning paradigm:** Different subsets are either generated when they are collected, or a single (usually larger) database is split to several subsets, following a certain criterion. Each base learner (with the same learning technique) processes different training subset. Typical examples of such a technique is bagging (Breiman, 1996) and boosting (Freund & Schapire, 1997).
2. **Different training parameters with a single learning paradigm:** Each learning algorithm is accompanied by various parameters that have to be setup. We can thus generate several base learners by changing these parameters, and use then the entire dataset for all the base learners, see for example Bruha (2004).
3. **Different learning paradigms:** The entire multi-strategy system consists of several base learners, each with different learning paradigm (learning system inducing decision trees, that inducing set of decision sets, artificial neural net, genetic algorithm, etc.). These base learners then can process the same database (Kotsiantis & Pintelas, 2004; LiMin et al., 2004).

It should be also noted that there is no uniform terminology in the knowledge-intensive systems (including DM, machine learning, and meta-learning); therefore, we use here usually not a single but several most common terms that can be found in literature.

KNOWLEDGE COMBINATION AND META-LEARNING

Knowledge Combination

A large research in ML focuses on improving topology of classifiers by combining various paradigms into one multi-strategy (hybrid) system which utilizes the advantages of each subsystem and tries to eliminate their drawbacks. There is a general belief that integrating results obtained from multiple lower-lever classifiers produce better performance. We can consider the boosting and bagging algorithms (Bauer & Kohavi, 1999) as already traditional topologies of this approach.

Generally speaking, the main advantages of such hybrid systems are: better performance than that of individual lower-level agents included, the ability to process multivariate data from different information sources, and better understanding of internal data processing when a complex task is solved.

Multi-level knowledge based techniques (called *knowledge integration* systems) can be divided into the following three 'subtechniques:'

1. **Knowledge combination/selection:** The input to such a system is usually formed by several knowledge bases (models) that are generated by various DM algorithms (learners). Each model (knowledge base) independently produces its decision about prediction. These results are then combined into a final decision (knowledge *combination*) or the best decision is selected according to a given statistical criterion (knowledge *selection*).
2. **Knowledge merging:** Several models (knowledge bases) are merged into one robust, usually redundant, model by utilizing statistics that accompany these models.
3. **Knowledge modification (also called revision, refining):** The input is an existing 'old' knowledge base and a 'new' database. A DM algorithm revises (modifies, refines) the current knowledge base according to the knowledge which is 'hidden' in the 'new' database. The new knowledge base thus gets over the 'old' knowledge by being updated by knowledge extracted from the 'new' database.

The first project in this field is evidently (Brazdil & Torgo, 1990); their system merges several decision trees generated by ID3 into a robust one. The already mentioned bagging and boosting algorithms can be viewed as representatives of multi-models. Another direction is formed by the system XCS that is a mixture of genetic algorithms and neural nets (Wilson, 1999). There are several extensions of this system, for example, NXCS (Armano et al., 2002). Another hybrid multisystem combines genetic algorithms with decision trees (Carvalho & Freitas, 2000). All these research projects have revealed that knowledge combination improves the performance of the base classifiers. Knowledge modification is quite often utilized in Inductive logic programming (ILP); they usually use the term 'theory refinement' (Haddawy et al., 2003).

(Fan et al., 1996) introduce the methodology of stacked generalizers and meta-combiners. It can be viewed as learning from information generated by a set of base learners, or using other words, as learning of meta-knowledge on the learned information. The base learners (each usually utilizing a different inductive strategy) induce base classifiers; the base classifiers applied to a training set of examples form so-called meta-database; it is then used by the meta-learner to derive a meta-classifier. The two-level structure of classifiers is then used for making decisions about the input objects.

There are many interesting issues in this field, for example, combining statistical/fuzzy data (probability distribution of classes, quality of decision/performance, reliability of each base classifier), cascade classifiers (Gama & Brazdil, 2000).

Application: Meta-Combiner

Here we present one particular topology and fundamental configuration of a knowledge combination system, a *meta-combiner*. It consists of a few different *base learners*, each utilizing a different inductive strategy. Each base learner generates a *base classifier* by processing a subset of training data. Afterwards, the base classifiers classify another subset of training objects, and their decisions (outputs) form a *meta-database*, a set of training meta-objects (examples). Afterwards, it is then processed by a *meta-learner* that then generates a *meta-classifier*.

It should be noted that the meta-classifier does not exploit the traditional 'select best' strategy (i.e., by selecting the best base classifier), nor by 'by vote' strategy (i.e., by selecting the class the majority of base classifiers predict). It rather combines the decisions (predictions, classes) of all the base classifiers.

Consequently, in the classification phase, the base classifiers first derive their predictions (classes, decisions); then a meta-object is derived from these predictions, which is then classified by the meta-classifier. The entire scheme of the meta-learner and the meta-classifier is called *meta-combiner*.

A training set is split for the meta-learning purposes into two subsets: the *genuine-training* and *examining* ones. The genuine-training subset is applied for inducing the base classifiers; the examining one for generating a meta-database.

We now specify the format of a meta-database. Let BCl_q be the q-th base classifier, q=1...,Q (where Q is the number of the base classifiers). Each examining example **x** of the examining subset generates corresponding meta-object **z** for the meta-database as follows. Let z_q be the decision (class) of the q-th base classifier for the examining object **x** (we call z_q a *meta*-attribute to distinguish them from *input* attributes A_n). Then the corresponding meta-object of the meta-database looks as follows:

$$\mathbf{z} = [z_1, \ldots , z_Q , Z]$$

where z_q , q=1...,Q is the decision the of q-th base classifier BCl_q , Z is the desired class of the examining object **x** .

Let T be a training set of K training examples, S be an integer in the range <2; K> . Let us have Q different base learners BL_q , q=1,...,Q, and a meta-learner ML. The flow chart of the meta-learning phase looks as follows:

procedure META-LEARNING-PHASE(T, S)

1. Partition the training set T randomly into S disjoint subsets of equal size (as equal as possible). Let T_s be the s-th such subset, s=1,...,S, card(T_s) the number of its objects (examples)
2. Form S pairs [T_s , $T \setminus T_s$] , s=1...,S ; for each s, $T \setminus T_s$ is the genuine-training subset and T_s the examining one, generated from the training set T .
3. Let *MetaDatabase* be empty
4. **For** s=1,...,S **do**
   - 4.1 Train all base learners BL_q using the genuine-training subset $T \setminus T_s$; the result is Q base classifiers BCl_q , q=1,...,Q
   - 4.2 Classify the examining objects from T_s by these base classifiers
   - 4.3 Generate card(T_s) meta-objects using the above class-combiner rule and add them to *MetaDatabase*

   enddo
5. Train the meta-learner ML using *MetaDatabase*; the result is a meta-classifier *MCl**
6. Generate the base classifiers BCl_q* , q=1,...,Q again but use the entire training set T (these classifiers are marked by star and will be used in the meta-classification)

Step 4.1 is depicted by Figure 1. The entire meta-learning phase (i.e., steps 4.2, 4.3, and 5) is portrayed on Figure 2.

Similar scenario is applied for the classifying an unseen object **x** :

procedure META-CLASSIFYING-PHASE(**x**)

1. Classify the unseen object **x** by all Q base classifiers BCl_q* ; let the output of the q-th base classifier BCl_q* be z_q , q=1,...,Q
2. Generate the corresponding meta-object **z** = [z_1, ..., z_Q]
3. Classify the meta-object by the meta-classifier *MCl**; its result (decision) is the class to which the given input object **x** is classified

We can now observe that the base classifiers BCl_q* and the meta-classifier *MCl** form the overall product that can be immediately used by an end-user for classifying unseen objects, see Figure 3.

The number S of split subsets is crucial for this system. Therefore, we use the term *S-fold meta-combiner*. (Fan et al., 1996) introduces two architectures: combiner and stacked generalizer. Their combiner corresponds to the two-fold and stacked generalized to the K-fold meta-combiner. The 'foldness' S of the meta-combiner is a decisive parameter of the entire system. We empirically observed that the value S=4 gives good results (Bruha, 2004).

Bruha (2004) also deals with inconsistency of the meta-databases; it can happen quite often there exists more identical meta-objects having different desired classes. The project uses

Figure 1. Base learners using the genuine-training subset $T \setminus T_s$ induce Q base classifiers

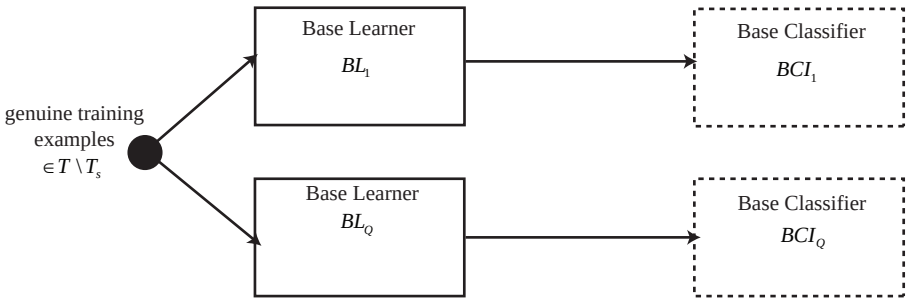


Figure 2. The entire meta-learning phase

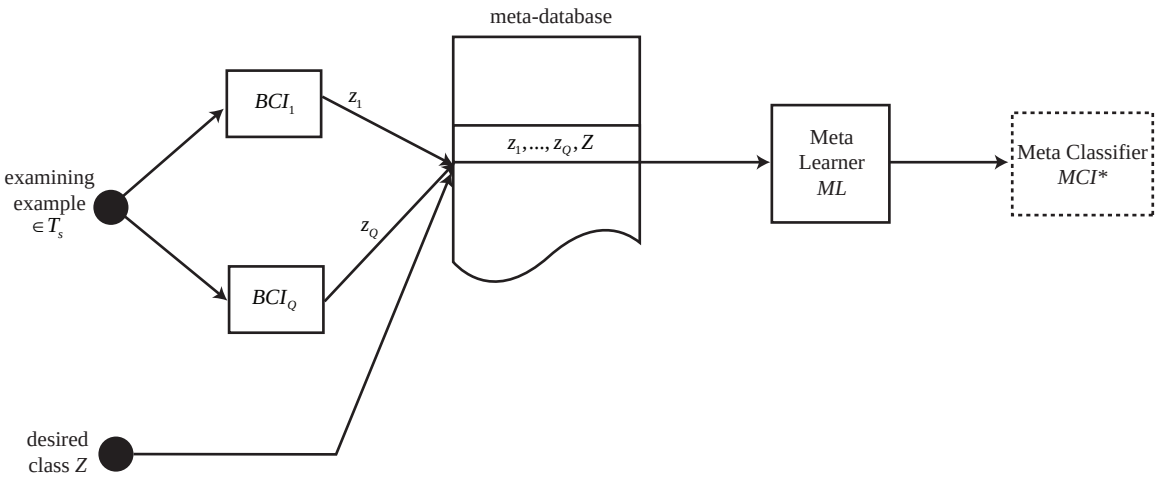
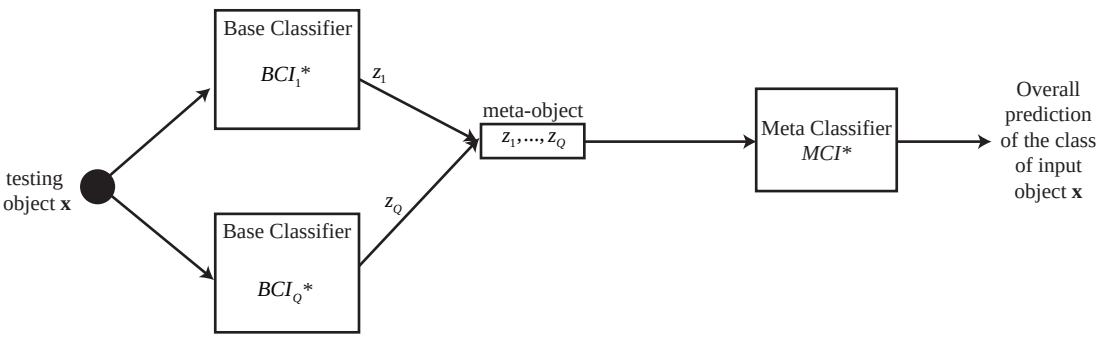


Figure 3. Meta-classifying phase



'purification' procedure that adds as many informative input attributes until the meta-database becomes consistent.

Meta-Learning

First, we introduce some definitions. A learning algorithm (learner) is learned by a training set of pre-classified examples [**x**, *Z*] where **x** is a training object (example) and *Z* its desired class (specified by a Designer of the entire learning task). The mapping DAC(**x**) = *Z* is an unknown designer-assigned classification. Since a training set is always finite and need not be consistent, the learning algorithm (learner) will not be able to induce a genuine form of the mapping DAC, but only its approximation. It conducts a search through a *hypothesis space* until it finds a hypothesis that approximates (according to a given criterion) the unknown DAC . A learning algorithm exhibits a set of assumptions (*bias*) that restrict the inductive process. First, it restricts the size of the hypothesis space, and second, it orders (ranks) the hypotheses in the hypothesis space.

A traditional (base) learner exhibits a fixed hypothesis space. Thus, it can learn efficiently only a limited number of learning tasks. Therefore, the previous meta-combiner is not recognized by many ML researches as a meta-learning system because it exhibits a fixed form of bias.

Vilalta and Drissi (2002) claim that the goal of the meta-learning is to choose the right learning algorithm for a particular learning task. A meta-learning system should indicate how to match learning algorithms with the properties of a learning task, that is, it should dynamically select a learning algorithm. Furthermore, a meta-learning system should be able to solve the problems of the learning tasks placed outside the scope of available learning algorithms.

One profound feature of meta-learning is to specify the ways of how to associate a learning algorithm with those learning tasks in which the learner works in an optimal fashion. Usually, a designer defines a set of learning tasks' characteristics (*meta-features*) that are essential to the learner's behavior. By utilizing these meta-features we can induce the conditions under which a learning algorithm works better than the others. For instance, the meta-features are the distribution of training examples among concepts (classes), number of examples, number of attributes, number of classes, and other statistics above a training set.

We can observe even further steps of exploiting these techniques: instead of selecting a learning algorithm for a particular domain of learning tasks, we can select a learning algorithm for each testing object; we assume that a learning algorithm exhibits a reasonably prestigious behavior around the neighborhood of the given testing object. Brazdil and Soares (2000) discuss the system in which learning algorithms utilizing the k-nearest neighbor strategy are ordered by the accuracy and the time complexity.

Meta-learning systems can also utilize the dynamic selection of bias for shifting a learner's applicability along a set of the learning tasks. The goal is to modify the hypothesis space to have better coverage of the given learning task. One possibility is to modify the size of the hypothesis space. Another possibility is to change the hypothesis space by adding or deleting the input attributes that represent the objects (examples). Adding new attributes increases the size of the hypothesis space and, thus, weakens the bias; and vice versa. Bruha (2000) utilizes the quality of decision rules for the actual classification; it updates the rule qualities within the testing.

Humans learn for the entire life by experience. This idea could be utilized also in machine learning. If a learning system wants to improve its performance, it should gather the knowledge about learning (*meta-knowledge*). Thrun (1998) and others discuss the idea of *learning-to-learn*. This strategy studies how to improve learner's behavior by detecting the characteristics among various learning tasks. Baum (1998) presents the idea of learning agents; learning agents collaborate and generate new learning agents.

Many other enhancements and applications of meta-learning can be found in Brazdil, Soares and Da Costa, (2003) and Todorovski and Dzeroski (2003).

FUTURE TRENDS

There are several ways of further research in these new disciplines of ML and DM. Let us survey the most promising trends:

- Knowledge combination can be extended to *higher-level* combining methods. For instance, (Fan et al., 2002) investigate classifier ensembles and multilevel tree structured combining methods. Bruha (2006) explores three-level combiners; the first two are formed by the meta-combiner discussed above, and the third level utilizes the decision of this meta-combiner and other classification systems, such as averaging, regression, best decision scenario, voting scenario, dynamic selective voting, naive Bayesian combination.
- *Meta features* that specify the way how a learner could be associated with a learning task is another direction in future work. They could be selected by a decision-making system that would choose the most optimal ones. Thus, we come to a strategy with two different levels of decision-making algorithms.
- *Learning-to-learn* is another direction of future work. Still, there is no general strategy how to improve learner's behavior by detecting the characteristics among various learning tasks. It will require to study

the characteristics of meta-knowledge, that is, collecting knowledge about learning.

- The idea of *learning agents* is also a promising trend. They can independently learn how a given learning algorithm processes a certain task and make up a decision for future which algorithms would be utilized for a new task.
- Vilalta and Drissi (2002) present a promising trend for meta-learning. It proposes the idea of *self-adaptive learning* algorithms. They are supposed to change their internal topology (bias) according to characteristics of a given learning task. Such a learner would accomplish it by continuous accumulation of meta-knowledge that would consist of suitable forms of bias for each learning tasks. At the beginning, possessing no meta-knowledge, a self-adaptive learner would use fixed bias, and as more learning tasks were analyzed and processed, the learner would accumulate and utilize the meta-knowledge. Consequently, it looks like a life-long learning process.
- Embedding *genetic algorithms* and similar optimization tools for generating more robust decision-making and knowledge-combining systems is another direction in this field. Particularly, genetic algorithms seem to be a very powerful and robust technique for inducing reliable and consistent knowledge bases (models). They can serve at both basic level and meta level as we discussed.
- So far, there was only a limited invention of the quality of base learners (or more general, low-level decision-making systems) that generate their decisions to a higher-level system. *Redundancy* of the base learners seems to be one of the solutions, as it complies with the human life. (We also have our knowledge stored in several 'pockets').
- Gama and Brazdil (2000) introduce their cascade algorithm. To our opinion, it is another direction of knowledge combination techniques. It combines statistical and logical techniques together in order to reach a better performance of the entire system.

CONCLUSION

This article discusses two relatively new directions in ML and DM research: knowledge combination and meta-learning. We have also presented one particular learner (meta-combiner) in details as an application in order to exhibit the complexity and topology of such learners.

We have also mentioned that different researchers in ML have different view about knowledge combination and meta-learning. There exists a common opinion that meta-learning algorithms exhibit flexible bias and dynamic search in hypothesis space, however, there is no sharp boundary between the above ML directions (Vilalta & Drissi, 2002). The article surveys several known topologies of meta-learners, too. A large source of papers on meta-learning and knowledge combination can be found on the internet.

REFERENCES

Armano. G., Murru, A., & Roli, F. (2002). Stock market prediction by a mixture of genetic-neural experts. *International J. Pattern Recognition and Artificial Intelligence*, *16*(5), 501-526.

Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, *36*, 105-142.

Baum, E.B. (1998). Manifesto for an Evolutionary Economics of Intelligence. *Neural Networks and Machine Learning*, (pp. 285-344). Springer-Verlag.

Brazdil, P., & Soares, C. (2000). Ranking Classification Algorithms Based on Relevant Performance Information. *ECML-2000, Workshop on Meta-Learning*, Barcelona.

Brazdil, P., & Torgo, L. (1990). Knowledge Acquisition via Knowledge Integration. In B. Wielinga (Eds.), *Current trends in knowledge acquisition*, (pp. 90-104). IOS Press.

Brazdil, P., Soares, C., & Da Costa, J.P. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning*, *50(*3), 251-277.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140.

Bruha, I. (2000). A feedback loop for refining rule qualities in a classifier: A reward-penalty strategy. *European Conference on Machine Learning (ECML-2000), Workshop Meta-Learning*, (pp. 15-27).

Bruha, I. (2004). Meta-learner for unknown attribute values processing: Dealing with inconsistency of meta-databases. *J. Intelligent Information Systems*, *22*(1), 71-84.

Bruha, I. (2006). Three-level tree-structured meta-combiner: A case study. Submitted to *Intelligent Data Analysis*, in press.

Carvalho, D.R., & Freitas, A.A. (2000). A hybrid decision tree/genetic algorithm for coping with the problem of small disjuncts in data mining. In *Proceedings of the Genetic and Evolutionary Computation (GECCO-2000)*, (pp. 1061-1068).

Fan, D.W., Chan, P.K., & Stolfo, S.J. (1996). A comparative evaluation of combiner and stacked generalization. *AAAI-96, Workshop Integrating multiple learning models*.

Fan, D.W., Wang, H., Yu, P.S., Lo, S., & Stolfa, S.J. (2002). Progressive Modelling. *2nd IEEE International Conference Data Mining (ICDM-2002)*.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 37-53.

Freund, Y., & Schapire, R. (1997). A decision-theoretical generalization of online learning and an application to boosting. *JCSS, 55*, 119-139.

Gama, J., & Brazdil, P. (2000). Cascade generalization. *Machine Learning, 41*(3), 315-343.

Haddawy, P., Ha, V., Restificar, A., Geisler, B., & Miyamoto, J. (2003). Preference elicitation via theory refinement. *J. Machine Learning Research*, *4*, 317-337.

Kotsiantis, S., & Pintelas, P. (2004). Selective voting. In *Proceedings of the 4th International Conference on Intelligent System Design and Applications*, Budapest, (pp. 397-402).

LiMin, W., SenMiao, Y., Ling, L., & HaiJun, L. (2004). Improving the performance of decision tree: A hybrid approach. *Lecture Notes in Computer Science, 3288,* 327-335.

Pratt, L., & Thrun, S. (1997). Second special issue on inductive transfer. *Machine Learning*, 28.

Thrun, S. (1998). Lifelong learning algorithms. *Learning to Learn* (pp. 181-209). Kluwer Academic Publ.,.

Todorovski, L., & Dzeroski, S. (2003). Combining classifiers with meta decision trees. *Machine Learning*, *50*(3), 223-249.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *J. Artificial Intelligence Review*, *18*(2), 77-95.

Wilson, S.W. (1999). Get real: XCS with continuous-valued inputs. In L. Booker, S. Forrest, M. Mitchell, & A. Riolo (Eds.),*Festschrift in honor of J.H. Holland* (pp. 111-121). Univ Michigan,.

KEY TERMS

Classifier: A decision-supporting system that given an unseen (to-be-classified) input object yields a prediction, for instance, it classifies the given object to a certain class.

Knowledge Combination: Its input is usually formed by several knowledge bases (models) that are generated by various Data Mining algorithms (learners). Each model (knowledge base) independently produces its decision about prediction; these results are then combined into a final decision—or the best decision is selected according to a given criterion.

Learner: Given a training set of (representative) examples (accompanied usually by their desired classes/concepts), a Learner induces concept description (model, knowledge base) for a given task.

Meta-Combiner: Its common topology involves base learners and classifiers at the first level, and meta-learner and meta-classifier at the second level. The meta-classifier combines the decisions of all the base classifiers.

Meta-Knowledge: Generally speaking, it is a knowledge about knowledge bases. In case of the meta-learning, it is a knowledge about the learning process(es) that is accumulated by a meta-learning algorithm within analyzing and processing various learning tasks.

Meta-Learning: It investigates how learning systems may improve their performance through experience in order to become flexible. Its goal is to search dynamically for the best learning strategy.

Model (knowledge base): Formally described concept of a certain problem; usually represented by a set of production rules, decision trees, semantic nets, frames.

Category: Bioinformatics

Knowledge Discovery from Genomics Microarrays

Lei Yu
Binghamton University, USA

INTRODUCTION

The advent of genomic microarray technology enables simultaneously measuring the expressions of thousands of genes in massive experiments, and hence provides scientists, for the first time, the opportunity of observing complex relationships between various genes in a genome. In order to extract biologically meaningful insights from a plethora of data generated from microarray experiments, knowledge discovery techniques, which discover patterns, statistical or predictive models, and relationships among massive data, have been widely applied in microarray data analysis. For example, clustering can be applied to identify groups of genes that are regulated in a similar manner under a number of experimental conditions or groups of samples that show similar expression patterns across a number of genes (Jiang, Tang, & Zhang, 2004). Classification can be performed to characterize the cellular difference between different samples, such as between normal and cancer cells or between cancer cells with different responses to treatment, and can potentially be used to predict the classes of samples based on their gene expression patterns (Statnikov, Aliferis, Tsamardinos, Hardin, & Levy, 2005). Feature selection or gene selection can help identify among thousands of genes a small fraction of genes that are relevant for discriminating between different sample types, and may potentially lead to the identification of a few biologically relevant "marker" genes for subsequent biological validation (Saeys, Inza, & Larranaga, 2007).

This article provides a brief introduction to the field of knowledge discovery and its applications in discovering useful knowledge from genomic microarray data. It describes common knowledge discovery tasks for genomic microarray data, presents representative methods for each task, and identifies emerging challenges and trends in knowledge discovery from genomic microarray data.

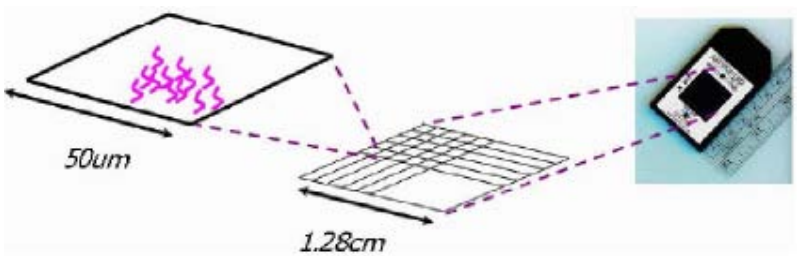
BACKGROUND

Knowledge discovery from data refers to the overall process of converting raw data into useful information, which consists of data preprocessing, data mining, and postprocessing of data mining results (Tan, Steinbach, & Kumar, 2005). The purpose of data preprocessing is to transform the raw input data into an appropriate format for subsequent data mining process. The tasks involved in data preprocessing include cleaning data to remove noise, duplicate, inconsistent, or missing information, integrating data from multiple sources, transforming data values into the right scale and format, and selecting instances and features that are relevant to the data mining task at hand. Data mining is the automatic process of extracting interesting patterns or knowledge from data. Data mining tasks can be generally divided into two major categories: predictive tasks and descriptive tasks. The objective of predictive tasks is to predict the values of a particular feature, based on the values of other features. The objective of descriptive tasks is to derive patterns that summarize the underlying relationships in data. These tasks are often exploratory in nature and frequently require postprocessing techniques which validate and explain the results. For example, visualizing the patterns allows analysts to explore the result from multiple viewpoints. Statistical tests can be used to validate the significance of the results and eliminate patterns that are generated by chance. For a comprehensive discussion on various knowledge discovery tasks, please refer to widely adopted text books on data mining (Han & Kamber, 2005; Tan et al., 2005). This article focuses on knowledge discovery tasks that are commonly performed on genomic microarray data introduced next.

Gene expression microarrays are silicon chips that simultaneously measure the mRNA expression levels of thousands of genes. Different types of microarrays use different technologies for constructing these chips and measuring gene expression levels. Detailed description of these technologies is beyond the scope of this article. Interested readers can refer to Draghici's book (Draghici, 2003) for an introduction. Here we briefly describe microarray technologies using Affymetrix arrays (shown in Figure 1) as an example, which are currently one of the most popular commercial arrays. However, the methodology for constructing other types of arrays would be similar, but would use different technology-specific data preparation and cleaning steps (Piatetsky-Shapiro & Tamayo, 2003).

Each Affymetrix array (GeneChip) contains probes for different genes tiled in a grid-like fashion. The simultaneous measure of expression levels of thousands of genes is done by hybridizing a complex mixture of mRNAs (derived from tissue or cells) to the probes. Hybridization events

Figure 1. An example of microarray: Affymetrix GeneChip (right), its grid (center), and a cell in the grip (left). Image courtesy of Affymetrix



are detected using a fluorescent dye and a scanner that can detect fluorescence intensities. The scanner and associated software perform various forms of image analysis to measure and report raw gene expression values. This allows for a quantitative readout of gene expression on a gene-by-gene basis (Piatetsky-Shapiro & Tamayo, 2003). To date, one microarray chip is capable of measuring expression levels for over 40,000 genes in the entire human genome.

The expression level of a specific gene among thousands of genes measured in an experiment is eventually recorded as a numerical value. Expression levels of the same set of genes under study are normally accumulated through multiple experiments on different samples (or the same sample under different conditions) and recorded in a data matrix. In data mining, data is often stored in the form of a matrix, of which each column is described by a feature or attribute and, each row consists of feature-values and forms an instance, also named as a record or data point, in a multidimensional space defined by the features. Figure 2 illustrates two ways of representing microarray data in a matrix form. In Figure 2 (a), each feature is a sample (*S)* and each instance is a gene (*G*). For each gene, its expression levels are measured across all the samples (or conditions), so f_{ij} is the measurement of the expression level of the i^{th} gene for the j^{th} sample where $i = 1,..., n$ and $j = 1,..., m$. In Figure 2 (b), the data matrix is the transpose of the one in Figure 2 (a), in which features are genes and instances are samples. Sometimes, data in Figure 2 (b) may have class labels c_i for each instance, represented in the last column. The class labels can be different cellular conditions of the underlying samples. A typical microarray data set may contain thousands of genes but only a small number of samples (often less than a hundred). The number of samples is likely to remain small at least for the near future due to the expense of collecting microarray samples.

Different data mining tasks can be performed on the two different forms of data shown in Figure 2. When genes are treated as instances (as in Figure 2 (a)), *gene clustering* can be performed to find similarly expressed genes across many samples. When samples are treated as instances (as in Figure 2 (b)), three different tasks can be performed: *sample clustering* which aims to group similar samples together and discover classes or subclasses of samples, *sample classification* which aims to classify diseases or phenotypes of novel samples based on patterns learned from training

Figure 2. Two views of a microarry data matrix

$$\begin{array}{cccccc} & S_1 & S_2 & \cdots & S_m \\ G_1 & f_{11} & f_{12} & \cdots & f_{1m} \\ G_2 & f_{21} & f_{22} & \cdots & f_{2m} \\ \vdots & & & & \vdots \\ G_n & f_{n1} & f_{n2} & \cdots & f_{nm} \end{array} \qquad \begin{array}{ccccccc} & G_1 & G_2 & \cdots & G_n & \\ S_1 & f_{11} & f_{21} & \cdots & f_{n1} & c_1 \\ S_2 & f_{12} & f_{22} & \cdots & f_{n2} & c_2 \\ \vdots & & & & \vdots & \vdots \\ S_m & f_{1m} & f_{2m} & \cdots & f_{nm} & c_m \end{array}$$

samples with known class labels, and *gene selection* which aims to select a small number of discriminative genes from thousands of genes.

KNOWLEDGE DISCOVERY FROM MICROARRAY DATA: TASKS AND METHODS

In this part, we present common knowledge discovery tasks from microarray data and representative methods for each task. These tasks include two descriptive data mining tasks: gene clustering and sample clustering, one predictive task, sample classification, and one preprocess task, feature selection.

Clustering

Clustering is a process of grouping similar samples, objects, or instances into clusters under some similarity measure (Tan et al., 2005). Clustering has proven to be helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. On one hand, identifying groups of genes with similar expression patterns across different samples (also called co-expressed genes) may help understand the functions of many genes for which information has not been previously available and give rise to hypotheses regarding the mechanism of the transcriptional regulatory network (Jiang et al., 2004). On the other hand, identifying groups of samples with similar expression patterns across a number of genes may reveal subcell types which are hard to identify by traditional morphology-based approaches (Jiang et al., 2004). We present three groups of frequently used clustering methods which can be applied for both gene clustering and sample clustering on microarray data.

Hierarchical clustering Each instance forms a cluster in the beginning, and the two most similar clusters are merged until all instances are in one single cluster. The clustering result is in the form of a tree structure, called dendrogram, which can be broken at different levels using domain knowledge. Tree structures are easy to understand and can reveal close relationships among resulting clusters, but they do not provide a unique partition among all the instances because different ways to determine a basic level in the dendrogram can result in different clustering results.

Partition-based clustering Unlike hierarchical clustering methods, partition-based methods divide the whole data into a fixed number of clusters. Examples are *K*-means, self-organizing maps, and graph-based partitioning (Jiang et al., 2004). The methods of *K*-means often require specification of the number of clusters *K* and the selection of *K* instances as the initial clusters. All instances are then partitioned into the *K* clusters, optimizing some objective function (e.g., inner-cluster similarity) by assigning each instance to the most similar cluster determined by the distance between the instance and the mean of each cluster in the current iteration. Self-organizing maps (SOMs) are variations of *K*-means methods and require specification of the initial topology of *K* nodes to construct the map. In graph-based partitioning methods, a Minimum Spanning Tree (MST) is often constructed and the clusters are generated by deleting the MST edges with the largest lengths. Graph-based partitioning methods do not heavily depend on the regularity of the geometric shape of cluster boundaries as *K*-means and SOMs do.

Fuzzy clustering Traditional clustering methods introduced before require that each instance belong to a single cluster, even though some instances may only be slightly relevant for the biological significance of their assigned clusters. Fuzzy *C*-means (Dembele & Kastner, 2003) applies a fuzzy partitioning method that assigns cluster membership values to instances. It links each instance to all clusters via a real-valued vector of indexes. The value of each index lies between 0 and 1, where a value close to 1 indicates a strong association to the corresponding cluster while a value close to 0 indicates no association. The vector of indexes thus defines the membership of an instance with respect to the various clusters.

Classification

Apart from clustering methods which do not require *a priori* knowledge about the classes of available instances, a classification method requires training instances with labeled classes, leans patterns that discriminate between various classes, and, ideally, correctly predict the classes of unseen instances. Many classification methods can be applied to predict diseases or phenotypes of novel samples from microarray data. We present four commonly used methods (Statnikov et al., 2005; Tan et al., 2005).

Linear discriminative analysis (LDA) For an m x n gene expression matrix $\boldsymbol{X}$ (m is the number of samples and n is the number of genes), it seeks linear combinations xa of sample vectors $x_i = (x_{i1},\ldots,x_{in})$ with large ratios of between-class to within-class sum of squares. In other words, it tries to maximize the ratio a^TBa/a^TWa, where B and W denote respectively the n x n matrices of between-class and within-class sum of squares.

Nearest neighbor (NN) usually does not learn during the training phase. Only when it is required to classify a new sample does NN search the data to find the nearest neighbor for the new sample and use the class label of the nearest neighbor to predict the class label of the new sample. *K*-NN makes the prediction for a new sample based on the most common class label among the *K* training samples most similar to the new sample.

Decision trees classify samples by building a tree-like structure. Specifically, they recursively split samples into two child branches based on the values of a selected feature, starting with all the samples. Each leaf node of the tree is pure in terms of classes, and the resulting partition corresponds to a classifier. By limiting the number of consecutive branches, they can produce more generalized classifiers.

Support vector machines (SVMs) try to separate a set of training samples of two different classes with a hyperplane in an *n*-dimensional space defined by *n* features (genes). If no separating hyperplane exists in the original space, a kernel function is used to map the samples into a higher-dimensional space where a separating hyperplane exists. Complex kernel functions that provide nonlinear mappings result in nonlinear classifiers. SVMs avoid overfitting by selecting a hyperplane that is maximally distant from the training samples of two different classes, called maximum margin separating hyperplane, from among many hyperplanes that can separate the two classes.

Gene Selection

Traditional classification and clustering methods are often designed for data where the number of instances is greatly larger than the number of features. In microarray data, however, the number of features (genes) is very huge and the number of instances (samples) is relatively small for sample classification and sample clustering tasks. The characteristic of high dimensionality (thousands of features) in microrray data presents a great challenge to the efficiency of many classification and clustering methods. Moreover, the shortage of instances in the context of high dimensionality often causes many classification methods to generate models that overly fit the known training data or causes many clustering methods to produce clusters purely by chance. Therefore, selecting a small number of genes from thousands of genes is essential for successful sample classification and clustering. Feature selection, the process of choosing an optimal subset of features from original ones according to a certain criterion, is a frequently used data preprocessing technique in knowledge discovery from data with high dimensionality. Liu and Yu (2005) provide a good introduction to the problem of feature selection and various feature selection methods for classification and clustering. Some of these methods have been applied on microarray data as an effective tool for gene selection task (Saeys et al., 2007).

Among various methods, ranking-based methods often evaluate genes in isolation without considering gene-to-gene correlation. They rank genes according to their individual relevance or discriminative power to the targeted class and select top-ranked genes. Some methods based on statistical tests or information gain have been employed (Li, Zhang, & Ogihara, 2004). However, simply combining a highly ranked gene with another highly ranked gene often does not form a good gene set because some highly correlated genes could be redundant to each other. Removing redundant genes among selected ones can achieve a better representation of the characteristics of the targeted class and lead to improved classification accuracy. Representative methods that handle gene redundancy can be found in Peng and Ding (2005) and Yu (2008).

FUTURE TRENDS

Traditional clustering methods introduced previously perform clustering along only one direction, either grouping genes using the set of samples as features or grouping samples using the set of genes as features. For a given data set to be clustered, along either direction, the feature space is globally determined and is shared by all resulting clusters. However, it is well known in molecular biology that only a small subset of the genes participates in any cellular process of interest and that any cellular process takes place only in a subset of the samples (Jiang et al., 2004). Recently, a number of bi-clustering or co-clustering methods have been proposed to capture coherent expression patterns exhibited in "blocks" within gene expression matrices, where a "block" is a submatrix defined by a subset of genes on a subset of samples. These methods aim to simultaneously cluster genes and samples under certain heuristic objective functions. Madeira and Oliveira (2004) provide a review of some recent work in this area.

Various classification methods coupled with feature selection methods have been applied to microarray sample classification and show promising accuracy in predicting the classes (diseases or phenotypes) of novel samples. However, existing classification and feature selection methods are not adequate for identifying leads for potentially useful biomarkers. The primary goal of biologists in conducting microarray experiments is sometimes not to build models for sample classification, but to detect a few biomarkers from many potential candidates for subsequent costly biological and clinical validation. The increasingly known gap between necessary genes for accurate classification models and biologically relevant genes for biomarker identification requires novel knowledge discovery tasks which can generate knowledge of both statistical significance and biological relevance and incorporate domain knowledge of experts (Berens, Liu, Parsons, Yu, & Zhao, 2005).

CONCLUSION

Gene expression microarrays are a revolutionary technology with great potential to provide accurate medical diagnostic, develop cures for diseases, and produce a detailed genome-wide molecular portrait of cellular states (Piatetsky-Shapiro

& Tamayo, 2003). Knowledge discovery methods are effective tools to turn massive raw data from microarray experiments into biologically important insights. In this article, we provide a brief introduction to knowledge discovery and genomic microarray data analysis and a concise review of various knowledge discovery methods for genomic microarray data.

REFERENCES

Berens, M.E., Liu, H., Parsons, L., Yu, L., & Zhao, Z. (2005). Fostering biological relevancy in feature selection for microarray data. *IEEE Intelligent Systems, 20*(6), 29-32.

Dembele, D., & Kastner, P. (2003). Fuzzy C-means method for clustering microarray data. *Bioinformatics, 19*(8), 973-980.

Draghici, S. (2003). *Data analysis tools for DNA microarrays*. Chapman & Hall/CRC.

Han, J., & Kamber, M. (2005). *Data mining: Concepts and techniques* (2nd ed.). Morgan Kaufmann.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering, 16(*11), 1370-1386.

Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics, 20*, 2429-2437.

Liu, H., & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 491-502.

Madeira, S.C., & Oliveira, L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 1*(1), 24-45.

Peng, H., & Ding, C. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology, 3*(2), 185-205.

Piatetsky-Shapiro, G., & Tamayo, P. (2003). Microarray data mining: Facing the challenges. *SIGKDD Explorations, 5*(2), 1-5.

Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics, 23*(19), 2507-2517.

Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., & Levy, S. (2005). A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics, 21*(5), 631-643.

Tan, P., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining.* Addison-Wesley.

Yu, L. (2008). Feature cluster selection for high-throughput data analysis. *International Journal of Data Mining and Bioinformatics, 2*, 1-15.

KEY TERMS

Classification: A process of predicting the classes of unseen instances based on patterns learned from available instances with predefined classes.

Clustering: A process of grouping instances into clusters so that instances are similar to one another within a cluster but dissimilar to instances in other clusters.

Data Mining: the automatic process of extracting interesting patterns or knowledge from data.

Feature Selection: A process of choosing an optimal subset of features from original features according to a certain criterion.

Gene: A hereditary unit consisting of a sequence of DNA that contains all information necessary to produce a molecule that performs some biological function.

Gene Expression Microarrays: Silicon chips that simultaneously measure the expression levels of thousands of genes.

Genome: All of the genetic information or hereditary material possessed by an organism.

Knowledge Discovery: The overall process of converting raw data into useful information, consisting of data preprocessing, data mining, and postprocessing of data mining results.

Knowledge Flow Identification

Oscar M. Rodríguez-Elias
University of Sonora, Mexico

Aurora Vizcaíno
University of Castilla-La Mancha, Spain

Ana I. Martínez-García
CICESE Research Center, Mexico

Jesús Favela
CICESE Research Center, Mexico

Mario Piattini
University of Castilla-La Mancha, Spain

INTRODUCTION

Knowledge management (KM) is an important factor in organizational competitive advantage (Ichijo & Nonaka, 2007). Unfortunately, traditional KM initiatives frequently fail when they are included in the work processes of organizations (Stewart, 2002). One of the factors responsible for this is that these initiatives are not well aligned to the real knowledge needs of the organization's knowledge workers. Thus, it is important to seek approaches to help to align KM initiatives to the real work processes of organizations (Maier & Remus, 2002), considering what is important for their knowledge workers (Dalkir, 2005; Wiig, 2004).

In this chapter, we describe the knowledge flow identification methodology (KoFI), a methodology, based on process engineering techniques, that has been developed to aid in the study of organizational processes from a knowledge flow perspective. The methodology proposes a set of steps and tasks that can be carried out to analyze knowledge flows in business processes; thus, helping to identify issues such as the knowledge workers' needs, the knowledge (and its sources) that is principally involved in the processes, the working tools that may (positively or negatively) affect the flow of knowledge in the process, or the problems that may be restricting the good flow of knowledge in the process. To exemplify the usefulness of the KoFI methodology, we provide a brief description of some of the results obtained from the application of the methodology, in real settings, in which it was helpful for various purposes, including: the design of a multiagent-based KM system, the development of a knowledge map for a process, the identification of the manner in which to integrate a tool currently used in an organization as a basis for a KM strategy, and for the development of an organizational knowledge portal.

BACKGROUND

The integration of KM into organizational processes has been considered one of the most important research approaches for the present and future of KM (Scholl, König, Meyer, & Heisig, 2004). It can be found in literature, some works addressing the integration of KM in organizational processes. Maier and Remus (2002), for instance, have studied different approaches for process-orientated KM strategies, and have developed a framework that is useful for characterizing them. Some other authors have proposed process-modeling approaches for studying the knowledge involved in organizational processes, most of which have been designed to aid in the development of KM systems (e.g., Bera, Nevo, & Wand, 2005; Kim, Hwang, & Suh, 2003; Nissen & Levitt, 2004; Papavassiliou & Mentzas, 2003; Smith & McKeen, 2004; Strohmaier & Tochtermann, 2005; Woitsch & Karagiannis, 2002). Most of the approaches we have found in literature are orientated towards developing specific KM systems, or they require special tools for using them. We have not found work focused on understanding the knowledge requirements of a process rather than on proposing specific solutions. Before proposing a specific approach for managing knowledge in an organization, it is important to analyze the organizations' work processes from a knowledge flow perspective (Nissen, 2002), since supporting knowledge flow should be the main focus of KM (Borghoff & Pareaschi, 1998).

The KoFI methodology, presented in this chapter, uses process-engineering techniques to analyze organizational processes from a knowledge flow perspective. The main differences, between our proposal and others, that we have found in literature, is that our approach focuses mainly upon understanding the flow of knowledge and the problems that affect it, and not upon developing specific types of KM

systems. However, more importantly, our approach takes a special interest in the current technologies that might be involved in the flow of knowledge within an organization, to consider them as a part, and perhaps as the basis, of the KM strategies. This is relevant since, as Davenport (2007) states, the most promising way to integrate KM into organizational processes is to embed KM into the systems that people use to do their jobs. Even in organizations that do not have explicit KM initiatives, employees frequently tend to apply certain KM activities implicitly, and the information systems that they use in their daily work may serve to partially support such activities. We argue that if we base KM strategies upon the work being done by the members of an organization, and use their current tools to do so, then we will be in a better shape to design KM systems that are really useful for organizations.

THE KOFI METHODOLOGY

The KoFI methodology was designed to aid in the analysis of software processes from a knowledge flow perspective (Rodríguez, Martinez-Garcia, Vizcaino, Favela, & Piattini, 2005; Rodríguez-Elias, Martinez-Garcia, Favela, Vizcaino, & Soto, 2007). It was defined to assist in three main areas: 1) to identify, structure, and classify the knowledge that exists in the process studied, 2) to identify the technological infrastructure that supports the process and affects the knowledge flow, and 3) to identify forms with which to improve the knowledge flow in the process.

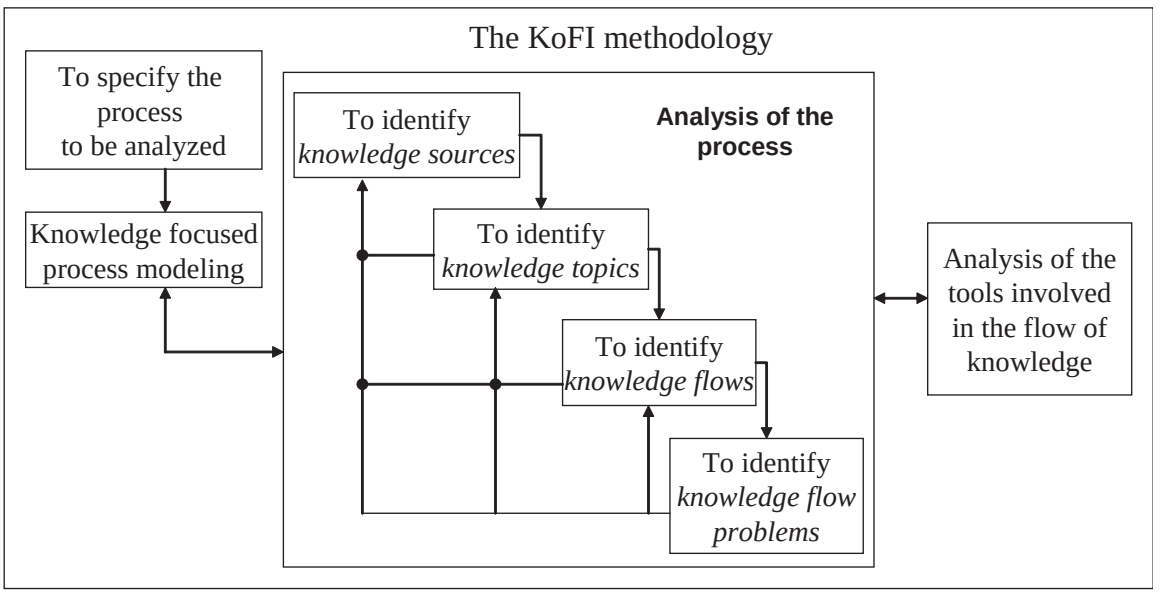
KoFI is orientated towards helping to analyze specific work processes. Therefore, it is necessary to define the specific process and model it. The process models are later analyzed following a four-stage process (see Figure 1) to finally identify and describe the tools that, positively or negatively, affect the flow of knowledge.

The process followed to apply the methodology is iterative, since each stage may provide information useful for the preceding stages. The process models are also capable of evolving while they are being analyzed in the different stages of KoFI. We shall now provide some directions about how each stage can be carried out.

The Knowledge-focused Process Modeling Phase

Traditional process modeling languages (PMLs) can be used to identify issues related to implicit knowledge flows, such as the information sources that are required, generated, or modified by an activity (Abdullah, Bennest, Evans, & Kimble, 2002). However, it is important that a PML used to analyze knowledge flow provides explicit representa-

Figure 1. General view of the KoFI methodology. KoFI has three main phases: knowledge-focused process modeling, analysis of the process (which include identification of knowledge sources, topics, and flows, and knowledge flow problems), and analysis of the tools affecting the knowledge flow.



tion of the knowledge consumed or generated in activities, the knowledge required by the roles participating in those activities, the sources of that knowledge, or knowledge dependencies. Unfortunately, there is a lack of PMLs that focus on the identification of knowledge involved in the processes (Bera et al., 2005). One way to address this situation is to adapt existing PMLs to integrate the representation of knowledge.

Modelling the process at different levels of abstraction is recommended. First, a general view of the process can be defined with a general and flexible process modeling technique. To perform a detailed analysis, a more formally constrained language should be used. It may also be helpful to use a PML that has been designed for the type of process that is being analyzed, since this language should provide primitives to represent specific elements involved in that type of processes, and the explicit representation of those elements will facilitate their analysis. In our case, we have used rich picture (Monk & Howard, 1998) to develop general models of the processes, and the software process engineering metamodel (SPEM) (Object Management Group, 2002) to develop the detailed models. The latter has been chosen due to the fact that our main domain area is software processes. Owing to space limitations, and since the focus of this chapter is not on the modeling languages, further details of this aspect are not presented; detailed examples can be found in Rodríguez-Elias et al. (2007). In this section, we shall limit ourselves to simply presenting the main activities that are carried out in each stage of KoFI.

The Process Analysis Phase

The analysis of the process is carried out in four stages, described in the following paragraphs.

Identifying Knowledge Sources

After modeling the process, the first step is to identify the main documents and people involved in that process. It is important that the sources identified be organized and classified. To this end, a taxonomy can be defined, which is one of the first steps in the development of KMSs (Rao, 2005). An ontology can be also developed to help define the relationships between the sources and the other elements of the process. This ontology can later be used to structure the knowledge base of the process.

Identifying Knowledge Topics

The topics of knowledge that can be obtained from the sources found in the previous stage are defined, together with the knowledge that the people involved in the process may have or require. In this stage, a taxonomy and an ontology can also help to classify the types of knowledge and define their relationships with other elements of the process. The ontology must define means to relate the knowledge sources with the knowledge areas or topics that can be found in them.

Identifying Knowledge Flows

In this stage, the process model is used to identify the way in which the knowledge and sources are involved in the activities performed in the process. The main activities in the processes must be identified, along with the decisions that the people performing those activities have to make. The process models are used to analyze the way the knowledge flows through the process while the people involved perform their activities; for example, the sources that could be being used as knowledge transfer mechanisms between activities or people. It is important to identify either flows of knowledge between activities or between sources, for example, the transfer of knowledge from a person to a document.

Identifying Knowledge Flow Problems

The knowledge flows, identified in the previous stage, are analyzed to discover the problems that may be affecting them. For example, whether the information generated from an activity is not captured, or whether there are sources that might be able to assist in the performance of certain activities, but that are not consulted by the people in charge of them. To do this, we propose using *problem scenarios*, which are stories describing a problem that is taking place in the process being analyzed (Rodríguez-Elias, Martínez-García, Vizcaíno, Favela, & Piattini, 2005). This story must show how the problems detected are affecting the knowledge flow. Then, one or more alternative scenarios must be defined to illustrate possible solutions and the manner in which those alternative solutions may improve the flow of knowledge. These possible solutions are then used as the basis for proposing the KM strategy or system.

The Tool Analysis Phase

The final phase of the methodology focuses on analyzing the manner in which the tools used to support the activities of the process affect the flow of knowledge. To this end, we have used the Rodriguez-Elias, Martínez-García, Vizcaino, Favela, and Piattini framework (2008). This framework proposes four main steps that can be used to analyze information systems as knowledge flow enablers. First, the application domain of the system is defined. This includes identifying the use, scope, and domain of the knowledge managed. The second step consists of identifying the structure of such knowledge. Later, the third step focuses on defining the KM activities being supported by the tool. Finally, the fourth step consists

of the definition of the main technical aspects considered important for the tools. For instance, the level of autonomy of the tool or the distribution of the knowledge managed.

After the application of the methodology, information should have been obtained that is useful, for example, in defining the knowledge base of the process, discovering the problems affecting the flow of knowledge and the mechanisms through which knowledge is flowing, and making proposals to improve the knowledge flow.

APPLYING KOFI

The KoFI methodology has been successfully applied in three case studies. The following paragraphs describe how it provided support in each case.

Designing a multiagent-based KM system. The first use of KoFI was in the analysis of a software maintenance process in which it was helpful in creating a knowledge map of the process that was later used to develop an agent-based KM system to automatically identify knowledge needs and search for knowledge sources that would be useful for solving specific change requests (Rodríguez et al., 2004).

Using a current system as a base for a KM strategy. The second case was focused on identifying the tools being used as knowledge flow enablers, also in a software maintenance process. From this study, the main tool identified was the one used to manage the maintenance requests sent by the clients of the organization studied (Rodríguez-Elias *et al.*, 2008). Some proposals were made for using this tool as basis for a KM strategy. The methodology helped us to identify the relationships between the different sources and the activities of the process, to define the manner in which to link the tool to the other sources of knowledge related to the maintenance requests.

Designing a knowledge portal. The third use of the methodology was to design a knowledge portal for an industrial company. In this case, the methodology was used by one of the members of the company to guide the identification of the main activities carried out in each of the departments of the company, the knowledge required in such activities, and the sources from which that knowledge could be obtained. With this information, a knowledge map was structured, which is now being used as the basis of a knowledge portal being developed by the company.

FUTURE TRENDS

One of the challenges of KM systems is to evaluate their possible benefits before investing heavily in their implementation. Unfortunately, the benefits of KM initiatives emerge in the long run, and are influenced by many factors that are difficult to control and measure. We state that one way to minimize the risk of investing in KM initiatives that are useless for the knowledge workers is to base such initiatives on the real work being done by the people who we expect to benefit from them, and to integrate the KM approaches within the tools in daily use. However, this does not ensure a successful KM approach. Therefore, it is still necessary to develop frameworks to help evaluate the possible usefulness and usability of KM approaches before investing in them. Because of the large set of different types of KM approaches, creating such a framework is a challenge. We consider that integrating such a framework into the KoFI methodology could make a great contribution to the usefulness of our approach. This is just one of the open issues that we should address in our future work.

CONCLUSIONS

Integrating knowledge into every day work is of ever-increasing interest to present-day KM experts. However, accomplishing this for specific organizations is not an easy task. In this chapter, we have presented a methodology that has been successfully used to improve knowledge flows in organizational processes. By taking the knowledge needs of employees into account, we expect to benefit process knowledge management. Moreover, we propose to include the tools used to support daily work as part of the KM strategies. In this chapter, we have summarized three case studies that show examples of how the methodology has been useful. However, it is necessary to perform more studies in different settings if we are to continue evaluating its benefits and limitations.

REFERENCES

Abdullah, M. S., Benest, I., Evans, A., & Kimble, C. (2002). *Knowledge modelling techniques for developing knowledge management systems.* Paper presented at the European Conference on Knowledge Management, Dublin, Ireland.

Bera, P., Nevo, D., & Wand, Y. (2005). Unravelling knowledge requirements through business process analysis. *Communications of the Association for Information Systems, 16*, 814-830.

Borghoff, U. M., & Pareschi, R. (Eds.). (1998). *Information technology for knowledge management*. Berlin, Germany: Springer.

Dalkir, K. (2005). *Knowledge management in theory and practice*. Amsterdam: Elsevier.

Davenport, T. H. (2007). Information technologies for knowledge management. In K. Ichijo & I. Nonaka (Eds.),

Knowledge creation and management: New challenges for managers (pp. 97-117). New York, NY: Oxford University Press.

Ichijo, K., & Nonaka, I. (Eds.). (2007). *Knowledge creation and management: New challenges for managers*. New York, NY: Oxford University Press.

Kim, S., Hwang, H., & Suh, E. (2003). A process-based approach to knowledge flow analysis: A case study of a manufacturing firm. *Knowledge and Process Management, 10*(4), 260-276.

Maier, R., & Remus, U. (2002). Defining process-oriented knowledge management strategies. *Knowledge and Process Management, 9*(2), 103-118.

Monk, A., & Howard, S. (1998). The rich picture: A tool for reasoning about work context. *Interactions, 5*(2), 21-30.

Nissen, M. E. (2002). An extended model of knowledge-flow dynamics. *Communications of the Association for Information Systems, 8*, 251-266.

Nissen, M. E., & Levitt, R. E. (2004). *Agent-based modeling of knowledge flows: Illustration from the domain of information systems design.* Paper presented at the Hawaii International Conference on System Science (HICSS 2004), Big Island, Hi, USA.

Object Management Group. (2002). *Software process engineering metamodel specification (SPEM), Version 1.0*. Retrieved October 29, 2004, from http://www.omg.org/technology/documents/formal/spem.htm

Papavassiliou, G., & Mentzas, G. (2003). Knowledge modelling in weakly-structured business processes. *Journal of Knowledge Management, 7*(2), 18-33.

Rao, M. (Ed.). (2005). *Knowledge management tools and techniques: Practitioners and experts evaluate KM solutions*. Amsterdam: Elsevier.

Rodríguez, O. M., Martínez, A. I., Favela, J., Vizcaíno, A., & Piattini, M. (2004). Understanding and supporting knowledge flows in a community of software developers. *Lecture Notes in Computer Science, 3198*, 52-66.

Rodríguez-Elias, O. M., Martínez-García, A. I., Favela, J., Vizcaíno, A., & Soto, J. P. (2007). *Knowledge flow analysis to identify knowledge needs for the design of knowledge management systems and strategies: A methodological approach.* Paper presented at the 9th International Conference on Enterprise Information Systems (ICEIS): special session on Business Intelligence, Knowledge Management and Knowledge Management Systems, Funchal, Madeira - Portugal.

Rodríguez-Elias, O. M., Martínez-García, A. I., Vizcaíno, A., Favela, J., & Piattini, M. (2005). Identifying knowledge flows in communities of practice. In E. Coakes & S. A. Clarke (Eds.), *Encyclopedia of communities of practice in information and knowledge management* (pp. 210-217). Hershey, PA: Idea Group Inc.

Rodríguez-Elias, O. M., Martínez-García, A. I., Vizcaíno, A., Favela, J., & Piattini, M. (2007.). *Organización de conocimientos en procesos de ingeniería de software por medio de modelado de procesos: Una adaptación de SPEM*. Paper presented at the VI Jornada Iberoamericana de Ingeniería del Software e Ingeniería del Conocimiento (JIISIC'07), Lima, Perú.

Rodríguez-Elias, O. M., Martínez-García, A. I., Vizcaíno, A., Favela, J., & Piattini, M. (2008). A framework to analyze information systems as knowledge flow facilitators. *Information and Software Technology, 50*(6), 481-498.

Scholl, W., König, C., Meyer, B., & Heisig, P. (2004). The future of knowledge management: An international Delphi study. *Journal of Knowledge Management, 8*(2), 19-35.

Smith, H. A., & McKeen, J. D. (2004). Knowledge-enabling business processes. *Communications of the Association for Information Systems, 13*, 25-38.

Stewart, T. A. (2002). The case against knowledge management. *Business 2.0, 3*, 80.

Strohmaier, M., & Tochtermann, K. (2005). B-kide: A framework and a tool for business process-oriented knowledge infrastructure development. *Journal of Knowledge and Process Management, 12*(3), 171-189.

Wiig, K. (2004). *People-focused knowledge management: How effective decision making leads to corporate success*. Amsterdam: Elsevier.

Woitsch, R., & Karagiannis, D. (2002). Process-oriented knowledge management systems based on KM-services: The promote approach. *International Journal of Intelligent Systems in Accounting, Finance & Management, 11*, 253-267.

KEY TERMS

Knowledge Flow: The transfer of knowledge from the place it is created or stored to the place it needs to be applied.

Knowledge Flow Facilitators: All those mechanisms or entities that interfere in the transfer of knowledge within an organization's processes.

Knowledge Management: A discipline focused on providing technologies or techniques to help organizations to store, process, disseminate, and reuse their knowledge in order to take advantage of it.

Knowledge Source: A source from which knowledge, with practical applications can be obtained, such as know how, know what, know where, and so forth.

Knowledge Topic: A definition of a particular area of knowledge that is useful for the members of an organization.

KoFI: Knowledge flow identification methodology

Ontologies: Conceptual models for specifying meanings of, or knowledge about, a common domain.

Problem Scenarios: Textual descriptions of problems observed in a specific situation. They are presented in the form of a story that illustrates the problem and offers possible alternative solutions.

Process Modeling: An activity in which a process is represented through a formalism called the process modeling language, which facilitates its analysis and abstracts the important aspects for the process analyzer.

Knowledge Management as Organizational Strategy

Cheryl D. Edwards-Buckingham
Capella University, USA

INTRODUCTION

"More than ever before, the effectiveness of organizations depends on their ability to address issues such as knowledge management, change management, and capability building, all of which could fall into the domain of the HR function" (Lawler & Mohrman 2003, p. 7). In its leadership role, Human Resources (HR) has many tasks and responsibilities. According to Lawler and Mohrman (2003), there are several key organizational challenges faced by HR departments. These challenges include improving productivity, increasing quality, facilitating mergers and acquisitions, improving new product possibilities, and knowledge management. Knowledge management (KM) is defined as the tools, techniques, and processes for the most effective and efficient management of an organization's intellectual assets (Davies, Studer, Sure, & Warren, 2005). Knowledge management consists of the combination of data and information processing capacity (i.e., information technologies), as well as the creative and innovative capacity of human resources. Knowledge management entails an organization viewing its processes as knowledge processes, in which these processes involve application of knowledge within the organization. Knowledge management focuses on the generation and application of knowledge, leveraging and sharing knowledge to increase the derived value, importing knowledge in the form of skilled employees, connecting knowledge workers, and motivating knowledge workers (Mohrman & Finegold, 2000). According to Robbins (2003) the process of knowledge management entails organizing and distributing an organization's collective wisdom so that the right information gets to the right people at the right time. As knowledge management becomes increasingly important, organizations must strive to understand the dynamics of knowledge management. This article will discuss the elements of knowledge management, in addition to presenting a case on how organizations can use knowledge management as strategy, where knowledge management is valued more than funding as a strategic resource.

BACKGROUND

According to Metaxiotis, Ergazakis, and Psarras (2005), knowledge management has its origins in several business improvement areas including Total Quality Management, Business Process Reengineering, Information Systems, and Human Resource Management. Historically knowledge management can be distinguished within three timeframe generations (Metaxiotis et al., 2005). The first generation encompasses the period between 1990-1995. Within this generation knowledge management initiatives focused on defining knowledge management, exploring its benefits, and designing KM-specific projects (Metaxiotis et al., 2005). In addition, artificial intelligence (AI) has been essential in the evolution of knowledge management, where AI is represented and defined as computer intelligence. AI has influenced knowledge management research pertaining to knowledge representation and knowledge storage. Knowledge storage (also referred to as knowledge repository) is based on the storage of knowledge and information for later use, both intellectually and physically (i.e., documents). The main goal of storage is . The second generation of knowledge management emerged around 1996, in which organizations began to establish jobs for KM specialists and knowledge workers. Within this generation, knowledge management research focused on knowledge definitional issues, business philosophies, systems, frameworks, operations and practices, and advanced technologies. The second generation of knowledge management emphasized knowledge management as a systematic organizational change, where management practices, measurement systems, tools, and content management needed co-development (Metaxiotis et al., 2005). The third generation of knowledge management encompasses present-day philosophies. According to Wiig (as cited by Metaxiotis et al., 2005), the difference in the third generation is the degree to which the third generation is integrated with the enterprise's philosophy, strategy, goals, practices, systems, procedures, and how it becomes part of each employee's work-life. The third generation of knowledge management plays on the link between knowing and action, where all knowledge is considered inherently social and cultural, and organizational knowledge can only be realized through change in organizational activity and practice (Metaxiotis et al., 2005).

The underpinnings of knowledge management encompass knowledge workers. Knowledge workers are considered experts in some abstract knowledge base, in which they identify with their profession and not a particular organization (Griffin & Ebert, 2002). This knowledge base can

consist of both educational and work experiences that add to the knowledge workers repository. Knowledge workers utilize acquired knowledge as raw materials and often rely on technology to design new products and business systems. In today's workforce the need for knowledge workers increases as the importance of information-driven professions increases. For knowledge workers, re-training and training updates are crucial in maintaining their skill set. Deficiencies in training and training updates can result in a loss of competitive advantage for the organization. As knowledge workers maintain their skills, they seek knowledge-based pay, also referred to as skilled pay. According to Griffin and Ebert (2002), knowledge-based pay is a performance pay plan based on rewarding employees for acquiring new skills or knowledge. This pay plan is based on the premise that as the knowledge worker acquires more knowledge and skill, he or she becomes more valuable to the company. Once knowledge workers are sustained, organizations can then engage in utilizing various classes of knowledge.

As organizations engage in knowledge management, it is beneficial to be aware of knowledge classes. Small and Sage (2005) state that there are two main classes of knowledge, tacit and explicit. Explicit knowledge is knowledge that can be codified. It is formal and systematic and can be found in books, enterprise repositories, databases, and computer programs. Zack (1999b) notes that explicit knowledge is more precise and formally articulated. Tacit knowledge is difficult to articulate and more personal, rooted in contextual experiences. "Tacit knowledge is subconsciously understood and applied, difficult to articulate, developed from direct experience and action, and usually shared through highly interactive conversation, storytelling, and shared experience" (Zack, 1999b, p. 2). Once organizations are familiar with the classes of knowledge, they can then evaluate their present knowledge base in an effort to use knowledge management as a tool for organizational strategy.

KNOWLEDGE MANAGEMENT AND ORGANIZATIONAL STRATEGY

Knowledge management entails how an organization can generate and communicate knowledge, where knowledge is defined as usable ideas (Behery, 2008). Knowledge management involves the process of managing knowledge to meet existing needs, identifying and exploiting existing and acquired knowledge assets, and developing new opportunities (Metaxiotis et al., 2005). From an organizational viewpoint, knowledge management focuses on exploiting and developing knowledge assets to further the company's goals and objectives. The objective of knowledge management is to identify and leverage collective knowledge in an effort to aid organizations in achieving a competitive advantage through utilization of organizational strategy. Strategy within an organization encompasses an organization's plans or actions for attaining a specific goal. Knowledge management can be used within an organization's strategy as a tactic for achieve its goals and objectives. As organizations seek competitive advantage, utilization of knowledge management can be beneficial. The benefits of organizations incorporating knowledge management into their organizational strategy include innovation, increased organizational learning, improved intellectual asset management, increased operational efficiency, time-to-market improvement, and continuous improvement (Demarest, 1997). Holsapple and Joshi (2000) list several factors that can influence an organization's inclusion of knowledge management in its strategy. These factors include culture, leadership, technology, organizational adjustments, employee motivation, and external aspects. These factors can be further organized into three categories, which include managerial influences, resource influences, and environmental influences. Once organizations have an understanding of the origins of knowledge and knowledge management, the organization can then create knowledge management projects as a strategy component to attain their organizational goals.

Davenport, DeLong, and Beers (1998) categorized four types of knowledge management projects based on organizational objectives. These categorizations, organized by objectives, include the following:

1. To create knowledge repositories, which store both knowledge and information, often in documentary form. Repositories can fall into three categories: those that include external knowledge, such as competitive intelligence; those that include structured internal knowledge, such as research reports and product-oriented marketing material as techniques and methods; and those that embrace informal, internal, or tacit knowledge, such as discussion databases.
2. To improve knowledge access, to provide access to knowledge, or to facilitate its transfer among individuals; here the emphasis is on connectivity, access, transfer, technologies (i.e., videoconferencing systems), sharing tools, and telecommunications networks.
3. To enhance the knowledge environment, so that the environment is conducive to more effective knowledge creation, knowledge transfer, and knowledge use. This involves tackling organizational norms and values as they relate to knowledge.
4. To manage knowledge as an asset, as well as recognize the value of knowledge to an organization. An organization's intangible assets, to which value can be assigned, can include assets (i.e., technologies that are sold under license or have potential value), customer databases, and detailed parts catalogs.

Organizations can then utilize these categorizations to develop teams and engage in the knowledge management process via knowledge sharing.

Organizations can utilize knowledge sharing in its stratagem. Knowledge sharing can be regarded as a catalyst for knowledge management. According to Dixon (2000) organizations are addressing the issue of knowledge sharing because knowledge sharing is more feasible as technology evolves, as well as due to the organization's growing awareness of the importance of knowledge to organizational success. Dixon (2000) posits that there are three myths to knowledge sharing which include: (1) build it and they will come (referencing knowledge sharing as a warehoused entity), (2) technology can replace face-to-face contact, and (3) first you have to create a learning culture to share knowledge. If organizations seek to engage in knowledge management, they must dispel these myths and engage in knowledge activities. There are two main types of knowledge activities which include finding effective ways to translate their ongoing experiences into knowledge (creating common knowledge) and transferring knowledge across space and time (leveraging common knowledge). Creating common knowledge entails the team performing a task, achieving the outcome, exploring the relationship between the action and outcome, and gaining common knowledge. Leveraging knowledge entails the team performing a task, achieving the outcome, exploring the relationship between the action and outcome, gaining common knowledge, selecting a knowledge transfer system, translating knowledge into a form usable by others, and lastly, adapting knowledge for its own use by the receiving teams. Once knowledge is created or exploited, leveraging (transferring) the knowledge is key in achieving organizational success.

In addition to knowledge sharing, exploiting existing knowledge and knowledge transferring can be used in the organization's stratagem. Exploiting existing knowledge and transferring knowledge within the organization can result in cost savings for an organization (Dixon, 2000). Cost savings can be a key component in an organization's strategy as it seeks to minimize costs and maximize profits. To achieve these results the organization must look at methods of knowledge transfer. There are several methods to transfer knowledge. Knowledge is transferred effectively when the transfer process fits the knowledge being transferred. According to Dixon (2000), there are five main methods to transfer knowledge as follows:

1. *Serial Transfer:* The knowledge a team has gained from doing its task in one setting is transferred to the next time that the team does the task in a different setting. Tacit and explicit knowledge are utilized.
2. *Near Transfer:* Explicit knowledge a team has gained from doing a frequent and repeated task is revised by the other teams doing similar work.
3. *Far Transfer:* Tacit knowledge a team has gained form doing a non-routine task is made available to other teams doing similar work in another part of the organization.
4. *Strategic Transfer:* The collective knowledge of the organization is needed to accomplish a strategic task that occurs infrequently but is critical to the whole organization. Utilizes tacit and explicit knowledge.
5. *Expert Transfer:* A team facing a technical question beyond the scope of its own knowledge seeks the expertise of others in the organization. Utilizes explicit knowledge.

Once an organization identifies the type of knowledge that will be exploited, the type of knowledge transfer that will yield the best results can be chosen. According to Dixon (2000), transferring and leveraging knowledge is critical for an organization's current and future viability.

According to Zack (1999a), organizations are viewing knowledge as their most valuable and strategic resource. To maintain a competitive edge, organizations must manage their intellectual resources and capabilities strategically. An organization's strategy can aid in guiding knowledge management. Its strategic framework can act to identify knowledge management initiatives that support the organization's mission, strengthen its competitive positions, and create shareholder wealth. For this reason, Zack (1999a) asserts that knowledge management can be considered the most important strategic resource, and the ability to acquire, integrate, store, shape, and apply it becomes the most important capability for building and sustaining a competitive advantage. As knowledge management takes the forefront, it appears that organizations are valuing knowledge, not funding, as the most strategically important resource. As a result, learning is viewed as the most strategically important capability for business success. Halawi, McCarthy, and Aronson (2006) posit that knowledge management is utilized as a strategy to improve organizational competitiveness. Strategy is vital to organizational success. According to Griffin and Ebert (2002), organizational strategy entails setting strategic goals (long-term goals derived from the organization's mission), analyzing the organization, analyzing the environment, matching the organization and its environment, and formulating strategy. Therefore, within the scope of knowledge management, strategy formulation should be considered.

Porter (1979) posits that the essence of strategy formulation is coping with competition. The state of competition in an industry depends on the five basic forces. These forces as depicted by Porter (1979) include threat of entry, bargaining power of suppliers, bargaining power of customers, threat of substitute products or services, and jockeying for position. Once organizations have assessed the forces that can affect competition, they can use this knowledge in strategy formulation and implementation. Porter (1998) asserts

that having knowledge of sources of competition provides groundwork for a strategic agenda of action. These actions include highlighting strengths and weaknesses, animating positioning of the company in its industry, and clarifying areas where strategic changes may yield the greatest payoff. Utilizing knowledge management in strategizing can ultimately result in operational effectiveness, in which an organization outperforms rivals in performing similar activities.

FUTURE TRENDS

According to Delmonte (2003), we have entered into an information age where the use of information technology could become the source of competitive advantage. Davies et al. (2005) affirm that knowledge is currently the key battleground for competition. In addition, knowledge is viewed as the result of requirements for highly skilled labor, new computing, telecommunications technologies, faster innovation, and shorter product cycles. To capitalize on this trend, organizations must consider their human resource capabilities and engage in knowledge management. Weatherly (2005) states that there has been a shift toward integrated human resource information technology and self-service functionality. This integration affects employees and HR personnel. The shift has been a direct response to the competitive challenges of today's economy. Advances in computer technology and Web-based technology have been attributed to advances in the economy. Weatherly (2005) posits the following:

"A critical success factor influencing an organization's ability to lead, simply languish or ultimately falter in the marketplace will eventually come to rest on the positive synergy the organization is capable of generating between the human capital assets in its employ and the judicious investments in technology that it makes in its efforts to remain abreast of competition." (p. 32)

In relation to knowledge management, HR has the task of utilizing informational assets within the organization. The organization's business processes and support systems must be parallel to continue to utilize and expand essential assets. As information technology is used in managing knowledge, support systems like automation become beneficial.

HR automation can also be referred to as self-service, where employees can essentially service themselves with mechanized processes. Automation comprises HR tasks that utilize technologies and computerization to replace the labor intensity of various tasks. The benefits of HR automation include reduced turnaround time per transaction, reduced cost per transaction, and reduced number of inquires to HR. Ultimately, automated service promotes greater efficiency and cost effectiveness for the organization. Examples of self-service include Web-based technologies used to automate and computerize administrative transactions. Administrative transactions include payroll, benefits, and training. The traditional approach to HR administration including costly processes, redundant data management, and paper-based processes can be replaced with automated processes. Utilizing automation allows organizations flexibility in allocating resources to functions that promote accomplishment of the company's mission and objectives. In today's business world the change from traditional approaches to encompassing the acquisition and management of information has become a competitive force (Raeside & Walker, 2001). Knowledge management systems, pooled with effective knowledge management strategy, can maximize the potential for an information system to contribute to sustained competitive advantage (Delmonte, 2003). As organizations engage in establishing a competitive advantage, the necessity for changes and evolution involving information technology are becoming apparent.

CONCLUSION

Knowledge management seeks to effectively and efficiently manage an organization's intellectual assets through the utilization of organizational tools, technology, and processes (Davies et al., 2005). Knowledge management has origins that encompass a gamut of business improvement areas including Total Quality Management, Business Process Reengineering, Information Systems, and Human Resource Management. The knowledge management process can include: (1) utilizing knowledge workers, (2) managing knowledge to meet existing organizational needs, (3) identifying and exploiting existing and acquired knowledge assets, and (4) developing new opportunities (Metaxiotis et al., 2005). Within the scope of knowledge management, organizations can share knowledge, exploit existing knowledge, and transfer knowledge as a part of their organizational strategy. As knowledge management becomes fundamental, organizations are valuing knowledge more than funding, making knowledge management strategically important for organizational success. As trends like the use and dependence upon information technology within the knowledge management process become ubiquitous, organizations must seek to engage in continuous improvement to maximize resources as they seek to remain competitive. Utilization and continued development of knowledge management within organizations will yield organizational benefits, including increases in productivity and profit generation, as well as establish competitive advantage.

REFERENCES

Behery, M.H. (2008). Leadership, knowledge sharing, and organizational benefits within UAE. *Journal of the American Academy of Business, 12,* 227-237.

Davenport, R., DeLong, D., & Beers, M. (1998). Successful knowledge management projects. *Sloan Management Review, 39*(2), 43-57.

Davies, J., Studer, R., Sure, Y., & Warren, P.W. (2005). Next generation knowledge management. *BT Technology Journal, 23*(3), 175.

Delmonte, A.J. (2003). Information technology and the competitive strategy of firms. *Journal of Applied Management and Entrepreneurship, 8*(1), 115-129.

Dixon, N.M. (2000). *Common knowledge: How companies thrive by sharing what they know.* Boston: Harvard Business School Press.

Griffin, R.W., & Ebert, R.J. (2002). *Business* (6th ed.). Upper Saddle River, NJ: Prentice Hall.

Halawi, L.A., McCarthy, R.V., & Aronson, J.E. (2006). Knowledge management and the competitive strategy of the firm. *The Learning Organization, 13*(4), 384-397.

Holsapple, C.W., & Joshi, K.D. (2000). An investigation of factors that influence the management of knowledge in organizations. *Journal of Strategic Information Systems, 9,* 235-261.

Lawler, E.E., & Mohrman, S.A. (2003). *Creating a strategic human resources organization.* Stanford, CA: Stanford University Press.

Metaxiotis, K., Ergazakis, K., & Psarras, J. (2005). Exploring the world of knowledge management: Agreements and disagreements in the academic/practitioner community. *Journal of Knowledge Management, 9*(2), 6-18.

Mohrman, S.A., & Finegold, D.L. (2000). *Strategies for the knowledge economy: From rhetoric to reality.* Retrieved October 1, 2007, from http://www.marshall.usc.edu/ceo/Books/pdf/knowledge_economy.pdf

Porter, M.E. (1979). How competitive forces shape strategy. *Harvard Business Review, 57*(2), 137-145.

Porter, M.E. (1998). *On competition.* Boston: Harvard Business School Publishing.

Raeside, R., & Walker, J. (2001). Knowledge: The key to organizational survival. *The TQM Magazine, 13*(3), 153-160.

Robbins, S.P. (2003). *Organizational behavior.* Upper Saddle River, NJ: Prentice Hall.

Small, C., & Sage, A. (2005). Knowledge management and knowledge sharing: A review. *Information Knowledge Systems Management, 5,* 153-169.

Weatherly, A. (2005). HR technology: Leveraging the shift to self-service—it's time to go strategic. *HR Magazine, 6*(50), 32.

Zack, M. (1999a). Developing a knowledge strategy. *California Management Review, 41*(3), 125-145.

Zack, M. (1999b). Managing codified knowledge. *MIT Sloan Management Review, 40*(4), 45-58.

KEY TERMS

Automation: The process and transformation of manual labor into automated and/or computerized processes.

Explicit Knowledge: Formal, systematic knowledge that can be codified (Zack, 1999b).

Knowledge: The collection of what has been learned or perceived.

Knowledge Creation: Translating ongoing experiences into knowledge (Dixon, 2000).

Knowledge Leverage: Transferring knowledge across space and time (Dixon, 2000).

Knowledge Management: Utilizing tools, techniques, and processes to regulate intellectual assets (Davies et al., 2005).

Knowledge Process: Application of creating and leveraging knowledge.

Organizational Strategy: An organization's tactics and actions to achieve the company's mission and objectives.

Tacit Knowledge: Knowledge rooted in contextual experiences that is subconsciously understood (Zack, 1999b).

Knowledge Management Challenges in the Non-Profit Sector

Paula M. Bach
The Pennsylvania State University, USA

Roderick L. Lee
The Pennsylvania State University, USA

John M. Carroll
The Pennsylvania State University, USA

INTRODUCTION

The concept of knowledge management is rooted in cognitive psychology and organizational theory. Knowledge management is concerned with the creation, storage, and distribution of knowledge by groups, organizations, and communities. Two theoretical frameworks are instrumental in shaping the knowledge management discourse: organizational knowledge creation (Nonaka, 1994) and organizational knowledge (Spender, 1996).

Widely cited in the literature is Ikujiro Nonaka's (1994) explication of the epistemological and ontological dimensions of organizational knowledge creation. Michael Polanyi (1966), makes a distinction between tacit and explicit (codified) knowledge in the epistemological dimension, whereas social interaction is the foundation of the ontological dimension. Over the years, the term knowledge management has been conflated with organizational learning and memory. Realizing that knowledge, memory, and learning are all interrelated, John-Christopher Spender (1996) proposed a knowledge-based theory of the firm. The knowledge-based theory of the firm is primarily concerned with the collective capabilities of generating, combining, and applying knowledge.

Given the advances in computing and telecommunications technologies, scholars have considered how information technologies can be used strategically to facilitate knowledge management (Alavi & Leidner, 2001). For example, wikis, blogs, content management systems, and the like provide dynamic infrastructures that support the creation, transfer, and application of knowledge. More importantly, these tools enhance organizational memory that can subsequently be shared across time and space. However, a knowledge friendly culture (Davenport & Prusak, 1998) precedes an effective knowledge management program.

The purpose of this article is to explore the challenges that arise in nonprofit settings, particularly the ways in which knowledge is stored and transmitted through an organization's culture. We propose two key challenges that influence organizational culture: acceptance of change and leaders' ability to develop a knowledge friendly culture. We conclude with a discussion on the role that these factors played in constraining a knowledge friendly culture in two case studies.

BACKGROUND

While the historical definition outlines knowledge management traditionally in firms where knowledge workers possess and share knowledge that is critical for the firm to capture, nonprofit organizations, in principle, benefit from some of the same goals. The goals of knowledge management in for-profit firms include competitive advantage, greater innovation, better customer experiences, consistency in good practices, and facilitating organizational learning. Although competitive advantage may not be a worthy goal for a nonprofit organization (NPO), consistency in good practices and facilitation of organizational learning are. Addressing the key knowledge management goals and challenges in a nonprofit setting has its own set of unique challenges.

Nonprofit organizations range from small, diverse community-based organizations that address local issues and rely primarily on volunteer labor, to large, nationally-based organizations such as the Red Cross. Nonprofit organizations consist of such groups as arts and culture, education and research, health services, social and legal services, religion, fraternities and sororities, civic and social services, and foundations. As such "their knowledge capital is heterogeneous, widespread, rarely formalized, and unstable" (Lettieri, Borga, & Safoldelli, 2004). These groups depend on temporary volunteers to help them work on specific projects (Boris, 1998). Yet some NPOs use paid labor. These workers can be classified as part- or full-time employees or consultants. Employees usually occupy a leadership position such as executive director and consultants and lead specific projects

based on their expertise such as technology infrastructure. Because nonprofits do not offer high salaries and are more strongly linked to a cause, they may not attract people with strong leadership skills. The lack of strong leadership and permanency pose a problem for knowledge management.

The above goals and challenges of nonprofit knowledge management are rooted in organizational culture. Organizational culture can be thought of as patterns of problem solving and ways of thinking that must be taught to new members (Schein, 1988). We argue that the culture of nonprofit organizations poses challenges for knowledge management. Challenges such as acceptance to change and leadership compromise consistency in good practices and facilitation of organizational learning. Knowledge management, therefore, must first be integrated into a nonprofit organization's culture.

CHALLENGES WITH KM IN NONPROFIT ORGANIZATIONS

In this section, we outline two cultural factors within nonprofit organizations that lead to challenges in knowledge management. One of these two factors also serves as a point for analysis in one of the two case studies examined in the next section.

Acceptance of Change

Two perspectives dominate the literature on organizational change: episodic and continuous (Weick & Quinn, 1999). Changes in personnel or technology trigger episodic change, whereas improvisation, learning, and adaptation trigger continuous change. The leader's role shifts from a change agent, creating change in the episodic view, to redirecting change in the continuous view. Each perspective illustrates a role for leadership and the subsequent importance of first modeling change in order to facilitate changes in the culture of the organization. Accepting change is a challenge in nonprofit organizations because values and norms do not support social structures and relationships that enhance knowledge management. Social structures that enhance knowledge management include heterarchy (instead of hierarchy) (Hedlund, 1994; Schutt, 2003) and knowledge sharing among the dynamic relationships of groups and individuals (Alavi & Leidner, 2001; Gilbert, 2002; Hurley & Green, 2005; Sahay & Robey, 1996; Schutt, 2003; Snowden, 2002).

In addition, volunteer culture complicates nonprofit social structure. Reliance on volunteers makes knowledge management a key challenge because new volunteers entering a nonprofit organization need to know just enough to get a job done. However, recurring volunteers may come in once or twice a month and need to know if anything has affected their task since the last time they volunteered. Volunteer turnover is also a problem for knowledge management. Volunteers working in a nonprofit setting may last only a few months, and without a knowledge management program, the knowledge they gain leaves with them. This constant change of incoming and outgoing volunteers poses challenges for managing volunteer knowledge.

Furthermore, for both volunteers and staff, organizational learning and socialization can enhance acceptance and sustainability of change (Farooq et al., 2005; March, 1991; Sahay & Robey, 1996; Schein, 1986, 1988). This socialization occurs as a result of organizational learning where mutual learning can be used for the development of knowledge in organizations. Organizational learning must be sustained (Farooq et al., 2005; Merkel et al., 2005) as part of routines that support the dynamic relationships of groups and individuals. Knowledge management can enhance sustainability, but NPOs must recognize the need for change as part of the organizational culture. As part of accepting change and supporting knowledge management, nonprofit organizations can explore knowledge management technologies.

Knowledge management technologies that are more likely to fit into the culture of NPOs are knowledge sharing technologies rather than knowledge application technologies because knowledge application technologies are domain specific and take more resources than most nonprofits have to dedicate to technology. Knowledge application technologies include expert systems and case-based reasoning systems. Expert systems capture knowledge and apply that knowledge to problems in a domain, resulting in a solution. On the other hand, case-based reasoning systems use knowledge from past experiences or cases to solve new problems. Because these systems take an enormous number of resources to design, develop, and maintain, such systems are beyond the financial scope of many nonprofit organizations. Yet, the scope of problems, for example, expert knowledge applied to fundraising problems, that nonprofit organizations face might be addressed using a knowledge application technology.

Knowledge sharing technologies, however, are already used by nonprofit organizations. Technologies such as the e-mail, search engines, document management systems, and databases store and allow users to share knowledge. Both knowledge application and knowledge sharing technologies must support the dynamic flow of knowledge and fit into the complex ecology of knowledge in organizations (Snowden, 2002). Technology acceptance occurs in the context of organizational culture that supports knowledge management initiatives through socialization and organizational learning. Accepting and sustaining a knowledge management culture remain key challenges for nonprofit organizations.

The goal of a business organization is to earn a profit, whereas the goal of a nonprofit is to provide a service. Structure and survival motivate change in for-profits. Competition

drives learning and change, whereas nonprofits struggle with learning and change because volunteer help is valuable. Pressuring volunteers too much to learn and accept change could result in them leaving. Nonprofits cannot leverage their volunteers as much as for-profits can leverage their workers because volunteers are not paid. This is a challenge for nonprofits because they have to learn how to provide efficient and effective services, but the volunteer culture can introduce inefficiencies. Strong leadership supports learning and change.

Leadership

Leadership is a key factor in creating and sustaining a knowledge management culture. In the nonprofit sector, a leader must bring the community together for a cause. Knowledge holders in an NPO include volunteers, employees, and board members (Lettieri et al., 2004). An NPO leader must understand the difference between the knowledge and skills required by, and the knowledge and skills available to, the organization (Lettieri et al., 2004) in order to create and sustain a knowledge management culture. Bringing knowledge management changes to an organization is challenging. As such leaders need to serve as change agents and understand that supporting an entire knowledge management socio-technical system requires a collective buy-in from many stakeholders (Young, 2001). Change to a knowledge management culture can result in anxiety and leaders must be able to support the psychological barriers to change such as the required change in assumptions (Schein, 1986, 1988) that include embracing the paradoxical nature of knowledge as both thing and flow (Snowden, 2002). In order to develop a knowledge friendly culture, leaders need to be aware of and model knowledge management values including support and nurturing of knowledge workers in the nonprofit socio-technical system: process, organizational culture, and information technology (Powell & Swart, 2005; Schutt, 2003).

CIVIC NEXUS AS THE NONPROFIT ORGANIZATION CONTEXT

We explore further the challenges that organizational culture poses for knowledge management in nonprofit organizations in the context of two case studies. The case studies were part of a larger research project called Civic Nexus. Civic Nexus was a participatory action research study in which we partnered with nonprofit organizations to understand and enhance their information technology practices. We worked with groups for 1 year. We began with fieldwork to understand how nonprofits use technology and how it fits with their work practices and values. This early fieldwork helped us to understand the socio-technical expertise and decision-making structure that a group already had in place. We leveraged this early understanding to help the groups choose a project on which we worked together.

Technology projects could include technical developments (e.g., hardware, software or Web site), transformations in procedures or practices (e.g., knowledge management), or organizing technical training sessions. Implementing projects provided a meaningful context from which nonprofits could learn to plan and execute technology-related activities. Because learning and sustainability were goals of the project, we embedded ourselves in the organization's structure and played the role of facilitator, became a member of a committee, or worked one-on-one with key staff members. After the year, we gradually "faded," staying in contact to check learning and sustainability by monitoring their ability to continue with the accomplishments we produced together over the year. The two groups, a local symphony and a grassroots historical initiative that we worked with, both have had specific problems with knowledge management. Although problems with leadership and change existed in both groups, the symphony mainly experienced problems with leadership and the historical group mainly experienced problems with change. Specific knowledge management technologies were related to the leadership and change challenges resulting in problems with technology use. Both organizations had access to knowledge sharing technologies. Whereas the historical group had problems with e-mail and a document management system, the symphony had problems with a database.

Case Study 1: The Nittany Valley Symphony

The symphony recently celebrated its fortieth anniversary and is the only classical music group local to central Pennsylvania. Over the course of the year we worked on a project to improve office flow and database access for the Executive Director (ED). A board of directors oversees the executive director, who was the only paid office staff member. Volunteers helped out in the office when special regular tasks needed to be accomplished, for example, preparing envelopes and letters to mail to past concert attendees or managing a poinsettia sale during the winter holiday season. One regular office volunteer works in the office two to four times a week to manage accounts received and another once a week to enter data. A committee, called management of information systems (MIS), consisting of a board member as chair and other volunteers, manages the information technology in the office. This committee designed and implemented a database, purchased computers and software, and ensures that the information systems are maintained.

Civic Nexus researchers worked closely with the MIS committee, executive director, and volunteers to research

and analyze a problem situation and devise a solution. The MIS committee decided to work on a project that assessed the information needs of the organization and whether the database was currently meeting those needs. The executive director was unable to track donations and send thank-you letters, among other tasks, using the database. The database, however, functionally could serve these tasks, but its usability was a major barrier for the executive director.

After working together with the Civic Nexus researchers, the MIS committee decided that part-time help, a technical office manager, might relieve some of the database problems. After a reworking of the budget and approval by the board, another staff member was hired to manage the database and other technical office issues that the ED was unable to address. The key knowledge management challenge, leadership, in the NVS is related to this office flow and suitability of IT project.

Leadership

The new ED was hired because of her 20 years of fundraising experience. Despite her experience as a fundraiser, she was unable to use the main knowledge management tool: the Access database. In an interview, the new ED shows her resistance to learning about the database.

Interviewer: And the database,

ED: The Access

Interviewer: Yea, Access, you are learning, what did you say?

ED: I don't want to, I know nothing about Access

Interviewer: I know the one day you were using tables [in Access]

ED: That's what I say and that's what I am going to stick to

Interviewer: Okay

ED: I do not want. But you know, the more you know, the more they [the board of directors] are going to expect you to do.

The current ED resists engaging in the sociotechnical structure of the organization, and this results in a disruption in the knowledge management practices, even if they were not optimal. The resistance indicates that the ED may not be equipped to lead the organization's knowledge management practices.

While the recent initiatives by the MIS committee showed some progress toward overcoming knowledge management challenges, no clear leadership role existed in the NVS. Little communication and knowledge sharing occurred between the office and the board. This disconnection in management caused some tension between the ED and the board and does not model leadership support for knowledge management.

Turnover with the ED also posed a challenge for knowledge management, and the recent turnover was particularly problematic. Although the former and new executive directors were able to work together for 2 weeks to help with knowledge transfer, the former ED was a young graduate of business school with database skills. Had this pairing of these two executive directors, a younger one with technical skills and an older one with fundraising knowledge, been longer, the skill transfer might have had more impact (Convertino, Farooq, Rosson, & Carroll, 2005) on the new ED's ability to lead knowledge management initiatives. In addition, the board did not recognize the value of pulling information out of the database to make decisions until recently when the MIS committee brought it up. Therefore, no change agent exists. Yet, the MIS committee seemed to be playing the role of change agent in the move toward better knowledge management.

Although the new awareness brought forth by the Civic Nexus involvement brought a positive impact on knowledge management within the NVS, the organization continues to have challenges with knowledge management related to contextual factors. For example, the strength of leadership depends on who occupies the president and executive director positions. Knowledge management opportunities and challenges depend on the current awareness of KM and the current resource pool. Furthermore, because "enormous 'gold mines' of knowledge are buried in databases" (Becerra-Fernandez, Gonzalez, & Sabherwal, 2004), importance lies in the symphony's ability to harvest this knowledge effectively. The MIS committee is making strides toward organizing office flow and IT practices, but their strides are slow. Although it is unlikely that the NVS will have a stable knowledge management culture, the MIS committee is working to extract knowledge from its database, and from our experience with other groups involved with the Civic Nexus project, the NVS has more sophisticated knowledge management practices than other groups we have worked with.

Case Study 2: The Underground Railroad

The Underground Railroad (UGGR) is a small, loosely connected, grass-roots group of volunteer historians in a small rural town in the Commonwealth of Pennsylvania. The group's goals are to identify and document sites that were involved in historic Underground Railroad activity and seek federal verification of the sites. No paid positions exist in

this group. In fact, the leader's book sales and independent fundraising efforts provide financial resources. We engaged in a mutual partnership to design and develop an online collaborative environment to support the goals of the group and to facilitate knowledge management. In this section, we highlight several factors that have inhibited adoption and productive use of the technology during the first 6 months of our partnership.

Acceptance of Change

Data from participant observation and interviews revealed that the group values timely responses to requests for information. Even though an IT infrastructure is now in place, adoption of the technology has undermined the productive use for knowledge management activities. In one case, use of the software is rather limited on two fronts. First, the group leader does not have Internet access at home and does not see any real value in having Internet access at home when she has access to the Internet at work. Second, another central actor in the local group is prohibited from accessing the Internet while at work. Therefore, no overlap occurs when accessing the online collaborative environment between two key members. As a result, members bypass the online system and continue to use e-mail and telephone as their primary modes of communication.

In a similar case, one of the actors involved lacks computer literacy and the motivation to learn to use computer technology. For instance, this subject matter expert has been a prominent figure in the local historical community for decades. During a training workshop, this individual demonstrated that he held quite a fount of tacit knowledge with respect to the UGRR history and knowledge of where documentation could likely be found. In addition, he knows which sites would be most attractive in terms of nominating them to the federal agency. However, this individual refused to use technology. In fact, he indicated that "he has gotten this far in life without using technology, so why start now." In order to capture some of his tacit knowledge, an undergraduate student was assigned to type his comments and selections into the discussion forum.

The second major challenge concerns the group leader's change in behaviors. In order to facilitate a knowledge friendly culture and bring the group together for a cause, the leader must serve as a change agent. Understanding this role presents a challenge for leaders in nonprofits. The leader must first undergo deep change in order to lead by example and model knowledge management values. As mentioned above, the group leader does not have Internet access at home and has continued to favor e-mail and verbal communication as opposed to using the online collaborative environment. When asked why, she responded, "That's just the way that we are used to doing things." For example, we held a workshop in order to train the users on the collaborative tools and give them an opportunity to use the tools. However, the group leader and other members favored verbal communication instead of computer-mediated communication. One member interrupted the flow and asked, "Aren't we supposed to be typing our comments into the system?" The room went silent for a moment and then the verbal dialogue between the group leader and two other participants continued.

Currently, this organization's culture is somewhat frozen. As a result, the organization is presently unaware of the need to change its culture in order to adopt technology and alter its values, norms, and routines. In the next action research cycle, we will intervene in order to assist the leader in understanding her role in facilitating a community of practice in order to develop a knowledge structure capable of stewarding knowledge for the organization.

FUTURE TRENDS

To support knowledge management within nonprofit organizations, future research on KM in nonprofit settings should focus on social-technical interventions that emphasize planned change in the organization's culture in the context of communities of practice. Future research on support networks and community information technology conferences where communities of practice emerge can elicit support for knowledge management among nonprofit organizations. In addition, we suggest that participatory action research is a malleable tool to facilitate and manage the trajectory and change in organizational culture. Finally, while the case studies showed that two small nonprofit organizations continue to face challenges with knowledge sharing technologies such as e-mail, a document management system, and a database, we caution larger, more resource rich nonprofit organizations to establish knowledge management practices before they implement a more sophisticated knowledge application technology.

CONCLUSION

We conclude that the two nonprofit organizations in this study are capable of neither episodic nor continuous change without outside intervention. Indeed, these case studies illustrate that changes in organizational culture may be more pronounced in nonprofit settings. Facilitating change in the culture of the organization requires a shift in culture and changes in leadership behavior. The leader must create conditions that are conducive to a knowledge friendly culture by first altering their underlying norms, values, and assumptions. Only then will change be effective.

REFERENCES

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Becerra-Fernandez, I., Gonzalez, A., & Sabherwal, R. (2004). *Knowledge management: Challenges, solutions, technologies*. Upper Saddle River, NJ: Pearson Prentice Hall.

Boris, E. (1998). *Myths about the nonprofit sector*. Retrieved December 14, 2007, from http://www.urban.org/url.cfm?ID=307554

Convertino, G., Farooq, U., Rosson, M. B., & Carroll, J. M. (2005). Old is gold: Integrating older workers in CSCW. In *Paper presented at the 38th Hawaii International Conference on System Sciences*, Honolulu, HI.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston, MA: Harvard Business School Press.

Farooq, U., Merkel, C. B., Nash, H., Rosson, M. B., Carroll, J. M., & Xiao, L. (2005). Participatory design as apprenticeship: Sustainable watershed management as a community computing application. In *Paper presented at the 38th Hawaii International Conference on System Sciences*, Waikoloa, HI.

Gilbert, M. C. (2002). Nonprofit knowledge management. *Nonprofit online news*. Retrieved December 14, 2007, from http://news.gilbert.org/NonprofitKM

Hedlund, G. (1994). A model of knowledge management and the N-form corporation. *Strategic Management Journal, 15*(Summer), 73-90.

Hurley, T. A., & Green, C. W. (2005). Knowledge management and the nonprofit industry: A within and between approach. *Journal of Knowledge Management Practice, 6*.

Lettieri, E., Borga, F., & Safoldelli, A. (2004). Knowledge management in non-profit organizations. *Journal of Knowledge Management, 8*(6), 16-30.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organizational Science, 2*(1), 71-87.

Merkel, C. B., Clitherow, M., Farooq, U., Xiao, L., Ganoe, C., Carroll, J. M., et al. (2005). Sustaining computer use and learning in community computing contexts: Making technology part of "Who they are and what they do." *The Journal of Community Informatics, 1*(2), 158-174.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

Polanyi, M. (1966). *The tacit dimension*. New York: Doubleday.

Powell, J. H., & Swart, J. (2005). This is what the full is about: A systematic modelling for organizational knowing. *Journal of Knowledge Management, 9*(2), 45-58.

Sahay, S., & Robey, D. (1996). Organizational context, social interpretation, and the implementation and consequences of geographic information systems. *Accounting, Management and Information Technologies, 6*(4), 255-282.

Schein, E. H. (1986). What you need to know about organizational culture. *Training and Development Journal, January*, 30-33.

Schein, E. H. (1988). *Organizational culture*. Retrieved December 14, 2007, from http://dspace.mit.edu/bitstream/1721.1/2224/1/SWP-2088-24854366.pdf

Schutt, P. (2003). The post-Nonaka knowledge management. *Journal of Universal Computer Science, 9*(6), 451-462.

Snowden, D. (2002). Complex acts of knowing: Paradox and descriptive self-awareness. *Journal of Knowledge Management, 6*(2), 100-111.

Spender, J. C. (1996). Organizational knowledge, learning and memory: Three concepts in search of a theory. *Journal of Organizational Change Management, 9*(1), 63-78.

Weick, K., & Quinn, R. (1999). Organizational change and development. *Annual Review of Psychology, 50*, 361-386.

Young, D. R. (2001). Organizational identity in nonprofit organizations. *Non-profit Management & Leadership, 12*(2), 139-157.

KEY TERMS

Communities of Practice: Knowledge-based structures that facilitate knowledge sharing and development, and support learning.

Knowledge Management: The creation, storage, and distribution of knowledge by groups, organizations, and communities.

Knowledge Management Systems: Information systems that are designed to support the creation, capture, storage, and distribution of expertise and knowledge.

Organizational Culture: The values, norms, and assumptions that are widely held by members of the organization that subsequently shape their behavior.

Organizational Knowledge: Tacit and explicit knowledge that is part of the organization's culture and identity, routines, standard operating procedures, and is expressed in documents.

Organizational Knowledge Creation: Knowledge that is created through a continuous dialogue between tacit and explicit knowledge through the processes of socialization, combination, internalization, and externalization.

Organizational Learning: A social process in which individuals in organizations enhance decision making and problem solving by improving knowledge and understanding.

Nonprofit Organization: An organization in the social sector whose main purpose is to either provide a service or support a cause.

Participatory Action Research: An interventionist method that involves close collaboration between the researcher and practitioners.

Knowledge Management for Production

Marko Anzelak
Alpen-Adria-Universität Klagenfurt, Austria

Gabriele Frankl
Alpen-Adria-Universität Klagenfurt, Austria

Heinrich C. Mayr
Alpen-Adria-Universität Klagenfurt, Austria

INTRODUCTION

Knowledge is one of the key drivers of innovation and success in the modern, information-based society. Consequently, knowledge has to be "operated" and "managed," which causes particular challenges due to the intangible nature of knowledge: "… it is fluid as well as formally structured; it is intuitive and therefore hard to capture in words or understand completely in logical terms. Knowledge exists within people, part and parcel of human complexity and unpredictability. " (Davenport & Prusak, 1998, p. 5) Being held in minds, knowledge is not easily accessible and hence, not manageable in the usual sense. Nevertheless, knowledge management (KM) tries to establish appropriate processes of externalizing, internalizing, and applying the knowledge of people involved in a given environment. Within that context, the notion of *knowledge* has undergone various definition attempts and interpretations.

From an economic and corporate perspective, knowledge was viewed as a commodity, like other products, to be packaged, archived, retrieved as needed, and sent across networks. An example of this approach is the "Wissenstreppe" (knowledge staircase), proposed by Klaus North (2002). This model proposes eight steps, each of which is linked to an instruction on how to reach the next step. The lowest level [1] consists of symbols. Combining these with rule-based syntax creates *data* [2], and the addition of semantics produces *information* [3]; information enriched by connectivity leads to *knowledge* [4]. Knowledge combined with applicability results in *ability* [5], which in combination with *willing* can be converted to *behaviour* [6]. Effective behaviour leads to *competence* [7]. Competences leading to a unique selling proposition (USP) create *competitive advantage* [8].

Knowledge became increasingly a decisive factor in competitive gain (e.g., Bryant, 2006), leading to an expanding demand for KM. However, manifold problems caused the failure of several KM initiatives, and led to the rediscovery of earlier approaches, such as that of Michael Polanyi (1973, 1985). Casselman and Samson (2005) extended the two types of knowledge, *explicit knowledge* and *tacit knowing*. Explicit knowledge can be represented by signs (symbols, text, and images), and thus stored electronically. As such, it is quite similar, or even might be seen as synonymous, to "information." Tacit knowing is always tied to a subject, that is, to a mind, and therefore, cannot be stored in a technical system. Nonetheless, it is possible to initiate processes that lead to the generation, externalisation, internalisation, and thus, to the sharing of tacit knowing.

Information technology (IT) is the natural enabler of managing *explicit knowledge* since it supports to store and handle signs: electronic content of any kind is easy to extend, rework, comment, structure, and complemented by metadata. These basic features of any document-based information management are strengthened in combination with standard or tailor-made KM Systems (KMS), like the one described in this chapter to support knowledge processes.

BACKGROUND

Tacit knowing can be understood as knowledge that is required to perform a behaviour, such as riding a bicycle, for which explicit knowledge is not mandatory (Dreyfus, Dreyfus, & Athanasiou, 1988): Even a child can learn to ride a bike without explicitly knowing specific rules or being able to articulate rules or formulas for balance calculations underlying bicycle riding. Knowledge that can be transformed into a skill is strongly embedded in experience, for example, gained from practising or sensing. Something can be understood comprehensively; reasons and connections can be recognised. Nevertheless, tacit knowing comprises aspects that are difficult to codify, such as personal convictions, perspectives, and values (Nonaka & Takeuchi, 1995).

The difficulty, and often even incapability, of articulating knowledge is one challenge of KM. Another one comes from the fact that even if *explicit knowledge* is available, it does not necessarily translate to action. In contrast to tacit knowing, explicit knowledge can be acquired via rote learning, which, however, can fail in its application or in conversion to intelligent behaviour. Learning facts without

relevant experience, understanding, or insight, results in the opposite of *tacit knowing*: lazy knowing, that is, knowledge that lacks capability.

The core concern of organisational KM, however, is the concrete use derived from (transforming or applying) knowledge, and not only to collecting and storing facts in databases, which would promote lazy knowing primarily. Therefore, KM, in the manufacturing industry, should focus on supporting and improving production processes. There are four categories of software-systems that can be used within that context (in sequence of increasing appropriateness):

Content management systems (CMS) and enterprise content management (ECM) transform structurally and semantically predefined (forms) information (content) into a desired uniform appearance (corporate identity). They are easy to handle, and establish tree structures of knowledge items, potentially complemented by metainformation.

Document management systems (DMS) store all kind of documents in a well-structured or extendable fashion. The documents have attached metadata, and they are indexed for rapid finding by full-text search. DMS have features like check in, check out, authorisation concept, workflow, user roles, history, and versions.

Learning management systems (LMS) are designed for knowledge transfer, but mostly concentrate on *internalisation* (learning). Content is created and arranged by experts.

Groupware (GW) systems focus on supporting the communication and collaboration of people having a common task or goal. They feature common workplaces for storing, exchanging, and processing documents, writing comments, and managing tasks.

When dealing with the introduction of knowledge management into an established organization, various social, technical, and organisational challenges arise that have to be considered for success:

1. ***Social challenges*** generally arise from changes inevitably coming up with the introduction of KM and KM-systems within an organisation. Change always requires energy in order to adapt to new circumstances, and also raises fear of failing to cope with the change. Specific KM-related social challenges are the willingness to share knowledge, knowledge externalization (how to articulate?), knowledge input, and retrieval.
2. ***Technical challenges*** refer to the functionality needed, and to the integration of KM mechanisms into the existing system landscape (platforms) that, in our case, required a tailor made solution. The KM user interface should behave (look and feel) similarly to the existing systems in order to be easily handled and used. In particular, the corporate design has to be conserved. Even in the case of a tailor-made solution, however, it should be flexible, w.r.t., both the user interface and the database interfaces. Another challenge comes from the key function of KM support, namely the search, if it is to be supported semantically and ontologically.
3. ***Organisational challenges*** primarily address individual learning: learning and knowledge processes mostly run self-organised; learners are stamped with personal conviction, perspectives, and values, forming a complex cognitive structure (Maturana & Varela, 1988; von Foerster, 1985/1999). Thus, structural changes can be achieved by controlling measures, but not instrumented or determined, and, knowledge processes cannot be completely planned.

Another organisational challenge is **search,** which can only be as good as the data to be searched. Consequently, all employees should use the same words ("language") for describing problems or solutions so that at least a wording-directive is needed. Another point is to train the users how to search.

Enhancing the level of education and further training of employees needs **time**, **freedom,** and "**knowledge rooms**": knowledge processes cannot be controlled by the parameters of economic efficiency. People need not only time to *internalize information*, they need also time for unconscious processes of internal assimilation and its linking, for understanding new and unplanned experiences, and for unexpected knowledge processes coming up with different types of acting.

Users have to be "won" without destroying their *intrinsic motivation*. Brain scientists, like Spitzer (2005), think that we have a natural predisposition to learn with fun, which is often distressed by education in school (Quinn, 2007). This could be pushed back if external rewards raise the knowledge sharing, retrieving, and transmission to something special. Thus, KM should offer to integrate knowledge processes in a natural way into the working day, where working and learning melts together.

Users easily can be stressed by huge systems coming with a big bang. Therefore, it is necessary to introduce a critical solution in small and easy-to-handle parts and steps: from function to function, and from department to department. The best way is to start in the area with the highest psychological strain, and with lead users, possibly supporting them by rapid prototyping and workshops.

Last, but not least, all the efforts put into motivating and training people, establishing, filling, and maintaining a knowledge base, must have a sustainable effect: this requires continuous attention of users and feedback, and appropriate activities in case of necessary modifications. A KM system, for example, should always exhibit something new to the user in order to attract his/her attention and motivation.

KNOWLEDGE FOR PRODUCTION (KFP)

Supporting knowledge management in the production industry is an important concern. We now focus on an exemplary approach for that domain, that is, a KM methodology and tools for a global operating craft paper production company having a three shift and 24/7 work process. Product quality here strongly depends on the knowledge each employee has about his/her workplace's share of the production process. Typically, this is *tacit knowing*, "stored" in the heads of the employees. Often, it cannot be explained or even will not by these people, because they are afraid of loosing competitive advantage, honour, identity, or power. In addition to that, the given setting mostly does not allow for the most effective way of sharing knowledge, namely face-to-face, think, for example, of distributed working areas, asynchronous attendance times due to shifts, and so forth.

INITIAL POSITION

Knowledge transfer from shift to shift is done using a "shiftbook" listing important events, using post-it notes for other information, and a few minutes talk (e.g., about observed or induced changes in chemical pulp composition). This is a source of information loss.

There are rules on how to enter data in the shiftbook; however, these are not consciously observed, and problem solutions are rarely added. Finding entries in the shiftbook reveals well-known problems associated with handwritten information, such as sequential search, illegible handwriting, no structure, paper quality, and so on.

If a problem occurs that cannot be solved by a certain employee, the practice is to ask for peer or expert help. Implicit knowledge about who might be asked in such a case may exist, but does not so in any case. Even when it is known whom to ask, this particular person might not be present: on annual leave, illness, retirement, or absence from work for other reasons.

The greatest psychological stressor, however, relates to knowing how to handle errors and production breakdowns. Check lists with procedures for standard production or for error handling do exist, but they cover only the top layer of guidelines and cannot meet all potential events. Moreover, the interpretation of rules may differ among employees who might proceed in different ways, depending on comprehension and level of education. It is known what skills are needed for a specific task; however, the competence level of the person to fulfil that task is not. Think of car driving, which is feasible for people having a driving license. This licence certificates a set of defined skills but gives no indication of the driver's quality in driving.

SOLUTION: SYSTEM ARCHITECTURE AND SYSTEM INFRASTRUCTURE

Based on the considerations presented in the Introduction and Background sections, at the eBusiness Institute of the Alpen-Adria-Universität, we developed a system family called "Knowledge for Production" (KFP), as shown in Figure 1.

ESP, the electronic shift protocol, was implemented first in 2002, and it aimed at providing an electronic shiftbook in order to resolve the afore-mentioned handover problems between shifts: ESP replaced the former handwritten shiftbook that was unstructured, mostly unreadable, and therefore very unclear.

Secondly, the module "Learning for Production" (LFP) was developed, which is to systematically collect, control, and manage the "Best Practice" knowledge about all production processes. The integration into the enterprise was supported by procedures and incentives agreed upon by the worker's council and the company's management.

The module E-Search supports a comprehensive search that makes it possible to find and combine information from all ESP and LFP knowledge bases.

During the realization phase, it turned out to be favourable to divide LFP into two subsystems with interlinked knowledge bases: the learning management system (LMS) and the knowledge management system (KMS). LMS provides a user tracking system, allowing him/her to control his/her level of education, needed courses, and important documents which he/she should or must read. In contrast to that, KMS supports a flexible unguided rummaging in any documents organized in a tree-structured knowledge base.

Figure 2 shows the user interface of the KMS application. The left-hand side features the navigational structure that maps the plant's production architecture. Documents linked to the nodes (machines/components) are shown in the middle right area. The KMS has a special access rights concept for the plant structure, that is, each user has different rights (read, write, and lock) on each subnode of the navigation tree. These rights will be inherited by the subnodes and may only be extended in the order specified. Further KMS features are the attributes of the respective document

Figure 1. KFP-architecture

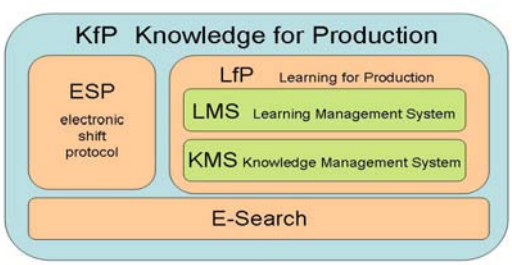
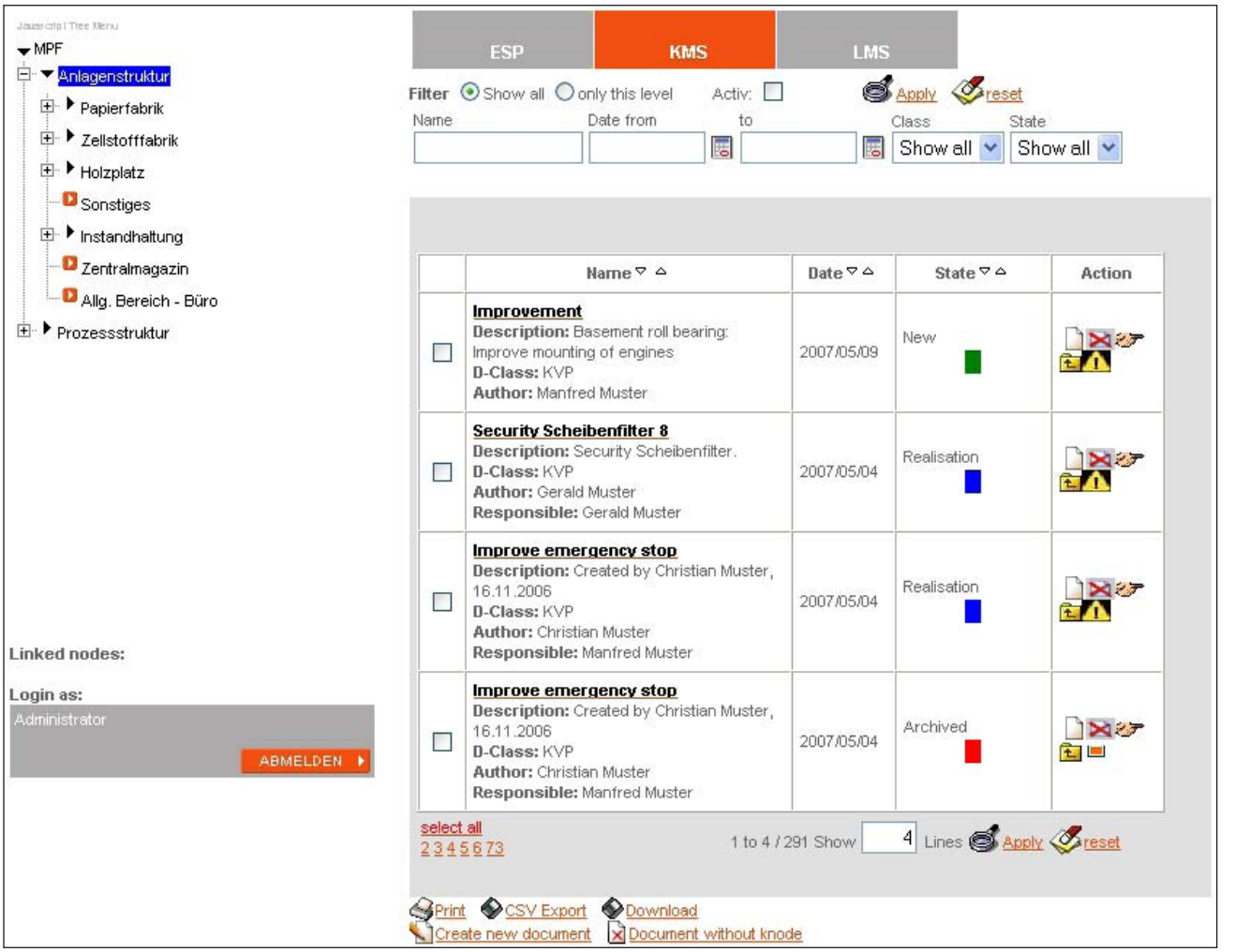


Figure 2. KMS-template



classes, which may be classified as mandatory, or optional. These fields are controlled, on the one hand, by the defined user roles and, on the other hand, by authorizations on the navigation tree. In addition, there are specific rights for the document owner, and for the particular role of the document class responsible person. There is also a freely configurable workflow for each document class. This workflow defines the order of the possible conditions of the document status as well as the visibility for the user. The workflow controls the possibility of editing metadata for documents.

FUTURE TRENDS

KM, in the production industry, mainly deals with knowledge close to the machines, production processes, and products. Mostly, this knowledge is not directly available at the shop flow workplace. Mobile, handheld, and embedded devices will help here. Potentials and limits of so-called "micro content," and the related "micro learning," are currently discussed (Hug, Lindner, & Bruck, 2006). Pervasive computing devices, combined with mechanisms for self organization, will be integrated to support pull and push strategies, and generally, will open new dimensions in KM.

The functioning of machines and processes is often hard to understand for the workers, in particular, if the relevant parts are hidden, due to casing or inaccessibility. This raises the safety for the workers but handicaps learning. In such cases, animated simulation, which also might visualize failure situations, can support knowledge acquisition. This becomes more and more good practise due to decreasing licence costs for simulation software.

Insufficient search algorithms are a bottleneck that could be overcome by the use of ontologies (Benjamins, Fensel, & Perez, 1998), which, however, require high initial development efforts and support by experts.

Another important aspect of future KM is the social collaboration: Shop floor workers usually are not trained in using communication tools like discussion boards, forums, blogs, wikis, and so on. They prefer face-to-face communication instead of fighting with problems in formulating texts. Thus, a new communication culture and speech/language processing mechanisms will have to be introduced.

CONCLUSION

Supporting and introducing KM into an organization is a sensitive issue that can hardly be standardized. Therefore, it seems to be the better way to adapt a KM system to the existing (proved) processes (best practises) instead of creating troubles by using the opposite way. People like and cultivate their habits. Changes consume energy and create fear or displeasure. Therefore, an important goal is to find the area with the highest psychological strain, and to start here by establishing the intended system step by step and department by department (always looking for the best moment to do so). The users should be within the main focus, and the system should support the user in the best way. Technical know-how of users differs, and there may be a huge gap in the ability of using KM systems. This can be balanced either by keeping the handling of the system as intuitive as possible, by training the users, and other by offering helping tools, for example, spell check or auto completion of data input, and so on.

There is no perfect strategy for KM system use and its initiation. Guidelines might help, enabling the user to intensify his/her work as he/she increases in competence and abilities.

The KFP tools we developed are now productive in practice, and the experience and expertise gained from the development are transferable to other knowledge management initiatives, especially to those in the manufacturing industry. For example, companies generally encounter knowledge-transfer-related problems in the event of annual or sick leave and retirement. Problems may also arise from limited opportunities for communication among employees because shift work reduces or even eliminates direct contact.

During content creation for KFP, experts and trainees were motivated to work together and thereby, took the advantage of collaborative processes and learning: The experts learned about the specific skills, limitations, and knowledge of the trainees thus, learning how to externalise, represent, and *transfer* ("teach") their *knowledge*. The trainees learned to ask the right questions in order to elicit the expert's knowledge.

REFERENCES

Benjamins V. R., Fensel D., & Perez A. G. (1998). Knowledge management through ontologies. In *Proceedings of the Second International Conference on Practical Aspect of Knowledge Management*.

Bryant, A. (2006). Knowledge management – The ethics of the agora or the mechanisms of the market? In *Proceedings of the 39th Hawaii International Conference on System Sciences*.

Casselman R., & Samson D. (2005). Moving beyond tacit and explicit: Four dimensions of knowledge. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences.*

Davenport, T. H., & Prusak, L. (1998). *Working knowledge. How organizations manage what they know*. Boston, MA: Harvard Business School Press.

Day, D. E. (2005). Clearing up "implicit knowledge": Implications for knowledge management, information science, psychology, and social epistemology. *Journal of the American Society for Information Science and Technology*, *56*(6), 630 – 635.

Dreyfus, H. L., & Dreyfus, S. E., & Athanasiou, T. (1988). Mind over machine: The power of human intuition and expertise in the era of the computer. New York: Free Press.

Frankl, G. (2008). *Win^n. Win-win-Konstellationen im Wissensmanagement*. Doctoral thesis, submitted.

Hug, T., Lindner M., & Bruck P. A. (Ed.). (2006). Microlearning. Emerging concepts, practices and technologies after e-learning. In *Proceedings of Microlearning 2005. Learning & Working in New Media Spaces.* Innsbruck: University Press.

Malhotra, Y., & Galletta, D. F. (2003). Role of commitment and motivation in knowldege management systems implementation. In *Proceedings of the 36th Hawaii International conference on system sciences.*

Maturana, H., & Varela, F. J. (1988). *The tree of knowledge: The biological roots of human understanding*. Boston, MA: Shambhala Publications Inc.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company*. Oxford: Oxford University Press.

North, K. (1999). *Wissensorientierte Unternehmensführung*. Wiesbaden: Gabler (2. Auflage).

Polanyi, M. (1973). *Personal knowledge. Towards a postcritical philosophy*. London: Routledge & Kegan Paul..

Polanyi, M. (1985). *The tacit dimension*. New York: Doubleday & Company.

Quinn, C. (2007). Hard fun: Cognition and emotion at play. In *The First IEEE International Workshop on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL'07)* (p. 4).

Siemens, G. (2006). Knowing knowledge. Morrisville: Lulu Enterprises.

Spitzer, M. (2005). *Vorsicht Bildschirm! Elektronische Medien, Gehirnentwicklung, Gesundheit und Gesellschaft*. Stuttgart [u. a.]: Klett.

Von Foerster, H. (1985/1999). Das Konstruieren einer Wirklichkeit. In P. Watzlawick (Ed.), *Die erfundene Wirklichkeit. Wie wissen wir, was wir zu wissen glauben? Beiträge zum Konstruktivismus* (pp. 16–38). München: Piper.

KEY TERMS

Explicit Knowledge: Knowledge that can be represented in signs (symbols, tests, and images). Thus, explicit knowledge can be stored in technical systems, and can be managed. As such, some see it as synonymous to **information**.

Information: Refers to quantitative, arranged, perceivable, and distinguishable (mental) units that have representations (signs). Long ago, Shannon defined information as the "reduction of uncertainty," Wiener made a disctinction from matter and energy.

Knowledge: Refers to quality and potential. It is related to understanding, experience, and expertise, and is gained from practising or sensing. Knowledge "can be seen as a culturally recognized set of performances called "knowing" that suggest that a person "has" the potential for further performances […] and, thus, is said to have "knowledge" of a certain form (Day, 2005, p. 631). For Siemens (2006) all knowledge is information, but NOT all information is knowledge.

Knowledge Management: Means the organization of explicit knowledge in technical systems and the enhancing of tacit knowing by supporting organizational knowledge processing.

Knowledge Management Systems (KMS): Software systems that provide features to collect, store, organize, distribute, and retrieve explicit knowledge in the form of information. KMSs are intended to support knowledge processing.

Lazy Knowledge: Simply based on facts. It lacks understanding, insight, and experience and therefore, lacks capability.

Tacit Knowing: The process of creating *knowing how*, of creating general understandings, insights, experience, and/or expertise from particular entities. *Knowing* is one psychological category of intentionality. (Day, 2005; Polanyi 1973, 1985)

Knowledge Management in E-Government

Deborah S. Carstens
Florida Institute of Technology, USA

LuAnn Bean
Florida Institute of Technology, USA

Judith Barlow
Florida Institute of Technology, USA

INTRODUCTION

Over the past decade, government has created innovative and complex systems connecting people to information by focusing on Knowledge Management (KM) practices. KM, described as the comprehensive management of an organization's expertise through collecting, categorizing and disseminating knowledge, leads to knowledge discovery through techniques such as data mining. These developments have transformed traditional access to public services into e-government. Ever increasing demand to access and information has also brought about e-government policy development challenges for integrative KM practices in public services (Riege & Lindsay, 2006). In particular, the size and complexity of governmental structures and the vast data stores have become problematic (Koh, Ryan, & Prybutok, 2005). Because government uses, collects, processes, and disseminates sensitive information containing personal, financial and medical data, it is very easy for organizations to reprocess the information and disseminate it (Hewett & Whitaker, 2002). Ebrahim and Irani (2005) state that the benefits gained by data mining and KM practices are erased when information is not viewed as confidential but instead as a commodity to be bought and sold. Therefore, e-government must uphold a higher standard of ethics in KM practices through continued development of codes of conduct and governance policies for data that build citizen trust and ensure success of e-government services and transactions (Verschoor, 2000). An excellent framework to effectively preserve this trust is a balanced scorecard (BSC), which was first introduced by Kaplan and Norton (1992, 1996a, 1996b). The framework serves to continuously improve the KM process when modified for e-government. Therefore, this chapter describes technological and organizational challenges faced by e-government in KM and retrieval and presents the BSC framework to overcome these challenges.

BACKGROUND

Government information systems have proved to be an efficient way to facilitate communication, provide information and deliver services to citizens. While database management systems lead to better data capture, storage, processing and sharing capabilities, extracting useful information from the data presents adversities for government (Robertson & Powell, 1999).

Electronic data capture and storage for later retrieval are major expenses associated with successful data mining applications (Robertson & Powell, 1999). Knowledge capture results from the use of data sources to find intelligent patterns in the data. The increase in the number existing government e-commerce sites that have the means to capture citizen data greatly reduces the cost of starting a data mining application (Tan, Steinbach, & Kumar, 2006).

Data mining software applies complex algorithms to massive data stores important for discovering relationships (Turban, Leidner, McLean, & Wetherbe, 2006). Within the public sector, data mining methods yield powerful information about the interrelationships between data elements (Robb & Coronel, 2006). Likewise, rich opportunities exist for uncovering "new" intelligence in government information.

While data mining technology advances, the potential impacts of mining citizens' private data present legal, social and ethical questions (Taipale, 2007). In addition to privacy, inherent risks of accuracy and integrity can result when data is merged from multiple sources. Safeguards for addressing privacy and integrity risks include formal codes of ethics, written ethics policies mandated employee training, and even audits that flag unauthorized access to data (von der Embse, Desai, & Desai, 2004). More specifically, legislators have enacted laws and regulations to protect citizen data (Glover & Owen, 2004). These include the Gramm-Leach-Bliley Act for financial data, Children's Online Privacy Protection Act for use/collection of child data online, the Electronic Communications Privacy Act, Privacy Act, Cable Communications Policy Act, and HIPAA regulations regarding medical information.

From a risk perspective, numerous technologies utilized in the e-government environment, as well as process/user interactions and system compatibility issues must be evaluated as potential areas of misuse or impediments. Technologically, the maze of perspective subcategories such as m-government (mobile government), u-government (ubiquitous government), and g-government (government GIS/GPS applications) provide complexities that are not much different from public sector information portals (Riley, 2007). However, the highly private nature of the data, ethical questions about its ownership, and IT infrastructure makes the stakes higher in an e-government configuration (Taipale, 2003).

From a process interaction perspective, risks can arise from the following types of systems: Government-to-Citizen (G2C), Government-to-Business (G2B), Government-to-Government (G2G) or Government-to-Employees (G2E) (Turban et al., 2006). For example, G2C systems support the majority of citizen data captured, including online systems for driver license renewal, vehicle tag renewal, voter registration, social benefits management, court records, and state property appraisal information. Despite the highly personal and private nature of G2C Web site exchanges, the G2B systems, which interface with business information systems, open even a more complex Pandora's Box of data security and data access solutions (Sagheb-Tehrani, 2007). G2G systems linking governments face problems with inconsistent policies/laws, intergovernment or interagency politics, language barriers, and customs (Jing & Pengzhu, 2007). Finally, G2E systems face privacy and safety issues of providing employees with access to employee benefit information and communication portals with human resources departments (Pardhasaradhi & Ahmed, 2007).

Besides technical and structural risks, the historic development of many e-government information systems in a "vacuum" rather than through coordinated efforts have led to incompatibilities with other systems and data redundancy (i.e., the same data is stored in more than one place). These incompatibilities result in citizen and legislative outrage, as well as increased costs, poor performance, and a lack of sharable components and data (Park & Ram, 2004).

KNOWLEDGE MANAGEMENT AND RETRIEVAL IN E-GOVERNMENT

Now that a discussion of technological and organizational challenges with government information systems faced by e-government in KM and retrieval has been presented, this section will discuss how the BSC can be used to ensure trusted e-government services and transactions.

Because uses within e-government environments vary widely (including even debt or courtroom trial management and litigation support), it is important to consider the skill-set and priority diversity of users, as well as organizational core capabilities. Integrated e-government systems must be able to manage data/document capture, convert data to digital format, allow for Internet/intranet document publishing, track correspondence and electronic information distribution, and complete action tracking functions (such as suspenses and corrections) in a systematic manner.

The balanced scorecard, first introduced by Kaplan and Norton (1992, 1996a, 1996b), can be modified for e-government use to provide one of the best frameworks to enhance the integrated retrieval functions of KM systems. One of the key advantages of this approach is that it provides order to the unstructured nature of electronic data or text mining through a systematic process of extracting knowledge. Secondly, the BSC modification continues to support, improve, and add value to the productivity of data warehouse retrieval capabilities. Using four perspectives (finance, customers, internal processes and training growth), the BSC promotes continuous improvement at each phase of the KM process for e-government. As Figure 1 shows, each step of the KM strategy addresses these perspectives and continually transforms the review and audit function to evaluate knowledge created.

Basically, the finance perspective addresses timely and accurate funding data, as well as issues of risk assessment and cost-benefit data associated with the KM system (Arveson, 1998). The kinds of customers, processes used to service these groups, value-added service delivery, and satisfaction metrics are all part of the customer perspective (Davison, Wagner, & Ma, 2005). The internal business process perspective focuses on how well the organization is performing its mission mission-oriented processes (i.e., specific and unique functions of the e-government environment) and support processes (i.e., more repetitive functions benchmarked using generic metrics) (Patton, 2007). Finally, the training growth perspective highlights employee training and attitudes toward the organizational culture can lead to strengthened user communications for problem solving (Wu, 2007).

The knowledge audit is both the beginning and ending point that closes the loop when establishing effective e-government KM strategies. At a minimum, the knowledge audit should: (1) determine what target areas of knowledge should be audited and what limitations exist; (2) identify benchmarks and how results will be measured and tracked against this reference; and (3) evaluate existing/missing knowledge in target areas (Daghfous & Al-Nahas, 2006). After identifying the parameters of the knowledge audit, the strategic steps will of KM will examine perspectives of reducing costs (financial), enhancing customer value (customer), leveraging advantageous organizational processes (internal processes), and promoting change in the KM system through continuous learning cultures (training growth). Figure 2 displays examples of each perspective that could be incorporated when developing the organization's

KM strategy. Realizing the advantages of this framework requires constant reinforcement of a "knowledge pull" position rather than a top-down "knowledge push" position. In the more successful "knowledge pull" model, e-government can tap into its intellectual resource strengths by rewarding its users for sharing, seeking and creating knowledge rather than creating frustrating experiences of pushing information where it is needed (Hauschild, Licht, & Stein, 2001).

FUTURE TRENDS

With the capabilities of computer hardware doubling every 18 months since the first computer (UNIVAC) was delivered to the U.S. Census Bureau in 1951, data mining in e-government without a doubt will continue to grow (Gray, Liu, Nieto-Santisteban, DeWitt, & Heber, 2005). Data mining algorithms that were once thought to be intractable can now easily be run on today's hardware. University and industry researchers continue to find new ways to automatically merge data from disparate sources while preserving data integrity and new algorithms to provide meaning within data patterns.

Unfortunately, these logistical and technical improvements do not address the legal, moral and ethical concerns. When data mining tools are able to discover sensitive e-government intelligence that is dangerous or hurtful, then citizens, legislators and the legal profession may choose to exercise more controls or pass legislative safeguards (Taipale, 2007). One safeguard is the newly revised Federal Rules of Civil Procedures (FRCP), which went into effect on December 1, 2006, that address which data is the relevant, discoverable, and deemed preservable in a court of law (U.S. Judicial Conference, 2005). While much of the data captured by e-government systems is not only of interest to the public but is also valuable to other government agencies and businesses (Pollach, 2007), rules on how to control access to this "private" data vary from state-to-state and country-to-country. For example, many state agencies provide free and easy access to information about traffic and

Figure 1. Knowledge management process for e-government

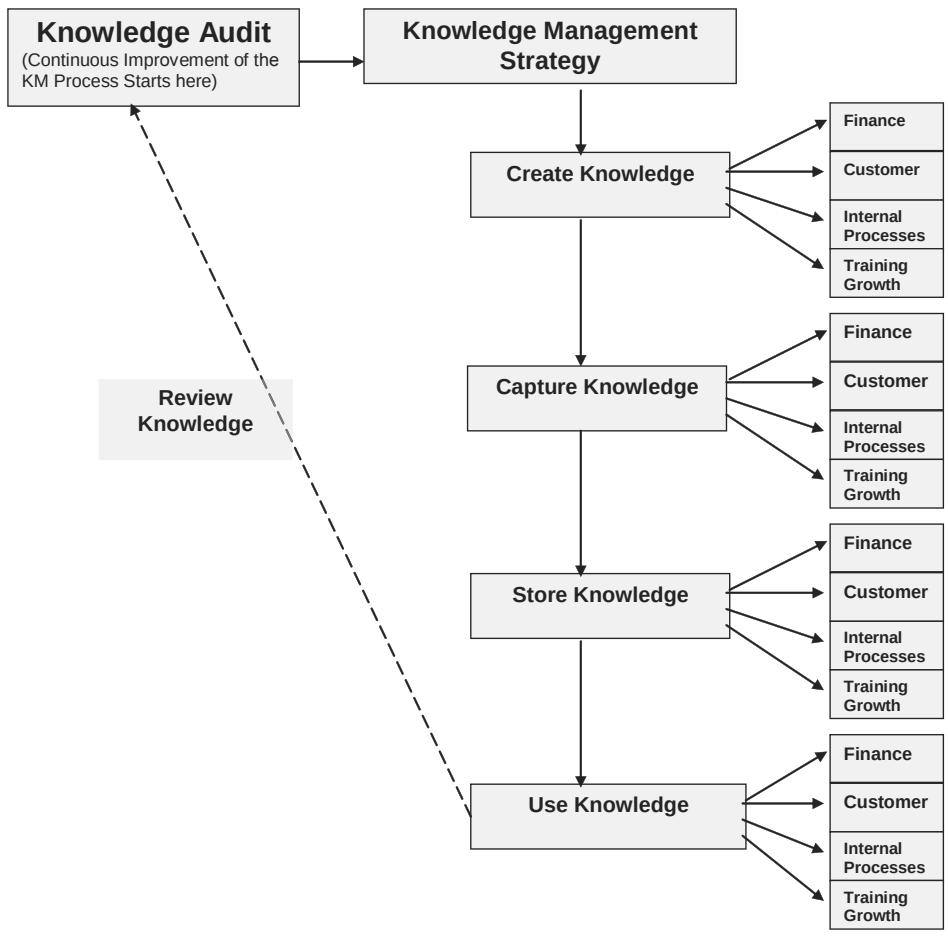


Figure 2. Creation, capture, store and use knowledge perspective in organization's KM strategy

CREATION OF KNOWLEDGE	
Finance	*Are there sufficient financial resources to support the information technology for the entire KM system?*
Customer	*What information could be presented to customers as value-added knowledge, by converting data currently gathered?*
Internal Processes	*Who are the people who play knowledge roles within the organization? Are they knowledge facilitators, curators, or engineers? Is an analysis of the interplay among interorganizational technology, social networks and competitive dynamics important for the organization?*
Training Growth	*How is knowledge created within the organization?*

CAPTURE KNOWLEDGE	
Finance	*How does unnecessary information capture affect service costs?*
Customer	*What knowledge hierarchy exists for serving customers in terms of the complexity, depth, security needs, etc. of captured data?*
Internal Processes	*Would experimental investigation of alternative navigation structures yield benefits? What repositories contain explicit knowledge (i.e., formal codified knowledge documented in reports, papers, rules, laws, patents, formulas, or books)?*
Training Growth	*Are there opportunities through training to advance tacit knowledge capture (i.e., informal uncodified knowledge that resides in people's heads over years of experience)?*

STORE KNOWLEDGE	
Finance	*What is the cost/benefit of storing certain data?*
Customer	*Which knowledge assets, if lost, would threaten users or the organization's existence?*
Internal Processes	*Who has access to the internal and external information sources?*
Training Growth	*By instituting certain employee policies (enhanced by training), can the organization eliminate valuable storage regarding electronic information?*

USE KNOWLEDGE	
Finance	*From a financial standpoint, what is the value of KM in productivity, agility, innovation and reputation for the organization?*
Customer	*Do the organizational structures and incentives foster knowledge sharing within and outside the organization?*
Internal Processes	*What processes/tool support is required to acquire, refine, index, store, retrieve, disseminate, and present knowledge? Are scheduled reviews or controls used to promote development and validation of the knowledge chain model in the organization?*
Training Growth	*What mechanisms exist for cooperative sharing among knowledge workers? What learning effects would benefit multiparticipant collaborative systems for the organization?*

criminal violations, vehicle registration information, and property data. While citizens may find it useful to review the accuracy of their own traffic, court, or property ownership information, there are risks of commercial exploitation of e-government data which can be used to identify lucrative customer relationships. Likewise, criminals may enjoy access to citizens' addresses, phone numbers, automobile data, arrest records, distance from schools and playgrounds, credit reports, and more.

One of the largest data-selling companies is Choice Point. When the Homeland Security Administration wanted to test the value of data mining in identifying post-9/11 terrorists, the FBI turned to Choice Point for help (Harris, 2005). Using data from e-government sources plus private databases owned by Choice Point, the company used propriety data mining methods in search of terrorists. After legislators and citizens expressed ethical concerns about citizen privacy and for-profit company involvement in the war on terrorism, the project was quickly terminated. While it is unclear if efforts were abandoned due to ethical concerns or unsuccessful results in terrorist identification, the point is that data-mining's value for e-government will continue to grow, be tested, and targeted for more applications of rich mine-able data (Strohm, 2005; Taipale, 2003, 2007).

CONCLUSION

Social, ethical and legal challenges in connection with information transformations are causing organizations, including e-governments, to examine privacy, piracy, safety, data security, data integrity, competence, honestly, loyalty and fairness (Chow, 2001; Himma, 2007). Legislators and the legal profession have stepped in to control the dissemination of information in light of social and ethical voids associated with limited or deficient knowledge policy development (Hewett & Whitaker, 2002). Organizations need to carefully review existing legislation as well as understand the legal implications of it and may benefit from seeking legal advice. Performing internal audits and updating organization's policies and procedures are recommended as new legislation evolves. For example, new information privacy legislation should alert an organization to ensure that their policies, procedures, controls, and systems are in compliance and receive continuous scrutiny. To assist with KM in e-government, a modified BSC provides a first step toward establishing a effective KM strategies (Kaplan & Norton, 1992, 1996a, 1996b). Using the BSC, continual improvement of the knowledge management process for e-government can be accomplished using four key perspectives. By transforming strategies into objectives, overcoming the barriers to knowledge management and retrieval in e-government can reduce costs (financial), enhance customer value (customer), leverage advantageous organizational processes (internal processes), and promote KM system change through continuous learning (training growth).

REFERENCES

Arveson, P. (1998). *What is the balanced scorecard?* Balanced Scorecard Institute. Retrieved May 29, 2008, from http://www.balancedscorecard.org/basics/bsc1.html

Chow, W.S. (2001). Ethical belief and behavior of managers using information technology for decision making in Hong Kong. *Journal of Managerial Psychology, 16*(4), 258-267.

Daghfous, A., & Al-Nahas, N. (2006). The role of knowledge and capability evaluation in e-business strategy: An integrative approach and case illustration. *S.A.M. Advanced Management Journal (Spring), 7*1(2), 11-22.

Davison, R.M., Wagner, C., & Ma, L.C.K. (2005). **From government to e-government: A transition model.** *Information Technology & People, 18*(3), 280-299.

Ebrahim, Z., & Irani, Z. (2005). E-government adoption: Architecture and barriers. *Business Process Management Journal, 11*(5), 58-611.

Glover, B., & Owen, E. (2004). Conference on computers, freedom and privacy: Going strong in its 14th year. *Library Hi Tech News, 21*(7), 5-10.

Gray, J., Liu, D., Nieto-Santisteban, A., Szalay, A., DeWitt, D., & Heber, G. (2005). Scientific data management in the coming decade. *ACM SIGMOD Record, 34*(4), 34-41.

Harris, S. (2005). *FBI, Pentagon pay for access to trove of public records*. Government Executive. National Journal. Retrieved May 29, 2008, from http://www.govexec.com/story_page.cfm?articleid=32802&dcn=e_gvet

Hauschild, S., Licht, T., & Stein, W. (2001). Creating a knowledge culture. *The McKinsey Quarterly, 1*, 74-81.

Hewett, W.G., & Whitaker, J. (2002). Data protection and privacy: The Australian legislation and its implications for IT professionals. *Logistics Information Management, 15*(5/6), 369-76.

Himma, K.E. (2007). Foundational issues in information ethics. *Library Hi Tech, 25*(1), 79-94.

Jing, F., & Pengzhu, Z. (2007, May 20-23). A case study of G2G information sharing in the Chinese context. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, Philadelphia, PA, (Vol. 228, pp. 234-235). ACM International Conference Proceeding Series, Digital Government Research Center.

Kaplan, R., & Norton, D. (1992). Putting the balanced scorecard to work. *Harvard Business Review, 71(*5), 134-47.

Kaplan, R.S., & Norton, D.P. (1996a). Linking the balanced scorecard to strategy. *California Management Review, 39*(1), 53-79.

Kaplan, R.S. and Norton, D.P. (1996b). Using the balanced scorecard as a strategic management system. *Harvard Business Review, 74*(1), 75-85.

Koh, C.E., Ryan, S., & Prybutok, V.R. (2005). Creating value through managing knowledge in an e-government to constituency environment. *The Journal of Computer Information Systems, 45*(4), 32-42.

Pardhasaradhi, Y., & Ahmed, S. (2007, December 10-13). Efficiency of electronic public service delivery in India: Public-private partnership as a critical factor. In *Proceedings of the 1st International Conference on Theory and Practice of Electronic Governance, ICEGOV'07*, Macao, China, (Vol. 232, pp. 357-365). New York: ACM.

Park, J., & Ram, S. (2004). Information systems interoperability: What lies beneath? *ACM Transactions on Information Systems, 22*(4), 595-632.

Patton, J. R. (2007). Metrics for knowledge-based project organizations. S.A.M. *Advanced Management Journal, 72*(1), 33-44.

Pollach, I. (2007). What's wrong with online privacy policies?. *Communications of the ACM 50*(9), 103-108.

Riege, A., & Lindsay, N. (2006). Knowledge management in the public sector: Stakeholder partnerships in the public policy development. *Journal of Knowledge Management, 10*(3), 24-39.

Riley, T.B. (2007). Strategies for the effective implementation of e-government projects. *Journal of Business and Public Policy, 1*(1). Retrieved May 29, 2008, from http://www.jbpponline.com/article/view/1024/817

Robb, P., & Coronel, C. (2006). *Database systems: Design, implementation, and management* (7th ed., p. 530). Boston: Course Technology Thomson Learning.

Robertson, S., & Powell, P. (1999). Exploiting the benefits of Y2K preparation. *Communications of the ACM, 42*(9), 42-48.

Sagheb-Tehrani, M. (2007). Some steps towards implementing e-government. *SIGCAS Computers and Society, 37*(1), 22-29.

Strohm, C. (2005). *Federal data-mining efforts fail to fully safeguard privacy, GAO says.* Government Executive. Retrieved May 29, 2008, from http://www.govexec.com/dailyfed/0805/083005c1.htm

Taipale, K.A. (2003). Data mining and domestic security: Connecting the dots to make sense of data. *Columbia Science & Technology Law Review, 5(*2). Retrieved May 29, 2008, from http://www.stlr.org/html/volume5/taipaleintro.php

Taipale, K.A. (2007, January 10). *The privacy implications of government data mining programs*. Data mining testimony before the U.S. Senate Judiciary Committee. Retrieved May 29, 2008, from http://data-mining-testimony.info/

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining* (p. 2). Upper Saddle River, NJ: Addison-Wesley.

Turban, E., Leidner, D., McLean, E., & Wetherbe, J. (2006). *Information technology for management: Transforming organizations in the digital economy* (5th ed., pp. 161-163). Edison, NJ: John Wiley & Sons.

U.S. Judicial Conference. (2005, September). *Report of the Civil Rules Advisory Committee 40* (amended July 25, 2005). Committee on the Rules of Practice & Procedure of the Judicial Conference of the U.S., Summary of Committee on Rules of Practice and Procedure: Agenda E-i8 app. C.

Verschoor, C. (2000). Can an ethics code change behavior?. *Strategic Finance, 82*(1), 26-28.

von der Embse, T.J., Desai, M., & Desai, S. (2004). How well are corporate ethics codes and policies applied in the trenches? *Information Management & Computers Security, 12*(2), 146-153.

Wu, E. (2007). A balanced scorecard for the People Development Function. *Organizational Development Journal, 25*(2), 113-116.

KEY TERMS

Balanced Scorecard: The Balanced Scorecard (BSC) was developed by Kaplan and Norton in 1992 to examine how an organization is using the activities and processes to meet its strategies, using the four key perspectives of finance, customers, internal processes, and training growth.

Data Mining: The use of algorithms to automatically search through massive data stores from different sources in order to find unknown patterns and interrelationships that ascribe meaning to the data.

Data Redundancy: The same data is stored in more than one place.

Knowledge Audit: An examination of KM systems for purposes of evaluating knowledge components against targeted benchmarks.

Knowledge Management: The comprehensive management of an organization's expertise that consists of collecting, categorizing and disseminating knowledge.

Knowledge Management (KM) Strategy: A knowledge management strategy identifies the key needs, issues, and components within an organization and provides a framework for addressing these. Specifically, knowledge management strategy components include how an organization creates, captures, stores and uses knowledge.

Knowledge Pull: Knowledge pull taps into the intellectual strengths of its users, who share, seek, and create knowledge efficiently and effectively at the optimal cost.

Knowledge Push: Knowledge push is typically associated with a traditional top-down approach to managing knowledge. In other words, the knowledge delivery is not initiated by the receiver and lacks collaborative potential of knowledge pull.

Knowledge Management Systems Acceptance

Fredrik Ericsson
Örebro University, Sweden

Anders Avdic
Örebro University, Sweden

INTRODUCTION

Knowledge management is a set of systematic actions that organizations can take to obtain the greatest value from the knowledge available to it (Davenport & Prusak, 1998). Systematic means that knowledge management is made up of intentional actions in an organizational context. Value means that knowledge management is measured according to how knowledge management projects contribute to increased organizational ability (see for example Prieto & Gutiérrez, 2001; see Goldkuhl & Braf, 2002, on the subject of organizational ability). The motivation for knowledge management is that the key to competitive advantage for organizations in today's business world is organizations' ability to manage knowledge (Nonaka & Takeuchi, 1995; Davenport & Prusak, 1998). Knowledge management as an intentional and value-adding action is not easy to accomplish in practice (Scarbrough & Swan, 1999). Scarbrough and Swan (1999) present several case studies in knowledge management, successful and unsuccessful in their respective knowledge management projects. A major point and lessons learned from the case studies is that prevalent approaches in knowledge management overstate technology and understate how technology is implemented and applied.

To succeed with knowledge management, encompassing development of information technology-based information system, some requirements have to be fulfilled. An important aspect in the development process is system acceptance. Implementation is at large a process of acceptance. Implementation is the process where the system becomes an integrated part of the users' or workers' work practice. Therefore implementation is essential to make a knowledge management project successful in order attain an increased organizational ability and to succeed with knowledge management.

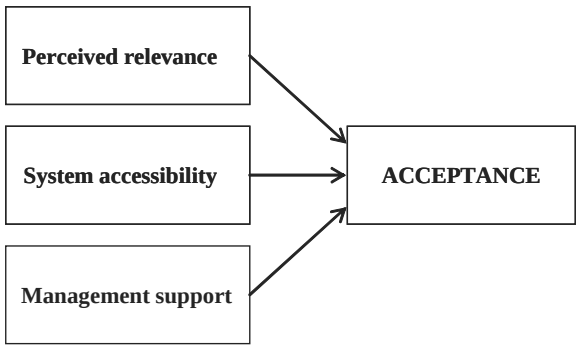
ISSUES OF KNOWLEDGE MANAGEMENT: SYSTEMS AND ACCEPTANCE

In this section we provide broad definitions and discussion of the topics to support our positions on the topics of knowledge management and systems acceptance.

MANAGING KNOWLEDGE

Work in knowledge management has a tendency to omit social or technological aspects by taking on one of two perspectives on knowledge management, the anthropocentric or the technocratic view (Sveiby, 2001; Swan, 1999). The anthropocentric and the technocratic views represent two contradictory views on knowledge management and can be summarized as technology can or technology cannot. The gap between the anthropocentric and technocratic view depends on a difference of opinions concerning the notion of knowledge. The technocratic view conceives knowledge to be some organized collection of data and information, and the anthropocentric view conceives knowledge to reside in humans, not in the collection (Churchman, 1971; Meredith & Burstein, 2000). Our conception of knowledge is that of the anthropocentric view. Taking on an anthropocentric view on knowledge management does not mean that we discard knowledge management technologies; we rather take on a balanced view on the subject. Information technology can support knowledge management in an organization through a number of different technological components, for example intranets, extranets, data warehouses, and database management systems (Borghoff & Pareschi, 1998; Tiwana, 2000; Ericsson & Avdic, 2002). The point in taking on an anthropocentric view of knowledge management is not to lose sight of the knower who gives meaning to the information and data found in IT-based knowledge management systems.

Figure 1. Requirements of Acceptance Model (Ericsson & Avdic, 2003)



KNOWLEDGE MANAGEMENT SYSTEMS

Information systems can include either operative or directive and decision support information (Langefors, 1966; Yourdon, 1989). Operative systems provide system users with information necessary in workers' daily work, while directive and decision support systems provide system users with information that improves the quality of decisions workers make in daily work. Knowledge managements systems are systems developed to manage knowledge directly or indirectly to give support for an improved quality of a decision made in workers daily work, and as an extension, an increased organizational ability. A knowledge management system typically includes directive information, for example in guiding a user's choice in a specific work situation. Such systems are often optional in the sense that users can deliberately refrain from using the system and/or refrain from taking the directed action. Accordingly, user acceptance is crucial for the degree of usage of knowledge management systems.

ACCEPTANCE OF TECHNOLOGICAL SYSTEMS

Technology acceptance has been subject of research by, for example, Davis, Bagozzi, and Warshav (1989), who developed the well-known Technology Acceptance Model (TAM) and later a revised version of the original model, TAM2 (Venkatesh & Davis, 2000). TAM is an explanative model explaining user behavior of computer technologies by focusing on perceived ease of use, perceived usefulness, attitude towards use, and behavioral intentions as determinants of user behavior. TAM2 is an extension of the original model including external factors related to perceived usefulness.

The framework for system acceptance, Requirements of Acceptance Model (RAM) have some resemblances with TAM and the later TAM2. RAM is in comparison with TAM descriptive in nature. Workers' work practice is treated as an integrated element of RAM, compared with not being treated as a determinant of system use in the original TAM and as an external factor in TAM2. Further, RAM covers acceptance of knowledge management systems, and TAM/TAM2 cover a broad range of computer technologies. RAM systematically acknowledges factors important in implementation of knowledge management systems to gain acceptance of such systems.

REQUIREMENTS OF THE ACCEPTANCE MODEL

We perceive acceptance to be a function of perceived relevance, systems accessibility, and management support. Together these elements constitute our framework RAM. In this section we present the requirements of acceptance in RAM. The Requirements of Acceptance Model is illustrated in Figure 1.

PERCEIVED RELEVANCE

The workers, who are to use the system, must perceive the knowledge management system as relevant. Since it is possible for workers to work without using the system, it has to be obvious that usage of the system implies adding value to the work result. An additional aspect of relevance related to perceived relevance is how the system should be integrated in running work, that is, to make the system an integrated part of the workers' work practice.

In summary, perceived relevance is about workers, who are to use the system, perceiving the system as (Ericsson & Avdic, 2003)

- adding value to the work results; and
- being integrated in running work.

ACCESSIBILITY

To obtain acceptance of knowledge management systems, accessibility has to be satisfactory. It must be accessible to the workers who are to use the system. Accessibility is a question of who is to be the user (type of workers concerning organizational position), what action and work the system is to support (daily work, product development, innovation, etc.), where users get access to the system (the physical access), when the system is ready to use, and how the system's interface fulfills the goal of the system.

In summary, systems accessibility is about (Ericsson & Avdic, 2003):

- knowing who the user is;
- systematizing the actions workers perform in the work practice the system is to support;
- deciding the system's physical access;
- securing a certain degree of usage before the system is put into operation; and
- ensuring the system's design meets the goals of the system.

MANAGEMENT SUPPORT

Management support is vital according to many models on information systems development, especially when the system is a directive/decision support system (Yourdon, 1989). Knowledge management systems are typically directive systems, and workers have a choice in deciding whether to use the system or not. Management support is important to stress the value for workers to use the system and to make conditions for workers to do so.

DEVELOPMENT IS A PROCESS OF ACCEPTANCE

There must be a fit between workers' work practice and technology to get acceptance of knowledge management systems. The technology used to create a knowledge management system must fit the actions workers perform in their work practice. On an overall level there must be a fit between technology and actions performed by individual workers, and between individual workers and the organization as a whole, thus forming a coherent whole. It is in the development of knowledge management systems that the requirements of acceptance are fulfilled. A common conception concerning information systems development is that it constitutes analysis, design, construction, and implementation of information systems (Hirschheim, Klein & Lyytinen, 1996).

The groundwork for acceptance is made during the design, but foremost when implementing the system. Workers who are to use the system should be engaged at an early stage of the development process. The point of including workers at an early stage is to acquaint users with the system and the purpose of the system. Further, this is an opportunity for workers to influence the system's design and content. The most prominent aspect addressed when involving workers at an early stage is that of choosing and determining the meaning of crucial concepts managed by the system. Crucial concepts managed by the system are the knowledge represented in the system, and by determining concepts, knowledge represented

Table 1. Summary of RAM (Ericsson & Avdic, 2003)

Perceived relevance—Workers, who are to use the system, have to perceive the system as:
- Adding value to work results
- Being integrated in running work

Systems accessibility—System accessibility is about:
- Knowing who the user is
- Systematizing actions workers perform in the work practice the system is to support
- Deciding the physical location where users get physical access to the system
- Securing usage of the system before it is put into operation
- The systems' design must meet up to the goals of the system

Management support—Fundamental because management authorizes development of systems

in the system takes on a systematized character. Further, by involving the workers in the process of choosing and determining the meaning of crucial concepts managed by the system, the knowledge represented in the system does not lose its origin or meaning. The point is to keep the knowledge represented in the system within a frame of understanding or meaning, as perceived by workers. A knowledge management systems should be seen as a tool developed to support workers in learning and acquiring knowledge about actions taking place at work. This requires closeness between how concepts are perceived by workers and how such concepts are represented in a system.

FUTURE TRENDS

Research on technology acceptance (i.e., Davis et al., 1989; Venkatesh & Davis, 2000) has focused on user behavior of computer technologies. RAM is developed for and is used to assess acceptance of knowledge management systems. Acceptance has not been a crucial issue within the knowledge management area. A problem with knowledge management systems is that they work in theory, but seldom in practice (Wickramasinghe, 2003). A contributing factor to that picture may very well be that of having overlooked usage-related problems connected to knowledge management systems. In that sense, knowledge management systems acceptance can be expected to be an area for further research in the future.

CONCLUSION

Acceptance of knowledge management systems is a function of perceived relevance, systems accessibility, and management support. Together these elements constitute our framework RAM. RAM is summarized in Table 1.

The Requirements of Acceptance Model point towards several important aspects concerning relevance, accessibility, and support. The groundwork for system acceptance is the development process. Development is very much a process of acceptance as a process of developing the system itself. Through requirements of acceptance, knowledge management systems can remain and continue to be a contributing factor for the organization's ability to do business.

REFERENCES

Borghoff, U.M. & Pareschi, R. (Eds.). (1998). *Information technology for knowledge management*. Berlin, Heidelberg: Springer-Verlag.

Churchman, C.W. (1971). *The design of enquiring systems: Basic concepts of systems and organization*. New York: Basic Books.

Davenport, T. & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston: Harvard Business School.

Davis, F.F., Bagozzi, R.P. & Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Ericsson, F. & Avdic, A. (2002). Information technology and knowledge acquisition in manufacturing companies: A Scandinavian perspective. In E. Coakes, D. Willis & S. Clarke (Eds.), *knowledge management in the socio-technical world. The grafitti continues.* London: Springer-Verlag.

Ericsson, F. & Avdic, A. (2003). Knowledge management systems acceptance. In E. Coakes (Ed.), *Knowledge management: Current issues & challenges* (pp. 39-51). Hershey, PA: Idea Group Publishing.

Goldkuhl, G. & Braf, E. (2002). Organisational ability: Constituents and congruencies. In E. Coakes, D. Willis & S. Clarke (Eds.), *Knowledge management in the socio-technical world. The grafitti continues* (pp. 30-42). London: Springer-Verlag.

Hirschheim, R., Klein, H.K. & Lyytinen, K. (1996). Exploring the intellectual structures of information systems development: A social action theoretic analysis. *Accounting, Management & Information Technology, 6*(1/2), 1-64.

Langefors, B. (1966). *Theoretical analysis of information systems*. Lund: Studentlitteratur.

Meredith, R. & Burstein, F. (2000). Getting the message across with communicative knowledge management. *Proceedings of the Australian Conference on Knowledge Management and Intelligent Decision Support* (ACKMID'2000) (pp. 43-55). Melbourne: Australian Scholarly Publishers.

Nonaka, I. & Takeuchi, H. (1995). *The knowledge creating company: How Japanese companies create the dynamics of innovation*. New York: Oxford University Press.

Prieto, I.M. & Gutiérrez, E.R. (2001). A contingency perspective of learning and knowledge management in organizations. In D. Remenyi (Ed.), *Proceedings of the 2nd European Conference on Knowledge Management* (pp. 487-502). Slovenia: Bled School of Management.

Scarbrough, J. & Swan, J. (Eds.). (1999). *Case studies in knowledge management*. London: Institute of Personnel and Development.

Sveiby, K.-E. (2001, April). *What is knowledge management?* Retrieved June 28, 2002, from *www.sveiby.com.au.*

Swan, J. (1999). Introduction. In J. Scarbrough & J. Swan (Eds.), *Case studies in knowledge management*. London: Institute of Personnel and Development.

Tiwana, A. (2000). *The knowledge management toolkit. Practical techniques for building a knowledge management system*. Upper Saddle River, NJ: Prentice-Hall.

Venkatesh, V. & Davis, F.D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46,* 86-204.

Yordon, E. (1989). *Modern structured analysis*. Englewood Cliffs, NJ: Prentice-Hall.

Wickramasinghe, N. (2003). Do we practice what we preach? Are knowledge management systems in practice truly reflective of knowledge management systems in theory? *Business Process Management Journal, 9*(3), 295-316.

KEY TERMS

Anthropocentric View of Knowledge: Knowledge resides in humans.

Information Systems Development: Constitutes analysis, design, construction, and implementation of information systems.

Knowledge: Knowledge is personal and talked about and may thus be public and shared among a group of people who have a common frame of reference, providing means for people to make sense of and apply knowledge in practice.

Knowledge Management: The name given to the set of systematic actions that an organization can take to obtain the greatest value from the knowledge available to it.

Knowledge Management Systems: Typically, directive systems developed to manage knowledge directly or indirectly to give support for an improved quality of a decision made in workers' daily work, and as an extension, an increased organizational ability.

Perceived Relevance: Workers who are to use the system perceive the system as adding value to the work results and being integrated in running work.

Systems Acceptance: A function of perceived relevance, systems accessibility, and management support.

Systems Accessibility/Development: Knowing who the user is, systematizing the actions workers perform in their work practice the system is to support, deciding the system's physical location, securing a certain degree of usage before the system is put into operation, and ensuring the system's design meets the goals of the system.

Technocratic View of Knowledge: Knowledge is an organized collection of data and information.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1778-1782, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Knowledge Management Technology in Local Government

Meliha Handzic
Sarajevo School of Science and Technology, Sarajevo

Amila Lagumdzija
Sarajevo School of Science and Technology, Sarajevo

Amer Celjo
Sarajevo School of Science and Technology, Sarajevo

INTRODUCTION

Increased interaction, interdependency and volatility on a global scale are rapidly changing local governments' external environment, their community characteristics, and their organisational orientation. In circumstances of high uncertainty and ambiguity, the success of local governments depends to a greater extent on how well they utilise knowledge resources in adjusting to contextual changes. This requires special attention to knowledge management (KM). The major challenge for KM in local government is to foster the development of an enriched knowledge base that will enable local actors to better deal with adjustment and development issues of importance to their communities (Anttiroico, 2006). The purpose of this article is to address technical issues in organisational KM.

Referring to the theoretical work by Handzic (2004), the article considers the role of various information and communication technologies (ICT) in facilitating the processes in which knowledge is created, transferred and utilised in local governments. Findings reported in the article are part of an ongoing research project into the adoption of KM principles and practices in public sector organisations in Bosnia and Herzegovina (BiH). The role of ICT in local government KM solutions addressed in this article is only one of several aspects covered by the research project. Further project details can be obtained elsewhere (Handzic, Lagumdzija, & Celjo, 2007).

BACKGROUND

The spectrum of views on the role of ICT in KM ranges from those that see knowledge as a uniquely human concept and consider that KM has little to do with technology, to those that see knowledge as an object and therefore KM as being mostly about technology (Swan, 2003). The integrated approach advocated by Handzic (2004) bridges the artificial divide between two extreme perspectives by considering KM as a socio-technical phenomenon with both technology and people playing an important role.

Within the integrated framework, technology is placed among major influencing factors on knowledge processes. The functionalities of ICT are perceived as significant in shaping organisational efforts for knowledge creation, transfer and utilisation, and thus for organisational learning, improvement and innovation. In order to better understand and appreciate the importance of technology in KM, this section surveys some ICT-based KM initiatives deployed in firms and their roles in supporting knowledge processes.

The KM literature offers a number of useful classifications of ICT tools for KM based on their functions and techniques (Binney, 2001; Tsui, 2003). Most recently, Handzic and Zhou (2005) developed a typology of KM technologies that includes seven categories based on the distinction of KM processes they support. They include: knowledge storage, access, search/retrieval, sharing/delivery, discovery/visualisation, utilisation and platform technologies. These categories are used to frame the discussion about the applications of ICT in KM in this article.

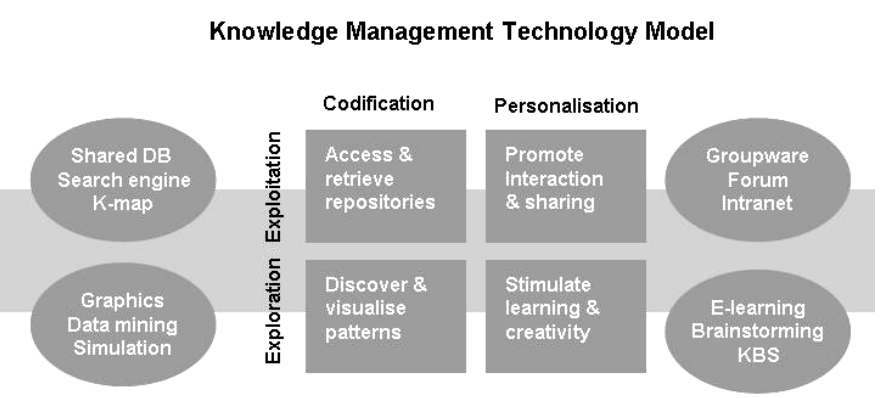
- *Knowledge storage technologies* cover databases, textbases, data warehouse, data marts and various multimedia systems used to capture and store organisational knowledge with the objective to enhance organisational memory and to provide broader access to knowledge resources (Alavi & Leidner, 2001). These technologies organise and make available knowledge in a variety of representational formats, and store current and retain historical and cross-functional aspects of knowledge.
- *Knowledge access technologies* such as knowledge maps, knowledge directories and yellow pages are tools used to improve access to knowledge stored in

knowledge repositories or facilitate knowledge transfer among individuals. These systems act as navigation aids that help knowledge seekers to quickly locate important explicit and tacit knowledge sources (Wexler, 2001).

- *Knowledge search/retrieval technologies* including search engines and intelligent agents are tools used to locate internal knowledge on intranets or external knowledge on the Internet, with the objective of increasing the speed and accuracy of knowledge search. These software programs enable access to unstructured information and can carry out search tasks with some degree of independence and autonomy (Tsui, 2003).
- *Knowledge sharing/delivery technologies* represent various applications that use ICT to facilitate peer-to-peer communication and knowledge sharing (Hansen, Nohria, & Tierney, 1999). E-mail systems, electronic bulletin boards, whiteboards, electronic forums, videoconferencing, voice mail, and groupware are some examples of such tools used to provide the right knowledge to the right person at the right time. Specialised groupware applications also offer support for collaborative processes.
- *Platform technologies* comprise net-based tools such as internet, intranets, extranets and portals that are used to provide connectivity and support knowledge sharing inside and outside the organisation. They are also commonly used by organisations to construct a single point of access to multiple sources of internal and external knowledge (Awad & Ghaziri, 2004). In general, they provide network platforms for knowledge collection, communication and analysis.
- *Knowledge discovery/visualisation technologies* describe applications that look for hidden patterns in data in order to discover and make visible previously unknown patterns (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Data mining, statistical tools, graphical representation and simulation technologies are technologies that use complex and sophisticated algorithms to extract and visualise new knowledge with a goal of supporting improvements and changes to the way knowledge is used, shared and transferred.
- *Knowledge utilisation technologies* such as knowledge-based systems, workflow systems, expert systems, rule induction and decision trees are tools used to enable knowledge workers to apply the best decision-making expertise and improve performance (Becerra-Fernandez, Gonzales, & Sabherwal, 2004). These systems harness technology by imbedding knowledge into work processes, with the objective of facilitating knowledge integration and application. They can also enable people to learn more easily through experience.

The classes of technologies illustrated above are not mutually exclusive. They can serve multiple purposes and can be combined in many ways to achieve synergic effects and tackle particular problems or support particular KM motives. The priority areas where technology can help organisations to deliver KM are summarised in Figure 1. These include support for "codification" (i.e., strategy focusing on explicit knowledge forms) or "personalisation" (i.e., strategy focusing on tacit knowledge forms) in processes of "exploitation" (i.e., knowledge use and sharing of the existing knowledge) or improving "exploration" (i.e., people's ability to discover and create new knowledge).

Figure 1. Theoretical KMT model



EMPIRICAL STUDY

Four inner-city and five outer-city municipal councils of Sarajevo were examined in this study as part of ongoing research into the current KM trends in local government in BiH. Inner-city municipalities have strong developmental orientation toward banking, finance, trade and consulting; boast pro-active universities and leading business partners; and have in common with other European cities a forward looking attitude and a high level of social capital (Anttiroico, 2006). In comparison, the outer-city areas suffer from having no university and being too dependent on domestically-oriented and labour-intensive industries. They are also sparsely populated areas with poorly resourced local administration. The initial study by Handzic et al. (2007) revealed differences in technical KM solutions among local councils. The intention of this study was to explore these differences more deeply. Results are presented in the following sections.

Types of ICT Used in KM

The results of the analysis presented in Table 1 indicate the extent of organisational usage of seven types of ICT in KM for nine local councils. The first four ICT types listed in the table are usually considered to have a more generic function, while the last three are seen as being more specific to knowledge management (Edwards & Shaw, 2004).

With respect to results by technology types, the form of ICT most widely used for knowledge management was storage technology, such as databases. This was identified as being in current extensive use (H) by all nine local councils. The second most common was sharing/delivery technology, such as e-mail systems mentioned as currently in extensive use by six local councils. Platform technologies were the other type of ICT mentioned as being in considerable use (H or M) for knowledge management in six councils. The other type of ICT being used considerably was search/retrieval technology, indicated by five councils. Three other classes of ICT were mentioned less frequently as being used significantly (H or M) for knowledge management. Access technologies received only four mentions and discovery/visualisation and utilisation technologies both received even fewer, with three mentions. These technologies were little (L) or not in current use at all (N).

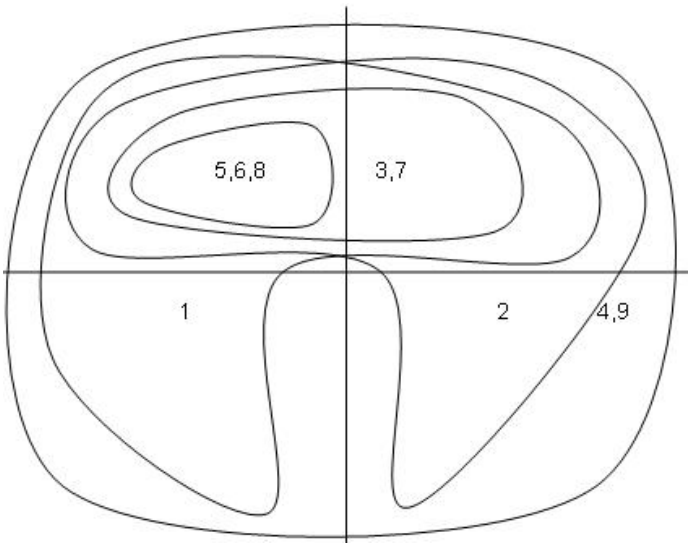
The analysis of results by councils reveals that, at one end of the scale are two councils (4, 9) who already make considerable use of all seven types of ICT in their knowledge management solutions. At the other end of the scale are three councils (5, 6, 8) where technology was seen as only of marginal relevance to knowledge management. In these councils, only storage technologies were in significant use and all other technologies only marginally or not at all, as in council 8. In between these two extreme groups are four councils (1, 2, 3, 7) who make very significant use of four generic ICT (sharing and platform, in addition to storage and retrieval), but make use of only some or none of three KM specific ICT (access, discovery and utilisation). The one exception is council 3, where several generic ICT systems were in use, but not search/retrieval. Overall, the results indicate that the majority of councils foster the use

Table 1. Usage of ICT for KM

Local council	Storage technology	Sharing technology	Platform technology	Retrieval technology	Access technology	Discovery technology	Utilisation technology
C4	H	H	H	H	H	H	H
C9	H	H	H	H	H	H	H
C1	H	H	H	M	H	H	L
C7	H	H	H	H	H	L	L
C2	H	H	M	H	L	L	H
C3	H	H	H	L	L	L	L
C5	H	L	L	L	L	L	L
C6	H	L	L	L	L	L	L
C8	H	L	L	N	N	N	N

N (=0), L (1-<3), M (3-<5), H (>=5)

Figure 2. Empirical KMT model



of technologies that are generic rather than specific to KM (72% vs. 37% significant mentions).

Roles of ICT in KM

The results reported here focus on the roles played by ICT in supporting KM strategies of knowledge exploitation or exploration, codification or personalisation. For analysis purposes, the empirical KM technology model presented in Figure 2 was built from local councils' ICT usage data.

Looking at the nature of ICT support for councils' KM strategies, there is a clear emphasis on facilitating exploitation of existing knowledge. All nine councils (1-9) demonstrate extensive use of technology to make available institutional knowledge repositories and to link organisational members to share personal knowledge. With respect to exploration, only two councils (4, 9) appear to use full technology potential to support creation of new knowledge, one council (1) makes some use of technology to support knowledge discovery and one (2) to facilitate experiential learning through knowledge application.

Further comparison of different ICT usage roles in local councils' KM as reflected in Figure 2 reveals a slightly greater emphasis placed on supporting codification than personalisation KM strategies. All nine councils (1-9) show extensive use of generic storage ICT in creating and disseminating knowledge artefacts and documents. However, only six councils (1-4, 7, 9) show considerable use of generic sharing and platform technologies to link people for communication and collaboration purposes. KM specific technologies appear to play a much lesser role in supporting either KM strategy, with only four councils (1, 4, 7, 9) using technology to facilitate knowledge access, three (1, 4, 9) to support knowledge discovery from data, and three (2, 4, 9) for learning by doing through knowledge application to the task.

Stages of KM Development

Further analysis of findings presented in Figure 2 reveals that local councils examined are currently in different stages of KM development. Three councils (5, 6, 8) appear to be in the first stage, concentrating mainly on capture and storage of codified knowledge. Such KM approach is consistent with the goal of minimising risk of losing valuable knowledge. Four councils (1, 2, 3, 7) appear to be in the second stage of KM development. They focus on sharing of the existing knowledge in people (not just in stores) through interaction. This approach is consistent with the goal of improving performance efficiency and effectiveness. The remaining two councils (4, 9) focus on creating knowledge environment conducive to innovation by balancing exploitation and exploration, and employing both codification and personalisation strategies. This approach suggests that they are in the third and final stage of KM development.

In general, inner-city councils are at higher stages of KM development than outer-city councils. Three out of four inner-city councils are in stage 2 and one in stage 3. In comparison, three out of five outer-city councils are in stage 1, one in stage 2 and one in stage 3. These findings are largely consistent with their contextual circumstances. It is clear that the majority of local councils in this sample have some way to go in developing and implementing effective KM strategies. From the follow-up interviews with two

innovative councils' officials, it is evident that these inner and outer city councils recognise new opportunities for local development (media-city and eco-tourism, respectively) and transform themselves to respond to new challenges.

FUTURE TRENDS

The findings on ICT usage in KM show clear emphasis on basic storage and sharing technologies that include files/databases and e-mail/net systems. These findings match earlier findings by Zhou and Fink (2003) and Edwards and Shaw (2004), reporting general preference for commonly known generic IT systems over those specifically labeled as being for KM. Possible reasons for this focus on simple generic systems may be found in familiarity, convenience and obvious benefits of these systems. In contrast, sophisticated and complex intelligent systems might have been rejected due to their "black-box" effect. This suggests an interesting line of inquiry for future research.

In terms of order of preference, storage technology was mentioned most often as being extensively used followed by sharing/delivery and platform technologies. This order is different from the Zhou and Fink (2003) Australian survey where e-mail, intranet and the Internet were rated as most effective ICT for KM. The fact that KM specific technologies often recommended elsewhere (Zyngier, Burstein, & Rodriguez, 2003) were rarely mentioned in this study suggests that people might have been unaware of such technologies or lacking skills to implement them. Indeed, the frequent comment received by participants in this research project concerns obstacles to effective KM in terms of inadequate ICT literacy of public sector employees. Future research should look deeper into these obstacles.

The study also considered how various ICT systems favoured by different local councils supported their organisational KM strategies. The findings indicate a clear emphasis on exploiting existing knowledge through shared repositories and personal interaction. Such strategic orientation is consistent with the currently stated business goal for BiH government administration of improving efficiency and effectiveness of its public services (Handzic et al., 2007). It is also consistent with empirical evidence from Australia showing its local government concerns with the loss of knowledge and the deployment of mechanisms for the sharing and reuse of knowledge (Martin, 2000). The challenge is to move toward innovation.

Furthermore, the findings indicate that ICT played a relatively greater role in supporting codification than personalisation KM strategy. This is not surprising, as codification approaches rely heavily on ICT to provide high-quality, reliable and fast systems to enable preservation and reuse of explicit knowledge (Hansen et al., 1999). In comparison, personalisation approaches rely more on people and invest moderately in ICT to facilitate conversations and the exchange of their tacit knowledge. Despite this, Edwards and Shaw (2004) found that personalisation and sharing of knowledge were prominent in the UK organisations. In this sample, physical proximity may have contributed to the lesser need for computer-mediated communication among employees. The role of contextual factors in KM needs to be further investigated.

Finally, the current state of ICT use in local councils is indicative of different stages in the evolutionary KM development. The majority of councils are in the early development stages, stressing reuse economics and explicit knowledge forms. These can ensure survival but not advancement, as suggested by Von Krogh, Ichijo, and Nonaka (2000). In addition, it is important to note that ICT deployment cannot resolve all KM problems and warrant effective performance. There is also need for leadership support, learning culture and continued measurement and adjustment.

CONCLUSION

This study identified current trends in usage of ICT in KM initiatives in the local government of Sarajevo. From the survey responses received, it is apparent that there was a marked preference for the use of technologies that are generic rather than specific to KM; this finding reinforces earlier ones from the UK and Australia. However, there was greater emphasis on information than on communication elements of ICT, which is contrary to UK and Australian studies and deserves further investigation.

Looking at the nature of ICT support for KM, it is obvious that technology was used more to facilitate exploitation of existing than exploration of new knowledge. This is one area for potential action by practice to move toward creation and innovation. It is also true that technology was more valuable and convenient in supporting codification than personalisation. In addition, using the KM development model, the study found that local councils in this study fitted into the categories of risk minimisers, efficiency-seekers and innovators. This leaves a significant scope for further improvement in KM among the first two groups.

Finally, it is obvious that the findings of this study are limited by choice of BiH context, local government organisations, and by the participants' self-reporting their opinions. Therefore, further research is necessary to address current limitations and ensure that these findings have broad applicability.

ACKNOWLEDGMENT

Finally, please note that the submitted article/research was funded by the Ministry of Science and Education, Canton

Sarajevo and was completed by three co-authors while working for SSST during 2007. The principal investigator/project leader was Meliha Handzic and research assistants Amila Lagumdzija and Amer Celjo. A modified/extended version of the article has been accepted for publication by IJKM (IGI journal).

REFERENCES

Alavi, M., & Leidner, D.E. (2001, March). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Anttiroico, A. (2006). Strategic knowledge management in local government (chap. 13). In A. Anttiroiko & M. Malkia (Eds.), *Encyclopedia of digital government.* Hershey, PA: IRM Press.

Awad, E.M., & Ghaziri, H.M. (2004). *Knowledge management*. Upper Saddle River, NJ: Pearson Education.

Becerra-Fernandez, I., Gonzales, A., & Sabherwal, R. (2004). *Knowledge management: Challenges, solutions, and technologies*. Upper Saddle River, NJ: Pearson Education.

Binney, D. (2001). The knowledge management spectrum: Understanding the KM landscape. *Journal of Knowledge Management, 5*(1), 33-42.

Edwards, J.S., & Shaw, D. (2004, July). Supporting knowledge management with IT. In *Proceedings of the DSS 2004 Conference*, Prato, Italy.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD-96*, Oregon.

Handzic, M. (2004). *Knowledge management: Through the technology glass.* Singapore: World Scientific Publishing.

Handzic, M., Lagumdzija, A., & Celjo, A. (2007). Auditing knowledge management practices in local government: Model and application. *Knowledge Management Research and Practice*.

Handzic, M., & Zhou, A.Z. (2005). *Knowledge management: An integrative approach.* Oxford: Chandos Publishing.

Hansen, M.T., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge?. *Harvard Business Review, 77*(2), 106-116.

Martin, B. (2000, October). Knowledge-based organisations: Emerging trends in local government in Australia. *Journal of Knowledge Management Practice*.

Swan, J. (2003). Knowledge management in action. In C.W. Holsapple (Ed.), *Handbook on knowledge management 1: Knowledge matters*. Berlin: Springer-Verlag.

Tsui, E. (2003). Tracking the role and evolution of commercial knowledge management software. In C.W. Holsapple (Ed.), *Handbook on knowledge management 2: Knowledge directions*. Berlin: Springer-Verlag.

Von Krogh, G., Ichijo, K., & Nonaka, I. (2000). *Enabling knowledge creation*. New York: Oxford University Press.

Wexler, M.N. (2001). The who, what and why of knowledge mapping. *Journal of Knowledge Management, 5*(3), 249-263.

Zhou, A., & Fink, D. (2003). Knowledge management and intellectual capital: An empirical examination of current practice in Australia. *Knowledge Management Research & Practice, 1*(2), 86-94.

Zyngier, S.M., Burstein, F., & Rodriguez, M.L. (2003). Knowledge management strategies in Australia: Analysis of uptake and understanding. In H. Hasan & M. Handzic (Eds.), *Australian studies in knowledge management.* Wollongong: UOW Press.

KEY TERMS

KM Development Model: a sequential evolutionary model of development in KM that includes three stages: retain knowledge to minimise risk, share knowledge to improve efficiency, and generate knowledge for innovation.

KM Strategies: codification (focuses on explicit knowledge) and personalisation (focuses on tacit knowledge).

KM Technology: any information or communication technology (ICT) used for the purpose of managing knowledge.

KM Technology Roles: support codification or personalisation strategies in processes of knowledge exploitation or exploration.

KM Technology Model: two-by-two matrix relating different types and roles of ICT in KM.

KM Technology Types: seven classes of ICT in KM based on the different knowledge processes they support (knowledge storage, access, search/retrieval, sharing/delivery, platform, discovery/visualisation, utilisation technologies).

Knowledge Management (KM): set of socio-technical initiatives and processes that move or modify knowledge stocks.

Category: Business Information Systems

Knowledge Sharing Tools for IT Project Management

Stacie Petter
Georgia State University, USA

Vijay Vaishnavi
Georgia State University, USA

Lars Mathiassen
Georgia State University, USA

INTRODUCTION

Information technology (IT) project disasters make worldwide headlines, and billions of dollars have been lost due to poor project implementations. The Standish Group, a research advisory firm, reports that only one-third of the over 13,500 IT projects evaluated in 2003 were successful, and half of the reported IT projects were classified as challenged, meaning they experienced cost and budget overruns (Larkowski, 2003). While the state of IT project management is improving, organizations must explore ways to reduce unnecessary spending that occurs because of failures, cost and schedule overruns on IT projects. One possibility is to improve knowledge sharing to avoid repeating mistakes and to build on successes from the past.

BACKGROUND

IT project management is demanding because of time pressure, restricted capital, and high degrees of uncertainty during projects and is comprised of complicated and ill-structured problems (Grupe, Urwiler, Ramarapu, & Owrang, 1998). However, valuable knowledge gained before, during, and after the completion of projects is rarely captured, shared, and utilized in future projects. As a result, projects suffer from reinvention of solutions, repetition of mistakes, and loss of process knowledge after project completion. These problems are further exacerbated by the turnover of project managers and the lack of technologies that effectively integrate relevant knowledge with existing project management software (Tiwana & Ramesh, 2001).

Knowledge is a combination of experience, values, contextual information, and insight used to create, absorb, and evaluate new experiences and information (Davenport & Prusak, 2000). Project managers rely on past experiences to make decisions that keep the project within schedule, budget, functionality, and quality targets; however, these experiences are rarely shared among project managers (Schindler & Eppler, 2003). Furthermore, a problem faced by many IT project managers is their own lack of experience as an IT project manager. Individuals are often promoted to an IT project manager position because of their superior programming skills. However, these experiences alone are not enough to guarantee success as an IT project manager (Standish Group, 2001). Fortunately, a variety of knowledge sharing tools can help inexperienced project managers acquire relevant knowledge. Even for IT project managers with extensive knowledge, such tools provide good opportunities to learn from others when confronted with a unique problem (Newell, 2004).

TOOLS FOR SHARING KNOWLEDGE

Many tools have been developed to assist IT project managers in avoiding project failures, including post-mortem analysis, knowledge management systems, and networking. Rather than focusing on specific tools for sharing knowledge, this section describes generic classes of knowledge sharing tools available to IT project managers. Each tool is described in terms of *what* type of knowledge is shared (i.e., available in documented form or emergent through interaction), *who* is the primary user of the tool (i.e., project manager or entire project team), *where* knowledge is shared within the organization (i.e., between individuals or organization-wide), and *why* the tool is used for sharing knowledge (i.e., exploitation of existing knowledge or a basis for exploration of new knowledge).

Post-Mortem Analysis

Post-mortem analysis is supported by a process and a series of documents to identify successes and failures for a given project (Sinofsky & Thomke, 1999). Good post-mortem analyses not only record the history of the project itself,

but also provide information on what went wrong during specific phases of the project's life cycle (Thomke & Fujimoto, 2000). Often, organizations only conduct post-mortem analyses on projects that have been abandoned or have failed (Esusi-Mensah & Przasnyski, 1995); however, there are benefits to conducting post-mortem analyses on successful IT projects. Organizations that perform post-mortems state that the knowledge gained is useful in avoiding repetition of past mistakes, improving processes on future projects, providing historical accounts of what went wrong, and enhancing performance on future projects (Esusi-Mensah & Przasnyski, 1995).

The output of a post-mortem analysis is a series of documents. These documents can be shared across the entire project team as they articulate the aspects of a project that were successful and the areas needing further improvement. Although the results of post-mortems can benefit the entire project team, a survey of post-mortem analyses found that IS managers, system developers, and new IT project managers were more likely to consult these documents than other groups such as programmers, senior management, and other functional managers (Esusi-Mensah & Przasnyski, 1995). Post-mortem analyses are shared across the organization, frequently through the use of new processes or management practices. The knowledge gained through post-mortem analyses is exploited throughout the organization to minimize the repetition of problems, better plan or manage new projects, and guide the development of new management procedures (Esusi-Mensah & Przasnyski, 1995).

KNOWLEDGE MANAGEMENT SYSTEMS

Within the knowledge management literature, there are three types of systems to create and share knowledge: codified, personalized, and collaborative systems. While codified knowledge systems transform knowledge into documented form, a personalized system helps people interact to create and communicate knowledge (Hansen, Nohria, & Tierney, 1999). Codified knowledge systems are basically shared databases; personalized knowledge systems encourage interaction. Collaboration systems combine personalized and codified knowledge and focus both on supporting interaction among colleagues and providing a repository to share knowledge that emerges through the collaboration.

Codified knowledge systems store knowledge in databases where it can be accessed and used easily by anyone in the organization, making knowledge available to all members of a project team and across project teams (Hansen et al., 1999). As a result, all project team members, from developer to project manager, can benefit from the codified knowledge. The purpose of developing and maintaining a codified knowledge system and populating it with knowledge is to disseminate knowledge throughout the organization (Hansen et al., 1999). By investing in codified knowledge and a tool to effectively share the stored knowledge, organizations can exploit knowledge obtained and captured during prior IT projects on future projects.

Organizations use codified knowledge systems in an effort to prevent knowledge from leaving the company by relying on technology for the storage of knowledge. Because IT projects are vulnerable to personnel changes, codified knowledge systems are created to exploit knowledge from past projects. Because knowledge is documented and stored within a shared system, temporal and geographical differences between the knowledge-seeker and knowledge-holder are irrelevant. The primary problem with these knowledge sharing systems is that the knowledge captured and stored often goes unused. The reason is that IT project managers must acknowledge that there is a need to seek out additional knowledge and this knowledge must be clearly communicated across project teams (Newell, 2004). Successful utilization of codified knowledge systems requires individuals and teams to seek out available knowledge as part of their project management activities.

Personalized knowledge systems help those seeking knowledge find people who have the needed knowledge or who can help create that knowledge (Davenport & Prusak, 2000). These systems are useful when the necessary knowledge is not easily documented; therefore, brainstorming sessions or conversations between an expert and the knowledge seeker are needed to provide the necessary knowledge (Hansen et al., 1999). The complex nature of IT project management suggests that some problems are better suited to this interactive method of knowledge sharing. A key limitation to this approach is that knowledge does not remain behind to benefit others when experts leave the company (Hansen et al., 1999).

Personalized knowledge systems provide a map or listing of individuals in the organization who possess different types of knowledge (Davenport & Prusak, 2000). Knowledge sharing is based on personal interaction rather than on explication of the knowledge to be shared (Hansen et al., 1999). All members of a project team, including the IT project manager, developers, and others, are able to solicit advice from those with expertise in a given subject matter. The purpose of developing and maintaining a personalized knowledge system, or knowledge map, is to communicate expertise across the organization. Personalized knowledge systems are often used for brainstorming and addressing unique problems that arise during a project (Hansen et al., 1999), and the reason for adopting these systems is to create new knowledge through the combination of existing personal insights.

Collaboration systems enable groups of people to share information, communicate, coordinate, and work together as a team (Lamont, 2004). These systems enable team members that may be geographically dispersed to work and share knowledge within the team. Incorporating collaboration sys-

tems into standard project management practices has helped organizations to better organize and manage projects (Stevens, 2001). Collaboration systems also enable version control for project plans, provide access to relevant information to project team members, and create awareness of issues that could affect the execution of a project (Stevens, 2001). Collaboration systems focus on creating and sharing knowledge electronically rather than paper-based documentation. The primary benefits of collaboration systems are the ability to bring people together within a team to solve problems and make decisions quickly and effectively and to provide them with repositories through which they can effectively share the knowledge that emerges through the collaboration. The purpose of these systems is to create, make available, and explore new knowledge to solve a unique problem. Collaboration systems are capable of supporting all members and stakeholders of the IT project team by allowing team members to communicate and brainstorm with colleagues that have knowledge relevant to the project.

Networking

A more informal method of knowledge sharing is working with others. Because projects have a finite duration, team members are disbanded at the end of a project and often placed on new teams with different members. This swapping of team members either within or across projects can enable knowledge sharing (Newell, 2004). People can learn from one another through the sharing of experiences (Davenport & Prusak, 2000). In addition, the informal relationships built from expanding one's social network can be used to share knowledge among IT project managers more effectively than extensive knowledge repositories (Newell, 2004). Effective networking can be enabled through systematic assignment of individuals to specific projects and through mentoring and apprenticeship arrangements.

Networking is often informal as one person seeks out another for knowledge, insight, and advice. Knowledge is emergent as individuals share stories and discuss experiences through conversation. Networking can be useful to any member of a project team, including IT project managers and developers. While organizations may arrange mentor relationships, the resulting knowledge sharing is occurring among colleagues. In a project environment, which often contains team members with diverse backgrounds, organizations have an opportunity to develop novel and creative ideas (Bresnen, Edelman, Newell, Scarbrough, & Swan, 2003). Formal methods, such as mentoring and apprenticeship, provide structured learning opportunities for the protégé to reuse knowledge (Swap, Leonard, Shields, & Abrams, 2001). Informal networking methods, such as communities of practice, promote knowledge creation as an integral part of project practices.

Templates

Templates can be used in projects to facilitate repeatability and consistency (Dvorak, 2003). Because specific processes are often similar across projects, templates are effective means to ensure that details are not overlooked and deliverables are complete and consistent (Dvorak, 2003). Templates can be collected and stored in a repository together with selected examples of how the template has been used. A significant portion of time is spent on documenting a project through development plans, quality plans, test plans, requirements specifications, and design specifications. Using shared templates can make such knowledge readily available to new projects and improve process productivity and product reliability (Ben-Menachem & Gelbard, 2002).

With this method of knowledge sharing, knowledge is documented as a template for easy reuse. Templates can be used by project managers to standardize their processes and documentation (Method123, 2003) or by software developers and team members for specifying requirements, writing code, or documenting the system (Ben-Menachem & Gelbard, 2002). Templates help individuals, teams, and organizations move forward quickly while maintaining consistency across projects (Dvorak, 2003). Templates are typically created for organizational level usage and are promoted as company standards; however, in other cases, templates serve as suggested practices. Eventually the success of templates relies on individuals who adopt them to avoid creating documentation from scratch. Exemplar project plans or requirements documents can often be reused and adapted to meet the needs of current projects quite similar to how generic, standardized letters can be used to quickly generate a personal response to a customer.

Best Practices

There are several sources available to assist project managers in managing IT projects. The Project Management Institute has developed the *Project Management Body of Knowledge* (PMBOK), which is a book of standards that describes the sum of knowledge within the profession of project management (Project Management Institute, 2000). The purpose of these standards is to provide project managers with a series of best practices applicable to a variety of projects. A second popular project management source is PRINCE (PRojects IN Controlled Environments), originally developed in 1989 by the Central Computer and Telecommunications Agency in the United Kingdom. PRINCE is used to "guide the project through a controlled, well-managed, visible set of activities to achieve the desired result" (PRINCE2, 2004). Both the *Project Management Body of Knowledge* and PRINCE2 contain best practices to guide project managers through a successful project. These sources are based on lessons learned through successful and failed projects. While PMBOK is a

generic set of standard best practices applicable to any type of project, the original PRINCE methodology was created specifically for IT projects. The updated version, PRINCE2, is a more generic set of best practices that can be applied to any type of project.

Best practices, as presented in sources like PRINCE2 and PMBOK, represent documented knowledge that is available via training materials, books, and Web sites for distribution to those interested in leveraging the knowledge. Standard best practices are specifically designed to enable project managers to better execute projects (PRINCE2, 2004; Project Management Institute, 2000) and are typically promoted and shared within the entire profession and across each organization. Organizations may have their own best practices or methodologies that are used for all projects within the enterprise. Developing these can often be done successfully through a combination of standard best practices and local traditions for project management.

FUTURE TRENDS

While not all knowledge sharing tools discussed here require the use of IT to accomplish their objective, the evolving nature of technology can change how these systems will be used in the future. Many software vendors for project management systems and knowledge management systems are incorporating collaboration technology into the software (Stevens, 2001). This enables team members to work together to solve problems regardless of geographic location. This integration of collaboration capabilities into standard tools enable a more seamless sharing of ideas. Another emerging trend is that as organizations find more benefits from knowledge sharing across IT projects, new roles are created to facilitate the documentation of knowledge gained during a project and the sharing of knowledge across other projects within the organization (Schindler & Eppler, 2003). Furthermore, as organizations learn about the benefits of networking to share knowledge, there will be a need to create more opportunities for IT project managers within the organization to interact and build relationships with one another (Newell, 2004). This can be accomplished via knowledge fairs, talk rooms, or conferences (Davenport & Prusak, 2000) and such initiatives can help IT project members appreciate the knowledge available among their colleagues as well as identify tools to capture and share this knowledge (Newell, 2004).

CONCLUSION

Knowledge sharing tools, such as the ones described in this article, provide organizations with the ability to share valuable lessons learned across IT projects (Schindler & Eppler, 2003). Care must be taken, however, to remember that simply instituting a knowledge sharing tool does not in and of itself lead to increased project success (Davenport & Prusak, 2000). Organizations should consider *what* type of knowledge to be shared, *who* the users of the knowledge are, *where* in the organization sharing is expected to take place, and *why* the organization is adopting specific tools for enhanced sharing of knowledge. Organizations can use one or more of these tools to enable knowledge sharing across software projects. Doing so creates an opportunity to capitalize on past successes and avoid prior mistakes in IT project management.

REFERENCES

Ben-Menachem, M., & Gelbard, R. (2002). Integrated IT management tool kit. *Communications of the ACM, 45*(4), 96-102.

Table 1. Summary of knowledge sharing tools

Tool Classes	What	Who	Where	Why
Post-Mortem Analysis	Documented	Team	Organizational	Existing Knowledge Exploitation
Personalized Knowledge Systems	Emergent	Team	Individual	New Knowledge Exploration
Codified Knowledge Systems	Documented	Team	Organizational	Existing Knowledge Exploitation
Collaboration Systems	Emergent	Team	Individual	New Knowledge Exploration
Networking	Emergent	Team	Individual	New Knowledge Exploration
Templates	Documented	Managers	Organizational	Existing Knowledge Exploitation
Best Practices	Documented	Managers	Organizational	Existing Knowledge Exploitation

Bresnen, M., Edelman, L., Newell, S., Scarbrough, H., & Swan, J. (2003). Social practices and the management of knowledge in project environments. *International Journal of Project Management, 21*(3), 157-166.

Davenport, T. H., & Prusak, L. (2000). *Working knowledge: How organizations manage what they know*. Boston: Harvard Business School Press.

Dvorak, P. (2003). Best practices for managing projects. *Machine Design, 75*(10), 54-56.

Esusi-Mensah, K., & Przasnyski, Z. H. (1995). Learning from abandoned information technology projects. *Journal of Information Technology, 10*(1), 3-14.

Grupe, F. H., Urwiler, R., Ramarapu, N. K., & Owrang, M. (1998). The application of case-based reasoning to the software development process. *Information and Software Technology, 40*(9), 493-499.

Hansen, M. T., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? *Harvard Business Review*, 106-116.

Lamont, J. (2004). Roundtable discussion: Collaboration. *KM World, 13*(1), 16-17.

Larkowski, K. (2003). *Latest Standish Group CHAOS report shows project success rates have improved by 50%*. West Yarmouth, MA: Standish Group International.

Method123. (2003). Project management templates. Retrieved December 14, 2004, from http://www.method123.com

Newell, S. (2004). Enhancing cross-project learning. *Engineering Management Journal, 16*(1), 12-20.

PRINCE2. (2004). *What is PRINCE2?* Retrieved December 17, 2004, from http://www.prince2.com/whatisp2.html

Project Management Institute. (2000). *A guide to the project management body of knowledge*. Newtown Square, PA: Project Management Institute.

Schindler, M., & Eppler, M. J. (2003). Harvesting project knowledge: A review of project learning methods and success factors. *International Journal of Project Management, 21*(3), 219-228.

Sinofsky, S., & Thomke, S. (1999). *Learning from projects: Note on conducting a postmortem analysis* (Harvard Business Case No. 9-600-021). Boston: Harvard Business School.

Standish Group. (2001). *Extreme CHAOS*. West Yarmouth, MA: The Standish Group International.

Stevens, L. (2001). At a moment's notice: Final mile introduces knowledge management to a project already under way. *Knowledge Management*, 26-27.

Swap, W., Leonard, D., Shields, M., & Abrams, L. (2001). Using mentoring and storytelling to transfer knowledge in the workplace. *Journal of Management Information Systems, 18*(1), 95-114.

Thomke, S., & Fujimoto, T. (2000). The effect of 'front-loading' problem-solving on product development performance. *Journal of Product Innovation Management, 17*(2)P, 128-142.

Tiwana, A., & Ramesh, B. (2001). A design knowledge management system to support collaborative information product evolution. *Decision Support Systems, 31*(2), 241-262.

KEY TERMS

Best Practices: Set of standards developed for specific activities to guide project managers in managing a successful project.

Codified Knowledge Systems: Documented and stored knowledge residing in a centralized database which serves as a repository of knowledge within the organization.

Collaboration Systems: Information systems used to enable colleagues to virtually meet to discuss, brainstorm ideas, and share the knowledge that emerge from their collaboration.

Knowledge: Combination of experience, values, contextual information, and insight used to create a framework to evaluate and absorb new experiences and information.

Knowledge Sharing: Knowledge that is communicated among people with similar job functions and backgrounds.

Networking: Social relationships that facilitate knowledge sharing through intense interaction, discussion, and sharing of practices and values.

Personalized Knowledge Systems: Knowledge map that enables people seeking knowledge to find people within the organization with the necessary knowledge and skills.

Project Management: Use of knowledge, skills, tools, and techniques to perform activities related to a temporary venture to develop a unique product or service according to stakeholder specifications.

Post-Mortem Analysis: Process to identify and document successes and failures for a given project.

Templates: Documents—such as project plans, budgets, or documentation—used on prior projects that are continually reused to maintain consistency across projects.

A Language/Action Based Approach to Information Modelling

Paul Johannesson
Stockholm University/Royal Institute of Technology, Sweden

INTRODUCTION

There are several different views of the role of information systems. Two of the most important are the data view and the communicative view. According to the data view, the primary purpose of an information system is to provide a model of a domain, thereby enabling people to obtain information about reality by studying the model. In this respect, an information system works as a repository of data that reflects the structure and behaviour of an enterprise, and the system provides data that can be used for decisions about the enterprise. In contrast, the communicative view states that the major role of an information system is to support communication within and between organisations by structuring and coordinating the actions performed by organisational agents. The system is seen as a medium through which people can perform social actions, such as stating facts, making promises, and giving orders.

The data and communicative views of information systems are mirrored by two different views of organisations: the functional view and the constructional view (Dietz, 2003a). The functional view focuses on the functions of an organisation with respect to its environment, in particular, the resources that the organisation consumes and produces. A model of an organisation from a functional perspective is a black-box model, as it shows the interactions with the environment but not the internal mechanisms. The constructional view, on the other hand, focuses on how behaviour and function are brought about by the operations and structure of an organisation. A model of an organisation from a constructional perspective is a white-box model as it shows the inner workings of the organisation.

In information systems design, the starting point has often been based on the data view and the functional view, though frequently augmented by concepts like reasoning and monitoring. However, these views easily lead to a computer- and technology-biased management of the communication taking place in an organisation, and they benefit from being complemented by the communicative and constructional views. A promising theoretical foundation for these views is the language/action approach, which is based on theories from linguistics and the philosophy of language. In the language/action approach, business actions are modelled on the notions of speech acts and discourses, which provide a basis for distinguishing between different communication phases, such as preparation, negotiation, and acceptance. The purpose of this chapter is to outline how the language/action approach can be used as a basis for the information modelling of communicative aspects in organisations.

BACKGROUND

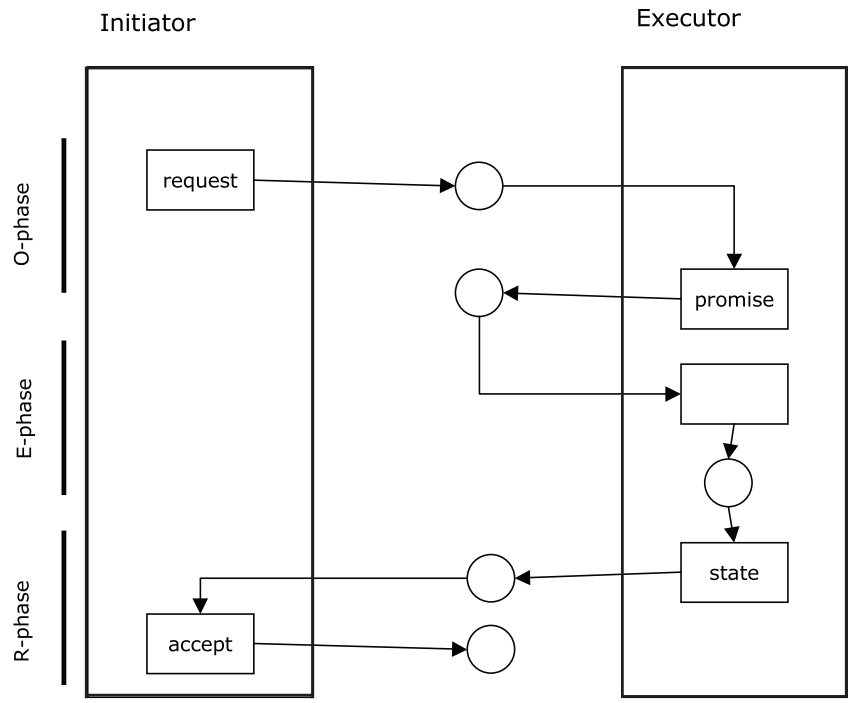
One important foundation of the language/action approach is speech act theory (Austin, 1962; Searle, 1969). The basic insight of speech act theory is that language can serve purposes other than that of representing the states of affairs of the world. Certain statements are equivalent to actions. For example, when someone says "I apologise," "I promise…," or "I name this ship…," the utterance changes the psychological or social reality. Statements such as these are called *speech acts*, and they enable people to use language as a means for acting as well as coordinating action.

In Searle (1969), a classification of speech acts is proposed based upon the way in which a speech act affects the social world. Searle identified five classes: assertive, commissive, directive, declarative, and expressive. An *assertive* is a speech act, the purpose of which is to convey information from a speaker to a hearer, e.g., "the cat is on the mat." A *commissive* is a speech act, the purpose of which is to commit the speaker to carry out some action or bring about some state of affairs, e.g., "I promise to bring it back." A *directive* is a speech act, where the speaker requests that the hearer carry out some action or bring about some state of affairs, e.g., "Please bring me the salt." A *declarative* is a speech act, where the speaker brings about some state of affairs by the mere performance of the speech act, e.g., "I hereby baptise you Samuel." An *expressive* is a speech act, the purpose of which is to express the speaker's attitude, e.g., "I like coffee."

In order to understand the role of speech acts, it is helpful to view human communication as taking place in three different worlds:

- The physical world—In this world, people carry out message actions. They utter sounds, wave their hands, send electronic messages, etc. Furthermore, other in-

Figure 1. OER pattern



strumental acts may take place in the physical world, such as repairing equipment.

- The communicative world—In this world, people express their intentions and feelings. They tell other people what they know and try to influence the behaviour of others through communication, i.e., they perform speech acts. These speech acts are brought about by means of message actions in the physical world. Note that a message action does not need to be verbal, as it can also be expressed by body language.
- The social world—In this world, people change the social and institutional relationships among them. For example, people become married or acquire possession of property. People perform such social actions by performing speech acts in the communicative world.

LANGUAGE/ACTION FOR BUSINESS PROCESS MANAGEMENT

The most important applications of the language/action approach have been made in the area of business process management (Lehtinen, 1986; Weigand, 2003). A language/action perspective provides a clear and well-founded basis for identifying and modelling recurring patterns in business processes. One such pattern is the order–execution–result (OER) pattern (Dietz, 2003b), which models a basic form of interaction that occurs in every business process (Figure 1). The interaction takes place between two parties—the initiator and the executor—and governs how they coordinate their actions. The interaction starts in the order phase by the initiator making a directive speech act, namely, a request to carry out some action (shown by a rectangle), which results in a state (shown by a circle in Figure 1), where there is an order from the initiator to the executor. The executor accepts the order by performing a commissive speech act (a rectangle labelled "promise" in Figure 1), resulting in a state where there is a commitment for the executor to carry out the action. This concludes the order phase, which is followed by the execution phase, where the executor actually performs the action (shown by an unlabelled rectangle in Figure 1) he or she is committed to. This action may be an instrumental action, e.g., delivering a package, or a declarative speech act, e.g., grading an exam. However, the execution phase is always concluded by a declarative speech act, where the executor states that he or she has carried out the committed action. The final phase is the result phase, where the initiator performs a declarative speech act and acknowledges that the executor has carried out the requested action in a satisfactory way.

The pattern introduced above shows the success path, where the initiator and executor agree to each other's actions. However, the pattern needs to be extended to handle cases where the parties do not agree, e.g., when the initiator does not accept the result of the execution phase. Larger business processes are typically built up by combinations of the OER pattern, e.g., when three or more parties are involved in a process or when reciprocal commitments are created in the process. The latter case occurs, for example, in e-commerce, as discussed in Lind (1997, 2003). Several variants of the OER pattern have been proposed in the literature, e.g., the conversation for basic action in Winograd (1986).

FUTURE TRENDS

The language action approach influenced another recent trend in the information systems community, agent-oriented information systems (AOIS Workshops, 2004). The language/action approach can be seen as one of the most active traditions within agent-oriented information systems. Agent concepts offer high-level abstractions addressing issues such as knowledge representation, communication, coordination, and cooperation among heterogeneous and autonomous parties. One precursor of agent-oriented information systems was the REA framework (McCarthy, 1982), which was designed for representing and reasoning about economic exchanges. The basic REA pattern models an exchange by three components: the *events* of the exchange, the *resources* that are exchanged, and the participating *agents*. In order to obtain a resource, an agent must give up some other resource. Therefore, an economic exchange always consists of two corresponding events, e.g., a purchase and a payment. Extended versions of the REA pattern also include commitments that are established, where a commitment is an obligation to perform a resource transfer some time in the future. These commitments are created through applications of the OER patterns.

One of the most comprehensive approaches to agent-oriented information systems is agent-object-relationship (AOR; Wagner, 2003), which is based on the notions of agents, events, actions, claims, and commitments. These notions form the basis for a general meta-model that can be used as a foundation for any information system. Another related work is that by Weigand et al. (1998), which proposes a layered architecture for information systems. At the lowest layer, we find the elementary acts—speech acts as well as instrumental acts. The next layer contains the business transactions. A business transaction is the smallest sequence of speech acts that results in a new deontic state, i.e., a state in which an obligation to carry out some action has been created, or a state where an authorisation for certain actions has been established, or a state where some action has been accomplished so that a previous obligation has become fulfilled. A single speech act is, in most cases, not sufficient to achieve a deontic effect. In general, at least two messages are required. For example, a customer requests that a supplier deliver a product, and the supplier promises to do so. These speech acts will result in two obligations: one for the supplier to deliver and one for the customer to pay upon delivery. The top layer is the workflow level, where a workflow is a set of linked business transactions that realise a business objective within an organisational structure.

CONCLUSION

The language/action approach to information modelling and systems provides a solid theoretical framework for analysing and designing communicative action in organisations. One foundation of the language/action approach is speech act theory, which investigates how language can be used to perform actions. The main application of the language/action approach has been in the area of business process design and management, where the approach can assist in creating complete and well-functioning processes. A recent trend is to apply the language/action approach for agent-oriented information systems, including applications to e-commerce and e-business (Bergholtz, 2003).

REFERENCES

AOIS Workshops. (2004). Retrieved from *www.aois.org*

Austin, J. L. (1962). How to do things with words. Oxford.

Bergholtz, M., Jayaweera, P., Johannesson, P., & Wohed, P. (2003). Reconciling physical, communicative and social/institutional domains in agent oriented information systems—A unified framework. In *Proceedings of the International Workshop on Agent Oriented Information Systems, ER 2003*, Chicago. Heidelberg: Springer.

Dietz, J. (2003a). Generic recurrent patterns in business processes. Business Process Management 2003, Eindhoven, LNCS 2678. Heidelberg: Springer.

Dietz, J. (2003b). The atoms, molecules and fibers of organizations. *Data and Knowledge Engineering, 47*(3).

Lehtinen, E., & Lyytinen, K. (1986). Action based model of information system. *Information System, 11*(4), 299–317.

Lind, M., & Goldkuhl, G. (1997). Reconstruction of different business processes—A theory and method driven analysis. *Conference on Language/Action Perspective '97*, Veldhoven.

Lind, M., & Goldkuhl, G. (2003). The constituents of business interaction—Generic layered patterns. *Data and Knowledge Engineering, 47*(3).

McCarthy, W. (1982). The REA accounting model: A generalized framework for accounting systems in a shared data environment. *The Accounting Review*, (July), 554–78.

Searle, J. R. (1969). *Speech acts—An essay in the philosophy of language*. London; New York: Cambridge University Press.

Wagner, G. (2003). The agent-object-relationship metamodel: Towards a unified view of state and behavior. *Information Systems, 28*(5).

Weigand, H., & de Moor, A. (2003). Workflow analysis with communication norms. *Data and Knowledge Engineering, 47*(3).

Weigand, H., & v d Heuvel, W. (1998). Meta-patterns for electronic commerce transactions based on FLBC. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS'98)*. Los Alamitos, CA: IEEE Press.

Winograd, T., & Flores, F. (1986). Understanding computers and cognition: A new foundation for design. Norwood, NJ: Ablex.

KEY TERMS

Business Process: This is comprised of business transactions that realise a business objective.

Business Transaction: This consists of speech acts that result in a deontic effect.

Deontic Effect: This is the establishment of an obligation or the fulfilment of an obligation.

Information System: This is a system for supporting communication within and between organisations.

Instrumental Act: This is an act performed in order to change the physical world.

OER Pattern: This is a basic pattern for business interaction based on order–execution–result phases.

Speech Act: This is a linguistic act performed with the intention of changing the social relationships between agents, in particular, creating deontic effects.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 7-10, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Leader–Facilitated Relationship Building in Virtual Teams

David J. Pauleen
Victoria University of Wellington, New Zealand

INTRODUCTION

How do virtual team leaders assess and respond to boundary crossing issues when building relationships with virtual team members? Virtual teams are a new phenomenon, defined as groups of people working on a common task or project from distributed locations using information and communications technology (ICT). With rapid advances in ICT allowing alternatives to face-to-face communication, virtual teams are playing an increasingly important role in organizations. Due to their global coverage, virtual teams are often assigned critical organizational tasks such as multinational product launches, negotiating global mergers and acquisitions, and managing strategic alliances (Maznevski & Chudoba, 2000). Their use, however, has outpaced the understanding of their unique dynamics and characteristics (Cramton & Webber, 2000).

Virtual team leadership remains one of the least understood and most poorly supported elements in virtual teams. Virtual team leaders are often the nexus of a virtual team, facilitating communications, establishing team processes, and taking responsibility for task completion (Duarte & Tennant-Snyder, 1999), and doing so across multiple boundaries. Recent research (Kayworth & Leidner, 2001-2002) has begun to look at virtual leadership issues and suggests that the trend toward virtual work groups necessitates further inquiry into the role and nature of virtual team leadership.

This article begins by briefly looking at the key concepts of virtual team leadership, relationship building and boundary crossing. Then, drawing upon the author's research, it examines the complexity inherent in building relationship across boundaries, and concludes with suggestions on how virtual team leaders can mediate this complexity.

BACKGROUND

Virtual Team Leadership

There has been extensive research on leadership in collocated teams and groups. Typically, leadership can be viewed in a number of ways, from a structured authoritative role to the ability of individuals to intrinsically or extrinsically motivate followers. It is generally agreed that leadership involves social influence and the use of communication activities in motivating teams to achieve goals. Barge proposes leadership as mediation in order to overcome the variety of task and relational problems that may be encountered by a group and explains that leadership "entails devising a system of helping the group get its work done, that is simultaneously stable and flexible and assists in managing the information shared among members and between the group and its external audience" (Barge, 1996, p. 319).

A key leadership skill in Barge's concept of leadership as mediation is that of relational management, which refers to the ability of leaders to "coordinate and construct interpersonal relations that allow an appropriate balance of cohesion, unity, and task motivation with a group" (Barge, 1996, p. 325). Cohesive teams tend to perform better and are more motivated to complete tasks. Of concern here is how team leaders can coordinate and construct interpersonal relations in a virtual environment to overcome the difficulty of multiple boundaries that do not exist in traditional collocated teams.

The Importance of Relationship Building in Virtual Teams

The link between team effectiveness and team member relationships is an important but underdeveloped area of study in virtual teams. Usually defined implicitly rather than explicitly, relationships develop over time through a negotiation process between those involved (Catell, 1948). While face-to-face meetings are the preferred way to build relationships and to deal with sensitive and complex situations, it is possible with the skillful and thoughtful application of virtual communication channels to effectively lead a completely virtual team. Research has found that computer-mediated teams do share relational information and are likely to develop relational links over time (Chidambaram, 1966; Warkentin, Sayeed & Hightower, 1997).

The role of the team leader is to move the team towards its objectives by encouraging collaboration. This is done through a sustained process of relationship building, idea generation, prioritisation and selection. The particular challenge to virtual team leaders is to manage this process through ICT. In virtual team research stronger relational links have been associated with higher task performance, more effective information exchange, enhanced creativity and motivation, increased

morale, and better decisions (Warkentin & Beranek, 1999; Warkentin et al., 1997). The building of relationships with virtual team members has been shown to be a fundamental concern of virtual team leaders (Pauleen, 2003-04).

Boundary Crossing in Virtual Teams

Boundary crossing is a defining characteristic of virtual teams. Contemporary organizations have highly permeable boundaries allowing substantial communication across boundaries (Manev & Sorenson, 2001). Boundary crossing is an important organizational activity that enhances the flow of information from the external environment. The role and activities of virtual teams leaders make them natural and strategic boundary crossers.

While traditional co-located teams may have members from different functions and cultures, sophisticated new synchronous and asynchronous ICT make it ever easier to form teams consisting of members from different functions, offices, organizations, countries and cultures. Furthermore, virtual teams must function across time and distance, often with team members having never met. These conditions present significant challenges to team leaders and members, team processes and ultimately team outcomes. Because virtual teams are still relatively new, outdated organizational HR and IT policies, which do not support virtual team performance, may be compounding the challenges (Jackson, 1999; Vickery, Clark & Carlson, 1999).

Boundary crossing in virtual teams can affect relationship-building efforts. Maznevski and Chudoba (2000) showed that deliberately addressing relationship building to develop shared views and trust across all types of boundaries could help virtual team performance. The more boundaries between leaders and team members at the start of a virtual team, the more likely higher levels of relationship with team members as well as more intensive relationship-building strategies will be needed.

MAIN THRUST OF THE CHAPTER

The Effects of Boundary Crossing on Relationship Building

The practical effect of working across distance means that teams can and do comprise members from different departments, head and branch offices, and organizations, as well as different countries and cultures. Indeed, access to different organizational, functional and cultural perspectives is a key reason for using virtual teams. These differences represent important conditions that team leaders will probably need to assess and accommodate before commencing a virtual team. According to team leaders, the development of personal relationships between themselves and team members is an important prerequisite in establishing and maintaining virtual working relationships across three conceptual boundary-crossing categories: (1) Organizational Boundary Crossing, (2) Cultural/Language Boundary Crossing and (3) Time/Distance Boundary Crossing (Pauleen, 2003-04). While organizational and cultural/language barriers exist in co-located teams, they are more likely to be found in virtual teams and to a have a more significant impact. Time and distance boundaries are unique to virtual teams.

Organizational Boundary Crossing

Organizational boundary crossing includes intra- and inter-organizational boundaries. Different functions, departments, and organizations may have diverse work cultures as manifested by deeply held core beliefs and assumptions (Kayworth & Leidner, 2000). Wiesenfeld, Raghuram and Garud (1998) suggested that organizational identification would be the psychological tie that binds virtual workers together into an organization, preventing workers from thinking of themselves as independent contractors, operating autonomously.

A strong organizational culture might influence the level of relationship building necessary in a team composed of members from within the same organization, even if they are located in different countries. Strong organizational cultures are exemplified by institution-based trust relationships (Nandhakumar, 1999; van der Smagt, 2000) and an anticipation of future association (Pauleen, 2003-04). The degree of relationship building necessary and the strategy for going about it are likely to be quite different when a team starts with a strong intra-organizational culture. Conversely, virtual teams with members from different organizations will need to be aware of and navigate the different organizational cultures.

Another aspect of organizational boundary crossing is the particular preferences of certain organizations for certain technologies, for example, communication channels such as e-mail or voice mail when leaving messages. Team leaders may experience difficulties trying to agree on common communication platforms with team members outside of the organization.

Cultural/Language Boundary Crossing

Cultural/language boundary crossing is another critical area. Cultural/language boundary crossing will most likely take place in global virtual teams, though it may also be a factor in national or even local virtual teams (Pauleen, 2003-04). The key point is whether a team leader is working with a team member from another nationality or ethnic culture. The effects of culture in team settings can be profound, and include, among other important issues, how individuals relate

to each other (Kayworth & Leidner, 2000). Misinterpretations or distortions may occur as team members and team leaders interpret communications through their own cultural programming (Lewis, 1996), a challenge that is greatly complicated when attempted through ICT. In all cases, team leaders need to assess the impact of cultural differences.

In some cases, there may be a strong cultural preference for the use of face-to-face communication to build relationships. Only after a certain level of comfort is attained can ICT be used with any effectiveness. This strategy supports Hall's (1976) theory of high and low context cultures, which states that for some cultures communication is more about context then the actual verbal message. In high-context cultures, messages have little meaning without an understanding of the surrounding context, which may include the backgrounds of the people involved, previous decisions, and the history of the relationship. People from low-context cultures prefer more objective and fact-based information. The message itself is sufficient.

Team leaders will need to consider the degree of personal relationship necessary to get the working relationship underway, as well as the use of appropriate communication channels along with appropriate messages delivered in an appropriate manner (Pauleen, 2003-04).

Time and Distance Boundary Crossing

Time and distance boundaries most obviously distinguish virtual teams from co-located teams. The effect of distance on relationship building strategies is proportional to how far the team leader and team members are from each other. The further away, the more difficult the use of face-to-face communication, which could be problematic in situations where face-to-face communication is the best or maybe only option.

The effect of time on relationship building strategies concerns the challenge of working across time zones. This may have little impact on the degree of relationship building that may be necessary, but a large effect on creating strategies to build relationships, as the time differences can restrict the kinds of ICT, particularly synchronous channels such as telephone and videoconferencing, available to the team leader. Probably not an issue that will make or break a virtual team of professionals, it is one of the conditions that must be carefully and fairly assessed by the team leader before creating relationship-building strategies.

If asynchronous communication channels are used, such as e-mail, the problem of pacing communication exchanges can become a serious consideration. Response times between team leaders and team members may differ, constraining communications, causing uncertainty and negatively impacting trust (Jarvenpaa & Leidner, 1999;Warkentin & Beranek, 1999). Time lags due to technical infrastructure and technological breakdowns, if not understood by the people involved, can cause the team leader or team member to attribute non-communication to lack of manners or conscientiousness, which can then seriously affect relationships (Cramton & Webber, 2000).

Problems associated with crossing time and distance have the potential to greatly disrupt relationship building in a virtual team, particularly with inexperienced team members. It is necessary for the team leader to carefully assess these potential obstacles before creating relationship-building strategies, as well as anticipate the problems that may be caused by time and distance.

Challenges and Solutions for Team Leaders and Organizations

This discussion points to significant differences in co-located and virtual teams in leadership-led relationship building across boundaries. The greater number and variety of boundaries and their deeper impact pose special challenges that virtual team leaders will need to mediate. Table 1 summarizes the key communication challenges faced by virtual team leaders. Two of these challenges are discussed. First the need for virtual team leaders to expand and hone their repertoire of skills in handling virtual communication channels and second, what organizations can do to improve their teams' virtual communication skills.

The strategic use of communication channels is one critical skill that virtual team leaders will need to hone. Table 2 illustrates ways in which virtual teams leaders need to mediate cultural/language boundaries by consciously selecting the most appropriate communication channels. Higher context cultures will tend to require media rich channels such as phone and video conferencing to build relationships, at least until the development of a sufficient level of trust. In contrast, lower context cultures, which are more task-oriented, will tend to be more tolerant of a wider range of communication channels. Indeed, being task-oriented, building relationships might be secondary to getting started on the task. Knowing the cultural composition of the team and team members' prior experiences working virtually and across boundaries in general are part of the complex web of factors that leaders need to determine and then mediate if they are to successfully build relationships with team members and ultimately complete team tasks.

The second challenge concerns organizational support structures to improve both virtual team leader and member skills. Organizations need to be willing to provide training and support to develop effective boundary crossing behaviors among virtual team members. Virtual team processes and dynamics are different from those of co-located teams and require special skills focusing on networking and establishing links across boundaries (Duarte & Tennant-Snyder, 1999). Team members, and particularly team leaders, will often need to play multiple roles as negotiators (with customers),

Table 1. Mediating complexity - Communication challenges when building relationships with virtual team members across boundaries (Pauleen & Rajasingham, 2004)

	Types of Boundaries Crossed		
	Organizational	**Cultural/Language**	**Time & Space**
Co-located Teams	Shared organizational culture supports relationship building, although functional culture can pose challenges.	Nonverbal cues can be understood by experienced leaders. Non-native speakers	The ability to regularly meet face-to-face supports relationship building across cultures of all types.
Virtual Teams	Differing organizational policies inhibit communication, as do organizational preferences for the use of different communication channels and ICT infrastructure. Lack of situational knowledge of team members can cause misunderstandings and mis-attribution, leading to potential difficulties.	Lack of nonverbal cues is difficult to overcome with most available communication channels. Cultural preferences for certain communication channels, often face-to-face (see Table 2 for more detailed analysis of the use of channels across cultures)	Building relationships across time and space requires concerted efforts and more time than in co-located contexts. Arranging synchronous meetings or even phone calls across time zones is often problematic. Asynchronous channels face the problem of pacing communications. Dealing with "silence" is a particularly difficult challenge.

Table 2. Mediating complexity - Guidelines for using communication channels to build relationships in virtual teams across cultural/language barriers (Pauleen & Rajasingham, 2004)

Preferred Communication Channels			
		Native Speakers	Non-native Speakers
High Context Cultures (relationship-oriented) (tend toward formality)	Media rich, Synchronous, Face-to-face, phone, video/audio	Media rich, Synchronous, asynchronous	Face-to-face with interpreter, Asynchronous written with editor
Low Context Cultures (task-oriented) (Tend toward informality with notable exceptions, e.g., Germans)	Flexible, as above, plus e-mail, fax, computer conferencing (online synchronous written chat, asynchronous discussion boards)	All channels – synchronous and asynchronous	All channels with translator, interpreter or editor as required

network and coalition builders (with other teams), lobbyists (with top management), and motivators (of team members). To be effective, team leaders will need training in boundary crossing, networking and relationship building skills (Yan & Louis, 1999).

FUTURE TRENDS

There has been a pressing need for rigorous conceptual and empirical work to examine factors that influence virtual teams (Pare & Dube, 1999), and it is only in the most recent literature that there have been systematic attempts to look at how virtual team leadership can support virtual team suc-

cess. This article has briefly looked at issues related to team leader-facilitated relationship building across boundaries. Its results suggest directions for future research and practice, including the development of virtual leadership mediation and communication skills.

Realizing effective leadership in group, team and multicultural interorganizational communications across diverse perspectives and global virtual environments presents new, real and compelling challenges to team leaders, but these challenges also present unparalleled opportunities for teams to explore new perspectives, approaches and ideas (Adler, 2002). Understanding these challenges and developing effective leadership and organizational processes to meet them presents opportunities for both practitioners and researchers.

CONCLUSION

Crossing organizational, cultural and time and distance boundaries requires training, experience and organizational support. These can help team leaders determine how to work across boundaries that may be present in their team. In addition to its effects on building relationships, boundary-crossing differences can affect team processes and performance in many ways. Ignoring them is an invitation to team failure. Leaders must learn to mediate the increased complexity inherent in virtual teams. A starting point for leaders is to approach the complexity introduced by boundary crossing by asking these two questions:

1) What are the boundary crossing influences of this situation?
2) How can they be understood and worked with so that a good people-oriented environment of assurance and trust can be maintained and productivity enhanced?

Perhaps the ultimate challenge for team leaders, particularly in long-term or on-going virtual teams, is to work to merge the individual cultures – functional, organizational, national, and so forth - of the team members into a team culture.

REFERENCES

Adler, N. (2002). *International dimensions of organizational behavior.* Cincinnati: South-Western.

Barge, J.K. (1996). Leadership skills and the dialectics of leadership in group decision making. In R.Y. Hirokawa & M.S. Poole (Eds.), *Communication and group decision making* (pp. 301-342). Thousand Oaks, CA: SAGE.

Catell, R. (1948). Concepts and methods in the measurement of group syntality. *Psychological Review, 55*, 48-63.

Chidambaram, L. (1996). Relational development in computer supported groups. *Management Information Systems Quarterly, 20*(2), 142-165.

Cramton, C., & Webber, S. (2000). Attribution in distributed work groups. In P. Hinds & S. Kiesler (Eds.), *Distributed work: New research on working across distance using technology* (pp. 191-212). Cambridge, MA: MIT Press.

Duarte, N., & Tennant Snyder, N. (1999). *Mastering virtual teams: Strategies, tools, and techniques that succeed.* San Francisco: Jossey-Bass Publishers.

Hall, E.T. (1976). *Beyond culture.* New York: Doubleday.

Jackson, P.J. (1999). Organizational change and virtual teams: Strategic and operational integration. *Information Systems Journal, 9*(4), 313-332.

Jarvenpaa, S.L., & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organizational Science, 10*(6), 791-815.

Kayworth, T., & Leidner, D. (2000). The global virtual manager: A prescription for success. *European Management Journal, 18*(2), 183-194.

Kayworth, T., & Leidner, D. (2001-2002). Leadership effectiveness in global virtual teams. *Journal of Management Information Systems, 18*(3), 7-40.

Lewis, R.D. (1996). *When cultures collide: Managing successfully across cultures.* London: Nicholas Brealey Publishing.

Manev, I.M., & Sorenson, W.B. (2001). Balancing ties: Boundary spanning and influence in the organization's extended network of communication. *The Journal of Business Communication, 38*(2), 183-205.

Maznevski, M.L., & Chudoba, K.M. (2000). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science, 11*(5), 473-492.

Nandhakumar, J. (1999). Virtual teams and lost proximity: Consequences on trust relationships. In P. Jackson (Ed.), *Virtual working: Social and organizational dynamics* (pp. 46-56). London: Routledge.

Pare, G., & Dube, L. (1999). Virtual teams: An exploratory study of key challenges and strategies. *Proceedings of the Twentieth International Conference on Information Systems*, Charlotte, NC (pp. 479-483).

Pauleen, D. (2003-2004). An inductively derived model of leader-initiated relationship building with virtual team

members. *Journal of Management Information Systems, 20*(3), 227-256.

Pauleen, D., & Rajasingham, L. (2004). Mediating complexity: Facilitating relationship building in start-up virtual teams. In D. Pauleen (Ed.), *Virtual teams: Projects, protocols and processes* (pp. 255-279). Hershey, PA: Idea Group Publishing.

Van der Smagt, T. (2000). Enhancing virtual teams: Social relations v. communication technology. *Industrial Management and Data Systems, 100*(4), 148-156.

Vickery, C.M., Clark, T.D., & Carlson, J.R. (1999). Virtual positions: An examination of structure and performance in ad hoc workgroups. *Information Systems Journal, 9*(4), 291-312.

Warkentin, M., & Beranek, P.M. (1999). Training to improve virtual team communication. *Information Systems Journal, 9*(4), 271-289.

Warkentin, M.E., Sayeed, L., & Hightower, R. (1997). Virtual teams versus face-to-face teams: An exploratory study of a Web-based conference system. *Decision Sciences, 28*(4), 975-996.

Wiesenfeld, B.M., Raghuram, S., & Garud, R. (1998). Communication patterns as determinants of organizational identification in a virtual organization. *Journal of Computer Mediated Communication, 3*(4).

Yan, A., & Louis, M.R. (1999). The migration of organizational functions to the work unit level: Buffering, spanning, and bringing up boundaries. *Human Relations, 52*(1), 25-47.

KEY TERMS

Asynchronous Communication Channels: Communication channels that support communication that usually requires a period of time to pass between communicative transactions. These channels include e-mail, discussion boards, fax, and so forth.

Boundary Crossing: Virtual teams are often characterized by their boundary spanning attributes; that is, they usually cross time and distance, and often include different national (ethnic), organizational and functional cultures.

Personal Relationships: The kind of relationship between people exemplified by shared understanding, mutual trust and social bonding. Communication in personal relationships is initially directed toward the exchange of personal information and later toward the sharing of mutual experiences.

Synchronous Communication Channels: Communication channels that allow real-time interaction. These include telephone, video conferencing, chat, and of course, face-to-face communication.

Virtual Team: A given number of people at distributed locations communicating and working to some degree via information and communication technologies on a set project or task, which may be of a limited or unlimited duration. Face-to-face meetings at the start-up of the team or at regular intervals are possible in a virtual team.

Virtual Team Leader: This is the person who functions as the hub of the team, holding it together. In the literature, this person may be termed a team facilitator, (virtual) project manager, coordinator or coach depending on the nuances of the role, the perspective of the researcher and organizational terminology. The team leader responsibilities may include all or some of the following: selecting team members; setting team tasks and team member roles; ensuring project or task completion; liaising with stakeholders and clients; establishing communication and team protocols, facilitating interpersonal and team communication, handling conflict, and managing technology and in general ensuring effective participation of all the team members.

Working Relationships: The kind of relationship exemplified by people who work together toward the completion of work-based tasks. It involves communication related to sharing information, coordinating tasks, meeting timelines, and so forth.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1793-1798, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Leapfrogging an IT Sector

Eileen M. Trauth
The Pennsylvania State University, USA

INTRODUCTION

Accompanying the global spread of the post-industrial society (Bell, 1973) are nations who see economic opportunity deriving from the development of an information economy to support it (Porat, 1977). But while advanced industrialized nations moved gradually from industrial to post-industrial work over a period of decades, newly industrializing countries are "leapfrogging" directly from agrarian to information-intensive work in a matter of years. Given this rapid labor force transformation, a critical consideration in the development of a global information sector is the development and management of information technology (IT) workers.

Ireland is an appropriate country for examination of this leapfrog phenomenon because it was one of the earliest examples of this phenomenon, having developed its information sector rapidly and successfully through inward investment by multinational firms during the 1970s to the 1990s. Thus, this case offers the point of view of both an advanced industrialized or "first wave" country and of a "second wave" country that is taking an alternate path into the information economy by rapidly moving directly from an agrarian or partially-developed industrial economy into an information economy. Since Ireland was one of the earliest examples of "leapfrogging", the Irish case has lessons applicable to other contexts (Trauth, 2000).

BACKGROUND

Ireland's rapid transformation from a poor, agrarian society to a robust information society fueled by its information economy was the result of policy initiatives, cultural compatibility with IT work and adaptive responses to opportunities and crises. Ireland's policy of economic development through inward investment was a direct reversal of the preceding policy of cultural and political sovereignty achieved largely through economic isolationism. But a combination of high emigration and high unemployment signaled the need for change (Trauth, 2001).

The multinational firms brought direct benefits through the jobs that kept people in Ireland and away from unemployment, and indirect benefit through the foreign investment that would provide both jobs and a new business climate. These outside influences were expected to help Ireland more quickly develop an indigenous entrepreneurial capacity. The long-term benefits would be the spillover effects from the development of technical and business expertise.

Ireland provided attractive economic incentives in the forms of tax relief and grants for equipping their factories, and training the work forces. These were the necessary conditions for establishing the multinational IT sector in Ireland. But the sufficient conditions were a societal infrastructure supportive of IT work and a qualified labor force to do it. Today, Ireland's software industry has emerged as a strong contender for multinational sites, along with Israel, India (Heeks, 1996) and Eastern Europe (Heavin, Fitzgerald, & Trauth, 2003). Ireland's software sector employs 30,000 people in both indigenous and multinational operations and creates revenues in excess of Euro 10 billion (Flood et al., 2002).

The Irish case offers two important sets of human resource issues. The first set relates to ensuring a supply of appropriately qualified IT workers. The second set relates to managing IT workers in a cross-cultural environment. To the extent that Ireland's experiences are typical of other second wave countries, the lessons learned apply to indigenous and multinational managers as well as government policy makers in other counties.

ENSURING A SUPPLY OF QUALIFIED IT WORKERS

(Re)Designing Societal Structures to Support IT Work

Among the societal infrastructures that were adapted to support the emerging IT sector, the most important was the educational infrastructure (Clancy, 1988). Irish policy makers recognized that the well-educated Irish population was a powerful resource that could be leveraged to support the emerging information sector. But there were two serious issues to overcome. The first was establishing equality of access to education. This was accomplished in 1968 when secondary education became state-funded. The other issue was enabling potential IT workers to acquire the specialized skills and knowledge for work in this sector. In the 1960's, the traditional university was not oriented toward vocational education much less vocational education of a technical nature. Consequently, in the 1970's and 1980's, two new

universities were established and the existing universities were adapted to incorporate business and IT skills into their curricula. Technical colleges were also established. Evening, adult-oriented programs were established for workers to develop their skills and employment prospects. Finally, the government-sponsored IT training programs for those with university degrees or who had been made redundant in other fields (Trauth, 1993).

It was also necessary to maintain alignment between the particular skill sets being developed in the schools and training programs, and the available types of jobs. An unintended consequence of Ireland's educational success was that Irish IT workers became a desired human resource in other countries. In response, industrial policy assessments recommended a closer match between the educational plans of the universities and the employment opportunities available in the country (Industrial Policy Review Group, 1992).

Addressing Barriers to a Wider Participation in IT work

There were barriers to full participation in Ireland's IT sector with respect to age, gender and social class. The perception of IT as a young person's field was reinforced by the extremely young population of the country. This age divide was exacerbated by the prior educational policy limiting access to secondary and, therefore, higher education. Thus, in the early 1990's, those over the age of thirty-five, those who had not had access to free secondary education, were fewer in number in the IT labor force.

Another type of barrier relates to gender. While women found the IT sector better for women than traditional industries, banking and the civil service, there was still a tension between opportunity and restriction. IT was a new industry without established patterns of gendered work. However, there was also an acknowledged stereotype of IT work as a male activity and recognition by both men and women that women were not full participants in the IT sector. The reasons have been typically linked to a culture of large families in which child rearing was a woman's responsibility (Kvasny & Trauth, 2002; Trauth, 1995).

The final type of barrier relates to social class. Despite Ireland's historic disdain for rigid social class categories, there is evidence of social class barriers in the information sector. The absence of free secondary education was a barrier to poor and working class individuals. Further, in family settings without a history of or value placed upon education, there can be pressure on young people to enter the work force as soon as possible in order to add to the family income. Other evidence of attitudinal barriers, coming from members of the middle class, was the importance of having the "correct" accent and address in order to secure employment in indigenous IT firms.

Managing IT Workers in a Cross-Cultural Work Environment

The second set of issues relates to managing IT workers in a cross-cultural work environment in which the "first wave" nation's culture is embedded in the corporate culture of the multinational firm (Trauth, 1996). A firm's corporate culture " its values, management style, method of operations and work environment" reflects the national culture in which it developed. Thus, the multinational IT workplace was a cross-cultural mix of American and Irish cultures, the IT culture, and the particular corporate culture of the firm. While the Irish workers welcomed the American management style and corporate culture, there was also tension over how far – and in what direction – the cultural influence should go. Not surprisingly, the American managers favored the American culture, believing that the multinational firms ought to have a significant cultural impact. On the other hand, Irish human resource managers argued for tailoring the corporate culture to the particular national context. The viewpoint was that while the multinationals were bringing certain values and attitudes to the workplace, there was also a significant contribution to be made by the Irish culture. But along with resistance to the multinational influence was the recognition that importing another work culture was part of the plan. By bringing in multinational firms Ireland would be able to import a well-established work ethic that would have taken considerably longer to develop if done indigenously.

FUTURE TRENDS

Managing IT workers in a cross-cultural environment requires the acknowledgment that two different national cultures are involved when a multinational IT firm sets up operations in a country. Both of these cultures have positive contributions to offer the workplace. Multinational managers should strive to understand work patterns and attitudes of the host country culture that, while different, may nevertheless be productive. They should also understand that introducing a corporate culture means introducing a different national culture. At the same time, the host country must recognize that when the intent is to import expertise in order to quickly introduce an IT sector, one side effect will be the changes in national culture.

This has been referred to in the literature as *situating culture* (Weisinger & Salipante, 2000; Weisinger & Trauth, 2002, 2003). According to this theoretical framework, culture is a socially negotiated, dynamic, practical, and locally situated process. Hence, it is a view of culture as "doing" (which places emphasis on the actual behaviors of people) rather than "thinking" (which places emphasis on shared cognitive schemas). This view sees the interaction among the group

members' different cultures as being situated in a particular context. Thus, by situating culture in a particular context, a manager may be better able to comprehend the emergence of unique local cultural processes that reflect distinct socially negotiated realities and workplace practices.

By recognizing that there is more than one way to achieve a management goal, multinational managers can develop procedures and management approaches that exploit the best features of the host country. In the presence of two distinct cultures, it is also necessary to acknowledge that cultural influence goes in both directions. It is natural that the home culture of the multinational IT firm will influence the society it enters. But it is also natural that the workers will influence the corporate culture to make it compatible with their own. By building a permeable wall, by the open exchange of values and norms, both cultures can be enriched.

For example, American managers in Ireland learned that a human relations problem could be diffused by utilizing an Irish cultural institution: the pub. By meeting in a setting that, according to cultural norms, conveys equal standing, a manager could more easily work out the problem with the employee. Clearly, this management approach fits well with Irish culture but may not suit another cultural context. Likewise, what works best in Japan or in America may not be what works best in Ireland. Another example is the approach taken to knowledge management. Knowledge management refers to the capturing and recording of the tacit organizational knowledge that has built up over time and that resides within the employees of a firm. One of the main objectives of engaging in knowledge management is to improve the flow of organizational knowledge to facilitate organizational learning of new employees and cross-training. In Ireland, managers can address some aspects of this goal by leveraging certain cultural characteristics instead of investing in expensive equipment. The cultural characteristic of interest in others and sociability leads naturally to employees knowing more about their colleague's work than might be the case in another culture that places less emphasis on sociability (Trauth, 2000). Hence, cross-training is a natural outcome of this interaction. Simply, by recognizing that this is occurring, management can reap knowledge transfer benefits.

CONCLUSION

The case of Ireland represents an early example of the issues associated with what is currently recognized as a trend toward global outsourcing. Because it was a leader in positioning itself as a destination for global outsourcing, Ireland's experiences offer insights for both nations and multinational companies interested in outsourcing. One insight is that a qualified labor force must be available. Because of the specialized skills required for employment in the information sector, education is the key to attracting outsourcing. But in order to ensure that all citizens are able to participate in this employment sector, there must be equal educational opportunity. In Ireland, and elsewhere, members of certain groups experience barriers that inhibit full and equal participation in the information economy. In order to ensure that the society can sustain its IT sector, the educational infrastructure needs to be aligned with the skills and knowledge required for IT work, and the educational plans of the universities should be coordinated with the employment objectives of the firms. Finally, management consideration needs to be given to the mix of cultures that is present when a multinational firm from one country sets up operations in another. In this new cultural context, both cultures come together to produce a new, *situated* culture. The public policy, human resource and management issues that are raised in this case reveal some of the issues that must be addressed as more and more countries "leapfrog" an IT sector in order to reap the benefits of global outsourcing. These issues must be considered from two perspectives: that of the firms engaged in outsourcing, and that of countries inviting firms to its shores.

REFERENCES

Bell, D. (1973). *The coming of post-industrial society: A venture in social forecasting.* New York: Basic Books.

Clancy, P. (1988). *Who goes to college: A second national survey of participation in higher education*. Dublin: Higher Education Authority.

Flood, P., Heffernan, M., Farrell, J., MacCurtin, S., O'Hara, T., O'Regan, P., & Carroll, C. (2002). *Managing knowledge-based organizations: Top management teams and innovation in the indigenous software industry.* Dublin: Blackhall Publishing.

Heavin, C., Fitzgerald, B., & Trauth, E.M. (2003). Factors influencing Ireland's software industry: Lessons for economic development through IT. In M. Korpela & R. Montealeagre (Eds.), *Information systems perspectives and challenges in the context of globalization.* Boston: Kluwer Academic Publishers (pp.235-252).

Heeks, R. (1996). *India's software industry.* New Delhi: Sage Publications.

Industrial Policy Review Group. (1992). *A time for change: Industrial policy for the 1990s.* Dublin: Ministry for Industry and Commerce.

Kvasny, L., & Trauth, E.M. (2002). The digital divide at work and home: Discourses about power and underrepresented groups in the information society. In E. Wynn, M.D. Myers, & E.A. Whitley (Eds.), *Global and organizational*

discourse about information technology. Boston: Kluwer Academic Publishers (pp.273-291).

Porat, M. (1977). *Information economy: Definition and measurement.* Washington, D.C.: Office of Telecommunications.

Trauth, E.M. (1993). Educating IT professionals for work in Ireland: An emerging post-industrial country. In M. Khosrowpour & K. Loch (Eds.), *Global information technology education: Issues and trends*. Harrisburg, PA: Idea Group Publishing (pp.205-233).

Trauth, E.M. (1995). Women in Ireland's information economy: Voices from inside. *Eire Ireland, 30*(3), 133-150.

Trauth, E.M. (1996). Impact of an imported IT sector: Lessons from Ireland. In E.M. Roche & M.J. Blaine (Eds.), *Information technology development and policy: Theoretical perspectives and practical challenges.* Aldershot, UK: Avebury Publishing Ltd (pp.245-261).

Trauth, E.M. (2000). *The culture of an information economy: Influences and impacts in the Republic of Ireland*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Trauth, E.M. (2001). Mapping information-sector work to the workforce: The lessons from Ireland, *Communications of the ACM, Special Issue on The Global IT Workforce, 44*(7), 74-75.

Weisinger, J.Y., & Salipante, P. (2000). Cultural knowing as practicing: Extending our conceptions of culture. *Journal of Management Inquiry, 9*(4), 376-390.

Weisinger, J.Y., & Trauth, E.M. (2002). Situating culture in the global information sector. *Information Technology and People, 15*(4), 306-320.

Weisinger, J.Y., & Trauth, E.M. (2003). The importance of situating culture in cross-cultural IT management. *IEEE Transactions on Engineering Management, Special Issue on Cross-cultural IT Management, 50*(1), 26-30.

KEY TERMS

Cross Cultural IT Management: Managing the IT function and its personnel in a globally distributed setting.

Global Outsourcing: The trend towards directing outsourcing—contracting with other firms to perform non-critical functions for a business—toward countries with low workforce costs.

Information Economy: That portion of the national economy that is based upon information processing and related activities (Porat, 1977).

Information Sector: A component of the information economy. The *primary information sector* includes those who develop hardware, software and information systems. The *secondary information sector* includes those engaged in information processing activities in the course of doing work related to some other primary activity such as insurance claims processing.

Information Society: A societal transformation in which information is the key resource.

Information Technology (IT) Worker: One who works in the primary or secondary information sector.

Post-Industrial Society: A society in which knowledge replaces capital as the key economic resource and the predominant type of work is in the service sector (Bell, 1973).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1799-1802, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Learnability

Philip Duchastel
Information Design Atelier, Canada

INTRODUCTION

Learnability is not exactly a new concept in information technology, nor in cognitive science. Learnability has been a key concept of usability (Folmer & Bosch, 2004) in the area of software system design, where it relates to such issues as consistency, familiarity and simplicity. It has also been a traditional concept in linguistics in relation to the ease of language learning (McCarthy, 2001) and in machine learning (Valiant, 2000).

The concept of learnability has recently been repurposed within the field of instructional technology (Duchastel, 2003), building on the concept of usability in Web site design (Nielsen, 2000), and it is that learnability that is considered here. Learnability in this new sense concerns how learnable some piece of instruction is. It deals with a facet of educational resources.

The basic question is this: What makes the content of an instructional site (or of some resource) learnable? Take any one of the many thousands of online learning courses currently available on the Web and ask yourself: Does this course seem difficult to learn (assuming you have the proper background for it)? What would improve it? What would the ideal online course in this area look like? These questions all underlie the learnability of the course.

What then is learnability? Could we say that it is defined by successful learning? That would mean that students who study the course thoroughly learn its content, as evidenced on a good test for instance. Or could we say that a main criterion is ease of learning? Meaning that students experience good intellectual flow and enjoy the course.

Both of these factors, success in learning and enjoyment of learning, can be considered criteria of learnability. Are there others? That is the issue of learnability.

The skeptic will immediately insist that learning takes place within a learner and that it is that locus that mainly determines learnability – that is, the curiosity, intelligence, motivation and persistence of the learner. These are what make or break learning. The teaching materials can only go so far, the learner has to make a go of it, make it succeed.

While there is some truth to that view, it is certainly not the full picture, nor the most useful picture. Consider traditional usability in Web sites or software products. There too, the user plays a role. If he is dull-witted, or perhaps too pressed for time (showing a lack of interest), or just resistant to learning the basics (jumping in and thrashing around – as often happens), there is little scope for success no matter how usable the site or program may have been made. But we do not give up on usability in Web site creation because of that.

The point is designers do not blame the user for incompetence, for ill-will or for the lack of success of their site or program. They maximize usability, realizing well enough that usability is certainly contextual. The same applies, as it should, to learnability: success in learning can be maximized through the product, over and beyond context issues, or in spite of them.

The product view of instruction is an important one, one that is emphasized here. An alternate view, much more widespread, is a process one: learning is a process, and so is instruction in the sense of manipulating the situation so as to facilitate learning. This is why the immense amount of research on learning and education over the past century has not dealt explicitly with learnability.

The process view is not to be denigrated, but a product view can incorporate processes and has definite design advantages. Learnability is best considered in this light.

LEADING QUESTIONS

The challenge before us is to identify those features of excellent learning materials. What makes something learnable? Very learnable, most learnable?

But first, why is it so difficult to pinpoint these features? What are the deep issues underlying learnability? There are three of them we need to consider. They are learning, design, and curriculum. Each is difficult in its own right and learnability involves considering them jointly – hence the magnitude of the challenge.

The first deep question is what is learning? The field of learning has long been a core issue in psychology and numerous theories of learning have been put forth in answer (Kearsley, 2004). The issue is far from settled, as practitioners such as educators well know. There is acknowledgment of different kinds of learning, with different factors at play, but no large agreement on these or on the overall picture.

The second deep question concerns teaching. How do you design for learning? There are general principles that have evolved over time, codified broadly in what is known as the field of instructional design (Reigeluth, 1999). But

here too, there is hardly agreement. All design theorists will subscribe to general systems principles like those found in software design or in HCI. All subscribe to the value of usability testing, the trying out of the materials designed with sample students in order to verify the strength of the design and capture any ways of improvement. But given divergences in views of learning, it is natural that hard disagreements will occur here too, in how to design for learning.

The third deep question concerns what to teach - the content. That was what led educators to determine and discuss taxonomies of learning objectives half a century ago (Bloom, 1956) and why this issue remains at the heart of much debate in education (Egan, 1997).

At first thought, you might think that this is an outside issue. That first, we decide what to teach, then only after that, how to teach it, how to design it. Or we might think that teachers and curriculum specialists, or professors and institutions, determine the content "to be covered". That learnability applies to any content, whatever it is determined it should be. But that overlooks the crucial notion that the *what* and the *how* of learning are inextricably linked (Carroll, 1990), just as in communication more generally. An instructional designer must fashion the content as much as the process, in the same way an information designer fashions information well beyond the graphic design aspect. Both are information architects, but that is not yet widely recognized, which creates difficulties for the acceptance of learnability.

In the next sections, I will address these leading questions by introducing some simple models that synthesize them in a nutshell. This remains a very cursory look at the issues, but nevertheless shows the direction in which they can be further explored, as is done in Duchastel (2003).

LEARNING - THE CIM MODEL

At its most general, learning is the process of internalizing information in memory, making that information available later on when needed. But learning the names of the bones in the body and learning the principles of acoustics are rather different forms of learning. We learn them in different ways. What are the commonalties? What are the differences?

There are three types of learning, conveniently contrasted in what we can call the CIM model. CIM stands for Comprehension, Interest, Memorizing, these being the three factors involved in the learning process.

Comprehension is based on our ability to reason, to fit things together, to see how they all work together. Comprehension is the process of generating internal models of the world in all its workings, large and small. We comprehend when we see how things fit together, how it all makes sense. Understanding is a process of rational model building.

Interest, the second element in CIM, is the attentional factor in learning. If something stands out from its context, it will be more easily remembered, as will things that are extremely vivid or of great personal importance. More often, we try to learn things that are only of mild interest and then, if attention wanders, learning suffers. Interest has the function of keeping us on task.

The third element, memorizing, handles things that do not fit well together, that have no basis in rationality. For instance, the name "cochlea" to represent one of the components of the ear is quite arbitrary to us – there is no reason for it [no reason that we know]. It is [to us] purely arbitrary and no amount of reasoning will assist in "understanding" it. We just have to associate the name and the component.

DESIGN - THE MOCAF MODEL

Based on the CIM model, we can see that there will be three types of elements that are needed within an instructional product: models, cases and facts. Combining these (and any product would have all three) leads to the acronym MoCaF for the design model appropriate for the creation of highly learnable instructional products.

Models are the tools of understanding; they are what lead to comprehension. Cases are the illustrative materials that instantiate the models in particular settings. They are the main means of grabbing and holding attention. As for facts, they are just the basics that need to be brutely memorized.

Models are what drive comprehension. The aim of design in this area is to create models that embody the disparate elements of content while synthesizing them in an artifact [the model] that clearly communicates and is easily learned. Models show how elements relate to one another; they capture relationships and interactions.

The craft of developing models is one of establishing the underlying structures in a field [content expertise is essential here] and of then representing those structures in synthetic form that facilitate communication and understanding (Wurman, 2000).

Cases are the illustrative material in instructional content. They embody the living problems and the living application of the models. They range from simple examples to complex case studies. Of particular interest are those relatively complex cases that mirror difficult real-life settings, such as those used in problem-based medical education or in business education.

Cases are multi-functional in an instructional application. At least three functions can be served:

1. To illustrate the content of a model, instantiating it and situating it in real life.
2. To provide practice to the student in applying knowledge.

3. To test the student's knowledge (either self-testing for monitoring purposes or formal testing for assessment and validation purposes).

Facts are those ill-fitting elements of knowledge that are considered important to know and that hang out there on their own, only incidentally attached to some model or other.

Facts are simple to state, for instance in a textbook or in a presentation. But that does not ensure they will be learned. While simple to state, facts are hard to enrobe in a context that will make them easily learnable. Practice or an eventful context is needed. There are means to accomplish this in instructional terms, such as through games, problems, contests, high-impact media, and so forth. Often, though, when these are not developed, the student is left pretty much to his or her own devices for rehearsing the facts to be learned. This is not an optimal situation.

Learning involves interacting with information. And so it is the design of that information that is crucial to learnability. We are dealing here with the content of the instructional product, that content being modeled through design into a certain form that makes it understandable, interesting and memorizable.

Models, cases, facts are all basically information content of particular sorts, information with which the learner will interact during learning. To a very large extent, then, instructional design is mostly a matter of information design, a notion that needs to become widely recognized. Even the more recent notion of interaction design [often applied to Web site design or to exhibition design] is largely a matter of information design involving models (structures), cases (events), and facts (impressions). In sum, learnability must focus primarily on the content to be included in an instructional product, that is, on information design. That is the key contribution of the learnability approach to design and its central usefulness in the practical design of knowledge artifacts.

FUTURE TRENDS

Perhaps the greatest trend emerging in the future with respect to learnability will be the continuing merging of instructional design into information design. As access to information becomes more ready, we will likely see a reduction of our need to memorize arbitrary information beyond the frequently used or crucial to know kind. Our external memory supports will fill the need for the less needed information.

This merging of the two traditionally distinct design worlds (information and instruction) is particularly informative for the learner-control issue in education. Adult learners like to have more control (or like to think they do) over what they learn, how they learn it and when they learn. They operate more in an access mode than in a traditionally receptive educational mode.

Well-designed information/instruction products will facilitate this approach, being used at times for informational purposes and at other times for instructional purposes. Informal learning (outside of academic structures) and formal instruction both involve learning, both involve interacting with information, both profit from good information design.

The design of e-learning materials also might offer more means of controlling interaction than does for instance the design of textbook or other printed materials. This may or may not be an advantage, depending on a whole host of factors, such as maturity of the learners, prior knowledge of the learners and other context factors. But it does raise once again the general issue of content vs. process.

This philosophical issue remains a challenging one, as well as a thorn in any attempt to devise an overarching theory of learning and instruction. One facet of this issue, and I will conclude with it, is why we speak of the learnability of instructional products. After all, we learn also a great deal from interacting with the world at large, not just with artifacts.

The way to come to grips with this issue is to adopt a wide conception of information, as does the field of semiotics. Information goes far beyond the written word, and beyond the world of illustration too. Information is structure that lies within the world around us, both in its structural elements and its processes (Duchastel, 2002). Some of these are found elements, others are designed artifacts, ones the design of which we can control. This is where we can affect the learnability of a product or of a structured process.

REFERENCES

Bloom, B. (Ed.) (1956). *Taxonomy of educational objectives, Handbook 1: Cognitive domain.* New York: Addison-Wesley.

Carroll, J. (1990). *The Nurnberg funnel*. Cambridge, MA: The MIT Press.

Duchastel, P. (2002). Information interaction. *Proceedings of the Third International Cyberspace Conference on Ergonomics*. Retrieved on March 24, 2004, CD-ROM available through *http://cyberg.wits.ac.za/backg' 2005.html*

Duchastel, P. (2003). Learnability. In C. Ghaoui (Ed.), *Usability evaluation of online learning programs* (pp.299-312). Hershey, PA: Idea Group Publishing.

Egan, K. (1997). *The educated mind.* Chicago: U. of Chicago Press.

Folmer, E., & Bosch, J. (2004). Architecting for usability: A survey. Journal of Systems and Software, *70*, 61-78.

Kearsley, G. (2004). *Explorations in learning & instruction: The theory into practice database*. Retrieved on March 24, 2004 from *http://tip.psychology.org/*

McCarthy, J. (2001). Optimal language learning. *Trends in Cognitive Sciences*, *5*, 132-133.

Nielsen, J. (2000). *Designing Web usability*. Indianapolis: New Riders Publishing.

Reigeluth, C. (1999) (Ed.). *Instructional design theories and models: A new paradigm of instructional theory, Volume II*. Mahwah, NJ: Erlbaum.

Wurman, R.S. (2000). *Information anxiety 2*. Indianapolis, IN: Que Publishing.

Valiant, L. (2000). Robust logics. *Artificial Intelligence*, *117*, 231-253.

KEY TERMS

Information Design: A similar soft technology applied to information more broadly for the purpose of successful access to information.

Instructional Design: The soft technology of organizing learning materials and events so that instruction will be most successful.

Interaction Design: A similar soft technology focusing on the processes of interacting with information, particularly in high-impact or strongly emotive contexts.

Learning: The processes used by organisms, including humans, to augment their knowledge base or their skill set for the purposes of better adaptation to their milieu.

Machine Learning: The processes used to fine-tune a program's performance or to augment its knowledge and functionality.

Online Course: A Web-based instructional program that organizes the learning of a student in a particular subject. Not all learning materials need be online and much of an online course involves dynamic interactions with other participants.

Usability: The ease with which a user can accomplish a desired task within a Web site. One also talks of a site being user-friendly.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1803-1806,

Learning Systems Engineering

Valentina Plekhanova
School of Computing and Technology, University of Sunderland, UK

INTRODUCTION

This chapter presents a project proposal, which defines future work in engineering the learning systems. This proposal outlines a number of directions in the fields of systems engineering, machine learning, knowledge engineering, and profile theory, that lead to the development of formal methods for the modeling and engineering of learning systems. This chapter describes a framework for formalisation and engineering the cognitive processes, which is based on applications of computational methods. The proposed work studies cognitive processes, and considers a cognitive system as a multi-agents system of human-cognitive agents. It is important to note that this framework can be applied to different types of learning systems, and there are various techniques from different theories (e.g., system theory, quantum theory, neural networks) can be used for the description of cognitive systems, which in turn can be represented by different types of cognitive agents.

BACKGROUND

Traditionally multi-agent learning is considered as the intersection of two subfields of artificial intelligence: multi-agent systems and machine learning. Conventional machine learning involves a single agent that is trying to maximize some utility function without any awareness of existence of other agents in the environment (Mitchell, 1997). Meanwhile, multi-agent systems consider mechanisms for the interaction of autonomous agents. Learning system is defined as a system where an agent learns to interact with other agents (e.g., Clouse, 1996; Crites & Barto, 1998; Parsons, Wooldridge & Amgoud, 2003). There are two problems that agents need to overcome in order to interact with each other to reach their individual or shared goals: since agents can be available/unavailable (i.e., they might appear and/or disappear at any time), they must be able to find each other, and they must be able to interact (Jennings, Sycara & Wooldridge, 1998).

Contemporary approaches to the modeling of learning systems in a multi-agent setting do not analyze nature of learning/cognitive tasks and quality of agents' resources that have impact on the formation of multi-agent system and its learning performance. It is recognized that in most cognitively driven tasks, consideration of agents' resource quality and their management may provide considerable improvement of performance process. However, most existing process models and conventional resource management approaches do not consider cognitive processes and agents' resource quality (e.g., Norman, Preece, Chalmers, Jennings, Luck & Dang, 2003). Instead they overemphasize the technical components, resource existence/availability problems. For this reason, their practical utilisation is restricted to those applications where agents' resources are not a critical variable. Formal representation and incorporation of cognitive processes in modeling frameworks is seen as very challenging for systems engineering research.

Therefore, future work in engineering the learning processes in cognitive system is considered with an emphasis on cognitive processes and knowledge/skills of cognitive agents as a resource in performance processes. There are many issues that need new and further research in engineering cognitive processes in learning system. New/novel directions in the fields of systems engineering, machine learning, knowledge engineering, and mathematical theories should be outlined to lead to the development of formal methods for the modeling and engineering of learning systems. This article describes a framework for formalisation and engineering the cognitive processes, which is based on applications of computational methods. The proposed work studies cognitive processes, and considers a cognitive system as a multi-agents system.

This project brings together work in systems engineering, knowledge engineering and machine learning for modelling cognitive systems and cognitive processes. A synthesis of formal methods and heuristic approaches to engineering tasks is used for the evaluation, comparison, analysis, evolution, and improvement of cognitive processes.

In order to define learning processes, cognitive processes are engineered via a study of knowledge capabilities of cognitive systems. We are not interested in chaotic activities and interactions between cognitive agents (since cognitive tasks require self-managing activities/work), nor interested in detailed tasks descriptions, detailed steps of tasks performance and internal pathways of thoughts. Rather, we are interested in how available knowledge/skills of cognitive agents satisfy required knowledge/skills for the performance of the cognitive tasks.

The proposed research addresses the problem of cognitive system formation with respect to the given cognitive tasks and considers the cognitive agent's capabilities and compatibilities factors as critical variables, because these factors have an impact on the formation of cognitive systems,

the quality of performance processes and applications of different learning methods.

It is recognised that different initial knowledge capabilities of the cognitive system define different performance and require different hybrid learning methods. This work studies how cognitive agents utilise their knowledge for learning the cognitive tasks. Learning methods lead the cognitive agent to the solution of cognitive tasks. The proposed research considers a learning method as a guide to the successful performance. That is, initial knowledge capabilities of cognitive agents are correlated with learning methods that define cognitive processes. An analysis of impact of different cognitive processes on the performance of cognitive agents is provided.

This work ensures support for a solution to resource-based problems in knowledge integration and scheduling of cognitive processes to form a capable cognitive system for learning the required tasks.

AIMS AND OBJECTIVES

The aims of the project are to develop a formal method for the modelling and engineering of cognitive processes. Capability and compatibility factors have an impact on the formation of cognitive systems, the performance processes and define different learning methods. Therefore this work studies cognitive processes and knowledge capabilities of cognitive systems to ensure the required level of the learning and performance of the cognitive systems. In order to support the formation of a cognitive system that will be capable of learning the required tasks within the given constrains, this work addresses problems of the knowledge integration and scheduling for cognitive system modeling, taking into account critical capability and compatibility factors. Study of learning conditions in cognitive systems defines an important task of the proposed project.

The individual measurable objectives are:

- Evaluation of knowledge integration and scheduling approaches in cognitive systems.
- Evaluation of existing machine learning approaches in cognitive systems.
- Determination of the impact of capability and compatibility factors on the formation of cognitive systems.
- Development of knowledge integration metrics.
- Development of knowledge integration models for the formation of the cognitive systems.
- Development of scheduling models for learning of cognitive systems.

METHODOLOGY AND JUSTIFICATION

In order to identify the best learning processes we analyze the cognitive processes. A scenario for engineering the cognitive processes is based on the following steps (Figure 1).

The methodology of the proposed project is based on the following new theoretical basis (Plekhanova, 2003).

Profile Theory and Machine Learning: For formal modeling of complex systems we utilise the profile theory (Plekhanova, 1999a). A profile is considered as a method for describing and registering multifaceted properties of objects. There are important practical applications of the profile theory (Plekhanova, 2000a, 2000b). For instance, internal properties of the system elements such as capability and compatibility factors are critical variables in modeling, design, integration, development, and management of most

Figure 1. A scenario for engineering the cognitive processes

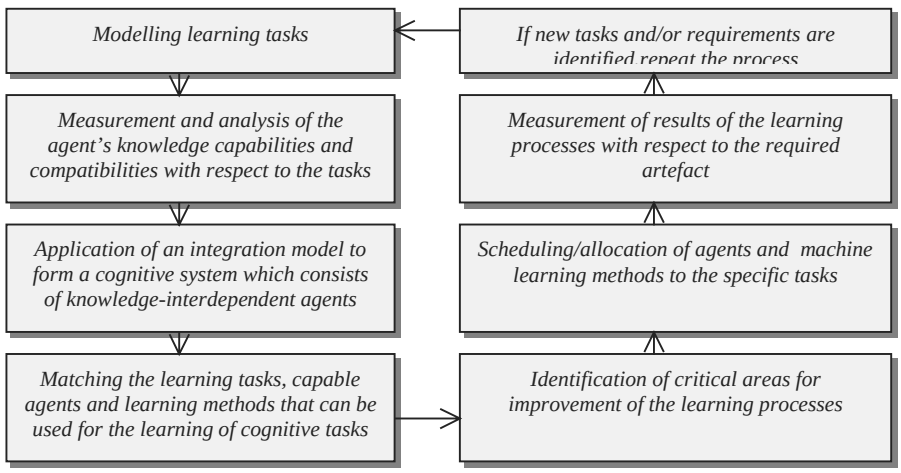


Table 1. A comparison of capability and compatibility problems

Technical Systems	Soft Systems
In *technical systems* the internal characteristics of technical elements are described in specifications, standards and formal documents [i.e. are known a-priori] from which it is not difficult to conclude whether combinations of elements are capable and compatible or not and whether they can be used for technical system design, development and construction. Each capability and compatibility factor can be represented by *one characteristic*.	There are a number of real world examples when an object factor cannot be described by just one characteristic. For example, systems such as human resources, software, information systems, cognitive agents, where the internal *multifaceted* properties can be changed with time, and capability and compatibility factors cannot be defined by one characteristic alone, and cannot be explicitly measured. These systems are termed *soft systems*. At the present time there are problems in the formal definition of specifications/standards and metrics that allow one to determine the capability and compatibility of such complex systems. For this reason, heuristic approaches are used for soft complex system modelling.

modern complex systems and their structure. Table 1 provides a comparison of capability and compatibility problems of technical and soft systems.

Profile theory is used for formal modeling of cognitive agents/systems since existing mathematical theories are limited. In particular, contemporary mathematical theories describe objects, where each internal factor is represented by one meaningful piece (e.g., set theory—an element) or two pieces of meaningful information (e.g., fuzzy set theory—an element and a membership function).

Knowledge factors are considered as basic internal factors in the modeling of cognitive agents, since agents must have particular knowledge capabilities to perform and learn their tasks. In a description of the knowledge of cognitive agents we identify the importance/weight of the factor for the performance of the task; time or factor existence/nonexistence; and other specific internal multifaceted properties, for example the property (level, grade, degree) of the factor.

In particular, knowledge of the cognitive agent is described by a set of knowledge factors; each factor is defined by multiple characteristics. A set of such factors forms a knowledge profile (Plekhanova, 1999a). Each factor is represented by qualitative and quantitative information. Quantitative description of the *i*th knowledge factor is defined by an indicator characteristic, property, and weight. In a simple way, a profile can be defined as follows (see Figure 2).

The profile theory is used for formalization of cognitive systems and cognitive processes, and for the identification of critical areas in learning performance where improvement should be taken. In particular, engineering the cognitive processes is considered to provide improvement of learning process by means of integrating adaptive machine learning into the profile theory. In order to model cognitive processes the profile theory is combined with machine learning methods, which are applied to the initial available knowledge capabilities of the cognitive system to define learning methods. (It is expected that different initial knowledge capabilities of the cognitive system require different hybrid learning methods.)

Machine learning methods are used for formalization and modeling of learning processes via applications of the profiles. It allows consideration of dynamics in learning processes (i.e., modeling of the *i*th profile factor $e_i(t) = < \varepsilon_i(t), v_i(t), w_i(t) >$ in the profile *b*). We should analyze existing machine learning methods, match them to learning tasks with relevance to available knowledge capabilities of cognitive agents and consider cognitive processes. A profile is considered as a model for the description of cognitive processes. That is, a new machine learning method will be developed and incorporated into an engineering framework for cognitive processes.

This research considers knowledge factors as critical variables in learning processes and addresses problems in the formation of a cognitive system that can be capable of learning. In particular, a cognitive system is defined by knowledge-interrelated agents, their flexible cognitive structure and cognitive processes. A teacher is defined as a learning oracle. Soft factors may be defined as a "noise" in data modeling for the training sets in machine learning.

This work addresses the problems of knowledge integration of the cognitive system in order to provide a better learning performance. A challenge for learning is to ensure

Figure 2. Definition of the profile

A profile b is defined as a set of factors $b_1, b_2, ..., b_n$: $b = \{b_i, i = \overline{1,n}\}$, where the ith factor b_i is represented by a pair $b_i = (\iota_i, e_i)$ with

- n - a number of factors
- ι_i - an identification of the ith factor, i.e. a name or label or type of the ith factor
- e_i - the 3-tuple of the ith factor as the Cartesian product: $e_i = <\varepsilon_i, v_i, w_i>$, where
  - ε_i - indicator characteristic, that indicates the factor presence in the description of a cognitive agent, the existence of certain conditions, e. g. ε_i may represent a binary case; a number of times of factor utilisation; or may be defined as a time characteristic $\varepsilon_i = \varepsilon_i(t)$
  - v_i - property of the ith factor: $v_i \geq 0$; v_i can be defined as a function of time $v_i = v_i(t)$
  - w_i - weight of a factor which defines either the factor importance or the factor priority: $w_i \geq 0$; w_i can be also considered as a function of time $w_i = w_i(t)$.

the existence of a desired level of performance of a cognitive system. There is a need to make a formal analysis of the available knowledge of cognitive agents in order to ensure the learning of the tasks at a desired performance level while utilizing the available knowledge capabilities effectively and efficiently.

This research deals with the problem of agent allocation in a cognitive system. This problem addresses not only task scheduling as in traditional approaches but also scheduling machine learning methods and knowledge of cognitive agents. The proposed project will develop a new scheduling approach where the agent allocation problem has specific emphasis on the following aspects: cognitive agents are allocated to tasks according to their multiple knowledge capabilities; the agent's knowledge capabilities must satisfy the particular combination of knowledge required for a task; agents of the cognitive systems should be compatible with each other (Plekhanova, 1999b); and learning methods are relevant to the available knowledge capabilities of the cognitive agents and system. The consideration of all these aspects defines a problem of knowledge integration in cognitive systems. This work will use formal methods for an integration of cognitive capabilities and compatibilities, and for an analysis of how system capabilities satisfy the learning of the tasks.

Capability and compatibility factors have considerable impact on the process of system integration. An integration model encompasses integration criteria, priorities of the knowledge profiles and knowledge integration goals. Knowledge integration goals are the improvement of available knowledge or generation of new/novel knowledge for better learning performance. This work addresses problems of effectiveness of the learning processes, their convergence (Vapnik, 1998), stability, and accuracy.

Therefore, the adventure in this research is that cognitive processes will be incorporated into multi-agent system development by a synthesis of systems engineering with knowledge engineering and machine learning methods. The combination of machine learning methods with profile theory will provide a more flexible adaptive framework for cognitive tasks performance. That is, the proposed method for the modeling and engineering of cognitive systems and cognitive processes can be used in systems engineering and machine learning for a formalization of cognitive processes, cognitive systems, and capability and compatibility aspects.

FUTURE TRENDS

The proposed project is particularly novel in its approach to learning processes that incorporate a synthesis of systems engineering, knowledge engineering and machine learning methods. There are no formal methods for knowledge integration and scheduling for learning of cognitive systems where capability and compatibility factors are critical variables. Existing machine learning approaches do not address scheduling problems in learning methods. We will develop a new scheduling approach where we consider scheduling machine learning methods and knowledge of cognitive agents vs. task scheduling in traditional approaches. New machine learning methods will be developed and incorporated into an engineering framework for cognitive processes. Moreover, the

proposed project brings together work in cognitive systems, systems engineering, knowledge engineering and machine learning for the modeling of cognitive processes.

The proposed project is timely because of the availability of new formal methods for engineering cognitive systems. The work is highly topical at present as demonstrated by a large interest in academia and great needs of industry, in particular, in:

- **Software/Systems Engineering:** Project resource capability and compatibility aspects have become the focus of performance process improvement. However, most contemporary approaches to the formation of project resources (Norman et al., 2003) do not examine their capability and compatibility factors. There is a need to develop evaluation techniques for people's capability, resource capability and compatibility in order to provide support for effective solutions to resource integration management in cognitive systems (Plekhanova, 2002). In particular, methods are of particular merit that incorporates a comparison of cognitive processes, resource capabilities, and compatibilities.
- **Scheduling:** Contemporary approaches to resource scheduling are based on the detailed description of tasks assuming that a resource pool is given and defined by a manager and resources are capable of performing any project task. Existing resource scheduling methods address the issues of resource availability and utilization, and are not concerned with the capability and compatibility of project resources. Furthermore, in traditional scheduling approaches, the objectives for the allocation of limited resources are to determine the allocation of resources that maximize total benefits subject to the limited resource availability. Contemporary approaches to resource allocation are founded on the assumption that different tasks require equal capability resources, and only one skill is involved. Hence, they cannot be successfully used for software projects where different software tasks require changing different sets of multiple knowledge and skill capabilities in an overall system (Plekhanova, 1998).
- **Software Tools for Resource Scheduling:** There are many scheduling tools that provide different approaches: event-oriented (PERT), activity-oriented (CPM), actions-oriented (TASKey PERSONAL), or offer a wide variety of scheduling options (SAP). Nevertheless, there are no tools that support an analysis of resource capabilities/compatibilities and their impact on project scheduling (Plekhanova, 2000c). Most existing tools (Microsoft Project, SAP, Up and Running) have facilities for entering new resources, but do not deal with an analysis of cognitive processes, and resource quality based on which resources can be added to the resource pool. Therefore, the existing scheduling tools cannot be effectively used for management of processes where resources are a critical variable.
- **Theory/Tools in Machine Learning:** Existing machine learning techniques (e.g., Boosting (Schapire, 1999), Lazy Learning (Aha, 1997), Neural Nets, Decision Tree Learning (Quinlan, 1990; Utgoff, 1989), Support Vector Machine (Vapnik, 1998), Reinforcement Learning (Sutton & Barto, 1998)), and contemporary machine learning tools (e.g., WEKA, AutoClass, mySVM) have not yet been examined in terms of agents' capability/compatibility and scheduling problems.

There is a direct relationship between the representation and the learning mechanisms. In many cases the underlying representations in machine learning have been of limited structure (e.g., vectors, trees, networks). Hybrid integration of various machine-learning mechanisms for engineering of structured objects is novel and will be examined in this project in the context of the profile theory.

CONCLUSION

Beneficiaries of the proposed research are:

Engineering the Complex Systems: Research in engineering of complex systems will provide insight into new methods and approaches to learning in cognitive systems. Research in machine learning will deliver adaptiveness to knowledge integration and scheduling of learning methods. Scientists in cognitive systems research will receive a formal method for modeling of cognitive processes. By developing integration metrics using the profile theory we can provide analysis, development, integration, modeling and management of complex systems and their elements where weight, time, and other internal multifaceted properties are critical variables. Further development of the profile theory will establish a new branch in mathematics and extend its applications.

Industry: New evaluation techniques could provide support for a solution to the resource-based problems in cognitive processes in software and IT projects such as team formation and integration in connection with process tasks.

The application of a new approach could provide learning organizations with:

- Superior management of resource capabilities and compatibilities;
- Streamlining of process development through better management of project resources and tasks;

- Increased opportunities for organizations to implement process improvement based on the constructive criticism derived from self-analysis.

It is apparent that there is a worldwide interest in the application of this research. Since most modern processes are cognitively driven our method can be used for the formal modeling of cognitive systems. It is important for the future competitiveness of the software and IT industry to employ a scientific (vs. heuristic) approach to the engineering of cognitive processes.

Technology: The profile-based approach assures a virtual prototyping of system development within different environment settings. An important application of this approach is that it gives the means of providing systemic methods of study, analysis, prediction, improvement, control and management of a system development. Moreover, this technology demonstrates a modeling flexibility that permits one to represent a fine-granularity of system components, as well as to generate different system models of a wide diversity of system development processes. Thus, any traditional system model becomes a special case of the capability- and compatibility-based modeling framework.

Formal modeling of the capability and compatibility of cognitive systems ensures the automation in system modeling. It leads to development of new technologies in system modeling. Some of the enhancements that we intend to offer through this method are to provide support for development and engineering of new knowledge capabilities of cognitive systems, that is, innovative technologies.

REFERENCES

Aha, D. (Ed.). (1997). *Lazy learning.* Dordrecht: Kluwer Academic Publisher.

Clouse, J. A. (1996). Learning from an automated training agent. In G. Weiß & S. Sen (Eds.), *Adaptation and learning in multiagent systems*. Berlin: Springer Verlag.

Crites, R. & Barto, A. (1998). Elevator group control using multiple reinforcement learning agents. *Machine Learning, 33,* 235-262.

Jennings, N. R., Sycara, K., & Wooldridge, M. A. (1998). Roadmap of agent research and development. In N.R. Jennings, K. Sycara & M. Georgeff (Eds.), *Autonomous Agents and Multi-Agent Systems Journal, 1*(1), 7-38. Boston: Kluwer Academic Publishers.

Mitchell, T. M. (1997). *Machine learning*. Boston: WCB/ McGraw-Hill.

Norman, T. J., Preece, A., Chalmers, S., Jennings, N. R., Luck, M., Dang, V. D., Nguyen, T. D., Deora, V., Shao, J., Gray, A. & Fiddian, N. (2003). CONOISE: Agent-based formation of virtual organisations. In *Proceedings of the 23rd SGAI International Conference on Innovative Techniques and Applications of AI* (pp. 353-366). Cambridge, UK.

Parsons, S., Wooldridge, M., & Amgoud, L. (2003). Properties and complexity of some formal interagent dialogues. *Journal of Logic & Computation,13*(3), 347-376.

Plekhanova, V. (1998). On project management scheduling where human resource is a critical variable. In *Proceedings of the Sixth European Workshop on Software Process Technology, Lecture Notes in Computer Science Series* (pp. 116-121). London: Springer-Verlag.

Plekhanova, V. (1999a). *A capability- and compatibility-based approach to software process modelling*. Unpublished doctoral thesis, Macquarie University, Sydney, Australia and the Institute of Information Technologies and Applied Mathematics, Russian Academy of Sciences.

Plekhanova, V. (1999b). Capability and compatibility measurement in software process improvement. In *Proceedings of the 2nd European Software Measurement Conference*, Amsterdam, Netherlands, Technological Institute Publications (pp. 179-188). Antwerp, Belgium.

Plekhanova, V. (2000a). Profile theory and its applications. In *Proceedings of the International Conference on Information Society on the 21st Century: Emerging Technologies and New Challenges* (pp. 237-240). The University of Aizu, Fukushima, Japan.

Plekhanova, V. (2000b). Applications of the profile theory to software engineering and knowledge engineering. In *Proceedings of the Twelfth International Conference on Software Engineering and Knowledge Engineering* (pp. 133-141). Chicago, Illinois.

Plekhanova, V. (2000c). On the compatibility of contemporary project management tools with software project management. In *Proceedings of the 4th World Multiconference on Systemics, Cybernetics* (Vol. I, pp. 71-76). Orlando, Florida.

Plekhanova, V. (2002). Concurrent engineering: Cognitive systems and knowledge integration. In *Proceedings of the 9th European Concurrent Engineering Conference* (pp. 26-31). Modena, Italy: SCS Europe (Society for Computer Simulation).

Plekhanova, V. (2003). Learning systems and their engineering: A project proposal. In J. Peckham & S. Lloy (Eds.), *Practicing software engineering in the 21st century* (pp. 164-177). Hershey, PA: Idea Group Publishing.

Quinlan, J. R. (1990). Probabilistic decision trees. In In Y. Kodratoff & R. S. Michalski (Eds.), *Machine learning:*

An artificial intelligence approach (Vol. 3, pp. 140-152). California: Morgan Kaufmann Publishers, Inc.

Schapire, R. (1999). Theoretical views of boosting and applications. In O. Watanabe & T. Yokomori (Eds.), In *Proceedings of the Tenth International Conference on Algorithmic Learning Theory* (pp. 13-25).

Sutton, R. S. & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge: MIT Press.

Utgoff, P. E. (1989). Incremental induction of decision trees. *Machine Learning, 4*, 161-186.

Vapnik, V. N. (1998). *Statistical learning theory.* Chichester: Wiley.

KEY TERMS

Agent: A complex system constituting elements that are individual performers, which can be described by their interrelationships, knowledge/skill, performance and constraints factors.

Agent's Compatibility: A capability of agent to work with other agents without adaptation, adjustment and modification.

Cognitive Process:

- The performance of some composite cognitive activity.
- A set of connected series of cognitive activities intended to reach a goal. Cognitive activities can be considered as a function of their embodied experience.

Cognitive System: A complex system that learns and develops knowledge. It can be a human, a group, an organization, an agent, a computer, or some combination. It can provide computational representations of human cognitive processes to augment the cognitive capacities of human agents.

Complex System: A collection of interrelated elements organized to accomplish a specific function or a set of functions. Complexity can be considered in terms of a number of elements and/or complexity of relationships.

Learning System: A complex system that learns and develops knowledge.

Machine Learning: The ability of a machine to improve its performance based on previous results.

Multi-Agent System: A set of interrelated agents that work together to perform tasks.

Legal Issues of Virtual Organizations

Claudia Cevenini
CIRSFID, University of Bologna, Italy

INTRODUCTION

In the present economic context, organizations, especially of small and medium dimensions, can draw a substantial advantage by collaborating and setting up flexible, temporary ICT-enabled networks.

Identifying the legal issues relevant for virtual organizations can provide a knowledge basis to regulate their activities, thus providing support for their creation and management.

BACKGROUND

The concept of virtual organization (VO) finds its origins in the United States in the early 1990s, when some authors start to give it a first theoretical outline.

Since then, a certain scientific debate has opened, and several attempts to define and concretise it have been made. The most active research sectors with respect to this appear to be business and computer science. Until recently, however, the legal research has substantially disregarded VOs, with few exceptions.

REGULATING VOS

VOs are far from being a consolidated reality; being fluid and flexible structures, they continually evolve over time and are difficult to grasp.

The starting point for their regulation is to provide a definition for the purposes of legal research: "VO's are ICT-enabled collaborations between legally independent subjects aimed at the joint provision of goods or services, where each partner contributes to specific activities. They do not aim at achieving an autonomous legal status but appear as one organization towards third parties."[1]

As a second step, a wide range of legal issues concerning them can be identified. By developing a legal taxonomy, it is possible to aggregate legal problems in major research areas. This makes it possible to focus on those issues most connected with the particular structure and nature of the VO.

The third step is to examine the identified issues in the light of the applicable legal framework at national and international level, considering the nationality of the partners and their reciprocal agreements.

A TAXONOMY OF VO-RELATED LEGAL ISSUES

Hereinafter, a synthesis of the most relevant issues is presented.

Identity and Nationality of the Virtual Organization

The VO does not embody a formal institution separate from its partners, although it may appear as a separate, autonomous entity.

National legal orders will tend to consider the VO as a structure without legal personality and, consequently, also without nationality, provided that the partners do not opt to formally adhere to a company type as foreseen by the national law of one of them.

Role of the Virtual Organization Broker

A VO can be set up and managed without the intervention of a VO broker. This, however, would imply higher coordination costs, more complex negotiations, and a slower speed of action.

The legal status of the VO broker depends upon its actual role and activities in the VO. The broker will be subject to and have to comply with the applicable legal framework set for the legal structure it has opted for in the state in which its head office is located, as well as with the state- and contract-based rules applicable to the same VO.

Virtual Organization Framework Agreement

The VO framework agreement is a set of rules aimed at governing the internal relationships between the partners of a VO.

It has to be signed before the beginning of any activity and is generally drafted with the support of the broker, who may propose business templates on the basis of which the detailed final provisions can be negotiated with the partners.

The absence of a clear agreement would possibly lead to difficulties in the management of the operation stage and, later, to possible disputes between the partners.

Contracting with Third Parties

Having no legal personality, the VO cannot directly close contracts with third parties. Therefore, if its members are to enter into contractual relationships, this will not be feasible for the VO as a separate subject; agreements can only be closed between third parties and some or all the individual members.

Once a partner—or a group of partners—has been selected, the other members can grant to it the power to act in their name and on their behalf to the purpose of closing contracts binding for all of them, or to take care of other jural acts, as it happens with mandates.

The Resolution in Disputes

The involvement in a lengthy dispute resolution procedure can cause severe economic damage or even disrupt a temporary entity like the VO. For this reason, before the final framework agreement between the partners is signed and, later, before the signing of every agreement with third parties, attention has to be placed upon developing and agreeing upon adequate dispute resolution mechanisms, which may range from legal actions before courts, to arbitration and mediation.

Liability Issues

It would be difficult to configure a liability on the VO as such, as there are no legal instruments to construct it. It thus appears more feasible to identify a liability on the individual partners. All of them may be held jointly and severally liable for the damages that can be imputed to the VO. Whenever one partner is held liable, those who have been sued without having contributed to causing the damage which is the object of the claim can resort to an internal redress.

Intellectual Property Rights

The VO activities are based on the reciprocal disclosure of relevant data and possibly on the sharing of immaterial goods. These can, in some cases, enjoy a precise legal protection, as happens with copyright, software, patents, and databases.

This applies both to the data and goods to which the individual partners are already entitled, as well as to the outcome of their collaboration. In the former case, the legitimate owner can grant to the other partners a right of economic exploitation, for example, through licensing contracts. As to the latter case, specific agreements are to be clearly set.

If the individual contributions are not to be clearly identified, it can be assumed that all partners will be entitled to the data and goods produced by the VO and of the relative rights, on the basis of a coownership.

Data Protection

All the different activities of the VO imply the processing of personal data. Within the present data protection framework, attention has to be placed, in particular, on a series of elements, such as the processing through automatic means, the disclosure of data to third parties, and the transfer of data between European Union (EU) and non-EU countries. The VO partners will actually carry out most processing with ICT tools and may need to process data originally collected by one or more of them and perform cross-border data transfers.

Competition Law

Should a VO achieve a substantial dimension in terms of its overall turnover, attention shall be placed by the partners to its compliance with rules on antitrust and the protection of competition applicable to the partners. Specific procedures may be imposed in order to get the authorization of antitrust bodies, as well as to verify the law-abidingness of the collaboration.

ICT-Related Issues

The nature of ICT-enabled entities possessed by VOs requires their compliance with the applicable rules on the use of specific technology tools, for example, with reference to security, electronic signatures, e-commerce, or teleworking. Should certain ICT-based interactions not be specifically regulated, reference shall be made to analogically applicable norms.

FUTURE TRENDS

The present economic and legal scenario does not point toward the drafting of an ad hoc legislation for VOs. However, the growing relevance of collaborative entities, such as industrial districts, coupled with the strong support of the techno-legal scientific community, will make VOs known to a wider audience and stimulate targeted initiatives.

CONCLUSION

A coherent and certain legal framework applicable to regulate VOs is presently still absent, and a clear qualification of their legal identity by the legislator is missing. Besides, their international character would require a level of harmonisation among contrasting rules that does not always appear easy to achieve, while their massive use of rapidly evolving information and communication technologies contrasts with the slow law-making process.

This makes it extremely difficult to envisage a state-based regulation for them, or the drafting of international treaties or agreements, at least in the short term.

This enhances the fundamental role of normative tools, such as codes of conduct and best practices, and especially of contractual agreements and intraorganisational rules drafted by the VO partners, developed on the basis of a clear identification of the relevant legal issues.

REFERENCES

Berwanger, E. (1999). The legal classification of virtual corporation according to German law. In Seiber, P., Griese, J. (Eds.), *Proceedings of the 2nd VoNet Workshop*, Simowa Verlag, Bern (pp. 158-170).

Cevenini, C. (2003). *Virtual enterprises: Legal issues of online collaberation between undertakings*. Milan: Giuffre, 79-80.

Conaway Stilson, A. E. (1997). The agile virtual corporation. *Delaware Journal of Corporate Law*, 22, 497.

Cousy, H., Van Schoubroeck, C., Droshout, D., & Windey B. (2001). *Virtual enterprise legal issues taxonomy*. Public deliverable D 03, ALIVE working group on Advanced Legal Issues in Virtual Enterprise.

Davidow, W. H., Malone, M. S. (1992). *The virtual corporation—Structuring and revitalizing the corporation for the 21st century*. New York: Harper.

Scholz, C. (1994). Virtuelle unternehmen—Faszination mit Rechtlichen Folgen. *jur-pc Zeitschrift,* (12), 2927-2935.

Sommerlad, K. W. (1996). Virtuelle Unternehmen—Juristisches Niemandsland? *Office Management*, (7–8), 22-23.

Van Schoubroeck, C., Cousy, H., Droshout, D. and Windey, B. (2001). *Virtual enterprise legal issue taxonomy.* In B. Stanford-Smith & E. Chiozza (Eds.), *E-work and e-commerce*. Novel solutions and practices for a global networked economy, I, IOS Press, Amsterdam, 609-615.

KEY TERMS

Arbitration: A private form of conflict resolution. The litigating parties may voluntarily submit a dispute to one or more independent, neutral experts (arbitrators), who decide upon the case similarly to a court, generally upon a shorter period of time. Arbitrations are usually regulated by law.

Intellectual Property Rights: Copyright and connected rights that include, *inter alia*, the right of copying, modifying, and distributing the protected work.

Legal Order: The set of legal norms that make up the legal system of a particular country.

Legal Personality: The capacity of a legal person (e.g., a corporation) to be holder of rights and duties.

Mediation: An alternative dispute resolution method. The litigating parties may voluntarily submit a dispute to a neutral, independent mediator. This latter does not issue a decision but supports the parties in finding a mutually agreed upon solution.

Personal Data: Any information concerning a natural or legal person that can identify it.

Virtual Organization: An ICT-enabled collaboration between legally independent subjects aimed at the joint provision of goods or services, where each partner contributes to specific activities. It does not aim at achieving an autonomous legal status but appears as one organization toward third parties.

VO Broker: A subject who acts as an intermediary for the setting up of the virtual organization by identifying a possible business opportunity, contacting the potential partners, and proposing agreement templates.

ENDNOTE

[1] Cevenini, C. (2003). Virtual Enterprises. Legal Issues of the Online Collaberation between Undertakings. Milan, Guiffre, 79-80

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1831-1833, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Leveraging Complementarity in Creating Business Value for E-Business

Ada Scupola
Roskilde University, Denmark

INTRODUCTION

The rapid developments of Internet and Web-based applications has shaped the era of the digital economy and changed the way enterprises operate. Internet is increasingly becoming part of the basic business model for many companies as organizations around the world are adopting new e-business models and integrated solutions to explore new ways of dealing with customers and business partners, new organizational structures, and adaptable business strategies (Singh & Waddell, 2004). According to Kalakota and Robinson (1999), e-business is the complex fusion of business processes, enterprise applications, and organizational structure necessary to create a high performance business model. E-business is therefore more than just having an Internet presence or conducting e-commerce transactions, it is a new business design that emphasizes a finely tuned integration of customer needs, technology, and processes (Kalakota et al., 1999). When discussing e-business, it is important to make a distinction between physical and digital products. A digital product is defined as a product whose complete value chain can be implemented with the use of electronic networks, for example it can be produced and distributed electronically, and be paid for over digital networks. Examples of digital products are software, news, and journal articles. The companies selling these products are usually Internet-based "digital dot coms" such as Yahoo and Google. On the contrary, a physical product cannot be distributed over electronic networks (e.g., a book, CDs, toys). These products can also be sold on Internet by "physical dot coms," but they are shipped to the consumers. The corporations adopting e-business are distinguished into "bricks and mortar" companies, hybrid "clicks and mortar" companies (such as Amazon.com) and pure dot coms (Barua & Mukhopadhyay, 2000a).

Many studies from the early days of deployment of information technology (IT) in organizations have struggled to measure the business value and profitability of information technology (Barua et al., 2000a). Many of these studies have showed that productivity gains are small or not existent and that the effects of information technology and e-commerce have to be often looked upon from a competitive advantage point of view (Barua, Konana, Whinston, & Yin, 2001; Porter & Miller, 1985; Scupola, 2003). Recent research has argued that to increase the business value of electronic commerce to a corporation is important to shift the focus from whether electronic commerce creates value to a company to "how to create value" and "how to optimize such value" (Barua et al., 2001). This can be achieved by exploring complementary relationships between electronic commerce, strategies and value chain activities (Scupola, 2002, 2003). Here this argument is taken further to show the importance of complementary relationships for the business value of e-business.

BACKGROUND

Since the early days of IT, use in commercial organizations, researchers, and professionals have struggled to understand how and to what extent the application of IT within firms leads to improved organizational performance. The research on IT business value has been characterized by diverse conceptual, theoretical, and analytic approaches as well as has adopted different research methodologies (Melville, Kræmer, & Gurbaxani, 2004). Six main areas of IT business value research can be distinguished: information economics-based studies, early IT impact studies, production economics studies that did not find positive impacts, microeconomics studies that found positive impacts of IT, business value studies and studies involving complementarity between IT and non-IT factors. The information economics-based studies date back to the 1960s and though relevant to the economic contribution of IT investments, they mainly focus on the changes in information due to IT use and their impact on the single decision-maker. Therefore, while the information economics approach is theoretically sound and rigorous, its unit of analysis, which is either the individual or team decision, makes it difficult to obtain meaningful and insightful results in broader organizational contexts (Barua et al., 2000a).

In the early 1980s, a stream of research emerges focusing on assessing the contribution of IT investments to performance measures such as return on investment and market share (Barua et al., 2000a, Barua et al., 2001). The majority of these studies did not find much positive correlation between IT investments and firm performance metrics up to the early 1990s. The lack of correlation between IT investments and firm productivity made Roach (1999) to coin the term "IT productivity paradox."

In the 1990s, the research on measuring the economic and performance contributions of IT can be divided into two main streams: one based on production economics and one based on "process oriented" models of IT value creation. The IT production studies based on production economics hypothesize that IT investments are inputs to a firm's production function. These studies (e.g., Brynjolfsson & Hitt, 1996) finally started finding signs of productivity gains from IT. For example, Brynjolfsson et al. (1996) identify three sources of IT value to a corporation: productivity, consumer value, and business profitability. The study shows that information technology contributes to increases in the productivity and consumer value, but not business profitability. Simultaneously, process-oriented studies started hypothesizing relationships between IT and other input factors to performance measures at various levels of aggregation. These studies (e.g., Kauffman & Kriebel, 1988) have laid the foundation of the business value approach to the impact of IT on firm performance. This approach on the contrary of the production function-based approach might have the explanatory power to point out where and how IT impacts are created and where management should act to increase the payoff from IT investments. These explanations are more difficult to get with production function-based approaches since they operate at a very high level of aggregation, thus making it difficult to distinguish between different types of IT investments and their impacts on specific areas of business. After having dispelled the productivity paradox, new refinements to existing approaches are emerging to measure the contribution of IT to business performance. An important stream of research is pointing to complementarity theory to investigate the interactions between IT and other organizational factors (e.g., Barua, Lee, & Whinston, 1996; Barua et al., 2000a; Barua et al., 2000b; Barua et al., 2001). In fact, production economics and business value approaches have mostly ignored the synergy between IT and other related factors such as the level of fit with business strategies, employee empowerment, and team orientation of business processes. Barua et al. (2000a) present a generalized business value complementarity model that explores the synergies among such factors. The basic idea of their business value complementarity model (BVC) suggests that investments in IT should be first related to intermediate performance measures such as time to market, customer service, response time and extent of product mass customization to be able to see any positive results from such investments. In a second moment, the intermediate performance measures can be related to high-level performance metrics such as profitability, return on investment (ROI), market share. The focal point of a business value complementarity model is the complementarity that potentially exists at each level of the model (Barua et al., 2000a; Barua et al., 2001; Scupola, 2003).

More recent studies are also investigating the impact of information technology on the financial performance of diversified firms (e.g., Shin, 2006), multi-business firms (e.g., Tanriverdi, 2006), and often take their starting point in the resource-based view of the firm as for example the theoretical study conducted by Melville et al. (2004).

The advent of the Internet, based on open standards and a universal Web browser, raises the question of whether investing more in Internet technology lead to a better financial performance in electronic commerce and e-business. In this regard, Zhu (2004) shows that there is a positive interaction between IT infrastructure and e-commerce capabilities suggesting that their complementarity positively contributes to firm performance in terms of sales per employee, inventory turnover, and cost reduction. Further Zhu (2004) provides "empirical evidence to the complementary synergy between front-end e-commerce capability and back-end IT infrastructure (Zhu, 2004, p. 167)." Yang, Yang, and Wu (2005) investigate the relationship between enterprise information portals (EIP) and e-business performance by conducting a survey of companies. Their results show that the implementation of enterprise information portals influence e-business performance. Barua et al. (2004) investigate the processes through which business value is created by Internet-enabled value chain activities. Their analysis suggests that "while most firms are lagging in their supplier-side initiatives relative to the customer side, supplier-side digitization has a strong positive impact on customer-side digitization, which in turn, leads to better financial performance. Further, both customer and supplier readiness to engage in digital interactions are shown to be as important as a firm's internal digitization initiatives, implying that a firm's transformation-related decisions include its customers' and suppliers' resources and incentives" (Barua et al., 2004, p. 585).

These studies point out the need of more attention to the specific business processes that have to be reengineered for e-business and the way they should support the company strategy (Scupola, 1999, 2003). In fact, as Pepper and Ward (2005) say IT has no inherent business value unless this value is unlocked and this process of unlocking business value from IT investments is a journey and not a destination and this journey requires careful planning.

A BUSINESS VALUE COMPLEMENTARITY MODEL OF E-BUSINESS

A business value complementarity model of e-business could be used as a methodology to optimize e-business initiatives when entering the e-business arena (Scupola, 2003). The business value complementarity (BVC) model presented here is based on the value chain (Porter, 1980), the theory of business value complementarity (Barua et al., 1996; Barua et al., 2000a; Barua et al., 2002; Milgrom & Roberts, 1990) and the concept of strategy (Porter, 1982).

In this model it is hypothesized that complementarity (represented in the figure with dotted lines) exist between the variables of the same level and different levels of the model. It is furthermore hypothesized that the exploration of complementarities and possible synergies between the company strategy, the primary activities of the value chain, corresponding business processes and supporting technologies should (1) maximize the business value of e-business to a corporation and (2) lead to a better fit between the overall organizational strategy, the business processes that have to be transformed for the online market place, and the e-business system that should be designed and implemented to support these strategies. The exploration of complementarities, it is hypothesized, can also contribute both to avoid investments into an e-business system that could not be used at a later point if new e-business processes should be added to the system and avoid the implementation of a business model that does not correspond to the corporation's strategy. It is argued that to develop a successful e-business model it is important to reengineer the parts of the value chain and the corresponding business processes relevant to the product in question and the company strategy.

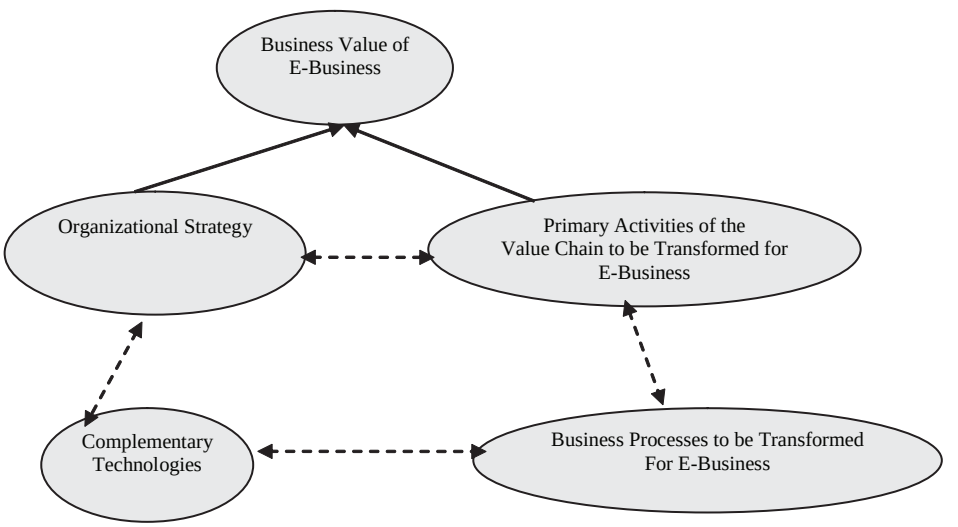
The main objective of the model is to make the business value of e-business as close to optimal as possible in terms of one of the performance measures, such as company profitability, competitive advantage, increase in market share, shareholder value or customer satisfaction. This can be done by exploring complementarities among the dependent variables of the model: the company strategy, the activities of the value chain, the corresponding business processes, and the technologies available to transform these activities and processes for the marketspace.

Furthermore, to succeed in e-business it is important to reengineer the parts of the value chain and the corresponding business processes relevant to the product in question and the company strategy. For example, the strategy or combination of strategies a company wants to pursue is relevant for the primary activities of the value chain, and the corresponding business processes that have to be implemented online. The strategy is also relevant to the classes of technologies that have to be chosen to enter the electronic market place. For example, a company can develop an e-business model to implement a cost leadership strategy, or to become the low cost producer in the industry. Once decided upon the strategy, it is important to explore complementarities between the strategy and the value chain activities in order to implement online all those activities that would support an optimal implementation of the strategy chosen.

The number of primary activities and corresponding business processes that should be transformed for the marketspace depends also on the company's type of product and strategy. It is important to take into consideration complementarities among the different activities of the value chain when designing an e-business model. The more activities of the value chain are simultaneously conducted online, the more likely it is that the business value of e-business will be optimized. The adoption of a holistic approach in redesigning the primary activities for the electronic marketplace should be a more successful strategy than reengineering only one or some at a time. This is due to potential complementarities between the different activities, which lead to a better performance in one if the others are also reengineered for online commerce.

Furthermore, each business process of each activity of the value chain could be re-engineered for e-business. This

Figure 1. Business value complementarity model of e-business



model argues that the exploration of complementarities among the different business processes and the simultaneous transformation of all the complementary processes of a particular activity for online commerce would lead to a higher business value than if only one or a casual numbers of processes were reorganized online (Scupola, 1999).

In the design phase, it is important to consider potential complementarities between the business processes that have to be redesigned for online commerce and the supporting technologies. The exploration of this complementarity should lead to an optimal system design that also offers possibilities for further expansion if other online business processes should be added in the future. For example, electronic search of the company's information will give more accurate and quicker results, the faster and more advanced the search engine is and the better built are the user interface and the repository systems.

Finally, the exploration of complementarities between the different technologies used to implement the e-business system could bring to a more robust and flexible computer system than a system built without the exploration of complementary relationships between the different component technologies. For example, end user interfaces and repositories are complementary technologies in the sense that the better designed the repository system, the simpler the user interface can be.

FUTURE TRENDS

The studies on IT productivity and business value conducted over the last decade have showed positive impacts of IT investments on firms' productivity both with respect to labor and other non-IT capital used by organizations (Barua et al., 2000a). However, Internet-based technologies, with their open standards and wide applicability, raise again the issue of profitability and business value of investing in such technologies. Furthermore, the fact that Internet is giving rise to a "new economy," raises a number of questions among which: how productive are the players in this new economy? Does e-business increase the profitability and business value of brick and mortars and hybrid click and mortars companies? For dot coms, do more investments in Internet commerce technologies necessarily lead to a better performance of the company? And especially, if all the companies have equal access to Internet-based technologies, what are the factors that differentiate their performance in e-business?

Recent literature investigating the business value and profitability of electronic commerce and e-business is focusing on the exploration of complementary relationships between Internet technologies and other factors in order to see positive returns from investments in these technologies (Barua et al., 2000a; Barua et al., 2000b; Barua et al., 2001; Scupola, 2003). For example, Barua et al. (2001) develops a framework of electronic commerce business value that identifies linkages between performance drivers such as Internet applications, processes and electronic business readiness of customers and suppliers and operational excellence and financial metrics. They argue that "firms engaged in electronic business transformation must make synergistic investments and commit resources not only in information technology, but also must align processes and customer and supplier readiness to maximize the benefits" (p. 1).

Similarly, an empirical investigation of the business value of e-commerce in small, medium, and large companies across Europe and USA (Barua et al., 2000b) identifies a set of key e-drivers such as system integration, customer orientation of IT, supplier's orientation of IT, and internal orientation of IT. The study concludes that high performance companies have invested more effort and resources in these e-business drivers than companies who have not benefited from e-business.

To conclude, these studies show that ignoring complementarities in research on business value measurement might lead to misleading results. On the other hand, from a managerial point of view, the non-exploration of complementary relationships between IT and related factors such as strategy, business processes, business models, incentives, etc. might lead to failure of investments in sophisticated e-business models and ventures. These considerations point to the need for more empirical as well as normative, prescriptive research on complementarity and business value of IT in general and e-business technologies in particular.

CONCLUSION

Many companies are very skeptical about investing into e-business technologies due to the lack of profitability, (or at least the difficulties to show positive return on IT investments) that until now has characterized the investments in IT, electronic commerce and e-business. Here a framework that can be used as a methodology to analyze organizational strategies and technology choices in reengineering for e-business has been presented. Companies should explore the potential complementarities existing between strategy, value chain activities, business processes, and supporting technologies when designing an e-business model. This should lead to investments in e-business systems that best support the company strategy, thus minimizing failures. This is a future challenge for corporations, industries, and researchers.

REFERENCES

Barua, A., & Mukhopadhyay, T. (2000a). Information technology and business performance: Past, present, and future.

In R. Zmud (Ed.), *Framing the domains of IT management, projecting the future through the past*.

Barua, A., Konana, P., Whinston, A. B., & Yin, F. (2001a). Driving e-business excellence. *Sloan Management Review, 43*(1), 36-44.

Barua, A., Konana, P., Whinston, A. B., & Yin, F. (2000b). *Making e-business pay: Eight key drivers for operational success.* IT Pro. IEEE Publisher, November-December. Retrieved from http://www.computer.org/portal/site/csdl/index.jsp

Barua, A., Konana, P., Whinston, A. B., & Yin, F. (2004). An empirical investigation of net-enabled business value. *MIS Quarterly, 28*(4), 585.

Barua A., Lee, S. C. H., & Whinston, A. B. (1996). The calculus of reengineering. *Information Systems Research, 7*(4), 409-428.

Barua, A., Pinnell, J., Shutter, J., Wilson, B., & Whinston, A. B. (1999). *The Internet economy indicators Part II*. Retrieved June 19, 2002, from http://www.internetindicators.com

Brynjolfsson, E., & Hitt L. M. (1996). Paradox lost? Firm-level evidence of the returns to information systems spending. *Management Science, 42*, 541-558.

Kalakota, R., & Whinston, A. B. (1996). *Frontiers of electronic commerce*. Addison-Wesley.

Kauffman, R. J., & Kriebel, C. H. (1988). Modeling and measuring the business value of information technologies. In P. A. Strassman, P. Berger, E. B. Swanson, C. H. Kriebel, & R. J. Kauffman (Eds.), *Measuring the business value of information technologies.* Washington, DC: ICIT Press.

Melville, N., Kræmer, K., & Gurbaxani, V. (2004). Review: Information technology and organizational performance: An integrative model of IT business value. *MIS Quarterly, 28*(2), 283.

Milgrom, P., & Roberts, J. (1990). The economics of modern manufacturing: Technology, strategy, and organization. *American Economic Review*, 511-528.

Pepper, J., & Ward, J. (2005). Unlocking sustained business value from IT investments. *California Management Review, 48*(1), 52.

Porter, M. (1982). *Competitive strategy*. The Free Press.

Porter, M. (1980). *Competitive advantage*. The Free Press.

Porter, M., & Miller, V. (1985). *How information gives you competitive advantage*. Harvard Business Review.

Ravi, K., & Robinson, M. (1999). *E-business: Roadmap for success.* Reading: Addison Wesley Longman.

Roach, S. S. (1989). America's white-collar productivity dilemma. *Manufacturing Engineering*, August, p. 104.

Scupola, A. (2003). Organization, strategy, and business value of electronic commerce: the importance of complementarities. In J. Mariga (Ed.), *Managing e-commerce and mobile computing technologies* (pp. 147-162). Hershey, PA: Idea Group Publishing.

Scupola, A. (2002). A business value complementarity framework of electronic commerce. In M. Khosrow-Pour (Ed.), *Issues and trends of information technology management in contemporary organizations*. Hershey, PA: Idea Group Publishing.

Scupola, A. (1999). The impact of electronic commerce on the publishing industry: Towards a business value complementarity framework of electronic publishing. *Journal of Information Science*, *25*(2).

Shin, N. (2006). The impact of information technology on the financial performance of diversified firms. *Decision Support Systems, 41*(4), 698.

Singh, M., & Waddell, D. (2004). *E-business innovation and change management*. Hershey, PA: Idea Group Publishing.

Tanriverdi, H. (2006). Information technology relatedness, knowledge management capability, and performance of multibusiness firms. *MIS Quarterly, 29*(2), 311.

Wigand, R. T. (1997). Electronic commerce, definition, theory, and context. *The Information Society*, *13*, 1-16.

Yang, S. M., Yang, M. H., & Wu, J. B. (2005). The impacts of establishing enterprise information portals on e-business performance. *Industrial Management + Data Systems*, *105*(¾), 349.

Zhu, K. (2004). The complementarity of information technology infrastructure and e-commerce capability: A resource-based assessment of their business value. *Journal of Management Information Systems*, *21*(1), 167.

KEY TERMS

Business Processes: The business processes corresponding to each activity of the value chain are the specific processes into which each primary activity of the value chain can be decomposed.

Business Value: It can be defined as the overall value that an investment brings to a corporation. Examples of performance measures of the business value of electronic commerce can be: (1) profitability, that is whether electronic

commerce contributes to an increase in the profitability of the corporation; (2) competitive advantage that could be measured as an increase in market share, shareholder value or customer satisfaction.

Complementarity: Several activities are mutually complementary if doing more of any one activity increases (or at least does not decrease) the marginal profitability of each other activity in the group. Complementarities among activities imply mutual relationships and dependence among various activities whose exploration can lead to higher profitability.

E-Business: Is more then e-commerce and is a new business design based on integration of technology, business processes and customer needs.

E-Commerce: Is the buying and selling of information, products and services via computer networks and especially the Internet.

Internet Economy: Is made up of a large collection of global networks, applications, electronic markets, producers, consumers, and intermediaries.

Re-Engineering: Is the redesign of a corporation's business processes (or part of them) to take place over the Internet. The main goal is reduced costs, lower product cycle times, faster customer response, and improved service quality.

Strategy: Strategy is a planning, rational process through which the company chooses a certain mode of development, among all the possible ones, and maintains that direction through a well defined period (design view). In the process view, strategy is a process that might change on the way, giving rise to an emergent strategy. The realized strategy might be different then the original intended strategy.

Value Chain: Represents the activities of a corporation such as procurement, production, marketing and sales, and customer support.

Category: Data Mining & Databases

Linguistic Indexing of Images with Database Mediation

Emmanuel Udoh
Indiana University – Purdue University, USA

INTRODUCTION

Computer vision or object recognition complements human or biological vision using techniques from machine learning, statistics, scene reconstruction, indexing and event analysis. Object recognition is an active research area that implements artificial vision in software and hardware. Some application examples are autonomous robots, surveillance, indexing databases of pictures and human computer interaction. This visual aid is beneficial to users, because humans remember information with greater accuracy when it is presented visually than when it originates in writing, speech or in kinesthetic form. Linguistic indexing adds another dimension to computer vision by automatically assigning words or textual descriptions to images. This augments content-based image retrieval (CBIR) that extracts or searches for digital images in large databases.

According to Li and Wang (2003), most of the existing CBIR projects are general-purpose image retrieval systems that search images visually similar to a query sketch. Current CBIR systems are incapable of assigning words automatically to images due to the inherent difficulty of recognizing numerous objects at once. This current situation is stimulating several research endeavors that seek to assign text to images, thereby improving image retrieval in large databases.

To enhance information processing using object recognition techniques, current research has focused on automatic linguistic indexing of digital images (ALIDI). ALIDI requires a combination of mathematical, statistical, computational, and graphical backgrounds. Many researchers have focused on various aspects of linguistic processing such as CBIR (Ghosal, Ircing, & Khudanpur, 2005; Iqbal & Aggarwal, 2002, Wang, 2001) machine learning techniques (Iqbal & Aggarwal, 2002), digital library (Witen & Bainbridge, 2003) and statistical modeling (Li, Gray, & Olsen, 20004, Li & Wang, 2003). A growing approach is the utilization of statistical models as demonstrated by Li and Wang (2003). It entails building databases of images to be used for supervised learning. A trained system is used to recognize and identify new images with statistical error margin. This statistical modeling approach uses a hidden Markov model to extract representative information about any category of images analyzed. However, in using computer to recognize images with textual description, some of the researchers employ solely text-based approaches. In this article, the focus is on the computational and graphical aspects of ALIDI in a system that uses Web-based access in order to enable wider usage (Ntoulas, Chao, & Cho, 2005). This system uses image composition (primary hue and saturation) in the linguistic indexing of digital images or pictures.

BACKGROUND

Current image indexing systems are text-based, relying on content-relevant text placed in proximity to images. There is need for Web-based automated linguistic indexing for digital images. This fact will likely accelerate the adoption of automated linguistic indexing for images in their native visual form, which basically assigns textual description automatically to images (Forsyth & Ponce, 2002; Li, Gray, & Olsen, 2000; Li & Wang, 2003). ALIDI is currently an active research area in data mining, and its application is growing in such fields as consumer photo managers, medical imaging databases and image search engines (Berman & Shapiro, 1997; Li & Wang, 2003; Tanev, Kouylekov, & Magnini, 2004; Zhang, Goldman, Yu, & Fritts, 2002).

In general, ALIDI systems aid computer object recognition and content-based image retrieval, despite inherent difficulties (Li & Wang, 2003). With statistical modeling and machine learning approach, especially supervised learning, much research progress has been achieved in this field, but it is obvious that no single method can be used to realize this endeavor. Due to varied uses, a complete linguistic indexing system is bound to implement different algorithms or methods for effectiveness. The focus of this project is on generic image composition, which can be applied for quality control in the food industry to determine intrinsic value of food products. The proposition involved parameter-based quality detection of consumable produce. Detecting such quality involved some analysis of primary hue and saturation of images. Color profiles would be constructed with acceptable levels of various quality-indicative color levels. Certain combinations of hue and saturation would yield unacceptable produce when compared with that which is acceptable. For instance, the apple fruit passes through different color stages in the ripening process. These stages can be captured and utilized in computer quality detection works. In this article,

the research is focused on assigning linguistic terms to any images based on composition, by using combinations of hue and saturation. It dwells on generic color identification and detection of images.

MAIN FOCUS

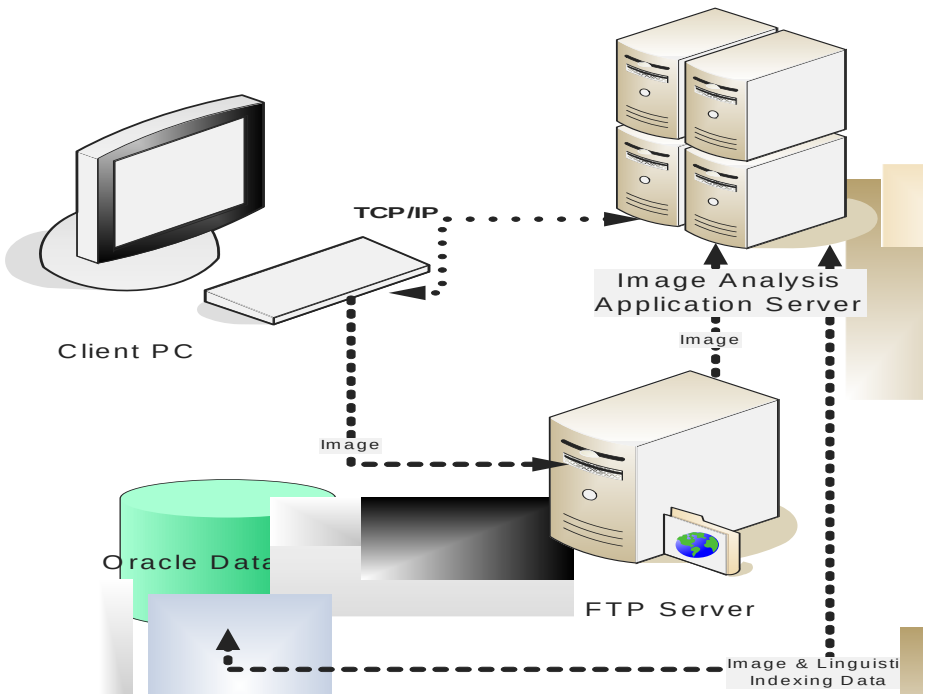
A current research theme on linguistic indexing is automating the process of assigning text to images using machine learning and statistical modeling techniques (Li & Wang, 2003). With some statistical considerations, this article focuses on the online assignment of color-based terms to images. The following describes the overall approach and Web access of the system developed by the author.

Overall Procedure

Linguistic indexing systems recognize images and textually describe them with human-readable terms like "beach" or "forest." In this article, the theoretical approach is to use image composition, and to describe the images using the color spectrum terms such as "blue" or "red." To index any image, the approach to decompose each image to pixels is adopted. After the decomposition, each pixel is assigned the values of triple composure using the red, green and blue (RGB) color model. The RGB value is then converted to the hue, saturation and brightness (HSB) color model before the color distribution in the image is determined. The quantification of the color value is carried out as opposed to the statistical similarity between the image and any concept. This approach has the potential to identify any digital image composition within a statistical error margin.

Elements of image composition require definition and human-encoded color names that are not necessarily native to computers, because computers process integers of bytes corresponding to varieties of hue, saturation, and brightness (HSB) or of the often overlapping values of red, green, and blue (RGB) from the additive color system. While RGB does reference colors directly in its indexing technique, HSB appeared to be most useful because it denotes definite color ranges with progressive hues. This informed our decision to convert RGB values to the HSB values with ease of manipulation. Due to the fact that RGB values do not translate exactly to HSB values, we statistically determined the errors associated with any conversion. It is worth mentioning that image composition offers two potential advantages over other architectural approaches: scalability and a simple programming model (Mello & Lins, 2002).

Figure 1. Implementation topology of a linguistic processing system



To enable a Web-based processing of images with associated linguistic terms, the topology shown in Figure 1 was implemented. This configuration could be considered to be three- or four-tier, depending on where the FTP server is situated. The system uses the concepts of portability, accessibility, and reusability. The system has three major components, the image analysis, the file upload and the database storage units. A combination of different technologies was used to achieve the objectives: PHP and Apache Tomcat application server (Harrison & McFarland, 2002) for Web-based programming, and the JPEG/Java technologies for image analysis. These two distinct technologies would depend on common database components and a computerized network for file transfer. Oracle 10g (Morrison, Morrison, & Conrad, 2006) served as the database for storage of the processed data, while an FTP server served as a temporary storage facility for the digital images that awaited processing. The overall topology with middleware, which handles the business rules and coordinates data access functions to the database server, provides added efficiency for a system that deals with large images and data set (Witten, Moffat, & Bell, 1999). Additionally, because of the seamless integration of the tiers of this application, any changes made to the application software or the database structure will not affect the users.

Web Access and Evaluation

To demonstrate this concept, a Web-enabled processing unit was implemented. Figure 2 shows the entry page, which leads to the page in Figure 3 that requests upload of the image to the FTP server en route to the image analysis server. The FTP server is designed to upload and then delete the image after analysis for space management. For this example, the image contains flowers, bouquet of pink roses (dimension: 640 x 480).

Assessing images based on color profiles generated pixel counts of each color identified in our spectrum. We converted these values to ratios and then to percentages for human-readable usability considerations. An error value would account for any rounding. An actual result set provided the following data from bouquet of pink roses (Figure 4). The results indicate a relatively large percentage of orange content, followed secondarily by red, with trace levels of yellow, green and blue. Such results will provide key insight into the composition of images at the pixel level for quality assurance detection routines.

We evaluated the system with about 100 different images in order to quantify its performance. The system was able to display the frequency of each color present in the image. About 95 images have error margins below 5%. Only two images can be considered to be outliers in terms of error values. Outliers exist for various reasons, and there are statistical methods to handle their effects (Li & Wang, 2003; Witten, Moffat, & Bell, 1999). Table 1 depicts the color identification result of some of the images. It lists the percentage of color in the images examined, while the last column holds the error values. This result does not depend on a categorized training database but on intrinsic color value. It is a process that can be enhanced for further studies in detection activities.

Figure 2. Web-based entry point to the linguistic system

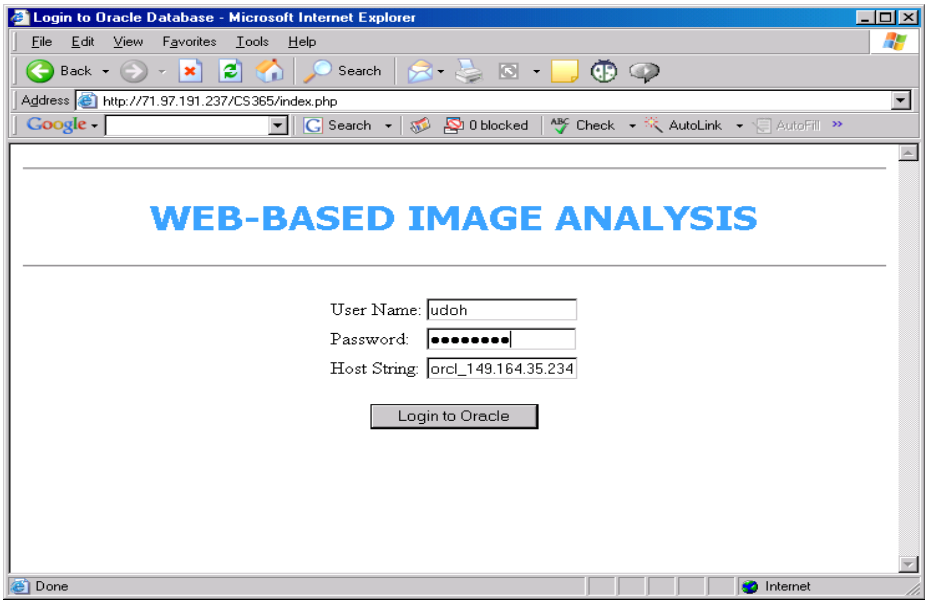


Figure 3. Upload of image to the processing servers

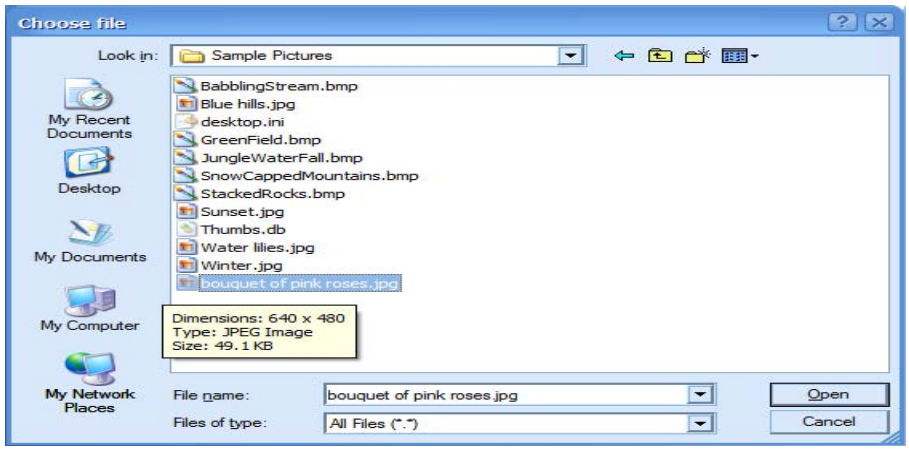


Figure 4. Display of processed image with descriptive linguistic terms (color ratio terms)

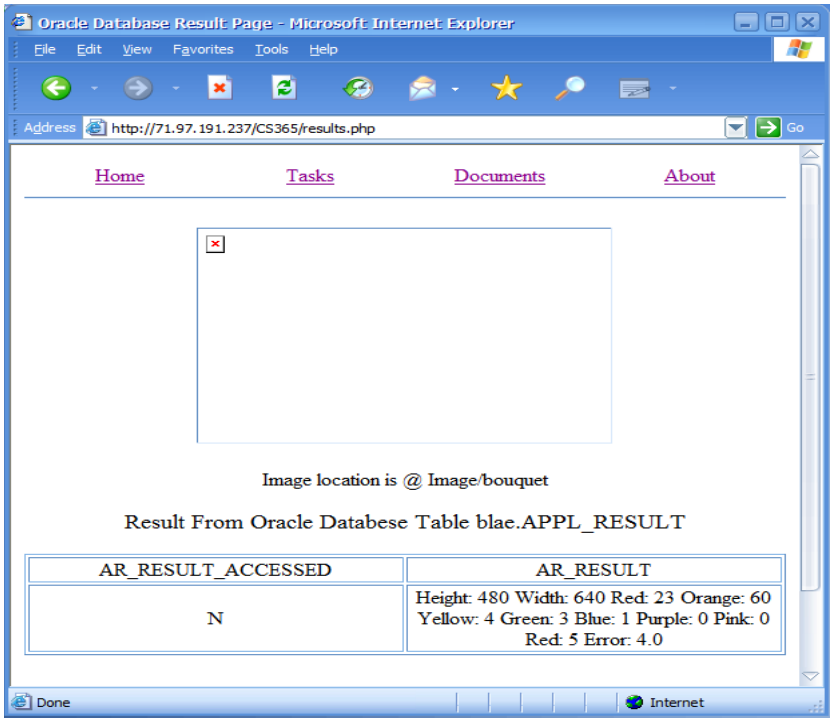


Table 1. Results of image identification experiments (color distribution in %)

Image	Red	Orange	Yellow	Green	Blue	Purple	Pink	Other	Error
Iris	2	65	5	4	20	1	1	0	2
Lawn	3	4	4	70	4	1	12	0	1
Flower	23	60	4	3	1	0	0	5	4
Forest	4	5	10	65	2	2	8	4	2

FUTURE TRENDS

The recognition of 2D or 3D objects by computers is important in various areas such as consumer photo managers, medical imaging databases, image search engines, surveillance, biomedicine, commerce, military, education and digital libraries. Computer vision complements or augments human vision activities. Linguistic indexing of images adds a positive component to object recognition that will continue to fuel research in this field. As a future trend, weights will be given to the assigned words to indicate the extent of the description appropriateness (Li & Wang, 2003). Furthermore, statistical modeling such as the Markov model, likelihood approximation and rule-based systems will impact the linguistic indexing of images and pictures. Smart technologies from artificial intelligence will also enhance progress in this field. Finally, a host of sophisticated applications will be developed to aid computer object recognition in different fields.

CONCLUSION

Widespread utilization of automated linguistic indexing for digital images will revolutionize access to visual information, especially in Internet search engines. In this article, we demonstrated a Web-based approach to assign linguistic terms to images using composition (image hue and saturation). To index any image, the image is decomposed to pixels. Each pixel is assigned the values of triple composure using the red, green and blue (RGB) color model. The RGB value is then converted to the hue, saturation and brightness (HSB) color model before the color distribution in the image is determined. To enable a Web-based processing of images with associated linguistic terms, a three- or four-tier topology is implemented on Tomcat and Oracle database servers. The evaluation of 100 different images for system performance shows that the system can be integrated into a larger linguistic processing system.

REFERENCES

Berman, A., & Shapiro, L. G. (1997). Efficient image retrieval with multiple distance measures. In *Proceedings of SPIE*, Newport Beach, CA, (Vol. 3022, pp. 12-21).

Forsyth, D. A., & Ponce, J. (2002). *Computer vision: A modern approach*. New York: Prentice Hall.

Ghosal, A., Ircing, P., & Khudanpur, S. (2005). Hidden Markov models for automatic annotation and content-based retrieval of images and video. In R.G. Baeza-Yates, R.G. Marchionini, Moffat, J. Tait, & N. Ziviani (Eds.), *28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 544-551).

Harrison, P., & McFarland, I. (2002). *Mastering Tomcat development*. Indianapolis, IN: John Wiley & Sons.

Iqbal, Q., & Aggarwal, J. K. (2002). Retrieval by classification of images containing large man-made objects using perceptual grouping. *Pattern Recognition Journal, 35*(7), 1463-1479.

Li, J., Gray, R. M., & Olshen, R. A. (2000). Multi-resolution image classification by hierarchical modeling, with two dimensional hidden Markov models. *IEEE Transactions on Information Theory*, *46*(5), 1826-1841.

Li, J., & Wang, J.Z. (2003). Automatic linguistic indexing of pictures by statistical modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(9), 1075-1088.

Mello, C. A. B., & Lins, R. D. (2002). Document reuse and semantics: Generation of images of historical documents by composition. In R. Furuta, J.I. Maletic, & E. Munson, (Eds.), *ACM Symposium of Document Engineering*, (pp. 127-133).

Morrison, J., Morrison, M., & Conrad, R. (2006). *Guide to Oracle 10g*. Boston: Course Technology.

Ntoulas A., Chao, G., & Cho, J. (2005). The infocious Web search engine: Improving Web searching through linguistic analysis. In J. Ellis & T. Hagino (Eds.), *14th International Conference on World Wide Web*, (pp. 840-849).

Tanev, H., Kouylekov, M., & Magnini, B. (2004). Combining linguistic processing and Web mining for question answering. In N. S. Sridharan, (Ed.), *11th Text Retrieval Conference*, Gaithersburg, MD, (pp. 1-10).

Wang, J. Z. (2001). *Integrated region-based image retrieval*. Dordrecht: Kluwer Academic.

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images*. San Francisco: Morgan Kaufman.

Witen, I. H., & Bainbridge, D. (2003). *How to build a digital library*. San Francisco: Morgan Kaufman.

Zhang, Q., Goldman, S.A., Yu, W., & Fritts, J.E. (2002). Content-based image retrieval using multiple-instance learning. In C. Sammut & A.G. Hoffmann (Eds.), *Proceedings of 19th International Conference on Machine Learning*, Sydney, Australia, (pp. 682-689).

KEY TERMS

Computer Vision or Object Recognition: A discipline concerned with the science and technology of artificial systems for extracting information from images and multidimensional data. Application areas include industrial robots, indexing databases of images and industrial inspection.

Concept: A generalization or abstraction of a particular set of instances or a particular category of images at the data level.

Content-Based Image Retrieval (CBIR): This approach retrieves or searches digital images from large databases using the content of the images themselves or syntactical image features without human intervention. To aid image retrieval, techniques from statistics, pattern recognition, signal processing, and computer vision are commonly deployed. Other terms used interchangeably for CBIR are query by image content (QBIC) and content-based visual information retrieval (CBVIR).

Image Composition: General makeup or the proportion of elements in an image, for example, color.

Linguistic Indexing: Assignment of textual description or words to images or pictures as a way of identifying the image.

Machine Learning: An area of artificial intelligence that allows computers to apply rules and algorithms in a learning process. It overlaps with data mining and statistics and has wide applications in areas such as object recognition, computer vision, robot locomotion and bioinformatics.

Pattern Recognition: Part of machine learning (with supervised learning underpinnings) that classifies or extracts patterns from raw data (measurements or observations) relying on the features of the data.

Category: IT Education

Linking Individual Learning Plans to ePortfolios

Susan Crichton
University of Calgary, Canada

INTRODUCTION

Throughout the 1990s, educators working in alternative schools explored the use of individual learning plans as support for at risk students and reluctant, returning adult learners (Crichton, 2005; Crichton & Kinsel, 2002). These early learning plans were strictly paper based. Each student had her/his own cardboard folder that contained goal personal statements, benchmarks, course process, and personal information (e.g., interests, preferred learning styles). Samples of completed work were included in the folders so students could see their improvement/progress. By 1998, there was interest in exploring the potential of technology to improve the paper portfolios, noting improvements in multimedia authoring and Internet access. It was found that electronic learning plans, complete with collaborative journals, showed promise (Kinsel, 2004). This chapter suggests that ePortfolios that draw on content from personal eJournals extend those early learning plans both in concept and impact on learning.

BACKGROUND

In many ways, electronic journals and portfolios are a natural extension of individual learning plans as they encourage authentic ways for individuals to demonstrate his/her growing understanding of a content area and a developing sense of self through demonstrations of learning. Sparks (1999) suggests that rich and meaningful learning opportunities should provide a bridge to the future by helping students to learn how to learn, "… so they can keep up with the rapidly changing world" (p. 20). This is consistent with the early work of Goffmann (1959, p. 20) who states, "We come into this world as individuals, achieve character, and become persons." The development of a positive sense of self appears key to learner success, and it must be recognized that it is something that individuals must create for themselves.

Initially, the rationale for the development of an individual learning plan was predicated on the understanding that

… the development of a complex, multi-faceted sense of self can increase student achievement and self-confidence. Individualized learning links the personal and social identities of students with the academic curriculum, mapping a pathway to activities appropriate to the needs and goals and the development of an increasingly complex sense of self (Crichton & Kinsel, 2002, p. 143).

The theoretical framework for the use of early paper-based learning plans was anchored in activity theory (Vygotsky, 1994), encouraging learners to think about what they want to do, how they learn best, what supports they need, and what their prior learning has afforded them. To engage this type of thinking, the natural evolution into electronic learning plans included an interactive, personal journal area where the learner and facilitator could communicate about experiences, successes, and failures, and generally document and reflect on the learning experience. This journaling area was designed to support the notion that learning must start at a personal level, and then gradually progress to the public and applied levels (Crichton & Kinsel, 2002).

In recent years, the literature (Barrett, 2003, 2005; Fox, Kidd, Painter, & Ritchie, 2006; Jafari & Kaufman, 2006) is rife with discussion about the value of portfolios (both paper and electronic) for educational purposes. Described by the National Learning Infrastructure Initiative as "a collection of authentic and diverse evidence, drawn from a larger archive representing what a person or organization has learned over time on which the person or organization has reflected, and designed for presentation to one or more audiences for a particular rhetorical purpose" (Barrett, 2005, p. 5), portfolios typically serve the purpose of assessment for learning, narrative of discovery, and tools for reflection. This, and the literature that follows, informed the initial design of the ePortfolio initiative shared in this chapter.

In general, the structure of ePortfolios varies. Throughout the literature, mention is made of a template as "a master or pattern from which similar things can be made" (Flanigan & Amirian, 2006, p. 111). Templates range from specific portfolio software to style pages in Dreamweaver that prevent ePortfolio development from being "… simply an exercise in Web design […suggesting the templates help to] … focus attention on developing the content of the ePortfolio" (Romance, Whitesell, Smith, & Louden, 2006, p. 534) without requiring advance experience/courses in HTML.

The literature suggests that portfolios can help faculty sustain and enhance the quality of pedagogy for preservice teachers by engaging them in conversation and reflection, and assessing those interactions in meaningful, authentic

ways. The ability to do this rests in the design of a portfolio experience that encourages risk taking, good questioning, and documenting the responses to those questions/experiences in manageable, sustainable, and meaningful ways (Black & Wiliam, 1998).

Commonly expressed in the literature, portfolios emphasize analysis and reflection by honoring the process not the product (Acosta & Liu, 2006). However, as Flannigan and Amirian (2006) note, "Documents, projects, and video that student felt represented their best works and abilities were collected as artifacts for the portfolios" (p. 105). This appears to place value on the quality of product over the process, which is consistent with the generally understood notion that portfolios are collections of quality work, and tends to come from the field of architecture and art, where portfolios are used as showcases of a body of work used by the architect or artist to gain work or additional study.

Implied in the literature is the potential to incorporate digital documentation as evidence for portfolio claims. Dahlberg, Moss, and Pence (1999) describe the importance of documentation for reflective practice, noting it "enables us to see how we ourselves understand and 'read' what is going on in practice; with this as a base, it is easier to see that our own descriptions as pedagogues are constructed descriptions. Hence, they become researchable and open for discussion and change" (p. 147). The documentation process also introduces students to authentic uses of technology to support information gathering and management as well as social networking, collaboration, and community building.

The research findings presented in this chapter challenge some of the notions presented in the literature while supporting others. This chapter will report on the integration of eJournals within an ePortfolio environment to encourage the development of a community of practice among student teachers, partner teachers, and university instructors. Building from research into the use of learning plans, this chapter extends the current thinking of ePortfolios, stressing the importance of an eJournal as a personal repository from which to draw meaningful portfolio items. Continuous developments in social software allow BLOGs to be an essential tool to build and sustain a community of practice.

The relevance of this topic rests in the link between journaling and portfolio development as well as the integration of technology for authentic purposes; a core competency for full participation in the 21st century.

LINK BETWEEN LEARNING PLANS AND EDOL

The link between learning plans and eJournals and portfolios rests in the potential they offer to help participants bridge their ability to learn how to learn (Sparks, 1999), and to critically reflect on that ongoing learning. An example of this type of reflective learning within a university setting is the eDOL (the electronic documentation of learning) project, a core component of the Teacher Preparation program at the University of Calgary. All pre-service teachers (n=800) maintain an eJournal and a series of ePortfolios.

During the two-year pilot study, five core points surfaced from the research:

1. The introduction of ePortfolios alone, without a journal option, placed a premium on the quality of products. Students were not prepared to place works-in-process or examples of challenging work in their portfolios. Because the students felt the word "portfolio" suggested best work, it appeared to limit ongoing discussions of challenges or frustrations that are essential for rich learning.
    - To remedy this situation, the project shifted from being about ePortfolios alone to focusing on eDOL (electronic documentation of learning). This shift has helped the three participant groups (student teachers along with their university instructors and partner teachers) to interact and share course content, journals, and experiences, thereby, honoring process as well as product.
2. Students needed an electronic journal that they could access anytime/anyplace. Access such as this allowed the students to write in their schools, at home, and basically anywhere there was Internet access. In addition, their partner teachers and university instructors would have the same access options. Because of this ease of access, the eJournal has become a rich archive from which the students could pull portfolio items at specific times in the program.
    - The login procedures and administrative controls keep the eJournals secure. This is very important as confidential/personal information is shared there, along with images from schools and classrooms.
    - Maintaining the integrity of the DRUPAL content on a university secure server has been paramount. This has required working closely with the university's information services department to verify user accounts, ensure each module added to the core DRUPAL site is secure, and backup the content in a secure and timely manner.
3. Digital documentation aimed at capturing issues of pedagogy and school environments was essential as a basis for reflection and social interaction as it provides evidence of specific events and contexts. The use of pedagogical documentation truly supports an emphasis on inquiry and reflection. Many forms of digital data can be recorded and placed in the students' journals. Examples include photographs of blackboard illus-

trations/directions, classroom seating arrangements, bulletin boards, student work, classroom activities.

- A concern with digital documentation is the need to recognize the sensitivities of filming in classrooms and obtaining the necessary permissions required to use the documentation in student journals and portfolios.
- Another concern that surfaced was the sharing of content from the eJournals in the university classrooms. Using an LCD projector, students and their instructors could share interesting text exchanges with the entire class or project an image as a prompt for discussion. Because of the Freedom of Information and Protect Act (FOIP) in Canada, permission must be secured in advance before classroom sharing of content can take place.

4. eJournals are developed using an open source, social software, DRUPAL (a blogging software). The architecture of DRUPAL is flexible enough to allow system administrators to customize it to meet emerging needs. DRUPAL can be password protected, securing the environment and allowing for issuing of accounts to registered users.
    - To customize and administer DRUPAL site, the Faculty of Education hired an ICT support person. This is an essential position, as security and integrity of the data are paramount.
5. Because DRUPAL works on the concept of groups, university instructors create a classroom group for all their students, and the students create individual journal groups into which they invite their partner teachers and university instructors.
    - Details concerning the various user checklists and user guides can be found on the main page of the eDOL site (http://education.ucalgary.ca/edol/front)
    - The students "own" their journal group, so they can maintain access to their own content/journal after the course has been completed.
    - These journals then become individual learning object repositories from which students can draw content for subsequent journal entries or for their portfolios.

eDOL would provides a way of understanding emerging student awareness of what inquiry-based teaching and learning looks like in actual practice. The originality of eDOL rests in the importance of establishing an eJournal to accompany the ePortfolio, challenging and adding to the existing ePortfolio literature.

At recent conferences (WestCAST 2007 and 2008), students shared their experiences. We focused on five core areas: 1) journals vs. portfolios (process and/or product); 2) value of digital documentation as evidence of student ownership of learning; 3) value of a unifying project to build program coherence; 4) importance of community within a program; and 5) the notion of "finishing with a place to start."

The first core area distinguishes eDOL from learning plans and contemporary ePortfolios, as it links the development of an individual repository of reflections, artifacts, evidence of learning, assignments, interactions with the development of a portfolio that requires the development of decision-making, critical reflection, and ICT skills.

The second area, digital documentation as evidence of student ownership of learning, has been fascinating. The process of recording, analyzing, editing, and preparing the documents has encouraged an authentic and appropriate integration of technology into the students' work. For many, this has been an introduction to ICT as it supports professional practice. The students have authored Web pages (the format for their ePortfolios) to showcase their work.

The third area, the notion of the portfolio as a unifying project that linked the pre-service program elements into a cohesive, personal learning experience, appears essential for student satisfaction. By drawing from their eJournals, students reported they were able to review their experiences, reflect on previous entries/activities, and see their personal growth. They found ePortfolios encouraged goal-oriented thinking and planning that invited them to consider audience, intention, and structure. This is consistent with the initial goals of the original work into learning plans. The process of selecting, editing, and developing portfolios resisted the traditional, fragmented university experience of independent courses that eventually lead to a degree, without necessarily promoting synthesis. In essence, eDOL supported the notion of synthesis through the process of reflection. The students stated that the unity of thought, activity, experience, and multiple perspectives helped them to begin to build their cohesive "beginning teacher" identities, allowing them to consider what was important and meaningful. Because the portfolios, and the archived data stored in their journals were electronic, students could easily modify and share it to a range of audiences (instructors, partner teachers, peers, and family), allowing others to view their work and regularly join in the ongoing conversation.

Fourth, eJournals provide a way to extend real-time social interactions by allowing users to reflect on face-to-face interactions. They also honor diverse learning styles, as not all students are comfortable or confident in verbal interactions. The flexible nature of BLOGS also allowed students to provide "evidence" of their claims by linking supporting images, text, or audio files.

In many ways, the fifth area is the most profound; the notion of finishing with a place to start. The students coined this phrase, and it is eloquent in its simplicity. As students leave the university and begin their careers, they take with them rich experiences, technology skills, a sustainable online

community of support and collegiality, and structure for a dynamic professional portfolio. They know they can maintain their connections with their mentors and peers through their eJournals, and they can continue to modify and add to their ePortfolios. By reviewing the ePortfolios and participating in the eJournals, faculty know that students recognize their increasing growth and professional maturity. The evidence provided in eDOL helps the students to make tangible the intangibles (Eisner, 1998). Journaling and maintaining a portfolio, like the learning plans, helps students to recognize both their individual and their professional selves and through that recognition, begin to reflect and grow.

FUTURE TRENDS

The use of ePortfolios for learning, employment, personal growth, and workplace certification is increasing. For example, "The Learning Innovations Forum d'Innovations d'Apprentissage (IfIA) in the Americas, together with the European Institute for E-Learning (ELfEL) in the European Union, actively advocate for:

- An ePortfolio for every citizen by 2010, and
- One ePortfolio for life" (Chang Barker, 2006, p. xxvi - xxvii).

The research from the eDOL example suggests that a portfolio, by itself, will probably serve workplace certification needs, but it may not address the needs for individual learning and personal growth. Those two areas require a venue for social interaction, collaboration, and reflection. Therefore, it is critical to use of both elements (journal and portfolio).

CONCLUSION

The relationship between learning plans and electronic documentation of learning experiences is clear. In many ways, ePortfolios are simply expanded, multimedia rich, learning plans. In the example of eDOL, the introduction of a reflective, interactive journal was a critical piece to enhance the value of the portfolio. Based on the findings from both studies (Crichton, Franks, O'Rourke, & Hodges, 2007; Kinsel, 2004), it appears that students must have a way to chart their progress, and make tangible and visible what it means to learn and grow professionally. Further, Marzano (2007) notes the importance of establishing and communicating learning goals, tracking student progress, and celebrating success. Findings from the use of learning plans and eDOL echo this as well. The potential afforded by electronic options, especially those using open source software, such as DRUPAL, simply improves access, ease of use, and the type of data and evidence that can be included.

REFERENCES

Acosta, T., & Liu, Y. (2006). ePortfolios: Beyond assessment. In A. Jafari & C. Kaufman (Eds.), *Handbook of research on ePortfolios* (pp. 15 - 23). Hershey, PA: IGA.

Barrett, H. (2003). *The eportfolio: A revolutionary tool for education and training*. Paper presented at the first International Conference on the ePortfolio, Poiters, France. Retrieved November 8, 2006, from http://electronicportfolios.org/portfolios/eifel.pdf

Barrett, H. (2005). *White paper: Researching electronic portfolios and learner engagement*. Retrieved June 23, 2006, from http://www.taskstream.com/reflect/whitepaper.pdf

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-75.

Chang Barker, K. (2006). Foreward. In A. Jafari & C. Kaufman (Eds.), *Handbook of research on ePortfolios* (pp. xxvi - xxviii). Hershey, PA: IGA.

Crichton, S. (2005). When just enrolling is not enough: A case for intentional learning plans. In S. Marshall, W. Taylor, & X. Yu (Eds.), *Encyclopedia of developing regional communities with information and communication technology*. London: Idea Group Reference.

Crichton, S., Franks, K., O'Rourke, E., & Hodges, Y. (2007). *Student teacher experiences with digital documentation: The introduction of eJournals and ePortfolios as support for reflection and inquiry.* WestCAST 2007, University of Manitoba, Manitoba, 2007 February 2007.

Crichton, S., & Kinsel, E. (2002). The importance of self and development of identity in learning. In *Proceedings of Lifelong Learning Conference* (pp. 143-151). Rockhampton, Queensland: Central Queensland University.

Dahlberg, G., Moss, P, & Pence, A. (1999). *Beyond quality in early childhood education and care.* Philadelphia, PA: Falmer Press, Taylor & Francis.

Dewey, J. (1904). The relation of theory to practice in education. In C. A. McMurray (Ed.), *The relation of theory to practice in the education of teachers* (Third yearbook of the National Society for the Scientific Study of Education, Part I, pp. 9 -30). Chicago: The University of Chicago.

Eisner, E. (1998). *The kind of schools we need*. Portsmouth, NH: Heinemann.

Flanigan, E., & Amirian, S. (2006). ePortfolio: Pathway from classroom to career. In A. Jafari & C. Kaufman (Eds.), *Handbook of research on ePortfolios* (pp. 15 - 23). Hershey, PA: IGA.

Fox, R., Kidd, J., Painter, D., & Ritchie, G. (2006). The growth of reflective practice: Teachers' portfolios as windows and mirrors. *The Teacher Educators Journal*, Fall, 2006. Retrieved from http://www.ateva.org

Goffmann, E. (1959). *The presentation of self in everyday life*. Toronto: Anchor.

Jafari, A., & Kaufman, C. (Eds.). (2006). *Handbook of research on ePortfolios*. Hershey, PA: IGA.

Kinsel, E. (2004). *Learning plans as support for personal success*. Unpublished Master's Thesis. Athabasca University. Spring 2004.

Marzano, R. (2007). *The art and science of teaching: A comprehensive framework for effective instruction.* Alexandria, VA: ASCD.

Romance, N., Whitesell, M., Smith, C., & Louden, A. (2006). Career ePortfolios in the IT Associates Program at DePauw University. In A. Jafari & C. Kaufman (Eds.), *Handbook of research on ePortfolios* (pp. 15 - 23). Hershey, PA: IGA.

Sparks, B. (1999). Critical issues and dilemmas for adult literacy programs under welfare reform. In L. G. Martin & J. C. Fisher (Eds.), *The welfare-to-work challenge for adult literacy educators* (pp. 15-25). San Francisco: Jossey Bass Publishers.

Vygotsky, L. (1994). *Thought and language*. Cambridge, MA: The MIT Press.

KEY TERMS

Authentic Learning: An approach that encourages students to explore, discover, discuss, and meaningfully construct concepts and relationships in contexts that involve real-world problems and projects that are relevant and interesting to the learner. Sometimes referred to as problem-based learning.

Demonstrations of Learning: Opportunities for students to **show** what they know rather than simply being evaluated using standard exams.

Digital Divide: Digital divide refers to the gap between those with regular, effective access to digital and information technology, and those without it. It encompasses both physical access to technology, hardware and, more broadly, the resources and skills needed to effectively participate as a digital citizen. In others words, it's the unequal access to some sectors of the community to information and communications technology, and the unequal acquisition of related skills (Wikipedia, 2007)

ePortfolios: An electronic portfolio, also known as an ePortfolio or digital portfolio, is a collection of electronic evidence assembled and managed by a user, usually on the Web. Such electronic evidence may include inputted text, electronic files, such as Microsoft Word and Adobe PDF files, images, multimedia, blog entries, and hyperlinks. EPortfolios are both demonstrations of the user's abilities and platforms for self-expression, and, if they are online, they can be maintained dynamically over time. Some ePortfolio applications permit varying degrees of audience access, so the same portfolio might be used for multiple purposes (Wikipedia, 2007).

eJournals: Electronic journals that allow multiple users to post comments within a secure, blogging environment. Other social software can be used, but in this chapter, DRUPAL, an open source blogging software, (http://drupal.org/) is used.

Learning Plans: A learning plan is in two parts. It is designed to make, much more explicit, the expectations of all parties (particularly student and Research Institute) in an attempt to avoid misunderstandings. In addition, students are required to keep much more detailed records of their activities as part of their research degree programme, which will help the RI better to monitor progress and become aware of any potential problems (http://www.keele.ac.uk/gradschool/learning.htm).

Open Source: Open source is a set of principles and practices that promote access to the design and production of goods and knowledge. The term is most commonly applied to the source code of software that is available to the general public with relaxed or nonexistent intellectual property restrictions. This allows users to create software content through incremental individual effort or through collaboration (Wikipedia, 2007).

Pre-Service Teaching: Typically called teacher preparation, it is a program within a college or university focused on the preparation for teachers, primary for the K-12 school environment.

Reflection: The initial act of reviewing and thoughtfully considering previous actions, events, and contexts as a form of constructive inquiry and personal growth and development.

Linking Information Technology, Knowledge Management, and Strategic Experimentation

V. K. Narayanan
Drexel University, USA

INTRODUCTION

Historically, the focus of IT infrastructure had been to capture the knowledge of experts in a centralized repository (Davenport & Prusak, 1998; Grover & Davenport, 2001; Nolan, 2001). The centralized databases contained knowledge that was explicit and historical (e.g., competitor pricing, market share), and the IT infrastructure served to facilitate functional decision making or to automate routine tasks (as in reengineering). The users of technology approached the repository to obtain data in a narrowly defined domain (Broadbent, Weill, & St. Clair, 1999). Consequently, IT originally played a significant, yet ultimately limited role in the strategy creation process. Management information systems (MISs) arguably generated information that was less applicable to strategy creation, as noted in early writings on the linkage between MIS and strategic planning (e.g., Lientz & Chen, 1981; Shank, Boynton, & Zmud, 1985; Holmes, 1985).

The active management of knowledge was similarly underdeveloped. Despite the fact that strategic decision makers had always emphasized the role of tacit knowledge, the actual importance of knowledge was not *explicitly* recognized. Formalized knowledge management (Davenport & Prusak, 1998; Dalkir, 2005), with its associated terminology and tools, is a recent development and as such did not inform the strategic planning process.

However, the shifts that have taken place in IT infrastructures over the last decade and the recent developments in knowledge management (KM) have brought them closer to the creators of strategy. Indeed, both IT and knowledge management are increasingly enablers in the contemporary strategic management practice:

1. IT infrastructure is transitioning in its focus from the functional work unit to a process orientation. Whereas computer systems were once the focal point, the new infrastructure is network centric, with an emphasis on business knowledge (Nolan, 2001). For example, traditional search engines utilized rule-based reasoning to identify elements matching specific search criteria; the "state-of-the-art" knowledge management systems employ case-based search techniques to identify all relevant knowledge components meeting the user's request (Grover & Davenport, 2001).
2. IT now takes into account contexts that include cross-functional experts, knowledgeable on a wide variety of potentially relevant issues. Additionally, there is greater emphasis on the integration of infrastructure with structure, culture (Gold, Malhotra, & Segars, 2001), and organizational roles (Awad & Ghaziri, 2004). In many ways, the newer IT infrastructures have enabled the garnering of explicit knowledge throughout the organization to speed up strategy creation.

The objective of this article is to outline how the developments in IT and KM are facilitating the evolution of strategic management to strategic experimentation to create quantum improvements in strategy creation and unprecedented developmental opportunities for the field if IT.

BACKGROUND

For the purposes of this article, *information technology* (IT) is defined as the physical equipment (hardware), software, and telecommunications technology, including data, image, and voice networks, employed to support business processes (Whitten & Bentley, 1998). The overarching plan for IT deployment within an organization is called the IT architecture. Technology infrastructure refers to the architecture—including the physical facilities, the services, and the management—that support all computing resources in an organization (Turban, McLean, & Wetherbe, 1996).

As used in this article, data are objective, explicit pieces or units, information is data with meaning attached, and knowledge is information with an implied element of action. According to Davenport and Prusak (1998):

"Knowledge is the fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers. In organizations, it often becomes embedded not only in documents or repositories but also in organizational routines, processes, practices, and norms." (p. 5)

Knowledge management is "a set of business practices and technologies used to assist an organization to

obtain maximum advantage from one of its most important assets—knowledge" (Duffy, 2000, p. 62). In other words, it is actively capturing, sharing, and making use of what is known, both tacitly, informally and explicitly, within the organization. IT often facilitates knowledge management initiatives by integrating repositories (e.g., databases), and indexing applications (e.g., search engines) and user interfaces. Awad and Ghaziri (2004) underscore the fact that KM also incorporates traditional management functions: building trust among individuals, allocating resources to KM, and monitoring progress.

The concept of "strategy" explicated in *strategic management* is one of marketplace strategy, that is, winning in the marketplace against competitors, entrenched or incipient. The underlying premise is that "to enjoy continued strategy success, a firm must commit itself to outwitting its rivals" (Fahey & Randall, 2001, p. 30). A large body of literature on strategic management has persuasively argued that effective strategy creation and execution are central to a firm's performance (e.g., Covin, Slevin, & Schultz, 1994).

Strategy creation involves both goal formulation—defined in terms of external stakeholders rather than operational milestones—and crafting of the strategic means by which to accomplish these goals (Hofer & Schendel, 1978). The means typically include business scope, competitive posture, strategic intent, and the organizational mechanisms for implementation. In practice, the process of strategy creation has often taken the form of strategic planning. Comprehensive strategic planning (Gluck, Kaufman, & Walleck 1978) has historically been practiced in large corporations: A celebrated example is the use of scenarios by Royal-Dutch Shell. It usually consisted of several sequential stages of decision making involving diagnosis, alternative development, evaluation and choice, and implementation. In each step, the strategic planners emphasized deliberate juxtaposition of "objective data" and careful analysis, with top management judgment, thus highlighting the role of tacit knowledge.

Strategic planning has evolved over the years. Writing in the 1970s, Gluck et al. (1978) identified four phases of evolution: budgeting, long-range planning, strategic planning, and strategic management. Each phase of evolution incorporated the lessons from the earlier phases, but also took into account the emerging realities faced by corporations. Gluck et al. (1978) noted that during the 1980s, the "strategic management" phase would represent the cutting edge of practice in the world.

TOWARD STRATEGIC EXPERIMENTATION

The 1990s witnessed a revolution in organizational environments often characterized as "hypercompetition." These environments have created three major imperatives for organizations: time compression, globalization, and technology integration (Narayanan, 2001). In addition, the increased environmental dynamism also contributes to an increase in the degree of uncertainty confronted by strategic managers, calling into question traditional planning practices. Consequently, a new type of strategy creation process is evolving which is termed "strategic experimentation." With this evolution, the relationship between strategy creation, knowledge management, and IT is undergoing a profound shift.

All the four phases of strategic planning documented by Gluck et al. (1978) incorporated a sequential approach to strategy creation and execution, leading to the identification of the one winning strategy that has the highest probability of success. Consequently, firms found it logical to commit the maximum available resources to the implementation of one winning strategy. The goal was to obtain a sustainable competitive advantage vis-à-vis the firm's rivals, and to reduce uncertainty ex ante using analytical forecasting techniques as well as market research. This approach to planning seems to have been effective during the 1980s, when the environment was moderately dynamic.

In hypercompetitive environments, market participants frequently confront great uncertainty over technological possibilities, consumer preferences, and viable business models. This high level of ambiguity often results in a situation where: (a) traditional methods of ex ante uncertainty reduction (e.g., market research) fail, and (b) the costs and risks of the traditional "big bet" strategic management approach outweigh its advantages in terms of focus, decisiveness, and concentrated resource commitment. It is in this situation that the emerging strategic experimentation approach holds significant promise.

Strategic experimentation (McGrath & MacMillan, 2000) draws on *real-options reasoning* (e.g., McGrath & Nerkar, 2003), discussions of *exploration vs. exploitation* (March, 1991), as well as *trial-and-error learning* (e.g., Van de Ven & Polley, 1992):

1. Companies engaging in strategic experimentation continually start, select, pursue, and drop strategic initiatives before launching aggressively those initiatives whose value is finally revealed (McGrath & MacMillan, 2000, p. 340).
2. Strategic initiatives thus serve as low-cost probes (Brown & Eisenhardt, 1998) that enable the discovery of product technology and market preferences. They also serve as a stepping stone *option* for future competitive activity in that particular product-market domain.
3. The role of the strategic manager is to administer a *portfolio of strategic initiatives* that represents an appropriate mix of high and low uncertainty projects, and to maximize the learning from these real options (McGrath & MacMillan, 2000).

Strategic experimentation represents a fundamentally different view of the practice of strategic planning and the path to competitive advantage. Movement is emphasized over position in this approach. Thus, competitive advantage is viewed as temporary at best, and hence innovation and learning are considered crucial to success. Strategic experimentation is especially appropriate for high-velocity environments such as emerging product markets with high uncertainty surrounding both technology and customer preferences (e.g., the early Personal Digital Assistant (PDA), Internet appliance, and satellite-based telephony markets). Here, low-cost probes can be very effective in gaining knowledge and reducing uncertainty while minimizing exposure to the results of faulty assumptions.

THE ROLE OF IT AND KNOWLEDGE MANAGEMENT IN THE ERA OF STRATEGIC EXPERIMENTATION

Since "strategic experimentation" represents the cutting edge of ideas in strategic management, we should expect significant advances in tool development and utilization in the next few years that will enable movement of the idea towards normal organizational practice.

Strategic experimentation necessitates several major functions to be performed by an organization. KM is critical in strategic experimentation; therefore, it is not surprising that many of the tools currently moving into practice have emerged from the KM. The four major strategic experimentation functions and the associated KM tools are:

- *Rapid Decision Making:* The ability to *quickly* garner tacit knowledge in all phases of decision making is a central requirement in strategic experimentation. Current KM tools to support this include visualization and prototyping, group decision facilitation, and knowledge representation. Each method attempts to reduce the time needed for a group to progress from problem identification to solution implementation. These tools help to coordinate the use of data, systems, tools, and techniques to interpret relevant information in order to take action (Ruggles, 1997).
- *Integration of Learning from Experiments:* Organizational learning, another core concept in strategic experimentation, requires that appropriate learning be distilled from each experiment. This orientation combines decision making and learning: initiatives judged to be failures are not merely weeded out, nor are successes simply alternatives to be financially backed. In contrast, failures become occasions for discovery of root causes; successes often generate potential best practices. Current KM tools in use for this purpose include learning histories (Roth & Kleiner, 1998), group brainstorming, and shared communication platforms (Dalkir, 2005).
- *Diffusion of Learning:* Finally, organizational learning must be diffused throughout the organization. Since formal organizational channels may stifle transmission of tacit knowledge (Williams, 2006), diffusion may require interactions among "communities of practice" (Wenger, McDermott, & Snyder, 2002; Narayanan, Douglas, Schirlin, Wess, & Geising, 2004). An organizational architecture, incorporating relevant tools and IT infrastructure, must be designed to support theses interactions. KM tools, such as knowledge maps identifying the experts in specific areas and repositories of case histories, are evolving to include dynamic updating of repositories and focused search tools to reduce information overload.
- *Managing a Portfolio of Strategic experiments:* Unlike in previous eras, strategic experimentation requires maintenance and management of a portfolio of initiatives (Narayanan, Buche, & Kemmerer, 2001). This has three major implications. First, the knowledge base for decisions must be broader and richer, simply due to the increase in the number of initiatives. Second, the knowledge base becomes much more complex, since the initiatives themselves differ in terms of the mix of tacit and explicit knowledge. Thus, newer initiatives are likely to be more dependent on tacit knowledge, whereas mature ones can be augmented by explicit knowledge. Finally, the sheer number of people involved in the process will be larger, given specialized pockets of tacit knowledge that would have grown up around specific strategic initiatives. DSSs and other rich data applications, including cognitive mapping, can be used to capture the knowledge and feedback.

IT can accelerate the development of strategic experimentation by designing infrastructures that accommodate the new KM demands (Smith & McKeen, 2003; King, Marks, & McCoy, 2002) imposed by this new mode of planning. Consider how each of the following functions can be enhanced by IT infrastructure development:

1. Future developments can significantly reduce the time expended in solution development through real-time displays and expand opportunities for geographically dispersed collaboration. Also, advanced multimedia and communication capabilities would increase the benefits of GSS and DSS tools.
2. Learning from experiments can be enriched by qualitative database construction, multimedia enhancements to communication applications, and open platforms to permit the sharing of knowledge over various communication channels, including wireless media.

3. Today, diffusion is hampered by information overload that has intensified competition for the user's attention (Hansen & Haas, 2001). To solve the problem, search tools should include separate parameters for content, rationale, and purpose of the query, in order to isolate salient responses. Additionally, knowledge repositories must be maintained to ensure the contents are accurate and of high quality. Also, maintenance, currently provided by intermediaries (Markus, 2001), might be performed by faster automated systems.
4. Expert systems or neural networks may be developed to manage and track portfolios, promoting reuse of the knowledge captured.

The significant implication for IT infrastructure from our discussion is the need for *technology integration* (Narayanan, 2001) with both hard and soft technologies. IT infrastructure should exploit the potential for integration with other *hard* technologies such as telecommunications to enhance the organizational capacity for speed and carrying capacity for *tacit* knowledge. Similarly, IT should seek to interface with *decision sciences* to embed AI-based processing tools, and with *cognitive theorists* to capture the tacit knowledge pervasive in organizations.

FUTURE TRENDS

Over the last decade, there has been a greater appreciation of both the importance of knowledge management and its links to IT. Indeed, KM as an organizational function and academic discipline is fast arriving at maturity; the linkage between KM and strategic management has been slow in building. We expect this to change in the coming decade. Platforms for strategic experimentation, built on overlay of KM and IT, are currently in the works among pioneering firms. We expect that the ideas around experimentation will begin to take off, with methods and approaches to identify, codify, and disseminate among organizational members being refined and utilized.

CONCLUSION

We have argued that the technological changes of the 1990s have ushered in the need for strategic experimentation as the metaphor for planning practice. Strategic experimentation involves: (a) maintaining a portfolio of strategic thrusts, (b) rapid decision making so that successful experiments are backed and failures are weeded out quickly, (c) learning from both successes and failures, and (d) diffusion of both explicit and tacit knowledge throughout the relevant segments of an organization. This phase requires fundamental shifts in our view of knowledge management: its significance, use, and tools. Finally, we have argued that the shift to strategic experimentation requires fundamental shifts in the development of IT infrastructure. Instead of developing in relative isolation to other disciplines, IT should focus on technology integration, by working in close collaboration with the telecommunication technologies, artificial intelligence community, and managerial cognition scholars.

REFERENCES

Awad, E.M., & Ghaziri, H.M (2004). *Knowledge management.* Englewood Cliffs, NJ: Prentice Hall.

Broadbent, M., Weill, P., & St. Clair, D. (1999). The implications of information technology infrastructure for business process redesign. *MIS Quarterly, 23,* 159-182.

Brown, S.L., & Eisenhardt, K.M. (1998). *Competing on the edge: Strategy as structured chaos.* Boston: Harvard Business School Press.

Covin, J.G., Slevin, D.P., & Schultz, R.L. (1994). Implementing strategic missions: Effective strategic, structural and tactical choices. *Journal of Management Studies, 31,* 481-505.

Dalkir, K. (2005). *Knowledge management in theory and practice.* New York: Elsevier.

Davenport, T., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Boston: Harvard Business School Press.

Duffy, J. (2000). The KM technology infrastructure. *Information Management Journal, 34,* 62-66.

Fahey, L., & Randall, R.M. (2001). *The portable MBA in strategy* (2nd ed.). New York: John Wiley & Sons.

Gluck, F.W., Kaufman, S.P., & Walleck, S. (1978, October). *The evolution of strategic management*. Staff Paper, McKinsey.

Gold, A.H., Malhotra, A., & Segars. A.H. (2001). Knowledge management: An organizational capabilities perspective. *Journal of Management Information Systems, 18,* 185-214.

Grover, V., & Davenport, T. (2001). General perspectives on knowledge management: Fostering a research agenda. *Journal of Management Information Systems, 18,* 5-21.

Hofer, C.W., & Schendel, D. (1978). *Strategy formulation: Analytical concepts.* St. Paul, MN: West.

Holmes, F.W. (1985). The information infrastructure and how to win with it. *Information Management Review, 1*(2), 9-19.

King, Marks, and McCoy. (2002). The most important issues in knowledge management. *Communications of the ACM, 35*(9), 93-97.

Lientz, B.P., & Chen, M. (1981). Assessing the impact of new technology in information systems. *Long Range Planning, 14*(6), 44-50.

March, J.G. (1991). Exploration and exploitation in organizational learning. *Organization Science, 2,* 71-87.

Markus, M.L. (2001). Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems, 18,* 57-93.

McGrath, R.G. (1998). Discovering strategy: Competitive advantage from idiosyncratic experimentation. In G. Hamel, C.K. Prahalad, H. Thomas, & D. O'Neal (Eds.), *Strategic flexibility: Managing in a turbulent environment* (pp. 351-370). Chichester: John Wiley & Sons.

McGrath, R.G., & Nerkar, A. (2003) Real options reasoning and a new look at the R&D investment strategies of pharmaceutical firms. *Strategic Management Journal, 25*(1), 1-21.

McGrath, R.G., & MacMillan, I. (2000). *The entrepreneurial mindset.* Boston: Harvard Business School Press.

Narayanan, V.K. (2001). *Managing technology and innovation for competitive advantage.* Prentice Hall College Division.

Narayanan, V.K., Buche, M., & Kemmerer, B. (2002). From strategic management to strategic experimentation: The convergence of IT, knowledge management, and strategy. In L. Joia (Ed.), *IT-based management: Challenges and solutions.* Hershey, PA: Idea Group.

Narayanan, V.K., Douglas, F., Schirlin, D., Wess, G., & Geising, D. (2004). Virtual communities as an organizational mechanism for embedding knowledge in drug discovery. *Journal of Business Chemistry, 1*(2), 37-47.

Nolan. (2001). *Information technology management from 1960-2000.* HBS Note # 9-301-147.

Roth, G., & Kleiner, A. (1998). Developing organizational memory through learning histories. *Organizational Dynamics, 27,* 43-60.

Shank, M.E., Boynton, A.C., & Zmud, R.W. (1985). Critical success factor analysis as a methodology for MIS planning. *MIS Quarterly, 9,* 121-129.

Smith, H., & McKeen, J. (2003). Developments in practice IX: The evolution of the KM function. *Communications of the AIS, 12,* 69-79.

Turban, E., McLean, E., & Wetherbe, J. (1996). *Information technology for management.* New York: John Wiley & Sons.

Van de Ven, A.H., & Polley, D. (1992). Learning while innovating. *Organization Science, 3,* 92-116.

Wenger, E., McDermott, R., & Snyder, W. (2002). *Cultivating communities of practice.* Boston: Harvard Business School Press.

Whitten, J.L., & Bentley, L.D. (1998). *Systems analysis and design methods* (4th ed.). Boston: Irwin McGraw-Hill.

Williams, R. (2006). Narratives of knowledge and intelligence…beyond the tacit and explicit. *Journal of Knowledge Management, 10*(4), 81-99.

KEY TERMS

Exploration and Exploitation: Exploration refers to process of discovery of knowledge, whereas exploitation refers to utilizing the knowledge. Similar to basic and applied research.

Information Technology (IT): The physical equipment (hardware), software, and telecommunications technology, including data, image, and voice networks, employed to support business processes.

Knowledge Management (KM): A set of business practices and technologies used to assist an organization in obtaining maximum advantage of its knowledge.

Options: A financial option owes the holder the right but not the obligation to trade in securities at prices fixed earlier. Options in the sense used here confer a firm the rights *and obligations* to choose a strategic alternative.

Strategic Experimentation: A form of strategic management in which firms continually start, select, pursue, and drop strategic initiatives before launching aggressively those initiatives whose value is finally revealed.

Strategic Management: The process of strategy creation and implementation. The concept of "strategy" as used here is one of marketplace strategy, that is, winning in the marketplace against competitors, entrenched or incipient. Strategy creation involves both goal formulation—defined in terms of external stakeholders rather than operational milestones—and crafting of the strategic means by which to accomplish these goals. Implementation refers to the means of execution of the created strategy.

Technology Integration: Activities involved in combining technologies. It could be combining process technolo-

gies as in incorporating biological process in mechanical equipment. Alternately it could be in product development, for example the merging of personal computers and satellite communications.

Trial-and-Error Learning: A process of learning through experimentation. Here strategic initiatives judged to be failures are not merely weeded out, nor are successes simply alternatives to be financially backed. In contrast, failures become occasions for discovery of root causes; successes often generate potential best practices.

Lip Extraction for Lipreading and Speaker Authentication

Shilin Wang
Shanghai Jiaotong University, China

Alan Wee-Chung Liew
Griffith University, Australia

INTRODUCTION

In recent years, there is a growing interest in using visual information for automatic lipreading (Kaynak, Zhi, Cheok, Sengupta, Jian, & Chung, 2004) and visual speaker authentication (Mok, Lau, Leung, Wang, & Yan, 2004). It has been shown that visual cues, such as lip shape and lip movement, would greatly improve the performance of these systems. Various techniques have been proposed in the past decades to extract speech/speaker relevant information from lip image sequences. One approach is to extract the lip contour from lip image sequences. This generally involves lip region segmentation and lip contour modeling (Liew, Leung, & Lau, 2002; Wang, Lau, Leung, & ALiew, 2004), and the performance of the **visual speech recognition** and **visual speaker authentication** systems depends much on the accuracy and efficiency of these two procedures.

Lip region segmentation aims to label the pixels in the lip image into lip and non-lip. The accuracy and robustness of the lip segmentation process is of vital importance for subsequent lip extraction. However, large variations caused by different speakers, lighting condition, or make-ups make the task difficult. The low color contrast between lip and facial skin, and the presence of facial hair, further complicate the problem. Given a correctly segmented lip region, the lip extraction process then involves fitting a lip model to the lip region. A good lip model should be compact, that is, with a small number of parameters, and should adequately represent most valid lip shapes while rejecting most invalid shapes. As most lip extraction techniques involve iterative model fitting, the efficiency of the optimization process is another important issue.

BACKGROUND

Accurate and robust lip region segmentation is of key importance for subsequent lip extraction. Techniques developed for lip segmentation are generally based on color space analysis, edge detection, Markov random field, or fuzzy clustering.

The color space analysis approach (Eveno, Caplier, & Coulon, 2001) identifies the lip pixels solely by their color information. However, color space-based methods are sensitive to poor color contrast and noise, and would give large segmentation error if the color distributions of lip and background regions overlap. The edge detection approach (Caplier, 2001) relies on the luminance or color edge information to detect the lip boundary. It works well when the speakers use lipstick or reflective markers. However, it would have difficulty dealing with unadorned lips. Markov random field (MRF) technique has also been used in lip region segmentation (Lievin & Luthon, 1999). MRF exploits local neighborhood information to enhance the robustness of the segmentation. However, MRF-based segmentation usually produces erroneous patches outside and inside the mouth region due to the presence of pixels with the wrong color distribution class.

Fuzzy clustering is another powerful tool for image segmentation. **Fuzzy clustering** attempts to assign a probability value to each pixel in order to minimize the fuzzy entropy. Since it is an unsupervised learning method, fuzzy clustering is capable of handling lip and skin color variation caused by make-up. Recently, we have proposed several novel fuzzy-clustering-based segmentation techniques that take the local (Liew, Leung, & Lau, 2000, 2003) and **global spatial information** (Leung, Wang, & Lau 2004; Wang, Lau, Liew, & Leung, 2007) into account to improve the segmentation performance. In our approaches, **spatial information** is seamlessly incorporated into the cost function and the optimization process.

Many techniques have been proposed for **lip modeling and extraction,** and they differ from each other in the following aspects:

- **The lip model used:** Active contour models (Snakes) (Eveno, Caplier, & Coulon, 2003; Lievin, Delmas, Coulon, Luthon, & Fristot, 1999), deformable templates (Hennecke, Prasad, & Stork 1994; Liew et al., 2002), active shape models (ASM) (Cootes, Hill, Taylor, & Graham, 1994; Luettin, Thacker, & Beet, 1996), and

active appearance models (AAM) (Cootes, Edwards, & Taylor, 2001; Matthews, Cootes, Bangham, Cox, & Harvey 2002) are some of the widely used lip models.

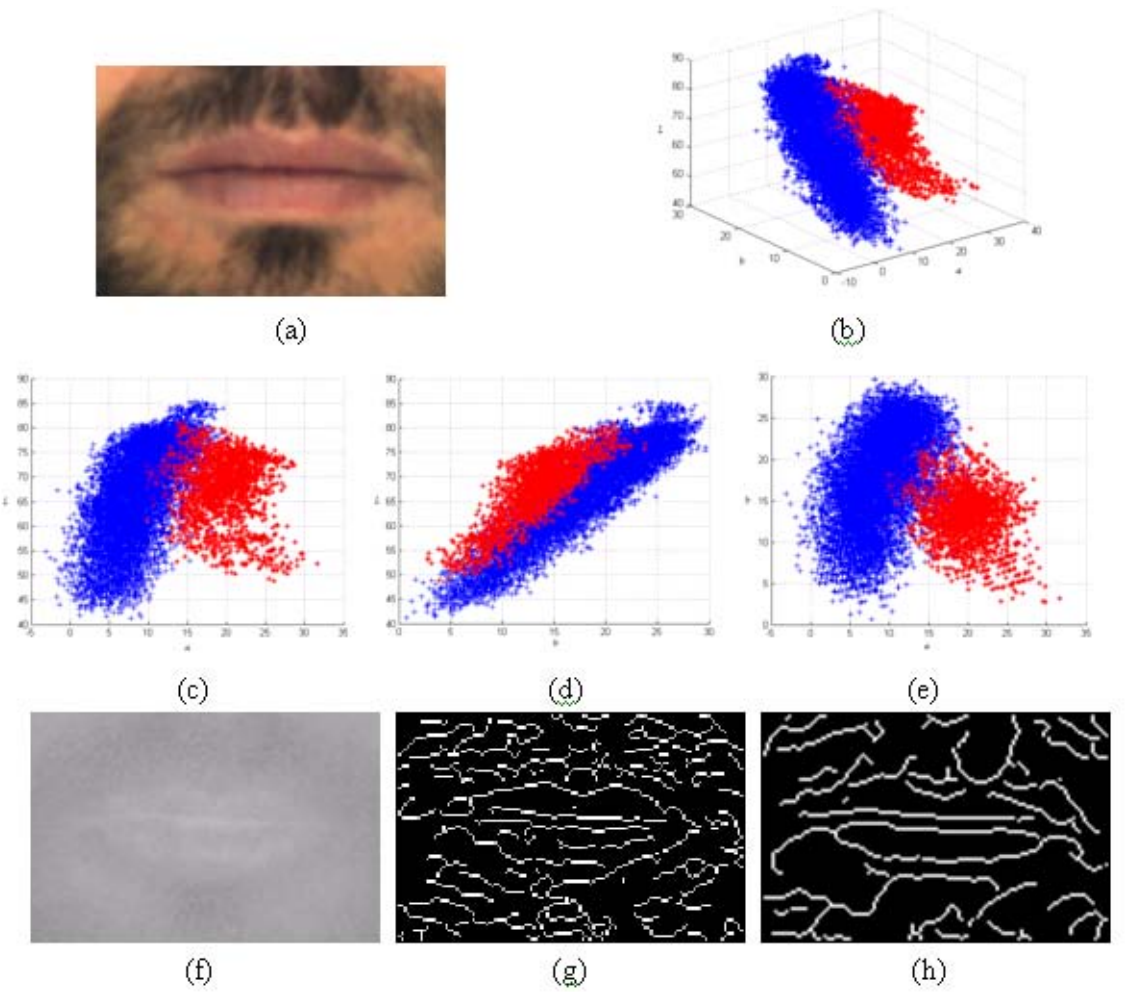
- **The cost function used:** Edge-based, intensity-based, and region-based cost functions are some of the typical approaches for evaluating the model fitness.
- **The optimization procedure:** Since iterative technique is often required to search for the best parameters, the convergence speed and the stability are the two important issues in optimization. The choice of cost function would affect the optimization scheme.

In our recent work (Wang, Lau, & SLeung, 2004), we have proposed a new lip modeling and extraction algorithm that fits a 16-point lip model by minimizing a region-based **cost function** using a point-driven **optimization scheme**.

FUZZY-CLUSTERING-BASED LIP IMAGE SEGMENTATION

Lip region segmentation is a difficult problem. Figure 1 shows a lip image and its corresponding color distribution for the lip and non-lip pixels in the CIE-1976 CIELAB color space, where * and + represent the lip and background (or non-lip) pixels, respectively. The hue image, with the hue definition given in Zhang and Mersereau (2000), and the edge map are also shown. We see that lip and non-lip pixels overlap severely in the color space, and cannot be easily

Figure 1. (a) Original lip image; (b) Color distribution in CIELAB color space; Color distribution projection on (c) L-a plane, (d) L-b plane, (e) b-a plane; (f) Hue map of the lip image; (g) Edge map based on hue information; (h) Edge map based on intensity information



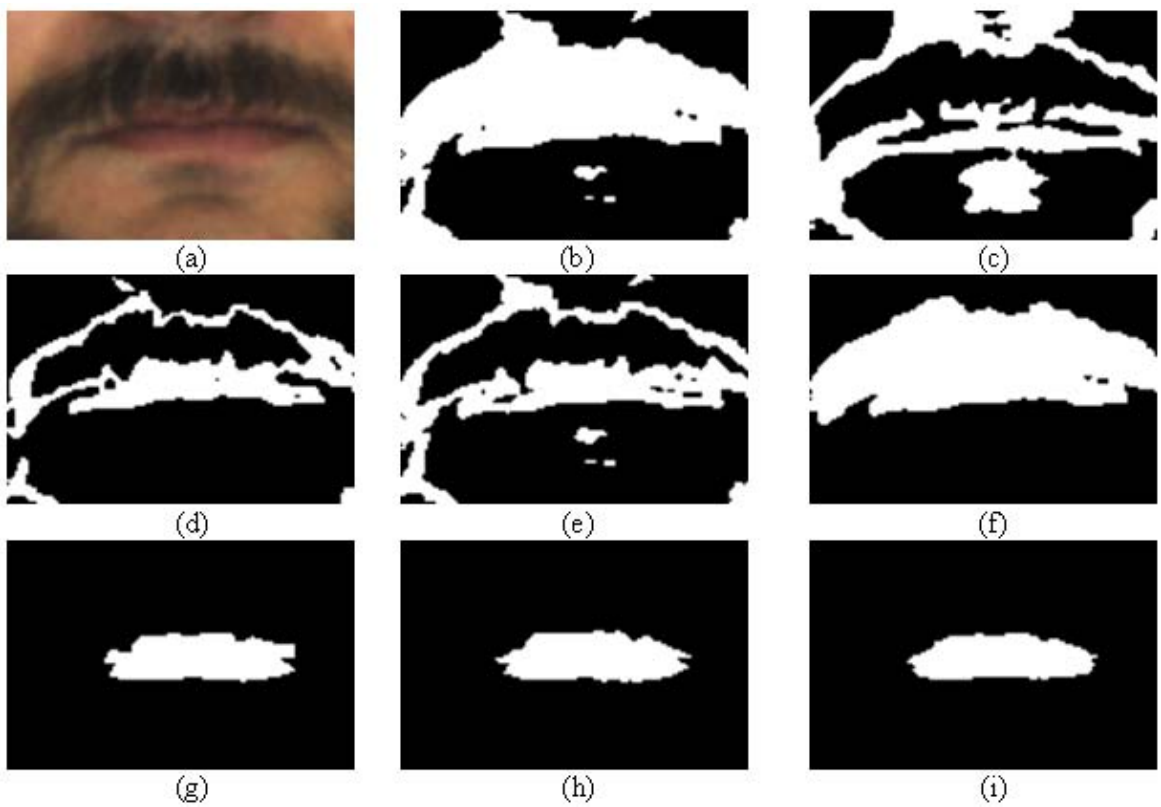
separated based on color information alone. The presence of facial hair results in many luminance and hue edges in the image, and the use of edge information for lip boundary detection becomes difficult. Moreover, the traditional two-class partitioning methods are utterly not appropriate, since the background region is too complex and inhomogeneous. In order to overcome these difficulties, we recently proposed a **fuzzy-clustering**-based lip-segmentation algorithm that can handle complex background, called multiclass, shape-guided fuzzy c-means (MS-FCM) clustering.

Fuzzy c-means (FCM) clustering aims to assign a membership value to each pixel, based on the distance to the cluster centroids in the feature space, in order to minimize a fuzzy entropy measure. In the conventional FCM algorithm, the clustering process depends solely on the feature space distribution, and each pixel is treated independently. However, for data with an inherent spatial ordering, such as image, additional information from the spatial connection among pixels should be exploited (Liew et al., 2000, 2003). Furthermore, for the specific task of **lip segmentation**, prior knowledge about the lip region could also be exploited (Leung et al., 2004; Wang et al., 2004, 2007).

Since lip pixels form a large patch, their distances from the lip center provide useful information to differentiate the lip region from the background. In MS-FCM, such prior shape information is seamlessly incorporated into the objective function to discriminate the non-lip pixels located at a distance away from the lip. **The spatial term** in the objective function penalizes the lip membership value of non-lip pixels outside the lip boundary, even though they might have similar color distribution as the lip pixels. The shape information utilized in the "shape-guided" algorithm helps to overcome the color overlap problem between the lip and background region.

The color information of the background pixels usually has a multimodal distribution due to background inhomogeneity. Modeling it using one spherical cluster would incur large segmentation error. In order to improve the modeling accuracy, multiple clusters are employed to describe the complex background region. An adaptive method is designed

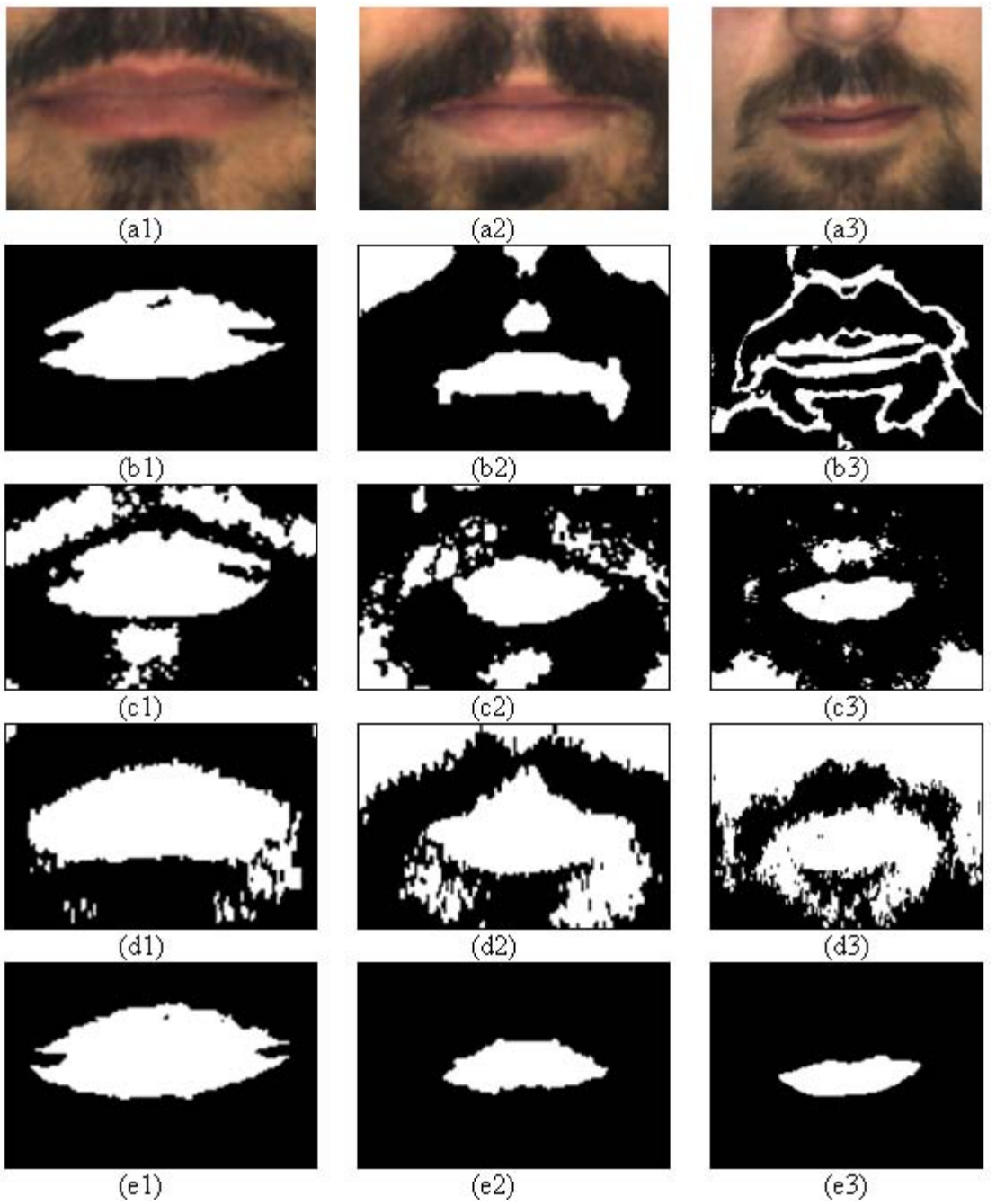
Figure 2. (a) Original lip image; lip-segmentation result obtained by the conventional FCM with (b) C=2, (c) C=3, (d) C=4, (e) C=5; lip-segmentation result obtained by MS-FCM with (f) C=2, (g) C=3, (h) C=4, (i) C=5



to select the appropriate number of background clusters to collectively model the complex background distribution. The "multiclass" feature of our algorithm helps to reduce misclassification caused by inadequate background modeling.

The "multiclass" and the "shape-guided" features are the two key novelties of our algorithm, and they work together to overcome the difficulties of segmenting lip images with complex background. Figure 2 shows the typical segmen-

Figure 3. (a1),(a2),(a3) Original lip images. Segmentation results of: (b1),(b2),(b3) conventional FCM, (c1),(c2),(c3) Lievin's method, (d1),(d2),(d3) Zhang's method, and (e1),(e2),(e3) MS-FCM.



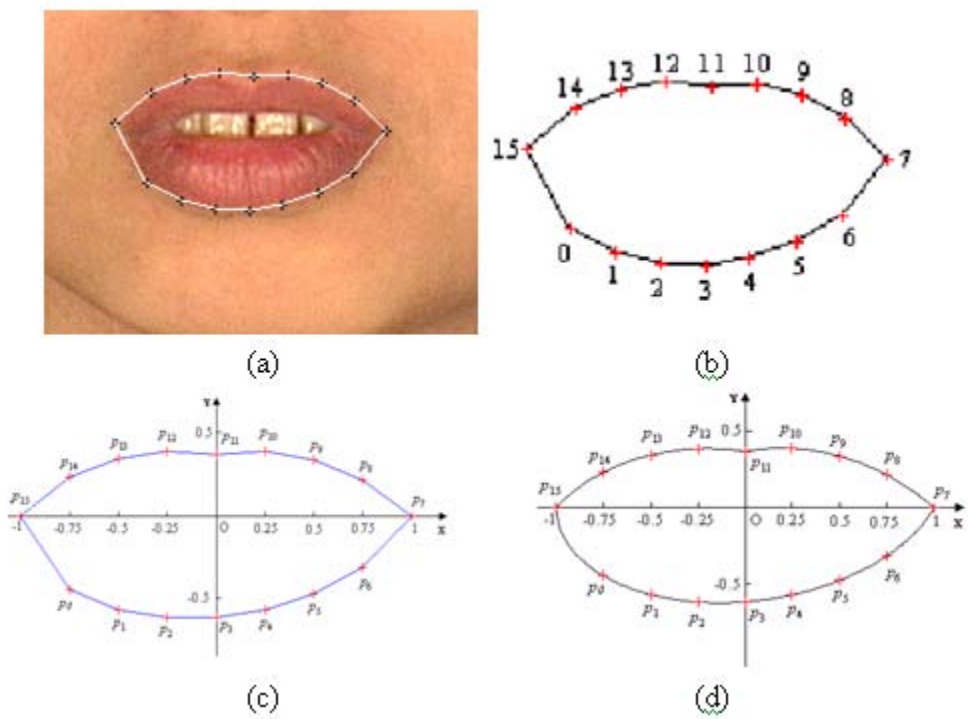
tation results obtained from the conventional FCM and the MS-FCM algorithm using different numbers of background clusters (where C is the total number of clusters). It is observed that the conventional FCM is unable to segment the lip region accurately for an image with complex background, even with different setting of C; whereas MS-FCM is able to clearly segment the lip region for C larger than 2 . The results indicated that the use of prior shape information and multiple background clusters are both necessary to ensure a good segmentation. Without using multiple background clusters, large segmentation error occurs due to insufficient background modeling, even if shape information is considered (Figure 2(f)). Without the prior shape information, poor segmentation result is obtained, even if multiple background clusters are employed (Figure 2(c)(d)(e)).

In Figure 3, three difficult lip images are used to compare the performance of the proposed algorithm with the conventional FCM, Lievin and Luthon's method (Lievin's for short) (1999), and Zhang and Mercereau's method (Zhang's for short) (2000). It can be seen that the conventional FCM can deliver acceptable results if the lip and the background are well differentiated (see Figure 3(b1) with three clusters). However, when the lip color and part of the background color are close, the conventional FCM is unable to produce good segmentation, even with more clusters (see Figure 3(b3) with five clusters). In Lievin's and Zhang's methods, basically two clusters are used to segment the image using the hue information. When the hue of the lip region and the background are close, the two-class assumption becomes inappropriate, resulting in poor segmentation (see third and forth rows in Figure 3). Zhang's method further makes use of edge information to aid the segmentation. It produces large segmentation errors, since the edge map is noisy for lip image with beards. Finally, the last row shows the segmentation results obtained from our algorithm with three clusters. It clearly outperforms the other three methods.

MODEL-BASED LIP CONTOUR EXTRACTION

In our method, a 16-point lip model is used to describe the lip contour, as shown in Figure 4(b). These points are divided into three groups: p_0 to p_7 and p_{15} describe the lower lip; p_7 to p_{11} and p_{11} to p_{15} describe the upper-right lip and the upper-left lip, respectively, p_7 and p_{15} are the lip corners, and

Figure 4. (a) Lip image with contour shown in white; (b) original 16-point lip model (before normalization); (c) normalized 16-point lip model; (d) lip contour points constrained by the geometric lip model.



p_{11} is the dip point. In order to reduce the number of free parameters in the model, a normalization process is used to translate the lip corner points p_7 and p_{15} to lie on the *x*-axis and the dip point p_{11} on the *y*-axis (see Figure4 (c)). Moreover, a geometric model, consisting of three connected quadratic curves, is used to constrain the extracted lip contour to be physically valid (see Figure 4 (d)). The geometric model is flexible enough to describe different lip shapes in that: (i) the asymmetry between the left and right sides caused by skewing is allowed; (ii) no bounded assumption is made about the relationship between the upper lip and lower lip except for the connections on the corner points and dip point. As the geometric model is only used to constrain the 16-point lip model, there is no need to explicitly estimate the parameters of the geometric model from the image.

Given the probability map generated by the MS-FCM, the initial lip model can easily be derived. The left and right lip corners can be found by scanning the columns. The top and bottom boundaries can then be derived from the center axis between the left and right corner points. Then three quadratic equations are used to fit the upper and lower lip points independently. Finally, the initial 16-point model λ_p is obtained from this initial fit.

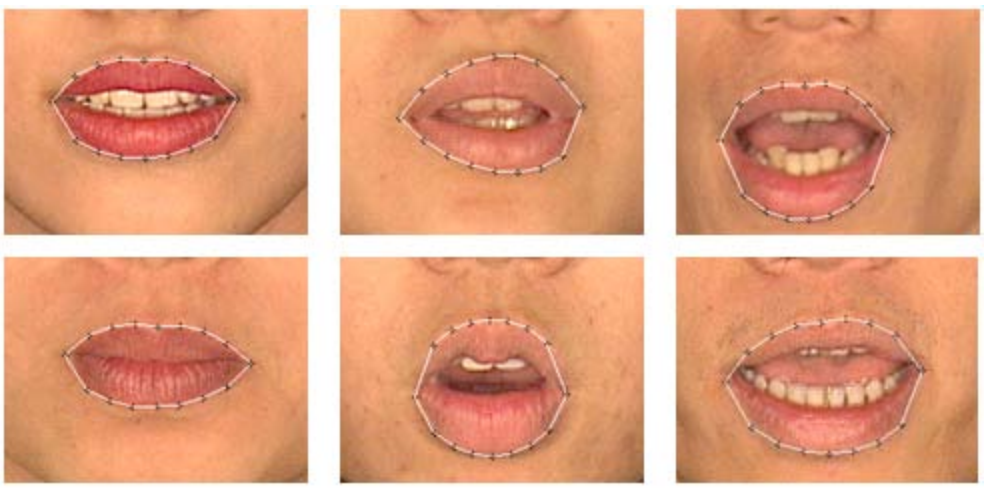
We use a region-based cost function to find the optimum model parameters such that the pixels inside the lip region have high probability of being lip pixels, while those outside have low probability. Region-based cost functions have larger capture range than their edge-based and intensity-based counterparts. They are also more robust and tolerant against changes in noise and illumination. We employ **an iterative point-driven optimization method** to adjust the lip-point locations. Since the optimization is point-driven rather than parameter-driven, the lip contour update becomes better controlled and, consequently, faster convergence is achieved. In each iteration, the displacement vector $\Delta\lambda_p$ and the new lip model can be updated by calculating the partial derivatives of the cost function with respect to λ_p (Sum, Lau, Leung, Liew, & Tse, 2001). Then after normalization, three quadratic curves are fitted to the 16-point model to obtain the parameters of the geometric model λ_g, which is subsequently used to constrain the 16-point model to form a valid lip shape.

Over 5,000 lip images of size 216×162 have been used to test the performance of our algorithm, and accurate lip contours can be obtained. Some lip extraction results are shown in Figure 5.

FUTURE TRENDS

Currently, the lip model parameters, found in the previous frame, are used as the initial parameters for the next frame. However, this simple approach is inadequate if the lip shape changes rapidly between successive frames due to fast articulation. Online prediction, based on Kalman filtering, can be used to provide a better initial estimate for tracking. The lip segmentation algorithm should also be made more robust to low image quality. In many applications, the captured videos are usually in compressed form. The degradation in image quality, due to compression, would undermine the fuzzy clustering process by biasing the color information. Transforming the color image to the S-CIELAB

Figure 5 Lip contour extraction results of different isolate lip images.



color space might be a possible solution for improving the segmentation accuracy.

CONCLUSION

Accurate lip extraction from lip image sequence is an important first step in automatic **lipreading** and **visual speaker authentication**. In this chapter, we described a lip segmentation and extraction algorithm that is accurate, robust, and efficient. In our segmentation algorithm, the shape information is incorporated into the fuzzy clustering process such that pixels with similar color, but in different regions, can be differentiated. In addition, multiple clusters are collectively used to model the complex background distribution. In our lip extraction approach, a 16-point lip model with geometric constraint is used to capture the lip shape. A point-driven optimization scheme is employed to speed up the iterative optimization process. Our algorithm is able to accurately extract the lip contour from video sequences in real time.

REFERENCES

Caplier, A. (2001). Lip detection and tracking. In *Proc. of 11th Int. Conf. on Image Analysis and Processing* (pp. 8-13).

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence, 23*(6), 681-685.

Cootes, T. F., Hill, A., Taylor, C. J., & Graham, J. (1994). Use of active shape models for locating structures in medical images. *Image and Vision Computing, 12*, 355-365.

Eveno, N., Caplier, A., & Coulon, P. Y. (2001). New color transformation for lips segmentation. In *Proc. of IEEE Fourth Workshop on Multimedia Signal Processing* (pp. 3-8).

Eveno, N., Caplier, A., & Coulon, P. (2003). Jumping snakes and parametric model for lip segmentation. *Proc. of ICIP'2003, 2*, 867-870.

Hennecke, M. E., Prasad, K.V., & Stork, D. G. (1994). Using deformable templates to infer visual speech dynamics. *Proc. of the 28th Asilomar Conference on Signals, Systems and Computers, 1*, 578-582.

Kaynak, M. N. , Zhi, Q., Cheok, A. D., Sengupta, K., Jian, Z. & Chung, K. C. (2004). Analysis of lip geometric features for audio-visual speech recognition. *IEEE Trans. on Systems, Man and Cybernetics, Part A, 34*(4), 564 – 570.

Leung, S. H., Wang, S. L., & Lau, W. H. (2004). Lip image segmentation using fuzzy clustering incorporating an elliptic shape function. *IEEE Trans. on Image Processing, 13*, 51-62.

Lievin, M., Delmas, P., Coulon, P. Y., Luthon, F., & Fristot, V. (1999). Automatic lip tracking: Bayesian segmentation and active contours in a cooperative scheme. *Proc. of IEEE International Conference on Multimedia Computing and Systems,.1,* 691-696.

Lievin, M., & Luthon, F. (1999). Lip features automatic extraction. *Proc. of IEEE ICIP'1999, 3*, 168-172.

Liew, A.W.C., Leung, S.H., & Lau, W.H. (2000). Fuzzy image clustering incorporating spatial continuity. *IEE Proceedings-Vision Image and Signal Processing, 147*(2), 185-192.

Liew, A. W. C., Leung, S. H., & Lau, W. H. (2002). Lip contour extraction from color images using a deformable model. *Pattern Recognition, 35*, 2949-2962.

Liew, A. W. C., Leung, S. H., & Lau, W. H. (2003). Segmentation of color lip images by spatial fuzzy clustering. *IEEE Trans. on Fuzzy Systems, 11*, 542-549.

Luettin, J., Thacker, N. A., & Beet, S. W. (1996). Visual speech recognition using active shape models and hidden Markov models. *Proc. of IEEE ICASSP'1996,2*, 817-820.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Trans. on Pattern Analysis and Machine Intelligence,24*(2), 198-213.

Mok, L. L., Lau, W. H., Leung, S. H., Wang, S. L., & Yan, H. (2004). Lip features selection with application to person authentication. *Proc. of IEEE ICASSP'2004, 3*, 397-400.

Sum, K. L., Lau, W. H., Leung, S. H., Liew, A. W. C., & Tse, K. W. (2001). A new optimization procedure for extracting the point-based lip contour using active shape model. *Proc. of IEEE ICASSP'2001, Salt Lake City, 3*, 1485-1488.

Wang, S. L., Lau, W. H., & Leung, S. H. (2004). Automatic lip contour extraction from color images. *Pattern Recognition, 37*, 2375-2387.

Wang, S. L., Lau, W. H., Leung, S. H., & Liew, A. W. C. (2004). Lip segmentation with the presence of beards. *Proc. of IEEE ICASSP'2004, 3*, 529-532.

Wang, S. L., Lau, W. H., Liew, A. W. C., & Leung, S. H. (2007). Robust lip region segmentation for lip images with complex background. *Pattern Recognition, 40*(12), 3481-3491.

Zhang, X., & Mersereau, R. M. (2000). Lip feature extraction towards an automatic speechreading system. *Proc. of IEEE ICIP'2000, 3*, 226-229.

KEY TERMS

Cost Function: Also called objective function. A function associated with an optimization problem which determines how good a solution is.

Fuzzy C-Means Clustering: Fuzzy C-means clustering is the most well-known partition-based clustering algorithm. The algorithm starts by choosing *k* initial centroids, usually at random. Then the algorithm alternates between updating the cluster membership value of each data point with different cluster centroids and updating the centroids based on the new clusters until convergence.

Lip Contour Extraction: A process to derive the lip contour information from the lip image.

Lip Modeling: An approach to describe the lip region by a number of model parameters. Model definition, cost function formulation and the optimization process are the key issues of the lip modeling technique.

Lip Region Segmentation: Lip region segmentation is a process by which all the pixels in the lip image are partitioned into two categories, that is, the lip pixels and the non-lip ones.

Local and Global Spatial Information (for lip segmentation): In lip region segmentation, spatial information is incorporated to improve the segmentation performance by the color information only. The local spatial information is referred to the spatial continuity among neighbouring pixels while the global spatial information is referred to the spatial position of a pixel with respect to the lip region (i.e., inside the lip region, outside the lip region or around the lip-background boundary).

Normalization Process: Normalization is a process to refine the input data initial data so as to reduce various undesirable effects (such as translation, rotation in our application).

Point-Driven Optimization (for lip modeling)**:** Point-driven optimization is an iterative procedure to fit the lip model to the actual lip contour. In each iteration, each contour point is adjusted to a better position so as to increase the cost function. Compared with the other optimization methods, such as the parameter-driven optimization, the point-driven optimization approach usually requires less number of iterations and thus is more efficient.

Visual Speaker Authentication: Visual speaker authentication aims to perform speaker's identity authentication based on the identity-relevant visual information, such as the dynamic visual information of the lip movement.

Visual Speech Recognition: Visual speech recognition, also called lipreading, is a process to perform speech recognition only by the speech-relevant visual perception, such as the dynamic visual information of the lip movement.

A Literacy Integral Definition

Norelkys Espinoza Matheus
University of Los Andes, Venezuela

MariCarmen Pérez Reyes
University of Los Andes, Venezuela

INTRODUCTION

Due to the lack of a unique definition of literacy and the need for redefining this conception in a context characterized by the changes generated by the inclusion of new technologies in all aspects of the society, this explicative research article is oriented toward proposing a definition of literacy from an integral conception which is based on three main kinds of literacy: functional, informational and ethical.

This integral conception must orient the basic contents in the school curricula in all current educational models, mainly at the university level. We consider that knowledge is unique; it should not be divided into pieces. Therefore, it is necessary to integrate the new technologies, from this new paradigm, in the contents of the school curricula.

The present article compiles some general considerations about literacy, proposes a new definition of literacy from an integral conception, as well as each one of its components.

BACKGROUND

It is undeniable that info technology has an influence on human issues. It is also true that this influence has become more intense since the end of the 20th century with the rise of the Internet, when several changes appeared in regard to information treatment and interpersonal relationships by offering communication facilities never observed before.

Two breakthrough inventions formed the information society's foundation: computers and telecommunications, which play roles similar to those that the steam engine and electricity played during the industrial revolution (Cellary, 2003). In this sense, the Internet and the current scientific and technological development are the results of society's evolution which has gone through dedifferentiated and clearly defined stages: agrarian society, industrial society, and currently, *information society*.

Cellary (2003) explains in a very simple way the form in which both inventions have a notorious influence on society. On one hand, he shows that although computers can only capture a fraction of their programmers' real intelligence, computers behave like people in that they make correct decisions based on the knowledge encapsulated in the programs they run. On the other hand, telecommunications ensure common access to all computers connected to the Internet, giving the entire society the chance to share and spread information.

This information revolution tends to deprive humans of their decision-making monopoly, given their tireless speed and mathematical precision; computers will always outdo humans in performing any intellectual activity that can be explicitly defined (Cellary, 2003) and the facility for sharing results through the Internet.

This is consistent with the ideas of Gutiérrez (2003) who affirms that this convergence of languages and technologies and the arising of cyberspace as a relational environment promote three important changes: 1) new kinds of predominant documents, 2) new forms of communication, and 3) new education and communication environments.

Additionally, the main features defining the *information society,* according to Castells (2001) are firstly, an informational and technological revolution as basis. Secondly, a socioeconomic reorganization process known as globalization. Thirdly, a change in organization processes (not less deep than the previous one) as the transition from vertical organizations to Web organizations. Gutiérrez (2003) adds that these three factors and the interaction among them, generate important social and cultural changes.

However, the process has not produced purely benefits. Even when science, and in particular technology, has been produced by social progress, and both have undeniably offered great contributions in different areas: health, communication, education, culture, socio-politics, among others, inconveniences have also been produced.

Nowadays, it is necessary to educate citizens from a new perspective: the new citizens must be literate in the use of new technologies, but at the same time they must have a critical thinking, a reflexive and participative sense and be committed to society in order to be able to consciously and responsibly use sources. In other words, they must be integrally literate.

INTEGRAL LITERACY

Through the decades of 1980s and 1990s, the definitions of literacy have been widened to take into account the challenges of globalization, including the repercussions of the new technologies, modern media and the rise of the economies of knowledge (UNESCO, 2004). The definition of literacy employed in the Education for All 2000 Assessment is the following: "Literacy is the ability to read and write with understanding a simple statement related to one's daily life. It involves a continuum of reading and writing skills, and often includes also basic arithmetic skills (numeracy)" (UNESCO, 2004, pp. 12-13).

A proposed operational definition was formulated during an international expert meeting in June 2003 at UNESCO. It states: "Literacy is the ability to identify, understand, interpret, create, communicate and compute, using printed and written materials associated with varying contexts. Literacy involves a continuum of learning in enabling individuals to achieve their goals, to develop their knowledge and potential, and to participate fully in their community and wider society" (UNESCO, 2004, pp. 13).

Other authors have defined literacy from different perspectives. Bawden (2001) suggests that the first and simplest meaning implies only the ability to read and write. The second certainly implies this ability, but also requires something beyond it. Gilster (1997, cited in Badwen 2001) advises that the concept of literacy goes beyond simply being able to read; it has always meant the ability to read with meaning, and to understand. It is the fundamental act of cognition.

Similarly, as Clifford (1984, cited in Bawden, 2001) suggests, expert opinion has abandoned the dichotomous framework, of literate or illiterate, in favor of the conception of literacy as a continuum; at one end lies some ability to reproduce letter combinations. At the other end, lie such language learning behaviors as logical thinking, higher order cognitive skills, and reasoning.

Bawden (2001) indicates that for most of the centuries the term has been in use, it has meant being well educated, well-read, versed in literature and "letters" (the "learned" aspect of the definitions quoted at the start). More recently, it has taken on a more prosaic meaning of being able to make effective use of *information*, gained from written material.

However, all authors cited agree in the nonexistence of a unique concept of literacy and express that this concept must be redefined in a context of the changes generated by the introduction of new technologies in all areas of society.

In this sense, the current technological revolution has favored the existence of a new definition of literacy that goes beyond the abilities to use computers and the Internet, and to create, save, broadcast or download information. For this reason, the most appropriate notion in the 21st century is the one related to "integral literacy" based on three main aspects: functional, informational and ethical literacy.

Integral Literacy includes not only the basic tools that an individual must have an individual like reading, writing and calculation; but also the socio-affective skills in combination with the new technologies used, in the search of changes of attitudes to construct a more humane and free society.

Functional Literacy

"Literacy lies at the heart of UNESCO's concerns and makes up an essential part of its mandate, being entwined with the right to education set forth in the Universal Declaration of Human Rights of 1948. These concerns have to do with promoting the meaningful acquisition and application of literacy in laying the basis for positive social transformation, justice, and personal and collective freedom. Despite tremendous progress made over the past 55 years, universal literacy remains a major challenge for both developing and developed countries in terms of commitment and action. There are over 800 million illiterate adults in today's world, a figure projected to remain unchanged in 2015 if current trends continue unabated" (UNESCO, 2004, pp. 1).

The term "Functional literacy" has been understood as the ability to read, write and to perform basic mathematical calculations. This term was originally introduced by UNESCO (1986), which states:

A person is literate who can with understanding both read and write a short simple sentence on his everyday life (…) A person is functionally literate who can engage in all those activities in which literacy is required for effective functioning in his group and community and also for enabling him to continue to use reading, writing and calculation for his own and the community's development. (UNESCO, 1986, p. 4)

Functional literacy is a matter to be developed by initial education; however, this is not totally achieved because, as Schleicher and Tamassia (2002) indicates, the students enrolled in the high school have very low levels of competence for written expression and reading comprehension. This problem is observed in developed and developing countries, getting worst in the latter, where indexes of reading are below OECD reading abilities statistical averages.

Therefore, it is considered that functional literacy as well as ethical and informational literacy also become a concern for higher education. In this sense, we consider that study plans at higher education must include subjects for reading and writing that guide students in the use of strategies, working on topics and texts of their interest, in order to overcome the limitations generated by this reality.

Informational Literacy

There is a broad discussion about the definition of "Informational Literacy." In this respect, Oxbrow (1998, cited in Badwen, 2001) sees information literacy as differing from, and going beyond, computer literacy by virtue of a changed *focus*: on "the content that flows through the technology - a focus on information and knowledge."

Similarly, Johnston and Webber (1999, cited in Badwen, 2001) suggest that information literacy is "emphatically not just computer literacy: but rather the ability to identify and evaluate information (using whatever tools are appropriate, such as those provided by IT) and learn to "read" information within its cultural and social context."

In relation to the above, we consider that informational literacy is referred to as the capacity to have access to information (in electronic and printed format), selecting, critically classifying as well as assimilating and turning it into knowledge, in other words, to build meaning. It is also referred to as the capacity to turn that meaning into new information (digital or printed). It also includes being able to make that information available for anybody in the cyber society by using the services for that purpose on the Internet.

This capacity to read digital and printed information with critical thinking is mandatory nowadays because the amount of information that currently exists is so large that a human being would spend a lifetime assimilating it and still would not finish. For this reason, according to Espinoza (2003) it is necessary to consider new schemata in which reliability of information can be stated in order to make its selection easier and more critical. This allows the development of higher processes of thought through the use of clearly defined criteria that demand the user to perform an analysis of the material. This analysis should include the document origin, author, topics and issues developed in it, the message itself and the potential audience to which it is addressed.

In this sense, the construction of meanings guarantees each individual the comprehension of the world as well as his or her participation in the processes of social transformations and in the construction of the world we expect to live in. This is another reason for which ethical literacy is needed.

Ethical Literacy

It is necessary to go toward a more humane information society, a society in which, according to Acosta (1996), we can find solutions to the ethical and moral problems we currently face. In this way, according to the aforementioned author, we can find solutions to all the other problems: world illiteracy rate, disappearing of natural resources (renewable, biological and mineral), pollution, demographic problems, hunger, undernourishment, energy supply, among others, as long as the situation seems to be a matter of values instead of a technological matter.

It is for that reason that not only is it necessary to be literate at the functional and informational level, but also at the ethical one. In this sense, Emery and Anderson (1995) indicate that reading and writing constitute simultaneous and interactive social processes that involve respect for human rights. A reader does not establish the meaning of text in an isolated manner, but rather does so through social interaction. The form in which the individuals speak or use what they have read reflects and gives form to their own cultural identities. Writing is also a social process in which writers construct speech from certain communities or sociocultural contexts using complex rhetorical strategies that allow them to express as much who they are as what they mean.

Ethical literacy refers to assuming reading and writing with a critical thinking, showing respect to other people's positions, with conscious use of the information, with respect to the intellectual property, and to rescue the moral principles and the ethical values established in the information society.

In this sense, at the present time it is necessary to form concerned, responsible citizens, committed to the social and economical development of society that satisfy the demands of the public and private economic sector, that contribute to technological and scientific research and newly direct moral and ethical values through the construction of a society based on solidarity, the shared use of information, free promotion and development of technology from a more human perspective and respect for human rights (De Gortari, 1984). These issues have been recognized by the Universal Declaration of Human Rights and become more important in the scientific and technological activities because the development of scientific research demands respect for those rights and the enjoyment of freedom (UN, 1948).

FUTURE TRENDS

Literacy from this new paradigm is extended to all levels of education and new technologies are an integral part of this process. This provides benefits in the teaching-learning process of the language in other dimensions, and gives new opportunities for its knowledge and use.

Additionally, integral literacy favors the understanding of the world in which each individual develops, so that he or she can act critically, creatively and consciously in that world.

CONCLUSION

In this *information society*, it is mandatory to start facing the challenge of integral literacy in all current educational models. To achieve this, we must develop a change of attitude because we cannot just be consumers of information. We

need to redefine our role by becoming actors and interacting in a dynamic, changing process demanded by the current model of society.

In this sense, literacy understood from an integral conception, based on these three kinds of literacy, functional, informational and ethical, acquires great importance.

Integral literacy lets citizens acquire better written expression as well as better reading comprehension in all the media. This fortifies one's autonomy, self-esteem, self-confidence, and attitude before the decision-making and solution of problems. Integral literacy also contributes to the enrichment of interpersonal relationships and the consideration of new forms of communication, which would contribute to one's personal and social development.

We are conscious that this proposal does not completely cover the issue. However, it opens new ways for reflection and debate about such a complex definition that is currently the object of interest for many researchers.

REFERENCES

Acosta, O. (1996). *Componente humanístico en la carrera de ingeniería*. Valencia, Venezuela: Universidad de Carabobo.

Badwen, D. (2001). Information and digital literacies: A review of concepts. *Journal of Documentation, 57*(2), 218-259. Retrieved December 14, 2007, from http://dlist.sir.arizona.edu/895/01/bawden.pdf

Castells, M. (1997). La era de la información. *Economía, sociedad y cultura I.* La sociedad en red. Madrid: Alianza.

Castells, M. (2001). La galaxia Internet. *Reflexiones sobre Internet, Empresa y Sociedad*. Barcelona: Areté.

Cellary, W. (2003). The profession's role in the global information society. *Computer, 9*(36), 122-124.

De Gortari, E. (1984). *Indagación crítica de la ciencia y de la tecnología*. México, D.F.: Editorial Grijalbo.

Emery, W., & Anderson, A. (1995). *Media files-introduction, en Ministère de l´éducation: Media files*. Québec: Gouvernement du Québec.

Espinoza, N. (2003). Criterios para la selección de información científica odontológica en la World Wide Web. *Acta Odontológica Venezolana, 41*(3), 251-257.

Goodman, K. (1994). Reading, writing, and written texts: A transactional sociopsycholinguistic view. In R. Ruddell, M. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed.) (pp. 1093-1131). Newark, DL: International Reading Association.

Goodman, K., & Goodman, Y. (1983). Reading and writing relationships: Pragmatic functions. *Language Arts, 60*(5), 590-99.

Gutiérrez, A. (2003). *Alfabetización digital. Algo más que ratones y teclas*. Barcelona, España: Editorial Gedisa.

Schleicher, A., & Tamassia, C. (2002). Measuring student knowledge and skills: The PISA 2000 experience. *Statistics Brief OECD, 4*, 1-8. París: OECD. Retrieved December 14, 2007, from http://www.oecd.org/dataoecd/44/63/33692793.pdf.

Smith, F. (1981a). *Writing and the writer*. New York: Holt, Rinehart & Winston.

Smith, F. (1981b). Myths of writing. *Language Arts, 58*(7), 792-798.

UNESCO. (1986). Revised recommendations concerning the international standardization of educational statistics. *UNESCO's Standard-setting Instruments, 3(*4), 1-9. Retrieved December 14, 2007, from UNESCO Web site: http://www.unesco.org/education/nfsunesco/pdf/STATIS_E.PDF

UNESCO. (2004). *The plurality of literacy and its implications for policies and programmes.* UNESCO Education Sector Position Paper. Retrieved December 14, 2007, from UNESCO Web site: http://unesdoc.unesco.org/images/0013/001362/136246e.pdf

United Nations. (1948, December 10). *Universal declaration of human rights*. United Nations General Assembly Resolution 217 A (III): United Nations document A/810 at 71. París, France.

KEY TERMS

Ethical Literacy: Refers to assuming reading and writing with a critical thinking, with respect toward the position of others, with conscious use of the information, and with respect to the intellectual property, and to rescue the moral principles and the values established in the information society.

Functional Literacy: This has been understood as the ability to read, write and to perform basic mathematical calculations.

Informational Literacy: Referred to as the capacity to access information (in electronic and printed format), select, critically classify as well as assimilate and turn it into knowledge; in other words, to build meaning.

Information Society: A specific form of social organization in which information generation, processing, and transmission are transformed into the fundamental sources

of productivity and power (Castells, 1997*)* due to two human inventions: computers and telecommunications.

Integral Literacy: It is based on three kinds of literacy: functional, informational and ethical. Includes the basic tools that an individual must have like reading, writing and calculation; and the socio-affective skills in combination with the new technologies used in the search of changes of attitudes to construct a more humane and free society.

Reading: A dynamic, transactional, sociopsycholinguistic process of constructing meaning and making sense of print (Goodman, 1994; Goodman & Goodman, 1983) and digital information.

Writing: Is a process of constructing meaning in which writers actively integrate thought and language (Smith, 1981a, 1981b; Goodman, 1994; Goodman & Goodman, 1983).

Category: Environmental Informatics

Location Information Management in LBS Applications

Anselmo Cardoso de Paiva
Federal University of Maranhão, Brazil

Erich Farias Monteiro
Empresa Brasileira de Correios e Telégrafos Regional Maranhão, Brazil

Jocielma Jerusa Leal Rocha
Federal University of Maranhão, Brazil

Claudio de Souza Baptista
University of Campina Grande, Brazil

Aristófanes Corrêa Silva
Federal University of Maranhão, Brazil

Simara Vieira da Rocha
Federal University of Maranhão, Brazil

INTRODUCTION

The mobile computing advent brings a set of new applications that benefit from the constant need of information, diminishing communication costs and favoring the popularization of mobile devices, to reach an increasing number of users.

The mobility characteristic opens a new area for software applications. Associated to the mobility we have the location identification, which turns into a critical attribute, once it allows the development of a great variety of new services and applications. The systems that benefit from the use of that location information are named location-based systems (LBS); alternatively, these applications are also known as location-aware, context-aware, or adaptive information systems

More precisely, we can define LBS as applications that use the location information to supply services, based on this position context, to their users (Kupper, 2005; Schiller & Voisard, 2004).

The user location information makes available completely new and innovative service concepts, offering information to the user based on its own context (e.g., climatic information in the region where the user is located), increasing considerably the utility of these services. We know that location-based applications increase the services effectiveness, as they give a customized access to the data based on the user's preferences and on its actual position. This enhances the personalization content, giving several benefits to users and to the application developers.

In our daily life, several activities may use these services, like the emergency call centers, the car navigation services, and even location-based friend finder.

We may verify that, beyond the already cited characteristics and benefits, what also gave the LBS applications a growing perspective were the location techniques modernization and the mobile devices popularization, enabling the offer of more precise, objective, and useful information. In Shiode et al. (Shiode, Li, Batty, Longley, & Maguire, 2002), research shows the trend of LBS market and the market potential reserved to this class of applications that, each year, turns out to be more important to the users, becoming the area that dominates the applications for mobile devices. According to Sayed (2005), the forecast annual revenues for location-based services was estimated in US $3.3 billions for United States in 2006/2007, and in US $11.7 billions on the other countries.

In summary, we may say that the positional information has the potential to explore the user's geographical context as one of the most important variables for content and services personalization for mobile devices users.

BACKGROUND

The processes to manage data in location-based applications are especially challenging, as we need to deal with information as the user is moving from one place to other, in an environment with limited resources and also with

heterogeneity. Diverse research areas contribute to ease the process of LBS application development to make possible to use all the spectrum of functionalities that can be implemented through these applications.

Developments in diverse areas such as databases, positioning technology, software engineering, and others, have been made, trying to ease the construction of LBS applications in a way to provide a large spectrum of services in this kind of applications.

Diverse works were proposed dedicated to developing frameworks that provide reuse in the development of LBS applications.

In Wolfson (1999), moving objects databases that store mobile objects location information are considered, especially the location information. The work concentrates on the query and update problems. For the update problem, an information cost model based on the communication cost and information accuracy is proposed.

Large-scale architecture for location services is proposed in Leonhardi and Rothermel (2002). The architecture presents a model of a location service (or generic API), defining the semantics of position queries (position), area queries (range), and proximity queries (nearest neighbor). To be scalable, it defines a distributed and hierarchical organization for the servers.

In Agre et al. (Agre, Akinyemi, Ji, Masuoka, & Thakkar, 2002), layer-based architecture is presented as basis for the construction of location-sensible systems. The paper defines a location service module (LSM) that is a middleware layer that simplifies the development of location-based services, as it gives a uniform interface that encapsulates the details related to the location information capture, and also gives several useful functionalities as location determination technology commutation, error estimation, location determination technology combination, and cooperative location determination

In Nord et al. (Nord, Synnes, & Parnes, 2002), an architecture for location aware applications is proposed, where positioning source devices, such as GPS, WaveLan, and Bluetooth, may be easily combined or intercalated to provide positioning services in a more precise way. The proposed architecture also supports a peer-to-peer communication, allowing the clients to know the others location, and also combining the location information with other contextual information.

Between several initiatives for the development of frameworks for location-based applications, we may highlight the ICING project (Kilfeather, Carswell, Gardener, Rooney, 2007). It has focused on multimodal, multidevice communications to provide enhanced services

Several location-based applications may be cited; among the most promising, we have the emergency call (Hargrave, 2000) (E911, 2001), navigation systems (ApontadorDuo, 2007), and support systems (Boondao, Esichaikul, & Kumar, 2003).

Also, there is an effort of standardization by the OpenGeoSpatial Consortium. The OpenLS (OpenLS, 2008) initiative, one of these efforts, is focused on the development of interfaces for the easiness of location information and other spatial information use over a wireless infrastructure. Its main aim is to integrate geospatial data and geoprocessing tools in location services, making available these functionalities for a large number of applications.

The management of location information is a central point in location-based applications development. Thus, all these applications share a common component, that is, in charge of the location information acquisition, communication, and storage, which we call location information management. This led us to the need for these functionalities reusing as a way to hasten the development process. In the next section, we describe an architecture that may be reusable, and a specification of a framework to permit the reutilization of the information location management.

LOCATION INFORMATION MANAGEMENT

Location information management is the management of location information of objects that may move, being in different locations in different times. This is an important aspect for several applications (Wolfson, 2008) such as flythrough visualization, context awareness, augmented reality, and cellular communication.

In the context of cellular communication systems, the location management problem has two main problems: point query to locate a cellular user that must be located, and point update to set the new cell that the user is when he/she moves beyond its original cell boundary. The problem, in this case, is restricted to how frequently it updates, and how to search a database of location records. For more information on this specific topic, see Pitoura and Samaras (2001).

In general, location-based applications, the problem is much broader. We need finer resolution, possibly with queries on the current, past, and future location, and triggers are very important; all these based on sets of objects. Actually, we see that a common approach for LBS applications development is to build a separate, independent location management component for each application. That results in significant complexity and duplication of efforts. Thus, as stated in Wolfson (2008), we need to develop location-management technology that addresses the common requirements and serves as a development platform.

To model the location of mobile objects, a commonly used approach is to model the location as a pair composed of location information (l) and time stamp (t) that is gener-

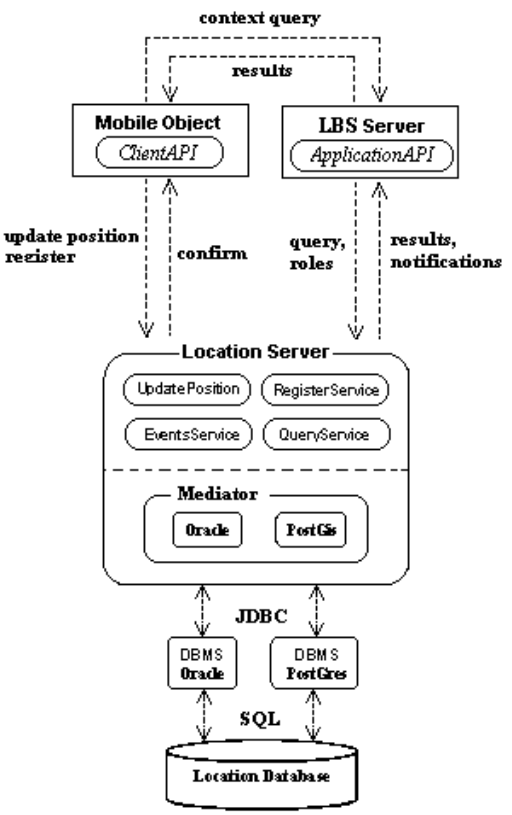
ated periodically. This pair is stored in a database, and SQL is used to retrieve the location information.

As the location information pair (l,t) is acquired on the mobile object, or computed by a network-based process, it has to be transmitted and stored in a location server. Both cases involve wireless communication. To control the transmission of location information, different update protocols can be used.

According to Leonhardi (2000), we can classify the updating protocols in three classes: querying, reporting, and combined protocols, and each class has a typical number of variations.

A protocol is classified as querying if the server decides when to request the location information from the information source. In this case, the implemented location information source becomes very simple. This becomes important for thin clients. This protocol results in a high location precision level, but may also result in a high number of messages if the location is frequently requested. The server response time is also comparatively high, as it needs to contact the client at each request. But, when the user privacy request does not allow its storing in the server, this protocol must be used.

Figure 1. General FRAGIL's architecture



When the server stores the last location information updates, we have the cached querying. In this case, when the object location information is requested, the server analyses if the information that is stored has enough precision to be used, requesting explicitly an update if necessary. Finally, there is the technique where the server requests the location information at each time interval t, named periodic querying.

To handle the questions that arise from the location information point management, we propose a specific framework that allows a complete reuse of the location information management functionalities that is independent of the database management system and is easy to be connected to an LBS application; it is named FRAGIL. The FRAGIL architecture (Monteiro, Paiva, Silva, & Baptista, 2005) is depicted in Figure 1; we can see that the framework is composed of three layers: location server, ClientAPI and ApplicationAPI.

The location server (LS) is responsible for the management of the mobile objects (mo) location information; offering to the LBS application, the services related to the mo position. The location server is composed of several components that are in charge of specific tasks, such as definition and control of the location information update protocol to optimize the access, register a new object to be tracked, queries submission, control of events related to objects location and location information updating.

The ClientAPI is a communication interface that is implemented by the mobile client to communicate to the LS. Using this API, the client requests its registration and send its location information according to the defined update policy. On the other side, the ApplicationAPI is an interface that is used by the LS to communicate with the server that has the LBS application semantics, here is named LBS Server (LBSS).

To make the framework independent from the spatial query languages specificity, a mediator component that is in charge of the translation of queries, was created; built in accordance with the OpenGeospatial standard to the language specific of each data source,. Initially, we have two implemented data sources for Oracle and PostGreSQL.

The LS provides the following services: UpdatePosition, which enables the mo updating; RegisterService manages the registration of new objects to be tracked, being in charge of the definition of location updating; QueryService provides a set of queries based on the mo location; and the EventService, which permits the set up and control of events related to the mo location.

One of the LS services is the RegisterService. This service is used to register a mobile object into the framework, necessary to initiate its location information management.

To control the information location transmission from the mobile object to the server, there are different update protocols. The FRAGIL architecture allows the use of diverse

updating politics, providing the application developers the option to choose the actualization technique better suited to the application requirements.

The ability to answer several queries related to the mobile objects location is implemented by the component named QueryService and its associated API. The application may submit the following queries: positionQuery, that needs the mo Id and returns the actual mo position; temporalPosition-Query, that returns the position of a mo related with a time in the past; temporalRegionQuery(Area,Time), in which the application submits a geographical area as parameter and gets as response a list of objects that are inside of the area; pathQuery, returns the path covered by a mo; and proximityQuery, that verify the proximity of other mo from a specified location or from other mo.

Another service made available is the registration of events based on the mo position related to fixed geographical locations or area and to others mo. Initially, the application server (LBSS) may, through the ApplicationAPI, register those events named RegionEvent and DistanceEvent. As it is registered a RegionEvent the client is notified when the mo enter o leave the pre-defined area. To the DistanceEvent, the notification is sent when the mo is at a specified distance of the other mo or geographical location. The events are defined as: RegionEvent and DistanceEvent.

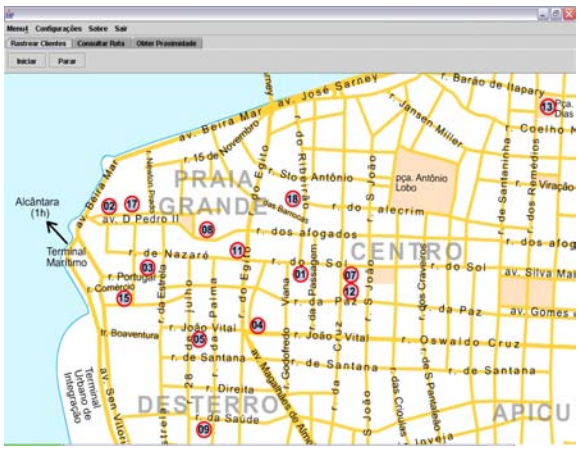
The framework uses an independent data schema for the mo location information that stores the mo location information along the time and its management based on sessions. There are four relations: OBJ_MOB_HST, OBJ_MOB, OBJ_SESSION and VW_MO_LOCATION.

The table OBJ_SESSION has the information of all tracking sessions, with a tuple that identifies the mobile object identification (obj_id), the session identification (id_session), the session init time (init), the session end time (finish), actual session status (status), and the actual update protocol (protocol). In the OBJ_MOB table are the location and time stamp of the mobile objects that have an open session. As the session is closed, the mobile object tuples are moved to the OBJ_MOB_HST.

Finally, we have the view VW_MO_LOCATION that represents the actual location of all objects with open session.

In order to validate the framework system, a prototype was developed to work as a central tracking station of users moving in the historical town of São Luis- MA, Brazil. The application provides a graphical user interface that makes possible to visualize the mobile users' position, determine proximity relations, query mobile users' path, and others. The experiment was done using the framework in a simulated environment, where the mobile users' paths were predefined. Figure 2 presents the application user interface.

Figure 2. Monitoring tool interface



FUTURE TRENDS

We may see that the location-based services are just the initial manifestations of the new applications classes. These new applications classes will, in a certain way, be a revolution in the way that we build and use the information systems. Then, we will reach the advent of pervasive and mobile computing.

One of the challenges that must be addressed in the location information management research is the creation of a technology-independent, high-level software application programming interface (API) for location sensing (Patterson, 2003). That is required as no single location-sensing technology is likely to become dominant, and its choice depends strongly on usage context.

Other aspects also need more investigation. We have a need for data models and query languages that are more adequate to the mobile objects location information nature. Also, there must be research into novel methods for mining objects that are moving in the geographical space. And finally, we believe that new strategies must be developed to include other sensory channels in the interaction of the mobile users to support augmented reality. This, we believe that will be the key application for the next generation of location based services.

CONCLUSION

The users need to get access to information anywhere and anytime. This characteristic leads to a natural demand of services that provide information to the users accounting

to their location context. This generates the class of applications named LBS. These applications allow the users, through a wireless connection, to use services based on their geographical location. Also, they have a set of functionalities that may be reused.

We described, in this work, the actual challenges in the location information management presenting a framework named FRAGIL, which is an alternative to help LBS software developers to create applications supporting the management of location information for a set of mobile objects. The proposed framework may be used by diverse LBs applications, being independent of the location determination technology and from the geographical context of the application.

The framework, in a general way, met the expectations with respect to its functionality, portability, and flexibility. The handling of their APIs furnished to the developers of the applications a total abstraction of the management process of the localization information realized by the LS. Such a deed let the efforts towards the system building process be concentrated only in the issues related with the application domain and with the users' interaction with the graphic interface, allowing a significant gain in the time spent for building the system.

REFERENCES

Agre, J., Akinyemi, A., Ji, L., Masuoka, R,. & Thakkar, P. (2002). A layered architecture for location-based services in eireless ad hoc networks. *IEEE Data Engineering Bulletin, 25*(2), 41-47.

ApontadorDuo. (2007). *Webraska do Brasil.* Retrieved May 2007, from http://www.apontadorduo.com.br/

Boondao, R., Esichaikul, V., & Kumar, N., (2003) *A model of location-based services for crime control.* Map Asia Confference..

E911. (2001). *Federal Communications Commission. Fact sheet: FCC Wireless 911 Requirements*.

Hargrave, S. (2000). Mobile location services: A report into the state of the market, white paper. Cambridge Positioning Systems.

Kilfeather, E., Carswell, J., Gardener, K., & Rooney, S. (2007). *Urban location-based services using mobile clients: The ICiNG approach.* GISRUK 2007, Geographical Information Science Research Conference, NUI Maynooth, Ireland.

Küpper, A., Treu, G,. & Linnhoff-Popien, C. (2006) TraX: A devicecCentric middleware framework for location-based. *IEEE Communications Magazine 44*(9), 114-120.

Leonhardi, A., & Rothermel, K. (2002). Lecture of a large-scale location service. In *Proceedings. of the 22nd International Confference on Distributed Computing Systems (ICDCS),* Vienna, Austria.

Monteiro, E. F., Paiva, A. C., Silva, F. J. S., & Baptista, C. S. (2005). Arquitetura de um Framework para o Desenvolvimento de Aplicações Baseadas em Localização. In *Conferência IADIS Ibero-Americana WWW/Internet 2005*, 2005, Lisboa (In Portuguese).

Nord J., Synnes K., & Parnes P., (2002). An architecture for location aware applications. 35th Annual Hawaii International Conference on System Sciences (HICSS'02) January 07 - 10, Big Island, Hawaii, volume 9, p.293.

Patterson, 2003

Pitoura, E., & Samaras, G. (2001). Locating objects in mobile computing. *IEEE Transactions on Knowledge and Data Engineering, 13*(4).

Sayed (2005)

Schiller, J., & Voisard, A. (2004). Location-based services. San Francisco: Morgan and Kaufman.

Shiode, N., Li, C., Batty, M., Longley, P., & Maguire, D. (2002) *The impact and penetration of location-based services. Centre for Advanced Spatial Analysis*, Working Paper, Centre for Advanced Spatial Analysis (UCL), London, UK.

Wolfson, O., Jiang, L. A., Sistla, P., Chamberlain, S., & Deng, M.(1999). Databases for tracking mobile units in real time. In *International Conference on Database Theory* (pp.169-186).

Wolfson, O. (2008). The opportunities and challenges of location information management, Tech. Report presented at Comp. Science and Telecommunications Board Workshop on the Intersection of Geospatial Information and Information Technology, 2002, , Retrieved April/2008, from http://www7.nationalacademies.org/cstb/wp_geo_wolfson.pdf

KEY TERMS

Framework: A set of software routines that provide a foundation structure for an application. Frameworks take the tedium out of writing an application from scratch.

Location-Based Services: Location-based services refers to a class of applications or services that are based on, or enhanced by, information about the spatial location of a user and/or device.

Location Update Protocol: This is the politic used by the location-based service to maintain the mobile object location information updated in a server.

Location Technology: The set of hardware and software tools that may be used to compute the location of a mobile object.

Middleware: Software that functions as a conversion or translation layer. It is also a consolidator and integrator. any programming that serves to "glue together" or mediate between two separate and often already existing programs.

Wireless Network: A type of network that uses radio waves rather than wires to communicate between nodes.

Location-Based Services

Ali R. Hurson
The Pennsylvania State University, USA

Xing Gao
The Pennsylvania State University, USA

INTRODUCTION

The past decade has seen advances in wireless network technologies and an explosive growth in the diversity of portable computing devices such as laptop computers, handheld personal computers, personal digital assistants (PDAs), and smart phones with Internet access. Wireless networking technologies and portable devices enable users to access information in an "anytime, anywhere" fashion. For example, a mobile user (MU) on the highway may query local weather, traffic information, nearby gas stations, next rest areas, or restaurants within 10 miles. Such new demands introduce a new type of services, ***location-based services*** (LBS), where certain location constraints (e.g., the user's current location) are used in the service provision.

The idea of queries with location constraints is originally introduced by Imielinski and Badrinath (1992), in which mobile users are likely to query information relating to their current positions, leading to the need for LBS. Such services are also termed as location dependent information services (LDIS) in Lee, Lee, Xu, and Zheng (2002). LBS system is the context sensitive systems in a mobile computing environment that consider the user's location as a significant and dynamic factor affecting the information and services delivered to the users. The major LBS applications include:

- Destination guides with maps, driving directions, and real time prompt
- Location-based traffic and weather alerts
- Wireless advertising and electronic coupons to nearby mobile devices
- Movie, theatre and restaurant location and booking
- Store locating applications helping users to find the desired services
- Telematics-based roadside assistance (e.g., OnStar from General Motors)
- Personal content and messaging (Live Chat with friends)
- Mobile Yellow Pages provide local information
- Information Services (News, Stocks, Sports)
- E911: (Wireless carriers provide wireless callers' numbers and locations.)

Generally, LBS services can be classified into three general categories: telematics LBS, Internet LBS, and wireless LBS (Telc).

Telematics LBS is the integration of wireless communications, vehicle monitoring systems, and location devices. Telematics LBS applications include automated vehicle location, fleet tracking, online navigation, and emergency assistance. For example, a trucking company can track all their fleet, proactively warn about traffic ahead, and estimate the arrival time. Commercial LBS providers are beginning to offer important management applications that help direct vehicle fleets and ensure optimal usage of key assets. Telematics LBS is a multibillion dollar service industry and is currently the largest segment of the LBS market (Telc).

Internet LBS provide Internet users the services relevant to their specified locations. Because they use a user-specified location instead of the user's current location, no positioning technology is required. For example, one can find turn-by-turn driving direction from one location to another and search for tour information about the destination. These services are targeting applications with stationary users, relatively powerful computers, and reliable network connections. As a result, Internet LBS support sophisticated services, such as local business searching and comparison, trip planning, online virtual tours, and so forth.

Wireless LBS deliver location relevant content to cell phones, PDAs, and other wireless devices. Equipped with automated positioning technologies, MUs can query local weather, nearby traffic information, and local businesses close to them. For example, a user can search neighboring post office or coffer shop from the PDA. The wireless LBS market is currently in a nascent stage, but it will potentially become the largest segment of the LBS market. The deployment of third generation (3G) mobile network, which support handsets that are both mobile and location sensitive, will lead to more wireless LBS subscribers and more useful LBS applications.

This article focuses on the discussion on wireless LBS system, and the term LBS refers to wireless LBS in the rest of this article. It compares LBS and traditional database system, introduces existing LBS systems, and reviews the related research works. Next, it describes a representative LBS system model and explains the functionality of the LBS

system. It introduces the major components, their roles, and interactions. The discussion also covers issues related to mobile devices, positioning technologies, spatial databases, location aware queries, and so forth. In particular, this article will provide a detailed review on location dependent query processing and caching. Issues such as query processing algorithms, validity region, and query result caching are discussed. Then, it foresees the new service demands, emerging applications, and trends in future LBS systems. Finally, the article provides a summary on the above discussion and concludes this article.

BACKGROUND

Compared to traditional database (DB) services, new characteristics of LBS lead to significant differences between LBS databases and traditional databases. A database in LBS is a ***spatial database*** (SDBS) (Guting, 1994), which is capable of representing, querying, and manipulating spatial data (such as point, line, and region) to efficiently process queries with spatial restrictions and support applications such as the ***geographical information system*** (GIS). An SDBS is required to handle continuously changing data, locations of moving objects, and provide location aware services to mobile users. LBS also face other research challenges (Jensen et al., 2001) in order to support the following features: nonstandard dimension hierarchies in database; imprecision and varying precision; movement constraints and transportation networks; multiresolution objects and maps in data modeling; spatial data mining on vehicle movement; and continuous location change in query processing techniques. Interested readers are referred to Jensen et al. (2001) for more details.

LBS have introduced two types of queries with location constraints. The ***location aware query*** (LAQ) is the query with certain location constraints (Seydim, Dunham, & Kumar, 2001). As a special type of LAQ, the ***location dependent query*** (LDQ) (Barbara, 1999) is the query whose result depends on querying location, that is, the mobile user's current location. For example, "Phone numbers of all McDonald's in New York City" is an LAQ, while "Phone number of the nearest McDonald's" is an LDQ. LDQ is one of the core functions of LBS. Two common types of LDQs are the ***nearest neighbor*** (NN) query, that retrieves the qualifying database object closest to the querying position, and the ***window query*** that retrieves all satisfying database objects within an axis-parallel rectangle centered at the querying position.

LDQ processing and result caching have new characteristics not observed in traditional database systems. An LDQ may have different results in different region called ***validity region*** (VR). LDQs can be answered by the cached result of the same LDQ, if the MU remains within the cached result's VR. There are several algorithms for the DB server to determine the VR for NN and window query result sets (Zheng, Lee, & Lee, 2004; Zheng & Lee, 2001). To improve the performance, in certain applications, the limited validity region for LDQ result sets needs also be considered in LDQ caching. The following section describes the LBS system model and important research issues including positioning technology, LDQ processing, and LDQ caching.

LBS SYSTEM

LBS System Model

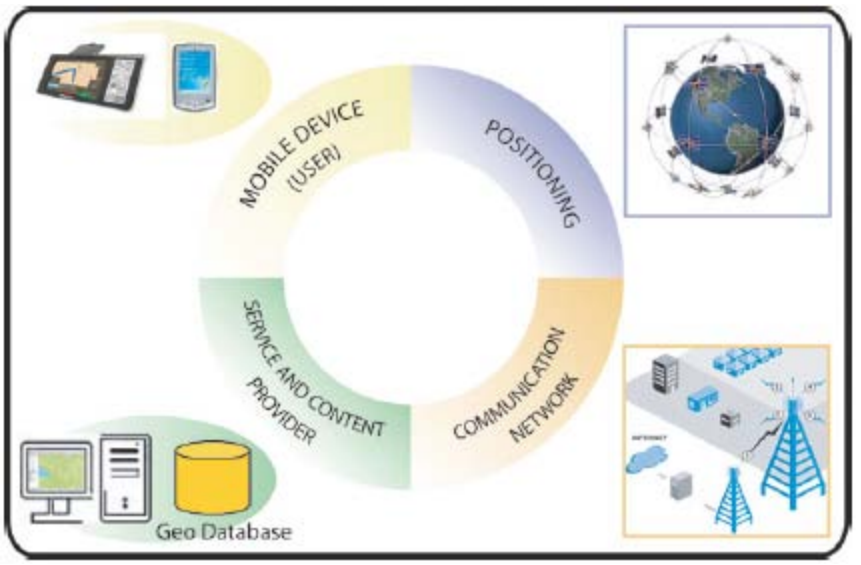
A typical LBS system consists of four components (Steiniger, Neun, & Edwardes, 2006) as shown in Figure 1:

- **Mobile devices:** LBS users request services and receive data using mobile devices, which can be personal digital assistants (PDA), mobile phones, laptops, and vehicle-mounted devices.
- **Communication network:** The communication network can be a wireless cellular network, wireless LAN, or other type of wireless network. It transfers the users' data and service requests to the service provider and forwards the requested information back to the users.
- **Positioning component:** User's location is an essential part of the LDQ, and it can be a symbolic entity (e.g., a street address) entered by the user or a geometric entity (e.g., the latitude and longitude coordinates) automatically acquired using positioning mechanisms. The user position can be obtained either by using the mobile communication network or by using the devices equipped with Global Positioning System (GPS). Further possibilities to determine the indoor position are active badges and radio beacons.
- **Service and content provider:** The service provider offers a number of different services to the user and is responsible for the service request processing. The requested data is usually stored and maintained at separated databases.

Positioning, Querying, and Caching Issues

- **Positioning technologies:** LBS require user's location, which can be input by the user manually or acquired by the device or network automatically. A user can input his/her street address or natural area code, which represents a location using alphanumeric characters code that is shorter than the latitude/longitude equivalent. Alternatively, one acquires a user's location via device-based techniques and network-based techniques.

Figure 1. LBS system components



The premium example of device-based techniques is the GPS and the Assisted GPS (A-GPS). GPS is the worldwide satellite-based radio navigation system consisting of 24 satellites launched by the U.S. Department of Defense. The mobile device equipped with a GPS receiver locates itself by comparing signals from four satellites. A GPS system has a high accuracy, ranging from 5 to 30 meters. In A-GPS, the mobile network or a third party service provider can assist the mobile device to achieve a very high accuracy, between 1 to 10 meters. A similar system is the GLObal NAvigation Satellite System (GLONASS) system comprised of 24 satellites launched by Russia. In 2005, the European Union launched its Galileo navigating system, which consists of 30 satellites and has a higher accuracy than GPS.

Due to the cost and power constraints of a GPS receiver, most handsets in cellular network are not equipped with GPS receivers, and network-based techniques are widely used in cellular networks to locate wireless subscribers. The commonly used network-based positioning techniques include Cell ID, angle of arrival (AOA), time of arrival (TOA), time difference of arrival (TDOA), and enhanced observed time difference (EOTD). Cell-ID, the simplest location approach in cellular networks, uses the serving base station (BS) to approximate the user's location. The accuracy is low and potential deviation depends on the radius of service cell, which is normally between 200 and 3000 meters. Other techniques use triangulation technology: finding the user's location by measuring at least three different signals between the user and fixed servers. Using AOA, the position can be determined if the user's signal is received by at least three BSs with additional electronics to detect the compass direction from which the signal arrives. The BSs send this compass data (i.e., angles) to a mobile switch to calculate the geographical location of the MU with the accuracy in the range of 100 to 200 meters. If the user's signal is received by at least three BSs with a synchronized atomic clock, TOA can compare the user's signal to compute the user's geographical location. The accuracy is between 100 and 200 meters. Zeimpekis, Giaglis, and Lekakos (2003) provides an in-depth analysis and evaluation of the commonly used indoor and outdoor wireless positioning technologies.

- **LDQ processing:** The two most common types of LDQs are nearest neighbor (NN) queries and window queries that have been intensively studied in the literature. A more general form of NN query *is* *k*-NN query that returns the *k* nearest neighbors to the querying location. The *k*-NN queries can be classified into six major categories: simple *k*-NN queries, approximate *k*-NN queries, reverse *k*-NN queries, constrained *k*-NN queries, *k*-join queries, and continuous *k*-NN queries. The most representative algorithm for NN query processing is a branch-and-bound R-tree traversal algorithm (Roussopoulos, Kelley, & Vincent, 1995). The algorithm searches R-tree in a depth first man-

ner, starting from the root node. It records the nearest neighbor found so far and the distance. This algorithm skips the nodes if the distance to node is farther than the shortest distance found so far. For the *k*-NN query processing, the procedure is similar expect maintaining *k* nearest neighbor and their distances to the querying location.

Window query retrieves all objects in an axis-parallel rectangle. This characteristic makes R-tree very attractive in efficient window query processing. A typical algorithm answers the window query in R-tree. The algorithm first retrieves the root node and compares the entries of its children nodes with the querying window. Nonintersecting entries will not contain qualifying points, and are therefore skipped. The searching algorithm recursively retrieves those entries intersecting with the querying window. When it reaches the leaf level, it will return the qualifying data object(s), if any.

- **LDQ caching:** Validity region is the essential information for LDQ caching schemes. Much research has been done in algorithms to generate LDQ result VR; and there are several algorithms for the DB server to determine the VR for an NN query result. Zheng and Lee (2001) built a static *Voronoi diagram* (VD) to index the objects in SDBS by partitioning the plane into Voronoi cells (VC), one for each object. The VC is a convex polygon that consists of the points closer to the object in this VC than to any other objects. As a result, VD is the most suitable mechanism to find the NN and the corresponding VR: the object in the same VC with the querying locations is the NN result, and the VC is the corresponding nearest neighbor validity region (NNVR). It is, however, expensive to maintain VD due to DB updates, and it is also inapplicable for the *k* nearest neighbor (*k*-NN) query when *k* is unknown. Even when *k* is known, an order-*k* VD is very expensive in terms of computational and storage overhead (Zhang, Zhu, Papadias, Tao, & Lee, 2003). Consequently, Zhang et al. (2003) introduced algorithms to calculate NNVRs during run time that avoids the large storage overhead at the expense of extra computing overhead.

A window VR is therefore the region within which the result set remains unchanged. The Minkowski region (MR) is introduced to examine the VR of a window query result set, which may contain zero, one, or multiple results. The MR of an object is an axis-parallel rectangle identical to the query window whose geometric center is the corresponding object. If the querying position is within the MR of an object *a*, then object *a* will be a result. Otherwise, object *a* is not in the window query result set. Thus, the VR of a window query result set is the area that is within the MRs of all result objects, and outside the MRs of all other objects.

LDQ caching has attracted much research attention. Ren and Dunham (2000) proposed a semantic caching scheme for location dependent results. An incoming query is decomposed into two disjointed parts: a probe query that can be answered by the cached data and a remainder query that has to be answered by the DB server. This scheme reduces the network traffic, allows query resolution during the disconnection, and in some cases allows partial query resolution. Zheng, Xu, and Lee (2002) presented algorithms for cache invalidation and cache replacement strategies that takes the validity of information into account. Zheng, Lee, and Lee (2004) presented a semantic NN caching scheme and addressed mobile user's intercell roam issues. The aforementioned caching schemes rely on DB servers to provide VRs for the LDQ results. Considering that DB servers may not always provide VRs for LDQ results due to the computation and storage overhead, Gao and Hurson (2005) and Gao, Sustersic, and Hurson (2006) suggested an LDQ proxy caching scheme where the proxy can calculate the estimated validity regions (EVR) for the LDQ result based on the querying history and the cache contents.

FUTURE TRENDS

In the next few years, LBS growth will continue with the following trends (Desiniotis, Markoulidakis, & Gaillet, 2006):

- **Market value:** Several reports published around 2000 claimed that LBS would become the most promising "killer applications," with billions of revenue in a few years. Unfortunately, those predictions were not met. A recent Juniper Research report estimates the total available market for mobile LBS will reach $8.5 billion by end of 2010 (Juniper, 2005) considering the fast growth of markets in Asia Pacific region.
- **Mobile devices:** The mobile devices will have more functions, better connection, and improved usability.
- **Vehicle navigation devices:** One major pushing power for the growth of LBS market is the increasing popularity of portable navigation devices such Garmin and Tom-Tom. These devices will be less expensive and equipped with more functions such as music player, hands-free phone kit, a Web browser, and so forth.
- **Improved accuracy:** Accuracy and compatibility of LBS are going to be critical competitive factors for carriers to compete for customers in the future. Most LBS carries currently using GPS system will move toward A-GPS that offers greater accuracy levels.

The research topics in LBS system will continue in LDQ query processing and caching. At the same time, more attention might be given to emerging topics or concerns. Vehicular users desire to exchange information between neighboring vehicles; thus, short range communication protocols need to be proposed for various applications. Security is always a concern and deserves continuous study in the future LBS system. LBS system tends to collect users' information and statistics to improve system performance, which raises the research topics on user privacy.

CONCLUSION

LBS have attracted much attention from both researcher and service providers in the past a few years. This article reviewed different types LBS services and described its system components. It also discusses several important research issues and the future trend of LBS.

REFERENCES

Barbara, D. (1999). Mobile computing and databases—a survey. *Knowledge and Data Engineering*, 108-117.

Desiniotis, C, Markoulidakis, J. G., & Gaillet, J.-F. (2006). Mobile LBS market. In *Proceedings of the LIAISON-ISHTAR Workshop*.

Gao, X., & Hurson, A. R. (2005). Location dependent query proxy. In *Proceedings of the ACM Symposium of Applied Computing*, (pp. 1020-1024).

Gao, X., Sustersic, J., & Hurson, A. R. (2006). Window query processing with adaptive proxy cache. *Mobile Data Management*.

Guting, R. (1994). An introduction to spatial database systems. *Very Large Databases Journal*, *3*(4), 357-399.

Imielinski, T., & Badrinath, B. (1992). Querying in highly mobile distributed environments. In *Proceedings of the Very Large Databases Conference*, (pp. 41-52).

Jensen, C.S., Friis-Christensen, A., Pedersen, T.B., Pfoser, D., Saltenis, S., & Tryfona, N. (2001). Location-based services—a database perspective. In *Proceedings of the Scandinavian Research Conference on Geographical Information Science*, (pp. 59-68).

Juniper Research. (2005). Mobile location based services: Information services, tracking, navigation, community & entertainment.

Lee, D.L., Lee, W.-C., Xu, J., & Zheng, B. (2002). Data management in location-dependent information services: Challenges and issues. *IEEE Pervasive Computing*, *1*, 65-72.

Ren, Q., & Dunham, M. (2000). Using semantic caching to manage location dependent data in mobile computing. *MobiCom*, 210-221.

Roussopoulos, N., Kelley, S., & Vincent, F. (1995). Nearest neighbor queries. *Management of Data*, 71-79.

Seydim, A., Dunham, M., & Kumar, V. (2001). Location dependent query processing. *MobiDE*, 47-53.

Steiniger, S., Neun, M., & Edwardes, A. (2006). *Foundations of location based services.* CartouCHe project lecture notes. Retrieved December 12, 2007, from www.geo.unizh.ch/publications/cartouche/lbs_lecturenotes_steinigeretal2006.pdf

Telc. Retrieved December 12, 2007, from http://www.telcontar.com/company/marketplace.html

Zeimpekis, V., Giaglis, G., & Lekakos, G. (2003). A taxonomy of indoor and outdoor positioning techniques for mobile location services. *SIGECOM Exchanges*, 19-27.

Zhang, J., Zhu, M., Papadias, D., Tao, Y., & Lee, D. (2003). Location-based spatial queries. *Management of Data*, 443-454.

Zheng, B., & Lee, D. (2001). Semantic caching in location-dependent query processing. In *Proceedings of the Symposium on Spatial and Temporal Databases*, (pp. 97-116).

Zheng, B., Lee, W.C., & Lee, D. (2004). On semantic caching and query scheduling for mobile nearest neighbor search. *Wireless Networks Journal*, *10*(6).

Zheng, B., Xu, J., & Lee. D. (2002). Cache invalidation and replacement strategies for location-dependent data in mobile environments. *Transaction on Computers, Special Issue on Database Management and Mobile Computing*, *51*(10), 1141-1153.

KEY TERMS

Geographic Information Systems: A computer system designed for storing, manipulating, analyzing, and displaying data in a geographic context.

Location Aware Query: is the query with certain location constraints.

Location-Based Services: Personalized services based on certain location constraints, normally the user's current location.

Location Dependent Query: The query whose result depends on the users' current location.

Nearest Neighbor: A query that returns the nearest objects to the user.

Spatial Database: A database system that offers spatial data types in its data model and query language and supports spatial data types in its implementation.

Validity Region: A regions where a location dependent query result remains valid.

Window Query: A query that returns all satisfying database objects within an axis-parallel rectangle centered at the querying position.

Machine Learning

João Gama
University of Porto, Portugal

André C P L F de Carvalho
University of São Paulo, Brazil

INTRODUCTION

Machine learning techniques have been successfully applied to several real world problems in areas as diverse as image analysis, Semantic Web, bioinformatics, text processing, natural language processing,telecommunications, finance, medical diagnosis, and so forth.

A particular application where machine learning plays a key role is data mining, where machine learning techniques have been extensively used for the extraction of association, clustering, prediction, diagnosis, and regression models.

This text presents our personal view of the main aspects, major tasks, frequently used algorithms, current research, and future directions of machine learning research. For such, it is organized as follows: Background information concerning machine learning is presented in the second section. The third section discusses different definitions for Machine Learning. Common tasks faced by Machine Learning Systems are described in the fourth section. Popular Machine Learning algorithms and the importance of the loss function are commented on in the fifth section. The sixth and seventh sections present the current trends and future research directions, respectively.

BACKGROUND

Machine learning can be seen as a subfield of artificial intelligence (Bratko, 1984) and is influenced by works on statistics (inference and pattern recognition [Duda & Hart, 1973; Fukunaga, 1990]), databases (analytical and multivariate databases [Berson & Smith, 1997]).

Machine learning is strongly linked to search, optimization, and statistics. Several models present optimization mechanisms, like support vector machines. Others are based on statistics inference, for instance, Bayesian classifiers.

Machine learning models have been extensively used in data mining. Data mining is concerned with the discovery of useful information in large databases. Very often, the observations need to be collected, selected, and preprocessed before machine learning techniques can be employed. It is important to mention that data mining relies not only on machine learning, but also on statistics, artificial intelligence, databases, and pattern recognition.

WHAT IS MACHINE LEARNING?

Informally speaking, the main goal of machine learning is to build a computational model from past experience of what has been observed. For such, machine learning studies the automated acquisition of domain knowledge looking for the improvement of systems performance as result of experience.

In the beginning of the 1980s, Michaslky, Carbonell, and Mitchell (1983) presented one of the first definitions of machine learning "Self-constructing or self-modifying representations of what is being experienced for possible future use" (p. 10).

The focus of this definition is on programs that modify themselves in response to feedback from their environment. This definition reflects the main research lines at that time: expert systems (Weiss & Kulikowski, 1991), automatic programming, and reinforcement learning (Sutton, 1998).

A more recent definition appears in (Hand, Mannila, & Smyth, 2001) "Analysis of observational data to find unsuspected relationships and to summarize the data in novel ways that is both understandable and useful for the data owner" (p. 1).

An even more recent definition is due to (Alpaydin, 2004), where machine learning is defined as "Programming computers to optimize a performance criterion using example data or past experience" (p. 3).

Clearly, the task here is much closer to a data analysis task, enlarging the range of practical applications, mainly industrial and commercial, where machine learning is frequently employed. In any case we can define machine learning as the acquisition of a useful (understandable) representation of a data set from its extensional representation.

MACHINE LEARNING TASKS

In a basic learning task, observations take the form of pairs. $\{\vec{x}, y\}$ The elements of the vector $\vec{x}$ are named *independent*

variables or *attributes* and the dependent variable $y=f(\vec{x})$ is an unknown function. The learning task is to obtain a predictive model or an approximation function $\hat{f}$ able to predict ŷ for future observations of the independent variables $\vec{x}$ (Mitchell, 1997). In this framework, we can consider two different problems: *classification problems,* whenever *y* takes values in a finite set of unordered values (e.g., $y \in \{C_1,...,C_n\}$), and *regression problems,* when *y* takes values in a subset of *R*. In machine learning theory, these observations are assumed to be independent and generated at random, according to a stationary probability distribution. This task is referred as *predictive or supervised learning*, because the value of the target variable *y* in the observations or training set is known.[1] When there is no clear target variable in the training set, we have an unsupervised learning task.

Several areas of human activity can involve *supervised machine learning*: predicting the use of land based on satellite images; assigning credit to individuals on the basis of financial information; sorting letters on the basis of machine readable post codes; preliminary diagnosis of a patient's disease; and so forth. Several problems in industry, commerce, and science are *decision problems* and require the analysis of complex and extensive data. Most of these problems can be analyzed from a supervized machine learning perspective (Witten & Frank, 2005).

In contrast to predictive learning, *descriptive or unsupervised learning models* provide compact representations for the whole data or for the process generating the data. Examples of such descriptions include models for density estimation, clusters analysis and segmentation, and models describing relations between variables. This task includes data synopsis or signatures (Arasu & Manku, 2004), data visualization, and cluster analysis (Berthold & Hand, 1999; Duda, Hart, & Stork, 2001).

MACHINE LEARNING ALGORITHMS

We can formalize a machine learning problem as either a parameter optimization problem or a hypothesis search problem. In the former, examples are points in a multi-dimensional space associated with a metric. The goal is to minimize a loss function. Illustrative examples following this approach are *k-nearest neighbor* (Aha, 1997), discriminant functions (Duda et al., 2001), and neural networks (Ripley, 1996). In the later, a language used to represent generalizations of examples defines a search space of possible hypothesis. The learning algorithm performs a search in this space. The goal is to find the best hypothesis, a state in the search space that maximizes some objective function. Illustrative examples of this approach are decision trees (Breiman, Friedman, Olshen, & Stone, 1984; Quinlan, 1993), decision rules (Quinlan, 1993), and Bayesian networks (Neapolitan, 2003).

Each approach employs a different representation schema and explores different search strategies. A representation is the decision structure of a certain type (i.e., a decision tree, a set of discriminant equations, a table of conditional probabilities, etc.) used to generalize the examples. The representation used by a learning algorithm restricts the set of hypotheses considered by the algorithm. Some authors call it the *restricted hypotheses space bias* (Mitchell, 1990).

A *search strategy* is the set of methods and heuristics used to explore the search space defined by the set of possible representations. Associated with each search strategy is the *evaluation component.* This component is used to guide the search, by either preferring one hypothesis over others or by ranking the set of possible hypotheses. Both the search strategy and the evaluation component have preferences on the possible set of hypotheses. Such preference is also known in the machine learning literature as *preference bias*, because it imposes a certain preference order on the elements of the hypotheses space. A commonly used heuristic is *to prefer general hypotheses over specialized ones*. For example, if we consider decision trees, this corresponds to the preference of small trees to larger ones.[2] This kind of preference that minimizes the syntactic complexity of the hypotheses representation reduces the chance of the model overfitting the training observations. Overfitting occurs when the model is over-adjusted to the training data. Overfitting decreases the model generalization, that is, its capacity to correctly classify new observations.

There is a strong relation between overfitting and the *Occam's razor* (Blumer, Ehrenfeucht, Haussler, & Warmuth, 1990), which states, "The simpler of two competing hypotheses should always be preferred."

One argument in favor of simplicity is that there are fewer simple hypotheses than complex ones (based on combinatorial arguments). As such, if both fit the data, we should prefer the simpler hypothesis because it is less likely to be a statistical coincidence (Mitchell, 1997). Domingos' (1998a) work presents a thorough discussion about the interpretation of Occam's razor in the context of machine learning. Domingos concludes "if a model with low training-set error is found within a sufficiently small set of models, it is likely to also have low generalization error" (p. 37). Therefore a fully adequate model evaluation is only possible if the search process by which the models are obtained is also taken into account (Domingos, 1998b).

Another general heuristic claim is, "Examples that are near in the instance space correspond to similar concepts." This heuristic is fully exploited in the so-called *instance-based learning*, also known as *lazy learning,* where learning consists in memorizing previous observations, as in k-nearest neighbor and *case-based reasoning* (Aamodt & Plaza, 1994). The term *lazy learning* is used because instead of estimating the target function once for the entire instance space, these algorithms delay learning until they need to output a pre-

diction. The approximate function is estimated locally and differently for each query example. Further developments of lazy learning lead to *local weighted regression* (Moore, Schneider, & Deng, 1997), *radial basis functions* (Powell, 1992), and so forth.

Other machine learning models extract general hypotheses from a set of previous observations, process named employ Inductive Learning. These models can obtain an arbitrarily good fit to the data by considering very complex hypotheses.

Some of these models, like decision trees, are symbolic, knowledge-based models. Decision trees approximate discrete functions by building a tree where each node tests the value of a particular attribute. Branches are associated to the possible results of this test (Mitchell, 1997). The hypothesis space of decision trees is within the disjunctive normal form (DNF) formalism. The conditions along a branch represent conjuncts, and the individual branches can be seen as disjuncts. Each branch forms a rule with a conditional part and a conclusion. The conditional part is a conjunction of conditions. There are clear relations between decision tree learning and rule-based learning algorithms (Wnek & Michalski, 1994). Both use the same representation language, and use a hill-climbing algorithm. While most of decision tree learning algorithms use a top-down, general-to-specific strategy, in rule learning both general-to-specific and specific-to-general (bottom up) are possible.

In order to avoid overfitting, they require a technique to balance the complexity of the hypothesis against the number of training examples misclassified by the hypothesis. One such technique is the *minimum description length* (Rissanen, 1983) which states, "The best hypothesis is the one that minimizes the total length of the hypothesis plus the description of the exceptions to the hypothesis" (p. 420).

The intuition is that the ad hoc hypothesis consists of a list of examples that make them no shorter than the examples. In contrast, good hypotheses reduce many examples to a single, general rule.

Other models have strong roots in statistics, particularly neural networks, support vector machines, hidden Markov models (HMM), and Bayesian networks. It must be observed that both machine learning and statistics infer hypothesis from observations. But unlike statistics, machine learning is also concerned with the analysis of the computational complexity of the algorithms employed.

Neural networks learning methods are based on the structure and functionalities of the nervous system (Haykin, 1998). Several different models have been proposed. The most popular being *multi-layer perception* is based on layers of processing units computing simple mathematical functions. These units are linked by weighted connections whose values are adjusted by the back propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). This algorithm overcomes problems pointed out by Minsky and Papert (1969).

Support vector machines are based on the *statistical learning theory* (Vapnik, 1995). Binary classifiers show high generalization capability by looking for a hyperplane that maximizes the separation margin between observations from different classes. The use of kernels allows their use for nonlinear problems.

HMM are probabilistic models, stochastic finite state automata, used for making a sequence of decisions [Rabiner, 1989]. They are and have been frequently applied to time series and linear sequences. They model a temporal problem by a set of discrete states and transition probabilities between the states. Thus, an HMM is a model that generates sequences. They are frequently used in speech recognition and prediction of protein structure.

Bayesian networks (Neapolitan, 2003) are probabilistic models for knowledge representation under uncertainty. They represent the joint probability distribution of a set of variables. Bayesian networks consist of a graphical model representing the set of independent assumptions and a set of probabilistic conditional tables representing $P(x \mid Parents(x))$. Graphical models can be used in a set of learning tasks, like prediction, diagnosis, and clustering.

Each machine learning algorithm can be characterized by the representation used for the acquired knowledge and the approach employed for hypothesis search. Both *representation* and *search* define a unique *learning bias* for each algorithm (Merz, 1998). A key issue in machine learning is the interdependence between representation and learning. In the next section we present some machine learning techniques used to extend the representational ability of systems.

Loss Functions

A requirement in any learning problem is to identify what we would like to minimize or approximate. We need to define a criterion to measure the quality of any estimator $\hat{f}$. In the machine learning literature, the criteria frequently considered include error rate, comprehensibility, compactness, learning speed, classification speed, and so forth.

In predictive learning, the most relevant dimension is the *error rate*. It is an estimator of the difference between $\hat{f}$ and the unknown *f*. We need to distinguish between two notions of error rate. One, denoted as *resubstitution error rate*, estimates the error rate from the training set. The other, the *generalization error rate*, estimates the error from an independent test set. What we usually wish to know is the generalization error rate, because this is the error that can be expected when applying the model to future examples. The error rate is computed as the ratio between the number of misclassifications and the total number of examples. In a most general scenario we are interested in measuring the loss incurred by a model, $\hat{f}$ given an example $\vec{x}$. The loss

measures the difference between $\hat{f}(\vec{x})$ and the true label $f(\vec{x})$. In the classification setting, the 0-1 loss function is the most commonly used. The loss is 0 if $\hat{f}(\vec{x}) = f(\vec{x})$ and 1 otherwise. In the regression setting, the *squared loss* defined as $(f(\vec{x}) - \hat{f}(\vec{x})^2$ is commonly used. The 0-1 loss function assumes equal loss for all misclassifications. To minimize this problem, a promising research line is the *receiver operational characteristic* (Provost, Fawcett, & Kohavi, 1998). The squared loss has interesting properties: It is derivable and decomposable into bias plus variance. Similar decompositions appear recently for 0-1 loss (Kohavi & Wolpert, 1996). The desirable characteristics of loss functions include: fast computation, derivable, decomposable, and must be convex in order to ensure uniqueness of the solution.

In 1995, Vapnik proposed another minimization criterion: the *risk minimization principle*. Algorithms like support vector machines, instead of minimizing a function of the error, minimize the margin, which is defined as the distance between the nearest examples (the support vectors) and the decision surface. This criterion has been shown to be very powerful in minimizing the generalization error (Duda et al., 2001; Hastie, Tibshirani, & Friedman, 2000).

CURRENT TRENDS IN MACHINE LEARNING

A main issue in current machine learning research is the model selection problem. This issue can be summarized as: For a given dataset what is the most suitable model? When the learning problem is defined by a fixed set of examples the most common strategy consists of estimating the error rate employing a pool of learning algorithms using some sampling method. The most common used sampling strategies are *cross validation*, *leave-one-out,* and *holdout*. In any case, we must guarantee that the error estimate is unbiased, meaning that it should be obtained as the error in a set of examples independent from those used to train the algorithm.

Any learning algorithm has its own area of applicability: There are problems where an algorithm exhibits good generalization performance and other problems without generalization capacity.

A promising research area is metalearning (Brazdil, Gama, & Henery, 1994), which try to characterize the applicability of learning algorithms using data characteristics, like statistical measures (e.g., class distribution, correlation between attributes, etc.) and data information based measures (mean entropy of attributes, noise-signal ratio, etc.). Characterizing the area of applicability of learning algorithms is useful for model selection in similar problems. Most of the works in metalearning use cross validation and statistical hypothesis testing to evaluate the significance of differences. Nevertheless, several authors refer that there is a trade-off between the systematic errors due to the representational language used by an algorithm (the *bias*) and the *variance* due to the dependence of the model to the training set (Breiman, 1998). When using more flexible representations the bias is reduced. Model selection using cross validation does not provide the complete picture. Whenever comparing algorithms with different degrees of complexity the bias-variance decomposition of the error rate could provide useful information.

Another approach to the model selection problem are the *ensemble models* (Kittler, 1998; Kuncheva, 2004). Instead of using a single learning algorithm, we should train a large set of classifiers, whose predictions are averaged to obtain a final prediction for a test example. Averaging over a large set of models has clear impact in the variance of individual models. This is the case of *bagging*, a successful technique when used in conjunction with decision trees (Breiman, 1998). Another popular ensemble technique able to reduce the bias and the variance components of the error is *boosting* (Schapire, Freund, Bartlett, & Lee, 1997).

Until now, we have assumed an attribute-value representation for data. Each example is described by a set of variables (a record in database terminology) and a set of examples is stored in a matrix (a relation or table in database terminology). Some of the most challenging applications of machine learning are: data are better described by sequences (i.e., DNA data), trees (XTML documents), and graphs (chemical components, network analysis). Inductive logic programming (Muggleton, 1992, 1998) goes behind the propositional representation of data by learning from relations and generating decision models in the form of logic programs. A very good review of the state of the art of graph-based mining appears in (Washio & Motoda, 2003). Related areas involve learning from scientific data (Bock & Diday, 2000), semi-structured text, or relational databases (Berson & Smith, 1997).

Most of the machine learning research in the past focuses on improving learning algorithms either by using more powerful representation languages or by improving search. Most of the algorithms were memory based and required all training data to fit in the memory. However, the exponential increase of data stored in large databases, usually collected over time during months or years, poses new algorithmic problems. Moreover, in some situations like sensor networks, TCPIP traffic, e-commerce, and Web sites, data flow at high speed. It is now more possible to store all data in memory and perform multiple scan over training data. To increase the difficulty of the learning problem the target concept can change over time. These observations lead to the development of research on scaling-up learning algorithms (Provost & Hennessy, 1996), sequential learning from data streams and dynamic environments (Guha, Meyerson, Mishra, Motwani, & O'Callaghan, 2003), and very fast machine

learning (Domingos & Hulten, 2000; Hulten, Spencer, & Domingos, 2001).

This and other related issues related to the design and analysis of machine learning models are the main focus of a research area known as *computational learning theory* (Valiant, 1984). This area investigates aspects as model complexity, accuracy estimation, and sample complexity.

FUTURE TRENDS

The key issue in the first machine learning definition, "Self-constructing or self-modifying representations of what is being experienced for possible future use" (p. 10) is the word *self*: systems that modify themselves to better fit or accommodate to the environment. The future of machine learning points to autonomous systems that can incorporate domain knowledge; learn from distributed sources in nonstationary and dynamic environments; and can transfer knowledge between learning systems.

CONCLUSIONS

In this text, we presented our view of the main aspects of machine learning, a research area that is facing a fast and steady growth. As the results of this growth, new research lines and application opportunities are continually emerging.

Machine learning research is fastly growing in both academic and industry. In academia, machine learning research groups can be found in different areas, like computer science and engineering, electrical engineering, statistics, economics, and business. In industry, several companies have started using machine learning either in the development of their products or as part of them.

Research in machine learning has benefited several social and economic sectors, providing new approaches for medical diagnosis, fault detection, time series analysis, and environmental monitoring, among several others.

ACKNOWLEDGMENT

The authors would like to thank the project ALESII sponsored by Fundação Ciência e Tecnologia, POSI/EIA/55340/2004, and CNPq for the financial support.

REFERENCES

Aamodt, A., & Plaza, E. (1994). Case-based reasoning; foundational issues, methodological variations, and systematic approaches. *AI Communications, 7,* 39-59.

Aha, D. W. (1997). Editorial on lazy learning. *Artificial Intelligence Review, 11,* 7-10.

Alpaydin, E. (2004). *Introduction to machine learning.* Cambridge, MA: MIT Press.

Arasu, A., & Manku, G. S. (2004). Approximate counts and quantiles over sliding windows. *ACM Symposium on Principles of Database Systems (PODS)* (pp. 286-296). ACM Press.

Berson, A., & Smith, S. (1997). *Data warehousing, data mining and olap*. MacGraw-Hill.

Berthold, M., & Hand, D. (1999). *Intelligent data analysis—An introduction*. Springer Verlag.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1990). Occam's razor. In *Readings in machine learning*. Palo Alto, CA: Morgan Kaufmann.

Bock, H., & Diday, E. (2000). *Analysis of symbolic data: Exploratory methods for extracting statistical information from complex data*. Springer Verlag.

Bratko, I. (1984). *Prolog, programming for artificial intelligence*. Addison-Wesley.

Brazdil, P., Gama, J., & Henery, R. (1994). Characterizing the applicability of classification algorithms using meta level learning. *Machine Learning—ECML-94*. (LNAI 784). Springer Verlag.

Breiman, L. (1998). Arcing classifiers. *The Annals of Statistics*, *26*, 801-849.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth International Group.

Domingos, P. (1998a). Occam's two razors: The sharp and the blunt. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (pp. 37-43) AAAI Press.

Domingos, P. (1998b). A process-oriented heuristic for model selection. *Proceedings of the 15th International Conference -ICML'98* (pp. 127-135).

Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. *Proceedings of the Six International Conference on Knowledge Discovery and Data Mining* (pp. 71-80). AAAI Press.

Duda, R., & Hart, P. (1973). *Pattern classification and scene analysis*. New York: Wiley and Sons.

Duda, R., Hart, P., & Stork, D. (2001). *Pattern classification*. Wiley & Sons.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. Academic Press.

Guha, S., Meyerson, A., Mishra, N., Motwani, R., & O'Callaghan, L. (2003). Clustering data streams: Theory and practice. *IEEE Transactions on Knowledge and Data Engineering, 15,* 515-528.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2000). *The elements of statistical learning, data mining, inference and prediction*. Springer Verlag.

Haykin, S. (1998). *Neural networks: A comprehensive foundation* (2nd ed.). Prentice Hall.

Hulten, G., Spencer, L., & Domingos, P. (2001). Mining time-changing data streams. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining,* San Francisco, CA (pp. 97-106). ACM Press.

Kittler, J. (1998). Combining classifiers: A theoretical framework. *Pattern Analysis and Applications, 1*(1), 18-27.

Kohavi, R., & Wolpert, D. (1996). Bias plus variance decomposition for zero-one loss functions. *Machine Learning, Proceedings of the 13th International Conference* (pp. 275-283). Morgan Kaufmann.

Kuncheva, L. I. (2004). Combining pattern classifiers. In *Methods and algorithms.* Indianapolis, IN: John Wiley & Sons.

Merz, C. (1998). *Classification and regression by combining models*. Unpublished doctoral dissertation, University of California, Irvine.

Michalsky, R., Carbonell, T. J., & Mitchell, T. M. (1983). *Machine learning—An artificial intelligence approach*. TIOGA Publishing Co.

Minsky, M. L., & Papert, S. A. (1969). *Perceptrons.* Cambridge, MA: MIT Press.

Mitchell, T. (1990). Generalization as search. In *Readings in machine learning*. Los Altos, CA: Morgan Kaufmann.

Mitchell, T. (1997). *Machine learning*. MacGraw-Hill.

Moore, A., Schneider, J., & Deng, K. (1997). Efficient locally weighted polynomial regression predictions. In *Proceedings of the 14th International Conference on Machine Learning, ICML '97*, Banff, Alberta, Canada (pp. 236-244). Morgan Kaufmann.

Muggleton, S. (Ed.). (1992). *Inductive logic programming*. San Diego, CA: Academic Press.

Muggleton, S. (1998). Knowledge discovery in biological and chemical domains. In *Proceedings of the First Conference on Discovery Science,* Berlin, Germany (pp. 58-59). Berlin: Springer-Verlag.

Neapolitan, R. (2003). *Learning Bayesian networks*. Upper Sadle River, NJ: Prentice Hall.

Powell, M. J. D. (1992). Advances in numerical analysis. In *The theory of radial basis function approximation* (105-210). UK: Oxford University Press.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). Building the case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th International Conference on Machine Learning-ICML'98,* Madison, WI (pp. 445-453).

Provost, F. J., & Hennessy, D. N. (1996). Scaling up: Distributed machine learning with cooperation. *American Association for Artificial Intelligence and International Association for Artificial Intelligence AAAI/IAAI* (Vol. 1, pp. 107-112).

Quinlan, R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. In *Proceedings of the IEEE,* (77(2), pp. 257-286).

Ripley, B. D. (1996). *Pattern recognition and neural networks*. UK: Cambridge University Press.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, *11*(2), 416-431.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318-362). Cambridge, MA: MIT Press.

Schapire, R., Freund, Y., Bartlett, P., & Lee, S. (1997). Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proceedings of the 14th International Conference on Machine Learning, ICML '97,* Banff, Alberta, Canada (pp. 322-330). Morgan Kaufmann.

Sutton, R. (1998). *Reinforcement learning*. Cambridge, MA: MIT Press.

Valiant, L. G. (1984) A theory of the learnable. *Communications of the ACM, 27*(11), 1134-1142.

Vapnik, V. (1995). *The nature of statistical learning theory*. Berlin, Germany: Springer-Verlag.

Washio, T., & Motoda, H. (2003). State of the art of graph-based data mining. *SIGKDD Exploration*, *5*, 59-68.

Weiss, S., & Kulikowski, C. (1991). *Computer systems that learn: Classification and prediction methods from statistics, neural nets, machine learning and expert systems*. San Mateo, CA: Morgan Kaufmann Publishers Inc.

Witten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Mateo, CA: Morgan Kaufmann.

Wnek, J., & Michalski, R. S. (1994). Hypothesis-driven constructive induction in AQ17-HCI: A method and experiments. *Machine Learning, 14*(2), 139-168.

KEY TERMS

Artificial Intelligence: The area of information technology concerned with the automation of reasoning, learning, and perception.

Bayesian Networks: Probabilistic models for knowledge representation under uncertainty.

Data Mining: The process of extraction of useful information in large databases.

Decision Tree: A symbolic learning classifier that represents a discrete function by a decision tree.

Hidden Markov Models: Stochastic models capable of performing a sequence of decisions.

Inductive learning: A learning approach based on the induction of a general concept from a limited set of observations.

LazyLearning: A learning approach where the instances are memorized.

Machine Learning: The programming of computers to optimize a performance criterion using example data or past experience.

Neural Networks: Learning models based on the structure and processing of the nervous system.

Support Vector Machines: Large margin models based on statistical learning theory.

ENDNOTES

1 We should note that *classification* is known in statistics as *discriminant analysis* (Hastie et al., 2000).

2 In nested trees smaller trees correspond to more general ones.

Machine Learning Through Data Mining

Diego Liberati
Italian National Research Council, Italy

INTRODUCTION

In dealing with information it often turns out that one has to face a huge amount of data, often not completely homogeneous and often without an immediate grasp of an underlying simple structure. Many records, each one instantiating many variables, are usually collected with the help of various technologies.

Given the opportunity to have so many data not easy to correlate by the human reader, but probably hiding interesting properties, one of the typical goals one has in mind is to classify subjects on the basis of a hopefully reduced meaningful subset of the measured variables. The complexity of the problem makes it worthwhile to resort to automatic classification procedures.

Then, the question arises of reconstructing a synthetic mathematical model, capturing the most important relations between variables, in order to both discriminate classes of subjects and possibly also infer rules of behaviours that could help identify their habits.

Such interrelated aspects will be the focus of the present contribution. The data mining procedures that will be introduced in order to infer properties hidden in the data are in fact so powerful that care should be put in their capability to unveil regularities that the owner of the data would not want to let the processing tool discover, like for instance, in some cases the customer habits investigated via the usual smart card used in commerce with the apparent reward of discounting.

Four main general purpose approaches will be briefly discussed in the present article, underlying the cost effectiveness of each one.

In order to reduce the dimensionality of the problem, simplifying both the computation and the subsequent understanding of the solution, the critical issues of selecting the most salient variables must be addressed. This step may already be sensitive, pointing to the very core of the information to look at.

A very simple approach is to resort to cascading a divisive partitioning of data orthogonal to the principal directions (PDDP) (Boley, 1998) already proven to be successful in the context of analyzing micro-arrays data (Garatti, Bittanti, Liberati, & Maffezzoli, 2007).

A more sophisticated possible approach is to resort to a rule induction method, like the one described in Muselli and Liberati (2000). Such a strategy also offers the advantage to extract underlying rules, implying conjunctions or disjunctions between the identified salient variables. Thus, a first idea of their even nonlinear relations is provided as a first step to design a representative model, whose variables will be the selected ones. Such an approach has been shown (Muselli & Liberati, 2002) to be not less powerful over several benchmarks than the popular decision tree developed by Quinlan (1994). An alternative in this sense can be represented by Adaptive Bayesian networks (Yarmus, 2003) whose advantage is also to be available on a commercial wide spread data base tool like Oracle.

Dynamics may matter. A possible approach to blindly build a simple linear approximating model is thus to resort to piece-wise affine (PWA) identification (Ferrari-Trecate, Muselli, Liberati, & Morari, 2003).

The joint use of (some of) such four approaches briefly described in this article, starting from data without known priors about their relationships, will allow to reduce dimensionality without significant loss in information, then to infer logical relationships, and, finally, to identify a simple input-output model of the involved process that also could be used for controlling purposes, even those potentially sensitive to ethical and security issues.

BACKGROUND

The introduced tasks of selecting salient variables, identifying their relationships from data, and classifying possible intruders may be sequentially accomplished with various degrees of success in a variety of ways:

- Principal components order the variables from the most salient to the least one, but only under a linear framework.
- Partial least squares do allow to extend to nonlinear models, provided that one has prior information on the structure of the involved nonlinearity; in fact, the regression equation needs to be written before identifying its parameters.
- Clustering may operate even in an unsupervised way without the a priori correct classification of a training set (Boley, 1998).
- Neural networks are known to learn the embedded rules with the indirect possibility (Taha & Ghosh,

1999) to make rules explicit or to underline the salient variables.
- Decision trees (Quinlan, 1994) are a popular framework providing a satisfactory answer to the recalled needs.

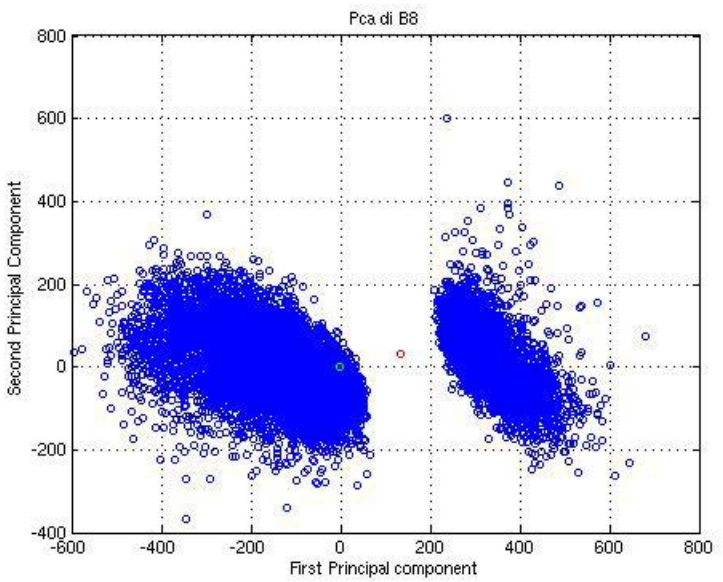
RECENT DEVELOPMENTS

Unsupervised Clustering

In this contribution, we will firstly resort to a quite recently developed unsupervised clustering approach, the Principal Direction Divisive Partitioning (PDDP) algorithm, proposed in Boley (1998). According to the analysis provided in Savaresi and Boley (2004), PDDP is able to provide a significant improvement of the performances of a classical k-means approach (Hand, Mannila, & Smyth, 2001; MacQueen, 1967), when PDDP is used to initialize the k-means clustering procedure. The approach taken herein may be summarized in the following three steps, the second of which is the core of the method, while the first one constitutes a preprocessing phase useful to ease the following tasks, and the third one is a postprocessing step designed to focus back on the original variables.

1. A principal component analysis defines a hierarchy in the transformed orthogonal variables according the principal directions of the data set. Principal Component Analysis (O'Connel, 1974; Hand et al., 2001) is a multivariate analysis designed to select the linear combinations of variables with higher intersubject covariances. Such combinations are the most useful for classification. More precisely, it returns a new set of orthogonal coordinates of the data space, where such coordinates are ordered in decreasing order of intersubject covariance.
2. The unsupervised clustering is performed by cascading a noniterative technique, the PDDP, (Booley, 1998) based upon singular value decomposition (Golub & van Loan, 1996), and the iterative centroid-based divisive algorithm k-means (MacQueen, 1967). Such a cascade, with the clusters obtained via PDDP used to initialize k-means centroids, is shown to achieve best performances in terms of both quality of the partition and computational effort (Savaresi & Boley, 2004). The whole dataset is thus bisected into two clusters, with the objective of maximizing the distance between the two clusters and, at the same time, minimizing the distance among the data points lying in the same clusters. The classification is achieved without using a priori information on the user (unsupervised learning), thus automatically highlighting the user belonging to a (possibly unknown) user class.
3. By analyzing the obtained results, the number of variables needed for the clustering may be reduced by pruning all the original variables that are not needed in order to define the final partitioning hyperplane, so that the classification eventually is based on a few variables only.

Figure 1. Clustering according to principal components



Binary Rule Inference and Variable Selection While Mining Data Via Logical Networks

Recently, an approach has been suggested—Hamming clustering—that is related to the classical theory exploited in minimizing the size of electronic circuits, with the additional care taken to obtaining a final function able to generalize from the training dataset to the most likely framework describing the actual properties of the data. In fact, the Hamming metric tends to cluster samples whose code is less distant. This is likely to be natural, if variables are redundantly coded via thermometer (for numeric variables) or only-one (for logical variables) code (Muselli & Liberati, 2000). The approach followed by Hamming clustering in mining the available data to select the salient variables and to build the desired set of rules consists of the three following steps:

Step 1: A critical issue is the partition of a possibly continuous range in intervals. In this way, the training process does not require floating point computation, but only basic logic operations. This is one reason for the algorithm speed and for its insensitivity to precision.

Step 2: In literature, classical techniques of logical synthesis are specifically designed to obtain the simplest AND-OR expression able to satisfy all the available input-output pairs without an explicit attitude to generalize. To generalize and infer the underlying rules, at every iteration, Hamming clustering groups together, in a competitive way, binary strings having the same output and close to each other. A final pruning phase does simplify the resulting expression, further improving its generalization ability. Moreover, the minimization of the involved variables intrinsically excludes the redundant ones, thus enhancing the very salient variables for the investigated problem. The low (quadratic) computational cost allows managing quite large datasets.

Step 3: Each logical product directly provides an intelligible rule, synthesizing a relevant aspect of the searched underlying system that is believed to generate the available samples.

The Hamming clustering approach has the following remarkable properties:

- It is fast, exploiting (after the mentioned binary coding) just logical operations instead of floating point multiplications.
- It directly provides a logical understandable expression (Muselli & Liberati, 2002), which is the final synthesized function directly expressed as the OR of ANDs of the salient variables, possibly negated.

Adaptive Bayesian Networks Under Minimum Description Length

An alternative learning strategy that looks for a trade-off between a high predictive accuracy of the classifier and a low cardinality of the selected feature subset may be derived according to the central hypothesis that a good feature subset contains features that are highly correlated with the class to be predicted, yet uncorrelated with each other. Based on information theory, the Minimum Description Length (MDL) principle (Barron, Rissanen, & Yu, 1998) states that the best theory to infer from training data is the one that minimizes the length (i.e., the complexity) of the theory itself and the length of the data encoded with respect to it. In particular, MDL can be employed as a criterion to judge the quality of a classification model.

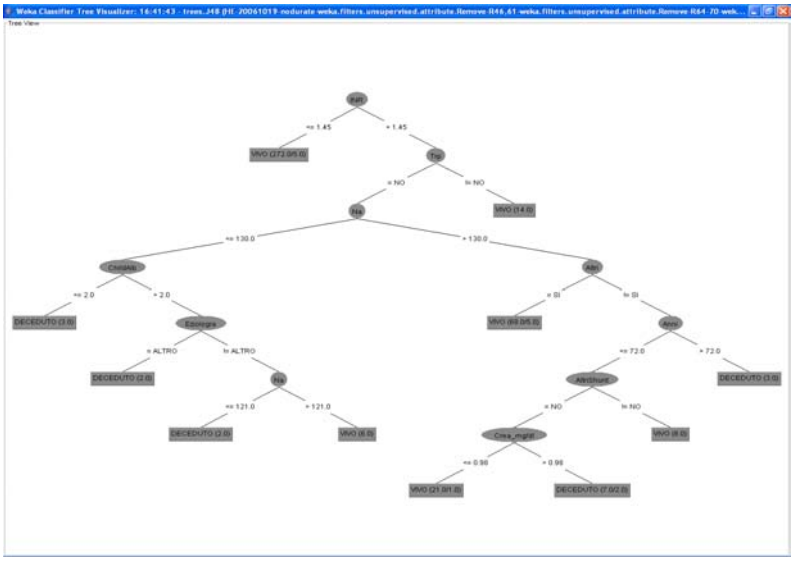
The motivation underlying the MDL method is to find a compact encoding of the training data. To this end, the MDL measure introduced in Friedman, Geiger, and Goldszmidt (1997) can be adopted, weighting how many bits one needs to encode the specific model (i.e., its length), and how many bits are needed to describe the data based on the probability distribution associated to the model.

This approach can be applied to address the problem of feature selection, by considering each feature as a simple predictive model of the target class. As described in Kononenko (1995), each feature can be ranked according to its description length, which reflects the strength of its correlation with the target. In this context, the MDL measure is given by Yarmus (2003), again weighting the encoding length, where one has one submodel for each value of the feature, with the number of bits needed to describe the data based on the probability distribution of the target value associated to each submodel.

However, once all features have been ordered by rank, now a priori criterion is available to choose the cut-off point beyond which features can be discarded. To circumvent this drawback, one can adopt a wrapper approach that starts with building a classifier on the set of the n-top ranked features. Then, a new feature is sequentially added to this set, and a new classifier is built, until no improvement in accuracy is achieved.

In the framework of Bayesian Networks, it is useful to step from Naïve Bayes (NB) to the Adaptive Bayesian Network (ABN), as recalled in the following. NB is a very simple Bayesian Network consisting of a special node (i.e., the target class) that is the parent of all other nodes (i.e., the features or attributes) that are assumed to be conditionally independent, given the value of the class. The NB network can be "quantified" against a training dataset of preclassified instances, that is, one can compute the probability associated to a specific value of each attribute, given the value of the class label. Then, any new instance can be easily classified

Figure 2. Decision tree obtained from Bayesian Networks. The rounded nodes of bifurcation are the salient variables, and the branches are labeled with their sets implying either outcome, represented by the squared leaves counting the corresponding events over the number of wrong attributions.



making use of the Bayes rule. Despite its strong independence assumption among variables is clearly unrealistic in most application domains, NB has been shown to be competitive with more complex state-of-the-art classifiers. (Cheng & Greiner, 1999; Friedman et al., 1997; Keogh & Pazzani, 2002).

In the last years, a lot of research has focused on improving NB classifiers by relaxing their full independence assumption. One of the most interesting approaches is based on the idea of adding correlation arcs between the attributes of a NB classifier. On these "augmenting arcs" are imposed specific structural constraints (Friedman et al., 1997; Keogh & Pazzani, 2002), in order to maintain computational simplicity on learning. The algorithm here proposed, the Adaptive Bayesian Network (Yarmus, 2003), is a greedy variant, based on MDL, of the approach proposed in Keogh and Pazzani (2002).

In brief, the steps needed to build an ABN classifier are the following. First, the attributes (predictors) are ranked according to their MDL importance. Then, the network is initialized to NB on the top k-ranked predictors, which are treated as conditionally independent. Next, the algorithm attempts to extend NB by constructing a set of tree-like multidimensional features.

Feature construction proceeds as follows. The top ranked predictor is stated as a seed feature, and the predictor that most improves feature predictive accuracy, if any, is added to the seed. Further predictors are added in such a way to form a tree structure, until the accuracy does not improve.

Using the next available top ranked predictor as a seed, the algorithm attempts to construct additional features in the same manner. The process is interrupted when the overall predictive accuracy cannot be further improved or after some preselected number of steps.

The resulting network structure consists of a set of conditionally independent multi-attribute features, and the target class probabilities are estimated by the product of feature probabilities. Interestingly, each multidimensional feature can be expressed in terms of a set of if-then rules enabling users to easily understand the basis of model predictions.

Piece-Wise Affine Identification Through a Clustering Technique

Once the salient variables have been selected, it may be of interest to capture a model of their dynamical interaction. A first hypothesis of linearity may be investigated, usually being only a very rough approximation, when the values of the variables are not close to the functioning point around which the linear approximations computed.

On the other hand, to build a nonlinear model is far from easy. The structure of the nonlinearity needs to be a priori known, which is not usually the case. A typical approach consists of exploiting a priori knowledge, when available, to define a tentative structure, then refining and modifying it on the training subset of data, and finally retaining the structure that best fits a cross-validation on the testing subset of data. The problem is even more complex when the collected data

exhibit hybrid dynamics (i.e., their evolution in time is a sequence of smooth behaviours and abrupt changes).

An alternative approach is to infer the model directly from the data without a priori knowledge via an identification algorithm capable of reconstructing a very general class of piece-wise affine model (Ferrari-Trecate et al., 2003). This method also can be exploited for the data driven modeling of hybrid dynamical systems, where logic phenomena interact with the evolution of continuous-valued variables. Such an approach will be described concisely in the following.

Piece-wise affine identification exploits k-means clustering that associates data points in multivariable space in such a way to jointly determine a sequence of linear submodels and their respective regions of operation without even imposing continuity at each change in the derivative. In order to obtain such a result, the following five steps are executed:

Step 1: The model is locally linear; small sets of data points close to each other likely belong to the same submodel. For each data point, a local set is built, collecting the selected points together with a given number of its neighbours (whose cardinality is one of the parameters of the algorithm). Each local set will be pure if made of points really belonging to the same single linear subsystem; otherwise, it is mixed.

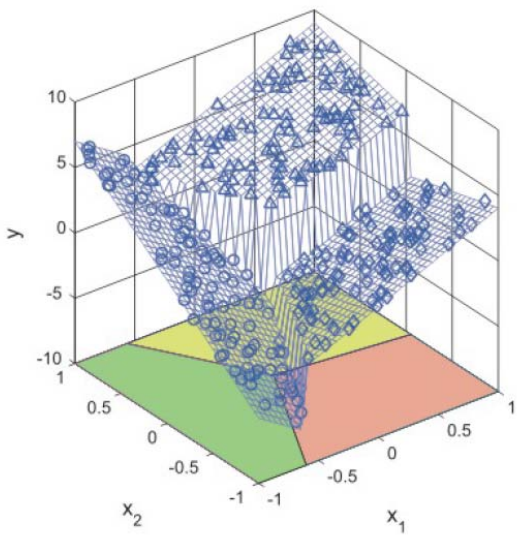
Step 2: For each local dataset, a linear model is identified through usual least squares procedure. Pure sets belonging to the same submodel give similar parameter sets, while mixed sets yield isolated vectors of coefficients, looking as outliers in the parameter space. If the signal to noise ratio is good enough, and if there are not too many mixed sets (i.e., the number of data points is enough more than the number of submodels to be identified, and the sampling is fair in every region), then the vectors will cluster in the parameter space around the values pertaining to each submodel, apart from a few outliers.

Step 3: A modified version of the classical k-means, whose convergence is guaranteed in a finite number of steps (Ferrari-Trecate et al., 2003), takes into account the confidence on pure and mixed local sets in order to cluster the parameter vectors.

Step 4: Data points are then classified, each being a local dataset one-to-one related to its generating data point, which thus is classified according to the cluster to which its parameter vector belongs.

Step 5: Both the linear submodels and their regions are estimated from the data in each subset. The coefficients are estimated via weighted least squares, taking into account the confidence measures. The shape of the polyhedral region characterizing the domain of each model may be obtained via linear support vector machines (Vapnik, 1998), easily solved via linear/quadratic programming.

Figure 3. Piece Wise Affine identification of dependence of measurements Y on conditioning variables X_1 and X_2



FUTURE TRENDS

The proposed approaches are now being applied in several contexts, such as bioinformatics, geo-informatics, neuro-informatics, systems biology and in general complex systems modeling. Nonetheless, on the methodological side, even further improvements are under consideration, like for instance Dynamics Bayesian Networks, aiming to join the inference capabilities of the Bayesian Networks with the identification of a dynamical model proper of approaches like the piece-wise affine one.

CONCLUSION

The proposed approaches are very powerful tools for quite a wide spectrum of applications in and beyond data mining, providing an up-to-date answer to the quest of formally extracting knowledge from data and sketching a model of the underlying process. The fact that a combination of different approaches, taken from partially complementary disciplines, proves to be effective may indicate a fruitful direction in combining in different ways classical and new approaches

to improving classification, making machine learning more and more sensitive to a variety of scientific issues.

REFERENCES

Barron A., Rissanen J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory, 44*, 2743-2760.

Boley, D.L. (1998). Principal direction divisive partitioning. *Data Mining and Knowledge Discovery, 2*(4), 325-344.

Cheng, G., & Greiner, R. (1999). Comparing bayesian network classifiers. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence.* San Francisco: Morgan Kaufmann.

Ferrari-Trecate, G., Muselli, M., Liberati, D., & Morari, M. (2003). A clustering technique for the identification of piecewise affine systems. *Automatica, 39*, 205-217.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine Learning, 29*, 131-161.

Garatti, S., Bittanti, S., Liberati, D., & Maffezzoli, P. (2007). An unsupervised clustering approach for leukemia classification based on DNA micro-arrays data. *Intelligent data analysis* (in print).

Golub, G.H., & van Loan, C.F. (1996). *Matrix computations*. Johns Hopkins University Press.

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data-mining*. Cambridge, MA: MIT Press.

Keogh, E., & Pazzani, M.J. (2002). Learning the structure of augmented bayesian classifiers, *International Journal on Artificial Intelligence Tools, 11*(4), 587-601.

Kononenko, I. (1995). On biases in estimating multivalued attributes. In *Proceedings of the International Joint Conference of Artificial Intelligence* (pp. 1034-1040).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, California.

Muselli, M., & Liberati, D. (2000). Training digital circuits with Hamming clustering. *IEEE Transactions on Circuits and Systems—I: Fundamental Theory and Applications, 47*, 513-527.

Muselli, M., & Liberati, D. (2002). Binary rule generation via Hamming clustering. I*EEE Transactions on Knowledge and Data Engineering, 14*, 1258-1268.

O'Connel, M.J. (1974). Search program for significant variables. *Comp. Phys. Comm., 8*, 49.

Quinlan, J.R. (1994). *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufmann.

Savaresi, S.M., & Boley, D.L. (2004). A comparative analysis on the bisecting K-means and the PDDP clustering algorithms. *International Journal on Intelligent Data Analysis,* 8(4), 345-363.

Setnes, M. (2000). Supervised fuzzy clustering for rule extraction. *IEEE Transactions on Fuzzy Systems, 8*, 416-424.

Taha, I., & Ghosh, J. (1999). Symbolic interpretation of artificial neural networks. *IEEE T Knowledge and Data Engineering, 11*, 448-463.

Vapnik, V. (1998). *Statistical learning theory*. New York: John Wiley.

Yarmus, J.S. (2003). *ABN: A fast, greedy bayesian network classifier*. Retrieved December 6, 2007, from http://otn.oracle.com/products/bi/pdf/adaptive_bayes_net.pdf

KEY TERMS

Hamming Clustering: A fast binary rule generator and variable selector able to build understandable logical expressions by analyzing the Hamming distance between samples.

Hybrid Systems: Their evolution in time is composed by both smooth dynamics and sudden jumps.

***k-means*:** Iterative clustering technique subdividing the data in such a way to maximize the distance among centroids of different clusters, while minimizing the distance among data within each cluster. It is sensitive to initialization.

Model Identification: Definition of the structure and computation of its parameters best suited to mathematically describe the process underlying the data.

PDDP (Principal Direction Divisive Partitioning): One-shot clustering technique based on principal component analysis and singular value decomposition of the data, thus partitioning the dataset according to the direction of maximum variance of the data. It is used here in order to initialize k-means.

Principal Component Analysis: Rearrangement of the data matrix in new orthogonal transformed variables ordered in decreasing order of variance.

Rule Inference: The extraction from the data of the embedded synthetic logical description of their relationships.

Salient Variables: The real players among the many apparently involved in the true core of a complex business.

Singular Value Decomposition: Algorithm able to compute the eigenvalues and eigenvectors of a matrix; also used to make principal components analysis.

Unsupervised Clustering: Automatic classification of a dataset in two of more subsets on the basis of the intrinsic properties of the data without taking into account further contextual information.

Making Sense of IS Failures

Darren Dalcher
Middlesex University, UK

INTRODUCTION

Researchers with a keen interest in information systems failures are faced with a double challenge. Not only is it difficult to obtain intimate information about the circumstances surrounding such failures, but there is also a dearth of information about the type of methods and approaches that can be utilized in this context to support such information collection and dissemination. The purpose of this chapter is to highlight some of the available approaches and to clarify and enhance the methodological underpinning that is available to researchers interested in investigating and documenting phenomena in context-rich and dynamic environments. The chapter concludes by introducing a new range of antenarrative approaches that represent future developments in the study of IS failures.

BACKGROUND

Contemporary software development practice is regularly characterized by runaway projects, late delivery, exceeded budgets, reduced functionality, and questionable quality that often translate into cancellations, reduced scope, and significant re-work cycles (Dalcher, 1994). Failures, in particular, tell a potentially grim tale. In 1995, 31.1% of US software projects were cancelled, while 52.7% were completed late, over budget (cost 189% of their original budget), and lacked essential functionality. Only 16.2% of projects were completed on time and within budget; only 9% in larger companies, where completed projects had an average of 42% of desired functionality (Standish, 2000). The 1996 cancellation figure rose to 40% (ibid.).

The cost of failed US projects in 1995 was $81 billion. In addition, cost overruns added an additional $59 billion ($250 billion was spent on 175,000 US software projects, however $140 billion out of this was spent on cancelled or over budget activities) (Standish, 2000). In fact, Jones (1994) contended that the average US cancelled project was a year late having consumed 200 percent of its expected budget at the point of cancellation. In 1996, failed projects alone totalled an estimated $100 billion (Luqi and Goguen, 1997). In 1998, 28% of projects were still failing at a cost of $75 billion, while in 2000, 65,000 of US projects were reported to be failing (Standish, 2000). As of 2004 partial failures still accounted for over 50% of all projects (Standish, 2004), whilst the figure for total failures continues to hover around the 20-25% mark.

The Standish Group makes a distinction between failed projects and challenged projects. Failed projects are cancelled before completion, never implemented, or scrapped following installation. Challenged projects are completed and approved projects which are over-budget, late, and with fewer features and functions than initially specified. Lyytinen and Hirschheim (1987) identify correspondence failures (where the system fails to correspond to what was required), process failures (failure to produce a system or failure to produce it within reasonable budgetary and time-scale constraints), interaction failures (where the system cannot be used, or is not satisfactory in terms of the interaction) and expectation failures (where the system is unable to meet a specific stakeholder group's expectations). Many situations contain behavioral, social, organizational, or even societal factors that are ignored and, therefore, the definition of failure needs to encompass a wider perspective. The general label "system failures" is often utilized in order to embrace a wider grouping of failures, including ones with undesirable side effects which may impact other domains and the organizational context (e.g., Fortune & Peters, 1995). As information becomes more embedded in other domains, the scope and impact of failure becomes more wide-reaching. This was clearly evident from the extensive effort to minimize the impact of the "year 2000 bug" from any system containing computers and underscores our interest in utilizing the term IS failure to describe a wider class of systems failures that impact on individuals, organizations and societal infrastructure.

IS failure investigations start with extensive attempts to collate relevant evidence. However, in most cases the researcher is exposed to specific information post-hoc, that is, once the failure is well established and well publicized and the participants have had a chance to rationalize their version of the story. Most of the available sources are, therefore, already in place and will have been set up by agencies other than the researcher.

The purpose of a forensic investigation is to explain a given failure by using available information and evidence. The term *forensic* is derived from the Latin 'Forensis', which is to do with making public. *Forensic science* is the applied use of a body of knowledge or practice in determining the cause of death. Nowadays extended to include any skilled investigation into how a crime was perpetrated, forensic systems engineering is the post-mortem analysis and study

of project disasters (Dalcher, 1994). The work involves a detailed investigation of a project, its environment, decisions taken, politics, human errors, and the relationship between subsystems. The work draws upon a multidisciplinary body of knowledge and assesses the project from several directions and viewpoints. The aim of forensic analysis is to improve the understanding of failures, their background, and how they come about (Dalcher, 1997). The concept of systems is a central tool for understanding the delicate relationships and their implications in the overall project environment.

Forensic systems engineering is primarily concerned with documentary analysis and (post-event) interviews in an effort to ascertain responsibility lines, causal links, and background information. The primary mode of dissemination of findings, conclusions, and lessons is through the publication of case study reports focusing on specific failures. However, there are limited research methods to explore the dynamic and fragmented nature of complex failure situations. Lyytinen and Hirschheim (1987) noted that more qualitative research methods were needed for IS failure research as well as more extensive case studies that explored problems in more detail and viewed solution arrangements in light of what transpired. The same methods also need to account for group issues and cultural implications. Sadly, twenty years on, the same constraints in terms of methods are still in evidence.

DESCRIBING FAILURE

Making sense of IS failures retrospectively is difficult. In general, there is very little objective quantitative failure information that can be relied upon. This makes the utilisation of quantitative methods less likely, until all relevant information is understood. Interpretation requires understanding of and engagement with the wider context. Indeed, a specific feature of failure is the unique interaction between the system, the participants, their perspectives, complexity and technology (Perrow, 1984). Lyytinen and Hirschheim (1987) pointed out that failure is a multifaceted phenomenon of immense complexity with multiple causes and perspectives. Research into failures often ignores the complex and important role of social arrangement embedded in the actual context. This is often due to the quantitative nature of such research. More recently, Checkland and Holwell (1998) argued that the IS field requires sensemaking to enable a richer concept of information systems.

Understanding the interactions that lead to failures likewise requires a humanistic stance that is outside the conventional positivist norm to capture the real diversity, contention, and complexity embedded in real life. Forensic analysis thus relies on utilizing qualitative approaches to obtain a richer understanding of failure phenomena in terms of action and interaction.

The fact that a failure phenomenon is being investigated, suggests that attention has already been drawn to the complexities, breakdowns, and messy interactions that such a situation entails (i.e., the investigation is problem-driven). Many such inquiries deal with subjective accounts including impressions, perceptions, and memories. The aim of the researcher is to increase, in a systemic way, the understanding of a situation, yet do so from a position that takes in the complexity of the entire situation and incorporates the different perspectives and perceptions of the stakeholders involved.

Overall, the purpose of a failure research method is to enable the researcher to make sense of the complexity of detail and the complexity of interaction, and chart the contributory role of different causes and issues in the build up to failure. However, the armoury of research methods in this domain is often limited to case studies.

The term "case study" is an umbrella term used in different contexts to mean different things that include a wide range of evidence capture and analysis procedures. Yin (1994, p.13) defines the scope of a case study as follows:

"A case study is an empirical inquiry that:

- investigates a contemporary phenomenon within its real-life context, especially when
- the boundaries between phenomenon and context are not clearly identified".

A case study can be viewed as a way of establishing valid and reliable evidence for the research process as well as presenting findings which result from research (Remenyi, 1998). According to Schramm (1971) the case study tries to illuminate a decision or a set of decisions and, in particular, emphasize why they were taken, how they were implemented, and with what results. A case study is likely to contain a detailed and in-depth analysis of a phenomenon of interest in context; in our case, the failure scenario. Table 1 summarizes some of the main advantages of using case studies.

The general aim of the case study approach is to understand phenomena in terms of issues in the original problem context by providing the mechanism for conducting an in-depth exploration. They often result from the decision to focus an enquiry around an instance or an incident (Adelman, Jenkins, and Kemmis., 1977), as they are principally concerned with the interaction of factors and events (Bell, 1999). The combination of a variety of sources offers a richer perspective which also benefits from the availability of a variety and multiplicity of methods that can be used to obtain new insights about this single instance. A case study allows the researcher to concentrate on specific instances in their natural setting and thereby attempt to identify the interacting perceptions, issues, and processes at work, ultimately resulting in in-depth understanding. Crucially, the focus on

Table 1. Main advantages of using case studies

✓	ability to identify and focus on issues
✓	richness of detail
✓	multiple perspectives
✓	multiple sources and types of data
✓	multiple explanations (no absolute truth)
✓	cross disciplinary remit
✓	ability to recognise and minimise inherent complexity
✓	ability to handle conflict, disparity and disagreement
✓	ability to show interactions
✓	ability to observe emerging patterns
✓	opportunity to focus on the particular
✓	gain real insight and understanding of a situation
✓	conducted in real-life (natural) setting
✓	encompasses original problem context
✓	ability to deal with interpretations
✓	features intensive analysis
✓	can extend the boundaries to include aspects of wider system environment
✓	can be accumulated to form an archive of cases
✓	can be strengthened and expanded with longitudinal features
✓	often retold in story format, which is more accessible to practitioners

a single incident thus enables the study of the particularity and complexity of a case, thereby coming to understand the activity within important circumstances (Stake, 1995).

There are a number of general objections that are associated with the use of case studies (see Table 2). However, one must recognize that case studies are more likely to be used retrospectively rather than as an on-going perspective (especially from a failure point-of-view), as researchers are unlikely to know the potential for useful results and interest from the outset and may have difficulty in negotiating access to the location.

Comprehensiveness of coverage is not necessarily a requirement. The richness of detail can be controlled through the careful placement of systems boundaries and consideration of the wider system environment that is relevant to the specific phenomenon under study. Case studies can

Table 2. Main objections to the use of case studies

❖	sometimes viewed as soft data (but some argue it is hard research)
❖	biases inherent in accepting views and perceptions
❖	questions about generalizability of findings (especially from a single case), but it is possible to build a library of such cases
❖	issues regarding objectivity of approach and perceived lack of rigour
❖	negotiating access to settings
❖	boundaries are difficult to define; but this could also be a strength!
❖	mainly retrospective
❖	sometimes viewed as likely to take too long and result in massive documentation
❖	the observer effect
❖	reliability of conclusions
❖	there is little control over events, but this may also be a strength

be utilized as a source of understanding, which is tolerant of ambiguity, paradox, and contradiction. A case study is viewed as interpretative when events in the real world are observed and then an effort takes place to make sense of what was observed, that is, when one tries to make sense of a failure from the perspectives of participants. They also offer the potential for generating alternative explanations from the different stakeholder perspectives thereby allowing the researcher to highlight contradictions, conflicts, and misunderstandings.

FROM CASE STUDIES TO CASE HISTORIES

The generally liberal use of the term *case study* requires a tighter definition of its meaning in failure research. While there may be a tradition of using case studies within the IS community, this is perhaps more often borrowed from the MBA culture than as a result of self-conscious effort to adopt them as a research approach (Cornford and Smithson, 1996; Walsham, 1993). Indeed, the case study is typically used more in its capacity as a teaching tool than as a *research tool*. The shift to studying the impact of issues within the organisational context renders case studies particularly useful for investigating failure scenarios. However, the use of the term often leads to some confusion.

Moreover, one of the major complications in failure investigations is in relating causes to effects (and possibly events) through extended time horizons (Dalcher, 2000). The implications of actions may not be witnessed for years, or even generations. Delays between making a decision and observing the result distort the causal link between the two. As a result, people tend to associate a different level of severity to events occurring following a delay. The perceived severity is, thus, diminished with the length of the delay further complicating the task of identifying patterns and interactions that contributed to a given failure. Failure researchers are, thus, required to provide adequate historical accounts of the interaction between actions, perceptions, and the passage of time.

Case studies have typically been used to explore issues in the present and the past and comprise of ethnographic studies, single case studies, and comparative case studies (Jankowicz, 2000), as well as action research, evaluative, exploratory, explanatory, and descriptive case studies (Bassey, 1999). In our experience there is a need to add the failure case study as a special example of a case study focusing primarily on the background, context, perception, interactions and patterns, especially as the failure investigation is likely to take place after the (failure) event. We, thus, propose the use of the label *case histories* to refer to the specialized historical research studies focusing on failure incidents.

The time dimension (sequencing) is critical to understanding interactions and identifying their impacts. Case histories are concerned with providing the background and context that are required to endow words and events with additional meaning. Background refers to previous history of the system itself, while context refers to interactions with the environment. As failures are time-and-place-dependent, the case history framework enables readers to obtain an understanding of the intimate context surrounding the main event. The primary tool available to the community is the case histories of failures (derived from the use of the case study method). These represent a detailed historical description and analysis of actual processes from a relevant perspective. Their value is in tracing decisions (and recorded rationale) to their eventual outcomes by utilizing techniques borrowed from decision analysis and systems engineering. Indeed, the historical description and presentation of a chronology of events infused with meaning, intention, and understanding are based on the recognition that real life is ambiguous, conflicting, and complex.

Case histories thus contain observations, feelings and descriptions. They can be used to construct, share, dispute and confirm meanings, interpretations and scenarios in the context of real events (e.g., Dalcher, 1995, 2004, 2007). Rather than simply highlight a chronicled sequence of happenings, they convey a story encompassing a specific perspective, focus, and possibly some inevitable biases. The interpretation plays a key part in transmutating the chronicle into a meaningful story with plot, coherence, and purpose. However, constructing a convincing narrative of a complex story with competing meanings, alternative perspectives, and inherent prejudices is a challenge in itself.

FUTURE TRENDS

Failures, in common with other activities that take place in organizations, are based on stories. The verbal medium is crucial to understanding behavior within organizations and systems, and researchers are, thus, required to collect *stories*, grounded in practice, about what takes place (Brown, Denning, Groh and Prusak, 2005; Denning, 2007; Easterby-Smith, Thorpe and Lowe, 2002; Gabriel, 2000; Gargiulo 2005; Matthews and Wacker, 2008; Simmons, 2007; White, 1973). Gargiulo (2005) further asserts that effective organizational communication and learning is dependent upon stories. Listening to them is critical to the success of the organization. Understanding failures often entails the retrospective untangling of complicated webs of actions and events, and emergent interaction patterns. Failure storytelling can, thus, be understood as a combination of narrative recounting of empirical events with the purposeful unlocking of meaningful patterns, or a plot.

Historically, storytelling has been an acceptable form of conveying and sharing ideas, norms, values, experience, and knowledge of context. It plays a key role in communicating the cultural, moral, or historical context to the listener. Indeed, Arendt, (1958) argued that the chief characteristic of human life is that it is always full of events which ultimately can be told as a story. There are even strong claims that the narrative is the main mode of human knowledge (Bruner, 1990; Schank, 1990), as well as the main mode of communication, learning, and thinking (Denning, 2001; Fisher, 1987; Gargiulo, 2005; Schank, 1990). Moreover, children are often initiated into culture (and its boundaries) through the medium of storytelling, offering models for emulation or avoidance.

In practice, the essence of any good case study revolves around the ability to generate an effective storyline, normally with a unique style, plot, or perspective. In a large case, a general theme can be obtained from selected excerpts weaved together to illustrate a particular story. Personal stories that form part of a case study can, thus, be viewed as a valid source of data organized to make sense of a theme or problem. This is particularly useful when the researcher is trying to portray a personal account of a participant, a stakeholder, or an observer in an incident, accident, or failure. The implication is that the need to address personal aspects of interaction and story is fulfilled by the development of a research-valid narrative. Indeed, Remenyi, et al. (1998) contend that a story, or a narrative description, is valid if the resulting narrative adds some knowledge. Furthermore, White (1973) describes a story as "the process of selection and arrangement of data from the unprocessed historical record in the interest of rendering the record more comprehensible to an audience of a particular kind" by inserting a sense of perspective and purpose.

Storytelling can endow listeners with different meanings as stories can be understood in multiple ways. Narratives are neither discovered, nor found; they are constructed. Understanding IS failures is, therefore, more complicated than the discovery of a simplistic chronology of events as stories are crystallized through infusion with meaning and context. Narrative inquiry is evolving into an acceptable research approach in its own right in the social sciences and in management research circles (Bell, 1999; Boje, 2001; Czarniawska, 1998, 2004; Gabriel, 2000; Easterby-Smith, et al., 2002) as the story format provides a powerful way of knowing and linking disparate accounts and perspectives. When different accounts are combined, the emerging story line benefits from the richness of multifaceted insights.

Developing a narrative requires plot as well as coherence as a story is made out of events and the plot mediates between the events and the story (Boje, 2001; Carr, 2001; Kearney, 2002). The narrative can, thus, become a powerful mechanism for eliciting and sharing experience through intimate reflection. In failure stories, the plot often emanates from the actions and perceptions of participants emerging out of the flux of events, in (direct) contradiction with expectations. The storyteller is concerned with the perspective and purpose of participants as well as with the plausibility of the emerging plot. The combination of plot, purpose, and perspective dictates the selection of elements, the filling in of links, and the removal of "irrelevant" noise.

Post-modern interpretation contends that most real life stories are fragmented, non-linear, discontinuous, multivariate, and incoherent. This has already been highlighted as a feature of failure stories. Such stories also tend to be dynamic, polyphonic (multi-voiced), and collectively produced as they occur in asymmetrical, random, and turbulent environments full of tensions and ambiguities. The stories are not plotted as such and they appear to flow, emerge, and network offering complex clustering of events, emergent phenomena, causes, and effects. Moreover, the accounts are often subjective, counter-intuitive, and contradictory. This leads to interacting and conflicting webs of narratives, characterized by coincidences, predicaments, and crises.

Generally, stories appear to be improperly told, as a story is an "ante" state of affairs existing previously to a carefully constructed narrative (Boje, 2001). The *antenarrative*, or the "real" story, is the fragmented, messy, and dynamic, multivocal, multi-plotted, multi-version, and complex tale. Indeed, modern storytellers look for new ways and mediums for weaving and depicting a multi-vocal reality, as exemplified by Mike Finggis's digitally shot film *Time's Arrow,* where the screen is split in four to allow for four separate perspectives and sub-stories that occasionally intersect or overlap. In the tradition of post-modern inquiry, a real life researcher is often faced with fragments rather than a whole story to tell; and many of the fragments may reflect contrary versions of reality. This is potentially more acute when the accounts attempt to justify roles of participants in the lead-up to disaster. It would also appear from past analysis that there are hierarchies of stories and stories that exist within, or interact with, other stories. Using the terminology provided by Boje, (2001), the purpose of narrative methods is to take a complex situation characterized by collective (yet often conflicting) memory and an antenarrative, and construct the plot and coherence that can be used to narrate the story of interest.

The reality in failure stories is of multi-stranded stories of experiences and reactions that lack collective consensus. Indeed, the discipline of decision making has also recognized that making choices is about forming and selecting interpretations from a mosaic of possibilities (March, 1994, 1997; Weick, 1995). Not surprisingly, disasters or traumatic stories are hard to narrate, understand, and justify. Stories have three basic properties: time, place, and mind (Boje, 2001), which interact and build up as the story evolves. In forensic case histories, these are further clarified through the

identification of the background and context which clarify and justify the interpretation in the context of the emerging phenomena.

Boje (2001) and Kearney (2002) contend that the current view is of sequential single voice stories and implies excessive reliance on the hypothetical-deductive approach (akin to simplistic causal pairings). Imposing a meaning is insufficient and inadequate as actors need to find their own voice and make collective sense of a situation. The answer is not to develop Harvard type case studies, but to rewrite stories as polyvocal tapestries enabling different perceptions and interpretations to exist, thereby explaining webs of actions and interactions. What is new in this approach is the antenarrative reading which enables narrative analysis methods to be supplemented by antenarrative methods, allowing previously fragmented and personal storytelling to be interpreted as a unified whole. This focus offers alternative discourse analysis strategies that can be applied where qualitative story analyses can help to assess subjective, yet "insightful" knowledge in order to obtain "true" understanding of complex interactions.

As for the long term future, good stories can also benefit from pictures. Once we have mastered the techniques of telling complex, modern stories, we need to focus on composing that information. Even the most gripping story needs to be made attractive and believable. Textual information needs additional support not only in "emplotting" and in maintaining coherence and perspective, but also in ascertaining the plausibility of constructed stories and in differentiating between noise and narrative. Developing improved techniques for organizing or visualizing knowledge (such as Net maps) can, therefore, help in untangling some of the fragmented strands as well as in making the stories more readable and understandable, and, ultimately, more appealing.

CONCLUSION

Stories provide a powerful research tool that can be used to reflect, share, and make sense. With the benefit of hindsight, it is possible to reconstruct a systematic re-telling of events that have led to a failure. The narrated structure provides an explanation as to how and why failures occur. The purpose of the structure is to make sense of a rich tapestry of interactions and connections by following an identified storyline that chronicles and links the relevant issues within the environment. Engaging with the world of others through the medium of a story enables a deeper and richer reflection. Indeed, recounted life may prize open perspectives that would have been inaccessible using ordinary methods and thinking arrangements. Moreover, failure tends to highlight missing and incorrect assumptions, and faulty defensive mechanisms, and can, therefore, serve as a pretext to updating the frame of reference or the context for understanding as the listeners learn to construct a meaning and make sense of a highly multifaceted, multilayered, and complex situation.

REFERENCES

Adelman, C., Jenkins, D. & Kemmis, S. (1977). Rethinking Case Study: Notes from the Second Cambridge Conference, *Cambridge Journal of Education*, 6, 139-150.

Arendt, H. (1958). *The Human Condition*. Chicago: University of Chicago Press.

Bassey, M. (1999). *Case Study Research in Educational Settings*. Buckingham: Open University Press.

Bell, J. (1999). *Doing Your Research Project, A Guide for First-time Researchers in Education and Social Science,* 3 Ed. Buckingham: Open University Press.

Boje, D. M. (2001). *Narrative Methods for Organisational & Communication Research*. London: Sage.

Brown, J. S., Denning, S., Groh, K. & Prusak, L. (2005). *Storytelling in Organizations: Why storytelling is transforming 21st Century organizations and management*. Burlington, MA: Elsevier.

Bruner, J. (1986). *Actual Minds, Possible Worlds*. Cambridge, MA: Harvard University Press.

Bruner, J. (1990). *Acts of Meaning*. Cambridge, MA: Harvard University Press.

Carr, D. (2001). Narrative and the Real World: An argument for Continuity, in Roberts G. (Ed.) *The History and Narrative Reader*. London: Routledge, 143-156.

Checkland, P. & Holwell, S. (1998). *Information, Systems and Information systems – Making Sense of the Field*. Chichester: Wiley.

Cornford, T. & Smithson, S. (1996). *Project Research in Information Systems: A Student's Guide*. Basingstoke: Macmillan.

Czarniawska, B. (1998). *A Narrative Approach to Organization Studies*. London: Sage.

Czarniawska, B. (2004). *Narratives in Social Science Research*. London: Sage.

Dalcher, D. (1994). Falling down is part of Growing up; the Study of Failure and the Software Engineering Community. *Proceedings of 7th SEI Education in Software Engineering Conference*, New York: Springer-verlag, pp. 489-496.

Dalcher, D. (1995). *The London Ambulance Service Fiasco, Proceedings of the 8th International Workshop on the Engi-*

neering of Computer Based Systems. Tucson, Arizone, IEEE Press, 1995, pp. 445-450.

Dalcher, D. (1997). The Study of Failure and Software Engineering Research. *Proceeding of the UK Software Engineering Association Easter Workshop*. London*:* Imperial College, April 1997, pp. 14-19.

Dalcher, D. (2000). Feedback, Planning and Control – A Dynamic Relationship. *FEAST 2000*. Imperial College, London, July 2000, pp. 34-38.

Dalcher, D. (2004). Still Waiting? Computerisation of Ambulance Despatch S ystems. *Annals of Cases on Information Technology, ACIT*, Volume VI, 2003/4, pp. 440-456.

Dalcher, D. (2007). Why the Pilot Cannot be Blamed: A Cautionary Note About Excessive Reliance on Technology. *International Journal on Risk Assessment and Management (IJRAM)*, Vol. 7, no. 3, pp. 350-366.

Denning, S. (2001). *The Springboard: How Storytelling Ignites Action in Knowledge-Era Organizations*. Boston: Butterworth-Heinemann.

Denning, S. (2007). *The Secret Language of Leadership*. San Francisco: Jossey-Bass.

Easterby-Smith, M., Thorpe, M. & Lowe, A. (2002). *Management Research*, 2 Ed. London: Sage.

Fisher, W. R. (1987). *Human Communication as Narration: Towards a Philosophy of Reason, Value and Action*. Columbia: University of South Carolina Press.

Fortune, J. & Peters, G. (1995). *Learning From Failure: The Systems Approach*. Chichester: John Wiley.

Gabriel, Y. (2000). *Storytelling in Organizations: Facts, Fictions and Fantasies*. Oxford: Oxford University Press.

Gargiulo, T. L. (2005). *The Strategic Use of Stories in Organizational .Communication and Learning* New York: M. E. Sharpe.

Jankowicz, A. D. (2000). *Business Research Projects*. 3 Ed. London: Business Press.

Jones, C. (1994). *Assessment and Control of Software Risks*. Englewood Cliffs, New Jersey: Prentice-Hall.

Kearney, R. (2002). *On Stories*. London: Routledge.

Luqi and Goguen, J.A. (1997). Formal Methods: Promises and Problems*. IEEE Software*. Vol 14(1), pp. 73-85.

Lyytinen, K. & Hirschheim, R. (1987). Information Systems Failures: A Survey and Classification of the Empirical Literature. *Oxford Surveys in Information Technology*. Vol. 4, 257-309.

March, J. G. (1994). *A Primer on Decision Making*. New York: Free Press.

March, J. G. (1997). Understanding How Decisions Happen in Organisations, in *Organisational Decision Making*, Shapira Z. (Ed.). Cambridge: Cambridge University Press, pp. 9-34.

Matthews, R. & Wacker, W. (2008). *What's Your Story?* Upper Saddle River, NJ: Pearson.

Perrow, C. (1984). *Normal Accidents, Living with High-Risk Technologies*. New York: Basic Books.

Remenyi, et. al. (1998). *Doing Research in Business and Management: An Introduction to Process and Method*. London: Sage.

Schank, R. C. (1990). *Tell Me a Story: Narrative and Intelligence*. Evanston: Northwestern University Press.

Schramm, W. (1971). *Notes on Case Studies of Instructional Media Projects*. Working paper for the Academy for Educational Development. Washington, DC.

Simmons, A. (2007) *Who Ever Tells the Best Story Wins*. New York: AMACOM,

Stake, R. E. (1995). *The Art of Case Study Research*. Thousand Oaks: Sage.

Standish_Group, (2000). *Chaos 2000*. Standish: Dennis, MA.

Standish_Group, (2004). *Chaos 2004*. Standish: Dennis, MA.

Walsham, G. (1993). *Interpreting Information Systems in Organizations*. Chichester: Wiley.

Weick, K. E. (1995). *Sensemaking in Organisations*. Thousand Oaks, CA: Sage Publications.

White, H. (1973). *Metahistory*. Baltimore: The John Hopkins University Press.

Wortmann, C. (2006). *What's Your Story? Using Stories to Ignite Performance and Be More Successful.* Chicago: Kaplan Books.

Yin, R. K. (1994). *Case Study Research: Design and Methods* (2nd Ed.). Newbury Park, CA: Sage.

KEY TERMS

Antenarrative: The fragmented and messy and dynamic stories of real life in their original context before a clear narrative is developed to explain away a certain aspect.

Case History: Specialized historical research focusing on failure incidents. Case histories emphasize the background and context that can help in untangling relationships and causes thus making sense of the events leading to the failure.

Case Study: Investigation of phenomena in naturalistic setting, conducted in order to enable in depth analysis of that phenomena.

Challenged Projects: Completed and approved projects which are late, over budget, and have fewer features and functions than originally specified. The degree of challenge depends on the way constrains are applied and interpreted within the organisation.

Failed Projects: Projects that are: cancelled before completion, are never implemented or are scrapped following installation. May also apply to projects that involve significant litigation.

Forensic Systems Engineering: Post-mortem analysis and study of a project failure or disaster aimed at uncovering causes and relationships.

Storytelling: A method of communicating and sharing ideas, experiences and knowledge in a specific context.

Management Considerations for B2B Online Exchanges

Norm Archer
McMaster University, Canada

INTRODUCTION

Information systems that link businesses for the purpose of inter-organizational transfer of business transaction information (inter-organizational information systems, or IOIS) have been in use since the 1970s (Lankford & Riggs, 1996). Early systems relied on private networks, using electronic data interchange (EDI) or United Nations EDIFACT standards for format and content of transaction messages. Due to their cost and complexity, the use of these systems was confined primarily to large companies, but low-cost Internet commercialization has led to much more widespread adoption of IOIS. Systems using the Internet and the World Wide Web are commonly referred to as B2B (business-to-business) systems, supporting B2B electronic commerce.

Technological innovations have led to several forms of B2B Internet implementations, often in the form of online exchanges. These are virtual marketplaces where buyers and sellers exchange information about prices, products, and service offerings, and negotiate business transactions. In addition to substituting proprietary lines of communication, emerging technologies and public networks have also facilitated new business models and new forms of interaction and collaboration, in areas such as collaborative product engineering or joint offerings of complex, modularized products. During the years 1999-2001 a number of online exchanges were introduced, but many of these failed (Gallaugher & Ramanathan, 2002), due mainly to an inability to attract participating business partners. Those that have survived are often owned by companies or consortia that are also exchange customers or suppliers.

The objective of this overview is to describe the evolution and the characteristics of B2B Internet implementations, and to discuss management considerations, the evaluation and adoption of B2B applications, and the technical infrastructure supporting these systems. We also indicate some of the open issues that remain as the technology and its adoption continues to evolve.

BACKGROUND

Although there are many classification schemes available for online exchanges (Choudhury, 1997; Kaplan & Sawhney, 2000), we will use a more generic and functional focus, with three categories: sell-side, buy-side, and neutral/market-type applications (Archer & Gebauer, 2001). Early B2B sell-side applications featured online catalogs, made available to the Internet community by distributors and manufacturers, often complemented by features such as shopping baskets and payment functionality. Many now provide customized and secure views of the data, based on business rules from contract agreements with individual customers. In some cases, buying processes of the customers are supported, including features such as approval routing and reporting. While some sophisticated applications exist to support collaborative forecasting or the configuration of complex products, many sell-side systems handle only the simpler transactions, such as maintenance, repair, and operation (MRO) supplies. Recently, more advanced features have become more widely available, such as CPFR (collaborative planning, forecasting, and replenishment) to support joint initiatives between customer and supplier (Holmstrom, Framling, Kaipia & Saranen, 2002).

Buy-side applications support procurement, moving order processes closer to the end user, and alleviating structured workloads in functional departments such as purchasing and accounts payable. For smaller companies, an affordable alternative is to work through hosted solutions, using Internet browsers to access procurement functionality provided by a third-party vendor or application service provider (ASP). Some applications provide functionality beyond the automation of highly structured procurement processes, including production tendering, and multi-step generation requests for proposals, as they are relevant for the procurement of freelance and management services. Interfacing purchasing systems to internal systems such as enterprise resource planning systems (ERP) makes it possible to automate many transactions, thus greatly increasing processing speed and reducing costs. Buy-side solutions that involve long-term inter-organizational relationships are typically set up by the purchasing organization, which then controls catalog content, data format, and back-end system functionality. Benefits

include a reduction in maverick buying, and freeing purchasing and accounts payable personnel from clerical work to handle more strategic tasks. Suppliers typically benefit from long-term relationships, and in many cases the relationships between the buyer and its suppliers were in place before the buy-side operation was established.

The third group of applications, often referred to as B2B electronic markets or hubs, can either bring together multiple buyers and sellers on an ad hoc basis involving various types of auctions, or support more permanent relationships (a many-to-many relationship, equivalent to IOIS). Those that have been more successful are likely to have been sponsored by a consortium (e.g., GlobalNetXchange, in the retail industry, sponsored by buying organizations, and Global Healthcare Exchange in the health care industry, sponsored by selling organizations). They may feature auctions, electronic catalogs, and auxiliary value-added functions, such as industry news and online forums. The initiator typically controls the catalog content, aggregates supplier input, and provides additional functionality and standardized data access to buyers. These marketplaces may eliminate the need for market participants to link directly to their business partners, circumventing costly value-added EDI network services. Their business models typically include service charges based on transaction volume and setup costs. They provide a standard for suppliers to deliver catalog content, increase flexibility if they support access to suppliers and customers outside pre-established relationships, and create customer value through competitive pressure. Participation in such marketplace solutions may also provide a low-cost alternative for SMEs (small and medium enterprises).

MANAGEMENT CONSIDERATIONS

A market assumes an intermediary role that supports trade between buyers and suppliers, including (Bailey & Bakos, 1997): a) matching buyers and sellers, b) ensuring trust among participants by maintaining a neutral position, c) facilitating market operations by supporting certain transaction phases, and d) aggregating buyer demand and seller information. Supporting the marketplace through an electronic exchange has characteristics of (Bakos, 1991): 1) cost reductions, 2) increased benefits with the number of participants, 3) potential switching costs, 4) capital investments but economies of scale and scope, and 5) significant uncertainties in benefits. Many of the management issues of B2B electronic commerce systems relate to the need to coordinate decisions and processes among multiple firms, often through differences in business processes, information systems, business models, and organizational cultures.

Early transaction cost theory recognized markets and hierarchies as the two main methods of governance for coordinating flows of goods and services. Markets such as stock exchanges coordinate the flow through supply and demand forces, with price as the main coordination vehicle. Hierarchies such as production networks consist of predetermined relationships among customers and suppliers, and rely on managerial decisions to coordinate flows. There are many intermediate forms of governance, such as network organizations and strategic alliances (Gulati, 1998). A common theme among all these governance structures is collaboration among the participants, but the level of collaboration varies. These levels can be described as cooperation, coordination, and collaboration (Winer & Ray, 1994). In cooperation, there is little sharing of goods, services, or expertise; coordination requires mutual planning and open communication among participants, who share resources; collaboration involves deeply synergistic efforts that benefit all parties. Collaboration at different levels between buyers and sellers is emphasized by online exchanges, but this can also take place among buyers and among sellers (Wang & Archer, 2004).

In recent years there has been a "move to the middle," with a growth of outsourcing arrangements and more cooperative, integrated long-term inter-organizational relationships with a relatively small number of preferred suppliers (Clemons, Reddi & Row, 1993). Distribution of market power is often an overriding factor. For example, auto manufacturers, as a concentrated industry, will be likely to adopt an approach that involves long-term collaborative relationships among business partners rather than the short-term market-driven relationships that traditionally characterized this industry. On the other hand, companies in fragmented industries such as construction are characterized by short-term relationships and low levels of trust, where transactions such as online procurement are more likely to be through B2B tendering and auctions (Stein, Hawking & Wyld, 2003).

EVALUATION AND ADOPTION

The task of evaluating an electronic exchange becomes difficult when network effects are taken into account (benefits from participating are usually positively related to the number of participants). As a result of complications such as these (strategic necessity, dependence on the commitment of business partners, additional risk, external effects, etc.), the evaluation of an electronic exchange is much more complex than for systems deployed within organizations (Gebauer & Buxmann, 2000). B2B market mechanisms focus on four factors that favor one market mechanism over another: degree of fragmentation, asset specificity, complexity of product description, and complexity of value assessment (Mahadevan, 2003). These have a significant impact on the choice of an appropriate market mechanism for B2B interactions.

From the organizational perspective of setting up a successful B2B application, there are initiators and (potential)

participants. Initiators bear the majority of the cost and risk, but on the other hand also enjoy the majority of the benefits, and they typically decide on technology infrastructure, type of systems used, corporate identity, representation of partners, and selection of participants. Success of the system depends on the participation of a critical mass of business partners. Supplier participation considerations in a buy-side solution, for example, include investments necessary to prepare and upload catalog data, integration with back-end systems, training of staff, and adjustments of business processes. Depending on individual arrangements, benefits include reduced time and costs for order processing, improved customer service, increased customer reach in a globalized marketplace, and an increase in revenues from long-term and trusted customer relationships. Neutral intermediaries in such markets face a difficult balancing task, as they have to be careful to satisfy suppliers as well as buyers, and a business model must be chosen that will attract the desired participants.

Although B2B exchanges have received a great deal of attention in the press and among researchers, their rate of adoption by business has not been high. While their aggregate transaction growth rate is higher, they are still (in 2004) outranked by at least a factor of 10 in transaction volume by EDI installations, which are still firmly in place in many large corporations (Jakovljevic, 2004). In addition, by utilizing EDI over the Internet, companies and organizations also benefit from the ability to facilitate a seamless bridging between XML and EDI that can now co-exist on the same infrastructure and use common protocols to handle electronic procurement, invoicing, and logistics information. With the need to electronically exchange volume-intensive catalogue and product specification information, organizations can significantly reduce the high cost of this exchange by using Internet EDI. Although EDI is able to support only text content, the combination with XML provides support for images and graphics (Hamdar, 2002).

SMEs can and often do handle B2B transactions through e-commerce solutions without fully automated transaction management systems, including hosted procurement applications. Supply-side solutions with Web access can also be used as parallel and partially automated channels for larger businesses that wish to deal with small suppliers or customers. In practice, SMEs execute small numbers of transactions and may not wish to make the investment in resources, training, and internal integration required to link to their business partners (Archer, Wang & Kang, 2003). Motivations for joining online exchanges include (Gebauer & Raupp, 2000): a) coercion (through market power), b) long-term commitment to business relationships and reduction of associated uncertainty, c) subsidies to support system installations for potential partners, and d) general system improvements that result in improved efficiencies and effectiveness. SMEs that link to online exchanges are most likely to be motivated through pressure from their larger partners and by long-term commitments. Many use alternative interactions utilizing a combination of manual and online functions. For example, a medium-sized value-added retailer might use ad hoc purchasing procedures such as searching the Web for catalogue information on major suppliers, and then use the telephone to negotiate prices and delivery schedules (Archer et al., 2003).

TECHNICAL INFRASTRUCTURE

Software products to support B2B interactions are continuing to mature as more complex functions are added, such as collaborative planning, forecasting, replenishment, negotiation and decision support, and procurement and asset management of complex and highly customizable items and systems (Paul et al., 2003). Linking data from many different sources, including legacy systems, through Web services (Iyer, Freedman, Gaynor & Wyner, 2003) is still in its infancy, but supported by diffusion of industry standard Extensible Markup Language (XML). Technical issues are complicated by the critical role of security and confidentiality in inter-organizational settings, particularly when using public networks such as the Internet as compared to the private networks that were traditionally used for EDI implementations. Meeting these needs may require significant investments in software, training, business process reengineering, technical support, and time, all of which favor larger organizations.

Complexity of transactions is an important adoption factor, and is determined by factors such as the number of sub-processes and organizational units that are involved, as well as their possible interactions, interdependencies and relationships with the process environment (Gebauer & Buxmann, 2000). This in turn depends on the type of goods or services. Acquiring indirect or non-production supplies and services is the least complex type of transaction, followed by direct goods, and capital goods and other types of ad hoc purchases tend to be the most complex. A high degree of automation is economically viable only for high volume, less complex transactions. As complexity increases and volume decreases, human intervention is more likely to be needed to handle exceptions and ad hoc transactions.

B2B applications can have major impacts on inter-organizational business processes, depending on the level of IOIS integration required (Stelzer, 2001). After planning and designing the system business model and infrastructure, a careful plan of how to implement it, how to train employees, and how to adapt business processes is the next step towards a successful project (Archer & Gebauer, 2001). Partner adoption, catalog management, and integration with a heterogeneous system of back-end applications are frequently listed as major stumbling blocks. For example, an organization could start out by reengineering and then automating an inefficient process that causes long lead times

and possibly frequent complaints, such as management approval of end-user requests. As a next step, putting together an online catalog that contains the offerings of preferred suppliers can be useful as a first step to reduce "maverick" buying outside pre-established contracts. While the exact steps will depend on the situation within the individual firm, the stepwise approach will also allow frequent adjustments during the project planning and implementation process, including the addition of new requirements. The adoption of a B2B e-commerce solution is a strategic company decision and it is important to evaluate the potential overall impact of this innovation on the firm before proceeding (Pant & Hsu, 1996), as it may require substantial reengineering before it will be effective (Maull, Childe, Smart & Bennett, 1995).

FUTURE TRENDS

There is little doubt that rapid growth will continue in the relative value of B2B transactions handled through electronic commerce solutions, especially as they become less costly and easier to implement in SMEs. However, most of the growth in such offerings is likely to be in sell-side or buy-side online exchanges, as there has been significant supplier resistance to participating in neutral market-type exchanges. Although these exchanges offer the greatest theoretical benefit because of the potential for standardized linkages among participating companies, and the collaborative functionalities offered by such systems, this does not outweigh resistance from suppliers who see declining profit margins due to transaction cost payments and competitive price bidding. Meanwhile, EDI systems continue to link many large companies, due to their reluctance to give up related investments or to change business processes to accommodate the newer solutions.

CONCLUSION

A clear understanding of the possibilities of emerging technologies is crucial to take advantage of new opportunities in the B2B marketplace. There have been failures in this environment due to a lack of consideration of the wide range of technical, managerial, and economic issues involved. No widely adopted frameworks have been developed to assist in the choice of level of integration or in reengineering boundary spanning business processes. Although technology continues to develop, it is still immature in many areas. In particular, the integration with current IT infrastructures is often extremely complex and difficult to justify for medium to low transaction rates. Meanwhile, B2B applications continue to evolve, changing the rules of the game in subtle ways, but providing fruitful areas for research and new developments.

REFERENCES

Archer, N., & Gebauer, J. (2001). B2B applications to support business transactions: Overview and management considerations. In M. Warkentin (Ed.), *Business-to-business electronic commerce: Challenges and solutions* (pp. 19-44). Hershey, PA: Idea Group Publishing.

Archer, N., Wang, S., & Kang, C. (2003). *Barriers to Canadian SME adoption of Internet solutions for procurement and supply chain interactions.* MeRC Working Paper #5. Hamilton, ON: McMaster eBusiness Research Centre.

Bailey, J., & Bakos, Y. (1997). An exploratory study of the emerging role of electronic intermediaries. *International Journal of Electronic Commerce, 1*(3), 7-20.

Bakos, J.Y. (1991, September). A strategic analysis of electronic marketplaces. *MIS Quarterly, 15,* 295-310.

Choudhury, V. (1997). Strategic choices in the development of interorganizational information systems. *Information Systems Research, 8*(1), 1-24.

Clemons, E.K., Reddi, S.P., & Row, M.C. (1993). The impact of information technology on the organization of economic activity: The "move to the middle" hypothesis. *Journal of Management Information Systems, 10*(2), 9-35.

Gallaugher, J.M., & Ramanathan, S.C. (2002). Online exchanges and beyond: Issues and challenges in crafting successful B2B marketplaces. In M. Warkentin (Ed.), *Business to business electronic commerce: Challenges and solutions* (chap. III). Hershey, PA: Idea Group Publishing.

Gebauer, J., & Buxmann, P. (2000). Assessing the value of interorganizational systems to support business transactions. *International Journal of Electronic Commerce, 4*(4), 61-82.

Gebauer, J., & Raupp, M. (2000). Zwischenbetriebliche elektronische katalogsysteme: Netwerkstrategische gestaltungoptionen und erfolgsfaktoren (Interorganizational electronic catalogs: Strategic options and success factors) - in German. *Informatik Forschung und Entwicklun, 15,* 215-225.

Gulati, R. (1998). Alliances and networks. *Strategic Management Journal, 19*(4), 293-317.

Hamdar, M. (2002, June). Catch-up: SMEs can benefit from Internet EDI strategy. *Purchasing B2B.*

Holmstrom, J., Framling, K., Kaipia, R., & Saranen, J. (2002). Collaborative planning forecasting and replenishment: New solutions needed for mass collaboration. *Supply Chain Management, 7*(3), 136-145.

Iyer, B., Freedman, J., Gaynor, M., & Wyner, G. (2003). Web services: Enabling dynamic business networks. *Communications of the Association for Information Systems, 11*, 525-554.

Jakovljevic, P.J. (2004, March 4). EDI versus XML: Working in tandem rather than competing? *TechnologyEvaluation.com, 4*.

Kaplan, S., & Sawhney, M. (2000). E-hubs: The new B2B marketplaces. *Harvard Business Review, 78*(3), 97-100.

Lankford, W.M., & Riggs, W.E. (1996). Electronic data interchange: Where are we today? *Journal of Systems Management, 47*(2), 58-62.

Mahadevan, B. (2003). Making sense of emerging market structures in B2B e-commerce. *California Management Review, 46*(1), 86.

Maull, R.S., Childe, S.J., Smart, P.A., & Bennett, J. (1995). Current issues in business process reengineering. *International Journal of Operations and Production Management, 15*(11), 37-52.

Pant, S., & Hsu, C. (1996). Business on the Web: Strategies and economics. *Computer Networks and ISDN Systems, 28,* 1481-1492.

Paul, J., Withanachchi, S., Mockler, R.J., Gartenfeld, M.E., Bistline, W., & Dologite, D.G. (2003). Enabling B2B marketplaces: The case of GE Global Exchange Services. In M. Khosrow-Pour (Ed.), *Annals of cases on information technology* (vol. 5, pp. 464-486). Hershey, PA: Idea Group Publishing.

Stein, A., Hawking, P., & Wyld, D.C. (2003). The 20% solution?: A case study on the efficacy of reverse auctions. *Management Research News, 26,* 1-20.

Stelzer, D. (2001). *Success factors of electronic marketplaces: A model-based approach*. Ilmenau, Germany: Technische Universitat Ilmenau.

Wang, S., & Archer, N. (2004). Supporting collaboration in business-to-business electronic marketplaces (in press). *Information Systems and e-Business Management*.

Winer, M., & Ray, K. (1994). *Collaboration handbook: Creating, sustaining, and enjoying the journey.* Saint Paul, MN: Amherst H. Wilder Foundation.

KEY TERMS

Application Service Provider (ASP): An ASP is a service company that can support and relieve a firm from the daunting challenges of finding, hiring, inspiring and training technical personnel to manage an application in-house. An ASP provides software applications on a pay-per-use or service basis via the Internet and leased lines.

Collaborative Planning, Forecasting, and Replenishment (CPFR): CPFR is a global, open, and neutral business process standard for value chain partners to coordinate the various activities of purchasing, production planning, demand forecasting, and inventory replenishment, in order to reduce the variance between supply and demand and share the benefits of a more efficient and effective supply chain.

Electronic Data Interchange (EDI): A standard used to govern the formatting and transfer of transaction data between different companies, using networks such as the Internet. As more companies are linking to the Internet, EDI is becoming increasingly important as an easy mechanism for companies to share transaction information on buying, selling, and trading. ANSI (American National Standards Institute) has approved a set of EDI standards known as the X12 standards. Although not yet a global standard, because of EDIFACT, a standard developed by the United Nations and used primarily in non-North American countries, negotiations are underway to combine the two into a worldwide standard.

Enterprise Resource Planning (ERP): A business management system that can integrate all facets of the business, including planning, manufacturing, sales, and marketing, through a common database. As the ERP methodology has become more popular, software applications have been developed to help business managers implement ERP in business activities such as inventory control, order tracking, customer service, finance and human resources.

Extensible Markup Language (XML): Document type definitions that can be used to specify or describe various types of objects. When a set of these is used on the Web to describe product information, it is referred to as cXML or commerce XML. It works as a meta-language that defines necessary information about a product, and standards are being developed for cXML in a number of industries, performing a function similar to that of EDI for non-Web-based systems. It will help to standardize the exchange of Web catalog content and to define request/response processes for secure electronic transactions over the Internet. The processes include purchase orders, change orders, acknowledgments, status updates, ship notifications and payment transactions.

Inter-Organizational Information System (IOIS): (Sometimes referred to as an IOS). An automated information system, built around computer and communication technology, which is shared by two or more companies. It facilitates the creation, storage, transformation, and transmission of information across a company's organizational boundaries to its business partners.

Maintenance, Repair, and Operations (MRO): Supplies and services purchased for use internally in the company, often referred to as indirect or non-production supplies and services (such as office supplies, computer equipment and repairs, cleaning supplies, etc.). These tend to be low unit cost, low volume, and off-the-shelf purchases.

Small and Medium Enterprise (SME): The definition of small and medium enterprises varies from country to country. If the definition is based on number of employees, SMEs in the U.S. have from 1 to 499 employees. The dividing line between a small and medium business is variously defined as being either 50 or 100 employees.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1858-1863, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Managing Converging Content in Organizations

Anne Honkaranta
University of Jyväskylä, Finland

Pasi Tyrväinen
University of Jyväskylä, Finland

INTRODUCTION

Content management is essential for organizational work. It has been defined as "a variety of tools and methods that are used together to collect, process, and deliver content of diverse types" (McIntosh, 2000, p. 1). Content management originates from document management. In fact, a great deal of contemporary content management system functionality has evolved from document management systems.

Documents are identifiable units of content, flexibly structured for human comprehension (Murphy, 2001; Salminen, 2003). They have traditionally been considered as containers for organizational content. *Document management* considers the creation, manipulation, use, publishing, archiving, and disposal of documents as well as the continuous development and design of these activities in organizational domains. In different domains, the requirements for document management differ accordingly. For example, manufacturing companies possess a bulk of technical drawings to be managed, and in e-government organizations, the document content may act as a normative reference that needs to be frozen and archived for long periods of time (Honkaranta, Salminen, & Peltola, 2005). Therefore document management in e-government is commonly split into two types: document management focusing on document production and the records management considering document repository management.

Research on document management in organizations has been carried out focusing on a multitude of issues, including document standardization (Salminen, 2003), document metadata (Murphy, 1998), document and information retrieval (Blair, 2002), the social role of documents for organizational groups (Murphy, 2001), as well as document engineering (Glushko & McGrath, 2005).

The wide selection of content management systems available has evolved mainly from document management systems (Medina, Meyers, Bragg, & Klima, 2002). They combine into single systems various functionalities developed separately in domains such as library sciences, text databases, information retrieval, and engineering databases. The essential features of document management systems cover:

- Library services and version management
- Management of user roles and access rights
- Text retrieval based on metadata and full-text search
- Support for document life-cycle and related workflows
- Management of *metadata*, as information about documents
- Multi-channel publishing for a multitude of devices and print

A survey on content management systems revealed that many of the systems still have a monolithic and closed architecture and their ability to adopt proprietary encodings is scarce (Paganelli & Pettenati, 2005). Contemporary content management systems' support for access management and for customizing workflows for integrating content into organizational processes may be modest. For example, the popular Microsoft SharePoint Server (http://www.microsoft.com/sharepoint/default.mspx) only assigns access rights to folders, not to individual files or units within the files. Content management software may include limited functionality for the design and management of an organization's Web site. The applicability of the document management approach and the systems for content management have been limited due to an orientation towards using documents as the only unit for managing content. As a consequence of this approach, long documents are difficult to browse through, portions of document content are difficult to reuse in other documents, and long documents are inconvenient for Web delivery (Honkaranta et al., 2005). At least two recent approaches on content management which aim at complementing these weaknesses can be identified. These are Web content management and the use of structured documents in the form of XML.

BACKGROUND

The Web Content Management (WCM) approach focuses on Web content publishing. A great deal of research efforts (e.g., McIntosh, 2000; Boiko, 2002; Murugesan & Ginige,

2005) are targeted specifically on Web content management. One focus in the approach is the reuse of content blocks, enforced either by an extensive use of metadata (Boiko, 2002) or by adopting a single-sourcing approach and XML (Rockley, Kostur, & Manning, 2003). The underlying approaches and the conceptual base used can be traced back to electronic publishing (e.g., Boiko, 2002; Rockley, Kostur & et al., 2003) and to database-oriented approaches (e.g., McIntosh, 2000). Technology-driven development and the growing adoption of open source software, such as Plone (http://plone.org/foundation/) and eZPublish (http://ez.no/products/ez_publish) are also characteristic to the Web content management approach. Therefore many researchers like Murugesan and Ginige (2005) call for more disciplined and more method-based development and maintenance for WCM and Web application development.

The conceptual base for the novel approach is inconsistent and immature (Grossniklaus & Norrie, 2002). The content life-cycle may involve different phases which are dealt with concepts and terms that are not yet stabilized. There are also differences in the content workflow. For example, according to McIntosh (2000, p.1), a content life-cycle consists of three main phases; 1) content assembly, 2) content production, and 3) content delivery. However, Boiko (2002) utilizes concepts such as content acquisition, aggregation and metatorial processing. The lack of a concise conceptual base may hinder requirements elicitation and cause communicational breakdowns between system analysts and the people in an organization.

The WCM approach focuses on managing the content delivered on the Web, while the content management approach manages the content right after its creation regardless of its (possibly multiple different) publishing channels. The Web may be just one additional channel for publication from the content management system perspective. Therefore the organization whose Web site is primarily meant for delivering textual and multimedia content rather than as an application interface should not only consider content management on the Web, but content management as a whole. Yet the Web as a delivery channel—just as any other—sets unique requirements for content presentation and organization. While document-based content management typically considers the content as an object to be presented via reader or editor software resembling a paper print, the WCM approach considers content units as a portion of Web site multi-frame layout. For defining content combinations and their positioning on the Web site, the WCM systems utilize *templates* containing placeholders for content units to be inserted.

Structured documents, such as XML (Bray et al., 2006), separate the content, its logical structure and visual layout from each other within documents by using markup delimiters. The logical structure is described by a schema such as a *document type definition* (DTD) or a *XML schema*. A schema defines the markup vocabulary and the structure for a class of XML documents. A great deal of contemporary research and utilization of XML for organizational content management has focused on the data-oriented use of XML, such as developing service-oriented architectures (SOAs), Web services, and markup languages for data exchange, such as HL7 for e-Health. For content management in organizations the structured documents provide means for document and data interoperability and unified and simplified maintenance procedures, and a standard format for data exchange between organizations. Although the tradition of structured documents is long, contemporary research on using XML (not XHTML) as a format for organizational content has remained scarce except for the field of electronic publishing (Fahrenholz-Mann, 1999) and utilization of XML in e-government (Salminen, 2005). Some other pieces of research on XML documents consider, for example, document type schema design (Jauhiainen & Honkaranta, 2006; Maler & El Andaloussi, 1996).

The benefits of using structured documents and related technologies have lately been recognized and adopted by both content management and WCM approaches, although the use of XML poses multiple challenges. There are systems and tools for managing XML documents as files or within databases. However, the support for XML in content management systems is varying or under development. Albeit many interoperability problems in systems and in service integration have been solved with XML, the hot topic is how to take advantage of the enhanced XML and custom schema support for document interoperability in new office software using XML as its native document format, such as Open Office (http://www.openoffice.org/) and Microsoft Office 2007 (http://msdn.microsoft.com/ office/understanding/xmloffice/tools/default.aspx). In addition, developers lack knowledge about research findings and practical experience on adaptations and implementations of domain-specific XML and related technologies in real-life organizations. This forms a novel and essential line of research for organizational content management.

CONVERGENCE OF CONTENT TYPES AND ORGANIZATIONAL WORK

We may identify at least two kinds of convergence taking place in organizational content management. First, the management of content is deeply intertwined with organizational work. Second, content as we know it is becoming exceedingly complex as it converges with different logical and physical entities to be managed.

Convergence of Organizational Work and Content Management

Despite there perhaps being one software for business (or workflow) process management and another for content management, the content and related organizational activities form a seamless, unified stream of actions from the user perspective. Yet, the requirements for managing business processes and content differ and consist of dissimilar entities. This article explores the components of content management and their relationship to organizational work, and the convergence of content types and metadata for content management.

The features of content management and its intertwined nature with organizational work are discussed via a framework consisting of the four components of content management: content, technology and systems, roles and processes (Honkaranta & Tyrväinen, 2005). Figure 1 illustrates the aforementioned framework for content management.

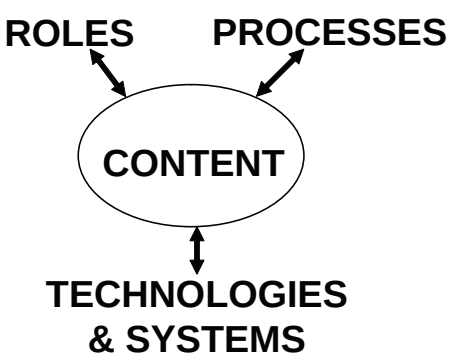
As presented by many researchers, the roles, processes, technologies and systems and content form a consistent whole in organizations. Therefore the components of the framework shown in the figure are interrelated. A change in one component will impact all other components of the framework. For example, adopting a content management system will impact the publication processes, the roles of the people, and the content units used (Honkaranta et al., 2005). A change in content form—such as transformation from paper to digital—may trigger the establishment of entirely new processes and roles in the organization (Eriksen & Ihlström, 2000).

For content management, people's *roles* should be separated with regard to content management development vs. operational roles. From the business process (operational) perspective a person may act as a manager, purchaser, accountant, and so on. From the content management perspective, the roles are attached to the content life-cycle. Thus, there are only a few roles to be considered. A person may act as a creator, modifier, viewer, signer, reviewer, publisher, and destroyer related to a content unit.

The operational business *processes* may be divided into two kinds: there are main processes related to the core of the organization's business and supporting processes. Traditionally, the core business processes have been supported by strategic information systems tailored for the organization, while the support processes have used standard systems. The technology for content management needs to be tailored to support the management of digital content in the main processes of an organization, where most of the communication volume appears (Tyrväinen, Kilpeläinen, & Järvenpää, 2005). As the business process consists of organizational activities, the "content process" concerns the activities leading to the change of content unit *state* during its life-cycle. These activities and states are typically similar to the roles discussed above, and allow few options (the following list has a notation of activity/state of a content unit): to create/a draft, to modify/a version, to review/a reviewed, to accept/an accepted, to publish/a published, to archive /an archived, to dispose/a disposed unit of content.

The *technology* of organizational content management includes standards, architectures, technologies, tools, and applications. Contemporary content management systems may be roughly divided into three categories: platform systems, horizontal systems, and vertical systems. Relational and XML databases and other generic *platforms* provide the base onto which content management applications and functionalities may be built on by custom software development projects. *Horizontal systems* are common-purpose document and content management systems, which already provide a base for managing content units as well as metadata and related processes and workflows. *Vertical systems* are systems that provide metadata and functionality related to the management of specific content units or for specific business processes. Examples of vertical systems are customer relationship management (CRM), supply chain management (SCM), and product data management (PDM) systems, as well as

Figure 1. Framework for organizational content management



patient information systems with specialized functionalities related to their subject area.

Convergence of Content Sources and Logical and Physical Content Units

The management of organizational content deals with the complexity of identifying and designing logical, physical and processable content units. *The logical units of content* are those identified by people according to their comprehension. *The physical units of content* to be managed may include a (virtual) folder of files, a file, or a field in a database. For example, a Web page containing text and multimedia files forms a (virtual) folder from the content management system perspective. The physical units of content may further contain *processable units of content,* such as XML elements.

The logical and physical content units in organizations may be analyzed by their *grain size* and *granularity* (Tyrväinen, 2003). The grain size of a logical content unit may vary from a single processable content element to content *aggregation* as a collection of documents. For example, a pension-system administrating organization may provide an aggregation of directives and guidelines related to the specific type of pension as a Web site folder for its customers (Honkaranta et al., 2005). A content unit considered as "a document" may be produced as *an assembly* of existing and novel content units, acquired from multiple content sources. For example, a budget of an organization may consist of various existing units of content—such as budgets of individual departments—which are combined with new content to form the budget of the whole company. While the grain size of the content units is increasing, the granularity is decreasing (Tyrväinen, 2003).

The amount of *metadata* related to the content increases along the number of the content units when the size of the managed unit decreases. Metadata is typically defined as "the sum total of what one can say about any information object at any level of aggregation" (Gilliland-Swetland, 1998). Commonly people associate metadata with content description for retrieval. It has almost become a norm to adopt a commonly used metadata standard, such as Dublin Core (1999), into a content management system. The metadata standard adopted defines the elements by which the content is described. Content may then be searched for and retrieved using the combination of metadata element names and their values specific to an instance content unit. Adoption of a metadata standard typically involves making organization-specific adjustments (e.g., Päivärinta, Tyrväinen, & Ylimäki, 2002).

A content management system utilizes metadata for a multitude of purposes. For example, a WCM system relies on metadata for automated publishing workflow and a CM system for the whole content workflow support. The role of metadata is increasingly essential as the sophisticated CM systems rely on metadata in virtually every aspect of content management: from access right management and workflow support to content creation and retrieval. Salminen (2005) has proposed categorizing metadata into three classes. Contextual metadata (such as metadata about user roles and access privileges) provides information about the context in which content units are created and used. Structural metadata provides information about the content units' relationships with respect to others (i.e., about the "has relationship," "is-a-part-of"—kind of hierarchical and non-hierarchical constructions), and semantic metadata describes the content with logical terms for search and retrieval.

With regard to content unit grain size, the WCM has a tendency to manage smaller units of content than those managed by content management systems. This means that WCM systems require extensive metadata for content assembly from smaller units of content and their management. Additional metadata also supports the reuse of content units but requires either much more manual work or a higher level of automation supported by sophisticated computerized management systems.

FUTURE TRENDS

The converging organizational content is produced from multiple heterogeneous data sources and delivered to multiple channels. Currently the Web dominates the research and will soon become the main delivery channel for content as well. Content creation includes the convergence of diverse systems, such as process and workflow management, content management, text and multimedia content production software, and Web Services, as well as the convergence of differing content types. As digitalization continues, we will witness an ever-increasing mix of combinations of speech, text, databases, spreadsheet data, and multimedia across and within the content units. Mobile and other handheld devices will gain popularity as secondary or perhaps also as alternative delivery channels into the Web. Since the Web as a delivery channel forced the use of smaller units of content, the new delivery channels may require the use of content units with an even smaller grain size than those used in the WCM, along with varying content formats and multiple variations of the content. For example, an organizational memorandum may be presented as text on the Web; the text may be automatically summarized by a natural-language processing system for mobile browsers, or transformed into speech for a mobile phone as requested by the user.

As the content convergence continues, the analysis and design of logical units of content in organizational settings becomes more complex. The theory of genres as prototypical models for communication (Yates & Orlikowski, 1992) provides a promising approach for identifying and analyzing logical content units in organizations. Thus far, genres

have been operationalized, for example, for identifying the content units in an organization and gathering metadata on them (Karjalainen, Päivärinta, Tyrväinen, & Rajala, 2000), as well as for analyzing the digitalization of organizational content (Tyrväinen et al., 2005).

Convergence of content and organizational work requires that process and workflow management, as well as other organizational systems, need to integrate with content management in a tighter and more sophisticated manner than before. There are a number of possibilities for technical integration of systems, most prominent of them being the utilization of Web services and service-oriented architectures allowing each system to be encapsulated from others and yet to interoperate. However, the system integration for content management requires that the metadata about content, activities, and businesses and operation logic should also be integrated for a consistent whole.

Metadata will be at the heart of research and development of content management. Extensive use of metadata is required for systems integration, supporting the workflows, and managing the ever-smaller units of content with varying forms. Convergence of enterprise taxonomies, corporate or business-level ontologies and metadata specifications across and in between organizations is required for the interoperability of systems and user groups.

Previously, organizations have commonly relied on standard metadata specifications for their semantic metadata, and on a set of system-specific or randomly defined metadata for other kinds of metadata. These standards may be appended by flexibly-built user-group developed folksonomies (Guy & Tonkin, 2006). A bundle of new and old technologies may be adopted to fulfill the requirements for advanced metadata management. Text databases have utilized text mining based on natural language processing, and text mining combined with ranking has been utilized in Web search engines. These technologies have emerged into organizations in the form of personal or desktop search engines, such as Google Desktop Indexing Software (http://desktop.google.com/) and Microsoft Desktop Search Engine (http://search.msn.com/ docs/toolbar.aspx?t= MSNT bar_CONC_About-SearchingYour Computer.htm). The CM systems will need to adapt to new kinds of hybrid retrieval modes combining metadata standard vocabularies and text processing. As people in organizations often need to find knowledgeable people instead of specific content, or they wish to look up resources by knowledgeable "authorities", the organizational content search may also be combined with metadata about persons or their resources. The aforementioned needs call for additional solutions, such as FOAF (Friend-of-a-Friend; a format for describing persons and their features), or joint book marking.

CONCLUSION

Content management has evolved from document management. Novel streams of research on the area focus on Web content management (WCM) and the utilization of structured documents in the form of XML documents. These approaches have not yet been consolidated into a uniform research area.

Characteristic to contemporary content management is a shift towards using units of content with varying grain sizes. A unit of content to be managed may be broader, similar, or smaller than a traditional document. Another focus of contemporary CM systems is the support for Web publishing. The Web as a publication channel is, however, different from document- or paper-based delivery. Web delivery enforces the use of content units smaller in grain sized than documents. The user interface is also different: instead of a document view, the WCM uses a site template which combines several units of content into a specific layout setting. Variance in content unit grain size and additional requirements for workflow support are reflected by an increasing variation and amount of metadata needed for managing and processing the content.

Content is deeply intertwined with organizational work. In the future, we may witness convergence of content units into new publishing channels such as mobile devices, and in ever-broadening formats, such as voice and new mixes of multimedia. Therefore, it may become difficult to identify the content units to be managed from the mix of publishing channels and media utilized for them. Research on content identification, analysis, and modularization will be needed to make sense of the content to be managed. Another avenue for further research is the management of new kinds of metadata for content management.

REFERENCES

Blair, D., C. (2002). The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. *Information Processing & Management, 38*(3), 273-291.

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yeargeau, F., & Cowan, J. E. (2006). *Extensible Markup Language* (XML) 1.1 (2nd ed.). W3C Recommendation 16 Aug 2006. Retrieved 1 Sept., 2006, from http://www.w3.org/TR/2006/REC-xml11-20060816/

DublinCore. (1999). *Dublin Core Metadata Element Set, Version 1.1*. Retrieved February, 17, 2000, from http://purl.org/DC/documents/rec-dces-19990702.htm

Eriksen, L. B., & Ihlström, C. (2000). Evolution of the web news genre—The slow move beyond the print metaphor. In R. H. Sprague (Ed.), *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS)* (pp. 10 pp.). Los Alamitos CA: IEEE Computer Society.

Fahrenholz-Mann, S. (1999). SGML for electronic publishing at a technical society— Expectations meets reality. *Markup Languages: Theory and Practice, 1*(2) (Spring 1999), 1-30.

Gilliland-Swetland, A. J. (1998). Defining metadata. In M. Baca (Ed.), *Introduction to metadata* (pp. 1-8). Los Angeles: Getty Information Institute.

Glushko, R. J., & McGrath, T. (2005). Document engineering: Analyzing and designing the semantics of business service networks. In *Proceedings of the IEEE EEE05 international workshop on Business services networks. ACM International Conference Proceeding Series; Vol. 87* (pp. 2-2). Piscataway, NJ, USA: IEEE Press.

Grossniklaus, M., & Norrie, M. C. (2002). Information concepts for content management. In *Proceedings of the Third International Conference on Web Information Systems Engineering (Workshop)*. IEEE.

Guy, M., & Tonkin, E. (2006). Folksomies. Tidying Up Tags? *D-Lib Magazine, 12*(1).

Honkaranta, A., Salminen, A., & Peltola, T. (2005). Challenges in the redesign of content management: A Case of FCP. *International Journal of Cases on Electronic Commerce, 1*(1), 53-69.

Honkaranta, A., & Tyrväinen, P. (2005). Content management in organizations. In M. Khosrowpour (Ed.), *Encyclopedia of Information Science and Technology* (pp. 550-555). Hershey: Idea Group Publishing.

Jauhiainen, E., & Honkaranta, A. (2006). A review on XML document schemas and methods for schema design. In *The Proceendings of the 9th International Conference on Business Information Systems (in cooperation with ACM SIGMIS)* (pp. 10). Klagenfurt, Austria.

Karjalainen, A., Päivärinta, T., Tyrväinen, P., & Rajala, J. (2000). Genre-based metadata for enterprise document management. In R. H. Sprague (Ed.), *Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS)* (HICSS Digital Library ed.). Los Alamitos CA: IEEE Computer Society.

Maler, E., & El Andaloussi, J. (1996). *Developing SGML DTDs. From text to model to markup.* Upper Saddle River, NJ: Prentice Hall.

McIntosh, M. (2000). *Content Management Using the Rational Unified Process®*. Rational Software White Paper (White Paper No. TP 164 09/2000). Cupertino, CA: Rational Software Corporation.

Medina, R., Meyers, S., Bragg, J., & Klima, C. (2002). *Doculabs evaluates document management for enterprise content management*. Retrieved April 2, 2003, from http://www.transformmag.com/db_area/archs/2002/02/tfm0202f1.shtml?contentmanagement

Murphy, L. D. (1998). Digital document metadata in organizations: Roles, analytical approaches, and future research directions. In R. H. J. Sprague (Ed.), *Proceedings of the 31st Annual Hawaii International Conference on System Sciences (HICSS)* (Vol. II, pp. 267-276). Los Alamitos, CA: IEEE Computer Society.

Murphy, L. D. (2001). Digital documents in organizational communities of practice: A first look. In R. H. Sprague (Ed.), *Proceedings of the Thirty-Fourth Hawaii International Conference on System Sciences* (CD ROM). Los Alamitos, CA: IEEE Computer Society.

Paganelli, F., & Pettenati, M. C. (2005). A model-driven method for the design and deployment of web-based document management systems. *Journal of Digital Information (JoDi), 6*(3), Article 360.

Päivärinta, T., Tyrväinen, P., & Ylimäki, T. (2002). Defining organizational document metadata: A case beyond standards. In *Proceedings of the Xth European Conference on Information Systems (ECIS). June 6-8, 2002* (pp. 10). Gdansk, Poland.

Rockley, A., Kostur, P., & Manning, S. (2003). *Managing enterprise content: A unified content strategy.* New Riders.

Salminen, A. (2003). Document analysis methods. In *Encyclopedia of Library and Information Science* (pp. 916-926). New York: Marcel Dekker, Inc.

Salminen, A. (2005). Building digital government by XML. In R. H. Sprague (Ed.), *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS)* (pp. 10). Los Alamitos, CA: IEEEE Computer Society.

Tyrväinen, P. (2003). Estimating applicability of new mobile content formats to organizational use. In *Proceedings of the 36th Hawaii International Conference on Systems Sciences.* (pp. CD ROM, 10 p.). Los Alamitos, CA: IEEE.

Tyrväinen, P., Kilpeläinen, T., & Järvenpää, M. (2005). Patterns and measures of digitalisation in business unit communication. *International Journal of Business Information Systems, 1*(2), 1999-1219.

Yates, J., & Orlikowski, W. J. (1992). Genres of organizational communication: A structurational approach to studying communication and media. *Academy of Management Review, 17*(2), 299-326.

KEY TERMS

Content Aggregation (noun/verb)**:** A set of existing content units collected together for a specific use purpose. An aggregation may contain several versions of the same unit of content, and its creation may require human involvement.

Content Assembly (noun/verb)**:** A collection of existing or new units of content which may be manipulated to produce content for a publication or for a specific target audience. May be produced (semi-)automatically or involve manual processing. A portion of training content for specialists only may be an assembly.

Content Seeding: Adding identifiers and metadata to content units or their parts to enable computerized assemblies and aggregations on the content.

Content Unit: The object with which the management metadata is associated. May be "a document," "a file," "a component," or "a section of a document" among others.

Metadata: Data describing the content from the viewpoint of humans or computers. Metadata may describe the domain in which content is used, as well as collections, classes, units or processable portions of the content and content instances. Metadata enables content search, classification, and processing. Whereas humans may use and interpret metadata about documents, the metadata about the small units of content is primarily meant for computerized manipulation of content.

Template: A combination of static content, references to units of existing content, and program code for producing the content and navigation aids for a Web site.

Web Content Management: Management of content intended primarily for web delivery. Grain size of the content unit is typically smaller (e.g., a page or a paragraph) than that of documents.

Managing IS Security and Privacy

Vasilios Katos
University of Portsmouth, UK

INTRODUCTION

The concept of privacy has received attention for over a century now and its definition—let alone, understanding—has been profoundly challenging. This is primarily attributed to the "incompatible" and rich set of characteristics privacy comprises. As Brunk (2002) states very sharply, "Privacy is a matter of intellectual and philosophical thought and retains few tangible characteristics, making it resistant to simple explanation."

Perhaps the first scholarly work on privacy was that of Warren and Brandeis (1980), who introduced the highly abstractive yet popular definition of privacy as the "right to be left alone." As privacy was recognized as a right, it primarily existed within a legal context. Legislation for protecting one's privacy exists in many countries and in some cases at a constitutional level (see for example the Fourth Amendment of the U.S. Constitution).

It was soon realized in the information revolution era that privacy and information are somewhat coupled. More precisely, emerging privacy concepts and metrics relate to the intentional or unintentional information flows. However, when it comes to studying, using, and investing in information, security appeared to have a higher priority over privacy. Security and privacy seemingly operate under different agendas; privacy is about protecting one's actions in terms of offering anonymity, whereas security includes the notion of accountability which implies that anonymity is waived. Still, security is a vital component of an information system, as it is well needed in order to protect privacy.

This contradictory relation between security and privacy has caused a considerable amount of debate, political and technical, resulting in a plethora of position and research papers. Accepting that there may be no optimum solution to the problem of striking a balance between security and privacy, this article presents a recently developed methodology that could support policy decision making on a strategic level, thus allowing planners to macro-manage security and privacy.

BACKGROUND

A thorough overview on the economics of privacy is maintained by Acquisti (2008). The 1970s was a decade marked by economists and their aspirations to develop an economic model to "decrypt" the market forces. Although Hirshleifer (1971) introduced the value of information in relation to privacy in the early 1970s, economics tools were ported to the privacy domain in the late 1970s and early 1980s (e.g., Posner, 1978; Stigler, 1980). However in the 1980s the concept of information sharing and the Internet were showing signs of potential, only to be interrupted by the Morris Worm in 1988 (Seeley, 1989), and security was added into the agenda. Initially this was done in the expense of privacy. For the following years information security received substantial attention—if the members of the private sector were to invest in electronic communications and technologies, trust needed to be restored.

Formal treatment of information security was initially in the domain of cryptography, but soon expanded to access control models and intrusion detection systems. The security goals of confidentiality, integrity, and availability were defined. The escape from security being equivalent to confidentiality was soon realized in the domain of cryptography, which was enforced with Rivest's (1990) definition of cryptography which "is about communication in the presence of adversaries." As such, the adversary would not necessarily be interested in eavesdropping on a communication, but could elect to interrupt, modify, fabricate, or replay messages. Formally, this omnipotent adversary was initially captured in Dolev and Yao's (1981) threat model, spawning research into cryptographic protocols.

To date, the body of knowledge for information security has fairly matured. The security domains include both technical and organizational aspects. Standards and methodologies emerged—see for example BS 7799 and ISO/IEC 17799 (BSI, 1995a, 1995b), ISO 27001 (ISO, 2005, aligning with BS 7799 part 3), and CobiT (IT Governance Institute, 2007). It can be seen from the directions taken by these standardization efforts that information security management was becoming an isomorphism of risk management: understanding that there is no absolute security, controls need to be in place in order to diversify the risks of unauthorized disclosure (breach of confidentiality), unauthorized modification (breach of integrity), and denial of service (breach of availability), accepting that there is an amount of residual risk that will be present after employing the security controls.

Research on privacy followed at a much slower pace. It could be argued that a valid reason for this is that privacy is upper bounded by security; security needs to be in place in order to offer privacy. Indeed, some security technologies

such as cryptography were branded as privacy enhancing technologies (PETs), emphasizing the synergetic relationship between security and privacy. As the number of privacy violations and intrusions was steadily increasing in the 1990s (Acquisti, 2008), research on privacy gained momentum. Similar to the security goals stated earlier, the privacy criteria of unobservability, pseudonymity, unlinkability, and anonymity (ISO, 1999; Fischer, 2001) were introduced. With respect to the economics of privacy, the work by Laudon (1996), Varian (1996), Huang (1998), and Posner (1999) set precedence leading to research in the formal application of micro-economic techniques to analyzing privacy. Representative work on the formal application of micro-economics on privacy was published by Acquisti (2004; Acquisti, Dingledine, & Syverson, 2003), Otsuka and Onozawa (2001), and Ward (2001).

However, it was realized by Katos and Patel (2008) that a micro treatment of privacy would be applicable in establishing operational management procedures, yet it had major limitations when attempting to understand the challenges in balancing security and privacy when engaging in policy-making activities. The authors argued that a detailed (micro) view of privacy and security would make it virtually impossible to track or predict the outcome of a policy decision, and suggested that a higher-level—or aggregate—view should be adopted. In fact, Odlyzko (2003) conjectured that the privacy problem is intractable. An analogy could be drawn with the stock market environment: although trends could be established on a macro level for the performance of a certain market, the actual assessment and prediction of the micro variables (stocks) would be substantially more challenging, error prone, and less informing.

The following section presents a methodology developed as a response to the shortcomings of the micro views on security and privacy. The section summarizes the main points of the model. For a more detailed explanation, the reader is referred to Katos and Patel (2008).

A MACRO TREATMENT OF INFORMATION SECURITY AND PRIVACY

Initially we accept that there is no universally accepted, objective measure for privacy. As privacy applies not only to the data, but also to the user's actions as he or she interacts with any given system, we can consider a space of events that could be expressed by a set as follows:

$$A=\{stay\_home, go\_shopping, use\_credit\_card, mortgage\_application, \ldots\}$$

If a metric $p(\ )$ on privacy existed, mapping the above set into a formal range, we could argue that as a very basic requirement, the metric would be on an ordinal scale, for example:

$$p(stay\_home) \leq p(go\_shopping) \leq p(use\_credit\_card) \leq p(mortgage\_application)$$

This would be a minimum requirement for this hypothetical metric to be meaningful—or fulfill the representation condition as expressed in the area of measurement theory (Fenton & Pfleeger, 1998). It should be obvious that the exact initial level as well as change of privacy cannot be established. This is not only because privacy is qualitative and perhaps subjective, but also because we have no control or knowledge of all variables affecting it. For example, how many monitoring technologies (such as CCTV elements) have invaded our private space, and to whom is the captured data available? For how long? What is the quality of the captured data and therefore the likelihood of positive identification? What protection does the legal system provide against third-party enquiries to access the data? It can be seen that not only the number and diversity of these questions can be exceedingly high, but also answering them is challenging in principle.

Determining qualitative variables in uncertain and open problem domains has been a major topic of interest in the discipline of macroeconomics. The well-known so-called cross methodology has significantly contributed to the understanding of the market forces of supply and demand (Dornbush & Fischer, 1998; Branson & Litvack, 1981). The remainder of this section deals with porting these proven techniques to the domain of privacy and security.

Initially we need to classify the relevant technologies in two "markets": the security technologies and the adversarial technologies. By security technologies we mean those that intend to support the confidentiality of our private data, such as firewalls, antivirus tools, and so on. In other words, these are defensive technologies and primarily access control measures. By adversarial technologies we mean those technologies that are used for testing our security technologies. These are hacking tools, such as vulnerability scanners, exploits, security assessment frameworks. These offensive security mechanisms are required in order to be able to assess the security level of an IT infrastructure. A key differentiator is the purpose or intention of use of a certain technology. In the security technologies market, the technologies can only be used for benign purposes, whereas in the adversarial technologies market, the technologies can be used for either benign or malicious purposes. "Ethical hacking" for instance is the term used for capturing the benign use of the adversarial tools.

Against the above we can now proceed in defining the two markets. Figure 1 shows the security technologies market. The process for establishing the respective relations is as follows. Our objective is to determine the relationship between price and privacy (i.e., quadrant Q1) in this market. To do this, we need to define Q2, Q3, and Q4; then, Q1 will be defined as an equilibrium by attaching all assumptions (functions) to the other three quadrants.

Starting with Q2, we assume that there is an inverse relationship between the aggregate demand of the security technologies and price—that is, the lower the price of security technologies (P), the higher the quantity demanded of security technologies (SD).

The assumption captured in Q4 shows that there is a positive relationship between security technologies supply and level of privacy. Indeed, we support the view that there can be no privacy if there are no security technologies in place to protect the relevant personal information (i.e., privacy is bounded by security). Hence the supply for security technologies function SS = *g*(V) is rationalized by the fact that the more important (higher) privacy (V) is the higher quantity supplied of security technologies (SS) to keep privacy at high levels.

Quadrant Q3 reflects the economists' views who suggest that market forces and economic laws, if left alone, will eventually push security technologies demand to equilibrium with security technologies supply, regardless of their initial allocation. This is represented by the identity function SD=SS, or *f*(P) = *g*(V).

By attaching all three assumptions in Q2, Q3, and Q4, we can derive the curve in Q1. Consider price P_0, which corresponds to demand SD_0, which in turn matches supply SS_0, which in turn rests at privacy P_0. Similarly, consider the path for a different price P_1, leading to demand SD_1, matching supply SS_1, resulting in privacy P_1. The two variable pairs (P_0,V_0) and (P_1,V_1) define the two points of the privacy-price curve A_0 and A_1 respectively. If this is done for the infinite number of (P,V) pairs, we eventually obtain curve SS-SD.

A similar line of reasoning is followed for the adversarial technologies market. Although Q2 and Q4 remain the same as in the security technologies market, the attention should be drawn to Q3, which in this case is substantially different (see Figure 2). More specifically, Q3 captures the assumption relating to the use or intention of the technology. Assuming that an adversarial system may be used either for benign or for malicious purpose, SS*=SM+SB would denote the total number of systems used, where SM is the total number of systems used for malicious purposes and SB is the number of systems used for benign purposes. The line drawn in Q3

Figure 1. The security technologies market

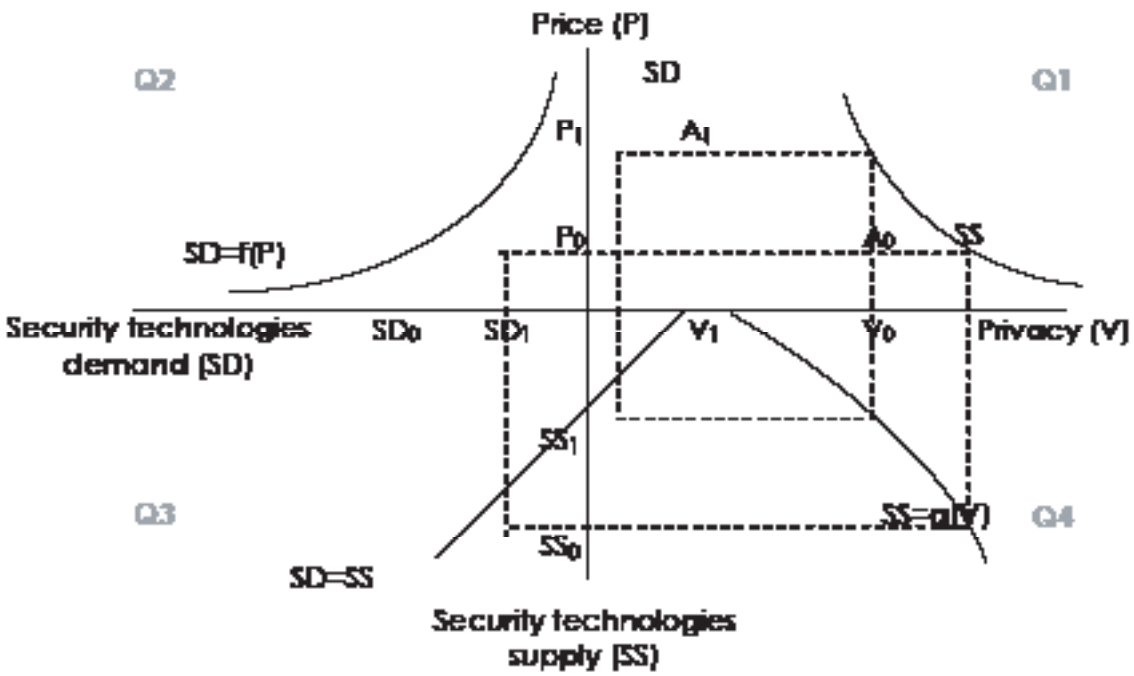
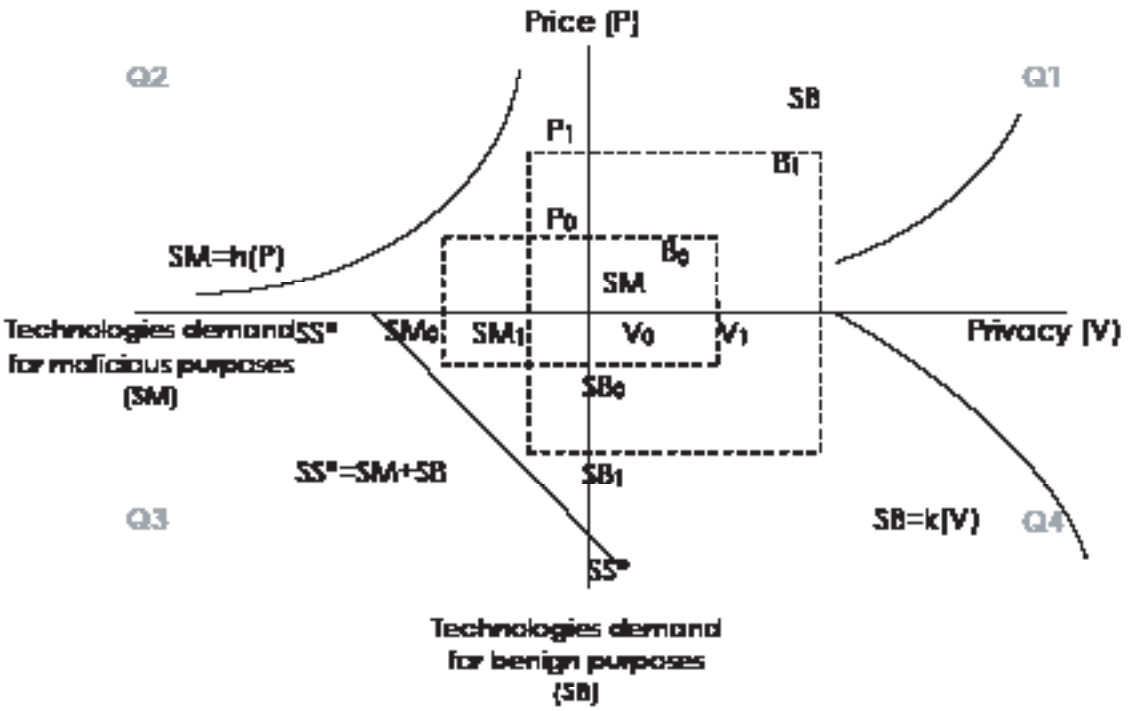


Figure 2. The adversarial technologies market



shows this as a constraint, and any point on this line sets the proportion between malicious and benign systems: due to the geometric nature of the -45º line, the two components of demand always add up to the total supply on each axis, so that the -45º line directly represents the equilibrium condition. Any point on this -45º line gives a demand for a malicious purposes technologies component plus a demand for a benign or privacy-enhancing security technologies component, which just add up to the total security technologies supply.

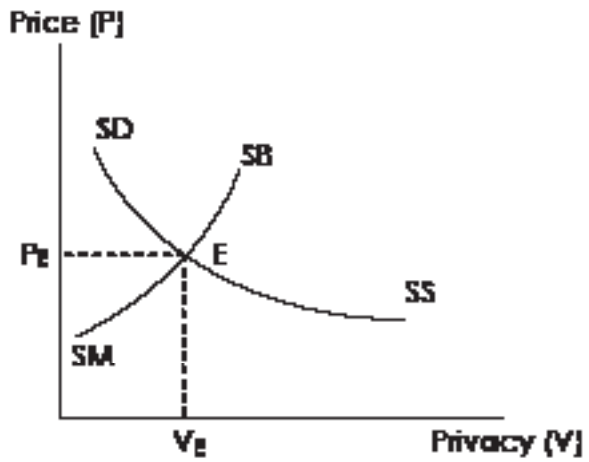
Following the process of identifying the (P,V) pairs by equating all assumptions in Q2, Q3, and Q4, we obtain the positively sloped SM-SB curve in Q1, which represents the technologies of the adversary market.

So far we have derived two pieces of geometric equipment. One gives the equilibrium pairs of P and V in Figure 1, that is, the SD-SS curve in the security technologies market, and the other gives the equilibrium pairs of P and V in Figure 2, that is, the SM-SB curve in the technologies of the adversary market. By placing these two curves on the same quadrant shown in Figure 3—that is, by solving the two equilibrium equations $f(P) = g(V)$ and $SS^* = h(P) + k(V)$ simultaneously—we can find the single (P, V) pair that gives equilibrium in both markets. This is shown as the equilibrium point $E(P_E, V_E)$ of intersection of the SD-SS and SM-SB curves in Figure 3.

Alone, this "snapshot" of the expected level of privacy is not very helpful, as we have still not escaped from the need to objectively measure privacy. However, the benefits of this method can be realized if we performed a comparative statics exercise.

Consider for example the case where a government decided to update a regulatory system by introducing stricter laws to hacking, that is, the malicious use of adversarial technologies. In this case, we can accept that the demand for security technologies for malicious purposes decreases, as indicated in Figure 4 by a right offset of the demand curve, from position $SF = h(P)$ to position $SF=h'(P)$. As a result it is shown, following the dashed lines in Figure 4, that curve SF-SG shifts to the right to position SF´-SG´ and thus the equilibrium point moves along the SD-SS curve from point E to point E´. Comparing points E and E´ (comparative statics), it is seen that the equilibrium price decreases from level PE to PE´ and privacy increases from level VE to level

Figure 3. The equilibrium

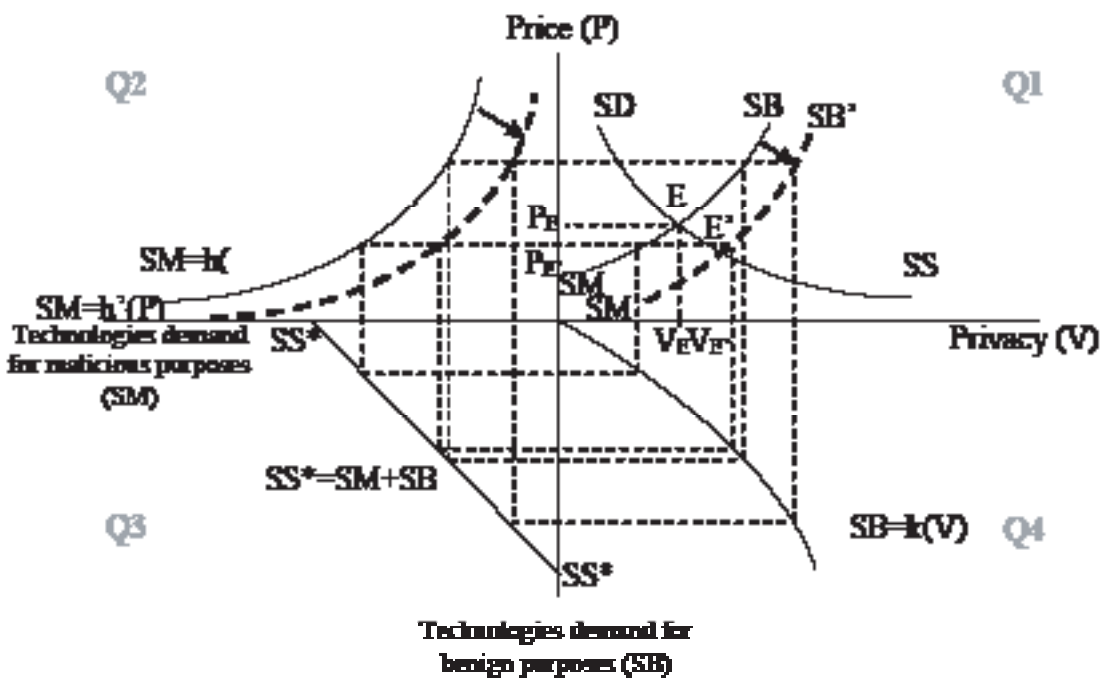


VE´. In other words, under the assumption of exogenously fixed supply of security technologies, the introduction of a stricter regulatory system results in a reallocation of security technologies demand between fraud and privacy-enhancing purposes (decreasing for malicious purposes and increasing for benign purposes), in lower levels of prices for security technologies, and correspondingly in higher levels of privacy. Intuitively the response from this model is correct. Although the model does not contribute to any knowledge on the objective and actual level of privacy (and price), it equips policymakers with enough information to allow them to make informed decisions on modifying or introducing new privacy and security policies.

FUTURE TRENDS

The macro view of security and privacy presented above seems to be a promising tool towards understanding the apparently complex and dynamic relationship between security and privacy. As this tool has been ported from the

Figure 4. The new equilibrium, following an introduction of a strict law against hacking



discipline of macroeconomics, its limitations and drawbacks are well known and have also been inherited. It is expected therefore that new models will be developed to overcome these problems.

It seems unreasonable to attempt a creation of an objective privacy metric, or measure privacy by absolute means. On the contrary, specialized hypothesis testing tools capable of running *what if* scenarios and simulators measuring privacy and security in a differential manner appear to be more realistic and result in a greater added value in managing security and privacy. Such an approach would provide adequate research space for a formal treatment of security and privacy.

CONCLUSION

The symbiotic yet apparently contradictory relationship between security and privacy has posed many challenges to the already rich in terms of characteristics concept of privacy. Starting from abstract and straightforward definitions of privacy, formal micro economic tools were introduced, arriving at a macro treatment in analyzing security and privacy. Nowadays strategic-level decision making is performed on a regular basis, affecting security, privacy, and their balance. It would be imperative therefore to work towards developing a tool for enabling policy planners to make informing decisions on societal, legal, and technological aspects influencing the delicate balance of security and privacy, and this article advocated a tool for doing so.

REFERENCES

Acquisti, A. (2004). Privacy in electronic commerce and the economics of immediate gratification. *Proceedings of the ACM Electronic Commerce Conference* (EC 04) (pp. 21-29). New York: ACM Press.

Acquisti, A. (2008). *The economics of privacy.* Retrieved January 11, 2008, from http://www.heinz.cmu.edu/~acquisti/economics-privacy.htm

Acquisti, A., Dingledine, R., & Syverson P. (2003). On the economics of anonymity. *Proceedings of the Conference on Financial Cryptography* (FC ''03). Berlin: Springer-Verlag (LNCS).

Branson, W.H., & Litvack, J.M. (1981). *Macroeconomics* (2nd ed.). New York: Harper & Row.

BSI. (1995a). *Information security management systems—specification with guidance for use.* British Standard Publication BS 7799, Part 2, British Standards Institute, UK.

BSI. (1995b). *Information technology—code of practice for information security management.* British Standard Publication BS 7799, British Standards Institute, UK.

Brunk, B. (2002). Understanding the privacy space. *First Monday, 7*(10).

Dolev, D., & Yao, A. (1981). On the security of public key protocols. *Proceedings of the IEEE 22nd Annual Symposium on Foundations of Computer Science* (pp. 350-357).

Dornbush, R., & Fischer, S. (1998). *Macroeconomics* (7th ed.). New York: McGraw-Hill.

Fenton, N., & Pfleeger, S. (1998). *Software metrics: A rigorous and practical approach.* Boston: PWS.

Fischer-Hubner, S. (2001). *IT security and privacy: Design and use of privacy enhancing security mechanisms.* Berlin: Springer-Verlag (LNCS 1958).

Hirshleifer, J. (1971). The private and social value of information and the reward to inventive activity. *American Economic Review, 61,* 561-574.

ISO. (1999). *Information technology—security techniques—evaluation criteria for IT security.* ISO/IEC 15408-2, International Organization for Standardization, Switzerland.

ISO. (2005). *Information technology—security techniques—information security management systems.* ISO/IEC 27001:2005, International Organization for Standardization, Switzerland.

IT Governance Institute. (2007). *Control objectives for information and related technology—CobiT 4.1.* Rolling Meadows, IL: Author.

Huang, P. (1998). *The law and economics of consumer privacy versus data mining.* Retrieved from http://ssrn.com/abstract=94041

Katos, V., & Patel, A. (2008). A partial equilibrium view on security and privacy. *Information Management and Computer Security,* (to appear).

Laudon, K. (1996). Markets and privacy. *Communications of the ACM, 39*(9).

Odlyzko, A. (2003). Privacy, economics, and price discrimination on the Internet. *Proceedings of the ACM 5th International Conference on Electronic Commerce* (pp. 355-366).

Otsuka, T., & Onozawa, A. (2001). Personal information market: Toward a secure and efficient trade of privacy. *Proceedings of the 1st International Conference on Human Society and the Internet* (p. 151). Berlin: Springer-Verlag (LNCS 2105).

Posner, R. (1978) An economic theory of privacy. *Regulation,* 19-26.

Posner, R. (1999). *Orwell versus Huxley: Economics, technology, privacy, and satire.* John M. Olin Law & Economics Working Paper No. 89, University of Chicago Law School, USA. Retrieved from *http://ssrn.com/abstract=194572*

Rivest, R. (1990). Cryptography. In van Leeuwen (Ed.), *Handbook of theoretical computer science.* Elsevier Science.

Seeley, D. (1989, February). A tour of the worm. *Proceedings of the 1989 Winter USENIX Conference,* San Diego, CA.

Stigler, G. (1980). An introduction to privacy in economics and politics. *Journal of Legal Studies, 9,* 623-644.

Varian, H. (1996). *Economic aspects of personal privacy.* Retrieved from *http://people.ischool.berkeley.edu/~hal/Papers/privacy/*

Ward, M. (2001). The economics of online retail markets. In G. Madden & S. Savage (Eds.), *The international handbook on emerging telecommunications networks.* Edward Elgar.

Warren, S., & Brandeis, L. (1890). The right to privacy. *Harvard Law Review, 4*(5), 193-220.

KEY TERMS

Access Control: All security processes and technologies that are responsible for determining and managing legitimate user access to data and system resources.

Adversarial Technologies: The technologies used in offensive security, such as hacking, penetration testing, and so forth.

Anonymity: The privacy goal of the inability to identify a user's identity when that user is performing an action or using a resource of a given system.

Pseudonymity: The privacy goal of hiding a user's identity by disguise, through the use of a pseudonym.

Side Channel: The unintentional flow of information through a probabilistic communication channel which facilitates information inference (or leakage).

Unlinkability: The privacy goal of allowing a user to perform multiple actions without others being able to link these actions together.

Unobservability: The privacy goal of allowing a user to perform an action or use a system resource without others being able to observe that the resource is being used.

Managing Organizational Knowledge in the Age of Social Computing

V. P. Kochikar
Infosys Technologies Ltd., India

INTRODUCTION

Technology, since the days of the Industrial Revolution, has been used by large corporations, such as factories and the railways, to great advantage. Starting around the end of the 19th century, technology began to be used directly by the consumer, but remained essentially a means of satisfying a *personal* need, such as lighting or listening to music. In the past decade, as technologies such as e-mail, Web, Weblogs (blogs), Wikis, and instant messaging have become pervasive, the way technology is used by individuals has changed—it has increasingly been put to use to meet *social* needs, such as interaction, sharing, and networking. This new paradigm of technology use, and the technologies that have enabled it, may be termed social computing.

By its very nature, social computing facilitates the sharing and leveraging of knowledge residing within a community of people. In this article, we discuss how social computing can act as the primary mechanism that enables the management of knowledge within an organization.

BACKGROUND: THE DISCIPLINE OF KNOWLEDGE MANAGEMENT

There are several ingredients that go into organizational success, and leveraging assets well is one of these. As intangible assets represent a rising proportion of total assets, they have come to represent an important area of management focus. The discipline of *knowledge management* (KM) thus encompasses the organizational activities directed toward the assimilation, dissemination, harvest, and reuse of knowledge. In simpler terms, KM is the answer to the question, "How can the organization update and use its knowledge more effectively?" (Kochikar, 2000).

Some of the world's most successful organizations, be they corporate, academic, or government, invest considerably in KM, and substantial benefits have been reported across industries (Berkman, 2001; Frappaolo, 2006; Kochikar & Suresh, 2005).

Knowledge Management Review magazine's survey of 400 global corporations revealed that the following are key objectives of KM programs (KM Review, 2002):

a. Increasing organizational communication
b. Gaining competitive advantage
c. Increasing collaboration among employees
d. Improving customer relationships
e. Raising efficiency
f. Innovating
g. Learning from mistakes and successes
h. Capturing and retaining tacit knowledge

Using the framework of Nahapiet and Ghoshal (1998), these objectives can be classified as improving *financial capital* (b, e); improving *social capital* (a, c, d); and improving *intellectual capital* (f, g, h).

Each organization must fashion a KM strategy that takes cognizance of its unique competencies, aspirations, and business context. Mechanisms for organizational KM typically take the form of setting up strongly engineered governance mechanisms, focusing on four key aspects: people, processes, technology, and content (see, e.g., Kochikar, Mahesh, and Mahind, 2002).

As an exemplar, Infosys Technologies (*NASDAQ: INFY*) has had a KM program since 1999, which aims to *empower every employee with the knowledge of every other employee*. Key elements of the KM architecture include the *Knowledge Currency Unit* scheme, a comprehensive mechanism for reward, recognition, and measurement of KM benefits; *KShop*, the corporate knowledge portal built in-house; and the *knowledge hierarchy*, a four-level taxonomy of over 2000 subject areas that constitute knowledge in the Infosys context (Kochikar et al., 2002).

For more on KM and its organizational uses, see work by Davenport and Prusak (1998), Drucker, Garvin, andLeonard (1998), Nonaka and Ichijo (2006), and Nonaka and Takeuchi (1995).

BACKGROUND: SOCIAL COMPUTING COMES OF AGE

Social computing is the name given to a slew of technologies that collectively allow people to pool their knowledge, keep in touch with, and interact better with others who belong to their community.

The stellar rise in the popularity of e-mail in the 1990s (the number of users skyrocketed from a few thousand at the beginning of that decade, to several hundred million at the end of it) clearly provides a pointer to the potential that social computing has—people are eager to take up technologies that will help them meet their social needs better. For example, there are as of May 2006 a total of 39 million blogs worldwide, with 75,000 being added each day (Klein, 2006). In an academic/research sense, social computing is a relatively new field—a fact reflected in the relative paucity of books and research papers in the reference section of this article.

What has spurred this gain in the importance of social computing? While there are several reasons, two in particular stand out:

- The steady march of advances in computing that have put more computing power in the hands of the users, allowing them to use it to achieve ends that they truly consider useful;
- Network effects as encapsulated in Metcalfe's Law: As the number of users of a particular technology that supports interaction or networking increases, the benefits perceived by all users accelerate significantly, causing even more users to adopt the technology.

Technologies that commonly go by the name of social computing include e-mail,instant messaging (IM), blogs, wikis, podcasting, and really simple syndication (RSS). They also include Web sites or portals supporting a variety of social interactions (examples include Yahoo!, Myspace, Flickr, del.icio.us).

A key sign of the coming of age of a new technology bubbling up from the masses is large corporations taking note of that technology. In the common view, technology diffuses by a "trickle-down effect," that is, a new technology first finds use within large corporations and then, as it becomes more affordable, trickles down to smaller businesses and finally becomes inexpensive enough to be used by the individual consumer. While such a top-down view is valid, it hardly represents the sole mechanism of technological diffusion. Equally, technology diffuses bottom-up too (Kochikar, 2006). The Internet was for decades used almost exclusively by researchers, then by academics, and subsequently (in the early 1990s) by individuals for publishing information using personal Web sites and so forth. Even when business uses were discovered for the Internet, it was small startups such as Amazon that leveraged it best—large corporations were in many ways the last to embrace the Internet. The same pattern can be seen with e-mail, instant messaging, gaming (which began with children and teenagers and is now finding uses in business such as for strategy formulation), and several other technologies. Other examples can be found in Kochikar (2006), which enumerates a few simple pointers for foreseeing emerging technologies that are "below-the-radar." Thus, large corporations must routinely monitor technologies that have not yet become visible on the corporate radar—that is, in use with small businesses, researchers, or individuals—or else they may miss an important source from where new technologies emerge. Social computing represents precisely such a "below-the-radar" technology.

Social computing is now beginning to find uses within large corporations and has elicited considerable enthusiasm from early adopters (BusinessWeek, 2006; McAfee, 2006).

SOCIAL COMPUTING: A NEW BACKBONE FOR ORGANIZATIONAL KM

Two key principles of social computing (or social software) are that

- It is highly participatory, or allows rich interaction between diverse and possibly dispersed members of a community, and
- It is evolutionary, or supports means for constant updating by the members of the community.

Together, these two characteristics indicate a mechanism for the collaborative creation and updating of content that constantly moves in such a direction as to better reflect the knowledge, beliefs, opinions, and/or aspirations of the community. This is precisely the goal of organizational KM—leveraging the combined knowledge of the organizational community.

To wit, a great deal of what has been learned and practiced by KM thinkers and practitioners over the past few years is finding expression now in the traction that social computing is getting. There has been recognition that social computing technologies can facilitate a new approach to KM. Say Caldwell and Linden (2004, p. 1):

Personal knowledge networking and social networks give individual knowledge workers direct control over the enterprise's intellectual capital and enable a new 'grassroots' approach to knowledge management. KM can happen without a lot of explicit governance.

While conventional KM systems often act as an additional "layer" on top of existing business processes and require people to devote time specifically for creating shareable content, or making existing content shareable, social computing technologies are more organic and integrate naturally into people's work habits or social needs. Harvard Professor Andrew McAfee writes (McAfee, 2006, p. 21):

There is a new wave of business communication tools including blogs, wikis and group messaging software—which may be dubbed, collectively, Enterprise 2.0 — that allow for more spontaneous, knowledge-based collaboration. These new tools may well supplant other communication and knowledge management systems with their superior ability to capture tacit knowledge, best practices and relevant experiences from throughout a company and make them readily available to more users.

Social computing technologies can help in a variety of organizational knowledge-sharing needs, such as

- Sharing useful materials, viz., documents, presentations, plans, and best practices
- Expressing opinions, reaching out, and getting feedback
- Finding experts in a specific domain
- Learning—sharing lessons learned, tips, and tricks
- Forming communities of like-minded individuals interested in a common activity, or sharing a common area of expertise; for example, a community dedicated to innovative product design may be formed
- Pooling knowledge about a product's market potential
- Finding products and services by means such as collaborative filtering

We illustrate with two or three sample technologies.

Wikis, which are one major component of social computing technologies, can find the following uses in the corporate activity of managing knowledge:

1. **Collaborative Publishing:** Creating documents that need input from multiple authors and reviewers, for example, user documentation for a product.
2. **Capturing Evolutionary Learning:** Creating and maintaining documents or reference material that change often and need to reflect knowledge being acquired on an ongoing basis. Examples include
   - Capturing tips and tricks on how to use a particular technology product or platform
   - Recording experiences of working in, or selling to, a new country
3. **Communication on Shared Concerns, Issues:** Providing a single location where a group of people can pool their views, opinions, and ideas on a concern that is common across the group, for example, how a new initiative can be designed to deliver results efficiently.

Blogging, another archetypal social computing technology, is essentially a "one-to-many" means of communication, while wikis are "many-to-many." Thus, while blogging can fulfill to a degree the three classes of use outlined previously, the quintessential nature of wikis, viz., the ability to capture multiple voices and multiple experiences easily and effectively, is absent. On the other hand, the ability to achieve more personal and individualistic ends, such as projecting one's expertise in a particular area, or enhancing one's credibility as an authoritative commentator on a specific topic or subject, is strongly supported by blogging.

A technology not commonly included in the list of social computing technologies is that of prediction markets. Since these are a mechanism for collecting and aggregating the opinions of a large number of people and making the output useful to the entire community, they can be seen as a form of social computing technology. And since diverse people form their opinions of the outcome based on the knowledge/expertise available to them, these markets, at a fundamental level, form a mechanism to bring together knowledge from across the organization—a goal fully consonant with that of KM. Sure enough, these are used at, for example, Microsoft (Kaihla, 2006) and Google (Google, 2005) to predict product launch dates by aggregating opinions of a large community of developers. Employees from across the company contribute knowledge and opinions, which are aggregated into a forecast by the market (Google, 2005). Eli Lilly Co. uses prediction markets to forecast the success of candidate drugs in the research pipeline (Giles, 2005). Interestingly, prediction markets have been found to be enormously accurate—often more so than conventional forecasting methods (Wolpers & Zitzevitz, 2006).

Can a social computing-based KM approach support KM programs at high maturity? Perhaps what social computing technologies do best within the organizational context is to allow content and human interaction to come together judiciously in carrying out organizational tasks. Referring to the five-level knowledge management maturity model (KMM) (Kochikar, 2000), one finds that two key result areas (KRAs or building blocks) at the highest levels of KM maturity—level 4 (Convinced) and level 5 (Sharing)—are *Content Enlivenment*, where content can be said to be truly "enlivened" with human expertise, and *Expertise Integration*, the notion that appropriate expertise is available to help understand content and tailor it to specific need. Kochikar (2000) states that,

Expertise integration represents the highest level of maturity of the sharing process, as true sharing requires a judicious mix of synchronous and asynchronous mechanisms, to achieve significant gains with optimal utilization of experts' time.

Social computing technologies, by acting as a mechanism to stitch together content and human interaction, facilitate both content enlivenment and expertise integration admirably, and are thus fully capable of supporting KM programs at high maturity.

CAVEATS

Most social computing technologies, by their very nature, involve the members of a community directly accessing and modifying content. Thus, they assume a certain degree of access to, and comfort with, computers, and so organizations/groups where the members do not meet this requirement may not find this technology very productive.

Accountability for content is an issue to be addressed. Accountability is direct in the case of some technologies such as blogging, as a blog typically is written by one person. However, with the appropriate security and auditing mechanisms, accountability can be very high with technologies that allow collaborative content creation, such as wikis, too.

Social computing also takes away a significant degree of responsibility, and hence control, from centralized corporate groups and puts it in the hands of a broader community. It thus engenders a mindset change in organizational governance, without which it may be difficult to absorb effectively.

Also, since a key tenet of social computing is trust, a culture of transparency is important.

FUTURE TRENDS

Considerable work is going into getting social computing technologies to better support various aspects of organizational knowledge sharing. One key area of focus is studying how groups and organizations "remember," and attempting to build systems to support various types of structured online activity such as design. The socially translucent system approach to KM, predicated on the notion that knowledge is discovered, shared, and used in a social context, is perhaps representative of the direction in which organizational KM systems will evolve (IBM, 2006).

Projects such as Cyc (http://www.cyc.com/) and MIT's Open Mind Project (MIT, 2006) also typify approaches for better organizing human knowledge and making machines more "aware," albeit in a societal rather than organizational context. However, the principles are valid in the organizational context, and likely to gain acceptance for managing organizational knowledge.

The Economist Intelligence Unit (Economist, 2006) identifies knowledge management as one of the five trends that will shape business and economy in the coming 15 years. However, in the era of social computing, corporate knowledge management systems will grow "lighter," ceding some of the hitherto centralized responsibility to the community.

CONCLUSION

In the final analysis, social computing is emblematic of the steady power shift that technology itself embodies—from governments, large institutions, and organizations to communities and the individual. Nobody decides on behalf of the user—the user decides what works. This stunningly simple principle is what has made social computing a phenomenon so powerful that it can be termed a profound social mega trend.

As this article illustrates, social computing is increasingly finding use in corporate environments, and is in the process of gaining traction as a mechanism for KM. This can only be a positive development, as ultimately the most effective way of ensuring that content represents what users want is to put the users in charge of managing it.

NOTE

Opinions and views expressed are personal and should not be taken to reflect those of the employer.

REFERENCES

Bartlett, C. A. (1998). *McKinsey & Company: Managing knowledge and learning. Case.* Cambridge, MA: Harvard Business School.

Berkman, E. (2001). When bad things happen to good ideas. *Darwin Magazine*, April. Retrieved July 31, 2006, from www.darwinmag.com/read/040101/badthings_content.html

Bontis, N., Crossan, M., & Hulland, J. (2002). Managing an organizational learning system by aligning stocks and flows. *Journal of Management Studies, 39*(4), 437-469.

Businessweek. (2006). CEO guide to technology. Retrieved July 31, 2006, from http://www.businessweek.com/technology/ceo_guide/

Caldwell, F., & Linden, A. (2004). PKN and social networks change knowledge management. Gartner Research. Retrieved July 31, 2006, from http://www.gartner.com/DisplayDocument?doc_cd=124178

Davenport, T. H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston: Harvard Business School Press.

Drucker, P. F., Garvin, D., Leonard, D. (1998). *Harvard business review on knowledge management* (1st ed.). Cambridge, MA: Harvard Business School Press.

Economist. (2006). *Foresight 2020: Economic, industry and corporate trends.* The Economist Intelligence Unit, London.

Frappaolo, C. (2006). *Knowledge management* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Giles, J. (2005). Wisdom of the crowd. *Nature, 438*(7066), 281.

Google. (2005). *Putting crowd wisdom to work*. Retrieved July 31, 2006, from http://googleblog.blogspot.com/2005/09/putting-crowd-wisdom-to-work.html

Gurteen. (2003). The Gurteen knowledge Web site. Retrieved July 31, 2006, from http://www.gurteen.com/gurteen/gurteen.nsf/0/17B666B9EE45086B80256CD500474AF0/

Hjelt, P. (2003). The world's most admired companies. *Fortune, 147*(3), 24-33.

IBM. (2006). *IBM research social computing page.* Retrieved July 31, 2006, from http://www.research.ibm.com/SocialComputing/AR.htm

Kaihla, P. (2006). Best kept secrets of the world's best companies. *Business 2.0 Magazine,* (March). Retrieved July 31, 2006, from http://money.cnn.com/magazines/business2/business2_archive/2006/04/01/8372806/index.htm

Klein, K. E. (2006). Does your small business need a blog? *Business Week,* (May). Retrieved July 31, 2006, from http://www.businessweek.com/smallbiz/content/may2006/sb20060515_027053.htm

KM Review. (2002). *KM Review: Industry survey*. London: Melcrum Publishing.

Kochikar, V. P. (2000, September 13-15). The knowledge management maturity model—A staged framework for leveraging knowledge. In *KMWorld 2000 Conference,* Santa Clara, CA. Retrieved November 9, 2006, from http://www.infotoday.com/KMWorld2000/presentations/default.htm

Kochikar, V. P. (2002). Creating the KM infrastructure at Infosys: The technology challenge. *IIMB Management Review, 13*(4), 104-110.

Kochikar, V. P. (2006, January 30). Re-engineering the crystal ball: Overcoming our deficiencies in foreseeing emerging technologies. *Computerworld*. Retrieved June 22, 2006, from http://www.computerworld.com/managementtopics/management/story/0,10801,108005,00.html?SKC=management-108005

Kochikar, V. P., Mahesh, K., & Mahind, C. S. (2002). Knowledge management in action: The experience of Infosys Technologies. In V. Hlupic (Ed.), *Knowledge and business process management* (pp. 83-98). Hershey, PA: Idea Group Publishing.

Kochikar, V. P., & Suresh J. K. (2005). Experiential perspective on knowledge management. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and information technology* (pp. 1162-1168). Hershey, PA: Idea Group Reference.

McAfee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration. *MIT Sloan Management Review, 47*(3), 21-28.

Metcalfe, R. (1996). *The Internet after the fad.* Retrieved June 21, 2006, from http://www.americanhistory.si.edu/csr/comphist/montic/metcalfe.htm

MIT. (2006). *The open mind project.* Retrieved July 29, 2006, from www.openmind.org

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital and the organizational advantage. *Academy of Management Review, 23*(2), 243.

Nonaka, I., & Ichijo, K. (2006). *Knowledge creation and management: New challenges for managers*. Oxford, UK: Oxford University Press.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford, UK: Oxford University Press.

Skyrme, D. J. (2003). *Measuring knowledge and intellectual capital: Models and methods to maximize the value of knowledge, intangibles and intellectual assets.* Retrieved from http://www.skyrme.com/pubs/measures2.htm

Storey, J., & Barnett, E. (2000). Knowledge management initiatives: Learning from failure. *Journal of Knowledge Management, 4*(2), 145-156.

Sveiby, K-E. (1997). *The new organizational wealth: Managing and measuring knowledge-based assets.* San Francisco: Berret-Koehler.

Wikipedia. (2006). *Prediction markets.* Retrieved July 21, 2006from http://en.wikipedia.org/wiki/Prediction_market

Wolpers, J. & Zitzevitz, E. (in press). Prediction markets in theory and practice. In L. E. Blume & S. N. Durlauf (Eds.), *The new Palgrave dictionary of economics* (2nd ed.). London: Palgrave Macmillan. Retrieved from http://bpp.wharton.upenn.edu/jwolfers/Papers/PredictionMarkets(Palgrave).pdf

KEY TERMS

Collaborative Filtering: A technique for producing recommendations that are likely to meet an individual's taste, by looking at the preferences of "like-minded" other people.

Intangible Assets: Organizational assets that do not have any physical manifestation, or whose physical measures have no bearing on their value. The following is a typical list of intangible assets:

- fragmented knowledge residing with individuals, or encapsulated in artifacts such as documentation and software code;
- codified and classified knowledge residing in repositories;
- unique systems, processes, methodologies, and frameworks that the organization follows;
- "formalized" intellectual property such as patents, trademarks, and brands; and
- relationships and alliances that the organization may have shaped (Kochikar, 2002).

Intellectual Capital (IC): The "stock" of knowledge that exists in an organization, that can be used for generating value for stakeholders (Bontis, Crossan, & Hulland 2002).

Knowledge Currency Units (KCU): A mechanism defined at Infosys Technologies to convert all knowledge-sharing activities to a common denominator, in order to enable their measurement in quantitative terms.

Knowledge Management (KM): The gamut of organizational processes, responsibilities, and systems directed toward the assimilation, dissemination, harvest, and reuse of knowledge (Kochikar, 2000).

Metcalfe's Law: The utility of a network rises in proportion to the square of the number of its users. This means that as more users get connected into a network, the marginal utility perceived by new users increases dramatically (Metcalfe, 1996).

Prediction Markets: Speculative markets that can aggregate the opinions of a large number of users regarding the outcome of a particular event (Wikipedia, 2006).

Social Capital: The resources available through and derived from the network of relationships possessed by an individual or social unit within an organization (Nahapiet & Ghoshal, 1998).

Managing Relationships in Virtual Team Socialization

Shawn D. Long
University of North Carolina at Charlotte, USA

Gaelle Picherit-Duthler
Zayed University, UAE

Kirk W. Duthler
Petroleum Institute, UAE

INTRODUCTION

The traditional organizational workplace is dramatically changing. An increasing number of organizations are employing workers who are physically and geographically dispersed and electronically dependent on each other to accomplish work (Gibson & Cohen, 2003; Griffith, Sawyer, & Neale, 2003). Recent technological advances, combined with more flexible job design, have helped increase the number of people working in distributed environments. Hence, more employees are working individually and on teams that seldom, if ever, meet face to face. These virtual employees have the same work responsibilities as traditional employees in addition to the challenge of operating within the dynamics of these newly designed mediated workplaces.

Rapid developments in communication technology and the increasing influence of globalization and efficiency on organizations have significantly accelerated the growth and importance of virtual teams in contemporary workplaces. Virtual teams are becoming more commonplace because of the possibilities of a more efficient, less expensive, and more productive workplace. Additionally, distributed teams are less difficult to organize temporal organizational members than traditional co-located teams (Larsen & McInerney, 2002; Lurey & Raisinghani, 2001; Piccoli & Ives, 2003).

Although there are apparent advantages of organizing work virtually, the challenge for new member integration lies in the fact that team members must communicate primarily through communication technology such as electronic mail, telephone, and videoconferencing or computer conferencing. This increased dependence on technology as a medium of communication significantly alters the way new members are socialized to work teams. Additionally, team members' ability to use complex communication technologies varies across individuals. This variation potentially may lead to inter- and intra-group conflict, as well as creating organizational work ambiguity, which refers to the existence of conflicting and multiple interpretations of a work issue (Miller, 2006). This article addresses the challenges of virtual team socialization with regard to newcomer assimilation and how newcomer encounter is an embedded process of virtual team assimilation.

BACKGROUND

Effective communication is central to organizational and team socialization. The way individuals are socialized in a team may determine his or her success within the team and the successful achievement of organizational and team goals. Team socialization and the communication practices associated with newcomer integration have been researched extensively (e.g., Brockmann, & Anthony, 2002; Lagerstrom & Anderson, 2003) since Jablin (1982) first explored this multilayered process. Socialization occurs when a newcomer of a team acquires the knowledge, behavior, and attitudes needed to participate fully as a member of that team. Jablin (1987) framed the stages of socialization as anticipatory socialization, organizational assimilation (encounter and metamorphosis), and organizational exit. Although there is an abundance of literature on traditional organizational socialization, research on virtual team socialization is beginning to emerge (Ahuja & Galvin, 2003; Picherit-Duthler, Long, & Kohut, 2004; Long, Kohut, & Picherit-Duthler, 2004).

NEWCOMER ASSIMILATION IN VIRTUAL TEAMS

Organizational assimilation is perhaps the most important, yet complicated, stage of virtual team socialization. Assimilation concerns the ongoing behavioral and cognitive processes of integrating individuals into the culture of an organization (Jablin, 1982). Assimilation is a dual-action process that consists of planned and unintentional efforts by the organization to "socialize" employees, while at the same time the organizational members attempt to modify their work roles and environment to coincide with their own

individual values, attitudes, and needs. Jablin (1987) suggests that organizational roles are negotiated and socially constructed by actively and reactively communicating role expectations by both the organization and its members. Newcomers typically enact this negotiation through information-seeking tactics.

Organizational culture also informs how newcomers are assimilated in virtual teams. Socialization is one of the most important processes by which organizations communicate their culture (Cheney, Christensen, Zorn, & Ganesh, 2004). While each member entering the organization learns the values, beliefs, and practices of the organization, they simultaneously shape the organization through their "reading" of those values. Because the spirit of virtual teams focuses on innovation, change, dynamic structure, and participant diversity, we should expect newcomers to be able to do more to shape the culture of their virtual team with their own values, beliefs, and practices than in the traditional team structure.

Organizational encounter as a phase of socialization is a time for newcomers to learn behaviors, values, and beliefs associated with their jobs and organizations (Schein, 1988). By entering a new situation, newcomers want to clarify their situational identity through their work roles (Berlew & Hall, 1966; Feldman, 1976), or through securing approval of others (Graen & Ginsburgh, 1977; Katz, 1978; Wanous, 1980). To reduce uncertainty, newcomers often search for information that allows them to adjust by defining the expectations of others and orienting their behavior to the behavior of others.

The speed that virtual teams form demands that workers deal with change rapidly. Although research on teamwork suggests that teams function optimally after they have worked together for a period of time, virtual teams may not have the luxury of establishing working relationships over an extended period of time (e.g., Furst, Blackburn, & Rosen, 1999; Mark, 2001). Hence, it is vital for newcomers to quickly establish and develop relationships with others in the work setting, especially with peers and supervisors (Jablin, 2001).

Among other things, organizational relationships provide newcomers with support that facilitates the learning process and reduces stress and uncertainty associated with adjusting to a new work environment (Jablin, 2001). Much of the research on relationship development in the organizational encounter stage focuses on information seeking and information giving (e.g., Boyd & Taylor, 1998), learning behaviors and attitudes through exchange activities (e.g., Comer, 1991), technical or social information (Comer, 1991; Morrison, 1995), and regulative and normative information (e.g., Galvin & Ahuja, 2001). Evidence suggests that formal and informal socialization practices may affect the level of organizational commitment (Berlew & Hall, 1966; Buchanan, 1974), longevity in the organization (Katz, 1978; Wanous, 1980), and satisfaction and feelings of personal worth (Feldman, 1976). In fact, Gibson and Gibbs (2005) propose that a supportive communication climate, defined as an atmosphere that encourages open, constructive, and honest and effective interaction (p .4), often enables innovation.

The next section examines the three central areas of relationship building in virtual teams: peer relationships, supervisory relationships, and mentoring relationships.

Peer Relationships

Working with others on a team may be problematic. Several questions arise when working with others in this context. Do individuals meet the expectations the team has of them? Are they easy to get along with? Are they competent? Peers help newcomers integrate disjointed pieces of information (Van Maanen, 1984) and communicate subtle values and norms that may not be explicitly expressed by their supervisors. Newcomers have more contact with coworkers, and as a consequence, more opportunities to share information with them and develop relationships (Jablin, 2001; Comer, 1991; Teboul, 1994). Sias and Cahill (1998) proposed a variety of contextual factors, including shared tasks and group cohesion (e.g., Fine, 1986), physical proximity (e.g., Griffin & Sparks, 1990), lack of supervisor consideration (Odden & Sias, 1997), and life events outside the workplace, as well as individual factors, such as perceived similarity in attitudes and beliefs as well as demographic similarity (Adkins, Ravlin, & Meglino, 1996; Duck, 1994; Glaman, Jones, & Rozelle, 1996; Kirchmeyer, 1995), that may affect the development of relationships with peers.

Trust is a key factor in developing close relationships. However, due to the lack of physical proximity and the reliance on communication technologies, our understanding of trust in virtual teams is different from the trust in traditional teams. Piccoli and Ives (2003) define team trust as the belief that an individual or group makes good-faith efforts to behave in accordance with any commitments both explicit and implicit. Cummings and Bromley (1996) further define trust as honesty in whatever negotiations preceded the commitment as well as not taking excessive advantage of another even when the opportunity is available (Cummings and Bromley, 1996). Meyerson, Weick, and Kramer (1996) coined the term "swift trust" to describe how virtual teams develop a different type of trust than in traditional teams. Due to the highly interdependent nature of task orientation of the team, newcomers develop trust more quickly. Team members are able to develop trust in the relationships on the basis of shared tasks rather than on the basis of similar demographics and/or physical proximity found in traditional teams (Jarvenpaa & Leidner, 1999).

However, swift trust is not enough to develop close peer relationships. Team members face a number of challenges including: technological mistrust by both newcomers and established members, intuitive fear of the misuse of archived

communication (e.g., e-mail trails), and the difficulty of sharing personal or non-work-related issues. Thus, virtual newcomers may be unable or unwilling to take advantage of the informal organizational development that appears central to organizational socialization in traditional teams. This clearly inhibits the development of close peer relationships in virtual teams, which in turn may inhibit constructive team cohesion. Similarly, opportunities to understand organizational politics are greatly reduced by the inherent dispersed nature of virtual teams. Unless the communication among team members is open, power alliances may form that foster certain behaviors such as social loafing, domination, and the formation of cliques to occur. Groups or individuals are alienated by these behaviors and may differ in their responses based on location or functional role. The outcome is the same — limited effectiveness of the team, low commitment, low loyalty, and mistrust. Other sources of information such as supervisors and mentors may prove more helpful in recognizing and adapting to political nuances.

Supervisor Relationships

Supervisors are important for assimilating newcomers to organizations by helping build a shared interpretive system that is reflective of assimilation (Berlew & Hall, 1966; Feldman, 1976; Graen, 1976; Kozlowski & Doherty, 1989; Ostroff & Kozlowski, 1992; Schein, 1988). Supervisors who frequently communicate with newcomers serve as a role model. These supervisors filter and interpret formal downward-directed management messages, have positional power to administer rewards and punishments, are a central source of information related to job and organizational expectations as well as feedback on task performance, and are pivotal in the newcomer's ability to negotiate his or her role (Ben-Yoav & Hartman, 1988; Jablin, 2001). According to Staples, Hulland, and Higgins (1998), workers who learn their communication practices by modeling their managers' behaviors have greater self-efficacy, better performance, and more positive job attitudes.

The supervisor-subordinate relationship is more important in virtual teams than in traditional teams due to the dislocated nature of the virtual structure (Long et al., 2005). The supervisor-subordinate relationship is complicated by the absence of a physical communication context that characterizes most traditional teams. The supervisor's coordination of virtual team activities is more difficult because of the distinct nature of technological feedback (synchronous vs. asynchronous), the lack of robust spontaneous information exchange between supervisor-subordinate, and the obvious reduction of face-to-face verbal and nonverbal communication cues. On the other hand, some findings suggest that assessment of team member contributions may be more accurate in virtual rather than face-to-face environments. For example, Weisband and Atwater (1999) found that ratings of liking contributed less bias to evaluations of contribution for virtual groups than face-to-face groups. Similarly, Hedlund, Ilgen, and Hollenbeck (1998) found that leaders of computer-mediated teams were better able to differentiate quality of decisions compared to leaders in face-to-face teams.

Regardless of whether the supervisor is part of the team or not, the effective supervisor-subordinate relationship depends in large part on whether the organization uses a traditional approach to managing the virtual team. In traditional teams, often supervisor-subordinate relationships are characterized by hierarchical embedded roles in responsibilities, more formalized rules, procedures, and structures (McPhee & Poole, 2001). However, in virtual teams there is a loosening of the rules and responsibilities in the supervisor-subordinate relationship. The virtual setting reduces tangible cues that distinguish the status and/or hierarchy of the team members. Thus, the supervisor-subordinate relationships in a virtual team rely more on co-orientation, which facilitates the socialization process more effectively. Mentoring relationships is also important to newcomers' adjustment to socialization efforts.

Mentoring Relationships

When discussing relationship building as part of the assimilation process, mentoring relationships is an important aspect. Mentors facilitate newcomer organizational adjustment by offering advice, support, and if appropriate, coaching behaviors to accomplish goals. Wigand and Boster (1991) suggest that "mentoring speeds up socialization into the work role, encourages social interaction, provides an opportunity for high-quality interpersonal interactions, and enhances identification with and commitment to the organization" (p. 16).

Mentoring relationships are formal and informal. Formal mentoring is a "deliberative pairing of a more skilled or experienced person with a lesser skilled or experienced one, with the agreed upon goal of having the lesser skilled or experienced person grow and develop specific competencies" (Murray & Owen, 1991, p. xiv). Several scholars (e.g., Allen, McManus, & Russell, 1999; Heimann & Pittenger, 1996; Seibert, 1999) acknowledge that newcomers who participate in formal mentoring relationships in traditional organizations realize greater benefits than those who do not have formal mentoring. Specifically, participation in formal mentoring increases the newcomers' understanding of various organizational issues and increases their level of organizational and job satisfaction.

Informal mentoring relationships develop naturally at the discretion of the mentor and protégé, and exist as long as the parties involved experience sufficient positive outcomes (Jablin, 2001). Newcomers who are informally mentored are privileged to information not directly associated with the job role or organizational tasks. This indirect communication includes organizational power and politics, involved

career-related support, "inside" information about various organizational issues, and increased social interaction outside of the workplace. As trust, commitment, and identification in the virtual team develops for both the newcomer and more experienced workers, informal mentoring will naturally occur. Virtual teams are more effective when communication barriers such as role uncertainty, task ambiguity, and tacit norms of the team are dismantled.

Organizations benefit when they recognize the value of both formal and informal mentoring relationships. Acknowledging the positive impact mentoring has on newcomer assimilation in a traditional team arrangement leads us to assume that mentoring will have similar impact on virtual team assimilation. However, due to the structural challenges of virtual teams, organizations should consider both formal and informal mentoring programs as tools to socialize newcomers to virtual teams.

In summary, virtual teams face an uncertain, but promising future. The socialization process of team members can become an enigma when building virtual teams. The next section outlines future trends in newcomer assimilation in virtual team socialization.

FUTURE TRENDS

Three interrelated relational aspects of virtual team assimilation are important to note when attempting to predict future trends in this emerging field of study. First, growing interest in virtual team socialization will lead to an accelerated interest in how trust is developed and maintained in virtual team relationships. Trust is a major factor influencing the cohesiveness among virtual team members (Sarker, Valacich, & Sarker, 2003). Instrument development and validation is of central concern as virtual teams become a ubiquitous aspect of organizational life. Scholars and practitioners both have a vested interest in fully developing this aspect of virtual team scholarship, as more individuals will work in this new organizational configuration and become more dependent upon the work of others in their virtual team. Working remotely clearly has its benefits, but opportunities will only be maximized when trust is fully realized among all organizational citizens, especially virtual team members.

Second, focused attention should be given to the amount of social contact individuals experience via mediated communication with their peers and supervisors. As workers become increasingly more isolated because of the flexibility technology affords them (e.g., telecommuting, flex hours), managers should be proactive in ensuring that many of the social components characteristic of working in traditional "brick-and-mortar" workplaces—the characteristics that keep individuals socially satisfied and committed to the organization—are transferred to the new dislocated work environment. Several scholars have suggested that the informal organization is equally, if not more powerful than the formal organization (e.g., Rogers & Kincaid, 1981; Monge & Contractor, 2001; Monge & Eisenberg, 1987). In order to maintain a consistently committed and talented workforce, deliberate attention should be earmarked to foster the intangible social relational aspect of virtual team functioning.

Finally, managers and scholars should work in tandem to fully realize the opportunities and potential for virtual mentoring. Networking and building informal coalitions and communities with others within and outside of their employing organization is a key strategy for upward mobility for individuals, especially minority organizational members (Bell & Nkomo, 2001; Parker, 2003). As more individuals are hired and socialized to work in more virtual team-based structures, it is critical that organizational leaders leverage the power of mentoring as a means to create a more committed workforce and reduce job transfers and turnover. Establishing a protégé and mentor relationship is critical, as technological uncertainty and task ambiguity increase due to the erosion of traditionally rich media forms such as face-to-face communication. Implementing formal and informal mentoring programs is certainly a future trend in virtual team socialization.

CONCLUSION

Organizations are turning to virtual teams as a way to remain competitive in an environment characterized by globalization, mergers, acquisitions, and dependence on information technologies. A great deal of attention is paid to how to provide adequate virtual team infrastructure such as hardware and software components. However, little attention has been devoted to the "human-structure" of virtual team organizing. Future research and organizational attention should focus on the methods of assimilating newcomers in virtual teams. This communication process is as important as the technology selected to accomplish work.

Some aspects of virtual team assimilation are similar to traditional team assimilation, but many are not congruent. Staples and Webster (2007) posit that the distinction between traditional and virtual teams is no longer needed, as all types of teams are characterized by varying degrees of virtuality. The primary difference is that virtual team members are typically more reliant on information technology as the medium for communicating, and virtual team members are more likely to be isolated from the rest of their team — hence, the importance of examining how virtual team members build relationships in this new organizational structure. Communication in peer, supervisor, and mentoring relationships is vital in optimizing organizational functioning, yet little attention is focused on these important relationships in the virtual team environment. In addition to the traditionally studied outcome variables such as costs, productivity,

and effectiveness, organizations should also be mindful of the importance of issues related to the internal team and interpersonal communication processes that embody and constitute organizations.

Effective virtual team assimilation, just like traditional team assimilation, fosters loyalty, commitment, trust, and potentially greater cohesiveness with the team. Virtual team socialization is a shared relational responsibility for the newcomer, the supervisor, and the organization. If little concern is given to building the relationships, then the long-term stability of the virtual team may be threatened.

REFERENCES

Adkins, C.L., Ravlin, E.C., & Meglinao, B.M. (1996). Value congruence between co-workers and its relationship to work outcomes. *Group & Organization Management, 21,* 439-460.

Ahuja, M.K., & Galvin, J.E. (2003). Socialization in virtual groups. *Journal of Management, 29,* 161-185.

Allen, T.D., McManus, S.E., & Russell, J.E.A. (1999). Newcomer socialization and stress: Formal peer relationships as a source of support. *Journal of Vocational Behavior, 54,* 453-470.

Bell, E., & Nkomo, S. (2001). *Our separate ways: Black and white women and the struggle for professional identity.* Boston: Harvard Business School Press.

Ben-Yoav, O., & Hartman, K. (1988). Supervisors' competence and learning of work values and behaviors during organizational entry. *Journal of Social Behavior and Personality, 13,* 23-36.

Berlew, D.E., & Hall, D.T. (1966). The socialization of managers: Effects of expectations on performance. *Administrative Science Quarterly, 11,* 207-223.

Boyd, N.G., & Taylor, R.R. (1998). A developmental approach to the examination of friendship in leader-follower relationships. *Leadership Quarterly, 9,* 1-25.

Brockmann, E.N., & Anthony, W.P. (2002). Tacit knowledge and strategic decision making. *Group and Organization Management, 27,* 436-455.

Buchanan, B. (1974). Building organizational commitment: The socialization of managers in work organizations. *Administrative Science Quarterly, 19,* 533-546.

Comer, D.R. (1991). Organizational newcomers' acquisition of information from peers. *Management Communication Quarterly, 5,* 64-89.

Cummings, L.L., & Bromley, P. (1996). The organizational trust inventory (OTI): Development and validation. In T.R. Tyler & R.M. Kramer (Eds.), *Trust in organizations: Frontiers of theory and research.* Thousand Oaks CA: Sage.

Duck, S. (1994). *Meaningful relationships: Talking, sense, and relations.* Thousand Oaks, CA: Sage.

Feldman, D.C. (1976). Contingency theory of socialization. *Administrative Science Quarterly, 21,* 433-452.

Fine, G.A. (1986). Friendships in the workplace. In V.J. Derlega & B.A. Winstead (Eds.), *Friendship and social interaction* (pp. 185-206). New York: St. Martin's.

Furst, S., Blackburn, R., & Rosen, B. (1999). Virtual team effectiveness: A proposed research agenda. *Information Systems Journal, 9,* 249-269.

Galvin, J.E., & Ahuja, M.K. (2001). Am I doing what's expected? New member socialization in virtual groups. In L. Chidambaram & I. Zigurs (Eds.), *Our virtual world: The transformation of work, play and life via technology* (pp. 40-55). Hershey, PA: Idea Group.

Gibson, C.B., & Cohen, S.G. (2003). *Virtual teams that work: Creating conditions for virtual collaboration effectiveness.* San Francisco: Jossey-Bass.

Gibson, C.B., & Gibbs, J.L. (2005, May). Unpacking the concept of virtuality: The role of supportive communication climate in facilitating team innovation. *Proceedings of the Meeting of the International Communication Association,* New York.

Glaman, J.M., Jones, A.P., & Rozelle, R.M. (1996). The effects of co-worker similarity on the emergence of affect in work teams. *Group & Organization Management, 21,* 192-215.

Graen, G. (1976). Role-making processes within complex organization. In M.D. Dunnette (Ed.), *Handbook of industrial/organizational psychology* (pp.1201-1245). Chicago: Rand McNally.

Graen, G., & Ginsburgh, S. (1977). Job resignation as a function of role orientation and leader acceptance: A longitudinal investigation of organizational assimilation. *Organizational Behavior and Human Performance, 19,* 1-17.

Griffin, E., & Sparks, G.G. (1990). Friends forever: A longitudinal exploration of intimacy in same-sex pairs and platonic pairs. *Journal of Social and Personal Relationships, 7,* 29-46.

Griffith, T.L., Sawyer, J.E., & Neale, M.A. (2003). Virtualness and knowledge in teams: Managing the love triangle of organizations, individuals, and information technology. *MIS Quarterly, 27*(2), 265-287.

Hedlund, J., Ilgen, D.R., & Hollenbeck, J.R. (1998). Decision accuracy in computer- mediated versus face-to-face decision-making teams. *Organizational Behavior & Human Decision Performance, 76,* 30-47.

Heimann, B., & Pittenger, K.K.S. (1996). The impact of formal mentorship on socialization and commitment of newcomers. *Journal of Managerial Issues, 8,* 108-117.

Jablin, F.M. (1982). Organizational communication: An assimilation approach. In M.E. Roloff & C.R. Berger (Eds.), *Social cognition and communication* (pp. 255-286). Beverly Hills, CA: Sage.

Jablin, F.M. (1987). Organizational entry, assimilation and exit. In F.M. Jablin, L.L. Putnam, K.H. Roberts, & L.W. Porter. (Eds.), *Handbook of organizational communication: An interdisciplinary perspective* (pp. 679-740). Newbury Park, CA: Sage.

Jablin, F.M. (2001). Organizational entry, assimilation, and disengagement/exit. In F.M. Jablin & L.L. Putnam (Eds.), *The new handbook of organizational communication: Advances in theory, research, and methods* (pp. 732-818). Thousand Oaks, CA: Sage.

Jarvenpaa, S.L., & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organization Science, 10,* 791-815.

Katz, R. (1978). Job longevity as a situational factor in job satisfaction. *Administrative Science Quarterly, 23,* 204-223.

Kirchmeyer, C. (1995). Demographic similarity to the work group: A longitudinal study of managers at the early career stage. *Journal of Organizational Behavior, 16,* 67-83.

Kozlowski, S.W.J., & Doherty, M.L. (1989). Integration of climate and leadership: Examination of a neglected issue. *Journal of Applied Psychology, 74,* 546-553.

Lagerstrom, K., & Anderson, M. (2003). Creating and sharing knowledge within a transnational team—the development of a global business system. *Journal of World Business, 38,* 84-95.

Larsen, K.R.T., & McInerney, C.R. (2002). Preparing to work in the virtual organization. *Information and Management, 39,* 445-456.

Long, S.D., Kohut, G.F., & Picherit-Duthler, G. (2005). Newcomer assimilation in virtual team socialization. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (vol. 1). Hershey, PA: Idea Group.

Lurey, J.S., & Raisinghani, M.S. (2001). An empirical study of best practices in virtual teams. *Information and Management, 38,* 523-544.

Mark, G. (2001). Meeting current challenges for virtually collated teams: Participation, culture, integration. In L. Chidambaram & I. Zigurs (Eds.), *Our virtual world: The transformation of work, play and life via technology* (pp. 74-93). Hershey, PA: Idea Group.

McPhee, R.D., & Poole, M.S. (2001). Organizational structures and configurations. In F.M.. Jablin & L.L. Putnam (Eds.), *The new handbook of organizational communication: Advances in theory, research, and methods* (pp. 503-542). Thousand Oaks, CA: Sage.

Meyerson, D., Weick, K.E., & Kramer, R.M. (1996). Swift trust and temporary groups. In R.M. Kramer & T.R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 166-195). Thousand Oaks, CA: Sage.

Monge, P., & Contractor, N. (2001). Emergence of communication networks. In F. Jablin & L. Putnam (Eds.), *The new handbook of organizational communication* (pp. 440-502). Thousand Oaks, CA: Sage.

Monge, P., & Eisenberg, E. (1987). Emergent communication networks. In F. Jablin, L. Putnam, K. Roberts, & L. Porter (Eds.), *Handbook of organizational communication* (pp. 204-342). Beverly Hills, CA: Sage.

Morrison, E.W. (1995). Information usefulness and acquisition during organizational encounter. *Management Communication Quarterly, 9,* 131-155.

Murray, M., & Owen, M. (1991). *Beyond the myths and magic of mentoring: How to facilitate an effective mentoring program.* San Francisco: Jossey-Bass.

Odden, C.M., & Sias, P.M. (1997). Peer communication relationships and psychological climate. *Communication Quarterly, 45,* 153-166.

Ostroff, C., & Kozlowski, S.W.J. (1992). Organizational socialization as a learning process: The role of information acquisition. *Personnel Psychology, 45,* 849-87.

Parker, P. (2003). Control, resistance, and empowerment in raced, gendered, and classed contexts: The case of the African American woman. In P. Kalbfleisch, (Ed.), *Communication yearbook* (vol. 27, pp. 257-291). London: Lawrence Erlbaum.

Piccoli, G., & Ives, B. (2003). Trust and the unintended effects of behavior control in virtual teams. *MIS Quarterly, 27,* 365-395.

Picherit-Duthler, G., Long, S.D., & Kohut, G. (2004). Newcomer assimilation in virtual team socialization. In S. Godar & S.P. Ferris (Eds.), *Virtual and collaborative teams: Process, technologies, & practice.* Hershey, PA: Idea Group.

Rogers, E., & Kincaid, D. (1981). *Communication networks: Toward a new paradigm for research.* New York: The Free Press.

Sarker, S., Valacich, J.S., & Sarker, S. (2003). Virtual team trust: Instrument development and validation in an IS educational environment. *Information Resources Management Journal, 16*(2), 35-55.

Schein, E.H. (1988). Organizational socialization and the profession of management. *Sloan Management Review, 30,* 53-65.

Seibert, S. (1999). The effectiveness of facilitated mentoring: A longitudinal quasi-experiment. *Journal of Vocational Behavior, 54,* 483-502.

Sias, P.M., & Cahill, D.J. (1998). From coworkers to friends: The development of peer friendships in the workplace. *Western Journal of Communication, 62,* 273-299.

Staples, D.S., Hulland, J.S., & Higgins, C.A. (1998). A self-efficacy theory explanation for the management of remote workers in virtual organization. *Journal of Computer-Mediated Communication, 3,* 4.

Staples, D.S., & Webster, J. (2007). Exploring traditional and virtual team members' "best practices": A social cognitive theory perspective. *Small Group Research, 38*(1), 60-97.

Teboul, J.C.B. (1994). Facing and coping with uncertainty during organizational encounter. *Management Communication Quarterly, 8,* 190-224.

Van Maanen, J. (1984). Doing new things in old ways: The chains of socialization. In J.L. Bess (Ed.), *College and university organizations* (pp. 211-247). New York: New York University Press.

Wanous, J.P. (1980). *Organization entry: Recruitment, selection, and socialization of newcomers.* Reading, MA: Addison-Wesley.

Weisband, S., & Atwater, L. (1999). Evaluating self and others in electronic and face-to-face groups. *Journal of Applied Psychology, 4,* 632-639.

Wigand, R.T., & Boster, F.S. (1991). Mentoring, social interaction, and commitment: An empirical analysis of a mentoring program. *Communications, 16,* 15-31.

KEY TERMS

Co-Located Team: A traditional team that shares a common goal and works toward that goal in a face-to-face, same-office environment.

Formal Mentoring: A deliberate pairing of a more skilled or experienced person with a lesser skilled or experienced one, with the agreed-upon goal of having the lesser skilled or experienced person grow and develop specific competencies.

Informal Mentoring: The non-assigned pairing of an experienced person who respects, guides, protects, sponsors, promotes, and teaches a younger, less experienced personnel member who develops naturally at the discretion of the mentor and protégé, and persists as long as the parties involved experience sufficient positive outcomes.

Organizational Assimilation: The processes by with individuals become integrated into the culture of an organization.

Newcomer Encounter: A time for newcomers to learn behaviors, values, and beliefs associated with their jobs and organizations.

Socialization: The process in which that member of a team acquires the knowledge, behavior, and attitude needed to participate fully as a member of the team.

Swift Trust: A type of trust that develops quickly on the basis of shared tasks rather than on the basis of similar demographics or physical proximity.

Virtual Team: A group of geographically and organizationally dispersed workers brought together across time and space through information and communication technologies.

Managing the Integrated Online Marketing Communication

Călin Gurău
GSCM – Montpellier Business School, France

INTRODUCTION

This chapter investigates the particularities of integrated marketing communication in the online environment. The study starts from the premise that the specific characteristics of the Internet transform the application of IMC principles from an alternative option to an absolute requirement for online organizations. Based on the analysis of the specific characteristics of the online environment and audiences, and on the primary data collected through face-to-face interviews with 19 marketing or communication managers of UK consumer retail firms, this article explores the opportunities and requirements for implementing integrated online marketing communication, proposing a theoretical model of that can be adopted by *Internet-active organizations.*

BACKGROUND

In the last 10 years, the concept of *integrated marketing communication* (IMC) has achieved notoriety and legitimacy both in the academic and in the professional environment (Schultz & Kitchen, 2000). This situation is proved by the numerous studies and debates centered on the subject of IMC (Cornelissen & Lock, 2001; Gould, 2004; Percy, Rossiter, & Elliott, 2001; Schultz & Kitchen, 2000).

The emergence and the development of IMC has been determined by a number of evolutionary trends in the areas of:

- *Marketing*—The increased fragmentation and segmentation of markets, relationship marketing, and direct marketing (Durkin & Lawlor, 2001; Eagle & Kitchen, 2000).
- *Information Technology*—The development of new communication technologies and database applications (McKim, 2002).
- *Communication*—Increased fragmentation of media audiences, multiplicity, and saturation of media channels (Smith, 2002).

From this perspective, the new paradigm of IMC can be represented as a strategic answer to the social and business conditions of the postmodern society (Proctor & Kitchen, 2002), which forced marketing organizations to move beyond functionally driven, internally focused approaches to marketing and communication (Cornelissen, 2003).

The concept of IMC was defined in many different, often contradictorial ways (Duncan, 2002; Shimp, 2000). The integration of marketing communication procedures was considered a result of centralized management, centralized budgeting, or message similarity across all communication channels, while other authors emphasized the integration of all the elements of the promotional mix in a coherent strategy (Pickton & Broderick, 2001). Many definitions emphasize that the integration of marketing communication should not be understood as a simple uniformity of the message transmitted across different channels (Kitchen, Brignell, Li, & Jones, 2004), but rather as the complex coordination and management of the information transmitted though complementary channels in order to effectively present a coherent image of the organization to the targeted audiences.

Beverland and Luxton (2005) argue that one of the main effects of IMC is the development of brand trust and credibility. Proctor and Kitchen (2002) emphasize that in the last 10 years there has been a significant move away from line branding towards corporate brand. The main reason is the desire to amortize communication across the entire portfolio as the cost of designing and supporting individual brands continues its upward curve. Board members and executives have come to realize that a major portion of shareholder value is brand equity, which requires careful development and management (Laczniak, 2005). In a fragmented and highly competitive marketplace, coherent images and messages lead to a greater impact on the perception and attitudes of targeted audiences (Moriarty, 1997). The integrated marketing communication effort can ensure that brand messages are strategically consistent and new communication technologies are effectively used to facilitate profitable interactions with customers and other stakeholders.

Despite the recognized impact of the Internet on integrated marketing communication, very few studies have investigated the specific requirements and opportunities for IMC in the online environment (Durkin & Lawlor, 2001), and the relation between IMC and customer relationship management (Grőnroos, 2004; Johnson & Schultz, 2004; Schultz, 2003).

This article attempts to investigate the particularities of implementing integrated marketing communication in an

online environment. The study considers that the specific characteristics of the Internet transform the application of IMC principles from an alternative option to an absolute requirement.

RESEARCH METHODOLOGY

In the first stage of this research project, a series of secondary sources of data have been accessed in the first instance, in order to collect general information about the evolution of the IMC concept, Internet characteristics, online communications, and online audiences.

In the second stage of data collection, a series of semi-structured interviews have been conducted with marketing or communication managers of Internet-active UK retailing firms. Using the contact information provided on their Web sites, 50 UK retailers specialized in consumer products (food, drinks, cosmetics, clothes, shoes) were contacted by phone or e-mail and invited to participate in this study. Twenty-four of these firms responded favorably, but only 19 interviews could be organized because of time restrictions in the interviewees' program. The face-to-face interviews took place from February to May 2006, and lasted between 40 and 60 minutes. The topics of discussion included the concept of integrated marketing communications, the opportunities and challenges created by the Internet concerning the corporate communication model, and the specific strategic model that can enhance the online marketing communication process. The analysis of the primary data was done manually, considering the limited number of interviews performed and the exploratory nature of the study.

THE IMPACT OF INTERNET TECHNOLOGY ON MARKETING COMMUNICATIONS

The rapid development of the Internet in the last 15 years has had a profound impact on traditional marketing paradigms and practices. But, most importantly, the Internet has changed the classical communication procedures, because of three specific and co-existent characteristics that differentiate it from any other communication channel:

- *Interactivity*—The Internet offers multiple possibilities of interactive communication, acting not only as an interface, but also as a communication agent (allowing a direct interaction between individuals and software applications).
- *Transparency*—The information published online can be accessed and viewed by any Internet user, unless this information is specifically protected.
- *Memory*—The Web is a channel not only for transmitting information, but also for storing information—in other words, the information published on the Web remains in the memory of the network until it is erased.

The networked world has increased exponentially the number of available channels of communication. We get messages from more different media: e-mail, voicemail, faxes, pages, cell phones, interoffice memos, overnight courier packages, television (with hundreds of channels), radio, Internet radio, and so forth. As a result, the media that used to provide an efficient channel of communication for practitioners have now become only noise that most of the audiences have learned to filter out. On the other hand, the *networked environment* provided the audiences with a new model, one in which they no longer accept every message a communicator wants to push to them, but they rather pull the information that suits their interests and needs. In the networked environment, information must be available where audiences can find it, and it needs to be customized or customizable (Rowley, 2001, 2004).

Therefore, in comparison with the traditional customer, the Internet user has more control over the communication process and can adopt a more proactive attitude, expressed by the capacity to:

1. easily search, select, and access information (using search and meta-search engines, intelligent agents, etc.);
2. contact online organizations or other individuals (using e-mail, chat, or discussion forums); and
3. express their opinions/views in a visible and lasting manner (creating and storing online content).

Taking advantage of the various online resources requires strategic thinking that recognizes that all these aspects of the networked world coexist. They must be coordinated to achieve specific, measurable objectives consistent with the goals of any marketing communications effort.

THE MEANING(S) OF INTEGRATED ONLINE MARKETING COMMUNICATION

The lack of a unifying definition for integrated marketing communication is one of the main barriers for the development and the practical application of this concept. One possible explanation of this theoretical crisis is the multitude of possible coexistent meanings for the IMC concept. This assumption might also be true in the case of Internet communication.

In order to identify the meaning(s) of integrated online marketing communication, the interviewees have been asked

Table 1. The practical meanings of integrated online marketing communication

Meaning	Frequency	Percentage
Combination of communication modes (one-to-one, one-to-many, many-to-many)	19	100
Integration of information types (text, sound, image)	19	100
Consistency of messages transmitted through the online communication mix (coherent meaning)	19	100
Integration of marketing and PR communication functions in the messages provided online	19	100
The coordination of the process—message conception, transmission, feedback reception, and analysis—in a closed loop	18	94.7
The direct connection of the corporate information system with the Internet	19	100
Coordination of internal, external, and internal-external flows of information	17	89.4
The integration of online marketing communication with the communication conducted through traditional channels	17	89.4
The consistency of the corporate message at global/international levels	11	57.8

to freely express the issues related with this concept in their practical activity. The responses have been semantically analyzed; the categories of meaning that have been identified are shown in Table 1.

As can be seen for all respondents, integrated online marketing communication represents a multi-faceted phenomenon, which comprises issues related to the message, the communication function, the management of information, and the specific mix of channels used for corporate communication. On the basis of these answers, the synergies and the challenges raised by the Internet are discussed in more detail in the following two sections.

INTERNET-BASED COMMUNICATION SYNERGIES

Internet technologies allow online-active organizations to implement three main communication synergies:

1. *The integration and coordination of communication modes*: The organizations can combine one-to-one (e-mail), one-to-many (list-based e-mail messages, Web pages), and many-to-many (discussion forums) communication in the online environment. This synergy increases the flexibility of the integrated communication approach, providing opportunities both for the personalization and the integration of messages (Rowley, 2001, 2004).
2. *The integration and coordination of various types of information* (Azzone, Bianchi, & Noci, 2000): The recent advances in information and communication technologies (broadband) allow organizations to transmit or receive a complex combination of information in the form of text, sound, and images (static and/or dynamic). This synergy has a direct effect on the complexity and clarity of the communication, enhancing the capacity of the organization to tailor its messages to the specific needs and requirements of various audiences.
3. *The integration and coordination of complex information flows between the organizational intranet and the Internet:* Organizations are now able to implement advanced software applications that connect their marketing and management information systems with the online environment, and to coordinate automatically the communication with various audiences (Basu, Poindexter, Drosen, & Addo, 2000). This capability has a powerful impact on multiple aspects of the communication process:
   - the capacity to capture and register automatically customer data (demographic or behavioral) and customer feedback;

- the capacity to analyze automatically the information collected about audiences, to a level of segmentation and detail that allows the implementation of one-to-one marketing communication; and
- the capacity to use the existing databases in order to automatically launch and coordinate highly targeted communication campaigns (automatic e-mail responses, automatic e-mail campaigns, personalized event marketing, promotional news, and newsletters).

INTERNET-BASED COMMUNICATION CHALLENGES

The online environment creates not only opportunities, but also challenges for the marketing communication process. The transparency of the Web makes online information available to all audiences, and reinforces the need for consistency in the planning, design, implementation, and control of online marketing communication (Hart, Neil, & Ellis-Chadwick, 2000).

On the other hand, since they are sharing the same channel and audiences, the marketing and the PR messages published on organizational Web sites are becoming more integrated (Ashcroft & Hoey, 2001). The corporate Web site is usually structured on various information categories, such as organizational profile, activity, products and services, financial reports, and other information for investors, job vacancies, contact, and related links.

The variety and multiplicity of information, sources, and interpretations available online raises an important challenge related to the management of the corporate image and identity. The voice of the corporation cannot be considered anymore as the dominant message, but only as a component in a mosaic of communication activities. The meaning is not simply transmitted, but must be negotiated separately with each online audience. The message needs to be adapted to the specific level of understanding and interpretation of each public, but on the other hand must express the same core organizational values, in order to display a coherent organizational image (Grőnroos, 2004). The various competing messages transmitted by other organizations, pressure groups, governmental agencies, or individuals must be taken into account and accommodated in such a way as the resultant effect to be favorable for the company.

The international dimension of the Internet creates another specific problem for communication practitioners. Complex choices must be made and implemented in terms of the communication strategy and tactics. If the company attempts also to reach foreign audiences, the message needs to be adapted to the cultural specificity of the overseas public. This raises important questions regarding the possibility of integrated online marketing communication in the global context.

A new strategic model must be adopted by any organization that attempts to present a coherent corporate identity in the online environment. The integrated marketing communication is the primary instrument to achieve this objective. However, the implementation of the IMC concept will have to accommodate the specific characteristics of the Internet, using the technological capabilities of the new medium to solve the specific challenges raised by the online environment and audiences.

A MODEL FOR IMPLEMENTING INTEGRATED ONLINE MARKETING COMMUNICATION

Based on the analysis of the specific characteristics of the online environment, applications, and audiences, a tentative model of integrated online marketing communications is proposed in this study. The model was designed and then refined as a result of the direct interaction and debate with 19 marketing or communication managers directly responsible for the online communication strategy of their companies.

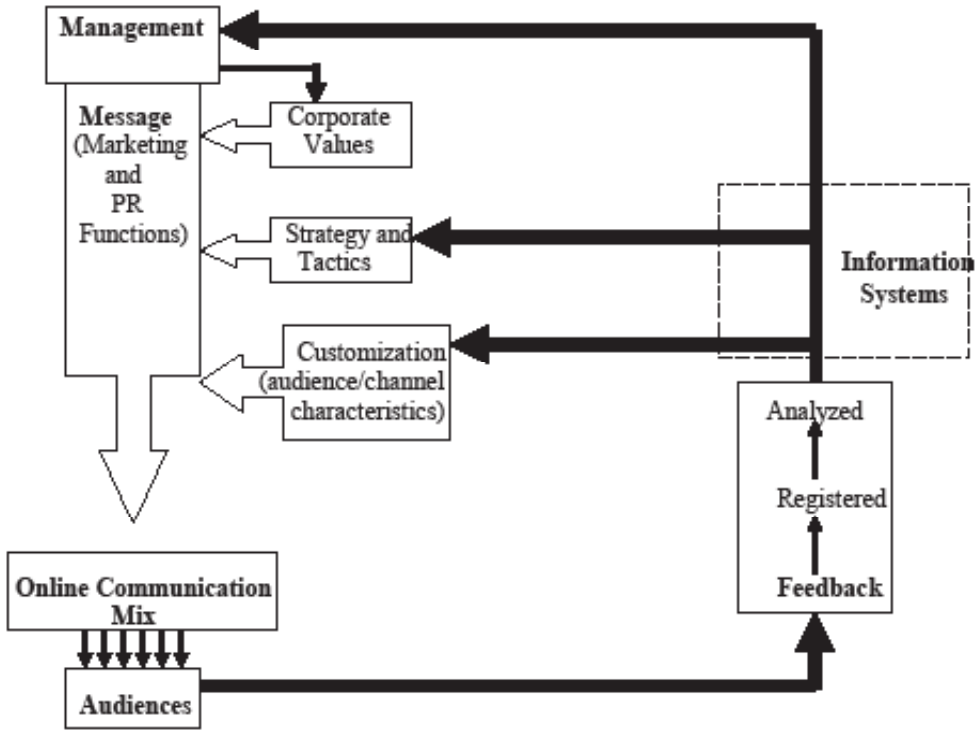
The messages sent by the company to its online audiences must be transformed/adapted in a three-stage process. First, the message should respect and integrate the core corporate values of the organization. Second, the message must be adapted in relation to the strategic and tactical objectives pursued through the online communication campaign. Third, the message should be transformed considering the specific characteristics of the targeted audience/channel. In the case of online communication, although the Internet can be considered as the main channel of communication, there are, in fact, various online applications or modalities of communication that can be combined and used as an online communication-mix (e-mail, chat, Web site, discussion forums, etc.). The online communication channels vary in terms of their degree of transparency, interactivity, memory, and selectivity, and these dimensions should be taken into account when establishing the proper *communication mix* for each targeted audience.

This process of message adaptation will preserve a flexible balance between continuity and customization, the consistency of the adapted communications being determined by the integration of the corporate core values in the structure of each message.

The respondents have also confirmed the tendency to integrate marketing and PR communication functions in the messages transmitted online, although there have been different opinions regarding the intensity of this integration.

On the other hand, the interactive dimension of the Internet forces the firm to adopt a more proactive attitude

Figure 1. A model of integrated online marketing communication



in searching, registering, and analyzing the direct and the indirect feedback transmitted by the targeted audiences—or even in some cases by all categories of relevant audiences connected to the Internet (considering the transparency and the memory of the Internet, even untargeted audiences can read and react to some of the online corporate messages).

The use of the feedback information collected and analyzed by the firm should represent a highly reactive process. The online environment is very dynamic, and any delay of an appropriate reaction to the messages sent by audiences can represent missed opportunities or aggravated situations. The company should therefore use the conclusions of the feedback analysis in order to define and better refine the strategic objectives of its communication campaigns, and the customization of the messages to audience/channel characteristics.

The feedback analysis is also transmitted to the company management, which can decide, if necessary, to modify the corporate core values in order to respond better to the market's requirements. However, this change should not be very frequent, in order to preserve the long-term coherence of corporate communications in line with the desired corporate image.

The respondents also emphasized the importance of an efficient information system that collects, selects, registers, and analyzes online input (feedback), and then acts directly on the adaptation of corporate communication strategy and tactics, as well as on the customization of online messages. In some cases, campaign management applications can use the feedback received directly and automatically for a more effective online message customization. On the other hand, the corporate information system represents the necessary basis for enhancing the customer relationship management capabilities of the firm (Grőnroos, 2004; McKim, 2002). The level of detail of customer-related data stored and analyzed by the internal information system defines the level of personalization that can be applied by the firm in its online communication and marketing campaigns. In fact, the modern database and campaign management applications permit the implementation of effective one-to-one marketing communication in the online environment.

FUTURE TRENDS

The rapid development of online communication tools requires a continuous adaptation and change of perspective from marketers. Search engines, newsgroups, blogs, discussion forums, advergames, and shared virtual reality environments create new opportunities and challenges for transmitting personalized messages to the online target audience. In these circumstances, the use of the company Web site as the main platform for integrating the online marketing communication must be complemented by the creative use of additional online communication facilities.

These new online applications permit a creative combination of above- and below-the-line communication methods, using a combination of text, images, and sound in order to transmit relevant messages to the target audience. On the other hand, in comparison with the corporate communication diffused through the company Web site, many of these online communication tools are open and interactive, outside the control of the company. In these conditions, online marketers should design and implement a formal process of technological and competitive intelligence comprising the following procedures:

- identifying the new online communication tools and evaluating their potential in relation to the targeted audiences;
- realizing a regular survey of the communication strategies applied online by competing firms;
- mapping the main online communication tools in terms of their effect on the corporate image;
- integrating the new online communication tools in the marketing communication portfolio of the firm; and
- assessing the performance of the new online communication tool and making corrections for increase efficiency and effectiveness.

CONCLUSION

This exploratory study has attempted to identify the major changes determined by the development of Internet technology in the area of marketing communications. The online environment raises a series of opportunities and challenges for communication practitioners.

The audiences become more fragmented and proactive, but, on the other hand, the company has the possibility to combine various modes and categories of information in a complex message. Online applications also permit the enterprise to collect, register, analyze, and use customer data and feedback for better targeting online audiences and customizing its messages.

In fact, the specific characteristics of the Internet are making the implementation of integrated online marketing communication both inevitable and efficient for an online organization. The transparency, interactivity, and memory of the Internet force the organization to adopt a proactive-reactive attitude in online communication, and to combine consistency and continuity with flexibility and customization.

These characteristics can be integrated by designing and implementing a specific model of integrated online marketing communication. The message communicated online should be first infused with the core corporate values, then adapted to the online strategy and tactics of the organization, and finally customized for a specific combination of target audience/online channel. The selection of the appropriate communication-mix needs to take into account the characteristics of the targeted audiences, but also the degree of transparency, interactivity, memory, and selectivity of each online channel.

The use of advanced online applications to collect customer data and feedback information is paramount for the success of the online communication campaign. Because of the high interactivity of the Internet, the communication process has become a real-time dialogue (Grőnroos 2004) that forces the company to quickly analyze the response of the audience to its messages and to restructure dynamically its message to the new circumstantial situation. Integrated online marketing communication is not an isolated and self-sufficient concept, but represents the main instrument for an effective implementation of online customer relationship management campaigns.

A number of important issues relevant for the implementation of online integrated marketing communication have not been addressed because of space and methodology limitations. These areas can represent the subject of future research projects investigating: the management process of integrated online marketing communication, the criteria used for selecting and combing various channels in the online communication mix, the relation between the organization and Web advertising agencies, or the challenges raised by the general integration and coordination of online and off-line (tradition) communication.

REFERENCES

Ashcroft, L., & Hoey, C. (2001). PR, marketing and the Internet: Implications for information professionals. *Library Management, 22*(1), 68-74.

Azzone, G., Bianchi, R., & Noci, G. (2000). The company's Web site: Different configurations, evolutionary path. *Management Decision, 38*(7), 470-479.

Basu, C., Poindexter, S., Drosen, J., & Addo, T. (2000). Diffusion of executive information systems in organizations and the shift to Web technologies. *Industrial Management and Data Systems, 100*(6), 271-276.

Beverland, M., & Luxton, S. (2005) Managing integrated marketing communication (IMC) through strategic decoupling: How luxury wine firms retain brand leadership while appearing to be wedded to the past. *Journal of Advertising, 34*(4), 1-15.

Cornelissen, J.P. (2003). Change, continuity and progress: The concept of integrated marketing communications and marketing communications practice. *Journal of Strategic Marketing, 11*(December), 217-234.

Cornelissen, J.P., & Lock, A.R. (2001). The appeal of integration: Managing communications in modern organizations. *Marketing Intelligence and Planning, 19*(6), 424-431.

Duncan, T. (2002). *IMC: Using advertising and promotion to build brands.* New York: McGraw-Hill.

Durkin, M., & Lawlor, M.-A. (2001). The implications of the Internet on the advertising agency-client relationship. *The Services Industries Journal, 21*(2), 175-190.

Eagle, L., & Kitchen, P.J. (2000). IMC, brand communications, and corporate cultures. *European Journal of Marketing, 34*(5), 667-686.

Gould, S.J. (2004). IMC as theory and as a poststructural set of practices and discourses: A continuously evolving paradigm shift. *Journal of Advertising Research, 44*(1), 66-71.

Grőnroos, C. (2004). The relationship marketing process: Communication, interaction, dialogue, value. *Journal of Business and Industrial Marketing, 19*(2), 99-113.

Hart, C., Neil, D., & Ellis-Chadwick, F. (2000). Retailer adoption of the Internet—implications for retail marketing. *European Journal of Marketing, 34*(8), 954-974.

Ihator, A.S. (2001). Communication style in the information age. *Corporate Communications: An International Journal, 6*(4), 199-204.

Johnson, C.R., & Schultz, D.E. (2004). A focus on customers. *Marketing Management, 13*(5), 20-27.

Kitchen, P.J., Brignell, J., Li, T., & Jones, G.S. (2004). The emergence of IMC: A theoretical perspective. *Journal of Advertising Research, 44*(1), 19-31.

Laczniak, R.N. (2005). From the editor. *Journal of Advertising, 34*(4), 6-1.

McKim, B. (2002). The difference between CRM and database marketing. *Journal of Database Marketing, 9*(4), 371-375.

Moriarty, S.E. (1997). The circle of synergy: Theoretical perspectives and an evolving IMC research agenda. In E. Thorson & J. Moore (Eds.), *Integrated communication: Synergy of persuasive voices.* Mahwah, NJ: Lawrence Erlbaum.

Percy, L., Rossiter, J.R., & Elliott, R. (2001). *Strategic advertising management.* New York: Oxford University Press.

Pickton, D., & Broderick, A. (2001). *Integrated marketing communications.* Essex: Pearson Education.

Proctor, T., & Kitchen, P.J. (2002). Communication in postmodern integrated marketing. *Corporate Communications: An International Journal, 7*(3), 144-154.

Rowley, J. (2001). Remodeling marketing communications in an Internet environment. *Internet Research: Electronic Networking Applications and Policy, 11*(3), 203-212.

Rowley, J. (2004). Just another channel? Marketing communications in e-business. *Marketing Intelligence and Planning, 22*(1), 24-41.

Schultz, D.E. (2003). Opinion piece: The next generation of integrated marketing communication. *Interactive Marketing, 4*(4), 318-319.

Schultz, D.E., & Kitchen, P.J. (2000). A response to 'theoretical concept or management fashion'? *Journal of Advertising Research, 40*(5), 17-21.

Shimp, T.A. (2000). *Advertising promotion: Supplemental aspects of integrated marketing communications* (5th ed.). Fort Worth, TX: Dryden Press.

Smith, P.R. (2002). *Marketing communications: An integrated approach* (3rd ed.). London: Kogan Page.

KEY TERMS

Communication Mix: The combination of various communication methods used by a company to transmit messages to its publics.

Customer Relationship Management (CRM): The procedures, methodologies, and tools that help businesses manage customer relationships in an organized and effective way.

Integrated Marketing Communication (IMC): A management orientation that integrates the use of various marketing communication methods such as advertising, sales promotion, public relations, and direct marketing in order to send a coherent and consistent message to its publics.

Marketing Communication: The messages and the related media channels used by a firm to communicate with its market.

Online Communication Channel: Online application (e-mail, chat, discussion forum, etc.) used by firms to communicate with its publics.

Public Relations Communication: The messages and the related media channels used by a firm to communicate with various publics, in order to develop a positive and coherent corporate image.

Target Audience: A clearly defined demographic group for which the company develops and transmits various communication messages.

Marketing Vulnerabilities in an Age of Online Commerce

Robert S. Owen
Texas A&M University, Texarkana, USA

INTRODUCTION

This article provides an overview of strategic and tactical threats to the marketing efforts of businesses engaged in online marketing activities. Marketing-related assets that are vulnerable to attack include networking and hardware resources, human resources, information resources, promotion resources, and brand equity and customer good will. Vulnerable areas that an organization should protect include its core network and computing infrastructure, its internal social infrastructure, domain name registrations related to its branding, and branding exploits on external social networks. Although hacks of networking and hardware resources are of concern, the focus of this article is on encouraging marketing managers and strategists to consider a wider variety of external and internal threats.

BACKGROUND

While the Internet has provided new opportunities for businesses and marketing, it has also created new vulnerabilities to attack. An organization's existing brand name can be taken hostage or destroyed via the online activities of third parties; opportunities to penetrate an online market with a new brand name can be diminished or eliminated by the actions of external third parties. Customer good will can be destroyed by the online activities of competitors or disgruntled customers. Technical, financial, and human resources can be diluted or consumed by the online activities of third parties. Confidential internal information can be compromised by employees who use e-mail and social networking Web sites for nonmission or personal uses.

This article attempts to outline such emerging strategic and tactical threats to online marketing efforts. While technical support people tend to focus on threats to hardware and networks, little guidance exists for marketing managers who should be interested in a wider variety of issues that can affect an organization's products, promotion, distribution, and costs (which affect pricing). "Scholarly" discussion on the subject is almost nonexistent, so this article attempts to compile, categorize, and discuss the sorts of issues that are starting to emerge in the popular press.

STRATEGIES AND TACTICS FOR ATTACK

The following are emerging strategies and tactics that have been enabled by online activities.

Fishing for Information

There are three basic ways to gain access to information within an organization: through exploits of the networking infrastructure, through exploits of the human social network, and through human mistakes. Networking exploits to obtain internal information could include system scans and probes, account and root compromises, packet sniffing, and malicious programming (cf. NIAC, 2004). A survey of 700 organizations by the Computer Security Institute and the U.S. Federal Bureau of Investigation found that unauthorized access amounted to $31.2 million in annual losses, and theft of proprietary information amounted to $30.9 million (Gordon, Loeg, Lucyshyn, & Richardson, 2005).

Attempts to fish for information do not have to be aimed directly at scanning and probing a networking system from the outside environment. A remote access Trojan is a hidden piece of malicious software that is attached to another seemingly innocent software application, such as a cute electronic greeting card or a more serious looking Excel spreadsheet (Vamosi, 2004). When these executable applications (greeting card, screen saver, spreadsheet, etc.) are opened, the Trojan is silently released to begin, say, covertly scanning files or logging keystrokes to be silently sent to another organization. These can be injected into an organization's computer if an employee opens an infected e-mail message or if an employee brings work that was infected on an online home computer. Once inside the organization, the Trojan can attach itself to internal applications that are exchanged, such as when employees exchange internal e-mail with executable attachments.

Microsoft employees, for example, reportedly received an infected e-mail which released a Trojan inside the Microsft organization; this in turn disguised itself as the Notepad text editor and sent information to a remote computer in Asia, with stolen passwords then used to gain access to the source code of Microsoft products (Thurrott, 2001). Attempts to

fish for information can be targeted to individual high-level executives, not just the organization as a whole. For example, an individual who opens an Excel spreadsheet attached to an e-mail message could unknowingly be installing a malicious program that now scans that person's files for information that is sent back to the criminal hacker (cf. Miller, 2003).

Information can be released through the mistakes or ignorance of employees. Members of the British Computer Society were sent a customer satisfaction survey that mistakenly contained the e-mail addresses of all recipients in the "to" field, allowing recipients to see the addresses of all other members (Oates, 2007). Information can also be obtained through simple employee ignorance. For example, employee names and e-mail addresses can be harvested through the use of chain e-mail (also known as a chain letter). Chain e-mail relies on social engineering, whereby one employee receives an e-mail message, for example, describing a cute lost puppy looking for a good home, and feels compelled to forward it to others within and outside of the organization. As each recipient forwards the seemingly harmless message to several others, a name and address list can be accumulated in each forward. This list can then be harvested when the chain letter eventually makes its way back out of the organization to the perpetrator; this allows a competitive intelligence researcher to find out who is employed by the organization, to find out who are partners or affiliates with the organization, or to find out who are the less-careful employees who are more likely to open e-mail of malicious intent. One of the more well-known incidents is the "Richard Douche Free CD" chain letter, in which the perpetrator offered a free CD to anyone who forwarded it to others with a CC to the perpetrator (Hoaxbusters, undated a).

Another way to harvest e-mail addresses is to send a message that contains a single unseen one pixel image tag. If the e-mail is received and opened, the hidden link accesses an external server and a record that this is a live e-mail address is made. The sender merely needs to guess at e-mail addresses and to use a subject line that is either motivating (social engineering) or appears to be official business in order to get a recipient to open it. The simple method of implementing this tactic is described by Voicenet Communications (undated). Organizations can use e-mail clients that block images, but this in turn creates problems for marketers (e.g., suppliers and other business partners) who send e-mail with legitimate product images (cf. Popov & McDonald, 2004).

Disruption and Consumption of Network and Hardware Resources

Disruption and consumption of networking and computing resources can temporarily inhibit an organization from conducting online commerce (cf. CERT, undated). Gordon et al. (2005) reported that annual business financial loss due to virus attacks were $42.8 million for 700 survey respondents, while denial of service (DOS) attacks were costing $7.3 million. Costs associated with attacks include not only the cost of defence, but also include the cost of lost business. After launching World Series online-only ticket sales, the Colorado Rockies baseball team received a malicious attack of 8.5 million hits. With their online resources swamped, they were forced to stop sales after only two hours and 500 tickets sold (Sports Illustrated, 2007).

In addition to shutting down an organization's Web site server, an organization's e-mail server could be swamped or crashed, temporarily disrupting communications with customers, suppliers, business partners, and employees. A former employee was convicted after crashing the server of a UK-based insurance company by swamping it with five million e-mail messages (BBC News, 2006). More difficult to trace, forged e-mails that contain hundreds of nonexistent recipient addresses in the "copy to" fields can cause some e-mail servers to then forward duplicate messages with huge attachments to those hundreds of e-mail addresses. Even if the recipient name is nonexistent, the server receiving those e-mails can be swamped to the point of crashing. Researchers tested the mail servers of all Fortune 500 companies and found that 30% could be used to make this kind of attack (Knight, 2004).

Disruption and Consumption of Human Resources

British mobile phone retailer Phones 4U started a ban on the use of e-mail on the belief that this would save each employee 3 hours per day. The company moved to communicating via phone and its internal intranet (Thomas, 2003). Deliberately flooding an organization with bogus e-mail messages can consume substantial amounts of recipient time as well as computing resources. Deliberately flooding an organization with electronic greeting cards, chain e-mail, and other such tactics that rely on social engineering can cause employees to waste time with activities that are not related to work. These could cause the organization to quit trusting e-mail or even to quit using e-mail communications altogether; if e-mail attacks are spoofed (faked) to appear to be from a particular business partner, an unscrupulous competitor might be able to cause the receiving organization to block incoming communications from that competitor.

Disruption of Promotion Strategies

These attacks are used to consume a competitor's promotion budget, to discourage a competitor from investing in online promotion activities, or to trick an organization into believing that its online marketing tactics are working better than actual performance. Traffic aggregation is the use of

a domain name, often multiple domain names, to redirect visitors to another Web site (cf. Marsan, 2002). When Web designers build a Web site, they might promise immediate traffic (visitors) to the Web site. Visitors to the client's Web site had believed they were going to some other Web site, but that Web site instead redirected them to the client Web site. A hit log might show a lot of visits to the client's Web site, but these redirected visitors did not end up at the Web site by choice or interest. Through traffic aggregation, the client organization is tricked by the Web site designer into believing that its marketing investment is generating legitimate Web site interest when it is not doing so.

Competitive click fraud is used to either drive up a competitor's online advertising budget, used to either deplete the competitor's promotion budget or to discourage the competitor from advertising online (cf. Claburn, 2005; Lee, 2005). With this tactic, one competitor can repeatedly click on another competitor's online advertising, driving up the advertising competitor's costs in pay-per-click advertising pricing. The competitor can also bid up the prices for pay-per-click rates, discouraging the advertising competitor from even advertising at all. Affiliate click fraud (Stricchiola, 2004) can also drive up an organization's pay-per-click costs, but in this case, the fraud is from advertising hosts rather than from competitors. Although done for financial gain by the advertising host and not necessarily to harm the advertising organization, this is still an external threat to the conduct of online commerce.

Damage to or Disruption of Branding Strategies

These attacks cause a loss of brand equity, customer good will, or brand penetration opportunities. A chain e-mail campaign, for example, is sometimes started to deliberately damage the reputation of a person or firm (Hoaxbusters, undated b). In one chain e-mail, an article is attributed to the popular American radio personality Paul Harvey to give it some level of credibility. The article purports to be providing "facts" which suggest that U.S. Senator Hilary Rodham Clinton, wife of former U.S. president Bill Clinton and a presidential candidate herself as this article is being written, was associated with the defence of Black Panther members accused of tortuous murders in 1969; Ms. Clinton, however, was merely a college student at the time (Snopes, 2005). This chain e-mail has the potential to damage not only the Senator and presidential contender's political reputation, but also the reputation of the radio personality Paul Harvey.

Domain names that attract traffic (visitors) have value and so are attractive to domain name grabbers whose activities can damage a brand name or the reputation of an organization. Domain grabbers register a domain name as it expires in order to capture its traffic. In some cases, the domain is allowed to expire accidentally, but in many cases, it is allowed to expire because the owner believes that it is no longer needed. The Catholic Diocese of Brooklyn, New York, for example, abandoned a domain name that was difficult to remember after registering a new one. It was later embarassed to find that the old name was being used as a host for adult content because it was still attracting traffic (Hardy, 2001). Domain name squatters register a domain name that has value in generating interest, and therefore traffic to the name. San Francisco mayor Willie Brown found his name on multiple domains, used by detractors to drive traffic to content that was unfavorable to the mayor (Learmonth, 1999).

Domain typosquatters register a mistyped variant of a brand name to gain traffic of people who had attempted to visit a Web site associated with the name (cf. Sullivan, 2000). Spoofing is done when someone registers and uses a domain name that is deceptively similar to a trademark or organizational name, potentially harming the brand or organizational name. The domain name irs.com looks very much like it would be the official Web site for the U.S. tax agency, Internal Revenue Service, which is instead located at irs.gov. Problems faced by users of the spoof Web site could ultimately reflect badly on the real organization. In an interesting twist, traffic that might go to the spoof irs.com is stolen by the typosquatter wwwirs.com (the typosquatter is getting visitors who had left out the dot after typing "www").

Social networking and consumer rating/complaint Web sites can also be harmful to an organization's marketing efforts. Phoney pages on the social networking Web site MySpace that purport to be posted by famous business people often contain malicious misinformation (Petrecca, 2006). U.S. presidential candidates Hilary Clinton and Barack Obama were both cast in unflattering roles in anonymously sponsored advertisements that appeared on YouTube (Fox News, 2007). Consumer rating and complaint Web sites, such as Epinions, RateItAll, PlanetFeedback, and My3Cents, can contain unflattering or false information about an organization or brand. Organizations also need to be concerned about unflattering or false information that is posted on forums, blogs, and personal Web sites (cf. Dozier, 2006). Urban legends, as in the chain e-mail which falsely describes Senator Clinton as having defended tortuous murderers, can be spread through social engineering.

FUTURE TRENDS

Due to online commerce, an organization's information resources, computing and human resources, promotion resources, and brand reputation are increasingly more vulnerable than was once the case. Organizations have much control over most of the strategic and tactical points that are outlined above. For example, information leakage and

lost productivity associated with social engineering could be halted altogether; employees could be prohibited from opening executable e-mail attachments even from colleagues if not mission-related, from exchanging mission-unrelated e-mail such as chain e-mail, and from conducting any online activities that are not mission-related. Organizations can dilute spoofing activities by maintaining their own spaces at social networking Web sites such as MySpace and by maintaining, by registering typosquatted and other variations on the domain names associated with new brands. Threats to branding and marketing will continue (and possibly worsen) if organizations do not become more proactive in protecting marketing assets depends on how well these threats can be publicized.

CONCLUSION

Marketing managers and strategists should monitor (and protect) the following:

- Monitor the Core Infrastructure for system probes, scans, and such; monitor for e-mail that is not related to mission-related activities; monitor for the installation of software applications that are not mission-related.
- Monitor the Internal Social Infrastructure for employee mission-unrelated activities that include the exchange of electronic greeting cards, the exchange of chain e-mail, personal e-mail activities, and visits to mission-unrelated Web sites; monitor employees for taking work outside of the physical and virtual confines of the organization; monitor employees for bringing any work or personal activities into the organization.
- Monitor External Social Infrastructures, consumer rating sites, and public forums for the spread of urban legends, unflattering information, and false information; maintain an organizational presence on social networking spaces such as MySpace and YouTube.
- Monitor Internal and External Domain Name Registration to ensure that names associated with organizational brands are not hijacked, grabbed, squatted, typosquatted, or spoofed.
- Monitor Online Promotion Exploits from competitors and online advertising affiliates; monitor potential attempts to siphon Web site visitors via traffic aggregation techniques.

REFERENCES

BBC News. (2006, August 23). Mass e-mail attack teen sentenced. BBC News. Retrieved May 29, 2008, from http://news.bbc.co.uk/1/hi/uk/5278772.stm

CERT. (n.d.). Denial of service attacks. CERT Coordination Center. Retrieved May 29, 2008, from http://www.cert.org/tech_tips/denial_of_service.html

Claburn, T. (2005, May 2). Click fraud threatens rising online ad revenue. InformationWeek. Retrieved May 29, 2008, from http://informationweek.com/story/showArticle.jhtml?articleID=162100620

Dozier, J.W. (2006). The online business cyber-attack: DEFAMATION. IPA's BusinessToday, 1(4). Retrieved May 29, 2008, from http://www.ipabusinesstodaymagazine.com/December06/CyberAttack.asp

Fox News. (2007, March 19). New anti-Hillary Clinton YouTube ad makes waves on Web. Fox News. Retrieved May 29, 2008, from http://www.foxnews.com/story/0,2933,259591,00.html

Gordon, L.A., Loeg, M.P., Lucyshyn, W., & Richardson, R. (2005). 2005 CSI/FBI computer crime and security survey. Computer Security Institute. Retrieved May 29, 2008, from http://i.cmpnet.com/gocsi/db_area/pdfs/fbi/FBI2005.pdf

Hardy, T. (2001, June 11). Porn Web sites using old domains of schools, churches. Scripps-McClatchy Western Service. Retrieved May 29, 2008, from http://www.knoxstudio.com/shns/story.cfm?pk=PORNSITES-06-11-01&cat=AN

Hoaxbusters. (n.d. a). Give away hoaxes. Computer Incident Advisory Capability, U.S. Department of Energy. Retrieved May 29, 2008, from http://hoaxbusters.ciac.org/HBGiveAways.shtml

Hoaxbusters. (n.d. b). Information about hoaxes. Computer Incident Advisory Capability, U.S. Department of Energy. Retrieved May 29, 2008, from http://hoaxbusters.ciac.org/HBHoaxInfo.html

Knight, W. (2004, April 6). E-mail attack could kill servers. New Scientist. Retrieved May 29, 2008, from http://www.newscientist.com/article/dn4858.html

Learmonth, M. (1999, August, 19-25). Invasion of the domain snatchers. Metro, Silicon Valley's Weekly Newspaper. Retrieved May 29, 2008, from http://www.metroactive.com/papers/metro/08.19.99/cover/domains2-9933.html

Lee, K. (2005, February 18). Click fraud: What it is, how to fight it. ClickZ Network. Retrieved May 29, 2008, from http://www.clickz.com/experts/search/strat/article.php/3483981

Marsan, C.D. (2002). Lurid links. Network World Fusion. Retrieved May 29, 2008, from http://www.nwfusion.com/news/2002/0304pornlinks.html

Miller, N. (2003, October 23). Organised crime goes to the top in online attacks. The Age. Retrieved

May 29, 2008, from http://www.theage.com.au/news/business/organised-crime-goes-to-the-top-in-online-attacks/2007/10/22/1192940985254.html

NIAC. (2004a, October 12). Prioritizing cyber vulnerabilities: Final report and recommendations by the council. National Infrastructure Advisory Council.

Oates, J. (2007, October 26). British computer society blunders on BCC. The Register. Retrieved May 29, 2008, from http://www.theregister.co.uk/2007/10/26/bcs_e-mail_gaffe/

Petrecca, L. (2006, September 25). If you see these CEOs on MySpace . . .. USA Today. Retrieved May 29, 2008, from http://www.usatoday.com/money/industries/technology/2006-09-24-fake-ceos-usat_x.htm

Popov, K., & McDonald, L. (2004, September 29). Blocked e-mail images. ClickZ. Retrieved May 29, 2008, from http://www.clickz.com/3413471

Snopes. (2005, September 18). Black Panthers. Snopes.com. Retrieved May 29, 2008, from http://www.snopes.com/politics/clintons/panthers.asp

Sports Illustrated. (2007, October 23). After cyber attack, Rockies will resume online ticket sales. Sports Illustrated. Retrieved May 29, 2008, from http://sportsillustrated.cnn.com/2007/baseball/mlb/wires/10/23/2010.ap.bbo.rockies.series.tickets.1st.ld.writethru.0480/

Stricchiola, J. (2004, July 4). Lost per click: Search advertising and click fraud. Search Engine Watch. Retrieved May 29, 2008, from http://searchenginewatch.com/searchday/article.php/3387581

Sullivan, B. (2000, September 23). Typosquatters turn flubs into cash. ZDNet News. Retrieved May 29, 2008, from http://news.zdnet.com/2100-9595_22-502915.html?legacy=zdnn

Thomas, D. (2003, September 23). Retailer's e-mail ban highlights the dangers of lost productivity. ComputerWeekly.com. Retrieved May 29, 2008, from http://www.computerweekly.com/Articles/2003/09/23/197399/retailers-e-mail-ban-highlights-the-dangers-of-lost.htm

Thurrott, P. (2001, February). Microsoft's internal network breached. IindowsIT Pro. Retrieved May 29, 2008, from http://www.windowsitpro.com/Articles/ArticleID/16435/16435.html?Ad=1

Vamosi, R. (2004, June 2004). Beware of keystroke-logging RATs. CNET. Retrieved May 29, 2008, from http://reviews.cnet.com/4520-3513_7-5138138-1.html

Voicenet Communications. (n.d.). Broadcast by e-mail—open tracking. Voicenet Communicatins. Retrieved May 29, 2008, from http://www.voicent.com/track-email-open.php

KEY TERMS

Chain E-Mail, Chain Letter: An e-mail letter that directs recipients to forward multiple copies of the same letter to others.

Click Fraud: Clicking on an online advertisement link for the premeditated purpose of causing a pay-per-click advertiser to pay for the click without the intent to take any other actions (such as buy a product).

Domain Name Grabbing: Registering an abandoned or lapsed domain name immediately after it is released by a registrar.

Domain Name Squatting: Registering a trademark, an organization's name, or person's name as a domain name with the intention to profit from traffic to an unrelated Web site or by reselling the domain name back to the person or organization.

Exploit: An action that takes advantage of weaknesses or vulnerabilities in software or hardware.

Social Engineering: Manipulating people through their natural trust or desire to help in order to trick them into divulging information or performing actions.

Spoofing: Pretending to be the owner of a trademark or organization name; registering and using a domain name that is deceptively similar to a trademark or organization name. This could be through transposing words or inverting a phrase.

Traffic Aggregation: Using a domain name, often multiple domain names, to drive traffic (visitors) to one Web site.

Trojan: A malicious program that is hidden within a seemingly useful and harmless program. Also known as a Trojan horse.

Typosquatting: Registering and using a domain name that is a misspelled variation of a trademark or organization name.

Urban Legend: A story that contains some measure of truth or fact but is embellished with misinformation and repeatedly passed from person to person.

Measurement Issues in Decision Support Systems

William K. Holstein
University at Albany, State University of New York, USA

Jakov Crnkovic
University at Albany, State University of New York, USA

INTRODUCTION

The past decade has seen tremendous progress in systems for information support–flexible and adaptable systems to support decision makers and to accommodate individual needs and preferences. These model- or data-driven or hybrid decision support systems (DSS), now often called business intelligence (BI) systems, incorporate diverse data drawn from many different internal and external sources. Increasingly, these sources include sophisticated enterprise resource planning (ERP) systems, customer relationship management (CRM) systems, data warehouses and other enterprise-wide systems that contain vast amounts of data and permit relatively easy access to that data by a wide variety of users at many different levels of the organization. Decision support, DSS and BI have entered our lexicon and are now common topics of discussion and development in large, and even in medium-sized, enterprises. Now that DSS is well established, attention is turning to measurement and the metrics that populate such systems.

BACKGROUND

Decision-making as we know it today, supported by computers and vast information systems, is a relatively recent phenomenon. But the concept has been around long enough to permit the methods and theories of decision-making to blossom into "a plethora of paradigms, research schools, and competing theories and methods actively argued by thousands of scientists and decision makers worldwide" (Robins, 2003).

Early computer systems focused primarily on accounting and financial data. It is said that information systems are about transforming data. We could say that early systems transformed data into aggregated or summarized data – for example, wage rates, hours worked, benefits and tax data, and so forth transformed into departmental or corporate payroll reports.

In the mid-1960's, the development of the IBM System 360 and rapidly proliferating competitive systems from other vendors ushered in the era of Management Information Systems (MIS). Applications quickly moved beyond finance and accounting data and into operations. Transaction processing systems began to generate order, usage, and customer data that could be analyzed with (what quickly became quite sophisticated) models. The transformation of data into information became commonplace. For example, data on sales and usage, costs, supplier lead times and associated uncertainties were transformed into reorder points, safety stocks, and comprehensive inventory management and production scheduling systems.

Despite the broader reach of MIS, such systems are characterized by highly structured, infrequent reports, often with standard formatting. Frequently, because it was "easier" (for the IT staff), each manager in a given function, for example, marketing, received the same voluminous report – even though a manager of activities in Japan could not careless about data relating to New Jersey. Despite the tremendous advance of MIS over previous-generation systems, contemporary MIS systems draw most of their data from enterprise resource planning (ERP) systems that contain mostly internal data on transactions, and therefore suffer from many of the same problems as older systems (an internal, historical, and financial focus).

Decision support systems "evolved from the theoretical studies of organizational decision making done at the Carnegie Institute of Technology during the late 1950s and early '60s and the technical work on interactive computer systems, mainly carried out at the Massachusetts Institute of Technology in the 1960s" (Keen & Scott Morton, 1978; Power, 2003). By the end of the 1970's, it was clear that model-based decision support had become a practical, useful tool for managers.

A 1970 article by John Little of MIT clarified the concept of decision support. In a 1979 paper he provided a definition that is paraphrased here:

A coordinated collection of data, systems, tools, and techniques along with requisite software and hardware, by which an organization gathers and interprets relevant information from the business and environment and turns it into a basis for action.

Another useful definition of a DSS is:

Interactive computer-based systems designed to couple the intellectual resources of individuals with the capabilities of the computer to utilize data and models to identify and solve semi-structured (or unstructured) problems and improve the quality of decisions (paraphrased from Gorry & Scott Morton, 1989)

In these two definitions, we see some important concepts–gathering and interpreting relevant information (related to the decision at hand, not just to transactions), using the intellectual resources of managers, and providing information that can be used as the basis for action. The "new idea" here was that managers need more than information, they need decision support. If provided with good data, and models and tools to transform the data into useful information, their effectiveness will improve.

As the field has evolved, the term *Business Intelligence* has come to be used for the types of systems that were previously referred to as DSS. A simple definition of business intelligence fits well with the DSS definitions given earlier: "Technologies that help companies make better business decisions" (www.orafaq.com/glossary).

METRICS OF BUSINESS AND MANAGEMENT PERFORMANCE

The definition of decision support, or the capturing of business intelligence, is supporting managers who are running the business. Increasingly, it refers to supporting middle-level managers who rely on a mix of internal and external data that is steadily tilting towards external data on customers, markets, competitors, and the political, regulatory and economic environment. If we define the process of *control* as tasks undertaken by middle- and lower-level managers to *ensure that plans come true*, we see clearly the role of data and information in decision support: managers use data and convert it into information to *monitor* the implementation of plans to ensure that strategic goals are met. If the monitoring indicates that plans will not be fulfilled, corrective *action* must be taken *in time* to ensure that the plan is in fact met. If the information from a decision support system cannot serve as the basis for action (i.e., cannot first help the decision-maker to decide to do something, and then help to decide what to do) the information will not be used and the system will therefore be useless.

The keywords in the previous paragraph that lead to action are *monitoring* and *in time*. Monitoring is the management function that is the primary target for DSS implementation. Timeliness is crucial; advance warning without enough time to steer around the iceberg, or to make the necessary changes to ensure that strategic plans are successful, is not the kind of decision support that managers seek.

In recent years, we have seen the emergence of *operational business intelligence*, the same concept as the older BI and DSS, but focused on shorter-term, operational decision-making.

Operational BI most differs from BI for management and control purposes in both the level of detail required and in the timeliness of the data. Operational BI may involve accessing a transaction system directly or through a data warehouse that is updated in near real-time multiple times throughout the day. Business intelligence for management and control purposes may also be in near real time, but can also be based on weekly or monthly data (Howson, 2008).

As we think about supporting management decision-making, we must think of how managers work at decision-making. What they do is easy to describe (despite the fact that it is fiendishly difficult to do it): managers abhor irregularities and plans that do not come true, yet they thrive on exceptions. They look for things that do not fit, that look funny, and that are out of line. Then they ask why. Much of their time is spent trying to answer that simple question and searching for actions that will make perceived problems disappear and bring things back to "normal expectations".

Examples of the "whys" that plague managers of large companies include:

- Why is it that Cadillac does not attract younger buyers?
- Why did the PC manufacturers who dominated the market in the 1990's lose so much share to Dell Computer?

Shorter-term examples include:

- Why did that ad campaign not work? (Think of Infiniti's ad campaign when the brand was introduced in 1989-90 – the ads never showed the car and while considered creative, it started the brand on an also-ran trajectory that to this day, has not caught up with its Japanese and European luxury-car rivals.)
- Where are the bottlenecks in our supply chain?
- What can we do to fix a supplier who is behind on delivery of a critical component? (Think of the current problems of the Airbus A380 and the Boeing 797).

For each of these questions, one can imagine a manager who is conjuring the question as a response to a perceived exception that needs to be "fixed".

- Which Cadillac sales manager thinks that the product is not attractive to young buyers?

- What ad manager thinks the campaign failed?
- What Boeing manager wants to contemplate dropping a supplier and finding another one immediately?

And yet, in all cases, the exception is obvious and something has to be done.

We cite this process and these questions to focus clearly on metrics of business performance and management performance. Measurement and metrics are the tools for identifying exceptions. Exceptions, in turn, drive management to seek and find actions that will deal with the exceptions and achieve strategic goals. But think for a moment about traditional metrics; the Cadillac manager knows how to measure the average age of Cadillac buyers. But how does the manager measure the potential attractiveness to younger buyers of a proposed new model? How does the manager measure the potential attractiveness of a proposed advertising campaign? How does the manager measure the potential attractiveness of a proposed discount or rebate program? This is where judgment, experience, and intuition come into play – and these are precisely the areas where managers need decision support.

In the current environment, measurement must be related to business matters, business strategies and goals – the stuff that managers deal with in their everyday environment. They are trying to formulate and monitor plans that reflect the strategic mission and goals of the business, that is, accomplish strategic tasks. They need IT that can add value to the business in ways that they can clearly understand.

We turn now to implementation issues surrounding these ideas and suggest some guidelines for DSS development incorporating new metrics.

IMPLEMENTATION GUIDELINES

Breadth in Measurement

A first guideline, as we have discussed, is that breadth in measurement (beyond financial and accounting measures) is important. The ideas behind the Balanced Scorecard (Kaplan & Norton, 1992) should be understood and implemented in decision support systems; breadth, however, does not imply complexity.

Consider the following quote – it may be true since it does not say "*only* three things" – and it fits the point here: "The ability of companies to boost profits depends on three things: how high they can lift their prices, how much they can increase output per worker, and how fast wages are rising" (Mandel, 2002).

The ability to raise prices, often described as "pricing power", is related to many different metrics and measures. Some of these measures are "hard", such as ROI, return on capital employed, turnover, profitability of other divisions, products and services, and capital structure; but many are not. Examples are "soft" metrics such as customer perception of the company's value proposition, features and benefits of products and services, embedded technology, total quality (not just product quality), competitive position in the marketplace, relationship with customers, suppliers, and workers. Increasing worker productivity involves similarly complex metrics.

Simplicity

A second guideline is simplicity. Metrics must be easy to understand and communicate. They must relate to relevant activities and tasks, and be "drivable", that is, managers must be able to use the measures to determine (drive) actions that will affect future results. The following quote makes this point very effectively and provides some detail on what the metrics should accomplish:

…Companies need an 'organizational magnifying glass' – something that focuses the work of employees so everyone is going in the same direction. Strong leaders do this. However, … they need more than just the force of their personality and experience to focus an organization. They need an information system that helps them clearly and concisely communicate key strategies and goals to all employees on a personal basis every day (Eckerson, 2005).

Selectivity

Our third suggestion relates to simplicity: DSS designers must avoid the impulse to measure everything. The focus must be only on the most important metrics from the user's (the decision maker's) point of view. Metrics should be built in hierarchies, with more details for lower-level managers, and fewer, more summarized measures for higher-level managers. Here is another quote from the previous source that emphasizes the importance of not only focus, but focusing on only the most important tasks:

The system should focus workers on tasks and activities that best advance the organization's strategies and goals. It should measure performance, reward positive contributions, and align efforts so that workers in every group and level of the organization are marching together toward the same destination. In short, what organizations really need is a performance dashboard that translates the organization's strategy into objectives, metrics, initiatives, and tasks customized to each group and individual in the organization (Eckerson, 2005).

Research and Learning

Fourth, DSS design should emphasize data and data collection, not just for reporting, evaluation and auditing, but for research and learning, for finding exceptions, for learning the root causes of exceptions and for exploring alternative courses of remedial action. "Research" is perhaps a strange term to use in the context of management practice, but research, in the best sense of the term (investigate, study, explore, delve into, examine) is required to find, first, meaningful measures of exceptions (*deviations* in the following quote) and, then, their causes.

Feedback is essential for control of any system. Without the feedback provided by sight, sound and touch, we humans would not be able to identify threatening or favorable situations and there would undoubtedly be fewer of us on the planet. The same is true for companies. When managers don't have timely and meaningful feedback, companies fail to recognize opportunities and become much more vulnerable to hazards that can threaten their existence.
The feedback provided by performance measures gives managers better control over their areas of responsibility, whether it is a department, a plant, or a division. With measures in place, deviations in performance are detected earlier, enabling managers to step in and minimize the damage or make the most of the opportunity. Performance measures also prevent managers from getting blind-sided with bad news (Kados, 1998).

Benchmarking

As a fifth guideline, we cannot overemphasize the importance of benchmarking against credible external targets. Indeed, without a firm connection to good external benchmarks (best practice, best-of-class indicators) companies can fall victim to *manumation*, simply automating old, outdated, manual processes. A formula for manumation has been around for years:

OP + NT = EOP
(**Old Processes** plus **New Technology** equals **Expensive Old Processes)**

Benchmark analysis can identify problems and suggest solutions and can serve as an excellent idea bank for new metrics. Benchmarks or comparatives are important because metrics need anchor points for comparison. Without a benchmark for "normal" or "best in class", how can you gauge results?

Simple benchmarks might include your own past performance, current goals, customer expectations for things like order-to-delivery time, percentage defects, or on-time delivery. In particular, you should know how you compare to others in your industry and leaders in your functional area. Be sure to think carefully about which comparatives will lead to valid conclusions and sensible action.

In some industries and functions, there are a growing number of highly useful benchmarks from trade associations, consulting companies and other organizations. An example of a widely used set of metrics is the Supply Chain Council's Supply Chain Operations Reference (SCOR) model (Supply-Chain Council, 2004). SCOR allows companies to objectively measure their supply chain practices and compare them against benchmark standards gathered from the more than 700 manufacturing and related companies that are members of the Supply Chain Council. The SCOR model groups supply chain functions into five process categories: Plan, Source, Make, Deliver, and Return. Metrics at each level of the model are supported by progressively more detailed metrics for processes at lower levels.

The following quote highlights the importance of metrics to support process change when new factors, such as Web-enabled processes, are introduced:

Updating performance metrics for Web-enabled supply chain operations became important the day your company migrated from an information-only Website to one with interactive customer capabilities. Whether you're tracking customer orders or collaborating with partners on products and processes, your new business dynamics likely don't fit the old criteria, and you may need to update your basic supply chain performance measurements to match an advanced level of attainable objectives (Schultz, 2001).

Benchmarking is useful beyond performance measurement. It can help to answer the "why"? when exceptions are identified. For example, is a manufacturer's frequent late delivery an inventory-level issue? Or is it caused by slow order-reaction-time on the part of one or more of the supply chain's participants? Benchmarking can help to target on the exact answer, which can then lead to needed adjustments.

Time

As a sixth and final guideline, we suggest careful consideration of collapsing time. Speeding up of business processes and reaction times has become almost a cliché, but time is the most important element in many new metrics to support decision-making. Consider this example from the previously quoted article:

If I said today that I need 30 percent more of something, how fast can my suppliers, and their component suppliers, deliver that? [Only] Part of the answer is in manufacturing lead time, [The other, equally important] part is administrative lead times – getting the information, sharing it, and synchronizing it.

FUTURE TRENDS

Pulling together what we have stated thus far, we see a situation that precludes significant progress in the development and implementation of decision support systems unless:

- New metrics that focus more clearly on business and management performance are developed and implemented, and
- More attention is given to metrics that focus on monitoring strategic activities, or activities that will have an important effect on the outcomes of strategic initiatives and high-level company goals.

The latter is particularly important. Without wishing to denigrate the importance of support for very-short-term operational decisions, the real prize is in supporting longer-term, more strategic decisions. This also means gathering and interpreting data and information from the lowest levels of the organization (the "front lines") and processing vast amounts of current information in a timely manner. Stated otherwise, it means taking advantage of supporting operational people responsible for short-term performance, but also making it the basis or platform for support for more long-term, strategic decisions.

A good example of this need is captured in the following quote from a recent book on dashboards, graphical interface displays of DSS data (Alexander, 2007):

…(many companies have) one dominate business, with smaller but different business units in the portfolio. Managers have a tendency to apply a single business model, expecting similar rations and performance across the businesses, which can result in dysfunctional decisions and missed opportunities. ... For example, there may be an opportunity to build a business based on the current product line, but requiring lower pricing and therefore lower costs. Managers may pass on this opportunity because of lower expected gross margins. However, it is possible that this product line may require lower levels of selling, general, and administrative spending and inventory. This may result in returns approximating or even exceeding the levels achieved by the high-end business.

This example raises interesting and provocative questions for systems designers: How can the elements of a "business model" be incorporated in a DSS or BI system, but not so deeply ingrained that it prevents good thinking about alternatives to the model? How can data from one business be "modeled" differently to see the effects of a different approach to the business? And how can that different model form the basis for a different system to support decision-making?

Early work at MIT focused on "Critical Success Factors" or "Key Performance Indicators" as the basis for metrics in information systems (Rockart, 1979; Rockart & Treacy, 1982). These metrics were defined to ensure successful competitive performance for the organization. They were incorporated in reports, and backup systems provided the ability to drill down to underlying detail, budget information, plans and objectives, competitive information, news, and more to determine the "drivers" of the success factors. More recently, the Balanced Scorecard and other initiatives have made it clear that traditional MIS systems and business metrics rely too heavily on financial and accounting measures (Kaplan & Norton, 2006). Today, the issue of metrics is very active with performance management systems, digital dashboards and cockpits and scorecards to drive organizational performance measurement and management down to the individual level in order to support alignment with business strategies (Akpan, 2007; Power, 2003).

Financial measures tend to be hard, historical, and internal, rather than soft (including judgmental data and estimates that, while "inaccurate", are vital to future planning), future-oriented (and therefore, by definition, "soft") and external (related to the customer, the market and the environment). We see, therefore, that an important question in the future of metrics in decision support systems and business intelligence is not just *what* to measure, but *how* to measure, and *in which areas of the business* to seek meaningful metrics.

CONCLUSION

The earlier quote about a 30% increase in output (Schultz, 2001) deftly summarizes much of what we have discussed. We see a clear exception – the need to increase output by 30%, quickly–and the immediate following need to organize the required information internally, and then to communicate that information externally to suppliers and on to their suppliers.

The key element in the processes that are involved here is time, not cost. So, first, metrics to ensure that these processes are working properly will be drawn largely from nonfinancial data, including a large slug of temporal (relating to time) data.

Next, there is a pressing need to focus only on the most important, but not necessarily the most obvious, issues and to collect relevant data to support understanding of how the task will be accomplished. Much of the learning will be based on intensive communication, with suppliers and their suppliers, probably much of it will be Web-based.

Previous benchmarking outside the organization with suppliers and processes would be highly valuable at this point–indicating who can perform and who cannot, who can respond quickly, who has people that we know and can trust.

Summarizing our implementation guidelines for the development of metrics for decision support or business intelligence systems:

- Think beyond ROI and payback metrics for IT investments (Renkema, 2000)
- Measurement for business and management performance is more important
- Think beyond financial measures
- Focus only on the most important, but not necessarily obvious, issues and collect relevant data to support exploration, analysis and understanding
- Benchmark outside the organization and build relevant knowledge to support change, and
- Do it fast, and make it possible for users of the system to work quickly as well.

REFERENCES

Akpan, E. O. (2007). *Strategic alignment: The business imperative for leading organizations*. Tate Publishing.

Alexander, J. (2007). *Performance dashboards and analysis for value creation.* Wiley Finance.

Eckerson, W. W. (2005). *Performance dashboards: Measuring, monitoring, and managing your business*. Wiley.

Gorry, G. A. & Scott Morton, M. S. (1989). A framework for management information systems. *Sloan Management Review, 13*(1)

Howson, C. (2008). *Successful business intelligence: Secrets to making BI a killer app*. McGraw-Hill.

Kaydos, W. (1998). *Operational performance measurement: Increasing total productivity.* CRC.

Kaplan, R. S. & Norton, D. P. (2006). *Alignment: Using the balanced scorecard to create corporate synergies*. Harvard Business School Press.

Kaplan, R. S. & Norton, D.P. (1992). The balanced scorecard—Measures that drive performance. *Harvard Business Review*

Keen, P. & Scott Morton, M. S. (1978). *Decision support systems: An organizational perspective.* Addison-Wesley, Inc.

Little, J. D. C. (1970). Models and managers: The concept of a decision calculus. *Management Science, 16*(8)

Little, J. D. C. (1979). Decision support systems for marketing managers. *Journal of Marketing, 43*

Mandel, M. J. (2002). More productivity, more profits? *Business Week*.

McCosh, A. M. & Scott Morton, M. S. (1978). *Management decision support systems.* London: Macmillan.

Power, D. J. (2003). *A brief history of decision support systems*. Retrieved June 15, 2008, from DSSResources.com

Renkema, T. J. W. (2000). *The IT value quest: How to capture business value of IT-based infrastructure.* John Wiley & Sons

Robins, E. (2003). *A brief history of decision-making*. White Paper from the Technology Evaluation Corp. Retrieved June 15, 2008, from http://www.technologyevaluation.com

Rockart, J. F. (1979). Chief executives define their own data needs. *Harvard Business Review*

Rockart, J. F. & Treacy, M. E. (1982). The CEO goes online. *Harvard Business Review*

Rockart, J. F. & DeLong, D. W. (1988). *Executive support systems: The emergence of top management computer use*. Homewood, IL: Dow Jones-Irwin.

Schultz, G. (2001). Advanced performance metrics for the e-era. *Technology Edge*.

Scott Morton, M. S. (1967). *Computer-driven visual display devices – Their impact on the management decision-making process*. Unpublished doctoral dissertation, Harvard Business School.

Scott Morton, M. S. & McCosh, A. M. (1968). Terminal costing for better decisions. *Harvard Business Review*

Supply-Chain Council (2004). *Supply-chain operations reference model overview* (Version 5.0). Retrieved June 15, 2008, from http://www.supply-chain.org

SUGGESTED ADDITIONAL REFERENCES

Friedlob, G. T., Schleifer, L. F., Plewa, F. J., Jr. (200). *Essentials of corporate performance measurement*. Wiley

Frost, B. (2000). Measuring performance : Using the new metrics to deploy strategy and improve performance. *Measurement International*.

Kaplan, R. S. & Norton, D. P. (196). The balanced scorecard: Translating strategy into action. *Harvard Business School Press*

Malik, S. (2005). *Enterprise dashboards: Design and best practices for IT.* Wiley

Marakas, G. M. (2006). *Decision support systems* (2nd ed.). Prentice Hall

Turban, E., Aronson, J. E., Liang, T. P., & Sharda, R. (2006). *Decision support systems and intelligent systems* (8th ed.). Prentice Hall

KEY TERMS

Decision Support System (DSS): Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. [They comprise] a computer-based system for management decision makers who deal with semi-structured problems (Gorry & Scott Morton, 1989).

Business Intelligence (BI): The process of gathering information in the field of business. Information is typically obtained about customer needs, customer decision making processes, the competition, conditions in the industry, and general economic, technological, and cultural trends. Business intelligence is carried out to gain sustainable competitive advantage, and is a valuable core competence in some instances. The term was first used by Gartner and popularized by analyst Howard Dresner. It describes the process of turning data into information and then into knowledge. The intelligence is claimed to be more useful to the user as it passes through each step (http://explanation-guide.info/meaning/Business-intelligence.html)

Hard Data: Historical, usually accurate, data, often from transaction processing systems.

Soft Data: Judgmental, qualitative data, often from external sources; any data involving the future.

Actionable Information: Information that can be used as the basis for a decision, or for taking action, usually to change something.

Metric: A predetermined measure that will be used as the basis for a measurement process. For example, percentage of customer calls answered within one minute.

Benchmark: A standard, usually from outside sources and usually representing the best, or better than average, performance against which an activity's metric is compared. For example, world-class competitors have 35 defects per unit within the first 6 months; we have 85.

Measurement: The process of determining values representing performance, or the results of the process. For example, the measurement process is now started, or the measurement was 35 days.

Balanced Scorecard: An approach for measuring business and management results that goes well beyond financial metrics. Several "perspectives" are suggested, for example financial, customer, internal processes and learning and innovation (Kaplan & Norton, 1992).

Measuring Collaboration in Online Communication

Albert L. Ingram
Kent State University, USA

INTRODUCTION

Collaboration has become a key concept in the workplace, in research laboratories, and in educational settings. Companies want members of different departments located far apart to work together. Various government agencies try to establish collaborative relationships with private organizations. Academics and corporate researchers collaborate with far-flung colleagues to produce new knowledge. Students at all levels of our educational system are increasingly being asked to learn collaboratively. In addition, more work is being done online. Businesses communicate over the Internet, and increasing numbers of educational experiences are being delivered at a distance. Virtual high schools, traditional and for-profit distance education institutions, and colleges and universities are all among the current users of the Internet in education.

In all of these situations—educational and non-educational, face-to-face, and online—several questions need to be addressed. First, what is collaboration? The word is sometimes used as if everyone already understands what it means, but we can find a variety of different definitions in the literature. Second, when we form groups to collaborate, how do we know when they have done so? Is it possible to measure the extent to which collaboration has occurred in a given group and setting? Third, what actions and conditions enhance the collaboration that does take place? And finally, does collaboration work? That is, do groups that are more collaborative produce better results or learning than groups that are less collaborative?

This brief article will not attempt to answer all these questions, but it will concentrate on a specific issue: What methods can be used to determine whether, and how much, collaboration has occurred in online groups in various settings? We will explain our preferred definition of collaboration, based on previous research, and then discuss some of the implications of these ideas for online collaboration and for research into that issue.

BACKGROUND

Collaboration can be generally described in a variety of ways, but perhaps a typical definition is "working in a group of two or more to achieve a common goal" (McInnerney & Roberts, 2004, p. 205). Such a general definition, however, does not tell us how reliably to identify when collaboration has taken place or, assuming that there can be degrees of collaboration, how much of it is going on. To make such measurements, we need an operational definition of collaboration. Recently, Hathorn and Ingram (2002) proposed such a definition. They maintained that collaboration consists of at least three key ingredients: interdependence (Johnson, Johnson & Smith, 1998), a product that is achieved through genuine synthesis of information and contributions from all members (Kaye, 1992), and independence from a single leader (Laffey, Tupper, Musser & Wedman, 1998). In education, this would likely be independence from the class instructor. In other settings, it would mean relative independence from supervisors or others who might otherwise control the process too tightly.

Under this definition, collaboration contrasts sharply with what can be called a *cooperative* way of working. In this characterization, cooperation occurs when a group agrees to divide the work among them, with each taking part of the project. The final product, then, is the sum of separate contributions from each member, rather than being a true synthesis as in a good collaborative effort (Hathorn & Ingram, 2002; Ingram & Hathorn, 2004; Dillenbourg, Baker, Blaye & O'Malley, 1996).

Hathorn and Ingram (2002) operationalized their definition by looking at ways of measuring each of the three components of collaboration. Positive interdependence occurs when group members share information and test their ideas on one another. When individuals in a collaborative group work toward their common goal, they often achieve things that would not have been possible individually (Henri, 1992; Kaye, 1992). Synthesis occurs as the group attains new insights as a result of working together (Henri, 1992; Kaye, 1992). Finally, independence requires that the group function on its own without too much centralized direction (Laffey et al., 1998). Otherwise, it is a directed project, not a collaboration among equals.

MEASURING COLLABORATION

In the literature we can find a variety of ways to measure collaboration that have been used by teachers and researchers. In general, these break down into a few major categories: teacher or leader observations, student and participant self-ratings and self-reports, and quantitative analysis of discussion transcripts. Here we look briefly at each of these in turn.

Teacher/Leader Observations

Sometimes an instructor or a team leader can have a very good "feel" for how well a group is collaborating. By scrutinizing the team in action and examining the products that result from the group work, these observers can often tell who is participating fully and contributing to the results, and who is not. Frequently, however, teachers and others may assume that simply putting people into groups automatically results in high-quality collaborative work. This assumption is false: good collaboration requires many factors, and casual observations may not reveal what is really going on. In many cases, online collaborative groups can be easier to observe than face-to-face groups, because all the conversations may be recorded automatically, depending on the software and systems used.

Student and Participant Self-Ratings and Self-Reports

In many instances, members of groups may know how well they are working together. For instance, a frequent complaint of students doing group work for classes is the uneven distribution of the workload. Finding ways to get clear and reliable self-reports from students and other participants in collaborative groups can lead to better understanding of how the groups operate. There is a danger in this, however, because group members may not have a clear understanding of what it means to collaborate effectively. This is especially true if they have never experienced high-quality collaboration themselves. Many groups, especially in education, seem to prefer a "divide-and-conquer" cooperative strategy that appears to them to be collaborative. In fact, it lacks both the interactions and synthesis necessary for good collaboration, because each member of the group works on just part of the whole project. Therefore, in order both to increase the actual collaboration among group members and to improve the reliability and validity of the self-reporting, it is necessary to teach people the characteristics of good collaboration, how to recognize those characteristics, and how to produce them.

Quantitative Measures

Finally, we look at quantitative measurements of whether collaboration has occurred and of its extent. One approach was taken by Wilczenski, Bontrager, Ventrone, and Correra (2001). They measured the behaviors in a group that facilitated and detracted from the collaboration, under the assumption that groups with more facilitative behaviors would be more collaborative. The study showed that groups exhibiting more facilitative behaviors did better on several measures.

Hathorn and Ingram (2002; Ingram & Hathorn, 2004) also took a quantitative approach. Based on the definition of collaboration cited above, they developed measures of its three main components: interdependence, synthesis of contributions, and independence. Specifically, they applied these concepts to asynchronous threaded discussions, although the same ideas could be useful in other contexts as well (e.g., synchronous online chats). They relied on close and detailed content analysis of the discussions themselves (Silverman, 1993). Rourke, Anderson, Garrison, and Archer (2001) noted that a key step is to develop a way of coding the discussions to illuminate the questions one wants to answer. Hathorn and Ingram (2002; Ingram & Hathorn, 2004) developed such a system for the construct of collaboration, noting the inadequacy of many previous schemes for analyzing online collaboration specifically.

In order to use these measures, one needs complete transcripts of the discussions. Online textual discussions are especially useful in this regard since the transcripts are usually kept automatically in both synchronous and asynchronous discussions. Conceivably, the actual medium of communication could be instant messaging/chat, e-mail (including listservs), threaded discussion boards, or other text-based systems, as long as the technology can keep complete logs of the discussions. In Hathorn and Ingram's (2002) system, coding is based on "statements" made in the discussions. Statements are sentences or complete ideas within sentences that represent individual idea units. A single message can contain just one statement or numerous statements on a variety of topics. Indeed, a single sentence can contain multiple statements.

Interdependence is identified using several criteria. First, it requires roughly equal participation among all members. Without that, it is difficult to see how the members can be meaningfully interdependent. Participation is measured primarily by the number of messages and/or statements contributed by each group member. The count of statements is probably a more accurate measure of actual participation than number of messages, sentences, or words would be. It is unlikely in any group that the members participate exactly equally by any measure, so the requirement for good collaboration is that there be at least roughly equal participation. A simple test for this is a chi-square analysis on the participation of the group members. If the test shows significance,

it is likely that the members were not participating equally enough for the group to be considered collaborative.

Beyond simple participation, we can measure interdependence by looking at interactions among the group members. In particular, Hathorn and Ingram (2002) focused on direct participation in the substantive discussion (as opposed to off-task comments and other such contributions). The patterns of discussion that indicate actual interaction are threads where different participants refer, explicitly or implicitly, to one another's comments. The minimum length of such a thread that would indicate true interaction is three statements: an initial comment, a response to that comment, and a synthesizing response. The more such threads (and longer ones) appear in a discussion, the more interaction there is and, all other things being equal, the more collaboration is taking place. A useful technique at this point is to diagram the discussion, using the statement as the unit. Each statement appears in a separate box, and lines connecting them show the patterns. It is important not to diagram the messages, because in our experience these do not reflect the substantive patterns that are revealed by the statement-level analysis.

Next, the question of synthesis arises. True collaborations result in products that cannot be identified as resulting from individual efforts. This can be measured in two ways. First, the patterns of interactions noted above include the necessity for synthesizing responses, so that was the first and more detailed measure of synthesis. The other measure focused on the final group product and whether it was written by an individual or the entire group.

Finally, independence from a central authority can be measured by examining both the basic participation of the group members in relation to the instructor and the number of interactions that take place without the instructor's or leader's participation.

Thus, in this quantitative view, the extent of collaboration is revealed by conducting a close content analysis based on the transcripts of the online discussions. Measuring the three elements of collaboration depends on several key pieces of data: overall participation based on the number of statements made by each individual, the types of statements they make (on-task, off-task, and others), and the patterns of interaction, represented especially by the basic unit of collaboration consisting of comment-response-synthesis. The nature of the actual product developed by the group is also important.

Measuring Collaboration in the Real World-An Example

Jim is project manager for a medium-sized information technology firm. His team is scattered around in several locations and needs to work together using a variety of different technologies. Specifically, his supervisor expects him to ensure that they are collaborating well in spite of the distance, the different time zones, and other obstacles. Naturally, the most important concern for Jim and his supervisor is the quality of the work. So far, this has been acceptable, but not outstanding. Jim thinks his team can do better if they learn to collaborate more effectively.

Being a systematic sort of person, Jim decides to approach the problem carefully. So far, simply exhorting the team to do better has not worked, so he wants to find out where they might be failing to collaborate well. To do that, he needs some data about their current efforts. One advantage he has is the fact that much of the work done by the team is online and text based; frequently there is an automatic record kept of the conversations that people have. For example, the team's private collaborative Web site includes a threaded discussion board where everyone can read all the messages and contribute to the conversation. In addition, all documents produced by the team are stored in the online repository in their space. Finally, there is the opportunity for synchronous chats among members while they are working on specific pieces of the project. Jim has decided to record and archive those chats for future use.

Jim's first step in deciding whether his team is collaborating productively is to scan the communications that they have made. At first glance, it looks as if they are all participating and making contributions, although he did not read everything in detail. He decides to try something else before he goes through that lengthy chore.

First, Jim sends an e-mail to everyone in the team asking them to assess the level of collaboration in the team and to e-mail him back by Friday. Very quickly he receives numerous questions about how he defines collaboration, what the request means, and so forth, so he decides to try another approach. This time, Jim does a little research about what is meant by collaboration and sends a more specific e-mail message to the group. Now, instead of asking a global question, he asks people to rate whether the members of the team participate roughly equally and respond well to one another, whether they synthesize each other's contributions, and whether they work independently of the leader. (That last one is tricky for team members, since Jim *is* the leader.)

This time, people are able to answer, and the responses are interesting. In general the team thought they were participating equally and responding to each other. They had mixed views on whether their resulting products were really the result of synthesizing ideas from many sources within the group. They all agreed (tongue in cheek?) that they could get along without Jim for most of their work.

Now Jim has a choice. Is the information he received from the group enough to make decisions about the direction he needs to take? One possibility is that he could arrange to train his team on collaborative techniques, especially using online technologies. Training can be expensive, though, and Jim does not relish taking people away from their jobs for the time it would take. If the team would benefit, then he

believes it would be worth it. He is still not sure that they will (or will not) gain enough from such training.

Perhaps the self-reports of his team members are good enough for Jim to make the decision. If the stakes are high enough, however, Jim might want to look further. In that case a more detailed content analysis might pay off. His team members might overestimate the extent to which they respond to each other's message. Often an online discussion proceeds more as a series of monologues than as a true collaborative conversation. Sometimes what seems to be a synthesis of ideas is really one person dominating the group. Or perhaps the team is not as independent of Jim as they think. If Jim can pinpoint more specific problems with their collaborative skills, then it is more likely that he can arrange to have them trained in the relevant skills. In turn, that could lead to more tangible benefits from the collaborative team structure that he has developed than he is seeing now. He will not really know until he finds a way to measure the actual collaboration taking place.

FUTURE TRENDS

Online collaboration is likely to become even more important in the future for education, research, business, and other fields. In many of these areas, it will be important to ensure that groups working online are actually collaborating. Often in job settings working groups are composed of people with different knowledge and skill sets. If they do not truly collaborate, then their products (reports, designs, and other substantive materials) may be incomplete and substandard. In educational settings group work may produce high-quality learning outcomes primarily in groups that collaborate, rather than those using a less effective "divide-and-conquer" strategy in which group members learn only the pieces that they work on.

One indication of the trend toward more online collaboration is the increasing number of software packages aimed at improving the efficiency and effectiveness of online collaboration. For example, Groove *(www.groove.net)* is inexpensive peer-to-peer software designed specifically for small groups and containing a variety of tools beyond basic synchronous and asynchronous discussion capabilities. Convea *(www.convea.com)* is similarly inexpensive Web-based software designed for much the same purpose. Such packages are increasingly under scrutiny as collaborative tools (Ingram & Parker, 2003; Ingram, Pretti-Frontczak & Parker, 2003).

These trends mean that we need more and better ways to assess the presence, amount, and nature of online collaborative processes. Such measures will be useful in research, in education, and in practical application. Which measures are used depends on the questions being asked and on the purposes to which the assessment will be put. Some measures are extremely subjective and not firmly grounded in theory and previous research. Others may rely too heavily on the untutored impressions of participants. On the other hand, one drawback of a detailed quantitative approach such as the one we outlined here is the time it takes to complete. An approach being pursued by the present author and his colleagues is using the detailed research approach to develop better surveys of team members that will allow the identification of good and poor collaboration more quickly and efficiently. Such tools would have wide applicability.

CONCLUSION

This article has discussed several ways that collaboration in small online working groups might be measured. The need for such measurements seems clear: high-quality work in a variety of fields demands good collaboration. In turn, managers, researchers, educators, and others need to be able to identify the presences, amount, and nature of the collaboration that does or does not take place in different circumstances.

All of the measurement methods discussed here have strengths and drawbacks. Some are quick and easy to use but may give misleading results. Others are precise and based on some of the best research available, but demand significant effort and time to implement. One direction for future research and practice may be to develop measurement tools that are both accurate and quick.

REFERENCES

Dillenbourg, P., Baker, M., Blaye, A. & O'Malley, C. (1996). The evolution of research on collaborative learning. In P. Reinman & H. Spada (Eds.), *Learning in humans and machines: Towards an interdisciplinary learning science* (pp. 189-211). New York: Pergamon.

Hathorn, L.G. & Ingram, A.L. (2002). Cooperation and collaboration using computer-mediated communication. *Journal of Educational Computing Research*, *26*(3), 325-247.

Henri, F. (1992). Computer conferencing and content analysis. In A.R. Kaye (Ed.), *Collaborative learning through computer conferencing* (pp. 117-136). Berlin: Springer-Verlag.

Ingram, A.L. & Hathorn, L.G. (2004). Methods for analyzing collaboration in online communications. In T.S. Roberts (Ed.), *Online collaborative learning : Theory and practice* (Chapter 10, pp. 215-241). Hershey, PA: Idea Group, Inc.

Ingram, A.L. & Parker, R.E. (2003). Collaboration and technology for teaching and learning. *Proceedings of the Annual Conference of the Ohio Learning Network,* Columbus, OH.

Retrieved from *www.oln.org/conferences/papers/Collaboration_and_Technology.pdf*.

Ingram, A.L., Pretti-Frontczak, K. & Parker, R. (2003). Comparisons of student and faculty use of online collaboration tools. *Proceedings of the Teaching Online in Higher Education Online Conference*. Retrieved from *www.ipfw.edu/as/2003tohe/*.

Johnson, D.W., Johnson, R.T. & Smith, K.A. (1998). Cooperative learning returns to college. *Change*, *30*(4), 26-35.

Kaye, A. (1992). Learning together apart. In A.R. Kaye (Ed.), *Collaborative learning through computer conferencing* (pp. 117-136). Berlin: Springer-Verlag.

Laffey, J., Tupper, T., Musser, D. & Wedman, J. (1998). A computer-mediated support system for project-based learning. *Educational Technology Research and Development*, *46*(1), 73-86.

McInnerney, J.M. & Roberts, T.S. (2004). Collaborative or cooperative learning? In T.S. Roberts (Ed.), *Online collaborative learning : Theory and practice* (Chapter 9, pp. 203-214). Hershey, PA: Idea Group, Inc.

Rourke, L., Anderson, T., Garrison, D.R. & Archer, W. (2001). Methodological issues in the content analysis of computer conference transcripts. *International Journal of Artificial Intelligence in Education, 12*(1), 8-22. Retrieved from *www.atl.ualberta.ca/cmc/publications.html.*

Silverman, D. (1993). *Interpreting qualitative data.* Thousand Oaks, CA: Sage Publications.

Wilczenski, F.L., Bontrager, T., Ventrone, P. & Correra, M. (2001). Observing collaborative problem-solving processes and outcomes. *Psychology in the Schools*, *38*(3), 269-281.

KEY TERMS

Asynchronous Discussion: Online discussions that occur independent of time and space. Participants do not have to be online simultaneously, and can read and contribute to the conversation on their own schedules.

Collaboration: Occurs when small groups of people work together toward a common goal in ways that produce new products and knowledge that are unlikely to be developed by individuals. Three essential elements of collaboration are interdependence, synthesis, and independence.

Cooperation: Cooperative groups work together on group projects in ways that do not necessarily result in high-quality interaction, and new products and knowledge. A typical cooperative strategy is to divide up the work among the members and stitch the various contributions together at the end of the project.

Independence: The independence of a group from a central authority, such as an instructor or manager, ensures that the group can truly collaborate among themselves and produce results that are unique and new.

Interaction: Interaction among members of a group is necessary to produce collaboration. It can be measured by examining the give-and-take nature of the discussion threads.

Interdependence: Interdependence among members of a small group is a necessary element of collaboration. It means that group members could not produce the results they did without one another.

Participation: The most basic requirement of collaboration; it may be measured by the number of postings made and read or by the number of statements contributed to a discussion.

Synchronous Discussion: Occur when all participants are online and actively involved in the discussion at the same time.

Synthesis: Occurs when the final product of a group effort includes information and other elements from all members in such a way that individual contributions are difficult or impossible to identify.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1912-1916, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Metrics for the Evaluation of Test-Delivery Systems

Salvatore Valenti
Università Politecnica delle Marche-Ancona, Italy

INTRODUCTION

Most solutions to the problem of delivering course content supporting both student learning and assessment nowadays imply the use of computers, thanks to the continuous advances of information technology. According to Bull (1999), using computers to perform assessment is more contentious than using them to deliver content and to support student learning. In many papers, the terms computer-assisted assessment (CAA) and computer-based assessment (CBA) are often used interchangeably and somewhat inconsistently. The former refers to the use of computers in assessment. The term encompasses the uses of computers to deliver, mark, and analyze assignments or examinations. It also includes the collation and analysis of data gathered from optical mark readers. The latter (that will be used in this paper) addresses the use of computers for the entire process, including assessment delivery and feedback provision (Charman & Elmes, 1998).

A typical CBA system is composed of the following.

- Test-Management System (TMS) - that is, a tool providing the instructor with an easy-to-use interface, the ability to create questions and to assemble them into tests, and the possibility of grading the tests and making some statistical evaluations of the results
- Test-Delivery System (TDS) - that is, a tool for the delivery of tests to the students. The tool may be used to deliver tests using paper and pencil, or a stand-alone computer on a LAN (local area network) or over the Web. The TDS may be augmented with a Web enabler used to deliver the tests over the Internet. In many cases, producers distribute two different versions of the same TDS: one to deliver tests either on single computers or on a LAN and the other to deliver tests over the WWW (World Wide Web). This is the policy adopted, for instance, by Cogent Computing Co. (2004) with CQuest LAN and CQuest Net.

The TMS and TDS modules may be integrated in a single application as, for instance, Perception developed by Question Mark Computing (2004), or may be delivered as separate applications as it occurs for MicroTest and MicroGrade developed by Chariot Software Group (2004).

BACKGROUND

The interest in developing CBA tools has increased in recent years thanks to the potential market of their application. Many commercial products, as well as freeware and shareware tools, are the result of studies and research in this field made by companies and public institutions.

Thus, for instance, 42 quiz software products are referenced by the Soft411 (2004) directory, 23 by the Educational Software (2004) directory, and 8 by Assessment System Co. (2004). Moreover, it must be noted that almost all course management systems (Edutools, 2004) provide facilities for CBA. This noteworthy growth in the market raises the problem of identifying a set of criteria that may be useful to an educational team wishing to select the most appropriate tool for their assessment needs. The literature on guidelines to support the selection of CBA systems seems to be very poor since no other up-to-date papers are available on the Internet apart from the works by the author and his colleagues (Valenti, Cucchiarelli, & Panti, 2002a, 2002b).

The purpose of this paper is to provide a framework for the evaluation of a test-delivery system.

METRICS FOR THE EVALUATION OF A TDS

Three main functional modules roughly compose a TDS: a student interface, a question-management unit, and a test-delivery unit. Therefore, our framework for the evaluation of a TDS is defined in terms of criteria that may support the evaluation of each functional module and other criteria for the evaluation of the whole system, as shown in Table 1.

The evaluation of the interface is a qualifying aspect for the evaluation of a CBA system and obviously for a TDS. This becomes dramatically true if we take into account the fact that neither the teacher nor the students involved in the use of a TDS necessarily have a degree in computer science, nor may be interested in acquiring skills in this field. According to Nielsen and Molich (1990), the interface must be easy to learn, efficient to use, easy to remember, error free, and subjectively pleasing. Some further criteria that may be adopted to evaluate the usability of the interface are summarized in the following list.

Table 1. Metrics for the evaluation of a TDS

Issue		Metrics
Component Level	Interface	Friendly GUI (graphical user interface)
	Question Management	Types of Questions
		Question Structure (retries, tutorial building)
	Test Management	Help and Hints
		Restricted Availability
		Grading
System Level		Security
		Survivability
		Communication

- speak the users' language (multilinguality and multiculturality)
- be accessible
- provide feedback
- provide clearly marked exit points

The question-management unit of a TDS can be evaluated with respect to its ability to provide

- multiple attempts at solving a question (retries),
- feedback and tutorials on the topic covered by the questions, and
- capabilities for the inclusion of multimedia in questions.

The ability of providing retries may be of great importance for self-assessment since it is useful to improve the knowledge of the student whilst reducing the need for providing feedback and/or tutoring. On the other hand, the impossibility to change the answer to a question during an examination is often perceived as unfair by the students (Valenti et al., 2002b). It is worth outlining that allowing multiple attempts at question answering may affect the use of adaptive systems whenever item presentation depends on previous responses.

The feedback may be provided after each question (this solution being preferable for self-assessment), after a set of questions covering a given topic, or at the end of the test, and can be based on the overall performance. Furthermore, the feedback may be used to indicate the correctness of the answer, to correct misconceptions, or to deliver additional material for deepening and/or broadening the coverage of the topic assessed by the question. Tutorials represent an extended approach to provide additional information to the students. The existence of some facility to ease inclusion of tutorials in the TDS represents an important feedback aid. As an example, Perception provides explanation-type questions that may be used for "information screens, title pages, or to display large bodies of text" (Question Mark Computing Ltd., 2004).

The use of questions incorporating multimedia, such as sound and video clips or images, may improve the level of knowledge evaluation. This aspect may be of great importance, for example, in language assessment, where the comprehension of a talk or a movie can be assessed by recurring to multimedia only. The use of multimedia can raise issues related to portability and interoperability since it may require special hardware and software, both for the server delivering the questions and for the client used by the students. Furthermore, it may raise the costs for the adopted solution. These issues may not represent a problem whenever a Web-enabled TDS is selected since the nature of the World Wide Web is inherently multimedial. In this case, the choice of standard plug-ins for the most common browsers may reduce risks of portability and of interoperability. Since most plug-ins used to grant access to multimedia sources are usually free of charge, their use may not interfere with cost problems.

Among the issues taken into account to evaluate the test-management unit of a TDS, we have identified the ability to

- provide help and hints,
- make tests available at a given time, and
- allow scoring procedures.

The capability of a TDS to provide directions about the completion of the test and hints that usually are related to the contents of the questions represents a further measure of the ease of use of the application from the student's point of view.

Tests can be made either available or unavailable at a specified date and time. This allows test designers to specify exactly when people can access a test. It should be possible to leave out either or both restrictions to provide maximum flexibility. This lends itself nicely to the computer-lab set-

ting where students are required to complete an online test during a specified time frame on a specified day.

Obviously, any software for assessment should be able to compute student grades. Furthermore, grades must be delivered as feedback to the course coordinator, to the instructor, and to the students. Each of these categories of stakeholders may require a different kind of feedback on the grades associated with a test. For instance, a student needs to know where he or she stands with respect to other students and to the class average besides his or her own individual and cumulative grades. This need raises obvious concerns about privacy that may be faced through the security facilities provided with the assessment tool.

Among the issues taken into account to evaluate a TDS from a systemic point of view, we have selected

- security,
- survivability, and
- communication with other software.

There is a wide range of security issues related to the use of TDSs. In more detail, many concerns on the security of the availability of the test material, on the HTML (hypertext markup language) code that implements testing, and on the access-control privileges do exist. With respect to security concerns about the test material and its HTML code, it must be outlined that while commercial programs usually implement encrypting approaches, a lot of issues are raised by freeware. In fact, most freeware applications rely either on Perl/CGI (common gateway interface) or on JavaScript. In particular, since a CGI program contains an executable code, the use of CGI-based applications is the equivalent of letting the world run a program on the server side, which is not the safest thing to do. Therefore, there are some security precautions that need to be implemented when it comes to using CGI-based applications. The one that will probably affect the typical Web user is the fact that CGI programs need to reside in a special directory so that the server knows to execute the program rather than just display it to the browser. This directory is usually under direct control of the webmaster, prohibiting the average user from creating CGI programs. On the other hand, the assessment program cannot be completely hidden whenever using a JavaScript code that runs on the client side of the application, so a "smart" student can access the source, discovering the right answer associated to each question. Some sophisticated techniques can be used to partially overcome this problem (Cucchiarelli, Panti, & Valenti, 2000).

The ability of a TDS to perform under adverse conditions (i.e., survivability as discussed in Valenti et al., 2002a) is of great importance. In particular, no termination procedures should result in any loss of data. To ensure this, both student and system files should be updated after each transaction so that no data is lost if the test is terminated because of machine or power failure. The possibility of providing examination printouts may further enforce the survivability of the system. Furthermore, after a crash, the system should be able to restart from the point of termination with all aspects of the original status unchanged, including the answers already given and the clock still displaying the time remaining.

Communication with other existing software may be very useful both for exporting answers and for calling external applications. Furthermore, this feature is required to allow the integration of a TDS with a TMS distributed by different producers. Exporting answers is usually performed through test files and data-conversion utilities. This may be useful to customize the reports generated by the application or whenever an in-depth analysis is needed to evaluate the results obtained. Moreover, many available TDSs enable the calling of a program as a block within a question. The called program returns a score in points that may be added to the test score. This may be useful for assessing abilities that cannot be evaluated through the basic question-answer paradigm of most assessment tools.

Finally, some tools allow external applications to be called at the very end of the test phase for printing certificates for all users who pass the test, for the electronic submission of the answer file to a central location for analysis and evaluation, and for the storage of the results in a file to be accessed by a user program (Question Mark Computing Ltd., 2004).

FUTURE TRENDS

The software-evaluation process is a very complicated task because many, often contradictory, attributes have to be taken into account. Issues such as selection of appropriate attributes, creation of their hierarchy, assignment of relative weights to them, and application of a sound decision-aid methodology arise frequently both for nonexperienced and skilled users. Therefore, a first step toward providing support for the evaluation of test-delivery systems could be the construction of a Web site listing all the available software on the market, providing an independently reviewed, objective source of information to help educators in making the best decision for their institution. A further step in this direction could be the implementation of a decision support system to be used as an (semi) automated aid for helping the educator in selecting the most appropriate criteria and constructing the evaluation model.

CONCLUSION

In this article we have discussed a framework that may be useful in assisting an educational team in the selection of a TDS. Three main functional modules roughly compose a TDS: a student interface, a question-management unit,

and a test-delivery unit. Therefore, we decided to organize our framework by identifying some metrics to support the evaluation of the functional modules and other metrics to support the evaluation of the system as a whole.

REFERENCES

Assessment System Co. (2004). Retrieved from http://www.assess.com

Bull, J. (1999). Computer-assisted assessment: Impact on higher education institutions. *Educational Technology & Society, 2*(3).

Chariot Software Group. (2004). Retrieved from http://www.chariot.com/home/index.asp

Charman, D., & Elmes, A. (1998). Computer based assessment: A guide to good practice (Vol. I). SEED Publications, University of Plymouth.

Cucchiarelli, A., Panti, M., & Valenti, S. (2000). Web-based assessment of student learning. In A. K. Aggarwal (Ed.), *Web-based learning and teaching technologies: Opportunities and challenges* (pp. 175-197). Hershey, PA: Idea Group Publishing.

Educational Software. (2004). Retrieved from http://www.educational-software-directory.net

Edutools. (2004). Retrieved from http://www.edutools.info

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *Proceedings of CHI 90*, 249-256.

Question Mark Computing Ltd. (2004). *Perception.* Retrieved from http://www.questionmark.com/home.htm

Soft411. (2004). Retrieved from http://www.soft411.com

Valenti, S., Cucchiarelli, A., & Panti, M. (2002a). Computer based assessment systems evaluation via the ISO9126 quality model. *Journal of Information Technology Education , 1*(3), 157-175.

Valenti, S., Cucchiarelli, A., & Panti, M. (2002b). Relevant aspects for test delivery systems evaluation. In M. Khosrow-Pour (Ed.), *Web-based instructional learning* (pp. 203-216). Hershey, PA: IRM Press.

KEY TERMS

Computer-based Assessment: addresses the use of computers for the entire process of assessment including production, delivery, grading, and provision of feedback

CGI: CGI is the acronym for common gateway interface. A CGI program is any program designed to accept and return data that conforms to the CGI specification. CGI programs are the most common way for Web servers to interact dynamically with users. The program could be written in any programming language including C, Perl, Java, or Visual Basic.

HTML: HTML is the acronym for hypertext markup language, the authoring language used to create documents on the World Wide Web. HTML defines the structure and layout of a Web document by using a variety of tags and attributes. HTML is derived from SGML, although it is not a strict subset.

JavaScript: a scripting language developed to enable Web authors to add dynamic content to sites. Although it shares many of the features and structures of the Java language, it was developed independently. It is supported by recent browsers from Netscape and Microsoft.

Multiculturality: the term used to address the measures now being taken to provide graphical user interfaces with ad hoc icons and texts according to the cultural heritage of the user

Multilinguality: the term used to address the measures now being taken to provide graphical user interfaces with features for internationalization, that is, support of the character sets and encodings used to represent the information being manipulated, and presentation of the data meaningfully

Test-Delivery System: a tool for the delivery of tests to students. The tool may be used to deliver tests using a stand-alone computer on a local area network or over the World Wide Web.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1945-1948,

Micro and Macro Level Issues in Curriculum Development

Johanna Lammintakanen
University of Kuopio, Finland

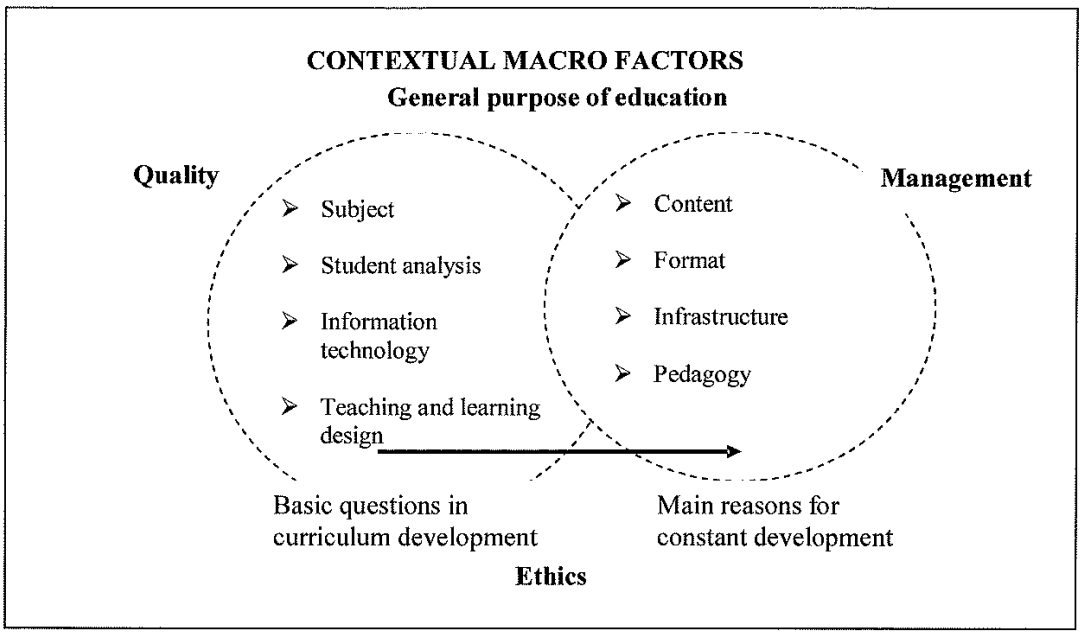
Sari Rissanen
University of Kuopio, Finland

INTRODUCTION

It is a well-known fact that an educational paradigm shift occurred in the course of the last decade, with a move from traditional to Web-based education at various educational levels (Harasim, 2000; Karuppan, 2001; Kilby, 2001). Web-based education (WBE) has advanced from the delivery of educational content to Web-based sites with interactive functions (Carty & Philip, 2001). Concurrently, new innovative kinds of pedagogical experiments have shifted the paradigm from teaching to learning (Pahl, 2003). However, there is a greater need for innovation in the area of pedagogy rather than that of technology (Littig, 2006). Indeed, educators have realized, as summarized by Armstrong (2001), that good Web-based educational theory and good educational theory are one and the same, the only difference being that WBE transcends the barriers of space and time.

The paradigmatic shift has occurred in both global education (including developing countries) and corporate training. The key impetus for this shift may vary in these areas, but the role of knowledge and intellectual capital, coupled with the needs of organizations and individuals to learn more rapidly, are apparent as the driving forces for WBE (e.g., Bell, Martin, & Clarke, 2004). The growth of WBE has been part of planned educational policy, but at the same time, good international or national experiences have also supported its growth. Furthermore, the cash crises in the western university sector (Bell et al., 2004) and the endeavors towards more coherent and cohesive educational systems and degrees, especially in the European context (Littig, 2006), can be identified as the other galvanizing factors for this shift.

Figure 1. Curriculum development as a continuous process



Aim and Structure of this Article

The aim of this article is to pursue the discussion of some essential micro- and macro-level issues in Web-based curriculum development, mainly at the level of higher education (see Figure 1). It was fairly often the case initially that the main concerns in curriculum development were related to students, the subject, new technology, and pedagogical issues. Curriculum development, however, must be seen as a *process* due to these issues, which are constantly evolving. Moreover, curriculum development does not happen in a vacuum—hence the two parts of the article. The first part focuses on the above-mentioned issues, while the second part presents a summary of the general purpose of education, ethics, quality, and management as important contextual concerns in WBE curriculum development.

Curriculum Development in Web-Based Education at the Micro Level

Web-based education, and curriculum development in particular, has taken a step forward in recent years. However, while a substantial body of research has concentrated on this new teaching medium, the results have been mixed and have shown no significant improvement in learning over traditional methods. In addition, the need for systematic and scientific knowledge remains, especially with regard to the effects and outcomes of WBE (Karuppan, 2001; Orr & Bantow, 2005).

Previous studies have shown that technology affects learning in many ways. Pedagogical choices, the design of the course Web site, and the interaction possibilities, for example, have different kinds of effects on learning outcomes (Romanov & Nevgi, 2006). One challenge is to integrate curriculum, technology, community, and learning in a manner that supports student motivation, self-regulation, and retention in WBE (Fisher & Baird, 2005). Unfortunately, technical matters and narrowly defined subject areas still receive the most emphasis from e-learning developers (Littig, 2006).

The identification of potential users and the analysis of their needs form the basis for curriculum development (Karuppan, 2001; Lammintakanen & Rissanen, 2003), but a stronger focus is needed more than ever before on learners and their needs (Littig, 2006). Moreover, Web architecture and learning materials should support the student's particular learning style in order to facilitate learning (Karuppan, 2001; Graff, 2006), while expectations concerning the technology dictate that it should be easy to access and easy to navigate at no extra cost to the learner.

Consideration of the expected level of learning is an important aspect in curriculum development. In other words, the course can be based on the assumption that learning is a process *for* acquiring information. However, the course can also be based on the assumption that learning is a process *of* acquiring information and processing experience, in which the learner selects and constructs useful and appropriate knowledge (Littig, 2006). Careful evaluation is needed during the curriculum development stage on whether or not the chosen technology supports teaching strategies that encourage active involvement and critical thinking, and fosters relationships between learners (Armstrong, 2001).

Curriculum development is time intensive and requires adequate financial and human resources in order to develop tightly organized courses. Web-based learning forces teachers to become course designers who make decisions based on their understanding of the probable needs, expectations, and behaviors of students on their own campuses. The role of tutor is a multi-faceted one, requiring organizational and management skills as well as the ability to motivate and encourage student interaction and facilitate learning and group processes (Packham, Jones, Thomas, & Miller, 2006).

Table 1. The factors of change promoting curriculum development (Pahl, 2003)

Content	The course subject evolves Changes in content to improve the material
Format	Changes in • Staff • Student body (qualifications, numbers, mode of learning) • Timetable (where and when the course takes place) • Syllabus (the content and organization of the course) • Curriculum (level, extent, prerequisites) • Legal and/or financial environment
Infrastructure	Improvements in hardware technology Systems and language technology face constant minor changes Learning devices are developing
Pedagogy	Knowledge acquisition, modeling of and access to educational knowledge Active learning in terms of engaging the student through interactive systems Collaborative learning supportive systems Autonomous learning Evolving instructional design

Previously, the challenges of curriculum development focused merely on the lack of both students' and staff members' skills, attitudinal problems, and suitable equipment for WBE. Maintenance aspects have been neglected in the design and development of new technologies. In fact, while in the previous stage, the main concerns were with regard to planning processes (i.e., how to begin with WBE); nowadays, important issues concern the how up to date the curriculum is and how to develop it further. In addition, online teachers' professional development has also become one of the topics of research (see Vrasidas & Zembylas, 2004).

Curriculum development can be based on students' evaluations, experiences of others, as well as previous studies. On the basis of previous research, Young and Norgard (2006), for example, summarized that the following areas are important for student satisfaction in an online course: interaction among students and the tutor, consistent course design across courses, the availability of technical support, and the flexibility of online courses compared to traditional learning and teaching. However, factors also exist that require us to develop the curriculum. Pahl (2003), for example, has provided a summary of both internal and external reasons why the curriculum needs to be constantly evolved (see Table 1). The evolvement of the design of a Web-based course can be affected by four dimensions; content, format, infrastructure, and pedagogy.

FUTURE TRENDS

The Macro Context of Curriculum Development

The previous section describes WBE curriculum development from the micro-level perspective. However, some existing and, concurrently, future challenges that have not yet been mentioned have become more important during the curriculum development process. These themes have been summarized into four interrelated categories: (1) the general purpose of education, (2) ethical and legal issues, (3) quality assurance and accreditation, and (4) managerial and organizational issues. These challenges became apparent in part from the literature and in part from practical experiences (e.g., Alexander, 2001; Kilby, 2001; Roffe, 2002), and concern both national and international curriculum development, since a more global perspective in design and courseware provision is expected (Kilby, 2001; e.g., MIT Open Courseware, in Potts, 2003; Bell et al., 2004).

The General Purpose of Education

The mass of information is growing rapidly and lifelong learning has been widely accepted at the societal level. However, critical analyses concerning how WBE fulfills the general purpose of education are rare. The theoretical literature of education shows that education has many other functions than the mere transfer of information. Bell et al. (2004) state that universities are not only "credentialing institutions or knowledge delivery mechanisms" (factors which remain the focus of major online and distance schools), but they also provide hugely beneficial learning communities in which students learn how to "be." This can mean, for instance, learning how to learn, or how to learn complex social strategies that cannot be learned in a virtual classroom. One of the crucial issues is how to combine the general purposes of education (e.g., sophisticated people acting in a civilized nation) and WBE education.

Ethical and Legal Issues

One interesting ethical issue concerns what happens to different cultures if education is globalized via Web-based learning. How, for example, do different countries maintain their identity, language, and culture in a globalized world? WBE solutions make it easily possible to assemble students from different socio-economical, political, and ethnical backgrounds. While this can bring added value for the learning and professional development (e.g., Akar, Öztürk, Tuncer, & Wiethoff, 2004), it may also cause some deep problems if the education is not carefully planned and supported by the tutors. Features of the WBE environment (e.g., effective communication, suitable instructions), for instance, may be interpreted and accepted more or less depending on the social and cultural background of the students (e.g., Lum, 2006). In addition, the faceless environment could, at its worst, facilitate student harassment if there is no supervision and proper norms and rules. In fact, from the ethical perspective, the students' privacy and confidentiality—and their respectful and dignified treatment on the Web-based environment—are imperative (Armstrong, 2001).

Web-based education provides good opportunities to make use of the many information sources available via the Internet (Jefferies & Hussain, 1998). The problem is, however, the quality of knowledge: how to cull information and how to avoid the use of misinformation (Calvert, 1998). At their worst, Web-based learning environments enhance fabrication, falsification, and plagiarism, all of which lead to copyright considerations.

Quality Assurance and Accreditation

Quality is a crucial concern in WBE: there are no common quality standards for course design, delivery, and evaluation, nor is there is an accreditation system. Some institutions and countries (such as the United Kingdom) have developed quality assurance protocols that demonstrate that the online programs are of equal quality to those delivered by traditional

methods (Roffe, 2002). As an example, the Massachusetts Institute of Technology (MIT) offers a standardized process for course modeling and encourages extending collaboration and interdisciplinary teaching. MIT Open Courseware is open and available all over the world (Potts, 2003). From the students' point of view, one important issue concerns how the courses can be accepted as part of the curriculum, how different educational institutes recognize the courses offered from other institutions nationally and internationally.

Managerial and Organizational Issues

The tradition of individualism in teaching remains part of the organizational culture at different educational institutions. At best, the organizational culture can support both Web-based curriculum development and joint teaching. The allocation of both human and technical resources and a clear strategic decision from managers are requirements for faculty development and the incorporation of WBE into the curriculum (Carty & Philip, 2001). It is essential that the organization and its managers have a positive attitude to WBE, as well as time and resources, and that they promote its implementation (e.g., Alexander, 2001; Lum, 2006). The promotion of WBE may threaten the status quo of the organization. Enuku and Ojugwu (2006), for instance, have noticed that the fear of change may be due to change in the routine at work, or fear of being left behind or replaced by others who have the relevant pedagogical and technological skills. These and other fears held by staff members and managers emphasize the management of change as a crucial factor in the promotion of WBE.

The managerial approach to guiding teaching and curriculum development work is, however, not yet very visible in various educational organizations. The study by McPherson and Nunes (2006) summarizes the critical success factors for WBE implementation from the leadership, structural, and cultural perspectives in higher education. Familiarity with the organizational culture, structure, corresponding, and potentially conflicting strategies is held as important before rushing into the design, development, and implementation of WBE solutions. In addition, human issues in terms of participation, information, training, and communication, as well as stakeholder involvement, are fundamental when engaging in WBE initiatives.

CONCLUSION

In conclusion, there is no consistent paradigm for WBE, rather there are multiple ways of making use of the Web in education, and these will vary depending on the subject being taught and the needs of the learner. Curriculum development requires a great deal of competence and effort from different stakeholders. Furthermore, both micro- and macro-level issues must be taken into account in Web-based education. Therefore, motivation and commitment to long-term WBE development strategies are needed at the personal, organizational, national, and global levels. In addition, Web-based education appears to support cultural cohesion and more rapid transfer of information. The attention of current research has tended to focus mainly on micro-level issues, while the more general aspects and impacts of Web-based education remain less studied. However, both technological and pedagogical competencies of different actors, as well as the body of research, are constantly evolving, and therefore, more promising outcomes of Web-based education will be achieved.

REFERENCES

Akar, E., Öztürk, E., Tuncer, B., & Wiethoff, M. (2004). Evaluation of a collaborative virtual learning environment. *Education and Training, 46*(6/7), 343-352.

Alexander, S. (2001). E-learning developments and experiences. *Education and Training, 43*(4/5), 240-248.

Armstrong, M.L. (2001). Distance education: Using technology to learn. In V. Saba & K.A. McCormick (Eds.), *Essentials of computers for nurses — information for the new millennium* (3rd ed., pp. 413-426). New York: McGraw-Hill.

Bell, M., Martin, G., & Clarke T. (2004). Engaging in the future of e-learning: A scenarios-based approach. *Education and Training, 46*(6/7), 296-307.

Calvert, P.J. (1999). Web-based misinformation in the context of higher education. *Asian Libraries, 8*(3), 83-91.

Carty, B., & Philip, E. (2001) The nursing curriculum in the information age. In V. Saba & K. A. McCormick (Eds.), *Essentials of computers for nurses — information for the new millennium* (3rd ed., pp. 393-412). New York: McGraw-Hill.

Enuku, U.E., & Ojugwu, C.N. (2006). Information and communication technology (ICT) in the service of the National Open University of Nigeria. *Education, 127*(2), 187-195.

Fisher, M., & Baird, D.E. (2005). Online learning design that fosters student support, self-regulation and retention. *Campus-Wide Information systems, 22*(2), 88-107.

Graff, M. (2006). Constructing and maintaining an effective hypertext-based learning environment. Web-based learning and cognitive style. *Education and Training, 48*(2/3), 143-155.

Harasim, L. (2000). Shift happens. Online education as a new paradigm in learning. *Internet and Higher Education, 3*(1-2), 41-61.

Jefferies, P., & Hussain, F. (1998). Using the Internet as a teaching resource. *Education and Training, 40*(8), 359-365.

Karuppan, C.M. (2001). Web-based teaching materials: A user's profile. *Internet Research: Electronic Networking Applications and Policy, 11*(2), 138-148.

Kilby, T. (2001). The direction of Web-based training: A practitioner's view. *The Learning Organization, 8*(5), 194-199.

Lammintakanen, J., & Rissanen, S. (2003). An evaluation of Web-based education at a Finnish university. In A. Aggarwal (Ed.), *Web-based education. Learning from experience* (pp. 440-453). Hershey, PA: Information Science.

Littig, P. (2006). New media-supported learning today and tomorrow: Recommendations for the next generation of education and training concepts supported by new learning media. *Industrial and Commercial Training, 38*(2), 86-92.

Lum, L. (2006). Internationally educated health professionals: A distance education multiple cultures model. *Education and Training, 48*(2/3), 112-126.

McPherson, M., & Nunes, M.B. (2006). Organizational issues for e-learning. Critical success factors as identified by HE Practitioners. *International Journal of Educational Management, 20*(7), 542-558.

Orr, S., & Bantow, R. (2005). E-commerce and graduate education. Is educational quality taking a nose dive? *International Journal of Educational Management, 19*(7), 579-586.

Packham, G., Jones, P., Thomas, B., & Miller, C. (2006). Student and tutor perspectives of online moderation. *Education and Training, 48*(4), 241-251.

Pahl, C. (2003). Managing evolution and change in Web-based teaching and learning environments. *Computers and Education, 40*(2), 99-114.

Potts, J.P. (2003, November 5). A new model for open sharing. *Proceedings of the WCET Annual Conference.* Retrieved March 31, 2004, from *http://ocw.mit.edu/OcwWeb/index.htm*

Roffe, I. (2002). E-learning: Engagement, enhancement and execution. *Quality Assurance in Education, 10*(1), 40-50.

Romanov, K., & Nevgi, A. (2006). Learning outcomes in medical informatics: Comparison of a WebCT course with ordinary Web site learning material. *Information Journal of Medical Informatics, 75,* 156-162.

Vrasidas, C., & Zembylas, M. (2004). Online professional development: Lessons from field. *Education and Training, 46*(6/7), 326-334.

Young, A., & Norgard, C. (2006). Assessing the quality of online courses from the students' perspective. *Internet and Higher Education, 9*(2), 107-115.

KEY TERMS

Computer and Information Literacy: The abilities to perform computer operations at a skill level high enough to meet the demands of the society, and to use the tool of automation in the process of accessing, evaluating, and utilizing information (Carty & Philip, 2001).

Cultural Differences: Not limited simply to differences in ethnicity or nationality; refers to patterns of thought, attitudes, and behaviors that vary according to the level of sameness shared by distinct groups (Lum, 2006).

Learning Style: The way in which individuals acquire and use information, strategies to process information in learning, and problem-solving situations (Karuppan, 2001).

Learning Tools: Tools included in Web-based learning environments for managing the course and geared to facilitating student learning in the environment.

Web-Based Education (WBE): Differs from traditional classroom teaching in two essential elements — physical distance and time — allowing the learner more flexibility. The most basic form of WBE is to deliver syllabi, lecture notes, reading materials, and assignments via the Internet. The more advanced level includes computer conference facilities, a help desk, linkage of conferencing and Web page assignment, testing and course management tools, and evaluation (Karuppan, 2001).

Web-Based Learning Environment: A specially developed program using Internet technology for the design and development of teaching and learning purposes. Trademarks include, for example, WebCT, WebBoard, Top Class, and Virtual-U.

Web-Based Misinformation: Describes information found on the Internet that does not fit normative patterns of "truth" — that is, it is incomplete, out of date, confused, or offers low consensus "knowledge" (Calvert, 1998).

Migration of Legacy Information Systems

Teta Stamati
National and Kapodistrian University of Athens, Greece

Panagiotis Kanellis
National and Kapodistrian University of Athens, Greece

Konstantina Stamati
National and Kapodistrian University of Athens, Greece

Drakoulis Martakos
National and Kapodistrian University of Athens, Greece

INTRODUCTION

In recent years, the accelerated competition in the global marketplace rendered the corporate environment more volatile than ever. The businesses are heavily relying on technological advancements to deliver a vast array of initiatives across a variety of industries. The firms' main partner in this increasingly complex and unpredictable journey is considered to be their information systems. Although the relevant industry offers an unprecedented rate of technological innovations, nevertheless there are cases where the information systems carry significant baggage from the past (Kelly, Gibson, Holland, & Light, 1999). There are aged systems that often form the central hub of the information flow within the organisation and are responsible for consolidating information about the business (Bisbal, Lawless, Wu, & Grimson, 1999; Sommerville, 2001) and thus they are called *mission-critical legacy information systems*.

The term "Legacy", according to the Oxford Dictionary, refers to any long-lasting effect of an event or process. The Legacy System describes an old system that remains in operation within an organisation. These systems often represent a massive, long-term business investment. Ulrich (1994) defined them as "stand-alone applications built during a prior era's technology, but they are perhaps more widely understood as software systems whose plans and documentation are either poor or non-existent" (Connall & Burns, 1993). Bennett (1995) referred to the legacy systems as, "large software systems that we do not know how to cope with but that are vital to the organisation", while Brodie and Stonebraker (1995) as "any information system that significantly resists modification and evolution to meet new and constantly changing business requirements". Finally, O'Callaghan (1999), drawing on the characteristics of legacy systems, described them as "a large system delivering significant business value today from a substantial pre-investment in hardware and software that may be many years old. Characteristically, it will have a long maintenance tail. It is, therefore, by definition a successful system and is likely to be one that is, in its own terms, well engineered. It is a business critical system which has an architecture which makes it insufficiently flexible to meet the challenges of anticipated future change requirements."

Legacy systems as a subject area is often overlooked in favour of areas such as new technology developments and strategic planning of information technology. In this context, the following sections present an overview of the legacy information systems problems in terms of their scale and definition. The legacy system issues include the required man-effort and costs of maintaining and evolving existing systems and the current methods of migrating complex legacy systems to new technology. It is shown that legacy systems present a critical area of study in both software engineering and business information systems. Taking into account that the role of technology is not merely supportive but affects the way enterprises conduct their business, it is shown that it is outdated to consider the migration process as the simple replacement of aged or problematic hardware and software. Thus, the migration should be approached as a planned change process that first and foremost requires an understanding and a methodology that covers the range of issues and organisational entities involved.

BACKGROUND

Legacy Information Systems

O'Callaghan (1999) refers to the adoption of an informational culture within the organisations in which "point solutions" were developed due to the widespread use of computer technology over several decades. There are cases where different divisions of the same organisation have developed individual applications in order to meet their perceived needs

in an application-by-application basis (O'Callaghan, 1999). In a similar way, there are applications in the same company that are running on different operating systems. Subsequently, such "point solutions", according to O'Callaghan (1999), became subject to localised optimisation, and uncontrolled maintenance, exacerbating the position further (Zou & Kontogiannis, 2002). These applications are unambiguously hard to maintain, improve, and expand because there is a general lack in their understanding. In addition, integration with newer systems may also be difficult because new business software may use completely different technologies (Wu, Lawless, Bisbal, Grimson, Wade, O'Sullivan, & Richardson, 1997). Due to the aforementioned reasons, there is a significant number of software engineers and practicing managers that consider the legacy systems to be potentially problematic (Bisbal et al., 1999).

On the other hand, according to Brodie and Stonebraker (1995), legacy systems do not always fit this stereotype. They propose that if a system was recently developed but cannot be readily modified to adapt to the constantly changing business requirements, then such a system can be regarded as a legacy system. Similarly, Randall (1999) stresses that "Legacy" is not just a problem encountered by organisations with aging mainframes and dated software, it is an issue from the moment a computer system becomes an integral part of any organisation's work (Randall, 1999).

The common rule is that if the legacy systems cannot support the business requirements, the business will not be able to remain competitive for long (Brodie & Stonebraker, 1995). Both a significant budget and person-hours will be monopolised by legacy systems maintenance. De Palma and Woodring (1993) referred to over 40% of the IT costs within an organisation being spent on maintaining its legacy systems, while Brodie and Stonebraker (1995) considered that the process of keeping these systems running takes 80-90% of the IT budget. Slee and Slovin (1997) gave an estimation in the area of 80% just for the routine maintenance activities.

Migration of Legacy Information Systems

The common sense solution to the legacy problem is migration. The definition of a successful information system migration according to Brodie and Stonebraker (1995) is as follows: "it begins with a mission-critical legacy system of a significant size in full operation and it ends with a fully operational, mission critical target application (or applications components) that replaces the essential aspects of the original legacy system." This involves replacing the problematic hardware and software, including the interfaces, applications, and databases that compose an information system infrastructure. Brodie and Stonebraker (1995) claim that legacy information system migration involves starting with a legacy information system and ending with a comparable target information system. This target system is significantly different from the original, but it contains substantial functionality and data from the legacy system.

The target system must be built using technological advancements in place of the legacy technology. For practical reasons, the target system may contain legacy components for which there is no adequate justification for their migration. According to Bisbal et al. (1999) the essence of legacy system migration is to allow the organisations to move their legacy systems to new environments, retaining the functionality of existing information systems without having to completely redevelop them.

In recent years, a significant number of organisations have initiated large-scale migration projects in order to improve their operations performance and to be compatible with the latest technological advancements. Particularly, the introduction of the Web browser technology, the so-called first-wave of the Internet (Dreyfus, 1998), forced the organisations to undertake migration projects in order to exploit the benefits of the shared information resources (Zou & Kontogiannis, 2002). Afterwards, the convergence of the Web and the distributed-object technologies extended the information Web-based applications to the services-based worldwide applications which was referred to as the Internet's second-wave (Dreyfus, 1998), and where the provided services and the content were distributed over the Internet (Zou & Kontogiannis, 2002). Moreover, the object-oriented technologies provided some valuable tools for the realisation of the services-based Web due to their inherent properties of encapsulation, polymorphism, and specialisation. In addition to the object orientation as a design paradigm, n-tier object computing was gradually being adopted by organisations as the preferred architecture for distributed applications because it allowed for the clear separation of business logic, representation logic, and back-end services (Zou & Kontogiannis, 2002).

Considering the case where the stakeholders of a large organisation (for instance, in the banking sector) have decided to maintain organisation's competitive edge and achieve conformation to the requirements posed by the need for flexibility and the minimization of time to market, the migration of company's legacy systems towards a new operating Web-based environment will be a considerably effective system evolution strategy. The strategic objectives will focus on leveraging the existing legacy software assets while minimising the risks involved in implementing from scratch its large scale mission-critical legacy applications (Umar, 1997). Thus, following the new era of distributed component-based applications, the organisation will face pressures to evolve its existing system (for instance, a large mainframe) in response to its customer expectations. The transformation from the mainframe computer to a multi-tier architecture will force the migration engineers to separate the integration logic and the legacy services to be stored in the middle-tier and the back-end tier, respectively. This architec-

ture will enable lightweight and thin clients to interact with servers in a fully customisable way. The component-based development will be based on the concepts of modularity, structured design, and object-orientation. The migration engineers and the practicing managers will emphasise on the design of the information systems in terms of well-defined modules that will be accessible to other modules through well-defined interfaces.

Current Migration Approaches

Each migration project could address the areas of reverse engineering, business reengineering, data transformation application development, human computer interaction, and testing (Ulrich, 2002). A generic process of legacy systems migration may include distinct phases (Bisbal et al., 1999) such as the Justification of the Migration process; the Understanding of the Legacy System; the Development of the Target System; the Testing and the Cut-Over (actual Migration). Within each task, general software and system engineering techniques can be applied.

Regarding the crucial phase of the actual migration process, various software engineers propose different methodological approaches.

Brodie and Stonebraker (1995) describe two strategies for migration: Cold Turkey and Chicken Little. The main drawback of the Cold Turkey strategy is the probable failure because the system will be written essentially from scratch. In that case, the risk level is high due to the fact that the development and evolution of the new system usually last for many years, in which the factors that affect the system may change. Chicken Little is a better strategy, according to Brodie and Stonebraker (1995), as it separates the migration process into step, and faces each one with a different aspect. Gateways play a significant role in the Chicken Little process.

The Butterfly methodology is being developed as part of the MILESTONE project (Wu et al., 1997). During the migration, the methodology eliminates the need for system users to simultaneously access both the legacy and target systems and, therefore, to keep consistency between these two heterogeneous information systems. The Butterfly methodology is based on the assumption that the data of the legacy system is logically the most important part of the system and that, from the viewpoint of the target system development, it is not the ever-changing legacy data that is crucial, but rather its semantics or schema(s). Thus, the Butterfly methodology separates the target system development and data migration phases, thereby eliminating the need for gateways.

BUSINESS "LEGACY" AND MIGRATION AS A BUSINESS CHANGE PROCESS

According to Laudon and Laudon (1998), today's business environment has been altered by the three powerful worldwide changes, namely, the emergence of globalisation; the transformation of industrial societies and economies into knowledge and information-based economies; and the transformation of the business enterprise whereby organisations are moving away from a hierarchical, centralised structure to become decentralised. The presence of business legacy within organisations can be identified by examining how they operate in their environment (Kelly et al., 1999).

Figure 1. Major activities in legacy system migration (Bisbal et al., 1999)

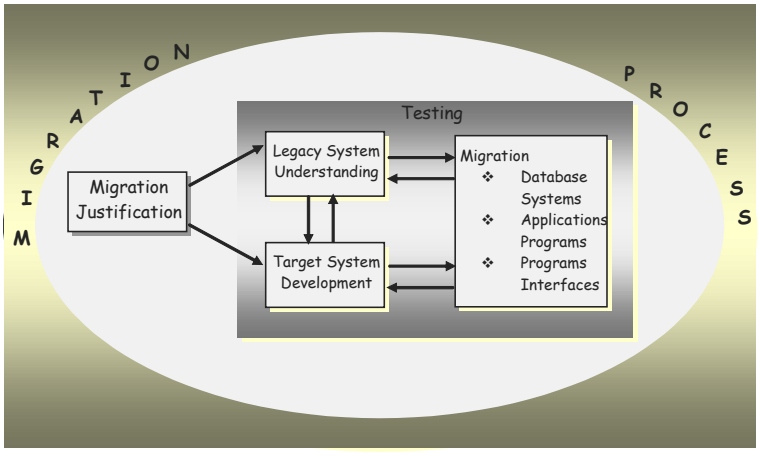


Figure 2. Chicken Little 11-step[1] strategy (Brodie & Stonebraker, 1995)

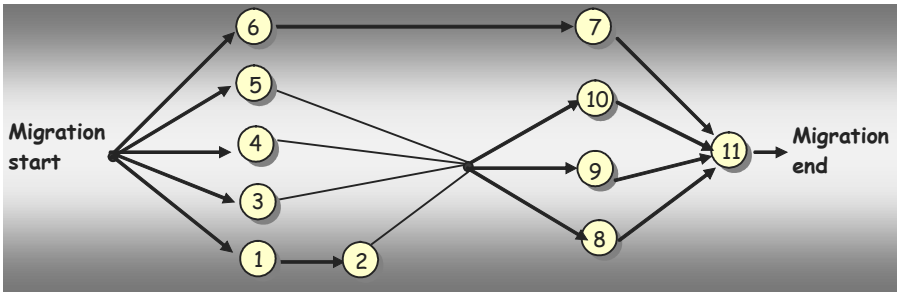
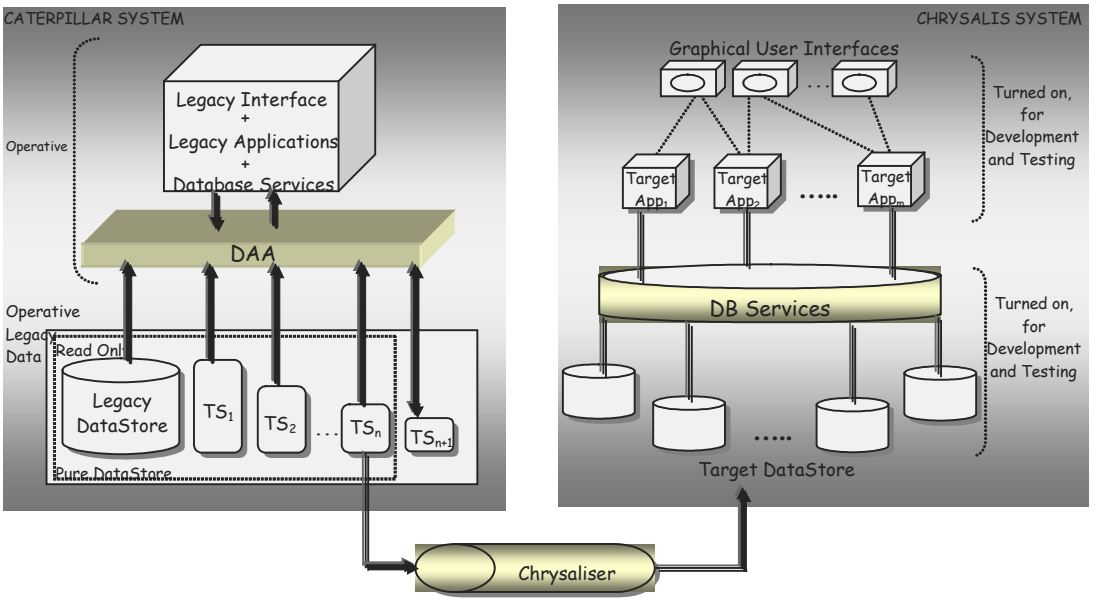


Figure 3. The Butterfly methodology (Wu et al., 1997)



Kelly et al. (1999) refer to the typical components of "business legacy" as the way firms perceive their business, their organisational structure, and the way that they perceive the market within which they operate. Moreover, the business objectives and the organisational strategy can be part of this "business legacy" as well as the way work is organised such as workflows and business processes (Kelly et al., 1999). In this context, the "business legacy" is embedded in the legacy information systems, and it is the inter-relatedness of business and information systems legacy that makes either business or systems change a very complex process.

In this context, Stamati, Kanellis, Stamati, and Martakos (2004) consider the fact that although the legacy migration is being viewed as the replacement of aged or problematic hardware and software, including applications, interfaces, and databases that compose an information system infrastructure, it does not take into account the role of technology today which is not merely supportive but pervades every aspect of the way enterprises conduct their business. Their position is that migration must be implemented as a planned change process that first and foremost requires an understanding of the range of issues and organisational entities involved (Stamati et al., 2004).

In this case, the main underlined hypothesis is that migration is a process which entails business change, and it is more than just the movement or reorganisation of database systems, application programs, and program interfaces. The consideration of these physical systems is only the informational view of migration. However, from a business perspective, migration must also account for the broader impact to

business change which occurs from the organisational and operational viewpoints.

The motivation for any migration process is the transition from an initial organisation situation A, which is unsatisfactory in some aspect, to a desired situation B where the problem is treated. Possible causes to such change include perceived opportunities, threats, or strategic decisions, including for example the opportunities offered by new technologies, the treatment of inefficient business processes, the increased customer demands, the globalisation of markets, or the decision to move to a different operating platform (Kavakli & Loucopoulos, 2004).

FUTURE TRENDS

Recent methodologies consider business requirements as "the process of transforming the need for organisational change into a requirements specification which can then serve as a framework for making the necessary change into the organisational domain" (Pohl, 1996). Considering the aforementioned hypothesis for a migration process, it should be stressed that the actual business structures and practices should be considered as a domain of potential change and new design. Thus, business requirements should be considered as a central part of any migration activity within a business environment (Kavakli & Loucopoulos, 2004). Migration requires cooperation and understanding between the business management and technology management. Once business management has determined its goals, technology management must seek ways to modify the technology to support these goals.

Migration, therefore, should not be an undirected process but a purposive activity driven by organisational change goals. Its effectiveness should depend on being able to make good decisions about what migration goals to pursue, on selecting the appropriate strategies for achieving the desired goals, and on guiding the application of the chosen strategies.

CONCLUSION

The concept of migration has shifted in order to include both organisational change as well as change in computer systems that enable such enterprise change. Such a movement has led to the emergence of new definitions that put emphasis on a broader process of migration, and include the cognitive, social, and technical context of migration. These definitions are based on the premise that the replacement of a software system in the organisation inevitably brings change in the way work is organised. In a similar manner, any organisational change should be reflected in the software system requirements. Therefore, migration is concerned both with the design and implementation of a new software system and the management of change in the business systems that might be supported by it.

REFERENCES

Bateman, A., & Murphy, J. (1994). *Migrating of legacy systems.* Working Paper CA-2894, School of Computer Applications, Dublin City University.

Bennett, K. H. (1995). Legacy systems: Coping with success. *IEEE Software, 12*(1), 19-23.

Bisbal, J., Lawless, D., Wu, B., & Grimson, J. (1999). Legacy information system migration: A brief review of problems, solutions and research issues. *IEEE Software, 16,* 103-111.

Brodie, M. L., & Stonebraker, M. (1995). *Migrating legacy systems.* Morgan Kaufmann Publishers.

Connall, D., & Burns, D. (1993). Reverse engineering: Getting a grip on legacy systems. *Data Management Review.*

De Palma, D. A., & Woodring, S. D., (1993). Breaking legacy gridlock. *Software Strategies, 4*(9), 1-16.

Dreyfus, D. (1998). The second wave: Netscape on usability in the services-based internet. *IEEE Internet Computing, 2*(2).

Ganti, N., & Brayman, W. (1995). *Transition of legacy systems to a distributed architecture.* John Wiley & Sons.

Holland, C. P., & Light, B. (1999). Focus issue on legacy information systems and business process change: Introduction. *Communications of AIS, 2*(9), 98-120.

Kavakli, E., & Loucopoulos, P. (2004). Goal modeling in requirements engineering: Analysis and critique of current methods. In J. Krogstie, T. Halpin, & K. Siau (Eds.), *Information modelling methods and methodologies* (pp. 102-124). Hershey, PA: Idea Group Publishing.

Kelly, S., Gibson, N., Holland, C. P., & Light, B. (1999). A business perspective of legacy information systems. *Communications of the Association for the Information Systems, 2*(7).

Laudon, K. C., & Laudon, J. P., (1998). *Management information systems, new approaches to organisation and technology* (5th ed.). New York: Prentice Hall.

Light, B., Holland, C., & Gibson, N. (1998). The influence of legacy information systems on business process change strategies. *Proceedings of the 4th American Conference on Information Systems Association for Information Systems, 2*(1), 527-529.

O' Challaghan, A. J. (1999). Migrating large-scale legacy systems to component-based and object technology: The Evolution of a pattern language. *Communications of the Association for Information Systems, 2*(3), 104-121.

Pohl, K. (1996). *Process-centered requirements engineering.* Research Studies Press Ltd.

Randall, D. (1999). Banking on the old technology: Understanding the organisational context of the "legacy" issues. *Communications of the Association for Information Systems, 2*(8), 208-221.

Slee, C., & Slovin, M. (1997). Legacy asset management. *Information Systems Management, 14*(1), 12-21.

Sommerville, I. (2001). *Software engineering* (6th ed.). Pearson Education.

Stamati, T., Kanellis, P., Stamati, K., & Martakos, D. (2004) Legacy migration as planned organizational change. *Sixth International Conference on Enterprise Information Systems (ICEIS), 3,* 501-508.

Ulrich, W. (1994). From legacy systems to strategic architectures. *Software Engineering Strategies, 2*(1), 18-30.

Ulrich, W. M. (2002). *Legacy systems: Transformation strategies.* Prentice Hall PTR.

Umar, A. (1997). *Application (re)engineering: Building Web-based applications and dealing with legacies.* Englewood Cliffs, NJ: Prentice Hall.

Wu, B., Lawless, D., Bisbal, J., Grimson, J., Wade, V., O'Sullivan, D. ,& Richardson, R. (1997). Legacy system migration. *Proceedings of the 17th International Database Conference* (pp. 129-138).

Zou, Y., & Kontogiannis, K. (2002). Migration to object oriented platforms: A state transformation approach. *19th IEEE International Conference on Software Maintenance (ICSM)* (pp. 530-539).

KEY TERMS

Component Architecture: A notion in object oriented programming where components of a program are completely generic. Instead of having a specialised set of methods and fields they have generic methods through which the component can advertise the functionality it supports to the system into which it is loaded.

Distributed Programming: The kind of programming that supports objects distributed across a network.

Gateway: A software module that is placed between other software modules. One of its roles for instance, is to simulate the old information system, while it is migrated, so it is still visible to the old user interfaces and application modules. Another role may be the transformation of the target information system in order for it to be visible to the old user interfaces, and the old legacy information system to be visible to the new user interfaces.

Mainframe: A machine designed for batch rather than interactive use, though possibly with an interactive time-sharing operating system retrofitted onto it.

N-Tier (or Multi-Tier) Architecture: This means splitting a system into more than just a client layer and a database layer. The server in this case refers to a custom written thing. The server then takes care of the business logic, and gets and returns the raw data to one or more database servers.

Object-Oriented Programming: The use of a class of programming languages and techniques based on the concept of an object which is a data structure encapsulated with a set of routines, called methods, which operate on the data. Operations on the data can only be performed via the methods, which are common to all objects that are instances of a particular class.

Reengineering: The examination and modification of a system to reconstitute it in a new form and the subsequent implementation of the new form.

Web Services Definition Language: An XML format for describing network services as a set of endpoints operating on messages containing either document oriented or procedure-oriented information. The operations and messages are described abstractly and then bound to a concrete network protocol and message format to define an endpoint. Related concrete endpoints are combined into abstract endpoints (services).

ENDNOTE

[1] S1: analyse the legacy information system, S2: decompose the legacy information system structure, S3: design the target interfaces, S4: design the target applications, S5: design the target database, S6: install the target environment, S7: create and install the necessary gateways, S8: migrate the legacy database, S9: migrate the legacy applications, S10: migrate the legacy interfaces, S11: cut over to the target information system.

Mobile Ad Hoc Network Security Vulnerabilities

Animesh K. Trivedi
Indian Institute of Information Technology, India

Rajan Arora
Indian Institute of Information Technology, India

Rishi Kapoor
Indian Institute of Information Technology, India

Sudip Sanyal
Indian Institute of Information Technology, India

Ajith Abraham
Norwegian University of Science and Technology, Norway

Sugata Sanyal
Tata Institute of Fundamental Research, India

INTRODUCTION

Mobile ad hoc networks inherently have very different properties from conventional networks. A mobile ad hoc network (MANET) is a collection of mobile nodes that are self configuring (network can be run solely by the operation of the end-users), capable of communicating with each other, establishing and maintaining connections as needed. Nodes in MANET are both routers and terminals. These networks are dynamic in the sense that each node is free to join and leave the network in a nondeterministic way. These networks do not have a clearly defined physical boundary, and therefore, have no specific entry or exit point. Although MANET is a very promising technology, challenges are slowing its development and deployment. Nodes in ad hoc networks are in general limited in battery power, CPU and capacity. Hence, the transmission ranges of these devices are also limited and nodes have to rely on the neighboring nodes in the network to route the packet to its destination node. Ad hoc networks are sometimes referred to as multi-hop networks, where a hop is a direct link between two nodes.

MANET has many important applications, including battlefield operations, emergency rescues, mobile conferencing, home and community networking, sensor dust and so forth.

Due to limited memory and computational power, nodes in MANETs have limited services and security provision. Unlike wired networks which have a higher level of security for gateways and routers, ad hoc networks have characteristics such as dynamically changing topology, weak physical protection of nodes, no established infrastructure or centralized administration and high dependence on inherent node cooperation. The routing protocols used in the current generation of mobile ad hoc networks, like Dynamic Source Routing (DSR), and Ad hoc On Demand Distance Vector Routing Protocol (AODV), are based on the principle that all nodes will cooperate, but dynamic and cooperative nature of MANETS presents substantial challenges to this assumption (Johnson, Maltz, & Broch, 2001; Perkins & Royer, 1999). Without node cooperation in a mobile ad hoc network, routes cannot be established, and packets cannot be forwarded. As a consequence, access control mechanisms, (similar to firewalls in wired networks) are not feasible. However, cooperative behavior, such as forwarding other node's messages, cannot be taken for granted because any node could misbehave. Misbehavior means deviation from regular routing and forwarding protocol assumption. It may arise for several reasons, non-intentionally when a node is faulty or intentionally when a node may want to save its resources. Cooperation in mobile ad hoc networks is a big issue of consideration. To save battery, bandwidth, and processing power, nodes should not forward packets for others. If this dominant strategy is adopted, the outcome is a nonfunctional network when multi-hop routes are needed, so all nodes are worse off. Without any counter policy, the effects of misbehavior have been shown to dramatically decrease network performance. Depending on the proportion of misbehaving nodes and their strategies, network throughput could decrease, and there could be packet losses, denial of

service or network portioning. These detrimental effects of misbehavior can endanger the entire network.

Wireless ad hoc networks are vulnerable to various attacks. These include passive eavesdropping, active interfering, impersonation, modification of packets and denial-of-service. Intrusion prevention measures, such as strong authentication and redundant transmission, can be used to tackle some of these attacks. However, these techniques can address only a subset of the threats, and moreover, are costly to implement due to the limited memory and computation power on nodes. We can identify two types of uncooperative nodes: faulty or malicious and selfish. Faulty or malicious behavior refers to the broad class of misbehavior in which nodes are either faulty and can therefore not follow a protocol, or are intentionally malicious and try to attack the system. Selfishness refers to no cooperation in certain network operations. In mobile ad hoc networks, the main threat from selfish nodes is dropping of packets (black hole), which may affect the performance of the network severely. Faulty, malicious and selfish nodes are misbehaved nodes.

ROUTING IN MANETs

Dynamic Source Routing is a popular routing protocol for ad hoc networks and was proposed for MANET by Johnson, Maltz and Broch (2001). In DSR, nodes do not store route to different nodes but they are discovered as they are needed. This type of routing is called R*eactive* routing and protocols used in this are called R*eactive* P*rotocols* (e.g., DSR, AODV, etc.). DSR works as follows: Nodes send out a ROUTE REQUEST (RREQ) message, all nodes that receive this message put themselves into the source route and forward it to their neighbors, unless they have received the same request before. If a receiving node is the destination, or has a route to the destination, it does not forward the request, but sends a REPLY (RREP) message containing the full source route. It may send that reply along the source route in reverse order or issue a ROUTE REQUEST including the route to get back to the source, if the former is not possible due to asymmetric links. After receiving one or several routes, the source selects the best (by default the shortest) route, stores it, and sends messages along that path. The better the route metrics (number of hops, delay, bandwidth, or other criteria) and the sooner the REPLY arrives at the source, the higher the preference given to the route and the longer it will stay in the cache. Because route to the destination is put into the packet, it is called source routing.

Attacks on DSR

There are a number of attacks possible on DSR protocol because there is no security measure and it assumes honest coordination of nodes among them and to protocol. A few attacks are outlined in this section and others are discussed in detail in the cited references.

- Dropping of packets by a node takes into account the following scenarios-Drop all packets not destined to it or perform only partial dropping. Partial dropping can be restricted to specific types, such as only data packets, or route control packets that contain it or packets destined to specific nodes.
- Avoid sending a ROUTE ERROR when having detected an error, to prevent other nodes from looking for alternative routes.
- By sending forged routing packets, an attacker can create a so-called black hole, a node where all packets are discarded or all packets are lost.
- Attempt to make routes that go through one appear longer by adding some virtual nodes to the route. Thus, a shorter route will be chosen, avoiding this node.
- Modify the nodes list in the header of a ROUTE REQUEST or a ROUTE REPLY to misroute packets and to add incorrect routes in the route cache of other nodes.
- Decrease the hop count (TTL) when receiving a packet, so that the packet will never be received by the destination. This attack could be detected by the previous node in route by enhanced passive acknowledgment.
- Initiate frequent ROUTE REQUEST to consume bandwidth and energy and to cause congestion.
- Send route replies with a time not proportional to the length of the route. This can give more priority to long routes, thus attracting routes to the attacker, or less priority to short routes, thus avoiding the attacker.

Listed above are some frequent attacks possible on DSR operating without any security measurements.

INTRUSION DETECTION SYSTEMS

Intrusion detection systems (IDS), especially those which are reputation-based, are a new paradigm and are being used for enhancing security in different areas. These systems are lightweight, easy to use and are capable to face a wide variety of attacks as long as they are observable. Among these mechanisms, some of the popular ones are CORE, CONFIDANT, OCEAN and SAFE.

Reputation-Based IDS

Reputation-based IDS do not rely on the conventional use of a common secret to establish confidential and secure communication between two parties. Instead, they are simply based on each other's observations (Buchegger & Le Boudec, 2005). To be more precise, every node in the network moni-

tors the packet emission of its neighboring nodes and derives a reputation value for them. If any misbehavior is detected, this information is broadcasted to the neighboring nodes in order to help them to protect themselves against this fraud (Buchegger & Le Boudec, 2003). Different architectures using the reputation concept for securing packet forwarding have been proposed so far (Resnick & Zeckhauser, 2002). The reputation herein is simply bound to how "good routers" the nodes are. Monitoring the packet loss carried out by the neighborhood is one of the main tasks of these reputation-based systems (Marti, Giuli, & Baker, 2000). The monitoring operation was implemented in CORE and CONFIDANT using a packet overhearing technique based on the promiscuous mode.

Issues Being Addressed

There are few basic problems in MANET that need to be kept in mind while designing any security solution. First, it is often very hard to differentiate intrusions and normal operations or conditions in MANET because of the dynamically changing topology and volatile physical environment. Second, mobile nodes are autonomous units that are capable of roaming independently in unrestricted geographical topology. This means that nodes with inadequate physical protection can be captured, compromised, or hijacked. Third, decision-making in ad hoc networks is usually decentralized and many ad hoc network algorithms rely on the cooperative participation of all nodes. Most ad hoc routing protocols are also cooperative in nature and hence can be easily misguided by false routing information (Yau & Mitchell, 2003).

It is observed that without countermeasure the effect of misbehavior dramatically decreases network performance. Intrusion prevention measures, such as authentication and encryption, can be used as the first line of defense against attacks in MANETs. However, even if these prevention schemes can be implemented perfectly, they still cannot eliminate all attacks, especially the internal or insider attacks. Also, they are costly to implement on mobile nodes from the point of view of limited computation power and energy needed. Another possible solution to this problem is similar to the concept of economic incentives, but the problem with them is that they need a centralized banking system and tamper proof hardware, and a more basic question is who will pay and how much ?

Architecture and Working Principle of Reputation-Based IDS

Reputation-based systems are used for enhancing security in ad hoc networks as they model cooperation between the nodes which is inspired from our social behavior. As in our daily life, when we meet somebody for the first time, we build a reputation about him or her from our personal (firsthand) and some body else's (secondhand) experience. Reputation-based systems are built on this principle. Such systems are used to decide who to trust, and to encourage trustworthy behavior. Resnick and Zeckhauser identify three goals for reputation systems (Resnick & Zeckhauser, 2002):

- To provide information to distinguish between a trustworthy principal and an untrustworthy principal,
- To encourage principals to act in a trustworthy manner, and
- To discourage untrustworthy principals from participating in the service the reputation mechanism is present to protect.

Watchdog and *Path-rater* are some essential components of any Reputation-based Intrusion detection System (Buchegger & Le Boudec, 2004). Complementing DSR with a watchdog increases throughput of mobile ad hoc networks. Misbehavior Detection and Reputation Systems may or may not be distributed. Here, fully distributed means whether information regarding one's reputation is immediately propagated in the whole network or not. In the latter case, nodes are fully dependent on their own personal view about other nodes reputation and behavior.

Distributed IDS protocols either rely only on firsthand information or on positive secondhand information. CONFIDANT and CORE fall into this category. Some basic problems with this approach of global reputation systems are:

- Every node has to maintain O(n) reputation information where n is number of nodes in network.
- Extra traffic generation in reputation exchange.
- Extra computation in accepting indirect reputation information (secondhand information), especially Bayesian Estimation.
- Security issues in reputation exchange such as reputation data packets can be modified.

CONFIDANT, proposed by Buchegger and Le Boudec, detects misbehaving nodes by means of observation or by ALARM signals from neighborhood (Buchegger & Le Boudec, 2002). CONFIDANT aggressively informs nodes in neighborhood about misbehavior of the malicious node. The weight-age of ALARM warning signal depends upon the level of trust that is believed by receiving node. CONFIDANT uses Bayesian Estimation for various measures and calculation of trust and reputation and thus IDS become complex. Each ad hoc running a CONFIDANT system comprises of a:

- *Monitor,* for observation purpose,
- *Reputation Manager,* for calculating reputation of other nodes,

- *Trust Manager,* for calculating level of trust to a particular node, which is used in calculating weightage of ALARM from that node, and
- *Path Manager,* for update path information in route cache as the reputation of neighborhood nodes changes. For example, Deletion of paths containing malicious node, selection of path from various available path option on particular situation and so forth.

CONFIDANT is vulnerable to false accusation if trusted nodes lie or if several liars collude.

Michiardi and Molva (2002) proposed a mechanism called CORE, to enforce node cooperation in mobile ad hoc network. In this mechanism, reputation is a measure of someone's contribution to network operations. Members that have a good reputation can use available resources while members with a bad reputation, because they refused to cooperate, are gradually excluded from the community. CORE defines three types of reputation:

1. Subjective reputation is a reputation value which is locally calculated based on direct observation.
2. Indirect reputation is secondhand reputation information which is established by other nodes.
3. Functional reputation is related to a certain function, where each function is given a weight as to its importance. For example, data packet forwarding may be deemed to be more important than forwarding packets with route information, so data packet forwarding will be given greater weight in the reputation calculations.

CORE reputation values range from positive (+1), through null (0), to negative (-1). CORE suffers from the problem of unwanted consequence of good reputation, where a good node may even wish to decrease its reputation by behaving badly to prevent its resources from being overused. The CORE mechanism assumes that every node will use the same reputation calculations and will also assign the same weights to the same functions. This is a potentially inappropriate assumption in heterogeneous ad hoc networks, where devices with different capabilities and roles are likely to place different levels of importance on different functions depending upon CPU usage, battery usage and so forth. One can take advantage of this situation and may perform only those functions which have higher preferences in calculating reputation.

A second type of IDS is one that solely depends upon the firsthand observation for reputation maintenance. Nodes make routing decision based on only the direct observation of its neighbor's node. This eliminates most of the trust manager complexity but in highly mobile ad hoc network it might not be appropriate to only depend solely upon personal observation. But also using secondhand information can significantly accelerate the detection and subsequent isolation of malicious nodes in mobile ad hoc networks.

OCEAN by Bansal and Baker relies exclusively on firsthand observations for ratings and avoids indirect (secondhand) reputation information. In OCEAN, the rating of each node is initialized to Neutral (0), with every positive action resulting in an increment (+1) of the rating, and every negative action resulting in a decrement (-2) of the rating (Bansal & Baker, 2003). Once the rating of a node falls below a certain faulty threshold (-40), the node is added to a faulty list. The faulty list represents a list of misbehaving nodes. If the rating is below the faulty threshold, the node is added to the faulty list. This faulty list is appended to the route request by each node broadcasting it to be used as an avoid list. A route is rated good or bad depending on whether the next hop is on the faulty list. In addition to the rating, nodes keep track of the forwarding balance with their neighbors by maintaining a chip count for each node.

OCEAN's approach is to disallow any secondhand reputation exchanges. Routing decisions are made based solely on direct observations of neighboring nodes behavior. This eliminates most trust management complexity. The basic problem with OCEAN is that it does not take secondhand information that can significantly improve detection of malicious nodes. Also, authors only consider individual bad behavior, not collusion of nodes.

CONCLUSION

Mobile ad hoc networks have a number of significant security issues which cannot be solved alone by Intrusion detection systems. Physical security of nodes is another very important issue. Reputation systems are used to establish trust and encourage trustworthy behavior and cooperation among nodes. In this article, we have critically examined the existing systems and outlined their strength and shortcomings.

REFERENCES

Bansal, S., & Baker, M. (2003). *Observation-based cooperation enforcement in ad hoc networks*. Retrieved December 10, 2007, from http://arxiv.org/pdf/cs.NI/0307012

Buchegger, S., & Le Boudec, J.Y. (2002). Performance analysis of the CONFIDANT Protocol: Cooperation of nodes—fairness in dynamic ad hoc networks. In *Proceedings of the IEEE/ACM Symposium on Mobile Ad Hoc Networking and Computing (MobiHOC),* Lausanne, CH, (pp. 226-236).

Buchegger, S., & Le Boudec, J.Y. (2003). The effect of rumor spreading in reputation systems for mobile ad hoc

networks. In *Proceedings of WiOpt `03: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Sophia-Antipolis, France.

Buchegger, C.T., & Le Boudec, J.Y. (2004). A test-bed for misbehavior detection in mobile ad hoc networks—how much can watchdogs really do? In *WMCSA: Proceedings of the Sixth IEEE Workshop on Mobile Computing Systems and Applications*.

Buchegger, S., & Le Boudec, J.Y. (2005). Self-policing mobile ad hoc networks by reputation systems. *IEEE Communications Magazine*.

Johnson, D.B., Maltz, D.A., & Broch, J. (2001). DSR: The dynamic source routing protocol for multi-hop wireless ad hoc networks. In C.E. Perkins (Ed.), *Ad Hoc Networking* (pp. 139-172). Addison-Wesley.

Marti, S., Giuli, T.J, Lai, K., & Baker, M. (2000). Mitigating routing misbehavior in mobile ad hoc networks. In *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking Table of Contents*, (pp. 255–265).

Michiardi, P., & Molva, R. (2002). Core: A collaborative reputation mechanism to enforce node cooperation in mobile ad hoc networks. In *Proceedings of the IFIP Communication and Multimedia Security Conference*.

Perkins, C.E., & Royer, E.M. (1999). Ad hoc on-demand distance vector routing. In *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications*, New Orleans, LA, (pp. 90-100).

Resnick, P., & Zeckhauser, R. (2002). Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system. In M. R. Baye (Ed.), *The economics of the Internet and e-commerce: Advances in applied microeconomics* (Vol. 11, pp. 127-157). Amsterdam, Elsevier Science.

Yau, P., & Mitchell, C.J. (2003). Reputation methods for routing security for mobile ad hoc networks. In *Proceedings of SympoTIC '03, Joint IST Workshop on Mobile Future and Symposium on Trends in Communications*, Bratislava, Slovakia.

KEY TERMS

Ad Hoc Network: A mobile ad hoc network (MANET) is a kind of wireless ad hoc network, and is a self-configuring network of mobile routers (and associated hosts) connected by wireless links—the union of which form an arbitrary topology.

Bandwidth: Bandwidth is a measure of frequency range and is typically measured in hertz. Bandwidth is related to channel capacity for information transmission.

Denial of Service (DoS): Is an attempt to make a computer resource unavailable to its intended users. Typically, the targets are high-profile Web servers where the attack is aiming to cause the hosted Web pages to be unavailable on the Internet. It is a computer crime that violates the Internet proper use policy as indicated by the Internet Architecture Board (IAB).

Firewalls: A logical barrier designed to prevent unauthorized or unwanted communications between sections of a computer network.

Gateway: A computer or a network that allows or controls access to another computer or network.

Intrusion Detection System (IDS): Is used to detect many types of malicious network traffic and computer usage that can't be detected by a conventional firewall.

Promiscuous Mode: Refers to a configuration of a network interface wherein a setting is enabled so that the interface passes all traffic it receives to the CPU rather than just packets addressed to it, a feature normally used for packet sniffing.

Routers: A router acts as a junction between two or more networks to transfer data packets among them.

Reputation: As a socially transmitted belief (i.e., belief about belief) concerns properties of agents, namely their attitudes toward some socially desirable behavior, be it cooperation, reciprocity, or norm-compliance.

Terminal: In the context of telecommunications, a terminal is a device which is capable of communicating over a line.

Category: Networking & Telecommunication

Mobile Ad Hoc Networks

Carlos Tavares Calafate
Technical University of Valencia, Spain

Pedro Pablo Garrido
Miguel Hernández University, Spain

José Oliver
Technical University of Valencia, Spain

Manuel Pérez Malumbres
Miguel Hernández University, Spain

INTRODUCTION

This chapter offers a state-of-the-art review in mobile ad hoc networks (MANETs). It first introduces the history of ad hoc networks, explaining the ad hoc network concept and referring to the main characteristics of these networks and their fields of application.

It then focuses on technologies and protocols specific to ad hoc networks. Firstly, it refers to relevant proposals targeting the PHY/MAC layers. Secondly, it discusses the different routing protocol proposals for ad hoc networks according to the category to which they belong. Finally, it includes an overview of the different protocols proposed for ad hoc networks at the transport layer. The chapter concludes with some remarks on future trends in these networks.

BACKGROUND

The history of wireless networks dates from the 1970s. In fact, radio communications and computer networks were first combined by the University of Hawaii, in 1971, in an experimental network named ALOHANET. That network offered bidirectional communications following a star topology, and its purpose was to allow communicating with US mainland. During the 1980s, the technology was improved, and towards the end of the 1990s, interest on wireless networks reached a peak, mainly due to the fast growth of the Internet.

Nowadays we can split existing wireless networks into different categories according to their scope and size. Wireless wide area networks (WWANs), such as GSM and UMTS (Ojanpera, T. & Prasad, R., 1998), usually cover hundreds of kilometers and use private frequency bands. Such networks are usually owned and maintained by telecommunications providers, and their purpose is to offer services in a country or a region of it. Wireless metropolitan area networks (WMANs), such as WiMax (IEEE 802.16 WG, 2004), typically have a range of a few kilometers, and can operate over both private and public frequency bands, so that both telecommunication companies and private users can take advantage of them. Wireless local area networks (WLANs), such as WiFi (IEEE 802.11 WG, 1999), usually cover areas between a few tens of meters up to a kilometer. They typically use public frequency bands so that users can freely install and use them. At the lower end, we have wireless personal area networks (WPANs), such as Bluetooth (IEEE 802.15 WG, 2005), which also use free frequency bands that are used to replace cables within a very limited area around a single user (few meters).

This chapter focuses on recent developments in terms of infrastructure-less wireless networks, more commonly known as ad hoc networks, that extend WLAN technologies to offer more flexible solutions. All nodes within an ad hoc network provide a peer-level multihopping routing service to allow out-of-range nodes to be connected. Unlike a wired network, nodes in an ad hoc network can move freely, thus giving rise to frequent topology changes.

Such a network may operate in a stand-alone fashion or be connected to the larger Internet. An ad hoc architecture has many benefits, such as self-reconfiguration and adaptability to highly variable characteristics, namely, power and transmission conditions, traffic distribution variations, and load balancing. However, those benefits come with many challenges. New algorithms, protocols, and middleware have to be designed and developed to create a truly flexible and decentralized network.

In terms of applications, ad hoc networks offer the required flexibility to adapt to situations where no sort of infrastructure is available. Examples of such situations are army units moving inside hostile territories, or organized teams, such as firemen, performing rescue tasks. In general, mobile ad hoc networks can be used on all those situations characterized by lack of fixed infrastructure, peer-to-peer communication, and mobility support.

TECHNOLOGIES AND PROTOCOLS FOR AD HOC NETWORKS

A. PHY/MAC Layer Technologies

Throughout the past few years, novel solutions for MAC/PHY layers have been sought in the wireless ad hoc networking field. In particular, there have been several proposals targeting the MAC layer (Kumar, Raghavan, & Deng, 2006).

Despite the many proposals available, very few have made it to the market. Nowadays, almost every ad hoc network relies on IEEE 802.11 technology (IEEE 802.11 WG, 1999), which defines both physical and MAC layers. Since this standard has gained much relevance, we now offer more details about it.

In 1997, IEEE group 802.11 was created. The purpose was to create a technology for wireless local area networks operating on ISM (industrial, scientific, and medical) frequency bands. With that purpose, a MAC layer and three different physical layers were defined, operating at 1 and 2 Mbit/s:

- Infrared (IR) – baseband
- Frequency hopping spread spectrum (FHSS) – 2.4 GHz band
- Direct sequence spread spectrum (DSSS) – 2.4 GHz band

In December 1999, the IEEE 802.11a standard was completed, proposing a different technique for the physical layer named orthogonal frequency division multiplexing (OFDM). This technology was able to offer up to 54 Mbit/s on the 5 GHz band. A year later, in January 2000, the IEEE 802.11b standard was completed, consisting basically of an extension to the original standard, offering up to 11 Mbit/s on the 2.4 GHz band. Only in July 2003 was the IEEE 802.11g standard completed, offering 54 Mbit/s speeds on the 2.4 GHz frequency band. Recently, the 802.11n group is proposing higher speed extensions to the standard, targeting data rates above 300 Mbit/s.

Concerning 802.11's MAC layer, its main functions are reliable data delivery, fair access to the wireless media, and data protection. Moreover, it is responsible for a correct operation in noisy, unreliable environments.

The 802.11 standard offers two different medium access mechanisms:

- Distributed coordination function (DCF), a mandatory access mechanism based on CSMA/CA. (*carrier sense multiple access with collision avoidance*).
- Point coordination function (PCF), optional, based on a polling method to support services with time restrictions.

Since the latter only applies to access points, in ad hoc networks, the DCF must be used instead. Despite that the ad hoc mode proposed by the IEEE 802.11 standard did not specifically target multihop ad hoc networks, it is widely used and offers relatively good performance.

B. Routing Protocols

A routing protocol is required when a packet must go through several hops to reach its destination. It is responsible for finding a route for the packet and making sure it is forwarded through the appropriate path.

Routing techniques used can be divided into three families: *distance vector* (Bellman, 1957; Ford & Fulkerson, 1962), *link state* (Dijkstra, 1959), and *source routing* (Estrin, Li, Rekhter, Varadhan, & Zappala, 1996).

Internet routing protocols, based on these techniques, generate periodic control messages, a procedure that is not adequate for a large mobile network with long routes since it would result in a large number of control messages. Reducing routing overhead is critical for mobile nodes since CPU use, as well as radio transmissions and receptions, would cause batteries to be quickly depleted.

We will now present different routing protocol proposals for MANETs that are currently available. We have organized them into three groups: proactive, reactive, and other strategies, being that the latter embraces all those that do not fall under the former two categories.

Proactive Routing Protocols

When using proactive routing protocols, all the nodes (routers) periodically exchange routing information, with the aim of maintaining a consistent, updated, and complete network view. Each node uses the exchanged information to calculate the costs towards all possible destinations. That way, if a destination is found, there will always be a route available towards it.

The main advantage of proactive routing schemes is that there is no initial delay when a route is required. On the other hand, these are usually related to a greater overhead and a larger convergence time than for reactive routing techniques, especially when mobility is high. To increase the performance in ad hoc networks, both *link-state* and *distance vector* algorithms were modified. Examples of routing protocols using *distance vector* techniques are the *destination-sequenced distance vector* (DSDV) (Perkins & Bhagwat, 1994) and the *wireless routing protocol* (WRP) (Murthy & Garcia-Luna-Aceves, 1996). Examples of *link-state* based protocols are the *optimized link state routing* (OLSR) (Clausen, Jacquet, Laouiti, Muhlethaler, Qayyum, & Viennot 2001), and the *topology broadcast reverse path forwarding* (TBRPF) (Bellur & Ogier, 1999).

Reactive Routing Protocols

Reactive routing does not depend, in general, on periodic exchange of routing information or route calculation. Therefore, when a route is required, the node must start a route discovery process. This means that it must disseminate the route request throughout the network and wait for an answer before it can proceed to send packets to the destination. The route is maintained until the destination is unreachable or until the route is no longer necessary. By following this strategy, reactive routing protocols minimize the resource consumption by avoiding the maintenance of unused routes. On the other hand, the route discovery process causes a significant startup delay and causes a considerable waste of resources. If the network is wide enough, the overhead will be similar or superior to that achieved with proactive routing protocols.

The most common routing algorithms found among reactive routing protocols are *distance vector* and *source routing*. Examples of reactive routing protocols are the *ad hoc on-demand distance vector* (AODV) (Perkins & Royer, 1999), the *dynamic source routing* (DSR) (Johnson, Maltz, & Hu, 2004), and the dynamic on-demand routing protocol (DYMO) (Chakeres & Perkins, 2008).

Other Strategies

There are other strategies proposed for the design of routing protocols. There are, for instance, hybrid solutions, such as the *zone routing protocol* (ZRP) (Pearlman & Hass, 1999), which uses both reactive and proactive concepts. Some protocols are based on *clustering* and hierarchical architectures, such as the *distributed mobility-adaptive clustering* (DMAC) (Basagni, 1999) and the *cluster-based energy saving algorithm* (CERA) (Cano, Kim, & Manzoni, 2003).

The LAR protocol (Ko & Vaidya, 1998) tries to avoid the flooding associated with route discovery by using GPS information so that only those nodes on a certain geographic area between source and destination must retransmit route requests.

Finally, *power aware routing* (PAR) (Singh, Woo, & Raghavendra, 1998) is a solution that intends to improve the power consumption by taking into account the battery lifetime, selecting those routes that minimize the energy consumption of the system.

C. Transport Protocols

The transmission control protocol (TCP) is perhaps the most important and widely used transport protocol in the Internet. Most applications, such as Web, mail, SSH, and peer-to-peer file exchange, depend on it for the reliable delivery of data on an end-to-end basis.

Since TCP was designed for the Internet environment, it is prone to suffer from poor performance in wireless networks, especially in mobile ad hoc networks. The main reasons have to do with packet losses and node mobility.

In the Internet environment, the physical media is very reliable, and the path traversed by packets is typically the same throughout the duration of a connection. So, losses are usually related to congestion. TCP's congestion control mechanisms act upon packet losses to regulate the data rate, being quite effective for the Internet. Contrarily to wired media, wireless transmission are prone to frequent bit errors due to fluctuations in signal-to-noise ratio (SNR), multipath and shadowing effects, and so forth. Such errors are not related to congestion, and so, should not receive a similar treatment at the transport layer.

Mobile ad hoc networks suffer from frequent topology updates that require highly adaptive routing protocols, as referred previously. Route maintenance, though, is not instantaneous, and often causes large groups of packets to be delayed and/or lost. This occurrence is not related to congestion either and, therefore, should also receive a differentiated treatment.

Due to the aforementioned problems, specific transport layer proposals for ad hoc network environments are available in the literature. They can be grouped into three different categories:

- Solutions that propose improvements to the TCP protocol
- TCP-aware cross layer solutions
- Transport protocols specific to ad hoc networks

In terms of solutions proposing improvements to the TCP protocol, the most relevant work in the field is ELFN (explicit link failure notification) (Holland & Vaidya, 1999). This solution mitigates the route discovery problem through explicit link failure notification from network nodes to the TCP sender. The sender then enters a hold state, periodically probing the network to assess if the path has been reestablished. When a new path is available, the TCP agent returns to its previous state (before the path was lost), hence, improving resource usage.

Concerning TCP-aware cross-layer solutions, the most relevant work in the field is the Atra framework (Anantharaman & Sivakumar, 2002). This proposal basically consists of three mechanisms, two at the routing layer and one at the MAC layer, that cooperate to improve the performance of TCP. At the MAC layer, there is a mechanism that predicts route failures to improve routing tasks. At the routing layer, it includes a mechanism, symmetric route pinning, to reduce the frequency of route failures. It also includes a proactive mechanism that informs all interested nodes about failing links, improving global performance.

Finally, in terms of protocols specific to ad hoc networks, the most relevant proposal in the field is the ad hoc transport protocol (ATP) (Sundaresan & Anantharaman, 2003). This protocol consists of a complete redesign of the transport layer for optimum performance in ad hoc network environments. Its main characteristics are the use of rate-based transmissions instead of TCP's sliding windows paradigm, a quick start mechanism, a delay-based congestion indicator, and a feedback mechanism from receiver to source that includes SACK (selective acknowledgements) blocks similar to those proposed in TCP-SACK (Mathis, Mahdavi, Floyd, & Romanow, 1996).

FUTURE TRENDS

The field of mobile ad hoc networks is still under intensive research. New application areas are emerging, such as vehicular ad hoc networks (VANETs), that rely on ad hoc connections between vehicles to improve road safety. The sensor networks area is also strongly related to ad hoc networks, and a merge of some of the ideas and solutions employed is prone to occur. Wireless mesh networking (WMN) is an area also intimately related to mobile ad hoc networks; the former are characterized by minimal or no mobility compared to the latter.

Concerning improvements to the MAC layers, the IEEE 802.11e standard represents an important enhancement to the MAC layer to offer quality of service (QoS) support. In the future, we expect to see MAC layer solutions that further improve QoS traffic discrimination at this layer.

In terms of routing, a merge of independent solutions is required to offer a protocol that takes into consideration issues such as power consumption, security, anonymity, QoS, as well as the physical and MAC layers used.

For the transport layer, a cross-layer solution specific to ad hoc networks offering efficient support to both best effort and real-time traffic is still one of the missing points. Since these networks are very prone to errors, enhancements to the transport layer are also expected to include advanced error correction techniques that completely avoid retransmissions up to a certain loss rate.

CONCLUSION

Mobile ad hoc networks are a field under intensive research due to their flexibility and lack of requirements in terms of infrastructure. Currently, several solutions are available for the different network layers involved, physical, MAC, routing, and transport, both in terms of theoretical and real-world implementations of protocols and technologies. Despite the ongoing efforts, there is still much room for improvement, since the performance of these networks is typically poor compared to other wireless technologies, such as UMTS, WiMax, and so forth.

In years to come, and with the advent of novel applications requiring these networks (e.g., VANETs), it is expected that this type of network will become widely adopted by the industry, resulting in the deployment of new products and solutions that rely on ad hoc networks to offer a set of functionalities and services that no other technology is able to offer. Once the technology becomes mature, it can be adopted also for critical missions, such as rescue, disaster, and military scenarios.

REFERENCES

Anantharaman, V., & Sivakumar, R. (2002). A microscopic analysis of TCP performance over wireless ad hoc networks. In *Proceeding of ACM SIGMETRICS 2002*, Marina del Rey (CA), USA.

Basagni, S. (1999). Distributed clustering for ad hoc networks. In *Proceedings of the IEEE International Symposium on Parallel Architectures, Algorithms, and Networks (I-SPAN)* (pp. 310–315), Perth, Western Australia.

Bellman, R. E. (1957). *Dynamic programming*, Princeton, NJ: Princeton University Press.

Bellur, B., & Ogier, R.G. (1999). A reliable, efficient topology broadcast protocol for dynamic networks. In *Proc. IEEE INFOCOM, The Conference on Computer Communications*, New York, USA.

Cano, J. C., Kim, D., & Manzoni, P. (2003). CERA: Cluster-based energy saving algorithm to coordinate routing in short-range wireless networks. In *Proc. International Conference on Information Networking (ICOIN)*, Jeju Island, Korea.

Chakeres, I., & Perkins, C. (2008). *Dynamic manet on-demand (DYMO routing*. Retrieved from http://www.ietf.org/internet-drafts/draft-ietfmanet-dymo-11.txt

Clausen, T., Jacquet, P., Laouiti, A., Muhlethaler, P., Qayyum, A., & Viennot L. (2001). Optimized link state routing protocol. *International Multi Topic Conference, Pakistan*.

Dijkstra, E. W. (1959). A note on two problems in connection with graphs, *Numer. Math., 1*, 269-271.

Estrin, D., Li, T., Rekhter, Y., Varadhan, K., & Zappala, D. (1996). *Source demand routing: Packet format and forwarding specification (Version 1)*, IETF RFC 1940.

Ford, L. R., & Fulkerson, D. R. (1962). *Flows in networks.* Princeton, NJ: Princeton University Press.

Holland, G., & Vaidya, N. H. (1999). Impact of routing and link layers on TCP performance in mobile ad hoc networks.

In *Proceedings of the IEEE WCNC*, New Orleans, USA.

IEEE 802.11 WG. (1999). *International standard for information technology - Telecom. and information exchange between systems – Local and metropolitan area networks - Specific requirements - Part 11: Wireless medium access control (MAC) and physical layer (PHY) specifications*, IEEE 802.11 WG, ISO/IEC 8802-11:1999(E) IEEE Std. 802.11.

IEEE 802.15 WG. (2005). *IEEE Standard for information technology--Telecommunications and information exchange between systems-- Local and metropolitan area networks--Specific requirements. Part 15.1: Wireless medium access control (MAC) and physical layer (PHY) specifications for wireless personal area networks (WPANs)*. IEEE Std. 802.15.1.

IEEE 802.16 WG. (2004). *IEEE standard for local and metropolitan area networks Part 16: Air interface for fixed broadband wireless access systems*. IEEE Std. 802.16.

Johnson, D. B., Maltz, D. A., & Hu, Y-C. (2004). *The dynamic source routing protocol*. Internet Draft, MANET Working Group, draft-ietf-manet-dsr-10.txt.

Ko, Y. B., & Vaidya, N. H. (1998). Location aided routing (LAR) in mobile ad hoc networks. In *The Annual International Conference on Mobile Computing and Networking (MOBICOM)*, Dallas (TX), USA.

Kumar, S., Raghavan, V. S., & Deng, J. (2006). Medium access control protocols for ad hoc wireless networks: A survey. *Elsevier Ad Hoc Networks, 4*(3), 326–358.

Mathis, M., Mahdavi, J., Floyd, S., & Romanow, A. (1996). *TCP selective acknowledgment options*, IETF RFC 2018.

Ojanpera, T., & Prasad, R. (1998). An overview of third-generation wireless personal communications: A European perspective. *IEEE Personal Communications, 5*(6), 59-65.

Pearlman, M.R., & Haas, Z.J. (1999). Determining the Optimal Configuration for the Zone Routing Protocol. *IEEE Journal on Selected Areas in Communications*, *17*(8), 1395-1414.

Perkins, C. E., & Bhagwat, P. (1994). Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers. *ACM Computer Communication Review, 24*(2), 234–244.

Perkins, C. E., & Royer, E. M. (1999). Ad hoc on-demand distance vector routing. In *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications,* New Orleans (LA), USA.

Singh, S., Woo, M., & Raghavendra, C. S. (1998). Power-aware routing in mobile ad hoc networks. In *Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking*, Dallas (TX), USA.

Sundaresan, K., & Anantharaman, V. (2003). ATP: A reliable transport protocol for ad hoc networks, In *Proceedings of ACM MOBIHOC*, Maryland, USA.

KEY TERMS

Distance Vector: Routing technique that maintains a table for the communication taking place, and employs diffusion (not flooding) for information exchange between neighbors. All the nodes must calculate the shortest path towards the destination using the routing information of their neighbors.

Link State: Routing protocols, based on this technique, maintain a routing table with the full topology. The topology is built by finding the shortest path in terms of link cost, cost that is periodically exchanged among all the nodes through a flooding technique.

MAC Layer: The medium access control layer is a protocol layer embedded within the link layer that is responsible for coordinating the access to a shared medium according to a set of rules.

Node: In the context of mobile ad hoc networks (MANETs), it usually refers to a mobile terminal, such as a PDA, laptop, smartphone, or other device with wireless communication capabilities that participates in the networks both as a traffic generator and traffic forwarder.

Source Routing: Technique where all the data packets have the routing information on their headers. The route decision is made on the source node, which avoids routing loops entirely.

SSH: Secure shell is a protocol that allows accessing a remote computer in a secure manner by employing cryptographic techniques. Usually, the term refers also to the client/server tools that support this protocol.

VANET: Vehicular ad hoc network, consisting of a network of vehicles, moving at a relatively high speed, that communicate among themselves with different purposes, being the main purpose that of improving security on the road.

Mobile Agent Authentication and Authorization in E-Commerce

Sheng-Uei Guan
National University of Singapore, Singapore

INTRODUCTION

With the increasing worldwide usage of the Internet, electronic commerce (e-commerce) has been catching on fast in a lot of businesses. As e-commerce booms, there comes a demand for a better system to manage and carry out transactions. This has led to the development of agent-based e-commerce. In this new approach, agents are employed on behalf of users to carry out various e-commerce activities.

Although the tradeoff of employing mobile agents is still a contentious topic (Milojicic, 1999), using mobile agents in e-commerce attracts much research effort, as it may improve the potential of their applications in e-commerce. One advantage of using agents is that communication cost can be reduced. Agents traveling and transferring only the necessary information save the bandwidth and reduce the chances of network clogging. Also, users can let their agents travel asynchronously to their destinations and collect information or execute other applications while they can disconnect from the network (Wong, 1999).

Although agent-based technology offers such advantages, the major factor that is holding people back from employing agents is still the security issues involved. On the one hand, hosts cannot trust incoming agents belonging to unknown owners, because malicious agents may launch attacks on the hosts and other agents. On the other hand, agents may also have concerns on the reliability of hosts and will be reluctant to expose their secrets to distrustful hosts.

To build bilateral trust in an e-commerce environment, the authorization and authentication schemes for mobile agents should be well designed. Authentication checks the credentials of an agent before processing the agent's requests. If the agent is found to be suspicious, the host may decide to deny its service requests. Authorization refers to the permissions granted for the agent to access whichever resource it requested.

In our previous work, we have proposed a SAFER (Secure Agent Fabrication, Evolution & Roaming) architecture (Zhu, 2000), which aims to construct an open, dynamic and evolutionary agent system for e-commerce. We have already elaborated agent fabrication, evolution, and roaming in Guan (1999, 2001, 2002), Wang (2001), and Zhu (2001). This article gives an overview of the authentication and authorization issues on the basis of the SAFER architecture.

BACKGROUND

Many intelligent agent-based systems have been designed to support various aspects of e-commerce applications in recent years, for example: Kasbah (Chavez, 1998), Minnesota AGent Marketplace Architecture (MAGMA) (Tsvetovatyy, 1997), and MAgNet (Dasgupta, 1999). Unfortunately, most current agent-based systems such as Kasbah and MAGMA are serving only stationary agents. Although MAgNet employs mobile agents, it does not consider security issues in its architecture.

D'Agents (Gray, 1998) is a mobile agent system, which employs the PKI for authentication purposes, and uses the RSA (Rivest, Shamir, & Adleman, 1978) public key cryptography (Rivest et al., 1978) to generate the public-private key pair. After the identity of an agent is determined, the system decides what access rights to assign to the agent and sets up the appropriate execution environment for the agent.

IBM Aglets (Lange, 1998; Ono, 2002) are Java-based mobile agents. Each aglet has a globally unique name and a travel itinerary (wherein various places are defined as context in IBM Aglets). The context owner is responsible for keeping the underlying operating system secure, mainly protecting it from malicious aglets. Therefore, he or she will authenticate the aglet and restrict the aglet under the context's security policy.

Ajanta is also a Java-based mobile agent system (Karnik, 1999, 2001, 2002) employing a challenge-response based authentication protocol. Each entity in Ajanta registers its public key with Ajanta's name service. A client has to be authenticated by obtaining a ticket from the server. The Ajanta Security Manager grants agents permissions to resources based on an access control list, which is created using users' Uniform Resource Names (URNs).

iJADE (intelligent Java Agent Development Environment) (Lee, 2002) provides an intelligent agent-based platform in the e-commerce environment. This system can provide fully automatic, mobile and reliable user authentication.

Under the public key infrastructure (PKI), each entity may possess a public-private key pair. The public key is known to all, while the private key is only known to the

key owner. Information encrypted with the public key can only be decrypted with the corresponding private key. In the same note, information signed by the private key can only be verified with the corresponding public key (Rivest, 1978; Simonds, 1996). The default algorithm that generates the key pairs is the digital signature algorithm (DSA), working in the same way as a signature on a contract. The signature is unique, so that the other party can be sure that you are the only person who can produce it.

MAIN THRUST OF THE ARTICLE

This article presents an overview of the architecture based on SAFER (Secure Agent Fabrication, Evolution & Roaming) (Zhu, 2000) to ensure a proper authentication and authorization of agent. Here, the public key infrastructure (PKI) is used as the underlying cryptographic scheme. Also, agents can authenticate the hosts to make sure that they are not heading to a wrong place. According to the level of authentication that the incoming agent has passed, the agent will be categorized and associated with a relevant security policy during the authorization phase. The corresponding security policy will be enforced on the agent to restrict its operations at the host. The prototype has been implemented with Java.

Design of Agent Authentication and Authorization

Overview of the SAFER Architecture

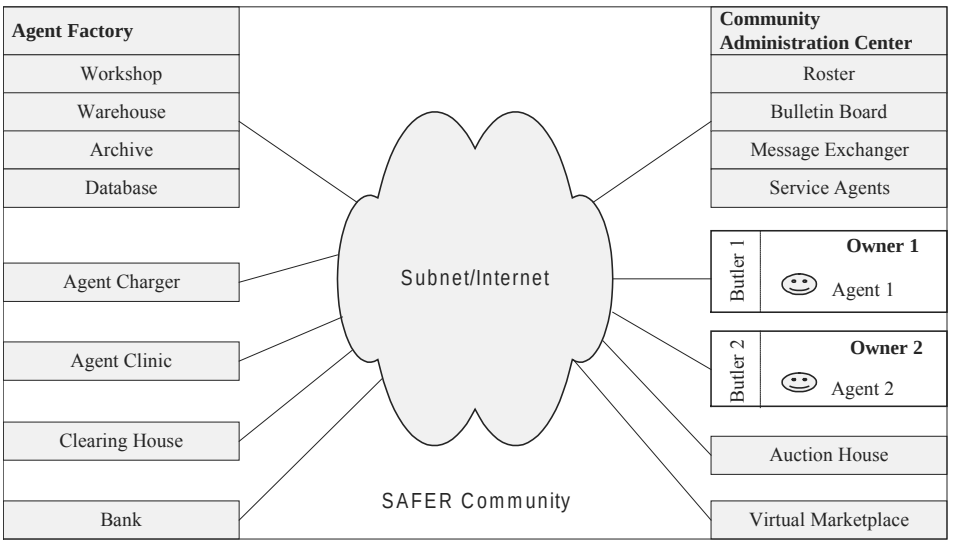
The SAFER architecture comprises various communities and each community consists of the following components (see Figure 1): Agent Owner, Agent Factory, Agent Butler, Community Administration Center, and so forth. The Agent Owner is the initiator in the SAFER environment, and requests the Agent Factory to fabricate the agents it requires. The Agent Butler is a representative of the Agent Owner authorized by the owner to coordinate the agents that are dispatched. Owner can go offline after dispatching his or her agents, and thereafter the butler can take over the coordination of the agents. The Agent Factory fabricates all the agents. This is the birthplace of agents and is thus considered a good source to check malicious agents. The Community Administration Center (CAC) is the administrative body, which has a roster that keeps the data of the agents that are in the community. It also collects information, such as addresses of new sites that agents can roam to.

Agent Structure and Cryptographic Schemes

In SAFER, mobile agents have a uniform structure. The agent credentials (hard-coded into the agent (Guan, 2000, 2001)) are the most important part of the agent body and are immutable. This part includes FactoryID, AgentID, Expiry Date, and so forth. The Agent Factory then signs this immutable part. When the receiving host accepts an agent, it can verify with the Agent Factory's public key whether the agent's credentials have been modified. The mutable part of the agent includes the Host Trace, which stores a list of names of the hosts that the agent has visited so far. Upon checking, if any distrusted host is found, a host may decide not to trust this agent and impose a stricter security policy on it.

In SAFER, the main cryptographic technology used is the PKI. The public keys are stored in a common database

Figure 1. SAFER architecture



located in CAC, where the public has read access, but no access to modify existing records.

Authentication Process

Authenticating Host

Before roaming to the next host, it is the duty of the agent to authenticate the next host to make sure that the host it is visiting is a genuine one. A hand-shaking method is devised to authenticate hosts. The agent sends its request to a host, asking for permission to visit. The host will sign on the agent's request message with its private key and send it back to the agent. The agent can then verify the host's identity by extracting the host's public key from the common database and authenticating the signature. If the authentication is successful, then the agent is communicating with the genuine host and starts to ship itself over to the host.

Authenticating Agent

Authentication of an agent involves two major steps: 1) to verify the agents' credentials, and 2) to verify the mutable part of the agent, checking whether it has been tampered with by anyone in its roaming process.

The authentication procedure is shown in Figure 2. Firstly, the agent will be checked for its expiry date. If it has not expired, its host trace will be examined to see if it has been to any distrusted host. If the agent passes these two tests, the final test is to check the trustworthiness of the factory that has manufactured it.

Authorization Process

After the host accepts an agent, it has to determine what resources the agent is allowed to access based on the level of authentication that the agent has passed. Four levels of authorization have been designed, with level 1 being the strictest and level 4 the most lenient. Level 1 authority is given to agents that the host does not have much trust in. An agent that passes all levels of authentication and is deemed to be trusted may be awarded the level 4 authority. Table 1 shows the four policies and the restrictions imposed on each policy. The permissions can be customized to meet the requirements of different hosts. Here, AWT stands for Abstract Window Toolkit (AWT), which allows our programs to create a Graphical User Interface (GUI) to interact with the users.

Implementation

Implementation of agent authentication and authorization was done using the Java programming language. The Java Security API and the Java Cryptography Extension were widely used

Figure 2. Authentication & authorization procedure

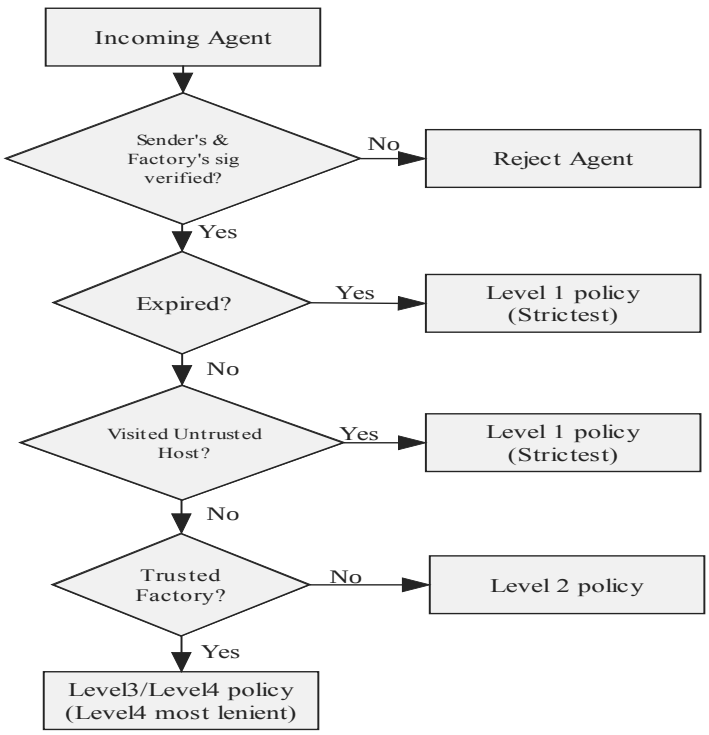


Table 1. Definition of the various security policies

Level of leniency	Policy name	Permissions
Level 4 (Most lenient)	Polfile.policy	FilePermission (Read, write)
		AWT Permission
		Socket Permission (Accept, Listen, Connect)
		Runtime Permission (create/set SecurityManager, queuePrintJob)
Level 3	Pol1.policy	FilePermission (Read, write)
		AWT Permission
		Socket Permission (Accept, Listen, Connect)
		Runtime Permission (create SecurityManager, queuePrintJob)
Level 2	Pol2.policy	FilePermission (Read only)
		AWT Permission
		Socket Permission (Accept, Connect)
		Runtime Permission (create SecurityManager)
Level 1 (Most Strict)	Pol3.policy	FilePermission (Read only)
		No AWT Permission
		No Socket Permission
		No Runtime Permission

in the implementation. The graphical user interfaces were designed using the Java Swing components.

Discussions

We do not intend to compare our prototype with mobile agent systems such as D'agents and Aglets on detail benchmarks, as the focus of our approach is on the security issues in the context of e-commerce applications. Here, we present our features in comparison to related work, and discuss the advantages and limitations of our system in the following subsections.

Our approach has some features that are similar to the related systems discussed in the second section. For example, the authentication mechanism is based on PKI. The authorization mechanism is implemented using the Java security manager and some user-defined security policies. The major features of our approach lie in the method of authorization. Some agent systems authorize agents based on a role-based access control list and some are based on the identity of the agent. The SAFER system is different in that it allocates a different security policy based on the identity of the agent and the level of authentication it has passed.

Advantages of Our Infrastructure

Storage of Keys

One of the principal advantages of the prototype implemented is that there is no sending of keys over the network. This enhances the security of the system since it is impossible that keys can get intercepted and replaced.

The database storage of the public keys also allows an efficient method of retrieval of keys. This facilitates the verification of all previous signatures in the agent by the current host. For example, the owner may want to verify the signatures of all the previous hosts that the agent has visited. Instead of having all these hosts append their public keys to the agent (which may be compromised later), the owner can simply retrieve the keys from the database according to the hosts' ID.

Examining the Agent's Host Trace

Every host is required to sign on the agent's host trace before it is dispatched to the next destination. The IDs of the hosts visited are compared with the distrusted host list that each host would keep. If a distrusted host were found, the host would then take special precautions against these agents by imposing a stricter security policy on its operations.

Automation of the Authentication and Authorization Process

The beauty of this system is that it can be automated or run manually when the need arises. In the automatic configuration, when an agent is sent over, the host will do the authentication and assign an appropriate security policy to the agent. If the execution is successful, the host signs the agent and adds its identity on the host trace, before sending it out to the next host. In the manual configuration, all the authentication and authorization procedures need prompting from the host owner. The advantage is that the host has more control on what methods to authenticate and what authorization level and policy to enforce on the agent.

Limitations of Our Infrastructure

Pre-Determined Security Policies

In the current design, the agent is assigned to the security policy based on the authentication process. Having pre-determined security policies may be stifling to the operations of an agent. It would be useless if the agent is denied the read access but instead granted other permissions that it does not need. The limitation here is an implementation choice because the mechanism to customize the permission for each agent has not been developed. Pre-determined security policies are simpler to implement for large-scale systems.

Difficulty in Identifying a Malicious Host

The current implementation does not have a way of identifying the host that is causing the attacks on the agent. The agent owner can only detect that certain information has been tampered with, but does not know which host exactly caused the disparity.

FUTURE TRENDS

The implementation of the prototype has provided a basic infrastructure to authenticate and authorize agents. We are improving our approaches and implementation in two aspects. Firstly, to make the system more flexible in enforcing restrictions on agents, a possible improvement is to let the agent specify the security policy that it requires for its operation at the particular host. It is desirable to have a personalized system with the agent stating what it needs and the host deciding on whether to grant the permission or not. Secondly, the protection of agents against other agents can be another important issue. The authentication and authorization aspects between communicating agents are similar to that of host-to-agent and agent-to-host processes. We are designing certain mechanisms for this type of protection.

CONCLUSION

The advantages of employing mobile agents can only be manifested if there is a secure and robust system in place.

In this article, the design and implementation of agent authentication and authorization are elaborated. By combining the features of the Java security environment and the Java Cryptographic Extensions, a secure and robust infrastructure is built. PKI is the main technology used in the authentication module. To verify the integrity of the agent, digital signature is used. The receiving party would use the public keys of the relevant parties to verify that all the information on the agent is intact. In the authorization module, the agent is checked regarding its trustworthiness and a suitable user-defined security policy will be recommended based on the level of authentication the agent has passed. The agent will be run under the security manager and the prescribed security policy.

REFERENCES

Chavez, A., & Maes, P. (1998). Kasbah: An agent marketplace for buying and selling goods. *Proceedings of First International Conference on Practical Application of Intelligent Agents and Multi-Agent Technology,* London (pp. 75-90).

Corradi, A., Montanari, R., & Stefanelli, C. (1999). Mobile agents integrity in e-commerce applications. *Proceedings of 19th IEEE International Conference on Distributed Computing Systems* (pp. 59-64).

Dasgupta, P., Narasimhan, N., Moser, L.E., & Melliar-Smith, P.M. (1999). MAgNET: Mobile agents for networked electronic trading. *IEEE Transactions on Knowledge and Data Engineering, 11*(4), 509-525.

Gray, R.S., Kotz, D., Cybenko, G., & Rus, D. (1998). D'Agents: Security in a multiple-language, mobile-agent system. In G. Vigna (Ed.), *Mobile agents and security. Lecture notes in computer science.* Springer-Verlag.

Greenberg, M.S., Byington, J.C., & Harper, D.G. (1998). Mobile agents and security. *IEEE Communications Magazine, 36*(7), 76-85.

Guan, S.U., & Yang, Y. (1999). SAFE: Secure-roaming agent for e-commerce. *Proceedings the 26th International Conference on Computers and Industrial Engineering,* Melbourne, Australia (pp. 33-37).

Guan, S.U., & Zhu, F.M. (2001). Agent fabrication and IS implementation for agent-based electronic commerce. To appear in *Journal of Applied Systems Studies.*

Guan, S.U., Zhu, F.M., & Ko, C.C. (2000). Agent fabrication and authorization in agent-based electronic commerce. *Proceedings of International ICSC Symposium on Multi-Agents and Mobile Agents in Virtual Organizations and E-Commerce,* Wollongong, Australia (pp. 528-534).

Hua, F., & Guan, S.U. (2000). Agent and payment systems in e-commerce. In S.M. Rahman & R.J. Bignall (Eds.), *Internet commerce and software agents: Cases, technologies and opportunities* (pp. 317-330). Hershey, PA: Idea Group Inc.

Jardin, C.A. (1997). *Java electronic commerce sourcebook.* New York: Wiley Computer Publishing.

Karnik, N., & Tripathi, A. (1999). *Security in the Ajanta mobile agent system.* Technical report. Department of Computer Science, University of Minnesota.

Karnik, N.M., & Tripathi A.R. (2001). Security in the Ajanta mobile agent system. *Software Practice and Experience, 31*(4), 301-329.

Lange, D.B., & Oshima, M. (1998). *Programming and deploying JAVA mobile agents with aglets.* Addison-Wesley.

Lee, R.S.T. (2002) iJADE authenticator - An intelligent multiagent based facial authentication system. *International Journal of Pattern Recognition and Artificial Intelligence, 16*(4), 481-500.

Marques, P.J., Silva, L.M., & Silva, J.G. (1999). Security mechanisms for using mobile agents in electronic commerce. *Proceedings of the 18th IEEE Symposium on Reliable Distributed Systems* (pp. 378-383).

Milojicic, D. (1999). Mobile agent applications. *IEEE Concurrency, 7*(3), 80-90.

Ono, K., & Tai, H. (2002). A security scheme for Aglets. *Software Practice and Experience, 32*(6), 497-514.

Oppliger, R. (1999). Security issues related to mobile code and agent-based systems. *Computer Communications, 22*(12), 1165-1170.

Pistoia, M., Reller, D.F., Gupta, D., Nagnur, M., & Ramani, A.K. (1999). *Java 2 network security.* Prentice Hall.

Poh, T.K., & Guan, S.U. (2000). Internet-enabled smart card agent environment and applications. In S.M. Rahman & M. Raisinghani (Eds.), *Electronic commerce: Opportunities and challenges* (pp. 246-260). Hershey, PA: Idea Group Inc.

Rivest, R.L., Shamir, A., & Adleman, L.M. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM.*

Simonds, F. (1996). *Network security: Data and voice communications.* McGraw-Hill.

Tripathi, A., Karnik, N., Ahmed, T. et al. (2002). Design of the Ajanta system for mobile agent programming. *Journal of Systems and Software.*

Tsvetovatyy, M., Mobasher, B., Gini, M., & Wieckowski, Z. (1997). MAGMA: An agent based virtual market for electronic commerce. *Applied Artificial Intelligence, 11*(6), 501-524.

Wang, T., Guan, S.U., & Chan, T.K. (2001). Integrity protection for code-on-demand mobile agents in e-commerce. To appear in *Journal of Systems and Software.*

Wayner, P. (1995). *Agent unleashed: A public domain look at agent technology.* London: Academic Press.

Wong, D., Paciorek, N., & Moore, D. (1999). Java-based mobile agents. *Communications of the ACM, 42*(3), 92-102.

Zhu, F.M., & Guan, S.U. (2001). Towards evolution of software agents in electronic commerce. *Proceedings of the IEEE Congress on Evolutionary Computation 2001,* Seoul, Korea (pp. 1303-1308).

Zhu, F.M., Guan, S.U., & Yang, Y. (2000). SAFER e-commerce: Secure agent fabrication, evolution & roaming for eE-commerce. In S.M. Rahman & R.J. Bignall (Eds.), *Internet commerce and software agents: Cases, technologies and opportunities* (pp. 190-206). Hershey, PA: Idea Group Inc.

KEY TERMS

Agents: A piece of software, which acts to accomplish tasks on behalf of its user.

Authentication: The process of ensuring that an individual is who he or she claims to be.

Authorization: The process of giving access rights to an individual or entity.

Cryptography: The act of protecting data by encoding them, so that they can only be decoded by individuals who possess the key.

Digital Signature: Extra data appended to the message in order to authenticate the identity of the sender, and to ensure that the original content of the message or document that has been sent is unchanged.

Java: A high-level programming language similar to C++ developed by SUN Microsystems.

Private Key: That key (of a user's public-private key pair) known only to the user.

Public Key: The publicly distributed key that if combined with a private key (derived mathematically from the public key), can be used to effectively encrypt messages and digital signatures.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1960-1966, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Mobile Agent-Based Information Systems and Security

Yu Jiao
Oak Ridge National Laboratory, USA

Ali R. Hurson
The Pennsylvania State University, USA

Thomas E. Potok
Oak Ridge National Laboratory, USA

INTRODUCTION

The rapid expansion of information and the high demand for timely data delivery have triggered the development of a large number of wireless information systems that enable users to access data from anywhere at anytime. These applications must face three major challenges: the limited bandwidth of wireless medium, intermittent network connectivity, and the fact that portable devices have limited CPU power, memory, and energy sources. Traditional distributed system design methods, such as the client/server-based computational model, cannot meet the aforementioned challenges very well. In contrast, a relatively new distributed system design paradigm, the mobile agent-based computation model, provides natural solutions to these problems. In this article, we will introduce the concept of mobile agent-based computing, review some examples of existing agent-based information systems, and discuss security issues that are related to them.

BACKGROUND

An *agent* is a computer program that acts autonomously on behalf of a person or organization (Lange & Oshima, 1998). A *mobile agent* is an agent that can move through the heterogeneous network autonomously, migrate from host to host, and interact with other agents (Gray, Kotz, Cybenko, & Rus, 2002). Agent-based distributed application design is gaining prevalence because it provides a single framework that allows a wide range of distributed applications to be implemented easily, efficiently, and robustly.

Mobile agents have many advantages (Lange & Oshima, 1998). We only highlight some of them that are closely related to distributed information system design.

- **Support Disconnected Operation:** Mobile agents can roam the network and fulfill their tasks without the owner's intervention. Thus, the owner only needs to maintain the physical connection during submission and retraction of the agent. This asset makes mobile agents desirable in the mobile computing environment where intermittent network connection is often inevitable.
- **Balance Workload:** By migrating from the mobile device to the core network, the agents can take full advantage of the high bandwidth of the wired portion of the network and the high computation capability of servers/workstations. This feature enables mobile devices with limited resources to support functions beyond their original capability.
- **Reduce Network Traffic:** Mobile agents' migration capability allows them to handle tasks locally instead of passing messages among the data sources. This implies fewer messages and, consequently, reduced chances for loss of messages and the overhead of retransmission.

One should note that the agent-based computation model also has some limitations. For instance, the overhead of mobile agent execution and migration can sometimes overshadow the performance gain obtained by reduced communication costs. In addition, the ability to move and execute code fragments at remote sites could introduce serious security implications.

MAIN THRUST OF THE CHAPTER

Mobile Agent-Based Information Systems

Papastavrou, Samaras, and Pitoura (2000) proposed the DBMA-Aglet Framework for World Wide Web distributed database access. The system uses mobile agents, between the client and the server machine, as a means of providing

database connectivity, processing, and communication. The DBMS-Aglet Multidatabase Framework is an extension of the DBMS-Aglet Framework that can perform parallel execution over multiple databases. In this framework, a coordinator DBMS-Aglet is responsible for creating and dispatching multiple DBMS-Aglets to different data sources. Finally, the coordinator DBMS-Aglet compiles the results and returns it to the client. The authors claimed that the DBMS-Aglet Framework allows the aglet to be portable, light, independent, autonomous, flexible, and robust.

Vlach, Lana, Marek, and Navara (2000) implemented a system called Mobile Database Agent System (MDBAS). The system intends to integrate heterogeneous databases under one virtual global database schema to transparently manage distributed execution. The MDBAS aims to preserve local autonomy and execute distributed transactions using the two-phase commit protocol. Based on the experiences gained in the development of MDBAS, the authors claimed that mobile agent technology will play an important role in the software industry in a short time.

Babaoglu, Meling, and Montresor (2002) proposed a Java-based multi-agent system, called Anthill, which is a framework for peer-to-peer (P2P) application development, deployment, and testing. Anthill adopts the concept of swarm intelligence (Kennedy & Eberhart, 2001), where there is no central coordination of activity and the collection of simple agents of limited capability achieves intelligent collective behavior. Performance evaluation of Anthill was done by using 10,000 queries collected from the Internet. Simulation results confirm that the performance of the system, in terms of the success rate for each search request and the number of hops necessary for the first reply to a search request, improves over time because of the learning and adaptive capabilities of agents.

The VIPAR (Virtual Information Processing Agent Research) system is a multi-agent system that uses agents and ontology to automatically monitor and manage newspaper articles in a manner comparable to humans (Potok, Elmore, Reed, & Shelton, 2003). It includes 13 information agents that manage 13 different newspaper sites. Results show that VIPAR can efficiently handle information gathering, analysis, and summarization tasks that are critical to the Virtual Information Center (VIC) at the U.S. Pacific Command. The deployment of such a system can drastically reduce the cost of these labor and resource intensive processes.

MAMDAS stands for Mobile Agent-based Mobile Data Access Systems (Jiao & Hurson, 2004). Its design aims to alleviate two major difficulties in large-scale mobile data access systems: heterogeneity and mobility of data sources and/or users. Experimental results have shown that under the same underlying multi-database configuration, the mobile agent-based computation mode can achieve better performance and robustness than the traditional client/server-based model. In addition, the authors also pointed out that, from a software engineering point of view, the use of mobile agents can significantly improve modularity and reusability and simplify the management of large complex systems. Therefore, the authors believe that mobile agent-based programming is an excellent solution to distributed information system design.

Spyrou, Samaras, Pitoura, and Evripidou (2004) also share the view that mobile agent technology has great potential in wireless computing applications. The authors proposed a general framework for dynamically configuring applications through the deployment of mobile agents. They believe that the use of mobile agents enhances the applicability of different software models to mobile wireless computing, and it makes applications more light-weight and tolerant to intermittent connectivity. The proposed framework was illustrated through a wireless Web-based data access application. Their simulation results show that, in the wireless environment, for average size transactions, the deployment of mobile agents provides a performance improvement of approximately a factor of 10.

Security Issues

Despite the advantages that mobile agents have demonstrated in building flexible distributed information systems, the success of mobile agent-based systems will depend on the development of robust security defense mechanisms. Due to the mobility and autonomy of mobile agents, designing such security mechanisms is a challenging task.

Within the scope of security, three key issues have been identified for mobile intelligent agent systems (Chess, 1998):

- Protection of the agent against malicious hosts;
- Protection of the host against malicious agents;
- Protection of the network communication.

Ideas from the distributed computing and operating systems can be borrowed to address the second issue (Farmer, Guttman, & Swarup, 1996; Greenberg, Byington, & Harper, 1998), and techniques such as the Secure Socket Layer (SSL) can resolve the third problem. However, protecting agents against malicious hosts is a new and difficult problem that is specific to mobile agents. In the following subsections, we will first summarize the security threats in mobile agent-based systems and then, briefly review solutions that have been proposed in the literature for both host and agent protection.

Security Threats in Mobile Intelligent Agent Systems

Types of attacks are often categorized as follows: damage, denial of service, unauthorized access, harassment, masquerade, and repudiation. These attacks may be launched by

mobile agents against hosts, by hosts against mobile agents, and among mobile agents.

Definition 1: Damage Attack is defined as destruction or subversion of a host's files, configuration, hardware, or of a mobile agent or its mission.

A mobile agent can launch damage attacks to a host by tampering with the services that the host provides, such as destroying the agent execution environment. A host can damage a mobile agent by modifying its code, providing wrong execution results, or simply by erasing the agent.

Definition 2: Denial of Service attack is defined as partially or completely impeding one or more computer services, or a mobile agent's access to some resources or services.

Malicious mobile agents can overload a host by constantly consuming CPU time, memory, or network connections. An overloaded host may fail to provide services to all other mobile agents. From the point of view of protecting the agents, a hostile host may refuse to execute or transport the agent and thus launch a denial of service attack against the mobile agent.

Definition 3: Unauthorized Access attack is defined as illegal or undesired access or removal of data from a host or a mobile agent.

A mobile agent may access and steal private information from the host. It may also use covert channels to transmit data in a hidden way that violates a host's security policy. Similarly, a host may observe or illegally copy secrets carried by mobile agents. Encryption may help alleviate the problem, but unfortunately, some information must be decrypted during execution.

Definition 4: Harassment attack is defined as annoying people with repeated attacks.

Some examples of this type of attack include mobile agents displaying unwanted messages on the host monitor, or hosts tracing a mobile agent and attempting to discover sensitive information about the agent's owner.

Definition 5: Masquerade attack is defined as one party disguising its own identity or claiming a false identity in order to deceive the other party.

A masquerading agent may pose as an authorized agent in order to gain access to services or resources to which it is not entitled. One agent host may masquerade as another in an effort to deceive a mobile agent. An agent may masquerade as another agent in order to lure the victim agent to communicate with it and, therefore, obtain information.

Definition 6: Repudiation attack is defined as one party participating in a transaction or communication, and later claiming that the transaction or communication never took place.

Repudiation may occur because of a misunderstanding or may be an intentional attack. One way to prevent this type of attack is to maintain records/logs that can help resolve any dispute.

Protection of the Host Against Malicious Agents

Techniques that protect the host include: authenticating credentials, access control, code verification, limitation techniques, and audit logging (Greenberg et al., 1998).

a. **Authenticating Credentials:** A mobile agent is digitally signed by one or more parties using one of a number of algorithms, such as a digital signature algorithm. By examining the signatures, a host can discover whether the agent has been tampered within transit. Note that forgery, cryptanalysis, theft of cryptographic keys, or poor implementation can compromise cryptographic techniques.
b. **Access Control:** After a mobile agent's identity is authenticated, a reference monitor is used to restrict the information, system resources, and services that the mobile agent is allowed to access and use. Some mobile agent systems put a restriction on the number of times an object can be accessed. The reference monitor consults a security policy to determine whether or not to grant permission to mobile agents based on their level of authorization as shown by their credentials.
c. **Code Verification:** A code verification program scans the mobile agent's binary image and determines if the mobile agent is a valid program. If illegal instructions are found, the code verification program will not invoke the execution layer. This technique protects hosts by protecting their execution layers that are vulnerable to subversion or sabotage by mobile agents executing within them.
d. **Limitation Techniques:** Time, range, and duplication limits are three limitation techniques that control the persistent survivability of mobile agents. Time limits control the amount of time that a mobile agent system allows a mobile agent to run in an execution layer. Range limits determine the number of destinations or network "hops" that a mobile agent is permitted to visit. Duplication limits regulate the number of times that a mobile agent can be transmitted or be cloned—a

large number of mobile agent clones can swamp the hosts on a network.

e. **Audit Logging:** An audit trail of the mobile agents' activities is kept so that after abuse is detected, the responsible party can be identified and called to account. Based on when the log is examined, auditing techniques can be classified into two categories: passive and active. Passive techniques perform a posteriori analysis and bring security violations to the auditor's attention. Active approaches, on the other hand, perform analysis in real time. Once a violation is detected or suspected, the intrusion detection system alerts the auditor and may take immediate measures for system protection.

Protection of the Agent Against Malicious Hosts

Mobile agent protection techniques can be categorized into two groups: fault tolerance based and encryption based techniques (Greenberg et al., 1998). Techniques based on fault tolerance aim to make mobile agents robust in unpredictable environment, thus protecting them from malfunctioning hosts, intermittent network connectivity, and so forth. Techniques based on encryption hide the mobile code or sensitive information so that it cannot be recognized and thus will be less likely to be stolen or misused.

Fault-tolerant based techniques include replication, persistence, and redirection.

- **Replication:** This technique ensures a mobile agent gets to its destination eventually. When traveling on a network, a mobile agent is replicated at each host for the purpose of fault masking. One obvious drawback of this approach is that it introduces a large amount of overhead in creating and storing the replicas.
- **Persistence:** This technique utilizes the temporary storage (hard disk) on the host to store the mobile agent and its execution state. In the case of host failure, the copy of mobile agent and its execution information persists in the storage. When the host restarts, it can reload the mobile agent from the temporary storage and resume services. Although this improves the mobile agent's tolerance to host failures, it may cause unwanted duplications if the mobile agent is already replaced by the sender assuming the agent is destroyed or forgotten.
- **Redirection:** Redirection refers to the mobile agent's capability of finding available paths around damaged hosts or network to accomplish a mission.

Encryption-based techniques mainly include append-only data logs, trail obscuring, code obfuscation, encrypted data manipulation, and state appraisal functions.

- **Append-Only Data Logs:** This scheme (Karnik & Tripathi, 2001) uses a combination of a digital signature and public key cryptographic primitives to ensure that data can only be appended to the log, and there is no way that data can be removed or modified without the owner being able to detect it. In addition, a host cannot observe data appended by the previous hosts.
- **Trail Obscuring:** This technique constantly modifies the mobile agent's binary image to make it hard to identify by pattern matching. It prevents hosts that are colluding in an attempt to track a specific agent from succeeding.
- **Code Obfuscation:** This is a method to obscure a mobile agent's code to make it hard to reverse engineer. This technique deters theft or subversion, but not destruction.
- **Encrypted Data Manipulation:** This technique encrypts the data carried by a mobile agent in a way that allows it to be manipulated while still encrypted (Sander, 1997). This allows the mobile agent to carry data that cannot be read by a host, but the host can still inspect the mobile agent's code through code verification. This technique was extended later by Lee, Alves-Foss, and Harrison (2004).
- **State Appraisal Functions:** While roaming the network, a mobile agent's execution state may change; a portion of its contents cannot be encrypted and is therefore vulnerable to subversion. State appraisal functions are used to ensure that a mobile agent's unencrypted dynamic data is not tampered with.

FUTURE TRENDS

The success of existing mobile agent-based information systems and the rapidly emerging new agent-based applications have led us to believe that the agent-based paradigm will become a prevalent solution to many distributed applications, especially in the mobile computing environment. We also note that applicability and security always go hand-in-hand. Security in mobile agent-based systems is especially difficult to achieve because mobile agents and hosts have a conflict of interest: Techniques that protect agents may pose potential threats to hosts, and measures that secure hosts may endanger mobile agents. Current mobile agent security solutions are far from being mature. Future research is required to properly address these issues.

CONCLUSION

Extensive research has been conducted to explore the potential of mobile agents in designing large distributed information systems and e-commerce systems. Mobile agents have also been utilized in network management and network intrusion detection tasks. Because mobile agents are distributed and intelligent in nature, it is envisioned that they are suitable for a broad range of future technologies such as semantic Web and sensor network management. This article is intended to provide an overview of applications of mobile agents in distributed information system design and security problems and solutions that are pertinent to them. We believe that if the challenges of agent security can be properly addressed, agent-based applications will proliferate in the near future.

ACKNOWLEDGMENTS

This work in part has been supported by the Office of Naval Research and the National Science Foundation under contracts N000014-02-1-0282 and IIS-0324835, respectively. We would also like to thank Mr. Fred Brenner for his invaluable input on the first draft of this article

REFERENCES

Babaoglu, O., Meling, H., & Montresor, A. (2002). Anthill: A framework for the development of agent-based peer-to-peer systems. In *Proceedings of the 22nd International Conference of Distributed Computing Systems* (pp. 15-22).

Chess, D. M. (1998). Security issues in mobile code systems. In G. Vigna (Ed.), *Mobile agents and security* (pp. 1-14). LNCS 1419, Springer-Verlag.

Farmer, W. M., Guttman, J. D., & Swarup, V. (1996). Security for mobile agents: Issues and requirements. In *Proceedings of the 19th National Info Security Conference (NISSC 96)* (pp. 591-597).

Gray, R. S., Kotz, D., Cybenko, G., & Rus, D. (2002). Mobile agents: Motivations and state-of-the-art systems. In J. Bradshaw (Ed.), *Handbook of agent technology*. AAAI/MIT Press.

Greenberg, M. S., Byington, L. C., & Harper, D. G. (1998). Mobile agents and security. *IEEE Communications Magazine, 36*(7), 76-85.

Jiao, Y., & Hurson, A. R. (2004). Application of mobile agents in mobile data access systems: A prototype. *Journal of Database Management*, *15*(4), 1-24.

Karnik, N., & Tripathi, A. (2001). Security in the Ajanta mobile agent system. *Software Practice & Experience*, *30*(4), 301-329.

Kennedy, J., & Eberhart, R.C. (2001). *Swarm intelligence*. Morgan Kaufmann Publisher.

Lange, D., & Oshima, M. (1998). *Programming and developing Java mobile agents with aglets.* Reading, MA: Addison Wesley Longman.

Lee, H., Alves-Foss, J., & Harrison, S. (2004). The use of encrypted functions for mobile agent security. In *Proceedings of the 37th Hawaii International Conference on System Sciences* (pp. 1-10).

Papastavrou, S., Samaras, G., & Pitoura, E. (2000). Mobile agents for World Wide Web distributed database access. *IEEE Transaction on Knowledge and Data Engineering*, *12*(5), 802-820.

Potok, T. E., Elmore, M., Reed, J., & Sheldon, F. T. (2003). VIPAR: Advanced information agents discovering knowledge in an open and changing environment. In *Proceedings of the 7th World Multiconference on Systemics, Cybernetics and Informatics, Special Session on Agent-Based Computing*, *IX* (pp. 28-33).

Sander, T. (1997). On cryptographic protection of mobile agents. In *Proceedings of the 1997 Workshop on Mobile Agents and Security*.

Spyrou, C., Samaras, G., Pitoura, E., & Evripidou, P. (2004). Mobile agents for wireless computing: The convergence of wireless computational models with mobile-agent technologies. *Mobile Networks and Applications*, *9*(5), 517-528.

Vlach, R., Lana, J., Marek, J., & Navara, D. (2000). MDBAS – A prototype of a multidatabase management system based on mobile agents. In *Proceedings of the 27th Conference on Current Trends in Theory and Practice of Informatics* (pp. 440-449).

KEY TERMS

Agent: A computer program that acts autonomously on behalf of a person or organization.

Damage Attack: Defined as destruction or subversion of a host's files, configuration, or hardware, or of a mobile agent or its mission.

Denial of Service Attack: Defined as partially or completely impeding one or more computer services, or a mobile agent's access to some resources or services.

Harassment Attack: Defined as annoying people with repeated attacks.

Masquerade Attack: Defined as one party disguising its own identity or claiming a false identity in order to deceive the other party.

Mobile Agent: An agent that can move through the heterogeneous network autonomously, migrate from host to host, and interact with other agents.

Repudiation Attack: Defined as one party participating in a transaction or communication, and later claiming that the transaction or communication never took place.

Mobile Commerce and the Evolving Wireless Technologies

Pouwan Lei
University of Bradford, UK

Jia Jia Wang
University of Bradford, UK

INTRODUCTION

The mobile phone industry has experienced an explosive growth in recent years. The emerging markets such as China, India, and Brazil contribute this growth. In China, the number of mobile subscribers has already surpassed the number of fixed landline phone subscribers. In Korea and Japan, there is an explosion of mobile and wireless services. The United States are joining too and there were 207.9 million subscribers in 2005 (CTIA, 2006). Mobile e-commerce (m-commerce) makes business mobility a reality; mobile users could access the Internet at any time, from anywhere with handheld devices or laptop. A 3G enabled smart phone enables you to access a wide range of services anywhere and anytime. For example, you can send and receive e-mail, make cinema and restaurant reservations and pay for them, check real train time, look at digital maps, download music and games, and also browse the Internet. Mobile and wireless services are ranging from mobile communication networks to wireless local area networks. The service provided by mobile communication systems has achieved huge success as mobile and wireless communication technologies are converging at fast speed. We will study mobile and wireless communication in relation to mobile phones. Hence, m-commerce is defined as electronic commerce carried out in handheld devices such as smart phone through mobile and wireless communication network.

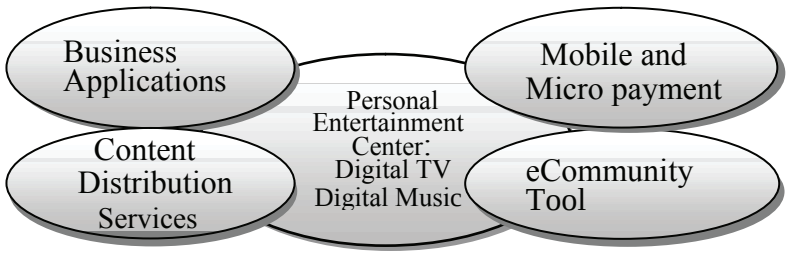
BACKGROUND

E-commerce was once characterized by e-marketplaces, online auction systems that act as the intermediary between buyers and sellers. Now it is evolving toward social network in which users take the control on the contents. Social networks such as YouTube (http://www.youtube.com) enable participants to share life experiences by posting video clips taken by mobile phones especially. The integration creates rich user-generated contents. At the same time, m-commerce also undergoes the evolution too. Many new business models have been established around the use of mobile devices. Mobile devices have the characteristics of portability, low cost, more personalization, GPS (global positioning system), voice etc. The new business models include micro payment and mobile payment, content distribution services, business services, and personal multimedia entertainment center (see Figure 1). Because of their existing customer base, technical expertise, and familiarity with billing, mobile telephone operators are the natural candidates for the provision of mobile and micro payment services. Micro payment involves small purchases such as vending and other items. In other words, the mobile phone is used as an ATM card or debit card. Consumers can pay for purchases at convenience stores or buy train tickets using their mobile phones.

Content distribution services are concerned with real time information, notification (e.g., bank overdraft), and using positioning systems for intelligent distribution of personalized information by location (e.g., selective advertising of locally available services and entertainment). Real-time information such as news, traffic reports, stock prices, and weather forecasts can be distributed to mobile phones via the Internet. The information is personalized to the user's interests. By using a positioning system, users can retrieve local information such as restaurants, traffic reports, and shopping information. Content distribution services with a greater degree of personalization and localization can be effectively provided through a mobile portal. Localization means to supply information relevant to the current location of the user. Users profiles such as past behavior, situation, and location should be taken into account for personalization and localized service provision. Notification can be sent to the mobile device too. Mobile network operators (MNOs) have a number of advantages over the other portal players (Tsalgatidou & Veijalainen, 2000). First, they have an existing customer relationship and can identify the location of the subscriber. Second, they have a billing relationship with the customer while the traditional portal does not. MNOs can act as a trusted third party and play a dominant role in m-commerce applications.

In addition, the mobile phone has become a personal entertainment center. A wide range of entertainment services are available, which consist of online game playing, ring tones download, watching football video clips, live

Figure 1. M-commerce applications



TV broadcasting, music download, and so on. Unsurprisingly, adult mobile service and mobile gambling service are among the fast growing services. Juniper Research Inc. estimates that worldwide revenue from adult mobile content will jump from US$500 million in 2004 to $2.5 billion in 2009 (Korzeniowski, 2005). In the market of mobile games, Juniper Research estimates that annual revenues will have passed the $3 billion mark by the end of 2006. It also acts as a community tool to generate a lot of revenue from SMS (short message service). SMS broadcasting is an ideal communication tool in community.

M-commerce also has a great impact on business applications, especially for companies with remote staff. Extending the existing enterprise resource planning (ERP) systems with mobile functionality will provide remote staff, such as sales personnel, with real-time corporate and management data. Time and location constraints are reduced and the capability of mobile employees is enhanced. Also it makes paperless office a reality so that off-site engineers or salesmen don't need to carry loads of paper such as delivery note to their clients. The logistic related business also benefits from the use of mobile inventory management applications. One interesting application is "rolling inventory" (Varshney & Vetter, 2002). In this case, multiple trucks carry a large amount of inventory while on the move. Whenever a store needs certain items/goods, a nearby truck can be located and just-in-time delivery of goods can be performed. M-commerce offers tremendous potential for businesses to respond quickly in supply chains.

CHALLENGES IN M-COMMERCE

M-commerce has a number of inherent complexities as it embraces many emerging technologies: wireless technologies, handheld devices, software, wireless protocols, and security (Ojanperä, & Prasad, 2001). These technologies have rapid product cycles and quick obsolescence. M-commerce, which is more complex than e-commerce, faces a number of challenges.

Economic Aspect

The delay in 3G mobile network operators (MNO) is in implementing their systems infrastructure. The success of m-commerce in Japan changes the concept of "free" Internet to "paid" Internet. Users are willing to pay for the service. MNOs anticipate a huge profit in taking control of the backbone of m-commerce—the wireless infrastructure. In addition, MNOs also play a dominant position in providing m-commerce applications. This has created an unreasonably high expectation from the services. Big companies in Europe such as Deutsche Telecom, France Télécom, Spain's Telefónica, and the UK's Vodafone spent an estimated US125 billion to US150 billion on 3G licenses (Garber, 2002). Many of them are burdened with high debts. In Europe, 3G was slowly rolled out in 2004 and the number of users was less than 1 million after a huge price cut campaign. As a result, MNOs are reluctant to build the infrastructure of 3G (e.g., O2 in UK).

Social Aspect

With the exception of Korea and Japan, there is a lack of interest in 3G mobile phone. The Western European market has reached saturation point, where mobile possession rate is close to 100% in some countries. In addition, mobile users have "upgrade fatigue" (i.e., they are reluctant to upgrade their mobile phones). In 2002, the mobile phone business pushed very hard on picture messaging, which required new expensive handsets. The response was poor. The mobile revenue mainly comes from the voice calls and SMS messaging. Recently, Apple iPod is a big success. It transforms the music industry and shows that consumers are willing to pay for download music. In 3 GSM world Congress 2005, mobile phone operators are unveiled to team up music companies and mobile phone manufacturers to offer digital music download and play in mobile phone. MNO hopes that they can attract the customers to use their network to download music.

Standards Aspect

The market for handheld devices is quite different from the personal computer (PC) market. For instance, Nokia, the handset manufacturer, not only produces handset hardware but also develops the Symbian software (the operating system (OS) of mobile phone) together with other handset manufacturers such as Motorola and the handset manufacturers have closed relationship with MNOs. The OS standards are under the control of handset manufactures and MNOs.

Technologies Aspect

Security is a major issue. Mobile communications offer users many benefits such as portability, flexibility, and increased productivity. The most significant difference between wired networks and mobile communication is that the airwave is openly exposed to intruders. The intruder eavesdrops on signaling and data connections associated with other users by using a modified mobile phone. In addition, their small size, relatively low cost, and mobility means that they are likely to be lost or stolen. Sensitive information such as the "private-key" is thus vulnerable (Karygiannis & Owens, 2002). A 3G mobile device, when connected to an IP network, is in the "always-on" mode. Both this "always-on" mode and bluetooth's "on" mode make the device susceptible to attack. Moreover, it also provides the opportunity to track a user's activity, which may be a violation of privacy.

Furthermore, handheld devices have many inherent limitations. Technological developments will increase the computing power and storage in handheld devices. However, insufficient battery life and power consumption will impede the potential growth of m-commerce even when 3G is widely available. At present, the battery life is very short (e.g., 2-4 hours for surfing). Fuel cells may be the answer to treble the power of a mobile phone. However, the liquid is inflammable, commercial viability is in doubt. Furthermore, the small screen is another limitation. The screen of a mobile phone is 7cm*5cm, which poses difficulty when surfing the Web. A low power, inexpensive, high-resolution color display would seriously increase the growth of m-commerce.

FUTURE TRENDS

Technology has historically advanced in waves of disruption. It is the same in the telecommunications. The disruptive technologies are Wi-Fi, WiMax, mesh networks, and powerline broadband. WiFi (wireless fidelity) allow users to surf the Internet while moving, are proliferating at astonishing speed on a global scale. Worldwide retail chains like Starbucks and McDonald offer wireless Internet access to their customers. It offers a fast and stable connection; the data rate is several times faster than 3G. The WiFi is an important, new, and disruptive technology to mobile telephone technology and it may be a watershed for all other m-commerce investment by telecom and content providers in the world of the mobile Internet (Lamont, 2001). In making use of this technology, mobile phone manufacturer Nokia and wireless network manufacturer Cisco have been working together closely to produce the Wi-Fi phone. In the future, the tariff of accessing the mobile Internet will be reduced to budget price.

WiMax (worldwide interoperability for microwave access), a low cost wireless broadband connection in wide area network (WAN) starts the rollout. It will offer an opportunity for new mobile phone operators to enter the industry at a small fraction of the cost. As the mobile and wireless technologies are converging, the competition will be keen. The low cost wireless mesh network competes in the market. It is a peer-to-peer wireless communication system that allows two user devices to communicate directly without routing through a central switch. End user devices in a mesh network not only send their data but also act as routers or repeaters, relaying signals for other devices. Powerline broadband has the potential to make an impact too. It is a technology delivering high-speed Internet access into electrical outlets via common electrical grid (Buvat, 2005). The transmission rates are currently better than DSL (digital subscriber line)/cable. It is important for MNOs to embrace those technologies in order to maintain the competition.

CONCLUSION

As the mobile and wireless technologies are evolving rapidly and sophisticated mobile phones becomes affordable, the mobile commerce will become part of our daily life. The mobile Internet is ideal for particular applications and has useful characteristics that offer a range of services and content. The widespread adoption of mobile commerce is fast approaching. In business, the mobile computing is changing the logic of business; businesses have to implement effective strategies to capture and retain increasingly demanding and sophisticated customers. Business needs to think critically about how to integrate mobile Web to wired Web.

REFERENCES

APEC (The Asia-Pacific Economic Cooperation). (2005). *U-Korea*. Retrieved April 23, 2005, from http://www.apec2005.org/apechome/eng/jsp/korea/u_mobile.jsp

Buvat, J. (2005). Two disruptive technologies. *Land Mobile*, *12*(4), 20-21.

CTIA, The Wireless Association. (2006). *Wireless quick facts*. Retrieved August 2, 2006, from http://www.ctia.org/

Garber, L. (2002). Will 3G really be the next big wireless technology? *Computer, IEEE*, *35*(1), 26-32.

Juniper Research. (2006). *Mobile fun & games III* (3rd ed.). White paper. Retrieved August 2, 2006, from www.juniper-research.com/pdfs/whitepaper_mobilegames.pdf

Karygiannis, T., & Owens, L. (2002). *Draft: Wireless network security: 802.11, BluetoothTM and the Handheld Devices*. National Institute of Standards and Technology, Technology Administration U.S. Department of Commerce, Special Publication 800-48.

Korzeniowski, P. (2005). *Adult entertainment on a cell phone near you*. TechNewsWorld. Retrieved August 2, 2006, from http://www.technewsworld.com/story/41140.html

Kwok, B. (2004). Watershed year for mobile phones. Companies and Finance. *South China Morning Post*, January 3.

Lamont, D. (2001). *Conquering the wireless world: The age of m-commerce*. Capstone Publishing Ltd (A Wiley Company).

Ojanperä, T., & Prasad, R. (2001). *WCDMA: Towards IP mobility and mobile Internet*. Artech House Publishers.

Screen Digest. (2005). *Mobile gaming gets its skates on*. Retrieved February 9, 2005, from www.theregister.com/2005/02/09/mobile_gaming_analysis

Tsalgatidou, A., & Veijalainen, J. (2000, September). Mobile electronic commerce: Emerging issues. In *Proceeding of EC-WEB 2000, 1st International Conference on E-Commerce and Web Technologies* (pp. 477-486). London, Greenwich. U.K. Lecture Notes in Computer Science, 1875, Springer Verlag.

Varshney, U., & Vetter, R. (2002). *Mobile commerce: Framework, applications, and networking support*. Mobile networks and applications, *7*, 185-198, 2002, Kluwer Academic Publishers.

KEY TERMS

3.5G: It is based on a technology called, HSDPA (high-speed downlink packet access). It will be upwardly compatible with 3G W-CDMA systems, but will enable more than 10X the peak data rate and more than 6X the capacity of initial 3G systems.

Global System for Mobile Communications (GSM): It is a world standard for digital cellular communications using narrowband TDMA (time division multiple access). It is the standard most commonly used in Europe and Asia, but not in the United States.

i-Mode: It is the packet-based service for mobile phones offered by Japan's leader in wireless technology, NTT DoCoMo. The i-mode protocol uses compact HTML (cHTML) as its markup language instead of WAP's wireless markup language (WML) to provide mobile phone voice service, Internet, and e-mail.

Short Message Service (SMS): It has grown very rapidly and is very popular in Europe. SMS messages are two-way alphanumeric paging messages up to 160 characters that can be sent to and from mobile phones

The Third Generation (3G): It will bring wireless transmission speeds up to 2Mbps, which permits high-quality wireless audio and video. It comprises three primary standards: W-CDMA (wide-band code division multiple access), CDMA2000, and TD-CDMA (time division CDMA).

The WAP: A standard for providing cellular phones, pagers, and other handheld devices with secure access to e-mail and text-based Web pages.

Wi-Fi (Wireless Fidelity): It is a popular term for 802.11b, a wireless local area network (WLAN) specified by the Institute of Electrical and Electronic Engineers (IEEE) and is based on the Ethernet protocol and CSMA/CA (carrier sense multiple access with collision avoidance) for path sharing. Wi-Fi supports a range of about 150 feet and data rates up to 11mbps.

WiMax (worldwide interoperability for microwave access): It is called 802.16 in industry standard, a wireless broadband connection in wide area network(WAN). It offers fast wireless data communications over distance up to about 30 miles.

Wireless Application Protocol (WAP): It is an open, global specification that empowers mobile users with wireless devices to easily access and interact with information and services instantly.

Mobile Commerce Technology

Chung-wei Lee
Auburn University, USA

Wen-Chen Hu
University of North Dakota, USA

Jyh-haw Yeh
Boise State University, USA

INTRODUCTION

With the introduction of the World Wide Web, electronic commerce has revolutionized traditional commerce and boosted sales and exchanges of merchandise and information. Recently, the emergence of wireless and mobile networks has made possible the admission of electronic commerce to a new application and research subject—mobile commerce, which is defined as the exchange or buying and selling of commodities, services, or information on the Internet through the use of mobile handheld devices. With services provided by mobile commerce, consumers may use the microbrowsers on their cellular phones or PDAs to buy tickets, order meals, locate and book local hotel rooms, even write contracts on the move.

In just a few years, mobile commerce has emerged from nowhere to become the hottest new trend in business transactions. NTT DoCoMo's i-mode (2003) is by far the most successful example of mobile commerce. Introduced in February 1999, i-mode has attracted over 36 million subscribers worldwide. With i-mode, cellular phone users can easily access more than 62,000 Internet sites, as well as specialized services such as e-mail, online shopping and banking, ticket reservations, and personalized ringing melodies that can be downloaded for their phones. The i-mode network structure not only provides access to i-mode and i-mode-compatible contents through the Internet, but also provides access through a dedicated leased-line circuit for added security. i-mode users are charged based on the volume of data transmitted, rather than the amount of time spent connected. In Spring 2001, NTT DoCoMo introduced its next-generation mobile system, based on wideband CDMA (W-CDMA), which can support speeds of 384Kbps or faster, allowing users to download videos and other bandwidth-intensive content with its high-speed packet data communications.

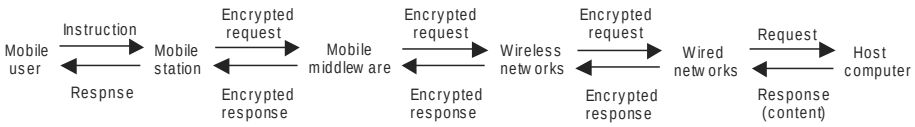
BACKGROUND

A mobile commerce system is very complex because it involves such a wide range of disciplines and technologies. In general, a mobile commerce system can be divided into six components: (1) mobile commerce applications, (2) mobile stations, (3 mobile middleware, (4) wireless networks, (5) wired networks, and (6) host computers.

To explain how these components work together, the following outline gives a brief description of a typical procedure that is initiated by a request submitted by a mobile user:

1. *Mobile commerce applications:* A content provider implements an application by providing two sets of programs: client-side programs, such as a user interface on a microbrowser, and server-side programs, such as database accesses and updating.
2. *Mobile stations:* Mobile stations present user interfaces to the end users, who specify their requests on the interfaces. The mobile stations then relay user requests to the other components and display the processing results later using the interfaces.
3. *Mobile middleware:* The major purpose of mobile middleware is to seamlessly and transparently map Internet contents to mobile stations that support a wide variety of operating systems, markup languages, microbrowsers, and protocols. Most mobile middleware also encrypts the communication in order to provide some level of security for transactions.
4. *Wireless networks:* Mobile commerce is possible mainly because of the availability of wireless networks. User requests are delivered to either the closest wireless access point (in a wireless local area network environment) or a base station (in a cellular network environment).
5. *Wired networks:* This component is optional for a mobile commerce system. However, most computers (servers) usually reside on wired networks such as the Internet, so user requests are routed to these servers

Figure 1. Flowchart of a user request processed in a mobile commerce system



using transport and/or security mechanisms provided by wired networks.

6. *Host computers:* This component is similar to the one used in electronic commerce, which includes three kinds of software. User requests are generally acted upon in this component.

To better illustrate the above procedure, Figure 1 depicts a flowchart showing how a user request is processed by the components in a mobile commerce system (Leavitt, 2000).

MOBILE COMMERCE SYSTEMS

Since each component in a mobile commerce system is large enough to be a research area by itself, only elements in components that are specifically related to mobile commerce are explained in this article. Related research on mobile commerce systems can be found in the article by Varshney, Vetter, and Kalakota (2000).

Mobile Commerce Applications

The applications of electronic commerce are already widespread; mobile commerce applications not only cover these but also include new ones. For example, some tasks that are not feasible for electronic commerce, such as mobile inventory tracking and dispatching, are possible for mobile commerce. Table 1 lists some of the major mobile commerce applications (Gordon & Gebauer, 2001; Sadeh, 2002), along with details of each.

Mobile Stations

A mobile station or a mobile handheld device, such as a personal digital assistant (PDA) or Web-enabled cellular phone, may embrace many of the features of computers, telephone/fax, e-mails, and personal information managers (PIMs), such as calendars and address books, and networking features. A mobile station differs from a PC or notebook due to its limited network bandwidth, limited screen/body size, and mobility features. The limited network bandwidth prevents the display of most multimedia on a microbrowser, while the limited screen/body size restricts the mobile stations of today to either a stylus or keyboard version. Table 2 lists some major mobile station specifications, although several table entries may be incomplete as some of the information is classified as confidential due to business considerations.

Table 1. Major mobile commerce applications

Mobile Category	Major Applications	Clients
Commerce	Mobile transactions and payments	Businesses
Education	Mobile classrooms and labs	Schools and training centers
Enterprise resource planning	Resource management	All
Entertainment	Games/images/music/video downloads and online gaming	Entertainment industry
Health care	Accessing and updating patient records	Hospitals and nursing homes
Inventory tracking and dispatching	Product tracking and dispatching	Delivery services and transportation
Traffic	Global positioning, directions, and traffic advisories	Transportation and auto industries
Travel and ticketing	Travel management	Travel industry and ticket sales

Table 2. Specifications of some major mobile stations

Vendor & Device	Operating System	Processor	Installed RAM/ROM	Input Methods	Key Features
Compaq iPAQ H3870	MS Pocket PC 2002	206 MHz Intel StrongARM 32-bit RISC	64 MB/32 MB	Touchscreen	Wireless email/Internet
Handspring Treo 300	Palm OS 3.5.2H	33 MHz Motorola Dragonball VZ	16 MB/8 MB	Keyboard/ Stylus	CDMA network
Motorola Accompli 009	Wisdom OS 5.0	33 MHz Motorola Dragonball VZ	8 MB/4 MB	Keyboard	GPRS network
Nokia 9290 Communicator	Symbian OS	32-bit ARM9 RISC	16 MB/8 MB	Keyboard	WAP
Nokia 6800	Series 40			Keyboard	Innovative keyboard integration
Palm i705	Palm OS 4.1	33 MHz Motorola Dragonball VZ	8 MB/4 MB	Stylus	Wireless Email/Internet
Samsung SPH-i330	Palm OS 4.1	66MHz Motorola Dragonball Super VZ	16 MB/8 MB	Touchscreen/ Stylus	Color screen
Sony Clie PEG-NR70V	Palm OS 4.1	66 MHz Motorola Dragonball Super VZ	16 MB/8 MB	Keyboard/ Stylus/ Touchscreen	Multimedia
Sony Ericsson T68i			800KB	Keyboard	Multimedia Messaging Service
Toshiba E740	MS Pocket PC 2002	400 MHz Intel PXA250	64 MB/32 MB	Stylus/ Touchscreen	Wireless Internet
Sony Ericsson Z1010			32MB	Keyboard	MP3, MMS, WAP2.0

Mobile Middleware

The term middleware refers to the software layer between the operating system and the distributed applications that interact via the networks. The primary mission of a middleware layer is to hide the underlying networked environment's complexity by insulating applications from explicit protocol handling disjoint memories, data replication, network faults, and parallelism (Geihs, 2001). Mobile middleware translates requests from mobile stations to a host computer and adapts content from the host to the mobile station (Saha, Jamtgaard, & Villasenor, 2001). According to an article in *Eurotechnology* entitled Frequently asked questions about NTT-DoCoMo's i-mode (2000), 60% of the world's wireless Internet users use i-mode, 39% use WAP, and 1% use Palm middleware. Table 3 compares i-mode and WAP, the two major kinds of mobile middleware.

Wireless Networks

Network infrastructure provides essential voice and data communication capability for consumers and vendors in cyberspace. Evolving from electronic commerce (EC) to mobile commerce (MC), it is necessary for a wired network infrastructure, such as the Internet, to be augmented by wireless networks that support mobility for end users. From the perspective of mobile commerce, wireless networks can be categorized into wireless local area networks (WLANs) and wireless cellular networks.

WLAN technologies are suitable for office networks, home networks, personal area networks (PANs), and ad hoc networks. In a one-hop WLAN environment, where an access point (AP) acting as a router or switch is a part of a wired network, mobile devices connect directly to the AP through radio channels. Data packets are relayed by the AP to the other end of a network connection. If no APs are

Table 3. Comparisons of WAP and i-mode

	WAP	i-mode
Developer	WAP Forum	NTT DoCoMo
Function	A protocol	A complete mobile Internet service
Host Language	WML (Wireless Markup Language)	CHTML (Compact HTML)
Major Technology	WAP Gateway	TCP/IP modifications
Key Features	Widely adopted and flexible	Highest number of users and easy to use

available, mobile devices can form a wireless ad hoc network among themselves and exchange data packets or perform business transactions as necessary. Many WLAN products are available on the market. In general, Bluetooth technology supports very limited coverage range and throughput. Thus it is only suitable for applications in personal area networks. In many parts of the world, the IEEE 802.11b (Wi-Fi) system is now the most popular wireless network and is used in offices, homes, and public spaces such as airports, shopping malls, and restaurants. However, many experts predict that with much higher transmission speeds, 802.11g will replace 802.11b in the near future.

Cellular system users can conduct mobile commerce operations through their cellular phones. Under this scenario, a cellular phone connects directly to the closest base station, where communication is relayed to the service site through a radio access network (RAN) and other fixed networks. Originally designed for voice-only communication, cellular systems are evolving from analog to digital, and from circuit-switched to packet-switched networks, in order to accommodate mobile commerce (data) applications. Currently, most of the cellular wireless networks in the world follow 2G or 2.5G standards. However, there is no doubt that, in the near future, 3G systems with quality-of-service (QoS) capability will dominate wireless cellular services. The two main standards for 3G are Wideband CDMA (WCDMA), proposed by Ericsson, and CDMA2000, proposed by Qualcomm.

Host Computers

A host computer processes, produces, and stores all the information for mobile commerce applications. This component is similar to that used in an electronic commerce system because the host computers are usually not aware of differences among the targets, browsers or microbrowsers they serve. It is the application programs that are responsible for apprehending their clients and responding to them accordingly. Most of the mobile commerce application programs reside in this component, except for some client-side programs such as cookies. Usually this component contains three major elements: a Web server, a database server, and application programs and support software.

FUTURE TRENDS

It is estimated that 50 million wireless phone users in the United States will use their handheld devices to authorize payment for premium content and physical goods at some point during the year of 2006. This represents 17% of the projected total population and 26% of all wireless users (The Yankee Group, 2001). Mobile commerce is an effective and convenient way to deliver electronic commerce to consumers from anywhere and at anytime. Realizing the advantages to be gained from mobile commerce, many major companies have begun to offer mobile commerce options for their customers in addition to the electronic commerce they already provide (Over 50% of large U.S. enterprises plan to implement a wireless/mobile solution by 2003, 2001).

However, without secure commercial information exchange and safe electronic financial transactions over mobile networks, neither service providers nor potential customers will trust mobile commerce systems. Mobile security and payment are hence crucial issues for mobile commerce. Security issues span the whole mobile commerce system, from one end to the other, from the top to the bottom network protocol stack, from machines to humans. For example, in WAP, security is provided through the Wireless Transport Layer Security (WTLS) protocol (in WAP 1.0) and IETF standard Transport Layer Security (TLS) protocol (in WAP 2.0). They provide data integrity, privacy, and authentication. One security problem, known as the "WAP Gap" is caused by the inclusion of the WAP gateway in a security session. That is, encrypted messages sent by end systems might temporarily become clear text on the WAP gateway when messages are processed. One solution is to make the WAP gateway resident within the enterprise (server) network (Ashley, Hinton, & Vandenwauver, 2001), where heavyweight security mechanisms can be enforced.

In an IEEE 802.11 WLAN, security is provided by a data link level protocol called Wired Equivalent Privacy (WEP). When it is enabled, each mobile host has a secret key that is shared with the base station. The encryption algorithm used in WEP is a synchronous stream cipher based on RC4. The ciphertext is generated by XORing the plaintext with a RC4 generated keystream. However, recently published literature has discovered methods for breaking this approach (Borisov, Goldberg, & Wagner, 2001; Fluhrer, Martin, & Shamir, 2001; Stubblefield, Ioannidis, & Rubin, 2002). The next version, 802.11i, is expected to have better security.

Payment on mobile commerce systems is another issue. Although the Secure Electronic Transaction (SET) protocol (SET Secure Electronic Transaction Specification, Version 1.0, 1997) is likely to become the global standard in the domain of electronic commerce over the Internet, a WAP client device normally does not have sufficient processing and memory capability to utilize SET software. A "thin" SET wallet approach (Jin, Ren, Feng, & Hua, 2002) has thus been proposed to adapt the SET protocol for WAP clients. Under the "thin" SET wallet model, most of the functionality of current "fat" SET wallets is moved to the wallet server. To support a SET payment, a WAP client installed with only a "thin" wallet securely connects with a wallet server, which communicates with other SET entities. When SET purchase requests arrive from the "thin" wallet, the wallet server takes over the responsibility of routing requests and managing digital keys and certificates.

CONCLUSION

The emerging wireless and mobile networks have extended electronic commerce to another research and application subject: mobile commerce. A mobile commerce system involves a range of disciplines and technologies. This level of complexity makes understanding and constructing a mobile commerce system an arduous task. To facilitate this process, this article divided a mobile commerce system into six components, which can be summarized as follows:

- Mobile commerce applications: Electronic commerce applications are already broad. Mobile commerce applications not only cover the existing applications, but also include new applications, which can be performed at any time and from anywhere by using mobile computing technology.
- Mobile stations: Mobile stations are limited by their small screens, limited memory, limited processing power, and low battery power, and suffer from wireless network transmission problems. Numerous mobile stations, such as PDAs or Web-enabled cellular phones, are available on the market, but most use one of three major operating systems: Palm OS, Microsoft Pocket PC, and Symbian OS. At this moment, Palm OS leads the market, although it faces a serious challenge from Pocket PC.
- Mobile middleware: WAP and i-mode are the two major kinds of mobile middleware. WAP is widely adopted and flexible, while i-mode has the highest number of users and is easy to use. It is difficult to predict which middleware will be the eventual winner in the end; it is more likely that the two will be blended somehow at some point in the future.
- Wireless and wired networks: Wireless communication capability supports mobility for end users in mobile commerce systems. Wireless LANs and cellular networks are major components used to provide radio communication channels so that mobile service is possible. In the WLAN category, the Wi-Fi standard with 11 Mbps throughput dominates the current market. It is expected that standards with much higher transmission speeds, such as 802.11g, will replace Wi-Fi in the near future. Compared to WLANs, cellular systems can provide longer transmission distances and greater radio coverage, but suffer from the drawback of much lower bandwidth (less than 1 Mbps). In the latest trend for cellular systems, 3G standards supporting wireless multimedia and high-bandwidth services are beginning to be deployed. WCDMA and CDMA2000 are likely to dominate the market in the future.
- Host computers: Host computers process and store all the information needed for mobile commerce applications, and most application programs can be found here. They include three major components: Web servers, database servers, and application programs and support software.

An important trend for mobile commerce is enhancing mobile security mechanisms and payment methods. Mobile commerce systems can prosper only if information can be securely exchanged among end systems (consumers and vendors). Security issues (including payment) include data reliability, integrity, confidentiality, and authentication and are usually a crucial part of implementation in wireless protocols/systems. Solutions are updated frequently, due to the lack of a comprehensive wireless security infrastructure and standard. A unified approach has not yet emerged.

REFERENCES

Ashley, P., Hinton, H., & Vandenwauver, M. (2001). Wired versus wireless security: The Internet, WAP and iMode for E-Commerce. In *Proceedings of Annual Computer Security Applications Conferences (ACSAC)*, New Orleans, LA, December 10-14, 2001 (p. 296).

Borisov, N., Goldberg, I., & Wagner, D. (2001). Intercepting mobile communications: The insecurity of 802.11. In *Proceedings of the 7th International Conference on Mobile Computing and Networking*, Rome, Italy, July 16-21, 2001 (pp. 180-189).

Fluhrer, S., Martin, I., & Shamir, A. (2001). Weakness in the key scheduling algorithm of RC4. In *Proceedings of the 8th Annual Workshop on Selected Areas in Cryptography*, Toronto, Ontario, Canada, August 16-17, 2001.

Frequently asked questions about NTT-DoCoMo's i-mode (2000). *Eurotechnology.* Retrieved December 16, 2002, from http://www.eurotechnology.com/imode/faq.html.

Geihs, K. (2001). Middleware challenges ahead. *IEEE computer*, 34(6), 24-31.

Gordon, P. & Gebauer, J. (2001). M-commerce: Revolution + inertia = evolution. *Working Paper 01-WP-1038*, University of California, Berkeley, CA.

i-mode (2003). *NTT-DoCoMo*. Retrieved November 28, 2002 from http://www.nttdocomo.com/.

Jin, L., Ren, S., Feng, L., & Hua, G. Z. (2002). Research on WAP clients supports SET payment protocol. *IEEE Wireless Communications*, 9(1), 90-95.

Leavitt, N. (2000). Will WAP deliver the wireless Internet? *IEEE Computer*, 34(5), 16-20.

Sadeh, N. (2002). *M-commerce: Technologies, services, and business models,* pp. 177-179. New York: John Wiley & Sons.

Saha, S., Jamtgaard, M., & Villasenor, J. (2001). Bringing the wireless Internet to mobile devices. *IEEE Computer*, 34(6), 54-58.

SET Secure Electronic Transaction Specification, Version 1.0 (1997). Retrieved October 11, 2002 from http://www.setco.org/.

Stubblefield, A., Ioannidis, J., & Rubin, A.D. (2002). Using the Fluhrer, Martin, and Shamir attack to break WEP. In *Proceedings of the Network and Distributed Systems Security Symposium*, San Diego, CA, February 6-8, 2002 (pp. 17-22).

Varshney, U., Vetter, R. J., & Kalakota, R. (2000). Mobile commerce: A new frontier. *IEEE Computer*, 33(10), 32-38.

WAP (Wireless Application Protocol) (2003). *Open Mobile Alliance Ltd*. Retrieved November 21, 2002 from http://www.wapforum.org/.

The Yankee Group (2001) Over 50% of large U.S. enterprises plan to implement a wireless/mobile solution by 2003. Retrieved December 10, 2002 from http://www.yankeegroup.com/public/news_releases/news_release_detail.jsp?ID=PressReleases/news_09102002_wmec.htm.

The Yankee Group publishes U.S. mobile commerce forecast (2001). *Reuters*. Retrieved December 16, 2002. from http://about.reuters.com/newsreleases/art_31-10-2001_id765.asp

KEY TERMS

i-mode: the full-color, always-on, and packet-switched Internet service for cellular phone users offered by NTT DoCoMo.

Mobile Commerce: the exchange or buying and selling of commodities, services, or information on the Internet (wired or wireless) through the use of mobile handheld devices.

SET: the Secure Electronic Transaction (SET) protocol is a technical standard designed to provide security for payment transactions among cardholders, merchants, payment gateways, and certification authorities in Internet.

Third Generation (3G): wireless system that can provide fairly high-speed (384 Kbps) packet-switched wide-area wireless Internet access to support multimedia applications.

Wi-Fi: IEEE 802.11b (Wi-Fi) is a wireless local area network standard. It operates in an unlicensed radio frequency band at 2.4 GHz and provides data access at 11 Mbps.

Wired Equivalent Privacy (WEP): a data link-level protocol that provides security for the IEEE 802.11 WLAN standards. The encryption algorithm used in WEP is a stream cipher based on

Wireless Application Protocol (WAP): an open, global specification that allows users with mobile devices to easily access and interact with information and services instantly

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1967-1972, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Mobile Location Services

George M. Giaglis
Athens University of Economics and Business, Greece

INTRODUCTION

The term "mobile era" as a characterization of the 21st century can hardly be considered an exaggeration (Kalakota & Robinson, 2001). Mobile phones are the fastest penetrating technology in the history of mankind, and global mobile phone ownership has surpassed even the ownership of fixed phones. Mobile applications, despite potentially being very different in nature from each other, all share a common characteristic that distinguishes them from their wire-line counterparts: they allow their users to move around while remaining capable of accessing the network and its services. In the mobility era, *location identification* has naturally become a critical attribute, as it opens the door to a world of applications and services that were unthinkable only a few years ago (May, 2001).

The term "mobile location services" (MLS) [or "location-based services (LBS), as they are sometimes also referred to] has been coined to group together applications and services that utilize information related to the geographical position of their users to provide value-adding services to them (Rao & Minakakis, 2003). This article provides a concise introduction to the major types of MLS and also introduces the reader to the most important positioning technologies that render the provision of MLS possible. Finally, the article also introduces a number of issues that are critical for the future of MLS, including privacy protection, regulation, and standardization.

CATEGORIES OF MOBILE LOCATION SERVICES

Mobile networks are quickly becoming ubiquitous. The ability to reach mobile phone users regardless of their location and, even more importantly, the ability to reach mobile phone users *based* on their location has created a new world of exciting and promising applications. While the possibilities for providing innovative MLS are limited only by one's imagination, we will outline the most important categories of such services in this section.

Emergency Management

Perhaps the clearest market application of MLS is the ability to locate an individual who is either unaware of his or her exact location or is not able to reveal it because of an emergency situation (injury, criminal attack, and so on). MLS are even applicable as a means of overcoming one of the most common problems of motorists, namely, the fact that, most often than not, they are unaware of their exact location when their vehicle breaks down. The ability of a mobile user to call for assistance and at the same time automatically reveal his or her exact location to the automotive assistance agency is considered one of the prime motivators for signing up subscribers to MLS (Hargrave, 2000).

Navigation Services

Navigation services are based on mobile users' needs for directions within their current geographical locations. The ability of a mobile network to locate the exact position of a mobile user can be manifested in a series of navigation-based services:

1. By positioning a mobile phone, an operator can let users know exactly where they are as well as give them detailed *directions* about how to get to a desirable destination.
2. Coupled with the ability to monitor traffic conditions, navigation services can be extended to include destination directions that take account of current *traffic conditions* (for example, traffic congestion or a road-blocking accident) and suggest alternative routes to mobile users.
3. The possibility to provide detailed directions to mobile users can be extended to support *indoor routing* as well. For example, users can be assisted in their navigation in hypermarkets, warehouses, exhibitions, and other information-rich environments to locate products, exhibition stands, and other points of interest.
4. Similarly, *group management* applications can be provided to allow mobile users to locate friends, family, coworkers, or other members of a particular group that are within close range and, thus, create *virtual communities* of people with similar interests.

Information Provision

Location-sensitive information services mostly refer to the digital distribution of content to mobile terminal devices based on their location, time specificity, and user behavior. The following types of services can be identified within this category:

1. *Travel services*, such as guided tours (either automated or operator-assisted), notification about nearby places of interest (for example, monuments), transportation assistance, and other services that can be provided to tourists moving around in unfamiliar surroundings.
2. *Mobile yellow pages* that provide a mobile user, upon request, with knowledge regarding nearby facilities.
3. *Infotainment services*, such as information about local events, location-specific multimedia content, and so on.

Advertising and Marketing

Mobile advertising is among the first trial applications of MLS, due to its promising revenue potential and its direct links to mobile-commerce activities. Furthermore, mobile advertising has gained significant attention because of the unique attributes, such as *personalization* (Kalakota & Robinson, 2001), that offer new opportunities to advertisers to place effective and efficient promotions on mobile environments. There are various mechanisms for implementing mobile advertising coupled with MLS. Examples of mobile advertising forms include *mobile banners*, *alerts* (usually dispatched as SMS messages), and *proximity-triggered advertisements*.

Tracking

Tracking services can be equally applicable to the consumer and the corporate markets. As far as consumers are concerned, tracking services can be utilized to monitor the exact whereabouts of, for example, children and elderly people. Similarly, tracking services can be effectively applied in corporate situations as well. One popular example refers to tracking vehicles so that companies know where their fleet and goods are at any time. A similar application allows companies to locate their field personnel (for example, salespeople and repair engineers) so that they are able, for example, to dispatch the nearest engineer and provide their customers with accurate personnel arrival times. Finally, the newfound opportunity to provide accurate product tracking within the supply chain offers new possibilities to mobile supply chain management (m-SCM) applications (Kalakota & Robinson, 2001).

Billing

Location-sensitive billing refers to the ability of a mobile service provider to dynamically charge users of a particular service depending on their location when using or accessing the service. For example, mobile network operators may price calls based on the knowledge of the location of the mobile phone when a call is made. Location-sensitive billing includes the ability to offer reduced call rates to subscribers who use their mobile phone when at their home, thereby allowing mobile operators to compete more effectively with their fixed telephony counterparts.

POSITIONING TECHNOLOGIES

The applications and services that were discussed in the previous section are based on underlying technological capabilities that enable the identification of the location of a mobile device, thereby making the provision of MLS possible. Positioning techniques can be implemented in two ways: *self-positioning* and *remote positioning* (Zeimpekis et al., 2003).

In the first approach (self-positioning), the mobile terminal uses signals, transmitted by the gateways/antennas (which can be either terrestrial or satellite) to calculate its own position. More specifically, the positioning receiver makes the appropriate signal measurements from geographically distributed transmitters and uses these measurements to determine its position. A self-positioning receiver, therefore, "knows" where it is, and applications collocated with the receiver can use this information to make position-based decisions, such as those required for vehicle navigation.

In the case of remote positioning, the mobile terminal can be located by measuring the signals travelling to and from a set of receivers. More specifically, the receivers, which can be installed at one or more locations, measure a signal originating from, or reflecting off, the object to be positioned. These signal measurements are used to determine the length and direction of the individual radio paths, and then the mobile terminal position is computed from geometric relationships.

Self-Positioning Techniques

Global Positioning System (GPS) and Assisted GPS (A-GPS): GPS is the worldwide satellite-based radio navigation system, consisting of 24 satellites, equally spaced in six orbital planes 20,200 kilometres above the Earth, that transmit two specially coded carrier signals: one for civilian use and one for military and government use (Djuknic & Richton, 2001). The system's satellites transmit navigation messages that a GPS receiver uses to determine its position.

GPS receivers process the signals to compute position in three dimensions—latitude, longitude, and altitude—with an accuracy of 10 meters or less. The main advantage of this technique is that GPS is already in use for many years. However, in order to operate properly, GPS receivers need a clear view of the skies and signals from at least three or four (depending on the type of information needed) satellites, requirements that exclude operation in indoor environments. As far as the A-GPS method is concerned, the mobile network or a third-party service provider can assist the handset by directing it to look for specific satellites and also by collecting data from the handset to perform location identification calculations that the handset itself may be unable to perform due to limited processing power. The A-GPS method can be extremely accurate, ranging from 1 to 10 meters (Giaglis et al., 2002).

Indoor Global Positioning System (Indoor GPS): This system focuses on exploiting the advantages of GPS for developing a location-sensing system for indoor environments. It should be noted that the GPS signal does not typically work indoors, because the signal strength is too low to penetrate a building (Chen & Kotz, 2000). Indoor GPS solutions can be applicable to wide space areas where no significant barriers exist. Indoor GPS takes into account the low power consumption and small size requirements of wireless access devices, such as mobile phones and handheld computers. The navigation signal is generated by a number of pseudolites (pseudo-satellites). These are devices that generate a GPS-like navigation signal. The signal is designed to be similar to the GPS signal in order to allow pseudolite-compatible receivers to be built with minimal modifications to existing GPS receivers. As in GPS, at least four pseudolites have to be visible for navigation, unless additional means, such as altitude aiding, are used (Giaglis et al., 2002).

Remote Positioning Techniques

Cell Identification (Cell-ID): The Cell-ID (or *Cell of Origin, COO*) method is the most widely used technique to provide location services and applications in second-generation mobile communication networks. The method relies on the fact that mobile networks can identify the approximate position of a mobile handset by knowing which cell site the device is using at a given time. The main benefit of the technology is that it is already in use today and can be supported by all mobile handsets. However, the accuracy of the method is generally low (in the range of 200 meters in densely covered areas and much lower in rural environments) (Giaglis et al., 2002).

Angle of Arrival (AOA): The basic idea is to steer in space a directional antenna beam until the direction of maximum signal strength is detected. In terrestrial mobile systems, the directivity required to achieve accurate measurements is obtained by means of antenna arrays (Sakagami et al., 1994). Basically, a single measurement produces a straight-line locus from the base station to the mobile phone. Another AOA measurement will yield a second straight line, and the intersection of the two lines gives the position fix for this system.

Time of Arrival (TOA): Positioning information is derived from the absolute time for a wave to travel between a transmitter and a receiver or vice versa. This implies that the receiver knows the exact time of transmission. Alternatively, this approach might involve the measurement of the round-trip time of a signal transmitted from a source to a destination and then echoed back to the source, giving a result twice that of the one-way measurement. This does not imply synchronization between the transmitter and the receiver and is the most common means of measuring propagation time.

Differential Time of Arrival (DTOA): The problem of having precisely synchronized clocks at transmitter and receiver is solved by using several transmitters synchronized to a common time base and measuring the time difference of arrival at the receiver. More specifically, each DTOA measurement defines a hyperbolic locus on which the mobile terminal must lie. The intersection of the hyperbolic loci will define the position of the mobile device.

CRITICAL ISSUES RELATED TO MLS

Further to the business and technological aspects of mobile location services discussed above, a number of other critical factors will also ultimately determine their successful application in mobile communication networks. Such issues include the need to protect sensitive personal information of individuals, as well as the role of regulation and standardization initiatives in establishing the right climate for market rollout and healthy competition.

Privacy Protection

According to Nokia (2001), “of all the challenges facing mobile location service providers, privacy is undoubtedly the biggest single potential barrier to market take-up" (p. 9). For example, mobile advertising based on a user’s location is a sensitive issue and has to be provided only with the explicit consent of the user.

However, even in such a case, the likely exchange of information between third parties (for example, network operators and advertising agencies) may hamper the privacy of user’s personal data. To ensure commercial success of mobile location services, user trust must be ensured. A

clear prerequisite of the trust-building mechanism is that the control over the use of location information is always on the hands of the user, not of the network operator or the service provider.

Regulation and Standardization

The role of regulatory and policy-making bodies is substantially enhanced in the case of mobile location services. It is not surprising that the initial boost to the market has come from such bodies (the US FCC mandate for emergency services) and that the European Commission has had a very active role in the development of the market on the other side of the Atlantic.

Standardization can also be a serious success or failure factor for any new technology, and mobile location services are not an exception to this rule. A number of bodies worldwide are working toward defining commonly accepted standards for the mobile industry, but prior experience has shown that standardization efforts may have a regional, rather than a global, scope. For example, the presence of incompatible standards for second-generation mobile telephony in Europe and the Americas has created considerable problems for users and the industry alike. Worldwide efforts to define universal standards for third-generation systems provide a more optimistic view of the future; however, the danger of incompatibility and technological "islands" remains. To this end, a number of standardization initiatives are underway, sometimes initiated by the industry. For example, Ericsson, Motorola, and Nokia, have joined forces to establish the *Location Interoperability Forum (LIF)*, with the purpose of developing and promoting common and ubiquitous solutions for mobile location services.

The importance of standardization becomes even more evident when we think of what can be termed as *the paradox of mobile location services*. Although these services are by definition local, any given user will most probably need them when in a nonlocal environment. We can envisage tourists outside the familiar surroundings of their local residences relying on MLS to obtain assistance and directions, and companies also utilizing MLS to track their goods in distant lands. To be useful to their users, mobile location services must, therefore, be provided in a location-independent and user-transparent fashion. From a standardization and technological point of view, this requirement poses a difficult problem: *service portability* and *roaming* issues have to be resolved in order for MLS to be compelling to users (UMTS, 2001).

REFERENCES

Chen, G., & Kotz, D. (2000). A survey of context-aware mobile computing research. Dartmouth Computer Science Technical Report TR2000-381.

Djuknic, G. M., & Richton, R. E. (2001). Geolocation and assisted GPS. *IEEE Computer, 34*(2), 123–125.

Giaglis, G. M., Kourouthanasis, P., & Tsamakos, A. (2002). Towards a classification framework for mobile location services. In B. E. Mennecke & T. J. Strader (Eds.), *Mobile commerce: Technology, theory, and applications*. Hershey, PA: Idea Group Publishing.

Giaglis, G.M., Pateli A., Fouskas, K., Kourouthanassis, P., Tsamakos, A. (2002). On the potential use of mobile positioning technologies in indoor environments. In C. Loebbecke, R. T. Wigand, J. Gricar, A. Pucihar, G. Lenart (Eds.) *The Proceedings of the 15th Bled Electronic Commerce Conference—E-Reality: Constructing the E-Economy*, Moderna Organizacija, Kranj, Slovenia, Vol 1, 413-429.

Hargrave, S. (2000). Mobile location services: A report into the state of the market. White paper. Cambridge Positioning Systems.

Kalakota, R., & Robinson, M. (2001). *M-business: The race to mobility*. New York: McGraw-Hill.

May, P. (2001). *Mobile commerce: Opportunities, applications, and technologies of wireless business*. London; New York: Cambridge University Press.

Nokia Corporation (2001) Mobile Location Services, White Paper of Nokia Corporation. Available online at *http://www.nokia.com/pc_files_wb2/mposition_mobile_location_services.pdf]*

Rao, B., & Minakakis, L. (2003). Evolution of mobile location-based services. *Communications of the ACM, 46*(12), 61–65.

Sakagami, S., (1994), Vehicle position estimates by multibeam antennas. In *Multi-path environments, IEEE Transactions on Vehicular Technologies, 43*(4), 902–908.

UMTS Forum. (2001). The UMTS third generation market—Phase II. UMTS Forum Report #13, April. Retrieved from *http://www.umts-forum.org*

Zeimpekis, V., Giaglis, G. M., & Lekakos, G. (2003). A taxonomy of indoor and outdoor positioning techniques for mobile location services. *ACM SIGECOM Exchanges, 3*(4), 19–27.

KEY TERMS

Angle of Arrival (AOA): A positioning technology in which the mobile network sends directional antenna beams

to locate a mobile device at the intersection of the directions of maximum signal strength.

Assisted Global Positioning System (A-GPS): A variation of the *global positioning system (GPS)* in which the mobile network or a third-party service provider assists the mobile handset in determining its geographical position (either by directing it to look for specific satellites or by collecting data from the handset to perform location identification calculations that the handset itself may be unable to perform due to limited processing power).

Cell Identification (Cell-ID): The Cell-ID method is the basic technique to provide location services and applications in second-generation mobile communication networks. The method relies on the fact that mobile networks can identify the approximate position of a mobile handset by knowing which cell site the device is using at a given time.

Differential Time of Arrival (DTOA): A positioning technology in which several transmitters (synchronized to a common time base) are used to measure time differences of arrival at the receiver and, hence, determine the receiver's geographical position.

Global Positioning System (GPS): GPS is the worldwide satellite-based radio navigation system. The system's satellites transmit messages that a receiver uses to determine its own geographical position.

Indoor Global Positioning System (Indoor GPS): A variation of the *global positioning system (GPS)* for use in indoor environments, where the normal GPS signal does not typically work, because the signal strength is too low to penetrate a building. Indoor GPS navigation signals are generated by a number of pseudolites (pseudo-satellites) and are sent to pseudolite-compatible receivers that use the information to determine their own geographical positions.

Location-Based Services (LBS): A synonym for *mobile location services (MLS)* denoting applications that utilize the knowledge of one's geographical position to provide added-value services.

Location Identification: The ability of mobile hosts to determine the geographical location of wireless access devices.

Mobile Location services (MLS): Applications provided over a mobile network that utilize information related to the geographical position of their users to provide added value to them.

Time of Arrival (TOA): A positioning technology where information is derived from the absolute time for a wave to travel between a transmitter and a receiver or vice versa.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1973-1977, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Mobile Positioning Technology

Nikos Deligiannis
University of Patras, Greece

Spiros Louvros
Technological Educational Institute of Messologgi, Greece

Stavros Kotsopoulos
University of Patras, Greece

INTRODUCTION

A radio mobile-position system operates by measuring, processing, and storing physical quantities related to radio signals travelling between a mobile terminal and a set of transceivers, for example, satellites or Base Stations (BSs). Positioning techniques in cellular networks are of great importance for supporting emerging services that require a sufficient, precise estimation of the position of the mobile terminal (MT) associated with a number of given base stations. The ability to support position location within wireless networks provides network operators with valuable services, as well as users with a host of new applications. This includes navigation, location-based services, network management, and security applications. Nowadays in GSM networks, there is no specific algorithm included in the software to locate subscribers. The only possibility to locate a subscriber is accomplished by using *GPS* technology or special firmware.

An innovative development should be the design and implementation of certain location positioning techniques, as *Time of Arrival* (ToA) and *Angle of Arrival* (AoA), over the existing GSM network. Although there are already next generation networks in use nowadays (UMTS, GPRS, WLAN), and there also is extensive research towards the fourth generation cellular networks, GSM seems to be the most popular network so far. GPRS network is a data network over GSM platform and it exists only with GSM architecture. The reason is that GPRS uses the GSM air interface (Radio Network Part) and it only diverts in the core network where it transmits the data packets towards a different switch. UMTS, on the other hand, is a unique network supporting cellular and voice-data applications, and is the evolution of GSM towards IP applications. Although it could be implemented separately from GSM, most of the operators preferred to implement it in a GSM convergence mode towards the core network for eliminating the investment. As a result, in most countries GSM is the major network with full geographical coverage and network location positioning techniques are most implemented in a network environment with a satisfactory number of Base Stations.

BACKGROUND

In 1991, European Telecommunication and Standardization Institute (ETSI) accepted the standards for a new upcoming mobile, fully digital, and cellular communication network (Figure 1).

The purpose of positioning the mobile is to provide location-based services (LBS), including wireless emergency services (Porretta, Nepa, Manara, Giannetti, Dohler, Ben, & Aghvami, 2004). The handset based positioning techniques require that the existing handsets have to be redesigned in order to meet new requirements, while the network based positioning techniques need adjustments only at the Base Stations (BSs) or switching centers. Furthermore, with the first approach, the MT utilizes transmitted signals from the BSs to estimate its own position while with the second approach the BSs measure the transmitted signals from the MT and relay them to a central site for processing.

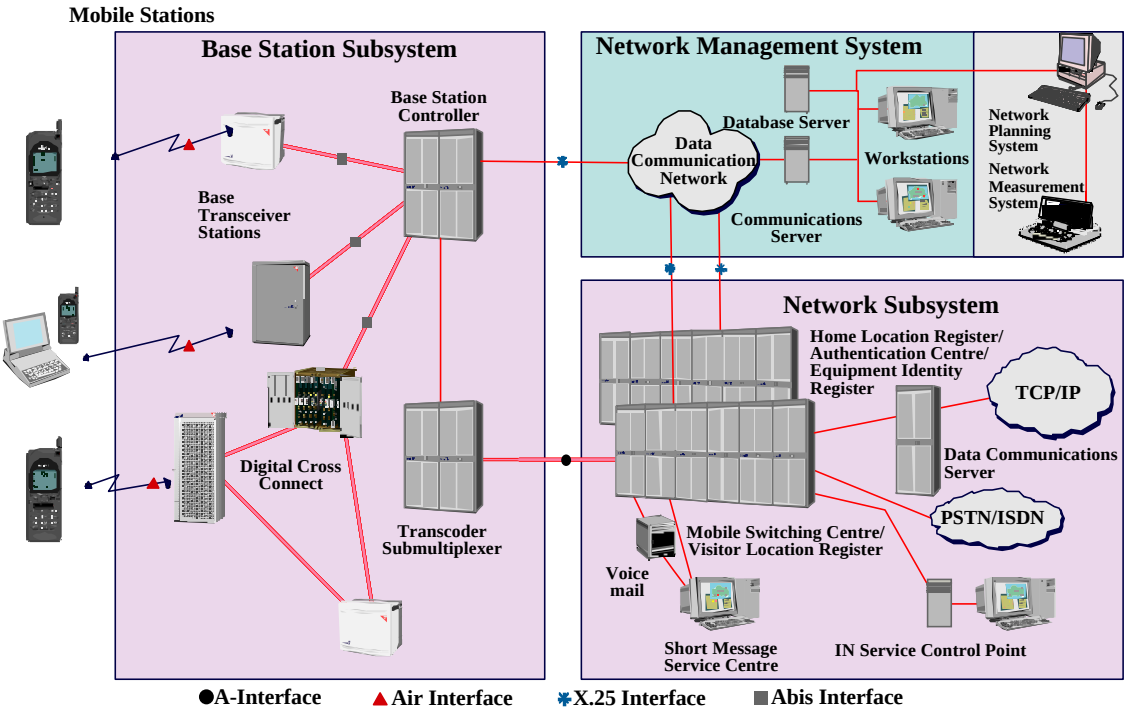
Handset-Based Mobile Positioning Technology

It is referred to as "handset based" because the handset itself is the primary means of positioning the user (Smith, 1991), although the network can be used to provide assistance in acquiring the mobile device and/or making position estimate determinations based on measurement data and handset based position determination algorithms (Kothris, Beach, Allen, & Karlsson, 2001). The representative techniques of handset based positioning technology are:

- **Enhanced Observed Time Difference (E-OTD)**

This technique is also encountered as handset based Time Of Arrivals and specially equipped handsets are required. Enhanced Observed Time Difference (E-OTD) technique operates by instating location receivers called location measurement units (LMU) at several places geographically dispersed in the radio coverage area of a cellular network.

Figure 1. GSM network architecture



- Global Positioning System (GPS)

Unambiguously, the main technology for the implementation of handset based positioning is the Global Positioning System (GPS) which has changed navigation and position forever. It is a universal system consisting of three interlocking segments: the space segment, the user segment, and the control segment. The space segment consists of 24 satellites each in its own orbit 20.000 km above the Earth which means that it takes 12 hours to orbit the Globe.

- Assisted Global Positioning System (AGPS)

GPS suffers position errors from satellite clock, satellite orbit, ephemeris prediction, ionospheric, and tropospheric delays, and so forth. In order to reduce these errors and correct the initial position estimation, additional information can be applied to GPS receivers. A substantial correction method is Differential GPS (DGPS). According to this method, a reference GPS receiver at a proper position is used to send correction data to the quested MT (Zhao, 2002).

Network Based Mobile Positioning Technology

It is called “network based” because the mobile network, in cooperation with network-based position determination equipment (PDE) is used to position the mobile terminal. It is also referred to as “unmodified handset” which means that there are no changes and, thus, additional cost for the subscribers in the mobile device. On the contrary, this technology requires changes in the infrastructure of the GSM elements.

- Fingerprint method

It is an implementation used mainly for indoor application (Pahlavan, Li, & Mäkelä, 2002), which makes use of

the multipath characteristics of the signal received at the Base Station in proportion to the handset position. During the implementation of the method a database, where the fingerprints of the received signals at the BS from each position in the cell area are stored, is created. Therefore, a received signal at the BS is matched with an entry and the BS estimates the location of the MT. The drawback of this method is its demanding implementation. As for each BS cell, extensive and accurate measurements are required for creating and updating the database. This is the reason why this method is used in indoor environments where the database has limited size.

- Doppler location method

It was proposed by N. J. Thomas, D. G. M. Cruickshank and D. I. Laurenson (2001). It requires a single BS and unmodified moving MTs. According to this approach, the position of the MT is calculated by using the Doppler shift related to each scatterer surrounding the BS. The advantages of the method are the implementation simplicity in addition to the signaling overhead, but the main disadvantage is the necessity for moving handsets.

- Cell of Origin (COO)

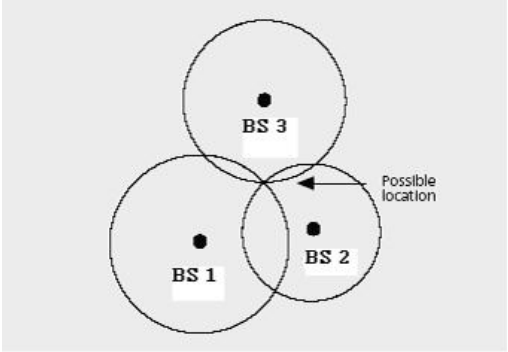
It is the easiest and the most natural technique to offer LBS because all mobile devices support this technology. According to the COO, the network uses the closer (to the handset) BS to identify the cell where the subscriber is. The accuracy depends upon the cell area and it can be up to 150 meters for an urban area. Moreover, depending on the radio architecture, the accuracy is approximately 150m – 5 Km in microcellular design, 500m – 150m in picocellular indoor or outdoor, or 35Km – 150m in macrocellular design.

TIME OF ARRIVAL POSITIONING TECHNIQUE

In the Time Of Arrival (TOA) technique, the location of the MT derives from measuring the time needed for a signal to travel from a number of BSs to the Mobile Terminal (MT). The equation $d = c \cdot t$ provides the distance of the MT from the BSs. Geometrically, the MT lies on a circle centerd at the BSs' location and radius the distance *d*. By using at least three BSs, the position of the MT is given by the intersection point of the three circles (Figure 2).

Due to measurement errors in time estimates, the circles do not intersect at a single point. In that case, location algorithms have been introduced in literature to resolve the problem (Chan & Ho, 1994). Moreover TOA method suffers from *non Line Of Sight* (nLOS) propagation which means that between the BS and the MT exist one or more obstacles.

Figure 2. The TOA method



In that case, the signal does not travel directly from the MT to the BS but it reaches the latter through reflections or diffractions on buildings, cars, other obstacles, and so forth. As a result, the signal takes a longer path in comparison to the direct path in LOS propagation. The typical location error caused by nLOS propagation in GSM networks has been calculated to approximately 400-700m. (Caffery & Gordon, 1998).

Several methods have been introduced to mitigate the location error caused by nLOS propagation. An effective one (Turin, Jewell, & Johnston, 1972), is to change the location algorithm taking into consideration that in nLOS, propagation the measured distance ct_{BS_i} is greater than the real one $c(t_{BS_i} - \Delta t)$ (Δt is the nLOS propagation error) and, therefore, the possible location of the MT lies inside the circle centred in BS's position. For the three BSs TOA

Figure 3. nLos Propagation error in TOA

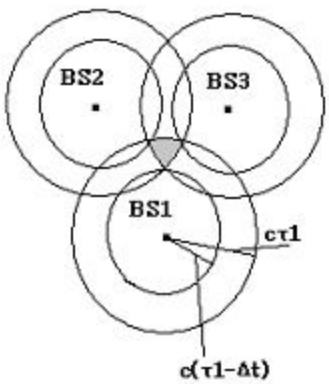
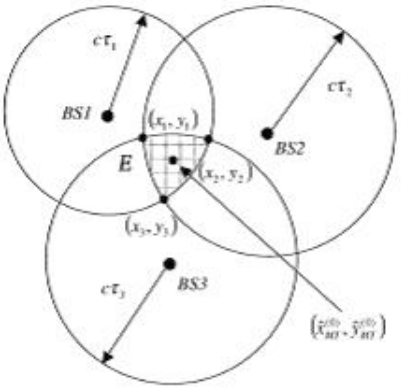


Figure 4. Location area for the three-TOA method



technique, the intersection of the three circles provides an area (feasible area in Figure 3) where the MT can possibly lie. An initial estimation of the MT location can be finding the center of the feasible area. In case of three BSs (Figure 4), supposing that the coordinates of the three points of intersection which set the feasible area E, are (x_1, y_1), (x_2, y_2), and (x_3, y_3) the coordinates of the center of the area E are (Porretta, et al., 2004):

$$(\hat{x}_{MT}^{(0)}, \hat{y}_{MT}^{(0)}) = (\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3})$$

In several cases where the MT is quite close to one BS and the circles do not intersect, the feasible area is estimated as a circle around the adjacent BS. Furthermore, a position nearby that BS is chosen as the initial guess. In order to enhance the initial guess and minimize the location error, a non-linear least square solution can be introduced. According to this, for each BS used in location process, the following function is formed (Porretta, et al., 2004):

$$g_i(x,y) = ct_{BS} - \sqrt{(x - x_{BS_i})^2 + (y - y_{BS_i})^2}$$

The feasible area E can be appointed by the following inequalities:

$$E = \{(x,y) \mid g_i(x,y) \geq 0 \forall i = 1,.....N_{BS}\}$$
$$=> E = \{(x,y) \mid (x - x_{BS_i})^2 - (y - y_{BS_i})^2 \leq (ct_{BS_i})^2 \forall i = 1,.....N_{BS}\}$$

N_{BS} is the number of the BSs.

The next step is to form the following cost function (Porretta, et al., 2004):

$$G(x,y) = \sum_{i=1}^{N_{BS}} a_i g_i(x,y),$$

α_i are weights reflecting the signal strength as received at the i_{th} BS, *(i=1,…..N_{BS})*.

If no information about signal strength is available or not taken into account, it is possible to set $a_i = 1 \forall (i = 1,.....N_{BS})$. The location estimate is finally given by the couple *(x, y)* that minimizes the cost function inside the feasible region.

In order to mitigate even more, the Turin's (1972) cost function from location errors due to nLOS propagation, a weight coefficient ℓ_i is proposed. To implement the weight ℓ_i the network has to meet the following requirements:

- BS antenna arrays in order to measure the angle of the received signal.
- A *Geographical Information System* (GIS) available to the network operator.

Another requirement is the introduction of a *First Obstacle Distance Function*, FODF(θ), which is defined as the Euclidean distance between the BS and the nearest obstacle found along the azimuth direction identified by angle θ of the received signal. FODF derives from GIS. If FODF(θ)<ct (t is the time estimate) then the MT is in nLOS with the BS. If FODF(θ)>ct the BSi is in LOS and then $\ell_i = 1$. If one BS is in nLOS then $\ell_i = \frac{1}{10}$. If two BSs are in nLOS and $FODF_1(\theta)-ct_1 < FODF_2(\theta)-ct_2$ then $\ell_1 = \frac{1}{10}$ and $\ell_2 = \frac{1}{100}$. In that case, the proposed cost function follows:

$$G'(x,y) = \sum_{i=1}^{N_{BS}} \ell_i a_i g_i(x,y)$$

TOA IMPLEMENTATION IN GSM NETWORK: AN ENHANCEMENT TECHNIQUE

GSM standards specify with accuracy the procedures and messages in order to implement specific applications. The purpose of implementing a TOA positioning technique in GSM network is to intervene in the GSM network's infrastructure as little as possible. In the following, a number of interventions are proposed in both busy and *idle mode* of the subscriber's MT in order to implement TOA technique.

Idle Mode

At the idle mode, the implementation begins with a periodic Location Update signaling for the MT (Figure 5). The timing parameter is sent to the MT through Broadcast Control channel (BCCH). In due course, a paging signaling follows in order

Figure 5. GSM method for the calculation of TOA time estimates (Idle mode)

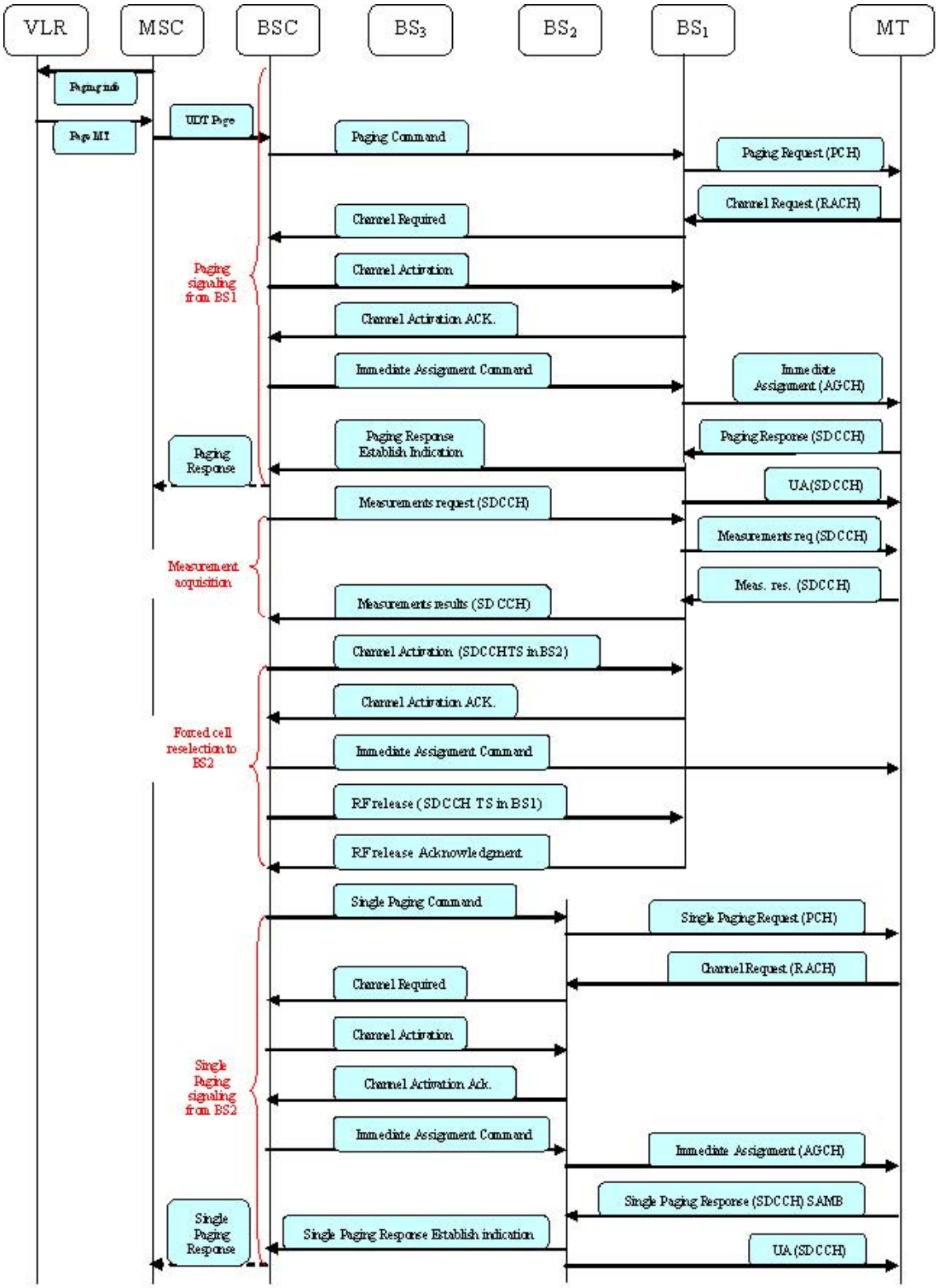
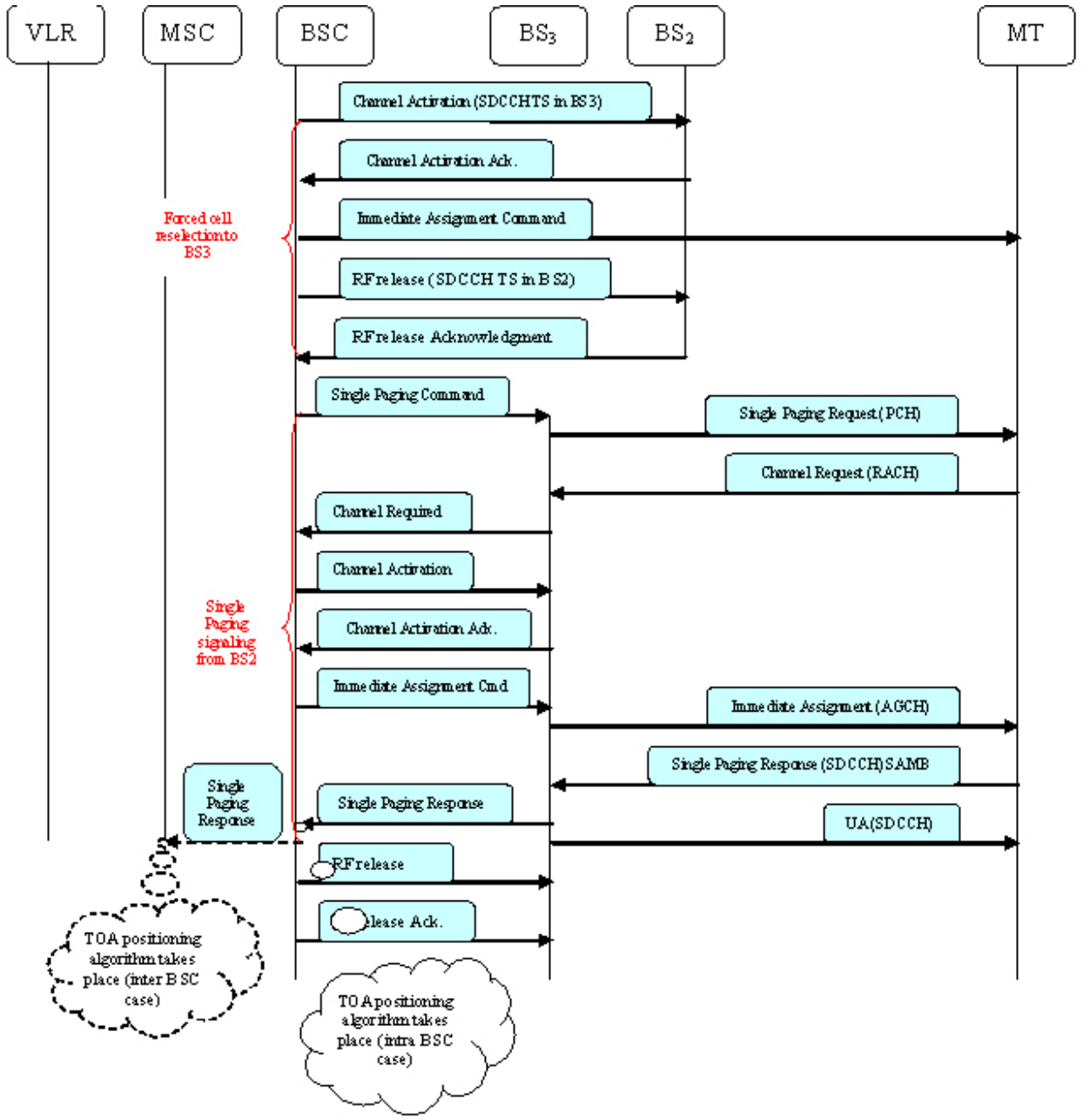
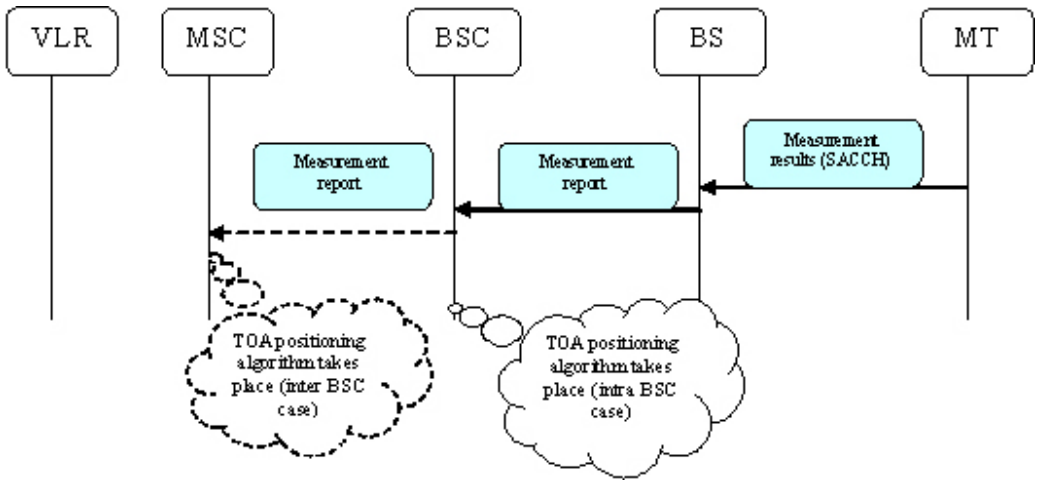


Figure 5. GSM method for the calculation of TOA time estimates (Idle mode) (continued)



to obtain the Timing Advance from the serving BS. In case of Idle Mode condition, the MT keeps for itself measurements about the received signal power from the nearest and the 6 adjacent BSs. Furthermore, to obtain the Timing Advance from the other 2 BSs a new Paging Command message is introduced. After that, the MT is sequentially coordinated to each of the 2 selected BSs by utilizing a proposed Forced Cell Reselection procedure. Finally, the three time estimates are sent back to the BSC where the position estimation of the MT according to TOA algorithm takes place.

Figure 6. GSM method for the calculation of TOA time estimates (Busy mode)



Busy Mode

In the Busy mode scenario (Figure 6) the MT transmits to the serving BS measurements about the received signal power level and the Timing Advance from the serving and the adjacent BSs, two times per second. Every 21 frames in the air interface, a TimeSlot (TS), corresponds to idle mode where the measurements take place in the MT while a second corresponds to SACCH where the measurements are sent. As a result, for a list of seven BSs (six adjacent plus one serving), a time space of: $7 \times 21\, frames \times 0.577\, {ms}/{frame} = 84.82 ms$ is required. This can be translated to a location error of 1.16m for a mean value of the subscriber's velocity.

ANGLE OF ARRIVAL POSITIONING TECHNIQUE

In AOA –or Direction Of Arrivals (DOA) in some literature- the MT position is calculated by using the triangulation technique. According to AOA method, the angles of arrivals of a signal from the MT at a pair -or more- BSs are measured by using antenna arrays. The position of the MT is defined by the intersection of at least two directional lines of bearing (Figure 7).

A linear antenna array consists of a set of similar dipoles, having the same direction and transpiring by a current of the same pattern but differences in the amplitude or the phase. By changing the number or distances between the dipoles or the amplitude and the phase of the current transpiring each one of them, the appropriate radiation pattern in proportion to the application, can be succeeded. Therefore, the received signal from the MT is translated to a current pattern (phase, amplitude) at the antenna array and the BS is able to compute the signal's AOA.

Location errors occur in AOA technique by reason of nLOS propagation and multipath. Due to nLOS propagation, the reflected signal received at BS antenna array has different AOA than the direction of the MT. Moreover, even in LOS propagation, multipath, which means scattered signals near and around the BS, would still alter the measured AOA. Because of measuring limitations of the devices, the higher the distance between MT and BS, the more the precision of the method decreases.

Due to nLOS propagation and multipath, it is wiser in location estimation using the AOA method to utilize more than two BSs. In Figure 8, the three BSs approach is shown. An

Figure 7. The AOA method

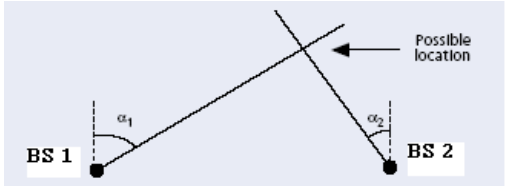
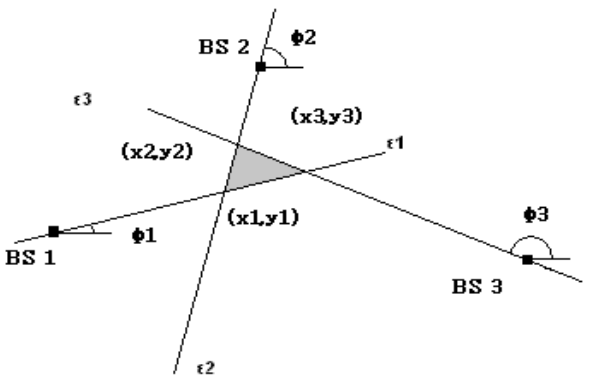


Figure 8. Three BTS AOA intersection area



initial suggestion, introduced by the writers, for the solution to the positioning problem could be the following:

$$(\hat{x}_{MT}^{(0)}, \hat{y}_{MT}^{(0)}) = (\frac{a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3}{a_1 + a_2 + a_3}, \frac{a_1 \cdot y_1 + a_2 \cdot y_2 + a_3 \cdot y_3}{a_1 + a_2 + a_3})$$

where a_i are weights related to signal strength received at the BS_i. If no information about signal strength is provided, it is possible to set $a_i = 1$.

In addition, the writers introduce a further, more precise algorithm that is based on the minimization of a cost function for AOA. To be more specific, as the MT's location lies on a beeline determined by the AOA of the signal received at the BS and BSs' coordinates (Figure 8), the following function can be proposed:

$$f_i(x, y) = (y - y_{BS_i}) - \tan \mathrm{f}_i \cdot (x - x_{BS_i})$$

Where (x_{BS_i}, y_{BS_i}) are the coordinates of the BS_i and f_i the AOA of the signal received at the BS_i from the MT. The coordinates of the MT are those that minimize the following quadratic form (cost function):

$$F(x, y) = \sum_{i=1}^{N_{BS}} a_i \cdot l_i \cdot f_i^2(x, y)$$

where, l_i are factors reflecting whether the MT is in LOS with BS_i (l_i=1) or not (l_i=0). Using a Geographical Interface System (GIS) and the TOA of the signals received, BS can determine whether there is LOS propagation or not and, thus, give the appropriate value to l_i factor. If none, BS is in LOS with the MT then we give the value l_i=0 only to one BS with the greater |ct-FODF(φ)|. The reason for this criterion is that the greater the difference |ct-FODF(φ)| is, the more possible it is that the signal reaches the BS after many reflections. Suggesting that the antenna array in BS uses differential reception and more than one BS are in LOS with the MT then the location error can be reduced to 0m as the MT coordinates zero the cost function.

In Turin, et al. (1972) a single BS AOA technique using the TOA ate the MPCs impinging at the BS is introduced.

For macrocells, where the BSs are usually above roof level, or at least above the terrain, the scatter signals are located close to the BS. As a result, the received signals AOA follow a narrow spread distribution. On the other hand, for microcells where the BSs are placed below roof level and, thus, surrounded by local scatterers following a large spread distribution. To sum up, AOA technique is more appropriate for macrocells than for microcells.

FUTURE TRENDS

Combined techniques TOA-AOA are interesting since they should provide location accuracy in different environments (urban, suburban, city center, suburbs), and the network applications would be independent of the environment the subscriber is moving. There is also another technique, the *Power of Arrival* (POA) (Hata & Nagatsu, 1980), which is extremely useful in environments of city centers and generally in microcellular and picocellular networks. According to this technique, the signal strength received at the BS from the MT is measured and, by the use of a known mathematical propagation model, the path attenuation loss is determined. Several propagation models have been introduced such as the Free Space Loss model, the Plane Earth model EGLI model, LEE model OKUMURA model, ext. In all propagation models, path loss attenuation is computed by using the distance between BS and MT and the height of the MT's antenna.

Since the measured signal strength is translated to a distance estimate, the MT's position lies on a circle centered at the BS. As in TOA method, the MT's location derives from the intersection of more than one circle. Errors in POA method are caused by multipath fading and shadowing phenomena. Degradation of the signal strength received at the BS can be as high as 30-40 dB on the order of a half of the wavelength, due to multipath. A solution to this apart from low mobility MTs, could be the utilization of the average strength of a set of measurements. In case of shadowing, error can be reduced by using premeasured signal strength contours centered at the BSs . However, this solution suggests a constant physical topography and needs that contours are mapped out for each BS.

According to POA, a future profitable solution should be a combination of TOA-AOA-POA in a unique algorithm

to predict position in a GSM network and, moreover, the implementation of this new algorithm with new or modified messages in the existing software configuration.

CONCLUSION

GSM network is a worldwide standard for cellular applications. Although new cellular networks have been developed, GSM is still the first choice for subscribers due to its simplicity, and for operators due to its cost effectiveness in implementation and development. As subscribers increase in GSM networks and ask for more precise applications dependent in location accuracy, the cost of implementation elimination is of great importance for the operators. The network location solution seems to be most preferred since it only demands software solutions. The implementation of TOA and AOA algorithms in the standard messages and signaling of GSM network is very important since it leaves unaffected the main GSM network architecture and it only requires a new software release to include the new procedures and messages.

REFERENCES

Caffery, J., & Gordon, L. (1998). Overview of Radiolocation in CDMA Cellular Systems. *IEEE Communications Magazine*. 120-126.

Chan, Y., & Ho, K. (1994). A Simple and Efficient Estimator for Hyperbolic Location. *IEEE Trans. Signal Processing*, 1905-1915.

Hata, M., & Nagatsu, T. (1980). Mobile Location Using Signal Strength Measurements in a Cellular System. *IEEE Transactions on Vehicular Technology*. VT-29, 245-251.

Kothris, D., Beach, M., Allen, B., & Karlsson, P. (2001). Performance Assessment of Terrestrial and Satellite Based Position Location Systems. *Proceedings of 2nd International Conference on 3G Mobile Communication Technologies*, 211–215.

Pahlavan, K., Li, X., & Mäkelä, J.P. (2002). Indoor Geolocation Science and Technology. *IEEE Communication Magazine*, 40, 112–118.

Porretta, M., Nepa, P., Manara, G., Giannetti, F., Dohler, M., Ben, A., Aghvami, A.H., (2004). A Novel Single Base Station Location Technique for Microcellular Wireless Networks: Description and Validation by a Deterministic Propagation Model. *IEEE Transactions on Vehicular Technology*. 53(5), 1502-1514.

Smith, A. (1991). Passive Location of Mobile Cellular Telephone Terminals. *Proceedings of IEEE 1st International Carnahan Conference on Security Technology*. 221-225.

Thomas, N.J., Cruickshank, D.G.M, & Laurenson, D.I. (2001). Calculation of Mobile Location Using Scatterer Information. *IEE Electronic Letters*. 37(19), 1193–1194.

Turin, G., Jewell, W., & Johnston, T. (1972). Simulation of Urban Vehicle-Monitoring Systems. *IEEE Transactions on Vehicular Technology*. VT-21, 9-16.

Zhao, Y.L. (2002)., Standardization of Mobile Phone Positioning for 3G Systems. *IEEE Communications Magazine*, 87-94.

KEY TERMS

Base Station: The module that produces the cellular coverage through electromagnetic radiation.

Base Station Controller: The switch that controls all the Radio resource management in the core network part.

Busy Mode: The subscriber's mode when the mobile is activated and a dedicated traffic channel is reserved and used for the transmission of voice data.

Cell: The geographical area covered by electromagnetic radiation from a transmitter in the GSM frequency band.

Idle Mode: The subscriber's mode when the mobile is activated and it listens to the network.

Line Of Site (LOS): The direct physical path between the mobile terminal and the antennae of the Base Station.

Location Area Code: The code that specifies uniquely a cell in a large geographical area.

Microcell: The cell with a coverage area distance of 1Km to 5 Km.

Mobile Switching Centre: The switch that controls all the mobility management in the core network part.

nLOS: Non Line of Site. The indirect path between the mobile terminal and the antennae due to reflections and scattering.

Picocell: The cell with a coverage area distance of 100m to 1 Km.

Subscriber Identity Module (SIM): The portable database that is carries inside the mobile station.

Transceiver: The hardware element of the Base Station that produces the electromagnetic coverage in a cell.

Mobile Spatial Interaction and Mediated Social Navigation

Mark Bilandzic
Technische Universität München, Germany

Marcus Foth
Queensland University of Technology, Australia

INTRODUCTION

The increasing ubiquity of location and context-aware mobile devices and applications, geographic information systems (GIS) and sophisticated 3D representations of the physical world accessible by lay users is enabling more people to use and manipulate information relevant to their current surroundings (Scharl & Tochtermann, 2007). The relationship between users, their current geographic location and their devices are summarised by the term "mobile spatial interaction" (MSI), and stands for the emerging opportunities and affordances that location sensitive and Internet capable devices provide to its users. The first major academic event which coined the term in its current usage was a workshop on MSI (see http://msi.ftw.at/) at the CHI 2007 (Fröhlich et al., 2007).

Mobile spatial interaction is grounded in a number of technologies that recently started to converge. First, the development of mobile networks and mobile Internet technologies enables people to request and exchange specific information from anywhere at anytime. Using their handheld devices people can, for example, check the latest news, request recent stock exchange values or communicate via mobile instant messaging. The second enabler is global positioning technology. Mobile devices with integrated Global Positioning System (GPS) receivers—soon to be joined by the Russian Global Navigation Satellite System (GLONASS) and the European Galileo system—are aware of their current latitude and longitude coordinates and can use this data as value added information for context-aware services, that is, mobile applications that refer to information relevant to the current location of the user. A possible use scenario for such an information request would be, for example, "find all clubs and pubs in a radius of 500 meters from my current position." The focus of this work is to enrich the opportunities given by such location aware services with selected Web 2.0 design paradigms (Beer & Burrows, 2007; Kolbitsch & Maurer, 2006) toward mobile social networking services that are bound to specific physical places. User participation, folksonomy and geotagging are three design methods that have become popular in Web 2.0 community-platforms and proven to be effective information management tools for various domains (Casey & Savastinuk, 2007; Courtney, 2007; Macgregor & McCulloch, 2006). Applying such a design approach for a mobile information system creates a new experience of collaboration between mobile users, a step toward what Jaokar refers to as the Mobile Web 2.0 (Jaokar & Fish, 2006), that is, a chance for mediated social navigation in physical spaces.

BACKGROUND

Applications based on mobile spatial interaction can be classified into four different categories (Fröhlich et al., 2007): 1) Systems that facilitate navigation and wayfinding in geographic places: This category is, for example, represented by car navigation systems that assist the driver, for example, with interactive maps, arrows or spoken instructions providing directions to the address of destination (Baus, Krüger, & Wahlster, 2002; Kray, Elting, Laakso, & Coors, 2003); 2) Mobile augmented reality applications such as the head-mounted display (HMD) of virtual information added to objects in the physical world, (Bruce, Piekarski, Hepworth, Gunther, & Demczuk, 1998; Piekarski & Bruce, 2002); 3) Applications creating; or 4) providing access to information that is attached to physical places or objects: For such applications, geotagging, a method to attach latitude and longitude identifiers, enables information resources such as text, pictures or videos to be put into a specific geographic context (Torniai, Battle, & Cayzer, 2005).

In categories 3 and 4, which represent the fields relevant to our work, most of the previous studies focus on techniques that allow people to create or access locative information and share their personal stories, thoughts, experiences and knowledge about specific places. Lancaster University's GUIDE project, for example, is an electronic tourist guide that provides users with context-aware information, depending on their profile, interests and location (Cheverst, Davies, Mitchell, Friday, & Efstratiou, 2000). On the other hand, GeoNotes (Espinoza et al., 2001) and Urban Tapestries (West, 2005) allow mobile users to not only read but

also create spatially contextualised content. They can attach virtual sticky notes to particular latitude and longitude coordinates. Equipped with Wi-Fi-enabled personal digital assistants (PDA), users can then see other users' notes that were left behind in their current immediate surroundings. E-graffiti, a context-aware application evaluated on a college campus, detects each participating student's location on the campus and displays notes that were left behind by other students (Burrell & Gay, 2002). Just-for-Us (Kjeldskov & Paay, 2005) helps a group of friends in a city to identify an appropriate place to meet depending on their individual current locations, and the George Square project (Brown et al., 2005) focuses on location, photography and voice sharing functions to let on-site and off-site users collaboratively explore a city sight.

Besides the various use scenarios, the applications primarily differ in the interaction design of specific features (Tungare, Burbey, & Perez-Quinones, 2006), for example, access virtual post-its from remote places (Espinoza et al., 2001; West, 2005) vs. in-situ access (Burrell & Gay, 2002; Rohs, 2005), push (Espinoza et al., 2001; Kjeldskov & Paay, 2005) vs. pull services, expiration dates of the messages or private vs. public messaging (Burrell & Gay, 2002; Espinoza et al., 2001). However, not much work has yet been carried out on studying different interaction techniques between the information provider and information consumer of geographically contextualised content. The focus of our work is on evaluating direct and indirect interaction methods, such as phone calls, text messages and whiteboard messages that can be described and retrieved via folksonomy tags.

MOBILE SPATIAL INTERACTION AS AN ENABLER FOR MEDIATED SOCIAL NAVIGATION IN PHYSICAL SPACES

Our physical world holds certain characteristics that enable us to interpret what other people have done, how they behaved and where they have travelled. Sometimes, we can see traces on physical objects that provide hints about people's actions in the past. Footprints on the ground left by previous walkers can show us the right way through a forest, or in a library, for example, dog-eared books with well-thumbed pages might be worthwhile reading, as they indicate the popularity of the text. The phenomenon of people making decisions about their actions based on what other people have done in the past or what other people have recommended doing, forms part of our everyday social navigation. The concept of social navigation describes the "moving towards a cluster of other people, or selecting objects because others have been examining them" (Dourish & Chalmers, 1994).

Social navigation in its classic sense is often restricted to visible traces that were intentionally or unintentionally left behind by earlier navigators, and indicate a former interaction between them and an object in the physical world. While some previous work use social navigation as a design concept to enhance navigation and wayfinding in virtual spaces such as Web browsing or online shopping (Dieberger, 1995, 1997; Dieberger, Dourish, Höök, Resnick, & Wexelblat, 2000; Dourish & Chalmers, 1994; Erickson & Kellogg, 2000; Forsberg, Höök, & Svensson, 1998; Höök, Benyon, & Munro, 2003; Svensson, 2002; Svensson, Höök, & Cöster, 2005), a different approach is to leverage multimedia and virtual information spaces to enhance social navigation in the physical world, which we refer to as "mediated social navigation." The technical infrastructure comprising mobile Internet and global positioning systems paves the way for mediated social navigation using mobile phones. They enable users to add multimedia content to physical places or objects and overlay the real world with a virtual information space that can then be requested by mobile users (Burrell & Gay, 2002; Jaokar & Fish, 2006), and more specifically, create a mediated social environment. Such applications not only provide hints about what somebody has done at a specific place, but he or she can document this with text, photos, audio and video recordings. Such an infrastructure based on user generated content makes use of mediated social navigation to enable mobile users to effectively exchange local knowledge (Foth, Odendaal, & Hearn, 2007).

In our study, we explored the appropriateness of principles that guide the design of a mobile phone application to support social navigation in physical environments (Bilandzic & Foth, 2007a). Targeting the specific domain of public inner-city places, we have designed "CityFlocks," a mobile system enabling urban residents to leave digital annotations with ratings, recommendations or comments on any place or physical object in the city. Thus, CityFlocks turns residents into in-situ amateur journalists for the benefit of visitors or other residents who have questions or need navigational aid related to any place in the city. Furthermore, it provides two interaction alternatives to let people collectively share information about places in their city or neighbourhoods, one following a direct and the other an indirect social navigation approach (Dieberger, 2003; Svensson, 2002). The direct social navigation feature enables information seekers to set up a direct voice link with a local resident who has agreed to voluntarily provide local information to visitors and other residents. The indirect approach produces a dynamic list of virtual, location-based messages, authored and rated by local residents that provide information about the respective place. In order to retrieve a virtual message or an appropriate voice-link partner, CityFlocks provides a built-in search function based on the folksonomy concept. Folksonomy is a user-generated taxonomy, initially applied to categorise and retrieve Web content such as Web pages or photographs, using open-ended labels called tags (Vander Wal, 2007). For every entry users submit, they can attach a number of

tags to describe the comment and place they are submitting. Other CityFlocks users can then search for such tags to find the recommendations and places they are looking for. The underlying database is designed to allow users to request tags that other people have used to describe similar places. In doing so, one can use the community's collective intelligence to find related places. Similar to Amazon's (http://amazon.com) recommendation feature ("customers who bought this item also bought this item"), CityFlocks identifies correlations between the user generated tags to propose places that are related to one's initial search request.

FUTURE TRENDS

There are two major research fields regarding future MSI-applications: On the one hand enhanced positioning methods and technologies need to be further developed. New approaches embrace a seamless mix of various positioning technologies, such as GPS, Wifi-Fingerprinting (Taheri, Singh, & Agu, 2004) and Radio Frequency Identification (RFID) to ensure precise location awareness where a single technology on its own would fail, for example, GPS in indoor locations or Wi-Fi fingerprinting in rural areas.

The second research field deals with usability issues and the creation of appropriate interaction techniques for the various types and use scenarios of MSI applications. In our CityFlocks project, for example, we propose an innovative interaction type for end-user information systems (Bilandzic & Foth, 2007b) by combining Web 2.0 techniques such as geotagging, folksonomy or user-generated content with context-aware features and direct voice link capabilities on mobile phones. The outcomes provide valuable input into the process of designing future community-driven, mobile information systems. This study continues and expands our work in the area of urban informatics (Foth, 2006, 2008, in press; Foth & Hearn, 2007; Klaebe, Foth, Burgess, & Bilandzic, 2007). Only if designers manage to create intuitive and easy-to-learn interfaces, MSI applications might be adopted and used by the broad mass of users. However, there are some other critical factors such as privacy and social acceptance issues where further studies need to be carried out.

CONCLUSION

Originally, social navigation was restricted to visible interaction histories that were naturally left behind and thus relied on earlier physical interaction between people and the respective object. People interpret these hints as a message, recommendation, warning or just a note telling them something about the type of interaction the previous navigator had with the object. With a clever mix of modern mobile information and communication technology and a set of Web 2.0 technologies, social navigation methods can be enhanced in physical spaces (Höök, 2003). There is an emerging trend that takes advantage of the network connectivity of mobile phone users to create a mediated social environment where people who are interested in particular geographic locations can exchange information, personal opinions and experiences with the respective place. This would, for example, enable visitors of a new city to access the knowledge and experiences from local residents about inner-city facilities. This mind-shift in designing mobile services toward a high engagement of individuals has great potential to enhance peoples' experience when navigating physical spaces (Höök, 2003; Jaokar & Fish, 2006). Turning mobile phone users into in-situ journalist who can upload location-based ratings, comments and recommendations to a shared community platform will eventually form a huge social knowledge repository decentralising control over information about local services.

The idea of the proposed service targets a community-driven urban information service and is meant to provide an infrastructure to let residents become authors of information regarding their own neighbourhoods and make them available for interested people in the city, for example, visitors and tourists. This design approach proposes a new type of mobile information systems, providing the benefits of mediated social navigation to MSI applications. It leverages Web 2.0 techniques toward an effective collaboration between mobile users in order to enhance social navigation in physical spaces.

REFERENCES

Baus, J., Krüger, A., & Wahlster, W. (2002). A resource-adaptive mobile navigation system. In *Paper presented at the International Conference on Intelligent User Interfaces*, San Francisco, CA, USA.

Beer, D., & Burrows, R. (2007). Sociology and, of and in Web 2.0: Some initial considerations. *Sociological Research Online, 12*(5).

Bilandzic, M., & Foth, M. (2007a). CityFlocks: Designing social navigation for urban mobile information systems. In *Paper presented at the ACM SIGCHI Designing Interactive Systems (DIS) Conference*.

Bilandzic, M., & Foth, M. (2007b). Transferring Web 2.0 paradigms to a mobile system for social navigation in public inner-city places. In *Paper presented at the Towards a Social Science of Web 2.0 Conference*.

Brown, B., Chalmers, M., Bell, M., Hall, M., MacColl, I., & Rudman, P. (2005, September 18-22). Sharing the square:

Collaborative leisure in the city streets. In *Paper presented at the Proceedings of the Ninth European Conference on Computer-Supported Cooperative Work*, Paris, France.

Bruce, T., Piekarski, W., Hepworth, D., Gunther, B., & Demczuk, V. (1998). A wearable computer system with augmented reality to support terrestrial navigation. In *Paper presented at the 2nd IEEE International Symposium on Wearable Computers*.

Burrell, J., & Gay, G. K. (2002). E-graffiti: Evaluating real-world use of a context-aware system. *Interacting with Computers, 14*(4), 301-312.

Casey, M. E., & Savastinuk, L. C. (2007). *Library 2.0: The librarian's guide to participatory library service*. Medford, NJ: Information Today.

Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou, C. (2000, April). *Developing a context-aware electronic tourist guide: Some issues and experiences.* In Paper presented at CHI 2000, The Netherlands.

Courtney, N. (2007). *Library 2.0 and beyond: Innovative technologies and tomorrow's user*. Westport, CT: Libraries Unlimited.

Dieberger, A. (1995). Providing spatial navigation for the World Wide Web. In A. U. Frank & W. Kuhn (Eds.), *Spatial Information Theory: Proceedings of Cosit '95* (pp. 93-106). Semmering, Austria: Springer-Verlag.

Dieberger, A. (1997). Supporting social navigation on the World Wide Web. *International Journal of Human-Computer Studies, special issue on innovative applications of the Web, 46*, 805-825.

Dieberger, A. (2003). Social connotations of space in the design for virtual communities and social navigation. In K. Höök, D. Benyon, & A. J. Munro (Eds.), *Designing information spaces: The social navigation approach* (pp. 293-313). London: Springer-Verlag.

Dieberger, A., Dourish, P., Höök, K., Resnick, P., & Wexelblat, A. (2000). Social navigation: Techniques for building more usable systems. *Interactions, 7*(6), n/a.

Dourish, P., & Chalmers, M. (1994). Running out of space: Models of information navigation. In *Paper presented at the HCI'94*.

Erickson, T., & Kellogg, W. A. (2000). Social translucence: An approach to designing systems that support social processes. *ACM Transactions on Computer-Human Interaction, 7*(1), 59-83.

Espinoza, F., Persson, P., Sandin, A., Nyström, H., Cacciatore, E., & Bylund, M. (2001). GeoNotes: Social and navigational aspects of location-based information systems. In *Paper presented at the Ubicomp 2001: Ubiquitous Computing, International Conference*.

Forsberg, M., Höök, K., & Svensson, M. (1998). Design principles for social navigation tools. In *Paper presented at the UI4All*, Stockholm, Sweden.

Foth, M. (2006). Facilitating social networking in inner-city neighborhoods. *IEEE Computer, 39*(9), 44-50.

Foth, M. (2008, in press). *Urban informatics: Community integration and implementation.* Hershey, PA: Information Science Reference, IGI Global.

Foth, M., & Hearn, G. (2007). Networked individualism of urban residents: Discovering the communicative ecology in inner-city apartment complexes. *Information, Communication & Society, 10*(5), 749-772.

Foth, M., Odendaal, N., & Hearn, G. (2007). *The view from everywhere: Towards an epistemology for urbanites.* In Paper presented at the 4th International Conference on Intellectual Capital, Knowledge Management and Organisational Learning (ICICKM), Cape Town, South Africa.

Fröhlich, P., Simon, R., Baillie, L., Roberts, J. L., Murry-Smith, R., Jones, M., et al. (2007). In *Proceedings of the Workshop on Mobile Spatial Interaction*, San Jose, CA, USA.

Höök, K. (2003). Social navigation: From the Web to the mobile. In G. Szwillus & J. Ziegler (Eds.), *Mensch & Computer 2003: Interaktion und Bewegung* (pp. 17-20). Stuttgart, Germany.

Höök, K., Benyon, D., & Munro, A. J. (2003). *Designing information Spaces: The social navigation approach*. London, New York: Springer-Verlag.

Jaokar, A., & Fish, T. (2006). *Mobile Web 2.0: The innovator's guide to developing and marketing next generation wireless mobile applications*. London: Futuretext.

Kjeldskov, J., & Paay, J. (2005). Just-for-us: A context-aware mobile information system facilitating sociality. In Proceedings of the *ACM International 7th International Conference on Human Computer Interaction with Mobile Devices & Services,* table of contents, (Vol. 111, pp. 23-30).

Klaebe, H., Foth, M., Burgess, J., & Bilandzic, M. (2007). Digital storytelling and history lines: Community engagement in a master-planned development. In *Paper presented at the 13th International Conference on Virtual Systems and Multimedia (VSMM'07),* Brisbane, Australia.

Kolbitsch, J., & Maurer, H. (2006). The transformation of the Web: How emerging communities shape the information we consume. *Journal of Universal Computer Science, 12*(2), 187-213.

Kray, C., Elting, C., Laakso, K., & Coors, V. (2003). Presenting route instructions on mobile devices. In *Paper presented at the International Conference on Intelligent User Interfaces*.

Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. *Library Review, 55*(5).

Piekarski, W., & Bruce, T. (2002). ARQuake: The outdoor augmented reality gaming system. *Communications of the ACM, 45*(1), 36-38.

Polanyi, M. (1966). *The tacit dimension*.

Rohs, M. (2005). Real-world interaction with camera phones. *Ubiquitous computing systems* (pp. 74-89). Berlin/Heidelberg: Springer-Verlag.

Scharl, A., & Tochtermann, K. (Eds.). (2007). *The geospatial Web: How geo-browsers, social software and the Web 2.0 are shaping the network society.* Heidelberg, Germany: Springer-Verlag.

Svensson, M. (2002). *Defining, designing and evaluating social navigation.* Stockholm University.

Svensson, M., Höök, K., & Cöster, R. (2005). Designing and evaluating Kalas: A social navigation system for food recipes. *Computer-Human Interaction, 12*(3), 374-400.

Taheri, A., Singh, A., & Agu, E. (2004). Location fingerprinting on infrastructure 802.11 wireless local area networks. In *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks*, (pp. 676–683).

Torniai, C., Battle, S., & Cayzer, S. (2005). *Sharing, discovering and browsing geotagged pictures on the Web*.

Tungare, M., Burbey, I., & Perez-Quinones. (2006). Evaluation of a location-linked notes system. In *Paper presented at the 44th ACM Southeast Regional Conference,* Melbourne, FL.

Vander Wal, T. (2007). *Folksonomy coinage and definition*. Retrieved May 31, 2008, from http://vanderwal.net/folksonomy.html

West, N. (2005). Urban tapestries: Experimental ethnography, technological identities and place. *Cultural Snapshot, 10*.

KEY TERMS

Mobile Social Navigation: The process of guiding activities aimed at determining our position and planning and following a specific route based on what other people have done or what other people have recommended doing, using mobile devices. First introduced by Dourish and Chalmers (1994), they describe social navigation as "moving toward a cluster of other people, or selecting objects because others have been examining them."

Local Folksonomies: In the context of the Web 2.0 discussion, a folksonomy (sometimes also known as a "tag cloud") is a user-generated taxonomy made up of key terms that describe online content. By assigning these freestyle keywords or so-called "tags," the semantics of various information resources can be described in a more flexible, decentralised, collaborative and participatory way than fixed categories allow for. The term has been coined by Thomas Vander Wal.

Geotagging: An approach which adds latitude and longitude identifiers as metadata to online content. It enables people to embed their information resources such as text, pictures or videos in a specific spatial and semantic context to augment the physical world with virtual information. Such a mediated social environment can help people navigate physical spaces by using location aware mobile devices.

Mobile Web 2.0: The suite of systems and mobile devices which either run existing Web 2.0 applications or re-appropriate according characteristics (tagging, user participation, mash-ups, personalisation, recommendations, social networking, collective intelligence, etc.) for the specific context of mobile use and mobile devices.

Context-aware Mobile Devices and Applications: Applications that react or provide information based on the user's context. The context is defined by a vector of selected data, for example, representing his/her current geographic location, emotional state, physical conditions (e.g. light, noise) or other variables that describe the user's current situation. Devices that run context-aware applications are usually equipped with sensors to perceive the environment or make assumptions via intelligent algorithms.

Local Knowledge: Knowledge, or even knowing, is the justified belief that something is true. Knowledge is thus different from opinion. Local knowledge refers to facts and information acquired by a person which are relevant to a specific locale or have been elicited from a place-based context. It can also include specific skills or experiences made in a particular location. In this regard, local knowledge can be tacitly held, that is, knowledge we draw upon to perform and act but we may not be able to easily and explicitly articulate it: "We can know things, and important things, that we cannot tell" (Polanyi, 1966).

Global Positioning System (GPS): A set of earth orbit satellites transmit microwave signals which can be received with so-called GPS-receivers. By comparing the timestamps of signals from different satellites, GPS-receivers can determine their geographic location. Each position can be described by a set of latitude and longitude coordinates.

Mobile Technology Usage in Business Relationships

Jari Salo
University of Oulu, Finland

INTRODUCTION

Business relationships have been studied for decades (Wilkinson, 2001). However, the literature has been criticized of the lack of focus on information technology (IT) usage within business relationships (Reid & Plank, 2000). As managers have started to employ digital tools such as the Internet, intranets, and extranets, buyer-seller relationship scholars have realized the need to focus on IT deployment within relationships. There is a growing body of research that focuses on the different types of technologies being employed such as electronic data interchange (EDI) (Naudé, Holland, & Sudbury, 2000), Internet-based EDI (Angeles, 2000), and extranet (Vlosky, Fontenot, & Blalock, 2000) and their influence on business relationships. Nevertheless, mobile technology usage within business relationships is a nascent field of scientific inquiry.

Besides buyer-seller relationship literature, mobile commerce (MC) (conducting commercial activities via mobile networks) literature also noticeably lacks academic research on business usage of mobile technology (Okazaki, 2005; Scornavacca, Barnes, & Huff, 2005). By combining these indications for further research from the buyer-seller relationship and MC fields it can be argued that there is a clear call for research in this area. Hence, I aim to bridge some aspects of the identified research gap. The research gap is filled in by discussing bonding within buyer-seller relationships to illustrate how mobile technologies create a novel bond in business relationships. It is acknowledged that some research on the adoption of mobile technology in the business context exists (see e.g., Kadyté, 2005).

The paper is organized as follows: First, a brief discussion of the background of business relationships, mobile technologies, and bonding is provided. Then, I highlight how mobile technologies are used within relationships with a case study. After that, future trends in this pertinent area are presented. The paper finishes with a concluding discussion.

BACKGROUND

Basically, it can be stated that both the popular as well as academic press has regularly indicated that the number of relationships that exist between buyers and sellers has decreased, but the amount of trade contracted within existing relationships has simultaneously increased. The fact remains that in many cases, it is not profitable to play dozens or even hundreds of competing suppliers or customers off against each other, but working directly with a few of them within a business relationship is profitable for all parties. This is because, as the number of possible partners increases so do the transaction costs. Thus, it is evident that existing business relationships are a vital area for research.

Within the business relationships domain there are multiple and overlapping fields of inquiry (Ritter & Gemünden, 2003) that have provided specific frameworks applicable to different types of problems. Here, I use the bonding discussion as it provides a means to evaluate changes occurring in business relationships. A bond can be seen as a building block for a relationship that is created through interaction between business parties. Academic literature to date has identified 10 bonds that are pertinent in business relationships: technical, time, knowledge, legal, economic, geographic, social, cultural, ideological, and psychological (Wendelin, 2004). Bonds have an important role in value creation and destruction in business relationships.

Social bonds were the starting point of studies focusing on bonding (McCall, 1970). Before a business relationship is built through business exchanges, there are many distances between the two interacting companies. Johanson and Wiedersheim-Paul (1975) identified social, cultural, technological, and time-related distances. For example, social distance measures the extent to which the actors are unfamiliar with each other's ways of thinking and working. Bonding is seen to reduce the distances. This paper focuses on technical bonds.

Technical bonds play a crucial role when business parties are interacting. For example, if company A produces mobile phones and company B is a supplier of components, over time company A and B will usually create interfacing processes in which, for example R&D teams can meet to plan how new products can be produced in the most effective way. Hence it might be the case that company B adjusts its production so that it is more suitable to company A or buy some machinery specifically to deliver the new subassembly to company A. This type of adaptation and mutual planning of the manufacturing process within the business relationship can be seen as one type of technical bonding that has a crucial role in the development of business relationships.

Technical bonds usually refer to connections in the manufacturing process (Wendelin, 2004); however, an exception to this view is provided by Perry, Cavaye, and Coote (2002). IT-based bonds are considered to be a subset of technological bonds. Bonding can have two opposing impacts depending on the context. It can either have a positive impact on a business relationship as it may enhance interfacing processes—this happens with mobile technologies in the business relationship of the Alpha-Zeta case study—or it can have a negative impact, as it may hinder cooperation with other parties. This can happen if a company has intensive bonding with directly competing companies. In the contract-based software development business for mobile phone manufacturers this may be the case.

Still, it can be stated that not all business relationships that managers engage their companies in are valuable and effective enough. Luckily, the current IT field provides many new applications that can be deployed interorganizationally to make existing business relationships even more valuable and profitable. One of these tools is mobile technology, more fashionably called the mobile solution. These mobile technologies provide a wide array of opportunities to enhance interorganizational processes. Here, I focus on wireless local area network (WLAN)-based infrastructure and personal digital assistant (PDA)-based mobile solutions.

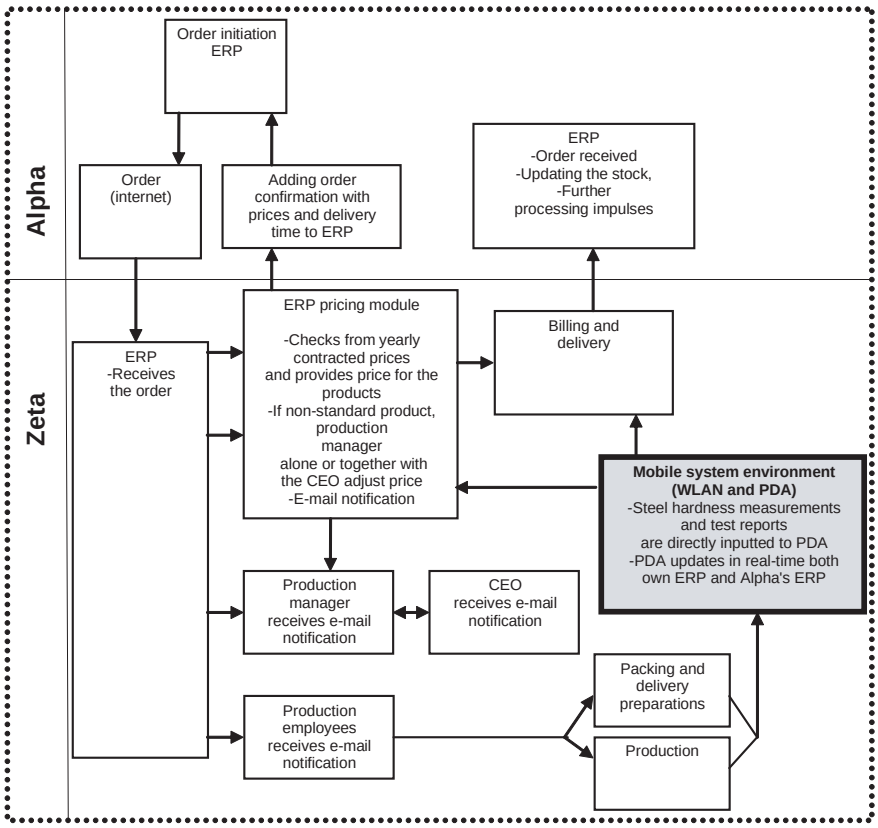
Currently, mobile technologies enable individual business people to check e-mails, place orders, and log in to company networks from the road (Aungst & Wilson, 2005). Some of the mobile devices are connected to Internet networks, such as wireless application protocol (WAP) phones, while others are unconnected, such as PDAs that are subscribed to services like Avantgo and Vindigo, which are unconnected services. More specifically, in a business context MC can be deployed internally or interorganizationally. Within an organization MC can be used to enhance selling activities in the form of sales force automation (SFA) (see Aungst & Wilson, 2005). Interorganizationally, MC can be used to mobilize customer relationship management (Sinisalo, Salo, Karjaluoto, & Leppäniemi, 2006) or it can be deployed with the help of WLAN or Wi-fi and smart devices (such as PDA, hybrid phones, Blackberries). WLAN deployed uses 802.11b standard, and it is connected to a local area network (LAN). MC and mobile technologies have been studied mainly from the consumers' point of view in recent years. The use of MC in relationships is studied even less and there are no studies, as far as I know, that employ empirical research. Based on this, it is essential to examine how mobile technologies can be employed to enhance order-delivery processes in a relationship.

DEPLOYMENT OF MOBILE TECHNOLOGIES IN THE ALPHA-ZETA BUSINESS RELATIONSHIP

In essence, methodological choices are guided by the purpose of research. The basic aim of this paper is to deepen and expand existing knowledge regarding business relationships and how mobile technology may be used to enhance those relationships. This goal calls for a case study method (see e.g., Yin, 1989), as a case study provides detailed and rich information of one focal phenomenon. More specifically, the case is a relationship that is composed of two companies interacting with each other. The perspective of both parties needs to be studied to be sure of the value of the findings (John & Reve, 1982). Based on thorough secondary information sources, I selected a business relationship from the steel industry. This industry was chosen as the empirical context because computerization has a long history in the industry and new technologies play a central role (Chaffey, 2004). Research on the usage of mobile technologies and solutions in the steel industry context is scarce. Therefore, it can be argued that the steel industry context is worth studying, especially when companies are employing novel mobile solutions to enhance their business. In this study, the steel mill is called Alpha and the steel workshop is labeled Zeta. To be more precise, the narrow context of the Alpha-Zeta business relationship is steel processing, that is, the hardening and marketing of steel plates and components. I used semistructured in-depth interviews that were transcribed and analyzed accordingly. I also used several additional sources of information such as documents, minutes of meetings, industry reports, and plant visits to validate research results (see e.g., Patton, 1987). The data triangulation employed here increases the validity and reliability of the research (Eisenhardt, 1989). The identities of the case study companies and the informants are not revealed for confidentially reasons.

The Alpha-Zeta business relationship is now 6 years old and is based on and developed from a previous 40-year business relationship that still exists between Alpha and Beta. Zeta was established to serve Alpha's needs by hardening steel qualities from 5mm up to 60mm. When the relationship was initiated, all activities related to the order-delivery process were manual, that is, handled physically. In brief, Alpha has relatively high IT skills and uses enterprise resource planning (ERP), customer relationship management (CRM), and supply chain management (SCM) systems as well as traditional office systems. At first Zeta had limited IT skills but managed to acquire and implement a small scale, first generation ERP system. Until 2004, almost the whole order-delivery process was digitized with the help of the ERP systems that were integrated over the Internet. From here on, I focus on the development and usage of mobile technologies.

Figure 1. Activities within the Alpha-Zeta business relationship after mobile solution adoption



The most recent technological addition to the relationship is the mobile technology discussed previously. The main idea is to use the mobile technology to speed up inventory control, test-report transmissions, and other nonroutine communications. In a technological sense the solution is based on WLAN infrastructure and employs handheld computers such as PDAs as wireless devices. The technology employed was acquired from a local mobile system developer that had provided some IT systems for Zeta earlier. Basically, the mobile solution renders paper-based and manual strength measurement reports obsolete and transforms these into a digital form that is easier to process for both parties in the relationship. Before this new technology was adopted, reports were first conducted by writing the required information down on paper, inputting this information to a system, printing the report, and then sending it to Alpha's administration who filed it. Now the mobile solution enables information to be input directly to a PDA which updates Zeta's ERP system and provides e-mail notification to Alpha about the new reports, which are essential for the documentation of the steel solutions delivered to customers. Furthermore, Alpha can now receive information about Zeta's hardening capacity, which is vital for generating new sales. Previously, a hardening capacity check was manual and information received by Alpha's sales department was usually too outdated to reliably act upon and thus information needed to be rechecked. Today, Alpha's employees, with access codes to Zeta's system, can retrieve information from the real-time database updated by Zeta's employees and the mobile solution. Figure 1 depicts the order-delivery process after implementing the mobile technology.

Figure 1 illustrates how the mobile technology employed is aligned with other technologies and how it has eased processes in the relationship. The benefits, in the form of faster information delivery which provides the means to sell more steel solutions, are enjoyed by both parties. Furthermore, the adoption of these novel technologies and the adaptations made a clear signal to Alpha that Zeta was willing to help in every way possible to make the business relationship more effective and efficient. Similarly, Alpha signaled to Zeta through increasing orders that the relationship was worth continuing.

FUTURE TRENDS

It is clear from current literature that future research will attempt even more than before to bridge the research void that exists concerning business usage of mobile solutions (Aungst & Wilson, 2005). Besides focusing on business usage of mobile solutions, specific emphasis will also be placed on the deployment of mobile solutions in business relationships (Salo, 2006). This is because all business organizations have a large number of relationships that need to be handled effectively. Thus, it is clear that IT-based systems such as mobile solutions will not render personal communication obsolete but those forms of communication will coexist and enhance communication in business and within business relationships.

Moreover, a wide array of different types of mobile solutions is applicable to business usage and business relationship usage. These systems might include Radio Frequency Identification RFID), mobile CRM (management of relationships via mobile gadgets), or mobile accessible SCM or ERP. It is evident that this topic merits further research from both MC and marketing viewpoints. In addition to academic studies more practical research is also needed to fully understand the nature of this emerging phenomenon.

CONCLUSION

This article provides insights into ways mobile technologies can be used to enhance existing processes in relationships. In addition to looking at MC literature this paper also provides ideas from a business relationship perspective. Basically, the paper identifies the emergence of mobile, technology-based bonds as a subset of technological bonds. Interestingly, technological bonds are usually perceived as manufacturing-related bonds but it can be seen here that mobile, technology-based bonds are based on communication via mobile devices. Mobile, technology-based solutions helped both companies to speed up inventory control, test-report transmissions, and other nonroutine communications. It also created more new sales possibilities as excess capacity became easier to identify and act upon. In general, this paper tackled some of the possible mobile technology benefits available from the adoption of technology into a business relationship.

Managerially this study has opened up a discussion on the possible usage of mobile technologies other than traditional e-business tools. This case illustrates how WLAN-based technology was used to effectively cut costs and streamline the order-delivery process as well as interfacing parts of that process. For managers interested in mobile solutions and how to proceed with ideas, a good reference point is an article by Aungst and Wilson (2005) in which they suggest 11 issues that should be covered when planning to adopt and use mobile solutions. The five most important ones are (1) coverage (WLAN or longer distance), (2) the mobile device platform, (3) the upgrade path, (4) the mobile application, and (5) issues with integration. Mobile technology such as m-ERP, mobile supply chain management, and WLAN-based internal systems provide novel tools to create leaner and meaner machines out of any organization that has the wisdom to grasp the mobile technology that will create the mobile future. Of course not all problems require a mobile solution but mangers ought to understand, and more importantly, recognize their problems and consider whether MC might provide a solution for their problems.

The main limitation of this paper is the exploratory and qualitative nature of the study. The second limitation is that only one business relationship in the steel industry context was studied. However, this study indicates that future studies should focus on how MC and mobile technologies are changing the logic of business relationships and how business relationships can excel in today's hypercompetitive landscape with the help of mobile systems. Future studies should employ a large scale survey about the usage of mobile technologies in business relationships to validate and broaden earlier findings.

REFERENCES

Angeles, R. (2000). Revisiting the role of Internet-EDI in the current electronic commerce scene. *Logistics Information Management, 13*(1), 45-57.

Aungst, S. G., & Wilson, D. T. (2005). A primer for navigating the shoals of applying wireless technology to marketing problems. *Journal of Business & Industrial Marketing, 20*(2), 59-69.

Chaffey, D. (2004). *E-business and e-commerce management* (2nd ed.). New York: Prentice Hall.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy of Management Review, 14*(4), 532-550.

Johanson, J., & Wiedersheim-Paul, F. (1975, October). The internationalization of the firm—Four Swedish cases. *Journal of Management Studies, 12,* 305-322.

John, G., & Reve, T. (1982, January). The reliability and validity of key informant data from dyadic relationships in marketing channels. *Journal of Marketing Research, 19,* 517-524.

Kadytė, V. (2005). Process visibility: How mobile technology can enhance business-customer care in the paper industry. *Proceedings of the International Conference on Mobile Business (ICMB).* Retrieved May 15, 2006, from http://ieeexplore.ieee.org

McCall, G. J. (1970). The social organization of relationships. In G. J. McCall, M. M. McCall, N. K. Denzin, G. D. Shuttles, & S. D. Kurth (Eds.), *Social relationship* (pp. 3-34). Chicago: Aldiline.

Naudé, P., Holland, C., & Sudbury, M. (2000). The benefits of IT-based supply chains-strategic or operational? *Journal of Business-to-Business Marketing, 7*(1), 45-67.

Okazaki, S. (2005). New perspectives on M-commerce research. *Journal of Electronic Commerce Research, 6*(3), 160-164.

Patton, M. Q. (1987). *Qualitative evaluation and research methods.* Beverly Hils, CA: Sage.

Perry, C., Cavaye, A., & Coote, L. (2002). Technical and social bonds within business-to-business relationships. *Journal of Business & Industrial Marketing, 17*(1), 75-88.

Reid, D. A., & Plank, R. E. (2000). Business marketing comes of age: A comprehensive review of the literature. *Journal of Business-to-Business Marketing, 7*(2/3), 9-178.

Ritter, T., & Gemünden, H. G. (2003). Interorganizational relationships and networks: An overview. *Journal of Business Research, 56*(9), 691-696.

Salo, J. (2006, May 21-24). Coping with business relationships: Use of mobile solutions to improve inter-organizational business processes. In *Proceedings of the Information Resources Management Association Conference,* Washington ,DC.

Scornavacca, E., Barnes, S. J., & Huff, S. (2005). *Mobile business research 2000-2004: Emergence, Current status and future opportunities*. Retrieved October 25, 2005, from www.m-lit.org

Sinisalo, J., Salo, J., Karjaluoto, H., & Leppäniemi, M. (2006). Managing customer relationships through mobile medium—Underlying issues and opportunities. *Proceedings of the 39th International Conference on System Sciences* Honolulu, HI.

Vlosky, R. P., Fontenot, R., & Blalock, L. (2000). Extranets: Impacts on business practices and relationships. *Journal of Business & Industrial Marketing, 15*(6), 438-457.

Wendelin, R. (2004). The nature and change of bonds in industrial business relationships. Doctoral thesis. Swedish School of Economics and Business Administration. Helsinki: Yliopistopaino.

Wilkinson, I. (2001). A history of network and channels thinking in marketing in the 20th century. *Australasian Journal of Marketing, 9*(2), 23-53.

Yin, R. K. (1989). *Case study research: Design and methods.* Newbury Park, CA: Sage.

KEY TERMS

Bond: Can be seen as a building block for a business relationship. Bonds are created through interaction between business parties.

Business Relationships: Relationships are created through the interaction between parties involved in exchange. Single acts and episodes, that is, the elements of interaction build the relationship over time. Satisfaction, commitment, and trust emerge as the distance between the parties is reduced. In due time, bonds are created on several levels of the relationship.

Mobile Commerce (MC): All commercial activities including sales and procurement that are conducted via mobile technology.

Mobile Technology: A form of information technology that can be used to mobilize various traditional activities. Includes sales force automation (SFA), mCRM, order pick-ups and other information and transaction flows between business parties.

Wi-Fi: 802.11.b based WLAN is often labeled Wi-Fi.

Wireless Local Area Network (WLAN): In short, a local area network without cables. Airwaves are used to transmit and receive data. The most commonly used WLAN is based on standard 802.11b.

Mobile Telecommunications and M-Commerce Applications

Clarence N.W. Tan
Bond University, Australia

Tiok-Woo Teo
Bond University, Australia

INTRODUCTION

This article presents an overview of prevailing trends and developments shaping mobile commerce (m-commerce) and the wireless economy. A review of wireless telecommunications infrastructure attempts to demystify the evolving technology landscape. Mobile Internet deployment and adoption demographics are highlighted, as are innovative wireless applications and current m-commerce issues.

BACKGROUND

The World Wide Web (WWW) and Web browser software brought mass market accessibility to the Internet. Riding on this ubiquity and reach is electronic commerce (e-commerce) in its many forms: inter-business dealing, intra-organization transactions and business-to-consumer trade, and so forth. E-commerce has witnessed impressive growth and continues to be a significant propellant of Internet progress. Participants have, however, hitherto been essentially tethered to fixed line connections. The development of novel wireless services and mobile adaptations of familiar applications (Ghini, 2000) is fueled by demand from increasingly nomadic users looking to access familiar online facilities, and the steady convergence of telecommunications and computing technologies (Messerschmitt, 1996).

Wireless telecommunications was conceived in the 1980s to carry voice, but has evolved to become data bearer, including Internet communications. The cellular telephone is now commonplace and more personal digital assistants (PDAs), hand-held computers and the like are sporting cellular radio connectivity. These devices form a sizable platform for deploying m-commerce applications. M-commerce refers to the ability to browse, interact with and make payment for goods and services directly from mobile terminals such as cell phones, PDAs and portable computers (Tan, 2002). Industry forecast statistics point to exponential growth in the sector:

- Worldwide shipment of Web-enabled wireless devices rose 796% in 2000 over 1999 and consumer transactions committed from such terminals will total US $1.8 trillion worldwide by 2005 (Source: Gartner Group).
- International wireless data market was expected to grow from 170 million to more than 1.3 billion subscribers between 2000–2004, equipping themselves with 1.5 billion wireless-capable handsets and other Internet appliances by end of 2004 (Source: Cahners In-Stat Group).
- Wireless Internet users in the Asia-Pacific region alone will rise 10-fold from 20 to 216.3 million between 2000–2007 (Source: Strategis Group).

As Internet and mobile communications converge, e-commerce evolves into m-commerce. The tremendous potential of "anytime" convenience and "anywhere" mobility in carrying out everyday online transactions will spur many unique mobile services yet.

TECHNOLOGY ROAD MAP

Early wireless telecommunications architecture in the late 1940s was modeled after television broadcasting. Tall, centralized transmitter towers provided radio coverage. Limitations like restricted user mobility and capacity, poor voice quality and high cost saw the introduction of new cellular technology in late 1970s—a superior architecture persisting to this day.

A cellular mobile communications system comprises a vast collective of low-power antenna subsystems, dispersed in small overlapping geographical units called cells. Individual cellular base stations provide local coverage and interconnect for a combined footprint that constitutes the wireless network. Modern implementations are typified by larger, sparse cells in rural areas and small, dense ones in metropolitan districts. The technology road map is demarcated by milestones corresponding to transmission bandwidth.

First Generation—1G: Analogue radio transmission characterized 1G cellular systems. The one-time de facto standard throughout the Americas and the Asia-Pacific was

the Advanced Mobile Phone Service (AMPS) introduced in the United States in 1983. Despite technical imperfections such as limited growth capacity, poor data transport and deficient transmission security, 1G systems maintained their popularity till the early 1990s. Improved derivatives of AMPS are still deployed in the remaining analogue cellular networks around the world today.

Second Generation—2G: Digital radio transmission heralded the 2G era. Information is digitized into a stream of computer binary coded data packets for transmission and reassembly at the receiving end. Two competing digitization schemes are Time Division Multiple Access (TDMA) and Code Division Multiple Access (CDMA). Better bandwidth utilization boosts network capacity, enhances coverage and improves voice quality. New service features such as data encryption, short text messaging, fax and data transmission can also be offered.

Launched commercially in 1991, the European developed, TDMA-based Global System for Mobile Communications (GSM) is the de facto international 2G standard today, with 863.6 million subscribers in 174 countries at end of May 2003 (Source: GSM Association, *http://www.gsmworld.com*). An American adaptation of GSM called PCS 1900 was launched in late 1995. CDMA-based networks began commercial operation in 1995 and are now the predominant standard in the Americas, Korea and Japan, with 164.1 million subscribers in 60 countries as at June 2003 (Source: CDMA Development Group, *http://www.cdg.org*).

Conventional fixed line Internet connectivity offers varying data rates:

- Up to 56 kbps (kilobits per second): Analog modem
- 64–128 kbps: Integrated Services Digital Network (ISDN)
- 256 kbps–1.5 Mbps (megabits per second): Optical fiber and Digital Subscriber Line (DSL)

In comparison, 2G data service provides 9.6–57.6 kbps throughput, with most network operators supporting only speeds no more than 14.4 kbps. This makes for poor overall user experience for consumers increasingly accustomed to higher-speed, broadband Internet access through fixed connections. Real-time multimedia applications such as live video are also impracticable on 2G architecture. Solutions better optimized for wireless data transmission are clearly needed to meet growing demand for quality mobile Internet access.

Second-and-a-Half Generation—2.5G: 2.5G systems extend 2G infrastructures for upgraded data throughput. New handsets, however, are required to tap the improvements and other added functionalities. The enhancement to GSM is called the General Packet Radio Service (GPRS). Capable of a theoretical maximum 171.2 kbps transmission, the average is 40–80 kbps deployed in practice. Like the Internet, GPRS networks are also based on the Internet Protocol (IP) standard, so GPRS terminals function just like other wired Internet sub-nodes with seamless access to familiar applications such as the WWW, e-mail, Telnet and FTP (File Transfer Protocol), and so forth. The first commercial GPRS service was inaugurated in the United Kingdom in June 2000. 2.5G CDMA technology is known as the IS-95B standard and offers ISDN-like data rates. Its first commercial debut in Japan in early January 2000 beat GPRS to market by some 6 months, and the 64 kbps throughput was superior to prevailing 2G GSM offerings.

While data speed boost is useful in time-sensitive applications like online transactions and credit authorization, there remain technical idiosyncrasies and transmission latencies detrimental to time-critical multimedia functions such as video decoding and playback. Hence, 2.5G is still not the ideal platform for deploying high quality, real-time video conferencing.

"2.75G": When wireless data rates approach those of conventional fixed line broadband connectivity, user experience will improve significantly and truly mobile Internet access would be a reality. Some have termed this stage of the road map loosely as "2.75G".

The Enhanced Data rates for Global Evolution (EDGE) standard is engineered as an extension to GSM technology and leverages past investments in TDMA and GPRS. Throughput of 384–553.6 kbps is theoretically possible and early EDGE-compatible handsets support rates in excess of 100 kbps. The world's first commercial EDGE service was deployed in the United States in July 2003. The equivalent evolution in CDMA technology is the CDMA2000 standards, providing network speeds from 144–614 kbps. From the first commercial service in Korea in October 2000, CDMA2000 subscribers have grown to number 54.1 million by June 2003 (Source: CDMA Development Group).

Third Generation—3G: Wireless broadband connectivity and industry standards harmonization characterize 3G systems. New handsets are again necessary to exploit multimedia capabilities and applications enabled by high-performance wireless networking at unprecedented 384 kbps–2 Mbps speeds.

More importantly, the vision is for 3G devices to roam seamlessly on enabled networks within a country and across continents, creating truly borderless mobile services. To this end, the International Mobile Telecommunications 2000 (IMT-2000) recommendations were promulgated in late 1999 with international participation. IMT-2000 prescribes how mobile service providers should evolve existing cellular networks towards full inter-network compatibility independent of underlying radio technologies. Disparate TDMA, CDMA, GSM systems and their derivatives will be accommodated. For example, under IMT-2000, GSM will evolve into the Universal Mobile Telecommunications System (UMTS) and employ a new transmission

technology called Wideband CDMA (W-CDMA). The first commercial UMTS 3G service was introduced in Japan in October 2001.

WIRELESS MESSAGING

SMS—Short Message Service: The Short Message Service is a wireless radio service for bidirectional transfer of alphanumeric messages of up to 160 characters each among mobile terminals in GSM and UMTS networks (European Telecommunications Standards Institute, 2002). Multiple text messages (up to 255) may be concatenated to form longer ones. Messaging is near instantaneous and operates on a store-and-forward scheme that guarantees delivery. Sent messages are routed to their destinations by an electronic intermediary in the cellular network called a Short Message Service Center, or SMSC (European Telecommunications Standards Institute, 2002). If the intended recipient is not immediately contactable, the message is stored for a stipulated time period to facilitate reattempted delivery. The SMSC can notify a message sender of the ultimate delivery success or failure via an SMS notice called a delivery report.

Introduced in 1991, SMS began as a supplement to voice calls, alerting subscribers to awaiting recorded messages. Its use has since grown to encompass notification services like paging alerts, new e-mail notice and calendar reminders; and information services like weather forecast, traffic watch, stock quotes, gaming results and horoscope readings, and so forth. Interpersonal communications still dominate SMS messaging volume by far today but SMS is also finding increasingly sophisticated uses in corporate applications such as job dispatch (Tan, 2002) and secure m-commerce transactions systems (Tan, 2003).

In 2002, an estimated 360 billion SMS messages were sent worldwide with some 30 billion sent monthly in December—double the rate 24 months ago (Source: GSM Association). The growth trend is expected to continue.

EMS—Enhanced Messaging Service: EMS is a minor extension to SMS whereby a mix of formatted text (e.g., bold, italics, etc.), simple animations, tiny pictures and short melodies can be included as message content. The specification provides for amalgamating multiple classic SMS messages to convey the increased EMS message complexity. Compatible handsets have been commercially available since 2001, but there has been a lack of full interoperability among different makes and models due to proprietary implementations of the EMS standard by manufacturers.

MMS—Multimedia Messaging Service: MMS is the latest generation wireless messaging standard that inherits the SMS store-and-forward schema and features improved security mechanisms for managing message encryption, authentication and privacy. MMS is devised to carry a full range of multimedia elements including text, audio clips, still pictures, animated images and full-motion video, all encoded using industry standard file formats (e.g., JPEG for photographs, MP3 for audio, MPEG for video, etc.). Such rich content increases message size and considerably more bandwidth is needed to deliver MMS messages expeditiously. Hence, the takeoff of MMS is only expected to dovetail on the rollout of 3G networks internationally. The world's first commercial MMS service was introduced in Hungary in April 2002.

INNOVATIVE APPLICATIONS

The sheer success of the mobile telephone as a consumer device is indisputable. Mobile phones have already outstripped conventional telephone subscriptions in many countries. As cellular telecommunications networks and the Internet converge, the spillover of Internet e-commerce will fuel m-commerce takeoff. Wireless broadband technology can be used to deploy not only mobile variants of existing Web-based applications, but also fundamentally new ones. This creates new value proposition for consumers and revenue streams for vendors. The true power of the mobile Internet lies in these innovative applications that will be spawned, some of which are listed next.

- **Location-based services (LBS):** LBS refers to the provisioning of value-added services to cellular subscribers based on the physical location of their mobile devices within the network. Potential applications: location-based call billing, emergency medical services, courier fleet tracking, proximity-based personalized marketing, locality information services like navigation and directory assistance, and so forth.
- **Mobile multimedia entertainment:** Higher bandwidth afforded by 2.5G and 3G networks creates new distribution channels for paid multimedia entertainment like computer games, music and video. For example, games can be downloaded wirelessly for play on enabled handsets or with other gamers on the network. Music and video vendors may stream content wirelessly on demand to mobile devices for instantaneous preview playback and subsequent purchase.
- **Wireless telemetry:** The facility for gathering data or remotely controlling devices over the footprint of a cellular network through two-way communication between remote equipment and a central facility. Potential applications: security alarm monitoring, climate control, meter readings and inventory status polling, and so forth.
- **Wireless electronic payment systems:** Cellular phones become secure, self-contained tools for instantly authorizing payment for goods and services wirelessly over the network. One example is the use

of SMS messages for making purchases at vending machines.

- **Telematics:** The integration of wireless telecommunications and in-vehicular monitoring and location systems will bring mobile Internet facilities, location-based services and multimedia entertainment to vehicle dashboards. Potential applications: online information services (e.g., traffic news, weather forecast, route assistance, etc.), target advertising based on vehicle locale and streaming media content, and so forth.
- **Wireless telemedicine:** Cellular telecommunications and wireless broadband multimedia technologies can deliver health care expertise and services remotely. Potential applications: remote monitoring of patients in ambulances and long-distant specialist teleconsultation for patients in remote regions.

CURRENT ISSUES

The realm of m-commerce is still in its infancy and to the extent that it bears roots in e-commerce, many technical, business, marketing and legal perspectives in m-commerce may be extrapolated from e-commerce ones. But as a fundamentally new business paradigm, m-commerce will elicit contentious issues of its own. A few have been highlighted as follows.

- **Wireless privacy:** Industry self-regulation and legislation are needed to guard against potential threat of undue surveillance and profiling, invasion to personal privacy and wireless spam resulting from abuse of location data gathered for provisioning LBS.
- **Wireless security:** Increased vigilance is necessary as computer viruses and like malicious attacks begin to target PDAs and mobile phones, as such devices gain more processing power and intelligence, and users become ever more reliant on them for voice and data services.
- **Wireless emission and public health**: Concerns persist on the long-term impact on public health from cellular emission, radiation and electromagnetic interference in the face of rapid consumer adoption of mobile phones. While no conclusive evidence exists to suggest significant risks, a more precautionary approach is prudent pending further research.

CONCLUSION

Mobile telecommunications have seen an evolution from a voice focus to data-centricity. For now, 3G networks present a viable channel for distributing practically any digital content. The future along the technology roadmap continues to emphasize quality multimedia delivery. Cellular devices are expected to improve in form and function, benefiting from advances in display, storage, imaging, processing and battery technologies. Historically, new handsets were required for each successive generation of wireless infrastructure and their timely volume availability at mass market prices remains crucial to the success of future networks.

The continued proliferation of mobile-enabled portable consumer electronics and the rising demand for ubiquitous computing will contribute to the volume and value of m-commerce, and begin to unleash the power of the mobile Internet.

REFERENCES

European Telecommunications Standards Institute. (2002). *ETSI TS 122 105 V5.2.0—UMTS Services and Service Capabilities.* France.

European Telecommunications Standards Institute. (2002). *ETSI TS 123 040 V5.5.1—Technical Realization of Short Message Service (SMS).* France.

Ghini, V., Pau, G., & Salomoni, P. (2000). Integrating notification services in computer network and mobile telephony. *Proceedings of the 2000 ACM Symposium on Applied Computing*.

Messerschmitt, D.G. (1996). The convergence of telecommunications and computing: What are the implications today? *Proceedings of the IEEE, 84*(1), 1167–1186.

Tan, C.N.W., & Teo, T.W. (2002). A short message service (SMS) enabled job dispatch system. In C.-H. Yeh & S. Tekinay (Eds.), *Proceedings of the International Conference on Wireless Networks 2002* (pp. 346-352). Pennsylvania: CSREA Press.

Tan, C.N.W., & Teo, T.W. (2002). From e-commerce to m-commerce: The power of the mobile Internet. In J.D. Haynes (Ed.), *Internet management issues: A global perspective* (pp. 27-53). Hershey, PA: Idea Group Publishing.

Tan, C.N.W., & Teo, T.W. (2003). An authenticated short message service (SMS)-based transactions system without SIM modification. In W. Zhuang, C.-H. Yeh, O. Droegehorn, C.K. Toh & H.R. Arabnia (Eds.), *Proceedings of the International Conference on Wireless Networks 2003.* Pennsylvania: CSREA Press.

KEY TERMS

LBS (Location-Based Services): The provisioning of value-added services to cellular subscribers based on the physical location of their mobile devices within the network.

M-Commerce (Mobile Commerce): The ability to browse, interact with and make payment for goods and services directly from mobile terminals such as cell phones, personal digital assistants (PDAs) and portable computers.

SMS (Short Message Service): A wireless service for bidirectional transfer of alphanumeric messages among mobile terminals in digital cellular networks.

WAP (Wireless Application Protocol): An open, extensible and unified global standard for delivering information and providing interactive services over cellular networks to mobile devices.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1984-1988

Mobile Payment

Gyõzõ Gódor
Budapest University of Technology and Economics, Hungary

Zoltán Faigl
Budapest University of Technology and Economics, Hungary

Máté Szalay
Budapest University of Technology and Economics, Hungary

Sándor Imre Dr.
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The widespread usage of new telecommunication technologies implies the demand on payment via Internet since the '90s. First, these solutions were applied only by pioneer users, while average men still chose traditional payment methods such as payment by cash, cheque, or bank transfer. In the latest decade, the notable improvement of mobile communications allowed the provision of customized services. A new payment method has appeared which is called mobile-payment. Consequently, increasing number of banks provide access to their services via mobile equipment.

Reliable network security is an essential prerequisite for the expansion of the rapidly growing world of electronic payment. Public key infrastructure (PKI) offers the capabilities needed to provide this security. Establishing trust in a wireless public key infrastructure (WPKI) is crucial for the success of applications that will exploit the opportunities created by handheld wireless devices. This trust is based on the reliability of the technology but also on a carefully implemented system of laws, policies, standards, and procedures.

The development of trusted electronic transactions is motivated by legislation. The EU adopted a legislative framework to guarantee the security and acceptance of electronic signatures in 1999. The U.S. adopted legislation for the recognition of electronic signatures in national and global trade in June 2000 (Sievers, 2000).

This article deals with mobile payment and mobile banking services and focuses particularly on the mobile side of the system. First, we introduce the technological background necessary for developing m-services, and we define the m-payment reference model. After that, the differences between chip-card and software based implementations will be presented. Finally, we conclude the article and summarize the main terms used in the article.

BACKGROUND

The Mobile Payment Forum (MPF) (2002) defines mobile-payment (m-payment) as the process of two parties exchanging financial value using a mobile device in return for goods or services. The trusted transactions of a mobile payment system are called mobile payment transactions. The main areas of use are the following:

- m-banking and m-payment, in case of performing banking and payment affairs;
- m-administration, when accomplishing administration tasks; and
- m-government, in case of arranging public administration affairs using the mobile electronic way.

The mobile device and the mobile network have two main roles in m-payment:

- they enable secure client authentication and identification; and
- they support the generation of digital signatures on the client side.

The user authentication means that a service provider determines the identity of a user (Kanniainen, 2001).

The digital signature is an electronic signature that can be used to authenticate the identity of the sender of a message or the signer of a document, and possibly to ensure that the original content of the message or document that has been sent is unchanged. Digital signatures can be used for many purposes, such as authorizing a subsequent transaction or creating a signature of the user with properties fulfilling the requirements of electronic signature laws.

The user authorization means that a service provider ensures that the user has viewed and accepted a transaction contract (Kanniainen, 2001).

The technological background exists for developing services based on trusted mobile transactions. The bandwidth of the mobile channel is only a small fraction of that of the Internet, but user authentication, digital signature transfer, authorization control require low bandwidth from the mobile network. Even low-end mobile devices support WAP functionalities (Wireless Application Protocol Forum [WAP Forum], 2001c) and text message sending. Their SIM card implements SIM Application Toolkit (SAT) and supports the necessary cryptographic algorithms at chip-card level. These are essentials to implement client-side banking applications (Van der Merwe, 2003).

Smart-phones with increased processing speed have already been introduced to the market. They support the development of software-based banking applications without using the functionalities of chip-card and relying only on the security of software-based encryption algorithms (Mobey Forum Mobile Financial Services Ltd. [Mobey], 2003).

In 2004 the Trusted Mobile Platform was introduced by IBM and Intel. It is a software and hardware requirement specification for mobile equipment (Trusted Mobile Platform, 2004) in secure environment.

Trusted mobile services are based on PKI technology. PKI offers strong authentication and encryption mechanisms and facilitates the secure exchange of sensitive messages in public information networks (Torvinen, 2000). PKI functions permit detection of messages that have been tampered with or altered during transmission. PKI summarizes the processes and techniques performing key-certification in public cryptography architecture. Each entity in the PKI system has at least one key-pair which consists of a private key and a public key. The private key is the secret of a given entity never discovered to others. Public keys are certified by a trusted third party called Certification Authority (CA) (WAP Forum, 2001a).

WPKI is a PKI where at least the user-side of the system uses wireless medium. WPKI uses more compact certificates then PKI and its certificate acquisition process is adapted to mobile environment (WAP Forum, 2001b).

Several international technical organizations were founded starting from the late ’90s, elaborating standard solutions or directives for trusted mobile transactions. Such organizations are Radicchio, Liberty Alliance, GSMA, ETSI M-COMM Working Group, OMA, MPSA, Mobey Forum, and Mobile Payment Forum, for example. Members of these organizations are mobile operators, financial institutions, research, developing, and standardization organizations (see Table 1).

Table 1. Organizations specifying trusted mobile transactions

Name	Foundation	Members
Radicchio (T2R)	Mars 2002	GSMA, Liberty Alliance, ETSI
Liberty Alliance	Sept 2001	VeriSign, Nokia, Sun, RSA, Vodafone, American Express, Novell
MPSA (SimPay)	Mars 2003	Vodafone, T-Mobile, Orange, Telefónica Móviles
Mobey Forum	Mai 2000	VISA, ABN-Amro Bank, Nokia, Deutsche Bank
MeT	April 2000	NEC, Nokia, Panasonic, SonyEricsso2n
Mobile Payment Forum (MPF)	Nov. 2001	American Express, JBC Co. Ltd., MasterCard International and Visa International
Open Mobile Alliance (OMA)	June 2002	Vodafone, Ericsson, WAP Forum, IT companies

MOBILE-PAYMENT SERVICES

An m-payment system involves a wireless device that is used and trusted by the customer. M-payment is not a new payment instrument but an access method to activate existing payment transactions processed by banks.

Mobile payment transactions (and systems) can be classified upon location basis or value basis. On location basis, local and remote environments are distinguish (Mobile electronic Transactions [MeT], 2001), on value basis, micro (<10Euros) and macro (>10Euros) payment (Mobey, 2003). In local environments transactions are usually initiated over a short-range wireless technology such as Bluetooth or RFID (Saleem, 2002). A typical application would be retail shopping using an account-based payment made from a mobile device. In remote environment the connection between the content server and mobile device is established via a Public Land Mobile Network, such as the GSM cellular network (Mobey, 2003).

The four main parties involved in a mobile payment transaction (see Figure 1)—the user, the network operator, the financial institution, and the merchant—share many of the same concerns that need to be addressed by a mobile payment standards body (Henkel, 2001; MPF, 2002).

- Consumers are mostly concerned with security, ease of use, and privacy. They also require the payment scheme to work across multiple devices, including mobile phones, PDAs, wireless tablets, and handheld computers.
- Mobile operators’ principal concerns revolve around standardization and interoperability. Operators want payment to be seamless, allowing them to compete on services and applications.

Figure 1. The mobile payment transaction

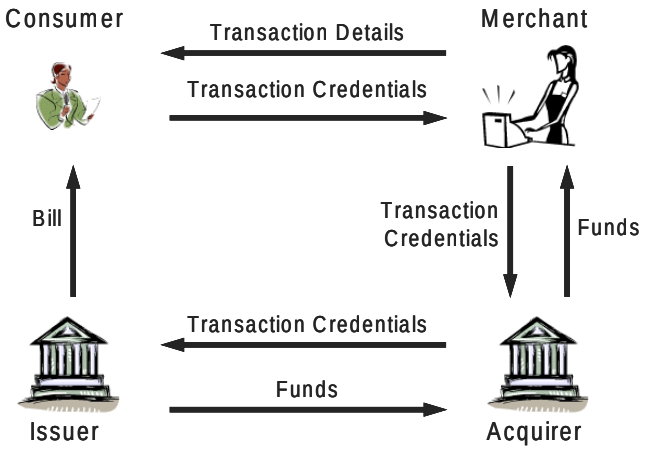
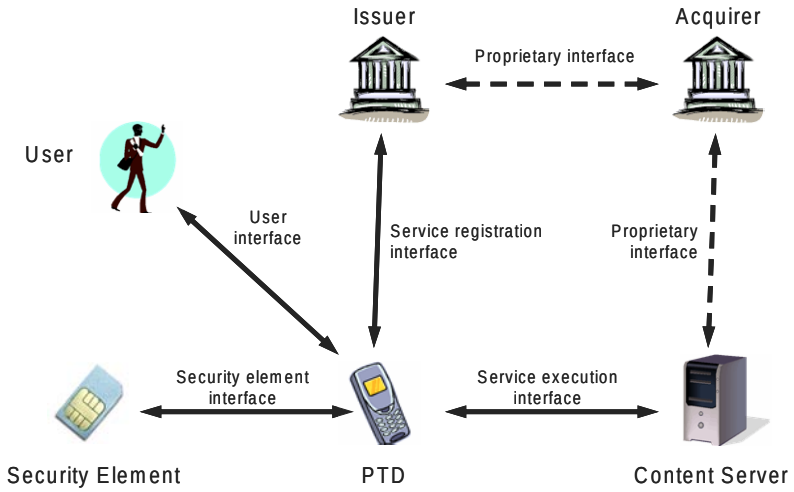


Figure 2. The reference model of the m-payment transaction



- Financial institutions, meanwhile, are primarily concerned with ensuring the integrity of the payment system and reducing the risk of fraud. M-payment solutions must have synergy with existing payment instruments and infrastructure, systems, processes, and rules.
- Merchants or content providers want the payment process to be transparent to the user, as this encourages greater usage and/or propensity to complete a purchase. They also want any payment scheme to facilitate swift and easy completion to ensure they get paid on time.

Figure 2 shows the reference model of mobile payment systems.

The main elements of the reference model are the following:

- **Mobile Device:** Its main functions are the authentication of the subscriber's identity and granting of the permission for the operation of a secure transaction. It is called Personal Trusted Device (PTD) by MeT (2001).
- **Security Element:** The mobile device includes a security element which provides a tamper-proof

environment to run the necessary cryptographic functions. It contains the user's private-public key pairs and root certificates (used to verify other certificates) (Mobey, 2003).

- **Issuer:** This is the user's bank or a provider via which the mobile electronic transactions are running. One of its main functions is to issue certificates for the subscriber when the user registers himself for the m-banking service. The issuer should produce at least two certificates for the user, which are the following: one is used during the authentication process; the other is used during the digital signing process.
- **Content Provider:** It corresponds to the merchant. The content provider supplies content to the user's mobile device. The application running on a content server may request the user to perform user authorization (sign a transaction).
- **Acquirer:** The acquirer provides a single point of contact between issuers and content providers. In mobile payment transactions, the acquirer's role is to provide the business rules and relationship among multiple content providers and issuers. As a consequence, banks do not need to make agreements with every merchant and vice versa.

A consumer decides to pay in a shop using mobile payment service. When the transaction is initiated, he sends the URL of his bank to the merchant. The merchant redirects the user to the issuer (bank) and sends him the total amount. The bank and the user authenticate each other; the bank sends the transaction data to the user. If the user decides to authorize the transaction, he signs the content with his digital signature by entering his signing PIN. Then the bank connects to the merchant and requests his authorization to finish the transaction. After this, the bank redirects the user's browser to the page of the merchant, and the merchant notifies the user about the successful termination of transaction (Vilmos & Karnouskos, 2003; Karnouskos, Hondroudaki, Vilmos & Csik, 2004).

TECHNOLOGICAL SOLUTIONS FOR M-PAYMENT SERVICES

The security element—defined in the reference model—can be implemented in many ways, depending on the agreement of banks, m-payment service issuers, and mobile operators. Two main categories exist: chip-card based solutions or software-based solutions. In chip-card based (Kanniainen, 2001) solutions, security-related functions and keys are stored on the chip-card. Applications not requiring a tamper-proof environment (e.g., the menu of the service) can be stored both on a chip-card and/or a mobile device. In software-based (Mobey, 2003) solutions all parts are implemented in software on the mobile device, including cryptographic functions, certificate and key storage. In case of chip-card implementations banks, and mobile operators have to decide either to put banking applications and keys on the same chip-card or on a separate one.

Security element implementations can also be classified by distinguishing between SIM-card dependent and SIM-card independent solutions referring to the role of mobile operator. In case of SIM-card dependent solutions, the applets, keys, and a part of the certificates for the banking and payment applications are stored on the chip-card under the control and authority of the mobile operator. In case of SIM-card independent solutions nothing from the banking application, private key and certificate for m-payment is stored on the chip-card for mobile operators.

SIM-Dependent Solutions

Advantages

SIM-dependent solutions are advantageous for the user because all functionalities are integrated into one chip-card and one device. The merchant can strongly authenticate the user through a trust chain starting from his bank and ending at the user's bank. Device manufacturers do not need to develop multi-card devices. The security level of this solution is high. The chip-card is portable from one device to another, can dispose standard interface via the device, and applications on the chip-card are protected against viruses.

Disadvantages

The main disadvantage of SIM-dependent banking applications from the bank's point of view is the dominant role of the mobile operator in the system. The bank and the mobile operator have to agree on their choice of chip-card manufacturer, the set of applications, and keys to upload to the card. The mobile operator can influence the choice of technical solution for the banking application. Besides these, it is the bank who has to handle the financial responsibility and risks. Mobile operators can charge the bank with costs of service. Registration for m-payment and m-banking service can be done in the stores of the mobile operator.

SIM-Independent Solutions

An SIM-independent solution means that the mobile operator does not have to change its SIM cards. There are two major types of SIM-independent solutions. The implementation of security element can be:

- a second chip-card (dual-chip device [Kanniainen, 2001]), which may be either one common chip-card

for all banks or one separate chip-card for each bank; or
- implemented in software as an application that is downloadable to the mobile device.

Dual-Chip Solution

Advantages
Banks become independent from the mobile operator because they provide the user a separate chip-card. The tasks of registration, chip-card distribution, and other services of the bank and of the mobile operator diverge. Later, newer applications can be downloaded over the air (OTA) to the chip-card provided by the bank without affecting the applications of the mobile operator. The mobile operator and the m-payment provider may choose chip-card manufacturer independently. The mobile operator does not have to extend its registration and chip-card distribution system. The device manufacturer gains a dominant role because new, dual-chip card devices are needed, and mobile operators and banks would support this activity. The user can feel himself more secure because his security data are not only on one card, but can also see the separation between mobile operator and banking services. The user does not have to change his SIM-card. The merchant can strongly authenticate the user through the trust chain of banks. High security applications can be implemented using this solution.

Disadvantages
In general, we cannot expect the introduction of these devices from device manufacturers. There are several reasons for this. The size of the device would grow, their hardware structure would change, and the production of new devices would be expensive. Another problem is that banks would have to handle chip-card management. In the case of one common chip-card for all banks, banks would need to agree on the application set installed on the chip-card, or they would have to choose the same m-payment service provider. Banks usually do not want to deal with the distribution of new chip-cards into the phones. This solution is costly for banks.

For the big-sized dual-chip devices, a future solution may be the use of secure RFID chips (Mobey, 2003). These chips do not need real slots; they are contactless. It is enough to put them near the device, on its back, for example. In this case, the bank would give an RFID chip to the user at the registration. Secure communication between the RFID chip and the mobile device is indispensable.

Software-Based Solutions

In the case of a software based security element (Kanniainen, 2001; Mobey, 2003), all the banking application are uploaded on the device without the need of any SIM-based function.

Advantages
This is a cheap and quick solution for banks and mobile operators. Device manufacturers tend to produce smart-phones capable of running these applications. Mobile operators will have traffic increase on their networks due to m-services. Theoretically, banks could upload their own applications separately from mobile operators, and banking services could work easily. In practice, they have to cooperate with mobile operators because operators have the only control over the communication channel. The applications are easy to upgrade but require high caution. The downloading of new applications to a coded device may require some agreement with the mobile operator.

Disadvantages
One of the disadvantages of software-based solutions, from a technical point of view, is the huge and rapidly increasing number of new software environments and operating systems. The client application must be prepared to run in several environments and should pass many audits during development. Besides this, if software-level encryption is used, viruses and malicious code pose a much higher risk than in the case of chip-card based solutions. Chip-cards filter out non-standard messages by default.

FUTURE TRENDS

A key issue concerning mobile payment is interoperability. As interoperability can be achieved by the use of (preferably international) standards, the intensive standardisation work is very important for mobile payment systems to get widely used.

Another important issue is a PKI as global as possible and trusted by all the players of the system. As Lannerström (2000) states, wireless devices offer tremendous flexibility when combined with a PKI. With the use of a PKI for wireless systems, a mobile device turns into an inexpensive but powerful device by which the user can be securely authenticated and digital signatures can be produced. As these features are absolutely necessary for mobile payment systems, PKI will also play an important role in the evolution of these systems.

CONCLUSION

In this article, the main concepts of m-payment were presented. The article gives a general overview of m-payment solutions, without going into too much technical detail but addresses technical issues. First, the mobile payment model was presented; the four participants were introduced. Then, we focused on the difference between chip-card based and

software-based client side banking applications. We analyzed them from the point of view of participants in the m-payment business.

REFERENCES

Henkel, J. (2001). Mobile payment. *The German and European Perspective*. Department of Computer Science and Engineering, Indian Institute of Technology, Bombay, India. Retrieved August 19, 2005, from Indian Institute of Technology Web site http://www.cse.iitb.ac.in/~anil/MTP/MobilePayment.pdf

Kanniainen, L. (2001). *The preferred payment architecture*. Technical Documentation. (Requirements for manufacturers and standardisation bodies, Version 1.0). Retrieved August 19, 2005, from Mobey Forum Web site http://mobeyforum.org/public/material/PPATechnical.pdf

Karnouskos, S., Hondroudaki, A., Vilmos, A., & Csik, B. (2004, July 12-13). *Security, trust and privacy in the SEcure MObile Payment Service*. The 3rd International Conference on Mobile Business 2004 (m>Business), New York.

Lannerström, S. (2000). *Wireless PKI: Opportunities.* Radicchio White Paper (Publication No. WP-SMD-002). Retrieved August 19, 2005, from Liberty Alliance Web site http://www.projectliberty.org/Radicchio/downloads/smd_002.pdf

Mobey Forum Mobile Financial Services Ltd. (2003). *Mobey forum white paper on mobile financial services 1.1*. Retrieved August 19, 2005, from Mobey Forum Web site http://mobeyforum.org/public/material/Mobey%20Forum%20White%20Paper%20on%20Mobile%20Financial%20Services%20v1_14.pdf

Mobile Electronic Transactions. (2001, February 21). MeT Core Specification. (Version 1.0). Retrieved July 15, 2005, from MeT Web site http://www.mobiletransaction.org/pdf/MeT-Core-Spec-20010221.pdf

Mobile Payment Forum Inc. (2002). *Enabling secure, interoperable, and user-friendly mobile payments*. Mobile Payment Forum White Paper. Retrieved August 19, 2005, from MPF Web site http://mobilepaymentforum.org/pdfs/mpf_whitepaper.pdf

Saleem, R. (2002). *Preferred payment architecture: Local payment*. (Local Payment Discussion Document 1.0). Retrieved August 19, 2005, from Mobey Forum Web site http://mobeyforum.org/public/material/Local%20Payments%20Discussion%20 Document%201.0.pdf

Sievers, M. (2000). *Legislation and PKI evolution*. Radicchio White Paper (Publication No. WP-LEG-001). Retrieved August 19, 2005, from Liberty Alliance Web site http://www.projectliberty.org/Radicchio/downloads/leg_001.pdf

Torvinen, V. (2000). *Wireless PKI: Fundamentals*. Radicchio White Paper (Publication No. WP-SMD-001). Retrieved August 19, 2005, from Liberty Alliance Web site http://www.projectliberty.org/Radicchio/downloads/smd_001.pdf

Trusted Mobile Platform. (2004). *Trusted Mobile Platform* (Hardware Architecture Description - Revision 1.0). Retrieved August 19, 2005, from Trusted Mobile Platform Web site http://www.trusted-mobile.org/TMP_HWAD_rev1_00.pdf

Van der Merwe, P.B. (2003). *Mobile commerce over GSM: A banking perspective on security*. MSc Theses, Faculty of Engineering, University of Pretoria, Pretoria, South Africa.

Vilmos, A. & Karnouskos, S. (2003, September 1-5). SEMOPS: Design of a new payment service. *International Workshop on Mobile Commerce Technologies & Applications (MCTA 2003). Proceedings of the 14th International Conference DEXA 2003*, Prague, Czech Republic.

Wireless Application Protocol Forum LTD. (2001a, April 24). Wireless application protocol. *Public key infrastructure definition*. (WAP-217-WPKI). Retrieved August 19, 2005, from Open Mobile Alliance Web site http://www.openmobilealliance.org/tech/affiliates/LicenseAgreement.asp?DocName=/wap/wap-217-wpki-20010424-a.pdf

Wireless Application Protocol Forum LTD. (2001b, May 22). *Wireless application protocol, WAP certificate and CRL profiles specification*. (WAP-211-WAPCert). Retrieved August 19, 2005, from Open Mobile Alliance Web site http://www.openmobilealliance.org/tech/affiliates/LicenseAgreement.asp?DocName=/wap/wap-211-wapcert-20010522-a.pdf

Wireless Application Protocol Forum LTD. (2001c, July 12). *Wireless application protocol architecture specification*. (WAP-210-WAPArch-20010712). Retrieved August 19, 2005, from Open Mobile Alliance Web site http://www.openmobilealliance.org/tech/affiliates/LicenseAgreement.asp?DocName=/wap/wap-210-waparch-20010712-a.pdf

KEY TERMS

Authentication: Proof of identity.

Digital Signature: An electronic signature based upon cryptographic methods of origin authentication. Usually it is appended to a message to assure the recipient of the authenticity and integrity of the message.

Mobile Payment: The process of two parties exchanging financial value using a mobile device in return for goods or services.

Mobile Transaction: Trusted transactions of a Mobile Payment system.

PKI: The abbreviation of Public Key Infrastructure, a set of policies, processes, server platforms, software, and workstations used to administer certificates and public-private key pairs, including the ability to issue, maintain, and revoke public key certificates.

Registration: A procedure where the account of the given services and the subscriber's identity are coupled.

RFID: Abbreviation of Radio Frequency Identification, a transponder technology for the contactless recognition of objects.

SAT (SIM Application Toolkit): A standard operational environment for applications stored on the SIM (and the third generation USIM).

SIM (Subscriber Identity Module): The subscriber dependent part of the mobile equipment.

Smart-Phone: Voice centric mobile phone with information capability.

Mobility-Aware Grid Computing

Konstantinos Katsaros
Athens University of Economics and Business, Greece

George C. Polyzos
Athens University of Economics and Business, Greece

INTRODUCTION

Grid computing has emerged as a paradigm for coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations (Foster, 2001). A grid computing system is essentially a large-scale distributed system designed to aggregate resources from multiple sites, giving to users the opportunity to take advantage of enormous computational, storage, or bandwidth resources that would otherwise be impossible to attain. Current applications of grid computing focus on computational-expensive processing of large volumes of scientific data, for example, for earthquake simulation, signal processing, cancer research, and pattern search in DNA sequences.

At the same time, the recent advances in mobile and wireless communications have resulted in the availability of an enormous number of mobile computing devices such as laptop PCs and PDAs (personal digital assistants). Thus, it is natural to extend the idea of resource sharing to mobile and wireless computing environments. Resource-sharing collaboration between mobile users appears as a promising research direction toward the alleviation of the inherent resource constraints present in mobile computing environments. Either in the context of mobile ad hoc networks (MANETs) or in wireless networks based on fixed infrastructure (i.e., cellular networks, wireless local area networks (WLANs), small- or large-scale communities of mobile users can form mobile grid systems and collaborate in order to either achieve a common goal (otherwise impossible to achieve) or simply overcome their individual limitations. In the following, we highlight the fundamental issues toward the realization of a computational mobile grid system.

BACKGROUND

The research area of mobile grid is relatively new compared to the traditional (fixed) grid and mobile and wireless computing research areas. A consensus on the exact character of mobile grid computing has not been reached yet. Hence, a classification of the existing approaches is provided in the following, aiming at the clarification of the mobile grid concept. At the same time, the most significant research directions within each approach are described.

As mentioned earlier, a primary distinction is made between various research efforts in the area based on whether mobile devices act exclusively as resource consumers or as resource providers as well.

Mobile Devices as Resource Consumers

In this case, research is motivated by the fact that mobile devices are considered to have limited computational and/or storage capabilities (Banavar, Beck, Gluzberg, Munson, Sussman, & Zukowski, 2000; Migliardi, Maheswaran, Maniymaran, Card, & Azzedin, 2002; Park, Ko, & Kim, 2003; Srinivasan, 2005). The grid, in this case, can provide the resources missing in mobile devices on demand. The emerging problems here stem from the mobile and wireless character of the devices and include intermittent connectivity, device heterogeneity (in terms of hardware and operating system), and limited battery life. The use of proxies is proposed in Park et al., which act as gateways to the grid. These proxies undertake the role of the mediator between the mobile device and the grid system, and try to hide device heterogeneity and intermittent connectivity by acting on behalf of the mobile device.

Other approaches target the provision of a "smart" environment for pervasive computing (Banavar et al., 2000; Srinivasan, 2005). Here, mobile devices are considered as pure access devices with no need for enhanced processing and/or storage capabilities (Migliardi et al., 2002). The role of the grid is to provide all the functionalities required by users, pushing this way the complexity of the system to the network rather than to the edges.

Mobile Devices as Resource Providers

In this case, mobile devices participate in grid systems as resource providers as well (Kurkovsky & Bhagyavati, 2003; Li, Sun, & Ifeachor, 2005; Litke, Skoutas, & Varvarigou, 2005; Park et al., 2003; Phan, Huang, & Dulan, 2002). Strong emphasis is given to two important factors. First, even though mobile devices have limited resources compared to their

stationary counterparts, they seem to increasingly gain sufficiently powerful CPUs and storage means. In effect, they are considered capable of providing useful resources. Second, since the number of mobile devices continuously increases, their aggregate resources cannot be considered negligible (Phan et al.). Again, mobility and device heterogeneity pose significant challenges, especially due to the mobility of the resource providers. Moreover, an important problem rises here: Since mobile devices are strictly personal and at the same time resource constrained, it is not a given whether their resources will be offered by the device owners for the sake of collaboration or not.

In a second-level classification, two fundamentally different architectures have been proposed in an effort to exploit resources relying on mobile devices. Their difference concerns the underlying networking topology.

Mobile Grids on Site

In mobile grids on site (Katsaros & Polyzos, 2007a; Kurkovsky & Bhagyavati, 2003; Kurkovsky, Bhagyavati, & Ray, 2004; Park et al., 2003; Phan et al., 2002), mobile devices residing in a well-defined area, such as a cell in cellular networks or a WLAN hot spot, are coordinated by a central entity (residing at the access point or base station, BS) in order to perform a task (computational grid). The advantage of this approach is that the BS does not suffer from the constraints imposed by mobility. Therefore, it is considered suitable to act as a mediator capable of hiding the heterogeneity of the participating devices from the requesting node, coordinating the overall execution of the submitted job, and even allowing the grid system to appear to the rest of the network as an ordinary grid node (Phan et al.). What is more, these networking environments are characterized by the concentration of a large number of mobile nodes yielding a potentially large amount of aggregated resources, usually under a single administrative domain (Katsaros & Polyzos).

Mobile Ad Hoc Grids

In the case of mobile ad hoc grids (Gomes, Ziviani, Lima, & Endler, 2007; Li et al., 2005; Marinescu, Marinescu, Ji, & Boloni, 2003), there is no stationary entity responsible for the coordination of the overall job execution. In environments such as MANETs, the absence of fixed nodes imposes difficulties in resource discovery, job scheduling, and monitoring. The instability of the network topology induces further difficulties due to unique ad hoc characteristics such as network partitioning and multihop routing. An approach toward overcoming this limitation is the formation of a virtual backbone consisting of a number of possibly more powerful mobile nodes responsible for coordinating the mobile nodes residing in a certain area of the overall ad hoc network (Li et al.; Marinescu et al.).

TOWARD THE REALIZATION OF A MOBILE GRID

Collaboration and Contribution

As implied earlier, the very essence of a mobile grid system depends on the actual availability of the existing mobile resources. Given that mobile devices are in principle personal devices with limited resources, at least in comparison to stationary ones, a critical issue concerns the willingness of the mobile users to offer their resources, that is, the willingness to collaborate. Therefore, a crucial target for the realization of mobile grids is to ensure this willingness, and this is subject to the incentives given to the owners of the mobile devices.

In the simplest case, the incentives may be inherent in the community of the collaborating mobile users; that is, mobile users may collaborate in order to achieve a common goal. However, mobile nodes may participate in a community where no common goal exists. In this case, mobile nodes would possibly engage in a mobile grid system to overcome their individual limitations, but it is not straightforward why they would also provide their own resources as well. Hence, in environments where users are considered rational (or selfish), proper incentive schemes are required to motivate collaboration. A simple approach, borrowed from the context of peer-to-peer systems, is based on reciprocity; that is, mobile users may consume resources offered by other mobile users as long as they also provide their own resources (Katsaros & Polyzos, 2007a). Apparently, a balance between consumed and offered resources must be struck for each participant in order to eliminate free riding (see also the "Security Issues" section).

This balance is strongly influenced by the fact that the willingness for collaboration is directly affected by the actual amount of the resources that are required to be offered. Special care must be given in order not to drain the already limited resources of mobile nodes as this would inevitably discourage users from participating in the mobile grid. In effect, the power of a mobile grid system must come from the aggregation of small amounts of resources from multiple mobile devices at a time. Additionally, mobile-device owners must be given the flexibility to control the degree of actual contribution to the mobile grid with respect to several criteria, such as available power and current device workload.

Operation in a Mobile Environment

As discussed earlier, the main target of a mobile grid system is to exploit the large number of participating mobile nodes by harvesting relatively small amounts of resources at a time from each of them. In the case of a computational mobile grid, what is desired is to take advantage of the parallel character of task execution in multiple nodes. The scheduler of the system is responsible for the decomposition of a submitted job into smaller tasks, which are then assigned to each available mobile node and executed in parallel. However, in a wireless and mobile networking environment, noise and intermittent connectivity impose delay and other overheads on the communication between the scheduler and the mobile nodes.

In this context, a fundamental design decision is whether a task executed by a mobile node must be aborted upon disconnection in the sense that it will be rescheduled, that is, submitted to another mobile node, or whether it must be carried out by the mobile node until connectivity with the scheduler is reestablished and the results can be returned back. In the architectures presented in Kurkovsky and Bhagyavati (2003) and Kurkovsky et al. (2004), the first approach is followed in view of the latency incurred by a handoff (the process of transferring a connection, established between a mobile node and the core network, to another point of attachment, i.e., another base station or access point). Obviously, even though recent research efforts have shown that, in certain networking environments, this approach yields better results (Katsaros & Polyzos, 2007a), this issue requires further investigation as it is subject to the connectivity characteristics of the participating mobile nodes. The bandwidth, storage, computational, and energy resources spent for the execution of a task are obviously wasted if the assigned task is aborted upon disconnection. From a system designer's point of view, this may lead to the underutilization of the available resources. From a mobile user point of view, this fact will lead to user annoyance unless a proper accounting mechanism ensures that he or she will indeed receive the credits for the already offered resources.

It is stressed at this point that the aforementioned approaches do not constitute the primary means for addressing the problem of intermittent connectivity. Other techniques such as load balancing and task replication (Katsaros & Polyzos, 2007b; Litke, Skoutas, Tserpes, & Varvarigou, 2007) can be employed in order to smooth the effects of the unstable networking environment. Moreover, it is noted that the above approaches are not mutually exclusive in a mobile grid system. For example, building connectivity profiles for each mobile node may assist an intelligent scheduler in dynamically choosing between these approaches.

Security Issues

In the considered context, distributed job execution comes with potential risks both for the integrity of the application and the resource provider.

The integrity of a mobile grid application is heavily affected by potential malicious mobile-node behavior. First, the integrity of the processing results is compromised insomuch as a malicious mobile node attempts to deliberately alter the results, possibly in a goal-oriented way. Second, the free-riding problem may reduce the overall system utility. Furthermore, input and output data confidentiality may also be a requirement. Promising research directions for the confrontation of these risks include the use of code and data encryption techniques (Sander & Tschudin, 1998), the introduction of redundancy in the system, and the employment of proper incentive schemes and reputation mechanisms (Kamvar, Schlosser, & Garcia-Molina, 2003) in order to discourage free riders.

On the other hand, there are serious security considerations regarding possible attacks performed by the mobile code against the mobile node on which the code is executed. To name a few of the risks, malicious mobile code may destroy files and applications, divulge confidential information, perform actions on behalf of the device owner without his or her consent and knowledge, and so forth. Possible countermeasures include the authentication of the mobile code dispatcher, code verification, and access control. The access rights of the mobile code on the hosting mobile node can be limited to a secure set of functions (i.e., the sandbox approach, a tightly controlled environment for the safe execution of untrustworthy code and programs). For example, the Java Virtual Machine (JVM) verifies byte-code properties before execution and additionally performs checks at run time.

FUTURE TRENDS

As already established, mobile grid computing is a new and promising research area. Even though current trends seem somehow vague, a clearer view of the research perspectives is becoming apparent as research continues.

First, the resource constraints of mobile devices impose the need for research on suitable schemes that will foster participation in the mobile grid. Initial approaches on this issue include game theoretic investigations of pricing strategies (Ghosh, Roy, & Das, 2007) as well as incentive schemes that tie the consumption of resources to the contribution of resources (Katsaros & Polyzos, 2007a). The personal character of mobile devices, with respect to ownership and usage, also requires further investigation so as to attain the

unobstructed personal use of a mobile device that participates in a mobile grid as a resource provider.

Furthermore, current research efforts focus on the feasibility of the mobile grid paradigm in mobile ad hoc networks (Gomes et al., 2007; Lima, Gomes, Ziviani, Endler, Soares, & Schulze, 2005; Zottl, Gansterer, & Hlavacs, 2007). The focus is on resource discovery and scheduling issues. Resource discovery has been widely studied in the context of ad hoc networks under the term *service discovery* (the process of finding the provider of a certain type of service or resource in a computer network; Gao, Yang, Zhao, Cui, & Li, 2006); however, under the scheduling requirements introduced by the mobile grid concept, the problem cannot be considered solved. The problem here is to discover a group of mobile nodes, rather than a single service provider, satisfying the resource requirements of the mobile grid. The discovery process may significantly vary, subject to these requirements and the nature of the application. For instance, the target may be to discover the suitable resource providers that will yield an optimal division of a job into independent tasks, with respect, for example, to the achieved response time. In another case, the desired outcome may be a set of resource providers that can efficiently handle the execution of interdependent tasks. Things may get more complex if the heterogeneity in resource provider characteristics and capabilities is taken into account. A further complication is introduced in multihop networks by data-intensive applications since collaboration is implicitly assumed by intermediate nodes that forward the volume of data to the actual computing nodes. In such scenarios, the selection of resource providers involves the intermediate nodes as well. The interconnection between resource discovery and scheduling is becoming apparent. Scheduling has recently drawn the attention of researchers, not only in the context of MANETs but also in networks with fixed infrastructure (Ghosh et al., 2007; Katsaros & Polyzos, 2007a). The main target here is to devise adequate scheduling policies that will try to conceal intermittent connectivity and yield enhanced performance, taking into account significant constraints such as energy limitations and human factors (e.g., personal device usage).

CONCLUSION

The emergence of the mobile grid paradigm has been characterized by the variance of the incipient research approaches. In this article, we have examined the commonalities and the differences among these approaches in an effort to unravel the strands of this research area. From our point of view, the ubiquity of mobile resources urges the investigation of resource sharing schemes in mobile communities of users. Toward this direction, we have highlighted the fundamental issues and problems emerging from the coalescence of the grid and the mobile computing paradigms.

REFERENCES

Banavar, G., Beck, J., Gluzberg, E., Munson, J., Sussman, J., & Zukowski, D. (2000). *Challenges: An application model for pervasive computing.* Paper presented at the Sixth Annual International Conference on Mobile Computing and Networking, Boston.

Foster, I. T. (2001). *The anatomy of the grid: Enabling scalable virtual organizations.* Paper presented at the Seventh International Euro-Par Conference on Parallel Processing, London.

Gao, Z.-g., Yang, Y., Zhao, J., Cui, J., & Li, X. (2006). Service discovery protocols for MANETs: A survey. In *Mobile ad-hoc and sensor networks* (Vol. 4325, pp. 232-243). Berlin, Germany: Springer.

Ghosh, P., Roy, N., & Das, S. K. (2007). *Mobility-based cost-effective job scheduling in an IEEE 802.11 mobile grid architecture.* Paper presented at the Seventh IEEE International Symposium on Cluster Computing and the Grid, Rio de Janeiro, Brazil.

Gomes, A. T. A., Ziviani, A., Lima, L. S., & Endler, M. (2007). *DICHOTOMY: A resource discovery and scheduling protocol for multihop ad hoc mobile grids.* Paper presented at the Seventh IEEE International Symposium on Cluster Computing and the Grid, Rio de Janeiro, Brazil.

Kamvar, S. D., Schlosser, M. T., & Garcia-Molina, H. (2003). *The eigentrust algorithm for reputation management in P2P networks.* Paper presented at the 12th International Conference on World Wide Web, New York.

Katsaros, K., & Polyzos, G. C. (2007a). *Optimizing operation of a hierarchical campus-wide mobile grid.* Paper presented at the 18th International Symposium on Personal Indoor and Mobile Radio Communications, Athens, Greece.

Katsaros, K., & Polyzos, G. C. (2007b). *Optimizing operation of a hierarchical campus-wide mobile grid for intermittent wireless connectivity.* Paper presented at the 15th IEEE Workshop on Local & Metropolitan Area Networks.

Kurkovsky, S., & Bhagyavati. (2003). *Wireless grid enables ubiquitous computing.* Paper presented at the 16th International Conference on Parallel and Distributed Computing Systems, Reno, NV.

Kurkovsky, S., Bhagyavati, & Ray, A. (2004). *A collaborative problem-solving framework for mobile devices.* Paper presented at the 42nd ACM Southeast Regional Conference, Huntsville, AL.

Li, H., Sun, L., & Ifeachor, E. C. (2005). *Challenges of mobile ad-hoc grids and their applications in e-healthcare.*

Paper presented at the Second International Conference on Computational Intelligence in Medicine and Healthcare.

Lima, L. d. S., Gomes, A. T. A., Ziviani, A., Endler, M., Soares, L. F. G., & Schulze, B. (2005). *Peer-to-peer resource discovery in mobile grids.* Paper presented at the Third International Workshop on Middleware for Grid Computing, Grenoble, France.

Litke, A., Skoutas, D., Tserpes, K., & Varvarigou, T. A. (2007). Efficient task replication and management for adaptive fault tolerance in mobile grid environments. *Future Generation Computer Systems, 23*(2), 163-178.

Litke, A., Skoutas, D., & Varvarigou, T. (2005). *Mobile grid computing: Changes and challenges of resource management in a mobile grid environment.* Paper presented at the Fifth International Conference on Practical Aspects of Knowledge Management, Vienna.

Marinescu, C. D., Marinescu, M. G., Ji, Y., & Boloni, L. (2003). *Ad hoc grids: Communication and computing in a power constrained environment.* Paper presented at the 22nd International Conference on Performance, Computing, and Communications, Phoenix, AZ.

Migliardi, M., Maheswaran, M., Maniymaran, B., Card, P., & Azzedin, F. (2002). Mobile interfaces to computational, data, and service grid systems. *Mobile Computing and Communications Review, 6*(4), 71-73.

Park, S.-M., Ko, Y.-B., & Kim, J.-H. (2003). *Disconnected operation service in mobile grid computing.* Paper presented at the First International Conference on Service Oriented Computing.

Phan, T., Huang, L., & Dulan, C. (2002). *Challenge: Integrating mobile wireless devices Into the computational grid.* Paper presented at the Eighth International Conference on Mobile Computing and Networking, Atlanta, GA.

Sander, T., & Tschudin, C. F. (1998). *Towards mobile cryptography.* Paper presented at the 1998 IEEE Symposium on Security and Privacy, Oakland, CA.

Srinivasan, S. H. (2005). *Pervasive wireless grid architecture.* Paper presented at the Second Annual Conference on Wireless on demand Network Systems and Services, St. Moritz, Switzerland.

Zottl, J., Gansterer, W. N., & Hlavacs, H. (2007). *A hierarchical two-tier information management architecture for mobile ad-hoc grid environments.* Paper presented at the Seventh IEEE International Symposium on Cluster Computing and the Grid, Rio de Janeiro, Brazil.

KEY TERMS

Accounting Mechanism: This is the process of recording (a summary of) the details of service consumption that usually follows successful authentication and authorization.

Free Riding: Free riding is a problem in P2P (peer-to-peer) systems, in which peers tend not to contribute resources in order to minimize their own costs while at the same time benefiting from the contributions of other peers.

Incentive Scheme: It is a mechanism used to provide the motive to the actors of a system for a certain behavior.

Load Balancing: Load balancing is the technique used to spread workload or resource consumption between many resource providers in order to improve resource utilization and/or performance.

Mobile Grid Computing: It is the computing paradigm for coordinated resource aggregation and sharing in which mobile computing devices act either as resource consumers or as resource providers or both.

Replication: Replication involves introducing redundant replicas of a resource in a system in order to improve reliability, performance, availability, and/or fault tolerance.

Reputation Mechanism: It is a mechanism used in P2P systems to associate the identities of peers with the opinions of other peers about the contribution of the former, and to acquaint the participating peers with this association. It is mostly used to discourage free riding.

A Model for Characterizing Web Engineering

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

The Internet, particularly the Web, has opened new vistas for many sectors of society, and over the last decade it has played an increasingly integral role in our daily activities of communication, information, and entertainment. This evidently has had an impact on how Web applications are perceived, developed, and managed.

The need to manage the size, complexity, and growth of Web applications has led to the discipline of Web engineering (Ginige & Murugesan, 2001). It is known (Kruchten, 2004) that conventional engineering practices cannot be simply mapped to software engineering without the engineer first understanding the nature of the software, and we contend the same applies to Web engineering. This article proposes a systematic approach to identify and elaborate the characteristics that make Web engineering a unique discipline, and considers the implications of these characteristics.

The rest of the article is organized as follows. We first outline the background and related work necessary for the discussion that follows, and state our position in that regard. This is followed by a model to uniquely posit the nature of Web applications based on the dimensions of project, people, process, product, and resources. Next, challenges and directions for future research are outlined. Finally, concluding remarks are given.

BACKGROUND

The notion of a Web application has evolved from its origins in the mid 1990s. For the sake of this article, by a Web application we will mean a Web site that behaves more like an interactive software system specific to a domain (such as health, entertainment, commerce, and so on) rather than a catalog. If the recent predictions (Jazayeri, 2007) are correct, then it is likely that the crosspollination of software engineering and Web applications will continue to flourish.

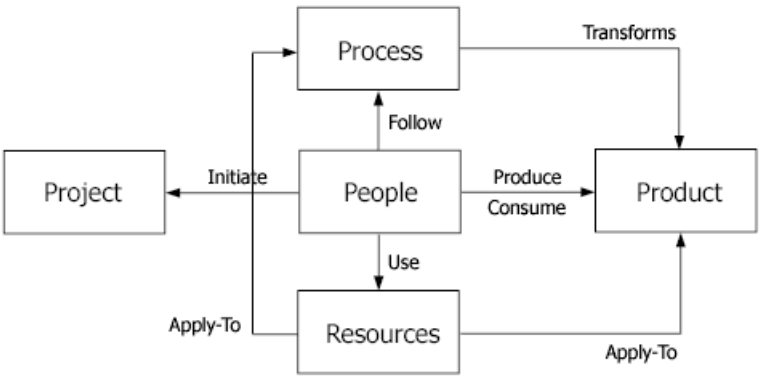
There has been some previous work that presents unique aspects of Web engineering, which we now discuss chronologically. It has been highlighted that Web applications differ from traditional software due to their focus on publishing, strong emphasis on quality attributes such as usability, and shorter initial delivery cycles (Overmyer, 2000). It has also been pointed out that the development of Web applications involves several social and technical disciplines and different sets of skills compared to conventional software development (Ginige & Murugesan, 2001), but stakeholders have not been considered as possessing one of the viewpoints. A model for the characterization of Web applications has been presented (Lowe, 2002), but details of individual characteristics are not given. It has also been noted that Web applications vary in many ways from traditional software including in the uncertainty of the domain, often shorter time to market, and rapid changes in technologies (Lowe, 2003; Ziemer & Stålhane, 2004); however, the arguments are often based on perception than technical reality. It has been pointed out that different types of Web applications vary along the lines of their nature, form, purpose, and development (Selmi, Kraïem, & Ghézala, 2005). An overview of the client-side properties of Web applications related to usability has been presented, and based on it, a more precise usability model has been derived (Bruno, Tam, & Thom, 2005). The variations between software engineering and Web engineering have been pointed out (Mendes & Mosley, 2006); however, the criteria focus on the development and underlying technologies rather than the stakeholders. Finally, the differences between Web applications and traditional software mentioned above have been recently amassed in a survey (Al-Salema & Samahab, 2007).

A MODEL FOR THE CHARACTERIZATION OF THE UNIQUE NATURE OF WEB ENGINEERING

In this section, we propose a model labeled henceforth as 4P+R for a characterization of the unique nature of Web engineering. The model, along with its high-level nonmutually exclusive elements, namely people, project, process, product, and resources, is illustrated in Figure 1.

The 4P+R model for Web engineering could be applied in a few different contexts. First, it could serve as a starting point for a reference model for Web engineering. Second, the existence of a body of knowledge is a sign of maturity of a discipline, and the 4P+R model could contribute to (and, once established, benefit from) the Web engineering body of knowledge (WEBOK), as is the case with the software engineering body of knowledge (SWEBOK) and the project management body of knowledge (PMBOK). Third, the 4P+R model could also be used as a basis for Web engineering pedagogy. In particular, it could be useful for deciding the

Figure 1. A high-level view of the elements of the 4P+R model of Web engineering



prerequisites and the selection of topics for an intensive course on Web engineering.

We do not claim that the 4P+R model is static or complete. Indeed, the model is subject to evolution along with the discipline of Web engineering and, indeed, the Web itself. We next discuss the elements of the 4P+R model in detail.

People Viewpoint

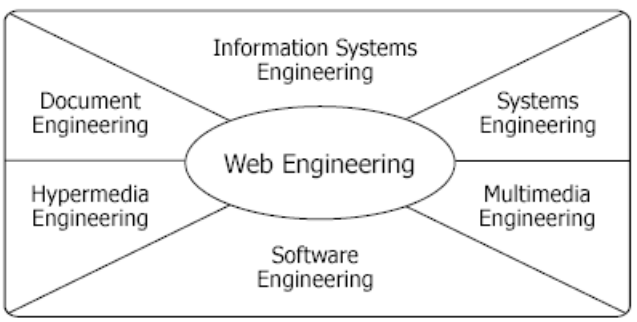
A stakeholder is a person or organization who influences a Web application or who is impacted by that Web application. In this section, we take the people view of a Web application and consider the challenges facing the stakeholders.

There are systematic approaches for the identification and refinement of stakeholder classes (Sharp, Galal, & Finkelstein, 1999). We identify two broad classes of stakeholders with respect to their roles in relation to a Web application, namely, a producer and a consumer. (There are other possible stakeholder classes such as legislators, but their characteristics are not unique to Web applications.) The mapping between stakeholders and roles is many to many. For a successful realization of the contract between producer and a consumer, the technical as well as the social differences between the development of traditional software and of Web applications need to be acknowledged and acted upon.

Producer

A producer (provider, project manager, marketing manager, engineer, media producer, graphic designer, or maintainer) is a person who owns, finances, develops, deploys, operates, or maintains the Web application. As shown in Figure 2, the desirable knowledge and skills (Kamthan, 2007) demanded of a producer go beyond what is part of the conventional training of a typical software engineer. Unfortunately, courses related to the Web offered at universities and training

Figure 2. The universe of engineering disciplines on which Web engineering draws upon



schools often tend to focus primarily on the manipulations of the moving target and popular client- and/or server-side technologies of the day rather than on the value added to the organization producing the Web application, human-oriented aspects of the Web application, or the fundamentals of analysis and design.

The scope in which a producer functions is limited by several factors, which often present unique challenges with respect to controllability and the domain of responsibility. Although the issues related to the network or domain name service are crucial, they are often not within the direct control and purview of an engineer. Furthermore, the use of hypermedia is at the heart of a Web application that strives for connection to the external information universe. However, a producer essentially also does not have any explicit control over the external resources pointed to by any (unidirectional) hyperlinks in the Web application. An engineer may perform timely checks but is not directly responsible for the integrity of external hyperlinks or for the content of the resources they point to (which can change without any notification). Unlike a desktop application, a producer essentially has little or no control over the consumer's environment: no explicit control over the end-user device or the user agent deployed by a user for accessing a Web application, and apart from some rudimentary data on the client side (such as the knowledge of the device, operating system, and user agent), little knowledge of the user preferences, particularly on a first-time use. An engineer is responsible for the information user interface of a Web application (Figure 3) but not for the user agent's user interface even though the rendering of the former (responsibility of the producer) intimately depends on the capabilities of the latter (external to the realm of responsibility of the producer).

Consumer

A consumer (novice or expert user) is a person who interacts with a Web application for some purpose. Aside from certain locale-specific legal restrictions, virtually any person of any age, of any cultural group, of any educational background, and so on could be a consumer of a Web application. This, at least in principle, creates a rather large consumer base for the producer of a Web application to contend with.

A consumer has absolute control over the choice of client-side computing environment, from its procurement to operation. However, a consumer has little or no control over the context in which a Web application operates or the availability of a Web application. For example, a Web application could be installed, upgraded, or even removed without prior notification.

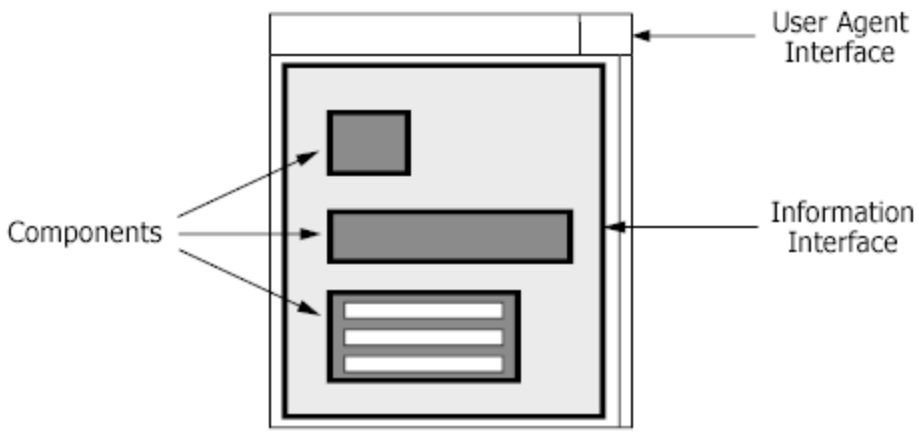
Project Viewpoint

In this section, we take the project view of a Web application. Based on that view, we consider the issues of legality and cost that are central to project management.

Legality

The stakeholders of a Web application and the Web application itself need not be colocated. For instance, they may be geographically located anywhere in the world: in different jurisdictions in the same country or in different countries. So, the laws that govern the producer and the consumer of that Web application may be different. This has direct implications toward the producers and consumers of commercial Web applications.

Figure 3. The user interface of a Web application explicitly depends on the capabilities of the user agent



Cost

The cost of production and delivery may be entailed with any type of software, including Web applications. However, there are differences between how cost and effort of conventional software and Web applications are estimated (Reifer, 2000). The equation has somewhat changed with the ascent of open-source software (OSS) and their widespread deployment in Web applications.

A Web application itself is never really purchased and owned by a consumer. However, there are three dimensions of cost for a consumer of a Web application: (a) connectivity to the Internet, (b), installation and use of client-side software to access the Web application, and (c) access to the Web application itself. Although many of the publicly available Web applications allow free-of-cost access, there are Web portals that charge for initial registration.

Process Viewpoint

In this section, we take the process view of a Web application and consider the issues underlying analysis and synthesis. The selection of a suitable development process for Web applications needs to consider several factors masquerading as potential risks, of which domain understanding and evolvability are critical.

Domain Understanding

A unique aspect of the Web is that it is open to any domain, and for some of these domains, interactive networked applications may never have been built before. The engineers have to deal with domains of which they do not have complete knowledge, and cannot elicit and specify all requirements at the inception of the project.

Evolvability

The frequent changes in the technological environment and their corresponding support in client-side devices and user agents can impact the requirements elicitation process.

Typically, a Web application, particularly if it provides time-sensitive information, requires frequent maintenance. At the same time, unlike most desktop applications, a Web application may have a predetermined, fixed life time known at the outset: Examples include Web applications for events such as a conference, the Olympics, or a music concert. This makes maintenance an integral, high-priority phase in the life cycle of a Web Application.

At the core of above-mentioned risks is unpredictability. Therefore, instead of developing the entire Web application in one cycle, it is preferable to build it iteratively and incrementally based on feedback from the customers and users. In particular, before moving on to implementation, frequent bidirectional exchange between requirements and design may be expected (Lowe, 2003; Zowghi & Gervsasi, 2001).

In recent years, agile methodologies (Highsmith, 2002) have emerged to mitigate risks by being human-centric (accommodating customers and users) and by introducing flexibility in the development processes. According to surveys (Khan & Balbo, 2005), agile methodologies have been successfully applied to Web applications. However, most agile methodologies are weak in their adoption of preventative approaches toward the improvement of the quality of a Web application.

Product Viewpoint

In this section, we take the product view of a Web application. Based on that view, we consider the issues of delivery, heterogeneity, and quality.

Delivery

The delivery context of a Web application is different from the stand-alone (desktop) software environment in many ways. Web applications are only delivered on request, in part, and one resource at a time, identified by a uniform resource locator (URL); they are not acquired and installed in entirety like desktop software.

Heterogeneity

A rich Web application has the ability to present information in a heterogeneous manner that includes text, mathematical symbols, graphics, audio, and video. For example, apart from its native information, a Web application may need to include advertisements from external sources. Indeed, markup languages based on the extensible markup language (XML) provide the capability to Web applications of being both document and data oriented.

The architecture of a Web application, by necessity, will require two or more subsystems to operate in a client-server environment, namely, a Web user agent (browser) and a Web server. Moreover, there may be other (not necessarily colocated or under the same ownership) subsystems such as an application server, multimedia server, or a database server. The quality of these subsystems will naturally impact the overall context in which a Web application operates.

Quality

There are different views of quality (Wong, 2006). From the ISO/IEC 9126-1:2001 standard, we define the quality of a

Web application to be the totality of characteristics that bear on its ability to satisfy stated and implied needs.

Although quality is a classical issue for most computer systems, there are unique quality-related concerns in the decentralized, heterogeneous networked environment of the Web. In particular, concerns related to certain quality attributes, including accessibility, interoperability, privacy, reliability, security, and usability are amplified in the context of a Web application. Indeed, low reliability, poor usability, and laxness in security have been identified as major causes of failures in Web applications (Pertet & Narasimhan, 2005). Also, managing the dichotomy of provision for personalization in the light of respecting privacy remains a major concern to many users (Paine, Reips, Stieger, Joinson, & Buchanan, 2007), where the benefits of respecting one can adversely affect the other. The conventional quality models such as the ISO 8402:1994 standard and the ISO/IEC 9126-1:2001 standard do not take into account all the aforementioned quality attributes and therefore are not sufficient for Web applications in their current form.

Resources Viewpoint

In this section, we take the resource view of a Web application. Based on that view, we consider the unique aspects that entail from it, particularly those related to reuse.

Every application will utilize resources during production and upon delivery. The resources during the production of a Web application could consist of expert entities of knowledge (principles, guidelines, and patterns), tools, and techniques. They are not fundamentally different from that of other applications except that some of them have yet to stabilize and mature.

Reuse

The distributed environment of the Web leads to a plethora of opportunities as well as to legal challenges associated with information reuse. For instance, the same database system could be used by multiple Web applications, and conversely, a single Web application could utilize multiple database systems. Although the request for a resource may have been made to a single specific address, not all or any parts of that resource that are delivered may originate from the same address.

Indeed, some of the parts of the resource may be hosted on external servers whose bandwidth is used without request. For example, this is possible via various forms of transclusion, such as framing (including and presenting information from external sources in a multiwindow document without permission from the original authors and without explicit knowledge of the users) or inlining images (including images from external sources without permission of the original authors).

FUTURE TRENDS

In this section, we consider two directions of evolution of the Web, namely, the Semantic Web and Web 2.0, and briefly outline their implications for the 4P+R model for Web engineering.

The Semantic Web has recently emerged as an extension of the current Web that adds technological infrastructure for better knowledge representation, interpretation, and reasoning (Hendler, Lassila, & Berners-Lee, 2001). Although there have been many advances toward enabling the technical infrastructure of the Semantic Web in recent years, there is much to be done in addressing the social issues (Lassila & Hendler, 2007). In particular, the cost-benefit ratio in the production of large domain-specific ontologies, the performance of these ontologies over the network, and the usability of query formulations on devices with restricted capabilities present but a few unique challenges (Kamthan & Pai, 2006) for a broad acceptance of Semantic Web applications.

The pseudonym Web 2.0 (O'Reilly, 2005) has been used to describe the apparent humanization and even socialization of the Web as it moves toward becoming a means of participation and collaboration. Indeed, applications like del.icio.us, Flickr, MySpace, Wikipedia, and YouTube are but a few examples of this phenomenon where a consumer becomes a coproducer in a social network. This has blurred the lines between the roles and responsibilities of a producer and a consumer with respect to the corrective and perfective maintenance of a Web application. It has also raised issues of legality and security and, in general, of the credibility of Web applications (Kamthan, 2007) not previously encountered. For example, the potential for the distribution of inaccurate medical information from unqualified sources has particularly had an acute impact on the user perception of health-related Web applications.

CONCLUSION

A Web application aims to provide some value to a stakeholder and it is the goal of Web engineering to help realize that. Although there have been many advances toward enabling the technological infrastructure of the Web in the past decade, there is much to be done in addressing the social challenges.

In conclusion, for a coherent evolution of the discipline of Web engineering, its unique nature needs to be approached systematically. To address that, we need to consider Web engineering from different, time-invariant viewpoints. The 4P+R model presents one direction in furthering the understanding of Web engineering. Although the model is subject to evolution, we believe that a suitable management of stakeholder dynamics as it manifests on both the client

and the server side will continue to play a crucial role in the model.

REFERENCES

Al-Salema, L. S., & Samahab, A. A. (2007). Eliciting Web application requirements: An industrial case study. *Journal of Systems and Software, 80*(3), 294-313.

Bruno, V., Tam, A., & Thom, J. (2005, November 21-25). *Characteristics of Web applications that affect usability: A review.* Paper presented at the 19th Conference of the Computer-Human Interaction Special Interest Group (CHISIG) of Australia on Computer-Human Interaction, Canberra, Australia.

Ginige, A., & Murugesan, S. (2001). Web engineering: An introduction. *IEEE Multimedia, 8*(1), 14-18.

Hendler, J., Lassila, O., & Berners-Lee, T. (2001). The Semantic Web. *Scientific American, 284*(5), 34-43.

Highsmith, J. (2002). *Agile software development ecosystems.* Addison-Wesley.

Jazayeri, M. (2007, May 19-27). *Trends in Web application development.* Paper presented at the 29th International Conference on Software Engineering (ICSE 2007), Minneapolis, MN.

Kamthan, P. (2007). Towards a systematic approach for the credibility of human-centric Web applications. *Journal of Web Engineering, 6*(2), 99-120.

Kamthan, P., & Pai, H.-I. (2006, May 21-24). *Human-centric challenges in ontology engineering for the Semantic Web: A perspective from patterns ontology.* Paper presented at the 17th Annual Information Resources Management Association International Conference (IRMA 2006), Washington, DC.

Khan, A., & Balbo, S. (2005, July 2-6). *Agile versus heavyweight Web development: An Australian survey.* Paper presented at the 11th Australian World Wide Web Conference (AusWeb 2005), Gold Coast, Australia.

Kruchten, P. (2004, April 13-16). Putting the "engineering" into "software engineering." Paper presented at the 15th Australian Software Engineering Conference (ASWEC 2004), Melbourne, Australia.

Lassila, O., & Hendler, J. (2007). Embracing "Web 3.0." *IEEE Internet Computing, 11*(3), 90-93.

Lowe, D. (2002, July 6-10). *Characterisation of Web projects.* Paper presented at the Eighth Australian World Wide Web Conference (AusWeb 2002), Sunshine Coast, Australia.

Lowe, D. (2003). Web system requirements: An overview. *Requirements Engineering, 8*(2), 102-113.

Mendes, E., & Mosley, N. (2006). *Web engineering.* Springer-Verlag.

O'Reilly, T. (2005). *What is Web 2.0: Design patterns and business models for the next generation of software.* O'Reilly Network.

Overmyer, S. P. (2000). What's different about requirements engineering for Web sites? *Requirements Engineering, 5*(1), 62-65.

Paine, C., Reips, U.-D., Stieger, S., Joinson, A., & Buchanan, T. (2007). Internet users' perceptions of "privacy concerns" and "privacy actions." *International Journal of Human-Computer Studies, 65*, 526-536.

Pertet, S. M., & Narasimhan, P. (2005). *Causes of failure in Web applications* (PDL Tech. Rep. No. PDL-CMU-05-109). Carnegie Mellon University.

Reifer, D. J. (2000). Web development: Estimating quick-to-market software. *IEEE Software, 17*(6), 57-64.

Selmi, S. S., Kraïem, N., & Ghézala, H. H. B. (2005, July 27-29). *Toward a comprehension view of Web engineering.* Paper presented at the Fifth International Conference on Web Engineering (ICWE 2005), Sydney, Australia.

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web revisited. *IEEE Intelligent Systems, 21*(3), 96-101.

Wong, B. (2006). Different views of software quality. In E. Duggan & J. Reichgelt (Eds.), *Measuring information systems delivery quality* (pp. 55-88). Idea Group.

Ziemer, S., & Stålhane, T. (2004, July 27). *The use of trade-offs in the development of Web applications.* Paper presented at the International Workshop on Web Quality (WQ 2004), Munich, Germany.

Zowghi, D., & Gervsasi, V. (2001, June 4-5). *Why is RE for Web-based software development easier?* Paper presented at the Seventh International Workshop on Requirements Engineering: Foundation for Software Quality (REFSQ 2001), Interlaken, Switzerland.

KEY TERMS

Agile Development: It is a philosophy that embraces uncertainty, encourages team communication, values customer satisfaction, vies for early delivery, and promotes sustainable development.

Delivery Context: It is a set of attributes that characterizes the capabilities of the access mechanism, the preferences of the user, and other aspects of the context into which a resource is to be delivered.

Personalization: It is a strategy enabling delivery that is customized to the user and the user's environment.

Quality: It refers to the totality of features and characteristics of a product or a service (such as a Web application) that bear on its ability to satisfy stated or implied needs.

Stakeholder: It is a person or organization who influences a Web application or who is impacted by that Web application.

Web Application: It is an application specific to a domain Web site that behaves more like an interactive software system. In general it will require programmatic ability on the server side and may integrate or deploy additional software (such as application servers, media servers, or database servers) for some purpose (such as dynamic delivery of resources).

Web Engineering: It is a discipline concerned with the establishment and use of sound scientific, engineering, and management principles, and disciplined and systematic approaches to the successful development, deployment, and maintenance of high-quality Web applications.

Modeling ERP Academic Deployment via Adaptive Structuration Theory

Harold W. Webb
The University of Tampa, USA

Cynthia LeRouge
Saint Louis University, USA

INTRODUCTION

Academic/industry collaboration can change learning processes and improve outcomes by integrating resources and creating opportunities not otherwise attainable (Wohlin & Regnell, 1999). However, each institution's culture and organizational objectives influence the collaborative relationships developed as advanced information technologies (e.g., computer aided software engineering tools, enterprise resource planning [ERP] systems, and database tools) are adopted. The challenge is to facilitate mutual understanding and acknowledge distinctions in addressing each organization's goals. The aim of these relationships is the appropriation of ERPs in a manner that enriches educational experiences, while providing industry benefit.

There are many quandaries associated with this phenomenon. How does the deployment of ERPs facilitate educational processes? To what degree should these resources be utilized? What tools and methods should be used? What is the role of the ERP vendor? Can academic independence be maintained?

Without a framework to identify relevant variables, it is daunting to begin to assess the impact of varying degrees of adoption, identify effective processes of deployment, and move toward assessing costs and benefits. Though some frameworks address academic/industry collaboration (Mead et al., 1999), few have considered the implications of ERPs on the evolution of inter-institutional collaborative relationships. This exposition augments a framework for understanding the forces at work when integrating ERPs into educational settings (LeRouge & Webb, 2002, 2005).

We begin our discussion by reviewing adaptive structuration theory (DeSanctis & Poole, 1994) as the foundation for the academic/industry ERP collaboration framework (LeRouge & Webb, 2002). We discuss academic/industry collaboration constructs and their relationships within the context of ERP systems and then integrate examples, findings, and issues from recent research.

USING AST TO MODEL ERP DEPLOYMENT IN THE ACADEMY

Adaptive structuration theory (AST), an extension of structuration theory (Giddens, 1982), has been used as a framework to study organizational change processes during advanced information technology adoption (Poole & DeSanctis, 1992). Adaptive structuration takes a socio-technical perspective. Human actors and organizational context are introduced within this perspective as moderators of technology impact. The adoption of an advanced technology, therefore, is a process of organizational change resulting from the mutual influence of the technology and social processes.

The premise at hand is that in academic settings, human actors and organizational context collectively moderate the processes by which ERPs are appropriated. Such dynamic processes affect not only institutional and industry outcomes resulting from the appropriation, but also the evolution of the relationship between industry and academia. The number of academic institutions adopting ERPs is increasing (Rosemann & Maurizio, 2005). However, use is not a perfect proxy for effectiveness, as ERPs serve some institutions better than others (Antonucci, Corbitt, Stewart, & Harris, 2004).

ERP system adoption within the context of colleges of business is of interest and has considerable impact for a number of reasons: market demand, level of commitment required, interdisciplinary functionality, and level of system sophistication. To provide insight, we reintroduce our AST-based model for organizing constructs and relationships for this phenomenon (see Figure 1). We augment this model and understanding by providing recent research examples, findings, and issues related to construct attributes (provided in Tables 1 through 9).

Advanced Information Technology Structure

Two ways to describe contributing social structures offered by advanced information technologies are "structural features," referring to the types of rules and resources embedded in the system, and "spirit," the intended purpose and utilization of

Table 1. Advanced information technology (ERP) structure and spirit attributes associated with ERP curriculum

Attributes (LeRouge & Webb, 2002)	Examples, Findings, and Issues Related to Construct
♦ *Structural features*—(Restrictiveness and Comprehensiveness) ♦ Capturing Spirit	EXAMPLES • *Capturing spirit*: Create learning modules focused on decision-making using data-rich business processes in an ERP environment—not software features (Strong, Johnson, & Mistry, 2004) • *Working within structure*: Students run a range of ERP functions in complex multi-semester business simulations (Draijer & Schenk, 2004) FINDINGS • *Working within structure*: Ensure students understand the business environment and avoid the "not seeing the forest for the trees" problem (Fedorowicz, Gelinas, Usaff, & Hachey, 2004) • *Capturing spirit*: Student background in understanding underlying business processes is critical to ERP classroom success (Rosemann & Maurizio, 2005) ISSUES • *Structural Features*: Complexity of ERP subject matter is a challenge (Rosemann & Maurizio, 2005) • *Capturing Spirit*: Recognizing there is a curriculum gap in engineering-based information technology programs—a failure to emphasize business processes (Peslak, 2005)

the system (DeSanctis & Poole, 1994). Regarding structural features, an ERP is a comprehensive database structured to support diverse organizational processes through a large number of application modules. Each module is geared toward a functional or industry-specific process. ERP systems challenge colleges with a level of cross-discipline sophistication and flexible feature sets that require substantial training.

The spirit of ERP systems can be described as the intention to process operational level transactions, support multi-level decisions, and aid strategic management of major corporations. The goals of ERP use in colleges of business are primarily educational and exploratory in nature and often focus on discipline-specific subsystems rather than cross-discipline business processes. With respect to technology spirit, the potential exists for a gap in appropriation between business use and academic use of ERP systems. This gap in system goals and values may have implications for academic/industry collaboration unless innovative cross-discipline approaches are taken as shown in Table 1.

External Environmental Structure

The demand for graduates to work with ERP systems is strong and acts as an external structuring force (Rosemann & Maurizio, 2005). Other external sources of influence are detailed curriculum guides for computer science (Lidtke & Stokes, 1999) and information systems (Gorgone & Gray, 2000). Table 2 highlights critical success factors and issues related to external structure.

Technology Infrastructure

Technology infrastructure is a major cost consideration for academic adoption of ERP systems (Becerra-Fernandez, Murphy, & Simon, 2000; Watson & Schneider, 1999). These systems typically cannot be deployed or maintained without considerable support. Industry may facilitate appropriation through donated services; however, colleges face additional costs. Table 3 lists collaborative relationships such as SAP's University Competence Center (UCC) hosting program, which is becoming a critical success factor for academic adoption (Rosemann & Maurizio, 2005).

Educational Organization Structure

The philosophies of appropriating ERP systems among educational entities vary widely. The overall philosophic quandary involves balancing conceptual technology education and the development of technology-specific skills. Table 4 identifies a number of critical success factors affecting educational organization structure.

Education Process

Structuration has at its core motivated and practical actions. Rules and resources embodied in social institutions are appropriated by participants and enter into the production and reproduction of a social system (Poole & DeSantis, 1992). Academic/industry collaborative interaction is embodied in the appropriation of the ERP into the educational process. Educators determine the curriculum strategy and the degree of appropriation for ERP systems. While the degree of appropriation has been addressed by academic institutions in a variety of ways ranging from inclusion of exemplary material within courses to new course creation to establishing new degree programs, issues related to the educational process remain (Table 5).

Table 2. External environmental structure attributes associated with ERP curriculum

Attributes (LeRouge & Webb, 2002)	**Examples, Findings, and Issues Related to Construct**
♦ Accreditation standards & curriculum studies ♦ Technology vendor market position ♦ Industry standards & technology trends ♦ Technology market competition & end user demands ♦ Technology-enabled labor supply ♦ Student interest (added 2006)	FINDINGS • *Technology-enabled labor supply*: Favorable job prospects is a top five learning critical success factor (CSF) (Rosemann & Maurizio, 2005) • *Student interest*: High student interest is a top five teaching CSF (Rosemann & Maurizio, 2005) • *Industry standards and technology trends*: Positive industry reaction to ERP specific knowledge (Rosemann & Maurizio, 2005) • *Accreditation standards and curriculum studies*: Emergence of ERP education centered conferences (Fedorowicz et al., 2004) ISSUES • *Accreditation standards and curriculum studies*: Lack of an established body of academic texts (Fedorowicz et al., 2004) • *Industry standards & technology trends*: Adaptation of vendor provided training material and courses require university resources including faculty time and travel (Fedorowicz et al., 2004)

Table 3. Technology infrastructure attributes associated with ERP curriculum

Attributes (LeRouge & Webb, 2002)	**Examples, Findings, and Issues Related to Construct**
♦ Software & hardware ♦ Internal maintenance & software support ♦ Database creation and maintenance ♦ Computer lab facility & student remote access ♦ Industry monetary donation or grants to support technology infrastructure ♦ Industry donation of ERP technician expert time (added 2006)	EXAMPLES • *Database creation and maintenance*: o SAP UCC hosting center with test and operational ERP instances for simulation development and testing (Draijer & Schenk, 2004). o Hosted solutions are increasing in popularity (Rosemann & Maurizio, 2005) o Hosting recommended to outsource non-competencies (Fedorowicz et al., 2004) FINDINGS • *Internal maintenance and software support*: Faculty support is a top five teaching CSF (Rosemann & Maurizio, 2005) • *Computer lab facility & student remote access*: Resourcing the technical infrastructure is one of the top four critical challenges (Strong et al., 2004) • *Internal maintenance and software support*: Systems support is a top five teaching CSF (Rosemann & Maurizio, 2005) • *Industry donation of ERP expert time*: Maintain technical support relationships with ERP experts (Fedorowicz et al., 2004)

Appropriation and Delivery Structure

The appropriation structure leads to instructional design choices. ERP systems are adaptable to many models of learning. The model of learning chosen may affect the instructor's approach and utilization of these tools. Conceptual presentation and demonstration may adequately support knowledge-level learning objectives, while experiential learning models may better support higher-order learning objectives (Leidner & Jarvenpaa, 1995). The application of appropriate learning models remains a challenge (see Table 6).

Emergent Forms of Educational Method

One emergent educational method is the use of problem-based learning. In one situation, graduate students work with industry partners to design ERP solutions for a business problem. Completed solutions are used as the basis of teaching cases to support undergraduate courses (Stewart & Rosemann, 2001). Table 7 highlights other innovative approaches and outstanding issues.

Joint Outcomes

The purpose of ERP appropriation and academic/industry collaboration is to achieve mutually beneficial joint outcomes (Stewart & Rosemann, 2001). The desired joint outcomes may include facilitating the educational mission, gaining competitive advantage, accessing educational resources, enhancing reputation, increasing revenue, and providing a staffing source (Mead et al., 1999). The academic institution, industry, or both may desire these goals. However, joint

Table 4. Educational organization structure attributes associated with ERP curriculum

Attributes (LeRouge & Webb, 2002)	**Examples, Findings, and Issues Related to Construct**
♦ Departmental structure ♦ Major program requirements ♦ Course objectives ♦ Instructor preferences ♦ Interdisciplinary cooperation (added 2006)	EXAMPLES • *Interdisciplinary cooperation*: Process-oriented ERP business simulations integrating multiple disciplines across academic department lines (Draijer & Schenk, 2004) • *Major program requirements*: Enterprise decision-making modules in integrated curricula (Strong et al., 2004) FINDINGS • *Instructor preferences*: Lecturer training is a top five teaching CSF (Rosemann & Maurizio, 2005) • *Departmental structure*: o Strong academic leadership and support is one of four critical challenges (Strong et al., 2004) o Need to share successful teaching materials among faculty (Fedorowicz et al., 2004) • *Interdisciplinary cooperation*: Faculty motivation and commitment is one of four critical challenges to cross-discipline faculty collaboration (Strong et al., 2004)

Table 5. Education process attributes associated with ERP curriculum

Attributes (LeRouge & Webb, 2002)	**Examples, Findings, and Issues Related to Construct**
♦ Learning models (collaborative learning, hands-on experience, simulations, conceptual presentations, programmed instruction, real-world exposure, case studies) ♦ Supporting technologies and resources (textbooks on technology, presentation tools, asynchronous communication tools, synchronous communication tools, computer-based training modules)	FINDINGS • *Supporting technologies and resources*: o Course material is a top five teaching and learning CSF (Rosemann & Maurizio, 2005) o Provide much support to students working on ERP assignments: face-to-face, e-mail, discussion lists, and online FAQ (Fedorowicz et al., 2004) • *Learning models*: o Learning approach is a top five learning CSF (Rosemann & Maurizio, 2005) o Mix of models: introductory hands-on exercises, routine uses of ERP in business operations via simulation, project development to implement changes in the ERP simulation environment (Draijer & Schenk, 2004) o Configuration matrices are suggested as a mechanism to specify how learning objectives are to be achieved in a specific learning context (LeRouge & Webb, 2004) ISSUES • *Learning models*: Most course modules are not in depth and do not cover specialized ERP capabilities (Rosemann & Maurizio, 2005) • *Supporting technologies and resources*: Vendor training materials are often outdated, too "hands-on," and lacking conceptual foundation (Rosemann & Maurizio, 2005)

Table 6. Appropriation and delivery structure attributes associated with ERP curriculum

Attributes (LeRouge & Webb, 2002)	**Examples, Findings, and Issues Related to Construct**
♦ Appropriation moves (direct use, relate to other structures, interpretation of structures, or judgment of features) ♦ Faithfulness ♦ Instrumental uses ♦ Attitude	FINDINGS • *Faithfulness*: Practical application is a top five learning CSF (Rosemann & Maurizio, 2005) • *Appropriation moves*: o Ability to change organizational structures during simulations to accommodate new technology and emerging theory is a strength (Draijer & Schenk, 2004) o By integrating the concerns-based adoption model from the field of education with AST, patterns of appropriation for a given instance can be measured that may impact outcomes, and be affected by a given profile that describes the state of AST structure constructs (LeRouge & Webb, 2004) ISSUES • *Instrumental uses*: Confounding technical implementation expertise with user expertise is one of the top four critical challenges. ERP should be used to link functional area concepts, and processes (Strong et al., 2004)

Table 7. Emergent forms of educational method attributes associated with ERP curriculum

Attributes (LeRouge & Webb, 2002)	Examples, Findings, and Issues Related to Construct
♦ Educators enrolling in corporate training programs ♦ Project/task specific internships ♦ Industry experts participating in classroom presentation ♦ Students/educators participating in ERP specific list serves ♦ Credit and/or increased access to technology training programs for students ♦ Industry development of targeted educational tools, databases, and exercises ♦ Phased learning (added 2006)	EXAMPLES • *Phased learning*: o Step-wise progression of roles of students introduced into increasingly complex set of interrelated business simulations (Draijer & Schenk, 2004) o Use of SDLC concept with multiple interactive exercises (case analysis, ERP failure presentation, and hands-on) (Grenci & Hull, 2004). • *Industry experts participating in classroom presentation*: o Include industry professionals integrated into curriculum as guest speakers (Fedorowicz et al., 2004) o Industry professionals as system consultants (Rosemann & Maurizio, 2005) FINDINGS • *General issue*: Integrating enterprise systems alters both course content and pedagogy (Fedorowicz et al., 2004) ISSUES • *Educators enrolling in corporate training programs*: Vendor ERP-specific training not directly transferable to university classes—more how to than why (Fedorowicz et al., 2004) • *Industry development of targeting educational tools*: Faculty must be willing to ask for help in adapting ERP materials (Fedorowicz et al., 2004)

Table 8. Joint outcomes attributes associated with ERP

Attributes (LeRouge & Webb, 2002)	Examples, Findings, and Issues Related to Construct
♦ Student learning/education in technology arena ♦ Increased work pool & employable students ♦ ERP market exposure ♦ Contribution to industrial research and development effort & academic research ♦ Continued/enhanced program attractiveness	EXAMPLES • *Increased work pool & employable students*: Favorable industry response—hiring students from a more sophisticated labor pool (Rosemann & Maurizio, 2005) • *Student learning/education in technology arena*: Students report learning system complexity while experiencing latest software (Draijer & Schenk, 2004) FINDINGS • *Student learning/education in technology arena*: Proposed ERP maturity model with metrics for assigning maturity level and maturity focus to a given curriculum (Antonucci et al., 2004) • *Continued/enhanced program attractiveness*: Outcomes generate feedback loops that sustain appropriation patterns or lead to adaptive change in level of use, structural variables, or both (LeRouge & Webb, 2004) ISSUES • *Continued/enhanced program attractiveness*: Recognition that some ERP implementations are foundering while others flourish (Antonucci et al., 2004) • *Student learning/education in technology arena*: Must manage both faculty and student expectations (Fedorowicz et al., 2004) • *Increased work pool & employable students*: Competitive advantage for colleges of business (Draijer & Schenk, 2004) • *Student learning/education in technology arena*: Metrics needed to evaluate effectiveness (Al-Mashari, 2003)

outcomes from academic appropriation are not guaranteed (see Table 8).

Structure of Academic/ Industry Collaboration

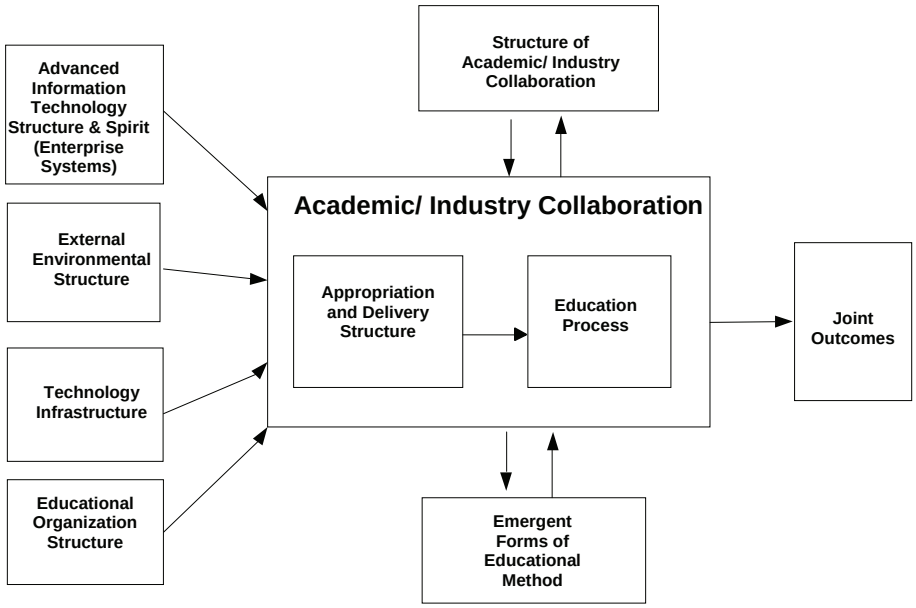
The collaborative system is not a recognized organization, but a structured social practice of interdependence that has broad spatial and temporal extension (Giddens, 1982). The existence of ERP alliance programs may be considered a representation of social practices affecting the relationship created between industry and academia. Representations of the implied social practice may be found in industry alliance program agreements and curriculum guides encouraging industry participation and the study of ERP system concepts (Gorgone & Gray, 2000; Lidtke & Stokes, 1999).

Practices suggested for a successful collaboration (Powell, Diaz-Herrera, & Turner, 1997) include centralized coordination, right mix of knowledge and experience, cooperative planning, curriculum flexibility, communica-

Table 9. Structure of academic/industry collaboration attributes associated with ERP curriculum

Attributes ***(LeRouge & Webb, 2002)***	**Examples, Findings, and Issues Related to Construct**
♦ Rules (*industry participation in curriculum development studies; inclusion of ERP in curriculum development research; academic participation in industry development; educator participation in corporate training programs)* ♦ Resources (*technology alliance programs; opportunities for field research*)	EXAMPLES • *Resources*: o SAP UCC program (Draijer & Schenk, 2004) o Academic alliances: SAP, Oracle, Microsoft • *Rules*: o Collaboration with multiple universities on an international level to develop advanced courses (Draijer & Schenk, 2004) o Collaborative curriculum development teams across institutions (Al-Mashari, 2003; Stewart & Rosemann, 2001) • ISSUES • *Resources*: Academic alliance membership is becoming increasingly difficult (Rosemann & Maurizio, 2005) • *Rules*: Most universities do not collaborate with other institutions (70%) but are interested in doing so (76%) (Rosemann & Maurizio, 2005)

Figure 1. Adaptive Structuration Theory Applied to Industry/Academic Collaborations adapted from DeSanctis and Poole (1994) (LeRouge & Webb, 2002)



tion, and objectivity. Mead et al. (2000) argue for the need to establish metrics to monitor and modify collaborative processes. Table 9 highlights trends in academic-industry collaboration structures.

FUTURE TRENDS

The appropriation of ERPs in colleges of business is a modern phenomenon that aspires to bridge the industry-academic gap while fulfilling educational goals. From a practical perspective, industry and students often desire opportune practical training and education from academic institutions (Stewart & Rosemann, 2001). However, the costs of appropriation may be high and the impact of appropriation on educational processes and collaborative relationships may be either positive or negative (Antonucci et al., 2004). Stakeholders should recognize the potential influence of structure and social context on desired outcomes when embarking on the process of academic/industry collaboration. ERP appropriation and associated collaboration decisions may affect the educational foundation and career prospects of the technological work force (Rosemann & Maurizio, 2005). Research studies cited support the 2002 model constructs (Figure 1). Further research is needed to evaluate relationships among the constructs and attributes identified in the

framework, refine metrics, and determine the effect on both academe and industry.

CONCLUSIONS

Adaptive structuration recognizes that technology appropriation may be a key factor in the evolution of affected social structures. Adaptive structuration theory gains predictive and descriptive power by identifying social complexities and relationships associated with the adoption of advanced information technologies. The major contribution of this exposition is further specification and support of an AST-based model to better understand the dynamics of academic/industry collaboration as they evolve on campus and in the workplace.

REFERENCES

Al-Mashari, M. (2003). Enterprise resource planning (ERP) systems: A research agenda. *Industrial Management + Data Systems, 103*(1/2), 22-27.

Antonucci, Y. L., Corbitt, G., Stewart, G., & Harris, A. L. (2004). Enterprise systems education: Where are we? Where are we going. *Journal of Information Systems Education, 15*(3), 227-234.

Becerra-Fernandez, I., Murphy, K., & Simon, S. J. (2000). Integrating enterprise in the business school curriculum. *Communications of the ACM, 43*(4), 1-4.

DeSanctis, G., & Poole, M. S. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organizational Science, 5*(2), 121-147.

Draijer, C., & Schenk, D. (2004). Best practices of business simulation with SAP R/3. *Journal of Information Systems Education, 15*(3), 261-265.

Fedorowicz, J., Gelinas, U. J. Jr., Usaff, C., & Hachey, G. (2004). Twelve tips for successfully integrating enterprise systems across the curriculum. *Journal of Information Systems Education, 15*(3), 235-244.

Giddens, A. (1982). *Profiles and critiques in social theory*. Berkeley, CA: University of California Press.

Gorgone, J. T., & Gray, P. (2000). MSIS 2000 model curriculum and guidelines for graduate degree programs in information systems. *Communication of the AIS, 3*(1), 1.

Grenci, R. T., & Hull, B. Z. (2004). New dog, old tricks: ERP and the systems development life cycle. *Journal of Information Systems Education, 15*(3), 277-286.

Leidner, D. E., & Jarvenpaa, S. L. (1995). The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly, 19*(3), 265-292.

LeRouge, C., & Webb, H. W. (2002). Theoretical foundations for enterprise systems technology collaborations: An adaptive structuration framework. In J. Lazar (Ed.), *Managing IT/community partnerships in the 21st century* (pp. 178-203). Hershey, PA: Idea Group Publishing.

LeRouge, C., & Webb, H. W. (2004). Appropriating enterprise resource planning systems in colleges of business: Extending adaptive structuration theory for testability. *Journal of Information Systems Education, 15*(3), 315-326.

LeRouge, C., & Webb, H. W. (2005). Modeling ERP academic deployment via AST. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 1989-1995). Hershey, PA: Idea Group.

Lidtke, D. K., & Stokes, G. E. (1999). An information systems-centric curriculum '99. *The Journal of Systems and Software, 49*(2, 3), 171-176.

Mead, N., Beckman, K., Lawrence, J., O'Mary, G., Parish, C., Unipingco, P., et al. (1999). Industry/university collaborations: Different perspectives heighten mutual opportunities. *The Journal of Systems and Software, 49*(2, 3), 155-162.

Mead, N., Unpingco, P., Beckman, K., Walker, H., Parish, C. L., & O'Mary, G. (2000, May). Industry/university collaborations. *Crosstalk: The Journal of Defense Software Engineering March 2000*. Retrieved April 7, 2004, from http://www.stsc.hill.af.mil/crosstalk/2000/03/mead.html

Peslak, A. R. (2005). Incorporating business processes and functions: Addressing the missing element in information systems education. *Journal of Computer Information Systems, 45*(4), 56-61.

Poole, M. S., & DeSanctis, G. (1992). Micro level structuration in computer-supported group decision making. *Human Communication Research, 19*(1), 5-49.

Powell, G. M., Diaz-Herrera, L., & Turner, D. J. (1997). Achieving synergy in collaborative education. *IEEE Software, 14*(6), 58-65.

Rosemann, M., & Maurizio, A. A. (2005). SAP-related education—Status quo and experiences. *Journal of Information Systems Education, 16*(4), 437-453.

Stewart, G., & Rosemann, M. (2001). Industry-oriented design of ERP-related curriculum—An Australian initiative. *Business Process Management, 7*(3), 234-242.

Strong, D. M., Johnson, S. A., & Mistry, J. J. (2004). Integrating enterprise decision-making modules into undergraduate

management and industrial engineering curricula. *Journal of Information Systems Education, 15*(3), 301-313.

Watson, E. E., & Schneider, H. (1999). Using enterprise systems in education. *Communications of the Association for Information Systems, 1*(9), 1-48.

Wohlin, C., & Regnell, B. (1999). Strategies for industrial relevance in software engineering education. *The Journal of Systems and Software, 49*(2, 3), 124-134.

KEY TERMS

Advanced Information Technology Structure: Rules and resources offered by systems such as enterprise resource planning systems that support the intended purposes and utilization of those systems.

Appropriation and Delivery Structure: Rules and resources that determine the choices made by educators regarding strategies for integrating learning model(s) and supporting technologies within a selected instructional design.

Educational Organization Structure: Rules and resources offered by the internal educational institution, that are derived from how it is organized as well as program requirements, curriculum, and course objectives.

Education Process: Use of learning models and supporting learning technologies to deliver the learning experience and/or training with students.

Emergent Forms of Educational Method: New methods and techniques employed or used in the education process.

External Environmental Structure: Rules and resources offered by outside interests including academic standards bodies, technology developers, industrial organizations, employers, and end users.

Joint Outcomes: The direct output and by-products of the education process including student learning, employable work force, market exposure, and contributions to research.

Structure of Academic/Industry Collaboration: Representation of social practices among the stakeholders affecting academic/industry collaboration that result in the establishment of rules of practice and the provision of resources.

Technology Infrastructure: Required supporting activities and facilities including network, hardware, software, development, and maintenance.

Modeling for E-Learning Systems

Maria Alexandra Rentroia-Bonito
Instituto Superior Técnico/Technical University of Lisbon, Portugal

Joaquim Armando Pires Jorge
Instituto Superior Técnico/Technical University of Lisbon, Portugal

INTRODUCTION

Computer-based instruction is touted as an effective tool to support knowledge dissemination within predefined learning environments. Indeed, many see it as a way to overcome geographical or social barriers to knowledge transmission and educational institutions. However, its domain of application has traditionally been restricted to basic skills and educational contexts. Recently, dynamic and complex business environments shaped by technological changes and the downsizing trend of the '90s placed new constraints on the underlying assumptions (Fuglseth, 2003). Organizations are now pushing for skill flexibility, demanding specialized knowledge and requiring faster learning curves from employees. Many advocate Internet-based education materials as one way to meet those challenges (Bernardes & O'Donoghue, 2003; Karoulis et al., 2004; Storey et al., 2002; Strazzo & Wentling, 2001). However, this raises important questions concerning both effectiveness and efficiency of such tools and materials. Indeed, developing interactive multimedia-based courseware remains pretty much a black art, consuming enormous resources. So far, there is a lack of established models to predict the performance and evaluate how adequately courseware can meet user needs. In fact, developing courseware should take into account the target constituency requirements, organizational context, and the stated educational or training goals. Developing the wrong training materials can lead to costly investments in creating and maintaining content to match the increasing expectations on e-learning. Perhaps this can explain the recent rash of failed e-learning projects—current results do not measure up to business and individual expectations yet.

A better understanding of the many factors affecting e-learning performance would allow individuals and organizations to achieve the expected benefits. In so doing, development teams need methods, techniques, and tools to evaluate in advance which features are needed to achieve higher outcomes, namely, performance and satisfaction. Thus, the need to develop predictive models to improve learning effectiveness is in order.

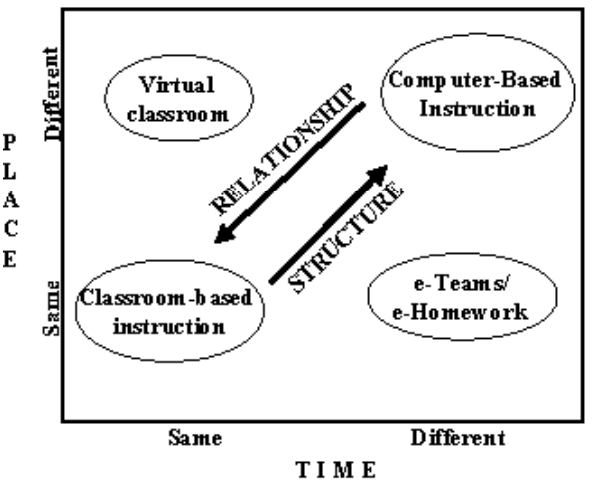
This overview includes four sections. "Background" presents a proposed e-learning theoretical framework to guide our analysis based upon the reviewed literature. "Key Issues" section describes main issues arising from the proposed e-learning conceptual framework. "Future Trends" describes our vision on how to approach e-learning initiatives and future trends. Finally, we present a general conclusion.

BACKGROUND

Organizational investment in e-learning strategies reflects strategic choices regarding skill development through e-learning. According to Wentling, Waight et al. (2000), e-learning involves acquiring and using distributed knowledge facilitated by electronic means in synchronous or asynchronous modes. As shown in Figure 1, knowledge could be distributed geographically within varying time frames.

Thus, the effective use of technology-based instruction would provide to organizations the ability to succeed at operational levels. This justifies the adoption of a holistic approach to courseware evaluation as a diagnostic and managerial tool. We propose a framework, shown in Figure 2, which comprises three basic entities, business processes, people, and information systems, and three main relationships: (a) interaction between people and systems, (b) process-based

Figure 1. Proposed types of e-learning in terms of time and place



roles played by people during this interaction, and (c) having the learning task be executed, as part of the e-learning experience, by people performing their process-based roles. This framework could lead to working techniques and approaches that assist development team members in designing work-related e-learning experiences within organizational contexts. To motivate a workable approach, we will now discuss each of these entities and relationships.

Reviewed literature strongly suggests that the external and internal fit among business strategies, culture, human resource practices, and leadership styles is critical to worker performance. Moreover, work contexts, for example, physical and technological conditions surrounding individual tasks, affect people's perceptions and, in turn, influence their motivation to engage into and perform learning tasks (Astleitner, 2001; Bandura, 2000; Chen, 2002; Dix et al., 1998; Kim, 2000; Liu & Dean, 1999; Reeves & Nass, 1996; Strazzo & Wentling, 2001; Vouk et al., 1999; Welbourne et al., 2000; Wentling et al., 2000).

Within the e-learning experience, business processes provide yardsticks to define educational or training goals and monitor outcomes. However, we need also to consider the roles people perform when interacting with courseware. Such process-based roles could be as diverse as e-learners, e-instructors, e-speakers, systems and courseware designers, supervisors, reviewers, human resource managers, and information technology officers among many others.

Human-computer interaction can model parts of the e-learning experience in accordance with Norman's extended model (Dix et al., 1998). Furthermore, the experience is also shaped by the way people relate to systems. This is supported by Reeves' and Nass' (1996) work, which suggests that people relate to media as they would relate to real people, treating them with affection and courtesy. Building on these findings, we argue that the more e-learning systems themselves are easy to use and learn and are "nicely behaved," the likelier e-learners will engage in the experience and profit from their outcomes.

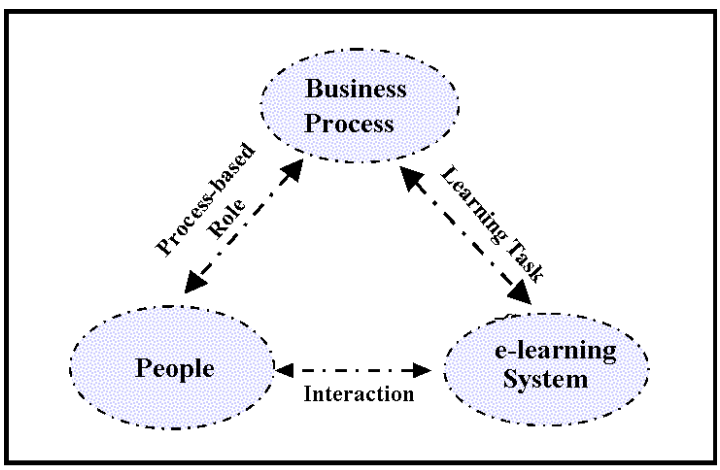
The interplay among these three relationships (process-based role, learning task, and interaction) relates to a just-in-time learning concept. Strategic knowledge acquisition should be enmeshed in current activities to support employees in learning new skills when performing day-to-day business tasks. We believe this concept can foster gradual alignment between learning outcomes, and technology with strategic aspects of business.

KEY ISSUES

We identify structure and relationship as the main issues within our framework as presented in the previous section. Figure 1 shows different modes of e-learning regarding the use of technology in education, both in terms of distance and time. As technology gets more extensively used for delivery, the need for course structure becomes higher and the relationship between instructor and e-learner turns increasingly weaker. Figure 1 also shows this relationship as defining three types of e-learning, which are set apart from conventional classroom instruction.

This shows that using technology to support learning requires higher course structure than traditional classroom-based instruction to be effective (Karoulis et al., 2004; Liu & Dean, 1999). However, current approaches take a one-size-fits-all method to provide courseware delivery regardless of differences in place and time. We cannot argue strongly enough that delivery needs to be tailored to context (space and time) to overcome the barriers imposed by structure and to improve the e-learning experience. This should be done differently for different students with diverse cognitive styles, roles, and tasks within organizational contexts. We will now discuss factors affecting structure and relationship.

Figure 2. Proposed e-learning framework



As for structure, organizations identify training needs taking into account work context, business process dynamics, individual tasks, objectives, and areas for performance improvement. A business-process approach driving the design and the development of interactive course contents should focus on skill gaps to define instructional objectives in order to meet performance standards. In this way, setting up appropriate goals for training evaluation poses the same requirements both for electronic and traditional media. However, as courseware becomes available and distributed through the Internet, quality of service (QoS) becomes an increasingly important factor to e-learner satisfaction. Thus, technology becomes another structural issue.

From the technology standpoint, three aspects are critical to e-learner satisfaction. The first is courseware evaluation (Chen, 2002; Karoulis et al., 2004; Kim, 2000; Liu & Dean, 1999; Storey et al., 2002; Strazzo & Wentling, 2001; Wentling et al., 2000). Indeed, users' perceptions of mismatch between content and structure reduce their motivation to learn and perform (Astleitner, 2001). Usability is a second aspect affecting both engagement and acceptance. It measures the extent to which a computer system can be used to complete well-defined tasks or achieve specified goals productively and satisfactorily for the intended users in a given context (Dix et al., 1998). Last, but not the least, user modeling completes the set of key technological aspects for e-learners' satisfaction. User modeling is the knowledge a system has about the user's level of knowledge and intentions, processed as users interact with systems. Knowledge of both user and task domains should allow intelligent and adaptable systems to properly respond to the competence levels and needs of the tasks within contexts of use (Dix et al., 1998). This holistic understanding would help developers take into consideration users' expectations at the early stages of system design. In this way, expected learner performance would supplement the metrics suggested by the literature (Dix et al., 1998; Wentling et al., 2000) concerning the implementation of strategies and quality-of-service goals.

Regarding relationship issues, two factors are relevant for this overview: cognitive styles and motivation. Cognitive styles are individual characteristics that serve as stable indicators of how learners perceive, think of, remember, and solve problems (Kim, 2000; Liu & Dean, 1999). This characteristic is consistent with major dimensions of individual differences and explains stable individual performance across tasks over time. Thus, it is a key variable in designing effective systems for a particular user group, especially at early stages of the interaction. Indeed, results show that cognitive styles can help in explaining usability problems when browsing hypermedia documents (Chen, 2002; Kim, 2000; Liu & Dean, 1999). However, research results are numerous and mixed and give no consistent evidence about the relationship between cognitive styles and learners' performance in computer-based settings (Shih & Gamon, 2001; Wentling et al., 2000).

Motivation can be defined as the internal set of processes (both cognitive and behavioral) through which human energy becomes focused on the direction of effort (Welbourne et al., 2000). This definition is twofold. First, the internal set of processes by which human energy becomes focused describes the "agentic" nature of individuals in interaction with their environment (Bandura, 2000; Liu et al., 2002). A second relevant aspect of motivation is the direction of effort, which implies individual goal orientation assessed using defined internal success criteria. These two motivation-related aspects require that development teams actively involve different user roles at early stages of system design to get high usability, learnability, acceptance, and usage levels in order to likely match specific individual, task, and contextual characteristics. In this way, organizations could be more effective in creating the conditions to foster knowledge acquisition, transfer, and reutilization across different learner groups.

FUTURE TRENDS

Currently, people perceive learning as a product rather than a process. Adopting the latter view would require a holistic approach to analyze e-learner experience. This may lead to a paradigm shift in dealing with process, especially since learning, as an activity, is deeply enmeshed in interdependent and rapidly evolving organizational processes. Our vision is to pursue useful endeavors to effectively support learning by people within organizational contexts during their performance of work-related tasks. In this sense, e-learning systems would become tools for people to use, as often as needed, to acquire new knowledge related to current or future tasks. For this to happen, we need to develop a close people-process-system fit. To achieve that, cost-effective, integrated tools for courseware and competence-building evaluation within specific knowledge domains are in order. This requires coordination of efforts, information, and competencies among the key players, including universities, companies, government, and research communities. Such coordination is fundamental to achieve skill development at individual, organizational, and societal levels, articulated with defined strategies. Table 1 summarizes the topics identified in our analysis and describes specific research goals toward the fulfillment of this vision.

CONCLUSION

We have discussed a holistic approach covering business processes, e-learners, and information systems fit, while stressing the need for quantitative models, especially in evaluation. The interplay among these entities defines the

Table 1. Identified research topics

Issues	Proposed research topics
Learning task	How do e-learning initiatives relate to organizational knowledge management practices?
E-learner	In what conditions could e-learning be effective for everybody? Are there usability metrics universal across roles, levels of technology experience, cognitive styles, job levels, organizational contexts, and cultures? Does task complexity level mediate users' perception about usability? Would the initial role assigned by people to media hold steadily throughout the learning process, or does it change over time, motivated by learners' habits?
Interaction	To what extent is learners' trust and performance affected by perceived support for privacy, confidentiality, and security at the system level? Could software agents be the new "hidden persuaders" (Packard, 1957/1981) to get learners to go further into the skill-development cycle, overcoming obstacles along the way? Should such concerns be incorporated into a design discipline centered on ethical and deontological roles?

organizational space for the just-in-time learning concept. The expected benefits lie in aligning learning outcomes with business strategies.

REFERENCES

Astleitner, H. (2001). Web-based instruction and learning: What do we know from experimental research? Retrieved November 2001 from http://rilw.emp.paed.uni-muenchen.de/2001/papers/astleitner.html

Bandura, A. (2000). Cultivate self-efficacy for personal and organizational effectiveness. In E. A. Locke (Ed.), *Handbook of principles of organizational behavior* (pp. 20-136). Oxford, United Kingdom: Blackwell.

Bernardes, J., & O'Donoghue, J. (2003). Implementing online delivery and learning support systems: Issues, evaluation and lessons. In C. Ghaoui (Ed.), *Usability evaluation of online learning programs* (pp. 19-39). Hershey, PA: Idea Group Publishing.

Chen, S. (2002). A cognitive model for non-linear learning in hypermedia programmes. *British Journal of Educational Technology, 33*(4), 449-460.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998). *Human-computer interaction* (2nd ed.). Prentice Hall Europe.

Fuglseth, A. M. (2003). A tool kit for measurement of organizational learning: Methodological requirements and an illustrative example. *Journal of Universal Computer Science, 9*(12), 1487-1499.

Karoulis, A., Tarnanas, I., & Pombortsis, A. (2003). An expert-based evaluation concerning human factors in ODL programs: A preliminary investigation. In C. Ghaoui (Ed.), *Usability evaluation of online learning programs* (pp. 84-97). Hershey, PA: Idea Group Publishing.

Kim, K. (2000). *Individual differences and information retrieval: Implications on Web design.* Retrieved September 2001 from http://citeseer.nj.nec.com/update/409393

Liu, Y., & Dean, G. (1999). Cognitive styles and distance education. *Online Journal of Distance Learning Administration, II*(III). Retrieved September 2001 from http://www.westga.edu/~distance/and23.html

Liu, Y., Lavelle, E., & Andris, J. (2002). Experimental effects of online instruction on locus of control. *USDLA Journal, 16*(6). Retrieved April 2004 from http://www.usdla.org/html/journal/JUN02_issue/article02.html

Packard, V. (1981). *The hidden persuaders.* Penguin Books. (Original work published 1957)

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Shih, C., & Gamon, J. (2001). Web-based learning: Relationships among student motivation, attitude, learning styles, and achievement. *Journal of Agricultural Education, 42*(4). Retrieved April 2004 from http://pubs.aged.tamu.edu/jae/pdf/Vol42/42-04-12.pdf

Storey, M. A., Phillips, B., Maczewski, M., & Wang, M. (2002). Evaluating the usability of Web-based learning tools. *Educational Technology and Society, 5*(3). Retrieved April 2004 from http://ifets.ieee.org/periodical/Vol_3_2002/storey.html

Strazzo, D., & Wentling, T. (2001). *A study of e-learning*

practices in selected Fortune 100 companies. Retrieved September 2001 from http://learning. ncsa.uiuc.edu/papers/elearnprac.pdf

Vouk, M., Bilzer, D., & Klevans, R. (1999). *Workflow and end-user quality of service issues in Web-based education.* NC: North Carolina State University, Department of Computer Science. Retrieved September 2001 from http://www.computer.org/tkde/tk1999/k0673abs.htm

Welbourne, T., Andrews, S., & Andrews, A. (2000). *Back to basics: Learning about energy and motivation from running on my treadmill.* Retrieved September 2001 from http://www.eepulse.com/pdfs/treadmill%20adobe%203.1.01.pdf

Wentling, T., Waight, C., Gallager, J., La Fleur, J., Wang, C., & Kanfer, A. (2000). *E-learning: A review of literature.* Retrieved September 2001 from http://learning.ncsa.uiuc.edu/papers/elearnlit.pdf

KEY TERMS

Business Process: A set of organized work-related tasks and resources to pursue a specific organizational objective influencing learning experiences by defining two specific relationships: process-based roles (between business process and people) and learning tasks (between business process and information systems).

E-Learning Experience: A process by which people identify work-related learning needs, formulate related goals and the associated internal level-of-success criteria, search for feasible online options to achieve defined learning goals, select and acquire choices, and engage into and complete them successfully by achieving the related goals in a productive and satisfactory manner.

E-Learning Framework: A formal construct to diagnose and manage learning outcomes in terms of the operational dynamic of three basic entities: business process, information systems, and people.

Just-In-Time Learning: Strategic knowledge acquisition enmeshed in business activities to support employees in learning new skills when performing day-to-day tasks, while fostering the alignment between learning outcomes, technological and strategic business issues.

Learning Task: A set of steps with a defined learning goal addressing specific training needs identified within business processes driving the definition of proper instructional design and e-learning system requirements.

Motivation to E-Learn: An individual variable denoting an internal set of processes (both cognitive and behavioral) by which human energy becomes focused on learning particular work-related content (whether by actively interacting with courseware, participating in a virtual class, self-studying, doing e-homework alone or in group) to achieve specific learning goals.

Process-Based Role: The combination of a set of expected work-related behaviors, responsibilities, and the associated set of required competencies to perform within business settings in order to achieve organizational goals.

People-Process-System Fit: A degree of consistency among learner groups, business processes, and e-learning systems that (a) reflects the target constituency requirements, organizational context, and the stated educational or training goals, (b) applies the principles and approaches of constructivism, user-centeredness, participatory design, quality management, and organizational development to instructional design of courseware, and (c) translates into expected performance levels.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 1996-2000,

Modeling Information Systems in UML

Peter Rittgen
University College of Borås, Sweden

INTRODUCTION

The first approaches to object-oriented modeling appeared already in the second half of the 1970s, but not much happened for more than a decade so there were still barely more than a handful of modeling languages at the end of the 1980s. It was the early 1990s that witnessed an ever-growing market in competing object-oriented methods so that potential users found it increasingly difficult to identify any single method that suited their needs. This phenomenon came to be known as the "method wars." Toward the end of 1994, two of the "big" players, Grady Booch and Jim Rumbaugh, decided to join forces by integrating their respective approaches, the Booch method and OMT (object modeling technique). In late 1995, Ivar Jacobson became a member of this team merging in his OOSE method (object-oriented software engineering). The efforts of the "three amigos" aimed at overcoming unnecessary differences between the individual approaches and also improving each of them by creating a common, standardized modeling language that could serve as an industry standard. The result was the release of the Unified Modeling Language (UML), version 0.9, in June 1996. The UML partners, an industry consortium, performed further work on UML. This led to the versions 1.0 and 1.1 being introduced in 1997. The latter was adopted by the OMG (Object Management Group) in the same year. The current version is 2.0 (OMG, 2005, 2006).

BACKGROUND

UML is a language to support the development of software systems. It features a common framework that provides a set of diagram types covering different aspects of an information system. Here we will focus on the ones we deem to be most important: class diagram, use case diagram, and activity diagram. The first is the principal diagram for the static view on a system. The second is often put forward as the central diagram in communication with the user (see, e.g., Dobing & Parsons, 2000). The latter plays a central role in the information-system (or business-oriented) perspective (see following sections). Elementary concepts of the UML are actor, activity (and state), object, and class.

An actor is a human being or a (computer) system who/which is able to perform activities on his/her/its own. The actor typically represents a group of similar human beings/systems and corresponds to the role the actor performs in the given context. *Teacher* is an example for such an actor. In UML, actors are shown as stick figures.

An activity refers to a logically connected set of actions that are carried out as a unit in some order. These actions might be executed sequentially, alternatively, concurrently, or in any combination thereof. *Grade exam* is an example for an activity.

"An object is an abstraction of a set of real-world things such that:

- all of the real-world things in the set—the instances—have the same characteristics,
- all instances are subject to and conform to the same rules" (Shlaer & Mellor, 1988, p. 14).

The structure (attributes) and behavior (operations) of similar objects are gathered in their common class. The values of the attributes at a certain point in time represent the state of the object. Classes are drawn as rectangles with the object identifier printed in bold face. Additional horizontal compartments can be introduced for the attributes and operations of the class. The object is depicted in the same way with the object identifier underlined (optionally followed by a colon and the respective class identifier). *Grade* is an example of a class, *F: Grade* that of an object.

Figure 1 shows how these concepts are related to each other: actors perform activities that involve objects. Each object is an instance of precisely one class. The class diagram shows classes and their relations (called associations). An

Figure 1. Language overview

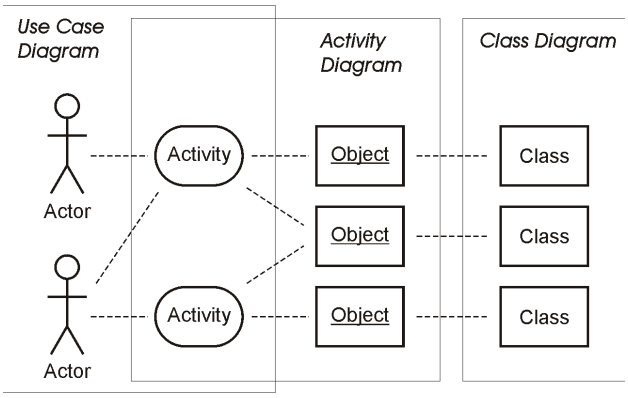
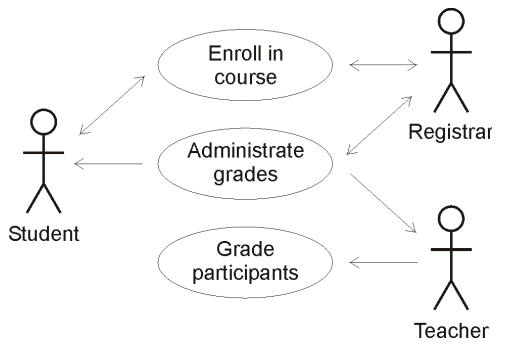


Figure 2. A use case diagram



activity diagram gives a detailed account of the order in which activities are executed. It can also show the relevant states of the involved objects. The use case diagram visualizes use cases (complex activities) and their relation to actors.

Use Case Diagram

Use cases have been introduced in 1992 by Jacobson and his colleagues. Their book is now available in the second edition (Jacobson, Christerson, Jonsson, & Övergaard, 2004). A use case describes a way in which an actor can interact with your organization. It forms part of a use case scenario, a business situation that is supposed to be supported by the information system in question. Figure 2 gives an example of a use case diagram involving three use cases and three actors.

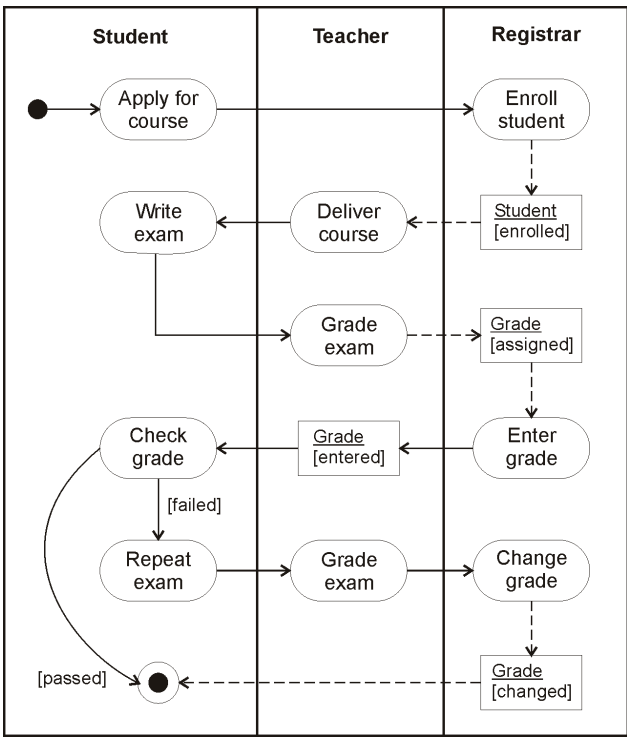
The arrows indicate the direction of the information flow. So the registrar can both read and enter/change grades, whereas teachers and students can only read them. Often the arrowheads are omitted to simplify modeling. Each use case is detailed by a description in structured English of the activities that make up this use case. The use case diagram and the use cases form the basis for the development of class diagrams and activity diagrams. The former specify the static structure of the information that is handled by the use cases, and the latter give a precise, formalized account of the process logic.

Activity Diagram

An activity diagram consists of the detailed activities that make up the use cases. Figure 3 gives an example of such a diagram in relation to the use cases of Figure 2.

The use case *Enroll in course* comprises the activities *Apply for course* (performed by the student) and *Enroll student* (performed by the registrar). *Grade participants* maps to *Grade exam* assuming that the grade for the course is determined by a final exam only. The use case *Administrate grades* involves the activities *Enter grade*, *Check grade*, and *Change grade*. Note that the activity diagram also contains

Figure 3. An activity diagram



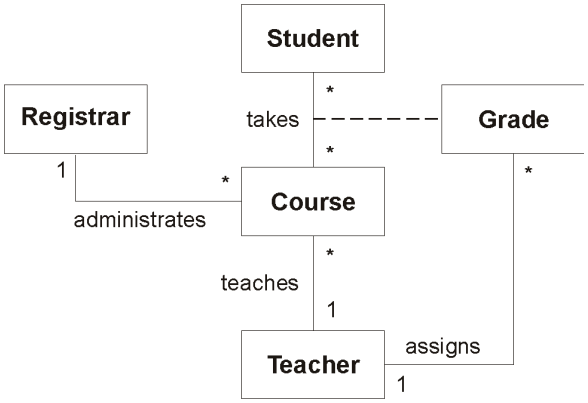
activities that are not present in the use case diagram, such as *Deliver course* and *Write exam*. That is because, in our example, they are not supposed to be supported by the information system and hence do not constitute use cases of such a system. But they are still an important part of the overall business process.

The business process in Figure 3 starts with the initial state (the black dot). The activities are in the boxes with round sides. The rectangles contain objects-in-state where the object identifier is underlined and the object's state is enclosed in square brackets. So after *Enter grade*, for example, the object *Grade* is in state *entered*. Alternative paths leaving an activity must be labeled by so-called guards that determine under which condition each path is taken. These guards are also enclosed in square brackets and should be mutually exclusive. The process terminates when the final state is reached (i.e., the circle containing the black dot). Actors are included in the form of so-called swim lanes. Concurrent execution is achieved by introducing synchronization bars. For more detailed information on activity diagrams we refer the reader to OMG (2005, p. 402 ff).

Class Diagram

A typical approach to identifying classes is that of looking at the nouns contained in a textual specification of the system.

Figure 4. A class diagram



Jacobson et al. (2004) suggest that we should limit our attention to the nouns contained in the use case diagram instead to avoid having to eliminate inappropriate class candidates. If we apply this to the use case diagram of Figure 2, we arrive at the class diagram of Figure 4 where all the nouns of the former map to a class of the latter. Observe that the noun *participants* maps to the class *Student*.

The class diagram also shows associations between the classes that indicate the way in which they are related to each other. A single line is drawn connecting the respective classes. It can be labeled with a verb that determines the nature of the association. In addition, we can also specify the cardinality with the asterisk referring to an unlimited number. In the example a *Teacher teaches* a number of *Course*s, but each *Course* is only taught by one *Teacher*. *Grade* is a so-called association class. It qualifies the relation between *Student* and *Course* and is therefore attached to the association path by a dashed line. For further details on class diagrams refer to OMG (2005, p. 134 ff).

MODELING INFORMATION SYSTEMS IN UML

UML has been designed primarily to support the development of software systems. In such systems the software artifact plays the central role, and the role of human beings is restricted to the specification of requirements and later on the use of the system. An information system, on the other hand, is a human activity system (sometimes also called a socio-technical system) where some part of the system *might* be *supported* by software (Buckingham, Hirschheim, Land, & Tully, 1987).

The focus being on software systems, it is natural that UML does not emphasize the aspects of an information system that aim at the value and support it can provide to the business, such as strategy (e.g., value chains and strategic goals) and organization (e.g., organizational charts and business processes). These issues are dealt with in "business modeling" (also called enterprise modeling), but UML is beginning to advance into that field, too. The OMG, for instance, claims that business processes can be represented in UML (OMG, 2005, p. 306). This claim can, however, be challenged in a number of ways. First of all, there is very little direct support for business processes in the basic UML diagrams (Bruno, Torchiano, & Agarwal, 2002). Instead many of the required concepts are "hidden" in extensions to UML, such as enterprise collaboration architecture (ECA) (OMG, 2004a). Another issue is whether UML models are indeed "understandable" enough to form a solid basis for the communication between "developers" and "users." An empirical study of UML's usability (Agarwal & Sinha, 2003) showed that UML diagrams score between 4.3 and 5.3 on a 7-point Likert scale, that is, the ease-of-use of UML is closer to indifference (4) than to the optimum (7). Understandability of UML is also weakened by construct redundancies, inconsistencies, and ambiguities (Shen & Siau, 2003).

Candidate languages for business processes (in native UML) are use cases and activity diagrams. The former have been criticized for their conceptual flaws (Dobing & Parsons, 2000) such as fragmentation of objects and for construct overload and ambiguity (Shen & Siau, 2003). In addition, they restrict the specification of process logic to a textual form, which is neither rigorous nor expressive enough for typical flows of control. Activity diagrams come closer to fulfilling the requirements of a business-process language and are therefore often used in that way (see the next section).

FUTURE TRENDS

So far, there is still an abundance of languages used for business modeling, for example, architecture of integrated information systems (ARIS) (Scheer, 1999), open system architecture for computer-integrated manufacturing (CIMOSA) (AMICE, 1993), integrated definition (IDEF*) (Ang, Pheng, & Leng, 1999), integrated enterprise modeling (IEM) (Spur, Mertins, & Jochem, 1995), dynamic essential modeling of organization (DEMO) (Reijswoud, Mulder, & Dietz, 1999), and GRAI integrated methodology (GIM) (Doumeingts, 1998). Often modelers also pick out particular model types from integrated methods or use "stand-alone" languages that were developed for specific purposes. In the area of business process modeling these include event-driven process chains (EPC) (from ARIS), Business Process Modeling Language (Arkin, 2002), IDEF3 (Mayer et al., 1997), Tropos (Castro, Kolp, & Mylopoulos, 2002), and petri nets (Aalst, Desel, & Oberweis, 2000).

But how can we bridge the gap between these enterprise-modeling languages and UML? One approach consists of devising mechanisms that "translate" domain models into UML models. Such a translation is by no means straightforward but typically involves the reconstruction of the domain model as a UML diagram by a modeler who is experienced in both the domain language and UML. It usually involves the loss of information that was present in the domain model and the necessity to add information not yet represented. This implies also the risk of introducing errors, or more precisely inconsistencies between the original and the translated model. Examples of such an approach can be found in many of the languages mentioned due to the popularity of UML, for example, EPC to UML (Nüttgens, Feld, & Zimmermann, 1998), IDEF and UML (Noran, 2000), Tropos and UML (Mylopoulos, Kolp, & Castro, 2001), DEMO and UML (Mallens, Dietz, & Hommes, 2001), or Petri Nets and UML (Gou, Huang, & Ren, 2000).

Another—and perhaps even more promising—approach is to equip UML itself with domain support. This can be done in at least two ways: extending existing concepts or introducing new ones (typically in the form of profiles). A major effort toward the latter is being made under the heading "Enterprise Distributed Object Computing (EDOC)" and was adopted by the OMG in February 2004 (OMG, 2004a-g), but work is still ongoing. Its core is the enterprise collaboration architecture ECA (OMG, 2004a). The former approach is followed by several researchers independently. In the area of business processes and workflows, they primarily investigate extensions of use cases and activity diagrams, for example, Rittgen (2003) or Dumas and Hofstede (2001).

CONCLUSION

The UML is a widely used modeling language in the area of software systems modeling, but it still has to gain ground in domain modeling, especially in business (or enterprise) modeling. Promising steps have already been taken that have so far led to results such as EDOC. But there is still a need for research that provides the core concepts and diagrams of the UML with domain-specific semantics to make them more useful for business modeling. On the other hand, add-on concepts such as EDOC processes require a tighter semantical integration with the existing models. Future research and development will show which of the paths toward integration is more promising and will hence be taken. But all approaches mentioned share a common goal: to improve the communication and understanding between the different people involved in shaping an information system and thereby also improving the system itself.

REFERENCES

Aalst, W. van der, Desel, J., & Oberweis, A. (2000). *Business process management—Models, techniques and empirical studies. Lecture notes in computer science* (vol. 1806). Berlin: Springer.

Agarwal, R., & Sinha, A. P. (2003). Object-oriented modeling with UML: A study of developers' perceptions. *Communications of the ACM, 46*(9), 248-256.

AMICE. (1993). CIMOSA: *Open system architecture for CIM* (2nd revised and extended ed.). Berlin: Springer.

Ang, C-L., Pheng, K. L., & Leng, G. R .K. (1999). IDEF*: A comprehensive modelling methodology for the development of manufacturing enterprise systems. *International Journal of Production Research, 37*(17), 3839-3858.

Arkin, A. (2002). *Business process modeling language*. Aurora, CO: BPMI.org.

Bruno, G., Torchiano, M., & Agarwal, R. (2002). UML enterprise instance models. In S. Iyer & S. Naik (Eds.), *CIT 2002, Proceedings of the Fifth International Conference on Information Technology* (pp. 135-138). New Delhi, India: Tata McGraw-Hill.

Buckingham, R. A., Hirschheim, R. A., Land, F. F., & Tully, C. J. (1987). Information systems curriculum: A basis for course design. In R. A. Buckingham, R. A. Hirschheim, F. F. Land, & C. J. Tully (Eds.), *Information systems education. Recommendations and implementation* (pp. 14-133). Cambridge, UK: Cambridge University Press.

Castro, J., Kolp, M., & Mylopoulos, J. (2002). Towards requirements-driven information systems engineering: The Tropos Project. *Information Systems, 27*(6), 365-389.

Dobing, B., & Parsons, J. (2000). Understanding the role of use cases in UML: A review and research agenda. *Journal of Database Management, 11*(4), 28-36.

Doumeingts, G. (1998). GIM—GRAI Integrated Methodology. In A. Molina, A. Kusiaka, & J. Sanchez (Eds.), *Handbook of life cycle engineering—Concepts, models and methodologies* (pp. 227-288). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Dumas, M., & Hofstede, A. H. M. ter (2001). UML activity diagrams as a workflow specification language. In M. Gogolla & C. Kobryn (Eds.), *<<UML 2001>>—The Unified Modeling Language. Fourth International Conference, Lecture Notes in Computer Science* (vol. 2185) (pp. 76-90). Berlin: Springer.

Gou, H., Huang, B., & Ren, S. (2000). A UML and Petri nets integrated modeling method for business processes in

virtual enterprises. In S. Staab & D. O'Leary (Eds.), *Bringing knowledge to business processes, papers from 2000 AAAI Spring Symposium* (pp. 142-144). Menlo Park, CA: AAAI Press.

Jacobson, I., Christerson, M., Jonsson, P., & Övergaard, G. (2004). *Object-oriented software engineering: A use case driven approach* (2nd ed.). Wokingham, UK: Addison-Wesley.

Mallens, P., Dietz, J., & Hommes, B-J. (2001). The value of business process modeling with DEMO prior to information systems modeling with UML. In J. Krogstie, K. Siau, & T. Halpin (Eds.), *EMMSAD'01—Sixth IFIP 8.1 Workshop on Evaluation of Modeling Methods in Systems Analysis and Design* (pp. 173-184). Oslo, Norway: SINTEF.

Mayer, R. J., Menzel, C. P., Painter, M. K., deWitte, P. S., Blinn, T., Perakath, B., et al. (1997). *Information integration for concurrent engineering (IICE)—IDEF3 Process Description Capture Method Report*. College Station, TX: Knowledge Based Systems Inc. & Wright-Patterson AFB, OH: Armstrong Laboratory.

McLeod, G. (2000). Beyond use cases. In K. Siau, Y. Wand, & A. Gemino (Eds.), *EMMSAD'00—Fifth IFIP 8.1 Workshop on Evaluation of Modeling Methods in Systems Analysis and Design*. Stockholm, Sweden.

Mylopoulos, J., Kolp, M., & Castro, J. (2001). UML for agent-oriented software development: The Tropos Proposal. In M. Gogolla & C. Kobryn (Eds.), *<<UML 2001>>—The Unified Modeling Language, Fourth Int. Conference. Lecture Notes in Computer Science* (vol. 2185) (pp. 422-441). Berlin: Springer.

Noran, O. (2000). *Business modelling: UML vs. IDEF*. Brisbane, Australia: Griffith University. Electronic version available at www.cit.gu.edu.au/~noran

Nüttgens, M., Feld, T., & Zimmermann, V. (1998). Business process modeling with EPC and UML: Transformation or integration? In M. Schader & A. Korthaus (Eds.), *The Unified Modeling Language—Technical aspects and applications* (pp. 250-261). Berlin: Springer.

OMG. (2004a). *Enterprise collaboration architecture (ECA) specification*. Version 1.0, February 2004. Needham, MA: OMG. Retrieved June 17, 2006, from http://www.uml.org

OMG. (2004b). *Metamodel and UML profile for Java and EJB specification*. Version 1.0, February 2004. Needham, MA: OMG. Retrieved June 17, 2006, from http://www.uml.org

OMG. (2004c). *Flow composition model (FCM) specification*. Version 1.0, February 2004. Needham, MA: OMG. Retrieved June 17, 2006, from http://www.uml.org

OMG. (2004d). *UML profile for patterns specification*. Version 1.0, February 2004. Needham, MA: OMG. Retrieved June 17, 2006, from http://www.uml.org

OMG. (2004e). *UML profile for enterprise collaboration architecture specification*. Version 1.0, February 2004. Needham, MA: OMG. Retrieved June 17, 2006, from http://www.uml.org

OMG. (2004f). *UML profile for metaobject facility (MOF) specification*. Version 1.0, February 2004. Needham, MA: OMG. Retrieved June 17, 2006, from http://www.uml.org

OMG. (2004g). *UML profile for relationships specification*. Version 1.0, February 2004. Needham, MA: OMG. Retrieved June 17, 2006, from http://www.uml.org

OMG. (2005). *Unified Modeling Language: Superstructure*. Version 2.0, August 2005. Needham, MA: OMG. Retrieved June 20, 2006, from http://www.uml.org/

OMG. (2006). *Unified Modeling Language: Infrastructure*. Version 2.0, March 2006. Needham, MA: OMG. Retrieved June 20, 2006, from http://www.uml.org/

Palanque, P., & Bastide, R. (2003). UML for interactive systems: What is missing. In G. W. M. Rauterberg, M. Menozzi, & J. Wesson (Eds.), *Human-computer interaction INTERACT '03. Ninth IFIP TC13 Int. Conference on Human-Computer Interaction* (pp. 96-99). Amsterdam, The Netherlands: IOS Press.

Reijswoud, V. E. van, Mulder, J. B. F., & Dietz, J. L. G. (1999). Speech act based business process and information modelling with DEMO. *Information Systems Journal, 9*(2), 117-138.

Rittgen, P. (2003). Business processes in UML. In L. Favre (Eds.), *UML and the unified process* (pp. 315-331). Hershey, PA: IRM Press.

Scheer, A-W. (1999). *ARIS—Business process modeling*. Berlin: Springer.

Shen, Z., & Siau, K. (2003). An empirical evaluation of UML notational elements using a concept mapping approach. In S. T. March, A. Massey, & J. I. DeGross (Eds.), *Proceedings of the 24th Int. Conference on Information Systems* (pp. 194-206). Atlanta, GA: AIS.

Shlaer, S., & Mellor, S. J. (1988). *Object-oriented systems analysis: Modeling the world in data*. Englewood Cliffs, NJ: Prentice-Hall.

Spur, G., Mertins, K., & Jochem, R. (1995). *Integrated enterprise modeling*. Berlin: Beuth.

KEY TERMS

Activity: A logically connected set of actions that are carried out as a unit in some order. It is associated with a state (called action state) in which the system remains while the activity is performed.

Actor: A person or (computer) system that can perform an activity. The actor does not refer to a particular individual but rather to a role (e.g., *Teacher*).

Attribute: A property of an object/class. The class *Car*, for example, can have an attribute *Color*, its value for the object *MyCar: Car* might be *blue*.

Class: A template for similar objects defining attributes and operations.

Object: An abstraction of a set of real-world things that has state, behavior, and identity. An instance of its class where the values of the attributes determine the state and the operations the behavior.

Operation: A function or transformation that may be applied to or by objects in a class. The class *Account*, for example, might have the operations *Open*, *Close*, *PayIn* (*Amount*), and *Withdraw* (*Amount*).

State: A certain combination of attribute values of an object.

Modeling Security Requirements for Trustworthy Systems

Kassem Saleh
American University of Sharjah, UAE

Ghanem Elshahry
American University of Sharjah, UAE

BACKGROUND

To increase users' trust in the systems they use, there is a need to develop trustworthy systems. These systems must meet the needs of the system's stakeholders with respect to security, privacy, reliability, and business integrity (Mundy, deVries, Haynes, & Corwine, 2002). The first major step in achieving trustworthiness is to properly and faithfully capture the stakeholders requirements. A requirement is something that the system must satisfy or a quality that the system must possess. A requirement is normally elicited from the system stakeholders, including its users, developers, and owners. Requirements should be specified before attempting to construct the system. If the correct requirements are not captured properly and faithfully, the correct system cannot be built. Consequently, the system will not be usable by its intended users. The success of any system depends on meeting requirements classified under two complementary types. First, the functional requirements are the system's operations from the user's perspective describing the visible and external interactions with the system under consideration. Second, the non-functional requirements (NFRs) are mainly the system's constraints imposing special conditions and qualities on the system to construct. Consequently, system acceptance testing must be based on both functional and non-functional system's requirements. Unfortunately, it is reported that about 60% of errors originate from the requirements and analysis activities (Weinberg, 1997).

Surveys have shown that large numbers of IT-based systems were implemented starting from their elicited functional requirements without a clear and formal consideration of their non-functional counterparts such as security requirements. Furthermore, system requirements engineers and analysts are not well-trained in capturing security requirements early in the system development process. Security assurances are often based on the traditional and ad hoc approach of conducting penetration tests followed by a patching process. This approach is very costly and endangers the fulfillment of the basic goals of system security, namely confidentiality, integrity, availability, and accountability. Recently, many researchers addressed security requirements engineering as an integral and essential element of systems engineering. Devanbu and Stubblebine (2000) propose a roadmap for software engineering for security, and Henning and Garner (1999) consider life cycle models for survivable and secure systems.

Non-functional requirements can be classified under three broad categories (Robertson & Robertson, 1999): system-related, process and project-related and human-related requirements.

The rest of this article is organized as follows. The next section overviews the security goals and requirements. The third section introduces security requirements modeling using the Goal-Oriented Requirements Language (GRL) (ITU, 2002) and UMLsec, a security extension to the Unified Modeling Language (Jurjens, 2005; Elshahry, 2005), and its modifications. The fourth section provides some examples of using GRL and UMLsec models for requirements specifications. We conclude in the final section and provide items for further investigation.

SECURITY GOALS AND REQUIREMENTS

The main system security goals to achieve are confidentiality, integrity, availability, and accountability. Confidentiality ensures that only authorized users or applications are allowed to interact with the system. Integrity ensures that critical data has not been changed in an improper way in the system. Availability ensures that the information and/or services are readily available to an authorized user on demand. Accountability ensures that once authorized users access the system, they are accountable for all of their actions (Whitman & Mattord, 2005). Normally, security requirements should not be specified in terms of the types of security mechanisms or controls that are currently used for implementation. To achieve the security goals, security requirements should be identified. These requirements can be structured around the 12 security requirement types identified in Firesmith (2003).

Table 1 maps and shows the contributions of the security requirements to the security goals. For example, survivability,

Table 1. Mapping security requirements to security goals

Security Requirement	Confidentiality	Integrity	Availability	Accountability
Identification Requirements	•	O		O
Authentication Requirements	•	O		O
Authorization Requirements	•	O		O
Immunity Requirements	O	•	O	
Integrity Requirements		•		
Intrusion Detection Requirements	•	O	O	
Intrusion Prevention Requirements	•	O	O	
Non-repudiation Requirements		O		•
Privacy/Secrecy Requirements	•			
Security Auditing Requirements		O		•
Survivability Requirements			•	
Physical Protection Requirements	•	•	•	O
System Maintenance Requirements	O	O	•	
Conformance Requirements	O	O	O	•
• main contribution	O partial contribution			

physical protection, and system maintenance requirements contribute to the system availability security goal.

MODELING SECURITY REQUIREMENTS

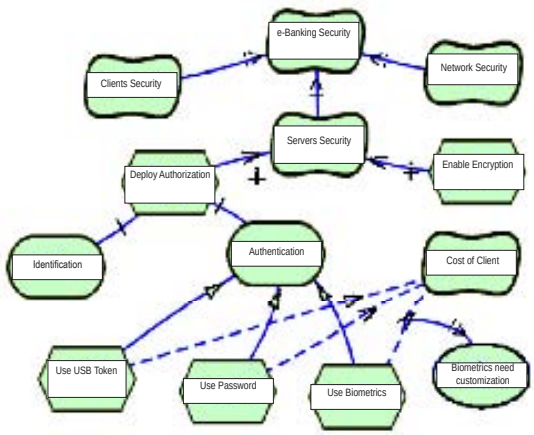
Modeling is an important process that can be used for specifying and analyzing requirements. There are many advantages of developing a model before starting the design process. Desirable modeling formalisms are executable and therefore allows the modeler to verify the model and its dynamic behavior before accepting it. An acceptable model is the basis upon which the design process can build. Moreover, correctness of the implemented system can be checked by verifying the conformity of the system to its specified requirements model. There are many existing modeling languages and formalisms. The most famous is the Unified Modeling Language (UML) (OMG, 2003). UML is a de facto standard in the software industry, and it is being generalized to model systems in general (OMG, 2005). However, we are aware of two modeling formalisms that are capable of expressing security requirements in the model. First, the Goal-Oriented Requirement Language (GRL) (ITU, 2002) is a language for supporting goal-oriented modeling and reasoning of requirements, especially for dealing with non-functional requirements such as security and performance. GRL provides the elements to express several concepts appearing during the requirement elicitation phase. Second, UMLsec, an extension to the UML, has been developed by Jurjens (2005) to express security requirements. It is worth mentioning here that another extension to UML, SecureUML, was introduced in Lodderstedt, Basin, and Doser (2002). However, this extension is only to specify role-based access control to support authorization requirements.

Goal-Oriented Requirement Language (GRL)

GRL is a graphical modeling language for capturing NFRs in general. There are three main categories of elements in GRL: intentional elements, links, and actors. The intentional elements in GRL are goals, tasks, softgoals, resources, and beliefs. These elements are used for models that allow answering questions such as why particular behaviors, informational and structural aspects were chosen to be included in the system requirement; what alternatives were considered; what criteria were used to deliberate among alternative options; and what the reasons were for choosing one alternative over the other. GRL supports the reasoning about scenarios by establishing mappings between intentional GRL elements and non-intentional elements in scenario models of the User Requirements Notation—Functional Requirements (URN-FR). Modeling goals and scenarios are complementary and may help identifying further goals, scenarios, and scenario steps important to stakeholders, thus contributing to the completeness and accuracy of the elicited requirements.

A GRL model consists of several goal model structures. Each structure represents a requirement category. For example, Figure 1 shows an e-banking security system as the root of the security requirements. A requirement can

Figure 1. E-banking requirements in GRL



be represented as a goal, softgoal, or task. A softgoal is a condition that the actor would like to achieve. However, unlike in the concept of goal, there are no clear-cut and subjective criteria for checking whether the condition has been achieved. The figure shows that Use USB Token, Use Password, and Use Biometrics are means for achieving the authentication security goal. Also, Deploy Authorization and Enable Encryption contributes positively to the Servers Security softgoal.

For a complete tutorial on GRL, the reader can refer to Amyot (2003).

Unified Modeling Language for Security (UMLsec)

The UMLsec extension has been developed by Jurjens (2005) based on his security-critical system development experience in industrial projects involving government agencies, banks, insurance companies, and smart cards. Security goals such as confidentiality, integrity, availability, and accountability are offered as specification elements and stereotypes in the UMLsec extension. UMLsec properties are used to evaluate different kinds of security-related diagrams. The UMLsec encapsulates knowledge on security engineering and makes it available to system engineers who may not be experts in security. UMLsec took advantage of three metalanguage mechanisms for extending the UML: stereotypes, constraints, and tagged values. Stereotypes are used with tags for security requirements formulation. The constraints criteria determine whether the requirements are met or not. The extension provides core profile that includes the main security requirements.

Stereotypes, included between <<>>, define new types of modeling elements extending the semantics of existing types in the UML metamodel. The new element can capture some aspect of a new domain, like security, not supported by the standard UML elements and stereotypes. Constraints supply conditions and restrictions for UML model elements. A constraint can be specified in any format as long as it is written inside braces. If, for example, a class has *password* as one of its attributes, constraint can be applied as {password can't be blank}. Tagged values are name-value pairs in curly brackets {} associating data with model elements. A tagged value is designed to explicitly define a property. For example, If nodeName represents the node where the component resides, {location = nodeName} may be attached to a component.

UMLsec Stereotypes

UMLsec stereotypes are used to describe security requirements for access control, confidentiality, integrity, and so forth. In the following, we describe the main security-related stereotypes introduced by Jurjens (Jurjens, 2005), in addition to our own additional stereotypes including the <<audit log>>, <<multiplicity>>, <<no misuse>>, and <<standard>> .

<<audit log>> stereotype can be used by the node elements to completely record any transaction information or activity. The logging fields include: event ID, event date, event time, event user, and event action (allowed or denied). Audit logs records are tamperproof since they include the integrity tag. The authors introduced <<audit log>> stereotype to fulfill several security requirements such as audit log, non-repudiation, and the intrusion detection system security requirements (Elshahry, 2005).

<<encrypted>>, <<internet>>, <<LAN>>, <<multiplicity>>, and <<wire>> link stereotypes are used in deployment diagrams to denote the corresponding types of communication links. It is required that each link carries at most one of these stereotypes except for <<multiplicity>>. The <<multiplicity>> link stereotype is introduced to fulfill the survivability security requirements. <<multiplicity>> stipulates multiple network links per device (Elshahry, 2005). <<Issuer Node>>, <<multiplicity>>, <<LAN>>, <<POS Device>>, and <<Smart Card>> node stereotypes are used in deployment diagrams to denote the respective kinds of system nodes. It is required that each node carries at most one of these stereotypes except for <<multiplicity>>. The <<*multiplicity*>> node stereotype is introduced to fulfill the survivability security requirements. <<*multiplicity*>> stipulates multiple active devices such as servers, routers, switches, or storage.

<<fair exchange>> stereotype stipulates that the interaction is performed in a way that prevents parties from cheating. For example, in the e-commerce context, it is the require-

ment that, after a prepayment, the buyer either receives the purchased good or the money can be reclaimed.

<<guarded access>> stereotype means that each <<guarded>> object in the subsystem can only be accessed through the object specified by the tag {guard} attached to the <<guarded>> object. However, the <<guarded>> stereotype labels objects that are supposed to be guarded in the scope of <<guarded access>> stereotype. It has a tagged value {guard} that defines the name of the corresponding guard object.

<<high>>, <<integrity>>, and <<secrecy>> stereotypes are used as dependency constraint for <<secure links>> stereotype in static structure or component diagrams. Dependencies, stereotypes, denote dependencies that are supposed to provide the respective security requirement for the data that is sent along them as arguments or returned values of operations or signals.

<<no misuse>> stereotype can be used as link or node stereotype to detect, alert, prevent any misuse of the system, such as a repeated request for system component in an irregular manner. The <<*no misuse*>> stereotype was introduced to fulfill the immunity security requirements.

<<standard>> stereotype is a node (system or subsystem) stereotype indicating the security standard the system or subsystem must conform to. <<*iso17799*>> is one instance of <<*standard*>> stereotype. ISO 17799 is an international standard for the code of practice for information security management (ISO/IEC 17799, 2000).

Table 2 lists the four stereotypes that were added to the existing UMLsec stereotypes. A table showing all UMLsec stereotypes and detailed description can be found in Jurjens (2005).

Table 3 shows the mapping of the security requirement types to the modified UMLsec stereotypes. For example, authorization requirements can be role-based (<<rbac>>) or subject-based (<<guarded access>>). Intrusion detection and prevention, non-repudiation, and auditing requirements are related to the <<audit log>> stereotype since audit logs are instrumental in meeting these requirements.

EXAMPLE

A local financial institution, consisting of head office and 20 branches, offers several banking services through branches, call centers, and automatic teller machines (ATM). The bank business unit, sales department, would like to deploy e-banking services to service their customers'; on the other hand they are faced with the risk element in terms of the e-banking transactions security. Along with a wide variety of features in e-banking services, today's customers also want assurance that their e-banking services are completely secure. The aim of this example is not to provide complete requirement specifications, but to give a flavor of using formal modeling languages. The following fitures were generated using the available tools for GRL (GRL, 2005) and UMLsec (Jurjens, 2004).

GRL Representation

The e-banking system security requirements are shown in Figure 2 from the system stakeholders perspective using the GRL. E-banking security softgoal is the root of the required security components. There are several ways for representing each component depending upon the scope of the design. Some requirements can be grouped into one softgoal; for example, identification, authentication and authorization are combined into the access control softgoal. Defense and hardening softgoal combines the intrusion detection system, physical security, intrusion prevention system, immunity (anti malicious software system), and system maintenance security requirements. The other security requirements such as confidentiality (for secrecy and privacy requirements), integrity, availability, and survivability, are represented individually. Non-repudiation and security auditing requirements are deployed as tasks contributing to the execution of the Tracking softgoal.

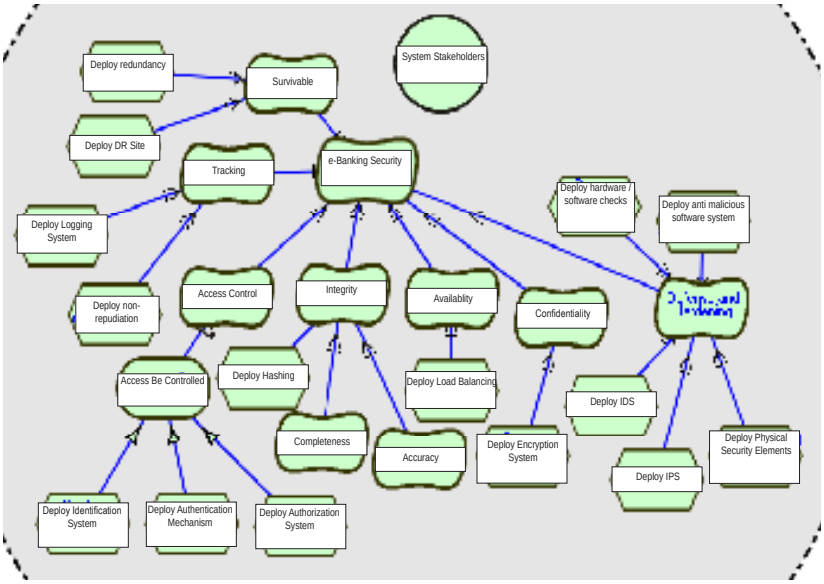
Table 2. Four stereotypes added to UMLsec

Stereotype	Base Class	Tags	Constraints	Description
<<*audit log*>>	Subsystems, node	integrity		Log event activity (audit security requirements)
<<*multiplicity*>>	Link, node			Load balancing and redundancy (survivability requirements)
<<*no misuse*>>	Subsystem		Prevent misuse	Antivirus and network traffic (immunity requirement)
<<*standard*>> *like* <<*iso17799*>>	System, subsystem			Standard to adhere to (standard conformance requirement)

Table 3. Mapping security requirements to UMLsec stereotypes

Security Requirements	Security Stereotypes
Identification Requirements	<<guarded access>>
Authentication Requirements	<<guarded access>>
Authorization Requirements	<<guarded access>>, <<rbac>>
Immunity Requirements	<<no misuse>>
Integrity Requirements	<<integrity>>, <<critical>>
Intrusion Detection Requirements	<<no misuse>>, <<audit log>>
Intrusion Prevention Requirements	<<no misuse>>, <<no down-flow>>, <<no up-flow>>, <<audit log>>
Non-repudiation Requirements	<<provable>>, <<fair-exchange>>, <<audit log>>
Privacy/Secrecy Requirements	<<encrypted>>, <<secure links>>, <<data security>>
Security Auditing Requirements	<<provable>>, <<audit log>>
Survivability Requirements	<<multiplicity>>
Physical Protection Requirements	<<secure links>>, <<LAN>>, <<wire>>, <<multiplicity>>
System Maintenance Requirements	<<no misuse>>, <<multiplicity>>
Security Conformance Requirements	Standard like <<iso17799>>

Figure 2. GRL diagram for the e-banking security requirements

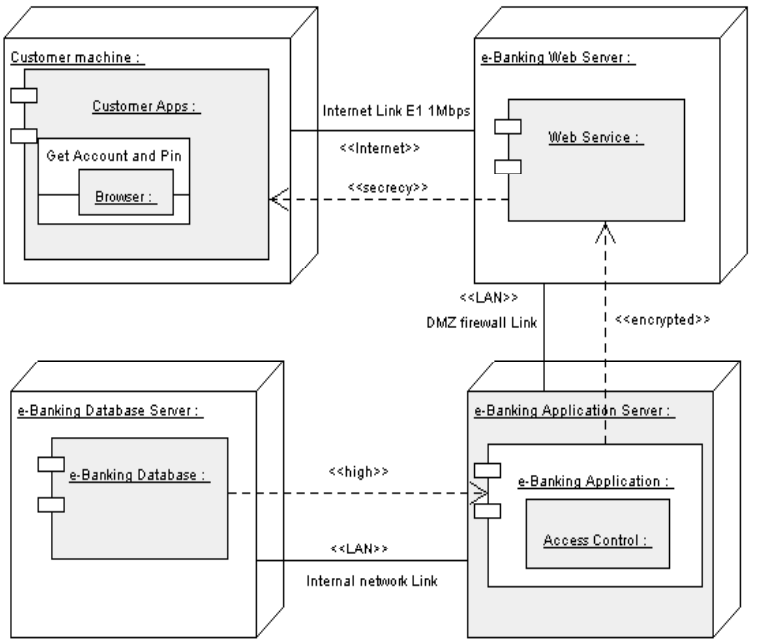


UMLsec Representation

Unlike a GRL diagram that can describe all elicited security requirements, a single UMLsec diagram can only represent related security requirements or a partial view of the modeled system security. Normally, many UMLsec diagrams would be needed to capture all system security requirements. Although GRL model can describe the security requirements and mechanisms in one diagram, it does not enforce or validate the requirements. On the other hand, UMLsec diagrams permit the expression of security-relevant characteristics within the model in the system specification and validate them using the available tools (Jurjens, 2004).

The e-banking Web service connectivity example in Figure 3 shows a deployment diagram with four node instances: customer machine, e-banking Web server, e-banking Application aerver, and e-banking database server.

Figure 3. E-banking Web connectivity deployment diagram



Customer machine and e-banking Web server nodes are connected via a link stereotyped <<Internet>>. There is also a <<secrecy>> stereotype dependency from the e-banking Web server component to "Get Account and Pin", thus the e-banking Web server is able to communicate with the Browser for getting the account and pin code. <<secrecy>> and similar dependency stereotypes, like <<integrity>> and <<high>>, denote dependencies that are supposed to provide the respective security requirement for the data that is sent along them as arguments or returned values of operations or signals. These stereotypes are used in the constraint for <<secure links>> stereotype. <<audit log>> stereotypes specifies the requirement that the application server must keep audit logs of all transactions it processes. Also, the <<multiplicity>> stereotype specifies the high availability and survivability requirement needed for the Web servers. Figure 4 shows an e-banking sequence diagram. In this diagram, a role-based access control (rbac) security mechanism is required, in which a legitimate client requests account balance inquiry. The user should be able to get the account status since account balance service is defined in the general clients' profile.

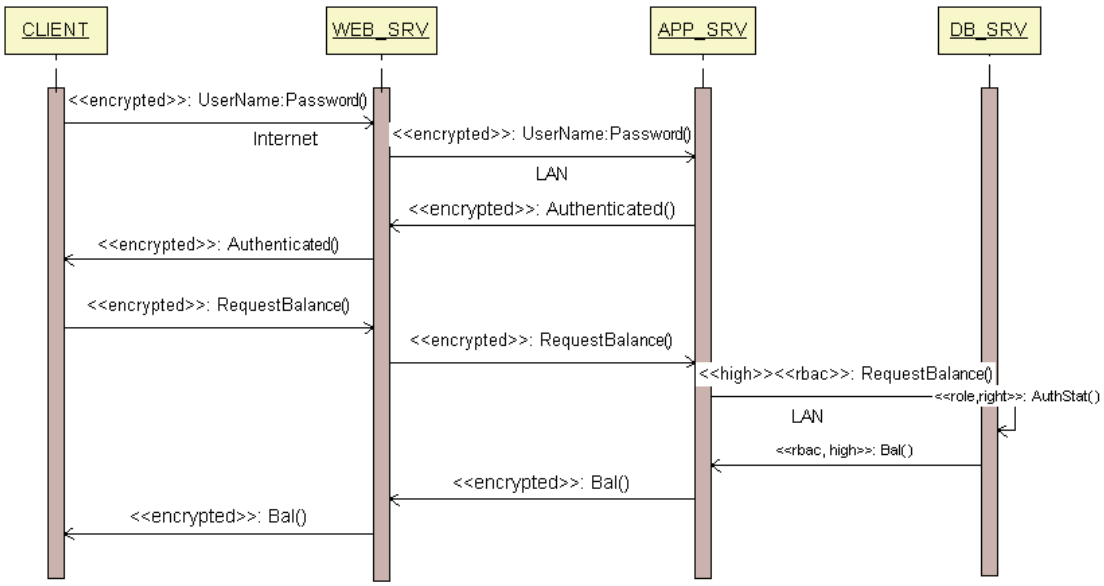
FUTURE TRENDS

A formal link between the two modeling languages needs to be established. In addition, a theoretical foundation to check some desirable properties of security requirements, such as consistency, correctness, conciseness, and completeness of requirements need to be developed. Moreover, the generation of security test cases starting from the formal security requirements specifications to check the conformity of the security implementations requires further investigation. Finally, the effects of security requirements on other non-functional requirements should be formally studied.

CONCLUSION

With the increasing dependency on information technologies for doing business using computers, palms, wireless devices, and the Internet, there is a need to formalize the process of embedding security while developing any information technology-based system. A formal process for the engineering of security in systems under development would lead to the delivery of trustworthy and dependable systems. This formal process for security engineering starts with the security requirements capture and elicitation. The use of formal modeling languages for the specification of security requirements has been addressed. The aim of this

Figure 4. E-banking sequence diagram specifying <<rbac>>



formalization is to increase the trustworthiness of the system under development, and hence, increase users' trust in the system they use to perform their tasks. Meeting security requirements captures many of the features and goals of trustworthy systems. The Goal-Oriented Requirement Language and extensions on the Unified Modeling Language to describe system security requirements have been introduced in a formal way.

REFERENCES

Amyot, D. (2003). Introduction to the user requirements notation: learning by example. *Computer Networks: The International Journal of Computer and Telecommunications Networking, 42*(3), 285-301.

Devanbu, P., & Stubblebine, P. (2000). Software engineering for security: A roadmap. *International Conference on Software Engineering* (pp. 227-239).

Elshahry, G. (2005). *A systematic approach to the management of system security reengineering process.* Master Thesis. American University of Sharjah.

Firesmith, D. (2003). Engineering security requirements. *Journal of Object Technology, 2*(1).

GRL. (2005). Retrieved from http://www.cs.toronto.edu/km/GRL

Henning, R., & Garner, L. (1999). *Using the system security life cycle plan to enforce a system security engineering process.* Government Communications Systems.

ISO/IEC 17799. (2000). International Organization for Standardization (ISO), Code of Practice for Information Security Management. Switzerland.

International Telecommunications Union ITU-T. (2002). Draft Recommendations Z.151—Goal-Oriented Requirements Language (GRL), Geneva.

Jurjens, J. (2004). Secure systems development with UML. *Viki UMLsec-Web interface tools.* Retrieved from http://www4.in.tum.de:8180/vikinew/vikiweb

Jurjens, J. (2005). *Secure systems development with UML.* Springer Verlag.

Lodderstedt, T., Basin, D., & Doser, J. (2002). *SecureUML: A UML-based modeling language for model-driven security.* 5th International Conference on the UML.

Mundy, C., deVries, P., Haynes, P., & Corwine, M. (2002). *Trustworthy computing.* Microsoft White Paper.

Object Management Group. (2003). OMG Unified Modeling Language Specification v1.5, OMG Document formal/03-03-01.

Object Management Group, Systems Modeling Language (SysML). (2005). Specification v. 0.9 Draft.

Robertson, S., & Robertson, J. (1999). *Mastering the requirements process.* New York: ACM Press.

Weinberg, J. (1997). *Quality software management.* New York: Dorset House.

Whitman, M., & Mattord, H. (2005). *Principles of information security* (2nd ed.). Thomson.

KEY TERMS

Functional Requirement: A functional requirement is a system feature or functionality that is directly visible to external system users.

Non-Functional Requirement: A non-functional requirement is a system feature that is not directly visible to the system users. However, it may have an impact of the quality of the visible system features of functional requirements.

Security Engineering Process: The security engineering process is the process of embedding system security starting from security requirements elicitation and ending with security implementation testing.

Security Requirement: A security requirement is a security feature required by system users or a quality the system must possess to increase the users trust in the system they use. In general, a security requirement is considered as a non-functional requirement.

System Model: A model is a high-level abstraction describing behavioral (dynamic) and architectural (static) aspects of a system. A desirable model is an executable and verifiable model.

Trust: Trust is a relative user's perception of the degree of confidence the user has in the system he/she uses.

Trustworthy System: A trustworthy system is a system that gains a high level of trust by its users by satisfying the specified security, privacy, safety, availability and business integrity requirements.

Models and Techniques for Approximate Queries in OLAP

Alfredo Cuzzocrea
University of Calabria, Italy

INTRODUCTION

Since the size of the underlying data warehouse server (DWS) is usually very large, response time needed for computing queries is the main issue in decision support systems (DSS). Business analysis is the main application field in the context of DSS, as well as OLAP queries being the most useful ones; in fact, these queries allow us to support different kinds of analysis based on a multi-resolution and a multi-dimensional view of the data. By performing OLAP queries, business analysts can efficiently extract *summarized knowledge*, by means of SQL aggregation operators, from very large repositories of data like those stored in massive DWSs. Then, the extracted knowledge is exploited to support decisions in strategic fields of the target business, thus efficiently taking advantage from the amenity of exploring and mining massive data via OLAP technologies. The negative aspect of such an approach is just represented by the size of the data, which is enormous, currently being tera-bytes and peta-bytes the typical orders of data magnitude for enterprise DWSs, and, as a consequence, data processing costs are explosive.

Despite the complexity and the resource-intensiveness of processing OLAP queries against massive DWSs, client-side systems performing OLAP and data mining, the most common application interfaces versus DWSs, are often characterized by small amount of memory, small computational capability, and customized tools with interactive, graphical user interface supporting *qualitative*, *trend analysis*. For instance, consider the context of retail systems. Here, managers and analysts are very often more interested in the product-sale plot in a fixed time window rather than to know the sale of a particular product in a particular day of the year. In others words, managers and analysts are more interested in the trend analysis rather than in the *punctual*, *quantitative analysis*, which is, indeed, more proper for OLTP systems. This consideration makes it more convent and efficient to compute approximate answers rather than exact answers. In fact, typical decision-support queries can be very resource intensive in terms of spatial and temporal computational needs. Obviously, the other issue that must be faced is the accuracy of the answers, as providing fast and totally wrong answers is deleterious. All considering, the key is proving fast, exploratory answers with some guarantees on their degree of approximation.

On the other hand, in the last few years, DSS have become very popular: for example, sales transaction databases, call detail repositories, customer services historical data, and so forth. As a consequence, providing fast, even if approximate, answers to aggregate queries has become a tight requirement to make DSS-based applications efficient, and, thus, has been addressed in research in the vest of the so-called approximate query answering (AQA) techniques. Furthermore, in such data warehousing environments, executing multi-steps, query-processing algorithms is particularly hard because the computational cost for accessing multi-dimensional data would be enormous. Therefore, the most important issues for enabling DSS-based applications are: (1) minimizing the time complexity of query processing algorithms by decreasing the number of the needed disk I/Os, and (2) ensuring the quality of the approximate answers with respect to the exact ones by providing some guarantees on the accuracy of the approximation. Nevertheless, proposals existent in literature devote little attention to the point (2), which is indeed critical for the investigated context.

BACKGROUND

Multi-dimensional models represent data as univocally associated to positions in a multi-dimensional space and support query and mining tasks according to a multi-resolution view of data. The growing attention towards such models was stirred up by recent advances of OLAP systems, which allow us to efficiently support data analysis for a wide range of modern application fields ranging from Business Intelligence to sensor network data management and QoS-based (quality of service) systems. Traditionally, OLAP technology was adopted for supporting just-in-time (summarized) knowledge extraction in decision-making processes of very large organizations, but its reliability and effectiveness has made OLAP engines a (very) popular component of a plethora of data-intensive systems. *Data cube* (Gray, Bosworth, Layman, & Pirahesh, 1996) is the fundamental data model of the OLAP technology, and effectively supports the above mentioned analysis goals. According to such a model, multi-dimensional data are organized in cubes that are characterized by a set of dimensions and a set of measures. The first set, which contains the analysis parameters or, more properly,

the *functional attributes*, allows us to univocally locate *data cells storing measures*, which are the values of interest for the target decision process, and on which various kinds of OLAP queries, which retrieve data by means of multi-dimensional aggregations, are executed.

Range queries (Ho, Agrawal, Megiddo, & Srikant, 1997) are an important class of OLAP queries that are very often executed on data cubes. They are defined as the application of a given *SQL aggregation operator* (such as SUM, COUNT, AVG, etc.) over a set of selected contiguous ranges in the domains of the dimensions. For instance, a *n*-dimensional range-SUM query over a *n*-dimensional data cube *A* can be generally formulated as follows:

$$SUM(l_0 : h_0, l_1 : h_1, \ldots, l_{n-1} : h_{n-1}) = \sum_{l_0 \le i_0 \le h_0} \sum_{l_1 \le i_1 \le h_1} \cdots \sum_{l_{n-1} \le i_{n-1} \le h_{n-1}} A[i_0][i_1]\ldots[i_{n-1}]$$

such that $\langle l_k : h_k \rangle$ is the range defined on the dimension d_k.

APPROXIMATE QUERY ANSWERING TECHNIQUES

There is a clear taxonomy of methods proposed for supporting approximate query answering: We distinguish between methods based on *pre-computation* and methods based on *online computation*. Methods belonging to the first class compute *synopses*, which are succinct representations of the original data, in an off-line mode, and we use them instead of the original data for answering queries, thus obtaining approximate answers. Methods belonging to the second class perform sampling at query time, so that answers can be continually improved (by progressively enlarging the data sample) under user control. Many techniques for compressing data cubes and evaluating range queries over their compressed representation have been proposed. Several compression models, which have been originally defined in different contexts, have been used to this end. In the following section, we focus our attention on both the most popular techniques such models are based on, that is, *histograms*, *wavelets*, and *sampling*.

Histograms

Ioannidis and Poosala (1999) introduced the use of histograms for providing approximate answers to set-valued queries. Histograms are data structures obtained by partitioning a given data domain into a number of mutually disjoint blocks, called *buckets*, and then storing for each bucket some aggregate information, such as the sum of their items. Histograms were originally proposed for query size estimation inside query optimizers, and to summarize multi-dimensional data distributions (Poosala & Ioannidis, 1997; Poosala, Ioannidis, Haas, & Shekita, 1996). In the latter case, the data distribution to be compressed consists of the frequencies of values of the attributes in a given relation, and queries are evaluated by performing linear interpolation on the stored aggregate values. Subsequently, histograms have been effectively used to estimate range queries in OLAP (Poosala & Ganti, 1999). Many techniques for building multi-dimensional histograms have been proposed in literature, each of them based on particular properties of the data distributions characterizing the input domains. Typically, statistical and error metrics-based properties are taken into account, and greedy algorithms are considered to mitigate the computational cost of processing very large data cubes. Among all the alternatives, we focus our attention on the following histograms, mainly because they describe approaches that we retain having a similar spirit to ours: Equi-Depth (Muralikrishna & DeWitt, 1998), MHist (Poosala & Ioannidis, 1997), and GenHist (Gunopulos, Kollios, Tsotras, & Domeniconi, 2000) histograms.

Given a *n*-dimensional data domain *D*, the Equi-Depth histogram $H_{E\text{-}D}(D)$ is built as follows: (1) fix an ordering of the *n* dimensions $d_0, d_1, \ldots, d_{n-1}$; (2) set $\alpha \approx n$-th root of desired number of buckets; (3) initialize $H_{E\text{-}D}(D)$ to the input data distribution of *D*; (4) for each *i* in $\{0, 1, \ldots, n\text{-}1\}$ split each bucket in $H_{E\text{-}D}(D)$ in α equi-depth partitions along d_i; and finally (5) return resulting buckets to $H_{E\text{-}D}(D)$. This technique presents some limitations: fixing α and a dimension ordering can result in poor partitions, and, consequently, there could be a limited level of *bucketization*.

The MHist histogram overcomes the Equi-Depth histogram performances, as stated by experimental results shown in (Poosala & Ioannidis, 1997). The MHist build procedure depends on the parameter *p* (specifically, such histograms are denoted by MHist-p histograms): contrarily to the previous technique, at each step, the bucket *b* in $H_{MH}(D)$ (i.e., the output histogram) containing the dimension d_i whose marginal is the most in need of partitioning is chosen, and it is split along d_i into *p* (e.g., $p = 2$) buckets.

GenHist histograms are a new class of multi-dimensional histograms that are different from the previous ones with respect to the build procedure. The key idea is the following: given an histogram *H* with h_b buckets on a *n*-dimensional input data domain *D*, the proposed technique builds the output histogram $H_{GH}(D)$ by finding n_b overlapping buckets on *H*, such that n_b is an input parameter. To this end, the technique individuates the number of distinct regions that is much larger than the original number of buckets h_b, thanks to a greedy algorithm that considers increasingly-coarser grids. At each step, the algorithm selects the set of cells *J* of highest density and moves enough randomly-selected points from *J* into a bucket to make *J* and its neighbor cells "close-to-uniform." Thus, the innovative contribution is to define a truly multi-dimensional split policy, based on the concept of *tuple density*.

Wavelets

Wavelets (Stollnitz, Derose, & Salesin, 1996) are a mathematical transformation which defines a hierarchical decomposition of functions (representing signals or data distributions) into a set of coefficients. They were originally applied in the field of image and signal processing. Recent studies have shown the applicability of wavelets to selectivity estimation (Matias, Vitter, & Wang, 1998), as well as to the approximation of both specific forms of query (like range queries) (Vitter, Wang, & Iyer, 1998), and "general" queries (Chakrabarti, Garofalakis, Rastogi, & Shim, 2000) (using join operators) over data cubes. Specifically, the compressed representation of data cubes via wavelets (Vitter et al., 1998) is obtained in two steps. First, a wavelet transform is applied to the data cube, thus generating a sequence of coefficients. At this step no compression is obtained (the number of wavelet coefficients is the same as the number of data points in the examined distribution), and no approximation is introduced, as the original data distribution can be reconstructed exactly applying the inverse of the wavelet transform to the sequence of coefficients. Next, among the N wavelet coefficients, only the $m << N$ most "significant" ones are retained, whereas the others are "thrown away," and their value is implicitly set to 0. The set of retained coefficients defines the compressed representation, called *wavelet synopses*. In order to execute the wavelet decomposition procedure for a given data domain D, first a wavelet basis function set must be chosen: *Haar Wavelets* (Chakrabarti et al., 2000; Vitter et al., 1998) are conceptually the simplest wavelet basis functions, and therefore, widely used in literature. It has been shown that wavelet-based techniques improve the histogram-based ones in the summarization of multi-dimensional data (Vitter et al., 1998), and, thus, they have been used for approximate query answering (Chakrabarti et al., 2000).

Sampling

Random sampling-based methods propose mapping the original multi-dimensional data domain in a smaller subset by sampling: this allows a more compact representation of the original data to be achieved. Query performances can be significantly improved by pushing sampling to query engines, with very low computational overheads. Traditionally, random sampling-based techniques have not been considered as performing as well as other more resource-intensive techniques such as multi-dimensional histograms and wavelet decomposition. However, they have recently been of renewed interest from the database and data warehousing research communities, due to their computational requirements, which are very low. (Hellerstein, Haas, & Wang, 1997) propose a system for effectively supporting online aggregate query answering, and also providing probabilistic guarantees about the accuracy of the answers in terms of *confidence intervals*. Such a system allows a user to execute an aggregate query and to observe its execution in a graphical interface that shows both the partial answer and the corresponding (partial) confidence interval. The user can also stop the query execution when the answer has achieved the desired degree of approximation. Therefore, the whole process is interactive. No synopses are maintained since the system is based on a random sampling of the tuples involved in the query. Random sampling allows an unbiased estimator for the answer to be built, and the associated confidence interval is computed by means of the Hoeffding's inequality (Hoeffding, 1963). The drawback of this proposal is that response time needed to compute answers can increase since sampling is done at query time. The absence of synopses ensures that there are not additional computational overheads because no maintenance tasks must be performed.

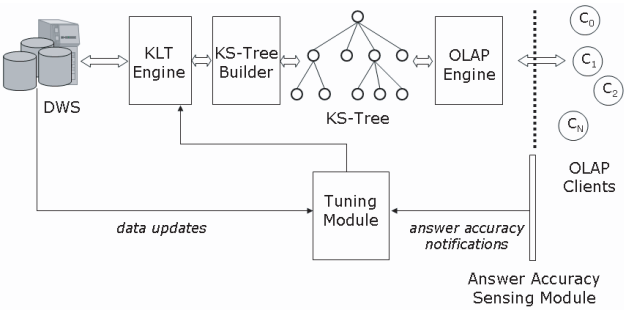
Recently, in Ganti, Lee, and Ramakrishnan (2000), a weighted sampling scheme that exploits workload information to continuously tune a representative sample of data is proposed. In Chaudhuri, Das, Datar, Motwani, and Rastogi, 2001, this novel approach is extended towards a more accurate scheme in which the presence of *data skew* (i.e., asymmetric peak values in the data distributions) and *low selectivity* of queries is considered; experimental results shown in (Chaudhuri et al., 2001) clearly show that the accuracy of the approximate answers is outperformed by taking into account a more "adaptive" scheme rather than a more "static" one. More recently, in Babcock, Chaudhuri, and Das (2003), the previous work is further extended by arguing that (1), for many aggregation queries, appropriately constructed biased samples can provide more accurate approximation than uniform samples, and (2) consequently, proposing an improved technique that dynamically constructs an ad hoc biased sample for each query by combining samples selected from a family of non-uniform samples built during the preprocessing phase.

S-OLAP:COMPRESSING MULTI-DIMENSIONAL DATA CUBES VIA PROBABILISTIC SYNOPSES

s-OLAP (see Figure 1) attempts to solve the aforementioned issues in dealing with querying massive data warehouses, provided by OLAP interfaces, and, on the basis of dimensionality reduction and *probabilistic synopses*, defines a methodology for obtaining fast, approximate answers to range queries, along with *probabilistic guarantees* on their degree of approximation. s-OLAP is a multi-user system that sits between a DWS and the OLAP clients that interact with the former by performing query and report tasks.

A very important issue in the approximate query processing research issue is just dealing with multi-dimensional data

Figure 1. The s-OLAP system



domains and queries. In fact, many interesting, state-of-the-art techniques focused on the approximation of set-valued queries on relational databases, one-dimensional and two-dimensional queries on data cubes have been proposed during the last years; nevertheless, presently, there are very few techniques that allow us to efficiently evaluate with approximation multi-dimensional queries on multi-dimensional data cubes. In order to address this critical challenge, in Cuzzocrea (2005), we propose a dimensionality reduction technique, based on the well known Karhunen-Loeve transform (KLT) (Jain, 1989). KLT allows us to obtain an *m*-dimensional data domain D_m from a given *n*-dimensional data domain D_n, such that $n >> m$, providing D_m a *summarized description* of D_n. The details of this transformation technique for obtaining dimensionality reduction of multi-dimensional data cubes are given in our article (Cuzzocrea, 2005). More specifically, in (Cuzzocrea, 2005), a MOLAP-based representation technique that allows us to significantly reduce the overall error caused by the projection of the input *n*-dimensional data cube over a smaller *m*-dimensional data domain is presented, along with some other effective optimizations for "tailoring" the KLT to efficiently support approximate query answering in OLAP.

In the s-OLAP system, our technique (Cuzzocrea, 2005) is used to obtain a *collection* of two-dimensional data domains (i.e., $m = 2$), said CoD_2, from the input *n*-dimensional data cube *A* (i.e., one transformed domain for each of the (MOLAP) data cubes representing *A*), and, the (KLT-based) *transformation relation* between *A* and the CoD_2 is exploited for building a R^+-tree like data structure, called *KS-Tree*, which is a synopsis data structure for *A*, that is, a persistent object that represents in a summarized fashion the knowledge kept in *A*. In other words, *KS-Tree* allows us to efficiently represent and query the (probabilistic) synopses built on *A*, and some others related information. Therefore, any query *Q* against the original data cube *A* is (1) redirect to the *KS-Tree* built on *A*, denoted by *KS-Tree*(*A*), and (2) evaluated against it, thus providing an approximate answer to *Q*, denoted by $\tilde{A}(Q)$.

The ratio of choosing the two-dimensional ones as output data domains of the KLT is suggested by the amenity of designing a very efficient query strategy, yet introducing low spatio-temporal costs because low-dimensional domains (i.e., 2) are processed. *KS-Tree* is also *self-adjustable*: s-OLAP periodically reconfigures it according to the *current* accuracy of the retrieved (approximate) answers, which is measured on the basis of an empirically-determined threshold τ.

The underlying theoretical framework for providing probabilistic guarantees on the degree of approximation of the retrieved answers is another important feature of s-OLAP. To this end, we adopt the so-called *tail inequalities* (such as Markov's inequality, Chebyshev's inequality, or Hoeffding's inequality (Hoeffding, 1963, etc.), which are important results coming from the theoretical statistics. Such inequalities give probabilistic bounds on the estimation of a parameter of a given statistics (such as mean value, variance, co-variance, etc.) defined on a set of random variables. It is well-recognized that, by using a tail inequality in the estimation of a given observed parameter *p*, the introduced error is at most ε with probability that is at least $1 - \delta$ (i.e., *with high probability*, as commonly intended), being ε and δ positive integers. We highlight that this property plays a critical role in our system: In fact, providing approximate answers without any bounds on their accuracy would be fruitless. To support this feature, *s*-OLAP builds probabilistic synopses starting from the original multi-dimensional data in such a way that these synopses "encode" the Hoeffding's inequality (Hoeffding, 1963) and computes the approximate answers against them, thus providing (theoretically-proofed) probabilistic bounds on the answers.

As shown in Figure 1, the main components of s-OLAP are the following:

- **KLT Engine:** It is the component that codifies our dimensionality reduction technique presented in Cuzzocrea (2005), and, given the target data cube, obtains the collection of two-dimensional data domains CoD_2.
- **KS-Tree Builder:** It is the component that takes the collection CoD_2 as input and builds the KS-Tree.
- **KS-Tree:** It is the synopsis data structure for the target data cube.
- **OLAP Engine:** It is the component that is in charge of (1) intercepting queries posed by OLAP clients; (2) parsing and re-writing them in terms of queries against the KS-Tree; (3) running the evaluation of the queries against the KS-Tree; and finally, (4) forwarding the approximate answers towards the OLAP clients.
- **Tuning Module:** It is the component that collects queries and answers on the s-OLAP interface, and periodically checks the current degree of accuracy of the answers, said *a*; if *a* is smaller than τ, then it activates the KLT Engine in order to compute a more accurate

(compressed) representation of the target data cube, thus improving the quality of the "next" answers.

CONCLUSION

Compressing multi-dimensional data cubes is an important research challenge for the database and data warehousing research communities, which still ask for solutions capable of capturing all the (complex) requirements described in this article—ranging from representation and storage issues to maintenance and query issues. At the same time, this research topic plays a leading role for the efficiency and the effectiveness of a wide range of modern IT applications, which also have a relevant impact on day-to-day real life. Given these considerations, we presented some interesting models and techniques for supporting OLAP-based summarized knowledge extraction from massive data warehouses.

FUTURE TRENDS

AQA techniques can be successfully applied in quality of answer (QoA)-based OLAP tools, which are gaining momentum in the context of next-generation DWS applications. The philosophy of such systems resembles QoS-based systems ones where client applications and service providers can mediate on the quality of available services, thus defining new service-oriented computing paradigms and drawing new scenarios for intelligent service delivering. In a similar spirit, in QoA-based OLAP tools, OLAP users/applications, and DWSs can mediate on the accuracy of the answers. This leading feature can be achieved through realizing that compressing data is a way to handle the accuracy of data (i.e., their degree of compression), and, as a consequence, the accuracy of the retrieved answers (i.e., their degree of approximation).

REFERENCES

Babcock, B., Chaudhuri, S., & Das, G. (2003). Dynamic sample selection for approximate query answers. *Proceedings of the 2003 ACM International Conference on Management of Data* (pp. 539-550).

Chakrabarti, K., Garofalakis, M., Rastogi, R., & Shim, K. (2000). Approximate query processing using wavelets. *Proceedings of the 26th International Conference on Very Large Data Bases* (pp. 111-122).

Chaudhuri, S., Das, G., Datar, M., Motwani, R., & Rastogi, R. (2001). Overcoming limitations of sampling for aggregation queries. *Proceedings of the 17th IEEE International Conference on Data Engineering* (pp. 534-542).

Cuzzocrea, A. (2005). Overcoming limitations of approximate query answering in OLAP. *Proceedings of the 9th IEEE International Conference on Database Engineering and Applications* (pp. 200-209).

Ganti, V., Lee, M., & Ramakrishnan, R. (2000). ICICLES: Self-tuning samples for approximate query answering. *Proceedings of the 26th International Conference on Very Large Data Bases* (pp. 176-187).

Gray, J., Bosworth, A., Layman, A., & Pirahesh, H. (1996). Data cube: A relational aggregation operator generalizing group-by, cross-tabs and sub-totals. *Proceedings of the 12nd IEEE International Conference on Data Engineering* (pp. 152-159).

Gunopulos, D., Kollios, G., Tsotras, V. J., & Domeniconi, C. (2000). Approximating multi-dimensional aggregate range queries over real attributes. *Proceedings of the 2000 ACM International Conference on Management of Data* (pp. 463-474).

Hellerstein, J. M., Haas, P. J., & Wang, H. J. (1997). Online aggregation. *Proceedings of the 1997 ACM International Conference on Management of Data* (pp. 171-182).

Ho, C.-T., Agrawal, R., Megiddo, N., & Srikant, R. (1997). Range queries in OLAP data cubes. *Proceedings of the 1997 ACM International Conference on Management of Data* (pp. 73-88).

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, *58*(301), 13-30.

Ioannidis, Y. E., & Poosala, V. (1999). Histogram-based approximation of set-valued query answers. *Proceedings of the 25th International Conference on Very Large Data Bases* (pp. 174-185).

Jain, A. K. (1989). *Fundamentals of digital image processing.* Prentice Hall.

Matias, Y., Vitter, J. S., & Wang, M. (1998). Wavelet-based histograms for selectivity estimation. *Proceedings of the 1998 ACM International Conference on Management of Data* (pp. 448-459).

Muralikrishna, M., & DeWitt, D. J. (1998). Equi-depth histograms for estimating selectivity factors for multi-dimensional queries. *Proceedings of the 1988 ACM International Conference on Management of Data* (pp. 28-36).

Poosala, V., & Ganti, V. (1999). Fast approximate answers to aggregate queries on a data cube. *Proceedings of the 11th*

IEEE International Conference on Statistical and Scientific Database Management (pp. 24-33).

Poosala, V., & Ioannidis, Y. E. (1997). Selectivity estimation without the attribute value independence assumption. *Proceedings of the 23rd International Conference on Very Large Databases* (pp. 486-495).

Poosala, V., Ioannidis, Y. E., Haas, P. J., & Shekita E. (1996). Improved histograms for selectivity estimation of range predicates. *Proceedings of the 1996 ACM International Conference on Management of Data* (pp. 294-305).

Stollnitz, E. J., Derose, T. D., & Salesin, D. H. (1996). *Wavelets for computer graphics.* Morgan Kaufmann.

Vitter, J. S., Wang, M., & Iyer, B. (1998). Data cube approximation and histograms via wavelets. *Proceedings of the 7th ACM International Conference on Information and Knowledge Management* (pp. 96-104).

KEY TERMS

Multi-Dimensional OLAP (MOLAP): An in-memory-storage model that represents a multi-dimensional data cube in form of a multi-dimensional array.

OLAP Engine: A software component positioned on the top of an OLAP server that is in charge of implementing OLAP query functionalities against DW data and providing answers to OLAP users/applications.

Online Analytical Processing (OLAP): A methodology for representing, managing, and querying massive DW data according to multi-dimensional and multi-resolution abstractions of them.

Online Transaction Processing (OLTP): A methodology for representing, managing, and querying DB data generated by user/application transactions according to flat (e.g., relational) schemes.

Query Engine: A software component positioned on the top of a data server (e.g., RDBMS server) that is in charge of implementing query functionalities against DB data and providing answers to users/applications.

Query Optimizer: A software component that extends the functionalities of a target query engine by optimizing the execution plan of a given query statement (e.g., SQL statement founding on joins over multiple relational tables).

Set-Valued Queries on Relational Databases: Queries returning as result a domain (or, equally, a set) of tuples extracted from relational tables.

Models in E-Learning Systems

Alke Martens
University of Rostock, Germany

INTRODUCTION

Models are everywhere. Terms like "modeling" and "model" are part of everyday language. Even in research, no overall valid definition of what a model is exists. Different scientific fields work with different models. Usually, the term "model" is used intuitively to describe something which is sort of "abstract". This is a rather vague concept, but all models have in common that they are abstractions in a broad sense and that they are developed for a certain purpose, for example, for testing and investigating parts of reality, theories or hypotheses, for communication, or for reuse. In e-learning the notion of models is frequently used in a rather naive and uncritical way. The main purpose of developing models seems to be lost in the overwhelming amount of available models. A situation has emerged where the development of a new special purpose model often seems to be much easier than the reuse, validation, or revision of existing ones.

In the following section approaches to define the term "model" will be sketched to provide a (historical) background in relation with computer science. Afterwards, an overview over existing models and different approaches to categorize e-learning models will be given. A future trend suggests a new categorization of e-learning models. The chapter closes with a conclusion.

BACKGROUND

In the 17th century the ancient Italian term *modello* became famous in fine arts. In contrast to its former narrow sense, nowadays the term is part of everyday language. "Models can be developed based on natural artifacts or things, on hypotheses, on theories, or even based on pure fiction. The modern interpretation of model is: the object which is the result of a construction process" (Martens, in press). However, the broad usage of the notion of model makes it difficult to exactly define the term. Mueller summarizes: "Each definition of 'model' is insufficient: It covers only a small range of the reach of use" (Mueller, 2005). Accordingly, the aim of the following sketch is not to give a definition of the term model, but to describe some perspectives on models and characteristics of models. Model is a cross-disciplinary concept – moreover, most models are inherently crossdisciplinary. Generally, models have in common that they are abstractions and interpretations. A model abstracts parts of the real world, or it sketches something new, which did not exist before. The model is always a summary of the main aspects of an original, as it abstracts from special parts and only takes into account what can be perceived as the generalization. Mathematically spoken, a model is a subset of a set of originals. Thus, a model is also a simplification and a reduction on the parts which are the most important for the model developer. As a model is an interpretation, the modeler's viewpoint, intention, and the purpose of the model also influence the model. A simple example might be the model of an ape—the designer of toy apes will use a completely different model of an ape than a scientist investigating ape behavior. Mueller (2005) describes the basic meaning of the term model as: "A model is a simplified part of reality or potentiality. It can be material or idealistic, graphic or abstract and describes a has-been, actual or future state". Stachowiak (1973) has summarized this in the three main characteristics of models, which are representation, reduction, and pragmatics. The *representation characteristic* of a model means that each model represents an original. This does not mean that a model must have its counterpart in reality (or the physical world). The original of a model can also be an assumption, a hypothesis, a theory, or a product of fantasy. The *reduction characteristic* implies that the model's attributes are a real subset of the attributes of the original. A model never comprises all attributes of the original. The *pragmatic characteristic* is that the model's purpose is to replace the original in a certain context, for example, to answer questions, for investigations, experiments, or under certain conditions.

Several different sources, for example, Flechsig (1983), Ludewig (2002), Mueller (2005), Reihlen (1997), and Troitsch (1990), agree about at least two perspectives on models. Models can bee seen alternatively as reproduction or representational interpretation of something (*descriptive model*), or as prescriptive interpretation of something (*prescriptive model*). This distinction focuses on two different perspectives of model development, that is, the model's background and the model's purpose. A *descriptive model* reproduces or represents a part of the real world; it is always based on an original. The model depicts something existing; it is a description and abstraction. The purpose of such a model is to document, to facilitate, to show, to allow for communication, etc. Instead of describing part of reality, *prescriptive models* describe something new, which does not exist before the model. The model itself is used to construct the original and not vice versa. A classical example for such

a model would be Charles Babbage's difference engine.

Ludewig (2002) describes yet another type of model, which he called the *transient model*. This model starts as a descriptive model which is modified and changed, and finally becomes a prescriptive model, as it not necessarily has a counterpart in reality any longer. This situation can be found if a state in the real world should be changed, but the modification might be dangerous or irreversible. Then a process might be to start with a descriptive model of the state, perform the modifications on the model (which changes the descriptive to a prescriptive model) and perform tests and experiments on the model. Later, the modifications might be applied to the state in the real world. Such a situation is a classical modeling and simulation situation.

In modeling and simulation, modeling is necessarily a part of research (see e.g., (Troitsch, 1990; Zeigler, Praehofer & Kim, 2000). Modeling and simulation is – roughly spoken – used to investigate existing or artificial systems (von Bertalanffy, 1969). The investigation takes place based on experiments performed on models of these systems. Thus, if someone wants to investigate an existing or to develop an artificial system, he usually starts with the design of a model of such a system. Some steps are required before the model can be designed, which are system identification, definition of the level of abstraction (e.g., system border, level of detail), definition of the model's purpose (e.g., investigate, experiment, teach), and decision about the model representation or the modeling language. All these steps influence how the model is designed, and how the model can be used, reused, and validated. After executing some experiments, the model is usually validated and probably refined or redesigned.

MODELS IN E-LEARNING SYSTEMS

Looking at e-learning, a large amount of different models can be found. Examples are student models (e.g., Wei, Moritz, Parvez & Blank, 2005), evaluation models (e.g., Daniel & Mohan, 2004), cognitive models (e.g., Schroeder, Moebus & Pitschke, 1995), expert knowledge models (e.g., Seitz et al, 1999), process models (e.g., Martens, 2005), and data models (e.g., LOM, 2002). These models are described in different ways, for example, graphical, formal, or verbal. Some of the models are based on modeling languages, old ones like the language of mathematic and newer ones, like the UML (Unified Modeling Language) (e.g., Booch, Rumbaugh & Jacobson, 1999). In some research papers, even the notion of metamodels occurs (e.g., Grob, Bensberg & Dewanto, 2005).

Baker (2000) has made an approach to describe roles of models in Artificial Intelligence and Education (AIED). He distinguishes between three major roles: models in AIED are used as scientific tools, as components or as basis for design. He observed that currently, these different roles of models are mixed. As Baker's distinction does no help to structure e-learning models, another approach is chosen. To structure the amount of models described earlier, Stachowiak's (1973) three characteristics can be taken into account. Optimally, for each model the model developer should explain in advance what the model represents, where abstraction took place, and what the model's purpose is. In this context, three different categories for e-learning models can be suggested: models for e-learning system development, educational models, and models of the application domain. *Models for e-learning system development* include standards (e.g., LOM, 2002), formal models like the Tutoring Process Model (e.g., Martens, 2005) or patterns (e.g., Harrer & Martens, 2006; Harrer & Martens, 2007), and software engineering models (e.g., Pawlowski, 2000). These models are used to represent (computer based) e-learning system. They abstract from the programming and realization of the e-learning program. The purpose of the models is to provide for content and implementation independent descriptions and to facilitate communication about (technical) parts, structure, and relations in the designed e-learning program. *Educational models* either have a pedagogical background or are related to educational research, for example, investigation of human learning and behavior, like cognitive models (e.g ACT-R, described at Anderson & ACT Research Group, 2001). They can for example represent theories of learning, pedagogy, and didactic. Necessarily they abstract from real human behavior in teaching and training situations. The purpose is again support in system design and communication about realization and evaluation of learning approaches which underlie e-learning systems, but on another level then the system development level. *Models of the application domain* are related to the teaching and training field. The models represent the knowledge structures on which the teaching and training content is based (e.g., Illmann, Martens, Seitzet al., 1999). However, they abstract from details. They are used to provide for content independent descriptions of the teaching and training material, and – again – are used as a basis for system design and communication about content independent knowledge structures, relations, and adaptation possibilities. An additional category might be *design models*, which are related with the HCI development (Human Computer Interfaces) and research in this area.

The distinction between descriptive, prescriptive and transient, as described in the previous section, can be applied on the three categories of models. In e-learning, *descriptive models* can be for example data models like (LOM, 2002), which document and thus help to facilitate the reproduction of learning materials. Usually, descriptive models in e-learning are models of the application domain (as sketched above), which are used for teaching and training. Unfortunately, these models are seldom communicated or made explicit, but they are implicitly represented in the way teaching and training

content is offered. Examples are case-based training (e.g., (Illmann, Martens, Seitz & Weber, 2001), the model of the inner ear (Kinshuk, Oppermann, Rashev & Simm, 1998), or a content model used in military teaching and training, for example (Rickel, Gratch, Hill, Marsella & Swartout, 2001). In educational games, learners can train models of behavior which they can later on apply if they are confronted with an equal real world situation (Prensky, 2001). In e-learning *prescriptive models* are often educational models (as sketched above). For example in (Künzel & Hämmer, 2006) the e-learning system itself is only the environment, in which a model of a learning theory or of a pedagogical strategy is tested. Another example is Krahmer and Martens (2003), where a planner is used to reason about cognitive processes based on the sequence of steps chosen by the learner in an Intelligent Tutoring System (ITS). Teaching and training systems, which are based on *transient models*, might be found in game-oriented education.

FUTURE TRENDS

The usage of a clearly defined and purpose oriented categorization of existing and new developed models, as described earlier, can facilitate communication about models – moreover, if a model category is extended by information about who is the developer and who is the potential user. It is a means to support model reuse and model validation. An approach towards model reuse is sketched by the component based plugin architecture described in (Oertel, Himmelspach & Martens, 2008). The component based design, based on model descriptions of high quality, will be a future trend.

Ongoing work leads in the direction of establishing and extending e-learning patterns as a means for design and development (as described in Harrer & Martens, 2007), and into developing a model matrix which combines the model categories with developers and potential users. Such a model matrix can then be used in e-learning system development, for example, by combining existing models or model ideas for developing a new type of system.

CONCLUSION

The intention of this chapter is to help structuring the broad field of "models", which exist in the context of e-learning. As abstractions, models are a means to facilitate communication. However, the situation is complicated if models for different purposes are compared and mixed – moreover if the application domain has a strong interdisciplinary character, as e-learning research. Thus, it is important to communicate not only the model itself, but also the model's purpose. Potentially, the intended user type of the model can be associated, for example, the computer scientist, the designer, the pedagogy expert, the learning psychologist, and the training domain expert.

A first step in this direction is the purpose oriented category which is described in this article. The category distinguishes between three main branches of model usage, that is, usage for software development, educational research, and describing an application domain. A potential fourth category would be models for design. These categories have been developed based on Stachowiak's (1973) three characteristics of models. The three main characteristics of models summarize main model properties, following the questions: What does the model represent? What does the model exclude? and What is the model's purpose? The distinction between prescriptive, descriptive and transient models, which is another general approach to distinguish models, can be used to sketch the role of existing models, but does not lend itself to be used to develop categories of models for an application domain. Baker's distinction between models as scientific tools, as components or as basis for design could be used to distinguish models, which are part of the categories described earlier. For example, a model of the application domain can be used as scientific tool, but also as basis for design (e.g., the model of the medical domain described by Illmann et al., 2001).

REFERENCES

Adorni, M., Bandini, S., Baresi, L. et al. (2003). Model requirements: Architectural model, functional model, context model, metamodel. Multichannel Adaptive Information Systems, UNIBS, R1.3.1.

Allen, R. B. (1997). Mental models and user models. Handbook of human-computer interaction (pp. 49-63). Elsevier Science

Anderson, J. R. (2000). *Cognitive psychology and its implications*. New York: W. H. Freeman and Company.

Anderson, J. R. & ACT Research Group (2001). Retrieved June 16, 2008, from http://act.psy.cmu.edu

Asendorpf, J. B. (1999). *Psychologie der persoenlichkeit*. Germany: Springer.

Baker, M. (2000). The roles of models in artificial intelligence and education research: A prospective view. *International Journal of Artificial Intelligence in Education, 11*, 122—143.

Bergin, R. A. & Fors, U. G. H. (2003). Interactive simulated patient—An advanced tool for student-activated learning in medicine and healthcare. *Journal Computers and Education, 40*(4), 361—376.

von Bertalanffy, L. (1969). *General system theory.* New York: George Braziller.

Booch, G., Rumbaugh, J., & Jacobson, I. (1999). The unified modeling language user guide. Longman, US: Addison Wesley.

Daniel, B. & Mohan, P. (2004). A model for evaluating learning objects. In *Proceedings of the International Conference on Advanced Technologies in Learning* (pp. 56—60). Finland.

Drosdowski, G. et al. (1990). *DUDEN fremdwoerterbuch*. Germany: Duden Verlag.

Flechsig, K.-H. (1983). Der Goettinger Katalog didaktischer Modelle, Germany.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1994). Design patterns: Elements of reusable object-oriented software., Reading, MA: Addison-Wesley.

Grob, H. L., Bensberg, F., & Dewanto, B. L. (2005). Model driven architecture (MDA): Integration and model reuse for open source e-learning platforms. *E-Learning Journal, 1*(online journal). Retrieved June 16, 2008, from http://eleed.camussource.de

Harrer, A. & Martens, A. (2006). Towards a pattern language for teaching and training systems. In *Proceedings of the Intelligent Tutoring Systems Conference, ITS 06*, Taiwan.

Harrer, A. & Martens, A. (2007). Using patterns for ITS development. In C. Pahl (Ed.), *Architecture solutions for e-learning systems*. Hershey, PA: Information Science Publishing, Idea Group Inc.

Illmann, T., Martens, A., Seitz, A. et al. (1999). A pattern-oriented design of a web-based and case-oriented multimedia training system in medicine. In *Proceedings of the 4th World Conference on Integrated Design and Process Technology*, Kusadasi, Turkey.

Illmann, T., Martens, A., Seitz, A., & Weber, M. (2001). Structure of training cases in web-based case-oriented training systems. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Kusadasi, Turkey.

Kaplan-Leiserson, E. (2002). Glossary. American society for training & development (ASTD). [Online Magazine] *All About e-Learning*. Retrieved June 16, 2008, from http://www.learningcircuits.org/glossary.html

Kinshuk, Oppermann, R., Rashev, R., & Simm, H. (1998). Interactive simulation based tutoring system with intelligent assistance for medical education. In *Proceedings of the Conference Education and Multimedia, Hypermedia and Telecommunications, AACE* (pp. 715-720).

Krahmer, M. & Martens, A. (2003). Reasoning about a learner's progress. In *Proceedings of the International Conference on Computers and Advanced Technology in Education*, Rhodos, Greece.

Künzel, J. & Hämmer, V. (2006). Simulation in university education: The artificial agent PSI as a teaching tool. *SCS Transactions* [Special Issue], *82*(11), 761-768.

Lakoff, G. & Núnez, R. E. (2000). *Where mathematics comes from*. New York: Basic Books.

LOM – IEEE Learning Technology Standards Committee P1484.12.1/D6.4, 03 (2002). *Draft standard for learning object metadata*. Retrieved June 16, 2008, from http://ltsc.ieee.org/wgs.html

Ludewig, J. (2002). Modelle im software engineering - eine Einfuehrung und Kritik. Proceedings of the Conference Modellierung. Tutzing, Germany.

Martens, A. (2003). Discussing the ITS architecture. In *Proceedings of the GI Workshop Expressive Media and Intelligent Tools for Learning*, Hamburg, Germany.

Martens, A. (2004). Case-based training with intelligent tutoring systems. In *Proceedings of the International Conference on Advanced Learning Technologies* (pp. 191-195). Joensuu, Finland.

Martens, A. (2005). Modeling of adaptive tutoring processes. In Z. Ma (Ed.), *Web-based intelligent e-learning systems: Technologies and applications*. Hershey, PA: Information Science Publishing, Idea Group Inc.

Martens, A. (in press). Simulation in teaching and training. In L. Tomei (Ed.), *Encyclopedia of information technology curriculum integration*.

Mueller, R. (2005). *Mueller science—Der modellbegriff: Definitionen*. Retrieved June 16, 2008, from http://www.muellerscience.com/

Oertel, M., Himmelspach, J., & Martens, A. (2008). Teaching and training system plus modeling and simulation – A plugin based approach. In *Proceedings of the European Simulation Conference, EUROSIM 08*, IEEE Computer Society Conference Publications, CPS (pp. 475-480).

Pawlowski, J. M. (2000). The Essen learning model— A multi-level development model. In *Proceedings of the International Conference on Educational Multimedia, Hypermedia and Telecommunications, ED-Media 00*, Montreal, Quebec, Canada.

Pierce, J. R. (1961). Symbols, signals and noise. New York: Harper and Brothers.

Prensky, M. (2001). Game-based learning. McGraw-Hill.

Reihlen, M. (1997). Ansaetze einer Modelldiskussion. Technical Report, University of Cologne. Cologne, Germany.

Rickel, J., Gratch, J., Hill, R., Marsella, S., & Swartout, W. (2001). Steve goes to Bosnia: Towards a new generation of virtual humans for interactive experiences. In *Proceedings of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*.

Schroeder, O., Moebus, C., & Pitschke, K. (1995). A cognitive model of design process for modelling distributed systems. In *Proceedings of the World Conference on Artificial Intelligence in Education*.

Seitz, A. et al. (1999). An architecture for intelligent support of the authoring and tutoring in multimedia learning environments. In *Proceedings of the International Conference Education and Multimedia, Hypermedia and Telecommunications*, Seattle, WA.

Stachowiak, H. (1997). Allgemeine Modelltheorie, Springer Verlag, Vienna, Austria.

Szyperski, C. (2002). Component software. *Component software series.* New York: ACM Press.

Troitsch, K. G. (1990). Modellbildung und Simulation in den Sozialwissenschaften, Westdeutscher Verlag GmbH, Opladen, Germany.

Wei, F., Moritz, S. H., Parvez, S. M., & Blank, G. D. (2005). A student model for object-oriented design and programming. In *Proceedings of the the 10th Annual Consortium for Computing Sciences in Colleges Northeastern Conference*, Providence, RI.

Zeigler, B. P., Praehofer, H., & Kim, T. G. (2000). Theory of modeling and simulation. London: Academic Press.

Zimbardo, P. G. (1998). Psychology and life. Glenview, IL: Scott, Foresman and Company.

KEY TERMS

E-Learning: Electronically supported learning.

E-Learning System: Is in the context of this chapter focused on computer-based teaching and training systems. Usually, the term covers a broad range of systems and techniques which are used in educational settings, for example, Peer Help Systems, Teleteaching, virtual classrooms, to name but a few. A more detailed definition can be found in Kaplan-Leiserson (2002).

Models for E-Learning System Development: Abstract from programming and realization, provides for content and implementation independent descriptions, and is used for communication about technical aspects of the e-learning system. Included are standards, formal models, patterns, and software engineering models.

Educational Models: Abstract from real human behavior in teaching and training. They are related to pedagogical or educational research, and can represent theories of learning, pedagogic, and didactic. They are used for communication and system design at the educational level.

Models of the Application Domain: Abstract from real content, and provide structures and relations in the teaching and training field of the e-learning system. They are used to communicate about underlying content related knowledge structures in the application domain.

Prescriptive Model: Describe something new, which does not exist before the model.

Descriptive Model: Depicts something existing. It reproduces or represents a part of the real world.

Transient Model: Starts as a descriptive model, which is performed to become prescriptive model.

Model-Supported Alignment of IS Architecture

Andreas L. Opdahl
University of Bergen, Norway

INTRODUCTION

An *information system (IS)* is a system that communicates, transforms, and preserves information for human users. An information system comprises one or more software applications and databases, and their relationships to their human users, operators, and maintainers.

A modern enterprise has many information systems that can be related in various ways. For example, information systems can be *related by exchange* because they exchange data through message passing or shared databases, or because they exchange functions through remote procedure calls or Web services. Information systems can also be *related by overlap* because they maintain the same data or provide the same functions. Information systems can be related in many other ways too, either *directly*, such as when one IS controls another, or *indirectly*, for example, because several ISs depend on the same run-time platforms or because they compete for their users' attention or for computer resources. In addition to being related to one another, information systems can be related to the *surrounding organization* in many ways. For example, organization units such as departments, individuals, or roles may be the *owners*, *users*, *operators,* or *maintainers* of ISs; organizational goals and strategies can be *realized by* ISs; organizational processes can be *supported* or *automated by* ISs; and so on.

The *information systems (IS) architecture* of an enterprise comprises its information systems, the relationships between those information systems, and their relationships to the surrounding organization. In addition to single enterprises, *alliances* of enterprises and *parts* of enterprises, such as divisions and departments, can have IS-architectures too. The above definition implies that *every* enterprise has an IS-architecture, even if that architecture is not explicitly talked about, described, or managed: 'IS-architecture' is a way to look at organizations and their information systems.[1]

IS-architecture alignment is the process of selecting an *IS-architecture vision* towards which the architecture should be incrementally but systematically evolved. This article will present a model-supported framework for aligning an IS-architecture with its surrounding organization (Opdahl, 2003a). The framework shows how an enterprise's *current* IS-architecture can be represented in an enterprise model, from which *candidate architecture visions* can then be generated, before one of them is selected as the enterprise's IS-architecture vision.

BACKGROUND

Zachman (1978) defines 'IS-architecture' as "the sum total of all information-related flows, structures, functions, and so on, both manual and automated, which are in place and/or required to support the relationships between the entities that make up the business." In the last few decades, several IS-architecture methods have been proposed in both industry and academia (Opdahl, 2003a).

A related term is *information architecture (IA)*, used by some authors (e.g., Periasamy & Feeny, 1997) as a synonym to 'IS-architecture', although IA can also be used to emphasize the information sharing and information management aspects of IS-architecture. Another related term is *enterprise architecture (EA)* (McGovern et al., 2004), sometimes called *enterprise information architecture (EIA)* (Cook, 1996), which, according to Chorafas (2002), "is to align the implementation of technology to the company's business strategy" and "to make technology serve innovation economics." 'EA'/'EIA' is sometimes used synonymously with 'IS-architecture', but can also be used to emphasize organizational aspects such as process structure and organizational roles.

IS-architecture alignment can also be understood as an intermediate step (or level) between ICT strategy and detailed IS planning (Brancheau & Wetherbe, 1986).

IS-ARCHITECTURE ALIGNMENT

A *good IS-architecture* should be *strategically* and *operationally fit* to the enterprise, *simple and well structured*, *well managed,* and *clearly and explicitly described*. These characteristics are explained as follows:

- *Strategically fit* means that the IS-architecture should support the enterprise in pursuit of its goals and strategies. This is of course the primary characteristic of a good IS-architecture.
- *Operationally fit* means that the IS-architecture should be integrated with the enterprise's *organizational structures*, such as its *market structure*, *product structure*, *process structure*, *function structure*, *organization structure,* and so on. Although operational fitness may not be a goal in itself, some degree of operational fitness is necessary to achieve strategic fitness.

- *Simple and well structured* means that the IS-architecture should not be unnecessarily complex, because a complex IS-architecture will be difficult to comprehend and understand, and difficult to change without unanticipated consequences. It will therefore be hard to manage.
- *Well managed* means that the principles, activities, roles, and responsibilities for IS-architecture maintenance and evolution should be well-defined and properly taken care of. An IS-architecture that is not explicitly and properly taken care of may start to drift and quickly become unnecessarily complex and/or strategically and operationally unfit.
- *Clearly and explicitly described* means that the enterprise should always document both its *current* IS-architecture and its IS-architecture *vision*. Whereas the current architecture should be represented by a sketch or blueprint,[2] the vision should additionally be documented by a set of higher-level evolution principles.[3]

IS-architecture alignment is the process of selecting such a set of higher-level principles—expressed as an IS-architecture vision—towards which the IS-architecture is to be incrementally but systematically evolved.

Henderson and Venkatraman's (1993) *strategic alignment model* distinguishes between the external and internal domains of businesses on the one hand, and between the business domain and the ICT domain on the other hand. In consequence, their framework distinguishes between *strategic integration*, which is "the link between business strategy and I/T strategy," and *functional integration*, which is "the link between organizational infrastructure and processes and I/S infrastructure and processes" (Henderson & Venkatraman,

Table 1. The core metatypes in the representation framework. For each metatype, a brief description is given, along with examples of possible sub-metatypes (from Opdahl, 2003a).

Metatype (Subtype examples)	***Description***
Goal (mission, vision, business objectives, etc.)	The motives/rationales for the **Activities** carried out by the **Organization Units** and for other **Phenomena**. ***Goals** can be either explicit statements or implicit ideas. They can be either shared or individual and either official or private.*
Strategy (business strategies, principles, plans and standards, etc.)	Guidelines for how **Organization Units** carry out **Activities**. *G**uidelines** can be either formal or informal.*
Organization Unit (divisions, business units, departments, work groups, employees, project groups, boards, committees, etc.)	One or more persons. *An **Organization Unit** can be either an individual or a group. It can be either permanent or temporary. Note that a **Role** is a subtype of **Organization Unit**, i.e., an individual unit at the type level. The **Role** subtype is so important in enterprise modeling that it should often have its own icon.*
Activity (functions, processes, tasks, some projects, etc.)	Actions or events that occur in the enterprise. ***Activities** can either be singular, continuous, or repeated.*
Information	A pattern of information or data that is used and/or produced in the enterprise. ***Information** can be on electronic or other formats, e.g., paper.*
Application	A software system that automates or supports an **Activity** in order to let an **Organization Unit** accomplish an **Goal**.
Database (electronic archives, libraries, etc.)	A collection of data or information in the enterprise. *A D**atabase** can be in electronic or other form.*
Basic Software (operating systems, protocols, etc.)	A group of cooperating programs that are used by Applications and Databases.
Computing Equipment (computers, peripherals, etc.)	A piece of hardware.
Network	A communication network that connects computers with other computers, peripherals, and/or networks.
Phenomenon	Any of the above, i.e., either an objective, a strategy, an organization unit, an activity, information, an application, a database, basic software, computing equipment, a network, or an instance of one of the extensional metatypes.

1993). Relative to Henderson and Venkatraman's model, IS-architecture alignment focuses on 'functional integration' and on the 'internal domain' of enterprises.

MODELING ORGANIZATIONS AND IS-ARCHITECTURES

The framework for model-supported alignment relies on a structured view of an *organization* (Opdahl, 2003a) as a collection of *organizational elements* with *organizational relationships* between them. There are different types of elements, such as *goals*, *organization units,* and *activities*. Table 1 lists the most important ones, with further details given in Opdahl (2003a). The terms used in Table 1 are only suggestions. In real projects, they should be refined into more organization-specific terms. For example, 'organization unit' could be replaced by terms such as 'division', 'department', 'project group', 'role', and 'position', and the term 'activity' could be refined into 'project', 'process group', 'process', 'process step', 'operation', 'business function', and so forth. The framework defines many types of organizational relationships too (not shown in Table 1), such as goals that are *realized-by* activities and organization units that *carry-out* activities and that are *responsible-for* goals.

According to the framework, a part of the organization forms an *IS-architecture*, which is correspondingly viewed as a collection of *IS-architecture elements* and *relationships*. Examples of IS-architectural element types are *applications* and *databases*, and an example of a relationship type between IS-architecture elements is that applications *manipulate* databases. There are also various types of relationships between IS-architecture elements and other organizational elements, such as databases that *store* information, and activities that are *supported-by* applications and that *manipulate* information.

IS-architecture elements form *IS-architecture areas*, which may comprise several different types of elements. A particularly important type of IS-architecture area is the enterprise's *information systems*, which are collections of tightly coupled applications and databases that are related to the rest of the organization. But the framework also allows for other types of IS-architecture areas that group other kinds of IS-architecture elements, for example *responsibility areas* that group information systems into larger clusters. IS-architecture areas are important because they can be used to make the IS-architecture simple and well structured, and thereby more manageable: in a good IS-architecture, the architecture areas support the enterprise's goals and strategies, and the elements in each architecture area are closely related to one another, whereas there are few relationships between elements that belong to distinct areas.

Figure 1 shows the most important organizational and IS-architectural element types in the framework. Goals and strategies are executed by the grey-shaded operational organization, whose three domains are shown as layers in Figure 1. The *organizational domain* comprises organization units, activities, and information; the *information systems domain* comprises IS-architecture areas, applications, and databases, whereas the *ICT-infrastructure domain* comprises basic software, computing equipment, and computer networks.

Importantly, many elements in the framework are *decomposable*. For example, an organization unit can have sub-units, and an information element can have sub-elements. Decomposable elements of the same type form *organizational structures*, similar to organizational *dimensions* (Armour, 2003). Examples of organizational structures are the hierarchy of organization units in an enterprise and its business function hierarchy.

Figure 1 is also a sketch of a more detailed underlying *metamodel* (Opdahl, 2003a). The metamodel can be seen as a model of a modeling language for representing IS-architectures and the organizations that surround them. The metamodel has been implemented in a graphical enterprise-modeling tool, supported by the Computas-Metis modeling and metamodeling tool family. The framework and tool can be used to represent an enterprise's *current* IS-architecture, as well as *candidate* and *selected* IS-architecture *visions*.

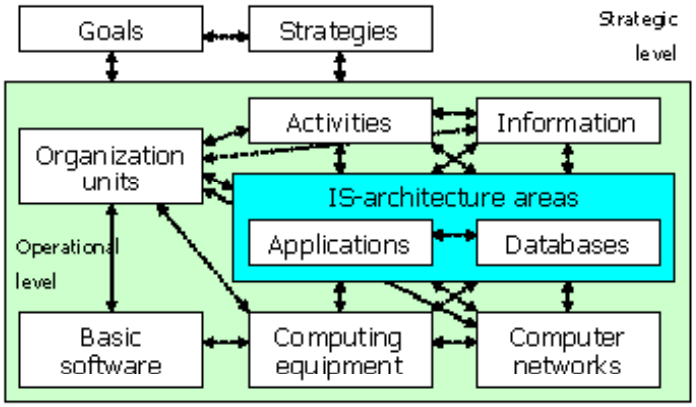
ALIGNING ORGANIZATIONS AND IS-ARCHITECTURES

The framework for model-supported alignment centers on *operational* and *strategic fitness*, the two most important characteristics of a good IS-architecture. Firstly, the framework systematically investigates alternative ways to achieve operational fitness by generating a candidate architecture vision for each of them. Secondly, the generated candidate visions are evaluated and compared according to their strategic fitness, before one of them is selected as the enterprise's IS-architecture vision. In other words, the framework investigates different ways to achieve operational fitness and selects one of them according to strategic fitness.

In the first step, different ways to align the IS-architecture with the surrounding organization are investigated systematically, inspired by an approach outlined by Kiewiet and Stegwee (1991). Each candidate architecture vision is expressed as a set of *IS-architecture principles* used for grouping IS-architecture elements into IS-architecture areas. Each principle prescribes one way to group one type of IS-architecture element and, together, the set of principles prescribes one way to group all the element types in the architecture.

A principle for grouping IS-architecture elements (of a particular type) comprises (a) an organizational structure,

Figure 1. The core of the IS-architecture representation framework. The core metamodel can be extended with metamodels that represent other metatypes, such as products and locations and IS architecture areas other than information systems (Opdahl, 2003a).



(b) possibly a hierarchical level of decomposition of that structure, and (c) a relationship, possibly indirect through several intermediate relationships, between that structure (at that level of decomposition) and an IS-architecture element (of that type). According to the principle, two IS-architecture elements will belong to the same group if they have the same relationship (c) to an element at the appropriate level of decomposition (b) in the organizational structure (a). For example, applications (a type of element) may be grouped into information systems (a type of architecture area) according to the business function (an organizational structure) that they support (a particular relationship). Furthermore, applications can be grouped into a few large information systems according to high-level business functions or into many smaller information systems according to lower-level, more decomposed functions (decomposition levels in the hierarchy of business functions) (Opdahl, 2003a).

This structured view of IS-architecture alignment, along with the equally structured views of organizations and IS-architectures presented in the previous section, makes it possible to systematically investigate all possible sets of principles, either manually or automatically. If the current IS-architecture and its surrounding organization are represented in an IS-architecture model, it is even possible to automatically generate models of each candidate architecture vision.

The first step of the alignment framework ensures that the generated candidate visions will be both *operationally fit, simple and well structured,* and *clearly and explicitly described*. But it does nothing to satisfy the most important criterion of all, *strategic fitness*. In the second step, the candidate architectures are therefore evaluated and compared according to strategic fitness. The framework does *not* currently support automatic selection of optimal or satisficing architectures leaving this task for manual assessment according to, for example, the enterprise's goals and strategies. However, automatic generation of enterprise models for candidate architecture visions should make selection easier by making assessments of and comparisons between candidates more concrete. Developing heuristics for critiquing, shortlisting, prioritizing, and selecting candidate architectures remains a topic for further work.

FUTURE TRENDS

This article has outlined how enterprise models can be used to help enterprises align their IS-architectures with the surrounding organization. But model-supported alignment of IS-architectures is only one among many important ways in which future enterprises can benefit from enterprise models. For example, new ICT systems will gradually become adaptable and manageable through enterprise models, as when a new customer resource management (CRM) system is tailored by modifying associated models of information processes and markets. Later changes to the enterprise model can then be automatically reflected in the running CRM system, so that both initial adaptation and subsequent management can be done at the model level.

When the enterprise's ICT systems thus become controllable through enterprise models, there is a danger that enterprises become more rigid and uncoordinated as a result, because the models are expressed in many different and unrelated modeling languages. This danger can be avoided by *integrating* the modeling languages and technologies

used to control different ICT systems. As a result, the enterprise can become better coordinated because its ICT systems become semantically integrated. The enterprise can also become more flexible because changing ICT systems through enterprise models is easier than changing them at the implementation level. There is therefore a need for theories and tools for tightly integrating a broad variety of enterprise models and modeling languages (Opdahl, 2003b; Opdahl & Sindre, 1997), including information and process models, actor and organization models, and goal and business-rule models. In this light, the alignment framework is only a partial contribution to developing theories and tools for tying together ICT systems in a flexible way through tightly integrated enterprise models.

CONCLUSION

The article has presented a model-supported framework for aligning an IS-architecture with the surrounding organization. One advantage of the framework is that it does not only derive blueprints of future architectures, but also generates *principles* for evolving the current IS-architecture towards the architecture vision. Another advantage is that the framework can in principle be supported by a tool, which can in the long run be developed into an *IS-architecture* (or *enterprise architecture*) *workbench*. The workbench could gradually be extended to supplement the framework with a variety of alternative approaches to ICT strategic alignment.

The framework needs to be developed further. When tool support becomes available, it must be validated by industrial case studies. Further research is also needed on how to compare candidate architectures and how to best represent and visualize IS-architectures. Also, the framework in its present form is *reactive* rather than *proactive*, because it takes the surrounding organization as a given. Although it should not be difficult to modify the framework to better support proactive use, this needs to be investigated.

Behind the alignment framework is a broader view of tomorrow's enterprises, whose ICT systems will be controlled by comprehensive enterprise models. As a result, *enterprise model integration* will become a prerequisite for ICT systems integration and thereby become the key to successful *enterprise integration*. In the enterprises of the future, the cost of establishing and maintaining large models will be shared by many different areas of use. For this to happen, new theories, technologies, and tools are needed to develop, maintain, and operate large models that integrate multiple perspectives on the enterprise and that are used for different purposes (Opdahl, 2003b; Opdahl & Sindre, 1997).

REFERENCES

Armour, P. (2003). The reorg cycle. *Communications of the ACM, 46*(2), 19-22.

Brancheau, J.C. & Wetherbe, J.C. (1986). Information architectures: Methods and practice. *Information Processing & Management, 22*(6), 453-463.

Chorafas, D.N. (2002). *Enterprise architecture and new generation information systems.* St. Lucie Press/CRC Press.

Cook, M.A. (1996). *Building enterprise information architectures—Reengineering information systems.* Englewood Cliffs, NJ: Prentice-Hall.

Henderson, J.C. & Venkatraman, N. (1993). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal, 32*(1), 4-15.

Kiewiet, D.J. & Stegwee, R.A. (1991). Conceptual modeling and cluster analysis: Design strategies for information architectures. In J.I. DeGross, I. Benbasat, G. DeSanctis & C.M. Beath (Eds.), *Proceedings of the 12th Annual International Conference on Information Systems* (pp. 315-326).

McGovern, J., Ambler, S.W., Stevens, M.E., Linn, J., Sharan, V. & Jo, E.K. (2004). *A practical guide to enterprise architecture.* Pearson.

Opdahl, A.L. (2003a). Model-supported alignment of information systems architecture. In K. Kangas (Ed.), *Business strategies for information technology management.* Hershey, PA: Idea Group Publishing.

Opdahl, A.L. (2003b). Multi-perspective multi-purpose enterprise knowledge modeling. In R. Jardim-Goncalves, J. Cha & A. Steiger-Garcão (Eds.), *Concurrent engineering: Enhanced interoperable systems—The vision for the future generation in research and applications* (pp. 609-617). A.A. Balkema Publishers.

Opdahl, A.L. & Sindre, G. (1997). Facet modeling: An approach to flexible and integrated conceptual modeling. *Information Systems, 22*(5), 291-323.

Periasamy, K.P. & Feeny, D.F. (1997). Information architecture practice: Research-based recommendations for the practitioner. *Journal of Information Technology, 12*, 197-205.

Zachman, J.A. (1978). The information systems management system: A framework for planning. *Data Base*.

KEY TERMS

Enterprise Model: A diagrammatic representation of an enterprise or part of an enterprise. An enterprise usually focuses on certain aspects of the enterprise, such as its goals and strategies, its business processes, its organization structure, its information and knowledge, etc.

Information System (IS): A system that communicates, transforms, and preserves information for human users. An information system comprises one or more computerized data systems along with their human users, operators, and maintainers.

Information Systems Architecture, IS-Architecture: The set of information systems in an organization, the relationships between those information systems, and the relationships between the information systems and the rest of the organization.

IS-Architecture Alignment: The process of selecting an IS-architecture vision that is strategically and operationally fit for the enterprise, simple and well structured, well managed, and clearly and explicitly described.

IS-Architecture Model: An enterprise model that focuses on the enterprise's IS-architecture and that can be used to represent a current architecture or to illustrate a candidate or selected architecture vision. An IS-architecture sketch is a high-level model, whereas an architecture blueprint is more detailed.

IS-Architecture Principle: A high-level rule that can be used to make decisions about developing and/or evolving individual ICT systems.

IS-Architecture Vision: A coherent set of IS-architecture principles that together guide all the aspects of IS-architecture evolution that are considered important.

ENDNOTES

1 This contrasts authors who define 'IS-architecture' as a 'blueprint' or 'sketch' of how the enterprise's ISs *are* or *should be* organized. In the terminology of this article, although a blueprint or sketch can *represent* or *describe* an IS-architecture, the blueprint or sketch *is not* the architecture.

2 The difference between the two is that a *sketch* is a rough representation that is used to communicate central aspects of an IS-architecture, whereas a *blueprint* is intended as a complete description of the architecture.

3 Principles are preferable to blueprints because a blueprint is only valid for as long as the rest of the organization stays roughly the same, whereas a principle can be useful even after the organization has undergone changes, although the enterprise's principles must of course be re-evaluated from time to time as it evolves.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2012-2017, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Moderation in Government-Run Online Fora

Arthur Edwards
Erasmus Universiteit Rotterdam, The Netherlands

Scott Wright
De Montfort University, UK

INTRODUCTION

A Dutch Internet dictionary has defined the moderator as "a person who exercises censorship on a mailing list or newsgroup."[1] Censoring the content of online discussion has often been considered as conflicting with the Internet's libertarian tradition of free speech and unrestrained communication (Tsagarousianou, 1998). However, as the famous PEN-experiment (public electronic network) in Santa Monica (1990-96) showed, the desirability of free speech must be weighed against other legitimate concerns such as the need to facilitate discussion and counteract possible abuses of the medium (Docter & Dutton, 1998).

This article analyses government-run online fora in which citizens and social organizations can discuss amongst themselves—or with government officials and elected representatives—issues of public concern. Effective moderation is considered crucial because the perceived anonymity in online fora weakens the norms of constitutive/self censorship that regulate face-to-face behaviour. It is thought that this can lead to "flame wars," polarized debates and dominant minorities. Thus, while the anonymity of online environments may diminish the psychological thresholds that can limit participation, it may also exacerbate them—inhibiting the social cooperation needed to accomplish complex communicative tasks. Moderators, it is suggested, can mitigate such problems by stimulating and regulating discussions—facilitating purposeful social action (Coleman & Gøtze, 2001; Edwards, 2002, 2004; Wright, 2006a).

Initial empirical analyses of online political discussion tended to focus on usenet newsgroups and found that debates were of poor deliberative quality and reinforced rather than changed pre-existing views (Davis, 1999; Hill & Hughes, 1998; Wilhelm, 2000). We must not extrapolate from this that all online political discussion is of poor quality—or, indeed, that all online discussion must be of high deliberative quality. The Internet provides us with a virtual commons upon which diverse interests can set up camp; the relative "free-for-all" provided by usenet can perform a useful socio-political function alongside regulated, government-led discussions. The two are not mutually exclusive. It is important that government-run online forums have clear aims, and are designed, structured, and moderated (or not) to ensure these are achieved (Wright, 2005; Wright & Street, forthcoming). A minimum level of moderation is normally required for legal reasons. Of course, this is balanced by local laws and rules on the right to free speech.

THE VARIED ROLES OF THE MODERATOR

A moderator can be defined as a person (or group of persons) who facilitates a discussion in view of its goals and agenda. Moderators can perform a wide range of functions from censorship to facilitation, dependent on the aims and context. The Guide for Electronic Citizen Consultation, published by the Dutch Ministry of the Interior (1998), mentions three moderator roles:

- **Host:** Guiding and making participants feel at ease
- **Discussion leader:** Progresses discussions and makes sure that all discussants have a chance to participate
- **Arbiter:** Designates which postings are inappropriate and removes them

Drawing on work by White (2002) and others, Coleman et al. (2001) have fleshed out this approach listing various metaphors to designate potential roles. These include: "social host," "project manager," "community of practice facilitator," "cybrarian," "help desk," "referee," and "janitor." White relates each role to specific types of communities. These designations are useful as they highlight the variety of potential functions.

Broadly speaking, two types of moderation have been adopted by governments: content moderation and interactive moderation (Wright, 2006a). To moderate the content of respondents' posts is to perform an act of censorship. Content can be moderated by electronic or human filters. Electronic filters are crude as they take no account of context and can be easily circumnavigated. Human moderation negates these problems, but raises further issues such as the subjectivity in making decisions. Content moderation is typically conducted silently: moderators do not reply to posts, facilitate discussions, or feed the discussions into the policy process. Furthermore, people whose messages

are considered inappropriate are not given an explanation for their message being deleted. This is, thus, a restricted and narrow approach to moderation. It is primarily suited to government-run discussions in which tens of thousands of messages are expected; where it would be unfeasible to adopt more interactive measures because of resource costs, and where pre-moderation would inhibit the flow of the discussion.

Governments have adopted various forms of interactive moderation by choosing specific roles from the list previously outlined to meet their aims. In this article, we go beyond a "pick and mix" approach by developing a "management" model of Internet discussions. The underlying claim of this model is that it specifies the principle tasks that have to be performed in the design and management of *decision-influencing* online policy fora. The model builds on theories of deliberative democracy, in which, citizens commit to resolve problems of collective choice through free public deliberation. Following Benhabib (1994), three principles can be derived that constitute a deliberative procedure. The first principle builds on Habermas' (1971) ideal speech situation and states that participation in deliberation is governed by norms of equality and symmetry; decisions are made by the force of arguments rather than power manoeuvres. Moderators advance these norms by attempting to promote discursiveness amongst participants and stopping the more active participants from dominating debates and agendas. They also encourage politicians and other institutional actors to participate. Benhabib's second principle states that all participants have the right to question the assigned discussion topic. This can be achieved by moderators being open to new or amended discussion topics both at the start and during the discussion. The third principle argues that everyone has the right to initiate reflexive arguments about the rules of the discourse procedure, and how they are applied. It suggests that the moderation policy should be transparent and negotiable. A users panel can be set up to resolve disputed decisions by moderators. Together with the agenda, the rules can be consolidated in a commonly agreed discussion group charter.

To specify possible moderator roles in interactive moderation, we use a management approach. This suggests that certain general "management functions" have to be performed. We distinguish (1) the strategic function, (2) the conditioning function, and (3) the process function (see Figure 1). The strategic function is to establish the boundaries of the discussion and to embed it in the political and organizational environment. This includes the following tasks:

- Establish the *goals* of the discussion, both for citizens and the institutional decision making system
- Establish and maintain the *substantive domain* of the discussion (i.e., the boundaries of the agenda within which themes and issues may be raised)
- Obtain *political and organizational support* for the discussion
- Establish the *status* of the discussion in terms of their influence on decision making
- Ensure that the *results* of the discussion will actually be carried over into the decision making process and to give feedback on this to the participants

The *conditioning* function involves the provision of all kinds of resources (including the recruitment of participants) to ensure the health of discussions such as:

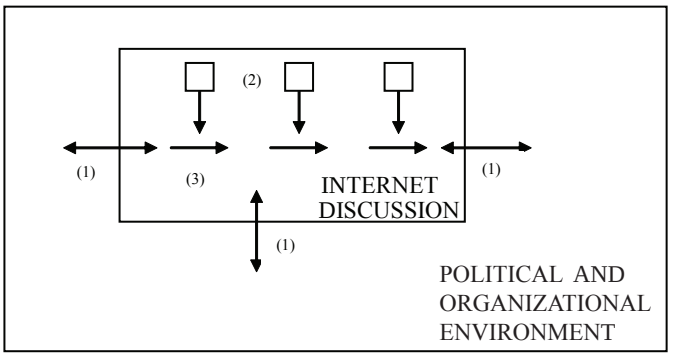
- Solicit people to join the discussion as participants
- Provide information
- Provide supporting technologies, such as moderation software, simulation models, and visualization

The *process* function includes all tasks that establish the discussion process as a cooperative, purposeful activity:

- Set the interactional goal of the discussion (i.e., the kind of results to be reached by the participants within the discussion, for instance, exploration of problem definitions or consensus about a proposal of policy measures)
- Specify the agenda of the discussion, within the substantive domain established in the strategic function: the questions, propositions, or themes
- Set the schedule of the discussion
- Manage the discussion process: its interactional goal, agenda, and schedule. For example, assign messages to discussion lines or open new discussion lines
- Facilitate the progress of the discussion by making summaries during the discussion
- Stimulate interactivity in the discussion by, for example, encouraging participants to take part in the discussion and to give reactions to specific contributions
- Set and maintain the rules of the game

As an analytical tool, this model can be used in two ways. First, in an actor-oriented way, it can be used as an instrument to discover what moderators do (Edwards, 2002). Second, in a process-oriented way, it can be used to ascertain how the different management functions are performed and which actors are involved. Used in this way, the model allows for contributions to the management of online discussions by actors other than the moderator. Especially important is the distinction between what the moderator does and what the initiators of the discussion do (Edwards, 2004—see next).

Figure 1. The management of government-run online fora (Edwards, 2002; reprinted with permission from IOS Press)



Note: 1 = strategic function; 2 = conditioning function; 3 = process function

EMPIRICAL FINDINGS ON MODERATION

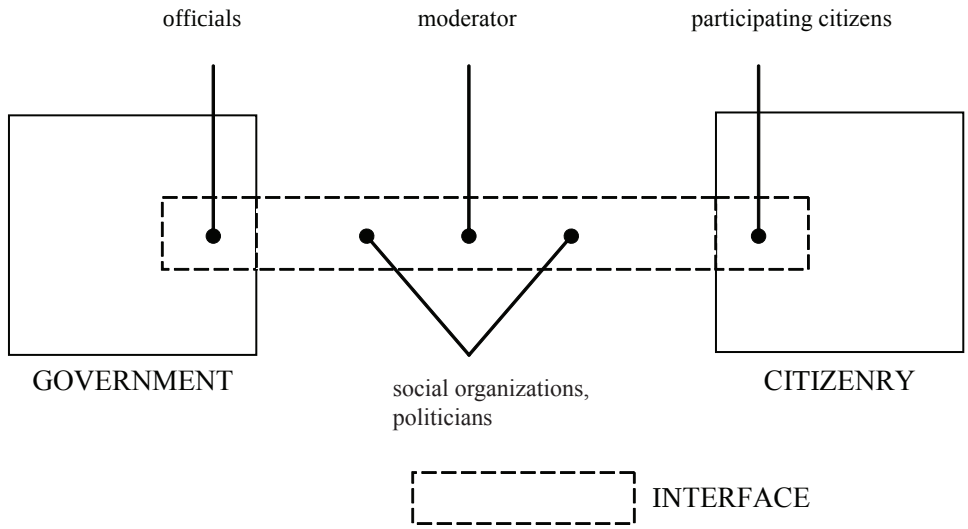
Empirical analyses have shown that governments have adopted a wide range of moderation strategies on their discussion fora—with varying degrees of success. Content moderation has proved the most controversial approach. It has led to conspiratorial atmospheres amongst participants (Coleman, Hall, & Howell, 2002); moderators receiving death threats (see Wright, 2006); and generated bad publicity for governments (Wright 2006a). Most notably, the Downing Street Web site's adoption of content moderation for its discussion fora led to accusations of control-freakery and censorship in *The Times* (Tom Baldwin, 18 March 2000), and technological naïvety in *The Observer* (Ros Coward, 20 February 2000). Empirical analysis (Wright, 2006a) has suggested that the problems were due to various unadvertised moderation policies such as removing all messages that replied to one subsequently deemed unacceptable.

Interactive moderation has proved a much more successful, and popular, method for facilitating discussion. Referring to our management model, case studies indicate that the moderator's primary involvement relates to the process function (Edwards, 2002, 2004; Lührs, Albrecht, Lübcke, & Hohberg, 2003; Jensen, 2003; Trénel, Märker, & Hagedorn, 2001; Wright, 2006a). Moderators have actively sought to enhance interactivity by, for instance, encouraging politicians and civil servants to contribute and reply. Edwards (2002) established that this activity is often conducted "behind the scene" (by the use of e-mail, for instance). In his analysis of the e-democracy forum, Wright (2006a) gives several examples of how the moderator directly intervened, thereby generating further discussion. Moderators also help to formulate discussion themes (often cooperating with the initiating agency and citizens). Furthermore, to keep the discussion on track as a purposeful activity, moderators manage discussion lines and make summaries. Finally, moderators are involved in setting and maintaining rules. The setting of the discussion rules can be a complicated process in which the initiators and the moderator may have different positions.

The extent to which moderators are required to filter offensive messages varies significantly. Edwards concluded from his research that, in practice, moderators had to intervene only infrequently. This finding is supported by analysis of the Nordpol debate in Denmark: only one message was deleted, while another user was forced to post an apology for making insulting comments about a private company. (Jensen, 2003, p. 46) Moreover, in the case of irrelevant postings or spam the debate tended to be quite self-regulatory. These findings contradict Wright's (2006a) analysis of two British central government-run fora. On the Downing Street Web sites content-moderated fora, 53.9% were missing at the end, while on the interactively moderated e-democracy forum some 26.25% of messages were blocked.[2] The differences may be explained by the facilitating activities of the moderator encouraging a more civil debate; less controversial topics; or the nature of the participants. As far as the conditioning function is concerned, in the cases analysed by Edwards moderators and initiators often cooperated in recruiting participants and in providing information—either directly or by inviting others, notably public agencies, interest groups, and political parties to provide their views on the subject.

With regard to the strategic function, there are numerous examples of moderators passing results and summaries of discussions to policy makers, although these are often subsequently ignored. However, it seems plausible that the strategic tasks, such as establishing the goals of the discussion, its status, as well as obtaining organizational and political support, are accomplished by the initiators themselves. Here we can indicate a clear demarcation between

Figure 2. Actors involved in government-run online fora (Edwards, 2004)



the management tasks that are fulfilled by the initiators and the moderators. We conclude that moderators are fulfilling important roles as intermediaries between citizens and public administrations: they enhance the interactivity and openness of online discussions as well as their relevance for public policy making.

The performance of moderators is, however, dependent on several *institutional and organizational factors*. Moderators do not function in isolation; they have to cooperate with other actors. Figure 2 depicts the interface of actors that are involved in the discussions.

Given the importance of their role, deciding who should moderate online fora is crucial. In government-initiated discussions, moderators can be civil servants, independent ("third party") moderators, or selected citizens. As intermediaries, moderators are agents of both the initiating government and participating citizens. This suggests that the relative autonomy of independent moderators places them in a better position to strike a balance. Independent moderators can still be biased though, for example when providing information or making summaries. Moderation by civil servants is acceptable if activities are transparent and negotiable; that is, if citizens have no reason to question their impartiality. Censorship is inherently subjective and open to accusations of bias—even where none exists. This has led Wright (2006a) to argue that it is necessary to separate out the censorial and facilitation roles of the moderator: censorship being conducted by an independent body with facilitation by civil servants. Involving citizens in the moderation process can also be valuable. In Bremen, an online discussion was moderated by a team consisting of four citizens, a civil servant, and an expert. They found that citizen moderators performed an important feedback function and that their local knowledge was important (Westholm, 2003). Citizen moderation and moderation by government staff are less appropriate where the issue is controversial.

Generally, the moderator's position is embedded in *organizational arrangements* in which the initiating government and sometimes also social organizations are represented. In the "project groups," "supervisory teams," or "editorial groups," the moderator, public officials, and representatives of other organizations work together in organizing and managing the discussion. These arrangements are of strategic importance, because they embed the discussion in the organizational and political environments. On the one hand, they impose some limits on the autonomy of the moderator, but on the other hand, they may enhance the relevance of the discussion, and its impact on the regular policy process.

The position of *social organizations* deserves particular attention. Social organizations can figure in various roles, for instance as initiators of a discussion, as information providers, or as participants. The involvement of social organizations in online discussions, their interaction with individual citizens, civil servants, and politicians, is an important subject for further research.

The participation of *politicians* is important, as they are the final decision makers in a representative democracy. Their commitment will be beneficial to the discussion's impact on political decision making. Politicians often show a reluctance to involve themselves in 'interactive policy exercises' with citizens, whether face-to-face or online

(Coleman et al., 2001). Moderators should encourage their participation. Equally, they must not dominate discussions as this may inhibit open citizen participation. According to Jensen (2003, p. 46), the presence of politicians "thus seems to be a two-edged sword." Here lies another example of the "sense of balance" that moderators have to show.

A possibility for strengthening the position of the *participating citizens*, and thereby enhancing the openness of the discussion, is to present the agenda and the moderation policy at the start of the discussion as proposals that are open to amendments. The outline of the discussion would then be consolidated in a commonly agreed upon charter. Also, procedural rules could be established for amendments or additions to the agenda during the discussion. Other provisions can be made to further the openness and negotiability of the moderation.[3] Furthermore, citizens can be invited to make suggestions as to relevant information or to forward information to be placed on the discussion site. Finally, the conclusions of the discussion could be consolidated in a collaboratively drafted summary.[4]

FUTURE TRENDS

A recent trend is the development of specific *discourse support systems* (i.e., systems of information and communication technologies for supporting a discussion) (Luskin, Fishkin, & Iyengar, 2004; Noveck, 2004; Sack, 2005). They provide a Web-based platform as well as a methodology and appropriate tools for enabling a fruitful dialogue. Generally, they also provide a set of features for moderating the discussion (Gordon & Richter, 2002; see also Coleman et al., 2001).[5] The moderator tasks will be more and more embedded in such systems of both social and technical components. More and more, the technical components will enable cooperative work and design by the participants.

CONCLUSION

Moderators are emerging as democratic intermediaries; facilitating government-citizen interactions in virtual environments. In this article, we focused on one such virtual environment: online discussion fora.[6] In so far as these forms will enter in the practice of democratic governance, moderators will establish themselves as elements of the information and communication infrastructure between the citizenry and public administration. Moderators can enhance the quality of online discussions and their relevance for political decision-making. We must add a note of caution though: a poorly designed and explained moderation policy can have negative consequences and thus moderation policies must be carefully considered in the light of the discussion goals.

The moderator's tasks will increasingly be embedded in socio-technical environments. The organizational arrangements, procedures, discussion rules and how the technological components are designed are all important. In developing these arrangements, due attention must be given to ensuring the openness of the discussion, and the transparency and negotiability of the moderation.

REFERENCES

Benhabib, S. (1994). Deliberative rationality and models of democratic legitimacy. *Constellations, 1*, 26-52.

Coleman, S., & Gøtze, J. (2001). *Bowling together. Online public engagement in policy deliberation.* London: Hansard Society. Retrieved July 25, 2006, from http://bowlingtogether.net/

Coleman, S., Hall, N., & Howell, M. (2002). *Hearing voices: The experience of online public consultations and discussions in UK governance*. London: Hansard Society.

Davis, R. (1999). *The Web of politics. The Internet's impact on the American political system.* New York, Oxford: Oxford University Press.

Docter, S., & Dutton, W. H. (1998). The First Amendment Online. Santa Monica's Public Electronic Network. In R. Tsagarousianou, D. Tambini, & C. Bryan (Eds.), *Cyberdemocracy. Technology, cities, and civic networks* (pp. 125-151). London: Routledge.

Edwards, A. R. (2004). The moderator in government-initiated Internet discussions: Facilitator or source of bias? In M. Mälkiä, A. V. Anttiroiko, & R. Savolainen (Eds.), *eTransformation in governance. New directions in government and politics* (pp. 150-167). Hershey, PA: Idea Group Publishing.

Edwards, A. R. (2002). The moderator as an emerging democratic intermediary. The role of the moderator in Internet discussions about public issues. *Information Polity, 7*(2002), 3-20.

Gordon, T. F., & Richter, G. (2002). Discourse support systems for deliberative democracy. In R. Traunmüller & K. Lenk (Eds.), *Electronic government. Proceedings EGOV 2002 Aix-en-Provence* (pp. 248-255). Berlin etc. Springer Verlag.

Habermas, J. (1971). *Toward a rational society.* London: Heineman.

Hill, K. A., & Hughes, J. E. (1998). *Cyberpolitics. Citizen activism in the age of the Internet.* Lanham: Rowman & Littlefield Publishers.

Jensen, J. L. (2003). Virtual democratic dialogue? Bringing together citizens and politicians. *Information Polity, 8*(2003), 29-47.

Lührs, R., Albrecht, S., Lübcke, M., & Hohberg, B. (2003). How to grow? Online consultation about growth in the city of Hamburg: Methods, techniques, success factors. In R. Traunmüller (Ed.), *Electronic government. Proceedings EGOV 2003, Prague* (pp. 79-84). Berlin etc.: Springer Verlag.

Luskin, R. C., Fishkin, J. S., & Iyengar, S. (2004). *Considered opinions on U.S. foreign policy: Face-to-face versus online deliberative polling*. Unpublished Paper. Retrieved from http://cdd.stanford.edu/research/papers/2004/online-fp.pdf

Ministry of the Interior. (1998). *Electronic Civic Consultation: First Experiences* (in Dutch), Den Haag.

Noveck, B. S. (2004). Unchat: Democratic solution for a wired world. In P. M. Shane (Ed.), *Democracy online: The prospects for political renewal through the Internet* (pp. 21-34). London: Routledge.

Sack, W. (2005). Discource architecture and very large-scale conversations. In R. Latham & S. Sassen (Eds.), *Digital formations: IT and the new architectures in the global realm*, (pp. 242-282). Princeton, NJ: Princeton University Press.

Trénel, M., Märker, O., & Hagedorn, H. (2001). *Bürgerbeteiligung im Internet-Das Esslinger Fallbeispiel*, FS II 01-308. Berlin: Social Science Research Center Berlin.

Tsagarousianou, R. (1998). Electronic democracy and the public sphere. Opportunities and challenges. In R. Tsagarousianou, D. Tambini, & C. Bryan (Eds.), *Cyberdemocracy. Technology, cities, and civic networks* (pp. 167-178). London: Routledge.

Westholm, H. (2003). Neue Medien für bessere Bürgerbeteiligung in der "Bürgerkommune"? Ein Praxisbericht. In W. Prigge & W. Osthorst (Eds.), *New media for better citizen development in the civic municipality? A report from practice.* Universität Bremen: Institut Arbeit und Wirtschaft.

White, N. (2002). *Facilitating and hosting a virtual community.* Retrieved July 25, 2006, from http://www.fullcirc.com/community/communityfacilitation.htm

Wilhelm, A. G. (2000). *Democracy in the digital age. Challenges to political life in cyberspace.* New York/London: Routledge.

Wright, S. (2005). Design matters: The political efficacy of government-run online discussion forums. In S. Oates, D. Owen, & R. Gibson (Eds.), *The Internet and politics: Citizens, voters, and activists* (pp. 80-99). London: Routledge.

Wright, S. (2006). Electrifying democracy: 10 years of policy and practice. *Parliamentary Affairs, 59*(2), 236-249.

Wright, S. (2006a). Government-run online discussion fora: Moderation, censorship and the shadow of control. *British Journal of Politics and International Relations*, *8*(4), 550-568.

Wright, S., & Street, J. (forthcoming). Democracy, deliberation and design: The case of online discussion forums. *New Media and Society*.

KEY TERMS

Content Moderation: A form of electronic or human filtering that blocks (censors) messages against pre-defined criteria. It is typically conducted silently and suitable only for discussions receiving thousands of messages.

Deliberative Procedure: A discussion that is governed by the norms of equality and symmetry in participation and the right of the participants to question the agenda and the discussion rules, as well as the way in which the agenda and rules are applied.

Discourse Support System: A system of information and communication technologies, providing a Web-based platform, a methodology, and appropriate tools for fruitful discussions.

Interactive Moderation: Moderators perform a variable suite of activities to facilitate productive debates such as bringing in external speakers, replying to messages, framing debates, and prompting further debate.

Moderator: A person or group of persons who facilitates a discussion in view of its goals and agenda.

The Conditioning Function of the Management of Online Discussions: Take care of all kinds of conditions and provisions to further the discussion.

The Strategic Function of the Management of Online Discussions: Establish the boundaries of the discussion and embed it in the political and organizational environment.

The Process Function of the Management of Online Discussions: All tasks that have to do with the discussion process as a cooperative, purposeful activity:

ENDNOTES

[1] Het Internet Woordenboek, Furore, 1999.

[2] The very high levels of deletions on the Downing Street site would appear to have been largely because stale threads were deleted in an attempt to speed up the operation of the software.

[3] In an online discussion that took place in the German city of Esslingen, there was a separate forum on the site for a reflexive 'meta-discussion' on the moderation, the relevance of the discussion, the user-friendliness of the technology, and similar issues (Trénel et al., 2001).

[4] Some of these suggestions are also formulated in a 'Dispute Resolution Flowchart' designed by the Centre for Information Technology and Dispute Resolution (1999), in: Wilhelm (2000).

[5] An example is the Delphi Mediation Online System (DEMOS): Retrieved 25 July 2006 from www.demos-project.org.

[6] ICTs can also be used to facilitate face-to-face discussions. For example, they were used in a discussion about the rebuilding of lower Manhattan in New York City (summer 2002): Retrieved 25 July 2006 from www.listeningtothecity.org.

Modern Passive Optical Network (PON) Technologies

Ioannis P. Chochliouros
Hellenic Telecommunications Organization, Greece

Anastasia S. Spiliopoulou
Hellenic Telecommunications Organization, Greece

INTRODUCTION

Presently, not only the European Union (EU) but the global community faces a decisive priority to "redesign" its economy and society, in order to meet a variety of challenges imposed by the expansion of innovative technological features, in the scope of the new millennium. The rate of investments performed and the rapid development of electronic communications networks-infrastructures, together with all associated facilities in the scope of broadband evolution, create novel major opportunities for the related market sectors (Chochliouros, & Spiliopoulou, 2005). Modern digital-based technologies make compulsory new requirements for next-generation components and for much wider electronics integration. This critical challenge also raises the issue for considering the "evolution" from current large legacy infrastructures towards new (more convenient) ones, by striking a "balance" between backward compatibility requirements and the need to explore disruptive architectures to appropriately build (and offer) future Internet, broadband, and related service infrastructures. More specifically, for the entire European market a number of evolutionary initiatives, as they currently have been encouraged by the latest EU strategic frameworks, relate first and foremost to the technological expansion and the exploitation of ubiquitous broadband networks, the availability/accessibility of dynamic services platforms, and the offering of "adequate" trust and security, all in the framework of converged and interoperable networked environments (European Commission, 2006).

However the global information society cannot deliver its major benefits without a "suitable" and appropriately deployed infrastructure, able to fulfill all requirements for increased bandwidth. During recent years, optics and photonics have become increasingly pervasive in a broad range of applications. Therefore, photonic components and subsystems are nowadays indispensable in multiple application areas, and consequently they constitute concerns of high-strategic importance for many operators. In this critical extent, fiber is constantly becoming an essential priority for wired access, as it can provide excessive bandwidth and additional advantages, if compared to similar alternative options of underlying infrastructures (Agrawal, 2002). There are several market and investment evidences demonstrating that a significant part of next-generation access networks will be based on optical access (Chochliouros, Spiliopoulou, & Lalopoulos, 2005).

This is due to the fact that we are presently witnessing an extraordinary expansion in bandwidth demand, mainly driven by the development of sophisticated services/applications, including video-on-demand (VoD), interactive high-definition digital television (HDTV), IPTV, multi-party videoconferencing, and many more. These facilities require both the existence and the use of a "fitting" underlying network infrastructure, capable of supporting high-speed data transmission rates that cannot be fulfilled by the "traditional" copper-based access networks. In fact, market actors are currently focusing on developing and deploying new network infrastructures (Leiping, 2005) that will constitute future-proof solutions in terms of the anticipated worldwide growth in bandwidth demand (reaching a rate of 50% to 100% annually), but at the same time be economically viable (Prat, Balaquer, Gene, Diaz, & Fiquerola, 2002). To this aim, fiber-access technologies evolve quite rapidly as they can guarantee "infinite" bandwidth opportunities, for all prescribed market needs, either corporate and/or residential.

BACKGROUND

A great majority of users currently benefit from rather high-speed communication services offered through DSL (digital subscriber line) access technologies. DSL's deployment has been widely supported by incumbent operators, as they were able to exploit their already laid copper infrastructure to offer broadband connectivity services to their customers, without being actually obliged to realize severe investments in access infrastructure. However, such schemes are considered as "short-term" market solutions, since the aging copper-based infrastructure is rapidly approaching its essential speed limits, while simultaneously, modern applications definitely "push" data rates beyond the capabilities of such networks. As a consequence, such networks "generate" a type of "limitation"

(or a "bottleneck") concerning requirements of bandwidth and service provision between the operator and the end user. In contrast to this option, optical access architectures allow communication via optical fibers and can provide significant advantages to the customers' needs mainly by providing a fully practical (and viable) solution to the access network "bottleneck problem," as they can support extremely high and symmetrical bandwidth to the end user (Green, 2006). Furthermore, they future-proof the network operator's CAPEX investment, as they offer simple and low-cost speed upscale, whenever necessary. While the cost of installing optical access networks has been considered as "extremely high" in the past, this has been falling progressively, and such infrastructures currently seem to be the main broadband access technology of the decade (Frigo, Iannone, & Reichmann, 2004). Optical access networks are not a new concept, as they have been considered as a potential solution for the subscriber access network for quite some time. Their deployment costs as well as their corresponding equipment costs have been dramatically reduced in recent years. Current experience has demonstrated that once fiber is installed, no significant additional investments (or reengineering) are likely to be required for the next few decades; in fact, fiber-based networks can offer fast and easy repair, low-cost maintenance, and simple upgrade. The specific category of Passive Optical Networks (PONs)—as explained in detail in the subsequent parts of this article—are now viewed as probably the "best solution" for bringing fiber to the home, since they are composed of only passive elements (fibers, splitters, splicers, etc.) and are therefore very low priced. In addition, a PON can support very high bandwidths and can function at long distances (of up to 20 km) significantly higher than these supported by high-speed DSL variants.

PON ARCHITECTURE AND DEPLOYMENT

The advent of video-on-demand and interactive gaming has prompted the deployment of immense broadband infrastructures. Because of its large bandwidth, passive optical networks are currently seen as a "proper" technology to make this happen. PON technology, nowadays being broadly adopted and deployed in multiple areas all over the world (with remarkable growth rates in North America and Japan where it provides the main solution for fiber-to-the-home (FTTH) exploitation), constitutes a convenient solution for exploiting the "undoubted" and beneficial usage of the broadband perspective (Gumaste, & Anthony, 2004; Cisco Systems, 2007).

PONs allow individual homes, larger residential or office buildings, and wider premises to be connected to public telecommunications networks directly via fiber with a high bit rate. Even across great distances, they provide users with a very high transfer capacity, which is essential for all modern data services such as high-resolution television reception or home entertainment services. PON is a very recent, *and still developing,* access technology based on the specification originally developed by the Full Service Access Network (FSAN) vendor consortium (http://www.fsanweb.org/) for the APON (ATM (asynchronous transfer mode)- based passive optical network) case. However, as discussed in a subsequent part of this article, several variants have been deployed, with distinct characteristics (Ramaswami, & Sivarajan, 2002).

A PON is a point-to-multipoint, fiber-to-the-premises network architecture where unpowered optical splitters are used to enable a single optical fiber to serve multiple premises (typically 32 different lines). A relevant configuration reduces the amount of fiber and central office (CO) equipment required, if compared with "traditional" point-to-point architectures. The "deletion" of active components implicates that the access network consists of one bi-directional light source and a number of passive splitters that divide the data stream into the individual links to each customer (Kramer & Mukherjee, 2000; Green, 2006). A PON system typically consists of optical line terminals (OLTs), optical network terminals (ONTs), optical network units (ONUs), and passive splitters, as shown in Figure 1.

The OLT is located in the network operator's CO in a telecommunications application, or in the CATV (cable TV) provider's head-end. The OLT can either generate optical signals on its own, or pass optical signals (e.g., synchronous optical network–SONET) from a collocated optical cross-connect or other device, broadcasting them downstream through one or more ports. The OLT provides the interface between the PON and the backbone network. These typically include: standard time division multiplexed (TDM) interfaces such as SONET/SDH (synchronous digital hierarchy) or PDH (plesiochronous digital hierarchy) at various rates, Internet protocol (IP) traffic over gigabit or 100 Mbit/s Ethernet, and ATM UNI (user-network interface) at 155-622 Mbit/s.

The ONU or the ONT terminate the circuit at the far end. An ONT is a single integrated electronics unit, and it is used to terminate the circuit inside the premises in an FTTP (fiber-to-the-premises) scenario, where it serves to interface the optical fiber to the copper-based inside wire. In fact, it presents the native service interfaces to the user.

An ONU is the PON-side half of the ONT, terminating the PON; it may present many converged interfaces (such as xDSL or Ethernet) towards the user. It typically requires a separate subscriber unit to provide native user services such as telephony, Ethernet data, or video. In practice, the difference between an ONT and ONU is frequently ignored, and either term is used generically to refer to both classes of equipment (Mukherjee, 1997). The ONU is used in an FTTC (fiber-to-the-curb) scenario, where the fiber stops at the curb, with the balance of the local loop being provisioned over

embedded copper (unshielded twisted pair-UTP) in conventional telecommunications networks and coaxial (coax) in CATV networks. An ONU also is used in an FTTN (fiber-to-the-neighborhood) scenario, in which it is positioned at a centralized position in the neighborhood, with the balance of the local loop being provisioned over embedded coax or UTP. While this scenario maximizes the use of embedded cable plant, *and therefore minimizes the costs associated with cable plant replacement,* it compromises performance to some extent.

The passive optical splitter "sits" in the local loop between the OLT and the ONUs (or ONTs). The splitter divides the downstream signal from the OLT at the network edge into multiple, identical signals that are broadcast to the subtending ONUs. Each OLT/ONU is responsible for determining which data are intended for it, and for ignoring all others (Keiser, 2006). Upstream signals are supported by a time-division multiple access scheme, with the transmitters in the ONUs operating in burst mode. FSAN supports both symmetric and asymmetric modes.

A PON uses small, inexpensive, low-power optical splitters, rather than the relatively large, expensive, "power-hungry" optical repeaters employed in more traditional optical networks. In particular, the neighborhood switches are replaced by cheap (or reasonably priced) passive (i.e., requiring no electric power) splitters, whose only core function is to split an incoming signal into many identical outputs (Gorshe, 2006).

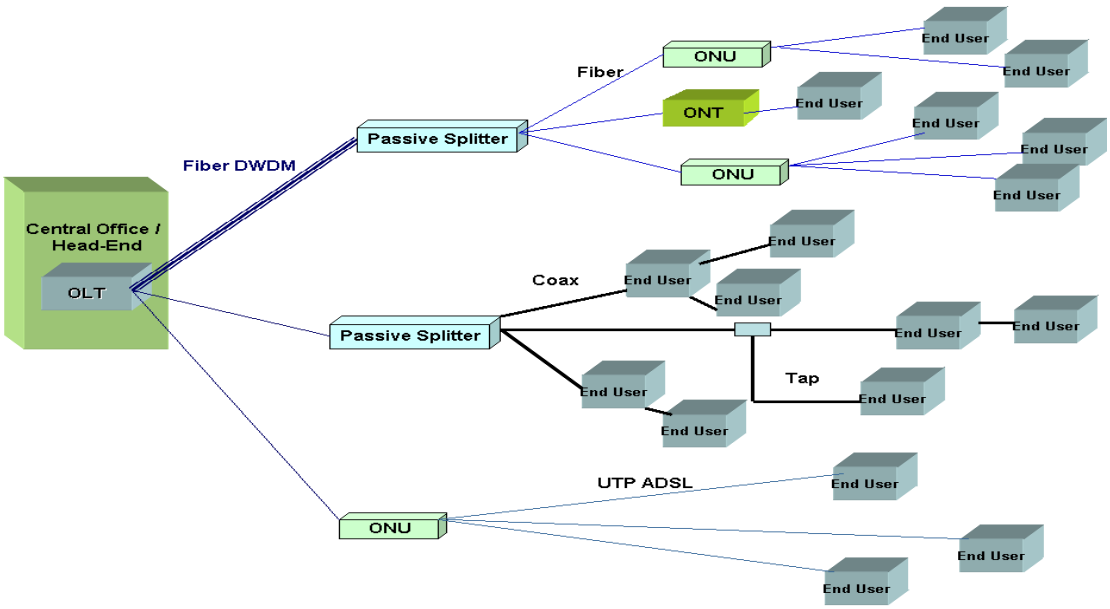
Downstream signals are broadcast to each premise by sharing a fiber. The OLT sends a single stream of downstream traffic that is "seen" by all ONTs. Each ONT only reads the content of those packets that are addressed to it, while encryption is used to prevent eavesdropping on downstream traffic. Upstream signals are combined using a multiple access protocol, invariably time division multiple access (TDMA). The OLTs "range" the ONUs in order to provide time slot assignments for upstream communication.

PON APPLICATIONS

A PON is a "pure" media network, which circumvents from any impacts caused by electromagnetic interference or lightning. As a consequence, a key reason to deploy it is to decrease the spectral interference created by copper-fed applications like asymmetric DSL (ADSL). Moreover, the fault rate is considerably decreased, bandwidth limitations are removed, reliability is strongly improved, and maintenance cost is significantly cut (as service is less expensive to maintain because there are no active loop devices and because fiber is less expensive to maintain in the long run than copper). In fact, a fiber-based PON solution using passive elements can deliver cost savings, which can add up to a 40-60% lower cash expense for labor. Savings mainly result from lower customer contacts associated with service orders and trouble reporting, outside plant operations, central office operations, and network operations (Rashid, 2004).

Simultaneously, a PON has a fine transparency and wide bandwidth, thus being applicable to a variety of signals of any format and of any bit rate. Moreover, it provides a very

Figure 1. Basic passive optical network (PON) architecture



good solution for technically and economically supporting "triple-play" services.

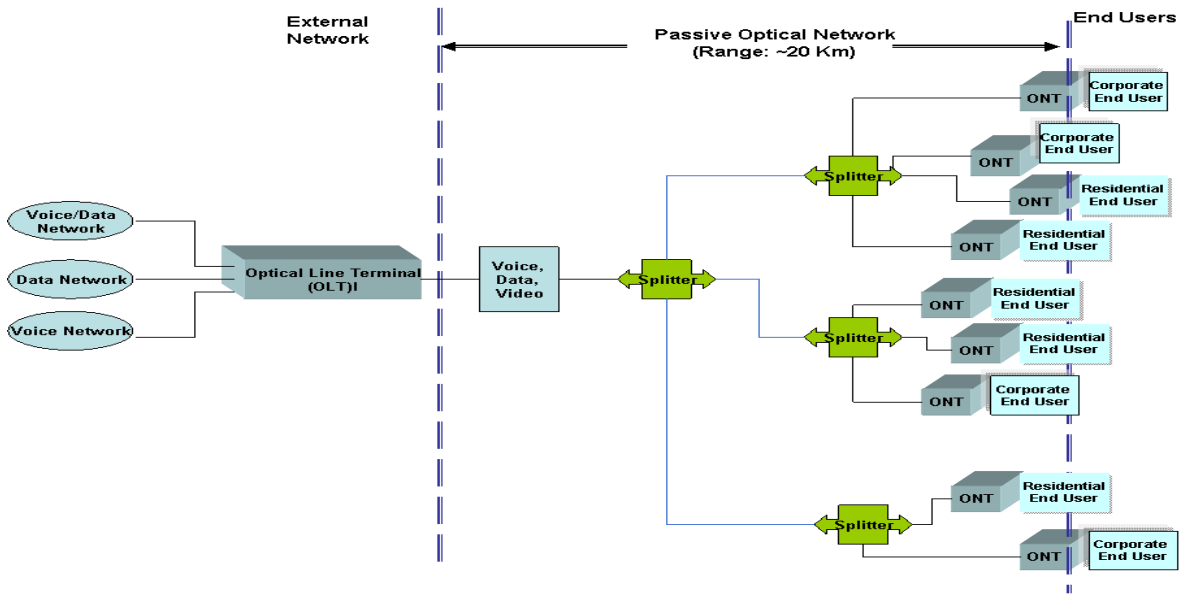
In a PON the entire downstream bandwidth is transmitted to the power splitter, and a portion of the optical power is delivered to each subscriber. Since bandwidth in a passive system is not dedicated to each subscriber, each user shares the total capacity of the system. Thus, potential customers/users have the opportunity to share end-office equipment and optical fibers, thus resulting in lower usage costs with shorter distance of the fiber, and less transmitting/receiving equipment (Nakano, 2006). However, PONs must physically restrict the number of subscribers on a power splitter to achieve higher throughputs. If the total network capacity is exhausted, then the electronics at each end (CO and CPE–customer premises equipment) must be upgraded to a newer technology.

PONs can also be used to backhaul traffic from remote DSLAMs (digital subscriber line access multiplexers) to CO-based DSLAMs, or for wireless backhaul between base station controllers and mobile switching centers. Furthermore, in a short-term, PON networks can be used in conjunction with other gear in the network. One possible configuration can suggest PON equipment to provide the backbone for an expanded DSL network, where PON extends the reach of DSL and brings it closer to the customer, allowing IPTV and VoD services to be deployed over existing copper connections to the home. Services offered can cover a broad range of network requirements like bit rate, symmetry/asymmetry or delay, and range from video distribution, *with varying degrees of interactivity,* to electronic data transfer, LAN interconnection, transparent virtual paths, and so forth.

PONs are especially attractive to today's carriers (or network providers) that have to reduce (or minimize) capital and operational costs, while maximizing the overall revenue per customer. These criteria must be achieved without sacrificing performance or network reliability, and a PON has all the ingredients to deliver these goods (Chanclou, Gosselin, Palacios, Álvarez, & Zouganeli, 2006). PONs are therefore considered as a "conformant solution" currently offered in the marketplace for bringing FTTH, since they comprise only passive elements and so they implicate low cost (Green, 2006). Moreover, when providing opportunities of high bandwidths, a PON can quite satisfactorily function at distances of up to 18-20 km in some cases, considerably higher than the distances supported by existing high-speed DSL variants (as shown in Figure 2). Besides, PONs do not engage the existence of any complex equipment (such as multiplexers/demultiplexers) in the local loop, in the proximity of the subscribers; this attribute drastically decreases the rates of installation and maintenance cost, and allows for uncomplicated upgrades to higher speeds, as such kinds of upgrades need only be performed centrally (i.e., at the network operator's central office) where the appropriate active equipment is established (ITU, 2005a).

Standardization of PONs is of fundamental meaning if they are indeed to be widely deployed and so to constitute a conformant future's broadband access infrastructure. From a network operator's perspective, standardization "translates" into cost reduction and adequate interoperability, while for a manufacturer it offers assurance that the products will successfully meet any probable market requirements, in order to be (widely) acceptable in international markets.

Figure 2. A passive optical network



Consequently, well-defined standardization effort provides certainty-guarantee for performing investments in the relevant field.

There are three fundamental PON standards in the marketplace: broadband PON (BPON), gigabit PON (GPON), and Ethernet PON (EPON). The first two generations of standards have been endorsed by the ITU (International Telecommunication Union), while the third is from the IEEE (Institute of Electrical and Electronic Engineers). The most significant differences between each type of PON technology are the supported line rates and the type of packet processing used, as discussed (among other features) in the subsequent sections.

Following the first release of the ATM-based passive optical network (APON) a few years ago, there were several releases of technical standards, all following the tracks of TDMA-based bandwidth sharing. APON was supposed to be a good solution and was intended primarily for business applications; however its use finally faded out, due to its high cost, limited service capacity, low bit rate, low efficiency, and most importantly, the decline of ATM technology in the global arena.

Broadband PON was based on APON (ITU, 2005a, 2005b). It added support for WDM (wavelength division multiplexing), dynamic and higher upstream bandwidth allocation, and survivability. It also created a standard management interface, called "OMCI," between the OLT and ONU/ONT, enabling mixed-vendor networks. A typical APON/BPON provides 622 Mbit/s of downstream bandwidth and 155 Mbit/s of upstream traffic, although the standard accommodates higher rates. BPON suffers from the very aggressive optical timing of ATM and the high complexity of the ATM transport layer.

The GPON standard (ITU, 2003) represented an improvement in both the total bandwidth and bandwidth efficiency through the use of larger, variable-length packets (PMC-Sierra, 2006). Again, the standard permits several choices of bit rate, but the industry has converged on asymmetrical operation of 2.488 Mbit/s of downstream bandwidth, and 1.244 Mbit/s of upstream bandwidth. The standard, instead of ATM encapsulation, uses GPON encapsulation method (GEM), which allows very efficient packaging of user traffic, with frame segmentation to allow for higher quality of service (QoS) for delay-sensitive traffic (such as voice and video communications). Moreover, GPON uses generic framing procedure (GFP) protocol to provide support for both voice- and data-oriented services. A big advantage over other schemes is that it interfaces to all provided main services while GFP-enabled networks packets belonging to different protocols can be transmitted in their native formats (Sims, 2007). GPON supports ATM, Ethernet, and WDM using a superset "multi-protocol" layer. The standard offers an abundant number of network management functions.

The emergence and deployment of IP technology has forwarded the concept of EPON, where APON's physical layer has been preserved, while replacing data link layer protocol (ATM) with Ethernet (Kramer, 2005). Thus, EPON was capable of providing a wider range of services with extended bandwidth, lower cost, lower complexity, and simplified timing (Hajduczenia, da Silva, & Monteiro, 2006). By sending and receiving signals within an Ethernet frame, less expensive and more versatile Ethernet parts can be used, thus helping to keep components simple and reduce costs. The basic features of this variant implicate that ATM and SDH layers have been removed. The IEEE 802.3 Ethernet PON standard (EPON or "GEPON" in order to emphasize the "gigabit" aspect of the service) was completed in 2004 (http://www.ieee802.org/3/) as part of the Ethernet First Mile Project (Beck, 2005). It uses standard 802.3 Ethernet frames with symmetric 1 Gbit/s upstream and downstream rates; however, recently (i.e., starting in early 2006), work began on a very high-speed 10 Gigabit/second EPON (XEPON or 10-GEPON) standard (http://www.ieee802.org/3/av/). EPON is applicable for data-centric networks, as well as full-service voice, data, and video networks.

However, with various emergent branches of PON technology, it is not yet possible to identify which one will be the most appropriate for next-generation optical access networks. Time to market, technology maturity, system availability, operational considerations, video compression performance, service requirements, engineering rules, and business impacts all need to be taken into account in making decisions regarding how to deploy PON (Nortel Networks, 2004).

The industry is looking at ways to deliver even more bandwidth over longer distances than ever before. Two ways of doing this are by increasing the number of optical wavelengths being used on the PON fiber, and by increasing the bandwidth and bandwidth efficiency of each wavelength. There are currently the following relative options deployed in the international environment: (i) the "traditional" TDMA-PON (Davey et al., 2006), and (ii) the "novel" WDM-PON (Banerjee et al., 2005; Lee, Sorin, & Kim, 2006) and OCDMA-PON.

FUTURE TRENDS

The main disadvantage of PONs is the requirement for complex mechanisms to allow shared media access to the subscribers so that data traffic collisions are avoided. This is due to the fact that although a PON is a point-to-multipoint topology from the OLT to the ONU (i.e., downstream direction), it is multipoint-to-point in the reverse (upstream) direction. This implies that data from two ONUs transmitted simultaneously will go into the main fiber link at the

same time, and thus will collide as the OLT is not able to discriminate them. Therefore, it is obvious that there needs to be an appropriate mechanism implemented in the upstream direction, so that data from each ONU can reach the OLT without colliding and getting distorted. An effective way to realize this can be via the usage of a wavelength division multiplexing (WDM) scheme, in which each ONU is allocated a dedicated wavelength to communicate with the appropriate OLT, so that all ONUs can use the main fiber link simultaneously (Park et al., 2004). In a similar way, the time division multiplexing (TDM) scheme can also be used, in which each ONU is allowed to transmit data only at a specific time window dictated by the OLT. The essential PON architecture permits simple service upscale whenever there is a need for more bandwidth, and this can be achieved through a combination of WDM and TDM schemes in the access fiber—that is, where different wavelengths carry different TDM PON streams or even better by using a WDM access scheme, where each subscriber is allocated a different wavelength (ITU, 2005b). In fact, WDM-PONs offer a very exciting perspective: they represent the next generation in PON development and promise to bring extended bandwidth to end users by fully utilizing the fiber's spectral windows. Although WDM-PONs truly have the potential of supporting enormous bandwidth rates (together with scalable functionality and ease of customization), their actual major drawback is associated with the high costs of the required equipment. Hence, there is currently a significant effort of research interest on finding appropriate ways to lower the high costs of such schemes, mostly by addressing the need for expensive broadband light sources. It should be expected that further progress in the area will entirely "transform" their adoption in the marketplace.

CONCLUSION

Growing demand for high-speed Internet is the primary driver for new access technologies that enable experiencing broadband. Fiber can sustain very high capacity, resulting in a high revenue potential (Hasegawa, Kuritani, Makino, Shimada, & Gorshe, 1990). A PON, utilizing passive splitters and shared-media configuration, can carry a veritable pipe of high bandwidth for users to share downstream, presenting a high-speed and inexpensive access scheme for multimedia service (Davey et al., 2006). The fundamental benefits of PON technology are flexibility, reliability, and simplicity.

The compelling advantages of PON for network operators and their customers are more than simply clear. These include a long-term life expectancy of the fiber infrastructure, lower operating costs through the reduction of "active" components, support for greater distances between equipment nodes, and most importantly, much greater bandwidth. PONs can also enable the use of new applications and services, such as high-resolution television, video telephony, e-learning, or even business applications. Furthermore, the passive splitting of the fibers means that expensive and high-maintenance active network elements are not necessary, while simultaneously the number of optical components is kept to a minimum.

Extended demand for bandwidth and progress in electro-optics have made passive optical networking into an attractive and quite convenient solution for bringing fiber to the customer. Since PONs eliminate active components on the loop, they can provide cost savings up to 10 times that of SONET, and they are now reaching cost equivalence with DSL and hybrid fiber coaxial (HFC). In addition, the combination of PON and WDM increases bandwidth availability and cost advantages even further.

Among the key reasons to deploy PONs is to decrease the spectral interference created by copper-fed applications (like ADSL). This results in a service that is less expensive to maintain, while PONs let operators go into new markets and share fibers among residential and small business customers. Other applications include buildings that are just out of reach of fiber in a metropolitan network, or even in-building networks to bring fiber to additional floors. Either way, once PON is deployed, it offers flexible bandwidth, which is important since business tenants commonly move within or out of buildings. PONs can also be used to backhaul traffic from remote DSLAMs, or for wireless backhaul between base station controllers and mobile switching centers of marketing at vendor Terawave.

During the deployment of the optical access network, market players are seriously taking into account OPEX (operational expenditure), which constitutes an integral part the lifecycle of the network. As a type of pure-medium network, the PON is nowadays considered by most operators as the "best technology that is currently available" and better oriented to synchronize with future optical access technology, although the CAPEX (capital expenditures) may be somewhat higher than that for peer-to-peer and copper wire access. In order to fulfill requirements imposed by both operators and users, PON technology may further develop according to the following alternatives:

1. Although hard in implementation due to the excessive requirements imposed on international industry, it is possible to expand the bandwidth of each single wavelength, to reach up to 10 Gbps.
2. The other option is to increase the number of wavelengths, that is from 16-wavelengths (CWDM) and 32-wavelengths (DWDM), to 64-wavelengths and 128-wavelengths, or more.

Nevertheless, crosstalk between channels brings several problems, especially when the number of wavelengths reaches a certain threshold, which results in an exponential cost increase. Consequently, WDM-PON technology alone will

not be able to fully satisfy the requirements of carriers/users in terms of the price performance ratio.

While PON has been discussed for decades, only today are network operators seriously deploying this exciting technology and increasing its penetration (Pesavento & Kelsey, 1999). Until very recently, PON deployment has been constrained by the pace of protocol standardization, equipment availability and cost, conservatism in moving to new technology, regulatory uncertainty, but most importantly by the cost of installing new fiber all the way to customer sites. However, a PON driver is the sharp decline of cost of fiber deployment in the access space over the last 10 years. In the mid-1990s, the capital cost for fiber access was around $7,500 per subscriber. This has fallen to less than $2,000 today and is expected to drop to less than $1,000 per subscriber in the near future. With this spectacular decrease in capital cost for fiber access, an enormous barrier for fiber-to-the-user (FTTU) has been virtually abolished.

Market researchers view PON as a rapidly growing market segment (Ernhofer, 2006). According to recent estimates (Rashid, 2004), the 2004 market volume amounted to approximately US$525 million, and this figure was expected to rise to 2.15 billion by 2008, corresponding to an average annual growth rate of 42%. Especially high growth is predicted for two regions: in particular North America, where high volumes of data must cross great distances; and Asia, where the next few years will see the emergence of modern urban areas with completely new infrastructures. Deployment is being driven primarily by established operators, offering a wider range of enhanced services to retain and increase their revenues. With the capability to support both today's and tomorrow's services, PON technology is an ideal solution for market players and consumers as well.

REFERENCES

Agrawal, G.P. (2002). *Fiber-optic communication systems.* New York: Wiley-Interscience.

Banerjee, A., Park, Y., Clarke, F., Song, H., Yang, S., Kramer, G., Kim, K., & Mukherjee, B. (2005). Wavelength-division-multiplexed passive optical network (WDM-PON) technologies for broadband access: A review. *Journal of Optical Networking, 4*(11)*,* 737-758.

Beck, M. (2005). *Ethernet in the first mile: The IEEE 802.3ah EFM standard.* New York: McGraw-Hill.

Chochliouros, I.P., & Spiliopoulou, A.S. (2005). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *International Journal of Infonomics, 1,* 5-21.

Chochliouros, I.P., Spiliopoulou, A.S., & Lalopoulos, G.K. (2005). Dark optical fiber as a modern solution for broadband networked cities. In M. Pagani (Ed.), *The encyclopedia of multimedia technology and networking* (pp. 158-164). Hershey, PA: IRM Press.

Cisco Systems. (2007). *Fiber to the home architectures—a white paper*. Retrieved July 14, 2007, from http://www.cisco.com/application/pdf/en/us/guest/netsol/ns547/c654/cdccont_0900aecd805df841.pdf

Chanclou, P., Gosselin, S., Palacios, J.F., Álvarez, V.L., & Zouganeli, E. (2006). Overview of the optical broadband access evolution: A joint operators in the IST network excellence e-photon/one. *IEEE Communications Magazine,* (August), 29-34.

Davey, R., Payne, D., Barker, P., Smith, K., Wilkinson, M., & Gunning, P. (2006). Designing a 21st and 22nd century fibre broadband access network. *BT Technology Journal, 24*(2), 57-64.

Ernhofer, B. (2006). *PON networks—the next evolution, white paper*. Ottawa, Canada: Zarlink Semiconductor. Retrieved June 14, 2007, from http://news.zarlink.com/in_the_news/papers.htm

European Commission. (2006). *Communication to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, on bridging the broadband gap [COM(2006) 129 final, 20.03.2006]*. Brussels, Belgium: European Commission.

Frigo, N.J., Iannone, P.P., & Reichmann, K.C. (2004). A view of fiber to the home economics. *IEEE Optical Communications, 42*(8), S16-S23.

Gorshe, S. (2006). *Introduction to passive optical networks (PON): White paper.* Retrieved July 27, 2007, from https://www.pmc-sierra.com/myPMC/download.html?res_id=13648&filename=2061015_013648.pdf

Green, P.E. (2006). *Fiber to the home: The new empowerment.* NJ: Wiley Interscience.

Gumaste, A., & Anthony, T. (2004). *First mile access networks and enabling technologies.* Indianapolis, IN: Cisco Press.

Hajduczenia, M., da Silva, H.J., & Monteiro, P.P. (2006). EPON versus APON and GPON: A detailed performance comparison. *Journal of Optical Networking, 5,* 298-319.

Hasegawa, T., Kuritani, K., Makino, K., Shimada, Y., & Gorshe, S. (1990). Optical customer access based on digital loop carrier. *Proceedings of IEEE ICC'90* (pp. 341.3.1-341.3.5).

ITU (International Telecommunication Union). (2003). *ITU-T recommendation G.984.1 (03/03): Gigabit-capable passive optical networks (GPON): General characteristics.* Geneva, Switzerland: Author.

ITU. (2005a). *ITU-T recommendation G.983.1 (01/05): Broadband optical access systems based on passive optical networks (PON).* Geneva, Switzerland: Author.

ITU. (2005b). *ITU-T recommendation G.983.2 (07/05): ONT management and control interface specification for B-PON.* Geneva, Switzerland: Author.

Keiser, G. (2006). *FTTX concepts and applications.* NJ: Wiley Interscience.

Kramer, G. (2005). *Ethernet passive optical networks*. McGraw-Hill Communications Engineering.

Kramer, G., & Mukherjee, B. (2000). *Design and analysis of PON.* Retrieved July 12, 2007, from http://wwwcsif.cs.ucdavis.edu/~kramer/papers/epon_sws.pdf

Lam, C. (2007). *Passive optical networks.* Elsevier.

Lee, C.H., Sorin, W.V., & Kim, B.Y. (2006). Fiber to the home using a PON infrastructure. *Journal of Lightwave Technology, 24,* 4568-4583.

Leiping, W. (2005). Development and prospects for fiber communication technologies. *Huawei Technologies, 18,* 1-8.

Mukherjee, B. (1997). *Optical communication networks*. New York: McGraw-Hill.

Nakano, Y. (2006, April 20-21). *Technologies and applications of passive optical networks (PON). Proceedings of the ITU-T Workshop on NGN and its Transport Networks*, Kobe, Japan. Retrieved June 21, 2007, from http://www.itu.int/ITU-T/worksem/ngn/200604/presentation/s6_nakano.pdf

Nortel Networks. (2004). *White paper: Pondering passive optical networking deployment.* Retrieved July 14, 2007, from http://www.nortel.com/solutions/brdbndaccss/collateral/nn109280-092304.pdf

Park, S.-J., Lee, C.-H., Jeong, K.-T., Park, H.-J, Ahn, J.-G., & Song, K.-H. (2004). Fiber-to-the-home services based wavelength-division-multiplexing passive optical network. *Journal of Lightwave Technology, 22*(11), 2582-2590.

Pesavento, G., & Kelsey, M. (1999). PONs for the broadband local loop. *Lightwave, PennWell, 16*(10), 68-74.

PMC-Sierra. (2006). *White paper: GPON FTTH market and technology overview.* Retrieved July 21, 2007, from http://www.unik.no/personer/aas/unik4350/GPON_FTTH_Market_and_Technology.pdf

Prat, J., Balaquer, P.E., Gene, J.M., Diaz, O., & Fiquerola S. (2002). *Fiber-to-the-home technologies.* Boston: Kluwer Academic.

Ramaswami, R., & Sivarajan, K.N. (2002). *Optical networks* (2nd ed.). San Francisco: Morgan Kaufmann.

Rashid, S. (2004). *PON delivers optical access to the masses.* Retrieved June 6, 2007, from http://www.alcatel.com/bnd/fttu/18282_FTTU_article_final.pdf

Sims, P. (2007). Preparing networks for GPON migration. *Telecom Infrastructure,* (January-February), 16-18.

KEY TERMS

Asymmetric Digital Subscriber Line (ADSL): Transmission technology that consists of modems attached to twisted-pair copper wiring that transmit from 1.5 Mb/s to 8 Mb/s downstream (to the subscriber) and up to 1.5 Mb/s upstream, depending on line distance.

Broadband PON: A term used to refer to the entire system described by the G.983.x family of ITU-T Recommendations. This includes a wide range of broadband services and goes beyond ATM access.

Ethernet: A large, diverse family of frame-based computer networking technologies that operates at many speeds (typically at 10, 100, or 1000 Mb/s) for local area networks (LANs). The name comes from the physical concept of the ether. It defines a number of wiring and signaling standards for the physical layers, through means of network access at the media access control (MAC)/data link layer, and a common addressing format. (Ethernet has been standardized as IEEE802.3.)

Ethernet Passive Optical Network (EPON): A type of PON technology that runs on the Ethernet protocol. EPON is applicable for data-centric networks, as well as full-service voice, data, and video networks.

FTTH (Fiber To The Home): A form of fiber optic communication delivery in which the optical signal reaches the end user's living or office space.

Optical Line Termination (OLT): The service provider endpoint of a passive optical network; placed at the central office or head end of a fiber-based system. Also called *optical line terminal.*

Passive Optical Network (PON): Network in which fiber optic cabling (instead of copper) brings signals all or most of the way to the end user. It is described as passive because no active equipment (electrically powered) is required between the central office (or hub) and the customer premises. Depending on where the PON terminates, the sys-

tem can be described as an FTTx network, which typically allows a point-to-point or point-to-multipoint connection from the central office to the subscriber's premises; in a point-to-multipoint architecture, a number of subscribers (for example, up to 32) can be connected to just one of the various feeder fibers located in a fiber distribution hub, dramatically reducing network installation, management, and maintenance costs.

Point-to-Multipoint (P2MP, PTMP, or PMP) Communication: Refers to communication that is accomplished via a specific and distinct type of multipoint connection, providing multiple paths from a single location to multiple locations.

SONET (Synchronous Optical NETwork): A protocol for backbone networks capable of transmitting at extremely high speeds and accommodating gigabit-level bandwidth. It has been standardized by the American National Standards Institute (ANSI).

Triple-Play Services: The ability of a telecommunications operator to supply voice, data, and video applications all at once. A typical example of a triple-play proposal would include one or multiple phone lines, a high-speed Internet connection, and television/video services (such as HDTV), all offered by the same provider. Also known as *bundled services.*

Monitoring Strategies for Internet Technologies

Andrew Urbaczewski
University of Michigan-Dearborn, USA

INTRODUCTION

Most large organizations that provide Internet access to employees also employ some means to monitor and/or control that usage (Reuters, 2002). A 2005 AMA report indicates that 76% of companies monitor worker's Web surfing, while 26% have fired workers for improper Internet usage (AMA, 2005). This chapter provides a classification and description of various control mechanisms that an organization can use to curb or control personal Internet usage. Some of these solutions are technical, while others rely instead on interpersonal skills to curb cyberslacking.

After a review of goals for a monitoring program, a list of different activities to monitor and/or control will also be provided. Then a discussion of different techniques for monitoring and associated products will be explored, followed by a discussion of fit between corporate culture and monitoring.

BACKGROUND

The Worker's Perspective

In this age of cell phones, pagers, wireless PDAs, e-mail, and home network links, many employees may feel like the employer owns them not just during the workday, but perhaps constantly. Though tiresome, the worker may accept this as an unfortunate circumstance of 21st century knowledge work. However, in the tit-for-tat that this availability demands, the employee may feel that he or she should be allowed to use the Internet at work to take care of quick business tasks such as paying bills, sending an e-mail, or checking that evening's movie listings. So long as it isn't excessive, the employee may wonder why the employer even cares. Employers can and do care for many reasons, some more profound than others.

Goals for Monitoring

Why do companies monitor their employees? Organizations monitor for many reasons including simply "because they can." An electronic monitoring effort is often difficult to establish and maintain, so before an organization begins such an effort, there should be clear monitoring goals.

The popular press is filled with stories of employees frittering away time on the Internet (Swanson, 2002). In the beginning, employees were likely to spend unauthorized time on the Internet at pornography and gambling sites, but now news and online shopping are more likely outlets for cyberslacking (Reuters, 2002). This is quite the opposite of employers' expectations when they implemented Internet connections.

Responding to these challenges, employers created acceptable use policies (AUPs). Some organizations already had AUPs implemented to keep out electronic games, and they simply modified those policies. Other organizations created new AUPs, which directly addressed the Internet's productivity threat. AUPs are useless without enforcement, but in today's litigious society it behooves accusers to be certain of transgressions before enforcing the policy. Monitoring tools create an irrefutable log of usage which can stand as legal evidence. Some employers hope the mere threat of punishment will keep employees from cyberslacking, often with some success (Urbaczewski & Jessup 2002). Listed next are some possible goals of a monitoring effort.

Increase Employee Productivity

The Internet was introduced into many organizations as a tool to increase employees' efficiency. While traditional IT packages provided few opportunities for employees seeking to slouch on employer time, the Internet posed an entirely different situation. Computers now had the capability to be an electronic equivalent of a water cooler, break room, or smokers' perch. To curb the potential problem of employees wasting time while appearing to be busy, an organization could implement a monitoring program which completely blocks and/or records the amount of time spent at non-work-related Internet sites. An alternative could be limiting access to frivilous sites to non-production hours only, such as during lunchtime.

Bandwidth Preservation

In some organizations, concerns are not productivity-based but rather that network bandwidth is being dominated by applications and instances not directly work related. An example might be listening to streaming audio or watching streaming video, both constant drains on bandwidth. People

can also engage in excessive file transfers across networks which results in reduced network performance. Two possible solutions to this problem are to purchase more bandwidth or limit the usage of existing bandwidth, with monitoring programs aiding in the latter solution.

Legal Liability Reduction

Along with productivity and bandwidth usage, organizations are also concerned about Internet usage from the potential exposure it brings to legal liability (Langin, 2005). Consider the following fictitious scenarios:

- "Organization X today was sued for negligence, as an employee was running a child pornography server inside the corporate network."
- "ABC corporation today was sued by a former employee who is now in treatment with Gambler's Anonymous. He is charging that ABC, by placing an unrestricted Internet terminal on his desktop, essentially gave him unfettered access to the virtual casinos thriving on the Internet."
- "Company B is defending itself today against a privacy lawsuit. It is charged that when an employee downloaded a file-sharing program, that program was equipped with a backdoor, which allowed malicious hackers entrance into Company B's networks. These hackers then took thousands of credit card numbers and personal data from the databases..." (a similar real-world incident happened with employees of the state of Oregon in May 2006, where about 2,200 taxpayers had their personal information compromised due to spyware picked up on a laptop computer by a worker surfing pornographic sites during downtime (Keizer, 2006).

Organizational Secret Protection

As blogging has become a popular way of communicaitng in the 21st century, many companies have realized that it is easy for employees to write their ideas and opinions about circumstances at their corporations, often with embarrassing results for the employee and/or the employer. Google and Microsoft are two well-known examples of companies that have dealt harshly with non-approved blogging by their employees, and a scandalous sex-for-money-and-favors blog detailing the life of a staffer in a U.S. Senator's office caused much embarrassment for those involved. Organizations may not want their internatl business practices or daily routines exposed to the outside world, so they use control of blog postings to keep this control. While their control over employee postings when not in the office may be murky, the organization may exert its right to keep this control during business hours.

Other possibilities like sexual harassment suits and industrial espionage make the legal risks mount. Organizations indeed may wish to monitor Internet connections to prevent any potential legal liabilities from allowing illegal activities to be conducted on their networks.

STRATEGIES AND TECHNIQUES FOR MONITORING

Different Monitoring Strategies

Once an organization decides it will monitor, it needs to know what to monitor. While Web pornography is probably the most reported off-topic use of the Internet in an organization, it is certainly not the only transgression that might come from an Ethernet card. Excessive personal e-mail, filesharing, instant messaging, multimedia streaming, and usenet browsing and posting are among other ways that employees use the corporate Internet connection for personal enjoyment.

There are several different control mechanisms that an organization might use, generally grouped into one of two categories: managerial and technical. The managerial techniques for monitoring are similar to ways that monitoring of employees has been done for decades: walking around and keeping one's eyes open. When a manager starts to wonder about an employee's performance or collegiality, then the manager starts to pay more attention to that employee's work habits.

Overall, however, the most popular means of monitoring employees is through technology. In many ways, this makes sense—a technical solution to a technological problem. Electronic monitoring operates like "big brother" (Zuboff, 1988), keeping a constant watchful eye on the network and its connected systems (or whatever subset of those systems/hours that a manager may choose to watch). Records can then be kept and offered as later "proof" of an offender's cyberslacking or lack thereof.

Electronic Monitoring Techniques

Logging at the Gateway

Many logging technologies are designed to capture and record packets as they enter and leave the organization, or at least the header information that indicates the sender, recipient, and content of the message. Gateway logging is useful in that it provides a central point of network control. However, it is difficult to accurately gauge how long an employee stares at a particular page, and if all that time he or she is actually staring at that page or if he or she has actually gone to lunch and returned later. Moreover, gateway logging can be

defeated by the use of encryption tools like PGP (www.pgp.com, see McCullagh (2001) for a more detailed description of an FBI case with the Phildelphia organized crime ring), or even tools like Anonymizer.com that allows a person to surf the Web anonymously using their gateways and encryption tools. In cases where these technologies are used, a separate technology might also be needed.

Blocking: Keeping Productive Employees Productive

Sixty-five percent of companies (AMA, 2005) utilize blocking software to keep employees from visiting certain Web sites. Organizations may not even look at the logs to see who was trying to view these sites--rather it serves simply as a friendly reminder that these sites are not to be viewed on company time. While the software is never foolproof, and relays can be used to get around the blocking software, it reminds the employees of the policies and makes a policy violator's defense of "I didn't know I wasn't supposed to do this" less believable. Akin to locking the doors on one's car that will not stop a determined thief, it does make the owner's intentions known and makes the target less appealing.

Spying at the Client

When gateway logging is insufficient, one can monitor and record connections directly at the source. A keystroke logging program can record everything that a person types on a computer, and many even include technologies to take screenshots or use the Web camera on the desk to prove that it was the person actually sitting at the computer and not someone who just walked up to the terminal.

Client sniffing programs are excellent at recording exactly what the user is doing with the computer at any given time. Many will record all the user's keystrokes, mouse movements, and active windows, allowing the reconstruction of the entire computing session. Moreover, they can capture other undesirable activity, such as playing games and typing job application letters. However, these programs are not without their own faults. First of all, the manager must install the program on the user's computer, which may not be as easy as it sounds, especially with laptop and other mobile computers. Second, the program must not be detectable (and thus deletable) by the monitored employees. Managers then must sift through mountains of captured data to determine if there is any untoward activity, or enough to warrant further investigation. However, products are available which meet the above concerns to varying degrees, and the next section will discuss some of those products.

Software Products for Controlling Internet Usage

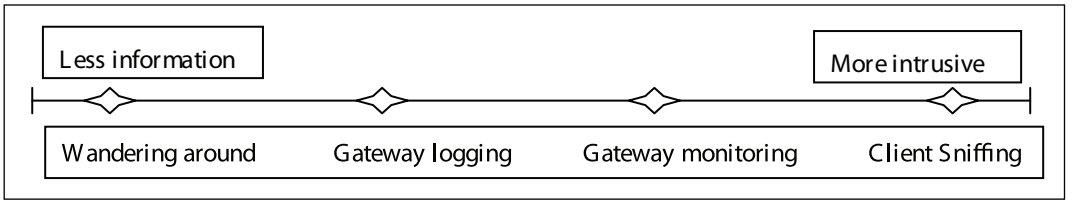
As previously mentioned, there are various products available to serve as control mechanisms. They are grouped below into five categories. Note that software products come and go, and the availability of certain products and companies are subject to market forces. The categories themselves should remain stable for the time being, although who knows what the future may hold. For example, if this chapter was being written in 1999, there would likely be no section on file-sharing.

Web Monitoring Products

As the popular press articles have largely focused on employee cyberslacking as a problem with personal Web usage, a number of products have been created to help employers manage these problems. The software products are all customizable to some degree, but there are two main classifications of these products: those that *monitor* and record Web usage, and those that actively *block* access to certain Web sites deemed inappropriate. The listing below, which is not intended to be exhaustive, details several of these products.

Cybersitter (www.cybersitter.com), *NetNanny* (www.netnanny.com), and Bsafe Online (www.bsafehome.com) are three programs that are geared largely at individuals, as they are installed on the client and maintain logs of Web pages seen by the users. *Websense* (www.Websense.com) however, is designed to monitor the Web usage of an entire network. It runs near the firewall and logs all Web requests leaving the network. All of these programs can be configured to block and/or record access to certain Web sites. Some

Figure 1. Monitoring strategies present a tradeoff between information and intrusiveness



of these programs can be tailored to allow different access rules at different times of day. For example, an organization may wish to allow its employees to use the Internet for shopping and other personal entertainment before and after normal business hours and on the lunch hour but not during the work day.

E-Mail Monitoring Products

E-mail can easily be monitored by simply examining accounts for incoming mail or logging the actions of the simple mail transport protocol (SMTP) server for outgoing mail. These logs are often difficult to read, especially with large volumes of mail and users. A series of products can assist in parsing the logs or searching for users or keywords, like *MIMESweeper* (www.mimesweeper.com) and *message inspector* (www.cyberguard.com). Encrypted e-mail can present its own challenges, and additional policies may be necessary regarding the use of strong encryption in an organization.

Monitoring e-mail sent through popular Web-based providers like Yahoo! or Hotmail can be difficult as well, because the message never passes through the SMTP servers for the organization, nor does the organization have direct control over users' mailboxes. Monitoring these types of mail services is usually done through a general monitoring tool, as listed in another section below.

File-Sharing Monitoring Products

File-sharing has a history of waxing and waning between one of the easiest applications to monitor to one of the toughest. Users of file-sharing services often try to devise ways to run their services around and through corporate attempts to halt them. Other problems were created by users demanding that they be allowed to use these programs, especially at high-profile universities like Yale and Indiana. In those cases, something had to be done to limit the amount of bandwidth these services could use, because other legitimate traffic was being delayed. A number of hardware and software solutions cropped up to aid network managers in their quest to reduce or eliminate file-sharing traffic.

On the software side, it was previously mentioned that already existing firewalls can be configured to block traffic on certain TCP (OSI layer 4) ports. Other programs, like *DynaCommI:scan* (http://www.futuresoft.com/products/antispyware/overview), are designed to examine the packets at the application layer (OSI layer 7) to determine the type of packet and whether or not to block it. Hardware solutions like *Packeteer* (www.packeteer.com) plug into the network to control the amount of bandwidth available to certain applications. Packeteer has been most popular at colleges and universities, which in general do not want to be accused of censorship or limiting access to resources, but still have to deal with bandwidth concerns amongst thousands of users.

Instant Messaging (IM) Monitoring Products

IM's "fire-and forget" nature made it one of the toughest applications to monitor for a long time. The problem was exacerbated because the employers generally did not control the IM servers or the clients. In 2002 applications were created which successfully monitor IM applications and content, implemented largely to comply with certain US Securities and Exchange Commission (SEC) requirements on keeping logs of all transactions between brokerage houses and their customers. Enterprise IM applications continue today with the likes of AIM Pro, an AOL and WebEx IM program designed for deployment within the organization and subject to its controls. IM is usually not monitored to conserve bandwidth, but it can become a productivity drain similar to e-mail.

Facetime (www.facetime.com) is probably the leader in monitoring organizational IM. The Facetime product can record all IM activity, important in SEC-regulated businesses and government agencies with freedom of information requirements. *Vericept* (www.vericept.com) also has all-purpose monitoring capabilities, but focuses largely on monitoring, blocking and recording IM activity. Organizations looking to block and monitor IM and communicate securely using encryption might turn to *Compliancer Hub* (www.communicatorinc.com).

General Monitoring (Spying at the Client) Tools

There are a series of more powerful, less-specific tools available for almost total user monitoring, classified as general monitoring tools. These tools are installed at the client and can create a record of almost everything a user does with the computer. Records can be written to a network database or even e-mailed to another account.

Spector CNE (www.spectorsoft.com) and Track4Win (www.track4win.com) are sniffing programs designed for enterprise monitoring. There are also many programs targeted at individuals, including *Pearl Echo* (www.pearlsw.com) and *eBlaster* (http://www.spectorsoft.com). Managers are often surprised at the sheer amount of data these programs provide, quite often more than a manager really wants to know. One should carefully consider the implications before implementing a general monitoring tool. Employees may react strongly to such total monitoring.

FUTURE TRENDS

Seeking the Recommended Fit between Goals and Monitoring Solutions

If productivity is a major concern, one might begin with a passive but comprehensive logging tool. Cyberslacking can be easily seen and measured, but it is done unobtrusively. When an employee's productivity falls, as observed traditionally, the technical data showing cyberslacking are available. This can be used for implementing positive disciplinary measures, or for supporting a termination. Periodically and when a situation occurs, employees should be reminded of the organization's policy about personal Internet usage and resulting enforcement actions. This is not to embarrass potential offending parties, but rather to keep the policy salient.

If legal liability is the major concern, minimally intrusive means can also be used for the monitoring and recording of transmitted data. In December 2002, five major Wall Street brokerage houses were fined $1.65 million for not keeping e-mails the required two years. An organization can avoid these types of penalties by simply logging and maintaining records of Internet traffic without any review, except on an as required basis. The RIAA and other entertainment industry groups in early 2003 began warning organizations to actively ensure that their employees were not using company networks to access copyrighted music and video files, lest the companies themselves be held liable. The RIAA has been supplying offenders' IP addresses and access times to companies and universities, idenitfying individuals who may have traded in music and video files. In the end, a company pursuing this strategy would be more concerned with record-keeping than record-reviewing.

An organization pursuing the third goal, bandwidth preservation can likely use the least passive of all monitoring tools--simply observing bandwidth usage spikes where they occur, and witnessing their relationship to organizational goals. A firm that sees bandwidth constantly full with apparently work-related material may want to investigate adding additional bandwidth resources. At the same time, organizations that suddenly block or "throttle" access to popular Internet destinations known for non-work related information will likely solve many problems. Employees are more likely to realize that they need to get to work or identify another means of entertainment than they are to complain and cause a ruckus over no longer being able to access these sites at all or at high speeds.

CONCLUSION

This chapter has detailed several types of control mechnisms. It has listed goals for monitoring, applications to monitor, and means of accomplishing the monitoring. Furthermore, it lists names of actual tools that can be used to accomplish the monitoring. What this chapter so far has not discussed is the viability and consequences of the monitoring.

In a series of studies, it was found (not surprisingly) that Internet monitoring indeed has a significant positive effect on keeping employees on task (Urbaczewski et al., 2002). However, it was also found that monitored employees were more likely to turnover and less likely to participate in other organizational activities. Could this happen in all organizations? Possibly, but the key to remember when establishing monitoring is:

Make the monitoring strategy fit the corporate culture.

Some organizations have a culture where monitoring of most (if not all) activities is simply expected. These industries generally include those handling cash or financially-related transactions (e.g., banks, casinos) or deal with physical and/or national security (CIA, FBI, R&D labs, etc.). In these cases, monitoring fits the culture, and if the organization is already monitoring employees in other ways, it also makes sense to monitor computer systems.

Other organizations have cultures which generally detest monitoring. Some industries, like publishing, academia, and other "civil liberties" organizations do not generally monitor their employees, as their foundations are centered on freedom of speech and the unfettered search for the truth. The introduction of monitoring in these industries will likely result in a culture clash between employees and management (Simmers, 2002).

The question then becomes "how does one then reap the benefits of monitoring without angering employees and creating unwanted stress in the organization?" Valli (2004) argues for provisions that enforce policies while at the same time not destroying the atmosphere of trust and fairness. Communication of the control mechanisms to the employees, with a clear understanding of how it supports corporate goals and principles, is key. Explicit statements of who will be monitored, what will be monitored, and when monitoring will occur should also be communicated to the employees, largely through AUPs.

REFERENCES

AMA. (2005). *Electronic monitoring and surveillance survey*. American Management Association. Retrieved May 18, 2005, from http://www.amanet.org/press/amanews/ems05.htm

Keizer, G. (2006). *Porn-surfing Oregon worker exposes 2,200 Taxpayer Ids*. TechWeb. Retrieved from http://www.techweb.com/wire/security/189401724

Langin, D. J. (2005). Employer liability for employee use of peer-to-peer technology. *Journal of Internet Law, 9*, 17-20.

McCullagh, D. (2001). *Scarfo: Feds plead for secrecy*. Wired Online. Retrieved from http://www.wired.com/news/politics/0,1283,46329,00.html

Reuters (author unknown). New sites top spots for work surfing, as printed in CNN.com. Retrieved from http://www.cnn.com/2002/TECH/internet/09/23/workplace.surfing.reut/index.html

Simmers, C. A. (2002). Aligning Internet usage with business priorities. *Communications of the ACM, 45*(1), 71-74.

Swanson, S. (2002). Employers take a closer look. *Information Week*, *901*, 40-41.

Urbaczewski, A., & Jessup, L. M. (2002). Does electronic monitoring of employee Internet usage work? *Communications of the ACM, 45*(1), 80-83.

Valli, C. (2004). Non-business use of the WWW in three Western Australian organizations. *Internet Research, 14*(5), 353-359.

Zuboff, S. (1988). *In the age of the smart machine: The future of work and power*. New York: Basic Books.

KEY TERMS

Acceptable Use Policy (AUP): A policy created in an organization to outline the permitted and restricted uses of the company's networks and computer systems.

Bandwidth: Colloquially, the amount of network capacity available for a connection.

Blocking: A means of disallowing access to Internet content and services by restricting access at the corporate gateway.

Cyberslacking: The process of using the Internet to waste time during a workday, similar to how an employee might spend time in a colleague's office or on the telephone.

Logging: Creating a record of all employee Internet usage.

OSI Layer 4: The transport layer of a network connection, which provides service differentiation and connection management. From the open systems interconnect model.

OSI Layer 7: The application layer of a network connection. It describes how applications work with the network operating system. From the open systems interconnect model.

Motivation for Using Microcomputers

Donaldo de Souza Dias
Federal University of Rio de Janeiro, Brazil

INTRODUCTION

Information technology implementation is an intervention we make in order to improve the effectiveness and efficiency of a sociotechnical system. Using microcomputers to help individuals perform their jobs and tasks is one of the most important actions we take when implementing this technology effectively. Information systems effectiveness has been extensively studied using, mainly, user satisfaction and quality of information constructs to evaluate users' acceptability (Iivari & Ervasti, 1994; Ives et al., 1983; Neumann & Segev, 1979). However, sometimes, the result of this intervention is not successful and may even generate difficulties related to people participation in the process. This leaves us with a question: What motivates individuals to use microcomputer technology in their daily activities?

Theorists and empirical researchers have been trying to understand the relevant motivators for the implementation and use of computer technology based on the idea that people make an effort if an activity is enjoyable or offers external rewards (Igbaria et al., 1996; Schwartz, 1983). They have been aiming to find out how individuals feel motivated to work with computers, and what motivates them to use computers in their daily activities.

BACKGROUND

Computer and information technology usage is determined by intrinsic as well as extrinsic motivation (Deci, 1975; Igbaria et al., 1996). The main driving forces considered in the literature as motivators for computer and information technology adoption are perceived usefulness, perceived ease of use, and perceived enjoyment (Davis, 1986, 1989; Igbaria et al., 1996). However, it is known that some individuals create personal obstructions to using technology (Pirsig, 1981), particularly, microcomputer technology (Igbaria & Parasuraman, 1989; Martocchio, 1994). They resist microcomputers usage and experience anxiety when they have to deal with them. We present results found in previous studies for relations and comparisons among the motivational forces above (Dias, 1998a, 1998b, 2002; Dias et al., 2002). The results presented here, all statistically significant at $p < 0.05$, were based on constructs measured using the instrument developed in Dias (1998a) and presented in the Appendix.

MAIN MOTIVATIONAL FORCES

Figure 1 shows the results for the relationships among perceived enjoyment, perceived ease of use, and perceived usefulness found in Dias (1998a). The author focused on the motivators perceived usefulness, perceived ease of use, and perceived enjoyment. The aim was to find out how Brazilian operations managers felt about using computer technology in their workplaces, how the perceived usefulness of computers is affected by ease of use and users' enjoyment in working with them, and how to find opportunities to act according to this acquired knowledge, in order to increase the quality of microcomputer technology usage in organizations. In his study, the author emphasized the relationships among these perceived motivators for using microcomputer technology. The impact of the motivators on systems usage or microcomputer adoption was considered to be beyond the scope of his research.

The path analysis model used was based on the natural precedence of intrinsic motivational factors over extrinsic motivational factors, as proposed by the Freudian theory of psychoanalysis (Freud, 1976).

The data for that study were gathered using a questionnaire administered personally to 79 Executive MBA students at a Brazilian university. Respondents held managerial positions in 55 companies, ranging from small firms to large corporations, located in Rio de Janeiro. The average age of respondents was 36, and they had an average of 11 years working experience. All of the participants were college graduates. Managers said they used microcomputer technology mainly because they perceived it as a useful tool to increase the quality of their work, to accomplish tasks more quickly, and to increase the productivity of their jobs.

Figure 2 shows results of a model in which computer anxiety and enjoyment were considered as antecedent variables to ease of use and usefulness (Dias, 1998b). We found that managers who were more anxious about computer technology tended to find it more difficult to use. On the other hand, enjoyment had a positive direct effect on ease of use and usefulness, as stated before.

In a study made at a private university located in Rio de Janeiro, with data gathered from 336 undergraduate computer information systems students, Dias et al. (2002) tested the influence of some antecedent variables on enjoyment, ease of use, and usefulness. They found that (a) the fact that a student worked part-time in an area related to information

Figure 1. Relationships among enjoyment, ease of use, and usefulness

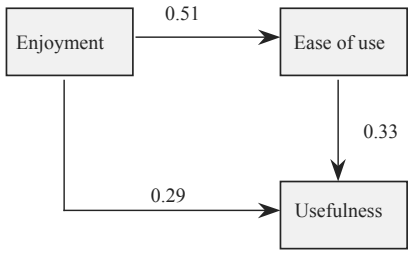


Figure 2. Anxiety and enjoyment as antecedent variables

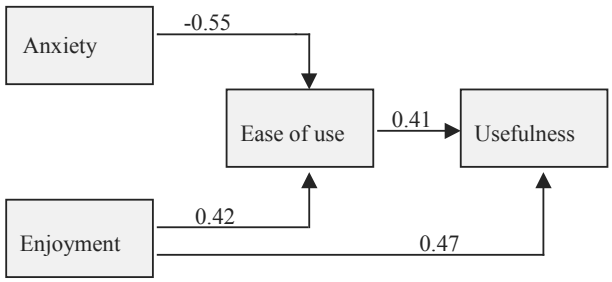
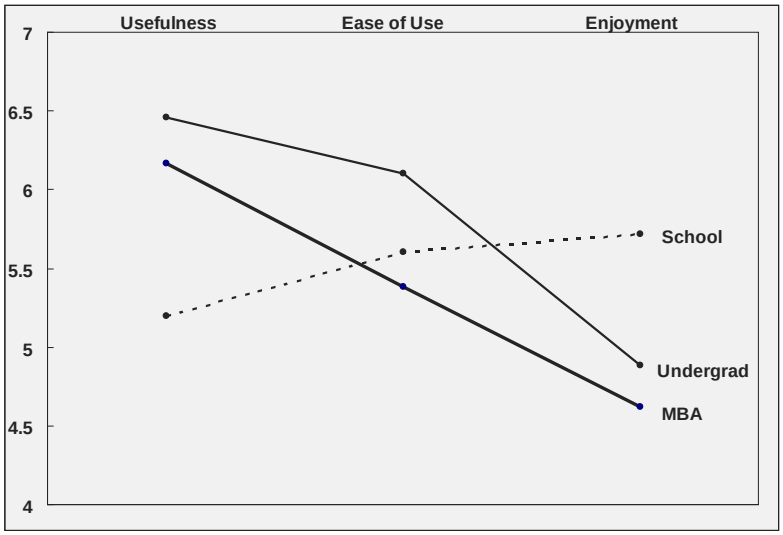


Figure 3. Motivation level for graduate, undergraduate, and elementary school students



technology positively influenced his or her perception of how easy it was to use microcomputers; (b) enjoyment with microcomputers seemed to decrease as students attained seniority in the university; and (c) older students perceived greater usefulness for microcomputers.

Level of education and age have shown influence on microcomputer attitudes (Igbaria & Parasuraman, 1989). Dias (2002) did a study on the motivation for using microcomputers among different classes of users. He aimed at finding out how graduate, undergraduate, and elementary school students, which represent very specific strata of educational level and age, would differ on the motivational factors examined here. The data for his study were gathered as follows:

- Fifty-three Executive MBA students of a leading Brazilian public university: The average age of respondents was 36, and they had an average of 11 years working experience, all participants were managers and had a college degree.
- Forty-six students aiming for degrees in Business Administration at a private university located in Rio de Janeiro: The average age of respondents was 22.
- Thirty-nine elementary schools students enrolled in the fourth to eighth grades of private (82%) and public schools located in the city of Rio de Janeiro: The students used microcomputers regularly at school, at home, or at relatives' homes.

Factor analysis confirmed that the statements for usefulness, ease of use, and enjoyment constituted three distinct perception constructs for the three classes of users. It confirmed the existence of three factors that accounted for 63.5% of the total variance in the 138 interviewees.

Figure 3 shows the motivational profile for graduate, undergraduate, and elementary school students on the usage of microcomputers.

Table 1. Actions for successful implementation of microcomputers

Action	Motivational Forces
Develop friendly systems	Ease of use/enjoyment
Encourage user participation in systems development	Usefulness/ease of use
Implement intensive user training	Ease of use
Sell the system to users	Usefulness
Facilitate access to microcomputers	Ease of use
Respond to user needs	Usefulness
Use up-to-date technology	Usefulness/enjoyment
Align business and information technology	Usefulness

Dias (2002) found that there was a statistically significant difference in the perceived enjoyment, perceived ease of use, and perceived usefulness of using microcomputers among MBA, undergraduate, and elementary school students. MBA and undergraduate students were most motivated by microcomputer technology usefulness. Elementary schools students mostly enjoyed the play aspect of microcomputers.

We did a study with managers in order to better understand these motivational forces and to generate recommendations for increasing the motivation for using microcomputers. Thirty-six managers were split into six working groups and asked to (1) discuss their experience in using microcomputers in their organizations (taking into account personal resistances for using computers in the workplace), situations in which this technology is used compulsively, motivators they thought were important for using computers for task execution, and organizational culture; and (2) propose actions they thought should be taken in the implementation of microcomputers in their workplaces that would lead to a more effective usage of this technology.

The main actions suggested and the corresponding motivational forces are presented in Table 1. We found that in order to attract adult users, information systems should be friendlier, use up-to-date interfaces, and be developed with user participation. In addition, it is important to offer intensive user training, to market the new systems to users heavily, and to fully meet the users' needs in the specification of the systems.

Although managers showed less unanimity on the motivating power of enjoying themselves while using microcomputers, this motivation factor clearly showed its significance. Enjoyment seems to serve more as self-motivation, while usefulness seems to be linked to a feeling of duty—something that the organization expects managers to do or to attain.

FUTURE TRENDS

According to the World Summit on the Information Society (WSIS, 2003):

The global Information Society is evolving at breakneck speed. The accelerating convergence between telecommunications, broadcasting multimedia and information and communication technologies is driving new products and services, as well as ways of conducting business and commerce. At the same time, commercial, social and professional opportunities are exploding as new markets open to competition and foreign investment and participation.
The modern world is undergoing a fundamental transformation as the Industrial Society that marked the 20th century rapidly gives way to the Information Society of the 21st century. This dynamic process promises a fundamental change in all aspects of our lives, including knowledge dissemination, social interaction, economic and business practices, political engagement, media, education, health, leisure and entertainment. We are indeed in the midst of a revolution, perhaps the greatest that humanity has ever experienced.

Several new concepts lay the foundation for prospering in the next form of information and communication technology. The future microcomputer usage will be based upon ubiquity, universality, uniqueness, and unison (Watson et al., 2002). The keys to managing network-driven firms will be based on this notion of ubiquitous networks. In addition, we have to consider the new generation of students entering business schools around the world. These students are quite different from the students of the last decades, and we have to consider the technological, social, and cultural changes that are developing. The implications for the next-

generation organization and information systems based on the Internet and mobile microcomputer technology have to be considered, and the motivational forces discussed here will probably have to be readdressed.

CONCLUSION

The research described here on motivation for using microcomputers offers several contributions for theory and practice. It confirmed that there are three positively interrelated motivators for computer technology usage: perceived usefulness, perceived ease of use, and perceived enjoyment. It also found that perceiving microcomputers as easy to use is negatively related to people's anxiety toward using them.

The research also compared the perceived reasons for using microcomputers among different classes of users and showed that there are significant differences in the motivation for using them among graduate, undergraduate, and elementary school students. Elementary school students have greater enjoyment in using microcomputers than MBA and undergraduate students in Business Administration. MBA and undergraduate students in Business Administration perceive greater usefulness in the usage of microcomputers than elementary school students, and they said that they used microcomputers mainly because this increased the quality of their work and allowed them to accomplish tasks more easily and quickly. Undergraduate students in Business Administration think it is easier to use microcomputers than do MBA and elementary school students. MBA students think it is most difficult to use microcomputers in their daily tasks.

We should emphasize that the motivational forces studied here could not apply equally to different countries, cultures, or organizational environments. A study made by Straub et al. (1997) compared the technology acceptance model (TAM) (Davis, 1986) across three different countries: Japan, Switzerland, and the United States. The study was conducted by administering the same instrument to employees of three different airlines, all of whom had access to the same computer technology innovation—e-mail. They used perceived ease of use and perceived usefulness as independent variables. The results indicated that the model holds both for the United States and Switzerland, but not for Japan, suggesting that the model may not predict motivation for technology use across all cultures. Harris and Davidson (1999) examined microcomputer anxiety and involvement of groups of microcomputer-using students in six developing countries: China, Hong Kong, Malaysia, New Zealand, Tanzania, and Thailand. Differences in computer anxiety were found to exist between some of the groups, which were probably attributable to demographic factors. Differences were found to exist between the microcomputer involvements of some of the groups, which could be attributed to cultural factors.

APPENDIX

Please indicate your agreement or disagreement with the following statements, related to the usage of computers in the workplace, using a 7-point scale ranging from *fully disagree* (1) to *fully agree* (7). Please refer to Appendix 1 on the previous page.

REFERENCES

Davis, F. (1986). *A technology acceptance model for empirically testing new end user information systems: Theory and results.* Unpublished Doctoral Dissertation. Boston, MA: MIT.

Appendix 1

	Fully disagree *Fully agree*
1. I do not see time go by when I am using a computer.	1 2 3 4 5 6 7
2. Using computers enables me to accomplish my tasks more quickly.	1 2 3 4 5 6 7
3. I feel motivated to perform activities using computers.	1 2 3 4 5 6 7
4. I find it is easy to use a computer to do my work.	1 2 3 4 5 6 7
5. Using computers is fun.	1 2 3 4 5 6 7
6. Using computers improves my job productivity.	1 2 3 4 5 6 7
7. Using computers makes it easier to perform my tasks.	1 2 3 4 5 6 7
8. Using computers is exciting.	1 2 3 4 5 6 7
9. Using computers increases the quality of my work.	1 2 3 4 5 6 7
10. I feel anxiety when I have to perform a task using microcomputers.	1 2 3 4 5 6 7
11. I think it is easy to use computers.	1 2 3 4 5 6 7
12. Using computers is pleasant.	1 2 3 4 5 6 7
13. I find computers useful for my job.	1 2 3 4 5 6 7
14. I think we should use computers as much as possible.	1 2 3 4 5 6 7

Davis, F. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340.

Deci, E. (1975). *Intrinsic motivation*. New York: Plenum Press.

Dias, D. (1998a). Managers' motivation for using information technology. *Industrial Management and Data Systems, 98*(7), 338–342.

Dias, D. (1998b). Intrinsic and extrinsic motivators for microcomputer usage. In M. Khosrow-Pour (Ed.), *Effective utilization and management of emerging information technologies* (pp. 664–667). Hershey, PA: Idea Group Publishing.

Dias, D. (2002). Motivation for using information technology. In E. Szewazak & C. Snodgrass (Eds.), *Human factors in information* systems (pp. 55–60). Hershey, PA: IRM Press.

Dias, D., Mariano, S., & Vasques, R. (2002). Antecedents of Internet use among Brazilian information systems students. *Issues in Information Systems, 3*, 144–150.

Freud, S. (1976). *Obras Completas*, Rio de Janeiro: Imago, 22, 75–102.

Harris, R., & Davidson, R. (1999). Anxiety and involvement: Cultural dimensions of attitudes toward computers in developing societies. *Journal of Global Information Management, 7*(1), 26–38.

Igbaria, M., & Parasuraman, S. (1989). A path analytic study of individual characteristics, computer anxiety and attitudes toward microcomputers. *Journal of Management, 15*(3), 373–388.

Igbaria, M., Parasuraman, S., & Baroudi, J. (1996). A motivational model of microcomputer usage. *Journal of Management Information Systems, 13*(1), 127–143.

Iivari, J., & Ervasti, I. (1994). User information satisfaction: IS implementability and effectiveness. *Information and Management, 27*(4), 205–220.

Ives, B., Olson, M., & Baroudi, J. (1983). The measurement of user information satisfaction. *Communications of ACM, 26*(10), 785–793.

Martocchio, J. (1994). Effects of conceptions of ability on anxiety, self-efficacy, and learning in training. *Journal of Applied Psychology, 79*(6), 819–825.

Neumann, S., & Segev, E. (1979). A case study of user evaluation of information characteristics for systems improvement. *Information and Management, 2*(6), 271–278.

Pirsig, R. (1981). *Zen and the art of motorcycle maintenance*. New York: Bantam Books.

Schwartz, H. (1983). A theory of deontic work motivation. *The Journal of Applied Behavioral Science, 19*(2), 204–214.

Straub, D., Keil, M., & Brenner, W. (1997). Testing the technology acceptance model across cultures: A three country study. *Information and Management, 33*, 1–11.

Watson, R., Pitt, L., Berthon, P., & Zinkhan, G. (2002). U-Commerce: Expanding the universe of marketing. *Journal of the Academy of Marketing Science, 30*(4), 333–347.

WSIS. (2003). World summit on the information society: Newsroom fact sheets: The challenge, Geneva, Switzerland. Retrieved May 6, 2004, from *http://www.itu.int/wsis/newsroom/fact/whynow.html*

KEY TERMS

Computer Anxiety: Degree to which an individual is nervous in his or her interaction with computers; the uneasiness some people feel when they have to use a microcomputer. Anxiety results from a danger or a danger threat. As a feeling, it has a clearly unpleasant character.

Computer Self-efficacy: A judgment of one's capability to use a computer. It incorporates judgments of an individual on his or her skills to perform tasks using a microcomputer.

Extrinsic Motivation: Motivation that derives from what you obtain from engaging in an activity. An example of extrinsic motivation for using microcomputers is using it because you think it is useful for your job.

Intrinsic Motivation: Motivation that derives from the activity itself. An example of intrinsic motivation for using microcomputers is using it because you enjoy it.

Microcomputers Ease of Use: User perception on how simple and easy it is to understand and use microcomputers; degree to which an individual believes that using a particular computer system would be free of physical or mental effort.

Microcomputers Enjoyment: The extent to which the activity of using microcomputers is perceived as being enjoyable in its own right, apart from any performance consequences. It encompasses the feelings of joy, elation, or pleasure associated by an individual to a particular act.

Microcomputers Usefulness: The degree to which an individual believes that using a particular computer system would enhance his or her job performance.

Sociotechnical System: One that focuses on the interaction between the technical and social subsystems that exists

in any work situation. As there is always a social system operating along with any technological system, we need to jointly design these two interrelated systems to get the best results from any work system.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2030-2035, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: IT Education

Motivational Matrix for Educational Games

Athanasis Karoulis
Aristotle University of Thessaloniki, Greece

INTRODUCTION

The study of the motivational factor in educational games (aka EduGames) has been limited up to now. A former study (Karoulis, 2004) discussed some aspects and proposed the adherence to the ARCS model of motivation proposed by Keller (Keller, 1983; Keller, 1998), which describes the motivation of any educational piece according to four factors: attention, relevance, confidence, and satisfaction.

Present study attempts to summarize the attributes of any EduGame, as they are encountered in the relative literature (including representations) and to match every one of those attributes to one (or more) of the ARCS-factors of motivation.

The benefit of this approach is a better understanding of the motivational nature of every attribute of every EduGame and an obvious extension is the evolvement of a set of design guidelines for designers of EduGames and educational software in general.

BACKGROUND

Keller's ARCS Model for Motivation

Motivation is the most overlooked aspect of instructional strategy, and perhaps the most critical element needed for employee-learners. Even the most elegantly designed training program will fail if the students are not motivated to learn. Without a desire to learn on the part of the student, retention is unlikely. Many students in a corporate setting who are forced to complete training programs are motivated only to "pass the test." Designers must strive to create a deeper motivation in learners for them to learn new skills and transfer those skills back into the work environment.

As a first step, instructional designers should not assume they understand the target audience's motivation. To analyze needs, the designer should ask prospective learners questions such as:

- What would the value be to you from this type of program?
- What do you hope to get out of this program?
- What are your interests in this topic?
- What are you most pressing problems?

The answers to these types of questions are likely to provide insight into learner motivation, as well as desirable behavioral outcomes.

Keller synthesized existing research on psychological motivation and created the ARCS model (Keller & Kopp, 1987). ARCS stands for attention, relevance, confidence, and satisfaction. This model is not intended to stand apart as a separate system for instructional design, but can be incorporated within Gagne's events of instruction (Gagne, 1985, 1987; Gagne, Briggs, & Wager, 1992).

- **Attention:** The first and single most important aspect of the ARCS model is gaining and keeping the learner's attention, which coincides with the first step in Gagne's model. Keller's strategies for attention include sensory stimuli (as discussed previously), inquiry arousal (thought provoking questions), and variability (variance in exercises and use of media).
- **Relevance:** Attention and motivation will not be maintained, however, unless the learner believes the training is relevant. Put simply, the training program should answer the critical question, "What's in it for me?" Benefits should be clearly stated. For a sales training program, the benefit might be to help representatives increase their sales and personal commissions. For a safety-training program, the benefit might be to reduce the number of workers getting hurt. For a software-training program, the benefit to users could be to make them more productive or reduce their frustration with an application. A healthcare program might have the benefit that it can teach doctors how to treat certain patients.
- **Confidence:** The confidence aspect of the ARCS model is required so that students feel that they should put a good faith effort into the program. If they think they are incapable of achieving the objectives or that it will take too much time or effort, their motivation will decrease. In technology-based training programs, students should be given estimates of the time required to complete lessons or a measure of their progress through the program.
- **Satisfaction:** Finally, learners must obtain some type of satisfaction or reward from the learning experience. This can be in the form of entertainment or a sense of achievement. A self-assessment game, for example, might end with an animation sequence acknowledging

the player's high score. A passing grade on a post-test might be rewarded with a completion certificate. Other forms of external rewards would include praise from a supervisor, a raise, or a promotion. Ultimately, though, the best way for learners to achieve satisfaction is for them to find their new skills immediately useful and beneficial on their job.

A Classification of Multimedia Representations

The locus of this work is to promote an understanding of the correlation between the four factors of the ARCS model and the characteristics of the employed representations. A well accepted taxonomy is the one by de Jong et al. (2004). However, in order to promote understanding of their employment in EduGames, one has to bear in mind a "virtual" game and rely to one's situated experience. To describe the employment of representations, examples of possible applications are used subsequently. Under this point of view, following categories are of interest in EduGames.

- **Multiple representations:** Text, animation, static image representations, sound, and video constitute the majority of the representations employed in EduGames. Usually, one cognitive aspect is presented by means of more than one representation (e.g., video with sound).
- **Code and modality:** The navigational elements are often icons, still or animated. They (the representing world) depict a navigational structure (the represented world), which is usual in educational environments: next, previous, home, exit, repeat, and help. The rest of the representations occur where interaction with the user is possible. A narration prompts the pupil to act and provides help on it. In such a case, we are talking in both cases of *depictive* and *non-equivalent* representations, which are however *multimodal*, as they employ aural, visual, and tactile modes to interact with the user (the user is often asked to type something). Although, from a usability perspective it could be debatable in how far the used navigation icons are intuitive to a novice user of younger age, the application of such a software often shows that children can easily overcome such burdens with little or not at all help. The exploratory nature of children permits them to explore the interface and discover their capabilities. The "prevent errors" usability factor is important here, in order to hinder fatal errors an exploring user could cause.
- Animation seems to provide potency for *dynamic* and *kinesthetic* (manipulable) types of representations. Often, only *concrete*, *pattern imagery,* and *symbolic elements* are represented. Animation for feedback is considered here to belong to the pattern category, as it only informs the user on the correctness or not of their action. It is obvious that we are dealing here with *depictive feedback* (if it is correct or not) and not with *constructive feedback* (in what direction one should seek for the correct solution).
- Affordances: Rarely do the visual representations provide concrete affordances, in helping to visualize the information. In this sense, they help to structure the cognitive activity and provide patterns for experimentation. However, in most cases animations and sound cues are used as feedback or as a helping facility (explaining narration).
- Concerning the underlying theoretical support, the theories of *dual coding* (Paivio, 1990) and *cognitive load* (Chandler & Sweller, 1991; Yeung, Jin, & Sweller, 1998) seem to be implicitly employed in the design of the majority of the systems. Dual coding theory is de facto implemented in any multimedia environment, and its ultimate purpose is to reduce cognitive load (Mayer, 2003), so it can be argued that the use of multimedia animations intends to benefit from these theories. In contrast, *multimedia design theories* seem to be explicitly employed in the design and construction of many interfaces.
- In regards to the *cognitive modeling* support, it is usually not apparent in the designers' intentions, although most interfaces do not provide any problems on it. Children usually can easily work in the interface, without any hindering. One remark must be stated here, concerning the *redundancy* principle and the claim "avoid presenting verbal information in both textual and narrative form especially when graphics are presented at the same time" (de Jong et al., 2004), and a claim stated by Juul (2000) that "it (the game) must not contain narration; everything must happen in the *now* of the playing."
- At this point, the provided *degrees of freedom* must be discussed. Many gaming environments, which simulate "the school" can not be characterized as a constructivist ones, as most of the exercises employed are already known to the pupils from school and must be performed in a pre-defined way. So, it can be argued that the used representations in such environments significantly reduce the degrees of freedom, while they provide only limited affordances.

These characteristics of the multimedia representations usually employed in EduGames are only a subset of the set of attributes to be considered while designing and constructing an EduGame.

MATCHING OF MOTIVATIONAL AND REPRESENTATIONAL CHARACTERISTICS

Accordingly, a combination of most encountered attributes in EduGames is attempted with the factors of the ARCS model. To facilitate this matching, this study proposes the employment of a matrix as follows.

Table 1 presents the four factors of the ARCS model on the top, analyzed in their dominant parameters. On the left-hand column are presented the main attributes of the EduGames, as described in the relative literature (the representations are included).

RESEARCH QUESTIONS

The initial and central question, according to which every cell of the table has to be filled, is: "Are the factors of the ARCS model supported by every attribute and how exactly?"

So, to fill out the table one must consider every attribute on the left-hand column and try to investigate its correlation (if any) to every factor of the ARCS modes, represented by their most dominant parameters. For every matching, a separate investigation must be set up, may it be a case study, a literature review or any other acknowledged approach.

These attributes are collected through an extensive literature overview on the characteristics of EduGames (e.g., Azar, 1998; Malone, 1980a, 1980b)

Representations in EduGames have also been considered as attributes and are listed on the table. A valuable source here are the works (among others) of Ainsworth and Van-Labeke (2004), van der Meij, J., and de Jong, T. (2004), and de Jong et al. (1998).

Another criterion to choose these attributes is their applicability. This is particularly difficult because of the given variety of EduGames. Games encompass many styles and subjects. For example, games may be competitive or cooperative, be played by individuals or groups, and touch on numerous themes, such as adventure, education, social interactions, science fiction, violence, and sexual circumstances. Simulations sometimes are considered games as well. Leemkuil, de Jong, and Ootes (2000) argue that games as learning environments are closely related to simulations, microworlds, adventures, and case studies. The definitions of these environments partially overlap. For instance, the distinction between simulation and games is often blurred, and many recent articles in this area refer to a single "simulation game" entity.

On the other hand, EduGames must underly some guidelines to be characterized as such. Harlow (2004) classifies them in two categories, process and reward. Process comprises the actual playing of the game such as the interface, levels, immersion, content, and interaction of game mechanics, while reward is either the internal game benefit or external feeling of satisfaction or success the player gains from the process. So, some attributes, such as "challenges," "choices," and "rewards" emerge from this approach.

Further on, a great deal of research focuses on computer literacy and basic programming skills. Playing computer games is a popular recreational activity for young people. Not surprisingly, many of these enthusiasts dream that one day they will develop computer games themselves. So why not use game design as a vehicle to teach youngsters computer science? Developing computer games involves many aspects of computing, including computer graphics, artificial intelligence, human-computer interaction, security, distributed programming, simulation, and software engineering. Game development also brings into play aspects of the liberal arts, the social sciences, and psychology. Creating a state-of-the-art commercial computer game is an incredibly difficult task that typically requires a multimillion-dollar budget and a development team that includes 40 or more people. But simpler alternatives—ones within the reach of students and hobbyists—exist. Budding game developers can have fun creating variations on Pac-Man, Space Invaders, or simple platform games (Overmars, 2004).

Under this point of view, attributes such as "assignments in layers" or "present abstract material in real world context" become also critical and are included in the list.

Another aspect is the social dimension of gaming. Kiefaber (2004) argues that gaming is inherently social and playing games has been closely linked with building relationships and social hierarchies throughout history. Not only games are social activities, but also many times the game itself is secondary to the social experience. Video games are now a permanent fixture to our culture, redefining the process by which children mature and develop. Over the past few years however, a great shift towards Internet-based games has been observed and recorded. The lure of multi-player gaming has been interacting with real people, not artificial intelligence. Live opponents or allies make gaming much more unpredictable and much more enjoyable. Nowadays, not only are potential game players connected to the Internet in large and rapidly growing numbers, but their composition is much more diverse and many new gaming possibilities have been created that were not available with non-Internet games. To conclude, attributes such as "cooperation," "competition," "fantasy," as well as almost all representational characteristics listed in the previous table, become also socially important, thus shifting their correct matching to the motivational parameters to an important task.

Table 1. Motivational matrix of educational representations

	Attention		Relevance		Confidence		Satisfaction	
	Curiosity (stimuli variability)	Interest (Inquiry arousal)	Perception of personal needs ac-complish-ment	Relation to a highly desired goal	Expectancy of final sucess	Sucess under control	Extrinsic rewards	Intrinsic com-patibility to anti-citations
Easy to understand								
Options for creativity								
A challenge (goal)								
Assignments (challenges) in layers								
Fantasy (extrinsic & intrinsic)								
Curiosity (sensory & cognitive)								
Multiple representations								
Code & modality								
Feedback (instant & summative)								
Affordances								
Dimensions								
Cognitive modeling								
Degrees of freedom								
Rewards (into the learning task)								
Abstract mate-rial in real world context								
Choices offering								
Competition & cooperation								
Decision making & problem solving								
Learning vs. Game objectives								
Graphics								

FUTURE TRENDS

It is obvious that further research has to elucidate the correct matching (if possible at all) and the matching strategy for every EduGame attribute to every (or more) ARCS factors. The cell will hence describe the details of every particular approach.

Research for every case can be in the form of literature review, case studies, experiments, or any other acknowledged approach.

However, this implies a huge amount of research, which is not feasible at this time. On the other hand, some research has already been done and some answers are already given. For example, in regards to the "graphics" attribute, literature gives an extensive focus on it, as it is stated that "they can capture the students' interest" almost in consensus (e.g., Walker, 2003)

So, a more realistic approach here would be to focus on some of the attributes, which are not yet studied by other researchers and set up some case studies or experiments to clarify their correlation to the ARCS parameters.

In more detail, some EduGames attributes must be chosen and implemented in an EduGame environment. Accordingly, a study, which has clear hypotheses, must investigate whether these attributes have any influence on the augmentation of any of the ARCS parameters, which in its turn would lead to the augmentation of any of the four factors of the ARCS model.

This could be a clear indication that the correct manipulation of the particular attribute can lead to an augmentation of one (or more) motivational factor, leading thus to an enhanced motivation of the EduGame, which is the Holy Grail for every educational designer.

Who could benefit from this approach? Educational designers, teachers, and any person involved in educational software could benefit from this research. Results are obviously of a broader applicability than only in EduGames, since the motivational factor is usually hardly understood and implemented in educational pieces. So, a reference to some "motivational guidelines" would be of great interest for designers as well as for evaluators of any instructional approach.

CONCLUSION

The matching of the encountered in the literature attributes as regards educational games with the four factors of the ARCS model proposed by Keller is a daunting task, which however could enlighten many hidden up to now relations between correct instructional design and motivation. Since it is commonly agreed that motivation is the *sine qua non* of successful learning (Prensky, 2003), it would be a valuable result to explicitly know relations of educational attributes that could enhance any motivational factor.

REFERENCES

Ainsworth, S., & VanLabeke, N. (2004). Multiple forms of dynamic representation. *Learning and Instruction*, (in press).

Azar, B. (1998). Research-based games enhance children's learning. *The APA Monitor*, *29*(8). Retrieved December 8, 2004, from http://www.apa.org/monitor/aug98/games.html

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715-730.

De Jong, T., Ainsworth, S., Dobson, M., van der Hulst, A., Levonen, J., Reimann, P., et al. (1998). Acquiring knowledge in science and mathematics: The use of multiple representations in technology based learning environments. In M. van Someren, P. Reimann, H. Boshuizen, & T. de Jong (Eds.) *Learning with multiple representations* (pp. 9-41). Oxford: Elsevier Science.

Gagne, R. (1987). *Instructional technology foundations*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Gagne, R. (1985). *The conditions of learning* (4th ed.). New York: Holt, Rinehart & Winston.

Gagne, R., Briggs, L., & Wager, W. (1992). *Principles of instructional design.* New York, Holt, Rinehart, and Winston. 1st edition in 1988, 4th edition in 1992.

Harlow, D. (2004). *Games as an educational tool*. Retrieved from http://www.gamedev.net/reference/articles/article2082.asp

Karoulis, A. (2004). Motivation and representation in educational games. In Kaleidoscope NoE JEIRP, *Interaction between learner's internal and external representations in multimedia environments*, State-of-the-art report, pp 296-312. Available from: http://aiges.csd.auth.gr/academica/pages_en/Science/sci_en_dislearn.html

Keller, J. M. (1998). Using the ARCS process in CBI and distance education. In M. Theall (Ed.), *Motivation in teaching and learning: New directions for teaching and learning*. San Francisco: Jossey-Bass.

Keller, J. M. (1983). Motivational design of instruction. In C. M. Reigeluth (Ed.), *Instructional design theories and models: An overview of their current status* (pp. 383-434). New York: Lawrence Erlbaum.

Keller, J. M., & Kopp, T. W. (1987). Application of the ARCS model to motivational design. In C. M. Reigeluth (Ed.) *Instructional theories in action: Lessons illustrating selected theories* (pp. 289-320). New York: Lawrence Erlbaum, Publishers.

Kiefaber, M. (2004). *Implications of online gaming*. Retrieved from http://www.units.muohio.edu/psybersite/syberspace/onlinegames

Leemkuil, H., de Jong, T., & Ootes, S. (2000). Review of educational use of games and simulations. *Knowledge Management Interactive Training System,* University of Twente. KITS concortium.

Malone, T. W. (1980a). *What makes things fun to learn? A study of intrinsically motivating computer games.* Technical report, Xerox Palo Alto Research Center, Palo Alto, CA.

Malone, T. W. (1980b). What makes things fun to learn? Heuristics for designing instructional computer games. In *Proceedings of the 3rd ACM SIGSMALL symposium and the first SIGPC symposium on Small systems* (pp. 162-169). Palo Alto, Cal.

Overmars, M. (2004). Teaching computer science through game design. *IEEE Computer, 37*(4), 81-83.

Prensky, M. (2003). Digital game-based learning. *Computers in Entertainment*, *1*(1).

Van der Meij, J., & de Jong, T. (2004). Examples of using multiple representations. In Kaleidoscope NoE JEIRP, *Interaction between learner's internal and external representations in multimedia environments*, State-of-the-art report, pp 66-80.

Walker, H. M. (2003). Do computer games have a role in the computing classroom? *Inroads-The SIGCSE Bulletin*, *35*(4), 18-20.

KEY TERMS

ARCS Model: A model describing motivation by means of the four factors of attention, relevance confidence, and satisfaction.

Attributes: Characteristics of an entity, clearly contributing to its utility.

Classification: A taxonomy according to dominant characteristics of some entities.

Educational Games: Computer-based electronic games with high educational value. They usually adhere to the constructivist theory of learning.

Motivational Matrix: A table combining the motivational characteristics with the attributes of other entities.

Multimedia Representations: Chunk of educational information depicting selected aspects of the depicted world by means of multimedia modalities.

Motivations for Internet Use

Thomas F. Stafford
University of Memphis, USA

INTRODUCTION

In light of the importance the Internet has as a channel of commerce, it is important to understand consumer motivations for Internet use (Eighmey & McCord, 1998; Lohse & Spiller, 1998; Schonberg, Cofino, Hoch, Podlaseck, & Spraragen, 2000). In the absence of motivations for Internet use, there can be no motivations for e-commerce use, so Internet use motivations are an important antecedent to e-commerce activities (Stafford, 2003b). The Internet is a telecommunications medium, but it is also far more than a computer-mediated communication channel. In its evolution, the Internet evolved from a basic telecommunications network, to a consumer communications and entertainment medium, to a converged channel of commercial and telecommunications media that combine the utilities of familiar entertainment and communications media such as telephones, radio, and television, along with emerging computer network functionalities.

While remaining at its core a network for the distribution of information and telecommunications services, it has evolved into a combined channel for the delivery of other, richer media—become a medium of conveyance for many separate media delivered simultaneously, or a *meta-medium* (Stafford, Stafford, & Shaw, 2002). In the past, understanding Internet motivations strictly related to computer use was sufficient to characterize Internet user motivations, but in the converged meta-medium of the modern day, we should consider a wider range of potential uses and motivating gratifications arising from use of this complex and converged medium.

Media uses and gratifications (U&G) has been a useful theoretical platform for understanding Internet use in this emerging age of media convergence. This perspective focuses on the process of using the Internet medium, and the gratifications related to the content provided by the network. More recently, Internet U&G research has demonstrated additional motivations for Internet use that expand beyond the traditional usage process and media content motivations found in U&G studies of conventional media. These motivations span usage process and content to include considerations of social motivations for network usage, which is a gratification that traditional media have not generally been able to supply to users (Stafford et al., 2002). These new and emerging media usage gratifications for the Internet are important for site and service operators to understand, if they wish to successfully motivate customer use of and loyalty to their resource. These new motivations are potential differentiators between operators *within* the Internet medium as well as *between* the Internet and conventional promotional media.

BACKGROUND: MEDIA USAGE AND GRATIFICATIONS PERSPECTIVES

It has long been known that media usage is not a random or undirected activity; like all rational human behavior, there are discernable motivations for media use (Katz, 1959). Early U&G research in the radio and television era determined that audiences are not passive consumers of media (Katz, Blumler, & Gurevitch, 1974; Rubin, 1981), and this "active use" tenet of U&G research means that media researchers can find theoretically compelling approaches to simulate more involved media use, which is surely a beneficial outcome to the media and their commercial sponsors.

Media choice is motivated by individual uses and the individual goals related to those individual uses, which has come to be characterized as media usage "gratifications" (Lin, 1977). In this sense, media are like any other product or service that might be marketed to customers. In markets, even media markets, consumers make choices about what to use in an active and selective manner (Levy & Windahl, 1984). This means that media operators cannot presume a captive audience and must strive to understand their audience usage motivations in order to provide a compelling and attractive media experience. This perspective has come to be called the "niche theory" of media use, wherein consumer time for media use is a finite and limited resource that must be actively competed for by available media in product-market fashion, on the presumption that time spent by consumers with one media reduces available time to be spent with other competing media (e.g., Dimmick, Chen, & Li, 2004).

In the competitive scenario of media niche theory, understanding user motivations becomes all the more critical, as media must compete with each other for the available audience, and good understanding of user motivations provides more able competition for scarce audience resources. U&G theory has been quite useful in providing clear and effective profiles of media user motivations in traditional media in the past, as well as in the emerging Internet meta-medium (Eighmey, 1997; McDonald, 1997; Newhagen & Rafaeli, 1996). In view of the multimedia aspect of the modern

Internet, it is considered that opportunities for user gratifications are greatly increased over traditional media (Dimmick et al., 2004); hence a clear understanding of Internet U&G will be useful for effective "niche-theory" marketing of the Internet medium to users.

Classic applications of U&G theory in the traditional media have consistently identified only two key areas of motivation for media use: media content uses and motivations (*gratifications*, in U&G parlance), and media usage process gratifications. Content gratifications concern the *messages* carried by the medium (which could be informative or entertaining), and process gratifications concern *actual use* of the medium, itself (Cutler & Danowski, 1980). The modern analogies would be the Web surfer, who is clearly motivated by the process of using the Internet, versus the highly focused online researcher, who is engaged in searches for very specific message content to support information needs (Stafford & Stafford, 2000).

U&G UP-TO-DATE

In research that has spanned the course of the past decade, an emerging stream of literature documenting Internet-specific media uses and gratifications is now reaching publication. Early research on U&G and the Internet was limited by the practice of adapting measurement scales for usage gratifications from television U&G studies (cf., Eighmey, 1997; Eighmey & McCord, 1998; Rafaeli, 1988). This was a useful transition stage approach for developing an initial understanding of the motivations of early Internet users, but the medium and its users have become far more sophisticated and complex in the intervening years.

Emerging Internet U&G research, which abandons traditional television-based measures and develops scales specifically tailored to the Internet experience, is based on the premise that Internet usage gratifications are different from the motivations that drive the use of other media (e.g., Stafford & Stafford, 1998). This article documents some of the more prominent findings in the process of developing these new Internet-specific U&G dimensions, as well as emerging "second-wave" Internet user motivation and behavioral studies.

The Initial Factors of Internet-Specific U&G

The initial approach to developing Internet-specific U&G profiles began with online qualitative research that sought to investigate the dimensionality of uses and gratifications; in classic measure development manner, Stafford and Stafford (2001a) leveraged their qualitative investigation of Web site users into an online user survey for purposes of identifying new Internet-specific gratifications through multivariate analysis. Exploratory factor analysis identified five Internet-specific U&G factors: searching, cognitive, new and unique, social, and entertainment.

The searching factor was not unexpected and is certainly intuitive to experienced users of the Internet. The cognitive factor was characterized by gratifications related to learning: education, information, learning, and research. The factor "new and unique" was representative of the still-new feeling of the medium at the turn of the century, characterized by user perceptions of the medium that included qualities such as "ideas," "interesting," "new," "progressive," and "relaxing." The entertainment factor (entertainment, fun, and games) was a media content gratification related to having fun with Internet site content. The social factor (chatting, friends, interaction, newsgroups, and people) was distinct in comparison to both traditional U&G dimensions of process and content gratifications, as well as early Internet applications of traditional U&G dimensions in exploratory research, since previous U&G research had never identified a social motivation for Internet use.

Applications of New Internet U&G Factors

With a newly identified social motivation for Internet use, important implications arise. Unlike any of the traditional media, the Internet can be considered as both an interpersonal and a mass exposure medium, with simultaneous gratifications along several mediated channels (Stafford & Stafford, 2001a).

Stafford's (2001) confirmatory analysis of Internet U&G dimensions was applied to investigate Internet diffusion in the consumer market (Stafford, 2003b). Significant differences between innovation adoption categories were found for specific Internet U&G factors. Internet laggards, for example, exhibited the lowest social gratification for online services, while Internet "innovators" (or, early adopters) exhibited the highest social gratifications. Innovators also appeared to be significantly more motivated by content gratifications for Internet use. In a study of AOL users and their uses and gratifications for Internet use (Stafford, 2003a), heavy users also scored higher than light users for both process and social Internet usage gratifications. AOL users were strongly motivated by Web browsing and the guided search for information, in addition to their appreciation for social gratifications related to Internet communications functionality. Interestingly, among all the potential indicators related to U&G factors examined in analysis, online shopping did not appear to be highly gratifying to AOL users (Stafford & Gonier, 2004).

The impact of social gratifications for the Internet appears to be a function of user experience, and a rising generalization is that heavy Internet users are more motivated by

social gratifications than light users (Stafford et al., 2002), which tends to fit with emerging views of Internet user demographics (e.g., Emmannouildes & Hammond, 2000). As an aspect of consumer media marketing strategy, heavy users are desirable consumer targets for ISPs, and this trend could have valuable implications for practice.

Internet U&G factors have also been applied to understand motivations for Internet use in distance education classrooms (Stafford, 2005; Stafford & Stafford, 2003). In the student use of distance course Internet resources, social gratifications were dominant; the conclusion was that students on the remote end of a distance education teleconference feel socially removed and isolated from colleagues in the live sections of classes (cf., Berger & Topol, 2001; Hamer, 2001). Internet technology, including chat and e-mail, as well as IP teleconferencing, acts to reduce the social isolation of distance education courses (Stafford, 2005).

Relationships between Internet-Specific U&G Dimensions

In a study seeking to establish trait validity for Interent U&G factors (Stafford, 2001), process and content gratifications for the Internet were found to be highly related ($\Phi = .72$), but there were much weaker relationships between the social factor and the more traditional process and content gratifications, which tends to reinforce the multi-modal conceptualization of the Internet as providing both personal entertainment and interpersonal communications channels in one meta-channel. The usage process gratification was only moderately related to social gratifications ($\Phi = .38$), and the content gratification was also only moderately related to social gratifications ($\Phi = .34$), so social gratifications for the Internet are distinct from (and not closely related to) standard process and content gratifications for Internet use. The social gratification construct did display excellent measurement qualities, indicated by measurement model fit indices (GFI = .97, AGFI = .95, RMSR = .11, SRMR = .043, NFI = .96, CFI = .97), so social gratifications appear to be a distinct and trait valid area of motivation for Internet use, if not directly related to the better-understood process and content gratifications.

THE SECOND WAVE OF INTERNET MOTIVATION RESEARCH

Now that the Internet is an accepted, even mundane, aspect of modern life, media gratifications studies of the medium are beginning to expand beyond basic identification of Internet-specific gratification dimensions, and to begin to deal with more applied behavioral perspectives of daily Internet use. More applied perspectives of Internet media user motivations now include themes of Internet uses and gratifications related to online advertising (cf., Ko, Cho, & Roberts, 2005; Yang, 2004), the motivations for Internet use among business channel supply chain members (Zank & Vokurka, 2003), and even maladaptive motivations related to over-use and Internet addiction (Song, LaRose, Eastin, & Lin, 2004).

U&G is a mature theoretical paradigm, dating from the advent of the radio era, and it has quickly been adapted to understanding aspects of Internet use. Researchers are also applying other mature theoretical models in similar fashion. For instance, the venerable technology acceptance model sees plenty of application in Internet use research, including explorations of recent evolutions of the model toward more affective components of judgment; playfulness and Internet acceptance have been investigated (Chung & Tan, 2004), and recent advances relayed in the recent unified model (Venkatesh, Morris, Davis, & Davis, 2003) related to extrinsic and intrinsic motivations have been expanded upon (Lee, Cheung, & Chen, 2005; Shang, Chen, & Shen, 2005). These Internet motivational schemes begin to take the study of Internet use away from the media-use motivations approach exemplified by the reference discipline perspectives of communications theory found in U&G and toward the more established technology use perspective favored by the community of IT scholars.

Future adaptations of reference discipline theories to the understanding of Internet use (and, indeed, both U&G and TAM approaches are adaptations of popular theories brought to the IT field from influential reference disciplines) will likely apply behavioral aspects of atmospherics, cultural artifacts, and social network theories, as we begin to understand the Internet as an established and influential multi-modal medium deeply integrated into the emerging modern lifestyle.

CONCLUSIONS

Media use theories have been useful in the intermediate days of Internet development, as scholars have undertaken to understand the multiple channel medium as a converged instance of communication and entertainment unlike others known in society previously. As Internet use becomes more mundane and widespread in modern society, scholars can begin to utilize theories of human and consumer behavior that anchor Internet use in lifestyle and workplace contexts at a more applied level. Yet, to this day, the importance of understanding the motivations for Internet use that arise from its characteristic as a multimodal mediated channel of information and entertainment will remain important as new and more spectacular uses are devised for this robust and rich channel.

Internet social gratifications will be particularly important in emerging perspectives of the network as a manifestation

of culture, lifestyle, and work, since the social gratifications afforded by the Internet are unlike those of any other mediated channel in previous experience, notwithstanding the numerous informational and entertainment gratifications the network provides. Unlike other media studied in the past, where sought content and enjoyment of media usage processes were the defining factors of motivation, the Internet provides standard media gratifications along with interpersonal connectivity, giving it a multi-modal appeal and influence in human life.

Interpersonal communication and interaction with other people over the Internet—the extension and maintenance of social networks via technologically mediated telecommunications networks—seems to characterize the social Internet gratification. Emerging perspectives of Internet user motivations in the recent second wave of Internet use research rely more on motivational bases than could be characterized in line with functional utility (content-based information resources, for example, and usefulness perspectives), or on the affective gratifications that derive from the enjoyable usage processes of the medium (playfulness, atmospherics, and various affective approaches to Internet use motivations relate strongly the Internet process gratifications).

Regardless of the use or gratification that a business may wish to impact in its provision of Internet service and utility, the important point is to realize that users actively interact with the medium, and that uses and gratifications is one of the best ways that exists in which to study active audience motivations for media use. Determining what audience members seek to do, and the benefit they expect to accrue from activity, is the important step in marketing and supporting Internet offerings.

Emerging product utility-based conceptualizations of Internet user motivations notwithstanding, U&G's "active-use" tenet continues to remind us to be customer-centric in our design and provision of products and services, in the realization that users want specific things and they want them because these things bring them enjoyment. Goal oriented activity is the lynchpin of motivation, and understanding the goals sought by our Internet customers will allow us to design and provide more compelling and satisfying offerings. Inevitably, business objectives will be more fully attained in a customer-centric media-based approach to the Internet industry.

REFERENCES

Berger, K. A., & Topol, M. T. (2001). Technology to enhance learning: Use of a Web site platform in traditional classes and distance learning. *Marketing Education Review, 11*(3), 15-26.

Chung, J., & Tan, F. B. (2004). Antecedents of perceived playfulness: An exploratory study on user acceptance of general information-searching Web sites. *Information & Management, 41*(7), 869-881.

Cutler, N. E., & Danowski, J. A. (1980). Process gratification in aging cohorts. *Journalism Quarterly, 57*(2), 269-277.

Dimmick, J., Chen, Y., & Li, Z. (2004). Competition between the Internet and traditional news media: The gratification-opportunities niche dimension. *The Journal of Media Economics, 17*(7), 19-33.

Eighmey, J. (1997). Profiling user responses to commercial Web sites. *Journal of Advertising Research, 37*(3), 59-66.

Eighmey, J., & McCord, L. (1998). Adding value in the information age: Uses and gratifications of sites on the World Wide Web. *Journal of Business Research, 41*(3), 187-194.

Emmannouildes, C., & Hammond, K. (2000). Internet usage: Predictors of active users and frequency of use. *Journal of Interactive Marketing, 14*(2), 17-32.

Hamer, L. O. (2001). Distance learning technologies as facilitators of learning and learning-related student activities. *Marketing Education Review, 11*(3), 55-67.

Katz, E. (1959). Mass communication research and the study of popular culture: An editorial note on a possible future for this journal. *Studies in Public Communication, 2*, 1-6.

Katz, E., Blumler, J. G., & Gurevitch, M. (1974). Uses of mass communication by the individual. In W. P. Davison & F. T. C. Yu (Eds.), *Mass communication research: Major issues and future directions* (pp. 11-35). New York: Praeger.

Ko, H., Cho, C., & Roberts, M. S. (2005). Internet uses and gratifications: A structural equation model of interactive advertising. *Journal of Advertising, 34*(2), 57-70.

Kukar-Kinney, M., Ridgway, N. M., & Monroe, K. B. (2006). The relationship between consumers' tendencies to buy excessively and their motivations to shop and buy on the Internet. In J. Johnson & J. Hulland (Eds.), *Proceedings of the 2006 American Marketing Association Conference* (pp. 36-37). Chicago, IL.

Lee, M. K. O., Cheung, C. M. K., & Chen, Z. (2005). Acceptance of Internet-based learning medium: The role of extrinsic and intrinsic motivation. *Information & Management, 42*(8), 1095-1104.

Levy, M. R., & Windahl, S. (1984). Audience activity and gratifications: A conceptual clarification and exploration. *Communication Research, 11*(1), 51-78.

Lin, N. (1977). Communication effects: Review and commentary. In B. Rubin (Ed.), *Communication yearbook 1* (pp. 55-72). New Brunswick, NJ: Transaction Books.

Lohse, G. L., & Spiller, P. (1998). Electronic shopping. *Communications of the ACM, 41*(7), 81-87.

McDonald, S. C. (1997). The once and future Web: Scenarios for advertisers. *Journal of Advertising Research, 37*(2), 21-28.

McGuire, W. J. (1974). Psychological motives and communication gratifications. In J. Blumler & E. Katz (Eds.), The uses of mass communications: Current practices on gratifications research. Beverly Hills, CA: Sage Publications.

Newhagen, J., & Rafaeli, S. (1996). Why communication researchers should study the Internet: A dialogue. *Journal of Communication, 46*(1), 4-13.

Rafaeli, S. (1988). Interactivity: From new media to communication. In R. Hawkins, J. Wieman, & S. Pingree (Eds.), *Advancing communication science: Merging mass and interpersonal processes* (pp. 110-134). Newberry Park, CA: Sage Publications.

Rubin, A. M. (1981). An examination of television viewing motivations. *Communication Research, 8*(3), 141-165.

Schonberg, E., Cofino, T., Hoch, R., Podlaseck, M., & Spraragen, S. L. (2000). Measuring success. *Communications of the ACM, 43*(3), 53-57.

Shang, R., Chen, Y., & Shen, L. (2005). Extrinsic versus intrinsic motivations for consumers to shop online. *Information & Management, 42*(3), 401-413.

Song, I., LaRose, R., Eastin, M. S., & Lin, C. A. (2004). Internet gratifications and Internet addiction: On the uses and abuses of new media. *CyberPsychology & Behavior, 7*(4), 384-394.

Stafford, M. R., & Stafford, T. F. (2000). Identifying the uses and gratifications of Web use. In M. Shaver (Ed.), *Proceedings of the 2000 American Academy of Advertising Conference*, Newport, RI (pp. 70-71). Lansing, MI: American Academy of Advertising.

Stafford, T. F. (1999). Consumer motivations for commercial Web site use: Antecedents to electronic commerce. In D. Nazareth & D. Goodhue (Eds.), *Proceedings of the Association for Information Systems 1999 Americans Conference on Information Systems*, Milwaukee, WI (pp. 544-546). Atlanta, GA: Association for Information Systems.

Stafford, T. F. (2001). *Motivations related to consumer use of online services.* Unpublished doctoral dissertation, University of Texas, Arlington.

Stafford, T. F. (2003a). Social and usage process motivations for Internet use: Differences between light and heavy users. In D. Galletta & J. Ross (Eds.), *Proceedings of the 2003 Americas Conference for Information Systems*, San Diego, CA (pp. 2248-2254). Atlanta, GA: Association for Information Systems.

Stafford, T. F. (2003b). Differentiating between innovators and laggards in the uses and gratifications for Internet services. *IEEE Transactions on Engineering Management, 50*(4), 427- 435.

Stafford, T. F. (2005). Understanding motivations for Internet use in distance education. *IEEE Transactions on Education, 48*(2), 301-306.

Stafford, T. F., & Gonier, D. (2004). Gratifications for Internet use: What Americans like about being online. *Communications of the ACM, 47*(11), 107-112.

Stafford, T. F., & Stafford, M. R. (1998). Uses and gratifications of the World Wide Web: A preliminary study. In D. Muehling (Ed.), *Proceedings of the 1998 American Academy of Advertising Conference*, Lexington, KY (pp. 174-182). Pullman, WA: American Academy of Advertising.

Stafford, T. F., & Stafford, M. R. (2000). Consumer motivations to engage in electronic commerce: Uses and gratifications of the World Wide Web. In S. Rahman & M. Raisinghani (Eds.), *Electronic commerce: Opportunities and challenges*. Hershey, PA: Idea Group Publishing.

Stafford, T. F., & Stafford, M. R. (2001a). Identifying motivations for the use of commercial Web sites. *Information Resources Management Journal, 14*(1), 22-30.

Stafford, T. F., & Stafford, M. R. (2001b). Investigating social motivations for Internet use. In O. Lee (Ed.), *Internet marketing research: Theory and practice* (pp. 93-107). Hershey, PA: Idea Group Publishing.

Stafford, T. F., & Stafford, M. R. (2003). Uses and gratifications for Internet use in the distance education classroom. In *Proceedings of the 2003 American Marketing Association Winter Educators Conference*. Orlando, FL.

Stafford, T. F., Stafford, M. R., & Shaw, N. (2002). Motivations and perceptions related to the acceptance of convergent media delivered thought the World Wide Web. In M. Khosrow-Pour (Ed.), *Advanced topics in information resources management* (pp. 116-126). Hershey, PA: Idea Group Publishing

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Yang, K. C. C. (2004). Effects of consumer motives on search behavior using Internet advertising. *CyberPsychology & Behavior, 7*(4), 430-442.

Zank, G. M., & Vokurka, R. J. (2003). The Internet: Motivations, deterrents, and impact on supply chain relationships. *SAM Advanced Management Journal, 68*(2), 33-40.

KEY TERMS

Active Audience: Uses and Gratifications theory presumes media users are actively involved in selection and use of media and are not passive recipients. This implies the need to specifically target media offerings to perceived user needs.

Content Gratification: Enjoyment of message specifics. Content can mean information, and often does, though it also includes entertainment in the form of medium-carried programming.

Gratifications: What people derive from use—the "why" of media use motivations.

Meta-Medium: A channel of channels, such as the Internet. This term conveys the sense of rich and complex media transmission across a multiplexed channel of conveyance. Numerous motivations for use could arise related to such a complex media venue.

Process Gratification: Enjoyment of media use, as distinguished from enjoyment of specific message content. This is much the same as channel surfing or Web browsing with no goal other than entertainment through engaging in the activity.

Uses: Things people do *with* media—the "how" of media use motivations.

Uses and Gratifications: Customer activities and the enjoyment that derives from such activities, particularly in a mass media context.

Multi-Agent Mobile Tourism System

Soe Yu Maw
University of Computer Studies, Myanmar

Ni Lar Thein
University of Computer Studies, Myanmar

INTRODUCTION

Nowadays, with the emergence of high-speed wireless networks, various portable devices such as personal digital assistant (PDAs), mobile phones, and other wearable equipment are widely used by people in their daily lives. Context-awareness plays a vital role in enabling smart environments, wearable computing, and wireless computing. This article presents the multi-agent system, which uses mobile technology to offer services in the tourism domain.

Agent-based systems are widely used for mobile and distributed information systems. Agents can also help in preventing the user from being overwhelmed by irrelevant information using personalization methods. This technology provides the integration of information from diverse sources, while personalization provides the filtering technique to deliver the relevant information to the users.

The system gives up-to-date information based on the user's preferences and other contextual information such as sight location, weather condition, and special functions that are arranged during the visit. The system consists of two types: Web-based and mobile-based. We design the system as client-server architecture, supporting desktop clients as well as mobile clients on a handheld device with appropriate interfaces. However, in this article, we now focus on the mobile-based tourism system. The handheld device or PDA is used for receiving information from a Web server.

In past years a broad spectrum of different Web-based tourism has been established. The acceptance and consequently the competitiveness of a tourism system are mainly determined by the quantity and quality of data it provides. Therefore, most existing tourism systems try to fulfill the tourist's request (interest) for an extensive data collection (Rumetshofer & Wob, 2005).

Tourism information (e.g., travel schedules, etc.) are distributed, dynamic, and heterogeneous. The users (tourists) may face difficulty using them when planning their trips.

Nowadays, the improvements in wireless communication technologies such as handheld devices to the Internet open up new prospects for e-commerce and e-tourism. Today, new technologies allow more flexible access to information booking services and other tourist support (Belz, Nick, Poslad, & Zipf, 2002). Tourism has been a popular area for mobile information systems. There are a number of obstacles to introducing new technology in tourism. Electronic guidebooks and maps have been a popular application area for mobile technology.

In the near future, a broader range of services will become available to users anywhere, at any time. People can receive their required information by interacting with their PDA from wherever they are. Kanellopoulos and Kotsiantis (2006) stated that the tourism industry makes efforts to implement techniques that can reduce travel cost and improve performance.

A major issue in offering mobile services to nomadic users is the limited display and networking capacity of mobile devices such as wireless application protocol (WAP) phones or PDAs. A possible solution for this is the adaptation of services and contents to the users' personal interests and their current location. The adaptation of services and contents to personal interests mainly filter the available information. Poslad et al. (2001) described the filtering process as based on a user profile describing the interests, abilities, and characteristics of the user.

Ding, Malaka, and Pfisterer (2002) described multi-agent systems as particularly well suited for mobile information systems, and some systems even allow for resource-aware computations in mobile and distributed environments.

Maw and Naing (2006) described the architecture and design of a multi-agent tourism system (MATS). MATS evaluated the similarity value and mean absolute error (MAE) to give the best recommendation to the user.

The central motivation of this article is to extend MATS suitable for the mobile user in the tourism domain. The objective is to give the user the most relevant and updated information according to the user's interest.

The next section provides the definitions and a discussion of the system, and reviews literature of some related works. The main focus of the multi-agent mobile tourism system architecture is then described, and the design considerations and future trends are also discussed, before we conclude the article.

BACKGROUND

In this section, we provide definitions and review some related works. There are a number of research projects related to the tourism system. The multi-agent system, personalization methods, and mobile technology are discussed in the literature.

Wooldridge (2000) defined multi-agent system as a collection of agents that work in conjunction with each other. In multi-agent system architecture, each agent is autonomous, cooperative, coordinated, intelligent, and able to communicate with other agents to fulfill the user's requirements.

The characteristic of agent proactiveness can provide the user with information related to his or her profile. Agents are autonomous, so they act on behalf of a user to reach a goal or solve a problem for the user. Agents filter information that is suitable for the user on the basis of a user profile that stores the user's interests. Agents need an agent platform, which provides communication for the agents, other services, and security features.

There are many advantages in the mobile devices of agent technology, such as providing services to the user in a personalized way.

Personalization means knowing who the user is, what the user's interests are, and recognizing a specific user based on a user profile. Willy (2001) and Rumetshofer and Wob (2005) defined personalization as a process of gathering and storing information about users, analyzing the information, and based on the analysis, delivering the information to each user at the right time.

Personalization is also an important feature of mobile services. Personalization adapts the services to the user location, user preferences, and user profile.

Poslad et al. (2001) described that "Creation of User-friendly Mobile services Personalized for Tourism" (CRUMPET) is a mobile application that uses multi-agent technology to construct a context-aware system. The system combines personalized services, multi-agent technology, location-aware services, and transparent mobile data communication, altogether in order to facilitate the users. The services provided by CRUMPET take advantage of integrating four key emerging technology domains and applying them to the tourist domain: location-aware services, personalized user interaction, seamlessly accessible multimedia mobile communication, and smart component-based middleware that uses multi-agent technology. Its use is mainly limited to providing query and recommendation services.

The GUIDE project (Simcock, Hillenbrand, & Thomas, 2003, Cheverst, Mitchell, Friday, & Davies, 2000) studies the electronic tourist guide system of Lancaster City. It obtains the user position by receiving location messages transmitted from non-overlapping WaveLAND cell base stations dispersed throughout the city. Wireless communication was used via a pen-based tablet computer. While this approach does not need additional hardware on the client side, it results in a lower resolution of positioning information.

The Cyberguide project (Abowd et al., 1997) is a handheld electronic tourist guide system that supplies the user with context-sensitive information. It was built in the mid-1990s, and its goal was to provide the information to a user based on the user's position and orientation. The application is hosted entirely on an Apple MessagePad and used infrared beacons for positioning using a Trimble GPS unit.

Hinze and Buchanan (2006) described the tourist information provider (TIP) as a mobile tourist information system that presents information to the user that is sensitive to the user's context, interest, and the related context of neighboring sights of interest. TIP provides map-based and browser-based information navigation, and uses contextual hierarchy to support outliner-style browsing that is efficient on small-screen, mobile displays (Hinze & Buchanan, 2005). The user dynamically interacts with the system by providing his or her current location while asking for information from the system. TIP also gives recommendation to users based on their current position and the information in their user profiles. They get a list of nearby sights which they might like to visit if they request the system's recommendation.

The IMAGE system proposed e-services for mobile users and introduced the personalized service feature, using agent technology in the context of the IST project IMAGE. In the IMAGE system all roles are implemented as agent types with the exception of the social role, which is realized by all agents that need to be acquainted with their collaborative partners. Their system integrated a set of intelligent agents having different functionalities (e.g., personalized assistance, travel information, and cultural events information), which are necessary in order to cover the needs presented by this specific application field called mobile personalized location-based services.

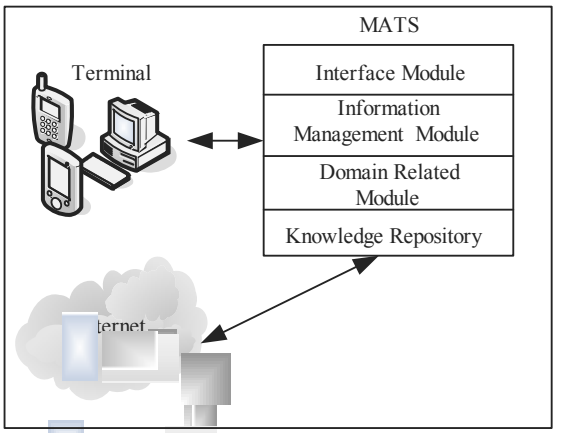
We also provide the services for mobile users by using personalized services with agent technology. The system uses GPS data to recognize the user location. The system gives the recommendation to the user according to the location of the user and user preferences.

MULTI-AGENT MOBILE SYSTEM ARCHITECTURE

A mobile system can be deployed in a wide range of physical environments to support users in diverse tasks. Advanced mobile systems need to exploit information about the environment in which they are working.

In this article, we extend our previous MATS to a mobile system. Figure 1 shows the general architecture of a multi-agent mobile system. A more detailed description is provided in Maw and Naing (2006).

Figure 1. General architecture of proposed system

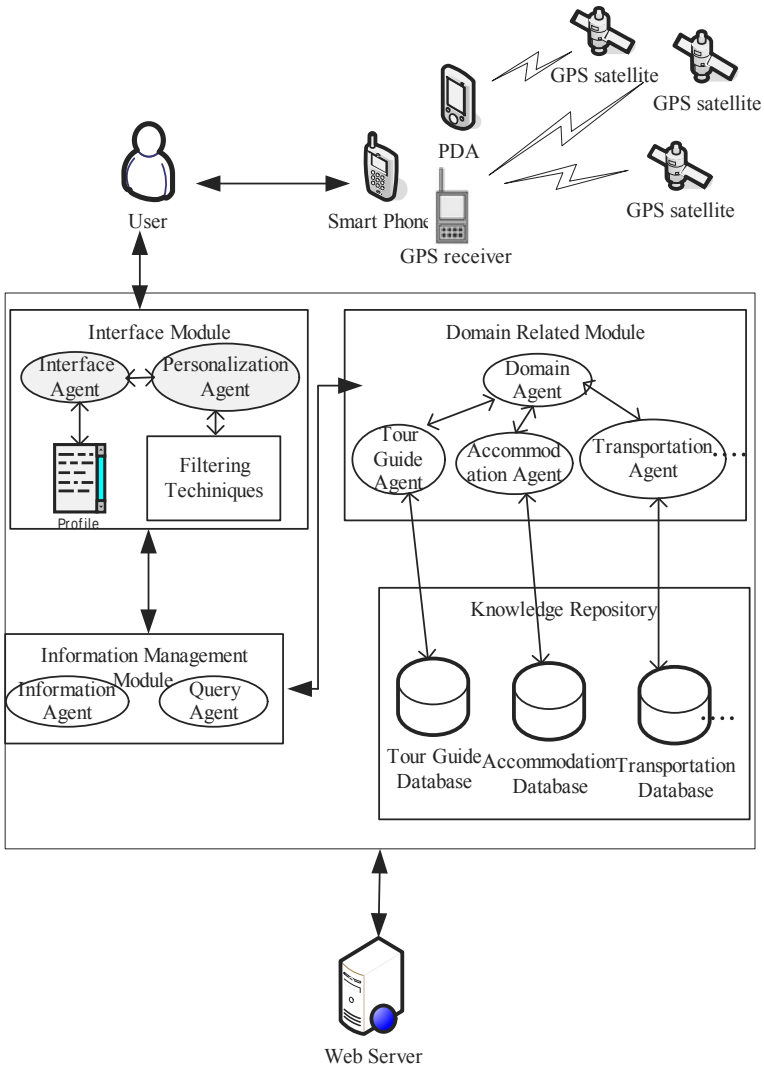


As shown in Figure 2 we provide the actual deployment of agents in smart or mobile devices (such as a PDA or smart phone). In this system, the client device is a handheld device or PDA used as a terminal for receiving information from a Web server and receiving signal from the GPS satellite.

The users are required to define their preferences (interests) as user profiles. Users' information can be changed whenever they want to revise their interests. We use the combination of rule-based personalization with collaborative filtering technique, which can give the user information effectively and efficiently according to the user's interests.

The architecture of multi-agent system is composed of four main modules:

Figure 2. Design architecture of multi-agent mobile system



- interface module,
- information management module,
- domain related module, and
- knowledge repository.

Each part is detailed in Maw and Naing (2006). A brief explanation of each module follows.

In the interface module, the interface agent gets users' information and creates user profiles. The personalization agent performs rule-based personalization with collaborative filtering technique. This technique computes the similarity measure between the users and gives recommendations to the users according to their profiles (Maw & Naing, 2006).

The information management module has two sub modules, namely information agent and query agent. The information agent manages the data to be stored in a corresponding database whenever changes or updates of information occur. The query agent retrieves the information from the knowledge repository through a domain-related module and returns the results to the personalization agent.

The domain-related module is composed of domain agent and many other agents such as tour guide agent, accommodation agent and transportation agent, and so forth. The domain agent performs database queries from the corresponding databases and delivers specific domain solutions.

The knowledge repository consists of tourism information for each domain in a standard structure. It maintains information and makes that information available on demand.

PERSONALIZATION SERVICES

Short (2000) stated that personalization is a key feature that facilitates the use of complex services on mobile devices.

The personalization agent performs the task of gathering the user information, interests, or preferences, and stores the context of information as the user profiles. The next step is to analyze the information and use rule-based personalization with collaborative filtering technique.

Rule-Based Personalization

Rule-based personalization is an important part to exposing relevant content to the user. This yields a very efficient reduction of the number of items that must be processed at a very early stage of the filtering process. It defines a set of rules to tailor the content based on the facts specified in the user profile. The advantage of rule-based personalization is that rules can be adopted for use in any platform and thus are not limited to the Web.

Collaborative Filtering

The collaborative filtering technique is used to obtain one or more numerical ratings for every item. The task is to predict the interest of an active user to a targeted item based on the user profiles. A collaborative filtering system looks for the users who share the same rating pattern with the active user and use the ratings from users to calculate a prediction for the active user. It is the process of computing personalized recommendation by finding users with similar taste. The Pearson correlation coefficient is used to compute the similarity measure (see details in Maw & Naing, 2006).

AGENT PLATFORMS FOR MOBILE DEVICES

An agent needs an agent platform to provide services such as communications of agents. The Foundation for Intelligent Physical Agent (FIPA) is a de facto standard for agent communication languages. To facilitate the development of a multi-agent system, we use the Java Agent Development Framework (JADE). The agents running on the mobile devices use a Light Extensible Agent Platform (LEAP) add-on. JADE-LEAP is used for enabling FIPA agents to execute on lightweight devices running Java. The graphic user interface (GUI) is developed using the Java 2 Micro Edition (J2ME) specification.

LOCATION-BASED SERVICES

Mobile users are moving from one place to another and passing through different locales. Location-based services can provide the location of the user. There are a variety of methods and devices for collecting and sensing the location of the user. Global positioning systems (GPSs), infrared sensors, Bluetooth, the Global System for Mobile Communications (GSM), and the universal mobile telecommunication system (UMTS) are positioning techniques that can be used for sensing location. The most well-known location-sensing system today is GPS.

One of the most powerful ways to personalize mobile services is based on location. We use GPS sensor data to determine the user's location. Far (2005) described that GPS-enabled devices can obtain latitude and longitude with accuracy of about 1.5 M. GPS devices use triangulation techniques by triangularity data points from the satellite constellation that covers the entire surface of the earth.

PDAs have definitely evolved over the years. Not only can they manage the personal information, such as contacts,

appointments, and to-do lists, but today's devices can also connect to the Internet, act as GPS devices, and run multimedia software.

For anyone with a GPS receiver, the system will provide location and time. A GPS provides accurate location and time information for an unlimited number of people in all weather, day and night, anywhere in the world.

The mobile device transmits the subscriber information to the network, and the GPS generates the location of the user. The system can give the information to the user relevant to the user's interest and the location of the user. By knowing the user's location, the system can give recommendations to the user of nearby places of interest.

FUTURE TRENDS

The field of context-aware mobile systems is an emerging research topic. There are many techniques to acquire user location, but mobile devices still have security and privacy issues. For our future work we will focus on the localization system that involves context-awareness of more dynamic information of user location, security, and privacy issues.

Although GPSs have many advantages, there have some limitations in that the issues of such systems are not solved efficiently, such as out-of-signal, inside building, and cost issues. Today's mobile phone comes with a built-in GPS receiver that can be used inside a building and can perform route planning and navigation. We will focus on solving the limitations of conventional the GPS system.

CONCLUSION

In this article, we have presented the architecture of a multi-agent system for mobile devices such as PDAs or handheld devices in the tourism domain. Multi-agent systems efficiently retrieve and filter information from sources that are spatially distributed. Personalization is also a prevailing means to facilitate the use of mobile systems. By using personalization methods, the system can give the best recommendation according to the user's interest. Contextual information such as a user's interests (preferences) is stored in user profiles. By performing the rule-based personalization with collaborative filtering technique, the system gives the relevant information to the user. We described enhancing MATS on mobile devices with GPS.

The main purpose of this article is to discuss the personalization method used in the tourism system and the benefits of the use of a personalized mobile tourism system.

To sum up, this article provides a multi-agent system architecture to offer services in the tourism industry, allowing different users the possibility to obtain up-to-date information about the places they will visit and to plan a specific day.

REFERENCES

Abowd, G., Atkeson, C., Hong, J., Long, S., Kooper, R., & Pinkerton, M. (1997). Cyberguide: A mobile context-aware tour guide. *ACM Wireless Networks, 3*(5), 421-433.

Belz, B., Laamanen, H., Poslad, S., & Zipf, A. (2003). Location-based mobile tourist services-first user experience. *Proceedings of the International Conference on Tourism and Communication Technologies,* Helsinki, Finland.

Belz. B., Nick, A, Poslad, S., & Zipf, A. (2002). Personalized and location-based mobile tourism services. *Proceedings of the 4th International Symposium on Human Computer Interaction with Mobile Devices.* Retrieved October 12, 2006, from http://www2.geoinform.fh-ainz.de/~zipf/mobileHCI-crumpet.pdf

Cheverst, K., Mitchell, K., Friday, A., & Davies, N. (2000). Sharing (location) context to facilitate collaborative between city visitors. *Proceedings of the IMC'00 Workshop on Interactive Applications of Mobile Computing,* Rostock, Germany.

Ding, Y., Malaka, R., & Pfisterer, D. (2002). An open framework for load balanced multi-agent systems. *Proceedings of the Workshop on Ubiquitous Agents on Embedded, Wearable, and Mobile Devices, held in conjunction with the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems* (AAMAS 2002), Bologna, Italy.

Far, R.B. (2005). *Mobile computing principles: Designing and developing mobile applications with UML and XML*. Cambridge: Cambridge University Press.

Hinze, A., & Buchanan, G. (2005). Context-awareness in mobile tourist information systems: Challenges for user interaction. *Proceedings of the International Workshop on Context in Mobile HCI at the 7th International Conference Human Computer Interaction with Mobile Devices and Services,* Salzburg, Austria.

Hinze, A., & Buchanan, G. (2006). The challenge of creating cooperation mobile services: Experiences and lessons learned. *Proceedings of the 29th Australasian Computer Science Conference* (ACSC 2000), Hobart, Australia.

Kanellopoulos, D., & Kotsiantis, S. (2006). Towards intelligent wireless Web services for tourism. *Proceedings of*

the IJCSNS International Journal of Computer Science and Network Security (vol. 6, p. 7B).

Maw, S.Y., & Naing, M.-M. (2006). Multi-agent tourism system. In *Social information retrieval systems: Emerging technologies and applications for searching the Web effectively.* Hershey, PA: Idea Group.

Poslad, S., Laamanen, H., Malaka, R., Nick, A., Buckle, P., & Zipf, A. (2001). CRUMPET: Creation of user-friendly mobile services personalized for tourism. *Proceedings of the 3G2001 Mobile Communication Technologies Conference* (pp. 28-32).

Rumetshofer, H., & Wob, W. (2005). Semantic maps and meta-data enhancing e-accessibility in tourism information systems. *Proceedings of the 16th International Workshop on Database and Expert Systems Applications* (DEXA'05).

Short, M. (2000). My generation: Third generation wireless mobile communication. *Electronics and Communication Engineering Journal, 12*(3), 119-122.

Simcock, T., Hillenbrand, S.P., & Thomas, B.H. (2003). Developing a location-based tourist guide application. *Proceedings of the Workshop on Wearable, Invisible, Context-Aware, Ambient, Pervasive and Ubiquitous Computing* (vol. 21), Adelaide, Australia.

Willy, C. (2001). *Web site personalization.* Retrieved May 31, 2005, from http://www.128.ibm.com/developerworks/websphere/library/techarticles/hipods/personalize.html

Wooldridge, M. (2002). *An introduction to multiagent system.* London: John Wiley & Sons.

KEY TERMS

Collaborative Filtering (CF): The process of filtering for information using techniques involving collaboration among multiple agents, viewpoints, data sources, and so forth. It is the method of making automatic predictions about the interests of a user by collecting taste information from many users.

Location-Based Service (LBS): Used to locate a user and provide services specific to the location the users are in at the time by using the power of mobile networks.

Mobile Device: A handheld device such as a smart phone, Bluetooth headset, personal digital assistant (PDA), pager, notebook PC, and so on. PDAs and smart phones are popular mobile devices.

Mobile Technology: The part of technology that involves mobility. Mobile technology includes general packet radio service (GPRS), multimedia messaging service (MMS), Bluetooth, 3G, wireless fidelity (WiFi), global positioning system (GPS), CLI, wireless application protocol (WAP), and short message service (SMS).

Multi-Agent System: A collection of agents that acts on behalf of user in an autonomous way, and in which each agent communicates with the network of agents and then makes decisions to match demand.

Personal Digital Assistant (PDA): A handheld computer that can access the Internet, intranet, or extranets via wireless wide area networks (WWANs) and global positioning systems (GPSs).

Personalization: A process of knowing who the user is, what the user wants, and recognizing a specific user based on a user profile. Delivers the information to each user at the right time.

Rule-Based Filtering: Refers to the personalization resulting from a match of a user profile with content profile based on rules. This form of personalization implements rules based on a user's profile.

Multi-Agent Simulation in Organizations: An Overview

Nikola Vlahovic
University of Zagreb, Croatia

Vlatko Ceric
University of Zagreb, Croatia

INTRODUCTION

Most economic and business systems are complex, dynamic, and nondeterministic systems. Different modeling techniques have been used for representing real life economic and business organizations either on a macro level (such as national economics) or micro level (such as business processes within a firm or strategies within an industry). Even though general computer simulation was used for modeling various systems (Zeigler, 1976) since the 1970s the limitation of computer resources did not allow for in-depth simulation of dynamic social phenomena. The dynamics of social systems and impact of the behavior of individual entities in social constructs were modeled using mathematical modeling or system dynamics.

With the growing interest in multi agent systems that led to its standardization in the 1990s, multi agent systems were proposed for the use of modeling social systems (Gilbert & Conte, 1995). Multi agent simulation was able to provide a high level disintegration of the models and proper treatment of inhomogeneity and individualism of the agents, thus allowing for simulation of cooperation and competition. A number of simulation models were developed in the research of biological and ecological systems, such as models for testing the behavior and communication between social insects (bees and ants). Artificial systems for testing hypothesis about social order and norms, as well as ancient societies (Kohler, Gumerman, & Reynolds, 2005) were also simulated.

Since then, agent-based modeling and simulation (ABMS) established itself as an attractive modeling technique (Klugl, 2001; Moss & Davidsson, 2001). Numerous software toolkits were released, such as Swarm, Repast, MASON and SeSAm. These toolkits make agent-based modeling easy enough to be attractive to practitioners from a variety of subject areas dealing with social interactions. They make agent-based modeling accessible to a large number of analysts with less programming experience.

BACKGROUND

Computer simulation modeling is an established method in scientific and industrial applications, appropriate for obtaining insight into the dynamics of organizations. Modeling is used to represent a part of reality in sufficient detail, and resulting model is an artificial system used for experimentation. There are several situations when replacement of the real system by an artificial one is helpful or even necessary.

- **Inaccessibility of the real world system:** Sometimes a part of the real world system that should be studied, is not accessible either because the system does not exist any more or is not yet put into operation.
- **Real world system is inappropriate for experimentation:** Some real world systems may be affected in undesired way by experimentation. Examining effects of drastic changes in taxing and pricing policies may for example, disturb the fiscal system, or discourage production and consumption.
- **Time scale or behavior of the system is inappropriate for observation:** A number of systems such as investments in some industries generate results over long periods of time, making it hard to collect enough data from the real system for a meaningful analysis. Simulation is using virtual time that can be accelerated or slowed down as needed in order to observe a particular phenomenon.
- **Intensive dynamics of the system:** All elements of simulation model can be taken under full control. This is especially important in economics for the purpose of studying the impact of changes in one factor on behavior of the whole system, while holding all other factors at the same level. This presumption cannot be achieved in a real life economic system (e.g., system of supply and demand).

The model should be able to answer questions directed to the real system. However, it can produce valid output only for the set of experiments defined by the *experimental frame* (Zeigler, 1976) determined in the early stages of

model development. After the model is successfully built, simulation experiments can be performed. In order to gain full control over the experiment, a simulation model is used in a predefined *artificial environment* and a predefined *virtual time* of simulation.

Treatment of virtual time is crucial for selection of the simulation method applied to the model. (1) If virtual time is continuous, then ***system dynamics*** is used. System dynamics focuses on feedback loops of the model whose behavior is represented by differential equations. Model is restricted to macro level, and its properties are described by attributes which represent the state of the system and its changes. System dynamics is used for analysis of the behavior of complex real systems on a macro level, in management, politics, economics, environmental change and so forth. Important advantage of system dynamics is its efficiency due to its high degree of abstraction.

(2) If virtual time is divided into a series of discrete periods, then **event-based simulation** is used (Seila, Ceric, & Tadikamalla, 2003). *Discrete event simulation* advances "time" to the moment (denoted by time stamp) in which at least one model entity needs to execute certain action or change its state. *Simulation clock* defines the beginning and the ending moments of time required for simulation execution. Discrete-event simulation models are primarily used for functional verification and performance evaluation of real world systems.

Standard methods for concurrent processes modeling are queuing networks, Petri nets and cellular automata. (3) **Queuing networks** and **Petri nets** do not include mechanisms for representing inhomogeneous space where the number of entities, their interactions and behavior change over time in dependence on their surroundings. If a conflicting situation occurs in the environment, then probabilistic factors or predefined fixed amounts of system resources and length of activities are used. (4) **Cellular automata** are purely space-based representations where each cell value is calculated on the basis of values of neighboring cells. However, modeling of individual behavior of an entity or modeling of deduction rules using cellular automata requires complex models and overwhelming computing resources (Klügl, Oechslien, Puppe, & Dornhaus, 2004).

Some of the shortcomings of system dynamics and discrete event simulation as well as other methods for modeling concurrent processes can be overcome by the **multi agent simulation** paradigm. The essential idea of agent-based modeling and simulation (ABMS) is that complex phenomena (such as economic systems and business organizations) can best be represented as systems of relatively simple autonomous agents that follow comparatively simple rules of interaction. These agents are capable of making independent decisions and interactions with the rest of the system. Therefore, multi agent simulation uses multi agent systems to represent the structure of simulation models. A multi agent system is composed of multiple interacting agents capable of achieving their goals, which are beyond their individual abilities, through mutual cooperation.

Multi agent simulation consists of simulated agents that "live" in a simulated—*artificial environment*, and in simulated—*virtual time*. Environment can play an important role as it frames the agents' behaviors and interactions. The difference between multi agent simulation and multi agent systems is that agent environment within multi agent systems is "sensed" by the agent as it is, while agent environment in multi agent simulation is artificially created as a part of the simulation model, thus allowing for much more control over the developed system.

There are several advantages of multi-agent simulation in comparison to system dynamics. System dynamics is restricted only to the macro level, making it incapable of answering questions concerning the relationships between different granulites of the observed system (such as relationships between the microeconomic activities and the macroeconomic aggregates) in a simplified way. ABMS, on the other hand, allows introduction of different layers of observations: individual level, population level and a number of customized intermediate levels. Another drawback of system dynamics is the assumption of homogeneity of entities and space, so that individualism and heterogeneity cannot be represented. However, agents may be capable of adoptive behavior and flexible interaction with other entities. They may change their environment and perceive those changes afterward. This kind of flexible feedback loops is not possible using system dynamics or any other simulation approach.

Multi agent simulation deals with systems with similar characteristics as discrete-event simulation does. Discrete-event simulation can be represented as directed graphs called queuing networks. These graphs can have two types of nodes: servers with or without a queue. Servers are used for processing jobs that pass through the graph. If a server is busy, other jobs must wait in line for the server to complete its task. Even though all queues have their own queuing discipline, no job entity may leave the queue due to its own decision. If the job has branching routes, probability-based decisions are made, because all job entities are the same, without internal structure that could allow them to make their own routing decisions. Besides, discrete-event simulation models do not allow for variable system structure (number of servers, paths and interactions are fixed) while in a multi agent simulation model variable system, structure can be achieved via intelligent job agent processing. Agents provide heterogeneity, where classical modeling approaches were based on homogeneity of behavior.

APPLICATIONS OF MULTI AGENT SIMULATION IN ORGANIZATIONS

Multi agent simulation has provided a number of results in exploration of relationships between different microeconomic factors and macroeconomic aggregates. These results were obtained through exploration of influence of individual behavior patterns on overall dynamics of systems, but also through exploration of influence of an agent's perception of overall system on changes in the agent's own behavior. Applications range from modeling agent behavior in stock markets and supply chains to modeling overall transportation of a country; from modeling overall hospital management tasks to predicting spread of epidemics; from modeling bacterial cell behavior to modeling social insect colony behaviors; from consumer behavior to understanding the fall of ancient civilizations. Recent publications offer a number of multi agent simulation applications in economics, business and related fields. These applications can be categorized into three groups depending on their purpose.

1. **Models for theoretic insight:** This category contains applications used to explain economic phenomena and provide better scientific insight and technical understanding of the interactions and interdependencies of system elements. A simulation model of supply chain management can measure deformation of demand information when this information is transmitted as orders move down the supply chain (Moyaux, Chaib-draa, & D'Amours, 2004). This phenomenon causes amplification of the order variability in a supply chain, which results in higher financial costs due to higher inventory levels and agility reduction. The simulation model, consisting of several companies linked in a supply chain, showed that a high level of collaboration between companies is required to achieve low supply chain costs. Other multi agent simulation models were developed with the purpose to provide insight in collaborative strategies of firms within a specific economic environment. A multi agent simulation model of economic systems and industries was used to simulate different populations of firms (such as reactive and adaptive firms) and study their interactions and the final impact on firm dynamics and industry demography (Guessoum, Reheb, & Durand, 2004).
   Some of the models concentrate on particular strategies and their limitations rather then on firm behavior, in order to explore the prevalence and severity of specific economic phenomena (like oligopolies, externalities, inflation, price wares, etc.). Simulation models of this type were developed for the purpose of determining the limitations of an economic environment during a price war with aggressive competitors (Wu & Durfee, 2004). This model suggested that thoughtfully constraining communication in an information economy can improve system performance in the segment of the market affected by a price war. A multi agent model AGEDASI TOF was employed to study price dynamics in a foreign exchange market. In the model a community of agent dealers derives their price quotes out of qualifying information from various news sources through a system of weights assigned to each information source. Dealer-agent's success was rated using transaction volume for each simulation experiment. The results exhibited interesting patterns in the formation of dealers' opinion trends that were used to establish a set of typical agents' strategies (Streltchenko, Yesha, & Finin, 2003).
   A number of models are aimed to provide insight in microeconomic elements of an economy. Even though most microeconomic models fall in the second category of multi agent models, some of them can be used to enhance stakeholders' perception of the processes they are involved in. These models are developed in close cooperation with the management, investors, workers or other professionals that can benefit from better understanding of the business processes at hand. A number of applications are available in health care (see Heine, Herrler, & Kirn, 2005, for further details), production lines and other segments of business.
2. **Prediction models:** This category of multi agent simulation applications contain more realistic models used for prediction purposes and planning. A number of models deal with scheduling tasks. For example, the multi agent simulation model of scheduling in a supply chain management is intended to help in planning the supply chains, ordering and capacities for a company that participates in several supply chains at the same time. This helps a company to improve its performance, especially if its activities and resources are costly and sensitive to scheduling errors. Developed models consist of agents that represent companies within supply chains. Each agent simulates production orders of its company by sending and receiving messages about purchase and sales orders. Scheduling of distributed supply chain does not require the company to uncover all information to its partner, but only information related to external orders between the company and its partner (Nurmilaakso, 2004). This is why the possibility that a competitor may impair a company's performance through suppliers and customers is minimized. Even though schedules generated during experimentation are not optimal, all of them are feasible.
   A number of models were developed with the purpose of reducing operating costs. A good example is the improvement of cargo routing of Southwest Airlines,

where an additional $10 million of revenues were gained from reduced costs of the personnel and multiple freight handling. Probably the most comprehensible multi agent simulation model developed is the model of 24 hour traffic simulation of the entire country of Switzerland, with about 7.5 million simulated traveling agents (Balmer, Raney, & Nagel, 2004). The goal of the model is to make traffic jam predictions.

3. **Environment models and self-organizing models:** This category of models is primarily focused on the simulated environment in which agents operate. Simulated environments can be used as test beds for implementation and testing of ideas about alternative solutions for more or less treatable problems. These electronic laboratories can be used to test assumptions about individual agents, their behaviors and interactions in great detail. If agents change not only their behavior, but also their properties as a response to the simulated environment, then the model represents a self-organizing system. Self-organizing capabilities of agents rely heavily on other agents and the simulated environment.

   A good example of benefits from this type of multi agent simulation model is simulation of the behavior of different work teams engaged in the development of a complex project. Due to a large number of factors such as individual characteristics, social characteristic and temporal and economic costs, finding an appropriate team that can successfully tackle all of the project tasks is a complex problem. The developed simulation model (Moreno, Valls, & Marin, 2003) consists of different classes of agents, some of which represent workers and managers, while others take care of assessments of teams and presentation of statistical data. The type of generated teams depends on the properties of tasks required by the project at hand. The system is able to generate information about expected cost of the project, expected number of days for project completion and expected rate of failure for each proposed team.

   Although numerous multi agent simulations in the field of financial markets were carried out for the purpose of theoretical analysis of dealer strategies, they are mostly used as test beds. These models can be used to test and predict outcomes of specific situations on the markets, or to test the actual multi agent systems designed for business applications, like e-commerce applications. An example of the first type of multi agent simulation model is MAFiMSi (Streltchenko et al., 2003), designed to be used for observation of the consequences of a range of investor behaviors. It can also be used to test different solutions for financial market automation, especially decision making tasks, as these are not fully automated yet. An example of the second case is a specially developed scenario that resulted in a simulated environment for testing multi agent systems intended for supply chain management. TAC SCM (Sadeh et al., 2003) is a simulated computer manufacturing scenario in which software agents tackle complex problems in supply chain management. In this model, agents represent personal computer manufacturers that compete in markets for components (supplies) as buyers and in markets for finished goods as sellers, with the purpose of maximizing profits over a simulated year. Agents have to solve a very complex combinatorial optimization problem, deciding which supplies to purchase and how to allocate the existing resources optimally. Moreover, they have to act in the face of tremendous uncertainty regarding the behavior of suppliers, clients, and competitor agents.

FUTURE TRENDS

Multi agent simulation proves to be the appropriate modeling paradigm for complex business systems as well as for various microeconomic and macroeconomic models. A number of macroeconomic problems and phenomena previously studied using mainly mathematical models are being studied using computer models that support a high number of elements and their interactions. Multi agent simulation models exhibit new advantages as they are computationally efficient and capable of providing significant answers about the behavior of these complex systems.

With the new software toolkits that replace coding with an intuitive visual interface and onscreen menus, this approach to simulation of large scale systems is rapidly becoming accessible to an ever wider range of professionals and scientists. Multi agent simulation models may provide insight in the emergent phenomena, for example, in unforeseen patterns or global behaviors (Holland, 1998) which are not derivable from properties of its constituents. Multi agent simulation will also provide required test beds for emerging technologies and services in the field of microeconomics, as well as enhancing further developments in electronic commerce and overall electronic business.

Some authors also suggest the need for modeling the "internal environment" of the agents in order to mimic the processing capacity and the cognitive structures of the social subjects in the model (Cioffi-Revilla et al., 2004). In this way different patterns of social factors' behavior that may not necessarily be economically efficient and optimum can be simulated. Emotions, for instance, have a great impact on the customer behavior, making it inefficient and theoretically unpredictable. Other applications in economics include simulation of fiscal systems and exchange rate policies,

introduction of government programs, and others. Applications in related fields (like health care, traffic, military, cargo routing, etc.) have important economic aspects.

CONCLUSION

Multi agent simulation paradigms for modeling economic systems combine traditional modeling approaches and agent-based systems, thus providing the capabilities that have not been previously possible. This novel approach offers a number of advantages that distinguish it from discrete-event simulation, system dynamics, queuing networks and Petri nets. The potential provided by multi agent simulation is used by a growing number of professionals from a number of different application areas. Applications include modeling for the purpose of theoretic insight, for predicting behavior of highly complex systems, as well as for creating self-organizing models and environment models for the purpose of testing new technologies and ideas. Problem specification is being less tied to user's knowledge of simulation tools or constraints of traditional modeling approaches, which represents a great promise for many interesting developments and opportunities in the future.

REFERENCES

Balmer, M., Raney, B., & Nagel, K. (2004). Large-scale multi-agent simulations for transportation applications. *Intelligent Transportation Systems Journal, 8*, 1-17.

Cioffi-Revilla, C., Paus, S.M., Luke, S., Olds, J.L., & Thomas. J. (2004). Mnemonic structure and sociality: A computational agent-based simulation model. In *Proceedings of the Conference on Collective Intentionality IV,* Siena, Italy, (pp. 1-11).

Gilbert, N., & Conte, R. (Eds.). (1995). *Artificial societies: The computer simulation of social life*. London: UCL Press.

Guessoum, Z., Rejeb, L., & Durand, R. (2004). Using adaptive multi-agent systems to simulate economic models. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2004).* New York: IEEE Computer Society.

Heine, C., Herrler, R., & Kirn, S. (2005). ADAPT@Agent. Hospital: Agent-based optimization & management of clinical processes. *International Journal of Intelligent Information Technologies (IJIIT), 1*(1), 30-48.

Holland, J.H. (1998). *Emergence*. Reading, MA: Helix Boox/Addison-Wesley.

Kohler, T.A., Gumerman, G.J., & Reynolds, R.G. (2005, July). Simulating ancient societies. *Scientific American*.

Klügl, F. (2001). *Multi-agent simulation–concept, tools, application*. Munich, Germany: Addison-Wesley.

Klügl, F., Oechslein, C., Puppe, F., & Dornhaus, A. (2004). Multi-agent modeling in comparison to standard modeling. *Simulation News Europe, 40*, 3-9.

Moreno, A., Valls, A., & Marin, M. (2003). Multi-agent simulation of work teams. In *Proceedings of the 3rd International Central and Eastern European Conference on Multi-agent Systems (CEEMAS03),* (pp. 51-60). Berlin: Springer-Verlag.

Moss, S., & Davidsson, P. (2001). *Multi-agent-based simulation*. Berlin: Springer-Verlag.

Moyaux, T., Chaib-draa, B., & D'Amours, S. (2004). Multi-agent simulation of collaborative strategies in a supply chain. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2004).* New York: IEEE Computer Society.

Nurmilaakso, J.-M. (2004). Supply chain scheduling using distributed parallel simulation. *Journal of Manufacturing Technology Management, 15*(8), 756-770.

Sadeh, N., Arunachalam, R., Eriksson, J., Finne, N., & Janson, S. (2003). TAC-03–a supply-chain trading competition. *AI Magazine, 24*(1), 92-94.

Seila, A.F., Ceric, V., & Tadikamalla, T. (2003). *Applied simulation modeling*. Thomson-Brooks/Cole.

Streltchenko, O., Yesha, Y., & Finin, T. (2003). Multi-agent simulation of financial markets. In O. Streltchenko, T. Finin, & Y. Yesha (Eds.), *Formal modeling in electric commerce*. Berlin: Springer-Verlag.

Wu, J., & Durfee, E.H. (2004). The impact of communication costs and limitations on price wars in an information economy. In *Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multi-agent Systems (AAMAS 2004).* New York: IEEE Computer Society.

Zeigler, B.P. (1976). *Theory of modeling and simulation*. New York: John Wiley & Sons.

KEY TERMS

ABMS: Abbreviation for agent-based modeling and simulation; a synonym for multi agent simulation.

Artificial Environment: A model of the environment where the simulation model is operating. Environment model

is completely controllable by the modeler. Particular environment models are highly relevant when modeling adaptive elements of the system or when using adaptive capabilities of the agents contained within the model.

Discrete-Event Simulation: A type of simulation where the simulation mechanism advances the simulation clock to discrete points in time. Time advancement can be round-based (simulation clock is advanced for a constant number of time units for each round of the simulation) or step-based (simulation clock is advanced to the time stamp of the next event in the event queue).

Experimental Frame: Establishes the set of experiments for which the model is valid. It has to be determined in early stages of model development.

Intelligent Software Agent: A system situated within a part of the environment that senses this environment and acts on it over time in pursuit of its own agenda.

Multi Agent Simulation: A simulation modeling paradigm that uses software agents to represent the entities of the modeled system that interact with each other and with the virtual environment. These interactions are used to model dynamics in functioning and structure of the real system.

Multi Agent System: A system consisting of a number of agents that interact with each other through communication, thus allowing them to achieve goals that are beyond their individual capabilities.

Simulation Modeling: An established method in science and industry used to map a part of reality in sufficient detail using a model. Developed model should be able to answer questions directed to the real system without disturbing functioning of the real system.

System Dynamics: A continuous simulation of systems exhibiting feedback loops. The feedbacks can either intensify activities of the system (positive feedback) or slow them down and stabilize the system (negative feedback).

Virtual Time: Denotes a time advancement paradigm used to handle the course of events within the simulation model. Event-based simulations use event queues that allow the simulation time to advance to the time stamp of the next event. In this way, the time scale can be stretched or compressed, depending on the needs of the model.

Multiagent Systems in the Web

Hércules Antonio do Prado
Brazilian Enterprise for Agricultural Research and Catholic University of Brasília, Brazil

Aluizio Haendchen Filho
Anglo-Americano College, Brazil

Míriam Sayão
Pontifical Catholic University of Rio Grande do Sul, Brazil

Edilson Ferneda
Catholic University of Brasília, Brazil

INTRODUCTION

The rapid evolution of Internet has opened a new era in the distributed systems scenery: the bigger part of the information systems currently developed is focused in Web applications. Typically, the information resources in Web systems are dynamic, distributed, and heterogeneous. Since these computing environments are opened, information resources can be connected or disconnected at any time. This ubiquity of Web and its distributed and interconnected characteristics represent a natural field for multiagent systems (MAS), spreading this kind of application. Software agents can dynamically discover, orchestrate, and compose services, check activities, run business processes, and integrate heterogeneous applications.

Most of the large organizations adopt heterogeneous and complex information systems. These systems must coordinate their applications in order to provide efficient support to business processes and consistent information management. Unfortunately, the operational software underlying these systems usually does not handle multitask distributed heterogeneous applications. Currently, enterprises are strongly interested in the strategic advantages of adopting distributed infrastructures that are designed to be dynamic, flexible, adaptable, and interoperable. In this context, the demand for agent-based applications has increased, opening new types of applications that include e-commerce, Web services, knowledge management, semantic Web, and information systems in general. Interesting solutions to B2B (business to business), e-business, and also applications that require interoperability based on knowledge about applications and business processes, will definitely benefit from the MAS technology. Also, intelligent information agents are regarded as one of the most promising areas for applying agents' technology. Intelligent information agents act in fields like collaborative systems on Internet, knowledge discovery from heterogeneous sources, systems for intelligent management of information, and so on. The Web can also be seen as a big distributed database having XML (extensible markup language) and its extensions or modifications as an underlying data model.

In this context, the MAS development has received support from new tools in order to make it easier for the developer to cope with specific requirements for Web architectures. It is accepted that these improvements in the technology, mainly by the new tools that are becoming available, will lead MAS technology to be explored in its full potential. So, we can state that the application domain of MAS is going to be strongly enlarged, defining a turning point in the systems development activity.

In this chapter, we provide an overview on MAS technology, discuss how this technology is impacting the Web context, and provide a sound description of the concepts that are relevant to the application developers and target users.

BACKGROUND

The adoption of agents' technology in distributed and concurrent systems derives from the idea that cooperation, flexibility, and intelligence of agents can contribute significantly to improve the overall performance and quality of information systems. An agent-based system is composed of autonomous computational entities that possess individual capabilities and goals, and can be grouped to work cooperatively, aiming to reach the system objective. There is not a universally accepted definition to the term "software agent." Wooldridge *et al*. (Wooldridge, Jennings, & Kinny, 1999) explains that this difficulty is partially due to the fact that, for each different application domain, the properties assigned to the agent concept take several important levels, therefore, it is possible to find many types of software agents with different

characteristics, such as mobility, autonomy, collaboration, persistence, and intelligence. The agents' behavior depends on, and is affected by, their properties. Based on previous studies carried by Kendall *et al.* (Kendall, Krishna, Pathak, & Suresh, 1999), the OMG (object management group) (2000), and Garcia *et al*. (Garcia, Silva, & Lucena, 2001) describe the following properties for software agents:

1. Interaction: an agent communicates to the environment and to other agents by means of *sensors* and *actuators*.
2. Adaptation: an agent must self adapt its state and behavior according to the environmental conditions.
3. Autonomy: an agent possesses its own control thread and can accept or refuse a request. Autonomy is understood as the agent capabability to perform its activities independently from the human intervention.
4. Capacity to learn: an agent can learn based on previous experiences when interacting with its environment.
5. Mobility: an agent must be able to transfer itself from one environment to another in order to achieve its goals.
6. Collaboration: an agent can cooperate with other agents in order to achieve its objectives and the system objectives.

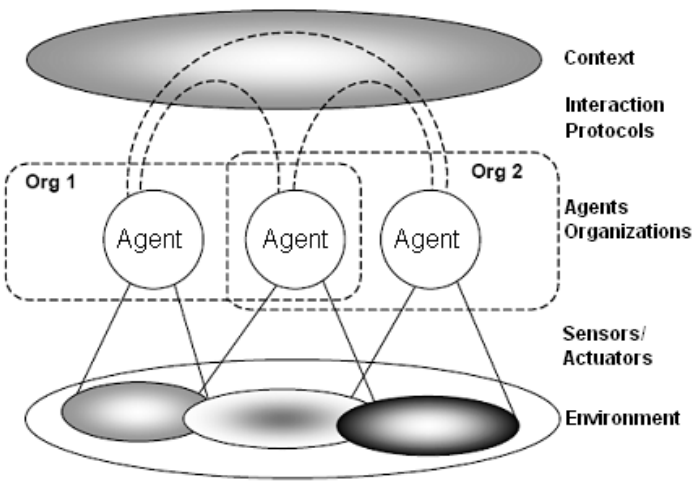
According to OMG (2000), autonomy, interaction, and adaptation can be considered fundamental properties of software agents, while capacity to learn, mobility, and collaboration are not strictly required properties to characterize agents. There is a consensus in the literature that autonomy is the key property for an agent. Agents must present, at least in some extension, independence. They are not completely preprogrammed, but can take decisions based on information from other agents or from the environment.

Ferber (2000) argues that an agent can be defined from the agency characteristics as a physical or a virtual entity, with the following properties or abilities:

- is capable to act in an environment;
- can communicate to other agents;
- is driven by a set of tendencies as individual goals;
- possesses its own resources;
- is capable to perceive its environment (in a certain extension);
- possesses only a partial representation of the environment;
- possesses abilities and can offer services;
- can be able to replicate itself;
- its behavior is driven by its goals, considering the amount of resources and abilities available, and depends on its environmental perception and the messages it receives.

By this definition, an agent can *act,* not only *reason*, and the effect of its actions in the environment can affect future decisions. There are a number of different criteria to classify agents in the literature. For example, Jennings *et al*. (Jennings, Atighetchi, Vincent, & Lesser, 1996) state that to act autonomously, agents must present the following abilities: perception, capacity to belief-based reasoning, to decision taking, and to plan, and ability to execute these plans, including message carrying. Jennings *et al.* (1996) categorize agents according to their levels to solve problems in the following types:

Figure 1. Multiagent system structure



1. *Reactive*: react to modifications in the environment or to messages from other agents. Do not exhibit capacity to reason on their intentions, only reacting to rules and stereotyped plans. Their actions comprise to update the fact base and to send messages to the environment or to other agents.
2. *Intentional*: they are able to reason on their intentions and beliefs, to create and execute action plans. They can reason on the plans, schedule actions, detect conflicts among plans, and also revise plans.
3. *Social*: intentional agents are considered social agents if they keep models from other agents over whom they reason to decide. This kind of agent refers to the idea that multiagent systems can be organized to simulate the behavior of social and organizational structures present in the real world.

Jennings *et al*. (1996) define a *multiagent system* (see Figure 1) as "a low-coupled network of problem solvers that work together to solve problems that are beyond the individual ability or knowledge of each problem solver." Usually, agents grouped in organizations compose multiagent systems.

Designing and developing MAS are not trivial tasks, due to the high level of complexity involved. For this reason, MAS usually are designed and developed using tools or platforms, such as JADE (Bellifemine, Poggi, & Rimassa, 2001) and SOMA (Corradi, Cremonini, & Stefanelli, 1998), among others. These tools provide support for the design, simplifying the development into two abstraction levels. First, in the generic architectural level, such tools abstract the application complexity by providing infrastructure services, as message transport (send/receive, pack/unpack, requests translate), the control over the platform and agents' life cycle, and tasks for services handling (services registry, publication and discovery). Second, in the agents' design level, those tools provide abstract classes that define the hot spots from which specific features of the concrete agents can be implemented. The MAS development has evolved to an increasing reuse of architectures, enabling the developer to focus in implementing the application particularities. The agents life cycle, its internal behavior, and interaction can be modeled by using the conventional UML (unified modeling language) (Booch, Rumbaugh, & Jacobsen, 2000) tools. The life cycle phases of the agents and applications can be designed using the activity diagram, the agent internal behavior can be represented using the state machine diagram, and the interactions of the agents with the architecture and other agents can be modeled by using the sequence diagram. An ACL (agent communication language) or the Blackboard (Ferber, 2000) pattern can be used to design the communication model of the agents. The Blackboard is one of the most widely used patterns in symbolic cognitive multiagent systems. It allows the indirect data communication among independent modules.

MAIN FOCUS OF THE ARTICLE

Due to the recent rising of Web-based technologies, their applications are spreading to domains like e-commerce (electronic commerce) (Rimmel, Clement, & Runte, 1999), B2B (business to business) (Boughaci & Drias, 2005), knowledge management (Jensen *et al.*, 1999), distributed information and database integration, intelligent information systems, and enterprise information systems in general (Adam & Mandau, 2004). E-commerce is the buying and selling of goods and services on the Internet, especially the World Wide Web. Software agents have been used for the individualization and automation of the marketing instruments applied. The use of methods from artificial intelligence enables agents to learn, making possible the automatic optimization of the marketing instruments to satisfy massive individualized needs of the demander. The B2B technology has been used in many business applications, for example, in the logistic support to the buying process (Lim, Park, & Kim, 2005), issuing information about suppliers, existing products in the market, distribution, the buying orders follow-up, payment management, and planning. There exists, nowadays, a considerable research effort to integrate knowledge representation technologies, semantic Web, and agents (McGuinness, & Silva, 2004; Sycara, Paolucci, Ankolekar, & Srinivasan, 2003). The Web semantic community effort to provide a semantic description of services represents a fundamental initiative to enable the larger-scale agents. One of its objectives is to build high-level procedures that can be reusable, and use agents' technologies to facilitate the coordination and dynamic composition of Web services. Decker *et al.* (Decker, Melnik, Harmelen, Fensel, Klein, & Broekstra, 2000) discusses the role of XML in the creation of the semantic Web. The use of agents in Web has also led to significant improvements in fields like information integration, and also those that require knowledge-based interoperability among heterogeneous applications and business processes. XML offers a path to the integration of distributed knowledge in the Internet, and to simplify this integration, collaborative agents have been used (Nodine, Perry, & Unruh, 1998). Integration information using XML has inspired efforts like MIX (mediaton of information using XML) (Baru, Gupta, Ludaescher, Marciano, Papakonstantinou, & Velikhov, 1999), which uses agents to integrate information distributed in multiple sources. The Web is regarded as a distributed database, and XML (and its modifications or extensions) is used as the underlying data model. Klusch (1999) identifies intelligent information agents as one of the most promising agent categories to apply agent's technology. Intelligent information agents act in areas like collaborative systems in the Internet, knowledge discovery

systems from heterogeneous sources, systems for intelligent management of information in corporate *intranets,* or even in the Internet. Enterprise information systems are another promising area to the application of agents technology in the Internet (Fox, Barbucenu, Teigen, 2001). The advances in IT (information technology) are opening opportunities to redesign information systems and processes management. Software agents have been used in conjunction with all the mentioned technologies. They can be designed to carry gathered data and to analyze schedules in different levels, promoting easy task coordination and interoperability in business applications.

The use of agents and MAS in the Internet represents a new generation of applications, based in emerging technologies and in a set of open standards to Web. The combination between agents and Web emerging technologies requires a specific software engineering (Griss & Kessler, 2003) and the evolution of related concepts and standards. The Web standards, established and governed by W3C (Berners-Lee, 1994) affect strongly the development of a wide set of applications and, consequently, the MAS development. The current agent platforms need to be adapted to handle specific Web services requirements and emerging Web technologies (Greenwood, 2005; Odell, 2005). The increasing adoption of Web services and the consequent necessity to establish standards led the W3C to propose the WSA (Web services architecture reference model). This model had its first version published in 2003 (W3C Working Draft 8 August 2003) (Booth, Champion, Ferris, McCabe, Newcomer, & Orchard, 2003), having its last version published in the W3C Working Draft manifesto of November 2004. WSA follows the SOA (services oriented architecture) architectural style, and introduces a set of concepts and abstractions to architectures that uses Web services. Its characteristics demand new concepts to the MAS development, when emerging technologies and open standards for Web are required. Standards used in MAS, like FIPA, conceived in the last decade, are not sufficient to promote the integration of the emerging technologies (FIPA, 2002) with the last established standards for Web.

FUTURE TRENDS

The technologies and systems, running in the Web, described in the previous section, are becoming more and more complex, requiring entities of widely divergent natures, numerous functionalities, and interaction among several devices. The dynamics of requirements changing in the Web are demanding new concepts and a transition from the monolithic architectures, based in passive software components, to more flexible and opened architectures (Garcia, 2005), composed by dynamic and proactive agents. The need for a painless adoption of these technologies induces a demand for new concepts and the internalization of open, flexible, and adaptive architectures. However, the agents technology did not reach its potential, and will not be largely used in business applications until adequate environments to the application development become available (Cowan & Griss, 2002; Curry, Chambers, & Lyons, 2003). A particular point is to understand those issues in the agent technology that make it difficult and/or improve the production of large open systems. Considering this context, we emphasize the following aspects to be approached in the near future in order to improve the agents' adoption:

1. In opposition to the strongly coupled structures used in the current MAS platforms, low-coupled structures should be promoted in a new generation of tools that complies with the new Web standards (Curry *et al*., 2003; Nagappan, Skoczylas, & Sriganesh, 2003).
2. A stronger support to integrate agent technology and the Web technologies using Web services and its open standards are required (FIPA, 2002; Griss *& Kessler,* 2003; Richards, Sabou, Splunter, & Brazier, 2003), in order to facilitate the interoperability among agents and heterogeneous applications; simplify the composition, orchestration, and reuse of distributed services.

SOA is one of the most recent evolutions in distributed computing, and defines an architectural style to build software applications that make use of services in a network like the Web. SOA requires a coordination level much simpler than the traditional architectures, enabling the developer to abstract local and remote distributed services requests. SOA also enables software components, including application functions, objects, and systems processes that are exhibited as services. All functions in SOA are grouped like reusable services: SOA is the contract to services identification, maintaining the rules to access them. All information on requests and answers, exception conditions, and functionalities are defined as part of this interface. The interface contains the necessary information to access the service without knowing its internal design, language, or implementation platform. It receives the requests that arrive in "envelops," that encapsulate the information about the services. In the context of SOA, agents do not communicate directly, there are no explicit messages among them. When an agent needs to request a service, the request is sent to a single server that locates the agent provider to forward the request. This makes the design and implementation simpler, since it is not necessary to know the name and the implementation details of the entity provider of the service. Only the name of the service, and its parameters must be known. In the SOA context, the explicitly message exchanging among agents can be achieved using the Blackboard pattern (Ferber, 2000).

CONCLUSION

New tools for the support of MAS development in the Web context can be considered the most important current trend. The most popular tools available are based in message exchange among agents, requiring a deep knowledge of the internal agents structure by the developers. They are demanded to build an intricate network of messages exchanging to accomplish the development objectives. There are few approaches using SOA in the context of MAS, such as AgentScape (Overeinder & Brazier, 2004) and Cougaar (Helsinger, Lazarus, Wright, & Zinky, 2003). However, none of these are based on a well-known and standardized architecture. The visionary promise of SOA paradigm (MW-4SOC, 2006) is a world of cooperating services, loosely coupled, to enable flexibility in creating dynamic business processes and agile applications that may span organizations and computing platforms and can, nevertheless, adapt quickly and autonomously to changes of requirements or context. Most of the current frameworks and platforms for MAS development adopt communication models associated to the message concept, and are based in an infrastructure characterized by synchronous communication and a strong coupling of components. These characteristics can affect, strongly and negatively, the architecture flexibility and adaptability, increasing the development complexity. New approaches and tools for MAS development are required to take advantage of SOA features, enabling large-scale multiagent systems and its integration with industrial and business applications.

REFERENCES

ADAM, E., & MANDIAU, R. (2004). Design of a MAS into a human organization: Application to an information multiagent system. In *Proceedings of the 5th Agent-Oriented Information Systems* (pp. 21-35), Chicago, IL, USA. Lectures in Artificial Intelligence LNAI 3030. Springer Verlag

BARU, C., GUPTA, A., LUDAESCHER, B., MARCIANO, R., PAPAKONSTANTINOU, Y., & VELIKHOV, P. (1999). XML-based information mediation with MIX. In *Demo Session, ACM-SIGMOD'99,* Philadelphia, PA. Retrieved from http://citeseer.ist.psu.edu/baru99xmlbased.html

BELLIFEMINE, F., POGGI, A., & RIMASSA, G. (2001). JADE: A FIPA2000 compliant agent development environment. In *Proceedings of the Fifth international Conference on Autonomous Agents (AGENTS '01)* (pp. 216-217). New York, NY: ACM Press..

BERNERS-LEE, T. (1994). *World wide Web standards and guidelines*. Retrieved from http://www.w3.org/Consortium/

BOOCH, G., RUMBAUGH, J., & JACOBSON, I. (2000). *UML – Guia do Usuário*. Editora Campus, São Paulo.

BOOTH, D., CHAMPION, M., FERRIS, C., McCABE F., NEWCOMER, E., & ORCHARD, D. (Eds.). (2003). *Web services architecture. W3C Working Draft* August 2003. Retrieved from http://www.w3.org/TR/2003/WD-ws-arch-20030514/

BOUGHACI, D., & DRIAS, H. (2005). An agent-based approach using the ebXML specifications for e-business. In *Proceedings of the 5th International Conference on Information Technology: Coding and Computing (ITCC'05)* - Volume II (pp. 766-767).

CORRADI, A., CREMONINI, M., & STEFANELLI C. (1998). Melding abstractions with mobile agents. In *Proceedings of the Cooperative Information Agents II CIA '1998. Paris, France, Lecture Notes in Artificial Intelligence LNAI* (pp. 37-51). Springer-Verlag.

COWAN, D., & GRISS, M. (2002). *Making software agent technology available to enterprise applications.* Technical Report HP Labs. Retrieved fromk http://www.hpl.hp.com/techreports/2002/HPL-2002-211.html

CURRY, E., CHAMBERS, D., & LYONS, G. (2003). A JMS message transport protocol for the JADE platform. In *Proceedings of the IEEE/WIC International Conference on Intelligent Agent Technology (IAT'03)* (pp. 396-405).

DECKER, S., MELNIK, S., HARMELEN, F. V., FENSEL D., KLEIN M., BROEKSTRA, J., ERDMANN, M., & HORROCKS, I. (2000). The semantic Web: The roles of XML and RDF. *IEEE Internet Computing , 4*(5), 63-73.

FERBER, J. (2000). Multiagent systems: An introduction to distributed artificial intelligence. Addison Wesley.

FOX, M. S., BARBUCEANU, M., & TEIGEN, R. (2001). Agent-oriented supply-chain management. *International Journal of Flexible Manufacturing Systems, 12*(2/3), 165-188.

GARCIA, A. F. (2005). Call for papers of the International Journal of Computer Systems Science and Engineering (CSSE 2005). Special issue on software engineering for multiagent systems. Retrieved from http://www.crlpublishing.co.uk/csse.htm

GARCIA A. F., SILVA, V. T., & LUCENA, C. J. P. (2001). *Engineering multiagent object-oriented software with aspect-oriented programming. Software: Practice & Experience*. Elsevier.

GREENWOOD, D. (2005) *JADE Web service integration gatway (WSIG). JADE AAMAS 2005 Workshop*. Retrieved from jade.tilab.com/doc/tutorials/JADE_WSIG_Guide.pdf

GRISS, M. L.; & KESSLER, R. R. (2003). Achieving the promise of reuse with agent component. In *Proceedings of the Software Engineering for Large-Scale MultiAgent Systems* (pp.139-147). Springer Verlag.

HELSINGER, A., LAZARUS, R., WRIGHT, W., & ZINKY, J. (2003). *Tools and techniques for performance measurement of large distributed multiagent systems*. AAMAS'03. Melbourne, Australia.

JENNINGS, N. R., & WOOLDRIDGE M. J. (1996). Applying agent technology. *Int. Journal of Applied Artificial Intelligence, 9*(4), 351-369.

JENSEN, D., ATIGHETCHI, M., VINCENT, R., & LESSER, V. (1999). Learning quantitative knowledge for multiagent coordination. In *Proceedings of the 16th National Conference on Artificial Intelligence* (AAAI-99), American Association for Artificial Intelligence (pp. 24-31).

KENDALL, E., KRISHNA, P., PATHAK, C., & SURESH C. (1999). A framework for agent systems. In M. Fayadd et al. (Eds.), *Implementing applications frameworks – Object-oriented framewoks at work*. John Wiley & Soons.

KLUSCH, M. (Ed.). (1999). *Intelligent information agents: Agent-based information discovery and management on the Internet*. Berlin: Springer Verlag.

Levine, D., & Dale, J. (2002). *FIPA Services work plan* no. f-in-00050. Retrieved from http://www.fipa.org/docs/input/f-in-00050/f-in-00050.html.

LIM, G. G., PARK, S. H., & KIM J. (2005). B-cart based agent system for B2B EC. In *Proceedings of the Third Asian Simulation Conference, AsianSim* 2004 (pp. 45-57). Jeju Island, Korea, LNCS Volume 3398. Berlin: Springer Verlag.

MCGUINNESS, D. L., & SILVA, P. P. (2004). Explaining answers from the semantic Web: The inference Web approach. In K. Sycara & J. Mylopoulos (Eds.), Web semantics: Science, services and agents on the World Wide Web. *International Semantic Web Conference 2003, 1*(4), 397-413..

MW4SOC. (2006). *Middleware for service-oriented computing. Workshop of the 7th International Middleware Conference 2006*. Retrieved from http://www.dedisys.org/mw4soc/

NAGAPPAN, R., SKOCZYLAS, R., & SRIGANESH, R. P. (2003). *Developing Java Web services*. Indiana: Wiley Publishing Inc.

NODINE, M., PERRY, B., & UNRUH, A. (1998). Experience with the info-sleuth agent architecture. In *Proceedings of the 15th National Conference on Artificial Intelligence, AAAI-98, Workshop on Software Tools for Developing Agents* (pp. 47-60).

ODELL, J. (2005). *Agent-based process management for SOA and WS applications. OMG Workshop on MDA, SOA and Web Services*, Orlando, FL USA. Retrieved from http://www.omg.org/news/meetings/mda-soa-ws/program.pdf

OMG - OBJECT MANAGEMENT GROUP. (2000). *Agent platform special interest group. Agent Technology – Green Paper, version 1.0.*

OVEREINDER, B. J., & BRAZIER, F. (2004). Scalable middleware environment for agent-based Internet applications. *Data Knowledge Engineering, 41*(2-3), 229-245.

RICHARDS, D., SABOU, M., SPLUNTER, S., & BRAZIER, F. (2003). Artificial intelligence: A promised land for Web services. In *Proceedings of the 8thAustralian and New Zealand Intelligent Information Systems Conference (ANZIIS2003),* Macquarie University, Sydney, Australia.

RIMMEL, G., CLEMENT, M., & RUNTE, M. (1999). *Intelligent software agents: Implications for marketing in e-commerce*. Göteborg, Department of Business Administration.

SYCARA, K., PAOLUCCI, M., ANKOLEKAR, A., & SRINIVASAN, N. (2003). Automated discovery, interaction and composition of semantic Web services. *Journal of Web Semantics, 1*(1), 27-46.

WOOLDRIDGE, M., JENNINGS, N., & KINNY, D. (1999). The Gaia methodology for agent-oriented analysis and design. In *Proceedings of the 3rd Int. Conference on Autonomous Agents*, Seatle, WA (pp. 27-42).

KEY TERMS

Agent Organizations: Can be understood as complex entities where a multitude of agents interact, within a structured environment aiming at some global purpose.

B2B (Business to Business): Exchange of products, services, or information between businesses instead of between businesses and consumers.

FIPA (Foundation for Intelligent Physical Agents): A collection of standards that are intended to promote the interoperation of heterogeneous agents and the services that they can represent.

MAS (Multiagent System): A computational system where agents cooperate or compete with others to achieve some individual or collective task.

SOA (Service-Oriented Architecture): A set of components that can be invoked, and whose interface descriptions can be published and discovered.

Software Agent: In a broad sense, and to have a working definition in a context in where the concept definition is not yet established, we assume that a software agent is an artificial agent that operates in a software environment.

UML (Unified Modeling Language): An object-oriented standard modeling language with a rich graphical notation, and comprehensive set of diagrams and elements.

W3C: An industry consortium that seeks to promote standards for the evolution of the Web and interoperability between WWW products by producing specifications and reference software.

Web Services: Services that are made available by developers (OR industry?) for Web users or other Web-connected programs.

WSA (Web Services Architecture): A set of requirements for Web services standard reference architecture.

XML (Extensible Markup Language): A flexible way to create common information formats and share both the format and the data on the World Wide Web, and intranets.

A Multidisciplinary View of Data Quality

Andrew Borchers
Kettering University, USA

INTRODUCTION

This article introduces the concepts of data quality as described in the literature of several disciplines and discusses research results on how individual perceptions of data quality are influenced by different media (in particular World Wide Web vs. print). A search of literature on "data quality" and "media creditability" reveals that researchers in many disciplines are separately studying the subject. These disciplines include accounting, advertising and public relations, information systems, scientific data collection, education, journalism fields, and others. While these threads have developed separately, these streams of research approach similar issues of how people view the quality of information they receive from different sources.

BACKGROUND

Data quality is an emerging area of research fundamental to the field of information systems. Indeed, the efficacy of systems is in large part driven by the quality of the data that they contain. With the Internet revolution, however, there have been fundamental changes in how information is collected and shared that have a potentially great influence on data quality. This challenge is accentuated with the recent move to "user-generated content" as a part of the broader evolution to Web 2.0 (Schwartz, 2007). In addition, younger generations immerse themselves in media more than their parents do. This has led to the label of the "M-generation." A study by the Kaiser Family Foundation and Stanford University finds young people spending on average 6.5 hours per day in media exposure. Increasingly, this exposure comes in multiple media at one time (Azzam, 2006).

However, with such access and participation comes a challenge as stated by Gilster (as cited in Flanigan & Metzger, 2000):

When is a globe spanning information network dangerous? When people make too many assumptions about what they find on it. For while the Internet offers myriad opportunities for learning, an unconsidered view of its contents can be misleading and deceptive.

Further, organizational responses to data quality have been largely ad hoc (Swartz, 2006) with the majority of firms relying on localized, ad hoc approaches to ensuring data quality.

Recent research and seminars underscore the importance of the topic of data quality. Interest in the discipline has spawned the creation of the International Association for Information and Data Quality, several annual conferences (e.g., *www.iqconference.org*), and the *ACM Journal of Data and Information Quality.* Indeed, Total Data Quality Management (TDQM) has evolved as a field of study extending the concepts of Total Quality Management (Radziwill, 2006). Data quality has emerged as a significant research area.

Information systems and journalism practitioners have echoed the importance of data quality for many years. Research by Redman (1998) summarizes the practical implications of poor data quality. He points out the consequences of poor data quality in areas such as decision making, organizational trust, strategic planning and implementation, and customer satisfaction. Redman conducted (1998) detailed studies and found increased costs of 8-12% due to poor data quality. Service organizations can find increased expenses of 40-60% (Redman, 1998). Strong, Lee, and Wang (1997) support the seriousness of this issue in their study of 42 data quality projects in three organizations. Early research by other authors note data quality issues in a number of settings including accounting (Xu, 2000; Kaplan, Krishnan, Padman, & Peters, 1998), airlines, healthcare (Strong et al., 1997), criminal justice (Laudon, 1986), and data warehousing (Ballou, 1999).

As for a formal definition of data quality, Umar, Karabatis, Ness, Horowitz, and Elmagardmid (1999) quote Redman (1992):

A product, service, or datum X is of higher quality than product, service, or datum Y if X meets customer needs better than Y.

Umar et al. (1999) go on to point out that this definition has been generally accepted and is consistent with the author's work. The definition is somewhat incomplete, however, as it does not delve into the various dimensions of data quality.

A number of authors in the information systems field have gone further than Redman and written conceptual articles on "data quality" (Wand & Wang, 1996; Wang, Reddy, & Kon, 1995; Wang & Strong, 1996; Strong et al., 1997). This work suggests that data quality is a multidimensional

concept (Wand & Wang, 1996) that researchers can view from a number of different perspectives. A panel discussion in 2000 (Lee, Bowen, Funk, Jarke, Madnick and Wand) found five different perspectives to discuss data quality. These included an ontological perspective (specification of a conceptualization) that included different views of reality based on actual observation vs. computer-influenced observations; an architectural perspective, a view that focuses on system infrastructure and its influence on data quality; a context mediation perspective, focusing on communication across space and time; a time-based e-commerce perspective, focusing on the real-time nature of e-commerce; and an information product perspective, focused on data as a product of an organization.

In talking about "data quality," a key beginning is to determine from the literature just what one means by the term. In a definitive work on the topic, Wang and Strong (1996) provide a conceptual framework for data quality. In a way consistent with Redman's (1992) customer perspective, they start by defining "high-quality data as data that is fit for use by data consumers." Using a two-stage survey and sorting process, Wang and Strong (1996) develop a hierarchical framework for data quality that includes four major areas: intrinsic, contextual, representational, and accessibility.

Intrinsic data quality refers to the concept that "data have quality in their own right" (Wang & Strong, 1996). Intrinsic dimensions include accuracy, objectivity, believability, and reputation. Contextual data quality is based on the idea that data does not exist in a vacuum—it is driven by context. Contextual dimensions include relevancy, timeliness, and appropriate amount of data. Representational data quality relates to the "format of the data (concise and consistent representation) and meaning of data (interpretability and ease of understanding)." Accessibility data quality refers to the ease with which one can get to data (Wang & Strong, 1996).

More recent research reinforces many of the concepts presented above. In information systems research, data quality is of particular interest to work on data warehouses and business intelligence. In a recent article noting "BI at age 17" (Martens, 2006), Howard Dresner, author of the term "business intelligence," notes the importance of data quality due to its impact on business process management and operational planning. In studying the maturity of data warehouse projects, Sen, Sinha, and Ramarmuthy (2006) note that data quality is a key determinant. Crie and Micheaus (2006) note that data quality management is a key step in the customer data to value information chain.

Beyond the information systems literature, journalism provides a second relevant body of literature. One of the focus points is on perceptions of Internet credibility (Flanigan & Metzger, 2000; Johnson & Kaye, 1998; Bucy, 2003). The major thrust of this literature is in comparing the Internet to traditional sources with respect to credibility. Note that when referring to "credibility," these authors say "the most consistent dimension of media credibility is believability, but accuracy, trustworthiness, bias and completeness of information are other dimensions commonly used by researchers" (Flanigan & Metzger, 2000, p. 521). Hence, there is a rough correspondence of thinking about "credibility" in the journalism literature to the concept of "intrinsic" and "contextual" data quality in the information systems literature. One author in this field (Bucy, 2003) goes on to differentiate "media" credibility from "source" credibility and suggests that researchers have viewed these two forms of credibility as being separate areas of research.

A third field has contributed to the same discussion, namely, advertising and public relations research. Working on a variety of topics, researchers have asked the question: "What impact does media credibility have in [an] organization's advertising and public relations efforts?" Huh, DeLorme, and Reid (2004) studied media credibility in the context of direct-to-consumer prescription drug advertising. They examined consumer perceptions of credibility based on age and media. Greary (2005) studied the impact on the public relations field of declining media credibility, reported in 2004 to be at a 30-year low. Finally, Cable and Yu (2006) studied job seekers and their organizational image beliefs of potential employers. In their work, they considered three different recruitment media and found media richness to be associated with job seekers' image beliefs.

The concept of data quality also appears in disciplines such as accounting and finance. With the passing of Sarbanes-Oxley legislation and an increased focus on the accuracy financial reporting, new standards focused on data quality are emerging. Clark (2006) reports on the adoption of corporate action standards and points out that data quality is a significant concern. Schwarzkopf (2007) examines source credibility and investors' attitudes toward financial and non-financial performance measures. Interestingly, he noted no difference between more and less experienced investors. He did note, however, that source credibility was most important to investors when viewing financial estimates compared to non-financial performance measures.

In yet another discipline, that of scientific data collection, similar dimensions appear. Radziwill (2006) quotes Loshin (2001) in dividing data quality into four areas: data models, data values, information domains, and data presentation. Within each of these four areas, Loshin gives further dimensions that are quite similar to Wang and Strong's (1996) work. In a similar fashion, Radziwill (2006) also quotes Graefe (2003) in describing data quality criteria in the context of decision process.

It is interesting to note how authors working in multiple disciplines have chosen many of the same dimensions in speaking about data quality. Table 1 summarizes the dimensions these authors have identified using Wang and Strong's (1996) framework.

Table 1. Dimensions of data quality in multiple disciplines

Discipline\ Dimension	Intrinsic Data Quality	Accessibility Data Quality	Contextual Data Quality	Representational Data Quality
Information Systems (Wang, 1997)	Accuracy, objectivity, believability, reputation	Accessibility, access security	Relevancy, value added, timeliness, completeness, amount	Interoperability, ease of understanding, concise representation, consistent representation
Scientific Data Products (Graefe, 2003, quoted in Radziwill, 2006)	Credibility, validity	Availability,	Novelty, relevance, pre-decision availability, validity, information value	Interpretability
Scientific Data Products (Loshin, 2001, quoted in Radziwill, 2006)	Precision, granularity, identifiability, accuracy,	Obtainability, stewardship, ubiquity	Comprehensiveness, robustness, relevance, completeness, currency/timeliness, appropriateness	Clarity, flexibility, essentiality, homogeneity, naturalness, simplicity, semantic consistency, structural consistency, consistency, correct interpretation, agreement of usage
Journalism (Flannigan, 2000)	Believability, accuracy, bias, trustworthy		Completeness	

COMPARISON OF RESEARCH FINDINGS

As noted, many researchers have undertaken work on data quality and media differences over the years. In this section the author will focus on work by Klein (1999, 2001), Flanigan and Metzger (2000), Bucy (2003), and Borchers (2003). All four authors have examined data quality in a similar way, focusing on perceived differences based on media (such as print vs. Internet). In addition, Flanigan and Metzger (2000) examine whether Internet users verify what they find. Borchers (2003) extends the discussion by examining the effect of personal involvement in the topic. Bucy (2003) looks at another aspect in studying the synergy effects of multiple media (TV and Web).

Klein (1999, 2001) has studied perceptions of data quality by surveying a sample of approximately 70 graduate business students conducting class projects. In one early study, Klein (1999) found Web-based material to be more timely, but less believable and of lower reputation, accuracy, and objectivity than printed material. In a more formal result, Klein (2001) found that her subjects perceived traditional text sources to be more accurate, objective, and to have higher reputation and representational consistency. She found Internet sources to be stronger in timeliness and appropriate amount. It would be interesting to repeat her work given the transition from Web 1.0's publication focus to Web 2.0's user collaboration focus (Schwartz, 2007).

Flanigan and Metzger's work (2000) focuses on three areas. First, he looks at the perceived credibility of television, newspapers, radio, and magazines compared to the Internet. The major finding, unlike Klein, is that there is little difference in credibility between media. Second, Flanigan and Metzger (2000) look at the extent to which Internet users verify what they receive. Here he finds that few Web users verify the information they receive. Those with limited Internet experience verify less than those with more experience. Third, and most important to this discussion, Flanigan and Metzger (2000) look at whether perceived credibility varies depending on the type of information viewers are seeking. They cite Gunther in suggesting that "greater involvement with the message results in, first, a wider latitude of rejection."

Bucy (2003) studied the credibility of online and broadcast news sources among younger and older viewers. The study sought to determine if cross-platform media use (that is online and broadcast) had a greater impact on audience perception than exposure to either source alone. Working with 167 subjects from young adults (ages 18-25) and older adults (ages 26-80), Bucy (2003) found significant synergy effects for older adults who viewed both TV and the Web. Between both age groups, credibility between TV and the Web were similar. Notably, older adults had significantly lower credibility scores in general compared to younger adults.

Borchers (2003) considered the literature cited above and examined a number of interesting questions. In keeping with Klein (2001) and Flanigan and Metzger (2000), he examined how people perceive Web-based material compared to printed material, considering dimensions such as "timely," "believable," "reputation," "accuracy," or "objectivity." Second, Borchers (2003) studied whether individuals with personal

involvement in a topic (e.g., cancer) are better discriminators of data quality than those who are not involved with a topic. Finally, Borchers (2003) explores whether women—given their role as healthcare acquirers (Bates & Gawande, 2000; Looker & Stichler, 2001)—are better discriminators of data quality than men on health-related topics such as cancer. This work provides an interesting point to compare with work by Huh et al. (2004) on direct to consumer advertising of drugs.

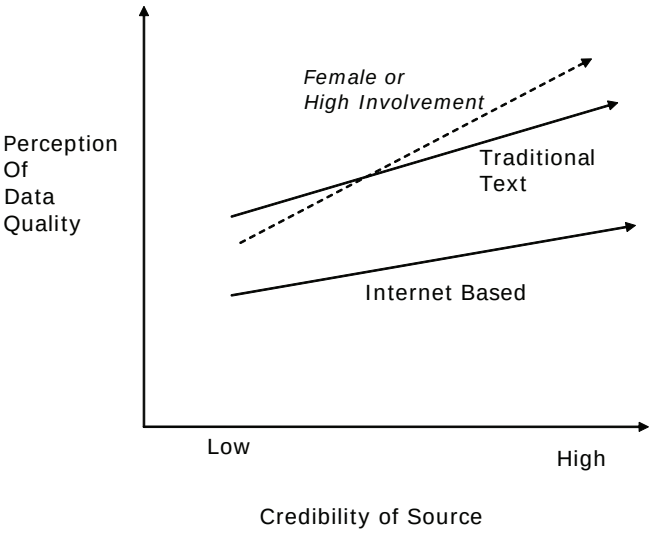
Figure 1 demonstrates what Borchers (2003) hoped to find. H0, his initial hypothesis, is that the perception of low creditable sources is significantly less than high creditable sources. Hence, the two lines for Internet-based and print-based text should have a positive slope. H1 suggests a significant gap between the lines for Internet-based sources and text-based sources on the timeliness, believability, reputation, accuracy, and objectivity dimensions. Borchers (2003) based this assertion on prior literature by Klein (1999). H2 suggests that the slope of the lines should vary based on one's personal involvement in cancer. This is to say, persons with high personal involvement in cancer should be better discriminators of data quality. Finally, H3 suggests that women are better able to differentiate credible from non-credible sources. Hence, the slope of the lines should vary based on gender.

Borchers (2003) studied 127 subjects on their perception of information on cancer based on exposure to Internet and print media. He drew subjects from mid-career students in MBA and MSIS classes at a Midwestern University. His sample was strongly multicultural, with significant U.S., Indian, and Chinese representation. Borchers (2003) randomly assigned subjects to one of four groups. These four groups were shown cancer information based on two sources of information presented in two different formats. One source was a Web site of a highly credible national cancer organization. The second source of cancer information was a Web site of low creditability, a site that touted alternative medical treatments. The third and fourth sources were identical to the first two, with the exception that the researcher presented information in printed form by way of a color document. Borchers (2003) then asked subjects about their perceptions of the data they viewed using Wang's intrinsic data quality dimensions (accuracy, objectivity, believability, and reputation), as well as contextual dimensions (timeliness, relevancy, and appropriate amount of information) and ease of use. Further, subjects were asked about their personal and family experience with cancer, as well as demographic questions (gender, age, and country of birth).

After Borchers (2003) collected data, he tested the dimensions using Cronbach's alpha to be sure they were reliable. Each dimension had a value of .8 or higher. He then generated a second set of statistics using a univariate ANOVA procedure to test each of the research hypotheses. H0 was tested for all eight measured data quality dimensions using the source reputation (high or low) and media (print or WWW) as fixed factors. In testing H2 and H3, he added cancer involvement or gender as random factors. The hypotheses were tested by looking at the product term for source reputation and cancer involvement (H2) or gender (H3). Table 2 summarizes Borchers' findings.

With respect to H0, Borchers (2003) found that his experimental design worked reasonably well. Subjects could easily discriminate between the low and high credible sources. In comparing Internet to print sources (H1), he observed no difference. In examining personal involvement (H2), he found that believability and reputation influenced personal involvement, but only for a subset of respondents. The work did not support the gender hypothesis (H3).

Figure 1. Research design (Borchers, 2003)



FUTURE TRENDS

Researchers should perform future work to extend the work of Klein (1999, 2001), Flannigan and Metzger (2000), Bucy (2003), and Borchers (2003), and to integrate newer literature in the field of data quality. First, researchers can address methodological issues. Readers should note that there are differences in research approach between these authors' work and that each work has methodological weaknesses that can be addressed in future research. Instrumentation in particular is a major concern. How can researchers more accurately measure people's perception of data quality and personal involvement in a topic? Researchers can also conduct studies on participants other than college students, as Bucy (2003) has done. Experimental designs such as Borchers' (2003) can be extended to more precisely test hypotheses. Other weaknesses include research designs. In Flanigan and Metzger's (2000) and Klein's (1999, 2001) work, subjects

Table 2. Hypothesis testing results (Borchers, 2003)

Hypothesis	Dimension	F-Ratio	Significance
H0—Initial Difference due to Reputation	Believable, accuracy, reputation, objectivity, and appropriate amount	10.526 to 24.489	.000 to .002
H0—Initial Difference due to Reputation	Timeliness, relevance, ease of use	< 1.4	> .35
H1—WWW Compared to Print	Timeliness Believable Reputation Accuracy Objectivity Appropriate Amount Relevance Ease Of Use	2.587 1.036 .340 .483 1.132 .617 .030 .620	.110 .311 .561 .489 .290 .484 .865 .484
H2—Personal Involvement with Cancer	Timeliness, Believable reputation, accuracy, objectivity, appropriate amount, relevance, ease of use	With all respondents, F-ratio < 4; U.S.-only respondents had significant interaction on believability and reputation	With all respondents, F-ratio > .05; U.S.-only respondents were significant on believability and reputation
H3—Gender	Believable, accuracy, reputation, objectivity, appropriate amount, relevance, ease of use	All < 4	All > .05

were asked to complete a one-time survey that asked about their perceptions in general of credibility of Internet and print sources. Bucy's (2003) work employed a control group, but did not control the material viewed. In Borchers' (2003) work, subjects were randomly placed in groups that saw exactly the same material in both Internet and text formats. In his study, however, cultural differences among subjects confounded the analysis. Each of these authors also uses surveys as a data collection technique, a potential weakness.

Second, early work focused on print and Internet (especially the WWW) as media. Further research could follow Flanigan and Metzger's (2000) and Bucy's (2003) work in other media, such as TV and radio, or emerging media. Third, researchers can extend tests of personal involvement to topics other than cancer. Huh et al.'s work (2004) on consumer perceptions of direct-to-consumer advertising is one promising topic. The concept of "personal involvement" needs to be developed, perhaps turning to the marketing literature for a base. Finally, as the general population becomes more computer literate and Internet savvy, researchers may find that perceptions of data quality between different media converge (or perhaps diverge). As the "M-generation" (Azzam, 2006) matures and computer technology permits users to multitask between different sources, media credibility may take on new importance.

Having noted these limitations and future areas of work, this line of research is important for several reasons. First, the Internet has become a de facto standard source of information for younger generations. Their perceptions of data quality, particularly compared to print, are a key factor in understanding how people will interpret what they see. Second, the question of personal involvement raises important concerns: Do people become more discriminating in evaluating data quality on topics that have significant influence to their lives? Finally, verification (as described by Flanigan & Metzger, 2000) is yet another interesting topic. To what extent do people verify what they read on the Internet or in print? Verification takes on extra importance with the rise of user-generated content (Schwartz, 2007). Future research should address these and other areas.

CONCLUSION

The work of these four researches (Klein, 1999, 2001; Flanigan & Metzger, 2000; Bucy, 2003; Borchers, 2003) seeks to understand how perceived data quality varies by different media. Flanigan and Metzger's (2000) work extends the discussion to verification of information. Bucy (2003) considers age of participants the synergy effects of multiple media. Borchers (2003) extends the case to include personal involvement.

Do people perceive data quality differently depending on the media that the data comes in? Flanigan and Metzger

(2000), Bucy (2003), and Borchers (2003) suggest that media is not a significant factor. This comes in contrast to Klein's work (1999, 2001), which suggested a difference in perceived data quality on five dimensions: accuracy, objectivity, reputation, timeliness, and appropriate amount.

Do Internet users verify what they see on the Internet? Flanigan and Metzger's (2000) work suggests that relatively few do so, and among inexperienced users even fewer verify what they find. This finding should be particularly distressing to academics and practitioners concerned with delivering high-quality data. As Schwartz (2007) notes, the move to user-generated data is a major trend, but one with questionable value: Will students be able to discern data quality in the myriad of Web resources available to them?

Do people become more discriminating of data quality for topics that they are personally involved in? Borchers' (2003) study provides a first look, at least for cancer information, and finds only limited support for this notion. Finally, does gender play a role in one's ability to discriminate between reputable and non-reputable sources of cancer information? Borchers' (2003) work would suggest that this is not so.

Data quality is an important topic that deserves continued research focus. The Internet revolution has fundamentally changed how people share information. How people perceive the quality of the information they view is an essential research topic for the information systems field.

REFERENCES

Azzam, A.M. (2006). A generation immersed in media. *Educational Leadership,* 63(7).

Ballou, D.P. (1999). Enhancing data quality in data warehouse environments. *Communications of the ACM, 42*(1)**.**

Bates, D.W., & Gawande, A.A. (2000). The impact of the Internet on quality measurement. *Health Affairs, 19*(6), 104-114.

Borchers, A.S. (2003). Intrinsic and contextual data quality: The effect of media and personal involvement. In *ERP & data warehousing in organizations: Issues and challenges.* Hershey, PA: IRM Press.

Bucy, E. (2003). Media credibility reconsidered: Synergy effects between on-air and online news. *Journalism and Mass Communications Quarterly, 80*(2).

Cable, D.M., & Yu, K. (2006). Managing job seekers' organizational image beliefs: The role of media richness and media credibility. *Journal of Applied Psychology, 91*(4).

Clark, T. (2006). Insight into actions—adoption of corporate actions standards is increasing, but data quality still is a concern and automation remains elusive. *Wall Street and Technology, 24*(11).

Crie, D., & Micheaus, A. (2006). From customer data to value: What is lacking in the information chain? *Journal of Database Marketing and Customer Strategy Management, 13*(4).

Flanigan, A.J., & Metzger, M.J. (2000). Perceptions of Internet information credibility. *Journalism and Mass Communications Quarterly, 77*(3), 515-540.

Greary, D.L. (2005). The decline of media credibility and its impact on public relations. *Public Relations Quarterly, 50*(3).

Huh, J., DeLorme, D., & Reid, L. (2004). Media credibility and informativeness of direct to consumer prescription drug advertising. *Health Marketing Quarterly*, *21*(3).

Johnson, T.J., & Kaye, B.K. (1998). Cruising is believing? Comparing Internet and traditional sources on media credibility measures. *Journalism and Mass Communications Quarterly, 75*(2), 325-340.

Kaplan, D., Krishnan, R., Padman, R., & Peters, J. (1998). Assessing data quality in accounting information systems. *Communications of the ACM, 41*(2).

Klein, B.D. (1999, October). Information quality and the WWW. *Proceedings of the Applied Business in Technology Conference,* Rochester, MI.

Klein, B.D. (2001). User perceptions of data quality: Internet and traditional text sources. *Journal of Computer Information Systems, 41*(4).

Laudon, K.C. (1986). Data quality and due process in large inter-organizational record systems. *Communications of the ACM, 29*(1).

Looker, P.A., & Stichler, J.F. (2001). Getting to know the women's health care segment. *Marketing Health Services, 21*(3), 33-34.

Martens, C. (2006). BI at age 17. *Computerworld, 40*(43).

Radziwill, N. (2006). Foundations for quality management of scientific data products. *The Quality Management Journal, 13*(2).

Redman, T. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM, 41*(2), 79-83.

Schwartz, E. (2007). User-generated debate. *InfoWorld,* (February 5).

Schwarzkopf, D.L. (2007). Investors' attitudes toward source credibility. *Managerial Auditing Journal, 22*(1).

Sen, A., Sinha, A., & Ramarmuthy, K. (2006). Data warehousing process maturity: An exploratory study of factors

influencing user perceptions. *IEEE Transactions on Engineering Management, 53*(3).

Smith, S.E. (1998). Reliable cancer resources on the Internet. *Information Today, 15*(6), 23, 28+.

Strong, D.M., Lee, Y.W., & Wang, R.Y. (1997). Data quality in context. *Communications of the ACM, 40*(5), 103-110.

Swartz, N. (2006). Ad hoc data quality processes don't cut it. *Information Management Journal, 40*(6).

Umar, A., Karabatis, G., Ness, L., Horowitz, B., & Elmagardmid, A. (1999). Enterprise data quality: A pragmatic approach. *Information Systems Frontiers, 1*(3), 279.

Wand, Y., & Wang, R.Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86-95.

Wang, R., Reddy, M.P., & Kon, H.B. (1995). Towards quality data: An attribute-based approach. *Decision Support Systems, 13*(3/4), 349-372.

Wang, R.Y., & Strong, D.M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33.

Yang, X. (2000, December). Data quality in Internet time, space, and communities. *Proceedings of the 21st International Conference on Information Systems.*

Xu, H. (2000, December). Managing accounting information quality. *Proceedings of the 21st International Conference on Information Systems.*

KEY TERMS

Accessibility Data Quality: An aspect of data quality that refers to the ease with which one can get to data.

Contextual Data Quality: A concept that data does not exist in a vacuum, it is driven by the circumstance in which data is used. Contextual dimensions include relevancy, timeliness, and appropriate amount of data.

Data Quality: A multifaceted concept in information systems research that focuses on the fitness for use of data by consumers. Data quality can be viewed in four categories: intrinsic (accuracy, objectivity, believability, and reputation), contextual (relevancy, timeliness, and appropriate amount of data), representational (format of the data), and accessibility (ease of access).

Internet Credibility: A multi-faceted concept in journalism research that consists of believability, accuracy, trustworthiness, and bias.

Intrinsic Data Quality: A concept that "data have quality in their own right" (Wang & Strong, 1996) including accuracy, objectivity, believability, and reputation dimensions.

Representational Data Quality: A concept that data quality is related to the "format of the data (concise and consistent representation) and meaning of data (interpretability and ease of understanding)" (Wang & Strong 1996).

User-Generated Content: A feature of Web applications that allows participants to add and modify information as contrasted to traditional sources such as publishers or broadcasters.

Multimedia Content Adaptation

David Knight
Brunel University, UK

Marios C Angelides
Brunel University, UK

INTRODUCTION

The previous decade has witnessed a wealth of advancements and trends in the field of communications and subsequently, multimedia access. Four main developments from the last few years have opened up the prospect for ubiquitous multimedia consumption: wireless communications and mobility, standardised multimedia content, interactive versus passive consumption and the Internet and the World Wide Web. While individual and isolated developments have produced modest boosts to this existing state of affairs, their combination and cross-fertilisation have resulted in today's complex but exciting landscape. In particular, we are beginning to see delivery of all types of data for all types of users in all types of conditions (Pereira & Burnett, 2003).

Compression, transport, and multimedia description are examples of individual technologies that are improving all the time. However, the lack of interoperable solutions across these spaces is holding back the deployment of advanced multimedia packaging and distribution applications. To enable transparent access to multimedia content, it is essential to have available not only the description of the content but also a description of its format and of the usage environment in order that content adaptation may be performed to provide the end-user with the best content experience for the content requested with the conditions available (Vetro, 2003).

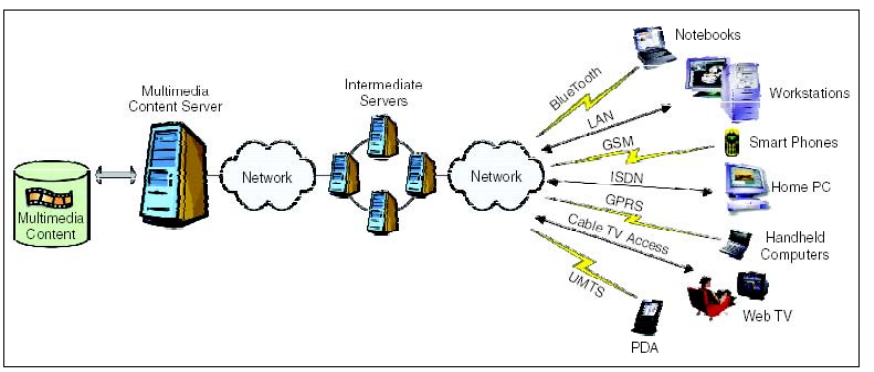
In the following sections, we will look at the background of multimedia content adaptation, why do we require it and why are present solutions not adequate. We then go onto the main focus of the article, which describes the main themes of modern multimedia content adaptation, such as present day work that defines the area and overviews and descriptions of techniques used. We then look at what this research will lead to in the future and what we can expect in years to come. Finally, we conclude this article by reviewing what has been discussed.

BACKGROUND

More and more digital audio-visual content is now available online. Also more access networks are available for the same network different devices (with different resources) that are being introduced in the marketplace. Structured multimedia content (even if that structure is still limited) increasingly needs to be accessed from a diverse set of networks and terminals. The latter range (with increasing diversity) from gigabit Ethernet-connected workstations and Internet-enabled TV sets to mobile video-enabled terminals (Figure 1) (Pereira & Burnett, 2003).

Adaptation is becoming an increasingly important tool for resource and media management in distributed multimedia

Figure 1. Different terminals access multimedia content through different networks



systems. Best-effort scheduling and worst-case reservation of resources are two extreme cases, neither of them well-suited to cope with large-scale, dynamic multimedia systems. The middle course can be met by a system that dynamically adapts its data, resource requirements, and processing components to achieve user satisfaction. Nevertheless, there is no agreement about questions concerning where, when, what and who should adapt (Bormans et al., 2003).

On deploying an adaptation technique, a lot of considerations have to be done with respect to how to realise the mechanism. Principally, it is always useful to make the technique as simple as possible, i.e., not to change too many layers in the application hierarchy. Changes of the system layer or the network layer are usually always quite problematic because deployment is rather difficult. Generally, one cannot say that adaptation technique *X* is the best and *Y* is the worst, as it highly depends on the application area.

The variety of delivery mechanisms to those terminals is also growing and currently these include satellite, radio broadcasting, cable, mobile, and copper using xDSL. At the end of the distribution path are the users, with different devices, preferences, locations, environments, needs, and possibly disabilities.

In addition the processing of the content to provide the best user experience may be performed at one location or distributed over various locations. The candidate locations are: the content server(s), any processing server(s) in the network, and the consumption terminal(s). The choice of the processing location(s) may be determined by several factors: transmission bandwidth, storage and computational capacity, acceptable latency, acceptable costs, and privacy and rights issues (see Figure 2).

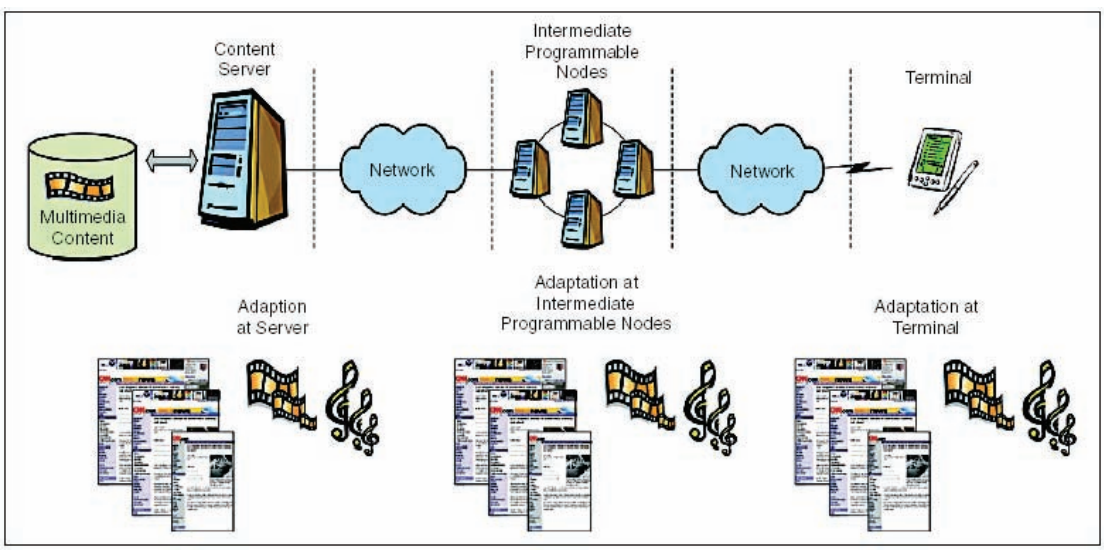
Present adaptation technologies concerning content adaptation mainly focus on the adaptation of text documents. Therefore, one text document will be adapted on demand to the capabilities of different devices or applications. To fulfill this functionality the structure of the content must be separated from its presentation, i.e., the source document is structured using XML (Extensible Markup Language) and then dynamically processed to generate a presentation tailored to the available resources. One possible use case scenario will be to present the same information either on a standard Web browser or a WAP browser.

Efficient adaptation requires that the participating components know from each other and take advantage of adaptation steps done by other components, which needs standardised media, metadata, and communication. Several standardisation bodies (W3C, MPEG, and WAP) have already been established or are currently under development, which have recognised the need to create a framework that facilitates the efficient adaptation of content to the constraints and preferences of the receiving end.

MPEG-7 (ISO/IEC 15938-5:2002) provides tools for content description, whilst capability description and negotiation is provided for with CC/PP (Composite Capabilities/ Preference Profiles, 2003) and UAProf (WAG User Agent Profile, 2001). MPEG-21 (ISO/IEC JTC 1/SC 29/WG 11), the "multimedia framework" includes Digital Item Adaptation (DIA), which enables standard communication of dynamic adaptation of both media resources and meta-data, enabling negotiation of device characteristics and QoS parameters. (Böszörményi et al., 2002)

In this section, the reasons for the need for interoperable and efficient multimedia content adaptation have

Figure 2. Adaptation may be performed at different places



been introduced. A number of standards groups (such as W3C, MPEG and WAP) that facilitate multimedia content adaptation by concentrating on the adaptation of associated XML-type documents have also been mentioned. The next section delves into the different technologies that help make up this exciting field.

MAIN THRUST OF THE CHAPTER

In this section we will look at the main themes found in multimedia content adaptation. We start with a look at a multimedia content adaptation architecture, a discussion on the present state of affairs regarding scalable coding and transcoding, an analysis of the effect the actual location point the adaptation takes place and a brief summary of the relevance of user profiling.

Multimedia Content Adaptation Architecture

The networking access paradigm known as Universal Multimedia Access (UMA) refers to the way in which multimedia data can be accessed by a large number of users/clients to view any desired video stream anytime and from anywhere. In the UMA framework, multimedia information is accessed from the network depending on the following three parameters: channel characteristics, device capabilities, and user preference.

Figure 3 gives an example of different presentations (to suit different capabilities such as formats, devices, networks, and user interests) of the same information.

One option for UMA is to provide different variations of the content with different quality, bit rate, media modality (e.g., audio to text), etc. The problem with this option is that it is not too efficient from the viewpoint of variation generations and storage space. On the other hand, real-time transformation of any content implies some delay for the processing and a lot of computing resources at the server (or proxy server) side. Pursuing either of these two options assumes the use of an adaptation engine. Figure 4 gives a bird's eye-view of such an adaptation engine architecture that is applicable to the adaptation of any type of content.

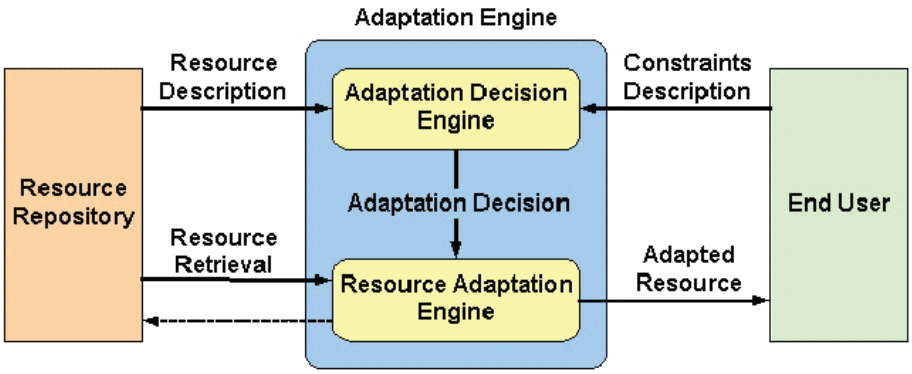
The architecture consists of an adaptation engine that can be located on the server, an intermediate network device such as a gateway, router, or proxy, or even on the client. This engine comprises of two logical engine modules, the adaptation decision engine and the resource adaptation engine. The adaptation decision engine receives the metadata information about the available content (context, format, and adaptation options) from the resource repository and the constraints (terminal and network capabilities, user characteristics, and preferences) from the receiving side. If there are multiple versions of the content pre-stored in the repository and one of these versions matches the constraints, then this version is selected, retrieved, and sent to the end user. However, if the available resource does not match the constraints, but can be adapted, then the adaptation decision engine determines the optimal adaptation for the given constraints and passes this decision to the resource adaptation engine. The resource adaptation engine retrieves the resource from the repository, applies the selected adaptation, and sends the adapted resource to the end user.

Constraints can be grouped into four broad categories: user and natural environment characteristics, terminal capa-

Figure 3. Different presentations of the same information

Workstation/LAN	PC/Dialup	TV Browser	Gray PDA	BW PDA	Text Browser
					"bridge"
38 KB	23 KB	8 KB	4 KB	0.6 KB	0.01 KB
24 bit color	24 bit color	256 colors	4 bit gray	B/W	-
256 x 256	192 x 192	128 x 128	96 x 96	64 x 64	-
22 sec	13.5 sec	4.7 sec	2.4 sec	0.35 sec	0.01 sec

Figure 4. Bird's-eye view of an adaptation engine architecture



bilities, and network characteristics. The terminal and network constraints will set an upper bound on the resources that can be transmitted over the network and rendered by the terminal. Information like the network's maximum bandwidth, delay and jitter, or the terminal's resolution, buffer size, and processing power, will help the adaptation engine determine the optimal version of the resource for the given network and terminal device. As the user is the actual consumer (and judge) of the resource, user-related information, including user preferences, user demographics, usage history and natural environment, is equally important in deciding which resource should be delivered to the terminal.

The adaptation engine needs to have sufficient information about the context and the format of the multimedia resources in order to make a decision whether the resource is suitable for the user or how it should be adapted in order to offer the user the optimal context and quality. The description should therefore include information on the resource type, semantics, available adaptation options, and characteristics of the adapted versions of the resource (Panis et al., 2003).

Scalable Coding

If content is scalable, then adapting content may be done using scalable coding. This removes or alters parts of resources in such a way as to reduce their quality in order to satisfy the receiver's capabilities and needs. Currently available scaling options depend on the coding format to be used. The Holy Grail of scalable video coding is to encode the video once, and then by simply truncating certain layers or bits from the original stream, lower qualities, spatial resolutions, and/or temporal resolutions could be obtained (Vetro, 2003).

Current scalable coding schemes fall short of this goal. MPEG-4 is currently the content representation standard where the widest range of scalability mechanisms is available, notably in terms of data types, granularities, and scalability domains (Pereira & Ebrahimi, 2002).

Transcoding

This is another more complex option that typically refers to transforming the resource from one coding format into another one, i.e., decoding the resource and encoding the resource using another codec (e.g., transcode an MPEG-2 video to an MPEG-1 video). According to Sun et al. (2003), the key design goals of transcoding are to maintain the video quality during the transcoding process and to keep complexity as low as possible. Cavallaro et al. (2003) identify three main approaches to video transcoding: content-blind transcoding, semantic transcoding, and description-based transcoding:

- **Content-blind transcoding** does not perform any semantic analysis of the content prior to conversion. The choice of the output format is determined by network and appliance constraints, independent of the video content (i.e., independent of the way humans perceive visual information). The three main content-blind transcoding categories are spatial conversion, temporal conversion, and colour-depth reduction.
- **Semantic (or intramedia) transcoding** analyses the video content prior to conversion. An example of such analysis is the separation of the video content into two classes of interest, namely foreground and background. Once this separation has been accomplished, the two classes can be coded differently to better accommodate the way humans perceive visual information, given the available network and device capabilities.
- **Description-based (or intermedia) transcoding** transforms the foreground objects extracted through semantic segmentation into quantitative descriptors. These quantitative descriptors are transmitted instead of the video content itself. In this specific case, video is transformed into descriptors so as to produce a textual output from the input video. Such textual output can be

used not only for transcoding, but also for annotating the video content and for translating the visual content into speech for visually impaired users. This transformation is also referred to as cross-media adaptation.

Location of Adaptation

As well as the technologies used, the location used for multimedia content adaptation needs to be addressed. Resource delivery and adaptation can be sender-driven, receiver-driven, or network-driven.

Sender-driven proceeds to adapt resources at the sender/server node depending on the terminal and/or network capabilities received beforehand. After successful adaptation the sender transmits the adapted version of the resource to the receiver. This action requires a serious amount of computational power at the server node and goes at the expense of latency between the receiver's request and the server's delivery.

In contrast, the receiver-driven approach decides what and how to adapt at the terminal side although the real adaptation could take place somewhere else, e.g., on a proxy node. Adaptation directly at the end node could fail due to insufficient capabilities. Additionally network bandwidth will be wasted, too. Nevertheless, adaptation on terminal devices should not be strictly excluded.

The pure network-driven approach is transparent where the network, i.e., the transport system, is responsible for adaptation only. Typical use case scenarios will cover all kind of adaptation approaches described so far, i.e., resource adaptability along the delivery chain, from resource provider to resource consumer. A high-performance server node will provide some kind of pre-processing in order to facilitate easy adaptation along the delivery chain across a wide range of network and terminal devices. Network nodes such as routers or gateways will then perform so-called light-weight adaptations using segment dropping or minor editing techniques whereas proxy nodes could utilise more complex adaptation techniques. Such complex adaptation techniques include not only scaling but also transcoding and cross-media. An adaptive terminal device could perform adjustments due to user and/or usage preferences. The complexity of these adaptations to be done in terminal devices depends on its capabilities, e.g., display resolution, computational power, local storage capacity, and buffer size.

User Profiling

In order for the personalisation and adaptation of multimedia content to take place, the users' preferences, interests, usage, and environment need to be described and modelled. This is a fundamental realisation for the design of any system that aims to aid the users while navigating through large volumes of audio-visual data. The expectation is that by making use of certain aspects of the user model, one can improve the efficacy of the system and further help the user (Kobsa et al., 2001).

This section described the components needed to make up a multimedia content adaptation architecture: transcoding, scalable coding, location of adaptation, and user profiling. The next section discusses the future of multimedia content adaptation by looking at UMA, transcoding, and scalability, specifically.

FUTURE TRENDS

The major problem for multimedia content adaptation is to fix the mismatch between the content formats, the conditions of transmission networks, and the capability of receiving terminals. A mechanism for adaptation needs to be created for this purpose.

Scalable coding and transcoding are both assisting in this. It can be seen that scalable coding and transcoding should not be viewed as opposing or competing technologies. Instead, they are technologies that meet different needs regarding multimedia content adaptation and it is likely that they will coexist.

Looking to the future of video transcoding, there are still quite a number of topics that require further study. One problem is finding an optimal transcoding strategy. Given several transcoding operations that would satisfy given constraints, a means for deciding the best one in a dynamic way has yet to be determined. Another topic is the transcoding of encrypted bit streams. The problems associated with the transcoding of encrypted bit streams include breaches in security by decrypting and re-encrypting within the network, as well as computational issues (Vetro, 2003).

The inherent problem with cross-media (description-based) adaptation is in preserving the intended semantics. What are required are not the blindfolded exchange of media elements and fragments, but their substitution by semantically equivalent alternatives. Unfortunately, current multimedia authoring tools provide little support for producing annotated multimedia presentations. Richly annotated multimedia content, created using document-oriented standards, such as MPEG-7 and MPEG-21 DIA, will help facilitate sophisticated cross-modal adaptation in the future.

For the implementation of UMA, "universal," scalable, video-coding techniques are essential components. Enhancements to existing video-coding schemes, such as MPEG-4 FGS (Fine-Granular-Scalability) and entirely new schemes will help drive the UMA ideal. More efficient FGS-encoders, tests on the visual impact of variability and more improved error resilient techniques are improvements that can be made to scalable coding schemes.

While some technologies such as content scalability and transcoding are fairly well established, there are still vital technologies missing for a complete multimedia content adaptation system vision. Many of these technologies are directly related to particular usage environments. While multimedia adaptation for improved experiences is typically thought of in the context of more constrained environments (e.g., mobile terminals and networks), it is also possible that the content has to be adapted to more sophisticated environments, e.g., with three-dimensional (3-D) capabilities. Whether the adaptation processing is to be performed at the server, at the terminal, or partially at both, is something that may have to be determined case–by-case, depending on such criteria as computational power, bandwidth, interfacing conditions, and privacy issues.

CONCLUSION

The development and use of distributed multimedia applications is growing rapidly. The subsequent desire for multimedia content adaptation is leading to new demands on transcoding, scaling, and, more generally, adaptation technologies. Metadata-based standards, such as MPEG-7 and MPEG-21, which describe the semantics, structure, and the playback environment for multimedia content are breakthroughs in this area because they can assist more intelligent adaptation than has previously been possible.

A prerequisite for efficient adaptation of multimedia information is a careful analysis of the properties of different media types. Video, voice, images, and text require different adaptation algorithms. The complex nature of multimedia makes the adaptation difficult to design and implement. By mixing intelligence that combines the requirements and semantic (content) information with low-level processing, the dream of UMA could be closer than we envision.

REFERENCES

Bormans, J., Gelissen, J., & Perkis, A. (2003). MPEG-21: The 21st century multimedia framework. *IEEE Signal Processing Magazine*, 20(2), 53- 62.

Böszörményi, L., Doller, M., Hellwagner, H., Kosch, H., Libsie, M., & Schojer, P. (2002). Comprehensive Treatment of Adaptation in Distributed Multimedia Systems in the ADMITS Project. ACM International Multimedia Conference, 429-430.

Cavallaro, A., Steiger, O., & Ebrahimi, T. (2003). Semantic segmentation and description for video transcoding. Paper presented at the Proceedings of the 2003 International Conference on Multimedia and Expo, 2003, ICME '03..

Composite Capabilities/Preference Profiles. (2003). Retrieved from the World Wide Web March 2003 at: http://www.w3.org/Mobile/CCPP/

Extensible Markup Language (XML). (2003). 1.0 (3rd Edition). Retrieved from the World Wide Web October 2003 at: www.w3.org/TR/2003/PER-xml-20031030/

Kobsa, A., Koenemann, J., & Pohl, W. (2001). Personalized Hypermedia Presentation Techniques for Improving Online Customer Relationships. *CiteSeer*.

ISO/IEC. (2002). ISO/IEC 15938-5: 2002: Information Technology—Multimedia Content Description Interface—Part 5: Multimedia Description Schemes.

ISO/IEC (2003). ISO/IEC JTC 1/SC 29/WG 11: MPEG-21 Digital Item Adaptation Final Committee Draft. Document N5845, Trondheim, Norway. Retrieved from the World Wide Web July 2003 at: http://www.chiariglione.org/mpeg/working documents.htm#MPEG-21

Panis et al. (2003). Bitstream Syntax Description: A Tool for Multimedia Resource Adaptation within MPEG-21, EURASIP Signal Processing. *Image Communication*, Special Issue on Multimedia Adaptation, 18(8), 721-74

Pereira, F., & Burnett, I. (2003). Universal multimedia experiences for tomorrow. *Signal Processing Magazine*, IEEE, 20(2), 63-73.

Pereira, F., & Ebrahimi, T., (2002). *The MPEG-4 Book*. Englewood Cliffs, NJ: Prentice-Hall.

Sun, H., Vetro, A., & Asai, K. (2003). Resource Adaptation Based on MPEG-21 Usage Environment Descriptions. Proceedings of the IEEE International Conference on Circuits and Systems, 2, 536-539.

Van Beek, P., Smith, J. R., Ebrahimi, T., Suzuki, T., & Askelof, J. (2003). Metadata-driven multimedia access. *Signal Processing Magazine*, IEEE, 20(2), 40-52.

Vetro, A. (2003). Visual Content Processing and Representation, Lecture Notes in Computer Science, (pp. 2849). Heidelberg: Springer-Verlag.

WAG. (2001.) User Agent Profile. Retrieved October 2001 from the World Wide Web at: http://www1.wapforum.org/tech/documents/WAP-248-UAProf-20011020-a.pdf

KEY TERMS

Bit Stream: The actual data stream, which is the transmission of characters at a fixed rate of speed. No stop and start elements are used, and there are no pauses between bits of data in the stream.

Content Scalability: The removal or alteration of certain subsets of the total coded bit stream to satisfy the usage environment, whilst providing a useful representation of the original content.

Cross-Media Adaptation: Conversion of one multimedia format into another one, e.g., video to image or image to text.

Multimedia Content Adaptation: The process of adapting a multimedia resource to the usage environment. The following factors make up this usage environment: users preferences, device, network, natural environment, session mobility, adaptation QoS, and resource adaptability.

Transcoding: The process of changing one multimedia object format into another.

UMA: How users can access the same media resources with different terminal equipment and preferences.

User Modelling: In the context of adaptation, the describing/modelling of the users preferences, interests, usage, and environment.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2051-2057, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Multimedia Information Filtering

Minaz J. Parmar
Brunel University, UK

Marios C Angelides
Brunel University, UK

INTRODUCTION

In the film *Minority Report* (20th Century Fox, 2002), which is set in the near future, there is a scene where a man walks into a department store and is confronted by a holographic shop assistant. The holographic shop assistant recognises the potential customer by iris-recognition technology. The holographic assistant then welcomes the man by his name and starts to inform him of offers and items that he would be interested in based on his past purchases and what other shoppers who have similar tastes have purchased. This example of future personalised shopping assistants that can help a customer find shopping goods is not too far away from becoming reality in some form or another.

Malone, Grant, Turbak, Brobst, and Cohen (1987) introduced three paradigms for information selection, *cognitive*, *economic*, and *social*, based on their work with a system they called the Information Lens. Their definition of cognitive filtering, the approach actually implemented by the Information Lens, is equivalent to the "content filter" defined earlier by Denning, and this approach is now commonly referred to as "content-based" filtering. Their most important contribution was to introduce an alternative approach that they called social (now also more commonly called collaborative) filtering. In social filtering, the representation of a document is based on annotations to that document made by prior readers of the document.

In the 1990s much work was done on collaborative filtering (CF). There were three systems that were considered to be the quintessential recommender systems. The Grouplens project (Miller, Albert, Lam, Konstan, & Riedl, 2003) initially was used for filtering items from the Usenet news domain. This later became the basis of Movielens. The Bellcore Video recommender system (Hill, Stead, Rosenstein, & Furnas, 1995), which recommended video films to users based on what they had rented before, and Ringo (Shardanand & Maes, 1995), which later was published on the Web and marketed as Firefly, used social filtering to recommend movies and music.

BACKGROUND

Filtering multimedia content is an extensive process that involves extracting and modeling semantic and structural information about the content as well as metadata (Angelides, 2003). The problem with multimedia content is that the information presented in any document is multimodal by definition. Attributes of different types of media vary considerably in the way the format of the content is stored and perceived. There is no direct way of correlating the semantic content of a video stream with that of an audio stream unless it is done manually. A content model of the spatial and temporal characteristics of the objects can be used to define the actions the objects take part in. This content model can then be filtered against a user profile to allow granular filtering of the content, allowing for effective ranking and relevancy of the documents.

Filtering has mainly been investigated in the domain of text documents. The user's preferences are used as keywords, which are used by the filters as criteria for separating the textual documents into relevant and irrelevant content. The more positive keywords contained in a document, the more relevant the document becomes. Techniques such as latent semantic indexing have found ways of interpreting the meaning of a word in different contexts to allow accurate filtering of documents using different syntax, but allow the same semantics to be recognised and understood.

Text documents adhere to the standards of the language they are written in. Trying to do the same for AV data streams, you are faced with the problem of identifying the terms in the content itself. The terms are represented as a series of objects that appear in the content, for example, a face in an image file. These terms cannot be directly related to the objects as there is no method of comparison, or if there is, it is complex to unlock. The title of the document and some information might be provided in the file description, but the actions and spatial and temporal characteristics of the objects will not be described to a sufficient level for effective analysis of relevancy.

MAIN THRUST OF ARTICLE

Information-filtering techniques have been applied to several areas including American football (Babaguchi, Kawai, & Kitahashi, 2001), digital television (Marusic & Leban, 2002), Web applications (Kohrs & Merialdo, 2000), and ubiquitous and pervasive device applications (Tseng, Lin, & Smith, 2002).

Filtering multimedia information requires different approaches depending on the domain and use of the information. There are two main types of multimedia information filtering: collaborative and content based. If the user wants a subjective analysis of content in order to find a recommendation based on their individual preference, then they use collaborative filtering, also known as social or community-based filtering. If, on the other hand, they require an objective decision to filter information from a data stream based on their information needs, then they use content-based filtering.

All of the above systems use either collaborative or content-based filtering or a combination of both (hybrid) as the techniques for recommending predictions on candidate objects. There are existing information-filtering models outside these classic techniques such as temperament-based filtering (Lin & McLeod, 2002), which looks at predicting items of interest based on temperament theory. It works on the same principle as social filtering. Unlike social filtering, the users are grouped on temperaments of the users and not on similar item selection.

Content-Based Filtering

Content-based filtering is suited to environments where the user requires items that have certain content features that they prefer. Collaborative filtering is unsuitable in this environment because it offers opinions on items that reflect preferences for that user instead of providing filtering criteria that tries to disseminate preferred content from a data stream based on a user's preference. Personalised video summaries are the perfect domain to use content-based filtering. The reason for this is that a user will be interested in certain content only within any video data stream. For example, when watching a football game, the user may only be interested in goals and free kicks. Therefore, users can state what content features and other viewing requirements they prefer and then filter the footage against those requirements.

The content-based approach to information filtering has its roots in the information retrieval (IR) community and employs many of its techniques. The most prominent example of content-based filtering is the filtering of text objects (e.g., mail messages, newsgroup postings, or Web pages) based on the words contained in their textual representations. Each object, here, text documents, is assigned one or more index terms selected to represent the best meaning of the document. These index terms are searched to locate documents related to queries expressed in words taken from the index language. The assumption underlying this form of filtering is that the "meaning" of objects and queries can be captured in specific words or phrases. A content-based filtering system selects items based on the correlation between the content of the items and the user's preferences as opposed to a collaborative filtering system that chooses items based on the correlation between people with similar preferences (van Meteren & Someren, 2000).

The main problem with content-based filtering is that it does not perform well in domains were the content of items is minimal and the content cannot be analysed easily by automatic methods of content-based retrieval (e.g., ideas and opinions). Users with eclectic tastes or who make ad hoc choices are given bad recommendations based on previous choices. For example, Dad, who usually buys classic rock CDs for himself, purchases a So Solid Crew album for his 12-year-old son. He may start getting recommendations for hardcore garage dance anthems every time he logs in. CF does not suffer this problem as it will rank on other users' recommendations of similar choices. Comparative studies have shown that collaborative-filtering recommender systems on the whole outperform content-based filtering.

Collaborative Filtering

A purely content-based approach to information filtering is limited by the process of content analysis. In some domains, until recently, the items were not amenable to any useful feature extraction with content-based filtering (such as movies, music, restaurants). Even for text documents, the representations capture only certain aspects of the content, and there are many others that would influence a user's experience, for example, in how far it matches the user's taste (Balabanovic, 2000).

Collaborative filtering is an approach to overcome this limitation. The basic concept of CF is to automate social processes such as "word of mouth." In everyday life, people rely on the recommendations from other people either by word of mouth, recommendation letters, and movie and book reviews printed in newspapers. Collaborative filtering systems assist and augment this process and help people in making decisions.

There are two main drawbacks to using collaborative filtering: the sparsity of large user-item databases and the first-rater problem (Rashid et al., 2002). Sparsity is a condition when not enough ratings are available due to an insufficient amount of users or too few ratings per user. An example of sparsity is a travel agent Web site, which has tens of thousands of locations. Any user on the system will not have traveled to even 1% of the locations (possibly thousands of locations). If a nearest-neighbour algorithm is used, the accuracy of any recommendation will be poor as a sufficient amount of peers will not be available in the user-item database. The

first-rater problem is exhibited when a new user is introduced that has not enough ratings. If no ratings have been given for an item or a new user has not expressed enough opinions, choices, or ratings, no predictions can be made due to the insufficient data available or bad recommendations will be made. In contrast, content-based schemes are less sensible to sparsity of ratings and the first-rater problem since the performance for one user relies exclusively on his or user profile and not on the number of users in the system.

Hybrid Filtering

Both content-based and collaborative filtering have disadvantages that decrease the performance and accuracy of the systems that implement them. If these methods are combined, then the drawbacks of one technique can be counteracted by the techniques of the other, and vice versa. There have been various implementations such as the following.

- By making collaborative recommendations, we can use others' experiences as a basis rather than the incomplete and imprecise content-analysis methods.
- By making content recommendations, we can deal with items unseen by others.
- By using the content profile, we make good recommendations to users even if there are no other users similar to them. We can also filter out items.
- We can make collaborative recommendations between users who have not rated any of the same items (as long as they have rated similar items).
- By utilizing group feedback, we potentially require fewer cycles to achieve the same level of personalisation.

User Profiles

In information filtering, a user's needs are translated into preference data files called user profiles. These profiles represent the users' long-term information needs (Kuflik & Shoval, 2000). The main drawbacks of using user profiles are creating a user profile for multiple domains and updating a user profile incrementally. The user profile can be populated by one or more of the following.

- *Explicit profiling*: This type of profiling allows users to let the Web site know directly what they want. Each user entering the site will fill out some kind of online form that asks questions related to a user's preferences (Eirinaki & Vazirgiannis, 2003). The problem with this method is the static nature of the user profile once it has been created. The stored preferences in the user profile cannot take into account the changing user's preferences.
- *Implicit user profiles*: This type of user profiles is created dynamically by tracking the user's behaviour pattern through automatic extraction of user preferences using some sort of software agent, for example, intelligent agents, Web crawlers, and so forth (Eirinaki & Vazirgiannis, 2003). All these usage statistics are correlated into a usage history that is an accurate interaction between the user and the system. This usage history is then analysed to produce a user profile that portrays the user's interests. The user profile can be updated every time the user starts a new session, making implicitly made profiles dynamic. The downside of this method is that the user initially will have to navigate and explore the site before enough data can be generated to produce an accurate profile.
- *Hybrid of implicit and explicit profiling*: The drawbacks of explicit and implicit profiling can be overcome by combining both methods into a hybrid. This allows the strong points of one technique to counteract the shortcomings of the other and vice versa. The hybrid method works by collecting the initial data explicitly using an online form. This explicitly created data is then updated by the implicit tracking method as the user navigates around the site. This is a more efficient method over both pure methods. In some instances, this hybrid method is reversed and the implicit tracking methods are used initially to produce a profile.
- *Stereotype profiling*: This can be achieved by data mining and analysis of usage histories over a period of visits. This provides accurate profiling for existing users with legacy data that is accurate. The disadvantage of this method is that it suffers from the same static nature as explicit profiling as the profile is created from archive data that might be obsolete, and therefore some updating might be necessary. The predefined user stereotype is a content-based user profile that has been created for a virtual user or group of users who have common usage and filtering requirements for consumption of certain material. The stereotyped profile will contain additional information about the stereotyped user such as demographic and social attributes. This additional information is then used to place new users to stereotyped profiles that match similar demographic and social traits. The new user without the need of any implicit or explicit tracking automatically inherits preference information.

FUTURE TRENDS

Content-based filtering in multimedia information filtering has one innate problem that researchers are trying to solve: How can we extract semantics automatically from structural content of the model? In collaborative filtering, the age-old

Table 1. Multimedia information filtering

	Description	Techniques Used	Advantages	Disadvantages	Future R&D
Information Filtering	filtering a dynamic information space using relatively stable user requirements	SDI systems recommender systems	allows user to constantly receive content they are interested in with minimal user effort	does not support ad hoc queries that are dynamic compared to the information space they are searching (information retrieval)	all of the below
Content-Based Filtering	filtering content from a data stream based on extracting content features that have been expressed in a content-based user profile	vector space model probabilistic/inference models latent semantic indexing	objective analysis of large and/or complicated (e.g., multimedia) sources of digital material without much user involvement	1. content dependent 2. hard to introduce serendipitous recommendations as approach suffers from "tunnel vision" effect	extracting semantics from the structure of the content automatically without human intervention
Collaborative Filtering	filtering items based on similarities between target user's collaborative profile and peer users/group	same as above	1. content independent 2. proves more accurate than content-based filtering for most domains of use enables introduction of serendipitous choices	1. sparsity: poor prediction capabilities when new item is introduced to database due to lack of ratings 2. new user: poor recommendations made to new users until they have enough ratings in their profiles for accurate comparison to other users	solving the sparsity and new-user problem finding other types of ratings schemes that do not use comparisons between users tastes (e.g., filtering using users temperament)
Hybrid Filtering	combines two or more filtering techniques	simple or rule based stereotype collaborative content based	to reduce weak points and promote strong points of each of the techniques used	weak points can outweigh strong points if the hybrid is created naively	using hybrid systems in domains where using one technique presents a large disadvantage/problem
User Profiles	log file containing user's preferences for consumption of content	content-based profiles collaborative profiles	1. user does not need to state preferences each and every time they use the system 2. user can maintain and update preferences with minimal effort compared to ad hoc methods	needs frequent updating or user preferences become stagnant	user profiles that are as ubiquitous and pervasive as the devices/systems that use them standardisation
Explicit User Profiles	user manually creates user profile by means of a questionnaire	questionnaires ratings	preference information gathered is usually of high quality	requires a lot of effort from user to update	collecting new user preferences that reduces user effort
Implicit User Profiles	system generates user profile from usage history of interactions between user and content	machine learning algorithms	minimal user effort required easily updateable by automatic methods	initially requires a large amount of interaction between user and content before an accurate profile is created	new machine learning algorithms for better accuracy when creating implicit user profiles
Hybrid User Profiles	combination of user profile techniques used to create a profile	explicit/implicit user profiles	to reduce weak points and promote strong points of each of the techniques used	N/A	finding effective strategies for deployment and use of hybrid profiles

problem of sparsity and the new-user problem are still the biggest hindrances to using this method of filtering. Sparsity is being solved presently by hybrid systems, and it appears that this will be favoured way of dealing with sparsity (Lin & McLeod, 2002). The most promising solutions appear to be collaboratively filtering, standardised content-based profiles, which allow flexibility for systems to use either pure content-based or collaborative filtering, or a hybrid of both interchangeably.

Current work on user profiles focuses on improving creation techniques such as improved machine learning algorithms that create implicit user profiles more rapidly so that they can be more reliable and accurate in a shorter amount of time. For explicit user profiling, there is the work on selecting items that increase the usefulness of initial ratings that we have already discussed. The main way forward here, though, appears to be hybrid user profiles that are initially explicitly created and then implicitly updated.

With the advent of digital television and broadband, consumers will be faced with a deluge of multimedia content available to them at home and at work. What they will require are autonomous, intelligent filtering agents and automated recommender systems that actively filter information from multiple content sources. These personalisation systems can then collaborate to produce ranked lists of recommendations for all purposes of information the user might require. The key to this kind of service is not in the implementation of these systems or the way they are designed, but rather on a standard metadata language that will allow systems to communicate without proprietary restrictions and aid in end-user transparency in the recommendation process.

CONCLUSION

In the coming years, as nearly all communication and information devices become digital, we will see the development of systems that will be able not only to recommend items of interest to us, but will be able to make minor decisions for us based on our everyday needs such as ordering basic shopping groceries or subscribing to entertainment services on an ad hoc basis. What is required is a model of the user that describes the user's preferences for a multitude of characteristics that define the user's information needs. This model can then be used to filter data and recommend information based on this complete view of the user's needs. This has been done for many years with text files using techniques such as content-based and collaborative filtering, but has always been a problem with multimedia as the content is diverse in terms of storage, analysis techniques, and presentation. In recent years, classical techniques used for text filtering have been transferred and used in the area of multimedia information filtering. New developments such as hybrid filtering and improved metadata languages have made filtering multimedia documents more reliable and closer to becoming a real-world application.

REFERENCES

20th Century Fox Pictures. (2002). *Minority report* [Motion picture]. 20th Century Fox Pictures.

Angelides, M. C. (2003). Guest editor's introduction: Multimedia content modelling and personalization. *IEEE Multimedia, 10*(4), 12-15.

Babaguchi, N., Kawai, Y., & Kitahashi, T. (2001). Generation of personalised abstract of sports and video. *IEEE Expo 2001*, 800-803.

Balabanovic, M. (2000). An adaptive Web page recommendation service. *First International Conference on Autonomous Agents*, 378-385.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology (TOIT), 3*(1), 1-27.

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995). Recommending and evaluating choices in a virtual community of use. *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, 194-201.

Kohrs, A., & Merialdo, B. (2000). Using category-based collaborative filtering in the active Web museum. *IEEE Expo 2000*.

Kuflik, T., & Shoval, P. (2000). Generation of user profiles for information filtering: Research agenda. *Proceedings of 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 313-315.

Lin, C., & McLeod, D. (2002). Exploiting and learning human temperaments for customized information recommendation. *Internet and Multimedia Systems and Applications*, 218-223.

Malone, T. W., Grant, K. R., Turbak, F. A., Brobst, S. A., & Cohen, M. D. (1987). Intelligent information sharing systems. *Communications of the ACM, 30*(5), 390-402.

Marusic, B., & Leban, M. (2002). The myTV system: A digital interactive television platform implementation. *IEEE Expo 2002*.

Miller, B. N., Albert, I., Lam, S. K., Konstan, J. A., & Riedl, J. (2003). MovieLens unplugged: Experiences with an occasionally connected recommender system. *Proceedings of ACM 2003 International Conference on Intelligent User Interfaces (IUI'03)*.

Rashid, M., Albert, I., Cosley, D., Lam, S. K., McNee, S. M., Konstan, J. A., et al. (2002). Getting to know you: Learning new user preferences in recommender systems.

Shardanand, U., & Maes, P. (1995). *Social information filtering: Algorithms for automating "word of mouth."* Proceedings of the CHI-95 Conference, Denver, CO.

Tseng, B. L., Lin, C.-Y., & Smith, J. R. (2002). Video summarization and personalization for pervasive mobile devices. *SPIE* (Vol. 4676). San Jose.

van Meteren, R., & Someren, M. (2000). *Using content-based filtering for recommendation.* Retrieved from http://citeseer.nj.nec.com/499652.html

Wyle, M. F., & Frei, H. P. (1989). Retrieving highly dynamic, widely distributed information. In N. J. Belkin & C. J. van Rijsbergen (Eds.), *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 108-115). ACM.

KEY TERMS

Collaborative Filtering: Aims at exploiting preference behaviour and qualities of other persons in speculating about the preferences of a particular individual

Content-Based Filtering: Organizes information based on properties of the object of preference and/or the carrier of information

Hybrid Filtering: A combination of filtering techniques in which the disadvantages of one type of filtering is counteracted by the advantages of another

Information Filtering: Filtering information from a dynamic information space based on a user's long-term information needs

Recommendation: A filtered list of alternatives (items of interest) that support a decision-making process

Recommender Systems: Assist and augment the transfer of recommendations between members of a community

User Profile : A data log representing a model of a user that can be used to ascertain behaviour and taste preferences

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2063-2068,

Multimedia Software Interface Design for Special-Needs Users

Cecilia Sik Lányi
University of Pannonia, Hungary

INTRODUCTION

Most software engineering companies do not develop for special users, because they do not see the potential in this limited market. But 10% of the population worldwide are handicapped. In the United States, 14% of the population are estimated to suffer from a disability. In the population aged over 65, this figure becomes 50%. Disabilities are strongly linked with age, and our societies are facing a growing number of people aged 75 and more, who are more likely to have impairments or disabilities. This group will comprise 14.4% of the population in 2040, compared with 7.5% in 2003—almost a twofold increase (EU Commission, 2003). It is not a simple task to assess the effectiveness of multimedia for all users with disabilities. The question is more complicated if the users have special needs.

This article provides a minimal requirements list that every software engineer, computer scientist, and Web designer should take into account if they develop a new multimodal software or a new Web site with multimedia elements.

BACKGROUND

Universal usability is sometimes tried to meeting the needs of users who are disabled or work in disabling conditions. This important direction is likely to benefit all users. The adaptability needed for users with diverse physical, visual, auditory, or cognitive disabilities is likely to benefit users with differing preferences, tasks, skills, hardware, and so on (Schneiderman, 2003, p. 41).

The present middle-aged user group, now using the computer for work or entertainment, will soon move into old age. It is the time to realize the problem and prepare for the solution. We should keep in mind today what we will experience when we grow old. We should design such a world now that will help us in the future!

A critical component in designing multimedia software is the production of educational programs. Obviously, it is not a simple task to assess the effectiveness of a multimedia teaching system. There are some organizations that published techniques for the evaluation of multimedia teaching software (Sik Lányi, Bacsa, Mátrai, & Kosztyán 2005a; Sik Lányi, Mátrai, Molnár, & Lányi, 2005b; Sik Lányi, 2006). The question is more complicated if the users have special needs. The literature is increasingly attentive to "Design for All" principles (NCSU, 2007). Several conferences run on the topic of how can computers and assistive technology help handicapped people. The most important ones are the following:

- International Conference on Computers Helping People with Special Needs (ICCHP), recent and upcoming meetings in Linz in 2006 and 2008.
- International Conference Series on Disability, Virtual Reality, and Associated Technologies (ICDVRAT), staged in Veszprém, Hungary in 2002; Oxford, UK in 2004; and Esbjerg, Denmark and Maia, Portugal in 2008.
- The **Association for the Advancement of Assistive Technology in Europe (AAATE)**, in Dublin, Ireland in 2003; Lille, France in 2005; and San Sebastian, Spain in 2007; will be staged in Florence, Italy in 2009.

What is Multimedia?

Multimedia refers to the use of computers to present text, graphics, animation, and sound in an integrated way. Long heralded as the future revolution in computing, multimedia applications were, until the mid-1990s, uncommon due to the expensive hardware required. With increases in computer performance and decreases in price, however, multimedia is now commonplace.

The term *multimedia* describes a number of diverse technologies that allow visual and audio media to be combined in new ways for the purpose of communicating. Applications include entertainment, education, and advertising. In recent years, the term multimedia has taken on many diverse meanings for an ever-increasing audience. Some of us have a form of multimedia "narrowcast" through digital cable. Home DVD editing software can be categorized as multimedia, along with the latest generation of mobile phones, which are capable of taking and sending voice annotated photos. The term multimedia will continue to evolve and take on as many new meanings as the technologies and applications it is being used to describe. From our viewpoint, multimedia is a means of communication that combines text with graphics, sound, animation, full-motion video, and so forth—usually

in a highly interactive way, and it also includes the use of the Internet (Sik Lányi, 2006).

What Does the Attributive Noun "Multimodal" Mean?

A software or hardware instrument is multimodal if it uses at least two media elements, and one or both must be time dependent—for example video, sound, and animation files; we also call this mixed media.

The so-called multimodal interaction provides the user with multiple modes of interfacing with a system beyond the traditional keyboard and mouse input/output. The most common such interface combines a visual modality (e.g., a display, keyboard, and mouse) with a voice modality (speech recognition for input, speech synthesis and recorded audio for output). Multimodal user interfaces are part of a research area in human-computer interaction.

Why Is Human Computer Interaction So Important?

Human computer interaction (HCI) is the study of how humans interact with computers and programs (this also used to be called 'man machine interface'). HCI is a discipline concerned with the design, evaluation, and implementation of interactive computing systems for human use, with the study of major phenomena surrounding them. From a computer science perspective, the focus is on interaction and specifically on interaction between human(s) and computer(s).

HCI is also a growing academic discipline. More than a dozen research journals in HCI are compiling practical results and theoretical frameworks to guide designers. These success stories in HCI and user interface design are paralleled and emulated in university courses, but change often comes slowly. The resistance comes from technology-centered researchers who value mathematical formalism more than psychological experimentation (Schneiderman, 2003, p. 71.).

Special-Needs Users

Disability is such a qualitatively difference of a human capability from its normal feature, which might be in-born or if acquired, can develop backwards only very slowly, or can be permanent and irreversible. The kinds of disabilities are: physical impairment, sensory impairment, cognitive impairment, intellectual impairment, and cumulative impairment.

In 1980, the World Health Organization (WHO) published its classification of impairments, disabilities, and handicaps in a document called, "ICIDH (International Classification of Impairments, Disabilities and Handicaps)." In this document, three levels of the impairments were distinguished. During recent years the ICIDH has been considerably revised. One of the main differences between the previous version and ICIDH-2, now called "International Classification of Functioning (ICF)", is that instead of *disability* and *handicapped,* new descriptions have been introduced. The ICF speaks about *activities* and *participation.* This means on one side that some more broad terms have to be used, and on the other that our attention must be focused on the still-available abilities instead of the disabilities (ICIDH, 1999). However different countries may use different terminology, we use 'special-needs user' in this article.

These people have special needs in daily requirements during all their life. Without assistive technology or special devices, they are not able to satisfy all their needs. The basic needs are eating, moving, communicating, and so on. This article deals only with using multimedia software for the above purposes, because most of the information and communication technology is based on it. If the handicapped users are not able to use these software (including the Internet) in average ways and means, their needs are special, so we call them 'special-needs users'.

Design for All, or Universal Design?

The concept of universal design is clear. Wikipedia (2007) gives the following definition of universal design:

"A relatively new paradigm that emerged from 'barrier-free' or 'accessible design' and 'assistive technology.' Barrier free design and assistive technology provide a level of accessibility for people with disabilities but they also often result in separate and stigmatizing solutions, for example, a ramp that leads to a different entry to a building than a main stairway. Universal design strives to be a broad-spectrum solution that helps everyone, not just people with disabilities. Moreover, it recognizes the importance of how things look. For example, while built up handles are a way to make utensils more usable for people with gripping limitations, some companies introduced larger, easy to grip and attractive handles as [a] feature of mass produced utensils. They appeal to a wide range of consumers. Universal design is a part of everyday living and is all around us. The 'undo' command in most software products is a good example."

But the author's opinion is a bit different; sometimes the everyday software products are not good enough and not easy to use by special-needs users.

Is the concept of Design for All (DfA) similar to universal design? Many specialists discuss the proper definition of DfA, but at the time of this writing, there was no consensus about a proper definition, therefore we give one for the sake of this article. DfA means design of products, services, systems (including information technology) to be accessible—that is,

they should be usable by the widest range of users as possible. Moreover these products, services, and systems should be usable by everybody, including future generations, regardless of age, gender, capabilities, or cultural background.

Laws and Guidelines

Numerous countries have passed legislation encouraging or even requiring accessibility in different settings, ranging from the general to the very specific. Different countries have different laws. One of the most critical laws is Section 508 of the Federal Rehabilitation Act in the United States (U.S. Section 508, 2001), which specifies that starting June 21, 2001, the U.S. government shall stop purchasing any information technology that is inaccessible to people with disabilities, nor will it make available for public use any information technology that is inaccessible (Hung, 2001).

There are "Web Content Accessibility Guidelines" (WCAG 1.0, 2007; WCAG 2.0, 2007), but they are not well known by designers. Moreover, neither addresses multimodal software; although multimedia surrounds us, we are not able to avoid it.

EXPERIENCES AND RECOMMENDATIONS

At the Virtual Environments and Imaging Technologies Laboratory of the University of Pannonia, we have been conducting research work on multimedia use for people with special needs, and developed several such programs (Sik Lányi et al., 2005a, 2005b; Sik Lányi, 2006). Some conclusions from this research follow.

What Are the Advantages of Multimedia Software?

- It is an audiovisual medium.
- It can be interactive.
- The treatment or situation can be reproduced, the same condition can be repeated several times.
- It can be adjusted to individual needs.
- Multimedia systems have an effect on more than one sense and can be more effective.
- It can help creativity, it can be varied.
- One can include the motivating qualities of games in multimedia programs.
- It can be designed to ensure the user experiences success.
- One can use motivating audio feedback.
- It can be used both in individual and small-group therapy.

Recommendations

According to the Census 2000 definition, types of disability are: visually impaired, partially sighted, deaf, hard of hearing, mentally retarded, and physically disabled users. In addition it distinguishes color-blind and elderly people as special-needs users. The following multimedia interface design issues for special-needs users will be summarized in this article.

Visual Impairment and Partially Sighted People

It is very important to keep on the developer's mind that the visual impaired and partially sighted people have no perfect vision. The visus of perfect vision is 1. A partially sighted person's visus is between 0.1 and 0.3.

For Blind Persons

- Ensure that all information can be accessed via text or sound, such that blind users can use screen readers or Braille display to access the information.
- GIVE pre-recorded audio as an alternative mean.
- Allow users to navigate the site by using a keyboard (the mouse is hardly used by blind users).
- Minimize the users' memory load because blind users can only hear one word at a time and need memory to integrate parts of the heard information (Hung, 2001).

For Partially Sighted Persons

- Ensure the text size is large enough, otherwise low-vision users usually need a screen magnifier to enlarge the text.
- Give an audio option to notify low-vision users about new information.
- Minimize the users' memory load because the effective screen size is very small while using a screen magnifier.
- For users with low vision, pictures must be drawn with thick contour lines. The user can be given the option to modify the contour line thickness of the objects. The user must be able to vary the color of the objects and background, and the speed of motion, and to stop the animation (Sik Lányi et al., 2005b).

Color-Deficient (Color-Blind) People

Color blindness is mostly neglected; even those impaired do not consider this a serious problem. Yet, it is quite common

to see combinations of background and foreground colors that make pages virtually unreadable for color-blind users. Background, text, and graphics colors should be carefully chosen to allow for people with color blindness. Designing for color-blind people is complicated. It is not a matter of green/red or yellow/blue combinations.

The most important issue in designing for color-blind users is not to rely on color alone to convey information and not to use color as a primary means to impart information (Karagol-Ayan, 2001).

If we have no possibility to test our software by the help of color-blind people, we have to see it in a grayscale setting, at least to check whether all the information is visible or not.

Deaf and Hard-of-Hearing People

People with impaired hearing may have a limited vocabulary. This is one of the problems with hearing-impaired people. Therefore new information and instructions must use simple language alongside cartoon-like presentation. They still require sounds to accompany the graphics. This also applies to anyone with any cognitive impairment.

Also:

- Give visual information (text and/or picture) that is redundant with audible information.
- Allow the users to configure frequency and volume of audible cues.
- Do not design interactions to depend upon the assumption that a user will hear audio information.
- For deaf and hard-of-hearing people to have access to multimedia applications, ways need to be developed to support the presentation of complex sounds and closed captioning for speech (Sik Lányi, 2006).

Physically Disabled Persons

The biggest problem for people with impaired fine motor ability is using the input devices. Remember:

- Do not design the navigation and input for use only by mouse, because the users might have bad motor ability.
- Do not design the navigation, input, and commands for use by voice input devices because of the control problem of facial muscles.
- Do not develop the navigation using multiple keys simultaneously.

Additionally, the multimedia software must be accessible via the keyboard, therefore it must be easy to use and have a good keyboard navigation system. Thus the task is to find the optimal navigation method for the mobility-impaired user. If the user does not has a special input device, navigation can be facilitated with a moving rectangle, the speed of which is adjustable, or through a voice-controlled navigation or command system (Sik Lányi, 2006).

Mentally Retarded

There is a wide variation of cognitive impairments that could be categorized as memory, perception, problem-solving, and conceptualizing disabilities. Memory disabilities include difficulty obtaining, recognizing, and retrieving information from short-term storage, as well as long-term and remote memory.

It is necessary to design multimedia software or Web pages in ways that minimize the skills and abilities required to navigate them. Auditory output might seem confusing to these users or be difficult for them to understand. The designers need to define terms that may not be known to cognitive disabled persons. Some suggestions include:

- Minimize the cognitive load while navigating in the software.
- Use graphics for navigation whenever possible.
- Avoid animated graphics and the use of overlaid large file sizes. Use animations and dynamic display with care.

Elderly Persons

As the years go by, everybody's ability, talent, and skills slowly but surely are wearing out; all is not lost that is delayed. An elderly person is lucky if he or she has only one health problem; in most cases, he or she has more. Therefore if designers develop multimedia software for the elderly, the designers must consider all the above-mentioned recommendations, as well as:

- For all special-needs users, software may have to include a higher degree of help than usual.
- To ensure software meets users' needs, the elderly must be involved in all stages of its development.

Schneiderman (2005, p. 15) states that designs should be based on careful observation of current users, refined by thoughtful analysis of task frequencies and sequences, and validated through early usability and thorough acceptance tests. Designers seek direct interaction with users during the design phase, development process, and throughout the system lifecycle. Iterative design methods that allow early testing of low-fidelity prototypes, revision based on feedback from users, and incremental refinements suggested by usability-test administrators are catalysts for high-quality systems. This system lifecycle is more proper if we develop for special-needs users. Finally, do not forget to consult

end users, teachers, and caretakers, and test the multimedia software with special-needs users too!

FUTURE TRENDS

The recent revolution in telemedicine, using the capabilities of telecommunication, multimedia, and information technologies to provide and support health care when distance separates participants and doctors, offers an opportunity to address the problem of the unequal geographical distribution of physicians and the lack of health care services in rural and urban districts. These communication systems are multimodal; the designer must develop barrier-free user interfaces that most of the users are able to use, or install them by individual needs. Ongoing advances in telemedicine (i.e., telecommunication, multimedia, and information technologies) are likely to change dramatically.

Current speech recognition systems have evolved technologically to begin having an impact in human computer interface designs that can be used by persons with motor disabilities. A real-time user-friendly programming HCI is able to convert voice commands into computer actions. These voice-controlled HCI systems would be the optimal support solutions for people with motor disabilities.

The future will be such multimedia systems, in which not only 2D information is available, but 3D as well. For example, a 3D virtual person (avatar) with the ability of artificial intelligence will help the user to control the interface, or to fill in an input window and so on.

Avatars—also called "embodied conversational agents" (Cassell, Sullivan, Prevost, & Churchill, 2000)—are the actors in a virtual reality world. They have the following attributes:

- They are similar to humans in their appearance—that is, on the exterior.
- Their movements are like real human movements and their gestures are also similar (hand gesture and face mimicry).
- They have a speech recognition module to understand human voices.
- They have artificial intelligence (Russel & Norvig, 2002) to give the user the correct answer, if there is one.
- They are able to utter an answer by means of a text-to-speech engine.
- They can simulate emotions in behavior and also in their facial expression. Their gestures change according to the way they feel.
- They have machine vision to recognize the user and the user's emotions, if this is necessary.
- The future multimedia systems are able to work together with assistive technology devices as well.

CONCLUSION

This article tried to answer some questions and made minimal guidelines to take into account when one wants to develop multimodal software. These guidelines are useful if one wants to develop multimodal software based on the nowadays-expected so-called "Design for All" or "universal design" theory. This chapter dealt with multimedia to be used by special-needs users and gave a minimal requirements' list for designers, software engineers, and computer scientists. Although research in the area of multimedia, HCI, disability, and AT has grown, and some important findings and guidelines have been proposed, our knowledge is still limited in some areas. There are a lot of unanswered questions.

REFERENCES

AAATE. (2007). Retrieved from http://www.aaate.net/index.asp?auto-redirect=true&accept-initial-profile

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (2000). *Embodied conversational agents* (pp. 212-214). Cambridge, MA: MIT Press.

Census 2000. (2007). *Census report on disability.* Retrieved from http://www.who.int/healthmetrics/tools/logbook/en/countries/zmb/2000_Census_Report_on_Disability.pdf

EU Commission. (2003). *2010: A Europe accessible for all.* Retrieved from http://europa.eu.int/comm/employment_social/index/final_report_ega_en.pdf

Hung, E. (2001). *Universal usability in practice, blind and low vision users.* Retrieved from http://www.otal.umd.edu/uupractice/vision/

ICCHP. (2007). *Homepage.* Retrieved from http://www.icchp.org/

ICDVRAT. (2007). *Homepage.* Retrieved from http://www.cyber.rdg.ac.uk/ISRG/icdvrat/home.htm

ICIDH. (1999). *Sustainable design.* Retrieved from http://www.sustainable-design.ie/arch/Beta2full.pdf

Karagol-Ayan, B. (2001). *Universal usability in practice: Color vision confusion.* Retrieved from http://www.otal.umd.edu/uupractice/color/

NCSU. (2007). *Universal design.* Retrieved from http://www.design.ncsu.edu:8120/cud/univ_design/princ_overview.htm

Russel, S.J., & Norvig, P. (2002). *Artificial intelligence. A modern approach* (pp. 32-53). Englewood Cliffs, NJ: Prentice Hall.

Schneiderman, B. (2003). *Leonardo's laptop, human needs and the new computing technologies.* Cambridge, MA: MIT Press.

Schneiderman, B. (2005, November 8). Leonardo's laptop, human needs and the new computing technologies. Proceedings of the 1st Usability Symposium on Empowering Software Quality: How Can Usability Engineering Reach These Goals?, Wien, Austria.

Sik Lányi, C. (2006). Multimedia medical informatics system in healthcare. In A. Ichalkaranje et al. (Eds.), *Intelligent paradigms for assistive and preventive healthcare* (pp. 39-91). Berlin: Springer-Verlag.

Sik Lányi, C., Bacsa, E., Mátrai, R., & Kosztyán, Z. (2005b). Developing interactive multimedia rehabilitation software for treating patients with aphasia, *International Journal on Disability and Human Development, 4*(3), 225-229.

Sik Lányi, C., Mátrai, R., Molnár, G., & Lányi, Z. (2005a). User interface design question of developing multimedia games and education programs for visual impairment children. *Elektrotechnik & Informationstechnik,* **(**12**),** 488-494.

U.S. Section 508. (2001). *U.S. Section 508.* Retrieved from http://www.section508.gov/

WCAG 1.0. (2007). *Web content accessibility guidelines 1.0.* Retrieved from http://www.w3.org/TR/WAI-WEBCONTENT/

WCAG 2.0. (2007). *Web content accessibility guidelines 2.0.* Retrieved from http://www.w3.org/TR/2006/WD-WCAG20-20060427/

Wikipedia. (2007). *Universal design.* Retrieved from http://en.wikipedia.org/wiki/Universal_design

KEY TERMS

Assistive Technology: A set of products, devices, or technical or software systems that are used by disabled, special-needs users. These could be serial or unique special products that help the disabled people's everyday life, and control or decrease the degree of the disability.

Design for All (DfA or D4All): Design of products, services, and systems (including information technology) to be accessible so that they are usable by the widest range of users as possible. Moreover these products, services, and systems should be usable by everybody including future generations, regardless of age, gender, capabilities, or cultural background.

Disability: A qualitative difference of a human capability from its normal feature, which might be in-born; if acquired, can develop backwards only very slowly; or can be permanent and irreversible. Types of disabilities include: physical impairment, sensory impairment, cognitive impairment, intellectual impairment, and/or cumulative impairment.

Health Condition: An alteration or attribute of the health of an individual that may lead to distress, interference with daily activities, or contact with health services; it may be a disease (acute or chronic), disorder, injury, or trauma, or may reflect other health-related states such as pregnancy, aging, stress, congenital anomaly, or genetic predisposition.

Human Computer Interaction (HCI): The study of how humans interact with computers and programs. HCI is a discipline concerned with the design, evaluation, and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.

Impairment: Indicates a loss or abnormality of a body part (i.e., structure) or body function (i.e., physiological function). The physiological functions include mental functions.

Multimedia: Manage in one unit text, graphics, animation, sound, and numeric data in an integrated way. This unit includes the production, storing, processing, transmission, representation, and reproduction of such information.

Multimodal: A software or hardware instrument is multimodal if it uses minimally two media elements, and one of them must be dependent on time—for example, video, sound, or animation files; the term mixed media is also used.

Special-Needs Users: Users who are not able to use the average devices or software in an average way. Handicapped people have special needs in daily requirements throughout their whole lives.

Music Score Watermarking

P. Nesi
University of Florence, Italy

M. Spinu
EXITECH S.r.L., Certaldo, Italy

INTRODUCTION

Music publishers, authors and/or distributors have high quantity of music scores in their archives. In classical music, the original music piece is normally kept in paper format, since its production goes back to many years ago. At present, only new light and popular music pieces are in symbolic notation formats. Light and popular music have a limited lifetime when compared with classical music pieces. The duration of the copyrights for that kind of music is about 60-80 years. Content owners are very cautious to transform their classical music pieces in digital format for e-commerce purposes, because they consider it as a highly risky process which could ultimately lose their copyright ownership. The situation is different when it comes to light and popular music, being market life shorter. According to content owners' opinion, e-commerce for music distribution cannot be accepted, unless adequate protection mechanisms are provided, as highlighted in WEDELMUSIC (www.wedelmusic.org) and MUSICNETWORK (www.interactivemusicnetwork.org). They accept to have their music protected only if it is possible to control while at the same time the users exploit content functionalities according to the established permissions and prices. To cope with these problems, mechanisms for protecting digital musical objects are used (see Table 1).

In this article, only problems and solutions for protecting and watermarking music scores are discussed.

Most music scores are still kept in paper format at publisher's archives. A first step to transform them into digital documents can be transforming them into images with a scanner. Another possible solution can be found in transforming them manually into symbolic music with a music editor. Obviously, this latter solution is very expensive, since the music has to be totally retyped. The use of very efficient Optical Music Recognition (OMR) software, similar to the Optical Character Recognition (OCR), seems to be quite unlikely in the next future. Currently, their recognition rate is close only to 90%, which makes this approach not too much reasonable when compared with retyping (www.interactivemusicnetwork.org, see assessment on the Working Group of Music Imaging).

Music images or symbolic music are obtained after music sheet digitalization. In the event of images, no further music manipulation is possible at the level of symbols. On the other hand, images can be easily viewed in any operating systems and with plenty of applications. The symbolic music gives several advantages in the score maintenance and manipulation; it allows the user to perform changes on the music, such as to justify it, change the page settings, add ornaments, accents, expressions, view single parts or the whole score, and so forth. The drawback consists in all these possible operations being performed only if the music editor is available: professional music sheets are produced by expensive and professional music editors.

It is well known that music sheets are distributed in paper format among musicians. Therefore, it seems that such digitizing process is useless. Practically speaking, Internet music sheet distribution, meaning from publishers to consumers, can only be achieved using digital formats. Distribution among users, as it occurs now with photocopies, could be made even via digital music sheets, as Napster did with audio files. Please note that on P2P (peer to peer) application there is also a quite significant distribution of music scores (www.interactive musicnetwork.org, read report on Music Distribution Models of the Working Group on Music distribution).

Table 1. Mechanisms for protecting digital musical objects

• encryption techniques to support any transferring of music objects;
• watermarking audio files in different formats;
• watermarking images of music score sheets;
• watermarking music sheets while they are printed from symbolic notation files.
• definition of digital rights management policies.

Whenever using digital formats, music could be converted again into paper (today musicians play music only from paper sheets).

BACKGROUND

The most relevant features for algorithms of score watermarking can be summed up into three categories (Monsignori, Nesi, & Spinu, 2003):

- *Content Requirements:*
  The embedded data may contain a simple identification code, which allows to recover the publisher and the distribution IDs simply by consulting a Web service. To this end, hiding about 100 bits is typically enough. The code can be encrypted, compressed and may include control and redundant bits to increase robustness.
- *Visual Requirements:*
  The watermark inserted in the printed music sheet has to be invisible for musicians or at least it should not bother musicians during their execution.
  The watermark has to be included in the music printed by the final user in any format if the music is available in symbolic format. Therefore, the watermark reading has not to depend on the availability of the original reference image of the music sheet.
- *Resistance Requirements:*
  The cost to remove watermark must be extremely expensive when compared to any regular purchase of the same music sheet.
  The watermark must resist against music sheet manipulation until the music printed becomes unreadable. Typically, five levels of photocopy are enough to make music unreadable or of a very bad quality.
  The watermark has to be readable when processing each single page or smaller part.

In addition, there are other parameters to be taken into account in order to analyze the technique capability.

- The amount of embedded information has a direct influence on watermark robustness. Typically, the hidden code is repeated several times in the same page; therefore, the bigger is the code, the lower is the number of times such code can be repeated, which means a decrease in the general robustness.
- Embedding strength " There is a trade-off between watermark embedding robustness and quality. Increased robustness requires a massive embedding of hidden bits. This increases music score degradation and watermark visibility.

Please note that watermarking images of scores or watermarking symbolic music lead to the same result: a watermarked music sheet. The watermarked music (symbolic or image) should be kept in some unchangeable digital file formats (like PDF) or in some formats difficult to change (PostScript), image format. The implementations of the algorithms for music watermarking in such two events are completely different (Busch, Nesi, Schmucker, & Spinu, 2002). In the first event, the watermarking is performed while the music score is printed by manipulating graphic primitives such as lines, music fonts, and so forth, and the process may generate a PostScript file or may send the information directly to the printer. In the latter case, the watermarking is performed by manipulating the B/W images.

In order to read the watermarked hidden code, the music sheet has to be scanned and the resulted image has to be elaborated with the watermark reader, to reconstruct the embedded code. The main advantages of distributing symbolic music sheets, instead of images are:

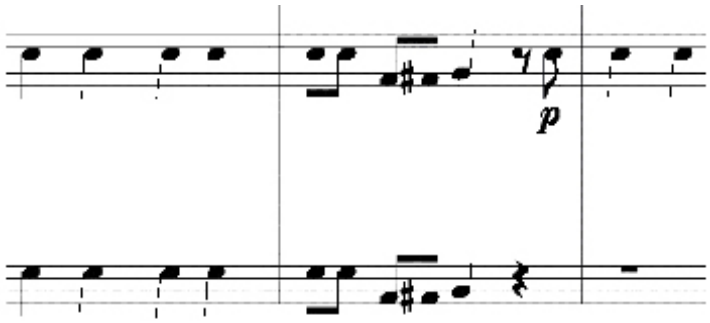
- Lower number of bytes for coding music, easier distribution, lower costs of download, and so forth;
- Higher quality of the printed music sheets, depending on the printer of the final user;
- Possibility of manipulating music notation for transposing, adding annotation, rearranging, and so forth; and
- Possibility of performing a direct music execution from symbolic format to produce MIDI or extended MIDI formats.

All of these features make the use of symbolic music more interesting for music distribution, and therefore its watermarking is very important for music protection.

APPROACHES

According to the user requirements, the printed music sheets must be produced at high resolution and quality. In appreciated music sheets, there is no noise, meaning that the information is in black and white, and therefore no space is left to hide information inside noise or in any kind of noise added-image. This means that the hidden code can be included only under the shape or in the position of music notation symbols. According to such purpose, some common elements of music sheets can be considered: staff lines, stems, note head, bar lines, and so forth. While stepping into such a direction, it is necessary to find a compromise between quality and watermark readability. Quality is very important for musicians and some minor changes could produce readability problems to musicians. They pay attention to the design of musical symbols, and any detectable variation may disturb the musician when playing. In general, the information to

Figure 1. Stem rotation approach



be hidden can be included in the changes considering both their presence and absence, for instance, coding 1 and 0 respectively. In some cases, the magnitude of the change can be used to hide more bits, for example in the orientation, the angle can be variable in order to add more bits.

Stem Rotation

The greatest problems of hiding information in the stem rotation (Busch, Rademer, Schmucker, & Wothusen, 2000) cope with the music score degradation and the low capacity in terms of hidden bits. As depicted in Figure 1, an untrained musician can identify that kind of changes in the music score. This method bothers many musicians when the music is read. In addition, the original music page is needed for watermark reading.

Beam Thickness Modification

By modifying the orientation or thickness of beam lines, it is possible to hide only a few bits. Another important problem has to deal with the presence of beams which is not guaranteed in the music page. Musicians may easily detect the thickness variation when the beam is placed near a staff line. Furthermore, this method requires the original music page in order to perform the watermark reading.

Noteheads Shifting

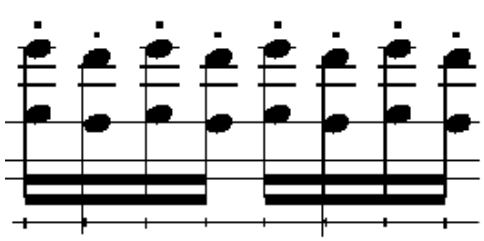
The approach chosen by Schmucker, Busch, and Pant (2001) consists in shifting note heads (see Figure 2). The distance among notes has a musical significance. Therefore, in several cases, the approach may disturb the music reading. In Figure 2, the second chord from the left was moved to left, and musicians may detect the missed alignment of the chords. The displacement has been highlighted with the line below the staff and the gray lines. The movement of notes may generate problems when notes are marked with ornaments, accents, expressions, and so forth. In such cases, the movement becomes evident, thus creating a misalignment of notes with the markers. The idea can work things out and hide a significant code length, if there are enough noteheads in the score page.

If considering the main score, the shifted notes are quite easy to be detected by musicians reading them (according to the needs of simultaneity among parts/layers/voices), while it turns out to be quite invisible in single parts. Such a watermark is easy to be detected by musicians in regular groups of notes, provided that the distance among successive notes of the same beam is non-regular/periodic.

Different Fonts for the Same Music Symbol

According to this technique, different fonts for the selected music symbols are used to hide either 1 or 0, depending on the font used. This implies that the font has to be easily recognized during watermark reading. The approach was proposed for text watermarking by Maxemchuk and Low (1997).

Figure 2. Shifting beamed notes approach



Watermarking Images of Music Sheets

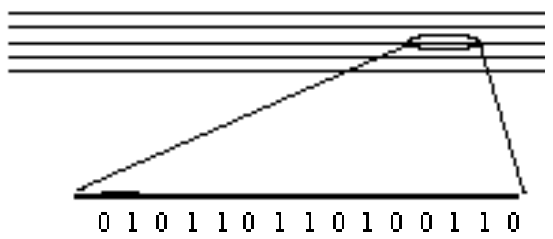
The proposed methods are based on the possibility of storing information by exploiting the relationship of black and white pixels in image segments (i.e., a block) as information carrier (Funk & Schmucker, 2001). The method was elaborated upon Zhao and Koch method (1995) which is based on blocks of distinct size. The ratio of white and black pixels in certain block/area is used to embed a watermark. These areas are treated differently in the process of flipping pixels. The final idea is to embed the watermark only on the black pixels belonging to the staff lines. The fact that the pixel is on a line does not guarantee that it is on the staff line. For this purpose, only horizontal segment having a length greater than a fixed threshold was considered.

Line Thickness Modulation Approach

Figure 3 shows an example of the line modulation. It consists in modifying the lines' thickness in order to insert a binary code made up of several bits. Modulated lines can be easily noted if their presence is known, whereas they are not perceived if their presence is unknown (Monsignori, Nesi, & Spinu, 2001a, 2001b). This approach allows to hide a considerable number of bits in several instances per page, thus making the solution particularly suitable and robust to permit the watermark reading, even out of small parts of the music sheet. This approach has been used in the WEDELMUSIC Editor.

The approach is robust with respect to staff bending, since the watermark is repeated on a large number of staff lines, and it can be read on bended lines. Moreover, a total of 108 bits can be hidden, and a certain number of CRC codes to increase the robustness has been added. This approach can be implemented only starting from the symbolic representation of music notation since the direct manipulation of staff lines on the image may introduce too much noise and produce line deformation.

Figure 3. Staff lines thickness modification



Line Mask Approach

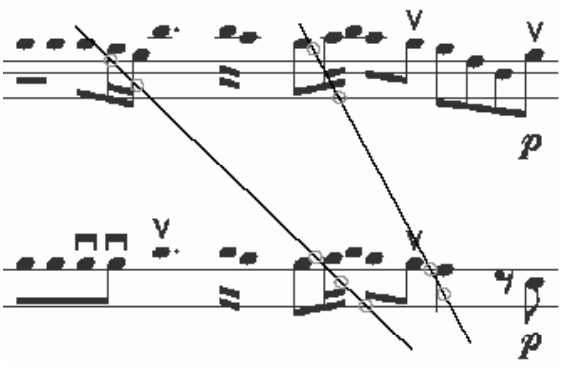
This watermarking approach can be applied to images of music sheet or during the print out of a music score from a symbolic music notation file. The approach consists in marking some points in the music score for virtually hiding a number of lines connecting them (Monsignori et al., 2001a, 2001b). The position and the orientation of the hidden lines are used as the vehicle to hide the watermark code. In particular, the angle between the hidden line and the vertical axis has been used for hiding the information. The idea is not based on writing black lines on the music score (this may only lead up to destroy the music sheet). The points identifying the hidden line may be placed in the intersection among the hidden line and the staff lines, like it occurred with the points in Figure 4, highlighted with circles (in reality, they are interruption on the staff line). In the solution taken, groups of the lines contributing to encoding the same code start from common points. The method allows to hide a large number of bits for each page, and the code can be repeated several time increasing the robustness.

APPROACH VALIDATION

For the validation of these solutions two different phases have to be followed (Monsignori et al., 2003). First, the validation has to be technically performed to assess the robustness against the attacks mentioned at the beginning of this article and to verify the effective coding of a large number of bits, repeated several times per page. As a second phase, the validation has to be focused on verifying the real applicability and acceptability of the solution from experts.

The experts' group has to cover the different needs: publishers, engravers, copyists, and many musicians which are the final users. Therefore, they are a very important category for the watermark validation. A specific watermark approach can be unacceptable for the musicians if the

Figure 4. Points chosen to be marked in the music score



music sheet is not readable or annoying for the presence of evident changes. Typically, copyists are the most exigent. The validation has to request the assessment of a sequence of several different music score pages. Some of them are watermarked; others are not. Different levels of photocopy of the same watermarked or not watermarked music sheets have to be included. Different resolutions (dpi of the printer) of the same music sheets have also been used to assess the minimum acceptance level people involved in the validation. All music sheets were printed at the same magnitude, thus the dimension of the staff line was constant. Its value has been chosen according to that most commonly used in printed sheets.

Experts were informed about the main concepts of watermarking and not about these specific changes made in the music score. They have to perform the assessment individually, without being left with the possibility of a comparison with different pages of music and an exchange of opinions among one another.

FUTURE TRENDS AND CONCLUSIONS

As discussed in this article, the technology of music sheet watermarking is quite mature. Several algorithms have been tested and validated on real applications. The effective value of these solutions is similar to the watermark of Audio file. The presence of a specific watermark in the music sheet may be used to demonstrate the ownership of a music piece over the simple presence of textual fingerprints. In addition, the presence of the watermark can discourage people from any possible and intentional copying action of the music sheet for business purposes. The simple copying of the music sheet among friends is not prevented. The future trends of this technology are mainly in its application for monitoring the distribution of music sheets. In fact, the score watermark can be used for hiding code that can be detected during the simple distribution. This permits the content owners to set up specific services to control the data flow and thus to control and detect the passage of their digital items on the network.

REFERENCES

Busch, C., Nesi, P., Schmucker, M., & Spinu, M.B. (2002). Evolution of music score watermarking algorithm. In E.J. Delp III & P. Wong (Eds.), *Proceedings of the Real-Time Imaging V (EI12) IS\&T/SPIE 2002, Workshop on Security and Watermarking of Multimedia IV*, Vol.4675, San Jose, CA, USA, pp.181-193.

Busch, C., Rademer, E., Schmucker, M., & Wothusen**,** S. (2000). Concepts for a watermarking technique for music scores. In *Proceedings of 3rd International Conference on Visual Computing, Visual 2000*, Mexico City.

Funk, W., & Schmucker, M. (2001). High capacity information hiding in music scores. In *Proceedings of the International Conference on WEB Delivering of Music, WEDELMUSIC2001*, pp.12-19. Florence: IEEE Press.

Maxemchuk, N. F., & Low, S. (1997). Marking text documents. In *Proceedings of International Conference on Image Processing, ICIP97*, 3. Santa Barbara: IEEE Press.

Monsignori, M., Nesi, P., & Spinu, M.B. (2001a). A high capacity technique for watermarking music sheets while printing. In *Proceedings of IEEE 4th Workshop on Multimedia Signal Processing, MMSP2001*, pp.493-498. Cannes: IEEE Press.

Monsignori, M., Nesi, P., & Spinu, M.B. (2001b). Watermarking music sheet while printing. In *Proceedings of the International Conference on WEB Delivering of Music, WEDELMUSIC2001*, pp.28-35. Florence: IEEE Press.

Monsignori, M., Nesi, P., & Spinu, M.B. (2003). Technology of music score watermarking. In S. Deb (Ed.), *Multimedia systems and content-based image retrieval*. Hershey, PA: Idea Group Publishing, pp. 24-61.

Schmucker, M., Busch, C., & Pant, A. (2001). Digital watermarking for the protection of music scores. In *Proceedings of IS&T/SPIE 13th International Symposium on Electronic Imaging 2001, Conference 4314 Security and Watermarking of Multimedia Contents III*, 4314, pp.85-95, San Jose: SPIE Press.

Zhao, J., & Koch, E. (1995). Embedding robust labels into images for copyright protection. In *Proceedings of the International Congress on Intellectual Property Rights for Specialized Information, Knowledge and New Technologies*, pp.242-251, Vienna.

KEY TERMS

Fingerprinting: Used for calling the hidden serial numbers or anything else that should allow to the copyright owner to identify which reseller broke the license agreement. It is used for the multilevel document distribution.

Fragile Watermarking: Techniques that do not guarantee the watermark presence after few document manipulations.

Image Score: An image obtained from a page of music sheet, it can include a main score or a part.

Optical Music Recognition (OMR): Optical recognition of music, transcoding of an image score to a symbolic score format by using a specific algorithm, called OMR.

Robust Copyright Marking: A term used for the techniques that assure a watermark persistence also after the original document was changed in different ways (in the case of the images: cropping, resizing, brightness modification, etc.).

Staff Line: Each single line of the music score staff. The pentagram is made of five staff lines.

Steganography: Techniques that allow secret communication, usually by embedding or hiding the secret information (called embedded data) in other, unsuspected data. Steganographic methods are based on the assumption that the existence of the covert communication is unknown and they are mainly used in secret point-to-point communication between trusting parties. As a result, steganographic methods are usually not robust, that is the hidden information cannot be recovered after data manipulation.

Symbolic Score: A representation of the music notation in symbolic, including a description of music symbols and their relationships. This can be done in some formal specific format such as Finale, Sibelius, WEDELMUSIC, HIFF, SMDL, and so forth.

Watermark: The code hidden into a digital or analog object containing an ID (identification) code or other pieces of information. The watermark is used for identifying the fields of embedded data (serial numbers, logos, etc.) that tell us who is the owner of the object or supply an ID in order to identify data connected with the digital object.

Watermarking: Process of inserting a hidden code or message into a digital or analog object. As opposed to steganography, it has the additional notion of robustness against attacks. As the name suggests, the additional data (the watermark) is added in order to protect the digital document from copyright infringements. Even if the existence of the hidden information is known, it has to be hard for an attacker to destroy the embedded watermark without destroying the data itself.

Watermark Reading: Process of extracting the watermarked code into the watermarked object.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2074-2079, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Neo-Symbiosis

Douglas Griffith
General Dynamics AIS, USA

Frank L. Greitzer
Pacific Northwest Laboratory, USA

INTRODUCTION

The purpose of this article is to re-address the vision of human-computer symbiosis expressed by J. C. R. Licklider nearly a half century ago, when he wrote: "The hope is that in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today" (Licklider, 1960). Unfortunately, little progress was made toward this vision over 4 decades following Licklider's challenge, despite significant advancements in the fields of human factors and computer science. Licklider's vision was largely forgotten. However, recent advances in information science and technology, psychology, and neuroscience have rekindled the potential of making the Licklider's vision a reality. This article provides a historical context for and updates the vision, and it argues that such a vision is needed as a unifying framework for advancing IS&T.

BACKGROUND

Licklider's statement is breathtaking for its vision, especially considering the state of computer technology at that time, that is, large mainframes, punch cards, and batch processing. It is curious to note that Licklider did not use the term symbiosis again, but he did introduce more visionary ideas in a symbiotic vein. An article he co-authored with Robert Taylor titled *The Computer As a Communication Device* made the bold assertion, "In a few years, men will be able to communicate more effectively through a machine than face to face" (Licklider & Taylor, 1968). Clearly, the time estimate was optimistic, but the vision was noteworthy. Licklider and Taylor described the role of the computer in effective communication by introducing the concept of "On-Line Interactive Vicarious Expediter and Responder" (OLIVER), an acronym that by no coincidence was chosen to honor artificial intelligence researcher and the father of machine perception, Oliver Selfridge. OLIVER would be able to take notes when so directed, would know what you do, what you read, what you buy and where to buy it. It would know your friends and acquaintances and would know who and what is important to you. This article made heavy use of the concept of "mental models," which was relatively new to the psychology of that day. The computer was conceived of as an active participant rather than as a passive communication device. Remember that when this article was written, computers were large devices used by specialists. The age of personal computing was off in the future.

Born during World War II, the field of human factors engineering gained prominence for its research on the placement of controls, widely known as knobology, which was an unjust characterization. Many important contributions were made to the design of aircraft, including controls and displays. With strong roots in research on human performance and human errors, the field gained prominence through the work of many leaders in the field who came out of the military: Alphonse Chapanis, a psychologist and a Lieutenant in the U.S. Air Force; Alexander Williams, a psychologist and naval aviator; Air Force Colonel Paul Fitts; and J.C.R. Licklider. Beginning with Chapanis, who realized that "pilot errors" were most often cockpit design errors that could be corrected by the application of human factors to display and controls, these early educators were instrumental in launching the discipline of aviation psychology and human factors engineering that led to worldwide standards in the aviation industry. These men were influential in demonstrating that the military and aviation industry could benefit from research and expertise of the human factors academic community; their works (Fitts, 1951) were inspirational in guiding research and design in engineering psychology for decades. Among the most influential early articles in the field that came out of this academic discipline was George Miller's (1956) "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity to Process Information," which helped to usher in the field of cognitive science and application of more quantitative approaches to the study of cognitive activity and performance.

An early focus of human factors engineering was to design systems informed by known human information processing limitations and capabilities, systems that exploit our cognitive strengths and accommodate our weaknesses (inspired by the early ideas represented in the Fitts' List that compared human and machine capabilities (1951). While the early HFE practice emphasized improvements in the design of equipment to make up for human limitations (reflecting a tradition of *machine centered computing*), a new way

of thinking about human factors was characterized by the design of the human-machine system, or more generally, *human-* or *user-centered computing* (Norman & Draper, 1986). The new subdiscipline of interaction design emerged in the 1970s and 1980s that emphasizes the need to organize information in ways to help reduce clutter and "information overload" and to help cope with design challenges for next-generation systems that will be increasingly complex while being staffed with fewer people.

There have also been theoretical developments in cognitive psychology that provide a foundation for Licklider's vision. Central here is the work by Kahneman (2002, 2003). In his effort to reconcile seemingly contradictory results in studies of judgment under uncertainty, he has advanced the notion of two cognitive systems introduced by Sloman (1996) and others (Stanovich & West, 2002). System 1, termed Intuition, is fast, parallel, automatic, effortless, associative, slow-learning, and emotional. System 2, termed Reasoning, is slow, serial, controlled, effortful, rule-governed, flexible, and neutral. Cognitive illusions, which were part of the work for which he won the Nobel Prize, as well as perceptual illusions, are the results of System 1 processing. Expertise is primarily a resident of System 1 as is most of our skilled performance such as recognition, speaking, and driving. System 2, on the other hand, consists of conscious operations and is commonly thought of as thinking.

System 1 is effective presumably due to evolutionary forces, massive experience, and by constraining context. Most of the time it works quite effectively. System 1 uses nonconscious heuristics to achieve these efficiencies, so occasionally it errs and misfires. Such misfires are responsible for perceptual and cognitive errors. One of the roles of System 2 is to monitor the outputs of System 1 processes.

NEO-SYMBIOSIS: A VISION AND FRAMEWORK FOR CONDUCTING RESEARCH

Licklider's notion of symbiosis does require updating. The term "man/computer symbiosis" is both politically incorrect and factually inaccurate. "Human/machine symbiosis" is preferable. There is also a problem with the term symbiosis itself. Symbiosis implies a co-equality between mutually supportive organisms. However, humans must be in the superordinate position. Dreyfus (1972, 1979, 1992) has made compelling arguments that there are fundamental limitations to what computers can accomplish, limitations that will never be overcome. In this case it is important that the human remain in the superordinate position so that these computer limitations can be circumvented. At the other extreme, Kurzweil (1999) has argued for the unlimited potential of computers. Should it be proven that computers, too, have this unlimited potential, then some attention needs to be paid to Bill Joy and his nightmarish vision of the future should technology go awry (Joy, 2000). In this case, we humans would need to be in the superordinate position for our own survival. Griffith (2005) has introduced the term neo-symbiosis for this updated version of symbiosis.

Kahneman's two system theory plays a central role in neo-symbiosis. It is the System 2 processes that require computer support, not only with respect to the pure drudgery and slowness of System 2 processes, but also with respect to the monitoring of System 1 processes. In most cases, it is a mistake to assign System 1 processes to the computer. This was the fundamental error in many automatic target recognition and image interpretation algorithms that attempted to automate the human out of the loop. The perceptual recognition processes of most humans are quite good. System design should capitalize upon these superb processes and provide support to other areas of human information processing such as search (to overcome a tendency to overlook targets), interpretation keys to provide support for the recognition processes. Other types of System 2 support could include the augmentation (not replacement) of human reasoning processes, support to facilitate adjusting to changes in context to maintain situational awareness and computational support.

A related approach is Joint Cognitive Systems (JCS's) (Hollnagel & Woods, 2005; Woods & Hollnagel, 2006), which represents a specific implementation of cognitive systems engineering. As the term JCS implies, this approach views the human-computer system as a combination of human and machine cognition. Another way of looking at this is that the human is a component of the computer architecture (consistent with our view of neo-symbiosis). In their two volumes, Hollnagel and Woods have developed a sophisticated approach to system design, but it does not draw much from either cognitive psychology or cognitive neuroscience. Neo-symbiosis draws liberally from both cognitive psychology and cognitive neuroscience. In our view, neo-symbiosis is a subset of cognitive systems engineering that may be applied to enrich the field through its focus on human cognition and the supervisory role of humans in joint cognitive systems.

Another related approach is hedonomics. Hedonomics (Hancock, Pepe, & Murphy, 2005) can most easily be thought of as designing technology to climb Maslow's (1970) Hierarchy of Needs. According to Maslow, human needs can be arranged in a hierarchy or pyramid beginning with physiological needs at the base, then proceeding up to safety, love and belonging, self-esteem, and ending with self-actualization at the top. An interesting exercise is to consider how technology can, and sometimes does, facilitate meeting these needs. Hedonomics is certainly in the spirit of neo-symbiosis. Both hedonomics and neo-symbiosis have the same destination. But in its present state, hedonomics presents what is effectively a *brochure* of the destination,

whereas neo-symbiosis provides some *direction* as to how to get to that destination.

Augmented cognition (Schmorrow & Kruse, 2004; Schmorrow, Stanney, & Reeves, 2006; see also Greitzer, 2005) holds much relevance and potential for neo-symbiosis. The program was initiated within the Defense Advanced Research Projects Agency (DARPA) Information Processing Technology Office (IPTO) with the aim to monitor and assess the user's cognitive state through behaviorally- and physiologically-derived measures acquired from the user while interacting with the system, and then to adapt or augment the computational interface to improve performance of the human-computer system. Clearly, the effort here is to address human information processing shortcomings, that is, to augment cognition. A theoretical framework is needed to identify what aspects of human cognition to augment. Greitzer and Griffith (2006) proposed neo-symbiosis, through Kahneman's two system model of cognition, for that theoretical framework. Thought also needs to be given to how human cognition can augment machine cognition in other than a trivial sense. Such human-computer collaboration is needed to realize Licklider's vision of an interaction between humans and computers so that produced levels of performance are achieved that have never been achieved before.

Although it is important to augment human cognition, it is equally important to consider how to leverage human cognition. Consider chess expertise, for example. Rather than trying to build chess playing systems that can defeat grandmasters, why not try to build systems with the grand master in the architecture to achieve unseen levels of chess playing performance. Tournaments could be run in which joint human/computer chess playing systems would compete. This calls to mind the Amazon Mechanical Turk Program (http://www.mturk.com/mturk/welcome) in which humans are invited to be used in programmatic ways by computers to solve problems that the computer can't solve.

FUTURE TRENDS

We shall consider alternative future trends from two different perspectives, a pessimistic perspective in which Licklider's vision is not fully realized, and an optimistic perspective in which desired outcomes and trends do occur. Let us first consider the trends we hope to see in the future.

A most significant facet of this trend will be the occurrence of a revolution in thinking. In a neo-symbiotic system, the human is superordinate. Because system design is theory driven, the system addresses human cognitive shortcomings and leverages human cognitive strengths. To do so requires a model of the human user. There are two levels of user models. One is a generic or nomothetic user model. Essentially, Kahneman's two system model of cognition provides the basis for a nomothetic user model, and there are many human factors handbooks that can assist in building nomothetic models. A second type of user model is an idiosyncratic user model, one that is specific to a particular human user. For example, the chess grand master system envisioned above requires an idiosyncratic model of a specific grand master. In the future, we hope to see systems that develop and refine models of specific human supervisor/partners (users) as a result of interaction with the computers. This is done to a limited extent today, but we envision that this will be taken to new levels of detail and sophistication. It is hoped that research in augmented cognition will become theory driven and will achieve advances to this end. That is, specific cognitive shortcomings would be remedied and specific cognitive strengths would be leveraged.

An article by Griffith and Greitzer (2007) develops a research agenda for neo-symbiosis. This is derived from Kahneman's two system model. Note that the identification of neurological correlates, although potentially helpful and holding vast potential, is not a requirement, nor is it the only enabler for neo-symbiosis. The article shows how requirements can be written to facilitate neo-symbiotic design. The authors also discuss the need for more attention on metrics to assess the extent to which neo-symbiosis has been achieved

Griffith (2006) has argued that neo-symbiosis can be achieved over a wide range of technological sophistication and describes some of its social and political ramifications. In business and industry, for example, the goal of replacing workers with technology would be superseded by using technology to enhance the potential and productivity of workers though their interaction with the technology.

So the bright future for neo-symbiosis is one of increased productivity and fulfillment: That is, people will not only be more productive, but also more fulfilled. They will proceed further up Maslow's hierarchy of needs. People will be happier. Life will be good.

But what if we ignore Licklider's vision? One could argue that the destination will remain the same, but that the journey will take longer. It seems that one could make an argument to this effect along evolutionary lines. A less happy possibility is that we never arrive. We continue to use technology in a non-optimal manner, failing even to address rudimentary usability issues. We fail to achieve the potential and happiness offered by technology as a result of deficient concepts regarding how to think about technology. Yet another scenario is that technology advances, but humanity declines. That is, humans become mentally lazy and potentialities decline as computers are assigned the challenging work. Of course, the most nightmarish vision is that of Bill Joy (2000). Here machines become intelligent and decide that we are no longer needed. The future continues without us.

CONCLUSION

Symbiosis was a metaphor for a vision advanced by Licklider early in the new era of IS&T. Neo-symbiosis is an updating of that vision. We have argued that advances in cognitive theory and computer technology have provided the basis for the realization of that vision. Related approaches have been reviewed. The JCS approach combines human and machine cognition into the architecture. Neo-symbiosis agrees that they need to be combined into the architecture. Hedonomics argues that technology should be viewed as a means for human fulfillment, as does neo-symbiosis. Augmented cognition places heavy emphasis of physiological indices to enhance cognition. Neo-symbiosis is interested in more than enhancing cognition, however. It is also interested in leveraging human cognition so that, in Licklider's words "… the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information processing machines…". Neo-symbiosis not only provides a goal, but also provides a theoretical basis for achieving that goal.

We have also described several alternative futures. If one takes an evolutionary perspective, we might eventually achieve Licklider's vision, but perhaps not as quickly compared to a more deliberate approach. We also speculated about some darker alternative futures in which humanity cedes the field to technology.

REFERENCES

Dreyfus, H.L. (1972, 1979, 1992). *What computers still can't do.* Cambridge, MA: MIT Press.

Fitts, P.M. (1951). Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1287-1340). New York: John Wiley.

Greitzer, F.L. (2005). Extending the reach of augmented cognition to real-world decision making tasks. In *Proceedings of the HCI International 2005/Augmented Cognition Conference*, Las Vegas, Nevada.

Greitzer, F. L., & Griffith, D. (2006). A human-information interaction perspective on augmented cognition. In D.D. Schmorrow, K.M. Stanney, & L.M. Reeves (Eds.), *Foundations of augmented cognition* (2nd ed.). Arlington, VA: Strategic Analysis.

Griffith, D. (2005). Beyond usability: The new symbiosis. *Ergonomics in Design, 13*, 3.

Griffith, D. (2006). Neo-symbiosis: A system design philosophy for diversity and enrichment. *International Journal of Industrial Ergonomics, 36*(12), 1075-1079.

Griffith, D., & Greitzer, F.L. (2007). Neo-symbiosis: The next stage in the evolution of human information interaction. *International Journal of Cognitive Informatics and Natural Intelligence, 1*(1), 39-52.

Hancock, P.A., Pepe, A.A., & Murphy, L. (2005). Hedonomics: The power of positive and pleasurable ergonomics. *Ergonomics in Design, 13,* 8-14.

Hollnagel, E., & Woods, D.D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering.* Boca Raton, FL: CRC Press Taylor and Francis Group.

Joy, B. (2000, April). Why the future doesn't need us. *Wired*, (8.04).

Kahneman, D. (2002, December 8). *Maps of bounded rationality: A perspective on intuitive judgment and choice.* Nobel Prize lecture.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58,* 697-720.

Kurzweil, R. (1999). *The age of spiritual machines: When computers exceed human intelligence.* New York: Penguin Group.

Licklider, J.C.R. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE,* 4-11.

Licklider, J.C.R., & Taylor, R.G. (1968, April). The computer as a communication device. *Science & Technology.*

Maslow, A.H. (1970). *Motivation and personality* (2nd ed). New York: Viking.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity to process information. *Psychological Review*, *63*, 81-97.

Norman, D.A., & Draper, S.W. (1986). *User-centered system design: New perspectives on human-computer interaction*. Mahwah, NJ: Lawrence Erlbaum.

Schmorrow, D.D., & Kruse. A.A. (2004). Augmented cognition. In W.S. Bainbridge (Ed.), *Berkshire encyclopedia of human computer interaction* (pp. 54-59). Great Barrington, MA: Berkshire Publishing Group.

Schmorrow, D.D., Stanney, K.M., & Reeves, L.M. (Eds.) (2006). *Foundations of augmented cognition* (2nd ed.). Arlington, VA: Strategic Analysis.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119,* 3-22.

Stanovich, K.E., & West, R.F. (2002). Individual differences in reasoning. Implications for the rationality debate. In T.

Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases.* New York: Cambridge University Press.

Woods, D.D., & Hollnagel, E. (2006). *Joint cognitive systems: Patterns in cognitive systems engineering.* Boca Raton, FL: CRC Press Taylor and Francis Group.

KEY TERMS

Augmented Cognition: This area of research seeks methods for addressing cognitive bottlenecks (e.g., limitations in attention, memory, learning, comprehension, visualization abilities, and decision making) to extend human information management capacity via technologies that assess the user's cognitive status in real time.

Cognitive Systems Engineering: This is a design philosophy that advances a broad system design perspective employing modeling concepts from engineering, psychology, cognitive science, information science, and computer science, emphasizing human cognitive processes in system design.

Hedonomics: This is a design philosophy that considers a hierarchy of human needs in system design.

Human Centered Design: This is a design philosophy that emphasizes the needs and abilities of the user in the design of a system.

Human Factors: This is the field devoted to understanding and applying the properties of human capabilities to the design and development of systems with the aim of improving operational performance and safety.

Joint Cognitive Systems (JCS's): This design philosophy regards a system as a whole comprising people and technology acting together.

Neo-Symbiosis: This is an updating of Licklider's vision in which technology is placed in a subordinate role to the human.

Symbiosis: This is Licklider's vision that in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today.

Two System Theory of Cognition: Although there are a number of two system theories, this refers to Kahneman's concept of Intuitive and Reasoning systems.

Category: Networking & Telecommunication

Network Effects of Knowledge Diffusion in Network Economy

Zhang Li
Harbin Institute of Technology, China

Yao Xiao
Harbin Institute of Technology, China

Jia Qiong
Harbin Institute of Technology, China

INTRODUCTION

Network industries are the central nervous system of the 21st century economy. During this time the newly developing "network economy" will act as the engine that will drive world development (Bao, 2001). The most valuable commodity in this economy has become information, and the economics of networks applies to almost all information products and services. Information can be consumed by more than one person. Most importantly, the total social value of information increases as it is shared with more consumers. Consumers of computers and software programs, cellular phones, faxes, and Internet services all have more valuable products as the use of these products by others increases. Whether we call this an "information economy" or a "network economy," the implication is the same—network economics accounts for an increasingly larger share of the economy. It is also the driving force behind many of the innovations and technological changes that occur (Balto, 2001).

At the same time, knowledge is nowadays considered to be a fundamental asset of the organizations. Although this concept is not new, in the few last years increasing attention has been devoted to knowledge and knowledge management (KM) issues within organizations. In fact, due to environmental factors such as the market globalization, the increased product complexity, and the turbulence of competitive scenarios, the powerful role of knowledge as a source of sustainable advantage has been considerably emphasized (Zack, 1999). The knowledge economy represents a strategic new era that human beings are entering. In this new environment of social and economic development, knowledge and information are recognized as being at least as important as physical capital, financial capital, and natural resources as a source of economic growth.

Network economy has provided an equal platform for the participation of all of society. It creates unique values and establishes an operational system in the globalization context, depending on the knowledge as core resource, utilizing the network as the fundamental mode, and taking the information industry as leadership. However, in the knowledge economy, networks are adapted better to knowledge-rich environments because of their superior information-processing capabilities. They minimize idiosyncratic investments in fixed assets and technology, and thus are more flexible and responsive to change. "In an economy where the only certainty is uncertainty, the only sure source of lasting competitive advantage is knowledge" (Nonaka, 1994). And one of the most important aspects is that network economy needs to utilize the knowledge diffusion to create more value. Because knowledge diffusion is the core process of knowledge management to explore more network effects and knowledge diffusion is important for total factor productivity, it is also important for international competitiveness. In consequence, knowledge diffusion should be regarded as one of the companies' core competencies.

BACKGROUND

Although interest in how knowledge diffusion plays a role during the network economy era is relatively new, considerable work has been done in the past on related topics. The study of knowledge diffusion and technology transfer is rooted in agriculture, the military, and education (Backer, 1991; Glaser, Abelson, & Garrison, 1983; Rogers, 1983, 1988, 1995). Perhaps one of the best known of technology transfer formulations is Rogers's (1983, 1995) "Diffusion of Innovations" theory. Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system" (Rogers, 1995, p. 5) The field of knowledge diffusion and technology transfer explores the strategies by which knowledge is disseminated and put to use, its benefits and shortcomings, and the policy and practice implications stemming from its application (Backer, 1991). This field represents a cross-disciplinary body of work that has produced an estimated

10,000 literature citations (Backer, 1991) and is widely used in the public health, education, and agricultural fields (Rogers, 1995). Initially conceptualized as a linear process, theories of diffusion and technology transfer have been modified to reflect the dynamic, interactive nature of knowledge dissemination and application (Smale, 1993). Moreover, the knowledge diffusion field has developed a number of important premises or conceptual frameworks. Chat can be useful in addressing the ways in which innovations or new technology is spread. Dunn, Holzner, and Zaltman (1985), backed by a body of literature and experiences in the field of knowledge diffusion dating back to Merton (1968), formulated the following four premises: (1) knowledge use is interpretative, (2) knowledge use is socially constrained, (3) knowledge use is systemic, and (4) knowledge use is transactive.

Network Economy and Knowledge Diffusion

The network economy means that, in view of the widespread use of computer networks in the socioeconomic era, all economic activities are based on the unification of the information transmission and processing in Internet platform, and the economic information cost sharply drops down. This phenomenon is leading to the idea that the information which replaces capital plays a dominant role in economy management and eventually becomes a globalization economic form of the core economic resources. It does not narrowly refer to the industries that are centering on the computer networks, or vigorous developing industries concerned, or the emerging industries. It also includes the combination and infiltration of computer networks and traditional industries, as well as the economic form that is reducing production and transaction costs and increasing productivity level; the result is information has turned to the most important power of industrial development and innovation. The greatest advantage of the network economy lies in accelerating the flow of information and reducing information costs. This influence penetrates the economic life of production, exchange, distribution, and consumption, which have changed the cost structure and other features of these links.

In the process of endogenous growth of knowledge, the endogenous economic growth model stressed the role of the transfer of knowledge. But many of the current studies focused on the role of technology transfer; it has not paid enough attention to the knowledge—the origin of technology. In terms of the implication of the knowledge economy, it includes four basic economics areas: knowledge production, knowledge accumulation, knowledge exchange, and knowledge distribution. These four basic aspects are inseparable from knowledge diffusion. The nature of the knowledge determines the knowledge production bred in the process of the dissemination and the use of knowledge. The knowledge exchange is only a typical state of knowledge diffusion, while the distribution and accumulation of knowledge is only the result of the knowledge diffusion.

There is an authoritative view that nobody will be able to develop better resources and have more wealth without more and better knowledge. Knowledge is not only a symbol of wealth, but also a source of wealth. Only when people can use their own comprehensive grasp of scientific knowledge can knowledge be transformed into wealth.

Tapscott, Lowy, and Ticoll (1998) thought the operational system of the network economy had the following characteristics:

- It is the knowledgeable and digital economic network that is gradually changing the economic metabolism of the world, changing the type of the organization and the institution and nature of economic activities.
- It is the molecular formula structure network.
- All stratifications of the network economic society already transformed from centralized form to discrete form.
- The individual dynamic portfolio gradually substitutes for the old large enterprise, and companies are weaving increasingly complex webs for each other in order to obtain more wealth.
- It is the globalization innovation network and the integration of production and consumption.
- The communication and the cooperation own the instantaneity.

In brief, “the network” and “the knowledge” already became the fundamental mode and the core soul of the network economy.

Network Effects of Knowledge Diffusion

Knowledge may be classified into various categories depending on the purpose of its use. Polanyi (1962) classified knowledge into explicit and tacit knowledge. Explicit knowledge refers to knowledge that is codified in formal, systematic language (encoded knowledge). It is knowledge that can be combined, stored, retrieved, and transmitted with relative ease and through various mechanisms. Tacit knowledge refers to knowledge that is so deeply rooted in the human body and mind that it is hard to codify and communicate. It is knowledge that can only be expressed through action, commitment, and involvement in a specific context and locality. Tacit and explicit knowledge are often treated as separate entities, but it is unlikely that a piece of knowledge will be exclusively explicit or tacit. Knowledge can exist in both forms simultaneously throughout an organization. Knowledge diffusion occurs routinely within organizations.

Knowledge is not static, but fluid; it is absorbed by individuals who interpret, modify, and use it for their own purposes. In particular, two patterns of knowledge diffusion have been documented. First, knowledge flows are geographically localized (Jaffe, Trajtenberg, & Henderson 1993). Second, knowledge diffuses more easily within a firm than between firms (Kogut & Zander, 1992).

Knowledge diffusion has three parts: besides the two parts of knowledge transfer (Davenport, Prusak, 1998)—transmission (sending knowledge to a recipient) and absorption (assimilation and use of knowledge by recipient)—there is a retransfer process. If absorption did not take place, knowledge cannot be said to have successfully transferred. Not to be ignored is that innovation occurred during the whole process of knowledge diffusion. As a result of the difference of the environment, knowledge foundation, principal part of knowledge exchange, innovation occurred during the process of knowledge diffusion; the process of knowledge diffusion is not only the transfer of originality but also the promotion of the second diffusion at the same time. The network effect precisely plays its effectiveness in the knowledge diffusion process. Knowledge value in the network has the accumulation and the transfer effect. The scattered, one-sided, disorderly materials and data may be processed by knowledge according to the users' requirements to review, analyze, and synthesize, thus forming the effective information resource. When the scope of knowledge use is large enough, knowledge networks utilize a special system that will allow each behavior to use the network, or contact the network automatically available and automatically integrated in the network. The sources of knowledge change into integrated knowledge of spontaneous generation or even a higher level, greater value. In addition to the cumulative effect, knowledge also has transfer effect. If knowledge average cost is fixed, the scope of knowledge use is larger and the benefits produced are greater. The online knowledge self-accumulation increment and transfer effect, as well as the work experience, can promote "learn by doing." A predecessor's research indicated that the cumulative effects caused by enterprise investment in information and knowledge produce a spillover effect on other enterprises and even the entire economy. Under certain conditions, marginal returns of knowledge unceasingly increase.

Positive feedback, network effects, and economies of scale are always based on the question of compatibility and network standards. For a network as an information technology, the larger the compatibility, the more transitional costs there are and the stronger are lock-in effects. The diffusion of human knowledge—following the self-expansion of the tongue and interpersonal media, the language and print media, and the digital and electronic media—realizes the optimal resource allocation and diversion and cooperation, and saves plenty of cost. People can share more information, and efficiency would be added. The faster the diffusion, the more expensive the value of the knowledge and corresponding media, even that becoming the most common standards. Media networks include the print media carrying language and electronic media carrying voices and pictures, as well as interpersonal media (face-to-face communications and friendly intercourses). Joining such a network is valuable precisely because many other households and businesses obtain components of the overall system. Diffusion is a process that transforms from the old knowledge network to the new knowledge network through the media, the large network that both types of knowledge share and which is compatible with the foundation environment of knowledge.

To realize the strategy of the network effects of knowledge diffusion requires the following two aspects. From the static aspect, the compatible strategy and tactics of revolution can be adopted. We can weigh and balance the relationship between function and compatibility, and open up and monopolize knowledge diffusion strategies in different market structures. Based on the relationship, there are four different strategies triggered by the new knowledge diffusion to the market that can stimulate the network effect: monopoly transfer, function performing, open-up transfer, and interruption.

From the dynamic aspect, we can find a complete knowledge diffusion process. Firstly, design the market strategies of knowledge diffusion, demonstrate the prominent function to corner the market. Secondly, improve compatibility and others' acceptance through repeated explanations and the establishment of extensive alliances. Finally, lead people to innovate the new knowledge. With the market changes, the strategies change responsively. Through the implementation of such strategies, a knowledge diffusion network will effectively transform knowledge into value, promote innovation, and accelerate the development of the knowledge economy.

FUTURE TRENDS

Peter Drucker (1991) has described the economy of the future as a network society. Business networks are not entirely new, but there has been a rapid evolution in their number, form, and complexity:

- Business networks are replacing traditional markets and vertically integrated companies.
- Global competition is pushing companies to focus on their core competitive competencies.
- IT is lowering transaction costs and providing tools to manage increasingly complex inter-company collaboration.
- Networks are better adapted to knowledge-rich environments because, compared with hierarchical orga-

nizations, they have superior information processing capacity and flexible governance.
- Empowered by new digital media, network organizations are expected to take a lead in creating economic and social innovations.

Therefore, firms are weaving increasingly complex webs of production and customer service, and of innovation and knowledge creation. Knowledge is becoming the most important source of growth as well as productivity. Information means competitive advantage, and knowledge leads to progress. The keys to the strong economic and cultural growth of a nation's future are successful generation, acquisition, diffusion, and exploitation of knowledge.

CONCLUSION

This article describes how knowledge diffusion plays a positive role in the network economy. First, this article analyzes the relationship between network economy and knowledge diffusion, and discusses the media network and characteristics of knowledge diffusion. Furthermore, this article carries on the analysis with the knowledge diffusion network effects by static analysis and dynamic analysis. Finally, the article discusses how to realize the network effect of knowledge diffusion. In conclusion, the elaborate integration of the network economy with knowledge diffusion to create network effects is proposed. This research can provide new vision for information technology to develop in the knowledge economy era.

The prevailing demands of the knowledge economy are challenging organizations worldwide to harness knowledge capital more extensively and innovatively to create greater socioeconomic value. For the coming, distinctively knowledge-based millennium, leaders have singled out intensive knowledge-driven innovation as the decisive success factor for all organizations. It should be reckoned that effective knowledge diffusion will enhance the benefits arising from strategic alliances and ensure enterprises achieve the strategic objectives of existing and future investment activities during the network economy.

REFERENCES

Achrol, R.S., & Kotler, P. (1999). Marketing in the network economy. *Journal of Marketing, 63*(4), 146-163.

Balto, D.A. (2001). Standard setting in the 21st century network economy. *Computer & Internet Lawyer, 18*(6), 5.

Bao, Z.H. (2001). An ethical discussion on the network economy. *Business Ethics: A European Review, 10*(2), 108-112.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Boston: Harvard Business School Press.

Drucker, P.F. (1991). Reckoning with the pension fund revolution. *Harvard Business Review, 69*(3/4), 106-114.

Ernst, D., & Kim L. (2002). Global production networks, knowledge diffusion, and local capability formation. *Research Policy, 31*(8/9), 1417-1429.

Garavelli, A.C., Gorgoglione, M., & Scozzi, B. (2002). Managing knowledge transfer by knowledge technologies. *Technovation, 22,* 269-279.

Herie, M., &, Garth, M.W. (2002). **Knowledge diffusion** in social work: A new approach to bridging the gap. *Social Work, 47*(1), 85-95.

Jaffe, A.B., Trajtenberg, M., & Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics, 108,* 578-598.

Katz, M.L., & Shapiro, C. (1994). Systems competition and network effects. *Journal of Economic Perspectives, 8*(2), 93-115.

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science, 3*(3), 383-397.

Martinez-Brawley, E.E. (1995). **Knowledge diffusion** and transfer of technology: Conceptual premises and concrete steps for human services innovators. *Social Work, 40*(5), 670-682.

Möller, K. (2006). Managing in the network economy. *European Business Forum, 4*(27), 30-35.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

Polanyi, M. (1962). *Personal knowledge: Towards a post-critical philosophy.* Chicago: University of Chicago Press.

Singh, J. (2005). Collaborative networks as determinants of knowledge diffusion patterns. *Management Science, 51*(5), 756-770.

Tapscott, D., Lowy, A., & Ticoll, D. (1998). *Blueprint to the digital economy: Creating wealth in the era of e-business.* McGraw-Hill Professional.

Zack, M.H. (1999). Developing a knowledge strategy. *California Management Review, 41*(3), 125-145.

KEY TERMS

Knowledge: Derived from individuals transforming data and information in a processing hierarchy that enables action (Wilson, 1996).

Knowledge Diffusion: The adaptations and applications of knowledge documented in scientific publications and patents.

Knowledge Economy: Based on the production, the assignment, and the application of knowledge and information.

Knowledge Innovation: A process to utilize experimental research and development activities and empirical practice activities to promote the knowledge that the technical innovation and the system innovation need.

Knowledge Management: Creating, acquiring, interpreting, retaining, and transferring knowledge to improve performance by purposefully modifying behavior based on new knowledge.

Knowledge Transfer: The process by which knowledge is transmitted to, and absorbed by, a user.

Network Economy: All economic activities based on the unification of information transmission and processing in an Internet platform.

Network Effect: A phenomenon whereby a service becomes more valuable as more people use it, thereby encouraging ever-increasing numbers of adopters.

Network Worms

Thomas M. Chen
Southern Methodist University, USA

Greg W. Tally
SPARTA Inc., USA

INTRODUCTION

Internet users are currently plagued by an assortment of malicious software (malware). The Internet provides not only connectivity for network services such as e-mail and Web browsing, but also an environment for the spread of malware between computers. Users can be affected even if their computers are not vulnerable to malware. For example, fast-spreading worms can cause widespread congestion that will bring down network services.

Worms and viruses are both common types of self-replicating malware but differ in their method of replication (Grimes, 2001; Harley, Slade, & Gattiker, 2001; Szor, 2005). A computer virus depends on hijacking control of another (host) program to attach a copy of its virus code to more files or programs. When the newly infected program is executed, the virus code is also executed. In contrast, a worm is a standalone program that does not depend on other programs (Nazario, 2004). It replicates by searching for vulnerable targets through the network, and attempts to transfer a copy of itself. Worms are dependent on the network environment to spread. Over the years, the Internet has become a fertile environment for worms to thrive.

The constant exposure of computer users to worm threats from the Internet is a major concern. Another concern is the possible rate of infection. Because worms are automated programs, they can spread without any human action. The fastest time needed to infect a majority of Internet users is a matter of speculation, but some worry that a new worm outbreak could spread through the Internet much faster than defenses could detect and block it. The most reliable defenses are based on attack signatures. If a new worm does not have an existing signature, it could have some time to spread unhindered and complete its damage before a signature can be devised for it.

Perhaps a greater concern about worms is their role as vehicles for delivery of other malware in their payload. Once a worm has compromised a host victim, it can execute any payload. Historical examples of worms have included:

- **Trojan horses:** Software with a hidden malicious function, for example, to steal confidential data or open a backdoor;
- **Droppers:** Designed to facilitate downloading of other malware;
- **Bots:** Software to listen covertly for and execute remote commands, for example, to send spam or carry out a distributed denial of service (DDoS) attack.

These types of malware are not able to spread by themselves, and therefore take advantage of the self-replication characteristic of worms to spread.

This article presents a review of the historical development of worms, and an overview of worm anatomy from a functional perspective.

BACKGROUND

The term "worm" was created by John Shoch and Jon Hupp at Xerox PARC in 1979, inspired by the network-based multisegmented "tapeworm" monster in John Brunner's novel, *The Shockwave Rider* (Shoch & Hupp, 1982). They were aware of an earlier self-replicating program, Creeper, written by Bob Thomas at BBN, which propelled itself between nodes of the ARPANET. They invented a worm to traverse their internal Ethernet LAN seeking idle processors after normal working hours for the purpose of distributed computing. Because the worms were intended for beneficial uses among cooperative users, there was no attempt at stealth or malicious payload. Their worms were designed with limited lifetimes, and responsive to a special "kill" packet. Despite these safeguards, one of the worm programs believed to have been accidentally corrupted ran out of control and crashed several computers overnight.

The most famous worm incident was the Morris worm in November 1988 that disabled 6,000 computers in a few hours (Spafford, 1989). Robert Morris Jr. was a student at Cornell University at the time. The damage was caused by the worm re-infecting computers that were already infected, until the computers slowed down and crashed. It was probably the first worm to use a combination of methods to spread quickly. First, it attempted to crack password files on Unix systems. The password file was encrypted but publicly readable. The worm could encrypt password guesses and compare them to the contents of the password file. Second, it exploited

the debug option in the Unix sendmail program. Third, it carried out a buffer overflow exploit taking advantage of a vulnerability in the Unix finger daemon program.

Worm development was relatively slow until 1999, when e-mail became a popular infection vector. In March 1999, Melissa spread to 100,000 computers in 3 days, setting a new record and shutting down e-mail for many companies using Microsoft Exchange Server (CERT advisory CA-1999-04, 1999). It was a Microsoft Word macro that used the functions of Word and Outlook e-mail to propagate. When the macro is executed in Word, it launched Outlook and sent itself to 50 recipients found in the address book. Additionally, it infected the Word normal.dot template, so that any Word document created from the template would carry the infection.

In the summer of 1999, the PrettyPark worm propagated as an e-mail attachment called "Pretty Park.exe" with the icon of a character from the television show "South Park." If executed, it installed itself into the system folder and modified the registry to ensure that it ran whenever any .exe program was executed. It e-mailed itself to addresses found in the address book. Another worm, ExploreZip, appeared to be a WinZip file attachment in e-mail but was not really a zipped file. When executed, it displayed an error message but the worm secretly copied itself into the systems folder and loaded itself into the registry. It e-mailed itself using Outlook or Exchange to recipients found in unread messages in the inbox. It monitored all incoming messages and replied to senders with a copy of itself.

The summer of 2000 saw more mass mailing worms. In May 2000, the Love Letter worm appeared with the subject line "I love you" and encouraged the recipient to read the attachment which was a Visual Basic script (CERT advisory CA-2000-04, 2000). When executed, the worm installed copies of itself into the windows and system directories and modified the registry to ensure that it would be run during bootup. It infected various types of files (.vbs, .jpg, .mp3, etc.) on local drives and networked shared directories. If Outlook is installed, the worm e-mailed copies of itself to anyone found in the address book. In addition, the worm sent copies of itself via IRC channels.

Appearing around the same time, NewLove was a Visual Basic script worm. It was interesting as a polymorphic worm that tried to change its appearance in every copy. The worm forwarded itself with a file name chosen randomly from "recent documents" to all addresses in the Outlook address book. The e-mail has no text but has a subject line including the new file name.

In October 2000, the Hybris worm spread as an e-mail attachment (CERT incident note IN-2001-02, 2001). If executed, it modified the "wsock32.dll" file in order to track all Internet traffic at the infected host. For every e-mail sent, it subsequently sent a copy of itself to the same recipient. It had the interesting capability to receive plug-ins dynamically by connecting to a preprogrammed newsgroup. The plug-ins were encrypted and updated the worm code. This capability is potentially dangerous because the worm functionality can be changed at any time by the worm author.

A new wave of more sophisticated worms began in early 2001. In March 2001, the Lion worm spread among Linux computers using the "pscan" application, a freely distributed network port scanner written in Perl. The worm used this port scanner in combination with the "randb" program to scan class B hosts listening on TCP port 53 that were vulnerable to the BIND buffer overflow vulnerability. It then attacked these hosts using an exploit called "name." After a system was compromised, the worm stole password files and other sensitive information (IP address, accounts) and sent these by e-mail. It also installed several things: the t0rn rootkit to evade detection, the DDoS agent TFN2K, a Trojanized version of SSH to listen on port 33568, and backdoor root shells on TCP ports 60008 and 33567.

In May 2001, the Sadmind worm first exploited a buffer overflow vulnerability in Sun Solaris systems. These compromised systems were then used to carry out an attack to compromise Microsoft IIS (Internet Information Services) Web servers.

In July 2001, the Code Red worm caused major damages by exploiting a buffer overflow vulnerability discovered in Microsoft IIS Web servers about a month earlier (Berghel, 2001; Moore, Shannon, & Brown, 2001). Specifically, the Index Server ISAPI vulnerability allowed a remote attacker to gain full system level access (Microsoft Security Bulletin MS01-033, 2001). The first version of the Code Red worm appeared on July 12. On infected systems, it set up 100 parallel threads, each an exact replica of the worm, in order to spread faster. It attempted to generate pseudorandom IP addresses but used a static seed which (apparently unintentionally) resulted in identical lists of IP addresses. Although 200,000 hosts were infected in 6 days, the worm was slowed down by the fact that the same targets were getting hit repeatedly. A second version of Code Red appeared on July 19. This version spread much faster because the static seed had been changed to a random seed, ensuring that each copy of the worm generated different IP addresses. More than 359,000 computers were reportedly infected by Code Red version 2 within 14 hours. By design, the worm stopped by itself on July 20. On August 4, a new worm self-named Code Red II used the same buffer overflow exploit but a different payload. It generated random IP addresses but they are not completely random; about 1 out of 8 are completely random; 4 out of 8 addresses are within the same class A range of the infected host's address; and 3 out of 8 addresses are within the same class B range of the infected host's address. On infected systems, it activated 300 parallel threads to spread faster. The enormous number of parallel threads created a flood of scans, resulting in serious network congestion.

In September 2001, the Nimda worm used a combination of five methods to spread quickly: e-mail to addresses from the host's Web cache and default MAPI mailbox, with random subject lines and an attachment named "readme.exe"; attacked random Microsoft IIS Web servers though a buffer overflow vulnerability published a year earlier; copied itself across open network shares; added Javascript to Web pages to infect Web browsers; and looked for backdoors left by previous Code Red II and Sadmind worms. It was able to spread to 450,000 hosts within 12 hours. Although none of the methods was new, the combination of so many methods in one worm was unusually complex.

In November 2002, the Winevar worm was an example showing the capability to protect itself by disabling antivirus software. It used a list of keywords to scan memory to stop recognized antivirus processes and scan the hard drive to delete associated files.

In January 2003, the SQL Slammer (or Sapphire) worm spread among Microsoft SQL servers (Moore et al., 2003). Interestingly, the worm consists simply of a 376-byte payload in a single 404-byte UDP packet. This is advantageous for fast spreading because infected hosts can generate these short UDP packets quickly. Unlike TCP, UDP is connectionless and does not require a host to wait for a connection set up. Infected hosts were put into a simple loop to send UDP packets to randomly generated IP addresses as fast as possible. The packets carried an exploit for a buffer overflow vulnerability in Microsoft SQL Server discovered 6 months earlier (Microsoft Security Bulletin MS02-039, 2002).

The week of August 11, 2003, has been called one of the worst weeks in worm history. First, the Blaster (or Lovsan) worm exploited a DCOM RPC (distributed component object model remote procedure call) vulnerability in Windows 2000 and Windows XP systems. On vulnerable systems, the worm opened a remote shell process that transfers a file "msblast.exe" from an infected host. Seven days later, the Welchi (or Nachi) worm used the same exploit along with an exploit for a WebDAV vulnerability in Microsoft IIS 5.0 servers. Interestingly, Welchi attempted to remove Blaster from infected systems and applied the Microsoft patch for the RPC vulnerability. It was programmed to self terminate on January 1, 2004. One day after Welchi, the Sobig.F worm, the fifth variant of the original Sobig.A worm discovered in January 2003, spread by mass mailing. It spoofed the "from" address in e-mails with a randomly chosen address found on the victim's computer. It had the capability to download arbitrary files to an infected computer. It was used to set up spam relay servers and steal confidential system information. At preprogrammed times, it contacted a number of master servers to get download instructions. Around the same day, the Dumaru worm pretended to be a Microsoft patch "patch.exe" attachment in e-mail. If opened, the worm copies itself into the system directory and installs a Trojan horse that listens to an IRC channel for commands from the worm author.

2004 was notable for a conflict between the authors of MyDoom, Netsky, and Bagle, evidenced by messages embedded in the worm codes. The MyDoom.A worm appeared in January. It e-mailed itself to addresses harvested from various types of files on the infected host, along with various subject lines and attachment names. The payload contains a DDoS agent and a backdoor to download arbitrary files. Soon afterward, the Bagle worm spread similarly by e-mail and installed a Trojan horse that opened a backdoor to allow remote control. The Netsky family of worms, also mass mailers, appeared shortly afterward with comments embedded in its code directed at the authors of MyDoom and Bagle, and some variants contained code to remove them from infected hosts.

Although worms have continued to evolve since 2004, there have not been "big" worm outbreaks on the scale of Slammer or Code Red. Worm writers have seemed to be spending more efforts toward exploring new infection vectors, such as instant messaging, Internet relay chat (IRC), peer-to-peer file sharing, or SMS/MMS (short message service/multimedia messaging service). It could be said that worms are still perceived as a major threat but fading in importance compared to other emerging malware threats. Since 2005, concern has been gradually shifting away from worms toward other types of malware, namely bots, spyware, and rootkits.

Worm Anatomy

Worms must have certain functions in their code for self replication:

- **Target identification:** To locate new targets
- **Infection mechanism:** To compromise a new target
- **Replication:** To transfer a worm copy to a target.

Optionally, worms might contain timing control and a payload. Timing might be controlled for self termination; downloading plug-ins or worm code updates; downloading new malware to infected systems; or activation of the payload.

Worms do not always carry a payload, and payloads can be virtually anything. Payloads such as a DDoS agent might be activated by the timing control (e.g., to start flooding at the same time) (Mirkovic, Dietrich, Dittrich, & Reiher, 2004).

Target Identification

The simplest method to find new targets is randomly chosen IP addresses (essentially 32-bit numbers). However, this approach is not efficient. As more hosts become infected,

the spreading rate slows down due to infected hosts hitting targets that are already infected. This inefficiency creates excessive traffic in the network, which slows down the spreading rate further.

Worms such as Blaster and Code Red II have used more complicated algorithms for IP addresses. Blaster chose a random IP address only 60% of the time; at other times, it attempted to find an address in the same local network as the infected victim. Code Red II chose random IP addresses 1 out of 8 times; 4 out of 8 addresses were within the same class A range; and 3 out of 8 addresses were within the same class B range as the infected host.

Another popular method is to harvest e-mail addresses from the victim host. Early mass mailing worms starting with Melissa found addresses from the address book. More sophisticated worms such as MyDoom can harvest e-mail addresses from many types of files located on a victim. The rationale for targeting addresses found from a victim is that recipients are more likely to read e-mail if it was apparently sent from an acquaintance.

Infection Vectors

It is apparent from the historical review that worms can spread by any number of ways. Since 1999, e-mail has been one of the most popular infection vectors because: worms often carry their own SMTP engine; e-mail can take advantage of social engineering; messages can be easily forged and mutated; e-mail can take advantage of social connections which may be more effective than random contacts.

Worms can also exploit vulnerabilities. The most common type of exploit is a buffer overflow because: it can usually be done remotely; it can give complete control over a target; and buffer overflow vulnerabilities are found in many operating systems and applications (Foster, Osipov, & Bhalla, 2005). Even the early Morris worm used a buffer overflow exploit.

Worms such as Lirva and Fizzer were able to spread by file sharing, namely the KaZaa peer-to-peer network. The worm resides in a shared folder, usually with a harmless name.

Worms can spread by messaging via instant messaging or IRC (Internet relay chat). The 2003 Lirva worm was able to spread by IRC. An infected file or URL is sent to a chat channel. In March 2005, the Kelvir family of worms began to spread by instant messaging via MSN Messenger. Random looking messages contained a link to a Web site which attempted to download files. When downloaded and executed, the worm continues to spread by sending instant messages to all found MSN messenger contacts.

Additional infection vectors include: password cracking (e.g., Morris worm); copy to open network shares; modification of Web sites for drive-by downloading; taking advantage of backdoors left by previous worms or Trojan horses (e.g., Nimda worm); and spreading by Bluetooth, SMS/MMS (short message service/multimedia messaging service), or other wireless connections (Hypponen, 2006).

Payloads

The optional payload of a worm is executed after a new victim has been compromised. Some worms have no payloads, and the reason is not known for certain. A payload could be virtually anything. In past cases, common payloads have included: bots to control a group of infected hosts as a bot net (Schiller & Binkley, 2007); spam relay servers to generate spam; backdoors to allow covert remote access; DDoS agents such as TFN2K; spyware, key loggers, and other Trojan horses; and rootkits to evade detection. While destructive payloads are entirely possible, it might be counterproductive, resulting in slower spreading and more attention from security experts.

The payload is often considered to be a clue to the worm author's motivations. When a payload is absent, the worm might be a proof of concept (e.g., to see how fast it could spread). Payloads for spamming or stealing personal information suggests a profit motive.

FUTURE TRENDS

Although widespread outbreaks of fast-spreading worms have been less common since 2003, worms are still a serious threat according to most surveys of organizations. The nature of the threat has simply continued to evolve.

First, worms continue to expand to new infection vectors. For example, the Cabir worm in June 2004 was the first to spread by Bluetooth between Symbian smartphones (Hypponen, 2006). This was followed by the ComWar worm in March 2005 using MMS as the infection vector. ComWar was followed shortly by the Mabir worm which was able to spread by both MMS and Bluetooth.

Second, there has been a growing prevalence of payloads oriented toward control (bots, backdoors, rootkits) and financial profit (spyware, keyloggers). Anti-spyware and rootkit detection programs are quickly becoming essential protection for computer users. These other types of malware do not have to be delivered by worms. For example, malware can spread by drive-by-downloading at a malicious Web site. But worms continue to be a popular vehicle to deliver a variety of malware.

Third, social engineering continues to be common. Malware writers have been quick to take advantage of interest in current events to entice e-mail recipients to read spam. Another example is one of the most prevalent worms in 2006 was the Nyxem (or Blackmal or "Kama Sutra") worm which offered sexually provocative subject lines and body texts.

CONCLUSION

Worm evolution has progressed from early experimentation to sophisticate vehicles for other types of malware. Worms are commonplace on the Internet and threaten to expand to other networking environments such as wireless.

Unfortunately, the nature of the worm threat is essentially similar to other criminal activities. Worms are created by criminals, and it is impossible to predict how new worms will be invented. Thus, defenses are always catching up to new attacks. There are natural questions that may always be somewhat uncertain. For instance, when will another major worm outbreak happen? How fast could a worm spread and what damage will be caused? Continued research is needed to address these questions.

REFERENCES

Berghel, H. (2001). The Code Red worm. *Communications of the ACM, 44*(12), 15-19.

CERT advisory CA-1999-04. (1999). *Melissa macro virus*. Retrieved December 8, 2007 from http://www.cert.org/advisories/CA-1999-04.html

CERT advisory CA-2000-04. (2000). *Love letter worm*. Retrieved December 8, 2007 from http://www.cert.org/advisories/CA-2000-04.html

CERT incident note IN-2001-02. (2001). *Open mail relays used to deliver Hybris worm*. Retrieved December 8, 2007 from http://www.cert.org/incident_notes/IN-2001-02.html

Foster, J., Osipov, V., & Bhalla, N. (2005). *Buffer overflow attacks*. Rockland, MA: Syngress Publishing.

Grimes, R. (2001). *Malicious mobile code: Virus protection for Windows*. Sebastopol, CA: O'Reilly & Associates.

Harley, D., Slade, R., & Gattiker, R. (2001). *Viruses revealed*. New York: Osborne/McGraw-Hill.

Hypponen, M. (2006). Malware goes mobile. *Scientific American, 295*(5), 70-77.

Microsoft Security Bulletin MS01-033. (2001). *Unchecked buffer in Index Server ISAPI extension could enable Web server compromise*. Retrieved December 8, 2007 from http://www.microsoft.com/technet/security/bulletin/MS01-033.asp

Microsoft Security Bulletin MS02-039. (2002). *Buffer overruns in SQL Server 2000 resolution service could enable code execution*. Retrieved December 8, 2007 from http://www.microsoft.com/technet/security/bulletin/MS02-039.asp

Mirkovic, J., Dietrich, S., Dittrich, D., & Reiher, P. (2004). *Internet denial of service: Attack and defense mechanisms*. Upper Saddle River, NJ: Prentice Hall.

Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., & Weaver, N. (2003). Inside the Slammer worm. *IEEE Security & Privacy, 1*(4), 33-39.

Moore, D., Shannon, C., & Brown, J. (2002). Code-Red: A case study on the spread and victims of an Internet worm. In *Proceedings of the ACM Internet Measurement Workshop 2002,* Marseille, (pp. 273-284).

Nazario, J. (2004). *Defense and detection strategies against Internet worms*. Norwood, MA: Artech House.

Schiller, C., & Binkley, J. (2007). *Botnets: The killer Web app*. Rockland, MA: Syngress Publishing.

Shoch, J., & Hupp, J. (1982). The "worm" programs—early experience with a distributed computation. *Communications of the ACM, 25*(3), 172-180.

Spafford, E. (1989). The Internet worm program: An analysis. *ACM Computer Communications Review, 19*(1), 17-57.

Szor, P. (2005). *The art of computer virus research and defense*. Upper Saddle River, NJ: Addison-Wesley.

KEY TERMS

Computer Virus: A set of program instructions capable of self replication by attaching to a normal host file or program.

Exploit: Code written to take advantage of a specific vulnerability.

Infection Vector: The transmission channel for spreading an infection.

Malicious Software (malware): The broad variety of software containing a harmful function, such as viruses, worms, and Trojan horses.

Payload: The part of a virus or worm that is executed after a target host has been successfully compromised and infected.

Social Engineering: A type of attack taking advantage of human gullibility.

Vulnerability: A weakness or bug in software programs that could lead to a security compromise if exploited.

Worm: An automated standalone program capable of self replication by copying itself to vulnerable hosts through a network.

Networked Virtual Environments

Christos Bouras
University of Patras, Greece

Eri Giannaka
University of Patras, Greece

Thrasyvoulos Tsiatsos
Aristotle University of Thessaloniki, Greece

INTRODUCTION

The inherent need of humans to communicate acted as the moving force for the formation, expansion and wide adoption of the Internet. The need for communication and collaboration from distance resulted in the evolution of the primitive services originally offered (i.e., e-mail) to advanced applications, which offer a high sense of realism to the user, forming a reality, the so-called virtual reality. Even though virtual environments were first introduced as stand alone applications, which could run on a single computer, the promising functionalities of this new form of representation and interaction as well as the familiarity of the users with it drew increased research interest. This fact resulted in virtual reality to be viewed as the solution for achieving communication and collaboration between scattered users, in various areas of interest, such as entertainment, learning, training, etc. This led to the creation of Networked Virtual Environments (NVEs). In particular, NVEs were first introduced in the 1980's and the first areas that exploited the newborn technology were military and entertainment applications. In particular, the U.S Department of Defense played an important role to the direction of applications, protocols and architectures for this promising technology. In the 1990's, where academic networks became a reality, NVEs drew increased academic research interest and a variety of applications and platforms were developed. In particular, the academic community has reinvented, extended, and documented what the Department of Defense has done. The evolution and the results extracted by research on this field were widely adopted from multiple areas of interest, with main representative the entertainment area.

Since 2000, where virtual reality technology, processing power of computers and the network were significantly improved, a wide variety of systems, protocols and applications were developed. In particular, the familiarization the end users with the Internet and the promising advantages and opportunities of Virtual Reality contributed to currently view NVEs as an effective tool for supporting communication and collaboration of scattered users. Currently, the application areas of NVEs have been widely expanded and their use can be found at military and industrial team training, collaborative design and engineering, multiplayer games (Zyda, 2005), mobile entertainment, virtual shopping malls, online tradeshows and conferences, remote customer support, distance learning and training, science, arts, industry, etc. Summarizing, NVEs nowadays tend to consist a powerful tool for communication and collaboration, with applications ranging from entertainment and teleshopping to engineering and medicine. To this direction, in the recent years important active research on this topic in both academic and industrial research is taking place.

BACKGROUND

NVE is a twofold term. Even though the "Virtual Environments" part prevails, the "networked" substance changes the meaning and nature of these environments. Regarding the Virtual Environment, it can be considered as a simulation generated by a computer, which can simulate either an imaginary or real world. Even though Virtual Environments can be two-dimensional, the term is mainly related to three-dimensional environments that aim at providing to the users a high sense of realism by incorporating realistic 3D graphics and stereo sound, to create an immersive experience. As far as it concerns the "networked" part of the term, this dimension is mainly related to the support of multiple concurrent users, scattered around the globe, even though NVEs can be single user applications. A definition provided by Singhal and Zyda (1999) states that "NVEs are software systems that can support multiple users, which can interact both with each other and with the environment in real time and aim at providing to the users a high-sense of realism by incorporating 3D graphics and multimedia."

The concept of a NVE is simple. Two or more users can view the Virtual Environment (VE) on their computer, having their own local copy of the virtual world. For achieving

high-sense of realism and maintaining consistency, when a user performs actions on one computer, these actions are propagated through the network to other participating computers for keeping all copies of the VE synchronized. The participants constitute active parts of the VE, usually represented by human-like entities, called avatars for enhancing the awareness (Joslin, Pandzic & Thalmann, 2003).

As mentioned earleir, the network constitutes the core of NVEs. However, NVEs can be further categorized by their architectural model or the nature, in terms of the kind of application they plan to support (Macedonia, 1997). In particular, regarding the architectural model, the most popular category of NVEs are the Distributed Virtual Environments (DVEs), where active parts of the virtual environment are scattered to different computers, which are connected through the network. Accordingly, in respect to the nature of these environments, one of the major categories are the Collaborative Virtual Environments (CVEs), where the users have the ability to meet and interact with others, with agents and the objects of the virtual environment.

MAIN ISSUES AND CONCEPTS IN NVES

A NVE constitutes a computer system, which generates virtual worlds, where the users can interact both with the system and the other connected users in real time. The users are connected to the Internet and working on different computers, access the same virtual scene. The simulation of the virtual scenes is realized through distributed and heterogeneous computational resources. The evolution of the software applications and services in combination to the melioration of the network allows for the development of networked applications, which are characterized by the enhancement and combination of many advanced features. For NVEs in particular, where the achievement of high realism constitutes a key concept, the realistic and detailed representation of the provided information is of high importance. Therefore, the potentialities that technology presents in combination to the increased needs of the users result in NVEs to adopt rich representation for the information in terms of graphics and media.

Despite the fairly simple concept, the design of NVE systems involves a complex interaction of several domains of Computer Science. In particular the interacting domains are the following: (a) networking, which is related to the transmission of various types of data with different requirements in terms of latency, bitrate, and so forth, (b) simulation, which is related to the virtual environment and involves visual database management and rendering techniques with real time optimizations, (c) human-computer interaction, which is related to the support of various types of devices, (d) virtual human simulation, which is related to the avatar's realistic representation in terms of facial expressions, motions, and so forth, and (e) artificial intelligence involving decision making processes and autonomous behaviors (Joslin et al., 2003).

This section will present the basic issues related to NVEs, in terms of the basic features they need to support, the components necessary, in terms of the hardware needed for their operation and interaction with the users, the most common architectures adopted for supporting such environments, the technologies and protocols for their development as well as the issues and factors that should be taken into account for assuring a good performance.

Basic Characteristics

As mentioned above, NVEs can represent either a real or imaginary world. Thus, the structure, the space, the objects and the functionalities provided in such an environment may significantly vary in respect to the concept they aim to support. However, for achieving a high sense of realism, NVEs are characterized by some common features. In particular, these environments should provide: (a) a shared sense of space, in terms of creating the illusion to the users that they are being located in the same place, (b) a shared sense of presence, which is mainly related to the virtual representation of the users that is commonly realized through human-like personas called avatars as well as to the visibility of others participants entering or leaving the environment, (c) a shared sense of time, in terms of being able to see other participants' actions when they occur, (d) a way to communicate, which can be achieved through gestures, typed text and voice and finally (e) a way to share, in terms of being able to interact realistically not only with other participants but also with the virtual environment itself (Singhal & Zyda, 1999). The support of the above-mentioned characteristics is critical for the successful simulation of reality and vital for the effective communication and collaboration of the participating users.

Basic Components

In terms of the hardware needed for NVEs, four components are found necessary for the correct and successful operation of these environments. In particular the components needed are: (a) graphics engines and displays, which constitute the cornerstone of the user interface and the users' "window" to the environment, (b) communication and control devices (e.g., keyboard, mouse, joystick, dataglove, head mounted display, motion detectors in full-body immersive environments), which allow and support the manipulation of the objects of the environment as well as the navigation and interaction of the user with the environment, (c) processing systems for computing and determining the transmission of the events that take place within a virtual environment and

last but not least (d) data network for the actual communication, transmission of information and sharing of data. The components work together for achieving and maintaining the sense of realism among the scattered users.

Architectures

From a more technical point of view, the architectures that support these types of software systems usually fall into one of the following cases: (a) client-server architectures, where the clients communicate their changes to one or more servers and these servers, in turn, are responsible for the redistribution of the received information to all connected clients and (b) peer-to-peer architectures, where the clients communicate directly their modifications and updates of the world to all connected clients (McGregor, Kapolka, Zyda & Brutzman, 2003). The case of the client-server model is the most simple but it is cannot support high scalability as there is a central point of failure, the server. As far as it concerns the peer-to-peer model the scalability is restricted by the network. It should be mentioned that hybrid solutions can be adopted, in regard to the specific needs and the type of the application that each system aims to support. However, there are hybrid architectures, which adopt the simple client-server model with peer-to-peer communication among groups of servers or with server hierarchies, where certain servers act as clients to others. In addition, the client-server and peer-to-peer structures can be integrated into peer-server architectures, where some data packets are transmitted through certain nodes using peer-to-peer while other data are transmitted though a server.

Technologies and Protocols

This subsection presents some of the commonly used technologies for the creation of 3D content as well as the protocols available for the support of the networking feature of the NVEs.

3D Internet Technologies for NVEs

There is a large number of technologies for the development of 3D content, each of which provides certain functionality. Some of the most know 3D technologies are (Diehl, 2001): VRML, Extensible 3D (X3D) and Java3D API. These technologies vary on the way an object/model is represented, on their ability to support animations, whether they provide a programming interface, whether they support streaming, and so forth. It becomes clear that the selection of an appropriate technology depends on the needs and requirements of the application developed. The main standard in this area is X3D, which is the open standard for Web-delivered three-dimensional graphics. It specifies a declarative geometry definition language, a run-time engine, and an application programming interface (API) that provide an interactive, animated, real-time environment for 3D graphics (Daly & Brutzman, 2007). As described in Bouras, Panagopoulos, and Tsiatsos (2005) there are some X3D enabled NVEs platforms as well as possible solutions for migrating from a VRML based multiuser platform to X3D available.

Protocols

The protocols used for the support of NVEs depend mainly on the networking solution that each system adopts. For NVEs the protocols most commonly used are the following: at the network layer the Internet Protocol (IP) and at the Transport Layer the Transmission Control Protocol (TCP), the User Datagram Protocol (UDP) and the Multicast IP protocol. It should also be mentioned that for Distributed Virtual Environments, which constitute a subset of NVEs there are additional protocols, which meet the specific needs of this type of applications and are the following: the Distributed Interactive Simulation (DIS) protocol, the Distributed Worlds Transfer and Communication Protocol (DWTP) (Broll, 1997), the Multi-User 3D Protocol (Mu3D) (Galli & Luo, 2000) and the Virtual Reality Transfer Protocol (VRTP) (Brutzman, Zyda, Watsen & Macedonia, 1997). As stated in (Diehl, 2001) there is no protocol able to serve all types of applications equally. Thus, based on the type and requirements of the developed application the appropriate protocol should be adopted for optimized performance and results.

Design and Development Challenges

The complexity of NVEs is mainly related to the need and desire to achieve a high-sense of realism. This fact results in applications that need to include multiple traditional software types, rich graphics, and compatibility with other applications. The networked nature of these environments is an additional factor that affects their complexity, in terms both of the development and deployment. In particular, NVE development is a difficult balancing act of trade-offs, as there are a number of factors that should be taken into account for optimizing the networking performance of the system (Diehl, 2001). These factors are: (a) the network bandwidth, which constitutes a limited resource and therefore the allocation of its capacity should be carefully determined, (b) heterogeneity, which is related to the quality of service that users with diverse equipment (e.g., processing system, network connection, graphic resolution) can achieve, (c) distributed interaction, which is related to the fact that the system must provide each user with the illusion that the entire environment is located on the local machine and that the actions of the users have a direct and immediate impact on the environment, (d) the real-time system design and resource management,

which defines the process and thread architecture of the application, (e) the failure management that concerns the reaction of the system in a possible failure and its impact on the users' view, (f) the scalability, which is related to the need for supporting a larger number of concurrent users and finally (g) the deployment and configuration, in terms of how the software will be accessible by the end users. It is very difficult to determine a formula that can satisfy all the aforementioned factors and resolve the limitations that each of them introduces, as the dependency among them is strong and improving one's behavior can affect other component's behavior as well (Singhal & Zyda, 1999). Therefore, based on the specific type of the application as well as its target group the developers need to specify their priorities for the design and development.

FUTURE TRENDS

As stated previously, NVEs are complex systems, which incorporate a number of applications and different technologies. In particular, the NVEs currently developed are prototyping the information infrastructure of the next century in terms of advanced networking, virtual reality, high performance computing, data mining, and human/computer interactions. Thus, there is a wide range of areas that can be further developed and improved for the optimization of these environments and their wider adaptation. Based on the fact that NVEs allow multiple participants to collaborate using high-speed networks connecting heterogeneous computing resources and large data stores, NVEs could further extend the human/computer paradigm so as to include human/computer/human collaborations. Another direction that draws increased interest for NVEs is the ability to efficiently support large-scale applications. The term "large-scale" refers both to the data size (in terms of virtual space and graphics) as well as to the concurrent number of users that can participate (Bouras, Giannaka, Panagopoulos & Tsiatsos, 2006). To this direction, research has already begun producing techniques and algorithms for achieving this challenging task. Moreover, the need for an advanced sense of realism seems to emerge, especially where the relationship between the virtual world and the everyday physical world is concerned (Benford, Greenhalgh Rodden & Pycock, 2001) while ubiquitous, mobile, and wearable computing promises to make access to digital information universal and continual. Finally, many ideas and technological solutions could be adopted by 3D games technology in order to use these environments to support other applications. As Zyda (2007) said, "the same technology that makes interactive 3D games so entertaining in the physical action domain is just as effective in education, training, and other more serious applications."

CONCLUSION

In this chapter we presented the basic issues of NVEs. The areas covered were: the basic characteristic and components of NVEs, the architectures, technologies and protocols available for their development as well as some design and development issues that should be taken into account when designing and developing a NVE. It is obvious that, as technological challenges are overcome, NVE systems tend to become more and more powerful communication and collaboration tools on various fields of interest.

REFERENCES

Benford S., Greenhalgh, C., Rodden, T., & Pycock, J. (2001). Collaborative virtual environments. *Communications of the ACM, 44*(7), 79 85.

Bouras, C., Panagopoulos, A., & Tsiatsos, T. (2005). Advances in X3D multi-user virtual environments. In *Proceedings of the 7th IEEE International Symposium on Multimedia*.

Bouras, C., Giannaka, E., Panagopoulos, A., & Tsiatsos, T. (2006), Distribution and partitioning techniques for NVEs: The case of EVE. In *Proceedings of the Challenges of Large Applications in Distributed Environments.* Paris, France.

Broll, W. (1997). Populating the internet: Supporting multiple users and shared applications with VRML. In *Proceedings of the 2nd Symposium on Virtual Reality Modeling Language* (p. 33). Monterey, CA.

Brutzman, D., Zyda, M., Watsen, K., & Macedonia, M. (1997). Virtual reality transfer protocol (VRTP) design rationale. In *Proceedings of the 6th Workshop on Enabling Technologies on Infrastructure for Collaborative Enterprises* (pp.179-186).

Daly, L. & Brutzman, D. (2007). X3D: Extensible 3D graphics standard. *Signal Processing Magazine, 24*(6), 130-135.

Diehl, S. (2001). *Distributed virtual worlds.* Springer.

Galli, R. & Luo, Y. (2000). Mu3D: A causal consistency protocol for a collaborative VRML editor. In *Proceedings of the 5th symposium on Virtual reality modeling language (Web3D-VRML)* (pp. 53-62). Monterey, CA.

Joslin, C., Pandzic, I. S., & Thalmann, N. M. (2003). Trends in networked collaborative virtual environments. *Computer Communication Journal, 26*(5), 430-437.

Joslin, C., Di Giacomo, T., & Magnenat-Thalmann, N. (2004). Collaborative virtual environments: From birth to standardization. *IEEE Communications Magazine, 42*(4), 28- 33.

Macedonia, M. & Zyda, M. (1997). A taxonomy for networked virtual environments. *IEEE Multimedia,4*(1), 48-56.

McGregor, D., Kapolka, A., Zyda, M., & Brutzman, D. (2003). Requirements for large-scale networked virtual environments. In *Proceedings of the 7th International Conference on Telecommunications ConTel 2003* (pp. 353-358). Zagreb, Croatia.

Singhal, S. & Zyda, M. (1999). *Networked virtual environments: Design and implementation*. ACM Press.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer, 38*(9), 25- 32.

Zyda, M. (2007). Introduction: Creating a science of games. *Communications of the ACM, 50,* (7), 26 – 29.

KEY TERMS

CVE: Collaborative Virtual Environment is an extension of a NVE which aims at a collaborative task. CVEs aim to provide an integrated, explicit and persistent context for cooperation that combines both the participants and their information into a common display space. These objectives create the potential to support a broad range of cooperative applications such as training.

DIS: Distributed Interactive Simulation is an open standard for conducting real-time platform-level wargaming across multiple host computers and is used worldwide especially by military organizations but also by other agencies such as those involved in space exploration and medicine.

DVE: Distributed Virtual Environment is an NVE where active parts of the virtual environment are scattered to different computers, which are connected through the network.

HLA: High Level Architecture is a general purpose architecture for distributed computer simulation systems. Using HLA, computer simulations can communicate to other computer simulations regardless of the computing platforms.

Java 3D API: The Java 3D API is a hierarchy of Java classes which serve as the interface to a sophisticated three-dimensional graphics and sound rendering system. Java 3D provides high-level constructs to create and manipulate 3D geometry, and to build the structures used to render that geometry.

NVE: Networked Virtual Environment is a virtual environment that allows a group of geographically separated users to interact in real time

X3D: Extensible 3D is the open standard for Web-delivered 3D graphics. It specifies a declarative geometry definition language, a run-time engine, and an application programming interface that provide an interactive, animated, real-time environment for 3D graphics.

VE: Virtual Environment is a computer-generated simulation with which the user can interact in such a way that he receives real time feedback aiming to provide its users with a sense of realism.

VR: Virtual reality is a technology which allows a user to interact with a computer-simulated environment.

VRML: Virtual Reality Modeling Language is a standard file format for representing 3D interactive vector graphics, designed particularly with the World Wide Web in mind.

Neural Networks for Automobile Insurance Pricing

Ai Cheo Yeo
Monash University, Australia

INTRODUCTION

In highly competitive industries, customer retention has received much attention. Customer retention is an important issue, as loyal customers tend to produce greater cash flow and profits, are less sensitive to price, bring along new customers and do not require any acquisition or start-up costs.

BACKGROUND

Various techniques have been used to analyse customer retention. Eiben, Koudijs and Slisser (1998) applied genetic programming, rough set analysis, Chi-square Automatic Interaction Detection (CHAID) and logistic regression analysis to the problem of customer retention modelling, using a database of a financial company. Models created by these techniques were used to gain insights into factors influencing customer behaviour and to make predictions on customers ending their relationship with the company. Kowalczyk and Slisser (1997) used rough sets to identify key factors that influence customer retention of a mutual fund investment company. Ng, Lui and Kwah (1998) integrated various techniques such as decision-tree induction, deviation analysis and multiple concept-level association rules to form an intuitive approach to gauging customers' loyalty and predicting their likelihood of defection.

Mozer and his co-researchers (2000) explored techniques from statistical machine learning to predict churn and based on these predictions to determine what incentives should be offered to subscribers of wireless telecommunications to improve retention and maximise profitability of the carrier. The techniques included logit regression, decision trees, neural networks and boosting. Besides Mozer and his co-researchers, others have also applied neural networks to churn prediction problems. Behara and Lemmink (1994) used the neural network approach to evaluate the impact of quality improvements on a customer's decision to remain loyal to an auto-manufacturer's dealership. Wray and Bejou (1994) examined the factors that seem to be important in explaining customer loyalty. They found that neural networks have a better predictive power than the conventional analytic techniques such as multiple regression. Smith, Willis and Brooks (2000) also found that neural networks provided the best results for classifying insurance policy holders as likely to renew or terminate their policies compared to regression and decision tree modelling.

PREDICTING RETENTION RATES

We have also used neural networks to learn to distinguish insurance policy holders who are likely to terminate their policies from those who are likely to renew in order to predict the retention rate prior to price sensitivity analysis. Policy holders of an Australian motor insurance company are classified into 30 risk groups based on their demographic and policy information using k-means clustering (Yeo, Smith, Willis & Brooks, 2001, 2003). Neural networks are then used to model the effect of premium price change on whether a policy holder will renew or terminate his or her policy. A multilayered feedforward neural network was constructed for each of the clusters with 25 inputs and 1 output (whether the policy holder renews or terminates the contract).

Several experiments were carried out on a few clusters to determine the most appropriate number of hidden neurons and the activation function. Twenty hidden neurons and the hyperbolic tangent activation function were used for the neural networks for all the clusters. A uniform approach is preferred to enable the straight-forward application of the methodology to all clusters, without the need for extensive experimentation by the company in the future. Input variables that were skewed were log transformed.

Some of the issues we encountered in using neural networks to determine the effect of premium price change on whether a policy holder will renew or terminate his or her policy were:

- Determining the threshold for classifying policy holders into those who terminate and those who renew
- Generating more homogenous models
- Clusters that had too few policy holders to train the neural networks

Determining Threshold

The neural network produces output between zero and one, which is the probability that that a policy holder will ter-

minate his or her policy. Figure 1 shows the probability of termination of Cluster 11. A threshold value is used to decide how to categorise the output data. For example a threshold of 0.5 means that if the probability of termination is more than 0.5, then the policy will be classified as terminated. Usually the decision threshold is chosen to maximise the classification accuracy. However, in our case we are more concerned with achieving a predicted termination rate that is equal to the actual termination rate. This is because we are more concerned with the performance of the portfolio (balancing market share with profitability) rather than whether an individual will renew or terminate his or her policy. The actual termination rate for cluster 11 is 14.7%. To obtain a predicted termination rate of 14.7%, the threshold was set at 0.204 (see Figure 1). The confusion matrix for a threshold of 0.204 is shown in Table 1. The overall classification accuracy is 85.3%.

Generating more homogeneous models

The confusion matrix provides the prediction accuracy of the whole cluster. It does not tell us how a given percentage change in premium will impact termination rates. To determine how well the neural networks were able to predict termination rates for varying amounts of premium changes, the clusters were then divided into various bands of premium as follows: decrease in premiums of less than 22.5%, premium decrease between 17.5% and 22.5%, premium decrease between 12.5% and 17.5% and so on. The predicted termination rates were then compared to the actual termination rates. For all the clusters the prediction accuracy of the neural networks starts to deteriorate when premium increases are between 10% and 20%. Figure 2 shows the actual and predicted termination rates for one of the clusters (Cluster 24).

In order to improve the prediction accuracy, the cluster was then split at the point when prediction accuracy starts to deteriorate. This is to isolate those policy holders with a significant increase in premium. It is believed that these policy holders behave differently due to a greater number of these policy holders who have upgraded their vehicles. Two separate neural networks were trained for each cluster. The prediction accuracy improved significantly with two neural networks as can be seen from Figure 3. The average absolute deviation decreased from 10.3% to 2.4%.

Combining Small Clusters

Some of the smaller clusters had too few policy holders to train the neural networks. We grouped the small clusters that had fewer than 7,000 policies. The criterion for grouping was similarity in risk. Risk in turn is measured by the amount of claims. Therefore the clusters were grouped according to similarity in claim cost. The maximum difference in average claim cost per policy was no more than $50. For the combined clusters, prediction ability is also improved by having two neural networks instead of one for each cluster.

PRICE SENSITIVITY ANALYSIS

Having trained neural networks for all the clusters, sensitivity analysis was then performed on the neural networks to determine the effect of premium changes on termination rates for each cluster. There are several ways of performing the sensitivity analysis:

One approach is based on systematic variation of variables (SVV). To determine the impact that a particular input variable has on the output, we need to hold all the other inputs to some fixed value and vary only the input of interest while we monitor the change in outputs (Anderson, Aberg & Jacobsson, 2000; Bigus, 1996).

A more automated approach is to keep track of the error terms computed during the backpropagation step. By computing the error all the way back to the input layer, we have

Figure 1. Determining the threshold value of the neural network output

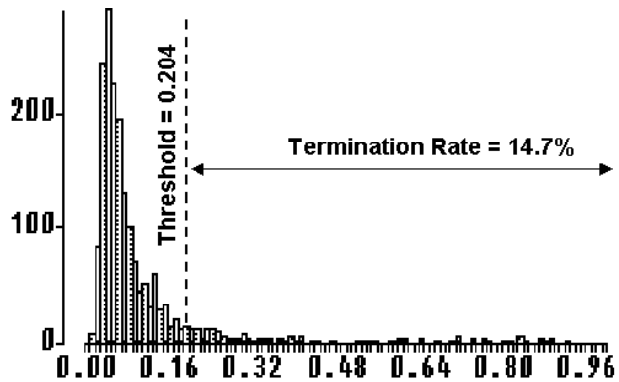


Table 1. Confusion matrix for cluster 11 with decision threshold = 0.204

	Classified as		
Actual	Terminated	Renewed	Total
Terminated	841 (50.0%)	842 (50.0%)	1,683
Renewed	845 (8.6%)	8,935 (91.4%)	9,780
Total	1,686	9,777	11,463
Overall Accuracy			85.3%

Figure 2. Prediction accuracy for one neural network model of cluster 24

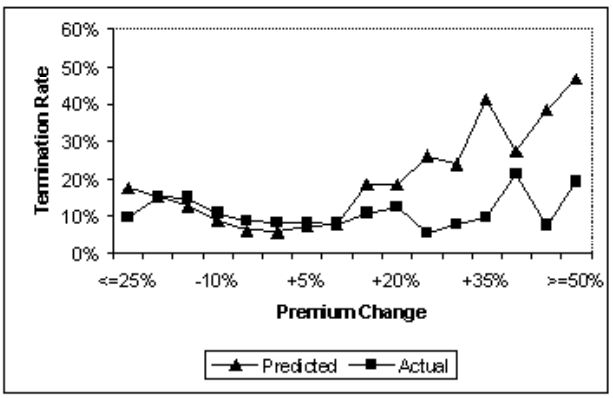


Figure 3. Prediction accuracy for two networks model of cluster 24

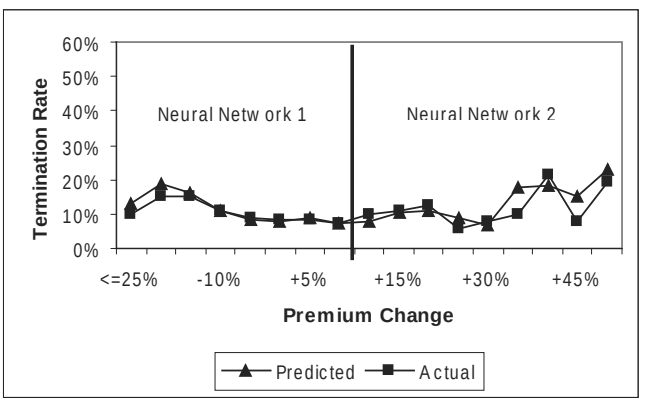
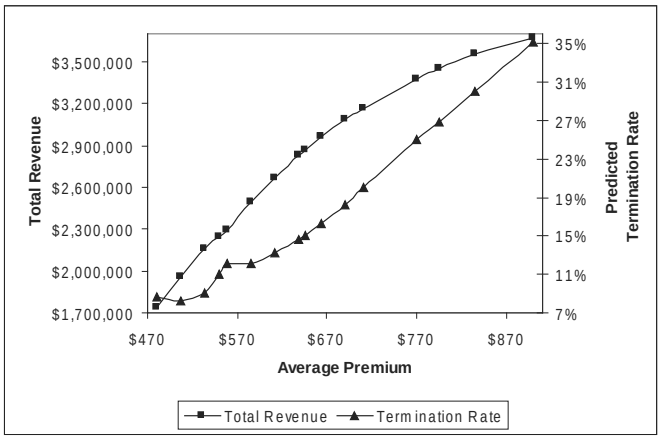


Table 2. Price sensitivity analysis – effect on termination rate (Cluster 24)

Scored Against	Average Change in Premium	Average Premium Amount ($)	Termination Rate
Neural Network 1	-8.3%	481	8.6%
	-3.3%	507	8.3%
	1.7%	533	9.1%
	6.7%	559	12.2%
Neural Network 2	11.7%	585	12.1%
	16.7%	612	13.2%
	21.7%	638	14.6%
	26.7%	664	16.3%
	31.7%	690	18.2%
	51.7%	795	26.9%
	71.7%	900	35.1%

Figure 4. Price sensitivity analysis - effect on termination rate and revenue (Cluster 24)



a measure of the degree to which each input contributes to the output error (Bigus, 1996).

The third approach is based on sequential zeroing of weights (SZW) of the connection between the input variables and the first hidden layer of the neural network (Anderson et al., 2000).

In their research, Anderson and his co-researchers (2000) found that neural networks are suitable not only for function approximation of nonlinear relationship but are also able to represent to a high degree the nature of input variables. The information generated about the variables using the SVV and SZW methods can serve as a guide to the interpretation of influence, contribution and selection of variables. We performed price sensitivity using the SVV approach by varying the premium information and holding all other inputs constant.

Separate data sets were created from each "half" cluster with all variables remaining unchanged except the new premium and related variables (change in premium, percentage change in premium and ratio of new premium to new sum insured). For example, cluster 24 was split into two "halves"; policy holders with premium decreases or increases of less than 10% and policy holders with premium increases of more than 10%. Data sets with varying percentage changes in premium were created and scored against the trained neural networks to determine the predicted termination rates. Results of price sensitivity analysis for Cluster 24 are shown in Table 2 and Figure 4 as an example. They show that as premiums increase the termination rates also increase.

CONCLUSION AND FUTURE RESEARCH

We randomly selected a few clusters to carry out experiments to determine the appropriate neural network architecture for the prediction of retention rates. More experiments could be

performed to find an appropriate neural network architecture for each of the clusters so that prediction accuracy could be improved. However, the insurance industry is unlikely to implement the proposed approach if it is difficult to implement. We have therefore adopted a straightforward approach that requires minimal fine tuning for specific data. While this research has been focused on premium pricing within the insurance industry, the methodology developed is quite general. In fact, the approach can be applied to any industry concerned with setting prices for products in competitive environments. This includes sectors such as retail and telecommunications.

REFERENCES

Anderson, F.O., Aberg, M., & Jacobsson, S.P. (2000). Algorithmic approaches for studies of variable influence, contribution and selection in neural networks. *Chemometrics and Intelligent Laboratory Systems, 51,* 61-72.

Behara, R.S., & Lemmink, J. (1994). Modelling the impact of service quality on customer loyalty and retention: A neural network approach. *1994 Proceedings Decision Sciences Institute, 3,* 1883-1885.

Bigus, J.P. (1996). *Data mining with neural networks: Solving business problems—from application development to decision support.* New York: McGraw-Hill.

Eiben, A., Koudijs, A., & Slisser, F. (1998). *Genetic modelling of customer retention.* Paper presented at the Genetic Programming First European Workshop, EruoGP'98, Paris, France.

Kowalczyk, W., & Slisser, F. (1997). *Modelling customer retention with rough data sets.* Paper presented at the Principles of Data Mining and Knowledge Discovery. First European Symposium, PKDD '97, Trondheim, Norway.

Mozer, M.C., Wolniewicz, R., Grimes, D.B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunication. *IEEE Transactions on Neural Networks, 11*(3), 690-696.

Ng, K.S., Lui, H., & Kwah, H.B. (1998). *A data mining application: Customer retention at the Port of Singapore Authority (PSA).* Paper presented at the 1998 ACM SIGMOD International Conference on Management of Data, Seattle, WA, USA.

Smith, K.A., Willis, R.J., & Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: A case study. *Journal of the Operational Research Society, 51*(5), 532-541.

Wray, B., & Bejou, D. (1994). *An application of artificial neural networks in marketing: Determinants of customer loyalty in buyer-seller relationships.* Paper presented at the Proceedings of Decision Sciences Institute 1994 Annual Meeting, Honolulu, HI.

Yeo, A., Smith, K., Willis, R., & Brooks, M. (2001). Clustering technique for risk classification and prediction of claim costs in the automobile insurance industry. *International Journal of Intelligent Systems in Accounting, Finance and Management, 10*(1), 39-50.

Yeo, A.C., Smith, K.A., Willis, R.J., & Brooks, M. (2003). A comparison of soft computing and traditional approaches for risk classification and claim cost prediction in the automobile insurance industry. In V. Kreinovich (Ed.), *Soft computing in measurement and information acquisition* (pp. 249-261). Heidelberg: Physica-Verlag.

KEY TERMS

Activation Function: Transforms the net input of a neural network into an output signal, which is transmitted to other neurons.

Association Rules: Predict the occurrence of an event based on the occurrences of another event.

Backpropagation: Method for computing the error gradient for a feedforward neural network.

Boosting: Generates multiple models or classifiers (for prediction or classification), and to derive weights to combine the predictions from those models into a single prediction or predicted classification.

Chi-Square Automatic Interaction Detection (CHAID): A decision tree technique used for classification of a data set. CHAID provides a set of rules that can be applied to a new (unclassified) data set to predict which records will have a given outcome. CHAID segments a data set by using chi square tests to create multi-way splits.

Confusion Matrix: Contains information about actual and predicted classifications done by a classification system.

Decision Trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.

Deviation Analysis: Locates and analyses deviations from normal statistical behavior.

Genetic Programming: Search method inspired by natural selection. The basic idea is to evolve a popula-

tion of "programs" candidates to the solution of a specific problem.

K-means Clustering: An algorithm that performs disjoint cluster analysis on the basis of Euclidean distances computed from variables and randomly generated initial seeds.

Logistic/Logit Regression: A technique for making predictions when the dependent variable is a dichotomy, and the independent variables are continuous and/or discrete.

Multi-Layered Feedforward Network: A layered neural network in which each layer only receives inputs from previous layers.

Multiple Concept-Level Association Rules: Extend association rules from single level to multiple levels. Database contents are associated together to the concepts, creating different abstraction levels.

Multiple Regression: A statistical technique that predicts values of one variable on the basis of two or more other variables.

Neural Network: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

Neural Network Architecture: A description of the number of layers in a neural network, each layer's transfer function, the number of neurons per layer, and the connections between layers

Neuron: The basic processing element of a neural network.

Regression Analysis: A statistical technique used to find relationships between variables for the purpose of predicting future variables.

Rough Sets: Rough sets are mathematical algorithms that interpret uncertain, vague, or imprecise information.

Sequential Zeroing of Weights: A sensitivity analysis method that involves sequential zeroing of weights of the connection between the input variables and the first hidden layer of the neural network.

Statistical Machine Learning: An approach to machine intelligence that is based on statistical modeling of data.

Systematic Variation of Variables: A sensitivity analysis method whereby the input of interest is varied while holding all other inputs to some fixed value to determine the impact that particular variable has on the output.

Weights: Strength of a connection between two neurons in a neural network.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2095-2099, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Neural Networks for Intrusion Detection

Rui Ma
Beijing Institute of Technology, China

INTRODUCTION

With the rapid expansion of computer networks, network security has become a crucial issue for modern computer systems. As an important and active defense technology, the intrusion detection system (IDS) plays an important role in defensive systems. IDSs provide real-time protection from interior attacks, exterior attacks, and invalid operations, and it can intercept intrusions and respond whenever the network system integrity is violated (Ma, 2004). Many intrusion detection approaches have been deeply researched and some widely deployed. But the diversification, complexity, and scale of intrusions raise new demands for IDSs. Neural networks are tolerant of imprecise data and uncertain information. With their inherent ability to generalize from learned data they seem to be an appropriate approach to IDSs (Hofmann, Schmitz, & Sick, 2003). This article discusses the detection of distributed denial-of-service (DDoS) attacks using artificial neural networks techniques. The implementation of a distributed intelligent intrusion detection system (DIIDS) is described, including both the data processing technique and neural networks approaches adopted.

BACKGROUND

Intrusion Detection System

Many IDSs are based on the general model proposed by Denning (1987). This model is independent of platform, system vulnerability, and type of intrusion. It maintains a set of historical profiles for users, matches an audit record with the appropriate profile, updates the profile whenever necessary, and reports any attacks detected.

IDSs can be divided into two types: (1) host-based IDSs and (2) network-based IDSs. Host-based IDSs evaluate information found on a single or multiple host systems, including contents of operating systems, system logs and application files. Network-based IDSs evaluate information captured from network communications, by analyzing the stream of packets traveling across the network. Packets are captured through a set of sensors placed at strategic points in the network (Jean & Philippe, 2001).

Intrusion detection schemes can be classified into two general categories (Ghosh, 1999a): (1) misuse detection and (2) anomaly detection. Misuse detection techniques assume that all kinds of intrusion behavior can be described as specific patterns, thus allowing the identification of intrusive behavior by comparing current user activity with specific patterns that have been observed previously during an attack. The most significant advantage of misuse detection techniques is that known attacks can be detected fairly reliably and with a low false positive rate. However, the key limitation of misuse detection techniques is that they cannot detect novel attacks.

Anomaly detection techniques assume that all kinds of intrusion behavior differ from normal user activities. Any current user behavior sufficiently deviant from the normal user activities will be flagged as anomalous and hence considered as a possible attack. The most significant advantage of anomaly detection techniques is that it directly addresses the problem of detecting novel attacks against systems. However, the most notable disadvantage of anomaly detection techniques is the high rates of false alarm.

In order to detect known attacks, subtle variations of known attacks, and novel attacks efficiently, IDSs should selectively combine aspects of both misuse detection techniques and anomaly detection techniques (Chen, 2004).

Neural Networks

An artificial neural network consists of a collection of processing elements that are highly interconnected and that transform a set of inputs to a set of desired outputs. The result of the transformation is determined by the characteristics of the elements and the weights associated with the interconnections among them. By modifying the connections between the nodes, the network is able to adapt to the desired outputs.

The neural network gains the necessary experience initially by being trained to correctly identify preselected examples of a problem. The response of the neural network is reviewed and the configuration of the system is refined until the neural network's analysis of the training data reaches a satisfactory level. In addition to the initial training period, the neural network also gains experience over time as it conducts analyses on data related to the problem.

The training algorithms of neural networks can be classified into two general categories: (1) supervised and (2) unsupervised. In the learning phase, a supervised algorithm

learns the desired output for a given input or pattern. Whereas an unsupervised algorithm learns without specifying the desired output (Jean & Philippe, 2001).

Neural Network Intrusion Detection Systems

Neural networks have capabilities of self-learning, self-organization, and self-adaptivity; as well as a capability to analyze fuzzy, nonstructured, imprecise, incomplete data and to generalize from previously observed behaviors. With the continuous development of network technologies and increasing multiplicity and novelty of network attacks, IDSs must be more flexible and efficient. Artificial neural networks offer the potential to resolve a number of the problems encountered by the other approaches to intrusion detection.

Various approaches to using neural networks for intrusion detection have been advocated. One approach is to create keyword, count-based, misuse detection systems with neural networks (Lippmann & Cunningham, 1999; Ryan, Lin, & Miikkulainen, 1998). The data that are presented to the neural network consist of attack-specific keyword counts in network traffic. Such an approach is close to a host-based IDS. Some researchers created a neural network to analyze program behavior profiles instead of user behavior profiles (Ghosh, 1999b). This method identifies the normal system behavior of certain programs and compares it to the current system behavior. Cannady (1998) developed a network-based neural network IDS in which packet-level network data were retrieved from a database and then classified according to the packet characteristics before being presented to a neural network.

In the case of DIIDS, three basic artificial neural networks approaches are described: (1) neural network expert system, which is mainly used to detect known attacks by using forward parallel inference which improves inference efficiency; (2) back-propagation neural network; and (3) adaptive resonance theory neural network, which are used to detect not only known attacks, but variations of them or indeed, unknown attacks. Finally, based on the neural network approaches just described, cooperative intrusion detection approaches are discussed.

NEURAL NETWORKS IN INTRUSION DETECTION SYSTEMS

The Distributed Intelligence Intrusion Detection System

DIIDS is based on neural networks techniques and is characterized by the following constituent components: data collection; detection and analysis; console; and database. DIIDS can be deployed in a distributed or integrated manner. In particular, DIIDS combines the capability of host-based and network-based IDSs and uses artificial neural networks as detective techniques.

The console is responsible for the management, monitoring, and reporting status of the system. Storage of feature data and detection results is undertaken by the database. Data collection assembles and processes both network-based and host-based data before identifying feature data. Detection and analysis then determines if the attack occurs through three basic neural networks approaches: (1) neural network expert system; (2) back-propagation neural network; and (3) adaptive resonance theory neural network. In addition, the cooperative intrusion detection technique is also used.

Data Processing

Three levels of data processing are conducted. Initially, data is selected from the available data set by the data collection component. The data collection component comprises a network engine, a host agent, and a data fusion facility. The network engine collects all data packages flowing in the network and then resolves them into the feature data set for the network. Eight elements of the feature data garnered from the network are: (1) protocol type, (2) source IP address, (3) destination IP address, (4) source port, (5) destination port, (6) sequence number, (7) acknowledgement number, and (8) raw data. The host agent collects the data that characterizes system performance, for example, network traffic, memory, and CPU usage. These form the original feature data. A second phase converts the data elements such as protocol type, source, and destination IP address into a standardized numeric representation. The third part is that the process concerns the association processing.

By analyzing the feature data, both the spatial and temporal associate relationships between intrusion behaviors can be determined. Some attacks, which originate from different sources and try to attack the same specific goal, send many packages which have identical destination IP addresses and different source IP addresses. The relationship of these packages is called the *spatial associate relationship*. The *temporal associate relationship* points to many packages which have same source IP address but try to attack the same specific goal in a certain time period. In order to embody the spatial and temporal associate relationship in a concrete detective record, the associate detection algorithm (Figure 1) must be used. This algorithm processes associate information of feature data and converts the stochastic associated attributes into the detection associated attributes. Based on a variable, continuous, and sliding temporal window, the algorithm calculates the count of any distinct value of stochastic associated attributes during this time. Then the algorithm compares the maximum count with the threshold to determine the value of the associate attributes (Lee, 1999). This approach can

effectively detect the associate relationship between intrusion behaviors, reduce the complexity of intrusion detection algorithms, and improve the intrusion detection capability. After the smooth transition of the original data, a total of 13 elements of the feature data are obtained which are in turn used by each neural networks approach.

The Neural Network Expert System Approach

The neural network expert system (Alpaslan & Tolun, 1994) intrusion detection approach converts the AND nodes and OR nodes in the rule-based system into the corresponding AND nodes and OR nodes represented by neural networks and constructs the neural network that is equivalent to the original rule set. The value of the weight is 1 if and only if the correspondence input node participates in the AND/OR operation. Otherwise, the value is 0. The threshold of the AND node is the total number of all input nodes that participate in the AND operation, whereas the threshold of the OR node is the invariable value 1.

From an architectural perspective, the neural network expert system is a three-layered feed forward network with full interconnection between layers. The three layers are the input layer, the hidden layer, and the output layer. The nodes in the input layer correspond to the detective feature data. The nodes in the hidden layer correspond to the "then-clause" of rules. All hidden nodes are AND nodes. The nodes in the output layer correspond to the different kinds of attacks. All output nodes are OR nodes.

The neural network expert system solves the problem by using forward parallel inference which improves inference efficiency. This approach achieves better results when behavior is normal and the attack is of a known type. However, it is not so effective when attacks are of a kind that have not been encountered previously.

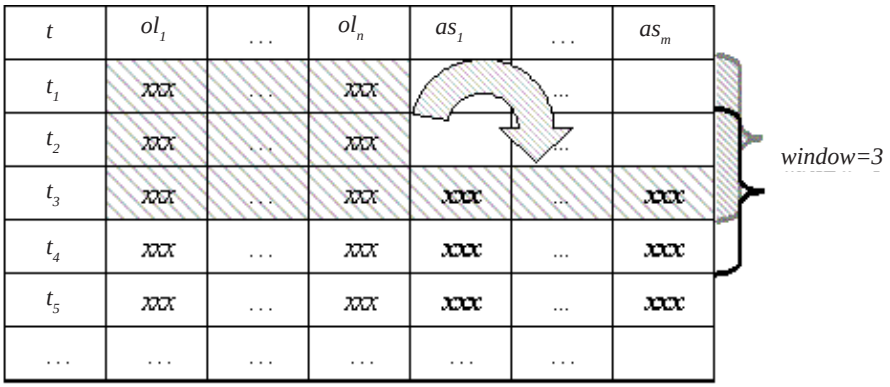
The Back-Propagation Neural Network Approach

The Back-Propagation (BP) neural network intrusion detection approach has the capability of both self-learning and self-adaptivity and can distinguish intrusion behaviors, albeit imprecisely, for both known and unknown attacks. The architecture of the BP neural network intrusion detection approach is the three-layered feed forward network with full interconnection between layers. Again, the three layers are the input layer, the hidden layer, and the output layer. There are 13 nodes in the input layer that correspond to the 13 elements of the detective feature data. There are 25 nodes in the hidden layer, which were determined by experiment. There are two nodes in the output layer that detect if an attack occurs.

First, the BP neural network translates the input feature values into continuous values using fuzzy functions. Then, it is trained using the supervised and improved BP algorithm (Hagan et al., 2002), which improves the rate of convergence through restricting the output of the sigmoid function and adjusts all weight values and thresholds separately. When the anticipant outputs are achieved, the approach calculates the distance between the real output and the anticipant output to decide which attack has occurred.

After the BP neural network is trained, it can be used to detect attacks online. This approach achieves better results for normal behavior and known attacks, but can also detect some variations of known attacks.

Figure 1. The associate detection algorithm



Legend: t-time, ol-original data with associate relationship, as-achieved associate data, window-time window

The Adaptive Resonance Theory 2 Neural Network Approach

The *adaptive resonance theory 2* (ART2) neural network (Carpenter & Grossberg, 1987) intrusion detection approach uses unsupervised learning; possesses the capabilities of self-learning and self-adaptivity; demonstrates better stability-plasticity trade-off; and can detect attacks, even unknown attacks, in real time.

The ART2 neural network intrusion detection approach adopts the competition and self-stabilization principle. It contains both attention subsystems and orienting subsystems. Attention subsystems establish the internal representation of familiar activities, while orienting subsystems respond to unfamiliar activities. Attention subsystems accomplish a competed selection from the bottom-up vectors, and a match comparison between bottom-up vectors and up-bottom vectors. Orienting subsystems verify the similarity between input vectors and anticipant output vectors. When a match value exceeds the vigilance parameter, the corresponding node is the winner.

In the ART2 neural network, there are 13 input nodes which correspond to the detective feature data, and five output nodes of which three correspond to the normal behavior and two kinds of known attacks, one to unknown attack, and one is reserved. This approach translates the input feature values into continuous values using fuzzy functions and classifies the output by adjusting the value of the vigilance parameter. This approach achieves better results during normal behavior, known attacks, and unknown attacks. However, it is particularly effective in unknown attack scenarios.

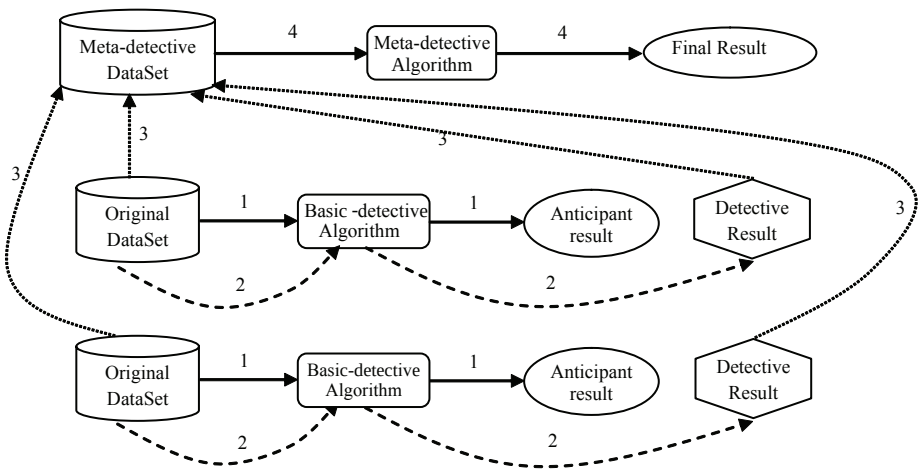
The Cooperative Intrusion Detection Approach

Any of the aforementioned detection approaches have advantages and disadvantages and can be successfully utilized for different attack modes. In order to utilize the characteristic of different detection approaches, the cooperative intrusion detection approach (Chan, 1996; Lee & Heinbuch, 2001; Parikh, 2001) is proposed (Figure 2). The kernel of the cooperative intrusion detection approach is the metadetection algorithm.

In this approach, the three individual neural network detection approaches are referred to as the basic intrusion detection algorithms. Every basic intrusion detection algorithm is a metadetector which has distributiveness, parallelism, independence, and an inherent separation of training and detection. The input data of the metadetection algorithm are the results obtained from metadetectors. The algorithm employs several detection strategies: voting, arbitration, and combing. Different metadetection algorithms are formed from combinations of these basic strategies.

Two categories of the cooperative intrusion detection approach exist: (1) local data-based and (2) the global data-based. The metadetectors in the local data-based approach are trained using local data. Instead of using global training data, the global data-based approach shares metadetectors (Figure 3). These metadetectors are trained by local data in one area and then used to detect attacks in another area. In this case, the metadetectors are referred to as remote metadetectors. Depending on the characteristics of neural networks, especially the ART2 neural network, the global data-based approach shares the global training data when it shares the

Figure 2. The architecture of cooperative intrusion detection approach



training results of the remote metadetectors. This approach solves the scalability issue and achieves good results.

FUTURE TRENDS

The most effective learning algorithms of neural networks need to be considered. The neural network expert system should have a learning capability. It is necessary to reduce the retraining time of the BP neural network when user behavior changes.

In order to have a highly adaptive capability to detect unknown attacks, research in self-organizing maps (SOMs) which offer an increased level of adaptability are needed. ART2 and SOM should be compared, and an improvement in both adaptability and detection capability for unknown attacks is required.

Because of the extensive number of vulnerabilities in computer systems and the creativity of attackers, intrusions have become increasingly diverse and complex. Using the cooperative intrusion detection approach as a basis, it is planned to develop a hierarchical IDS and research how efficiency may be improved.

CONCLUSION

The distributed intelligence intrusion detection system has been deployed in a small test network. After collecting and analyzing the feature data, it was characterized into four parts: (1) normal, (2) Trinoo attack, (3) TFN2K attack, and (4) unknown attacks. Normal means there is no attack. Trinoo and TFN2K are known attacks and have appeared in the training data. Unknown attacks have not appeared in the training data. The results show that known attacks can be detected better by the neural network expert system; unknown attacks can be detected better by the BP neural network and the ART2 neural network, especially the ART2. In the cooperative intrusion detection technique, the global data-based approach achieves promising results.

REFERENCES

Alpaslan, F. N., & Tolun, M. R. (1994). Connectionist expert system. *Artificial Neural Networks and Artificial Life Symposium,* METU Ankara, Turkey, December 15.

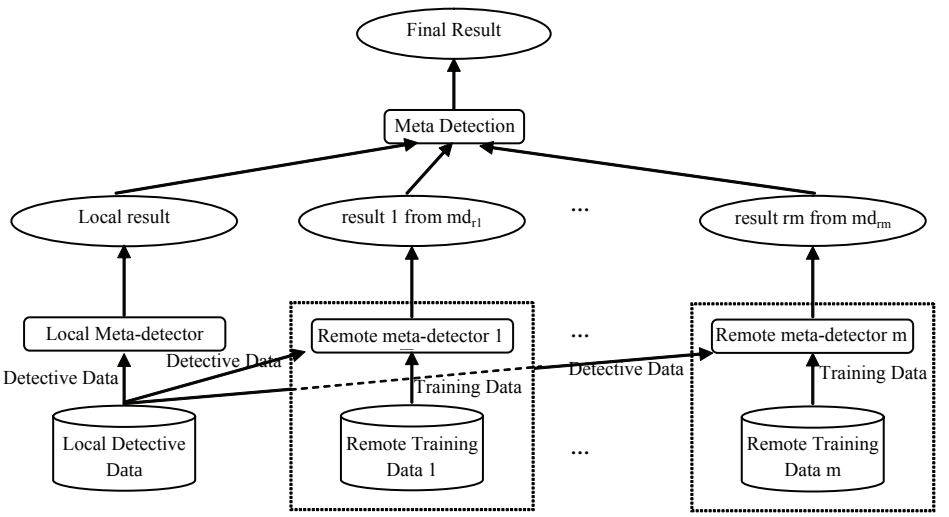
Cannady, J. (1998, October 5-8). Artificial neural networks for misuse detection. *The 21st National Information Systems Security Conference, Arlington* (pp. 392-397).

Carpenter, G. A., & Grossberg, S. (1987). ART2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics, 26*(23). 4919-4930.

Chan, P. K. (1996). *An extensible meta-learning approach for scalable and accurate inductive learning.* New York: Columbia University.

Chen, T. M. (2004). Intrusion detection for virus and worms. *IEC annual review of communications,* Chicago (Vol. 57). Chicago, IL: International Engineering Consortium. Retreived from http://engr.smu.edu/^tchen/papers/iec2004.pdf

Figure 3. The global data-based cooperative intrusion detection



Denning, D. E. (1987). An intrusion detection model. *IEEE Transactions on Software Engineering, 13*(2), 222-232.

Ghosh, A. K., & Schwartzbard, A. (1999a, August 23-26). A study in using neural networks for anomaly and misuse detection. *Proceedings of the 8th USENIX Security Symposium,* Washington, DC.

Ghosh, A., Schwartzbard, A., & Shatz, M. (1999b, April 9-12). Learning program behavior profiles for intrusion detection. *Proceedings of the 1st USENIX Workshop on Intrusion Detection and Network Monitoring,* Santa Clara, CA.

Hagan, M. T., Demuth, H. B., & Beale, M. H. (2002). *Neural networks design.* Beijing, China: China Machine Press.

Hofmann, A., Schmitz, C., & Sick, B. (2003, June 26-29). Intrusion detection in computer networks with neural and fuzzy classifiers. *International Conference on Artificial Neural networks ICANN 2003,* Istanbul, Turkey.

Jean, & Philippe. (2001). *Application of neural networks to intrusion detection.* Bethesda, MD: SANS Institute. Retreived from http://www.sans.org/reading-room/whitepapers/detection/336.php

Lee, W. (1999). *A data mining framework for constructing features and models for intrusion detection systems.* New York: Columbia University.

Lee, S. C., & Heinbuch, D. V. (2001). Training a neural-network based intrusion detector to recognize novel attacks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 31*(4), 294-299.

Lippmann, R. P., & Cunningham, R. K. (1999, September 7-9). Improving intrusion detection performance using keyword selection and neural networks. *Web Proceedings of the 2nd International Workshop on Recent Advances in Intrusion Detection (RAID'99),* Purdue University, West Lafayette, IN.

Ma, R. (2004). *Research of intrusion detection technologies based on fuzzy neural networks.* Beijing, China: Beijing Institute of Technology.

Parikh, S. (2001). *A framework of system integrator for MAIDS.* Ames: Iowa State University.

Ryan, J., Lin, M. J., & Miikkulainen, R. (1998, December 2-4). Intrusion detection with neural networks. In *Proceedings of neural information processing systems (NIPS '97)*, Denver, CO.

KEY TERMS

Adaptive Resonance Theory: This is a kind of neural network. The basic ART system is an unsupervised learning model and typically consists of comparison and recognition fields (one each) of neurons, a vigilance parameter, and a reset module. There have been several types. ART2 supports continuous inputs.

Artificial Neural Networks: A type of artificial intelligence that attempts to imitate the way a human brain works. Rather than using a digital model, in which all computations manipulate zeros and ones, a neural network works by creating connections between processing elements, and the organization and weights of the connections determine the output.

Back-Propagation Neural Network: This is a kind of feed forward neural network. It consists of multiple layers of computational units and is fully interconnected between layers. Each neuron in one layer has directed connections to the neurons of the subsequent layer. It usually applies the sigmoid function as an activation function.

Distributed Denial of Service Attack (DDoS): A DDoS attack is an attack on a computer system or network that causes a loss of service to users, typically the loss of network connectivity and services by consuming the bandwidth of the victim network or by overloading the computational resources of the victim system.

Intrusion Detection System (IDS): An IDS is a security system that monitors computer systems and network traffic. It analyzes the traffic for possible hostile attacks originating from outside the organization as well as for system misuse, and attacks originating from inside the organization.

Neural Network Expert System: Expert systems are an artificial intelligence application that uses a knowledge base of human expertise for problem solving. In a neural network expert system, the knowledge is encoded in the weight, and the artificial neural network generates inference rules.

Tribe Flood Network 2000 (TFN2K): This is a kind of distributed DDoS attack. TFN2K uses a client/server mechanism where a client issues commands simultaneously to a set of TFN2K servers. The servers then conduct the DDoS attacks against the victim(s).

Trinoo: This is a kind of distributed DDoS attack. Trinoo is the attack server. Trinoo waits for a message from a remote system and, upon receiving the message, launches a DDoS attack against a third party.

Category: Business Information Systems

Neural Networks for Retail Sales Forecasting

G. Peter Zhang
Georgia State University, USA

INTRODUCTION

Forecasting of the future demand is central to the planning and operation of retail business at both macro and micro levels. At the organizational level, forecasts of sales are essential inputs to many decision activities in various functional areas such as marketing, sales, and production/purchasing, as well as finance and accounting (Mentzer & Bienstock, 1998). Sales forecasts also provide basis for regional and national distribution and replenishment plans. The importance of accurate sales forecasts for efficient inventory management has long been recognized. In addition, accurate forecasts of retail sales can help improve retail supply chain operations, especially for larger retailers who have a significant market share. For profitable retail operations, accurate demand forecasting is crucial in organizing and planning purchasing, production, transportation, and labor force, as well as after sales services.

Barksdale and Hilliard (1975) examined the relationship between retail stocks and sales at the aggregate level and found that successful inventory management depends to a large extent on the accurate forecasting of retail sales. Agrawal and Schorling (1996) and Thall (1992) also pointed out that accurate demand forecasting plays a critical role in profitable retail operations, and poor forecasts would result in too-much or too-little stocks that directly affect revenue and competitive position of the retail business. The importance of accurate demand forecasts in successful supply chain operations and coordination has been recognized by many researchers (Chopra & Meindl, 2007; Lee, Padmanabhan, & Whang, 1997).

Retail sales often exhibit both seasonal variations and trends. Historically, modeling and forecasting seasonal data is one of the major research efforts, and many theoretical and heuristic methods have been developed in the last several decades. Different approaches have been proposed, but none of them has reached consensus among researchers and practitioners. Until now, the debate is still not abated in terms of what the best approach to handle the seasonality is.

On the other hand, it is often not clear how to best model the trend pattern in a time series. In the popular Box-Jenkins approach to time series modeling, differencing is used to achieve stationarity in the mean. However, Nelson and Plosser (1982) and Pierce (1977) argued that differencing is not always an appropriate way to handle trend, and linear detrending may be more appropriate. Depending on the nature of the non-stationarity, a time series may be modeled in different ways. For example, a linear or polynomial time trend model can be used if the time series has a deterministic trend. On the other hand, if a time series exhibits a stochastic trend, the random walk model and its variations may be more appropriate.

In addition to controversial issues around the ways to model seasonal and trend time series, one of the major limitations of many traditional models is that they are essentially linear methods. In order to use them, users must specify the model form without the necessary genuine knowledge about the complex relationship in the data. This is the main reason for the mixed findings reported in the literature regarding the best way to model and forecast trend and seasonal time series.

One non-linear model that recently received extensive attention is the neural network (NN) model. The popularity of the neural network model can be attributed to their unique capability to simulate a wide variety of underlying non-linear behaviors. Indeed, research has provided theoretical underpinning of neural network's universal approximation ability. In addition, few assumptions about the model form are needed in applying the NN technique. Rather, the model is adaptively formed with the real data. This flexible data-driven modeling property has made NNs an attractive tool for many forecasting tasks, as data are often abundant while the underlying data generating process is hardly known.

In this article, we provide an overview on how to effectively model and forecast consumer retail sales using neural network models. Although there are many studies on general neural network forecasting, few are specifically focused on trending or seasonal time series. In addition, controversial results have been reported in the literature. Therefore it is necessary to have a good summary of what has been done in this area and more importantly to give guidelines that can be useful for forecasting practitioners.

It is important to note that the focus of this article is on time series forecasting methods. For other types of forecasting methods used in retail sales, readers are referred to Dominique (1998), Green (1986), and Smith, McIntyre, and Dale (1994).

BACKGROUND

Neural networks are computing models for information processing. They are very useful for identifying the func-

tional relationship or pattern in the retail sales and other time series data. The most popularly used neural network model in practice for retail sales is the feedforward multi-layer network. It is composed of several layers of basic processing units called neurons or nodes. For an in-depth coverage of NN models, readers are referred to Bishop (1995) and Smith (1993). A comprehensive review of the NNs for forecasting is given by Zhang, Patuwo, and Hu (1998).

Before it can be used for forecasting, the NN model must be built first. Neural network model building (training) involves determining the order of the network (the architecture) as well as the parameters (weights) of the model. NN training typically requires that the in-sample data be split into a training set and a validation set. The training set is used to estimate the parameters of some candidate models, among which the one that performs the best on the validation set is selected. The out-of-sample observations can be used to further test the performance of the selected model to simulate the real forecasting situations.

The standard three-layer feedforward NN can be used for time series forecasting in general and retail sales in particular. For one-step-ahead forecasting, only one output node is needed. For multiple-step forecasting, more output nodes should be employed. For time series forecasting, the most important factor in neural networks modeling is the number of input nodes, which corresponds to the number of past observations significantly auto-correlated with the future forecasts. In a seasonal time series such as the retail sales series, it is reasonable to expect that a forecasting model should capture the seasonal autocorrelation that spans at least one or two seasonal periods of, say, 12 or 24 for monthly series.

Therefore, in modeling seasonal behavior, it is critical to include in the input nodes the observations separated by multiples of seasonal period. For example, for a quarterly seasonal time series, observations that are four quarters away are usually highly correlated. Although theoretically, the number of seasonal lagged observations that have autocorrelation with the future value can be high, it is fairly small in most practical situations, as empirical studies often suggest that the seasonal autoregressive order be one or at most two (Box & Jenkins, 1976).

There are many other parameters and issues that need to be carefully considered and determined in neural network model building for retail and other time series. These include data preparation, data division and sample size, network architecture in terms of number of hidden and input nodes, training algorithm, model evaluation criteria, and so forth. Practical guidelines can be found in many references in the literature including Adya and Collopy (1998), Kaastra and Boyd (1996), and Zhang et al. (1998).

MAIN FOCUS OF THE ARTICLE

Modeling seasonal and trend time series has been one of the main research endeavors for decades. In this article, we provide a review of the research work on trend and seasonal time series forecasting with a focus on recent studies for retail sales time series with neural networks. In the early 1920s, the decomposition model along with seasonal adjustment was the major research focus due to Persons (1919) work on decomposing a seasonal time series. Different seasonal adjustment methods have been proposed, and the most significant and popular one is the Census X-11 method developed by the Bureau of the Census in the 1950s and 1960s, which has evolved into the current X-12-ARIMA program. Because of the *ad hoc* nature of the seasonal adjustment methods, several model-based procedures have been developed. Among them, the work by Box and Jenkins (1976) on the seasonal ARIMA model has had a major impact on the practical applications to seasonal time series modeling. This model has performed well in many real world applications and is still one of the most widely used seasonal forecasting methods. More recently, neural networks have been widely used as a powerful alternative to traditional time series modeling.

In neural network forecasting, little research has been done focusing on seasonal and trend time series modeling and forecasting. In fact, how to effectively model seasonal time series is a challenging task not only for the newly developed neural networks, but also for the traditional models. One popular traditional approach to dealing with seasonal data is to remove the seasonal component first before other components are estimated. Many practitioners in various forecasting applications have satisfactorily adopted this practice of seasonal adjustment. However, several recent studies have raised doubt about its appropriateness in handling seasonality. Seasonal adjustment has been found to lead into undesirable non-linear properties, severely distorted data, and inferior forecast performance (Ghysels, Granger, & Siklos, 1996; Plosser, 1979). De Gooijer and Franses (1997, p. 303) pointed out that "although seasonally adjusted data may sometimes be useful, it is typically recommended to use seasonally unadjusted data." On the other hand, mixed findings have also been reported in the limited neural network literature on seasonal forecasting. For example, Sharda and Patil (1992) found that, after examining 88 seasonal time series, NNs were able to model seasonality directly and pre-deseasonalization is not necessary. Alon, Qi, and Sadowski (2001, p. 154) also found that NNs are able to "capture the dynamic non-linear trend and seasonal patterns, as well as the interactions between them." However, Farway and Chatfield (1995) and Nelson, Hill, Remus, and O'Connor (1999), among others, found just the opposite. Their findings suggest that neural networks are not able to directly model

seasonality, and pre-deseasonalization of the data is necessary to improve forecasting performance.

Chu and Zhang (2003) compared the accuracy of various linear and neural network models for forecasting aggregate retail sales. Using multiple cross-validation samples, they found that the non-linear models outperform their linear counterparts in out-of-sample forecasting, and prior seasonal adjustment of the data can significantly improve forecasting performance of the neural network model. The overall best model is the neural network built on deseasonalized time series data. While seasonal dummy variables can be useful in developing effective regression models for predicting retail sales, the performance of dummy regression models may not be robust. In addition, they found that trigonometric models are not useful in aggregate retail sales forecasting.

Zhang and Qi (2005) examined the issue of how to use neural networks more effectively in modeling and forecasting a seasonal time series with a trend component. The specific research questions they addressed are (1) whether neural networks are able to directly model different components of a seasonal and trend time series and (2) whether data pre-processing is necessary or beneficial. Instead of focusing solely on the seasonal component alone as in previous studies (e.g., Nelson et al., 1999), they took a systematic approach on the data preprocessing issue to study the relevance of detrending and deseasonalization. Using a large number of simulated and real time series, they evaluated the effect of different data preprocessing strategies on neural network forecasting performance. Results clearly show that without data preprocessing neural networks are not able to effectively model the trend and seasonality patterns in the data, and either detrending or deseasonalization can greatly improve neural network modeling and forecasting accuracy. A combined approach of detrending and deseasonalization is found to be the most effective data preprocessing that can yield the best forecasting result.

Therefore, it is recommended that data preprocessing is performed before building neural network models for retail sales data that contain seasonal and trend components. If the time series contains only the seasonal variation, deseasonalization should be the best choice. However, if both trend and seasonal fluctuations are evident, a combined approach of detrending and deseasonalization should be used.

FUTURE TRENDS

Sales forecasting is critical for most business operations. Developing new forecasting methods and models will continue to be an important endeavor in the future. Although neural networks provide an excellent modeling tool for forecasting researchers and practitioners, there are still many ways to improve their performance in the future. For example, while this article focuses on time series approaches, it may be valuable to consider both time series and cross-sectional data in building sales forecasting models. Many external factors including both micro- and macro-variables can be useful in sales forecasting. In addition, combining multiple models to improve forecasting accuracy is an important area to explore.

CONCLUSION

Neural networks become an important tool for retail sales forecasting. Due to complex nature in model building for retail sales, it is necessary to preprocess the data first. Several recent studies strongly suggest that for retail sales that contain both trend and seasonal variations, it is not appropriate to directly model sales time series with neural networks. Rather, both seasonal effect and trend movement can have significant effect in accurately modeling retail sales. In order to build the best neural network model, forecasters should use a combined approach of deseasonalization and detrending to removing the seasonal and trend factors first. Another benefit with both seasonality and trend removed is that more parsimonious neural networks can be constructed.

REFERENCES

Adya, M., & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation, *Journal of Forecasting, 17*(5-6), 481-495.

Agrawal, D., & Schorling, C. (1996). Market share forecasting: An empirical comparison of artificial neural networks and multinomial logit model. *Journal of Retailing, 72*(4), 383-407.

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services, 8*(3), 147-156.

Barksdale, H. C., & Hilliard, J. E. (1975). A cross-spectral analysis of retail inventories and sales. *Journal of Business, 48*(3), 365-382.

Bishop, M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting, and control*. San Francisco: Holden Day.

Chopra, S., & Meindl, P. (2007). *Supply chain management: Strategy, planning, and operation* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Chu, C., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics*, forthcoming.

De Gooijer, J. G., & Franses, P. H. (1997). Forecasting and seasonality. *International Journal of Forecasting, 13*(3), 303-305.

Dominique, H. M. (1998). Order forecasts, retail sales, and the marketing mix for consumer durables. *Journal of Forecasting, 17*(34), 327-348.

Farway, J., & Chatfield, C. (1995). Time series forecasting with neural networks: A comparative study using the airline data. *Applied Statistics, 47*(2), 231-250.

Franses, P. H., & Draisma, G. (1997). Recognizing changing seasonal patterns using artificial neural networks. *Journal of Econometrics, 81*(11), 273-280.

Ghysels, E., Granger, C. W. J., & Siklos, P. L. (1996). Is seasonal adjustment a linear or nonlinear data filtering process? *Journal of Business and Economics Statistics, 14*(3), 374-386.

Green, H. L. (1986). Retail sales forecasting systems. *Journal of Retailing, 62*(3), 227-230.

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing, 10*(3), 215-236.

Lee, H. L., Padmanabhan, V., & Whang, S. (1997). The bullwhip effect in supply chains. *Sloan Management Review, 38*(3), 93-102.

Mentzer, J. T., & Bienstock, C. C. (1998). *Sales forecasting management*. Thousand Oaks, CA: SAGE.

Nelson, M., Hill, T., Remus, T., & O'Connor, M. (1999). Time series forecasting using NNs: Should the data be deseasonalized first? *Journal of Forecasting, 18*(5), 359-367.

Nelson, C. R., & Plosser, C. I., (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics, 10*(2), 139-162.

Persons, W. M. (1919). Indices of business conditions. *Review of Economics and Statistics, 1*(1), 5-107.

Pierce, D. A. (1977). Relationships—and the lack of thereof—between economic time series, with special reference to money and interest rates. *Journal of the American Statistical Association, 72*(1), 11-26.

Plosser, C. I. (1979). Short-term forecasting and seasonal adjustment. *Journal of the American Statistical Association, 74*(1), 15-24.

Sharda, R., & Patil, R. B. (1992). Connectionist approach to time series prediction: An empirical test. *Journal of Intelligent Manufacturing, 3*(5), 317-323.

Smith, M. (1993). *Neural networks for statistical modeling*. New York: Van Nostrand Reinhold.

Smith, S. A., McIntyre, S. H., & Dale, A. D. (1994). Two-stage sales forecasting procedure using discounted least squares. *Journal of Marketing Research, 31*(1), 44-56.

Thall, N. (1992). Neural forecasts: A retail sales booster. *Discount Merchandiser, 23*(10), 41-42.

Zhang, G., Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35-62.

Zhang, G. P., & Qi, M. (2002). Predicting consumer retail sales using neural networks. In K. Smith & J. Gupta (Eds.), *Neural networks in business: Techniques and applications*. Hershey, PA: Idea Group Publishing.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting of seasonal and trend time series. *European Journal of Operational Research, 160*(2), 501-514.

KEY TERMS

Box-Jenkins Approach: A very versatile linear time series modeling approach that can model trend, seasonal, and other behaviors by using moving averages, autoregression, and difference equations.

Census II X-11: A method that systematically decomposes a time series into trend, cyclical, seasonal, and error components. It was developed by the Bureau of the Census of the Department of Commerce and is widely used in deseasonalizing economic data.

Deseasonalization: Sometimes also called seasonal adjustment. A process of removing seasonality from the time series. Most governmental statistics are seasonally adjusted to better reflect other components in a time series.

Detrending: A process of removing trend from the time series through either differing of time series observations or subtracting fitted trends from actual observations.

Feedforward Neural Network: A special type of neural network where processing elements are arranged in layers and the information is one directional from input layer to hidden layer(s) to output layer.

Neural Networks: Computing systems that are composed of many simple processing elements operating in parallel

whose function is determined by network structure. They are used mainly to model functional relationship among many variables.

Retail Sales: A measure of the total receipts of retail stores. Retail sales data provide valuable information about consumer spending. Aggregate retail sales are reported monthly by the Commerce Department.

Seasonality: Periodic pattern that typically occurs within a year and repeats itself year after year. Most retail data exhibit seasonal variations that repeat year after year.

Trend: Long-term movement in a time series that represents the growth or decline over an extended period of time.

New Perspectives on Rewards and Knowledge Sharing

Gee-Woo (Gilbert) Bock
National University of Singapore, Singapore

Chen Way Siew
IBM Consulting Services, Singapore

Young-Gul Kim
KAIST, Korea

INTRODUCTION

Of the 260 responses from a survey of European multinationals, 94% believed that knowledge management requires employees to share what they know with others within the organization (Murray, 1999). Among the processes of knowledge management—creation, sharing, utilization and accumulation of knowledge—sharing is what differentiates organizational knowledge management from individual learning or knowledge acquisition.

However, the process of sharing knowledge is often unnatural to many. Individuals will not share knowledge that is regarded to be of high value and importance. In fact, the natural tendency for individuals is to hoard knowledge or look suspiciously at the knowledge of others. Thus, incentive schemes—where employees receive incentives as a form of compensation for their contributions—are common programs in many organizations. Such schemes have met their fair share of success as well as failure in the field of knowledge management. On the one hand, the carrot and stick principle used in Siemens' ShareNet project turned out to be a success (Ewing & Keenan, 2001). On the other hand, the redemption points used in Samsung Life Insurance's Knowledge Mileage Program only resulted in the increasingly selfish behavior of its employees (Hyoung & Moon, 2002).

Furthermore, despite the plethora of research on factors affecting knowledge sharing behavior, little concerns discovering effective ways to encourage individuals to voluntarily share their knowledge. Early studies on knowledge management began by trying to discover key factors pertaining to knowledge management in general, instead of knowledge sharing in particular, as summarized in Table 1. Although research on knowledge sharing started around the mid 1990s, it focused mainly on knowledge sharing at the group or organizational level in spite of the fact that knowledge itself actually originates from the individual. Even at the group or organizational level, most studies dealt with a specific knowledge type, such as best practices (Szulanski, 1996) or a specific context, such as between dispersed teams (Tsai, 2002). In addition, factors such as trust, willingness to share, information about the knowledge holder, and the level of codification of knowledge were considered in abstract. Although these factors are valuable, they require further empirical research before they could be used to explain the individual's fundamental motivation to share knowledge. Thus, this study aims to develop an understanding of the factors that support or constrain the individual's knowledge sharing behavior in the organization, with a special interest in the role of rewards. This is done according to Fishbein and Ajzen's (1975) Theory of Reasoned Action (TRA), a widely accepted social psychology model that is used to explain almost any human behavior (Ajzen & Fishbein, 1980).

BACKGROUND

Due to the fact that knowledge is a resource that is locked in the minds of humans, knowledge sharing does not occur with the sole implementation of information systems. As such, an investigation into the individual's motivation behind knowledge sharing behavior, coupled with a firm foundation in social psychology, should take precedence. Accordingly, the TRA is adopted so as to provide a well-established explanation for such volitional, rational, systematic decision logic as that of knowledge sharing.

The TRA assumes that human beings are usually rational in thinking, and would systematically use available information (Fishbein & Ajzen, 1975). In the TRA, the individual's attitude toward and subjective norm regarding a behavior jointly determine the behavioral intention that results in the individual's decision to engage in a specific behavior. In this study, we focus only on the salient beliefs that affect the knowledge sharing attitude because knowledge sharing behavior is assumed to be motivated and executed mainly at the individual level. Since the TRA can be applied to almost any behavior, the nature of the beliefs operative for a particu-

Table 1. Factors affecting knowledge management and knowledge sharing

	Factors	References
Knowledge Management	Knowledge management system, Network, Knowledge worker, Clear vision and goals, Middle-up-down management, Organizational change, Monitoring and support, Knowledge infrastructure, Knowledge repository and map, Organizational culture, Top manager's support	Davenport, De Long, and Beers (1998); Davenport and Prusak (1998); Earl (1996); Nonaka and Takeuchi (1995); Ulrich (1998); Wiig (1997)
Knowledge Sharing	The Group and Organizational Level Level of trust between groups, Arduous relationship between source and the recipient, Role of top managers, Characteristics of knowledge, Prior experience on knowledge transfer, Channel richness, Openness of the organization	Butler (1999); Gupta and Govindarajan (2000); Kogut and Zander (1993); Nelson and Cooprider (1996); Szulanski (1996); Wathne, Roos and Krogh (1996)
	The Individual Level Trust between individuals, Willingness to share, Information about the knowledge holder, Level of codification of knowledge	Hansen (1999); Kramer (1999); Moreland (1999); Stasser, Stewart, and Wittenbaum (1995); Tsai and Ghoshal (1998)

lar behavior are left unspecified. Following the elicitation recommendations suggested by Fishbein and Ajzen (1975), free response interviews to elicit five to nine salient beliefs were conducted with chief knowledge officers (CKO) and chief information officers (CIO) of the subject population in April 1999. Once these salient beliefs surfaced, the research model was developed.

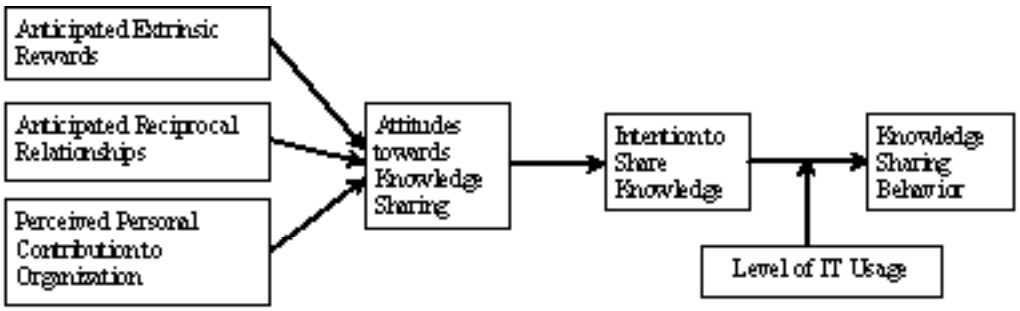
We propose three factors that are consistently emphasized throughout the interviews: anticipated extrinsic rewards, anticipated reciprocal relationships, and perceived personal contribution to the organization, as the antecedents of the attitudes towards knowledge sharing. According to the interdependence theory, individuals will behave according to rational self-interest. Knowledge sharing occurs when the rewards exceed the costs (Constant, Keisler & Sproull, 1994; Kelley & Thibaut, 1978), implying that anticipated extrinsic rewards will positively affect the individual's attitude. Concerning intrinsic rewards, the social exchange theory states that social exchanges entail unspecified obligations (Blau, 1967). As employees are seen to believe that their relationship with others can be improved through sharing knowledge, the anticipated reciprocal relationships positively affect the individual's attitude. In addition to these, the self-motivation theory (Deci, Connell & Ryan, 1989; Iaffaldano & Muchinsky, 1985; Schwab & Cummings, 1970) finds that feedback from others on shared knowledge can form a self-motivational factor and serve as another major determinant of the attitude toward knowledge sharing. Eisenberger and Cameron (1996) note that one's sense of competence actually increases due to the feedback concerning the quality of one's output. Employees who are able to link instances of past knowledge sharing with an understanding of how these actions contribute to others' work, and/or improvements in organizational performance are likely to develop more favorable attitudes toward knowledge sharing than employees who are unable to construct such linkages. Finally, following Fishbein and Ajzen's (1975) argument about the possibility of several external variables affecting intention to perform a behavior, we introduced an aspect of information technology (IT) into our model. Since IT is considered to be an important enabler in knowledge management (O'Dell & Grayson, 1998; Ruggles, 1998), we examined how the individual's level of IT usage affects knowledge sharing behavior.

Data were collected through the utilization of a survey. A total of 900 questionnaires were distributed in October and November 1999 to employees in 75 departments of four large government-invested organizations in South Korea. Of this total number, 861 responses were received, of which 467 were usable. We found that the anticipated reciprocal relationship provided for the individual's positive attitude towards knowledge sharing, and resulted in a positive influence of intention and behavior. However, contrary to many researchers' expectations, anticipated extrinsic rewards were found to have a negative effect on such an attitude.

FUTURE TRENDS

This negative correlation—which might prove important for future research—can be explained with the results of research in the pay-performance area. Kohn (1993) found that there is either no relationship or a negative relationship

Figure 1. Research model



between rewards and performance, although many assume that people will do a better job if they are promised some form of reward. Kohn cited six reasons as to why rewards fail, three of which can also be considered within the knowledge-sharing context.

First, rewards are seen to have a punitive effect because, as compared to outright punishment, they are manipulative in nature. Not receiving an expected reward is seen to be indistinguishable from being punished. Both result in movement, but not motivation (Herzberg, 1968). Rewards are seen to destroy relationships because for any one winner, many others would feel that they have lost. When there exists a limited number of rewards, competition between employees will ensue. Second, rewards are at times used as a simpler alternative to addressing underlying issues, such as the lack of an ideal knowledge-sharing culture within the organization. The ideal culture mentioned should include providing useful feedback, social support, and room for self-determination. Third, rewards could be an undermining factor towards intrinsic motivation. Interest in knowledge sharing would decrease with an increase in one's perception of being controlled (Levinson, 1973). Employees might assume that the task at hand is not something they would want to do if they have to be bribed to do it. As such, with the increase in incentive offered, the negative perception towards the task at hand becomes greater.

Another explanation can also be found in organizational citizenship behavior literature. According to Katz and Kahn (1966), any critical voluntary behavior that is beyond the scope of one's job description is a direct result of one's identification with and internalization of individual and organizational values, rather than the involvement of any external factors. Furthermore, Constant et al. (1994) stated that experienced workers perceive the process of sharing knowledge as part of normal business activity. These workers hold a negative view of any extrinsic rewards given in return for sharing knowledge. With such strong support for the negative effect view of extrinsic rewards on the attitude towards knowledge sharing, would it be right to completely discard extrinsic rewards?

Eisenberger and Cameron (1996) found that extrinsic rewards could both positively and negatively influence motivation—knowledge sharing in this case. They find that rewards can be divided into two broad types, namely task-contingent and quality-dependent rewards. Quality-dependent rewards positively influence organization initiatives, as they do not reduce one's intrinsic motivation. In fact, due to the feedback concerning the quality of one's output, one's sense of competence actually increases. Task-contingent rewards, on the other hand, undermine any task because of their negative influence on intrinsic motivation. A possible design for such a scheme is a knowledge market where "buyers and sellers of knowledge negotiate a mutually satisfactory price for the knowledge exchanged" (Ba, Stallaert & Whinston, 2001, p. 232). In this way, the reward of individuals would be based on the usefulness of their knowledge, thus ensuring the creation of high-quality knowledge.

In addition, a reward that is less than what employees feel their performance justify could threaten their self-esteem. This is due to the fact that the self-ratings done by employees are usually higher than those done by the management. According to Meyer (1975), a common way for employees to cope with such a problem is to "downgrade the importance of the activity on which the threat is focused" (p. 44). Hence, an incentive scheme should be well-designed so as to reward individuals who are deemed deserving, as the rewarding of contributors aids in positively influencing the sharing of knowledge. Additionally, the scheme's design should discourage self-centered behavior, which is detrimental to the organization's health (Michailova & Husted, 2003); otherwise, the scheme would only produce temporary compliance, and might decrease an individual's intrinsic motivation (Deci, 1971,1972a). When intrinsic interests exist among employees for a particular task, the attachment of

incentives to the performance of the task only results in the decrease of that interest (Deci, 1972b). "When pay becomes the important goal, the individual's interest tends to focus on that goal rather than on the performance of the task itself" (Meyer, 1975, p. 41), thus resulting in employees striving to increase incentives at the cost of output quality.

Furthermore and most importantly, an incentive scheme needs to be incorporated with proper organization norms (Markus, 2001). If the norms are not in place, employees will not share their knowledge even if there is in place, a comprehensive incentive scheme. According to O'Dell and Grayson (1998), "if the process of sharing and transfer is not inherently rewarding, celebrated and supported by the culture, then artificial rewards [will not] have much effect and can make people cynical" (p. 168). Husted and Michailova (2002) also stated that "unless knowledge sharing is built into the expectation of the individual and is reflected in the reward mechanism, sharing will not take place".

CONCLUSION

In summary, of the two views posed by past research, our recent study to discover the influence of extrinsic rewards in knowledge sharing supports the negative view. This implies that, when the management of an organization is motivated to embrace knowledge sharing but its employees are not, using incentives to influence knowledge sharing would only result in the employees placing emphasis on the incentives. This could result in the sharing of low-quality knowledge and undermine the whole knowledge-sharing effort. Furthermore, the continuous use of incentives "could actually be encouraging hoarding behavior and competitive actions, diminishing the free flow of knowledge in the organization" (Wasko & Faraj, 2000, p. 162). Therefore, extrinsic rewards should be coupled with other factors, such as organizational norms, to bring about benefits. The reconciliation of this disparity in views should provide new grounds for future research.

REFERENCES

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Alavi, M., & Leidner, D.E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly*, *25*(1), 107-136.

Ba, S., Stallaert, J., & Whinston, A.B. (2001). Introducing a third dimension in information systems design-The case for incentive alignment. *Information Systems Research*, *12*(3), 225-239.

Blau, P. (1967). *Exchange and power in social life*. New York: Wiley.

Bock, G.W., & Kim, Y.G. (2002). Breaking the myths of rewards: An exploratory study of attitudes about knowledge sharing. *Information Resources Management Journal*, *15*(2), 14-21.

Bock, G.W., Zmud, R.W., Kim, Y.G., & Lee, J.N. (2003). *Determinants of the individual's knowledge sharing behavior: From the theory of reasoned action*. University of Minnesota: Minnesota Symposium on Knowledge Management, 3.

Butler, J.K. (1999). Trust expectations, information sharing, climate of trust, and negotiation effectiveness and efficiency. *Group & Organization Management*, *24*(2), 217-238.

Constant, D., Keisler, S., & Sproull, L. (1994). What's mine is ours, or is it? A study of attitudes about information sharing. *Information Systems Research*, *5*(4), 400-421.

Davenport, T.H., De Long, D.W., & Beers, M.C. (1998). Successful knowledge management projects. *Sloan Management Review*, *39*(2), 43-57.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge*. Boston, MA: Harvard Business School Press.

Deci, E.L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology*, *18*(1), 105-115.

Deci, E.L. (1972a). Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personality and Social Psychology*, *22*(1), 113-120.

Deci, E.L. (1972b). The effects of contingent and noncontingent rewards and controls on intrinsic motivation. *Organizational Behavior and Human Performance*, *8*, 217-229.

Deci, E.L., Connell, J.P., & Ryan, R.M. (1989). Self-determination in a work organization. *Journal of Applied Psychology*, *74*(4), 580-590.

Earl, M.J. (1996). Knowledge as strategy. In L. Prusak (Ed.), *Knowledge in organizations*. Newton, MA: Butterworth-Heinemann.

Eisenberger, R., & Cameron, J. (1996). Detrimental effects of reward: Reality or myth? *American Psychologist*, *51*(11), 1153-1166.

Ewing, J., & Keenan, F. (2001). Sharing the wealth. *Business Week, 3724,* EB36-EB40.

Fishbein, M., & Ajzen, I. (1975). *Beliefs, attitude, intention and behavior: An introduction to theory and research*. Philippines: Addison-Wesley Publishing Company.

Gupta, A.K., & Govindarajan, V. (2000). Knowledge flows within multinational corporations. *Strategic Management Journal*, *21*(4), 473-496.

Hansen, M.T. (1999). The search–transfer problem: The role of weak ties in sharing knowledge across organization subunits. *Administrative Science Quarterly*, *44*(1), 82-111.

Herzberg, F. (1968). One more time: How do you motivate employees? *Harvard Business Review*, *46*(1), 53-62.

Husted, K., & Michailova, S. (2002). Knowledge sharing in Russian companies with western participation. *International Management*, *6*(2), 17-28.

Hyoung, K.M., & Moon, S.P. (2002). Effective reward systems for knowledge sharing. *Knowledge Management Review*, *4*(6), 22-25.

Iaffaldano, M.T., & Muchinsky, P.M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin*, *97*(2), 251-273.

Katz, D., & Kahn, R.L. (1966). *The social psychology of organizations*. New York: Wiley.

Kelley, H.H., & Thibaut, J.W. (1978). *Interpersonal relations: A theory of independence*. New York: Wiley.

Kogut, B., & Zander U. (1993). Knowledge of the firm and the evolutionary theory of the multinational corporation. *Journal of International Business Studies*, *24*(4), 625-645.

Kohn, A. (1993). Why incentive plans cannot work. *Harvard Business Review*, *71*(5), 54-63.

Kramer, R.M. (1999). Social uncertainty and collective paranoia in knowledge communities: Thinking and acting in the shadow of doubt. In L.L. Thomson, J.M. Levine & D.M. Messick (Eds.), *Shared cognition in organizations, the management of knowledge*. London: LEA Inc.

Levinson, H. (1973). Asinine attitudes toward motivation. *Harvard Business Review*, *51*(1), 70-76.

Markus, M.L. (2001). Toward a theory of knowledge reuse: Types of knowledge reuse situations and factors in reuse success. *Journal of Management Information Systems*, *18*(1), 57-93.

McDermott, R., & O'Dell, C. (2001). Overcoming cultural barriers to sharing knowledge. *Journal of Knowledge Management*, *5*(1), 76-85.

Meyer, H.H. (1975). The pay-for-performance dilemma. *Organization Dynamics*, *3*(3), 39-50.

Michailova, S., & Husted, K. (2003). Knowledge-sharing hostility in Russian firms. *California Management Review*, *45*(3), 59-77.

Moreland, R.L. (1999). Transactive memory: Learning who knows what is work groups and organizations. In L.L. Thomson, J.M. Levine & D.M. Messick (Eds.), *Shared cognition in organizations, the management of knowledge*. London: LEA Inc.

Murray, P. (1999, March 8). How smarter companies get results from KM. *Financial Times*, 15.

Nelson, K.M., & Cooprider, J.G. (1996). The contribution of shared knowledge to IS group performance. *MIS Quarterly*, *20*(4), 409-429.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company*. New York: Oxford University Press.

O'Dell, C., & Grayson, C.J. (1998). If only we knew what we know: Identification and transfer of internal best practices. *California Management Review*, *40*(3), 154-174.

Ruggles, R. (1998). The state of notion: Knowledge management in practice. *California Management Review*, *40*(3), 80-89.

Schwab, D.P., & Cummings, L.L. (1970). Theories of performance and satisfaction. *Industrial Relations*, *9*(4), 408-430.

Stajkovic, A.D., & Luthans, F. (1998). Social cognitive theory and self-efficacy: Going beyond traditional motivational and behavioral approaches. *Organizational Dynamics*, *26*(4), 62-74.

Stasser, G., Stewart, D.D., & Wittenbaum, G.M. (1995). Expert roles and information exchange during discussion: The importance of knowing who knows what. *Journal of Experimental Social Psychology*, *31*(3), 244-265.

Stenmark, D. (2000). Leveraging tacit organizational knowledge. *Journal of Management Information Systems*, *17*(3), 9-24.

Szulanski, G. (1996). Exploring internal stickiness: Impediments to the transfer of best practice within the firm. *Strategic Management Journal*, *17*(Winter), 27-44.

Triandis, H.C. (1971). *Attitude and attitude change*. New York: John Wiley & Sons, Inc.

Tsai, W. (2002). Social structure of "coopetition" within a multiunit organization: Coordination, competition, and intraorganizational knowledge sharing. *Organization Science*, *13*(2), 179-190.

Tsai, W., & Ghoshal, S. (1998). Social capital and value creation: The role of intrafirm networks. *Academy of Management Journal*, *41*(4), 464-476.

Ulrich, D. (1998). Intellectual capital = Competence x commitment. *Sloan Management Review*, *39*(2), 15-26.

Wasko, M.M., & Faraj, S. (2000). "It is what one does": Why people participate and help others in electronic communities of practice. *The Journal of Strategic Information Systems*, *9*(2-3), 155-173.

Wathne, K., Roos, J., & Krogh, G. (1996). Toward a theory of knowledge transfer in a cooperative context. In G. Krogh & J. Roos (Eds.), *Managing knowledge*. London: Sage.

Wiig, K.M. (1997). Knowledge management: Where did it come from and where will it go? *Expert Systems with Applications*, *13*(1), 1-14.

KEY TERMS

Explicit Knowledge: Knowledge that has been captured and codified into manuals, procedures, and rules, and is easy to disseminate (Stenmark, 2000).

Extrinsic Rewards: Incentives that are mediated outside of a person, such as praises and monetary compensation (Deci, 1972b).

Implicit Knowledge: Knowledge that can be expressed in verbal, symbolic, or written form but has yet to be expressed (Bock, Zmud, Kim, & Lee, 2003).

Intrinsic Rewards: Incentives that are mediated within a person, such as satisfaction (Deci, 1972b).

Knowledge Management System: A knowledge repository, shared knowledge base or knowledge based system, which is a class of information systems developed to support and enhance the organizational processes of knowledge creation, storage / retrieval, transfer and application (Alavi & Leidner, 2001).

Knowledge Sharing: Voluntary activities of transferring or disseminating knowledge between people or groups in an organization (Bock, Zmud, Kim & Lee, 2003).

Organizational Norm: Organization culture or climate, which consists of the shared values, beliefs and practices of the people in the organization (McDermott & O'Dell, 2001).

Tacit Knowledge: Knowledge that cannot be easily articulated, and thus only exists in people's minds, and is manifested through their actions (Stenmark, 2000).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2110-2115, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

New Technologies in Hospital Information Systems

Dimitra Petroudi
National and Kapodistrian University of Athens, Greece

Nikolaos Giannakakis
National and Kapodistrian University of Athens, Greece

INTRODUCTION

A hospital information system (HIS), variously also called clinical information system (CIS), is a comprehensive, integrated information system designed to manage the administrative, financial, and clinical aspects of a hospital. This encompasses paper-based information processing as well as data processing machines.

As an area of medical informatics, the aim of an HIS is to achieve the best possible support of patient care and administration by electronic data processing.

It can be composed of one or few software components with specialty specific extensions, as well as of a large variety of subsystems in medical specialties (e.g., laboratory information system, radiology information system).

CISs are sometimes separated from HISs in that the former concentrate on patient and clinical state-related data (electronic patient record), whereas the latter keeps track of administrative issues. The distinction is not always clear, and there is contradictory evidence against a consistent use of both terms.

Types of HIS

1. Central or exocentric: The difference is supported in whether the information is kept in central computer or is distributed in other computers in all the hospital.
2. Oriented or not to the patient: Even if both of this two types deal with the data of patient, the orientation of HIS can influence the processes and the general "character of" HIS.
3. With terminals or workstations: They are two appliances that resemble and usually are not separated. Terminals are electronic appliances that allow the users to communicate with the computer. Generally, they are connected with mini-computers or mainframes that can find themselves far or near. If they are alone, they have few possibilities, and generally they are not capable to make anything if they are not connected with a functional computer. Workstations are computers drawn for professional use from an individual each time. They are completely functional computers, and they can be connected with other workstations, mainframes, or mini-computers.

An HIS can be placed:

1. Next to the bed of patient: Its placement next to the patient's bed is essential for the monitor and control appliances. For the recording of situation of patient, nevertheless, there is no advantage. In a study, the results were the recording of data was not improved when the system was found next to the patient, since the bigger part of recording was done outside the room, or in the rooms of other patients. Nevertheless, its placement in this point improved the use of automated drawings of care, the calculation of situation of patient, and the pricing for the care of services.
2. In the corridor near the patient's bed: Its placement in the corridor is continuously increasing. It allows the nurses to record, very shortly afterwards, their removal from the patient, without the detachment of attention from the presence of patient and the potential requirement of attention. However, there is danger for the safety, since someone can receive information about the situation of a patient simply looking at, indiscreetly, the hour of recording.
3. In a staff's room: Its placement in regions, where the staff is only allowed, has the advantage of bigger safety. However, it is uncomfortable and time consuming, since the staff should walk enough each time it needs information.
4. Other possibilities: Electronic clipboards. The unique disadvantage is found in that the users perhaps forget where they left it.

Expected profits from the hospital information systems

1. Reduction or repression of registrations
2. Reduction of office duties for the medical and nursing staff

3. Easier access to the medical data
4. Reduction of duration of staying in the hospital
5. Minimisation the insufficient medical recipes
6. Minimisation of errors in the recording of results
7. Redeployment, reorientation or reduction of staff
8. Improvement of quality of registrations
9. Improvement of quality of care
10. Better communication
11. Reduction of hospital cost
12. Increase of satisfaction of nurses
13. Growth of common hospital database
14. Improvement of perception of patients on their care
15. Improvement of general appearance of hospital

HARDWARE TECHNOLOGIES

The patients entrust the organisms of healthcare in order to offer them the higher level of care with the smaller probability of error. The existing technologies help standards of health to be strengthened, but the hardware solutions will save an important cost, also measured in money and in time. Solutions, such as electronic forms, can exclude the problems that come up in a handwritten system. These technologies give the doctors/nurses, and the other clinical, a lot of time in order to focus in what they know better.

Bar Coding

The bar coding is a low risk technology (cost, application) that makes it a practical choice for a organisation. Also, in order to be used effectively, it does not need intensive education. The positive results of bar code in the safety of patient are widely acceptable in the healthcare system.

The initiative of FDA (Food and Drug Administration) for the safety of the patient was completed in February 2004, and it will ask for the hospitals to use bar codes in the hospitals the next 3 years. The technology uses bar code for the patients, the medicines, the blood, and the vaccines. A tool collects all the codes, and is immediately connected with the medical file of patient. Following the system with the bar codes, the FDA calculated that this will involve economy of 3-9 billion dollars.

The solution offers:

- Effectiveness
- Safety
- Reduction of cost
- Reciprocity
- Correctness
- Precision

Touch Screen Technology

This technology allows a nurse to easily export conclusions for the patients using a friendly environment (touch screen). The nurse simply touches upon the screen that she/he wishes. The button in the screen shows the room number at the same time as the name of patient. In this easy two-step access, the nurse shows the category from which she/he asks information.

This solution offers:

- Easy to use
- Precision in the import of data
- Efficient administration
- Easy completion of data.

Browser-Based Technology

Three factors should appear for using the browser-based technologies. Firstly, the technology should contribute to the cost decrease. Secondly, the technology should have competitive advantage. Finally, the technology should have an impact in the improvement of patient. An installation with this technology needs only a Web browser, such as Internet Explorer, Netscape, Mozilla, for running the software.

The solution offers:

- Portability - the browser-based software offers an easy access to the information with point and click
- Easy access each hour (day-night), from everywhere (office, home, everywhere Internet exists)
- Less hour of employment - It is decreased because the system functions permanently all day
- Friendly to the user - does not need learning new applications only point and click.

Document Imaging

The solution offers:

- Focus to the customer service - the important data can be easily filed. The solution allows the hospital to vindicate the medical data
- Decreased of printing - All the information is stored in a filed system that is fast approachable
- Easy Web access - the doctors, as long as the other users can have always access and from everywhere via Web
- Not other lost files - the e-files give the possibility to the hospital to compose an electronic recording of all given data and afterwards, place the disk in a database

- Completed copy of HIS - the medical directives can immediately be placed in a filed e-file with easy access and reuse
- Better functional space - Better management of existing spaces in the hospital and utilisation of these for more effective ways.

Mobile Carts

They make the worker feel comfortable with the utilisation of technology. The professionals can be hesitant in the use of the technology if the programs are not easily used. The unwillingness can be obvious in each employee, which can decrease his/her productivity or increase the overtime and create an unpleasant working environment. With a portable workstation, the aim is for the workers to feel comfortable. The solution offers:

- Faculty of transport
- Flexibility.

Pen Tablets

This technology helps hospitals to decrease the functional costs and the handwritten work, as well as to improve the healthcare. In this technology, there is a pen that is used as the mouse. Pen tablets allow the professionals of health to store information fast and easily, allow them to focus on the patient and spend less time in tedious debits. The professionals, with this technology, can collect the vital points, statistical data, and other information near to patient's bed.

The solution offers:

- Decreased handwritten work
- It improves the provided healthcare
- It increases the productivity.

E-Forms

E-forms allow the health professionals to organise their registrations without the enormous volume of paper that is usually in the bookshelves of an office.

The solution offers:

- Bigger flow - e-forms give the opportunity to the hospital for research, decreasing the repeated exit for handwritten work.
- Correctness (precision) - the solution increases the safety of patient by decreasing the errors that exist in a handwritten system.
- It improves the flow of work - E-forms offer to the hospital processes in order to improve the effectiveness. It releases the personnel from tedious obligations.
- Seamless integration - the solution offers completion to the hospital without extra programs and maintenance.
- It increases the professionalism - Improving the appearance and the drawing of medical forms can improve the picture of hospital for the public.

Wireless Technology

The medical community continues searching ways in order to maintain the precision, as well as to improve the functional efficiency. For example, the doctors can have access and renew the medical file using computers that are found next to the patient's bed. The portable computers, the PDA, the Pen Tablets and the mobile telephones were proved a new powerful tool for the improvement of efficiency and effectiveness of services that are provided in the hospital environment. The information with regard to the patient, his/her care, and the programmed examinations is immediately available to the professionals of health in any place in the hospital. More important is the fact that with the import of data straight in the network of hospital, are limited the double written data and errors are decreased at the copy of data.

The solution offers:

- Mobility
- Flexibility
- Speed
- Safety

BENEFITS

For the patients:

- It builds an electronic medical file
- It improves the quality of care

For the hospital doctors:

- It helps the decision-making via the easier access in the information
- It is capable to program and record the clinical activities

For the governors:

- It provides the comprehensive daily lists of activity
- It simplifies the planning of energies for the care of patient
- It improves the communication with the colleagues

For the management:

- It helps with the required reports at the meetings
- It improves the efficiency via the automation
- It helps with the improvement of management of staff

CONCLUSION

The future of HIS depends on many factors, like: 1) the hardware becomes cheaper, 2) the software does not become more expensive, 3) the money spent by the government is not increased, 4) there have not been determined the models for the storage and exchange of data, and 5) it demands the education of the staff.

URL REFERENCES

http://en.wikipedia.org/wiki/Hospital_information_system

http://courses.wccnet.edu/computer/mod/m30c.htm

http://www.healthcare.siemens.com/soarian/int/index_flash.html

http://www.datamed.gr/products/index.asp?Cat_ID=108&pageID=113

http://www.datamed.gr/products/index.asp?Cat_ID=109&pageID=119

http://www.datamed.gr/products/index.asp?Cat_ID=110&pageID=114

http://www.tsystem.com/ED-Information-System/nurse-charting.asp

http://www.hmstn.com/solutions/?content_id=320

http://www.business2005.gr/ec_pageitem.asp?id=6466

http://www.pressreleases.gr

http://www.presspoint.gr

http://www.medicomsoft.com

http://www.e-m3i.com/content/HospitalInformationSystem.asp

http://www.cdacindia.com/html/his/sushrut.asp

http://www.stockell.com

KEY TERMS

Barcode (also **bar code**): A machine-readable representation of information in a visual format on a surface.

Hardware: The general term that is used to describe physical artifacts of a technology.

Laboratory Information System (LIS): A class of software that handles receiving, processing, and storing information generated by medical laboratory processes.

Radiology Information System (RIS)**:** Used by radiology departments to store, manipulate, and distribute patient radiological data and imagery.

Software Component: A system element offering a predefined service and able to communicate with other components.

Touch Screens, **Touch Panels:** Display overlays that have the ability to display and receive information on the same screen.

Next-Generation Enterprise Systems

Charles Møller
Aalborg University, Denmark

INTRODUCTION

"ERP is dead - long live ERP II" was the title of a path breaking research note from Gartner Group (Bond, Genovese, Miklovic, Wood, Zrimsek, & Rayner, 2000). In this research note, Gartner Group envisions how the ERP vendors respond to market challenges and how ERP and ERP strategies evolved by 2005. Gartner Group defines ERP II as a transformation of ERP (Enterprise Resource Planning), and today the major vendors have adopted this concept in their contemporary ERP packages.

ERP (Enterprise Resource Planning) is an important concept to industry. Enterprises are increasingly implementing packaged ERP systems. A recent study confirmed that over 90% of the 500 largest Danish enterprises have adopted one or more ERP system. Further, the study found the systems to be of an average age of 2.8 years and decreasing (Møller, 2005a).

ERP is a standardized software package designed to integrate the internal value chain of an enterprise (Klaus, Rosemann, & Gable, 2000). In 2002, the five major ERP vendors were: (i) SAP; (ii) Oracle; (iii) Peoplesoft; (iv) SAGE; and (v) Microsoft Business Solutions. They controled almost 50% of the ERP market (c.f. Table 1) and consequently the corporate infrastructure is dominated by the design of these systems and the vendors. By 2006, the market is consolidated and many of the smaller vendors have been merged with larger vendors. Oracle acquired PeopleSoft and JD Edwards and the global market seems to be dominated by SAP, Oracle and Microsoft.

According to Nah (2002) the American Production and Inventory Control Society (APICS) defines ERP as: "a method for the effective planning and controlling of all the resources needed to take, make, ship and account for customer orders in a manufacturing, distribution or service company." This definition expresses ERP as a tool but ERP is also a management vision and an agency of change and ERP has been attributed to almost any good or bad that IT may bring about in business.

In the late 1990s, the ERP hype was primarily motivated by companies rushing to prepare for Y2K (Calloway, 2000). Then, after a short recession the adoption of ERP has continued. Davenport's sequel on enterprise systems (Davenport, 1998, 2000; Davenport & Brooks, 2004) illustrates the changing business perspective on ERP and the ERP hype.

Davenport (1998) sums up the first wave of experiences from implementing ERP systems in a much cited paper on "putting the enterprise system into the enterprise," and points to the new potential business impact of the ERP systems. The discussion evolved over the first enthusiastic expectations, continued over a growing number of horror stories about failed or out-of-control projects, toward a renewed hype of expectations on e-business and SCM.

The ERP II concept is the software industry's perception of the new business challenges and the vision addresses the issues of e-business integration in the supply chain. ERP II is the next-generation ERP concept and in a few years from now the ERP II vision is going to be institutionalized into the infrastructure of most enterprises. This article will portray the conceptual framework of ERP II.

Table 1. Top 5 worldwide ERP software application new license revenue market share estimates for 2002. Source: Gartner Dataquest (June 2003)

Vendor	*2002 Market Share (%)*	*2001 Market Share (%)*
SAP AG	25.1	24.7
Oracle	7.0	7.9
PeopleSoft	6.5	7.6
SAGE	5.4	4.6
Microsoft Business Solutions	4.9	4.6
Others	51.1	50.3
Total Market Share	100.0	100.0

Table 2. Enterprise systems in retrospective

Decade	Concept	Function
1950s	Inventory Control Systems (ICS)	Forecast and inventory management
1960s	Material Requirement Planning (MRP)	Requirement calculations based on Bill-of-Material (BoM)
1970s	Manufacturing Resource Planning (MRP/II)	Closed-loop planning and capacity constraints
1980s	Computer Integrated Manufacturing (CIM)	Automation, Enterprise models
1990s	Enterprise Resource Planning (ERP)	Integrated processes

BACKGROUND: THE EMERGENCE OF THE ERP CONCEPT

The ERP II concept may be understood by taking a closer look at the development of the ERP concept. Enterprise systems have often been explained through the historical evolution of ERP (Chen, 2001; Klaus, Rosemann & Gable, 2000; Wortmann, 1998). The concept of Enterprise Systems (ES) has evolved over almost 50 years, driven by the changing business requirements, the new information technologies, and by the software vendor's ability to provide standardized solutions.

The fundamental structure of ERP has its origin in the 1950s and in the 1960s with the introduction of computers into business. The first applications were automating manual tasks such as bookkeeping, invoicing and reordering. The early Inventory Control (ICS) systems and Bill of Material (BOM) processors gradually turned into the standardized Material Requirements Planning (MRP). The legacy of the IBM's early COPICS specifications can be found in the structure of the systems even today.

The development continued in the 1970s and in the 1980s with the MRP II and the CIM concept. During the 1970s MRP caught on like wildfire, and was fueled by the "MRP Crusade" of the American Production and Inventory Control Society (APICS). But gradually industry came to the understanding that neither of these concepts was able to meet the expectations. Even though the CIM ideas failed in many aspects, the research, for example, on IS development (ISD) and enterprise models, provided the background for gradually integrating more areas into the scope and of the information systems (Wortmann, 2000). This development peaked in early 1990s with the advent of the Enterprise Resource Planning (ERP) systems, often embodied in SAP R/3 (Bancroft, Seip, & Sprenge,l 1997), along with the other major vendors: Oracle, Peoplesoft, JD Edwards and Baan; the so-called JBOPS. Although the ERP systems have other legacies like accounting, the prevailing planning and control philosophy is deeply rooted in manufacturing and in MRP.

ENTERPRISE RESOURCE PLANNING II

The ERP market experienced a hype based on the Y2K problem, but after Y2K the ERP market soured. It was doubted that traditional ERP could meet the e-business challenge (Mabert, Soni, & Venkataramanan, 2001). New vendors of the "bolt-on" systems like, for example, i2 Technology with SCM and Siebel with CRM emerged on the scene (Calloway, 2000) and Application Integration (EAI) became a critical issue (Evgeniou, 2002). New delivery and pricing methods like ASP (Application Service Provider) and ERP rentals were conceived (Harell, Higgins, & Ludwig, 2001) and the traditional ERP vendors were challenged.

The ERP II concept is a vision original conceived by Gartner Group in 2000. Gartner Group, who also put the name on the ERP concept, defines ERP II as: "a business strategy and a set of industry-domain-specific applications that build customer and shareholder value by enabling and optimizing enterprise and inter-enterprise, collaborative-operational and financial processes" (Bond et al., 2000).

ERP II builds on ERP and thus the concept excludes the "bolt-on" vendors like i2 or Siebel from this vision (Mello, 2001). AMR Research does not restrict their competing vision on Enterprise Commerce Management (ECM) to the ERP vendors and define ECM as: "a blueprint that enables clients to plan, manage, and maximize the critical applications, business processes and technologies they need to

support employees, customers, and suppliers" (http://www.amrresearch.com/ECM). GartnerGroup has later resigned on this requirement and today ERP II is a framework which includes enterprise systems based on "Best of Breed" systems and EAI (Light, Holland, & Willis, 2001) as well as "Single Vendor" solutions.

ERP II includes six elements that touch business, application and technology strategy: (i) the role of ERP II, (ii) its business domain, (iii) the functions addressed within that domain, (iv) the kinds of processes required by those functions, (v) the system architectures that can support those processes, and (vi) the way in which data is handled within those architectures. With the exception of architecture, these ERP II elements represent an expansion of traditional ERP. ERP II is essentially componentized ERP, e-business and collaboration in the supply chain (Bond, 2001).

Throughout the ERP industry, the new philosophies of e-business was gradually incorporated into the legacy ERP systems, and system architectures were redesigned and modularized, for example, like SAP did with their NetWeaver platform. Consequently, the contemporary standard systems today incorporate the ERP II vision. The ERP industry survived the challenge and a recent market analysis does not render any signs of market fragmentation, but rather a consolidation. Today, we see an ERP market consisting of one dominant actor, a handful of major vendors and a larger number of vendors of minor significance (also c.f. Table 1).

Today, all the major vendors have adopted the ERP II concept, either partly or to the full extent. The evolution has been driven by the emerging business requirements and by the possibilities offered by the new information technology, exactly the same as the evolution of the ERP concept.

The new technologies are not necessarily inventions of the ERP vendors, but rather the technologies emerge as stand alone systems and after a while they are adopted by the major vendors and then incorporated into the standard systems. That happened to, for example, application frameworks (.NET or J2EE), the databases (Oracle or MS SQL) or Decision-Support Systems (DSS). Business Intelligence (BI) is an example of an analytical DSS technology previously associated with add-ons, like Data warehouse systems based on OLAP tools, and are now integrated into the core of the standard databases. BI refers to a broad category of analytical applications that helps companies make decisions based on the data in their ERP systems. Another example is the Internet standards like XML, originally conceived outside the control of the major vendors but gradually adopted into the infrastructure of the ERP systems. Other examples are the Supply Chain Management (SCM) systems or the Customer Relationship Management (CRM) systems from third-party vendors like i2 or Siebel. Those third-party vendors experienced a short explosive growth, but then when the technologies are incorporated into the standard ERP system the potential business benefit increases.

FUTURE TRENDS: THE NEXT-GENERATION ERP

There is an emerging pattern of stable generic application architectures which we choose to portray as the ERP II con-

Table 3. The four layers of ERP II

Layer		*Components*
Foundation	Core	Integrated Database (DB) Application Framework (AF)
Process	Central	Enterprise Resource Planning (ERP) Business Process Management (BPM)
Analytical	Corporate	Supply Chain Management (SCM) Customer Relationship Management (CRM) Supplier Relationship Management (SRM) Product Lifecycle Management (PLM) Employee Lifecycle Management (ELM) Corporate Performance Management (CPM)
Portal	Collaborative	Business-to-consumer (B2C) Business-to-business (B2B) Business-to-employee (B2E) Enterprise Application Integration (EAI)

cept. Calloway (2000) and Weston (2003) have attempted to frame this overall development and partial aspects have further been dealt with, for example, by Wortmann (2000) and by Møller (2005b). The conceptual framework of ERP II illustrated in Figure 1 consists of four distinct layers, as exhibited in Table 3: (i) the foundation layer, (ii) the process layer; (iii) the analytical layer and (iv) the e-business layer consisting of the collaborative components.

Core Components

The foundation layer consists of the core components of ERP II. The core components shape the underlying architecture and provide a platform for the ERP II systems. ERP II does not need to be centralized or monolithic. One of the core components is the **integrated database**, which may be a distributed database. Another core component is the **application framework,** likewise potential distributed. The integrated database and the application framework provide an open and distributed platform for ERP II.

Central Components

The process layer of the concept is the central component of ERP II, and this layer reflects the traditional transaction-based systems. ERP II is **Web-based open and componentized**. This is different from being Web-enabled, and the ultimate ERP II concept may be implemented as a set of distributed Web services.

Enterprise Resource Planning (ERP) is one of the central components in the ERP II conceptual framework. The backbone of ERP is the traditional ERP modules like financials, sales and distribution, logistics, manufacturing, or HR. ERP still makes up the backbone of ERP II along with the additional integrated modules aimed at new business sectors outside the manufacturing industries.

The ERP II concept is based on **Business Process Management** (**BPM**). ERP has been based on "Best-practice" process reference models but ERP II systems build on the notion of the process as the central entity and ERP II include tools to manage processes: design (or orchestrate) processes, to execute and to evaluate processes (Business Activity Monitoring) and in ERP II redesigning processes will have effect in real-time.

The BPM component allows for ERP II to be accommodated to suit different business practices for specific business segments that otherwise would require problematic customization. ERP II further includes vertical solutions for specific segments like apparel and footwear or the public sector. Vertical solutions are sets of standardized preconfigured systems and processes with "add-ons" to match the specific requirements in, for example, a business sector.

Corporate Components

The analytical layer consists of the corporate components that extend and enhance the central ERP functions by providing decision support to manage relations and corporate issues. Corporate components are not necessarily synchronized with the integrated database and the components may easily be "add-ons" acquired by third-party vendors. The most common components such as Supply Chain Management (SCM) systems and Customer Relationship Management (CRM) systems are listed in Table 3 and illustrated in Figure 1, and they will further be explained under the terms and definitions at the end of this article.

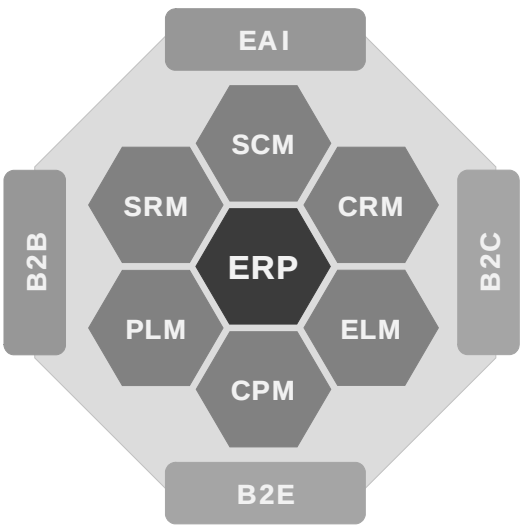
In the future, we can expect new breeds of corporate components. Product Lifecycle Management is already established as the R&D equivalence to ERP and Employee Lifecycle Management is emerging as an example of a people-oriented component.

Collaborative Components

The e-business layer is the portal of the ERP II systems and this layer consists of a set of collaborative components. The collaborative components deal with the communication and the integration between the corporate ERP II system and actors like customer, business partners, employees, and even external systems.

The most common and generic components are listed in Table 3 and illustrated in Figure 1, and will further be explained under the terms and definitions at the end of this article.

Figure 1. The conceptual framework for ERP II



CONCLUSION

ERP II is a vision of the next-generation ERP and the conceptual framework for ERP II is a generic model of the emerging architecture of the contemporary enterprise systems. All major ERP vendors have implemented major parts of this concept (SAP NetWeaver, Oracle Fusion and Microsoft Dynamics). ERP II is aimed at extending the reach of integration into the supply chain and the business benefits of the systems are only realized when the integration occurs (Davenport, Harris, & Cantrell, 2004). Business managers will therefore need to consider their entire range of enterprise systems into a supply chain integration context, and future research will deal with interorganizational integration based on the ERP II.

REFERENCES

Bancroft, N. H., Seip, H., & Sprengel, A. (1997). *Implementing SAP R/3: How to introduce a large system into a large organization*. Greenwich: Manning.

Bond, B., Genovese, Y., Miklovic, D., Wood, N., Zrimsek, B., & Rayner, N. (2000). *ERP is dead—long live ERP II* (No. Strategic Planning SPA-12-0420). GartnerGroup.

Callaway, E. (2000). *ERP—the next generation: ERP is WEB enabled for e-business*. South Carolina: Computer Technology Research Corporation.

Chen, I. J. (2001). Planning for ERP systems: Analysis and future trend. *Business Process Management Journal, 7*(5), 374.

Davenport, T. H. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review, July/August*, 121-131.

Davenport, T. H. (2000). The future of enterprise system-enabled organizations. *Information Systems Frontiers, 2*(2), 163-180.

Davenport, T. H., & Brooks, J. D. (2004). Enterprise systems and the supply chain. *Journal of Enterprise Information management, 17*(1), 8-19.

Davenport, T. H., Harris, J. G., & Cantrell, S. (2004). Enterprise systems and ongoing process change. *Business Process Management Journal, 10*(1), 16-26.

Evgeniou, T. (2002). Information integration and information strategies for adaptive enterprises. *European Management Journal, 20*(5), 486-494.

Harrell, H. W., Higgins, L., & Ludwig, S. E. (2001). Expanding ERP application software: Buy, lease, outsource, or write your own? *Journal of Corporate Accounting & Finance, 12*(5).

Klaus, H., Rosemann, M., & Gable, G. G. (2000). What is ERP? *Information Systems Frontiers, 2*(2), 141-162.

Light, B., Holland, C. P., & Wills, K. (2001). ERP and best of breed: A comparative analysis. *Business Process Management Journal, 7*(3), 216.

Mabert, V. A., Soni, A., & Venkataramanan, M. A. (2001). Enterprise resource planning: Common myths versus evolving reality. *Business Horizons, 44*(3), 69-76.

Mello, A. (2001, October 25). Battle of the labels: ERP II vs. ECM. *ZD TECH Update*.

Møller, C. (2005a). Unleashing the potential of SCM: Adoption of ERP in large Danish enterprises. *International Journal of Enterprise Information Systems, 1*(1), 39-52.

Møller, C. (2005b). ERP II: A conceptual framework for next-generation enterprise systems? *Journal of Enterprise Information Management, 18*(4), 483-497.

Nah, F. F.-H. (Ed.). (2002). *Enterprise resource planning solutions and management*. IRM Press.

Weston, E. C. T. (2003). ERP II: The extended enterprise system. *Business Horizons, 46*(November/December), 49-55.

Wortmann, J. C. (1998). Evolution of ERP systems. In U. S. Bititchi & A. S. Carrie (Eds.), *Strategic management of the manufacturing value chain*. Kluwer Academic.

KEY TERMS

Enterprise Resource Planning (ERP) II systems are second generation ERP systems. ERP II extends on the ERP concept (see Figure 1). The ERP II vision is framed by Gartner Group and in practice defined by the contemporary systems of the major ERP vendors.

Supply Chain Management (**SCM**) systems provide information that assist in planning and managing the production of goods. For instance, SCM assists in answering questions such as where the good is to be produced, from what the parts are to be procured and by when it is to be delivered.

Customer Relationship Management (**CRM**) systems facilitate the managing of a broad set of functions relating to managing customers relations that primarily include the categories of customer identification process and customer service management.

Supplier Relationship Management (**SRM**) is the vendor side analogy to CRM aimed at the effective management of the supplier base. SRM facilitates the management of the supplier relations in its entire life-cycle.

Product Lifecycle Management (**PLM**), including Product Data Management (PDM), enables enterprises to bring innovative and profitable products to market more effectively, especially in the evolving e-business environment. PLM enables enterprises to harness their innovation process through effective management of the full product definition lifecycle in their extended enterprises.

Employee Lifecycle Management (ELM) is the integration of all aspects of information and knowledge in relation to an employee from the hiring to the retirement from the company. ELM enables enterprises to effectively manage their portfolio of competencies.

Corporate Performance Management (CPM), or sometimes Enterprise Performance Management (EPM), is an umbrella term that describes the methodologies, metrics, processes and systems used to monitor and manage an enterprise's business performance. Thus, CPM provides the managements with an overall perspective on the business.

Consumer to business (B2C), or e-commerce systems, deals with the carrying out of commercial transactions with businesses or with individual customers by using the Internet as an electronic medium. This requires an extensive infrastructure of which the main features are a catalogue, online ordering facilities and status checking facilities.

Business to business (B2B), or e-procurement systems, improves the efficiency of the procurement process by automating and decentralizing the procurement process. The traditional methods of sending Request for Quotes (RFQ) documents and obtaining invoices and so forth, are carried out over the Web through purchasing mechanisms such as auctions or other electronic marketplace functions, including catalogues.

Business to employee (B2E), intranets or knowledge management systems, provide the employee with an updated personalized portal to the enterprise on his desktop. The perspectives of the intranet and knowledge management systems increase in the context of the ERP II concept.

Enterprise Application Integration (EAI), or extranets, provides the ERP II system with a portal and a platform for integration with other systems inside or outside the corporation. EAI provides the support for automating processes across various IT platforms, systems and organizations.

The Nomological Network and the Research Continuum

Michael J. Masterson
USAF Air War College, USA

R. Kelly Rainer, Jr.
Auburn University, USA

INTRODUCTION

Social science and management information systems (MIS) research have been criticized for failure to integrate theory construction and theory testing (see e.g., Subramanian & Nilakanta, 1994). In particular, concerns with MIS as a cohesive research discipline have long included inadequate construct development and lack of valid, reliable measuring instruments for those constructs (Keen, 1980). Understanding the theoretical basis of constructs and how they are developed and tested across the research continuum are fundamentals of a cohesive academic discipline. To provide a common research framework for the growth of MIS as a scientific discipline, this chapter proposes a framework for an integrated research continuum across the life cycle of the research process.

BACKGROUND

The growth of any scientific discipline entails the development of a system of specialized, abstract concepts that define the lawful relationships (or hypotheses) that represent a discipline's theories. Scientists recognize underlying concepts in observed phenomena and build those concepts into the lawful relationships of a scientific theory (Blalock, 1982, Hempel, 1952).

This search for lawful relationships among natural events is the primary function of science, and focuses on two objectives: (1) describing specific events, objects, or phenomena in the world of experience and (2) establishing theories, or general principles, to explain or predict the specified events, objects, or phenomena (Feigl, 1970 Hempel, 1952).

Theories of a science must support explanation and prediction. If a scientific discipline lacks explanatory and predictive theories, no connection can be established between descriptions of different events, objects, or phenomena. Such a nontheory based discipline is unable to predict or prepare for future occurrences. The lack of explanatory principles permits no use of theory for practical application. Practical application requires theories or principles that explain what particular effects occur when specific changes occur in a given system. In addition, comparability tests of the theory by other researchers would be impossible (Blalock, 1982).

SCIENTIFIC THEORY AS A SPATIAL NETWORK AND THE RESEARCH CONTINUUM

To develop precise theories of wide scope and high empirical confirmation, scientific disciplines create and evolve comprehensive systems describing lawful relationships of theoretical constructs (Hempel, 1952). Some variables of interest within these theoretical constructs cannot be directly observed. Thus, constructs contain theorized unobservable, latent factors measured by empirically observed indicators (Nunnally, 1978).

A comprehensive scientific theory can be represented as a spatial network in which hypotheses use correspondence rules to link theoretical constructs to derived and observed empirical concepts, which acquire meaning through operational definitions. The unobservable (latent) theoretical constructs are anchored to the empirical environment by rules of interpretation (the correspondence rules). By virtue of these interpretive connections, the network can function as a scientific theory (Blalock, 1982). The ability to interpret unobservable, underlying constructs transforms a theoretical, spatial network into an empirically testable theory.

An important implication of this view of the structure of theory is that the integration of theory construction and theory testing across the research continuum is of major importance (Feigl, 1970; Hempel, 1952; Hughes, Price, & Marrs, 1986; Trochim, 1996). Figure 1 (below) is a graphic depiction of this spatial framework, developed from the work of Hempel (1952), Blalock (1982), and Trochim (1996).

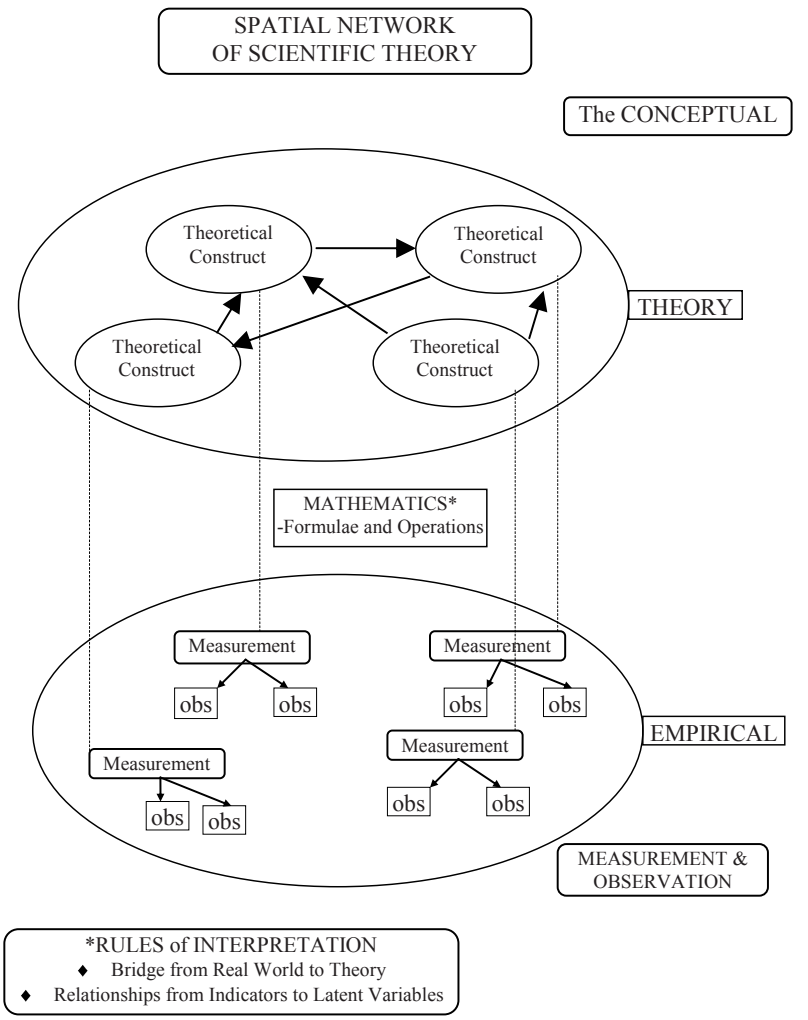
Figure 1 represents a 1:1 correspondence between theoretical constructs and measurements, which is not always the case. Observable measurements, which are qualitatively different in the empirical world, can overlap or measure the same thing if their positions in this spatial network link them to the same theoretical construct. Also, a measurement may serve as the interpretive link to multiple theoretical

constructs, though the researcher would specify the operative definition in use for that interpretive relationship, for example, measuring self-esteem requires accounting for elements of self-image and ego (Cronbach & Meehl, 1955; Hempel, 1952; Trochim, 1996).

In a spatial network of theory and observation, internal principles define basic entities (concepts and constructs) of the theory, and the hypotheses describe interrelationships of these theoretical constructs, either within the same theory or with other theories. (see "The CONCEPTUAL" in Figure 1). Bridge principles are the mathematical formulae and operations, stated as the rules of interpretation, that link the processes proposed by the theory to empirical phenomena (measures and observations). Bridge principles enable a theory to be used for explanations and/or prediction (see "EMPIRICAL" and "RULES of INTERPRETATION" in Figure 1). Without these principles, a theory has no explanatory power or practical application, and no empirical test is possible.

A fundamental requirement of science is that a statement of fact from an appropriate theory made by one scientist must be independently verifiable by other scientists. This principle of scientific generalization gains wide belief and support for any theory and is central to all scientific work (Nunnally, 1978).

Figure 1.



Unlike the physical sciences, social and behavioral sciences study living organisms to define and classify behavior, and the hypotheses linking behavior to conditions that control it. Psychological measurement, or psychometrics, is inseparable from this process (Nunnally, 1978). Evaluating the adequacy of a theoretical spatial network requires judging empirical testability (Popper, 1959), verifiability (Dodd, 1968), and confirmability (Clark, 1969). Meeting these standards requires a clear and explicit specification of theoretical construct definitions and operationalizations across the research continuum (Hughes et al., 1986). Statistics is the area of mathematics used in evaluations of theoretical constructs.

Scientific behavioral studies begin with observations or questions about some object, event, or phenomenon, in a search for the functional relationships, the laws, and the principles of behavior. This behavior must be described in a manner that fully communicates all observations to others so that they may verify the observation. (Blalock, 1982).

The Nomological Network. Cronbach and Meehl (1955) defined a nomological network as the interlocking system of hypotheses, principles, and laws linking the constructs that constitute any theory. A nomological network encompasses the theoretical constructs being measured, how the concept is going to be measured, and the specification of the interrelationships between the theoretical and empirical planes (see Figure 1). All three aspects of the spatial network are present in a nomological network: constructs, empirical observations/measures, and the mathematical rules of interpretation that link theory to observation. Scientifically, the purpose of the nomological network is to clarify all parts of a theory, so a theory's laws of explanation and prediction can be exploited.

As a standard for tests of social and behavioral constructs, the nomological network focuses on sampling statistics of content (test items), not sampling statistics of people. As empirical measures of a theory are developed, the theory formulation that guides the development of the mathematical formulae and operations focuses on sampling content (Blalock, 1982; Nunnally, 1978). The goal of sampling content is to generalize findings over populations of test items. The researcher addresses what, not who, is being measured (Cronbach, 1975). Thus, the specific focus of the nomological network is not the subject. Rather, the nomological network measures observable properties, the relationship of different theoretical constructs to each other, and the relationship of the theoretical constructs to the observable properties via the mathematical rules of interpretation (Cronbach, 1975; Cronbach & Meehl, 1955). Also, at least some of the laws in the nomological network must involve observables, otherwise no bridge could exist to the theoretical constructs. (Cronbach & Meehl, 1955; Trochim, 1996).

Cronbach and Meehl (1955) defined construct validity as the process of following all these principles of the nomological network to ensure generalizability of scientific principles. Construct validity specifically addresses the question, "Does the test measure the attribute it is said to measure?" Construct validation identifies theoretical constructs defined by a network of relations, all of which are anchored to observables, making the constructs testable (Cronbach, 1975). This process is a direct identification of the spatial network (see Figure 1) and the nomological network, including the constructs (concepts) of interest in a theory, their observable manifestations, and the interrelationships among the constructs, measures, and observations.

The social and behavioral sciences, including MIS, face difficulties that vary considerably from those encountered in the physical sciences. These difficulties stem from two fundamental issues, which are:

1. the relationship between theory and research, and
2. the development of theoretically defined constructs and measurement procedures.

When the phenomena of interest cannot be directly observed, the development of latent constructs and instruments to operationalize them provide the theoretical basis for research in a discipline (Venkatraman & Grant, 1986). To meet this basic criterion of science, the focus must be on generalizability and the comparability of one's measurements across diverse settings (Hughes et al, 1986; Nunnally, 1978). Blalock (1982) identifies this focus as one of conceptualization and measurement, where

1. Conceptualization refers to the theoretical process by which researchers move from ideas or constructs to suggesting appropriate research operations.
2. Measurement refers to the linkage process between the physical operation, on the one hand, and a mathematical language on the other.

Systems theory is commonly used as a generic concept to represent the elements represented in this triple linkage (Blalock, 1982; Trimmer, 1950). The essential point in connecting theoretical constructs, physical measurement operations, and mathematical formulae and operations, is that the basic concepts or variables be specified with sufficient clarity so that

1. the "systems" or unit to which they are intended to refer is known;
2. it is clear how to distinguish them from other stimuli, properties, or responses with which they are likely to be confounded;

3. it has been specified how concepts that refer to one type of system are related definitionally, or by some presumed causal law, to at least some concepts that refer to any other systems that are also expected to play important roles in substantive theories.

This last specification implies that it is often desirable to use information provided at one level of analysis (e.g., about individual behaviors) to support, refute, or clarify theories formulated at other, unobservable levels. Psychologists successfully integrated the methodological tools of empirically minded investigators presented in the scaling and measurement literature. Unfortunately, this has resulted in an imbalance between theory development and methods of measurement, with development of measures occupying a much larger place in the research process than conceptualization of theory (Ackoff, 1984). This creates the necessity for explicit theoretical formulations that clearly state which theoretical assumptions are being made in each measurement decision across the research continuum (Blalock, 1982).

The Research Continuum. The life cycle of the research process functions as an integrated continuum moving from the most exploratory, descriptive kinds of observation, to generalization, and finally to experimental applied research focused on complete, specific answers (Ackoff, 1984; LaMott, 1997; Vandendorpe, 1997). The clear focus on the nonhierarchical developmental aspect of basic research supports a model of the research continuum in which phases of the research life cycle are integrated, rather than performed separately. This holistic approach lets the researcher determine research strategy, orientation, and criteria (Cronbach & Meehl, 1955; Yin, 1994). A researcher determines the type of research strategy by choosing one of the following familiar series of questions to ask: "who," "what," "where," "how," and "why" (Hedrick et al., 1993; Yin, 1994).

"What" questions are exploratory and deal with conceptual discussions and initial theory formulation examining observed phenomena. "Who" and "where" questions favor survey strategies or analysis of archival records to seek generalizable patterns. Frameworks (i.e., a taxonomy of classification based on standard guidelines) are developed to explain, organize, and manage identified patterns (see e.g., Kochen, 1986; Merrill & Tenneyson, 1977). Frameworks in the field of MIS, for example Nolan and Wetherbe (1980), are typically validated by mapping MIS research publications to the framework to determine if the framework properly classifies the published material. This method organizes the research and enables recognition of conceptual patterns. These strategies support research goals that are either descriptive or predictive. "How" and "why" questions are more explanatory, and lead to the use of single or multiple case studies and experiments as the preferred research strategy. The experiments could be either field experiments or laboratory experiments, again determined by the researcher's orientation (Latane & Darley, 1969; Yin, 1994).

Surveys help researchers generalize their findings. One survey development example in MIS research is the development of the end-user computing construct (Rainer & Harrison, 1993), which used Churchill's (1979) survey development process.

When researchers focus on specific "how" and "why" criteria, field experiments are in order. Data are collected from subjects in the environment where the construct of interest exists. Finally, with the proper constructs of interest (e.g., computer skills) identified, researchers set up laboratory experiments to obtain observable outcomes (see e.g., Miller, 1994).

Theory building takes place in the exploratory and generalization phases of the research process, while theory testing occurs in the generalization and experimental phases. This blending of phases, where theory development or testing are accomplished, demonstrates the integration of research across the continuum, as depicted in Figure 2. The research continuum (LaMott, 1997; Yin, 1994) supports theory building and theory testing in an interactive model.

The process of basic research as a complex, pluralistic feedback model integrates the research life cycle across a broad continuum. Research does not follow a rigidly structured, waterfall life cycle. Instead, a researcher can spiral across and down the continuum on a specific research track. In this spiral process, different strategies and phases of the research process are determined by the orientation and criteria set by the researcher. This process includes the peculiarity of an observation being applied, in practice, with no theoretical foundation (Kochen, 1986).

In 1997, *R&D Magazine*'s basic research survey of 4,000 researchers, 67.3% stated that the most critical element of basic research developed an understanding of scientific principles or phenomena. They noted that the second most critical element was creation of a foundation for future development. These two critical elements point out the importance and currency of the spatial network of theory development and the nomological network (Vandendorpe, 1997).

Different strategies can be employed across the research continuum. The changing orientation and criteria of the researcher determines strategy. The nature of the following components and dimensions of the research continuum become more specific and increase in importance in moving from conceptual studies to laboratory experiments (Blalock, 1982; Nunnally, 1978):

1. knowledge of subjects and environment;
2. questions and objectives and data collection procedures;
3. knowledge of and accuracy in operationalizing variables and relationships;
4. potential internal validity;

Figure 2.

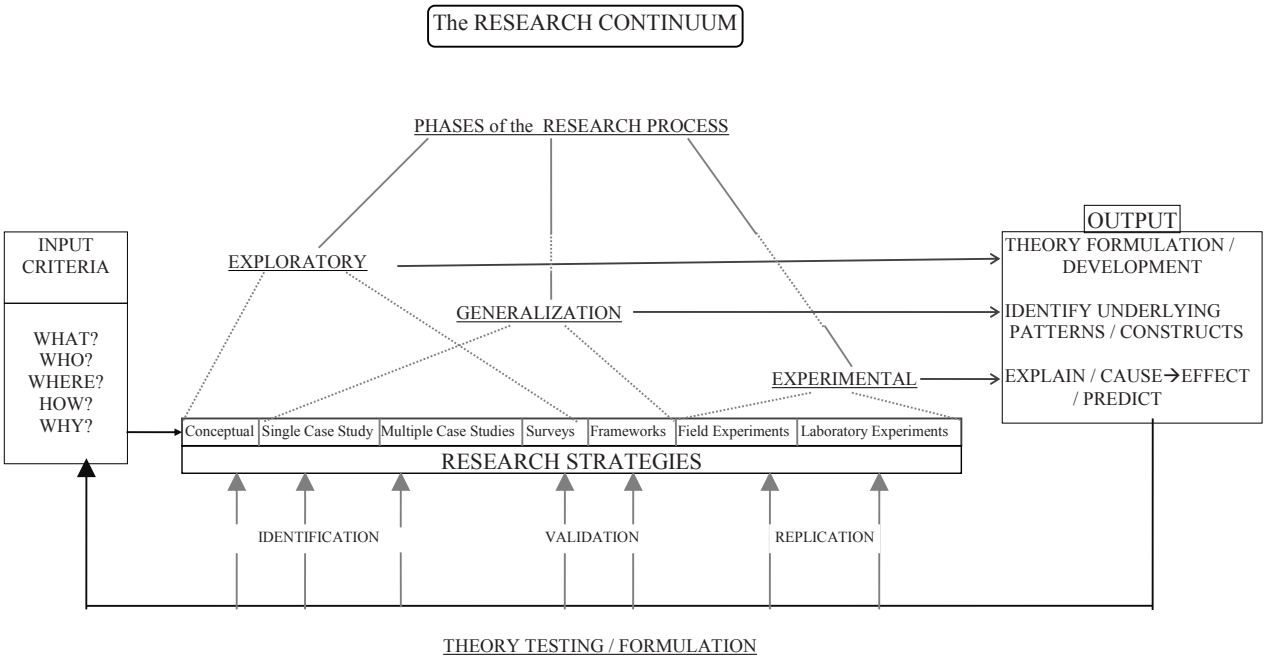
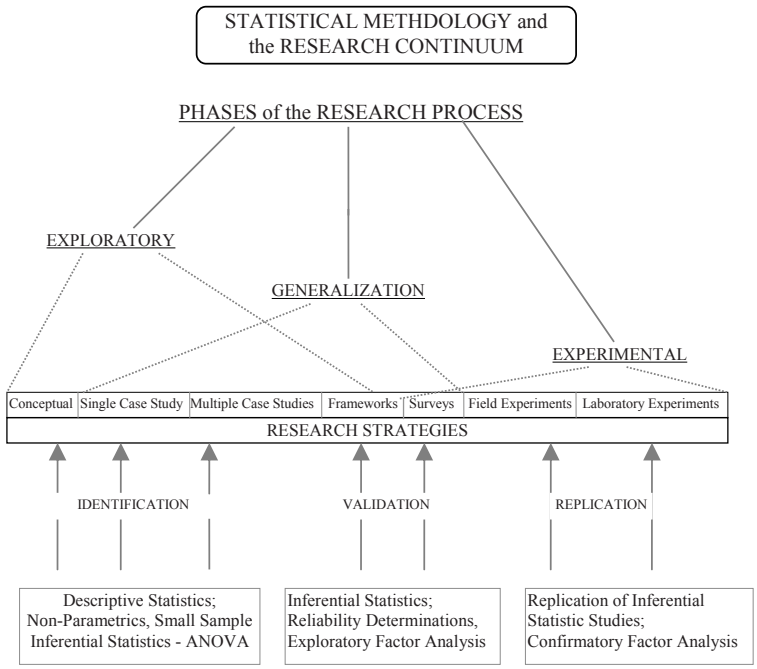


Figure 3.



5. potential external validity;
6. potential reliability;
7. possible sources of bias.
8. ability to predict;
9. ability to uncover causal models.

Also, the statistical methodologies employed by a researcher change across the research continuum from conceptual studies to laboratory experiments (see Figure 3). The change is from descriptive to inferential statistics, from initial descriptive identification of observations to independent confirmation and replication critical to the process of scientific generalization.

FUTURE TRENDS

Although the spatial, nomological network and the research continuum have high face validity, more work needs to be done in these areas. It is interesting to note that the references here are relatively old. The fact is that researchers have not done recent work in these areas.

One direction for future research would be to validate the nomological network and research continuum proposed here. Researchers could take a subject area in a discipline (e.g., decision support systems in MIS), map its constructs to the nomological, spatial network and map its published articles to the research continuum.

CONCLUSION

Social and behavioral sciences face two very different issues from those encountered in the physical sciences: (1) the relationship between theory and research, and (2) the development of theoretically defined constructs and measurement procedures. By understanding the integrated and interactive nature of the research continuum across the life cycle of research, and the fundamental requirement for development of theory out of that research, investigators can advance the rigor of their chosen academic disciplines, and build a testable body of work to support and grow their chosen scientific discipline.

This research continuum, grounded in the triple linkage depicted in the spatial network, is robust and flexible. However, the challenge in not having readily observable and measurable data outcomes across the research continuum makes the operationalization and measurement of constructs for a researcher difficult, challenging, and tedious (Blalock, 1982). Understanding the theoretical basis of a construct, how it is developed and tested within a nomological network, and the interplay across the research continuum, are fundamental in contributing to the establishment and future development of MIS and social sciences as cohesive academic disciplines.

REFERENCES

Ackoff, R. L. (1984). *Scientific method: Optimizing applied research decisions*. Malabar, FL: Robert E. Krieger Publishing Company, Inc.,

Blalock, H. M. (1982). *Conceptualization and measurement in the social sciences*. Beverly Hills, Ca: Sage Publications.

Churchill, G. A. Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*, 64-73.

Clark, J. T. (1969). The philosophy of science and the history of science. In M. Clayett (Ed.), *Critical problems in the history of science*, (pp. 103-140). Madison, WI: University of Wisconsin Press.

Cronbach, L. J. (1975). The validation of educational measures. In D. A. Payne & R. F. McMorris, (Eds.), *Educational and psychological measurement: Contributions to theory and practice*, (pp. 75-88). General Learning Corporation, General Learning Press.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-303.

Dodd, S. C. (1968). Systemmetrics for evaluating symbolic systems. *Systemmatics, 6*, 27-49.

Feigl, H. (1970). The "orthodox" view of theories: Remarks in defense as well as critique. In M. Radnor, & S. Winokur (Eds.), *Minnesota Studies in the Philosophy of Science*, (pp. 4). Minneapolis, MN: University of Minnesota Press.

Hedrick, T., Bickman, L., & Rog, D. J. (1993). *Applied research design.* Newbury Park, CA: Sage Publications.

Hempel, C. G. (1952). Fundamentals of concept formation in empirical science. *Foundations of the Unity of Science*, vol. II, #7. Chicago: University of Chicago Press.

Hughes, M. A., Price, R. L., & Marrs, D. W. (1986). Linking theory construction and theory testing: Models with multiple indicators of latent variables. *Academy of Management Review, 11*, 128-144.

Keen, P. G. W. (1980). MIS research: Reference disciplines and a cumulative tradition. In E. R. McLean (Ed.), *Proceedings of the First International Conference on Information Systems* (pp. 9-18).

Kochen, M. (1986). Are MIS frameworks premature? *Journal of Management Information Systems, 2*, 92-100.

LaMott, E. E. (1997). *The research continuum*. Retrieved from http://www.csp.edu/scp/people/lamott/sac581/Research1/index.htm

Latane, B., & Darley, J. M. (1969). Bystander apathy. *American Behavioral Scientist, 57*, 244-268.

Merrill, M. D., & Tennyson, R. D. (1977). *Teaching concepts: An instructional design guide*. Englewood Cliffs, NJ: Educational Technology Publications.

Miller, M. D. (1994). *The extended technology acceptance model: Theory and empirical test*. Unpublished Dissertation Auburn, Al: Auburn University.

Nolan, R. L., & Wetherbe, J. C. (1980). Toward a comprehensive framework for MIS research. *MIS Quarterly, 4*, 1-19.

Nunnally, J. C. (1978). *Psychometric theory*, [2nd ed.]. New York: McGraw-Hill Book Company.

Popper, K. R. (1959). *The logic of scientific discovery*. New York: Harper & Row.

Rainer, R. K., Jr., & Harrison, A. W. (1993). Toward development of the end user computing construct in a university setting. *Decision Sciences, 24*, 1187-1202.

Studt, T. (Ed.). (1997). Basic research white paper. *R&D Magazine*.Highlands Ranch, Co: Reed Elsevier, Inc.

Subramanian, A., & Nilakanta, S. (1994). Research measurement: A blueprint for theory-building in MIS. *Information & Management, 26*, 13-20.

Trimmer, J. D. (1950). *Response of physical systems*. New York: Wiley.

Trochim, W. M. K. (1996). The multitrait-multimethod matrix.Retrieved from http://trochim.human.cornell.edu/kb/mtmmmat.htm

Vandendorpe, L. (1997). Basic research white paper. *R&D Magazine*. Highlands Ranch, Co: Reed Elsevier.

Venkatraman, N., & Grant, J. H. (1986). Construct measurement in organizational strategy research: A critique and proposal. *Academy of Management Review, 11*, 71-87.

Yin, R. K. (1994). *Case study research: Design and methods*, [2nd ed.]. Thousand Oaks, CA: Sage Publications.

KEY TERMS

Basic Research: Research performed without thought of practical ends, producing general knowledge and an understanding of nature and its laws.

Bridge Principles: Mathematical formulae and operations, stated as the rules of interpretation, that link the processes proposed by the theory to empirical phenomena (measures and observations).

Conceptualization: Refers to the theoretical process by which researchers move from ideas or constructs to suggesting appropriate research operations.

Construct Validity: Identifies theoretical constructs defined by a network of relations, all of which are anchored to observables, making the constructs testable.

Internal Principles: Define basic entities (concepts and constructs) of the theory and the lawful relationships, hypothesized as part of theory, describing interrelationships of these theoretical constructs, either within the same theory or with other theories.

Measurement: Linkage process between the physical operation, on the one hand, and a mathematical language on the other.

Measurement by Proclamation: Theoretical claims of lawful relationships with no link to empirical tests (Blalock, 1982).

Nomological Network: Interlocking system of hypotheses, principles, and laws linking the constructs that constitute any theory.

Theories: General principles to which individual events, objects, or phenomena conform, and by which the occurrence of these events, objects, or phenomena is systematically anticipated.

Nonlinear Approach to Brain Signal Modeling

Tugce Balli
University of Essex, UK

Ramaswamy Palaniappan
University of Essex, UK

INTRODUCTION

Biological signal is a common term used for time series measurements that are obtained from biological mechanisms and basically represent some form of energy produced by the biological mechanisms. Examples of such signals are electroencephalogram (EEG), which is the electrical activity of brain recorded by electrodes placed on the scalp; electrocardiogram (ECG), which is electrical activity of heart recorded from chest, and electromyogram (EMG), which is recorded from skin as electrical activity generated by skeletal muscles (Akay, 2000).

Nowadays, biological signals such as EEG and ECG are analysed extensively for diagnosing conditions like cardiac arrhythmias in the case of ECG and epilepsy, memory impairments, and sleep disorders in case of EEG. Apart from clinical diagnostic purposes, in recent years there have been many developments for utilising EEG for brain computer interface (BCI) designs (Vaughan & Wolpaw, 2006).

The field of signal processing provides many methods for analysis of biological signals. One of the most important steps in biological signal processing is the extraction of features from the signals. The assessment of such information can give further insights to the functioning of the biological system.

The selection of proper methods and algorithms for feature extraction (i.e., linear/nonlinear methods) are current challenges in the design and application of real time biological signal analysis systems. Traditionally, linear methods are used for the analysis of biological signals (mostly in analysis of EEG). Although the conventional linear analysis methods simplify the implementation, they can only give an approximation to the underlying properties of the signal when the signal is in fact nonlinear. Because of this, there has been an increasing interest for utilising nonlinear analysis techniques in order to obtain a better characterisation of the biological signals.

This chapter will lay the backgrounds to linear and nonlinear modeling of EEG signals, and propose a novel nonlinear model based on exponential autoregressive (EAR) process, which proves to be superior to conventional linear modeling techniques.

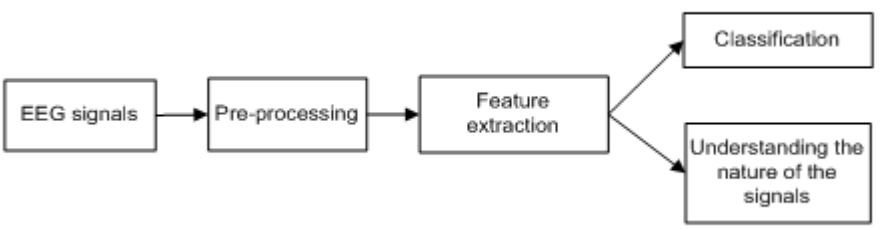
BACKGROUND INFORMATION

EEG Signal Processing

In recent years, the field of biological signal processing has seen an explosive growth. In particular, there have been many research studies on EEG signals for:

- Diagnosis of certain neurological conditions such as sleep disorders, memory impairments and epilepsy;
- Extracting relevant features for classification of different mental states;

Figure 1. The basic steps in EEG signal analysis



- Understanding the dynamics and underlying mechanisms of the brain.

Figure 1 shows the basic steps in the analysis of EEG signals, these are: *preprocessing* which includes the removal of noises such as the baseline noise, powerline interference and eye blink contamination; *feature extraction,* which extracts representative values of the signals through modeling techniques, and *classification,* where the extracted features are classified in specific for the application, such as discrimination between different mental states or neurological conditions. Note that the feature extraction step is not necessarily followed by classification—the features can also be used in understanding the nature and underlying dynamics of the signals, for example in investigating a certain brain disorder. The selection of appropriate feature extraction methods for obtaining a better representation of the EEG signals is the most challenging step in EEG signal processing. This can be approached in two ways namely the linear and nonlinear modeling techniques.

Utilising Linear Modeling Techniques for Analysis of EEG Signals

Since its discovery by Hans Berger in 1929 (Sanei & Chambers, 2007) the EEG signals have been used extensively in research studies for diagnosis of certain neurological conditions (such as memory impairments, sleep disorders, and epilepsy). Traditionally linear modeling techniques like autoregressive (AR) modeling and power spectral estimation (PSD) have been extensively used for the analysis of EEG signals (Sanei & Chambers, 2007).

Palaniappan (2005) used second order AR model coefficients as features for the classification of EEG signals recorded from alcoholic and control subjects. The EEG signals were recorded from subjects while they were exposed to visuals selected from Snodgrass and Vanderwart picture set. The feature sets were classified using three different classification algorithms namely the simplified Fuzzy ARTMAP (SFA) neural network (NN), multilayer-perceptron trained by the backpropagation algorithm (MLP-BP) and Linear Discriminant (LD). The results of this study indicated that the classifiers were able to discriminate the alcoholic and control subjects with average discrimination error of 2.6%, 2.8% and 11.9% for LD, MLP-BP and SFA classifiers respectively.

In another study, Subasi, Kiymik, Alkan, and Koklukaya (2005) characterised and classified EEG segments recorded from epilepsy patients and healthy subjects using PSD values as feature sets. Two different methods were utilised for PSD estimation namely the AR spectral estimation and FFT-based spectral estimation. The feature sets were classified using multilayer feedforward neural network with backpropagation algorithm (MLP-BP). The results of this study indicated an average classification accuracy of 92.3% for AR spectral estimation and 91.6% for FFT-based spectral estimation. The authors also suggested that utilizing nonlinear methods instead of the conventional linear methods would improve the classification accuracy.

Apart from diagnostic purposes, in the last decade there has been an increasing interest in utilising EEG for Brain Computer Interface designs. Keirn and Aunon (1990) were one of the first groups that suggested using EEG as an alternative mode of communication between disabled people and their environment. The different pairs of mental tasks were classified (i.e., baseline, maths, letter composing, geometric figure rotation, and visual counting) using a Bayesian quadratic classifier. They used power asymmetry ratio for creating the feature sets since the mental tasks were identified as belonging to right or left hemisphere of the brain. In addition, they used AR model coefficients as feature sets. Their study showed that the AR method was superior to asymmetry ratios where the most significant result was 84.6% classification accuracy for discrimination of two different mental tasks.

Utilising Nonlinear Modeling Techniques for Analysis of EEG Signals

The individual neurons in the brain behave in a nonlinear manner. There are many research studies reporting more or less successful attempts to apply nonlinear methods to biological time series data (Babloyantz, Salazar & Nicolis, 1985; Bukkapatnam et al, 2002; Gautama, Van Hulle & Mandic, 2003; Lehnertz, Mormann, Kreuz, Anderzak, Rieke & David, 2003; Stepien, 2002).

One of the first studies on nonlinear EEG analysis was by Babloyantz et al. (1985). In this study it was shown that certain nonlinear measures (i.e., Correlation Dimension) change during low-wave sleep patterns. In other words, different sleep stages could be discriminated using these nonlinear measures. After this study the nonlinear methods began to attract the interest of many researchers. Nonlinear methods have been applied mainly to areas such as diagnosis of epileptic seizures and sleep disorders (Chippa & Bengio, 2003).

Bukkapatnam (2007) characterized and classified two different mental conditions from EEG signals using the theory of nonlinear dynamical systems. In this study, 64 channel signals of length 256 samples recorded from 20 people were used. Out of 20 EEG signals used, 10 were obtained from people under alcoholic influence and the remaining ten were recorded from people in a normal (non-alcoholic) condition. The feature sets were created by calculating the correlation dimension of the EEG segments (where this measure quantifies the nonlinear complexity of the signals) (Sanei & Chambers, 2007). The created feature sets were used as an input to a two layer back propagation neural network. The

classifier was able to distinguish between subjects under alcoholic influence and the control subjects with 90% accuracy. The results of this study indicated that EEG signals could be described as noise-contaminated, nonlinear, and perhaps chaotic dynamic systems.

In Stepien et al. (2002), an analysis of spontaneous EEG of 21 healthy subjects recorded when they were resting was conducted. The EEG signals were tested if they were generated by a nonlinear process using surrogate data method (Theiler, Eubak, Longtin, Galdrikian & Farmer, 1992), where the nonlinear prediction error was used as a test statistic. Out of 336 (from 21 subjects with 16 channels) EEG segments, only 17 (5%) of them were found to be nonlinear. The results of this study indicated very low percentage of nonlinearity in the EEG signals recorded from healthy subjects. However, the existence of nonlinearity in various pathological states like epilepsy is indicated by Lehnertz et al. (2003) and Gautama et al. (2003). The existence of this distinguishing feature between normal and diseased cases would allow improved classification of EEG signals using nonlinear methods.

In another previous study done by Gautama et al. (2003), the nonlinearity of EEG signals recorded from healthy and epilepsy patients was investigated. In total, five sets of EEG data were utilised where the sets A and B were recorded from healthy subjects with eyes open and closed, the sets C and D were recorded from epilepsy patients during seizure-free interval from epileptogenic zone and from outside of epileptogenic zone, respectively. And the set E contained the EEG segments recorded from epilepsy patients during seizure activity recorded from seizure generating areas. The nonlinearity of the EEG segments was assessed by surrogate data method (Theiler et al., 1992) where the delay vector variance, third order autocorrelation and asymmetry due to time reversal methods were used for the characterisation of time series (Gautama et al., 2003). The results of this study indicated that the percentage of nonlinearity is lower for EEG segments recorded from healthy subjects (i.e., with eyes open and closed: sets A and B) compared to epilepsy patients (i.e., during seizure and seizure free intervals: sets C, D, and E). These results show that there are clear differences in dynamical properties of the electrical activity of the brain recorded from different physiological and pathological brain states.

Lehnertz et al. (2003) indicated in his article that there are plenty of evidences in the literature that nonlinear EEG analyses are able to characterize the neuronal behavior in the brain and provide a tool for detecting the preictal state in the epilepsy patients. However the sufficiency of sensitivity[a] and specificity[b] of these analysis techniques are still subject to current research. The development of new time series analysis techniques that will sufficiently represent the nonlinear, chaotic and multidimensional behavior of the EEG signals will improve the understanding of the brain dynamics. Once enough specificity and sensitivity is obtained from these analysis techniques, more extensive clinical studies and the implementation of such systems can be considered in the future.

EXPONENTIAL AUTOREGRESSIVE MODEL–A RECENT NONLINEAR MODELING TECHNIQUE FOR ANALYSIS OF EEG SIGNALS

Haggan and Ozaki (1981) introduced EAR algorithm for modeling nonlinear fluctuations in time series. They stated that the analysis of stochastic processes have been mostly done using some form of linear time series modeling and this can only provide an approximation to the underlying properties of the signals. Besides, it is found that many signals exhibiting random vibrations display nonlinear behavior; hence a nonlinear model that gives a good approximation to the underlying properties of a signal is required.

The EAR model exhibits certain features of random vibrations that do not occur in linear models namely the amplitude-dependent frequency, jump phenomena and limit cycle (Haggan & Ozaki, 1981).

An EAR model of order p is defined by;

$$x_t = \sum_{k=1}^{p} (\varphi_k + \pi_k \cdot e^{-\gamma x_{t-1}^2}) \cdot x_{t-k} + e_t \tag{1}$$

where φ, π, γ are autoregressive coefficients, x_t is data at sampled point t, p is the model order and e_t is Gaussian white noise with mean zero.

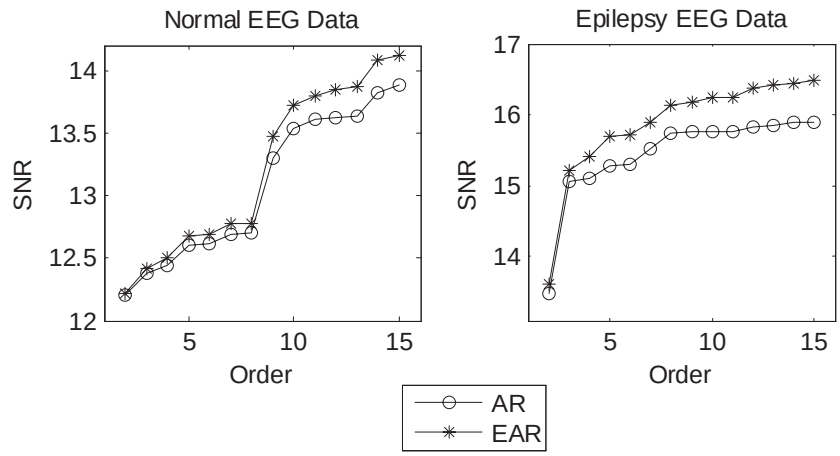
The nonlinearity of the EAR model comes from the exponential term, $e^{\gamma \cdot x(n-k)^2}$, which makes the series globally nonlinear. If nonlinear parameter γ is set to 0, the equation will become an ordinary linear AR model with coefficients $a_p = \varphi_p + \pi_p$ such that;

$$x_t = \sum_{k=1}^{p} a_k \cdot x_{t-k} + e_t \tag{2}$$

The estimation of the $2p$+1 coefficients $\{\gamma, (\varphi_i, \pi_i, i=1,2,\ldots,p)\}$ of the EAR model is a nonlinear optimisation problem, hence is complicated especially with increasing model order. In order to achieve this task, binary genetic algorithms (BGA) hybridized with recursive least squares (RLS) algorithm can be used (Shi & Aoyama, 1997).

Genetic algorithms are search algorithms inspired by the natural selection and natural genetics which can be used to solve optimisation problems. Initially, there is a population of candidate solutions to the optimisation problem and the solutions evolve toward better solutions according to the

Figure 2. Example of SNR result when linear/nonlinear methods were applied to EEG modeling



principles of natural selection (i.e., survival of the fittest) (Goldberg, 1989). For the selection of the fittest chromosome, a fitness function that measures the performance of a chromosome in the population must be defined according to the optimisation problem to be solved.

In our study here, the nonlinear coefficient γ of the EAR model is determined by BGA and once the nonlinear coefficient is obtained, the model will become a linear regression problem in which the linear coefficients, $\{\varphi_i, \pi_i, i=1,2,\ldots,p\}$ will be determined by RLS algorithm. Moreover, the model order is selected as the order with minimum Akaike Information Criterion (AIC) value (Akaike, 1974).

Figure 2 shows an example of signal to noise ratio (SNR) results obtained by applying conventional linear AR modeling and EAR modeling to EEG data from a healthy subject and an epilepsy patient. Note that the SNR values were calculated by reconstructing time series with corresponding AR coefficients (for both AR and EAR modeling techniques) and calculating the SNR between original and reconstructed signals. The figure clearly indicates an improved modeling when EAR was used.

These initial results are promising since it appears that the EAR method can provide an improved characterisation of time series. It is hoped that this method will lead to a better representation of the EEG signals when used in various applications.

FUTURE TRENDS

The preliminary results obtained from EAR model are promising since they indicated an improved modeling of EEG signals recorded from healthy subjects and epilepsy patients. However, further experiments should be conducted to investigate the representative ability of EAR method for the classification of different classes of EEG data (i.e., EEG data from epilepsy patients during seizure and seizure free intervals, EEG data from healthy subjects, mental task EEG data, etc).

The proposed improved EAR method could also be explored for other biological signal analysis applications, such as electrophysiological analysis of cognitive processes, prediction of epilepsy onset, abnormal heart sound and beat detection, heart rate variability monitoring, and so forth.

CONCLUSION

The characterisation (i.e., feature extraction) of EEG signals is one of the most challenging steps towards the design of a real time biological signal analysis system. In order to achieve that task, knowledge of the underlying dynamics of the EEG signals is necessary so that suitable modeling techniques could be utilised for characterisation of the EEG signals. In recent years, nonlinear time series analysis techniques in particular largest Lyapunov exponent, correlation dimension and nonlinear prediction error measures along with surrogate data method were repeatedly applied to some of the biological signals (specifically, EEG) in order to understand the nature of the signals. The results of these studies suggested presence of highly significant nonlinearities in EEG signals, especially the signals recorded from patients with neurological disorders. However the sensitivity and specificity of the utilized nonlinear measures are still subject to current research. It is believed that the development of new time series analysis techniques that will sufficiently represent the nonlinear, chaotic and multidimensional behavior of the EEG signals will improve the understanding of the brain dynamics.

REFERENES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Akay, M. (2000). *Nonlinear biomedical signal processing, fuzzy logic, neural networks, and new algorithms*. Wiley-IEEE Press.

Babloyantz, A., Salazar, J., & Nicolis, C. (1985). Evidence of chaotic dynamics of brain activity during the sleep cycle. *Physics Letters A, 111*, 152-156.

Bukkapatnam, S. (2007). *Nonlinear EEG analysis for mental condition monitoring*. Retrieved June 18, 2008, from http://www.okstate.edu/ceat/iem/iepeople/ bukkapatnam/ Papers/31journalpublic.pdf

Chippa, S. & Bengio, S. (2003). *Nonlinear analysis of cognitive and motor-related EEG signals*. IDIAP Research Report, 03-14, Martigny, Switzerland. Retrieved June 18, 2008, from http://citeseer.ist.psu.edu/chiappa03nonlinear.html

Gautama, T., Van Hulle, M. M., & Mandic, D. P. (2003). Indications of nonlinear structures in brain electrical activity. *Physical Review, E*, 67, 046204.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning.* Addison-Wesley.

Haggan, V. & Ozaki, T. (1981). Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model. *Biometrika, 16(*1), 189-196.

Kantz, H. & Schreiber, T. (2004). *Nonlinear time series analysis* (2nd ed.) Cambridge: University Press.

Keirn, Z. A. & Aunon, J. I. (1990). A new mode of communication between man and his surroundings. *IEEE Transactions on Biomedical Engineering, 37*(12), 1209-1214.

Lehnertz, K., Mormann, F., Kreuz, T., Anderzak, R. G., Rieke, C., David, P., & Elger, C. E. (2003). Seizure prediction by nonlinear EEG analysis. *IEEE Engineering in Medicine and Biology Magazine, 22*(1), 57-63.

Olbrich, E. & Achermann, P. (2005). Analysis of oscillatory patterns in human sleep EEG using a novel detection algorithm. *J. Sleep Res, 14*, 337-346.

Palaniappan, R. (2005). Improved automated classification of alcoholics and non-alcoholics. *International Journal of Information Technology, 2*(3), 182-186.

Sanei, S. & Chamber, J. A. (2007). *EEG Signal Processing.* Wiley.

Shi, Z. & Aoyama, H. (1997). Estimation of exponential autoregressive time series model by using genetic algorithm. *Journal of Sound and Vibration, 205*(3), 309-321.

Stepien, R. A. (2002). Testing for non-linearity in EEG signal of healthy subjects. *Acta Neurobiol. Exp., 62*, 277-281.

Subasi, A., Kiymik, M. K., Alkan, A., & Koklukaya, E. (2005). Neural network classification of EEG signals by using AR with MLE preprocessing for epileptic seizure detection. *Mathematical and Computational Applications, 10*(1), 57-70.

Theiler, J., Eubak, S., Longtin, A., Galdrikian, B., & Farmer, J. D. (1992). Testing for nonlinearity in time series: The method of surrogate data. *Physica D, 58*, 77-94.

Vaughan, T. M. & Wolpaw, J. R. (2006). Guest editorial: Third international meeting on brain-computer interface technology. *IEEE Trans. Neural Systems and Rehabilitation Engineering, 14*(2), 126-127.

KEY TERMS

AR: Autoregressive model, a linear prediction model where each data point in time series is defined to be linearly related to its previous data points.

EAR: Exponential Autoregressive model, a nonlinear extension of Autoregressive model.

Linear Regression: A technique that attempts to model a set of data points by fitting a linear equation to the data.

Linear System: A system f(·) that obeys the superposition and scaling property is said to be linear such that; for $a, b \in R : f(a \cdot x + b \cdot y) = a \cdot f(x) + b \cdot f(y)$.

Linear Signal: A linear signal is generally defined as the output of a linear shift invariant system that is driven by Gaussian white noise.

Nonlinear Signal: A nonlinear signal is generally defined as the signal generated by the system that does not obey superposition and scaling properties.

Power Spectral Density: Power spectral density shows the power per unit frequency of a signal.

Shift-Invariant System: A shift-invariant system is known as a system that input-output relationship does not vary with time such that; let $y[n]$ be the response of the system to input $x[n]$, for any delay t, the response of the system to input $x[n-t]$ will be $y[n-t]$.

Stochastic (Random) Process: Opposite of deterministic processes in which the future states of the system can not be predicted precisely. In other words, even if the initial states of

the process are known there are many states that the process can go where some states are more probable than others.

White Noise: A random signal that has equal amount of power at all frequency bands.

ENDNOTES

[a] Sensitivity is a statistical measure of how well a classification test correctly identifies a condition.

[b] Specificity is a statistical measure of how well a classification test correctly identifies the negative cases, or those cases that do not meet the condition under study.

Nonspeech Audio-Based Interfaces

Shigueo Nomura
Kyoto University, Japan

Takayuki Shiose
Kyoto University, Japan

Hiroshi Kawakami
Kyoto University, Japan

Osamu Katai
Kyoto University, Japan

INTRODUCTION

Visual and auditory imagery combination offers a way of presenting and communicating complex events that emulate the richness of daily experience (Kendall, 1991). It is notable that sound events arise from the transfer of energy to a sound object in everyday life. Even in childhood, we learn to take the following attitudes about the sound events:

- Recognize the occurrence of sound events and relate them to physical events.
- Classify and identify heterogeneous sound events through a lifetime of experience.

Important distinctions in the data can be communicated by exploiting simple categorical distinctions of sound events. Taste, smell, heat, and touch are not suitable channels for data presentation because our perception of them is not quantitative. However, the auditory system constitutes a useful channel for data presentation (Yeung, 1980).

Furthermore, sounds play an important role in the study of complex phenomena through the use of auditory data representation according to Buxton (1990) and Kendall (1991). It is known that our ears and brains can extract information from nonspeech audio that cannot be, or is not visually displayed (Buxton, 1990).

BACKGROUND

Nonspeech Audio

According to Wall and Brewster (2006), nonspeech audio can be delivered in a shorter time than synthetic speech. Synthetic speech audio can be laborious and time consuming to listen to and compare many values through speech alone. So, nonspeech audio can be a better means at providing an overview of the data.

Researchers, such as Bronstad, Lewis, and Slatin (2003) have investigated whether the use of nonspeech audio cues can reduce cognitive workload to users performing very complex tasks that they would otherwise find impossible.

Nonspeech audio researchers have investigated sounds more complex than ubiquitous interrupting beeps to provide information about spatial structure to computer users. In this way, the vOICe Learning Edition (Jones, 2004) is an actual example of interface that translates arbitrary video images from an ordinary camera into nonspeech sounds. However, the artificial sounds adopted by the vOICe have no analogs in everyday listening. So, this kind of interfaces has required extensive trials from users before their effective use.

Everyday Listening

According to Buxton (1990) and Gaver (1988), everyday listening is the experience of listening to events rather than sounds. This experience is different than that enjoyed by traditional psychoacoustics. It consists of hearing which things are important to avoid and which might offer possibilities for action. Instead of perceiving attributes, such as frequency, spectral content, amplitude of sounds, everyday listening is concerned with the attributes of events in the world. Examples of events are the speed of an approaching automobile, the force of a slammed door, and the direction of a walking person (Gaver, 1988). Everyday tasks, such as driving and crossing the street, are examples due to everyday listening. Listening to airplanes, water, birds, and footsteps are other examples of everyday listening. Also, Gaver (1993) has investigated several studies of everyday listening based on the ecological perspective.

Surprisingly, everyday listening skills have been virtually ignored in computer-based interfaces. Listening to events is not well understood by traditional audition approaches.

Studies suggest that a comprehensive account of everyday listening has yet to emerge. We have investigated a first study on comprehensive account of everyday listening by using nonspeech sounds as events (Nomura, Utsunomiya, Tsuchinaga, Shiose, Kawakami, Katai, & Yamanaka, 2007a).

We have concentrated on such everyday listening and ecological approach to perception in our previous works (Nomura, Shiose, Kawakami, Katai, & Yamanaka, 2004; Nomura, Yamanaka, Katai, Kawakami, & Shiose, 2004; Shiose, Ito, & Mamada, 2004). One of the works concerned with perceiving an approaching automobile is useful in understanding the scope of an ecological approach to perception (Shiose et al., 2004). The auditory system captures the experience of everyday listening by the idea that a given sound provides information about the interaction of materials at a location in an environment. Also, alterations in loudness caused by alterations on distance from a source may also provide information about time-to-contact in an analogous fashion to alterations in visual texture (Shaw, McGowan, & Turvey, 1991).

Echolocation

Spallanzani discovered that blinded bats could fly, avoid obstacles, land on walls and ceiling, and survive in nature as well as sighted bats. However, this discovery remained unanswered as how bats possess a sixth sense for orientation and navigation (Raghuram & Marimuthu, 2005). Griffin answered the question in 1938 and called this sixth sense "echolocation" (Griffin, 1958).

Some bat-like sonar systems have been developed using echolocation (Barshan & Kuc, 1992; Bitjoka & Takougang, 2007; Waters & Vollrath, 2003).

On the other hand, Daniel Kish (Roberts, 2006), as an example of human echolocation, lost his sight as an infant and taught himself to "see" with sonar by clicking his tongue. He learned to see without sight.

In our recent work (Nomura, Chiba, Honda, Shirakawa, Shiose, Katai, Kawakami, & Yamanaka, 2008), we looked at the possibility to emerge a comprehensive account of human echolocation that enables spatial structure perception of the environment using acoustics.

Virtual 3-D Acoustic Environment

Virtual 3-D acoustic environments created by computers are an emerging technology that may be used to teach blind children in actual acoustic environment (Inman, Loge, & Cram, 2000).

Advantages of using virtual 3-D acoustic training environments are to provide learners with guided and unguided practice controlling audio parameters by software. The parameters can be adjusted to suit the specific needs of a learner's auditory experience (Inman, Loge, & Cram, 2001).

The virtual 3-D acoustic environment is modeled on the binaural human hearing system (Inman, Loge, & Cram, 2001) and described by a complex response function (head-related transfer function - HRTF) (Mφller, Sφrensen, Hammershφi, & Jensen, 1995). HRTFs contain all the information about the sound source's location (its direction and distance from the listener), and can be used to generate binaural cues (interaural time differences - ITDs; interaural intensity differences - IIDs). Consequently, measured and implemented HRTFs can generate virtual 3-D acoustic environments.

In our previous work (Nomura, Utsunomiya, Tsuchinaga, Shiose, Kawakami, Katai, & Yamanaka, 2007b), we have investigated an approach to enhance spatial conceptualization performances of subjects using the virtual 3-D acoustic environment system.

WORK'S PURPOSE

We propose novel nonspeech audio-based user interfaces to provide visually impaired and elderly people with the opportunity to perceive and conceptualize spatial structure through echolocation. We hope that the users have opportunities to freely access the available facilities and friendly interaction with these targets (spatial structures) without depending on such expert systems as pattern recognizer, spoken language converter, or voice synthesizer.

Also, the purpose is not only or simply based on converting inaudible ultrasound echoes into audible sounds like the sonar system modeled by Bitjoka and Takougang (2007).

In this way, we believe that the user may take advantage of the similar skills employed by bats in everyday listening through echolocation. For example, the association of reverberations with empty environment is a kind of such skill. When a room is reverberant, it means that it is spacious, considering all other things equal. The idea is that the users can utilize their experience and familiarization with everyday listening to friendly conceptualizing spatial structure information without hard cross-modal training. The strategy is based on skill transfer process (Shiose, Sawaragi, Nakajima, & Ishihara, 2004) to embody the cues for spatial structure conceptualization process by avoiding heavy computational load, due to the previously mentioned expert systems (like the vOICe).

We suppose that the eventual interface users can embody the skills employed in everyday tasks by navigating and training in a virtual 3-D acoustic environment. Then, we experimentally evaluate performances of subjects on tasks to perceiving spatial structures represented by various aural surfaces in the 3-D acoustic environment.

EXPERIMENTAL APPROACH

Figure 1 presents a schematic overview of our experimental approach toward nonspeech audio-based interfaces to support users on spatial structure (target) perception and conceptualization.

The apparatus of the experimental approach is constituted by an ultrasound system and a 3-D acoustic environment system, described in the following sections.

Ultrasound System

The ultrasound system is represented by a conventional Pioneer robot, as shown in Figure 2. The robot is constituted by a panel with eight ultrasonic (transmitter and receiver) sensors in a line at its front. This set of sensors is responsible for getting spatial structure information, such as distance from the target.

In this work we do not use ultrasounds to convert them into audible sounds, as done in conventional approaches.

Figure 1. Schematic overview of the experimental approach

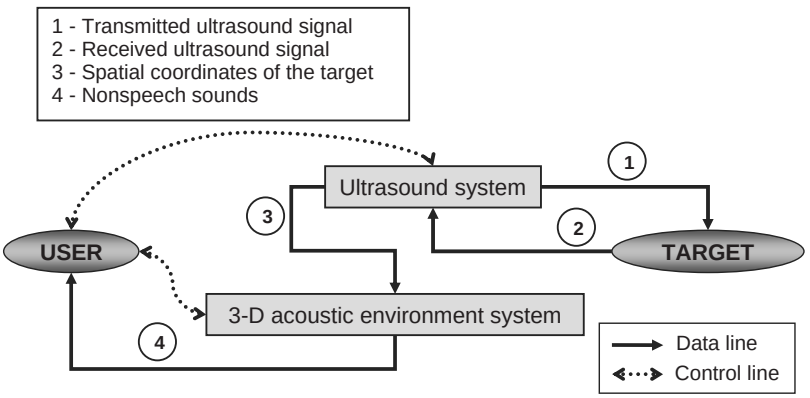
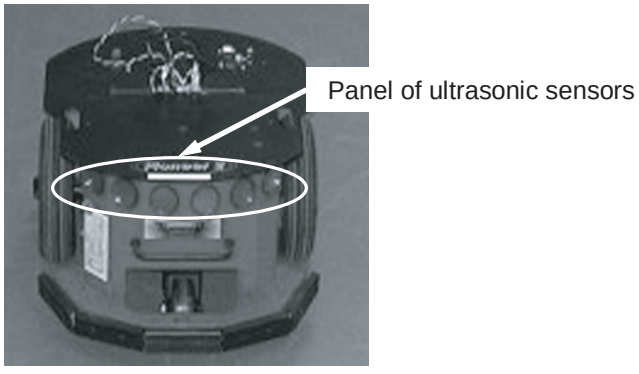


Figure 2. Conventional Pioneer robot with a panel of ultrasonic sensors



We only take advantage of the Pioneer robot to detect and obtain spatial structure coordinates of a target. According to Figure 1, the ultrasound system interacts with the target by transmitting (line 1) and receiving (line 2) ultrasound signals.

3-D Acoustic Environment System

The second part of the experimental apparatus consists of a 3-D acoustic environment system mainly based on sound space processors to add necessary effects such as movement, reflection, reverberation into sound source. In this way, the system creates different sets of experiments characterized by different virtual acoustic environments with nonspeech sounds.

A front view of devices that compose the 3-D acoustic environment system is presented in Figure 3.

Figure 4 presents a schematic overview of the 3-D acoustic environment system with a list of devices and their respective models. According to this schematic overview, the nonspeech sound generation process occurs as follows:

- The audio recorder (AR) plays a fan noise representing the sound source of nonspeech sounds in everyday world. This sound source was saved using precise microphones (Nomura, et al., 2007).
- The sound space processor (SSP) adds effects such as movement, reverberation, and reflection into this sound source. The movement of each sound is guided by the spatial coordinates as output of the ultrasound system (first part of our experimental approach). According to the previous work (Shiose, Ito, & Mamada, 2004), reverberation and reflection levels were adjusted to -30 dB to generate nonspeech sounds and create the virtual hallways described in Figure 5.
- The sounds with effects are mixed by mixing devices (MXP and MXC).
- Then, the resulting sounds are heard by the subjects wearing headphones (DSH) during the experiments.

The apparatus to create the virtual 3-D acoustic environment is only for training, enabling, and simulating echolocation skill transfer process (Shiose, Sawaragi, Nakajima, & Ishihara, 2004) by subjects. The system does not represent a prototype to be implemented on the eventual interface devices.

Virtual Hallways

The upper view of a typical virtual hallway created by the experimental approach is shown in Figure 5. The virtual hallway is formed by aural surfaces, and each aural surface is constituted by the generated nonspeech sounds with the experimental approach.

The interval (e) corresponds to the part of hallway, measured in meters, where there are no alterations on aural surfaces during the navigation task by subjects.

The value of this interval is given by the following equation:

$$e = v.t, \quad (1)$$

Figure 3. Front view of devices composing the 3-D acoustic environment system (Nomura, Tsuchinaga, Nojima, Shiose, Kawakami, Katai, & Yamanaka, 2007)

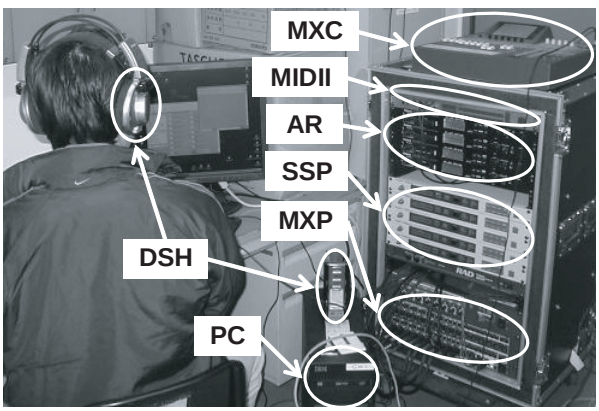


Figure 4. Schematic overview of the apparatus

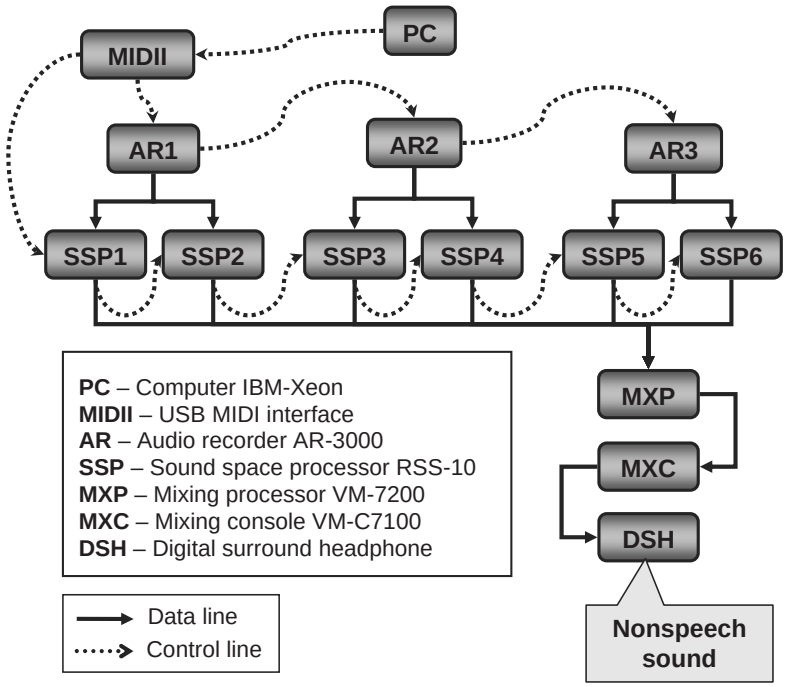


Figure 5. Detailed upper view of a virtual hallway

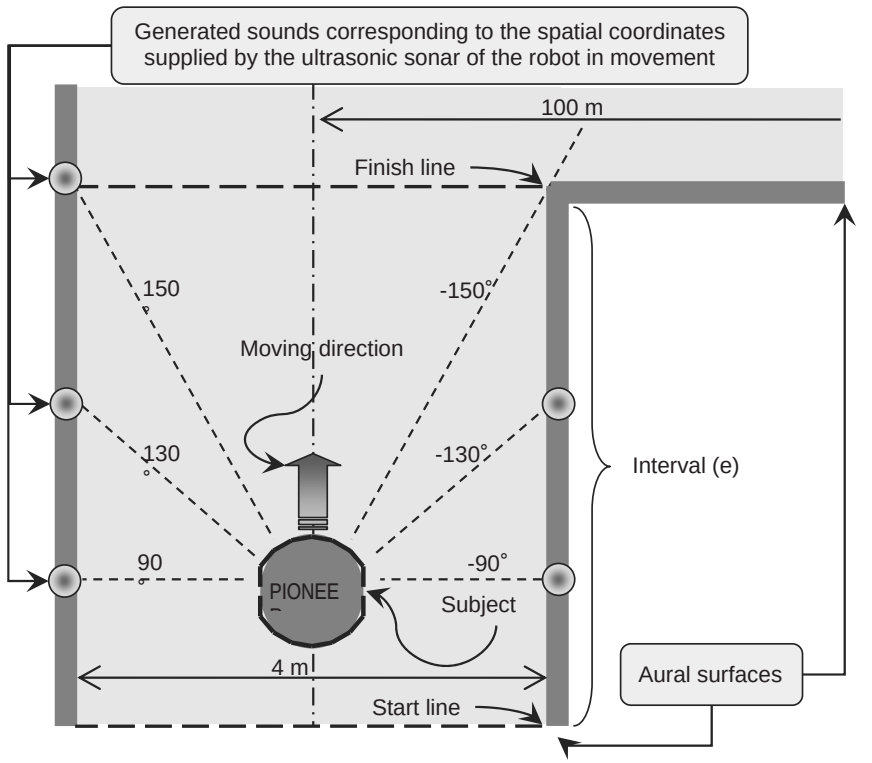


Figure 6. Types of virtual hallways modified by different appearances of aural surfaces

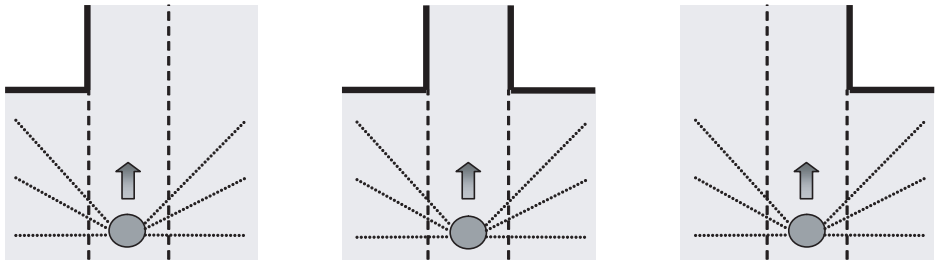
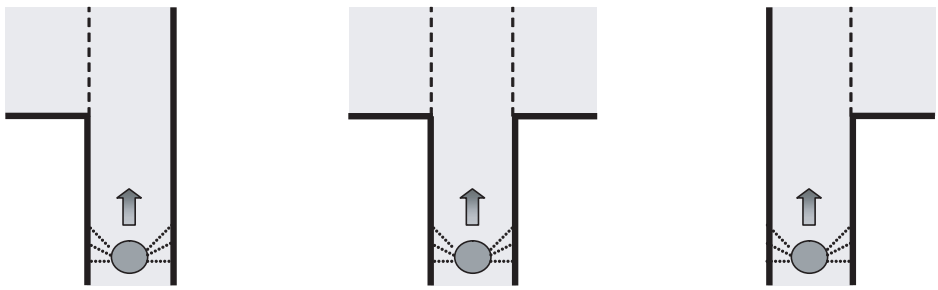


Figure 7. Types of virtual hallways modified by different disappearances of aural surfaces



where v represents the moving speed in km/h and t represents the time in seconds that is randomly set during the experiments.

The virtual hallways created by left (A1), left and right (A2), and right (A3) appearances of aural surfaces are shown in Figure 6. Figure 7 shows the virtual hallways created by left (D1), left and right (D2), and right (D3) disappearances of aural surfaces.

EXPERIMENTS

The experiments aim to evaluate performances of subjects on tasks to perceiving alterations on aural surfaces belonging to a virtual hallway. They used their auditory system skills to hear nonspeech sounds as events. In this way, the subjects' task consisted of navigating virtual hallways created by the experimental approach.

Fourteen subjects participated in the following two sets of experiments:

- Set 1 characterized by appearances of aural surfaces representing the types A1, A2, and A3 of virtual hallways.
- Set 2 forming the types D1, D2, and D3 for virtual hallways modified by disappearances of aural surfaces.

All subjects had no audition problems. No visual information was necessary to participate in the experiments.

The experiments consisted of two sessions as follows:

- A training session provided subjects' familiarization with the three types of appearances for set 1 and the other three types of disappearances for set 2.
- A testing session got results from subjects traveling the different types of virtual hallways and trying to categorize the type of the traveling hallway by anticipation.

In the testing session, we instructed the subjects in pressing a key on a keyboard as follows:

- "←" when they perceive that the left aural surface is going to appear as shown in Figure 6(A1) for experimental set 1, or disappear as shown in Figure 7(D1) for set 2.
- "↓" when they perceive that the left and right aural surfaces are going to appear as shown in Figure 6(A2) for experimental set 1, or disappear as shown in Figure 7(D2) for set 2.
- "→" when they perceive that the right aural surface is going to appear as shown in Figure 6(A3) for experimental set 1, or disappear as shown in Figure 7(D3) for set 2.

Also, in this testing session, an important point is that the subjects should anticipate the type of the traveled virtual hallway. In other words, they should make a decision by the period randomly selected from the set {3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0} in seconds. This period corresponded to the time spent by the subject to move from start to finish lines as indicated in Figure 5. The moving speed was set at 4 km/h in all the experiments.

Then, the experimental results correspond to anticipatory times. To obtain these results, we just have computed the values (negative ones) related to the decisions made by the subjects before crossing the finish line.

Results

In Table 1, we verify that the subjects found serious difficulties in anticipating and perceiving the virtual hallways modified by disappearances of aural surfaces. The averaged performance rate was equal or less than 30%.

However, Table 2 shows that the subjects achieved successful performances on anticipating and perceiving appearances of aural surfaces as virtual hallways. We verify that the averaged performance rate was equal or more than 96%.

FUTURE TRENDS

In this work we have concentrated on investigating a nonspeech audio-based interface that does not require instrumentation of the environment with tags or any additional communication infrastructure in contrast to the conventional user interfaces.

Table 1. Performances of subjects in tasks to perceive disappearances of aural surfaces

	D1	D2	D3
Averaged anticipatory time (s)	-0.58	-1.11	-0.72
Maximum anticipatory time (s)	-1.25	-2.00	-2.20
Averaged performance (%)	21.4	20.0	30.0

Table 2. Performances of subjects in tasks to perceive appearances of aural surfaces

	A1	A2	A3
Averaged anticipatory time (s)	-2.45	-2.29	-2.44
Maximum anticipatory time (s)	-2.75	-2.70	-2.75
Averaged performance (%)	100.0	96.0	100.0

It is relevant that sighted and nonsighted people can get remarkable echolocation abilities to use reflected or reverberated sound patterns to move around and locate their position in rooms and corridors in spite of such irregularities. In other words, users can also locate moving and nonmoving sound-producing objects as well as non-sound-producing obstacles in the environment, and they can often describe the size, shape, and texture of these targets from the pattern of the perceptual consequences.

In the future, we hope to contribute with design of eventual user interfaces that take advantage of a low computational cost provided by the proposed concept. We believe in the realistic practicalities of our proposed concept on future aural user interfaces.

CONCLUSION

Experimentally, we evaluated the subjects' ability to anticipate and perceive alterations on aural surfaces of virtual hallways by generating nonspeech sounds in a 3-D acoustic environment.

Early experimental results showed that more exhaustive experiments could be useful to accurately evaluate how sensitive a subject is to anticipating and perceiving alterations on aural surfaces of virtual hallways.

Surprisingly, the subjects performed well in anticipating and perceiving alterations on aural surface of virtually created hallways without hard cross-modal training. Comparison of experimental results concerning the sets of experiments showed that subjects performed better in tasks conceptualizing virtual hallways modified by appearances of aural surfaces. A possible reason for this difference could be explained by an ecological psychology approach based on direct perception of everyday listening events. We had two event situations in the sets of experiments. First, an event started in a silent situation (without sound) corresponding to the aural surface appearance of experimental set 1. Second, an event stopped in a noisy situation (with initial nonspeech sounds) corresponding to the aural surface disappearance of experimental set 2. Then, we could infer that a starting sound event was easier to be perceived by the subject inserted in first situation. In this situation, there was no initial nonspeech sound to disturb the direct perception of events.

We found that the comprehensive account of everyday listening by taking advantage of subjects' experience to listening to events, rather than sounds, was primordial for achieving these results.

The results were promising to conclude that the exploration of auditory perceptual skills, using nonspeech sounds as events, can be a great alternative to developing novel nonspeech audio-based interfaces. In this way, we believe that upcoming interface users, such as visually impaired or elderly people, have the opportunity to friendly recover the freedom to access the existing public environment. We hope that they can access the facilities without depending on help from others or additional instruments into the infrastructure.

Furthermore, we expect to contribute to the request of impaired persons that do not like to be wired to the complex expert systems of conventional interfaces requiring mental workload and hard training. An immediate contribution can be into the development of applications such as "driving a car in the dark" or "detecting room layouts in the dark" systems.

REFERENCES

Barshan, B., & Kuc, R. (1992). A bat-like sonar system for obstacle localization. *IEEE Transactions on Systems, Man and Cybernetics, 22*, 636-646.

Bitjoka, L. & Takougang, N. A. C. (2007). A sonar system modeled after spatial hearing and echolocating bats for blind mobility aid. *International Journal of Physical Sciences, 2*(4), 104-111.

Bronstad, P. M, Lewis, K., & Slatin, J. (2003). Conveying contextual information using nonspeech audio cues reduces workload. In *Proceedings of Technology and Persons with Disabilities Conference*, USA.

Buxton, W. (1990). Using our ears: An introduction to the use of nonspeech audio cues. In E. Farrell (Ed.), *Extracting meaning from complex data: processing, display, interaction. Proceedings of the SPIE, 1259*, 124–127.

Gaver, W. W. (1988). *Everyday listening and auditory icons*. Doctoral Dissertation, University of California, San Diego.

Gaver, W. W. (1993). What in the world do we hear? An ecological approach to auditory event perception. *Ecological Psychology, 5*, 1–29.

Griffin, D. R. (1958). *Listening in the dark: The acoustic orientation of bats and men.* New Haven, CT: Yale University Press.

Inman D. P., Loge K., & Cram A. (2000). Teaching orientation and mobility skills to blind children using computer generated 3-D sound environments. *Proceedings of the International Community for Auditory Display, Atlanta, Georgia*. Retrieved from http://www.icad.org

Inman D. P., Loge K., & Cram A. (2001). Acoustic virtual training for the blind. *Soundscape: The Journal of Acoustic Ecology, 2*(1), 20–22.

Jones, W. D. (2004). Sight for sore ears. *IEEE Spectrum, 41*(2), 13-14.

Kendall, G. (1991). Visualization by ear: Auditory imagery for scientific visualization and virtual reality. *Computer Music Journal, 15*(4), 70–73.

Mφller, H., Sφrensen, M. F., Hammershφi, D., & Jensen, C. B. (1995). Head-related transfer functions of human subjects. *Journal of the Audio Engineering, 43*(5), 300–321.

Nomura, S., Chiba, G., Honda, A., Shirakawa, T., Shiose, T., Katai, O., Kawakami, H., & Yamanaka, K. (2008). Affordable echolocation-based user interfaces in accessing chaotic environments. In *Intelligent User Interfaces for Developing Regions - IUI4DR Proceedings*, Canary Islands, Spain (pp. 17-22).

Nomura, S., Shiose, T., Kawakami, H., Katai, O., & Yamanaka, K. (2004). A novel "sound visualization" process in virtual 3-D space: The human auditory perception analysis by ecological psychology approach. In *Proc. of 8th Asia Pacific Symposium on Intelligent and Evolutionary Systems*, Cairns, Australia (pp. 137–149).

Nomura, S., Tsuchinaga, M., Nojima, Y., Shiose, T., Kawakami, H., Katai, O., & Yamanaka, K. (2007). Novel nonspeech tones for conceptualizing spatial information. *Artificial Life and Robotics, 11*, 13-17.

Nomura, S., Utsunomiya, T., Tsuchinaga, M., Shiose, T., Kawakami, H., Katai, O., & Yamanaka, K. (2007a). Toward novel interfaces using nonspeech sounds as events for human perception. In *Proc. of the Second International Workshop on Image Media Quality and its Applications*, Chiba, Japan (pp. 189-194).

Nomura, S., Utsunomiya, T., Tsuchinaga, M., Shiose, T., Kawakami, H., Katai, O., & Yamanaka, K. (2007b). Designing an aural user interface for enhancing spatial conceptualization. In *Proc. of the Second IASTED International Conference on Human-Computer Interaction*, Chamonix, France (pp. 205-210).

Nomura, S., Yamanaka, K., Katai, O., Kawakami, H., & Shiose, T. (2004). Towards a novel "sound visualization" via virtual 3-D acoustic environmental media. In *Proc. of International Workshop on Intelligent Media Technology for Communicative Intelligence*, Warsaw, Poland (pp. 121–124).

Raghuram, H., & Marimuthu, G. (2005). Donald Redfield Griffin : The discovery of echolocation. *Resonance*, 20-32.

Roberts, J. (2006). *A sense of the world: How a blind man became history's greatest traveler.* HarperCollins Publishers.

Shaw, B. K., McGowan, R. S., & Turvey, M. T. (1991). An acoustic variable specifying time-to-contact. *Ecological Psychology, 3*, 253–261.

Shiose, T., Ito, K., & Mamada, K. (2004). The development of virtual 3-D acoustic environment for training perception of crossability. In *Proc. of the 9th International Conference on Computers Helping People with Special Needs* (Vol 3118, pp. 476-483.). Paris, France.

Shiose, T., Sawaragi, T., Nakajima, A., & Ishihara, H. (2004). Design of interactive skill-transfer agent from a viewpoint of ecological psychology. *International Journal of Human-Computer Interaction, 17*(1), 69–86.

Wall, S. A., & Brewster, S. A. (2006). Tac-tiles: Multimodal pie charts for visually impaired users. In *4th Nordic Conference on Human-Computer Interaction, ACM International Conference Proceeding Series, 189*, 9–18.

Waters, D. A., & Vollratch, C. (2003). Echolocation performance and call structure in the megachiropteran fruit-bat Rousettus aegyptiacus. *Acta chiropterologica, 5*(2), 209-219.

Yeung, E. S. (1980). Pattern recognition by audio representation of multivariate analytical data. *Analytical Chemistry, 52*, 1120–1123.

KEY TERMS

Aural Surface: It consists of a kind of virtual wall constituted by the generated nonspeech sounds in the 3-D acoustic environment.

Conceptualization: It is the result of spatial structure categorization after processing the visual, aural, or echo information captured in nonspeech audio cues.

Echolocation: It is based on the principle that listeners can process the returning echo information when they emit sound waves and listen to the echoes that return from a target.

Everyday Listening: It is the experience of listening to events rather than sounds. Listening to airplanes, water, birds, and footsteps are some examples of everyday listening. Everyday tasks such as driving and crossing the street are examples due to everyday listening too.

Nonspeech Audio: It can be delivered in a shorter time than synthetic speech. It is a better means at providing an overview of the data. Since nonspeech audio can be heard from 360 degrees, auditory information is better captured than visual information.

Spatial Structure Perception: It refers to perception of size, shape, and texture of targets (objects) by processing the returned echoes through echolocation.

3-D Acoustic Environment: It provides learners with guided and unguided practice controlling audio parameters by software. These parameters can be adjusted to suit the specific needs of a learner's auditory experience during a computer simulation.

Virtual Hallway: It is a kind of virtual corridor in the 3-D acoustic environment where a listener travels to anticipate and perceive alterations on aural surfaces during the navigation task. There are three types of virtual hallways characterized by appearances of aural surfaces and another three types corresponding to disappearances.

Object Classification Using CaRBS

Malcolm J. Beynon
Cardiff Business School, UK

INTRODUCTION

The notion of uncertain reasoning has grown relative to the power and intelligence of computers. From sources which are uncertain information and/or imprecise data, it is importantly the ability to represent uncertainty and reason about it (Shafer & Pearl, 1990). A very general problem of uncertain reasoning is how to combine information from independent and partially reliable sources (Haenni & Hartmann,forthcoming). With data mining, understanding the confirming and/or conflicting information from characteristics describing objects classified to given hypotheses is affected by their reliability. Further, the presence of missing values compounds the problem, since the reasons for their presence may be external to the incumbent reliability issues (Olinsky, Chen, & Harlow, 2003; West, 2001).

These issues are demonstrated here using the classification technique: Classification and Ranking Belief Simplex (CaRBS), introduced in Beynon and Buchanan (2004) and Beynon (2005). CaRBS operates within the domain of uncertain reasoning, namely in its accommodation of ignorance, due to its mathematical structure based on the Dempster-Shafer theory of evidence (DST) (Srivastava & Mock, 2002). The ignorance here encapsulates incompleteness of the data set (presence of missing values), as well as uncertainty in the evidential support of characteristics to the final classification of the objects.

This chapter demonstrates that a technique such as CaRBS, through uncertain reasoning, is able to uniquely manage the presence of missing values by considering them as a manifestation of ignorance, as well as allowing the possible unreliability of characteristics to be inherent. Importantly, the described process removes the need to falsely transform the data set in any way, such as through imputation (Huisman, 2000).

The example issue of credit ratings considered here has become increasingly influential since its introduction in around 1900 with the Manual of Industrial and Miscellaneous Securities (Levich, Majnoni, & Reinhart, 2002). The rating agencies shroud their operations in particular secrecy, stating that statistical models cannot be used to replicate their ratings (Singleton & Surkan, 1991), hence advocating the need for alternative analyses, including those utilising uncertain reasoning.

BACKGROUND

DST is a methodology for evidential reasoning, manipulating uncertainty, and capable of representing partial knowledge (Kulasekere, Premaratne, Dewasurendra, Shyu, & Bauer, 2004; Scotney & McClean, 2003). Early after its introduction it was considered as a generalisation of Bayesian theory.

The traditional terminology within DST begins with a finite set of hypotheses Θ (frame of discernment). A *mass value* (basic probability assignment) is a function $m: 2^{\Theta} \rightarrow [0, 1]$ such that: $m(\varnothing) = 0$ and $\sum_{A \in 2^{\Theta}} m(A) = 1$ (2^{Θ} the power set of Θ). Any $A \in 2^{\Theta}$, for which $m(A) > 0$ is called a *focal element* and represents the exact belief in the proposition depicted by A. From one source of evidence, a set of focal elements (and mass values) is defined as a body of evidence (BOE).

To collate two or more sources of evidence, DST provides a method to combine them, using Dempster's rule of combination. If $m_1(\cdot)$ and $m_2(\cdot)$ are independent BOEs, then the function $m_1 \oplus m_2: 2^{\Theta} \rightarrow [0, 1]$, defined by:

$$[m_1 \oplus m_2](y) = \begin{cases} 0 & y = \varnothing \\ (1-k)^{-1} \sum_{A \cap B = y} m_1(A) m_2(B) & y = \varnothing \end{cases}$$

where $k = \sum_{A \cap B = \varnothing} m_1(A) m_2(B)$, is a mass value associated with $y \subseteq \Theta$. The term $(1 - k)$ can be interpreted as a measure of conflict between the sources (Murphy, 2000) and is made up of one minus the sum of the products of mass values from the two pieces of evidence with empty intersection (often the k is also called the level of conflict and not $1 - k$). The associated problem with conflict is the larger the value of k the more conflict in the sources of evidence, and subsequently the less sense there is in their combination (Murphy, 2000).

To demonstrate DST, the example of the murder of Mr. Jones is considered, where the murderer was one of three assassins, Peter, Paul, and Mary, so Θ = {Peter, Paul, Mary}. There are two witnesses. Witness 1, is 80% sure that it was a man, the concomitant BOE, defined $m_1(\cdot)$, includes m_1({Peter, Paul}) = 0.800. Since we know nothing about the remaining mass value it is allocated to Θ, m_1({Peter, Paul, Mary}) = 0.200. Witness 2, is 60% confident that Peter was leaving on a jet plane when the murder occurred, a BOE defined

Table 1. Intermediate combination of BOEs $m_1(\cdot)$ and $m_2(\cdot)$

$m_2(\cdot) \setminus m_1(\cdot)$	$m_1(\{\text{Peter, Paul}\}) = 0.800$	$m_1(\{\text{Peter, Paul, Mary}\}) = 0.200$
$m_2(\{\text{Paul, Mary}\}) = 0.600$	{Paul}, 0.480	{Paul, Mary}, 0.120
$m_2(\{\text{Peter, Paul, Mary}\}) = 0.400$	{Peter, Paul}, 0.320	{Peter, Paul, Mary}, 0.080

$m_2(\cdot)$, includes $m_2(\{\text{Paul, Mary}\}) = 0.600$ and $m_2(\{\text{Peter, Paul, Mary}\}) = 0.400$.

Defining the combination of these sources of evidence the BOE $m_3(\cdot)$, using Dempster's combination rule, the intermediate set intersections of focal elements of the two BOEs and multiplication of the respective mass values are given in Table 1.

In Table 1, the noticeable result is that no intersections of focal elements produce the empty set, so $k = \sum_{A \cap B = \varnothing} m_1(A) m_2(B) = 0$. It follows, with the measure of conflict $(1 - k) = (1 - 0) = 1$, then the values in Table 1 identify the combined BOE $m_3(\cdot)$ has the form, $m_3(\{\text{Paul}\}) = 0.480$, $m_3(\{\text{Peter, Paul}\}) = 0.320$, $m_3(\{\text{Paul, Mary}\}) = 0.120$ and $m_3(\{\text{Peter, Paul, Mary}\}) = 0.080$. This combined evidence has a more spread-out allocation of mass values to varying subsets of {Peter, Paul, Mary}. Further, there is a general reduction in the level of ignorance associated with the combined evidence. Smets (2002) offers a comparison of a variation of this example with how it would be modelled using traditional probability and Transferable Belief Model.

MAIN THRUST

The main thrust of this chapter is the description and application of the CaRBS system for object classification (Beynon, 2005), which operates in the DST environment. It operates on n_O objects (o_1, o_2, …), each described by n_C characteristics (c_1, c_2, …) and classified to a given hypothesis x or its compliment $\neg x$. For the object o_j ($1 \leq j \leq n_O$) and i^{th} characteristic c_i ($1 \leq i \leq n_C$), a *characteristic* BOE, defined $m_{j,i}(\cdot)$, has the mass values, $m_{j,i}(\{x\})$, $m_{j,i}(\{\neg x\})$ and $m_{j,i}(\{x, \neg x\})$. Following Gerig, Welti, Guttman, Colchester, and Szekely (2000), they are given by:

$$m_{j,i}(\{x\}) = \frac{B_i}{1-A_i} cf_i(v), \quad \frac{A_i B_i}{1-A_i} \quad m_{j,i}(\{\neg x\}) = \frac{-B_i}{1-A_i} cf_i(v) + B_i$$

and $m_{j,i}(\{x, \neg x\}) = 1 - m_{j,i}(\{x\}) - m_{j,i}(\{\neg x\})$,

where $cf_i(v) = \frac{1}{1+e^{-k_i(v-\theta_i)}}$ is the confidence value associated with a characteristic value supporting evidence on the association of objects to the given hypothesis and its complement, and k_i, θ_i, A_i and B_i are incumbent control variables. Importantly, if either $m_{j,i}(\{x\})$ or $m_{j,i}(\{\neg x\})$ are negative they are set to zero, and the respective $m_{j,i}(\{x, \neg x\})$ then calculated. In Figure 1, a characteristic value v is shown to be first transformed into a confidence value (Figure 1a), then deconstructed into its characteristic BOE (Figure 1b) and finally represented as a single simplex coordinate $p_{i,j,i,v}$ in a simplex plot (Figure 1c).

The group of characteristic BOEs $m_{j,i}(\cdot)$ $i = 1, \ldots, n_C$ associated with an object o_j and its classification to x and $\neg x$ can be combined using Dempster's combination rule into an *object* BOE, defined $m_j(\cdot)$, from which its final classification can be found.

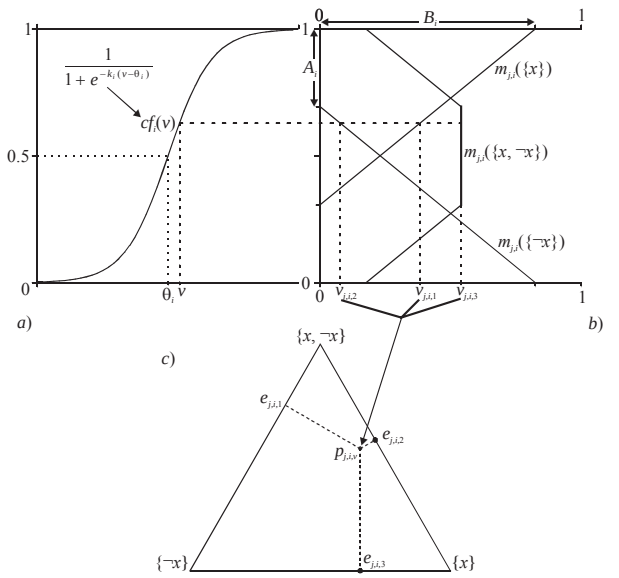
The CaRBS system depends on the assignment of values to the incumbent control variables (with standardised characteristic values, their example domains are; $-1 \leq k_i \leq 2$, $-1 \leq \theta_i \leq 1$, $0 \leq A_i < 1$ and $B_i = 0.4$, see Beynon, 2005). The configuration process then becomes a constrained optimisation problem, solved here using Trigonometric Differential Evolution (TDE) (Fan & Lampinen, 2003), with operation parameters; amplification control $F = 0.99$, crossover constant $CR = 0.85$, trigonometric mutation probability $M_t = 0.05$ and number of parameter vectors $NP = 360$. In summary, TDE develops possible optimum solutions by perturbing previous solutions with the differences between two other previous solutions.

The employed objective function (OB) attempts to minimise the ambiguity in the classification of objects but not the inherent ignorance. For sets of objects making up the equivalence classes, $E(x)$ and $E(\neg x)$, namely those associated with the hypothesis and not the hypothesis, respectively, the optimum solution is to maximise the weighted difference values $(m_j(\{x\}) - m_j(\{\neg x\}))$ and $(m_j(\{\neg x\}) - m_j(\{x\}))$, respectively. The subsequent OB is given by:

$$\text{OB} = \frac{1}{4}\left(\frac{1}{|E(x)|} \sum_{o_j \in E(x)} (1 - m_j(\{x\}) + m_j(\{\neg x\})) + \frac{1}{|E(\neg x)|} \sum_{o_j \in E(\neg x)} (1 + m_j(\{x\}) - m_j(\{\neg x\})) \right)$$

which has domain $0 \leq \text{OB} \leq 1$. Each $(m_j(\{x\}) - m_j(\{\neg x\}))$ and $(m_j(\{\neg x\}) - m_j(\{x\}))$ difference value measures the ambiguity in each classification, there is no attempt to minimise the $m_{j,i}(\{x, \neg x\})$ values so no inclination to directly minimise the concomitant ignorance in each objects' classification. Correct classification is graphically defined by which side of the vertical dashed line down from the $\{x, \neg x\}$ vertex, in a simplex plot, an object BOE's simplex coordinate is positioned (classifying to x (right) and $\neg x$ (left)).

Figure 1. Graphical representation of stages in CaRBS for a characteristic value



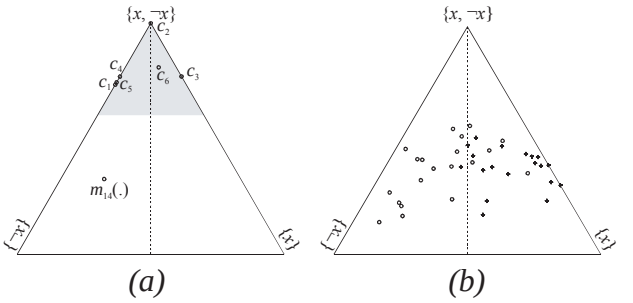
The example results briefly presented concern the Bank Financial Strength Rating (BFSR) of banks in the U.S. (Moody's, 2004). Here, a sample of 40 U.S. banks is considered, described by the six characteristics: (1) Net Income Revenue/Average Assets (c_1); (2) Non Interest Expense/Average Assets (c_2); (3) Equity/Assets (c_3); (4) Net Loans/Assets (c_4); (5) Loan Loss Reserves/Gross Loans (c_5); and (6) Dividend Pay-Out (c_6). The binary classification of the individual banks is based on the rating groups: "B– and above" (x - 20 banks) and "C+ and below" ($\neg x$ - 20 banks).

The first results shown are based on the optimisation of the classification of the 40 banks, using all their characteristic values and known BFSR ratings, see Figure 2.

In Figure 2(a), the classification results of a single bank are shown, namely o_{14}, known to have a BSFR rating "C+ and below." The shaded region denotes the domain that each characteristic BOE can exist in (constrained by $B_i = 0.4$). The simplex coordinates of the characteristic BOEs describing it (labelled c_1, …, c_6), in this case those to the left and right of the vertical dashed line are offering confirming and conflicting evidence, respectively. Its final object BOE, defined $m_{14}(\cdot)$, is to the left of the vertical line again suggesting "C+ and below," the correct classification in this case.

In Figure 2(b), the final classification of all 40 banks is presented, each cross and circle representing banks known to be "B– and above" and "C+ and below," respectively. Pertinent to this investigation, the different heights in the simplex plots that the circles and crosses are found demonstrate the different levels of ignorance associated with the predicted classifications of the banks (their $m_j(\{x, \neg x\})$ mass

Figure 2. Simplex plot representations of results for 40 banks (with no missing values)



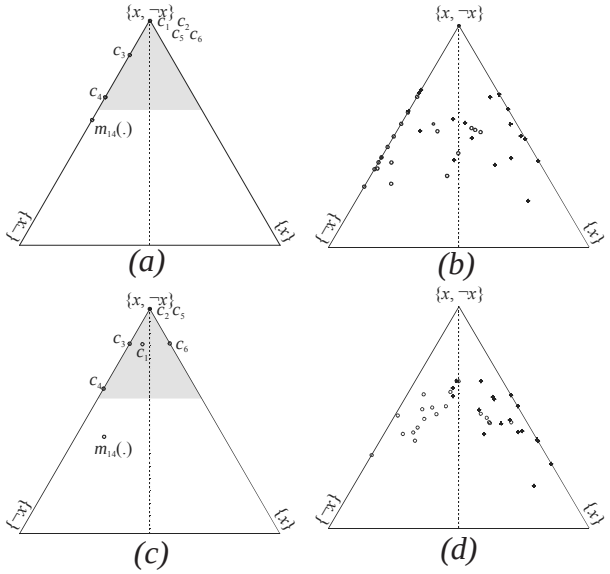
values). Inspection of these results shows an overall 85.0% classification accuracy.

An incomplete data set with missing values causes a problem in data mining due to the uncertainty in what to do with their presence, since most data analysis techniques were not designed for their presence (Schafer & Graham, 2002). Moreover, any imputation-based management of their presence alters the data set considered and mitigates the interpretive power of the concomitant results. In the case of the CaRBS system, there is no requirement to manage the missing values, instead they are retained and considered ignorant values, a direct consequence of its basis on uncertain reasoning.

To illustrate, 50% of the characteristic values in the bank data set are removed and defined as missing. With data mining, this would be a critical situation (Shen & Lai, 2001), since any traditional management dramatically alters the data set considered. All those missing values offer ignorant evidence for which their associated characteristic BOEs are made up of; $m_{j,i}(\{x\}) = 0$, $m_{j,i}(\{\neg x\}) = 0$ and $m_{j,i}(\{x, \neg x\}) = 1$. As with the complete data set, a CaRBS system was configured on this incomplete data, see the top row of simplex plots in Figure 3.

The results in Figure 3(b) show many of the object BOEs are further up the simplex plot indicating more ignorance associated with their final BFSR classifications, a consequence of the number of missing values present which now only confer ignorance evidence. Many of the simplex coordinates are towards the edges of the simplex plots, due to the reduced number of character values that may offer conflicting evidence to a banks BFSR classification (overall 77.5% classification accuracy). In Figure 3(a), for the bank o_{14}, it is noticeable that the characteristic BOEs of c_1, c_2, c_5 and c_6 are at the $\{x, \neg x\}$ vertex, due to them being missing in this case.

Figure 3. Simplex plot representations of results for banks (with missing values retained [top row] and managed using mean imputation [bottom row])



By way of comparison, the missing values in the incomplete data set are replaced through mean imputation (Huisman, 2000): see the bottom row of simplex plots in Figure 3. The results in Figure 3(c) show the inhibiting information of the c_1 and c_6 imputed characteristic values (previously missing), which illustrates the effect of transforming an incomplete data set. That is, since they now are considered to have numerical values (not missing now instead the mean of the characteristic values still present), the optimisation process to assign values to the control variables means they now contribute some evidence rather than just ignorance as previously mentioned.

The results in Figure 3(d) indicate 77.5% classification accuracy, with less simplex coordinates of the object BOEs at the edges of the simplex plot, a consequence of more conflicting evidence from the characteristics, since all characteristic values are not missing. The classification accuracies when there are missing values present are the same, and only slightly worse than when there were no missing values. Importantly when the missing values are retained the rich analysis is on the original data and allows further interpretability unlike when the missing values are imputed.

FUTURE TRENDS

The development of data mining techniques is for the improvement of understanding aspects of everyday life. General uncertain reasoning-based methodologies, such as DST, fuzzy set theory, and rough set theory, offer novel opportunities for evolving the nature of intelligent modelling, bringing different advantages to the subsequent analyses. In the case of DST, the concomitant Transferable Belief Model is taking the theoretical issues forward (Smets, 2005), with particular techniques such as CaRBS utilising the DST environment practically when undertaking the more applied data mining. Perhaps the future lies in the collaboration of the general methodologies (like those previously mentioned), which could augment the advantages, hence realism that subsequent techniques can utilise.

CONCLUSION

The nascence of the CaRBS system described and employed in this chapter illustrates how the development of data mining techniques is an ongoing process. With its underlying mathematical structure being based around the DST of evidence, the utilisation of uncertain reasoning here allows the classification of objects that is allowing for the presence of ignorance in the results.

The advantageous traits associated with the CaRBS system also highlight the incumbencies of effective data mining. Firstly the interpretability of the results, be it with a consistent visual domain (such as the simplex plot), and secondly the lack of a need to externally manage the presence of missing values. This second point is particularly pertinent since perhaps there exists a too relaxed attitude toward their management without real concerns for the effects to real interpretation that can be afforded.

REFERENCES

Beynon, M. J. (2005). A novel technique of object ranking and classification under ignorance: An application to the corporate failure risk problem. *European Journal of Operational Research, 167,* 493-517.

Beynon, M. J., & Buchanan, K. L. (2004). A novel approach to gender classification under ignorance: The case of the European barn swallow (Hirundo Rustica). *Expert Systems With Applications, 27*(3), 403-415.

Fan, H.-Y., & Lampinen, J. (2003). A trigonometric mutation operation to differential evolution. *Journal of Global Optimization, 27,* 105-129.

Gerig, G., Welti, D., Guttman, C. R. G., Colchester, A. C. F., & Szekely, G. (2000). Exploring the discrimination power of the time domain for segmentation and characterisation of active lesions in serial MR data. *Medical Image Analysis, 4,* 31-42.

Haenni, R., & Hartmann, S. (forthcoming). Modelling partially reliable information sources: A general approach based on Dempster-Shafer theory. *Information Fusion.*

Huisman, M. (2000). Imputation of missing item responses: Some simple techniques. *Quality & Quantity, 34,* 331-351.

Kulasekere, E. C., Premaratne, K., Dewasurendra, D. A., Shyu, M.-L., & Bauer, P. H. (2004). Conditioning and updating evidence. *International Journal of Approximate Reasoning, 36,* 75-108.

Levich R. M., Majnoni, G., & Reinhart, C. (2002). *Ratings, rating agencies and the global financial system.* Boston: Kluwer.

Moody's. (2004). Rating definitions—Bank financial strength ratings. Retrieved March 14, 2005, from www.moodys.com

Murphy, C. K. (2000). Combining belief functions when evidence conflicts. *Decision Support Systems, 29,* 1-9.

Olinsky, A., Chen, S., & Harlow, L. (2003). The comparative efficacy of imputation methods for missing data in structural equation modelling. *European Journal of Operational Research, 151,* 53-79.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Scotney, B., & McClean, S. (2003). Database aggregation of imprecise and uncertain evidence. *Information Sciences, 155,* 245-263.

Shafer, G., & Pearl, J. (1990). *Readings in uncertain reasoning*. San Mateo, CA: Morgan Kaufmann.

Shen, S. M., & Lai, Y. L. (2001). Handling incomplete quality-of-life data. *Social Indicators Research*, *55,* 121-166.

Singleton, J. C., & Surkan, J. S. (1991). Modeling the judgement of bond rating agencies: Artificial intelligence applied to finance. *Journal of the Midwest Finance Association, 20,* 72-80.

Smets, P. (2002). Decision making in a context where uncertainty is represented by belief functions. In R. P. Srivastava & T. J. Mock (Eds.), *Belief functions in business decisions* (pp. 17-61). Heidelberg, Germany: Springer-Verlag.

Smets, P. (2005). Belief functions on real numbers. *International Journal of Approximate Reasoning, 40*(3), 181-223.

Srivastava, R. P. & Mock, T. J. (2002). *Belief functions in business decisions*. Heidelberg, Germany: Springer-Verlag.

West, S. G. (2001). New approaches to missing data in psychological research: Introduction to the special section. *Psychological Methods, 6*(4), 315-316.

KEY TERMS

Confidence Value: A function to transform a value into a standard domain, such as between 0 and 1.

Equivalence Class: A set of objects considered the same subject to an equivalence relation (e.g., those objects classified to *x*).

Evolutionary Algorithm: An algorithm that incorporates aspects of natural selection or survival of the fittest.

Focal Element: A finite nonempty set of hypotheses.

Imputation: Replacement of a missing value by a surrogate.

Mass Values: A positive function of the level of exact belief in the associated proposition (focal element).

Objective Function: A positive function of the difference between predictions and data estimates that are chosen so as to optimise the function or criterion.

Simplex Plot: Equilateral triangle domain representation of triplets of nonnegative values which sum to one.

Uncertain Reasoning: The attempt to represent uncertainty and reason about it when using uncertain knowledge, imprecise information, and so forth.

Object-Oriented Software Reuse in Business Systems

Daniel Brandon, Jr.
Christian Brothers University, USA

INTRODUCTION

"Reuse [software] engineering is a process where a technology asset is designed and developed following architectural principles, and with the intent of being reused in the future" (Bean, 1999). "If programming has a Holy Grail, widespread code reuse is it with a silver bullet. While IT has made and continues to make laudable progress in our reuse, we never seem to make great strides in this area" (Grinzo, 1998). "The quest for that Holy Grail has taken many developers over many years down unproductive paths" (Bowen, 1997). This article is an overview of software reuse methods, particularly object oriented, that have been found effective in business systems over the years.

BACKGROUND

Traditional software development is characterized by many disturbing but well documented facts, including:

- Most software development projects "fail" (60%) (Williamson, 1999).
- The supply of qualified IT professionals is much less than the demand (www.bls.gov).
- The complexity of software is constantly increasing.
- IT needs "better," "cheaper," "faster" software development methods.

Over the years, IT theorists and practitioners have come up with a number of business and technical methods to address these problems and improve the software development process and results thereof. Most notable in this sequence of techniques are CASE (computer-aided software engineering), JAD (joint application development), prototyping, 4GL (fourth generation languages), and Pair/Xtreme programming. While these methods have often provided some gains, none have provided the improvements necessary to become that "silver bullet." CASE methods have allowed development organizations to build the wrong system even faster, "wrong" in the sense that requirements are not met and/or the resulting system is not maintainable or adaptable. JAD methods tend to waste more of everyone's time in meetings. While prototypes can help better define user requirements, the tendency (or expectation) that the prototype can be easily extended into the real system is very problematic. The use of 4GL languages only speeds up the development of the parts of the system that were easy to make anyway, while unable to address the more difficult and time consuming portions. Pair programming has some merits but stifles creativity and often requires more time and money.

The only true "solution" has been effective software reuse. Reuse of existing proven components can result in the faster development of software with higher quality. Improved quality results from both the use of previous "tried and true" components and the fact that standards (technical and business) can be built into the reusable components (Brandon, 2000). This improved quality results in lower lifecycle maintenance costs, and since two thirds of software product lifecycle costs are in post-delivery maintenance, this cost savings aspect of reusability is the most rewarding (Schach, 2005). There are several types of reusable components that can address both the design and implementation process. These come in different levels of "granularity" and in both object oriented and non-object oriented flavors.

Software reuse received much attention in the 1980s but did not catch on in a big way until the advent of object oriented languages and tools" (Anthes, 2003). In Charles Darwin's theory of species survival, it was the most adaptable species that would survive (not the smartest, strongest, or fastest). In today's fast moving business and technical world, software must be adaptable to survive and be of continuing benefit. Object oriented software offers a very high degree of adaptability. "Object technology promises a way to deliver cost-effective, high quality and flexible systems on time to the customer" (McClure, 1996). "IS shops that institute component-based software development reduce failure, embrace efficiency and augment the bottom line" (Williamson, 1999). "The bottom line is this: while it takes time for reuse to settle into an organization—and for an organization to settle on reuse—you can add increasing value throughout the process" (Barrett & Schmuller, 1999). We say "object technology" not just adopting an object oriented language (such as C++, Java, or PHP), since one can still build poor, non-object oriented, and non-reusable software, even using a fully object oriented language.

TYPES AND APPLICATIONS OF REUSE

Radding (1998) defines several different types of reusable components, which form a type of "granularity scale":

- **GUI Widgets:** Effective, but only provide modest payback.
- **Server-Side Components:** Provide significant payback but require extensive up-front design and an architectural foundation.
- **Infrastructure Components:** Generic services for transactions, messaging, and database … require extensive design and complex programming.
- **High-Level Patterns:** Identify components with high reuse potential.
- **Packaged Applications:** Only guaranteed reuse—may not offer the exact functionality required. This includes COTS (commercial off the shelf software).

An even lower level of granularity is often defined to include simple text files that may be used in a number of code locations such as "read-me" and documentation files, "help" files, Web content, business rules, XML schemas, test cases, and so forth. Among the most important recent developments of object oriented technologies is the emergence of design patterns and frameworks, which are intended to address the reuse of software design and architectures (Xiaoping, 2003). The reuse of "patterns" can have a higher level of effectiveness over just source code reuse. Current pattern level reuse includes such entities as a J2EE Session Façade or the .Net Model-View-Controller pattern.

Reuse has two types. The first is called opportunistic (or accidental) reuse, where developers realize that a component from a previous project could be used in the current project. The second is systematic (or deliberate) reuse, where components are built to be reused (Schach, 2005). Reusing code also has several key implementation areas: application evolution, multiple implementations, standards, and new applications. The reuse of code from prior applications in new applications has received the most attention. However, just as important is the reuse of code (and the technology embedded therein) within the same application.

Application Evolution

Applications must evolve even before they are completely developed, since the environment under which they operate (business, regulatory, social, political, etc.) changes during the time the software is designed and implemented. This is the traditional "requirements creep." Then after the application is successfully deployed, there is a constant need for change.

Multiple Implementations

Another key need for reusability within the same application is for multiple implementations. The most common need for multiple implementations involves customizations, internationalization, and multiple platform support. Organizations whose software must be utilized globally may have a need to present an interface to customers in the native language and socially acceptable look and feel ("localization"). The multiple platform dimension of reuse today involves an architectural choice in languages and delivery platforms.

Corporate Software Development Standards

Corporate software development standards concern both maintaining standards in all parts of an application and maintaining standards across all applications. "For a computer system to have lasting value it must exist compatibly with users and other systems in an ever-changing information technology (IT) world" (Brandon, 2000). As stated by Weinschenk and Yeo, "Interface designers, project managers, developers, and business units need a common set of look-and-feel guidelines to design and develop by" (Weinschenk & Yeo, 1995). In the area of user interface standards alone, Appendix A of Weinschenk's book presents a list of these standards; there are over 300 items (Weinschenk, Jamar, & Yeo, 1997). Many companies today still rely on some type of printed "Standards Manuals."

EFFECTIVE SOFTWARE REUSE

Only about 15% of any information system serves a truly original purpose; the other 85% could be theoretically reused in future information systems. However, reuse rates over 40% are rare (Schach, 2004). "Programmers have been swapping code for as long as software has existed" (Anthes, 2003). Formal implementation of reuse in various forms of software reuse has been a part of IT since the early refinements to 3GLs (Third Generation Languages). COBOL had the "copy book" concept, where common code could be kept in a separate file and used in multiple programs. Most all modern 3GL's have this same capability, even today's Web-based languages like HTML and JavaScript on the client side, and PHP (on the server side). HTML has "server side includes"; JavaScript has ".js" and ".css" files; and PHP has "require" files (".inc"). Often used in conjunction with these "include" files is the procedure capability where some code is compartmentalized to perform a particular task, and that code can be sent arguments and possibly also return arguments. In different 3GLs this might be called "subroutines"

or in modern languages "functions." A function "library" is a separate file of one or more functions and depending on the language may be pre-compiled.

Object oriented methods are concerned with the design and construction of modules of code that can exhibit certain characteristics. The key characteristics are encapsulation, composition, genericity, inheritance (generalization and specialization), and polymorphism. The code modules are typically called "object types" at the design level and "classes" at the implementation level. These classes contain both form (data) and functionality (functions). Encapsulation involves public "access functions" mediate access to the private data of a class to preserve both data security and integrity. Object oriented methods foster reusability in several ways:

- A class can be "composed" of other classes, and this provides one form of code reuse.
- Generic classes (or templates) can be created, which provide generic definitions and functionality that are common to a number of specific classes, and this is often used to implement types of data structures.
- Interfaces can be specified that define, but do not implement, functionality. Using interfaces, one can design common functionality that may be realized (implemented) by a number of classes (Arlow & Neustadt, 2005).

A class (or interface) can also be derived from another class (a more general "super" or "base" class). Derived classes (more specific classes) inherit the form and functionality of their base class and can add additional form or functionality and also modify functionality; this is the polymorphism aspect. This principle of "inheritance" is the cornerstone of reusability in object oriented systems (Oestereich, 1999).

As an example of object oriented application, consider a large GUI application that may have hundreds of windows. Suppose the corporate design is to have each window show the company name on the top left and the date/time on the top right. Later a new boss decides he or she wants the company name on the top right and the date/time on the top left. If the application had been written in a non-object oriented language, then we would have to go back and modify hundreds of windows. If instead we had used an environment that supported inheritance (such as C++, Java, or PHP) and we had derived all our windows from a base window class, then we would only have to change the one base class. If we used only an "include file" from which to get the company name, that would allow us to easily change the company name, but not to change the place where the name appeared in the window (unless we had clairvoyantly foreseen such a possibility, and designed a name location parameter in our include file).

In most organizations, software reusability is still a goal that is very elusive, as said by Bahrami (1999), "a most difficult promise to deliver on." Radding (1998) stated, "Code reuse seems to make sense, but many companies find there is so much work involved, it's not worth the effort. … In reality, large scale software reuse is still more the exception than the rule." Bean (1999) in "Reuse 101" states that the current decreased "hype" surrounding code reuse is likely due to three basic problems:

- Reuse is an easily misunderstood concept.
- Identifying what can be reused is a confusing process.
- Implementing reuse is seldom simple or easy to understand.

Grinzo (1998) also listed several reasons and observations on the problem of reuse, other than for some "difficult to implement but easy to plug-in cases" such as GUI widgets: a "nightmare of limitations and bizarre incompatibilities," performance problems, "thorny psychological issues" involving programmers' personalities, market components that are buggy and difficult to use, fear of entrapment, component size, absurd licensing restrictions, or lack of source code availability.

Schach (2005) lists and describe the impediments to reuse as:

- Too many IS professionals would rather rewrite than reuse.
- Many IS professionals are skeptical on the safety and quality of using components built for other applications.
- Reuse artifacts are not stored and cataloged in a useful manner.
- Reuse can be expensive (building a reuse process, maintaining the reuse library, and verifying the fitness of the artifact for the new application).
- Legal issues may arise with contract software.
- COTS products are seldom available in source code format.

Some organizations try to promote software reusability by simply publishing specifications on class libraries that have been built for other in-house applications or that are available via third parties; some dictate some type of reuse, and other organizations give away some type of "bonus" for reusing the class libraries of others (Bahrami, 1999). But more often than not, these approaches typically do not result in much success. "It's becoming clear to some who work in this field that large-scale reuse of code represents a major undertaking" (Radding, 1998). "An OO/reuse discipline entails more than creating and using class libraries. It requires *formalizing* the practice of reuse" (McClure, 1996).

There are generally two key components to *formalizing an effective software reuse practice* both within an applica-

tion development and for new applications (Brandon, 2002). These components are:

1. Defining a specific "Information Technology Architecture" within which applications would be developed and reuse would apply
2. Defining a very specific object oriented "Reuse Foundation" that would be implemented within the chosen IT architecture

Once the technical issues of an architecture and a reuse foundation are addressed, then management issues need to resolved, namely "procedures, disciplines, and tools for tracking, managing, searching, and distributing software assets" (Anthes, 2003). Procedures and disciplines have to be formulated and enforced as part of the software development management and organizational "culture." Software can be built or procured to handle the searching, tracking, cataloging, and distribution issues.

FUTURE TRENDS

Several future trends will be key to the continuing success of software reuse. The first trend is for companies to adopt the necessary management principles that foster reuse including programming incentives. Programmers must be rewarded for reusing software, and also producing software that can be reused.

The initial design of systems in an object oriented manner will include several reuse considerations. Lee and Tepfenhart (2002) itemize the basic steps in identifying the system object types as:

- Using the things to be modeled
- Using the definitions of objects, categories, and types
- Using object decomposition (composition)
- Using generalization (working up an inheritance hierarchy)
- Using subclasses (working down the inheritance hierarchy)
- Using domain analysis (reusing domain level functionality)
- Reusing an application framework
- Reusing class hierarchies (including template classes)

Another trend is for software to be developed that facilitates "library" functions for reusable components. Vendors and products in this area today include CMEE (Component Manager Enterprise Edition) from Flashline (http://www.ejbean.com), Logidex byLogicLibrary, and Component Manager by Select Business Solutions (www.selectbs.com). International standards like UML (Unified Modeling Language) and RAS (Reusable Asset Specification) (www.rational.com) will make these tools more interoperable. As well as "library" type tools, other technologies and tools will facilitate reuse that is based on standards such as "Web services" and other messaging technologies, which will let software be reused "where it sits."

Still another trend is for companies to devote programming resources specifically to developing reusable components, while other programmers use the components built by this devoted group.

CONCLUSION

"If you want reuse to succeed, you need to invest in the architecture first" (Radding, 1998). "Without an architecture, organizations will not be able to build or even to buy consistently reusable components." In terms of general IT architectures for business systems, there are historically several types as Central Computer, File Services, Two or Three Tier Client Server, and Two or Three Tier Internet- (Browser-) based. Various transaction processing and database vendors have their own "slants" on these basic approaches, which may depend upon how business logic and the database are distributed.

Today companies are mainly interested in the last of these categories. Internet-based applications are becoming the preferred way of delivering software-based services within an organization (intranets), to the worldwide customer base via browsers and "net appliances" (Internet), and between businesses (extranets). Vendor independent and "open" architectures are often preferred, and the "multiple platform"

Figure 1.

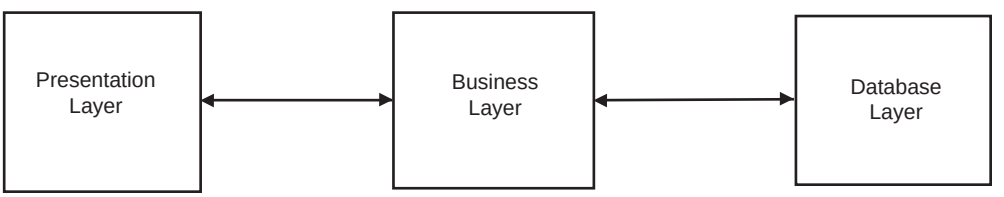


Figure 2.

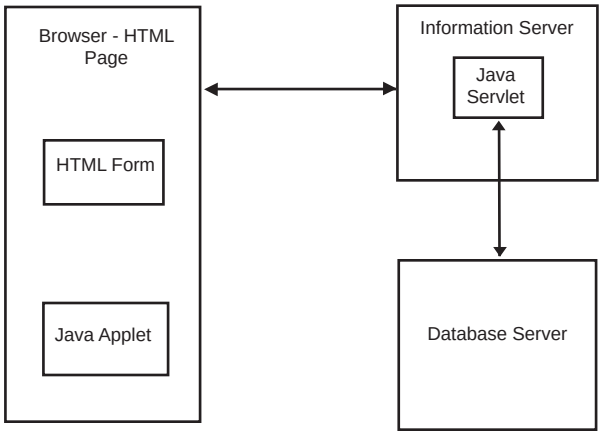


Figure 3.

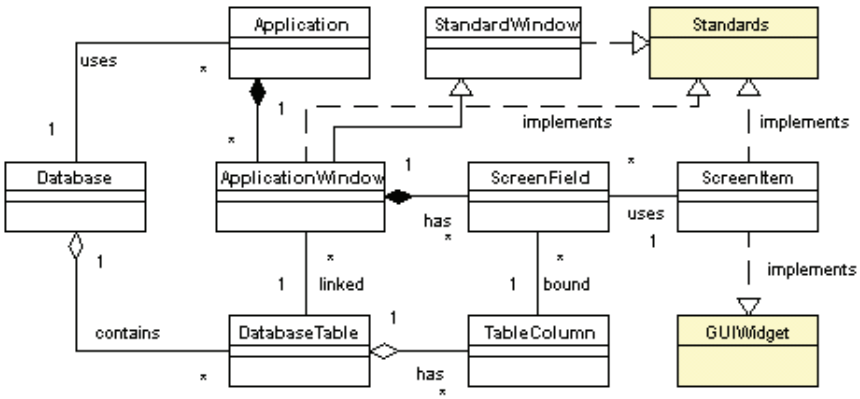
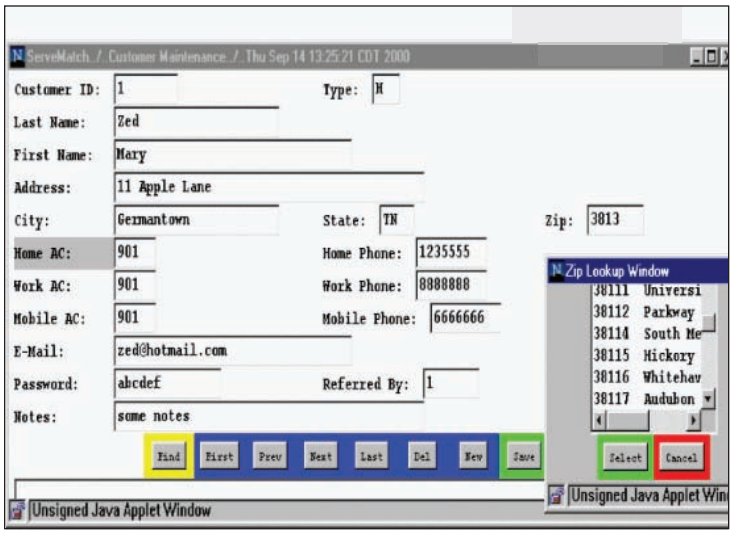


Figure 4.



dimension of reusability is handled by using server generated HTML and JavaScript (via Java or PHP programs on the server).

As has been concluded by several authors, "A reuse effort demands a solid conceptual foundation" (Barrett & Schmuller, 1999). One such foundation was presented by Brandon (2002). It is based on the key object oriented principles of inheritance and composition. By establishing this foundation, an organization can effectively begin to obtain significant reusability since programmers must inherit their class from one of the established classes, and they must only compose their class of the established pre-built components.

REFERENCES

Arlow, J., & Neustadt, I. (2005). *UML 2 and the unified process*. Addison Wesley.

Anthes, G. (2003). Code reuse gets easier. *Computerworld*.

Bahrami, A. (1999). *Object oriented systems development*. Irwin McGraw Hill.

Barrett, K., & Schmuller, J. (1999). Building an infrastructure of real-world reuse. *Component Strategies,* (October).

Bean, J. (1999). Reuse 101. *Enterprise Development,* (October).

Bowen, B. (1997). *Software reuse with Java technology: Finding the Holy Grail*. Retrieved from www.javasoft.com/features/1997/may/reuse.html

Brandon, D. (2000). An object oriented approach to user interface standards. In *Challenges of information technology in the 21st century*. Hershey, PA: Idea Group Publishing.

Brandon, D., Jr. (2002). Achieving Effective Software Reuse for Business Systems. In S. Valenti (Ed.), Successful Software Reengineering (pp. 92-98), Hershey, PA: IRM Press.

Grinzo, L. (1998). The unbearable lightness of being reusable. *Dr. Dobbs Journal,* (September).

Lee, R., & Tepfenhart, W. (2002). *Practical object-oriented development with UML and Java*. Prentice Hall.

McClure, C. (1996). Experiences from the OO playing field. *Extended Intelligence*.

Oestereich, B. (1999). *Developing software with UML*. Addison Wesley.

Radding, A. (1998). Hidden cost of code reuse. *Information Week,* (November 9).

Reifer, D. (1997). *Practical software reuse*. Wiley Computer Publishing.

Schach, S. (2004). *Introduction to object oriented analysis and design*. Irwin McGraw Hill.

Schach, S. (2005). *Object oriented and classical software engineering*. Irwin McGraw Hill.

Weinschenk, S., Jamar, P., & Yeo, S. (1997). *GUI design essentials*. John Wiley & Sons.

Weinschenk, S., & Yeo, S. (1995). *Guidelines for enterprise wide GUI design*. John Wiley & Sons.

Williamson, M. (1999). Software reuse. *CIO Magazine,* (May).

U.S. Department of Labor, Bureau of Labor Statistics. (2003) Retrieved from www.bls.gov

Component manager, Flashline. (2004) Retrieved from www.ejbean.com/products/related/flashline_cm2.html

Reusable asset specification, Rational Rose. (2001) Retrieved from www.rational.com/rda/ras/preview/index.htm

Component manager, Select Business Solutions. (2004) Retrieved from www.selectbs.com

Xiaoping, J. (2003). *Object oriented software development using Java*. Addison Wesley.

KEY TERMS

Class: A program construct representing a type of thing (abstract data type), which includes a definition of both form (information or data) and functionality (methods).

Composition: A new class in an objected programming language that is composed of other classes.

Encapsulation: The ability to insulate data in a class so that both data security and integrity is improved.

Framework: A software foundation that specifies how a software system is to be built. It includes standards at all levels both internal construction and external appearance and behavior.

Function: A programming construct where code that does a particular task is segregated from the main body of a program; the function may be sent arguments and may return arguments to the body of the program.

Implementation: The code placed inside of methods. For some languages this code is pre-compiled or interpreted.

Include: Some code stored separately from the main body of a program so that this code can be used in many programs (or multiple places in the same program).

Inheritance: A feature of object oriented languages that allows a new class to be derived from another class (a more general class); derived classes (more specific classes) inherit the form and functionality of their base class.

Interface: The specification for a method ("what" a method does); how that function is called from another program. Interfaces are provided in source form as opposed to implementations, which are secure. This allows one to use a method without regard for "how" that method is coded. It also allows multiple implementations of the same interface.

Libraries: A group of functions and/or classes stored separately from the main body of the main program; an "include" file consisting of functions and/or classes.

Method: A function defined inside of a class.

Packages: Similar to a library, but just containing classes.

Patterns: A software library for a common business scenario. A framework may be a design framework (possibly expressed in UML) or an implementation framework (possibly in C++, Java, or PHP).

Polymorphism: The ability of object oriented programs to have multiple implementations of the same method name in different classes in an inheritance tree. Derived classes can override the functionality defined in their base class.

Reuse: Reuse (software) is a process where a technology asset (such as a function or class) is designed and developed following specific standards, and with the intent of being used again.

Separation: The separation of what a method does (interface) from how the method does it (implementation).

Observations on Implementing Specializations within an IT Program

Erick D. Slazinski
Purdue University, USA

INTRODUCTION

With a projected 2.26 million additional jobs to fill in various computer fields by the year 2010, there are and will continue to be ample job opportunities in the computer industry. However, the computer field is far too broad for one individual to be an expert in the entire field. Therefore it may be more useful for students to have the opportunity to concentrate their studies in a specific interest area within a broader Information Technology (IT) degree.

IT educators throughout the United States (US) have paid attention to the needs and demands of the IT industry. To address the need for IT graduates with specialized skills, many of the leading universities have created programs which allow undergraduate students to specialize or focus their studies.

This chapter will discuss findings on the state of IT programs with regards to their course offerings. One area of specialization, or track, is presented as an example. It will be noted that even within a specialty area, there can be further specializations. In addition to supporting the students pursuing the specialty area, general knowledge courses must also be offered to those pursuing other specialty areas.

BACKGROUND

The Bureau of Labor Statistics reported 2.9 million computer-related jobs in 2000, with an expected 4.89 million computer jobs by the year 2010. Considering new jobs as well as replacements, over 2.26 million additional people will be needed to fill these jobs (Hecker, 2001). The fluid nature of the IT industry makes generalizations difficult. Therefore, skills are often categorized or grouped together into skill sets (or job descriptions). The most common clustering of skills has been summarized in Table 1.

Of these pathways (or specialization tracks), two of the top occupations (as predicted by the US Dept of Labor) are systems analysis and database administrators (which have been grouped in the EDC Information Support and Service Pathway). See Table 2 for a listing of the top growth occupations.

WHERE ARE THE SPECIALIZED IT PROGRAMS?

Published curriculum from the institutes who attended the Conference for IT Curriculum (CITC) II held in April of 2002 were used as the sample set. The conference attendees were primarily IT educators from around the US, who had an interest in IT curriculum issues. An IT curriculum is focused on the application of technologies to solve problems. To differentiate, a traditional Computer Science curriculum is focused on algorithm design.

Table 3 illustrates, that out of the 28 programs studied, 50% (14) had some specialization available for students. Of the 14 programs that offered specializations, 45% (6) of those offered at least a database specialization similar to our sample track.

NEED FOR A DATABASE TRACK

The same data from the Bureau of Labor Statistics (Hecker, 2001) indicates there were 106,000 jobs for database administrators (DBAs) in 2000, with a projected 176,000 openings to fill by the year 2010. In addition to DBAs, there are also

Table 1. Summary of educational pathways

- Network Systems Pathway
- Information Support and Service Pathway
- Programming and Software Development Pathway
- Interactive Media Pathway

From EDC (2002)

Table 2. Percentage change in employment, projected 1998-2008

Occupation	Percent change
Computer engineers	108
Computer support specialists	102
Systems analysts	94
Database administrators	77

From U.S. Dept. of Labor

database professionals who specialize in database architecture and database programming.

ANATOMY OF A DATABASE TRACK

Though there are various job titles given to database activities in the workplace, such as these listed on the ITWORKS-OHIO website: Data Analyst, Database Administrator, Database Analyst, Database Developer, and Database Specialist—many others exist in the marketplace. However, based on the author's opinion, when the job descriptions are examined, one can find that there are really three, inter-related roles. Figure 1 illustrates these roles using the structure of a house as an example. The DBA (as the foundation) keeps a database available, secure, and healthy. Without the DBA, the rest of the database team could not function. The database developer (as the framing of the house) encodes the business logic in the database. Additionally the database developer often develops the interface layer between the system software and the database engine. Lastly, the database architect (as the roof of the house) is often a senior staff member. Typical

Table 3. Programs and specializations

INSTITUTION NAME	SPECIALAZATION	DATABASE SPECIALIZATION
Ball State	NO	
Bentley	NO	
Brigham-Young University (BYU)	NO	
BYU-Hawaii	NO	
BYU-Idaho	NO	
Capella	YES	NO
Drexel	YES	YES
Florida State University	YES	NO
Georgia Southern	NO	
George Mason University	YES	NO
Hawaii at Manoa	NO	
Houston	NO	
Indiana University	YES	NO
Indiana University Purdue University at Indianapolis	YES	NO
Macon State	YES	YES
New Jersey Institute of Technology	YES	NO
Northern Alabama	YES	NO
Pace Univerisy	NO	
Pennsylvania College of Technology	YES	NO
Purdue University	YES	YES
Purdue University - Calumet	YES	YES
Rochester Institute of Technology	YES	YES
Southern Alabama	YES	YES
State University of New York (SUNY) Morrisville	NO	
Towson University	NO	
University of Baltimore	NO	
University of Cincinnati-Clermont	NO	
USCS	NO	

duties include determining the amount of business logic to be encoded in the database, developing the database design (including distribution of data, data flows, etc.) and often overseeing the implementation of the database. Like the components of a house, all three roles are co-dependant and necessary. In addition to the specific duties, each member must maintain several lines of communication within the development team and organization.

The database administrator must often work with the system/network administrators when it comes to the purchasing of hardware, determining network traffic and bandwidth requirements, coordinating the use of backup devices, etc. They often work with the database architects in providing recommendations for physical implementation issues. For the database developer, they must interact with the database architect (they are implanting the design from the architect) as well as the system developers and ensure that the developers have access to the data they need while maintaining the integrity of the database. The database architect must work with the software architect. Together they determine how to partition the business logic of the system. The database architect works with the DBA in determining the best physical implementation of the supporting database. And lastly the architect often coordinates the activities of the database developers.

Based on these responsibilities and communication channels, the database track was designed to produce students prepared for these multi-faceted jobs. During the first two years of our program, all students gain a broad overview of the Information Technology field, taking introductory courses in programming, Internet technologies, architectures, telecommunications, database, and systems analysis and design

- **Introduction to Application Development:** Introduction to system development using MS Access.
- **Introduction to Computer Programming:** Application development using Visual Basic.
- **Internet Foundations and Technologies:** Web page development using XHTML.
- **Information Technology Architectures:** Explores the history, architecture, and development of the Internet and the World Wide Web.
- **Systems Software and Networking:** Introduction to data communications and Network Operating systems.
- **Programming for the Internet:** Internet application development using scripting languages.
- **Database Fundamentals:** Normalization, SQL, and application interfaces to databases.
- **Systems Analysis and Design Methods:** Introduction to information systems development.

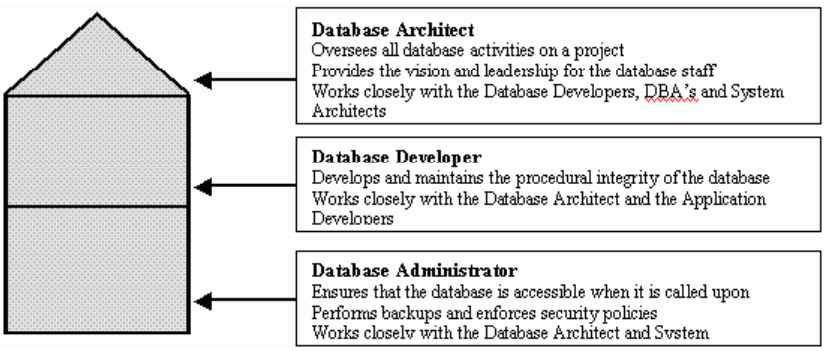
COURSES IN THE DATABASE TRACK

Students specializing in the database area must complete an additional 25 credit hours of technical courses focused on getting the database student prepared for one of the three roles described above. The progression of the courses is shown in Figure 2. Then it will be shown how these courses, in combination with other technical electives, prepare our students for future jobs.

A brief description of these database courses follows.

- **Introduction to Application Development (all students):** Introduces the development of information systems through the use of a database. Topics include business information systems, system and application development, database management systems, problem solving, logic, data types, and programming using database technology. Given a database design and application requirements, students design, construct, and test a personal computer information system.
- **Database Fundamentals (all students):** Looks at relational database concepts, including data design,

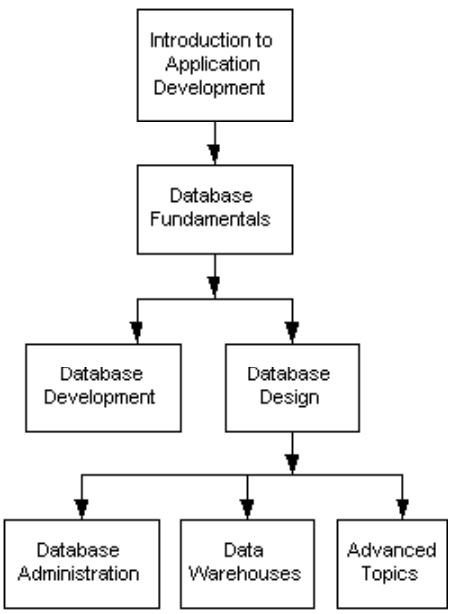
Figure 1. Database roles



modeling and normalization. Students use SQL to query, define, populate, and test a database. Expands on previous courses by accessing databases from programs and the Web, and discusses practical issues that database developers must handle.

- **Database Development:** Explores some of the programmatic extensions to SQL supported by leading Relational Database Management Systems (RDBMS) vendors. Topics include stored procedure and trigger design and implementation, query optimization to enhance performance, and data transformation to enhance interoperability of data.
- **Database Design and Implementation:** Deals with advanced design techniques and physical issues relating to enterprise-wide databases. Topics include advanced normalization, data distribution, distributed database design and replication, storage estimation and allocation, usage analysis, partitioning of very large tables, metadata analysis, data conversion, and load techniques.
- **Database Administration (elective):** Explores tools and techniques for managing an organization's database technology. Topics include database architecture, database technology installation, database creation and maintenance, RDBMS operations and troubleshooting, and database performance tuning.
- **Data Warehousing (elective):** Studies the design and implementation of data warehouses (including data marts and operational data stores) using current database technologies. Topics include data modeling for warehouses, data warehousing infrastructure and tool selection, data exploration, data synthesis and reduction, organizational metadata, and data warehouse administration.
- **Advanced Topics in Database Technology (elective):** Explores contemporary issues in the database arena. These issues may be related to new or breakthrough concepts, technologies, or techniques.

Figure 2. Database courses



Rounding Out the Students

Since database development has a significant analysis and design component, students are required to take a Systems Requirements Discovery and Modeling course, offered by our systems analysis group. Likewise many database personnel are often relegated to a distinct group with a project that has management requirements, so a Project Management course is also required. To support the further specialization into the database architect, database programmer and database administrator roles, two separate selective areas - database management selectives and database technology selectives - were defined. The students must choose a minimum of six credit hours (two courses) in each of the selective areas. The database management selectives are the elective courses listed with the database course flow in Figure 2 (above) – database administration, data warehousing and advanced topics in database management, as well as any graduate level database course offered during the student's senior year. The database technology selectives provide additional areas of exploration on topics that piqued the students' interest in their first two years. Again they are allowed to choose any two courses from the following:

- **Object-Oriented Programming:** use object-oriented programming languages (Java) in the development of modern, business applications.
- **Advanced Design Techniques:** advanced study of system design methods and techniques used by systems analysts to develop information systems.
- **Software Development Methodologies:** methodologies and practices commonly used in contemporary software development projects.
- **Enterprise Application Development:** component development and reuse, distributed object technologies, multi-tier applications.
- **Senior Software Development Project:** integrates the software development technologies and techniques taught in prior courses.
- **E-Commerce:** components of e-commerce.
- **Automatic identification and data capture:** real-time data feeds.
- Any other database management selective course(s) not already taken.

MODEL CURRICULA

At the time this chapter was written, there are at least three curricula models available from which to build from: the IEEE/ACM Computing Curricula; the EDC Academic Foundations and Pathways (EDC, 2002) and the ACM SIG ITE Curriculum Proposal. These models share more similarities than differences and selecting a model may be more a matter of preference of one set of authors over another. Considering the most mature model, the author has chosen to compare the specified database track against the IEEE / ACM Computing Curricula 2001.

The most recent IEEE / ACM Computing Curricula report, available at the ACM Web site, identifies Information Management (IM) as a knowledge area within their Computing Curricula 2001. The IM area of knowledge includes 14 components. Three are considered core, and 11 are viewed as electives. The courses in our database track provide significant coverage of the IM areas identified. Additionally, a Database Administration course that is beyond the model is offered. Table 4 shows the mapping from the IM components to our DB-track courses.

FUTURE TRENDS

For our University, the new curriculum implementing the track concept started in Fall 2001. In Spring 2003 the first cohort of Freshmen (who entered the program with the tracks in place) were required to select their track of interest. However, as news of the program's availability became known, students under the old plan of study transferred to the new plan of study. Additionally, transfer students were admitted under the new plan of study. These students plus an enthusiastic faculty caused the transition to the new course models to occur immediately after the new program (with tracks) was ratified by the University curriculum committee instead of waiting until the Spring 2003 semester.

This caused several registration issues which our counselors had to handle: multiple plans of studies; multiple course equivalencies (as courses were being brought online and retired); and pre-requisite issues. Of these the issue of tracking pre-requisites has been the most challenging because certain courses exist in multiple curriculums whose content has changed every time the curriculum has changed. Often counselors would have to have professors validate that a student had the appropriate pre-requisite knowledge. This often led to questions such as "who taught this pre-requisite course?" or "did you cover this topic?"

CONCLUSION

Since the IT field is too broad to be comprised of only generalists, the IT workforce is becoming increasingly specialized. In response to this, IT educators have started offering areas of specialization or tracks that meet industry's needs. However, designing and implementing a tracked curriculum is a demanding task. The implementation costs may be higher than first estimated. If students are to get

Table 4. Comparison of model curricula to database track courses

IM Knowledge Areas	DB Track
IM1: Information Models and Systems (core)	In 1st year course
IM2: Database Systems (core)	In 1st year course
IM3: Data Modeling (core)	2nd year: Database Fundamentals 3rd year: Database Design
IM4: Relational Databases (elective)	2nd year: Database Fundamentals 3rd year: Database Design
IM5: Database Query Languages (elective)	2nd year: Database Fundamentals (not OO queries)
IM6: Relational Databases Design (elective)	2nd year: Database Fundamentals 3rd year: Database Design
IM7: Transaction Processing (elective)	2nd year: Database Fundamentals 3rd year: Database Development
IM8: Distributed Databases (elective)	3rd year: Database Design 4th year Data Warehousing
IM9: Physical Database Design (elective)	3rd year: Database Design
IM10: Data Mining (elective)	4th year Data Warehousing
IM11: Information Storage and Retrieval (elective)	4th year Data Warehousing (partial)
IM12: Hypertext and Hypermedia (elective)	Candidate for 4th year Adv Topics
IM13 Multimedia Information and Systems (elective)	Candidate for 4th year Adv Topics
IM14: Digital Libraries (elective)	Candidate for 4th year Adv Topics

through a curriculum with multiple tracks in a reasonable time frame, the number of courses that need to be offered every semester can increase. This increase will require more resources could put a drain on other departmental offerings (such as service/graduate courses).

REFERENCES

Bachelor of Science in Information Technology. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://ite.gmu.edu/bsit/degree.htm*

Computer and Information Sciences. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://wwwnew.towson.edu/cosc//*

Computer Information Systems Programming and Database Processing Associate of Applied Science Degree. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.pct.edu/degrprog/pd.shtml*

Department of Computer Information Systems. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://www2.una.edu/business/cis.html*

Department of Information Systems. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.byui.edu/Catalog/2002-2003/_jim2. asp?departmentID=1888#680*

Education Development Center, Inc. (EDC). (2002, August) Information Technology Career Cluster Initiative Academic Foundations and Pathway Standards ValidationStudies. Retrieved November 29, 2004 from: *http://webdev2.edc.org/ewit/materials/ITCCIVSFinal.pdf*

Finkelstein, L., and Hafner, C. (2002). *The Evolving Discipline(s) of IT (and their relation to computer science): A Framework for Discussion*. Retrieved June 30, 2004 from the World Wide Web at: *http://www.cra.org/Activities/it-deans/resources.html*

Free On-Line Dictionary of Computing (FOLDOC). (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://wombat.doc.ic.ac.uk/foldoc/foldoc.cgi*

Georgia Southern University – College of Information Technology. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://cit.georgiasouthern.edu/*

Hecker, D. (2001, November). Occupational employment projections to 2010. *Monthly Labor Review*. Retrieved January 6, 2002 from the World Wide Web at: http://*www.bls.gov/opub/mlr/2001/11/art4full.pdf*

Indiana School of Informatics – Undergraduate Program. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://www.informatics.indiana.edu/academics/undergrad.asp*

Information Systems. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://soc.byuh.edu/is/course.htm*

Information Systems and Operations Management. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://www.bsu.edu/web/catalog/undergraduate/programs/Programs02/isom02_cb.html*

Information Technology – Macon State College. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.maconstate.edu/it/it-bachelors.asp*

IT Concentrations – Information Technology Program. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.it.njit.edu/concentrations.htm*

Joint Computer Society of IEEE and Association for Computing Machinery. (2001, August 1). *Computing Curricula 2001 – Steelman Draft (August 1, 2001)*. Retrieved January 6, 2002 from the World Wide Web at: *http://www.acm.org/sigcse/cc2001/steelman/*

Morrisville State College – CIT. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.cit. morrisville.edu/index.html*

Occupational Area Definitions for Information Systems and Support. (2002). Retrieved July 11, 2002 from the World Wide Web at: *http://www.itworks-ohio.org/ISSdefinit.htm*

Purdue University Computer Technology. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.tech.purdue.edu/cpt/*

Required Courses for BSIS year 2000-02 Entrants. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.cis.drexel.edu/undergrad/bsis/required_2000+_3.asp#dms*

School of Computer and Information Sciences. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.southalabama.edu/bulletin/cis.htm*

School of Computer Science and Information Systems at Pace University. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://csis.pace.edu/csis/index.html*

School of Technology – BS in Information Technology. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://www.capella.edu/aspscripts/schools/technology/bs_general.asp*

SIGITE. (2002, Sept 27). *IT Curriculum Proposal – Four Year Degrees*. Retrieved November 29, 2003 from the World Wide Web at: *http://site.it.rit.edu/*

Specialization in Computer Information Systems. (2002). Retrieved July 13, 2002 from the World Wide Web at: *http://*

business.ubalt.edu/DegreePrograms/ungrad_prog/special.html#CIS

Undergraduate Programs in ICS. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://www.ics.hawaii.edu/academics/degree_programs/undergrad/index.jsp*

University of Cincinnati Undergraduate Programs Information Systems. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://www.uc.edu/programs/viewprog.asp?progid=1492*

U.S. Department of Labor Bureau of Labor Statistics. (2003). *Working in the 21st Century*. Retrieved November 29, 2003 from the World Wide Web at: *http://www.bls.gov/opub/working/home.htm*

USCS Computer Information Systems. (2004). Retrieved June 30, 2004 from the World Wide Web at: *http://www.uscs.edu/academics/cas/mcs/Catalog_info/ba_cis.html*

KEY TERMS

DBA: The title database administrator (DBA) represents an IT professional who ensures the database is accessible when it is called upon, performs maintenance activities, and enforces security policies.

IT Discipline: The ***intellectual gap*** in our educational frameworks for students who are interested in computing careers but find computer science too narrow, mathematical, and physical-science oriented, while MIS is insufficiently deep in technical content and too focused on traditional business topics and culture (Finkelstein, 2002).

Knowledge Area: Represents a particular sub-discipline that is generally recognized as a significant part of the body of knowledge that an undergraduate should know (IEEE, 2001).

Relational Database Management System (RDBMS): A suite of programs which typically manages large structured sets of persistent data, offering ad hoc query facilities to many users, which is based on the relational model developed by E.F. Codd (FOLDOC).

Selective List: A set of courses, grouped together for the purpose of filling an educational skill gap.

Systems Analysis: The design, specification, feasibility, cost, and implementation of a computer system for business (FOLDOC).

Systems Design: The approach used to specify how to create a computer system for business (FOLDOC).

Track: A series of courses designed around a topical area which is structured in a manner to efficiently develop a student's skill set.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp.2161-2167,

Offshore Software Development Outsourcing

Stephen Hawk
University of Wisconsin - Parkside, USA

Kate Kaiser
Marquette University, USA

OFFSHORE SOFTWARE DEVELOPMENT OUTSOURCING EVOLUTION

Until the global economic downturn of the new millennium, demand for information technology (IT) professionals exceeded supply mostly due to specific skill sets such as integrating legacy applications with Web development, project management, telecommunications, mobile commerce, and enterprise resource planning. More firms are turning externally not only to local vendors but also to services across the globe (Carmel, 1999). Staff supplementation from domestic contractors has evolved to a sophisticated model of partnering with offshore/nearshore software development firms. Many of these relationships evolved from a short-term project need for select skills to a long-term commitment of resources, cultural diversity efforts, and dependencies that integrate vendors as partners.

The most pervasive IT project, Year 2000 (Y2K), had constraints of skill sets, time, and budget. IT managers had to look at many alternatives for achieving compliance. Firms that planned as early as the mid-1990s had time to experiment and build new relationships. With governmental sanction and support, some countries and their business leaders recognized the competitive advantage their labor force could offer (O'Riain, 1997; Heeks, 1999; Trauth, 2000). An unusual need for services because of Y2K, economic disparity within a global workforce, and proactive efforts of some governments led to the fostering of offshore software development.

Early companies to outsource offshore were software vendors. Managing offshore development smoothly took years and a certain type of project management expertise in addition to a financial commitment from executives. The activity involved new applications and integrating software development with existing domestically built applications. The Y2K investment and intense cultural communication paid off for firms willing to work through the challenges. Not only did initial offshore projects provide a solution to the skill shortage, they also yielded substantial cost savings when compared to outsourcing the work domestically. Such factors resulted in these relationships continuing past Y2K, where some companies now regard their offshore arm as partners.

The IT outsourcing market was estimated at over US$100 billion by 2001 (Lacity and Willcocks, 2001). Although outsourced IT services can include call centers and facilities management, this discussion focuses on outsourcing software development. "Offshore" software development typically refers to engaging workers from another continent. Examples are U.S. companies using Indian contractors. "Nearshore" software development refers to vendor firms located in nearby countries often on the same continent, for example Dutch firms engaging Irish software developers. For our purposes, discussion of offshore software development issues is assumed to apply to nearshore outsourcing, since most of the issues are the same.

Most firms already have had experience supplementing their staff. Dealing with offshore firms, however, is relatively new for firms whose main business is not software development. Distance, time zones, language and cultural differences are some key issues that differentiate offshore software development from the use of domestic contractors or consultants (Carmel, 1999). Nearly 1 million IT-jobs will move offshore from the United States by 2017 (Gaudin, 2002).

The most common reasons for outsourcing are cost reduction, shortages of IT staff, reduced development time, quality of work, and internationalization (see Table 1).

APPROACHES TO OFFSHORE SOFTWARE DEVELOPMENT

The approaches taken to offshore development describe the basic features of how firms have structured their relationships with offshore software developers. The basic dimensions include the type of organization that provides offshore services, the nature of the working relationship that clients have with that organization, and whether offshore staff are on a defined project or in a staff augmentation role.

Type of Organization Providing Offshore Development

An important dimension for describing the relationship with foreign software developers would be the basic types of

Table 1. Reasons for engaging in offshore software development

- Cost reduction - due primarily to lower wages and secondarily from tax incentives. Accounting for indirect costs, such as travel and teleconferencing, overall average savings may be close to 25% (Overby, 2001).
- Access to a global pool of talent - quicker staffing thereby enabling faster time to delivery.
- 24×7 Productivity – the global delivery model allows a "follow-the sun" approach for round-the-clock handing-off of work to developers in different time zones.
- Quality – rigorous requirements of certifications (CMM, ISO) or cultural dedication to detail and quality.
- Localization - adapting software to local requirements may be best handled by resources in or near the target location.

organizations with which client firms form an outsourcing arrangement. The basic options are to contract directly with an offshore IT services firm, to engage a domestic IT services firm with offshore capabilities, or to set up a wholly owned software development center in an offshore location.

- Direct contract with an offshore firm – The client firm contracts directly with the offshore software firm. This option represents a range of alternatives.

  • Long-distance contracting – the offshore firm maintains no permanent staff in the country of their clients. Clients and offshore firms use long distance communications technologies such as phone, e-mail, and short-term visits.
  • Limited presence – Offshore firms locate some staff in client countries by setting up satellite offices in one or more cities. Most of the work is performed offshore, but a local project manager and a small local team answers questions and function as liaisons to solve problems.
  • Integrated on-site – the export firm provides a mix of extensive onsite expertise to handle business and requirements analysis, system design, and project management, and coordinates with offshore staff, normally working closely with client management. As many as 70 firms from Ireland have set up subsidiaries in the U.S. with Irish government financial support (Cochran, 2001).

- Contract with a domestic IT services firm with offshore capabilities – Many domestically-based IT services firms have the ability to send work offshore. This may take the form of an IT firm that acts as an intermediary, or a firm that itself owns an offshore satellite location. Some domestic firms employ both approaches. An intermediary acts as a liaison with offshore software firms, and already has subcontracting or partnering arrangements with them. The IT services firm works with clients in much the same way as other domestic IT services firms, but differs in that it uses offshore resources for some of the software development. Intermediaries may negotiate a cut of the cost savings their clients achieve by going offshore. The advantage in giving up the full savings of dealing directly with outsourcers is that the intermediaries buffer clients from most aspects of dealing with offshore providers. For small to midsize companies new to outsourcing, brokers offer the big benefit of offshore projects—good work done more cheaply, but without cross-cultural, long–distance, vendor management headaches.

- Wholly owned Satellite – The firm operates in an offshore location to directly hire IT professionals to become its employees. This is model is almost exclusively used by companies that sell software or by companies whose products have a large software component.

Length of Relationship and Type of Project

The working relationship between clients and vendors can be described by whether the relationship is short or long term, and the type of work the vendor performs.

Relationships with an offshore IT services firms can range from a few months to many years. Short-term relationships may be a single-hit project with no plan for follow-up projects. Long-term relationships may become important business partnerships where an offshore firm is responsible for an ongoing stream of projects.

The nature of work given to offshore vendors can range from well-defined projects to relatively unstructured ones. Typical examples of well-defined projects are technical design, implementation, maintenance, or software conversion. In other cases, vendors may take on significant responsibility

for project initiation, requirements determination, and project management involving new systems or major enhancements to existing systems that have strategic impact to the firm.

Carmel and Agarwal (2002) propose a four-stage model of offshore outsourcing that can be described in terms of time frame and nature of work. They describe firms moving from short-term structured engagements to longer-term structured engagements, and finally to long-term unstructured engagements as a maturation process. A company, however, may have outsourcing efforts that fall into different categories. For example an organization may engage offshore firms for both an ongoing stream of structured and unstructured projects, while at the same time evaluating other vendors, or outsourcing special needs projects on a one-time basis.

Defined Project vs. Staff Augmentation

The previous discussion of offshore development focused on contracting with the outsourcer for the delivery of a defined product. It's possible to tap some outsourcing firms as sources for augmenting staff. Companies contract outsourcing services on the basis of developer time instead of a fixed price software deliverable. Offshore staff may travel to the client's site and be managed closely by the client or may remain in the offshore location managed remotely. Offshore firms in this case have limited responsibility for managing projects.

SOFTWARE EXPORTING COUNTRIES

Table 2 lists dominant players by order of reported revenues, industry software organizations, and their Web sites. Many reports of exports are not standardized. It is not clear what is included in sales figures and who is reporting them. Even if these figures were comparable, the projection for evaluating different countries is not. Some countries are producing more software with fewer developers than others. The types of applications delivered may not compare. Converting a legacy customer database is not similar to customizing an enterprise system. More importantly, within a country there may be many development vendors and generalizations that are not applicable to each company.

CRITICAL ISSUES OF OFFSHORE SOFTWARE DEVELOPMENT

Challenges associated with offshore development are different than domestic development. In most cases the issues lead to creative solutions that some companies have tackled and incorporated to their advantage. Special circumstances of offshore development present factors that practitioners have found to be of special importance in achieving success.

Quality Certification

The Software Engineering Institute's Capability Maturity Model (CMM) (CM-SEI, 2002) or the International Organization for Standardization's ISO9001: 2001 standard (ISO, 2002) serves as an independent evaluation of the quality of a vendor's development processes. Certification should not be taken at face value. Investigation can clarify its meaning. When it was acquired? Was it qualified by project or by subsidiary? Who provided the evaluation? Was it self-assessed? There are a number of questions to raise.

Table 2. Information Technology Industry Organizations by Country (Enterprise Ireland promotes Irish exports, including software services.)

India	National Association of Software and Services Companies -NASSCom	www.nasscom.org
Philippines	Information Technology Association of the Philippines - ITAP	www.itaphil.org
Ireland	Enterprise Ireland*	www.enterprise-ireland.com
Israel	Israeli Association of Software Houses – I.A.S.H.	www.iash.org.il
Russia	RUSSOFT - The National Software Development Association of Russia	www.russoft.org
China	Software Offshore Business Union of Shanghai - SOBUS Ministry of Information Industry - MII	www.sobus.com.cn www.mii.gov.cn/mii/index.html
Canada	Information Technology Association of Canada	www.itac.ca
Mexico	AMITI - Mexican Association for the Information Technologies Industry	www.amiti.org.mx/

Onsite Presence

Despite the ability to operate virtually, the tenuous nature of offshore development is optimized when face-to-face contact, handshaking, and social/professional relationships nurture the engagement. The limited presence and integrated on-site approaches address the distance resistance.

Managing Globally

The downside of being productive 24×7 is that some part of the team is often not available for real-time contact because they are not at work at the same time. Travel by someone is a necessity in almost all offshore arrangements. The nature of the relationship will dictate how many and how often, from where to where. Sometimes this may involve extended stays.

Trust and Confidence

Relationship management is the overriding success factor when minimal levels of other factors are in place. IT services firms accent their marketing on these traits. For offshore firms to be seen as important business partners and take on work beyond routine development work, clients need to have a high level of trust and confidence in the outsourcers' abilities and intentions.

Political Environment

The stability of the offshore country's government prevents some firms from exploring offshore outsourcing (Nicholson and Sahay, 2001). Israel's thriving development curtailed with the Palestinian war. The Pakistan-India border disputes caused the U. S. State Department to ban travel to India. Although firms contract with a software firm and not the country, political environment of the country can influence the feasibility of working with firms located there.

Intellectual property concerns are usually dissuaded due to protection by law. Most countries have laws that are comprehensive enough to deal with software as ownership. However, having laws in place does not mean that governments enforce them.

Cultural Awareness and Integration

Close working relationships between offshore vendors and clients from another culture may require significant time, not only for becoming aware of each others' cultures, but to integrate it in practice. Firms with close working relationships not only achieve integration in software methods and delivery but also achieve integration in social aspects of team building.

Nationalism

Organizations that hire offshore developers may have to justify why they are not hiring domestically when many people are unemployed. Issuing of U.S. H1-B visas raised concern as many Americans became unemployed. The IT community was sensitive to the impact of hiring foreigners with special skills. Although visas are no longer granted as they once were, many organizations have had to deal with the same issue that arises with offshore software development. Human rights and workers' ability to earn a living wage are not comparable to client countries' workers—but no published reports of abuse or exploitation surfaced. Offshore software development offers governments ways to gain power and enhance their standing in the world.

Capacity of the Workforce

The availability of human resources to deliver IT demand is debatable. There is mixed evidence about the ability to meet demand (Overby, 2001; Carmel and Agarwal, 2002). Determining capacity is tricky. One must not only review population numbers but the educational, telecommunications, and managerial infrastructure that allows growth.

Risk Management

The most obvious threat relates to the political environment. Because some firms perceive security breaches as more vulnerable using offshore developers, outsourcers probably take extra precaution to sell their services. Y2K and September 11th reinforced the need for business continuity planning.

Risk management can also affect selection of firms in part on the basis of location. Developing relationships with a geographically diverse portfolio of outsourcers could minimize the risk of regional political and economic disturbances.

FUTURE TRENDS

Increased Offshore Outsourcing of Higher-Level System Life Cycle Tasks

Greater experience in using offshore developers on the part of client firms coupled with the maturation of technologies and methodologies for global software delivery allow firms to outsource more high-level development and realize the benefits of outsourcing across a wider variety of IT work.

Emergence of Major Contenders to India

India has been by far the major destination for offshore software development outsourcing, with Israel and Ireland usually considered "major players." China, Philippines, and Russia are often noted as potential contenders. Outsourcing industries are emerging in a number of unexpected locations such as Viet Nam, Mexico, Ukraine, and Iran due to low entry costs.

Increasing Use of Offshore Workers by Domestic Software Firms

To remain competitive, domestic IT outsourcing firms increasingly use offshore developers either through wholly-owned development centers offshore, or by subcontracting or partnering with foreign outsourcing firms (Overby, 2001).

Increased Sensitivity to Outsourcing Domestic Jobs

Reduced employment in importing countries creates a potential for a backlash. Ensuing restrictive changes in visas for foreign workers could limit higher-level work outsourced offshore since it requires greater onsite presence.

Increased Use of Third-Party Offshore Software Development Suppliers

Rather than bear project management costs directly, more importers will engage third-part consultants to act as a buffer and minimize relationship costs

CONCLUSIONS

Offshore software development outsourcing potentially provides considerable benefits. Some businesses are in the early stages of offshore experimentation, while others have formed mature business partnerships with offshore software firms. Many businesses will outsource more IT work offshore in order to cut costs and reduce time to market in addition to the benefits of quality. An increasing share of the IT work being outsourced is found at higher levels. These forces, however, may be compensated by factors such as political agenda of domestic employment, a desire to avoid negative public relations, and visa limitations that increase the challenges for offshore firms to provide on-site staff.

REFERENCES

Carmel, E. (1999). *Global Software Teams: Collaborating Across Borders and Time Zones.* New York: Prentice-Hall.

Carmel, E., and Agarwal, R. (2002). The Maturation of Offshore Sourcing of Information Technology Work. *MIS Quarterly Executive.* 1(2), 65-77.

CM-SEI. (2002). Carnegie Mellon Software Engineering Institute's Capability Maturity Model® for Software (SW-CMM®). Retrieved December 1, 2003 from the World Wide Web at: *http://www.sei.cmu.edu/cmm/cmm.html*

Cochran, R. (2001). Ireland: A Software Success Story. *IEEE Software*. 18(2), 87-89.

Gaudian, S. (2002, November 19). Nearly 1 Million IT Jobs Moving Offshore. *Datamation.* Retrieved December 1, 2003 from the World Wide Web at: *http://itmanagement.earthweb.com/career/article.php/1503461*

Heeks, R.B. (1999). Software Strategies in Developing Countries. *Communications of the ACM*. 42(6), 15-20.

ISO. (2002). International Organization for Standardization's "Selection and Use of the ISO 9000:2000 family of standards" includes ISO-9001. Retrieved December 1, 2003 from the World Wide Web at: *http://www.iso.org/iso/en/iso9000-14000/iso9000/selection_ use/selection_use.html*

Lacity, M. C., and Willcocks, L.P. (2001). *Global Information Technology Outsourcing: In Search of Business Advantage*. New York: Wiley.

Nicholson, B., and Sahay, S. (2001). Some Political and Cultural Issues in the Globalisation of Software Development: Case Experience from Britain and India. *Information and Organization,* 11(1).

O'Riain, S. (1997). The Birth of a Celtic Tiger. *Communications of the ACM*. 40(3), 11-16.

Overby, C. S. (2001, September). The Coming Offshore Service Crunch. *Forrester Report,* 1-21.

Trauth, E.M. (2000). The *Culture of an Information Economy: Influences and Impacts in The Republic of Ireland.* Netherlands: Kluwer, Dordrecht.

KEY TERMS

Capability Maturity Model (CMM): A methodology used to evaluate an organization's software development process. The model describes a five-level evolutionary path

of increasingly organized and systematically more mature processes.

Global Delivery Model: To distribute and manage software development across multiple global locations. GDM requires infrastructure, project management, cultural sensitivity, and process guidelines to support communication and coordination between locations.

Intellectual property rights: Laws and enforcement mechanisms to afford the creator of an intellectual property (e.g., software) the means of controlling how their work is used, and ensuring that the creator is properly rewarded and recognized for their work**.**

ISO 9001:2001: Provides standards used to assess an organization's ability to meet its customer's requirements and achieve customer satisfaction. Software firms use it as an alternative to CMM.

Localization: The process of adapting software to a particular language, culture, and desired local "look-and-feel". This may include local sensitivities, geographic examples, and adhering to local legal and business requirements.

Staff Augmentation: When an employee of an outsourcing firm is lent to a client company for work. Normally a staff augmentation employee works at the client's worksite with the technical direction coming from the client company.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2174-2179,

OMIS-Based Collaboration with Service-Oriented Design

Kam Hou Vat
University of Macau, Macau

INTRODUCTION

The success of today's enterprises, measured in terms of their ability to learn and to apply lessons learned, is highly dependent on the inner workings and capabilities of their information technology (IT) function. This is largely due to the emergence of the digital economy (Ghosh, 2006; Turban, Leidner, McLean, & Wetherbe, 2005), characterized by a highly competitive and turbulent business environment, inextricably driven by the intra- and inter-organizational processes and the knowledge processing activities they support. One consequence is the increase in organizations' efforts to deliberately manage knowledge (Tapscott, 1997), especially the intellectual capital (Stewart, 1997) of their employees (De Hoog, van Heijst, van der Spek, et al., 1999), which necessarily deals with the conceptualization, review, consolidation, and action phases of creating, securing, combining, coordinating, and retrieving knowledge. In fact, such efforts must be instrumental to creating an efficient organization model based on some innovative initiative, and then enable the organization to launch and learn. In a knowledge-creating organization (Nonaka & Takeuchi, 1995), employees are expected to continually improvise, and invent new methods to deal with unexpected problems and share these innovations with other employees through some effective channels of communications or knowledge transfer mechanisms. The key is collaboration, implying that organizational knowledge is created only when individuals keep modifying their knowledge through interactions with other organizational members. The challenge that organizations now face is how to devise suitable information systems (IS) support to enable such collaboration, namely, to turn the scattered, diverse knowledge of their people into well-documented knowledge assets ready for reuse to benefit the whole organization. This article presents some service-oriented perspectives of employee-based collaboration through the design of specific IS support called the Organizational Memory Information System (OMIS) in light of the peculiar open-source development initiative of Wiki technology (Leuf & Cunningham, 2001).

BACKGROUND

Lately, an organization's ability to learn is often considered a process of development to organizational memory. By organizational memory (Walsh & Ungson 1991), we are referring to various structures within an organization that hold knowledge in one form or another, such as databases and other information stores, work processes, procedures, and product or service architecture. As a result, organizational memory (OM) must be nurtured to assimilate new ideas and transform those ideas into action and knowledge, which could benefit the rest of the organization (Ulrich, Von Glinlow, & Jick 1993). Through understanding the important components of the OM (Vat, 2001), an organization can better appreciate how it is currently learning from its key experiences, to ensure that relevant knowledge becomes embedded within the future operations and practices of the organization. In practice, creating and using an OM is a cooperative activity necessarily involving many members of an organization. If those individuals are not adequately motivated in contributing to the OM initiative, and the organizational culture does not support knowledge sharing (Orlinkowski, 1992), it is not likely to turn the scattered, diverse knowledge present in various forms into well-structured knowledge assets ready for deposit and reuse in the OM.

Consequently, it is important to distinguish between the organizational memory (OM encompassing people) and the OMIS that captures in a computational form only part of the knowledge of the organization. The OM captures the knowledge of the organization. The associated OMIS makes part of this knowledge available either by providing direct access to it (e.g., codified knowledge assets such as experience reports) or indirectly by providing knowledge maps (e.g., tacit knowledge assets such as personnel with specific expertise). Managing the OM deals first of all with the question of "Which knowledge should go into the OMIS?" Answering this question requires determining what knowledge is owned by the members of the organization, what knowledge is needed now, what is going to be needed in the future, and for what purposes. This helps the organization to define not only a strategy for acquiring the needed knowledge, but also to establish validation criteria in relation to the defined goals. Besides, we also need to deal with "who needs the knowledge, when and why," as well as the policies for accessing and using the OMIS. This contextualization of the OMIS with respect to the organization's ability to learn is essential to implement the mechanisms of organizational knowledge transfer, examples of which are discussed in Vat (2006). In fact, in this modern age of information technology and swift change, learning has become an integral part of

the work of an organization run along principles intended to encourage constant reshaping and change. An OMIS-based organization can be characterized as one that continuously transforms itself by developing the skills of all its people and by achieving what Argyris (1992) has called *double-loop learning,* which helps transfer learning from individuals to a group, provide for organizational renewal, keep an open attitude to the outside world, and support a commitment to knowledge. One of the missions of the OMIS is to facilitate and bring about the fundamental shifts in thinking and interacting and the new capabilities needed in the organization.

SERVICE-ORIENTED DESIGN FOR OMIS

When designing an OMIS to nurture an organization's ability to learn (Vat, 2001, 2002), we consider the following modes of learning behavior: (1) individual, (2) group, and (3) repository. Individual learning is characterized by knowledge being developed, and possibly the result of combining an insight with know-how from other sources in the organization, but it is often not distributed and is not secured for reuse. Group learning is centered around the concept of communication in two possible modes: supply-driven or demand-driven. The former is characterized by an individual who has found a way to improve the work process and communicates this to one's coworkers. The latter refers to a worker who has recognized a problem in the current process and asks fellow workers whether they have a solution for this problem. In each case, knowledge is developed, distributed, and possibly combined with knowledge from other parts of the organization, but it is seldom secured. In repository learning, the communication element is replaced by collection, storage, and retrieval of knowledge items. Namely, it is typified by storing lessons learned in some information repository so that they can be retrieved and used when needed. Overall, in repository learning, knowledge is developed, secured, distributed, and is possibly the result of knowledge combination. It is convinced that the requirements of an OMIS design should be formulated in terms of some typical usage scenarios. Namely, an OMIS should facilitate individual workers to access the knowledge required by combination, to submit a lesson learned, and to decide which of the coworkers would be interested in a lesson learned. Also, there should be criteria to determine if something is a lesson learned, how it should be formulated and where it should be stored, and how to distribute some newly asserted knowledge piece to the workers in need. The perceived technical issues, nevertheless, could include the following: How are we to organize and index the OM to enhance its diffusion? How does an organization retrieve relevant elements of the OM to answer a user request or proactively push relevant elements towards users? How does an organization adapt the answer to users, in particular to their tasks, according to the knowledge contexts? These problems are largely related to the OM framework for knowledge distribution, whose goal is to improve organizational learning, with the aid of the previously mentioned OMIS support whose discussion through the idea of service-orientation is our major concern in the following section.

The Context of Service-Orientation

The term "service" has existed for some time (Chesbrough & Spohrer, 2006), and its attendant "service-oriented" connotation has also been used in different contexts and for different purposes (Rust & Miu, 2006). According to Erl (2005), one constant characteristic of this term currently identified among the research community is that it represents a distinct approach for separating concerns. Simply stated, the effort or logic required to solve any problem can be better constructed, executed, and managed if it is decomposed into a collection of smaller, related pieces. Each of these pieces addresses a concern or a specific part of the problem. Indeed, this thinking is not new and it does transcend technology and automation solutions, especially in the IT field, but what distinguishes the service-oriented approach to separating concerns is the manner in which it achieves separation. Consider our city that is full of service-oriented businesses, each of which provides a distinct service that can be used by multiple consumers. Collectively, these businesses comprise a community, decomposable into specialized, individual outlets, providing all possible business services. More importantly, individual outlets are encouraged to interact and leverage one another's services. Nonetheless, we want to avoid a model in which outlets form tight connections that result in constrictive inter-dependencies. Preferably, businesses are empowered to self-govern their individual services so as to evolve and grow relatively independent of each other. Meanwhile, it is also important to ensure that service providers must adhere to certain baseline conventions that standardize key aspects of each business for the benefit of the consumers without significantly imposing on the individual provider's ability to exercise self-governance.

The Promise of Service-Oriented Computing

With the rapid increase of software applications for the daily running of modern businesses, service-oriented computing (SoC) (Dijkman & Dumas, 2004) is emerging as a promising paradigm for enabling the flexible interconnection of autonomously developed applications operating within and across organizational boundaries (Alonso, Casati, Kuno, & Machiraju, 2003). Under the SoC paradigm, the functionality of existing applications can be expressed as services or a network of services called service compositions (Casati & Shan,

2001; Benatallah, Sheng, & Dumas, 2003). Currently, the SoC paradigm is mainly associated with enabling technology founded on such standards as SOAP (Simple Object Access Protocol), WSDL (Web Services Description Language), WS-Security, and BPEL (Business Process Execution Language). Such technology enables businesses to describe the services they offer, to publish these descriptions online, to find other services based on their descriptions, and to build applications using those services. The term "service-oriented design" was coined by Dijkman and Dumas (2004) to represent the set of modeling languages, methods, and techniques used to design such services, to verify the conformance of services to their requirements, and to enable a model-driven approach to service development and composition.

The Challenge in Service-Oriented Design

In anticipation of the emerging service opportunities to provide enterprise solutions that can extend or change on demand, Dimmermann, Krogdahl, and Gee (2004) enumerated three major levels of abstractions to be managed within the service-oriented design process:

- **Operations:** These are transactions that represent single logical units of work (LUWs). Execution of an operation will typically cause one or more persistent data records to be read, written, or modified. SoC operations are typically comparable to object-oriented (OO) methods. They have a specific, structured interface, and return structured responses.
- **Services:** These represent logical groupings of operations. For example, if we consider *KnowledgePortfolio* as a service, then *Lookup knowledge objects by reference number, List knowledge item by name and call reference,* and *Save data for new knowledge* represent the associated operations.
- **Business Processes:** These represent a long-running set of actions or activities performed with specific business goals in mind. Business processes typically encompass multiple service invocations. Examples include: *Initiate New Student, Create StudentPortfolio, Showcase StudentPortfolio,* or *View StudentPortfolio.* In SoC terms, a business process consists of a series of operations executed in an ordered sequence according to a set of business rules. The sequencing, selection, and execution of operations is termed a service or process choreography. Typically, choreographed services are invoked in response to business events.

From a modeling standpoint, the challenge in service-oriented design for OMIS is how to characterize in a well-specified manner those operations, services, and process abstractions systematically for such architectural components as individual learning, organizational learning, and intellectual property management.

FUTURE TRENDS

Much of earlier literature review (Ghosh, 2006; Badaracco 1991; Hamel & Prahalad 1994; Quinn 1992; Pinchot & Pinchot 1994) supports the supposition that intellectual material in the form of information, knowledge, and any other form of intellectual property is a valued organizational asset, and organizations are increasingly dependent on information technology (IT) for the transfer of knowledge and information. Conspicuously missing, however, is often a discussion of collaboration (Tabaka, 2006; Schrage 1990) as a regenerative source of ideas that will advance organizations to learn, change, and excel (Menon, 1993; Stewart, 1994). To collaborate is to work in a joint intellectual effort, to partition problem solving to produce a synergy such that the performance of the whole exceeds that of any individual contributor. The central issue in organizational learning is how individual learning is transferred to the organizational level. In this regard, the use of Wiki technology (*www.wiki.org*) as a collaborative tool within an organizational setting renders an excellent example. Yet, only with a clear understanding of the transfer process can we manage learning processes consistent with organizational goals, issues, and values. If this transfer process was indeed actualized in the design and practice of the OMIS, we could well have a knowledge organization with the capability of capturing learning in its different paths and incorporating that learning into the running of its daily operations.

The Service-Oriented Aspects of Wiki Technology

Wiki technology is based on open-source software. The software that operates any Wiki is called a Wiki engine (Kille, 2006). A variety of free Wiki engines (also known as Wiki clones) are available from the Web (*www.wiki.org*). There are also Wiki hosts offering Wiki service with a minimal fee, such as the Seedwiki (*www.seedwiki.com*), and JotSpot (*www.jot.com*). The first Wiki application invented by Ward Cunningham in 1995 was to publish information collaboratively on the Web (Leuf & Cunningham, 2001), and this first Wiki Web site (*c2.com/cgi/wiki*) is still actively maintained today. Leuf and Cunningham (2001, p. 14) define a Wiki (Hawaiian word meaning *quick*) as a freely expandable collection of interlinked Web pages, a hypertext system for storing and modifying information. Cunningham's original vision was to create a Wiki as the simplest online database that could possibly work. Today, Wikis are interactive Web sites that can offer numerous benefits to users (Wagner, 2004), in the form of a simple editing and publishing interface that can

be used and understood easily. Anyone can create a new Wiki page, add or edit content in an existing Wiki page, and delete content within a page, without any prior knowledge or skills in editing and publishing on the Web. In fact, the major distinguishing factor between Wikis and regular Web sites is the ability of Wiki users to easily edit all aspects of a Wiki Web site. Fuchs-Kittowsk and Kohler (2002, p. 10) interpret a Wiki as an open author system for a conjoined construction and maintenance of Web sites. They suggest that Wiki technology can facilitate cooperative work and knowledge generation in such contexts as content management systems, discussion boards, and other innovative forms of groupware. Indeed, members of a Wiki community can build and develop meaningful topic associations by creating numerous links among Wiki pages. To make the Wiki technology useful for collaborative work in organizations, Wagner (2004, p. 270) suggested 11 principles that govern the functional design of a Wiki application:

- **Open:** If a Wiki page is found to be incomplete or poorly organized, any reader can edit it as he or she sees fit.
- **Incremental:** Wiki pages can cite other pages, including pages that have not been written yet.
- **Organic:** The structure and text content of the site is open to editing and evolution.
- **Mundane:** A small number of (irregular) text conventions will provide access to the most useful but limited page markup.
- **Universal:** The mechanisms of editing and organizing are the same as those of writing, so that any writer is automatically an editor and organizer.
- **Overt:** The formatted (and printed) output will suggest the input required to reproduce it.
- **Unified:** Page names will be drawn from a flat space so that no additional context is required to interpret them.
- **Precise:** Pages will be titled with sufficient precision to avoid most name clashes, typically by forming noun phrases.
- **Tolerant:** Interpretable (even if undesirable) behavior is preferred to error message.
- **Observable:** Activity within the site can be watched and reviewed by any other visitor to the site. Wiki pages are developed based on trust.
- **Convergent:** Duplication can be discouraged or removed by finding and citing similar or related content.

The Potential Benefits as a Collaborative Tool

According to Wagner (2004) and Raman, Ryan, and Olfman (2005), the use of Wiki technology can address some knowledge management goals for collaborative work and organizational learning. Here, a knowledge management system refers to any IT-based system that is developed to support and enhance the organizational processes of knowledge creation, storage, retrieval, transfer, and application (Alavi & Leidner, 2001, p. 114). In particular, any Wiki clone can be designed to support such basic functions as searching and indexing capabilities for effective retrieval and storage of knowledge attributes. The most often cited benefits of using Wikis to support collaborative work thereby include the simplicity of learning and working with the technology, and the free download through the Wiki engines of all the necessary knowledge items of interest throughout the organization. More importantly, Davenport and Prusak (1998) provide three essential reasons why organizations need such a technology to implement its knowledge management systems:

1. To enhance visibility of knowledge in organizations through the use of maps, hypertexts, yellow pages, and directories;
2. to build a knowledge-sharing culture, namely, to create avenues for employees to share knowledge; and
3. to develop a knowledge infrastructure, not confined solely to technology, but to create an environment that permits collaborative work.

CONCLUSION

If designed and implemented effectively, Wiki technology can support a portion of an organization's collaboration and knowledge management requirements — specifically, knowledge sharing, storing, and support for the communication process within organizations. A key advantage of using Wikis to support knowledge management initiatives is that the technology is free. Nonetheless, issues such as sufficient user training, the availability of resources and skills to support the technology, and effective customization of Wiki features must be considered before the value of using the technology to support collaborative work within any organization is to be realized. Meanwhile, the use of service-oriented design is yet to be explored in terms of a more systematic methodology to enable enterprises to describe, publish, and compose application services (in the specific area of collaborative work and knowledge management), and to communicate with applications of other enterprises according to their service descriptions. The current development of SoC (Quartel, Dijkman, & Sinderen, 2004) promises to deliver the methods and technologies to help business partners link their software applications. This should facilitate the introduction of richer and more advanced applications (other than the Wiki applications), thereby offering new collaborative opportunities. Currently, we consider service-oriented design as the process of designing application support for one or more

intra- and/or inter-organizational processes using the SoC paradigm, which is characterized by the explicit identification and description of the externally observable behavior (service) of an application. Thereby, applications can then be linked, based on the description of their externally observable behaviors. According to this paradigm, developers in principle do not need to have any knowledge about the internal functioning of the applications being linked. This peculiar feature of separation of concerns forms the basis of service-orientation that has been elaborated in this article as a promising means of designing collaborative work within an organizational setting in the immediate future.

REFERENCES

Alavi, M., & Leidner, D.E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issue. *MIS Quarterly, 25*(1), 107-136.

Argyris, C. (1992). *On organizational learning.* Cambridge, MA: Blackwell Business.

Alonso, G. Casati, F, Kuno, H., & Machiraju, V. (2003). *Web services: Concepts, architectures and applications.* Berlin: Springer-Verlag.

Badaracco, J. (1991). *The knowledge link.* Boston: Harvard Business School Press.

Benatallah, B., Sheng, Q., & Dumas, M. (2003). The self-serv environment for Web services composition. *IEEE Internet Computing, 7*(1), 40-48.

Casati, F., & Shan, M.C. (2001). Dynamic and adaptive composition of e-services. *Information Systems, 26*(3), 143-162.

Chesbrough, H., & Spohrer, J. (2006). A research manifesto for services science. *Communications of the ACM*, *49*(7), 35-40.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Boston: Harvard Business School Press.

De Hoog, R., van Heijst, G., van der Spek, R., Edwards, J.S., Mallis R., van der Meij, B., & Taylor, R.M. (1999). Investigating a theoretical framework for knowledge management: A gaming approach. In J. Liebowitz (Ed.), *Knowledge management handbook*. Berlin: Springer-Verlag.

Dijkman, R., & Dumas, M. (2004). Service-oriented design: A multi-viewpoint approach. *International Journal of Cooperative Information Systems, 13*(4), 337-378.

Dimmermann, O., Krogdahl, P., & Gee, C. (2004). Elements of service-oriented analysis and design: An interdisciplinary modeling approach for SOA projects. Retrieved December 11, 2006, from http://www-128.ibm.com/developerworks/webservices/library/ws-soad1/

Erl, T. (2005). *Service-oriented architecture: Concepts, technology, and design.* New York: Prentice Hall.

Fuchs-Kittowsk, F., & Kohler, A. (2002). Knowledge creating communities in the context of work processes. *ACM SIGCSE Bulletin, 23*(3), 8-13.

Ghosh, R.A. (2006). *CODE: Collaborative ownership and the digital economy.* Boston: MIT Press.

Hamel, G., & Prahalad, C. (1994). *Competing for the future.* Boston: Harvard Business School Press.

Kille, A. (2006). *Wikis in the workplace: How Wikis can help manage knowledge in library reference services.* Retrieved December 11, 2006, from http://libres.curtin.edu.au/libres16n1/Kille_essayopinion.htm

Leuf, B., & Cunningham, W. (2001). *The Wiki way: Quick collaboration on the Web.* New York: Addison-Wesley.

Menon, A. (1993). Are we squandering our intellectual capital? *Marketing Research: A Magazine of Management, 5*(3), 18-22.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company: How Japanese companies create the dynamics of innovation.* Oxford: Oxford University Press.

Orlikowski, W.J. (1992). Learning from notes: Organizational issues in groupware implementation. *Proceedings of the 1992 ACM Conference on Computer-Supported Cooperative Work* (CSCW'92) (pp. 362-369).

Pinchot, G., & Pinchot, E. (1994). *The end of bureaucracy and the rise of intelligent organization.* San Francisco: Berrett Koehler.

Quartel, D., Dijkman, R., & Sinderen, M. (2004, November 15-19). Methodological support for service-oriented design with ISDL. *Proceedings of the 2nd International Conference on Service-Oriented Computing* (ICSOC'04) (pp. 1-10).

Quinn, J.B. (1992). *Intelligent enterprise.* New York: The Free Press.

Raman, M., Ryan, T., & Olfman, L. (2005). Designing knowledge management systems for teaching and learning with Wiki technology. *Journal of Information Systems Education, 16*(3), 311-320.

Rust, R.T., & Miu, C. (2006). What academic research tells us about service. *Communications of ACM, 49*(7), 49-54.

Schrage, M. (1990). *Shared minds.* New York: Random House.

Senge, P.M. (1990). *The fifth discipline: The art and practice of the learning organization.* London: Currency Doubleday.

Stewart, T. (1997). *Intellectual capital: The new wealth of organizations.* New York: Doubleday.

Stewart, T. (1994). Measuring company I.Q. *Fortune, 129*(2), 24.

Tabaka, J. (2006). *Collaboration explained: Facilitation skills for software project leaders.* Boston: Addison-Wesley.

Tapscott, D. (1997). *The digital economy: Promise and peril in the age of networked intelligence.* New York: McGraw-Hill.

Turban, E., Leidner, D., McLean, E., & Wetherbe, J. (2005). *Information technology for management: Transforming organizations in the digital economy.* New York: John Wiley & Sons.

Ulrich, D., Von Glinlow, M., & Jick, T. (1993). High-impact learning: Building and diffusing a learning capability. *Organization Dynamics,* Autumn, 22 (2): 52-66.

Vat, K.H. (2002). Designing organizational memory for knowledge management support in collaborative learning. In D. White (Ed.), *Knowledge mapping and management* (pp. 233-243). Hershey, PA: IRM Press.

Vat, K.H. (2006). Knowledge synthesis framework. In D. Schwartz (Ed.), *Encyclopedia of knowledge management* (pp. 530-537). Hershey, PA: Idea Group.

Vat, K.H. (2001, November 1-4). Towards a learning organization model for knowledge synthesis: An IS perspective. *CD-Proceedings of the 2001 Information Systems Education Conference* (ISECON2001), Cincinnati, OH.

Wagner, C. (2004). Wiki: A technology for conversational knowledge management and group collaboration. Communications of the *AIS, 13*, *(19)*, pp. 256-289.

Walsh, J.P., & Ungson, G.R. (1991). Organizational memory. *Academy of Management Review, 16*(1), 57-91.

KEY TERMS

Collaboration: To facilitate the process of shared creation involving two or more individuals interacting to create shared understanding where none had existed or could have existed on its own.

Double-Loop Learning: Together with single-loop learning, describes the way in which organizations may learn to respond appropriately to change. Single-loop learning requires adjustments to procedures and operations within the framework of customary, accepted assumptions, but fails to recognize or deal effectively with problems that may challenge fundamental aspects of organizational culture, norms, or objectives. Double-loop learning questions those assumptions from the vantage point of higher-order, shared views, in order to solve problems.

Knowledge Management: The broad process of locating, organizing, transferring, and using the information and expertise within the organization, typically by using advanced information technologies.

Learning Organization: An organization that focuses on developing and using its information and knowledge capabilities to achieve the following: to create higher-value information and knowledge, to modify behaviors to reflect new knowledge and insights, and to improve bottom-line results.

Organizational Learning: A process of leveraging the collective individual learning of an organization to produce a higher-level organization-wide intellectual asset. It is a continuous process of creating, acquiring, and transferring knowledge accompanied by a modification of behavior to reflect new knowledge and insight, and produce a higher-level asset.

Organizational Memory: A learning history that tells an organization its own story that should help generate reflective conversations among organizational members. Operationally, an organizational memory has come to be a close partner of knowledge management, denoting the actual content that a knowledge management system purports to manage.

Organizational Memory Information System (OMIS): An information system supporting the development of organizational memory, whose design philosophy is often organization specific. An example philosophy is to consider the OMIS as a meaning attribution system in which people select certain resource items out of the mass potentially available and get them processed to make them meaningful in a particular context in order to support their purposeful actions.

Service-Oriented Computing: A field of research focusing on the development of such technology that enables enterprises to describe the services they offer in a textual, mostly XML-based form, to publish these descriptions online and find services of other enterprises according to these descriptions, to compose services into new services, and to communicate with applications of other enterprises according to their service descriptions.

Service-Oriented Design: The process of designing software application support for one or more business processes, using the service-oriented computing paradigm.

Wiki Technology: Technology based on open-source software in the form of a Wiki engine. The Hawaiian word "Wiki" means "quick," with the connotation that this technology is easy to use once installed. Wikis run over the World Wide Web and can be supported by any browser. The technology is governed by an underlying hypertext transfer protocol (HTTP) that determines client and server communication. Wikis are able to respond to both requests for data (GET) and data submission (POST), in a given Web front, based on the HTTP concept.

On a Design of Narrowband FIR Low-Pass Filters

Gordana Jovanovic Dolecek
INSTITUTE INAOE, Puebla, Mexico

Javier Diaz Carmona
INSTITUTE ITC, Celaya, Mexico

INTRODUCTION

Stearns and David (1996) states that "for many diverse applications, information is now most conveniently recorded, transmitted, and stored in digital form, and as a result, digital signal processing (DSP) has become an exceptionally important modern tool." Typical operation in DSP is digital filtering. Frequency selective digital filter is used to pass desired frequency components in a signal without distortion and to attenuate other frequency components (Smith, 2002; White, 2000). The pass-band is defined as the frequency range allowed to pass through the filter. The frequency band that lies within the filter stop-band is blocked by the filter and therefore eliminated from the output signal. The range of frequencies between the pass-band and the stop-band is called the transition band and for this region no filter specification is given.

Digital filters can be characterized either in terms of the frequency response or the impulse response (Diniz, da Silva & Netto, 2002). Depending on its frequency characteristic, a digital filter is either low-pass, high-pass, band-pass, or band-stop filters. A low-pass (LP) filter passes low frequency components to the output, while eliminating high-frequency components. Conversely, the high-pass (HP) filter passes all high-frequency components and rejects all low-frequency components. The band-pass (BP) filter blocks both low- and high-frequency components while passing the intermediate range. The band-stop (BS) filter eliminates the intermediate band of frequencies while passing both low- and high-frequency components.

In terms of their impulse responses digital filters are either infinite impulse response (IIR) or finite impulse response (FIR) digital filters. Each of four types of filters (LP, HP, BP, and BS) can be designed as an FIR or an IIR filter (Ifeachor & Jervis, 2001; Mitra, 2005; Oppenheim & Schafer, 1999).

The design of a digital filter is carried out in three steps (Ingle & Proakis, 1997):

- Define filter specification
- Approximate given specification
- Implement digital filter in hardware or software.

The topic of filter design is concerned with finding a magnitude response (or, equivalently, a gain) which meets the given specifications. These specifications are usually expressed in terms of the desired pass-band and stop-band edge frequencies ω_p and ω_s, the permitted deviations in the pass-band (pass-band ripple) R_p, and the desired minimum stop-band attenuation A_s, (Mitra, 2005). Figure 1 illustrates a typical magnitude specification of a digital low-pass filter.

In many applications it is often advantageous to employ FIR filters, since they can be designed with exact linear phase, and exibits no stability problems. However FIR filters have a computationally more intensive complexity compared to IIR filters. During past several years, many design methods have been proposed to reduce complexity of the FIR filters, (Chen, Chang and Vinod, 2006; Jovanovic-Dolecek & Mitra, 2007; Lian and Yang, 2001; Rodrigues & Pai, 2005; Yang and Lian, 2003;Yang and Lian, 2006; Zou & Saramaki, 2004).

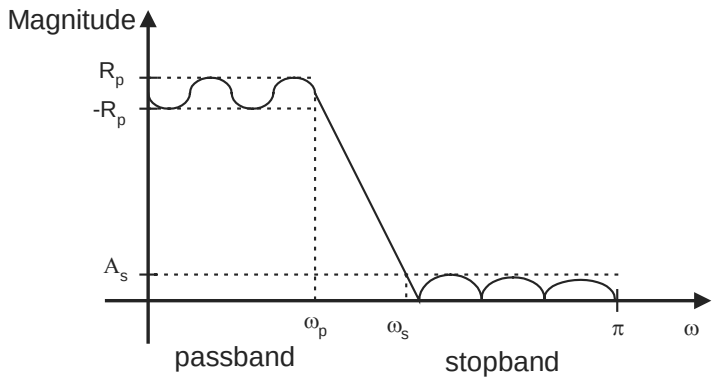
This chapter describes a class of digital filters, called IFIR (interpolated finite impulse response filters), that can implement narrowband low-pass FIR filters with a significantly reduced computational complexity.

The IFIR filter $H(z)$ is a cascade of two filters, an expanded shaping or model filter $G(z^M)$ and an interpolator or image suppressor $I(z)$. In this manner, the narrowband FIR prototype filter $H(z)$ is designed using lower order filters, $G(z)$ and $I(z)$. For more details on the IFIR structure see Neuvo, Cheng, and Mitra (1984) and Jovanovic Dolecek (2003).

An increase in the interpolation factor results in the increase of the interpolation filter order as well as in the decrease of the shaping filter order.

The proposal in Jovanovic Dolecek and Diaz-Carmona (2003) decreases the shaping filter order as much as possible, and efficiently implements the high order interpolator filter. To do so, in Jovanovic Dolecek and Diaz-Carmona (2005) the use of a sharpening recursive running sum (RRS) is proposed as an interpolator in the IFIR structure. The transfer function of an RRS filter with length M is given by

Figure 1. Low-pass filter magnitude specification



$$H_{RRS}(z)=\left(\frac{1}{M}\sum_{k=0}^{M-1}z^{-k}\right)^{L}=\left(\frac{1}{M}\frac{1-z^{-M}}{1-z^{-1}}\right)^{L} \qquad (1)$$

where L is the number of stages. The RRS filter is a linear-phase low-pass filter that requires no multipliers and two additions per output sample, but it has a high pass-band droop and a low stop-band attenuation. The filter sharpening technique, (Hartnett & Boudreaux, 1995; Kaiser & Hamming, 1984; Samadi, 2000), can be used to improve the frequency characteristic of the RRS filter.

We describe here how to find the interpolation factor M, and the number of the stages L, for a given sharpening polynomial, so that the specifications are satisfied with as low as possible complexity.

BACKGROUND

Filter Sharpening

The sharpening technique, which was first proposed by Kaiser and Hamming (1984), attempts to improve both the pass-band and stop-band of a linear FIR filter by using multiple copies of the same filter. This technique is based on the use of a polynomial approximation which maps a transfer function before sharpening, to a new transfer function. The sharpening polynomial is obtained using the closed formula proposed by Samadi (2000). The polynomial coefficients are always integers, which can be implemented as Shift-and-Add multipliers (Parhi, 1999).

Algorithm Description

This section exposes the algorithm for computing the main design parameters for the design of a narrowband low-pass filter using an IFIR structure, where the interpolator filter is the sharpening recursive running (RRS) filter.

The algorithm is based on an iterative design of the model filter G(z) using following specifications:

$$\omega_{Gp}=M\omega_p,\quad \omega_{Gs}=M\omega_s,$$
$$\delta_{Gp}=\alpha\delta_p,\quad \delta_{Gs}=\delta_s, \qquad (2)$$

where ω_{Gp} and ω_{Gs} are the normalized pass-band and stop-band frequencies, respectively, δ_{Gp} the maximum pass-band ripple and δ_{Gs} the minimum stop-band attenuation of the model filter $G(z)$, and ω_p, ω_s, δ_p and δ_s are the corresponding design specifications of the desired filter. The parameter α is a key factor in this design. The higher value of α the smaller model filter order, and vice versa. For a given value of M the maximum value of α is determined by the sharpened RRS filter pass-band droop. Our experience says that it is useful to take the values of α in the range $0.5 \le \alpha \le 0.9$.

In Figure 2 is shown the flow diagram of the algorithm. At first step, the maximum possible interpolation factor, $M = M_{max} = \pi/\omega_s$ and the minimum possible number of RRS stages $L = 1$, are selected. The choice of M is a key factor to satisfy the pass-band specification, while the choice of L is a key factor to satisfy the stop-band specification of the designed overall filter. Consequently, in the next steps the factor M is increased until the pass-band specification is satisfied, interpolation factor searching loop, while the parameter L is increased until the stop-band specification is reached, RRS stages searching loop. If as a result, an even

value of M and an odd value of L are obtained from these searching loops, the factor M is reduced by one to avoid a branch fractional delay in the sharpened RRS filter (Webb & Munson, 1996). In the following steps the algorithm looks for a possible further model filter order reduction by increasing the model pass-band ripple δ_{Gp}, while keeping satisfied the given overall filter specification.

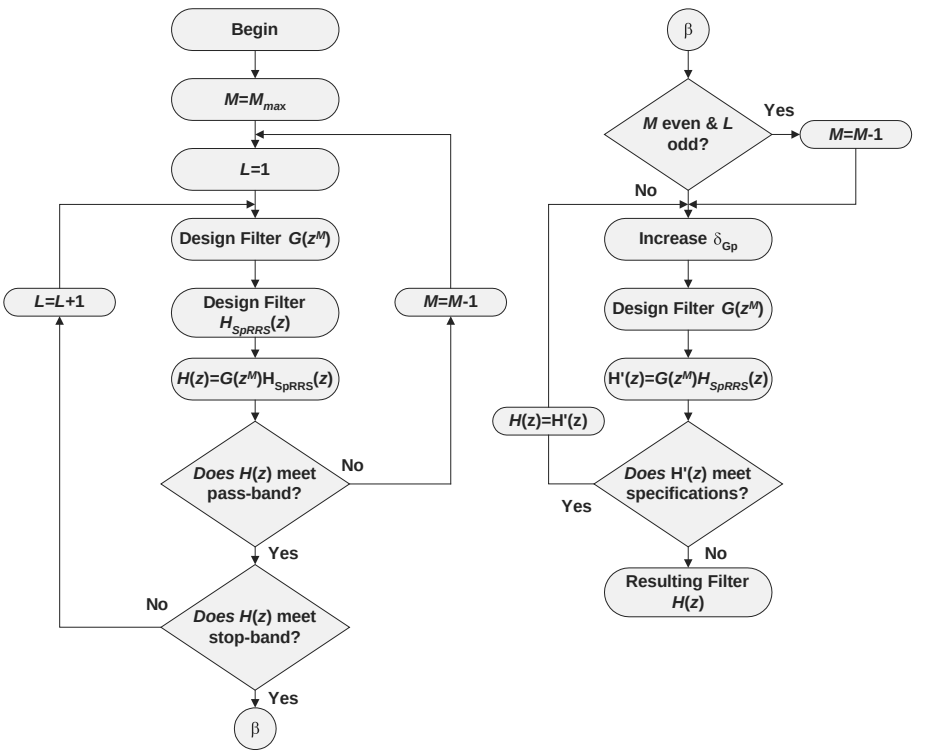
The algorithm complexity, number of operations performed, is strongly determined by the interpolation factor and RRS stages searching loops, this is the number of iterations required to meet the desired pass-band and stop-band specifications, respectively. The main iteration computational workload is the design of both the model filter, by Parks McClellan algorithm, and the sharpened RRS filter, by the sharpening polynomial. Hence the total algorithm complexity depends on how fast the parameters are found for both pass-band and stop-band specifications. This speed is determined basically by two factors:

1. Desired filter specifications: in this sense the narrower pass-band the smaller number of iterations in the interpolation factor searching loop, and the smaller stop-band attenuation the smaller number of iterations in the RRS stages searching loop.
2. Sharpening polynomial structure: due the sharpening polynomial modifies the RRS pass-band and stop-band frequency magnitude response. For instance the higher the sharpening polynomial improves the RRS pass-band the smaller number of iterations in the interpolation factor searching loop, and the higher the sharpening polynomial improves the RRS stop-band the bigger number of iterations in the RRS stages searching loop.

Accordingly to the algorithm there can be possible no convergence cases:

1. Any interpolation factor M meets the pass-band specification.
2. The increment of L in the RRS stages searching loop causes a not pass-band meeting for any M in the interpolation factor searching loop. If the design is limited to narrowband filters the first case is completely avoided and the second case is less prone to occur. Both no convergence cases are detected with a fixed maximum number of iterations.

Figure 2. Flow diagram of the algorithm



Design Example

The described algorithm was implemented in MATLAB and different filters have been designed in a PC PIV at 1.7 GHz. We show as design example the low-pass linear-phase FIR filter, proposed in Saramaki, Neuvo, and Mitra (1988), with next specifications: ω_p=0.05π, ω_s=0.1π, δ_p=0.01, δ_s=0.001. According to the proposal by Mehrnia and Willson (2004) the optimal interpolation factor M=4, results in the total number of products per output sample of 25 and the total number of additions of 47.

The design parameters obtained using the described algorithm for the sharpening polynomial with σ =δ =0, m=n=1 and α = 0.6 are: L= 2, M= 3 with a number of floating point operations of 112885030 FLOPS and a computation time of 3.4 seconds. The model filter order is N_G = 35. The design parameters for the sharpening polynomial with σ =δ = 0, m= 2, n = 1 and α = 0.6 are: L=3, M= 5, a number of floating point operations of 106420351 FLOPS and a computation time of 4.0 seconds. In this case we have N_G = 21. In Figures 3 and 4 are shown the frequency magnitude responses for both cases.

As a matter of filter complexity comparison Table 1 shows the total number of products per output sample NPS_T, the total number of additions per out sample NAS_T, and the total delay units $NDel_T$ required for the proposed algorithm, and for algorithms proposed in (Gustafsson, Johansson & Wanhammar, 2001; Saramaki et al., 1988; Webb & Munson, 1996). Note that the smallest total number of products is obtained with the proposed algorithm using the sharpening polynomial with the parameters m= 2, n= 1. The method in (Saramaki et al., 1988) with three stages exhibits lower complexity than the proposed method with the sharpened polynomial m= n=1. However, the method (Saramaki et al., 1988) does not propose how to choose the design parameters.

FUTURE TRENDS

The use of IFIR filter design techniques is now a rather well-developed topic in the field of digital filters. Nonetheless, an analytical technique for obtaining a design employing the minimum number of filter-tap multipliers has not previously appeared and presents an important research task (Mehrnia & Willson, 2004). The future research also includes a multistage design and finding an optimal choice of the corresponding interpolation factors.

Another important issue that needs to be addressed is how to choose the sharpening polynomials (as a trade-off between the values of tangencies m and n and the number of cascaded basic RRS filters) that will yield an improved characteristic of the RRS filter and a decreased complexity of the overall design.

CONCLUSION

A direct algorithm to find the design parameters of an IFIR structure with a low complexity has been described, where the interpolation filter is a sharpened RRS filter. The proposal is based on an iterative design of the model filter $G(z)$ and the sharpened RRS filter until the filter desired specifications are met. As shown in the design example for a given sharpening polynomial, the proposed algorithm finds the maximum interpolator factor and the minimum number of RRS stages to meet desired specifications with small number of products per output sample. The proposed method is useful for narrowband FIR filter design.

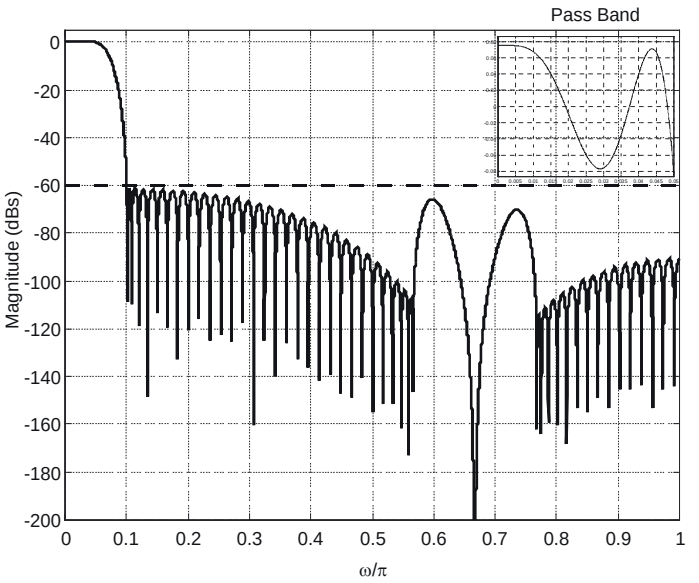
REFERENCES

Chen, J. C. H. & Vinod, A. P. (2006). Design of high-speed, low-power FIR filters with fine-grained cost metrics. In *Proceedings of the IEEE Conference APCCAS 2006, Singapore* (pp. 757-760).

Table 1. Filter complexity comparison between the described algorithm and other design methods

Reference	Method	NPS_T	NAS_T	$NDel_T$
(Saramaki et al., 1988)	1 stage IFIR	18	34	119
	3 stages IFIR	15	24	127
(Webb & Munson, 1996)	Pseudo IFIR	24	46	122
(Gustafsson et al., 2001)	Two stages	20	36	90
	Three stages	24	45	120
Described Algorithm	m=n=1	18	50	125
	m=2, n=1	11	55	198

Figure 3. Example magnitude response obtained with described algorithm with $\sigma = \delta = 0$, $m=n=1$



Diaz-Carmona, J., Jovanovic Dolecek, G., & Padilla, J. A. (2006). An algorithm for computing design parameters of IFIR filters with low complexity. *Journal Computacion Y Sistemas, 10*(2), 99-106.

Diniz, P. S. R., da Silva, E. A. B., & Netto, S. L. (2002). *Digital signal processing, system analysis and design*. Cambridge: Cambridge University Press.

Gustafsson, O., Johansson, H., & Wanhammar, L. (2001). Narrow-band and wide-band single filter frequency masking FIR filters. In *Proceeding of the IEEE Conference ISCAS 2001* (pp. 181-184). Sidney, Australia.

Hartnett, R. J. & Boudreaux, G. F. (1995, December). Improved filter sharpening. *IEEE Transactions on Signal Processing, 43*, 2805-2810.

Ifeachor, E. C. & Jervis, B. E. (2001). *Digital signal processing: A practical approach* (2nd ed.). NJ: Prentice Hall

Jovanovic-Dolecek, G. (2003). Design of narrowband Highpass FIR filters using sharpening RRS filter and IFIR structure. In J. Peckham & S. J. Lloyd (Eds.), *Practicing software engineering in the 21st century* (pp. 272-294). Hershey, PA: Idea Group Publishing.

Jovanovic Dolecek, G. & Diaz-Carmona, J. (2003). One method for design of narrowband lowpass filters. In J. Peckham & S. J. Lloyd (Eds.), *Practicing software engineering in the 21st century* (pp. 258-271). Hershey, PA: Idea Group Publishing.

Jovanovic Dolecek, G. & Diaz-Carmona, J. (2005). One method for design of narrowband low-pass filters. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology.* Hershey, PA: Idea Group Publishing.

Jovanovic Dolecek, G. & Mitra, S. K. (2007). Computationally efficient multiplier-free FIR filter design. *Computacion y Sistemas, 10*(3), 251-268.

Kaiser, J. F. & Hamming, R. W. (1984, October). Sharpening the response of a symmetric non-recursive filter by multiple use of the same filter. IEEE Transactions on Acoustics Speech and Signal Processing, ASSP-25, 415-422.

Lian, Y. & Yang, C. (2001). A new structure for design narrowband lowpass FIR filters. *IEEE Catalogue, No. 01CH37239* (pp. 274-277).

Mehrnia, A. & Willson, A. N., Jr. (2004). On optimum IFIR filter design. In *Proceedings of the IEEE Conference ISCAS 2004* (pp. 13-136). Vancouver, Canada,

Mitra, S. K. (2005). *Digital signal processing: A computer-based approach* (3rd ed.). New York: McGraw-Hill.

Milic, L. & Lutovac, M. (2002). Efficient multirate filtering. In G. Jovanovic-Dolecek (Ed.), *Multirate systems: Design and applications*. Hershey, PA: Idea Group Publishing.

Neuvo, Y., Cheng-Yu, D., & Mitra, S. K. (1984, June). Interpolated finite impulse response filters. *IEEE Transactions on Acoustics Speech and Signal Processing, 32*, 563-570.

Rodrigues, J. & Pai, K. R. (2005). New approach to the synthesis of sharp transition FIR digital filter. In *Proceedings of the IEEE International Conference ISIE 2005* (pp. 1171-1173). Dubrovnik, Croatia.

Samadi, S. (2000, October). Explicit formula for improved filter sharpening polynomial. *IEEE Transactions on Signal Processing, 9*, 2957-2959.

Saramaki, T. Y., Neuvo, S., & Mitra, S. K. (1988). Design of computational efficient interpolated FIR filters. *IEEE Trans. on Circuits and Syst., 35*, 70-88.

Smith, S. (2002). *Digital signal processing: A practical guide for engineers and scientists*. New York: Newnes.

Webb, J. & Munson, D. (1996). A new approach to designing computationally efficient interpolated FIR filters. *IEEE Transactions on Signal Processing, 44*, 1923-1931.

White, S. (2000). *Digital signal processing: A filtering approach*. Delmar Learning.

Yang, C. Z. & Lian, Y. (2003). Reduce the complexity of frequency-response masking filter using multiplication free filter. In *Proceedings of the IEEE Conference ISCAS 2003* (pp. 181-184).

Yang, C. Z. & Lian, Y. (2006). New structures for single filter based frequency-response masking approach. In *Proceedings of the IEEE Conference APCCAS 2006* (pp. 69-72). Singapore.

Zou, Y. & Saramaki, T. (2004). Design of computationally efficient narrowband linear-phase FIR filters. In *Proceedings of the IEEE International Conference MELECON 2004* (pp. 255-259). Dubrovnik, Croatia.

KEY TERMS

Digital Filter Design: It is carried out in three steps: Definition of filter specification, approximation of given specification, and implementation of digital filter in hardware or software.

Expanded Filter: Expanded filter $G(z^M)$ is obtained by replacing each delay z^{-1} in the filter $G(z)$ with M delays z^{-M}. In the time domain this is equivalent to inserting M-1 zeros between two consecutive samples of the impulse response of $G(z)$.

FLOPS: Measure of the computational complexity of an algorithm expressed in terms of Floating Point Operations per second.

Frequency Selective Filters: Digital filters which pass desired frequency components in a signal without distortion and attenuate other frequency components.

Interpolator: The filter $I(z)$ which is used to eliminate the unwanted spectrum introduced by expansion of the model filter in an IFIR structure.

Model or Shaping Filter: The filter $G(z)$ in the IFIR structure which has M times higher both the pass-band and the stop-band frequencies, than the prototype filter.

MPS: Measure of the computational complexity of a digital filter expressed in terms of Multipliers per Output Sample.

Pass-Band: The frequency range allowed to pass through the filter.

Pass-Band Ripple: The permitted deviation in the pass-band.

Stop-Band: The frequency band that is blocked by the filter.

Stop-Band Attenuation: The desired minimum attenuation in the stop-band.

Transition Band: The range of frequencies between the pass-band and the stop-band.

One Organization, One Strategy

Kevin Johnston
University of Cape Town, South Africa

INTRODUCTION

Most organizations have multiple levels of strategic plans (de Kluyver & Pearce, 2006), one of which is the Information Technology (IT) strategic plan. The alignment of an organization's business strategy with its IT strategy has been a concern of CIOs (Benson & Standing, 2008; Croteau & Bergeron, 2001; Johnston, Muganda, & Theys, 2007; Luftman Kempaiah, & Nash, 2006), CEOs (Armstrong, Chamberlain, Moore, & Hart, 2002; O'Brien & Marakas, 2006), academic researchers (Henderson & Venkatraman, 1999; Kangas, 2003; Pearlson & Saunders, 2004; Reich & Benbasat, 2000), and research companies (Broadbent, 2000; Croteau & Bergeron, 2001; Meta Group, 2001) since the age of vacuum tubes. The Society for Information Management (SIM) studies reveal that 'IT and Business Alignment' was the number one management concern in 2003, 2004 and 2005, and has been one of the top 10 concerns since 1983(Luftman et al., 2006).

IT and business strategies should not be separate or aligned; organizations should simply have one business strategy: one organization, one strategy.

BACKGROUND

Organizations need to recognize changing business climates, fluctuating resources, and the need to expand or grow (Bocij, Chaffey, Greasley, & Hickie, 2006). Organizations that plan, and then move in the right direction at the right time, survive. "Strategic planning enables a company to focus on what is important" (Benson & Standing, 2008, p. 207). All factors including IT must be considered and taken into account holistically to create value (de Kluyver & Pearce, 2006).

Key tasks of most managers within an organization are to acquire, develop, and allocate the organization's resources, and to develop and exploit the organization's capacity for innovation (Burgleman, Maidique, & Wheelwright, 2001). The acquisition, development, allocation, and exploitation of IT should be part of any business strategy. Many new products/services and ideas like e-commerce and BPR have been based on IT (Benson & Standing, 2008).

IT can contribute to the overall performance of the organization in many ways, including making/saving money, quality improvements, productivity gains, and providing new services/functions (McKeen & Smith, 2004). Strategic, management, operational, and functional support benefits can and do arise from the IT contribution (Ward & Daniel, 2006). It is unlikely that these contributions and possible innovations will occur by chance; they need to be planned.

O'Brien and Marakas (2007, p. 42) state that IT is more than a support for business, and that IT is "no longer an afterthought in forming business strategy." Linear planning is useless; organizations must plan holistically (Hartman, Sifonis, & Kador, 2000) or harmoniously in order to survive. Organizations should have a single strategic plan.

WHERE, WHY, WHAT, HOW?

Strategy is about creating options; strategic thinking focuses on taking different approaches, on choosing different sets of activities, on choosing a unique competitive position (de Kluyver & Pearce, 2006). Strategy is about choice, change and conclusions, where to do business (location and industry), why (reasons), what to do/offer, what not to do/offer, and how to do it.

Where should an organization do business, locally, nationally, or internationally, and in which industries? Organizations should be thinking in terms of where it makes good business sense, where the organization will survive.

A vision stating the goals of the organization (provides broad future focus) and which provides guidance and motivation needs to be developed (de Kluyver & Pearce, 2006) and communicated to all staff.

A business strategy needs to be defined which includes all the capabilities (forces/tools/resources) of an organization, so that approved plans may be executed as effectively as possible (Henderson & Venkatraman, 1999). Strategy articulates ways in which opportunities can be exploited using the organization's capabilities and resources (Burgleman et al., 2001). Strategy without capabilities is meaningless (Burgleman et al., 2001), and excluding the IT capability from the organizations strategy renders the strategy less effective at best. Similarly, having capabilities without strategy makes them aimless (Burgleman et al., 2001). The IT strategy and capability must therefore be an integral part of the overall strategy, or IT will become an aimless capability of the organization or at best will be run according to the CIO's aims. Managing the IT resource is a basic business function (Burgleman et al., 2001), which should be the responsibility of all managers within an organization.

Organizations create a strategy to anticipate change beyond the control of the organization, so changes within the organization (such as to business processes and organizational structure) can be initiated and controlled (Ivancevich & Matteson, 1999). A number of forces must fit together in a balanced way in order for an organization to function effectively. The framework developed by Scott Morton (1991) illustrates the interrelationship between strategy and four other forces, namely business processes, organizational structure, people, and IT. External environments (socioeconomic and technological) influence the organization. Changes in any one or more of these forces upsets the equilibrium of the organization (Turban, McLean, & Wetherbe, 2004); a change in any one force may require changes in some or all of the forces. Essentially, Scott Morton's framework implies that strategy should be developed in a holistic fashion, taking all forces into consideration. So IT (and business processes, organizational structure, people, and their roles) should not be merely aligned with the overall strategy or be there to support the overall strategy, they each contribute and form an essential part of one organizational strategy.

Organizations need to define exactly where they are going, why they wish to go there, what resources they will use to get there, and how they will utilize the resources to get there. Management needs to ensure that all stakeholders are aware of the plan.

IT STRATEGY

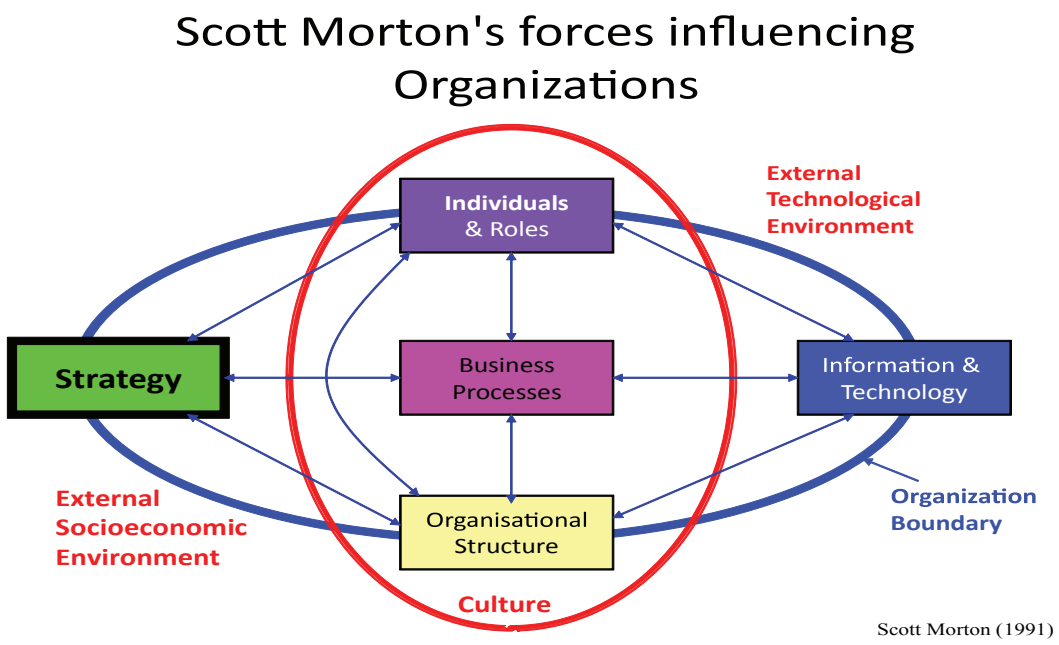
Several authors (Boddy, Boonstra, & Kennedy, 2005; Bocij et al., 2006; Burgleman et al., 2001; Croteau & Bergeron, 2001; McKeen & Smith, 2003; Reich & Benbasat, 2000; Ward & Daniel, 2006) agree that it is important to align IT strategy with the organization's business strategy. Although the importance of strategic alignment of IT is acknowledged and widely accepted, it remains an issue within many organizations (Armstrong et al., 2002; McKeen & Smith, 2003).

Gates (1999) wrote: "It is impossible to align IT strategy with business strategy if the CIO is out of the business loop." Lucas (2005) wrote that in many organizations the CIO is "kept in the dark about corporate strategy."

The lack of IT alignment with business can result in late market entry, lost market opportunities, or an unsustainable market advantage (Conarty, 1998) or business failure (Bocij et al., 2006).

Some authors (Benson & Standing, 2008; Bocij et al., 2006; Ward & Daniel, 2006) view IT strategy and business strategy as two distinct strategies, with IT strategy either supporting or influencing business strategy. Ward and Peppard (2002) suggested guidelines to align business and IT strategies. Huber, Piercy, and McKeown (2008) state that the different strategic views within an organization need to be interlinked. Pukszta (1999) stressed that IT strategy must

Figure 1



Scott Morton (1991)

be completely and seamlessly integrated (co-adapted) with business strategy at all organizational levels. O'Brien and Marakas (2007) compare the two approaches (alignment and co-adaptation).

IT and business strategies should not be aligned; organizations should have one harmonious strategy. IT strategy (and other strategies such as e-business, supply-chain, technology, etc.) must lose its distinctness; in this way it will gain prominence and exert greater influence (Pukszta, 1999; O'Brien & Marakas, 2007) within organizations. Each organization should have one harmonized strategy: one organization, one strategy.

IT planning should be in harmony, not merely aligned with business strategy.

CONFIGURATION

Organizations need to decide on internal structures, processes, and people that optimize the strategic plan. Structure, processes, and order are necessary for the survival of the organization.

"A gap has developed between the power and choice enjoyed by individuals as consumers and citizens on the one hand, and that available to them in the workspace on the other" (Chowdhury, 2000). This gap must be reduced. Employees have to be included in decisions regarding the strategy, structure, and business processes of organizations.

Organizational design and structure have always been important factors that influence the behavior of groups and individuals. It is through structure that management establishes expectations of achievements for individual employees and departments, and decides how the organization's strategy is to be measured. The purpose of structure is to regulate, or reduce, uncertainty in the behavior of employees (Ivancevich & Matteson, 1999). Where and how IT is placed in the organizational structure determine the role and influence of IT.

Each organization requires a structure, and when that is ignored, the organization will not be able to crawl, much less fly! No organization will last if everyone acts independently; a structure needs to be developed preferably with the employees, which will enhance the strategy and vision of the organization.

Employees cannot survive alone within an organization; they need to be in some sort of formation. Organizations only 'fly' when all the employees are in formation.

Organizations need tradition, ritual, and structure to retain their identity. A department in which the author worked had a daily meeting in which all employees met each other and offered encouragement, support, guidance, and feedback. The meetings were stand-up, 15- to 30-minute affairs, with a fixed regular program. On Mondays, projects and work for the week ahead were discussed; Tuesdays were thinking days and employees had to solve puzzles in groups; Wednesdays were learning days where one employee had to teach the others something; Thursdays were to announce and discuss change; and Fridays were for external focus or external speakers. Each employee had an opportunity to lead a Tuesday and Wednesday meeting. In organizations where there is encouragement, involvement, and participation, the production is much greater. Goleman (2007, p. 312) states that all people believe warm human relationships to be the core feature of 'optimal human existence'. Human relationships drive people's ability to be and work at optimal levels (Goleman, 2007). It is therefore imperative that organizations involve and include all employees in debates and decisions about strategy, IT, business processes, organizational structures, and people.

Organizations need to fly with employees who want to be in the organization, who feel part of the organization, who know where the organization is headed, and who want to fly in formation.

TIMING

Organizations understand that they cannot change or fight the seasons. Organizations need to develop strategies anticipating 'winter' and unseasonal changes. Organizations must develop rapid responses to external demands; be able to adapt their organizational structures, IT, and business processes; and involve their employees to achieve competitive advantage.

Organizations need to understand that IT change is only one factor that affects competitive position. In order to survive, organizations need to adapt to change in a timely manner. Therefore strategic planning must be a dynamic process, and IT and IT change must form part of the process, as must other changes such as people and structural changes. Successful organizations balance a well-defined business focus with the willingness, and the will, to undertake major and rapid change.

Organizations need to be ready to respond to continual changes. Prerequisites are leadership, governance, competencies, and technology (Hartman et al., 2000). The first prerequisite is leadership; outstanding companies are associated with their leaders (such as Welch, Gates, and Bezos). Leaders create a vision that is shared and accepted within the organization. Governance is the operating model that defines the organization. The formation or structure of the organization must be clear. People's roles, responsibilities, and authority levels must be defined. Organizations need to have methods for assessing, selecting, allocating, and monitoring resources. Competencies are the ways in which the organization responds to change, exploits available resources and opportunities, and accommodates reality. Technology needs to be robust and comprehensive.

"The reality of a strategy lies in its enactment, not in those pronouncements that appear to assert it" (Burgleman et al., 2001). Strategic intent needs to be converted into strategic action to be meaningful.

Organizations need strategies to deal with obstacles, and may in some cases need to revise their strategies in order to be successful.

SUCCESS

The organization must expect some casualties, but if the organization has survived and looks likely to continue surviving, the strategy was successful. To succeed, the energy, creativity, and resources of the organization must have been used. If there are parts of the organization that have not being used, the question must arise, why are they in the organization?

Burgleman et al. (2001) ask and answer the question: "What strategies, policies, practices, and decisions result in successful management of high-technology enterprises?" Six themes of success are listed: (1) business focus, (2) adaptability, (3) organizational cohesion, (4) entrepreneurial culture, (5) sense of integrity, and (6) hands-on top management. All are controlled or influenced by the organizational strategy. Unless each and every one of the organizations' resources are in harmony, an organization cannot have succeeded. If IT strategy and all other strategies are harmonized into a single business strategy, then the organization can claim to have a holistic business focus, to have organizational cohesion, and to have a sense of integrity. A disharmonized strategy certainly cannot be regarded as being honest, fair, or open—other attributes of integrity.

FUTURE TRENDS

Organizations will develop a single evolving harmonized strategy, to which all employees have contributed. The organizational strategy will be a continual, changing, probing, never-ending cycle involving all in the organization. Organizations will have one integrated strategy that has involved all employees.

CONCLUSION

Organizational success must include the ability to harmonize the organization and mobilize the workforce (Reich & Benbasat, 2000). Organizations and individuals can only realize their potential for greatness and goodness when they join the flock, fly in formation, and contribute something for the common good. Business and IT strategies will be in harmony when IT is seen as contributing positively to the organization's business strategy. IT strategy must form a harmonized part of the organization's overall strategy (Ward & Peppard, 2002).

REFERENCES

Armstrong, T., Chamberlain, G., Moore, B., & Hart, M. (2002). *Key information systems management issues for CEOs and other executives in South Africa 2002*. Unpublished Honours Empirical Research, University of Cape Town, South Africa.

Benson, S., & Standing, C. (2008). *Information systems—a business approach* (3rd ed.). Australia: John Wiley & Sons.

Bocij, P., Chaffey, D., Greasley, A., & Hickie, S. (2006). *Business information systems, technology, development and management for the e-business.* Pearson Education, Harlow.

Boddy, D., Boonstra, A., & Kennedy, G. (2005). Managing information systems: An organisational perspective. Pearson Education, Harlow.

Broadbent, M. (2000). *Today's CIO energizes, enables, executes and exploits*. Retrieved March 2002 from http://www4.gartner.com/UnrecognizedUserHomePage.jsp

Burgleman, R.A., Maidique, M.A., & Wheelwright, S.C. (2001). *Strategic management of technology and innovation.* Singapore: McGraw-Hill.

Chowdhury, S. (Ed.). (2000). *Management 21C.* London: Prentice Hall.

Conarty, T.J. (1998). *Alignment for success: Information technology and business strategy.* Retrieved March 2002 from http://www.worldsteel.org/events/proceed/IISI-32_1998/PR_conarty1.html

Croteau, A., & Bergeron, F. (2001). An information technology trilogy: Business strategy, technological deployment and organisational performance. *Journal of Strategic Information Systems, 10*(2), 77-99.

De Kluyver, C.A., & Pearce, J.A. II. (2006). *Strategy: A view from the top (an executive perspective).* Englewood Cliffs, NJ: Pearson Education Limited/Pearson Prentice Hall.

Gates, B. (1999). *Business @ the speed of thought.* London: Penguin.

Goleman, D. (2007). *Social intelligence.* London: Arrow Books.

Hartman, A., Sifonis, J., & Kador, J. (2000). *Net ready.* New York: McGraw-Hill.

Henderson, J.C., & Venkatraman, N. (1999). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal, 38*(2-3), 472-484.

Huber, M.W., Piercy, C.A., & McKeown, P.G. (2008). *Information systems, creating business value.* New York: John Wiley & Sons.

Ivancevich, J.M., & Matteson, M.T. (1999). *Organisational behaviour and management.* Singapore: McGraw-Hill.

Johnston, K.A., Muganda, N., & Theys, T. (2007). Key issues for CIOs in South Africa. *Electronic Journal of Information Systems in Developing Countries, 30*(3), 1-11.

Kangas, K. (Ed.). (2003). *Business strategies for information technology management.* Hershey, PA: IRM Press.

Lucas, H.C. Jr. (2005). *Information technology strategic decision making for managers.* London: John Wiley & Sons.

Luftman, J., Kempaiah, R., & Nash, E. (2006, June). Key issues for IT executives 2005. *MIS Quarterly Executive, 4*(2).

Meta Group. (2001). *Top CIO issues for 2001.* Retrieved from http://www.metagroup.com/cgi-bin/inetcgi/search/displayArticle.jsp?oid=23211

McKeen, J.D., & Smith, H.A. (2004). *Making IT happen, critical issues in IT management.* New York: John Wiley & Sons.

O'Brien, J.A., & Marakas, G.M. (2006). *Management information systems* (7th ed.). New York: McGraw-Hill.

O'Brien, J.A., & Marakas, G.M. (2007). *Enterprise information systems* (13th ed.). New York: McGraw-Hill.

Pearlson, K.E., & Saunders, C.S. (2004). *Managing and using information systems—a strategic approach.* New York: John Wiley & Sons.

Pukszta, H. (1999). Don't split IT strategy from business strategy. *Computerworld, 33*(2), 35.

Reich, B., & Benbasat, I. (2000). Factors that influence the social dimension of alignment between business and information technology objectives. *MIS Quarterly, 24*(1), 81-113.

Scott Morton, M.S. (Ed.). (1991). *The corporation of the 1990s. Information technology and organizational transformation.* Oxford: Oxford University Press.

Turban, E., McLean, E., & Wetherbe, J. (2004). *Information technology for management* (4th ed.). New York: John Wiley & Sons.

Ward, J., & Daniel, E. (2006). *Benefits management. Delivering value from IS & IT investments.* London: John Wiley & Sons.

Ward, J., & Peppard, J. (2002). *Strategic planning for information systems.* London: John Wiley & Sons.

KEY TERMS

Alignment: The arrangement or position of different, separate elements (strategies) in relation to each other.

Business Process: A collection of business activities that take several inputs and create one or more outputs.

Business Strategy: A description of the plans, actions, or steps an organization intends to take in order to strengthen and grow itself.

CIO (Chief Information Officer): The head of the IS department in an organization.

Harmony: A pleasing combination of elements in a whole. The combination of elements intended to form a connected whole, as opposed to alignment where the elements remain separate.

IT (Information Technology): A collection of all systems in an organization**.**

IT Strategy: A description of the plans, actions, or steps an organization intends to take in order to make the best use of IT within itself.

Strategy: Plans to create and manage change, and exploit opportunities.

Online Communities and Community Building

Martin C. Kindsmüller
Berlin University of Technology, Germany

Sandro Leuchter
Berlin University of Technology, Germany

Leon Urbas
Berlin University of Technology, Germany

INTRODUCTION

"Online community" is one of today's buzzwords. Even though superficially it is not hard to understand, the term has become somewhat vague while being extensively used within the e-commerce business. Within this article, we refer to online community as being a voluntary group of users who partake actively in a certain computer-mediated service. The term "online community" is preferred over the term "virtual community," as it denotes the character of the community more accurately: community members are interacting online as opposed to face to face. Furthermore, the term "virtual community" seems too unspecific, because it includes other communities that only exist virtually, whereas an online community in our definition is always a real community in the sense that community members know that they are part of the community.

Nevertheless, there are other reasonable definitions of online community. An early and most influencing characterization (which unfortunately utilizes the term "virtual community") was coined by Howard Rheingold (1994), who wrote: "…virtual communities are cultural aggregations that emerge when enough people bump into each other often enough in cyberspace. A virtual community is a group of people […] who exchanges words and ideas through the mediation of computer bulletin boards and networks" (p. 57). A more elaborated and technical definition of online community was given by Jenny Preece (2000), which since then, has been a benchmark for developers. She stated that an online community consists of four basic constituents (Preece, 2000, p. 3):

1. Socially interacting people striving to satisfy their own needs.
2. A shared purpose, such as interest or need that provides a reason to cooperate.
3. Policies in the form of tacit assumptions, rituals, or rules that guide the community members' behavior.
4. A technical system that works as a carrier that mediates social interaction.

Not explicitly mentioned in this characterization but nevertheless crucial for our aforementioned definition (and not in opposition to Preece's position) is voluntary engagement.

BACKGROUND

Just because everybody is now talking about them, online communities are, historically seen, neither an implication of the World Wide Web — which dates back to 1991 (Berners-Lee et al., 1992) — nor dependent on the Internet as a transport infrastructure. In fact, online communities emerged at times when ARPAnet—the predecessor of the Internet — was still restricted to military-funded institutions. They were based on computerized bulletin boards first introduced by Christensen and Suess (1978). Their system was called CBBS (computerized bulletin board system) and followed the idea of a thumbtack bulletin board hosted electronically on a computer. Other computer hobbyists were able to connect with their home computers via a dial-up modem connection and could "pin" messages to a shared "board." The first online communities developed through other participants responding to those messages, creating ongoing discussions. At that time, computer hobbyists and scientists were more or less the only ones who owned computers and modems. Therefore, most topics on CBBS were within the realm of computers, but in the long run, the discussions broaden. Within the 1980s, similar systems appeared that were now subsumed as BBS (bulletin board system). The most well known were "The Well" (Whole Earth 'Lectronic Link) and FidoNet (Rheingold, 2000).

Apparently, at the very same time, the technological and social environment was ready for online communities, as there were at least two other independent developments concerning this matter:

1. The Usenet was invented by computer science students at Duke University and the University of North

Carolina, using a simple scheme by which these two computer communities could automatically exchange information via modems at regular intervals.

2. The first MUDs appeared at the University of Essex (UK) creating playful and imaginative online communities. MUDs (short for Multi-User Dungeon/Dimension/Domain) are computer-implemented versions of text-based role-playing games, in which multiple persons can take virtual identities and interact with one another. Early MUDs were adventure games played within old castles with hidden rooms, trapdoors, etc.

Nowadays, most online communities are using the Internet as carrier, and most of them are Web based, using HTTP as a protocol for transportation and the DHTML standard for presentation. But there are still communities that employ other systems and protocols, like newsreaders using NNTP and mail-groups using SMTP or IRC (Internet relay chat) based chatting systems (IRC). Some online communities even use multiple systems and protocols to communicate and cooperate.

ONLINE COMMUNITIES

The conditions in pure online communities highly differ from a computer-mediated communication situation within a company. Whereas employees in a computer-supported cooperative work (CSCW) context usually meet online as well as face-to-face, members of online communities have, as a general rule, never met each other. Working in a highly standardized company context, employees have to focus on task fulfillment within a certain time frame. Superiors evaluate their achievements, and they are accordingly paid by the company. Online communities thrive on volunteers. Usually none of the community members can be forced to do something, and there are no tangible incentives. Basic research in motivation psychology (Franken, 2001) even shows that incentives tend to be counterproductive.

Community members usually show a high degree of intrinsic motivation to participate actively in the development of an online community. It is still open to discussion where this motivation comes from. Simple rules like "It's all based on trying to maximize the potential personal benefit" seem to fail, as long as one has a simplistic concept of the term "personal benefit." As the attention-economy-debate (i.e., Aigrain, 1997; Ghosh, 1997; Goldhaber, 1997) shows that personal benefit is a complex entity if one relates it to online activities in the World Wide Web.

The likelihood of taking an active part in a community increases with the potential personal benefit that could be gained within that community. This is directly related to the quality of the contents offered. As, e.g., Utz (2000) stated, the likelihood of submitting high quality contributions increases with the quality and the manifoldness of the existing entries. Appropriate solutions of these quality assurance aspects are rating systems.

A "killer-feature" for such an application generates immediate benefit for a user as soon as he or she contributes to the community, even without anybody else contributing. In addition to such a feature, or even as a partial replacement, one can follow best practices. After analyzing numerous well-working online communities, Kollock (1999) came to the conclusion that there are basically two states of motivation: self-interest (what seems to be the common motivation found) and altruism. Self-interest as a motivational state is linked to expectation of reciprocity: people are willing to help or cooperate with others if they can expect a future quid pro quo.

A widely discussed issue in the context of community building is the so-called public goods dilemma: if people can access public goods without restriction, they tend to benefit from these goods and, therefore, from others' contributions without contributing in the same way. If, on the other hand, most members of a community are led into temptation, the public good will vanish (Kollock & Smith, 1996). The main problem is to keep the balance between the individual and common interest: an individually favorable and reasonable behavior turns out to be harmful for the others, and in the long run, disastrous for the community (Axelrod, 1984; Ostrom, 1990).

Owing to these circumstances, it is not surprising that a great deal of all online community building projects fail, even though much effort has been put into these projects due to the high profit opportunities within the field as, for instance, Hagel and Armstrong (1997) predicted.

ONLINE COMMUNITY BUILDING

Recipe-based fabrication of online communities is, at least, a bold venture if not an illusionary enterprise. Social relationships and group momentum are particularly hard to predict. As Rheingold (2000) explicated, online communities grow organically and tend to follow their own rules. Therefore, controlling efforts always have to be adjusted to the current group context. Nevertheless, some well-approved principles could be derived from findings that were discussed in the last paragraph.

According to Kollock (1999), cooperation within an online community can only be successful if individuals:

1. Can recognize each other, i.e., they are not operating anonymously within the community.
2. Have access to each others interaction history.
3. Share the presumption of a high likelihood of a future encounter within the online community.

This leads to the conclusion that online communities have to offer possibilities of creating and managing relationships by supporting continuous interaction between their members. Therefore, it is advantageous if the system has a memory, in the sense that every community member and every item stored in the system holds a personal history.

People tend to act from self-interest if they are aware that their actions have effects on their reputations: high-quality contributions, impressive knowledge, and the perception of being willing to help others enhance the prestige of the community member. Although altruism as a motivational state for taking part in an online community is less common in comparison with self-interest, it is still frequent enough to be addressed if one thinks about community building. People with altruistic motivation try to meet the needs of the group or certain group members. This motivational state can be satisfied by establishing a public space where these needs can be stated, communicated, and discussed.

Considering the public goods dilemma, it is essential to introduce a role concept to clearly communicate the borderline between being in a group and out of a group. To get full access to all group resources, one has to join the group. Several functionalities are only accessible for registered and authorized users. The commitment that is required to join the group leads to a comprehensible demarcation between members and nonmembers, who, in turn, facilitate the togetherness of the group and the identification of the members within the group. Three further selective measures address the public goods dilemma: personal presence, personal responsibility, and personal history. Anonymity and lack of continuity among the members promotes egoistic behavior. Therefore, modifying actions should be tagged with users' login names, which, in turn, should be associated with records of personal data. Tagging entries and actions with user login names makes it easy to recognize people and enhances the constitution of personal relationships and online cooperation among the community members. Seeing all modifying actions supports the actors' personal responsibility. If the system has a memory of every action, this gives a personal history to every community member, as well as to the community's artifacts. Information about past interactions of the members again increases personal responsibility, whereas information about interaction with the artifacts facilitates getting up-to-date within a new area of interest and familiarizing new members with the online community's etiquette and who-is-who.

"Content is king" is commonplace for virtually all Web-based efforts. This is notably true for online communities operating on a user-as-editors base. To implement a reasonable quality assurance system, it is crucial to apply technical as well as social means. Technically, this can be done by employing a content rating system. Employing a "tiger team" of highly motivated volunteers can, on the other hand, help the online community to start up by delivering good content. For all content producers, it has to be as easy as possible to feed new content into the system. The best way of avoiding barriers is through continuous usability testing.

Introducing dedicated and active moderators seems to be the most important step to nourish motivation of the community members. Moderators can enhance group activities and increase the efficiency of the group. They are responsible for communicating the group standards (etiquette), acting as confessors for new community members, and helping in preserving continuity. Rojo (1995) showed that an active moderator can, to some extent, compensate for the lack of active members in an online community.

As opposed to face-to-face communities, where only members can start a conversation, online communities can initiate communication. This opportunity should be brought into play by implementing change awareness functions as software agents that collect relevant information for users and present it in e-mails and personalized portal pages. An important item in using profiles for personalized services is keeping profiles up-to-date. Experience shows that interests continuously change over time. Profile setting dialogues are often accessed once and then forgotten. Thus, there is risk for personalized services to decrease subjectively in quality over time. Hence, it is important to monitor user behavior and let the agents ask from time to time if interpretations of observations of, for example, changing interests, are correct.

The open-source movement has become very successful in recruiting new developers who start their own projects or join existing software development efforts. Today, most if not all open-source software development communities use online services for cooperation. In the remainder of this section, the application of the requirements for software supporting online community building is demonstrated with examples of open-source software development communities:

- **"Killer-feature":** Open-source projects are often founded to solve an existing problem, i.e., the killer-feature is the product of the online community. When others join the project, the work becomes even more effective.
- **Recruitment:** Open-source communities produce software systems that are not only intended for use by themselves but also for external clients. These external users, i.e., users not actively involved in the open-source online community, can be made active developers that modify the source and give it back to the project. To foster this process of developer recruitment, online communities should provide transparent rules for becoming actively involved. The projects that do not seem to be hermetic have better chances of growing their developer base.

- **Transparency:** The most important possibility of gaining transparency is through a public archive of the project's mailing lists. Because it is often not easy to scan large e-mail archives, open-source communities should provide text documenting guidelines and standards.
- **Policy:** The Debian community that maintains an open-source Linux distribution is a good example of a growing community that has given itself a substantial set of roles, rules, and guidelines. They achieve transparency of their standards by publication of documents on their Web server — other projects, such as UserLinux, use a Wiki for that purpose which makes such standards more vivid and activates community members' attendance.
- **Trust:** Debian has a twofold quality assurance system. There are no anonymous additions to the system, and approved maintainers electronically sign all modifications. Bugs are reported by all users. The bug lists are available to the public.
- **Cooperation and usability:** Cvs is a configuration management software for source code trees in distributed development teams. cvs is an example of cooperation software and its usability issue: Based on rcs (revisions management for single files), it is efficient to use for everyday tasks in open-source software development.
- **Awareness:** Workflows can provide awareness. Examples include automated e-mail distribution of users' bug reports and e-mail notifications of cvs commits.

This exploration into open-source projects and their online coordination and cooperation tools reveals that a voluntary community approach works, and the infrastructures and supporting tools of these projects can be taken as a best practice reference case.

FUTURE TRENDS

Recently, the term "socialware" was proposed for software systems dedicated to enhance social relations. According to Hattori et al. (1999), socialware are systems which aim to support various social activities by means of computer networks. This is done by linking people with others, fostering communication within a community and connecting the community's information. Initially intended for CSCW systems that are used by stable communities, the socialware approach seems suitable for implementing software for online communities as well. It uses rules of interpersonal communication and transfers these structures into software. The technical concept associated with socialware is a multiagent system architecture. The CSCW functionality is achieved through coordination and cooperation of a distributed set of software entities (agents). Users of a community system have personal agents for gathering and exchanging information, visualizing contexts, and supporting decisions. Personal agents and the users they belong to are seen as personal units. Personal units interact with community agents that have the function of providing shared information and mediating communication between personal units. This approach also makes it possible to link several partially overlapping online communities.

A current development in online communities is the transformation of the virtuality of computer networks into the real world. There are different enabling technologies for mobile and ad hoc communities. An important factor is localizability in cellular phone networks or with global positioning systems (GPSs). Using the localization information as part of the application context allows for mobile communities. They are often based on asynchronous communication, like Internet online communities. An example for such a mobile community is the petrol station price comparison community. In 2000, the German Research Center for Information Technology offered car drivers a location aware service for obtaining the locations of petrol stations together with their prices, in exchange for other stations' current petrol prices.

The availability of new short-range radio networking technologies, such as Bluetooth or wireless LAN, enables new synchronous mobile communities. This gives users the ability to connect devices ad hoc (i.e., without a server infrastructure), permitting mobility and interaction. As with other Internet online communities, besides professional applications such as disaster management (Meissner et al., 2002), game playing is an important technology driver, e.g., pervasive group games are being developed (Pennanen & Keinänen, 2004) that could build up social structures in some ways comparable to online communities.

CONCLUSION

Advanced software solutions like the aforementioned socialware approach can help to build and maintain stable online communities. In the long run, though, it is not the technology, it is the people that make an online community work. The most advanced technology is neither sufficient nor, as early BBS/MUD approaches show, necessary for stable online communities. People will always make creative use of technology by using it in other ways than were originally intended by the designers. This will, once in a while, generate possibilities for new online communities.

Nevertheless, the most important factor for successful online communities is providing awareness about changes in the communities' databases to members. Awareness functions provide an understanding of the others' activities and

the communities' goals and progress to relate and evaluate the users' own activities accordingly (Dourish & Bellotti, 1992).

REFERENCES

Aigrain, P. (1997). Attention, media, value and economics. *First Monday, 2*(9). Retrieved March 15, 2004, from http://www.firstmonday.dk/issues/issue2_9/aigrain/

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Berners-Lee, T. J., Cailliau, R., Groff, J. -F., & Pollermann, B. (1992). World-Wide Web: The information universe. *Electronic Networking: Research, Applications and Policy, 2*(1), 52–58.

Christensen, W., & Suess, R. (1978). Hobbyist computerized bulletin board. *Byte Magazine, 3*(11), 150–158.

Dourish, P., & Bellotti, V. (1992). Awareness and coordination in shared work spaces. *In Proceedings ACM Conference on Computer-Supported Cooperative Work CSCW'92* (pp. 107–114), Toronto, Canada.

Franken, R. E. (2001). *Human motivation* (5th ed.). Pacific Grove, CA: Brooks/Cole.

Ghosh, R. A. (1997). Economics is dead. Long live economics! A commentary on Michael Goldhaber's "The Attention Economy." *First Monday, 2*(5). Retrieved March 15, 2004, from http://www.firstmonday.dk/issues/issue2_5/ghosh/

Goldhaber, M. H. (1997). The attention economy and the Net. *First Monday, 2*(4). Retrieved March 15, 2004, from http://www.firstmonday.dk/issues/issue2_4/goldhaber/

Hagel, J., & Armstrong, A. G. (1997). *Net gain: Expanding markets through virtual communities*. Boston, MA: Harvard Business School Press.

Hattori, F., Ohguro, T., Yokoo, M., Matsubara, S., & Yoshida, S. (1999). Socialware: Multi-agent systems for supporting network communities. *Communication of the ACM, 42*(3), 55–61.

Kollock, P. (1999). The economies of online cooperation. Gifts and public goods in cyberspace. In M. A. Smith & P. Kollock (Eds.), *Communities in cyberspace*. London, UK: Routledge.

Kollok, P., & Smith, M. A. (1996). Managing the virtual commons: Cooperation and conflict in computer communities. In S. Herring (Hrsg.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 109–128). Amsterdam, The Netherlands: John Benjamins.

Meissner, A., Luckenbach, T., Risse, T., Kirste, T., & Kirchner, H. (2002). Design challenges for an integrated disaster management communication and information system. *First IEEE Workshop on Disaster Recovery Networks (DIREN 2002),* June 24, 2002, New York City. Retrieved October, 04, 2004 from: *http://comet.columbia.edu/~aurel/workshops/diren02/IEEE_DIREN2002_Meissner_DesignChallenges.pdf*

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.

Pennanen, M., & Keinänen, K. (2004). Mobile gaming with peer-to-peer facilities. *ERCIM News, 57*, 31–32.

Preece, J. (2000). *Online communities: Designing usability and supporting sociability*. Chichester, UK: John Wiley & Sons.

Rheingold, H. (1994). A slice of life in my virtual community. In L. M. Harasim (Ed.), *Global networks: Computers and international communication* (pp. 57–80). Cambridge, MA: MIT Press.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier* (revised edition). Cambridge, MA: MIT Press.

Rojo, A. (1995). *Participation in scholarly electronic forums*. Unpublished Ph.D. thesis, Ontario Institute for Studies in Education, University of Toronto, Canada. Retrieved March 14, 2004, from http://www.digitaltempo.com/e-forums/tabcont.html

Utz, S. (2000). Identifikation mit virtuellen Arbeitsgruppen und Organisationen. In M. Boos, K. J. Jonas, & K. Sassenberg (Eds.), *Computervermittelte Kommunikation in Organisationen*. Göttingen: Hogrefe.

KEY TERMS

Community Building: All activities related to building and maintaining online communities.

CSCW (Computer-Supported Cooperative Work): Software tools and technology as well as organizational structures that support groups of people (typically from different sites) working together on a common project.

Online Community: An online community is a voluntary group of active users that partake actively in a certain computer-mediated service.

Socialware: Socialware aims to support various social activities on a network. Rules of interpersonal communication are used and transferred into community software.

UaE (User-As-Editors) Approach: The community members are responsible for supplying new content and for the quality assurance of existing content, as well as for creating and maintaining the etiquette of the community.

Virtual Community: A featureless and, therefore, often misleading term usually regarded as synonymous to online community. The term "online community" is preferable, as it denotes the character of the community more accurately.

Wiki: Internet service based on HTTP and HTML providing "open editing" of Web pages with a Web browser. Hyperlinks between documents are supported with simple textual references. By default, everybody is allowed to edit all available pages.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2203-2208, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Online Communities and Online Community Building

Martin C. Kindsmüller
University of Lübeck, Germany

André Melzer
University of Lübeck, Germany

Tilo Mentler
University of Lübeck, Germany

INTRODUCTION

In this article, we define and describe the concept of online communities, outline the essential conditions under which they emerge and present some means that foster the building of online communities.

"Online community" is one of the buzzwords in the age of Web 2.0. Within this article, we refer to online community as a voluntary group of users who partake actively in a certain computer-mediated service. The term "online community" is preferred over the term "virtual community," as it denotes the character of the community more accurately: community members are interacting online as opposed to face-to-face. Furthermore, the term "virtual community" seems too unspecific, because it includes other communities that only exist virtually, whereas, an online community in our definition is always a real community in the sense that community members know that they are a part of their community.

Nevertheless, there are other reasonable definitions of online community. An early and most influencing characterization (which unfortunately utilizes the term "virtual community") was coined by Howard Rheingold (1994). He wrote: "…virtual communities are cultural aggregations that emerge when enough people bump into each other often enough in cyberspace. A virtual community is a group of people […] who exchanges words and ideas through the mediation of computer bulletin boards and networks" (p. 57). A more elaborate and technical definition of online community is given by Jenny Preece (2000), which acts as a benchmark for developers since then. She states that an online community consists of four basic constituents (Preece, 2000, p. 3):

- Socially interacting people striving to satisfy their own needs;
- A shared purpose like an interest or need that provides a reason to cooperate;
- Policies in the form of tacit assumptions, rituals, or rules that guide the community members' behavior; and
- A technical system that works as a carrier that mediates social interaction.

Not explicitly mentioned in this characterization, but nevertheless crucial for our aforementioned definition (and not in opposition to Preece's position), is voluntary engagement (see also Janneck, Finck, & Oberquelle, 2005).

As Preece's (2000) definition indicates, the basic constituents of online communities include individual issues, group-related issues, as well as technology-related issues. Online communities thus comprise the participants' basic individual motivation, the social interaction processes entailed to "bundle" individual needs to increase efficiency, and the implemented technical functions that support these processes.

In the light of the aforementioned role of social processes, it is not surprising that, with respect to online communities, findings from voluntary groups of active user communities outside computer-based systems are also a highly relevant source of information (see e.g., Baumeister & Bushman, 2008). In the section devoted to online community building, we will present Kraut's (2003) suggestion of a highly-sophisticated application of social psychology theory to address some well-known problems in online communities.

BACKGROUND

Just because everybody is now talking about them, online communities are historically seen neither as a repercussion of the World Wide Web—which dates back to 1991 (Berners-Lee, Cailliau, Groff, & Pollermann, 1992)—nor as dependent on the Internet as a transport infrastructure. In fact, online communities emerged at the time when ARPAnet—the predecessor of the Internet—was still restricted to military-

funded institutions. Some of these online communities were based on computerized bulletin boards first introduced by Christensen and Suess (1978). Their system was called CBBS (computerized bulletin board system) and followed the idea of a thumbtack bulletin board hosted electronically on a computer. Other computer hobbyists were able to connect with their home computers via a dial-up modem connection and could "pin" messages to a shared "board." The first online communities developed when other participants responded to those messages and created ongoing discussions. At that time, computer hobbyists and scientists were more or less the only ones who owned computers and modems. Therefore, most topics on CBBS were within the realm of computers, but in the long run, the topics of discussion broadened. By the 1980s, similar systems appeared that were now called BBS (bulletin board system). The most well known BBSs were "The Well" (Whole Earth 'Lectronic Link) and FidoNet (Rheingold, 2000).

Apparently, at the very same point in time, the technological and social environment was ready for online communities, as there were at least two other independent developments emerging:

1. The Usenet was invented by computer science students at Duke University and the University of North Carolina. They used a simple scheme by which these two computer communities could automatically exchange information via modems at regular intervals.
2. The first MUDs appeared at the University of Essex (UK) creating playful and imaginative online communities. MUDs (Multi-User Dungeon/Dimension/Domain) are computer-implemented versions of text-based role-playing games, in which multiple gamers can take virtual identities and interact with one another. Early MUDs were adventure games played in a labyrinth of dark dungeons with hidden rooms, trapdoors, and so forth.

Nowadays, most online communities are using the Internet as a carrier. Most of them are Web-based, using HTTP as a protocol for transportation and a combination of XHTML, CSS and JavaScript for presentation. But there are still communities that employ other systems and protocols, like newsreaders using NNTP and mail-groups using SMTP- or IRC- (Internet relay chat) based chatting systems. Some online communities even use multiple systems and protocols to communicate and cooperate.

A multiple group of new Web-based services like instant messaging, forums, chats, Web logs (or blogs), wikis, social bookmarking services and several types of other sharing services (e.g., for photos, videos, audio-files, or files in general) has recently been developed. Some of these services like instant messaging, forums or chats are typical applications within the field of computer-mediated communication and therefore foster online communities. Other types of services like, for example blogs, are at first sight not made to be platforms to house online communities. But as soon as these services are enriched with comment functions, RSS feeds and linkbacks (linkbacks are means to obtain notifications when other documents are linked to a certain document) they can be used as such. The latest developments are platforms like Facebook or MySpace, often summarized under the somewhat vague label Web2.0. They typically combine several of the aforementioned services to create rich communication media that could be used by online communities.

ONLINE COMMUNITIES

The conditions in pure online communities highly differ from a computer-mediated communication situation within companies and corporations. Whereas employees in a computer-supported cooperative work (CSCW) context usually meet online as well as face-to-face, members of online communities have, as a general rule, never met each other. Working in a highly standardized company context, employees have to focus on task fulfillment within a certain timeframe. Superiors evaluate their achievements, and they are accordingly paid by the company.

Online communities live from their volunteers. Usually none of the community members can be forced to do something, and there are no tangible incentives. Basic research in motivation psychology (Franken, 2001) even shows that incentives tend to be counterproductive.

Community members usually show a high degree of intrinsic motivation to participate actively in the development of an online community. It is still open to discussion where this motivation comes from. Simple rules like "It's all based on trying to maximize the potential personal benefit" seem to fail, if the concept of the term "personal benefit" is too simplistic. The attention-economy-debate (e.g., Aigrain, 1997; Ghosh, 1997; Goldhaber, 1997) shows that personal benefit is a complex entity if one relates it to online activities in the World Wide Web.

The likelihood of taking an active part in a community increases with the potential personal benefit that could be gained within that community. This is directly related to the quality of the contents offered. As Utz (2000) stated, the likelihood of submitting high quality contributions increases with the quality and the manifoldness of the existing entries. Appropriate solutions for quality assurance are rating systems.

A "killer feature" for such an application generates immediate benefit for users as soon as they start using the application, even without anybody else contributing. Unfortunately, this kind of feature can't always be found and implemented. As a (partial) replacement for such a feature, one can follow best practices. After analyzing numerous

popular online communities, Kollock (1999) came to the conclusion that there are basically two sources of motivation: self-interest (what seems to be the most common motivation) and altruism. Self-interest as a motivator is linked to expectations of reciprocity: people are willing to help or cooperate with others if they can expect a future quid pro quo. Altruistic behavior, in contrast, denotes people's motivation to increase another's welfare without expecting anything in return (Baumeister & Bushman, 2008).

A widely discussed issue in the context of community building is the so-called public goods dilemma: if people can access public goods without restriction, they tend to benefit from these goods and, therefore, from others' contributions without contributing reciprocally. If the majority of community members are tempted to behave that way, the public good will vanish (Kollock & Smith, 1996). The main problem is to keep the balance between the individual and common interest: An individually favorable and reasonable behavior turns out to be harmful for the others, and in the long run, disastrous for the community (Axelrod, 1984; Ostrom, 1990).

Owing to these circumstances, it is not surprising that a great deal of all online community building projects have failed, even though much effort has been put into these projects due to the high profit opportunities within the field as, for instance, Hagel and Armstrong (1997) predicted.

ONLINE COMMUNITY BUILDING

Recipe-based fabrication of online communities is, at least, a bold venture if not an illusionary enterprise. Social relationships and group momentum are particularly hard to predict. As Rheingold (2000) explicated, online communities grow organically and tend to follow their own rules. Therefore, controlling efforts always have to be readjusted to the current group context and dynamics. Nevertheless, some well-approved principles could be derived from the findings that were discussed in the last chapter.

Kim (2000) presents a membership lifecycle which describes five successive stages and levels of participation:

1. Visitors (people not involved in the community processes);
2. Novices (new community members who are still trying to find their way);
3. Regulars (community members who are consistently involved in the community life);
4. Leaders (community members who keep the community running and bear responsibility as well as acquired rights); and
5. Elders (long-time community members who share their knowledge and communicate the community culture).

We have already stressed the importance of findings of "off-line" groups for online communities. In this respect, Kraut (2003) suggested applying social psychology theory to some of the well-known problems existing in online groups. In particular, Kraut addresses the problem of under-contribution in groups. This issue refers to a common characteristic of online groups, namely their highly uneven distribution of contributions with a small number of members contributing most of the content, and the majority of members acting as so-called lurkers or read-only subscribers. Typically, however, lurkers do not doubt the significance and usefulness of the online group they partake in; they simply do not contribute actively.

To overcome the problem of under-contribution, Kraut (2003) borrows from current social psychology theories like, for example, Karau and William's (1993) theory of social loafing. In particular, he suggests design guidelines to increase participation rates in this group. Kraut's guidelines include the identifiability of members, task attractiveness, group attractiveness, the group's overall size, and the recognition of the uniqueness and high significance of one's own contribution (compared to other members' contributions) as key variables to collective effort in online communities.

With respect to these key variables, Kraut (2003) suggests various design implications or strategies for optimization. For instance, identifiability is known to be an indispensable prerequisite to the success of online communities: Only if anonymity is not allowed, any change or progress being made will be displayed and connected to individual group members (see also Janneck et al., 2005). Individual behavior will thus become accountable.

In addition, to increase the attractiveness of contributing, the underlying software should provide interactive elements. Elements like, for example, a chat function supports mutual communication, which is more attractive and requires less effort than asynchronous communication.

Taken together, identifiability and providing interactive elements address, and eventually reduce, a broad range of group problems like social loafing or production blocking, because they act as a motivating effect on perceiving one's own performance.

Kollock (1999) also focuses on personal identifiability, which he links to the memory functions of a community supporting technological system. More precisely, he argues that cooperation within an online community can only be successful if individuals:

1. Can recognize each other, that is, they are not operating anonymously within the community;
2. Have access to each others interaction history; and
3. Share the presumption of a high likelihood of a future encounter within the online community.

This leads to the conclusion that online communities have to offer possibilities of creating and managing relationships by supporting continuous interaction between their members. Therefore, it is advantageous if the system has a memory, in the sense that every community member and every item stored in the system holds a personal history.

People tend to act from self-interest if they are aware that their actions have effects on their reputations: high-quality contributions, impressive knowledge, and the perception of being willing to help others enhance the prestige of the community member. Although altruism as a motivational state for taking part in an online community is less common in comparison with self-interest, it is still frequent enough to be addressed if one thinks about community building. People with altruistic motivation try to meet the needs of the group or certain group members. This motivational state can be satisfied by establishing a public space where these needs can be stated, communicated, and discussed.

Considering the public goods dilemma, it is essential to introduce a role concept to clearly communicate the borderline between being in a group and being out of a group. To get full access to all group resources, one has to join the group. Several functionalities are only accessible for registered and authorized users. The commitment that is required to join the group leads to establishing a comprehensible boundary between members and nonmembers. This, in turn, facilitates the togetherness of the group and the identification of the members within the group. The membership itself constitutes a strong coupling feature.

Three further selective measures address the public goods dilemma: personal presence, personal responsibility, and personal history. Anonymity and lack of continuity among the members promotes egoistic behavior. Therefore, modifying actions should be tagged with users' login names, which, in turn, should be associated with records of personal data. Tagging entries and actions with user login names makes it easy to recognize people and enhances the development of personal relationships and online cooperation among the community members. Seeing all modifying actions supports the participator's personal responsibility. If the system has a memory of every action, this gives a personal history to every community member, as well as to the community's artifacts. Information about past interactions of the members again increases personal responsibility. Whereas, information about interaction with the artifacts allows members to get up-to-date, or introduce members to new areas of interest, as well as familiarize new members with the online community's etiquette and Who's Who.

"Content is king" is commonplace for virtually all Web-based efforts. This is notably true for online communities operating on a user-as-editors base. To implement a reasonable quality assurance system, it is crucial to apply technical, as well, as social devices. On a technical level, this can be done by employing a content rating system. Employing an "A team" of highly motivated volunteers can, on the other hand, help the online community to start up by delivering good content. For all content producers, it has to be as easy as possible to feed new content into the system. The best way of avoiding barriers is through continuous usability testing.

Introducing dedicated and active moderators seems to be the most important step to nourish motivation of the community members. Moderators can enhance group activities and increase the efficiency of the group. They are responsible for communicating the group codex (etiquette), acting as role models for new community members, and helping in preserving continuity. Rojo and Ragsdale (1997) show that an active moderator can, to some extent, compensate for the lack of active members in an online community.

In face-to-face communities only members can start a conversation. In online communities on the other hand, the technical systems can initiate communication as well. This opportunity should be brought into play by implementing awareness functions as software agents that collect relevant information for users and present it in e-mails, RSS feeds or personalized portal pages. These agents generally base their information gathering and presenting strategies on keywords or categories stored in configuration files (profiles). It is crucial to keep these profiles up-to-date. Experience shows however, that the members' interests continuously change over time. Profile setting dialogues are often accessed once and then forgotten. Thus, there is risk for personalized services to decrease in quality over time. Hence, it is important to monitor user behavior and let the agents ask from time to time if their interpretations of observations of, for example, changing interests, are correct. Furthermore, people may change their general attitude toward the community. This can be connected to an altered degree of involvement and must be considered during the design phase (i.e., modifiable feature sets).

The open-source movement has become very successful in recruiting new developers who start their own projects or join existing software development efforts. Today, most, if not all, open-source software development communities use online services for collaboration. In the remainder of this section, the application of the requirements for software that supports online community building is demonstrated by the following examples of open-source software development communities:

- **"Killer feature":** Open-source projects are often founded to solve an existing problem, that is, the killer feature is the product of the online community. When others join the project, the work becomes even more effective.
- **Recruitment:** Open-source communities produce software systems that are not only intended for use by original members but also for external clients. These

external users, that is, users not actively involved in the open-source online community, can be made active developers that modify the source and give it back to the project. To foster this process of developer recruitment, online communities should provide transparent rules for becoming actively involved. The projects that do not seem to be hermetic have better chances of growing their developer base.

- **Transparency:** The most important possibility of gaining transparency is through a public archive of the project's mailing lists. Because it is often not easy to scan large e-mail archives, open-source communities should provide text documenting guidelines and standards.
- **Policy:** The Debian community, for example, maintains an open-source Linux distribution and is a good example of a growing community that has given itself a substantial set of roles, rules, and guidelines. They achieve transparency of their standards by publication of documents on their Web server: Other projects, such as UserLinux, use a Wiki for that, which makes such standards more vivid and activates community members' attendance.
- **Trust:** Debian has a twofold quality assurance system. There are no anonymous additions to the system, and approved maintainers electronically sign all modifications. Software bugs are reported by all users. The bug lists are available to the public.
- **Cooperation and usability:** CVS and Subversion are systems for configuration management for source code trees for distributed development teams. They are both good examples of cooperation software with a high usability that are efficient to use for everyday tasks in software development.
- **Awareness:** Workflows can provide awareness. Examples include automated e-mail distribution of users' software bug reports and e-mail notifications or RSS feeds of CVS or Subversion commits.

This exploration into open-source projects and their online coordination and cooperation tools reveals that a voluntary community approach works, and the infrastructures and supporting tools of these projects can be taken as a best practice reference case.

FUTURE TRENDS

Recently, the term "Socialware" was proposed for software systems dedicated to enhance social relations. According to Hattori, Ohguro, Yokoo, Matsubara, and Yoshida (1999), Socialware denotes systems which aim to support "various social activities on network communities." Supports include linking people with others, smooth communication in a community and information integration for a community. The Socialware approach was initially intended for CSCW systems that are used by stable communities. This approach seems suitable for implementing software for online communities as well.

It uses rules of interpersonal communication and transfers these structures into software. The technical concept associated with Socialware is a multiagent system architecture. The CSCW functionality is achieved through coordination and cooperation of a distributed set of software entities (agents). Users of a community system have personal agents for gathering and exchanging information, visualizing context information, and supporting decisions. Personal agents and the users they belong to are seen as personal units. Personal units interact with community agents that have the function of providing shared information and mediating communication between other personal units. This approach also makes it possible to link different partially-overlapping online communities.

A current development in online communities is the transformation of the virtuality of computer networks into the real world. There are different enabling technologies for mobile and ad hoc communities. An important factor is the ability to locate in cellular phone networks or with global positioning systems (GPSs). Using the positioning information as part of the application environment allows for mobile communities. They are often based on asynchronous communication, like Internet online communities. An example for such a mobile community is the petrol station price comparison community. In 2000, the German Research Center for Information Technology offered car drivers a location awareness service which gave a comparison price list of the varying petrol rates of all the petrol stations.

The availability of new short-range radio networking technologies, such as Bluetooth, WiFi or WiMAX, enables new synchronous mobile communities. This gives users the ability to connect devices ad hoc (i.e., without a server infrastructure), permitting mobility and interaction. As with other Internet online communities, game playing is an important technology driver, for example, pervasive group games are being developed (Pennanen & Keinänen, 2004) that could build up social structures in some ways comparable to online communities.

Finally, most characteristics that are prototypical to online communities can be found in so-called guilds in Massively Multiplayer Online Role-Playing Games (MMORPG). The gamers follow strict rules according to behavior and tasks for ensuring personal and common progress. Oral and written communication within and outside the game is essential for coordination and team play. Collaboration always has a short-term as well as a long-term perspective. Associated goals vary from coping with short but stressful in-game situations (e.g., fighting the "final enemy") to preparing and organizing activities months in advance (e.g., acquir-

ing needed resources). In spite of the "role-playing" label, serious and reliable relationships can be established within the MMORPG community (Yee, 2006).

CONCLUSION

Advanced software solutions like the aforementioned Socialware approach can help to build and maintain stable online communities. In the long run, though, it is not the technology; it is the people that make an online community work. Using the most advanced technology is neither sufficient nor, as early BBS/MUD approaches show, necessary to assure the building of a stable online community. People will always make creative use of technology by using it in other ways than were originally intended by the designers. This will, once in a while, generate possibilities for new online communities.

Nevertheless, the most important factor for successful building and maintaining an active online community is providing awareness about changes in the communities' databases to members. Awareness functions provide an understanding of the others members' activities and the communities' goals and progress; the user can thus relate and evaluate their own activities accordingly.

REFERENCES

Aigrain, P. (1997). Attention, media, value and economics. *First Monday, 2*(9). Retrieved May 28, 2008, from http://www.firstmonday.org/issues/issue2_9/aigrain/index.html

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Baumeister, R. F., & Bushman B. J. (2008). *Social psychology and human nature*. Belmont, CA: Thomson Wadsworth.

Berners-Lee, T. J., Cailliau, R., Groff, J. -F., & Pollermann, B. (1992). World-Wide Web: The information universe. *Electronic Networking: Research, Applications and Policy, 2*(1), 52-58.

Christensen, W., & Suess, R. (1978). Hobbyist computerized bulletin board. *Byte Magazine, 3*(11), 150-158.

Franken, R. E. (2001). *Human motivation* (5th ed.). Pacific Grove, CA: Brooks/Cole.

Ghosh, R. A. (1997). Economics is dead. Long live economics! A commentary on Michael Goldhaber's "The attention economy." *First Monday, 2*(5). Retrieved May 28, 2008, from http://www.firstmonday.org/issues/issue2_5/ghosh/index.html

Goldhaber, M. H. (1997). The attention economy and the Net. *First Monday, 2*(4). Retrieved May 28, 2008, from http://www.firstmonday.org/issues/issue2_4/goldhaber/index.html

Hagel, J., & Armstrong, A. G. (1997). *Net gain: Expanding markets through virtual communities*. Boston, MA: Harvard Business School Press.

Hattori, F., Ohguro, T., Yokoo, M., Matsubara, S., & Yoshida, S. (1999). Socialware: Multi-agent systems for supporting network communities. *Communication of the ACM, 42*(3), 55-61.

Janneck, M., Finck, M., & Oberquelle, H. (2005). Social identity as an agent of technology-use mediation in virtual communities. *I-com, 4*(2), 22-28.

Karau, S. J., & William, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality & Social Psychology*, *65*(2), 681-707.

Kim, A. J. (2000). *Community building on the Web: Secret strategies for successful online communities*. Berkeley, CA: Peachpit Press.

Kollock, P. (1999). The economies of online cooperation. Gifts and public goods in cyberspace. In M. A. Smith, & P. Kollock (Eds.), *Communities in cyberspace* (pp. 220-242). London: Routledge.

Kollock, P., & Smith, M. A. (1996). Managing the virtual commons: Cooperation and conflict in computer communities. In S. Herring (Eds.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 109-128). Amsterdam, Netherlands: John Benjamins.

Kraut, R. E. (2003). Applying social psychological theory to the problems of group work. In J. M. Carroll (Ed.), *HCI models, theories, and frameworks: Toward a multidisciplinary science* (pp. 325-356). San Francisco: Morgan Kaufmann.

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. New York: Cambridge University Press.

Pennanen, M., & Keinänen, K. (2004). Mobile gaming with peer-to-peer facilities. *ERCIM News, 57*, 31-32.

Preece, J. (2000). *Online communities: Designing usability and supporting sociability*. Chichester, UK: John Wiley & Sons.

Rheingold, H. (1994). A slice of life in my virtual community. In L. M. Harasim (Ed.), *Global networks: Computers and international communication* (pp. 57-80). Cambridge, MA: MIT Press.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier* (rev. ed.). Cambridge, MA: MIT Press.

Rojo, A., & Ragsdale, R. G. (1997). A process perspective on participation in scholarly electronic forums. *Science Communication*, *18*(4), 320-341.

Utz, S. (2000). Identifikation mit virtuellen Arbeitsgruppen und Organisationen. In M. Boos, K. J. Jonas, & K. Sassenberg (Eds.), *Computervermittelte Kommunikation in Organisationen*. Göttingen, Germany: Hogrefe.

Yee, N. (2006). The psychology of MMORPGs: Emotional investment, motivations, relationship formation, and problematic usage. In R. Schroeder & A. Axelsson (Eds.), *Avatars at work and play: Collaboration and interaction in shared virtual environments* (pp. 187-207). London: Springer-Verlag.

KEY TERMS

Community Building: All activities related to building and maintaining online communities.

CSCW (Computer-Supported Cooperative Work): Software tools and technology as well as organizational structures that support groups of people (typically from different sites) working together on a joint project.

MMORPG (Massively Multiplayer Online Role-playing Games): Role-playing games played online by a large number of players at the same time. Participants are represented by customized avatars and solve different tasks (quests) on their own or in coordinated groups.

Online Community: An online community is a voluntary group of active users that partake actively in a certain computer-mediated service.

Socialware: Socialware aims to support various social activities on a network. Rules of interpersonal communication are used and transferred into community software.

UaE (User-as-Editors) Approach: The community members are responsible for supplying new content and assuring the quality of existing content, as well as for creating and maintaining the etiquette of the community.

Virtual Community: This is a featureless and, therefore, often misleading term usually regarded as synonymous to online community. The term "online community" is preferable, as it denotes the character of the community more accurately.

Wiki: Internet service based on HTTP and HTML providing "open editing" of Web pages with a Web browser. Hyperlinks between documents are supported with simple textual references. By default, everybody is allowed to edit all available pages.

Online Learning as a Form of Accommodation

Terence Cavanaugh
University of North Florida, USA

INTRODUCTION

An estimated three billion people, representing approximately half of the planet's population, are in some way affected by disabilities, which includes an estimated 150 million from the United States of America (Half the Planet, 2001). According to the *Twenty-Third Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act* (U.S. Department of Education, 2002a), concerning students with special needs between the ages of three and 21, the U.S. and its outlying areas are currently serving educationally more than 6,272,000 students classified as having a disability. The inclusion model, in which a special needs student participates in the "regular" classroom, has become the current classroom education standard. Today's special needs students have increasing impacts on the general education teacher as, during the past 10 years, the percentage of students with disabilities served in schools and classes with their non-disabled peers has gradually grown to over 90% in 1998 (U.S. Department of Education, 2000b). Because of the large and increasing number of special needs students, assistive educational technology is growing in importance. The population of postsecondary students with disabilities has increased over the past two decades, and currently there are approximately one million persons in postsecondary institutions who are classified as having some form of disability (U.S. Department of Education, 2000b). In 1994, approximately 45% of the adult population who reported having a disability had either attended some college or had completed a bachelor's degree or higher, as compared to only 29% in 1986 (National Center for Educational Statistics, 1999a).

BACKGROUND

Changes in the Population of Schools

While the makeup of the student population (K-20) has changed, because more students have been classified as having a disability and are now included in the general educational population, so too have the possibilities of the educational setting changed. For the 1999-2000 school year, the number of U.S. students with disabilities served was 588,300 preschool children and 5,683,707 students ages 6 through 21, representing an increase of 2.6% over the previous year (U.S. Department of Education, 2002a). Instructors now have on hand instructional tools that include forms of interactive telecommunication, such as the Internet and two-way video communication, as options for the delivery of instruction. While schools may not have been planning, designing, and creating distance learning courses and programs to meet the needs of students with disabilities, many students' needs were met through such a delivery system nonetheless. Electronic learning in and of itself is a form of instructional accommodation. Additionally, a range of assistive technology can support the student in the distance learning environment. The online class can be an assistive technology tool that students can use who would otherwise not be able to participate in a classroom for physical, health, or other reasons.

The number of students with disabilities is growing in the online education environment. A 1999 Canadian study of students with disabilities attending community colleges and universities found that an overwhelming majority of respondents (95%) indicated that they used a computer in their education situation, to the most noted reason for using the Internet was for doing research (Fichten, Asuncion, Barile, Fossey & De Simone, 2000). Thompson's (1998) summarizing report states that approximately 5% of the undergraduates at Open University of the United Kingdom have disabilities, with their population increasing at a rate of approximately 10% per year. The growth is ascribed to the convenience of home study and the ability of technology to overcome barriers to learning for students with disabilities. According to the U.S. Department of Education's (2002a) National Postsecondary Student Aid Study of 1999-2000, more than 8% of all undergraduates took at least one distance learning course, and 9.9% of those students identified themselves as having some form of disability.

Accommodations or Modifications Needed for Disabled Access

There is a difference between accommodations and modifications for students with special needs. Accommodations are considered to be provisions made in how a student accesses and/or demonstrates learning. The term accommodations focuses on changes in the instruction, or how students are expected to learn, along with changes in methods of assessment that demonstrate or document what has been learned. The use of an accommodation does not change the educational goals, standards, or objectives, the instructional level,

or the content, and provides the student with equal access and equal opportunity to demonstrate his or her skills and knowledge (State of Florida, Department of State, 2000). Accommodations assist students in working around the limitations that are related to their disabilities and allow a student with a disability to participate with other students in the general curriculum program. Accommodations can be provided for: instructional methods and materials; assignments and assessments; learning environment; time demands and scheduling; and special communication systems. By comparison a modification is a change in what a student is expected to learn and demonstrate. The use of a modification for a student changes the standard, the instructional level, or the content to be learned by the student (Beech, 2000).

According to the Assistive Technology Education Network (ATEN) of Florida (2000), instructors of any classes that have students with disabilities should provide students with:

- opportunities to interact with others,
- varied models of print use,
- choices—and then wait for the student to respond,
- opportunities to communicate, and
- expectations that students will communicate, this may require the use of an alternate or augmentative form of communication.

Online instruction, especially through asynchronous Internet presentation, provides all of ATEN's requested opportunities for students. In a "traditional" course, a teacher or professor would be in a classroom, with the students sitting at tables or desks, and there would be lectures, demonstrations, possibly videos and slideshows, handouts, and readings. In an online course in an asynchronous course model, these interactions could still take place, but without the limitations of specific time and location (Picciano, 2001). In such a distance learning course, the main interactions between the student and the instructor take place using course Web pages, streaming audio and video, forums, e-mail, and online books. Assistive tools that a student with special needs may require could be more easily applied in the online environment; such assistive tools may include speech-to-text programs, environmental control devices, or assistive hearing devices. Additionally the asynchronous course design allows the students to access the information at the course Web site and learn at a time convenient to them (Barron, 1999). Within online course sections there could be forums or discussions in which students can participate, allowing each and every student the opportunity and appropriate time to develop and share responses, again without the time restrictions of the standard class period.

The Law, the IEP, and Education

Federal laws and their directives charge that each student classified as having any form of disability have an individual education plan (IEP) developed specifically for that student, that assistive technology devices and services must be considered, and that the student must be taught in the least restrictive environment (Individuals with Disabilities Education Act, 1992). The IEP will be developed by a team of people including teachers, administrators, counselors, parents, outside experts (as needed), and often the student. Distance learning can be considered an adapted form of instruction that through the use of telecommunication technology (usually the Internet) allows a student to participate in a class, meeting the classification of assistive technology. While some students with special needs may choose distance learning courses because these courses provide the necessary accommodations or educational modifications that they need in order to function in that form of "classroom," that in and of itself is not enough. It is up to educators to make sure that the accommodations and educational modifications necessary for these students to function in our classrooms exist or can be made available to these students as they need them. The educational charge extends to ensuring that distance learning classes are also accessible. These distance learning courses or situations must be designed, accommodated, or modified to allow students with special needs to be able to effectively participate.

Distance Learning and Students with Disabilities

A recent survey of seven open enrollment distance learning schools (state, public or private, or college/university) that offered Internet-based instruction may indicate trends in the current status of distance learning programs and special needs students. The distance learning population of the responding schools ran from 300 to 5,000 full- or part-time students, with an average of approximately 1,000 students. Most schools indicated that they did not have records or tracking methods for identifying students with disabilities. Schools that did identify these students indicated that special needs populations ran between 2% and 10%. With the exception of the responding university school, all the K12 distance learning schools indicated that their teachers have participated in IEPs for students. Half of the respondent schools indicated that they currently had students taking courses as part of the student's IEP, as recommended or required by the student's home school IEP team. The schools also indicated that they did not participate as IEP team members, but that the school or distance learning environment was written in as a service in the student's IEP. A consistent thread in the responses was that the distance education schools were

sure that they had special needs students, but that they were not identified. When identifying their accommodations for students with special needs, all of the schools responded that they were willing to make accommodations, and the most common accommodation occurring in their programs was extending required time on tests and assignments. When questioned about the virtual school's faculty, only half of the respondents indicated that they had both counselors and exceptional education teachers on staff. Others indicated that they depended on counselors or other support personal to work with the student's home school. Interestingly, all responding virtual schools did indicate that distance learning instructors have already had, or currently have access to, training concerning accommodations for educating special needs students. Only two of the responding distance learning schools indicated that their Web-based instructional pages were compliant with either national or international accessibility guidelines, with one indicating that it was in the process of becoming so.

Hospital/Homebound Students

According to U.S. government statistics, there are more than 26,000 students classified as hospital/homebound students across the nation (U.S. Department of Education, 2002a). How these students are being served at a distance from their "home" school is a distance learning strategy question. The classic hospital/homebound program has a visiting teacher who acts as intermediary between a student's regular teacher and the student. The hospital/homebound teacher contacts the student's classroom teacher or teachers to collect assignments and directions to deliver to the hospital/homebound student, and visits the student to provide instruction and assistance. The more common approaches for hospital/homebound education are the "teleclass" phone model for secondary students and hospital/home visitation for the elementary level. In phone-based or teleclass instruction, all students taking a course dial into a common number, and a teacher provides oral instruction. Students have the opportunity to ask questions and interact through voice (G.A. Ball, personal phone communication, October 2, 2002). This instruction by phone service qualifies as a true distance learning program, as education is provided to the students through a telecommunication system. Other school systems also use online instruction, audio/videotaped instruction, and CD-ROMs, where the districts provide the needed hardware and software; some district school systems have even placed fax machines in students' homes for students to use to receive and submit assignments (C. Bishop, personal e-mail communication, January 2003).

FUTURE TRENDS

Online Education

Distance learning is a growing educational option. The numbers of both public and private distance learning intuitions are on the rise, along with the numbers of students attending. An instructional example would be the Florida Virtual School, going from 227 students in 1998 to 8,200 in 2002, an increase of 3,612% in just five years. School systems and students need flexible options to increase educational success, and using distance learning as an alternative can do that. We are already seeing experiments with asynchronous distance learning being used as an alternative for the hospital/homebound, but research still needs to be done. Virtual schools need to track the special needs students and identify effective methods that are working with them. Hospital/homebound programs need to investigate options for students who must be out of school for extended periods of time and determine the effectiveness of online skill-building education versus asynchronous module instruction of courses. Also, can distance learning assist such programs by allowing small or rural districts to band together and use common certified teachers, who are in short supply and are needed for No Child Left Behind requirements, to be shared between districts? In addition to human concerns, the future brings hardware and software issues, since most of the course management systems (CMSs) are not "disabled" accessible according to US 508 or W3C accessibility guidelines. Schools need to either create their own accessible course systems or apply pressure so that the currently available systems become compliant with access guidelines.

CONCLUSION

Distance Learning as an Accommodation

Electronic learning in and of itself is a form of accommodation, and there is also a range of assistive technology that can support the student in the distance learning environment. The online class can be an effective assistive technology tool that students can use who would otherwise not be able to participate in a classroom for physical, health, or other reasons. While distance learning may not be appropriate for every course or accommodating to every form of disability, it does provide a viable instructional option for many. As the inclusive education of all students occurs more frequently within the standard K12 classroom and through the electronic environment at the college level, it is reasonable to expect that more students with disabilities will participate in online education. While many accommodations are already avail-

able online, instructors need to insure that online courses should adhere to accessibility guidelines as proposed by state, national, or international organizations to allow all students access. These educational environments should be designed for all students, even those students who may need modifications, accommodations, and assistive technology.

REFERENCES

Assistive Technology Education Network of Florida (ATEN). (2000). *Assistive technology: Unlocking human potential through technology.* Presentation at the University of South Florida, USA.

Barron, A. (1999). *A teacher's guide to distance learning.* Tampa, FL: Florida Center for Instructional Technology.

Beech, M. (2000). *Accommodations and modifications*: *What parents need to know*. Florida Developmental Disabilities Council, Inc. ESE10753.

Fichten, C.S., Asuncion, J.V., Barile, M., Fossey, M. & De Simone, C. (2000, April). *Access to educational and instructional computer technologies for postsecondary students with disabilities: Lessons from three empirical studies.* EvNet Working Paper. Retrieved from evnet-nt1.mcmaster.ca/network/workingpapers/jemdis/jemdis.htm.

Half the Planet. (2001). *Half the Planet foundation information.* Retrieved from www.halftheplanet.com.

Individuals with Disabilities Education Act. (1992). Pub. L. No. 101-476. Retrieved from frWebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=105_cong_public_la.

National Center for Educational Statistics. (1999a). *Students with disabilities in postsecondary education: A profile of preparation, participation, and outcomes.* Retrieved from nces.ed.gov/pubs99/1999187.pdf.

Picciano, A.G. (2001). Distance learning: Making connections across virtual space and time. Upper Saddle River, NJ: Prentice-Hall.

State of Florida, Department of State. (2000). *Developing quality individual educational plans.* Document ESE9413, Bureau of Instructional Support and Community Services, Florida Department of Education.

Thompson, M.M. (1998). Distance learners in higher education. *Global Distance Education Net.* Retrieved from wbWeb5.worldbank.org/disted/Teaching/Design/kn-02.html.

U.S. Department of Education. (2000b). To assure the free appropriate public education of all children with disabilities. *Twenty-Second Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act.* Retrieved from: www.ed.gov/offices/OSERS/OSEP/Products/OSEP2000AnlRpt/index.html.

U.S. Department of Education. (2002a). *Twenty-Third Annual Report to Congress on the Implementation of the Individuals with Disabilities Education Act*. Retrieved from www.ed.gov/offices/OSERS/OSEP/Products/OSEP2001AnlRpt/index.html.

U.S. Department of Education. (2002b). The percentage of undergraduates who took any distance education courses in 1999-2000, and among those who did, the percentage reporting various ways in which the courses were delivered. *1999-2000 National Postsecondary Student Aid Study (NPSAS:2000).* NEDRC Table Library. Retrieved from nces.ed.gov/surveys/npsas/table_library/tables/npsas22.asp (number of disabled students taking distance learning).

KEY TERMS

Accommodations: Provisions made in how a student accesses and/or demonstrates learning. The term focuses on changes in the instruction, or how students are expected to learn, along with changes in methods of assessment that demonstrate or document what has been learned. The use of an accommodation does not change the educational goals, standards, or objectives, the instructional level, or the content, and provides the student with equal access and equal opportunity to demonstrate his or her skills and knowledge.

Assistive Technology: "…any item, piece of equipment, or product system, whether acquired commercially off the shelf, modified, or customized, that is used to increase, maintain, or improve functional capabilities of individuals with disabilities…." (20 U.S.C. 1401 (33)(250))

Asynchronous: Communications between the student and teacher which do not take place simultaneously.

Disabled Student: From the *U.S. Federal Register*: child/student has been evaluated as having mental retardation, a hearing impairment including deafness, a speech or language impairment, a visual impairment including blindness, serious emotional disturbance (hereafter referred to as emotional disturbance), an orthopedic impairment, autism, traumatic brain injury, an other health impairment, a specific learning disability, deaf-blindness, or multiple disabilities, and who, by reason thereof, needs special education and related services (IEP or 504).

Inclusion: A classroom design where all students should take part and attend "regular" classes. Generally, an ESE and regular education teacher work together with the same group of students, including students with disabilities and

general education students. Both of the teachers share the responsibility for all of the students.

Individualized Education Program (IEP): A written statement for each child with a disability that is developed, reviewed, and revised in accordance with this section. (20 U.S.C. 1414 (d)(1)(A)) (Individuals with Disabilities Education Act, 1997)

Modification: A change in what a student is expected to learn and demonstrate. The use of a modification for a student changes the standard, the instructional level, or the content to be learned by the student.

Specific Learning Disability: Term meaning a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, that may manifest itself in an imperfect ability to listen, think, speak, read, write, spell, or to do mathematical calculations, including conditions such as perceptual disabilities, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia.

Teleclass: Voice-only communications linking two or more sites. A standard method used is to connect multiple telephone lines for an audio conference through a phone bridge. A telephone bridge where the conference is established by having all of the distant sites call in to a common bridge telephone number.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 47-52, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Online Student and Instructor Characteristics

Michelle Kilburn
Southeast Missouri State University, USA

Martha Henckell
Southeast Missouri State University, USA

David Starrett
Southeast Missouri State University, USA

INTRODUCTION

As technological advances become mainstream in higher education, many universities have begun delving into online learning as an effective means of course delivery. Transitioning from the Industrial Age to the Digital Age of learning has forced some evaluators to rethink standards of success and the idea of productivity and learning (Leonard, 1999).

Understanding the positive attributes of students and instructors in the online environment will contribute to the understanding of how we can enhance the learning experience for the student and the teaching experience for the instructor. This article will also assist students and instructors in understanding the differences that may be experienced in the online environment vs. the face-to-face environment and provide the opportunity to consider whether online learning or teaching is a "good fit" for them. Understanding why students or instructors might choose the online environment will also assist administrators in developing successful, quality online programs that enrich the experiences for both students and instructors.

BACKGROUND

In 1981, the first online classes were developed at the School of Management and Strategic Studies at Western Behavior Sciences Institute in La Jolla, California. An evaluation of the program, and the discussions that took place, revealed that the quality of the online course was higher than the information collected in the traditional classroom setting (Feenberg, 1999). Jung, Choi, Lim, and Leem (2002) also found that online instruction showed significantly better results on examinations, complicated problems, or student's perception of learning outcomes.

With the popularity of the Internet, and the continuing demand for online courses, many college and university administrators might find it challenging to incorporate online technology. Many may feel pressured to jump on the "online bandwagon" in order to keep up with the student demand for these types of courses.

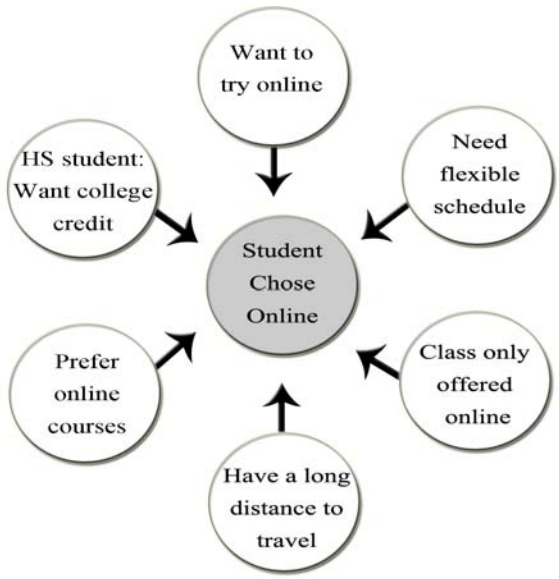
Kilburn (2005) developed the following conceptual map (see Figure 1) regarding student motivations to take an online course at a particular university in the Midwest:

In the upcoming section, an examination of student and instructor characteristics and how each of those different roles contributes to the quality of an online course will help provide insight into the foundational underpinnings of Web-based learning.

STUDENT CHARACTERISTICS

It is estimated that five out of six students taking an online course are employed and would not be able to attend tradi-

Figure 1. Student motivations to take an online course at a Midwest university



tional classes (Thomas, 2001). Moore and Kearsley (1996) and Hardy and Boaz (1997) found that most distance learners are working adults, primarily female. Literature suggests that the growth in online courses is based on attracting new students rather than "stealing" from students enrolled in current on-campus programs (Mangan, 2001; Thomas, 2001).

Some researchers have attempted to identify student abilities that will suggest whether a student will complete an online course, or be less satisfied with an online course, in comparison to the traditional classroom setting. Kilburn (2005) found that positive characteristics identified in studies stress the importance of an active vs. passive student role in an online course and include: self-motivation and the ability to organize thought (Hardy & Boaz, 1997), prior experience with technology (Richards & Ridley, 1997), positive attitude regarding the subject matter (Coussment, 1995), learning and personality styles (Saunders, Malm, Malone, Nay, Oliver, & Thompson, 1998), self-selection of online courses vs. forced-choice (Thomerson & Smith, 1996), intrinsic motivation, and self-reported explorative behavior (Martens, Bastiaens, & Kirschner, 2007).

One attraction to online learning is the presumed capacity to increase access and equity to learners by removing some of the barriers to participation (Harasim, 1990). Sullivan (2001) found that the online classroom is often more welcoming for quiet or shy students than the traditional classroom. Online learning has also been advanced as a powerful, yet neutral, tool for enhancing the potential of distance education (Weisband, 1992).

Sullivan (2001) asserts that nontraditional students, particularly female with children or other familial responsibilities, value online courses. Research has suggested that the asynchronous nature of the online environment might encourage a more reflective type of interaction that changes the dynamics of classroom discussion in a way that female students find rewarding (Selfe, 1999). Another aspect of online learning that female students appear to find appealing is the relative anonymity online affords and changes to the social dynamics inherent in a traditional course (Sullivan, 1999).

Most students may initially select distance education as a result of perceived convenience (Klesius, Homan, & Thompson, 1997). These conveniences may include travel, compatibility with personal schedules, and opportunity for self-paced work. Online courses can accommodate all of these conveniences.

Busy work schedules and the expense of completing a higher education degree have often been the typical roadblocks encountered by students in traditional educational arenas (Holt, 1993). Online education has helped to remove some of these hurdles by saving commuting costs, equalizing classroom participation, and offering convenient scheduling for working students.

The age of a student generally is related to the course completion rate and the type of study in a distance education course. On average, students over 30 and under 50 years old are more likely to finish a distance course than traditional students age 18 to 29 (Willis, 1993). Technical skills also appear to play a role in the success of students taking online courses (Kerka, 1996). Students should possess the ability to navigate the Internet and deal effectively with computer software and hardware difficulties.

In the online environment, successful students shift from a more passive role in the exchange of knowledge to a more active role. Several studies purport that the following are the most influential factors affecting a student's active participation in online learning: (1) prior knowledge of online learning, (2) knowledge of a given subject area, (2) information overload, (3) personality traits, (4) instructor facilitation, and (5) appropriate feedback (Lim, 1999; Vonderwell & Zachariah, 2005).

INSTRUCTOR CHARACTERISTICS

The relationship between student-teacher interactions and learning outcomes has been well documented in the traditional classroom (Powers & Rossman, 1985). Of particular importance in traditional classrooms is teacher "immediacy tendencies." Immediacy alludes to the psychological distance between student and instructor (Weiner & Mehrabian, 1968). Research suggests that a teacher's verbal and nonverbal immediacy behaviors can help diminish the apparent distance between themselves and their students. By lessening the perception of disconnect between the instructor and student in a course, instructors can help facilitate (directly or indirectly) effective learning.

With the importance of interactions established as a crucial component of learning, one might assume it would be equally important online. Certain researchers have suggested that asynchronous media are less capable of representing the social presence of participants (Short, Williams, & Christie, 1976). Researchers with experience teaching online contest this view, arguing that rather than being impersonal, computer-mediated communication often seems to be hyperpersonal (Walther, 1994). According to Kilburn (2005), research indicates that participants in online courses communicate, argue, and create social presence by projecting immediacy behaviors (LaRose & Whitten, 2000; Rourke, Anderson, Garrison, & Archer, 2001).

Previous studies (Thompson & Chute, 1998) suggest that in order to enhance learning motivation, small group activities are important in online courses. One might argue that learning motivation is more important in distance education courses because distance learners with low motivation have more of a tendency to drop out or fail when

compared to traditional on-campus courses (Keller, 1999). Interactions among students seem to clearly matter in online discussions. The development of social presence and the perceived interaction with others is one of the cornerstones for the development of online communities (Rourke et al., 2001). Lu and Jeng (2006) discovered that instructors who strived to serve as both facilitator and co-participant were helpful in "enhancing knowledge construction," particularly in discussion forums (p. 196).

Eastmond (1995) points out that online communication should not be assumed to be intrinsically interactive but it depends on the nature, frequency and timeliness of discussions. Hawisher and Pemberton (1997) found there was a potential correlation with the success of an online course and the value the instructor placed on discussions. Picciano (1998) found that students' perceptions of learning were related to the amount of discussion that took place in online course. Likewise, Jiang and Ting (2000) report correlations between perceived learning in online courses and the specificity of the instructors' discussion, instructions and the percent of course grades based on discussion responses. These studies exemplify the magnitude of importance that is placed on the instructor to assure the quality of an online course.

Researchers have begun to look at the changing roles of teachers in online classrooms. Coppola, Hiltz, and Rotter (2001) assert that in any delivery medium teachers have three roles: cognitive, affective, and managerial. They found with online courses the cognitive role often becomes more complex and paramount. The affective role requires instructors to find new tools to express emotion and the managerial role requires greater attention to detail, more structure, and additional student monitoring. Easton (2003) reported that prior to the start of class, the instructor role was more like an instructional designer and subject matter expert; however, once the class began, the instructor's activities took on more of an interactive, facilitator role. Darabi, Sikorski, and Harvey (2006) describe the various instructors' roles as managerial, social and technical.

Janicki and Liegle (2001) evaluated the work of a wide range of instructional design professionals and developed a list of 10 concepts believed to support effective design of Web-based instruction. The findings included: (1) instructors acting as facilitators, (2) usage of a variety of presentation styles, (3) multiple exercises, (4) hands-on problems, (5) learner control of pacing, (6) frequent testing, (7) clear feedback, (8) consistent layout, (9) clear navigation, and (10) available help screens.

Interaction has been acknowledged as an important component of learning in conventional and distance education (Moore, 1993). Studies have argued for the importance of instructors' social or interpersonal feedback when attempting to improve learning achievement in online courses (Jung, 2000: Leem, 1999). Dennen, Darabie, and Smith (2007) have suggested effective online instructors adopt the following practices: Maintaining frequency of contact, having a presence in class discussion, and making expectations clear to learners.

Instructors must balance academic interaction related to the subject matter with more personal student interaction that will cultivate a sense of relationship and community (Moller, 1998). Immediate and frequent feedback must be sufficient enough to communicate a sense of co-presence and a willingness to be responsive to student needs (Boettcher, 1999). Dennen (2007) purports that "instructors establish a persona via both presence (amount of instructor posts) and position (interaction relative to those in the student role)" (p. 95). Some online learners find it challenging to stay on task. Consequently, they fail to complete coursework. Prompt e-mail responses may help students remain on task and on schedule (Klesius et al., 1997).

FUTURE TRENDS

As online learning continues to grow in the higher education arena, more research needs to be done on the changing roles of students and instructors in the online environment. Time is no longer spent in the classroom passively listening to a lecture; students are required to play a more active role in their learning. Instructors often struggle to adapt their teaching approaches to develop active interaction and engagement in their online courses.

Rethinking quality assurance, how to measure student success and what constitutes quality interaction will be a crucial step in moving instructors and administrators from the traditional teaching and classroom management frame of reference to the "virtual campus." Encouraging instructors and students to "think outside the box" and experiment with ways to enhance interaction will help shape the future of online learning.

CONCLUSION

Students will need to adjust to a more self-directed, student-centered approach to learning, while instructors will need to move toward a facilitating model of instruction. Understanding how to assist both students and instructors to adjust to more nontraditional roles in academe is critical. Instructional designers in higher education have the critical role of helping outstanding face-to-face instructors translate their expertise, knowledge and personality into outstanding online courses. This discussion has reinforced the argument that developing an online course is not a matter of simply putting lecture notes and presentations online. Instructors must assume both an affective and managerial role. The instructional designer must help them stock a new "tool box" to effectively transfer knowledge, communicate with their students and develop a

sense of community in the online environment. Administrators may also need to rethink how to evaluate instructors in the promotion and tenure process in order to acknowledge instructors who excel in the online environment and make that critical connection with their students.

Understanding the premises and concepts in this discussion will not only assist administrators and instructors in developing quality online courses and programs, but also provide an understanding of how online learning has resulted in an evolution of student and instructor roles. Successful students take on a more active and participative role and successful instructors take on a more facilitative and supportive role.

REFERENCES

Berge, Z.L., Collins, M.P., & Day, M. (1995). *Computer mediated communication and the online classrooms*. Cresskill, NJ: Hampton Press.

Boaz, M., Elliott, B., Foshee, D., Hardy, D., Jarmon, C.G., & Olcott, D. (1999). *Teaching at a distance: A handbook for instructors*. San Diego, CA: Archipelago.

Boettcher, J.V. (1999). How many students are "just right" in a Web course? *Syllabus, 21*(1), 45-49.

Clark, R. (1983). Reconsidering research on learning from media. *Review of Educational Research, 53*(4), 445-459.

Coppola, N.W., Hiltz, S.R., & Rotter, N. (2001). *Becoming a virtual professor: Pedagogical roles and ALN.* Piscataway, NJ: Institute of Electrical and Electronics Engineers Press.

Coussement, S. (1995). *Educational telecommunication: Does it work? An attitude study*. (ERIC No. ED 391 465)

Darabi, A., Sikorski, E., & Harvey, R. (2006). Validated competencies for distance teaching. *Distance Education, 27*(1), 105-122.

Dennen, V.P. (2007). Presence and positioning as components of online instructor persona. *Journal of Research on Technology in Education, 40*(1), 95-108.

Dennen, V.P., Darabi, A.A., & Smith, L.J. (2007). Instructor-learning interaction in online courses: The relative perceived importance of particular instructor actions on performance and satisfaction. *Distance Education, 28*(1), 65-79.

Eastmond, D.V. (1995). *Alone but together: Adult distance study through computer conferencing*. Cresskill, NJ: Hampton Press.

Easton, S.S. (2003). Clarifying the instructor's role in online distance learning. *Communication Education, 52*(2), 87-105.

Feenberg, A. (1999). No frills in the virtual classroom. *Academe, 85*(2), 26-31.

Harasim, L. (1990). *Online education: Perspectives on a new environment*. New York: Praeger.

Harasim, L.M. (1995). *Learning networks: A field guide to teaching and learning online*. Houston, TX: MIT Press.

Hardy, D.W., & Boaz, M.H. (1997). *Teaching and learning at a distance: What it takes to effectively design, deliver, and evaluate programs*. San Francisco: Jossey-Bass.

Hawisher, G.E., & Pemberton, M.A. (1997). Writing across the curriculum encounters asynchronous learning networks or WAC meets up with ALN. *Journal of Asychronous Learning Networks*.

Holt, L. (1993). Learning that transcends time and place. *Educational Technology, 79*(4), 50-56.

Janicki, T., & Liegle, J.O. (2001). Development and evaluation of a framework for creating Web-based learning modules: A pedogogical and systems approach. *Journal of Asychronous Learning Networks, 5*, 189-206.

Jiang, M., & Ting, E. (2000). A study of factors influencing students' perceived learning in a Web-based course environment. *International Journal of Educational Telecommunications, 6*, 317-338.

Jung, I.S. (2000). Technology innovations and the development of distance education. *Open Learning, 15*(3), 217-231.

Jung, I., Choi, S., Lim, C., & Leem, J. (2002). *Effects of different types of interaction on learning achievement, satisfaction and participation in Web-based instruction*. Retrieved May 29, 2008, from http://www.tandf.co.uk/journals

Jung, I.S., & Leem, J.H. (1999). Design strategies for developing Web-based training courses in a Korean corporate context. *International Journal of Educational Technology, 1*(1), 107-121.

Keller, J.M. (1999). Motivation in cyber learning environments. *International Journal of Educational Technology, 1*(1), 7-30.

Kerka, S. (1996). *Distance learning, the Internet, and the World Wide Web*. Columbus, OH: Office of Educational Research and Improvement.

Kilburn, M. (2005). *A descriptive analysis of online learning at a Midwest university*. Doctoral dissertation, University of Missouri, Columbia. OCLC 70724933.

Klesius, J., Homan, S., & Thompson, T. (1997). Distance education compared to traditional instruction: The students'

view. *International Journal of Instructional Media, 24*(3), 207-220.

LaRose, R., & Whitten, P. (2000). Re-thinking instructional immediacy for Web courses: A social cognitive exploration. *Communication Education, 49*, 320-338.

Leem, J.H. (1999). *Effects of small group learning strategies in Web-based problem solving environment on discussion, participation, and problem-solving*. Unpublished master's thesis, Seoul National University.

Leonard, D.C. (1999). The Web, the millennium, and the digital evolution of distance education. *Technical Communication Quarterly, 8*(1), 9-20.

Lim, C.I. (1999). Integrated approach in designing interactive WBI. *Korea Journal of Educational Technology, 15*(1), 3-24.

Lu, L.L., & Jeng, I. (2006). Knowledge construction in in-service teacher online discourse: Impacts of instructor roles and facilitative strategies. *Journal of Research on Technology in Education, 39*(2), 189-202.

Mangan, K.S. (2001). Expectations evaporate for online MBA programs. *Chronicle of Higher Education, 48*(6), A31.

Martens, R., Bastiaens, R., & Kirschner, P.A. (2007). New learning design in distance education: The impact on student perception and motivation. *Distance Education, 28*(1), 81-93.

Moller, L. (1998). Designing communities of learners for asynchronous distance education. *Educational Technology Research and Development, 46*(4), 115-122.

Moore, M.G. (1993). *Distance education: New perspective*. London: Croom Helm.

Moore, M.G., & Kearsley G. (1996). *Distance education: A systems view*. Belmont, CA: Wadsworth.

Picciano, A. (1998). Developing an asynchronous course model at a large, urban university. *Journal of Asychronous Learning Networks, 2*, 157-189.

Powers, S., & Rossman, M. (1985). Student satisfaction with graduate education: Dimensionality and assessment in college education. *Psychology: A Quarterly Journal of Human Behavior, 22*(2), 46-49.

Reeves, T.C., & Reeves, P.M. (1997). The effective dimensions of interactive learning on the WWW. In B.H. Khan (Ed.), *Web-based instruction* (pp. 59-66). Englewood Cliffs, NJ: Educational Technology.

Richards, C., & Ridley, D. (1997). Factors affecting college students' persistence in online computer-managed instruction. *College Student Journal, 31*, 490-495.

Rourke, L., Anderson, T., Garrison, D. R., & Archer, W. (2001). Assessing social presence in asynchronous text-based computer conferencing. *Journal of Distance Education, 14*(3), 51-71.

Saunders, N., Malm, L., Malone, B., Nay, F., Oliver, B., & Thompson, J. (1998). Student perspectives: Responses to Internet opportunities in a distance learning environment. *Mid-Western Educational Researcher, 11*(4), 8-18.

Selfe, C. (1999). Technology and literacy: A story about the perils of not paying attention. *College Composition and Communication, 50*, 411-436.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. Toronto, Canada: John Wiley.

Sullivan, P. (1999). Gender and the online classroom. *Teaching English in the Two-year College, 26*, 361-371.

Sullivan, P. (2001). Gender differences and the online classroom: Male and female college students evaluate their experiences. *Community College Journal of Research and Practice, 25*, 805-818.

Thomas, K.Q. (2001). Local colleges providing online learning programs. *Rochester Business Journal, 16*(43), 28.

Thomerson, D., & Smith, C. (1996). Student perceptions of affective experiences encountered in distance learning courses. *American Journal of Distance Education, 10*(3), 37-48.

Thompson, M.M., & Chute, A.G. (1998). A vision for distance education: Networked learning environments. *Open Learning, 13*(2), 4-11.

Vonderwell, S., & Zachariah, S. (2005). Factors that influence participation in online learning. *Journal of Research on Technology in Education, 38*(2).

Walther, J. (1994). Interpersonal effects in computer mediated interaction. *Communication Research, 21*, 460-487.

Weiner, M., & Mehrabian, A. (1968). *Language within language: Immediacy, a channel in verbal communication*. New York: Appleton-Century-Crofts.

Weisband, S.P. (1992). Group discussion and first advocacy effects in computer-mediated and face-to-face decision making groups. *Organizational Behavior and Human Decision Processes, 53*(3), 352-380.

Willis, B. (1993). *Distance education: Strategies & tools*. Englewood Cliffs, NJ: Educational Technology Publications

Woods, R. (2002). How much communication is enough in online courses? *International Journal of Instructional Media, 29*(4), 377-355).

KEY TERMS

Academic Interaction: Occurs when learners study materials and get task-oriented feedback from the instructor (Moller, 1998; Moore, 1993).

Asynchronous Learning: Electronic communication in which the student and teacher interact via e-mail and listservs, but do not do so by being on the Internet at the same time (Berge, Collins, & Day, 1995).

Distance Education: "Any formal approach to learning in which a majority of the instruction occurs while educator and learner are at a distance from one another" (Clark, 1983, p. 8).

Distance Learning: Learning that occurs when the instructor and students are separated by physical distance and technology is used to bridge the instructional gap (Boaz, Elliott, Foshee, Hardy, Jarmon, & Olcott, 1999).

Online Learning/Course: A context for learning in which students interact using technology and do not meet in a physical classroom with the instructor.

Synchronous Learning: Adjective used to describe an operation performed at the same time as another event (Boaz et al., 1999).

Web-Based Instruction: A media-rich online environment allowing people to interact with others asynchronously or synchronously in collaborative and distributed environments (Harasim, 1995), to gain access to remote multimedia databases for active and resource-based learning (Jung & Leem, 1999), and to manage self-paced individual learning in a flexible way (Reeves & Reeves, 1997).

Organization of Home Video

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

With the progress of electronic equipments and computer technology for taking motion pictures and processing huge data, an increasing number of people now own and use camcorders to make home videos that capture their experiences and document their lives. Home video has no time limits and no restriction in content (Lienhart, 2000), so these videos easily add up to many hours of material. However, the organization and edition of the large amount of information contained in home videos present technical challenges due to the lack of efficient tools. Though a number of prototype systems for content-based video analysis and retrieval have been constructed, for example, as shown in (Wactlar, 1996; Chang, 1998), the development of tools and systems specialized for addressing home video, that is for extracting, representing, organizing, browsing, querying and retrieving video, is just on a preliminary stage (Huang, 2005; Wu, 2005).

Several tasks are needed to confront to make the organization of home video possible and feasible. Home video has certain particular characteristics. The organization of home video should be based on the understanding of video structures, and by taking advantages of this structure. Home video are completed and stored straight in compressed domain. In order to save both time and space, techniques that manipulate home videos directly in compressed domain should be considered (Wang, 2003). Some typical techniques working on compressed domain could be found in (Taskiran, 2004). Home video are made by shot after shot without storyline, these shots may or may not have immediate relationship. To group shots, the visual features should be extracted from every shot (Gatica-Perez, 2003)

Facing these tasks and difficulties, a novel technique is described in this article. It is based on the analysis of characteristics of home video, on the detection of motion attention regions in compressed domain, on the time weighting based on camera motion, and on a novel two-layer shot clustering approach and organization strategy. Experiments made on two home videos from MPEG-7 data set provide encouraging results.

BACKGROUND

Video analysis is an important branch of content-based video retrieval (CBVR). Compared to other types of video programs, home video has some particularities according to the persons in shoot and objects to be screened [Lienhart 1997]. The study on home video analysis may benefit from its unique characteristics.

In general, a typical home video has certain structure characteristics: it contains a set of scenes, each composed of ordered and temporally adjacent shots that can be organized in clusters conveying semantic meaning. The fact is that home video recording imposes temporal continuity. Unlike other video programs, home video just records the life but not composes story, so every shot (clip of video captured in one place without interruption) may have the equal importance. In addition, filming home video with a temporal back-and-forth structure is rare. For example, on a vacation trip, people do not usually visit the same site twice. In other words, the content tends to be localized in time. Consequently, discovering the scene structure above shot level plays a key role in home video analysis. Video content organization based on shot clustering provides an efficient way of semantic video accessing and fast video editing.

Home video is not prepared for very large audience (like broadcasting TV), but for relatives, guests and friends. To analyze home video, the purpose and filming tact should be considered. For example, the subjective feeling transferred by video information can be decomposed into two parts: one from motion region that attract the attention of viewers, another from the general impression of environment. In the same time, different types of camera motions should also be considered. Different camera motions may signify different changes of attentions. For example, zoom in makes the attention of viewers more on motion regions, while pan and tilt make the notice of viewers more on environments.

Home video may be made by different persons in different circumstances. How to reflect the viewers' visual perception during the filming needs to be considered. In fact, different viewer's attention will never stay equal all through the watching of home video. The importance of certain frames or clips should be weighted reasonably by their relative importance according to subjective perception. Sometimes, users are required to assign these weights to express their real preference (Chang, 1998; Babu, 2002). Other approaches include temporally making weights in key frame selection procedure by assuming that the greater number of skipped frames, the more weight is assigned to the current frame (Tan, 1999), or building up a generic user attention model by integrating a set of audio-visual attention model features extracted from video sequence (Ho, 2003).

MAIN FOCUS OF THE CHAPTER

Detection of Attention Regions

Structuring video needs the detection of motion attention regions. This is not equal to the detection of video objects. The detection of video objects requires accurately determination of the boundary of objects and quickly following of the change of objects. For example, an object-based approach first imposes spatial-temporal segmentation to get individual regions (Achanta, 2002). The video structuring stresses more on the subjective feeling of human beings in viewing video. In this regard, the influence of region detection on subjective feeling is more important than just accurate segmentation (Zhai, 2005).

It is know that most object detection methods would fail if there were no specific clear-shaped object in the background. In order to circumvent the problem of actual object segmentation, a different concept: "attention region" could be employed. An attention region does not necessarily correspond to a real object, but denotes the region with irregular movement compared to camera motion. Based on the assumption that different movement from the global motion attracts more attention (supported by the common sense that irregular motion tends to be easily caught by human eyes in a static or regular moving background), these regions are regarded as somewhat important areas in contrast to the background image. The detection of attention region requires less precision in general, since more emphasis has been placed on the degree of human attention than on the accurate object outline. This task could be simply completed by detecting the outliers in frame. In addition, the tracking process for attention region is easier, too.

The first step of attention region detection is to segment a "dominant region" from a single frame, which is illustrated in Figure 1. Figure 1(a) is a typical home video frame. This frame can be decomposed into two parts: the running boy on grassplot, as shown in Figure 1(b); and the background grasses and trees, as shown in Figure 1(c). The latter represents the environment and the former corresponds to the attention region.

The detection of attention regions does not require at very high precision. Therefore, the detection of attention regions can be performed directly in MPEG compressed domain. Two types of information in compressed domain can be used (Jiang, 2005):

DCT Coefficients of Macro Block

Among DCT coefficients, DC coefficient is easy to get. It is the directly obtainable component of image block, and its value equals to eight time of the average value in block. It roughly reflects the brightness and color information.

Motion Vectors of Macro Block

Motion vectors correspond to sparse and coarse motion field. They reflect approximately the general motion information in the block.

With the motion vectors, a simple but effective four-parameter global motion model can be used to estimate simplified camera motions (zoom, rotate, pan and tilt):

$$\begin{cases} u = h_1 x + h_2 y + h_3 \\ v = h_2 x + h_1 y + h_4 \end{cases} \quad (1)$$

A common least-square fitting algorithm is imposed to optimize the model parameters h_1, h_2, h_3, and h_4. This algorithm recursively examines the error produced by the current estimate of the camera parameters and generates an outlier mask, consisting of macro-blocks with motion vectors not following the camera motion model. Then, the camera parameters are re-computed and new outlier mask are formed. This process iterates until the parameters are stabilized. As it operates directly on MPEG video, motion

Figure 1. Illustration of attention region and environment

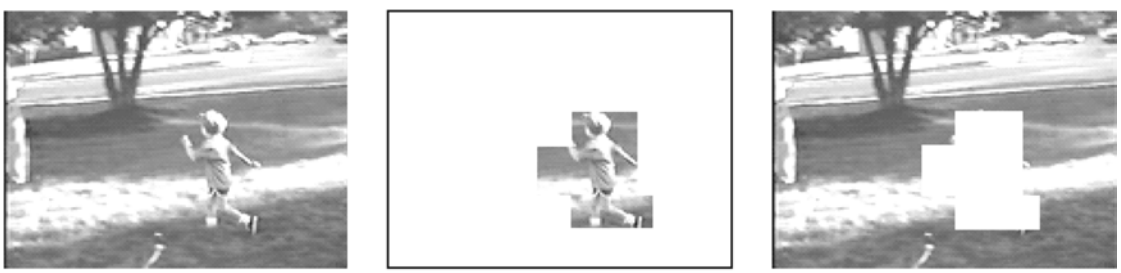
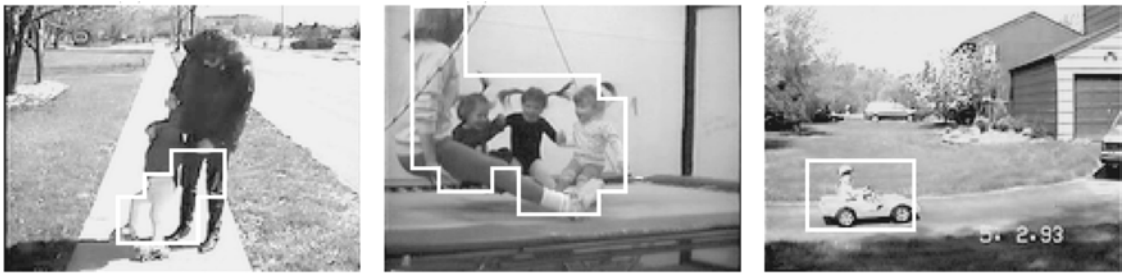


Figure 2. Detected attention regions



(a) (b) (c)

vector pre-processing has to be taken to get a dense motion vector field.

Some examples of attention regions detected are shown in Figure 2. All of them are macro-blocks characterized by different movement from global motion. Note that these regions are quite different from real objects. They could be a part of real object (kid's legs in Figure 2(a)) or several objects as a whole (woman and kids jumping on trampoline in Figure 2(b), kid and little car on the road in Figure 2(c)). Reasonably, one region drawing for viewer's attention is not necessary to be a complete accurate semantic object.

Time Weighted Model Based on Camera Motion

The detection of attention regions splits video content into two parts both spatially and temporally: attention regions and remnant regions. Two types of features can be used to represent the content of shot: one is the feature in attention regions, which emphasize the part attracting the audience; another is the features for other regions, which stress the global impression for the environment.

Color is often considered as the most salient visual feature that attracts viewer's attention, since the color appearance, rather than the duration, trajectory or area of region, is counted more in mind. Thus, an effective and inexpensive color representation in compressed domain - DC color histogram - is used to characterize each shot of video. In contrast to calculating an overall average color histogram, DC histograms of macro-blocks constituting both the attention regions and background are computed, respectively. The two types of histograms form a feature vector for each shot, which holds more information than a single histogram describing the global color distribution.

As each attention region and the background appears in several frames, the histograms along time can be accumulated to form a single histogram. Instead of the common average procedure, a camera motion based weighting strategy is used, giving different importance to histograms at different time. It is known that camera motion are always utilized to emphasize or neglect a certain objects or a segment of video, that is, to guide viewers' attentions. Actually, it can imagine a camera as a narrative eye. For example, a camera panning imitating an eye movement either to track an object or to examine a wider view of a scene, while a close-up of camera indicates the intensity of an impression. In order to reflect the different impression of viewers affected by camera movement, a camera attention model can be defined.

Camera movement controlled by photographer is useful for formulating viewer attention model. However, the parameters in the four-parameter global motion model do not represent the true camera move, scale and rotation angles. Thus, these parameters have to be first converted to real camera motion for attention modeling (Tan, 2000):

$$\begin{cases} S = h_1 + 1 \\ r = h_2 / (h_1 + 1) \\ L = \sqrt{p^2 + t^2} = \sqrt{h_3^2 + h_4^2} \,/ (h_1 + 1) \\ \theta = \arctan[t/p)] = -\arctan(h_4 / h_3) \end{cases} \quad (2)$$

where S indicates the inter frame camera zoom factor ($S > 1$, zoom in; $S < 1$, zoom out), r is the rotation factor about the lens axis, p and t are the camera pan and tilt factors, L is the magnitude of camera panning (horizontal and vertical) and θ is the angle.

The next step is mapping camera parameters to the effect they have on viewer's attention. Camera attention can be modeled with the help of the following assumptions (Ma, 2002):

(1) Zooming is always used to emphasize something. Zoom-in is used to emphasize the details, while zoom-out is used to emphasize an overview. The faster the zooming speed is, the more important the content focused is.

(2) Situations of camera panning and tilting should be divided into two types: Region-tracking (with attention region) and Surroundings-examining (without attention region). The former corresponds to the situation of a camera tracking a moving object, thus much attention is given to the attention region and little to the background. Camera motion of the latter situation tends to attract less attention since horizontal panning is often applied to neglect something (e.g. change for another view), if no attention region exists. The faster the panning speed is, the less important the content is. Additionally, unless a video producer wants to emphasize something, vertical panning is not used since it brings viewer's unstable feeling.

(3) If the camera pans / tilts too frequently or too slightly, it is considered as random or unstable motion. In this case, the attention is determined only by zoom factor.

In frames with attention region, viewer's attention is supposed to be split into two parts, represented by weight of background W_{BG} and weight of attention region W_{AR}.

$$W_{BG} = 1 / W_{AR} \tag{3}$$

$$W_{AR} = \begin{cases} S & L < L_0 \\ S(1 + L / R_L) & L \geq L_0 \end{cases} \tag{4}$$

where W_{AR} is proportional to S and enhanced by panning, L_0 is the minimal camera pan (panning magnitude less than L_0 is regarded as random), R_L is a factor controlling the panning power to affect attention degree. In this model, a value that is bigger than 1 means emphasis and a value that is smaller than 1 means neglect.

In frames without attention region, only background attention weight W_{BG} is computed.

$$W_{BG} = \begin{cases} S & L < L_0 \\ S /(1 + f(\theta) L / R_L) & L \geq L_0, \quad \theta < \pi/4 \\ S(1 + f(\theta) L / R_L) & L \geq L_0, \quad \theta \geq \pi/4 \end{cases} \tag{5}$$

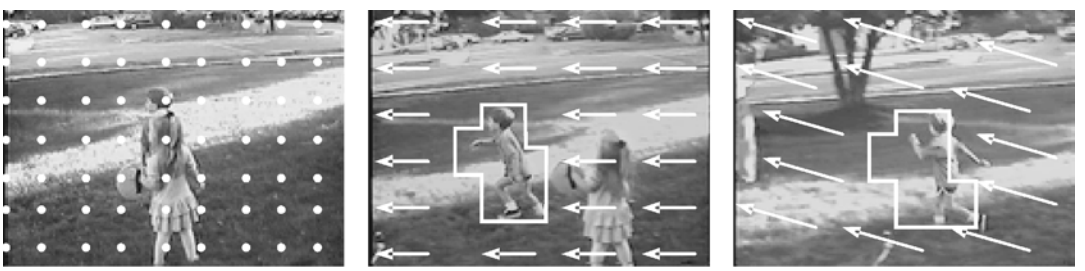
The attention degree decreases in situation of horizontal panning ($\theta < \pi/4$) and increases in situation of vertical panning ($\theta \geq \pi/4$). Function $f(\theta)$ is a function representing the effect of panning angle on decrease or increase rate. It becomes smaller while the panning angle gets nearer to $\pi/4$ because panning in this diagonal direction tends to have little effect on attention.

Figure 3 shows some examples of time-weighted modeling based on camera motion. Figures 3 (a), (b), and(c) are three frames extracted from the same shot of kids playing on lawn. Although they share a similar background, the background attention weight W_{BG} differs in the attention model according to different camera motions of the three frames. Figure 3(a) is almost stationary (without attention region), with attention just determined by zoom factor. Figure 3(b) is panning left tracking a running kid (with an attention region), thus it has less weight on background than on the attention region. Figure 3(c) shows also the region tracking but has a higher panning speed than that of Figure 3(b), thus it has an even smaller background weight. It is observed in Figure 3 that visual contents are spatially split into two attention parts, while temporally weighted by camera motion parameters.

Strategy for Shot Organization

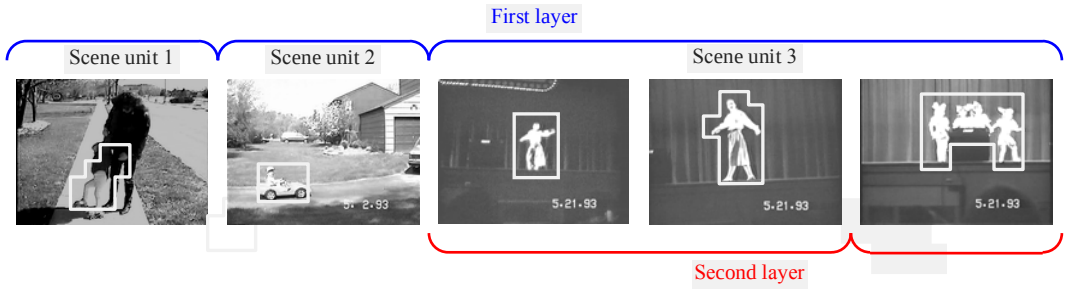
Using shot features and weights obtained above, a feature vector for each shot is composed. The visual similarity between two shots is then computed. In particular, similarities of background and of attention regions are computed separately, for example, by using the normalized histogram

Figure 3. Time weighted modeling based on camera motion



(a) (b) (c)

Figure 4. Two layer clustering of shots



intersection. Based on the similarity among shots, similar shots can be grouped. A two-layer shot organization strategy can be used. In the first layer, scene transition is detected. The place where both attention region and environment change gradually indicates the change of location and scenario. In the second layer, either attention regions or environment is changed (not both). This change can be the change of focus (different moving object in same environment) or the change of object position (same moving object in different environment). The change in the second layer is also called change of sub-scene.

One example of two-layer shot organization is shown in Figure 4. Five frames are extracted from five consecutive shots. White sashes mark detected attention regions. The first layer of organization clusters the five shots into three scene units, which are different both in attention regions and in environments. The last three shots in scene unit 3 of first layer can be further analyzed. The third and fourth shots in Figure 4 can be clustered together, which have similar background and similar attention regions; while the fifth shot has different attention region from the third and fourth shots. In short, the first layer clusters shots with the semantic event, while the second layer distinguishes different moving objects.

FUTURE TRENDS

Further researches are needed in organization of home videos as it is still in a preliminary stage

(1) To make the moving-object detection more robust (Pan, 2007).
(2) To extract visual features in multiple attention regions.
(3) To also make use of other media information, especially audio ones.

CONCLUSION

In this chapter, the organization of home video is discussed. Some concluding points are:

(1) Discovering the scene structure above shot level plays a key role in home video analysis.
(2) Assuming that viewer's attention is spatially divided into two parts: one for a dominant region in the frame that attracts most attention, the other for the rest background that expresses an overall impression, two metrics of shot similarity are developed for grouping shots.
(3) The detection of motion attention regions and the treating of attention regions and other remaining regions separately are important.
(4) The process taken directly in the compressed domain reduced the computational complexity.
(5) Considering a scene as a collection of semantically related and temporally adjacent shots, a high-level concept or story can be depicted and conveyed.

ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation under Grants NNSF-60573148 and the Ministry of Education under Grants SRFDP-20050003013.

REFERENCES

Achanta R., Kankanhalli M., Mulhem P. (2002). Compressed domain object tracking for automatic indexing of objects in MPEG home video. Proceedings of ICME, 61-64.

Babu R. V., Ramakrishnan K. R. (2002). Compressed domain motion segmentation for video object extraction. In: Proc. ICASSP 4: 3788-3791.

Chang S. F., Chen W., Meng H. J., Sundaram H., Zhong D. (1998). Fully automated content-based video search engine supporting spatiotemporal queries. IEEE Transactions on Circuits and Systems for Video Technology, 8(5): 602-615.

Gatica-Perez D., Loui A., Sun M. T. (2003). Finding structure in home videos by probabilistic hierarchical clustering. IEEE Trans. CSVT., 13(6): 539-548.

Ho C.C., Cheng W. H., Pan T. J., Wu J. L. (2003). A user-attention based focus detection framework and its applications. Proc. 2003 Joint Conference of the 4th IC on Information, Communications and Signal Processing and 4th Pacific-Rim Conference on Multimedia. 3(15): 1315-1319.

Huang S. H., Wu Q. J., Chang K. Y., et al. (2005). Intelligent home video management system. Proc. 3rd International Conference on Information Technology: Research and Education, 176-180.

Jiang F., Zhang Y. J. (2005). Camera attention weighted strategy for video shot grouping. SPIE, 5960: 428-436.

Lienhart R., Pfeiffer S., Effelsberg W. (1997). Video abstracting. Communications of ACM. 40(12): 54-62.

Lienhart R. (2000). Dynamic video summarization of home video. SPIE, 3972: 378-389.

Ma Y. F., Lu L., Zhang H. J., Li M. J. (2002). A user attention model for video summarization. Proc. ACM International Multimedia Conference and Exhibition, 533-542.

Pan Z., Ngo C. W. (2007). Moving-object detection, association, and selection in home videos. IEEE Transactions on Multimedia, 9(2): 268-279.

Tan Y. P., Saur D. F., Kulkarni S. R., et al. (2000). Rapid estimation of camera motion from compressed video with application to video annotation. IEEE-CSVT, 10(1): 133-146.

Taskiran C., Chen J. Y., Albiol A., Torres L., Bouman C. A., Delp E. J. (2004). ViBE: A compressed video database structured for active browsing and search." IEEE Transactions on Multimedia, 6(1): 103-118.

Wactlar H. D., Kanade, T., Smith M. A., Stevens S. M. (1996). Intelligent access to digital video: Informedia project. Computer, 29(5): 46-52.

Wang H., Divakaran A., Vetro A., Chang S. F., Sun H. (2003). Survey of compressed-domain features used in audio-visual indexing and analysis." Journal of Visual Communication and Image Representation, 14 (2): 150-183.

Wu P., Obrador P. (2005). Personal video manager: Managing and mining home video collections. SPIE, 5960, 775-785.

Zhai Y., Shah M. (2005). Automatic segmentation of home videos. ICME, 9-12.

KEY TERMS

Content-Based Image Retrieval (CBIR): A process framework for efficiently retrieving images from a collection by similarity. The retrieval relies on extracting the appropriate characteristic quantities describing the desired contents of images. In addition, suitable querying, matching, indexing and searching techniques are required.

Content-Based Video Retrieval (CBVR): A process framework for efficiently retrieving required clip from video. The retrieval relies on the organization of video and non-linear search techniques.

Content-Based Visual Information Retrieval (CBVIR): A combination of CBIR and CBVR.

Image Engineering: An integrated discipline/subject comprising the study of all the different branches of image and video techniques.

MPEG-7: An international standard named "Multimedia content description interface" (ISO/IEC 15938). It provides a set of audiovisual description tools, descriptors and description schemes for effective and efficient access (search, filtering and browsing) to multimedia content.

Object Segmentation: A process of image analysis. Its purpose is to extract the interesting region from image (corresponding to the interesting objects in scene).

Pan: One type of camera motion forms in capturing the scene image. It consists of the movement of camera around the vertical axis in the imaging plan.

Scene: In video analysis, it is composed of a series of consecutive shots that are coherent from the narrative point of view.

Tilt: One type of camera motion forms in capturing the scene image. It consists of the movement of camera around the horizontal axis in the imaging plan.

Organizational Aspects of Cyberloafing

Elisa Bortolani
University of Verona, Italy

Giuseppe Favretto
University of Verona, Italy

INTRODUCTION

The introduction of new technologies at workplaces causes the emergence of new organizational productivity threats. These threats are both inside and outside the organizations themselves. More often, organizations regret[1] programming and administrative errors; system and technical failures; sabotages; unauthorized accesses; disruption, manipulation, or loss of data and programs, not due to cyber-criminality (intrusions, employees' disloyalty, etc.); widespread virus; and issues caused by wireless devices. External threats, on the other hand, are more related to natural catastrophes (flooding, earthquakes, etc.), fires, industrial espionage, cyber-criminality, viruses, unfair competition, and physical damages to structures.

It is necessary for organizations to protect themselves from both intrusion attempts and employees' technology misuse. A United States survey[2] revealed that 35% of companies interviewed about suffered attacks in 2004 said that the prevalence was from insiders; on the other hand, only 26% revealed a prevalence of outsider attacks. Compared to the previous year, the trend was inverted. In 2003, in fact, insider attacks were around 14% and outsider attacks were about 23%. This, it is possible to think that insider threats will become more and more frequent and dangerous.

According to Radcliff (2004), internal data thefts are estimated to be 75% of total data thefts. An employee, for example, can copy and misappropriate a customer's database before passing it to the competitors. Another possible scenario is referred to as waste of efficiency caused by business e-mail abuse or Internet access misuse.

The FBI's Computer Crime Squad affirms that it is not necessary to blame corrupt or vindictive employees for all intrusion issues. Many problems, in fact, can be traced back to an improper use of IT business resources. Actually, for example, many companies that had put up with employees surfing the Internet for non-work-related activities for years now regretted Internet misuse, characterized by pornography, mp3, and illegal software downloading.

More than this, illicit software downloading and surfing insecure sites allow virus and malware introduction. This software, if installed on strategic machines, can make the company vulnerable. And so, costs are not limited to loss of business resources (e.g., working time), but are also related to damages caused by illicit and careless online employees' activities.

If, on one hand, the opportunity to work online helps in increasing several organizations' productivity (Anandarajan, Simmers, & Igbaria, 2000), on the other hand it causes an addition in number and level of risks. So, a lot of Internet access issues are related to information download (copyrighted software, offensive material, infected files, etc.), but the loss of productivity related to this habit does not seem secondary. In other words, without leaving their desks and without social control risk, the employees may, more easily than in the past, give themselves up to surfing the Net for non-work-related purposes.

Some years ago, a U.S. survey affirmed that 30-40% of daily business Internet traffic was attributed to surfing the Net for personal purposes.[3] In another research, carried out in 2001,[4] 51% of Italian employees affirmed access to non-work-related sites — more than English (44%), German (41%), and French (29%). And finally, another U.S. survey[5] shows that the average time spent online by respondents who admitted to using the Internet for personal purposes (58%) is about three-and-a-half hours per week.

Despite many people already having Internet access at home, about half of all online shopping and 70% of the global pornographic traffic[6] are registered during working time.[7] The favorite activities for surfing the Net are: holiday booking (52%), culture (42%), hobbies (41%), shopping (28%), sports (30%), and job searching.

The average user spends about two hours per day online, and 31% is for non-work-related surfing.[8] Seventy percent of employees admit either visiting "for adult" sites or sending personal e-mails during work time, 64% also send offensive or politically incorrect messages, and 57% admit surfing online decreases their own productivity.

A survey conducted in 2004[9] reports that of 3,245 respondents belonging to 750 employers, 40% answered to spending 40% of his or her working time cyberloafing, with an increase of one hour per day in the last year.

Organizational expenses in loss of productivity are estimated at billions of dollars per year (Greengard, 2000; Gordon, Loeb, Lucyshyn, & Richardson, 2005). After all, as evidenced by IDC Research (2000), U.S. and European Internet users seem to spend more time online when they are at work than when they are at home. This is probably due

to two characteristics of the workplace: perceived privacy and higher speed to link.

This new form of productive deviance, defined by Lim (2002) as *cyberloafing*, consists of business Net access during working time to surf for personal ends and/or to manage personal e-mails.

Siau, Nah, and Teng (2002) identify 11 categories of Internet abuses: general e-mail abuses, unauthorized usage and access, copyright infringement/plagiarism, newsgroup postings, transmission of confidential data, pornography, hacking, non-work-related download/upload, leisure use of the Internet, usage of external ISPs, and e-moonlighting (side jobs).

Especially, focusing on e-mails, Whitty and Carr (2006) affirm that cyberloafing is not always an intentional choice but sometimes is induced by other people. Ninety percent of their respondents, in fact, perceive chain e-mail, sent by friends or known people in any case, as more unpleasant than e-mail spam, and 17% perceive joke e-mails as objectionable. The modality to manage e-mails is increasingly an emerging issue and requires new and socially shared rules.

BACKGROUND

Lim (2002) categorizes cyberloafing as productive deviance (see Robinson & Bennett, 1995) but, for its consequences, it may also be included in:

- Property deviance (Hollinger & Clark, 1983);
- personal aggression (psychological harassment, e.g., sexual or racist e-mails); and
- political deviance, defined as the use of incorrect means to put someone at a political disadvantage in comparison with someone else (see Robinson & Bennett, 1995).

Mastrangelo (2002) suggests three dimensions below a counterproductive computer use. Each of them corresponds to the necessity of:

a. Social linking (personal e-mail, instant messaging, chat);
b. doing an errand; or
c. indecent behavior.

Lim (2002) proposes a cyberloafing explanation model starting from considerations about treatment equity (see Adams, 1965; Foa & Foa, 1976) and distributive justice (Deutsch, 1985; Skarlicki & Folger, 1997) in the organizations. She identifies in neutralization (Sykes & Matza, 1957) the theory construct to explain cyberloafing. Neutralization, in her opinion, is the individual attempt to rationalize a situation to convince himself and the others to be right and that deviant behavior is understandable. This mechanism of rationalization aims at building and preserving one's image.

Through neutralization, deviant behavior does not appear as a revenge to the employee. This feeling, in fact not acceptable for the person, is mediated from a rational explanation. After all, cyberloafing becomes a way to reestablish a trade-off in the person-organization relationship. Loss of equity (in terms of economic, relational, or symbolic treatment) is restored through a cyberloafing behaviour, so the individual feels rewarded for the time and energy he or she spent for the company that was not recognized.

Anandarajan et al. (2000), instead, describe both organizational and personal factors involved in Internet misused at work. They applied Fishbein and Ajzen's (1975) Theory of Reasoned Action (TRA) to Internet use at work. The authors affirm that the use of the Net is influenced by perceptions, personal attitudes, and social influences. TRA is extended by authors to the Technology Acceptance Model (TAM), which focuses on information technologies' perceived usefulness. In TAM, the factors that motivate a person to use a computer could be categorized into two groups: extrinsic motivators (perceived advantages, social pressure, etc.) and intrinsic motivators (playfulness, distraction, etc.). The proposed model shows four kinds of multidimensional variables. Research outcomes refer a more frequent, easier access and more time spent online by men than by women. Moreover, the ability to use the Internet and the Web are related to improvement in job characteristic perception (significance, autonomy, heterogeneity, job control, etc.). Playfulness can lead to perception of job characteristics improvement, more job satisfaction, and more global productivity. Actually playfulness, is a double-edged weapon; in fact, it can lead easily to Internet misuse with negative consequences in terms of increasing loss of time, necessity of redoing work because of a loss of accuracy, and a longer period of time in order to complete a task. These factors contribute to ineffectiveness and loss of productivity. On the other hand, social pressure and organizational support are associated to a kind of intimidation, which implies a lower use of the Internet. It suggests that management's commitment and support in the use of the Net can reduce the abuse. In addition, employees with high-structured tasks are less involved in improper use of the Internet at their workplace. The same result has emerged for people who have a low task variability. These considerations imply that employees with less structured tasks have a higher use of the Internet for personal scopes.

Henle and Blanchard (2005) suppose that, at first, employee cyberloafing is a modality of stress coping and that they do it only if the perception of sanctions by the organization is low. In fact, cyberloafing increases when there are no sanctions, but contrary to their initial expectations, it increases when workload is low.

It is interesting to quote the research of Lara, Tacoronte, and Din (2006). Their study shows that the variable leader

physical proximity (LPP) is an antecedent positively associated to fear of formal punishment (FFP) and perceived organizational control (POC). This last factor decreases cyberloafing, whereas Fear of Formal Punishment increases it. Fear of Formal Punishment and Perceived Organizational Control moreover mediate the relationship between Leader Physical Proximity and cyberloafing.

So, is employee monitoring a solution? Nowadays, solutions to cyberloafing are mainly based on monitoring of employee surfing and on filtering Web sites considered to be inappropriate. Chalykoff and Kochan (1989) showed that satisfaction of employees with computer-aided monitoring has a strong impact on job satisfaction. Recently, moreover, research revealed that employees perceive monitoring software as invasive, and this decreases their job satisfaction (Stanton & Weiss, 2000). With regard to this, Urbaczewski and Jessup (2000) affirm that subjective motivation has an important role both in productivity and in performance quality, and in monitoring satisfaction. Employees, in fact, accepted monitoring when it was used to give a feedback about their job quality; when, on the contrary, monitoring was used as a way of control, as in Web monitoring, the impact on organizational climate was negative.

Davis, Flett, and Besser (2002) affirm that an alternative way to cope with cyberloafing is to prevent the phenomenon, not to apply sanctions. So, authors suggest that attention be paid to some individual characteristics during the selection process.[10] For example, Wyatt and Phillips (2005), contrarily to expectations and previous literature, found that extraverted[11] people used and abused the Internet in the workplace as a means of socializing and developing relationships, and moreover that disagreeable people spent more time on the Internet than agreeable[12] people.

Contextually many companies are attempting to solve the problem by internal policies that clarify in detail permitted, tolerated, forbidden, and sanctioned behaviors. Often, in fact, the employee does not perceive the real risks to which he or she exposes his or her company by surfing the Web.

The flexibility favored by the new communication and information technologies imposes to give strong responsibilities to human resources in terms of autonomy and control on their own job, and this is possible only if appropriate employee training is provided. Although there are many discordant opinions about policies' preventive usefulness (Young & Case, 2004),[13] we think it is necessary to make employees responsible for an awareness use of technologies. These policies, if they include clear behavioral rules, may be a good first step to helping employees build a cognitive map to make head of their company. The lack of clearness about the existence of a specific managerial way to administer Web surfing and e-mail, in fact, can become a large organizational cost both in terms of dissatisfaction and distress (with the subsequent turnover, absenteeism, etc.; see Favretto, 1994).

It is interesting to report an important result of Whitty's (2004) study: 19% of her sample (N=524) ignored if their company had a policy, and 17% did not know if the Web sites were filtered. Other relevant data is that 19% of the respondents state that they do not have a policy, according to 17% of Greenfield and Davis's (2002) research. In reality, our research (Bortolani, 2005), conducted in Italy (N=190), shows that 25% of the respondents affirm that their companies do not have any forms of monitoring, and 31% state that they do not know if any type of monitoring exists. So, a good policy, shared with all employees (and if possible, different according to the various professional families), can give at least the awareness about this issue and, contextually, give the right to the company to contest the debit to the defaulting employee.

FUTURE TRENDS

In the future, it would be interesting to study in depth personality characteristics as well as organizational factors that favor cyberloafing. Moreover it could be interesting to investigate what kind of professional figures, what kind of workplaces, and what moments of the working day are more exposed to this activity. If cyberloafing, as shown in Henle and Blanchard (2005) and Bortolani (2006) surveys, reveals itself as an expression of employment unease (boredom and monotony, distress and social and physic isolation), it would be useful to remove favoring conditions through *job redesign* techniques. This process allows the change of tasks perceived as poor, boring, and deprived of meaning by employees. In addition, we report another consideration: in Bortolani's (2006) study the variable *loneliness/physic and social isolation* and the variable *mistrust organization and perception of unfair treatment* are more often chosen by public sector employees than by private sector employees.

CONCLUSION

The article analyzes cyberloafing (also known as cyberslacking or cyberbludging), a new kind of productive deviance in organizations. The phenomenon concerns Internet misuse at work, surfing the Net, sending personal e-mails, and other non-work-related activities during work time. Loss of productivity, damage to the organization's image, and lawsuit problems are some of the cyberloafing-related issues.

Managerial actions are necessary in order to lead human resources to the phenomenon containment. Potentially, all employees that have a computer linked to the Internet can easily engage themselves in cyberloafing activities. Research, carried out in a manufacturing industry (LaPlante, 1997) with 400 employees, pointed out that in a typical workday (eight hours), more than 250,000 Web sites are visited, 90%

of which are non-work-related. If, on one hand, it is very difficult to estimate the real entity of the phenomenon, on the other it is interesting to understand the meaning that employees assign to this activity.

Bortolani (2006), for example, found that younger employees spend more time cyberloafing (more than five hours per week) than older employees (less than one hour per week), and they affirm to do it when they are distressed and frustrated for personal reasons (not due to the work environment). Moreover, younger employees (less than 30 years old) state that cyberloafing increases perceptions of both job productivity and job satisfaction (this last data according to Anandarajan et al., 2000; Mastrangelo, 2002). Despite that we are talking about perceptions, we think this is useful information to plan actions more aimed at this definite target (e.g., during work socialization).

REFERENCES

Adams, J.S. (1965). Inequity in social exchange. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 267-297). New York: Academic Press.

Anandarajan, M., Simmers, C.A., & Igbaria, M. (2000). An exploratory investigation of the antecedents and impact of Internet usage: An individual perspective. *Behavior and Information Technology, 19*(1), 69-85.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Bortolani, E. (2005, December 1-2). Nuove forme di devianza produttiva nelle organizzazioni: L'assenteismo virtuale. *Proceedings of the 2nd Conference of the Italian Chapter of AIS: Organizing Information Systems — Evolution of Studies in the Field*, Verona.

Bortolani, E. (2006). L'utilizzo di Internet dai luoghi di lavoro per scopi personali: Devianza produttiva o strategia di coping? In *Esperti e utenti a confronto: Significati psico-sociali e aspetti organizzativi dell'interazione uomo-tecnologia*, Unpublished PhD Thesis.

Chalykoff, J., & Kochan, T. (1989). Computer-aided monitoring: Its influence on employee job satisfaction and turnover. *Personnel Psychology, 42,* 807-834.

Davis, R.A., Flett, G.L., & Besser, A. (2002). Validation of a new scale for measuring problematic Internet use: Implications for pre-employment screening. *Cyberpsychology & Behavior, 5*(4), 331-345.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues, 31,* 137-149.

Favretto, G. (1994). *Lo stress nelle organizzazioni*. Bologna: Il Mulino.

Foa, E.B., & Foa, U.G. (1976). Resource theory and social exchange. In J.S. Thibaut, J. Spence, & R. Carson (Eds.), *Contemporary topics in social psychology.* Morristown, NJ: General Learning Press.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention, and behavior: An introduction to theory and research.* Ontario: Addison-Wesley.

Gordon, L.A., Loeb, M.P., Lucyshyn, W., & Richardson, R. (2005). *2005 CSI/FBI computer crime and security survey.* Retrieved from http://www.cpppe.umd.edu/Bookstore/Documents/2005CSISurvey.pdf

Greenfield, D.N., & Davis, R.A. (2002). Lost in cyberspace: The Web @ work. *CyberPsychology and Behavior, 5*(4), 347-353.

Greengard, S. (2000). The high cost of cyberslacking. *Workforce, 79*(12), 22-24.

Henle, C.A., & Blanchard, A.L. (2005). Cyberloafing as a coping method: Relationship between work stressors, sanctions, and cyberloafing. *Proceedings of the Annual Meeting of the Academy of Management,* Honolulu, HA.

Hollinger, R.C., & Clark, J. (1983). *Theft by employees.* Lexington: Lexington Books.

LaPlante, A. (1997). Start small, think infinite. *The Premier 100 Supplement to Computerworld,* (February 24), 24-26.

Lara, P.Z.M., Tacoronte, D.V., & Ding, J.M.T. (2006). Do current anti-cyberloafing disciplinary practices have a replica in research findings? A study of the effects of coercive strategies on workplace Internet misuse. *Internet Research: Electronic Networking Applications and Policy, 16*(49), 450-467.

Lim, V.K.G. (2002). The IT way of loafing on the job: Cyberloafing, neutralizing and organizational justice. *Journal of Organizational Behavior, 23,* 675-694.

Mastrangelo, P.M. (2002). *The misuse of work computers: theory, data and policy.* Retrieved from http://www.ipmaac.org/mapac/newsltrs/winternl2002.pdf

Radcliff, D. (2004). What are they thinking? *Network World,* (March 3).

Robinson, S.L., & Bennett, R.J. (1995). A typology of deviant workplace behaviors: A multidimensional scaling. *Academy of Management Journal, 38,* 555-572.

Skarlicki, D.P. & Folger, R. (1997). Retaliation in the workplace: The roles of distributive, procedural and interactional justice. *Journal of Applied Psychology, 82,* 434-443.

Siau, K., Nah, F.F., & Teng, L. (2002). Acceptable Internet use policy. *Communications of the ACM, 45*(1), 75-79.

Sykes, G., & Matza, D. (1957). Techniques of neutralization: A theory of delinquency. *American Sociological Review, 22,* 664-670.

Stanton, J.M., & Weiss, E.M. (2000). Electronic monitoring in their own words: An exploratory study of employees' experiences with new types of surveillance. *Computers in Human Behavior, 16,* 423-440.

Urbaczewski, A., & Jessup, L.M. (2000). *An examination of the effects of electronic monitoring of employee Internet usage.* Retrieved from *portal.acm.org/citation.cfm?id=931434*

Whitty, M.T. (2004). Should filtering software be utilized in the workplace? Australian employee's attitudes towards Internet usage and surveillance of the Internet in the workplace. *Surveillance and Society, 2*(1), 39-54.

Whitty, M.T., & Carr, A.N. (2006). New rules in the workplace: Applying object-relations theory to explain problem Internet and email behavior in the workplace. *Computers in Human Behavior, 22,* 235-250.

Wyatt, K., & Phillips, J.G. (2005, November 23-25). Internet use and misuse in the workplace. *ACM International Conference Proceeding Series, Proceedings of OZCHI 2005* (vol. 122, pp. 1-4), Canberra, Australia.

Young, K.S. (1998). *Caught in the Net: How to recognize the signs of Internet addiction and a winning strategy for recovery.* New York: John Wiley & Sons

Young, K.S., & **Case**, C.J. (2004). Internet abuse in the workplace: New trends in risk management. *CyberPsychology & Behavior, 7(1), 105-111.*

KEY TERMS

Cyberloafing: The activity of surfing the Net, sending and reading e-mails in the workplace for personal ends during working time. Also called *cyberslacking* or *cyberbludging.*

Job Satisfaction: The pleasure feeling derived from one's own job and from the fact that the job meets some personal needs.

Job Redesign: Process that aims to reorganize the elements of work to enrich the job and to allow an employee to best match with the job.

Monitoring Software: Software used by employers to record computer activities of their employees; examples include e-mail, chat, instant messaging, and visited Web sites.

Netiquette: A set of non-written behavioral rules developed by the Internet community to facilitate social interactions and avoid reproaches.

Policy: A set of behavioral and procedural written rules that a company drafts as guidelines for its employees.

Social Control: Mechanism, present in all societies, that regulates individual behavior to lead to conformity with social rules. This mechanism — including the influence of family, moral values, beliefs, and so forth — aims to guarantee social order.

Theory of Reasoned Action (TRA): Model theorized by Fishbein and Ajzen (1975) that says human behavior springs from intensions and that these intensions are shaped by positive or negative attitudes towards something. This theory focuses on three components: individual attitude, reference group influence, and subjective propensity to allow that external influences may affect one's own choices.

ENDNOTES

1 SPACE Bocconi, at *http://www.clusit.it/infosecurity2004/mansutti.pdf*

2 In "Internal threats concern financial institutions," *Internal Auditor,* August 2005.

3 International Data Corporation Research, 2000

4 Websense Data, Web@work2001

5 Websense Data, Web@work2005

6 Sex Tracker Data

7 Internet use statistics, *www.webspy.com,* 2002

8 Angus Reid Group's data, 2000, at *http://www.webspy.com/files/publications/InternetUseStatistics.pdf*

9 At *http://www.management-issues.com/display_page.asp?section=research&id=1417n*

10 Online Cognitive Scale (OCS), for example, is a questionnaire that aims at surveying a possible problematic relationship with technology, studying the most used Internet applications. Young (1998), in fact, affirms that Internet Addiction Disorder reveals in the preference for interactive applications that provide synchronous communication (e.g., chat lines, instant messaging), rather than applications that provide asynchronous communication (e.g., e-mail, newsgroups).

11 Extraversion concerns traits of warmth, gregariousness, assertiveness, activity, excitement-seeking, and positive emotions (Anastasi & Urbina, 1997).

12 Agreeableness concerns trust, altruism, compliance, and modesty (Anastasi & Urbina, 1997).

[13] Young and Case (2004) realized that only 40% of human resources managers respondent to an online survey believed that their policies were effective.

Organizational Assimilation Capacity and IT Business Value

Vincenzo Morabito
Bocconi University, Italy

Gianluigi Viscusi
University of Milano, Italy

INTRODUCTION

IT business value represents important outcomes in firms (Banker & Kauffman, 2004; Gable, Darshana, & Chan, 2003; Ravichandran & Chalermsak Lertwongsatien, 2005) whereas information systems (IS) integration represents a relevant amount of the IT spending. Notwithstanding, while most firms are making major investments in information technology, particularly in information systems integration (e.g., ERP and data warehouse solutions), not all of them apply IT effectively in their business activities (Brynjolfsson, McAfee, Zhu, & Sorell, 2006; Dehning & Stratopoulos, 2003; Jason, Vijay, & Kenneth, 2003) obtaining IT business value and organizational competitive advantage.

This research is based on an integrative model of IT business value, aiming to evaluate the mediating effect of an "IT organizational assimilation capacity" between IS integration and organization competitive advantage. Taking into account the theoretical premises that IT business value is generated by the exploitation of both IT and organizational resources, we develop a research model and propose two research hypotheses.

The model and the related hypotheses are based on a large-scale sample survey (Francalanci & Morabito, 2006). The responses were obtained from 466 CIOs and senior business executives, who were members of the firms' top management teams in Italian companies.

BACKGROUND

The term IT business value is commonly used to refer to the organizational performance impacts of IT resources at both the intermediate process level and the organization-wide level, comprising both efficiency and competitive impacts (Melville & Kraemer, 2004). In fact, IT resources generate business value when they are "assimilated" as a routinized element of firms' value-chain activities and business strategies (Aral, Brynjolfsson, & Wu, 2006).

Researchers have employed several theoretical paradigms in examining the organizational performance impacts of IT, including microeconomics (Brynjolfsson & Hitt, 2003; Tanriverdi, 2005; Wade & Hulland, 2004), industrial organization theory (Belleflamme, 2001; Mahnke, Overby, & Vang, 2005), sociology and socio-political paradigms (Chatfield & Yetton, 2000; Devaraj & Kohli, 2003; Hatami, Galliers, & Huang, 2003) and, finally, strategic perspective (Bharadwaj, 2000; Caldeira & Ward, 2003).

Analyzing this stream of studies by focusing on the "focal firm", we can summarize that if the right IT is applied and assimilated within the right business process, the IT application improves processes and organizational performance/competitiveness, conditional upon appropriate investments in complementary organizational resources. In particular, the competitive environment, including industry characteristics and trading partners, as well as the macro environment are relevant to IT business value generation (Melville et al., 2004). Our research is focused at firm level, where IT business value is generated by the deployment of IT resources (including both technology IT resources and human IT resources) through a process that involves the deployment of complementary organizational resources within business processes. Referring to technological IT resources, there are studies that aggregate diverse technological IT resources into a single measure, and studies that examine specific information systems and types of IT.

In the first case, scholars use large-sample data sets, finding support for a positive association between aggregate measures of the technological IT resource and organizational performance (Bharadwaj, 2000; Kohli & Devaraj, 2003). The idea that the technological IT resource confers economic value is preserved when considering alternative econometric specifications, assumptions, data sets, and time frames (Aral et al., 2006). Early evidence indicates both a positive impact (Brown et al., 2000) and no association between technological information resources (TIR) and sustainable performance advantages (Powell & DentMicallef, 1997). Whereas others point out the relevance of managerial IT

skills and culture in order to confer a competitive advantage (Hafeez, Zhang, & Malak, 2002; Mata, Fuerst, & Barney, 1995; Zahra, Hayton, & Salvato, 2004).

Considering the second research approach introduced earlier, scholars have examined specific information systems and types of IT. Several studies find a positive impact on cost reduction, for example, in the context of a production data management system in the clothing industry (Tatsiopoulos, Ponis, Hadzillias, & Panayiotou, 2002), and supply chain management in the food industry (Hill & Scudder, 2002). Enterprise resource planning systems are associated with higher financial market valuation, although short-term effectiveness is reduced after implementation (Hitt & Wu, 2002).

Focusing on organizational performance, the firm's absorptive capacity (Cohen et al., 1990) plays a strategic role, exploiting IT business value by transforming into performance the IT-driven change of the organization, and mediating the assimilation of new external knowledge (Malhotra, Gosain, & El Sawy, 2005; Zahra & George, 2002). In this context, there are studies that assess the degree to which complementary organizational resources mediate organizational performance/competitive advantage impacts, and studies that analyze the highly contextual value generation process.

In the first category, mainly by using quantitative empirical methods applied to large samples of firms, the findings confirm that firms must not only customize technological systems and deploy and maintain them, but also must manage teams of IT and non-IT resources, together generating greater value than they do alone (Brynjolfsson & Hitt, 2000). Non-IT resources include organizational practices and structures that complement the functions of information systems. Empirical analyses discover decentralization of decision authority in firms with higher levels of IT, these firms having disproportionately higher market valuations (Brynjolfsson, Hitt, & Yang, 2002). Another set of organizational resources that may be complementary to IT are firm characteristics, such as worker composition, size, financial condition, and culture. Francalanci and Galal (1998) find that IT business value, as measured by productivity, differs according to employee category: firms with higher IT investment that have also decreased their clerical and professional ranks have higher productivity.

Focusing on the second category of studies introduced previously, the well known example from Clemons and Row (1988) documents IT-enabled efficiencies at McKesson, where customers benefit from rationalizing operations in preparation for the new order entry and distribution system adopted by McKesson. Other case and field studies examine, for example, package delivery (Williams & Frolick, 2001). The co-introduction of IT and complementary organizational changes may not result in immediate success, due to adjustment costs, learning, and other factors (Melville et al., 2004).

In summary, empirical evidence supports the claim that the technological IT resource (TIR) has economic value (Kohli et al., 2003), and that complementary organizational resources interact with IT in the processes of IT assimilation value generation.

Taking these findings into account, *we propose that competitive advantage can result from a specific information systems characteristic* (also realized by mixing different technological IT resources, i.e., ERP and EAI as in our case) *that is, in our case, the "information systems integration"*. But how "IS integration" improves organizational performance? In line with this question, we propose that "IS integration" effect is mediated by a specific complementary organizational resource: the "IT organizational assimilation capacity".

ORGANIZATIONAL ASSIMILATION CAPACITY AND IT BUSINESS VALUE

Based on theoretical proposition that IT business value is generated by the deployment of IT and complementary organizational resources, we develop a research model and propose two hypotheses. In particular, taking into account the role of firm's absorptive capacity (Cohen et al., 1990; Malhotra et al., 2005; Zahra et al., 2002), we propose that the IS integration business value is generated by the mediating effect of the IT organizational assimilation capacity. Although it is possible to apply IT for improved organizational performance with few organizational changes (McAfee, 2002), the IT business value is often generated by the deployment of IT and complementary organizational resources within business processes (Melville et al., 2004). In addition, firm-specific organizational resources tend to be tacit and deeply embedded in the organization's social fabric and history; what is not understood is what specific resources qualify the complementary effect, under what conditions, and how are the attributes of complementary resources related to business process and organizational performance impacts. Due to these issues, we propose a multi-variables concept defined "IT organizational assimilation capacity". We introduce four distinct groups of constructs that represent the elements of an IT organizational assimilation capacity: *training* (or *knowledge*) *orientation*, *change orientation*, *flexibility orientation* and *process orientation*. Further, we point out that the presence of IT organizational assimilation capacity largely influences how well the organization assimilates the business potential of IT, that is, in our case, the "information system integration". This "assimilation capacity" includes devising new IT-ways in which knowledge is distributed and accessed throughout the organization, tasks are accomplished, and so forth. As a result, IT organizational assimilation capacity may amplify or enhance the organizational effects of IT in general and information system integration in particular. Indeed, we propose that:

Hypothesis 1: *Stronger IT organizational assimilation capacity leads to a higher level of firm competitiveness.*

In particular, we argue that IT organizational assimilation capacity mediates the effect of information system integration on firm competitive advantage. Information systems integration implies that all functional information systems exchange data each others and that functional activities are interrelated and should be handled together. This "system integration" should be coupled with an "organizational integration". In particular, the achievement of an information system integration needs to be "assimilated" by the organization. In order to realize an effective information system integration, firms should work by process, be open to changes required by the integration, have skilled workers, and be flexible in its operations. Thus, we propose that:

Hypothesis 2: *Stronger information systems integration leads to a higher level IT organizational assimilation capacity.*

To fully account for the differences among organizations, we include *organization size* as control variable. We use the number of employees as measure of organization size. Organization size is an important control variable for another reason. IT/IS vendors and systems integrator could have to define different implementation programs in larger client or SME ones. IT organizational assimilation capacity (ITACP) refers to the firm's ability to identify, assimilate, and exploit the business potential from IT/IS solutions or IT/IS characteristics, in our case, the information system integration. The IT organizational assimilation capacity is a strategically valuable capability because it is path dependent, firm-specific, and socially embedded. It lowers the transaction costs of adopting and using IT/IS solutions, and its focus is on the individual firms' internal capabilities.

To capture the readiness of the organization towards assimilating IT/IS solutions or characteristics, we have identified four distinct groups of dimensions that represent the elements of the IT organizational assimilation capacity*: training (or knowledge) orientation*, *change orientation*, *flexibility orientation,* and *process orientation*. These dimensions should serve as the bases upon which organizations can be differentiated in their ability to identify (training orientation), assimilate (change orientation and process orientation), and exploit (flexibility, process and change orientation) IS integration.

Focusing on sustainable competitive advantage (SCA), the actual term emerged in 1985, when Porter discussed the basic types of competitive strategies firms can possess (low-cost or differentiation) to achieve SCA. However, no formal conceptual definition was presented by Porter in his discussion. Further, Barney (1991) offers the following definition: "A firm is said to have a sustained competitive advantage when it is implementing a value creating strategy not simultaneously being implemented by any current or potential competitors and when these other firms are unable to duplicate the benefits of this strategy". Taking these definitions into account, we define SCA as the prolonged benefit of implementing some unique value-creating strategy not simultaneously being implemented by any current or potential competitors along with the inability to duplicate the benefits of this strategy (in our case, implementing an information system integration). Concentrating on "prolonged benefit" we refer to superior organizational performances that, in our study, are designed as four items scale. Three items are referred to as a subjective measure of financial and economic performance over the previous 3-year period; one item is referred to as subjective perception of future sustainability of organization superior performances themselves.

In using subjective performance measures, given the senior executives involved (Francalanci et al., 2006), we assumed that respondents had sufficient perspective and information to assess their firms' performances relative to competitors. Subjective measures have been widely used in organizational research (Dess, 1987; Powell et al., 1997), and are often preferred to economic and financial statement data, since firms may adopt varying accounting conventions in areas such as inventory valuation, depreciation, and officers' salaries.

FUTURE TRENDS

Future research should test alternative measures of assimilation capacity by focusing on key business processes or units. We also acknowledge that the relationship between IS Integration, assimilation capacity, and competitive advantage may unfold through cyclical causal relationships. For example, a high competitive advantage may facilitate further developments of assimilation capacity and IS integration. Future trends must address these cycles of causal relationships.

CONCLUSION

Our thesis is that IS integration generates an organizational competitive advantage through the mediation effect the "IT organizational assimilation capacity". This capacity encompasses four distinct groups of dimensions: training (or knowledge) orientation, change orientation, flexibility orientation, and process orientation. These dimensions should serve as the basis upon which organizations can be differentiated in their ability to obtain a competitive advantage from investing in IS integration. Further, our data suggest that information system integration investments do not merge themselves automatically with others' organization resources, but they require what we define as "IT organizational assimilation capacity". From a theoretical point of view, this study contributes theory-based conceptual synthesis and empirical evidence to an IT literature still dominated by anecdotes and consultants' IT implementation models. Finally, we can sustain that "organization matters" in generating IT business value from IT/IS investments

REFERENCES

Aral, S., Brynjolfsson, E., & Wu, D. J. (2006). *Which came first, IT or productivity? The virtuous cycle of investment and use in enterprise systems*. Working paper. SSRN. Retrieved from http://ssrn.com/abstract=942291

Banker, R. D., & Kauffman, R. J. (2004). The evolution of research on information systems: A fiftieth-year survey of the literature in management science. *Management Science, 50*(3), 281-298.

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99-120.

Belleflamme, P. (2001). Oligopolistic competition, IT use for product differentiation and the productivity paradox. *International Journal of Industrial Organization, 19*(1-2), 227-248.

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly, 24*(1), 169-196.

Brynjolfsson, E., & Hitt, L. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives, 14*(4), 23-48.

Brynjolfsson, E., & Hitt, L. (2003). Computing productivity: Firm level evidence. *Review of Economics and Statistics, 85*(4), 793-808.

Brynjolfsson, E., Hitt, L., & Yang, S. (2002). *Intangible assets: Computers and organizational capital*.Unpublished manuscript.

Brynjolfsson, E., McAfee, A., Zhu, F., & Sorell, M. (2006). *Scale without mass: Business process replication and industry dynamics*. SSRN.

Caldeira, M. M., & Ward, J. M. (2003). Using resource-based theory to interpret the successful adoption and use of information systems and technology in manufacturing small and medium-sized enterprises. *European Journal of Information Systems, 12*(2), 125-139.

Chatfield, A. T., & Yetton, P. (2000). Strategic payoff from EDI as a function of EDI embeddedness. *Journal of Management Information Systems, 16*(4), 195-224.

Clemons, E. K., & Row, M. C. (1988). McKesson drug company: A case study of economost: A strategic information system. *Journal of Management Information Systems, 5*(1), 35-50.

Cohen, W., & Levinthal, D. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quaterly, 35*, 128-152.

Davenport, T. H. (1993). *Process innovation: Reengineering work through information technology*. Cambridge, MA: Harvard Business School Press.

Dehning, B., & Stratopoulos, T. (2003). Determinants of a sustainable competitive advantage due to an IT-enabled strategy. *Journal of Strategic Information Systems, 12*, 7-28.

Dess, G. (1987). Consensus on strategy formulation and organizational performance: competitors in a fragmented industry. *Strategic Management Journal, 8*(3), 259-277.

Devaraj, S., & Kohli, R. (2003). Performance impacts of information technology: Is actual usage the missing link? *Management Science, 49*(3), 273-289.

Francalanci, C., & Galal, H. (1998). Information technology and worker composition: Determinants of productivity in the life insurance. *MIS Quarterly, 22*(2), 227.

Francalanci, C., & Morabito, V. (2006, October 26-27). *IS integration and business performance: The mediation effect of organizational absorptive capacity in SMEs.* Paper presented at the ItAIS 2006, Università Bocconi, Milano.

Gable, G. G. a. S., Darshana, & Chan, T. (2003). *Enterprise systems success: A measurement model.* Paper presented at the Twenty-Fourth International Conference on Information Systems, Seattle, USA.

Hafeez, K., Zhang, Y., & Malak, N. (2002). Core competence for sustainable competitive advantage: A structured methodology for identifying core competence. *IEEE Transactions on Engineering Management, 49*(1), 28-35.

Hasselbring, W. (2000). Information system integration. *Communications of the ACM, 43*(6), 33-38.

Hatami, A., Galliers, R. D., & Huang, J. (2003). *Exploring the impacts of knowledge (re)use and organizational memory on the effectiveness of strategic decisions: A longitudinal case study*. Proceedings of the 36th Annual Hawaii International Conference on System Sciences (HICSS'03), IEEE Computer Society Washington, DC, USA.

Hill, C. A., & Scudder, G. D. (2002). The use of electronic data interchange for supply chain coordination in the food industry. *Journal of Operations Management, 20*(4), 375-387.

Hitt, L. M., & Wu, X. Z. (2002). Investment in enterprise resource planning: Business impact and productivity measures. *Journal of Management Information Systems, 19*(1), 71.

Jason, D., Vijay, G., & Kenneth, L. K. (2003). Information technology and economic performance: A critical review of the empirical evidence. *ACM Comput. Surv., 35*(1), 1-28.

Jhingran, A. D., Mattos, N., & Pirahesh, H. (2002). Information integration: A research agenda. *IBM System Journal, 41*, 555-562.

Kohli, R., & Devaraj, S. (2003). Measuring information technology payoff: A meta-analysis of structural variables in firm-level empirical research. *Information Systems Research, 14*(2), 127-145.

Mahnke, V., Overby, M. L., & Vang, J. (2005). Strategic outsourcing of IT services: Theoretical stocktaking and empirical challenges. *Industry & Innovation, 12*(2), 205-253.

Malhotra, A., Gosain, S., & El Sawy, O. A. (2005). Absorptive capacity configuration in supply chains: Gearing for partner-enabled market knowledge creation. *MIS Quarterly, 29*(1), 145-187.

Mata, F. J., Fuerst, W. L., & Barney, J. B. (1995). Information technology and sustained competitive advantage: A resource-based analysis. *MIS Quarterly, 19*(4), 487-505.

McAfee, A. (2002). The impact of enterprise information technology adoption on operational performance: An empirical investigation. *Production and Operations Management, 11*(1), 33-53.

Melville, N., & Kraemer, K. (2004). Review: Information technology and organizational performance: An integrative model of it business value. *MIS Quarterly, 28*(2), 283-322.

Powell, T. C., & DentMicallef, A. (1997). Information technology as competitive advantage: The role of human, business, and technology resources. *Strategic Management Journal, 18*(5), 375-405.

Ravichandran, T., & Chalermsak Lertwongsatien, C. L. (2005, Spring). Effect of information systems resources and capabilities on firm performance: A resource-based perspective. *Journal of Management Information Systems, 21*(4), 237-276.

Tanriverdi, H. (2005). Information technology relatedness, knowledge management capability, and performance of multibusiness firms. *MIS Quarterly, 29*(2), 311-334.

Tatsiopoulos, I. P., Ponis, S. T., Hadzillias, E. A., & Panayiotou, N. A. (2002). Realization of the vertical enterprise paradigm in the clothing industry through e-business. *Production and Operations Management, 11*(4), 516-530.

Wade, M., & Hulland, J. (2004). Review: The resource-based view and information systems research: Review, extension, and suggestions for future research. *MISQ Quarterly, 28*(1), 107-142.

Williams, M. L., & Frolick, M. N. (2001). The evolution of EDI for competitive advantage: The FedEx case. *Information Systems Management, 18*(2), 47-53.

Zahra, S. A., & George, G. (2002). Absorptive capacity: A review, reconceptualization, and extension. *Academy of Management Review 27*(2), 185-203.

Zahra, S. A., Hayton, J. C., & Salvato, C. (2004). Entrepreneurship in family vs. non-family firms: A resource-based analysis of the effect of organizational culture. *Entrepreneurship Theory and Practice, 28*(4), 363–381.

KEY TERMS

Absorptive Capacity: Absorptive capacity is defined as the firm's ability to identify, assimilate, and exploit external knowledge to commercial ends (Cohen & Levinthal, 1990); upon a resource-based view perspective, absorptive capacity represents the ability of a company to translate a change in a combination of input resources into organizational performance (Malhotra et al., 2005; Zahra et al., 2002).

Business Process: A business process is the specific ordering of work activities across time and space, with a beginning, an end, and clearly identified inputs and outputs (Davenport, 1993).

IS Integration: IS integration is defined as the outcome of initiatives leading to greater technical standardization and broader user access to a common set of technical resources, infrastructure, data, or software applications (Hasselbring, 2000; Jhingran, Mattos, & Pirahesh, 2002).

IT Business Value: IT business value is the organizational performance impacts of information technology at both the intermediate process level and the organization wide level, and comprising both efficiency impacts and competitive impacts (Melville et al., 2004).

Organizational Performance: Organizational performance denotes IT-enabled overall firm performance, including productivity, efficiency, profitability, market value, and competitive advantage (Melville et al., 2004).

Sustainable Competitive Advantage: A firm is said to have a sustained competitive advantage when it is implementing a value creating strategy not simultaneously being implemented by any current or potential competitors and when these other firms are unable to duplicate the benefits of this strategy (Barney, 1991).

Technological IT Resource (TIR): TIR can include both hardware and software, and can be categorized into IT infrastructure, and business applications using the infrastructure (Melville et al., 2004)

Organizational Hypermedia Document Management Through Metadata

Woojong Suh
Inha University, Korea

Garp Choong Kim
Inha University, Korea

INTRODUCTION

Web business systems, the most popular application of hypermedia, typically include a lot of hypermedia documents (hyperdocuments), which are also called Web pages. These systems have been conceived as an essential instrument in obtaining various beneficial opportunities for CRM (customer relationship management), SCM (supply chain management), e-banking or e-stock trading, and so forth (Turban et al., 2004). Most companies have made a continuous effort to build such systems. As a result, today the hyperdocuments in the organizations are growing explosively.

The hyperdocuments employed for business tasks in the Web business systems may be referred to as organizational hyperdocuments (OHDs). The OHDs typically play a critical role in business, including the forms of invoices, checks, orders, and so forth. The organization's ability to adapt the OHDs rapidly to ever-changing business requirements may impact on business performance. However, the maintenance of the OHDs increasing continuously is becoming a burdensome task to many organizations; managing them is as important to economic success as is software maintenance (Brereton et al., 1998).

An approach to solve the challenge of managing OHDs is to use metadata. Metadata are generally known as data about data (or information about information). Concerning this approach, this article first reviews the previous studies and discusses perspectives desirable to manage the OHSs and then provides metadata classification and elements. Finally, this article discusses future trends and makes a conclusion.

BACKGROUND

The hyperdocument is a special type of digital document based on the interlinking of nodes such as multimedia components and sets of data elements derived from databases. For digital document, metadata have typically been employed for the access to media- and application-specific documents, such as for information discovery (Anderson & Stonebraker, 1994; Glavitsch et al., 1994; Hunter & Armstrong, 1999). Also, most of the previous studies on metadata for hyperdocuments have also been interested in information discovery from a content-oriented perspective (Lang & Burnett, 2000; Li et al., 1999; Karvounarakis & Kapidakis, 2000). Especially, a set of hyperdocument metadata, the Dublin Core (Dublin Metadata Core Element Set) (Weibel et al., 1995; Weibel & Koch, 2000), has been paid attention to as a standard for Web information resources and also focuses on the information discovery. However, besides this perspective, for the OHDs metadata, the organizational perspectives also need to be considered to satisfy various managerial needs of organizations.

First, a process-oriented perspective needs to be considered. It is also pointed out that the perspective needs to be reflected on defining metadata of corporate digital documents (Murphy, 1998). OHDs as corporate digital documents are closely related to business tasks and information for them in an organization. Generally, corporate documents are produced in undertaking an organizational process (Uijlenbroek & Sol, 1997); furthermore, most businesses are based on, or driven by, document flow (Sprague, 1995). Thus, documents and business processes may be considered simultaneously in the analysis of a corporate information system (Frank, 1997). In this context, the OHDs may affect the speed of communications to perform business process. Accordingly, the OHDs should be designed to support collaboration among workers in business processes. Also, the OHDs can be rapidly improved to fit ever-changing business requirements.

Second, the metadata for OHDs are to be considered from a technical perspective. The system resources linked to the OHDs, such as program files and data components dynamically cooperated, are a considerable part of the organizational assets. The links between such resources and OHDs are very complex. Accordingly, managing the resources and the links through metadata can result in the efficient use of the organizational asset; the metadata related to the technical components can help developers change and improve the OHDs more efficiently.

Third, in the long term, the metadata role of OHDs should be extended toward organizational memory (OM), because organizational digital documents are a major source of OM (Murphy, 1998; Sprague, 1995). The OM techniques concentrate on managing an organization's information or knowledge of the past (Stein & Zwass, 1995; Wijnhoven, 1998). The metadata for OHDs need to play a critical role in managing a variety of histories in terms of business functions, communication mechanisms, technical artifacts, and contents. The memory may provide knowledge to support various decisions for controlling communication mechanisms in a business process, linking to the previous responsible workers, or maintaining the hypermedia applications.

Considering all the perspectives discussed previously, metadata roles for OHDs can be summarized in three levels--operation, system, and organization--as shown in Table 1. In fact, we believe that these roles can also be applied to other kinds of corporate digital documents.

METADATA CLASSIFICATION AND ELEMENTS FOR OHDS

Metadata classification can be perceived as a fundamental framework for providing metadata elements. According to the our perspective on the metadata for OHDs described in the previous section, the following categories of metadata need to be considered:

- *Content-dependent Metadata*: These metadata are used to enable understanding of the content of documents. The metadata include information that depends on (i) the content directly, and (ii) semantic meanings based on the content of the document indirectly.
- *Workflow-dependent Metadata*: These metadata provide information about workflow related to an organizational hyperdocument. These metadata are concerned with process-related factors such as workers, tasks, or business rules.
- *Format-dependent Metadata*: These metadata describe information on formats related to organizational hyperdocuments as well as hypermedia components such as nodes, anchors, interface sources, and database attributes.
- *System-dependent Metadata*: These metadata provide information concerned with storage- and software-related information on system resources such as hyperdocuments, interface sources, and databases.
- *Log-dependent Metadata*: These metadata describe information on the history and the status of organizational hyperdocuments.

Content-dependent metadata are essential for discovering information in OHDs. Workflow-dependent metadata can contribute to increasing the ability to control business processes through the efficient adaptation of OHDs to ever-changing organizational requirements. Format-dependent metadata can provide an understanding of the hypermedia features in terms of structures and operational mechanisms, so that they can be useful in the technical maintenance of OHDs. System-dependent metadata can also play a critical role in technical maintenance by providing information on hardware and location, and software technologies applied to the hyperdocuments. This meta-information is essential for sharing and reusing system resources. Finally, log-dependent metadata may contribute to organizational memories. Thus, the metadata in this category should be specified in order to capture the history of OHDs. According to this classification, detailed metadata elements may be specified under the classification suggested in this article, as shown in Table 2.

Content-dependent classification consists of elements that enable users to understand the content of the hyperdocuments. The document domain may be in terms of content and

Table 1. Metadata roles for OHDs

Level	Metadata Roles
Operation	● Easy and fast access ● Increased accuracy
System	● Interoperability under heterogeneous environment ● Document maintenance ● Document distribution
Organization	● Increased reusability of information and knowledge resources ● Increased capacity of business management ● Increased organizational memory

Table 2. Metadata elements of OHDs

Classifications	Elements
Content-dependent	[Document] Title, Description, Document Domain Name, Conceptual Attribute Name [Anchor] Name [Data Node] Title [Interface-Source] Name
Workflow-dependent	Task Domain Name, Task, Agent Domain Name, Agent Object Name, Business Rule
Format-dependent	[Document] Type [Anchor] Type [Node] Type, [Data Node] Property [Interface-Source] Property [DB] Physical Attribute Type
System-dependent	[Document] File Name, H/W Name, Location Path, S/W Technology [Data Node] File Name, H/W Name, Location Path [Interface-Source] File Name, Storage, Location Path [Database] Name, H/W Name, Location Path, Table Name, Table Type, Physical Attribute Name, DBMS Name
Log-dependent	Document Number, Version Number, Loading Date, Withdrawal Date, Update Date, Update Description, Director, Operator

roles. The conceptual attributes, as data items represented on a hyperdocument, are connected to a corporate database. Interface sources are primarily multimedia components such as image or animation that are represented on interfaces.

A node, an essential factor of hypermedia, has been defined as the fundamental unit of hypertext (Nielsen, 1993), fragments of information (Fluckiger, 1995), or basic information containers (Schwabe & Rossi, 1994). This article defines a node as any navigational object with hyperlinks. An object may a type of media, such as image, sound, animation, or a hyperdocument itself. Nodes may be categorized into two types from the perspective of their properties: document nodes and data nodes. Nodes are also of two types from the perspective of link directions: source and destination. The fundamental definitions for nodes are summarized in Table 3.

An interface may consist of one or more hyperdocuments. Accordingly, a document node, a hyperdocument, can be either only a part of an interface or an interface itself. From these definitions, the element of node type in format-dependent metadata may take a document node or data node as its value.

The information of a hyperdocument in terms of a process can be obtained effectively by the use of a document-based workflow concept. The workflow concept typically includes common essential factors in terms of a unit of a work, a tool of a work, and a person for a work (Lee & Suh, 2001). In the document-based workflow approach, an OHD is regarded as a tool of a work. A task, as a work unit consisting of a workflow, may be described as operations or descriptions of human actions with a hyperdocument. An agent refers to a person who performs the task, and is expressed by hierarchical status in an organization. An agent domain can be defined as a group of agent objects having common tasks or organizational objectives. The agent domain can be typically conceived as a department of an organization. The task

Table 3. Types of nodes

Perspectives	Types	Descriptions
Properties	Document Node	A unit of an HTML document, which may be a whole interface or a part of it.
	Data Node	A unit of multimedia data that may be accessed from a document node.
Link Direction	Source Node	Nodes that can access a current node.
	Destination Node	Nodes that a current node can access.

domain is a set of tasks corresponding to an agent domain.

The format-dependent metadata are concerned with type or properties of hyperdocuments, anchors, nodes, interface sources, and databases. The types of anchors can be static or dynamic depending their value. The definitions of these types are as follows:

- *Static anchor:* One fixed in a hyperdocument.
- *Dynamic anchor:* One generated by data stored in a database; that is, it refers to data transformed into and represented as an anchor when the data are accessed by a hyperdocument according to any event that occurs as a function or another anchor.

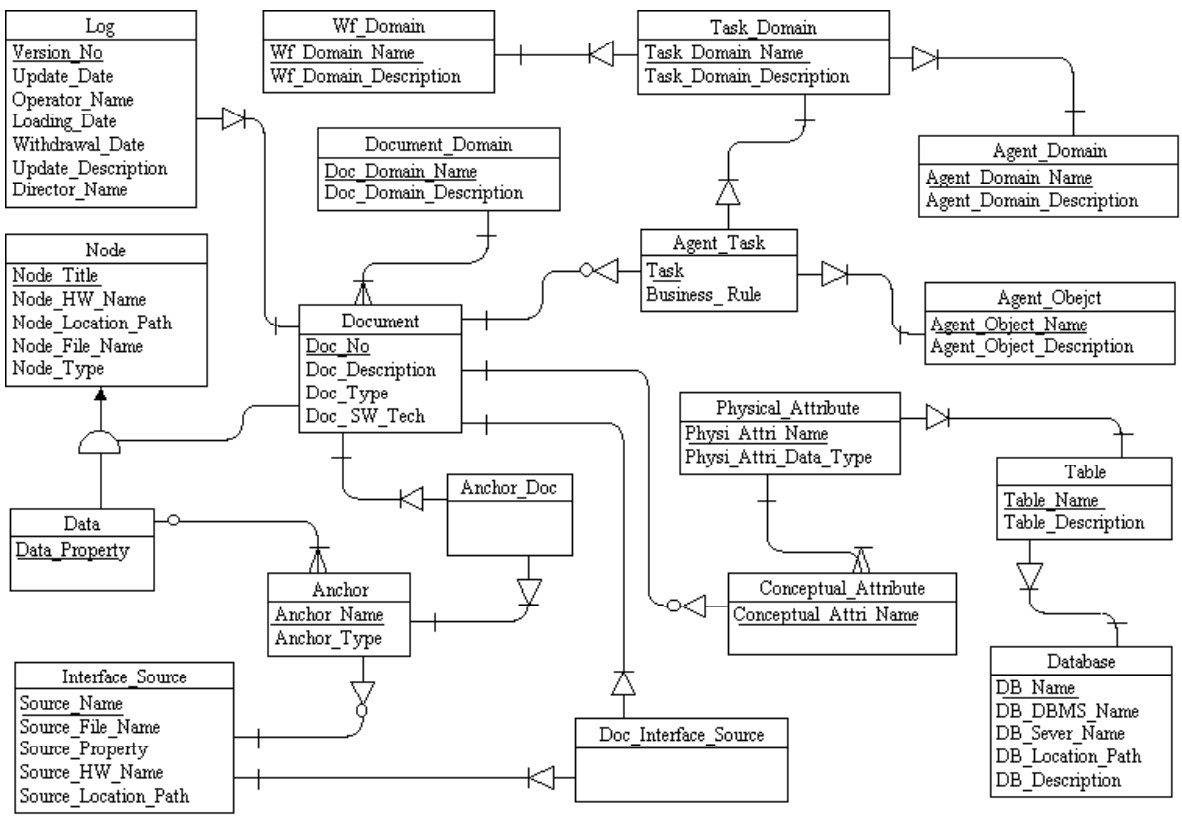
The types of OHDs can be categorized into three: control, processing, and referential, according to their roles in a hypermedia application. These types are defined as follows:

- *Control Type:* Hyperdocuments that typically guide users to other hyperdocuments of processing or referential types. Homepages or index pages are examples of this type.
- *Processing Type:* Hyperdocuments that typically contain data attributes connected with a database in the style of a form.
- *Referential Type:* Hyperdocuments that provide supplementary information about work instructions, business rules, news, or products.

Properties of interface sources are multimedia properties such as images or animation. The properties of data nodes are the same as those of interface sources. The physical attribute type of a database implies the data properties of the attribute.

System-dependent metadata focus on storage-related information. The storage-related information can be found in various applications, but they are not integrated, so it is difficult to create a synergetic effect. However, if metadata of all the components of hypermedia systems, such as hyperdocuments, data nodes, interface sources, and databases, are integrated into a repository, it is possible to manage a hypermedia system effectively. Software technology is a major factor in determining the capacity and characteristics of a system. Recently, for example, a variety of emerging

Figure 1. Metadata DB schema of OHDs



software technologies, such as ASP (Active Server Page), Java scripts, Visual Basic scripts, or Perl, have had a considerable impact on the improvement of hypermedia systems. Accordingly, the information on software technologies applied to a hyperdocument may contribute to the maintenance of a hypermedia application.

Log-dependent metadata are used for tracing the history of hyperdocuments for the maintenance of their system. Although there may be log information captured automatically by an operating system or an application program, it is typically separated, so it is difficult to obtain a synergetic effect in maintenance. Furthermore, it is insufficient to maintain a hypermedia system effectively. Therefore it is necessary to capture the information about changes of hyperdocuments synthetically. Some hyperdocuments may be operated temporally, depending on their purposes. Accordingly, version- or time-related information should be managed. The loading date is a date pushing a hyperdocument into its system. The withdrawal date is a date removing a hyperdocument from the system for the expiration of its periodic role or for updating. Information on responsible operators and directors for a hyperdocument may be required for responsible management, or questions by new staff members.

The metadata elements in Table 2 can be designed as shown in Figure 1. The schema was implemented in a meta-information system for the maintenance of OHDs (Suh & Lee, 2002). The system consists of two main modules: metadata management and a supporting module. The metadata management module is responsible for metadata handling such as creating, editing, or deleting. The supporting module serves two types of functions: searching an OHD and reporting its meta-information.

FUTURE TRENDS

Concerning the challenge to cope with a flood of OHDs, today content management system (CMS) is drawing great attention; the market size of content management solutions are forecasted to grow explosively (WinterGreen Research, Inc., 2003). CMSs help organizations develop Web applications, and support content lifecycle including the following phases: creation, approval, publishing, deployment, delivery and removal. Now, CMSs are conceived as an essential infrastructure for not only developing but also maintaining Web applications; most Web projects include implementing CMSs along with Web applications. Furthermore, CMSs are often incorporated into the existing infrastructures of Web applications to support the maintenance of the applications.

The lifecycle of content is supported primarily by the engines of workflow and personalization on the basis of a repository that includes the metadata as well as actual content. The metadata are a critical component required to operate those engines and other functions of CMSs, so their importance has been addressed (Perry, 2001). In fact, current CMS solutions originate from some different backgrounds, so that the functional features and metadata schema are also different from each other. Accordingly, in the near future, the metadata standard may be required in the CMS solution industry as it has been in other application domains of information technologies so far.

CONCLUSION

Recently, many organizations have expanded their business workplaces through Web business systems. Organizational hyperdocuments (OHD), critical information resources, are growing explosively in such organizations. Accordingly, the maintenance of these documents is becoming a burdensome task. For this challenge, this article offers an approach based on metadata required from the managerial perspectives as well as content-oriented. The managerial perspectives address the adaptability of Web business systems to the ever-changing business requirements and the efficiency of developers' works. The presented metadata are expected to help manage organizational memory in the long term through implementation into a metadata-based system.

REFERENCES

Anderson, J.T., & Stonebraker, M. (1994). Sequoia 2000 metadata schema for satellite images. *ACM SIGMOD Record, 23*(4), 42-48.

Brereton, P., Budgen, D., & Hamilton, G. (1998). Hypertext: The next maintenance mountain. *IEEE Computer, 31*(12), 49-55.

Fluckiger, F. (1995). *Understanding networked multimedia: Applications and technology.* Englewood Cliffs, NJ: Prentice Hall.

Frank, U. (1997). Enhancing object-oriented modeling with concepts to integrate electronic documents. *Proceedings of the 30th Hawaii International Conference on System Sciences, 6,* 127-136.

Glavitsch, U., Schauble, P., & Wechsler, M. (1994). Metadata for integrating speech documents in a text retrieval system. *ACM SIGMOD Record, 23*(4), 57-63.

Hunter, J., & Armstrong, L. (1999). A comparison of schema for video metadata representation. *Computer Networks, 31,* 1431-1451.

Karvounarakis, G., & Kapidakis, S. (2000). Submission and repository management of digital libraries, using WWW. *Computer Networks, 34,* 861-872.

Lang, K., & Burnett, M. (2000). XML, metadata and efficient knowledge discovery. *Knowledge-Based Systems, 13,* 321-331.

Lee, H., & Suh, W. (2001). A workflow-based methodology for developing hypermedia information systems. *Journal of Organizational Computing and Electronic Commerce, 11*(2), 77-106.

Li, W., Vu, Q., Agrawal, D., Hara, Y., & Takano, H. (1999). PowerBookmarks: A system for personalizable Web information organization, sharing, and management. *Computer Networks, 31,* 1375-1389.

Murphy, L.D. (1998). Digital document metadata in organizations: Roles, analytical, approaches, and future research directions. *Proceedings of the 31st Hawaii International Conference on System Science, 2,* 267-276.

Nielsen, J. (1993). *Hypertext and hypermedia.* Boston: Academic Press Professional.

Perry, R. (2001). Managing the content explosion into content-rich applications. *Internet Computing Strategies, Report, 6*(2). Yankee Group.

Schwabe, D., & Rossi, G. (1994). *From domain models to hypermedia applications: An object-oriented approach.* Technical Report MCC 30/94. Dept. of Information, PUC-Rio.

Sprague, R.H. (1995, March). Electronic document management: Challenges and opportunities for information systems managers. *MIS Quarterly, 19,* 29-49.

Stein, E.W., & Zwass, V. (1995). Actualizing organizational memory with information systems. *Information Systems Research, 6*(2), 85-117.

Suh, W., & Lee, H. (2002). Managing organizational hypermedia document: A meta-information system. In K. Siau (Ed.), *Advance topics in database research* (vol. 1, pp. 250-266). Hershey, PA: Idea Group Publishing.

Turban, E., Lee, J., King, D., & Chung, H.M. (2004). *Electronic commerce 2004: A managerial perspective.* Prentice Hall.

Uijlenbroek, J.J.M., & Sol, H.G. (1997). Document based process improvement in the public sector: Aettling for the second best is the best you can do. *Proceedings of the 30th Hawaii International Conference on System Sciences, 6,* 107-117.

Weibel, S., Godby, J., Miller, E., & Daniel, R. (1995). OCLC/NCSA metadata workshop report. Retrieved March 21, 2004, from *http://www.oasis-open.org/cover/metadata.html*

Weibel, S., & Koch, T. (2000). The Dublin Core metadata initiative: Mission, current activities, and future directions. *D-Lib Magazine, 6*(12). Retrieved March 21, 2004, from *http://www.dlib.org/dlib/december00/weibel/12weibel.html*

Wijnhoven, F. (1998). Designing organizational memories: Concept and method. *Journal of Organizational Computing and Electronic Commerce, 8*(1), 29-55.

WinterGreen Research, Inc. (2003). *Content management market opportunities, market forecasts, and market strategies, 2004-2009.* Report.

KEY TERMS

Guided Tours: A navigation type that leads users to a predefined trail of nodes without freely explorative navigation, using, for example, previous and next anchors.

HTML: HyperText Markup Language. It is markup language using tags in pairs of angle brackets, for identifying and representing the Web structure and layout through Web browsers; it is not a procedural programming language like C, Fortran, or Visual Basic.

Hyperspace: Information spaces interlinked together with hypermedia structures. Concerning World Wide Web, cyberspace is sometimes used instead of hyperspace.

META Tag: A type of HTML tags with a word, "META". Web "spiders" read the information contained within META tags to index Web pages. The information consists of keywords and a description of its Web site.

Open Hypermedia: As a more generalized concept addressing interoperability among hypermedia services, it has the following characteristics: it should be platform independent and distributed across platforms, and users should be able to find, update, make links to, and exchange the information, unlike hypermedia titles on CD-ROM.

SGML: Standard Markup Language. Document standard from ISO (reference ISO 8897). It is a meta-language that can define document logical structure by using Document Type Definition (DTD) component. An example of document types defined using the DTD of SGML is HTML.

XML: eXtensible Markup Language. It is quite different from HTML in that XML gives document authors the ability

to create their own markup. XML is flexible in creating data formats and sharing both the format and the data with other applications or trading partners, compared with HTML.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2236-2242, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Organizational Project Management Models

Marly Monteiro de Carvalho
University of São Paulo, Brazil

Fernando José Barbin Laurindo
University of São Paulo, Brazil

Marcelo Schneck de Paula Pessôa
University of São Paulo, Brazil

INTRODUCTION

Project management plays an important role in the competitive scenario, and achieved in the 1990s the status of methodology (Carvalho & Rabechini, Jr., 2005). Nowadays, there are more than 100,000 practitioners that earned the Project Management Professional (PMP®) certification from the Project Management Institute (PMI). This indicator highlights the increasing interest in project management area, especially in the IT companies, which are one of the top five industries in PMI's membership numbers (PMI, 2005).

The widely spread framework proposed by PMI called Project Management Body of Knowledge (PMBoK), now in the third edition (PMBoK, 1996, 2000, 2004), has been adopted by several kinds of project-driven organization (PMI, 2004). PMBoK clusters the main project management best practices in nine key areas.

Nevertheless, a research carried out by Standish Group (2003) showed high failure level in IT project in North America. The research involved about 13.522 projects, of which only 34% can be considered a success. The main causes for IT projects failure were related to user's commitment, manager support and requirement definition. It is important to emphasize that, regarding the project success measure in historical perspective, the success rate improved if compared to the first similar research carried out in 1999, which was just 16%.

Based on this scenario, this chapter presents the main organizational project management models in order to help companies to upgrade project performance.

BACKGROUND

Several project management models had been discussed in the academic literature concerning its effectiveness and efficiency. The models focus on project efficiency, balancing scope expectations and the available resources (Carvalho & Rabechini, Jr., 2005). However, the project management efficiency models, such as PMBoK framework, cannot provide a standard benchmark for project management competences and maturity enhancing. Thus, in order to extend the efficiency models to an effectiveness perspective, several PM organizational models have been proposed.

Nevertheless, project management efficiency models focus on the project and not on organizational issues. As Engwall (2003, p. 789) states "no project is an island" and to achieve success in this area it is important to fit project management best practices to organizational environment.

On the other hand, the effectiveness issue encompasses the organizational project management models, which promotes the strategic alignment between this area and the organizational vision. It means providing an appropriate strategic alignment and portfolio analysis, project management organizational structure, methodology and project manager carrier (Carvalho & Rabechini, Jr., 2005; Carvalho, Laurindo, & Pessoa, 2003, 2005; Rabechini, Jr., Gelamo, & Carvalho, 2005; Shimizu, Carvalho, & Laurindo, 2006).

The implementation of formal efficiency and effectiveness procedures is quite new in IT projects and organizations. There are different approaches and this article focuses on the organizational project management models. The theoretical models selected to discuss this issue are the Capability Maturity Model (CMM) (Humphrey, 1989; Paulk, Weber, Curtis, & Chrissis, 1995), Project Management Maturity Model (PMMM) (Kerzner, 2000, 2001); the Quality Systems to software ISO9000-3 (2001) and ISO 12207 (1995); and the Organizational Project Management Maturity Model (OPM3) (PMI, 2003).

CAPABILITY MATURITY MODEL (CMM)

Humphrey (1989) identifies maturity levels in the IT project development process, based on the managerial behavior found in companies. The fundamental concepts of the process maturity derive from the belief that the development management process is evolutionary. Paulk et al. (1995) identify the distinguishing characteristics between immature and mature organizations, as shown in Table 1.

Table 1. Immature organization x mature organization (Paulk et al., 1995)

IMMATURE ORGANIZATION	MATURE ORGANIZATION
• *Ad hoc*: improvised process by practitioners and managers • Not rigorously followed and not controlled • Highly dependent on personal knowledge • Little understanding of progress and quality • Compromising product functionality and quality to meet schedule • High risk when new technology is applied • High maintenance costs and unpredictable quality	• Coherent with action plans: the work is effectively achieved • Processes are documented and continuously improved • Perceptible top and middle management commitment • Well controlled assessment of the process • Product and process measures are used • Disciplined use of technology

The CMM (Humphrey, 1989; Paulk et al., 1995; Pessôa & Spinola, 1997) was developed by SEI (the Software Engineering Institute of Carnegie Mellon University) and presents five maturity levels, each corresponding to a set of structural requirements for key process areas (Figure 1).

Although each project is unique, it could be organized in a process to be applied in other projects. IT projects managers used to apply a "methodology," that is, they established the steps to be followed in order to develop a system. Another singular characteristic is the dynamic technologies breakthrough that demands continuous improvements in the development methods and management of changing process, as described in the CMM model at level 5, the highest level of maturity.

The CMM second level has a consistent project management structure and the goal of this level is to deliver projects on time. To perform this, the model has several points that must be achieved, like effort and size estimation, strong process control (such as periodic meetings between technical people and managers), and several measures to show project status more clearly.

CMM is not an adequate reference for the assessment of internal methodologies, because it was not conceived to perform this kind of analysis. ISO 15504 (1998) proposed the standard project SPICE as a more appropriated model to evaluate maturity level of specific processes. While CMM level of maturity specifies a set of processes that have to be performed, ISO 15504 establishes maturity levels for each individual process: level 0-incomplete; level 1-performed; level 2-managed; level 3-established; level 4-predictable; and level 5-optimizing. This is a different approach of CMM, because an organization does not perform a maturity level, but has a maturity profile: A maturity level is measured for each specific process. This new approach is very useful to the organization perspective because one can easily measure strong and weak points of their process and plan improvement activities. Furthermore, from the companies' point of view, it is easier to understand staged levels, as the performed processes are already predefined.

The SPICE approach defined in standard ISO 15504 (1998) had firstly influenced *CMM for Systems Engineering*, published in 1995, and more recently influenced CMM I (CMM-I1; CMM-I2), just published in 2002. CMM-I, the integration model, was enhanced in two dimensions: *scope dimension* and *evaluation dimension*.

In the scope dimension, this new model incorporated other published models and covered all project activities, not only software, as the original software CMM did, but also other engineering fields. In the evaluation dimension, CMM-Il incorporated both approaches: the traditional (called staged CMM) and the maturity profile (called continuous CMM). Figure 2 shows the continuous CMM-I representation to be compatible with the ISO/IEC 15504 standard.

Figure 1. Maturity levels (Paulk et al., 1995)

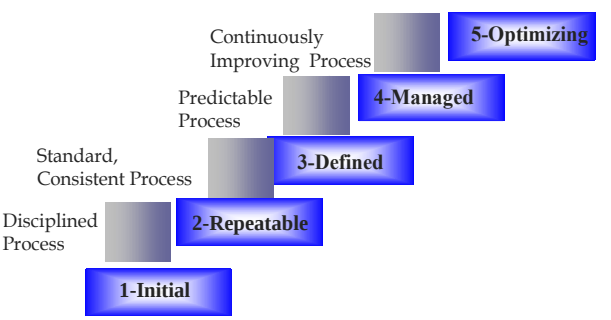
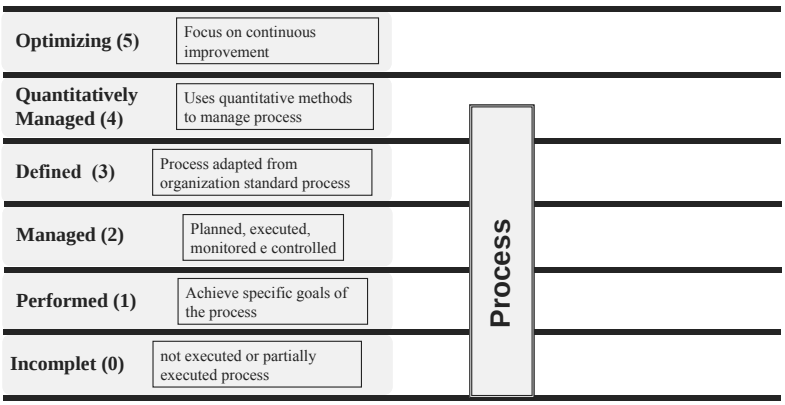


Figure 2. Continuous maturity process representation in CMM-I (CMM-I1, 2002)



CMM-I (and software CMM) considers that maturity level is an organizational characteristic and it is independent of the professionals involved. Nowadays, there is a strong tendency toward the adoption of CMM-I models which were sponsored by the Department of Defense (DoD); meanwhile, ISO standards are less used.

PROJECT MANAGEMENT MATURITY MODEL

In order to extend the capability maturity model (CMM) to project management, Kerzner (2000) and (2001) proposes a Project Management Maturity Model (PMMM).

The PMMM differs in many aspects from the CMM, but this framework also introduces benchmarking instruments for measuring an organization's progress along the maturity model, detailing five levels of development for achieving maturity (Carvalho et al., 2003). Figure 3 shows PMMM's levels.

It is important to highlight the differences in terminology between the CMM and PMMM (compare Figures 2 and 3), which could lead to misunderstanding when both models are being implemented in the IT domain of the same organization.

PMMM addresses the key knowledge areas across the project management process, in compliance with PMBoK, and integrates them with the Project Management Office (PMO) in the strategic level.

Kerzner (2000) identifies a life cycle in PMMM level 2, common processes, which could be broken into five phases, as shown in Figure 4. It is important to note that some simultaneity among the phases can occur.

Figure 3. Project Management Maturity Model (Adapted from Kerzner, 2001)

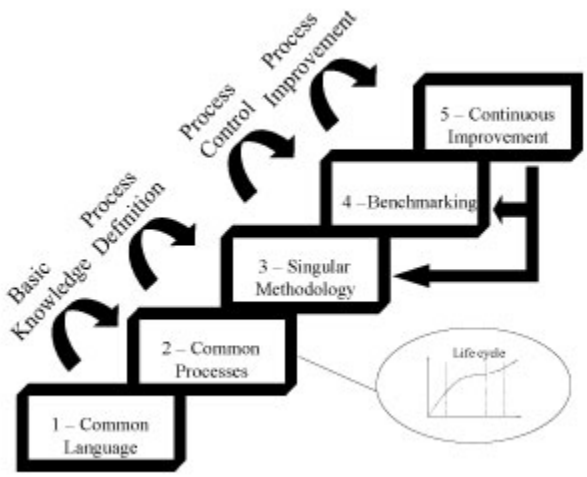
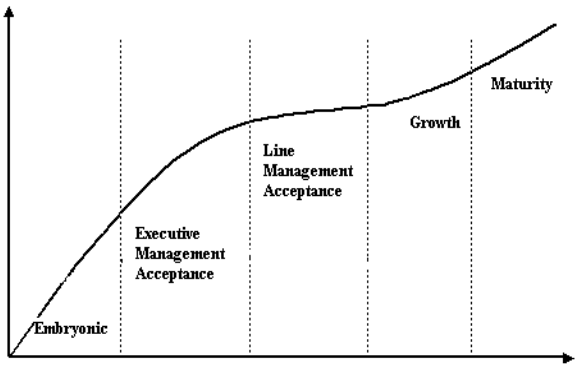


Table 2. Life cycle phases characteristics (Kezner, 2001)

Phase	Characteristics
embryonic	• recognizing the need for PM • recognizing PM's potential benefits • applications of PM to the business • recognizing the changes necessary to implement PM
executive management acceptance	• visible executive support • executive understanding of PM • project sponsorship • willingness to change the way the company does business
line management acceptance	• visible line management support • line management commitment to PM • line management education • release of functional employees for PM training programs
growth	• development of company PM life cycles • development of a PM methodology • a commitment to effective planning • minimization of scope • selection of PM software to support methodology
maturity	• development of a management cost/schedule control system • integration of schedule and cost control • development of an educational curriculum to support PM

Figure 4. Life cycle phases (Kerzner, 2000)



The embryonic phase means that the organization starts to recognize the benefits of project management (PM), usually by lower and middle levels of management. The two next phases are achieved when the PM concepts are accepted and have visible support and commitment by executive and line management.

Kerzner (2001) emphasizes the growth phase as the most critical, because it is the beginning of the creation of the PM process, and warns that different methodologies for each project should be avoided.

The last life cycle phase—maturity—is difficult to achieve due to several factors such as organizational resistance to project cost control and horizontal accounting.

The main characteristics of these life cycle phases emphasized by Kerzner (2001) are described in Table 2.

QUALITY SYSTEMS

It is important to note that the adoption of systems models, such as ISO 9000, focuses on the creation and maintenance of a quality system, applied to any process. The ISO 9001:2000 new version, published in the year 2000, was fully restructured to have a more clear process-focused approach. Other ISO standards offer an overview of these standards to the software field and contribute to deploying this approach to specific processes, such as software products (ISO 9126-NBR 13596), quality requirements for software packages (ISO 12119), and the software life cycle process (ISO 12207).

ISO 9000-3 (2001) is a guide to help with ISO 9001 interpretation for the software field (Pessôa & Spinola, 1997). The previous versions of this guide were developed by the ISO/TC/SC2 committee, the quality branch of ISO. Nowadays, this ISO 9000-3 guide is being revised by ISO/IEC JTC1/SC7, the information technology branch. The ISO9001:2000 new structure made this task easier than previous versions and it is incorporating a map of the relationship between IT standards (ISO/IEC JTC1/SC7) and its respective quality systems described in ISO 9001 to this standard. For example, ISO 9001 specifies that the organizations have to identify their processes and ISO 12207 (ISO12207) proposes a set of processes involving software

development, supply, acquisition and maintenance. In addition to ISO 12207, other standards are referenced, such as ISO 9126 for software products, ISO 12119 for quality requirements for software packages and ISO 15504 for software evaluation. This was considered an improvement of the standards structure that matches quality system standards with technical and specific standards.

ISO 9001:2000 standards have the purpose of certifying organizations, in which Quality Systems comply with the standards and provide a structure to manage quality independent of the organizational activity. This is not enough for specific fields of application and, for this reason there are complementary sets of standards in some areas, like QS-9000 for the automobile industry and TL-9000 for the telecommunications industry.

The standards from ISO/IEC/ JTC1/SC7 have the purpose of complementing the quality system for the IT specific area, not focusing on applications as with the automobile or telecommunications industry, but considering the specificities of software and systems development.

In general, ISO 9000 can be a good starting point for implementing a quality system. This system allows the organizations to disseminate quality culture and create the initial structure to implement more specific quality systems.

ORGANIZATIONAL PROJECT MANAGEMANT MATURITY MODEL

In 2003, the *Project Management Institute* (PMI) concluded the *Organizational Project Management Maturity Model* (OPM3) development, started in 1998 (PMI, 2003). According to PMI (2003), the meaning of the OPM3 can be defined as follows. *Organizational* increases the work domain, leaving the context of the project itself, which is a matter for PMBoK to the organizational perspective. The meaning of the word *maturity* implies that the capacities should grow during the project management implementation period but also demonstrates how success occurs and ways of correcting or preventing common problems. The model implies change and progression in 4 stages: *standardize measure, control* and *continuously improv*e.

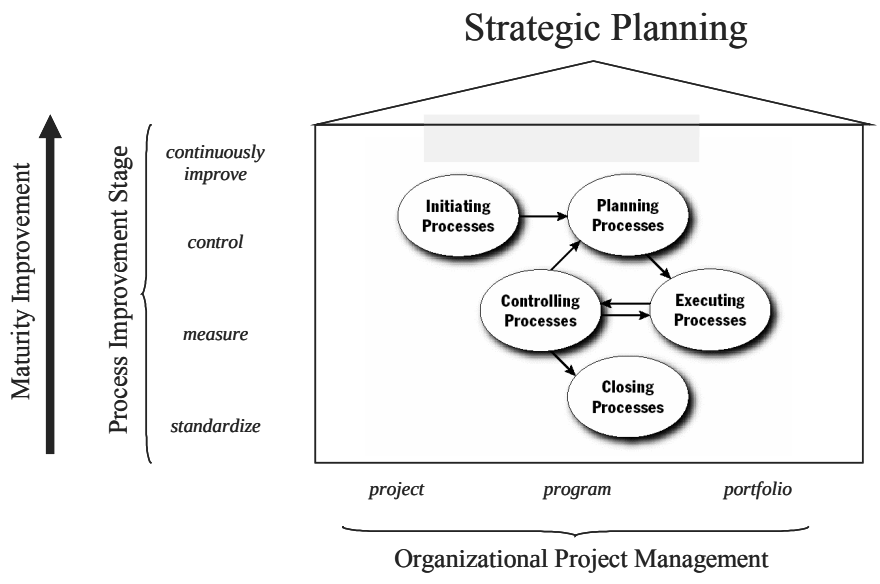
The OPM3 model is organized into the five groups of processes, proposed in PMBoK, as following: *initiating, planning, executing, controlling* and *closing* processes. Besides, the model classifies improvement stages: *standardize, measure, control* and *continuously improve.* Finally, this maturity model considers three constructs: *project, program* and *portfolio.* Figure 5 shows the relation among life cycle processes, improvement stages and constructs.

The OPM3 seems to be more detailed than the others and has specific recommendations concerning portfolio and programs.

FUTURE TRENDS

All the organizational models analyzed propose a step by step guide in order to drive organizations to achieve project management maturity. However, competences and organizational structure issues should be incorporated in order to achieve project success.

Figure 5. The OPM3 (Adapted from PMI, 2003)



The competences in professional, teams and organizations levels should be designed and developed. In each of these levels further research is necessary.

The appropriate organizational structure to support maturity progress in all constructs also demands more research. This important issue encompasses several factors in order to design the most appropriate organizational structure to organize project activity such as: projects typology (size, complexity, and uncertainty), organizational culture and philosophy, flexibility demanded and responsiveness.

Thus, the integration of competences build and the design of organizational structure in the maturity models should be considered in futures research.

Finally, companies are spending significant quantities of resources in the implementation of organizational project management models, keeping the discussion about the results controversial, when considering the return over the investments. Thus, the impact of the adoption of organizational models in the results of organization should be measured in further research.

CONCLUSION

In spite of different approaches regarding the organizational project management, there is a general consensus about the importance of the following widely used models: the CMM (Humphrey, 1989; Paulk et al., 1995), the Project Management Maturity Model (PMMM) (Kerzner, 2000, 2001), Quality Systems for software ISO9000-3 (2001) and ISO 12207 (1995), and the Organizational Project Management Maturity Model (OPM3) (PMI, 2003). All of them are empirical and were developed based on methodologies and best practices adopted in organizations.

Although they have different approaches in their conception, they are rather more complementary than conflictive. These models consider that project management maturity level is an organizational characteristic and it should not depend on individual professional expertise and skills but in organizational procedures, methodologies and culture.

In general, the Quality System (ISO 9000) and the maturity models (CMM, PMMM and OPM3) models have the possibility of mutual and complementary synergy, maintaining consistency with their fundamental points. On the other hand, there are important differences among these models, especially concerning the degree of abstraction (Carvalho et al., 2003, 2005; Pessôa, Spinola, & Volpe, 1997; Shimizu et al., 2006; Tingey, 1997).

It is important to highlight the differences in terminology between these models which could lead to misunderstanding when these models are being implemented in the IT domain of the same organization.

REFERENCES

Carvalho, M.M., Laurindo, F.J.B., & Pessôa, M.S.P. (2003). Information technology project management to achieve efficiency in Brazilian companies. In S. Kamel (Ed.), *Managing globally with information technology* (pp. 260-271). Hershey, PA: Idea Group.

Carvalho, M.M., Laurindo, F.J.B., & Pessôa, M.S.P. (2005). Project management models in IT. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 2353-2358). Hershey, PA: Idea Group.

Carvalho, M.M., & Rabechini Jr., R. (2005). Construindo Competências para gerenciar projetos. São Paulo: Editora Atlas, publicação prevista para agosto de 2005, 317.

CMM-I-1. (2002). *Capability maturity model integration, version 1.1 for systems engineering and software engineering: Continuous representation CMU/SEI/SW, V1.1 – CMU/SEI–2002-TR01*. Retrieved May 27, 2008, from www.sei.cmu.edu 02-02-2002

CMM-I-2. (2002). *Capability maturity model integration, version 1.1 for systems engineering and software engineering: Staged representation CMU/SEI/SW, V1.1 – CMU/SEI–2002-TR02*. Retrieved May 27, 2008, from www.sei.cmu.edu 2-02-2002

Engwall, M. (2003). No project is an island: Linking projects to history and context. *Research Policy, 32*, 789-808.

Humphrey, W.S. (1989). *Managing the software process*. Reading, MA: Addison-Wesley.

ISO 12207. (1995). *ISO/IEC 12207:1995: Information technology, software life cycle processes – ISO*.

ISO 9000-3. (2001, May). *Software engineering:Guidelines for the application of ISO 9001:2000 to software* (working draft WD4 ISO/IEC JTC-1 /SC7/WG18 N48).

ISO/IEC/TR15505-2–SPICE. (1998). *Information technology, software process assessment, part 2: A reference model for processes and process capability* (Tech. Rep., 1st ed.).

Kerzner, H. (2000). *Applied project management best practices on implementation*. USA: John Wiley & Sons.

Kerzner, H. (2001). *Strategic planning for project management—using a project management maturity model*. New York: John Wiley & Sons.

Laurindo, F.J.B., Carvalho, M.M., & Shimizu, T. (2003). Information technology strategy alignment: Brazilian cases. In K. Kangas (Org.), *Business strategies for information technology management* (pp. 186-199). Hershey, PA: Idea Group.

Paulk, M.C., Weber, C.V., Curtis, B., & Chrissis, M.B. (1995). *The capability maturity model: Guidelines for improving the software process/CMU/SEI*. Reading, MA: Addison-Wesley.

Pessoa, M.S.P., & Spinola, M.M. (1997). Qualidade de Processo de Software: um novo paradigma. In Iv Inftel – Congresso Petrobrás De Informática and Telecomunicações, São Paulo, 1 a 5/12/1997. Anais. São Paulo.

Pessôa, M.S.P., Spinola, M.M., & Volpe, R.L.D. (1997). Uma experiência na implantação do modelo CMM. In Simpósio Brasileiro De Engenharia De Software, 11., WQS´97 - Workshop Qualidade De Software, Fortaleza, 14/10/1997. Anais. Fortaleza, UFC, 49-57.

Pressman, R.S. (1987). *Software engineering, a practitioner's approach 2a* (2nd ed.). McGraw-Hill.

Project Management Institute. (2003). *Project Management Institute: Organizational project management maturity model (OPM3*). Maryland: Project Management Institute.

Project Management Institute I. (2004). *Project Management Institute: A guide to the project management body of knowledge (PMBoK)* (3rd ed.). Maryland: Project Management Institute.

Rabechini, Jr., R., & Carvalho, M.M. (1999). Concepção de um programa de gerência de projetos em instituição de pesquisa. Revista Valenciana Dèstudis Autonòmics. Espanha: Valência.

Rabechini, Jr., R., Gelamo, R.P., & Carvalho, M.M. (2005). Organizing project management maturity in a system integrating company. In *Proceedings of IRMA 2005: Information Resource Management Association International Conference*, San Diego.

Shimizu, T., Carvalho, M.M., & Laurindo, F.J.B. (2006). *Strategic alignment process and decision support systems: Theory and case studies*. Hershey, PA: Idea Group.

Tingey, M.O. (1997). *Comparing ISO 9000, Malcolm Baldrige, and the SEI CMM for software: A reference and selection guide*. Englewood Cliffs, NJ: Prentice Hall.

KEY TERMS

CMM I (CMM-I1, CMM-I2): A model, which has been enhanced in two dimensions: *scope dimension* and *evaluation dimension*. The CMM-Il incorporated both approaches, the traditional (called staged CMM) and the maturity profile (called continuous CMM).

CMM: The *Capability Maturity Model* is a framework to achieve maturity in project activities in the software field, which presents five maturity levels, each corresponding to a set of structural requirements for key process areas.

ISO 15504: An international standard that proposes the standard project SPICE, which establishes maturity levels for each individual process: level 0-incomplete; level 1-performed; level 2-managed; level 3-established; level 4-predictable; and level 5-optimizing.

ISO 9000-3: A guide to help ISO 9001 interpretation for the software field, that is, the development, supply, acquisition and maintenance of software.

PMBoK: The *Project Management Body of Knowledge* provides a framework to manage project efficiency, balancing scope expectations and the available resources. This model proposes the following nine key areas: (i) integration; (ii) scope; (iii) time; (iv) cost; (v) quality; (vi) human resource; (vii) communication; (viii) risk; and (ix) procurement.

PMMM: The Project Management Maturity Model is a framework that introduces benchmarking instruments for measuring an organization's progress along the maturity model, detailing five levels of development for achieving maturity: level 1 - common language; level 2 - common processes; level 3 - singular methodology; level 4 – benchmarking; and level 5 - continuous improvement, as shown in Figure 3.

Project Life Cycle: Common processes identify in PMMM, level 2, which could be broken into five phases: embryonic; executive management acceptance; line management acceptance; growth; and maturity.

OPM3: The Organizational Project Management Maturity Model is an organizational model, proposed by PMI in 2003, that encompasses five groups of processes, four improvement stages, and three constructs in order to drive organizations to achieve maturity in project management.

Category: IT Education

An Overview of Asynchronous Online Learning

G. R. Bud West
Regent University, USA

Mihai Bocarnea
Regent University, USA

INTRODUCTION

Distance education typically refers to a process where students complete their coursework at a location other than a primary campus. Effectively, this method first developed in the mid-19th century in the form of correspondence courses in the United Kingdom. The correspondence course design includes the instructor and the student mailing assignments back and forth between the university and the student's location. In many cases, the use of the Internet has replaced the correspondence-by-mail method of instructional delivery.

With the advent of television and the further development of radio, some colleges and universities saw an opportunity to present classes via these media. By these methods, various instructors present lectures during set broadcast times while students continue to conduct assignments via correspondence. Additionally, some primary and secondary schools also began at the same time to provide information via television, mainly to supplement and reinforce standard pedagogical instruction in the classroom. Similar to correspondence courses, television and radio instruction generally decreased after the introduction of the Internet as an educational delivery vehicle. However, a notable exception currently exists in some university programs where instructors broadcast live satellite feeds to and from their classroom with classrooms located in regional community colleges, military installations, and other locations. In Virginia, Old Dominion University's TELETECHNET initiative represents an example of one such effort where students both regionally and around the world sit in local centralized meeting places and access instructors while they teach classes at the home campus in real time through two-way television broadcasts.

Once described as the wave of the future, some educators, researchers, and educational administrators suggest that online Internet instruction represents one educational process that has truly come of age. The use of the Internet as an instructional delivery system is exploding in the new millennium. With that explosion comes recognition of the existence of both opportunities and challenges for its effective use as a conduit for meaningful and structured education. In that regard, several researchers describe distinct time and location advantages in the use of Internet instruction, as well as its usefulness in developing knowledge about knowledge (Blair, 2002; Hung, Tan, & Chen, 2005). However, upon review of some examples of online coursework, one may witness a broad range of approaches and quality in online educational programs. In fact, experts specifically note that some online courses lack pedagogical emphasis and design and that universal promises of limitless Internet instruction fail the rationality test (Hung et al.; Wojnar, 2002). This suggests the importance of the Internet as a conduit of learning, but it also suggests a significant and, in some cases, unmet responsibility for those who would help mold and shape lives by instructing and helping to educate people through Internet and intranet mediums. In that regard, dialogue or online discussion has proven valuable in enhancing the online educational process (Blair; Dennen, 2005). Although many university programs use several methods of online discussion with varying degrees of success, some benchmarks have emerged and proven their effectiveness.

BACKGROUND

Different experts consider dialogue differently, but most would describe it in either or all of three primary categories: (a) one-on-one synchronous discussion, (b) group synchronous discussion, and (c) asynchronous discussion. As with most differing methodologies, each of these processes has its own distinct advantages and disadvantages.

One-on-one synchronous discussion refers to what many people call a chat. Popular delivery vehicles for one-on-one chat include programs like Yahoo! Messenger and Microsoft's Instant Messenger, among others. To use this technology, one need only start the application that resides on their local computer, select the name or pseudonym of the targeted individual, and begin typing when the window on the computer's screen opens. Real-time, one-on-one discussions provide some of the same benefits in application that phone or face-to-face conversations allow. Advantages of one-on-one chat include the ability to conduct real-time question-and-answer sessions that provide for a more personal approach than group discussions or those found in typical classroom settings.

Group synchronous discussion refers to what many people call chat rooms. Yahoo!, Microsoft, Google, AOL, and a host

of other service providers offer chat-room delivery platforms for both social and professional purposes. Most chat rooms require users to subscribe to their respective services. Some charge a nominal fee and others charge no fee to use their services. When starting up a particular service, a program downloads to the random-access memory of the user's computer. The user then selects and logs into the room of choice using his or her name or pseudonym. This method allows several people to join a discussion at the same time as information appears in a real time, bulletin-board manner. One of the main advantages of group chat includes the provision for people to brainstorm ideas and receive feedback across vast distances in real time. Additionally, clear documentation exists after the fact, regarding who says what and how much. Depending on hardware, software, bandwidth, and other infrastructure limitations, either form of synchronous discussion might additionally offer two-way voice and streaming video communications directly over the Internet or intranet. Downsides of either form of synchronous discussion include the logistical coordination of events and the need to enter the discussion well prepared.

Asynchronous discussion refers to both e-mail and to the classic bulletin-board approach where members of particular board groups post sequential comments and responses on given topics. Although instructors and students usually use e-mail on ongoing bases, researchers have conducted few studies that identify any relevant theoretical constructs concerning e-mail as it pertains to enhancing online education. However, researchers have conducted a substantial number of investigations on asynchronous bulletin-board discussions as an educational method. Some advantages of this method include allowing members of the group to read and respond to the information on the board at their leisure, lower demand for activity coordination among members, the time allowed for responders to develop deeper and more thoughtful responses to existing posts, and the likelihood that more people will actually participate in the event. Some researchers suggest that it also offers better opportunities for critical thinking and deep learning, resulting in metacognition (Havard, Du, & Olinzock, 2005; Weigel, 2001).

THE ASYNCHRONOUS METHOD

Several of the current leaders in distance education, including Cappella University, The University of Phoenix, and Regent University, along with scores of other colleges and universities, employ the asynchronous method of instruction as their central delivery systems with Blackboard and WebCT providing popular instructional delivery system packages that include bulletin-board conduits. These institutions usually require students to complete some formal training on the methods of delivery and response early in their programs. Then, before each course, instructors and administrators provide students with course requirements, syllabi, and other pertinent information through e-mail, bulletin boards, or both. Typically, during grading periods, students submit and instructors or other graders provide feedback on both major and minor projects via e-mail. Concurrently, instructors or other designated moderators periodically post topics for discussion. They also solicit participation from and provide feedback for students concerning those given topics. Additionally, through many software products, students have access to communication and course tools that provide for announcements, collaboration with other students on group projects, the development of group and individual Web pages, grade checking, and course materials beyond those initially provided (Johnson & Rupert, 2002; Hutchins, 2001).

The stated purpose of much asynchronous dialogue includes helping students to understand the relevance of concepts. Students demonstrate their mastery by contextualizing these concepts through discussion and application. Moderators can run asynchronous discussion boards using several different styles. Some styles that have demonstrated significant value include (a) point-counterpoint-response, (b) open forum, (c) two-sided debate, and (d) multiple posts of specific word counts. Point-counterpoint-response offers the opportunity for an initiator to post an initial point and justification, a respondent to counter the initial point, and the initial poster to rebut the counter. An open forum provides opportunities for students to develop relevant, theoretical conceptualizations and to post concise, significant contributions to the topic. Two-sided debate requires students to divide into two groups and to debate the strengths and weaknesses of given positions germane to the given topic. Multiple posts of specific word counts challenge students to move more deeply into forum concepts as they seek clarification and challenge other students' positions on the given topic. In each of these processes and in many other asynchronous bulletin-board learning environments, student participants use the process of threaded dialogue or discussion to enhance the information exchange process. In a threaded dialogue, the software package groups contributed messages together to form an easily traced thread of information. When using an asynchronous bulletin board, threading discussion entries allows respondents the opportunity to consider thoughtfully their responses before posting, thus potentially aiding in the quality and depth of the respondent's contribution to the forum. Researchers offer that threading might actually result in improved student contributions to discussions, critical thinking, and ease of communicating (Blair, 2002; Dennen, 2005; Smith, Ferguson, & Caris, 2002).

Preparation

To ensure success, both instructors and students have several items to consider in using a given asynchronous discussion method. Instructors first prepare by mastering the particular

delivery vehicle deployed by the educational institution. This may include software packages like those already mentioned or it may include some other software solution outside of the mainstream. However, once the instructor masters the appropriate package, preparation of the structure and content of the pedagogy begins and continues for as long as the institution offers educational programs online. In that regard, several researchers find that online instructing requires a different approach from classroom instructing. In fact, some experts imply a requirement for a paradigm shift for instructors. Some suggest that with asynchronous, online learning, instructors now have to focus on facilitating learning in new ways and with different approaches for online class discussion, including initiation, conclusion, and feedback, and that they might possibly provide very little in the way of content (Dennen, 2005; Smith et al., 2002). Finally, other scholars suggest that instructors preparing to enter into the world of online instruction, especially for the first time, should best experience it by fully immersing themselves in the process. They recommend that instructors should learn to teach online by first participating in training classes actually conducted online (Cook, 2007; Hewett & Ehmann, 2004).

Likewise, students must prepare for online dialogue or discussion by learning the requirements of the delivery vehicle, as well as the expectations, procedures, rules, and standards developed and enforced by their particular institution. Moreover, structure or the lack of structure directly affects students' eagerness and aptness to participate in asynchronous discussions. Other considerations for effective preparation include the following: (a) Students participate more in relevant and goal-based activities, (b) students participate more when they understand an activity's associated learning objective, (c) moderate instructor presence contributes to the learning process, especially if it includes well-designed, conversational contributions, and (d) appropriate feedback designed to meet students' needs contributes to the learning process (Dennen, 2005).

Performance

Many instructors cite the central challenge of online instruction as the time commitment involved. Some experts say that online courses require much more of a time investment than traditional, in-class educational courses. However, others suggest that online instruction requires no more time, but that it seems like it does because the timing now spreads across a 24-hour day (Dennen, 2005; Young, 2002). Other identified issues in performance include the inability of instructors to observe students and their reactions, and the inability of instructors to use classroom theatrics prevent on-the-spot, as-needed adjustments that in classroom settings could otherwise enhance learning processes (Blair, 2002; Dennen). Conversely, researchers have shown that for any limitations one may find regarding Internet education, the Internet finally provides an avenue to enable students to realize the educational dialogues that educators and researchers consider fundamental to situated cognition and social constructivism (Hung et al., 2005).

In the Internet age, students often demonstrate outstanding skills in computer and Internet usage. However, comparing performance success on a computer or on the Internet to that of using dialogue or a discussion board might correspond to the similarities and differences between communication and rhetoric. One construct might prepare an individual to accomplish the other, but the foundation might not provide an accurate indicator for the outcome. Researchers have identified significant differences in students' approaches and involvement in Net-based educational opportunities and have further identified some factors for consideration. These include the following.

1. Not all students appreciate the possibilities for learning through collaboration and some students lack the ability to cooperate.
2. A student's self-confidence (self-efficacy) concerning success in studying contributes to his or her involvement and success in completing dialogue work on the Internet.
3. Students with nonacademic backgrounds demonstrate higher levels of success than those with traditional academic backgrounds, as measured by test scores.
4. Students who participate more in quantity to the dialogue process generally contributed higher quality submissions (Dennen, 2005; Jakobsson, 2006).

In any educational endeavor, grades are at least as important for some people as the learning itself. Testing has consistently concerned distance educators, primarily because of the lack of direct educator supervision. Some institutions address these issues by requiring proctored examinations on campus or at some acceptable testing center. However, experts suggest that one can easily assess the quality and quantity of threaded discussions and that it thereby provides a relatively objective method for graders to assign grades based on participation. They further suggest that when students write all of their communication, most graders could easily assess whether or not the student submitting the examination actually wrote it based on writing style alone (Smith et al., 2002).

FUTURE TRENDS

It is clear that the volumes of distance education generally and learning by asynchronous discussion in particular will only increase as the future unfolds. On an almost daily basis, the number of institutions offering greater numbers of courses in online formats seems to increase. This means that in the

not-so-distant future, the competition will also increase between institutions for instructors who have mastered the processes associated with asynchronous learning. Moreover, as the Internet provides a truly global marketplace, educational institutions will continue to expand into cities, towns, and villages around the world, presenting opportunities to groups and individuals who might never have otherwise been able to experience the quality of instruction provided online. Some universities already offer students opportunities to participate in asynchronous discussions that cover core competencies, which all students in the curriculum must master. However, those students in places all around the globe must also meet together in pockets close to home where they experience learning events that help them integrate what they have learned online into knowledge that applies in local applications. These programs and others like them are sure to provide continued opportunities to form a global educational neighborhood.

Technologically, the future appears to hold further integration of classroom and asynchronous learning support services as instructors produce Web-based assignments for local classrooms in addition to distance programs. Education providers also will continue to integrate databases and systems to provide broader ranges of services in both real-time and asynchronous environments, for both unique and routine interactions. The concerns that some experts express regard the fact that the culture of education has not kept pace with the technological explosion. Consequently, administrative and regulatory barriers will likely have a neutering effect on the full utilization of technology. Conversely, some theorists project that rather than broadening and deepening existing programs, the technological boom will enable educators and institutions to offer highly customized curriculums designed to meet specific needs of individuals and small groups of learners. To these ends, the technology will continue to offer greater levels of student-to-instructor and student-to-student interactions, learning activities, design flexibility, and cost effectiveness (Garrison & Anderson, 1999; Saba, 2005).

CONCLUSION

This article provided a topical review of correspondence courses, television and radio, and online distance education delivery systems. It specifically discussed synchronous and asynchronous processes of online dialogue, including preparation by instructors or curriculum developers and students. It also provided a review of the best practices and potential drawbacks generally associated with educational dialogue and particularly associated with asynchronous dialogue as used in educational processes. It specifically covers the capabilities, advantages, and disadvantages of the three primary categories of online educational dialogue. It also defined threaded dialogue and provided an overview of suggested methodologies and other structural components, which if used by the educator and students, could enhance the overall online educational process. It further discussed the ongoing requirement that exists for pedagogical development after instructors master the techniques associated with the particular delivery systems employed by their institutions. Understanding the advantages and limitations of these processes, as well as mastery of the techniques listed herein, can provide both teachers and students with an overall better learning environment and enhanced online educational experiences.

REFERENCES

Blair, J. (2002). The virtual teaching life. *Education Week, 21*(35), 31-34.

Cook, K. C. (2007). Immersion in a digital pool: Training prospective online instructors in online environments. *Technical Communication Quarterly, 16*(1), 55-82.

Crow, S. M., Cheek, R. G., & Hartman, S. J. (2003). Anatomy of a train wreck: A case study in the distance learning of strategic management. *International Journal of Management, 20*(3), 335-341.

Dennen, V. P. (2005). From message posting to learning dialogues: Factors affecting learner participation in asynchronous discussion. *Distance Education, 26*(1), 127-148.

Fullick, P. L. (2006). Synchronous Web-based communication using text as a means of enhancing discussion among school students. *Campus-Wide Information Systems, 23*(3), 159-170.

Garrison, R. D., & Anderson, T. D. (1999). Avoiding the industrialization of research universities: Big and little distance education. *American Journal of Distance Education, 13*(2), 48-63.

Havard, B., Du, J., & Olinzock, A. (2005). Deep learning: The knowledge, methods, and cognition process in instructor-led online discussion. *Quarterly Review of Distance Education, 6*(2), 125-135, 182-183.

Hewett, B. L., & Ehmann, C. (2004). *Preparing educators for online writing instruction.* Urbana, IL: NCTE.

Hung, D., Tan, S. C., & Chen, D. (2005). How the Internet facilitates learning as dialog: Design considerations for online discussions. *International Journal of Instructional Media, 32*(1), 37-46.

Hutchins, H. M. (2001). Enhancing the business communication course through WebCT. *Business Communication Quarterly, 64*(3), 87-94.

Jakobsson, A. (2006). Students' self-confidence and learning through dialogues in a Net-based environment. *Journal of Technology and Teacher Education, 14*(2), 387-405.

Johnson, A., & Ruppert, S. (2002). An evaluation of accessibility in online learning management systems. *Library Hi Tech, 20*(4), 441-451.

Saba, F. (2005). Critical issues in distance education: A report from the United States. *Distance Education, 26*(2), 255-272.

Smith, G. G., Ferguson, D., & Caris, M. (2002). Teaching over the Web versus in the classroom: Differences in the instructor experience. *International Journal of Instructional Media, 29*(1), 61-67.

Weigel, V. (2001). *Deep learning for a digital age: Technology's untapped potential to enrich higher education.* San Francisco: Jossey-Bass.

Wojnar, L. (2002). Research summary of a best practice model of online teaching and learning. *English Leadership Quarterly, 25*(1), 2-9.

Young, J. R. (2002). The 24-hour professor: Online teaching redefines faculty members' schedules, duties, and relationships with students. *The Chronicle of Higher Education, 48*(38) A.31.

KEY TERMS

Asynchronous: In distance education, asynchronous refers to communication in a learning process that is not necessarily immediate. This includes methods like Web logs (blogs), bulletin boards, e-mail, and correspondence.

Distance Education: It refers to the processes of teaching and learning in environments other than those associated with traditional, on-campus classrooms. These processes include correspondence, usually through the postal service, Internet, or intranet via synchronous learning like Webcasts and real-time messaging; via asynchronous methods like e-mail and bulletin boards; and through two-way audio and video feeds.

Self-Efficacy: It is the belief one possesses about his- or herself that suggests he or she has the wherewithal to achieve a given outcome.

Situated Cognition: Stemming from pragmatism, theories regarding situated cognition suggest that one learns better in contextual circumstances than in classroom or laboratory settings.

Social Constructionism: This refers to a theory of knowledge that suggests people collectively develop their methods of interacting with each other. Members of a given society or culture might assume or otherwise perceive one of the particular methods they employ as a natural way to think or feel. However, upon deconstruction, one might discover that the method originated through a number of decision-making processes.

Synchronous: In distance education, synchronous refers to real-time communication in a learning process. This includes methods like individual and group chat and parallel audio and video streaming.

TELETECHNET: This refers to a distance education tool pioneered by Old Dominion University (ODU) in Norfolk, Virginia. TELETECHNET consists of an integrated system of computers and two-way audio and video feeds whereby ODU links with other institutions to provide live classroom instruction, at both the undergraduate and graduate level, to place-bound students throughout the world.

Thread: The topic thread refers to hierarchical, asynchronous postings that focus on a particular topic. Threading generally promotes reflection and considered responses, and therefore overall deeper learning than that found in similar, synchronous methods.

Overview of Electronic Auctions

Patricia Anthony
Universiti Malaysia Sabah, Malaysia

INTRODUCTION

Online auctions are one of the most popular and effective ways of trading goods over the Internet (Bapna, Goes, & Gupta, 2001). Thousands of items are sold on online auctions everyday including books, toys, computers, antiques, and even services. As an example, eBay (http://www.ebay.com), the largest online auction house, has more than 241 million registered users today, and in the year 2006 alone, eBay recorded consolidated net revenue of $6 billion (eBay, 2007). On any given day, there are more than 78 million items listed on eBay, and approximately 6 million listings are added per day. It has also spread its wings to other countries outside the USA, and to date it is present in 24 countries. In addition to eBay, there are more than 2,600 auction houses that conduct online auctions (common sites include Amazon.com, http://www.amazon.com; Yahoo! Auctions, http://auctions.shopping.yahoo.com; priceline.com, http://www.priceline.com; and uBid, http://www.ubid.com). These auction houses conduct many different types of auctions according to a variety of rules and protocols.

BACKGROUND

An auction is defined as a bidding mechanism described by a set of auction rules that specify how the winner is determined and how much he or she has to pay (Wolfstetter 2002). Auctions are widely used in many transactions including the sale of arts, wine, fresh products, diamonds, and real estate. In fact, auctions are not new. They were used to allocate scarce resources in Babylon from about 500 B.C. (Shubik, 1983). During those times, an annual Babylon marriage market was conducted where men had to bid for their prospective wives. The richest Babylonians who wished to wed had to bid against each other for the loveliest maiden, while the poorer ones had to settle for the less beautiful ones. Moreover, in ancient Rome, auctions were used in commercial trade to liquidate property and to sell surplus spoils of war on the battlefield (including plundered booty). Since then, auctions have flourished in many other civilizations, but they became more prominent in the 17th century where they were used to liquidate goods and to sell unsalable goods (such as domestic animals, tobaccos, natural resources, horses, and slaves) at the end of the season.

The practice of auctioning goods has been popular throughout the years because auctions are an extremely effective way of allocating resources to the individuals who value them most highly (Reynolds, 1996). This effectiveness means that very many variants have been produced (Wurman, Wellman, & Walsh, 2001); however, there are four main types of single-sided auctions that are commonly and traditionally held (Klemperer, 1999):

- the ascending-bid auction (also called the open, oral, or English auction),
- the descending-bid auction (also called Dutch auction),
- the first-price sealed-bid auction, and
- the second-price sealed bid auction (also called Vickrey auction).

In more detail, in an English auction, the auctioneer starts the auction with a low price that is then successively raised until one bidder remains. That remaining bidder wins the object and pays a value equivalent to his or her bid value. This type of auction is commonly used when selling antiques, artwork, and houses. The descending auction is the opposite of the ascending-bid auction. The auctioneer starts at a very high price and this price is then progressively lowered until there is a call from any bidder to claim the item. The first bidder who calls out wins the object at the current price. This auction is also called the Dutch auction because it is used in the sale of flowers in the Netherlands (van Heck & Ribbers, 1997). Fish and tobaccos are also sold in a similar way in Spain, Israel, and Canada (Klemperer, 1999).

The last two auctions are sealed-bid auctions. In the first-price sealed-bid auction, each bidder submits a single bid independently without knowing what the others bid. When the auction closes, the bids are opened and the winner is the bidder with the highest bid; he or she gets the item at a price equivalent to the bid value. First-price sealed-bid auctions are used in auctioning mineral rights in government-owned land and are also sometimes used in the sales of artworks and real estate (Klemperer, 1999). Finally, the second-price sealed-bid auction works in the same way as the first-price sealed-bid auction, but the price paid by the winner is equivalent to the second-highest bid. This type of auction is widely used for auctioning stamps, autographs, and Civil War memorabilia by mail (Lucking-Reiley, 2000b; Rothkopf, Teisberg, & Kahn, 1990).

Against this background, an online auction can be defined as an Internet-based version of a traditional auction (Jansen, 2003). The major difference between the two types is the additional degree of flexibility in the way the online variety is conducted. Specifically, in a traditional auction setting, the auctioneer and the bidders gather in one room at a given time to decide who gets the item and at what price. Such auctions generally last only for a few minutes or even seconds for each item sold. This rapid process gives very little time to the auction participants to make decisions, so they may decide not to bid in the auction. As a consequence, the sellers may not get the highest possible price for their goods (Turban, Lee, King, & Chung, 2000). Apart from that, bidders are usually required to come to the auctions, and this practice leaves out many potential bidders that may not be able to attend the auction sessions. It may also be very difficult for the sellers in traditional auctions to move their goods to the auction site (especially when the items are bulky), and there may be a large cost associated with operating the auction since the sellers have to rent the auction site, auctioneers and other employees need to be hired, and the auction needs to be advertised in advance.

Things are somewhat different in the online case. In particular, many of the geographical and temporal limitations of the traditional auctions are removed (Lucking-Reiley, 2000a). Specifically, the consumers can be sitting in the comfort of their homes while participating in an online auction that may be located many thousands of miles away. Moreover, online auctions generally last for days and weeks, giving the bidders more flexibility about when to submit bids. Online auctions also allow sellers to sell their goods efficiently and with little action or effort required. Apart from that, sellers have fewer problems of getting a large group of bidders together on short notice because of the availability of a large number of online bidders distributed across the globe. This creates a larger market for the goods on sale.

In summary, online auctions provide a selection of goods that Internet communities can buy or sell, allowing the consumers a greater chance of getting their goods and the sellers a greater chance of selling their goods. Since online auctions offer a wide array of items, generally speaking, there are likely to be many online auctions that sell similar items. For example, a search for the term *digital camera* using the auction search engine AuctionSeek (http://www.auctionseek.net) yields 49,768 results. The current bid prices range between $1.00 and $18,000, and the duration of these auctions range between less than a minute and a week. This means it is quite possible for buyers to get a bargain based on the fact that there is a wide selection of choices with a variety of prices. Online auctions have also been used by sellers to dispose of aging items, unwanted items, and excess items to the community of buyers who may require them. This is the reason why online auctions are an ideal place to search for collectors' items and for hard-to-find items (Turban, 1997). From the sellers' perspective, online auctions are the perfect place to trade goods due to the presence of a huge number of bidders that are distributed globally (Lucking-Reiley, 2000a). There are of course some disadvantages of online auctions such as the bidder's inability to physically view the item being auctioned off, or the possibility of fraud. However, generally speaking, the benefits outweigh the risks.

ONLINE AUCTION ISSUES AND CHALLENGES

Online auctions have been the focus of attention of many academic researchers because of its complexities and its dynamics. Historical data from eBay are downloaded and analyzed to investigate the bidder behaviors and the bid arrivals, and to predict the closing price of a given auction. However, in order to study these, one has to understand the auction settings, the bidding and selling processes, as well as the different types of auctions that may be employed by the auction houses.

The most common format of online auctions that are used by most auctions houses is the ascending-bid auctions constrained by time. Each auction is given a start time and an end time, and bidders are free to post bids anytime while the auction is on going. The winner is the one that posts the highest bid. However, this type of auction can be further classified into a variety of options. For example, eBay offers five options: reserve-price auction, private auction, multiple-item auction, "buy it now," and best offer. The reserve-price auction allows the seller to fix a reserve price, which is the minimum price that the seller is willing to accept for the item being sold. The seller is not obligated to sell the item if the reserve price is not met.

The seller can also choose to create a private auction in which the buyer's user ID does not appear in the listing or in the listing's bid history. Only the seller is authorized to view the buyer user IDs associated with the listing. This kind of auction is usually used to sell high-priced items or approved pharmaceutical products. The multiple-item auction, also known as the Dutch auction (this is not the same as the traditional Dutch described earlier), allows the seller to offer multiple identical items for sale. Each bidder is required to enter the quantity required and the bid amount for each item. The winning bidders are determined in order of bid price per item and will pay a price equal to the lowest winning bid. A variation of this auction style is the Yankee auction in which the successful bidders pay what they bid.

The buy-it-now option enables the bidder to purchase an item when he or she wants it at a known set price without having to wait for the online auction to end. During the bidding period, the buy-it-now option is only available for a limited time. The best-offer feature allows sellers to receive price-based offers from buyers that can be accepted at the

discretion of the seller. This feature is available for listings in fixed-price auctions and classified ads. Once a buyer's best offer is accepted by the seller, the listing ends.

Another auction format that is commonly used is the reverse auction. In this auction, the bidder is the seller and not the buyer. It is a fixed-duration bidding event hosted by a single buyer, in which multiple suppliers compete for business. The bid reflects how much the buyer is being asked to pay and not how much the good or service is being sold for. These kind of Web-based reverse auctions have become extremely popular for purchasing everything from accounting services to securing raw materials.

Bidding in online auctions is time consuming. Bidders are always looking to simplify the bidding process as well as to ensure that they get good deals. On the other hand, sellers are always looking for ways to dispose of their items with profits. Hence, most auction sites provide a tool for the bidders that can automate the bidding process. This tool, referred to as the proxy bidder, only requires the bidders to key in the maximum bid for a given item and will place bids on the bidder's behalf starting with the next bid increment for the auction. It will only bid as much as necessary to make sure that the bidder remains the highest bidder and will keep bidding until the bidding reaches the maximum amount or until the auction is closed. Despite the convenience, this tool does not give the freedom to bidders to specify the time they wish to place their bids (Bapna, 2003). Moreover, the auction host may try to offer a price very close to the bidder's maximum amount to guarantee that he or she obtains as much profit as possible.

Another popular method employed by bidders is sniping (or also known as late bidding). Sniping refers to the practice of bidding at the last opportunity in online auctions with fixed closing times. Some bidders snipe manually during the last seconds before the auction closes while others use software (known as snipers) to perform the auction sniping. Placing bids at the end of the auction ensures that no other bidders are able to place another bid and almost always guarantee a win in the auction. This practice is not favored by the bidding communities, resulting in some auction houses like Amazon.com and uBid to extend the auction if there are any bids within 10 minutes of the auction close time; there is no limit to the number of times an auction can be extended. This practice of sniping also converts the auction into a single-shot sealed-bid auction.

Those sellers looking for more profit in online auctions resort to all sort of tricks and tactics. One such tactic is shill bidding. Shilling occurs when a seller bids on his or her own auction in an effort to increase the price other bidders need to pay to win the auction (Kauffman & Wood, 2005). It usually occurs in the form of competitive shilling, where the seller starts bidding so that the final bidder is forced to bid higher than would otherwise be necessary to win the auction item. Shilling is considered a criminal fraud in the United States and elsewhere, and the offence is punishable by a jail term or heavy fines. Unfortunately, shilling is one of the online auction frauds that are the hardest to detect.

As discussed, bidding in an online auction is complicated since the bid changes from time to time and the closing price of a given auction is only known when the auction is concluded. First, the bidder needs to decide what item to buy and in which auction to participate. To do this, he or she will need to monitor several auctions to analyze the progression of the auctions. In any given auction, the bidder can access information about the auction such as the current bid value, the bids history, the bidding duration, and the seller's rating. So, using this information, the bidder needs to decide when to bid and at what price to bid. Bidders usually have their own reservation price, which is the maximum price they are willing to bid for a given item. Even if the bidder is able to determine these parameters, this does not guarantee a win in the auction (some other bidder may bid at the last second with a higher bid value). This process needs to be repeated several times until the bidder succeeds in obtaining the item.

Even if the item is obtained, the bidder may end up paying more than the intrinsic value of the item. This is called the winner's-curse phenomenon. This is one of the reasons why prices determined by auctions deviate from standard purchases: Auctions tend to favor those bidders who have most overestimated the value of the item for sale (since the winning bidder is always the one with the highest bid). The winner, therefore, may have overpaid relative to the true value of the item.

To maximize the chances of winning in an auction, bidders can always place bids simultaneously in multiple auctions. While this may reduce the likelihood of not winning an auction, this will also increase the risk of the bidders in that they may end up getting more than a single item. It would be desirable to have software to monitor bids, deciding on which auction to bid as well as determining how much to bid.

On the other hand, sellers want to auction off their items at a reasonable price with a reasonable profit. The timing and the placement of the item on certain auction sites also need to be considered to ensure that the item is sold in a manner consistent with the seller's requirement. A seller may get a higher profit in an auction that has many bidders and when the bidding time is longer.

In view of the auction complexities, many researchers have embarked on a variety of research-related online auctions with varying objectives in order to address the issues highlighted in the previous section. Some notable work includes exploring factors that affect final prices as well as predicting the closing price for a given auction (Ghani, 2005; Jank & Shmueli, 2005; Lucking-Reiley, Bryan, Prasad, & Reeves, 2007), modeling the bidder arrival process (Akula & Menasce, 2005; Shmueli, Russo, & Jank, 2007), studying reputation and trust in online auctions (Dewan & Hsu, 2001; Houser & Wooders, 2006), finding empirical evidence

for late bidding (sniping; Bapna, 2003; Roth & Ockenfels, 2002), analyzing and detecting shill bidding, investigating bidding strategies (Anthony & Jennings, 2003; Jiang & Leyton-Brown, 2007; Shah, Joshi, Sureka, & Wurman, 2003), studying bidders' behavior (Dewally & Ederington 2004; Walley & Fortin, 2005), and developing strategies for sellers (Anthony, 2006; Kim, 2007).

FUTURE TRENDS

Due to the popularity of online auctions, in the future we expect that an emerging technological force, in the form of smarter ubiquitous open-source bidding agents, is likely to have a significant influence on their future practice (Bapna, 2003). Online auctions will be populated by software agents that will bid and sell on behalf of the users. More research on online auctions will be undertaken that will address issues such as online auction closing-price prediction, bidder behavior prediction, the development of bidding strategies that can be utilized by bidders, and the development of seller strategies. More complicated auction protocols are expected to emerge to counter shill bidding and sniping activities.

CONCLUSION

The online auction is still the major contributor to the Net economy and is still growing. Millions of people are hooked on eBay with similar objectives: to look for bargains and to sell goods. We foresee that online auction sites will be populated by software agents that will bid and sell on the customer's behalf. Hence, the economic impact of having software agents in the online auction environment should be further analyzed and investigated.

REFERENCES

Akula, V., & Menasce, D. A. (2004). An analysis of bidding activity in online auctions. In *E-commerce and Web technologies* (LNCS 3182, pp. 206-217). SpringerLink.

Anthony, P. (2006). Strategy for seller agent in multiple online auctions. *International Journal of Intelligent Information Technologies, 2*(4), 1-17.

Anthony, P., & Jennings, N. R. (2003). Developing a bidding agent for multiple heterogeneous auctions. *ACM Transactions on Internet Technology, 3*(3), 185-217.

Bapna, R. (2003). When snipers become predators: Can mechanism design save online auctions? *Communications of the ACM, 46*(12), 152-158.

Bapna, R., Goes, P., & Gupta, A. (2001). Insights and analyses of online auctions. *Communications of the ACM, 44*(11), 43-50.

Dewally, M., & Ederington, L. H. (2004). *What attracts bidders to online auctions and what is their incremental price impact?* Retrieved from http://ssrn.com/abstract=589861

Dewan, S., & Hsu, V. (2001). *Trust in electronic markets: Price discovery in generalist versus specialty online auctions* (working paper). WA: University of Washington.

eBay. (2007). *eBay Inc. announces fourth quarter 2006 and full year 2006 financial results.* Retrieved January 28, 2007, from http://files.shareholder.com/downloads/ebay/BayInc-EarningsReleaseQ42006.pdf

Ghani, R. (2005). Price prediction and insurance for online auctions. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (Chicago, Illinois, USA, August 21-24, 2005): KDD '05* (pp. 411-418). New York: ACM Press.

Houser, D., & Wooders, J. (2006). Reputation in auctions: Theory, and evidence from eBay. *Journal of Economics & Management Strategy, 15*(2), 353-369.

Jank, W., & Shmueli, G. (2005). *Profiling price dynamics in online auctions Using curve clustering* (working paper). MD: Smith School of Business, University of Maryland.

Jansen, E. (2003). *Netlingo the Internet dictionary.* Retrieved January 28, 2007, from http://www.netlingo.com/

Jiang, A. X., & Leyton-Brown, K. (2007). Bidding agents for online auctions with hidden bids. *Machine Learning, 67*(1-2), 117-143.

Kauffman, R. J., & Wood, C. A. (2005). *Electronic Commerce Research and Application, 4*, 21-34.

Kim, Y. S. (2007). Maximizing sellers' welfare in online auction by simulating bidders' proxy bidding agents. *Expert Systems with Application,* 32, 289-298.

Klemperer, P. (1997). Auction theory: A guide to literature. *Journal of Economic Surveys, 13*(3), 227-286.

Lucking-Reiley, D. (2000a). Auctions on the Internet: What's being auctioned, and how? *Journal of Industrial Economics, 48*(3), 227-252.

Lucking-Reiley, D. (2000b). Vickrey auctions in practice: From nineteenth century philately to twenty-first century e-commerce. *Journal of Economic Perspectives, 14*(3), 183-192.

Lucking-Reiley, D., Bryan, D., Prasad, N., & Reeves, D. (2007). Pennies from eBay: The determinants of price in

online auctions. *Journal of Industrial Economics, 55*(2), 223-233.

Reynolds, K. (1996). *Auctions going, going, gone! A survey of auction types.* Retrieved January 28, 2007, from http://www.agorics.com/Library/auctions.html

Roth, A. E., & Ockenfels, A. (2002). Last-minute bidding and the rules for ending second-price auctions: Evidence from eBay and Amazon auctions on the Internet. *The American Economic Review, 92*(4), 1093-1103.

Rothkopf, M. H., Teisberg, T. J., & Kahn, E. P. (1990). Why are Vickrey auctions rare? *The Journal of Political Economy, 98*(1), 94-109.

Shah, H. S., Joshi, N. R., Sureka, A., & Wurman, P. R. (2003). Mining eBay: Bidding strategies and shill detection. In *Proceedings of the Conference on Mining Web Data for Discovering Usage Patterns and Profiles (WEBKDD 2002)* (pp. 17-34).

Shmueli, G., Russo, R. P., & Jank, W. (2007). *The BARISTA: A model for bid arrivals in online auctions* (hdl:1902.1/10643). Institute for Mathematical Statistics.

Shubik, M. (1983). *Auctions, biddings, and markets: An historical sketch.* New York: New York University Press.

Turban, E. (1997). Auctions and bidding on the Internet: An assessment. *Electronic Markets, 7*(4), 30-34.

Turban, E., Lee, J., King, D., & Chung, H. M. (2000). *Electronic commerce: A managerial perspective.* Prentice Hall.

van Heck, E., & Ribbers, P. M. (1997). Experiences with electronic auctions in the Dutch flower industry. *Electronic Markets, 7*(4), 30-34.

Walley, M. J. C., & Fortin, D. R. (2005). Behavioral outcomes from online auctions: Reserve price, reserve disclosure, and initial bidding influences in the decision process. *Journal of Business Research, 58*, 1409-1418.

Wolfstetter, E. (2002) Auctions: An introduction. *Journal of Economic Surveys, 10*, 367-420.

Wurman, P., Wellman, M. P., & Walsh, W. E. (2001). A parametrization of the auction design space. *Games and Economic Behavior, 35*, 304-338.

KEY TERMS

Auction: It is a mechanism described by a set of auction rules that specify how the winner is determined and how much he or she has to pay.

Bid: It is the price offered by a given buyer in an auction at a given time while the auction is on going.

Bid Sniping: This occurs when bids are placed just a few minutes or seconds before an auction ends. This type of bidding is applicable in English auctions that have fixed deadlines (such as the ones conducted in eBay).

Dutch Auction: It is an auction in which the auctioneer starts at a very high price that is then progressively lowered until there is a call from any bidder to claim the item.

English Auction: It is an auction in which the auctioneer starts the auction with a low price that is then successively raised until one bidder remains.

Online Auction: it is an Internet-based version of a traditional auction in which there is no time and location constraints.

Reservation Price: This is the maximum price the bidder is willing to pay for the item being auctioned.

Reverse Auction: It is an auction where the bidder is the seller and not the buyer. The buyer will indicate the requirements and sellers will submit bids.

Sealed-Bid Auction: This is an auction in which each bidder submits a single bid independently without knowing what the others bid.

Shill Bidding: This occurs when a seller bids in his or her own auction in an effort to increase the price other bidders need to pay to win the auction.

Software Agents: These are software capable of making decisions on behalf of the user and are usually endowed with some level of autonomy and intelligence.

An Overview of Enterprise Resource Planning for Intelligent Enterprises

Jose M. Framinan
University of Seville, Spain

Jose M. Molina
University of Seville, Spain

INTRODUCTION

Enterprise resource planning systems can be defined as customizable, standard application software which includes integrated business solutions for the core processes and administrative functions (Chan & Rosemann, 2001). From an operative perspective, ERP systems provide a common technological platform unique for the entire corporation allowing the replacement of mainframes and legacy systems. This common platform serves to process automation as well as to simplify current process either by an explicit reengineering process or by the implicit adoption of the system "best practices" (Markus & Tanis, 2000). Finally, the common centralized platform allows the access to data that previously were physically or logically dispersed. The automation of the processes and the access to data allows the reduction of operating times (thus reducing operating costs) while the latter serves to a better support of business decisions (see e.g., Umble, Haft & Umble, 2003 for a detailed review of ERP benefits). ERP is considered to provide businesses with new opportunities to acquire knowledge (Srivardhana & Pawlowski, 2007), being the sources of knowledge the aforementioned best practices from the ERP, and the ERP software company's staff during the implementation phase.

At present, ERP systems are either used or implemented in a large number of enterprises. According to Genoulaz and Millet (2006), up to 74% of manufacturing companies and up to 59% of service companies use an ERP system. In addition, more than 70% of Fortune 1000 companies have implemented core ERP applications (Bingi, Sharma, Godla, 1999; Yen, Chou & Chang, 2002). The objectives for implementing an ERP system can be classified as operational, strategic, dual (operational plus strategic), or without objective (Law & Ngai, 2007). The adoption of an ERP system with operational objectives is aimed at improvement operating efficiency together with the reduction of costs, while companies implementing ERP with a strategic objective would experience a change in business processes, improving sales and market expansion.

A widespread critique to ERP systems is their high total cost of ownership (Al-Mashari, Al-Mudimigh & Zairi, 2003) and hidden costs in implementation (Kwon & Lee, 2001). Besides, ERP systems impose their own logic on an organization's strategy and culture (Davenport, 1998), so ERP adopters must adapt their business processes and organization to these models and rules. Consequently, organizations may face difficulties through this adaptation process which is usually carried out without widespread employee involvement. This may cause sore employees, sterile results due to the lack of critical information usually provided by the employees; and also late delivery, with reduced functionality, and/or with higher costs that expected (Kraemmeraard, Moeller & Boer, 2003). Additionally, some analysts have speculated that widespread adoption of the same ERP package in the same industry might lead to loss of competitive advantage due to the elimination of process innovation-based competitive advantage (Davenport, 1998). This has been observed, for instance, in the semiconductor manufacturers sector (Markus & Tanis, 2000).

The early stage of ERP was carried out through Materials Requirement Planning (MRP) systems (Umble, Haft & Umble, 2003). The next generation of these systems, MRP II (Manufacturing Resources Planning), crossed the boundaries of the production functionality and started supporting not only manufacturing, but also finance and marketing decisions (Ptak & Schragenheim, 2000). Current ERP systems appeared in the beginning of the 1990's as evolved MRP II, incorporating aspects from CIM (Computer Integrated Manufacturing) as well as from EDP (Electronic Data Processing). Therefore, ERP systems become enterprise-wide, multilevel decision support systems. ERP systems continue evolving, incorporating Manufacturing Execution Systems (MES), Supply Chain Management (SCM), Product Data Management (PDM), or Geographic Information Systems (GIS), among others (Kwon & Lee, 2001).

BACKGROUND

Most enterprise resource planning systems share a number of common characteristics, both from a technological as well as a business perspective. These include:

- **Client/server, open systems architecture**. Most ERP packages adopt an open systems architecture that separates data (database server), application (ERP server), and presentation (user interface/ERP client) layers, guaranteeing cross-platform availability and systems integration (Basoglu, Daim & Kerimoglu, 2007). In order to interoperate with existing business applications or information systems, most ERP adhere to the majority of common standards for data exchange or distributing processing.
- **Enterprise-wide database**. One of the most distinguishable characteristics of ERP is the strong centralization of all relevant data for the company (Al-Mashari, Al-Mudimigh & Zairi, 2003). When physical centralization is not possible, communication and/or replication protocols among the different databases should be implemented in order to ensure data consistency and accessibility throughout the entire enterprise.
- **Kernel architecture**. Some ERP systems support more than 1,000 different business functionalities (Bancroft, Seip & Sprengel, 1998), covering nearly all-relevant business aspects for most of the enterprises. As all these functionalities cannot be loaded in the ERP server at the same time, the majority of ERP systems employ a so-called "kernel architecture". In this architecture, most functionalities are stored in the ERP database, usually in the form of source code of a proprietary, fourth generation, programming language. When certain functionality is required by an ERP client, the ERP server loads it from the database and compiles the corresponding code so the functionality is made available for the clients. Once it is not required, the functionality is removed from the ERP server. Note that this mechanism also allows for an easy enhancement/updating of existing functionalities as well as for the construction of new ones.
- **Process-oriented, business reference model**. ERP is process-oriented software that has been developed starting from an implicit or explicit business reference model in order to appropriately describe the relevant business functions covered by the ERP system. For most ERP vendors, this model is explicit and takes the form of the best practices extracted form the ERP vendor experience (Markus & Tanis, 2000). This can be used to analyze and evaluate current business processes in the enterprise prior to the implementation of the ERP package, serving thus as benchmark processes for business process reengineering (BPR).
- **Adaptation to the enterprise**. In order to meet the specific requirements of different enterprises, ERP systems are highly configurable. This potential for customization is considered to be one of the main differences between ERP and other standard software packages (Kraemmeraard, Moeller & Boer, 2003). The customization process may take several months, or even years, depending on the enterprise.
- **Modularity**. Although the term "ERP system" is usually employed to design a system covering all corporate functions (Slater, 1998), generally an ERP system is composed of a set of ERP modules. An ERP module is a group of function-oriented, tightly integrated, functionalities which in many cases can be separately purchased and installed. Typical ERP modules are the financial-accounting module, production-manufacturing module, sales-distribution module, or human resources module. This allows enterprises to purchase only these modules strictly required as well as the possibility of integrating them with existing information systems.

An intelligent enterprise is an organization which acts effectively in the present and is capable to deal effectively with the challenges of the future (Wiig, 1999). Since most enterprises operate today in a complex and dynamic environment characterized by increasing competition and continuous changes in products, technology and market forces, an intelligent enterprise should be proactive, adaptable, knowledgeable, and well-resourced (Kadayam, 2002). In order to achieve this behavior, it is expected that all employees in the intelligent enterprises not only deliver the work products that are directly associated with their functions, but that they also innovate to improve customer relationships, enterprise capabilities, and envision opportunities for new products and services (Wiig, 1999). Therefore, it is clear that an intelligent organization should have timely access to all critical information in order to gain insight into its performance and should be able to provide effective decision support systems. Hence, one of the requisites for the intelligent enterprise is the availability of all relevant data in the organization. Indeed, access to the right information is considered to be one of the key characteristics of intelligent enterprises (Smirnov, Pashkin, Chilov & Levashova, 2003).

FUTURE TRENDS

One of the main problems presented by ERP systems is the large amount of time needed to carry out their implementation. The length of these periods tends to be reduced even to a period of weeks and 2-3 months at most (Jacobs & Weston, 2007), as the majority of companies have realized the benefits gained from short implementation cycles. This reduction of the implementation period is also achieved by means of the so-called "pre-customized" packages, which essentially include tailored modules with default values usually depending on the different sectors of activity.

Up to date, ERP systems lack or have very little "intelligence". Data mining and intelligence tools would lead to an increase in the use of these systems in the decision-making business, thanks to the expert systems and advanced planning (optimization) (Jacobs & Weston, 2007). If we adopt the generic intelligent enterprise architecture by Delic and Dayal (2002), ERP addresses issues of supply chain efficiency and back-office optimization, and provide the basis for Enterprise Knowledge Management (EKM). At the same time, the evolution of enterprises to the form of intelligent organizations requires the cooperation of independent companies into a virtual multitier enterprise (Olin, Greis & Kasarda, 1999), the Internet providing the glue for their heterogeneous information systems (Delic & Dayal, 2002). In order to achieve this, one of the main trends followed by most ERP systems vendors is the introduction of Internet (Chan & Rosemann, 2001; Kwon & Lee, 2001). The use of Web services within the scope of these systems is believed to provide two major advantages: firstly, the ease of integration, and the cost reduction through the hosted application model (Tarantilis, Kiranoudis & Theodorakopoulos, 2006). The adoption of the Internet can be seen from two viewpoints, i.e. the user interface viewpoint, and the internal/external communication viewpoint. With respect to the user interface, ERP systems are transaction-oriented. However, the connectionless nature of the Internet protocols (i.e., the connection between the Web server and the browser is not maintained after the former has sent the requested data to the later) makes it not well suited for transactions. Therefore, it is intrinsically difficult to adapt the ERP internal structure to the Internet. As a consequence, most of the ERP vendor's effort is on creating reliable gateways between the ERP system and an Internet server.

Regarding the internal or external communication of the ERP system, the emphasis is done in the adoption of the Internet standards for data exchange. This is done with respect to both the exchange of data among the different ERP modules and to the exchange of data among the ERP system and external applications. Hopefully, this effort will result in the adoption of a common communication standard that will allow the integration of the information systems of customers and/or providers in a supply chain. Additionally, it will make feasible the so-called "component ERP": that is, the acquisition of the "best-of-breed" modules from every ERP vendor (Fan, Stallaert & Whinston, 2000). Since communications among ERP modules has been driven by proprietary protocols, the ERP market has been forcing the enterprises to purchase all modules of the ERP system from the same vendor or face huge costs in developing interfaces for modules from different ERP systems. This may be greatly simplified by the adoption of a public, common, protocol standard such as those in the Internet. Even in the most likely case that interfaces between modules from different vendors are still required; the decrease in the cost of their development may render it affordable for the enterprises (Appleton, 1997; Kwon & Lee, 2001).

Another trend is caused by the high cost of acquisition and maintenance of ERP packages, which makes them too expensive for SMEs. To solve this, there are several open source ERPs. The implementation of these open source packages has the same complexity than those of proprietary ERP systems, but with lower costs of licenses, reduces dependency on a limited group of suppliers, and increased adaptability of the software. At present, the open source ERP packages include Compiere, ERP5, Openbravo ERP, Fisterra, OFBiz, among others (Serrano & Sarriegi, 2006).

THE ERP MARKET: VENDORS AND MARKET TRENDS

The world ERP market has been growing at a rate between 3%-13% per year in the period between 2000 and 2004 (AMR research reported in 2005 that the market grew by 14% in 2004 and became a $23.6 billion business, see Basoglu, Daim & Kerimoglu, 2007). AMR research expects that the annual growth of the market between the years 2006 and 2009 will be in the range of 6%-7%. Nowadayas, the ERP market is the largest segment of company's application budget (34%), and it is expected to remain so (Aloini, Dulmin & Mininno, 2007; Basoglu, Daim & Kerimoglu, 2007; Scott & Shepherd, 2002; Somers & Nelson, 2004). This growth has been boosted by a number of reasons, such as globalization, market maturity in developed countries, and advances in information and communication technologies, among others.

Market analysis shows the enormous fragmentation of the ERP market, where more than 100 products are targeted as ERP (see APICS, 2000). With respect to the main market players, by 2002 it was composed by the following companies: SAP, Oracle, PeopleSoft, and JD Edwards, as some former major players such as Baan had disappeared (Jacobs & Weston, 2007). At the end of 2002, JD Edwards (whose products were strong in manufacturing, accounting, and finance) and PeopleSoft (expert on human resources) contemplated the possibility of merging, resulting in a company larger than Oracle, its main competitor along with SAP. The merger was announced in June 2003. A few days later, Oracle spears hostile to a takeover PeopleSoft, causing the breakdown of the agreement that came with JD Edwards. After a period of hesitation, because the takeover offer by Oracle had raised antitrust issues, the takeover was consummated in 2005. Currently, the market is composed of two large companies such as SAP and Oracle, but with the abilities of the five initial companies. It is not easy to offer precise information about market shares, since the comparative analysis of the published results show a great dispersion depending on the sources. However, there is consensus that SAP AG is the world market leader with the product SAP

R/3, being its market share around 40%, while the market share of Oracle is approximately 22% (Basoglu, Daim & Kerimoglu, 2007).

THE ERP IMPLEMENTATION PROJECT: RISKS AND KEY SUCCESS FACTORS

The ERP acquisition and implementation constitutes a risky project that may result, in a high number of cases, in unsatisfactory, if not failed, system implementations. It has been reported that nearly three-fourths of ERP implementation projects are judged unsuccessful by the ERP implementing form (Griffith, Zammuto & Aiman-Smith, 1999). ERP implementation involves a significant commitment of time and money (King & Burgess, 2006), ranging from $300,000 to several million dollars depending on the size of the enterprise (Heizer & Render, 2003). Cases of failures in well-known organizations such as Boeing (Stein, 1997) or Siemens (Seidel & Stedman, 1998) have been described. Furthermore, these cases involve software of all primary ERP vendors (Motwani, Mirchandani, Madan & Gunasekran, 2002). More detailed reports (see in Cunningham, 1999 a study of 7,400 IT projects) confirm that nearly 35% of ERP implementation projects are late or over budget, 31% are abandoned, scaled or modified, and only 24% are completed on time and in budget. For this reason, studies that seek to identify factors that are critical (so-called Critical Success Factors or CSFs) for achieving a successful implementation have been carried out (see e.g., Akkermans & van Helden, 2002; Finney & Corbett, 2007; Holland & Light, 1999; Hong & Kim, 2002; Remus, 2007; Somers & Nelson, 2001). In this line, Finney and Corbett raised a methodology to identify CSFs, and they found 26 types of CSFs that may be classified according to the classification by Holland and Light (1999) into strategic and tactical. Strategic CSFs include top management commitment and support, visioning and planning, build a business case, project champion, implementation strategy and timeframe, vanilla ERP, project management, change management, and managing cultural change. Tactical CSFs include balanced team, project team: the best and brightest, communication plan, empowered decision makers, team morale and motivation, project cost planning and management, BPR and software configuration, legacy system consideration, IT infrastructure, client consultation, selection of ERP, consultant selection and relationship, training and job redesign, troubleshooting/crises management, data conversion and integrity, system testing, post-implementation evaluation. In addition, it is possible to organize CSFs into organizational and technological (see Esteves & Pastor, 2000).

According to Tarantilis, Kiranoudis, and Theodorakopoulos (2006); Genoulaz and Millet (2006), and Amoako-Gyampah (2007), the implementation of an ERP system is associated with a range of benefits for the company. Some of these include the following:

- Through streamlining, improving, and controlling business processes more meaningful, there is an increase in the operation of the company.
- Cost reduction and execution times of the most important business processes.
- Support for the management of the supply chain through the integration and synchronization of all activities within it.
- Using the module management, the enterprise has the ability to handle matters related to personnel and costs.
- Improving the use of the system of quality management of the company, allowing it to avoid paperwork, reduce staff tasks related to quality, among others.
- By complementing the management module, it gets a production scheduling more flexible and efficient.
- Design of an integrated information system that would eliminate multiple data sources, facilitating proper information flows and communication among different departments of the company in real time.
- Exploiting abilities to control sales and promotional activities, receiving data on each technique promotional to increase the efficiency of the sales department

CONCLUSION

ERP systems have been considered one of the most noteworthy developments in information systems in the past decade. ERP systems are present in most big companies that operate in the new millennium. Their advantages in terms of access to information or the integration of business functions has been outlined. However, ERP implementation projects are not risk-free: rates of ERP implementation failures are rather high. Although the failure figures may be partly explained by the intrinsic complexity of the ERP implementation project, some others may be minimized by the consideration of the ERP implementation as a strategic decision in the enterprise, and thus a principal, long-term project rather than a single information system change.

ERP systems will play a central role in the intelligent enterprise of the future. ERP vendors are continuously adding new features and providing an easy integration with other information systems as well as among modules from different vendors. Success in the latter issue is claimed to be

crucial for maintaining the now outstanding ERP position in the new enterprise.

REFERENCES

Al-Mashari, M., Al-Mudimigh, A., & Zairi, M. (2003). Enterprise resource planning: A taxonomy of critical factors. *European Journal of Operational Research, 146*, 352-364.

Aloini, D., Dulmin, R., & Mininno, V. (2007). Risk management in ERP project introduction: Review of the literature. *Information & Management, 44*, 547-567.

Amoako-Gyampah, K. (2007). Perceived usefulness, user involvement and behavioural intention: An empirical study of ERP implementation. *Computers in Human Behavior, 23*, 1232-1248.

APICS (2000). *APICS survey on ERP 2000.* Retrieved June 15, 2008, from www.apics.com

Appleton, E. L. (1997). How to survive ERP. *Datamation*, 50-53.

Bancroft, N., Seip, H., & Sprengel, A. (1998). *Implementing SAP R/3*. Greenwich, CT: Manning,

Basoglu, N., Daim, T., & Kerimoglu, O. (2007). Organizational adoption of enterprise resource planning systems: A conceptual framework. *Journal of High Technology Management Research, 18*, 73-97.

Bingi, P., Sharma, M., & Godla, J. K. (1999). Critical issues affecting and ERP implementation. *Information Systems Management, 16*(3), 7-14.

Buckhout, S., Frey, E., & Nemec, J. (1999). Making ERP succeed: Turning fear into promises. *Strategy and Business, 15*, 60-72.

Chan, R. & Rosemann, M. (2001). Managing knowledge in enterprise systems. In *Proceedings of the 5th Pacific Asia Conference on Information Systems* (pp. 916-932), Seoul.

Cunningham, M. (1999). It's about the business. *Information, 13*(3), 83.

Davenport, T. (1998). Putting the enterprise into the enterprise system. *Harvard Business Review, July-August*, 121-131.

Delic, K. A. and Dayal, U. (2002). The rise of the intelligent enterprise. *Ubiquity – ACM IT Magazine & Forum, 3*(45).

Esteves, J. & Pastor, J. (2000). Towards the unification of critical success factors for ERP implementations. In *Proceedings of the 10th Annual BIT Conference*, Manchester.

Fan, M., Stallaert, J., & Whinston, A. B. (2000). The adoption and design methodologies of component-based enterprise systems. *European Journal of Information Systems, 9*, 25-35.

Finney, S. & Corbett, M. (2007). ERP implementation: A compilation and analysis of critical success factors. *Business Process Management, 13*(3), 329-347.

Genoulaz, V. & Millet, P., (2006). An investigation into the use of ERP systems in the service sector. *International Journal of Production Economics, 99*, 202-221.

Griffith, T. L., Zammuto, R. F., & Aiman-Smith, L. (1999). Why new technologies fail? *Industrial Management, 41*, 29-34.

Heizer, J. & Render, B. (2003). *Operations management-International edition* (7th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Jacobs, F. R. & Weston, F.C., Jr (2007). Enterprise resource planning (ERP)- A brief history. *Journal of Operations Management, 25*, 357-363.

Holland, C. & Light, B. (1999). A critical success factor model for enterprise resource planning implementation. *IEEE Software, 16*(3), 30-35.

Kadayam, S. (2002). The new business intelligence. *KM-World, (January)*, S6-S7.

King, S. F. & Burgess, T. F. (2006). Beyond critical success factors: A dynamic model of enterprise system innovation. *International Journal of Information Management, 26*, 59-69.

Kraemmeraard, P., Moeller, C., & Boer, H. (2003). ERP implementation: An integrated process of radical change and continuous learning. *Production Planning & Control, 14*(4), 338-348.

Kwon, O. B. & Lee, J. J. (2001). A multi-agent system for efficient ERP maintenance. *Expert Systems with Applications, 21*, 191-202.

Law, C. C. H. & Ngai, E. W. T. (2007). ERP systems adoption: An exploratory study of the organizational factors and impacts of ERP success. *Information & Management, 44*,418-432.

Markus, M. & Tanis, C. (2000). The enterprise systems experience: From adoption to success. In R.W. Zmud (Ed.), *Framing the domains of IT research glimpsing the future through the past*. Cincinnati, OH: Pinnaflex Educational Resources.

Motwani, J., Mirchandani, D., Madan, M., & Gunasekran, A. (2002). Succesful implementation of ERP projects: Evidence

from two case studies. *International Journal of Production Economics, 75*, 83-96.

Olin, J. G., Greis, N. P., & Kasarda, J. D. (1999). Knowledge management across multi-tier enterprises: The promises of intelligent software in the auto industry. *European Management Journal, 17*(4), 335-347.

Ptak, C. & Schragenheim, E. (2000). *ERP: Tools, techniques and applications for integrating the supply chain*. Boca Raton, FL: St. Lucie Press.

Remus, U. (2007). Critical success factors for implementing enterprise portals. A comparison with ERP implementations. *Business Process Management, 13*(4), 538-552.

Scott, F. & Shepherd, J. *The steady stream of ERP investments*. AMR Research.

Seidel, B. & Stedman, C. (1998). Siemens cuts PeopleSoft loose for SAP. *Computerworld (Online)*, October 5.

Serrano, N. & Sarriegi, J. M. (2006). Open source software ERPs: A new alternative for an old need. *IEEE Computer Society*, 94-97.

Slater, D. (1998). The hidden costs of enterprise software. *CIO Magazine, 12*, 30-37.

Smirnov, A.V., Pashkin, M., Chilov, N., & Levashova, T. (2003). Agent-based support of mass customization for corporate knowledge management. *Engineering Applications of Artificial Intelligence, 16*, 349-364.

Somers, T. M. & Nelson, K. G. (2004). A taxonomy of players and activities across the ERP project cycle. *Information & Management, 41*, 257-278.

Sprott, D. (2000). Componentizing the enterprise application packages. *Communications of the ACM, 43*(4), 63-69.

Srivardhana, T. & Pawlowski, S. D. (2007). ERP systems as an enabler of sustained business process innovation: A knowledge-based view. *Journal of Strategic Information Systems, 16*, 51-69.

Stein, T. (1997). Boeing to drop Baan's software. *Information Week, August 25*.

Tarantilis, C. D., Kiranoudis, C. T., & Theodorakopoulos, N. D. (2006). A web-based ERP system for business services and supply chain management: Application to real-world process scheduling. *European Journal of Operational Research*.

Umble, E. J., Haft, R. R., & Umble, M. M. (2003). Enterprise resource planning: Implementation procedures and critical success factors. *European Journal of Operational Research, 146*, 241-257.

Wheatley, M. (2000). ERP training stinks. *CIO Magazine*, June.

Wiig, K. M. (1999*). The intelligent enterprise and knowledge mangament*. Knowledge Research Institute Working Paper.

Yen, D. C., Chou, D. C., & Chang, J. (2002). A synergic analysis for web-based enterprise resource planning systems. *Computer Standards and Interfaces, 24*(4), 337-346.

KEY TERMS

Best Practices: Process procedures of recognized excellence, usually obtained from companies' experience and or process optimization analysis.

Big-Bang ERP Implementation: ERP implementation strategy consisting in implementing all required modules and features at once.

BPR: Business Process Reengineering (BPR) radically rethinks key enterprise process in order to achieve substantial process improvement.

Client/Server Architecture: Computer network model separating computers providing services (servers) from computers using these services (clients).

EKM: Enterprise Knowledge Management (EKM) is aimed to inject knowledge into business processes and to enable reuse of human expertise through the creation of common data objects and definitions that can be used with equal ease and success by all employees in the enterprise.

ERP: Enterprise Resource Planning (ERP) denotes packaged software to support corporate functions such as finance, human resources, material management, or sales and distribution.

Intelligent Enterprise: Organization capable to act effectively in the present and to deal effectively with the challenges of the future by being proactive, adaptable, knowledgeable, and well-resourced.

Modularity: Most ERP packages decompose their functionality in modules grouping typical business functions such as finance, sales, manufacturing, among others.

Process Orientation: Recognition of series of functions that carry out an overriding task by providing the customer with a meaningful result.

Category: Business Information Systems

An Overview of Executive Information Systems (EIS) Research in South Africa

Udo Richard Averweg
eThekwini Municipality and University of KwaZulu-Natal, South Africa

INTRODUCTION

Executive information systems (EIS) are designed to serve the needs of executive users in strategic planning and decision-making. Sometimes the terms "executive information systems" and "executive support systems" are used interchangeably (Turban, McLean, & Wetherber, 1999). Definitions of EIS are varied but all identify the need for information that support decisions about the organization. EIS can be defined as "a computerized system that provides executives with easy access to internal and external information that is relevant to their critical success factors" (Watson, Houdeshel, & Rainer, 1997).

This article is organized as follows: The background to EIS implementation is given. EIS research studies undertaken in South Africa are then described. Some future EIS trends are then suggested.

BACKGROUND TO EIS IMPLEMENTATION

A number of possible indicators for a successful information system (IS) have been suggested in various implementation studies (see, for example, Laudon & Laudon,1998). The definition of implementation includes the concept of success or failure. Implementation is a vital step in ensuring the success of new systems.

The EIS implementation process is defined as the process used to construct an EIS in an effective manner (Srivihok, 1998). Different factors have been suggested by various researchers as influencing successful EIS implementation (see, for example, Rainer & Watson, 1995). However, there is no agreement on which factors play key roles in EIS implementation. A large number of success factors have been repeatedly suggested by practitioners and researchers, even though empirical studies on the success factors are rare.

EIS are high-risk application systems that are expensive to build and maintain (Strydom, 1994). For example, in October, 1997, the largest water utility in South Africa, Rand Water, took a decision to build an EIS (based on Oracle® products) and invested ZAR4.5m in revamping its IT infrastructure to support that deployment. In the case of Rand Water, the organization's EIS eventually played a major role in providing its executives with benchmarking information helping them track Rand Water's overall performance against a set of objective criteria.

EIS RESEARCH UNDERTAKEN IN SOUTH AFRICA

A review of previously conducted EIS research at universities in South Africa is undertaken. From this collection, the nature of EIS research for each study is discussed. South African databases were searched for research literature (essays, technical reports, thesis, dissertations, etc) with the keywords "Executive Information Systems" in the title. Nine successful "hits" were found. Those research articles are reflected in chronological publication sequence in Table 1.

The nature of each of the nine EIS studies in South Africa is now briefly discussed:

- **Researcher No 1: Design and Implementation of Executive Information System (EISs):** DeWitt (1992) discusses critical success factors (CSFs) for EIS development and states that the type of EIS for an organization will depend on the information requirements of the organization. It should be driven by the CSFs that are unique to a particular business. From previous studies, DeWitt (1992) identifies nine CSFs for an EIS (see Table 2) and notes that there "are differences of opinion in the literature regarding the selection of the right technology" as a CSF.
  This study was undertaken with sixteen large Cape Town companies from various industry sectors. The findings from Watson's international survey (Watson, Rainer, & Koh, 1991) were compared against the local (South Africa) survey findings. The findings indicate (1) congruences between the literature search and survey

Table 1. Research literature (essays, thesis or dissertations) with the keywords "executive information systems"

No	Researcher(s)	Publication Date	Research title	Report Type	Qualification and Institution
1	DeWitt, P.	May 1992	Design and Implementation of Executive Information Systems (EISs)	Technical Report	B Com (Honours): University of Cape Town
2	Twemlow, S. Hoffmann, U. and Erlank, S.	October 1992	An Assessment of the Penetration of Executive Information Systems in South Africa	Technical Report	B Com (Honours): University of Cape Town
3	Strydom, I.	April 1994	Executive Information Systems: A Fundamental Approach	Thesis	Doctor Commercii (Informatics): University of Pretoria
4	Steer, I.J.	January 1995	The Critical Success Factors for the Implementation of Executive Information Systems in the South African environment	Dissertation	M Com: University of Witwatersrand
5	Faure, S.	June 1995	The Impact of Executive Information Systems on the User	Essay	B Com (Honours): University of Cape Town
6	Chilwane, L.	November 1995	Critical Success Factors for the Management of Executive Information Systems in Manufacturing	Research report	M Com: University of Witwatersrand
7	Khan, S.J.	February 1996	The Benefits and Capabilities of Executive Information Systems	Research report	MBA: University of Witwatersrand
8	Baillache, S.	April 1997	The Experiences Gained by Users of Executive Information Systems	Dissertation	MBA: University of Witwatersrand
9	Averweg, U. R. F.	December 2002	Executive Information Systems Usage: The Impact of Web-based Technologies	Dissertation	M Science: University of Natal

findings; (2) major conflicting results between the local survey, the international survey and literature search; and (3) major problems encountered in developing EIS.

- **Researchers No 2: An Assessment of the Penetration of Executive Information Systems:** Twemlow et al. (1992) carried out an exploratory study that showed the extent of EIS penetration in South Africa. The sample (61 companies) was selected from the 1992 Financial Mail survey (a reputable weekly financial publication) of "top" companies in South Africa. The research instrument was designed to evaluate EIS as a significant business trend, the extent of penetration of this trend in the organization and perceived impact on the business. From these researchers' findings, the problems experienced by companies during the implementation and use of their EIS is reflected in Table 3.
  Twemlow et al. (1992) suggest that even though studies have been performed to determine the nature of executive work and their information requirements, there is still uncertainty in this area. Twemlow et al. (1992) note that "it is not surprising" that the first two out of the top four problems associated with EIS implementation were concerned with the complex and changing executive information needs.
- **Researcher No 3: Executive Information Systems: A Fundamental Approach:** Strydom's (1994) research investigated the problems concerning EIS "from a fundamental research perspective." Based on the results of the research an augmented EIS was proposed and referred to as a computer supported executive system (CSES). Strydom (1994) discussed the role of training in successful implementation of IS and focuses on computer supported learning for EIS.
- **Researcher No 4: Critical Success Factors for Executive Information Systems Implementation:** Steer's (1995) study used the findings of research undertaken by Harris (1993) and others. The basis of Steer's research "was to identify the critical success factors for the successful implementation of an Executive Information System ..... where an EIS had been implemented." Seventeen well-established organizations in Gauteng (a province in South Africa) that have EIS experience were targeted and surveyed. The analysis of Steer's findings indicate 21 major concepts that were raised by interviewed respondents in relation to the CSFs for implementing EIS. The top 10 CSFs (in descending order) that were identified in this study for the successful implementation of EIS are reflected in Table 4.
  Steer indicates that although "the remaining 11 concepts of the 21 discussed during the research are not the most important critical success factors of implementing an EIS, they are still important, and should therefore be considered when implementing an EIS." Steer (1995) labels these CSFs as "secondary" CSFs for the successful implementation of EIS. These secondary CSFs are reflected in Table 5.

- **Researcher No 5: The Impact of Executive Information System on the User:** The focus of Faure's (1995) research was to highlight "the key features of an EIS, the benefits that can be achieved from implementing an EIS and the development methodologies that can be adopted to achieve success in the implementation of an EIS."
- **Researcher No 6: Critical Success Factors for the Management of Executive Information Systems in Manufacturing:** The aim of Chilwane's (1995) research was "to identify those critical issues, which when managed properly, will ensure that the system remains providing and meeting the needs of the executives. Ten interviews were conducted from business organizations in order (*sic*) identify these factors." Table 6 reflects the CSFs for managing an operating EIS "as seen by respondents who organizations have implemented EIS" (Chilwane, 1995). Chilwane (1995) states that ensuring "that these factors are monitored will contribute to sustaining the investment an organization has made in this technology."
- **Researcher No 7: The Benefits and Capabilities of Executive Information Systems:** The objective of Khan's (1996) research was to identify and evaluate the organizational benefits derived from EIS and to establish which of its capabilities contribute to the realization of the benefits. Khan (1996) notes that a major problem when implementing an EIS is determining the information requirements for the system (Watson & Frolick, 1993). For these researchers a major developmental problem is determining the information to include in the system.

  Khan (1996) notes that practitioners find it difficult to get executives to specify what they want and to keep abreast of executives' changing information desires and needs. Khan's (1996) findings identify six major benefits of EIS and five major capabilities of EIS.
- **Researcher No 8: Experiences Gained from Executive Information Systems:** Baillache's (1997) research investigated the experiences gained by South African users of EIS. The results are seen as important in identifying problem areas that negatively affected the evolution of EIS in South Africa. Some 30 companies participated in this survey. Four users from each company surveyed were requested to complete a questionnaire.

  The findings indicate that (1) some important capabilities had been omitted in systems; (2) users expectations of benefits were far greater than benefits delivered; (3) key CSFs did not occur during the implementation of the project; and (4) the growth of the system by new users was not strongly correlated to the CSFs. Baillache's (1997) summary of results of CSFs implemented is reflected in Table 7.

Table 2. DeWitt's (1992) nine CSFs for an EIS

A committed and informed executive sponsor
An EIS driver
A clear link to business objectives
Carefully defined system requirements
Ensure feasibility of data availability
An active team approach to ensure spread to additional users
An evolutionary development approach
Quick response and user friendliness
Managing organisational resistance

Table 3. Problems with Implementation and Use of EIS (Source: Adapted from Twemlow et al., 1992)

Complex information needs of EIS users
Changing needs of EIS users
Insufficient management support
Lack of clarity of EIS purpose
Data availability
Failure to incorporate EIS into management processes
Hardware compatibility
Software compatibility
Unexpected increase in costs
Failure to meet the user's expectations

Table 4 . The top ten CSFs for the successful implementation of EIS (Source: Adapted from Steer, 1995)

Concept
An EIS needs a project champion
An EIS must support the cross-functional integration of information
An EIS has to link to the organisation's business strategy
An EIS should be implemented using a phased approach
An EIS project champion should be a steering committee rather than one person
Resistance from the information users must be managed
An EIS must have the capability to access external information
Resistance from the information providers must be managed
The project champion should change during the project
An EIS must support "drill down" facilities

Table 5. Secondary CSFs for the successful implementation of EIS (Source: Adapted from Steer, 1995)

Concept
An EIS should be made available to everyone
An EIS must have "what if" and simulation facilities
Resistance from IT people must be managed
An EIS must support trend analysis
The user must be able to interact with and manipulate the information
An EIS must support exception reporting
It must be possible to track actuals against plans
An organisation must develop a formalized business strategy before it embarks on an EIS project
An EIS must have a good graphical user interface
An EIS should be for executives only
An EIS must be able to access financial information

Table 6. CSFs for the management of an operating EIS (Source: Adapted from Chilwane. 1995)

Executives or users should provide regular feedback on the EIS either formally or informally
Continued alignment of EIS ensures that the system remains useful to the users
Continued executive involvement ensures success of the system
An EIS should be flexible to accommodate the dynamic business environment
As EIS spreads new requirements should be reflected in the system
ISD should provide somebody who knows the business to look after the users
An EIS should help individual managers to monitor their individual CSFs
EIS data should always be consistent with the operational data it summarizes
There should be prompt attention to user queries and requirements
An EIS should be portable that is loaded on a notebook and accessed offline

- **Researcher No 9: Executive Information Systems Usage: The Impact of Web-based Technologies:** The objective of Averweg's (2002) study was *inter alia* to identify and rank Web-based technologies in order of their perceived future impact on EIS.
  Only 6.4 percent of organizations surveyed reported that it is unlikely that the intranet will impact future EIS implementations. Almost half of organizations surveyed reported that it is unlikely that e-commerce (business-to-consumer) will impact future EIS implementations. WAP and other mobile technologies have similar unlikely future impact levels. It is striking to note that 67.7 percent of respondents indicated that it is *extremely unlikely* that other technologies (such as portal) will impact future EIS implementations. There was a positive impact level trend for all Web-based technologies on future EIS implementations. The largest trend increase was the Intranet rising from 32.2 percent to 87.1 percent. Averweg (2002) suggests that this "should occur as the use of Web-based technologies in the distribution of information becomes more widespread."

DISCUSSION OF PREVIOUS EIS RESEARCH UNDERTAKEN IN SOUTH AFRICA

From Table 1, four EIS researchers (Nos 1, 4, 6 and 8) dealt with CSFs for EIS implementation. A synopsis of the results and findings indicates that there is no consistent "shopping basket" of CSFs for EIS implementation for use by South African practitioners.

Like other systems, EIS are constantly changing. Khan (1996) suggests an investigation into new technologies being employed in the IT area and "to what extent advances in technology have influenced ... EIS." Khan's (1996) EIS research "identified the employment of new technologies as the most important future trend of EIS." The EIS research undertaken by Averweg (2002) serves to fill that gap.

The Web serves as the foundation for new kinds of IS (Laudon & Laudon, 1998). As the Web grows in direct usage by executives, existing EIS implementation models may need to be revisited. While there is no single listing of key variables for EIS success factors (Rainer & Watson, 1995), strong human factors are nevertheless associated with EIS research. These are influenced by cultural, political and other "soft" human factors. It is therefore neither possible nor valid to generalize experiences on other continents to South Africa's conditions. This makes relevant local studies (in South Africa) of EIS implementation and usage.

FUTURE EIS TRENDS

Information Technology (IT) is more than just computer systems and it is rapidly changing and developing, especially due to the Web, altering the way in which an IS is built. With the Internet, information is no longer a scarce resource. It has changed the way in which organizations are doing business and the way in which they compete.

Table 7. Summary of results of CSFs implemented

CSFs supported by the research	CSFs Implemented
	Having an executive sponsor on the project
	System reliability was ensured
	Quality data
	The skills of the system designers
CSFs not supported by the research	Having an operating sponsor on the project
	Local representation of software companies for support
New CSFs which emerged from the research	There was a clear link between the EIS and business objectives
	Appropriate resources were used from the information systems function
	Appropriate technology was used
	Users specified their own information requirements
	The first deliverable by the ITD was information that was highly valuable
	The EIS contained information of much value to me
	The EIS was implemented as quickly as possible
	The hardware used was reliable
	Pilot sites were used in the implementation

The environment for EIS is undergoing upheaval based on the emergence of Web-based technologies. The following trends for EIS implementation are envisaged:

- Data warehouses store data that have been extracted from the various operational databases of an organization over some years. The intrinsic design of the Web resembles that of a data warehouse bringing access to data collected and provided by a host of users outside a specific organization. The immediacy of the Web is *not* seen as a factor for improved decision-making since EIS are rarely used by executives in emergency or critical time-modes;
- With the increasing amount of IT investment and substantial evidence of failures (Remenyi & Lubbe, 1998), many managers and researchers feel that IS justification and evaluation has become a key management issue. It is contended that wise judgement is needed when deciding on the selective use of IT and feel that this is particularly relevant to EIS in South Africa;
- There will be a significant degree of EIS diffusion to lower organizational hierarchical levels and use by these lower levels. EIS in organizations will spread to managers at various levels such as functional areas and other levels of management (Singh et al., 2002). This will be in keeping with international trends where EIS are being diffused in organizations as EIS is becoming less strictly defined to support professional decision-makers throughout the organization;
- Web-based technologies have enabled EIS to become available to more management levels in the organization. The Web browser has become the common interface to end-user access. While applications can now be accessed by browsers, the capabilities long associated with decision support software are still found (Averweg & Erwin, 2000). Nowadays vendors of decision support software are making their products Web-enabled;
- Xu et al. (2003) suggest that the internal information orientation is the main reason for dissatisfaction with EIS. In order to overcome this dissatisfaction, it is felt that greater use will be made from data from *external* sources (for an organization's CSFs); and
- Special care will be needed when implementing EIS because of its major potential importance to an organization's performance. Failure can lead to long delays in further attempts to use such technology effectively.

This information is particularly useful for IT practitioners in the planning of future EIS implementations in South Africa. An understanding of Web-based technology taxonomies is important to EIS researchers and practitioners.

CONCLUSION

Organizations must "start simple, grow fast" (McKenna Group, 1999) using technologies that will enable it to build on what it has, link to legacy systems, rather than throwing away what has been achieved and developed through each new enhancement iteration. EIS will be impacted by these

change catalysts as EIS become integrated with Web-based technologies not specifically designed for EIS usage.

EIS is going through a major change to take advantage of Web-based technologies in order to satisfy information needs of an increasing group of users (Averweg & Roldán, 2006). Web technologies are often not just a single technical solution, rather a host of an industry specific with inter-connective capabilities that pull together people, processes and technology infrastructure. EIS is being catalyzed through a major change as technical barriers disappear.

REFERENCES

Averweg, U.R.F., & Erwin, G.J. (2000). Executive Information Systems in South Africa: A Research Synthesis for the Future. *Proceedings of the South African Institute of Computer Scientists and Information Technologists Conference (SAICSIT-2000), Cape Town, South Africa*, 1-3 November.

Averweg, U.R. (2002). *Executive Information Systems Usage: The Impact of Web-based Technologies.* Master of Science dissertation, Faculty of Science & Agriculture, University of Natal, Pietermaritzburg, South Africa.

Averweg, U.R., & Roldán, J.L. (2006). Executive information system implementation in organisations in South Africa and Spain: A comparative analysis. *Computer Standards & Interfaces, 28*(6), 625-634.

Baillache, S.C. (1997). *The Experiences Gained by Users of Executive Information Systems.* MBA dissertation. University of Witwatersrand, Johannesburg, South Africa.

Chilwane, L. (1995). *The Critical Success Factors for the Management of Executive Information Systems in Manufacturing.* M Com dissertation. University of Witwatersrand, Johannesburg, South Africa.

DeWitt, P. (1992). *Design and Implementation of Executive Information Systems (EISs). B Com (Honours).* University of Cape Town, Cape Town, South Africa.

Faure, S. (1995). T*he Impact of Executive Information Systems on the User. B Com* (Honours). University of Cape Town, Cape Town, South Africa.

Harris, J. (1993). Is your EIS too stupid to be useful? *Chief Information Officer Journal.*

Khan, S.J. (1996). *The Benefits and Capabilities of Executive Information Systems.* MBA dissertation, University of Witwatersrand, Johannesburg, South Africa.

Laudon, K.C., & Laudon, J.P. (1998). Management Information Systems. New Jersey: Prentice-Hall, Inc.

McKenna Group. (1999). *Becoming an e-business.* White Paper, 30 March.

Rainer, R.K., & Watson, H.J. (1995). The keys to executive information system success. *Journal of Management Information Systems, 12*(2), 83-98.

Remenyi, D., & Lubbe, S. (1998). Some Information Systems issues in South Africa and Suggestions as to how to deal with them. Cited in S. Lubbe (Ed.), *IT investment in developing countries: An assessment and practical guideline*. Hershey: Idea Group Publishing.

Singh, S.K., Watson, H.J., & Watson, R.T. (2002). EIS support for strategic management process. *Decision Support Systems, 33*, 71-85.

Srivihok, A. (1998). *Effective Management of Executive Information Systems Implementations: A Framework and a model of successful EIS implementation.* PhD dissertation. Central University, Rockhampton, Australia.

Steer, I.J. (1995). *The Critical Success Factors for the Successful Implementation of Executive Information Systems in the South African Environment.* M Com dissertation, University of Witwatersrand, Johannesburg, South Africa.

Strydom, I. (1994). *Executive Information Systems: A Fundamental Approach.* Doctor Commercii (Informatics), University of Pretoria, Pretoria, South Africa.

Turban, E., McLean, E., & Wetherbe, J. (1999). *Information technology for management.* New York: John Wiley & Sons, Inc.

Twemlow, S., Hoffmann, U., & Erlank, S. (1992). *An Assessment of the Penetration of Executive Information Systems in South Africa*. B Com (Honours). University of Cape Town, Cape Town, South Africa.

Watson, H.J., & Frolick, M.N. (1993). Determining information requirements for an executive information system. *MIS Quarterly, 17*(3), 255-269.

Watson, H.J., Houdeshel, G., & Rainer, R.K. Jr. (1997). *Building executive information systems and other decision support applications*. New York: John Wiley & Sons, Inc.

Watson, H.J., Rainer, R.K., & Koh, C.E. (1991). Executive information systems: A framework for development and a survey of current practices. *MIS Quarterly, 15*(1), 13-30.

Xu, X.M., Lehaney, B., Clarke, S., & Duan, Y. (2003). Some UK and USA comparisons of executive information systems in practice and theory. *Journal of End User Computing, 15*(1), 1-19.

KEY TERMS

Critical Success Factors (CSF): Those key areas of activity in which favorable results are *absolutely necessary* for a particular manager to reach his or her goals.

Data Cube: In a multidimensional database, data can be viewed and analyzed from different views or perspectives, known as business decisions. These dimensions form a cube.

Data Warehouse: A repository of subject-oriented historical data that is organized to be accessible in a form readily acceptable for analytical processing activities.

Executives: Corporate knowledge workers responsible for corporate strategic management activities.

Executive Information System: A computerized system that provides executives with easy access to internal and external information that is relevant to their critical success factors.

Information Systems (IS): A combination of technology, people and processes to capture, transmit, store, retrieve, manipulate and display information.

Web-Based Technology: A technology that did not exist prior to the World Wide Web ("the Web") and utilizes core Internet and Web technologies as the platform on which the solution operates.

World Wide Web ("the Web"): An information space consisting of hyperlinked documents published on the Internet.

An Overview of Knowledge Translation

Chris Groeneboer
Learning and Instructional Development Centre, Canada

Monika Whitney
Learning and Instructional Development Centre, Canada

INTRODUCTION

Knowledge translation (KT) was traditionally framed as a problem of moving research results into policy and practice. The impetus for the flow of knowledge originated with researchers constructing new knowledge and seeing its utility, or with policymakers and administrators seeing problems in practice and looking to researchers for solutions.

In the 1970s, a shift in focus away from knowledge use was exemplified by Caplan's (1979) two-communities theory, which posits that researchers and policymakers comprise two different communities with two different languages (Jacobson, Butterill, & Goering, 2003). A shift back to knowledge use with a new focus on user-centered design is evident in more recent KT models that provide frameworks for researcher and user interaction in order to build better understanding between diverse groups.

The flow of knowledge from its construction in one context to its use in another context has been variously termed knowledge translation, knowledge exchange, knowledge transfer, research transfer, technology transfer, knowledge transformation, knowledge dissemination, knowledge mobilization, knowledge utilization, and research utilization. The terms are often used synonymously, but a specific term is sometimes used because it highlights a particular component of the knowledge flow process. For example, *knowledge exchange* implies a sharing of information between partners of equal value and focuses on the movement of knowledge between them, whereas *research utilization* implies the transformation of research results into usable knowledge and focuses on embedding the usable knowledge in practice.

Information technologies have the potential to support knowledge translation in powerful ways. Key processes in the translation of knowledge include: (1) knowledge creation, management, and dissemination; (2) recognition of links between existing knowledge and its potential application to problems or practice; (3) translation into usable knowledge in practice; and (4) change in practice.

Information technologies are a natural solution for these knowledge translation processes. For example, group and social software such as blogs and wikis support collaborative construction and sharing of knowledge; knowledge management systems support capture, storage, accessibility, and maintenance of constructed knowledge; and most Internet-based technologies support dissemination of information. Well-designed virtual communities provide online environments for the kinds of human interaction that enable collaborative exploration of ideas, that foster recognition of potential links between existing knowledge and its application to solve problems or change practice, and that inspire people to transform their practice. Data mining and artificial intelligence techniques can be used to enhance identification of potential links between knowledge in one context and problems in another context.

BACKGROUND

A variety of approaches to knowledge translation have been developed, most focusing on the interaction of researchers, practitioners, and policymakers to move research results into practice. KT is not inherently unidirectional (research to practice), and Lavis et al. (2001) have argued that researcher-user interaction should become standard practice in research contexts, not simply an add-on. This practice has the potential to open new communication channels from knowledge constructed in practice to new research questions and hypotheses. The five models described below demonstrate a variety of KT approaches in use, from a national initiative to a framework for individual researchers.

The Canadian Institutes of Health Research

The Canadian Institutes of Health Research (CIHR) were created in June 2000 by the government of Canada with a mandate that included health research and knowledge translation defined as:

the exchange, synthesis and ethically-sound application of knowledge—within a complex system of interactions among researchers and users—to accelerate the capture of the benefits of research for Canadians through improved health, more effective services and products, and a strengthened health care system. (http://www.cihr-irsc.gc.ca/e/29418.html)

This definition acknowledges the importance of interaction between researchers and users in order to develop a sound

translation. CIHR (2004) also recognizes that "knowledge translation strategies and activities vary according to the type of research to be translated…and the intended user audience…."

The Knowledge Translation Strategic Plan 2004–2009 (CIHR, 2004) identifies four strategic directions to promote knowledge translation at a national level:

1. Support KT research, i.e., research on KT concepts and processes;
2. contribute to building KT networks, i.e., networks of researchers and research users;
3. strengthen and expand KT at CIHR, i.e., improve capability to support KT research and, with partners, KT itself; and
4. support and recognize KT excellence, i.e., build and celebrate a culture of KT.

The Ottawa Model of Research Use

Logan and Graham (1998) developed the Ottawa Model of Research Use (OMRU), a holistic, interactive approach to knowledge translation intended for use by policymakers to increase utilization of health research results and by researchers interested in the integration of research results into practice. The six key elements include the practice environment, potential adopters of the evidence, evidence-based innovation, research transfer strategies, the use of the evidence, and health-related and other outcomes.

These elements are continuously evaluated in order:

(1) to identify potential barriers and supports to research use related to the practice environment, potential adopters, and the evidence-based innovation; (2) to provide direction for selecting and tailoring transfer strategies to overcome the identified barriers and enhance the supports; (3) to track the progress of the transfer effort; and (4) to evaluate the actual use of the evidence-based innovation and its impact on outcomes of interest. (Logan & Graham, 1998, p. 230)

Research Implementation Approach

Grol and Jones (2000), the National Cancer Institute (2002), and Caburnay, Kreuter, and Donlin (2001) have developed research implementation approaches. Grol and Jones (2000) describe an iterative process of research implementation and evaluation consisting of "Research evidence → Develop concrete proposal for change → Analysis of target social and organizational context, obstacles to change → Link interventions to obstacles → Develop plan → Carry out plan and evaluate progress" (p. S33).

Based on KT implementation research results, four factors that influence the uptake and continued use of clinical guidelines were identified (Grol & Jones, 2000). These factors include:

1. Features of the guidelines (such as the underlying research evidence and the language of the guidelines),
2. features of the target group,
3. features of the social context/setting, and
4. features of the organizational context.

Lavis, et al.'s Framework

Lavis, Roberston, Woodside, McLeod, and Abelson (2003, p. 222) developed a framework for knowledge transfer based on five questions:

1. What should be transferred to decision makers (the message)?
2. To whom should research knowledge be transferred (the target audience)?
3. By whom should research knowledge be transferred (the messenger)?
4. How should research knowledge be transferred (the knowledge-transfer processes and supporting communications infrastructure)?
5. With what effect should research knowledge be transferred (evaluation)

The framework was derived from a review of the research literature across the five questions, four target audiences (general public/service recipients, service providers, managerial decision makers, and policy decision makers at federal, state/provincial, and local levels) and a range of disciplinary perspectives and methodological approaches (Lavis et al., 2003). For example, with regard to question 1 (What should be transferred to decision makers?), they concluded that action should be taken to transfer knowledge based on a body of research results as opposed to a single published paper to assure validity.

Jacobson, Butterill, and Goering's Framework

Jacobson et al. (2003) developed a framework for knowledge translation focused on building understanding between researchers and user groups. The framework was derived from a review of the literature and the authors' experiences. Articles related to user groups and the knowledge translation process were coded into conceptual categories that emerged from the data. The synthesis of this analysis resulted in a framework containing five domains: user group, issue, research, researcher-user relationship, and dissemination strategies. Each domain consists of a series of questions to guide researchers toward increased understanding of the

intended user context. The framework also provides the research results from which the questions were derived.

KNOWLEDGE TRANSLATION CHARACTERISTICS

KT involves the interaction of elements in a complex knowledge ecosystem of people, contexts, bodies of knowledge, ideologies, methodologies, organizational structures, cultures, and languages. Five key components underlie each of the KT models described above:

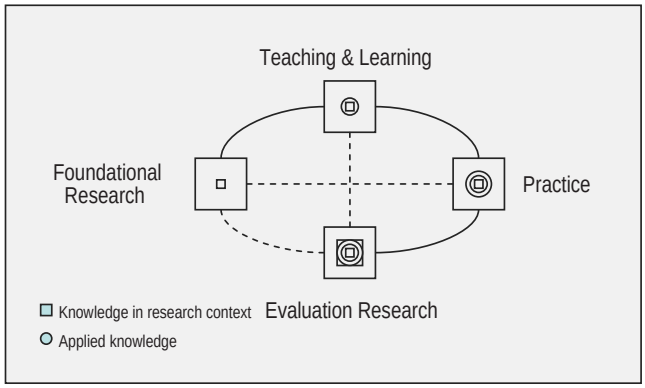
1. Knowledge construction;
2. knowledge contexts;
3. knowledge movement, or flow;
4. knowledge adaptation, or transformation; and
5. evaluation of the translation.

Knowledge construction refers to the building of knowledge using paradigms from a community of practice within a particular context. In order to interpret its meaning for a target context, a translator needs to understand how the knowledge was constructed in the source context. The translator also needs to understand how the problem was framed and constructed in the target context.

Four broad categories of knowledge contexts include:

- Foundational research typically conducted in universities and industry labs,
- education and training typically conducted in universities and other organizations,
- practice in the real world including policymaking, and
- evaluation research assessing the translation.

Figure 1. Knowledge contexts and flow (Groeneboer, 2006)



Knowledge constructed in one context may be applied in another context as shown in Figure 1. The solid links demonstrate typical knowledge flow from research to education to practice, whereas the dashed links indicate connections that need to be strengthened. One difficulty in creating channels between contexts is that they are often ideologically and geographically divided.

To summarize knowledge flow, suppose that knowledge K_R was constructed in research context R, and that problem Q_P was identified in real-world context P. First, someone must know about K_R and Q_P, and recognize that K_R is a potential solution for Q_P. Once that connection is made, knowledge K_R needs to be translated into a localized form of knowledge K_{RP} that can be utilized by the intended user group in their context P. The translation is necessary but not sufficient in that the user group needs to adopt practical knowledge K_{RP} and change their individual practice. Finally, the translation

Table 1. General questions to orient knowledge flow activities

#	Knowledge Flow Activity	Questions
1.	Construct source knowledge K_A in context A	How was the knowledge K_A constructed in context A? For example, is it generalizable? Does K_A describe a correlational or causal relationship?
2.	Construct target problem Q_B in context B	How was the problem Q_B framed in context B? What does it mean in context B?
3.	Make connection between $K_A \leftrightarrow Q_B$	What does K_A mean? Does it have meaning in context B? Does it address Q_B effectively and affordably?
4.	Plan the translation $K_A \rightarrow K_{AB}$	What was the context in which K_A was constructed? What are the characteristics of context B and the user community that are relevant to the design of K_{AB}? How do we get from A to B?
5.	Translate $K_A \rightarrow K_{AB}$	Given contexts A and B, how should K_A be interpreted and localized for context B? What action in context B does K_A imply?
6.	Support adoption of K_{AB} in target context B	How can individuals and organizations be supported in changing their practice to adopt K_{AB}? What incentives or rewards are there? Are there existing policies that might dissuade adoption of the new practice?
7.	Evaluate the process	Does K_{AB} effectively solve Q_B? What worked and what did not work in designing and implementing the translation plan? What would you do differently next time?

process needs to be evaluated to determine whether the translation successfully solved problem Q_P.

General questions need to be addressed at each activity in the knowledge translation flow as shown in Table 1.

A major barrier to #3 (making connections between $K_A \leftrightarrow Q_B$) is that there are very few communication channels bridging contexts. For example, foundational researchers rarely hear the experiences of practitioners in the field, yet this type of interaction early in a cycle can lead to construction of knowledge more readily adapted to a new context. Examples of established environments supporting interaction are teaching hospitals where research, education, training, and practice contexts have been integrated. Cross-disciplinary channels are also important for KT because real-world problems experienced in practice do not usually fall squarely in one discipline, and results obtained in one discipline may have useful applications in other fields.

KNOWLEDGE TRANSLATION STRATEGIES

The goal of the knowledge translation process is to enhance the multi-directional, cross-disciplinary flow of knowledge between contexts so that knowledge constructed in one context can be utilized in others. Conceptually, a space is needed that fosters dialogue regarding research, education and training, policy and practice, and evaluation. This space is referred to as the *nexus*.

The nexus forms a means of connection between different contexts as shown in Figure 2. Here researchers, educators, practitioners, policymakers, and others 'meet' to share information and gain a better understanding of the cultures and contexts in which each operates. Virtual spaces provide a natural solution for a nexus of this type because:

1. Technologies are available to enable key processes of knowledge translation such as online environments for human interaction and tools for collaborative creation and sharing of knowledge,
2. virtual space overcomes geographical and time-based barriers, and
3. there are precedents of successful virtual communities from which lessons can be learned.

Well-designed virtual environments also offer flexibility so that participants shape the structure of the space to fit their needs over time. For example, subgroups may form around shared interests and subspaces may be developed for special activities and events.

KT requires a new kind of professional, what Cernada (1982) called *linkers*—people who can facilitate the translation of research results into practice. He argued that linkers are necessary but not sufficient for research utilization, and suggests researchers and practitioners themselves need to come together. Professional development of linkers is a key strategy for KT, as linkers need to understand both source and target contexts and the construction and use of knowledge in each context.

The nexus becomes a point of convergence for a *metacommunity* of researchers, educators, practitioners and policymakers, evaluators, and linkers. The nexus also serves as a point where hierarchical roles are put aside, and the ultimate goal of knowledge acquisition and application is of primary importance. Table 2 summarizes some of the strategies that could be used at the nexus and linker levels to overcome potential barriers and support the knowledge flow activities of Table 1.

Figure 2. The nexus of research, teaching and learning, practice, and evaluation (Groeneboer, 2006)

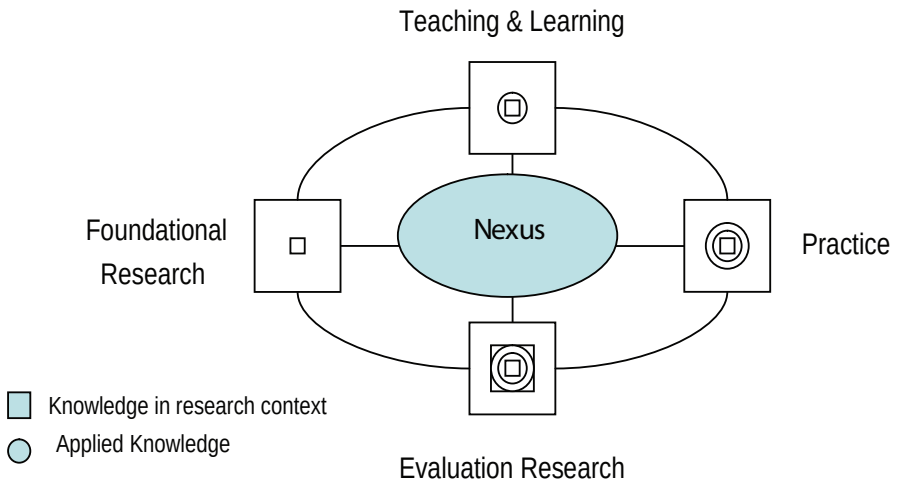


Table 2. Summary of knowledge translation strategies

Level	#	Strategy
Responsibilities/ Roles of the Nexus	1.	Create collections of knowledge constructed in different contexts for sharing (e.g., a nexus knowledge base)
	2.	Create virtual spaces and events to bridge people and knowledge from different communities
	3.	Professionally develop linkers who can translate and support the interaction of diverse communities in the nexus
Responsibilities/ Roles of the Linker	1.	Scan published results in target and related fields
	2.	Conduct meta-analysis of potential applicable research
	3.	Interpret research for use—what do the results really mean? — create a resource for general use in target contexts
	4.	Create a framework/template learning design for teaching the interpreted knowledge
	5.	Localize knowledge for a particular identified context
	6.	Create case studies of translated knowledge embedded in practice
	7.	Share evaluation results through the nexus
	8.	Support community building, events, awareness-raising activities, etc.
	9.	Scaffold researchers, educators, and practitioners in becoming linkers themselves
	10.	Create future scenarios to inspire others and articulate a vision of the ways in which good KT practice can help

FUTURE TRENDS

Future trends in KT include:

1. Given the current climate of user-centered design and interaction, more systems-theoretic approaches will emerge to deal with the growing complexity of a dynamic knowledge ecosystem (see Wingens, 1990, for an early example).
2. Given the current general consensus that KT flow should be multi-directional and cross-disciplinary, more strategies will need to be developed to support flow and overcome barriers in non-traditional directions.
3. If Lavis et al.'s (2001) argument that researcher-user interaction should become standard practice—not simply an add-on— in research contexts is realized, this could open new communication channels from knowledge constructed in practice to new research questions and hypotheses.
4. New techniques and strategies in knowledge management will be required to handle knowledge bases of knowledge objects and associated metadata to enhance knowledge translation.

CONCLUSION

Knowledge translation is the movement and transformation of knowledge constructed in one context for the purpose of utilization in another context. It is generally agreed that interaction between researchers and intended user groups is essential to the success of translation and adoption into practice, and that knowledge flow should be multi-directional and cross-disciplinary. However, most of the work to date has focused on translation from knowledge constructed in a research context to usable knowledge embedded in practice.

The KT process consists of: (1) construction of source knowledge K_A in context A, (2) construction of target problem Q_B in context B, (3) recognition of potential connection between $K_A \leftrightarrow Q_B$, (4) planning of the translation $K_A \rightarrow K_{AB}$, (5) translation of $K_A \rightarrow K_{AB}$, (6) support for the adoption of K_{AB} in target context B, and (7) evaluation of the process. Strategies are required at multiple levels to increase the success of each of these activities, and information technologies can provide powerful support for the implementation of these strategies.

REFERENCES

Backer, T.E. (1991). Knowledge utilization: The third wave. *Knowledge: Creation, Diffusion, Utilization, 12*(3), 225-240.

Balas, E.A. (2001). Information systems can prevent errors and improve quality. *Journal of the American Medical Informatics Association, 8(4), 398-399.*

Caburnay, C.A., Kreuter, M.W., & Donlin, M.J. (2001). Disseminating effective health promotion programs from prevention research to community organizations. *Journal of Public Health Management & Practice, 7*(2), 81-89.

Caplan, N. (1979). The two-communities theory and knowledge utilization. *American Behavioral Scientist, 22,* 459-470.

Cernada, G.P. (1982). *Knowledge into action: A guide to research utilization.* Farmingdale, NY: Baywood.

Choi, B.C.K. (2005). Understanding the basic principles of knowledge translation. *Journal of Epidemiology and Community Health, 59*(2), 93.

CIHR. (2004). *Knowledge translation strategy 2004-2009: Innovation in action.* Ottawa, ON: Canadian Institutes of Health Research.

Dasgupta, S. (1989). The structure of design processes. In M. Yovitz (Ed.), *Advances in computers.* San Diego, CA: Academic Press.

Glasgow, R.E., Lichtenstein, E., & Marcus, A.C. (2003). Why don't we see more translation of health promotion research to practice? Rethinking the efficacy-to-effectiveness transition. *American Journal of Public Health, 93*(8), 1261-1267.

Greiner, A.C., & Knebel, E. (Eds.). (2003). *Health professions education: A bridge to quality.* Washington, DC: National Academy Press.

Groeneboer, C. (2006, June 14-17). Knowledge translation in teaching and learning. *Proceedings of the Society for Teaching and Learning in Higher Education Conference 2006*: *Knowledge and its Communities,* Toronto, Ontario, Canada.

Grol, R., & Jones, R. (2000). Twenty years of implementation research. *Family Practice, 17,* S32-S35.

Ho, K., Chockalingam, A., Best, A., & Walsh, G. (2003). Technology-enabled knowledge translation: Building a framework for collaboration. *CMAJ, 168*(6), 710-711.

Holden, N.J., & von Kortzfleich, H.F.O. (2004). Why cross-cultural knowledge transfer is a form of translation in more ways than you think. *Knowledge and Process Management, 11*(2), 127-136.

Huberman, M. (1987). Steps toward an integrated model of research utilization. *Knowledge: Creation, Diffusion, Utilization, 8*(4), 586-611.

Jacobson, N., Butterill, D., & Goering, P. (2003). Development of a framework for knowledge translation: Understanding user context. *Journal of Health Services Research & Policy, 8*(2), 94-99.

Lavis, J., Robertson, D., Woodside, J.M., Mcleod, C.B., & Abelson, J. (2003). How can research organizations more effectively transfer research knowledge to decision makers? *Milbank Quarterly, 81*(2), 221-248.

Lavis, J., Ross, S., Hurley, J., Hohenadel, J., Stoddart, G., Woodward, C., & Abelson, J. (2001). *Reflections on the role of health-services research in public policy-making.* Paper 01-06.

Logan, J., & Graham, I.D. (1998). Toward a comprehensive interdisciplinary model of health care research use. *Science Communication, 20*(2), 227-246.

Lomas, J. (1997). Improving research and uptake in the health sector: Beyond the sound of one hand clapping. *Policy Commentary*, *C97-1*(November).

National Cancer Institute. (2002). *Making health communication programs work.* NIH Publication No. 89-1493, Public Health Service, U.S. Department of Health and Human Services, USA.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company.* New York: Oxford University Press.

Nutley, S., Walter, I., & Davies, H.T.O. (2003). From knowing to doing: A framework for understanding the evidence-into-practice agenda. *Evaluation, 9*(2), 125-148.

Oxford English Dictionary. (n.d.). *Nexus.* Retrieved from *http://dictionary.oed.com/knowledge, n.http://dictionary.oed.com/cgi/entry/50127602?query_type=word&queryword=knowledge&first=1&max_to_show=10&sort_type=alpha&result_place=1&search_id=hPOM-sXd4Qo-5006&hilite=50127602; nexus, n. http://dictionary.oed.com/cgi/entry/00324112?single=1&query_type=word&queryword=nexus&first=1&max_to_show=10*

Wingens, M. (1990). Toward a general utilization theory: a systems theory reformulation of the two-communities metaphor. *Knowledge: Creation, Diffusion, Utilization, 12,* 27-42.

KEY TERMS

Adoption: The process by which practitioners accept new knowledge translated from research results and transform their personal practice to incorporate the translated knowledge.

Applied Knowledge: Knowledge constructed in one context, often a research context, and translated to another context to solve a problem or improve practice. Applied knowledge is distinguished from theoretical or abstract knowledge and implies practical use.

Blog: Short form of *Web log*, a chronology-based Web application for sharing information and commenting on the shared information. It is usually organized around a particular topic with most-recent entries displayed at the top. It is sometimes used as an online personal diary in which the owner posts entries and invites others to comment.

Implementation Research: Research designed to reduce disparity or space between any and all steps from research to practice.

Knowledge Construction: The creation of knowledge, typically using the paradigm(s) of the community in which the knowledge builder is situated.

Knowledge Context: The contexts in which knowledge is constructed, exchanged, utilized, and between which knowledge flows, including broad categories such as research, education and training, practice and policy, and evaluation.

Knowledge Flow: The process that transforms knowledge from constructed knowledge in the source context to translated knowledge embedded in practice in the target context.

Knowledge Transformation: The process of changing the form of the knowledge constructed in the source context to a form that is usable in the target context.

Knowledge Translation: The movement and transformation of knowledge constructed in one context for the purpose of utilization in another context. This broad definition attempts to maintain direction-neutrality. In contrast, the CIHR (2004) defines KT within the context of health care:

Knowledge translation is the exchange, synthesis and ethically-sound application of knowledge — within a complex system of interactions among researchers and users — to accelerate the capture of the benefits of research for Canadians through improved health, more effective services and products, and a strengthened health care system. (http://www.cihr-irsc.gc.ca/e/29418.html)

Synonyms include: knowledge exchange, knowledge transfer, research transfer, technology transfer, knowledge transformation, knowledge dissemination, knowledge mobilization, knowledge utilization, and research utilization.

KT Model: A framework designed to facilitate knowledge translation. Criteria for a good model (adapted from Dasgupta, 1989) include: (1) a structural form that describes the components of the process and their interrelationships, (2) an explanation of the process so that design decisions can be inferred, (3) predictive power so that impacts of certain changes in the system can be anticipated, (4) the model should serve as a basis for analysis and criticism of the process, and (5) the model should serve as a basis for exploration and testing of design options.

Linker: A person who facilitates the translation of knowledge from one context to another by bridging ideas and people in different domains. A key role is in recognizing a link between knowledge constructed in one context and its potential translation into usable knowledge in another context.

Nexus: The *Oxford English Dictionary* (n.d.) defines nexus as a means of connection between things, a connected group, a central point of convergence; a focus; a meeting place. Used conceptually to describe the *space* needed to support knowledge translation.

Wiki: A resource-based Web application that supports collaborative creation of content.

An Overview of Semantic-Based Visual Information Retrieval

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

Content-based image retrieval (CBIR) could be described as a process framework for efficiently retrieving images from a collection by similarity. The retrieval relies on extracting the appropriate characteristic quantities describing the desired contents of images. Content-based video retrieval (CBVR) made its appearance in treating video in the similar means as CBIR treating images. Content-based visual information retrieval (CBVIR) combines CBIR and CBVR together (Zhang, 2003).

With the progress of electronic equipments and computer techniques for visual information capturing and processing, a huge number of image and video records have been collected. Visual information becomes a well-known information format and a popular element in all aspects of our society. The large visual data make the dynamic research to be focused on the problem of how to efficiently capture, store, access, process, represent, describe, query, search, and retrieve their contents. In the last years, CBVIR has experienced significant growth and progress, resulting in a virtual explosion of published information. It has attracted many interests from image engineering, computer vision and the database community.

The current focus of CBVIR is around capturing high-level semantics, that is, the so-called Semantic-based Visual Information Retrieval (SBVIR). This article will first show some statistics about the research publications on SBVIR in recent years to give an idea about its developments statue. It then gives an overview on several current centers of attention, by summarizing results on subjects such as image and video annotation, human-computer interaction, models and tools for semantic retrieval, and miscellaneous techniques in applications. Finally, some future research directions, the domain knowledge and learning, relevance feedback and association feedback, as well as research at even high levels, such as cognitive level and affective level, are pointed out.

BACKGROUND

To get a general idea about the scale and progress of research on CBVIR and SBVIR for the past years, several searches in EI Compendex database (http://www.ei.org) for papers published in English from 1995 through 2005 have been made. In Table 1, the results of two searches in the title field for the numbers of English published papers (records) are listed: One term used is "image retrieval (IR)" and other term is "semantic image retrieval (SIR)." The papers found out by the second term should be a subset of the papers found out by the first term. Both numbers are increasing in that period, as seen from Table 1.

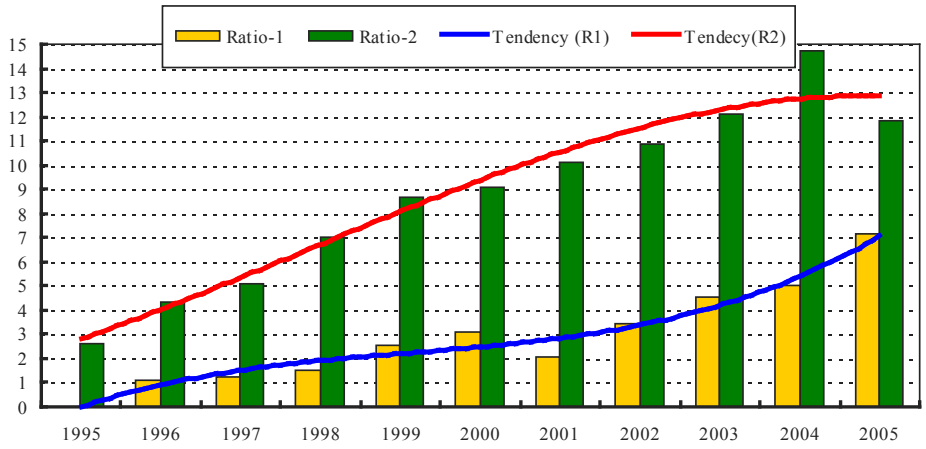
Table 1. List of English records found in the title field of EI Compendex

Searching Terms	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Total
(1) Image Retrieval	70	89	80	131	155	161	191	233	241	358	417	2126
(2) Semantic Image Retrieval	0	1	1	2	4	5	4	8	11	18	30	84
Ratio of (2) over (1)	0	1.12	1.25	1.53	2.58	3.11	2.09	3.43	4.56	5.03	7.19	3.95

Table 2. List of English records found in the subject/title/abstract field of EI Compendex

Searching Terms	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	Total
(1) Image retrieval	421	580	531	640	718	871	1080	1203	1267	2196	2174	11681
(2) Semantic image retrieval	11	25	27	45	62	79	109	131	153	324	257	1223
Ratio of (2) over (1)	2.61	4.31	5.08	7.03	8.64	9.07	10.09	10.89	12.08	14.75	11.82	10.47

Figure 1. Comparison of two groups of ratios



Other searches take the same terms as used for Table 1, but are performed in the field of title/abstract/subject. The results are shown in Table 2.

The numbers of records in Table 2 for both terms are increasing in that period, too.

Comparing the ratios of SIR over IR in two tables, these ratios in Table 2 are much higher than those ratios in Table 1. This difference indicates that the research for SIR is still in an early stage (many papers have not put the word "semantic" in the title of papers) but this concept has started to get numerous considerations and attracts much attention ("semantic" appeared already in the abstract or subject parts of these papers).

To have a closer comparison, these ratios in Table 1 and Table 2 are plotted together in Figure 1. In Figure 1, light bars represent ratios from Table 1 and dark bars represent ratios from Table 2. In addition, the tendencies of ratio developments are approximated by a third order polynomial. It is clear that many papers have the "semantic" concept in mind although they do not always use the word "semantic" in the title.

MAIN FOCUS OF THE CHAPTER

Recently, a book specially contributing to SBVIR has been published (Zhang, 2007) from which, the current advancements on SBVIR on various topics can be perceived.

From Feature to Semantics

It is recognized that high-level research often relies on low-level investigation, so the development of feature-based techniques would considerably help semantic-based techniques.

A suitable start for going into the complex problem of content representation and description can be found in Konstantinidis, Gasteratos, and Andreadis (2007). Considering IR as a collection of techniques for retrieving images based on features (in its general sense), both low-level feature and high-level feature, and especially their relations, are discussed. An efficient way to present these features is by means of a statistical tool capable of bearing concrete information, such as the histogram. A number of IR systems using histograms is presented in a thorough manner and some experimental results are discussed. The steps in order to develop a custom IR system, along with modern techniques in image feature extraction, are also presented.

Among the existing CBIR techniques based on different perceptual features, shape-based ones are particularly challenging due to the intrinsic difficulties in dealing with shape localization and recognition problems. Nevertheless, there is no doubt that shape is one of the most important perceptual features, and successful shape-based techniques would significantly improve the spreading of general-purpose systems. A shape-based image retrieval approach, which is able to efficiently deal with domain independent images with

possible cluttered backgrounds and partially occluded objects, is proposed in Sangineto (2007). It is based on an alignment approach proven robust in rigid object recognition, which has been modified in order to deal with inexact matching between the stylized shape input by the user as query and the real shapes represented in the system's database.

Video has a large data volume and complex structure. Automatic video segmentation into semantic units is important in effectively organizing long videos. A statistical video scene segmentation approach, which detects scenes boundaries in one pass by fusing multimodal audio-visual features in a symmetrical and scalable manner, is derived (Parshin & Chen, 2007). The approach deals properly with the variability of real-valued features and models their conditional dependence on the context. It also integrates prior information concerning the duration of scenes. Two kinds of features, video coherence and audio dissimilarity, extracted both from visual and audio domains, are used in the process of scene segmentation. This approach effectively fuses multiple modalities with higher performance as compared with an alternative rule-based fusion technique.

Image and Video Annotation

Currently, the topic of image and video annotation gets a lot of attention from the SBVIR research community. Text could be considered as a compact medium that expresses more abstract sense than image and video do. By annotating image and video with characteristic textural entities, their semantic contents would be efficiently represented and be used in retrieval.

A novel framework for image categorization and automatic annotation has been proposed in Xu and Zhang (2007). Image classification and automatic annotation could be treated as effective solutions to enable keyword-based SIR. To choose representative features for obtaining information from images, a feature selection strategy is proposed and visual keywords are constructed, by using both discrete and continuous methods. Based on the selected features, the Integrated Patch (IP) **model** is proposed to describe the properties of different image categories. As a generative model, the IP model describes the appearance of the mixture of visual keywords, in considering the diversity of the object composition. The parameters of IP model are then estimated by EM algorithm. Some experimental results on Corel image dataset and Getty Image Archive demonstrate that the proposed feature selection and image description model are effective in image categorization and automatic image annotation, respectively.

Automatic and semi-automatic techniques for image annotation have also been investigated in Shah, Benton, Wu, and Raghavan (2007). When retrieving images, users may find that it is easier to express the desired semantic content with keywords than with visual features. Current methods for automatically extracting semantic information from images can be classified into two classes. One is text-based methods, which use metadata such as ontological descriptions or texts associated with images, to assign or refine annotations. Although highly specialized in domain (context) specific image annotations, the text-based methods are usually semi-automatic. Another is image-based methods, which focus on extracting semantic information directly and automatically from image content, though they are domain independent and could deliver arbitrarily poor annotation performance for certain applications. By identifying some currently unsolved problems in these two classes, several suggestions for future research directions are pointed.

To provide a methodology allowing the integration of the results of content analysis of visual information, the adaptive metadata generation is proposed in Sasaki and Kiyoki (2007). It consists of the content descriptors and their text-based representation to attain the semantically precise results of keyword-based image retrieval operations. The main visual objects under discussion are images, which do not have any semantic representations therein. Those images demand textual annotation of precise semantics that is to be based on the results of automatic content analysis but not on the results of time-consuming manual annotation processes. The technical background and literature review on a variety of annotation techniques for visual information retrieval has been outlined. The proposed method and its implemented system for generating metadata or textual indexes to visual objects by using content analysis techniques are then described.

Human-Computer Interaction

Human beings play an important role in semantic level procedures. By putting humans into the loop of computer routine, it is quite convenient to introduce the domain knowledge into description module and to improve the performance of retrieval system greatly.

A typical form of human-computer interaction is that of relevance feedback whereby users supply relevance information on the retrieved images. This information can subsequently be used to optimize retrieval parameters and to enhance retrieval performance. A comprehensive review of existing relevance feedback techniques and a number of limitations for browsing is discussed in Heesch and Rueger (2007). Browsing models where the merit of hierarchical structures and networks for interactive image search are also evaluated. This exposition provides many details to enable the practitioner to implement many of the techniques.

A method of semi-automatic ground truth annotation for benchmarking of face detection in video is proposed in Tsishkou, Chen, and Bovbel (2007). It aims to illustrate the solution where an image processing and pattern recognition expert is able to label and annotate facial patterns in video

sequences at the rate of 7500 frames per hour. These ideas are extended to the semi-automatic face annotation methodology, where all object patterns are categorized into four classes in order to increase flexibility of evaluation results analysis. A guide to speed up manual annotation process by 30 times is presented.

To overcome the limitations of keyword- and content-based visual information access, an ontology-driven framework is developed in Dasiopoulou, Doulaverakis, Mezaris, Kompatslaris, and Strintzis (2007). Under the proposed framework, an appropriately defined ontology infrastructure is used to drive the generation of manual and automatic image annotations and to enable semantic retrieval by exploiting the formal semantics of ontology. In this way, the descriptions considered in the tedious task of manual annotation are constrained to named entities, because the ontology-driven analysis module can automatically generate annotations concerning common domain objects of interest.

Models and Tools for Semantic Retrieval

New models and tools for semantic retrieval have continuously been incorporated in recent years. As for other disciplines or subjects, the progresses of research on SBVIR should also get support from a variety of mathematic models and technique tools. Several models and tools utilized in SBVIR are thus introduced and some pleasing results are obtained.

A machine learning based model for constructing index of retrieval system is proposed (Hacid & Zighed, 2007). A suitable index makes it possible to group data according to similarity criteria. An effective method for locally updating neighborhood graphs, which constitute the structure of required index, is first introduced. This structure is then exploited in order to make the retrieval process using queries in an image form more easy and effective. In addition, the indexing structure is used to annotate images in order to describe their semantics. The proposed approach is based on an intelligent manner for locating points in a multidimensional space.

A critical literature review of the use of neural networks for CBIR systems is presented in Verma and Kukkarni (2007). It shows how neural networks and fuzzy logic can be used in various retrieval tasks, such as interpretation of queries, feature extraction and classification of features by describing a detailed research methodology. It investigates a neural network-based technique in conjunction with fuzzy logic to improve the overall performance of the CBIR systems. The results of the investigation on a benchmark database with a comparative analysis are presented. The methodologies and results presented will allow researchers to improve and compare their methods and will allow system developers to understand and implement the neural network and fuzzy logic-based techniques for CBIR.

A new emotion-based video scene retrieval method is proposed in Yoo (2007). Five features extracted from a video are represented in a genetic chromosome and target videos that users have in mind are retrieved by the interactive genetic algorithm through the feedback iteration. After selecting the videos that contain the corresponding emotion from the initial population of videos by the proposed algorithm, the feature vectors extracted from them are regarded as chromosomes, and a genetic crossover is applied to those feature vectors. Next, new chromosomes after crossover and feature vectors in the database videos are compared based on a similarity function to obtain the most similar videos as solutions of the next generation. By iterating this process, a new population of videos that users have in mind are retrieved.

Miscellaneous Techniques in Applications

Research on SBVIR is still in its infancy, and a large number of special ideas and exceptional techniques have been applied. These works provide new sights and fresh views from various sides, and enrich the technique pool for treating the process on semantics.

The functionalities of multimedia databases that are not present in traditional databases are discussed in Picariello and Sapino (2007). Multimedia data are inherently subjective. For retrieval purposes, such subjective information needs to be combined with objective information obtained through (generally imprecise) data analysis processes. Therefore, the inherently fuzzy nature of multimedia data, both at subjective and at objective sides, may lead to multiple, possibly inconsistent, interpretations of data. A fuzzy nonfirst normal form (FNF^2) data model that is an extension of the relational models is presented. It takes into account subjectivity and fuzziness. It enables user-friendly information access and manipulation mechanisms.

A new approach with multiple steps for improving image retrieval accuracy by integrating semantic concepts is presented in Kherfi and Ziou (2007). First, images are represented according to different abstraction levels. At the lowest level, they are represented with visual features. At the middle level, they are represented with a set of very specific keywords. At the highest level, they are represented with keywords that are more general. Second, visual content together with keywords are used to create a hierarchical index. A probabilistic classification approach is proposed, which allows to group similar images into the same class. Finally, this index is exploited to define three retrieval mechanisms: text-based, content-based, and a combination of both.

Finally, a comprehensive review of analysis algorithms to extract semantic information from multimedia content is presented in Magalhães and Rueger (2007). It considered SBVIR as a combination of multimedia understanding, information extraction, information retrieval and digital

libraries. Some statistical approaches to analyses image and video contents are described and discussed.

FUTURE TRENDS

With the recent advancements, some future research work can still be conducted.

(1) Domain Knowledge and Learning: Domain knowledge is critical in the SBVIR system to make the system more competent to handle the semantics of the query. The knowledge, which can be in the form of rules, heuristics or constraints, could be acquired by using various learning techniques, such as active learning, or multiple instance learning. An active role of statistical learning will bridge the gap between the user's desire and the system's reply.
(2) Relevance Feedback and Association Feedback: Feedback has been used for refining the retrieval results. The use of relevance feedback based on high-level content description in the object level could further improve the performance. In addition, association feedback tries to find out the associated parts between the existing interest (user intent) and new target (related to the current retrieval results), which bridges to the new retrieval (Xu & Zhang, 2003).
(3) Higher Level Exploration: Semantic level is higher than feature level, while affective level is higher than semantic level (Hanjalic, 2001). Affection is associated with some abstract attributes, which are quite subjective.

CONCLUSION

An introduction to SBVIR, some statistics about its publications, and several practical approaches, such as multi-level model, classification and annotation, machine-learning techniques, human-computer interaction, as well as various models and tools have been discussed. Few potential research directions, such as the domain knowledge and learning, relevance feedback and association feedback, as well as research at even high level, are discussed.

ACKNOWLEDGMENT

This work has been supported by the Grants NNSF-60573148 and SRFDP-20050003013.

REFERENCES

Dasiopoulou, S., Doulaverakis, C., Mezaris, V., Kompatslaris, I., & Strintzis, M. G. (2007). An ontology-based framework for semantic image analysis and retrieval. *Semantic-based visual information retrieval* (pp. 208-228).

Hacid, H., & Zighed, A. D. (2007). A machine learning based model for content based image retrieval. *Semantic-based visual information retrieval* (pp. 230-251).

Hanjalic, A. (2001). Video and image retrieval beyond the cognitive level: The needs and possibilities. *SPIE, 4315*, 130-140.

Heesch, D., & Rüeger, S. (2007). Interaction models and relevance feedback in image retrieval. *Semantic-based visual information retrieval* (pp. 160-186).

Kherfi, M. L., & Ziou, D. (2007). A hierarchical classification technique for semantics-based image retrieval. *Semantic-based visual information retrieval* (pp. 311-333).

Konstantinidis, K., Gasteratos, A., & Andreadis, I. (2007). The impact of low-level features in semantic-based image retrieval. *Semantic-based visual information retrieval* (pp. 23-45).

Magalhães, J., & Rüeger, S. (2007). Semantic multimedia information analysis for retrieval applications. *Semantic-based visual information retrieval* (pp. 334-354).

Parshin, V., & Chen, L. (2007). Statistical audio-visual data fusion for video scene segmentation. *Semantic-based visual information retrieval* (pp. 68-88).

Picariello, A., & Sapino, M. L. (2007). Managing uncertainties in image databases. *Semantic-based visual information retrieval* (pp. 292-310).

Sangineto, E. (2007). Shape based image retrieval by alignment. *Semantic-based visual information retrieval* (pp. 46-67).

Sasaki, H., & Kiyoki, Y. (2007). Adaptive metadata generation for integration of visual and semantic information. *Semantic-based visual information retrieval* (pp. 135-158).

Shah, B., Benton, R., Wu, Z., & Raghavan, V. (2007). Automatic and semi-automatic techniques for image annotation. *Semantic-based visual information retrieval* (pp. 112-134).

Tsishkou, D., Chen, L., & Bovbel, E. (2007). Semi-automatic ground truth annotation for benchmarking of face detection in video. *Semantic-based visual information retrieval* (pp. 187-207).

Verma, B., & Kulkarni, S. (2007). Neural networks for content based image retrieval. *Semantic-based visual information retrieval* (pp. 252-272).

Xu, Y., & Zhang, Y. J. (2003). Semantic retrieval based on feature element constructional model and bias competition mechanism. *SPIE*, *5021*, 77-88.

Xu, F., & Zhang, Y. J. (2007). A novel framework for image categorization and automatic annotation. *Semantic-based visual information retrieval* (pp. 90-111).

Yoo, H. W. (2007). Semantic-based video scene retrieval using evolutionary computing. *Semantic-based visual information retrieval* (pp. 273-290).

Zhang, Y. J. (2003). *Content-based visual information retrieval.* Beijing, China: Science Publisher.

Zhang, Y. J. (2007). *Semantic-based visual information retrieval.* IRM Press.

KEY TERMS

Content-Based Image Retrieval (CBIR): A process framework for efficiently retrieving images from a collection by similarity. The retrieval relies on extracting the appropriate characteristic quantities describing the desired contents of images. In addition, suitable querying, matching, indexing and searching techniques are required.

Content-Based Video Retrieval (CBVR): A process framework for efficiently retrieving required clip from video. The retrieval relies on the organization of video and nonlinear search techniques.

Content-Based Visual Information Retrieval (CBVIR): A combination of CBIR and CBVR.

Feature-Based Image Retrieval: A branch of CBIR, which is based on specific visual characteristics called features, such as color, texture, shape, and so forth, and is considered at a low abstraction level.

Image Engineering: An integrated discipline/subject comprising the study of all the different branches of image and video techniques.

MPEG-7: An international standard named "Multimedia content description interface" (ISO/IEC 15938). It provides a set of audiovisual description tools, descriptors and description schemes for effective and efficient access (search, filtering and browsing) to multimedia content.

Semantic Gap (SG): The discrepancy between the perceptual property and semantic meaning of images in the context of CBVIR. The semantic gap is also considered as a gap between current techniques and human requirements.

An Overview of Software Engineering Process and Its Improvement

Alain April
École de Technologie Supérieure, Montréal, Canada

Claude Laporte
École de Technologie Supérieure, Montréal, Canada

INTRODUCTION

The software engineering process is concerned with the definition, implementation, measurement, change, and improvement of software processes.

This short article presents software engineering process knowledge along the lines of the software engineering body of knowledge (International Organization for Standardization & International Electrotechnical Commission [ISO/IEC], 2005b). The objective of the software engineering process is to implement new or better processes in current software engineering practice.

BACKGROUND

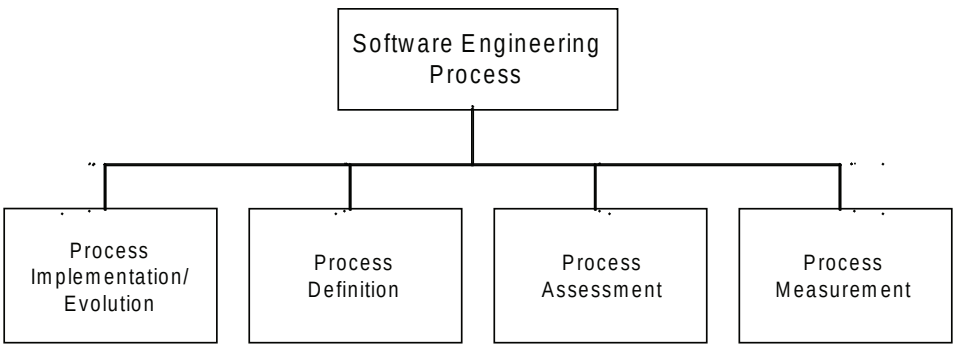
Software engineering is a young discipline, and many authors maintain that process engineering is crucial to its success, as well as being key to software quality assurance activities. This article presents generally accepted knowledge about the software engineering process. This knowledge has been adapted from industrial engineering, the management sciences, and human resources management. We have witnessed the emergence of software engineering process literature during the past 20 years and watched as some process topics have appeared while others have disappeared. This article presents four key topics (see Figure 1) that represent the fundamental concepts that must be acquired by all software engineers.

PROCESS IMPLEMENTATION AND EVOLUTION

Process implementation and change concern the initial deployment of processes and ongoing changes designed to improve and develop a supporting infrastructure (software process assets). Software engineering process activities typically follow a life cycle in which some process models are used as a reference, and certain practical considerations must be considered to ensure their success.

The first section of this article introduces concepts relating to the initial deployment of processes and to the improvement of current processes. In both cases, existing software engineering practices have to evolve. If the evolution process is extensive, then the possibility of cultural changes within

Figure 1. Key topics in the software engineering process and its improvement



the organization may need to be addressed to lower the risk of resistance and failure.

Software process improvement typically follows an improvement life cycle composed of four activities: (a) Establish the process infrastructure and assets, (b) plan the implementation (or improvement), (c) implement and evolve the process, and (d) evaluate the process. Improvement is often a project in and of itself, requiring appropriate planning, resources, monitoring, and review. Completing these life cycle activities permits continuous feedback and improvement of the software process. The first activity, establishing the process infrastructure and assets, involves establishing commitment to the process implementation and change, and acquiring the appropriate resources and personnel. The objective of the second activity, planning the implementation (or improvement), is to describe and communicate the improvement project's objectives and process needs following an assessment of the strengths and weaknesses of the current processes. The third activity, implementing and evolving the process, involves executing the planning step and deploying new processes or evolving existing processes, or both. This activity will often require piloting the new or enhanced processes. The last activity, evaluating the process, is concerned with measuring the resulting process and assessing how well it has achieved the initial objectives. This information is then used as input for subsequent improvement cycles.

The need for an appropriate software engineering infrastructure should always be considered in process improvement. This includes having the resources as well as a clear assignment of process ownership. Management commitment is essential to the success of the process improvement effort. Having an individual or an isolated group develop and evolve the software engineering processes in isolation, sometimes using proven practice handbooks, may not be the best approach as it often creates the impression that the process has been imposed by an individual or a specific organization (like quality assurance). It would be better to establish mixed-group committees that are involved in the software engineering process definition and evolution as this will ensure better representation and involvement of all software engineering staff. Two examples of such committees are the Software Engineering Process Group (Fowler & Rifkin, 1990) and the Experience Factory (Basili, Caldiera, McGarry, Pajerski, Page, & Waligora, 1992).

Moitra (1998) presents guidelines for process implementation and evolution within software engineering organizations. Hutton (1994) debates the importance of change agents in the case of major process evolution. Organizational change can also be viewed from the perspective of technology transfer (Rogers, 1983). Krasner (1999) presents examples of software definition and evolution initiatives.

PROCESS DEFINITION

The defining of processes can be represented in models, as well as in the automated process infrastructure. In an organization, a process is often composed as a procedure, a policy, or a standard, and software engineering processes are defined to harmonize software engineering activities and communication, as well as to support process improvement and its automation. In the software engineering body of knowledge, a process is defined in terms of four perspectives: life-cycle models, software life-cycle processes, notations, and automation.

Life-cycle models serve as high-level definitions of the phases that occur during development, maintenance, and operations. They are not aimed at providing detailed definitions, but rather at highlighting the key activities and their interdependencies. Examples of life-cycle models in practice are the waterfall model, the prototyping model, the evolutionary model, incremental or iterative development, the spiral model, and the reusable software model, among others (Comer, 1997).

Definitions of life-cycle processes tend to be more detailed than framework models, and unlike the standards associated with the latter, life-cycle process standards do not attempt to order their processes in time. Therefore, in principle, life-cycle processes can be arranged to fit any of the life-cycle models. The main reference in this area is ISO/IEC (1995).

Other important standards providing process definitions include the following.

- Institute of Electrical and Electronics Engineers (IEEE) Standard 1074: developing software life cycle processes (IEEE, 1991)
- ISO/IEC Standard 14764: software maintenance (ISO/IEC, 2006)
- ISO/IEC Standard 19759: software measurement process (ISO/IEC, 2005b)

To meet some certification criteria, like ISO9001 (ISO, 2000), the CMMi (capability maturity model integration; Software Engineering Institute [SEI], 2006), or the Sarbanes-Oxley Act (SOX; Securities and Exchange Commission [SEC], 2002), the definition of software processes must be compliant with quality management standards and other reference guides. ISO9001 provides requirements for quality management processes. Specifically, for the software industry, ISO/IEC 90003 (ISO, 2004) interprets each ISO9001 clause, and ISO/IEC 20000-1 (ISO/IEC, 2005a) has recently been released to address the IT service quality management system.

Processes can be defined at different levels of abstraction (Pfleeger, 2001). Various elements of a process can be

defined, for example, as roles and responsibilities, activities, controls, artifacts, and resources. Madhavji et al. made a proposal in 1994 that sets out the types of information required to define software engineering processes. In addition to the type of information, processes are always presented using a particular representation.

A number of representations are used to define processes (Software Productivity Consortium [SPC], 1992) and, more recently, the software domain ontology (Kitchenham et al., 1999). They vary as to the type of process structure and the components, symbols, and information they define, capture, and use. While these notations are gradually being normalized (Object Management Group [OMG], 2006), current approaches a software engineer should be aware of are data flow diagrams (Yourdon & DeMarco), state charts (Harel & Politi, 1998), and ETVX (Radice, Roth, O'Hara, & Ciarfella, 1985), and for representing business processes, DFD (Gane & Sarson, 1979), office support system analysis and design (OSSAD; Commission des Communautées Européennes, 1992), and more recently, BPMN (OMG).

Process automation supports human activities by means of a set of services describing the environment's capabilities. The software engineering environment (SEE) is defined as "a set of tools providing full or partial automated support to software engineering activities." Software process automation is a new technology with significant promise (Christie, Levine, Morris, & Zubrow, 1996)

Automated tools either support the execution of the process definitions, or they provide guidance to humans performing the defined processes. There exist tools that support each notation, and these tools can execute the process definitions to provide automated support to the actual processes or to fully automate them in some instances. An overview of process modeling tools is presented by Finkelstein, Kramer, and Nuseibeh (1994), and one of process-centered environments is given by Garg (Garg & Jazayeri, 1996).

PROCESS ASSESSMENT

ISO/IEC 15504 (ISO/IEC, 2004) defines an exemplar assessment model and conformance requirements on other assessment models. Popular process assessment models available are the following.

- SW-CMMi for software development processes (SEI, 2006)
- S3M for small software maintenance processes (April & Bran, 2008)
- Information Technology Infrastructure Library (ITIL, 2007) and CMMi (SEI, 2007) for IT services (data center processes)

Many more capability maturity models have been developed over the years. Process assessment using a capability maturity model is often referred to as process maturity assessment. In order to perform a maturity assessment, a specific assessment method needs to be followed to produce a quantitative score that characterizes the capability of the process (also referred to as the maturity of the organization). For example, SCAMPI (standard CMMi appraisal method for process improvement; SEI, 2000) focuses on assessments for the purpose of process improvement using the CMMi. The activities performed during an assessment, the distribution of effort on these activities, and the atmosphere during an assessment are different if the purpose is improvement and not a contract award.

ISO9001 is another process model that has been applied by software organizations. ISO9001 conformity is assessment using an audit process rather than an assessment method per se. There are five key differences between the maturity assessment and the ISO9001 audit.

a. Level of involvement: The organization's level of involvement is greater with the maturity assessment.
b. Review method: Maturity assessment reviews are performed in small groups, which facilitates and stimulates communication. These reviews do not focus as much on quality documentation.
c. Results reporting: Maturity assessment results are presented in their initial and final form in a presentation to all employees of the organizational group that was evaluated. The ISO9001 audit report is presented to the management representative only.
d. Assessment: Maturity assessment requires a chief evaluator and an assessment team (five to nine people), whereas the ISO9001 audit requires one auditor.
e. Certification of the evaluators: Who evaluates the evaluators to ensure the quality of their evaluations? ISO/IEC prescribes four evaluator insurance levels: (a) ISO9001 auditor courses, (b) evaluation of auditor performance by the registrar, (c) evaluation of the registrars through accreditation, and (d) evaluation of the accreditation by the International Accreditation Forum. Maturity assessment, by contrast, has only three levels: (a) courses, (b) certification, and (c) registration of the chief evaluators once they have been supervised for a few evaluations.

PROCESS MEASUREMENT

Software engineering process measurement can be performed to support the initiation of process implementation and change, or to evaluate the consequences of process imple-

mentation and change. Key terms on software measures and measurement methods have been defined in ISO/IEC 15939 (ISO/IEC, 2007).

Process measurement, as used here, means that quantitative information about the process is collected, analyzed, and interpreted. It can be used for process conformance assessment, process improvement, evaluation of suppliers' process capability, and benchmarking with other organizations. Measurement is used to identify the strengths and weaknesses of processes, to assess conformance, and to evaluate processes after they have been implemented and/or changed. The steps for deploying a software engineering measurement program are described in ISO/IEC 15939 (ISO/IEC, 2007). Software engineering process measures can be aimed at many different outcomes: quality, progress, productivity, and reliability. Therefore, measurement programs tend to measure multiple process outcomes that are important to the organization's business and to customer satisfaction.

Software process quality measures focus on error removal and its effectiveness. Progress measures report on the completion of milestones. Productivity measures represent the amount of work performed (in lines of code or function points) per person-month. A comparison of productivity can be achieved in a benchmarking exercise (International Software Benchmarking Standards Group [ISBSG], 2003).

In general, we are most concerned about process outcomes. However, in order to achieve the process outcomes that we desire (e.g., better quality, better maintainability, greater customer satisfaction), we have to measure the particular process.

Of course, it is not only the process that has an impact on outcomes. Other factors, such as the capability of the staff and the tools used, play an important role. Furthermore, the extent to which the process is institutionalized or implemented (i.e., process fidelity) is also important as it may explain why good processes do not necessarily lead to the desired outcomes.

FUTURE TRENDS

We have presented here an overview of current knowledge on software engineering process. Future trends are developing in five main directions: first, the ongoing debate regarding the use of lightweight and open-source life cycles (Agile, Open UP); second, the need for greater conformity of IT to rules and regulations (SOX, CobiT); third, the challenge of applying quality paradigms to software engineering processes (i.e., Six Sygma); fourth, the introduction of new international standards that will recommend the process support services for SEEs; and finally, we predict that there will be a consolidation of reference models for IT processes in the near future, and that simplicity and ease of use will prevail.

CONCLUSION

This short article has presented the software engineering process body of knowledge (ISO/IEC, 2005b).

REFERENCES

April, A., & Abran, A. (2008). Software maintenance management: Evaluation and continuous improvement. In *Software engineering best practice* (Vol. 1). John Wiley & Sons.

Basili, V., Caldiera, G., McGarry, F., Pajerski, R., Page, G., & Waligora, S. (1992). The software engineering laboratory: An operational software experience factory. In *Proceedings of the International Conference on Software Engineering* (pp. 370-381).

Christie, A. M., Levine, L., Morris, E. J., & Zubrow, D. (1996). *Software process automation: Experience from the trenches* (Tech. Rep. No. CMU/SEI-96-TR-013). Carnegie Mellon University.

Comer, E. (1997). Alternative software life cycle models. In M. Dorfmann & R. Thayer (Eds.), *Software engineering.* IEEE CS Press.

Commission des Communautées Européennes. (1992). *Office support system analysis and design (OSSAD): Project Esprit #285.* Retrieved from http://dumas.univ-tln.fr/Ossad/Appel%20vol%201.htm

Finkelstein, A., Kramer, J., & Nuseibeh, B. (1994). *Software process modeling and technology.* Research Studies Press Ltd.

Fowler, P., & Rifkin, S. (1990). *Software engineering process group guide* (Tech. Rep. No. CMU/SEI-90-TR-24). Software Engineering Institute. Retrieved from http://www.sei.cmu.edu/pub/documents/90.reports/pdf/tr24.90.pdf

Gane, C. P, & Sarson, T. (1979). *Structured system analysis: Tools and techniques.* Prentice Hall.

Garg, P., & Jazayeri, M. (1996). Process-centered software engineering environments: A grand tour. In A. Fuggetta & A. Wolf (Eds.), *Software process.* John Wiley & Sons.

Harel, D., & Politi, M. (1998). *Modeling reactive systems with statecharts: The statemate approach.* McGraw-Hill.

Hutton, D. (1994). *The change agent's handbook: A survival guide for quality improvement champions.* ASQC Quality Press.

Information Technology Infrastructure Library (ITIL). (2007). *Central Computer and Telecommunications Agency* (Version 3). Norwich, United Kingdom: Controller of Her

Majesty's Stationery Office. Retrieved from http://www.itsmf.org/

Institute of Electrical and Electronics Engineers (IEEE). (1991). *IEEE standard for developing software life cycle processes* (IEEE Std 1074-1991). IEEE Computer Society.

International Organization for Standardization (ISO). (2000). *ISO9001:2000, quality management systems: Requirements* (3rd ed.). Author.

International Organization for Standardization (ISO). (2004). *Software engineering: guidelines for the application of ISO9001:2000 to computer software. ISO/IEC Standard 90003:2004.* International Organization for Standardization & International Electrotechnical Commission.

International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). (1995). *ISO/IEC 12207: Information technology. Software life cycle processes.* Author.

International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). (2004). *ISO/IEC 15504-1: Information technology. Process assessment: Part 1. Concepts and vocabulary: ISO/IEC Standard 15504-1.* Author.

International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). (2005a). *ISO/IEC 20000-1: 2005 information technology. Service management: Part 1. Specification.* Author.

International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). (2005b). *ISO/IEC TR 19759: 2005 information technology. Software measurement process.* Author.

International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). (2006). *ISO/IEC 14764: Software engineering. Software maintenance: ISO/IEC Standard 14764.* Author.

International Organization for Standardization & International Electrotechnical Commission (ISO/IEC). (2007). *ISO/IEC 15939: 2007 software engineering. Guide to the software engineering body of knowledge (SWEBOK).* Author.

International Software Benchmarking Standards Group (ISBSG). (2003). Retrieved from http://www.isbsg.org

Kitchenham, B., Guilherme, H., et al. (1999). Towards an ontology of software maintenance. *Journal of Software Maintenance: Research and Practice, 11*, 365-389.

Krasner, H. (1999). The payoff for software process improvement: What it is and how to get it. In K. El-Emam & N. H. Madhavji (Eds.), *Elements of software process assessment and improvement.* IEEE CS Press.

Madhavji, N., et al. (1994). *Elicit: A method for eliciting process models.* Paper presented at the Third International Conference on the Software Process.

Moitra, D. (1998). Managing change for software process improvement initiatives: A practical experience-based approach. *Software Process: Improvement and Practice, 4*(4), 199-207.

Object Management Group (OMG). (2006). *Business process management initiative Version 1.0.* Retrieved from http://www.bpmn.org

Pfleeger, S. L. (2001). *Software engineering: Theory and practice* (2nd ed.). Prentice Hall.

Radice, R., Roth, N., O'Hara, A. Jr., & Ciarfella, W. (1985). A programming process architecture. *IBM Systems Journal, 24*(2), 79-90.

Raghavan, S., & Chand, D. (1989). Diffusing software-engineering methods. *IEEE Software*, pp. 81-90.

Rogers, E. (1983). *Diffusion of innovations.* Free Press.

Securities and Exchange Commission (SEC). (2002). *SOX: Sarbanes-Oxley Act* (Pub. L. No. 107-204, 116 Stat. 745). Retrieved from http://www.sec.gov/rules/interp/2007/33-8810.pdf

Software Engineering Institute (SEI). (2000). *Standard CMMi appraisal method for process improvement (SCAMPI): Method description. Version 1.0* (Tech. Rep. No. CMU/SEI-2000-TR-009). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

Software Engineering Institute (SEI). (2006). *CMMI product development team: Capability maturity model integration for software engineering. Version 1.2* (Tech. Rep. No. CMU/SEI-2006-TR-008). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University.

Software Engineering Institute (SEI). (2007). *CMMi for services.* Retrieved from http://www.sei.cmu.edu/cmmi/models/CMMI-Services-status.html

Software Productivity Consortium (SPC). (1992). *Process definition and modeling guidebook* (SPC-92041-CMC). Author.

KEY TERMS

Audit: It is an independent examination of a work product or set of work products to assess compliance with specifications, standards, contractual agreements, or other criteria.

Capability Maturity Model: The model is a description of the stages through which organizations evolve as they define, implement, measure, control, and improve their processes.

Maturity Level: It is a well-defined evolutionary plateau toward achieving a mature software acquisition process. The typical five maturity levels are initial, repeatable, defined, quantitative, and optimizing.

Measurement Process: This is a set of interrelated resources, activities, and influences related to a measurement.

Measurement Program: A measurement program is the set of related elements for addressing an organization's measurement needs. It includes the definition of organization-wide measurements, methods, and practices for collecting organizational measurements and analyzing data, and measurement goals for the organization.

Process Assessment: It is a disciplined evaluation of an organization's software processes against a model compatible with the reference model.

Process Assets: They are a collection of items, maintained by an organization, for use by programs in developing, tailoring, maintaining, and implementing their processes.

Software Process Assets: They are a collection of entities, maintained by an organization, for use by projects in developing, tailoring, maintaining, and implementing their software processes.

An Overview of Threats to Information Security

R. Kelly Rainer, Jr.
Auburn University, USA

INTRODUCTION

Organizations and individuals have many information assets, which are subject to an increasing number of threats. The purpose of this article is to provide (1) an overview of the factors that are increasing the threats to information security and (2) an overview of the threats to information security.

BACKGROUND

Several factors are contributing to today's dangerous threat environment. The first factor is the evolution of the information technology resource from mainframe-only to today's highly complex, interconnected, interdependent, wirelessly networked business environment. This environment exposes organizations and individuals to a world of untrusted networks and potential attackers.

The second factor results from the fact that modern computers and storage devices (e.g., thumb drives) continue to become smaller, faster, cheaper, and more portable, with greater storage capacity. These characteristics make it much easier to steal or lose a computer or storage device that contains huge amounts of sensitive information. Also, far more people are able to afford powerful computers and connect inexpensively to the Internet, thus raising the potential of an attack on information assets.

The third factor is that the computing skills necessary to be a hacker are *decreasing*. The reason is that the Internet contains information and computer programs (called scripts) that users with few skills can download and use to attack any information system connected to the Internet.

The fourth factor is that international organized crime is turning its attention to cybercrime, which are illegal activities taking place over computer networks, particularly the Internet. These crimes can be committed from anywhere in the world, at any time, effectively providing an international safe haven for cybercriminals. Computer-based crimes cause billions of dollars in damages to businesses each year, including the costs to repair information systems, the costs of lost business, and the loss of customer confidence.

The fifth factor is downstream liability. Downstream liability occurs in this manner. If company A's information systems were compromised by a perpetrator and used to attack company B's systems, then company A could be liable for damages to company B. Note that company B is "downstream" from company A in this attack scenario. A downstream liability lawsuit would put company A's security policies and operations on trial. Under tort law, the plaintiff (injured party or company B) would have to prove that the offending company (company A) had a duty to keep its computers secure and failed to do so, as measured against generally accepted standards and practices.

At some point, all companies will have some minimal set of standards that they have to meet when operating information systems that connect to the Internet. The models already exist in the form of regulations and laws (e.g., Gramm-Leach-Bliley Act and the Health Insurance Portability and Accountability Act). Contractual security obligations, particularly *service level agreements* (SLAs), which spell out very specific requirements, might also help establish a security standard. Courts or legislatures could cite typical SLA terms, such as maintaining up-to-date antivirus software, software patches, and firewalls, in crafting minimum security responsibilities.

A company being sued for downstream liability will have to convince a judge or jury that its security measures were reasonable. That is, the company must demonstrate that it had practiced due diligence in information security. Due diligence can be defined in part by what your competitors are doing, which defines best practices.

The sixth factor is increased employee use of unmanaged devices, which are devices outside the control of an organization's IT department. These devices include customer computers, business partners' mobile devices, computers in the business centers of hotels, computers in Starbucks and Paneras, and many others.

The final factor is lack of management support. For the entire organization to take security policies and procedures seriously, senior managers must set the tone and provide necessary resources. Ultimately however, lower-level managers may be even more important. These managers are in close contact with employees every day and thus are in a better position to determine whether employees are following security procedures.

THE THREAT ENVIRONMENT

Whitman and Mattord (2003) classified threats into five general categories to enable us to better understand the complexity of the threat problem. Their categories are natural disasters, technical failures, management failures, unintentional acts, and deliberate acts.

Natural disasters include floods, earthquakes, hurricanes, tornados, lightning, and in some cases, fires. In many cases, natural disasters can cause catastrophic loss of systems and data. Technical failures include problems with hardware and software. The most common hardware problem is a crash of a hard disk drive. The most common software problem is errors, called bugs, in computer programs. We discussed management failures in the Introduction.

Unintentional Acts

Unintentional acts are those with no malicious intent and consist of human errors, deviations in the quality of service by service providers, and environmental hazards. Of these three, human errors are by far the most serious threats to information security.

Before we discuss the various types of human error, we categorize organizational employees. The first category consists of regular employees. There are two important points to be made about regular employees. First, the higher the level of employee, the greater the threat the employee might pose to information security. Higher-level employees may have greater access to corporate data and enjoy greater privileges on organizational information systems. Second, employees in two areas of the organization pose significant threats to information security. Human resources employees generally have access to sensitive personal information about all employees. Likewise, information systems employees not only have access to sensitive organizational data, but they often control the means to create, store, transmit, and modify that data.

The second category of employee includes contract labor, consultants, and janitors and guards. Contract labor, such as temporary hires, may be overlooked in information security. However, these employees often have access to the company's network, information systems, and information assets. Consultants, while technically not employees, do work for the company. Depending on the nature of their

Table 1. Human mistakes

Human Mistake	**Description and Examples**
Tailgating	A technique designed to allow the perpetrator to enter restricted areas that are controlled with locks or card entry. The perpetrator follows closely behind a legitimate employee and, when the employee gains entry, asks them to "hold the door."
Shoulder surfing	The perpetrator watches the employee's computer screen over that person's shoulder. This technique is particularly successful in public areas such as airports, commuter trains, and on airplanes.
Carelessness with laptops and portable computing and storage devices	Losing or misplacing them, or using them carelessly, so that malware can be introduced into an organization's network.
Opening questionable e-mails	Opening e-mails from someone unknown, or clicking on links embedded in e-mails (see phishing attacks below).
Careless Internet surfing	Accessing questionable Websites; can result in malware and/or alien software being introduced into the organization's network.
Poor password selection and use	Choosing and using weak passwords (see strong passwords).
Carelessness with one's office	Unlocked desks and filing cabinets when employees go home at night; not logging off the company network when gone from the office for extended period of time.
Carelessness with discarded equipment	Discarding old computer hardware and devices without completely wiping the memory; includes computers, cell phones, Blackberries, and digital copiers and printers.

work, these people may also have access to the company's network, information systems, and information assets.

Finally, janitors and guards are the most frequently ignored people in information security. They are usually present when most – if not all – other employees have gone home. They typically have keys to every office, and nobody questions their presence in even the most sensitive parts of the building. In fact, an article from the Winter 1994 edition of *2600: The Hacker Quarterly* described how to get a job as a janitor for the purpose of gaining physical access to an organization.

Human errors or mistakes by employees pose a large problem as the result of laziness, carelessness, or a lack of information security awareness. This lack of awareness comes from poor education and training efforts by the organization. Human mistakes manifest themselves in many different ways, as we see in Table 1.

Deviations in the quality of service by service providers consist of situations in which a product or service is not delivered to the organization as expected. For example, heavy equipment at a construction site cuts a fiber optic line to your building or your Internet Service Provider has availability problems. Organizations may also experience service disruptions from various providers, such as communications, electricity, telephone, water, wastewater, trash pickup, cable, and natural gas.

Environmental hazards include dirt, dust, humidity, and static electricity, which are harmful to the safe operation of computing equipment.

Deliberate Acts

Deliberate acts include espionage or trespass; information extortion; sabotage and vandalism; theft of equipment and information; social engineering and reverse social engineering; identity theft; compromises to intellectual property; software attacks; and supervisory control and data acquisition attacks.

Espionage or trespass occurs when an unauthorized individual attempts to gain illegal access to organizational information. When we discuss trespass, it is important that we distinguish between competitive intelligence and industrial espionage. Competitive intelligence consists of legal information-gathering techniques, such as studying a company's Website and press releases, attending trade shows, and so on. In contrast, industrial espionage crosses the legal boundary.

Information extortion occurs when an attacker either threatens to steal, or actually steals, information from a company. The perpetrator demands payment for not stealing the information, for returning stolen information, or for agreeing not to disclose the information.

Sabotage and vandalism are deliberate acts that involve defacing an organization's Website, possibly causing the organization to lose its image and experience a loss of confidence by its customers. One form of online vandalism is a hacktivist or cyberactivist operation. These are cases of high-tech civil disobedience to protest the operations, policies, or actions of an organization or government agency.

Theft of equipment and information result from computers and storage devices becoming smaller yet more powerful with vastly increased storage (e.g., laptops, Blackberries, smart phones, digital cameras, thumb drives, and iPods). As a result, these devices are becoming easier to steal and easier for attackers to use to steal information.

The uncontrolled proliferation of portable devices in companies has led to a type of attack called pod slurping. In *pod slurping*, perpetrators plug portable devices, such as an iPod, into a USB port on a computer and download huge amounts of information very quickly and easily.

Another form of theft, known as *dumpster diving*, involves the practice of rummaging through commercial or residential trash to find information that has been discarded. Files, letters, memos, photographs, IDs, passwords, credit cards, and other forms of information can be found in dumpsters. Unfortunately, many people never consider that the sensitive items they throw in the trash may be recovered and used for fraudulent purposes.

Social engineering is an attack where the perpetrator uses social skills to trick or manipulate a legitimate employee into providing confidential company information (e.g., passwords). The most common example of social engineering occurs when the attacker impersonates someone else on the telephone, such as a company manager or information systems employee. The attacker says he forgot his password and asks the legitimate employee to give him a password to use. Other common exploits include posing as an exterminator, air conditioning technician, or fire marshal.

In one example of social engineering, a perpetrator entered a company building wearing a company ID card that looked legitimate, and put up signs on bulletin boards saying, "The help desk telephone number has been changed. The new number is 555-1234." He then exited the building and began receiving calls from legitimate employees thinking they were calling the company help desk. Naturally, the first thing the perpetrator asked for was user name and password. He now had the information necessary to access the company's information systems.

In social engineering, the attacker approaches legitimate employees. In reverse social engineering, the employees approach the attacker. For example, the attacker gains employment at a company and, in conversations, lets it be known that he is "good with computers." As is often the case, they ask him for help with their computer problems. While he is helping them, he loads Trojan horses on their computers that e-mail him with their passwords and information about their machines.

Table 2. Types of software attacks

	Description
Virus	Segment of computer code that performs malicious actions by attaching to another computer program.
Worm	Segment of computer code that performs malicious actions and will replicate, or spread, by itself (without requiring another computer program).
Trojan Horse	Software programs that hide in other computer programs and reveal their designed behavior only when they are activated.
Back Door	Typically a password, known only to the attacker, that allows the attacker to access a computer system at will, without having to go through any security procedures (also called trap door).
Logic Bomb	Segment of computer code that is embedded with an organization's existing computer programs and is designed to activate and perform a destructive action at a certain time or date.
Dictionary Attack	Attacks that try combinations of letters and numbers that are most likely to succeed, such as all words from a dictionary.
Brute Force Attack	Attacks that use massive computing resources to try every possible combination of password options to uncover a password.
Denial of Service Attack	Attacker sends so many information requests to a target computer system that the target cannot handle them successfully and typically crashes (ceases to function).
Distributed Denial of Service Attack	An attacker first takes over many computers, typically by using malicious software. These computers are called *zombies*, or *bots*. The attacker uses these bots (which form a *botnet*) to deliver a coordinated stream of information requests to a target computer, causing it to crash.
Phishing Attack	Phishing attacks use deception to acquire sensitive personal information by masquerading as official-looking e-mails or instant messages.
Zero-day Attack	A zero-day attack takes advantage of a newly discovered, previously unknown vulnerability in a software product. Perpetrators attack the vulnerability before the software vendor can prepare a patch for the vulnerability.
Rootkit	Software that enables an attacker to have administrator-level access (meaning complete control) to a computer or computer network.
Blended Threats	Combine the characteristics of viruses, worms, Trojan horses, and rootkits to initiate, transmit, and spread an attack.
Alien Software	Clandestine software that is installed on your computer through duplicitous methods. Uses valuable system resources and can capture your Web surfing habits and personal information. Includes spyware,
Keystroke Logger	Software and/or hardware that records your keystrokes. Purposes range from criminal (e.g., theft of passwords) to annoying (e.g., recording your Internet browsing history for targeted advertising).
Screen Scrapers	Software that records a continuous "movie" of a screen's Contents rather than simply recording keystrokes.

Identity theft is the deliberate assumption of another person's identity, usually to gain access to another person's financial information or to frame another person for a crime. Techniques for obtaining information include stealing mail, dumpster diving, stealing personal information in computer databases, infiltrating organizations that store large amounts of personal information (e.g., data aggregators), and impersonating a trusted organization in an electronic communication (e.g., phishing).

Recovering from identity theft is costly, time-consuming, and difficult. Victims report difficulties in obtaining credit and obtaining or holding a job, as well as adverse effects on insurance or credit rates. Victims also note that is is difficult to remove negative information from their records, such as their credit reports.

Compromises to intellectual property is a vital issue for people who make their livelihood in knowledge fields. I**ntellectual property** is the property created by individuals or corporations that is protected under trade secret, patent, and copyright laws.

The most common intellectual property related to IT deals with software. Copying a software program without making payment to the owner—including giving a disc to a friend to install on the owner's computer—is a copyright violation. Not surprisingly, this practice, called **piracy,** is a major problem for software vendors. The global trade in pirated software amounts to hundreds of billions of dollars.

Software attacks have evolved from the outbreak era, where malicious software tried to infect as many computers worldwide as possible, to the profit-driven, Web-based attacks of today. Cybercriminals are heavily involved with malware attacks to make money, and they use sophisticated, blended attacks typically via the Web. Table 2, although not inclusive, shows a variety of well-known software attacks.

A Supervisory Control and Data Acquisition (SCADA) system uses sensors to monitor and control chemical, physical, or transport processes in oil refineries, water and sewage treatment plants, electrical generators, and nuclear power plants. The sensors connect to physical equipment. They read status data such as the open/closed status of a switch or a valve, as well as measurements such as pressure, flow, voltage, and current. By sending signals to equipment, sensors control equipment, such as opening or closing a switch or valve or setting the speed of a pump.

The sensors are networked, and each sensor typically has an Internet Protocol (IP) address. In a SCADA attack, the attacker tries to gain access to the network in order to, for example, disrupt the power grid over a large area or disrupt the operations of a large chemical plant. Such actions could have catastrophic results.

FUTURE TRENDS

Unfortunately, the threat environment continues to be bleak. Attackers are moving from mass malware attacks to stealthy, blended, targeted attacks, where the attacker tries to control as many computers for as long as possible without the users being aware of it. In fact, botnets have become today's biggest threat.

CONCLUSION

The contributions of this chapter are threefold. First, it addresses the factors increasing today's level of threats. Second, it provides a brief overview of the numerous threats to information security. Third, it addresses the lengthy topic of information security threats in a concise, readable manner.

REFERENCES

Brenner, B. (2007). *How Russia became a malware hornet's nest*. Retrieved June 16, 2008, from http://searchsecurity.techtarget.com/originalContent/0,289142,sid14_gci1275987,00.html?bucket=NEWS&topic=306900

Dubie, D. (2007). *E-mail boosts productivity; IM poses threats, survey says*. Network World. Retrieved June 16, 2008, from www.networkworld.com/news/2007/101007-email-boost-productivity.html

Espiner, T. (2007). *Public sector lacks IT security sense*. BusinessWeek. Retrieved June 16, 2008, from www.businessweek.com/globalbiz/content/feb2007/gb20070202_558265.htm

Gaudin, S. (2007). *Fewer companies suffer security breaches, but they're much more severe*. Information Week. Retrieved June 16, 2008, from www.informationweek.com/showArticle.jhtml;?articleID=202100132

Greenemeier, L. (2007). *InformationWeek 500: How MassMutual got its security data under control.* Information Week. Retrieved June 16, 2008, from www.informationweek.com/security/showArticle.jhtml?articleID=201806190

Hamm, S. & D. Kopecki. (2006). *Tech's threat to national security*. BusinessWeek. Retrieved June 16, 2008, from www.businessweek.com/technology/content/nov2006/tc20061102_797312.htm

Hesseldahl, A. (2006). *Security threats come a-callin'*. BusinessWeek. Retrieved June 16, 2008, from www.businessweek.com/technology/content/aug2006/tc20060802_454386.htm

Martin, R. (2007). *Cyberthreats outpace security measures, says McAfee CEO*. Information Week. Retrieved June 16, 2008, from www.informationweek.com/showArticle.jhtml;?articleID=201807230

McMillan, R. (2008). *CIA says hackers have cut power grid*. PC World. Retrieved June 16, 2008, from http://www.pcworld.com/article/id,141564-c,hackers/article.html

Pfleeger, C. & S. Lawrence. (2002). *Security in computing* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Preston, R. (2007). *Pacific northwest national lab does cybersecurity*. Information Week. Retrieved June 16, 2008, from www.informationweek.com/blog/main/archives/2007/10/pacific_northwe.html

Soat, J. (2007). *What makes a CIO shudder?* Information Week. Retrieved June 16, 2008, from www.informationweek.com/blog/main/archives/2007/08/what_makes_a_ci.html

Stallings, W. (2005). *Cryptography and network security* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Vijayan, J. (2007). *House committee grills DHS on information security*. Computerworld. Retrieved June 16, 2008, from www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9018399

Westervelt, R. (2007). *Cybercriminals employ rootkits in rising numbers to steal data.* SearchSecurity.com. Retrieved June 16, 2008, from http://searchsecurity.techtarget.com/originalContent/0,289142,sid14_gci1271024,00.html?bucket=NEWS&topic=306900

Whitman, M. E. & Mattord, H. (2003). *Principles of information security.* Boston: Course Technology.

KEY TERMS

Downstream Liability: A potential liability incurred by a company whose computer systems are compromised by an attacker and used to attack another company's systems.

Distributed Denial of Service Attack: An attacker takes over many computers (called *zombies* or *bots*), typically by using malicious software. The attacker uses these bots (which form a *botnet*) to deliver a coordinated stream of information requests to a target computer, causing it to crash.

Phishing Attack: Attacks that use deception to acquire sensitive personal information by masquerading as official-looking e-mails or instant messages.

Rootkit: Software that enables an attacker to have administrator-level access (meaning complete control) to a computer or computer network.

Social Engineering: An attack where the perpetrator uses social skills to trick or manipulate a legitimate employee into providing confidential company information.

Tailgating: The perpetrator follows closely behind a legitimate employee and, when the employee gains entry, asks them to "hold the door" so the perpetrator can enter restricted areas that are controlled with locks or card entry.

Trojan Horse: Software programs that hide in other computer programs and reveal their designed behavior only when they are activated.

Virus: Computer code that performs malicious actions by attaching to another computer program.

Worm: Computer code that performs malicious actions and will replicate, or spread, by itself (without requiring another computer program).

Zero-Day Attack: An attack that takes advantage of a newly discovered, previously unknown vulnerability in a software product. Perpetrators attack the vulnerability before the software vendor can prepare a patch for the vulnerability.

An Overview of Trust Evaluation Models within E-Commerce Domain

Omer Mahmood
University of Sydney, Australia
Charles Darwin University, Australia

INTRODUCTION

This chapter outlines various models which can be used to predict users' trust on online shopping, to enhance user's trust on online vendor, and to estimate the risk in an online transaction. The discussed models are selected to provide an overview of different aspects which can be used by the service providers and developers to identify the factors which impact user's online trust. The factors that have been identified can be further used as a guide to enhance user's trust levels. The rest of the article is organized as follows. In the next section, four models are discusses starting from Cheung and Lee (2000) conceptual model of trust in electronic environment to the model presented by Mahmood (2006a) that relies on mathematical equations to assist user to compare and evaluate online retailers. After discussing the presented models, the impact and effect of Web 2.0 technologies are discussed in future directions. The potential use of FOAF and RDF to create completely decentralized repository of users' trust evaluations which can be tapped into any application that uses Web 2.0 is also discussed in future directions. Concluding remarks and model comparison are presented the conclusion.

BACKGROUND

In electronic commerce the gap between payment and delivery of service or product is substantial as compared to physical transactions. These gaps impact the users' estimated trust on the service provider as well as on user's decision to commit or abort a transaction. Although the estimation of trust is based on personal, subjective properties, recently, several models have been proposed to translate trust into quantifiable terms. Such models target to assist the user's to establish the degree of trust so that the user can make an informed decision while committing a transaction in electronic environment. The right degree of trust has been defined as the risk that the user accepts in case of failure is estimated to be less than the expected subjective utility in case of success (Castelfranchi & Falcone, 2000).

AN OVERVIEW OF TRUST MODELS

Trust in Internet Shopping: Model and Measurement

Cheung and Lee (2000) showed that consumer trust in online shopping can be predicted by following two sets of experiences: factors which contribute to the sense of merchant's trustworthiness, and external factors which are linked to the external environment. The merchant's associated sense of trustworthiness is linked to its perceived integrity, competence, security and privacy controls in place. The external environment contributing factors include third party recognition (e.g., trusted party seals of approval) and legal framework. The model showed that the combined effect of both sets of factors on consumer's overall trust belief is mediated by the consumer's tendency to trust. The model also acknowledges the relationship between perceived risk and online consumer's trust response.

A Model of Internet Trust from the Customer's Point of View

Ang, Dubelaar, and Lee (2001) proposed a model of trust which focuses on consumer's perception on online retailer's trustworthiness. The authors used the following equation to describe the process by which a consumer makes a decision to commit a transaction or not:

$$G_b = p_b L_b,$$

where

G_b = Gain to the user from the transaction
p_b = Subjective probability that the online merchant will be untrustworthy
L_b = The loss the consumer will suffer in case of fraud

Ang et al., in their study, argued that in order to further capture the market share the online retailer will either have to increase the LHS or reduce the RHS. For example, the

Figure 1. A conceptual model of trust in Internet shopping (Cheung & Lee, 2000)

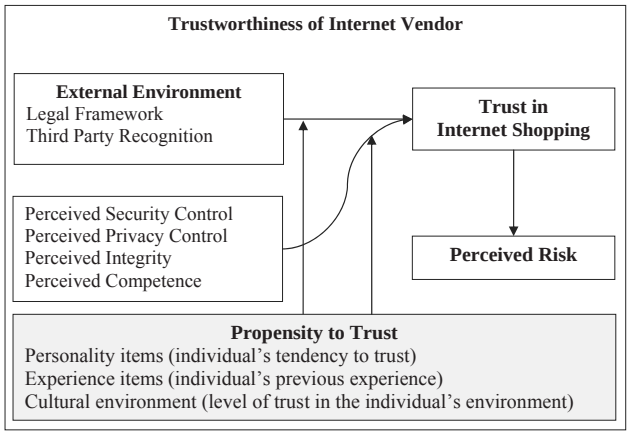


Table 1. Trust variables and corresponding levels

	Trust Variables		
Levels	**Ability to deliver**	**Willingness to rectify**	**Personal privacy**
	Known Brand	30-days guarantee	Has privacy policy statement
	Unknown Brand	No guarantee	No Privacy policy statement

LHS can be increased by giving more discounts to the customer. This will also reduce the computer value of RHS as the value of Lb will decrease. RHS can also be reduced by portraying a sense of trustworthiness. In the study following three aspects of trustworthiness in an online transaction were identified:

1. Ability of the merchant to deliver the product or service. This aspect could be highlighted with the help of previous customer's comments, access to detailed information regarding the product and availability of widely known respected brands
2. The willingness of the merchant to rectify the problem and honor its commitments for example, money-back guarantee, access to physical customer service centre or phone support system, and so forth
3. Clear statement from the merchant on the use of user's personal information that is, whether the information will be shared with third parties or not

The following table summarizes the identified trust variables and their corresponding levels used in the study

A Model of Trust for E-Commerce System Design

Egger (2000) states that "Traditional HCI analysis and design methods can be employed effectively to address usability aspects of ecommerce interfaces, but they may fail to deliver when it comes to designing trust-inducing features susceptible to convert users into customers. Indeed, the discipline of HCI currently lacks substantive knowledge about how trust is formed, maintained and lost in B2C e-commerce" (pp. 101). The author identified six factors which were later grouped into three following categories:

1. **Purchase knowledge:** Knowledge acquired before the user interacts with the system, this could be the reputation of the system, referral or user's previous experience
2. **Interface properties:** Interface properties consist of two sub properties, namely familiarity and attitude. Familiarity refers to both experience in terms of both navigation and online technology while attitude is the first impression the system makes on the user with respect to information presentation
3. **Information content:** Information content consists of risk, transparency and cooperation components. The risk component outlines the financial risks and security measures for example, insurance, and so forth. Transparency component refers to openness of the merchant in terms of privacy policy and business policies, while the cooperation content of information focuses on assisting effect of merchant and user.

Trust Evaluation in Electronic Environment

Mahmood (2006a) based his trust evaluation model on the notion that in business to consumer e-commerce trust consists of two non-separable aspects. First, it involves the trust in other party and second, trust in the transaction medium. The view of trust, adopted by Mahmood, is also consistent with the generic model of trust presented by Tan and Theon (2001). In the same study Mahmood (ibid), suggested that while committing an online transaction the user mainly considers functional and financial risk factors. The risk factors are evaluated on the basis of user's subjective initial trust assessment of functional and financial aspects of a transaction. Functional risk factors primarily relate to the business, that is, trust in party, rather than the actual

Figure 2. Revised MoTEC model (Egger, 2000)

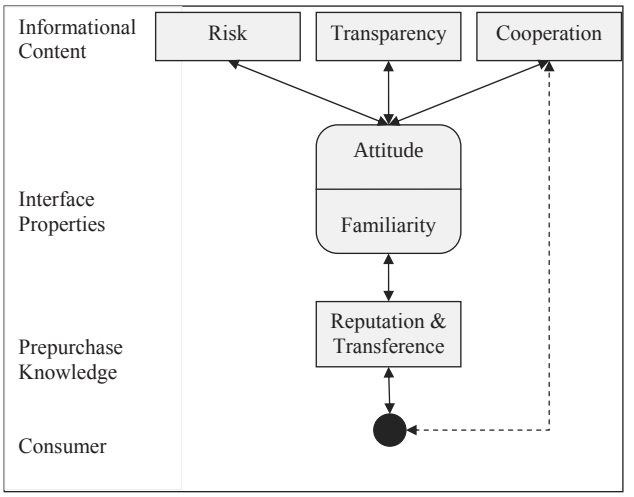
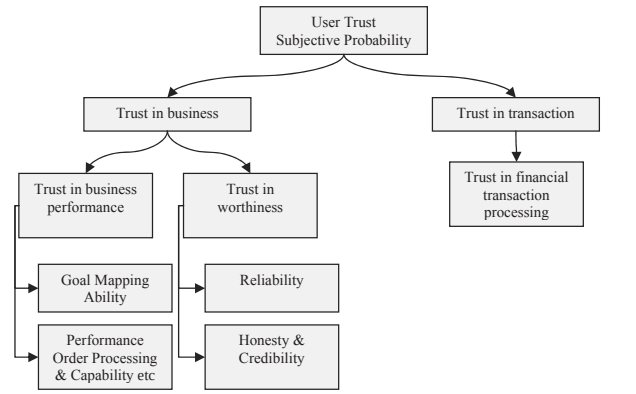


Figure 3. Dynamics of initial trust in electronic transactions (Mahmood, 2006a)



exchange of funds. The study outlined following aspects which impact user's functional trust evaluation:

- Business's capability and ability to deliver the desired good or service
- Business's ability to deliver the service and goods within desired time frame
- The business can be relied in fulfilling the desired goals
- The business will use and maintain the user's personal information with honesty
- The business is professional in conducting business

Besides functional evaluation the financial initial trust evaluation of the transaction includes risk of losing funds, either the whole or part of the transaction amount. The financial risks can be significantly reduced by providing alternate dispute resolution, transaction insurance and appropriate use of online security technologies that increase user's perceived trust on transaction medium.

From this, Mahmood concluded that trust in the e-environment consists of two main parts, firstly, trust in transaction where the transaction includes exchange of funds and second, trust in the party or business. The model further divided the user's perceived trust in online business and online transactions into subjective trust levels on the business's performance and honesty and trust in financial transaction processing respectively. The following figure outlines the involved dynamics of trust in an electronic transaction:

The study proposed following variables which represent different dynamics of trust in e-business. Each proposed variable represents the subjective probability that is, user's level of trust on each dynamic (i.e., aspect) of the transaction. The probabilities are appropriately aggregated in order to assist the user to judge and evaluate the factors involved and their determinants.

Initial Trust in Business:

p_b: Subjective probability (trust) based on business performance in terms of business ability (i.e., capability) to deliver and order processing (does not include dispute resolution as that comes after the transaction). This is dependant on online reputation, trusted referral(s) and structure and verbal communication of the Web site.

p_h: Subjective probability (trust) based on business honesty in terms of reliability and credibility. This is based on trusted referral(s), presence of third party privacy seals and Web interface.

Initial Trust in Transaction:

p_t: Subjective probability (trust) in financial transaction processing in terms of whether the transaction is insured by the bank, credit card service provider or by some third party trusted party, for example, PayPal. This is based on alternate dispute resolution, presence of third party security seals and transaction insurance from trusted credit card service provider.

Since the user may assign high value to business performance in the case of some transactions and/or may trust few e-businesses over others in the case of certain types of transactions each of these probabilities will have varied levels of impact on the user's final decision. For example, Albert may trust a business to buy small household items but may feel a higher level of risk when buying a car or while ordering certain types of services from the same e-business. Therefore weights have to be assigned to each subjective probability in order to represent the importance of each initial trust dynamic in an online transaction. Previous studies support that the user may feel more confident to buy a known brand from an

e-business than an unknown brand (Ang et al., 2001). The model defined following subjective weights:

Business related weights:

w_b: Subjective weight of business performance
w_h: Subjective weight of business honesty and credibility

Transaction related weight:

w_t: Subjective weight of financial transaction processing

Each of these subjective weights value will range from zero to 10, where the value zero represents that the user's final decision is not effected by user's level of trust on a particular aspect of the transaction and the value ten reflects that the user's decision is very sensitive to a particular subjective probability. The defined range between zero to 10 is to make it easy for the user to evaluate and to assign subjective value. Higher weight value range would result in same final computed result as the weights are fractioned in the final equation, discussed later. By using subjective weights and probabilities the weighted subjective probabilities or user's weighted trust values are calculated as follows:

Weighted Trust in business:

Weighted trust in business performance (w_{bp}) = $p_b * w_p$

Weighted trust in business honesty (w_{bh}) = $p_h * w_h$

Other than the trust levels (subjective probability of each trust evaluation component) and weights in a transaction, the total uninsured part of investment (u_i) that is, the maximum dollar amount a user may lose in case of online fraud (financial loss) will have to be incorporated in order to logically estimate the user's possibility to commit a transaction. In order to balance out the absolute amount of investment the fraction of uninsured investment is used.

t_i = Total Investment
i_i = Insured Investment
u_i = Uninsured Investment (the maximum amount a user can lose) where $u_i = t_i - i_i$

Equation 1: Fraction of uninsured investment.

$$f_{ui} = \frac{u_i}{t_i}$$

Where f_{ui} is the fraction of uninsured investment.

The user's subjective probability on financial transaction processing 'p_t' and user's subjective weight of financial transaction processing 'w_t' only affects the uninsured portion of transaction i.e. 'f_{ui}'. Therefore the user's weighted subjective probability of losing uninsured investment has to be incorporated in order to evaluate user's total trust level in an online transaction. The following equation is composed in order to calculate ratio of weighted probability of losing uninsured investment:

Equation 2: Ratio of weighted probability of losing uninsured investment

$$rw_{ui} = \frac{(1 - p_t) * f_{ui} * w_t}{m_{wt}}$$

Where:
'$1 - p_t$' represents the user's subjective probability of losing the uninsured portion of investment
'm_{wt}' represents the maximum weight which can be assigned to financial transaction processing. 'm_{wt}' will always be 10

The weighted probabilities have to be grouped as they correspond to two different aspects of an electronic transaction, that is, user's trust in business and user's trust in transaction. Moreover weighted trust in business performance (w_{bp}) and weighted trust in business honesty (w_{bh}) have a positive affect on the user's decision while rw_{ui} has a negative impact on user's decision to commit an online transaction. The following equation is composed in order to calculate Worthy of Investment '*WoI*' value:

Equation 3: Worthy of this Investment.

$$WoI = \left(\left(\frac{w_{bp} + w_{bh}}{w_p + w_h} \right) - rw_{ui} \right) * 100$$

The computed value of *WoI* can be used by the user to get an indication of whether they should commit to a certain online transaction or not, on the basis of weights and subjective probabilities assigned by the user for each dynamic (aspect) of the transaction. Moreover, *WoI* can also be used to compare two or more e-businesses or online available options.

FUTURE TRENDS

Since a certain level of uncertainty or risk is prerequisite for trust to exist (Koller, 1988), this notion of trust suggests that when consumers willingly become vulnerable to a Web retailer, they consider both the characteristics of the Web retailer and the characteristics of the related technological infrastructure. Although the user's perceived risk of technological infrastructure can be significantly reduced by employing appropriate security technologies but effective

techniques to develop and enhance user's trust in Web retailer are yet to be developed.

In future, the use of Web 2.0 (O'Reilly, 2006) technologies and social networking applications, for example, use of friend of a friend (FOAF) (Brickley & Miller, 2005) network to connect the online users who are interested in contributing to and sharing online trust evaluations, will emerge. Such applications will enable the users to add friends and families information to the network on the basis of their personal trust and relationship. The users will use FOAF network to exchange contact information, views and trust evaluations of companies, businesses and parties with whom they have already completed transactions. Such information will be exchanged openly and freely within and across networks. Such applications will use FOAF, Atom (Nottingham & Sayre, 2005), RDF (W3C, 2004a) and RDF Schema—RDFS (W3C, 2004b). This will eventually result in an open and decentralised repository of users' trust evaluations, which will be used by potential customers as assistive guide before they commit e-transaction.

CONCLUSION

The chapter outlines the proposed models which can be used to predict user's trust in online shopping, enhance user's trust in online vendor and estimate the risk in an online transaction. The Cheung and Lee (2000) model proposes the external factors, user's propensity to trust and user's perceived trust worthiness in determining overall user's trust in online merchant. Ang et al. model focuses on merchant's ability to deliver; bands offered by the merchant, dispute resolution process and privacy statements as the determining factors of user's trust. The model proposed by Egger (2000) uses user's previous purchase knowledge, interface of the Web site and the contents of the Web site as the foremost factors which impact users' online trust. All these models suggest different aspects of online transaction which contribute to user's level of trust in a transaction. The proposed factors can be used by online businesses to concentrate on the aspects of online businesses which contribute most to user's online trust.

The model proposed by Mahmood (2006b) recommends that the user's trust in an electronic business can be further enhanced if the e-business portrays professionalism, credibility and honesty in its practices and portrays capability and performance to deliver the desired service or product. Mahmood presented the proposed trust evaluation model into a mathematical form. The model can be used by the e-business to recognize the areas which can be improved in order to increase users' level of online trust, to recognize and resolve the users' online trust reducing factors and also to map the user's desired outputs from the transaction to impacting factors.

The discussed models highlight excellent opportunities for improvements to the service providers. By focusing on areas which impact user's online trust the service and product providers can really make the online shopping experience easy and more practical for the users. In future, it is expected that online users' will use social networking to exchange and gather information on e-merchants in a decentralized open manner.

REFERENCES

Castelfranchi, C., & Falcone, R. (2000). Trust is much more than just subjective probability: Mental components and sources of trust. *National Research Council—Institute of Psychology*. Retrieved August 5, 2006 from http://www.istc.cnr.it/T3/publications/index.html

Ang, L., Dubelaar, C., & Lee, B.-C. (2001, June25-26). *To Trust or Not to Trust? A Model of Internet Trust From the Customer's Point of View*. In Proceedings of the 14th Bled Electronic Commerce Conference, Bled, Slovenia, 2001, (pp 40–52)._

Brickley, D., & Miller, L. (2005). FOAF Vocabulary Specification. *W3C* Retrieved September 16, 2006 from http://xmlns.com/FOAF/0.1/

Cheung, C., & Lee, M. (2000, August 3-5). *Trust in Internet shopping: A Proposed Model and Measurement Instrument.* In Proceedings of the 2000 Americas Conference on Information Systems (AMCIS), (pp. 681-689).

Egger, F.N. (2000, April 1-6).. Trust me, i'm an online vendor: Towards a model of trust for e-commerce system design*. In G. Szwillus & T. Turner (Eds.), *CHI2000 extended abstracts: Conference on human factors in computing system.,* The Hague (The Netherlands), (pp. 101-102).

Koller, M. (1988). Risk as a determinant of trust. *Basic and Applied Social Psychology, 9*(4),265–276.

Mahmood, O. (2006a). Modelling trust recognition and evaluation in electronic environment. *International Journal of Networking and Virtual Organisations (IJNVO). Special Issue on Trust for Virtual Organisations and Virtual Teams*.1741-5225.

Mahmood, O. (2006b). Trust: From sociology to electronic environment. *Journal of Information Technology Impact, 6*(3), 119-128.

Nottingham, M., & Sayre, R. (2005). The Atom Syndication Format. *Internet Society Taskforce (IEFT) FRC 4287*. Retrieved September 19, 2006 from http://tools.ietf.org/html/rfc4287

O'Reilly, T. (2006). Levels of the Game: The Hierarchy of Web 2.0 Applications O'Reilly radar. Retrieved December 3, 2006 from http://radar.oreilly.com/archives/2006/07/levels_of_the_game.html

Tan, Y-H., & Thoen, W. (2001). Toward a generic model of trust for electronic commerce. *International Journal of Electronic Commerce, 5*(2), 61–74.

W3C. (2004a). RDF Specification Development. *W3C.org* Retrieved November 8, 2006 from http://www.w3.org/RDF/

W3C. (2004b). RDF Vocabulary Description Language 1.0: RDF Schema. *W3C.org*. Retrieved November 8, 2006from http://www.w3.org/TR/rdf-schema/

KEY TERMS

Collective Trust Transfer: Collective trust transferring techniques rely on combined effort of several users, service providers and communities. Such techniques have much wider impact on potential customers, as they involve large number of contributing parties and are widely available.

Online Reputation: The online information regarding an e-business from the past direct or indirect experiences of a large body of users. It is the general opinion of the users toward a person, a group of people, or an organization in the cyberspace.

Privacy Statement: a statement posted on the company's or individual's Web site that explains the personal information being collected with or without a visitor's consent, the reasons it is being collected, and how the collected information will be used or shared. The privacy statement also states that how the information provided by someone else is used and shared.

Trust: The subjective estimation by which an individual, A, estimates about how likely another individual, B, performs a given task on which its welfare (interests) depends. It also consists of the elements of dependence, competence, disposition and fulfillment.

Trusted Referral: The information regarding a product or physical or online business, service or individual acquired from either the user's physical or online trusted network. It impacts the user's initial and subsequent levels of trust in an online business. The impact is directly related to the user's level of trust on the source in terms of source's credibility, honesty and ability.

Web 2.0: A network platform, enabling the utilization of distributed services such as social networking and communication tools. It is also referred as the architecture of participation.

An Overview of Wireless Network Concepts

Biju Issac
Swinburne University of Technology, Sarawak Campus, Malaysia

Wireless networks and the subsequent mobile communication are growing by leaps and bounds in the past years and the demand for connection without cables is certainly high. Nowadays, wireless networks are quite common and can be found on university campuses, corporate offices and in public places like hotels, airports, coffee shops and so forth. Not only are mobile devices getting smaller and cheaper, they are also becoming more efficient and powerful, capable of running applications and network services. This is causing the uncontrollable growth of mobile computing as we are witnessing today. Among the many number of applications and services that are executed by mobile devices, network and data services are in high demand. Brief descriptions of some selective wireless technologies that help mobile computing, like IEEE 802.11 networks (with infrastructure mode and ad-hoc mode), Bluetooth, HomeRF, WiMAX and cellular technologies are given below.

IEEE 802.11 INFRASTRUCTURE NETWORK

Wireless local area network (WLAN) which is also known as Wi-Fi (Wireless Fidelity) networks, requires an infrastructure network that could provide the services of accessing other networks, along with forwarding functions and medium access control. The Institute of Electrical and Electronics Engineers (IEEE) in 1997 initiated the first WLAN standard and they called it 802.11. But, 802.11 only supported a maximum bandwidth of 2 Mbps, which is quite slow for most applications. The IEEE 802.11 family consists of different standards. The initial standard was approved in 1997 and it backed wireless LAN Medium Access Control (MAC) and Physical layer (PHY) specifications that supported 1 Mbps and 2 Mbps data rate over the 2.4 GHz ISM band.

In a wireless infrastructure setup, there are two basic components, access points and wireless stations. An access point or base station functions as a bridge by connecting to a wired LAN (Local Area Network) through Ethernet cables. It receives data, buffers and transmits data between the wireless and the wired network infrastructure. A single access point supports on average 20 users and has a coverage varying from 20 meters in areas with obstacles like walls, stairways, elevators, and so forth, and up to 100 meters in areas where there is clear line of sight. The design of infrastructure-based wireless network is rather simpler as most of the network functionality lies within the access point. Transmission and reception of wireless communication can happen in different channels.

A building may require several access points to provide complete coverage and allow users to roam seamlessly between access points. A wireless network adapter connects users via an access point to the rest of the LAN. A wireless station can be using a PC card in a laptop, an ISA or PCI adapter in a desktop computer, or can be fully integrated within a handheld device. Security of a WLAN is of great concern with WEP (Wired Equivalent Privacy) encryption as static WEP keys could be easily recovered because of a design flaw (Stubblefield, Ionnidis, & Rubin, 2002). RADIUS Server authentication which uses EAP protocol (Extensible Authentication Protocol) with TKIP (Temporal Key Integrity Protocol) encryption is proposed as interim solution in WPA (Wi-Fi protected Access) standard, with AES encryption option being looked into as long term solution (Gast, 2002). Figure 1 shows a simple wireless network that uses RADIUS server authentication.

There are different wireless LAN technologies that the IEEE 802.11 standard supports in the unlicensed bands of 2.4 and 5 GHz. They share the same MAC (Medium Access Control) over two PHY layer specifications: Direct-Sequence

Figure 1. The wireless network in an organization showing mobile laptops

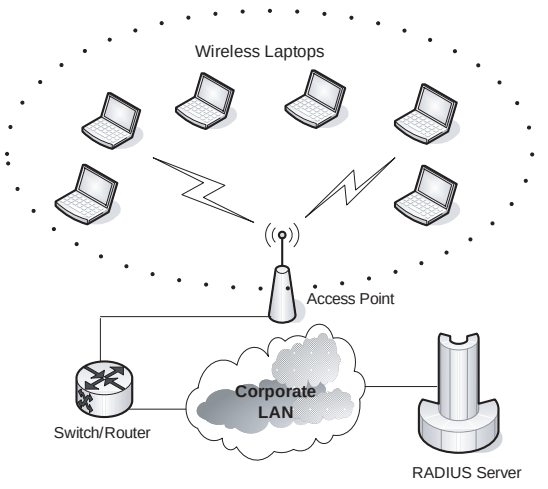


Table 1. Popular IEEE 802.11 comparisons

IEEE Standard	Maximum Speed	Frequency band	No. of nonoverlapping channels	Notes
802.11 (legacy)	1 Mbps to 2 Mbps	2.4 GHz	n/a	First standard (ratified in 1997). Uses FHSS and DSSS.
802.11a	54 Mbps	5 GHz	8 to 14 (or more in future)	Second standard (ratified in 1999). Uses OFDM.
802.11b	11 Mbps	2.4 GHz	3 (Channel 1,6 and 11)	Third and the most common standard (ratified in 1999). Uses DSSS.
802.11g	54 Mbps	2.4 GHz	3 (Channel 1,6 and 11)	Popular standard (ratified in 2003). Uses OFDM.

Spread Spectrum (DSSS) and Frequency-Hopping Spread Spectrum (FHSS) technologies. Infrared technology though supported, is not accepted by any manufacturer. Data rates of up to 2 Mbps were achieved initially by IEEE 802.11 systems operating at the 2.4 GHz band. Their wide acceptance initiated new versions and enhancements of the specification. The different extensions to the 802.11 standard use the radio frequency band differently. Popular 802.11 standards like 802.11a, 802.11b and 802.11g are listed in table 1.

As a strong and robust standard, 802.11i deals with the limitations of WEP encryption that was used with 802.11b and enhances the overall wireless security. The architecture uses 802.1x for authentication (with the use of EAP and an authentication server that uses 4-way handshake), includes improvements in key management and the Advanced Encryption Standard (AES) for encryption. Other 802.11 extensions include 802.11c that focuses on MAC bridges, 802.11d that focuses on worldwide use of WLAN with operation at different power levels, 802.11e that focuses on Quality of Service, 802.11f that focuses on access point interoperability and 802.11h that focuses on addressing interference problems when used with other communication equipments. Table 1 shows the comparison of the popular 802.11 standards (Held, 2003; Issac, Hamid, & Tan, 2006).

IEEE 802.11 AD HOC NETWORK

A wireless ad-hoc network is a network that uses wireless links where each node is willing to forward data to the other neighbouring nodes dynamically, based on the network connectivity. Types of wireless ad-hoc networks include Mobile ad-hoc networks (MANET), Wireless Sensor Networks (WSN) and Wireless Mesh Networks (WMN). A mobile ad-hoc network can be defined as a network of computer nodes that happens to be in proximity with each other, having no fixed infrastructure. A wireless sensor network (WSN) is a wireless network that makes use of distributed autonomous devices that uses sensors to measure or monitor environmental conditions like temperature, motion, sound, vibration, pressure and so forth, in a cooperative fashion. Wireless mesh networking (WMN) is mesh networking that is implemented on top of a wireless LAN in a decentralized (with no central server) way or centralized way (with a central server). Mesh networks are also extremely reliable with its redundant links, as each node is connected to several other nodes. If one node shuts down due to hardware errors or due to some other reason, its neighbours can easily find another route. Mesh networks can involve either fixed or mobile nodes.

Generally, in any ad-hoc network, each node can directly communicate with other nodes and so no access point or controlling station is needed. It is a self configuring network of routers along with associated hosts connected by wireless links. This union of network nodes or devices forms an arbitrary topology. This type of network provides great flexibility as it can be used for unplanned meetings, fast replacements of communication scenes far away from any infrastructure. Nodes or devices may look or rather search for target nodes that are out of vicinity by flooding the network with broadcasts packets that would be forwarded by each node. Wireless connections are even possible through multiple nodes forming a multihop ad-hoc network. Routing protocols then provide reliable connections even if nodes are moving around.

The routers are free to move randomly and organize themselves arbitrarily, making unpredictable changes in network's wireless topology. There is no need for access point and if one station working in ad-hoc mode is connected to wired network, stations forming ad-hoc network have a wireless access to Internet. IEEE 802.11 technology can be used to implement single-hop ad-hoc networks where the stations need to be in the same transmission radius to be able to communicate. But in multihop ad-hoc networking,

routing mechanisms can be enabled to extend the range of the ad-hoc network beyond the transmission radius of the single source station. Routing solutions for wired network doesn't apply to ad-hoc networks, because of its dynamic topology.

Some common problems that affect ad-hoc network performance are as follows: (1) stations cannot receive network frames outside the fuzzy wireless boundaries, (2) interference from other signals and (3) the channel has time varying and asymmetric propagation properties.

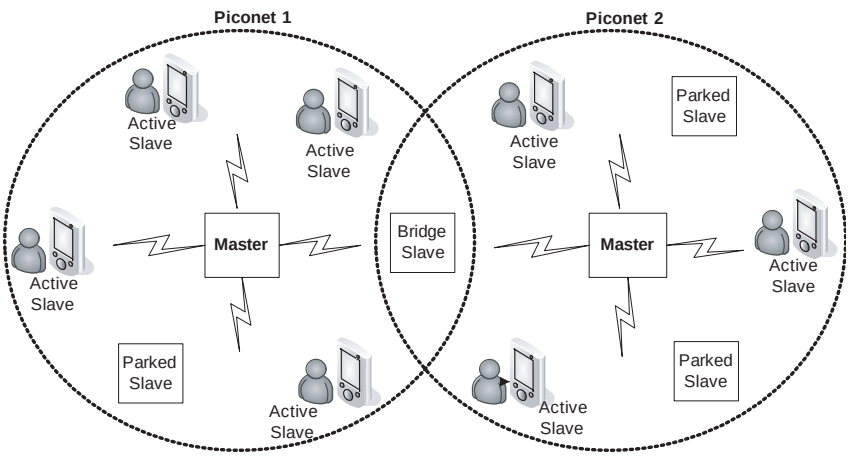
Ad-hoc routing protocols are getting popular with the increase in mobile computing. Ad-hoc networks include resource-starving devices, low bandwidth, high error rates and a topology that is continuously changing. Some of the design goals with ad-hoc routing protocols are minimal control overhead, minimal processing overhead, multihop routing capability, dynamic topology maintenance and loop prevention. The protocols should operate in a distributed manner. The nodes should operate either in proactive or reactive mode. Proactive protocols are table-based and maintain routes for the entire network within each node. The nodes must be fully aware of the changing topology. For topologies that are overtly dynamic, this approach can introduce a considerable overhead. Reactive or on-demand protocols trade off this overhead with increased delay. A route to destination is established when it is needed based on an initial discovery between the source and destination. Security of ad-hoc networks are a great concern with WEP (128 bits) encryption and 802.1x authentication offers some temporary solution (Basangi, Conti, Giordano, & Stojmenovic, 2004).

BLUETOOTH NETWORK

Bluetooth standard for wireless personal area networks is defined by the Bluetooth Special Interest Group (SIG) founded in 1998 by five companies; namely, Ericsson, Intel, IBM, Nokia and Toshiba. Currently, the SIG includes hundreds of other member companies. Bluetooth is a radio standard that is primarily designed for short range communication and for low power consumption, based on low-cost transceiver chips embedded in portable devices. It is named after Denmark's first king Harald Blaatand (or Bluetooth in English). Bluetooth provides a convenient way to connect and exchange information between devices such as personal digital assistants (PDAs), mobile phones, laptops, printers, digital cameras and so forth, via a secure, low cost, globally available short range radio frequency. The range of radio is from 10 meters to 100m (through optional amplifier). It aims to be a cable replacement technology and it defines three topologies: point-to-point, single cell (piconet) and multicell (scatternet). The basic unit of networking in Bluetooth is a piconet.

A device can be either a slave or master in piconet and can thus be part of more than one piconet. This network overlapping is called scatternet, as shown in Figure 2. As one of the devices in the piconet can act as a master and the other as slaves, the master decides the hopping pattern and the slaves have to synchronize to this pattern. Each piconet has a unique hopping pattern. Two other devices can also exist in the piconet, namely parked devices and standby devices. Parked devices which can be around 200 numbers maximum, cannot actively participate in the piconet, but can be reactivated within some milliseconds delay. Standby devices do not participate in the piconet. Each piconet has

Figure 2. Two piconets connected to form a scatternet



exactly one master and up to seven slaves. The frequency hopping (FH) used in Bluetooth provides resistance to interference and multiple path effects and provides multiple accesses from colocated devices in different piconets. The first step in forming a piconet involves the master sending its clock and device ID, which is a 48-bit unique identifier. After synchronizing its internal clock with the master's clock, a device may join a piconet whereby it is assigned a 3-bit active member address (AMA). The parked devices use an 8-bit parked member address (PMA). Devices on standby do not use any address (Schiller, 2003).

The Bluetooth version 2.1 specification includes the following features: *extended inquiry respons*e that provides more information during the inquiry process that would allow better filtering of devices before connection, *sniff subrating* that reduces the power consumption when devices are in the sniff low-power mode, *encryption pause resume* that enables an encryption key to be refreshed, enabling much stronger encryption and *secure simple pairing* that radically improves the pairing process securely for Bluetooth devices. Bluetooth provides support for three general wireless application areas like data and voice communication, cable replacement and ad-hoc networking.

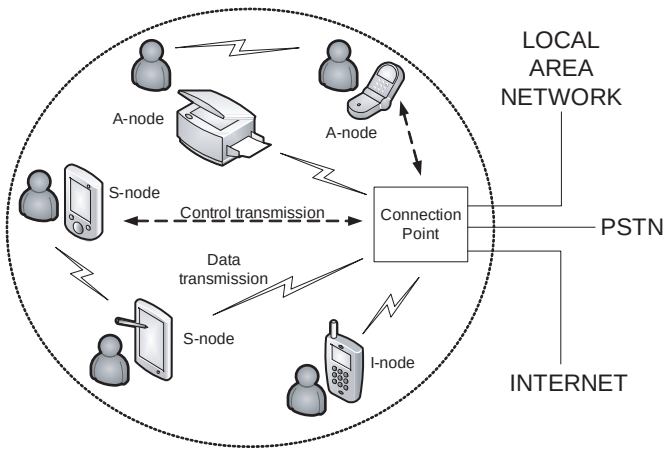
The security mechanism in Bluetooth uses a 48-bit address defined by IEEE, a 128-bit authentication key, a 128-bit symmetric encryption key and a generated random number. In order to add more value to this prospective radio technology, two potential aspects should be looked into to strengthen the Bluetooth platform for future applications; they are Bluetooth-enabled secure access to wide area networks and the interference issues with other devices operating within the same frequency band. In recent years, the research communities have been looking into the possibility to expand the coverage of Bluetooth networks for greater coverage and applications.

HomeRF NETWORK

HomeRF 2.0 specification defines a common interface that supports wireless voice and data networking within a home. It was developed by the HomeRF Working Group, a consortium of mobile wireless companies that included Siemens, Proxim, National Semiconductor, Motorola and more than 100 other companies. It supports three types of data services: asynchronous data service (that provides best effort service), priority asynchronous data service (that provides priority service) and isochronous data service (that provides support for applications with strict QoS requirements). HomeRF created a new class of mobile consumer devices using personal computers and the Internet. It is also referred to as the last 50 meters access. The availability of high speed last mile access technologies such as xDSL, cable modem and ISDN has also created a high speed demand on home networking. The Shared Wireless Access Protocol (SWAP) that has been developed by the Home Radio Frequency Working Group (HomeRF WG) was launched in March 1998 as a single specification for consumer devices interoperability (HomeRF, n.d.; Issac, Chee, & Mohammed, 2005).

HomeRF 2.0 delivers Ethernet equivalent bandwidth to support wider wireless voice and data networking applications at home. It allows 800 kbps for isochronous voice and a maximum of 10 Mbps for asynchronous data transfer. The standard categorizes the HomeRF devices into 3 types: Asynchronous data device (also called A-node), Streaming data device (S-node) and Isochronous data device (I-node).

Figure 3. Managed HomeRF network



The Connection Point (CP) provides service management for the A-node, S-node and I-node. CP supports A-nodes with power savings and data access to Internet or other network computers, S-nodes with session setup and priority channel access to other S-nodes and finally I-nodes with connection setup, resource allocation and connectivity to PSTN. The HomeRF standard defines two kinds of network topologies: Ad-hoc network and Managed Network. An ad-hoc network is a distributed network in which the devices can make a peer-to-peer communication that includes only A-nodes without any CP. The Managed Network as shown in Figure 3 is a network managed by the CP and the devices can either communicate peer-to-peer (data services) or can establish communication through the CP (voice services) depending on the type of HomeRF devices used. HomeRF 2.0 supports 128 bit encryption so that all the data traveling across the radio waves is encrypted. The major drawback of this implementation is the lack of a centralized key management for the network. But this is not a concern for home users because it is a common assumption that the manual key management needs to be done only for a few devices in a home network. Similar to some of the competing technologies such as Bluetooth and WiFi, HomeRF operates in the license free 2.4GHz frequency band and utilizes frequency hopping spread spectrum technology to achieve a secure and robust communication (Dean, 2004; Ganz, Ganz & Wongthavarawat, 2004).

WiMAX: THE BROAD BAND WIRELESS NETWORK

WiMAX (Worldwide Interoperability for Microwave Access) forms the set of standards developed by IEEE 802.16 working group in 2001 that defines a broad band wireless network. 802.16 networks are geared toward point to multipoint (PMP) communication topology in which a central base station communicates with a number of subscriber stations and all the communication will have to go through the base station, as in Figure 4. As the wavelengths are short in the 10 to 66 GHz frequency bands, the standard requires line of sight between the communicating points. It supports data rates in excess of 120 Mbps (theoretically).

There are two types of WiMAX, namely fixed and mobile, typified by 802.16d and 802.16e. 802.16d provides 2 Mbps or more data rates within a distance of up to 5 miles and its aim is to provide wireless connection between one central base station and a set of fixed networks, like connecting a set of offices to a central office without WAN links. 802.16e is intended to help mobile users. It provides multiple channels with an effective data rate of 5 Mbps with a distance limit of up to 6 miles (Ganz, Ganz, & Wongthavarawat, 2004).

2G/2.5G/3G CELLULAR NETWORKS

Mobile phones that use cellular technologies have traversed three generations with three different technologies: analog voice, digital voice and digital voice with data. The cells are modeled as hexagons (or roughly circular), as in Figure 5, with transceiver towers in them. A mobile device can move from one cell to the other cell, keeping connected. The first generation (1G) mobile phones used analog voice that used AMPS (Advanced Mobile Phone System). The second generation (2G) used digital voice and had used D-AMPS (Digital AMPS), GSM (Global System for Mobile Communication) and CDMA (Code Division Multiple Access) systems. The third generation (3G) uses W-CDMA (Wideband CDMA)

Figure 4. Point to multipoint network topology in WiMAX

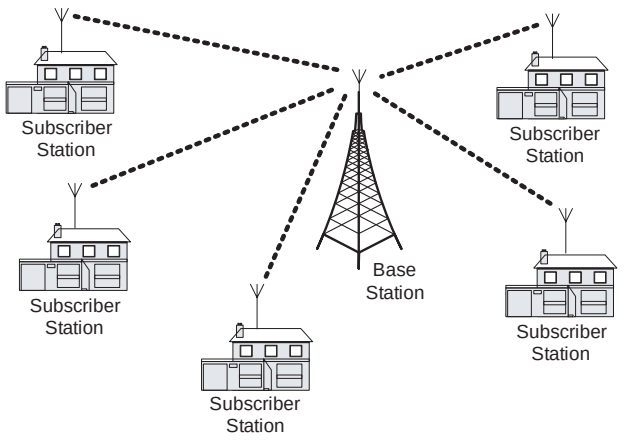
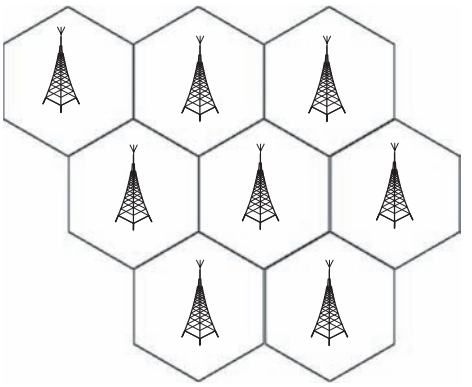


Figure 5. Cell sites with transceiver towers



and CDMA2000. GSM is the dominant 2G mobile system worldwide primarily due to its roaming capability, and due to its early standardization and deployment by European countries. GSM uses TDMA (Time Division Multiple Access) technology in three different bands: 900MHz & 1800 MHz in most countries, and 1900MHz in USA. GSM data rate is around 9.6 Kbps. CDMA IS-95a (CDMAone) is also a 2G cellular wireless technology and the first commercial CDMA network was mainly deployed in USA and South Korea and was developed by Qualcomm. IS-95a supports data rates of up to 14.4 Kbps and uses GPS satellites for network system timing.

GPRS (Global Packet Radio Service) is a 2.5G technology based on GSM. 2.5G is a term used for a 2G technology that has been modified to give it the packet data qualities of 3G. It has improved data rates but it is still not near the maximum rate of 2Mbps of a 3G system. GPRS data rate is 20 to 40 Kbps (with 144 Kbps theoretical maximum). CDMA IS-95b is also a 2.5G descendant of IS-95a. It has low deployment costs and theoretically provides data rates up to 115Kbps, but providing less than 64 Kbps in practice.

EDGE (Enhanced Data for Global Evolution) is also a 2.5G technology, but it is an enhanced version of GPRS. It uses enhanced modulation to include more data into the available radio bandwidth. Its data rate is much more comparable to 3G data rates.

Wideband-CDMA network is a 3G system that was proposed by Ericsson and it uses direct sequence spread spectrum modulation. It is designed to interwork with GSM networks and runs in 5 MHz bandwidth. The other 3G technology is CDMA2000 proposed by Qualcomm. It is backward compatible with IS-95, but doesn't work with GSM. CDMA2000 also uses direct sequence spread spectrum and uses 5MHz bandwidth. Work is in progress toward 4G systems that promises much higher bandwidth, ubiquity, seamless integration with wired network and high service quality for multimedia (Tanenbaum, 2004).

COMPARISION OF WIRELESS TECHNOLOGIES

Table 2 compares the features of different wireless technologies like modulation, frequency range, distance range, network type and data rate (Ferro & Potorti, 2004).

CONCLUSION

Brief descriptions of popular wireless technologies were presented through the discussions done. The world is increasingly becoming mobile and wireless networking is getting greater focus in the recent past, especially with hardware technology breakthroughs and price drops. As wireless networks facilitate mobile computing, the future of the above mentioned technologies are being watched with eagerness. Mobility when combined with computing power makes a deadly combination and that's why relentless research is going on to make wireless technologies even better. The convergence of all these wireless technologies to form one single network is the greatest challenge ahead, as they all

Table 2. Wireless technology comparison

Standard / Feature	802.11a/b/g (Wi-Fi)	802.11n (Wi-Fi)	Bluetooth	HomeRF	802.16 (WiMax)	2G/2.5G/3G (Cellular)
Modulation	OFDM or DSSS	DSSS	Adaptive FHSS	FHSS (2 or 4 Level FSK)	QPSK, 16-QAM, 64-QAM	D-AMPS /CDMA/ GSM/
Operating Frequency	2.4 GHz (b/g) and 5 GHz (a)	2.4 GHz or 5 GHz	2.4 GHz	2.4 GHz	10-66 GHz	869-894 MHz
Distance Range	~30 to 110m	~70m to 160m	~10m	~150 feet	~2 km to ~10 km	Cell range of 1 to 5 miles (approx.)
Network Type	IP and peer-to-peer	IP and peer-to-peer	peer-to-peer (point to multi-point)	peer-to-peer	IP (point to multi-point)	IP
Data Rate	2 Mbps to 54 Mbps	200Mbps	721Kbps to 3 Mbps	800kbps to 10Mbps	5 Mbps to 70 Mbps	>128kbps to 2Mbps

use different networking technologies and that could take us to a world where seamless and ubiquitous computing will become a definite reality and not a future dream.

REFERENCES

Basangi, S., Conti, M, Giordano, S., & Stojmenovic, I. (2004). *Mobile ad-hoc networking*. Wiley InterScience.

Dean, T. (2004). *Network+ Guide to networks* (3rd ed.). Course Technology.

Ferro, E., & Potorti, F. (2005). Bluetooth and Wi-Fi wireless protocols: A survey and a comparison. *IEEE Wireless Communications Magazine*, 12-26.

Ganz, A., Ganz, Z., & Wongthavarawat, K. (2004). *Multimedia wireless networks–technologies, standards and QoS*. NJ, USA: Prentice Hall.

Gast, M.S. (2002). *802.11 wireless networks–the definitive guide*. CA: O'Reilly.

Held, G. (2003). *Securing wireless LANs*. Sussex: John Wiley & Sons.

HomeRF. (n.d.). *Wikipedia, the free encyclopedia*. Retrieved December 13, 2007, from http://www.reference.com/browse/wiki/HomeRF

Issac, B., Chee, V.K.M., & Mohammed, L.A. (2005). Security considerations in the design of wireless networks. In *Proceedings of the International Conference on Wireless Networking and Mobile Computing (ICWNMC 2005),* Chennai, India, (pp. 136-141).

Issac, B., Hamid, K., & Tan, C.E. (2006). Analysis of single and mixed 802.11 networks and mobility architecture. In *Proceedings of the International Conference on Computing and Informatics (ICOCI 2006),* Malaysia.

Schiller, J. (2003). *Mobile communications* (pp. 269-297). Harlow: Pearson Education.

Stubblefield, A., Ionnidis, J., & Rubin, A.D. (2002). Using the Fluhrer, Mantin, and Shamir attack to break WEP. In *Proceedings of the Network and Distributed Systems Security Symposium (NDSS 2002),* (pp. 17-22).

Tanenbaum, A.S. (2004). *Computer networks* (4th ed.). NJ, USA: Prentice Hall.

KEY TERMS

Access Point: The central or master device through which an infrastructure wireless node makes a connection to the local area network. It acts more like a bridge between wireless node and LAN.

Bluetooth: It is an industrial specification and standard for wireless personal area networks (W-PANs).

Direct Sequence Spread Spectrum: A form of spread spectrum in which each bit in the original signal is represented by multiple bits in the transmitted signal, using a spreading code.

Encryption: To convert plain text or data into unreadable form by means of a reversible mathematical computation.

Frequency Hopping Spread Spectrum: It is a spread-spectrum method of transmitting radio signals by rapidly switching a carrier among many frequency channels, using a pseudorandom sequence known to both transmitter and receiver.

HomeRF: Short form for "home radio frequency." It is designed specifically for wireless networks in homes, in contrast to 802.11, which was created for use in businesses.

Medium Access Control (MAC): For a communication network, the method of determining which station has access to the transmission medium at any time.

Piconet: A small network of communication devices connected in an ad-hoc fashion using Bluetooth technology.

Scatternet: A scatternet is collection of piconets connected through sharing devices.

WiMax: Short form for Worldwide Interoperability for Microwave Access, to provide wireless data over long distances from point to point links to full mobile cellular type access.

Wireless Transmission: Electromagnetic transmission through air, vacuum or water by means of antenna.

OWL: Web Ontology Language

Adélia Gouveia
University of Madeira, Portugal

Jorge Cardoso
SAP Research CEC Dresden, Germany
University of Madeira, Portugal

INTRODUCTION

The World Wide Web (WWW) emerged in 1989, developed by Tim Berners-Lee who proposed to build a system for sharing information among physicists of the CERN (*Conseil Européen pour la Recherche Nucléaire*), the world's largest particle physics laboratory.

Currently, the WWW is primarily composed of documents written in HTML (hyper text markup language), a language that is useful for visual presentation (Cardoso & Sheth, 2005). HTML is a set of "markup" symbols contained in a Web page intended for display on a Web browser. Most of the information on the Web is designed only for human consumption. Humans can read Web pages and understand them, but their inherent meaning is not shown in a way that allows their interpretation by computers (Cardoso & Sheth, 2006).

Since the visual Web does not allow computers to understand the meaning of Web pages (Cardoso, 2007), the W3C (World Wide Web Consortium) started to work on a concept of the Semantic Web with the objective of developing approaches and solutions for data integration and interoperability purpose. The goal was to develop ways to allow computers to understand Web information.

The aim of this chapter is to present the Web ontology language (OWL) which can be used to develop Semantic Web applications that understand information and data on the Web. This language was proposed by the W3C and was designed for publishing, sharing data and automating data understood by computers using ontologies. To fully comprehend OWL we need first to study its origin and the basic blocks of the language. Therefore, we will start by briefly introducing XML (extensible markup language), RDF (resource description framework), and RDF Schema (RDFS). These concepts are important since OWL is written in XML and is an extension of RDF and RDFS.

BACKGROUND

Everyday, the Web becomes more attractive as an information sharing infrastructure. However, the vast quantity of data made available (for example, Google indexes more than 13 billion pages) makes it difficult to find and access the information required by the wide diversity of users. This limitation arises because most documents on the Web are written in HTML (HTML, 2007), a language that is useful for visual presentation but which is semantically limited. As a result, humans can read and understand HTML Web pages, but the contents of Web pages are not defined in a way that computers can understand them. If computers are not able to understand the content of Web pages it becomes impossible to develop sophisticated solutions to enable the interoperability and integration between systems and applications.

The aim of the Semantic Web is to make the information on the Web understandable and useful to computer applications and in addition to humans. *"The Semantic Web is an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation"* (Berners-Lee et al., 2001). The Semantic Web is a vision for the future of the Web, in which information is given explicit meaning, making it easier for machines to automatically process and integrate the information available on the Web.

One of the corner stones of the Semantic Web is the OWL. OWL provides a language that can be used by/on applications that need to understand the meaning of information instead of just parsing data for display purposes. Nowadays, several projects already rely on semantics to implement their applications. Example include semantic wikis (Campanini et al., 2004), social networks (Ding, et al., 2005), semantic blogs (Cayzer & Shabajee, 2003), and Semantic Web services (McIlraith et al., 2001),

THE SEMANTIC WEB STACK

The Semantic Web identifies a set of technologies and standards which form the basic building blocks of an infrastructure

that supports the vision of the Web associated with meaning. Figure 1 illustrates the different parts of the Semantic Web architecture. It starts with the foundation of URI (universal resource identifier) and Unicode. URI is a formatted string that serves as a means of identifying abstract or physical resources. For example, http://dme.uma.pt/jcardoso/index.htm identifies the location from where a Web page can be retrieved and urn:isbn:3-540-24328-3 identifies a book using its ISBN. Unicode provides a unique number for every character, independent of the underlying platform, program, or language.

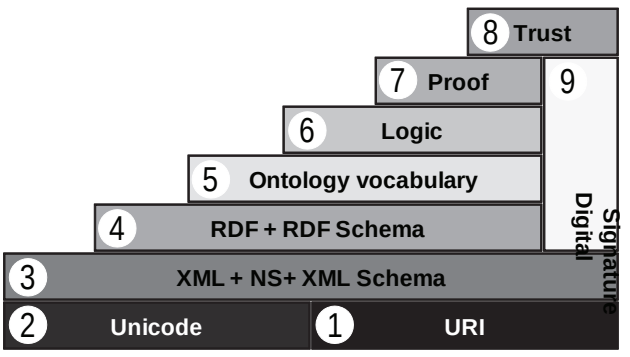
Directly above URI and Unicode we find the syntactic interoperability layer in the form of XML, which in turn underlies RDF and RDFS. Web ontology languages are built on top of RDF and RDFS. The last three layers are logic, proof, and trust, which have not been significantly explored. Some of the layers rely on the digital signature component to ensure security.

In the following sections we briefly describe the most relevant layers (XML, RDF, and RDFS). While the notions presented have been simplified, they give a reasonable conceptualization of the various components of the Semantic Web.

XML

The extensible markup language (XML) (Decker et al., 2000; XML, 2007) was originally pictured as a language for defining new document formats for the WWW. An important feature of this language is the separation of content from presentation, which makes it easier to select and/or reformat the data. SGML (standard generalized markup language) and XML are text-based formats that provide mechanisms for describing document structures using markup tags (words surrounded by '<' and '>'). Both HTML and XML representations use tags such as <h1> or <name>, and information between those tags, referred to as the content of the tag. However, there are significant differences between HTML and XML.

Figure 1. Semantic Web layered architecture (Berners-Lee et al., 2001)



XML is case sensitive while HTML is not. This means that in XML the start tags <Table> and <table> are different, while in HTML they are the same. Another difference is that HTML has predefined elements and attributes whose behavior is well specified, while XML does not. Instead, users can create their own XML vocabularies that are specific to their application or business' needs.

The following structure shows an example of an XML document identifying a 'Contact' resource. The document includes various metadata markup tags, such as <first_name>, <last_name>, and <e-mail>, which provides various details about a contact.

```
<Contact contact_id="1234">
      <first_name> Jorge </first_name>
      <last_name> Cardoso </last_name>
      <organization> University of Madeira </organization>
      <email> cardoso@uma.pt </email>
      <phone> +51 291 705 156 </phone>
</Contact>
```

While XML has gained much of the world's awareness, it is significant to identify that XML is simply a way of standardizing data formats. But from the point of view of semantic interoperability, XML has restrictions. One important characteristic is that there is no way to recognize the semantics of a particular domain because XML aims at a document structure and enforces no common interpretation of the data. Although XML is simply a data-format standard, it is part of a set of technologies that constitute the foundations of the Semantic Web.

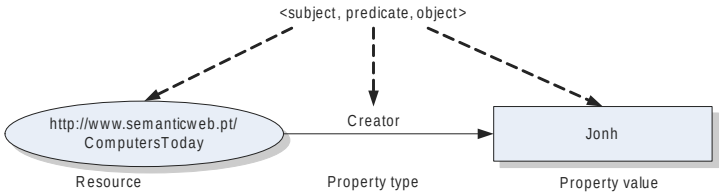
RDF

Resource description framework (RDF) (RDF, 2002), was developed by the W3C to provide a common way to describe information so it could be read and understood by computer applications. RDF was designed using XML as the underlying syntax language. RDF provides a model for describing resources on the Web. A resource is an element (document, Web page, printer, user, etc.) on the Web that is uniquely identifiable by a URI. The RDF model is based upon the idea of making statements about resources in the form of a subject-predicate-object expression, a 'triple' in RDF terminology.

- Subject is the resource, that is, the thing that is being described;
- Predicates are aspects about a resource, and expresses the relationship between the subject and the object;
- Object is the value that is assigned to the predicate.

RDF has a very limited set of syntactic constructs, no other constructs except for triples is allowed. Every RDF

Figure 2. RDF graph



document is equivalent to an unordered set of triples. Let us write a RDF triple that describes the following statement:

"The creator of the page named ComputersToday is John."

In this example, 'http://www.semanticweb.pt/ComputersToday' is a resource, and it has a property, 'Creator,' with the value 'John.' The resulting RDF statement is:

st = (http://www.semanticweb.pt/ComputersToday, Creator, John)

The statement can also be graphically represented as illustrated in Figure 2.

One way to represent the statement in Figure 2 using RDF language is the following:

```
<? xml version="1.0" ?>
<RDF xmlns = "http://w3.org/TR/1999/PR-rdf-syntax-19990105#"
      xmlns:DC = «http://dublincore.org/2003/03/24/dces#»>

      <Description about =
         "http://www.semanticweb.pt/ComputersToday">
      <DC:Creator>John</DC:Creator>
   </Description>
</RDF>
```

The first lines of this example use namespaces to explicitly define the meaning of the notions that are used. The first namespace xmlns:rdf="http://w3.org/TR/1999/PR-rdf-syntax-19990105#" refers to the document describing the syntax of RDF. The second namespace http://dublincore.org/2003/03/24/dces# refers to the description of the Dublin Core (DC) (DC, 2005), a basic ontology about authors and publications.

RDF SCHEMA

RDF schema (RDFS) (XMLSchema, 2005) is technologically more advanced when compared to RDF. RDFS describes the resources with classes, properties, and values. RDFS associates the resources in classes and states the relations between these classes, or declares properties and specifies the domain and range of these properties. RDFS' specification consists of some basic classes and properties that can be extended to any given domain.

Classes in RDFS are much like classes in object oriented programming languages. These allow resources to be defined as instances of classes, and subclasses of classes. Properties can be seen as attributes that are used to describe the resources by assigning values to them. RDF is used to declare a property and RDFS can extend this capability by defining the domain and the range of that property (however, RDFS has some limitations but these have been resolved with the introduction of OWL).

THE WEB ONTOLOGY LANGUAGE

The Web ontology language (OWL) (OWL, 2004) is one of the most important ontology languages. It enables the interoperability of applications and allows computers to understand the Web's content. In this respect it is more expressive than XML, RDF or RDF Schema due to providing additional vocabulary along with formal semantics.

OWL Flavors

There are three OWL sublanguages: OWL Lite, OWL DL, and OWL Full. An important feature of each sublanguage is its expressiveness. OWL Lite is the least expressive and the OWL Full is the most expressive sublanguage. OWL DL is more expressive that OWL Lite but less expressive than OWL Full. In other words, this entails that every legal OWL Lite ontology is a legal OWL DL ontology; every legal OWL DL ontology is a legal OWL Full ontology; every valid OWL Lite conclusion is a valid OWL DL conclusion; and every valid OWL DL conclusion is a valid OWL Full conclusion.

OWL Full is the most expressive of the OWL sublanguages and it uses the entire OWL language primitives. It is intended to be used in situations where very high expressiveness is more important than being able to guarantee the

Figure 3. OWL sublanguages

decidability or computational completeness of the language. This sublanguage is meant for users who want maximum expressiveness and the syntactic freedom of RDF, but with no computational guarantees.

OWL DL is a sublanguage of OWL Full that restricts the application of OWL and RDF constructors. OWL DL (DL stands for description logics) is not compatible with RDF, in the same way that not every RDF document is a legal OWL DL document, although every legal OWL DL document is a legal RDF document. This sublanguage supports those users who want the maximum expressiveness without losing computational completeness and decidability.

OWL Lite is syntactically the simplest sublanguage. It is intended to be used in situations where only a simple class hierarchy and constraints are needed. This sublanguage supports those users primarily needing a simple classification hierarchy and constraint features.

Note that every OWL Lite ontology or conclusion is a legal OWL DL ontology or conclusion, but not the inverse, and so on for OWL DL and OWL Full, as showed in Figure 3:

The choice between OWL Lite and OWL DL may be based upon whether the simple constructs of OWL Lite are sufficient or not. The choice between OWL DL and OWL Full may be based upon whether it is important to be able to carry out automated reasoning on the ontology or whether it is important to be able to use highly expressive and powerful modeling facilities.

OWL Syntax

In this section we describe the syntax of OWL. We illustrate step-by-step how to build an ontology using OWL. We also explain how to define the heather of ontology, its classes, properties and relationships. After reading this section the reader should be able to recognize an ontology written in OWL and identify some of its components.

Header

The first element in an OWL document is an rtf:RDF element which specifies a set of XML namespace's declarations that provide a means to unambiguously interpret identifiers and make the rest of the ontology presentation much more readable. For example,

```
<rdf:RDF
    xmlns="http://apus.uma.pt/~adelia/RUD.owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xml:base="http://apus.uma.pt/~adelia/RUD.owl">
```

A namespace is composed by: reserved XML attribute xmlns, a prefix that identify the namespace and the value.

Information Version

After the namespace declaration, an OWL document specifies a collection of assertions that are grouped under an owl: ontology element and offers details about the ontology:

- **owl:versionInfo:** Provides information about the current ontology.
- **owl:priorVersion:** Indicates an earlier version of the current one.
- **owl:backwardCompatibleWith:** Contains a reference to an ontology that is a prior version of the containing ontology that is backward compatible with it.
- **owl:incompatibleWith:** Indicates that the containing ontology is not backward compatible with the referenced ontology.
- **owl:imports:** Only this assertion has a formal meaning to the ontology and represents a set of other ontologies that are considered to be part of the current ontology. Note that owl:imports is a transitive property because if ontology A imports ontology B, and ontology B imports ontology C, then ontology A also imports ontology C.

The following is a simple example:

```
<rdf:RDF>
...
<owl:Ontology rdf:about="">
    <rdfs:comment> University Ontology </rdfs:comment>
    <owl:versionInfo> v.1 2006-9-05 </owl:versionInfo>
    <owl:priorVersion>
    <owl:Ontology rdf:about = "http://apus.uma.pt/~adelia/RUD.owl"/>
    </owl:priorVersion>
    <rdfs:label> University Ontology </rdfs:label>
</owl:Ontology>
...
</rdf:RDF>
```

Class Element

Classes are a collection of individuals, a way of describing part of the world. They are defined in an OWL document with the owl:Class element. For example, the class "Teacher" can be define as follows,

```
<owl:Class rdf:ID="Teacher">
    <rdfs:subClassOf rdf:resource="#Person"/>
</owl:Class>
```

Note that the rdf:ID element defines the name of the class. If we want to make reference to a class we use the rdf: resource element. An OWL ontology can represent the hierarchy between classes using the element owl:subClassOf. For example, the class "Teacher" is a subclass of "Person."

Between the two classes it is possible to establish relations using owl:equivalentClass and owl:disjoinWith elements. The assertion owl:equivalentClass when applied to two classes A and B, represents that class A has the same individuals as class B. For example, the class "faculty" is equivalent to the "academicStaffMember" class:

```
<owl:Class rdf:ID="faculty">
  <owl:equivalentClass rdf:resource="#academicStaffMember"/>
</owl:Class>
```

The owl:disjoinWith element applied on two classes A and B suggest that class A and B disjoin, that is, if an instance is member of class A it cannot be an instance of class B. For example, a "Full Professor" cannot be an "Associate Professor" at the same time.

```
<owl:Class rdf:about="#AssociateProfessor">
    <owl:disjoinWith rdf:resource="#FullProfessor"/>
</owl:Class>
```

Complex Class

Another way to create classes in OWL is to combine simple classes using Boolean operators (union, intersection, and complement) and create complex classes. The members of the class are completely specified by the Boolean operators. The owl:unionOf element applied on classes A and B creates a new class that contains all members from class A and B. For example, the combination of the class "staff members" and the class "student" create the new class "peopleAtUni."

```
<owl:Class rdf:ID="peopleAtUni">
    <owl:unionOf rdf:parseType="Collection">
        <owl:Class rdf:about="#staffMember"/>
        <owl:Class rdf:about="#student"/>
    </owl:unionOf>
</owl:Class>
```

The owl:intersectionOf element creates a new class from the two classes A and B which has elements that were both in class A and class B, which follows as,

```
<owl:Class rdf:ID="facultyInDME">
  <owl:intersectionOf rdf:parseType="Collection">
    <owl:Class rdf:about="#faculty"/>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#belongsTo"/>
      <owl:hasValue rdf:resource="#DMEDepartment"/>
    </owl:Restriction>
  </owl:intersectionOf>
</owl:Class>
```

The individuals of the new class created in this example are those individuals that are members of both the classes "faculty" and the anonymous class created by the restriction on the property "belongTo."

The owl:complement element selects all individuals from the domain that do not belong to a certain class,

```
<owl:Class rdf:about="#course">
  <rdfs:subClassOf>
    <owl:Class>
      <owl:complementOf rdf:resource="#staffMember"/>
    </owl:Class>
  </rdfs:subClassOf>
</owl:Class>
```

In this example, the class "course" has as its members all individuals that do not belong to the" staffMember" class.

Property

Properties let us describe a kind of relationship between members of classes. In an OWL document two types of properties are distinguished:

- Object properties which relate objects to other objects, i.e. instances of a class with instances of another class. In the next example the object property "isTaughtBy" relates the class "course" with the class "academicsStaffMember." This means that a "course" "isTaughtBy" an instance of the "academicStaffMember" class.

  ```
  <owl:ObjectProperty rdf:ID="isTaughtBy">
      <owl:domain rdf:resource="#course"/>
      <owl:range rdf:resource="#academicStaffMember"/>
  </owl:ObjectProperty>
  ```

- Datatype property which relates objects to data type values. OWL does not have predefined data types, but it allows one to use the XML Schema data types. In following example, the year in which a tourist was born is specified using the "http://www.w3.org/2001/XLMSchema#nonNegativeInteger" data type from the XML Shema.

```
<owl:DatatypeProperty rdf:ID="ageYear">
    <rdfs:domain rdf:resource="#Person"/>
    <rdfs:range rdf:resource=
http://www.w3.org/2001/XLMSchema#nonNegativeInteger"/>
</owl:DatatypeProperty>
```

Note that both kinds of properties can use the rdfs:domain and rdfs:range element to restrict the relation.

Property Restrictions

More elaborate boundaries can be made by applying restrictions to a property, this results in the subclasses of individuals that satisfy that condition. There are two kinds of restrictions: values constraints and cardinality constraints. Examples of values constraints include owl:allValuesFrom, owl:someValuesFrom, and owl:hasValues.

owl:allValuesFrom: Defines the set of individuals, for which all the values of the restricted property are instance of a certain class:

```
<owl:Class rdf:about="#firstYearCourse">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#isTaughtBy"/>
      <owl:allValuesFrom rdf:resource="#professor"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

In this example, the individuals that are members of the class "firstYearCourse" are all the courses that have the property "isTaughtBy" assigned to a "professor"

owl:someValuesFrom: Defines the set of individuals that have at least one relation with an instance of a certain class, for example:

```
<owl:Class rdf:about="#academicStaffMember">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#teaches"/>
      <owl:someValuesFrom rdf:resource="#undergraduateCourse"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

owl:hasValues: Defines a set of individuals for which the value of the restricted property is equal to a certain instance. For example the individuals of the class "mathCourse" can be characterized as those that are taught by the professor "949352:"

```
<owl:Class rdf:about="#mathCourse">
    <rdfs:subClassOf>
        <owl:Restriction>
            <owl:onProperty rdf:resource="#isTaughtBy"/>
            <owl:hasValues rdf:resource="#949352"/>
        </owl:Restriction>
    </rdfs:subClassOf>
</owl:Class>
```

Cardinality constraints point out how many times the property can be used on an instance. Examples include owl: maxCardinality, owl:minCardinality, and owl:cardinality.

owl:maxCardinality: Defines the set of individuals that have at the most N distinct values of the property concerned. For example, we can specify that the class "department" has at the most 30 members:

```
<owl:Class rdf:about="#department">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#hasMember"/>
        <owl:maxCardinality rdf:datatype="&xsd;nonNegative
        Integer">
        30
      </owl:maxCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

owl:minCardinality: Defines the set of individuals that have at least N distinct values of the property concerned. For example, a course must be taught at least by one teacher. In OWL it is defined as follows:

```
<owl:Class rdf:about="#course">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#isTaughtBy"/>
        <owl:minCardinality rdf:datatype="&xsd;nonNegative
        Integer">
        1
      </owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

owl:cardinality: Defines the set of individuals that have an exact number of distinct values of the property concerned. This element is used to specify a precise number, that is, to express that a property has a minimum cardinality which is equal to the maximum cardinality.

Properties' Characteristics

Properties' characteristics add more expressivity to the OWL language. owl:equivalentProperty and owl:inverseOf elements are examples of those characteristics. The owl: equivalentProperty element associate properties that have the same range and the same domain. For example, the property "leacturesIn" is equivalent to "teaches" and in OWL this can be represented as:

```
<owl:ObjectProperty rdf:ID="leacturesIn">
    <owl:equivalentProperty rdf:resource="#teaches"/>
</owl:ObjectProperty>
```

The owl:inverseOf element can be used to define inverse relation between properties. If property P' is stated to be the inverse of property P'', then if X'' is related to Y'' by the P'' property, then Y'' is related to X'' by the P' property. For example, "teacher teaches a course" is the inverse of "a course is taught by a teacher." This can be expressed in OWL as:

```
<owl:ObjectProperty rdf:resource="#teaches">
    <owl:inverseOf rdf:resource="#isTaughtBy"/>
</owl:ObjectProperty>
```

The property element has some properties that can be defined directly:

- **Function property:** The function property (owl: FunctionProperty) defines a property that has at the most one value for each instance.
- **InverseFuncionalProperty:** In OWL, by using the InverseFuncionalProperty (owl:InverseFuncional-Property) it is possible to define properties that have different values to different instances, that is, two different instances can not have the same values.
- **Transitive property:** The transitive property is understood as: if the pair (x, y) is an instance of the transitive property P, and the pair (y, z) is an instance of P, we can infer the pair (x, z) is also an instance of P.
- **Symmetric property:** The symmetric property (owl: SymmetricProperty) is interpreted as follows: if the pair (x, y) is an instance of A, then the pair (y, x) is also an instance of A.

The following example illustrate the application of the owl:SymmetricProperty and owl:TransitiveProperty elements.

```
<owl:ObjectProperty rdf:ID="hasSameGradeAs">
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
    <rdf:type rdf:resource="&owl;SymmetricProperty"/>
    <rdfs:domain rdf:resource="#student"/>
    <rdfs:range rdf:resource="#student"/>
</owl:ObjectProperty>
```

FUTURE TRENDS

Many researchers worldwide have recognized that the Semantic Web (or Web3.0) is the key to develop the new generation of information systems. The number of international conferences organized every year on this topic clearly shows the interest and importance of this new technology. In this context, OWL is the most widespread language to develop a new breed of the Semantic Web-based applications. OWL has been used in many areas; some applications and tools use this conceptual approach to build Semantic Web based systems. According to TopQuadrant (TopQuadrant, 2005), a consulting firm that specializes in Semantic Web technologies, the market for semantic technologies will grow at an annual growth rate of between 60 percent and 70 percent until 2010

In the near future, we will see the use of OWL to implement applications raging from semantic social networking, semantic RSS, semantic podcasts, semantic wikis, semantic blogs to semantic mashups. As you can see, we will be devising a solution that matches most of executives' future product acquisitions strategies. Adding semantics to these types of applications is important since in a survey of 8,300 executives from McKinsey it was found that when asked about their plans to invest in tools in the future, the answers given included those applications.

Enterprise Information Integration (EII) is another area that will benefit from the Semantic Web and OWL. Today, integration is a top priority for many European and worldwide enterprises. Most organizations have already realized that the use of Semantic Web technologies (Berners-Lee et al., 2001) is the best solution to support cross-organizational cooperation for SMEs that operate in dynamically changing work environments. Semantic Web technologies are already viewed as a key technology to resolve the problems of interoperability and integration within the heterogeneous world of ubiquitously interconnected systems with respect to the nature of components, standards, data formats, protocols, etc. Moreover, we also believe that Semantic Web technologies can facilitate not only the discovery of heterogeneous components and data integration, but also the communication, coordination and collaboration behavioral of employees and individuals. semantics can help not only the system, but also human integration and interoperability.

Managing information in enterprises faces three barriers that have to be overcome: the diverse data formats, the disparate nature of content and the need to drive "intelligence" from this content. The Semantic Web helps to surpass these limitations by providing a way to add semantic metadata to documents. Metadata allows software programs to automatically understand the full context and meaning of each document. So, it is accurate to say that semantics will enable information integration and analyses in the following tasks:

- Extracting, organizing and standardizing information from many disparate and heterogeneous content sources and formats.
- Identifying interesting and relevant knowledge from heterogeneous sources and formats.
- Making efficient use of the extracted knowledge and content by providing tools that enables fast and high-quality querying, browsing and analysis of relevant and actionable information.

Finally, programming the Semantic Web with OWL can reduce and eliminate terminological and conceptual confusion by defining a shared understanding, that is, a unifying framework enabling communication and cooperation amongst people in reaching a better inter-enterprise organization. Presently, one of the most important roles ontology plays in communication is that it provides unambiguous definitions for terms used in a software system, but semantics needs to be applied rapidly to human integration to enable communication, coordination, and cooperation. The use of ontologies for improving communication has already been shown to work in practice.

CONCLUSION

The Semantic Web is the future vision of the current Web, where information will have a precise meaning. Currently, the WWW is primarily composed of documents written in a language (HTML) that is useful for visual presentation, but not for computerized understanding. The Semantic Web is not a separate Web, but an extension of the current one, in which information is a given well-defined meaning, enabling computers and people to work better in cooperation. To make possible the creation of the Semantic Web it is important to have a language that: (1) describes the concepts of a given domain, and (2) creates ontologies. One of the most prominent ontology languages to achieve those two tasks is OWL (ontology Web language) which can be used to develop Semantic Web applications. These applications will constitute a new wave of enhanced systems that will understand better the domain in which they are working and with which they interact. OWL defines a common set of terms that are used to describe and represent a specific domain. Thus, standard OWL enables the Web to be a global infrastructure for sharing both documents and data, which makes searching and reusing information easier and more reliable as well. OWL can be used by applications to improve search engines on the Web and tools to manage knowledge. In this chapter we have laid out the foundations of the Semantic Web, its associated languages and standards. These elements are the basic building blocks of any Semantic Web application.

REFERENCES

Berners-Lee, T., J. Hendler, Lassila, O. (2001 May). The semantic web. *Scientific American. 279*(5). 34-43.

Campanini, S.E., Castagna, P. and Tazzoli, R. (2004). Platypus wiki: a semantic wiki wiki web. In semantic Web Applications and Perspectives. *In proceedings of 1st Italian semantic Web Workshop.*

Cardoso, J. (2007). Semantic Web Services: Theory, Tools and Applications. New York, NY, USA, IGI Global, ISBN:978-1-59904-045-5.

Cardoso, J., & A. Sheth (2005). Semantic Web Process: powering next generation of processes with semantics and Web services. Heidelberg, Lecture Notes in *Computer Science*, Springer-Verlag, Vol. 3387.

Cardoso, J., & A. Sheth (2006). *Semantic web services, processes and applications.* Springer.

Cayzer, S., & Shabajee, P. (2003). Semantic Blogging and Bibliography Management. *Blogtalk the First European Conference on Weblogs (Blogtalk 2003*) Vienna, Austria.

DC. (2005). The Dublin Core Metadata Initiative. Retrieved May 9, 2007, from http://dublincore.org/

Decker, S.; Melnik, S.; van Harmelen, F.; Fensel, D.; Klein, M.; Broekstra, J.; Erdmann, M.; Horrocks, I. (2000). The semantic web: The roles of XML and RDF. *Internet Computing 4*(5), 63-74.

L. Ding, T. Finin, and A. Joshi. (2005). Analyzing social networks on the semantic web. *IEEE Intelligent Systems 1*(9).

HTML. (2007). Hyper Text Markup Language. Retrieved May 9, 2007, from http://www.w3.org/html/

McIlraith, S.A.; Son, T.C.; Honglei Zeng. (2001). Semantic web services. *IEEE Intelligent Systems 16*(2), 46-53.

OWL. (2004). Web Ontology Language (OWL). Retrieved May 9, 2007, from http://www.w3.org/TR/owl-features/

RDF. (2002). Resource Description Framework (RDF). Retrieved May 9, 2007, from http://www.w3.org/RDF/

TopQuadrant. (2005). TopQuadrant. 2005. Retrieved 15 May 2007, from http://www.topquadrant.com

XML. (2007). Extensible Markup Language (XML). Retrieved May 9, 2007, from http://www.w3.org/XML/

XMLSchema. (2005). XML Schema. Retrieved May 9, 2007, from http://www.w3.org/XML/Schema

KEY TERMS

Metadata: Data that describe other data. Generally, a set of metadata describes a single set of data, called a resource.

Ontology: Is a description of concepts and relationships that can be used by people or software agents that want to share information within a domain. An ontology document defines the terms used to describe and represent a domain.

OWL: A markup language for publishing and sharing data using ontologies on the Internet. OWL is a vocabulary extension of the RDF and is derived from the DAML+OIL Web Ontology Language.

RDF: Resource description framework is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata model using XML but which has come to be used as a general method of modeling knowledge, through a variety of syntax formats.

RDFS: RDF schema is an extensible knowledge representation language, providing basic elements for the definition of ontologies, otherwise called RDF vocabularies, intended to structure RDF resources.

Semantic Web: The Semantic Web provides a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries. It is a collaborative effort led by W3C with the participation of a large number of researchers and industrial partners.

XML: The extensible markup language (XML) is a simple, very flexible text format derived from SGML (ISO 8879). XML is accepted as a standard for data interchanged on the Web, allowing for the structuring of data but without meaning.

Category: High Performance Computing

Parallel and Distributed Visualization Advances

Huabing Zhu
National University of Singapore, Singapore

Lizhe Wang
Institute of Scientific Computing, Forschungszentrum Karlsruhe, Germany

Tony K. Y. Chan
Nanyang Technological University, Singapore

INTRODUCTION

Visualization is the process of mapping numerical values into perceptual dimensions and conveying insight into visible phenomena. With the visible phenomena, the human visual system can recognize and interpret complex patterns. One can detect meaning and anomalies in scientific data sets. Another role of visualization is to display new data in order to uncover new knowledge. Hence, visualization has emerged as an important tool widely used in science, medicine, and engineering.

As a consequence of our increased ability to model and measure a wide variety of phenomena, data generated for visualization are far beyond the capability of desktop systems. In the near future, we anticipate collecting data at the rate of terabytes per day from numerous classes of applications. These applications can process a huge size of data, which are produced by more sensitive and accurate instruments, for example, telescopes, microscopes, particle accelerators, and satellites (Foster, Insley, Laszewski, Kesselman, & Thiebaux, 1999). Furthermore, the speed of the generation of data is still increasing.

Therefore, to visualize large data sets, visualization systems impose more requirements on a variety of resources. For most users, it becomes more difficult to address all requirements on a single computing platform, or for that matter, in a single location. In a distributed computing environment, various resources are available, for example, large volume data storage, supercomputers, video equipment, and so on. At the same time, high speed networks and the advent of multi-disciplinary science mean that the use of remote resources becomes both necessary and feasible (Foster et al., 1999).

BACKGROUND

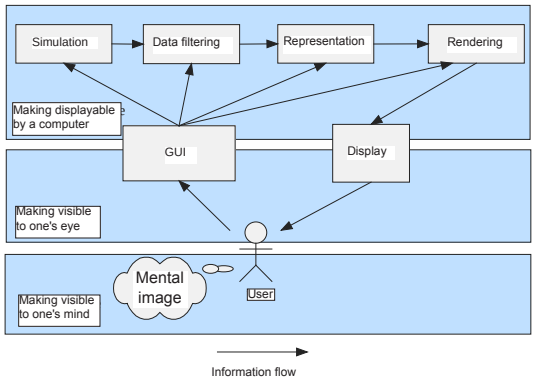
Visualization Process

The process of visualization is decided by the choice of representation. The process of visualization with 3-D, real-time interaction is much more complicated than one with still images. No matter how a specific process behaves, it can be considered in three different but interrelated semantic contexts (see Figure 1). They are *making displayable by a computer*, *making visible to one's eyes* and *making visible to one's mind* (Brodlie et al., 2004). This section provides an overview of the process of creating a 3-D, real-time, interactive visualization.

Interactive means that the system responds quickly enough for users to adjust the controlling inputs in rapid response to the output (Mueller, 2001). *Real time* means that the system must always keep updating outputs within a certain small fixed amount of time (Mueller, 2001). Generally, *real time interactive visualization* is defined as the process of creating images at rates between approximately 1 and 100 frames per second. In particular, if the latency exceeds beyond approximately 1 second, humans would feel that the computer's responses is too slow to interact continuously. However, increasing the frame rate from 100 frames per second to 1000 frames per second is of no use due to perceptual characteristics of the human brain (Igehy, 2000).

Figure 1 shows the visualization flow chart of real-time, interactive visualization. In this diagram, information flow follows the arrows. Firstly, data are generated from some simulation systems such as a mathematically based computational model or a collection of observed values. Data filtering involves a wide range of operations, such as removing noise, replacing missing values, and so on, and makes data readable to visualization software. These operations are also used to refine the data by sifting the most relevant aspects and removing unnecessary values.

Figure 1. Interactive visualization process



After the data are filtered, the representation procedure maps data to some geometric form. At this point, a geometric scene graph will be setup. The scene graph involves the geometric objects, the color value of the objects, the materials, and so on. These parameters can also be driven by computational models within the constraints imposed by visualization software.

The rendering procedure takes the information of scene graph and computes the 2-D image for human eyes to see. These images will be stored in a color buffer temporal for display. The resultant images will be displayed on computer screens or other output devices, such as project wall, Head Mounted Display (HMD).

Using graphical user interfaces (GUI), users can steer visualization systems. One can adjust the visualization system by modifying variables and data in the simulation, data filtering, representation, and rendering stages and get real-time response. It is important for real-time interaction that both simulations and graphics systems should provide real-time performance and allow for user input.

Towards Distributed Visualization

There are trends to increase not only the complexity of the models that are displayed, but also the resolution of the images. However, users still require that the system must always respond with updated output within a certain small fixed amount of time (Mueller, 2001). Then hundreds of megaFLOPS of performance and memory bandwidth of gigabytes per second are demanded. These requirements are far beyond capabilities of a single desktop system (Molnar, Cox, Ellsworth, & Fuchs, 1994). Therefore, research efforts have been made to build high performance architectures for visualization. Several recent developments have demonstrated how graphics hardware of a PC cluster can accelerate a graphics and visualization task (Muraki, Lum, Ma, Ogata, & Liu, 2003). In order to speed up the development of distributed visualization systems, some toolkits are developed for distributed parallel visualization.

WireGL (Humphreys, Buck, Eldridge, & Hanrahan, 2000; Humphreys et al., 2001) is the first sort-first cluster rendering system. It includes an efficient network protocol, a geometry bucketing scheme, and an OpenGL (http://www.opengl.org) state tracking algorithm. It supports heterogeneous hardware platforms. WireGL divides the computational nodes into clients and rendering servers. It replaces the OpenGL driver on client machines, intercepts the OpenGL calls, and sends the calls over a high-speed network to servers, which render the geometry. WireGL classifies each OpenGL call into one of three categories: (1) geometry, (2) state, or (3) special. Each rendering server receives OpenGL commands from remote clients and renders a portion of the screen space. Chromium (Humphreys et al., 2002), the new version of WireGL, is a stream-oriented framework for processing streams of OpenGL commands on parallel architectures, for example, clusters. It can support sort-first, sort-last, and hybrid parallelization strategies by the use of stream processing units. Some researches, such as VTK and OpenRM (Bethel, Humphreys, Paul, & Brederson, 2003), integrated Chromium with visualization software.

Mesa is an open source software implementation of OpenGL without using a hardware accelerator. PMesa (Mitra & Chiueh, 1998) is a parallel version of Mesa. PMesa is built with sort-last architecture. The geometry data are arbitrarily assigned to rendering processors. Mitra and Chiueh developed a general parallel compositing algorithm, which is integrated with both binary swapping composition (Ma, Painter, Hansen, & Krogh, 1994) and parallel pipeline composition (Lee, Raghavendra, & Nicholas, 1996). According to the number of composition processors, it divides the processors into several groups. Inside one group, it performs parallel pipeline composition. Between groups, it runs binary sweeping composition. In this way, it not only keeps all processors at high utilization efficiency, but also cuts down the step number for composition to optimize the performance.

AnyGL (Yang, Shi, Jin, & Zhang, 2002) is a software implementation of a sort-anywhere architecture.

Sort-Both (Zhu, Chan, Wang, & Jegathese, 2004) is a hybrid architecture which includes both sort-first and sort-last stages in the rendering pipeline. Other systems, such as OpenSG (http://www.opensg.org), Aura (Van der Schaaf, Renambot, Germans, Spoelder, & Bal, 2002), and PGL (Crockett & Orloff, 1993), have achieved a set of valuable results in parallel rendering.

TECHNOLOGIES OF PARALLEL AND DISTRIBUTED VISUALIZATION

Design Methods for Parallel Visualization

Parallelism

When scenes are complex, or when high quality images or high frame rates are required, rendering processes become computationally demanding. To provide necessary levels of performance, parallel computing technologies have been applied in the research field. Parallel visualization refers to the exploitation of parallelism in performing the visualization computations. Several different types of parallelism can be applied in visualization processes. These include functional parallelism, data parallelism, and temporal parallelism (Crockett, 1998). Each type is suitable for a set of applications or specific rendering methods.

- **Functional Parallelism:** Different processors perform different functions in the visualization pipeline. It is used in pipelines of processors, where the operations to be performed are distributed among a set of processors. As a processor completes work on one data item, the processor forwards the data item to the next processor, and then it receives a new item from its upstream neighbor. The degree of parallelism achieved is proportional to the number of functional units in the pipeline. Therefore, the available parallelism is limited to the number of stages in the rendering pipeline.
- **Data Parallelism:** The data are split up among processors, and each processor performs operations on a portion of the data. The parallelism that is achievable with this approach is not limited by the number of stages in the rendering pipeline. However the performance is challenged by the network bandwidth and the number of processors which can be incorporated into a single system. Data parallelism can be classified into two categories—object parallelism and image parallelism (Crockett, 1998). Because the data parallelism approach can take advantage of larger numbers of processors, it has been adopted in one form or another by most of the software rendering implementations. Data parallelism has been implemented for general purpose massively parallel systems. Data parallelism also lends itself to scalable implementation. It can allow the number of processing elements to vary with some factors, for example, scene complexity, image resolution, and desired performance levels. Data parallelism is the most suitable for interactive visualization with massive data sets.
- **Temporal Parallelism:** Temporal parallelism is also mentioned as frame parallelism. In animation visualization, the tasks can be decomposed in the time domain. Different processors work on successive frames in an animation sequence or interactive session. The main advantage of frame parallelism is its linear speedup—the usual performance is proportional to the number of processors. However, the performance increase is not accompanied by a reduction in network latency.

Temporal parallelism and function parallelism is naturally load balanced and easy to be implemented. However, they are not suitable for real–time, interactive visualization. The following discussion will focused on data parallelism.

Parallel Visualization Strategy

For data parallelism visualization systems, Molnar et al. (1994) have classified parallel rendering strategies into sort-first, sort-middle, and sort-last. The classification is based on where the data redistribution takes place in the graphics pipeline.

- **Sort-First:** In sort-first architectures, also known as *image space parallel*, the image space is partitioned into nonoverlapping 2-D regions (i.e., tiles) and each processor does just enough geometry processing to determine the region of the raster image which a primitive will belong to. The primitive is then sent to the appropriate processor to perform both the geometry processing and rasterization for that region of the raster image. The main advantage of sort-first is low communication requirements. Another advantage is that the processors implement the entire rendering pipeline for a portion of the screen. These improve the performance of distributed systems. However, it is susceptible to load imbalance with the reason of the random distribution of primitives in the image space. Another significant disadvantage is that it produces extra rendering work. Therefore, the scalability of sort-first systems is limited.
- **Sort-Middle:** Sort-middle architectures are those that perform sorting and data redistribution at the obvious place between geometry processing and rasterization. In a sort-middle strategy, each geometry processor computes screen coordinates for each primitive that has been assigned to it. It then determines the rasterizing processor to which it will send a primitive's scan line information, based on the raster image's partition among processors. One problem with sort-middle strategy is that they are susceptible to load imbalance of rasterizing processors due to nonuniformly distributed primitives. Another problem is high communication

cost if the tessellation ratio is high. Sort-middle is best suited for tightly coupled systems that use a fast, global interconnection to send primitives between geometry and rasterization processors.

- **Sort-Last:** Sort-last architectures, also known as *object space parallel*, defer sorting until the last stage. Primitives are arbitrarily distributed to available processors. Each processor then performs both geometry processing and rasterization of its assigned primitives, regardless of where they fall within the raster image. The final stage is to compose (sort and merge) the raster images from all rendering processors to form the final image with depth information. The advantage of the sort-last strategies is that it is less prone to load imbalance and is scalable. The main disadvantage is that it usually requires an image composition network with very high bandwidth and processing capabilities to support transmission and composition of overlaps (Molnar et al., 1994).

Task Decomposition

Task partitioning is important for parallel and distributed systems. It is particularly critical for distributed memory architectures, for example, Beowulf clusters, where partitioned data are distributed to different processing nodes.

For sort-middle and sort-last strategies, the geometry data can be arbitrarily and evenly partitioned before sending to the graphics pipeline. Sort-first strategy subdivides the screen and the divided regions are assigned to different processors. Because of the arbitrary distribution of geometry data in screen space, it is critical to keep load balancing. The task partitioning methods for sort-first architectures can be broadly categorized as either static or adaptive.

- **Static Data Partitioning:** The general static approach is to divide the screen into more regions than number of processors and assign the regions to the processors in interleaved fashion. Region shapes have been based on scan lines, horizontal strips (Eldridge, 2001; Whelan, 1985), vertical strips (Whelan, 1985), and rectangular areas (Chen, Stoll, Igehy, Proudfoot, & Hanrahan, 1998; Cox, 1997). The key idea is that if the screen is divided finely enough, each processor would have similar portions of both the populated and the sparse areas of the screen, and thus they should have nearly equal loads. The first interleaved partitioning algorithm is proposed by Fuchs and Johnson (1979). Figure 2 shows the processors are assigned regions in an interleaved pattern.

Figure 2. Assignment in interleaved pattern (based on Ellsworth, 1996)

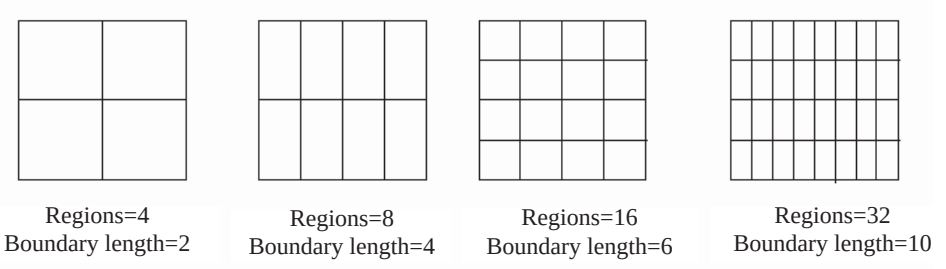
1	2	1	2	1	2	1	2
3	4	3	4	3	4	3	4
1	2	1	2	1	2	1	2
3	4	3	4	3	4	3	4
1	2	1	2	1	2	1	2
3	4	3	4	3	4	3	4

Obviously, in this way, load balance is difficult to guarantee, since it is still possible that most of the primitives are concentrated into one region. Then, the overloaded processor will be the bottleneck of the system. In order to address this problem, some techniques have been utilized, such as level of detail (LOD) (Clark, 1976) control and culling algorithms. Another problem is that a large number of partitions will introduce large redundant rendering. Primitives that overlap region boundaries must be processed in multiple regions. This inefficiency is expressed in terms of the overlap factor—the number of regions covered by a typical primitive. Both the overlap factor and primitive traffic are proportional to the total linear length of the region boundaries. As shown in Figure 3, it is intuitively easy to see that more boundaries will provide more opportunities for primitives to cross them. Therefore, a large number of partitions will dramatically increase computation and communication overhead and do harm to the system scalability. Therefore, the number of data partitions, namely granularity ratio, is important for static sort-first algorithms. For overhead considerations a small granularity ratio is better, whereas a larger granularity ratio would generally improve the load balance.

- **Adaptive Data Partitioning:** Adaptive approaches dynamically partition the screen space. There are a variety of solutions. Adaptive solutions offer the possible benefit of keeping the number of divisions to a minimum, but at the cost of increased overhead and complexity.

Samanta, Zheng, Funkhouser, Li, and Singh (1999) divide adaptive decomposition approaches into three types, top-down, bottom-up, and optimization. Top-down approaches start from the screen space as a whole and divide it recursively into regions based on estimated workloads. Bottom-up approaches start from a large number of predetermined small regions and combine them into larger regions that are then assigned to processors. Optimization approaches begin with some initial decomposition and assignment (e.g., a static one or the one from the previous frame) and adjust it to balance workloads by cutting out and reassigning smaller

Figure 3. Boundary length and number of regions (Muller, 2001)



regions from existing partitions to meet some load balancing criterion.

Whelan's (1985) median-cut method is a typical top-down approach. Median-cut splits the screen into subregions based on the onscreen position of the centroids of each polygon. The algorithm recursively splits the screen (along the longer dimension of the given region) until the number of regions equals the number of processors. Since the algorithm is based only on the centroids of primitives, it cannot accurately take into account the onscreen area overlapped by the primitives. Therefore, it is difficult to achieve good and steady load balance in this way.

Muller's (1995) mesh-based adaptive hierarchical decomposition (MAHD) algorithm is another top-down decomposition approach. Primitives are first tallied up according to how their bounding boxes overlap a fine mesh. For each cell covered by a given primitive's bounding box, an amount proportional to the rendering costs for that primitive is tallied. Once all the primitives have been counted, the cells are added up to form a summed area table. Then, using this data table as a hint, screen space tiles are recursively split along their longest dimensions until the number of regions equals the number of processors.

Samanta, Funkhouser, Li, and Singh (2000) introduced a hybrid sort-first and sort-last task decomposition algorithm. It simultaneously decomposes the 2-D screen into regions and the 3-D polygonal model into groups and assigns them to PCs to balance the load and minimize overheads. However, different regions may overlap each other. The overlapped area will do image composition like sort-last. It may cause large communication overhead when a large partition number is needed.

Image Composition

Image composition transforms parallel streams into a useful output (usually a single image). It is often the bottleneck of algorithms, especially sort-last algorithms, in high performance visualization (Brodlie et al., 2004).

In sort-last approaches, primitives are arbitrarily distributed to available processors. Each processor then performs both geometry processing and rasterization of its assigned primitives, regardless of where they fall within the raster image. The final stage is to compose (sort and merge) raster images from all nodes to form a final image with depth information.

The naive composition approach is to have a designated processor accept the contributions from all of the other processors, performing the appropriate Z-buffering or composing operations for each contribution. Obviously, in this way, the composing processor will become the bottleneck. With a large number of processors, it will be overloaded. Ma et al. (1994) developed the binary-swap composing algorithm to parallelize the image composition among processors. It first distributes primitives arbitrarily among processors. Then, each processor renders its primitives into a whole frame image with Z-buffer. After that, as its name implies each processor exchanges data with another processor and composes the image. Each processor sends half of its remaining color and Z-buffer to the other processor and composites its remaining buffer with the incoming data. The resulting smaller color-buffer and Z-buffer are then used in the next iteration until all the processors have exchanged data with each other. Figure 4 shows an example of the binary-swap composing algorithm using four processors.

Lee et al. (1996) developed a family of image composition algorithms, named *parallel pipeline composition*, for sort-last rendering systems. In this approach, the color-buffer and Z-buffer at each processor is divided into N portions, where N is the number of processors. Processors are organized in a circular ring and the subimages are accumulated in a pipelined fashion along the ring with each processor involved in each stage. At the end, each processor holds a fraction of the final image. Figure 5 shows the procedure of parallel pipeline composition with four processors. In order to improve the performance of the parallel pipeline composition, many optimization methods have been developed, including bounding box optimization, direct pixel

Figure 4. Binary-swap compositing algorithm (based on Ma et al., 1994)

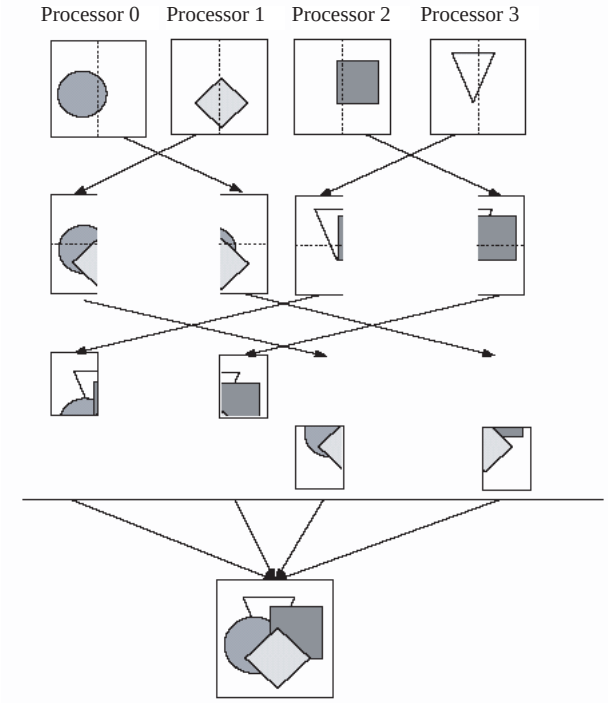
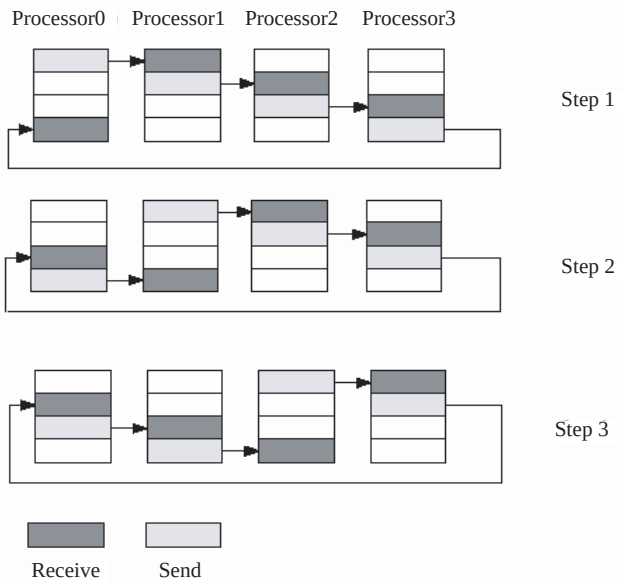


Figure 5. Parallel pipeline composition with four processors (based on Lee et al., 1996)



forwarding, interleaved composition region, and adaptive task scheduling.

Nguyen, Peery, and Zahorjan (2001) developed a partition algorithm—Image Layer Decomposition (ILD), which is similar to sort-last in that it uses an object partition. However, unlike sort-last, for each frame, ILD repartitions the scene objects into *P* nonmutually occlusive subsets for a system with *P* nodes and assigns each subset to a node for rendering. Since subsets are nonmutually occlusive, each node generates a coherent image layer; to compose the final image. ILD simply layers these image layers on top of one another according to the visibility order of the subsets (and thus does need the Z-buffer generated at each node).

FUTURE TRENDS

Grid computing (Foster, Kesselman, & Tuecke, 2001) now becomes a new computing paradigm and solution for high performance distributed computing. Computational grids can provide dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities for high performance scientific and engineering applications.

Several research challenges for large scale distributed visualization system, for example, large data set management, network performance improvement, and adaptation to heterogeneous environments are needed to be addressed. State-of-the-art grid technologies provide numbers of solutions for these problems (Karonis et al., 2003).

CONCLUSION

This article firstly introduces the background of visualization systems, with specific regard on distributed visualization technologies and systems. Sort-first, sort-middle, and sort-last algorithms are presented, analyzed, and evaluated. Recent popular parallel and distributed visualization systems are introduced and investigated. Some interesting observations can be found after the study. The strategies used in parallel and distributed visualization are in multiple levels. In the pipeline of visualization, parallelism can be exploited by executing the steps in the pipeline currently in the distributed resources. In the separate steps, parallelism can be exploited by partitioning the task and mapping them to multiple compute elements.

Visualization is a useful approach for computational scientists and engineers to analyze and represent data. As modern computational science evolves, increasing requirements are imposed on visualization systems, for example, large volume data, geographically distributed research environments, more complex computational models, and so forth. On the other hand, development of computing technologies, for example, high speed network, large size data storage, and especially grid technology, propose a promising solution for the research challenges—large scale distributed visualization on computational grids.

REFERENCES

Bethel, W., Humphreys, G., Paul, B., & Brederson, J. (2003). Sort-first, distributed memory parallel visualization and rendering. *Proceedings of the IEEE Symposium on Parallel and Large Data Visualization and Graphics.*

Brodlie, K., Brooke, J., Chen, M., Chisnall, D., Fewings, A., Hughes, C. , et al. (2004). Visual supercomputing tecnologies, applications and challenges. In *EUROGRAPHICS.*

Chen, M., Stoll, G., Igehy, H., Proudfoot, K., & Hanrahan, P. (1998). Simple models of the impact of overlap in bucket rendering. In *Proceedings of the Eurographics/SIGGRAPH Workshop on Graphics Hardware.*

Clark, J. H. (1976). Hierarchical geometric models for visible surface algorithms. *Communications of the ACM, 19*(10), 547-554.

Cox, M. (1997). Architectural implications of hardware accelerated bucket rendering on the PC. In *Proceedings of the SIGGRAPH/Eurographics Workshop on Graphics Hardware.*

Crockett, T. W. (1994). *Design considerations for parallel graphics libraries.* Institute for Computer Applications in Science and Engineering (ICASE). (NASA CR-194935 ICASE Rep. No. 94-49).

Crockett, T. W. (1998). Parallel rendering. In *SIGGRAPH Course Notes: Parallel Graphics and Visualization Technology,* (42), 150-207.

Crockett, T. W., & Orloff, T. (1993). A MIMD rendering algorithm for distributed memory architectures. *Proceedings of the Parallel Rendering Symposium.*

Eldridge, M. (2001). *Designing graphics architectures around scalability and communication.* Unpublished doctoral thesis, Stanford University.

Ellsworth, D. A. (1996). *Polygon rendering for interactive visualization on multicomputers.* Unpublished doctoral thesis, University of North Carolina at Chapel Hill.

Foster, I., Insley, J., Laszewski, G., Kesselman, C., & Thiebaux, M. (1999). Distance visualization: Data exploration on the grid. *IEEE Computer Magazine.*

Foster, I., Kesselman, C., & Tuecke, S. (2001). The anatomy of the grid: Enabling scalable virtual organizations. *International Journal of Supercomputer Applications, 15*(3).

Fuchs, H., & Johnson, B. (1979). An expandable multiprocessor architecture for video graphics. In *Proceedings of the 6th ACM/IEEE Symposium on Computer Architecture.*

Humphreys, G., Buck, I., Eldridge, M., & Hanrahan, P. (2000). Distributed rendering for scalable displays. In *IEEE Proceedings of the Supercomputing 2000.*

Humphreys, G., Eldridge, M., Buck, I., Stoll, G., Everett, M., & Hanrahan, P. (2001) Wiregl: A scalable graphics system for clusters. *Proceedings of the SIGGRAPH 2001* (pp. 129-140).

Humphreys, G., Houston, M., Ng, R., Frank, R., Ahern, S., Kirchner, P. D., et al. (2002). Chromium: A stream-processing framework for interactive rendering on clusters. *ACM Transactions on Computer Graphics, 21*(3), 693-702.

Igehy, H. (2000). *Scalable graphics architectures: Interface & architecture.* Unpublished doctoral thesis, Standard University.

Karonis, N., Papka, M., Binns, J., Bresnahan, J., Insley, J., Jones, D., et al. (2003). High-resolution remote rendering of large datasets in a collaborative environment. *Future Generation of Computer Systems*.

Lee, T. Y., Raghavendra, C. S., & Nicholas, J. B. (1996). Image composition schemes for sort-last polygon rendering on 2D mesh multi-computers. *IEEE Transactions on Visualization and Computer Graphics, 2*(3), 202-217.

Ma, K., Painter, J. S., Hansen, C. D., & Krogh, M. F. (1994). Parallel volume rendering using binary-swap compositing. *IEEE Transactions on Computer Graphics and Applications, 14*(4), 59-68.

Mitra, T., & Chiueh, T. (1998). A breadth-first approach to effecient mesh traversal. In *Proceedings of 13th ACM SIGGRAPH/Eurographics Graphics Hardware Workshop.*

Molnar, S., Cox, M., Ellsworth, D., & Fuchs, H. (1994). A sorting classification of parallel rendering. *IEEE Computer Graphics and Applications: Special Issue on Rendering, 14*(4), 23-32.

Mueller, C. A. (1995). The sort-first rendering architecture for high performance graphics. In *Proceedings of the Symposium on Interactive 3D graphics.*

Mueller, C. A. (2001). *The sort-first architecture for real-time image generation.* Unpublished doctoral thesis, University of North Carolina at Chapel Hill.

Muraki, S., Lum, E., Ma, K., Ogata, M., & Liu, X. (2003). A PC cluster system for simultaneous interactive volumetric modeling and visualization. In *IEEE Symposium on Parallel and Large Data Visualization and Graphics.*

Nguyen, T. D., Peery, C., & Zahorjan, J. (2001). Drrrraw: A prototype distributed 3d real-time rendering toolkit for commodity clusters. In *Proceedings of the International Parallel and Distributed Processing Symposium.*

Samanta, R., Funkhouser, T., Li, K., & Singh, J. P. (2000). Hybrid sort-first and sort-last parallel rendering with a cluster of PCs. In *Proceedings of SIGGRAPH/Eurographics Workshop on Graphics Hardware.*

Samanta, R., Zheng, J., Funkhouser, T., Li, K., & Singh, J. P. (1999). Load balancing for multi-projector rendering systems. In *Proceeings of SIGGRAPH/Eurographics Workshop on Graphics Hardware.*

Van der Schaaf, T., Renambot, L., Germans, D., Spoelder, H., & Bal, H. (2002). Retained mode parallel rendering for scalable tiled displays. *Proceedigns of the 6th annual Immersive Projection Technology (IPT) Symposium.*

Whelan, D. S. (1985). *Animac: A multiprocessor architecture for real-time computer animation.* Unpublished doctoral thesis, California Institute of Technology, Pasadena.

Yang, J., Shi, J., Jin, Z., & Zhang, H. (2002). Design and implementation of a large scale hybrid distributed graphics system. In *Proceedings of the 4th Euro graphics Workshop on Parallel Graphics and Visualization.*

Zhu, H., Chan, K. Y., Wang, L., & Jegathese, C. R. (2004). A distributed 3D rendering application for massive data sets. *IEICE Transaction of Information and Systems* E87-D(7).

KEY TERMS

Computational Grid: Computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities. (Foster & Kesselman, 1998).

Distributed Memory: Distributed memory means the memory is associated with individual processors and a processor is only able to address its own memory. Some authors refer to this type of system as a *multicomputer*, reflecting the fact that the building blocks in the system are themselves small computer systems complete with processor and memory.

Graphics Pipeline: Graphics pipeline has two major steps. Starting with primitives (polygons) in object space, a geometry processing step transforms the primitives into screen space. This is followed by a rasterization step to convert the primitives into a set of screen pixels. They finish with a set of appropriately colored pixels in the frame buffer. Each step includes several computationally intensive procedures.

Rasterization: Rasterization is the process of converting a vertex representation to a pixel representation; rasterization is also called *scan conversion.*

Rendering: Rendering is the computational process of generating an image from the abstract description of a scene (Crockett, 1994).

Scene Graph: Scene graph is a data structure used to hierarchically organize and manage the contents of spatially oriented scene data.

Z-Buffer: Z-buffer is an area in graphics memory reserved for storing the Z-axis value of each pixel.

Pattern-Oriented Use Case Modeling

Pankaj Kamthan
Concordia University, Canada

INTRODUCTION

The majority of the present software systems, such as those that run on automatic banking machines (ABMs), on mobile devices, and on the Web, are interactive in nature. Therefore, it is critical to precisely understand, identify, and document the services that an interactive software system will provide from the viewpoint of its potential users. A large and important class of models that these services encapsulate is *use cases* (Jacobson, Christerson, Jonsson, & Övergaard, 1992).

In the last few years, use cases have become indispensable as means for behavioral modeling of interactive software systems. They play a crucial role in various software development activities, including estimating development cost (Anda, 2003), eliciting behavioral requirements, and defining test cases.

It is well known that addressing quality *early* is crucial to avoid the propagation of problems to later artifacts (Moody, 2005). With the increasing deployment of use cases as early artifacts in software process environments, the question of *how* these models should be developed so as to attain high quality arises. In response, this article focuses on the use case modeling process (the *act* of constructing use case models) and, based on the notion of patterns (Appleton, 1997), proposes a systematic approach towards the development of use case models.

The rest of the article is organized as follows. The background and related work necessary for the discussion that follows is outlined. This is followed by the presentation of a pattern-oriented use case modeling process for systematically addressing the semiotic quality of use case models in a feasible manner. Next, challenges and directions for future research are outlined, and finally, concluding remarks are given.

BACKGROUND

In this section, the terminology necessary for the discussion that follows is presented and the significance of quality in the development of use case models is briefly reviewed.

A Primer on Use Case Models

A *use case* models the behavior of a software system, which yields an observable result of value to an actor of the system (Jacobson et al., 1992). In doing so, a use case intends to capture typical interactions between the actors and the software system being built. There can be multiple use cases of a system and they can be related. The use cases can be classified into analysis use cases (those addressing the problem domain) and synthesis use cases (those addressing the solution domain). There are three possible types of (binary and non-reflexive) relationships among use cases: "include," "extend," and "generalization."

An *actor* is an external entity that interacts with each instance of a use case. An actor could be a (human) user or another program. Each actor plays a unique *role* with respect to a use case from the viewpoint of the system. There can be multiple actors of a system, and they can also be related. The actors can be classified into primary actors (those initiating a use case) and secondary actors (those supporting the goal of a primary actor). There is one possible type of (binary and non-reflexive) relationship among actors: "generalization."

A use case is an abstraction of the real-world use of a system. A *scenario* is a concrete realization or instance of a use case; it can be classified as either normal or non-normal (including alternates and exceptions).

The s*ystem boundary* is a means to illustrate the separation of actors and use cases. The set of all actors and use cases describing the complete usage of a system is known as the system's *use case model.*

There are two common means of representing use cases: as structured text and as a graphic. Each means of representation has its own advantages and limitations (Jacobson, 2003); a detailed discussion of this issue is beyond the scope of this article.

The Unified Modeling Language (UML) (Booch, Jacobson, & Rumbaugh, 2005) is a standard language for modeling the structure and behavior of object-oriented software systems. UML provides explicit support for use case models (Bittner & Spence, 2003): the Use Case Diagram is a commonly used UML diagram type to graphically represent use case models, the Activity Diagram is used to represent the sequential order in which use cases are executed, and the Sequence Diagram is used to represent scenarios.

A View of the Quality of a Use Case Model

Using ISO/IEC 9126-1:2001 Standard, the quality of a use case model could be formally but broadly defined as the totality of characteristics of a use case model that bear on its ability to satisfy stated and implied needs.

There are different views of quality (Wong, 2006) of which an intersection of the product, user-based, manufacturing, and value-based views are applicable in our case. That is because a use case model must satisfy certain quality attributes, be communicable to the user, must conform to the language specification it is expressed in, and its development must be feasible.

Related Work on Quality-Centered Use Case Modeling

There are various reasons why quality of a use case model can be compromised, including lack of understanding of the underlying domain, lack of knowledge or skills in the modeling language, or limitations imposed by modeling tools. There have been a few initiatives addressing the quality of textual and graphical use case models (Rosenberg & Scott, 1999; Fantechi, Gnesi, Lami, & Maccari, 2003; El-Attar & Miller, 2006; Törner, Ivarsson, Pettersson, & Öhman, 2006). However, these efforts focus on the product, not on the process.

It was previously suggested that the development of use case models needs to be carried out iteratively (Rosenberg & Scott, 1999; Bittner & Spence, 2003; Leffingwell & Widrig, 2003). However, details are sketchy, and the emphasis on the improvement of quality of use case models is lacking. In particular, there is no use of patterns. A model-driven requirements engineering process that integrates certain metrics for the improvement of quality of use case models has been proposed (Berenbach & Borotto, 2006). However, the rationale for the selection of quality attributes is unclear.

A SYSTEMATIC APPROACH FOR THE DEVELOPMENT OF USE CASE MODELS OF HIGH SEMIOTIC QUALITY

In this section, a use case modeling process for systematically addressing the semiotic quality of use case models is proposed.

A Pattern-Oriented Use Case Modeling Process

The motivation for a use case modeling process (henceforth labeled as UCMP for brevity) is based on the assumption that a "good" process will lead to a "good" outcome of the process (Nelson & Monarchi, 2007), namely the use case model.

Characteristics of a Use Case Modeling Process

UCMP is expected to have the following characteristics:

1. UCMP must be cost-effective. The advantages of adopting UCMP must substantially outweigh the costs.
2. UCMP is a *sub-process* of a user-sensitive software development process. Examples of these include Crystal Methodologies (Cockburn, 2005) and the Unified Process (UP) (Jacobson, Booch, & Rumbaugh, 1999). The processes for domain modeling and requirements elicitation can impact the "velocity" of UCMP.
3. UCMP must be *both* iterative and incremental. They are complementary: being iterative enables re-visitation of a use case model, and being incremental facilitates the progress of a model.
4. UCMP must have *explicit* support for model quality assurance and model evaluation. This is necessary for construction of use case models to attain desirable quality.
5. UCMP should have a quantifiable and feasible stopping criterion.

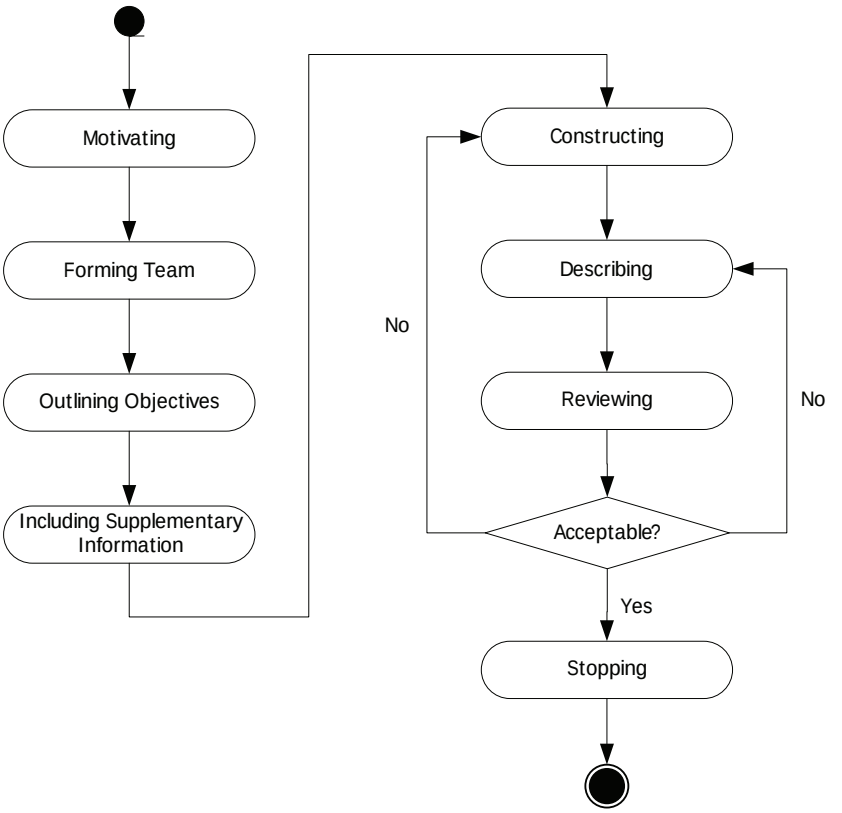
Use Case Modeling Process, Quality, and Patterns

The reliance on past experience and expertise is critical to any development. A pattern is a proven solution to a recurring problem in a given context (Appleton, 1997). A unique aspect of a pattern is that it not just describes how but *why* a certain solution works, the scope within which it works,

Table 1. A framework for addressing the semiotic quality of use case models using patterns

Product	Semiotic Level	Means for Quality Assurance	Decision Support
Use Case Model	• **Pragmatic:** Maintainability (Modifiability, Portability, Reusability), Usability (Comprehensibility, Readability) • **Semantic:** Completeness, Validity • **Syntactic:** Correctness	Patterns	Feasibility

Figure 1. The steps of a pattern-oriented use case modeling process



and it is preventative rather than curative (Dromey, 2003) in its approach towards quality improvement.

A general framework for the semiotic quality of use case models (FSQUCM) was presented by Kamthan (2008). As shown in Table 1, FSQUCM suggests the use of patterns as one means for improving the semiotic quality of use case models.

In the rest of the section, a pattern-oriented use case modeling process (POUCMP) that complies with UCMP characteristics 1-5 is proposed. Indeed, POUCMP is carried out with FSQUCM in consideration.

Steps of POUCMP

As illustrated in Figure 1, POUCMP consists of a nonlinear sequence of steps. At each step, POUCMP utilizes (process and product) patterns available from different collections (Adolph, Bramble, Cockburn, & Pols, 2003; Björnvig, 2003; Övergaard & Palmkvist, 2005).

The following assumes that the prerequisites to POUCMP have been satisfied: the team responsible for use case analysis has carried out feasibility study including domain understanding, and resource allocation (selection of a modeling language, selection of a modeling tool, and so on) has taken place. The names of patterns are expressed in *italics* to distinguish them from the surrounding text.

Step 1. Motivating: The prospective team may need motivation to get started. The *Know-HowKickoff* pattern suggests that identifying the market and making a business case to the team may help.

Step 2. Forming Team: The modeling starts with a *SmallWritingTeam* for small projects or a *BalancedTeam* for large projects. This pattern could be realized in practice using Pair Modeling (Kamthan, 2005), a practice that involves two people such that one person works on the model, while the other provides input and critical feedback on all aspects of the model as it evolves. Using the *ParticipatingAudience* pattern, one can involve the stakeholders and solicit their feedback. A systematic approach for identifying and classifying stakeholders is available (Sharp, Galal, & Finkelstein, 1999).

Step 3. Outlining Objectives: It has long been recognized that requirements elicitation is a social process (Macaulay, 1993). Thus, the team needs to have a *SharedClearVision* with respect to objectives and scope of the system.

Step 4. Including Supplementary Information: The *Adornments* pattern allows the association of information that is not central to the definition of the use case model but important nevertheless. This supplementary information, for example, could include details of non-functional requirements. It could also include metadata information such as project title, author name (using the *WritersLicense* pattern), date/time, version, copyright (Perry & Kaminski, 2005), and so on.

The process of use case modeling can lead to contentious issues (for example, due to the involvement of multiple people or otherwise) that may not be resolved instantly. These issues could also be documented separately (to be visited later in step 5) using the *ReadinessReflectionsList* pattern.

Step 5. Constructing: The actual contents of a use case model are derived as follows:

Step 5.1: The system boundary is set as a *VisibleBoundary* to have a clear separation between what is and what is not part of the system.

Step 5.2. The actors are identified using the *ClearCastOfCharacters* pattern. In the first iteration of the use case model, the primary actors can be elicited. The *MultipleActors:DistinctRoles* pattern can be used if two or more actors play different roles toward a single use case. If two or more actors play the same role toward a use case, then this role is represented by a single actor inherited by the actors sharing the role using the *MultipleActors:CommonRole* pattern.

Step 5.3: The team considers the *BreadthBeforeDepth* pattern during development to get an overview of the possible use cases before delving into details.

Step 5.4: Each use case models a complete usage of the system and corresponds to an atomic goal using the *CompleteSingleGoal* pattern. The *GoalsDefineNumber* pattern can be used to manage the number of goals and therefore the number of use cases.

Step 5.5. The pre-conditions (including triggers) of a use case are determined by *DetectableConditions.*

Step 5.6. Each use case is defined using the *UserValuedTransactions* and labeled using the *VerbPhraseName.*

Step 5.7. The (normal, alternate, or exceptional) flows are determined by the following steps.

Step 5.7.1. According to the *ScenarioPlusFragments* pattern, each flow is given the following structure: the normal behavior is described first, followed by the description of the possible alternates and exceptions (including failures).

Step 5.7.2. Using the *ExhaustiveAlternatives* pattern, it is made sure that alternates take into account all possibilities and therefore are indeed exhaustive.

Step 5.7.3. It is made sure that each (normal, alternate, or exceptional) flow ranges from 3-9 steps and all the steps in a single flow are roughly at the same level of abstraction using the *LeveledSteps* pattern, and that each step shows clearly which actor is performing the action and what the actor gets accomplished using the *ActorIntentAccomplished* pattern.

Step 5.8. Any relationships across use cases are captured using *CommonSubBehavior* (for the "include" relationship), *InterruptsAsExtensions* (for the "extend" relationship), and *CapturedAbstraction* (for the "generalization" relationship) patterns.

Step 6. Describing: The team adopts one of the many *MultipleForms* to describe the use case model, using text or using UML. This description takes the format of an *EverUnfoldingStory* so that details are revealed on a need-to-know basis. It is assured during the documentation that each use case in the use case model is *PreciseAndReadable* and *TechnologyNeutral.*

Step 7. Reviewing: The team performs a *TwoTierReview* of the use case model and revisits the previous steps 5 and/or 6 as necessary. The reason for revisiting a use case model is evaluation and the reason for subsequent refinement is the discovery of model "smells" or an unacceptable degree of one or more semiotic quality attributes. This is reminiscent of the use of patterns as "targets" for refactoring.

For example, during a review one could focus on reducing the details of each step of a flow to a minimum so as not to hinder the progress of an actor interacting with the system using the *ForwardProgress* pattern, restructuring (rather large use cases are decomposed into smaller use cases via the *RedistributeTheWealth* pattern and very small use cases may need to be merged into another use case using the *MergeDroplets* pattern), focus on the removal of redundant use cases as described by the *CleanHouse* pattern, and so on.

Step 8. Stopping: Any refinement of the use case model can take place via the *SpiralDevelopment* pattern until the stopping criterion such as that set by the *QuittingTime* pattern has been reached. Indeed, the *QuittingTime* pattern assists in making POUCMP feasible.

Note that not all steps of POUCMP are mandatory. For example, step 1 could be skipped if deemed unnecessary. Furthermore, not all steps will require the same time for completion; it is likely that steps 5-7 are the most time consuming.

Use of Anti-Patterns in POUCMP

An anti-pattern is a frequently faced "negative" solution to a recurring problem (Appleton, 1997). Indeed, if a pattern reflects a "best practice," then an anti-pattern reflects a "lesson learned." There are some anti-patterns available for use cases (Övergaard & Palmkvist, 2005; El-Attar & Miller, 2006). An anti-pattern will not explicitly improve

any of the quality attributes of a use case model; instead avoiding the anti-pattern will simply allow some quality attribute not to get worse.

For example, the anti-pattern of having *a single use case that does everything for a software system* is an impediment to comprehensibility. To mitigate this, one can for example use the *LargeUseCase:MultiplePaths* pattern (Övergaard & Palmkvist, 2005), where each of the longer flows can be modeled as a separate use case.

Scope and Limitations of POUCMP

There are inevitable constraints associated with respect to allocation of resources (i.e., time, effort, budget, or personnel) in any initiative towards quality improvement, and the same applies to use case models. In this section, the scope and certain limitations of POUCMP are highlighted.

Steps 4 and 6 of POUCMP are oriented towards textual use case models, and a graphical counterpart requires some modifications. If step 7 of POUCMP is carried out rigorously, then it is prone to the limitations inherent to inspections (Wiegers, 2002). The description of POUCMP can evolve and become more granular with the introduction of patterns from other collections. The resources pertaining to any software project are limited: there is no *a priori* guarantee that an organization may be able to allocate resources (such as the time or budget for training personnel) to deploy patterns as means for improvement of the quality of use case models. Furthermore, for a given problem there simply may not be any suitable pattern available. In general, the level of organizational process maturity (Paulk, Weber, Curtis, & Chrissis, 1995) may inhibit the extent (if at all) of an adoption of patterns.

FUTURE TRENDS

The work presented in this article can be extended in a few different directions, which is briefly discussed next.

First, POUCMP could benefit from empirical validation in both academic and industrial contexts, in particular alignment with requirements engineering processes. Second, it appears that for certain aspects of a use case process (such as conducting interviews with the client or suggesting what is to be done if the use case modeling team is not working optimally) or of a use case model (such as post-conditions or for a pre-requisite/co-requisite to a use case model such as a domain model), there are currently no corresponding patterns available. Therefore, these aspects have not been addressed in POUCMP and constitute a direction for future research.

CONCLUSION

If there is any constant in the evolution of today's information-driven software systems, it is the movement towards interaction. The conceptual models of these services, namely the use case models, must strive for high quality to be amenable to their stakeholders throughout the software development process.

A pattern-oriented approach encapsulates both the artistic (Lieberman, 2007) and the scientific aspects of conceptual modeling. A systematic approach towards modeling such as POUCMP that incorporates past expertise and experience in the form of patterns and anti-patterns can, when deployed appropriately, help construct high-quality use case models.

In conclusion, use case models are becoming first-class members of organizations and software process environments that embrace them. An investment in a quality-centric approach to use case modeling can benefit all stakeholders and in the long term can outweigh the costs.

REFERENCES

Adolph, S., Bramble, P., Cockburn, A., & Pols, A. (2003). *Patterns for effective use cases.* Boston: Addison-Wesley.

Appleton, B.A. (1997). Patterns and software: Essential concepts and terminology. *Object Magazine Online, 3*(5), 20-25.

Berenbach, B., & Borotto, G. (2006, May 20-28). Metrics for model driven requirements development. *Proceedings of the 28th International Conference on Software Engineering* (ICSE 2006), Shanghai, China.

Bittner, K., & Spence, I. (2003). *Use case modeling.* Boston: Addison-Wesley.

Björnvig, G. (2003, June 25-29). Patterns for the role of use cases. *Proceedings of the 8th European Conference on Pattern Languages of Programs* (EuroPLoP 2003), Irsee, Germany.

Booch, G., Jacobson, I., & Rumbaugh, J. (2005). *The Unified Modeling Language reference manual* (2nd ed.). Boston: Addison-Wesley.

Cockburn, A. (2005). *Crystal clear: A human-powered methodology for small teams.* Boston: Addison-Wesley.

Dromey, R.G. (2003). Software quality — prevention versus cure? *Software Quality Journal, 11*(3), 197-210.

El-Attar, M., & Miller, J. (2006, September 11-5). Matching antipatterns to improve the quality of use case models. *Pro-*

ceedings of the 14th International Requirements Engineering Conference (RE 2006), Minneapolis-St. Paul, MN.

Fantechi, A., Gnesi, S., Lami, G., & Maccari, A. (2003). Applications of linguistic techniques for use case analysis. *Requirements Engineering, 8*(3), 161-170.

Jacobson, I. (2003). Use cases: Yesterday, today, and tomorrow. *IBM developerWorks,* (November 20).

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The Unified Software Development Process*. Addison-Wesley.

Jacobson, I., Christerson, M., Jonsson, P., & Övergaard, G. (1992). *Object-oriented software engineering: A use case driven approach.* Boston: Addison-Wesley.

Kamthan, P. (2005, January 14-16). Pair modeling. *Proceedings of the 2005 Canadian University Software Engineering Conference* (CUSEC 2005), Ottawa, Canada.

Kamthan, P. (2008). A framework for understanding and addressing the semiotic quality of use case models. In J. Rech & C. Bunse (Eds.), *Model-driven software development: Integrating quality assurance.* Hershey, PA: IGI Global.

Leffingwell, D., & Widrig, D. (2003). *Managing software requirements: A use case approach* (2nd ed.). Boston: Addison-Wesley.

Lieberman, B. A. (2007). *The Art of Software Modeling*. Auerbach Publications.

Macaulay, L. (1993, January 4-6). Requirements capture as a cooperative activity. *Proceedings of the 1st IEEE International Symposium on Requirements Engineering,* San Diego, CA.

Moody, D.L. (2005). Theoretical and practical issues in evaluating the quality of conceptual models: Current state and future directions. *Data and Knowledge Engineering, 55*(3), 243-276.

Nelson, H.J., & Monarchi, D.E. (2007). Ensuring the quality of conceptual representations. *Software Quality Journal, 15*(2), 213-233.

Övergaard, G., & Palmkvist, K. (2005). *Use cases: Patterns and blueprints.* Boston: Addison-Wesley.

Paulk, M.C., Weber, C.V., Curtis, B., & Chrissis, M.B. (1995). *The Capability Maturity Model: Guidelines for improving the software process.* Boston: Addison-Wesley.

Perry, M., & Kaminski, H. (2005, July 6-10). A pattern language of software licensing. *Proceedings of the 10th European Conference on Pattern Languages of Programs* (EuroPloP 2005), Irsee, Germany.

Rosenberg, D., & Scott, K. (1999). *Use case driven object modeling with UML: A practical approach.* Boston: Addison-Wesley.

Sharp, H., Galal, G.H., & Finkelstein, A. (1999, August 30-September 3). Stakeholder identification in the requirements engineering process. *Proceedings of the 10th International Conference and Workshop on Database and Expert Systems Applications* (DEXA 1999), Florence, Italy.

Törner, F., Ivarsson, M., Pettersson, F., & Öhman. P. (2006, September 11-15). An empirical quality assessment of automotive use cases. *Proceedings of the 14th International Requirements Engineering Conference* (RE 2006), Minneapolis-St. Paul, MN.

Wiegers, K. (2002). *Peer reviews in software: A practical guide.* Boston: Addison-Wesley.

Wong, B. (2006). Different views of software quality. In E. Duggan & J. Reichgelt (Eds.), *Measuring information systems delivery quality* (pp. 55-88). Hershey, PA: Idea Group.

KEY TERMS

Pattern: A proven solution to a recurring problem in a given context.

Refactoring: A structural transformation that provides a systematic way of eradicating the undesirable(s) from an artifact while preserving its behavioral semantics.

Semiotics: The field of study of signs and the communicative properties of their representations.

Software Engineering: A discipline that advocates a systematic approach of developing high-quality software on a large scale while taking into account the factors of sustainability and longevity, as well as organizational constraints of time and resources.

Use Case Model: A behavioral model that captures typical interactions between actors and the software system under development.

Use Case Model Quality: The totality of characteristics of a use case model that bear on its ability to satisfy stated and implied needs.

Use Case Modeling Process: A set of activities and practices that are used to develop and maintain a use case model.

Category: Software & Systems Design

Patterns in the Field of Software Engineering

Fuensanta Medina-Domínguez
Carlos III Technical University of Madrid, Spain

Maria-Isabel Sanchez-Segura
Carlos III Technical University of Madrid, Spain

Antonio de Amescua
Carlos III Technical University of Madrid, Spain

Arturo Mora-Soto
Carlos III Technical University of Madrid, Spain

Javier Garcia
Carlos III Technical University of Madrid, Spain

INTRODUCTION

In the mid 1960's, the architect Christopher Alexander (1964) came up with the idea of Patterns, as "a solution to a problem within a defined context" and developed this concept. He explains, in a very original way, his ideas of urban planning and building architecture, using patterns to explain the "what", "when", and "how" of a design.

Alexander invented a Pattern Language that is the fundamental to good building and city designs, and describes it in a collection of repetitive schemas called *patterns*.

In Computer Science, software is susceptible to conceptual patterns. Consequently, Ward Cunningham and Kent Beck, used Alexander's idea to develop a programming pattern language composed of five patterns as an initiation guide for Smalltalk programming. This work was presented at the Object-Oriented Programming, Systems, Languages & Applications Conference (OOPSLA) in 1987.

In the early 1990's, Erich Gamma and Richard Helm did a joint research that resulted in the first specific design patterns catalog. They identified four patterns: Composite, Decider, Observer, and Constrainer patterns.

According to many authors, OOPSLA '91 highlighted the evolutionary process of design patterns. The synergy between Erich Gamma, Richard Helm, Rala Johnson, and John Vlissides (better known as the "Gang of Four" or GoF) and other reputable researchers (Ward Cunningham, Kant Beck or Doug Lea) definitively launched the study of and research into Object Oriented Design Patterns.

At the same time, James Coplien, another software engineer, was compiling and shaping a programming patterns catalogue in C++, which was a significant advance in the implementation phase in software development. Coplien's catalog was published in 1991 under the title "Advanced C++ Programming Styles and Idioms".

Between 1991 and 1994 the concept of pattern design was discussed at international congresses and conferences. All of these encounters culminated in OOPSLA '94. The GoF took advantage of this event to present their compilation (Gamma, Helm, Johnson & Vlissides, 1995). This publication, considered at that time as the best book on Object Orientation, compiled a 23-pattern catalog, founding the basis of patterns design.

The number of pattern-related works, studies and publications in general, but especially in design, has exponentially grown since. However, the different research groups being born must be cataloged into three fundamental paradigms:

- Theoretical approximations to the software pattern design concept and pattern languages. Coplien's work (Coplien, 1996; 2004; Coplien & Douglas, 1995) stands out in this field.
- Analysis and compilation of software applications design patterns. Rising's efforts (Rising, 1998; 2000) and Buschmman (Buschmann, Meunier, Rohnert, Sommerlad & Stal, 1996; Buschmann, Rohnert, Stal & Schmidt, 2000) are included in this classification.
- The study of special purpose patterns, like antipatterns (Brown, 1998).

As has been explained, the pattern concept has clear origins, and an important value as a reuse tool. The main

problem is that the word pattern has been used almost for everything, thus losing its original meaning. The goal of this work is to go back to the definition of patterns and present how software engineering is working with this concept.

The remainder of this chapter is structured as follows. Section two provides both, formal and informal definitions of pattern as well as the formats used to describe them. Section three presents a classification of existing patterns in the field of software engineering. In section four, the authors describe their conclusions and present the future trends in section five. A selection of key terms is defined at the end of the chapter.

BACKGROUND

Pattern Definition

The knowledge and use of pattern improves communication between the designer and the developer. According to Erich Gamma et al. (1995):

"Designers know that you do not have to solve each problem starting from scratch….you must reuse solutions which previously worked. When you find a good solution, you must use it continuously. This experience makes you an expert."

Although software engineers knew about design patterns, it was a tremendous boost for them when design patterns were systematized and categorized by four engineers Erich Gamma, Richard Helm, Ralph Johnson, and John Vlissides, known as the Gang of Tour (GoF):

They schematized 23 software design patterns through templates and used the Unified Modeling Language (UML) to describe them. They also provided examples of implementation written in Smalltalk and in C++. These patterns were later written in oriented-object language such as Java (Cooper, 1998).

FORMAL DEFINITION OF SOFTWARE PATTERNS

In this section, the most significant definitions of the software pattern have been gathered.

The first ever definition, is the one proposed by Alexander:

"A recurring solution to a common problem in a given context and system of forces" (Alexander, 1979).

Less literary but more concrete is the one proposed by Riehle and Zullighoven (1996):

"A pattern is the abstraction from a concrete form which keeps recurring in specific non-arbitrary contexts."

Nevertheless, the most precise one many authors followed is that of Gabriel (1998):

"Each pattern is a three-part rule, which expresses a relation between a certain context, a certain system of forces which occurs repeatedly in that context, and a certain software configuration which allows these forces to resolve themselves."

Gabriel defined concepts that make up the terminology of software patterns:

- Forces System: a set of objects and restrictions that have to be satisfied by the application, such as portability, flexibility, reuse, and so forth.
- Software Configuration: a set of design rules to be applied to solve the problem forces.

James Coplien (1996) enumerated the requirements that a "good" pattern has to carry out:

- Solve a problem: the patterns capture solutions, not principles or strategy.
- Provide tested solutions: the patterns show neither theories nor speculations. Simple solutions are not provided.
- Describe a relation: the patterns describe systems, structures and mechanism. They do not provide a simple module.
- Have a human component: the best patterns have to be useful.

PATTERN DESCRIPTION

Patterns must be described formally so that their content is available to all. A pattern format is a template with sections, a formal structure that eases learning, comparison among other patterns and their use. There are different formats for describing patterns such as: the Alexander, GoF and canonical formats.

The Alexander Format

Alexander explained his patterns, in a narrative style, in terms of problem to be solved, described the context in which the pattern is applied and the proposed solution. So, each Alexander's pattern is described according to the following elements:

- Name
- Problem
- Context
- Forces
- Solution

Table 1. GoF format

Field	Description
Pattern name and classification	The pattern's name conveys the essence of the pattern succinctly. The proposal and field is defined.
Intent	A short statement that answers the following questions: What does this pattern do? What is its rationale or intent? What particular issue or problem does it address?
Also Known As	Other well-known names for the pattern, if any.
Motivation	A scenario that illustrates the problem.
Applicability	Situations where the pattern can be used.
Structure	A graphical representation of the structure of classes and objects in the patterns.
Participants	The classes and/or objects participating in the pattern and their responsibilities.
Collaborations	How the participants collaborate to carry out their responsibilities.
Consequences	How does the pattern support its objectives? What are the trade-offs and results of using the pattern? What aspect of system structure does it let you vary independently?
Implementation	Techniques or tricks that you must considerer in the implementation.
Sample Code	Code that illustrates the implementation of pattern.
Known Uses	Examples of pattern in real systems.
Related Patterns	The relationships between this pattern and others and the differences among them.

The GoF Format

The Gang of Four described the patterns catalogued in their work (Gamma et al., 1995) as shown in Table 1.

The Canonical Format

The canonical format, also known as the POSA format, was proposed by Buschmann et al. (1996). Basically, it takes into account the Alexander format but is more detailed. This format is described in Table 2.

PATTERNS IN THE FIELD OF SOFTWARE ENGINEERING

In this section, we describe a classification of existing patterns in the field of software engineering.

PATTERNS CLASSIFICATION

As the Pattern concept is used in different disciplines, there are many kinds. This work has gathered the main patterns in the field of software engineering. Seven kinds of patterns with different domain applications in this area have been identified:

- Implementation patterns

1. Reference architectural patterns
2. Architectural patterns
3. Analysis patterns
4. Design patterns

- Process and improvement patterns

1. Process patterns
2. Software process improvement patterns

Table 2. Canonical format

Section	Description
Name	A meaningful and memorable way to refer to the pattern.
Problem	A description of the problem indicating the intent in applying the pattern, the intended goals and objectives.
Context	The preconditions under which the pattern is applicable.
Forces	A description of the relevant forces and constraints and how they interact/ conflict with each other and with the intended goals and objectives.
Solution	A description, using text and/or graphics of how to achieve the intended goals and objectives.
Examples	One or more sample applications of the pattern.
Resulting Context	The postconditions after pattern application..
Rationale	An explanation / justification of the pattern as a whole, or of individual components within it, indicating how the pattern actually works.
Related Patterns	The relationships between this pattern and others
Known uses	Known applications of the pattern within existing systems.

- Configuration management patterns

1. Software configuration management patterns

Implementation Patterns

Reference Architectural Patterns

These patterns correspond to the most abstract level in the category of patterns. A reference architecture is an architecture created for a specific domain.

In software terms, it can be said that these patterns affect the subsystems architecture. There is no single classification of these patterns, but a list of the most important ones follows:

- Two tier
- Multitier
- Workflow
- Forms oriented
- Hypertext oriented
- Video/imaging/multimedia
- Data warehouse
- Real time systems
- Process control systems
- Thin client

These kinds of patterns are widespread and widely used because systems architectures are identified and described in depth. However, the main inconvenience with using these patterns is the nonexistence of tools to allow the integration among subsystems.

Architectural Patterns

An Architectural Pattern expresses a fundamental structural organization or schema for software systems. It provides a set of predefined subsystems, specifies their responsibilities, and includes rules and guidelines for organizing the relationships among them (Buschmann et al., 1996).

The technical literature in this field is complicated because of the fact that many people in the software area use the term “architecture” to refer to software, and many patterns described as “architecture patterns” are high-level software design patterns.

Although the interest in architectural patterns is recent, it has been increasing in this field—extending the principles and concepts of design patterns to the architecture domain. There are two main approaches to architectural patterns proposed by Buschmann et al. (1996) and Shaw and Garlan (1996). Both describe more or less the same patterns with slight differences. According to Buschmann et al. (1996), an architectural pattern defines the following elements:

- A vocabulary of components and connectors.
- A topology that describes the way components and connectors must be combined.
- A set of requirements and constraints on components and connectors combinations.
- Semantic models that allow the specification of the global system behavior starting with the behavior of the components.

The main feature of architectural patterns is that it allows the systems designs, focusing on system features such as performance, reliability and efficiency.

Examples of Architectural patterns are Model View Controller (MVC), Presentation Abstraction Control (PAC).

These patterns are properly described and documented based on software engineering design techniques of separation. The main deficiencies of these patterns are the natural language used to define them and the absence of tools that support their implementation.

Analysis Patterns

Analysis patterns (Fowler, 1997) represent business structures that embody a similar internal structure. These patterns are linked to interfaces or data, but not to implementation. This is one the most important differences with other kinds of patterns.

There are many classifications of analysis patterns, but the most relevant ones are described by Fowler (1997).

Fowler does not use a formal format to represent the analysis pattern. However, in 2001, Geyer-Schulz and Hahslerwe defined a template for this pattern:

- Pattern Name (Gamma et al., 1995; Buschmann et al., 1996)
- Intent (Gamma et al., 1995)
- Motivation (Gamma et al., 1995)
- Forces and Context (Alexander, 1979)
- Solution (Buschmann et al., 1996)
- Consequences (Gamma et al., 1995; Buschmann et al., 1996)
- Design (Geyer-Schulz & Hahslerwe, 2001) How can the analysis pattern be realized by design patterns? Sample design suggestions.
- Known Uses (Gamma et al., 1995;Buschmann et al., 1996)

Kodaganallur and Shim (2006) developed a taxonomy of analysis patterns used for specifying system requirements.

Table 3. Analysis patterns classification

Group	Description	Pattern
Accountability	When a person or organization is responsible to another	Party Organization hierarchies Hierarchic accountability Operating scopes
Observations and Measurements	Record information about measurements	Quantity Conversion ratio Measurement Observation
Referring to Objects	Identify objects and schemes	Name Identification scheme Object equivalence
Inventory and Accounting	Basis for financial accounting, inventory or resource management	Account Transactions Summary account Posting rules
Trading	Buying and selling of goods	Contract Portfolio Quote

They explained the key characteristics, which are patterns used in the analysis phase, and whose main purpose is to help specify the requirements and relate the objects and concepts in the real world.

Design Patterns

Design patterns systematically define and explain a general design for a recurrent design problem in an object-oriented system. These patterns describe the problem, the solution, the instant the solution must be applied and the consequences. Solution is an ordered set of objects and classes to solve the problem in a context (Gamma et al., 1995). The patterns also provide tips and examples of implementation, although they are independent of the programming language.

Gamma et al. (1995) defined 23 software design patterns, independent of language and platform. They used examples which were written in Smalltalk and C++. Some years later, these 23 patterns were written in Java (Cooper, 1998). These patterns focused on common problems in object-oriented development. Currently, these design patterns are widely used and are classified by their functionality in:

- Creational patterns: to deal with the creation of object instants or classes.
- Structural patterns: to define the design of objects to develop specific tasks.
- Behavioral patterns: to define how a set of objects interact among themselves to develop specific tasks.

UML is used to express the pattern description, and in some object-oriented languages another element of design patterns is “application examples”.

The main advantages of design patterns are:

- High-quality design results: due to the experience acquired.
- Improvement of communication within the development life cycle: providing names of design structures which establish a common vocabulary to reduce the complexity.
- Design information is captured and preserved: design decisions are faster and documentation is improved.
- Restructuring is made easier: design is more flexible.
- Design patterns are related among themselves so they can be used as parts of a structure, as for example, the patterns defined in J2EE (Alur, Crupi & Malks, 2003). Assembling design patterns allow you to define the framework of an application.

On the other hand, design patterns have some negative aspects. Some of them are:

- To name everything pattern: this eliminates the capability of abstraction and reuse of pattern. The correct pattern must be chosen through the rules of selection to avoid this (to specify benefits, to demonstrate extend applicability, to check if there are two real examples or more, etc.).
- To apply the patterns blindly: sometimes the design patterns are applied without the right criteria, increasing the cost and complexity. In this way, the design of patterns is more expensive and difficult than the development of an object- oriented system without patterns.

Process and Improvement Patterns

Process Patterns

A process pattern provides a guide to show how carry to out specific tasks in the development of a process. A process pattern (Garson, 2006) is an approach to a specific task that has been tried with positive results. The experiences of other people who have performed a specific task in several ways are stored in a process pattern so that users of process patterns will benefit from these experiences.

According to Ambler (1998) a process pattern describes a proven successful approach in a set of actions in software development. The focus here is on process pattern for object-oriented and three levels or types of patterns have been provided: task, stage, and phase process patterns. The aim of these patterns is to guide the developer in developing object-oriented software at different levels.

The formal documentation of a process pattern includes the following elements:

- Name.
- Intent.
- Type.
- Initial Context.
- Solution.
- Resulting Context.
- Related Patterns.
- Known Uses/Examples.

Process patterns researchers disagree on how to document the process patterns elements, but all the patterns involve variations of the previously defined elements

Process patterns are organized in taxonomies that allow the user to find alternative patterns which are more appropriate for a specific project or phase. There are several approaches to formalize process patterns: Hagen and Gruhn (2004) defined a language, called PROPEL, based on Unified Modeling Language (UML) to describe Process Patterns; a metamodel was defined to represent process model and

process patterns in different levels of abstraction (Tran, Coulette & Dong, 2006)

The authors of this contribution think that Ambler's patterns are not intuitive so, the efficiency of use is low, and they do not allow the user to work fast, easily and dynamically.

Software Process Improvement Patterns

Software process improvement is a mechanism for continuous, improvement of quality applying the good practices and eliminating problematic ones. The good practices are those which produce people good results when applied (Appleton, 1997).

The elements to define software process improvement patterns are:

- Name
- Context
- Problem
- Forces
- Solution
- Resulting Context
- Rationale
- Related Patterns
- Known Uses

Appleton classified the software process improvement patterns into organizational structure, and process and communication patterns (Table 4):

These patterns are appropriate for an organization or department in which management has already made a genuine commitment to sponsor and support process improvement efforts. The main deficiencies of these patterns are the absence of tools to support their implementation.

CONFIGURATION MANAGEMENT PATTERNS

Software Configuration Management Patterns

Software configuration management (SCM) must be developed during the whole lifecycle and must be integrated with the rest of the development processes. Berczuk (2003) described a pattern language that allows the implementation of software configuration management practices through the use of a set of heuristics. These patterns are not described formally so they can be used as the basis for the understanding of the SCM process and as an inspiration to apply this process.

FUTURE TRENDS

The conclusions extracted from the existing software patterns demonstrate their low efficiency. It is true that patterns are more interesting for beginners than for experts but the authors believe that they should not be written to teach beginners only. Patterns are intended to empower the reuse of knowledge in order to help software engineering be more an "engineer" than an "artisan". The main cue for the success of patterns is that they should be accessible and operative. This means:

- They must follow a formal description in order to have a repository of them and, thus, can be shared with the rest of the interested community if needed. Knowledge management mechanisms should be applied in order to retrieve or upload them.

Table 4. Process improvement patterns classification

Organization Patterns	Process and Communication Patterns
Local Heroes Center PEG PIT also Practices Dedicated Improvement Processors Improvement Action Teams	Process is Product (process) Virtual Forum (communication) Process follows Practice (process) Improvement follows Process (process) Improvement follows Spiral (process)

- They must also provide, on the spot, all they information necessary for their execution. This means that even when a formal representation is being used, the descriptive information in natural language must be provided.
- Their execution must be software supported
- The patterns execution must provide feedback in order to:
  - Provide measurements of the pattern execution
  - Provide new knowledge after its execution, or even new patterns.

The low success of the patterns described in this work is based on the fact that it was forgotten that patterns should be used by humans while developing projects. The human factor is a very important element for the success of a software project.

Our vision of software patterns, oriented to attaining the above-mentioned features, is focused on patterns to help people in projects development. In this sense the authors believe in product patterns as a tool to help in the development of any software product throughout the software development lifecycle, independently of the nature of the product. Product patterns are designed to gather the know-how of the enterprise, that is, where they are going to be deployed so that everybody in the organization is involved on their use (Amescua, Garcia, Segura-Sanchez & Medina-Dominguez, 2006; Medina-Dominguez, Sanchez-Segura, Amescua & Garcia, 2007). These patterns are designed and implemented to be useful for all.

Product patterns could cohabit with existing software engineering patterns because:

- Architectural patterns provide the knowledge to define system features.
- Process patterns provide the methodology, phases and tasks framework to execute a project.
- Design patterns are the most mature ones and provide the knowledge to reuse solutions at a low level design

The rest of the existing patterns could be embedded in the term product pattern in order to be used. Otherwise, existing patterns is knowledge that, although available in electronic format, does not reach the people involved in the development of a project when needed. As a result, they are hardly ever going to be used.

CONCLUSION

In this work the authors summarized the origin of patterns, their definitions and descriptions in software engineering and present a classification of software engineering patterns, including a description of each.

It is clear that the origin of software engineering patterns and the concept have been widely used. However, the formalization of patterns is very poor and, without formalization, there is no systematization, which is one of the bases of software engineering.

ACKNOWLEDGMENT

This work has been partially funded by the Spanish Ministry of Science and Technology through the TIC2004-7083 project.

REFERENCES

Alexander, C. (1964). *Notes on the synthesis of form*. Cambridge, MA: Harvard University Press.

Alexander, C. (1979). *The timeless way of building*. Oxford: Oxford University Press

Appleton, B. (1997). Patterns for conducting process improvement. In *Proceedings of the Pattern Languages of Programs Conference*, Monticello, Illinois.

Alur, D., Crupi, J., & Malks, D. (2003). *Core J2EE patterns*. New Jersey: Prentice Hall

Ambler, S. W. (1998). *Process patterns, building large-scale systems using object technology*. Cambridge: Cambridge University Press.

Amescua, A., Garcia, J., Segura-Sanchez, M., & Medina-Dominguez, F. (2006). A pattern-based solution to bridge the gap between theory and practice in using process models. *LNCS, 3966*, 97-104. Springer.

Appleton, B. (1997). *Patterns and software: Essential concepts and terminolgy*. Retrieved June 16, from CM Crossroads Website: http://www.cmcrossroads.com/bradapp/docs/patterns-intro.html

Berczuk, S. P. & Appleton, B. (2003). Software configuration management patterns. *Efective Teamwork, Practical integration.* Addison-Wesley.

Brown, W. J., McCormick, H. W., & Thomas, S. W. (1998). *Antipatterns: Refactoring software, architecture, and projects in crisis.* New York: John Wiley & Sons.

Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., & Stal, M. (1996). Pattern oriented software architecture. *A system of Patterns*. New York: John Wiley & Sons.

Buschmann, F., Rohnert, H., Stal, M. & Schmidt, D. (2000). Pattern oriented software architecture. *Patterns for concurrent and networked objects.* New York: John Wiley & Sons.

Cooper, J. W. (1998). *The design patterns: Java companion.* Retrieved June, from The Design Patterns Website: http://www.patterndepot.com/put/8/JavaPatterns.htm

Coplien, J. O. (1996). The column without a name: The human side of patterns. *C++ Report*, *8*(1), 81-85.

Coplien, J. O. (2004). Patterns of engineering. *Potentials IEEE, 23*(2), 4- 8.

Coplien, J. O. & Douglas, C. S. (1995). *Pattern languages of program design.* Addison Wesley

Fowler, M. (1997). *Analysis patterns: Reusable object models*. Addison Wesley.

Gabriel, R. P. (1998). *Patterns of software: Tales from the software: Community*. Oxford: Oxford University Press.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software.* Addison Wesley.

Garson, E. (2006). *A whirlwind introduction to process patterns*. Retrieved June 16, 2008, from Consulting D.T. Website http://www.dthomas.co.uk/dtalm/downloads/process_patterns_intro.pdf

Geyer-Schulz, A. & Hahsler, M. (2001). *Software engineering with analysis patterns technical report 01/2001*. Retrieved June 16, 2008, from Institute for Information Business Website: http://wwwai.wu-wien.ac.at/~hahsler/research/virlib_working2001/virlib/

Kodaganallur, V. & Shim, S. (2006). Analysis patterns: A taxonomy and its implications. *Information Systems Management, 23*(3), 52-61.

Hagen, M. & Gruhn, V. (2004). Towards flexible software processes by using process patterns. In *Proceedings of the IASTED Conference on Software Engineering and Applications* (pp. 436-441),

Medina-Domínguez, F., Sanchez-Segura, M., Amescua, A., & Garcia, J. (2007). Extending Microsoft team foundation server architecture to support collaborative product patterns. LNCS 1-11.

Riehle, D.,& Zullighoven, H. (1996). Understanding and Using Patterns in Software Development. *Theory and Practice of Object Systems,*2 (1), 3-13.

Rising, L. (1998). *The patterns handbook*. Cambridge: Cambridge University Press.

Rising, L. (2000). *The pattern almanac 2000*. Addison Wesley.

Shaw, M. & Garlan, D. (1996). *Software architecture, perspectives on an emerging discipline*. New Jersey: Prentice Hall.

Tran, H. N., Coulette, B., & Dong, B. T. (2006). A UML-based process meta-model integrating a rigorous process patterns definition. Volume 4034/2006. 429-434

KEY TERMS

Analysis Pattern: Is an idea that has been useful in one practical context and will probably be useful in others (Fowler, 1997).

Design Pattern: Defines and explains systematically a general design to a recurrent problem of design in object oriented system.

Pattern: Is a recurring solution to a common problem in a given context and system of forces (Alexander, 1979).

Process: Is a set of sequential practices that are functionally coherent and reusable for software engineering organization, implementation, and management. It is usually referred to as the software process, or simply the process.

Process Improvement: Is an activity that seeks to identify and rectify "common causes" of poor quality in software systems by making basic changes in the underlying software management process (available at: http://www.sei.cmu.edu/opensystems/glossary.html).

Process Pattern: Provides a guide to show how carry out specifics tasks in the development of a process. A process pattern (Garson, 2006) is an approach to a specific task that has been tried with good results.

Product: Is any thing to be produced during the whole software development process.

Product Pattern: Is an artefact that encapsulates the knowledge of software engineering experts to obtain a specific software product.

Software Engineering: Is the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software.

Pedagogical Perspectives on M-Learning

Geraldine Torrisi-Steel
Griffith University, Australia

INTRODUCTION

The advent of multimedia on desktop computers in the late 1980s and early 1990s heralded an era of educational technology that held the promise of revolutionising the business of teaching and learning by facilitating a shift from traditional teacher-centred methods to more effective student-centred approaches. During the mid-late 1990s the popularisation of the Internet, added to educational technology a new dimension of "connectedness" between people and between people and information resources. Online learning and e-learning became icons of the era**.** In late 1990s and early 2000s major players in the mobile phone industry worked on developing a wireless infrastructure to allow for wireless communication between devices, WAP (wireless application protocol) being one of the principle outcomes. This set the stage for the wireless Internet and for another new dimension to educational technology, mobility**.** Thus, the maturation of multimedia, the Internet and communication technologies together with development and availability of ubiquitous computing devices and wireless networking birthed the notion of mobile learning (m-learning) or "learning on the move."

Like many other media technologies before, **m-learning** is considered to have the potential to reshape teaching and learning, in this instance, holding promise of unprecedented connectivity and learning interactions between learners, learners and educators, information and computing resources, anywhere, anytime. This article seeks to facilitate the realisation of the pedagogical potential of m-learning by proposing a model for the construction of m-learning spaces**.** The proposed model is founded upon a **pedagogical framework** directing attention to guiding philosophies, technology integration, and the capabilities of mobile devices.

BACKGROUND

The belief underlying the following discussion is that although technology use in educational contexts is not a requisite for positive change in teaching and learning practice, some degree of change in teaching practice is a requisite for effective technology use in educational contexts**.** The effective use of technology in educational contexts should precipitate significant and positive changes in teaching practice (Tearle, Dillon & Davis, 1999). History has shown that, the adoption of new technologies frequently occurs at a superficial level consequently failing to make significant impact on teaching and learning environments (Cuban, 1986; Hammond, 1994; Nichol & Watson, 2003; Conlon & Simpson, 2003). New technologies used inappropriately or in ways replicating traditional teacher centred approaches contribute little to improving the quality of the learning environment. From this perspective, effective integration of technology in the curriculum results from teaching practice informed by an awareness of available technologies within the context of pedagogical frameworks**.**

The manner in which m-learning is defined fosters certain perceptions and beliefs about its implementation (Laouris & Eteokleous, 2005). Of fundamental importance to pedagogical discussions surrounding m-learning is the provision of a teaching "centric" rather than "techno-centric" definition for m-learning (Laouris & Eteokleous, 2005). Techno-centric definitions of m-learning accentuate the technology as the focus rather than teaching and learning**.** The motivation for implementation of mobile technologies should be not be driven by the technology but rather driven by two phases of activities: Firstly, reflection on current teaching practice and learning outcomes in order to identify deficiencies or new avenues for new effective strategies. Secondly, consider if and how any of the array of **mobile devices** can be exploited in order to achieve more effective strategies and more effective, meaningful learning outcomes (Torrisi-Steele, 2004).

Congruent with this approach, m-learning may be defined as:

the integrative use of mobile devices into the curriculum in order to facilitate active and meaningful learning through the creation of learning spaces extending outside the physical and temporal constraints of the traditional classroom. These learning spaces (m-learning spaces) are characteristically dynamic, collaborative and focused on individual learner needs in the current context. (adapted from Torrisi-Steele, 2006)

The term "mobile devices" refers to laptop computers, tablet PCs, PDAs, mobile phones, smart phones, MP3 players and any other small portable or handheld devices technically capable of connectivity (ideally wireless) to each other, other devices or Internet.

M-learning is considered here as an extension of e-learning (Brown, 2005). M-learning may include all the features of e-learning (multimedia, information access, Internet

capability, collaboration) but with the distinguishing feature of being ubiquitous and mobile.

PEDAGOGICAL FRAMEWORK

In alignment with the definition of m-learning proposed above, the pedagogical framework for the design of effective m-learning spaces incorporates three key aspects for discussion: guiding philosophies, technology integration, and capabilities of mobile devices**.** Following a discussion of each of these aspects, a model for the implementation of m-learning spaces is provided.

Guiding Philosophies

It is well established in literature that constructivist approaches that actively engage learners by presenting them with authentic learning activities, lead to more meaningful learning outcomes and are congruent with lifelong learning goals (Strommen, 1999, p. 2). Emerging from the work of theorists including Piaget (1952), Bruner (1985), and Vygotsky (1978), the constructivist perspective describes a "theory of development whereby learners build their own knowledge by constructing mental models, or schemas, based on their own experiences" (Tse-Kian, 2003, p.295). **Constructivist learning** supports a learner centred philosophy. Learner centred philosophy promotes and allows for a high degree of learner control and the individual construction of learning pathways.

Meaningful learning is being used to refer to learning resulting in a deep understanding of complex ideas, and it is relevant to learners. Jonassen, Peck, and Wilson (1999) define meaningful learning to have the following characteristics:

- **Active:** Created by interactions and manipulations with the environment
- **Constructive:** Knowledge created by reflection and interpretation
- **Intentional:** Activity directed toward trying to achieve a goal encourages thinking and learning
- **Authentic:** Contextual clues found in "real situations" assist understanding and learning
- **Cooperative:** Conversation and interaction with others promotes understanding and exposure to ideas of others; negotiation of knowledge.

M-learning spaces are well suited to supporting principles meaningful learning and constructivist philosophies (Table 1). Mobile devices support a variety of personalised experiences. The mobility attribute enables learners to explore knowledge and situations in their own way, in a variety of places and often outside the time constraints of traditional classroom-based teaching. Mobile devices also increase motivation, provide for interactive leaning and facilitate control of the learning process and emphasise its relationship with the real world (Zurita & Nussbaum, 2004).

The ability of mobile devices to support ubiquitous communication brings the social aspects of learning into focus. M-learning is thus proving to be the catalyst for growing emphasis on **social constructivism**, and learning communities (Evans, 2005). M-learning allows for greater exploitation of collaboration and conversation as powerful learning strategies (Brown, 2005). Learning participants are able to communicate outside the bounds of physical locations and often from diverse learning contexts. Brown (2005) maintains that m-learning optimises the opportunities for interaction among learners, among educators, and among educators and learners. Consequently, communication and interaction should be exploited as critical factors for success of m-learning.

From this perspective, a valuable addition to the pedagogical toolbox for construction of m-learning spaces is the **conversational** framework proposed by **Diana Laurillard** (1993). The framework places emphasis on the role and importance of interactions in learning**.** The basic premise is that learner is more effective when learners converse with

Table 1. Congruency between aspects of constructivist learning principles and attributes of mobile devices

Constructivist principles supporting meaningful learning	**Attributes of mobile devices**
Learner-centred	Personal
Active	Includes tools for data gathering while on location e.g. image recording, sound recording, databases, spreadsheets. Impact of manipulations on environment can be recorded immediately often in multimodal manner
Constructive	Access to information through wireless to other information or mobile web allows interpretation of results in light of additional information resources (includes materials and human resoures
Intentional	Ubiquitous, goal directed activity specific to the context.
Authentic	Ubiquitous, explore information in the real world, in real contexts. Contextual clues assist understanding. Support for real-world case based learning rather than pre-determined sequence.
Co-operative	Connectivity May support multimodal interaction with others e.g. voice, video, text, images. Data can be shared and discussed spontaneously by learning participants in different locations. Allows for negotiation of knowledge through collaborative activity.

a partner and iteratively refining understanding by sharing and questioning their understandings. The "partner" may be another student or a teacher**.** The conversational space in which interactions take place can be enhanced by the use of mobile devices**.** Mobile devices free participants from the restrictions of physical location. Furthermore, mobile devices may enrich the conversational environment by providing access to data gathering, data analysis, visualisation and information retrieval tools**.**

Technology Integration

As was alluded to earlier, in order for technology to go beyond the simple acquisition of information and facilitate meaningful learning, it must be integrated into the curriculum, not used merely at a superficial level (Tearle et al., 1999). Integration of mobile technology into the curriculum presupposes reflection on teaching practice. Reflection in turn, precipitates a degree of change/innovation in existing practice.

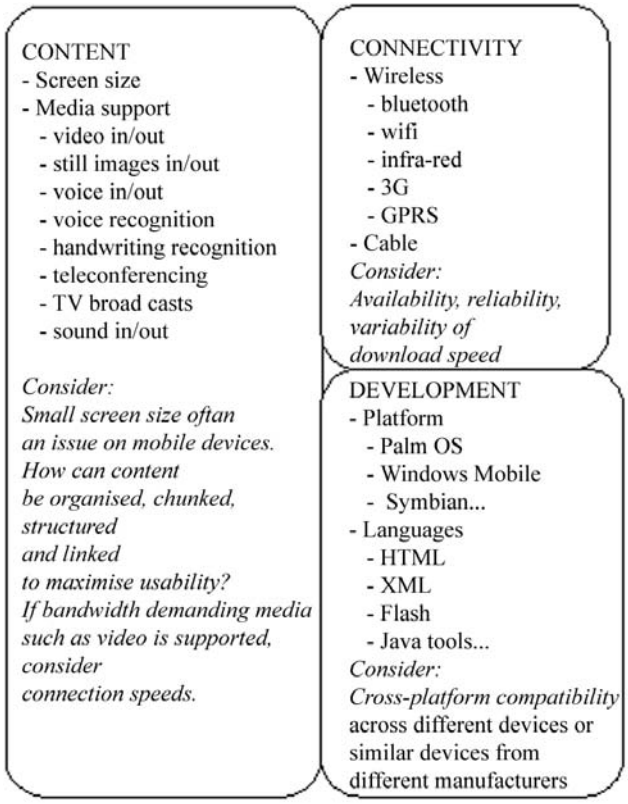
The process of reflection, change and curriculum integration is driven by the perceived need to improve the quality of teaching. The learning environment is constructed on the foundation of knowledge of the learning context (discipline requirements, learning outcomes and goals, learner needs) intertwined with knowledge of the attributes of available mobile devices. The design and development of m-learning spaces is driven by questions such as: What strategies have been used? Have they been effective in producing meaningful and desired learning outcomes? How can they be made more effective? What new strategies can be devised, given the new tools that will enrich the learning environment? How can these new strategies exploit the attributes of mobile devices in order to provide for engaging learning and more meaningful learning outcomes? (Torrisi-Steele, 2006). It is the educators' knowledge, assumptions and perceptions regarding the technology and its implementation in the specific learning context, rather than the technology itself, that will determine its implementation and hence its effectiveness (Jackson & Anagnotopoulou, 2000; Bennet, Priest, & Macpherson, 1999).

Capabilities of Mobile Devices

In order to assist the selection of mobile devices for enriching teaching and learning environments, it is useful to consider the capabilities of mobile devices in terms of three categories: content, connectivity and development (Figure 1).

In a discussion of mobile device selection it must be noted that while the capabilities of mobile devices may be desirable as driving the decision of selection, the principle driving force is often cost. The cost of devices and infrastructure costs are key challenges. Mobile devices are being identified as a means

Figure 1. Categorising the capabilities of mobile devices to assist in device selection and key relevant considerations



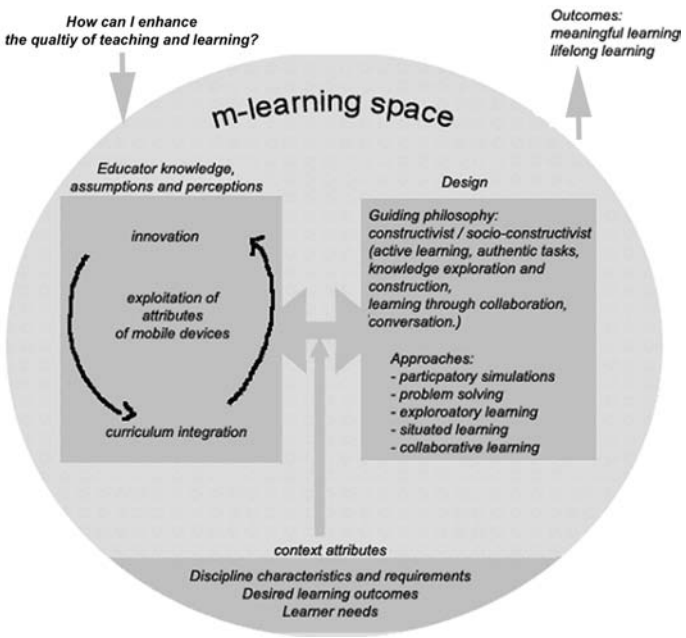
to narrow the "digital divide" given that PDAs and mobile phones are generally cheaper than computers**.** The United Nations at its World Summit on Information Society in 2005 recognised mobile devices as a way to disseminate learning among disadvantaged groups (Thomas, 2006a)**.** However, the cost and politics of establishing infrastructure to support wireless connectivity is still a significant consideration. How to design and implement a cost effective and secure system that is fast and reliable within a large service area remains a major area of consideration for educational institutions wishing to implement m-learning environments within their local area networks (Thomas, 2006b).

A Model for the Design of M-Learning Spaces

Against the background of the preceding discussion, the following model to guide the design of m-learning spaces precipitates.

The model proposed above, may be refined further by drawing attention to Parsons and Ryu's (2006) guide for pedagogical design requirements for mobile environments. The guide draws attention to ubiquity and personalisation as

Figure 2. A model for the design of m-learning spaces. The model emphasises: firstly, the process is driven by a motivation to improving effectiveness of learning environments rather than by the technology itself; secondly, the requirement for curriculum integration and innovation in practice; and thirdly, constructivist, particularly socioconstructivist philosophies to guide the design of the m-learning space (Torrisi-Steele, 2006).



key aspects of mobile environments and suggests m-learning environments should:

- Be sensitive to learner needs, learning styles, because m-learning tends to be targeted to individuals
- Be up to date in terms of content – the ubiquitousness of m-learning demands dynamic content
- Be highly interactive
- Enable mutual feedback with instructors in order to provide learner support and guidance
- Enable learners to explores knowledge
- Enable learners to collaborate with peers
- Enable learners to take a higher degree of control over their own learning
- Be goal directed and specific to the learning context at the time of use

FUTURE TRENDS

The vast majority of today's learners are digitally literate and increasingly community-orientated (Oblinger, 2004). "Connectedness" is becoming a strong theme in everyday life for these learners and will conceivably continue to do so. The inclusion of m-learning strategies in learning environments is as thus as much a response to learner expectations and the "way they do things" as it is to the search for more effective ways of conducting teaching and learning. Thus, the key concern for future discussion on m-learning is not focused on whether or not to use but rather on "how it can best be successfully used."

Two forces will shape the effectiveness of future of m-learning scenarios: technical and pedagogical. Technically, following on from current trends there is the expectation of increases in sophistication and capabilities. Context aware devices and options for synchronisation are two key areas of development. Existing networks for mobile phones are offered by a variety of service providers and support a variety of protocols. To some extent the future of m-learning lies in the ability of service providers to provide high quality, cheap, reliable and compatible networks.

From a pedagogical perspective, now more than ever, in such a varied and dynamic environment there is a need for ongoing reflection on teaching practice, focus on appropriate professional development and on fostering continuous aspirations for improving the quality of teaching and learning. Professional practice must place emphasis on teaching and learning rather than devices themselves. Knowledge of device capabilities is essential but useless if it is disconnected from the business of teaching and learning. Useful

future research will thus focus on gaining insight into the experience of learners using mobile devices, the evaluation of the effectiveness of m-learning devices in a variety of learning contexts and the documentation of cases illustrating how mobile devices have been successfully or unsuccessfully integrated into learning environments. The design and implementation of m-learning spaces must be inextricably linked with evaluation if they are to be successful (Kukulska-Hulme & Traxler, 2005).

CONCLUSION

The preceding discussion has sought to highlight some of the key pedagogical aspects of implementing m-learning. Three key themes related to pedagogy arose:

Firstly, m-learning must be considered from a learning-centric rather than technology-centric perspective. The motivation for using m-learning strategies stems from a desire to enrich the learning environment, and thus it should be integrated into meaningful learning activities rather than used superficially.

Secondly, the integration of mobile devices into learning contexts occurs as result of reflection on practice and careful consideration of discipline requirements, learner needs, context attributes along with knowledge of the capabilities of various mobile devices. Positive change in practice is seen as an unavoidable consequence of such reflection and consideration.

Thirdly, Social constructivism and conversation should form the guiding philosophy for shaping m-learning spaces. The defining characteristics of mobile devices are "connectedness" and "mobility." Social constructivist principles and conversational approaches to learning are heavily congruent with these characteristics. Furthermore, today's learners are immersed in "connectedness" and a sense of community is becoming embedded in their lives.

The notion of learning outside the constraints of the traditional classroom is not a novel approach, especially when one considers the degree of informal learning experienced outside the classroom. Learning in the field activities have been undertaken for a long time as part of many learning contexts. Mobile devices offer unprecedented access to information, communication, collaboration and manipulation of data, and thus have the capacity to enrich learning experiences outside the confines of physical and temporal parameters. The use of mobile devices in teaching and learning contexts is still in its infancy. Effective use of the technology will be driven by informed practice, attention to issues of content delivery, instructional design, as well as issues of network capabilities and infrastructure.

REFERENCES

Bennet, S., Priest, A., & Macpherson, C. (1999). Learning about online learning: An approach to staff development for university teachers. *Australian Journal of Educational Technology*, *15*(3), 207-221.

Brown, T.H. (2005). Towards a model of m-learning in Africa. *Journal of Educational Multimedia and Hypermedia*, *4*(3), 299-316.

Bruner, J. S. (1985). Models of the learner. *Educational Researcher*, *14*(6), 5-8.

Conlon, T., & Simpson, M. (2003). Silicon Valley versus Silicon Glen: The impact of computers upon teaching and learning: A comparative study. *British Journal of Educational Technology*, *34*(2), 137-150.

Cuban, L. (1986). *Teachers and machines: The classroom use of technology since 1920*. New York: Teachers College Press.

Evans. (2005). *Potential uses of wireless and mobile learning.* Retrieved December 7, 2007, from http://www.jisc.ac.uk/eli_proj_landscape.html

Hammond, M. (1994). Measuring the impact of IT on learning. *Journal of Computer Assisted Learning*, *10*, 251-260.

Jackson, B., & Anagnostopoulou, K. (2000). *Making the right connections: Improving quality in online learning. Teaching and learning online: New pedagogies for new technologies*. Retrieved December 7, 2007, from http://webfeedback.mdx.ac.uk/_lmlseminar/_private/_abstract14/finland.htm

Jonassen, D.H., Peck, K.L., & Wilson, B.G. (1999). *Learning with technology*. Upper Saddle River, NJ: Merrill Publishing.

Kukulska-Hulme, A., & Traxler, J. (Eds.). (2005). Mobile learning: A handbook for educators and trainers. London: Routledge.

Laouris, Y., & Eteokleous, N. (2005). We need an educationally relevant definition of m-learning. In *Paper presented at the M-Learn Conference*, Cape Town, South Africa. Retrieved December 7, 2007, from http://www.mlearn.org.za/papers-full.html

Laurillard, D. (1993). *Rethinking university teaching: A framework for the effective use of educational technology.* London and New York: Routledge.

Nichol, J., & Watson, K. (2003). Editorial: Rhetoric and reality—the present and future of ICT in education. *British Journal of Educational Technology, 34*(2), 131-136.

Oblinger, D. G. (2004). The next generation of educational engagement. *Journal of Interactive Media in Education,* (8). Retrieved December 7, 2007, from http://www-jime.open.ac.uk/2004/8/oblinger-2004-8-disc-paper.html

Parsons, D., & Ryu, H. (2006). *A framework for assessing the quality of mobile learning*. Retrieved December 7, 2007, from www.massey.ac.nz/~hryu/M-learning.pdf

Piaget, J. (1952). *The origins of intelligence in children*. New York: International Universities Press.

Rettie, R. (2003). Connectedness, awareness and social presence. In *Proceedings of the 6th International Presence Workshop*, Aalborg. Retrieved December 7, 2007, from http://www.kingston.ac.uk/%7Eku03468/docs/Connectedness,%20Awareness%20and%20Social%20Presence.pdf

Strommen, D. (1999). C*onstructivism, technology, and the future of classroom learning.* Retrieved December 7, 2007, from http://www.ilt.columbia.edu/ilt/papers/construct.html

Tearle, P., Dillon, P., & Davis, N. (1999). Use of information technology by English university teachers. Developments and trends at the time of the National Inquiry into Higher Education. J*ournal of Further and Higher Education, 23*(1), 5-15.

Thomas, M. (2006a, May). Book review: M-learning. Mobile learning and performance in the palm of your hand. *The knowledge tree* (9th ed.). Retrieved December 7, 2007, from http://education.guardian.co.uk/elearning/comment/0,10577,1490476,00.html

Thomas, M. (2006b). E-learning on the move. *Education Guardian, 23*, May 2005. Retrieved December 7, 2007, from http://education.guardian.co.uk/elearning/comment/0,10577,1490476,00.html

Torrisi-Steele, G. (2004). Toward effective use of multimedia technologies in education. In S. Mishra & R.C. Sharma (Eds.), *Interactive multimedia in education and training* (pp. 25-46). Hershey, PA: Idea Group.

Torrisi-Steele, G. (2006). The making of m-learning spaces. In Paper presented at the Queensland University of Technology OLT2006 Conference, Brisbane, Australia.

Tse-Kian, K.N. (2003). Using multimedia in a constructivist learning environment in the Malaysian classroom. *Australian Journal of Educational Technology, 19*(3), 293-310.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.

Zurita, G., & Nussbaum, M. (2004). A constructivist mobile learning environment supported by a wireless handheld network. *Journal of Computer Assisted Learning, 20*, 235-243.

KEY TERMS

Constructivism: A philosophy for teaching and learning based on the notion of individuals generating their own understanding of the world in "their own way." Constructivist theories of learning espouse learner-centred approaches that actively engage the learner in order to facilitate meaningful learning**.** Key theorists include Piaget and Bruner.

Connectedness: The feeling of "being in touch." "It is an emotional experience invoked by, but independent of others' presence (Rettie, 2003). Feelings of connectedness can range from general awareness of the presence of others (e.g., being online) to stronger awareness of social presence (e.g., exchange of text messages).

Meaningful Learning: Achieving a deep understanding of complex ideas**.** Meaningful learning implies that knowledge can be manipulated and applied to a variety of situations and contexts.

Mobile Devices: Any small portable or handheld computing devices technically capable of connectivity (ideally wireless) to each other, other devices or Internet. Includes laptop computers, tablet PCs, PDAs, mobile phones, smart phones and MP3 players.

Mobile Learning (M-Learning): The integrative use of mobile devices into the curriculum in order to facilitate active and meaningful learning through the creation of learning spaces which extend outside the physical and temporal constraints of the traditional classroom.

M-Learning Space: The learning environment within which learning activities are enriched by the use of mobile devices. M-learning spaces are supportive of meaningful learning. M-learning spaces are characteristically dynamic, collaborative and focused on individual learner needs in the current context.

Socioconstructivism: Like constructivism active, learner-centred learning is emphasised**.** Interaction and communication are considered important for engaging in the active construction of knowledge through a process of negotiation**.** Collaborative learning is an important strategy within this view. Socioconstructivism is becoming an important guiding philosophy for the design of m-learning spaces.

Ubiquitous: Present and available anywhere, anytime**.** Used in computing to refer to access to computing technology away from the constraints of the desktop**.** M-learning is characterised by being ubiquitous.

Peer-to-Peer Computing

Manuela Pereira
University of Beira Interior, Portugal

INTRODUCTION

The term peer-to-peer (P2P) was originally used to refer to network protocols where all the nodes had the same role and there were no nodes with specific responsibilities to act as the administrators or supervisors of a network (Ye, Makedon, & Ford, 2004). However, with the evolution of Internet as the dominant architecture for applications, contents, and services, applications and services have gradually migrated from the client-server paradigm to the edge services paradigm and now to the P2P computing paradigm. Therefore, nowadays, the term P2P refers to a class of systems and applications that use distributed resources to perform some function in a decentralized manner, where every participating node can act as both a client and a server (Ye et al., 2004).

This article provides an overview of P2P computing, being focused on the types of multimedia distribution services and cooperation models in P2P systems. These models are classified regarding the functionality, the degree of decentralization, and the degree of structure of the information system.

BACKGROUND

The P2P Working Group (http://www.peer-to-peerwg.org), a consortium for the development of P2P technology, defines P2P as the sharing of computer resources by direct exchange.

P2P systems have advantages regarding client-server systems, namely: (1) improved scalability and reliability since they avoid the dependency of centralized servers, which are often points of failure; (2) cheaper infrastructures due to direct communication among peers; and (3) easiness of resource aggregation in order to provide, for instance, massive processing power (Ye et al., 2004). However, P2P systems also have some drawbacks namely considerably more complex searching and node organization and security issues (Aberer, Punceva, Hauswirth, & Schmidt, 2002).

P2P networks have been deployed in several application areas, such as distributed grid computing (http://www.entropia.com), storage (Cohen, 2003), Web cache (Dabek, Kaashoek, Karger, Morris, & Stoica, 2001), and service directory (Iyer, Rowstron, & Druschel, 2002; Ratnasamy, Francis, Handley, Karp, & Shenker, 2001; Stoica, Morris, Karger, Kaashoek, & Balakrishman, 2001). However, P2P systems were popularized due to the applications of file sharing: Many different P2P file sharing systems, such as Gnutella (http://www.gnutella.com), KaZaA (http://www.kazaa.com.), eDonkey (http://www.overnet.com), and BitTorrent (http://bitconjurer.org/bittorrent/) have recently experienced dramatic growth in popularity and are currently responsible for a large amount of the Internet traffic (Saroiu, Gummadi, & Gribble, 2002; SD-NAP, 2002). As a result of the increasing popularity, P2P file sharing systems became more complex in order to provide services to millions of users. The original centralized architecture of Napster (http://www.napster.com) has been replaced by unstructured decentralized systems such as Freenet (http://freenet.sourceforge.net) and Gnutella. A detailed performance evaluation of the main features of current unstructured P2P architectures may be found in Benevenuto, Ismael, and Almeida (2004). Due to scalability limitations of the unstructured P2P approaches, structured P2P systems have been developed to manage huge amounts of data in a scalable way in overlay networks. One type of structured P2P systems is Distributed Hash Tables (DHTs) (Rieche, Wehrle, Landsiedel, Gotz, & Petrak, 2004). Examples of these DHTs include Chord (Stoica et al., 2001), Content-Addressable Network (CAN) (Ratnasamy et al., 2001), DKS(N,k,f) (Alima, El-Ansary, Brand, & Haridi, 2003), or Pastry (Rowstron & Druschel, 2001).

MULTIMEDIA DISTRIBUTION SERVICES

The demand of delivering multimedia content over the Internet has become increasingly high for scientific, educational, entertainment, and commercial applications. However delivering streaming media content over best effort, packet-switched networks has to deal with high bit rates, delay, loss sensitivity, and heterogeneous client resources. The expensive growth of multimedia applications over the Internet lead to an increasing interest to provide low cost, efficient, and scalable multimedia distribution services. Recently, P2P systems have received a great amount of interest as a promising scalable and cost-effective solution for next-generation multimedia content distribution.

Multimedia distribution services may be classified into three categories (Xiang, Zhang, Zhu, Zhang, & Zhang, 2004) as follows:

1. **Centralized Multimedia Distribution:** A centralized multimedia server is deployed to support a client to access multimedia content across the Internet. In order to extend the storage and Input/Output capacity of the centralized server and to improve the service availability, server clustering or mirroring are often used. This strategy is widely used in traditional Web-based distribution services, although they are unable to reduce the network bottleneck problem, which has significant impacts on the performance of multimedia distribution. As reported by Kangasharju, Roberts, and Ross (2002), a centralized system is not suitable neither scalable for multimedia distribution services. The use of proxy caches (Xiang, Zhang, Zhu, & Zhong, 2001; Zhang, Wang, Du, & Su, 2000) can alleviate the bottleneck problem by caching popular contents from origin servers to proxy servers located at the edge of network. Clients receive the content from edge servers without consuming the network bandwidth. However, the cache policies that influence the effectiveness of proxy caching are suboptimal for streaming media since they were not developed with new video coders in mind. Moreover, proxy caching has scalability limitations for multimedia distribution services.
2. **Multimedia Distribution Based on Content Distribution Networks:** This technique is a server-oriented approach based on content distribution networks (CDNs) (also known as content delivery networks) platforms. The original server is replicated and placed locally or remotely in geographical or network spaces. CDN-based architectures have a limited performance for large-scale multimedia distribution services, since the capacity of the edge server is not large enough to support multimedia services, specially the streaming media service. Furthermore, the decision of the number and location of edge servers is a difficult problem, which has not yet been solved efficiently (Chen, Katz, & Kubiatowicz, 2002; Cohen, Katzir, & Raz, 2002; Qiu, Padmanabhan, & Voelker, 2001)
3. **P2P Networks:** In P2P networks, clients host contents in their local storage and distribute contents to other clients, allowing the sharing of data and resources by a large community at low cost and small network management. Furthermore, the availability of distribution services relies on the reliability of each peer, but peers may not guarantee service persistence. Some current P2P systems also have scalability limitations such as Napster, CenterSpan (http://www.centerspan.com), and Vtrails (http://www.vtrails.com), which are centralized. However, new P2P scalable frameworks have also been developed, and P2P-based multimedia distribution services have started to appear and are considered as the most scalable, efficient, and low-cost solution for future multimedia applications and services. The next section is devoted to P2P systems.

P2P COMPUTING

Since there are about 70 different P2P applications, this section provides an overview of those P2P applications regarding their models of cooperation and how they may be classified according to the following criteria: functionality, degree of decentralization, and degree of structure of the information system.

Functional Classification

P2P systems may be classified from a functional point of view into three basic subcategories: (1) management and contents-sharing applications; (2) distributed processing and; (3) collaboration and communication (Benayoune & Lancieri, 2004). However, there are also platforms, such as JXTA (Gong, 2001), and Globus, that aim at facilitating the development of these applications by offering a set of common basic services such as the authentication or research and routing services. Table 1 summarizes this classification. The file-sharing applications are extremely popular on the Internet and have a large user base. Recent statistics show that the activities of these applications consume more than 60% of Internet service provider (ISP) traffic.

Degree of Decentralization

The Internet is nowadays largely based on the client-server paradigm but the use of central servers leads to a waste of

Table 1. Functional classification of P2P systems

Management and Contents Sharing Applications	Distributed Processing	Collaboration and Communication
Napster	Seti@Home	Groove,
Audiogalaxy, GNUtella	Genome@Home	NextPage,
KaZaA	Folding@Home	Kanari,
Grokster	Evolutionary@Home	Magi,
Morpheus	XPulsar@home	Jabber,
Blubster	Life Mapper	AIMster,
DirectConnect	ChessBrain	MSN,
BitTorrent	FightAIDS@Home	AOL Chat,
Freenet	Avaki	NetMeeting
Aimster	Jivalti	
IMesh	Axceleon	
EMule	Entropia	
eDonkey2000	GridSystems	
OpenNap (WinMX)		
LimeWire		
Shareaza		
XoLoX		
Chord		
Tapestry		
Pastry		
Tornado		
CAN		

resources, creates bottlenecks, and is very sensitive to failures in the server. P2P systems increase the scalability and fault tolerance due to the decentralization. P2P systems may be classified into three categories regarding the degree of decentralization (Benayoune & Lancieri, 2004; Tsoumakos & Roussopoulos, 2003):

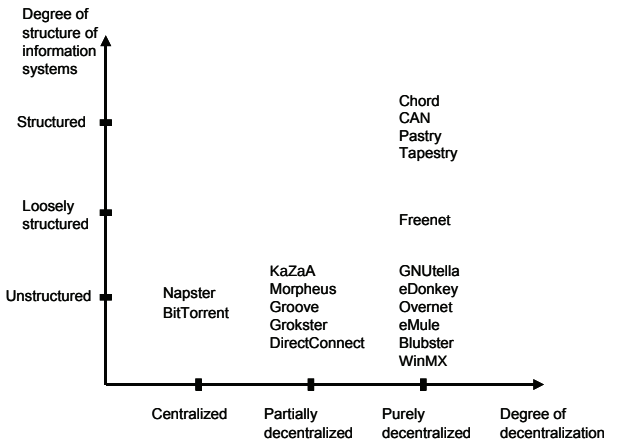
1. **Purely Decentralized Systems:** These systems, also called pure P2P, became popular with GNUtella, Freenet, Plaxton, Blubster, and so forth. The nodes within such systems communicate together without any intermediate central point. In fact, each node termed Servent (**server** and **client** at the same time) performs research and routing functions. This model is more robust than the centralized one because the failure of any particular node does not have impact on the system, resulting in high service availability at reduced costs. However, this model has two main drawbacks. First, the localization of an object is not guaranteed because of the directory decentralization. Second, the mechanism of research by flooding (set of broadcast) wastes high amounts of network bandwidth. In fact, the increase of the number of the peers generates an exponential increase of requests, which may cause network congestion, slow down downloads and poses scalability problems.
2. **Partially Decentralized Systems:** A new wave of P2P systems is born from a purely decentralized system combined with a centralized one (e.g., FastTrack: KaZaA, Morpheus, Grokster, iMech, etc.). This model is based on a Super-Node (or Super Peer) that acts as a centralized server for a set of nodes. The nodes send their queries to the Super-Node that provide the content or relay requests to other nodes. Such a system solves the problem of networks extensibility of purely decentralized systems by keeping the efficiency of the centralized ones. In case of failures of the primary Super-node, others are defined to automatically replace it. However, the process of selection of the Super-Nodes is still complex. They can be elected according to their own capacities, in particular in terms of bandwidth and persistence in the system (time of connection), but each user can decide to be or not to be a Super-Node. Moreover, being a Super-Node can lead to increased security problems.
3. **Centralized Systems:** In this model a central server is responsible for answering the queries, hence all the query traffic is directed to it (e.g., Napster, BitTorrent, etc.). Before using the network, a peer must connect to a central server, which manages a central directory of shared resources and users that indicates from which node the files should be downloaded. This mode is a very efficient way to locate resources and to have a complete view of the network. However, although shared files are more easily managed, there is a single access point which may lead to several problems such failure, overload, copyright, or security. Moreover, the central server limits the network scalability.

Structure of the Information System

P2P systems use several techniques of references publication, contents search, and routing requests. Therefore, it is possible to build another classification of P2P systems according to the mode of management of meta-information in order to discover distributed data or evaluate its level of accessibility. P2P systems may be classified into three categories regarding the degree of structure of the information system (Benayoune & Lancieri, 2004; Tsoumakos & Roussopoulos, 2003):

1. **Unstructured Systems:** In some P2P systems such as GNUtella, KaZaA, Morpheus, DirectConnect, BitTorrent, and so forth, nodes have no information about the location of the required files. Therefore, the research will be random and the used discovery methods are known as blind. Nodes relay requests from node to node towards all the neighbors in order to seek the maximum of objects on the maximum of nodes. Thus, the network will be flooded by duplicated messages and the partial answers will be limited to a zone of locality defined by the time-to-live (TTL) field.
2. **Structured Systems:** The hash-based systems such as Chord (Stoica et al., 2001), CAN (Ratnasamy et al., 2001), Tapestry (Zhao et al., 2004), or Tornado are supposed to correct the lack of scalability of the prior systems. The objective is to add more dynamics to P2P network by proposing at the same time an algorithm of localization and routing in an entirely distributed environment. In these kinds of networks, the files or their references are placed in quite precise places. Node and files have identifiers independent of their localization and contents semantics. Each node of the network has a routing table containing two parts: (1) set of neighbors and (2) set of pointers towards the nodes. The major weakness of these systems is that it is difficult to maintain the structure in a context of a changing population where users often join and leave the system.
3. **Loosely Structured Systems:** P2P networks like Freenet can be classified between the two preceding types of P2P systems. Indices are provided on the localization of the files to select the next peer to query. Therefore, the searches will be guided but without guaranties of success. Freenet is also fault tolerant due to the strategy for the replication of files, since every

Figure 1. Overview of the classification of P2P systems regarding both the degree of decentralization and the degree of structure of the information system



time a file is requested it is replicated in the neighborhood of nodes that frequently make requests.

Figure 1 presents an overview of the last two classifications.

FUTURE TRENDS

Novel frameworks for multimedia distribution services based on P2P networks are now starting to appear. Most of them are based on new multimedia coding technologies that allow to reach scalability, robustness, and to circumvent frequent peer going downs (Pouwelse, Taal, Lagendijk, Epema, & Sips, 2004; Xu, Wang, Panwar, & Ross, 2004). Some of these techniques are layered coding and/or multiple description coding (MDC). Layer coding is usually used to construct a scalable framework, while MDC can also provide adaptivity to packet loss and variable transmissions (Pereira, Antonini, & Barlaud, 2003). Coopnet (Padmanabhan, Wang, Chou, & Sripanidkulchai, 2002) considers MDC which enables graceful QoS degradation when packet losses occur due to node or link failure or link state change.

Although most P2P systems do not yet take security into account, security and trust are key issues in P2P computing. Besides the occurrence of improper handling by users of sensitive personal information, a number of security attacks on P2P systems are possible including denial-of-service attacks, replay attacks, collaborated or un-collaborated attacks by malicious nodes, incorrect routing updates, black hole attacks (modifying the routing message to say a node has the shortest path to some nodes), and worm hole attacks (collusion by two malicious nodes to make the packets they want flow through them). In addition, management of trust information of nodes and access control to resources are two important issues that need to be addressed (Sieka, Kshemkalyani, & Singhal, 2004).

CONCLUSION

Multimedia distribution services have been discussed and P2P systems were identified as the most scalable, efficient, and low-cost solution for future multimedia applications and services. These systems were analyzed and classified regarding the functionality, the degree of decentralization, and the degree of structure of the information system. The improvements of the new generation of P2P frameworks were tackled. These new frameworks pretend to be scalable, robust, reliable, and secure. Some of the techniques used to reach each one of these goals were discussed.

REFERENCES

Aberer, K., Punceva, M., Hauswirth, M., & Schmidt, R. (2002). Improving data access in P2P systems. *IEEE Internet Computing, 6*(1), 58-67.

Alima, L. O., El-Ansary, S., Brand, P., & Haridi, S. (2003). DKS(N, k, f): A family of low communication, scalable and fault-tolerant infrastructures for P2P applications. *Proceedings of The 3rd International Workshop on Global and P2P Computing on Large Scale Distributed Systems (CCGRID 2003),* Tokyo, Japan.

Benayoune, F., & Lancieri, L. (2004). Models of cooperation in peer-to-peer networks, a survey. In M. Freire, P. Chemouil, P. Lorenz, & A. Gravey (Eds.), *Universal multiservice networks. Lecture Notes in Computer Science* (pp. 327-336). Springer.

Benevenuto, F., Ismael, J., Jr., & Almeida, J. (2004). Quantitative evaluation of unstructured peer-to-peer architectures. *Proceedings of the 2004 International Workshop on Hot Topics in Peer-to-Peer Systems (HOT-P2P 2004)*.

Chen, Y., Katz, R., & Kubiatowicz, J. (2002). Dynamic replica placement for scalable content delivery. *Proceedings of First International Workshop on Peer-to-Peer Systems (IPTPS 2002)* (pp. 306-318).

Cohen, B. (2003). *Incentives build robustness in BitTorrent*.

Cohen, R., Katzir, L., & Raz, D. (2002). Scheduling algorithms for a cache prefilling content distribution network. *Proceedings of IEEE InfoCom 2002.*

Dabek, F., Kaashoek, M. F., Karger, D., Morris, R., & Stoica, I. (2001). Wide-area cooperative storage with CFS. In *Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP 2001),* Chateau Lake Lou, Banff, Canada.

Gong, L. (2001). JXTA: A network programming environment. *IEEE Internet Computing, 5*(3), 88-95.

Iyer, S., Rowstron, A., & Druschel, P. (2002). Squirrel: A decentralized peer-to-peer Web cache. *Proceedings of ACM Symposium on Principles of Distributed Computing (PODC 2002),* Monterey, CA.

Kangasharju, J., Roberts, J., & Ross, K. W. (2002). Object replication strategies in content distribution networks. *Computer Communications, 24*(4), 367-383.

Padmanabhan, V. N., Wang, H. J., Chou, P. A., & Sripanidkulchai, K. (2002). Distributing streaming media content using cooperative networking. *Proceedings of NOSSDAV,* Miami Beach, FL.

Pereira, M., Antonini, M., & Barlaud, M. (2003). Multiple description coding for Internet video streaming. *Proccedings of IEEE International Conference on Image Processing (ICIP 2003),* Barcelone, Spain.

Pouwelse, J. A., Taal, J. R., Lagendijk, R. L., Epema, D. H. J., & Sips, H. J., (2004). Real-time video delivery using peer-to-peer bartering networks and multiple description coding. *IEEE International Conference on Systems, Man and Cybernetics.*

Qiu, L., Padmanabhan, V., & Voelker, G. (2001). On the placement of Web server replicas. *Proceedings of IEEE Infocom 2001* (pp. 1587-1596).

Ratnasamy, S., Francis, P., Handley, M., Karp, R., & Shenker, S. (2001). A scalable content-addressable network. *Proceedings of the 2001 ACM SIGCOMM Conference,* San Diego, CA.

Rieche, S., Wehrle, K., Landsiedel, O., Gotz, S., & Petrak, L. (2004). Reliability of data in structured peer-to-peer systems. *Proceedings of the 2004 International Workshop on Hot Topics in Peer-to-Peer Systems (HOT-P2P'04).*

Rowstron, A., & Druschel, P. (2001). Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. *Proceedings of IFIP/ACM International Conference on Distributed Systems Platforms (Middleware),* Heidelberg, Germany.

Saroiu, S., Gummadi, P., & Gribble, S. (2002). A measurement study of peer-to-peer file sharing systems. *Proceedings of Multimedia Computing and Networking (MMCN 2002),* San Jose, CA.

SD-NAP. (2002). T*op applications (bytes) for subinterface: Sd-nap traffic, in CA/DA workload analysis of SD-NAP data.* Retrieved from http://www.caida.org/analysis/workload/byapplication/sdnap/index.xml

Sieka, B., Kshemkalyani, A. D., & Singhal, M. (2004). On the security of polling protocols in peer-to-peer systems. *Proceedings of the Fourth International Conference on Peer-to-Peer Computing (P2P04).*

Stoica, I., Morris, R., Karger, D., Kaashoek, M., & Balakrishman, H. (2001). Chord: A scalable peer-to-peer lookup service for Internet applications. *Proceedings of ACM SIGCOMM,* San Diego, CA.

Tsoumakos, D., & Roussopoulos, N. (2003). Adaptive probabilistic search for peer-to-peer networks. *Proceedings of the Third International Conference on Peer-to-Peer Computing (P2P'03).*

Xiang, Z., Zhang, Q., Zhu, W., Zhang, Z., & Zhang, Y. (2004). Peer-to-peer based multimedia distribution service. *IEEE Transactions on Multimedia, 6*(2), 343-355.

Xiang, Z., Zhang, Q., Zhu, W., & Zhong, Y. (2001). Cost-based replacement policy for multimedia proxy across wireless Internet. *Proceedings of IEEE Global Telecommunications Conference (Globecom 2001),* San Antonio, TX.

Xu, X., Wang, Y., Panwar, S. S., & Ross, K. W. (2004). A peer-to-peer video-on-demand system using multiple description coding and server diversity. *Proceedings of IEEE Inter-national Conference on Image Processing (ICIP 2004),* Singapore.

Ye, S., Makedon, F., & Ford, J. (2004). Collaborative automated trust negotiation in peer-to-peer systems. *Proceedings of the Fourth International Conference on Peer-to-Peer Computing (P2P 2004).*

Zhang, Z.-L., Wang, Y., Du, D. H. C., & Su, D. (2000, August). Video staging: A proxy server-based approach to end-to-end video delivery over wide-area networks. *IEEE/ACM Transactions on Networking, 8,* 429-442.

Zhao, B. Y., Huang, L., Stribling, J., Rhea, S. C., Joseph, A. D., & Kubiatowicz, J. (2004). Tapestry: A resilient global-scale overlay for service deployment. *IEEE Journal on Selected Areas in Communications, 22*(1), 41-53.

KEY TERMS

Distributed Hash Table (DHT) Scheme: The basic idea of a DHT scheme is to use a hash table-like interface to locate the objects and to distribute the duty of maintaining

the hash table data structure, in the face of node joins/leaves, to all participating P2P nodes.

Gnutella: Gnutella is a decentralized file-sharing system whose participants form a virtual network and communicate peer-to-peer via the Gnutella protocol for distributed file search. To participate, a peer first connects to a known Gnutella host. Upon receiving a message, the servent decrements the time-to-live (TTL) field of the message. If the TTL is greater than 0 and the servent has never seen the identifier of the message (loop detection), it resends the message to all known peers. The servent also checks whether it should respond to the message. If it receives a Query, for example, it checks its local file store and responds with a QueryHit if it can satisfy the request. Responses are routed along the same path as the originating message.

JXTA: JXTA is a suite of protocols that facilitates P2P communication. It provides a common network architecture layer for a wide variety of network services and applications. Peers on a JXTA network are arbitrary devices and can be anything with an electronic heartbeat.

JXTA Search: JXTA search is a decentralized peer-to-peer search engine. It provides a common distributed query interface for peer devices, exposing services and content for a network of information providers and consumers.

Peer, Party, or Node: These terms are used interchangeably and all refer to an entity in a P2P system.

Peer-to-Peer (P2P): This term refers to any exchange system characterized by direct interaction and data exchange between its entities (called nodes or peers).

Servent: This word is composed by the first four and last three letters of **server** and **client**. It is used to designate a node that is server and client at the same time.

Performance Implications of Pure, Applied, and Fully Formalized Communities of Practice

Siri Terjesen
Queensland University of Technology, Australia
Max Planck Institute of Economics, Germany

INTRODUCTION

Interest in knowledge-based perspectives on the firm has grown in both practitioner and academic realms, spurred by management bestsellers such as Senge's *Fifth Discipline* (1990) and the acknowledgement that intangible assets are key to the firm's sustainable competitive advantage. Knowledge management tools and processes are used by organizations to identify, create, represent, and distribute knowledge for reuse, awareness, and learning. One component of knowledge management is the "communities of practice" (CoPs) concept. CoPs are informal networks of individuals who possess various levels of a common capability and apply their knowledge in pursuit of a similar endeavor (Brown & Duguid, 1991). For example, Xerox technicians solve problems by relying on informal communication with colleagues in addition to formal user user manuals. Created as a response to bureaucratization, CoPs emerge from individuals' passions for a particular activity and the term is used to describe a formal of organization that is distinct from traditional formal boundaries around geographic and functional business units or other institutional affiliations and divisions.

For the most part, managers use the CoP concept to encourage informal, situated learning (e.g., Hildreth & Kimble, 2004). However, some managers developed highly formalized structures with regulated membership, prescribed roles, scheduled meetings, and technical tools. This formalization distorts the original concept—that CoPs are created as a response to bureaucracy and are, by definition, emergent. The formalization of CoPs defeats both the original intent and the ability to reap full benefits for the firm. The chapter reviews three models of communities of practice — pure, applied, and formalized — and explores how coordination, opportunity, and knowledge flow costs in formalized CoPs can impede organizational performance.

BACKGROUND

In practice, CoPs can take three forms. In the first "pure" case, the original construct is adhered to and CoPs are emergent in nature. A second, mid-spectrum "applied" group exists when original CoP theory has been slightly tweaked. Next, a mid-spectrum "applied" group in which original CoP Theory has

Table 1. CoPs: Pure, applied, and fully formalized constructs

	CoP: Pure Construct	CoP: Applied Construct	CoP: Fully Formalized Construct
Organization Type	Emergent Community	Supported Community	Formalized Community
How is the CoP born?	Emergent, from individuals' passion, bottom-up	Emergent, especially in firm-enabled spaces	Emergent, but with strong top-down directives
Who are CoP members?	Self-selected individuals; choice	Both self-selected and strongly encouraged by others	Corporate assignations; restricted membership
How many CoP members?	Small core group	Small to medium-sized group	Small to large group
What is the goal of the CoP?	Learn and share knowledge about passionate individual interest	Share knowledge about area of strong individual interest that the firm also deems interesting	Share knowledge about area of interest that the firm also deems especially interesting
Who is in charge?	Individuals	Individuals and organization	Organization and individuals
What holds CoP together?	Shared interest and passion	Interest oriented to project goal	Some interest, also job requirement
Where is resource level?	Individual's own time	Individual time, some funding at various organization levels	Funding at various organization levels, especially corporate
When are CoP interactions?	Spontaneous interactions	More regular, but spontaneous interactions possible	Scheduled meetings; spontaneous interactions if time
What type of learning?	Situated	Situated and classroom	Classroom
When does the CoP die?	Naturally, when interest fades	When project completed	When firm resources extinguished

been slightly tweaked. For example, a firm might establish a common area such as a water cooler, coffee pot, or plate of cookies where individuals meet spontaneously. In the third fully "formalized" case, CoPs are no longer chiefly fueled by individual passion, but rather by organizational mechanization. For example, spontaneous get-togethers are supplemented by set monthly meetings and agendas. Thus formally recognized CoPs function just like any other formal unit within the firm. The shift from the original pure intent to applied (in which there are some costs and benefits) and fully formalized (only negatives) construct is depicted in Table 1.

Pure Construct

Lave and Wenger (1991) formally coined the term CoPs, which was later incorporated into an organizational framework by Brown and Duguid (1991). CoP theory is based on the value of informal structures to organizational development, learning, and performance (Barnard, 1938, among others), epistemological perspectives on the importance of tacit and action-oriented knowledge (Polanyi, 1966), and the key role of situated learning, social processes (March & Olsen, 1975), and community (Daft & Weick, 1984).

As emergent organizations, CoPs encourage informal situated learning that is unobtainable in a structured organizational bureaucracy. Brown and Duguid (1991) note this difference: "Work practice and learning needs to be understood not in terms of the groups that are ordained (e.g., 'task forces' or 'trainees') but in terms of the communities that emerge" (p. 49). Wenger (1998) also notes: "Unlike more formal types of organizational structures, it is not so clear where [CoPs] begin and end…Whereas a task force or a team starts with an assignment and ends with it, a community of practice may not congeal for a while after an assignment has started, and it may continue in unofficial ways far beyond the original assignment" (p. 96).

CoPs extend beyond traditional classrooms to work environments, hobbies, and families (Lave & Wenger, 1991; Wenger, 1998). Individuals become members of CoPs through narration, social construction, and collaboration (Brown & Duguid, 1991). Narration involves the telling of stories and encourages individuals to develop a socially constructed world. Through collaboration, individuals learn from one another. Strong communities are characterized by trust and a sense of identity and belonging. Knowledge transfer is both "leaky" within and "sticky" across communities (Brown & Duguid, 2001). CoPs are distinct from (but are subsets of) large groups which perform similar activities but are not in direct contact. These groups have been variously termed "networks of practice" (Brown & Duguid, 1991), "occupational groups" (Van Maanen & Barley, 1984), "social worlds" (Strauss, 1978), or "constellations of practice" (Wenger, 1998). In these large networks, individuals share knowledge and practice, but are unknown to one another except through Web sites, listservs, or other indirect communication.

Applied and Corrupted Construct

In *Cultivating Communities of Practice*, Wenger, McDermott, and Snyder (2002) provide a gardening analogy for these emergent organizations:

A plant does its own growing, whether its seed was carefully planted or blown into place by the wind. You cannot pull the stem, leaves, or petals to make a plant grow faster or taller. However, you can do much to encourage healthy plants: till the soil, ensure they have enough nutrients, supply water, secure the right amount of sun exposure, and protect them from pests and weeds. There are also a few things we know not to do, like pulling up a plant to check if it has good roots. Similarly, some communities of practice grow spontaneously while others may require careful seeding. (p. 12-13)

In formalizing CoPs, some practitioners have tugged, over-watered, or otherwise too zealously attended to these emergent communities.

The mid-spectrum group can be described as "applied," the gray area in which original CoP theory has been slightly tweaked. The organization may establish systems enabling CoPs to emerge naturally. For example, a firm might establish a common area such as a water cooler, coffee pot, or plate of cookies where individuals can meet spontaneously. In the garden analogy, this is the equivalent of cultivating the soil by adding nutrients.

In the corrupted construct, CoPs are no longer chiefly fueled by individual passion, but by organizational mechanization. During this process, pure CoP theory is mutated into prescriptive formulas bureaucratizing these emergent communities. For example, some management consultants advocate regulated membership, prescribed roles, scheduled meetings, and even distribute CoP-printed pins and pens to identify members. Spontaneous get-togethers are supplemented by set monthly meetings and agendas. Once CoPs are formally recognized, they become just like any other formal unit within the firm.

An example of formalized CoPs existed at the Fairfax, Virginia-based global consultancy, American Management Systems (AMS). In the late 1990s, then-CEO Charles Rossotti asked business units to nominate "thought leaders" who were then mandated to establish CoPs. AMS paid for two to three weeks per year of the leaders' time. CoP membership was a privilege and extended only to those individuals recognized as "experts" by their managers. Every CoP member was required to write one knowledge white paper per year. Business units funded participation, meeting attendance, projects, and an annual conference with members of all CoPs. At one point, 900 of AMS' 9,000 employees were members of one

of the CoPs. AMS estimated that these efforts saved $2 to $5 million per year. In the January/February 2000 *Harvard Business Review,* Wenger and Snyder highlighted AMS as an example of CoP success. Barely six months after the article was published, AMS completely disbanded CoPs in favor of a more flexible organization enabling individuals to work across different areas. AMS' Director of Knowledge Management Susan Hanley moved to another consultancy, and most of the identified CoP leaders are also no longer with the firm.

MAIN FOCUS: HOW CAN FORMALIZATION DECREASE PERFORMANCE?

Formalization describes the degree of activity bureaucratization such as the number of scheduled meetings, extent of reporting, regulation of member selection and retention, and mandated linkages to firm hierarchy (such as division reporting relationships). The decrease in organizational performance of CoP formalization can be attributed to hefty coordination, opportunity, and knowledge flow costs.

Coordination Costs

In complex environments, organizations require more integration and coordination efforts such as rules, plans, and mutual adjustment. Rational organizations create hierarchies to alleviate organization uncertainty from complex coordination and interdependence requirements, lower communication costs, and align interests. If interdependence is not limited by this organization, remaining coordination problems are assigned to community, task, project, or other self-contained teams. While bureaucracy does minimize coordination costs, CoPs emerged as a rebellion against bureaucracies. CoP members accept entropy in the system because perceived benefits of this workaround outweigh costs of dealing with the hierarchy. For example, CoP formalization reduces the costs of interacting with a given set of people, but individuals may still need to interface with others informally. Formalization of CoP routines may upset the internal CoP and drive members to look outside the CoP for new knowledge-oriented routines. Finally, CoPs are a social context characterized by spontaneous interactions that have naturally low costs. Additional formalization becomes unnecessary. In formalized CoPs, individuals would continue to seek social contexts outside of the formal structure, creating additional coordination costs.

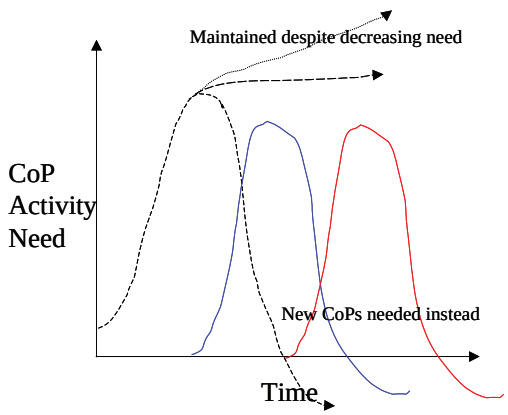
Opportunity Costs

Organizations have a limited amount of resources available for internal distribution and use. As formalized entities, CoPs receive resources such as dedicated employee time, physical space, and technical tools. These resources may be budgeted for an entire fiscal year, even if the need only exists for the next two to three weeks or months. Meanwhile, the budget process may not be able to aid CoPs that emerge outside the fiscal cycle. When reinforced with resources, formalized CoPs are made artificial and do not follow the normal course of spontaneous birth, growth, maturity, decline of usefulness, and death. For example, a firm may formalize a CoP dedicated to Java programming when individual interests (and perhaps also future competitiveness) have shifted to a new technology such as ActiveX. Firm resources would maintain the old CoP, even in the face of decreasing need. Meanwhile, the ActiveX CoP could not emerge because individuals are unable to access key resources. There are dire opportunity costs for individuals and the firm. Figure 1 illustrates how a formalized CoP is not able to follow a normal lifecycle. In the normal cycle of a CoP, resource activity needs decrease over time (Gongla & Rizzuto, 2004). In the case of a formalized CoP, however, resources are maintained despite decreasing need and the organization is perpetuated. Meanwhile, new CoPs are not able to gain the resources they need.

Furthermore, human resource requirements may pull individuals away from the actual work at hand and push them into community maintenance roles. A recent study examining 10 different formal and informal CoP roles found that members, depending on their roles, spent between 5.0% (mentor) and 53.4% (facilitator) of their time on community activities. Across all 10 roles (subject matter expert, core team member, community member, community leader, sponsor, facilitator, content coordinator, mentor, administrative/events coordinator, and technologist), each CoP member spent an average of 20% of her or his time on community activities (Fontaine, 2001).

Finally, an organization might formalize a CoP and effectively kill it, deterring other emergent CoPs. Individual and organizational resources could be devoted to other activities,

Figure 1. Formalized CoP receives resources despite decreasing need



such as the emergence of potentially new CoPs. The dedication of formalized CoP resources, as a percentage of total firm resources available, is a dire opportunity cost.

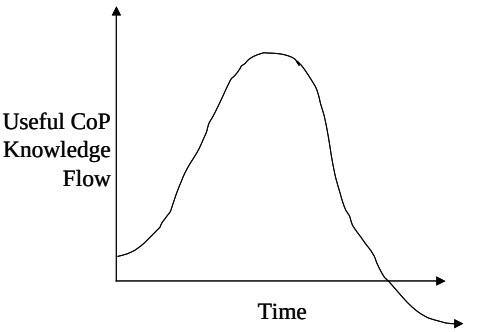
Knowledge Flow Costs

Communities maintain themselves by adding new members who introduce new practices. Early on, formalized CoPs benefit from new joiners. Structures in which individuals are in closed circles can create "silos" that prevent knowledge sharing among groups in an organization (Pfeffer & Sutton, 2000). Over time, formalized CoPs may become cliquish and suffer from the "liability of competence" as new members are unable to join because they do not speak the same advanced language. Formalization reduces the costs of interacting with certain individuals. Individuals may choose to interact only with those with whom it is easiest to interact, and not with those to whom they most need to speak.

In this community, individuals may spend their time talking about work (or other issues), rather than performing required work. For example, in formalized CoP meetings, individuals are no longer necessarily engaged in the "practice" of learning by doing, such as the Xerox technicians who discussed problems and solutions one on one or in small groups with the copiers in front of them. Instead, formalized CoP members might sit in large groups listening to one speaker, akin to a traditional classroom environment. Furthermore, formalized CoPs often require members to document knowledge that is subsequently archived in electronic repositories. Storing and transferring information is not the same as understanding and using knowledge. Employees may have limited attention, but unlimited access to good information. CoPs may add more to the information explosion than they dedicate to needed specialized attention.

Furthermore, formalized CoPs may suffer from a "liability of expertise" as firm experts become primarily one-way (outward) communicators. For example, in a CoP of programmers, a few key individuals may serve as expert resources. These individuals may spend the bulk of their day playing "expert" and solving others' problems, leaving little time to perform their real tasks. Meanwhile, "receivers" benefit from this information and may come to rely on experts rather than learn how to sort out problems themselves. When CoPs cross firm boundaries such as the Experts Exchange (*www.experts-exchange.com*), individual experts from Firm A may spend most of their day informing Firm B receivers, with no reciprocation. Figure 2 depicts the possibility of diminishing returns to scale from formalized CoP activities.

Figure 2. Useful CoP knowledge flows decrease over time in formalized CoPs



FUTURE TRENDS

Viewing organizations as embedded in their environment, institutional theory examines how isomorphic pressures from the environment lead to homogeneity among firms, for example, within a particular industry. The cultivation of CoPs (in the original construct) and formalization of CoPs could be described as a case of institutionalization using DiMaggio and Powell's (1983) three institutional pressure framework: (1) coercive (from political influence and problems of legitimacy), (2) mimetic (response to uncertainty), and (3) normative (result from professionalization). Coercive institutionalism describes the influence of formal and informal pressures and cultural expectations. In the case of CoP cultivation, informal pressures such as individual interest may lead to their formation. In the case of corrupted formalized CoPs, well-meaning CEOs could mandate the establishment of new communities or formalize support of existing communities. Mimetic institutionalism describes how uncertainty may, with or without intent, drive an organization to mould itself after other organizations. A firm might replicate a successful industry competitor's CoP initiative. In the worse-case scenario, this institutionalization might include mimicking extremely formalized CoPs. Finally, professionalism describes how the spread of formal and informal professional networks help diffuse organizational norms rapidly. Informed management consultants who are well versed in the original theory may help firms put processes into place, such as creating a common water cooler-type space, to cultivate CoPs. Other consulting professionals may institute formalized CoPs through prescriptives for formulaic rules around membership and process.

CONCLUSION

Communities of practice are informal networks of individuals who possess various levels of a common capability and apply this knowledge in pursuit of a similar endeavor. In practice, managers have implemented CoPs in three forms: the original "pure" construct of informality, an applied model, and a fully formalized structure. Formalized CoPs have high

coordination, opportunity, and knowledge flow costs that impede organizational performance. Managers interested in reaping the benefits of informal knowledge sharing should resist the temptation to formalize CoPs through mandatory meetings, roles, and routines.

Orr's (1990) seminal doctoral thesis cautioned against corporate interference in communities:

The process of working and learning together creates a work situation which the workers value, and they resist having it disrupted by their employers through events such as a reorganization of the work. This resistance can surprise employers who think of labor as a commodity to arrange to suit their ends. The problem for the workers is that this community which they have created was not part of the series of discrete employment agreements by which the employer populated the work place, nor is the role of the community in doing the work acknowledged. The work can only continue free of disruption if the employer can be persuaded to see the community as necessary to accomplishing work. (p. 48; also cited by Brown & Duguid, 1991)

Wenger (1998, p. 243) also warned against formalization:

Institutionalization must be in the service of practice. Institutionalization in itself cannot make anything happen. Communities of practice are the locus of 'real work'…excessive institutionalization stalls the organization insofar as the practices end up serving the institutional apparatus, rather than the other way around.

REFERENCES

Barnard, C. (1938). *The functions of the executive.* Cambridge, MA: Harvard University Press.

Brown, J.S., & Duguid, P. (1991). Organizational learning and communities of practice. *Organizational Science, 2*(1), 40-57.

Daft, R., & Weick, K. (1984). Toward a model of organizations as interpretation systems. *Academy of Management Review, 9*(2), 284-295.

DiMaggio, P., & Powell, W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review, 48,* 147-160.

Duguid, P. (2005). The art of knowing: Social and tacit dimensions of knowledge and the limits of the community of practice. *The Information Society, 21*(2), 109-118.

Fontaine, M. (2001). Keeping communities of practice afloat. *Knowledge Management Review,* (September/October), 16-21.

Gongla, P., & Rizzuto, C. (2004). Where did that community go? In P. Hildreth & C. Kimble (Eds.), *Knowledge networks: Innovation through communities of practice.* Hershey, PA: Idea Group.

Hildreth, P., & Kimble, C. (2004). *Knowledge networks: Innovation through communities of practice.* Hershey, PA: Idea Group.

Lave, E., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge, UK: Cambridge University Press.

March, J., & Olsen, J.P. (1975). The uncertainty of the past: Organizational learning under ambiguity. *European Journal of Political Research, 3,* 147-171.

Orr, J. (1990). *Talking about machines: An ethnography of a modern job.* PhD Thesis, Cornell University, USA.

Pfeffer, J., & Sutton, R. (2000). *The knowing-doing gap.* Boston: Harvard Business School Press.

Polanyi, M. (1966). *The tacit dimension.* New York: Doubleday.

Strauss, A. (1978). A social world perspective. *Studies in Symbolic Interactions, 1,* 119-128.

Van Maanen, J., & Barley, S. (1984). Occupational communities: Culture and control in organizations. In B. Staw & L. Cummings (Eds.), *Research in organizational behavior* (vol. 6, pp. 287-365). Greenwich, CT: JAI Press.

Wenger, E. (1998). *Communities of practice: Learning, meaning, & identity.* Cambridge, MA: Cambridge University Press.

Wenger, E., McDermott, R., & Snyder, W. (2002). *Cultivating communities of practice.* Boston: Harvard Business School Press.

Wenger, E., & Snyder, W. (2000). Communities of practice: The organizational frontier. *Harvard Business Review,* (January-February), 139-145.

KEY TERMS

Community of Practice: Informal network of individuals who possess various levels of a common capability and apply their knowledge in pursuit of a similar endeavor.

Coordination Cost: Cost of processing information in an organization.

Formalization: Act of making formal, often for the sake of official or authorized acceptance.

Knowledge Management: Set of practices used by organizations to identify, create, represent, and distribute knowledge for reuse, awareness, and learning.

Opportunity Costs: Set of costs of passing up the next best choice when making a decision.

Tacit Knowledge: Knowledge that people carry in their minds and is difficult to share with others.

Personalization in the Information Era

José Juan Pazos-Arias
University of Vigo, Spain

Martín López-Nores
University of Vigo, Spain

INTRODUCTION

We are witnessing the development of new communication technologies (e.g., DTV networks [digital TV], 3G [third-generation] telephony, and DSL [digital subscriber line]) and a rapid growth in the amount of information available. In this scenario, users were supposed to benefit extensively from services delivering news, entertainment, education, commercial functionalities, and so forth. However, the current situation may be better referred to as *information overload;* as it frequently happens that users are faced with an overwhelming amount of information. A similar situation was noticeable in the 1990s with the exponential growth of the Internet, which made users feel disoriented among the myriad of contents available through their PCs. This gave birth to *search engines* (e.g., Google and Yahoo) that would retrieve relevant Web pages in response to user-entered queries. These tools proved effective, with millions of people using them to find pieces of information and services. However, the advent of new devices (DTV receivers, mobile phones, media players, etc.) introduces consumption and usage habits that render the search-engine paradigm insufficient. It is no longer realistic to think that users will bother to visit a site, enter queries describing what they want, and select particular contents from among those in a list. The reasons may relate to users adopting a predominantly passive role (e.g., while driving or watching TV), the absence of bidirectional communication (as in broadcasting environments), or users feeling uneasy with the interfaces provided. To tackle these issues, a large body of research is being devoted nowadays to the design and provision of personalized information services, with a new paradigm of *recommender systems* proactively selecting the contents that match the interests and needs of each individual at any time. This article describes the evolution of these services, followed by an overview of the functionalities available in diverse areas of application and a discussion of open problems.

BACKGROUND

The development of personalized information services brings together diverse areas of technology, plus a significant body of legislation. The following subsections group these aspects into five major topics.

User Modeling

To identify the most suitable contents for a user, it is necessary to handle profiles that capture the user's preferences and needs. Such profiles can be represented in many different ways: consumption histories, ontologies, neural networks, decision trees, and so on (Conati, McCoy, & Paliouras, 2007). Besides these, there exist many *de jure* or *de fact*o application-dependent standards to manage data: *learner information packaging* (LIP) to track educational activities, TV-Anytime for TV programs, *Integrating the Healthcare Enterprise* (IHE) for health records, and so forth.

The initialization of the user profiles can be done manually, with the user entering a description of his or her interests, or automatically, with the program retrieving information from his or her interactions with other systems (e.g., commonly, from Web navigation histories). Whichever the approach, a recommender system must implement mechanisms to capture new data (*relevance feedback*) and discard obsolete information (*gradual forgetting*). For relevance feedback, some systems use explicit mechanisms that require the user to rate contents as interesting or not interesting; others provide implicit mechanisms that infer information from ongoing interactions (Montaner, López, & de la Rosa, 2003). In gradual forgetting, the common approach is to measure the obsolescence of data in the profiles as a function of time.

Context Awareness

Context awareness aims at acquiring information about physical and social situations to maximize the value of the information delivered to the user. Knowledge about context is added to that in the user's profile to drive the selection of contents. Regarding format and length, it is necessary to match the time the user will have to read or watch material, the size of the screens where it will be presented, the input mechanisms available, and so forth. As for the semantics, the goal is to identify the topics the user may welcome at a given moment.

The first possibility explored was to develop location-sensitive mobile applications that would display different contents following the users' moves in indoor environments (e.g., museums) or outdoors (e.g., city tours) (Baldauf, Dustdar, & Rosenberg, 2007). Other dimensions were progressively added, like the informational context (e.g., inferred from the words on screen), infrastructure (e.g., surrounding communication resources), and physical conditions (noise, light, etc.). Recent works on affective computing (Picard & Daily, 2005) bring the user's feelings (mood, stress, etc.) into consideration, too.

Characterization of Contents and Services

To enable automatic selection of contents, it is necessary to characterize the available resources regarding format, length, and semantics. The MPEG-7 standard (Manjunath, Salembier, & Sikora, 2002) here appears as a common umbrella for previously existing purpose-specific metadata specifications, such as the Dublin core for Web pages, TV-Anytime for programs, or *shared content object repository model* (SCORM) for educational resources. MPEG-7 provides descriptors for low-level audiovisual features (like color or texture) that can be annotated automatically, and for high-level characteristics of objects, events, and concepts. Furthermore, contents can be split into segments that can be handled separately: audio clips, video sequences, 3-D objects, and many others.

In the characterization of interactive services, the reference is the architecture of Web services (Cerami, 2002), which includes solutions for automatic discovery, invocation, and composition, covering practically the whole spectrum of applications over the Internet. At the core of most proposals, the Web service description language (WSDL) allows one to describe the operations offered by a service and the formats to follow in their invocation. Other initiatives, like the ontology Web language for services (OWL-S), focus on describing what the services do with a common conceptualization to reason about the functionalities delivered to the users. These solutions are now being extended to other platforms (e.g., mobile devices or DTV set-top boxes), where the services have been typically monolithic and purpose built.

Filtering

Personalization is achieved by matching the information in the user's profile with his or her context and the contents available. The first possibility explored in this regard was content-based filtering to make recommendations by looking at contents that gained the user's interest in the past (Balabanovic & Shoham, 1997). This approach is easy to adopt, but the recommendations tend to be repetitive for considering that a user will always appreciate the same kind of content. Alternatively, collaborative filtering (Mobasher, Jin, & Zhou, 2003) evaluates not only the profile of the target user, but also those of users with similar interests (his or her neighbors). This approach solves the lack of diversity, but faces problems like sparsity when the number of contents is high (which makes it hard to find users with similar evaluations for the same contents) or problems in the treatment of users whose preferences are dissimilar to the majority. To neutralize these shortcomings, there exist hybrid approaches (Burke, 2002), such as recommending contents similar to the ones stored in the target user's profile, but considering two items similar if the users who show interest in the one tend to be interested in the other (item neighborhoods).

Regardless of the filtering strategy, the first recommender systems relied on syntactic matching techniques, comparing textual strings. This approach missed the ability to discover semantic relationships between preferences, context, and contents. Nowadays, research is focused on applying techniques from the Semantic Web (Antoniou & van Harmelen, 2004), which enable processes that gain insight into the meaning of words and sentences. The reasoning can be done directly over textual contents, and over metadata descriptions in the case of images, audiovisual material, or interactive programs.

Legal Aspects

Personalization is typically opposed to privacy. The effectiveness of the former depends on accumulating information about a user, while the latter aims at restricting the management of personal information by commercial or administrative entities. Since privacy is a right in itself, there is a legal framework that must be taken in account when implementing personalization. Globally, this is delimited by OECD guidelines, which lie at the core of more specific laws for different countries worldwide. These guidelines pose restrictions to the management of user data in such aspects as the necessary levels of the user's consent, the exchange and aggregation of data from different sources, the limits to granting access to third parties or to trade information, the requirement to keep the user informed of what his or her data will be used for, and so forth. In-depth details about these requirements can be found in Wang, Zhaoqi, and Kobsa (2006).

PERSONALIZED INFORMATION SERVICES

Personalized information services are emerging in diverse areas, considering a range of consumer devices and com-

munication networks. Next, we provide a brief survey of the functionalities available.

Adaptive Hypermedia

The first studies on personalized services arose on adaptive hypermedia to improve the usability of the World Wide Web. There have been countless approaches to Web page recommenders (Chen & Magoulas, 2005), usually relying on navigation histories as the main source of information about the users' interests and context; a few ones would also tailor the presentation of contents to device characteristics such as screen size and input facilities (Bandelloni & Paternò, 2004). Recent works have extended the scope to embrace incipient research on the automatic composition of interactive services (O'Keeffe, Conlan, & Wade, 2006).

Personalized Learning

Personalized learning over the Internet developed hand in hand with adaptive hypermedia, with many systems exploring the whole spectrum of personalization technologies (Papanikolaou & Grigoriadou, 2003) and developing standards for student profiling, content description, and so on. These standards have been adopted in personalized TV-based learning (t-learning), which delivers education through simple interactions and audiovisual contents (Rey-López, Fernández-Vilas, & Díaz-Redondo, 2006). Also, many authors are working to provide personalized learning through mobile devices (m-learning), with systems that analyze the user's preferences and location to recommend the best-suited educational activities (Silander & Rytkönen, 2005).

Personalized Commerce

Personalized e-commerce has been around on the Internet for some years (Markellou, Mousourouli, Sirmakessis, & Tsakalidis, 2005), but it was absent until very recently in other platforms. The first developments arose in digital TV, with approaches to present personalized advertisements retrieved from broadcast emissions (Lekakos & Giaglis, 2004). López-Nores, Pazos-Arias, García-Duque, and Blanco-Fernández (2007) recently presented an approach to automatically assemble interactive services that provide personalized commercial functionalities. Also, systems have been demonstrated that deliver personalized and location-aware advertisements over devices like PDAs (personal digital assistants) and car radios (IST, 2006).

Recommenders of Audiovisual Contents

Maybe the first application area for personalization beyond the PC was that of personalized programming guides to help TV viewers find interesting programs among the growing number of channels available. This is now a relatively mature field of research in what concerns TV watching at home, with various recommenders featuring semantic reasoning capabilities (Ardissono, Kobsa, & Maybury, 2004). Some systems can adapt their recommendations depending on whether the user is watching TV alone or in a group. As for the delivery of audiovisual contents over mobile devices, there are incipient approaches to build music recommenders and personalized video channels, for example, for podcasting (Billsus & Pazzani, 2007).

Personalized E-Government

The need for personalization in e-government arises from the fact that users often get lost in the information space of portals, needing specific hints that are easily obtained in administrative buildings (e.g., finding the office responsible for a given service or asking for assistance to fill out certain fields of a form). Approaches to fight this problem include strategies to help discover administrative proceedings one has to fulfill (Sacco, 2007) and solutions to identify contexts in which a given proceeding is applicable (Grandi, Mandreoli, Martoglia, Ronchetti, Scalas, & Tiberio, 2006). User profiles here usually consist of structures to gather demographical data such as age, gender, job, or level of income.

Personalized Health Care

Personalized health care aims at supporting users' well-being and medical treatments, considering the individual's unique biological, social, and cultural characteristics. The related research ranges from the design of wearable or implantable devices that sense physiological parameters (Winters & Wang, 2003) to technologies for ambient-assisted living (López-Nores, Pazos-Arias, García-Duque, & Blanco-Fernández, 2008) and expert systems for information processing and diagnosis (James, Wilcox, & Naguib, 2001). Although the research activity is hectic, health care remains one of the areas where information technologies have attained lowest penetration, yet it is also the one in which personalization may report the greatest benefits in social and economical terms.

Other Services

Personalization is progressively gaining new areas of application. Just to enhance the picture of the preceding subsections, we can cite works about location-aware restaurant recommenders (Lee, Kim, Jung, & Jo, 2006), personalized tourism (Kang, Kim, & Cho, 2006), and personalized friend making (Lo & Lin, 2006).

FUTURE TRENDS

There is much place for innovation in the development of personalized information services to enhance technical aspects of user modeling, context awareness, and filtering. Notwithstanding, the greatest challenge may be to find sustainable exploitation models for the already existing solutions. Currently, content and service providers bear the brunt of all the personalization tasks, which poses problems related to increases in production costs and computational requirements as the number of users grows. To solve them, it is necessary to move part of the complexity to the users. For example, the users' devices could keep their profiles and run the filtering algorithms, which would foster privacy while harnessing the power of millions of consumer devices. However, this idea precludes the application of collaborative filtering techniques, which depend on matching multiple profiles to delimit neighborhoods. Besides this, it would be necessary to enhance many consumer devices with data processing capabilities, which may be hard to realize in some cases.

Possibly the most complex part of gaining users' involvement is to get them to contribute metadata for recommender systems to reason about. This is necessary because content and service providers are already incapable of characterizing the growing amount of information available. Therefore, just as new tools such as blogs and YouTube have turned users into active producers of information, there is an urgent need for means that will make them active agents for information retrieval as well. After all, without a scalable approach to produce metadata, most of the research on personalization is doomed to collapse for lack of input. The recent advances on collaborative tagging (Grahl, Hotho, & Stumme, 2007) and the experiences of social sites like Flickr and del.icio.us provide insight on how this could be solved on the Internet, but other platforms may require much further innovation.

CONCLUSION

The success of information technologies depends ultimately on the services provided to the users, which can be rendered useless by the growth in the amount of information available. Personalization aims at solving this problem by providing users with information services that match their preferences and needs in any context within the reach of technology. This requires a shift from the traditional search engines to recommender systems working in behalf of each user, within the limits imposed by privacy concerns. The range of applications is enormous, covering many new communication networks, consumer devices, and consumption models. However, it is still an open issue to find viable and scalable exploitation models for personalized information services to become a ubiquitous reality.

REFERENCES

Antoniou, G., & van Harmelen, F. (2004). *A Semantic Web primer.* Cambridge, MA: The MIT Press.

Ardissono, L., Kobsa, A., & Maybury, M. (Eds.). (2004). *Personalized digital television: Targeting programs to individual users.* Norwell: Kluwer Academic Publishers.

Balabanovic, M., & Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Communications of the ACM, 40*(3), 66-72.

Baldauf, M., Dustdar, S., & Rosenberg, F. (2007). A survey on context-aware systems. *International Journal of Ad Hoc and Ubiquitous Computing, 2*(4), 263-277.

Bandelloni, R., & Paternò, F. (2004). Migratory user interfaces able to adapt to various interaction platforms. *International Journal of Human-Computer Studies, 60*, 621-639.

Billsus, D., & Pazzani, M. (2007). Adaptive news access. In *Lecture notes in computer science* (Vol. 4321, pp. 550-570).

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction, 12*(4), 331-370.

Cerami, E. (2002). *Web services essentials.* Sebastopol: O'Reilly Media.

Chen, S., & Magoulas, G. (Eds.). (2005). *Adaptable and adaptive hypermedia systems.* Hershey, PA: IRM Press.

Conati, C., McCoy, K., & Paliouras, G. (Eds.). (2007). *Proceedings from UM'07: The 11th International Conference on User Modeling.* Berlin, Germany: Springer.

Grahl, M., Hotho, A., & Stumme, G. (2007). *Conceptual clustering of social bookmarking sites.* Paper presented at the Seventh International Conference on Knowledge Management, Graz, Austria.

Grandi, F., Mandreoli, F., Martoglia, R., Ronchetti, E., Scalas, M., & Tiberio, P. (2006). Semantic Web techniques for personalization of egovernment. In *Lecture notes in computer science* (Vol. 4231, pp. 435-444).

IST. (2006). *iPointer platform.* Retrieved December 20, 2007, from http://www.i-spatialtech.com/ipointer.htm

James, A., Wilcox, Y., & Naguib, R. (2001). A telematic system for oncology based on electronic health and patient records. *IEEE Transactions on Information Technology in Biomedicine, 5*(1), 16-17.

Kang, E.-Y., Kim, H., & Cho, J. (2006). Personalization method for tourist point of interest (POI) recommenda-

tion. In *Lecture notes in computer science* (Vol. 4251, pp. 392-400).

Lee, B.-H., Kim, H.-N., Jung, J.-G., & Jo, G. (2006). Location-based service with context data for a restaurant recommendation. In *Lecture notes in computer science* (Vol. 4080, pp. 430-438).

Lekakos, G., & Giaglis, G. (2004). A lifestyle-based approach for delivering personalised advertisements in digital interactive television. *Journal of Computer-Mediated Communications, 9*(2).

Lo, S., & Lin, C. (2006). *WMR: A graph-based algorithm for friend recommendation.* Paper presented at the International Conference on Web Intelligence, Hong Kong, China.

López-Nores, M., Pazos-Arias, J., García-Duque, J., & Blanco-Fernández, Y. (2007). *Spontaneous and personalized advertising through MPEG-7 markup and semantic reasoning.* Paper presented at the Second International Conference on Signal Processing and Multimedia Applications, Barcelona, Spain.

López-Nores, M., Pazos-Arias, J., García-Duque, J., & Blanco-Fernández, Y. (2008). *A smart medicine manager delivering health care to the networked home and beyond.* Paper presented at the International Conference on Health Informatics, Funchal, Portugal.

Manjunath, B., Salembier, P., & Sikora, T. (2002). *Introduction to MPEG-7: Multimedia content description language.* Hoboken, NJ: Wiley.

Markellou, P., Mousourouli, I., Sirmakessis, S., & Tsakalidis, A. (2005). *Personalized e-commerce recommendations.* Paper presented at the IEEE International Conference on E-Business Engineering, Beijing, China.

Mobasher, B., Jin, X., & Zhou, Y. (2003). Semantically-enhanced collaborative filtering on the Web. In *Lecture notes in computer science* (Vol. 3209, pp. 57-76).

Montaner, M., López, B., & de la Rosa, J. (2003). A taxonomy of recommender agents on the Internet. *Artificial Intelligence Review, 19*(4), 285-330.

O'Keeffe, I., Conlan, O., & Wade, V. (2006). A unified approach to adaptive hypermedia personalization and adaptive service composition. In *Lecture notes in computer science* (Vol. 4018, pp. 303-307).

Papanikolaou, K., & Grigoriadou, M. (2003). *An instructional framework supporting personalized learning on the Web.* Paper presented at the IEEE International Conference on Advanced Learning Technologies, Athens, Greece.

Picard, R., & Daily, S. (2005). *Evaluating affective interactions: Alternatives to asking what users feel.* Paper presented at the CHI Workshop on Evaluating Affective Interfaces, Portland, OR.

Rey-López, M., Fernández-Vilas, A., & Díaz-Redondo, R. (2006). A model for personalized learning through IDTV. In *Lecture notes in computer science* (Vol. 4018, pp. 457-461).

Sacco, G. (2007). *Interactive exploration and discovery of e-government services.* Paper presented at the Eighth International Conference on Digital Government Research, Philadelphia.

Silander, P., & Rytkönen, A. (2005). *An intelligent mobile tutoring tool enabling individualization of students' learning processes.* Paper presented at the Fourth World Conference on M-Learning, Cape Town, South Africa.

Wang, Y., Zhaoqi, C., & Kobsa, A. (2006). *A collection and systematization of international privacy laws.* Retrieved December 20, 2007, from http://www.ics.uci.edu/~kobsa/privacy/intlprivlawsurvey.html

Winters, J., & Wang, Y. (2003). Wearable sensors and telerehabilitation. *IEEE Engineering in Medicine and Biology Magazine, 22*(3), 56-65.

KEY TERMS

Collaborative Filtering: Collaborative filtering includes techniques to estimate the relevance of a given service or piece of content for a target user, considering the ratings given by other users with similar profiles to the same resource or to similar ones.

Content-Based Filtering: This includes techniques to estimate the relevance of a given service or piece of content for a target user, considering that user's previous ratings for resources consumed in the past.

Context Awareness: It is the ability of an information system to adapt to a user's changing attention, location, and physical environment.

Information Overload: It is a situation in which it is hard for a user to stay informed or make decisions about a given topic due to being exposed to an excessive amount of information.

Personalized Information Service: It is a system that delivers information tailored to the interests, preferences, needs, and context of a user or a group of users.

Recommender Systems: These are software systems that proactively look for services and pieces of content that may be relevant for a target user or group of users, matching

user profiles, contextual information, and metadata describing the resources available.

User Profiles: User profiles are data structures containing information that may be used to characterize the interests, preferences, and needs of a user.

Personalization Technologies in Cyberspace

Shuk Ying Ho
The University of Melbourne, Australia

INTRODUCTION

Hundreds of thousands of companies worldwide are using the Web as a major channel to interact with their customers for brand promotion, product marketing, order fulfillment, and after-sales support. Competition is extremely keen among online merchants.[1] In doing business online, the question that lurks in the back of their mind is, are we maximizing our business opportunities*?*

With the high interactivity of e-commerce, online merchants now adopt various differentiating strategies to attract and retain customers in the hope of remaining competitive. To provide a differentiated service, online merchants first identify each individual, and then acquire more information about each individual's interests. Then, they can tailor Web content directly to a specific user by having the user provide information to the Web site either directly or through tracking devices on the site. The software can then modify the content to the needs of the user. Ultimately, highly focused and relevant products or services are delivered to each customer, who is treated in a unique way to fit marketing and advertising with his or her needs. This process is generally named *personalization*.

There is a wide range of personalization strategies used nowadays. For instance, My Yahoo! provides a personalized "space" for each user. It automatically generates personalized content (e.g., information on the horoscope for the correct star sign) matched with users' profiles (e.g., a person's date of birth). Apart from automatic personalization, it also presents the users with an array of choices and allows the users to select what is of interest to them. The users can personalize not only the content (e.g., weather, finance) but also the layout (e.g., color, background). My Yahoo! was considered to be one of the forerunners among the growing number of personalized Web sites that have been springing up on the Internet over the last few years (Manber, Patel, & Robison, 2000). Amazon.com greets returning customers with a personalized message and offers a hyperlink to book recommendations congruent with their past purchases. These recommendations are generated based on the customers' previous purchases and the preferences of like-minded people, and there is no extra work imposed on the customers. Amazon continues to establish its personalization system, and more filtering mechanisms are being added to make the book recommendations be more relevant and useful. Recently, there has been the introduction of a personalized search engine, A9.com by Amazon.com, which recommends relevant Web sites to each individual by analyzing his or her browsing history and bookmarks. Expedia.com asks users for their desired destinations and then e-mails them information about special discounts to the place where they like to travel. It is expected that corporate investment in personalization technologies will continue to surge in the future (Awad & Krishnan, 2006; Poulin, Montreuil, & Martel, 2006; Rust & Lemon, 2001).

Given the proliferation of personalization, this chapter will address the key issues related to personalization and provide definitions to some keywords, such as rule-based personalization and collaborative filtering.

BACKGROUND

Definition of Personalization

Personalization is one of the rapidly emerging technologies in the field of information systems (IS) and is drawing increasing attention in academia (Adomavicius & Tuzhilin, 2005). Due to its fast emergence, there is still no globally agreed definition from researchers and practitioners as to what personalization actually is. Here are some of the representative ones:

- Wikipedia defines personalization to be a means of changing Web pages based on the interests of an individual. Personalization implies that the changes are based on implicit data, such as items purchased or pages viewed.
- As defined by Personalization Consortium in 2003, personalization is "the use of technology and customer information to tailor electronic commerce interactions between a business and each individual customer."
- According to Bitpipe,[2] personalization is a process of creating a means of communication (Web site, letter, etc.) that is specific to your readers in order to improve customer relationships and loyalty.

Generally speaking, we agree with the previous definitions. In this chapter we would reference personalization to be a process of matching products, services, and advertising content with each individual. The matching process is

based on what an online merchant knows about a user. With this knowledge, online merchants construct a user profile, which defines users' preferences and their interaction behaviors. Personalization technologies can be applied not only on e-commerce, but also mobile commerce or other business channels in the hope of generating more business opportunities.

The technology enabler is generally referred to as a *personalization agent*, which is a collection of software modules that provides tools to collect and analyze user data and adapt the content to Web users' objectives and facilitate their navigation or buying process (Cingil, Dogac, & Azgin, 2000). This is accomplished by deploying pattern recognition software to collect and analyze Web semantics, navigation activities, and purchase transactions of the individuals. Then, content and presentation format is adapted for each individual. Examples of software modules include customer relationship management, data mining, collaborative technology, and clickstream analysis components. With the help of personalization agents, online merchants can now exert control and manipulate content-related parameters at a very fine level not previously possible. In this way, they can ensure the right person receives the right content in the right format at the right time.

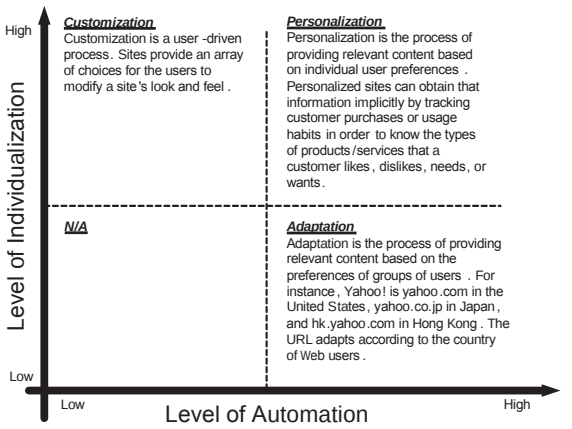
Adaptation, Customization, and Personalization

When researchers and practitioners start a project on personalization, they might ask whether the terms, *adaptation*, *customization,* and *personalization*, are interchangeable. So far, there are few studies that clearly distinguish these three terms. We modify the definition by Ho (2006) and distinguish these terms with two dimensions: level of automation and level of individualization. Level of automation refers to user control in the process of generating recommendations. The two ends are user-driven and machine-driven. Level of individualization refers to the degree of differentiation of recommendations from one person to another person. Some recommendations are offered for a group of people, whereas some are tailored for each individual.

Customization focuses on direct user control. In this user-controlled process, the user is given a set of options, and he or she chooses specific interests on a checklist so that the site can display the requested information. In some sophisticated settings, based on these user preferences, the merchant's system recommends additional products to the user. This technique is fairly complicated because the merchant has to map among different product categories in advance.

Adaptation is driven by intelligent software programs, and a group of users receives the same outcome of adaptation. For instance, a global brand differentiates itself in each market segment by the domain name. For example, Yahoo is yahoo.com in the United States, and yahoo.co.jp in Japan. Each regional portal provides local news and weather reports relevant to the local users.

Figure 1. Taxonomy of customer relationship management systems for e-commerce



Personalization is driven by the software package. The online merchants extract, combine, and integrate data taken from multiple sources before personalization becomes operational. Then personalization software packages mine a Web site's data and attempt to serve up individualized pages to the user based on a model of that user's preferences. By definition, personalization will generate different recommendations to different individuals.

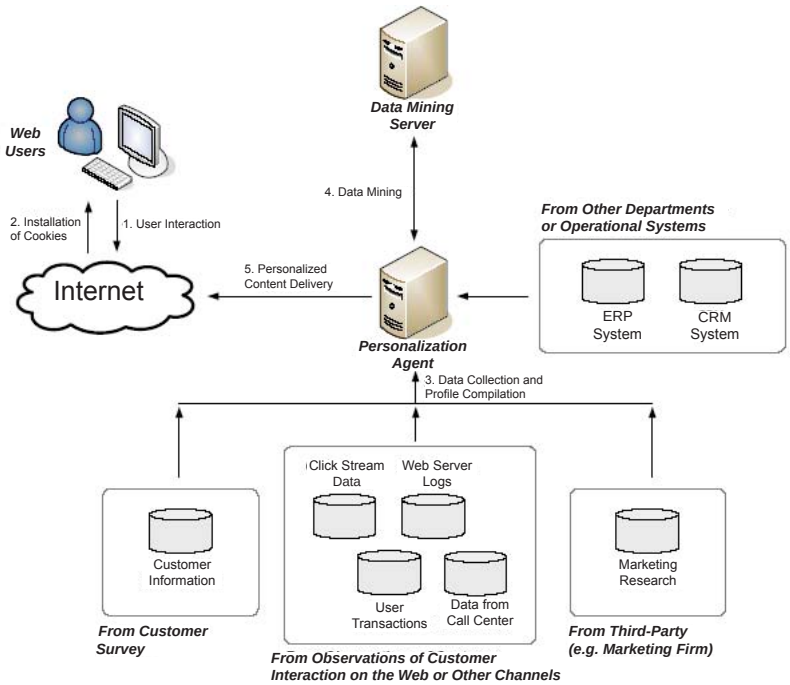
Figure 1 presents a summary of these three terms. Nowadays, these technologies are powerful tools in the battle for customer loyalty.

Personalization Process

Figure 2 shows the process to operate personalization. It also specifies the data and the machine components to support the process. The five key steps are:

1. **User Interaction:** Web users interact with the Web site and gradually provide information that profiles them in terms of browsing habits and product needs. In many cases, the site requests the users to fill out a survey stating their preferences.
2. **Installation of Cookies:** A cookie is a set of small text files stored on users' hard drives, and it is usually installed without the users' consent when the users first visit a Web server (Goldsborough, 2005). When the users revisit the site and request a page from that Web server, the cookie sends a message to that server. In this way, online merchants can identify each user.

Figure 2. Procedures of personalization process



A cookie allows the merchants to implement a simple personalization strategy, such as name greeting.

3. **Data Collection and Profile Compilation:** The processes of extracting, transforming, and loading are activated. Online merchants extract data from various sources to compile a user profile.

   a. User Surveys. Questionnaires and interviews are used to solicit information directly from the users. General information, such as user demographics, is collected. With a well-designed survey instrument, the data is easy to analyze (Ho, 2005).

   b. Observations. Server logs provide descriptive statistics on the page popularity, and user transactions are important indicators of what users are actually interested in. Richer data include clickstream data, which refers to lines of code stored in a file every time a user views a page. Possible measurements include the pages they browse, how long they navigate on each page, the products explored and bought. Clickstream analysis makes it possible for an online merchant to construct all users' browsing transitions. As a whole, observations provide more objective and reliable data than a user survey (Ho, 2005).

   c. Third-Party Data. Marketing firms, such as ACNielsen.com,[3] provide services to online merchants to help them research the market segment. The amount of data used and the kinds of data to be used are dependent on online merchants' constraints and their needs.

4. **Data Mining:** With a compiled profile, online merchants can use data mining techniques to analyze large volumes of data and discover subtle relationships between data items (Dobler, 2005). Patterns that accurately predict behaviors in users are identified (Eirinaki & Vazirgiannis, 2003). The prediction rules mined are sent to the personalization agent.
5. **Personalized Content Delivery:** After the production details and marketing strategies retrieved from other operational systems are integrated with the prediction rules, personalization agents can generate and deliver recommendations to the users.

Personalization Approaches

Once an online merchant knows a user's browsing habits and preferences, it would be useful if the merchant could predict what other products or services this user might enjoy. Much personalization work focuses on how to use existing data to infer if users are interested in other products or services.

This refers to step 4 in the previous section. This prediction is based on special formulas derived from behavioral science. The merchant usually does not ask the users; otherwise, this may impose an extra workload on the users and reduce their satisfaction. In the following sections, we introduce two typical approaches used in personalization.

Collaborative Filtering Personalization

Collaborative filtering systems usually take two steps: First, it keeps track of users' behaviors and transactions across the Web. The software interprets their preferences by comparing the information about a user's behavior against data gathered about other users with similar behaviors. Second, it finds the closest peers for each user, that is, people with the most similar preferences, and uses the ratings from those like-minded users found in the previous step to calculate a prediction for this user. A typical example is the feature "customers who bought this book also bought …" provided by Amazon.com. Currently, many personalization systems are based on collaborative filtering. One of these commercial systems is Firefly, which is now embedded in Microsoft's Passport System.

Rule-Based Personalization

The merchants ask users a series of predefined questions. Certain behavioral patterns are predicted using the collected information. For instance, an insurance company can use a rule, "If customer age is greater than 35 with annual income greater than $200,000, then we should ask this customer to buy a premium insurance package." Users are divided into segments based on business rules that generate certain types of information from a user's profile. One of the popular commercial products is BroadVision,[4] which invites users to fill out a questionnaire to determine the type of product or service they like. A user profile is formed based on the questionnaire result and stored in the database by user segment (e.g., age and income). The decision to give personalized information is based on these business rules.

IS PERSONALIZATION EFFECTIVE?

Debates on the Effectiveness of Personalization

Online merchants adopt personalization technology in the hope of better communicating with their users and generating more business opportunities. Nowadays, while personalization technology has become a crucial component of relationship management solutions, our understanding of the effectiveness of personalization is far from conclusive.

Advocates of personalization claim that personalization agents have changed the Web into a personal communication medium. By providing individualized content, offerings, and services, personalization helps to control aimless surfing activity (Hanson & Crayne, 2005; Light & Maybury, 2002; Loia, Pedcrycz, Senatore, & Sessa, 2006; Shahabi & Banaei-Kashani, 2003) and to facilitate business-to-consumer interaction (Ardissono, Goy, Petrone, & Segnan, 2002; Tam & Ho, 2005). Also, personalizing Web content empowers merchants to deliver customer value and to achieve profitable growth (Greer & Murtaza, 2003). It was reported that e-commerce sites using personalization technology have seen annual revenue increases of up to 52% (Parkes, 2001).

On the other hand, there remains skepticism on the prospects of personalization. On the side of consumers, a report by Jupiter Research (2003) indicated that only 14% think that personalized offers or recommendations on shopping Web sites lead them to purchase more frequently. Nunes and Kambil (2001) conducted a survey and found that half of the online users of various e-commerce Web sites would rather customize a Web site themselves than have it automatically personalized for them. On the side of online merchants, Festa (2003) remarked that online merchants seeking to personalize their Web sites in the hope of boosting online sales are not getting the expected payback. It costs about four times as much to personalize a Web site than to run a comparable adaptive site. Nielsen (1998) also had the criticism that most badly-designed Web sites were incorporated with personalization agents in the hope of smoothing user navigation. However, this was unnecessary, and a redesign of information architecture of the Web site could solve the problem equally well.

Gaps in Personalization Research

Why are there conflicting findings in personalization research? We identify three reasons which lead to these inconsistencies. First, most studies on personalization focus on one single system, and these systems use various strategies to personalize the users and are applied in different contexts. Their users have different needs, interests, knowledge, goals, and working tasks. This greatly reduces the ability to generalize their findings. For instance, Manber et al. (2000) intensively studied the personalization functions in My Yahoo!. The users are provided with a choice collection and customize the content and the layout of their individual site on their own. Also, they use My Yahoo! for different proposes. Some may look for specific information, whereas some may just randomly browse to kill time. Their paper highlighted the strength of information architecture and user-friendly functions of this system. However, they did not provide any empirical evidence showing the user

evaluation on My Yahoo! In a study by Smyth and Cotter (2000), a personalization system, which tailored television listings service for real users, was examined. Rather than user-driven customization, their personalization engine integrated collaborative filtering and content-based filtering to generate a personalized list and *predict* what the users want. The users browse the personalized list for television program recommendations.

Second, these studies use a variety of tasks and measurements to justify the effectiveness of personalization. For instance, in the study by Smyth and Cotter (2000), subjects were not required to perform any specific tasks, and they were asked to write down their perception of the personalization system in terms of content precision, ease of use, and speed of service. Though the evaluation of each quality variable was high and they argued that the personalization system attracted more than 20,000 registered users one year after launch, this might not be a good indicator of its success. A better judgment of a user's interest in the personalized program is whether the user actually watches the program. Besides, Te'eni and Feldman (2001) conducted a study on how an adaptive Web facilitates the searching process in a decision-making context. Performance and user satisfaction were used as the dependent variables. When working with an adaptive Web site, subjects were found to have better performance, but less user satisfaction because of the interface inconsistencies. In their study, the Web site provided content personalization, but at the same time, degraded layout personalization. A better approach is to personalize content, but minimize the change of layout (i.e., fix the layout personalization variable). This may come up with different findings from theirs. The study by Festa (2003) focused on advice-giving personalization systems and evaluated the revenue-generating variables to evaluate personalization systems, but ignored user satisfaction, loyalty, and shopping decision performance.

Last, personalization research lacks any adequately developed theoretical basis. Researchers do not have a common ground for developing hypotheses and interpreting results, and hence, the lack of underlying theory leads to the current state of inconclusive results in the IS literature. The current trend in personalization research is on the information architecture and technical implementation of the systems (e.g., Cingil et al., 2000; Desouza, 2003; Eirinaki & Vazirgiannis, 2003) and case analysis of individual system performance (e.g., Manber et al., 2000; Perkowitz & Etzioni, 2000). Researchers put little effort to build relatedness among studies. To improve this situation, researchers can achieve an acceptable level of understanding of personalization and ultimately formulate an underlying theory by first building a framework that defines the boundary for research to be conducted.

FUTURE TRENDS

Although most examples in this chapter are based on e-commerce, we believe that personalization is also useful in other communication channels, such as mobile commerce. Further research can be conducted to improve the process of personalization. Attempts by researchers and practitioners can be made to use artificial intelligence to match the product with users' needs. For instance, users have different preferences at different times and contexts, and thus it is difficult for personalization to be perfect all of the time. User context is an important factor in a personalization model. Additional variables, such as personality traits, can also be incorporated into the model to increase prediction accuracy. Metrics should be developed to measure the impact of personalization.

In order to generate personalized recommendations highly matched with users' preferences, online merchants have to collect much information from users. If the data collection is without users' knowledge or consent, then this raises ethical and legal concerns, such as invasion of privacy issues (Awad & Krishnan, 2006). In the future, research efforts could be spent on developing permission-based personalization tools. With these, online merchants and users can control how much information is given from the users to the merchants for personalization.

CONCLUSION

This article describes the concept of personalization, particularly the personalization process and its major components. We also provide explanations to address one of the most controversial topics in personalization—the effectiveness of personalization. Currently, personalization is still considered to be a costly and highly sophisticated technology, which requires knowledge of human behavior, statistical techniques, and marketing strategies. With advances in technologies and development of good metrics to measure its impact, we believe better personalization solutions will emerge in the near future.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2005). Personalization technologies: A process-oriented perspective. *Communications of the ACM, 48*(10), 83-90.

Ardissono, L., Goy, A., Petrone, G., & Segnan, M. (2002). Personalization in business-to-customer interaction. *Communications of the ACM, 45*(5), 52-53.

Awad, N. F., & Krishnan, M. D. (2006). The personalization privacy paradox: An empirical evaluation of information

transparency and the willingness to be profiled online for personalization. *MIS Quarterly, 30*(1), 13-28.

Cingil, I., Dogac, A., & Azgin, A. (2000). A broader approach to personalization. *Communications of the ACM, 43*(8), 136-141.

Desouza, K. C. (2003). Technical opinion: Barriers to effective use of knowledge management systems in software engineering. *Communications of the ACM, 46*(1), 99-101.

Dobler, C. P. (2005). Data mining: Next generation challenges and future directions. *Journal of the American Statistical Association, 100*(472), 1467.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transactions on Internet Technology, 3*(1), 1-27.

Festa, P. (2003). *Report slams Web personalization*. Retrieved May 21, 2004, fromhttp://www.news.com.com/2100-1038-5090716.html

Goldsborough, R. (2005). The benefits, and fear, of cookie technology. *Tech Directions, 64*(10), 9.

Greer, T. H., & Murtaza, M. B. (2003). Web personalization: The impact of perceived innovation characteristics on the intention to use personalization. *Journal of Computer Information Systems, 43*(3), 50-55.

Hanson, V. L., & Crayne, S. (2005). Personalization of Web browsing: Adaptations to meet the needs of older adults. *Universal Access in the Information Society, 4*(1), 46-58.

Ho, S. Y. (2005, August). An exploratory study of using a user remote tracker to examine Web users' personality traits. In *Proceedings of International Conference of Electronic Commerce*, Xi'an, China.

Ho, S. Y. (2006). The attraction of internet personalization to Web users. *Electronic Markets, 16*(1), 41-50.

Jupiter Research. (2003, October 14). *Jupiter research reports that Web site "personalization" does not always provide positive results*. Retrieved November 30, 2005, from http://www.jupitermedia.com/corporate/releases/03.10.14-newjupresearch.html

Light, M., & Maybury, M. T. (2002). The adaptive Web: Personalized multimedia information access. *Communications of the ACM, 45*(5), 54-59.

Loia, V., Pedcrycz, W., Senatore, S., & Sessa, M. I. (2006). Web navigation support by means of proximity-driven assistant agents. *Journal of the American Society for Information Science and Technology, 57*(4), 515-527.

Manber, U., Patel, A., & Robison, J. (2000). Experience with personalization on Yahoo!. *Communications of the ACM, 43*(8), 35-39.

Nielsen, J. (1998). *Personalization is over-rated.* Retrieved November 30, 2005, from http://www.useit.com/alertbox/981004.html

Nunes, P. F., & Kambil, A. (2001, April). Personalization? No thanks. *Harvard Business Review,* 2-3.

Parkes, C. (2001). *The power of personalization: Web technologies get personalized to increase profits*. Retrieved May 21, 2004, from http://esj.com/Features/article.aspx?EditorialsID=35

Perkowitz, M., & Etzioni, O. (2000). Adaptive Web sites. *Communications of the ACM, 43*(8), 152-158.

Poulin, M., Montreuil, B., & Martel, A. (2006). Implications of personalization offers on demand and supply network design: A case from the golf club industry. *European Journal of Operational Research*, *169*(3), 996-1009.

Rust, R. T., & Lemon K. N. (2001). E-Service and the customer. *Internal Journal of Electronic Commerce, 5*(3), 85-101.

Shahabi, C., & Banaei-Kashani, F. (2003). Efficient and anonymous Web-usage mining for Web personalization. *Journal of Computing, 15*(2), 123-147.

Smyth, B., & Cotter, P. (2000). A personalized television listings service. *Communications of the ACM, 43*(8), 107-111.

Tam, K. Y., & Ho, S. Y. (2005). Web personalization as a persuasion strategy: An elaboration likelihood model perspective. *Information Systems Research, 16*(3), 271-293.

Te'eni, D., & Feldman, R. (2001). Performance and satisfaction in adaptive Websites: An experiment on searches within a task-adapted Website. *Journal of the Association for Information Systems, 2*(3), 1-28.

KEY TERMS

Adaptation: It is the process of providing relevant content based on the preferences of groups of users.

Collaborative Filtering: It is a process to keep track of users' behaviors and transactions across the Web, and finds the closest peers for each user. Recommendations are made based on the behaviors of the closest peers.

Customization: It is a user-driven process, and Web sites provide an array of choices for the users to modify a Web site's look and feel.

Data Mining: It is a process to use statistical techniques to analyze large volumes of data and discover subtle relationships between data items.

Personalization: It is the process of providing relevant content based on individual user preferences. The objective is to ensure the right person receives the right content in the right format at the right time.

Personalization Agent: It is the technology enabler for personalization, and it is a collection of software modules that provides tools to collect and analyze user data and adapt the content to Web users' objectives.

Rule-Based Personalization: Users are asked a series of predefined questions and the answers are divided into segments. Recommendations are given based on the business rules.

User Profile: It defines users' preferences and their interaction behaviors on a Web site.

ENDNOTES

1 Examples of online merchants include content providers, online shops, information portals, and auction sites whose primary clients are individual Web users.
2 See: http://www.bitpipe.com/tlist/Personalization.html
3 See: http://www.acnielsen.com/
4 See: http://www.broadvision.com

Perspectives of Transnational Education

Iwona Miliszewska
Victoria University, Australia

INTRODUCTION

In recent years, a particular stream of distance education called *transnational education* has become widespread (Davis, Olson, & Bohm, 2000; van der Vende, 2003). Transnational education, often referred to as offshore education, describes all programs in which the learners are located in a country different from the one where the awarding institution is based. This article discusses various aspects of transnational education. It reviews the definition of transnational education, its typology, growth, factors determining demand and supply, and characteristics of typical programs. The article concludes with a discussion on the role that face-to-face interaction plays in transnational programs.

BACKGROUND

Reviewing recent studies of transnational education reveals that there is little agreement about what to include in this category. Similarly, there is no agreement on the various subdefinitions that inform the subject. For the purpose of this article, a working definition of transnational education produced by UNESCO and the Council of Europe for their Code of Practice in the Provision of Transnational Education was used (UNESCO & Council of Europe, 2001). This states that transnational education includes:

All types of higher education study programme, or sets of courses of study, or educational services (including those of distance education) in which the learners are located in a country different from the one where the awarding institution is based. Such programmes may belong to the educational system of a State different from the State in which it operates, or may operate independently of any national system. (UNESCO & Council of Europe, 2001)

This definition includes education that is provided by collaborative arrangements, such as franchising, twinning, joint degrees where study programs are provided in collaboration with a partner institution, as well as noncollaborative arrangements such as branch campuses, offshore institutions, and corporate universities.

The Australian Department of Education Science and Training (DEST, 2005) provides a definition of *Australian Transnational Education*; this definition includes two additional requirements:

1. That the transnational program be delivered or assessed by an accredited Australian provider; and
2. That the delivery include a face-to-face component.

It further stresses that, in contrast to distance education provided in purely distance mode, transnational education includes a physical presence of instructors offshore, either directly by the Australian provider, or indirectly through a formal agreement with a local institution (DEST, 2005).

Transnational Education: Perspectives and Characteristics

There are a great number of different relationships between different types of transnational education providers, delivery mechanisms, and programs/awards. Charting these types is a difficult task, as the constantly evolving, highly complex situation includes an array of partnerships, consortia, articulation agreements, modes of delivery, public, private, off-shore, for-profit and corporate elements. Various models of teaching can also be found, ranging from full program delivery at an offshore campus, combined face-to-face and flexible delivery option, and e-learning (Goodfellow, Lea, Gonzales, & Mason, 2001).

Typology of Transnational Education

Transnational education is constantly evolving. Wilson and Vlãsceanu (2000) distinguished between three interrelated perspectives of this evolution adding that:

all these new developments in higher education share certain common characteristics and similarities, mainly in terms of the ways they cross the borders of national higher education systems. It is for this reason that they are usually identified by the generic phrase of transnational education. (Wilson & Vlãsceanu, 2000, p. 75)

The first perspective relates to the delivery mechanisms and arrangements including franchising, corporate universities, international institutions, distance learning, and virtual universities (Machado dos Santos, 2002). Wilson and Vlãsceanu (2000) noted:

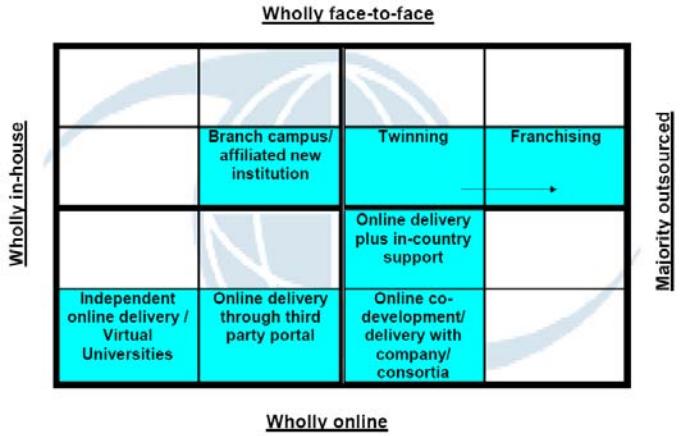
One form of development refers to a modality of delivering an educational programme (i.e., distance education), others to ways of establishing a programme/institution (i.e., franchising or twinning/branch campus), and others again to ways of offering primarily continuing education to certain new groups of students. There seems to be no limit to the proliferation of such modalities or arrangements, as long as the demand for higher education is still growing, and the possibilities for a global market continue to emerge. (Wilson & Vlãsceanu, 2000, p. 78)

The second perspective relates to the institutional and organisational arrangements that result from the adopted delivery mechanisms. This can be either a new institution, a branch, or a franchised program or course of study offering an award within an existing institution or other organisation. The third perspective refers to the nature and quality of qualifications awarded through transnational education, for example, degrees, certificates, or study credits (Wilson & Vlãsceanu, 2002).

Following on Wilson and Vlãsceanu's (2002) categorisation of transnational programs according to their delivery mechanism, Adam (2001) and Vignoli (2004) have described the most common forms of such programs as follows:

- **"Franchising:** The process whereby a higher education institution (franchiser) from a certain country grants another institution (franchisee) in another country the right to provide the franchiser's programmes/qualifications in the franchisee's host country, irrespective of the students' provenance; in many cases, the franchisee only provides the first part of the educational programme, which can be recognised as partial credits toward a qualification at the franchiser's in the context of a *programme articulation*.
- **Programme articulations:** Inter-institutional arrangements whereby two or more institutions agree to define jointly a study programme in terms of study credits and credit transfer, so that students pursuing their studies in one institution have their credits recognised by the other in order to continue their studies (*twinning programmes, articulation agreements*, etc.). These may—or may not—lead to joint or double degrees.
- **Branch campus:** A campus established by a higher education institution from one country in another country (host country) to offer its own educational programmes/qualifications, irrespective of the students' provenance; the arrangement is similar to franchising, but the franchisee is a campus of the franchiser.
- **Off-shore institution:** An autonomous institution established in a host country but said to belong, in terms of its organisation and educational contents, to the education system of some other country, without (necessarily) having a campus in the mother country.
- **Corporate universities:** They are usually parts of big transnational corporations and organise their own higher education institutions or study programmes offering qualifications that do not belong to any national system of higher education.
- **International institutions:** Institutions offering so-called *international* programmes/qualifications that are not part of a specific education system.
- **Distance learning arrangements and virtual universities:** Where the learner is provided with course material via post or Web-based solutions, and self-administers the learning process at home; the only contact with the student is by remote means." (Vignoli, 2004, p. 2)

Figure 1. Types of transnational provision (as presented in Bjarnason, 2005)



Bjarnason (2005) further qualified the various types of transnational education programs according to two scales. One scale indicated the extent of online reliance of a program and ranged from *wholly face-to-face* to *wholly online*; the other indicated the extent of institutional involvement in program development and delivery, ranging from *wholly in-house* to *majority outsource*. Figure 1 presents placement of transnational programs according to the two scales.

Davis, Olsen, and Böhm (2000) suggested a typology for Australian transnational education programs in which they separate the provider dimension from the student dimension. The provider dimension of the model spans a range where the increasing responsibility of the partner institution varies across academic teaching, assessment, and support; curriculum; provision of study location; student support; financial administration; and marketing and promotion. The student dimension includes various modes of delivery from fully face-to-face through supported distance and independent distance to fully online. The two-dimensional model is presented in Figure 2.

According to Davis et al. (2000), this two-dimensional model offers several advantages. First, it gives the ability to examine transnational programs without having to draw distinction between the student perspective and the provider perspective. Second, it separates the characteristics that describe business models from those that describe teaching and learning models. And finally, it enables the examination of the relationship between the transnational program provider and its partner institution.

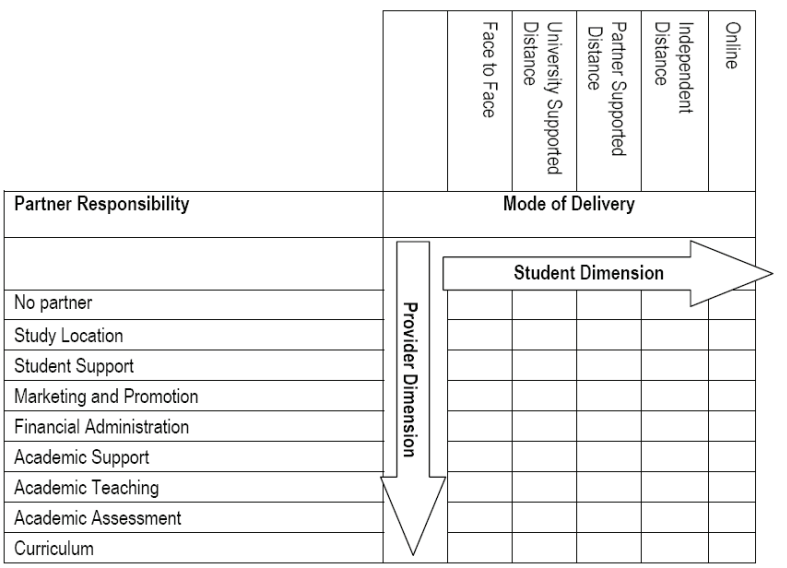
Factors Determining the Demand for and Supply of Transnational Education

The changing nature of demand and supply in transnational tertiary education that has emerged since the late 1990s has been described as the "business of borderless education" (Cunningham et al., 2000). The demand varies between countries, whereby countries with more rigid education systems tend to attract more transnational providers. Here, it often acts as a significant access route to higher education and the acquisition of internationally recognised qualifications (although not necessarily nationally recognised ones). According to Adam (2001), the main determinants of demand include: cost of the program; brand name of the provider and product; value-added from the program; reputation, quality and perceptions of the program; the national/international recognition of the program; the convenience and nature of delivery; and, the level of competition (dissatisfaction/failings of traditional education provision). These determinants can be further separated into *pull factors* that attract students to imported education and *push factors* that repel students from home provision.

According to Marginson (2004) demand for transnational, or *cross-border*, tertiary education in Asia-Pacific is driven by three factors: (1) insufficient supply of places in local universities, (2) globalisation of work force, and (3) potential status and mobility associated with and acquisition of a foreign degree.

Demand is driven by three factors. First, in many nations there are insufficient places in reputable degree-granting institutions at home. Second, there are expanding opportu-

Figure 2. Two-dimensional model of offshore provision (as presented in Davis et al., 2000, p. 41)



nities for globally mobile labour in fields such as business services, ICTs and scientific research. Education in the USA or another English-language nation provides favourable positioning in global labour markets. Third, graduates can use foreign degrees to secure status and mobility benefits. They enhance employment potential at home and abroad, and may open the way to migration to the nation of education or elsewhere. (Marginson, 2004, p. 85)

Research reveals three main determinants of the supply of transnational education: costs of production of programs (that decrease with increasing scale); the nature of the national market; and, the existence of legal regulation and controls (Marginson, 2002). According to Knight (2004), much of the impetus for transnational education comes directly from the need to raise income by both traditional and *for profit* education providers–the former are increasingly seeking new ways to increase their funding. The supply of transnational education provision is also encouraged by the increasing technical ease of delivery through the use of the Internet and other technologies.

Typical Transnational Program: Operational Characteristics

According to the Confederation of European Union Rectors' Conferences (2001) report, transnational education in Europe is largely confined to business subjects (especially MBAs), information technology, computer science and the teaching of widely spoken languages, for example, English, Spanish, and German. A typical transnational program offered by Australian universities is also in the field of the study of business, information technology, and education (Davis et al., 2000; Welch, 2002); in the past few years, health has emerged as a popular field of study for transnational students (AVCC, 2005).

Davis et al. (Davis et al., 2000), having conducted a survey of Australia's offshore programs, provide a list of further characteristics of a typical Australian transnational education program. Such a program is offered in Hong Kong, Malaysia or Singapore; these countries host the largest number of Australian transnational programs, as evidenced in Table 1; they also provide the largest number of transnational students, as evidenced in Table 2; together, these markets account for 65% of students in Australian transnational programs.

Table 1. Current offshore programs of Australian universities (by year of first intake), pre-2000–2003 (AVCC, 2005, p. 11)

COUNTRY	Pre - 2000	2000	2001	2002	2003	Total
China	98	30	22	24	24	**200**
Hong Kong	154	21	26	23	16	**227**
Indonesia	15	3	2	1	3	**25**
Malaysia	174	59	28	24	29	**321**
Singapore	194	43	30	58	53	**375**
Other	260	62	39	43	18	**421**
TOTAL	895	218	147	173	143	1,569

Table 2. International students: Top 5 markets by detailed transnational mode (IDP Education Australia, 2004, p. 12)

Rank	Distance online	Number	Growth	Offshore on-campus	Number	Growth
1	Malaysia	3,846	-29%	Singapore	19,986	3%
2	Singapore	2,952	-16%	Hong Kong	9,351	-17%
3	Hong Kong	1,952	-25%	Malaysia	8,126	17%
4	China	1,867	23%	China	5,472	18%
5	Canada	807	-15%	Vietnam	955	47%
	Total	16,053	-15%	Total	41,162	1%

A typical Australian transnational education program involves full-time attendance and, in terms of delivery mode, relies on face-to-face teaching or supported distance education; involves a partner which is a private education institution or public education institution; and, awards an Australian qualification (Davis et al., 2000). Recent statistics, presented in Table 2, confirm the prevalence of transnational programs that rely on face-to-face interaction and, with the exception of the Hong Kong market, their increasing growth; the figures also indicate a decline, with the exception of China, in the demand for online programs. Overall, in 2004 the number of distance online students declined by 15% on semester two, 2003, while there was a 1% growth in on-campus students (IDP Education Australia, 2004).

In terms of responsibility, the Australian university is responsible for curriculum, teaching assessment, and quality assurance, and allocates to the offshore partner responsibility for provision of study location, marketing, promotion and financial administration. Although, on the whole, the Australian university is responsible for the quality assurance of the program, partner institutions, overseas governments, and international organisations also participate in this responsibility (IDP Education Australia, 2000).

The Importance of the Face-to-Face Component

Although many universities view online learning as an economic alternative to face-to-face teaching (Davis & Meares, 2001), online learning cannot be regarded as a suitable alternative in transnational settings (Emil, 2001). This view is also supported by Tomasic who claims that:

Electronic delivery of courses to off-shore destinations is unlikely to be seen as an acceptable substitute for face-to-face delivery, although greater use of electronic means to deliver parts of courses may be acceptable. (Tomasic, 2002, p. 11)

Fully-online provision of transnational programs raises many concerns regarding the learning experience, particularly about the extent of feedback and guidance that can be provided to students (Knipe, 2002). Debowski (2003) agrees that fully-online provision of offshore programs is generally perceived to be less effective than options including a face-to-face component. She emphasises the strong recognition of the value of (Australian) academics meeting and interacting with their offshore student population; such regular teaching input by these academics significantly enriches the transnational program (Debowski, 2003).

Another aspect of transnational education that benefits from face-to-face interactions is localisation of teaching. As Ziguras (2000) pointed out, the curriculum of a transnational program is usually standardised across several campuses, which may be located in different countries. While the curriculum is sometimes tailored to local conditions, the modifications are usually minimal; they may only involve assignment questions, for example. In such circumstances, teachers, through face-to-face interaction, can play an important role contextualising and interpreting the content of study materials to make it useful for their students.

The importance of the face-to-face communication and the need for localisation of transnational programs was also raised by Evans and Tregenza (2002) who examined a range of transnational programs offered in Hong Kong by Australian universities in collaboration with Hong Kong partner institutions. They concluded that Hong Kong students seek and expect face-to-face contact. These findings are supported by Miliszewska (2007) who reports on a recent study of the perceptions of transnational students in Hong Kong on fully-online provision of transnational programs. The study found that the students overwhelmingly opposed an online-based delivery model–the opposition ranged from 84% to 100%–and, instead, preferred a blended delivery format; they emphasized the importance of face-to-face interaction, and regarded the Internet as a useful, but only supplementary, means of support.

FUTURE TRENDS

Given the importance of face-to-face interaction and decreasing interest in transnational programs if provided fully online, the future belongs to programs that include face-to-face interaction facilitated largely by an offshore partner of the educational provider (Davis & Meares, 2001; Emil, 2001; Tomasic, 2002; Ziguras & Rizvi, 2001). Ziguras (2002) uses the term *joint delivery* to describe such programs:

Evidence internationally shows that fully on-line delivery is proving unpopular except in small niche programmes, due to the lack of face-to-face contact ... Perhaps the best approach, both in terms of mode of delivery and financial risk, is seen to be "joint delivery" with local, established partners, using on-line delivery in some form. (Ziguras, 2002)

The shift in perception of providing transnational education is also likely to continue. Already, the emphasis has moved from educational aid and promotion of international understanding to educational trade; the focus is on expanding access, and packaging and marketing higher education offshore (Leask, 2004; Marginson, 2004). De Vita and Case (2003, p. 384) further argue that transnational education in particular is a consequence of the marketisation of higher education and *the competitive rush for international students and their money*. This view is also supported by Feast and

Bretag who, commenting on the increasing financial motivation of transnational education programs, concluded:

Distasteful as it may be to the many educators working in transnational settings who are committed to genuine cross-cultural exchange, transnational education is a multi-million dollar 'business', motivated as much by profits as by teaching and learning objectives. (Feast & Bretag, 2005, p. 64)

CONCLUSION

The growth of the transnational education market is set to continue, particularly in South East Asia. It is estimated that the demand for transnational higher education in Asian countries (excluding China) will reach nearly 500,000 students by 2020 (GATE, 2000). In addition, a 2002 report by IDP Education Australia (2002) predicts that the demand for international education will increase four-fold from 1.8 million students in the year 2000 to 7.2 million students in 2025, and Hyam (2003) concludes that *by 2025 approximately half of all international students enrolled in Australian universities will be transnational* (p. 8). With rapid expansion of the transnational education market, more and more universities will join the ranks of transnational education providers, or expand their transnational education offerings (Leask, 2004).

Advances in technology, and the Internet in particular, have created new ways of delivering education, and fully-online provision of transnational programs has been viewed by many providers as an economic alternative to face-to-face teaching (Davis & Meares, 2001). However, it appears that despite earlier predictions that globally offered fully-online programs would dominate the transnational education market, Web-supported face-to-face delivery is likely to continue as a principal model of transnational tertiary education programs. As Ziguras and Rizvi (2001) pointed out:

Transnational education providers need to remember that the habitual ways of teaching and learning are resilient not because they are the most effective means of 'delivering information', but because of the richness of the learning relationship that are developed through ongoing face-to-face interaction. (Ziguras & Rizvi, 2001, p. 10)

REFERENCES

Adam, S. (2001). *Transnational education project: Report and recommendations*. Confederation of European Union Rectors' Conferences, University of Westminster. Retrieved December 8, 2007, from http://www.crue.org/espaeuro/transnational_education_project.pdf

AVCC (Australian Vice Chancellors' Committee). (2005, January). *Report*. Retrieved December 8, 2007, from http://www.avcc.edu.au/documents/publications/stats/International.pdf

Bjarnason, S. (2005). MBA in higher education management. Seminar presentation. *Observatory on higher borderless education*. Retrieved December 8, 2007, from http://www.obhe.ac.uk/resources/speeches/mba_higher_education.pdf

Confederation of European Union Rectors' Conferences. (2001). Transnational Education Project: Report and recommendations. In *Proceedings of the Conference on Transnational Education*, Malmö, Sweden. Retrieved December 8, 2007, from http://www.unesco.org/education/studyingabroad/highlights/global_forum/reference/tne.doc

Cunningham, S., Ryan, Y., Stedman, L., Tapsall, S., Bagdon, K., Flew, T., & Coaldrake, P. (2000). *The business of borderless education* (pp. 18-23). Canberra: DETYA.

Davis, D., & Meares, D. (Eds.). (2001). *Transnational education: Australia online – critical factors for success*. Sydney: IDP Education Australia.

Davis, D., Olson, A., & Bohm, A. (Eds.). (2000). *Transnational education providers, partners and policy: Challenges for Australian institutions offshore*. Canberra: IDP Education Australia.

Debowski, S. (2003). Lost in internationalised space: The challenge of sustaining academics teaching offshore. In *Proceedings of the 17th IDP Australian International Education Conference: Securing the Future for International Education*. Melbourne, Australia. Retrieved December 8, 2007, from http://www.idp.com/17aiecpapers/

DEST (Department of Education, Science and Training). (2005). *A national quality strategy for Australian transnational education and training: A discussion paper*. Retrieved December 8, 2007, from http://aei.dest.gov.au/AEI/GovernmentActivities/QAAustralianEducationAndTrainingSystem/QualStrat_pdf.pdf

De Vita, G., & Case, P. (2003). Rethinking the internationalisation agenda in UK higher education. *Journal of Further and Higher Education*, *27*(4), 383-398.

Emil, B. (2001). Distance learning, access, and opportunity: Equality and e-quality. *Metropolitan Universities*, *12*(1), 19.

Evans, T., & Tregenza, K. (2002). Academics' experiences of teaching Australian "non-local" courses in Hong Kong. In *Paper presented at the Australian Association for Research in Education Conference, Crossing Borders: New Frontiers for Educational Research*, Brisbane, Australia. Retrieved December 8, 2007, from http://www.aare.edu.

au/02pap/eva02510.htm

Feast, V., & Bretag, T. (2005). Responding to crisis in transnational education: New challenges for higher education. *Higher Education Research and Development*, *24*(1), 63-78.

GATE (Global Alliance for Transnational Education). (2000). *Demand for transnational education in the Asia Pacific*. Washington: Global Alliance for Transnational Education.

Goodfellow, R., Lea, M., Gonzalez, F., & Mason, R. (2001). Opportunity and e-quality: Intercultural and linguistic issues in global online learning. *Distance Education*, *22*(1), 65-84.

Hyam, L. (2003). Australian higher education and quality: International issues, challenges and opportunities. Keynote address. In *Proceedings of the Australian Universities Quality Forum 2003*. AUQA occasional publication. Retrieved December 8, 2007, from http://www.auqa.edu.au/auqf/2003/program/papers/Hyam.pdf

IDP Education Australia. (2002). *The global student mobility 2025: Forecasts of the global demand for international higher education* (Report). Canberra: IDP Education Australia.

IDP Education Australia. (2004). *International students in Australian universities* (Report, semester 2). Canberra: IDP Education Australia. Retrieved December 8, 2007, from http://www.idp.com/research/fastfacts/Semester%20Two%202004%20-%20Key%20Outcomes_Web.pdf

Knight, J. (2004). Internationalism remodeled: Definition, approaches and rationales. *Journal of Studies in International Education*, *8*(1), 5-31.

Knipe, D. (2002). The quality of teaching and learning via videoconferencing. *British Journal of Educational Technology*, *33*(3), 301-311.

Leask, B. (2004). Transnational education and intercultural learning: Reconstructing the offshore teaching team to enhance internationalization. In *Proceedings of the Australian Universities Quality Forum 2004,* (pp. 144-149). Retrieved December 8, 2007 from http://www.auqa.edu.au/auqf/2004/proceedings/AUQF2004_Proceedings.pdf

Machado dos Santos, S.M. (2002). Regulation and quality assurance in transnational education. *Tertiary Education and Management*, *8*(2), 97-112.

Marginson, S. (2002). The phenomenal rise of international degrees Down Under. *Change*, *34*(3), 34-43.

Marginson, S. (2004). Don't leave me hanging on the Anglophone: The potential for online distance higher education in the Asia-Pacific region. *Higher Education Quarterly*, *58*(2/3), 74-113.

Miliszewska, I. (2007). Is it fully "on" or partly "off?" The case of fully-online provision of transnational education. In *Journal of Information Technology Education, 6,* 499-514.

Tomasic, R. (2002, October 2-4). Guanxi and sustainable teaching and research programs in business and law in the People's Republic of China. In *Proceedings of the 16th Australian International Education Conference*, Hobart. Retrieved December 8, 2007, from www.businessandlaw.vu.edu.au/cicgr/Tomasic_p.pdf

UNESCO & Council of Europe. (2001). Code of good practice in the provision of transnational education. Bucharest: UNESCO-CEPES. Retrieved December 8, 2007, from *http://www.cepes.ro/hed/recogn/groups/transnat/code.htm*

van der Vende, M.C. (2003). Globalisation and access to higher education. *Journal of Studies in International Education*, *7*(2), 193-206.

Vignoli, G. (2004). *What is transnational education? Online document*. Retrieved December 8, 2007, from http://www.cimea.it/servlets/resources?contentId=2831&resourceName=Inserisci%20allegato

Welch, A. (2002). Going global? Internationalizing Australian universities in a time of global crisis. *Comparative Education Review*, *46*(4), 433-473.

Wilson, L., & Vlãsceanu, L. (2000). Transnational education and the recognition of qualifications. In *Internationalization of higher education: An institutional perspective* (pp. 75-85). Bucharest: UNESCO-CEPES Papers on Higher Education.

Ziguras, C. (2000). *New frontiers, new technologies, new pedagogies. Educational technology and the internationalisation of higher education in South East Asia*. Melbourne, Australia: Monash Centre for Research in International Education.

Ziguras, C. (2002, October). *Education beyond our shores: Defining the way forward* (workshop report). International Policy & Development Unit, New Zealand Ministry of Education. Retrieved December 8, 2007, from http://www.minedu.govt.nz/web/downloadable/dl7382_v1/workshop-report-final.doc

Ziguras, C., & Rizvi, F. (2001). Future directions in international online education. In D. Davis & D. Meares (Eds.), *Transnational education: Australia online* (pp. 151-164). Sydney, Australia: IDP Education Australia.

KEY TERMS

Educational Program/Course: A set of units/subjects, which lead to an academic qualification, for example, a degree.

Franchise Programmes: Study units of one higher education institution adopted by and taught at another institution, although the students formally obtain their qualification from the originating institution.

Fully-Online Education: Mode of education with no traditional campus component and no face-to-face interaction. All interactions with study content, as well as staff and students is conducted online.

Joint Degree: A degree awarded by more than one higher education institution.

Offshore Provision: Offshore provision is the export of higher education programs from one country to another.

Transnational Education: All programs in which students are studying in a country other than the one in which the institution providing the program is located. Australian transnational education includes a mandatory face-to-face component.

Web-Supported Education: Mode of education in which online information is used to supplement traditional forms of delivery (face-to-face), and student participation online is optional.

Category: Networking & Telecommunication

Pervasive Wireless Sensor Networks

David Marsh
University College Dublin, Ireland

Song Shen
University College Dublin, Ireland

Gregory O'Hare
University College Dublin, Ireland

Michael O'Grady
University College Dublin, Ireland

INTRODUCTION

Throughout the history of computing, there has been a trend for the ratio of processing elements to people to increase, resulting in the creation and popularization of new usage paradigms. At the start of the modern computer age, many individual users shared a single mainframe in one central location. In the early 1980s, however, significant developments in microprocessor technologies ushered in the desktop era, resulting in a one-to-one correspondence between individual users and their computers. Computer resources were now intrinsically distributed. The growth of the internet allowed these resources to connect to each other. The *pervasive computing* paradigm is the next logical stage in this trend, resulting in the original computer-human ratio reversing, so that multiple computational devices are available to each individual user. In reality, this point was passed a number of years ago. Mobile phones, personal digital assistants (PDAs), portable music players, as well as numerous embedded devices that people now take for granted, has resulted in computing technologies being embedded into the fabric of everyday life. Thus, for the first time, the desire of computing resources being available on an anywhere, anytime basis is a realistic objective.

In addition to computing being available everywhere, *pervasive computing* has a second key element. This tenet states that user interaction with these universal computing elements should occur in as natural and intuitive a manner as possible. Thus, *pervasive computing* technology should be assimilated transparently into the user's natural environment.

Rather than deal with the entirety of this broad topic, the focus of this article is to provide an overview of the key developments on one particular technology which is essential to the realization of the *pervasive computing* vision: the *wireless sensor network*.

BACKGROUND

The original vision for *pervasive computing*, originally termed ubiquitous computing and frequently referred to as such, was articulated in 1988 by *Mark Weiser* (1952-1999), then of Xerox's *Palo Alto Research Center* (PARC) in California. The fundamental concept underlying his proposal was that a person's interactions with computers should be as natural and intuitive as possible. One important consequence of this is that interactions should not be localized to a desktop-style interface, but rather, should be embedded within everyday objects, thus facilitating access to computational resources when and where necessary. In essence, it is a fusion of the anytime, anywhere computing concept augmented with an inherent need for embedded and intelligent user interfaces. Weiser observed that the technologies which have the greatest impact are those that people do not regard as technology *per se*, but, rather, as an integral component of their environment. He used writing, perhaps the original precursor to the information technology revolution, to illustrate his concept (Weiser, 1991, p. 94).

Pervasive computing technologies should be seen as an extension of an individual's own capabilities, rather than an interface to a restricted set of predefined abilities. Instead of limiting people to a standard interface, *pervasive computing* envisages many different kinds of devices and interfaces for a myriad of tasks (Abowd, Mynatt, & Rodden, 2002). As an demonstration, *Weiser's* team at *PARC* developed three kinds of devices corresponding to the inch, the foot, and the yard scale, which they entitled tabs, pads, and boards, respectively. These devices were designed to emulate commonly used office objects like post-it notes, paper note books, and bulletin boards, while providing enhanced computational capacities tailored to the scale of the device in question, and portability where appropriate (Weiser, 1991, p. 103).

With the benefit of hindsight, it can be seen that *Weiser's* vision was ahead of its time. Early attempts to construct pro-

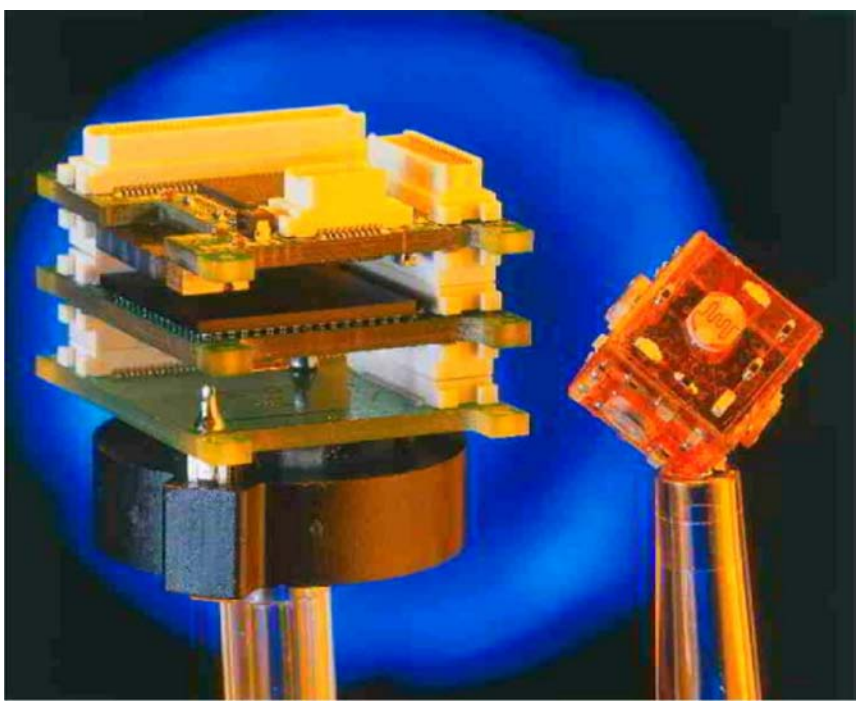
totype systems ran into constant technological hurdles, both in the hardware and software realms. Among the problems encountered were the non-existence of high-capacity wireless networks, display units which struggled to produce output of acceptable quality, a lack of software support for roaming *user contexts* (which had to be developed from scratch), and the absence of practical, easily deployable *sensor networks* to collect the prerequisite information necessary to determine the prevailing context for pervasive systems to adapt their services accordingly. Because of this, the initial focus of pervasive computing researchers was primarily addressed towards remedying the perceived technological deficiencies, rather than on design and usability issues. However, a cross-disciplinary approach is necessary to realize the full *pervasive computing* vision. Thus, research must continue in diverse areas such as Human Computer Interaction (HCI) and software development methodologies.

In the time since Weiser's paper, the parameters and goals of *pervasive computing* have become more clearly defined, so much so that it is already being incorporated into related research areas, such as ambient intelligence (AmI) (Raisinghani, Benoit, Ding, Gomez, Gupta, Gusila, Power, & Schmedding, 2004). With regard to the engineering and software design issues, *pervasive computing* can be seen as an extension of distributed systems and mobile computing (Satyanarayanan, 2001, p. 11). The *wireless sensor network* is a clear product of these areas, as well as energy-efficiency research, *sensor* miniaturization, and a widespread desire for better information gathering technologies. This article is concerned with the technical engineering and software design issues of WSNs, rather than any anthropocentric matters.

WIRELESS SENSOR NETWORKS

The study of *wireless sensor networks* (WSNs) started in earnest in the late 1990s (Pottie, 1998). It was only then that a practical combination of processing, communications, and sensing capabilities could be integrated into a single battery-powered miniature device (see Figure 1). Early research focused on creating networks of these devices, usually through automated, cooperative means initiated by the *sensor nodes* themselves (Intanagonwiwat, Govindan, & Estrin, 2000). Later, research expanded into network security, power management, ensuring adequate sensing coverage, distributed signal processing, and *sensor* fusion, querying and reprogramming nodes, and *sensor* localisztion. Although ongoing development unveils large-scale *sensor*

Figure 1. Examples of sensor nodes: 25mm and 10mm wireless sensor modules - courtesy of Tyndall National Institute, Cork, Ireland. Copyright 2006 Tyndall.



applications on military, industrial, environmental, and home tasks (Akyildiz, Su, Sankarasubramaniam, & Cayirci, 2002; Hill, 2003), the hardware generally available is tailored more towards research environments rather than being optimized for widespread deployment. Indeed, there are differing lines of thought on the direction which this development should follow; some advocate a reduction of *sensors* to near microscopic size (Pister, Kahn, & Boser, 1999), with reliability coming from redundancy, while others see a need for larger, more capable devices (Patra, Kot, & Panda, 2000). Realistically, both of these viewpoints must coexist to provide the continuum of devices needed to fulfil the constraints of the diverse pervasive applications foreseen by *Weiser*.

One example where *wireless sensor networks* have been integrated into a pervasive information system is at the Great Duck Island project (Mainwaring, Polastre, Szewczyk, Culler, & Anderson, 2002). *Sensors* were deployed to monitor environmental weather conditions (such as temperature and humidity) in a nature reserve where a population of cormorants were being studied. The network was designed so that the function of the *nodes* could be altered remotely after deployment. This was very important so that the researchers could avoid disturbing the wildlife. The data gathered from these *sensors* was made available both to PC users (through relational databases) and via a PDA interface, allowing field access to the scientists involved.

Context-Awareness and Adaptation

Contextual awareness and adaptation has been identified as one of the most important features of a *pervasive computing* application (Mattern, 2000). The most commonly cited example of a person's *context* is their actual physical location. When this is known, appropriate information can delivered to the person that is pertinent to their immediate location. A classic example is that of an electronic yellow pages service that sorts its entries according to their distance from the user. However, there are many other elements that can constitute a person's *context*, including those characterized as "when, what and why" (Abowd, et al., 2002, p. 52), as well as with whom, and so on. Without an effective and transparent ability to collect data for use in determining the salient aspects of a user's *context*, a *pervasive computing* application is significantly functionally restricted. While collection of *context* can occasionally occur in an implicit manner, in practice, a *sensor network* is essential for many applications. Traditional *sensor networks* usually have a centralized architecture, with the *sensors* wired to a central work station where all the processing takes place. However, the requirements of *pervasive computing* are such that a less obtrusive sensor network, one that employs wireless communication, is preferable in order to provide the necessary level of sampling density and ease of deployment that is not feasible with wired sensor networks.

POWER MANAGEMENT IN WIRELESS SENSOR NETWORKS

Pervasive computing systems consist of electronic components, many of which are mobile. *Power management* of such mobile systems, where reliable power is not guaranteed, is of great importance. The energy issue is addressed through either intelligently reducing power consumption or developing new energy sources (Want, Farkas, & Narayanaswami, 2005).

Energy scarcity is the single most crucial issue in WSN research at present. All other concerns are viewed within the context of energy consumption. At present, commercially available hardware operates for only days or weeks without any power *management*. While battery capacity is increasing, it is starting from a woefully inadequate level from the point of view of the long term deployment of pervasive systems. Two key strategies have been identified for tackling this problem – seeking alternative energy resources, and reducing power consumption, respectively.

Acquiring Energy From Environmental Sources

Acquiring energy from environmental sources, such as solar or vibrational energy, has potential (Roundy, Wright, & Rabaey, 2003). Unfortunately, these sources usually only generate tens/hundreds of microwatts per second, falling short of the power consumption of a typical *sensor node* by orders of magnitude. This necessitates the use of batteries as a supplementary (or often sole) power source. Because of their fixed energy capacity, relying on batteries limits the energy available to the *sensor node*, and hence, the operational lifetime. Because of this energy constraint, almost all WSN research has been framed by the trade-off between performance and operational lifetime.

Reducing Power Consumption

While creating more efficient hardware helps reduce the demand placed on a battery, the primary method for reducing power consumption is to activate a *sensor node's* components only when they are needed (Hill, 2003). This can extend as far as suspending the processor until an external stimulus reactivates it. The choice to activate/deactivate components (radio, sensors, etc.) is not difficult for a single *sensor* to evaluate. However, as part of a network, each element must also consider the effects of this choice on the operation of neighbouring nodes, and of the network as a whole. This requires coordination between *nodes* to ensure the network can continue operating at a sufficient level to provide the required quality of service demanded by the pervasive application. However, network-wide coordination is con-

sidered harmful in the majority of cases since the energy and bandwidth cost exceed the benefits gained by using a network-level optimal solution. This leads to the use of distributed algorithms and protocols which don not oblige the *sensor nodes* to know any more about the network than they can detect locally (so called "localized algorithms") (Qi, Kuruganti, & Xu, 2002, p. 286).

Because of the many competing factors involved in making a decision about whether to activate/deactivate a component, or indeed an entire *sensor node*, and the lack of system-wide information, *power management* methods profit from the use of intelligent decision making. One technology proposed to handle this is intelligent *agents* (Tynan, O'Hare, Marsh, & O'Kane, 2005; Fok, Roman, & Lu, 2005). *Agents* are independent, autonomous software entities capable of cooperative behavior. They can deliberate on the set of information they possess, and can act on this information in a way that will bring about their goals. Given the dual goals of reducing local power usage and preserving the operation of the surrounding area of the network, appropriately programmed *agents* can find the balance point that best satisfies these conditions.

The issue of *power management* covers many separate and often competing concerns. All *sensor nodes* have a radio, a CPU with memory, and a set of *sensors*. Other components, such as various actuators for influencing the environment of the pervasive application, may also be included. The radio is often the most power hungry component, so minimizing the time that the radio is active is often the first step in reducing battery depletion. Striking a balance between low radio usage and maintaining the communications capacity of the network is a much studied topic (Papadopoulos & McCann, 2004). If a node decides when it will deactivate its radio independently of other matters such as information generated by its *sensors*, it may not be able to communicate when most needed. Furthermore, if a *node* is deemed to be redundant for the purposes of preserving the network topology by its network control software, this does not mean it can be deactivated. Sensing coverage must also be sustained.

AUTONOMIC WIRELESS SENSOR NETWORKS

Since one of the primary goals of pervasive computing is to make the computer "disappear", *sensor networks* must be designed in such a way so as to require minimum human intervention. Self-management, encompassing self-configuration, self-optimization, self-healing, and self-protection (Kephart & Chess, 2003), are integral to the autonomous operation of this core component of a pervasive system. These standards could be applied to the rest of the system too, but since a *sensor network* is likely to have the greatest number of elements, each with a relatively high chance of failing, it is of particular importance in this part of the system. Most protocols for *WSNs* are self-configuring. Self-optimization is also attempted, though usually within the constraints of local knowledge. Self-healing, that is the ability of the network to overcome damage, is usually achieved by ensuring redundancy (deploying more than the minimum number of sensor nodes that are needed). These additional nodes can be activated when a subset of the network is lost, for instance through energy exhaustion. There are many protocols for networking (Cerpa, & Estrin, 2002; Hsin & Liu, 2004) and sensing coverage (Zhang & Hou, 2005; Wang, Xing, Zhang, Lu, Pless, & Gill, 2003) which exploit this mechanism. Self-healing can also be realized if there is a reserve of self-deploying mobile *nodes* available to replace any lost nodes. Self-protection, mainly dealing with security issues, has received much attention (Perrig, Szewczyk, Wen, Cullar, & Tygar, 2001), not least because *sensor networks* have potential as a military tool.

Mobility and Intelligence

Increasingly, personal computers and mobile devices in the form of portable telephones, personal digital assistants, and embedded microchips in cars, appliances, and other technical artefacts with which people interact everyday, are emerging as crucial enabling elements for *pervasive computing*. In order to make these devices useful, such mobile computers should be sufficiently smart to interact with each other effectively.

There are three forms of mobile intelligence which are relevant to *pervasive computing*. These are:

1. when an intelligent device is passively mobile, as in a mobile phone or PDA;
2. when a piece of software is capable of migrating from device to device in a network of some kind; and
3. when a device is independently mobile, for instance an autonomous unmanned aerial vehicle.

In all instances, *agents* provide a useful computational paradigm which offers intelligence and mobility (when they are migration-capable) in an integrated package. In the first case, the software can use information such as location, orientation, velocity, and so on, as inputs to location-reliant pervasive applications. It must have an idea of the objectives of the user, for example, to get to a train station by a certain time, and relies on the user to act on this information.

The second instance is especially useful for pervasive networks in general, and *sensor networks* in particular, because a mobile *agent* can travel around a network, deciding where to move to next based on the latest information in order to perform management functions or collect data.

This is in contrast to a client-server architecture, in which a reliable connection is required between the server and any other devices which must be controlled. The mobile *agent* method is more robust and has a lower latency (because the agent is located on or near the *nodes* with which it must interact), and so is better equipped to respond to events in the network.

In the third case, the software on the device has its own set of goals, and is capable of performing actions which will fulfill these goals. Wireless *sensor nodes*, once laid out in the environment, might be damaged by environmental factors or expire from energy depletion. Therefore, it is desirable to have *sensor nodes* that are geographically mobile in nature. These *nodes* can move to an area that has become impaired through *node* loss, thus supplementing the extant fixed *nodes*. Additionally, the *nodes* can move to areas where events of interest are occurring to boost the sensing quality there.

FUTURE TRENDS

Driven by the technological achievements on microelectronics, *pervasive computing* will remain a hot topic in the information industry. Future trends will likely deal with the following crucial problems so that *pervasive computing* can be truly ubiquitous. One of the most vital concerns is the management of billions of software entities with different purposes, deployed in multiple locations, and based on different platforms. A further problem is how to protect hosts and applications from malicious attacks. Security of *pervasive computing* will attract more and more attention with the explosion of the amount of information available via pervasive technologies and the increasing maliciousness of attacks. If the public's perception of ubiquitous services sours because of weak security measures, the wide-spread roll-out of *pervasive computing* systems could be delayed by many years, thus, robbing the world of a technology of great potential.

CONCLUSION

Wireless sensor networks will form an essential component of the embedded electronic infrastructure necessary to realize the *pervasive computing* vision. The successful realization of this vision is anticipated to result in significant economic benefits. More than this, however, is the potential of *WSNs* to provide a solid foundation for context-awareness, adaptivity, and personalization of services. In doing this, significant benefits will accrue to end users as a new era of dedicated services, which up until now could only be speculated about, will become embedded in the fabric of everyday life.

REFERENCES

Abowd, G. D., Mynatt, E, D., & Rodden, T. (2002). The human experience. *Pervasive computing*, 1(1), 48-57. New York: IEEE Press.

Akyildiz, I., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: a survey. *Computer networks*, 38(4), 393-422. Berlin: Elsevier.

Cerpa, A., & Estrin, D. (2002). ASCENT: Adaptive self-configuring sensor network topologies. *Computer communication review*, 32(1), 62. New York: ACM Press.

Fok, C. L., Roman, G. C., & Lu, C. (2005). Mobile agent middleware for sensor networks: An application case study. *4th International conference on information processing in sensor networks*, 382-387. New York: ACM Press.

Hill, J. (2003). *System architecture for wireless sensor networks*. Unpublished PhD thesis. University of California, Berkeley, CA.

Hsin, C., & Liu, M. (2004). Network coverage using low duty-cycled sensors: random & coordinated sleep algorithms. *Third international symposium on information processing in sensor networks*, April, 2004. Berkeley, California. New York: ACM Press.

Intanagonwiwat, C., Govindan, R., & Estrin, D.(2000). Directed diffusion: A scalable and robust communication paradigm for sensor networks. *MobiCom*. Boston, MA. August 2000. New York: ACM Press.

Kephart, O. J., & Chess, D. M. (2003). The Vision of Autonomic Computing. *IEEE Computer*, 36(1), 41-50. New York: IEEE Press.

Mainwaring, A., Polastre, J., Szewczyk, R., Culler, D. E., & Anderson, J. (2002). Wireless sensor networks for habitat monitoring. *ACM Workshop on sensor networks and applications*, Atlanta, GA. September, 2002. New York: ACM Press.

Mattern, F. (2000). *State of the art and future trends in distributed systems and ubiquitous computing*. Vontobel TeKnoBase. August, 2000. Retrieved July 31, 2006 from http://www.vs.inf.ethz.ch/publ/papers/DisSysUbiCompReport.pdf.

Papadopoulos, A., & McCann, J. A. (2004). Towards the design of an energy-efficient, location-aware routing protocol for mobile, ad-hoc sensor networks. *15th International workshop on database and expert systems applications*. August, 2004. Zaragoza, Spain. New York: IEEE Press.

Patra, J. C., Kot, A. C., & Panda, G.(2000). An Intelligent Pressure Sensor Using Neural Networks. *IEEE Transactions*

on instrumentation and measurement, Vol. 49(4), 829-835. New York: IEEE Press.

Perrig, A., Szewczyk, R., Wen, V., Cullar, D., & Tygar, J. D. (2001). SPINS: Security protocols for sensor networks. *MobiCom*. Rome, Italy. July, 2001. New York: ACM Press.

Pister, K. S. J., Kahn, J. M., & Boser, B. E.(1999). Smart dust: Wireless networks of millimeter-scale sensor nodes. *Highlight article in 1999 Electronics Research Laboratory research summary*. Berkeley, CA: University of California Press.

Pottie, G. J. (1998). Wireless sensor networks. *IEEE Information theory workshop*. June, 1998. Killarney, Ireland. New York: IEEE Press.

Qi, H., Kuruganti, P.T., & Xu, Y.(2002). The development of Localized algorithms in wireless sensor networks. *Sensors*. Vol. 2, 286-293. New York: New York: IEEE Press.

Raisinghani, M. S., Benoit, A., Ding, J., Gomez, M., Gupta, K., Gusila, V., Power, D., & Schmedding, O. (2004). Ambient intelligence: Changing forms of human-computer interaction and their social implications. *Journal of digital information*. Vol. 5(4), Article No. 271. August 2004. Retrieved August 4, 2006 from http://jodi.ecs.soton.ac.uk/Articles/v05/i04/Raisinghani/.

Roundy, S., Wright, P. K., & Rabaey, J. M. (2003). *Energy scavenging for wireless sensor networks: with special focus on vibrations*. New York: Springer.

Satyanarayanan, M. (2001). Pervasive computing: Vision and challenges. *IEEE Personal communications*. Vol. 8(4), 10-17. August, 2001. New York: IEEE Press.

Tynan, R. O'Hare, G.M.P., Marsh, D., & O'Kane, D. (2005). Multi-agent system architectures for wireless sensor networks. *5th International Conference on Computational Science*. Emory University. Atlanta, GA. May, 2005. Berlin: Springer-Verlag.

Wang, X., Xing, G., Zhang, Y., Lu, C., Pless R., & Gill C. (2003). Integrated coverage and connectivity configuration in wireless sensor networks. *SenSys*, November, 2003. Los Angeles, CA. New York: ACM Press.

Want, R., Farkas, K. I., & Narayanaswami, C. (2005). Energy harvesting and conservation. *IEEE Pervasive Computing*. Vol. 4(1), 14-17. New York: IEEE Press.

Weiser, M.(1991). The computer for the 21st century. *Scientific American*, 265(3), 94-104. Stuttgart, Germany: Holtzbrinck.

Zhang, H., & Hou, J. C. (2005). Maintaining sensing coverage and connectivity in large sensor networks. *International journal of wireless ad hoc and sensor networks*, vol. 1(1-2) 89–124. Philadelphia, PA: Old City Publishing.

P

KEY TERMS

Actuator: In the context of sensor networks, any output device. Actuators allow a WSN node to influence its environment, providing a feedback channel through which its decisions can be enacted.

Agents: In this article, the word *agent* refers to software entities which are capable of displaying autonomous, cooperative, and flexible behavior directed towards achieving a set of internal goals. A group of agents which operate together is called a multi-agent system (MAS).

Context-Awareness: The property of a system that allows it to adjust its behavior based on physical environmental cues, such as location, or user presence or absence. This necessitates the use of sensor networks which can monitor the relevant properties of the environment.

Human Computer Interaction (HCI): The study of how people and computers interact, the effects of these interactions on both the users and the computers, and the design and testing of new user interfaces for the purpose of improving the computer's user-friendliness. HCI combines aspects of computer science, engineering, psychology, sociology, aesthetics, and ergonomics, as well as many other fields.

Pervasive Computing: Also known as ubiquitous computing, this is the study of how computing can be integrated into the environment in a way that makes it easily accessible to users. It includes an emphasis on ease and naturalness of use, and unobtrusiveness is paramount.

Sensor: In the context of sensor networks, any input device other than the communications transceiver, for example, a microphone or barometer. The sensors and transceiver produce a sensor node's input data, which is used to inform its actions.

Wireless Sensor Network: A network comprised of small, communication-enabled microcontrollers with an array of sensors and actuators which rely on batteries or ambient environmental energy for power. These components are often referred to as sensor nodes.

Physiologic Adaptation by Means of Antagonistic Dynamics

Juergen Perl
University of Mainz, Germany

INTRODUCTION

In particular in technical contexts, information systems and analysing techniques help a lot for gathering data and making information available. Regarding dynamic behavioral systems like athletes or teams in sports, however, the situation is difficult: data from training and competition do not give much information about current and future performance without an appropriate model of interaction and adaptation.

Physiologic adaptation is one major aspect of target-oriented behavior, in physical training as well as in mental learning. In a simplified way it can be described by a stimulus-response-model, where external stimuli change situation or status of an organism and so cause activities in order to adapt. This aspect can appear in quite different dimensions like individual biochemical adaptation that needs only milliseconds up to selection of the fittest of a species, which can last millions of years.

Well-known examples can be taken from learning processes or other mental work as well as from sport and exercising. Most of those examples are characterized by a phenomenon that we call antagonism: The input stimulus causes two contradicting responses, which control each other and – by balancing out – finally enable to reach a given target. For example, the move of a limb is controlled by antagonistic groups of muscles, and the result of a game is controlled by the efforts of competing teams.

In order to understand and eventually improve such adaptation, models are necessary that make the processes transparent and help for simulating dynamics like for example, the increase of heart rate as an reaction of speeding up in jogging. With such models it becomes possible not only to analyze past processes but also to predict and schedule indented future ones.

In the Background section, main aspects of modeling antagonistic adaptation systems are briefly discussed, which is followed by a more detailed description of the developed PerPot-model and a number of examples of application in the Main Focus section.

BACKGROUND

Undisturbed limited growth processes in biological systems often are asymptotic, oriented in specific target values. This behavior reflects adaptation to limited resources and delays caused by resource production and transport. Processes of this type theoretically can be modeled rather easily by means of exponential functions – for example, $f(t) = c - \exp(-s \times t/d)$, where c is the target value, s characterizes the deceleration, and d characterizes the delay (see Banister & Calvert, 1980; Banister, Calvert, Savage & Bach, 1975). In practice, however, situations are more complex (see Lames, 1996; Viru & Viru, 1993): The growth process normally is disturbed (stopped, restarted, reduced, intensified) by external effects; capacity limitations cause changes of the temporary process type (phase changes); buffers cause delays of the internal dynamics; seemingly constant parameters turn out to be time-depending. Therefore, often such processes cannot be modeled using continuous functions (e.g., as solutions of differential equations) but have to be calculated iteratively using discrete level-rate-equations, which only piecewise could be approximated by exponential functions.

Physiologic adaptation is a kind of limited biological growth and therefore can be modeled and simulated using such an iterative approach – as we have successfully done with load-performance-interaction and learning in sport. To this aim we developed an approach (PerPot: Performance Potential Metamodel, see Mester & Perl, 2000; Perl, 2002; Perl & Mester, 2001), the central idea of which is that of antagonism: A load input flow feeds in the same way two internal buffers – the strain potential and the response potential. These buffers are connected with a performance potential, the level of which is decreased by a negative strain flow and increased by a positive response flow. Both flows are delayed. The relations between the strain delay and the response delay characterize the interaction of load input and performance output. In case of training or learning the strain delay can be interpreted as fatigue delay, while response delay stands for the delay of recovery.

MAIN FOCUS OF THE CHAPTER

Applying such a model to a test person, after a short phase of calibration the delay parameters are known and the behavior of the test person can be analysed and simulated using PerPot. This approach has successfully been used for detecting striking features like contra-productive overtraining, doping abuse, or threatening collapses as results of overload. PerPot

can predict performance output depending on training or learning input and so can be used for scheduling optimal training or learning processes.

The PerPot Concept

The metamodel PerPot describes physiologic adaptation on an abstract level as an antagonistic process, as is shown in Figure 1. An input flow (which usually is called "load" rate) is feeding identically a strain potential as well as a response potential. From the response potential the performance potential is increased by a positive flow, while the strain potential reduces it by a negative flow. Additionally, there are the following two effects:

If the strain potential is filled over its upper limit, it produces an overflow, which acts on the performance potential as a reducing negative flow. In turn, the difference between the upper limit of the strain potential and its current level indicates how far the situation is from such a dangerous overflow. This difference is called the "reserve" of the system.

Finally, in order to model atrophy, the performance potential continuously looses substance. By mathematical reasons, this loss has to be fed back to the response potential to preserve the potential balance of the system (see Perl, 2003). (From the mathematical point of view, the load rate only plays the role of a pump, which moves the system potential around without violating this balance property.)

All flows show specific delays modelling the time that components need to react. Possible physiologic interpretations are production and transport of biochemical stuff on the micro-level or fatigue and recovery effects on the macrolevel.

Delays are model parameters the model behavior depends on in a quite characteristic way. For instance, the fitness of an athlete can be measured by the correlation of fatigue and recovery delays.

PerPot-Based Simulation and Analysis

Based on the model, a simulation tool has been developed, which in an iterative way calculates the model's behavior in order to follow individual and temporary profiles of load and delays. Basic simulations and analyses can be run using the PerPot-tool by just varying delay parameters and load profiles. As is shown in Figure 2, there are mainly three types of characteristic behavior, which normally are mixed up to a complex behavioral profile:

In the left graphic, the normal adaptation is shown, depending on the relation between the delays DR and DS: If DS is less than DR then the performance reducing strain comes faster then the performance increasing response, therefore causing the well-known super-compensation effect (see Clijsen, van de Linden, Welbergen & Boer, 1988; Friedrich & Moeller, 1999). In turn, if DS is greater than DR, the response effect comes first, causing increasing performance, which later is balanced out with increasing strain.

The graphic in the middle demonstrates how a switch on of load starts the performance development (like in the left graphic) and a switch off causes a characteristic decrease of development, which is a combination of recovery and atrophy.

The right graphic shows the effect of overload, which causes internal load overflow with collapsing reserve and performance (see Hartmann & Mester, 2000).

Figure 1. PerPot: Structure and parameters

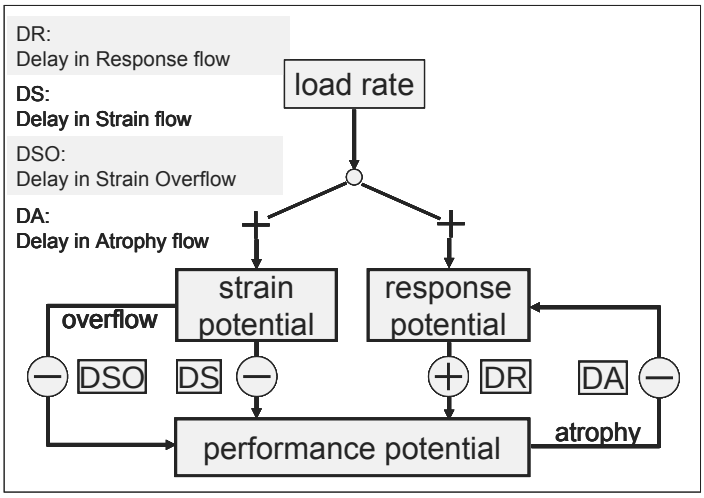


Figure 2. Characteristic types of PerPot behavior

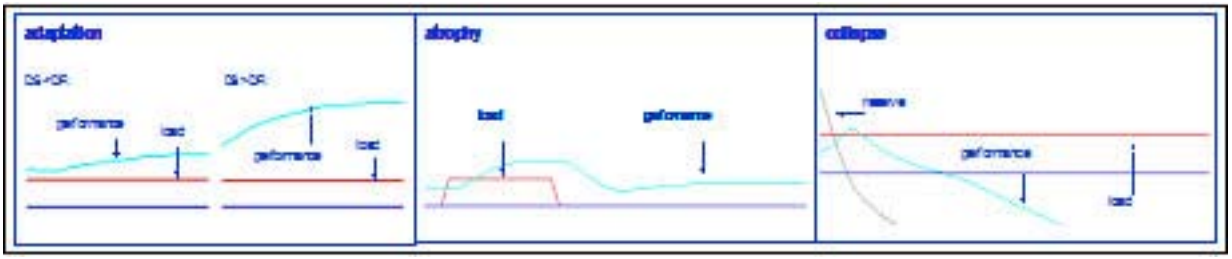
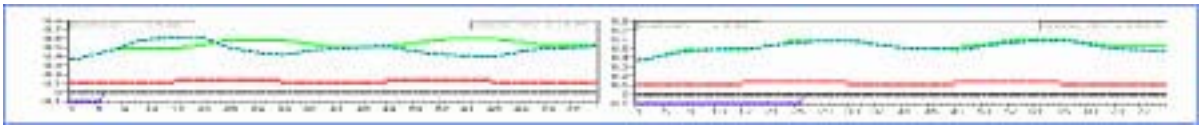


Figure 3. Precision of prediction depending on the load dynamics and/or the number of data used for delay calculation (violet dots on the bottom mark the interval of used data)



By means of the described simulation features, also future behavior of an adapting system can be predicted – if the respective load profile, as well as the delay time series is known for example, from a schedule.

PerPot-Based Prediction

In the simplest way, constant delay values can be deduced from the past load and performance profiles by calibration. Consequently it can be asked how many "current" data are necessary to calculate delay parameters, which not only allow for simulating the current but also the future performance values (see Ganter, Witte & Edelmann-Nusser, 2006; Perl, 2004; Perl, 2005).

In Figure 3 it is demonstrated that 5 data seem not to be sufficient for a proper prediction, while 25 data seem to fit rather well. The truth is that the typical dynamic changes, for example, from low to high and/or from high to low load, are necessary to give the model enough information about the characteristic adaptation behavior. Moreover, this method of prediction can be used for online adaptation, where every new pair of load and performance values can be used for adapting the delay values to the changing situation respectively to the changing status of the athlete.

One reason for a changing status of the athlete can be a decrease of condition and should influence the delay values. In turn, changing delay parameters can be indicators of effects like this. For example, a closer look to the right graphic of Figure 3 shows a significant deviation between original and simulated performance in the "far future" (rightmost area). One could think that prediction becomes worse the more the point in time is different from the available values. The truth is that the calculated mean delay values do not reflect the situation correctly, even if all values are used for delay calculation (marked by violet dots on the bottom line): Instead, stepwise online adaptation results in a nonconstant DS-profile which gives a rough idea of the tendency of decreasing shape.

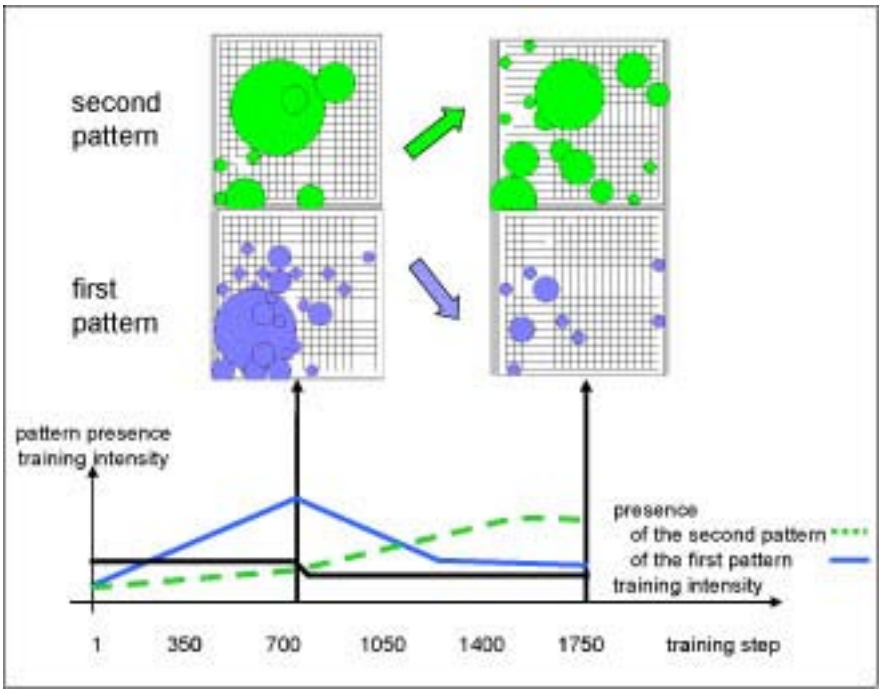
Example "Jogging": Online Speed Control

Load-performance-interaction can be analyzed by means of a special tool that was derived from PerPot (see Perl & Endler, 2006), where speed is taken as load input and heart rate is taken as performance output. Additional impact parameters are positive and negative slopes from the course profile, age and fitness of the athlete, and current position from the run. For example, the recovery delay DR will increase by fatigue, which depends on age and fitness as well as on the time already used on the run.

Based on this information, the optimal speed profile can be calculated given a heart rate profile as objective function, which can help for improving respectively saving the runner's performance and health.

In turn, the correspondence between speed and heart rate can be used for controlling the run just watching the heart rate – that is, optimizing the speed by matching the intended heart rate. This approach exemplarily was got to work successfully in a test, where the test person in the first hill marathon missed the scheduled time of three hours by only 10 minutes or 5.6% – using a heart rate meter as only control instrument.

Figure 4. Cooperative learning (left) vs. replacing learning (right)



Example "Learning": Dynamic Network Control

Because of its positive effect on configuring learning processes, PerPot has been used successfully for improving the functionality of neural networks of type Kohonen Feature Map (see Kohonen, 1995): Supporting every neuron with an individual PerPot, which controls the learning process of the neuron, results in a Dynamically Controlled Network (DyCoN, see Perl, 2002) that can learn continuously and can adapt its learning behavior dynamically to the flow of learning stimuli (see Memmert & Perl, 2006).

Not least, this approach enables for analysing and simulating types of learning processes like cooperative or replacing learning (see Perl & Weber, 2004), as is demonstrated in Figure 4. In the first phase the blue pattern, which contains a lot of information of the green one, is trained to the net. In the second phase the green pattern, which contains only few information of the blue one, is trained to the net. Even though the second training is done with rather low intensity, the green pattern quickly becomes dominant, while the blue one begins to fade out by atrophy.

Obviously, those effects remind on aspects of natural learning and therefore have been subject to respective investigations.

Example "Interaction": Feed Back Control in Games

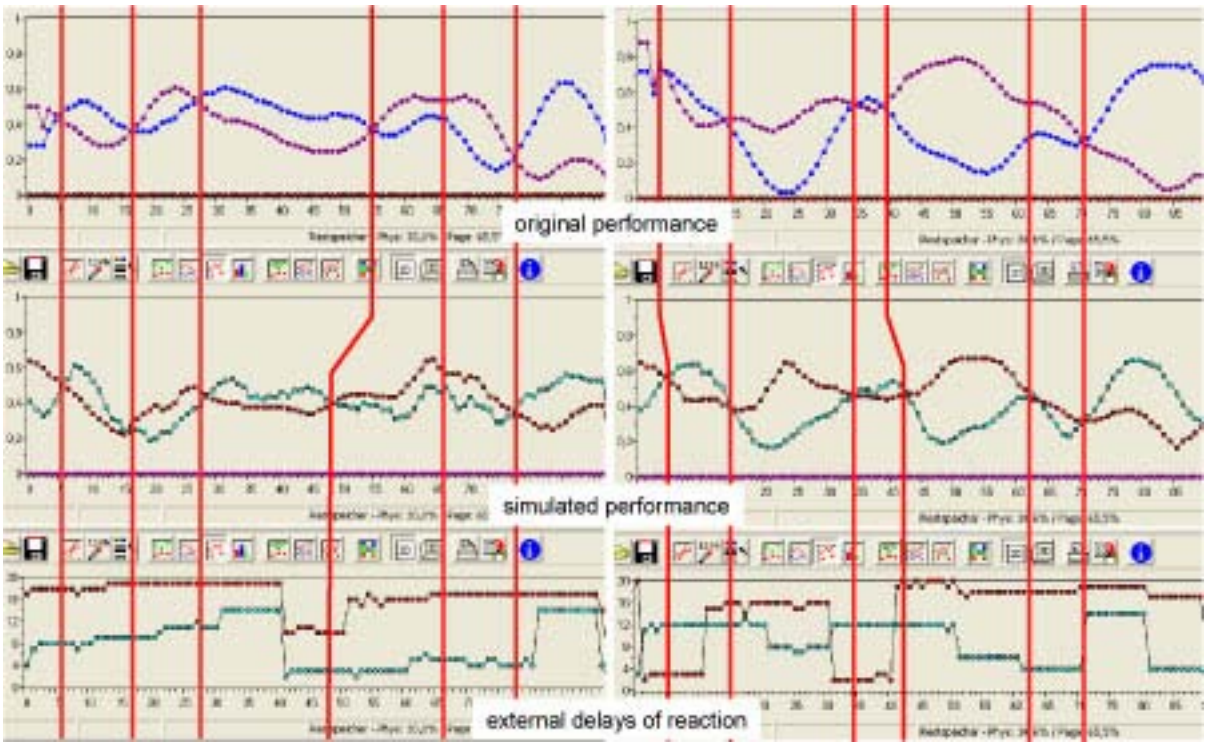
The dynamics of team behavior in games show rather inhomogeneous distributions of the levels of activity and effectiveness, where phases of great effort are followed by phases of reduced activities, which are used by the opponent team for increasing its pressure (see Lames, 2006).

Based on the interpretation that the subsequent effectiveness of the one team is a kind of performance, which in a delayed way is caused by the past and present pressure or load from the opponent team – that is, its activities and scoring – the game can be understood as a symmetric process of load-performance-interaction. Such symmetric feed-back-systems very often show an oscillating-circuit-behavior, where both components show oscillations that are in antiphase to each other. Therefore an approach has been used for modeling, where two exemplars of PerPot model the respective dynamics of each team, the interaction of which is controlled by team-specific delays of reaction (see Perl, 2006).

In Figure 5, two games of Germany haven been analyzed exemplarily:

In the first game against Denmark, the German team shows two long phases of large delay of reaction with respectively corresponding maxima of performance. In contrast, the Danish team shows smaller delays that cause comparably small fluctuations of performance.

Figure 5. Left: Germany (red) vs. Denmark (blue) (22 : 20). Right: Germany (blue) vs. Croatia (red) (23 : 23)



In the second game the Croatian team plays more or less the role of Germany from the first example: Large external delays cause comparably high performance. In contrast, the German team shows smaller delay values, which however are not as small as those of Denmark. The result is a German performance profile, which lies between those of Denmark and Croatia and the frequency of which is smaller than that of Denmark

The main question is whether there in general are significant differences between the profiles of reaction delays, which then could be used for classifying and predicting a team's tactical behavior in a game.

FUTURE TRENDS

As some of our advanced projects show, the presented approach can improve the understanding of delayed antagonistic reaction and interaction not only in sports but also in a wide range of "real-life"-processes like learning or weight-watching, where knowledge acquisition or weight-reduction are delayed antagonistic reactions to knowledge resp. food input. Last but not least first investigation has successfully been done in the field of rehabilitation, where the patient's state is characterized by a complex adaptation process the prediction of which could be improved if its antagonistic dynamics could be better understood and modeled.

CONCLUSION

Antagonistic dynamics makes a system's behavior complicate and difficult to predict. Simple mathematical approximation or stochastic analyses like crosscorrelation approaches cannot provide process-related interpretations and understanding and therefore fail in prediction.

There are two major fields of application of antagonistic adaptation analysis: From the theoretical point of view, complex system behavior can be analysed and understood in a more dynamic way, were the last two approaches dealing with learning and handball can be taken as examples for. From the practical point of view, predicting developed performance depending on given load profiles can improve training scheduling, exercising, and even leisure activities a lot. Model-based scheduling of adaptation processes so can improve efficiency and effectiveness of exercises and training processes and help for avoiding overload and critical situations.

In general, the experience from a lot of applications show that the developed approach can help for improving analysis as well as prediction of the behavior of antagonistic adaptation systems.

REFERENCES

Banister, E. W., Calvert, I. W., Savage, M. V., & Bach, I. M. (1975). A system model of training for athletic performance. *Australian Journal of Sports Medicine, 7*(3), 57-61.

Banister, E. W. & Calvert, T. W. (1980). Planning for future performance: Implications for long term training. *Canadian Journal of Applied Sport Sciences, 5*(3), 170-176.

Clijsen, L. P. V. M., van de Linden, J., Welbergen, E., & Boer, R. d. W. d. (1988). Supercompensation in external power of well-trained cyclists. In E. R. Burke & M. M. Newsom (Eds.), *Medical and scientific aspects of cycling* (pp. 133-144).

Friedrich, W. & Moeller, H. (1999). Zum Problem der Superkompensation. Leistungssport, 29(5), 52-55.

Ganter, N., Witte, K., & Edelmann-Nusser, J. (2006). Einsatz von antagonistischen Trainings-Wirkungs-Modellen zur Leistungsprädiktion im Radfahren. J. Edelmann-Nusser & K. Witte (Eds.), Sport und Informatik IX, 43-48.

Hartmann, U. & Mester, J. (2000). Training and overtraining markers in selected sport events. *Medicine and Science in Sports and Exercise, 1*, 209-215.

Lames, M. (1996). Die komplexe sportliche Leistung - Ein nichtlineares dynamisches System? J.-P. Janssen, K. Carl, W. Schlicht & A. Wilhelm (Eds.), Synergetik und Systeme im Sport, 179-197.

Lames, M. (2006). Modelling the interaction in game sports – Relative phase and moving correlations. *Journal of Sports Science and Medicine, 5*(2), 556-560.

Kohonen, T. (1995). *Self-organizing maps*. New York: Springer.

Memmert, D. & Perl, J. (2006). Analysis of game creativity development by means of continuously learning neural networks. In E. F. Moritz & S. Haake (Eds.), *The enginieering of sport, 6*(3), 261–266.

Mester, J. & Perl, J. (2000). Grenzen der Anpassungs- und Leistungsfähigkeit aus systemischer Sicht – Zeitreihenanalyse und ein informatisches Metamodell zur Untersuchung physiologischer Adaptationsprozesse. Leistungssport, 30(1), 43-51.

Perl, J. (2002). Adaptation, antagonism, and system dynamics. In G. Ghent, D. Kluka & D. Jones (Eds.), *Perspectives – The Multidisciplinary Series of Physical Education and Sport Science, 4*, 105-125.

Perl, J. (2003). On the long term behaviour of the performance-potential-metamodel PerPot: New results and approaches. *International Journal of Computer Science in Sport, 2*(1), 80-92.

Perl, J. (2004). PerPot – A meta-model and software tool for analysis and optimisation of load-performance-interaction. *International Journal of Performance Analysis of Sport-e, 4*(2), 61-73.

Perl, J. (2005). Dynamic simulation of performance development: Prediction and optimal scheduling. *International Journal of Computer Science in Sport, 4*(2), 28-37.

Perl, J. (2006). Qualitative analysis of team interaction in a game by means of the load-performance-metamodel PerPot. *International Journal of Performance Analysis in Sport, 6*(2), 34-51.

Perl, J. & Endler, S. (2006). Training- and contest-scheduling in endurance sports by means of course profiles and PerPot-based analysis. *International Journal of Computer Science in Sport, 5*(2), 42-46.

Perl, J. & Mester, J. (2001). Modellgestützte Analyse und Optimierung der Wechselwirkung zwischen Belastung und Leistung. Leistungssport, 31(2), 54-62.

Perl, J. & Weber, K. (2004). A neural network approach to pattern learning in sport. *International Journal of Computer Science in Sport, 3*(1), 67-70.

Viru, A. & Viru, M. (1993). Der Mechanismus von Training und Adaptation. Leistungssport, 23(5), 5-8.

KEY TERMS

Adaptation: If the situation of a physiologic system is changed to a nonoptimal or unstable one – for example by external stimuli – the system tries to balancing out the disturbance by changing its state. This process is called adaptation.

Antagonism: Normally, in particular in technical systems, a target position or situation is reached by an asymptotic approach "from one side", where one-directional impulses with decreasing intensity control the adaptation process. In most physiologic (and also in some technical) systems the control is completed by a countercontrol, where the same input activates two contradicting activities, which are controlling each other.

Delay: Adaptation processes can be thought of as time-consuming (stepwise or continuous) changes of the system's state. This means that the reaction on a state disturbance does not takes place at once but shows a certain delay.

Strain Potential: If a load input stream is fed into an organism, it normally cannot be handled at once but is stored

in buffers (e.g.,, organs) for being processed after certain delays. The abstract model of the load buffer here is called strain potential.

Response Potential: The antagonistic counter-part of the strain potential here is called the response potential.

Performance Potential: The result of load input under antagonistic control is modeled by the performance potential, which indicates the current performance state of the modeled physiologic system.

PerPot: PerPot is a model of dynamic adaptation, where an input flow feeds an internal strain potential as well as an internal response potential, from which an output potential is fed by specifically delayed flows. Since the strain flow is negative and the response flow is positive, resulting in an oscillating stabilizing adaptation, the model is called antagonistic (see Perl, 2002).

Kohonen Feature Map (KFM): A KFM consists of a (normally: 2-dimensional) matrix of neurons, each of which can contain information. During a training phase, those neurons are fed with information, which then is learned and organized by means of training algorithms. In the productive or test phase the learned information can be used for attaching or classifying test input (see Kohonen, 1995).

DyCoN: A DyCoN is a KFM-type network, where each neuron contains an individual PerPot-based self-control of its learning behaviour. The DyCoN-concept enables for continuous learning and therefore supports continuous training and testing, training in phases and with generated data, online-adaptation during tests and analyses, and flexible adaptation to new information patterns (see Perl, 2002).

Policy Frameworks for Secure Electronic Business

Andreas Mitrakas
Ubizen, Belgium

INTRODUCTION

Terms conveyed by means of policy in electronic business have become a common way to express permissions and limitations in online transactions. Doctrine and standards have contributed to determining policy frameworks and making them mandatory in certain areas such as electronic signatures. A typical example of limitations conveyed through policy in electronic signatures includes certificate policies that Certification Authorities (CAs) typically make available to subscribers and relying parties. Trade partners might also use policies to convey limitations to the way electronic signatures are accepted within specific business frameworks. Examples of transaction constraints might include limitations in roles undertaken to carry out an action in a given context, which can be introduced by means of attribute certificates. Relying parties might also use signature policies to denote the conditions for the validation and verification of electronic signatures they accept. Furthermore, signature policies might contain additional transaction-specific limitations in validating an electronic signature addressed to end users. Large-scale transactions that involve the processing of electronic signatures in a mass scale within diverse applications rely on policies to convey signature-related information and limitations in a transaction. As legally binding statements, policies are used to convey *trust* in electronic business. Extending further the use of policy in transaction environments can enhance security, legal safety, and transparency in a transaction. Additional improvements are required, however, in order to render applicable terms that are conveyed through policy and enforce them unambiguously in a transaction. The remainder of this article discusses common concepts of policies and certain applications thereof.

BACKGROUND

An early example of a transaction framework is open EDI (Electronic Data Interchange) that aims at using openly available structured data formats and is delivered over open networks. While the main goal of open EDI has been to enable short-term or *ad hoc* commercial transactions among organisations (Kalakota & Whinson, 1996), it has also aimed at lowering the entry barriers of establishing structured data links between trading partners by minimising the need for bilateral framework agreements, known as interchange agreements. One specific requirement of open EDI is to set up the operational and contract framework within which a transaction is carried out. Automating the process of negotiating and executing agreements regarding the legal and technical conditions for open EDI can significantly lower the entry barriers, especially for non-recurrent transactions (Mitrakas, 2000).

Building on the model for open EDI, the Business Collaboration Framework is a set of specifications and guides, the centre of which is the UN/CEFACT; it aims at further lowering the entry barriers of electronic commerce based on structured data formats. The need for flexibility and versatility to loosely coupled applications and communication on the Internet has led to the emergence of Web services. A Web service is a collection of protocols and standards that are used to exchange data between applications. While applications can be written in various languages and run on various platforms, they can use Web services to exchange data over the Internet.

In Web services, using open standards ensures interoperability. These standards also include formal descriptions of models of business procedures to specify classes of business transactions that all serve the same goal. A trade procedure stipulates the actions, the parties, the order, and the timing constraints on performing actions (Lee, 1996). In complex business situations, transaction scenarios typically might belong to a different trade partner that each one owns a piece of that scenario. Associating a scenario with a trade partner often requires electronic signatures. When a trade partner signs with an electronic signature, she might validate or approve of the way that individual procedural components might operate within a transaction. The signatory of an electronic document or a transaction procedure depends on the performance of complex and often opaque-to-the-end-user systems.

Trust in the transaction procedures and the provision of services is a requirement that ensures that the signatory eventually adheres to transparent contract terms that cannot be repudiated (Mitrakas, 2003). Policy is seen as a way to formalise a transaction by highlighting those aspects of a

transaction that are essential to the end user (Mitrakas, 2004). The immediate effect of using policies to convey limitations is that the party that relies on a signed transaction adheres to the limitations of that policy. Policy is, therefore, used to convey limitations to a large number of users in a way that makes a transaction enforceable. While these limitations are mostly meaningful at the operational or technical level of the transaction, they often have a binding legal effect and are used to convey contractual terms. Although these terms are not necessarily legal by nature, they are likely to have a binding effect. Sometimes they can be more far reaching by constraining relying parties that validate electronic signatures. Limitations might be mandated by law or merely by agreement, as in the case of limitations of qualified signatures according to European Directive 1999/93/EC on a common framework for electronic signatures (ETSI TS 101 456).

POLICY CONSTRAINTS IN ELECTRONIC BUSINESS

Electronic signatures have been seen as a lynchpin of trust in electronic transactions. The subject matter of current electronic signature regulation addresses the requirements on the legal recognition of electronic signatures used for non-repudiation and authentication (Adams & Lloyd, 1999). Non-repudiation is addressed in both technical standards such as X.509 and legislation. Non-repudiation addresses the requirement for electronic signing in a transaction in such a way that an uncontested link to the declaration of will of the signatory is established. Non-repudiation is the attribute of a communication that protects against a successful dispute of its origin, submission, delivery, or content (Ford & Baum, 2001). From a business perspective non-repudiation can be seen as a service that provides a high level of assurance on information being genuine and non-refutable (Pfleeger, 2000). From a legal perspective non-repudiation, in the meaning of the Directive 1999/93/EC on a common framework on electronic signatures, has been coined by the term, *qualified signature*, which is often used to describe an electronic signature that uses a secure signature creation device and is supported by a qualified certificate. A qualified signature is defined in the annexes of the directive and is granted the same legal effect as hand-written signatures where law requires them in the transactions.

Policies aim at invoking trust in transactions to ensure transparency and a spread of risk among the transacting parties. Policies are unilateral declarations of will that complement transaction frameworks based on private law. Policies can be seen as guidelines that relate to the technical organizational and legal aspects of a transaction, and they are rendered enforceable by means of an agreement that binds the transacting parties.

In Public Key Infrastructure (PKI), a CA typically uses policy in the form of a certification practice statement (CPS) to convey legally binding limitations to certificate users, being subscribers and relying parties. A CPS is a statement of the practices that a CA employs in issuing certificates (ABA, 1996). A CPS is a comprehensive treatment of how the CA makes its services available and delimiting the domain of providing electronic signature services to subscribers and relying parties. A certificate policy (CP) is sometimes used with a CPS to address the certification objectives of the CA implementation. While the CPS is typically seen as answering "how" security objectives are met, the CP is the document that sets these objectives (ABA, 2001). A CP and a CPS are used to convey information needed to subscribers and parties relying on electronic signatures, in order to assess the level of trustworthiness of a certificate that supports an electronic signature. By providing detailed information on security and procedures required in managing the life cycle of a certificate, policies become of paramount importance in transactions. Sometimes, a PKI Disclosure Statement (PDS) distils certain important policy aspects and services the purpose of notice and conspicuousness of communicating applicable terms (ABA, 2001). The Internet Engineering Task Force (IETF) has specified a model framework for certificate policies (RFC 3647).

Assessing the validity of electronic signatures is yet another requirement of the end user, most importantly, the relying parties. A signature policy describes the scope and usage of such electronic signature with a view to address the operational conditions of a given transaction context (ETSI TR 102 041). A signature policy is a set of rules under which an electronic signature can be created and determined to be valid (ETSI TS 101 733). A signature policy determines the validation conditions of an electronic signature within a given context. A context may include a business transaction, a legal regime, a role assumed by the signing party, and so forth. In a broader perspective, a signature policy can be seen as a means to invoke trust and convey information in electronic commerce by defining appropriately indicated trust conditions.

In signature policies it is also desirable to include additional elements of information associated with certain aspects of general terms and conditions to relate with the scope of the performed action as it applies in the transaction at hand (Mitrakas, 2004). A signature policy might, therefore, include content that relates it to the general conditions prevailing in a transaction, the discreet elements of a transaction procedure as provided by the various parties involved in building a transaction, as well as the prevailing certificate policy (ETSI TS 102 041).

Trade parties might use transaction constraints to designate roles or other attributes undertaken to carry out an action within a transaction framework. Attribute certificates are used to convey such role constraints and are used to indicate a

role, a function, or a transaction type constraint. Attribute policies are used to convey limitations associated with the use and life cycle of such attributes (ETSI TS 101 058).

Processing signed electronic invoices is an application area of using policies. By means of a signature policy, the recipient of an invoice might mandate a specific signature format and associated validation rules. The sender of the invoice might require that signing an invoice might only be carried out under a certain role; therefore, an attribute certificate issued under a specific attribute policy might be mandated. This attribute policy complements the certification practice statement that the issuer of electronic certificates makes available. It is expected that certificate policies shall influence the requirements to make a signature policy binding (Mitrakas, 2003).

BINDING POLICIES IN ELECTRONIC BUSINESS

Communicating and rendering policies binding has been an issue of significant importance in electronic transactions. Inherent limitations in the space available for digital certificates dictate that policies are often conveyed and used in a transaction by incorporating them by reference (Wu, 1998). Incorporation by reference is to make one message part of another message by identifying the message to be incorporated, providing information that enables the receiving party to access and obtain the incorporated message in its entirety, expressing the intention that it be part of the other message (ABA, 1996). The incorporation of policies for electronic signatures into the agreement between signatory and recipient can take place by referencing the intent to use such policy in transactions. When the recipient accepts the signed document of the signatory, he implicitly agrees on the conditions of the underlying signature policy. In practice, incorporating policy into the agreement between signatory and recipient can also be effected by:

- Referring to a policy in a parties' agreement that explicitly refers to such policy.
- Accepting a signed document and implicitly agreeing on the conditions of the underlying policy, although this option might be more restrictive in case of a dispute.

An issue arises with regard to how and under which conditions a particular policy framework can be incorporated into an agreement of a signatory in a way that binds a relying party, regardless of its capacity to act as consumer or business partner. Incorporation of contract terms into consumer contracts and incorporation of contract terms into business contracts follow different rules. Incorporation by reference in a business contract is comparatively straightforward, whereas in a consumer contract stricter rules have to be followed as mandated by consumer protection regulations. Limitations to the enforceability of legal terms that are conveyed by means of policy are applied as a result of consumer protection legislation. In Europe, consumer protection legislation includes the Council Directive 93/13/EC on unfair terms in consumer contracts, Directive 97/7/EC on the protection consumers in distance transactions, and Directive 1999/44/EEC on certain aspects of the sale of consumer goods and associated guarantees (Hoernle, Sutter & Walden, 2002). In an effort to proactively implement these legal requirements, service providers strive to set up specific consumer protection frameworks (GlobalSign, 2004).

Sometimes the scope of the underlying certificate policy frameworks is to equip the transacting parties with the ability to use a certificate as evidence in a court of law. It is necessary to also provide transacting parties with assurance that allows a certificate to be admitted in legal proceedings and that it provides binding evidence against the parties involved in it, including the CA, the subscriber, and relying parties (Reed, 2000). Qualified electronic signatures in the meaning of Directive 1999/93/EC establish a rebuttable presumption that reverses the burden of proof. In other words the court may at first admit a signature that claims to be qualified as an equivalent of a handwritten signature. The counter-party is allowed to prove that such signature does not meet the requirements for qualified signatures, and could therefore be insecure for signing documents requiring a handwritten signature (UNCITRAL, 2000). To further answer the question of admissibility, it is necessary to examine the admissibility of electronic data as evidence in court, which is a matter that has been addressed in Directive 2000/31/EC on electronic commerce. Consequently, electronic data can be admitted as evidence as long as certain warranties are provided with regard to the production and retention of such data. In assessing the reliability of a certificate, a Court will have to examine the possibility of a certificate being the product of erroneous or fraudulent issuance, and if is not, the Court should proclaim it as sufficient evidence against the parties involved within the boundaries of conveyed and binding policy.

FUTURE TRENDS

While case law is expected to determine and enhance the conditions of admissibility and evidential value of policy in transactions based on electronic signatures, additional technological features such as the use of object identifiers (OIDs) and hashing are expected to further enhance the certainty required to accept policies. Remarkably, to date there has been little done to address in a common manner the practical aspects of identifying individual policies and distinguishing among the versions thereof. Additionally, mapping and reconciling policy frameworks in overlapping

transactions also threaten transactions, which are based on the use and acceptance of varying terms. A typical hard case might involve for example overlapping policy conditions, which apply to certificates issued by different CAs. The situation is exacerbated if those CAs do not have the means to recognise one another, while they issue certificates that can be used in the same transaction frameworks (ETSI TS 102 231). Although such certificates may well be complementary to a transaction framework, the varying assurance levels they provide might threaten the reliability of the transaction framework as a whole. The immediate risk for the transacting parties can be an unwarranted transaction environment that threatens to render otherwise legitimate practices unenforceable. Reconciling the methods used across various electronic signing environments is likely to contribute to creating trust in electronic business.

An additional area of future attention may address policy frameworks related to the application layer in a transaction. As present-day requirements for transparency are likely to be further raised, it is expected that online applications will increasingly become more demanding in explaining to the end user what they do and actually warranting the performance. To date general conditions and subscriber agreements cover part of this requirement; however, it is further needed to provide a comprehensive description of the technical features and functionality of the online application. In electronic business, consumers and trade partners are likely to benefit from it. Policies for the application layer are likely to become more in demand in electronic government applications, where the requirement for transparency in the transaction is even higher than in electronic business. Finally, specifying policies further to meet the needs of particular groups of organisations is an additional expectation. Again in electronic government it is expected that interoperability will be enhanced through best practices and standards regarding policy in specific vertical areas.

CONCLUSION

While policies emerge as a necessary piece in the puzzle of invoking trust and legal safety in electronic transactions, policy frameworks can still have repercussions that reach well beyond the scope of single transaction elements and procedures in isolated electronic business environments. Formal policy frameworks require additional attention to ensure that apparently loosely linked policy elements do not threaten to overturn the requirements of transaction security and legal safety, which are the original objectives of using policy frameworks. Electronic transaction frameworks for diverse application areas can benefit from the processing of data on the basis of policy-invoked constraints among the parties involved. Large-scale processing that requires policy to convey operational and legal conditions in electronic transactions benefits from a combination of policy instruments, including certificate polices, signature policies, attribute certificate policies, and so forth, to enhance the outlining of the transaction framework and allow the transacting parties to further rely on electronic business for carrying out binding transactions.

NOTE

The views expressed in this article are solely the views of the author.

REFERENCES

American Bar Association. (1996). *Digital signature guidelines.* Washington, DC.

American Bar Association. (2001). *PKI assessment guidelines.* Washington, DC.

Adams, C. & Lloyd, S. (1999). *Understanding public key infrastructure.* Macmillan Technical Publishing, London.

ETSI TS 101 733. (2000). *Electronic signature formats.* Sophia-Antipolis.

ETSI TS 101 456. (2001). *Policy requirements for CAs issuing qualified certificates.* Sophia-Antipolis

ETSI 102 041. (2002). *Signature policy report, ETSI.* Sophia-Antipolis.

ETSI TS 101 058. (2003). *Policy requirements for attribute authorities.* Sophia-Antipolis.

ETSI TS 102 231. (2003). Provision of harmonized trust service provider status information. Sophia-Antipolis.

Ford, W. & Baum, M.(2001). *Secure electronic commerce* (2nd edition). Englewood Cliffs, NJ: Prentice-Hall.

GlobalSign. (2004). *Certification practice statement.* Retrieved from.*http://www.globalsign.net/repository*

Hoernle, J., Sutter, G. & Walden, I. (2002). Directive 97/7/EC on the protection of consumers in respect of distance contracts. In A. Lodder & H.W.K. Kaspersen (Eds.), *eDirectives: Guide to European Union Law on e-commerce.* Kluwer Law International, The Hague.

IETF RFC 3647. (2003). *Internet X.509 public key infrastructure—certificate policies and certification practices framework.* Retrieved from http://www.faqs.org/rfcs/rfc3647.html

ITU-T Recommendation X.509, ISO/IEC 9594-8. *Infor-*

mation technology—open systems interconnection—the directory: Public-key and attribute certificate frameworks. Draft revised recommendation. Retrieved from *http://www.iso.ch/iso/en/Catalogue DetailPage.CatalogueDetail ?CSNUMBER =34551 &ICS1=35*

Kalakota, R. & Whinson A. (1996). *Frontiers of electronic commerce*. Boston: Addison-Wesley.

Lee, R. (1996). *InterProcs: Modelling environment for automated trade procedures.* User documentation, EURIDIS, WP 96.10.11, Erasmus University, Rotterdam.

Mitrakas, A. (2003). Policy constraints and role attributes in electronic invoices. *Information Security Bulletin, 8*(5).

Mitrakas, A. (2004). Policy-driven signing frameworks in open electronic transactions. In G. Doukidis, N. Mylonopoulos & N. Pouloudi (Eds.), *Information society or information economy? A combined perspective on the digital era*. Hershey, PA: Idea Group Publishing.

Mitrakas, A. (2000). Electronic contracting for open EDI. In S.M. Rahman & M. Raisinghani (Eds.), *Electronic commerce: Opportunities and challenges.* Hershey, PA: Idea Group Publishing.

Pfleeger, C. (2000). *Security in computing.* Englewood Cliffs, NJ: Prentice-Hall.

Reed, C. (2000). *Internet law: Text and materials.* Butterworths.

United Nations. (2000). *Guide to enactment of the UNCITRAL uniform rules on electronic signatures.* New York.

Wu, S. (1998). Incorporation by reference and public key infrastructure: Moving the law beyond the paper-based world. *Jurimetrics, 38*(3).

KEY TERMS

Certification Authority: An authority such as GlobalSign that issues, suspends, or revokes a digital certificate.

Certification Practice Statement: A statement of the practices of a certificate authority and the conditions of issuance, suspension, revocation, and so forth of a certificate.

Electronic Data Interchange (EDI): The interchange of data message structured under a certain format between business applications.

Incorporation by Reference: To make one document a part of another by identifying the document to be incorporated, with information that allows the recipient to access and obtain the incorporated message in its entirety, and by expressing the intention that it be part of the incorporating message. Such an incorporated message shall have the same effect as if it had been fully stated in the message.

Public Key Infrastructure (PKI): The architecture, organization, techniques, practices, and procedures that collectively support the implementation and operation of a certificate-based public key cryptographic system.

Relying Party: A recipient who acts by relying on a certificate and an electronic signature.

Signature Policy: A set of rules for the creation and validation of an electronic signature, under which the signature can be determined to be valid.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2288-2292, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Policy Options for E-Education in Nigeria

Wole Michael Olatokun
University of Ibadan, Nigeria

INTRODUCTION

Information and communication technology (ICT) has turned the world into a global village, and its impact is being felt in all spheres of life. Though it has been rightly said that what is wrong with education cannot be fixed with technology; there is no doubt that modern life is dominated by technology. In today's globalized world, there is a universal recognition of the need to use ICT in education because the free flow of information via satellite and the Internet hold sway in global information dissemination of knowledge. The application of ICT to education brought about the concept of **e-education**. This chapter considers the concept of e-education *vis a vis* the provisions of the national policy for information technology, and gives a state of the art with regard to some e-education initiatives that have been embarked upon by the government, nongovernmental organizations, and other stakeholders in the country. It also identified the challenges constraining the effective deployment and exploitation of ICT for teaching and learning in the Nigerian education system, and recommends some policy options for the development of e-education in the country. The next section gives a background to the concept of e-education.

BACKGROUND

E-education and e-learning are terms sometimes used interchangeably to describe learning through electronic devices or media. That is, a technologically based enhanced learning mechanism that is packaged and targeted towards a broad based population. In other words, the scope of e-education is enhanced on efforts at reaching a widely dispersed population. According to Mac-Ikemenjima (2005), e-education is "an electronic mode of knowledge sharing and transmission, which may not necessarily involve physical contact between teachers and students" (p. 5). Thus, e-education is an alternative to conventional classroom educational system of face-to-face, dynamic, ongoing interaction between teachers and learners, and it is both a computer-aided teaching and computer-aided learning, which ultimately lead to computer-aided instruction. Nenad, Tibor, and Sabina (2005) have noted that e-learning is characterized by the following terms (which give the "e" in e-learning):

- Electronic learning – the main medium of the learning is a computer, with all the advantages of the Internet, intranet, database systems, and applications that make the system easy-to-use and easy-to-manage,
- Everywhere learning – the student is not bound to one place; the computer can be used wherever there is a computer and Internet access,
- Enterprise learning – education is of the utmost importance for the academic community, this way we can offer the materials even to the graduated students to help them improve their knowledge even after graduation,
- Experience learning – the system treats the student as a solver of the problem; it simulates real life situations, thus making the education interactive and exciting; it enables the student to test the knowledge and, if needed, to get help from the mentor (Nenad et al., 2005).

In essence, e-education is a learning system that rises above the confines of space and time, as it does not require physical contact between teacher and student. In this case, learning could be done without classrooms, since instructions and learning materials are accessed through the Internet, CD-ROM, specialized software, and other media. Information and communication technology (ICT) provides the platform on which e-education runs. The use of ICT in e-education makes schools more efficient and creative in knowledge transmission. It produces a variety of tools to enhance teachers' professional skills. However, as Hicks, Reid, and George (1999) have noted, technology by itself, or in itself, is inadequate to provide quality learning. Accordingly, quality is perceived as the function of the way technology is deployed to provide access to germane learning opportunities at the appropriate time. According to them, the main features of online education include computer-mediation, potential for accessing large amounts of dynamic information through WWW, use of hypertext and working with materials in nonlinear way, and access to real-world contexts via Internet. Others include capacity to communicate via e-mail and other electronic technologies with lecturers and other students, new methods for administration of learning, for example, submitting assignments, getting results, networking, and internationalizing the curriculum (Hicks et al., 1999). Also, NetTOM (2007) submitted that cognitive gains from e-learning include hypertext learning, which is nonlinear and

can be structured to engage learners into making greater use of critical thinking skills.

MAIN THRUST OF THE ARTICLE

E-Education and Nigeria's National Policy for Information Technology

In 2001, the Federal Government of Nigeria approved a **National Policy for Information Technology (IT)**. Its implementation started in the same year. The vision statement of the policy is "to make Nigeria an IT capable country in Africa and a key player in the information society by the year 2005, using IT as the engine for sustainable development and global competitiveness." Its mission statement is "to use IT for education, creation of wealth, poverty eradication, job creation and global competitiveness" (Nigerian National Policy for Information Technology, 2001). Some questions worth asking are: What are the provisions of the policy vis-à-vis the implementation of e-education in Nigeria? How adequate is the policy for the integration of ICT in the Nigerian education system? The policy, judging from its mission, general objectives and strategies recognized the importance of ICT in education. Some deficiencies observed however are as follows:

a. The document lacks any sectoral application to education. While sectoral application exist for health, governance, agriculture, legislation, and others, issues on education are grouped under sectoral application for human resources development.
b. The impact of ICT on education is limited to its economic competitiveness. In the sectoral application for human resource development, its objective is basically for students to learn about computers and prepare them to acquire knowledge and skills that would position them for future competitive jobs. It ignores the potential of ICT as a means of solving issues in teaching and learning. No focus is given to the integration of ICT for the development and management of teaching and learning in Nigerian schools.
c. Students cannot acquire knowledge in a vacuum; they must be taught by teachers. However, the policy does not address the issue of teachers' ICT education. Many teachers are incompetent in using ICT to impart knowledge because they also lack such education.
d. Fourth, the national IT policy does not acknowledge the need to develop nationally relevant context software for the education system. The available software in the country are foreign with no local content.

The reality is that the national policy is not focused on the basic issues involved in quality ICT application in education (Olatokun, 2006). It is limited to the market-driven goal of the application of ICT in education, and this can be seen from its emphasis on learning about ICT and not learning through ICT. However, the simple fact is that application of e-education can only be successful when not just learning, but learning and teaching through ICTs are encouraged. In the next section, we describe some e-education initiatives in Nigeria.

E-Education Initiatives in Nigeria

1. Government Initiatives

National Open University of Nigeria

Mac-Ikemenjima (2005) reported that the National Open University (NOU) was initially established on 22nd July 1983 but became functional in April 2001. The aim of the institution is to train professionals in various disciplines through the distance learning mode. The institution was set up on the premise that every year almost 1.5 million students apply to the various universities in the country but only about 20% can be absorbed by the university system. NOU is expected to take care of the remaining 80%. The course delivery is through a combination of Web-based modules, textual materials, audio and video tapes, as well as CD ROMs. The university currently has 18 study centers and plans to have at least one study center in each of the 774 local governments of Nigeria. It runs programmes in education, arts and humanities, business and human resource management, and science and technology.

The NOU is designed to increase the access of all Nigerians to formal and nonformal education in a manner convenient to their circumstances, and cater to the continuous educational development of professionals. The range of target clientele is elastic and is to be continually reviewed to meet Nigeria's ever-changing needs (National Open University of Nigeria, 2008). It employs a range of delivery methods to take education to the people and make learning an enjoyable activity, including printed instructional materials, audio, video tapes, and CD-ROMs; television and radio broadcast of educational programmes and electronic transmission of materials in multimedia (voice, data, graphics, video) over fixed line (telephone or leased lines), terrestrial, and VSAT wireless communication systems. For the take off of the university, pioneer student enrolment was 32,400 and it is believed that with time, more people will benefit from the programmes of the NOU (National Open University of Nigeria, 2008).

Nigerian Universities Network (NUNet) Project

The **NUNet** project was conceived in 1994 by the **National Universities Commission** (NUC) when it was realized that Nigerian academic staff and students should not be isolated from each other and from the global academic community (Mac-Ikemenjima, 2005). The decision was made that all Universities should be linked at least through e-mail through Trieste via a dial-up system. Apart from e-mail, the project is also to solve the problems of resource sharing among Nigerian Universities and their counterparts all over the world, provide access to electronic journals and books, many of which are only available in such forms, and serve as vehicle to expand access to education at minimal cost. Thus, in 1996, with the help of the International Centre for Theoretical Physics (ICTP), NUNet established the e-mail gateway with the purpose of providing an dedicated international direct dial (IDD) line for NUC, together with a mail exchange and server. While this dial-up technology has been surpassed by other ICT developments, NUNet still keeps it operational (and indeed some universities still use it as their sole electronic communications means) as a back-up system. By 2000, NUNet had established its own VSAT set-up to real-time Internet access (at least within the NUC building). In an attempt to transform itself, NUNet is concentrating on IT training policies for universities (UNESCO, 2002). In order to meet these needs, several partnerships were built with various institutions, and NUC staffs were trained in various IT-related skills. The NUNet system has been fluctuating in its operations but it is hoped that within a few years, its performance would have improved. It is expected that partnership among Nigerian institutions would improve and drive down the cost of Internet connections through economies of scale.

2. Virtual Library Initiatives

The objectives of the national virtual library include improving the quality of teaching and research institutions through the provision of current books, journals, and other library resources; enhancing access of academic libraries to global library and information resources; enhancing scholarship and lifelong learning through the establishment of permanent access to shared digital archival collections; provision of guidance to academic libraries on ways of applying appropriate technologies for production of digital library resources; and to advance the use and usability of globally distributed networks library resources. Virtual library initiatives in Nigeria include:

- **The National Virtual (Digital) Library Project** of the Ministry of Education, which is supervised by the National Universities Commission (National Universities Commission, 2007).
- The National Virtual Library Project of the Ministry of Science and Technology is supervised by the National Information Technology Development Agency (NITDA)
- An ongoing effort by UNESCO to develop a virtual Library for all Nigerian Higher Education Institutions in Nigeria (National Universities Commission, 2007).

3. Civil Society (NGO led) E-education Initiatives

Several NGOs have had a notable impact on the development of **telecenters**. Telecenters are places that usually have rudimentary ICT facilities designed to acquaint people with these technologies, and to begin educational programs to teach people in marginalized regions about how to use them (Mac-Ikemenjima, 2005). Examples include community teaching and learning centers, Lagos Digital Village by Junior Achievement Nigeria, Owerri Digital Village by Youth for Technology Foundation, Computer Literacy for Older Persons Programme by Mercy Mission, and so forth. The impact of these telecentres is felt in the immediate rural location, where they are sited in terms of increasing access to ICT facilities. However, due to the fact that they are very few, they are grossly inadequate in providing access to ICT facilities to the country's large rural populations. Other challenges include low level of general and ICT literacy, poor energy supply, and lack of local content.

Challenges of E-Education in Nigeria

Nigeria, like most developing nations, is yet to fully integrate ICT into its educational system and is thus incapacitated to benefit from e-education. Although the benefits of the implementation of e-education in any society are numerous, yet there are also many attendant challenges. The problem most often noted is the availability of adequate ICT infrastructure and poor maintenance culture (Mac-Ikemenjima, 2005; Ololube, 2005a; Yusuf, 2005). An efficient e-educational system requires adequate functional specialized software, unlimited access to the Internet, and the stability of other infrastructure, such as electricity and telecommunications (Aduwa-Ogiegbaen & Iyamu, 2005). The quality of teachers is known in virtually all countries to be a key predictor of student learning (Ololube, 2005a; Ololube, 2005b). Teachers need training not only in computer literacy, but also in the application of various kinds of educational software in teaching and learning (Ololube, 2006). It is pertinent to note that in some societies like Nigeria, many teachers lack access to basic ICT infrastructure, like computer hardware. Worse still, many hardly come in contact with ICT-aided instructional materials. Most of the identified challenges of

e-education are those associated with the use of ICT infrastructure. These hindrances are discussednext.

Inadequate ICT Infrastructure

The implementation of e-education would only have the desirable result if people have direct access to ICT infrastructure, a situation which is not obtainable in Nigeria at present. Most times, users of ICT infrastructure depend on the limited facilities available at their workplaces, while a large number patronize commercial providers of such facilities; these are either unreliable or inefficient. The situation is worse for the rural dwellers. It is pertinent to say that most rural dwellers lack basic ICT infrastructure such as radio, television, and telephone services. It is all the more so with more complex ICT infrastructure, such as computer hardware, software, and other accessories required to facilitate their access to e-education.

Inadequate Manpower Skills

Nigeria, like its other African counterparts, is basically a consumer of ICT. Not only does it lack adequate ICT infrastructure, but also has a dearth of experienced skills needed to manage such infrastructure (Aduwa-Ogiegbaen, & Iyamu, 2005). There is severe scarcity of qualified computer software designers and other trained personnel in the management of operating systems. Networking and local workers to service and repair computer facilities are also grossly inadequate. Worthy of note is the fact that educators, who are expected to be the main force for the implementation of e-education, similarly lack the skills to fully utilize ICT in curriculum implementation.

Poor Energy Supply

While ICT is the platform on which e-education is implemented, electricity is the backbone for all ICT infrastructure. It follows then that there cannot be any efficient e-educational system without a functional electricity facility. However, it can be observed that no area in Nigeria is confident of regular electricity supply except areas inhabited by high level government officials. The government has not succeeded in providing stable and reliable electricity supply to every part of the country (Olatokun, 2006). In addition is the fact that, in Nigeria, most rural inhabitants do not have access to electricity, hence, implementation of e-education in such areas becomes a hurdle.

Poor Telecommunication Facilities

The Nigerian telecommunication sector is one of the fastest growing among African countries (Aduwa-Ogiegbaen, & Iyamu, 2005). This has especially been the case with the liberalization of Nigerian telecommunication sector in 2001, and the subsequent licensing of the **Global System of Mobile Communication (GSM)** and other fixed wireless operators. Prior to 2001, there were only some 700,000 lines in the country, with only 450 connected. However, since 2001, more than 3 million landlines and 5 million GSM subscribers have been added to the existing telephone capacity. Despite this increment in the number of land and GSM lines, they are still inadequate for Nigeria's over 140 million population (National Population Commission, 2006). Besides, most users of the GSM and landlines are in the urban areas. In addition, there is the high cost of using these facilities. Although this has significantly reduced over the years, it is still beyond the comfortable reach of many Nigerians. However, the recent acquisition of the nation's first national carrier, Nigeria Telecommunication, by Transnational Corporation has raised the hope of Nigerians that there would be further increase in competition among telecom operators, and ultimately, reduction in prices (Aduwa-Ogiegbaen, & Iyamu, 2005).

Underfunding of the Educational System

In Nigeria, the overall educational system is dependent on the government. However, it is apparent that the government lacks what it takes to adequately provide for the entire education system. As a result of underfunding, most institutions are not oriented towards the provision of e-education. Rather, such limited resources are used in meeting more urgent and survival needs of the institutions. Moreover, overdependence of educational institutions on government for virtually everything has beclouded their ability to partner with the private sector or other alternative funding sources for ICT education initiatives, except in a few private institutions.

Limited Internet Access

A major challenge to e-education in Nigeria is how to establish reliable cost-effective Internet connectivity that ensures easy access to information at all times (Aduwa-Ogiegbaen, & Iyamu, 2005). The limitation stems from the fact that in Nigeria, there are no indigenous ISPs. The available ones are usually Nigerian companies in partnership with foreign IT companies. Most times, the relationship does not translate into efficient services, as one would expect, but poor services and exploitation of customers. The few reputable ISPs that have invested so much usually charge high fees due to very small customer base occasioned by high fees. A contributory factor to the high Internet connectivity fees is the high cost of telecommunication services (Aduwa-Ogiegbaen, & Iyamu, 2005). The monopoly in the Nigerian telecommunication sector for many years negatively affected the penetration of Internet services in the country.

FUTURE TRENDS

Various challenges to e-education have been identified. Deficiencies in the national policy on information technology in relation to implementation of e-education have also been described. What then does the future portend for Nigeria?

The E-Education Initiatives

To answer these questions, it is very important to place the future of e-education in Nigeria within the context of the overall objectives of various e-education initiatives. The stated objectives are a step in the right direction. Aside from this, the spread of **National Open University** in over 18 study centers, and the projection for the establishment of a study center in each of the 774 local governments in Nigeria, is a welcome development. Also, the recent and ongoing proposal of 100-dollar laptops, equivalent of thirteen thousand five hundred naira (117 naira is equivalent to 1 dollar), is an attempt to increase wider access to media technology.

Increasing the Level of Awareness of Nigerians about ICT

It is important also to observe the increase in level of awareness of Nigerians (public, private) of various IT products in relation to available opportunities. Awareness often leads to curiosity, familiarity, and eventually interactions, which in turn leads to skill acquisition and development. The contribution of Nigerian software development companies is becoming noticeable. This would produce the much-emphasized local content in the areas of applications and packages, thereby bridging the much-referenced digital divide existing between the developed and less developed or developing nations.

Human Resources Development and Skills Upgrading

The introduction of computer education in the academic curriculum and campaign for mass training and participation in information technology is an effective tool in the development of highly skilled workforce and drastic reduction in the level of illiteracy prevalent in the country both in the traditional and formal education system, but especially e-education (Aduwa-Ogiegbaen, & Iyamu, 2005).

Liberalization of the Telecommunication Sector

Access to reliable telecommunications systems is critical to the implementation of e-education. The liberalization of the telecommunication sector is a positive move towards achieving a stable telecommunication network and expanding access. This has broken the monopoly enjoyed by the Nigerian Telecommunications and stirred up competition among the different operators. Although the cost of connectivity to these facilities is still high, it is expected that in the nearest future, more operators would be forced to reduce their prices, especially with the acquisition of the Nigerian Telecommunications by Transnational Corporation of Nigeria Plc. It would be remarkable to witness considerable reduction in telecommunication tariffs arising from stiff competition among the operators.

Investments in Telecenters

The activities of some nongovernmental organizations (NGO) have been observed in the provision of **telecenters,** which are places that have basic ICT facilities designed to acquaint people with these technologies in order to increase the flow of information to and from the poorest and the marginalized. It has been effective in bridging the gap between those who have access to computer and Internet and those who do not. Telecenters are particularly useful for rural development. As discussed earlier, there are few NGOs that have been providing such telecenters. One such is the Owerri Digital Village, which is based in the Eastern part of the country. Owerri Digital Village was launched in September 2001 by the Youth for Technology Foundation (YTF), an international nonprofit organization based in the United States and Nigeria. It is a community technology and learning center that offers IT skills development and training for Nigerian youth in an effort to develop entrepreneurial spirit and passion for technology. Community teaching and learning center is another initiative by the nongovernmental organizations. It is an initiative of Teachers Without Borders (TWB), an international US-based NGO that seeks to connect teachers globally, to each other and to resources through ICT. The centers currently exist in various parts of Nigeria and are yielding some results in terms of increasing access.

CONCLUSION

E-education is about IT infrastructures and content. The factors of inadequate infrastructures to support the IT of e-education in Nigeria have been established. To harness the opportunities inherent in e-education, conscious effort is required and necessary to create an enabling environment that would promote the participation of all stakeholders in terms of funding, training, and development. Increase campaign in collaboration with NGOs, communities, and relevant agents would assist in the realization of the vision of e-education. In addition, it is very crucial that every citizen has access

to the pursuit of education irrespective of age, level of experience and qualification, and location. It is obvious that Nigeria is on the right track towards the implementation of e-education owing to the numerous initiatives discussed. Specific sectoral applications to education in the objectives of the national policy have been suggested. The Government should focus on the integration of ICT into every aspect of teaching and learning. Since educational institutions are generally underfunded, government should increase funding of these institutions. Specific budget for ICT should be set aside annually for all institutions of learning, and its use should effectively be monitored. Additionally, all institutions should intensify efforts to generate more funds from donor organizations to supplement whatever they receive from the Federal Government. Private sector partnership should be considered in the application of ICT. If these suggestions could be adopted with the right policies, Nigeria will soon begin to reap the benefits and opportunities offered by e-education.

REFERENCES

Aduwa-Ogiegbaen, S. E., & Iyamu, E. O. S. (2005). Using information and communication technology in secondary schools in Nigeria: Problems and prospects. *Educational Technology and Society, 8*(1), 104-112. Retrieved 12th November 2006, from http://www.ifets.info/journals/8_1/13.pdf

Hicks, M., Reid, J., & George, R. (1999). *Enhancing online teaching: Designing responsive learning environments*: Conference paper, 1999 HERDSA. Annual International Conference paper, Melbourne 12-15 July.

Mac-Ikemenjima, D. (2005). E-education in Nigeria: Challenges and prospects. In *Harnessing the Potential of ICT for Education – A Multistakeholder Approach: Proceedings of the 8th Conference of United Nations Information and Communications Technology Task Force* held in Dublin, Ireland from 13-15 April, 2005 Retrieved 12th June 2007, from http://www.onevillagefoundation.org/dpi/downloads/e-education.doc

National Open University of Nigeria. (2008). Retrieved 16 March 2008, from http://www.nou.edu.ng/noun/index.htm

National Population Commission (NPC). (2006). *Nigerian population facts and figures.* Retrieved 11th February 2008, from http://www.population.gov.ng/factsandfigures.htm

National Universities Commission. (2007). *National Virtual Library of Nigeria.* Retrieved 12 November 2007, from http://www.nigerianvirtuallibrary.com/:

Nenad, K., Tibor, S., & Sabina, S. (2005). Computer science education: Differences between e-learning and classical approach. Retrieved 16th February 2008, from http://www.claroline.net/dlarea/Zagrebpaper1336.pdf

NetTom. (2007). *Why is e-learning important*? Retrieved 11/2/2008, from http://ebdd.wsu.edu/eder/NetTom_ToT/unit.1/whyelearning.htm

Nigerian National Policy for Information Technology. (2001). Retrieved 21st March 2008, from http://forum.org.ng/system/files/Nigeria_IT_Policy.pdf

Olatokun, W. M. (2006). National information technology policy in Nigeria: Prospects, challenges and a framework for implementation. *African Journal of Library, Archives and Information Science, 16*(1), 9-18.

Ololube, N. P. (2005a). Benchmarking the motivational competencies of academically qualified teachers and professionally qualified teachers in Nigerian secondary schools. *The African Symposium, 5(3), 17-37. Retrieved 12th November 2006, from http://www2.ncsu.edu/ncsu/aern/TASS.3Ololube.pdf*

Ololube, N. P. (2005b). School effectiveness and quality improvement: Quality teaching in Nigerian secondary schools. *The African Symposium,. 5(4), 17-31. Retrieved 16th August 2007, from http://www2.ncsu.edu:8010/ncsu/aern/TASS.4Ololube.pdf.*

Ololube, N. P. (2006). Appraising the relationship between ICT usage and integration and the standard of teacher education programs in a developing economy. *International Journal of Education and Development using ICT, 2*(3). Retrieved 20th November 2006, from http://ijedict.dec.uwi.edu/viewarticle.php?id=194&layout=html.

United Nations Educational Scientific and Cultural Organization (UNESCO). (2002). Mission report with the purpose of launching the feasibility study for the development of a virtual library for universities and institutions of higher learning in Nigeria. Retrieved 27 March 2008, from http://portal.unesco.org/fr/files/13024/10560949311Nigeria_Mission_Report.doc/Nigeria%2BMission%2BReport.doc

Yusuf, M. O. (2005). An investigation into teachers' self-efficacy in implementing computer education in Nigerian secondary chools. *Meridian: A Middle School Computer Technologies Journal, 8*(2). Retrieved 12th November 2006, from http://www.ncsu.edu/meridian///////sum2005/computer_ed_nigerian_schools/index.html

KEY TERMS

E-Education: Use of computers and electronics to assist learning

E-Learning: Learning that is facilitated by the use of digital tools and content. Typically, it involves some form of interactivity that may include online interaction between the learner and their teacher or peers.

Information and Communication Technology (ICT): Includes technologies such as desktop and laptop computers, software, peripherals, and connections to the Internet that are intended to fulfill information processing and communications functions.

Information Technology Policy: Policy guidelines that concern all forms of technology used in processing and disseminating information.

Initiative: An organized and coordinated strategy to address the needs, issues, or desires of a population or community.

Policy: An overarching plan of action developed for the achievement of a set of goals.

Strategy: A plan or method for obtaining a specific result.

Predictive Data Mining: A Survey of Regression Methods

Sotiris Kotsiantis
University of Patras, Greece & University of Peloponnese, Greece

Panayotis Pintelas
University of Patras, Greece & University of Peloponnese, Greece

INTRODUCTION

Data mining is the extraction of implicit, previously unknown, and potentially useful information from data. The idea is to build computer programs that sift through databases automatically, seeking regularities or patterns. Strong patterns, if found, will likely generalize to make accurate predictions on future data. Machine learning (ML) provides the technical basis of data mining. It is used to extract information from the raw data in databases—information that is expressed in a comprehensible form and can be used for a variety of purposes.

Every instance in any data set used by ML algorithms is represented using the same set of features. The features may be continuous, categorical, or binary. If instances are given with known labels (the corresponding correct outputs), then the learning is called supervised in contrast to unsupervised learning, where instances are unlabeled (Kotsiantis & Pintelas, 2004). This work is concerned with regression problems in which the output of instances admits real values instead of discrete values in classification problems.

BACKGROUND

A brief review of what ML includes can be found in Dutton and Conroy (1996). A historical survey of logic and instance-based learning is also presented in De Mantaras and Armengol (1998). The first step of predictive data mining is collecting the data set. If a requisite expert is available, then he or she can suggest which fields (attributes, features) are the most informative. If not, then the simplest method is that of "brute force," which means measuring everything available in the hope that the right (informative, relevant) features can be isolated. However, a data set collected by the brute-force method is not directly suitable for induction. It contains, in most cases, noise and missing feature values, and therefore requires significant preprocessing (Zhang, Zhang, & Yang, 2002). Hodge and Austin (2004) have recently introduced a survey of contemporary techniques for outlier (noise) detection. Depending on the circumstances, researchers have a number of methods to choose from to handle missing data (Batista & Monard, 2003). Feature subset selection is the process of identifying and removing as many irrelevant and redundant features as possible (Yu & Liu, 2004). The fact that many features depend on one another often unduly influences the accuracy of supervised ML models. This problem can be addressed by constructing new features from the basic feature set (Markovitch & Rosenstein, 2002).

The problem of regression consists of obtaining a functional model that relates the value of a target continuous-variable y with the values of variables x_1, x_2...x_n (the predictors). This model is obtained using samples of the unknown regression function. These samples describe different mappings between the predictor and the target variables. The traditional approach for prediction of a continuous target is the classical linear least-squares regression (Fox, 1997). The model constructed for regression in this traditional approach is a linear equation. By estimating the parameters of this equation with a computationally simple process on the training set, a model is created. However, the linearity assumption between input features and predicted value introduces a large bias error for most domains. That is why most studies are directed to nonlinear and nonparametric techniques for the regression problem.

Murthy (1998) provided an overview of work in decision trees and a sample of their usefulness to newcomers as well as practitioners in the field of data mining. Decision trees are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Regression trees are the counterpart of decision trees for regression tasks. A regression tree, or any learned hypothesis h, is said to overfit training data if another hypothesis h´ exists that has a larger error than h when tested on the training data, but a smaller error than h when tested on the entire data set. There are two common approaches that regression-tree induction algorithms can use to avoid overfitting training data: (a) Stop the training algorithm before it reaches a point at which it perfectly fits the training data, and (b) prune the induced regression tree.

Most algorithms use a pruning method. Model trees generalize the concepts of regression trees, which have constant values at their leaves (Torgo, 2000). Thus, they are analogous to piecewise linear functions (and hence nonlinear functions). The major advantage of model trees over regression trees is that model trees are much smaller than regression trees, the decision strength is clear, and regression functions do not normally involve many variables. The most well-known model-tree inducer is the M5´ (Wang & Witten, 1997). Model trees can tackle tasks with very high dimensionality—up to hundreds of attributes; however, computational requirements grow rapidly with dimensionality.

Regression trees can be translated into a set of rules by creating a separate rule for each path from the root to a leaf in the tree (Torgo, 1995). The M5´ rules algorithm produces propositional regression rules using routines for generating a decision list from M5´ model trees (Witten & Frank, 2005). The algorithm is able to deal with both continuous and nominal variables, and obtains a piecewise linear model of the data. However, rules can also be directly induced from training data using a variety of rule-based algorithms. Furnkranz (1999) provided an excellent overview of existing work in rule-based methods. Regression rules represent each result by a disjunctive normal form (DNF). A *k*-DNF expression is of the form $(X_1 \wedge X_2 \wedge \ldots \wedge X_n) \vee (X_{n+1} \wedge X_{n+2} \wedge \ldots X_{2n}) \vee \ldots \vee (X_{(k-1)n+1} \wedge X_{(k-1)n+2} \wedge \ldots \wedge X_{kn})$, where k is the number of disjunctions, n is the number of conjunctions in each disjunction, and X_n is defined over the alphabet $X_1, X_2 \ldots X_j \cup \sim X_1, \sim X_2 \ldots \sim X_j$. The goal is to construct the smallest rule set that is consistent with the training data. A large number of learned rules is usually a sign that the learning algorithm is attempting to remember the training set instead of discovering the assumptions that govern it. The difference between heuristics for rule learning and heuristics for regression trees is that the latter evaluate the average quality of a number of disjointed sets (one for each value of the feature that is tested), while rule learners only evaluate the quality of the set of instances that is covered by the candidate rule (Torgo, 1995).

Artificial neural networks (ANNs) are another method of inductive learning based on computational models of biological neurons and networks of neurons as found in the central nervous system of humans (Witten & Frank, 2005). A multilayer neural network consists of a large number of units (neurons) joined together in a pattern of connections. Units in a net are usually segregated into three classes: input units, which receive information to be processed; output units, where the results of the processing are found; and units in between known as hidden units. Feed-forward ANNs allow signals to travel one way only, from input to output. First, the network is trained on a set of paired data to determine input-output mapping. The weights of the connections between neurons are then fixed and the network is used to predict the value of a new set of data. Generally, properly determining the size of the hidden layer is a problem. An excellent argument regarding this topic can be found in Camargo and Yoneyama (2001). Kon and Plaskota (2000) also studied the minimum amount of neurons and the number of instances necessary to program a given task into feed-forward neural networks. ANN depends upon three fundamental aspects: input and activation functions of the unit, network architecture, and the weight of each input connection. Given that the first two aspects are fixed, the behavior of the ANN is defined by the current values of the weights. The weights of the net to be trained are initially set to random values, and then instances of the training set are repeatedly exposed to the net. The values for the input of an instance are placed on the input units and the output of the net is compared with the desired output for this instance. Then, all the weights in the net are adjusted slightly in the direction that would bring the output values of the net closer to the values for the desired output. There are several algorithms with which a network can be trained (Neocleous & Schizas, 2002). However, the most well-known and widely used learning algorithm to estimate the values of the weights is the back-propagation (BP) algorithm.

Instance-based learning algorithms are lazy learning algorithms as they delay the induction or generalization process until the regression process is performed. *k*-nearest neighbor (k-NN) is based on the principle that the instances within a data set will generally exist in close proximity with other instances that have similar properties (Aha, 1997). The k-NN algorithm first finds the closest instances to the query point in the instance space according to a distance measure, and then outputs the average of the target values of those instances as the prediction for that query instance. Many different metrics for the relative distance between instances have been presented (Witten & Frank, 2005). For more accurate results, several algorithms use weighting schemes that alter the distance measurements and voting influence of each instance. A survey of weighting schemes is given by Wettschereck, Aha, and Mohri (1997). As the prediction of the target value of a query instance requires one to measure its distance to all training instances, which might be a very huge set, the prediction in k-NN is very costly.

An excellent survey of support vector machines (SVMs) can be found in (Burges, 1998). The sequential minimal optimization (SMO) algorithm has been shown to be an effective method for training support vector machines on classification tasks defined on sparse data sets (Platt, 1999). SMO differs from most SVM algorithms in that it does not require a quadratic programming solver. In Shevade, Keerthi, Bhattacharyya, and Murthy (2000) and Flake and Lawrence (2002), SMO is generalized so that it can handle regression problems. This implementation globally replaces all missing values and transforms nominal attributes into binary ones.

MAIN FOCUS OF THE ARTICLE

No single learning algorithm can uniformly outperform other algorithms over all data sets. When faced with the question "Which algorithm will be most accurate on our regression problem?" the simplest approach is to estimate the success rate of the candidate algorithms on the problem and select the one that appears to be most accurate. The success of regression can be judged by trying out the concept description that is learned on an independent set of test data for which the true values are known but not made available to the machine. The success rate on test data gives an objective measure of how well the concept has been learned. For the regression methods, there is more than one regressor criterion. Table 1

Table 1. Regressor criteria (p is the predicted value, and a is the actual value, $\bar{a} = \frac{1}{n}\sum_i a_i$)

Mean absolute error	$\frac{\lvert p_1 - a_1\rvert + \ldots + \lvert p_n - a_n\rvert}{n}$
Root mean squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_1 - a_1)^2}{n}}$
Relative absolute error	$\frac{\lvert p_1 - a_1\rvert + \ldots + \lvert p_n - a_n\rvert}{\lvert a_1 - \bar{a}\rvert + \ldots + \lvert a_n - \bar{a}\rvert}$
Root relative squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_1 - a_1)^2}{(a_1 - \bar{a})^2 + \ldots + (a_1 - \bar{a})^2}}$

*Table 2. Comparing learning algorithms (Note: **** the best performance, * the worst performance)*

	Regression Trees	Neural Networks	k-NN	SVM	Rule Learners
Accuracy in general	**	***	**	****	**
Speed of learning with respect to number of attributes and the number of instances	***	*	****	*	**
Speed of prediction	****	****	*	****	****
Tolerance to missing values	***	*	*	**	**
Tolerance to irrelevant attributes	***	*	**	****	**
Tolerance to redundant attributes	**	**	**	***	**
Tolerance to noise	**	**	*	**	*
Dealing with danger of overfitting	**	*	***	**	**
Explanation ability/transparency of knowledge/predictions	****	*	**	*	****
Model parameter handling	***	*	***	*	***

represents the most well known. Fortunately, it turns out that for in most practical situations, the best regression method is still the best no matter which error measure is used.

Generally, logic-based algorithms (regression trees and rule learners) are all considered very easy to interpret, whereas neural networks and SVMs have notoriously poor interpretability. k-NN is also considered to have very poor interpretability because an unstructured collection of training instances is far from readable, especially if there are many of them. Moreover, k-NN is generally considered intolerant of noise; its similarity measures can be easily distorted by errors in attribute values, thus leading it to misclassify a new instance on the basis of the wrong nearest neighbors. Contrary to k-NN, rule learners and regression trees are considered resistant to noise because their pruning strategies avoid overfitting the data in general and noisy data in particular. Features of learning techniques are compared in Table 2.

FUTURE TRENDS

The use of multiple regression models has gained momentum in the recent years, and researchers have continuously argued for the benefits of using multiple regression models to solve complex problems (Hjort & Claeskens, 2003). The main motivation for combining models is based upon the assumption that different models using different data representations, different concepts, and different modeling techniques are likely to arrive at results with different patterns of generalization. As most combination functions benefit from disagreement to errors of individual models, the greater this disagreement, the lower the impact of individual errors on the final decision, and effectively the lower the combined error.

Ensemble learning consists of two problems: ensemble generation, which is how the base models are generated, and ensemble integration, which is how the base models' predictions are integrated to improve performance. Ensemble generation can be characterized as homogeneous if each base learning model uses the same learning algorithm or heterogeneous if the base models can be built from a range of learning algorithms.

Bagging (Breiman, 1996) is a "bootstrap" ensemble method that creates individual regression models by training the same learning algorithm on a random redistribution of the training set. Each regression model's training set is generated by randomly drawing, with replacement, N instances, where N is the size of the original training set. Many of the original instances may be repeated in the resulting training set while others may be left out. After the construction of several regression models, taking the average value of the predictions of each regression model gives the final prediction. A more sophisticated version of bagging is described in Breiman (2001). Another method that uses different subsets of training data with a single learning method is the boosting approach (Duffy & Helmbold, 2002). The boosting approach uses the base models in sequential collaboration, where each new model concentrates more on the examples where the previous models had high error.

The simplest approach for building heterogeneous ensembles is to use a variety of learning algorithms on all of the training data and combine their predictions according to an averaging scheme (Kotsiantis & Pintelas, 2005). Among the combination methods, the averaging rule is the simplest to implement since it requires no prior training (Hjort & Claeskens, 2003). Stacked generalization (Ting & Witten, 1999), or stacking, is another heterogeneous approach. Stacking combines multiple regression models to induce a higher level regression model with improved performance. A learning algorithm is used to determine how the outputs of the base regression models should be combined. The original data set constitutes the Level 0 data. All the base regression models run at this level. The Level 1 data are the outputs of the base regression models. Another learning process occurs using as input the Level 1 data and as output the final prediction. Multiresponse linear regression (MLR) was used for metalevel learning (Ting & Witten).

Finally, there are some open problems in ensembles of learners, such as how to understand and interpret the decision made by an ensemble of learners because an ensemble provides little insight into how it makes its decision. For learning tasks such as data mining applications where comprehensibility is crucial, averaging methods normally result in incomprehensible learners that cannot be easily understood by end users.

CONCLUSION

The key question when dealing with ML regression problems is not whether a learning algorithm is superior to others, but under which conditions a particular method can significantly outperform others on a given application problem. One of the most active areas of research in supervised learning has been to study methods for constructing good ensembles of learners. Even if the learning algorithm can in principle find the best hypothesis, we actually may not be able to find it. Building an ensemble may achieve a better approximation, even if no assurance of this is given.

REFERENCES

Aha, D. (1997). *Lazy learning.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Batista, G., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence, 17*, 519-533.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(3), 123-140.

Breiman, L. (2001). Using iterated bagging to Debias regressions. *Machine Learning, 45*(3), 261-277.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery, 2*(2), 1-47.

Camargo, L. S., & Yoneyama, T. (2001). Specification of training sets and the number of hidden neurons for multilayer perceptions. *Neural Computation, 13*, 2673-2680.

De Mantaras & Armengol, E. (1998). Machine learning from examples: Inductive and lazy methods. *Data & Knowledge Engineering, 25*, 99-123.

Duffy, N., & Helmbold, D. (2002). Boosting methods for regression. *Machine Learning, 47*(2-3), 153-200.

Dutton, D., & Conroy, G. (1996). A review of machine learning. *Knowledge Engineering Review, 12*, 341-367.

Flake, G. W., & Lawrence, S. (2002). Efficient SVM regression training with SMO. *Machine Learning, 46*(1-3), 271-290.

Fox, J. (1997). *Applied regression analysis, linear models, and related methods.* Sage Publications.

Furnkranz, J. (1999). Separate-and-conquer rule learning. *Artificial Intelligence Review, 13*, 3-54.

Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association, 98*(464), 879-899.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85-126.

Kon, M., & Plaskota, L. (2000). Information complexity of neural networks. *Neural Networks, 13*, 365-375.

Kotsiantis, S., & Pintelas, P. (2004). Recent advances in clustering: A brief survey. *WSEAS Transactions on Information Science and Applications, 1*(1), 73-81.

Kotsiantis, S., & Pintelas, P. (2005). Selective averaging of regression models. *Annals of Mathematics, Computing & Teleinformatics, 1*(3), 66-75.

Markovitch, S., & Rosenstein, D. (2002). Feature generation using general construction functions. *Machine Learning, 49*, 59-98.

Murthy. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery, 2*, 345-389.

Neocleous, C., & Schizas, C. (2002). Artificial neural network learning: A comparative review. In *Lecture notes in artificial intelligence* (Vol. 2308, pp. 300-313). Berlin, Germany: Springer-Verlag.

Platt, J. (1999). Using sparseness and analytic QP to speed training of support vector machines. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems* (p. 11). MA: MIT Press.

Shevade, S., Keerthi, S., Bhattacharyya, C., & Murthy, K. (2000). Improvements to the SMO algorithm for SVM regression. *IEEE Transactions on Neural Networks, 11*(5), 1183-1188.

Ting, K., & Witten, I. (1999). Issues in stacked generalization. *Artificial Intelligence Research, 10*, 271-289.

Torgo, L. (1995). Data fitting with rule-based regression. In J. Zizka & P. Brazdil (Eds.), *Proceedings of the Workshop on Artificial Intelligence Techniques (AIT'95)*, Brno, Czech Republic.

Torgo, L. (2000). Inductive learning of tree-based regression models. *AI Communications, 13*(2), 137-138.

Wang, Y., & Witten, I. H. (1997). Induction of model trees for predicting continuous classes. In *Proceedings of the Poster Papers of the European Conference on ML*, Prague, Czech Republic (pp. 128-137). Prague, Czech Republic: University of Economics, Faculty of Informatics and Statistics.

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review, 10*, 1-37.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). San Francisco: Morgan Kaufmann.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *JMLR, 5*, 1205-1224.

Zhang, S., Zhang, C., & Yang, Q. (2002). Data preparation for data mining. *Applied Artificial Intelligence, 17*, 375-381.

KEY TERMS

Artificial Neural Networks: They are nonlinear predictive models that learn through training and resemble biological neural networks in structure.

Data Cleansing: This is the process of ensuring that all values in a data set are consistent and correctly recorded.

Nearest Neighbor: It is a technique that predicts the value of each record in a data set based on a combination of the values of the *k* record(s) most similar to it.

Predictive Model: A predictive model is a structure and process for predicting the values of specified variables in a data set.

Regression Analysis: It is a technique that examines the relation of a dependent variable to specified independent variables.

Regression Tree: It is a tree-shaped structure that represents a set of decisions.

Rule Induction: It is the extraction of useful if-then rules from data based on statistical significance.

A Primer on Text-Data Analysis

Imad Rahal
College of Saint Benedict & Saint John's University, USA

Baoying Wang
Waynesburg College, USA

James Schnepf
College of Saint Benedict & Saint John's University, USA

INTRODUCTION

Since the invention of the printing press, text has been the predominate mode for collecting, storing and disseminating a vast, rich range of information. With the unprecedented increase of electronic storage and dissemination, document collections have grown rapidly, increasing the need to manage and analyze this form of data in spite of its unstructured or semistructured form. **Text-data analysis** (Hearst, 1999) has emerged as an interdisciplinary research area forming a junction of a number of older fields like machine learning, natural language processing, and information retrieval (Grobelnik, Mladenic, & Milic-Frayling, 2000). It is sometimes viewed as an adapted form of a very similar research field that has also emerged recently, namely, data mining, which focuses primarily on structured data mostly represented in relational tables or multidimensional cubes.

This article provides an overview of the various research directions in text-data analysis. After the "Introduction," the "Background" section provides a description of a ubiquitous text-data representation model along with preprocessing steps employed for achieving better text-data representations and applications. The focal section, "Text-Data Analysis," presents a detailed treatment of various text-data analysis subprocesses such as ***information extraction***, ***information retrieval*** and ***information filtering***, ***document clustering*** and ***document categorization***. The article closes with a "Future Trends" section followed by a "Conclusion" section.

BACKGROUND

Text-data analysis is defined as the computerized process of automatically extracting useful knowledge from enormous collections of natural text documents (a.k.a. document collections) usually coming from various dynamic sources. It is a broad process embedding a number of subprocesses, all of which deal with textual resources which are naturally unstructured or semistructured, as in the case of HTML (HyperText Markup Language) and XML (eXtensible Markup Language) documents; a fact that makes it extremely difficult to apply computational solutions to real life text-based problems.

In order to alleviate the difficulty faced by computers when dealing with the unstructured nature of text-data resources, a process called *indexing* is utilized. This process is normally preceded by a number of preprocessing steps that attempt to optimize the indexing process mainly by feature reduction, as explained in this section.

Indexing Textual Data

Indexing is the process of mapping a document into a structured format that captures its content. It can be applied to the whole document or some parts of it, though the former is usually the case. In indexing, the terms occurring in the given collection of documents are used to represent the documents. It is widely known that text documents contain large numbers of terms that have no significant relationship to the context in which they exist. Using all the terms would certainly result in high inefficiencies; therefore, many unrelated terms are usually eliminated through some preprocessing steps, as we shall discuss later.

One very widely used indexing model is the *vector space model* (Salton & Buckley, 1988) which is based on the bag-of-words (or set-of-words) approach. This model has the advantages of relative computational efficiency and conceptual simplicity (Salton & Buckley, 1988); nonetheless, it suffers from the loss of important information about the original text, such as information on the order of the terms in the text or about the boundaries between sentences or paragraphs. In this model, each document is represented as a vector, the dimensions of which are the terms in the initial document collection. The set of terms used as dimensions is referred to collectively as the *term space*. Each vector coordinate is a term having a numeric value representing its relevance to the corresponding document with higher values implying higher relevance. The process of giving numeric

values to vector coordinates is referred to as *weighting*. From an indexing point of view, weighting is the process of giving more emphasis to more important terms.

Three popular weighting schemes have been thoroughly studied in the literature: *binary*, *term frequency (TF)*, and *term frequency by* ***inverse document frequency*** (*TF*IDF*). For a term *t* in document *d*, the binary scheme records binary coordinate values, where a 1 is given to *t* if it occurs at least once in *d*, and a 0 is given otherwise. The TF scheme records *t*'s frequency of occurrence in *d*. It is common to normalize TF measurements in order to help overcome problems associated with document sizes. Normalization may be achieved by dividing all coordinate measurements for every document by the highest coordinate measure for that document. The TF*IDF scheme simply weights TF measurements with a global weight, the IDF (inverse document frequency) measurement. The IDF measure for a term *t* is defined as $log_2 (N/N_t)$, where *N* is the total number of documents in the collection, and N_t is the total number of documents containing at least one occurrence of *t*. The reader should note that IDF increases as N_t decreases, that is, as the uniqueness of the term among the documents in the given collection increases. As with TF, normalization is usually done here too. To normalize measurements based on the TF*IDF scheme, the cosine normalization is usually utilized as shown below:

$$W_{tk,dj} = \frac{TF * IDF(tk, dj)}{\sqrt{\sum_{s=1}^{|T|}(TF * IDF(ts, dj))(TF * IDF(ts, dj))}},$$

where *tk* and *dj* are the term and document under consideration, respectively, TF*IDF (*ts*,*dj*) is the coordinate measure of *ts* in *dj*, and |T| is total number of terms in term space.

Preprocessing

Various preprocessing steps are usually performed on the text corpus prior to indexing in order to optimize the indexing process primarily by reducing the number of terms used, thus leading to faster processing at the application level later on.

Case folding is the process of converting all the characters in a document into the same ***case***, either all upper ***case*** or lower ***case***. This step has the advantage of speeding up comparisons during the indexing process. ***Stemming*** is the process of removing prefixes and suffixes from words to reduce them to ***stem***s, thus eliminating tag-of-speech and other verbal or plural inflections. For example, the words "Computing," "Computer," and "Computational" all map to "Compute." It is worth noting that stemming algorithms have the disadvantage of requiring a great deal of linguistics and are, thus, language dependent. ***Stop words*** are words having no significant semantic relation to the context in which they exist. Stop words can be terms that occur frequently in most of the documents in a given collection (i.e., have low uniqueness and thus low IDF measurements) and as a result, must not be included as indexing terms. For example, articles and propositions such as "the," "on," and "with" are usually stop words. Stop words may also be document-collection specific. For example, the word "blood" would probably be a stop word in a collection of articles addressing blood infections, but certainly not in a collection describing the events of the 2006 FIFA World Cup that took place Germany.

TEXT-DATA ANALYSIS

Text-data analysis (a.k.a. text mining) is a very broad process that can be refined into a number of task-oriented subprocesses. A treatment of the major **text-data analysis** subprocesses follows.

Information Extraction

Many regard *information extraction (IE)* (Cowie & Lehnert, 1996) as the central text-data analysis subprocess largely owing to the success of its applications. IE has emerged as a joint research area between text-data analysis and natural language processing (NLP). It is the process of extracting predefined information on known entities and relationships among those entities from streams of documents and usually storing this information in predesigned templates. Information extraction is associated with streams of documents rather than static collections. One popular application of IE is the extraction of promotions and sales from streams of newspaper documents; the extracted information might be the event, the companies involved, or the event dates. Systems employing such technologies are usually referred to as news-skimming systems.

The IE process is twofold; first, it divides every document into relevant and irrelevant portions, and then, fills the predefined templates with the information extracted from the relevant portions. Simple IE applications, such as extracting proper names or companies from text, is currently being performed with very high precision; however, this is still not the case for more complex tasks, like determining the sequences of events from a document. In such complex tasks, IE systems are usually defined and applied on very restricted domains, normally with the help of domain experts which obviously hinders their portability to other domains. To summarize, IE systems scan streams of documents in order to transform the associated documents into much smaller bits of extracted relevant information that can be more easily maintained and comprehended. A number of very popular IE applications are briefly outlined as follows.

Template filling fills templates with information extracted from a document. Those templates are then stored in structured environments such as databases for fast information retrieval later. Zechner (1997) discusses template-filling systems.

Question answering is a variant of template filling where applications can answer questions like, "Where is Lebanon?" This area is still in its infancy and cannot answer more complex questions such as, "Which country had the lowest inflation in 1999?" However, some systems might be able to point the user to documents that discuss the posed question. BORIS (Lehnert, Dyer, Johnson, yang, & Harley, 1983) is a question-answering IE system which attempts to understand short domain-specific documents and answer questions on them.

Summarization maps documents into extracts (Neto, Santos, Kaestner, & Freitas, 2000) which are machine-made summaries, as opposed to abstracts which are man-made summaries. Applications usually extract a group of highly relevant sentences and present them as summaries. Some systems for this purpose are described in Zechner (1997).

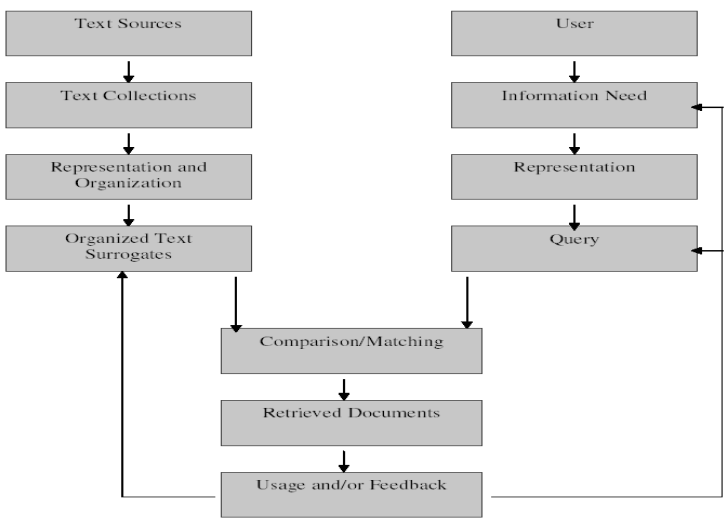
Information Retrieval and Information Filtering

Information retrieval (**IR**) and *information filtering* (IF) (O' Riordan & Sorensen, 2003) are two distinct processes having the same underlying goals. Given an *information need* represented by the user in a suitable manner, they are concerned with discovering a set of documents that satisfy that need. An information need is represented via *queries* in IR systems and *profiles* in IF systems. In the literature, IR is viewed as the ancestor of IF (and of text-data analysis as a whole). The reason for this is that IR is older, and IF bases much of its foundations on IR. Most of the research done in IR has been adapted to fit IF.

In spite of the fact that these two processes are very similar from a foundations point of view, many differences exist between the two that render them as two distinct subprocesses. A comparison between IR and IF is drawn next to highlight the main differences between the two in regard to goals, users, usage, and the type of the data on which they operate. First, the primary goal of IR is to collect and organize documents that match a given query according to some ranking function. IF, on the other hand, operates on document streams and is more concerned with routing newly received documents to interested users. Second, IR systems are usually used once by a one-time query user (Belkin & Croft, 1992) with short-term needs while IF serves a subscribed audience whose needs do not change much over long periods of time. Third, IR systems are developed to tolerate some inadequacies in the query representation of the information need; on the other hand, profiles are assumed be highly accurate in IF systems. Finally, IR systems usually operate over static collections of documents, while IF systems deal with dynamic streams of documents.

IR retrieves a set of documents from a collection that match a certain query. The retrieved documents are then rank ordered and presented to the user. Figure 1 presents a general model for IR as described in Belkin (Belkin & Croft, 1992). In this model, a user with some information need presents a query to the IR system. As mentioned earlier, a query is a simple representation of the user's information need in a language understood by the system. This representation is considered as an approximation due to the difficulty associated with representing information needs accurately. The query is then matched against the documents, which are organized into *text surrogates*. The collection of text surrogates

Figure 1. A general IR model



can be viewed as a summarized structured representation of unstructured text data, such as document vectors in the vector space model. Thus, they provide an alternative to the original documents as they take far less time to examine, and, at the same time, encode enough semantic cues to be used for matching instead of the original documents. As a result of matching, a set of documents would be selected and presented to the user. The user either uses the returned documents or provides *relevance feedback* to the system resulting in modifications to the query and the original information need or, in rare cases, to the text surrogates (Belkin & Croft, 1992). Google is one of the most popular systems using the IR model.

IF systems deal with large streams of incoming documents, usually broadcasted via remote sources. It is sometimes referred to as *document routing*. The system maintains profiles created by subscribed users to describe their long-term interests. Profiles may describe what the user likes or dislikes. New incoming documents are removed from the stream routed to a subscribed user if those documents do not match the user's profile. As a result, the user only views what is left in the stream after the mismatching documents have been removed; an e-mail filter, for example, removes all "junk" e-mail. Figure 2 depicts a general model for IF (Belkin & Croft, 1992).

The first step in using an IF system is to create a profile for a new subscribed user. A profile represents a user's (or a group of users') information need, which is assumed to be static over a long period of time. Whenever a new document is received through the document stream, the system represents it as text surrogates and compares it against every profile stored in the system. If the document matches a profile, it will be routed to the corresponding user. The user can then use the received documents or provide relevance feedback. The feedback provided may lead to modifications in the profile and the information need as Figure 2 shows.

Document Clustering

Document clustering (Neto et al., 2000; Steinbach, Karypis, & Kumar, 2000) is the process of grouping similar documents into partitions where documents within the same partition exhibit higher degree of similarity among each other than to any other document in any other partition. Clusters are usually mutually exclusive and collectively exhaustive. Figure 3 depicts a general clustering scheme. Clustering has been employed in a number of text-based applications such document categorization (Rahal & Perrizo, 2004) and IR (Kowalski, 1997). Some examples of applications that employ document clustering include browsing a collection of documents (Cutting, Karger, Pedersen, & Tukey, 1992), organizing documents returned by a search engine for some query (Zamir, Etzioni, Madami, & Karp, 1997), and automatically generating hierarchical clusters of documents (Koller & Sahami, 1997).

The two most popular techniques for document clustering are *hierarchical* and *k-means* clustering. Hierarchical clustering techniques produce a hierarchy of partitions with a single partition including all documents, at one end, and singleton clusters, each composed of an individual document, at the other end. The tree depicting the hierarchy of clusters is referred to as a *dendrogram*. Each cluster along the hierarchy is viewed as a combination of two clusters from the next lower or higher level, depending on the type of hierarchical clustering used, that is *divisive* or *agglomerative*, respectively. Agglomerative clustering starts with the set of all singleton clusters (i.e., each cluster includes one document) at the root of the tree, and then combines the most similar pair of clusters together at every tree level

Figure 2. A general IF model

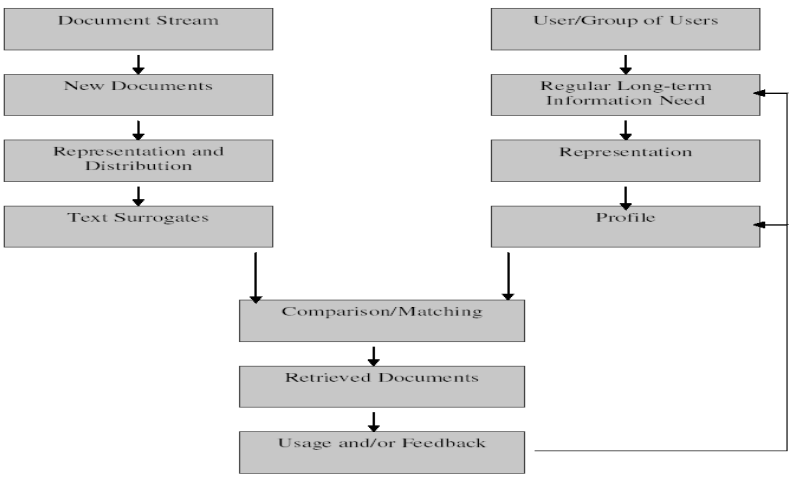
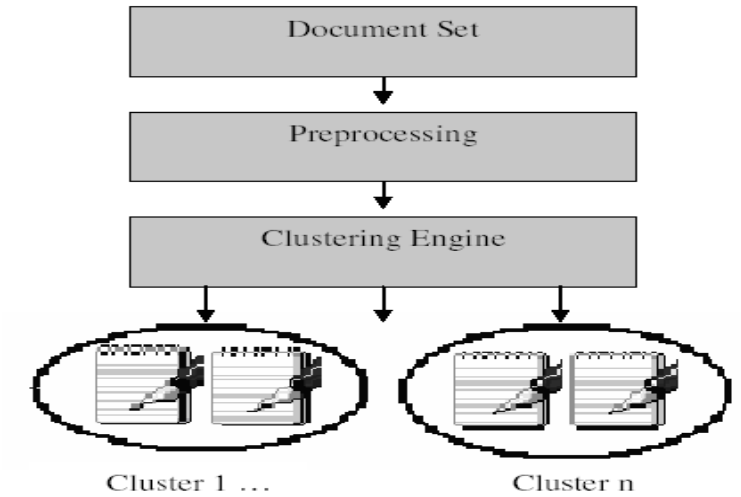


Figure 3. A general depiction of the clustering process



until it forms one cluster containing all documents at the leaf level. Divisive clustering, on the other hand, starts with a single cluster containing all documents at the root level and then splits one cluster into two clusters at every level of the tree until it forms a set of clusters, at the leaf level, each containing one and only one document.

Given a fixed number k, k-means clustering creates a set of k clusters and distributes the set of given documents among those clusters using the similarity between the document vectors and the cluster means. A mean for a given cluster is the average vector of all document vectors in that cluster. Every time a document is added to a cluster, the mean of that cluster must be recalculated. Similarity between a document and a mean can be computed using cosine similarity between the corresponding vectors. The mean does not always correspond to an actual document; as a result, a variant of k-means, called *k-mediods*, requires the mean to be an actual document vector. To do this, the document vector in the cluster that is closest to the mean is selected to serve as the mean of the cluster (Han & Kamber, 2006, pp.383-406).

In general, it is believed that hierarchical clustering techniques produce better cluster quality than their k-means counterparts while suffering from quadratic time complexity. On the other hand, k-means clustering and its variants have complexities linear to the number of documents, but produce clusters with lower quality (Steinbach, et al., 2000). The entropy (Steinbach et al., 2000) and the F-measure (Steinbach et al., 2000) measures are two very popular schemes used in evaluating the quality of the clusters produced by a clustering technique.

Document Categorization

Given a set of predefined document labels or categories, *document categorization* (Rahal & Perrizo, 2004), sometimes referred to as *topic spotting* or *text classification*, is the process of labeling unlabeled text documents with their corresponding category (or categories) based entirely on their content; in other words, the categorization process relies entirely on endogenous knowledge (knowledge extracted from the document) as opposed to exogenous knowledge (knowledge extracted from external resources). Categories are chosen to correspond to the topics or themes of the documents. The ultimate purpose of document categorization is automatic organization. Some categorizations systems (or classifiers) return one category per document, while others return multiple categories. Sometimes, a classifier might return no category or a number of categories but with very low confidence which results in flagging the document for manual inspection. The following is a brief account of two major applications in document categorization.

In *document organization* applications, document categorization is used for the purpose of organizing documents into categories for personal or corporate use. For example, newspapers receive a number of ads everyday and would benefit from an automatic system that can assign those ads to their corresponding categories, such as Real Estate or Autos. An important consequence of document organization is ease of search. Larkey (1999) provides an interesting example of this type of application.

Word sense disambiguation (WSD) is the process through which a system is able to find the sense of an ambiguous word (i.e., has more than one meaning) based on the context in which it occurs. For example, the word “class” might

mean a place where students study, or a category. Given a text containing an ambiguous word, a WSD application returns which of the meanings associated with the word is intended by the given text. WSD is heavily used in natural language processing (NLP) and in indexing documents by using word senses. See Roth (1998) for more details.

Regardless of the application, a text classifier can be either *eager* or *example-based*. In eager categorization, a set of precategorized documents is used to build the classifier. This data set is divided into a *training set*, TR, and *testing set*, TS. The two partitions need not be equal in size (and usually are not). The TR is used to build or train the classifier, and then the TS is used to test the accuracy of the classifier. This process is repeated until an acceptable classifier is attained. Some popular approaches that fall under this category include, *decision-tree* classifiers such ID3, C4.5 and C5 (Han & Kamber, 2006, pp. 291-309), *Rocchio* classifiers (Cohen & Singer, 1999; Sebastiani, 2003), *neural networks* (NN) (Lam & Lee, 1999), and *support vector machines* (SVMs) (Joachims, 1999).

On the other hand, in example-based categorization, no classifier is built in advance; rather, the set of all labeled samples is used to categorize new unlabeled samples. One very popular classifier under this category is the *k*-nearest neighbors (*k*NN) classifier (Yang, 1994) which, given a new sample d_j, tries to find k documents that are the most similar to d_j among the given set of samples using some distance measure, such as the *Euclidean* distance. Then a process, such as *plurality voting*, is performed by the selected neighbors to decide on the most appropriate label for d_j.

It is worth noting that almost all current text classifiers suffer from performance considerations related to accuracy and speed. Perhaps this fact is the main reason for the huge amount of research currently being invested in this promising area.

FUTURE TRENDS

With this vast and continuous spread of text information over the Internet, text-data analysis continues to attract attention with the promise of delivering critical hidden nuggets of knowledge extracted from huge text resources. From a business standpoint, this might mean better decision-making processes leading to increased profitability in market competitions based on Aristotle Onassis's maxim: "*The secret of success is to know something that nobody else knows*." (Prionas et al., 1996). This potential coupled with the present unsatisfactory text-application reliability will continue to push researchers in the near future to dedicate more financial and time resources to achieve higher degrees of accuracy for what once was viewed as a seemingly far-fetched objective.

The World Wide Web (WWW) is a dynamic entity that continues to evolve in order to accommodate our needs and ambitions. A recent observed trend has been the spread of multimedia data over the Internet intermixed with text data. The result is push for the adaptation and expansion of text-data analysis to deal with new forms of "text," such as audio, video and images.

CONCLUSION

This article has presented an overview of text-data analysis including text indexing, text preprocessing, and various text-data analysis subprocesses. Text-data analysis is an infant interdisciplinary field that lies at the junction of a number of older and more established fields like machine learning, natural language processing, information retrieval and the like. Its remarkable functional similarity to data mining is balanced off by the fact that each of those two fields deals with a different type of data: unstructured text data vs. structured data represented via relational tables or multidimensional cubes. Consequently, text-data analysis faces an additional challenge due to the unstructured or semistructured nature of text data.

REFERENCES

Belkin, N.J., & Croft, W.B. (1992). Information filtering and information retrieval: Two sides of the same coin? *Communications of the ACM, 35*(12), 29-38.

Cohen, W.W., & Singer, Y. (1999). Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems, 17*(2), 141-173.

Cowie, J., & Lehnert, W. (1996). Information extraction. *Communications of the ACM, 39*(1), 80-91.

Cutting, D.R., Karger, D.R., Pedersen, J.O., & Tukey, J.W. (1992). Scatter/gather: A cluster-based approach to browsing large document collections. In *Proceedings of the ACM SIGIR, International Conference on Research and Development in Information Retrieval, Copenhagen, Denmark*, (pp. 318-329).

Grobelnik, M., Mladenic, D., & Milic-Frayling, N. (2000). Text mining as integration of several related research areas: Report on KDD'2000 Workshop on Text Mining. *ACM SIGKDD Explorations, 2*(1), 99-102.

Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann.

Hearst, M.A. (1999). Untangling text data mining. In *Proceedings of the ACL, the Annual Meeting of the Association*

for Computational Linguistics, College Park, MD, (pp. 20-26).

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of the ICML, International Conference on Machine Learning*, Beld, Slovenia, Germany, (pp. 200-209).

Koller, D., & Sahami, M. (1997). Hierarchically classifying document using very few words. In *Proceedings of the ICML, International Conference on Machine Learning*, Nashville, TN, (pp. 170-178).

Kowalski, G. (1997). *Information retrieval systems: Theory and implementation*. Norwell, MA: Kluwer Academic.

Lam, S.L., & Lee, D.L. (1999). Feature reduction for neural network based test categorization. In *Proceedings of the IEEE DASFAA, International Conference on Database Advanced Systems for Advanced Applications*, Hsinchu, Twaiwan, (pp. 195-202).

Larkey, L.S. (1999). A patent search and classification system. In *Proceedings of the ACM DL, Conference on Digital Libraries*, Berkley, CA, (pp. 90-95).

Lehnert, W. G., Dyer, M. G., Johnson, P. N., Yang, C. J. & Harley, S. (1983). BORIS—an experiment in in-depth understanding of narratives. *Artificial Intelligence Journal, 20*(1), 15-62.

Neto, J.L., Santos, A.D., Kaestner, C.A., & Freitas, A.A. (2000). Document clustering and text summarization. In *Proceedings of the PADD, International Conference on Practical Applications of Knowledge Discovery and Data Mining*, London, England, (pp. 41-55).

O' Riordan, C., & Sorensen, H. (1997). *Information filtering and retrieval: An overview*. Retrieved December 11, 2007, from http://citeseer.ist.psu.edu/483228.html

Prionas, E, Kiriazis, C, Elisofon, M, Roberts, A, & Salter, A. (1996). *The life of Aristotle Onassis: The man, the myth, the legend*. Retrieved December 11, 2007, from http://www.greece.org/poseidon/work/modern times/onassis.html

Rahal, I., & Perrizo, W. (2004). An optimized approach for kNN text categorization using P-tees. In *Proceedings of the ACM SAC, Symposium on Applied Computing, Nicosia*, Cyprus, (pp. 613-617).

Roth, D. (1998). *Learning to resolve natural language ambiguities: A unified approach*. In *Proceedings of the AAAI, Conference of the American Association for Artificial Intelligence*, Madison, WI, (pp. 806-813).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513-523.

Sebastiani, F. (2002). Machine learning in automated text categorisation. *ACM Computing Surveys, 34*(1), 1-47.

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. In *Proceedings of the ACM KDD Workshop on Text Mining*, Boston, MA. Retrieved December 11, 2007, from http://www.cs.cmu.edu/~dunja/KDDpapers/Steinbach_IR.pdf

Yang, Y. (1994). Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the ACM SIGIR, International Conference on Research and Development in Information Retrieval*, Dublin, Ireland, (pp. 13-22).

Zamir, O., Etzioni, O., Madami, O., & Karp, R.M. (1997). Fast and intuitive clustering of Web documents. In *Proceedings of the ACM SIGKDD, International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, (pp. 287-290).

Zechner, K. (1997). *A literature survey on information extraction and text summarization.* Term paper, Carnegie Mellon University, Pittsburgh, PA. Retrieved December 11, 2007, from http://www.cs.cmu.edu/~zechner/infoextr.pdf

KEY TERMS

Case Folding: The process of converting all the characters in a document into the same case, either all upper case or lower case, in order to speed up comparisons during the indexing process.

Document Categorization: The process of labeling unlabeled text documents with their corresponding category based entirely on their content.

Document Clustering: The process of grouping similar documents into partitions where documents within the same partition exhibit higher degree of similarity among each other than to any other document in any other partition.

Indexing: Indexing is the process of mapping a document into a structured (tabular) format that captures its content.

Information Extraction: The process of extracting predefined information on known entities and relationships among those entities from streams of documents and storing this information in pre-designed *templates.*

Information Filtering: The process of eliminating documents from a document stream routed to a subscribed user based on the user's profile.

Information Retrieval: The process of discovering a set of documents from static collection which satisfy a user's temporary information need defined by a query.

Stemming: The process of removing prefixes and suffixes from words to reduce them to stems thus eliminating tag-of-speech and other verbal or plural inflections.

Principles of Digital Video Coding

Harilaos Koumaras
University of the Aegean, Greece

Evangellos Pallis
Technological Educational Institute of Crete, Greece

Anastasios Kourtis
National Centre for Scientific Research "Demokritos", Greece

Drakoulis Martakos
National and Kapodistrian University of Athens, Greece

INTRODUCTION

Multimedia applications and services have already possessed a major portion of today's traffic over communication networks. The revolution and evolution of the World Wide Web has enabled the wide provision of multimedia content over the Internet and any other autonomous network.

Among the various types of multimedia, video services (transmission of moving images and sound) are proven dominant for present and future communication networks. Although the available network bandwidth and the corresponding supporting bit rates continue to increase, the raw video data necessitate high bandwidth requirements for its transmission. For example, current commercial communication networks throughput rates are insufficient to handle raw video in real time, even if low spatial and temporal resolution (i.e., frame size and frame rate) has been selected. Towards alleviating the network bandwidth requirements for efficient transmission of audiovisual content, coding techniques have been applied on raw video data, which perform compression by exploiting both temporal and spatial redundancy in video sequences.

Video coding is defined as the process of compressing and decompressing a raw digital video sequence, which results in lower data volumes, besides enabling the transmission of video signals over bandwidth-limited means, where uncompressed video signals would not be possible to be transmitted. The use of coding and compression techniques leads to better exploitation and more efficient management of the available bandwidth.

Video compression algorithms exploit the fact that a video signal consists of sequence series with high similarity in the spatial, temporal, and frequency domain. Thus, by removing this redundancy in these three different domain types, it is possible to achieve high compression of the deduced data, sacrificing a certain amount of visual information, which however it is not highly noticeable by the mechanisms of the *human visual system*, which is not sensitive at this type of visual degradation (Richardson, 2003).

Thus, the research area of video compression has been a very active field during the last few years by proposing various algorithms and techniques for video coding (International Telecommunications Union [ITU], 1993; ITU 2005a, 2005b; Moving Picture Experts Group [MPEG], 1998; MPEG, 2005a, 2005b). In general video compression techniques can be classified into two classes: (1) the *lossy* ones and (2) information preserving (*lossless*). The first methods, although maintaining the video quality of the original/uncompressed signal, do not succeed high compression ratios, while the lossless ones compress more efficiently the data volume of initial raw video signal with the cost of degrading the perceived quality of the video service.

The lossy video coding techniques are widely used, in contrast to lossless ones, due to their better performance. More specifically, by enhancing the encoding algorithms and techniques, the latest proposed coding methods try to perform in a more efficient way both the data compression and the maintenance of the deduced perceived quality of the encoded signal at high levels. In this framework, many of these coding techniques and algorithms have been standardized, encouraging by this way the interoperability between various products designed and developed by different manufacturers.

This article deals with the fundamentals of the video coding process of the lossy encoding techniques that are common on the great majority of today's video coding standards and techniques.

BACKGROUND

The majority of the compression standards have been proposed by the ITU and the International Organization for Standardization (ISO) bodies, by introducing the fol-

lowing standards H.261, H.263, H.263+, H.263++, H.264, MPEG-1, MPEG-2, MPEG-4 and MPEG-4 Advanced Video Coding (AVC).

Some of the aforementioned standards were developed in partnership of ITU with MPEG, exploiting similar coding techniques developed by each body separately.

Each standard was designed and proposed targeting a specific service and application, featuring therefore specific parameters and characteristics. For example H.261 was proposed in 1990 for transmission of video signals over Integrated Services Digital Network (ISDN) lines on which data rates are multiples of 64 kbit/sec. The H.263 standard was designed as a low bit rate encoding solution for video-conferencing applications.

Similarly MPEG-1 was proposed by MPEG in order to be used by the video compact disc (VCD) medium, which stores digital video on a compact disc (CD) with a quality almost similar to that of an analog VHS video. In 1994 MPEG-2 was proposed for encoding audio and video for broadcast signals, exploiting interlace format. MPEG-2 is also the coding format used by the widely successful commercial digital versatile disc (DVD) medium.

Regarding the latest H.264, or MPEG-4 Part 10 AVC, it was proposed in common by the ITU Telecommunication Standardization Sector (ITU-T) Video Coding Experts Group (VCEG) and the ISO MPEG as the outcome of a joint venture effort known as the Joint Video Team (JVT). The scope of H.264/AVC project is to create a standard that would be capable of providing broadcast video quality at very low bit rates on a wide variety of applications, networks and systems.

Finally, in order to create a framework, which will reassure the interoperability of the codec implementation among the various developers, the standards include the concept of profiles and levels, defining a specific set of capabilities to be defined and implemented for a specific subset of applications and services.

VIDEO CODING

All the aforementioned video coding standards are based on the same basic coding scheme, which briefly consists of the following stages: (1) the temporal, (2) the spatial, (3) the transform, (4) the quantization and (5) the entropy coding stage.

The temporal stage exploits the similarities between successive frames with scope to reduce the temporal redundancy in a video sequence. The spatial stage exploits spatial similarities located on the same frame, reducing by this way the spatial redundancy. Then the output parameters of the temporal and spatial stages are further quantized and compressed by an entropy encoder, which removes the statistical redundancy in the data, producing an even more compressed video stream.

More analysis of each stage follows.

TEMPORAL STAGE

As input to the temporal stage of the encoding process, the uncompressed video sequence is used, which contains a lot of redundancy between its successive frames. The scope of this stage is to remove this redundancy by constructing a prediction of each frame based on previous or future frames, enhanced by compensating for fine differences between the selected reference frames. Depending on the prediction level, by which each frame is constructed, each frame is classified to three discrete types, namely:

(1) Intra-frame (I), (2) Predictive (P) and (3) Bidirectional predictive (B), widely referred as I, P, and B. The I-frames are also called Intra frames, while B and P are known as Inter frames.

- I-frames do not contain any prediction from any other coded frame.
- P-frames are coded based on prediction from previously encoded I- or P-frames.
- B- frames are coded based on prediction from previously or future encoded I- or P-frames.

The pattern of successive types of frames like IBBPBBPBBP… forms a Group of Pictures (GOP), whose length is mainly described by the distance of two successive I-frames.

Therefore, in order to perform this temporal compression, two discrete processes are performed at this stage—the motion estimation and motion compensation. Both these processes are usually applied on specific rectangular regions of a frame, called blocks if their size is 8 x 8 pixels or macroblocks if they are 16 x 16 rectangular pixel regions. At the latest standards (i.e., H.264) as Figure 1 depicts, variable block sizes are used for motion compensation depending on the content, achieving better coding efficiency.

During motion estimation, the encoding algorithm searches for an area in the reference frame (past or future frame) in order to find a corresponding matching region. The process of specifying the best match between a current frame and a reference one, which will be used as a predictor of the current frame, is called motion estimation. This is performed by comparing specific rectangular areas (i.e., blocks/macroblocks) in the reference and current frame, until the best match is detected. Due to this, their spatial differences are calculated, using the Sum of Absolute Differences (SAD):

$$SAD(d_x,d_y)=\sum_{i=0}^{15}\sum_{j=0}^{15}|f(i,j)-g(i-d_x, j-d_y)|$$

Figure 1. Example of variable block size coding

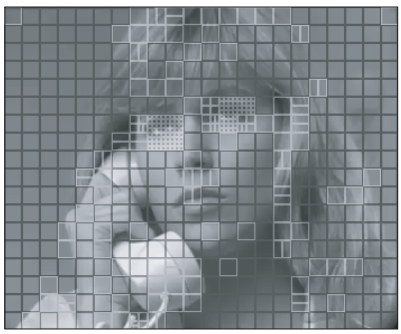
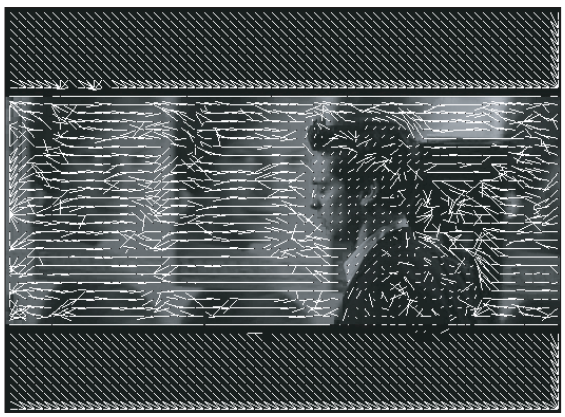


Figure 2. A frame where motion vectors appear denoting the position of the best matching region



where *f(i,j)* and *g(i,j)* denote the luminance pixels of the current rectangular area (in this case a Macroblock) and the reference one respectively. The reference area is relatively defined by the current one using the motion vectors (d_x,d_y), denoting the position of the best matching region (see Figure 2).

When the best match has been performed, then the motion compensation follows. During this process the selected optimal matching region in the reference frame (i.e., the region that sets the SAD minimum) is subtracted from the corresponding region in the current frame with scope to produce a luminance and chrominance residual block/macroblock that is transmitted and encoded along with the reference motion vectors. The deduced frame by the motion compensation process is called residual frame, which contains the result of the subtraction of the reference regions from the corresponding ones of the current frame. In the residual frame the static areas correspond to difference equal to zero, while darker areas denote negative differences and lighter areas positive differences respectively. A typical example of a residual frame is represented in Figure 3.

Thus motion compensation enhances the efficiency of the motion estimation by adding at the predicted frame the fine differences that may contain the motion estimated predicted regions in comparison to the actual frame. Thus, during motion estimation the best matches between reference and current frames are detected and this match is further improved by motion compensation, which calculates the residuals of the motion estimated frame and the actual frame. So, by adding this motion compensated residual information on the motion estimated frame, an accurate and efficient prediction of the current frame can be performed, using regions of past or future frames.

SPATIAL STAGE

Similarly to the temporal stage, where predictive coding is performed between successive frames, a prediction of an image region may be also performed based on samples located within the same image or frame, which is usually referred to as Intra coding. At spatial stage, the encoder performs a prediction for a pixel-based pattern on a combina-

Figure 3. A residual frame denoting the differences between two successive frames for the fireman reference sequence

tion of previously coded pixels located on the same frame. Especially for frames that contain homogeneous areas, the spatial prediction can be quite efficient. In the case of a good prediction then the residual energy is small and the corresponding compression ratio high.

TRANSFORM CODING STAGE

At this stage the spatially/temporally encoded frames or the motion compensated, residual data are converted into another domain, usually called the transformed domain, where the optically correlated data become decorrelated. The use of transformation facilitates the exploitation in the compression technique of the various psycho-visual redundancies by transforming the picture to a domain where different frequency ranges with dissimilar sensitivities at the human visual system (HVS) can be accessed independently (Winkler, 2005).

The most commonly used transformation is the Discrete Cosine Transform (DCT). The DCT operates on an X block of N X N image samples or residual values after prediction and creates Y, which is a N X N block of coefficients. The action of the DCT can be described in terms of a transform matrix A. The forward DCT is given by:

$$Y=AXA^T$$

where X is a matrix of samples, Y is a matrix of coefficients and A is an N X N transform matrix. The elements of A are:

$$A_{ij} = C_i \cos\frac{(2j+1)i\pi}{2N} \text{ where } C_i \begin{cases} \sqrt{1/N}, & i=0 \\ \sqrt{2/N}, & i>0 \end{cases}$$

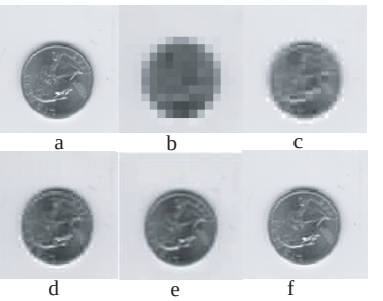
Therefore the DCT can be written as:

$$Y_{xy} = C_x C_y \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} X_{ij} \cos\frac{(2j+1)y\pi}{2N} \cos\frac{(2i+1)x\pi}{2N}$$

The advantage of the DCT transform is that it is possible to reconstruct quite satisfactorily the original image, applying the reverse DCT on a subset of the DCT coefficients, without taking under consideration the rest coefficients with insignificant magnitudes (see Figure 4).

Thus, with the cost of some quality degradation, the original image can be satisfactorily reconstructed with a reduced number of coefficient values. This DCT property is exploited by the following stage where quantization of the DCT coefficients is performed.

Figure 4. Example of DCT efficiency. Figure a is the source image, while b is reconstructed using only 1 DCT coefficient, b exploits 4 coefficients, c uses 8 coefficients, d uses 12, e uses 18, and f 32 out of the 64 total DCT coefficients for each block (i.e., 8 x 8).



QUANTIZATION

Quantization is the process of approximating the continuous range of DCT coefficients by a relatively small set of discrete integer values. The best-known form of quantization is the scalar quantizer, which maps one sample of the input to one quantized output value. A scalar quantization operator *Q()* can be mathematically represented as

$$Q(x) = g(\lfloor f(x) \rfloor)$$

where

- x is a real number
- $\lfloor x \rfloor$ is the floor function
- *f(x)* and *g(i)* are arbitrary real-valued functions.

The integer value $i=\lfloor f(x) \rfloor$ is the representation that is usually stored or transmitted, but the final interpretation may be further modified using also *g*(*i*). Thus, typically during a scalar quantization process, it has performed the rounding of a fractional number to its nearest integer:

$$Y=QP\ round\left(\frac{X}{QP}\right)$$

where QP is the quantization parameter (i.e., quantization step size), X the initial integer value, and Y the deduced quantized number. Table 1 depicts some representative examples of the scalar quantization process for various quantization parameters.

Table 1. Quantization examples

X	Y		
	QP=1	QP=2	QP=3
0	0	0	0
1	1	0	0
2	2	2	3
3	3	2	3
4	4	4	3
5	5	4	6
6	6	6	6
7	7	6	6
8	8	8	9
9	9	8	9

Applying quantization on the aforementioned DCT coefficients is the main reason for the quality degradation and the appearance of artifacts, like the blockiness effect, at the digitally encoded videos. The blockiness effect refers to a block pattern of size 8 x 8 pixels in the compressed sequence, which is the result of the independent quantization of individual blocks of block-based DCT. Due to the quantized DCT coefficients, within a block (8 x 8 pixels), the luminance differences and discontinuities between any pair of adjacent pixels are reduced. On the contrary, for all the pairs of adjacent pixels, which are located across and on both edge sides of the border of adjacent DCT blocks, the luminance discontinuities are increased, by the coding process. This happens because the quantization process is lossy (i.e., not totally reversible) since it is not possible to determine the accurate fractional number from the deduced rounded integer. So it is somewhat equivalent with the case of not exploiting the entire DCT coefficient set for the reconstruction of the original image, as in Figure 4, because some low DCT values may have been quantized to zero.

It must be noted that the quantization stage is the only lossy stage at the described coding chain and is mainly responsible for any visual artifact and quality degradation that may appear on the deduced coded video signal.

ENTROPY CODING STAGE

At this final stage it has performed a transformation of the video sequence symbols into a compressed stream. The term video sequence symbol stands for all the aforementioned encoding parameters, such as quantization coefficients, motion vectors, and so forth. Basically, two widely known variable length coding techniques are exploited at this stage: The Huffman Coding (Huffman, 1952) and the Arithmetic Coding (Witten, Neal, & Cleary, 1987).

The variable length coding methods assign to each video sequence symbol a variable length code, based on the probability of its appearance. Symbols appearing frequently are represented with short variable length codes while less common symbols are represented with long variable length codes. Over a large number of encoded symbols, this replacement of video sequence symbols by variable length codes leads to efficient compression of the data. (Held & Marshall, 1991)

BLOCK DIAGRAM OF A DIGITAL VIDEO ENCODER

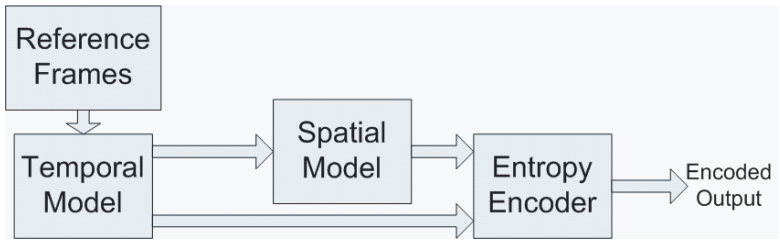
Based on the aforementioned description and analysis of the digital video coding stages, the following generalized block diagram of a video coded is presented in Figure 5 (Richardson, 2003).

In this diagram, we present the interconnection between the various stages that a typical video encoder follows during the digital video coding process, considering the transform and the quantization stage as internal processes of the spatiotemporal stages.

FUTURE TRENDS

Digital video coding techniques will prevail in the upcoming multimedia services and applications, because they will

Figure 5. Generalized block diagram of digital video encoder



enable the provision of such content over various bandwidth limited means and terminals. Towards this, the research community has been focused on developing video coding methods, which will be based on the aforementioned described encoding chain, but will be able to adapt dynamically to the coded video stream depending on the available bandwidth and the terminal characteristics. Dynamic video adaptation schemes will enable the seamless distribution of multimedia content over heterogeneous networks and terminal devices.

CONCLUSION

Video compression algorithms exploit the redundancy that a video signal contains in the spatial, temporal, and frequency domain. Thus, by removing this redundancy in these three different domain types, it is possible to achieve high compression of the deduced data. Briefly the encoding process consists of the following stages: (1) the temporal, (2) the spatial, (3) the transform, (4) the quantization, and (5) the entropy coding stage. This article has presented, explained, and analyzed the principles of each encoding stage, which remain the common basis over any existing video coding standard.

REFERENCES

Held, G., & Marshall, T. R. (1991). *Data compression.* Chichester, England: Wiley.

Huffman, D. (1952). *A method for the construction of minimum redundancy codes. Proceedings of the IRE, 40,* (pp. 1098-1101).

International Telecommunications Union-Radiocommunications (ITU-R). (1993, March). *Video codec for audiovisual services at p x 64 kbit/s.* (Recommendation H.261).

International Telecommunications Union-Radiocommunications (ITU-R). (2005a, January). *Video coding for low bit rate communication.* (Recommendation H.263).

International Telecommunications Union-Radiocommunications (ITU-R). (2005b, March). *Advanced video coding for generic audiovisual services.* (Recommendation H.264).

Moving Picture Experts Group (MPEG). (1998). *Coding of moving pictures and associated audio for digital storage media at up to about 1,5 Mbit/s.* (MPEG-1 ISO/IEC 11172-5:1998).

Moving Picture Experts Group (MPEG). (2005a). *Generic coding of moving pictures and associated audio information.* (MPEG-2 ISO/IEC 13818-5:2005).

Moving Pictures Experts Group (MPEG). (2005b). *MPEG-4 Coding of audio visual objects.* (MPEG-4 ISO/IEC 14496-5:2001/Amd.6:2005).

Richardson, I. E. G. (2003). *H.264 and MPEG-4 video compression.* Chichester, England: Wiley.

Winkler, S. (2005). *Digital video quality—Vision models and metrics.* Chichester, England: Wiley.

Witten, H., Neal, M., & Cleary G. (1987). Arithmetic coding for data compression. *Communications of the ACM, 30*(6), 520-540.

KEY TERMS

Bit Rate: Bit rate is the frequency at which bits are passing over a given physical medium. It is quantified by using the *bit per second* (*bit/s*) unit.

Frame: Frame is one of the many still images that as a sequence compose a video signal,

Integrated Services Digital Network (ISDN): ISDN is a type of circuit-switched, telephone network system designed to allow digital transmission of voice and data over ordinary telephone copper wires resulting in better quality and higher speeds than available with analog systems.

International Organization for Standardization (ISO): ISO is an international standard-setting body composed of representatives from national standards bodies. Founded in 1947, the organization produces worldwide industrial and commercial standards.

Moving Picture Experts Group (MPEG): MPEG is a working group of ISO charged with the development of audiovisual encoding standards. MPEG includes many members from various industries and universities related to audiovisual coding research.

Multimedia: Multimedia is the several different media types (e.g., text, audio, graphics, animation, video).

Pixel: A pixel is considered the smallest sample of a digital image or video.

Video Codec: Video codec is the device or software that enables the compression/decompression of digital video.

Video Coding: Video coding is the process of compressing and decompressing a raw digital video sequence.

Process-Aware Information Systems for Virtual Teamwork

Schahram Dustdar
Vienna University of Technology, Austria

INTRODUCTION

The question of the "right" organizational form and the appropriate information systems support remains of paramount importance and still constitutes a challenge for virtually all organizations, regardless of industrial background. Organizations distribute their required work activities among groups of people (teams), with teams constituting the main building block for implementing the work (tasks). In most cases, team members are organized as "virtual (project) teams." These teams are under heavy pressure to reduce time to market of their products and services and lower their coordination costs. Some characteristics of distributed virtual teams are that team (member) configurations change quite frequently and that team members report to different managers, maybe even in different organizations. From an information systems' point of view, distributed virtual teams often are self-configuring networks of mobile and "fixed" people, devices, as well as applications. A newly emerging requirement is to facilitate not just mobility of content (i.e., to support a multitude of devices and connectivity modes) to team members, but also to provide contextual information on work activities to all distributed virtual team members (Dustdar, 2002a, 2002b, 2002c). By context, we mean traceable and continuous views of associations (relationships) between artifacts (e.g., documents, database records), resources (e.g., people, roles, skills), and business processes. Context is composed of information on the "who, when, how, and why." The remainder of this chapter is organized as follows: The next section provides an overview of related work on classification systems of collaborative systems and provides an overview on evaluation aspects of current collaborative systems for virtual teamwork. Section 3 discusses some issues and problems related to the integration of artifacts, resources, and processes. Section 4 presents one proposed solution. Finally, Section 5 discusses some future trends and concludes the chapter.

FUNCTIONAL CLASSIFICATION OF COLLABORATIVE SYSTEMS

There has been a lot of work on classification models for collaborative systems. However, there is no one-and-agreed-upon taxonomy of analyzing and understanding collaborative systems. Academia and industry suggest various classification schemes. In industry, for example, people frequently use the term *e-mail* and *groupware* interchangeably. More generally, there is the tendency to classify categories of collaborative systems by naming a product (e.g., many use the terms *Lotus Notes* and *groupware* interchangeably). Academic research has suggested many different classification models. For a recent survey of collaborative application taxonomies, see Bafoutsou and Mentzas (2002). DeSanctis and Gallupe (1987), Ellis, Gibbs and Rein (1991), and Johansen (1988) suggest a two dimensional matrix based on time and place, where they differentiate between systems' usage at same place/same time (e.g., electronic meeting rooms), same place/different time (e.g., newsgroups), different place/different time (e.g., workflow, e-mail), different place/same time (e.g., audio/video conferencing, shared editors). This classification model helps one to easily analyze many tools on the market today; however, it fails to provide detailed insights on collaborative work activities themselves, as well as their relationship to business processes. Ellis (2000) provides a functionally oriented taxonomy of collaborative systems that helps one to understand the integration issues of workflow and groupware systems. The classification system of Ellis (2000) provides a framework in which to understand the characteristics of collaborative systems and their technical implementations.

The first category (Keepers) provides those functionalities related to storage and access to shared data (persistency). The metaphor used for systems based on this category is a "shared workspace." A shared workspace is basically a central repository where all team members put (upload) shared artifacts (in most cases, documents) and share those among the team members. Technical characteristics of "keepers" include database features, access control, versioning, and backup/recovery control. Examples of popular systems include *BSCW* (Bentley et al., 1997), IBM/Lotus *TeamRoom* (IBM, 2002), and the peer-to-peer workspace system *Groove* (Groove, 2002). The second category (Communicators) groups all functionality related to explicit communications among team members. This boils down to messaging systems (e-mail). Its fundamental nature is a point-to-point interaction model where team members are identified only by their name (e.g., e-mail address) and not by other means (e.g., skills, roles, or other constructs, as in some advanced workflow systems).

The third category (Coordinators) is related to the ordering and synchronization of individual activities that make up a whole process. Examples of Coordinator systems include workflow management systems. Finally, the fourth category (Team-Agents), refers to semi-intelligent software components that perform domain-specific functions and thereby help the group dynamics. An example of this category is a meeting scheduler agent. Most systems in this category are not off-the-shelf standard software. Both evaluation models presented above provide guidance to virtual teams on how to evaluate products based on the frameworks. Current systems for virtual teamwork have their strength in one or two categories of Ellis' framework. Most systems on the market today provide features for Keepers and Communicators support or are solely Coordinator systems (e.g., Workflow Management Systems) or Team-Agents. To the best of our knowledge, there is no system that integrates at least three of the above categories into one system. In the following section, we evaluate current collaborative systems categories for their usage in virtual teams and summarize their shortcomings with respect to the requirement for virtual teamwork.

Evaluation of Collaborative Systems for Virtual Teamwork

Cooperative tasks in virtual teams are increasing, and, as a consequence, the use of collaborative systems is becoming more pervasive. In recent years, it has increasingly become difficult to categorize systems according to the frameworks discussed previously, due to the increasing fuzziness of systems boundaries and to recent requirements for virtual teamwork. Traditional systems in the area of interest to virtual teamwork are groupware, project management (PM) and workflow management systems (WfMS). These system categories are based on different metaphors. Groupware systems mainly can be categorized along two lines (metaphors)—the *communications* or the *workspace* metaphor.

Communications-oriented groupware supports unstructured work activities using communications as the underlying interaction pattern. One very popular instance of communications-oriented groupware is e-mail. When e-mail is used as the main medium for virtual teams (as in most cases), data and associated information (e.g., attachments) remain on central mail servers and/or personal inboxes without any *context* information in which those e-mail communications were used (i.e., involved business processes, performed activities, created artifacts). Enterprise groupware systems generally focus on enterprise-wide messaging and discussion databases and do not support organizational components and structures, such as people and their associated roles, groups, tasks, and skills. This leads to "organizationally unaware" systems that treat all messages alike (semantically) and without any awareness of underlying business processes that are essential for efficient collaboration in project teams.

Workspace-oriented groupware, on the other hand, allows team members to upload or download artifacts using files and folders to organize their work. Groupware, as previously indicated, usually does not implement an underlying organizational model (i.e., providing information on the structure of a team, such as team members and their roles, skills, tasks, and responsibilities). The lack of explicit organizational structuring is a disadvantage and an advantage at the same time. It is disadvantageous because traditional groupware has no "hooks" for integrating business process information, which is important in order to integrate artifacts, resources, and processes. This will be discussed in more depth in the next section. The advantage of the lack of explicit organizational structure information is that these systems may be used in all organizational settings without much prior configuration efforts, and they lead to increased personal flexibility, as the proliferation of e-mail systems in teamwork demonstrates.

The second category, which we will briefly investigate in this section, is *project management systems*. As we have stated, virtual teamwork is, in most cases, organized as project work. Projects have well defined goals and are defined by their *begin* and *end* dates, as well as by the required resources and their tasks (work breakdown structure). It is interesting to note, however, that PM systems traditionally support the work of the project manager as the main (and sometimes the only) user of the PM system. They do not support dynamic interaction (instantiation) of processes. More recently, project management systems combine with information sharing tools (shared workspaces) to provide a persistent storage for artifacts. The enactment of the task by team members, as defined by the project manager, is not supported by PM systems. In other words, we can conclude that PM systems are not geared towards virtual teamwork, but focused more on the planning aspect. They provide "static" snapshots (usually in the form of GANNT charts) of projects and how they "should" be. There is no support for the work activities performed by the virtual team members.

The purpose of *workflow management systems* is to support the notion of processes within and, in some cases, between organizations (Aalst & Kumar, 2001; Bolcer, 2000; Bussler, 1999). However, WfMS' first requirement is to model a business process (build time) and then to enact this model (run time). This leads to substantial inflexibility for virtual teams (Ellis, 1995). In business, "exceptions are the rule;" therefore, modeling a process (project) is often not possible for creative, innovative virtual teams of knowledge workers such as in product development or consulting teams. A business process can be unstructured (ad hoc), semi-structured, or highly structured (modeled). For example, a business process such as "customer order entry" can be modeled using traditional WfMS. However, highly structured processes only can be enacted (instantiated) as they were designed. If an exception occurs, a workflow administrator needs to remodel the process before the execution can continue. This limits the

usability of WfMS in a world where constant adaptation to new situations is necessary and where teams are increasingly mobile and distributed. An example of an *ad hoc* process is discussion of a project's design review using Groupware. A semi-structured process consists of groups of activities that are modelled; however, in contrast to a structured (modelled) process, it may also consist of activities that are not predefined. A process is semi-structured when there might be one or more activities between already modeled activities such as *assign process*, which are not known beforehand and therefore cannot be modeled in advance.

It is important to note that requirements for virtual teamwork do not follow the traditional boundaries of systems already presented. We differentiate between synchronous and asynchronous technologies for teamwork support. During our case study requirements analysis, we came to the conclusion that distributed product development in virtual communities requires a blend of synchronous and asynchronous systems support for communications, as well a basic support for asynchronous coordination of team members and their activities. In summary, the requirements for virtual teams cannot be met simply by using a combination of traditional synchronous and asynchronous systems, since the criteria for successful systems in this area differ substantially with traditional "enterprise information systems." We identified and implemented (see Section 4) four fundamental feature sets for our virtual team software—device independence; process-awareness; integration of artifacts, resources, and processes; and organizational awareness. Most systems on the market do not cater to the requirements of virtual teams; namely, *dynamic views of relationships* between artifacts, resources, and process awareness are vital to the work organization of virtual teamwork.

ON THE INTEGRATION OF ARTEFACTS, RESOURCES, AND PROCESSES

Organizations increasingly define the work activities to be fulfilled in virtual teams where team members from within the organization cooperate (i.e., communicate and coordinate work activities) with outside experts, and therefore form virtual teams, which in many cases operate as geographically dispersed teams. In fact, team members work on business processes; however, in many instances, team members view their own work as a project and not necessarily as part of a larger business process fulfilling a business goal in a larger context. The work of virtual team members often results in artifacts (e.g., documents) that need to be shared among virtual team members. The underlying assumption of this chapter is that *process-awareness* is increasingly important to virtual teams. Teamwork is a fundamental property of many business processes. Business processes have well defined inputs and outputs and serve a meaningful purpose, either within or between organizations. Business processes in general and their corresponding workflows, in particular, exist as logical models (e.g., weighted directed graphs). When business process models are executed, they have specific instances. A business process consists of a sequence of work activities. An activity is a distinct process step and may be performed either by a human agent or by a machine (software). A workflow management system enacts the real world business process for each process instance (Craven & Mahling, 1995; Dayal et al., 2001; Schal, 1996). Any activity may consist of one or more tasks. A set of tasks to be worked on by a user (human agent or machine) is called *work list*. The work list itself is managed by the WfMS. The WfMC (WfMC, 1995) calls the individual task on the work list a work item. Software systems for workflow management—Groupware—process modeling (Puustjärvi & Laine, 2001), and project management has been used to automate or to augment business processes in organizations (Casati et al., 2001; Hausleitner & Dustdar, 1999). Workflow management systems have been defined as "technology based systems that define, manage, and execute workflow processes through the execution of software whose order of execution is driven by a computer representation of the workflow process logic" (WfMC, 1995). Workflow systems generally aim at helping organizations' team members to communicate, coordinate, and collaborate effectively, as well as efficiently. Therefore, WfMS possess temporal aspects such as activity sequencing, deadlines, routing conditions, and schedules. WfMS are typically "organizationally aware" because they contain an explicit representation of organizational processes (process model). However, traditional WfMS present a rigid work environment consisting of *roles* and their associated *activities* and *applications*. In this context, they do not provide support for virtual teams such as frequently changing process participants, ad hoc formation of groups collaborating on a business process, and device independent support of group activities. Unfortunately, today's WfMS assume that each *work item* is executed by a *single* worker (Aalst & Kumar, 2001). Most WfMS focus on automating structured (modeled) intra-organizational business processes. Groupware, on the other hand, typically does not contain any knowledge or representation of the *goals* or underlying business *processes* of the group. We argue that, considering the top three problems occurring in virtual teamwork, increasing contextual information in the form of building relationships between artifacts, resources, and business processes solves the fundamental problems and, as an implication, the most dominant problems such as "difficulties in team communications" and "unclear work activities." Our approach for integration of artifacts, resources, and processes comprises a communications and coordination building block where team members exchange "enriched" messages. Workflow

research has shown that modeling organizational structures has substantial benefits for business processes. Therefore, we allow modeling of organizational constructs such as groups, roles, skills, and organizational units. Each team member can be associated with those constructs, as shown in Section 4. Furthermore, an integrated database allows for attaching database objects to the communications and coordination activities of virtual team members, enabling integration of resources (organizational constructs) and artifacts. The process modeling component allows the creation of directed graphs consisting of tasks and their relationships with organizational constructs. The next section, therefore, discusses implementation issues on how to make context information (i.e., information about process instances), the team configuration (i.e., participants and their roles), their associated artifacts, and connectivity modes of group members (fixed, mobile, or ad hoc) accessible to all virtual team members.

THE CASE OF AN INTEGRATED INTERACTION MANAGEMENT SYSTEM FOR VIRTUAL TEAMS

In the following section, we will provide an overview of integration issues with which we are concerned, and design an integrated system for virtual teams called *Caramba* (Caramba Labs, 2002). An in-depth presentation of the conceptual foundations, the architecture, or the components is beyond the scope and focus of this chapter and can be found in Dustdar (2002b, 2002c, 2004). The Caramba software architecture is composed of multiple layers—middleware, client suite, and persistence store. Objects and services are accessed through the Transparent Access Layer (TAL) from the CarambaSpace platform (middleware). Depending on access mechanisms and the requested services (e.g., via Java client with RMI protocol or via Web browser with http), Caramba provides a unique way to handle requests using a metamodel framework to describe content and separate presentation, logic, and data. This model permits high flexibility and enables customization and extensions, as well as adopts new devices or technologies. The goal of this layer is to offer transparent access to a CarambaSpace. The TAL utilizes various services to transform, describe, manipulate, and observe objects. All objects managed through a CarambaSpace are described well using a metamodel description framework. Objects can be customized in their structure (e.g., adding columns to tables, adding relations to objects) and in their presentation by adopting their metamodel description. Any changes are dynamically reflected by client components. Based on the metamodel description framework Caramba enables various options to customize data and content and to integrate data from different resources (e.g., corporate databases). This layer also provides facilities for fine-grained object notification services and the implementation of customized services based on object observers. The middleware does not manage states and persistence of objects. Objects are stored, manipulated, and retrieved via the Persistence Layer (PEL). Caramba leverages and adopts standard Java-based technologies (e.g., JDBC, JNDI, HTTP, etc.) to access and integrate data.

An overall conceptual overview of how Caramba implements the requirements and how a work scenario of virtual teamwork may look is depicted in Figure 1. Virtual teams have one or more project managers and several resources (people) with various skill sets and professional backgrounds, as well as possibly different organizational affiliations. The daily teamwork consists of meetings, exchange of documents, and many communications (tasks, e-mails) being sent back and forth. For each project (business process), meetings, documents, and communications occur, and the trail of communications and interactions is depicted as lines between the team members. Without appropriate virtual team software, the relationship between artifacts, resources, and business processes is only available in the "heads" of the team members. For example, each team member has to remember *when* a particular document was sent to *whom* (e.g., a customer) and *why* (i.e., as a part of a particular business process). The goal of virtual team software should be to explicitly provide this relationship information to all team members based on their views and interests.

In order to provide one example of what an implementation looks like, we present the Caramba components. The ObjectCenter component provides mechanisms to link activities with artifacts. Based on a metamodel, Caramba provides a set of organizational objects: Persons, Roles, Groups, Skills, Units, Organization, Tasks, and Documents (i.e., Templates). Utilizing these organizational constructs, an administrator can model any organizational structure, such as hierarchical, flat, or matrix. Each object class consists of attributes describing the object. The object class *Persons* contains attributes about the person, such as name, address, and so forth. The object class *Roles* allows definition of organizational roles such as "Head of IT." The object class *Groups* defines project settings such as "Product Team IT-Solutions." *Skills* enables the definition of required skill sets such as "Certified Java Developer." *Units* describes permanent departments such as "Marketing." The Object Center provides a means (by drag and drop) to link rows of object classes with each other, as depicted in Figure 2. It allows users to view relationships between who (organizational constructs) is performing which activities (Tasks) and using what (Documents). A business process modeller component enables a project manager to model a process template, which may be instantiated later using the built-in Workflow engine. Exceptions to the model are possible, without the need to remodel the process template, by choosing the communications (coordination)

Figure 1. Conceptual view on virtual team software support

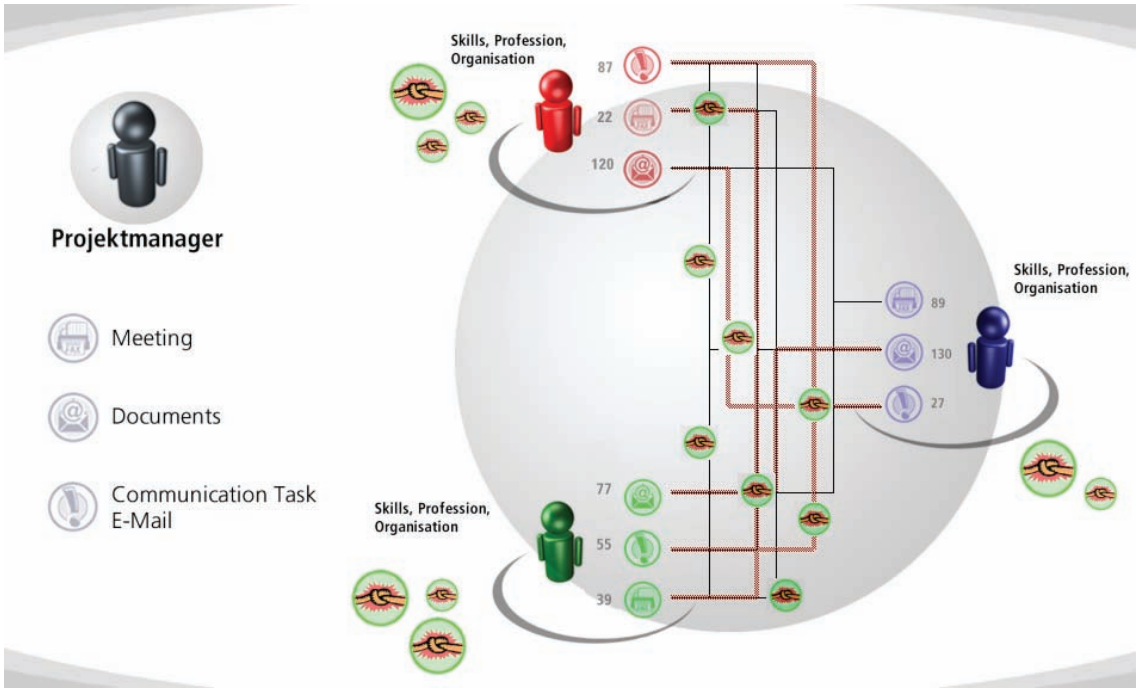
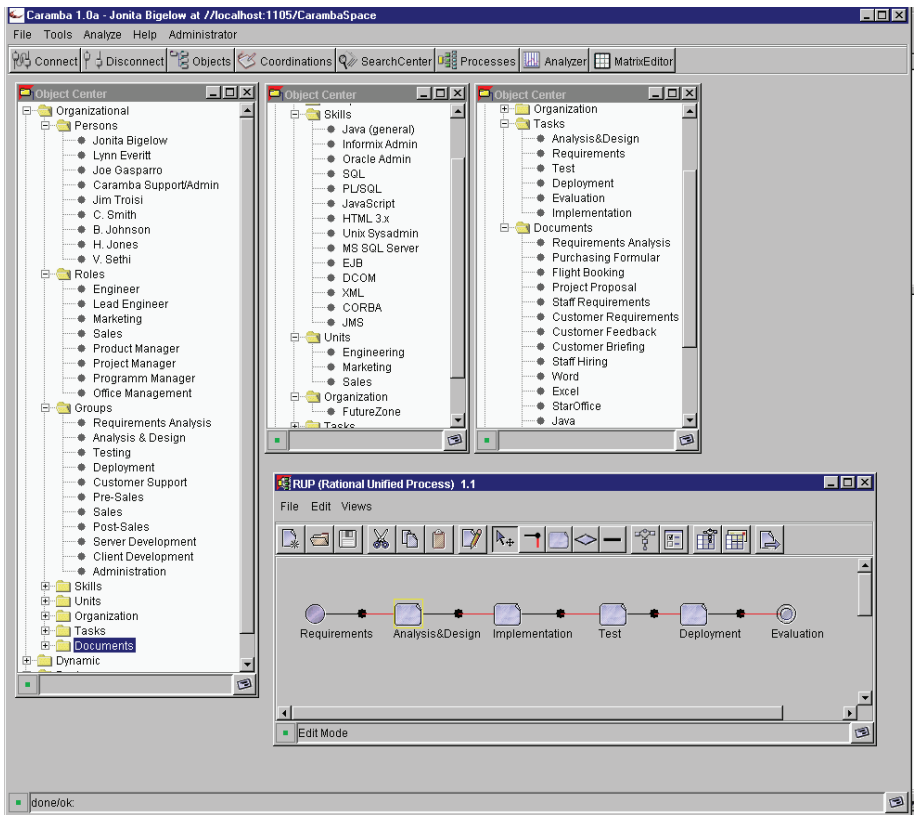


Figure 2. Modelling organizational resources and processes



partner (from the ObjectCenter). The receiving end can read the appropriate message in his or her inbox.

CONCLUSION

During the last few years, virtually all business processes changed regarding their requirements for flexibility, interconnectivity, and coordination styles. Most business processes are based on teamwork. Most teams are organized as virtual teams, with team members coming from different organizations. In this chapter, we discussed the requirements of modern virtual teamwork and the problems associated with using traditional groupware, project, and workflow management systems for virtual teamwork. A fundamental need for distributed virtual teamwork is to have access to contextual information (i.e., to see a "knowledge trail" of who did what, when, how, and why). We presented the underlying conceptual issues and one implemented information system (Caramba) to support the integration of artifacts, resources, and business processes for virtual teams. Future virtual team systems should provide mechanisms for the integration of organizational models with artifacts and business processes in loosely coupled information systems. In our future work, we plan to design and implement support for definition, configuration, and composition of processes for virtual teams based on Web services. A Web service is an interface that describes a collection of operations that are network accessible through standardized XML messaging using Web servers or Application servers. A Web service is described using a standard, formal XML notion, called its *service description*. It can be published and found by other Web services. To summarize our recommendations and lessons learned, we think that for typical mid-size (e.g., 15-person) virtual teams (geographically dispersed), process-awareness, organizational awareness, and the integration of artifacts, resources, and processes is crucial. In most cases, we found that asynchronous systems support is of paramount importance when there are more team members in a virtual team and when more work occurs across different time zones.

ACKNOWLEDGMENT

The author thanks all team members of Caramba Labs Software AG for the fruitful and constructive discussions.

REFERENCES

Aalst, W.M.P., & Kumar, A. (2001). A reference model for team-enabled workflow management systems. *Data & Knowledge Engineering, 38*, 335-363.

Akademie für Führungskräfte (2002). Probleme bei der teamarbeit [Report]. Germany.

Bafoutsou, G., & Mentzsa, G. (2002). Review and functional classification of collaborative systems. *International Journal of Information Management, 22,* 281-305.

Bentley, R. et al. (1997). Basic support for cooperative work on the World Wide Web. *International Journal of Human-Computer Studies, 46*, 827-846.

Bolcer, G.A. (2000, May and June). Magi: An architecture for mobile and disconnected workflow. *IEEE Internet Computing*, 46-54.

Bussler, C. (1999). Enterprise-wide workflow management. *IEEE Concurrency, 7*(3), 32-43.

Caramba Labs Software AG (2002). Retrieved January 15, 2002, from http://www.CarambaLabs.com

Casati, F. et al. (2001). Developing e-services for composing e-services. In *Proceedings CaiSE 2001.* Springer Verlag.

Craven, N. &Mahling, D.E. (1995). Goals and processes: A task basis for projects and workflows. In *Proceedings COOCS International Conference*. Milpitas, CA.

Dayal, U., Hsu M., & Ladin R. (2001). Business process coordination: State of the art, trends, and open issues. *Proceedings of the 27th VLDB Confererence*. Rome, Italy.

DeSanctis, G., & Gallupe, R.B. (1987). A foundation study of group decision support systems. *Management Science, 23*(5), 589-609.

Dustdar, S. (2002a). Mobility of context for project teams. *Proceedings of the International Workshop on Mobile Teamwork at the 22nd International Conference on Distributed Computing Systems*.

Dustdar, S. (2002b). Collaborative knowledge flow – Improving process-awareness and traceability of work activities. *4th International Conference on Practical Aspects of Knowledge Management*.

Dustdar, S. (2002c). Reconciling knowledge management and workflow management: The activity-based knowledge management approach. In H. Nemati, P. Palvia, & R. Ajami (Eds.), *Global Knowledge Management: challenges and opportunities*. Hershey, PA: Idea Group Publishing.

Dustdar, S. (2004, January). Caramba – A process-aware collaboration system supporting ad hoc and collaborative processes in virtual teams. *Distributed and Parallel Databases, 15*(1), 45-66.

Johansen, R. (1988). Groupware. Computer-support for business teams. *The Free Press*. New York.

Ellis, C.A. et al. (1995). Dynamic change within workflow systems. *Proceedings of COOCS International Conference.* Milpitas, CA.

Ellis, C.A. (2000). An evaluation framework for collaborative systems [Report]. University of Colorado at Boulder Technical Report CU-CS-9001-00.

Ellis, C.A., Gibbs, S.J., & Rein, G.L. (1991). Groupware: Some issues and experiences. *Communications of the ACM, 34*(1).

Groove (2002). http://www.groove.net

IBM (2002). http://www.ibm.com

Puustjärvi, J., & Laine, H. (2001). Supporting cooperative inter-organizational business transactions [Lecture]. *Proceedings of DEXA 2001.* Springer Verlag.

Schal, T. (1996). *Workflow management systems for process organizations*. New York: Springer.

Workflow management specification (1995). *Workflow Management Coalition*. Retrieved January 15, 2002, from http://www.wfmc.org/standards/docs/tc003v11.pdf

Workflow management specification glossary (1995). *Workflow Management Coalition (WfMC)*. Retrieved from http://www.wfmc.org

KEY TERMS

Knowledge Trail: Provides information on who did what, when, how, and why.

Interaction Management System: An information system providing an environment for communications and coordination of work activities for virtual teams.

Process: Indicates what tasks must be performed and in what order to complete a case.

Process-Awareness: (See knowledge trail)

Role: In order to perform tasks, skills are required. A role is a collection of complementary skills.

Task: An atomic process that is not divided further and is a logical unit of work.

Workflow: Comprises cases, resources, and triggers that relate to a particular process.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2314-2320, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Process-Based Data Mining

Karim K. Hirji
AGF Management Ltd, Canada

INTRODUCTION

In contrast to the Industrial Revolution, the Digital Revolution is happening much more quickly. For example, in 1946, the world's first programmable computer, the Electronic Numerical Integrator and Computer (ENIAC), stood 10 feet tall, stretched 150 feet wide, cost millions of dollars, and could execute up to 5,000 operations per second. Twenty-five years later, Intel packed 12 times ENIAC's processing power into a 12–square-millimeter chip. Today's personal computers with Pentium processors perform in excess of 400 million instructions per second. Database systems, a subfield of computer science, has also met with notable accelerated advances. A major strength of database systems is their ability to store volumes of complex, hierarchical, heterogeneous, and time-variant data and to provide rapid access to information while correctly capturing and reflecting database updates.

Together with the advances in database systems, our relationship with data has evolved from the prerelational and relational period to the data-warehouse period. Today, we are in the knowledge-discovery and data-mining (KDDM) period where the emphasis is not so much on identifying ways to store data or on consolidating and aggregating data to provide a single, unified perspective. Rather, the emphasis of KDDM is on sifting through large volumes of historical data for new and valuable information that will lead to competitive advantage. The evolution to KDDM is natural since our capabilities to produce, collect, and store information have grown exponentially. Debit cards, electronic banking, e-commerce transactions, the widespread introduction of bar codes for commercial products, and advances in both mobile technology and remote sensing data-capture devices have all contributed to the mountains of data stored in business, government, and academic databases. Traditional analytical techniques, especially standard query and reporting and online analytical processing, are ineffective in situations involving large amounts of data and where the exact nature of information one wishes to extract is uncertain.

Data mining has thus emerged as a class of analytical techniques that go beyond statistics and that aim at examining large quantities of data; data mining is clearly relevant for the current KDDM period. According to Hirji (2001), data mining is the analysis and nontrivial extraction of data from databases for the purpose of discovering new and valuable information, in the form of patterns and rules, from relationships between data elements. Data mining is receiving widespread attention in the academic and public press literature (Berry & Linoff, 2000; Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Kohavi, Rothleder, & Simoudis, 2002; Newton, Kendziorski, Richmond, & Blattner, 2001; Venter, Adams, & Myers, 2001; Zhang, Wang, Ravindranathan, & Miles, 2002), and case studies and anecdotal evidence to date suggest that organizations are increasingly investigating the potential of data-mining technology to deliver competitive advantage.

As a multidisciplinary field, data mining draws from many diverse areas such as artificial intelligence, database theory, data visualization, marketing, mathematics, operations research, pattern recognition, and statistics. Research into data mining has thus far focused on developing new algorithms and tools (Dehaspe & Toivonen, 1999; Deutsch, 2003; Jiang, Pei, & Zhang, 2003; Lee, Stolfo, & Mok, 2000; Washio & Motoda, 2003) and on identifying future application areas (Alizadeh et al., 2000; Li, Li, Zhu, & Ogihara, 2002; Page & Craven, 2003; Spangler, May, & Vargas, 1999). As a relatively new field of study, it is not surprising that data-mining research is not equally well developed in all areas. To date, no theory-based process model of data mining has emerged. The lack of a formal process model to guide the data-mining effort as well as identification of relevant factors that contribute to effectiveness is becoming more critical as data-mining interest and deployment intensifies. The emphasis of this article is to present a process for executing data-mining projects.

BACKGROUND

The fields of machine learning, pattern recognition, and statistics have formed the basis for much of the developments in data-mining algorithms. The field of statistics is one of the oldest disciplines concerned with automatically finding structure in examples. Discriminant analysis (Fisher, 1936), for example, is the oldest mathematical classification technique used to separate data into classes by generating lines, planes, or hyperplanes. Through the pioneering work on classification and regression trees (CART) by Breiman, Friedman, Olshen, and Stone (1984), the statistical community has made an important contribution in legitimizing

the use of decision trees, in data mining, for classification and regression. Pattern-recognition research emphasizes the creation of machines that can perform tasks more accurately, faster, and cheaper than humans (Fukunaga, 1972; Ripley, 1993), and has made an important contribution to data mining by popularizing the use of neural networks. A feed-forward neural network is a network in which the nodes (or processing units) are numbered so that all connections go from a node to one with a higher number. In practice, the nodes are arranged in layers with connections only to higher layers. Back propagation is an implementation for a feed-forward neural network in which error terms, from the output layer, are propagated back to the input layer so that the resulting connection weights at each node adjusted can be adjusted by means of an error-minimization method called gradient descent.

The multitude of data-mining algorithms can be linked to three main data-mining-problem approaches: clustering, association and sequential pattern discovery, and predictive modeling. Clustering (or segmentation) is concerned with partitioning data records into subsets. The *K*-means clustering algorithm is used for demographic clustering because categorical data are predominant. This algorithm, which is efficient for large databases, clusters a data set by determining the cluster to which a record fits best. Once clusters have been found in a data set, they can be used to classify new data. To uncover affinities among transaction records consisting of several variables, association algorithms are used. These algorithms are used to solve problems where it is important to understand the extent to which the presence of some variables implies the existence of other variables and the prevalence of this particular pattern across all data records. Sequential-pattern-discovery algorithms are related to association algorithms except that the related items are spread over time. Finally, the predictive-modeling data-mining-problem approach involves the use of a number of algorithms (e.g., binary decision tree, linear discriminant function analysis, radial basis function, back-propagation neural network, logistic regression, and standard linear regression) to classify data into one of several predefined categorical classes or to use selected fields from historical data to predict target fields.

The initial implementation of data-mining applications has been in the banking, consumer marketing, insurance, and telecommunications industries. Credit scoring, direct-mail target marketing, policy-holder risk assessment, and call graph analysis are but a few of the "killer" applications of data mining in these respective industries. As a result of some of the realized benefits of data mining, new applications are emerging in a number of areas including biomedicine where molecular data are combined with clinical medical data to achieve a deeper understanding of the causes for and treatment of disease, national security where unusual patterns and fraudulent behavior play a role in identifying and tracking activities that undermine security, pharmaceuticals where interest in understanding the 3D substructure of a molecule and how it interacts with the target is a crucial step in the design of new drug molecules, and ecology where large amounts of climate data, terrestrial observations, and ecosystem models offer an unprecedented opportunity for predicting and possibly preventing future ecological problems. Although the frontiers of data-mining applications continue to expand, focus on developing a data-mining process has not met with similar enthusiasm.

DATA-MINING PROCESS OVERVIEW

New product development (NPD) is a well-researched area (e.g., Hauptman & Hirji, 1999) and, thus, it is the foundation for the data-mining process model because NPD projects, by their very nature, are the most complex as they include systems, subsystems, components, and modules, as well as physical product and software aspects. Focusing on the NPD literature and synthesizing the elements of the various process models allows for the development of an information-centric process model for performing data-mining projects. Table 1 provides a baseline of what an inclusive process for performing data-mining projects might look like.

The phases in the baseline data-mining process include Phase 0, Phase 1, Phase 2, and Phase 3. Phase 0 is the *discovery* phase that supports the subsequent three phases. The set of proposed activities in this phase include (a) assessment of the organization's orientation toward data-centricity, (b) assessment of the capability of the organization to apply a portfolio of analytical techniques, and (c) strategy development for the use of analytics throughout the department or organization. Phase 1 is the *entry* phase. The underlying intent of this phase is to define the candidate business problem that is solvable and that can at least partially use existing data resident in the organization's databases. Prospecting and domain analysis, business problem generation and preliminary assessment, and data sensing are the proposed set of activities in this phase. Data sensing in particular is concerned with the representational faithfulness of the data set in question. Phase 2 is the *launch* phase. In this phase the data-mining project becomes a formal project with an associated capital and operational budget. The set of proposed activities in this phase include (a) secure project sponsorship, (b) project planning and project-team resourcing, (c) business problem refinement and validation of business assumptions and constraints, (d) development of the data strategy (i.e., explore the possible need for purchasing data from a third-party provider), and (e) formulation of the data-mining approach. Phase 3, the final phase, is the *execution and infusion* phase. The actual execution of data-mining algorithms takes place here as well as results analysis. The proposed set of activities in this final phase are (a) detailed

business-problem definition, (b) data sourcing and enrichment, (c) execution of data-mining algorithms, (d) results interpretation, synthesis, and validation, and (e) information harvesting and business-strategy formulation. This final activity focuses on developing a strategy to tactically leverage the new insight and to communicate this on a department-, division-, or perhaps enterprise-wide basis.

APPLICATION IN BUSINESS

The 1990s are referred to as the "heyday" of the mutual-fund industry in North America. Falling yields from fixed-rate investments, new product creation, and the hope of higher returns drove investors to purchase mutual funds. It is not surprising that mutual-fund companies therefore experienced double-digit growth rates in the range of 25% to 65%. The picture today is vastly different as industry consolidation, a mature business cycle, and flat market growth are causing gross sales to increase at a decreasing rate. Retail investors are now more than ever before scrutinizing management expense ratios (MERs) and fund performance reporting and returns. Competition among the various fund companies is fierce and thus there is a renewed focus on both stealing market share and developing a compelling story about future sales growth rates across multiple product lines. Some of the tactics employed to achieve various business goals are product innovation and quality, superior fund performance, and exceptional sales, marketing, and client service.

The experiences to date of the application of the proposed data-mining process to an analytics project in the mutual-fund industry suggests that the model is playing a role in contributing to a successful outcome. The project is in the early aspects of Phase 3 and therefore because it is ongoing and no actual positive net present value has been realized, conclusions cannot be made at this time. However, some of the key accomplishments that can be attributed to the process model are as follow.

- Project has been formally approved as a result of the capital-budgeting process
- Project team and stakeholders are in place
- Business problem related to effective channel management is understood and defined
- Business entity and data models have been developed
- Strategies to close data gaps and augment data where necessary are in place
- Data-mining approach with an emphasis on clustering exists

FUTURE TRENDS

The need for near-real-time information to support business decision making is becoming more and more critical. Data are benign and therefore there is more focus on understanding the "story" behind the data and the relationships among them. From the vantage point of practitioners and managers, there appear to be new patterns emerging specifically in three areas. The software and enterprise-wide-package-implementation industry is experiencing major consolidation, and software companies are realigning their sales efforts to focus on the ever-growing mid-market segment. Once this shakedown is complete, many business intelligence software and solution

Table 1. A summary of a data–mining process

Phase Name	Phase Identifier	Key Activity
Discovery	0	• Assess data centricity • Assess analytics capability • Develop analytics strategy
Entry	1	• Prospecting and domain analysis • Problem generation • Problem assessment • Data sensing
Launch	2	• Project sponsorship • Project planning and core project team • Problem refinement • Problem validation • Data strategy • Data-mining approach
Execution & Infusion	3	• Business-problem definition • Data sourcing and enrichment • Run data-mining algorithms • Results interpretation, synthesis, and validation • Information harvesting • Business-strategy formulation

providers will begin to bundle data-mining solutions with their existing offerings as a response to the already-established demand by customers in this market segment.

The data-mining process today is more akin to craft than interdisciplinary team-based problem solving. Once data mining gains acceptance and has a critical mass in organizations, there will undoubtedly be a shift to using this technology as part of a repertoire of tools to assist planners, product developers, managers, and others to develop strategies and tactical implementation plans for the organization. In this respect, improvement to existing tools in the areas of human interaction and visual presentation of information are expected. Finally, with the realities of virtual teams now solidified, the question of how virtual teams can effectively execute data-mining implementation projects will become relevant.

CONCLUSION

The paramount objective of publicly traded organizations is to focus on creating and maximizing shareholder wealth. However, recently well-publicized corporate shenanigans have shown that single-mindedly pursuing largesse at the expense of corporate social responsibility is not only fundamentally wrong, but also something that cannot be sustained indefinitely. Managers, C-suite executives, and corporate directors are now once again facing increasing scrutiny about revenue, profit, and earnings quality; adoption of accounting rules; and timely disclosures regarding the going concern of business entities. As managers craft, implement, and execute various growth, talent management, cost reduction, and competitive differentiation strategies, information technology will without a doubt continue to play an important role in enabling, supporting, leading, and transforming business, government, and not-for-profit organizations. Data mining as a technology is not an end, but rather a means to an end. Through the use of a disciplined and structured process, data-mining benefits can be obtained by the operationalization of data-mining results, via a business strategy, to achieve specific business-unit and enterprise-wide objectives.

REFERENCES

Alizadeh, A. A., et al. (2000). Distinct types of diffused large B-cell lymphoma identified by gene expression profiling. *Nature, 403*, 503-511.

Berry, M., & Linoff, G. (2000). *Mastering data mining*. New York: John Wiley & Sons, Inc.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* CA: Wadsworth & Brooks.

Dehaspe, L., & Toivonen, H. (1999). Discovery of frequent datalog patterns. *Data Mining and Knowledge Discovery, 3*(1), 7-36.

Deutsch, J. M. (2003). Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics, 19*, 45-54.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*(11), 27-34.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

Fukunaga, K. (1972). *Introduction to statistical pattern recognition.* Boston: Academic Press.

Hauptman, O., & Hirji, K. K. (1999). Managing integration and coordination in cross-functional teams: An international study of concurrent engineering product development. *R&D Management, 29*(2), 179-192.

Hirji, K. K. (2001). Exploring data mining implementation. *Communications of the ACM, 44*(7), 87-93.

Jiang, D., Pei, J., & Zhang, A. (2003). Toward interactive exploration of gene expression patterns. *SIGKDD Explorations, 5*(2), 79-90.

Kohavi, R., Rothleder, N. J., & Simoudis, E. (2002). Emerging trends in business analytics. *Communications of the ACM, 45*(8), 45-48.

Lee, W., Stolfo, S., & Mok, K. (2000). Adaptive intrusion detection. *Artificial Intelligence Review, 14*, 533-567.

Li, T., Li, Q., Zhu, S., & Ogihara, M. (2002). A survey of wavelet applications in data mining. *SIGKDD Explorations, 4*(2), 49-68.

Newton, M., Kendziorski, C., Richmond, C., & Blattner, F. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology, 8*, 37-52.

Page, D., & Craven, M. (2003). Biological applications of multi-relational data mining. *SIGKDD Explorations, 5*(1), 69-79.

Ripley, B. D. (1993). Statistical aspects of neural networks. In O. E. Barndorff-Nielsen et al. (Eds.), *Networks and*

chaos: Statistical and probability aspects. London: Chapman & Hall.

Spangler, W., May, J., & Vargas, L. (1999). Choosing data mining methods for multiple classification: Representational and performance measurement implications for decision support. *Journal of Management Information Systems, 16*(1), 37-62.

Venter, J., Adams, M., & Myers, E. (2001). The sequence of the human genome. *Science, 291*, 1304-1351.

Washio, T., & Motoda, H. (2003). State of the art of graph-based data mining. *SIGKDD Explorations, 5*(1), 59-68.

Zhang, L., Wang, L., Ravindranathan, A., & Miles, M. (2002). A new algorithm for analysis of oligonucleotide arrays: Application to expression profiling in mouse brain regions. *Journal of Molecular Biology, 317*, 227-235.

KEY TERMS

Classification Trees: Type of decision tree that is used to predict categorical variables, whereas regression trees are decision trees used to predict continuous variables.

Cluster: Subset of data records; the goal of clustering is to partition a database into clusters of similar records such that records that share a number of properties are considered to be homogeneous.

Data Mart: Scaled-down version of an enterprise-wide data warehouse that is created for the purpose of supporting the analytical requirements of a specific business segment or department.

Data Mining: Analysis and nontrivial extraction of data from databases for the purpose of discovering new and valuable information, in the form of patterns and rules, from relationships between data elements.

Data Warehouse: A platform consisting of a repository of selected information drawn from remote databases or other information sources, which forms the infrastructural basis for supporting business decision making.

Information: Interpreted symbols and symbol structures that reduce both uncertainty and equivocality over a defined period of time.

Knowledge: Information combined with experience, context, interpretation, and reflection.

Operational Data Store: An integrated repository of transaction-processing systems that uses data-warehouse concepts to provide "clean" data in support of day-to-day operations of a business.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2321-2325, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Project Management and Graduate Education

Daniel Brandon, Jr.
Christian Brothers University, USA

INTRODUCTION

Project Management is "the application of knowledge, skills, tools, and techniques to the project activities in order to meet or exceed stakeholder needs and expectations from a project" (Duncan, 1996). A project is defined as "a temporary endeavor undertaken to create a unique product or service" (Duncan, 1996). This article provides an overview of the coverage of the project management discipline in academic graduate education.

BACKGROUND

A number of professional organizations have developed around the world to address and foster this specific discipline. Most notable is the Project Management Institute (PMI, www.pmi.org) with about 100,000 members worldwide. Other major international organizations are the Association for Project Management (APM) and the International Project Management Association (IPMA) (Morris, 2001). These organizations have recognized there is a distinct skill set necessary for successful project managers, and the organizations are devoted to assisting their members develop, improve, and keep current these skills (Boyatzis, 1982; Caupin, Knopfel & Morris, 1998).

Several universities have also recognized the fact that project management involves distinct skills, and that the traditional degree programs and courses in both business schools and other schools do not adequately cover and/or integrate these skills. The *Chronicle of Higher Education* recently reported that seven Philadelphia-area corporations established ties with four universities in that region to improve the business skills of computer science and IT students; most of these key skills involved the project management skill sets, which are specifically identified later in this document (*Chronicles of Higher Education*, 2001).

Perhaps self-evident from the previous paragraph is the fact that the knowledge and training needed by project managers covers both traditional business disciplines and disciplines involved with building or making things. Often the skills involved with building or making things would be found in an engineering curriculum, and also in information technology or computer science curriculums.

Since the skill sets needed by project managers are extensive, and since these skills involve both business and engineering disciplines, and also since most candidate students are degreed working adults, most schools have developed their project management curriculums as graduate school programs. A number of universities also have a single "project management" course offered as a graduate or undergraduate course.

TYPES OF GRADUATE DEGREE PROGRAMS

An analysis of universities currently offering graduate project management programs indicates several types of programs being offered:

1. A master's level general degree program (such as an MBA) with a specialization in Project Management;
2. A full masters level (generally MS) program in project management; and
3. A "certification program" of several project management courses.

Some universities offer more than one of these program types. Also in some universities the program is offered in the School of Business (or Management) and in some schools the program is offered in the School of Engineering. In most universities, many of the courses appeared to be shared with other graduate degree programs; in other words, not all of the courses in the program are focused on project management.

PMI (and the other international project management organizations) have a certification program, and for PMI the designation is "Project Management Professional" (PMP). To obtain PMP certification, an individual must have 4,500 hours of documented project management experience over a period of six years, have a BS level college degree, and pass a rigorous 4-hour examination. The first PMP exam was given in 1984 to about 30 people, and today there are over 30,000 PMPs worldwide (Foti, 2001). Once the PMP status is obtained, an individual must earn 60 PDUs (Professional Development Units) every three years. Some universities offer a PMP Exam Preparation course or cover exam prep material in one of their project management courses. However, most graduate programs do not cover exam prep; in

Figure 1. Institutions offering graduate credit programs in project management

University	Organize	School	Certificate Program		MBA/MS Specialization		PM Masters Degree	
			# Courses	# PM	# Courses	# PM	# Courses	# PM
Amberton	KA	Business	4	4				
American Graduate Univ.	KA	Business					12	7
Boston University	KA	Business	8	8				
City University	KA	Business	6	6				
Colorado Technical University	Step	Both	6	6	13	6		
George Washington University	Step	Business					12	3
Int'l School of Info. Mgmt.	Step	Business	3	3	12	4		
Keller School of Management	KA	Business	6	4			14	6
Northwestern	Step	Engineering					12	4
Regis University	PG	Business			13	6		
Stevens Inst. Of Technology	Step	Business	4	4	12	4	12	6
U. of Management & Tech.	KA	Both	7	7				
U. of Wisconsin - Madison	KA	Business	6	6				
U. of Wisconsin - Platteville	Step	Business					12	5
University of Central Florida	Step	Engineering	5	1				
University of Maryland	Step + KA	Engineering					10	5
University of Texas - Dallas	Step	Business	6	1	10	1		
Western Carolina University	PG	Business					12	6
Wright State University	Step	Business			12	3		

fact, the graduate programs studied herein are more geared to providing the PDU credits for PMPs.

Figure 1 summarizes the program types for most of the U.S. universities offering project management programs "certified" by PMI. The list of such schools is on the PMI website (www.pmi.org). Out of the 19 schools listed, 11 offer a certificate program, six offer an MBA/MS specialization, and eight off a full Master's is project management. In 14 of the 19 schools, the program is entirely in the Business (or Management) school.

PROJECT MANAGEMENT KNOWLEDGE ORGANIZATION

PMI has developed an index of project management skills and knowledge called the "Project Management Body of Knowledge" (PMBOK). The PMBOK has been developed through several iterations over many years; the first version was developed in 1976 (Cook, 2004). The latest version (PMBOK, 2000) has been released (for certification testing beginning 1/2002) (PMI, 2000). It defines nine "knowledge area" which are organized into 37 "processes". The processes are grouped into five "process groups". This is illustrated in Figure 2 (for PMBOK, 1996) (Duncan, 1996).

Since so many resources have been put into the development and refinement of the PMBOK and it has been so well received by the project management community, it seemed prudent to us to organize our graduate program courses around the processes defined within PMBOK. The issue then became how do we "slice and dice" the processes as shown in Figure 2 into distinct (but integrated) courses. The PMBOK document itself organizes its write-up by knowledge area. However, most classic overall project management books and textbooks are organized by process groups (Badiru, 1989; Cleland & King, 1988; Hajek, 1984; Kerzner, 1980; Meredith & Mantel, 1989; Royce, 1988; Verzuh, 1999). There are however a number of books concerning particular parts of project management, and these cover particular knowledge areas, but they are not specifically written as "textbooks" (Fisher & Fisher, 2000; Fleming & Koppelman, 2000; Pinto & Trailer, 1999; Schuyer, 2001; Verma & Thamhain, 1996).

Looking at the universities currently offering degree programs to see how their curricula were organized, we defined three general types of organization:

1. "Step" – Courses are organized in the traditional manner from less depth to more depth over most of the knowledge areas. For example, the first course might be "Introduction to Project Management"; the next might be "Intermediate Project Management"; and the next would be "Advanced Project Management";
2. "KA" – Follows the PMBOK knowledge areas (Scope, Time, Cost, …); and
3. "PG" – Follows the PMBOK process groups (Initiation, Planning, …).

Figure 2. PMI process groups and knowledge areas

	Initiation	Planning	Executing	Controlling	Closing
Integration		Project Plan Development	Project Plan Execution	Overall Change Control	
Scope	Initiation	Scope Planning	Scope Verification	Scope Change Control	Scope Verification
		Scope Definition			
Time		Activity Definition		Schedule Control	
		Activity Sequencing			
		Activity Duration Estimation			
		Schedule Development			
Cost		Resource Planning		Cost Control	
		Cost Estimating			
		Cost Budgetting			
Quality		Quality Planning	Quality Assurance	Quality Control	
Human Resources		Organizational Planning	Staff Acquisition	Team Development	
Communications		Communications Planning	Information Distribution	Performance Reporting	Administrative Closure
Risk	Risk Identification	Risk Identification		Risk Response Control	
		Risk Quantification			
		Risk Response Development			
Procurement		Procurement Planning	Solicitation	Contract Administration	Contract Closeout
		Soliciation Planning	Source Selection		
			Contract Administration		

Most programs do not fit entirely into one of these molds, but they were categorized according to the best fit. Overall, out of the 19 schools, 10 use primarily the Step method, six use primarily the KA method, and two use the PG area.

For schools offering certification, five use the Step method, six use the KA method, and none use the PG method. For schools offering the MBA/MS specialization, none use the KA method, one uses the PG method, and the rest use the Step method. For schools offering the full MS in Project Management, two use KA's, one uses PG's, and the rest (five) use the Step method.

The issue of course material organization is a difficult one for a university. As discussed earlier, universities offering these programs are taking different approaches in this area. We feel the "Step" approach is only useful for programs that have two or three project specific courses. The "KA" approach requires much more "course preparation" time, textbooks are limited, and instructors need depth in these skills. One possible curriculum design would be to use a combination of "PG" and "KA". For "PG", separation into two process "super-groups" may be appropriate: project planning and project control; both covering scope, time, and cost. Separate "KA" courses would likely involve: procurement, risk, quality, and human resources/communications.

PROJECT MANAGEMENT CONTENT IN PROGRAMS

As can be seen from Figure 1, not all of the courses in a project management program are project management specific courses. For most schools, the certification offering is made up of mostly project management specific courses (the #PM in Figure 1 is the number of project management specific courses). For the project management specialization, most schools use three to six project management specific courses. For the full MS Project Management degree, the number of project management specific courses is about one-third to one-half of the courses. These non-specific courses in the full MS degree program vary widely from school to school especially if the degree is in the Engineering school instead of the Business school. Some of these non-project management specific courses are typically: general management, organizational behavior, leadership, managerial accounting, information technology, finance, human resources, quantitative methods, quality assurance, procurement and contracting, and risk management.

DELIVERY

Some universities are offering some, all, or portions of their courses in the form of "distance learning". So the issue be-

comes: "where on the spectrum from 'bricks to clicks' should a program position itself." There are many pros and cons on both sides of this issue, and most of those pros and cons depend on exactly how a course is made available "online" and the university's overall vision, mission, and tradition. This issue encompasses most degree programs (not just project management), so we are not going to further debate it here, except to indicate it is highly dependent on a particular school's mission, tradition, and demographics. As discussed in the following section, the potential students for such a graduate program are working adults, so attention has to be given to the best delivery for that market. Many schools are holding classes on weekends or evenings to accommodate the adult audiences for these types of programs (*San Diego Business Journal*, 2001).

PROGRAM STAKEHOLDERS AND THEIR NEEDS

We have discussed and surveyed the needs of the stakeholders of a graduate program in our region. The external stakeholders we identified were those companies and those individuals who would benefit from such a program. The companies would benefit by the introduction or reinforcement of the specific project management methodologies into their organizations; this has both an educational and training perspective.

Our individual stakeholders are primarily degreed working professionals. This is similar to the market served by the other universities we investigated, since those other universities like ourselves are located in large metropolitan areas. These individuals benefit from a "continuing education" perspective that makes them individually more valuable. Those individuals having earned PMI PMP Certification would have another way to earn PDU credits (a credit course at a university earns 15 PDUs per semester credit hour). Currency of methods and tools is also quite important to both corporations and individuals.

FUTURE TRENDS

For future university programs in project management, four dimensions can be defined. The PMI PMBOK focuses on the dimension of breadth of the knowledge areas (and the 37 processes) but intentionally does not go into much depth. Going into depth gets into method and tool specifics. Thus, a future trend in university programs would be to address not only the breadth but also the depth of these key processes.

The next dimension identified is industry particulars. While there is much commonality to project management in all industries, there is also much that is specific to each area. For example, task estimation for an IT project is much different than task estimation in a construction project. So this should be considered as another added dimension to new programs, certainly not for all industries but for the major ones in a school's geographic region.

The next dimension we identified was that of time or "currency". This not only includes the use of current tools, but the practice of project management in the current business and technical environment. Issues such as "virtual teams", international coverage, and Web-based systems would be included in this dimension.

CONCLUSION

Herein, we have examined content, approach, and logistics issues in graduate education for project management. The university programs surveyed were all relatively new programs, so there is little or no data available for a statistical or comparative historical analysis at this time. In the future, one may be able to survey graduates from the different types of programs to determine the pros and cons of each type of program organization and the "best practices" for project management education.

REFERENCES

Badiru, A.B. (1989). *Project management in manufacturing and high technical operations*. Wiley Interscience.

Boyatzis, R. (1982). *The competent manager: A model for effective performance*. John Wiley & Sons.

Caupin, G., Knopfel, H., & Morris, P. (1998). *ICB IPMA competence baseline*. Zurich: International Project Management Association.

Chronicles of Higher Education. (2001, August 10). *47*(48), A45.

Cleland, D.I., & King, W.R. (1988). *Project management handbook*. Van Nostrand Reinhold.

Cook, D.L. (2004). Certification of project managers – Fantasy or reality. *Project Management Quarterly, 8*(2), 32-34.

Duncan, W. (1996). *A guide to the project management body of knowledge.* Project Management Institute.

Fisher, K., & Fisher, M. (2000). *The distance manager: A hands on guide to managing off-site and virtual teams*. McGraw-Hill.

Fleming, Q., & Koppelman, J. (2000). *Earned value project management*. Project Management Institute.

Foti, R. (2001, September). The case for certification. *PM Network*.

Hajek, V.G. (1984). *Management of engineering projects*. McGraw-Hill.

Kerzner, H. (1980). *Project management. A systems approach to planning, scheduling, and controlling*. Van Nostrand.

Meredith, S.R., & Mantel, S.J. (1989). *Project management, A management approach*. John Wiley & Sons.

Morris, P. (2001, September). Updating the project management bodies of knowledge. *Project Management Journal*.

Pinto, J., & Trailer, J. (1999). *Essentials of project control*. Project Management Institute.

PMI. (2000). *A guide to the project management body of knowledge*. Project Management Institute.

Royce, W. (1988). *Software project management*. Addison-Wesley.

San Diego Business Journal. (2001, August 6). *22*(32), 23.

Schuyler, J. (2001). *Risk and decision analysis in projects*. Project Management Institute..

Verma, V., & Thamhain, H. (1996). *Human resource skills for the project manager*. Project Management Institute..

Verzuh, E. (1999). *Fast forward MBA in project management*. John Wiley & Sons.

KEY TERMS

APA: Association for Project Management.

IPMA: International Project Management Association.

Knowledge Area (KA): Project Management Knowledge Area; For PMI there are nine knowledge areas: Integration, Scope, Time, Cost, Quality, Human Resources, Communications, Risk, and Procurement.

Process Group (PG): A grouping of project management processes; for PMI, there are five process groups: Initiation, Planning, Executing, Controlling, Closing.

Professional Development Unit (PDU): A unit of continuing education for certified project managers.

Project Management: The application of knowledge, skills, tools, and techniques to project activities in order to meet or exceed stakeholder needs and expectations of that project.

Project Management Body of Knowledge (PMBOK): A consolidation and organization of information about project management including “best practices”.

Project Management Institute (PMI): The largest of the professional organizations which foster the discipline of project management.

Project Management Professional (PMP): The highest level of professional certification of project managers by PMI.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2348-2352, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Software & Systems Design

Project-Based Software Risk Management Approaches

Subhas C. Misra
Carleton University, Canada

Vinod Kumar
Carleton University, Canada

Uma Kumar
Carleton University, Canada

INTRODUCTION

The last few decades—especially the end of 20th century and the beginning of 21st century—have shown an increase in the interest in automation of different activities. Automation is dependent in its core on sound functional software. The complexity of software development has increased significantly over the years. Articles showing the failure of projects in the software industry are not surprising. Standish Group (1994) reports show that about 53% of projects get completed, but they do not meet the cost and schedule requirements, and about 31% are canceled before the completion of the projects. These failure reports are significantly alarming.

With the tremendous growth in the complexity of software development in the last 10 to 15 years, the management of risks in software engineering activities is becoming an important and nontrivial issue from three perspectives: project, process, and product. Therefore, researchers and practitioners are continually trying to find effective risk management approaches.

This article should help the academicians, researchers, and practitioners interested in the area of risk management in software engineering to gain an overall understanding of the area.

BACKGROUND

Meaning of Risk Management

Simply put, risk management is a way to manage risks. In other words, it concerns all activities that are performed to reduce the uncertainties associated with certain tasks or events. Risk management reduces the impacts of undesirable events on a project or the final product. Risk management in any project requires undertaking decision-making activities.

Origin of Risk Management

Risk management has its roots in probability theory and decision making under uncertainty. Three well-known theories in these areas—*expected utility theory* (Bernoulli, 1954; Hogarth, 1987), *theory of bounded rationality* (Simon, 1979), and *prospect theory* (Kahneman & Tversky, 1973; Kahneman, Slovic, & Tversky, 1982)—were of the greatest influence. These theories may be considered as disciplines by themselves. Therefore, to put our discussions on risk management in context, we briefly state hereafter only what each of these theories propose.

In brief, the expected utility theory discusses how people make choices from different alternatives, based on their expected utility. The theory of bounded rationality states that for real life events the outcomes and their associated probabilities are very limitedly understood by people to make the required decisions to maximize their expected utility. Therefore, people have a tendency to set up targets of aspiration in life by eliminating alternatives from the different options they have. This theory is useful for modeling the behavior of project management personnel in charge of risk management. Prospect theory, which has its origin in psychology, helps to model how the perceptions of human beings influence their choices from the given options. Thus, it helps for understanding and estimating the utility losses of different alternatives while analyzing risks in risk management.

Purpose of Risk Management

Risk management in software has different uses. It helps to save projects or products from failing due to different factors such as noncompletion of projects within the specified schedule and budget constraints and not meeting the customer expectations of the final product.

In the context of projects, risk management looks at projects from different perspectives to ensure that the threats

to the projects are identified and analyzed, and appropriate strategies are undertaken to mitigate and control risks. The mitigation strategies may not necessarily mean the cancellation of tasks that involve risks. Many tasks are undertaken in the software industries even after knowing that undertaking them involves taking high risks. The high-risk tasks are sometimes important to provide the industries a leading edge over their competitors.

Software risk management takes a preventative approach leading to completion of projects or the development of products within predictable time, money, and according to the product specifications. In fact, risk-managed projects and products have the ability to reduce costs and time of completion and increase the overall quality of the project and product deliverables. Without these, organizations could risk loss of revenue and customer trust in an average case, or a complete bankruptcy of the participating organizations in the worst.

RISK MANAGEMENT IN SOFTWARE PROJECTS

The software development projects in the early years of the last century conducted risk management using different ad hoc approaches, without following any systematic methodologies. However, with the increasing complexity of software development, industries have realized the importance of risk management, because it helps in reducing the uncertainties involved in developing software and decreasing the chances of project or product failures.

In the context of projects, before applying any risk management method, the team members should be clear about the following dimensions of risks in their projects (Smith & Pichler, 2005):

- The nature of *uncertainty* involved, and the likelihood with which the risk will occur.
- The *loss* that will be incurred if the risk occurs. Loss in software projects can take many forms including loss of revenue, loss of market share, and loss of customer goodwill.
- The *severity* of the loss.
- The *duration* of the risks.

Different Approaches

Project Risk Management

Several software project risk management approaches have been proposed in the past, most of which assess risks during all the phases of software development, by integrating risk management practices along with the software development process. As a result, in these approaches the risk management approaches follow a disciplined process. These approaches are listed as follows:

- Boehm's risk management model (win-win) (Boehm & Ross, 1989; Boehm & Bose, 1994; Boehm et al. 1998),
- SEI's software risk management model (SRE Version 2.0) (Williams et al., 1999),
- Hall's risk management model (P^2I^2) (Hall, 1998)
- Karolak's risk management model (Just-In-Time Software) (Karolak, 1998), and
- Kontio's riskit methodology (Kontio, 2001).

A "horizontal" comparison of all of these approaches may not be fair because although each of them addresses risk management, they were developed under different circumstances for solving—may be related but different issues. For example, Hall's P^2I^2 was developed from a risk management capability modeling perspective. On the other hand, Boehm's win-win model (Boehm & Ross, 1989; Boehm & Bose, 1994; Boehm *et al*. 1998) was developed primarily as a novel software development process model ("spiral" development) taking a risk-based approach. However, we provide later on an overview of the characteristics of all these approaches.

Of all these approaches, Boehm's win-win (Boehm & Ross, 1989; Boehm & Bose, 1994; Boehm *et al*. 1998) is perhaps the most influential software engineering risk management process model, which became popular during the early 1990s. He developed the first software engineering risk management process model, which integrates seamlessly into the software development lifecycle.

SEI's software risk evaluation approach (called, SRE) (Williams et al., 1999) is also quite popular in practice. It has been applied in several software development projects of several government, and nongovernment organizations. SEI's SRE provides a systematic, detailed, and step-wise approach one could use in software development and acquisition projects. It is based on the idea of continuous risk management. Another characteristic of SRE is that it integrates team risk management principles into the core framework.

Hall (1998) proposed a framework from a different perspective. She proposed a comprehensive framework based on the notion of risk management capability maturity. Her approach is based on four critical success factors of risk management, namely, people, process, infrastructure, and implementation (P^2I^2). However, it is the "process" component of the framework which discusses the risk management processes.

Karolak (1996, 1998) looked at software engineering risk management from the just-in-time viewpoint, the idea of which was popular in the traditional manufacturing industries. Like Hall, he also provided a complete framework that one

could use for risk management in software development. His framework first identifies a set of highlevel risk categories, associates them with risk factors, specifies risks assessment measures for each of these factors to obtain quantitative estimates of risks.

Kontio took a stakeholder-oriented approach to risk management. He proposed a thorough process model that recognizes and manages risks by balancing the stakeholder expectations. According to this approach, the stakeholder goals and expectations are modeled as essential entities for defining risks.

Recently, there has been few mentionable works conducted in the area of software risk management. In this article, we mention below some of the following recent approaches:

- Software risk assessment model of Foo and Murugananthan (2000): It takes a quantitative approach to predict risks using situational factors.
- Source-based software risk assessment methodology of Deursen and Kuipers (2003): It is based on the collection of different types of facts.
- ProRisk risk management framework of Roy (2004): It provides a complete framework for risk management based on the Australian AS/NZS 4350 standard.
- One-minute risk assessment tool of Tiwana and Keil (2004): It provides a tool that project managers could use to assess risks in a very short time.

FUTURE TRENDS

Many of the approaches discussed in this article are limited by the lack of empirical evidences supporting them. This is an important area in which future work should be targeted. Focus should be made on comparing the competing approaches with respect to a predefined set of evaluation criteria.

Software development often involves contractors. None of the risk management approaches clearly address the issues related to such resources. Similarly, they do not address the several telecommuters who may be working on the project remotely. More so, the impact of recent changes like offshoring and outsourcing may have several impacts on software development, and their influence in the context of risk management in software engineering should be investigated. The major challenges lie in social and cultural differences between the different players on the project in an outsourced project environment. It might so happen that in a project there are two or more software developers doing similar jobs with vastly different cultural settings and vastly different pay scales. There are issues like time zone differences and, above all, the "perceived" quality by the customers due to outsourcing.

For most of the proposed approaches we need controlled case studies and actual field trials for assessing their effectiveness and applicability under modern contexts and shifting paradigms.

The volatility of software project risks has some negative impact on the acceptance of the risk models that suggest different risk mitigation strategies. Thus, we should perform future studies on software risks keeping the aforementioned factors in mind.

CONCLUSION

It is conjectured that the management of risks can lead to the success of projects. Risk management has been popular in non-software domains for several decades. However, it is primarily in the last few years that risk management in software domains has become popular. However, at present, risk management in software is a developing discipline—it is poorly understood and practiced. Compared to the risk management literature available in other disciplines (e.g., insurance and manufacturing), the volume of risk management literature available in software is scarce. In this article, we attempted to review the fundamentals of software risk management and the different popular risk management project and product-based approaches.

We have reviewed the principles of software project risk management, reviewed some of the risk management approaches popular in the software engineering community, provided a summary of some of the important works conducted recently in this area in the last 5 years, and finally, provided some thoughts on future works that can be done.

The article should help the academicians, researchers, and practitioners interested in the area of risk management in software engineering to gain an overall understanding of the area. The article should be of immense help to the software engineering community.

The implications for practitioners is that they can use risk management approaches to know all possible risks in a project, assess their severity and consequence, and then determine resolution steps depending on the nature of the risks. The idea is to minimize any unforeseen and unexpected issues arising during the course of the project or product by properly planning for eventualities. Proper planning leads to minimizing uncertainties, which might lead to a "turbulent" completion, or a complete cancellation of the projects.

REFERENCES

Basili, V. R. (1993). The experience factory and its relationship to other improvement paradigms. *Proceedings of the 4th European Software Engineering Conference.* Springer-Verlag.

Bernoulli, D. (1954). Exposition of new theory on the measurement of risk. *Econometrica, 22,* 23-36.

Boehm, B. W. (1988, May). A spiral model of software development and enhancement. *Computer,* 61-72.

Boehm, B. W. (1991). Software risk management: Principles and practices. *IEEE Software*, *8*(1), 32-41.

Boehm, B. W. et al., (1998). Using the win-win spiral model: A case study. *IEEE Computer, 31*(7), 33-44.

Boehm, B. W., & Bose, P. (1994). A collaborative spiral software process model based on theory W. *Proceedings of the 3rd International Conference on Software Process,* IEEE Computer Society, Washington.

Boehm, B. W., & Ross, R.. (1988). Theory-W software project management: A case study. *Proceedings of the 10th International Conference on Software Engineering* (pp. 30-40). Singapore.

Boehm, B. W., & Ross, R. (1989). Theory W software project management: Principles and examples. *IEEE Transactions on Software Engineering, 15*(7), 902-916.

Chung, L., Nixon, B. A., Yu, E., & Mylopoulos, J. (2000). *Non-functional requirements in software engineering.* Kluwer.

Deursen, A. V., & Kuipers, T. (2003). Source-based software risk assessment. *Proceedings of the International Conference on Software Maintenance (ICSM'03),* Amsterdam, The Netherlands.

Donzelli, P. (2002). Agents, goals, and quality in a structured requirements engineering framework—A case study. *Proceedings of CAiSE'02,* Toronto, Ontario.

Dorofee, A. J. et al. (1996). *Continuous risk management guide book, SEI.* Pittsburgh, PA: Carnegie Mellon University.

Foo, S.-W., & Muruganantham, A. (2000). Software risk assessment model. *Proceedings of the 2000 IEEE International Conference on Management of Innovation and Technology, 2,* 536-544.

Giunchiglia, F., Mylopoulos, J., & Perini, A. (2002). The tropos software development methodology: Processes, models and diagrams. *Proceedings of AOSE-2002,* Bologna, Italy.

Glutch, D. P. (1994). *A construct for describing software risks.* (Tech. Rep. No. CMU/SEI-94-TR-14). Pittsburgh, PA: Carnegie Mellon University, Software Engineering Institute.

Hall, E. M. (1998). *Managing risk: Methods for software systems development.* Reading, UK: Addison-Wesley.

Hogarth, R. M. (1987). *Judgment and choice.* New York: John Wiley & Sons.

Higuera, R. P., & Haimes, Y. Y. (1996). *Software risk management.* (Tech. Rep. No. CMU/SEI-96-TR-012). Pittsburgh, PA: Carnegie Mellon University, Software Engineering Institute.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases.* New York: Cambridge University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychology Review, 80,* 237-251.

Karolak, D. (1996). *Software engineering risk management.* Los Alamitos, CA: IEEE Computer Society Press.

Karolak, D. (1998). Software engineering risk and just-in-time development. *International Journal of Computer Science and Information Management, 1*(4).

Karunanithi, N., & Whitley, D. (1992, July). Using neural networks in reliability prediction. *IEEE Software.*

Kontio, J. (1997). *The Riskit method for software risk management. (Version 1.00).* (Tech. Rep. No. CS-TR-3782/ UMIACS-TR-97-38). College Park: University of Maryland, Computer Science Department.

Kontio, J. (2001). *Software engineering risk management: A method, improvement framework, and empirical evaluation.* Unpublished doctoral thesis, University of Technology, Helsinki, Finland.

Lanning, D. L. (1995). A neural network approach for early detection of program modules having high risk in the maintenance phase. *Journal of Systems and Software, 29.*

Leishman, T. R., & VanBuren, J. (2003). *The risk of not being risk conscious: Software risk management basics, STSC Seminar Series.* Clearfield, UT: Hill AFB.

Lyu, M. (1995). *Software reliability engineering.* IEEE Computer Society Press.

McManus, J. (2004). *Risk management in software development projects.* Elsevier.

Misra, S. C., Kumar, V., & Kumar, U. (2005a, May 25-28). Modeling strategic actor relationships to support risk analysis and control in software projects. *Proceedings of the 7th International Conference on Enterprise Information Systems,* Miami, FL (pp. 288-293).

Misra, S. C., Kumar, V., & Kumar, U. (2005b, April 20-22). Strategic modeling of risk management in industries undergoing BPR. *Proceedings of the 8th International Conference on Business Information Systems*. Poznan, Poland, (pp. 368-385).

Misra, S. C., Kumar, V., & Kumar, U. (2005c). An approach for modeling information systems security risk assessment. *Proceedings of the 3rd International Workshop on Security in Information Systems,* (pp. 253-262). Miami, FL, May 24-25. (The journal version of the paper is under review by *Journal of Systems and Software,* Elsevier.)

Musa, J. (1998). *Software reliability engineering: More reliable software, faster development and testing.* McGraw-Hill.

Roy, G. G. (2004, April). A risk management framework for software engineering practice. *Proceedings of the 2004 Australian Software Engineering Conference (AAWEC '04),* IEEE Computer Society, Melbourne, Australia.

Simon, H. A. (1979). Rational decision making in business organizations. *The American Economic Review, 69*(4), 493-513.

Smith, P. G., & Pichler, R. (2005). Agile risks/agile rewards. *Software Development Magazine,* 50-53.

Standards Australia. (1999). *Risk management.* (AS/NZS 4360: 1999).

Standish Group. (1994). *The chaos report.* Retrieved from http://www.standishgroup.com/sample_research/chaos_1994_1.php

Tiwana, A., & Keil, M. (2004). The one-minute risk assessment tool. *Communications of the ACM, 47*(11), 73-77.

Williams, R. C. et al. (1999). *Software risk evaluation (SRE) method description (Version 2.0).* (Tech. Rep. No. CMU/SEI-99-TR-029). Pittsburgh, PA: Carnegie Mellon University, Software Engineering Institute.

KEY TERMS

P^2I^2: Elaine Hall's approach for risk management in projects. It is based on four critical success factors of risk management, namely, people, process, infrastructure, and implementation.

Product Risks: Risks related to products developed. These risks have the potential to affect the successful operation of the products. They are often associated with the reliability of operation of the products.

Project Risks: Risks related to projects. These risks have the potential to affect the successful completion of the projects. They are associated with project parameters such as the project time lines and budgets.

Risk: "Risk refers to a possibility of loss, the loss itself, or any characteristic, object, or action that is associated with that possibility" (Kontio, 2001).

Risk Management: The disciple of managing risks using strategies such as planning, assessment, analysis, and control of risks.

Software Reliability: A branch of software engineering dealing with the evaluation of how reliably a software system will perform when functional.

Software Risk Management: The disciple of managing risks in software projects, processes, and products.

PROLOG

Bernie Garrett
University of British Columbia, Canada

INTRODUCTION

Prolog is a logic based programming language, and was developed in the early 1970s and is a practical programming language particularly useful for knowledge representation and artificial intelligence (AI) applications. Prolog is different from many common computer languages in that it is not a procedural language (such as Basic, C, or Java). It is an interpreted logic based declarative language and as such has no loops, jumps, type declarations or arrays, and no fixed control constructs. In the past this has led to the impression that Prolog is a restricted language, useful only for highly specialized programming tasks by enthusiasts (Callear, 1994; Krzysztof, 1997). However, this is not the case and modern versions of Prolog are well equipped and versatile, and can be used for any programming task. The latest generations of the language (e.g., Visual Prolog) can also be integrated into more common object oriented languages.

BACKGROUND

Origins

The development and growth in the use of prolog has followed the expansion of interest in artificial intelligence and knowledge based/expert systems. These are computer systems that simulate human cognitive processes, and incorporate large volumes of information in a database using rules to attempt to encapsulate this information as knowledge (or the knowledge of a human expert in the case of expert systems).

Prolog was developed by Alain Colmerauer of Marseilles University, and Robert Kowalski of the University of Edinburgh, in the early 1970s as an alternative to the American Lisp programming languages (early mathematical notation based languages), and Planner (a procedural language representing "knowledge" in the form of high level procedural plans). Kowalski, was a primary advocate in the logic paradigm community (see Fundamental Ideas), and in collaboration with Alain Colmerauer they created a subset of the language "micro planner" called Prolog, Kowalski hoped to demonstrate with Prolog that the logic paradigm was a viable approach to programming. It was Philippe Roussel (also at Marseilles University) who came up with the name as an abbreviation for "PROgrammation en LOGique" to refer to this software tool which was originally devised as a man-machine interface using natural language.

Fundamental Ideas

Prolog is a declarative language in that all the facts and data relating to the subject domain are stored and statically declared in a Prolog database. Rules are created that draw out the information from the database as necessary. Problem solving is achieved from the perspective of the data rather than the procedure, and this can be highly efficient (Bratko, 1996). We can contrast this with the conventional procedural paradigm where the computer performs a sequence of instructions or procedures to resolve a problem. Prolog does not specify any data types in its structure in the way common programming languages do. It therefore has a very open data structure and does not distinguish integers from real numbers, for example. Prolog has two basic functional components. Firstly, a query interpreter program that searches the second component, a Prolog database of facts and rules. The database or program is normally in the form of a text file.

The Logic Paradigm

John McCarthy (1958) originally proposed that mathematical logic be used for representing the nature of knowledge in computer systems. Marvin Minsky and Seymour Papert developed a different approach based on procedural implementations at MIT where the program simply contains a series of computational steps to be carried out to reach a goal (Hewitt, 2006). The logic programming paradigm developed as an alternative to the procedural paradigm and incorporates the invocation of procedures from inferential and deductive processes. Many people were involved in the endeavor of deriving a computer programming language from the discipline of logic, notably Robert Kowalski at Edinburgh University.

Unlike most procedural languages, Prolog programs are not written in a way that models how a computer works, but incorporate techniques that reflect the logical principals of problem solving. In Prolog rather than describing how to compute a solution, the program consists of a data base of facts (or defined predicates about something) and logical relationships (rules) which describe the relationships which hold between those facts. Rather then running a program based on a set of procedures to find a solution to a problem, the logic paradigm makes the user ask a question. A runtime system then searches through a database of facts and rules using logical deduction to determine the answer to this

question, and then invocates a predetermined procedure as a result.

In reality Prolog is not a full implementation of logic programming as this would be purely declarative. It is more accurate to say that Prolog is a programming language based on logic as its implementation has distinct procedural aspects, such as backtracking (Hewitt & Agha, 1988). However, this is a useful aspect of the language as it makes it straightforward to write conventional computer programs in Prolog.

Backward and Forward Chaining

There are two main methods of reasoning when using inference rules in computer applications. These are forward and backward chaining. Forward chaining starts with the available data and uses inference rules to extract more data until a goal is reached. In backward chaining, the system starts with a list of goals and works backwards to see if there are data available that will support any of these goals. An inference engine using backward chaining searches all the inference rules until it finds one which has a then clause that matches a desired goal. If the if clause of that rule is not known to be true, then it is added to a list of goals to be searched for and the search continues until all the goals are met (or fail to be met). Backward chaining attempts to match the action rather than the conditions during its operation, and works from goals to facts. It eliminates the need to solve every possible outcome for a given set of rules. Forward chaining inference is often called data driven in contrast to backward chaining inference, which is referred to as goal driven reasoning. Prolog is based on mathematical logic, and the basis for this claim is that Prolog uses backward chaining processes in its operation from goal to sub-goal.

This process can be represented by the following code in Prolog:

$goal_X$:- $subgoal_1$, ..., $subgoal_n$.

This states that in order to prove $goal_X$, then you must prove $subgoal_1$ through to $subgoal_n$.

Unification

Unification is the built-in pattern-matching algorithm in Prolog, and one of the main concepts in behind it (Sterling & Shapiro, 1994). It is the mechanism by which variables are bound (or instantiated) to unique assignments. In Prolog, this operation is denoted by the equal symbol (=).

Queries in Prolog work by pattern matching. The query pattern is the goal, and if a fact in the Prolog database matches this goal, then the query succeeds and Prolog responds with 'yes.' If there is no matching fact, then the query fails and Prolog responds with 'no.'

Example:
In the following example Prolog unifies the variable "what" with the atom (a constant string of characters) "trees."

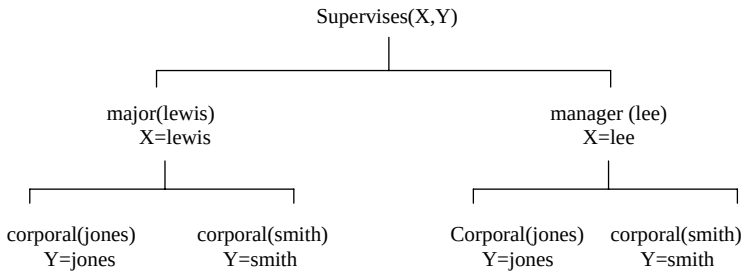
```
?- climbs(bear,What).
What=trees.
Prolog responds: yes
```

Prolog uses a capital letter to indicate a variable and a lowercase letter to indicate an atom. In older versions of Prolog a variable which is has not been instantiated yet can be unified with any atom, term, or another uninstantiated variable. In more modern versions a variable cannot be unified with a term that contains it. Binding a variable to a structure containing that variable can results in a cyclic structure which would cause the unification to loop forever. For example, A=f(A). This is a type of recursion. Modern versions of Prolog include an "occurs check" to prevent this happening.

Backtracking
In Prolog backtracking is a process that allows it to work through all the sub-goals in a rule if one sub-goal fails. In this case Prolog does not give up immediately and make the rule fail but it backtracks to previous sub-goals to try other instances of them in the database, then move forward again and see whether this causes the failed sub-goal to succeed. In this way it goes through a process that tries all the possible combinations of solutions, and finds the successful ones, before it finally reports that a rule has failed (Coehlo & Cotta, 1988).

In order to cope with the very limited memory systems and sequential computer architecture that were available when the language was developed, an efficient backtracking control structure was implemented so that only one possible computational path had to be stored at a time. This backtracking process is a method that has to be used on a sequential computer, which can only do one thing at a time and has to work through all the possibilities systematically. As it searches, Prolog leaves markers at points in the database to which it returns if a path down a particular branch fails to yield a resulting match. This exhaustive search method used by Prolog is called a depth first search method (Nilsson & Maluszynski, 1995). Diagram 1 demonstrates how this works in practice. The tree shows possible solutions for a supervision rule. Prolog finds "major(lewis)" first and then explores all possibilities of the "corporal" predicate before moving on to "major(lee)." It then explores all the possibilities of "corporal" again going as deep down the branches as possible.

Diagram 1. Depth first searching



Tail Recursion

In Prolog recursion involves a rule using itself as a sub-goal, with a fresh call made to the sub-goal each-time. This is memory stack intensive. Tail recursion (also known as tail-end recursion) is a type of recursion that converts a recursive function into an iterative function (O'Keefe, 1990). Iteration is commonly used for repetition of the same code in programs and is efficient. If a call is tail-recursive nothing has to be done after the call returns, that is, when the call returns, the returned value is immediately returned from the calling function. Here is an example.

For the rule "Int (Small, Big)" Big and Small are integers and Small =< Big. "Int" is uninstantiated and will be bound successively on backtracking to Small, Big+1, and Big. The recursive call 'for (Int, Next, Big), the last goal in the body of this clause, is tail-recursive:

```
for(Int, Int, _Big).
for (Int, Small, Big):-
      Small < Big,
      Next is Small + 1,
      for(Int, Next, Big).
```

Tail recursion is a form of recursion that can be implemented much more efficiently than general recursion. In Prolog it facilitates an efficient method of making a section of code repeat as there are no such conventional looping constructions such as UNTIL, WHILE, FOR, DO or GOTO.

CHALLENGES

Arguments over standards for Prolog implementations have been common since its inception. The original Edinburg standard was developed early on provides a basic minimum version of the language. This is sometimes also refered to as the Clocksin and Mellish standard (Callear, 1994; Cloksin & Mellish, 1994). However, there have been a variety of versions of Prolog and interpreters over the years such as LPA Prolog, and AMZI Prolog to name but two. Some versions (such as Visual Prolog) have implemented structural data types, and other procedural structures. These have radically altered the original open architecture of the language. All these versions of Prolog operate slightly differently depending on whose version you use. Therefore work on an international standard began in Britain in 1984, and the ISO Prolog standard: ISO/IEC 13211-1 was published in 1995. This standard defines a portable set of Prolog built-in predicates and their semantics. The original intention was, to standardize the existing practices of the many implementations of Prolog. However, the standard introduced an IO system which departed radically from the " see " and " tell " predicates of original Edinburgh Prolog. For these reasons the ISO Prolog standard has been suggested to not be taken seriously by the Prolog development community (Bagnara, 1999), and the different versions that still arise would tend to support this view.

In the early 1980s Prolog was selected (some would argue erroneously) as the language for the Fifth Generation Computer System project by the Japan Information Processing Development Center (JIPDEC). The project envisaged a parallel processing computer running massive databases using a Prolog to access the data (Feigenbaum & McCorduck, 1983). A primary problem was that their selected language, Prolog, did not support concurrency, (Shapiro, 1989) and this amongst other problems led to the eventual demise of this project in1993.

FUTURE TRENDS

Arguments between proponents of the logic paradigm and the procedural paradigm continue. Declarative database

query languages have been criticized by the computer research community (Hewitt, 2006; Manthey, 1990). Database programming and object-oriented database developers have recommended ending declarative means for organizing retrieval. On the converse deductive database and logic programmers want to extend declarative query languages into full programming languages while retaining declarative features. It is our most likely that neither approach will prevail, and both approaches offer significant advantages (Hewitt, 2006).

CONCLUSION

The unique and open nature of Prolog offers significant advantages in flexibility and efficiency for AI and natural language applications. Because of this niche in the programming world, the variety of open source versions available, and the ability to develop complex applications quickly with Prolog, it continues to remain popular it is likely that it will continue to be used and develop in the future. Indeed, Prolog remains one of the most common languages taught in Universities today.

REFERENCES

Bagnara, R. (1999). *Is the ISO Prolog standard taken seriously*? HTML. Retrieved from on January 10, 2006, from http://www.cs.unipr.it/~bagnara/Papers/Abstracts/ALPN99a

Bratko, I. (1996). *Prolog programming for artificial intelligence*. London: Addison Wesley.

Callear, D. *Prolog programming for students: With expert systems and artificial intelligence topics.* London: DP Publications.

Clocksin, W.F., & Mellish, C.S. (1994). *Programming in prolog* (4th ed.). New York: Springer-Verlag.

Coehlo, H., & Cotta, J.C. (1988). *Prolog by example*. New York: Springer-Verlag.

Colmerauer, A., & Roussel, P. (1992). The birth of prolog. In *The second ACM SIGPLAN conference on History of programming languages*, (pp. 37-52).

Feigenbaum, E.A., & McCorduck, P. (1983). *The fifth generation: Artificial intelligence and Japan's computer challenge to the world*. London: Michael Joseph.

Hewitt, C. (2006). The repeated demise of logic programming and why it will be reincarnated What Went Wrong and Why: Lessons from AI Research and Applications. *Technical Report SS-06-08*. AAAI Press.

Hewitt, C., & Agha, G. (1988). Guarded Horn clause languages: are they deductive and Logical? *Proceedings of the International Conference on Fifth Generation Computer Systems*, Ohmsha 1988. Tokyo.

Kowalski R. (1988). The early years of logic programming. *CACM* January 1988.

Krzysztof, R. (1997). *From logic programming to prolog*. London: Prentice Hall.

Manthey, R. (1990). Declarative languages—paradigm of the past, or challenge for the future. In A.W. Schmidt & A.A. Stogney (Eds.), *Proceedings of the 1st East/West Database Workshop, Kiev, Ukraine, Next Generation Information Systems Technology, LNCS 504,* (pp. 1-16). New York: Springer Verlag.

McCarthy, J. (1959). John McCarthy. Programs with Common Sense. *In Proceedings of the Symposium on the Mechanization of Thought Processes,* Vol. 1, (pp. 77-84). London

Nilsson, U., & Maluszynski, J. (1995). *Logic programming and prolog* (2nd ed.). New York: John Wiley.

O'Keefe, R. (1990). *The craft of prolog*. Massachusetts: MIT Press.

Shapiro, E. (1989). The family of concurrent logic programming languages. *ACM Computing Surveys*. September 1989.

Sterling, L., & Shapiro, E. (1994). *The art of prolog* (2nd ed.). Massachusetts: MIT Press

KEY TERMS

Atom: A group of alphabetical characters in Prolog, similar to a string in other languages

Argument: A word or phrase that occurs in brackets after the head of the predicate, that makes up a fact in the Prolog database (really a predicate consists of a head and one or a number of arguments) for example, in this predicate "animal(mammal)." Mammal is the argument.

Facts: A type of statement made in Prolog. After you supply a database of facts and rules; and can then perform queries on the database. Facts consist of a predicate head and argument. For example "cat(meows)." could be a fact entered in the database that a cat meows. Facts together with rules are also known as "clauses."

Iteration: In computing is the repetition of a process within a computer program. It can be used both as a general term, synonymous with repetition, and to describe a specific form of repetition with a mutable state.

Instantiation: The process in prolog of making a variable equal to a constant, or giving it a value. For example, "X is 2+2" introduces X as a variable (as it is uppercase), and X is instantiated to 2+2. Prolog will respond X=4.

Occurs Check: A system built into some versions of Prolog to avoid never-ending loops with unification. Whenever an attempt is made to unify a variable with a compound term, a check is made by the system to see if the variable is contained within the structure of the compound term, and if this is so, the unification will fail. For efficiency most versions of Prolog do not perform an occurs check.

Predicate: Consists of a word in the Prolog database which succeeds in the Prolog interpreter by writing its argument. It consists of a predicate head (an atom) and one, or a number of arguments. The predicate is the basic unit of Prolog and is always defined to be true. Prolog has some standard built in Predicates to help in programming, such as "write" for example "write(jane)." Will succeed and Prolog will respond "Jane yes."

Recursion: In Prolog, recursion appears when a predicate contain a goal that refers to itself. When a rule is called Prolog creates a new query with new variables. A recursive definition always has at least two parts: facts that act like a stopping condition, and a rule that calls itself.

Rules: Are another type of statement made in Prolog (along with facts) and are also called "clauses." Rules are really extensions of facts in Prolog, with added sub-goals that also have to be satisfied to be found true by the interpreter. A rule consists of a head (a predicate and argument) and a body (sub-goals) separated by the :- symbol. For example, fault (electric):- car_will_not_start,no_lights. This rule has two sub-goals which must both be found in the database for the rule to succeed (the car will not start and has no lights).

Variables: Are quantities that can take any value, and in Prolog are introduced as strings of characters starting with a capital letter. For example, "animal" is an atom in Prolog whereas "Animal" is a variable and can be instantiated for any animal.

Extended and Derivative Versions of Prolog

Datalog: Is actually a subset of PROLOG. It is limited to relationships that may be stratified such that solutions on a large knowledge base return in finite time.

F-Logic: An extended version of Prolog with frames/objects for knowledge representation.

HiLog and λProlog: Are extended versions of Prolog with additional higher-order programming features.

InterProlog: A programming library bridge between Java and Prolog, incorporating bi-directional predicate/method calling between both languages.

Logtalk: An open source object-oriented extension to the Prolog programming language. Integrating logic programming with object-oriented and event-driven programming, it is compatible with most Prolog compilers. It supports both prototypes and classes. In addition, it supports component-based programming through category-based composition.

OW Prolog: A version of Prolog created in order to answer Prolog's lack of graphics and interface.

Prova: Provides native syntax integration with Java, agent messaging and reaction rules. Prova positions itself as a rule-based scripting (RBS) system for middleware. The language breaks new ground in combining imperative and declarative programming.

SWI-Prolog and YAP: Are both open source versions of Prolog offering a full development environment, including graphics, libraries and many interface packages and written in the C/C++. YAP is a high-performance Prolog compiler developed at LIACC/Universidade do Porto, Portugal.

Visual Prolog: Formerly known as PDC Prolog and Turbo Prolog: a strongly-typed object-oriented version of Prolog, which is considerably different than standard Prolog. Incorporates bi-directional predicate/method calling between Prolog and some other common object- oriented languages.

Prolonging the Aging of Software Systems

Constantinos Constantinides
Concordia University, Canada

Venera Arnaoudova
Concordia University, Canada

INTRODUCTION

The evolution of programming paradigms and languages allows us to manage the increasing complexity of systems. Furthermore, we have introduced (and demanded) increasingly complex requirements because various paradigms provide mechanisms to support their implementation. As a result, complex requirements constitute a driving factor for the evolution of languages which in turn can support system complexity. In this circular relationship, the maintenance phase of the software life cycle becomes increasingly important and factors which affect maintenance become vital.

In this chapter we review the notions of software aging and discuss activities undertaken during maintenance. We also discuss challenges and trends for the development of well-maintained systems as well as for aiding in the maintenance of legacy systems.

BACKGROUND

Aging in Software

In the literature, many authors tend to have drawn analogies between software systems and biological systems (ISO/IEC 12207:1995(E); Jones, 2007; Parnas, 1994). Two such notable examples are the widely used notions of aging and software life cycle, implying that we can view software systems as a category of organisms. This analogy is convenient because it creates certain realizations about software. First, we note that systems exist (by operating as a community of intercommunicating agents) inside a given environment. Furthermore, much like their biological counterparts, they evolve (to adapt to their environment) and they grow old. Finally, when speaking of the life cycle of software, we also imply the unavoidable fact that software systems eventually die.

However, the causes of software aging are very different from those of biological organisms or those that cause aging in other engineering artifacts. Unlike biological organisms (such as humans) software systems are not subjected to fatigue or physical deterioration. Unlike other engineering products (such as machinery and structures), software systems are not subjected to physical wear caused by factors such as friction and climate. Aging in software systems is predominantly (but not always) caused by changes that take place in their surrounding (operating) environment.

In his seminal paper on aging, author David Parnas (1994) describes two causes of software aging: The first factor, referred to as lack of movement, is the failure of owners to provide modifications to the software in order to meet changing needs (requirements) of its environment which results in end-users changing to newer products. The second factor, referred to as ignorant surgery, is the careless introduction of changes in the implementation which can cause the implementation to become inconsistent with the design, or even to introduce new bugs. This latter factor is associated with two significant implications: The first is a bloating of the implementation, resulting in a reduction in performance (memory demands, throughput and response time). This weight gain makes new changes difficult to be introduced quickly enough to meet market demands. The second implication is a phenomenon known as bad fix injection (Jones, 2007), which refers to the introduction of errors during maintenance resulting in a decrease in reliability. As a result, software systems become unable to be competitive in the market, thus loosing customers to newer products.

Measures to Prolong Aging

Certain measures are proposed in the literature (Parnas, 1994) to prolong aging such as:

1. The quality of documentation can be upgraded (retroactive documentation). For example, reverse engineering is a model transformation activity which can read implementation and produce an up-to-date design model.
2. Since we cannot really predict the actual changes, predictions can be made about the types of changes, such as changes to the graphical user interface. Parnas (1994) recommends re-organizing the software in such a way so that elements which are most likely to change, such as the user interface, are confined to small amounts of code (retroactive modularization). A similar view is shared by Fayad and Altman (2001) through an architectural pattern to support software stability where the architecture is built around two notions, conceptualized as two concentric circles. In

the inner circle, we have aspects of the environment that will not change. These aspects will constitute a stable core design (and thus a stable software product). In the outer circle, or periphery, we define a design which will allow changes to be introduced.

3. Eliminate components which are of very low quality (amputation).
4. Eliminate redundant components (major surgery).

These measures take place during the period of operability of a software system and are explicitly treated as a separate phase of the software life cycle which is discussed subsequently.

Software Maintenance

ISO/IEC and IEEE define maintenance as the modification of a software product after delivery to correct faults, improve performance (or other attributes) or to adapt the product to a modified environment (ISO/IEC 14764:2006(E); IEEE Std 14764-2006). The importance of maintenance lies on the following observations: (1) Surveys indicate that it is an activity which tends to consume a significant proportion of the resources utilized in the overall life cycle (consequently consuming a large part of the costs) and (2) Reliable changes to software tend to be time consuming. Prolonged delays during software change may result in a loss of business opportunities.

The objective of maintenance is not to stop the unavoidable effects of aging, but to provide techniques and tools to understand its causes, to limit its effects and to prolong the life of software systems.

Maintenance is not a uniform activity and as the type of required changes may vary, four different types of maintenance can be identified which are also defined in the ISO/IEC; IEEE international standard. Corrective maintenance includes all changes made to a system after deployment to correct problems. Preventive maintenance includes all changes made to a system after deployment to correct faults in order to prevent failures. Adaptive maintenance includes all changes made to a system after deployment to address new requirements. Perfective maintenance includes all changes made to a system after deployment to support operability in a different (software or hardware) environment. The ISO/IEC; IEEE international standard provides a classification scheme by grouping the former two under correction and the latter two under enhancement. Adaptive and perfective types of maintenance are shown in the literature to consume a significantly large proportion of all maintenance effort. Corrective and preventive types of maintenance are reported to consume a relatively small proportion of the overall maintenance effort. It is important to note, that the different types are not mutually exclusive but rather they can be combined concurrently to be mutually supportive.

Also, the four maintenance types do not refer to single activities. Jones (2007) lists 23 discrete topics which involve a modification of an existing system often described under maintenance.

Stages of Maintenance and the Staged Model of the Software Life Cycle

In the literature, Bennett and Rajlich (2000) define a model whereby a software system undergoes distinctive stages during its life: Initial development, evolution, servicing, phase-out, and closedown.

Initial development would produce a deployable system (the first operating version). After deployment, evolution would extend the capabilities of the system, possibly in major ways. Once evolution is no longer viable, the software would enter the servicing stage (often referred to as maturity, or most commonly legacy stage). As the term suggests, only small changes are possible during this stage.

Maintainers often encounter what Bennett (1995) describes as the legacy dilemma: On one hand, a system (or component) is valuable and replacing it may not be a viable (cost effective) solution (e.g., large volumes of data may have to be converted). On the other hand, the cost of maintenance is becoming high and requests for changes cannot be sustained. When faced with legacy systems, organizations have to adopt a strategy which is based on economics (i.e., cost of coping with the current system vs. the cost of investment of improvement) and management (e.g., a replacement system would normally require training of end-users). Table 1 summarizes various options based on two factors, namely business value and quality (adopted from Sommerville, 2007).

Finally, once servicing is no longer viable the system enters a phase-out stage where deficiencies are known but not addressed. At closedown, the system is withdrawn from the market. In an alternative model (versioned staged model), during evolution a version is publicly released and subsequently enters the servicing stage whereas the system continues to evolve in order to produce the next version.

Central to any maintenance activity is the notion of change, discussed in the next subsection.

SOFTWARE CHANGE

Whether new requirements are introduced or existing requirements are refined or dropped, the notion of change is a fundamental activity during evolution and servicing. Bennett and Rajlich (2000) describe a change mini-cycle as one which involves a number of activities: request for change, planning phase, change implementation, verification, and documentation update.

Table 1. Maintenance options based on business value and quality factors

		QUALITY	
		Low	High
BUSINESS VALUE	Low	Expensive to keep, with a low return. Scrap. Would require a modification of business processes which rely on it.	Continue with maintenance while it remains cost effective. Once maintenance becomes expensive, then scrap.
	High	Expensive to maintain. Options include reengineering or replacement by off-the-shelf systems.	No special measures required. Continue normal system maintenance.

1. Request for change: originates from a number of sources including bug reports, system enhancement requests from end-users or changes in standards.
2. Planning phase: Involves comprehension and change impact analysis.

Comprehension: It refers to activities that humans perform in order to understand, conceptualize and reason about a program or software system. It normally consumes a large proportion of resources during the overall maintenance phase.

Often it is the case that maintainers are not the initial designers. Furthermore, design artifacts, if at all present, are often incorrect or incomplete. For systems that have been undergone ignorant surgery, comprehension becomes particularly difficult, as it has to bridge the gap between what the original designers had envisioned (initially represented by a coherent and structured description of a model) and the actual system (whose structure may have disintegrated over time). This tendency of systems to experience a structural disintegration over time (also known as entropy) and the subsequent increase in complexity was discussed by Lehman (1980) as one of the laws that govern evolving systems (known as Lehman's second law).

One method to achieve comprehension is through reverse engineering (sometimes the two terms are used interchangeably), which transforms a representation of the system to a higher level of abstraction. This activity involves analysis only, without any modification. The objectives of reverse engineering include the recovery of information which is either not apparent in the code or it is difficult to detect as well as to facilitate reuse by detecting candidates for reusable components. A variation (or specialization) of reverse engineering is design recovery whose objective is to identify meaningful high-level abstractions, which are not identifiable by examining the implementation. Reverse engineering falls under the notion of translation in the model transformation taxonomy (Mens & Van Gorp, 2006), as the source and target languages are at a different level of abstraction.

For complex systems, comprehension methods rely on the study of the dependencies between program (or software) elements (see Tables 2 and 3), such as program slicing and formal concept analysis.

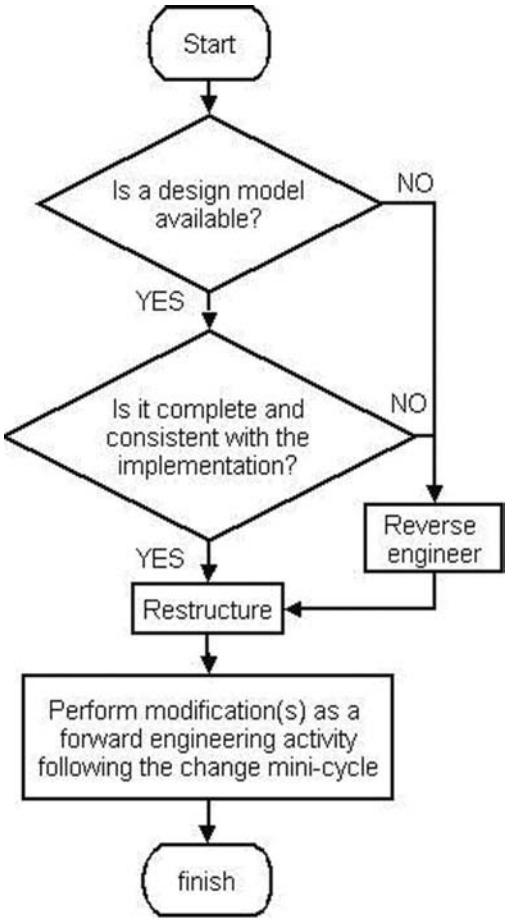
Change impact analysis: To identify impacted components and to assess the impact of change.

3. Change implementation: Involves restructuring and change propagation.

Often deployed to perform preventive maintenance and to simplify or optimize a model, restructuring (or its object-oriented equivalent: refactoring) involves the transformation of one representation form into another at the same level of abstraction. As such, it falls under the notion of rephrasing in the model transformation taxonomy. One important property of restructuring is that it must guarantee the preservation of the behavior of the system. It may also detect problems which would call for changes. In the object-oriented context, strategies for refactoring have been extensively documented in the literature (Fowler, Beck, Brant, Opdyke & Roberts, 1999) and are currently supported by a number of tools[a], perhaps the most notable of which is currently the Eclipse IDE[b].

Change propagation involves changes which may have to be made to dependent components to make sure that

Figure 1. Integration of reengineering activities in the change mini-cycle



changes made in the previous step have not violated the integrity of the system.

In implementing changes, we often involve a sequence of three activities: reverse engineering – restructuring – forward engineering, which are together referred to as reengineering (or renovation). Reengineering addresses system functionalities which need to be added, refined or deleted. Essentially reengineering uses program comprehension to reimplement the program in a new form (Figure 1). Reengineering can be advantageous only when a return of investment can be projected (cost of reengineering vs. increase of product and process quality and business value) (Sneed, 1995) and best candidates to undergo this transformation are components with low quality but strategic to the organization (high business value) (Table 1).

4. Verification: Once a change has been implemented, we need to argue about the preservation of the system behavior and correctness.
5. Documentation update: Artifacts need to retain their interconsistency. Computer-aided software engineering (CASE) tools[c] can automate this process.

FUTURE TRENDS

Reports in the literature indicate a disproportionately large amount of resources and time devoted to maintenance. Of that, a large proportion is reported to be spent on comprehension.

To ease the maintenance phase, we can follow two paths, which we feel exist in mutual support: The less obvious (or perhaps less appreciated) one is to look into development and focus on the provision of quality attributes which can affect maintainability. The obvious one is to improve current methods which are utilized to perform the various maintenance activities. We address challenges and related research along these two paths (we refer to them as pre and postdeployment) in the next subsections.

Predeployment Challenges

The challenge that we face during development is how to encourage and support the consideration of quality attributes which can later have an effect on maintenance. We will restrict the rest of the discussion to the context of object-oriented systems even though many of the ideas are applicable to other paradigms, such as procedural and functional programming.

Maintainability as a quality factor: The degree to which a software system can be easily modified (or alternatively, the effort required for it to be modified) is referred to as maintainability and it is defined as one of the six characteristics of the ISO/IEC 9126-1 quality model (2001) with the subcharacteristics of analyzability (the ease with which the cause of a failure can be detected), changeability (how easy it is to change the software), stability (low risk of modification having unexpected effects), testability (establishing test criteria and the performance of tests) and maintainability compliance (adhering to standards or conventions). Maintainability should be explicitly addressed as part of the nonfunctional requirements of a system. A related term is adaptability which is defined as the ability of a software system to allow changes. In their 1996 article, authors Fayad and Cline discuss factors which affect adaptability. High-level changes are addressed by extensibility (change amount of capabilities but not kind) and flexibility (change kind of capabilities). Low-level changes are addressed by performance tunability and fixability (the ability to fix one thing without breaking another).

Separation of concerns and modularity: The principle of separation of concerns (Parnas, 1972) refers to the real-

ization of system concepts into separate software units and it is a fundamental principle to software development. The associated benefits include better analysis and understanding of individual concerns, high readability of modular code, high-level of reuse, easy adaptability and good maintainability. Despite the success of object-orientation in the effort to achieve separation of concerns, certain properties cannot be directly mapped in a one-to-one fashion from the problem domain to the solution space, and thus cannot be localized in single modular units. Their implementation ends up cutting across the inheritance hierarchy of the system. Crosscutting concerns (or aspects) include persistence, authentication, synchronization, and logging. The crosscutting phenomenon creates two implications: (1) the scattering of concerns over a number of modular units and (2) the tangling of code in modular units. As a result, developers are faced with a number of problems including low level of cohesion of modular units, strong coupling between modular units and difficult comprehensibility, resulting in programs that are more error prone.

Aspect-Oriented Programming (AOP) (Kiczales, Lamping, Mendhekar, Maeda, Lopes, Loingtier et al., 1997) explicitly addresses those concerns which cannot be cleanly encapsulated in a generalized procedure (i.e., object, method, procedure, API) by introducing the notion of an aspect definition, which is a new modular unit of decomposition. There are currently a growing number of approaches and technologies to support AOP. One notable technology is AspectJ (Kiczales, Hilsdale, Hugunin, Kersten, Palm & Griswold, 2001), a general-purpose aspect-oriented extension to the Java programming language, which has influenced the design dimensions of several other general-purpose aspect-oriented languages, and has provided the community with a common vocabulary based on its own linguistic constructs. It is important to note that AOP is neither limited to Object-Oriented Programming (OOP) nor to the imperative programming paradigm.

During the construction of software, developers aim to produce a clean (tangled-free) system and achieve the maximum benefits of advanced separation of concerns. To meet this objective, the design artifacts themselves must in turn explicitly address crosscutting concerns. To this end, we need to provide the means to identify and model crosscutting concerns from the early stages of the software life cycle. As a result, the explicit capture of crosscutting concerns in code should be the natural consequence of good and clean modularity and not the result of a corrective measure (refactoring activity) due to a tangled implementation. One future development we should expect is the extension of the official UML specification to provide support for aspects[d].

Reusability and component-based software: As software products need to satisfy both technical and nontechnical criteria, developers find it essential to combine theory and experience in order to reuse proven designs. The importance of reuse lies on the fact that it can speed up the development process, cut down costs, increase productivity and improve the quality of software. Design-level reuse is viewed as the attempt to share certain aspects of an approach across various projects. Object and component-based systems offer a wide spectrum of techniques to reuse designs at different levels, ranging from frameworks and patterns that constitute approaches on how to best program in-the-large to libraries and programming languages that constitute approaches on how to best program in-the-small (Szyperski, 2002). On the higher level of the reuse spectrum, the literature for object-oriented systems includes discussions on frameworks (Fayad, Johnson & Schmidt, 1999) and design patterns (Gamma, Helm, Johnson & Vlissides, 1994). Discussions on frameworks (Constantinides, Elrad & Fayad, 2002) and design patterns (Hannemann & Kiczales, 2002) are also found in the literature for aspect-oriented systems. On the lower level of the reuse spectrum, the importance of the mechanism of inheritance and the contribution of OOP towards reusability has been extensively discussed in the literature. Inheritance allows non-invasive (incremental) modification of class definitions. Subclasses may introduce new attributes and methods whose definitions may be based on those of the superclasses. Furthermore, methods may be overloaded (providing different semantics for the same method names) or overridden (providing new semantics for the same method signatures).

Postdeployment Challenges

Both OOP and AOP introduced their own maintenance problems, particularly for comprehension of the entire system which comes as a tradeoff for increasing the level of separation of concerns and the increase of the level of comprehension of individual modules. Analysis methods have been deployed for OOP, but not much work has been done on AOP, particularly to address the difficulty to identify portions of the core functionality affected by the aspectual behavior.

There are currently a number of methods and tool support to ease the activities involved, focusing on comprehension, which tends to consume a large proportion of time. The analysis of artifacts can be categorized into static and dynamic (Table 2) based on the view over which an artifact is investigated (structural or behavioral respectively). Static analysis is performed by examining design or implementation artifacts and reasoning over possible behaviors that might arise during execution. Dynamic analysis is performed by executing a program and observing the executions. The latter is precise, because no approximation or abstraction need to be done and the analysis can examine the exact runtime behavior of the program. The analysis of an artifact is performed based on the dependencies between its elements. There are currently several different analysis methods avail-

Table 2. Analysis types

	SOURCE CODE	MODEL	EXECUTION TRACES
STRUCTURAL (STATIC ANALYSIS)	Yes	Yes	Not Applicable
BEHAVIORAL (DYNAMIC ANALYSIS)	Yes	Yes	Yes

Table 3. Analysis methods

DESCRIPTION	STATIC ANALYSIS	DYNAMIC ANALYSIS
Database: Relational schema.	Xiao & Pham (2004).	Parsamanesh *et al.* (2006).
Declarative	De Volder (2006); Mousavi Eshkevari *et al.* (2008).	Richner *et al.* (1998).
Control-Flow Graph (CFG)	Robillard & Murphy (2002).	Breu & Krinke (2004).
Slicing	Horwitz *et al.* (1988); Zhao (2002).	Korel & Rilling (1997).
Formal Concept Analysis (FCA)	Snelting & Tip (1998).	Eisenbarth *et al.* (2003); Tonella & Ceccato (2004).

Table 4. Aspect-mining tools

Feature Exploration and Analysis tool (FEAT)	Robillard & Murphy (2007).
Aspect Mining Tool (AMT)	Hannemann & Kiczales (2001).
Clone detection	Bruntink *et al.* (2005).
PRISM	Zhang & Jacobsen (2004).

able (Table 3). Furthermore, a number of tools exist which focus on analysis for aspect mining (Table 4).

Maintenance model: Unlike development which is currently supported by two major process types (linear and iterative) under a number of various different models (such as the spiral model, the unified software development process, or extreme programming and agile methods), there are no equivalent models for maintenance. The question we need to answer is whether a model is required, or should maintenance be integrated as a sequence of postdeployment iterations in an existing model.

Reengineering of (object-oriented) legacy systems: One consequence of the adoption of UML support for aspects, is that we will be able to achieve reengineering of legacy object-oriented systems and their migration to an aspect-oriented context through aspect mining, refactoring, reverse engineering, and forward engineering. As an example, legacy Java code can be rejuvenated into AspectJ.

The following two quality attributes of methods and tools for comprehension should also be taken into consideration.

Adaptability: The degree to which current methods, techniques and tools can be adapted to address new aspects of comprehension, such as crosscutting concerns.

Scalability: In discussing difficulties associated with reverse engineering, Rugaber (1992) pointed out the need for powerful tool support. The increasing complexity of systems has motivated automation and today we can say that comprehension is largely automatic with a large collection of tools available[e]. However, a challenge is to make sure that current methods and tools to support comprehension can scale for large systems.

CONCLUSION

Not only the life of a software system does not end at the time of its deployment, but it can be argued that this is the time when software begins its productive life. In this chapter we discussed the notion of software aging as the driving force behind maintenance. As we face legacy systems, maintenance and evolution are becoming increasingly important and complex. The complexity of maintenance lies on the fact that comprehension of large systems is tedious, particularly in the cases where the source code is not supported by a consistent design model.

We have discussed challenges of maintenance from two viewpoints: first, during development, we aim at designing for change and capturing quality attributes to support well-maintained systems. We are, of course, aware that our ability to design for change can only be approximate, since it depends on our ability to predict the future and, despite the successful implementation of quality attributes, we cannot prevent aging but can merely prolong it. Second, for systems currently in operation, our goal is to design technologies to ease comprehension through static and dynamic analysis, such as declarative reasoning, program slicing and formal concept analysis.

REFERENCES

Bennett, K. (1995, January). Legacy systems: Coping with success. *IEEE Software, 12*(1), 19 – 23.

Bennett, K. H. & Rajlich, V. T. (2000). Software maintenance and evolution: A roadmap. In *Proceedings of the 22nd International Conference on Software Engineering, Future of Software Engineering Track* (pp. 73 – 87). New York: ACM.

Breu, S. & Krinke, J. (2004). Aspect mining using event traces. In *Proceedings of the 19th IEEE International Conference on Automated Software Engineering* (pp. 310 – 315). Washington, DC: IEEE Computer Society.

Bruntink, M., van Deursen, A., van Engelen, R., & Tourwé, T. (2005, October). On the use of clone detection for identifying crosscutting concern code. *IEEE Transactions on Software Engineering, 31*(10), 804 – 818.

Constantinides, C. A., Elrad, T., & Fayad, M. (June 2002). Extending the object model to provide explicit support for crosscutting concerns. *Software Practice and Experience, 32*(7), 703 – 734.

De Volder, K. (2006). JQuery: A generic code browser with a declarative configuration language. In *Proceedings of the 8th International Symposium on Practical Aspects of Declarative Languages* (pp. 88 – 102). Berlin/Heidelberg, Germany: Springer (LNCS 3819).

Eisenbarth, T., Koschke, R., & Simon, D. (March 2003). Locating features in source code. *IEEE Transactions on Software Engineering, 29*(3), 210 – 224.

Fayad, M. E. & Altman, A. (2001, September). Thinking objectively: An introduction to software stability. *Communications of the ACM, 44*(9), 95 – 98.

Fayad, M. & Cline, M. P. (1996, October). Aspects of software adaptability. *Communications of the ACM, 39*(10), 58 – 59.

Fayad, M. E., Johnson, R. E., & Schmidt, D. C. (Eds.) (1999). *Building application frameworks: Object-oriented foundations of framework design*. New York: John Wiley & Sons.

Fowler, M., Beck, K., Brant, J., Opdyke, W., & Roberts, D. (1999). *Refactoring: Improving the design of existing code*. Menlo Park, CA: Addison Wesley Longman, Inc.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. M. (1994). *Design patterns: Elements of reusable object-oriented software*. Upper Saddle River, NJ: Addison-Wesley Professional.

Hannemann, J. & Kiczales, G. (2001). Overcoming the prevalent decomposition of legacy code. In *Proceedings of the 23rd International Conference on Software Engineering Workshop on Advanced Separation of Concerns.* Retrieved June 16, 2008, from http://www.cs.ubc.ca/~jan/amt/

Hannemann, J. & Kiczales, G. (2002, November). Design pattern implementations in Java and AspectJ. *ACM SIGPLAN Notices, 37*(11), 161 – 173.

Horwitz, S., Reps, T. W., & Binkley, D. (1988, July). Interprocedural slicing using dependence graphs. In *Proceedings of the ACM SIGPLAN Conference on Programming Language Design and Implementation* (pp. 35 – 46). New York: ACM.

International Standards Organization/ International Electrotechnical Commission (1995). *ISO/IEC 12207:1995(E): International standard: Information technology – Software life cycle processes* (1st ed.). Geneva, Switzerland: ISO/IEC.

International Organization for Standardization/ International Electrotechnical Commission (2001). *ISO/IEC 9126-1:2001 Software engineering - Product quality - Part 1: Quality model*. Geneva, Switzerland: ISO/IEC.

International Standards Organization/ International Electrotechnical Commission; Institute of Electrical and Electronics Engineers (2006). *ISO/IEC 14764:2006(E); IEEE Std 14764-2006: International standard: Software Engineering – Software life cycle processes – maintenance* (2nd ed.). Piscataway, NJ: ISO/IEC; IEEE.

Jones, C. (2007, December). Geriatric issues of aging software. *CrossTalk - The Journal of Defense Software Engineering, 20*(12), 4 – 8.

Kiczales, G., Lamping, J., Mendhekar, A., Maeda, C., Lopes, C. V., Loingtier, J.-M., et al. (1997). Aspect-oriented programming. In M. Akşit & S. Matsuoka (Eds.), In *Proceedings of the 11th European Conference on Object-Oriented Programming* (pp. 220 – 242). Berlin/Heidelberg, Germany: Springer (LNCS 1241).

Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J., & Griswold, W. G. (2001). An overview of AspectJ. In J. L. Knudsen (Ed.), In *Proceedings of the 15th European Conference on Object-Oriented Programming* (pp. 327 – 353). Berlin/Heidelberg, Germany: Springer (LNCS 2072).

Korel, B. & Rilling, J. (1997). Dynamic program slicing in understanding of program execution. In *Proceedings of the 5th International Workshop on Program Comprehension* (pp. 80 – 89). Washington, DC: IEEE Computer Society.

Lehman, M. M. (1980, September). Programs, life cycles and laws of software evolution. *Proceedings of IEEE: Special Issue on Software Engineering, 68*(9), 1060 – 1076.

Mens, T. & Van Gorp, P. (March 2006). A taxonomy of model transformation. *Electronic Notes in Theoretical Computer Science*, 152. Retrieved June 16, 2008, from http://www.sciencedirect.com

Mousavi Eshkevari, L., Arnaoudova, V., & Constantinides, C. (2008). Comprehension and dependency analysis of aspect-oriented programs through declarative reasoning. In P. Hudak & D. S. Warren (Eds.), In *Proceedings of the 10th International Symposium on Practical Aspects of Declarative Languages* (pp. 35 – 52). Berlin/Heidelberg, Germany: Springer (LNCS 4902).

Parnas, D. L. (1972, December). On the criteria to be used in decomposing systems into modules. *Communications of the ACM, 15*(12), 1053 – 1058.

Parnas, D. L. (1994). Software aging. In *Proceedings of the 16th International Conference on Software Engineering* (pp. 279 – 287). Los Alamitos, CA: IEEE Computer Society Press.

Parsamanesh, P., Foumani, A., & Constantinides, C. (2006). Mining anomalies in object-oriented implementations through execution traces. In J. Filipe, B. Shishkov & M. Helfert (Eds.), In *Proceedings of the International Conference on Software and Data Technologies* (pp. 177 – 189). Setubal, Portugal: INSTICC Press.

Richner, T., Ducasse, S., & Wuyts, R. (1998). Understanding object-oriented programs with declarative event analysis. In S. Demeyer and J. Bosch (Eds.), In *Object-Oriented Technology. ECOOP'98 Workshop Reader. ECOOP'98 Workshops, Demos, and Posters* (pp. 78 – 79). Berlin/Heidelberg, Germany: Springer (LNCS 1543).

Robillard, M. P. & Murphy, G. C. (2002). Concern graphs: Finding and describing concerns using structural program dependencies. In *Proceedings of the 24th International Conference on Software Engineering* (pp. 406 – 416). New York: ACM.

Robillard, M. & Murphy, G. (2007, February). Representing concerns in source code. *ACM Transactions on Software Engineering and Methodology, 16*(1), Article 3.

Rugaber, S. (1992). Program comprehension for reverse engineering. In *Proceedings of the AAAI Workshop on AI and Automated Program Understanding*.

Sneed, H. M. (January 1995). Planning the reengineering of legacy systems. *IEEE Software, 12*(1), 24 – 34.

Snelting, G. & Tip, F. (1998). Reengineering class hierarchies using concept analysis. In *Proceedings of the 6th ACM SIGSOFT International Symposium on the Foundations of Software Engineering* (pp. 99 – 10). New York: ACM.

Sommerville, I. (2007). *Software engineering* (8th ed.). Harlow, UK: Pearson Education Limited.

Szyperski, C. (2002). *Component software: Beyond object-oriented programming* (2nd ed.). Harlow, UK: Pearson Education Limited.

Tonella, P. & Ceccato, M. (2004). Aspect mining through the formal concept analysis of execution traces. In *Proceedings of the 11th IEEE Working Conference on Reverse Engineering* (pp. 112 – 121). Washington, DC: IEEE Computer Society.

Xiao, S. & Pham, C. (2004). Performing high efficiency source code static analysis with intelligent extensions. *In Proceedings of the 11th Asia-Pacific Software Engineering Conference* (pp. 346 – 355). Washington, DC: IEEE Computer Society.

Zhang, C. & Jacobsen, H.-A. (2004). PRISM is research in aspect mining. In J. M. Vlissides & D. C. Schmidt (Eds.), In *Proceedings of the Companion to the 19th Annual ACM SIGPLAN Conference on Object-Oriented Programming, Systems, Languages, and Applications* (pp. 20 – 21). New York: ACM.

Zhao, J. (2002). Slicing aspect-oriented software. In *Proceedings of the 10th International Workshop on Program Comprehension* (pp. 251 – 260). Washington, DC: IEEE Computer Society.

KEY TERMS

Control-Flow Analysis (CFA): An investigation to determine properties of the program control structure such as possible control flow paths (branches) and to find basic blocks and loops.

Data-Flow Analysis (DFA): A process for collecting run-time information about data in a computer program without actually executing it. A program's control-flow graph is deployed to determine those parts of a program to which a particular value assigned to a variable might propagate.

Formal Concept Analysis (FCA): Mathematical technique for analyzing binary relations, capturing conceptual structures among data sets.

Program Dependency Graph (PDG): A refinement of DFA for a program routine with CFA information.

Program Slicing: A derivative of Program Dependency Graph (PDG) and System Dependency Graph (SDG).

Reengineering: An activity to reimplement a software in a new form. Normally comprised by reverse engineering (for comprehension), restructuring, and forward engineering.

Rephrasing: A model transformation activity where the source and target languages are at the same level of abstraction.

Reverse Engineering: A model transformation activity where the target language is at a higher level of abstraction.

System Dependency Graph (SDG): A refinement of DFA for a program with CFA information. It captures the union of all PDGs in a program.

Translation: A model transformation activity where the source and target languages are at a different level of abstraction.

ENDNOTES

a A list of tools for refactoring is available at http://www.refactoring.com/tools.html
b http://www.eclipse.org
c A list of CASE tools is available at http://www.cs.queensu.ca/Software-Engineering/tools.html
d Currently, several proposals exist; cf. the *Proceedings of the Aspect-Oriented Modeling (AOM) Workshop (2002 – 2007)*; Jacobson, I. & Ng, P-W. (2005). *Aspect-oriented software development with use cases*. Upper Saddle River, NJ: Pearson Education, Inc.
e A list of tools for reverse engineering is available at http://www.laatuk.com/tools/documentation_tools.html.

Promotion of E-Government in Japan and Its Operation

Ikuo Kitagaki
Hiroshima University, Japan

INTRODUCTION

In Japan, e-government has been considered, since 2001, as one of the strategies of so called "e-Japan" (Ohyama, 2003). It had been decided that e-government shall be constructed within the fiscal year 2003. Preparation in terms of the legal system and technological developments made steady progress towards that goal. The construction of e-government should alleviate residents' burdens in terms of bureaucracy, enhance service quality rationalization, lean and transparent administrative agencies, countermeasures for natural calamities, more participation in policy making and administration by residents, and so forth. Various tasks have been carried out at many places. For example, in Autumn of 2002 a "basic residential register network" was established. Its initiation enjoyed broadly smooth operation. Residents had received administrative services only within certain jurisdiction limits until then. Now they are free to enjoy access to any administrative services from anywhere in Japan thanks to this e-system. Some local authorities introduced electronic tenders to enhance transparency of administration. Some local authorities adopted an electronic voting system in part of their areas. This paper explains the details of how the construction of the e-government came about and what the status of its operation is.

In constructing an e-government, basic researches in respect of relevant individual electronic chores are necessary. In reality, however, planning and drawing up an idea will often be brought about, depending on certain actual domestic social circumstances of the legal systems or certain consensus within and between relevant representative bodies of the government. Because of such circumstances, we have decided to list up general magazines easily available which report often on these themes and the most up-to-date URLs of relevant organizations of the Japanese government.

BACKGROUND

Information processing on the part of the public administration has progressed in line with each development of computers (Makiuchi, 2003). It followed the progress of information processing in the private sector. The first steps included, as early as 1970s, information processing for specified jobs such as accounting, salary payments. It also controlled systems of government offices on the basis of one PC per one person, as personal computers spread around 1990, and Internet technologies were introduced in the course of 1990s, and so forth. In particular, the Patent Office in Japan adopted an on-line system for patent application to speed up the processing of patent applications. And it did substantially increase.

Nevertheless, given the progress and more use of the information processing for administration, the mentality that it only simply meant computers would replace mechanical style jobs kept on living a long time. In addition, it exerted a negative influence in that the conventionally vertical administration confined the progress of the information processing for administration to specific ministries and offices. And it made services difficult to access from the standpoint of the people.

The Mori Cabinet came into power in 2000. He decided to shift the nation to the use of state-of-the-art information under his "e-Japan strategies". He would emphasis the importance of information processing in Japanese society in order to get out of the economic slump. Objectives were placed in five fields: broadband infrastructure, education, public/administration services, e-commerce, and security. Among the items, administration services are selected so that e-government will be realized within the fiscal year 2003. Preparation in terms of legal systems and technological developments have made progress in that respect. Legal systems were reviewed. Now, there are the On-line Transmission Regulations, Revised Basic Residential Register Law, Public Individual Certification Law, and so forth. On the technological front, there were also some new introductions of electronic systems for administrative procedures, for example all-purpose reception system for applications, the introduction of a one-stop system for import/export harbor and taxation procedures as well as the electronic tender system. Electronic systems for revenue bookings allows payment per Internet and an on-line transfer system. The purposes of the introduction of information processing for administrative chore and business operations are: simplification of work for human resources, salaries and so forth, and connection of the National Networks with LGWAN (Local Government Wide Area Network) owned by communalities.

Electronic procedures will be put into practice in various fields as described above. From their own perspectives, users consider more or less that the essential part of e-government is mainly the given opportunity of the electronic application and its ease.

E-government is one of the e-Japan Strategies in the Five Year Plan which started in 2000. Preparatory activities on legal systems and system construction are to be completed by March, 2004. On the other hand, as regards set-up of broadband infrastructure, the objective was set that 30 million households would use it by the end of the fiscal year 2005. In reality, though, 50 million households (the actual number of subscribers is about 5.70 million) were able to use it as early as June, 2002, due to the rapid progress of information telecommunication technologies. Partly as a result of such thanks to these unexpected circumstances, e-Japan Strategies could enjoy only after one and a half years of their inception another review.

SYSTEM DEVELOPMENT CASES

Here are some cases of e-governments established on national scheme or at local authorities.

Basic Residential Register Network System

Resident registration cards, tax payment certificates, and so forth, are normally applied for at the desk of an agency. The regional jurisdiction of a resident is responsible and issues them. E-government, any application at home or application and issuance from an office outside the resident district. Confirmation of the subject person beyond an administrative boundary and the necessary information exchange between administrative bodies in this respect becomes easier with the use of Basic Residential Register Network. The Basic Residential Register Law regulates undertakings for this system and how its operation should look at what its practical use is like. Information contained in this system are these four items: name, birth date, gender, and address. It also contains resident card code number and information for any changes to them. The resident card code allows an easy access to these four data items. It consist of eleven digits, selected at random, and will be issued as single and only in Japan without duplication for one person throughout Japan. This code can be changed at the request of the subject person. The system constellation is three-fold in vertical; the levels consist of national, prefecture and municipalities. Depending on the distance between the "clients" making use of the common contents from each other, the option would be made which of the three levels of networks be used. If two towns sit in the same prefecture, the communication server of one of the two towns, and the dedicated line in connection with this server on the level of the prefecture, would be used so as to come into communication with the other town magistrate. If the communication should go to the level of prefecture-prefecture, the network of this constellation level would be working plus the national network involved. Communication servers have firewalls both inside and outside to prevent illegal access (Inoue, 2003; Yoshida, Mizokami, & Watanabe, 2003)

Electronic Tender

Here is a case of Yokosuka City in Kanagawa Prefecture (Hirokawa, 2003). Up until 1997, order placement of civil works in the City took place mainly by public tender in which 7-10 specified and pre-selected bidders could join the tender. However, it became known that there had been some collusions on the bidding. It was considered that reform is necessary. The tender system was changed in the fiscal year 1998 to a conditionally open tender where any business entity could participate, if inspection standards prescribed were satisfied. In the fiscal year 1999, information on order placements started to be shown on the agency's homepage. It became possible for companies wishing to participate to confirm general picture of an order. Where is the site? What is the nature of this work? Firstly, a company wishing to bid will transmit a tender application form to the contracting section by facsimile. If confirmed that this company is qualified to participate, it will purchase design documents from a designated printing company and draw up an estimate. Then it sends the tender documents *poste restante* at the Yokosuka Post Office by registered mail with a delivery certificate by the deadline. On the date appointed, the contract section of the magistrate will open all the tender documents collected so far and decide the successful bidder with representatives of the participating companies witnessing. All the tender results will be publicized on the homepage after that. The effect is that the number of tender companies increased by 2.5 times on average before and after the reform. At the same time, the annual average success ratio (the ratio of a successful bid to the expected price) declined from 95.7% before the reform to 85.7%.

In the fiscal year 2001, the series of procedures as above were to be carried out by the Internet. The purpose is to further widen the door to bidding companies and save on administrative chore of the contract section. Companies put in a tender bid on the tender document transmission screen first and then transmit it by the Internet. The tender documents will at once be transmitted to the server of the authentication bureau, where the guarantee of originality and perception time stamps are added. Then it will be transmitted to the contract section. In other words, the function conventionally carried out at a post office has been transferred to an authentication

bureau. The average success bid ratio for this fiscal year fell as far as to 84.8%.

Noteworthy is the fact that the reform of the tender system was brought about only with the strong will on the part of the personnel. With the mayor on the top down to all those who worked on the reform who did not give up facing various forms of resistance.

Electronic Voting

For electronic voting there are two options: voting made from a household PC and another made at the ballot. Both are under examination now. The former is carried out where an electronic voting paper is obtained by a household from a voting server. Here, voting contents are filled in and transmitted to a sum-up server. However, a security problem has been pointed out in cases of electronic voting by household. Some people held the view that it should be limited to persons with physical disability. On the other hand, electronic voting at a ballot is carried out in that on voting contents decisions will be made via touch panels. Also an electronic pen can be used at the voting terminal. This has been put into practice in Niimi City, Okayama Prefecture, and part of the Hiroshima City. Various technologies for voting have been proposed, but none of them has been developed enough in any of the business in e-government. The reason appears to be that the frequency of an election itself is low. At most it is once in a year. Therefore, cost efficiency of system development is not clear as a result. Further, security must be perfect, and so forth.. A voting logic where contents, choice of an individual voter, will not be known, not only to others but also to the system itself, is under study. The system must cope with bribery and corruption, and misuse of strong power. There remain difficult issues to be solved (Kitagaki, 2003).

Various Resident Services

In Yokosuka City (Hirokawa, 2003) Geographical Information Service (GIS) was introduced to enhance and reach greater efficiency of administrative services in order to facilitate residents' participation in the regional community activities. Common use on the level of the municipal office was possible by integrating information owned by each division and bureau in the municipal office at one place, except for that information which would violate the protection of personal data. In the fiscal year 2002, Yokosuka City arranged so that citizens could obtain maps, using Web-GIS from home. That was a result of an alliance with a map company in the private sector. On the Web site, information on medical institutions, bargain sales at shopping centers, sightseeing information,, and so forth, became available. Many organizations and individuals such as the Chamber of Commerce and Industry, medical associations, SOHO workers, and so forth, are engaged in preparation of these data and operation in cooperation.

In Kochi Prefecture (Ishikawa, 2003), the information super-highway has been in operation since the fiscal 1997. Further improvement has been sought to accompany the preparation of the administrative networks further and the changes in the information environment. In the fiscal year 2001, telecommunication services at the private sector started to be carried out as well under a new basic conception. The main circuit of the information highway was changed from the previous 50Mbps to 2.4Gbps. Preservation control and stable operation throughout the year were guaranteed. In addition, main circuits were open to Internet service providers (ISP) and businesses in the private sector providing CATV services. They became compatible with IP v 6, and so forth.

Sapporo City in Hokkaido showed its basic policy in the fiscal year 2002 amid the progress of various businesses with information technology in mind. It considered a "city for cooperation between administration, citizens and enterprises". It should be an ideal state of city management. The city listens carefully to residents' voices. It started to operate a call center business to reflect this in administration. This is the center to deal with various kinds of inquiries from citizens. The city has made experiments with citizens monitoring. Trial operation has further expanded since the fiscal year 2003. On the other hand, the city attempted to distribute IC cards. For example, it has prepared the environment to be able to use them for all the municipally operated subways.

Further, payment of administrative services with IC credit cards (Endo, 2003), provision of various electronic municipality software developed with the residents' standpoint (Shimada, 2003) as well as technological issues of its operation (Maeda, Okawa, & Miyamoto, 2003) are under examination.

DISCUSSION AND CONCLUSIONS

The plan of e-Japan has been administrated during five years. Based upon the result, in January of 2006, the New Innovative Strategies was publicized. It declares that, by 2010, it will complete the innovation using IT, then change Japan to the cooperative society which is possible to develop sustainably, autonomic and enable everybody to independently participate in. To note, as for the e-government, it aims to build a society that computerizes administrative service as much as possible and to realize "the small government" that is convenient, efficient and transparent. The New Innovative Strategies consists of three parts below.

1. It is to realize the e-administration (e-government, e-local authorities) which actually feels convenience and

in its service improvement, then by 2010 to make the on-line use rate, more than 50 percent, in the process of application and notification for the government and local authorities.

2. It is to arrange the system of evaluation and installation of information systems, to arrange the evaluation method of the information system, then to optimize total task and system in all the government to realize the efficient e-government. It is to arrange the similar system in local authorities.
3. Relating to the system of government and local authorities, it is to maintain its reliability and safety while considering user's convenience, to propose the advance of security, then to foster and pervade the foremost technology through Japan's advance in e-administration.

For the objectives above, IT Strategic Headquarters also set several indicators in order to get how much the objectives would have been attained.

1. On-line use rate in application and notification and so on.
2. Time and cost which a user needs in application and notification.
3. The browsed number of the government portal site.
4. Reduced cost of IT relation. Reduced time and reduced capacity for processing the task.
5. Installation of IC card in public service and the improvement of the public service in its use.

REFERENCES

Endo, C. (2003). Credit smart card payment using NICSS token method in electronic government. *Information Processing Society of Japan*, 44(5), 489-493.

Fujisaki E., Ohta, K. & Okamoto, T. (1993). Practical Secret Voting Schemes Using Polling Places. The *Institute of Electronics,Information and Communication Engineers*, ISEC93-24, 7-21.

Hirokawa, S. (2003). Approach to the local e-government. *Information Processing Society of Japan*, 44(5), 461-467.

Inoue, M. (2003). The basic residential registers network system for e-government and e-local governments. *Information Processing Society of Japan*, 44(5), 468-472.

Ishikawa, Y. (2003). Toward the implementation of local e-government. A case in Kochi prefectual government. *Information Processing Society of Japan*, 44(5), 480-483.

Kitagaki. I. (2003). Full Confirmation Electronic Voting Model: Countermeasures to Forcible or Internal Illicit Improprieties. *Proceedings of The Information Resources Management Association*, 230-231.

Maeda, M., Okawa, Y., & Miyamoto, S. (2003). On the construction of electronic local governments from the standpoint of a technical vendor company. *Information Processing Society of Japan*, 44(5), 499-502.

Makiuchi, K. (2003). E-government, version 2. *Information Processing Society of Japan*, 44(5),461-467.

Ohyama, N. (2003). Progress of e-government in Japan. *Information Processing Society of Japan*, 44(5),455-460.

Segawa, M. (2003). IT turns to realization of e-collaboration city. *Information Processing Society of Japan*, 44(5), 484-488.

Shimada, H. (2003). For the early realization of eletronic local government. *Information Processing Society of Japan*, 44(5), 494-498.

Yoshida, T., Mizokami, M., & Watanabe, T. (2003). Electronic filing to governments. *Information Processing Society of Japan*, 44(5), 473-475.

URLs

The Strategic Headquarters for the Promotion of an Advanced Information and Telecommunications Network Society (January 22, 2001) e-Japan strategies. [WWW document]. URL http://www.kantei.go.jp/jp/singi/it2/dai1/1siryou05_2.html

The official residence of the Prime Minister (February 12, 2004). The Liaison Conference of Chief Information Officers (CIO) from Representative Bodies of the Government. [WWW document]. URL http://www.kantei.go.jp/jp/singi/it2/cio/index.html

Administrative Management Bureau, Ministry of Public Management, Home Affairs, Posts and Telecommunications (February 16, 2004). The general window of e-Government, [WWW document]. URL http://www.e-gov.go.jp/

Federal Citizen Information Center, Office of Citizen Services and Communications U.S. General Services Administration (February 16, 2004). *The U.S. Govenments Official Web Portal* [WWW document]. URL http://www.firstgov.gov.

IT Strategic Headquarters (Jan. 19, 2006), the New Innovative Strategies [WWW document].http://www.kantei.go.jp/jp/singi/it2/kettei/060119honbun.pdf

KEY TERMS

Basic Residential Register Network: A network to confirm the being of a person, a subject. Its use is common throughout Japan and jointly operated by local authorities.

E-Government: The government and local authorities which construct an electronic processing system for administrative procedures with full command of information telecommunication technologies. It provides various kinds of services for residents in line with these.

E-Japan Strategies: The national strategies decided in the first conference of the Strategic Headquarters for the Promotion of an Advanced Information and Telecommunications Network Society (the IT Strategic Headquarters) held on the January 22, 2001. It was laid on the IT basic strategies decided at the Joint Conference of the IT Strategic Conference and the Strategic Headquarters for Information Telecommunications (IT) Technologies (November 27, 2000).

Electronic Tender System: A system to carry out a series of works from notification of order placement information on the homepage,, and so forth, application for tender participation, sending a tender document, opening tender documents to the public announcement of the result.

Electronic Voting System: A system to vote electronically, using an information terminal of a personal computer, and so forth. Various technologies have been created to prevent illegal actions such as alteration of voting contents or the use of abstainers' votes.

On-Line Communication Regulation Law: A law to enhance application in the administration, which is decided by laws exceeding 50,000, and cover all these laws for conducting application electronically. It is considered that this communication regulation law has completed the basic legal frameworks necessary for the e-government such as electronic signature law.

Public Individual Certification Law: A law to prepare a system to provide individual certification and authorization services in order to process on-line the public application and notification on-line.

Proxy Caching Strategies for Internet Media Streaming

Manuela Pereira
University of Beira Interior, Portugal

Mário M. Freire
University of Beira Interior, Portugal

INTRODUCTION

Media streaming consists in the viewing of dynamic media information while being downloaded by clients. With the explosive growth of the Web and the mature of digital video technology, *media streaming* has received a great deal of interest as a promising solution for multimedia delivery services. This approach allows that media objects can be accessed in a similar way to conventional text and images using a download-and-play mode. However, unlike static text-based content, proxy caching has difficulty in delivering streaming media content because media objects are usually very large and its transmission consumes a great amount of network resources, prolongs startup latency, and threatens the playback continuity. The size of a conventional *Web object* is typically on the order of 1–100 kbytes and, therefore, a decision regarding either caching or not an object in its totality is an easy task (Liu & Xu, 2004). However, the size of *media objects* is very large, reaching a size on the order of several hundreds of Mbytes or even Gbytes. Therefore, caching a whole media object at a Web proxy optimized for delivering conventional small-size Web objects is not feasible, since large streams would quickly exhaust the capacity of the proxy cache. Besides, the streaming of *media objects* requires a significant amount of resources such as disk space and network bandwidth, which need to be maintained during a long period of time. Moreover, the long playback duration of a streaming may allow several client-server interactions. Therefore, access rates might be different for different parts of a stream, which makes cache management potentially more complex, as pointed out by Liu and Xu (2004). On the other hand, a download-before-playing solution provides continuous playback, but it also introduces a large startup delay.

An effective solution to reduce client-perceived latencies and network congestion is to cache data at proxies widely deployed across the Internet. This solution, besides inexpensive, also leads to an improvement of both availability of objects and packet losses since redundant network transmission decreases while transmission efficiency increases. However, proxies are generally optimized for delivering conventional small-size Web objects, which may not satisfy the requirements of streaming applications. Due to these particular features of media objects, novel caching strategies have been proposed.

With the evolution of the Internet as the dominant architecture for applications, contents, and services, these are gradually migrating from the *client-server paradigm* to the edge services paradigm and to the peer-to-peer (P2P) computing paradigm. Recently, P2P system has received a great amount of interest as a promising scalable and cost-effective solution for next-generation multimedia content distribution. This kind of systems have advantages regarding systems based on the client-server paradigm, namely improved scalability and reliability, cheaper infrastructures due to direct communication among peers, and easiness of resource aggregation in order to provide, for instance, massive processing power (Ye, Makedon, & Ford, 2004). However, P2P systems also have some drawbacks, namely the considerably more complex searching and node organization and security issues (Aberer, Punceva, Hauswirth, & Schmidt., 2002). Therefore, this article limits the discussion to low-cost proxy caching strategies for *media streaming* over Internet.

BACKGROUND

As discussed earlier, media caching has different requisites regarding conventional Web caching due to the special features of media streaming. Since the content of a media object is rarely updated, management issues like cache consistency and coherence are less critical in media caching. However, it requires an effective management of proxy cache resources due to the resource requirements of media objects (Liu & Xu, 2004).

Roughly, there are two main types of caching strategies: the strategies focused on homogeneous clients and the strategies focused on heterogeneous clients. Most of the proposed strategies are focused on homogeneous clients, which have identical or similar configurations and capabilities behind a proxy. Figure 1 presents an overview of caching strategies

for media streaming. A brief description of these strategies is provided in the next sections.

CACHING STRATEGIES FOR HOMOGENEOUS CLIENTS

Strategies for homogeneous clients can be classified, regarding the parts of media objects to cache, as *prefix caching*, *sliding-interval caching*, *segment-based caching*, and *rate-split caching*. A brief description of these strategies follows.

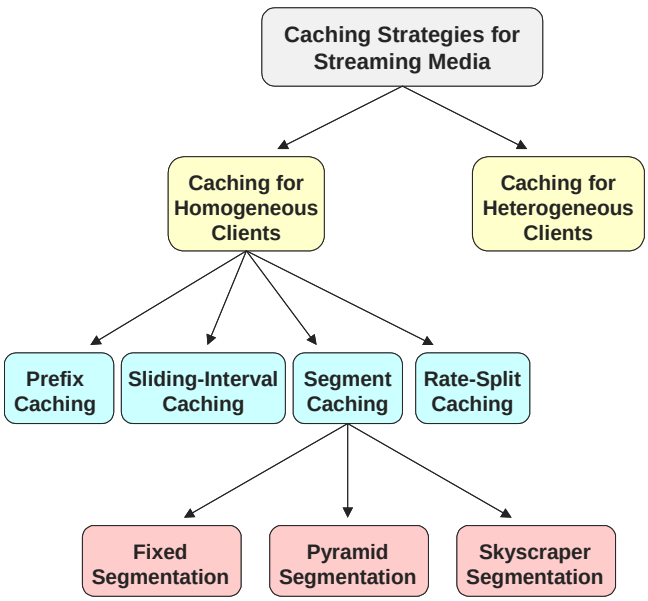
Prefix Caching

According to this strategy, the media object is divided into two parts: the prefix and the suffix. The prefix is cached at a proxy. After the reception of a client request, the proxy immediately delivers the prefix to the client and fetches the suffix from the source server to be further delivered to the client. This strategy leads to a significant reduction of the startup delay for a playback since the proxy is generally closer to the clients than the source server (Liu & Xu, 2004; Miao & Ortega, 1999; Sen, Rexford, & Towsley, 1999). In this strategy, the prefix size is a key issue for the system performance. In general, this strategy leads to a moderate bandwidth reduction but to a high startup latency reduction.

Sliding-Interval Caching

According to this strategy, a sliding interval of a media object is cached in order to exploit the sequential access of a streaming media. For instance, if two consecutive requests for the same object are received, the first request may access the object from the server and incrementally store it into the proxy cache, while the other request only needs to access the cached portion and release it after the access. In general, if multiple requests for a given object are received in a short period of time, a set of adjacent intervals may be grouped to form a run, of which the cached portion will be released only after the satisfaction of the last request. This strategy can substantially reduce the consumption of network bandwidth and the startup delay for subsequent accesses. However, it involves high disk bandwidth utilization because the cached portion is dynamically updated with the playback. Besides, the effectiveness of the sliding-interval caching strategy diminishes with increased access intervals. Moreover, in the case of the access interval of a given object be longer than the duration of the playback, it degenerates to the case of full-object caching (Liu & Xu, 2004; Tewari, 1998). These limitations may be mitigated if the cached content is retained over a relatively long period of time. Nevertheless, for a good cache design, this strategy may lead to a high bandwidth reduction and to a high startup latency reduction.

Figure 1. Overview of caching strategies for media streaming



Segment-Based Caching

Segment-based caching is a generalization of the prefix caching strategy by partitioning a media object into a set of segments, differentiating their respective utilities, and making a caching decision accordingly (Liu & Xu, 2004). Several segment-based caching approaches have been proposed (Chen, Wang, Zhang, Shen, & Wee, 2005; Wu, Yu, & Wolf, 2004). Here, we consider the fixed segmentation (also called uniform segmentation), the pyramid segmentation (also called exponential segmentation), and the skyscraper.

In the fixed segmentation approach, the size of segments is fixed. In the pyramid segmentation (Viswanathan & Imielinsky, 1996; Wu, Yu, & Wolf, 2004), segment size increases exponentially from the first segment. For instance, objects are segmented in a form that the size of a succeeding segment can double the size of its preceding one. Therefore, in this approach, the size of segments increases exponentially. This approach favors the caching of first segments of media objects due to the following reasons (Wu, Yu, & Wolf, 2004): First, the last portion of a media object is generally the least valuable for caching; second, it is more cost effective to quickly remove a large portion of a cached media object whose caching priority typically decreased from the beginning of the viewing. A hybrid approach taking into account both uniform lengths and exponentially increasing lengths has been considered by Chae, Guo, Buddhikot, Suri, and Zegura (2002). The skyscraper segmentation approach (Wu, Yu, & Wolf, 2004) is a variation of the pyramid segmentation, in which the segment size increases more slowly. For example, the size of a given segment is either the same as or twice of the previous segment. Skyscraper segmentation leads to a larger number of segments than pyramid segmentation, but it has less segments than fixed segmentation.

Using event-driven simulation, Wu, Yu, and Wolf (2004) evaluated the three segmentation approaches and have compared them with a full-object approach and with a prefix caching approach. They have shown that: 1) the three segment-based caching strategies are more effective in not only increasing the byte-hit ratio (i.e., reducing total traffic) but also lowering the fraction of requests that require delayed start; 2) pyramid segmentation is the best approach among the three segment-based caching strategies; and 3) segment-based caching strategies are particularly effective when the cache size is limited, when the set of hot media objects changes over time, when requests spread over a large number of media objects, when the size of the media file is large, and when there are a large number of distinct media objects (Wu, Yu, & Wolf, 2004).

Rate-Split Caching

Unlike the previously discussed caching strategies, in which the partition of a media object occurs horizontally along the time axis, the rate-split caching strategy partitions it vertically along the rate axis. It is established as a cutoff rate, whose part above the cutoff rate will be cached at the proxy, while the part below the cutoff rate will remain stored at the origin server. This type of partitioning is particularly attractive for variable bit rate streaming, since only the lower part of a nearly constant rate has to be delivered through the entire network path. A critical issue in this strategy is the selection of the cutoff rate (Liu & Xu, 2004; Zhang, Wang, Du, & Su, 2000). In general, this strategy leads to a moderate bandwidth reduction and to a moderate startup latency reduction.

CACHING STRATEGIES FOR HETEROGENEOUS CLIENTS

Due to diversity of network environments and device configurations, clients behind the same proxy often have quite different requirements for the same media object in terms of streaming rates or encoding formats. In order to accommodate such heterogeneity, a simple solution is to produce replicated streams of different rates or formats, each targeting the requirements of a subset of clients. Although widely used in commercial streaming systems, this approach is unattractive due to the storage and bandwidth demands (Liu, Chu, & Xu, 2004). An alternative strategy is based on the transcoding of a media object from one form to another of a lower rate or a different encoding format on demand (Tang, Zhang, & Chanson, 2002). However, the large computational overhead of transcoding limits the capability of the proxy to support a large and diverse client population (Liu & Xu, 2004).

A more efficient approach to this problem is based on the use of layered encoding and transmission. A layered coder compresses a media object into some layers. The most significant layer, the base layer, contains the data representing the most important features of the object, whereas additional enhancement layers contain data that can progressively refine the quality. Thus, a client can subscribe to a subset of cumulative layers to reconstruct a stream according to its capability or according to the desired quality (Kangasharju, Haranto, Reisslein, & Ross, 2002; Liu & Xu, 2004).

FUTURE TRENDS

Recently, two new approaches for segment-based caching have been reported: adaptive and lazy segmentation and active prefetching (Chen et al., 2005). However, those approaches need to be improved in a cost-effective way.

Proxy caching strategies discussed earlier are focused on unicast delivery. Although the discussed strategies can effectively reduce access latencies and network bandwidth, the scalability of proxy caching strategies is limited for media

content delivery. One strategy to alleviate this problem is the use of proxy caching with multicasting (Ramesh, Rhee, & Guo, 2001). However, this approach may lead to larger startup latencies and is geographically limited (Liu & Xu, 2004). Recently, Xu, Guo, Pan, and Wang (2004) proposed a dynamic cache-multicast algorithm for streaming media, which seems to effectively reduce the utilization of network resources and enhances the byte hit ratio of a proxy cache. An alternative to overcome these limitations is to purchase the services of proprietary content delivery networks (CDNs). However, besides expensive, this solution may have a limited performance for large-scale multimedia distribution services (Xiang, Zhang, Zhu, Zhang, & Zhang, 2004).

An alternative approach is based on the peer-to-peer content distribution in which a proxy and its clients can be structured into a peer-to-peer system (Guo et al., 2004) or where clients cooperate to distribute content (Padmanabhan, Wang, Chow, & Sripanidkulchai, 2002) using advanced video coding schemes such as multiple description coding (Pereira, Antonini, & Barlaud, 2003).

CONCLUSION

A discussion about particular requirements of media objects for media streaming was presented. Proxy caching strategies for media streaming have been discussed for both homogenous and heterogeneous clients. Regarding homogeneous clients, particular attention was paid to the following approaches: prefix caching, sliding-interval caching, segment-based caching, and rate-split caching. Advantages and limitations of these strategies were pointed out. A discussion about limitations of caching strategies and their evolution was also provided.

REFERENCES

Aberer, K., Punceva, M., Hauswirth, M., & Schmidt, R. (2002). Improving data access in P2P systems. *IEEE Internet Computing, 6*(1), 58-67.

Chae, Y., Guo, K., Buddhikot, M. M., Suri, S., & Zegura, E. W. (2002). Silo, rainbow, and caching token: schemes for scalable fault tolerant stream caching. *IEEE Journal on Selected Areas in Communications, 20*(7), 1328-1344.

Chen, S. Wang, H., Zhang, X., Shen, B.,& Wee, S. (2005). Segment-based proxy caching for Internet streaming media delivery. *IEEE Multimedia, 12*(3), 59-67.

Guo, L, Chen, S., Ren, S., Chen, X., & Jiang, S. (2004). PROP: A scalable and reliable P2P assisted proxy streaming system. In *Proceedings of the 24th International Conference on Distributed Computing Systems (ICDCS'04),* Tokyo, Japan.

Kangasharju, J., Hartanto, F., Reisslein, M., & Ross, K.W. (2002). Distributing layered encoded video through caches. *IEEE Transactions on Computers, 51*(6), 622-36.

Liu, J., Chu, X., & Xu, J. (2004). Proxy cache management for fine-grained scalable video streaming. *Proceedings of IEEE INFOCOM 2004*, Hong Kong, China.

Liu, J., & Xu, J. (2004). Proxy caching for media streaming over the Internet. *IEEE Communications Magazine, 42*(8), 88-94.

Miao, Z., & Ortega, A. (1999, March 31-April 2). Proxy caching for efficient video services over the Internet. *Proceedings of International Web Caching Workshop*, San Diego.

Padmanabhan, V. N., Wang, H. J., Chou, P. A., & Sripanidkulchai, K. (2002, May 12-14). Distributing streaming media content using cooperative networking. *Proceedings of NOSSDAV'02*, Miami, FL.

Pereira, M., Antonini, M., & Barlaud, M. (2003). Multiple description coding for Internet video streaming. *Proccedings of IEEE International Conference on Image Processing (ICIP 2003)*, Barcelona, Spain.

Ramesh, S., Rhee, I., & Guo, K. (2001, April). Multicast with cache (Mcache): An adaptive zero-delay video-on-demand service. *Proceedings of IEEE INFOCOM 2001*, Anchorage, AK.

Sen, S., Rexford, J., & Towsley, D. (1999). Proxy prefix caching for multimedia streams. *Proceedings of IEEE INFOCOM 1999*, New York.

Tang, X., Zhang, F., & Chanson, S. T. (2002). Streaming media caching algorithms for transcoding proxies. *Proceedings of the 31st International Conference on Parallel Processing (ICPP)* (pp. 287-295). IEEE Computer Society Press.

Tewari, R. (1998, January). Resource-based caching for web servers. *Proceedings of MMCN '98*, San Jose, CA.

Viswanathan, S., & Imielinski, T. (1996). Metropolitan area video-on-demand service using pyramid broadcasting. *Multimedia Systems, 4*(4), 197-208.

Wu, K. L., Yu, P. S., & Wolf, J. L. (2004). Segmentation of multimedia streams for proxy caching. *IEEE Transactions on Multimedia, 6*(5), 770-780.

Xiang, Z., Zhang, Q., Zhu, W., Zhang, Z., & Zhang, Y. (2004). Peer-to-peer based multimedia distribution service. *IEEE Transactions on Multimedia, 6*(2), 343- 355.

Xu, Z., Guo, X., Pang, Y., & Wang, Z. (2004, October). The dynamic cache-multicast algorithm for streaming me-

dia. In P. Chemouil, M. M. Freire, A. Gravey, & P. Lorenz (Eds.), *Universal Multiservice Networks, Lecture Notes in Computer Science, LNCS 3262* (pp. 275-284). Berlin Heidelberg: Springer-Verlag.

Ye, S., Makedon, F., & Ford, J. (2004). Collaborative automated trust negotiation in peer-to-peer systems. *Proceedings of the Fourth International Conference on Peer-to-Peer Computing (P2P2004)*.

Zhang, Z.-L., Wang, Y., Du, D.H.C., & Su, D. (2000). Video staging: A proxy-server-based approach to end-to-end video delivery over wide-area networks. *IEEE/ACM Trans. Net., 8*(4), 429-42.

KEY TERMS

Content Delivery Networks (or Content Distribution Networks (CDNs): In this kind of networks, the origin server is replicated and placed locally close to the clients or remotely in suitable geographical or network spaces.

Media Streaming: Consists in the viewing of dynamic media information while being downloaded by clients

MPEG: Moving Picture Experts Group.

Multicast: The sender generates only a single data stream that will be transmitted to selected multiple receivers who have joined the appropriate multicast group. A multicast-enabled router will forward a multicast message to a particular network only if there are multicast receivers on that network.

Peer-to-Peer System: This term refers to any exchange system characterized by direct interaction and data exchange between its peers.

ProxyCache: Memory to store media objects.

Web: Also referred as World Wide Web, it is a multimedia system based on hypertext applications that use the client/server model to access to Internet resources.

Qualitative Methods in IS Research

Eileen M. Trauth
The Pennsylvania State University, USA

INTRODUCTION

As information technologies have evolved, so too has our understanding of the information systems that employ them. A significant part of this evolving understanding is the role of the human contexts within which information systems are situated. This, in turn, has led to the need for appropriate methods of studying information systems in their context of use. Just as decisions about information systems need to be considered within their contexts of use, so also do choices about the appropriate research methodologies to employ for studying them. Increasingly, qualitative methods are chosen as an appropriate method for studying contextual aspects of information systems development, use and impact.

Qualitative research refers to research methods that engage in the interpretation of words and images rather than the calculation of numbers. These methods include: ethnography, case study, action research, interviews, and text analysis (i.e., conversation analysis, discourse analysis, and hermeneutics). Qualitative research can be theory-driven in much the same way as quantitative analysis. However, it can also employ grounded theory techniques in order to develop theory (Glaser & Strauss, 1967).

Following some early uses of qualitative methods in the 1980s (e.g., Benbasat et al., 1987; Kaplan & Duchon, 1988; Lee, 1989; Mumford et al., 1985), there has been a significant growth in the use of qualitative methods for information systems research since the 1990s (e.g., *Journal of Information Technology*, 1998; Lee et al., 1997; *MIS Quarterly,* 1999, 2000; Nissen et al., 1991; Trauth, 2001).

Accompanying the increased use of qualitative methods for IS research has been a discussion of various methodological issues. Among the key aspects of this dialogue are discussions about the suitability of qualitative methods for various types of research and issues arising from a particular type of qualitative methods: interpretive methods. This article presents a reflection on some these discussions in the form of a consideration of five factors that can influence the choice of qualitative (particularly interpretive) methods for information systems research.

FACTORS INFLUENCING THE DECISION

The Research Problem

The research problem, *what* one wants to learn, should determine *how* one goes about learning it. Heaton (1998) chose observation, interview and document analysis to examine the social construction of computer-supported cooperative work in two different cultures in order to learn how the meaning of "culture" was reflected in the design of systems. Trauth (2000) used ethnographic methods to explore the influence of socio-cultural factors on the development of a nation's information economy. Bentley et al.'s (1992) ethnographic study of the work practices of air traffic controllers informed their design of an interface to an air traffic control database. Walsham and Sahay (1999) conducted extensive interviews to gain an in-depth understanding of the implementation of geographical information systems for administrative applications in India. Phillips (1998) employed public discourse analysis to reveal the way in which concerns about anonymity, surveillance, and privacy are integrated into public understanding of a consumer payment system.

The Epistemological Lens

Orlikowski and Baroudi (1991) considered the influence of the epistemological lens – positivist, interpretive or critical – on the conduct of IS research. While there is some positivist, qualitative IS research (e.g., Lee 1989), most qualitative IS research is either interpretive or critical because of the assumption that "our knowledge of reality is a social construction by human actors" that precludes obtaining objective, value-free data (Walsham, 1995, p. 376). The interpretive epistemology has also spawned IS research employing hermeneutic methods (e.g., Boland, 1985, 1991, and Trauth & Jessup, 2000). Ngwenyama and Lee (1997) used the critical lens to examine information richness theory.

The Uncertainty Surrounding the Phenomenon

According to Galliers and Land (1987), the added complexity from including relations with people and organizations in a view of information systems introduces greater imprecision

and the potential for multiple interpretations of the same phenomenon. Hence, alternatives to quantitative measurement are needed. Others argue that the less that is known about a phenomenon the more difficult it is to measure it. Benbasat et al. (1987) explained that the case study approach is appropriate for IS research areas in which few previous studies have been carried out. Paré and Elam (1997) built theories of IT implementation using case study methods. Orlikowski's (1993) rationale for choosing qualitative methods and grounded theory to study the adoption of CASE tools rested on the absence of systematic examination of the organizational changes accompanying the introduction of CASE.

The Researcher's Skills

The absence of formal study of qualitative methods may serve as a barrier to choosing these methods. Orlikowski (1991) suggested that institutional conditions have inhibited the teaching of qualitative methods because of the functionalist/positivist perspective of American business schools where IS is typically taught. These institutional conditions, within which doctoral studies are conducted and dissertations are written, have inhibited the use of alternative research paradigms and methodologies with long-term implications for the choice of methods used in IS research. Schultze's (2001) reflection on her decision to choose interpretive methods for her dissertation illustrates the importance of institutional influence. Exposure to advisors with expertise in interpretive methods gave her methodological opportunities not available to other students.

The Academic Politics

The choice of research methods is influenced by the country in which one works, whether one has completed the PhD, whether one has a tenured position, one's academic rank, and the particular inclinations of the university at which one works. The norms and values of the field are reinforced during one's education and beyond. What is taught in research methods seminars sets the standard for "acceptable" research. Advice to junior faculty, peer review of journal articles and the tenure review process all reinforce those norms and values. Fitzgerald and Howcroft (1998) described the polarization of positions into "hard" and "soft" perspectives. Klein and Myers (1999) contributed to closing this methodological divide by developing a set of principles for conducting and evaluating interpretive field studies.

FUTURE TRENDS

As our understanding of the context of information systems grows, our desire to understand and explain contextual factors will motivate researchers to explore new ways to employ qualitative methods. Therefore, we can expect greater use of the critical epistemological lens in the use of qualitative methods in IS research. We can also expect to see the increased use of *virtual* qualitative research methods. That is, the traditional face-to-face methods of data generation used in qualitative research will find increasing analogues in the virtual world. We can expect to see, for example, "virtual ethnographies," "virtual participant observation" and "online interviews".

CONCLUSION

The primary advantage of using qualitative, particularly interpretive, methods is the flexibility it affords the researcher during data generation and analysis. The main disadvantage of qualitative, particularly interpretive, methods is overcoming concerns about validity and generalization of findings. The concepts of both statistical validity and statistical generalization need to be redefined for qualitative research. The *MIS Quarterly* "Special Issue on Intensive Research" has addressed the validity issue by publishing exemplar research papers that provide evaluative criteria for other researchers to use in establishing validity. The generalizability issue is being addressed in thoughtful pieces such as the recent article by Lee and Baskerville (2003).

Despite these issues, the acceptance of qualitative methods for IS research is evidence of a growing consensus that these methods make a valuable contribution to the study of information systems in context. In making the decision to use qualitative methods a number of factors must be taking into consideration. These factors relate to the characteristics of the research problem, the researcher and the research environment.

REFERENCES

Benbesat, I., Goldstein, D.K., & Mead, M. (1987). The case research strategy in studies of information systems. *MIS Quarterly, 11*(3), 369-386.

Bentley, R., Hughes, J.A., Randall, D., Rodden, T., Sawyer, P., Shapiro, D., & Sommeville, I. (1992). Ethnographically-informed systems design for air traffic control. In J. Turner & R. Kraut (Eds.), *Sharing perspectives: Proceedings of ACM Conference on Computer-Supported Cooperative Work* (pp.123-129). New York: ACM Press.

Boland, R.J. (1985). Phenomenology: A preferred approach to research on information systems. In E. Mumford, R.A. Hirschheim, G. Fitzgerald & T. WoodHarper (Eds.), *Research methods in information systems* (pp.193-201). Amsterdam: NorthHolland.

Doolin, B. (1998). Information technology as disciplinary technology: Being critical in interpretive research on information systems. *Journal of Information Technology, 13*(4), 301-312.

Fitzgerald, B., & Howcroft, D. (1998). Towards dissolution of the IS research debate: From polarization to polarity. *Journal of Information Technology, 13*(4), 313-326.

Galliers, R.D., & Land, F.F. (1987). Choosing appropriate information systems research strategies. *Communications of the ACM, 30*(11), 900-902.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory.* Chicago: Aldine Publishing Co.

Heaton, L. (1998). Talking heads vs. virtual workspaces: A comparison of design across cultures. *Journal of Information Technology, 13*(4), 259-272.

Journal of Information Technology. (1998). *Special Issue on Interpretive Research in Information Systems, 13*(4).

Kaplan, B., & Duchon, D. (1988). Combining qualitative and quantitative methods in information systems research: A case study. *MIS Quarterly, 12*(4), 571-586.

Klein, H.K., & Myers, M.D. (1999). A set of principles for conducting and evaluating interpretive field studies in information systems. *MIS Quarterly, 23*(1), 67-93.

Lee, A.S. (1989). A scientific methodology for MIS case studies. *MIS Quarterly, 13*(1), 33-50.

Lee, A.S., & Baskerville, R.L. (2003). Generalizing generalizability in IS research. *Information Systems Research, 14*(3), 221-243.

Lee, A.S., Liebenau, J., & DeGross, J.I. (Eds.). (1997). *Information systems and qualitative research.* London: Chapman & Hall.

MIS Quarterly. (1999). *Special Issue on Intensive Research, 23*(1).

MIS Quarterly. (2000). *Special Issue on Intensive Research, 24*(1).

Mumford, E., Hirschheim, R.A., Fitzgerald, G., & Wood-Harper, T. (Eds.). (1985). *Research methods in information systems.* Amsterdam: NorthHolland.

Ngwenyama, O.K., & Lee, A.S. (1997). Communication richness in electronic mail: Critical social theory and the contextuality of meaning. *MIS Quarterly, 21*(2), 145-167.

Nissen, H.-E., Klein, H.K., & Hirschheim, R. (Eds.). (1991). *Information systems research: Contemporary approaches and emergent traditions.* Amsterdam: North-Holland.

Orlikowski, W.J. (1991). *Relevance versus rigor in information systems research: An issue of quality — the role of institutions in creating research norms.* Panel Presentation at the IFIP 8.2 Working Conference on the Information Systems Research Challenges, Copenhagen, Denmark.

Orlikowski, W.J. (1993). CASE tools as organizational change: Investigating incremental and radical changes in systems development. *MIS Quarterly, 17*(3), 309-340.

Orlikowski, W.J., & Baroudi, J.J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information Systems Research, 2*(1), 1-28.

Paré, G., & Elam, J.J. (1997). Using case study research to build theories of IT implementation. In A.S. Lee, J. Liebenau & J.I. DeGross (Eds.), *Information systems and qualitative research* (pp. 542-568). London: Chapman & Hall.

Phillips, D. (1998). The social construction of a secure, anonymous electronic payment system: Frame alignment and mobilization around ecash. *Journal of Information Technology, 13*(4), 273-284.

Schultze, U. (2001). Reflexive ethnography in information systems research. In E.M. Trauth (Ed.), *Qualitative research in IS: Issues and trends* (pp. 78-103). Hershey, PA: Idea Group Publishing.

Trauth, E.M. (2000). *The culture of an information economy: Influences and impacts in the Republic of Ireland.* Dordrecht, the Netherlands: Kluwer Academic Publishers.

Trauth, E.M. (2001). *Qualitative research in IS: Issues and trends.* Hershey, PA: Idea Group Publishing.

Trauth, E.M., & Jessup, L. (2000). Understanding computer-mediated discussions: Positivist and interpretive analyses of group support system use. *MIS Quarterly, 24*(1), 43-79.

Walsham, G. (1995). The emergence of interpretivism in IS research. *Information Systems Research, 6*(4), 376-394.

Walsham, G., & Sahay, S. (1999). GIS for district-level administration in India: Problems and opportunities. *MIS Quarterly, 23*(1), 39-65.

KEY TERMS

Case Study: An examination of a phenomenon in its natural setting using fixed boundaries such as time.

Critical Research: Critique of the status quo through the exposure of what are believed to be deep-seated, structural contradictions within social systems.

Ethnography: Research characterized by an extended period in the field and which involves the researcher being immersed in the community being studied.

Grounded Theory: A method used to systematically derive theories of human behavior from empirical data.

Interpretive Research: Exploring the deeper structure of a phenomenon within its cultural context by examining the subjective meanings that people create.

Positivist Research: Research premised on the existence of a priori fixed relationships among phenomena.

Qualitative Methods: Methods for data collection and analysis that focus on understanding of text and observation rather than numeric calculation.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2378-2381, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Qualitative Spatial Reasoning

Shyamanta M. Hazarika
Tezpur University, India

INTRODUCTION

Artificial Intelligence (AI) has, as one of its central topics, the ability to represent and reason with *common sense* knowledge. Early forays into common sense reasoning about the physical world involved solving textbook problems on physics and mathematics. These were not adequate for reasoning about most commonplace physical scenarios.

A system suggested by DeKleer, involving both quantitative knowledge and qualitative information concerning the physical situation marked the starting point for *qualitative physics* (Weld & DeKleer, 1990). Hayes' *Naive Physics Manifesto* (Hayes, 1985) paved the way for establishing qualitative physics (meantime re-christened *qualitative reasoning*) as an important topic of research within AI.

Qualitative Reasoning (QR) is an approach for dealing with common sense knowledge without recourse to complete quantitative knowledge. Representation of knowledge is through a limited repository of *qualitative abstractions*.

Space and spatial change is an important part of common sense reasoning. *Naive Physics Manifesto* proposed to represent space-time with four-dimensional *histories*. Despite early forays such as the *Naive Physics Manifesto*, representation of space within QR has been ill addressed. Nevertheless, there has been an increasing interest over the last few years in *qualitative spatial reasoning* - reasoning about space using qualitative abstractions.

BACKGROUND

Qualitative Spatial Representation and Reasoning is concerned with providing calculus which allow a machine to represent and reason with spatial entities without resort to traditional quantitative techniques. Reasoning is concerned with methods and techniques for decision-making using spatial knowledge and developing efficient algorithms for doing so. The term *Qualitative Spatial Reasoning* (QSR) subsumes both the sub-fields of representation and reasoning.

Conventional mathematical theories of space consider points as primitive spatial entities. Within QSR there is a strong tendency to take regions of space as the primitive spatial entity. The nature of the embedding space, that is, the universal spatial entity, is another important ontological commitment. One might take this to be R^n for some n, but one can imagine applications where discrete, finite or non-convex universes might be useful.

QUALITATIVE SPATIAL REASONING

Within QSR, *qualitative spatial representations* addressing different aspects of space including topology, orientation, shape, size, and distance have been put forward.

Different Approaches to QSR

Topology

Topology is the most elemental aspect of space and holds promise as a fundamental facet of qualitative spatial reasoning. Mathematical topology is too abstract to be of relevance to those attempting to formalize common sense spatial reasoning. QSR is concerned with reasoning and not just representation, and this has been paid little attention in mathematics.

One existing approach to topology, which has been espoused by QSR, is the work to be found in philosophical logic (Clarke, 1981; De Laguna, 1922). This work has built axiomatic theories of space which are predominantly topological in nature, and which take regions rather than points as primitive. In particular, the work of Clarke (1981) has led to the development of the *Region Connection Calculus* (Cohn, Bennett, Gooday, & Gotts, 1977; Randell, Cui & Cohn, 1992) and has also been a basis for theory of common sense geometry (Asher & Vieu, 1995).

Mereotopology

Mereology is the theory of parts and whole. Mereology is not sufficient by itself and there are theories in the literature, which have proposed integrating topology and mereology. The notion of *connection*, which is the key topological notion for the qualitative description of space, cannot be defined in terms of the mereological *part-whole* relation alone. Therefore, topological notions have to be added to mereology to provide an adequate qualitative theory of space. Such combination of the disciplines of mereology and topology is referred to as *mereotopology.*

Orientation

Orientation relations describe where objects are placed relative to one another. Of the qualitative orientation calculi to be found in the literature, certain calculi have an explicit triadic relation (Freksa, 1992), while others presuppose an extrinsic frame of reference (Frank, 1992).

Recently Dehak, Bloch, & Maitre (2005) have described a probabilistic method of inferring the position of a point with respect to a reference point knowing their relative spatial position to a third point. They address this problem in the case of incomplete information, where only the angular spatial relationships are known.

Distance and Size

Qualitative representation of distance is based on either some *absolute* scale or some kind of *relative* measurement. De Laguna's *Geometry of Solids* (De Laguna, 1922) is the earliest among the *relative* kind of representations. Distance is closely related to the notion of orientation, for example, distances cannot usually be summed unless they are in the same direction. It is perhaps not surprising that there have been a number of calculi, which are based on *positional information*: a primitive, which combines distance and orientation information.

Shape

Shape is an important characteristic of an object, and particularly difficult to describe qualitatively. Qualitative formalisms for describing shape can either be *constructive representations* or certain *constraining approaches*. Within the constructive representation of qualitative shape, structured combinations of primitive entities describe complex shapes. Approaches that work by describing the boundary of an object include sequence of different types of *curvature extrema* (Leyton, 1988). Meathrel & Galton (2000) present a general theory of qualitative outlines in 2D.

Region-Based Theories of Space

Early Theories

De Laguna's *Geometry of Solids* (De Laguna, 1922) is based on a triadic primitive CanConnect(x,y,z): x can connect y and z. CanConnect(x,y,z)is true if a body x can connect y and z by simple translation i.e., without scaling, rotation or shape change. The primitive is extremely expressive and it is easy to define notions such as *connectedness* and *relative distance* measures. Mereology as understood today is a formulation due to Tarski (1959) and is built on the single primitive relation P(x,y): x is a part of y. Tarski gave a theory of the *Geometry of Solids*, embedded by means of definition into an axiomatization of elementary Euclidean geometry.

Clarke's Calculus of Individuals

Clarke's formalism is based on connectedness (Clarke, 1981). Clarke took as his primitive C(x,y): the notion of two regions x and y being connected. C(x,y)is axiomatized to be reflexive and symmetric. From the C(x,y) relation, Clarke defines the relation of *part to whole* and several other useful spatial relations as enumerated in Table 1 below.

Region Connection Calculus

The Region Connection Calculus (RCC) is a modification and development of Clarke's original theory (Cohn, et al., 1997; Randell, et al., 1992). The basic part of the formal theory assumes a dyadic relation: C(x,y) to mean that region x is connected to region y.

The mereological relation of *parthood*, P(x,y) is defined from the connection relation C(x,y), which together is used to define a number of relations as enumerated in Table 2. DC(x,y) through NTPP(x,y) with the inverses for the last two, that is, $TPP^{-1}(x,y)$ and $NTPP^{-1}(x,y)$ constitute a *Jointly Exhaustive* and *Pair wise Disjoint* (JEPD) set of base relations referred to as RCC-8.

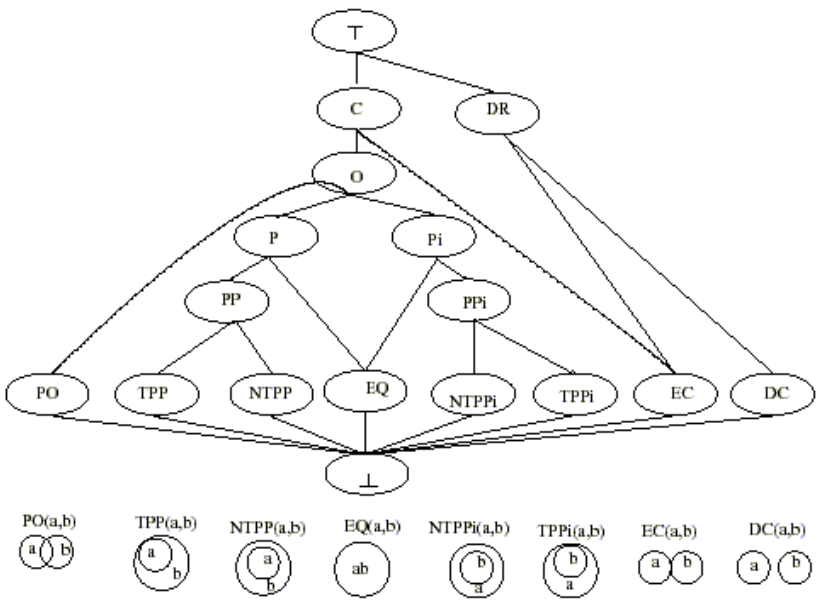
Table 1. Defined relations in Clarke's theory

Relation	Interpretation
DC(x,y)	x is disconnected from y
P(x.y)	x is part of y
PP(x,y)	x is proper-part of y
O(x,y)	x overlaps y
DR(x.y)	x is discrete from y
EC(x,y)	x is externally connected to y
TP(x.y)	x is a tangential part of y
NTP(x,y)	x is a non-tangential part of y

Table 2. Defined relations in Region Connection Calculus

Relation	Interpretation
PP(x,y)	x is proper-part of y
O(x,y)	x overlaps y
DR(x,y)	x is discrete from y
DC(x,y)	x is disconnected from y
EC(x,y)	x is externally connected to y
PO(x,y)	x partially overlaps y
EQ(x,y)	x is equal to y
TPP(x,y)	x is a tangential proper part of y
NTPP(x,y)	x is a non-tangential proper part of y

Figure 1. Lattice defining the subsumption hierarchy of dyadic relations defined in terms of C. The pictorial representation of the eight base relations (referred to as RCC-8) is included below the lattice.



Relations defined in terms of C(x,y) can be embedded in a relational lattice with the top element interpreted as tautology and the bottom element as contradiction. The relational lattice along with the representations of the eight base relations is shown in Figure 1.

Asher & Vieu's Theory

Asher & Vieu (1995) gave a mereotopological system based on Clarke's *Calculus of Individuals*. A significant feature of the theory is the notion of *weak contact* and *strong contact*. They qualify the standard interpretation of connection and make distinction between connection such as "relation between a glass and the table on which it is standing" with that from "relation between the stem of the glass and the cup of the glass" (Asher & Vieu, 1995). The former is an example of weak contact, whilst the latter is of strong contact.

Contrary to the RCC interpretation, Asher & Vieu (1995) argue that differentiating between an individual, its closure, and its interior is cognitively important. Asher

& Vieu 's mereotopological theory incorporated notions of open and closed sets (from point-set topology) to make such distinctions.

Muller's Extension

Muller (2002) has taken over the theory of Asher & Vieu (1995) and extends it to *space-time*. Taking up the idea of Hayes' (1985) spatio-temporal *histories*, Muller presents a mereotopological model in which the primitive entities are spatio-temporal regions, on which spatio-temporal and temporal relations are defined.

The spatio-temporal relations are an extension of spatial relations to space-time. Additional temporal relations are introduced to add further structural specification. Besides a classical *temporal precedence* relation, a primitive *temporal connection* (a connection with almost the same behaviour as C(x,y) but only on a temporal level) is introduced. With these it is possible to distinguish a temporal overlap from a simple temporal contact. Perhaps the most important contribution of Muller's mereotopological theory of space-time was an explicit definition of *qualitative continuity*.

Other Region-Based Theories

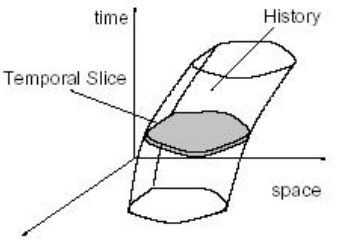
An alternative approach to representing and reasoning about topological relations is Egenhofer's n-intersection Model (Egenhofer & Franzosa, 1991). Three sets of points are associated with every region: its interior, boundary, and complement. A 3x3 matrix called the 9-intersection can characterize the relationship between any two regions[1]. Taking into account the physical reality of 2D space and some specific assumptions about the nature of regions, there are exactly 8 matrices, corresponding to the RCC-8 relations.

Qualitative Spatio-Temporal Reasoning

Spatial *configurations* tend to change. Reasoning about space often involves reasoning about change in spatial configurations. Driven by *cognitive* approaches that characterize the processing of spatial information in QSR, there has been work in other areas within AI such as computer vision, robotics, and so forth. Qualitative representation and reasoning about spatial change (Galton, 2000) and spatial interactions (Hazarika & Cohn, 2002) have been explored. Qualitative Spatio-Temporal Reasoning (QSTR) encompasses all such techniques.

There are two basic approaches to reasoning with qualitative spatial data over time: take a *snapshot* viewpoint and describe dynamic behaviour as a set of temporal states, or view the world as spatio-temporal histories (Hayes, 1985).

Figure 2. A space-time history is a n+1 dimensional volume for n-D space.



Qualitative Space-time Histories

In order to add time to space, an obvious and straightforward choice is to interpret entities in space-time rather than in space alone. In fact, Clarke's intended interpretation of his region-based calculus was spatio-temporal (Clarke, 1981). More recently (Hayes, 1985; Hazarika & Cohn, 2002; Muller, 2002) consider entities to be *histories* as shown in Figure 2.

Qualitative Motion

In spite of a large amount of work in mereotopological theories as a basis for common sense reasoning, very little work has been done on motion in a qualitative framework. Galton (2000) and more recently Muller (2002) and Hazarika & Cohn (2002) have looked at motion in the more cognitive kind of approach characterized by processing spatial information. Muller's (2002) Theory is an enrichment of Asher & Vieu's (1995) theory to achieve a formal theory for reasoning about motion. The expressive power of the theory allows for definition of complex motion classes such as those expressed by motion verbs in natural language.

Other Approaches

Qualitative representation and reasoning over *episodes* in space (El-Geresy, Abdelmolty, & Jones, 2000) is the closest to the spatio-temporal *histories* of a mereotopological theory of space-time. The episode of an object is the consistent behaviour of a spatial object within duration of time when this behavior can be described as being consistent (i.e., described by a single function). The approach is limiting, as only *well behaved* approximations of representation of spatio-temporal relations are possible.

Many qualitative spatial and temporal calculi arise from a set of JEPD relations. Ligozat & Renz (2004) examine the construction of such formalisms in order to make apparent the formal algebraic properties of all formalisms of that type. Malek (2004) propose a logic-based framework for representing and reasoning about qualitative spatial relations over moving agents in space and time. The framework would find applications in intelligent transportation system and mobile autonomous navigation systems.

FUTURE TRENDS

Use of QSR for other areas such as the *semantic web* (Katz & Grau, 2005) and *bio-medical ontologies* (Donnelly, Bittner, & Rosse, 2005) is being explored in a big way. The Web Ontology Language (OWL) has not been designed for representing spatial information. A translation of the RCC-8 calculus into OWL makes it possible to adapt OWL based tools for representing and reasoning on qualitative spatial information. In particular, ontology editors could be equipped with suitable user interfaces for spatial modeling and spatial *knowledge bases* could be published and shared on the Semantic Web.

CONCLUSION

Even though much work has been done in generating spatial representational calculi, there remain a number of theoretical questions. Their inventors have not given a formal semantics for many calculi. The calculus given by Clarke (1981) and all related calculi such as the first order theory of RCC (Randell, et al., 1992) and of Asher & Vieu (1995) are undecidable. The constraint language of RCC-8 has been shown to be decidable. This was achieved by encoding each RCC-8 relation as a set of formulae in *intuitionistic proportional calculus*, which is a decidable calculus.

Renz & Ligozat (2005) show under which conditions algebraic closure can be used to decide consistency in a qualitative calculus, how weak consistency affects different important techniques for analyzing qualitative calculi, and under which conditions these techniques can be applied. All their results are general and can be applied to all existing and future qualitative spatial and temporal calculi. Renz & Ligozat (2005) also gave a road map of how qualitative calculi should be analyzed.

An issue that has not been much addressed yet in the QSR literature is the issue of cognitive validity. Claims are often made that qualitative reasoning is akin to human reasoning, but with little or no empirical justification. More work needs to be done in this direction.

REFERENCES

Asher, N. & Vieu, L. (1995). Towards a geometry of common sense: A semantics and a complete axiomatization of mereotopology. *Proceedings of 14th International Joint Conference on AI*, 846–852.

Clarke, B. L. (1981). A calculus of individuals based on "connection." *Notre Dame Journal of Formal Logic*, 22(3), 204–218.

Cohn, A. G., Bennett, B., Gooday, J., & Gotts, N. M. (1997). RCC: A Calculus for Region based Qualitative Spatial Reasoning. *GeoInformatica*, 1(3), 275–316.

Dehak, S. M. R., Bloch, I., & Maitre, H. (2005). Spatial Reasoning with Incomplete Information on Relative Positioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9), 1473-1484.

De Laguna, T. (1922). Point, line and surface as sets of solids. *The Journal of Philosophy*, 19, 449–461.

Donnelly, M., Bittner, T., & Rosse, C. (2005). A Formal Theory for Spatial Representation and Reasoning in Biomedical Ontologies. *Artificial Intelligence in Medicine.*

Egenhofer, M. J. & Franzosa, R. (1991). Point-set topological spatial relations. *International Journal of Geographical Information Systems*, 5(2), 161- 174.

El-Geresy, B. A., Abdelmoty, A. I., & Jones, C. B. (2000). Episodes in space: Qualitative representation and reasoning over spatio-temporal objects. *International. Journal on Artificial Intelligence Tools*, 9(1), 131–152.

Frank, A. (1992). Qualitative spatial reasoning about distance and directions in geographic space. *Journal of Visual Languages and Computing*, 3, 343–373.

Freksa, C. (1992). Temporal reasoning based on semi-intervals. *Artificial Intelligence*, 54, 199–227.

Galton. A. P. (2000). *Qualitative Spatial Change*. Oxford University Press.

Hayes, P. J. (1985). *The Second Naive Physics Manifesto. Formal Theories of the Commonsense World*, Ed., J. R. Hubbs and R. C. Moore, Ablex Publishing Corporation, NJ.

Hazarika, S. M. & Cohn, A. G. (2002). Abducing qualitative spatio-temporal histories from partial observations. Principles of Knowledge Representation and Reasoning: *Proceedings of KR 2002*, Ed., D. Fensel, F. Guinchiglia, D. McGuinness, and Mary-Anne Williams, 14–25. Morgan Kaufmann.

Katz, Y., & Grau, B.C. (2005). Representing Qualitative Spatial Information in OWL-DL. *In proceedings of OWL: Experiences and Directions*, Galway, Ireland.

Leyton, M. (1988). A Process Grammar for Shape. *Artificial Intelligence*, 34, 213–247.

Ligozat, G., & Renz, J. (2004) What Is a Qualitative Calculus? A General Framework. *Proceedings of PRICAI* 2004, 53-64.

Malek, M. R. (2004). *A Logic-Based Framework for Qualitative Spatial Reasoning in Mobile GIS Environment. S. Tsumoto et. al.* (Eds.) Rough Sets and Current Trends in Computing, 4th International Conference, Uppsala, Sweden. LNCS-3066, 418-426.

Meathrel, R., & Galton, A. P. (2000). Qualitative representation of planar outlines. *Proceedings of 14th European Conference on AI*, 224–228.

Muller, P. (2002). Topological spatio-temporal reasoning and representation. *Computational Intelligence*, 18(3), 420–450.

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. *Proceedings of 3rd International Conference on Knowledge Representation and Reasoning*, Ed., B. Nebel, C. Rich, and W. Swartout, 165–176.

Renz, J., & Ligozat, G. (2005). Weak Composition for Qualitative Spatial and Temporal Reasoning. *Proceedings of Principles and Practice of Constraint Programming - 11th International Conference*. Ed. P van Beek, LNCS 3709, 534-548.

Tarski, A. (1959). *What is elementary geometry? The Axiomatic Method* (with special reference to geometry and physics), Ed., L. E. J. Brouwer, E. W. Beth and A. Heyting, 16–29, Amsterdam.

Weld, D. S., & De Kleer, J. (1990). *Readings in Qualitative Reasoning About Physical Systems*. Morgan Kaufman, San Mateo.

KEY TERMS

Mereology: Mereology is the theory of parts and whole.

Mereotopology: Topological notions have to be added to mereology to provide an adequate qualitative theory of space. Such combination of the disciplines of mereology and topology is referred to as mereotopology.

Qualitative Motion: Description of motion in a more cognitive kind of approach characterized by processing spatial information.

Qualitative Reasoning: Approach for dealing with common-sense knowledge without recourse to complete quantitative knowledge. Representation of knowledge is through a limited repository of qualitative abstractions.

Qualitative Spatial Reasoning: Qualitative Spatial Reasoning is concerned with providing calculus which allow a machine to represent and reason with spatial entities without resort to traditional quantitative techniques. Representation is concerned with different forms of spatial knowledge and reasoning is concerned with methods and techniques for decision-making. The term Qualitative Spatial Reasoning subsumes both the sub-fields of representation and reasoning.

Qualitative Spatio-Temporal Reasoning: Qualitative Spatio-Temporal Reasoning encompasses all techniques of qualitative representation and reasoning about spatial change and spatial interactions.

Region Connection Calculus: RCC is a mereotopological theory of space. The topological primitive of connection is primal from which the mereological primitive of parthood is defined. The theory has a set of eight jointly exhaustive and pair-wise disjoint base relations referred to as RCC-8.

Spatio-Temporal History: Space-time regions traced by objects over time are termed space-time history or spatio-temporal history.

ENDNOTE

[1] A simpler 2x2 matrix known as the 4-intersection featuring just the interior and the boundary is sufficient to describe the eight RCC relations. The 3x3 matrix allows more expressive sets of relations to be defined since it takes into account the relationship between the regions and its embedding space.

Quality Assurance Issues for Online Universities

Floriana Grasso
Liverpool University, UK

Paul Leng
Liverpool University, UK

INTRODUCTION

Online delivery of degree-level programmes is an attractive option, especially for working professionals and others who are unable to contemplate full-time residential university attendance. If such programmes are to be accepted, however, it is essential that they attain the same standards and quality as conventionally delivered degrees. The key challenge is to find ways to ensure that the qualities that make university education attractive are preserved in the context of a new and quite different model of delivery.

Many systems have been developed to support online learning (see, e.g., Anderson & Kanuka, 1997; Davies, 1998; Persico & Manca, 2000; Suthers & Jones, 1997; Yaskin & Everhart, 2002). These systems may or may not mimic conventional lecture-room teaching, but will necessarily involve major differences in the ways in which teaching and student support are organised. Furthermore, the Internet lends itself naturally to an internationalisation of education delivery, but this also poses challenges for universities that have developed their structures within the framework of national education systems. To address these issues, it may be desirable for the university to work in partnership with other agencies, for example to provide local support services for students. This too, however, may introduce new problems of quality control and management. We will discuss here what structures are required to ensure the quality of the education provided and the standards of the degrees offered in this context.

BACKGROUND

The emergence of the Internet as a way of delivering higher education has led to examinations of its implications for education policy in many national and international contexts. A set of benchmarks for quality of online distance education was developed by the (U.S.-based) Institute for Higher Education Policy (2000). This identified a total of 24 benchmarks, in seven categories. In the UK, the Quality Assurance Agency for Higher Education has issued guidelines on the Quality Assurance of Distance Learning (QAA, 2000a), with a similar scope to those of the IHEP. A comparison of the main headings of the two frameworks is illustrated in Table 1. Also relevant, when the delivery model involves partnership with external agencies, is the QAA Code of Practice in relation to Collaborative Provision (QAA, 2000b). Similar issues are examined in an Australian context by Oliver (2001), and from Hong Kong by Yeung (2002). Yorke (1999) discusses quality assurance issues in relation to globalised education, touching especially on collaborative provision. Other perspectives are offered by Pond (2002), Little and Banega (1999), and Davies et al. (2001).

Much of the work in this field reflects "an implicit anxiety that the 'values' of traditional teaching may somehow be eroded" (Curran, 2001). There is consequently, in most prescriptions, a strong emphasis on replicating in an online context the characteristics of quality that we might expect to (but do not always) find in conventional teaching. Thus, one of the precepts of (QAA, 2000a) calls for "…..managing

Table 1. Comparison of U.S. and UK QA frameworks

IHEP (USA)	QAA (UK)
Institutional Support	System Design (i.e., institutional issues)
Course Development	Programme Design (course development and structure)
Course Structure	
Teaching and Learning	Programme Delivery
Student Support	Student Support
Faculty Support	
Evaluation and Assessment	Student Assessment
	Student Communication and Representation

the delivery of each distance learning programme of study in a manner that safeguards the academic standards of the award"; and one of the benchmarks of the IHEP specifies that "Feedback in student assignments is provided in a timely manner". Unexceptionable as they are, these requirements are not peculiar to online distance learning. The key issue is not, therefore, one of defining quality assurance criteria, but rather that of providing structures to ensure their implementation.

QUALITY ASSURANCE FOR ONLINE DEGREES

Pedagogic Issues

Before examining quality assurance as such, we will first consider questions relating directly to the pedagogic approach used in online learning. In this respect, the premise that quality in online learning involves only a replication of on-campus characteristics is, we believe, limiting. We start, instead, from the standpoint that lecture-based teaching, whatever its merits, is not necessarily an ideal which online teaching must emulate. Students all too frequently attend lectures in an entirely passive mode, expecting to listen and receive the information they require while making no positive contribution themselves. Interaction between lecturer and students, and within groups of students is low, especially in the large classes that are typical of modern universities.

Conversely, online teaching makes it possible to recreate, through the medium of moderated online discussion, an atmosphere that is closer to that of a small-group on-campus seminar, and, paradoxically, can be far more involving and interactive than is typically the case in on-campus teaching. Two broad principles inform the approach: *constructivism* (Wilson, 1996), and *collaborative enquiry.* Collaborative enquiry via Internet-mediated communication provides a framework for this mode of learning (Stacey, 1998). The aim is to use the medium to foster the creation of a *learning community* (Hiltz & Wellman, 1997) that will enable dialogue between participants, sharing of information, and collaborative project work.

Moderated discussion (Collins & Berge, 1997) is a key feature of the teaching paradigm here, and serves a number of purposes that are relevant to the question of quality. Most obviously, it provides the means by which students may share information and experience, comment on the course materials and assignments, raise questions, and bring to the class knowledge and expertise that is outside the experience of the course teacher. To a significant extent, the students thus participate actively in the teaching process, augmenting the overall learning experience. Less obviously, there are other issues of quality in which classroom discussion can have a role; we will discuss these next.

Quality Assurance Issues

Key issues of quality assurance in an online degree programme include:

- Academic control
- Academic standards
- Staff appointment and training
- Monitoring of programme delivery
- Assessment procedures
- Student identity and plagiarism
- Student progression and support

Our review of these issues, next, draws on our experience with the online degree programmes at the University of Liverpool (Gruengard, Kalman & Leng, 2000).

Academic Control

The primary requirement of the frameworks defined by the QAA and other bodies is that all academic aspects of an online degree programme should remain the responsibility of the parent university, which should have structures and procedures that are effective in discharging this responsibility. The issue here is that the academic standards and quality of the programme may be threatened by differences between the parties involved in its delivery, especially when there is only an indirect relationship between the university and some of the people involved (for example, regional partner organisations, or locally-based tutors).

In principle, these problems can be resolved by placing online degree programmes firmly within the framework defined by the university for approving and managing its courses. To oversee this, we have at Liverpool established a dedicated organisational unit within the university, the e-Learning Unit.

Academic Standards

A corollary this is that, wherever possible, the quality management of an online programme should follow procedures that are comparable to those established for other degrees of the university, especially in respect to those procedures that define and maintain the academic standards of the degree. These will include course and module approval and review procedures, assessment criteria, and so forth. In most cases, it should be possible to exactly replicate the procedures that apply on campus.

Q

Staff Appointment and Training

In the Liverpool University online programmes, teaching is principally carried out by part-time *instructors* who are based throughout the world. All appointments, however are subject to University approval using the same procedures and criteria that apply for the approval of non-established staff to teach on internal programmes. All instructors are required to first undertake an (online) training programme, over a period of 6 weeks, in which they are instructed on the use of the software platform and the methodology and pedagogic approach used in the programme. Further to this, in the first module they teach, the instructor is overseen by an academic *mentor* whose role is to advise and guide the novice.

Monitoring of Programme Delivery

A key aspect of quality assurance, especially when instructors as well as students are not campus-based, is the monitoring of module delivery. Interestingly, it is easier to do this effectively than is the case for on-campus teaching. In the Liverpool University model, *all* significant communications between staff and students are made electronically, as postings in or through a "virtual classroom". Thus, all are subject to scrutiny by "lurking" in the classroom while the module is being delivered, and/or by examining the recorded history of the class subsequently. In this way the academic staff of the e-Learning Unit monitor the delivery of each module to ensure that the appropriate standards and quality criteria are met. The module monitor is required to complete a report that also incorporates a review of feedback from students and the comments of the instructor. These reports are reviewed routinely by a Board of Studies.

Assessment Procedures

Assessment in this model is readily subject to moderation by the module monitor, who has access to all the relevant information contained in the record of the virtual classroom. Review of assessment outcomes is an explicit part of the end-of-module procedures, and these are finally subject to confirmation by the Board of Examiners, which in the UK framework includes an external examiner from another university. The external examiner also has full access to the virtual classroom and so, in practice, has the opportunity for a more detailed examination of assessment processes and standards than is usually possible in on-ground teaching.

Strict management of the assessment process is in our experience essential when instructors are drawn from a wide variety of cultures. In this case consistency of assessment can only be achieved by defining very clear grading descriptors and criteria, and by insisting that the interpretation of these is moderated firmly within the university's assessment model.

Student Identity and Plagiarism

One of the questions that most exercises organisations considering online learning is that of how to confirm the identity of participants, together with the related question of protecting against plagiarism. In this respect a key role is played by discussion in the virtual classroom. Participation in discussion provides a means of monitoring the effective involvement of each student, and assists in preventing impersonation and plagiarism. Research has shown (Klemm & Snell, 1996; Lai, 1997) that involvement in online discussion is rarely wholly effective unless moderated by external facilitators, and we believe that it is important that it be made a requirement, equivalent to the attendance requirements of on-campus degrees.

The fact that all communications take place online, and are recorded and preserved indefinitely, provides further protection against plagiarism. It is easy to apply programs that perform comparisons of work submitted in the virtual classroom, or use services that perform checks against plagiarism throughout the Web.

Student Progression and Support

The requirement to participate in online discussion also provides a means of monitoring student progress. If a student is failing to keep up with the requirements of the programme, this rapidly becomes apparent as his/her contributions to the discussion falter. At this point the instructor can intervene to investigate and take action if required.

FUTURE TRENDS

A key precept for quality assurance of online degree programmes is that, wherever possible, procedures and structures that are thought to be necessary on campus should have an online equivalent. As the previous discussion reveals, creating equivalent processes is usually possible, but some significant differences emerge, highlighted in Table 2. The comparison of conventional lecture-based teaching with our model of online learning is not, we believe, to the disadvantage of the latter.

Although online degree programmes will not replace campus-based education, we believe there will be a strong future trend towards programmes that will meet particular areas of demand, especially the needs of working adults. Examples of successful online higher education programmes are, so far, relatively few, especially in Europe. We believe firmly that the successful programmes of the future will

Table 2. Comparison of on-campus and online characteristics

	On Campus	Online
Teaching/learning mode	Predominantly lecture-based	Predominantly seminar-based
Interpersonal interactions	Low: via classroom discussions/questions	High: via moderated e-mail dialogue
Verification of student identity	Personal appearance	Textual/linguistic characteristics
Student support	Face-to-face meetings with tutors	E-mail interactions with tutors
Review of standards of attainment	Scrutiny of examination scripts	Inspection of work in online classrooms
Mentoring/monitoring of staff performance	Inspection of selected lectures/classes	"Lurking" in ongoing online classes

be those that focus on pedagogy, and give precedence to academic standards and quality assurance, rather than those that emphasise technological aspects or focus on low-cost delivery.

CONCLUSION

Quality in online degree programmes is often perceived to imply a replication of on-campus characteristics. We believe, conversely, that in some respects online delivery provides an opportunity to enhance the quality of learning opportunities for students. Especially this is so when the learning paradigm encourages a high degree of discussion and interaction between staff and students in the virtual classroom.

In other respects, it is indeed necessary to maintain comparability with on-campus degrees, especially in relation to academic standards. Here, however, the key issues do not relate to the definition of standards and quality criteria, but to the creation of mechanisms to uphold them. We believe it is necessary to establish well-defined and rigorous monitoring procedures for this purpose. Again, a learning environment that emphasises classroom discussion is a help in many aspects of quality management.

REFERENCES

Anderson, T., & Kanuka, H. (1997). On-line forums [1]: New platforms for professional development and group collaboration. *Journal of Computer-Mediated Communication, 3*(3).

Collins, M.P., & Berge, Z.L. (1997). *Moderating online electronic discussion groups.* Paper presented at the American Educational Research Association, Chicago. Retrieved on September 20, 2004 from *http://www.emoderators.com/moderators/sur_aera97.html*

Curran, C. (2001). The phenomenon of on-line learning. *European Journal of Education, 36*(2), 113-132.

Davies, G. (Ed.). (1998). Teleteaching '98: Distance learning, training and education. *Proc XV IFIP World Computer Congress,* Vienna/Budapest.

Davies, G., Doube, W., Lawrence-Fowler, W., & Shaffer, D. (2001). Quality in distance education. *Proc 32nd SIGCSE Technical Symposium on Computer Science Education*, ACM (pp. 392-393).

Gruengard, E., Kalman, Y.M., & Leng, P (2000). University degrees via the Internet: A new paradigm for public-private partnership. *Innovations Though Electronic Commerce* (Proc IEC2000), Manchester (pp. 46-53).

Hiltz, S.R., & Wellman, B. (1997). Asynchronous learning networks as a virtual classroom. *Comm ACM, 40*(9), 44-49.

Institute for Higher Education Policy. (2000). *Quality on the line: Benchmarks for success in Internet-based distance education.*

Klemm, W.R., & Snell, J.R. (1996). Enriching computer-mediated group learning by coupling constructivism with collaborative learning. *Journal of Instructional Science and Technology, 1*(2).

Lai, K-W. (1997). Computer-mediated communication for teenage students: A content analysis of a student messaging system. *Education and Information Technologies, 2*, 31-45.

Little, D.L., & Banega, B.H. (1999). Development of standards or criteria for effective online courses. *Educational Technology and Society, 2*(3), 4-15.

Oliver, R. (2001). Assuring the quality of online learning in Australian higher education. *Proceedings of Moving Online Conference II.*

Persico, D., & Manca, S. (2000). Use of FirstClass as a collaborative learning environment. *Innovations in Education and Training International, 37*(1), 34-41.

Pond, W.K. (2002). Distributed education in the 21st century: Implications for quality assurance. *Online Journal of Distance Learning Administration, 5*(2).

QAA. (2000a). (Quality Assurance Agency for Higher Education): *Distance learning guidelines.* Retrieved on September 20, 2004 from *http://www.qaa.ac.uk*

QAA. (2000b). (Quality Assurance Agency for Higher Education): *Code of practice for the assurance of academic quality and standards in higher education: Collaborative provision.* Retrieved on September 20, 2004 from *http://www.qaa.ac.uk*

Stacey, E. (1998). Learning collaboratively in a CMC environment. In G. Davies (Ed.), *Teleteaching '98, Proceedings of XV IFIP World Computer Congress,* Vienna/Budapest, Austrian Computer Society (pp. 951-960).

Suthers, D., & Jones, D. (1997). An architecture for intelligent collaborative educational systems. *Proceedings of AI&ED'97,* Kobe, Japan.

Wilson, B.G. (1996). *Constructivist learning environments: Case studies in instructional design.* Educational Technology Publications.

Yaskin, D., & Everhart, D. (2002). Blackboard Learning System (Release 6): Product overview. Retrieved on September 20, 2004 from *http://www.blackboard.com/docs/wp/LSR6WP.pdf*

Yeung, D. (2002). Towards an effective quality assurance model of Web-based learning: The perspective of academic staff. *Online Journal of Distance Learning Administration,* 5(2).

Yorke, M. (1999). Assuring quality and standards in globalised higher education. *Quality Assurance in Education, 7*(1), 14-24.

KEY TERMS

Constructivism: A form of learning in which students construct their own unique understanding of a subject through a process that includes social interaction, so that the learner can explain understandings, receive feedback from teachers and other students, clarify meanings, and reach a group consensus.

Globalised Education: Educational programmes in which both students and educators may be globally distributed.

Higher Education Programme: The processes, learning materials, and associated procedures and facilities that lead to the completion of a degree or related qualification.

Moderated Discussion: Discussion that is supervised, partly directed, and evaluated by a programme tutor.

Moderation of Assessment: External oversight of an assessment process to ensure that appropriate standards are maintained.

Online Learning: A programme of education that is carried out wholly or primarily through the medium of the Internet.

Quality of Education: Refers to the effectiveness of the programme in promoting student learning and achievement, in relation to its expected outcomes and standards.

Standards (of Education): Refer to the level of attainment achieved by students, in relation to the qualification awarded.

Virtual Classroom: An internet-mediated forum for distribution of learning materials, classroom discussion, and collaborative working.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2382-2386, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Networking & Telecommunication

Quality-of-Service Routing

Sudip Misra
Cornell University, USA

INTRODUCTION

The area of quality-of-service (QoS) routing is concerned with selecting routing paths while meeting strict end-to-end service requirements involving resource constraints, while achieving optimum throughput in the network. The usefulness of QoS routing is not new. QoS routing is quite popular in the telecommunications industry because of the increased demand for satisfying multiple customer demands and obtaining increased utilization of network resources, while satisfying the varied user requirements.

BACKGROUND

Two basic considerations in QoS routing in integrated services packet-switched networks concern: (1) routing traffic with bandwidth guarantees, and (2) routing traffic with delay guarantees. Some of the algorithms proposed traditionally for solving the former class of problems are the widest-shortest path (WSP) (Guerin, Orda, & Williams, 1997), and the shortest-widest path (SWP) (Wang & Crowcroft, 1996) algorithms. More recently, Vasilakos, Saltouros, Atlassis, and Pedrycz, (2003) proposed the stochastic estimator learning automata (SELA) routing algorithm for QoS routing in asynchronous transfer mode (ATM) networks.

Online routing using traffic engineering (TE) principles has recently drawn considerable attention. We mention here a few TE algorithms proposed in the literature (e.g., Iliadis & Bauer, 2002; Kar, Kodialam, & Lakshman, 2000; Suri, Waldvogel, Bauer, & Warkhede, 2003; Szeto, Boutaba, & Iraqi, 2002; Wang, Su, & Chen, 2002). Of all the online TE algorithms, we believe that the one that has attracted the most attention is the minimum interference routing algorithm (MIRA) designed by Kar et al. (2000).

In addition to the MIRA algorithm, there are a few other TE routing algorithms that were proposed by other researchers, some of which are: the profile-based routing (PBR) (Suri et al., 2003); the dynamic online routing algorithm (DORA) (Szeto et al., 2002); Iliadis and Bauer's (2002) algorithm; Wang et al.'s (2002) algorithm; and the random races-based traffic engineering routing algorithm (RRATE) (Oommen, Misra, & Granmo, 2006). Some of these are also described in the sections to follow.

QUALITY-OF-SERVICE ROUTING

A network is said to support QoS (Guerin et al., 1997; Peterson & Davie, 2000), if it has the capability of treating different packets differently. QoS technology has enabled service providers to support different levels of service to different customers, thereby capacitating them with the option to provide better levels of paid services to some customers more than to others. For example, some groups of customers may be concerned with a service that guarantees packet delivery, even if that means paying a higher price for these services, others may just as well be satisfied with relatively less reliable data transfer by paying less for their subscribed services. Networks that transport multimedia traffic, that is, voice, data, and video need differential treatments of different packets—while voice traffic is highly sensitive to time delay and the orderly delivery of packets, data traffic is relatively less sensitive to these, and videoconferencing traffic requires a dedicated connection for a fixed amount of time for the real-time, orderly delivery of packets.

Typical QoS routing-based performance metrics are bandwidth, delay, and throughput. While some applications require bandwidth guarantees, some others mandate the satisfiability of strict end-to-end delays, and others still require a high throughput, or a combination of both of these criteria (Ma, 1998).

Routing protocols in the pre-QoS era did not consider QoS requirements of connections (e.g., delay, bandwidth, and throughput). Furthermore, optimization of resource utilization was not a primary goal. As a result, while there were flow requests that were rejected because of nonavailability of sufficient underlying resources, there were some other resources that remained available. To address such deficiencies with conventional routing protocols, QoS routing algorithms were devised that could locate network paths which satisfy QoS requirements and which made better use of the network. Routing of QoS traffic requires stringent performance guarantees of the QoS metrics (e.g., delay, bandwidth, and throughput) over the paths selected by the routing algorithms. Accordingly, whereas the traditional shortest path algorithms, for example, Dijkstra (1959)'s or Bellman (1958)'s, indeed, have the potential of selecting a feasible path for routing, QoS routing algorithms must consider multiple QoS and resource utilization constraints, typically making the problem intractable. QoS routing is, thus,

different from that of routing in traditional circuit-switched and packet-switched networks.

In ATM networks, for example, a connection is accepted or not by the Connection Admission Control (CAC), depending on whether or not sufficient resources (e.g., bandwidth) are available. The CAC operates by taking into account factors such as the incoming QoS requests and the available resources in the network. The CAC operates the QoS routing algorithms, which identify whether the different possible candidate paths satisfy the QoS requirements or not. The network architects and the engineers want to choose such a routing algorithm that will help the service providers for maximizing the network resources (e.g., the amount of bandwidth to be routed), while satisfying the requirements from the customers. Therefore, the design of any good QoS routing algorithm takes into account factors such as satisfying the QoS requirements, optimizing the consumption of the network resources (e.g., buffer space, link bandwidth), and balancing the traffic load across different paths. In addition to these, a good QoS routing algorithm should characterize itself by its ability to adapt to the periodic dynamic behavior of the network.

Conventional QoS Routing Algorithms

In this section we present an overview of some of the traditional algorithms for QoS Routing.

- **SWP:** The SWP algorithm was proposed by Wang and Crowcroft (1996). They proposed two variants of the algorithm—the *distance-vector-based SWP*, and the *link-state based SWP*—depending on whether the SWP algorithm is governed by *distance-vector-based routing* or by *link-state-based routing*. Essentially, in both variants of the SWP algorithm, the algorithm finds a route with the widest path, that is, the path with the maximum bottleneck bandwidth. When there is more than one choice available, the algorithm chooses the path that has the minimum length, that is, the one that has the shortest propagation delay (Wang & Crowcroft, 1996). If still there is a tie between one or more such path(s), one of the prospective paths is randomly chosen.
- **WSP:** The WSP algorithm was proposed by Guerin et al. (1997). Unlike the SWP algorithm, WSP first attempts to compute the shortest path; if there is more than one alternative, the algorithm chooses the one with the largest residual bandwidth in the bottleneck link (i.e., the widest path). If there is still a tie with one or more such path(s), one of the prospective paths is randomly chosen.
- **SELA:** The SELA routing algorithm was proposed by Vasilakos et al. (2003), and is based on the concept of SELA (Vasilakos & Papadimitriou, 1992). SELA is a QoS-based dynamic source routing algorithm where each source node maintains a database of the topology information between it and its k-shortest path neighbors (to each reachable destination node) (Cormen, Leiserson, & Rivest, 1990). In SELA, a learning automaton is stationed at each node in the network for determining how each call is to be routed between every node to every other reachable node in the network. For establishing a call, the source node selects one of the precomputed shortest path routes that can potentially be accepted by the algorithm based on the QoS requirements and traffic parameters. If no such path can be found, SELA rejects the request. In the heart of the SELA routing algorithm is the design of a function that is used by SELA to estimate the environmental feedback of the path selected by the algorithm. Further details of the algorithm can be found in Vasilakos et al. (2003).

Traffic Engineering Routing

TE mandates to optimize the performance of traffic handling and resource utilization on existing physical network topologies. This is, in principle, engineered by minimizing the over utilization of network capacity and distributing the traffic load on costly network resources such as, links, routers, switches, and gateways (Osborne & Simha, 2002). In the context of routing, TE is of great usefulness because traditional routing techniques are based on greedy shortest path computation techniques that lead to the over utilization of certain network resources, even when other resources remain under utilized.

Multi-protocol Label Switching (MPLS) has recently emerged for many professionals as a de facto standard in TE. MPLS is an Internet Engineering Task Force (IETF) standard which merges the layer 2 information of bandwidth, latency, and utilization of network links, with the control protocols used in layer 3 Internet protocol (IP), in order to simplify the exchange of IP packets. At the heart of the idea is the usage of a *label* (or a *tag*) to calculate shortest paths to all destinations within an autonomous system, thereby expediting the forwarding of packets. A label can be perceived as a simplified representation of an IP packet's header, with the additional advantage of enabling core backbone networks to operate at high speeds because of the exclusion of the need to re-examine each packet's IP header in detail. This, in turn, permits the differentiating between packets on an individual basis and facilitates the support of QoS. The destination of a packet is determined by observing the label and not the IP address it is destined to. MPLS helps network operators manage network route failures and makes the system able to decongest bottleneck links by providing a detour for the incoming traffic. It can also help service providers manage the

provisioning of different kinds of traffic according to different service plans, service, and policies of the customers.

We review the basic MPLS terminologies by using a hypothetical network as shown in Figure 1 and describe hereafter, in brief, how a packet is transmitted. In actuality, the steps involved are much more complex. A Label Switched Router (LSR) is a backbone router in the physical network topology that runs the existing layer 3 IP protocol. The particular cases of LSR, which are the edge routers, are called the label edge routers (LERs). These routers often serve as the *ingress* and *egress routers*, which can be thought of as synonymous to the source and destination routers for packets passing through the MPLS network. Packets are tagged with fixed-length labels at the ingress router and untagged off the labels at the egress routers. The paths (or routes) along which the labeled packets are transmitted are termed as the labeled switched paths (LSPs), and the protocol that monitors the negotiation and the exchange of labels is called the label distribution protocol (LDP). In a typical transmission of a packet through the MPLS network, after the backbone routing protocols, (such as the open shortest path first protocol (OSPF)), ascertains the reachability of a destination node, the LDP protocol establishes a mapping between a label and the destination node; the ingress node then receives a packet and labels it with a fixed length label (as described earlier) and transmits it towards the egress node, which upon receipt of the labeled packet, removes its label and delivers it (Osborne & Simha, 2002).

During the LSP setup phase mentioned previously, the intermediate LSRs between the ingress-egress nodes are specified. During this phase, the paths for a given flow are explicitly specified, thereby bestowing the service providers with a tool for engineering the incoming traffic to be routed and also supporting QoS, optimizing network utilization, and minimizing the number of rejected LSP setup requests, as this is what is typically mandated by the traffic engineers.

The traditional routing algorithms that forward packets based on the information about the destination address are only shortsighted and are, therefore, constrained by the following limitations (as identified by Suri et al., 2003):

- They do not take into consideration the current flows or expected future flow demands in the network. Those algorithms are greedy, in the sense that they would route a request through the default shortest path, the shortest-widest path, or the widest-shortest path, and would reject a demand when the precomputed routes using those algorithms are congested, even when other alternate routes are free to accept more requests.
- They do not take into consideration information about network infrastructure, network topologies, or traffic profiles to avoid loading bottleneck links in a network that might lead to rejection of future demands.
- The previous algorithms will perform negatively when they operate in an online routing situation, where the tunnel setup requests arrive one at a time, and the future demand is unknown. Those algorithms require the knowledge of future demands to operate successfully.
- The previous algorithms are not adaptive to possible link failures. Therefore, in a situation where a link fails, those algorithms will not be able to route requests through alternative routes.

Online TE Routing Algorithms

Kar et al. (2000) identified a set of properties that a "practically useful" TE algorithm (based on MPLS) should have. Some of these properties are

- The algorithm should be based on an *online* routing model, where LSP setup requests arrive one at a time (not all at once), and the future demand is unknown. On the other hand, in an *off-line* model, all LSP setup requests are known a priori, and there are no demands for future LSP setup requests.
- It should be able to use knowledge about the *physical locations* of the ingress-egress router pairs through which an LSP is set up.
- The algorithm should be able to *adapt to possible link failures* in the network. In other words, a good TE algorithm should be able to reroute post-failure requests through alternate routes.
- It should be able to route the requested bandwidth *without splitting the demands* as much as possible through multiple paths. This is necessary because it often occurs that the nature of the traffic does not allow a demand to be split. Splitting traffic is, however, a common practice in scenarios like load balancing and network performance improvement.
- The algorithm should, if possible, support a *distributed implementation*, where instead of performing the route computation in a centralized server, the computations of each LSP's route request is distributed at the local ingress node.
- It is quite desirable that such an algorithm is capable of using different *policy constraints*. For example, service level agreements might impose a restriction that LSPs with less than a threshold value of flow guarantees should not be accepted (Kar et al., 2000; Suri et al., 2003).
- Such an algorithm should operate under strict bounds of computational complexity. The algorithms should be very fast and execute within a fixed time constraint. The amount of computation involved with LSP setup requests should be minimized so that the algorithm

can be implemented on a router or a route-server (Kar et al., 2000; Suri et al., 2003).

The MIRA Algorithm

The most influential and primitive online routing algorithm in TE was the MIRA algorithm of Kar et al. (2000). MIRA is online because it does not require a priori knowledge of the tunnel requests that have to be routed. The algorithm is targeted towards applications such as the setup of LSPs in MPLS networks. The algorithm helps service providers setup bandwidth guaranteed tunnels at ease in their transport networks. The terminology *minimum interference* used in the nomenclature of the algorithm indicates that the tunnel setup request is to be routed through a path (or a path segment) that must not interfere too much with future tunnel setup requests. The algorithm aims to protect the "critical links" in a network from being overloaded and thereby reducing the chances of rejection of future requests. The critical links are identified by the algorithm to be those that if congested because of heavy loading might lead to rejection of requests. Unlike its predecessor algorithms, MIRA uses any available information about ingress-egress nodes for potential future demands. MIRA is based on the core concepts of the *max-flow* and the *min-cut* computations of the area of network flows (Ahuja, Magnanti, & Orlin, 1993).

The RRATE Algorithm

The RRATE algorithm, proposed by Oommen et al. (in press), is perhaps one of the most representative state-of-the-art algorithms. They proposed a new class of solutions by incorporating the family of stochastic random-races (RR) algorithms. The most popular previously proposed TE solutions attempt to find a superior path to route an incoming path setup request. Their algorithm, on the other hand, tries to learn an optimal ordering of the paths through which requests can be routed according to the *rank* of the paths in the order learned by the algorithm. They showed that their algorithm has better performance than the important algorithms in the literature. Their conclusions are based on three important performance criteria: (1) the rejection ratio, (2) the percentage of accepted bandwidth, and (3) the average route computation time per request. While some of the previously proposed algorithms were designed to achieve low rejections and high throughput of route requests, they are unreasonably slow. Their algorithm, on the other hand, in general, attempts to reject the least number of requests, achieves the highest throughput, and computes routes in the fastest possible time, as compared to the algorithms we used as benchmarks for comparison.

FUTURE TRENDS

The future work in this area could involve the investigation of the following issues:

- testing the performance of all the aforementioned algorithms on complex massively sized networks; and
- testing the algorithms on real networks, under real traffic conditions.

Most of the aforementioned algorithms were evaluated on idealistic situations where there were no network link failures. With regard to future work, it would be interesting to consider situations where there are realistic link-ups and downs occurring in a network. It would also be interesting to see how all the algorithms can perform "rerouting" of paths if a desired path fails.

CONCLUSION

This article discussed the concepts of QoS and TE routing in networks and also some of the popular algorithms in this area. QoS and TE are very popular among networking researchers and practitioners because of the increased demand for satisfying multiple customer demands and obtaining increased utilization of network resources, while satisfying the varied user requirements. Different directions of future work for network professionals are also articulated.

REFERENCES

Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). Network flows: Theory algorithms, and applications. Upper Saddle River, NJ: Prentice Hall.

Bellman, R. (1958). On a routing problem. *Quarterly of Applied Mathmatics, 16*, 87-90.

Cormen, T. H., Leiserson, C. E., & Rivest, R. L. (1990). *Introduction to algorithms.* Cambridge, MA: MIT Press.

Dijkstra, E.W. (1959). A note on two problems in connection with graphs. *Numerishe Mathmatik* 1(2), 69-271.

Guerin, R., Orda, A., & Williams, D. (1997). QoS routing mechanisms and OSPF extensions. *Proceedings of the Global Internet Miniconference, November.*

Iliadis, I., & Bauer, D. (2002). A new class of online minimum-interference routing algorithms. *NETWORKING 2002,* (LNCS 2345, pp. 959-971).

Kar, K., Kodialam, M., & Lakshman, T. V. (2000). Minimum interference routing of bandwidth guaranteed tunnels with MPLS traffic engineering applications. *IEEE Journal of Selected Areas in Communications, 18*(12), 2566-2579.

Ma, Q. (1998). *Quality-of-service routing in integrated services networks.* Unpublished doctoral thesis, Carnegie Mellon University, Pittsburgh, PA.

Oommen, B. J., Misra, S., & Granmo, O.-C. (2006). A stochastic random-races algorithm for routing in MPLS traffic engineering. *Proceedings of IEEE INFOCOM, 2006.*

Osborne, E., & Simha, A. (2002). *Traffic engineering with MPLS.* Indianapolis, IN: Pearson Education Cisco Press.

Peterson, L., & Davie, B. (2000). *Computer networks: A systems approach* (2nd ed.). San Franciso, CA: Morgan Kauffmann.

Suri, S., Waldvogel, M., Bauer, D., & Warkhede, P. R. (2003). Profile-based routing and traffic engineering. *Computer Communications, 26,* 351-365.

Szeto, W., Boutaba, R., & Iraqi, Y. (2002). Dynamic online routing algorithm for MPLS traffic engineering. *Proceedings of NETWORKING 2002,* (LNCS 2345, pp. 936-946).

Vasilakos, A. V., & Papadimitriou, G. (1992). Ergodic discretized estimator learning automata with high accuracy and high adaptation rate for nonstationary environments. *Neurocomputing, 4,* 181-196.

Vasilakos, A., Saltouros, M. P., Atlassis, A. F., & Pedrycz, W. (2003). Optimizing QoS routing in hierarchical ATM networks using computational intelligence techniques. *IEEE Transactions on System, Man, and Cybernetics., Part C, 33*(3), 297-312.

Wang, Z., & Crowcroft J. (1996). Quality-of-service routing for supporting multimedia applications. *IEEE Journal of Selected Areas in Communications, 14*(7), 1228-1234.

Wang, B., Su, X., & Chen, P. (2002). Efficient bandwidth guaranteed routing algorithms. In *Proceedings of the IEEE International Conference on Communications (ICC '2002),* New York, USA.

KEY TERMS

Algorithm: An algorithm is a set of clear steps that is used to define how a task or a set of tasks can be accomplished.

Label Switched Path (LSP): The paths along which the labeled packets in MPLS networks are transmitted.

Label Switched Router (LSR): An LSR is a backbone router in the physical network topology that runs the existing layer 3 IP protocol.

Multi-Protocol Label Switching: A popular protocol for traffic engineering operating in layer 3, in synchronization with the Internet protocol (IP).

QoS Routing: QoS routing is concerned with selecting routing paths while meeting strict end-to-end service requirements involving resource constraints, while achieving optimum throughput in the network.

Routing: The mechanism by which a path or a set of paths is selected to send information or commodities.

Traffic engineering (TE): An engineering technique or techniques where information and goods can be efficiently transmitted or transported.

Quantum Cryptography Protocols for Information Security

Göran Pulkkis
Arcada Polytechnic, Finland

Kaj J. Grahn
Arcada Polytechnic, Finland

INTRODUCTION

Quantum cryptography will have a severe impact on information security technology. The objective of this article is to present state-of-the-art and future possibilities of two quantum cryptography protocol types. These protocols are for absolutely secure distribution of symmetric encryption/decryption keys and for creating secure digital signatures.

BACKGROUND

In the early 1980s, it was observed that the stochastic parallelism of quantum states cannot be simulated efficiently on a classical computer (Feynman, 1982). This observation started research on using quantum mechanical effects for more efficient information processing than is achievable with classical computers. During the 1980s, the operating principles and implementation possibilities of quantum computing were outlined at Oxford University (Deutsch, 1985). In 1984, a quantum protocol for information transfer with provable confidentiality was proposed (Bennett & Brassard, 1984). This protocol, called the BB84 protocol, uses quantum states implemented by randomly polarized photons.

In the 1990s, efficient algorithms based on the operating principles of quantum computing were proposed. Shor's quantum algorithm for integer factorization (Shor, 1994) has polynomial (cubic) computational complexity and has been experimentally verified in a quantum computer with seven qubits implemented using nuclear magnetic resonance (Vandersypen, Steffen, Breyta, Yannoni, Sherwood, & Chang, 2001). Search with Grover's (1996) quantum algorithm in an unsorted database has only square root computational complexity. The BB84 protocol has been used for secure distribution of symmetric encryption/decryption keys (Quantum Key Distribution, QKD) in research networks (BBN Technologies, 2005; Elliot, 2004; Quellette, 2005;). For some years also commercial QKD technology has been available (id Quantique Portal, 2005; MagiQ, 2005).

The RSA algorithm is often used to create digital signatures. RSA is secure because the best known factorization method for large integers has superpolynomial computational complexity in a classical computer. False signed messages could however be created with a sufficiently large quantum computer since the private signing key could easily be computed with Shor's algorithm from the corresponding public key. On the other hand, secure digital signatures could be created by manipulation and measurements of quantum states (Gottesman & Chuang, 2001; Lu & Feng, 2005).

INFORMATION REPRESENTATION WITH QUANTUM STATES

Quantum states are energy levels of molecules, atoms, and photons. Two quantum states for which a state transition exists can be used to represent an information bit, if the energy levels of both states can be measured. A bit defined by quantum states is called a quantum bit or qubit. However, quantum states are probabilistic. When the energy level of a molecule, an atom, or a photon is measured, the outcome is one of all possible energy levels, and each possible outcome is associated with a probability. The sum of the probabilities of all possible measurement outcomes is, of course, 1. A qubit is thus also probabilistic. The binary values 0 and 1 are represented by two possible quantum states of a qubit. If the measurement probabilities of these two quantum states are p_0 and p_1, then $p_0+p_1=1$.

Properties of Quantum Bits (Qubits)

A qubit can be treated mathematically by linear algebra as a 2-dimensional vector. The orthogonal base vectors $(1,0)^T$ and $(0,1)^T$ represent the quantum states associated with the binary values 0 and 1, respectively. Usually the **Dirac Notation** is used for qubits as well as for these two orthogonal base vectors, thus $(1,0)^T = |0\rangle$ and $(0,1)^T = |1\rangle$.

A qubit is a superposition of $|0\rangle$ and $|1\rangle$ and both values are simultaneously present. A measured qubit is set to the measured value, which is $|0\rangle$ or $|1\rangle$. Let $|\psi\rangle$ be a qubit in Dirac Notation. Then

$$|\psi\rangle = a\cdot|0\rangle + b\cdot|1\rangle \quad (1)$$

where a and b are complex numbers for which

$$\langle\psi||\psi\rangle = (a^*,b^*)\cdot(a,b)^T = a^*\cdot a + b^*\cdot b = |a|^2 + |b|^2 = 1, \quad (2)$$

{a*,b*} are complex conjugates of {a,b}, and $\{|a|^2,|b|^2\}$are the probabilities to measure the values 0,1, respectively. A qubit has thus three dimensions because complex numbers have two dimensions. From this follows further, that a qubit can be presented geometrically as a point on a 3-dimensional unit sphere.

A fundamental qubit property is the No Cloning Property, according to which it is impossible to clone an unknown quantum state (Nielsen & Chuan, 2002).

Multiple Qubits

The quantum state of **2 qubits** is a column vector with 2^2 = 4 components. For the qubits

$$|\psi_1\rangle = a\cdot|0\rangle + b\cdot|1\rangle \text{ and } |\psi_2\rangle = c\cdot|0\rangle + d\cdot|1\rangle \quad (3)$$

the quantum state is

$$|y_1y_2\rangle = |\psi_1\rangle \otimes |\psi_2\rangle = a\cdot c\cdot|00\rangle + a\cdot d\cdot|01\rangle + b\cdot c\cdot|10\rangle + b\cdot d\cdot|11\rangle \quad (4)$$

where $\otimes$ is the **tensor product** of two column vectors. In such a product, the result vector is obtained by multiplying the latter vector with each component in the first vector, see the following examples: $|00\rangle = |0\rangle\otimes|0\rangle = (1,0,0,0)^T$, $|01\rangle = |0\rangle\otimes|1\rangle = (0,1,0,0)^T$, $|10\rangle = |1\rangle\otimes|0\rangle = (0,0,1,0)^T$, and $|11\rangle = |1\rangle\otimes|1\rangle = (0,0,0,1)^T$ are called the **base vectors** of the 2 qubit state.

An **N qubit** quantum state $|\psi_1\psi_2\ldots\psi_N\rangle$ is a superposition of 2^N base vectors. For a three qubit quantum state $|\psi_1\psi_2\psi_3\rangle$ the 2^3 base vectors are {|000>,|001>,|010>,|011>, |100>,|101>, |110>,|111>}, where $|001\rangle = |0\rangle\otimes|0\rangle\otimes|1\rangle = (0,1,0,0,0,0,0,0)^T$, and so forth.

The qubit state $|\psi_1\psi_2\ldots\psi_N\rangle$ is an **entangled state** if there exists no $|\psi_1\rangle,|\psi_2\rangle,\ldots,|\psi_N\rangle$ for which $|\psi_1\psi_2\ldots\psi_N\rangle = |\psi_1\rangle\otimes|\psi_1\rangle\otimes\ldots\otimes|\psi_N\rangle$.

Example: The 2 qubit state $2^{-½}$ (|00> + |11>) is entangled.

Proof: $(a\cdot|0\rangle+b\cdot|1\rangle)\otimes(c\cdot|0\rangle+d\cdot|1\rangle) = a\cdot c\cdot|00\rangle+a\cdot d\cdot|01\rangle +b\cdot c\cdot|10\rangle+b\cdot d\cdot|11\rangle \neq 2^{-½}\cdot(|00\rangle + |11\rangle)$, since one of {a,d} and one of {b,c} must be 0.

Physical Implementation of Qubits

Qubits have been implemented by ion traps, by cavity quantum electrodynamics (QED), by nuclear magnetic resonance (NMR), and by quantum dots (Nielsen & Chuang, 2002).

A qubit is implemented by the **polarization state** of a photon in practical quantum cryptography. A polarization state consists of all planes in which the electromagnetic wave of a photon propagates. The polarization of a randomly polarized photon is a superposition of any pair of orthogonal states. Examples of orthogonal polarization state pairs are:

- horizontal and vertical polarization;
- +45° and -45° diagonal polarization.

The polarization of a photon can thus be modeled by a qubit $|\psi\rangle$ for which

$$|\psi\rangle = a\cdot|\text{horis}\rangle + b\cdot|\text{vert}\rangle = c\cdot|+45°\rangle + d\cdot|-45°\rangle \quad (5)$$

where a,b,c,d are complex numbers and $|a|^2 + |b|^2 = |c|^2 + |d|^2 = 1$. One polarization state must be interpreted as |0> and the other state as |1> for a chosen orthogonal state pair. For example,

- $|\text{horis}\rangle = |0\rangle$ and $|\text{vert}\rangle = |1\rangle$
- $|+45°\rangle = |0\rangle$ and $|-45°\rangle = |1\rangle$.

Notice also that (c,d) can be calculated from (a,b) and vice versa since

$$|+45°\rangle = 2^{-½}(|\text{horis}\rangle + |\text{vert}\rangle),\ |-45°\rangle = 2^{-½}(|\text{horis}\rangle - |\text{vert}\rangle). \quad (6)$$

The polarization of a photon is measured with a filter. After measurement, only the component defined by the filter can pass through. For example, the polarization will change to $|\psi\rangle = |\text{horis}\rangle$ when a photon with polarization $|\psi\rangle = a\cdot|\text{horis}\rangle + b\cdot|\text{vert}\rangle$ passes through a horizontal filter. If a=0 and b=1, then the photon is absorbed by a horizontal filter.

A consequence of the No Cloning Property of a qubit is that an unknown polarization state of a photon cannot be copied to any other photon.

QUANTUM INFORMATION PROCESSING

Quantum Gates and Circuits

Qubit states are changed with quantum gates. Basic single qubit quantum gates are:

- Identity Gate I, defined as $I\cdot|\psi\rangle = |\psi\rangle$ for a qubit state $|ø\rangle$
- Negation Gate X, defined as $X\cdot|0\rangle = |1\rangle$ and $X\cdot|1\rangle = |0\rangle$
- Phase Shift Gate Z defined as $Z\cdot|0\rangle = |0\rangle$ and $Z\cdot|1\rangle = -1\cdot|1\rangle$
- Hadamard Gate, H defined as $H\cdot|0\rangle = 2^{-½}\cdot(|0\rangle + |1\rangle)$ and $H\cdot|1\rangle = 2^{-½}\cdot(|0\rangle - |1\rangle)$.

Basic 2 qubit quantum gates are:

- Identity Gate I, defined as $I\cdot|\psi_1\psi_2\rangle = |\psi_1\psi_2\rangle$ for a two qubit state $|\psi_1\psi_2\rangle$
- Controlled-NOT C_{not}, defined as $C_{not}\cdot|0\psi\rangle = |0\psi\rangle$ and $C_{not}\cdot|1\psi\rangle = |1\rangle\otimes X\cdot|\psi\rangle$ for a qubit state $|\psi\rangle$ where $\otimes$ is the tensor product.

Also N qubit gates for N>2 can be defined, for example an N qubit Identity Gate. An N qubit gate has a $2^N * 2^N$ matrix representation G for which $G\cdot G^* = I$ where G^* is the complex conjugate matrix of G.

A quantum circuit is an interconnection of quantum gates. In Figure 1, a quantum circuit for teleportation of an unknown qubit state is shown. The state of the qubit $|\psi\rangle = a\cdot|0\rangle + b\cdot|1\rangle$ is teleported to the qubit $|\theta\rangle$. The upper qubit is the control bit in both C_{not} gates. m1 and m2 are measured qubit values. $X^{m2} = I$ for m2 = 0, $X^{m2} = X$ for m2 = 1, $Z^{m1} = I$ for m1 = 0, and $Z^{m1} = Z$ for m1 = 1. Notice that teleportation is not in contradiction with the No Cloning Property of a qubit, since the original state of the qubit $|\psi\rangle$ is lost in the measurement m1 and the measurement result m1 is needed in the teleportation.

Quantum Algorithms

An algorithm in quantum information processing is based on qubit state modifications with quantum gates and qubit measurements. Any algorithm in quantum information processing can be implemented by:

- quantum circuits of interconnected single qubit and Controlled-NOT quantum gates;
- qubit state measurements.

QUANTUM CRYPTOGRAPHY

Quantum cryptography means development and application of cryptographic protocols based on measurement and manipulation of quantum states. Technology for absolutely secure distribution of session and transaction keys to cryptographic network software with quantum protocols has been available for some years. Quantum protocols for signing both classical bit strings and quantum states have been proposed and thoroughly examined for future technology solutions (Gottesman & Chuang, 2001; Lu & Feng, 2004).

Quantum Key Distribution (QKD)

In this section, proposed QKD protocols are surveyed and integration of these protocols in the TCP/IP network protocol stack is outlined. Available QKD technology and present QKD applications are described. Security threats to QKD applications are also discussed.

QKD Protocols

The communication architecture and the phases of the BB84 protocol are shown in Figure 2.

1. Alice sends over an optical channel a sequence of polarized photons to Bob. Each photon is randomly polarized to a state in the set {|horis>, |vert>, |+45°>, |-45°>}. Bob guesses the polarization base (horizontal/vertical or diagonal) of each photon and measures the polarization in the guessed base. Polarization is correctly measured only for a correctly guessed polarization base, since measurement outcome is with equal probability one of the two orthogonal states of a wrongly guessed polarization base.

Figure 1. A quantum circuit for teleportation of the unknown qubit state $|\psi\rangle$ to the qubit $|\theta\rangle$

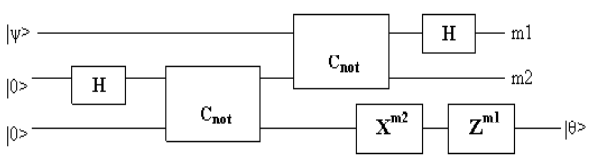
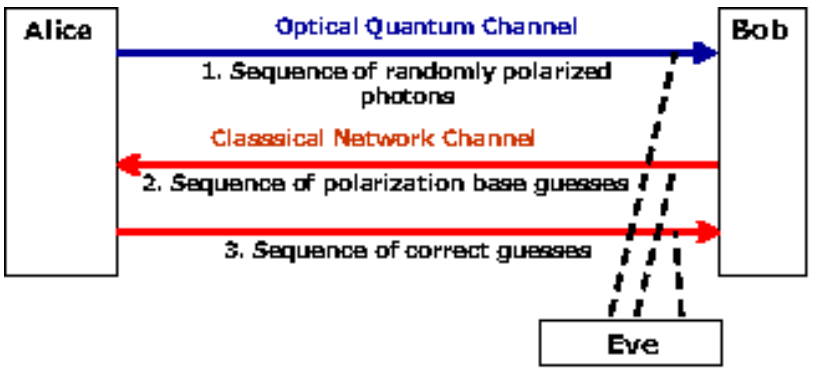


Figure 2. Principle of the BB84 QKD protocol



2. Bob sends over a classical network channel the sequence of polarization base guesses to Alice.
3. Alice returns to Bob the sequence of correct guesses.

Now Alice and Bob can use Bob's correctly measured polarizations as a shared secret (key), which consists of the polarization states of about 50% of the photons sent by Alice.

For the notation

- **u** is horizontal/vertical polarization base with **h** = |horis> as 0 and **v** = |vert> as 1
- **d** is diagonal polarization base with + = |+45°> as 0 and - = |-45°> as 1
- **r** is measurement of 0 or 1 with equal probability, **t** is true and **f** is false.

Assume, that

- Alice sends +hv++h-vvhv+—vh-v-
- Bob guesses and sends uuuuuudduuududddduud
- Bob measures r01rr01r1010r11rrr11
- Alice sends fttffttfttttfttfffft.

Then Alice and Bob share the secret 010110101111, since these bit values are correctly measured and they have not been revealed in the communication between Alice and Bob.

Assume that Eve (eavesdropper) measures the polarization of each photon in Alice's sequence in a guessed basis and forwards the measured photons to Bob. This means that a random choice of 0 or 1 is forwarded when Eve's polarization basis guess is incorrect. Since the probability of an incorrect guess is 50 %, about half of the bits are randomized in the bit sequence, which Alice and Bob think is a shared secret. Let **e** be an incorrect guess of Eve:

- Alice sends +hv++h-vvhv+—vh-v-
- Eve guesses dudddduudduduuuudddd
- Eve sends +he++eeveev+eeeve-e-
- Bob guesses and sends uuuuuudduuududddduud
- Bob measures r0erreeree10reerrre1
- Alice sends fttffttfttttfttfffft.

Now Alice thinks that she shares the secret 010110101111 with Bob, but Bob thinks that he shares the secret 0eeeee10eee1 with Alice. Here e is 0 or 1 with equal probability because of the intervention of Eve. About four of the eight e bits are probably correct but the eavesdropper Eve also knows these 4 bits.

The original BB84 protocol is presently the Sifting Phase of a QKD protocol for production of the raw keys of Alice and Bob. Sifting means that photons, which fail to reach the receiver, are omitted from the raw keys. Practical use of the BB84 protocol must include a Key Distillation process, which consists of Error Correction and Privacy Amplification. The shared secret will then include the bits extracted from the raw keys in the Key Distillation process. Error Correction means that the bits, which are different in the two raw keys, are detected and removed. Privacy Amplification means that as many as possible of the correct bits known by a possible eavesdropper are dropped from both raw keys. If both raw keys are equal, then any same bit subset of the required shared secret key size can be taken from both raw keys. The reason is that the probability for an eavesdropper to correctly guess the polarization basis of all photons corresponding to the bits in the raw keys is $O(2^{-N/2})$. Here, N is the length of the sequence of randomly polarized photons (Elliott, Pearson & Troxel, 2003).

Since the BB84 protocol is based on random choice from four different polarization states, it has been called a 4-state QKD protocol. It can be proved that only two non-orthogonal polarization states are actually needed to implement a perfectly secret QKD protocol (Bennett, 1992). Also,

perfectly secret 6-state QKD protocols have been proposed and analyzed (Bruss, 1998).

QKD Protocol Integration

Quantum Repeaters

With optical communication technology, one can transmit polarized photons with reasonable error rates over distances up to about 70…100 km (MagiQ, 2005; Stucki, Gisin, Guinnard, Ribordy, & Zbinden, 2002). For longer distances the current QKD protocols would need quantum repeaters (see Figure 1 for functional principle) on the physical layer of the network protocol stack. However, technology for optical quantum repeaters has still not been developed. Research efforts to implement a quantum repeater are presented in (Curcic et. al, 2004).

QKD Key Agreement

Secure network communication protocols for TCP/IP networks like IPSec, TLS/SSL, and SSH presently use the Diffie-Hellman protocol or RSA key transport for session key agreement. The use of a QKD protocol in these client/server communication protocols may require additional optical network links with own Ethernet interfaces.

QKD protocol integration in the IKE Deamon of the IPSec protocol is described in Elliott, Pearson, and Troxel (2003). Protocol phases are Sifting, Error Correction, Privacy Amplification, and Authentication before a session key can be included in a new Security Association (SA).

New cipher suites defining key exchange are prerequisites for QKD protocol integration in the TLS/SSL Handshake Protocol. New ServerKeyExchange and ClientKeyExchange messages must also be defined. The same phases as for QKD protocol integration in IPSec are needed before a shared secret is available in the SSL client and server. This shared secret could be used as Premaster Secret in the TLS/SSL handshake protocol.

Attacks Against QKD Systems

Security of quantum cryptography has been proven for any eavesdropping attack limits of security have been established for the case of a noisy environment and imperfect detection as well as of using non-ideal light sources. In all these cases, Eve performs her measurements on the quantum states transmitted from Alice to Bob (Vakhitov, Makanov, & Hjelme, 2001).

The real security of a quantum cryptography system will also be determined by technological implementations, technical measures, and by loopholes in the optical system. Attacks do not deal with quantum states but use loopholes and imperfections in implementations (Gisin, Ribordy, Tittel, & Zbinden, 2002). Optical loophole attacks are:

- **Large Pulse Attack:** A strong light pulse launched into a QKD system may be partly reflected. Measuring characteristics of pulses reflected via internal modulators of the system may allow the intruder to detect transmitted quantum states unambiguously (Bethune & Risk, 2000).
- **High-power destruction of optical components:** Damaging an optical component by using high-power external pulses could, for instance, make beam splitting and other quantum attacks more efficient.
- **Light emission from avalanche photodiodes (APDs, see Key Terms):** During an avalanche, an APD emits light over a broad spectrum and a part of this light leaks back into the system. An intruder may detect this leak.
- **Faked States Attack:** A faked states attack is an intercept-and-resend attack. Imperfections in Bob's scheme are utilized by Eve to generate light pulses not discovered by any alarm and legitimate parties are fooled.

QKD Technology and Applications

QKD works through optical telecommunication fibers or through the atmosphere. Data is transmitted faster in free space systems, but the distance is a limiting factor. Key data rates close to 5,000 bits/s have been reached at distances of up to 50 km through optical fiber and 10 to 20 km through atmosphere (Elliott, 2004). Single photons allow only slow real-time data transmission because of the limited detection rate of single photons.

The first commercial quantum cryptography products, for example single photon detectors and random-number generators, were introduced in 2002. Encoding was based on the phase of the photon not on the polarization state. In 2005, a new quantum cryptography Link Encryptor was presented, which combines QKD with AES (Advanced Encryption Standard) and securely bridges two Fast Ethernet networks. Key refresh rates up to 100 times per second and a transmission distance of 100 km in an optical fiber are supported (id Quantique Portal, 2005).

A quantum cryptography based solution called QPN™ Security Gateway adds layers of VPN security and classical data encryption to QKD. Protocols like BB84, IPSec, and 256 bit AES are implemented. A key refresh rate up to 100 keys/second, a transmission distance of 100 km, and full-duplex 10/100 Ethernet ports are supported (MagiQ, 2005).

The world's first quantum cryptographic network interconnects BBN Technologies, Harvard, and Boston University. This network has 10 nodes and uses different QKD protocols with a bandwidth up to 5 Mbit/s (BBN Technologies, 2005).

Quantum Digital Signature

Classical digital signatures can be created from any one-way function. A sender (Alice) chooses two binary private keys {***k0, k1***} and publicly announces ***f***, {***0,f(k0)***}, {***1,f(k1)***}, where f is the used one-way function. To sign a single bit ***b***, Alice sends {***b,kb***} to receivers, who can compare ***f(kb)*** with the earlier announcement. The signature is verified if ***f(kb)=f(k0)*** or ***f(kb)=f(k1)***, since only Alice can send a correct ***kb***. The public keys can, however, only be used once (Lamport, 1979).

A secure quantum digital signature scheme based on the use of quantum one-way functions is proposed in Gottesman and Chuang, (200)1. Alice

1. chooses a one-way quantum circuit ***f***, which maps a classical bit string ***k*** of length ***L*** to an ***n*** qubit state $|f_k\rangle$.
2. chooses ***M*** different pairs of private bit strings {***k0,k1***}. ***M*** depends on the application environment.
3. announces the structure of the quantum circuit ***f*** and ***M*** different $\{(0,|f_{k0}\rangle),(1,|f_{k1}\rangle)\}$ pairs to ***T*** recipients.

To sign a single bit message ***b***

1. Alice sends ***b*** in combination with ***M kb*** values over an insecure classical communication channel to all ***T*** recipients and thus reveals half of her private bit strings. Each recipient checks all ***M kb*** values with a swap quantum circuit (Buhrman, Cleve, Watrous, & De Wolf, 2001) to find out for each ***kb*** if the qubit state $|f_{kb}\rangle$ is identical with $|f_{k0}\rangle$ or with $|f_{k1}\rangle$. The check fails if $|f_{kb}\rangle$ is different from both $|f_{k0}\rangle$ and $|f_{k1}\rangle$.
2. Each recipient counts the number ***s*** of failed checks.
3. All ***T*** recipients must also know two numbers, c_1 and c_2, thresholds for acceptance and rejection. Both threshold values depend on the application environment. Each recipient
   - accepts the message ***b*** as valid and transferable (***result 1-ACC***) if $s<c_1 \cdot M$.
   - rejects message ***b*** as invalid (***result REJ***) if $s>c_2 \cdot M$.
   - otherwise concludes the message b to be valid but not necessarily transferable to other recipients (***result 0-ACC***).
4. All used and unused keys are discarded.

The number of recipients ***T*** must be less than ***L/n***, since measurements of the public state $|f_k\rangle$ can reveal up to ***n*** bit values in ***k***. When ***s*** is large, the message has been heavily tampered and may be invalid. c_1 can be zero in the absence of noise. When ***s*** is small, the message cannot have been changed very much from what Alice sent. ***s*** is similar for all recipients, but ***s*** values must not be identical. Forgery is prevented by c_2, and cheating by Alice is prevented by the difference between c_2 and c_1. The results ***1-ACC*** and ***0-ACC*** verify that Alice has sent the message. ***1-ACC*** means verification to the recipient that any other recipient will also find the message valid. Thus, the message is "transferable". ***0-ACC*** means, that a second recipient might find the message invalid. Result ***REJ*** means that the recipient cannot safely reach any conclusion about the authenticity of the message. It should be required that any recipient who receives a correct (message, signature) pair always reaches conclusion ***1-ACC***.

The quantum digital signature scheme (Gottesman & Chuang, 2001) is for signing only classical bit strings. Another quantum digital signature scheme based on quantum one-way functions for signing general quantum states is proposed in (Lu & Feng, 2004). Authentication of quantum messages can be based on schemes for signing quantum states. Quantum message authentication issues are examined in (Barnum, Crepequ, Gottesman, Smith, and Tapp, 2002).

FUTURE TRENDS

Today, the maximum guaranteed transmission distance is short because optical fibers are not perfectly transparent. Random noise degrades the photon stream. Therefore, continued research and development in the field of quantum repeaters is necessary. Researchers are also looking beyond optical fibers as the medium to distribute quantum keys. By optimizing free-space technology, it might be possible to build systems that transmit and receive signals reaching satellites in low earth orbit.

A major factor limiting the development of QKD systems is the lack of fast photon detectors. New detectors based on cryogenic niobium nitride superconductors may be a solution (Quellette, 2005).

The work on ideal single-photon sources continues. If more than one photon is emitted, the system will be vulnerable. An intruder will have the possibility to discover the used polarization and then send the rest of the photons further onto the receiver.

The ongoing EU Sixth Framework Programme, including research on Information Society Technologies, has a strong focus on Quantum Information Processing and Communications in Future and Emerging Technologies (CORDIS ISTweb Future, 2005). An example is the SECOQC ("Development of a Global Network for Secure Communication based on Quantum Cryptography") project, see (CORDIS ISTweb Network, 2005).

CONCLUSION

Quantum cryptography is an emerging information security technology, which is presently already available but still

neither fully integrated in current network technology nor standardized. Technological obstacles regarding quantum transmission distance and quantum device technology must be removed. Two potential information security areas are quantum key distribution and quantum digital signatures. Needed innovations will renew the classical information security technology.

REFERENCES

Barnum, H., Crepeau, C., Gottesman, D., Smith, A., Tapp, A. (2002). Authentication of quantum messages. In *Proceedings of the 43rd IEEE Symposium on the Foundations of Computer Science* (pp. 449-458).

BBN Technologies. (2005). *News and events*. Retrieved May 4, 2005, from http://www.bnn.com

Bennett, C.H. (1992). Quantum cryptography using any two nonorthogonal states. *Phys. Rev. Lett*., 68, 3121-3124.

Bennett, C.H., & Brassard, G. (1984, December 10-12). Quantum cryptography: Public key distribution and coin tossing. *International Conference on Computers, Systems & Signal Processing*, Bagalore, India (pp. 175-179).

Bethune, D.S., & Risk, W.P. (2000). An autocompensating fiber-optic quantum cryptography system based on polarization splitting of light. *IEEE J. Quantum Electron*., *36*(3), 340-347.

Bruss, D. (1998). Optimal eavesdropping in quantum cryptography with six states. *Phys. Rev. Lett., 87*.

Buhrman, H., Cleve, R., Watrous, J., & de Wolf, R. (2001). Quantum fingerprinting. *Phys. Rev.Lett. 87*.

CORDIS ISTweb Future and Emerging Technologies. (2005). *Quantum information processing & communications*. Retrieved April 30, 2005, from http://www.cordis.lu/ist/fet/qipc.htm

CORDIS ISTweb Network and Communiocation Technologies. (2005). *ICT for trust and security*. Retrieved April 30, 2005, from http://www.cordis.lu/ist/directorate_d/trust-security/projects.htm

Curcic, T., Filipkowski, M.E., Chtchelkanova, A., D'Ambrosio, P. A., Wolf, S.A., et al. (2004). Quantum networks: From quantum cryptography to quantum architecture. *ACM SIGCOMM Computer Communications Review, 34*(5).

Deutsch, D. (1985). Quantum theory, the church-turing principle and the universal quantum computer. *Proceedings of the Royal Society of London Series A A400* (pp. 97-117).

Elliot, C. (2004, July / August). Quantum cryptography. *IEEE Security & Privacy*, *2*(4), 57-61

Elliott, C., Pearson, D., & Troxel, G. (2003). Quantum cryptography in practice. In *Proceedings of the ACM SIGCOMM* (pp. 227-238). ACM Press.

Feynman, R. (1982). Simulating physics with computers. *International Journal of Theoretical Physics, 21*(6&7), 467-488.

Gisin, N., Ribordy, G., Tittel, W., & Zbinden, H. (2002). Quantum cryptography. *Rev. Mod. Phys*., *74* (1), 145-195.

Gottesman, D., & Chuang, I. (2001). *Quantum digital signatures*. Retrieved August 10, 2005, from http://arxiv.org/abs/quant-ph/0105032

Grover, L. K. (1996, May 22-24). A fast quantum mechanical algorithm for database search. In *Proceedings of the 28th Annual ACM Symposium on the Theory of Computing*, Philadelphia (pp. 212-219).

id Quantique Portal. (2005). Retrieved May 4, 2005, from http://www.idquantique.com

Lamport, L. (1979, October). *Constructing digital signatures from a one-way function*. Technical Report CSL-98, SRI International.

Lu, X., & Feng, D. (n.d.). Quantum digital signatures based on quantum one-way functions. In *Proceedings of the 7th International Conference on Advanced Communication Technology ICACT* (Vol. 1, pp. 514-517).

MagiQ. (2005). *Quantum information solutions for the real world*. Retrieved May 4, 2005, from http://www.magiqtech.com

Nielsen, M.A., & Chuan, I.L. (2002). *Quantum computation and quantum information*. UK: Cambridge University Press.

Quellette, J. (2005, December/January). Quantum key distribution. *The Industrial Physicist*, 22-25.

Shor, P. W. (1994, November). Algorithms for quantum computation: Discrete log and factoring. In *Proceedings of the 35th Annual Symposium on Foundations of Computer Science* (pp. 124-134).

Stucki, D., Gisin, N., Guinnard, O,. Ribordy G., & Zbinden, H. (2002). Quantum key distribution over 67 km with a PlugPlay system. *New Journal of Physics, 4*, 41.1-41.8.

Vakhitov, A., Makarov, V., & Hjelme, D. R. (2001). Large pulse attack as a method of conventional optical eavesdropping in quantum cryptography. *Journal of Modern Optics*, 2023-2038.

Vandersypen, L., Steffen, M., Breyta, G., Yannoni, C., Sherwood, & Chuang, I. (2001). *Experimental realization of Shor's quantum factoring algorithm using magnetic resonance.* 414, p. 883.

KEY TERMS

Avalanche Photodiode: A photodetector which can be regarded as the semiconductor analog to a photomultiplier.

Photon: A discrete packet of electromagnetic energy.

Polarization: The plane in which an electromagnetic wave propagates.

QPN: A registered abbreviation for Quantum Private Network, which is a VPN link where session key agreement is based on a QKD protocol.

Quantum Gate: A device for changing the state of one or of multiple qubits.

Raw Key: Two parties agree on a shared binary secret from a sequence of randomly polarized photons.

Teleportation: Transfer of an unknown qubit state.

Real Options Analysis in Strategic Information Technology Adoption

Xiaotong Li
University of Alabama in Huntsville, USA

INTRODUCTION

Many information resource managers have learned to be proactive in today's highly competitive business environment. However, limited financial resources and many uncertainties require them to maximize their shareholders' equity while controlling the risks incurred at an acceptable level. As the unprecedented development in information technology continuously produces great opportunities that are usually associated with significant uncertainties, technology adoption and planning become more and more crucial to companies in the information era. In this study, we attempt to evaluate IT investment opportunities from a new perspective, namely, the real options theory. Its advantage over other capital budgeting methods like static discounted cash flow analysis has been widely recognized in analyzing the strategic investment decision under uncertainties (Amram & Kulatilaka, 1999; Luehrman, 1998a, 1998b). Smith and McCardle (1998, 1999) further show that option pricing approach can be integrated into standard decision analysis framework to get the best of the both worlds. In fact, some previous IS researches have recognized the fact that many IT investment projects in the uncertain world possess some option-like characteristics (Clemsons, 1991; Dos Santos, 1991; Kumar, 1996). Recently, Benaroth and Kauffman (1999) and Taudes, Feurstein and Mild (2000) have applied the real options theory to real-world business cases and evaluated this approach's merits as a tool for IT investment planning.

As all real options models inevitably depend on some specific assumptions, their appropriateness should be scrutinized under different scenarios. This study aims to provide a framework that will help IS researchers to better understand the real options models and to apply them more rigorously in IT investment evaluation. As the technology changes, the basic economic principles underlying the real options theory do not change. We do need to integrate the IT dimension into the real options based investment decision-making process. Using electronic brokerage's investment decision in wireless technology as a real-world example, we show the importance of adopting appropriate real options models in IT investment planning. By specifically focusing on the uncertainties caused by IT innovation and competition, our study also gives some intriguing results about the dynamics between IT adoption and the technology standard setting process.

REAL OPTIONS THEORY

It is generally believed that the real options approach will play a more important role in the highly uncertain and technology driven digital economy. Before reviewing the real options literature body that is growing very rapidly, we use an example to give readers an intuitive illustration of the values of real options and their significance in financial capital budgeting.

Pioneer Venture: The Value of a Growth Option

In this example, the management of a large pharmaceutical company wants to decide whether to acquire a young biomedical lab. If they decide to acquire it, they should provide $100,000 funding to cover the initial costs for the pioneer venture. Five years after the initial funding, the management will decide whether to stop the pioneer venture or to expand it significantly according to the market situation at that time. If they choose to expand it, additional $1,000,000 is needed. The cost of capital is assumed to be 15%. Five years after acquisition of the lab, the management will face two scenarios. The good scenario will occur with 60% likelihood, while the bad one will have 40% likelihood of happening. All expected future cash flows during the next 10 years are given in Table 1. Using standard capital budgeting method, we can find that the NPV for the pioneer venture is -$15,215. For the period of large-scale production, the NPV is -$71,873. As the NPVs for both periods are negative, it seems that the management should give up the acquisition. However, the acquisition will be a good investment if we consider the growth option associated with it. By acquiring the lab, the company also buys a growth option that enables it to expand the lab when the conditions are favorable 5 years later. In this case, the good scenario will occur with 60% likelihood. After simple calculation, it is easy to find that the growth option has a value of $28,965. Combining

Table 1. Projected cash flows in the example of pioneer venture project

Year	Pioneer Stage	Larger Scale Stage	Total Cash Flows	Discount Rate
0	-$100,000		-$100,000	15%
1	$10,000		$10,000	
2	$10,000		$10,000	
3	$50,000		$50,000	
4	$50,000		$50,000	
5	$20,000	-$1,000,000	-$980,000	
6		$100,000	$100,000	
7		$100,000	$100,000	
8		$500,000	$500,000	
9		$500,000	$500,000	
10		$200,000	$200,000	

	Large Scale Stage	Good Scenario	Bad Scenario	Prob (good)
5	-$1,000,000	-$1,000,000	-$1,000,000	0.6
6	$100,000	$130,000	$55,000	
7	$100,000	$130,000	$55,000	
8	$500,000	$650,000	$275,000	
9	$500,000	$650,000	$275,000	
10	$200,000	$260,000	$110,000	

NPV Pioneer Stage	-$15,215.42
NPV Large Scale Stage	-$71,872.54
NPV with Growth Option	$13,749.98
Value of the Option	$28,965.40

its value with the negative NPV during the pioneer venture period, the adjusted NPV of the acquisition is $13,750, which means this investment is strategically plausible.

Many researchers recognized the potential of this options pricing theory in capital budgeting because traditional DCF (discounted cash flows) technique has its inherent limitation in valuing investments with strategic options and many uncertainties. Table 2 gives a comparison between an American call option on a stock and a real option on an investment project. Despite the close analogy, some people may still question the applicability of option pricing theory on real options that are usually not traded in a market. However, Cox, Ingersoll and Ross (1985) and McDonald and Siegel (1984) suggest that a contingent claim on a non-traded asset can be priced by subtracting a dividend like risk premium from its growth rate.

Recent development in real option theory focuses on the valuation of more complicated real options like shared options, compounded options and strategic growth options. Dixit and Pindyck (1994) examine the dynamic equilibrium in a competitive industry. Their model suggests that a firm's option to wait is valuable when uncertainty is firm-specific. For industry-wide uncertainty, there is no value to wait because of the asymmetric effects of uncertainty.

FOUR CATEGORIES OF IT INVESTMENT OPPORTUNITIES

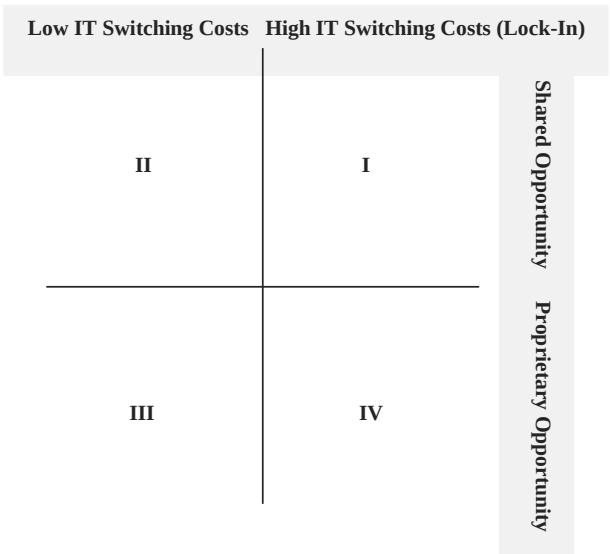
As shown in Figure 1, we have four types of IT investment opportunities based on the two criteria: (i). Shared opportunities with high IT switching costs; (ii). Shared opportunities with low IT switching costs; (iii). Proprietary opportunities with low IT switching costs; (iv). Proprietary opportunities with high IT switching costs. It is worth noting that each category has distinctive requirements on the application of real options models. We use the continuous-time model developed in McDonald and Siegel (1986) as a benchmark to show why we differentiate IT investment opportunities based on the two criteria. It basically suggests that the option to defer uncertain investment is very valuable and should be taken into account when a company makes investment decisions. A major assumption of this model is that there is no competitive erosion; in other words, the investment

Table 2. Comparison between an American call option and a real option on a project

AMERICAN CALL OPTION ON STOCK	REAL OPTION ON A PROJECT
Current Stock Price	Present Value of Expected Cash Flows
Option Exercise Price	Investment Cost of a Project
Right to Exercise the Option Earlier	Right to Invest in the Project at any time before the Opportunity Disappears
Stock Price Uncertainty	Project Value Uncertainties
Option Price	Value of Managerial Flexibility Associated with the Project
Expiration Time	Time Window of the Investment Opportunity
Traded in Financial Market	Usually not Traded
Easy to Find a Replicating Portfolio	Hard to Find a Replicating Portfolio

project is a proprietary opportunity. Without this assumption, the value of the project should not follow the symmetric geometric Brownian motion described in their model. The reason is simple: the existence of potential competition makes the distribution of future project value asymmetric, with high project value less likely to occur. It is worth noting that the well-known Black-Scholes option pricing formula is also based on the assumption that the underlying asset price follows the geometric Brownian motion. In the real business world, most investment opportunities are shared or at least partially shared. Especially in the IT business sector where intensive competition is pervasive, those real options models assuming symmetric uncertainty in investment opportunity value are generally inappropriate. Intuitively, competition pressure will decrease the value of the option to defer an investment. There are usually two approaches to deal with this issue. One approach is to model the competitive entries as exogenous shocks. For examples, Dixit and Pindyck (1994) and Trigeorgis (1991) use a Poisson Jump process to describe the competitive arrivals. Their studies show that the effect of the competitive erosion can be expressed as the following equation

Figure 1. Four categories of IT investment opportunities



Strategic NPV= NPV+ (Value of Option to Wait - Competitive Loss).

In other words, strong competition will restrict managerial flexibility if the investment opportunity is shared. We need to evaluate the investment opportunity in the context of the real options theory by considering the growth option and the waiting option simultaneously. Alternatively, we can incorporate the preemptive effect into the standard real options models. For example, Li (2001) proposes a real options model with strategic consideration based on the model in McDonald and Siegel (1986).

The other criterion we used to categorize different IT investment opportunities is the IT switching cost. We all know that future uncertainty makes the options embedded in an investment opportunity valuable. Theoretically, there is no need to single out technology uncertainty from all other uncertainties in the real options model. All these uncertainties have the same effect: they make the future

payoff of an investment project less predictable. However, we will concentrate on the technology uncertainty in this study because it plays a pivotal role in affecting IT investment payoff. Perhaps the most important question that management faces before committing an IT investment is whether the technology adopted is the right choice. More specifically, will the adopted technology be the best solution to maximize the expected investment payoff? Clearly there is not a simple answer to this question because there are so many uncertainties involved. Some very promising or popular IT solutions may become obsolete in a few years. In some other cases, some neglected IT solutions may evolve to be the standard solution. Nevertheless, most technology uncertainties can be resolved as the process of technology competition goes forward. A typical process of technology competition includes:

- Problem identification: An important problem is identified and new technology is sought to solve it.
- Technology solutions proposition: Several technology developers/vendors propose different solutions to solve the problem.
- Solution testing and comparison: Different technology solutions are competing in the market and their effectiveness is tested and compared.
- Technology standardization: The best solution will flourish over time. Based on it, the technology to solve the problem will be standardized.

For many IT investment projects, decision makers face an uncertain technology environment where several IT solutions are competing in the market. Obviously, the future successes of these projects will to some extent depend on whether the IT solutions adopted will win the technology competition. Consequently, decision makers do have an incentive to use the deferring option to let more technology uncertainties be resolved. Under this scenario, many option-to-wait models can be easily extended to find the optimal investment strategy. However, to apply these real options models we must presume that there are significant technology switching costs once an IT solution is adopted. Otherwise, the uncertainties in technology competition will not make the option to wait valuable because the decision makers can easily switch to other IT solutions after they implement the investment project. As pointed out by Shapiro and Varian (1998), the IT switching costs are very significant in many cases. They use the term "technology lock-in" to describe the situation where management has little flexibility to switch to other technology solutions once they have adopted one IT solution.

Now it should be clear why we use IT switching cost as the second criterion to classify different IT investment opportunities. When the IT switching cost is significant (technology lock-in), the option to wait is valuable. Therefore, real options analysis should concentrate on the managerial flexibility in deferring an IT investment to let more technology uncertainties be resolved. When the switching cost is low, high IT uncertainties cannot be used to justify the wait-and-see policy. On the contrary, we should use real options analysis to quantify the value of the option to switch that usually makes an investment opportunity more appealing to the management.

To summarize our discussion, let us look at the four categories of IT investment opportunities based on the two criteria.

- Category I: Shared investment opportunity with high IT switching cost. For this type of IT investment opportunity, we must consider both the strategic benefit of early preemptive investment and the valuable option to wait. Potential competitive pressure forces investors to be proactive. However, preemptive investment will incur the loss of the valuable option to wait. So for this type of IT investment opportunity, the key in the real options analysis is to consider the strategic growth option and the option to wait at the same time. By balancing the two contradictory effects, we can find the optimal investment point at which the expected investment payoff will be maximized.
- Category II: Shared investment opportunity with low IT switching cost. For this type of IT investment opportunity, early preemptive investment is usually the best strategy. As we discussed before, it is beneficial to invest early to preempt potential competitors. Moreover, IT uncertainties will not make the wait-and-see strategy more appealing because the IT switching cost is low. Therefore, real options models should be used to quantify the values of the growth option and the switching option embedded in the IT investment opportunity.
- Category III: Proprietary investment opportunity with low IT switching cost. It is worth noting that the option to wait is a very valuable component of a proprietary investment opportunity. However, technology uncertainty will not contribute a lot to the value of the option to wait because the IT switching cost is low for investment opportunities in this category. So in the real options analysis we should pay attention to other business uncertainties that may increase the value of the option to wait.
- Category IV: Proprietary investment opportunity with high IT switching cost. Wait-and-see is the dominant strategy for this type of IT investment opportunity. So real options analysis should concentrate on the option to defer an investment. With the presence of technology lock-in, decision makers should be more patient before they commit a proprietary investment.

In the real business world, an IT investment opportunity may dynamically evolve from one category to other ones. So decision makers should be very cautious when they conduct real options analysis. In the next section, we use a real-world case to show the importance of adopting appropriate real options models as the IT investment opportunity evolves.

CONCLUSION

Although some recent studies recognized the potential of real options theory in evaluating strategic IT investment opportunities, we believe that the applicability of various real options models should be scrutinized under different scenarios. Standard real options models assuming symmetric uncertainty in future investment payoffs cannot be directly applied to the shared opportunities because of the competitive erosion. With the presence of potential competitive entry, real options analysis should balance the strategic benefit of preemptive investment and the value of the option to wait. IT switching cost is another important factor we must consider when we conduct real option analysis. As high IT switching cost or technology lock-in is very common in the digital economy, decision-makers should pay more attention to the technology uncertainties before committing early investment to preempt their competitors.

REFERENCES

Amram, M., & Kulatilaka, N. (1999). *Real options, managing strategic investment in an uncertain world.* Boston: Harvard Business School Press.

Benaroth, M., & Kauffman, R.J. (1999). A case for using real options pricing analysis to evaluate information technology project investments. *Information Systems Research, 10*(1), 70-88.

Clemons, E.K. (1991). Evaluating strategic investments in information systems. *Communications of the ACM*, *34*(1), 22-36.

Cox, J., Ingersoll, J., & Ross, S. (1985). An intertemporal general equilibrium model of asset prices. *Econometrica, 53*, 363-84.

Dixit, A., & Pindyck, R. (1994). *Investment under uncertainty.* Princeton University Press.

Dos Santos, B.L. (1991). Justifying investment in new information technologies. *Journal of Management Information Systems, 7*(4), 71-89.

Kumar, R. (1996). A note on project risk and option values of investments in information technologies. *Journal of Management Information Systems, 13*(1), 187-93.

Li, X. (2001). *Optimal timing for brokerage to go wireless-a real options approach.* Unpublished PhD dissertation. The University of Mississippi.

Luehrman, T. (1998a, July-August). Investment opportunities as real options: Getting started with the numbers. *Harvard Business Review,* 51-64.

Luehrman, T. (1998b, September/October). Strategy as a portfolio of real options. *Harvard Business Review,* 89-99.

McDonald, R., & Siegel, D. (1984). Option pricing when the underlying asset earns a below-equilibrium rate of return: A note. *Journal of Finance, 39*(1), 261-265.

Shapiro, C., & Varian, H. (1998). *Information rules: A strategic guide to network economy.* Harvard Business School Press.

Smith, J., & McCardle, K. (1998). Valuing oil properties: Integrating option pricing and decision analysis approach. *Operations Research, 46*(2), 198-218.

Smith, J., &. McCardle, K. (1999). Options in the real world: Some lessons learned in evaluating oil and gas investments. *Operations Research, 47*(1), 1-15.

Taudes, A., Feurstein, M., & Mild, A. (2000). Options analysis of software platform decisions: A case study. *MIS Quarterly, 24*(2), 227-43.

Trigeorgis, L. (1991). Anticipated competitive entry and early preemptive investment in deferrable projects. *Journal of Economics and Business, 43*(2), 143-145.

KEY TERMS

Black-Scholes Option Pricing Model: A model that is used to calculate the value of an option by taking into account the stock price, strike price and expiration date, the risk-free return, and the standard deviation of the stock's return.

Deferred Option: Option to defer a project or an investment gives a firm an opportunity to make an investment at a later point in time.

Managerial Investment Flexibility: Flexibility in the timing and the scale of an investment provided by a real investment option.

Net Present Value (NPV): The present value of an investment's future net cash flows minus the initial investment.

Option: The right, but not the obligation, to buy or sell an asset by a pre-specified price on before a specified date.

Real Options Theory: Financial valuation tool that helps in calculating the value of managerial flexibility under uncertainties.

Switching Costs: Switching costs refer to the hidden costs consumers face when switching from one product or technology to another in the marketplace.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2397-2402, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Real Time Interface for Fluidized Bed Reactor Simulator

Luis Alfredo Harriss Maranesi
University of Campinas, Brazil

Katia Tannous
University of Campinas, Brazil

INTRODUCTION

Nowadays, the world witnesses a large technological revolution which has brought new information distribution forms, interpretation and storage. With that, computational tools can be used to sustain the education, as with the learning objects case. A learning object is any digital product that could be re-used for knowledge acquisition, with significant economy and reduction of computer time.

Learning objects have led to new solutions, which resulted in good structured and safe programs. Hereby, they rend possible creations of simple units, and the objects, which are associated with each other, can produce large units. Some of them are distinguished among the presence or absence of simulation functions.

The software SEREA has been developed to reach undergraduate and graduate chemical engineering for studies about fluid dynamics of fluidized bed reactors motivating students in order to acquire a successful learning process. Motivating students is certainly a stimulating and challenging problem, and is always present in teaching methodologies (Tannous, 2007). This article will present a comparison between two methodologies for interface creations, to sustain the chemical engineering learning and other correlated fields.

BACKGROUND

Since 2002, the Laboratory of Particle Technology and Multiphase Flow at State University of Campinas have developed new learning objects, mainly simulator modules, to evaluate their limitations as educational software.

Several denominations are found in the literature about the concepts of learning objects such as: instructional object, educational object, knowledge object, intelligent object and data object (Gibbons, Nelson, & Richards, 2000). Nevertheless, it does not matter what denomination has been improved, as the object can be practically the same.

The IEEE Learning Technology Standard Committee (2002) defines learning objects "*as any entity, digital or non-digital, which can be used, re-used, or referenced during technology supported learning.*" Chronological and instructional texts, class activities, books and revision aids are some examples of nondigital learning objects.

However, concerning digital learning objects, the main idea is to break the contents in small pieces that can be re-used in different learning environments, following the "spirits" of oriented-objects programming (Wiley, n.d., Verbert & Duval, 2004). According to Downes (2001), the idea of object-oriented tends toward the development of real pattern that, once defined, are copied and used in a part of the software. In this way, the simulators associated with the object-oriented programming can be classified in this definition.

According to Logmire (2001), for designing and developing material to be reused as learning objects, it should consist of features such as flexibility, easy to update, search and management, customization, interoperability, facilitation of competency-based learning, and increased value of content. All these characteristics show that the learning object models can make an easier and enhanced quality of learning, providing several facility tools for professors, students and administrators.

The simulation is a learning resource that allows the students to observe the different system behaviors through mathematical graphics or symbolical modeling of the phenomenon. In this context, the simulations have an important role to minimize the problems due missing equipment and laboratories for undergraduate students.

Tannous (2005) and Rimoli, Assis, and Tannous (2006) described some of the strategies and methodologies applied to develop learning objects (simulators with or without instructional program). It is important to remark that, in general a few works are applied to chemical engineering.

DEVELOPMENT OF LEARNING OBJECTS

General Information

SEREA (Fluidized bed reactors modules) is simulator software for undergraduate chemical engineering students. It was developed to simulate the fluid dynamics parameters

of different fluidized bed reactors, being divided by behavior of particles and project of distributors. As SEREA expanded, it was split in one real time process named "slipping controls."

The following sections cover two modules for basic parameters that consist of the determination of minimum fluidization velocity and porosity, and bed expansion. The fluidization engineering concepts are based upon Geldart (1986), Kunii and Levenspiel (1991) and Martin (1998), and are also covered in other texts (Tannous, 1993).

Required Hardware

For an educational tool to be effective, it must be readily accessible to students. SEREA has been developed on personal computers sufficiently supplied with the necessary amount of memory and processing of operational systems. The exact hardware chosen was processor Intel Pentium 4 1.6 GHz, 768 MB of RAM memory, CD recorder, Monitor of 15," video board Nvidia Riva TNT 2, mouse and keyboard.

Software and Operating System

Nowadays, several technologies can be used to build the learning objects, including Java applets, flash, Modellus, Javascript and those more powerful and innovative resources as Java and C++ languages. Also, we found the most popular object-oriented programming languages to be Visual basic and Delphi. Each language has their own advantages and disadvantages depending upon the developers requirements. For our case a feasibility analysis was constructed identifying our problem and indicating available resources within the department to assist with our costs/benefits analysis. Justifiably, Borland Delphi 2005 IDE was chosen to develop the SEREA simulator. It has been successfully run in Windows XP.

The integration of this programming language with **graphics** resources causes easier and faster creation of software, with better results. Some of these characteristics are: elements repository, automatic inclusion of declarations of variables and classes upon the insertion of an element using the graphic mode, list of properties that allow modification of the parameters for each object easily and quickly, compilation and execution of codes with automatic syntax and verification of semantic error and integration with the Paradox 7.0 database, that allows for easy interaction between different modules of the software (Rimoli et al., 2006).

Other software employed in the development of SEREA was the open source application package OpenOffice.org which includes text, presentations, diagrams, mathematical equations and spreadsheet editors. The equation editor (OpenOffice.org Math) creates complex mathematical equations with only one line of code and OpenOffice.org Writer elaborates explanation texts for each field in the software.

Methodology Applied

The methodology adopted for creating the software modules followed the software engineering steps (Schaerer & Schauer, 1991; Rimoli et al., 2006): Analysis, Design, Coding, Testing, Production and Maintenance.

Analysis: it is the process of defining the problem. The requests are evaluated, selecting which factors are limited for the development and viability of the software. The factors are: cost, time, in and out data format, and concepts of the problem. If the software is practicable, we have a product of this process and the project specifications will be made for the next step.

Design: it starts with a known basic model to determinate an outline without details of product and a plan for the development of the software, making appropriate choices concerning the graph design, mathematical methods, languages and tools (e.g., IDE). The decision making in this process defines the final characteristics of the software, and can guarantee the success or failure of the whole project. However, it may be necessary after advancing to the next phase to return to this step within the software development cycle, in order to succeed in codification and to solve possible problems identified by the developer.

Implementation and final product: codification, function tests and others structures created, satisfying all requirements demanded in previous steps without execution errors, composes the program process. As a final product, the software is completed once it meets the initial requirements and solves the question in a satisfactory way. The project is then ready to be integrated into the rest of the system. In due course the final product will require continual maintenance.

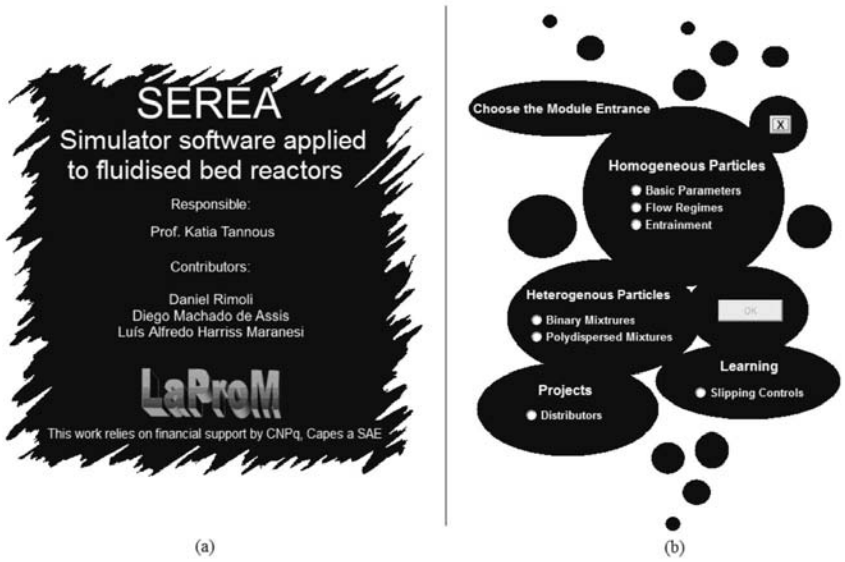
User Interface

Users interface with SEREA using a mouse to activate objects and request information about parameters and identify the equations. Each result can be saved for the next module, keeping all data introduced or only one individual module in different order.

Getting Started

The simulator modules were developed to be objective and intuitive aiding the students to achieve better results in a complex mathematical analysis. The program SEREA is composed of five modules (traditional) with the same structure but with differences in the entrance variables, correlations and results: the module of basic parameters of fluidized beds (minimum fluidization velocity and porosity, bed expansion), the module of fluidisation regimes (transition velocities of different regimes), and entrainment of particles. All of them were applied for the hydrodynamic of homogeneous particles. For heterogeneous particles, until now, we developed

Figure 1. Software simulator: (a) initial screen (b) menu



minor modules considering the mixtures and segregation parameters and basic parameters. Distributors are essential for sustaining the particles on the bed. So, we propose two distributor designs: perforated plate and tuyere.

A special module was developed defined as "real time simulator-slipping controls" to compare with the first module of homogeneous particles and understand how the knowledge can reach the students.

Figure 1 shows the open screen of the software SEREA marking the (a) project collaborators and sponsorships and (b) the simulator module menu. The latter shows modules for fluid dynamics behaviour of fluidized bed reactors: three modules for homogeneous particles, two modules for heterogeneous particles and one module for distributors. All of them present the same structure but with differences in the entrance variables, correlations and results.

Besides that, we have also developed one alternative module, called "learning." In this article, we will compare the latter and the first module, concerning the basic parameters for fluidized bed rectors, to calculate the minimum fluidization velocity and bed expansion.

The comparison is related concerning the purpose of interface using practically the same programming developed for the simulator.

RESULTS AND DISCUSSION

The learning objects developed until now are disposable in the Laboratory of Particle Technology and multiphase Flow and in the intranet of the School of Chemical Engineering at State University of Campinas. For each module concluded, the learning objects are applied in different courses as Operation Units I (practical and theoretical) and Fluidization Technology. In this article, we will present a comparison between two different styles of interfaces to reach the students during the chemical engineering courses.

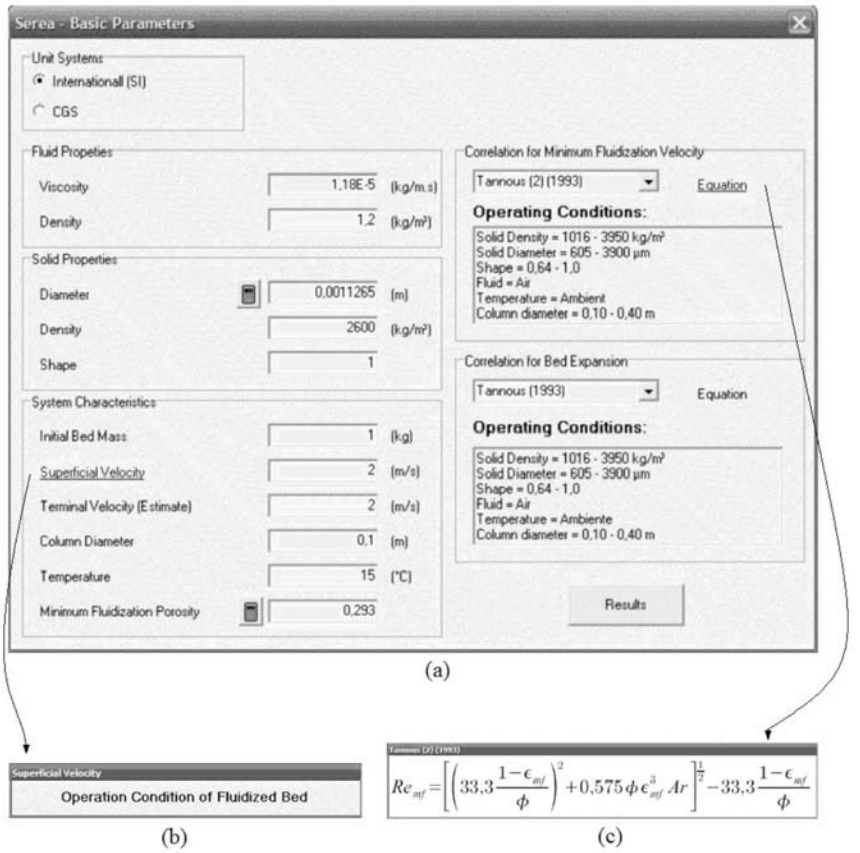
Traditional Module

The first module of the SEREA software (Figure 2a) intends to give users the immediate results of these parameters, allowing to diagnose and to change the variables of process interferences (Assis & Tannous, 2006; Rimoli et al., 2006). For example, in the laboratory experiments, it will allow the user to compare the experimental and theoretical results in the same time and to verify the mistakes during the runs. The users have total control regarding the data and no limitation for applying the models. They are responsible for right or wrong results.

The interface proposed concerning the seven focal areas: four areas in the right side and two in the left side. In the right side, it presents the unit systems (top), fluid and solid properties (middle) and characteristic system (bottom). In the right side, it shows the correlations about minimum fluidization velocity and bed expansion, and at the bottom, the results button.

For each variable present in the screen (Figure 2a), there is a space besides its name where the value can be inserted manually in the International System (IS) or Centimeter-

Figure 2. Screens of basic parameters module of homogeneous particles: (a) simulator interface (b) variable definition (c) minimum reynolds number equation (Tannous, 1993)



gram-second system (CGS) of Units. Passing the mouse under the name of solid and fluid physical properties, and the characteristics of the system, a new window with a small explicative text will appear, as shown in Figure 2b (e.g., superficial velocity).

The variables, mean particle diameter and minimum fluidisation porosity, with a calculator icon besides, can be calculated through an auxiliary window choosing correlations and introducing experimental data (Rimoli et al., 2006).

The minimum fluidisation velocity is obtained from the solid and fluid properties, but in some cases it needs specific variables such as minimum fluidisation porosity and particle sphericity. There is a correlations list in a selection box. Each correlation has an operational range that can be seen in the box below the drop-down menu. By moving the mouse over the word "equation" the formulas can be seen in Figure 2c.

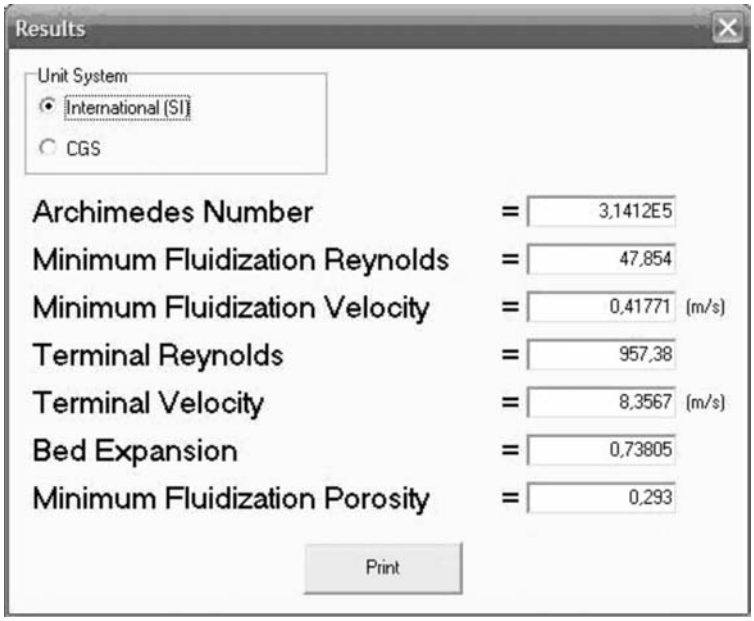
To calculate bed expansion, the user can again choose between several correlations. There is no standard form in these correlations, so one or more extra functions are required in each case to get the results.

If the users choose incompatible system units between the correlations, an error message appears and allows amending of the problem before obtainment of wrong results.

When the button "results" is put into action, the program verifies if the entrance values comply with the restrictions, producing an error message when necessary. Otherwise, the values are copied as internal variables that follow a sequence to calculate the results exhibited on the screen shown in Figure 3: the Archimedes number, the Reynolds number and velocity at minimum fluidisation, the Reynolds number and terminal velocity, bed expansion, and minimum fluidization porosity.

All results are saved in a database to be accessed by other program modules or just finally results. The screen shows a printer button that allows one to print a page containing a table with all the entrance data and graphs produced in the calculation of the mean diameter (Rimoli et al., 2006).

Figure 3. Final results screen of basic parameters



After that, the user is responsible for the interpretation of results.

Real Time – Slipping Controls

The development of a new module, on real time, comparative of traditional software emerged in relation to the difficulties of students in understanding the magnitudes in the elaboration and execution of projects in fluidized beds. The use of the calculator in the engineering courses does not contribute to comprehension of phenomena as well as the knowledge of empirical or semi-empirical models.

In general, the students rarely find additional information about the physical meaning of results leading to excessive mistakes. Then, to solve this problem, we have adopted the use of slipping controls simultaneously with results presentation (Figure 4). The interface proposed looks like a control panel used in chemical industries located in special rooms, controlling the operational conditions of equipments connected in the processes.

The interface was built considering the same structure of focal areas cited in the traditional module. Six focal areas are considered: the correlations about minimum fluidization velocity and bed expansion (top), fluid and solid properties and characteristic system (middle), the results and storage (bottom).

The help functionalities were maintained once the mouse was passed upon the text of the exhibited equations and on the variables (entrance fields).

The main difference between the two modules is the method in which data entry is made into the form and the presentation of the result. The entrance of data is filled by adjustment of slipping controls or introducing the specific place of a variable. We remark that each correlation has it own range conditions.

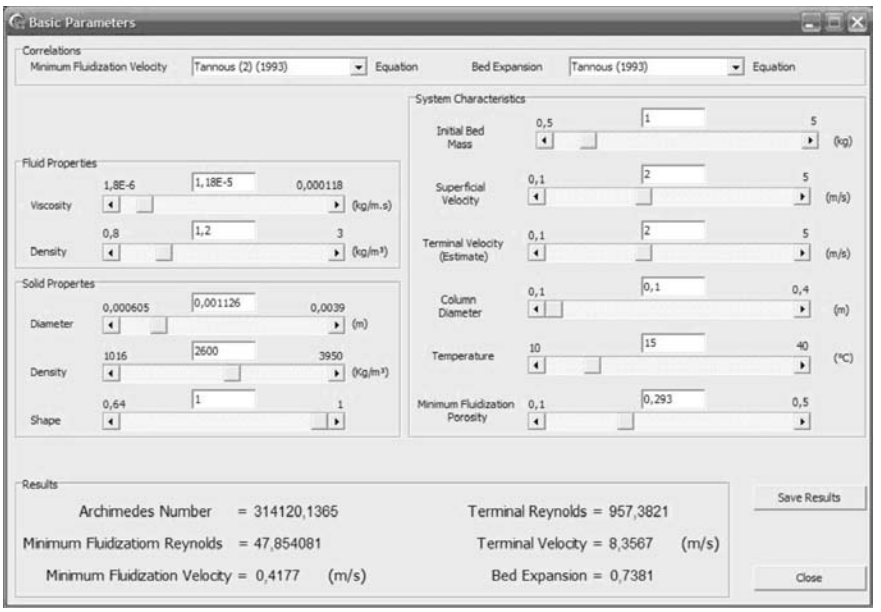
This solution allows the user to examine with details the influence of each data entry upon the final results, obtaining more sensibility about the phenomena inside of reactors.

In the case of disagreement between parameter ranges, for example, the density and the column diameter ranges, a new screen shows the incompatibly of that for users. The methodology applied guarantees that the user will acquire right decisions at the end of simulations.

The option "save results," located in the left bottom of the window (Figure 4), offer the user the possibility of saving the results obtained during the simulations. For each press of a button, the program stores in its memory the values tested and finally a spreadsheet file (MS-Excel) organizes these values, as shown in Figures 5a and 5b. Saving the results make possible to analyze all simulations together and make a correct opinion about the process and data chosen.

In this way, the software can apply in different courses to understand the fluidized bed processes and assist the students for elaboration of general reports.

Figure 4. Module interface of basic parameters on real time



FUTURE TRENDS

For future works, the software should keep the objects created and follow the same characteristics of interfaces known, constructing easily the development of new modules. Moreover, the software has been prepared for effective application of deterministic modeling representatives of solid-fluid systems. The integration between Object Pascal programming language and other languages, for example, C++ language, can create new auxiliary components for simulations.

CONCLUSION

A learning object can be present in different features stimulating students through texts, videos, tutorials, graphs, simulators and educative programs where they are distributed electronically and applied for the learning process. In this way, the information distribution makes it easily accessible, far reaching and economically more feasible for a higher number of people. The objects are found in the literature distinguished by presence or absence of simulation functionalities, and may perhaps be used individually or complementary of conventional learning.

The learning object, SEREA, allows easy simulation of the main parameters of different processes applied to fluidized bed reactors. In general, the software demonstrates that the decisions adopted and the implementation process was appropriate. The whole decision was in agreement with the literature and also provided adequate results applied in chemical engineering courses.

The module "Learning" with sliding controls allows the users to manipulate the data immediately after operation, gaining the sensibility of measurements and results involved in the fluidized bed processes.

ACKNOWLEDGMENT

The authors gratefully acknowledge the contribution of many students who were involved in this project. The authors also wish to thank CAPES and SAE/UNICAMP for their sponsorship.

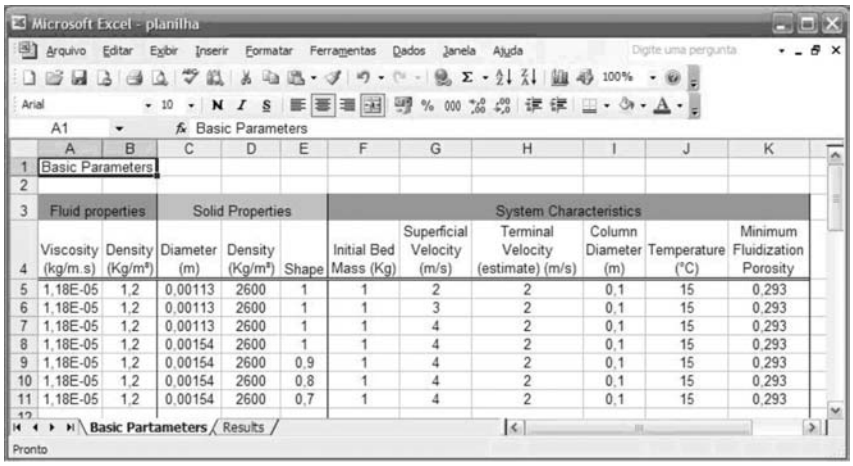
REFERENCES

Assis, D.M., & Tannous, K. (2006, July 24-27). Development of simulator software in fluidized bed. In *Proceedings of the VI Brazilian Congress in Chemical Engineering*, Campinas, Brazil.

Downes, S. (2001). Learning objects: Resources for distance education worldwide. *International Review of Research in Open and Distance Learning, 2*(1). ISSN: 1492-3831. Retrieved May 28, 2008, from http://www.irrodl.org/index.php/irrodl/article/view/32/81

Geldart, D. (1986). *Gas fluidization technology*. New York:

Figure 5a. Spreadsheet of entrance data



Microsoft Excel - planilha

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

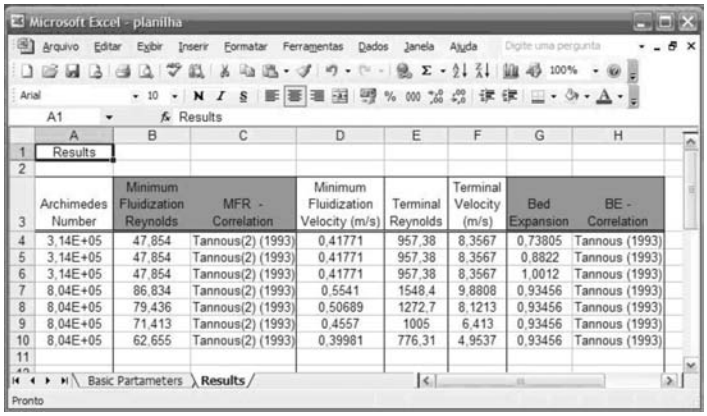
A1 Basic Parameters

	A	B	C	D	E	F	G	H	I	J	K
1	Basic Parameters										
2											
3	Fluid properties		Solid Properties			System Characteristics					
4	Viscosity (kg/m.s)	Density (Kg/m³)	Diameter (m)	Density (Kg/m³)	Shape	Initial Bed Mass (Kg)	Superficial Velocity (m/s)	Terminal Velocity (estimate) (m/s)	Column Diameter (m)	Temperature (°C)	Minimum Fluidization Porosity
5	1,18E-05	1,2	0,00113	2600	1	1	2	2	0,1	15	0,293
6	1,18E-05	1,2	0,00113	2600	1	1	3	2	0,1	15	0,293
7	1,18E-05	1,2	0,00113	2600	1	1	4	2	0,1	15	0,293
8	1,18E-05	1,2	0,00154	2600	1	1	4	2	0,1	15	0,293
9	1,18E-05	1,2	0,00154	2600	0,9	1	4	2	0,1	15	0,293
10	1,18E-05	1,2	0,00154	2600	0,8	1	4	2	0,1	15	0,293
11	1,18E-05	1,2	0,00154	2600	0,7	1	4	2	0,1	15	0,293

Basic Partameters / Results

Pronto

Figure 5b. Spreadsheet of results



Microsoft Excel - planilha

Arquivo Editar Exibir Inserir Formatar Ferramentas Dados Janela Ajuda

A1 Results

	A	B	C	D	E	F	G	H
1	Results							
2								
3	Archimedes Number	Minimum Fluidization Reynolds	MFR - Correlation	Minimum Fluidization Velocity (m/s)	Terminal Reynolds	Terminal Velocity (m/s)	Bed Expansion	BE - Correlation
4	3,14E+05	47,854	Tannous(2) (1993)	0,41771	957,38	8,3567	0,73805	Tannous (1993)
5	3,14E+05	47,854	Tannous(2) (1993)	0,41771	957,38	8,3567	0,8822	Tannous (1993)
6	3,14E+05	47,854	Tannous(2) (1993)	0,41771	957,38	8,3567	1,0012	Tannous (1993)
7	8,04E+05	86,834	Tannous(2) (1993)	0,5541	1548,4	9,8808	0,93456	Tannous (1993)
8	8,04E+05	79,436	Tannous(2) (1993)	0,50689	1272,7	8,1213	0,93456	Tannous (1993)
9	8,04E+05	71,413	Tannous(2) (1993)	0,4557	1005	6,413	0,93456	Tannous (1993)
10	8,04E+05	62,655	Tannous(2) (1993)	0,39981	776,31	4,9537	0,93456	Tannous (1993)
11								

Basic Partameters / Results

Pronto

John Wiley & Sons.

Gibbons, A.S., Nelson, J., & Richards, R. (2000). *The nature and origin of instructional objects*. Retrieved May 28, 2008, from http://www.reusability.org/read/chapters/gibbons.doc

IEEE Learning Technology Standard Committee P1484.12. (2002). *Learning object metadata*. Retrieved May 28, 2008, from http://ltsc.ieee.org/wg12

Kunii, D., & Levenspiel, O. (1991). *Fluidization engineering*. New York: John Wiley & Sons.

Longmire, W. (2001). *A primer on learning objects*. Retrieved May 28, 2008, from http://www.learningcircuits.org/2000/mar2000/Longmire.htm

Martin, R. (1998). *Introduction to particle technology*. Chichester, England: John Wiley & Sons.

Rimoli, D., Assis, D.M., & Tannous, K. (2006, August 27-31). Simulator software applied to fluidized bed reactors – SEREA. In *Proceedings of the 5th International Conference for Conveying and Handling of Particulate Solids*, Sorrento, Italy.

Schaerer, D.E., & Schauer, H. (1991). List and graph algorithms in object pascal. *Journal of Microcomputer Applications, 14*, 229-261.

Tannous, K. (1993). *Contribution a l´etude hydrodynamique des lits fluidisés de grosses particules*. Doctoral thesis in Process Engineering, ENSIGC – IPT, France.

Tannous, K. (2005). Interactive learning in engineering education. In S. Mishra & R. C. Sharma (Eds.), *Interactive multimedia in education and training* (pp. 289-305). Hershey, PA: Idea Group.

Tannous, K. (2007). Project-based learning in chemical engineering education using distance education tools. In S. Mishra & R. C. Sharma (Eds.), *Cases on global e-learning practices successes and pitfalls* (pp. 202-217) Hershey, PA: Idea Group.

Verbert, K., & Duval, E. (2004). *Toward a global component architecture for learning objects: A comparative analysis of learning object content models*. Retrieved May 28, 2008, from http://www.cs.kuleuven.ac.be/~hmdb/publications/files/pdfversion/41315.pdf

Wiley, D.A. (n.d.). *Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy*. Retrieved May 28, 2008, from http://reusability.org/read/ chapters/wiley.doc

KEY TERMS

Chemical Engineering: The branch of engineering that deals with the application of physical science (e.g., chemistry and physics), with mathematics, to the process of converting raw materials or chemicals into more useful or valuable forms.

Fluidization: The operation by which the fine and large solids are transformed into a fluid-like state through contact with a gas or liquid.

Fluidized bed reactor: Consists of a particle-laden vessel, with the particle bed kept fluidized by feeding a gas flow through it. It is widely used in the chemical, petrochemical, and pharmaceutical industries, and also for desulphurization of flue gases, calcinations, gasification and combustion of waste/biomass.

Interface: Defines the communication boundary between two entities, such as a piece of software, a hardware device, or a user. The interface between a human and a computer is called a user interface.

Learning Object: Digital product that could be re-used for the knowledge acquisition

Object-Oriented Programming: A programming paradigm that uses "objects" and their interaction to design applications and computer programs.

Object Pascal: An object-oriented derivative of Pascal mostly known as the primary programming language of Borland Delphi.

Database: A structured collection of data that can be used to answer queries and retrieve previously stored data.

Simulator: The mathematical representation of the interaction of real-world objects.

Slipping Controls: Artifacts used in software interface to implement controls that look like real life machine controls.

Really Simple Syndication (RSS)

Kevin Curran
University of Ulster, UK

Sheila McCarthy
University of Ulster, UK

INTRODUCTION

E-mail has been one of the major reasons for the broad acceptance of the Internet, and although e-mail is still a vitally important communication tool, it suffers from an increasing number of problems as a medium for delivering information to the correct audience in a timely manner. The increasing volume of spam and viruses means that e-mail users are forced into adopting new tools, such as spam-blocking and e-mail-filtering software, that attempt to prevent the tirade of unwanted e-mails. Many users are also becoming increasingly reticent to divulge their e-mail address for fear of an impending spam influx. Further to this, recent studies suggest that up to 38% of bona fide e-mail messages are being erroneously blocked by filtering software. In reality, this means that more than a third of e-mails, newsletters, special offers, and event announcements are not reaching their intended audience (Patch & McKinlay-Key, 2004). Therefore, the combination of e-mail issues, such as the increasing difficulties associated with multimedia downloads, such as delays, compression, and data integrity maintenance, could be seen as creating a demand for an alternate, effective, and secure communication methodology. One such alternative technology is Really Simple Syndication (RSS), previously known as Rich Site Summary. RSS allows some elements of Web sites, such as headlines, to be transmitted in unembellished form. When devoid of all elaborate graphics and layouts, such minimalist headlines are quite easily incorporated into other Web sites. In other words, third-party Web sites can insert this content on their site through embedded RSS news readers and thus, provide active news feeds quite easily to their clientele. RSS, termed a lightweight content syndication technology, offers many advantages over streaming and e-mail, and for the consumer, no more difficult to access as the RSS readers are akin to e-mail clients (Byrne, 2003). There is no question that the media is keen to adopt a new communications option, and RSS most certainly can comply.

RSS solves a myriad of problems Web masters commonly face, such as increasing traffic, and gathering and distributing news (BBC, 2008). RSS can also be the basis for additional content distribution services (Kerner, 2004). The real benefit of RSS, apart from the added benefit of receiving news feeds from multiple sites, simultaneously, in the viewer, is that all the news feeds (i.e., news items) are chosen by the user. With thousands of sites now RSS-enabled and more on the way, RSS has become perhaps one of the most visible Xtensible Mark-up Language (XML) success stories to date. RSS formats are specified using XML, a generic specification for the creation of data formats. Although RSS formats have evolved since March 1999, the RSS icon ("") first gained widespread use in 2005/2006. RSS democratizes news distribution by making everyone a potential news provider. It leverages the Web's most valuable asset, content, and makes displaying high-quality relevant news on a site relatively easy (King, 2004). It must be recognized, however, that RSS cannot entirely replace the primary function of e-mail, which is to provide person-to-person asynchronous communications, but it does compliment it in some interesting ways.

BACKGROUND

RSS can be found as an acronym for, *Rich Site Summary*, *Resource Description Framework (RDF) Site Summary,* or indeed *Really Simple Syndication*, the latter is used here (Oasis-open, 2004). The RSS format was created to facilitate "channels" on Netscape Netcenter (Netscape, 2005), and was made available to the general public in March of 1999. Channels were a "pull" type mechanism where users requested certain information from various channels. The original RSS, version 9.0, was created by Netscape as a method of building portals to major news sites for news headlines. Portals are Web sites dedicated to specific topics. It was, however, soon replaced by the 0.91 version that stripped out many of the less important features, as Netscape believed 0.90 proved simply too intricate for this undemanding task. The newly established 0.91 itself was promptly dropped by Netscape as their interest in the portal-making business declined. The now obsolete 0.91 was swiftly adopted by the competition, UserLand Software, and employed as the foundation for all its Web-based concepts. Shortly after this, RSS version 1.0, a new version based on Resource Description Framework (RDF), was developed by a third-party spin-off, a group of designers who built their version modeled closely on the concepts and framework of the initial, original 9.0 (prior to its

simplification into version 0.91). The Resource Description Framework (RDF) integrates a variety of applications from library catalogs and worldwide directories to syndication and aggregation of news, software, and content to personal collections of music, photos, and events, using XML as an interchange syntax. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web (Nilsson, 2001; Van der Vlist, 2001). As a result of this, Userland, indignant at being omitted from the latest increment, ignored version 1.0 and continued to advance their own brand of RSS, developing versions 0.92, 0.93, 0.94 through to their current 2.0. In reality, this means there are seven different formats to contend with. A feed aggregator, also known as a feed reader, news reader, or simply as an aggregator, is client software or a Web application that aggregates syndicated Web content, such as news headlines, blogs, podcasts, and vlogs in a single location for easy viewing (Wikipedia, 2008). Aggregators must be flexible and comprehensive, and must be able to recognise and deal with all versions. RSS version 2.0 is currently offered by the Berkman Centre for Internet and Society, at Harvard Law School.

REALLY SIMPLE SYNDICATION (RSS)

RSS has rapidly developed into a prevalent means of sharing content between Web sites. Many sites already use RSS, and as word spreads, new sites incorporate this feature into their sites daily. RSS looks set to become a dominant force. Numerous news sites, including BBC, Yahoo! and Wired, currently use RSS to provide their subscribers with the latest headlines. Indeed the Web sites of many mainstream "giants" also incorporate RSS in a bid to keep their subscribers notified of announcements, events, and advertisements. As yet, only sites that currently offer news in RSS format may be read using a news aggregator. To ascertain if a site utilizes RSS is generally simple. Sites make no secret of the fact and proudly display RSS feed pictograms, such as ("  ") throughout their pages, indicating which sections are available in RSS format. Right clicking on such an icon, copying the shortcut (URL), and adding it to an aggregator, creates a feed. This establishes a subscription to that particular Web site for the desired information. Channels to numerous sites can be created, maintained, and removed, if desired, using most aggregators with minimal effort.

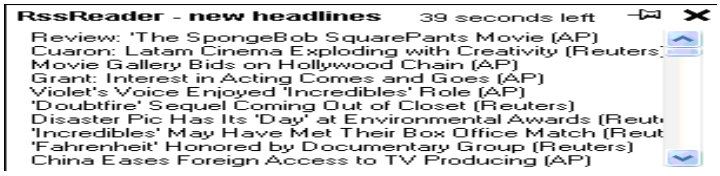
An RSS text file contains both static and dynamic information. At a high level, an RSS document is an rss element, with an obligatory attribute called version, this attribute specifies the version of RSS that the document conforms to. Here an element is a piece of data within a document that may contain either text or other subelements describing the RSS data. Succeeding the rss element is a single channel element that contains information about the channel (metadata) and its contents. Metadata is commonly defined as "data about data" or data describing context, content, and structure of records and their management through time. A channel may, in turn, contain any number of items. Items are subelements that are enclosed in matching XML start and end tags, and appear as subelements of channel, listed before the closing/channel tag. Each item is identified with an opening item tag, and concluded with a closing/item tag. All child elements of an item are optional, however, at least one element must be present, either title or description. An item may be a snippet of information that represents a larger article, much in the same way as a headline represents a newspaper article. If this is the case, the item's description is a synopsis of the story, and the link points to the full story. An item may also be complete in itself; if so, the description contains the full text, and the link and title may be omitted. In this way, an RSS channel can contain many items that, in turn, may incorporate many differing subelements. When design and coding is complete, the validated RSS file can be registered with various aggregators, allowing the feed to be "sucked up" by discerning subscribers. Any amendments or updates made to the RSS file will automatically be relayed to all subscribing clients.

RSS Enclosures

RSS version 2.0 encompasses a powerful feature; it allows an item to have an enclosure. This can, in simplistic terms, be likened to an e-mail having an attachment. In reality, enclosures hold huge potential, and represent another step in the evolution of content syndication (Kerner, 2004). By incorporating an enclosure subelement into an item, any RSS element can then describe a video or audio file. The enclosure feature has three attributes, the first, "url," says where the multimedia file is located, the second "length" determines the size of the file in bytes, and the last "type" describes the Multipurpose Internet Mail Extension (MIME) type of the multimedia file. In this way, an aggregator can determine the payload attributes prior to any communication, and can then apply the appropriate scheduling and filtering rules.

Primarily, the most attractive feature of RSS is that it enables information from numerous Web sites to be viewed simultaneously, all on one page; consequently, numerous sites can be scrutinized in seconds rather than having to be tediously downloaded independently. A free newsreader is *RssReader*. Like other aggregators, the RssReader aggregator can sustain numerous channels, scouring each of the user's designated Web sites for updated feeds at regular intervals. When RssReader gathers updated headlines from the various sites, it displays an amalgamation of such in a list box positioned in the bottom right of the user's desktop (see Figure 1; which displays headlines from Yahoo's entertainment news feed).

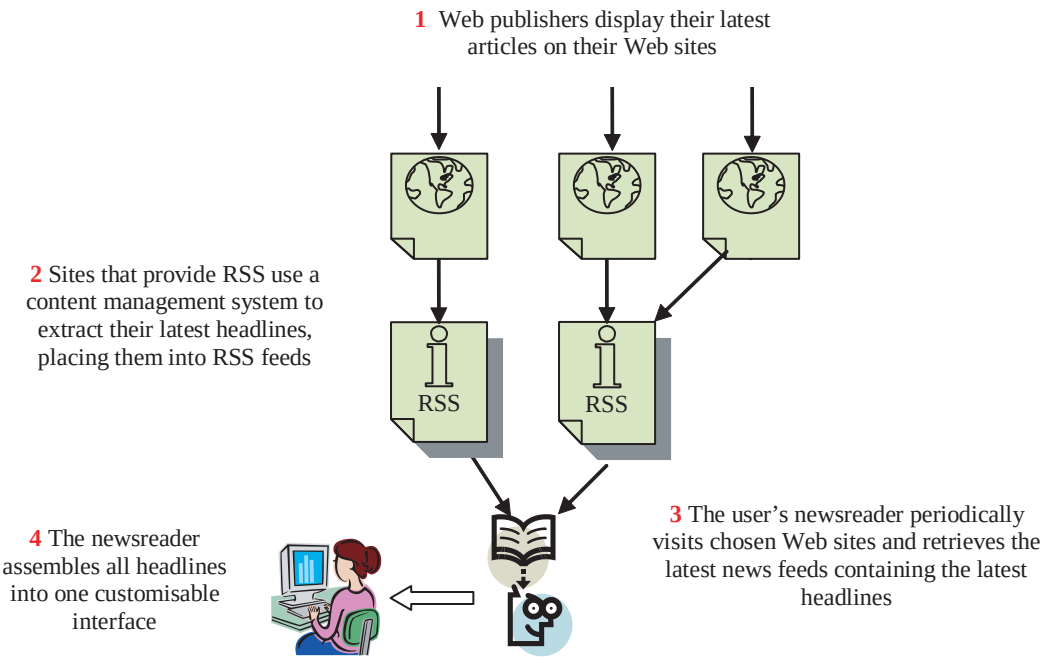
Figure 1. RssReader's "headline alert" screen



If a user wishes to select a headline from the list, aggregators will provides features to open and provide a synopsis of each news article, if the user wishes to read further on any given topic a link is provided to the specific article on each of the initiating Web sites. This way, Web publishers are able to channel tremendous traffic towards their sites. No longer having to wait for passing traffic, RSS bestows a means of advertising ware on a much wider stage indeed. This way, by employing a news aggregator, a client can subscribe to any sites of their choosing that provide RSS feeds. The Web sites that do not offer this facility may be disadvantaged, and must wait for the client to visit their page directly, if at all, see Figure 2.

Perhaps the most compelling feature of RSS feeds is the ability to keep track of changes on the Web. It is not difficult to ascertain what Web sites are available today, what is difficult is to ascertain when such sites make crucial changes. RSS feeds provide us with the necessary aptitude to overcome such problems.

Figure 2. How RSS feeds work



Podcasting

Podcasting is a term derived from Apple's portable MP3 player,that is, the iPod. Podcasting is the preparation and distribution of predominately audio for download to digital music players, such as the iPod player. A podcast is created from a digital audio file that must be saved in an MP3 format and then uploaded to the Web site of a service provider. The MP3 file then receives its own URL, which is inserted into an RSS XML document as an enclosure within an XML item tag. Once a podcast has been created, it is usually registered with content aggregators, such as podcasting.net or ipodder.org, for inclusion in podcast directories. Interested parties can then browse through these categories, or subscribe to specific podcast RSS feeds that will, in turn, download to their audio players automatically when they next connect. Although podcasts are generally audio files created for digital music players, the same technology can be used to prepare and transmit images, text, and video to any capable device, this is the approach taken for this project. Podcasting could be described as the first application based on RSS enclosures to capture the imagination of users and developers. Podcasting allows users to listen to selected podcasts whenever they like, similar to the way time shifting allows viewers to watch television programs when it suits them. The cultural milieu supporting podcasting is sometimes referred to as the podosphere, just as the cultural environment surrounding the blog is called the blogosphere. Anyone using RSS to distribute information can potentially make use of enclosures, for example, a company currently distributing a newsletter using an RSS feed could upgrade this completely with the inclusion of a promotional video clip as an enclosure in the feed.

COMMERCIAL RSS SYSTEMS

There are an increasing number of Web sites that offer RSS format producing software. Two of the more popular are Nooked and Radio Userland.

Nooked (Nooked, 2008) is an online service that enables users to create, publish, and maintain their own RSS channels, with the minimal amount of effort and at low costs. Nooked makes RSS potentially available to users of all abilities by shielding the technical intricacies of how RSS is configured. They have a FeedWizard on their site that is Web based. It allows a feed to be created quite quickly and support podcasting and flashcasting similar to the system described here.

Radio UserLand is one of the most popular Web logging tools. Radio supports the publication of Web logs that can optionally include enclosures, and also allow customers use of its own built-in news aggregator. Users can subscribe via this aggregator to feeds, allowing Radio users to achieve more than is possible with similar services that offer only Web log hosting. Radio comprises features such as the "category" feature that supports blogging on varied topics, and an "author Web log tool" that permits multiple authors to contribute to a community Web log.

FUTURE TRENDS

It should not be forgotten that one of the main objectives of all RSS modules is to extend the basic XML schema that was established for more robust content syndication allowing for wider–ranging, yet standardized, transactions without modifying the core RSS specification. This can be achieved with an XML namespace that give names to concepts and relationships between those concepts. We can expect to see an increasing number of RSS 2.0 modules with established namespaces appearing in the near future, such as Media RSS Module - RSS 2.0 Module (MRSS, 2007).

Really Simple Syndication is central to some of the leading companies such as Google, Yahoo, and Microsoft (Rubel, 2006). Yahoo has stated in the past that almost 3 out of 10 Internet users consume RSS-syndicated content on personalized start pages without knowing that RSS is the enabling technology (Hrastnik, 2005). We expect that feed reading will become even easier than it is now, and should be incorporated into all kinds of connected devices especially mobile phones and home media units.

We can expect to see it become more embedded into bit-torrent-based peer-to-peer applications where even now, feeds (also known as *Torrent/RSS-es* or *Torrentcasts*) allow client applications to download files automatically from the moment the RSS reader detects them. The term for this currently is Broadcatching.

News site, such as the Guardian (http://www.guardian.co.uk), will continue to increasingly provide RSS feeds for a selection of news services and sites, automatically updating as stories are added across the network, flagging up what's new as it breaks (Guardian, 2008) .

We also expect to see diverse uses of RSS, such as the encoding of location in RSS, as performed by georss.org. Here, location is described in an interoperable manner so that applications can **request, aggregate, share,** and **map** geographically tagged feeds. This site was created to promote a relatively small number of encodings that meet the needs of a wide range of communities in the hope that building these encodings on a common information model would result in an "upwards-compatibility" across encodings.

CONCLUSIONS

E-mail suffers from an increasing number of problems as a medium for delivering information to the correct audience in a timely manner due to the volume of spam and viruses arriving in our in-boxes on an hourly basis. RSS is a way of receiving constantly updated links to selected Web sites. Once a connection is setup to a Web site, then a list of all the stories currently shown on a certain page or section of that site can be retrieved. There are several ways of receiving RSS feeds, but a common method is to download a program called a "News Reader," which can then be setup to receive RSS information from Web sites offering it, and browse headlines and story summaries that link through to the full story on the Web site. Alternatively, newer Web browsers offer similar functionality already built-in that will detect whether the Web site one is currently browsing offers an RSS feed, and will then let you create a constantly-updated list of links in the "bookmarks" menu. Perhaps the most compelling feature of RSS feeds is the ability to keep track of changes on the Web, as it can be difficult to ascertain when sites make crucial changes. RSS feeds, however, provide us with the necessary aptitude to overcome such problems.

REFERENCES

BBC. (2008). *News feeds from the BBC*. Retrieved from http://news.bbc.co.uk/1/hi/help/3223484.stm

Byrne, T. (2003). Content syndication: Ready for the masses? *All Business Magazine*. Retrieved from http://www.allbusiness.com/information/internet-publishing-broadcasting/955591-1.html

Guardian. (2008). Retrieved from http://www.guardian.co.uk/webfeeds

Hrastnik, R. (2005). *Analyzing the New Yahoo RSS - Whitepaper for Marketers*, *RSS Statistics*, October 10, 2005. Retrieved from http://rssdiary.marketingstudies.net/content/analyzing_the_new_yahoo_rss_whitepaper_for_marketers.php

Kerner, S. (2004). The RSS Enclosure Exposure – It's really simple stuff: audio feeds and the rise of RSS. *Internet News – Realtime IT*. Retrieved from http://internetnews.com/xSP/article.php/3431901

King, A. B. (2004). *Webref and the future of RSS, Introduction to RSS , WebRef and RSS.* Retrieved from http://www.webreference.com/authoring/languages/xml/rss/intro/3.html

MRSS. (2007). *Media RSS Module - RSS 2.0 Module*. Retrieved from http://search.yahoo.com/mrss

Netscape. (2005). Retrieved from http://my.netscape.com

Nilsson, M. (2001). *The semantic Web: How RDF will change learning technology standards,* Center for User-Oriented IT-design, Royal Institute of Technology, Stockholm September 27, 2001.

Nooked. (2008). Retrieved from http://www.nooked.com

Oasis-open. (2004). *Technology Reports, RDF Rich Site Summary (RSS).* Retrieved from http://www.oasis-open.org/cover/rss.html

Patch & McKinlay-Key. (2004). *Netiquette Note.* Retrieved from, http://www.synergywise.com/internet.html

Radio Userland. (2008). Retrieved from http://www.userland.com

RSS Reader. (2007). Retrieved from http://www.rssreader.com

Rubel, S. (2006). *Trends to watch, Part III: "RSS Inside," Micropersuasion blog*, December 2006. Retrieved from http://www.micropersuasion.com/2005/12/2006_trends_to__1.html

Van der Vlist, E. (2001). Building a semantic Web site. *XML.com*, May 2001.

Wikipedia. (2008). Retrieved from http://en.wikipedia.org/wiki/Aggregator

KEY TERMS

E-Mail: The term electronic mail shortened itself to E-mail, e-mail, and now email as it became an everyday process. E-mail is a cheap, fast text message delivered electronically over the Internet, or indeed local area networks.

Podcasting: The preparation and distribution of predominately audio for download to digital music players, such as the iPod player. A podcast is easily created from a digital audio file that must be saved in an MP3 format and then uploaded to the Web site of a service provider. The MP3 file then receives its own URL, which is inserted into an RSS XML document as an enclosure within an XML item tag. Once a podcast has been created, it is usually registered with content aggregators, such as podcasting.net or ipodder.org, for inclusion in podcast directories. Interested parties can then browse through these categories, or subscribe to specific podcast RSS feeds that will, in turn, download to their audio players automatically when they next connect

Resource Description Framework (RDF): RDF integrates a variety of applications from library catalogs and worldwide directories to syndication and aggregation of news, software, and content to personal collections of music, photos, and events using XML as an interchange syntax. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web**.**

RSS: An acronym for, *Really Simple Syndication*. Also known in parts by the terms *Resource Description Framework (RDF) Site Summary,* or *Rich Site Summary*. RSS has rapidly developed into a prevalent means of sharing content between Web sites.

RSS Enclosures: RSS version 2.0 encompasses a powerful feature; it allows an <item> to have an enclosure, this can, in simplistic terms, be likened to an e-mail having an attachment. In reality, enclosures hold huge potential and represent another step in the evolution of content syndication. By incorporating an <enclosure> subelement into an <item>, any RSS element can then describe a video or audio file.

RSS Readers: RSS aggregators can scour designated Web sites for updated feeds at regular intervals. An aggregator can gather updated headlines from the various sites and display them in a variety of ways for users.

Real–Time Thinking in the Digital Era

Yoram Eshet-Alkalai
The Open University of Israel, Israel

INTRODUCTION

In 2004, Eshet-Alkalai published a 5-skill holistic conceptual model for digital literacy, arguing that it covers most of the cognitive skills that users and scholars employ in digital environments, and therefore providing scholars, researchers, and designers with a powerful framework and design guidelines. This model was later reinforced by task-based empirical research (Eshet-Alkalai & Amichai-Hamburger, 2004). Until today, it is considered one of the most complete and coherent models for digital literacy (Akers, 2005); it is used as the conceptual design infrastructure in a variety of educational multimedia companies and was also described in the *Encyclopedia of Distance Learning* (Eshet-Alkalai, 2005). The conceptual model of Eshet-Alkalai consists of the following five digital literacy thinking skills:

1. **Photo-Visual Digital Thinking Skill:** Modern graphic-based digital environments require scholars to employ cognitive skills of "using vision to think" in order to create photo-visual communication with the environment. This unique form of digital thinking skill helps users to intuitively "read" and understand instructions and messages that are presented in a visual-graphical form, as in user interfaces and in children's computer games.
2. **Reproduction Digital Thinking Skill:** Modern digital technologies provide users with opportunities to create visual art and written works by reproducing and manipulating texts, visuals, and audio pieces. This requires the utilization of a digital reproduction thinking skill, defined as the ability to create new meanings or new interpretations by combining preexisting, independent shreds of digital information as text, graphic, and sound.
3. **Branching Digital Thinking Skill:** In hypermedia environments, users navigate in a branching, nonlinear way through knowledge domains. This form of navigation confronts them with problems that involve the need to construct knowledge from independent shreds of information that were accessed in a nonorderly and nonlinear way. The terms *branching* or *hypermedia thinking* are used to describe the cognitive skills that users of such digital environments employ.
4. **Information Digital Thinking Skill:** Today, with the exponential growth in available information, consumers' ability to assess information by sorting out subjective, biased, or even false information has become a key issue in training people to become smart information consumers. The ability of information consumers to make educated assessments requires the utilization of a special kind of digital thinking skill, termed *information skill*.
5. **Socio-Emotional Digital Thinking Skill:** The expansion of digital communication in recent years has opened new dimensions and opportunities for collaborative learning through environments such as knowledge communities, discussion groups, and chat rooms. In these environments, users face challenges that require them to employ sociological and emotional skills in order to *survive* the hurdles that await them in the mass communication of cyberspace. Such challenges include not only the ability to share formal knowledge, but also to share emotions in digital communication, to identify pretentious people in chat rooms, and to avoid Internet traps such as hoaxes and malicious Internet viruses. These require users to acquire a relatively new kind of digital thinking skill, termed *socio-emotional*, because it primarily involves sociological and emotional aspects of working in cyberspace.

The publication of Eshet-Alkalai's (2004) model of digital thinking skills has led to an extensive debate within the community of instructional technology designers, researchers, and educators, as to its validity and completeness, and a special panel in *ED MEDIA2005 Conference* (Montreal) was dedicated to it. This discussion (Aviram & Eshet-Alkalai, 2006) confirmed the validity and value of the model, but indicated that it lacked a sixth thinking skill: the *real-time thinking skill*, which relates to the ability of users to perform effectively in advanced digital environments, mainly high-tech machines, multimedia games, and multimedia training environments that require the user to process simultaneously large volumes of stimuli which appear in real time and at high speed. In the present article, real-time thinking is introduced as the sixth digital thinking skill, which completes the conceptual model of digital thinking skills.

BACKGROUND

The rapid development in digital technologies in recent decades confronts users with situations that require them to master a variety of technical, sociological, and cognitive skills, collectively termed *digital literacy* (Hargittai, 2002; Lanham, 1995), that are necessary to perform effectively in digital environments. Digital literacy is more than just the technical ability to operate digital devices properly; it comprises a variety of cognitive skills that are utilized in executing tasks in digital environments, such as surfing the Web, deciphering user interfaces, and chatting in chat rooms. Digital literacy has become a "survival skill" in the technological era—a key that helps users to work intuitively and effectively in executing complex digital tasks (Lazar, Bessiere, Ceaparu, Robinson, & Shneiderman, 2003).

In recent years, extensive efforts have been made to establish models that describe the cognitive skills that users employ in digital environments (e.g., Hargittai, 2002; Wegerif, 2004; Zins, 2000). Unfortunately, these efforts are usually local, focusing on a selected and limited variety of skills, mainly information-seeking skills (Zins, 2000) and, therefore, they do not cover the full scope of digital literacy. The present article presents real-time thinking, as an additional, sixth skill, in Eshet-Alkalai's (2004) holistic conceptual model of thinking skills in the digital era.

WORKING IN REAL-TIME ENVIRONMENTS

Imagine a pilot flying a jet, a driver driving a car, or a child playing a video game. In all these situations the users are exposed to a large flux of stimuli that *bombard* their cognition in real time, at very high speed, and in random temporal and spatial distribution. In all these situations the key to the users' successful performance is their ability to manage and synchronize these stimuli effectively. When operating such environments, users need to split their attention, reacting to various kinds of stimuli that appear simultaneously in different places on the monitor; they have to be able to execute different tasks simultaneously (multitasking); they need to be able to rapidly change their angle of view and perspective of the environment; and they have to respond to feedback that appears in real time. And above all—they have to quickly and effectively synchronize the chaotic multimedia stimuli into one coherent body of knowledge.

Today, situations that require real-time and high-speed processing of simultaneous large fluxes of information have become common in our lives, mainly in operating multimedia computer programs and advanced machines. This requires that users of today's digital environments master a special kind of thinking skill, here termed *real-time thinking*. Of course, real-time thinking is not new; it has been utilized ever since humans began to think and to synchronize information simultaneously in order to create knowledge. But in the digital era, with the central role of fast computers, multimedia environments, and devices that can process and present information in real time and at high speed, real-time thinking has become a critical skill. Real-time thinking situations usually require the utilization of split-attention skills in order to manage simultaneously large volumes of stimuli (text, sound, and images) that appear in real time and at a very high speed.

Today, most studies of real-time situations are conducted by researchers in the field of operations research, who explore human performance in aircrafts (Hamblin, Naidu, & Miller, 2006; Roessingh, 2005), cars (Casimir & Gilchrist, 2002) and other real-time working environments. Very little research has been done on the "soft" pedagogic aspects of real-time learning, such as digital games and language acquisition in real-time environments (Eshet-Alkalai, & Chajut, 2006; Pemberton, Fallahkhair, & Masthoff, 2004).

DIMENSIONS OF REAL-TIME THINKING

Simultaneous Synchronization

According to the *dual channel model* (Mayer, 2001), in multimedia environments, real-time stimuli are processed in parallel, independent channels (auditory-verbal and visual-pictorial), and the users' ability to synchronize them effectively is a major factor in their performance (Gopher, Weil, & Bareket, 2004). This model is useful for describing information processing in most multimedia environments such as microworld simulations (e.g., flight or driving simulations, in which the users operate a simulated aircraft or car). In the operation of these simulations, the users employ real-time thinking as they process large volumes of digital information simultaneously. Studies show that practicing real-time thinking in such real-time simulations is useful for improving synchronization ability, and therefore performance, of pilots (Gopher et al., 2004) and drivers (Barkan, Zohar, & Erev, 2003). However, information processing is not limited to these two channels only; in real-life situations, people utilize additional channels for processing real-time information, such as emotions and tactile information, which makes real-time thinking much more complex. Leuchter and Urbas (2002) showed that effective real-time thinkers are capable of synchronizing many channels of information processing simultaneously. One of the common applications of real-time synchronization is the case of language acquisition from subtitled films and from *living books*—computer-based storytelling multimedia programs (http://www.livingbooks.com). In living books, children simultaneously hear a story and watch its text on the monitor. Studies have shown that

children acquired a foreign language by synchronizing subtitles with narration in both living books (Eshet-Alkalai & Chajut, 2006) and interactive television (Pemberton et al., 2004).

High Speed

In addition to the simultaneous synchronization of information, modern, real-time digital environments are characterized by the fact that information passes through the information processing channels at very high speed. In car racing, fighting video games, or in flight simulations, users are *bombarded* with a large variety of high-speed moving stimuli (e.g., enemies, snipers, and road obstacles) as they move quickly from place to place and from scene to scene in the game environment. This requires that users not only master simultaneous synchronization skills, but also be able to respond quickly to obstacles. The response rate aspect of real-time thinking was investigated in a variety of real-time situations such as games (e.g., Erev, Luria, & Erev, 2006), simulations that require real-time thinking (Seagull, Wickens, & Loeb, 2001) and real-time situations drawn from daily life such as car driving (Barkan et al., 2003). These studies showed that the best performers were those who were able to process information quickly, indicating the crucial role of high-speed information processing in real-time situations. Roessingh (2005) found that flight simulators were useful in increasing the response rate of pilots and their ability to manage fast-moving stimuli. Eshet-Alkalai, and Chajut (2005) investigated the performance of gamers in shooting and car racing, real-time game environments, in which the users' success depends on their ability to respond in real time to fast-moving obstacles and stimuli. They found that the younger participants were the best real-time performers in their response rate and their ability to retain a high level of performance over a long period.

Attention Management and Multi-Tasking

In most real-time environments, users are required to be able to split their attention and respond simultaneously to stimuli that appear in different areas on the screen, and at the same time to conduct multitasking, physical activities such as operating the keyboard and the mouse simultaneously. For example, in *FIFA*—a real-time soccer-playing environment (http://fifa06.ea.com/) multiple events take place simultaneously on the screen: One player leads the ball, another runs after him, and on the other end of the yard, the gatekeeper is away from the gate. In such an environment, gamers must be able to split their attention among independent events in order to analyze the situation and adopt the most appropriate playing strategy. Gopher (1982) and Gopher et al. (2004) studied attention management and multitasking problems that relate to the management of real-time environments such as flight and basketball training simulations. Participants were tested on their ability to simultaneously split attention and relate to different types of stimuli such as sound, pictures, and text. Split attention and multitasking management—the ability of users to turn their attention simultaneously to different stimuli and to conduct different simultaneous tasks, was found to be the most significant factor in determining the level of performance and real-time thinking, and the best predictor of success in a flight course (Gopher, 1982). The combination of constant attention shifting and multitasking in real-time environments requires that real-time thinkers have a high degree of cognitive flexibility that helps them in shifting perspectives effectively. Erev and Gopher (1999) summarized the major aspects of selective attention strategies and showed that the ability to perform effectively in real-time environments is closely related to a high level of the user's selective attention management.

Multiple Perspectives

Many real-time environments, especially multimedia games and simulations, provide users with the ability to shift their perspectives, angles of view, and degree of resolution, as they work. For example, in FIFA, the user can shift constantly between different views: the single-player view; the gatekeeper view, and the entire system view; in Flight Simulator, the user can shift between the interior in-plane view and exterior view. The flexible shifting of perspectives in real-time environments may improve users' performance (Erev & Gopher, 1999) but it also requires them to be able to synthesize these multiple perspectives, accessed in real time, into a coherent decision or body of knowledge. For example, in Flight Simulator, the multiple perspectives that the users gain help them to select the flight route and make appropriate combat decisions. These situations require the utilization of a high level of real-time thinking. Roessingh (2005) found that pilots' ability to employ multiple perspectives in real-time flight simulations is closely related to their ability to successfully create mental models of the flight route and the interior of the plane.

Real-Time Feedback

Just-on-time feedback is an integral part of most real-time environments. These environments provide users with constant feedback that helps to improve their performance. For example, in typist training programs (e.g., http://www.21stsoftware.com/SS_Typing.htm), in car racing, multimedia simulations, or sport games, the system provides vocal and visual feedback on users' mistakes and achievements. This enables gradual improvement in performance through a process of successive approximation. Such a pro-

cess requires that users be able to manage a steady flow of real-time feedback of a different nature, in order to perform effectively. The positive impact of real-time feedback on the performance of users was illustrated in recent years by various empirical studies (e.g. Barron & Erev, 2004; Erev et al., 2006).

FUTURE TRENDS

Real-time digital thinking is a skill that evolves side by side with present-day technological developments. In the future, with the rapid evolution of technologies, mainly multimedia and game technologies, we will face increasingly more complex real-time environments, in which more stimuli occur much faster and users are required to manage more multiple-perspective and multitasking situations. This will open new frontiers for users, but at the same time, confront them with an ever-growing assortment of problems and challenges that must be addressed in order to perform effectively.

CONCLUSION

Real-time digital thinking is a cognitive skill that helps users of present-day digital environments to work effectively and to create knowledge from large volumes of information that are introduced simultaneously and at high speed. It is a pivotal skill for users of many present-day digital environments such as multimedia environments, digital games, and advanced machines (e.g., aircraft and cars). Studies from the field of operations research show that a high level of real-time digital thinking is critical for successfully operating real-time environments.

In the present article, real-time thinking is presented as a sixth digital thinking skill that completes Eshet-Alkalai's (2004) holistic conceptual mode of digital thinking skills, in which an attempt was made to represent digital literacy and human thinking skills in the digital era with five digital thinking skills.

Effective real-time thinkers are able to successfully process simultaneous stimuli in real time, manage stimuli that appear at very high speed, split their attention effectively, conduct multitasking jobs, simultaneously manage the multi-perspective representations of the environment, and utilize effectively the just-on-time feedback provided by the system.

REFERENCES

Akers, C. (2005). IRT's top 20. *Library Instruction Roundtable News, 27*(4), 8.

Aviram, R., & Eshet-Alkalai, Y. (2006). Towards a theory of digital literacy: Three scenarios for the next steps. *European Journal of Open Distance E-Learning. European Journal of Open Distance E-Learning, 2,* Retrieved June 26, 2006, from http://www.eurodl.org/

Barkan, R., Zohar, D., & Erev, I. (2003). Accidents and decision making under uncertainty: A comparison of four models. *Organizational Behavior and Human Decision Processes, 74,* 118-144.

Barron, G., & Erev, I. (2004). Small feedback based decisions and their limited correspondence to description based decisions. *Journal of Behavioral Decision Making, 16,* 215-233.

Casimir, J. L., & Gilchrist, I. (2002). Stimulus-driven and goal-driven control over visual selection. *Journal of Experimental Psychology: Human Perception and Performance, 28*, 902-912.

Erev, I., & Gopher, D. (1999). A cognitive game theoretic analysis of attention strategies, ability and incentives. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and applications*. Cambridge, MA: MIT Press.

Erev, I., Luria, A., & Erev, A. (2006, March 1). On the effect of immediate feedback. In Y. Eshet-Alkalai,, A. Caspi, & Y. Yair (Eds.), *Learning in the technological era. Proceedings of the Chais Conference* (pp. 26-30). Raanana, Israel: Open University Press.

Eshet-Alkalai, Y. (2004). Digital literacy: A conceptual framework for survival skills in the digital era. *Journal of Multimedia and Hypermedia, 13*(1), 93-106.

Eshet-Alkalai, Y. (2005). Thinking skills in the digital era. In C. Haward, J. V. Bottcher, L. Justice, K. Schenk, P. L. Rogers, & G. A. Berg (Eds.), *Encyclopedia of distance learning* (Vol. 1, pp. 1840-1845). London: Idea Group Reference.

Eshet-Alkalai, Y., & Amichai-Hamburger, Y. (2004). Experiments in digital literacy. *Cyberpsychology & Behavior, 7,* 421-429.

Eshet-Alkalai, Y., & Chajut, E. (2006, March 1). Living books: On the acquisition of reading skills in multimedia environments. In Y. Eshet-Alkalai, A. Caspi, & Y. Yair (Eds.), *Learning in the technological era. Proceedings of the Chais Conference* (pp. 61-66). Raanana, Israel: Open University Press.

Gopher, D. (1982). A selective attention test as a predictor of success in flight training. *Human factors, 24*(2), 173-183.

Gopher, D., Weil, M., & Bareket, T. (2004). Transfer of skill from computer game trainer to flight. *Human Factors, 36*(3), 387-405.

Hamblin, C. J., Naidu, S., & Miller, C. (2006, February). Usability analysis of a computer-based avionics system. *Uability News, 8*(1). Retrieved June 26, 2006, from http://www.usabilitynews.org

Hargittai, E. (2002). Beyond logs and surveys: In-depth measures of people's Web use skills. *Journal of the American Society for Information Science and Technology, 53*(14), 1239-1244.

Lanham, R. (1995). Digital literacy. *Scientific American, 273,* 253-255.

Lazar, J., Bessiere, K., Ceaparu, I., Robinson, J., & Shneiderman, B. (2003, Winter). Help! I'm lost: User frustration in web navigation. *IT & Society, 1*(3). Retrieved June 26, 2006, from www.ITandSociety.org

Leuchter, S., & Urbas, L. (2002, September 9-12). Simulation based situation awareness training for control of human-machine-systems. In V. Petrushin, P. Kommers, D. Kinshuk, & I. Galeev (Eds.), *IEEE International Conference on Advanced Learning Technologies. Media and the Culture of Learning,* Kazan, Russia. Palmerston North, New Zealand: IEEE Learning Technology Task Force. Retrieved June 26, 2006, from http://www.zmms.tu-berlin.de/~sandro/doc/icalt2002.pdf

Mayer, R. E. (2001). *Multimedia learning*. Cambridge, UK: Cambridge University Press.

Pemberton, L., Fallahkhair, S., & Masthoff, J. (2004). Towards a theoretical framework for informal language learning via interactive television. In D. Kinshuk, G. Sampson, & P. Isaias, (Eds.), *Cognition and exploratory learning in the digital age (CELDA 2004)* (pp. 27-34). Lisbon, Portugal: IADIS Press.

Roessingh, J. M. (2005). Transfer of flying manual skills from pc-based simulation to actual flight-comparison of in-flight measured data and instructor ratings. *International Journal of Aviation Psychology, 1,* 67-90.

Seagull, J., Wickens, D. D., & Loeb, R. G. (2001, October 8-12). When is less more? Attention and workload in auditory, visual and redundant patient-monitoring conditions. *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society*, Santa Monica, CA .

Wegerif, R. (2004). Literature review in thinking skills, technology and learning. *Nesta Futurelab Series* (Report #2). Retrieved June 26, 2006, from http://www.nestafuturelaborg/research/reviews/ts01/.htm

Zins, C. (2000). Success, a structures search strategy: Rationale, principles and implications. *Journal of the American Society for Information Science, 51*, 1232-1247.

KEY TERMS

Channel Model: A model which suggests that in multimedia environments information is processed in parallel independent channels: verbal and visual-pictorial.

Digital Era: A term used to describe today's era, in which digital technologies are used in almost every aspect of life.

Digital Literacy: A term used to describe the ability of users to perform in digital environments.

Digital Thinking Skills: A refinement of the term *digital literacy*, describing the variety of thinking skills that comprise digital literacy.

Real-Time Digital Thinking: A term used to describe the thinking skill that is employed in many of today's digital environments. In such environments, the user needs to manage large volumes of information, perspectives, and tasks that are introduced in real time and very high speed.

Real-Time Digital Environment: A term used to describe digital environments in which users need to manage in real-time simultaneous stimuli, multiple perspectives, and multitasking.

Synchronous Stimuli: Stimuli that are emitted simultaneously by the environments. For example, simultaneous sound, text, and images to which users of digital environments are exposed.

Category: Multimedia Technology

Recent Progress in Image and Video Segmentation for CBVIR

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

A simple search from EI Compendex by using the term "image segmentation" only in title field could produce around 5000 records (Zhang, 2006). However, as no general theory for image segmentation for different application domains, particular algorithms have been developed. The domain of Content-Based Image Retrieval (CBIR) is such a typical example, where many specific techniques have been proposed. An introduction focused on research works before 2004 can be found in Zhang (2005). This paper is an up-to-date and extended version from CBIR to CBVIR (Content-Based Visual Information Retrieval) by including CBVR (Content-Based Video Retrieval), which focused on the progress in last 3 years, and especially on video segmentation.

BACKGROUND

A Formal Definition of Image Segmentation

Image segmentation is the first step and also one of the most critical tasks of image analysis. It is often described as the process that subdivides an image into its constituent parts and extracts those parts of interest (objects).

A formal definition of image segmentation, supposing the whole image is represented by $f(x, y)$, and $f_i(x, y)$, $i = 1, 2, \ldots, n$ are disjoint non-empty regions of $f(x, y)$, consists of the following conditions (Fu, 1981):

1. $\bigcup_{i=1}^{n} f_i(x, y) = f(x, y)$;
2. For all i and j, $i \neq j$, there exits $f_i(x, y) \cap f_j(x, y) = \varnothing$;
3. For $i = 1, 2, \ldots, n$, it must have $P[f_i(x, y)] = TRUE$;
4. For all $i \neq j$, there exits $P[f_i(x, y) \cup f_j(x, y)] = FALSE$; where $P[f_i(x, y)]$ is a uniformity predicate for all elements in $f_i(x, y)$ and $\varnothing$ represents an empty set. Considering the real situation in practice, the following condition can be added:
5. For all $i = 1, 2, \ldots, n$, $f_i(x, y)$ is a connected component.

In the above conditions, condition (1) points out that the summation of segmented regions could include all pixels in an image; condition (2) points out that different segmented regions could not overlap each other; condition (3) points out that the pixels in the same segmented regions should have some similar properties; condition (4) points out that the pixel belonging to different segmented regions should have some different properties; and, finally, condition (5) points out that the pixels in the same segmented region are connected.

Definition Extension to Video Segmentation

If a 2-D still gray level image is represented by $f(x, y)$, then its extension to 3-D moving images or sequences of images (video) can be represented by $f(x, y) \Rightarrow f(x, y, t)$. In video domain, two kinds of segmentation can be distinguished: *spatial segmentation* and *temporal segmentation*. In *spatial segmentation*, each frame of $f(x, y, t)$ can be denoted as $f_t(x, y)$, which is a 2-D still image and the above formal definition for image segmentation can still be used.

The *temporal segmentation* of video can be defined as follows. Given a video sequence $f(x, y, t) = \{f_1(x, y), f_2(x, y), \ldots, f_i(x, y), \ldots, f_n(x, y)\}$, the k-th partition of $f(x, y, t)$ can be denoted as $g_k(x, y) = \{f_i(x, y), f_{i+1}(x, y), \ldots, f_{i+ik-1}(x, y)\}$, where i_k is the number of frames in the k-th partition, and $\sum_{k=1}^{m} i_k = n$. The formal definition of temporal video segmentation consists of the following conditions:

1. $\bigcup_{k=1}^{m} g_k(x, y) = g(x, y)$;
2. For all k and l, $k \neq l$, there exits $g_k(x, y) \cap g_l(x, y) = \varnothing$;
3. For $k = 1, 2, \ldots, m$, it must have $P[g_k(x, y)] = TRUE$;
4. For all $k \neq l$, there exits $P[g_k(x, y) \cup g_l(x, y)] = FALSE$.

Compared with the definition of image segmentation, the corresponding condition (5) has already been included in the definition of $g_k(x, y)$. In other words, the frames in the same shot are connected in time.

MAIN THRUST

In recent years, many researches on image and video segmentation for CBVIR are carried out. Two of them are described with some detail in the following, some others are just briefly indicated.

Color Image Segmentation in Feature and Image Spaces

In CBVIR, color information plays an important role. Early retrieval algorithms for CBVIR are often based on the color information of image or object. Many current retrieval algorithms are still using color information to derive semantic description. Therefore, efficient color segmentation techniques are critical for CBVIR.

One color segmentation technique based on *watershed* and feature space analysis is described below. This technique made the combination of *watershed* transform and feature space analysis.

In most cases, the *watershed* algorithm is applied on image domain (usually on the edge image). It focuses on local color feature instead of global color distribution. In edge image, the local minima exist in the interior of objects and high altitude appears on the boundary of objects. After a flooding process, dams (*watershed* lines) will be constructed on object boundary and different objects are separated. In this way, it captures only information of local color feature instead of global color distribution.

In feature space, one obstacle of segmentation is the difficulty relies on color clustering. Researchers have noticed (Park, 1998; Pauwels, 1999) that color distribution in 3-D color space cannot be well approximated by the traditional parameter based clustering algorithm (such as *K*-mean model or Gaussian mixture model). For example, *K*-mean is unable to handle unbalanced or elongated clusters. Gaussian mixture model is not appropriate for cluster with irregular shape. One example is shown in Fig. 1. The original image PEPPERS is in Fig. 1(a), its pixel distribution in color RGB space is projected onto 2-D plane as in Fig. 1(b) and its pixel distribution in color $L^*a^*b^*$ space is projected onto 2-D plane as in Fig.1(c). In Fig. 1(b), the distribution has irregular shapes, some clusters are sharp and compact, and some are flat. In Fig. 1(c), the clusters seem more salient, but it is still hard to get the boundary between clusters with parametric models.

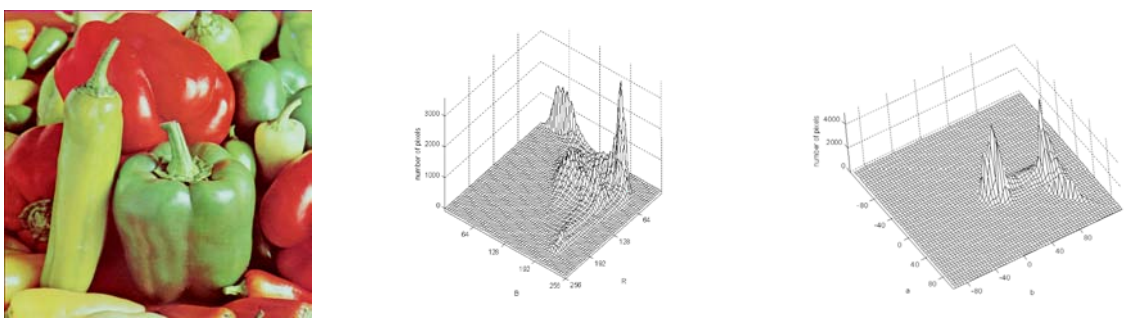
To solve the problem, the *watershed* algorithm is applied on a 3-D $L^*a^*b^*$ histogram $H(x, y, z)$ to capture the feature space information. A labeling process with the following steps is used to cluster the color histogram (Dai, 2006):

1. Get the reverse histogram $H'(x, y, z) = -H(x, y, z)$ $(0 \le x < u, 0 \le y < v, 0 \le z < w)$.
2. Get all local minimum of the reverse histogram H', label them as 1, 2, 3, …, m.
3. Find the unlabeled bin in H' with minimum value and label it according to its neighbors:
   i. If more than one label appears in its neighborhood, it is a "dam" bin, and it will be labeled as 0.
   ii. If else, label it the same as its labeled neighbor.
4. Go to step (3) until all non-zero bins are labeled.

After obtaining the *watershed* in the color histogram, the results can be brought back to the image space. The following post-process steps are used to get continuous homogeneous regions with meaningful size.

1. Get all pixels with corresponding color bins labeled

Figure 1. Pixel distribution of image PEPPERS in color space



(a) *(b)* *(c)*

Figure 2. Examples of segmentation result

(a)

(b)

as 0, and label them as 0 in image domain.

2. Get all continuous (4-neighborhood) regions, and label them the same value in image domain.
3. Label all pixels in regions with size smaller than a predefined threshold as 0.
4. (i) Label the left regions as 1, 2, …, *m*, and all pixels in the labeled regions are labeled the same as their correspondent regions. (ii) Label the rest pixels the same as the label of their correspondent histogram.

The initial segmentation result is thus obtained (see Fig. 2(a)). It is composed of two parts: some continuous homogeneous regions with meaningful size, and some scattered pixels labeled as zero, which are called uncommitted pixels. A Markov Random Field algorithm is followed to refine the initial result. Two efficient energy minimization techniques are applied to get sub-optimal results. One is Highest Confidence First (HCF), which can enforce continuity properties. Another is Graph Cuts, which can efficiently avoid the problem of over-smoothness. The final result obtained by HCF is shown in Fig. 2(b).

Video Segmentation with Parametric Model

In CBVR, a preliminary step is shot boundary detection that partitions video data into fundamental units called shots that are composed of a sequence of frames taken by one camera without interruption. Recently, a parametric model for shot boundary detection (and key frame extraction) is proposed (Chen, 2008). The *autoregressive* (AR) modeling is used to model the frame feature sequence over time and to make the future analysis in the AR parametric space.

Autoregressive model is a simple time-series model widely used for prediction and modeling. An AR model with order p (4 ~ 24) can be expressed as follows (Martin, 2000):

$$\boldsymbol{x}_n = \sum_{j=1}^{p} \boldsymbol{a}_j \boldsymbol{x}_{n-j} + \eta_n \tag{1}$$

where η_n is the uncorrelated noise with variance σ.

The parameters in AR model can be estimated by using the direct-form of recursive least squares (RLS) FIR adaptive filter as follows:

- Initial: $\boldsymbol{a}(0) = 0$, $\boldsymbol{P}(0) = \delta^{-1}\boldsymbol{I}$, where δ is a small positive number (usually 0.01), $\boldsymbol{I}$ is an identity matrix.
- Update: for $n = 1, 2, \ldots$

$$\boldsymbol{e}(n) = \boldsymbol{d}(n) - \boldsymbol{a}^T(n-1)\boldsymbol{u}(n) \tag{2}$$

$$\boldsymbol{k}(n) = \frac{\boldsymbol{P}(n-1)\boldsymbol{u}(n)}{\lambda + \boldsymbol{u}^{\mathrm{T}}(n)\boldsymbol{P}(n-1)\boldsymbol{u}(n)} \tag{3}$$

$$\boldsymbol{P}(n) = \frac{1}{\lambda}\left[\boldsymbol{P}(n-1) - \boldsymbol{k}(u)\boldsymbol{u}^T(n)\boldsymbol{P}(n-1)\right] \tag{4}$$

$$\boldsymbol{a}(n) = \boldsymbol{a}(n-1) + \boldsymbol{k}(n)\boldsymbol{e}(n) \tag{5}$$

where λ is a forgetting factor (can be selected from 0.92 to 0.98), $\boldsymbol{a}(n)$ is the AR coefficient vector, $\boldsymbol{u}(n)$ is the input vector, $\boldsymbol{d}(n)$ is the output vector, $\boldsymbol{k}(n)$ is the gain for the recursive procedure, $\boldsymbol{e}(n)$ is the model prediction error, $\boldsymbol{P}(n)$ is the inverse of the input covariance matrix.

The problem of shot boundary detection can be viewed as an estimation of the on-line *autoregressive* system parameters. It is believed that there would be a shot boundary

at the place where the system structure changes. This can be shown by focusing on the AR prediction errors (APE) in the recursive parameter estimation procedure. In the AR model, APE can measure the accumulated errors as long as the orders of the model. When the APE grows big, it means the present model's parameter cannot fit the current frame well, and a shot change may happen. The transition between two adjacent shots can be divided into two classes: abrupt shot boundary where the change takes place over a single frame, and gradual shot boundary where the change of video content occurs over a sequence of frames gradually. Both of them can be detected with this method.

In order to detect possible abrupt shot boundary, an adaptive threshold on that parameter should be selected:

$$T_b = km \tag{6}$$

where m is the local APE mean values on a 1-D temporal window of size p, and k is suggested to be 2 to 3 according to some empirical studies.

For the gradual transitions, the frame sequences change progressively, causing the correspondent evolvement of the APE among consecutive frames. Such a change can be detected by combining the twin-comparison method (Furht, 1995) with APE.

The proposed method has been tested by using a variety of test videos, as listed in Table1. Among the test videos, the movie (V1) is from the film "Star Wars: The Phantom Menace". The cartoon (V2) is from an episode of "The Incredible". The news (V3) and the advertisement (V4) are from the CCTV-9, and the sports video (V5) is a table tennis game in Olympic Games. The movie and the cartoon are characterized by significant camera parameter changes like zoom-in/out, pans, abrupt camera movement, as well as significant object motion, but the shot change patterns are mostly abrupt ones. The news and the advertisement have nearly all kinds of shot boundaries, including cut, fade in/out, and dissolves.

Two measures: recall (R) and precision (P), are used to evaluate the performance of detection algorithm.

$$R = \frac{\#\text{ of hits}}{\#\text{ of hits} + \#\text{ of misses}} \times \% \tag{9}$$

$$P = \frac{\#\text{ of hits}}{\#\text{ of hits} + \#\text{ of false alarms}} \times \% \tag{10}$$

The detection performance obtained for all test videos are shown in Table 1, too. The proposed method is based on color histogram which is robust to camera motion as well as object motion, and not sensitive to the transition and rotation of the view axis. This intrinsic property of the color histogram and the robust of the RLS to the white noise make the proposed method not sensitive to the motion interruption. In the meanwhile, the parametric model can distinguish motion from some shot changes by the long sequence modeling. Therefore, the performance of this method is in general satisfactory. On the other side, histograms are sensitive to illumination changes, so the scenes with explodes in the movies were falsely detected as shot boundaries.

More Technique Examples

The current research works on CBVIR are toward Semantic-Based Visual Information Retrieval (SBVIR), see (Zhang, 2007). One important stage in SBVIR consists of extracting objects from image and video. An algorithm has been devised for fast, fully automatic, and reliable object segmentation from live video for scenario with static camera (Ong, 2006). It adaptively determine the threshold for change detection; generates robust stationary background reference frame; select the reference frame to improve segmentation results; and refine the spatial change detection mask by incorporating information from edges, gradients, and motion.

Shot boundary detection is the first step for video analysis. After shots are obtained, related shots can be further grouped into scenes with some help of spatial and/or temporal constraints. This process is called shot grouping or

Table 1. Experiments and results on test videos

Video	Frame	Shot #	Cut			Fade in/out			Dissolve		
			#	R(%)	P(%)	#	R(%)	P(%)	#	R(%)	P(%)
V1	10000	188	188	90.5	94.5	0	—	—	0	—	—
V2	7000	130	130	85.0	96.6	0	—	—	0	—	—
V3	10500	54	27	92.6	96.2	15	93.3	100	12	83	100
V4	8000	199	191	96.3	94.4	2	100	100	6	83	83
V5	7500	48	42	97.6	97.6	0	—	—	6	100	100

scene segmentation. Both shot boundary detection and shot grouping are inherently similar tasks. One method has been proposed to formulate both shot boundary detection and shot grouping as the problem of sequential change detection (Lu, 2006). The difference is that in shot boundary detection the feature sequence is extracted from video frame (e.g.,, using frame color histogram as features) while in shot grouping the feature sequence is extracted from shots (e.g., using shot color histograms as features).

Image semantics can be represented more accurately by keywords than by low-level visual features. In this regard, automated image classification and annotation are considered to be promising. Besides obtaining text annotation, a successful image categorization will significantly enhance the performance of the content-based image retrieval system by filtering out images from irrelevant classes during matching.

Recent progresses in object recognition and image annotation have shown that local salient features are more informative in describing image content than global features (Csurka, 2004; Fergus, 2003). In the process of image classification, features are required to be common for the same class and discriminative for different classes. In the object semantic concept learning, the model should emphasize the object in an image category.

Two approaches have been proposed for detecting local salient features (Xu, 2007), which are capable of selecting the most informative features based on the detected salient patches. One approach detects the salient patches by the local salient feature detector proposed in (Kadir, 2001). This detector finds regions that are salient over both location and scale. For each point in an input image, a number of intensity histograms are calculated in circular regions of different radiuses (scales). The entropy of each histogram is then calculated and the local maximum is selected as a candidate region. The regions with highest saliency over the image provide the features. In practice, this method gives stable identification of features over a variety of sizes and copes well with intra-class variability. Only intensity information is used to detect and represent features. Another approach transforms image data into scale-invariant coordinates relative to local features by SIFT (Scale Invariant Feature Transform; Lowe, 1999). SIFT is built by selecting key locations at maxima and minima of a difference of Gaussian function applied in scale space. Maxima and minima of this scale-space function are determined by comparing each pixel in the pyramid to its neighbors.

FUTURE TRENDS

The following research directions should be considered in the future:

1. Both spatial and temporal information are included in video, how to combine them or fuse them in segmentation is an interesting topic.
2. The domain of content-based visual information retrieval has particular requirements for extracting useful information; specific segmentation techniques should be developed.
3. With the trends of more research focused on *SBVIR* (Zhang, 2007), semantic image and video segmentation will play an even significant role.

CONCLUSION

Some recent progress in image and video segmentation for *CBVIR* are reported. Compared to other application areas of segmentation, CBVIR has certain particularities. Further development of image and video segmentation techniques should take them into consideration.

It is expected that the techniques of image and video segmentation will evolve with the advancement of CBVIR and it will also push CBVIR to an even high level.

ACKNOWLEDGMENTS

This work has been supported by Grants NNSF-60573148 and SRFDP-20050003013.

REFERENCES

Chen, W., Zhang, Y.J. (2008). Parametric model for video content analysis. *Pattern Recognition Letters*, 29 (3), 181-191.

Csurka, G.., Dance, C.R., Fan, L., et al. (2004). Visual categorization with bags of keypoints. In: *Proceedings of 8ECCV*, 11-14.

Dai, S.Y., Zhang, Y.J. (2006). Color image segmentation in both feature and image space. In: *Advances in Image and Video Segmentation*, Zhang Y-J ed., IRM Press, Chapter 10 (209-227).

Fergus, R., Perona, P., Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In: *Proceedings of CVPR*, 2: 264-271.

Fu, K.S., Mui, J.K. (1981). A survey on image segmentation. *Pattern Recognition, 13*: 3-16.

Furht, B., Smoliar, S. W., Zhang, H. J. (1995). *Video and Image Processing in Multimedia Systems*. Boston, MA:

Kluwer.

Kadir, T., Brady, M., (2001). Scale, saliency and image description. *International Journal of Computer Vision* 45(2): 83-105.

Lowe, D.G. (1999). Object recognition from local scale-invariant features. In: *Proceedings of ICCV*, 2: 1150-1157.

Lu, H., Li, Z.Y., Tan, Y.P., et al. (2006). Video shot boundary detection and scene segmentation. In: *Advances in Image and Video Segmentation*, Zhang Y-J ed., IRM Press, Chapter 9 (188-207).

Martin, J.R. (2000). A metric for ARMA processes. *IEEE Trans. Signal Processing*. 48(4): 1164-1170.

Ong, E.P., Lin, W.S., Tye, B.J., et al. (2006). Fast automatic video object segmentation for content-based applications. In: *Advances in Image and Video Segmentation*, Zhang Y-J ed., IRM Press, Chapter 7 (140-160).

Park, S. H., Yun, I. D., Lee, S. U. (1998). Color image segmentation based on 3-D clustering: Morphological approach. *Pattern Recognition*, *31*(8): 1061-1076.

Pauwels, E.J., Frederix, G.. (1999). Finding salient regions in images: Nonparametric clustering for image segmentation and grouping. *Computer Vision and Image Understanding*, *75*(1/2): 73-85.

Xu, F., Zhang, Y.J. (2007). A novel framework for image categorization and automatic annotation. In: *Semantic-Based Visual Information Retrieval*, Zhang Y-J ed., IRM Press, Chapter 5 (90-111).

Zhang, Y.J. (2005). New Advancements in Image Segmentation for CBIR. In: *Encyclopedia of Information Science and Technology,* Idea Group Reference, Mehdi Khosrow-Pour *ed.*, 4: 2105-2109.

Zhang, Y.J. (2006). An Overview of Image and Video Segmentation in the Last 40 Years". In: *Advances in Image and Video Segmentation*, Zhang Y-J ed., IRM Press, Chapter 1 (1-15).

Zhang, Y.J. (2007). *Semantic-Based Visual Information Retrieval*. IRM Press.

R

KEY TERMS

AR Model Order: It characterizes how close the relation is for the frame subsequence in the visual content.

Clustering: A process to group, based on some defined criteria, two or more terms together to form a large collection. In the context of image segmentation, clustering is to gather several pixels or groups of pixels with similar property to form a region.

Feature Space Analysis: Image analysis performed in the feature space, in which each point corresponds to a feature value extracted from image.

Gaussian Mixture Model: A traditional parameter-based clustering model. All samples are assumed to belong to one of several Gaussian distributions, the ensemble can be described by the sum of several Gaussian distribution.

Graph Cuts: A widely used technique in energy minimization and clustering. When there are only two labels, the energy minimization problem can be reduced directly to a problem of computing the max-flow (or min-cut) of a graph.

Highest Confidence First: An efficient technique for energy minimization. It is a deterministic minimization algorithm, which is suitable for assigning label to pixels with unknown label, because it introduces an uncommitted label.

***K*-mean Model: A** traditional parameter-based clustering model. *K* initial classes are represented by their mean values, the following iterations try to minimize the distance between the samples with their respect means.

Watershed: A transform or an algorithm for image segmentation. It is traditionally classified as a region-based segmentation approach. The idea underlying watershed transform comes from geography.

Reconciling the Perceptions and Aspirations of Stakeholders in a Technology Based Profession

Glenn Lowry
United Arab Emirates University, UAE

Rodney Turner
Monash University, Australia

INTRODUCTION

Information systems professionals help to achieve business and organizational goals through the use of information technology.[a] The information systems (IS) profession is team-oriented and project-based. It involves a blend of business knowledge and understanding, technical skills, and working relationships with business and technical professionals. The skills and knowledge involved range from traditional computing, wide ranging business related studies, to "soft" skills useful in working with individuals and teams to achieve organizational objectives.

IS students are first and foremost concerned with future employability. Employers, on the other hand, often indicate that they want new graduates who can be immediately productive in their environment.

Are the aspirations of students and employers fundamentally incompatible? How can IS educators help to find a workable and satisfying balance? How can information systems educators achieve a better fit between the workplace and the university "studyplace"?

BACKGROUND

The past decades have been characterized by a rapidly and constantly changing business environment. Lee, Trauth, and Farwell (1995) argued that technological and sociological developments facilitated by evolving information technology and changing business needs has made it necessary for IS professionals to develop a wider range of nontechnical skills than was previously the case. Similar views have been expressed by many others, including Burn, Ng, and Ma (1995), Cafasso (1996); Lowry, Morgan, and FitzGerald, (1996); Morgan, Lowry, and FitzGerald (1998). Beise, Niederman, Quan, and Moody (2005) saw a need for the reform of undergraduate IS programs to specifically target the global IT environment by adding a global business perspective to existing curricula or by developing new curricula focusing on globalized information management. The perpetual global competition for skilled information systems professionals continues unabated (Florida, 2005; Schwarzkopf, Saunders, Jasperson & Croes, 2004).

The preparation of IS professionals must encompass a body of knowledge and a repertoire of technical skills identified by various professional bodies (ACM-AIS, 2002; Cheney, Hale, & Kasper, 1990; Cohen, 2000; Davis, Gorgone, Feinstein & Longenecker, 1997; Gorgone & Gray, 1999; Lidtke, Stokes, Haines & Mulder, 1999; Lyytinen & King, 2004; Mulder & van Weert, 2000; Underwood, 1997). IS curricula must take cognizance of the greater diversity within the IT labour force as a result of globalization (Trauth, Huang, Morgan, Quesenberry & Yeo, 2006).

The persistent research finding that employers want graduates who possess better business skills has often been interpreted by academics to mean that more traditional, formal business subjects such as accounting, economics, business finance, and marketing should be taught alongside traditional technical or "hard" skill subjects such as systems analysis & design and programming in particular languages. (Amarego, 2005; Gardiner, 2005; Holt, MacKay & Smith, 2004; Lee, 2005; Leong & Tan, 2004; Litecky, Arnett & Prabhakar, 2004; Medlin, 2004; Trauth, Farwell, & Lee, 1993; Van Slyke, Kittner, & Cheney, 1997). Beachboard and Parker (2003) observed that course requirements in model curricula likely contain more technical material than can be covered in an undergraduate course. On the other hand, "soft" areas such as teamwork, communication skills, ability to accept direction, and others are expected to be somehow "picked up" along the way by students through an unspecified, osmotic process and not addressed as part of a curriculum. Unfortunately, anecdotal evidence continues to suggest that at least some new graduates continue to lack "soft skills" (Maiden, 2004). Berghel and Sallach (2004) maintain that a curriculum must take account of developments in technologies, business models, and applications to enable students to build the necessary competencies.

The work presented here is part of an ongoing research program that investigates the views of major IS curriculum stakeholders including employers, IS practitioners, currently

enrolled students, recent graduates, and academics. The data were gathered from surveys of IS practitioners and IS decision makers in Australia and covered all industry sectors as well as business unit sizes. We argue that IS practitioners, employers, and students see little value in some of the more formal business subject areas that often form the core of an IS degree offered in business or commerce faculties. These stakeholder groups see more value in the development of "soft skills" useful in client interaction, often through cooperative education in which students are placed in real-world roles as novice business analysts (Dressler & Keeling, 2004; Fincher, Clear, Petrova, Hoskyn, Birch & Claxton, 2004; Gallivan Truex & Kvasny, 2004). The findings have serious implications for IS educators and IS curriculum design (Turner & Lowry, 1999a; 1999b; 2000; 2001; 2002; 2003; Turner, Fisher & Lowry, 2004a; 2004b; 2005a; 2005b; Lowry & Turner, 2005a; 2005b; Turner, Lowry & Fisher, 2005; Turner, Fisher & Lowry, 2005a; 2005b; Lowry, Turner & Fisher, 2006).

IS CURRICULUM CONTENT AND DELIVERY IN THE FIRST DECADE OF THE 21ST CENTURY

In a 1999 study, the authors began to suspect that the "other business skills" desired of new IS graduates were not syn-

Table 1. Comparative ratings of academic subjects by IS practitioners and employers

	IS/IT Professionals		IS/IT Employers	
Skills	**Mean**	**sd**	**Mean**	**sd**
Communications & Report Writing	6.02	1.05	6.09	0.81
Analysis & Design	5.87	1.09	5.63	1.26
Client server applications	5.67	0.92	5.37	1.15
Business Applications	5.65	1.11	5.76	1.20
Use operating systems	5.60	1.10	5.39	1.27
Database design	5.55	1.25	5.12	1.16
Management	5.54	1.03	5.20	1.10
Knowledge of PC apps	5.43	1.22	5.41	1.37
Project Management	5.43	1.16	5.60	1.24
E-Commerce/E-business development	5.33	1.23	4.78	1.38
Apply OOPs	5.26	1.25	4.61	1.51
LAN & Data Communications	5.22	1.27	5.55	1.22
Large System experience	5.12	1.19	4.54	1.53
Business Ethics	5.07	1.57	5.23	1.52
Web design/development	4.96	1.54	4.67	1.15
Organizational Behavior	4.90	1.41	4.92	1.34
Data mining/Data warehousing	4.76	1.36	4.66	1.36
Apply 3GLs	4.70	1.41	4.15	1.58
CASE applications	4.51	1.32	3.80	1.42
Knowledge base/Expert systems	4.49	1.42	4.20	1.37
ERP implementations & operations	4.48	1.39	4.33	1.61
Marketing	4.35	1.52	4.39	1.34
Business Finance	4.30	1.50	4.54	1.47
Operations Research	4.29	1.26	4.32	1.26
Mathematical Modeling	4.25	1.44	3.97	1.49
International Business	4.24	1.59	3.69	1.56
Business Statistics	4.18	1.40	4.33	1.38
Accounting	4.13	1.55	4.68	1.38
Business or Commercial Law	4.07	1.55	4.12	1.45
Psychology	3.70	1.76	3.85	1.46
Economics	3.63	1.50	3.68	1.47
Foreign Languages	3.15	1.78	3.04	1.46
n=	**136**		**138**	

Table 2. Comparative importance of soft skills by IS/IT professionals and employers

	IS/IT Professionals		IS/IT Employers	
Skills	**Mean**	**sd**	**Mean**	**sd**
Work as a team	6.52	0.66	6.39	0.81
Problem solving skills	6.44	0.57	6.37	0.68
Work under pressure	6.42	0.78	6.27	0.75
Quickly acquire new skills	6.37	0.64	6.15	0.73
Independently acquire new skills	6.35	0.72	6.23	0.71
Meet deadlines	6.35	0.68	6.13	0.81
Work independently	6.27	0.94	6.22	0.65
Time management	6.21	0.95	5.98	0.84
Problem definition skills	6.18	0.74	6.14	0.74
Ongoing professional development willingness	6.18	0.89	5.93	0.89
Written communication skills	6.18	0.85	6.04	0.76
Client focused service ethic	6.16	1.00	6.09	0.94
Handle concurrent tasks	6.16	0.81	6.08	0.81
Interact with people of different backgrounds	6.13	0.73	6.03	0.85
Think creatively	6.08	0.89	6.09	0.71
Work with people from different disciplines	6.04	0.74	6.10	0.82
Accept direction	6.03	0.89	5.98	0.84
Information seeking skills	5.83	0.96	5.82	0.93
Oral presentation skills	5.79	1.07	5.56	0.88
Place organizational objectives first	5.73	0.95	5.74	0.97
Business analysis skills	5.63	1.03	5.51	1.04
Leadership potential	5.18	1.08	4.99	0.94
Good sense of humor	5.15	1.35	5.58	1.14
Able to prepare multimedia presentations	4.73	1.25	4.32	1.38
n=	**136**		**138**	

onymous with traditional business curriculum subjects. Study results indicated that, of nine business subjects that are typically included in IS curricula, only three, Accounting, Business Ethics, and Management were judged to be important by students and employers alike. Follow-up studies were conducted in 2001 and 2002 to further explore the "other business skills" aspect of the IS curriculum (Turner & Lowry, 1999b; 2001; 2002; 2003). Tables 1 and 2 show comparative ratings of "hard" or traditional IS and business subjects and of "soft" skills by information systems and technology professionals and employers.

Comparative Ratings of Academic Subjects by IS Practitioners and Employers

Table 1 shows the mean and standard deviations of ratings of academic subjects and soft skills by (IS/IT) Professionals and IS/IT Employees.

The research instruments contained two sections pertaining to the academic preparation of graduates. These sections separately covered the technical areas of an IS business degree and the other academic areas that are not specific to IS. Subjects responded to each question using a seven point Likert scale.

Of the 14 subjects/skills that achieved a mean rating of 5.0 or more, the highest rating by both practitioners and employers was achieved by Communications & Report Writing—a soft skill. 11 technical subjects and 2 "other business subjects", management and business ethics, achieved mean ratings of 5.0 or more, consistent with earlier findings by the authors (Turner & Lowry, 1999b).

Comparative Importance of Soft Skills by IS/IT Professionals and Employers

A third section in the survey solicited rankings of the importance of a range of so-called "soft skills". The results are presented in Table 2, which shows the ratings by IS/IT Professionals and IS/IT Employees for soft business skills.

Table 2 shows a marked similarity between practitioners and employers in their ratings of the importance of soft business skills. Only the ability to prepare multimedia presentations failed to achieve a mean rating of 5.0. All other soft business skills were highly rated by both IS practitioners and employers.

Table 3. Extracted factors: Academic subjects

Rotated Component Matrix [a]

Data mining/Data warehousing	.765						
ERP implementations & operations	.735						
Project Management	.588						
Knowledge base/Expert systems	.526						
Apply 3GLs		.821					
Apply OOPs		.752					
CASE applications		.630					
Communications & Report Writing			.785				
Management			.728				
Organisational Behaviour			.643				
Accounting				.835			
Business Finance				.780			
Business Statistics				.599			
Foreign Languages					.824		
Psychology					.762		
International Business					.512		
Use operating systems						.730	
Knowledge of PC apps						.714	
LAN & Data Comms						.661	
Mathematical Modelling							.850
Operations Research							.657
Analysis & Design							.715
Database design							.614

Extraction Method: Principal Component Analysis.

a.

Table 4. Focus of academic subjects factors

Factor	Focus
hard1	enterprise software development and management subjects
hard2	software engineering and programming subjects
hard3	soft skills development subjects
hard4	more valued non-IS academic subjects
hard5	less valued non-IS academic subjects
hard6	IT networking and operations subjects
hard7	applied planning and modelling subjects
hard8	fundamental systems development subjects

Differences

Eight hard academic subject factors and four soft skills factors were identified through factor analysis. The academic subject factors accounted for 66% of the variance. The soft factors accounted for 58% of the variance extracted. These factors were used to establish composite variables and these composite variables were then used in comparing the two groups. Table 3 shows the academic subject factors.

The focus of each of the hard factors shown in Table 3 is characterised in Table 4.

Table 5 shows subject factors for the soft skills rated by IS/IT Professionals and Employers in Table 2.

The focus of each of the soft skill factors shown in Table 5 is characterized in Table 6.

Table 7 shows rankings of hard and soft skills factors by IS/IT Professionals and Employers.

Table 5. Extracted factors – Soft skills

Rotated Component Matrix [a]

Independently acquire new skills	.741			
Problem solving skills	.740			
Problem definition skills	.700			
Quickly acquire new skills	.692			
Work under pressure		.678		
Place organizational objectives first		.628		
Work independantly		.603		
Think creatively		.557		
Accept direction		.529		
Handle concurrent tasks		.514		
Oral presentation skills			.765	
Leadership potential			.651	
Able to prepare multimedia presentations			.554	
Written communication skills			.544	
Business analysis skills			.524	
Able to interact with people of different background				.877
Able to work with people from different disciplines				.813

Extraction Method: Principal Component Analysis.

a.

Table 6. Focus of Soft Skills Factors

Factor	Focus
soft1	intellectual skills
soft2	work environment skills
soft3	communication and public persona
soft4	effectiveness across disciplines and cultures

Ranking of Soft Skills by IS/IT Employers and Professionals

Table 8 shows the rankings of soft skills by IS/IT Employers and Professionals.

IS/IT Employers ranked the **soft4** factor, *effectiveness across disciplines and cultures* highest, followed by **soft2,** *work environment skills* and **soft1**, *intellectual skills*, rating **soft3**, *communication and public persona* last.

IS/IT Professionals ranked all factors higher than IS/IT employers. This group ranked factor **soft3**, *communication and public persona***,** first. Factor **soft1,** *intellectual skills*, was next, followed by **soft2,** *work environment skills*. The **soft4** factor, *effectiveness across disciplines and cultures*, received the lowest rating by this group.

IS/IT Employers and IS/IT Professionals ranked all four of the soft factors differently. Both groups ranked factor **soft4,** *effectiveness across disciplines and cultures* around the same, middle level, but whilst **soft4** was ranked **highest** by IS/IT Employers, it was ranked **lowest** by IS/IT Professionals. The inverse order of soft skills rankings by the two groups is cause for concern by all stakeholder groups.

Table 8. Ranking of soft skills by IS/IT employers and professionals

Factor	Factor Name	Employers Ranking	Professionals Ranking
soft1	intellectual skills	3	2
soft2	work environment skills	2	3
soft3	communication and public persona	4	1
soft4	effectiveness across disciplines and cultures	1	4

Table 9. Ranking of hard skills by IS/IT employers and professionals

Factor	Factor Name	Employers Ranking	Professionals Ranking
hard1	enterprise software development and management subjects	3	6
hard2	software engineering and programming subjects	8	1
hard3	soft skills development subjects	4	5
hard4	more valued non-IS academic subjects	1	8
hard5	less valued non-IS academic subjects	6	3
hard6	IT networking and operations subjects	2	7
hard7	applied planning and modelling subjects	5	4
hard8	fundamental systems development subjects	7	2

Table 10. Hard and soft skills differences

Test Statistics [a]

				Asymp. Sig.
soft1	8271.000	17862.000	-1.720	.085
soft2	8619.000	18210.000	-1.173	.241
soft3	7532.500	17123.500	-2.836	.005
soft4	9292.000	18608.000	-.147	.883
hard1	9197.000	18788.000	-.286	.775
hard2	6699.500	16290.500	-4.116	.000
hard3	8859.000	18450.000	-.808	.419
hard4	7722.000	17038.000	-2.553	.011
hard5	8217.500	17808.500	-1.796	.073
hard6	8703.000	18019.000	-1.046	.295
hard7	8748.000	18339.000	-.984	.325
hard8	7558.000	17149.000	-2.834	.005

a.

Ranking of Hard Skills by IS/IT Employers and Professionals

Once again, IS/IT Employers and IS/IT Professionals ranked all four of the "hard" factors differently. Table 9 shows the rankings of hard skills by IS/IT Employers and Professionals.

Employers ranked **hard4,** *more valued non-is academic subjects*, first. This appears to be consistent with the value placed by employers on the **soft4** factor, *effectiveness across disciplines and cultures***.** As with soft skills, rankings of hard skill factors by employers and professionals are inverse images of one another throughout.

IS/IT Employers ranked the **hard6** factor, *IT networking and operations* ***subjects***, second. This was followed by

hard1, *enterprise software development and management subjects*, **hard3,** *soft skills development subjects*, **hard7**, *applied planning and modelling subjects*, **hard5**, *less valued non-is academic subjects*, **hard8**, *fundamental systems development subjects*, with **hard2**, *software engineering and programming subjects*, rated last.

IS/IT Professionals ranked the **hard2,** *software engineering and programming* subjects factor first, followed by **hard8**, *fundamental systems development subjects*, **hard5**, *less valued non-is academic subjects*, **hard7**, *applied planning and modelling subjects*, **hard3**, *soft skills development subjects*, **hard1**, *enterprise software development and management subjects*, **hard6**, *IT networking and operations subjects*, and **hard4**, *more valued non-IS academic subjects*, rated lowest by this group. As with the soft skills ratings, the inverse order of hard skills rankings by the two

Table 11. Issues and Recommendations for IS Curriculum Refinement

Issue	Recommendation
Confusion: Research results that call for more "business skills" have handily and traditionally been interpreted as meaning exposure of students to additional, formal business subjects. While an IS student may well gain knowledge and skill in marketing, economic analysis, or international business in that way, our findings suggest that it is the "soft" skills, rather than formal academic skill, wanted by IS/IT practitioners and employers.	**Reducing confusion:** Local course advisory and professional bodies can provide invaluable insight into the mix of technical and non-technical formal courses and "soft" skills appropriate for a given institution's service area. Focus groups and local replication of available studies should provide targeted, timely, and authoritative guidance for ongoing curriculum evolution.
Tradition and inertia - Content: It is easy and tempting to offer traditional business subjects as they are already being taught anyway. In many institutions, the existing IS academic staff would have to acquire the additional academic background and skills needed to introduce a substantive "soft skills" emphasis into the IS curriculum. In most instances, there may be insufficient time or interest to do so. **Tradition and inertia – Delivery Method:** Many, if not most, subjects are taught in a familiar lecture / practical mode. As a project and team-oriented profession, information systems programs may well achieve a better match between the workplace through the "study place" through the adoption of active-learning, student-centered delivery methods such as Problem-Based Learning (PBL) and Work-Integrated Learning (WIL). (Bentley, Sandy, & Lowry, 2002)	**Overcoming tradition and inertia:** To some extent, prospective students have taken matters into their own hands by opting in larger numbers to bypass traditional university courses in favour of industry-sponsored/sanctioned entry gateways such as those offered by Microsoft, Oracle, SAP, amongst others. It is possible that students electing the non-academic alternative see more value in industry-focused training than in academic education, which they may see as irrelevant to their career aspirations. A large number of traditional business subjects were rated low in importance by IS practitioners and employers in the tables above. Tertiary educators need to reconsider the value of these traditional business subjects and will have to develop and deliver revitalised curricula that squarely address the expressed desire for "soft" skills that have been identified in a number of studies.
Resources: If a substantial portion of an undergraduate degree program was shifted from traditional business subjects to the acquisition and development of "soft skills", who would develop, teach, and assess the new "soft skills" curriculum component?	**Finding resources:** Resources are easier to obtain from a body of satisfied clients, such as the firms who employ our graduates. If we are seen to consult with, listen to, and serve the interests of those firms, they will follow their self-interest and become rich sources of guidance in curriculum planning and development, work experience for students, consultancies for academics, equipment, money, political weight in our own institution.
We must learn to master what we teach about building client ownership to enlist influential industry partners.	
Vested interest: Some academic institutions supplement the enrolment in less relevant or popular subjects through inclusion of those subjects in a popular curriculum such as Information Systems. In many institutions, economic incentives exist for students to be enrolled in subjects within a single administrative unit, such as a business or IT faculty.	**Neutralising vested interests:** University senior managers may oppose substantive IS curriculum reform such as that discussed in this chapter for a number of reasons, including loss of revenue if students enroll in "soft" skill courses provided by another administrative unit. It is up to IS educators to develop strategies to address the turf issues that preoccupy some administrators. Building effective partnerships with the right industries and accrediting bodies in our service area can provide a powerful voice to speak on our behalf to senior management.

groups is cause for concern by all stakeholder groups. These findings are consistent with findings from an earlier study by the authors (Turner & Lowry, 1999b).

Analysis of the differences in ratings between the two groups, shown in **Table 10**, indicates significant differences between four factors, **soft3,** *communication and public persona***, hard2,** *software engineering and programming subjects,* **hard4,** *more valued non-IS academic subjects* and **hard8**, *fundamental systems development subjects*. Professionals rated **soft3** highest whilst Employers rated **soft3** lowest. Employers rated **hard2** eighth, whilst Professionals rated **hard2** first**.** The **rankings** were reversed for **hard4** and Employers rated **hard8** seventh and Professionals rated this second. As no other differences between groups were significant, stakeholders may more productively focus on understanding and reducing the gaps in rankings between stakeholder groups.

FUTURE TRENDS

Barriers to meaningful reform and evolution of the IS curriculum include confusion, tradition and inertia, scarcity of resources, and vested interests. Table 11 summarizes these issues and offers recommendations, in increasing order of difficulty.

CONCLUSION

There has long been agreement that the IS curriculum should be comprised of some combination of technical subjects and nontechnical business subjects, and that graduates also need soft business skills. There is far less agreement about what the mix between these should be and how best to prepare students in some areas, notably in the development of soft business skills.

While we agree with the general view that soft skills have become increasingly important, we argue that the traditional business subjects are *not* the business skills primarily sought in studies of the IS marketplace. Does the study of traditional business subjects such as marketing, business law, or economics directly help the students to develop a repertoire of soft business skills? The findings suggest that in reality it is not more core business subjects that are needed but an appreciation of business processes and activities that are not always covered in IS degree programs.

Some formidable barriers exist to substantive revision of IS curricula to emphasise acquisition and development of soft business skills. In increasing order of difficulty, Table 11 summarizes some of the issues and barriers to meaningful reform and evolution of the IS curriculum might be surmounted over time and with sufficient dedication. There are, of course, no easy solutions to resolve these issues.

While study after study has called for soft skills acquisition and development by IS students, some IS programs have a clearer and better-developed vision than others of what those skills are and how they may be introduced and cultivated. The growing emphasis on soft skills in IS education is an indication that what began as a fundamentally technology-oriented discipline is, indeed, evolving into a technology-based profession. We can watch someone else claim that knowledge and the opportunities that it offers or we can embrace it ourselves, hoping that there is still time.

REFERENCES

ACM-AIS (2002). *IS 2002 - Model curricula and guidelines for undergraduate degree programs in information systems*. Association for computing machinery association for information systems. Retrieved June 15, 2008, from http://www.acm.org/education/curricula.html#IS2002

Armarego, J. (2005). Educating agents of change. In *Proceedings of the CSEE&T2005 18th Conference on Software Engineering Education and Training*, Ottawa.

Beachboard, J. C. & Parker, K. R. (2003). How much is enough? Teaching information technology in a business-oriented IS curriculum. In *Proceedings of the Ninth Americas Conference on Information Systems*, Tampa, FL.

Beise, C., Niederman, F., Quan, J., & Moody, J. (2005). Revisiting global information systems management education. *Communications of the Association for Information Systems, 16*, 625-641.

Bentley, J., Sandy, G., & Lowry, G. (2002). Problem-based learning in information systems analysis and design. In E. Cohen (Ed.), *Challenges of information technology education in the 21st century* (pp. 100-123). Hershey, PA: Idea Group Publications.

Berghel, H. & Sallach, D. L. (2004). A paradigm shift in computing and IT education. *Communications of the ACM, 47*(6), 83-88.

Burn, J. M., Ng Tye, E. M. W., & Ma, L. C. K. (1995). Paradigm shift - Cultural implications for development of IS professionals. *Journal of Global Information Management, 3*(2), 18-28.

Cafasso, R. (1996). Selling your soft side helps IT. *Computerworld, 18*(35), 60-61.

Cheney, P., Hale, D., & Kasper, G. (1990). Knowledge, skills and abilities of information systems professionals: Past, present and future. *Information Management, 9*(4), 237-247.

Cohen, E. (2000). *Curriculum model 2000 of the information management association and the data administration mangers association*. Retrieved June 15, 2008, from http://www.irma-international.org/downloads/pdf/irma_dama.pdf

Davis, G., Gorgone, J. T., Feinstein, D. L., & Longenecker, H. E. (1997). *IS'97 model curricula and guidelines for undergraduate degree programs in information systems*: Association of Information Technology Professionals.

Dressler, S. & Keeling, A. E. (2004). Student benefits of cooperative education. In R.K. Coll & C. Eames (Eds.), *International handbook for cooperative education* (pp. 217-236). Hamilton, New Zealand: World Association for Cooperative Education.

Evans, N. (2003) Informing clients in education about instructional offerings and |careers in the ICT industry. In *Proceedings of the InSITE Conference "Where Parallels Intersect"*, Pori, Finland

Fincher, S., Clear, T., Petrova, K., Hoskyn, K., Birch, R., Claxton, G., et al. (2004). Cooperative education in information technology. In R. K. Coll & C. Eames (Eds.), I*nternational handbook for cooperative education: An international perspective of the theory, research and practice of work-integrated learning* (pp. 111-121). Hamilton, New Zealand: World Association for Cooperative Education.

Florida, R. (2005). *The flight of the creative class: the new global competition for talent*. New York: HarperCollins Publishers.

Gallivan, M. J., Truex, D. P., & Kvasny, L. (2004). Changing patterns in IT skill sets 1988-2003: A content analysis of classified advertising. *The Data Base for Advances in Information Systems, 35*(3), 64-87.

Gardiner, C. (2005, September 12 – October 9). Finding the right ICT skills. *Telecommunications Review, 31,* 32-34.

Gorgone, J. T. & Gray, P. (1999). *Graduate IS curriculum for the 21st century.* In Proceedings of the 32nd Hawaii International Conference on Systems Science, Maui, Hawaii. Retrieved June 16, 2008, from http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/proceedings/&toc=comp/proceedings/hicss/1999/0001/01/0001toc.xml

Holt, D., MacKay, D., & Smith, R. (2004). Developing professional expertise in the knowledge economy. *Asia-Pacific Journal of Cooperative Education, 5*(2), 1-11.

Lee, C. K. (2005). Transferability of skills over the ICT career path. In *Proceedings of the Annual Conference of Special Interest Group on Computer Personnel Research* (pp. 85-93). Atlanta, Georgia: Association for Computing Machinery.

Lee, D. M., Trauth, E. M., & Farwell, D. (1995). Critical skills and knowledge requirements of IS professionals: A joint academic/industry investigation. *MIS Quarterly, 19*(3), 313-340.

Lee, D. M. S. (2004). Organizational entry and transition from academic study: Examining a critical step in the professional development of young IS workers. In M. Igbaria & C. Shayo (Eds.), *Strategies for managing IS/IT personnel* (pp. 113-141). Hershey, PA: Idea Group.

Leong, K.-C. & Tan, M. T. K. (2004). The long road to being an IS professional: A newcomer perspective. In T. Leino, T. Saarinen & S. Klein (Eds.), *Proceedings of the Twelfth European Conference on Information Systems*. Turku, Finland: University of Turku, School of Economics and Business Administration.

Lidtke, D., Stokes, G., Haines, J., & Mulder, M. (1999). *ISCC '99: An information systems-centric curriculum'99: Guidelines for educating the next generation of information systems specialists*. Retrieved June 16, 2008, from http://www.iscc.unomaha.edu/

Litecky, C. R., Arnett, K. P., & Prabhakar, B. (2004). The paradox of soft skills vs. technical skills in IS hiring. *Journal of Computer Information Systems, 45*(1), 69-76.

Lowry, G. R., Morgan, G. W., & FitzGerald, D. G. (1996). Organizational characteristics, cultural qualities and excellence in leading Australian-owned information technology firms. In *Proceedings of the 1996 Information Systems Conference of New Zealand* (pp. 72-84). Palmerston North, New Zealand: IEEE Computer Society Press.

Lowry, G. & Turner, R. (2005a). Information systems education for the 21st century: Aligning curriculum content & delivery with the professional workplace. In D. Carbonara (Ed.), *Technology literacy applications in learning environments* (pp. 171-202). Hershey, PA: IRM Press.

Lowry, G. & Turner, R. (2005b). Softening the MIS curriculum for a technology-based profession. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 2539-2545). Hershey, PA: Information Science Publishing.

Lowry, G., Turner, R., & Fisher, J. (2006). The contribution of employment satisfaction factors to recruiting, retaining, and career development of information systems and technology professionals. *The Review of Business Information Systems, 10*(1), 137-150.

Lyytinen, K. & King, J. (2004). Nothing at the center? Academic legitimacy in the information systems field. *Journal of the Association for Information Systems, 5*(6). Retrieved June 16, 2008, from http://jais.isworld.org/articles/5-8/

Maiden, S. (2004, June 10). Graduates failing the Uni of Life. *The Australian*.

Medlin, B. D. (2004). Skills crucial to the information technology professionals in the global business environment: An empirical study in the United States. *International Journal of Human Resources Development and Management, 4* (2).

Morgan, G. W., Lowry, G. R., & FitzGerald, D. G. (1998). Development staff characteristics and service stability in leading Australian-owned information technology firms. In *Proceedings of the 1998 International Conference on Software Engineering: Education and Practice* (pp. 96-103). Dunedin, New Zealand: IEEE Computer Society Press.

Mulder, M. & van Weert, T. (2000). *Informatics curriculum framework 2000 for higher education* (ICF-2000). Paris: UNESCO.

Schwarzkopf, A. B., Saunders, C., Jasperson, J., & Croes, H. (2004). Strategies for managing IS personnel: IT skills staffing. In M. Igbaria & C. Shayo (Eds.), *Strategies for managing IS/IT personnel* (pp. 143-164). Hershey, PA: Idea Group Publishing.

Trauth, E., Farwell, D., & Lee, D. (1993). The IS expectation gap: Industry expectations versus academic preparation. *MIS Quarterly, 17*(3), 293-307.

Trauth, E. M., Huang, H., Morgan, A. J., Quesenberry, J. L., & Yeo, B. (2006). Investigating the existence and value of diversity in the global IT workforce: An analytical framework. In F. Niederman & T. Ferratt (Eds.), *IT workers: Human capital issues in a knowledge-based environment* (pp. 331-360). Greenwich, CT: Information Age Publishing.

Turner, R. & Lowry, G. (1999a). The complete graduate: What students think employers want and what employers say they want in new graduates. In S. Lee (Ed.), *Preparing for the global economy of the new millennium.* In *Proceedings of Pan-Pacific Conference XVI* (pp. 272-274). Fiji: Pan-Pacific Business Association.

Turner, R. & Lowry, G. (1999b). Reconciling the needs of new information systems graduates and their employers in small, developed countries. *South African Computer Journal, 24*, 136-145.

Turner, R. & Lowry, G. (2000). Motivating and recruiting intending IS professionals: A study of what attracts IS students to prospective employment. *South African Computer Journal*, *26*(4), 132-137.

Turner, R. & Lowry, G. (2001). What attracts IS students to prospective employment: A study of students from three universities. *Managing information technology in a global economy* (pp. 448-452). Hershey, PA: IRMA.

Turner, R. & Lowry, G. (2002). The relative importance of "hard" & "soft" skills for IT practitioners. In M. Khrosrow-Pour (Ed.), *Issues and trends of information technology management in contemporary organizations* (pp. 1-10). Seattle,WA: IRMA.

Turner, R. & Lowry, G. (2003). Education for a technology-based profession: Softening the information systems curriculum. In T. McGill (Ed.), *Issues in information systems education* (pp. 156-175). Hershey, PA: Idea Group Publishing.

Turner, R., Fisher, J., & Lowry, G. (2004a). Describing the IS professional: A structural model. In *Proceedings of the Eighth Pacific-Asia Conference on Information Systems*, Shanghai, China.

Turner, R., Fisher, J., & Lowry, G. (2004b). A structural model of the information systems professional: Comparing practitioners, employers, students, and academics. In *Proceedings of the International Federation for Information Processing Conference*, Melbourne, Australia.

Turner, R., Fisher, J., & Lowry, G. (2005a). Age related variation in a two stage regression model describing the IS professional. In *Proceedings of the XXII Pan-Pacific Conference*, Shanghai, China.

Turner, R., Fisher, J., & Lowry, G. (2005b). Gender variations in a structural model of the information systems professional. In *Proceedings of the Ninth Pacific Asia Conference on Information Systems*, Bangkok, Thailand.

Turner, R., Lowry, G., & Fisher, J. (2005). A structural model of the information systems professional: Comparing practitioners, employers, students, and academics. In T. V. Weert & A. Tatnall (Eds.), *Information and communication technologies and real-life learning: New education for the knowledge society* (Vol. ISBN 0-387-25996-3, pp. 243-254). New York: Springer.

Underwood, A. (1997). *The ACS core body of knowledge for information technology professionals*. Retrieved June 16, 2008, from http://www.acs.org.au/index.cfm?action=show&conID=200509022309270170

KEY TERMS

Hard Skills: Measurable capabilities and academic knowledge acquired through traditional tertiary study. Current MIS curriculum examples such communications and report writing, systems analysis and design, client/server applications, and business applications, are shown in rank-order in Table 1**.**

Inquiry-Based Learning: A student-centered, active learning approach focusing on questioning, critical thinking, and problem-solving. IBL is expressed by the idea "involve me and I understand". The IBL approach is more focused on using and learning content as a means to develop information-processing and problem-solving skills. The system is more student-centered, with the teacher as a facilitator of learning. There is more emphasis on "how we come to know" and less on "what we know". Students are involved in the construction of knowledge through active involvement. The more interested and engaged students are by a subject or project, the easier it will be for them to construct in-depth knowledge of it. Learning becomes easier when reflects their interests and goals and piques their natural curiosity.

Problem-Based Learning: Is an active learning strategy that may be suitable for better preparing information systems students for professional practice. In the problem-based approach, complex, real world problems or cases are used to motivate students to identify and research concepts and principles they need to know in order to progress through the problems. Students work in small learning teams, bringing together collective skill at acquiring, communicating, and integrating information in a process that resembles that of inquiry.

Project-Based Learning: An active learning approach that focuses on developing a product or creation. The project may or may not be student-centered, problem-based, or inquiry-based. Project-based learning uses open-ended assignments that provides students with a degree of choice, and extends over a considerable period of time. Teachers act as facilitator, designing activities and providing resources and advice to students. Instruction and facilitation are guided by a broad range of teaching goals. Students collect and analyze information, make discoveries, and report their results. Projects are often interdisciplinary.

Soft Skills: Cultivated elements of professionalism that derive from example, reflection, imitation, and refinement of attitudes, personal capabilities, work habits, and interpersonal skills and are expressed in consistent and superior performance, characterized by a customer service and team orientation. Current MIS curriculum examples such as ability to work as a member of a team, well-developed oral and written presentation skills, and the ability to work independently, are shown in rank-order in Table 2.

Student-Centered/Active Learning: Places the student into active, self-directed learning, learning by enquiry and ownership of the learning goals. Active-learning strategies include Problem-Based Learning (PBL), Project-Based Learning; and Inquiry-Based Learning, and Work-Integrated Learning.

Teacher-Centered Learning: Is characterized by didactic teaching, passive learning, and the teacher as the "expert". Control of learning rests with the instructor. The learning of the students is directed by the instructor and is often based on what the instructor believes the student needs to learn. This approach can lead to students focusing on determining what knowledge they require to pass the subject rather than the instructional objectives of the program, a phenomenon well known to educators.

Work-Integrated Learning: WIL is a hybrid approach that achieves learning outcomes through a combination of alternating periods of traditional academic pedagogy with extended periods of practical experience in the professional workplace. Work-Integrated Learning is a mature pedagogical strategy that is often referred to as "sandwich" and "end-on" courses.

ENDNOTE

[a] The term "information systems" will be used to mean "business information systems", "management information systems", "informatics", and similar variations throughout this entry.

Reconfigurable Computing Technologies Overview

Kai-Jung Shih
National Chung Cheng University, ROC

Pao-Ann Hsiung
National Chung Cheng University, ROC

INTRODUCTION

Reconfigurable computing is breaking down the barrier between hardware and software design technologies. The segregation between the two has become more and more fuzzy because reconfigurable computing has now made it possible for hardware to be programmed and software to be synthesized. Reconfigurable computing can also be viewed as a trade-off between general-purpose computing and application specific design. Given the architecture and design flexibility, reconfigurable computing has catalyzed the progress in hardware-software codesign technology and a vast number of application areas such as scientific computing, biological computing, artificial intelligence, signal processing, security computing, and control-oriented design, to name a few.

In this article, we briefly introduce why and what is reconfigurable computing in the introduction section. Then, the resulting enhancements of hardware-software codesign methods and the techniques, tools, platforms, design and verification methodologies of reconfigurable computing will be introduced in the background section. Furthermore, we will introduce and compare some reconfigurable computing architectures. Finally, the future trends and conclusions will also be given. This article is aimed at widespread audiences, including both a person not particularly well grounded in computer architecture and a technical person.

Why Reconfigurable Computing?

With the popularization of the use of computers, computer-aided computing can be roughly divided into two technical areas, one of which is general-purpose computing and the other is application-specific integrated circuit (ASIC) computing.

On one extreme, general-purpose computing was accomplished by the world's first fully operational electronic general-purpose computer, called *Electronic Numerical Integrator and Calculator* (ENIAC), built by J. Presper Eckert and John Mauchly. But it is well-known as *von Neumann computer* because ENIAC was improved by John von Neumann (Hennessy & Patterson, 2007). A general-purpose computer is a single common piece of silicon, called a *microprocessor*, that could be programmed to solve any computing task. This means many applications could share commodity economics for the production of a single integrated circuit (IC). This computing architecture has the flexibility and superiority that the original builders of the IC never conceived (Tanner Research, 2007).

On the other extreme, an ASIC is an IC specifically designed to provide unique functions. ASIC chips can replace general-purpose commercial logic chips, and integrate several functions or logic control blocks into one single chip, lowering manufacturing cost and simplifying circuit board design. Although the ASIC has the high performance and low power advantages, its fixed resource and algorithm architecture result in drawbacks such as high cost and poor flexibility.

As a tradeoff between the two extreme characteristics, reconfigurable computing has combined the advantages of both general-purpose computing and ASIC computing. A comparison among the different architecture characteristics is illustrated in Table 1 (Tredennick, 1996; Tessier & Burleson, 2001).

From Table 1, we observe that reconfigurable computing has the advantage of programmable or configurable computing resources, called configware (TU Kaiserslautern, 2007a), as well as configurable algorithms, called flowware (Hartenstein, 2006; TU Kaiserslautern, 2007b). Further, the performance of reconfigurable systems is better than general-purpose systems and the cost is smaller than that of ASICs. The main advantage of reconfigurable system is its high flexibility, while its main disadvantage is its high power consumption. The design effort in terms of nonrecurring engineering (NRE) cost is between that of general-purpose processor and ASICs.

Because reconfigurations of underlying resources help achieve the goals of balance among performance, cost, power, flexibility, and design effort. The reconfigurable computing architecture has enhanced the performances of large variety of applications, including embedded systems, SoCs, digital signal processing, image processing, network

Table 1. Comparison of representative computing architecture

Computing Architecture	Programming source		Advantage				
	Resources	**Algorithms**	**Performance**	**Cost**	**Power**	**Flexibility**	**Design effort (NRE)**
General-purpose	Fixed	Software	Low	Low	Medium	High	Low
ASIC	Fixed	Fixed	High	High	Low	Low	High
Reconfigurable	Configware	Flowware	Medium	Medium	High	High	Medium

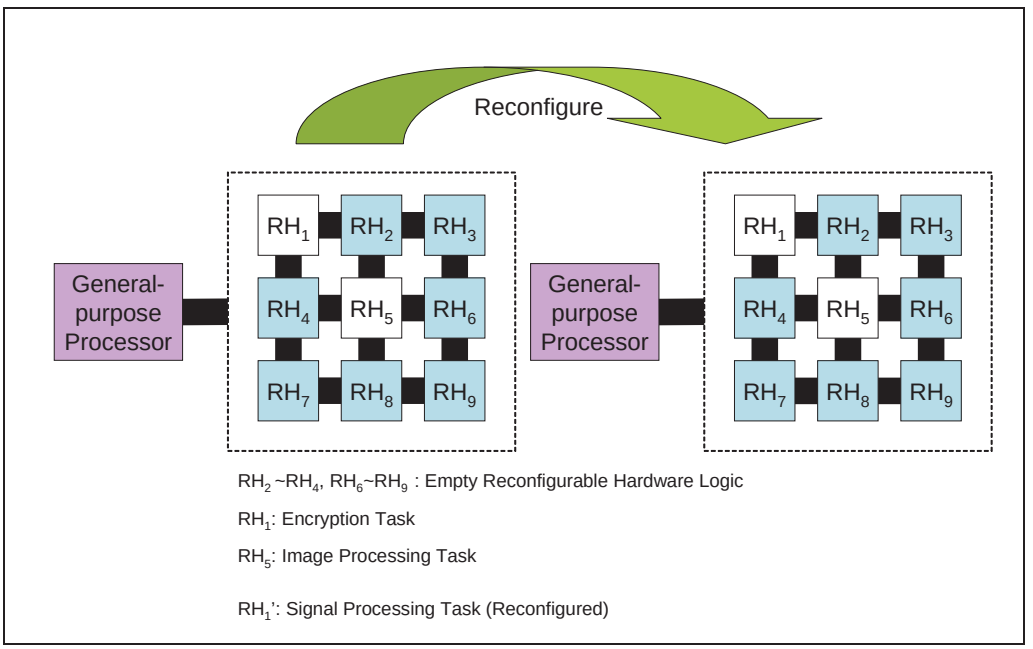
security, bioinformatics, supercomputing, boolean SATisfiability (SAT), spacecrafts, and military applications. We can say that reconfigurable computing will widely, pervasively, and gradually impact human lives.

What is Reconfigurable Computing?

In 1960, Estrin (1960) first proposed the term "reconfigurable computing." The reconfigurable computing architecture is composed of a general-purpose processor and reconfigurable hardware logic. The reconfigurable computing architecture can be concisely defined as *Hardware-On-Demand* ™ (Schewel, 1998), *general purpose custom hardware* (Goldstein et al., 2000) or a *hybrid approach between ASICs and general-purpose processors* (Singh et al. 2000).

We illustrate a general reconfigurable computing architecture in Figure 1. In this architecture, the reconfigurable hardware logic executes application-specific computation intensive task, such as encryption (RH_1) and image processing (RH_5) as shown in Figure 1. The processor is used to

Figure 1. Reconfigurable computing



control the behavior of the task running in the reconfigurable hardware and some other functions such as external communications. When a reconfigurable hardware has finished its computation, such as the encryption task in RH_1, the processor reconfigures the hardware to execute another task such as the signal processing task. During this reconfiguration process, the image processing task continues to execute in RH_5 without interruption.

From the above illustration, we can also define reconfigurable computing as a discipline in which system or application functions can be changed by configuring a fixed set of logic resources through memory settings (Hsiung & Santambrogio, 2008). Functions may be transforms, filters, codec, and protocol, the fixed set of logic resources may be logic block, I/O block, routing block, memory block, and application-specific block, and the memory settings mean configuration bits.

BACKGROUND

Materials science and technology progress has resulted in the maturity and development of reconfigurable computing. To understand the development of reconfigurable computing, an important perspective is to view it from the transition of hardware-software (HW-SW) codesign technology to reconfigurable computing. In the following, we will first introduce reconfiguration techniques and the *Field-Programmable Gate Arrays* (FPGA) (Gokhale & Graham, 2005). Furthermore, we will introduce some reconfiguration tools and platforms used in the academia and the industry. Finally, the design and verification methodologies will also be introduced.

From Codesign to Reconfiguration

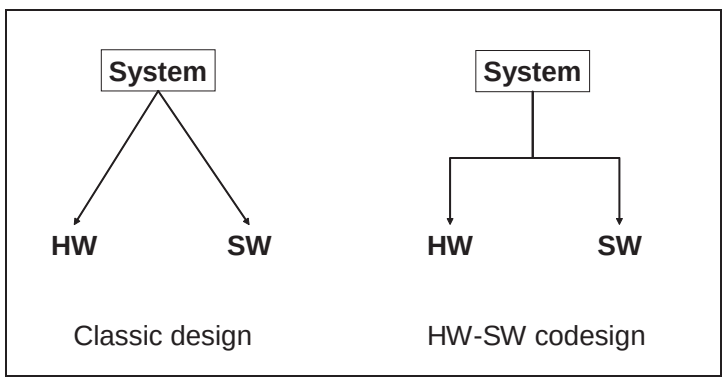
HW-SW *codesign* is an emerging topic that highlights a unified view of hardware and software (Vahid & Givargis, 2002). It is a system design methodology different from *classic design,* as illustrated in Figure 2.

The classic design partitions a system into hardware and software. Because a software designer may not know the final hardware architecture design and the hardware designer may also be unacquainted with the software design flow, hardware and software and are often implemented independently and then integrated toward the end of the design flow. If problems crop up during integration, changing either the hardware or the software could both be quite difficult. This will increase the maintenance difficulty and also delay the marketing time.

To address the problems mentioned above, a system design methodology called HW-SW codesign was proposed, which emphasizes the consistency and the integration between hardware and software. It is based on a system-level view, and thus eases both software verification and hardware error detection. The HW-SW codesign methodology reduces the cost of design and also shortens the time-to-market.

Nevertheless, the high cost in hardware design is a major issue of the HW-SW codesign flow because hardware must go through a time-consuming flow including design, debug, manufacturing, and test. The inconvenient hardware manufacturing forces designers to search for alternate ways. One way is to use modeling languages such as SystemC (Black & Donovan, 2004) to simulate hardware and software. Another method is to use concurrent process models to simulate hardware and software tasks. However, simulation speed is a major bottleneck (Schewel, 1998). To overcome the drawback, prototyping using reconfigurable architectures has

Figure 2. Classic design and HW-SW codesign



become the most appropriate choice for HW-SW codesign. In contrast to ASIC design, reconfigurable hardware can be much easily used to design hardware prototypes that can be integrated with software to obtain performance and functional analysis results much more efficiently and accurately. We can thus say that reconfigurable computing has accelerated and enhanced the HW-SW codesign flow.

Reconfiguration Techniques

From the mid 1980s, reconfigurable computing has become a popular field due to the FPGA technology progress. An FPGA is a semiconductor device containing programmable logic components and programmable interconnects (Compton & Hauck, 2002) but no instruction fetch at run time, that is, FPGA do not have a program counter (Hartenstein, 2006). In most FPGAs, the logic components can be programmed to duplicate the functionality of basic logic gates or functional intellectual properties (IPs) and also include memory elements composed of simple flip-flops or more complete blocks of memories (Barr, 1998).

Besides FPGA, the *reconfigurable data-path array* (rDPA) is another reconfiguration technique. In contrast to FPGA having single bit programmable logic blocks, rDPAs have *multiple* bits wide (e.g., 32 bit path width) reconfigurable data-path units (rDPUs). An rDPA is structurally programmed from configware sources, compiled into pipe networks to be mapped onto the rDPA. The term *reconfigurable data-path array*, or rDPA, had been proposed by Rainer Kress in 1993 at TU Kaiserslautern (Hartenstein, 2006). For further details on the comparison between FPGA and rDPA, readers can refer to the section on "Fine-grained vs. Coarse-grained Reconfiguration."

The main part of a reconfigurable system is the configware such as FPGA or rDPA. Besides configware, the software is another essential part that can control and thus incorporate the configware into a reconfigurable system. Although configware can provide resources for high performance computation, complex control must be implemented in software. Reconfiguring hardware implies software must also be appropriately reconfigured, and thus we need reconfigurable software design too (Compton & Hauck, 2002; Voros & Masselos, 2005).

Reconfiguration Tools and Platforms

To construct a reconfigurable computing system, designers need computer-aided design (CAD) tools for system design and implementation, such as a design analysis tool for architecture design, a synthesis tool for hardware construction, a simulator for hardware behavior simulation, and a placement and routing tool for circuit layout. We may build these tools ourselves or we can also use commercial tools and platforms for reconfigurable system design, such as the *Embedded Development Kit* (EDK) from Xilinx, which is a common development tool. The EDK integrates both the software and

Table 2. Commercial reconfiguration tools

Functionality	Tool Name	FPGA/EDA Company
Design Analysis	PlanAhead	Xilinx
FPGA Suite Tools	ISE Foundation	Xilinx
	Quartus	Altera
	FPGA Advantage	Mentor Graphics
FPGA Synthesizer	Synplify Pro	Synplicity
	FPGA Compiler	Synopsys
	Leonardo Spectrum	Mentor Graphics
	Precision Synthesis	Mentor Graphics
Simulator	ModelSim	Mentor Graphics
	NC SIM	Cadence
	Scirocco Simulator	Cadence
	Spexsim	Verisity
	VCS	Synopsys
	Verilog-XL	Cadence

the hardware components of a design to develop complete systems (Donato et al., 2005). In fact, EDK provides developers with a rich set of design tools, such as Xilinx Platform Studio (XPS), gcc, and Xilinx Synthesizer (XST). It also provides a wide selection of standard peripherals required to build systems with embedded processors, like MicroBlaze or IBM PowerPC (Xilinx, 2007). Besides Xilinx EDK, we list commonly used commercial FPGA and electronic design automation (EDA) tools in Table 2.

After designers build a reconfigurable system, a platform for operating and testing is needed. We can use the platforms developed in the industry or in the academia, such as the Caronte Architecture (Donato et al., 2005) and the Kress-Kung Machine (Hartenstein, 2006). The Caronte Architecuture is entirely implemented in the FPGA device and constituted by several elements such as a processor, memories, a set of reconfigurable devices and a reconfiguring device. It implements a module-based system approach based on an EDK system description and provides a low cost approach to the dynamic reconfiguration problem. The Kress-Kung is a data-stream-based machine. Instead of rDPAs, it has no *Central Processing Unit* (CPU) or program counter.

Design and Verification Methodologies

To design reconfigurable computing systems, we need some appreciation of the different costs and opportunities inherent in reconfigurable architectures. Currently, most systems are designed based on our past experiences. We can use the design patterns identified and cataloged by DeHone et al. (DeHon et al., 2004). Each pattern description has a name, intent, motivation, applicability, participants, consequences, implementation, known uses, and related patterns. They also cataloged the design patterns into several classification types, such as patterns for area-time tradeoffs and patterns for expressing parallelism. This classification is a good start for constructing reconfigurable systems. In the following, we present a typical design methodology and a typical verification methodology for illustration purpose.

Tseng and Hsiung (2005) proposed a UML-based design flow for *Dynamically Reconfigurable Computing Systems* (DRCS). This design flow is targeted at the execution speedup of functional algorithms in DRCS and at the reduction of the complexity and time-consuming efforts in designing DRCS. The most notable feature of the design flow is a HW-SW partitioning methodology based on the UML 2.0 sequence diagram, called *Dynamic Bitstream Partitioning on Sequence Diagram* (DBPSD). In DBPSD, partitioning guidelines are also included to help designers make prudent partitioning decisions at the class method granularity. The enhanced sequence diagram in UML 2.0 is capable of modeling complex control flows, and thus the partitioning can be done efficiently on the sequence diagrams.

After design and implementation, we need to verify that the system design is correct and complete. Correctness means that the design implements its specification accurately. Completeness means that our specification described appropriate output responses to all relevant input sequences (Vahid & Givargis, 2002).

Hsiung, Huang, and Liao (2006) proposed a SystemC-based performance evaluation framework, called *Perfecto*, for dynamically partially reconfigurable systems, which is an easy-to-use system-level framework. Perfecto is able to perform rapid explorations of different reconfiguration alternatives and to detect system performance bottlenecks. In their framework, a system designer can detect performance bottlenecks, functional errors, architecture defects, and other system faults at a very early design phase.

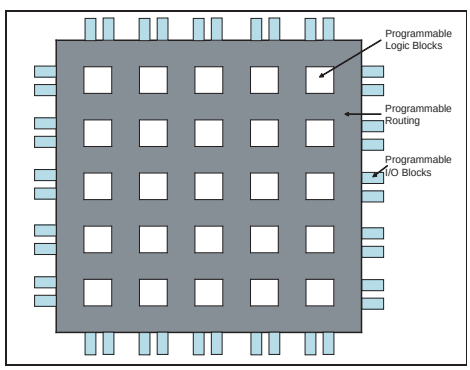
RECONFIGURATION ARCHITECTURES

As illustrated in Figure 3, FPGA is constructed from a large number of programmable logic structures, called programmable logic blocks, which can be interconnected to each other through programmable routing resources. If we want to connect the programmable logic blocks to the external, we can also interconnect them to the programmable I/O blocks through the programmable routing resources (Gokhale & Graham, 2005).

The programmable logic block can be configured as the desired circuit functionality and the programmable I/O blocks can be configured for communicating with outer devices. Between programmable logic blocks or between programmable logic blocks and I/O blocks, programmable routing can be configured for interconnection (Gokhale & Graham, 2005).

In this section, we will go through different classification schemes of reconfigurable architectures. In terms of reconfiguration granularity, we have fine-grained and

Figure 3. A generic FPGA architecture



coarse-grained reconfiguration. If we consider the time when reconfigurations are performed, we have static and dynamic reconfigurations. Considering the amount of logic resources that can be reconfigured, we have full and partial reconfigurations. Considering the model of reconfiguration supported, we will introduce column-based and tile-based reconfiguration. Finally, we give a comparison among different reconfiguration architectures.

Fine-Grained vs. Coarse-Grained Reconfiguration

To understand what fine-grained and coarse-grained mean, we can refer to Table 3 as formulated in Hartenstein (2006). In Table 3, the data path width indicates the granularity of the configware, that is, fine-grained or coarse-grained. The data path of FPGA is about 1 bit wide and rDPA is about 32 bits, so we call FPGA fine-grained and rDPA coarse-grained.

Besides the difference in data path width, the reconfigurable unit is also different. Fine-grained reconfiguration uses look-up table (LUT) as the typical reconfigurable unit and coarse-grained uses ALU-like unit. LUT is the most common implementation method for combinational logic. The characteristic of LUT is to use a multiplexer to select the input data. For example, a four inputs and one output multiplexer, denoted as a 4x1 multiplexer, is used as an LUT, which is shown in Figure 4.

In Figure 4, the selector will select which input will be connected to the output. For example, if the A=1 and B=0, the output port will be connected to input port I_{10} and output data is 0. For the example in Figure 4, the LUT represents a logic AND gate. When AB = "00", "01", "10", "11", the output will be connected to corresponded input I_{00}, I_{01}, I_{10}, I_{11}, respectively, and result in "0", "0", "0", "1". These are the values in a logic AND truth table. Based on this characteristic, a LUT can reserve, that is, reconfigure, the input to obtain any logic circuit.

The rDPA is an array of reconfigurable data unit (rDPU) and can be illustrated as in Figure 5. A typical example of rDPU is an arithmetic-logic unit (ALU) which is found in a von Neumann computer. The ALU is a digital circuit that calculates arithmetic and logic operations. The rDPUs constitute the set of coarse-grained programmable logic blocks.

Figure 4. A two-input look-up table

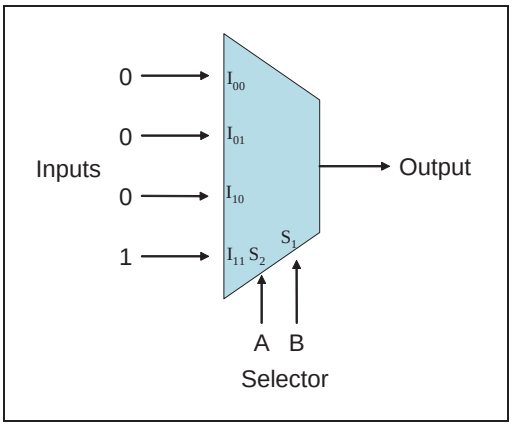
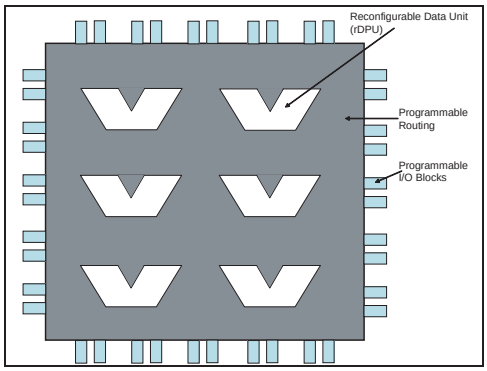


Table 3. Comparison of reconfiguration granularities

	Fine-grained	**Coarse-grained**
Configware	Field-programmable gate array (FPGA)	Reconfigurable data-path array (rDPA)
Data path width	~ 1 bit	~ 32 bits
Physical level of basic reconfigurable units	Gate level	RT level
Typical reconfigurable units examples	LUT (look-up table)	ALU-like
Configuration time	Milliseconds	Microseconds
Clock cycle time	~ 0.5 GHz	~1 - 3 GHz

Figure 5. Reconfigurable data array



Static vs. Dynamic Reconfiguration

As we mentioned earlier, traditional configwares can be configured for a hardware design as required. If we want to replace a new hardware design in the configware, the configware needs to be sent the RESET signal for reconfiguration use. Because the reset action usually consumes a lot of time, we would want to reduce the number of times this action is taken. In other words, we can say that traditional configwares are reconfigurable, but not run-time reconfigurable (Barr, 1998). We can classify these configwares as *static reconfiguration* unit.

Besides static reconfiguration, dynamic reconfiguration has also resulted from progress in new reconfiguration technologies. Run-time reconfiguration has added another dimension of flexibility in such systems. When we want to place a hardware design into a dynamically reconfigurable configware, we can just stop the clock of the region we need, reset the hardware resources in that region, and then configure the desired the hardware design and start the clock for this region. The other regions on this configware will still work unrestrictedly. Thus, it is called *dynamic* or *run-time reconfiguration*. Dynamic reconfiguration technique reduces the response time and the configuration overhead compared to static reconfiguration. Nowadays, industry products, such as Xilinx Virtex-II, Virtex-II Pro, Virtex-4, and later versions of the Virtex series all support dynamic reconfiguration.

Full vs. Partial Reconfiguration

In traditional reconfiguration, we integrate a set of one or more hardware design and configure them using a single reconfiguration action into a configware. We can call this type of reconfiguration as *full reconfiguration*. If a hardware design uses much fewer resources than that available in the configware, full reconfiguration will result in low resource utilization.

If the configware can be configured partially, resource utilization can be increased and portions of the chip can be reconfigured while other parts can still continue running and computing.

Partial reconfiguration is illustrated in Figure 6, where the resource usage of hardware design A is smaller than that available in the configware. The configware is partially configured with the bitstream of design A. When another hardware design B is required, the configware will again be partially configured with hardware design B. During the configuration of B, A will run without any glitches. Thus, the partial reconfiguration technology results in better configware resource utilization than that in full reconfiguration.

Currently, the dynamically reconfigurable architectures also support partial reconfiguration, including Xilinx Virtex-II, Virtex-II Pro, Virtex-4, and the later versions of the Virtex series.

Column-Based vs. Tile-Based Reconfiguration

Traditionally, configware was designed for column-based reconfiguration, that is, the basic unit of configuration is a *column* that crosses the chip. Thus, in implementing a partially reconfigurable system, the configware is modeled as a one-dimensional area of resources in which hardware designs can be configured. A typical example of configware with column-based reconfiguration is the Xilinx Virtex II series

Figure 6. An example of partial reconfiguration

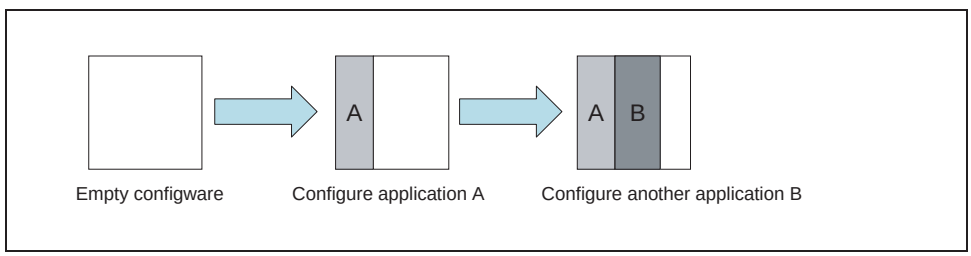
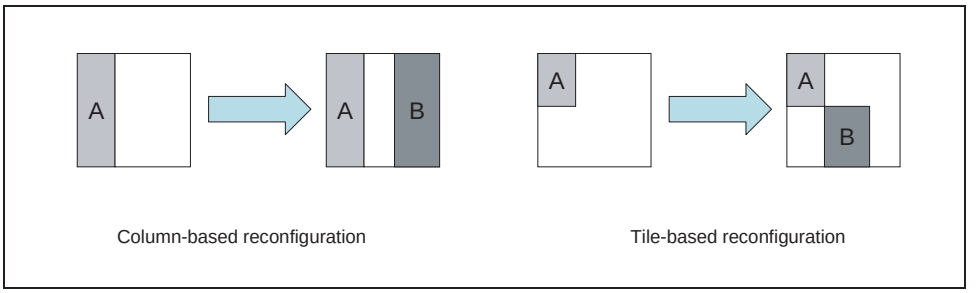


Figure 7. Column-based vs. tile-based reconfiguration



of FPGA chips. Tile-based reconfiguration is another more flexible architecture that can be found in the Xilinx Virtex 4 series of FPGA chips. These two types of reconfiguration are illustrated in Figure 7.

A *tile* is the area smaller than a column. In the Xilinx Virtex 4 series of FPGA chips, each column is partitioned into two tiles. Thus, designers can have the two-dimensional vision of their reconfigurable system. As shown in Figure 7, if the sizes of hardware designs A and B are the same in the column-based and the tile-based instances, then they can be configured with more flexibility with the tile-based one. The resource utilization of the configware is also increased with tile-based reconfiguration.

Reconfiguration Architecture Comparisons

Summarizing and analyzing the discussions from above, we can compare the reconfiguration architectures as in Table 4, where the traditional reconfiguration architecture supports static and full reconfiguration, while the modern architecture supports dynamic and partial reconfiguration. When the resource requirements of a hardware design can be met by a configware, the performance and power will be the same for both traditional and modern reconfiguration architectures because dynamic and partial reconfiguration techniques are not needed. However, when a hardware design cannot fit in a configware, these modern techniques will be required, resulting in higher performance and power.

The cost, flexibility, and design effort are all quite high for the modern architecture because the dynamic and partial reconfiguration techniques require additional hardware support, tool support, scheduling support, and user expertise, while providing greater flexibility in system design.

FUTURE TRENDS

With technology progress, not only has the gate count capacity of FPGA increased rapidly, but chip reconfiguration can now be performed at *run-time* and *partially*. On one hand, the large gate capacity enables several hardware tasks to run concurrently in a single FPGA chip, which could also interact with software tasks running on a microprocessor. On the other

Table 4. Comparison of configuration

Reconfiguration Architecture	Reconfiguration Classification		Advantage				
	Temporal	Spatial	Performance	Cost	Power	Flexibility	Design effort
Traditional	Static	Full	Low	Low	Low	Low	Low
Modern	Dynamic	Partial	High/Low	High	High/Low	High	High

hand, the partial run-time reconfigurability allows a system to dynamically change some of its hardware functionalities such as in mobile networking, wearable computing, and networked embedded systems. The complexity in designing such systems drives the need for an *operating system* that not only manages software tasks and resources, but also manages hardware tasks and related FPGA resources.

As an important feature trend, in the design and implementation an *operating system for reconfigurable systems* (OS4RS), we need to complete the following tasks.

1. OS4RS architecture design, including services, components, abstractions, performance, and hardware support mechanism
2. OS4RS kernel design, including the main kernel, system call interfaces, device drivers, loader, partitioner, scheduler, process manager, memory manager, placer, and router
3. OS4RS loadable modules design, including power manager, virtual memory manager, and other drivers
4. Set up the experiment platform and related software and *computer-aided design* (CAD) toolchains.

CONCLUSION

Reconfigurable computing has become an important field of research in the academia and the industry. Reconfigurable architecture requires changes in both computer architectures and software systems leading to many research topics. A reconfigurable system is composed of a standard processor and a set of reconfigurable hardware. The reconfigurable hardware executes application-specific computing intensive task and is reconfigured by the standard processor. Reconfigurable computing combines the advantages of both general-purpose and ASIC computing, including performance, cost, power, flexibility and design effort. Due to its reconfigurable hardware characteristics, reconfigurable computing has accelerated and enhanced the HW-SW codesign flow.

Reconfigurable computing architecture can be classified according to the granularity of configware. One is fine-grained and the other is coarse-grained. The most representative of fine-grained configware is FPGA and rDPA is the representative of coarse-grained configware. To increase flexibility of reconfiguration, configwares have been enhanced from static/full reconfiguration to dynamic/partial reconfiguration.

To develop a reconfigurable system, we can use the design patterns collected by DeHone et al. (2004) as an initial solution. Several tools and platforms developed in the academia and in the industry support reconfigurable computing and thus will help us to implement reconfigurable systems easily and rapidly. When the reconfigurable system is implemented, we can use the verification methodologies to verify the system.

With technology progress, not only has the gate count capacity of FPGA increased rapidly, but chip reconfiguration can now be performed at *run-time* and *partially*. To design and implement an *operating system for reconfigurable systems* (OS4RS) will be an important future trend in reconfigurable computing.

REFERENCES

Barr, M. (1998). A reconfigurable computing primer. *Multimedia System Design, 9*, 44-47.

Black, D. C., & Donovan, J. (2004). *SystemC: From the ground up*. Kluwer Academic.

Compton, K., & Hauck, S. (2002). Reconfigurable computing: A survey of systems and software. *ACM Computing Surveys, 34*(2), 171-210.

DeHon, A., Adams, J., DeLorimier, M., Kapre, N., Matsuda, Y., Naeimi, H., et al. (2004, April). Design patterns for reconfigurable computing. In *Proceedings of the 12th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM'04),* Napa, CA, USA, (pp. 13-23).

Donato, A., Ferrandi, F., Redaelli, M., Santambrogio, M. D., & Sciuto, D. (2005, April). Caronte: A complete methodology for the implementation of partially dynamically self-reconfiguring systems on FPGA platforms. In *Proceedings of the 13th Annual IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM'05),* Napa, CA, USA, (pp. 321- 322).

Estrin, G. (1960, May). Organization of computer systems: The fixed plus variable structure computer. In *Proceedings of the Western Joint Computer Conference*, San Francisco, USA, (pp. 33-40).

Gokhale, M., & Graham, P. S. (2005). *Reconfigurable computing –accelerating computation with field-programmable gate arrays*. Springer-Verlag.

Goldstein, S.C., Schmit, H., Budiu, M., Cadambi, S., Moe, M., & Taylor, R.R. (2000). PipeRench: A reconfigurable architecture and compiler. *Computer, 33*(4), 70-77.

Hartenstein, R. (2006, March). Why we need reconfigurable computing education. In *Proceedings of the 1st International Workshop on Reconfigurable Computing Education,* Karlsruhe, Germany, (pp. 1-11).

Hennessy, J. L., & Patterson, D. A. (2007). *Computer architecture—a quantitative approach* (4th ed.). Morgan Kaufmann.

Hsiung, P.-A., Huang, C.-H., & Liao, C.-F. (2006, August). Perfecto: A system C-based performance evaluation frame-

work for dynamically partially reconfigurable systems. In *Proceedings of the 16th IEEE International Conference on Field Programmable Logic and Applications (FPL 2006),* Madrid, Spain, (pp. 190-198). IEEE CS Press.

Hsiung, P.-A., & Santambrogio, M. (2008). *Reconfigurable system design and verification*. USA: CRC Press.

Schewel, J. (1998, March). Hardware/software codesign system using reconfigurable computing technology. In *Proceedings of the 12th International Parallel Processing Symposium & 9th Symposium on Parallel and Distributed Processing (IPPS/SPDP),* Orlando, FL, USA, (pp. 620-625).

Singh, H., Lee, M.-H., Lu, G., Kurdahi, F.J., Bagherzadeh, N., & Filho, E. M. C. (2000). MorphoSys: An integrated reconfigurable system for data-parallel and computation-intensive applications. *IEEE Transactions on Computers, 49*(5), 465-481.

Tanner Research, Inc. (2007). *Reconfigurable computing*. Retrieved December 14, 2007, from http://www.reconfigcomputing.com/default.htm

Tessier, R., & Burleson, W. (2001). Reconfigurable computing for digital signal processing: A survey. *Journal of VLSI Signal Processing, 28*(1-2), 7-27.

Tredennick, N. (1996). The case for reconfigurable computing. *Microprocessor Report, 10*(10), 25-27.

Tseng, C.-H., & Hsiung, P.-A. (2006, December). A UML-based design flow and partitioning methodology for dynamically reconfigurable systems. In *Proceedings of the 2005 IFIP International Conference on Embedded and Ubiquitous Computing (EUC'2005*), Nagasaki, Japan, (pp. 479-488). Lecture Notes in Computer Science (LNCS) 3824.

TU Kaiserslautern. (2007a). *The Configware page*. Retrieved December 14, 2007, from http://configware.org

TU Kaiserslautern. (2007b). *The Flowware page*. Retrieved December 14, 2007, from http://flowware.net

Vahid, F., & Givargis, T. (2002). *Embedded systems design—a unified hardware/software introduction*. John Wiley & Sons.

Voros, N. S., & Masselos, K. (2005). *System level design of reconfigurable systems-on-chip*. Springer-Verlag.

Xilinx Inc. (2007). *Xilinx : Virtex Series FPGAs*. Retrieved December 14, 2007, from http://www.xilinx.com/products/silicon_solutions/fpgas/virtex/index.htm

KEY TERMS

Codesign: The meeting of objectives by exploiting the trade-offs between hardware and software in a system through their concurrent design.

Configware: Source programs for configuration like field-programmable gate arrays (FPGA) or reconfigurable data path array (rDPA).

Dynamic/Static Reconfiguration: A reconfiguration technology that allows resources in a configware to be programmed without/with resetting the configware.

Field-Programmable Gate Arrays (FPGA): A programmable integrated circuit and contains a set of gate array that is programmed in the field.

Flowware: A data-stream-based software. It is the counterpart of the traditional instruction-stream-based software.

Full/Partial Reconfiguration: A traditional/modern reconfiguration technology which forces all/partial resources of a configware to be programmed during each configuration.

Reconfigurable Data Path Array (rDPA): A programmable integrated circuit with coarse-grained granularity.

Reconfiguration: The process of physically altering the location or functionality of configwares with new ones.

System on Chip/SoC: A chip which is complete to constitute an entire system or major subsystem.

Wearable Computing: A small portable computer that is designed to be worn on the body during use.

Referential Constraints

Laura C. Rivero
Universidad Nacional del Centro de la Provincia de Buenos Aires, Argentina
Universidad Nacional de La Plata, Argentina

INTRODUCTION

Inclusion dependencies support essential semantic aspects of the standard relational data model. An inclusion dependency is defined as the existence of attributes in a table whose values must be a subset of the values of the corresponding attributes in another table (Codd, 1990; Abiteboul, Hull, & Vianu, 1995; Connolly & Begg, 2004). Formally, it can be expressed as $R[X] \subseteq S[Z]$. R and S are relation names. With X and Z as compatible attributes, R[X] and S[Z] are the inclusion dependency's left and right sides respectively. When Z is the primary key of S or it is restricted by a unique clause, the inclusion dependency is key-based (also named referential integrity restriction, *rir*). In this case, X is a foreign key (FK) for R. On the contrary, if Z does not constitute the key of the relation, the inclusion dependency is non-key-based (simply, an inclusion dependency, *id*). Both *rir*s and *id*s are referential constraints.

*Rir*s are important because they contain basic local semantic characteristics, which have been elicited from the Universe of Discourse (UofD). They are sufficient to symbolize many natural semantic links such as the relationships and hierarchies that are captured by semantic models (Abiteboul et al., 1995). Conversely, *id*s do not appear as a product of the translation 'conceptual schema→logical schema', but because of ad-hoc changes made in the phase of detail design, some denormalization degree, or the presence of complex n-ary relationship constructs. In this scenario, *id*s frequently misrepresent objects and their corresponding inter-object relationships.

*Rir*s can be declaratively defined via the SQL foreign key clause (SQL:1999-2, 1999) and are enforced by most current database systems:

FOREIGN KEY (<referencing column list>) REFERENCES <referenced table name> [(<referenced column list>)]

[MATCH <match type>]
[ON UPDATE <update referential action>]
[ON DELETE <delete referential action>]

If <referenced column list> is omitted, the foreign key refers to the primary key of <referenced table name>.

The *rir*s can be specified with respect to different match types: SIMPLE (implicit if no match option is declared), PARTIAL, and FULL. As stated in the SQL:1999 standard document: If <match type> is not specified, then for each row in the referencing table, either the referencing column has at least one null value or its value matches the value of a corresponding row in the referenced table. If PARTIAL is specified, then for each row in the referencing table the value of each foreign key column is null, or it has at least one non-null value that equals the corresponding referenced column value. Finally, FULL means that, for each row in the referencing table, either all foreign key values are null or they equal the value of the corresponding referenced column (Türker & Gertz, 2001; SQL:1999-2).

When an integrity restriction is violated, the usual response of the system is the *rollback* of the data manipulation intended by the user. In the case of *rir*s, some other alternative actions are possible. These actions, named referential actions or referential rules, specify the behavior of the left and right relations under the deletion or the updating of a referenced row (a row in the right table), or the insertion of a row in the referencing (left) table. Possible actions are: *cascade*, *restrict*, *no action*, *set null*, *set default* (Markowitz, 1994; SQL:1999_2, 1999; Türker & Gertz, 2001). With the *cascade* option, the referencing rows will be deleted (updated) together with the referenced row. With the *set null* (*set default*) option, all references to the deleted (updated) row will be set to null (default) values. The deletion (update) of a referenced row is disallowed by *restrict* and *no action* rules, whenever at least a row in the left table is pointing to it. The unique referential rule for insertions is *restrict*: inserting a row into the referencing table is possible only if the referenced tuple already exists in the right term.

*Id*s may be defined with general CHECK statements having the semantics of assertions or triggers. Erroneously *id*s are frequently specified by means of an attribute-based CHECK constraint associated to the referencing table, requiring the existence of the referred-to value.

CHECK (<referencing column> IN (SELECT <referenced column>
FROM <referenced table>))

Since these constraints are checked whenever any tuple changes the value for that attribute, an update of the referred-to value in the referenced table would result in the attribute-based CHECK constraint becoming violated.

Triggers are a widespread way of implementation, although they usually complicate the development of ap-

plication programs and make the integrity maintenance quite difficult (Date & Darwen, 1997; Elmasri & Navathe, 2000; Connolly & Begg, 2004).

BACKGROUND

The comprehension of the semantic issues related to referential constraints is facilitated by the study of the syntactic structure of their terms.

Structure

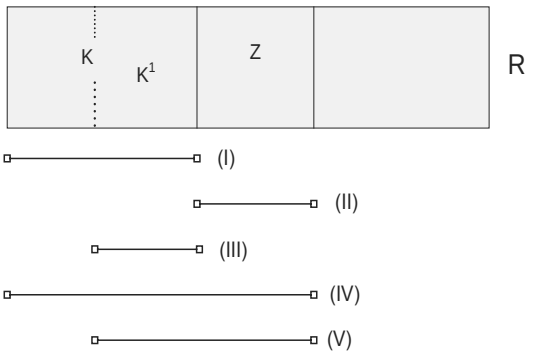
Considering a relation shape, there are five possible placements of a non-empty set of attributes with regard to the key placement. With W as such a set of attributes and K the primary key of R, the five placements are depicted in the Figure 1: (I) $W \equiv K$ (W coincides with K); (II) $W \equiv Z$, where Z is a subset of non-key attributes; (III) $W \equiv K^1$, where K_1 is a proper subset of K, $K_1 \neq \varnothing$; (IV) $W \equiv K \cup Z$; and finally (V) $W \equiv K_1 \cup Z$, $K_1 \neq \varnothing$ (W and K partially overlap). In all cases, $Z \neq \varnothing$.

As a consequence, 25 possible configurations of $R[W_R] \subseteq S[W_S]$ can be derived (see Table 1). The five cases having $S[W_S]$ as the primary key for S (numbered 1 to 5 in Table 1) correspond to *rir*s.

Semantic Perspective

*Rir*s of types I, II, and III represent typical relationships in semantic models (Abiteboul et al., 1995). Type I depicts subtype relationships; type II corresponds to designative relationships such as 1:1, N:1, or n-ary relationships with at least one 1 cardinality; and type III appears in associative relationships such as N:N, n-ary relationships, and weak entities (Elmasri & Navathe, 2000; Teorey, 1990). *Rir*s of types IV, V, and *id*s deserve a different analysis, as they cannot be derived from a conceptual model. They appear as the specification of well-typified business rules with the semantics of an inclusion in latter stages of the logical design (Rivero, Doorn, & Ferraggine, 2001, 2004).

Figure 1. Placements of a set of attributes in correlation with the key



K_*(key); Z_* (non-key attributes); K_*^1 (a proper subset of K_*); $_* = l, r$ (left, right)

Table 1. Possible structures for referential constraints

W_l \ W_r	I) Key (K_r)	II) Non Key (Z_r)	III) Part of a Key (K_r^1)	IV) Key + Non Key ($K_r \cup Z_r$)	V) Part of a Key + Non Key ($K_r^1 \cup Z_r$)
(I) Key (K_l)	**1.** $K_l << K_r$	**6.** $K_l \subseteq Z_r$	**11.** $K_l \subseteq K_r^1$	**16.** $K_l \subseteq K_r \cup Z_r$	**21.** $K_l \subseteq K_r^1 \cup Z_r$
(II) Non Key (Z_l)	**2.** $Z_l << K_r$	**7.** $Z_l \subseteq Z_r$	**12.** $Z_l \subseteq K_r^1$	**17.** $Z_l \subseteq K_r \cup Z_r$	**22.** $Z_l \subseteq K_r^1 \cup Z_r$
(III) Part of a Key (K_l^1)	**3.** $K_l^1 << K_r$	**8.** $K_l^1 \subseteq Z_r$	**13.** $K_l^1 \subseteq K_r^1$	**18.** $K_l^1 \subseteq K_r \cup Z_r$	**23.** $K_l^1 \subseteq K_r^1 \cup Z_r$
(IV) Key + Non Key ($K_l \cup Z_l$)	**4.** $K_l \cup Z_l << K_r$	**9.** $K_l \cup Z_l \subseteq Z_r$	**14.** $K_l \cup Z_l \subseteq K_r^1$	**19.** $K_l \cup Z_l \subseteq K_r \cup Z_r$	**14.** $K_l \cup Z_l \subseteq K_r^1 \cup Z_r$
(V) Part of a Key + Non Key ($K_l^1 \cup Z_l$)	**5.** $K_l^1 \cup Z_l << K_r$	**10.** $K_l^1 \cup Z_l \subseteq Z_r$	**15.** $K_l^1 \cup Z_l \subseteq K_r^1$	**20.** $K_l^1 \cup Z_l \subseteq K_r \cup Z_r$	**25.** $K_l^1 \cup Z_l \subseteq K_r^1 \cup Z_r$

CHARACTERISTICS AND APPLICATIONS OF REFERENTIAL CONSTRAINTS

A description of main issues about referential constraints and relevant applications of this concept follows.

Referential Actions and Global Semantics

Update operations promote the execution of specialized triggers—the referential actions—for the programmed maintaining of *rir*s. As mentioned, the standard actions are: *cascade, restrict, no action, set null, set default* (Markowitz, 1994; SQL:1999_2, 1999; Türker & Gertz, 2001).

Despite the fact that the local effect of such rules is precisely defined, when update operations are executed on the database state, the global effects of those interacting actions may show ambiguities (Markowitz, 1994; Reinert, 1996; May & Lüdascher, 2002). This problem has been—and currently is—a matter of profuse research from the beginning of the relational databases era. Markowitz (1994) described the anomalies caused by the use of referential actions (in some cases interacting with null constraints) and presented some solutions for its treatment. In addition, Reinert (1996) revised this study and showed that the problem of deciding if a relational schema may or may not have an instance leading to ambiguities is undecidable. The whole understanding of the global behavior of referential actions based on the local characterization of the rules was formalized in May and Ludäscher (2002).

Referential actions for *id*s have not been defined, but they can be inferred from those corresponding to *rir*s. All standard local actions can be directly applied when the deleted (updated) tuple is the one that contains the last instance of a referred value. Naturally, the global effect of these extended referential actions and the interaction with other restrictions must be studied in the context of the coexistence of triggers and declarative restrictions. In other words, it depends on the adherence of each product to the SQL:1999 execution model (Türker & Gertz, 2001).

Inclusion Dependencies and Functional Dependencies: Interaction, Inference Rules, and the Implication Problem

Levene and Loizou (1999) affirm that "the interaction between functional and inclusion dependencies is a complex problem, and there is not a complete and sound system of axioms for functional and inclusion dependencies at all." While some interactions may result in a new (functional or inclusion) dependency being obtained, other cases derive in redundant attributes. Pullback, Collection, and Attribute-Introduction rules capture significant interactions between such restrictions (Levene & Vincent, 2000; Levene & Loizou, 1999). Levene and Loizou (1999) discuss the relational database theory needed to understand the inference rules and the implication problem for *id*s. The axiom systems for *id*s and the interaction between *id*s and functional dependencies were originally introduced in Mitchell (1983) and Casanova, Fagin, and Papadimitriou (1984). Particularly, the pullback rule is one of the mechanisms useful to solve specific schema reengineering cases.

Use of Inclusion Dependencies as Domain Constraints

Some relationships have the semantics of *id*s or *rir*s symbolizing, essentially, domain restrictions that indicate the legal values for an attribute. UofD business rules associated to specific domain restrictions over dynamic and voluminous sets of values are frequently written as *id*s or *rir*s. The right term may be a table containing only the valid values for an attribute or an arbitrary set of columns of another table. For example, CHECK (LeftAttrList IN (SELECT RightAttrList FROM R)) indicates that the set of allowable values for LeftAttrList is conformed by the current set of instances of RightAttrList in R (Rivero et al., 2001, 2004). Naturally, as the specification of this constraint involves a subquery, under certain circumstances it could be violated. This problem was addressed in the Introduction.

Equality Constraint

Like other constraints, equality restrictions embody an important portion of the UofD business logic. An equality constraint between two sets of values is shorthand for two inclusion dependencies, one in either direction. Such a constraint indicates that the population of one column (simple or composite) in a table must always be equal to the values of the corresponding column (simple or composite, respectively) in another table. They would have to be modeled at the conceptual phase, but its specification is often deferred to late stages of the logical design. However, the analysis of the syntactic structures of the equality constraints terms, similar to the one made for the inclusion dependencies, allows the designers to find adequate forms of specification during the conceptual design. In addition, the study of inclusion dependencies helps to comprehend relevant issues of the equality constraints as they express comparability of domains for two different sets of attributes (Halpin, 1998; Rivero et al., 2004).

Basis for the Database Reengineering

This application is strongly related to the previous two issues, as inclusion dependencies and equality constraints enable different kinds of schema transformations. The structure of the left and right terms of *id*s is a key concept for the reengineering of relational database schemas. *Rir*s of types IV and V, and all cases of *id*s reveal specific business logic specified as inclusion relationships. A careful examination of the structure of these constraints allows the transformation of a subject schema into an enhanced one by applying the pullback rule, specific heuristic methods, and/or refinements of the final product (Tari, Buhkres, Stokes, & Hammoudi, 1998; Rivero et al., 2001, 2004). With respect to the equality constraints, the reiterated application of such restructuring mechanisms helps finding design patterns that can be captured with the standard facilities provided by any CASE tool for the relational conceptual design.

Referential Integrity in Object-Oriented and Object-Relational Databases

From its beginnings, relational technology has experienced a fantastic process of evolution to finally become a mature environment. Main commercial and open source software offer referential constraints and triggers to maintain referential integrity, even though support for SQL:1999 characteristics for referential integrity varies by product.

Referential integrity is a concept that also applies to post-relational environments. Object-oriented (OO) literature typically uses the term "relationship" to mean relationships supported by foreign keys in a relational system. This implementation may be used in an object-relational DBMS (ORDBMS) as well (Stonebraker, 1999). Object-relational (OR) systems, in compliance with the current standard—SQL:1999—provide the *reference* as a natural surrogate for primary key-foreign key relationships. In these systems, a column in a table may contain a value that is a *reference* to an instance of a type stored in another table. This implementation is supported by the unique object identifier (*oid*) of rows (SQL:1999-2).

This is a pragmatic implementation since the user has no way to update or generate unintentionally *oid*s. In this way, the system can guarantee the referential integrity since a reference is valid when it is created. On the other hand, being able to perform delete operations raises the problem of the integrity of the references. Referential actions, defined in the relational context, must be efficiently implemented in the OR and OO context. Different strategies suggest implementing integrity rules following the principles of the relational side of ORDBMS. Others, coming from the OO technology, recommend the implementation following the business logic—that is, within the objects.

There are three standard maintenance strategies: (a) the reference can be deleted, by placing null in the reference pointer; (b) the deletion of the object could be rejected if it is referenced by other objects, or (c) the referencing object could be deleted as well. Other solutions could be implemented to cope with the problem of dangling references: to allow delete operations freely (causing exceptions whenever the deleted object is being referenced) or a reference count (an object could not be deleted unless its reference count is zero).

As mentioned, programming code (triggers) for referential integrity is an alternative to the use of foreign key constraints, even though these ones are preferable to explicit triggers since they are declarative, more concise, and then more productive (Blaha, 2005). The decision about the choice of the best strategy taking into account the structure of relations, their variability, and other performance issues pose a challenging research area.

FUTURE TRENDS

Referential integrity concepts have usually been associated to implementation issues in databases. However, referential integrity has a deeper meaning related to dependencies among objects (Blaha, 2005) and makes this concept applicable to modeling methodologies. Moreover, modern software development works with languages that implement classes and their relationships; as a consequence, application programs must deal with referential integrity as well.

While OO programming and semantic models are now the standard for best developers and programmers, OO databases constitute just a small portion of the market, albeit they have gained popularity in the last few years. One of the reasons why the OO paradigm has been adopted to extend the relational technology is because it allows developers to cope with the growing capacity of current software development. Consequently, the OO paradigm has been the basis for current research on models that integrate the relational data model and SQL query languages with features coming from the object-oriented world. The result of this integration is the current generation of OR database systems (ORDBMS), which emerged in the last decade (Stonebraker, 1999). As mentioned, this evolution has posed new challenges related to the referential integrity maintenance.

The degree of distribution and heterogeneity of the DBMS and the variety of data that current DBMSs can manage are two of the current directions of developments and advances in DBMS technology. With respect to the first issue, the definition and implementation of a consistency checking service is particularly difficult whenever data is stored in a heterogeneous collection that is not controlled by a single DBMS (Hamidah, 2005).

On the other hand, multimedia databases (Dunckley, 2002), spatio/temporal databases (Rodríguez-Tastets, 2005),

XML databases (Graves, 2001), and active databases (Zaniolo et al., 1997; Paton & Diaz, 1999) are recent examples that illustrate the variety of data information that a DBMS should be able to manage. Efficient implementation of integrity issues and mainly referential integrity constraints to deal with heterogeneous data sources in centralized or distributed environments represent a challenge and a promising research area.

The World Wide Web, lacking specific guidelines to enforce referential integrity (Aldana, Yagüé, & Gómez, 2002), and newer multilevel security strategies posing a different scenario to the integrity maintenance (Lee, Kim, & Kim, 2004); digital libraries (Martínez-González, 2005) and the recent development of applications and DBMSs for mobile databases (Pitoura & Samaras, 1998)—all have given rise to interesting challenges and research opportunities in the area of database management and particularly in the area of data consistency.

CONCLUSION

This work presents an overview of main issues and applications of inclusion dependencies and the particular case of referential integrity restrictions. Some topics related to its structure and semantics have been sketched. Relevant results related to both inclusion dependencies interacting with functional dependencies and the application of these concepts in the context of database reengineering have been outlined.

Atypical behaviors promoted by design flaws should encourage users to consider referential constraints not only as merely integrity constraints but also as the basis for a good practice on database design. With respect to this problem, recent findings in the context of object-oriented applications, reengineering patterns, and modern software development theory were briefly explained.

Finally, current and future directions of database evolution are succinctly mentioned to describe the scenarios in which the integrity issues must mature.

REFERENCES

Abiteboul, S., Hull, H., & Vianu V. (1995). *Foundations on databases.* Boston: Addison-Wesley.

Aldana, J., Yagüé, M., & Gómez, A. (2002). Integrity issues in the Web: Beyond distributed databases. In J. Doorn & L. Rivero (Eds.), *Database integrity: Challenges and solutions.* Hershey, PA: Idea Group.

Blaha, M. (2005, November). *Referential integrity is important for databases.* Retrieved from http://www.odbms.org/download/007.02 Blaha Referential Integrity Is Important For Databases November 2005.PDF

Casanova, M., Fagin, R., & Papadimitriou, C. (1984). Inclusion dependencies and their interaction with functional dependencies. *Journal of Computer and System Sciences, 28,* 29-59.

Codd, E. (1990). *The relational model for database management. Version 2.* Boston: Addison-Wesley.

Connolly, T., & Begg, C. (2004). *Database systems: A practical approach to design, implementation and management* (4th ed.). Boston: Addison-Wesley.

Date, C., & Darwen, H. (1997). *The SQL standard* (4th ed.). Boston: Addison-Wesley.

Dunckley, L. (2002). *Multimedia databases: An object relational approach.* Pearson Education.

Elmasri, R., & Navathe, S. (2000). *Fundamentals of database systems.* Boston: Addison-Wesley.

Graves, M. (2001). *Designing XML databases.* Englewood Cliffs, NJ: Prentice Hall.

Lee, S.-W., Kim, Y.-H., & Kim, H.-Y. (2004). The semantics of an extended referential integrity for a multilevel secure relational data model. *Data & Knowledge Engineering, 48,* 129-152.

Halpin, T. (1998). UML data models from an ORM perspective: Part five. *Journal of Conceptual Modeling*, (5). Retrieved from http://www.inconcept.com/jcm/October1998/halpin.html

Hamidah, I. (2005). Checking integrity constraints in a distributed database. In J. Doorn, L. Rivero, & V. Ferraggine (Eds.), *Encyclopedia of database technologies and applications.* Hershey, PA: Idea Group Reference.

Levene, M., & Vincent, W.M. (2000). Justification for inclusion dependency normal form. *IEEE Transactions on Knowledge and Data Engineering, 12*(2), 281-291.

Levene, M., & Loizou, G. (1999). *A guided tour of relational databases and beyond.* London: Springer-Verlag.

Martínez-González, M. (2005). Approaches to the document versioning issue in digital libraries. In J. Doorn, L. Rivero, & V. Ferraggine (Eds.), *Encyclopedia of database technologies and applications.* Hershey, PA: Idea Group Reference.

May, W., & Ludäscher, B. (2002). Understanding the global semantics of referential actions using logic rules. *ACM Transactions on Database Systems (TODS), 27*(4), 343-397.

Markowitz, V. (1994). Safe referential integrity and null constraint structures in relational databases. *Information Systems, 19*(4). 359-378.

Mitchell, J. (1983). Inference rules for functional and inclusion dependencies. *Proceedings of the 2nd ACM SIGACT-*

SIGMOD Symposium on Principles of Database Systems (pp. 58-69), Atlanta, GA.

Paton, N.W., & Diaz, O. (1999). Active database systems. *ACM Computing Surveys, 31*(1), 63-103.

Pitoura, E., & Samaras, G. (1998). *Data management for mobile computing.* Kluwer.

Reinert, J. (1996). Ambiguity for referential integrity is undecidable. In G. Kuper & M. Wallace (Eds.), *Constraint databases and applications* (pp. 132-147). Berlin: Springer-Verlag (LNCS 1034).

Rivero, L., Doorn, J., & Ferraggine, V. (2001). Inclusion dependencies. In S.A. Becker (Ed.), *Developing quality complex database systems: Practices, techniques and technologies* (pp. 261-278). Hershey, PA: Idea Group.

Rivero, L., Doorn, J., & Ferraggine, V. (2004). Enhancing relational schemas through the analysis of inclusion dependencies. *International Journal of Computer Research, 12*(4).

Rodríguez-Tastets, M.A. (2005). Challenges of consistency in spatial databases. In J. Doorn, L. Rivero, & V. Ferraggine (Eds.), *Encyclopedia of database technologies and applications.* Hershey, PA: Idea Group Reference.

SQL:1999-2. (1999). *Database language SQL, part 2.* Document ISO/IEC 9075-2, SQL Foundation, USA.

Stonebraker, M. (1999). *Object-relational DBMSs. Tracking the next great wave.* San Francisco: Morgan Kauffman.

Tari, Z., Buhkres, O., Stokes, J., & Hammoudi, S. (1998). The reengineering of relational databases based on key and data correlations. In S. Scappapietra & F. Maryanski (Eds.), *Searching for semantics: Datamining, reverse engineering, etc.* Chapman & Hall.

Teorey, T.J. (1990). *Database modeling and design. The entity-relationship approach.* San Francisco: Morgan Kaufmann.

Türker, C., & Gertz, M. (2001). *Semantic integrity support in SQL-99 and commercial (object-)relational database management systems.* Retrieved from http://www.db.cs.ucdavis.edu/papers/TG00.pdf

Zaniolo, C., Ceri, S. et. al. (1997). *Advanced database systems.* San Francisco: Morgan Kauffman.

KEY TERMS

Axiom System for Inclusion Dependencies: Set of inference rules that axiomatize *id*s. Let **ID** be a set of *id*s over a database schema $\mathbf{R} = \{R_1, R_2, \ldots, R_k\}$, **FD** a set of functional dependencies over **R**, and $sch(R_i)$ the set of attributes of R_i:

(id-1) Reflexivity: If $X \subseteq sch(R_i)$, then $R_i[X] \subseteq R_i[X] \in \mathbf{ID}$.

(id-2) Projection and Permutation: if $R_1[X] \subseteq R_2[Y] \in \mathbf{ID}$, with $X = < A_1, A_2, ..., A_m > \subseteq sch(R_1)$, $Y = < B_1, B_2, ..., B_m > \subseteq sch(R_2)$, and $i_1, i_2, ..., i_k$ is a sequence of distinct natural numbers over {1, ..., m}, then $R_1[A_{i_1}, A_{i_2}, ..., A_{i_k}] \subseteq R_2[B_{i_1}, B_{i_2}, ..., B_{i_k}] \in \mathbf{ID}$.

(id-3) Transitivity: If $R_1[X] \subseteq R_2[Y]$; $R_2[Y] \subseteq R_3[Z] \in \mathbf{ID}$, then $R_1[X] \subseteq R_3[Z] \in \mathbf{ID}$.

Axiom System for the Interaction Between Inclusion and Functional Dependencies: Comprises the following three inference rules:

(if-1) Pullback: If $R_1[X,Y] \subseteq R_2[U,V]$; $R_2: U \rightarrow V \in \mathbf{FD} \cup \mathbf{ID}$, with $X,Y \subseteq sch(R_1)$ and $U,V \subseteq sch(R_2)$, then $R_1: X \rightarrow Y \in \mathbf{FD} \cup \mathbf{ID}$, where $|X|=|U|$.

(if-2) Collection: If $R_1[X,Y] \subseteq R_2[U,V]$; $R_1[X,Z] \subseteq R_2[U,W]$; $R_2: U \rightarrow V \in \mathbf{FD} \cup \mathbf{ID}$, then $R_1[X,Y,Z] \subseteq R_2[U,V,W] \in \mathbf{FD} \cup \mathbf{ID}$, where $|X|=|U|$.

(if-3) Attribute Introduction: If $R_1[X] \subseteq R_2[U]$; $R_2: U \rightarrow W \in \mathbf{FD} \cup \mathbf{ID}$, then $R_1[X,Z] \subseteq R_2[U,W] \in \mathbf{FD} \cup \mathbf{ID}$, with Z an attribute newly added to $sch(R_1)$.

Business Rules: Statements that model the reaction to events that occur in the real world, having tangible side effects on the database content.

Database Schema Reengineering: The process of analyzing a subject database schema to recover its components and their relationships. It guides the reconstitution of such a system into one with a higher level of abstraction and semantically closer to the Universe of Discourse.

Equality Constraint: Shorthand for two inverse inclusion dependencies, $R[X] \subseteq S[Z]$ and $S[Z] \subseteq R[X]$. It can be specified as $R[X]=S[Z]$ (the populations of the attributes R.X and S.Z must always be equal).

Inclusion Dependency (*id*): The existence of attributes in a relation whose values must be a subset of the values of the corresponding (compatible) attributes in another (or the same) relation: $R[X] \subseteq S[Z]$. R and S are relation names (possibly the same); R[X] and S[Z] are the left and right sides, respectively.

Referential Action: Specific reactions to compensate referential integrity violations. SQL:1999 standard defines *Cascade*, *Restrict*, *No Action*, *Set Default*, and *Set* Null actions explicitly for referential integrity restrictions.

Referential Integrity Restriction (*rir*): Particular case of an inclusion dependency, when Z is the primary key K of S: R[FK]<<S[K]. X constitutes a foreign key FK for R.

Relating Cognitive Problem–Solving Style to User Resistance

Michael J. Mullany
Northland Polytechnic, New Zealand

INTRODUCTION

This chapter explores cognitive problem-solving style and its impact on user resistance, based on the premise that the greater the *cognitive difference* (*cognitive gap*) between users and developers, the greater the user resistance is likely to be. Mullany (1989, 2003) conducted an empirical study demonstrating this. This study contradicts the findings of Huber (1983) and supports Carey (1991) in her conclusion that cognitive style theory, as applied to IS, should not be abandoned. Mullany's findings, in fact, are the opposite. Kirton (1999, 2004) supported Mullany's results. In particular, Mullany made use of Kirton's (2004) adaption–innovation theory. The emergent instrument, called the Kirton adaption–innovation inventory (KAI; Kirton, 1999, 2004), was used by Mullany as his measure of cognitive style.

Mullany's study also investigated the relationship between user resistance and user ages and lengths of service in the organisation. It failed to show any relationship between these factors and user resistance. This countermands the findings of Bruwer (1984) and dismisses any intimation that older or longer-serving employees are necessarily more resistant to change as myths.

BACKGROUND

Ever since the early 1980s, experts have identified user resistance to new systems as an expensive time overhead (see studies by Hirschheim & Newman, 1988, and Markus, 1983). Some authors suggest the greater importance of age and length of service. Bruwer (1984), for instance, claimed to have demonstrated that the older or longer-serving an employee, the more resistant he or she is likely to be to a new computer system. Clarification of issues surrounding user resistance has also highlighted *cognitive style theory* as potentially important, but to date, its impacts have only been sparsely researched in relation to user resistance, many of the prior studies being open to question. This research, on the other hand, proposes that a system will fail when the developer and user differ significantly in their problem-solving approaches. To reduce user resistance, it thus makes sense to recommend system designs that suit the user's approach to problem solving.

This issue appears only to have been studied empirically by Mullany (1989, 2003). He formulated the research question, "Is there a relationship between user resistance to a given information system and the difference in cognitive style between the user and the developer?" With the aid of his own instrument for measuring user resistance and the Kirton adaption–innovation instrument (Kirton, 1999) to measure the cognitive styles of users and associated system developers, he found a highly significant relationship between developer–user cognitive style differences and the level of user resistance to systems.

Why no other studies along similar lines have been reported in credible current research is difficult to explain. One possibility is that the literature contains speculative studies, such as that by Huber (1983), that discredit cognitive-style theory as a tool in understanding system success. Other studies, such as that by Carey (1991), while encouraging the continued use of cognitive-style theory in studying system phenomena, do not demonstrate its predictive success in information systems (IS). The remainder of this chapter thus examines the meaning and measure of cognitive style, the measure of user resistance, the specific findings of Mullany (1989, 2003), and outlooks for the future in this area of research.

THE MEANING AND MEASURE OF COGNITIVE PROBLEM-SOLVING STYLE

Liu and Ginther (1999) defined *cognitive style* as, "An individual's consistent and characteristic predispositions of perceiving, remembering, organizing, processing, thinking and problem-solving." Schroder, Driver, and Streufert (1967), in a discussion of human information processing, suggested that organisms "either inherit or develop characteristic modes of thinking, adapting or responding and go on to focus upon adaptation in terms of information processing." In short, an individual exhibits characteristic ways of processing information (and, hence, solving problems), known as his or her "cognitive style." Table 1 gives an historic summary of key experts over the years who have endeavoured to name and measure the construct of cognitive style. Of these, the MBTI (Myers–Briggs type indicator) is the most used in current, credible research literature, followed by the KAI

Table 1. Cognitive-style constructs: Key studies

Reference	Cognitive-Style Construct	Instrument
Kelly (1955)	Cognitive complexity or simplicity	**RepGrid** (Repertory grid)
Jung (1960)	Jungian typology	**MBTI** (Myers–Briggs type indicator)
Witkin et al. (1967)	Field dependence or independence	**EFT** (Embedded figures test)
Hudson (1966)	Converger or diverger	None
Schroder et al. (1967)	Cognitive complexity	**DDSE** (Driver's decision-style exercise)
Ornstein (1973)	Hemispherical lateralisation	Brain scan
Kirton (1976)	Adaptor–innovator continuum	**KAI** (Kirton adaption–innovation inventory)
Taggart (1988)	Whole-brain human information processing	**HIP** (Human information-processing instrument)

(Kirton, 1976, 1984). As previously stated, the only evident effort made to relate cognitive style to user resistance was carried out by Mullany (1989) using the KAI. The reason for his preferred use of the KAI stemmed from its ability to provide a near-continuous, bipolar scale, convenient for finding correlations and associations. The MBTI, by contrast, yields only certain cognitive classifications, where no mutual order is evident. The correlation with other factors would then have been more difficult to show statistically.

Turning to the theory behind the KAI, Kirton (1999) identified two extremes of cognitive style; namely, the *adaptor* and the *innovator*. The adaptor tends to follow traditional methods of problem solving, while the innovator seeks new, often unexpected, and frequently less-accepted methods. The adaptor tends to "do well" within a given paradigm, where the innovator tends to "do differently," thus transcending accepted paradigms. The adapter is prepared to be wedded to systems, solving problems "in the right way," but is often seen as "stuck in a groove." The innovator has little regard for traditions, is often seen as creating dissonance, and elicits comments such as, "He wants to do it his own way, not the 'right' way." All humans, Kirton proposed, can be located on a continuum between the extremes of these two cognitive styles.

Both cognitive extremes can be highly creative, can resist change, and can act as agents for change. Adaptors support changes to the conservative, back to the "good old ways," and resist changes to novel methodologies. Innovators support changes toward unprecedented systems and technologies and resist changes to the traditional.

Kirton's instrument, the KAI, has been widely demonstrated to be a successful measure of his construct of cognitive problem-solving style. The instrument takes the form of a questionnaire, on which the respondent has to rate himself or herself against 33 character traits. KAI scores can range from 32 to 160, with a mean of 96 and a standard deviation of about 16. A person scoring above the mean of 96 is considered to be an innovator; conversely, a person scoring below 96 is rated as an adaptor. However, in the range of 80 to 112 (that is, within one standard deviation of the mean), a third cognitive style can be identified—the mid-scorer. Such persons tend to have human rather than technical problem-solving preferences and can relate better to the extreme scorers than either can to the other.

A DESCRIPTION AND MEASURE OF USER RESISTANCE

Mullany (1989) measured user resistance at personal interviews with the key user of each system selected for investigation. The user was asked to list the problems that he or she recalled had occurred during the system's development and implementation. They were asked, in effect, to make complaints, in confidence, against the system and its manner of implementation. Then they were requested to rate the severity of each complaint on a seven-point scale (with seven representing the most severe weighting). The sum of severities of all the complaints measured the respondent's

resistance score or *R-score*. Obvious criticisms of the R-score method are as follows:

1. It may be highly influenced by the cognitive style of the interviewer.
2. At an interview, the user might forget certain crucial problems that had been experienced.

Mullany refuted (1) on the grounds that the same person (himself) did all the interviewing in his study. He assumed (2) to be of limited impact, because the object of the R-score method is to observe the user in the process of complaining. Consequently, the resistant user is capable of exaggerating or even inventing complaints, making the issue of those forgetten less relevant. However, he conceded the limitation that there are covert forms of resistance, such as absenteeism and withdrawal, that are not necessarily related to overt complaints.

To investigate a relationship between cognitive-style differences and user resistance, Mullany (1989) set out to collect data from a suitable sample of computer system developers and users. Bivariate data were to be collected, namely, the analyst–user KAI difference and the R-score for each user. The association between these was then to be measured. For his association measure, he used both the Kendall-*t* and the more traditional Spearman-*r*, which are equally reliable for significance testing (Liebetrau & Kendall, 1970). According to Kendall (1970; who invented the Kendall-*t* measure of association) and as confirmed by Liebetrau (1983), sample sizes of 10 to 20 are sufficient for such tests. The author thus selected a much larger sample size of 34 systems in 10 South African organizations. However, the following further criticisms were identified and addressed:

1. The sample size is small compared with some other studies in IS.
2. A user who champions a system may point out deficiencies in the hopes of improving that system.

Referring to the first of these criticisms, one should be alerted to the fact that sample representivity is more important than size in obtaining reliable results. In fact, the larger the sample, the less the researcher is likely to be able to guarantee a lack of significant bias. For example, suppose that with the aid of a suitable instrument, one sets out to measure the diligence of some human population. If a large sample size is sought through a postal, Web-based, or e-mail survey (the only practicable methods for really large samples), only the most diligent respondents are likely to respond, giving a bias to the more diligent and, thus, casting serious doubt on the results. To reduce this effect, Mullany (1989) collected all the data at personal interviews with the analysts and users. Furthermore, organisations he approached were requested to provide a fair spread of systems in use. He thus used legitimate power lent to him by the organisations to interview those as he required.

The second criticism was addressed by obtaining approval to keep all employees' responses confidential and to make this clear at each interview. This meant that a user would be unlikely to complain to Mullany in the hopes of achieving some system improvement, as he or she knew that no information would be relayed to the rest of the organisation. Every effort was made to preserve standard interviewing conditions: these being freedom from pressure or interruption and complete assurance of confidentiality. In short, interviewing conditions similar to those of face-to-face counselling were achieved. Furthermore this technique of measuring user resistance has been confirmed by respected researchers. First, both Markus (1983) and Hersheim et al. (1988) identified complaint as an overt symptom of, if not even a form of, resistance. Kirton (2004), in a discussion of Mullany's study, confirmed the technique as valid.

THE RELATIONSHIP BETWEEN USER RESISTANCE AND THE DIFFERENCES IN COGNITIVE STYLES BETWEEN THE USER AND THE DEVELOPER

The key developer and key user of each were interviewed. In each case, measures were obtained for the developer KAI score, user KAI score, and user R-score. At the same time, demographic data were collected; most particularly, the ages and lengths of service of the respondents, in order to test the findings of Bruwer (1984). A relationship as an association was found for the user R-scores versus the absolute differences between developer and user KAI scores. The association (with $p < 0.005$) proved to be strong, suggesting that user resistance can be minimized by matching a user with a developer of similar cognitive style. However, no significant associations were found between the ages and lengths of service of users and their R-scores, in contradiction of Bruwer's (1984) results. Rosen and Jerdee's (1976) study, which sought a similar result for occupational groups in general, agrees with Mullany's findings in this respect.

An interpretation of the R-score was demonstrated based on a near-perfect direct proportion that proved to exist between the weighted and nonweighted numbers of the users' complaints. In this relationship, the constant of proportionality was found to be 3.913 (that is, nearly 4). The R-score can thus be described as approximately four times the number of complaints a user will make retrospectively, in private, concerning a system and its manner of implementation.

FUTURE TRENDS

This study reignites the issue of cognitive style as an important issue in IS and completely countermands the conclusions of Huber (1983). It substantially strengthens the case made by Carey (1991) that cognitive-style issues in IS research should not be abandoned. Further, it suggests that user resistance and the related constructs of user dissatisfaction and system success can be predicted from cognitive-style measures (that is, KAI scores) prior to system development.

Areas for further research centre upon the main limitation of this study. For instance, there is little known regarding how the developer-user cognitive gap influences the system development life cycle (SDLC) over a significant passage of time, and neither this study nor any other found in the literature has achieved this. In fact, the literature is devoid of any attempts to conduct such research. A longitudinal study where SDLC curves are compared with the developer-user cognitive gap would be of immense importance and interest. New rules for system development based on cognitive-style testing would be expected to emerge.

CONCLUSION

It is clear that cognitive problem-solving style, as defined by Kirton and measured using the KAI, impacts user resistance. The greater the cognitive gap between users and developers, the greater the user resistance is likely to be. This contradicts the findings of Huber (1983) and supports Carey (1991) in her conclusion that cognitive-style theory, as applied to IS, should not be abandoned. Mullany's findings, in fact, are the opposite.

The failure to show any relationship between users' ages and lengths of service, and their resistance ratings, countermand the findings of Bruwer (1984) and suggest that organisations should be alerted to the danger of discriminating against older or longer-serving users or dispensing with their services on such grounds.

As mentioned above, areas for further research centre upon the main limitation of this study. A longitudinal study where SDLC curves are compared with the developer–user cognitive gap would be of great importance and interest.

REFERENCES

Bruwer, P. J. S. (1984). A descriptive model of success for computer-based information systems. *Information and Management, 7*, 63-67.

Carey, J. M. (1991). The issue of cognitive style in MIS/DSS research. In J. Carey (Ed.), *Human factors in information systems. An organizational perspective* (pp. 337-348). Norwood, NJ: Ablex Publishing Corp.

Hirschheim, R., & Newman, M. (1988). Information systems and user resistance, theory and practice. *The Computer Journal*, *31*(5), 398-408.

Huber, G. P. (1983). Cognitive style as a basis for MIS designs: Much ado about nothing? *Management Science, 29*(5), 567-579.

Hudson, L. (1966). *Contrary imaginations*. United Kingdom: Methuen.

Jung, C. G. (1960). *The basic writings of CG Jung*. New York: Pantheon.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York: Norton.

Kendall, M. G. (1970). *Rank correlation methods*. London: Charles Griffin & Co.

Kirton, M. (1976). Adaptors and innovators: A description and measure. *Journal of Applied Psychology, 61*(5), 622-629.

Kirton, M. (1984). Adaptors and innovators—Why new initiatives get blocked. *Long Range Planning, 17*(2), 137-143.

Kirton, M. (1999). *KAI manual* (3rd ed.). Berkhamstead: UK: Occupational Research Centre.

Kirton, M. (2004). *KAI Certification Course*. Birkhamstead, Hertfordshire: Occcupational Research Centre.

Liebetrau, A. M. (1983). *Measures of association*. Beverly Hills, CA: Sage Publications.

Liu, Y., & Ginther, D. (1999). Cognitive styles and distance education (invited submission). *Online Journal of Distance Learning Administration, 2*(3), 118.

Markus, M. L. (1983). Power, politics, and MIS implementation. *Communications of the ACM, 26*(6), 430-444.

Mullany, M. J. (1989). An analysis of the relationship between analyst–user cognitive style differences and user resistance to information systems. Master's thesis. Cape Town, South Africa: University of Cape Town.

Mullany, M. J. (2003). Forecasting user satisfaction. *The International Principal, 7*(1), 10-12.

Ornstein, R. E. (1973). *The nature of human consciousness*. San Francisco, CA: Viking Press.

Rosen, B., & Jerdee, T. H. (1976). The nature of job-related age stereo-types. *Journal of Applied Psychology, 61*(2), 180-183.

Schroder, H. M., Driver, M. J., & Streufert, S. (1967). *Human information processing. Individuals and groups functioning in complex social situations*. New York: Holt, Rinehart and Winston.

Taggart, W. M. (1988). A human information processing model of the managerial mind: Some MIS implications. In J. Carey (Ed.), *Human factors in information systems. An organizational perspective* (pp. 253-268). Norwood, NJ: Ablex Publishing Corp.

Witkin, H. A., Moore, C. A., Goodenough, D. R., & Cox, P. W. (1977). Field dependent and field independent cognitive styles and their educational implications. *Review of Educational Research, 47*, 1-64.

KEY TERMS

Adaptor: An adaptor tends to follow traditional methods of problem solving, tending to "do well." He or she is often seen as "stuck in a groove" (Kirton, 1999).

Association: A relationship between two statistical variables. Unlike a *correlation*, an association does not yield a quantitative result but is contingent upon the ranking of the bivariate data values only.

Cognitive Gap: The difference in cognitive problem-solving style between two people, especially two people who are obliged to interact as members of a group or team.

Cognitive Problem-Solving Style: The position an individual occupies between two extremes of cognitive problem-solving style personality; namely, *the adaptor* and *the innovator*.

Cognitive Style: An individual exhibits characteristic ways of processing information and, hence, solving problems, known as his or her "cognitive style."

Innovator: The innovator seeks new, often unexpected, and frequently less acceptable methods. He or she has little regard for traditions, is often seen as creating dissonance, and elicits comments such as, "He wants to do it his own way, not the 'right' way" (Kirton, 1999).

KAI (Kirton Adaption-Innovation Inventory): An instrument that measures cognitive problem-solving style. It takes the form of a questionnaire, on which the respondent is asked to rate himself or herself against 32 character traits.

R-Score (Resistance Score): A method of measuring user resistance where, at personal interviews with the key user of a given system, the user is asked to list system problems and then to rate the severity of each on a seven-point scale. The sum of severities of all the complaints measures his or her *R-score* (Mullany, 1989).

User Resistance: Any user behaviour, action, or lack of action that inhibits the development, installation, or use of an information system.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2414-2418, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Reliability Growth Models for Defect Prediction

Norman Schneidewind
Naval Postgraduate School, USA

INTRODUCTION

In order to continue to make progress in software measurement as it pertains to reliability, there must be a shift in emphasis from design and code metrics to metrics that characterize the risk of making requirements changes. By doing so, the quality of delivered software can be improved because defects related to problems in requirements specifications will be identified early in the life cycle. An approach is described for identifying requirements change risk factors as predictors of reliability problems. This approach can be generalized to other applications with numerical results that would vary according to application.

BACKGROUND

Several projects have demonstrated the validity and applicability of applying metrics to identify fault prone software at the code level (Khoshgoftaar & Allen, 1998; Khoshgoftaar, Allen, Halstead, & Trio, 1996a; Khoshgoftaar, Allen, Kalaichelvan, & Goel, 1996b; Schneidewind, 2000). This approach is applied at the requirements level to allow for early detection of reliability and maintainability problems. Once high-risk areas of the software have been identified, they would be subject to detailed tracking throughout the development and maintenance process (Schneidewind, 1999).

Much of the research and literature in software metrics concerns the measurement of code characteristics (Nikora, Schneidewind, & Munson, 1998). This is satisfactory for evaluating product quality and process effectiveness once the code is written. However, if organizations use measurement plans that are limited to measuring code, the plans will be deficient in the following ways: incomplete, lack coverage (e.g., no requirements analysis and design), and start too late in the process. For a measurement plan to be effective, it must start with requirements and continue through to operation and maintenance. Since requirements characteristics directly affect code characteristics and hence reliability, it is important to assess their impact on reliability when requirements are specified. As will be shown, it is feasible to quantify the risks to reliability of requirements changes—either new requirements or changes to existing requirements.

Once requirements attribute that portend high risk for the operational reliability of the software are identified, it is possible to suggest changes in the development process of the organization. To illustrate, a possible recommendation is that any requirements change to mission critical software—either new requirements or changes to existing requirements--would be subjected to a *quantitative* risk analysis. In addition to stating that a risk analysis would be performed, the policy would specify the risk factors to be analyzed (e.g., number of modifications of a requirement or *mod level*) and their threshold or critical values. The validity and applicability of identifying critical values of metrics to identify fault prone software at the code level have been demonstrated (Schneidewind, 2000). For example, on the space shuttle, rigorous inspections of requirements, design documentation, and code have contributed more to achieving high reliability than any other process factor. The objective of these policy changes is to prevent the propagation of high-risk requirements through the various phases of software development and maintenance. The payoff to the organization would be to reduce the risk of mission critical software *not* meeting its reliability goals during operation.

Definitions

- **D_{ir}:** *Cumulative* defects (e.g., discrepancy reports: program executes incorrect path due to excessive requirements conflicts), corresponding to risk variable r during operating time interval i.
- **r_i:** Requirements risk (e.g., risk of excessive size, excessive conflicting issues)
- **rmin:** Minimum value of r_i
- **rmax:** maximum value of r_i

Risk Variables:

- **Sloc:** *Cumulative* source lines of code (sloc): number of source lines of code written for a Shuttle release.
- **Issues:** *Cumulative* number of possible conflicts among requirements (e.g., an aggregation of requirements to provide greater Shuttle thrust conflicts with the fact that more thrust requires more engine weight). An aggregation of conflicting requirements issues causes reliability risk.
- **t:** Operating time (e.g., execution, wall clock, calendar time)

- **t_c:** Critical value of t = maximum value of t
- **i:** Operating time interval
- **a and b:** Coefficients of D_{ir} obtained through regression analysis
- **Decision maker:** Software quality control manager

Research Ideas on Models

Models are representations of states, objects, and events. They are less complicated than reality, hence easier to use. This is due to the fact that only relevant properties are included in the model, as explained by Ackoff, Gupta, & Minas, 1962).

Models of Problem Situations (Ackoff et al., 1962)

$V = f(X_i, Y_j)$, where

V = measure of the value of the decision that is made (i.e., action that is taken)

Example: Probability that software will survive for an operational time > t (i.e., reliability R (t))

X_i = the variables that are subject to control by the decision. The decision variables define alternative courses of action.

Example: amount of time T allocated to testing software

Discrete or continuous decision variables

Example: $D_{ri} = a\ r_i^b$ has the continuous variable defect count D_{ri}

Y_j: the factors (variable or constant) that affect performance, which may or may not, be subject to control by the decision maker. These are called parameters.

Example: Failure rate parameters α and β in reliability model

f = functional relationship between independent variables and parameters X_i and Y_j and dependent variable V.

Example: $D_{ri} = a\ r_i^b$

A model has two essential characteristics: At least one of the decision variables X_i is subject to control by the decision maker.

Example: Amount of time T allocated to testing software Second, the value V must be a measure of alternative courses of action.

Example: Probability that software will survive for an operational time > t (i.e., reliability R (t)). The decision maker sets the value of t, and, hence, the value of R (t).

Models with the above properties are called decision models.

Some models contain constraints:

Example: R (t) > R* (t), where R* (t) is the minimum allowable reliability.

Models as approximations of the real world

Therefore, we start with a small and simple defect prediction model and build upon this to achieve more complex and accurate models.

Churchman, Ackoff, and Arnoff, (1957) offers the following advice concerning how to view models:

Viewed generically, a model is a representation of some subject of inquiry such as a process, like the software defect reduction process. The model is used for prediction (e.g., predict defect count as a function of risk variables) and control (i.e., control software quality by managing defect occurrence). The primary purpose is explanatory rather than descriptive. For example, we could want to show how defect count varies as a function of risk variables and operating time interval. A great advantage of such models is the ability to manipulate the model without having the change the system that is being modeled. Thus, we would not want to change the software quality control system just to experiment with how defects vary with requirements risk! Rather, we might change the system depending on the results of our model experiments.

Box, Hunter, and Hunter (1978) states the following with respect to experimental model building:

1. The experimenter must construct a flexible model, subject to change, in case the model assumptions turn out to be erroneous. For example, we may find that defect count does not vary exponentially with requirements risk variables. With this outcome, we want a model that could be changed easily, for example, to a linear model.
2. Frequently, the mechanism underlying the model is not completely understood, or is too complicated, to allow an exact model to be postulated from theory. In this case, a simplified model with a response over a limited range of the variables is appropriate.

The manner in which defects occur is a complicated process involving the complexity of software requirements, quality of personnel, the quality of the software development process, and the nature of the application and its operating environment. Therefore, obviously, a simple model like $D_{ri} = a\, r_i^{b}$, which is only a function of requirements risk, cannot possibly capture all of the foregoing attributes. However, before we rush to judgment about simple models, we must be cognizant of the fact that if a model is too complex, it will be difficult to use and understand. Furthermore, there is the principle of *surrogate* variables that can substitute for the variables of interest. This principle is illustrated by requirements risk substituting for complexity of requirements in our model.

Hillier and Lieberman (2001) state the following principles for model development:

1. It is important to validate the model before it is put into practice. One check is to ensure that the variables and parameters have consistent dimensions, as recommended by For example, in the defect count model, Dri = a rib, where Dri has the dimension of defects and ri has the dimension, for example, of sloc, a must have the dimension of defects per sloc and b must be a dimensionless parameter.
2. Additionally, the model builder should try extreme values of the variables on the model—outside the range of the collected data—to test the reaction of the model to these outliers.

An example would be to try extremely large values of the risk variables *sloc* and *issues* and see whether the predictions of D_{ri} are plausible. The most important validation method is to compare model predictions against *future* actual values. For example, we would want to compare predictions produced by $D_{ri} = a\, r_i^{b}$ with *actual defect count* in time interval greater than the range of *i* for estimating the parameters a and b. This is not to suggest that *retrospective* testing is not valuable. It is only to urge that model builders not assume that because they obtain a good fit with historical data that the model is valid. Unfortunately, such "validation" will only be legitimate to the extent that the future is like the past!

Objective of Research

We have several objectives in this research. One is to suggest a methodology for software reliability development using templates designed by researchers in the field operations research and statistical modeling, as described in the foregoing section.

Another objective is to develop and demonstrate a new model for predicting risk (i.e., probability that defect counts exist on a release), as a function of risk variables. To develop and evaluate this model, we use defect and risk data from the NASA Space Shuttle Flight Software. We build upon our previous research, Schneidewind (2005), to develop predictors of risk (e.g., discrepancy report counts on a release), as a function of risk variables (sloc and issues). These predictors are significant because, with them, the risk to reliability of making requirements changes can be predicted early, when the cost and effort of correcting problems is low.

Desirable Properties of Model

Application of Ackoff et al. (1962) to the Defect Modeling Problem

Models are representations of states, objects, and events. They are less complicated than reality, hence easier to use. This is due to the fact that only relevant properties are included in the model.

$V = f(X_i, Y_j)$, where

V = measure of the value of the decision that is made (i.e., action that is taken)

An example of V is defect count D that is a measure of the quality of the software that would result from reduction D.

X_i = the variables that are subject to control by the decision maker. The decision variables define alternative courses of action.

Continuing the example, X_i could represent the controllable variable r_i, the risk variable i, which could be, for example, requirements risk such as Sloc: *cumulative* source lines of code and issues: *cumulative* number of possible conflicts among requirements. In this example, the r_i are continuous decision variables.

Now, consider the Y_j factors, which in our case, are the parameters a_i and b_i. These are under the control of the decision maker because their values would be obtained from regression analysis, using the r_i risk variables (i.e., X_i variables).

f = functional relationship between independent variables and parameters X_i and Y_j and dependent variable V.

Example: $D_{ri} = a\, r_i^{b}$, which expresses the dependence of defect count D_{ri} on the risk variables r_i.

Objective of Defect Count Metric

The defect count metric $D_{ri} = a\, r_i^{b}$ should be able to predict D_{ri} within the confidence intervals, as given by equation, with a specified α (.01):

$$(\bar{D}_{ri} - Z_\alpha S) \le D_{ri} \le (\bar{D}_{ri} + Z_\alpha S)$$

$$S = \sqrt{\sum_{i=1}^{N} \frac{(D_{ri} - \bar{D}_{ri})^2}{N-1}} \qquad (0.1)$$

Looking at Figure 1, it can be seen that this objective is satisfied.

Constraint of Defect Count Metric

There is a constraint levied on D_{ri}:

The objective $(\bar{D}_{ri} - Z_\alpha S) \le D_{ri} \le (\bar{D}_{ri} + Z_\alpha S)$ must be achieved in an operating time $t < t_c$

Figure 1 shows that constraint 1) is satisfied.

FUTURE TRENDS

Reliability predictions that are made in the test phase, when failure data are available, are useful, but it would be more useful to predict at an earlier phase--preferably during requirements analysis—when the cost of error correction is relatively low. Thus, there is great interest in the software reliability and metrics field in using static attributes of software in reliability modeling and prediction. Presently, the software engineering field does not have the capability to make early predictions of reliability problems. Early predictions would allow errors to be discovered and corrected when the cost of correction is relatively low. In addition, early detection would prevent poor quality software from getting into the hands of the user. As a future trend, the focus in research and practice will be to identify the attributes of software requirements that cause the software to be unreliable.

CONCLUSION

Risk factors that are statistically significant can be used to make decisions about the risk of making changes. These changes affect the reliability of the software. Risk factors that are not statistically significant should not be used; they do not provide useful information for decision-making and cost money and time to collect and process. This methodology can be generalized to other risk assessment domains, but the specific risk factors, their numerical values, and statistical results may vary.

REFERENCES

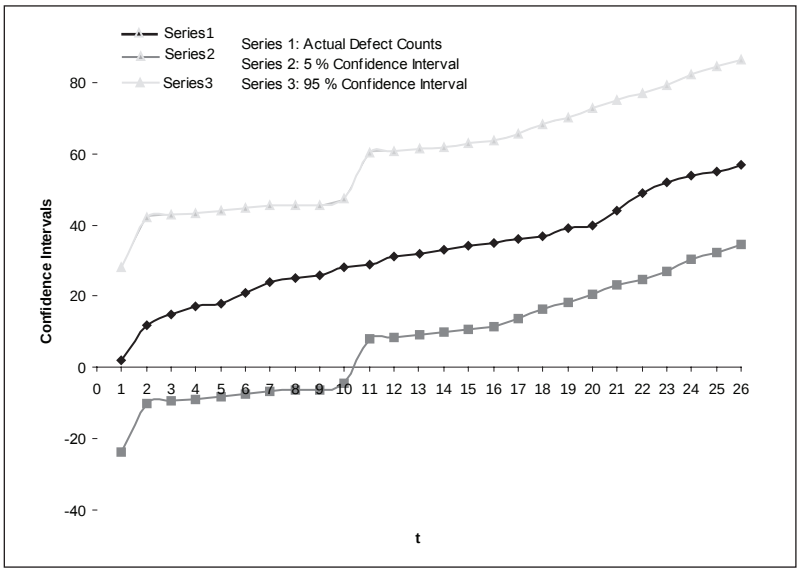
Ackoff, R. L., Gupta, S. K., & Minas, J. S. (1962). *Scientific method: Optimizing applied research decisions*. John Wiley & Sons.

Churchman, C. W., Ackoff, R. L., & Arnoff, E. L. (1957). *Introduction to operations research*. John Wiley & Sons.

Box, G. E. P., Hunter, W. G., & Hunter, J. S. (1978). *Statistics for experimenters*. John Wiley & Sons.

Hillier, F. S., & Lieberman, G. K. (2001). *Introduction to operations research* (7th ed.). McGraw Hill.

Figure 1. Confidence intervals of cumulative defect count D_{ri} (risk variable r_i = issues) vs. time intervals t



Khoshgoftaar, T. M., & Allen, E. B. (1998). Predicting the order of fault-fault-prone modules in legacy software. In *Proceedings of the 9th International Symposium on Software Reliability Engineering* (Vol. 7, pp. 344-353). Paderborn, Germany.

Khoshgoftaar, T. M., Allen, E. B., Halstead, R., & Trio, G. P. (1996a). Detection of fault-prone software modules during a spiral life cycle. In *Proceedings of the International Conference on Software Maintenance* (pp. 69-76). Monterey, CA.

Khoshgoftaar, T. M., Allen, E. B., Kalaichelvan, K., & Goel, N. (1996b). Early quality prediction: A case study in telecommunications. *IEEE Software, 13*(1), 65-71.

Nikora, A. P., Schneidewind, N. F, & Munson, J. C. (1998). *IV&V issues in achieving high reliability and safety in critical control software, final report.* Pasadena, California Jet Propulsion Laboratory, National Aeronautics, and Space Administration.

Schneidewind, N. F. (1999). Measuring and evaluating maintenance process using reliability, risk, and test metrics. *IEEE Transactions on Software Engineering, 25*(6), 768-781.

Schneidewind, N. F. (2000). Software quality control and prediction model for maintenance. *Annals of Software Engineering, 9*, 79-101. Baltzer Science Publishers.

Schneidewind, N. F. (2005). Predicting risk as a function of risk factors. *The R & M Engineering Journal, 25*(1).

KEY TERMS

Failure: The inability of a system or component to perform its required functions within specified performance requirements.

Metric: A quantitative measure of the degree to which a system, component, or process possesses a given attribute.

Quality: The degree to which a system, component, or process meets specified requirements.

Reliability: The ability of a system or component to perform its required functions under stated conditions for a specified period of time.

Requirement: A condition or capability needed by a user to solve a problem or achieve an objective.

Risk: The chance of injury, damage, or loss.

Representational Decision Support Systems Success Surrogates

Roger McHaney
Kansas State University, USA

INTRODUCTION

Rapid and frequent organizational change has been a hallmark of business environments in the past two decades. Frequently, technology and new software development are embraced as aspects of complex strategies and tactical plans. Without sufficient analysis, the unforeseen consequences of change can result in unexpected disruptions and the loss of productivity. In order to better control these contingencies, modern managers often employ a variety of decision support aids. One such aid, classified as a representational decision support system, is discrete event simulation (DES).

BACKGROUND

In its purest form, DES is considered to be a branch of applied mathematics. Its considerable popularity is due in part to the availability of computers and improvements in simulation languages and simulator software packages. DES is often the technique of choice when standard analytical or mathematical methodologies become too difficult to develop. Using a computer to imitate the operations of a real-world process requires a set of assumptions taking the form of logical relationships that are shaped into a model. These models assess the impact of randomly occurring events. Experimental designs are developed and the model manipulated to enable the analyst to understand the dynamics of the system. The model is evaluated numerically over a period of time, and output data are gathered to estimate the true characteristics of the actual system. The collected data are interpreted with statistics allowing formulation of inferences as to the true characteristics of the system. Table 1 lists primary features of a DES application.

While the value of DES in organizational settings has been accepted and is evidenced by the varied and growing market of related products, not every DES application is suited for every problem domain. For this reason, information systems researchers have worked to identify salient characteristics of DES and its usage, and then measure the relationship between its application and successful organizational outcomes. These studies have been conducted in different ways with focuses on independent and dependent variables.

INDEPENDENT VARIABLE RESEARCH

Much DES research has focused on identifying and evaluating software and project characteristics (independent variables) and then producing recommendations that suggest either how a project can be successfully implemented or how failure can be avoided. A variety of useful practitioner-focused articles have been published in this area (Hlupic & de Vreede, 2005; Swain, 2003). These studies provide recommendations, based in part on first hand experience, of consultants and simulation analysts. In many cases, these recommendations list the packages currently available and focus on the pros and cons of each.

Academic studies of independent variables have also been conducted. In one of the first studies using information systems as a basis for DES, McHaney and Cronan (2000) extended a framework of general decision support system (DSS) success factors identified by Guimaraes, Igbaria, and Lu (1992). The developed contingency model was theoreti-

Table 1. DES features

Feature	Description
Statistics Collection	Tools that gather data for purposes of inferential statistics about the model
Resource Modeling	A means for the representation of a constrained resource in the model
Transaction	A means for representation of the participants in the simulation model
Simulation Clock	Tools for analysis and step processing of the coded model
Random Number Generators	A means for producing random number streams for randomization of events within the simulation model
Model Frameworks	Generalized frameworks for the rapid development of a model

Table 2. List of DES success factors

Factor 1: Software Characteristics
Factor 2: Operational Cost Characteristics
Factor 3: Software Environment Characteristics
Factor 4: Simulation Software Output Characteristics
Factor 5: Organizational Support Characteristics
Factor 6: Initial Investment Cost Characteristics
Factor 7: Task Characteristics

cally derived from the simulation literature and empirically tested. The results indicated a seven-factor model that was structured as shown in Table 2.

Academic research by Robinson (1999) approached the problem from the opposite perspective and identified sources of simulation inaccuracy that may result in project failure. Other studies, such as one by McHaney and White (1998), developed a DES software selection framework that matched salient DES characteristics with software package features. The importance of evaluation and selection of appropriate DES packages were determined in relation to the success of simulation implementation. As might be expected, the choice of the wrong simulation package often correlated with simulation system failure. This study provided a set of criteria to be systematically considered when evaluating DES software. The taxonomy for simulation evaluation together with importance ratings provided by the collective expertise of a large number of DES users were used as guidelines in deciding the relative weighting to give various software package capabilities.

Other research (McHaney, White, & Heilman, 2002) attempted to develop an understanding of DES project success by determining which characteristics of a simulation project were more likely to be present in a successful simulation effort. Potential success factors were derived from the simulation literature and used to develop a questionnaire. Based on the findings, projects perceived as failing were often characterized by high costs, model size constraints, and slow software. Successful projects were characterized by teamwork, cooperation, mentoring, effective communication of outputs, high quality vendor documentation, easily understood software syntax, higher levels of analyst experience, and structured approaches to model development. By understanding the simulation process and characteristics of successful simulations, practitioners will find it easier to avoid common mistakes that can ruin a modeling effort.

DEPENDENT VARIABLE RESEARCH

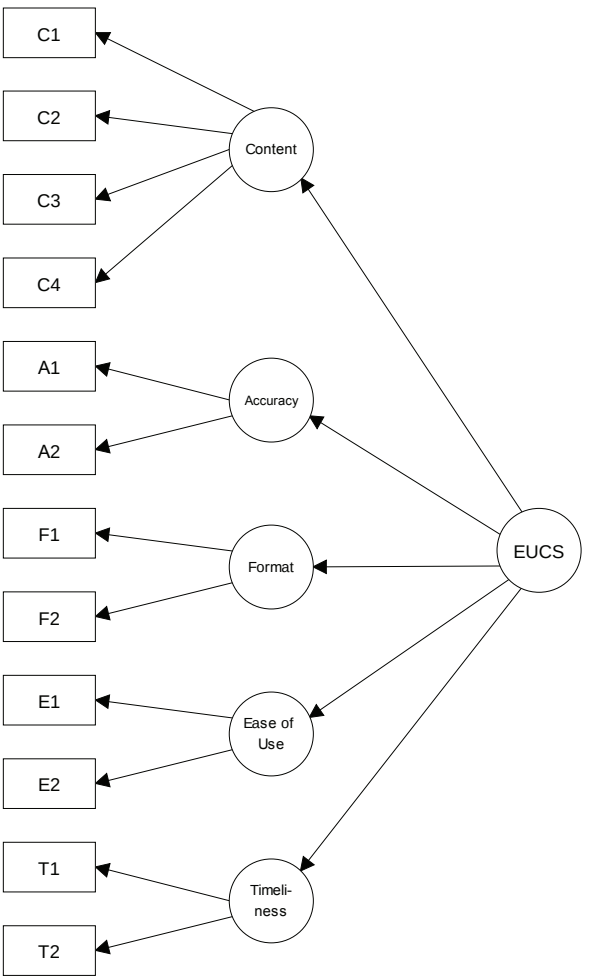
The studies mentioned in the previous section focused on understanding independent variables defining successful representational decision support system applications. A problem with research that considers only the independent side of the equation is the lack of objective measures of whether suggested recommendations correlate with desired outcomes. This dilemma has been problematic throughout information systems research in general and has been the subject of academic debate (Delone & McLean, 1992). DES researchers recognize that the identification of a meaningful, reliable, and robust dependent variable is central to being able to conduct accurate comparisons between competing tools, techniques, software implementations, project approaches, and modeling perspectives.

A wide variety of dependent variables have been investigated in the broader fields of information systems (Delone & McLean, 1992) and decision support systems. Among these, DES researchers have focused on success surrogates and investigated the possibility of assessing DES success or failure. Representational DSS researchers have extended the use of these surrogates.

The first information system success surrogate validated in the context of DES was the end-user computing satisfaction instrument (Doll, Deng, Raghunathan, Torkzadeh, & Xia, 2004; Doll & Torkzadeh, 1988). The EUCS instrument, shown in Figure 1, is of particular interest because most applications of discrete event computer simulation can be categorized as end-user computing. McHaney and Cronan (1998) collected data from 411 participants using a variety of discrete event computer simulation software packages. The analysis indicated the instrument retained its psychometric properties and provided a valid success surrogate for end users beyond the introductory stages of using representational DSSs. The study established the use of information systems surrogate success measures for applied instruments. The results suggest EUCS can reliably and confidently be used in the investigation of competing tools, features, and technologies in the area of DES. Later, EUCS was the subject of a test-retest reliability study that was distributed to users of DES through a mail survey. One month later, follow-up surveys were administered. The original respondents were asked to again evaluate the same system. The two data sets were compared, and the results supported the instrument's internal consistency and stability (McHaney, Hightower, & White, 1999).

Another success surrogate, the technology acceptance model (TAM), was also investigated within the context of representational decision support systems. TAM was developed by Davis (1989) to provide a theoretical explanation of factors influencing technology usage. Davis' theory is derived from the theory of reasoned action (TRA) model (Fishbein & Ajzen, 1975). The TRA model explains actions by identifying connections between various psychological constructs such as attitudes, beliefs, intentions, and behaviors. TRA posits that an individual's attitude toward a given behavior is determined by the belief that this behavior will result in a particular outcome. TAM expands upon this framework

Figure 1. EUCS instrument



C1: Does the simulation system provide the precise information you need?
C2: Does the simulation output information content meet your needs?
C3: Does the simulation provide reports that seem to be just about exactly what you need?
C4: Does the simulation system provide sufficient information?
A1: Is the simulation system accurate?
A2: Are you satisfied with the accuracy of the simulation system?

F1: Do you think simulation output is presented in a useful format?
F2: Is the simulation output information clear?

E1: Is the simulation system user friendly?
E2: Is the simulation system easy to use?

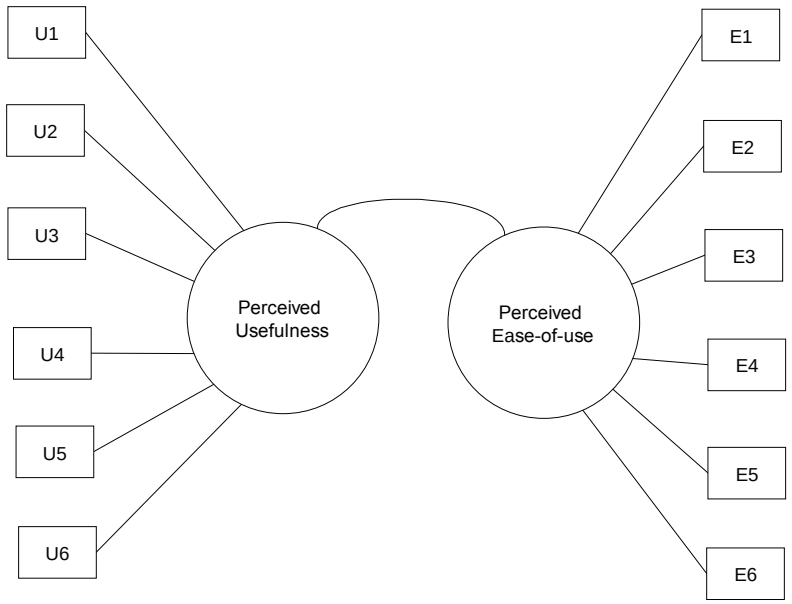
T1: Do you get the output information you need in time?
T2: Does the simulation system provide up-to-date information?

to provide a theoretical explanation of factors influencing technology usage. However, instead of relying on purely attitudinal determinants, TAM hypothesizes that perceived ease of use and perceived usefulness influence a person's intention, which in turn determines actual technology usage. In essence, Davis' research identifies the external variables that influence attitude toward voluntary use of technology. TAM has been validated and tested in a variety of technology-related studies (Adams, Nelson, & Todd, 1992; Davis, Bagozzi, & Warshaw, 1989; Mathieson, 1991). Figure 2 illustrates the Davis Instrument and TAM. McHaney and Cronan (2001) focused on external validity aspects of TAM when applied to users of DES. The findings suggested that TAM provided a reasonable surrogate measure for success within this domain. An updated version of this study was developed into a book chapter (McHaney & Cronan, 2002).

FUTURE TRENDS

Representational decision support systems will continue to become more important to managers, particularly as computing power increases and costs decrease. Web-enabled DES, applications embedded in enterprise computing systems, visualization, and use of DES to fine-tune manufacturing operations and supply chain operations on the fly will become more commonplace. Although questions have been raised concerning the direction and relevance of DSS research in general (Arnott & Pervan, 2005), representational decision support system research should grow in importance. Areas requiring additional work include the search for better dependent variables and related surrogates that enable competing tools and techniques to be evaluated and compared. User interface design and development environments will also

Figure 2. Davis instrument



E1: Learning to operate simulation would be easy for me.
E2: I would find it easy to get simulation to do what I want it to.
E3: My interaction with simulation would be clear and understandable.
E4: I would find simulation to be flexible to interact with.
E5: It would be easy for me to become skillful at using simulation.
E6: I would find simulation easy to use.

U1: Using simulation in my job would enable me to accomplish tasks more quickly.
U2: Using simulation would improve my job performance.
U3: Using simulation in my job would increase my productivity.
U4: Using simulation would enhance my effectiveness on the job.
U5: Using simulation would make it easier to do my job.
U6: I would find simulation useful in my job.

be important. Mainstream DSS and information system research will continue to contribute methodologies and techniques that can be adopted for use in this growing field, and representational DSS systems will find wider user bases as the capabilities offered by simulation are integrated with popular spreadsheets and ERP systems software.

CONCLUSION

Representational decision support systems, particularly as manifested in DES, have become an important tool in a modern manager's decision-making arsenal. In order to determine the effectiveness of these applications, researchers have investigated characteristics of the technology and determined the suitability of various success surrogates to enable comparisons between competing packages, approaches, and methodologies. Future research in this field may determine whether other surrogates are suitable for use with DES and will provide managers with quantitative comparisons to aid in their software selection processes.

REFERENCES

Adams, D. A., Nelson, R. R., & Todd, P.A. (1992). Perceived usefulness, ease of use, and usage of information technology: A replication. *MIS Quarterly, 16*(2), 227-247.

Arnott, D., & Pervan, G. (2005). A critical analysis of decision support systems research. *Journal of Information Technology, 20*(2), 67-87.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(4), 319-340.

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Delone, W. H., & McLean, E. R. (1992). Information success: The quest for the dependent variable. *Information Systems Research, 3*(1), 60-95.

Doll, W. J., Deng, X., Raghunathan, T. S., Torkzadeh, G., & Xia, W. (2004). The meaning and measurement of user satisfaction: A multigroup invariance analysis of the end-user computing satisfaction instrument. *Journal of Management Information Systems, 21*(1), 227-262.

Doll, W. J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly, 12*(2), 259-274.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Guimaraes, T., Igbaria, M., & Lu, M. (1992). The determinants of DSS success: An integrated model. *Decision Sciences, 23*(2), 409-430.

Hlupic, V., & de Vreede, G-J. (2005). Business process modelling using discrete-event simulation: Current opportunities and future challenges. *International Journal of Simulation & Process Modelling, 1*(1/2), 72-81.

Mathieson, K. (1991). Predicting user intentions: Comparing the technology acceptance model with the theory of planned behavior. *Information Systems Research, 2*(3), 173-191.

McHaney, R. W., & Cronan, T. P. (1998). Computer simulation success: On the use of the end-user computing satisfaction instrument. *Decision Sciences, 29*(2), 525-536.

McHaney, R. W., & Cronan, T. P. (2000). Toward an empirical understanding of computer simulation implementation success. *Information & Management, 37*(3), 135-151.

McHaney, R. W., & Cronan, T. P. (2001). A comparison of surrogate success measures in on-going representational decision support systems: An extension to simulation technology. *Journal of End User Computing, 13*(2), 15-25.

McHaney, R. W., & Cronan, T. P. (2002). Success surrogates in representational decision support systems. In M. A. Adams (Ed.), *Advanced topics in end user computing 1* (pp. 243-262). Hershey, PA: Idea Group Publishing.

McHaney, R.W., Hightower, R., & White, D. (1999). EUCS test-retest reliability in representational model decision support systems. *Information & Management, 36*(2), 109-119.

McHaney, R. W., & White, D. (1998). Discrete event simulation software selection: An empirical framework. *Simulation & Gaming, 29*(2), 228-250.

McHaney, R. W., White, D., & Heilman, G. (2002). Simulation project success and failure: Survey findings. *Simulation & Gaming, 33*(1), 49-66.

Robinson, S. (1999). Three sources of simulation inaccuracy (and how to overcome them). In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 Winter Simulation Conference* (pp. 701-708). Piscataway, NJ: IEEE.

Swain, J. J. (2003). Simulation reloaded: Sixth biennial survey of discrete-event software tools. *OR/MS Today, 30*(4), 46-57.

KEY TERMS

Dependent Variable: A value representing the presumed effect or consequence of various states of related independent variables. In other words, a dependent variable is the condition for which an explanation is sought.

Discrete Event Simulation (DES): Use of a computer to mimic the behavior of a complicated system and thereby gain insight into the performance of that system under a variety of circumstances. Generally the system under investigation is viewed in terms of instantaneous changes due to certain sudden events or occurrences.

End-User Computing Satisfaction: A widely accepted information systems success surrogate that measures the degree to which a technology provides the user with a sense of satisfaction that meaningful usage has been affected.

Independent Variable: A value representing a presumed cause of a particular outcome.

Representational Decision Support System: Computer-based information system that combines models with data in a fashion that closely resembles the system that is being studied. Computer simulation is a form of representational decision support system.

Success Surrogate: A proxy for information systems success that takes the form of measurable values. A success surrogate is a dependent variable.

Technology Acceptance Model: A theoretical explanation of factors influencing technology usage that hypothesizes perceived ease of use and perceived usefulness influence a person's intention, which in turn determine actual technology usage.

Theory of Reasoned Action (TRA): This theoretical model explains actions by identifying connections between various psychological constructs such as attitudes, beliefs, intentions, and behaviors, then posits that an individual's attitude toward a given behavior is determined by the belief that this behavior will result in particular outcome.

A Requirement Elicitation Methodology for Global Software Development Teams

Gabriela N. Aranda
Universidad Nacional del Comahue, Argentina

Aurora Vizcaíno
Universidad de Castilla-La Mancha, Spain

Alejandra Cechich
Universidad Nacional del Comahue, Argentina

Mario Piattini
Universidad de Castilla-La Mancha, Spain

INTRODUCTION

Failures during the elicitation process have been usually attributed to the difficulty of the development team in working on a cooperative basis (Togneri, Falbo, & de Menezes, 2002), but today there are other points that have to be considered. In order to save costs, modern software organizations tend to have their software development team geographically distributed, so distance between members becomes one of the most important issues added to the traditional problems of the requirement elicitation process (Brooks, 1987; Loucopoulos & Karakostas, 1995).

So far, literature has widely analysed real life Global Software Development (GSD) projects and pointed out the main problems that affect such environments, especially related to communication. As a complementary view, we have focused our research on analysing how cognitive characteristics can affect people interaction in GSD projects, especially during the requirement elicitation process, where communication becomes crucial.

In this article, we present the main characteristics of requirements elicitation in GSD projects and introduce a cognitive-based requirement elicitation methodology for such environments.

BACKGROUND

Advantages and challenges of GSD have been widely analyzed in literature. As part of the advantages, the most cited are:

- Taking advantage of time difference to extend productive hours (Herbsleb & Moitra, 2001);
- Minimizing development costs (Lloyd, Rosson, & Arthur, 2002);
- Locating developers closer to the customers (Damian & Moitra, 2006); and
- Taking advantage of diversity of stakeholders' knowledge and experiences (Ebert & De Neve, 2001).

On the other hand, the challenges that GSD must face are (Damian & Zowghi, 2002):

- the loss of communicative richness, affected by the lack of face-to-face interaction;
- the time difference between sites, that introduce delays in the project;
- cultural diversity, as a source of misunderstandings; and
- knowledge management, because of the need of maintaining information from many distributed sources.

Looking for solutions to improve communication in GSD, concepts from CSCW (Computer-Supported Cooperative Work) become important because this research area concerns the development of software for enabling communication between cooperating people (*groupware*), that can be simple systems (like e-mail or plain-text chat), more complex ones (like videoconferencing), or the combination of more than one of them. To be more specific, when talking about groupware we follow a convention: We refer to every simple communication technology (e-mail, chat, videoconference) as groupware tools, and to the systems that combine them as groupware packages (Gralla, 1996). Doing so, the most common groupware tools used during multisite developments are e-mails, newsgroups, mailing lists, forums, electronic notice boards, shared whiteboards, document sharing, chat,

instant messaging, and videoconferencing (Damian & Zowghi, 2002; Gralla, 1996).

Another research area related to the distributed requirements elicitation process is Cognitive Informatics (CI), a transdisciplinary research area that encompasses informatics, computer science, software engineering, mathematics, cognition science, neurobiology, psychology and philosophy, and knowledge engineering (Chiew & Wang, 2003). In CI, there is a bidirectional relationship between cognitive sciences and informatics (Wang, 2002):

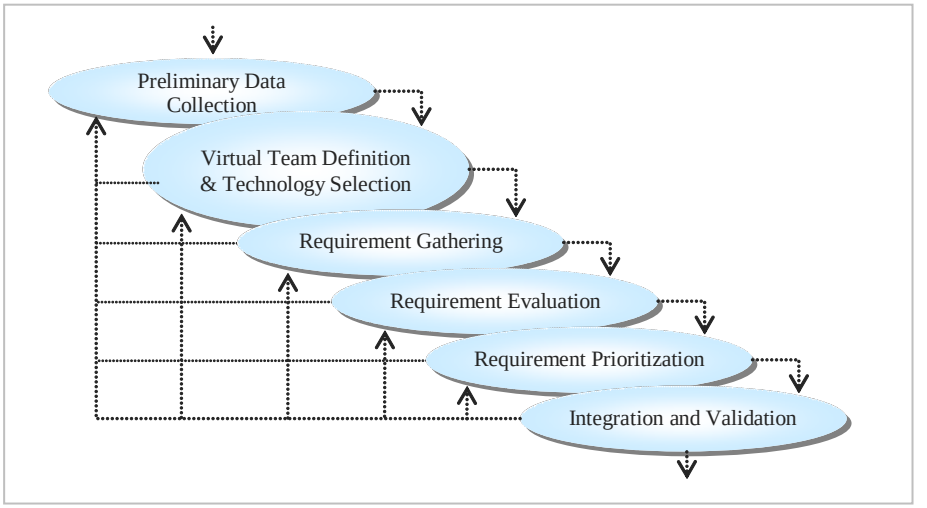
1) using computing techniques to investigate cognitive science problems like memory, learning, and thinking; and
2) using cognitive theories to investigate informatics, computing, and software engineering problems.

In our research, we have followed the second point of view, using concepts from cognitive psychology to improve the requirement elicitation process. Doing so, our research focused on learning styles models (LSMs), a cognitive psychology theory based on Jung's theory of psychological types published in 1921 (Miller & Yin, 2004), that classify people according to the ways they perceive and process information. These models have been discussed in the context of analyzing relationships between instructors and students, but we propose applying them to a virtual team that deals with a distributed requirement elicitation process, considering an analogy between stakeholders and roles in LSM, because during the elicitation process everybody "learns" from others (Martin, Martinez, Martinez, Aranda, & Cechich, 2003), and stakeholders play the role of students or instructors alternatively, depending on the moment or the task they are carrying out.

After analyzing five LSM in Martin et al. (2003), we found out that every item in the other models was included in the model proposed by Felder-Silverman (Felder & Silverman, 1988), so that we may build a complete reference framework choosing this as a foundation. The Felder and Silverman (F-S) model classifies people into four categories, each of them further decomposed into two subcategories (*Sensing – Intuitive*; *Visual – Verbal; Active – Reflective; Sequential – Global*). To know their cognitive profile, people must fill in a multiple-choice test (available at http://www.engr.ncsu.edu/learningstyles/ ilsweb.html), that returns a rank for each category. Depending on the circumstances people may fit into one category or the other, being, for instance, "sometimes" active and "sometimes" reflective, so preference for each category is measured as *strong*, *moderate*, or *mild*.

Most of related works use learning and psychological style models with educational purposes, while few works use them to solve problems in software engineering. One work which uses cognitive styles as a mechanism for software inspection team construction is described in Miller and Yin (2004). They use the MBTI method, an instrument similar to the F-S model. Their intent is different from ours because they use the cognitive styles to set which people seem to be more suitable to work together, while we try to give the best solution (concerning technology) for an already chosen group of people.

Figure 1. RE-GSD methodology



OUR METHODOLOGY: RE-GSD

In order to define the basis for a methodology for requirement elicitation in GSD projects, we have analyzed methodologies used in colocated development and proposed extending them from a cognitive point of view, using the F-S model as a basis for defining a model for technology selection.

We have called our methodology RE-GSD (Requirement Elicitation for Global Software Development projects) and, as a starting point for it, we have selected the models proposed by Christel and Kang (1992) and Hickey and Davis (2003). This selection is due to the fact that both models share a generic view of the selection of requirement elicitation techniques, which fits our intention of defining what to use according to stakeholders' personalities. Both models have been extended and adapted to a distributed environment, so that our methodology can be expressed as follows (see Figure 1):

1. Preliminary data collection: Further decomposed into two categories: (1) about the stakeholders, and (2) about the system and the domain.
2. Virtual team definition & Technology selection: Before starting the requirement gathering, it is important to determine who is going to participate in that stage, because not all the stakeholders in the project are required to participate in every iteration of the elicitation process. Then, the selection of appropriate technology should be carried out, that means, choosing the most appropriate set of requirement elicitation techniques and groupware tools for a given group of people, by taking into account their personal characteristics.
3. Requirement gathering: Once technology has been defined, it is time to apply the requirement elicitation techniques (which are combined with appropriate groupware tools) to obtain a new list of requirements, trying to answer "*what*" is to be built (Christel & Kang, 1992).
4. Requirement evaluation: In this stage, requirements lists must be analyzed in order to determine consistency between different statements.
5. Requirement prioritization: Once requirements are defined, it is important to give them an order of relative importance so as to know when they should be addressed in relation to other requirements (Christel & Kang, 1992). Specially designed tools for distributed requirement inspection (Lanubile, Mallardo, & Calefato, 2003), which allow synchronous and asynchronous discussion, voting, and so forth, can be used to address both this step and the previous one.
6. Requirement integration and validation: In this step, the new requirement list must be integrated to the requirements collected in the previous iterations, looking for inconsistencies also with the system's goals and organizational factors initially defined.

The following sections present the first two phases of our model that are related to our proposal of technology selection according to the stakeholders' cognitive styles.

PHASE 1: PRELIMINARY DATA COLLECTION

a) About the stakeholders
   1) Identify people whose participation is important for the requirement elicitation process, including people from different levels of the organization.
   2) Get personal information about stakeholders, using the form shown in Figure 2. Some important points for distributed environments are, for instance, distinguishing which is the given name and the family name, because different cultures use different order (for instance, in China the family name goes first, while in most of occidental countries, the family name is the last one). Also, recording information about the country of origin and the country of residence is important to identify their mother language and possible differences in cultural background, as well as the foreign languages stakeholders know to choose a second language.
   3) Get information about stakeholders' job, roles, responsibilities and schedules. And because they are distributed, obtaining information about each team member's location (time difference with other sites, work hours, lunch time, etc.) is relevant for other members to know how to contact each other. The form is shown in Figure 3.

b) Get information about the structure, culture, and internal politics of the organization (SWEBOK, 2004), to answer the following questions:
   4) *About the groupware tools*: Which groupware tools are commonly used in the organization? Have stakeholders received training in the use of groupware tools? Which ones do they know better? Which ones have they not used before? Is there any policy that limits the use of groupware tools? Are stakeholders willing to learn how to use other groupware?
   5) *About the requirements elicitation techniques*: Which requirement elicitation techniques are commonly used in the organization? Have stakeholders received training in the use of

Figure 2. Stakeholder's personal information form

Stakeholder's Personal Information Form						
Complete Name (as written in the ID card)						
Given Name		Family Name				
Nickname (optional)						
Birthday						
Gender						
Mother Language		Country of Origin				
		Country of Residence				
Academic degree		University / College		Years of study		
For each foreign language (mark with an X your level of knowledge)	<Language>	low	low-interm	interm	high- interm	high
	Writing					
	Reading					
	Speaking					
Felder and Silverman preferences		Active Reflexive	Sensitive Intuitive	Visual Verbal	Sequential Global	

Figure 3. Stakeholder labour information form

Stakeholder's Labour Information Form						
Role during RE						
Job description	Position		Time in such a position: ……..... years …….. months			
	Place		Time difference (GM)			
Daily timetable	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Arrival time	◷	◷	◷	◷	◷	◷
Dismissal time	◷	◷	◷	◷	◷	◷
Coffee-break(s)	◷	◷	◷	◷	◷	◷
Lunch break	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to
Time I prefer to be called	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to	◷ from ◷ to
Contact information (write the number or user login name)	❍ **Telephone** (number) Country code …….. City code …… Numbers (1) ……. (2) ……. (3) …….			❍ **Fax** (number) Country code …….. City code …… Numbers (1) ……. (2) ……. (3) …….		
	❍ **E-mail** (user name) (1) ……. (2) ……. (3) …….			❍ **Instant messaging** (user name) MSN: ……..……..…. Yahoo messenger: ……..…… Skype: …..……..…… Other: ……...….		
I have also the possibility to use (check)	❍ videoconference ❍ audio conference ❍ others: ……					
If I can choose, I prefer using (put a value of preference between 1 and 10)	…….. e-mail …….. telephone …….. instant messaging …….. discussion forums			…….. shared whiteboards …….. audio conference …….. videoconference		

requirement elicitation techniques? Which ones do they know better? Which ones have they not used before? Are stakeholders willing to learn new techniques?

6) *About the organizational culture*: Do organizational policies allow stakeholders to communicate with others in the virtual team freely, or is there a person that must act as a mediator?

c) About the system and the domain
   7) Get information about the domain and the system in construction.
   8) Determine the system goals.
   9) Identify similar systems

PHASE 2: VIRTUAL TEAM DEFINITION AND TECHNOLOGY SELECTION

As we have mentioned before, we aim at defining strategies that analyze the personal characteristics of stakeholders with the objective of selecting the best groupware tools and requirement elicitation techniques for them. For that reason, selecting appropriate groupware tools is related to our goal of providing the stakeholders with the possibility of communicating with others in a manner closer to the way in which they perceive and process information. That means giving them the chance to feel comfortable with the way in which they interact (synchronously or asynchronously) and the kind of information they interchange (based on words, based on diagrams, etc.).

Even when the technology selection is done for a given virtual team before the elicitation process, obtaining the preference rules is a previous step. Preference rules are obtained from a generic set of people that works on GSD projects, and then the resulting sets can be used in many different GSD projects as well as different virtual teams, at the same time that new information can be added to improve the source of knowledge of the fuzzy logic system.

Obtaining Preference Rules

In order to support personal preferences toward groupware tools, in Aranda, Cechich, Vizcaíno, and Castro-Schez (2004) a model based on fuzzy logic and fuzzy sets to obtain rules from a set of representative examples has been proposed. The obtained rules represent patterns of behavior that indicate the preferences of stakeholders in their daily use of groupware tools and requirements elicitation techniques, according to their classification in the F-S model. To do so, we have collected examples of people preferences and applied a machine learning algorithm. The algorithm we chose (Castro, Castro-Schez, & Zurita, 1999) finds a finite set of fuzzy rules to reproduce the input-output system's behavior. Using this machine learning algorithm over a set of examples that represent the preferences of many stakeholders, we obtained a set of rules (Aranda, Vizcaíno, Cechich, Piattini, & Castro-Schez, 2006). The resulting set of preference rules can be applied to choose the best suite of groupware tools for a group of people by analyzing the results and combine them appropriately, as we will explain in the next section.

The Technology Selection Strategies

According to the analysis of real life projects, analysts are those who choose the techniques for requirement elicitation

Figure 4. Requirement technique selection according to analyst's preferences (π)

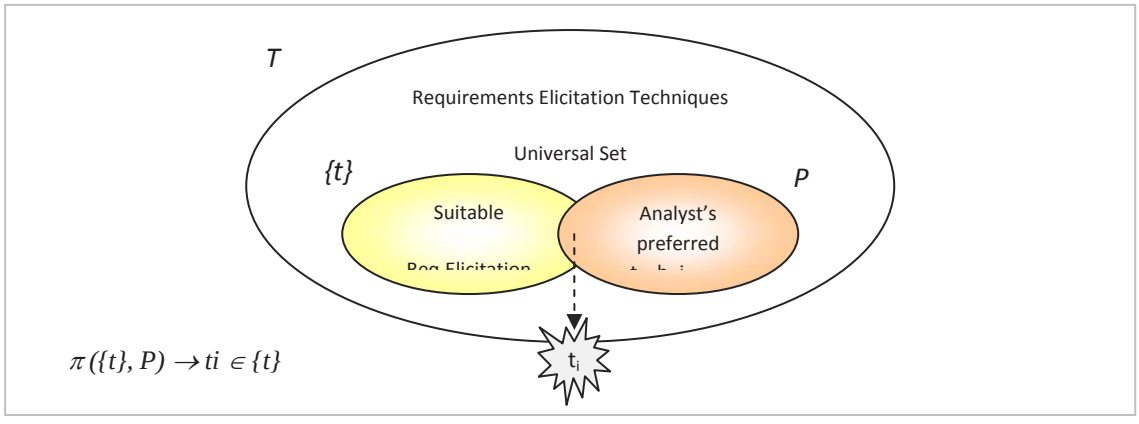


Figure 5. Requirement technique selection according to the most common preference (π^*)

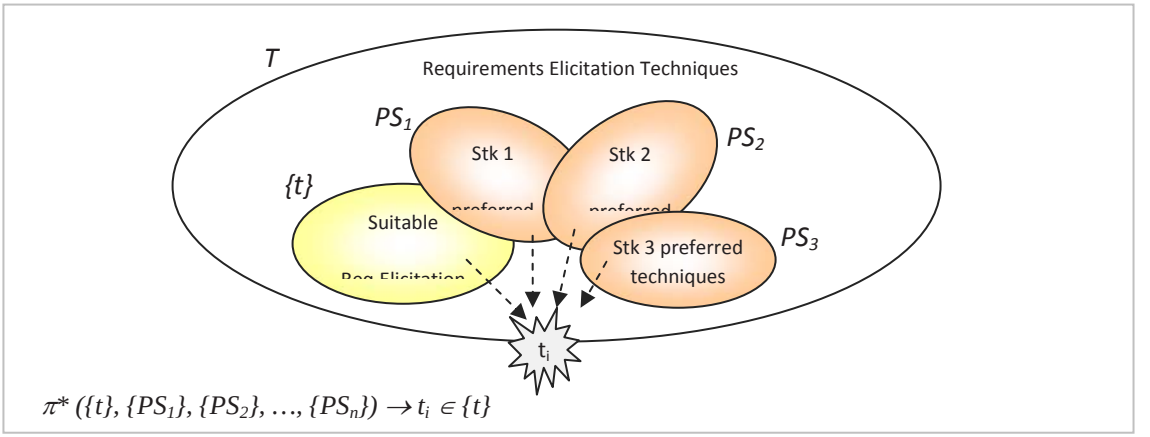
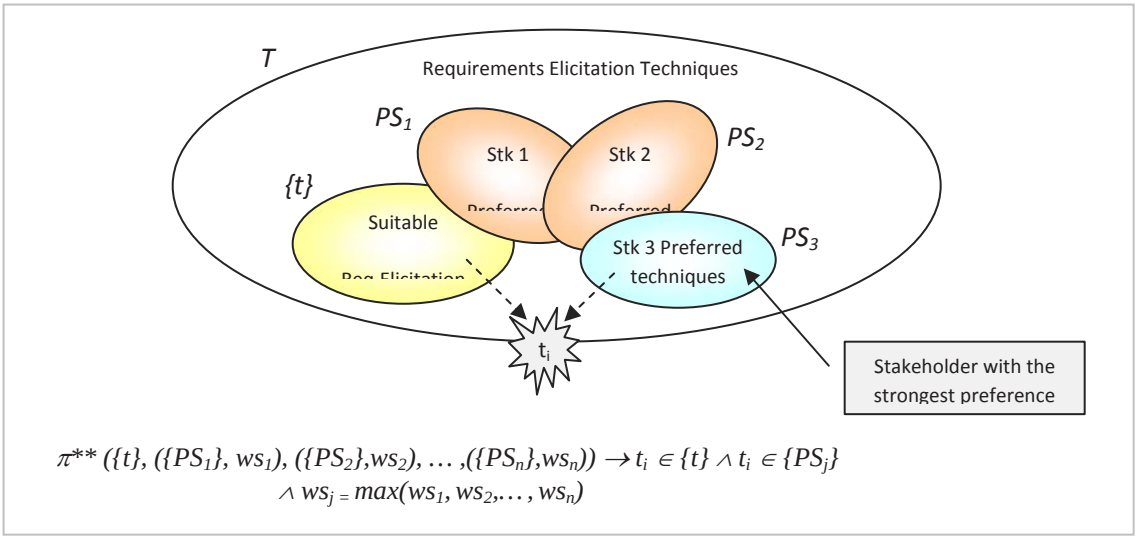


Figure 6. Requirement technique selection according to the strongest preferences in the group (π^{**})

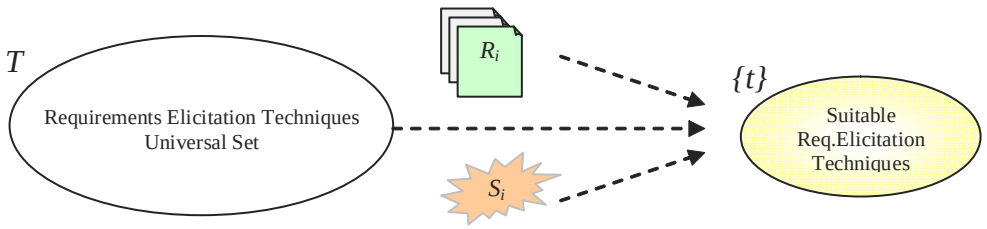


(Hickey & Davis, 2003). To model such a selection, Hickey and Davis' model considers a personal selector function π that returns a technique t_i, from a given a set of techniques *{t}* and a set *P* that represents the personal preferences of the analyst. That means only the analyst's preferences are taken into account, as it is shown in Figure 4.

A first attempt to adapt the previous generic model to our cognitive point of view analyzes the preferences of all the stakeholders and chooses the technique that has more adherents (Aranda, Vizcaíno, Cechich, & Piattini, 2005). This extension of the π function, called π*, is shown in Figure 5. In this formula PS_i represents a set of techniques that fit the *i-th* stakeholder's preferences (defined by the mechanisms based on fuzzy logic and fuzzy sets we have described previously), and $t_i \in \{t\}$ is the technique that appears in most of the PS_i. In this case, analysts are considered without any priority over the rest of the stakeholders.

A later improvement of π* considers the relative importance of stakeholders' preferences by means of weighting them. Its purpose is that, if some stakeholders' preferences are stronger than the rest, the preferences that should be primarily considered are those of the first group of stakeholders. Also, the different weights might be used to prioritize preferences according to stakeholders' roles. The resulting function, called π**, is shown in Figure 6. In this case, ws_i represents the weight (how strong the preferences are), and the resulting t_i is a technique that is appropriate for the current

Figure 7. Definition of the requirements elicitation techniques most suitable for a given situation in a collocated environment (σ)



$\sigma(R_i, S_i, \chi(T)) \rightarrow \{t \in T \mid t \text{ is applicable in situation } S_i \text{ when the current state of requirements is } R_i\}$

situation and is also appropriate for the stakeholder whose personal preferences are the strongest.

PHASE 3: REQUIREMENT GATHERING

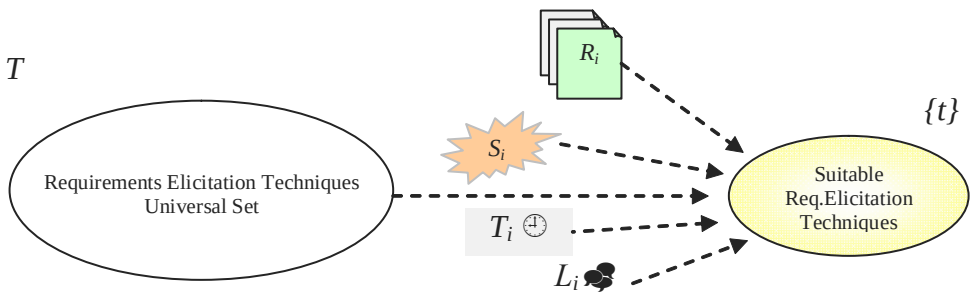
At the moment of deciding how to conduct the requirement gathering, the first step is deciding which of the possible techniques is best, given a current state of knowledge and a particular situation. To do so, the Hickey and Davis model defines the selector function σ (shown in Figure 7).

In a scenario where stakeholders are distributed along many geographically distanced sites, the selector function σ must also consider other aspects. The most important are time difference and the level of knowledge of a common language.

Time difference is important because when timetable does not overlap, or overlaps for a very short period of time, synchronic communication is not possible, then the selection process should prioritize those techniques that work better on asynchronous basis. Similarly, when stakeholders do not share the same language, some of them would need more time to read, think quietly, look for some vocabulary in the dictionary, and so forth.

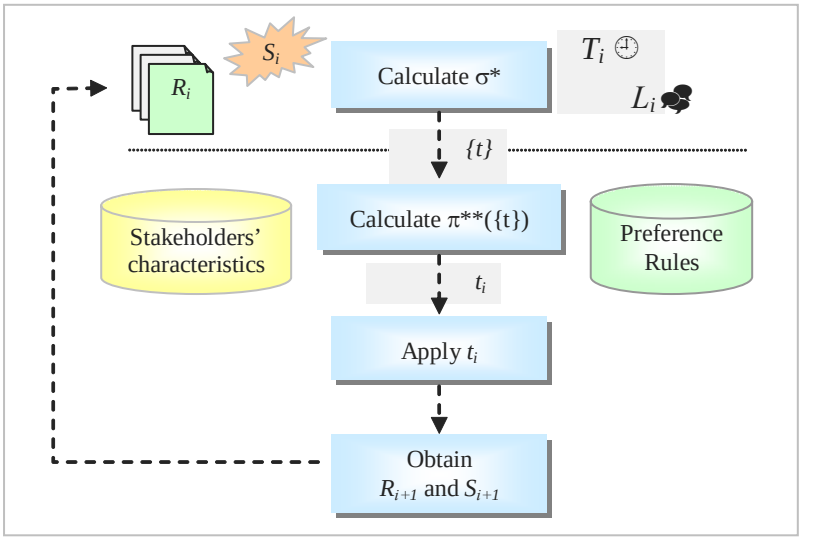
With those considerations in mind, we extended the selector function σ as it is shown in Figure 8, where T_i (time difference) indicates in which degree synchronous communication is possible between the sites that need to interact; and L_i (degree of knowledge of a common language) indicates the level of fluency of communication.

Figure 8. Definition of the requirements elicitation techniques most suitable for a given situation in a distributed environment (σ)*



$\sigma^*(R_i, S_i, \chi(T), T_i, L_i) \rightarrow \{t \in T \mid t \text{ is applicable in situation } S_i \text{ when the current state of requirements is } R_i, \text{ according to restrictions } T_i \text{ and } L_i\}$

Figure 9. Requirement elicitation in GSD, as an iterative process of technique selection and application



Finally, a graphical representation of the requirement elicitation process in a distributed environment is shown in Figure 9.

FUTURE TRENDS

For many years, research on GSD has focused on improving technology as a supporting media for spreading global communication. Those technology topics are really necessary indeed, but not enough for a society challenged by cultural diversity and worldwide-located working teams. Currently, research efforts are looking at GSD as a human-intensive process, where new strategies to deal with socio-cultural differences and distance are bringing psychological and cognitive aspects into the arena. Human factors seem more complex than technological ones, so probably we will see many strategies in the future trying to find the most effective way people understand, choose, and communicate during GSD. Certainly, cognitive informatics will be one of them.

CONCLUSION

GSD projects are a common way of work these days. Looking for solutions to improve communication in virtual environments has led us to analyse human interaction and how to apply well-known techniques from the field of cognitive psychology (called learning style models) as a base for technology selection.

Based on a cognitive-based technology selection approach, we propose a methodology for requirement elicitation in GSD projects, which focuses on improving communication between distant stakeholders. Considering cognitive aspects of stakeholders is significant because the selection of requirement elicitation techniques according to personal characteristics of all the members of a virtual team (instead of just the analyst) might affect positively the quality of the information gathered during the requirement elicitation process. In addition, as stakeholders might feel more comfortable expressing themselves when using a groupware tool closer to the way they perceive and reason about the world, communication in virtual teams is expected to be more fluid and personal satisfaction higher.

In this article, we show the first three phases of such a methodology, called RE-GSD, which are the most clearly different from traditional requirement elicitation methodologies. Current work is focused on defining experiments, in academic and industrial scenarios, to analyze its performance.

ACKNOWLEDGMENT

This work is partially supported by the ENIGMAS (PIB-05-058), and MECENAS (PBI06-0024) project, Junta de Comunidades de Castilla-La Mancha, Consejería de Educación y Ciencia, and the ESFINGE project (TIN2006-15175-C05-05) Ministerio de Educación y Ciencia, Dirección General de Investigación, Fondos Europeos de Desarrollo Regional (FEDER), from Spain. The CompetiSoft project (CyTED 3789); and the 04/E059 project, Universidad Nacional del Comahue, Argentina.

REFERENCES

Aranda, G., Cechich, A., Vizcaíno, A., & Castro-Schez, J. J. (2004). Using fuzzy sets to analyse personal preferences on groupware tools. In *Proceedings of the 10th Argentine Congress of Computer Science, CACIC 2004*, San Justo, Argentina, (pp. 549-560).

Aranda, G., Vizcaíno, A., Cechich, A., & Piattini, M. (2005). A cognitive-based approach to improve distributed requirement elicitation processes. In *Proceedings of the 4th IEEE International Conference on Cognitive Informatics (ICCI'05*), Irvine, USA, (pp. 322-330).

Aranda, G., Vizcaíno, A., Cechich, A., Piattini, M., & Castro-Schez, J. J. (2006). Cognitive-based rules as a means to select suitable groupware tools. In *Proceedings of the 5th IEEE International Conference on Cognitive Informatics (ICCI'06),* Beijing, China, pp.

Brooks, F. P. (1987). No silver bullet: Essence and accidents of software engineering. *IEEE Computer, 20*(4), 10-19.

Castro, J. L., Castro-Schez, J. J., & Zurita, J. M. (1999). Learning maximal structure rules in fuzzy logic for knowledge acquisition in expert systems. *Fuzzy Sets and Systems, 101*(3), 331-342.

Chiew, V., & Wang, Y. (2003). From cognitive psychology to cognitive informatics. In *Proceedings of the Second IEEE International Conference on Cognitive Informatics, ICCI'03*, London, UK, (pp. 114-120).

Christel, M., & Kang, K. (1992). *Issues in requirements elicitation*. Pittsburgh, PA: Carnegie Mellon University.

Damian, D., & Moitra, D. (2006). Guest editors' introduction: Global software development: How far have we come? *IEEE Software, 23*(5), 17-19.

Damian, D., & Zowghi, D. (2002). The impact of stakeholders geographical distribution on managing requirements in a multi-site organization. In *Proceedings of the IEEE Joint International Conference on Requirements Engineering, RE'02*, Essen, Germany, (pp. 319-328).

Ebert, C., & De Neve, P. (2001). Surviving global software development. *IEEE Software, 18*(2), 62-69.

Felder, R., & Silverman, L. (1988). Learning and teaching styles in engineering education. *Engineering Education, 78*(7), 674-681.

Gralla, P. (1996). *How Intranets work*. Emeryville, CA: Ziff-Davis Press.

Herbsleb, J. D., & Moitra, D. (2001). Guest editors' introduction: Global software development. *IEEE Software, 18*(2), 16-20.

Hickey, A. M., & Davis, A. (2003). Requirements elicitation and elicitation technique selection: A model for two knowledge-intensive software development processes. In *Proceedings of the 36th Annual Hawaii International Conference on Systems Sciences (HICSS),* (pp. 96-105).

Lanubile, F., Mallardo, T., & Calefato, F. (2003). Tool support for geographically dispersed inspection teams. *Software Process: Improvement and Practice, Wiley InterScience, 8*(4), 217-231.

Lloyd, W., Rosson, M. B., & Arthur, J. (2002). Effectiveness of elicitation techniques in distributed requirements engineering. In *Proceedings of the 10th Anniversary IEEE Joint International Conference on Requirements Engineering, RE'02*, Essen, Germany, (pp. 311-318).

Loucopoulos, P., & Karakostas, V. (1995). *System requirements engineering*. New York, USA.

Martin, A., Martinez, C., Martinez, N., Aranda, G., & Cechich, A. (2003). Classifying groupware tools to improve communication in geographically distributed elicitation. In *Proceedings of the Ninth Argentine Congress on Computer Science, CACIC 2003*, La Plata, Argentina, (pp. 942-953).

Miller, J., & Yin, Z. (2004). A cognitive-based mechanism for constructing software inspection teams. *IEEE Transactions on Software Engineering, 30*(11), 811-825.

SWEBOK. (2004). *Guide to the software engineering body of knowledge*.

Togneri, D. F., Falbo, R. d. A., & de Menezes, C. S. (2002). Supporting cooperative requirements engineering with an automated tool. In *Proceedings of the Workshop em Engenharia de Requisitos, WER02*, Valencia, España, (pp. 240-254).

Wang, Y. (2002). On cognitive informatics. In *Proceedings of the First IEEE International Conference on Cognitive Informatics, ICCI'02*, Calgary, Alberta, Canada, (pp. 34-42).

KEY TERMS

Cognitive Informatics: It is an interdisciplinary area that applies concepts from psychology and other cognitive sciences to improving processes in engineering disciplines, such as informatics, computing, and software engineering.

CSCW (Computer-supported Cooperative Work): It is a research area that focuses on how people work together and the design and development of software (groupware) that may help their work as a group.

Distributed Software Development: It is a way of developing software that allows the stakeholders to be distributed in geographically distanced sites.

Global Software Development: When the distribution of the members of a distributed software development team exceeds the frontiers of a country.

Groupware: Software that supports and improves group work. It can be a simple text-based technology like e-mail or more sophisticated like videoconferencing.

Learning Style Model: It is a cognitive psychology theory that classifies people according to a set of behavioural characteristics pertaining to the ways they perceive and process information. They can be used to improve the way people learn a given task.

Requirements Elicitation: It is the first stage in the process of understanding the problem the software has to solve. It is crucially based on human communication between the development team and the customer. Other terms that are used as synonyms are "requirements capture," "requirements discovery" and "requirements acquisition."

Stakeholders: They are all the actors that have some interest on a system. They can be the people that pay for it, work on its development or people whose task will be affected by the system, among others.

Requirements Prioritization Techniques

Nadina Martinez Carod
Universidad Nacional del Comahue, Argentina

Alejandra Cechich
Universidad Nacional del Comahue, Argentina

INTRODUCTION

As part of Requirements Engineering, "Elicitation" is the phase where an analyst collects information from the stakeholders, clarifies the problems and the needs of the customers and users, tries to find the best solutions, and makes its planning on what software system will be developed. During elicitation, to get well-defined requirements, a consensus among the different stakeholders is needed. There are several elicitation techniques in the literature; however every technique faces the same problem: each stakeholder has different requirements and priorities, which potentially produces conflicting situations. Therefore, this situation points out Requirements Prioritization as a relevant research area to define the requirements' level of importance.

Nevertheless, often the strategies implemented to solve conflicts among stakeholders are inadequate; for example, weighting requirements can be problematic because sometimes weights are inconsistent and lead to confusion about which are the most essential customer requirements. The prioritizing process must hold stakeholder satisfaction considering high-priority requirements first. However, practical experience shows that prioritizing requirements is not as straightforward task as the Literature suggests. In any case, clearly defining a way of balancing preferences on requirements is essential to the elicitation process.

The remainder of this chapter is structured as follows. Section 2 describes a conceptual framework to describe several prioritization proposals, which are characterized in Section 3. Future trends are presented afterwards.

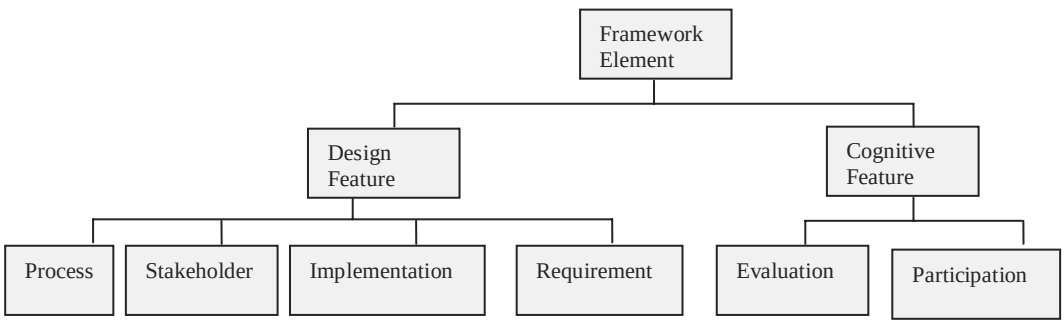
BACKGROUND

Some comparisons of elicitation methods have clarified common features. Firstly, the comparative study by Thomas and Oliveros (2003) is centralized in properties and limitations of five of the most significant methods for eliciting requirements in goal-oriented requirements engineering. This comparison is organized from the viewpoint of goal acquisition with especial emphasis in goal elicitation. Secondly, based on an evaluation framework and influenced by an industrial application (Karlsson & Ryan, 1997), characterizes six different methods for prioritizing software requirements. The objective of Karlsson's evaluation is outlining the methods' behavior for a particular experience, thus the results obtained are not supposed to be generalized by any environment for any application. This evaluation framework is based on inherent characteristics, objective measures and subjective measures.

Our classification framework (Figure 1) is structured into two building blocks – *design features* and *cognitive features*.

The *Design category* is composed of four elements that consider different aspects: *Process, Stakeholders, Implementation* and *Requirements*. The specific features of each

Figure 1. A conceptual framework for comparison



prioritization method are categorized by the *Process* element. It considers answering some questions, such as: Does the process detect inconsistency? Is the process referred to as a systematic or a rigorous process? How we address the problem of dealing with different priorities? Conceptually, is it based on goal decomposition? Does it use a priority or an importance order? The framework also characterizes how prioritizing methods consider *stakeholders*. There are two parameters to be analyzed here: the former refers to the kind of information the method provides with respect to stakeholders. Does the method analyze which stakeholder prioritized a goal, and which priority degree was assigned? The second parameter considers stakeholders geographically distributed.

The *implementation category* depends on the method's scalability and dynamism, that is, usability. It is influenced by how many and which calculus the method uses, and by the performance of the method with a huge number of requirements. It is considerably important whether tools, as well as a reference to spread projects, were applied to support the method. The framework considers information that can demonstrate the method's success in pilot studies. *Requirements* analyze functional (FR) and nonfunctional requirements (NFR) as well as interactions among requirements – *interdependency* represents requirements interaction. Some methods calculate cost and benefit figures for individual requirements, but if there are significant interactions among requirements, the situation becomes more complex. As an example, if two requirements in a method can be achieved by sharing the same solutions to subproblems, then the cost of attaining both of them may be significantly less than the sum of their individual costs. Therefore, the main key is whether the method can handle requirements' interdependencies. The *requirement category* also analyses if the methods deal with functional—FR—and nonfunctional requirements—NFR.

Cognitive aspects cover the evaluation of cognitive features as participation and negotiation among stakeholders during the whole process (Chiew & Wang, 2003). *Evaluation* studies what personal characteristics serve to establish priorities. *Participation* includes defining how priorities were assigned (subjective or objective) from personal experiences and interviews to ensure the success of the developed method. The cognitive aspects no cover the cognitive techniques for knowledge acquisition of knowledge based system.

FEATURES FOR COMPARISON

We can identify two kinds of features: those present in any strategy, and those that may be present or not. Although the first group of features is present in any method, the way these characteristics are maintained is specific of each method. Our appreciation focuses on the second group, which makes the comparison more interesting. By taking into account these concerns, we established three levels of increasing importance to analyze the features: desirable "D", highly desirable "HD" and mandatory "M".

Simple Features

The simple features we considered to analyze processes are:

Consistency: Many times two stakeholders agree on requirements with opposite meanings, which turns impossible the implementation of those requirements. These requirements inconsistencies arise as a result of conflicts between requirements. We consider the action of *detect inconsistencies mandatory* because it is the key of a successful project.

R*igorous:* If a method is *rigorous* and *systematic* it provides robust and comprehensive steps and handling requirements consistently and effectively, which became this feature *highly desirable*. It aids in the validity and verification and it is related intimately to the consistency of requirements.

Goal decomposition: The process based on *goal decomposition* is *desirable* in a prioritization process. The reason is that goals help do not assure successfulness. Clarifying conflicting terms can reduce conflicts, even if the technique does not support goal decomposition.

Priority: Discussion of requirements priorities improves communication between the customer and the developer and helps resolve conflicts. Therefore, we consider *mandatory* the process of deriving an order relation on a given set of requirements, in order to assign a *priority order*, with the ultimate goal of obtaining a shared rationale for partitioning them into subsequent product releases.

Requirements Interdependence: The different occurrences of requirements changes throughout the life cycle points out some dependencies among functional requirements. Understanding these dependencies may improve the requirements process. An approach assumption implies that if two functions are modified due to the same fault report, then there are some *requirements interdependencies* between them. Thus an analysis of such identified fault reports is *desirable* as it may give additional information about requirements.

Objective: One disadvantage detected is that in many methods only one stakeholder has the responsibility of estimating the relative requirements value, which becomes the process subjective. We suppose as *desirable* that the process considers the search of solutions to be as *objective* as possible because the quality requirements always are influenced by analysts' opinions.

Table 1. Characterization in terms of simple features

	Consistency (M)	Rigorous/ Systematic (HD)	Goal Decomposition (D)	Priority (M)		Objectivity (D)
AGORA	By attaching attribute values as preference matrices.	Rigorous process	It uses AND-decomposition and OR-decomposition	Priorities are based on conflicting goals	Only in goal	Attribute values are attached subjectively. But techniques as AHP can be used to obtain more objective values
AHP	By redundancy of pair-wise comparison	Systematic and rigorous method	No	Compares requirements in three hierarchical levels	No	Objective because it represents each term respect to other term.
Cost-Value	By redundancy of pair-wise comparison	Systematic and rigorous method	No	Idem as AHP	No	Idem as AHP
Win-Win	By analyzing the priorities with a Conflict Consultant tool.	Not rigorous or systematic	No	Detects priorities between the requirements	No	Objective because it must reach consensus among the stakeholders
QW-W	Between pairs of requirements (AHP process), eliminating some of them and checking the resulting set.	Systematic process	No	Detects priorities between the requirements	No	It is more objective than Win-Win because it adds a quantitative analysis
ReqInt	Although it detects inconsistencies, it does not have an explicit methodology to correct them.	Not rigorous or systematic	No	Requirement precedence can be given	The process is based on requirements	It is subjective

	Consistency	Rigorous/ Systematic	Goal Decomp.	Priority	Req. Dependence	Objective
QFD	It does not detect inconsistencies.	Not rigorous	No	Precedence can be given because it is based on assigning a numeric value to each requirement	No	Priorities are given subjectively
MPARN	It does not detect inconsistencies.	Systematically negotiated agreements using the multi-criteria preference analysis techniques	No	It considers a precedence for each option through the preference function	No	Although the priorities occur in subjective form they return unfailingly objective

continued on the following page

Table 1. continued

VI	It does not detect inconsistencies.	Nor systematic or rigorous	No	It considers a precedence that can be shared by one or several requirements	No	Priorities are given subjectively
GSP	It does not detect inconsistencies.	Nor systematic or rigorous	Each goal is a node in a goal graph, and is decomposed in OR/AND relationships into sub goals	It considers a precedence when evaluating the alternatives	No	It is subjective. The first part of the process (identification of objectives) can be made by using any elicitation technique
Psy. SR	Although it detects divergence between the stakeholders, it does not detect inconsistencies.	Nor rigorous or systematic	No	The precedence is subjective and is not well defined	No	It is subjective

Table 2. Characterization in terms of compound features

	Traceability (M)	**Distributed Stk (HD)**	**Tools (D)**	**Experience (D)**	**Cognitive aspects (HD)**	**Human experience (D)**	**NFR (D)**
AGORA	It allows to maintain information of objectives prioritized by each stakeholder, using the preference matrix, but not why	No	It is still not supported by computational tools	It has not been used in spread projects. The example proposed is a user accounting system on the Web	None	Although it requires little experience, also requires many interviews	It considers only functional require-ments
AHP	The process involves almost all the stakeholders, so it does not maintain information of who considered each priority or why.	No	An extensive bibliography of reference and several computational tools has been generated	It is applied by main companies and world-wide institutions	None	Although it does not need much experience, it needs several interviews to coordinate the relative values between the stakeholders	Although it is usually used for functional require-ments, it could also be used for non-functional ones.
Cost-Value	It does not maintain information of whom considered each priority or why	No	The second phase of the method is supported by a program written in language C	It was used in several industrial projects	None	Interviews are necessary to coordinate the relative values between the stakeholders and to review the results of the cost-value diagrams	It is adapted for both types of require-ments

continued on the following page

Table 2. continued

	Traceability	Distributed Stakeholders	Tools	Experience	Cognitive aspects	Human experience	NFR
QWW	It is possible to obtain which participants prioritized certain objectives, but not why	No, this method is fed up on the co-participation of the stakeholders to consider new require-ments	Some specific tools not widely used such as. Boehm also created a prototype for his Win-Win spiral model	It was used in spread projects. It is widely used in industry, independently from the domain	None	Although it does not require too much experience, it requires too many interviews	It can be adapted to both types of require-ments
ReqInt	It does not maintain information of who assigned each priority or why	Yes, since stakeholders choose products independently	Parts of the method are supported by tools, nevertheless it does not exist a general software that fully support this methodology	It was used in spread projects, usually in industry	It considers the political status of the stake-holders	It needs experience to make the process successful	It can be adapted to both types of require-ements
QFD	It does not maintain any type of information from the stakeholders	The geometric nature of the process allows working better with isolated groups	This technique is partially supported by tools.	It has been applied successfully from 1991 in the industry of health	It considers the political status of the stake-holders	It needs experience to make the process successful.	It can be adapted to both types of requirements
MPARN	Yes, as in the Win-Win method, it is possible to obtain which participants prioritized certain objectives, but not why. Preference analysis can be a useful tool	No	The MPARN offers supports for generation and negotiation planning, for criteria exploration and assessment of scores and criteria	It does not mention any spread project	None	Similar to the Win-Win method. It does not require too much experience	It can be adapted to both types of requirements
VI	Although the different priorities assigned from each requirement are known, it is not possible to know who assigns each priority or why	Yes, authors are even working to improve this item	Currently working on the elabo-ration of supporting tools, inspired by (DCPT)	It has not been used in real-world projects for case studies	None	Although it does not need much experience, it needs several interviews to negotiate priorities	It is thought for functional requirements

continued on the following page

Table 2. continued

	Traceability	Distrib Stk	Tools	Experience	Cognit. Aspects	Human experience	NFR
GSP	No. As the criteria of all the participants are joined together, it does not register who prioritized each requirement	No	There is no tool yet. It is an on-going project.	It is applied to a case study involving traumatic brain injury patients	Yes, but it does not use it as a weight to mediate.	It needs much experience and many interviews to determine, for each user, goals, skills and preferences	It is developed only for functional require-ments
PsySR	No. As the criteria of all the participants are joined together, it does not register who prioritized each requirement	No.	It does not make calcula-tions of any type. It is not supported by tools	It is used in many small projects, but it is not used in great projects.	It does not consider cognitive characteristics of any of the partici-pants	It does not need much experience, which is obtained in two or three days of training	It can be adapted to both types of require-ments

Compound Features

Our specific set of compound features is:

Traceability: It involves providing techniques and tools for controlling the impact of changes in different parts of the project. Typical changes in requirements specifications include adding or deleting requirements. The process for dealing with requirements changes, as the environments that support this process, is considered *mandatory* because it helps to scope the possible impact of changes.

Distributed stakeholders: Many organizations have adopted a decentralized, team-based, distributed structure, whose members communicate and coordinate their work through information technology. As the groups are several and heterogeneous, the process which support distributed stakeholders, allowing powerful ways of communication and allowing groups to develop distributed software engineering activities, is highly desirable.

Computational tools: Is our intention that computational tools support the prioritization process; the argument for that is that sometimes this feature avoids paralyzing the process or making the process not too hard.

Experience: The process must be proved in real projects, which implies practical experience. At least this is a *desirable* item since many processes are good theoretically but they are practically impossible to be implemented.

Cognitive aspects: Cognitive techniques may be satisfactory used in personal requirements evaluation. Therefore, we consider *highly desirable* the evaluation of *cognitive aspects* to establish priorities weights in order to solve requirements conflicts.

Human experience: Requirements specialists are generally more familiar than other development staff with recent technology advances and also can help eliciting real customer needs and expectations. Sometimes developers have not enough experience, (or the ones which have enough experience are too expensive for specific projects). Anyway, it would be *desirable* to have methods or processes, which imply less *human experience*.

NFR: The software requirements specification serves as a container for both the functional requirements and the nonfunctional requirements. It is *desirable* that quality attributes – *nonfunctional requirements* such as usability, efficiency, portability, and maintainability – can be elicited from users during the prioritization process.

Characterizing Approaches

Nowadays, a broad spectrum of elicitation techniques are practiced in different software development projects (Hickey & Davis, 2003; Leoucopoulos & Karakostas, 1995; Young, 2002). To reduce conflicting situations, methods as (Boehm, Grünbacher & Briggs, 2001; Grüenbacher, 2000; Ruhe, Eberlein & Pfahl, 2002) must negotiate the "right requirements".

Systematic methods, such as the AHP, and the Cost-Value (Karlsson & Ryan, 1997; Saaty, 1990), have received some interest in the application of elicitation procedures, and simpler decision-making techniques (In, Olson & Rodgers, 2001), or visualization techniques (In & Roy, 2001) have

been found out to be appropriate to resolve disagreements promoting a cost-effective use.

On the other hand, the requirements elicitation techniques have widely used a family of goal-oriented requirements analysis (GORA) methods (Antón, 1996; Dardenne, van Lamsweerde & Fickas, 1993; GRL homepage, I* homepage, KAOS homepage) as approaches to refine and decompose the needs of customers into more concrete goals that should be achieved. Particularly, a proposal called AGORA (Kaiya, Horai & Saeki, 2002) extends a version of a Goal-Oriented Requirements Analysis Method by considering detecting and resolving conflicts on goals; the work considers greater priority when there exists a dependency between requirements, and these interdependencies can be identified before they are negotiated More recently, the Goals-Skills-Preferences Framework (Hui, Lisakos & Mylopoulos, 2003) is used to generate a customizable software design; or techniques from Cognitive Informatics try to find solutions to communication problems during all stages of software engineering.

Table 1 characterizes the following proposals in terms of the framework's simple features we have introduced in section 2: Attributed Goal-Oriented Requirements Analysis Method –*AGORA*- (Kaiya et al., 2002), Analytical Hierarchy Process –*AHP*- (Saaty, 1990), *Cost-Value Approach* (Karlsson & Ryan, 1997), *Win-Win* (Grüenbacher, 2000), Quantitative Win-Win –*QWW*- (Ruhe et al., 2002), Requirements Interdependencies –*ReqInt*- (Giesen & Völker, 2002), Quality Function Deployment method –*QFD*- (Dean, 1992), Multi-Criteria Preference Analysis Requirements Negotiation –*MPARN*- (In et al., 2001), Visualization technique –*VT*- (In & Roy, 2001), Goals-Skills-Preference –*GSP*- (Hui et al., 2003), Psychotherapy for System Requirements -*PsySR*- (Goetz & Rupp, 2003; Rupp, 2002)

Table 2 characterizes the same proposals in terms of the framework's compound features we have introduced in section 2.

FUTURE TRENDS

Nowadays, requirements prioritization techniques give much importance to minimize personal efforts maintaining an agreement between stakeholders. However, we have identified a number of interesting issues that can be the basis for more targeted research projects. The tendency is to systematize as much as possible, with computational tools, the activities implied in each technique. The goal is obtain a set of requirements strong and consistent enough to avoid future pitfalls. Currently, some of the proposals are improving to work well in a distributed setting. Generally, the approaches use cognitive aspects only during the negotiation phase, where the analyst must reach mutual consensus, but the assignment of cognitive weights to each stakeholder may help assess a candidate group to be involved in prioritizing a set of requirements. Some research efforts on this line have introduced learning style models to improve communication and perception of goals (Martinez Carod & Cechich 2007), showing promissory results.

CONCLUSION

From *Table 1* and *Table 2* we can observe there is not a complete, simple, fast and reliable prioritizing approach. Neither of them provides specific tools to solve conflicts. Some approaches as Goals-Skill and Preferences (GSP) and AGORA are based on goals, others such as the Win-Win, Quantitative Win-Win and Visualization Issue technique, on negotiation processes, and some others such as QDF and MPARN are based on industrial and decision-making techniques. On the other hand, AHP and Cost-Value are based on pair wise comparison, and the Psychotherapy for System Requirements method is based on human interaction using natural language and is the only method that cannot establish priorities between requirements. AGORA, as the methods based on negotiation processes, can detect such inconsistencies. In these methods, we can see both win conditions and candidate requirements as initial goals. Considering this aspect, only the GSP and AGORA approaches allow decomposition from needs of the customers into subgoals. Although both AHP as Quantitative Win-Win are reliable, they require a large number of mathematical calculations to prioritize few requirements. Only the Psychotherapy from System Requirements takes cognitive aspects into account allowing people specify what they really mean. However, it is not a formal or systematic method.

Now, we focus our attention on both mandatory and highly desirable features. Even *consistency checking* is considered the most important feature, only few of them (those based on negotiation, AGORA and ReqInt) provide support to it. Besides, only systematic methods (such as AGORA, AHP, and MPARN) and Win Win, ReqInt and GSP have the ability to define priorities among requirements. Analyzing *traceability,* AGORA, Win Win, MPARN, VT have the property of knowing whose participants prioritized certain objectives, but neither of them maintains the information about requirements changes. Respect of *distributed Stakeholders,* AGORA, AHP, MPARN, GSP and PsySR do not provide this characteristic; in particular there are some proposals to incorporate geographically distributed participants to AGORA. Only the PsySR takes *cognitive aspects* into account allowing people specify what they really mean, not from the viewpoint of assigning cognitive weights to each stakeholder, but for knowledge acquisition. Only AHP and MPARN are essentially systematic and rigorous.

The comparison framework introduced in this chapter allowed us to analyze interesting features of most representative requirements prioritization techniques, revealing their strengths and weaknesses. Hope our work will help clarify the field and identify opportunities for improvement.

REFERENCES

Antón, A.(1996). Goal based requirements analysis. In *Proceedings of the 2nd International Conference on Requirements Engineering (ICRE '96) IEEE Software* April 15 - 18, 1996.

Boehm, B. W., Grünbacher, P., & Briggs, B. (2001). *Developing groupware for fequirements negotiation: Lessons learned*. IEEE Software, May/June 2001, pp. 46-55.

Chiew, V. & Wang, Y. (2003). From cognitive psychology to cognitive informatics. In *Proceedings of the Second IEEE International Conference on Cognitive Informatics (ICCI'03)*(pp. 114-120). London, UK.

Dardenne, A., van Lamsweerde A., & Fickas, S. (1993). Goal-directed requirements acquisition. *Science of Computer Programming, 20*, 3-50.

Giesen, J. & Völker, A. (2002). Requirements interdependencies and stakeholders preferences. *IEEE Joint International Conference on Requirements Engineering (RE'02)* (pp. 206-212).

Goetz, R. & Rupp, C. Psychotherapy for system requirements. In *Proceedings of Second IEEE International Conference on Cognitive Informatics (ICCI' 03)*.

GRL homepage, http://www.cs.toronto.edu/k-m/GRL/

Grüenbacher, P. (2000). Collaborative requirements negotiation with EasyWinWin. In *Proceedings of the 2nd International Workshop on the Requirements Engineering Process, Greenwich, London IEEE Computer Society* (pp. 954-690).

Hickey, A. M. & Davis, A. M. Requirements elicitation and elicitation technique selection: A model for two knowledge-intensive software development processes. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*.

Hui, B., Lisakos, S., & Mylopoulos, J. (2003). Requirements analysis for customizable software: A goals-skills-preferences framework. In *Proceedings of the 11th IEEE International Requirements Engineering Conference* (pp. 117–126)

I* homepage, http://www.cs.toronto.edu/km/istar

In H., Olson, D., & Rodgers, T. (2001). A requirements negotiation model based on multi-criteria analysis In *Proceedings of the 5th IEEE International Symposium on Requirements Engineering (RE '01)* (p. 312). Toronto, Canada.

In, H. & Roy, S. (2001). Visualization issues for software requirements negotiation. *In Proceedings of the IEEE International Computer Software and Applications Conference (COMPSAC 2001)* (pp. 10-15). Chicago, Illinois.

Kaiya, H., Horai, H., & Saeki, M. (2002). AGORA: Attributed goal-oriented requirements analysis method. In *Proceedings of the IEEE International Conference on Requirements Engineering* (pp. 13-22).

KAOS homepage, http:www.info.ucl.ac.be/research/projects/AVL/ReqEng.html

Karlsson, J. & Ryan, K. (1997). A cost-value approach for prioritizing requirements. *IEEE Software, 14*(5), 67-74.

Leoucopoulos, P. & Karakostas, V. (1995). *System requirements engineering.* Mc Graw-Hill.

Martinez Carod, N. & Cechich, A. (2007). A cognitive psychology approach for balancing elicitation goals. In *Proceeding of the Sixth IEEE International Conference on Cognitive Informatics (ICCI'07)*, California.

Ruhe, G., Eberlein, A., & Pfahl, D. (2002). *Quantitative WinWin - A quantitative method for decision support in requirements negotiation.* Germany: Fraunhofer IESE.

Rupp, C. (2002). *Requirements and psychology*. IEEE (Softwarepp.16-18

Saaty, T. L. (1990). *The analytic hierarchy process*. McGraw-Hill.

Thomas, P. & Oliveros, A. (2003). *Elicitación de Objetivos, un estudio comparativo"*. IX Congreso Argentino en Ciencias de la Computación, CACIC 2003, La Plata, 6-10 Octubre 2003, pp.990-1002

Young, R. (2002). Recommended requirements gathering practices. *CroossTalk The Journal of Defense Software Engineering*. pp. 9 -12

KEY TERMS

Candidate Requirements: A set of requirements which probably requirements engineer would decide to be implemented first.

Cognitive Features: The use of cognitive sciences such cognitive psychology and cognitive informatics to characterize people according to some personal characteristics or cognitive profiles.

Conflicts Between Requirements: Are conflicts between stakeholders on satisfactory level of a requirement, imprecise requirements or opposite candidate requirements.

Distributed Stakeholders: Participants geographically situated in distant places that are involved in a software

development project.

Goal Decomposition: When a goal can be expressed in terms of logical combinations of subgoals.

Negotiation Process: When at least two people groups interchange opinions in order to resolve conflicts between stakeholders with satisfaction for each group.

Nonfunctional Requirements: There are quality attribute describing systems needs, such as performance, safety, security, traceability.

Requirements Interdependencies: When one requirement depends in a big percentage on the implementation of others requirements.

Requirements Traceability: The ability of manage all concern about a requirement. The reason for its existence, which constraints has the implementation of it, which dependent requirement has, the person who has defended it and so on.

Rigorous Method: A method which has consistent and well defined steps to follow.

Category: Business Information Systems

Researching Technological Innovation in Small Business

Arthur Tatnall
Victoria University, Australia

INTRODUCTION

The introduction of a new information system into a small business, or upgrading an existing system, should be seen as an innovation and considered through the lens of innovation theory. The most widely accepted theories of how technological innovation takes place are provided by innovation diffusion (Rogers, 1995) and the technology acceptance model (Davis, 1986), but most of the research based on these models involves studies of large organizations or societal groups. This article argues that another approach, innovation translation, has more to offer in the case of innovations that take place in smaller organizations (Burgess, Tatnall, & Darbyshire, 1999; Tatnall, 2002; Tatnall & Burgess, 2004).

BACKGROUND

There are important differences in the processes by which small and large enterprises choose to adopt or reject computers (Tatnall, 2002, 2005a), and this article concerns itself only with issues related to small business. To begin, however, it is important to distinguish between invention and innovation. Whereas invention can be seen in the discovery or creation of new ideas, innovation involves putting these ideas into commercial or organizational practice (Maguire, Kazlauskas, & Weir, 1994). Invention does not necessarily invoke innovation, and it fallacious to think that invention is necessary and sufficient for innovation to occur (Tatnall, 2005b).

Changing the way things are done is a complex affair (Machiavelli, 1515) and one that is difficult to achieve successfully. The dominant paradigm, by far, in innovation research is that of *innovation diffusion*, and no discussion would be complete without consideration of this approach. Innovation diffusion has had success in describing how innovations diffuse through large populations (Rogers, 1995). There are occasions, however, when diffusion does not occur, and the diffusion model finds these difficult to explain (Latour, 1996). Another common approach is to use the technology acceptance model proposed by Davis, Bagozzi, & Warshaw (1989) that looks at user perceptions of technology as a basis for adoption or non-adoption. The approach offered in *innovation translation*, informed by actor-network theory (ANT), is also worthy of consideration. In the translation model the key to innovation is creating a powerful enough consortium of actors to carry it through, and when an innovation fails, this can be considered to reflect on the inability of those involved to construct the necessary network of alliances amongst the other actors. This article will compare these models of technological innovation.

INNOVATION DIFFUSION

Rogers (1995), perhaps its most influential advocate, approaches the topic of innovation diffusion by considering a variety of case studies, the prime concern of which is the identification of factors that affect the speed with which an innovation is adopted, or that cause it not to be adopted at all.

In diffusion theory the existence of an innovation is seen to cause uncertainty in the minds of potential adopters causing a lack of predictability and of information. Rogers (1995) asserts that a technological innovation embodies information and that this has the potential to reduce uncertainty. Diffusion is thus considered to be an information exchange process amongst members of a communicating social network driven by the need to reduce uncertainty (Lepa & Tatnall, 2002). There are four main elements of the theory of innovation diffusion (Rogers, 1995):

Characteristics of the Innovation Itself

Rogers argues that the attributes and characteristics of the innovation are important in determining the manner of its diffusion and the rate of its adoption and outlines five important characteristics of an innovation that affect its diffusion: relative advantage, compatibility, complexity, trialability, and observability. The attributes of the potential adopter are also seen as an important consideration, and Rogers maintains that these include social status, level of education, degree of cosmopolitanism, and amount of innovativeness.

Nature of the Communications Channels

Acts of communication are a necessary part of any change process, and to reach a potential adopter the innovation must be diffused through some communications channel. Channels

involving mass media are the most rapid and efficient means of spreading awareness of an innovation, but interpersonal channels are generally more effective in persuading someone to accept a new idea.

The Passage of Time

In common with earlier researchers, Rogers found that different individuals in a social system do not necessarily adopt an innovation at the same time. Borrowing from work by Deutschmann and Fals Borda (1962), he proposes that adopters can be classified by their degree of "innovativeness" into five categories—innovators, early adopters, early majority, late majority, and laggards—and that if the number of individuals adopting a new idea is plotted over time, it usually follows a normal curve.

The Social System

Diffusion occurs within a social system in which the structure constitutes a boundary inside which this diffuses. Rogers argues that the system's social structure affects diffusion through the action of social norms, the roles taken by opinion leaders and change agents, the types of innovation decisions that are taken, and the social consequences of the innovation.

TECHNOLOGY ACCEPTANCE MODEL (TAM)

The main goal of TAM is "to provide an explanation of the determinants of computer acceptance that is general, and capable of explaining user behavior across a broad range of end-user computing technologies and user populations, while at the same time being both parsimonious and theoretically justified" (Davis et al., 1989, p. 985). Davis's (1986) conceptual framework proposed that a user's motivational factors are related to actual technology usage and hence act as a bridge between technology design (e.g., system features and capabilities) and actual technology usage. In his conceptual framework, Davis (1986) assumes that stimulus variables (e.g., system features and capabilities) trigger organism factors (e.g., user motivation to use the technology), and in turn users respond by actually using the technology. Davis identifies the following major determinants of technology acceptance:

- Perceived usefulness
- Perceived ease of use

INNOVATION TRANSLATION

An alternative view is that of innovation translation, which draws on the sociology of translations, more commonly known as actor-network theory (ANT). The core of the actor-network approach is translation (Law, 1992), which can be defined as "... the means by which one entity gives a role to others" (Singleton & Michael, 1993, p. 229).

Essentialism

Diffusion theory asserts that a technological innovation embodies "information": some essential capacity or "essence" instrumental in determining its rate of adoption. A significant problem with an essentialist paradigm like this arises if a researcher tries to reconcile the views of all parties involved in the innovation on what *particular* essences are significant. The difficulty is that people often see *different* "essential attributes" in any specific technological or human entity, making it hard to identify and settle on the ones that allegedly were responsible for the diffusion.

To illustrate this difficulty, consider the case of a small business deciding whether to purchase their first computer. Researchers using an innovation diffusion model would begin by looking for innate characteristics of the PC that would make a potential adopter more likely to accept it. They would consider the relative advantages of a PC over alternatives like a filing-cabinet. An examination of the compatibility, trialability, and observability of a PC with this older office technology would show good reasons for acceptance. An examination of the PC's complexity would, however, bring out some reasons for reluctance in its adoption. The researchers would then investigate characteristics of the potential adopters, considering factors like their educational background, innovativeness, and how they heard about the innovation. If, however, you *ask* small business people why they purchased their first PC, the answers often do not match with this view.

Actor-Network Theory: The Sociology of Translations

Rather than recognizing in advance the essences of humans and of social organizations and distinguishing their actions from the inanimate behavior of technological and natural objects, ANT adopts an anti-essentialist position in which it rejects there being some difference in essence between humans and non-humans. ANT considers both social and technical determinism to be flawed and proposes instead a socio-technical account (Latour, 1986) in which neither social nor technical positions are privileged. To address the need

to properly consider the contributions of both human and non-human actors, actor-network theory attempts impartiality toward all actors in consideration, whether human or non-human, and makes no distinction in approach between the social, the natural, and the technological (Callon, 1986).

Mechanisms of Translation

The process of translation has four aspects or "moments" (Callon, 1986), the first of which is known as *problematization*. In this stage, a group of one or more key actors attempts to define the nature of the problem and the roles of other actors so that these key actors are seen as having the answer and being indispensable to the solution of the problem. In other words, the problem is re-defined (translated) in terms of solutions offered by these actors (Bloomfield & Best, 1992). The second moment is *interessement* and is a series of processes that attempt to impose the identities and roles defined in the problematization on the other actors. It means interesting and attracting an entity by coming between it and some other entity. Here the enrollers attempt to lock the other actors into the roles proposed for them (Callon, 1986) and to gradually dissolve existing networks, replacing them with a network created by the enrollers themselves.

If the interessement is successful, then the third moment, *enrollment*, will follow through a process of coercion, seduction, or consent (Grint & Woolgar, 1997), leading to the establishment of a solid, stable network of alliances. Enrollment, however, involves more than just one set of actors imposing their will on others; it also requires these others to yield (Singleton & Michael, 1993). Finally, *mobilization* occurs as the proposed solution gains wider acceptance and an even larger network of absent entities is created (Grint & Woolgar, 1997) through some actors acting as spokespersons for others.

An Example: Adoption of a Slide Scanner by a Small Publishing Company

To illustrate the use of the two innovation models, consider the case of DP Pty Ltd—a small publishing company where four people work on the publication of information systems textbooks. DP is a very small business with a well established market. Members of the company do much of the writing and all of the work involved in publication of their books but send the work off for printing and binding. Most of DP's print runs are small. All those involved in the work of the company are computer literate and make good use of IT. None of them, however, has much knowledge of computer graphics.

Several years ago the company decided it needed to improve the appearance of the covers on its textbooks. Several options were considered until someone thought of using a photograph. The brother of one of the directors is a landscape photographer who was able to provide a suitable color photograph. This was supplied in print form, and a problem arose in how to convert it (or its negative) into a suitable format to print on the cover along with the cover text. A digital image seemed to be the answer. The photograph was scanned (using a flat-bed scanner) and the digital image inserted into Microsoft Word so that text could easily be added. The final result was then converted into an Acrobat file and sent off to the printer. This process, however, proved to be quite a bother. Today the company makes use of a Nikon slide and negative scanner, considerably improving the quality of the covers, but also making the process of producing them much simpler. This device is capable of producing digital images of slides and negatives at various resolutions. The question is: why did DP decide to adopt this *particular* item of technology?

Consider first the application of a diffusion model. Here the company would have been mainly influenced by attributes and characteristics of the technology itself. The directors would have considered the relative advantage, compatibility, complexity, trialability, and observability of this technology compared with the alternatives. Two of the directors of the company certainly did see some relative advantage in using the slide scanner, particularly as they both had large numbers of color slides that they had taken over the years. There was, however, one major disadvantage of this technology, and that was its high cost: the slide scanner was around four times as expensive as a good flat-bed scanner. The slide scanner did not come out well on compatibility or complexity, as it was quite different and difficult to use. The company arranged trial use of the scanner, which was lucky as it proved difficult to find anyone else using one—its observability was low. On this basis it is difficult to see why DP would have adopted the slide scanner at all.

When a translation model is applied, the situation is seen quite differently. The socio-technical network consisting of the publishing company personnel, their computers, and their books was destabilized by the need to find a new way of producing book covers. The addition of a new actor, the landscape photographer, introduced new possibilities that worked to further destabilize this network. The slide scanner (also seen as an actor seeking to enter the network) offered new possibilities in which existing (and future) slides and negatives could easily be turned into digital images. This included *any* of the directors' old slides, not just those required for book covers. As well as the main application of producing book covers, both directors quickly saw advantages in a device that could also easily convert the old slides and negatives they had each taken of their children and of their holidays into digital format. It was thus a combination of factors, some business-related and others rather more personal, that the translation modes suggests could be seen as leading to the adoption of this technology.

Table 1. Innovation diffusion versus innovation translation (Adapted from McMaster, Vidgen, & Wastell, 1997)

	Innovation Diffusion	**Innovation Translation**
Innovation	A technology perceived to be new by the potential adopter.	A technology that has yet to be "black-boxed."
Communication	Communication channels can be categorized as cosmopolite or localite, and mass media or interpersonal. Innovations are transferred through these channels.	Translations are made by actors in enrolling the innovation.
Time	Speed of decision to innovate, earliness of adoption, and rate of adoption are important.	Network dynamics in enrollment, control, and dissemination are what matter.
The Social System	Homophily vs. heterophily. Sharing of interests of human actors.	Interessement between actants, both human and non-human, and goals. Black boxes form when interests move in the same direction.
The Technology	Changes are made to the form and content of the technology as a result of experiences during implementation (re-invention).	The technology is translated through being enrolled, regardless of whether its form or content is modified.
Socio-Technical Stance	The social system and the technology are separate. Diffusion is the adoption of technology by a social system. Technology transfer requires the bringing together of social and technical elements.	The social system and the technology are inseparable. Successful innovation and technology transfer give the appearance of separation, but this is merely evidence that the actor-network has stabilized.

It should also be pointed out that a significant translation occurred from the slide scanner advertised by Nikon in the magazine article and on their Web page to the device adopted by the publishing company. DP was not interested in using the scanner in all the ways that Nikon offered. They wanted a device to digitize slides and negatives for two reasons: the easy creation of attractive book covers and the conversion of their own color slides into a format that could easily be displayed on a video monitor. They did not adopt the scanner advertised by Nikon, but a device for creating book covers and formatting their slides for video display.

FUTURE DIRECTIONS: TRANSLATION VS. DIFFUSION

Many small businesses are family-related concerns with several members working in the business. Investigations (Tatnall, 2002; Tatnall & Burgess, 2002; Tatnall & Pliaskin, 2005) suggest that a common reason why a small business first acquired a PC is that it was, at least partly, intended for family use. In this regard, one reason that the PC was obtained was in the belief that it would assist with their children's education. Once obtained, other uses are almost always also found for the technology, but the question remains: would the characteristics of the PC and of the people involved in its adoption have been identified by a diffusion model? Table 1 summarizes differences in approaches between the diffusion and translation frameworks.

CONCLUSION

The translation approach to innovation details the construction of networks and alliances to support and embed changes in order to make them durable. Innovation diffusion and innovation translation are based on quite different philosophies, one stressing the properties of the innovation and the change agent and routed in essentialism, and the other emphasizing negotiations and network formation.

Innovation translation is more attuned to the human and political issues involved in small business decision making, and so offers a useful approach to modelling innovation in small business. Most current writings on technology uptake and innovation employ either no theoretical model at all, or else use a diffusion model. The author suggests that future research should further investigate the use of a translation model to explain the adoption of information technology in small business and encourages future researchers to consider this approach.

REFERENCES

Bloomfield, B. P., & Best, A. (1992). Management consultants: Systems development, power and the translation of problems. *The Sociological Review, 40*(3), 533-560.

Burgess, S., Tatnall, A., & Darbyshire, P. (1999). *Teaching small business entrepreneurs about computers*. EuroPME—Entrepreneurship: Building for the Future, Rennes, France, Groupe ESC, Rennes.

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the rishermen of St. Brieuc Bay. In J. Law (Ed.), *Power, action & belief. A new sociology of knowledge?* (pp. 196-229). London: Routledge & Kegan Paul.

Davis, F. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results. PhD*. Boston: MIT.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 10*(3), 318-340.

Davis, F. D., Bagozzi, R., & Warshaw, P. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1003.

Deutschmann, P. J., & Fals Borda, O. (1962). *Communication and adoption patterns in an Andean village*. San Jose, Costa Rica: Programa Interamericano de Informacion Popular.

Grint, K., & Woolgar, S. (1997). *The machine at work—Technology, work and organisation*. Cambridge: Polity Press.

Latour, B. (1986). The powers of association. In J. Law (Ed.), *Power, action and belief. A new sociology of knowledge? Sociological Review monograph 32* (pp. 264-280). London: Routledge & Kegan Paul.

Latour, B. (1996). *Aramis or the love of technology*. Cambridge, MA: Harvard University Press.

Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy and heterogeneity. *Systems Practice, 5*(4), 379-393.

Lepa, J., & Tatnall, A. (2002). *Older people adopting the GreyPath Village Lyceum: An analysis informed by innovation diffusion*. Queensland: AusWeb.

Machiavelli, N. (1515). *The prince* (1995 ed.). London: Penguin Classics.

Maguire, C., Kazlauskas, E. J., & Weir, A. D. (1994). *Information services for innovative organizations*. San Diego, CA: Academic Press.

McMaster, T., Vidgen, R. T., & Wastell, D. G. (1997). Towards an understanding of technology in transition. Two conflicting theories. In K. Braa & E. Monteiro (Eds.), *Information Systems Research in Scandinavia, IRIS20 Conference*. Hanko, Norway: University of Oslo.

Rogers, E. M. (1995). *Diffusion of innovations*. New York: The Free Press.

Singleton, V., & Michael, M. (1993). Actor-networks and ambivalence: General practitioners in the UK cervical screening programme. *Social Studies of Science, 23*(2), 227-264.

Tatnall, A. (2002). Modelling technological change in small business: Two approaches to theorising innovation. In S. Burgess (Ed.), *Managing information technology in small business: Challenges and solutions* (pp. 83-97). Hershey, PA: Idea Group Publishing.

Tatnall, A. (2005a). Technological change in small organisations: An innovation translation perspective. *International Journal of Knowledge, Culture and Change Management, 4*(1), 755-761.

Tatnall, A. (2005b). To adopt or not to adopt computer-based school management systems? An ITEM research agenda. In A. Tatnall, A. J. Visscher, & J. Osorio (Eds.), *Information technology and educational management in the knowledge society* (pp. 199-207). New York: Springer.

Tatnall, A., & Burgess, S. (2002). Using actor-network theory to research the implementation of a B-B portal for regional SMEs in Melbourne, Australia. In C. Loebbecke, R. T. Wigand, J. Cricar, A. Pucihar, & G. Lenart (Eds.), *15th Bled Electronic Commerce Conference—E-Reality: Constructing the E-Economy* (pp. 179-191). Bled, Slovenia: University of Maribor.

Tatnall, A., & Burgess, S. (2004). Using actor-network theory to identify factors affecting the adoption of e-commerce in SMEs. In M. Singh & D. Waddell (Eds.), *E-business: Innovation and change management* (pp. 152-169). Hershey, PA: IRM Press.

Tatnall, A., & Pliaskin, A. (2005). Technological innovation and the non-adoption of a B-B portal. In H. Erhan (Ed.), *Second International Conference on Innovations in Information Technology*. Dubai, UAE: UAE University.

KEY TERMS

Actor-Network Theory (ANT): An approach to research in which networks' associations and interactions between actors (both human and non-human) are the basis for investigation.

Innovation: The application, in any organization, of ideas new to it, whether they are embodied in products, processes, or services.

Invention: The discovery or creation of new ideas.

Innovation Diffusion: A theory of innovation in which the main elements are characteristics of the innovation itself, the nature of the communication channels, the passage of time, and the social system through which the innovation diffuses.

Innovation Translation: A theory of innovation in which, instead of using an innovation in the form it is pro-

posed, potential adopters *translate* it into a form that suits their needs.

Small Business: For the purpose of this article, small businesses are considered to be those businesses that have from 1-20 employees.

Socio-Technical Research: Involving both social and technical interactions, occurring in such a way that it is not easily possible to disentangle them.

Sociology of Translations: Another term used to refer to actor-network theory.

Technological Innovation: The introduction or alteration of some form of technology (often information technology) into an organization.

Risk Management in the Digital Economy

Bob Ritchie
University of Central Lancashire, UK

Clare Brindley
Nottingham Trent University, UK

INTRODUCTION

Digital Economy and Risk: A Two-Edged Sword

Business processes have been transformed by radical changes predicated on the digital economy. Every business sector has witnessed changes in the competitive structure of the marketplace, consumer preferences, buying habits, marketing and promotional strategies, production operations, internal administration systems, supply chain arrangements, and the opening up of the global economy. These changes provide an array of risks for managers. A study of 500 financial executives in Europe and America (FM Global, 2007) concluded that they expected an increase in overall business risks in the foreseeable future. Similarly, Harland, Brenchley, and Walker (2003) concluded that the risk exposure of organizations will heighten due to increased complexity, turbulence, and the dynamic and changing supply chain context. However, the digital economy is a two-edged sword in the sense that the information and communication technologies (ICTs) generating the additional uncertainties and risks also provide the means to enable decision makers to manage them more effectively. The key to survival in the digital economy depends on the abilities of managers to utilize ICTs effectively to manage uncertainties and risks.

ICTs have largely been seen as helping to enhance database access, analytical powers, and the communications capacity of managers. The justification for these efforts has been based on the premise that more and better quality information will result in reduced uncertainty and improved risk perceptions in decision situations. In short, the outcome would be reflected in "better quality" decisions in terms of risk assessment and resolution.

Key topic areas presented in this article include:

- primary elements of the digital economy
- overview of risk and risk management,
- risk and uncertainty,
- individual/organizational response to resolving risk,
- role of information search and corporate intelligence,
- contribution of the digital economy to risk resolution,
- individual characteristics and risk perceptiveness,
- management of risks,
- risk perception,
- information processing and risk resolution, and
- risk management within the digital economy.

BACKGROUND

The Digital Economy

The term "digital economy" reflected in the title of this article may be viewed from a variety of perspectives:

1. *Technology* developments, especially those relating to the digital communication of data and information, are usually considered the primary driver in the creation of the digital economy. Increases in speed, improvements in capacity, accuracy, reliability, general quality, and ease of use are all features that have enabled the widespread adoption and development of digital technologies. The developments in audiovisual communication technologies and wireless technologies have opened up further opportunities for the transmission and exchange of business intelligence. Bluetooth technology /WiFi capabilities are now features of most hotels, gyms, educational centers, and other social spaces.
2. *Socio-economic* changes have been equally influential in the rapid adoption of the new technologies. Individuals of all ages, and educational and social backgrounds are prepared to regularly use mobile communications, access the Internet, and engage in interactive video communications, often with friends and family in different parts of the globe. The impact that these changes in individual and group social behaviors have had on the rate of adoption of new technologies should be fully appreciated. This is evidenced by the growth in social networking sites, such as MySpace and Facebook, and the ubiquitiness of MP3 technologies. The reasons underlying such changes are multifaceted, complex, and

beyond the scope of our present discussion, although they have been broadly influenced by individual social and economic needs.

3. *Micro-economic* factors at the level of the individual organization have been responsible for "pulling" and "pushing" organizations and their management towards increased attention and adoption of the digital economy. Significant "pull" factors include demands from end users of the product or service (e.g., requests for more detailed information on the product/service prior to and subsequent to the purchase, in terms of performance, maintenance, modifications, upgrades). The "push" factors are typically associated with the business organization seeking to maintain its competitive position by offering services equivalent to those of its main competitors, especially if these may be viewed as providing a distinctive competitive advantage (e.g., providing detailed product information via the Web and enabling customers to order directly). Some of the issues involved will be discussed further in later sections of the article.
4. *Macro-economic* factors are particularly significant in enabling the development of the digital economy, although they are often less evident when exploring individual product/market developments. Changes in legislation affecting consumer rights, guarantees of financial transactions, security of information held on computer systems, and commercial contracts negotiated via the Internet are all examples of the changes in the macro-economic environment needed to facilitate and support the development of the digital economy. Without such changes, individual organizations and customers might consider the risks of such commercial transactions to be too high. In essence, the responsiveness of governments and other similar institutions have lagged behind many of the demands placed on them by the rate of change in the digital economy. However, the introduction of the 2007 Gambling Act by the UK government illustrates how legislation is beginning to take into account digital developments, such as the growth in online gambling.

It may be argued that defining the term "digital economy" remains problematic due to the number of perspectives from which this term may be viewed and due to the number and interactive nature of the variables involved.

Overview of Risk and Risk Management

The key dimensions of risk and its management are:

- risk and uncertainty,
- risk resolution,
- role of information search and corporate intelligence,
- contribution of the digital economy,
- individual characteristics,
- management of risk, and
- risk perception.

Risk and Uncertainty

This seemingly simple term "risk" has proved somewhat problematic in arriving at an agreed definition. Most academic fields and researchers (e.g., Dowling & Staelin, 1994; Knight, 1921) provide variations on the theme, though most would agree that risk relates to two dimensions: the *likelihood* of a particular event occurring (i.e., probability), and the *consequences* should this event occur.

In the case of the consequences, it has been common to assume that these are generally undesirable, for example, financial loss or even loss of life. Sitkin and Pablo (1992, p. 9) define risk as "the extent to which there is uncertainty about whether potentially significant and/or disappointing outcomes of decisions will be realised." Similarly, MacCrimmon and Wehrung (1986) identified three components of risk: the magnitude of loss, the chance of loss, and the potential exposure to loss. However, it is important to recognize that there would be no point in taking risks unless there were some benefits to compensate for the possible negative outcomes (Blume, 1971). An associated feature is that of differing risk perceptions. Different individuals, groups of individuals, and organizations may view or perceive the risks (i.e., the likelihood of occurrence, nature and scale of negative consequences, and the potential rewards) differently (e.g., Forlani & Mullins, 2000; March & Shapira, 1987).

Uncertainty is viewed by many authors as a special case of the risk construct (Paulsson, 2004). The term "uncertainty" typically reflects the ambiguity surrounding the decision situation in terms of the precise nature of the situation, its causes, possible solutions, and the reaction of others to possible actions taken. Rowe (1977) has defined uncertainty as the absence of information concerning the decision situation and the need to exercise judgment in determining or evaluating the situation.

Risk Resolution

A natural reaction by decision makers facing uncertainty and risk is to seek to resolve the uncertainty and risk inherent in the decision situation. There are several actions that may be taken:

- seek to understand the parameters of the "problem" or situation,
- assess the predictability of the parameters,

- consider the possibility of eliminating or ameliorating the risks, and
- assess the attitudes of the decision makers towards the risks.

These elements represent the process that individuals and organizations develop to manage the risk, to position the organization, and to develop appropriate strategies to manage the impact of such events (Bettis, 1982). Brindley (2004) suggests that global competition, the continuous search for competitive advantage, and the multi-faceted relationships that have evolved are the primary motives behind organizations turning towards risk management approaches. Giannakis, Croom, and Slack (2004) also evidenced the emergence of risk management as an important contributor to most fields of management decision and control.

Role of Information Search and Corporate Intelligence

The decision maker confronted with a risky decision situation naturally follows a process of gathering more information, processing this in different ways, and evaluating this in relation to the risks faced. This information is then used—maybe consciously or unconsciously—by the decision maker to assess to what extent the risk has been removed, reduced, or resolved. If the decision maker is not satisfied that he or she has achieved sufficient resolution of the risks involved, then further information searches, analysis, and processing will occur. However, this intelligence-gathering process need not necessarily result in improved understanding and risk resolution. Uncovering more information may reveal more influencing variables, causing perceptions of greater uncertainty, both as a consequence of uncovering new influencing variables and an increasing sense of complexity. Ritchie and Marshall (1993) argued that the quality and sufficiency of the information available will influence the perception of risk by those involved in the decision-making process. Ritchie and Brindley (1999) argued that risk perception is both the consequence of information search and analysis as well as its determinant.

Contribution of the Digital Economy

Information can be considered as a risk-reduction or risk-insulating tool, on the basis that more and better information would result in improved decision making and more effective risk management. The rapid pace of ICT developments (Lumpkin & Dess, 2004; Kalakota & Robinson, 2000; Rycroft & Kash, 1999) and the emerging digital economy (Brindley & Ritchie, 2001) demonstrate that both individuals and organizations have easier access to more information that is more immediate and arguably more relevant. The decision makers can therefore access internal data, external market reports, competitor information, and so forth through their own desktops. Swash (1998) recognized that it is essential to overcome the problem of information overload by ensuring the Internet is used frequently, thus improving the users' surfing proficiency and their knowledge of "useful" sites. Improvements in the voluntary exchange of information may produce better knowledge of the situations surrounding the dynamics of a business or commercial relationship. This provides greater potential for detecting, averting, and managing risks. Indeed, organizations believe that the best approach is to accept that they will be exposed to risks, and the best strategy is to become more aware and proactive of the risks and better prepared to respond more quickly should such risks materialize (Kovoor-Misra, Zammato, & Mitroff, 2000).

Individual Characteristics and Risk

It is generally accepted that there are differences in individual risk perception, information-processing styles, and decision-making attributes (e.g., Chung, 1998). There is perhaps less agreement on the nature and consequences of such individual differences. For example, differences in gender have been posited as the reason for women being more meticulous in information search, more responsive to decision cues, and more risk averse than men (e.g., Chung, 1998). Others have failed to establish significant differences (e.g., Masters & Meier, 1988), suggesting that other contextual variables (e.g., social, educational, and experiential background) may be more relevant to any differences in behavior. For example, the decision to start a business may generate differences in risk-taking behavior that is not gender, culture or age related (Busenitz, 1999; Shapira, 1995). A number of authors (see Forlani & Mullins, 2000; Ghosh & Ray, 1997; Kahneman & Lovallo, 1993) have suggested that the provision of structured decision approaches and information search frameworks can provide the appropriate mechanism to overcome any biases and improve the quality of decisions. Ritchie and Brindley (2007) have posited a framework for managing risks within the supply chain.

Management of Risk

The strands of risk together with business intelligence in the digital economy may be captured in the three dimensions of individual/group/organizational behavior when confronted with risk in given decision situations:

- willingness to seek further resolution of the decision situation faced, both in terms of the context and the decision-specific variables;
- desire to identify and measure the risks in some way, either objectively or subjectively; and

- information search and processing to support the decision.

These three dimensions are closely inter-related and are employed to modify the risk perceptions of the decision makers. In addition to these decision-making activities is the range of activities associated with risk management. Some risk management activities may be undertaken prior to the decision itself (e.g., insuring against certain risks) or after the decision (e.g., effective management of relationships with customers to reduce the likely incidence of disputes). While activities of this type may be employed fairly readily with local markets, many managers may find it more difficult to avoid the risks resulting from increased global competition in their home or local markets, consequential of the digital economy.

Risk Perception

Throughout the decision process, an individual is seeking information to remove the uncertainties and to resolve the risks involved in the decision. The factors that contribute to the perceptions of uncertainty and risk experienced by the individual may be represented as (Brindley & Ritchie, 2001):

- direct implications in terms of costs and benefits and the scale of these;
- other decision outcomes, known with less certainty (e.g., reactions of peers and colleagues);
- time available in which to take the decision, typically constrained;
- funding available to undertake the research, analysis, and evaluation processes involved;

Figure 1. Information processing and risk resolution (adapted from Ritchie & Marshall, 1993)

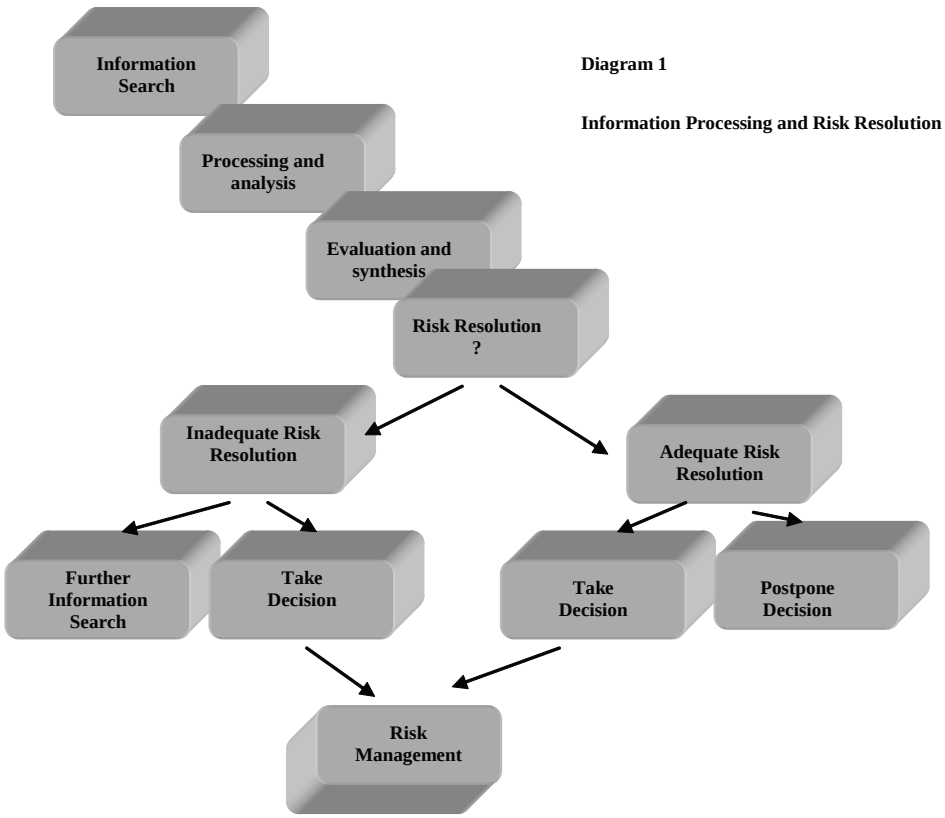


Diagram 1

Information Processing and Risk Resolution

- achievement of personal goals that decision maker(s) seek to satisfy; and
- data quality, which may significantly influence the nature of the risks perceived.

The integration and interaction of many of these components increases the complexity of understanding both risk perception and decision behavior. For example, although funding may be available to undertake intensive research, the pressures of time to respond to a given situation may preclude such activities even though they may be considered desirable.

Information Processing and Risk Resolution

The information-processing behavior of the individual may be presented in terms of the process described in Figure 1. The decision maker seeking to resolve the risks perceived will search for appropriate information, process, analyze, evaluate, and synthesize this with other information relating to the situation. The contribution of the digital economy is evident not only in providing the initial intelligence, but in assisting the decision maker in the processing and manipulation of the data prior to the decision-making stage. It is suggested that the decision maker either consciously or subconsciously assess whether the risks have been sufficiently resolved or not. In many respects, this may be posing the question concerning one's own degree of confidence in proceeding with the decision, though it may also be influenced by the factors identified.

An assessment that the risks had not been adequately resolved would normally result in further information search, analysis, synthesis, and so forth, repeating this cycle until the individual believed the risks had been sufficiently resolved to proceed. It is unlikely that the decision maker would achieve the situation where he or she was fully confident concerning risk resolution. The probable outcome, even after seeking further risk resolution, is that of having to make the decision even though the risks are not fully resolved, as a consequence of pressure of time, lack of resources to pursue further information, or the feeling that further search is unlikely to prove cost effective. This is probably a very common outcome reflected in the statements of many decision makers that "there are many other angles that we ought to explore but we no longer have the time to do so."

In situations where the decision maker considers that the risks have been adequately resolved, one might conclude that the normal procedure would be to proceed with the decision. Another possible outcome could be suggested where the decision maker decides to delay the decision. This may be simply a reluctance to commit oneself even if one is fairly clear and confident of the decision to be made. Alternatively, it may reflect a more deliberate tactic of delay to await changes in the decision situation, to assess the likely responses of other participants in the competitive situation, or perhaps in the hope that someone else may take the responsibility. The making of the decision is no longer viewed as the concluding stage in the process. Attention is increasingly being directed to the post-decision activities that seek to ensure that the opportunity is taken to avoid potential risks anticipated in the decision process, to avoid risks seen as less likely to impact on the situation, and to minimize the likelihood and consequences of those risks seen as more likely to occur. A generic term for these activities is *risk management.* Examples include:

- insuring against the financial consequences of particular outcomes;
- developing formal contractual arrangements with trading partners to limit scope for non-predicted behavior;
- training and staff development;
- communication of intentions within the appropriate internal and external networks to ensure involvement and commitment (e.g., relationship marketing);
- detailed planning, monitoring, and control at all stages;
- building close relationships with key partners to generate a sense of common purpose and trust (e.g., partnering); and
- developing more effective risk management strategies through more effective communications at all levels, both within the organization and with external partners.

FUTURE TRENDS

Risk Management Within the Digital Economy

The digital economy is impacting on organizations irrespective of their size, geographic location, or sector, a trend that is set to continue (Ritchie & Brindley, 2005) While such developments will engender greater uncertainty and risks, they will also facilitate solutions and opportunities for organizations to resolve and manage the new risks. Specific challenges for the organizations in the digital economy are:

1. Fundamental changes occur in the marketplace as a consequence of providing more direct communications with the consumer, greater heterogeneity in the market in terms of consumer needs and behavior, and a movement in the balance of power towards the consumer as opposed to the manufacturer or service provider.
2. Individuals and organizations are unlikely to behave in a rational and structured manner to resolve risks.

The previous emphasis on predictable patterns of information search, corporate business intelligence, and evaluation of information is likely to be replaced by less predictable demands as the nature of the problems encountered are less predictable. This will have significant implications for the development and design of business intelligence systems.

3. Predictions that the digital economy—by providing improved access to information and processing capabilities—will lead to improved decisions is unlikely to be sustainable.
4. The modus operandi of competitive and business relationships within the digital economy will evolve as the nature and extent of risks faced change, both in terms of novel risk situations encountered and the rate at which these occur. This is evident in the changing nature of supply chains (Brindley, 2004).
5. The development and training of the individual in terms of the appropriate knowledge, skills, and behavior patterns provides an important requirement to utilize the information available effectively.
6. The nature of what constitutes "effective" risk management is changing in the digital economy (Ritchie & Brindley, 2007).

CONCLUSION

The digital economy has produced a fundamental change in the nature of the risks faced and increased the likelihood that the marketplace will remain turbulent, unstable, and risk prone. Managers now need to have the capability—through improved knowledge, skills, and understanding—to identify, analyze, and manage these competitive developments and the associated risks. Associated with the improved capability to manage the risks is the ability to implement a wider range of risk management strategies to ameliorate the consequences of the incidence of risks and their consequences. The digital economy and the associated ICTs improve the opportunities to ensure effective risk management.

REFERENCES

Bettis, R.A. (1982). Risk considerations in modeling corporate strategy. *Academy of Management Proceedings* (pp. 22-25).

Blume, M.E. (1971). On the assessment of risk. *Journal of Finance, 26*(1), 1-10.

Brindley, C.S. (Ed.), (2004). *Supply chain risk.* Hampshire, England: Ashgate.

Brindley, C.S., & Ritchie, R.L. (2001). The information-risk conundrum. *Marketing Intelligence and Planning, 19*(1), 29-37.

Busenitz, L.W. (1999). Entrepreneurial risk and strategic decision making: It's a matter of perspective. *Journal of Applied Behavioral Science, 35*(3), 325-340.

Chung, J.T. (1998). Risk reduction in public accounting firms: Are women more effective? *International Review of Women and Leadership, 4*(1), 39-45.

Dowling, R.G., & Staelin, R. (1994).A model of perceived risk and intended risk-handling activity. *Journal of Consumer Research, 21*(1), 119-125.

FM Global. (2007). *Managing business risk—through 2009 and beyond*. Windsor, Berks, UK: FM Insurance Company Ltd.

Forlani, D., & Mullins, J.W. (2000). Perceived risks and choices in entrepreneurs' new venture decisions. *Journal of Business Venturing, 15* (4), 305-322.

Ghosh, D., & Ray, M.R. (1997). Risk, ambiguity, and decision choice: Some additional evidence. *Decision Sciences, 28*(1), 81-104.

Giannakis, M., Croom, S., & Slack, N. (2004). Supply chain paradigms. In S. New & R. Westbrook (Eds.), *Understanding supply chains* (pp. 1-22). Oxford: Oxford University Press.

Harland, C., Brenchley, R., & Walker, H. (2003). Risk in supply networks. *Journal of Purchasing and Supply Management, 9,* 51-62.

Kahneman, D., & Lovallo, D. (1993). Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Science, 39*(1), 17-31.

Kalakota, R., & Robinson, M. (2000). *E-business.* Reading, MA: Addison-Wesley Longman.

Knight, F.H. (1921). *Risk, uncertainty and profit.* Boston/New York: Houghton Mifflin.

Kovoor-Misra, S., Zammato, R., & Mitroff, I.I. (2000) Crisis preparation in organisations: Prescription versus reality. *Technological Forecasting and Social Change, 63,* 43-62.

Lumpkin, G.T., & Dess G.G. (2004). E-business strategies and Internet business models: How the Internet adds value. *Organizational Dynamics, 33*(2), 161-173.

Paulsson, U. (2004). Supply chain risk management. In C. Brindley (Ed.), *Supply chain risk* (pp. 79-96). Hampshire, England: Ashgate.

MacCrimmon, K.R., & Wehrung, D.A. (1986). *Taking risks: The management of uncertainty.* New York: The Free Press.

March, J.G., & Shapira, Z. (1987).Managerial perspectives on risk and risk taking. *Management Science, 33* (11), 1404-1418.

Masters, R., & Meier, R. (1988). Sex differences and risk-taking propensity of entrepreneurs. *Journal of Small Business Management, 26*(1), 31-35.

Ritchie, B., & Brindley, C. (2005). ICT adoption by SMEs: Implications for relationships and management. *New Technology Work and Employment, 20*(3), 205-217.

Ritchie, B., & Brindley, C.S. (2007). An emergent framework for supply chain risk management and performance measurement, *Journal of the Operational Research Society.*

Ritchie, R.L., & Brindley, C.S. (1999, April 22-23). Relationship marketing as an effective approach to organisational risk management strategies. *Proceedings of the 4th International Conference on the Dynamics of Strategy* (pp. 313-323), Surrey, UK.

Ritchie, R.L., & Marshall, D.V. (1993). *Business risk management.* London: Chapman and Hall.

Rowe, W.D. (1977). *Anatomy of risk.* New York: John Wiley & Sons.

Rycroft, R.W., & Kash, D.E. (1999). *The complexity challenge.* London: Pinter.

Shapira, Z. (1995). *Risk taking: A managerial perspective.* New York: Russell Sage.

Sitkin, S.B., & Pablo, A.L. (1992). Reconceptualizing the determinants of risk behaviour. *Academy of Management Review, 17*(1), 9-38.

Swash, G. (1998). UK business information on the Internet. *New Library World, 99*(1144), 238-242.

KEY TERMS

Decision Support: The tools, techniques, and information resources that can provide support to the decision maker in improving the efficiency and effectiveness of his or her decisions. Many of these decision support tools may employ ICTs and be part of the management information system itself.

Digital Economy: Accepts as its foundation the ICT developments and represents the impact that these have had on the conduct of business and commercial activities. Changes in markets and supply chains as well as increasing global competition all represent what is encapsulated within the term 'digital economy'.

Information and Communication Technologies (ICTs): A generic term used to encapsulate the diverse range of technological developments (e.g., computer storage and retrieval, computing capacity, wired communications, wireless communications, portable technologies) that have enhanced the internal and external activities of organizations. Especially important is the manner in which these strands of technological development have been integrated to provide greater synergy.

Management Information: A term that covers a wide variety of sources and types of information that may provide valuable to the decision making, management, and control of an organization. This term would include quantitative and qualitative information types, internal and externally sourced information, as well as classifying the information in terms of its quality (e.g., accuracy, detail, relevance, timeliness).

Risk: In a limited manner, the decision situation in which the full range of possible outcomes are known with certainty and the probability of their occurrence can be assessed accurately, usually by some objective means (e.g., rolling the dice is a classic risk decision situation). More usually, the probabilities must be assessed subjectively, often based on previous experiences or intuition, and the outcomes themselves may not be fully identifiable. The term "risk" is used commonly to generally define decision situations that are really a combination of classical risk and uncertainty – that is, the more normal decision situation in organizations.

Risk Management: The range of activities that may be taken to avoid the occurrence of an undesirable event or to modify, minimize, or eliminate the consequences should the event occur (e.g., an insurance policy against particular risks would not prevent the occurrence, but would compensate for the financial and other consequences of the outcome).

Risk Perception: The term used to express how a situation is viewed or seen by the decision maker(s). Individual characteristics, experiences, and beliefs may influence the way in which we might view a given situation as being either more or less risky. Usually this is measured on a subjective and relative scale (i.e., Situation A is perceived as riskier than B) rather than on an objectively measurable scale.

Uncertainty: The situation where less-than-perfect knowledge exists about a particular problem or decision requirement. There exists a wide variation in terms of degrees of uncertainty from extreme uncertainty (i.e., very limited knowledge of outcomes or likelihood of their occurrence)

to near certainty (i.e., almost complete knowledge of the outcomes and the likelihood of occurrence). Generally, an uncertain decision situation refers to one containing ambiguity about part or all of the decision parameters.

A Road Map for the Validation, Verification and Testing of Discrete Event Simulation

Evon M. O. Abu-Taieh
The Arab Academy for Banking and Financial Sciences, Jordan

Asim Abdel Rahman El Sheikh
The Arab Academy for Banking and Financial Sciences, Jordan

INTRODUCTION

The aim of this chapter is to give an elaborate reasoning for the motivation for Validation, Verification, and Testing (VV&T) in Simulation. Thereby, defining Simulation in its broadest aspect as embodying a certain model to represent the behavior of a system, whether that may be an economic or an engineering one, with which conducting experiments is attainable. Such a technique enables the management, when studying models currently used, to take appropriate measures and make fitting decisions that would further complement today's growth sustainability efforts, apart from cost decrease, as well as service delivery assurance. As such, the Computer Simulation technique contributed in cost decline; depicting the "cause and effect," pinpointing task-oriented needs or service delivery assurance, exploring possible alternatives, identifying problems, as well as proposing streamlined, measurable, deliverable, solutions, providing the platform for change strategy introduction, introducing potential prudent investment opportunities, and finally, providing a safety net when conducting training courses. Yet, the simulation development process is hindered due to many reasons.

Like a rose, Computer Simulation technique, does not exist without thorns, of which the length, as well as the communication during the development life cycle. Simulation reflects real-life problems; hence, it addresses numerous scenarios with handful of variables. Not only is it costly, as well as liable for human judgment, but also, the results are complicated and can be misinterpreted.

BACKGROUND

There are four characteristics, which distinguish simulation from any other software intensive work, that also makes

Table 1. Published research of V & V in WSC adapted from (Abu-Taieh & ElSheikh, 2006)

Year	V & V Papers	Total Published Papers	Percentage
1997	15	280	**5%**
1998	22	236	**9%**
1999	26	244	**11%**
2000	19	280	**7%**
2001	15	224	**7%**
2002	1	119	**1%**
2003	12	263	**5%**
2004	6	280	**2%**
2005	11	412	**3%**
2006	10	412	**2%**
Total	**137**	**2750**	**5%**
Average	**13.7**	**275**	**5%**

distinction VV&T for simulation from VV& T for other software. The four characteristics were discussed by Page and Nance (1997): *time*, *correctness*, *computational intensive*, and *the uses of simulation*. In simulation there is an indexing variable, called *TIME,* that "establishes an ordering of behavioral events" (p. 91). The objective of *correctness* is very special to simulation software for this simple reason: how useful is simulation program "if questions remain concerning its validity" (p.91). Simulation is *computational intensiveness;* therefore, the execution efficiency is essential due to the repetitive sample generation for statistical analysis and testing alternatives. *Uses of simulation*: the uses of simulation are not typical; in fact there is "No typical use for simulation can be described" (p.91).

Accordingly, validation and verification methods and techniques that relate to simulation have been thoroughly discussed by 137 research papers in the Winter Simulation Conference (WSC) over the years 1997 through 2006, as seen in Table 1 and Figure 1, highlighting the fact that such numbers clearly indicate the importance, inimitable, and unique case of validation, verification, and testing of simulation software.

In this context, it is worth noting that validation, verification, testing, as well as experimentations, execution, and design are so important that Shannon (1998) suggested giving 40% of the project time for these steps.

Why Do Simulation Projects Fail?

The arising issue of simulation projects falling short of being labeled as successful can be attributed to many reasons. An answer by Robinson (1999) has been put forth, as he listed three main reasons: the first being "poor salesmanship when introducing the idea to an organization," (p. 1702) which includes too much hope in too little time, while identifying the second reason as "lack of knowledge and skills particularly in statistics, experimental design, the system being modeled and the ability to think logically" (p. 1702), and pinpointing the third reason as "lack of time to perform a study properly" (p. 1702). Nevertheless, simulation inaccuracy has become a recurrent condition that instigated a thorough query into its sources.

Sources of Simulation Inaccuracy

There are three sources of inaccuracy the simulation project might be developing during the three major steps that are in the simulation life cycle, namely; modeling, data extraction, and experimentation (see Figure 2).

In this regard, the modeling process includes a subordinate set of steps, namely, the modeler understanding the problem, then developing a mental/conceptual model, and finally the coding. Note that, from these steps, some problems might mitigate themselves, such as (i) the model could misunderstand the problem, (ii) the mental/conceptual model could be erroneous, and (iii) the conversion from mental/conceptual model to coding could be off beam.

Furthermore, during the modeling process, the data collection/analysis is a key process element, particularly since the data collected is really the input of the simulation program. If the data is collected inaccurately, then the principle of *Garbage In Garbage Out* is clearly implemented; likewise, the data analysis, while using the wrong input model/distribution (see Leemis, 2003) is also a problem.

Figure 1. V & V research papers in WSC adapted from (Abu-Taieh & ElSheikh, 2007)

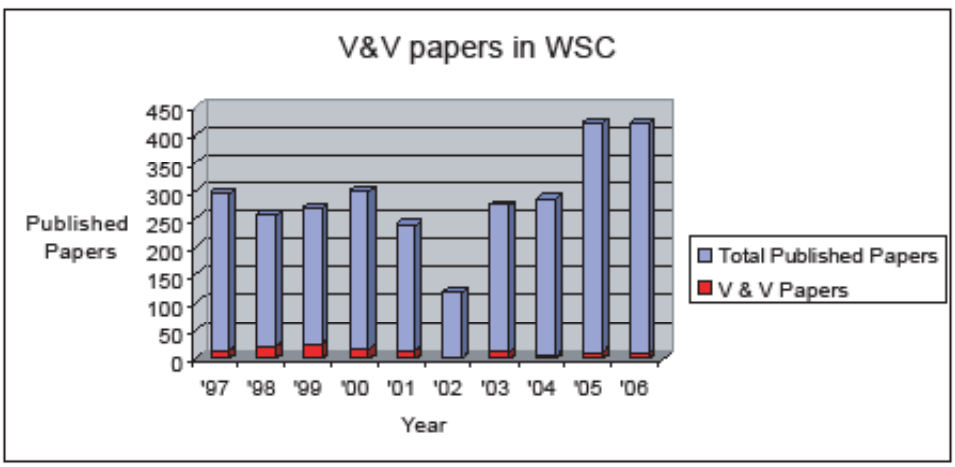
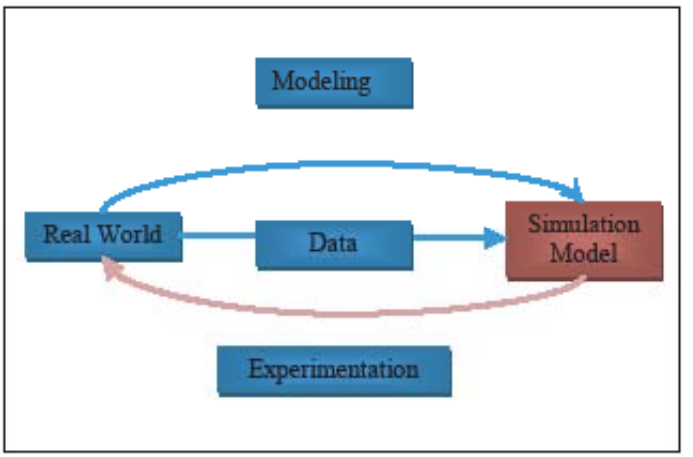


Figure 2. The simulation modeling process (simple outline) (Robinson, 1999, p. 1702)



Last, but not least, the third source of inaccuracies is experimentation, which is using the collected data used in the simulation system and comparing the end result to the real world, given that experimentation inaccuracies can result from ignoring the initial transient period, insufficient run-length or replications, insufficient searching of the solution space, not testing the sensitivity of the results.

Within this context, ignoring the initial transient period, labeled as the first inaccuracy source during experimentation process, has been identified by Robinson (1999), when stating that "Many simulation models pass through an initial transient period before reaching steady-state" (p. 1705). The modeler, suggests Robinson, can either take into account such period or set the simulation system so that the system has no transient period. Moreover, insufficient run-length or replications, particularly as running the simulation system requires long enough is essential, in order to reach the results that reflect real life, therefore, Robinson suggests two remedies; (i) run the simulation system long enough, or (ii) do many reruns (replications). In addition, insufficient searching of the solution space is the third and last source, which in turn would incite the modeler to "only gain a partial understanding of the model's behavior" (Robinson, 1999, p. 1702), such inaccuracy, obviously leads to erroneous conclusion.

Acknowledging that, clearly, errors in the simulation world do not originate only from one source, therefore, validation, verification, and testing are not only considered a necessity, but also considered to be imperative and crucial, as Osman Balci, the well-known simulation scientist declared 15 principles, see (Balci, 1995, p. 149-151), demonstrating the fundamentality and inevitability of conducting VV&T to the simulation world.

VV&T Definition and Distinction (Main Focus)

In order to be able to discuss the VV&T, first, they must be defined, based on their original definition, fully comprehending the true denotation of VV&T, as related to simulation systems. Sommerville, (2001) defines validation by raising the question "are we building the right product?"(p. 420), while defining verification by raising another question, "are we building the product right?" (p. 420), as a pure software engineering point of view, noting that the simulation perspective on the definitions of V & V as similar yet not the same. On another note, Pidd (1988) defines Validation as "a process whereby we asses the degree to which the lumped model input: output relation map onto those real systems" (p. 157).

Likewise, Pidd (1998) distinguishes validation from verification by referring to verification as "a process by which we try to assure ourselves that the lumped model is properly released in the computer program. The lumped model is defined as an explicit and simplified version of the base model and is the one that will be used in management science" (p. 157). While Smith (1998) paraphrases validation as answering the question "Are we building the right product?" (p. 806) and verification as answering the question, "Are we building the product right?"(p. 806). (Balci, 1994, 1995 , 2003,) and (Kleijnen, Bettonvil, & Gmenendahl

1996) define *model validation* as follows:" *Model validation* is substantiating that the model, within its domain of applicability, behaves with satisfactory accuracy consistent with the study objectives" (Balci, 1995, p. 147).

Defining validation as true simulation scientist when stressing the words domain and satisfactory. Also, Balci (1995) defines *model verification* as in the following": *"Model verification* is substantiating that the model is transformed from one form into another, as intended, with sufficient accuracy" (p. 147). Again, Balci's (1995) definition is stressing here sufficient accuracy. Then Balci defines *model testing* as follows: "*Model testing* is demonstrating that inaccuracies exist or revealing the existence of errors in the model" (p. 147).

As such, the mere distinction between verification and validation is minute, yet substantive, acknowledging that validation ensures that the product caters to the needs, whereas, verification ensures the correctness and aptness of the product. Within this context, the word *testing* would be the tool to examine these two aspects.

Validation, Verification and Testing (VV&T) Taxonomy

Back in 1994, Balci (1994) categorized 45 VV&T techniques to informal, static, dynamic, symbolic, constraint, and formal (Balci, 1994), later he categorized 115 VV&T techniques to three families: conventional, adaptive, and specific VV&T (Balci, 1997), of which, 77 VV&T techniques for conventional simulation was again categorized to informal, static, dynamic, and formal, the remaining 38 VV&T techniques for adaptive and specific VV&T to be used in object-oriented simulation. Nevertheless, others, like Pressman (2005) and Hoffer (Hoffer, George, & Valacich, 2005) categorized the VV&T techniques based on the life-cycle phase. Within this context, following is a thorough discussion of both Balci's categorization.

On one hand, the conventional VV&T, in Figure 3 of Balci, had been before the object-oriented simulation idea in which the 45 VV&T techniques were categorized: informal, static, dynamic, symbolic. Within this context, the *informal VV&T technique* includes tools and approaches that "relay heavily on human reasoning" (Balci, 1994, p.217) rather than "mathematical formalism" (Balci, 1994, p.217), this category includes audit, desk checking, inspection, reviews, Turing test, walkthroughs (Freedman & Weinberg, 2000; Hoffer et al., 2005; Zammit, 2005), given the fact that the word *informal* should not reflect lack of structure or formal guidelines, as Balci (1994) says.

Along the same lines, *static VV&T techniques* concentrate on the source code of the model and need not the execution of the code (Hoffer et al., 2005), more importantly, Balci has stated that automated tools and the language compilers relay on this type of VV&T (Balci, 1994), since the *static VV&T techniques* category comprises of consistency checking, data flow analysis (Pressman, 2005), graph-based analysis (DMSO, 2005), semantic analysis, structural analysis, and syntax analysis (Sebesta, 2003).

Whereas, *dynamic VV&T techniques* distinguishing characteristic is for the model execution in order to evaluate the model as Balci (1994) states, as this category comprises of the following VV&T techniques: black-box testing (DMSO, 2005), bottom-up testing, debugging, execution monitoring, execution profiling, execution tracing, field testing, graphical comparisons, predictive validation, regression testing (Orso, Shi, & Harrold, 2004, p. 241),.and Mansour, (Mansour & Brardhi, 2001), sensitivity analysis, statistical techniques, stress testing, submodel testing, symbolic debugging, top-down testing, visualization, and white-box testing.

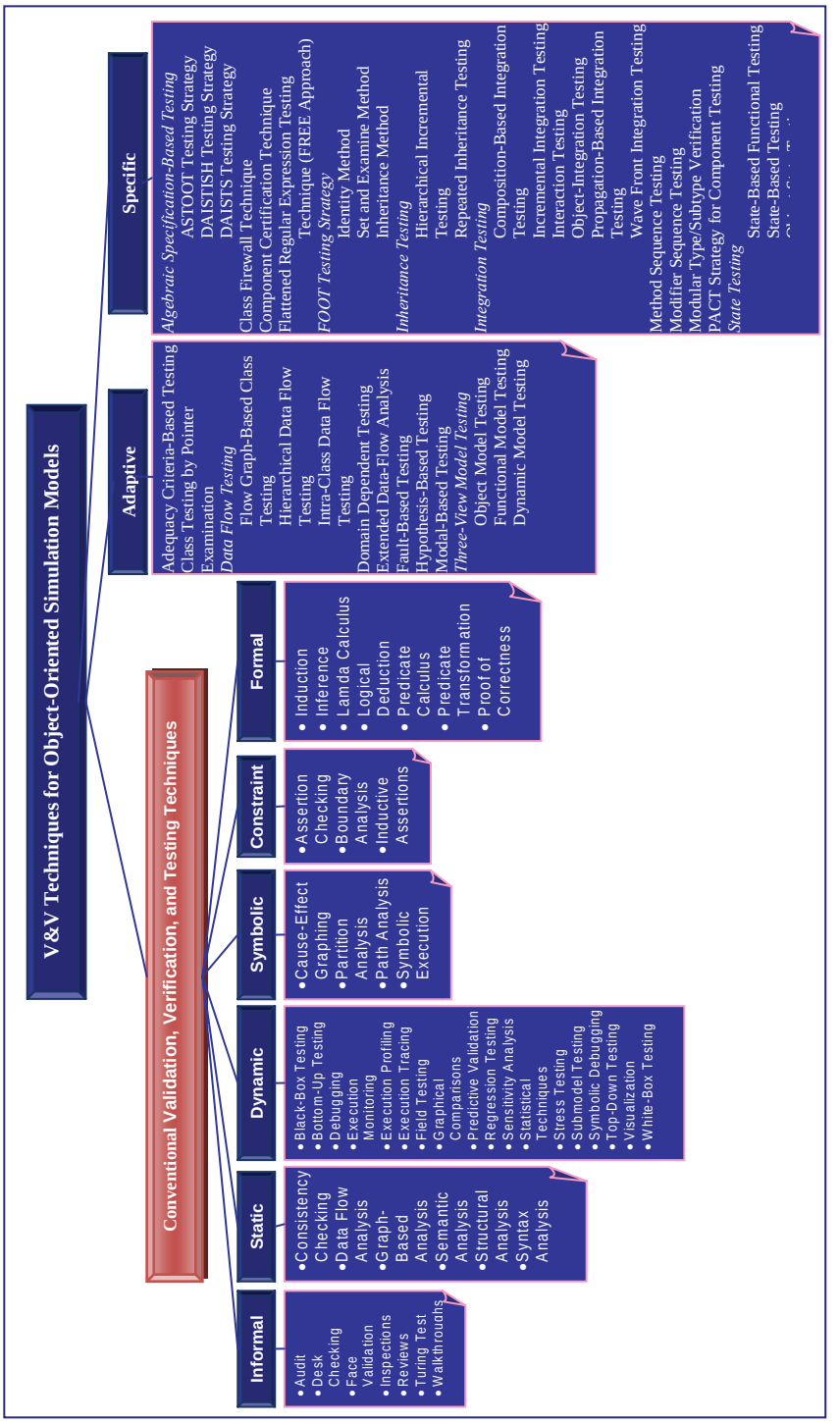
Nonetheless, the *symbolic VV&T techniques* category has been used to asses the model using VV&T techniques like cause–effect graphing, partition analysis, path analysis, and symbolic execution, noting that later in Balci's (1997), this category was incorporated with static category and dynamic category, while cause–effect graphing, and symbolic execution were incorporated in the static category, partition analysis, and path analysis were incorporated in the dynamic category. On another note, Balci has stated that *"Constraint VV&T techniques* are employed to assess model correctness using assertion checking, boundary analysis, and inductive assertions." (Balci, 1994, p. 218), later, however, such category also disappeared and was incorporated with others.

Balci, among others, admits that *formal VV&T techniques* are based on mathematical proof, stating that "Current state-of-the-art formal-proof of correctness techniques are simply not capable of being applied to even a reasonably complex simulation model" (Balci, 1994, p. 218).

Balci (1997) has developed taxonomy of VV&T techniques, Figure 3, that classified the VV&T techniques into three categories: conventional, specific, and adaptive. Moreover, the taxonomy defined 77 V&V techniques and 38 V&V techniques for object-oriented simulation, noting that most of the techniques are derived from the software engineering world, and modified to the needs of object-oriented simulation systems.

Within this context, *conventional techniques* refer mainly to VV&T techniques used in the object-oriented simulation without any adaptation to OOP, which is further classified as informal, static, dynamic, and formal categories, highlighting the fact that such categories were previously discussed. Likewise, *adaptive techniques*, the second category in the taxonomy, which refers to techniques that were adapted to object-oriented theory including adequacy criteria-based testing, class testing by pointer examination, data flow testing, (which includes flow graph-based class testing, hierarchical data flow testing, intra-class data flow testing), domain dependent testing, extended data-flow analysis, fault-based testing, hypothesis-based testing, modal-based

Figure 3. Taxonomy of validation, verification, and testing techniques (Balci, 1995, p. 152) and (Balci, 1997, p.140)



testing, three-view model testing (Pezz & Young, 2004) (which include object model testing, functional model testing, dynamic model testing).

Finally, *specific techniques*, the third category in the taxonomy, which are newly created based on object-oriented formalism and used for object –oriented software. Taking into consideration that those techniques are algebraic specification-based testing, ASTOOT testing strategy (Doong & Frankl, 1994), DAISTISH testing strategy (Hughes & Stotts, 1996), DAISTS testing strategy (Hughes & Stotts, 1996, p.1), class firewall technique, component certification technique, flattened regular expression testing, technique (FREE approach), FOOT testing strategy, identity method, set and examine method, inheritance method, inheritance testing, hierarchical incremental testing, repeated inheritance testing, integration testing, composition-based integration testing, incremental integration testing, interaction testing, object-integration testing, propagation-based integration testing, wave front integration testing, method sequence testing, modifier sequence testing, modular type/subtype verification, PACT strategy for component testing, state testing, state-based functional testing, state-based testing, object state testing, graph-based class testing.

FUTURE TRENDS

There are two major trends in the VV&T world: automation and correctness. The importance of correctness, which emphasizes the 100% correct software, is obvious in the use of Z specification language. Where the major idea is to prove mathematically that software is 100% error free.

The second trend is the automation of the VV&T. Indeed, testing is tedious, mind-numbing, lackluster, repetitive, and long process. Just as compilers started years ago to locate syntax errors, VV&T tools are becoming. In fact, there are many testing software tools that test software.

Whether the VV&T technique is static or dynamic, formal or informal, many tools have been developed to do the grubby work for the developer. Some tools, like *ProofPower,* go as far as supporting proof in higher order logic (HOL) and in the Z notation. As such, those mentioned hereinafter will be part of our life in validation, verification, accreditation, and surly independence.

CONCLUSION

In conclusion, this chapter showed the importance of VV&T that pertains to simulation through the eyes of research and researchers in the most famous front to simulation, Winter Simulation Conference. Then the chapter discussed why simulation fails and how VV&T of simulation is different than VV&T of the rest of software. Furthermore, this chapter has given a comprehensive synopsis of the simulation VV&T techniques, whether that may be procedural or object-oriented simulation, putting forth various taxonomies for both types of VV&T techniques.

REFERENCES

Abu-Taieh, E. & ElSheikh, A. (2006). Verification, validation and testing in software engineering. In Dasso and Funes (Eds.), *Discrete event simulation process validation verification and testing*. Hershey, PA: Idea Group Inc.

Arthur, J., & Nance, R. (2000). Verification and validation without independence: A recipe for failure. In J. A. Joines, R. R. Barton, K. Kang, & P. A. Fishwick (Eds.), *Proceedings of the Winter Simulation Conference* (pp. 859-856), December 10-13, Orlando, FL, United States. San Diego, CA: Society for Computer Simulation International.

Balci, O, Nance, R., Arthur, J., & Ormsby, W. (2002). Expanding our horizons in verification, validation, and accreditation research and practice. In E. Yücesan, C.-H. Chen, J. L. Snowdon, & J. M. Charnes (Eds.), *Proceedings of the 2002 Winter Simulation Conference* (pp. 653-663), December 8-11, San Diego, CA. Piscataway, NJ: IEEE.

Balci, O. (1994). Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of Operations Research, 53*, 215-220.

Balci, O. (1995). Principles and techniques of simulation validation, verification, and testing. In C. Alexopoulos, K. Kang, W. R. Lilegdon, & D. Goldsman (Eds.), *Proceedings of the 1995 Winter Simulation Conference* (pp. 147-154). New York, NY: ACM Press.

Balci, O. (1997). Verification, validation and accreditation of simulation models. In S. Andradóttir, K. J. Healy, D. H. Withers, & B. L. Nelson (Eds.), *Proceedings of the Winter Simulation Conference* (pp. 135-141)**,** DECEMBER 7—10, Atlanta, Georgia, United States.

Balci, O. (2003). Verification, validation, and certification of modeling and simulation applications. *ACM Transactions on Modeling and Computer Simulation*, *11*(4), 352–377.

Banks, 1999

Doong, R. & Frankl, P. (1994). The ASTOOT approach to testing object-oriented programs. *ACM Trans. on Software Engineering and Methodology*, *3*(2), 101-130.

DMSO. (1996). Department of Defense verification, validation and accreditation (VV&A) recommended practices guide. In O.Balci, P. A.Glasow, P. Muessig, E. H. Page, J. Sikora, S. Solick, & S.Youngblood, *Defense modeling and*

simulation office. Alexandria, VA, Nov. Retrieved May 25, 2005, from http;//vva.dmso.mil/Mini_Elabs/VVtech-dynamic.htm

Freedman, D. P., & Weinberg, G. M. (2000). Handbook of walkthroughs, inspections, and technical reviews: Evaluating programs, projects, and products, 3rd edition. New York, NY: Dorset House Publishing Co., Inc.

Hoffer, J. A., George, J. F., & Valacich, J. S. (2005). Modern systems analysis and design, 4 ed. Prentice Hall.

Hughes, M., & Stotts, D. (1996). Daistish: Systematic algebraic testing for OO programs in the presence of side-effects. In *Proceedings of the ACM SIGSOFT Int. Symp. On Software. Testing and Analysis* (pp. 53 -61), 1996. . Retrieved June 12, 2005, from http://rockfish.cs.unc.edu/pubs/issta96.pdf

Kleijnen, J., Bettonvil, B., & Gmenendahl, W. (1996). Validation of trace-driven simulation models: Regression analysis revisited. In J. M. Ckrnes, D. J. Morrice, D. T. Runner, & J. J. Swain (Eds.), *Proceedings of the Winter Simulation Conference* (pp.352-359).

Leemis, L. (2003). Input modeling. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice, *Proceedings of the 2003 Winter Simulation Conference*, 14-24, December 7-10, New Orleans, New Orleans, LA, United States.

Mansour, N., & Brardhi, B R. (2001). Empirical comparison of regression test selection algorithm. *The Journal of System and Software, 57*, 79-90.

Orso, A., Shi, N., & Harrold, M. J. (2004). Scaling regression testing to large software systems**. In** *SIGSOFT'04/FSE-12* (pp.241-251), Oct. 31–Nov. 6, 2004, Newport Beach, CA.

Page, H., & Nance, R. (1997). Parallel discrete event simulation: A modeling methodological perspective. *ACM Transactions on Modeling and Computer Simulation*, *7*(3), 88-93.

Pezz, M., & Young, M. (2004). Testing object oriented software. In *Proceedings of the 26th International Conference on Software Engineering* (ICSE'04).

Pidd, M. (1998). Computer simulation in management science, 4th ed. Chichester, England, John Wiley & Sons.

Pressman, R. (2005). Software engineering: A practitioner's approach, 6th ed. McGraw Hill.

Pretschner, A. (2005). Model based testing. In *Proceedings of ICSE'05* (pp.722-723), May 15–21, 2005, St. Louis, Missouri, USA.. ACM 1581139632, May/0005.

Richardson, D. J., O''Malley, O., & Tittle, C. (1989). Approaches to specification-based testing. In *Proceedings of ACM SIGSOFT Symposium on Software Testing, Analysis and Verification* (pp.86-96), December.

Robinson, S. (1999). Three sources of simulation inaccuracy (and how to overcome them).In P. A. Farrington, H. B. Nembhard, D. T. Sturrock, & G. W. Evans (Eds.), *Proceedings of the 1999 Winter Simulation Conference* (pp. 1701-1708), December 10-13, Orlando, FL, United States.

Sargent, R. (2003). Verification and validation of simulation models. In S. Chick, P. J. Sánchez, D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the Winter Simulation Conference* (pp.39-48), December 7-10, New Orleans, LA, United States.

Sebesta, 2003

Shannon, R. (1998). Introduction to the art and science of simulation. In D. J. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the Winter Simulation Conference* (pp. 7-14), December 13-16, Washington D.C.

Smith, R. (1998). Essential techniques for military modeling and simulation. In D. J. Medeiros, E. F. Watson, J. S. Carson, & M. S. Manivannan (Eds.), *Proceedings of the Winter Simulation Conference* (pp. 805-812), December 13-16, Washington D.C.

Sommerville, I. (2001). *Software engineering*, 6th ed. Addison-Wesley, Pearson Education Ltd.

Withers, D. (2000). Software engineering best practices applied to the modeling process. In J. A. Joines, R. R. Barton, K. Kang, & P. A. Fishwick (Eds.), *Proceedings of the Winter Simulation Conference* (pp.432-439), December 10-13, Orlando, FL.

Zammit, J. (2005). Correct system, Web site for information systems engineering, University Of Malta. Retrieved June, 18, 2005, from http://www.cis.um.edu.mt/~jzam/vv.html

KEY TERMS

Adaptive Techniques: Refers to *VV&T* techniques that were adapted to object-oriented theory.

Constraint VV&T Techniques: Employed to assess model correctness using assertion checking, boundary analysis, and inductive assertions." (Balci, 1994, p. 218).

Conventional Techniques: Refer mainly to VV&T techniques used in the object-oriented simulation without any adaptation to OOP.

Dynamic VV&T Techniques: Distinguishing characteristic is for the model execution in order to evaluate the model as Balci (1994) states.

Formal VV&T Techniques: Based on mathematical proof, stating that "Current state-of-the-art formal proof

of correctness techniques are simply not capable of being applied to even a reasonably complex simulation model" (Balci, 1994, p. 218).

Informal VV&T Technique: Includes tools and approaches that "relay heavily on human reasoning" (Balci, 1994, p.217) rather than "mathematical formalism" (Balci, 1994, p.217).

Simulation: "Is the imitation of the operation of a real-world process or system over time" (Banks, 1999).

Specific Techniques: Which are newly created based on object-oriented formalism and used for object –oriented software.

Static VV&T Techniques: Concentrate on the source code of the model and need not the execution of the code (Hoffer et al., 2005), more importantly, Balci has stated that automated tools and the language compilers relay on this type of VV&T (Balci, 1994).

Symbolic VV&T Techniques: Techniques that uses graphical or symbols to represent the VV&T process and to simplify it.

Validation: Raising the question "are we building the right product?".

Verification: Answering the question, "Are we building the product right?".

Robustness in Neural Networks

Cesare Alippi
Politecnico di Milano, Italy

Manuel Roveri
Politecnico di Milano, Italy

Giovanni Vanini
Politecnico di Milano, Italy

INTRODUCTION

The robustness analysis for neural networks aims at evaluating the influence on accuracy induced by perturbations affecting the computational flow; as such it allows the designer for estimating the resilience of the neural model w.r.t perturbations. In the literature, the robustness analysis of neural networks generally focuses on the effects of perturbations affecting biases and weights. The study of the network's parameters is relevant both from the theoretical and the application point of view, since free parameters characterize the "knowledge space" of the neural model and, hence, its intrinsic functionality.

A robustness analysis must also be taken into account when implementing a neural network (or the intelligent computational system into which a neural network is inserted) in a physical device or in intelligent wireless sensor networks. In these contexts, perturbations affecting the weights of a neural network abstract uncertainties such as finite precision representations, fluctuations of the parameters representing the weights in analog solutions (e.g., associated with the production process of a physical component), ageing effects or more complex, and subtle uncertainties in mixed implementations.

BACKGROUND

The sensitivity/robustness issue has been widely addressed in the neural network community with a particular focus on specific neural topologies. In particular, when the neural network is composed of linear units, the relationship between perturbations and the induced performance loss can be obtained in a closed form (Alippi & Briozzo, 1998). Conversely, when the neural topology is non-linear, we have either to assume the small perturbation hypothesis or particular assumptions about the stochastic nature of the neural computation (e.g., see Alippi (2002a), Alippi et al. (1998), and Pichè, 1995); unfortunately, such hypotheses are not always satisfied in real applications. Another classic approach requires expanding the neural computation with Taylor around the nominal value of the trained weights. A subsequent linearized analysis follows, which allows the researcher to solve the sensitivity issue problem (Pichè, 1995). This last approach has been widely used in the implementation design of neural networks where the small perturbation hypothesis abstracts small errors introduced by finite precision representations of the weights (Dundar & Rose, 1995; Holt & Hwang, 1993). Again, the validity of the analysis depends on the validity of the small perturbation hypothesis.

Differently, other authors avoid the small perturbation assumption by focusing the attention on very specific neural network topologies and/or by introducing particular assumptions regarding the distribution of perturbations, internal neural variables, and inputs as done for Madalines neural networks (Alippi, Piuri, & Sami, 1995; Stevenson, Winter, Widrow, 1990).

Some other authors tackle the robustness issue differently by suggesting techniques leading to neural networks with improved robustness ability by acting on the learning phase (e.g., see Alippi, 1999) or by introducing modular redundancy (Edwards & Murray, 1998); though, no robustness indexes are suggested there. The robustness of neural networks with respect to hardware implementations were also studied in Hereford and Kuyucu (2005) and Nugent, Kenyon, and Porter (2004) where authors proposed evolutionary and adaptive approaches.

Again, the study of robustness over training time has been evaluated for neural networks in the large, without assuming the small perturbation hypothesis (Alippi, Sana, & Scotti, 2004). In this direction, other authors have addressed the issue of the robustness analysis during the training phase (Manic & Wilamowski, 2002; Qin Juanyin, Wei Wei, & Wang Pan, 2004) by suggesting a genetic approach or by considering the use of the regression theory.

An overview of the sensitivity issues in neural networks can be found in Ng, Yeung, Xi-Zhao, and Cloete, (2004).

In this article, we suggest a robustness/sensitivity analysis in the large (i.e., without assuming constraints on the size or nature of the perturbation); as such, the small perturbation hypothesis becomes only a subcase of the theory. The suggested

sensitivity/robustness analysis can be applied to ALL neural network models (including recurrent neural models) involved in system identification, control signal/image processing and automation-based applications without any restriction to study the relationship between perturbations affecting the knowledge space and the induced accuracy loss.

A ROBUSTNESS ANALYSIS IN THE LARGE

In the following we consider a generic neural network implementing the $\hat{y} = f(\theta, x)$ function where $\hat{\theta}$ is the weight vector of the trained neural network.

In several neural models, and in particular in those related to system identification and control, the relationship between the inputs and the output of the system are captured by considering a regression vector φ, which contains a limited time-window of actual and past inputs, outputs, and -possibly- predicted outputs.

Of particular interest are those models, which can be represented by means of the model structures $\hat{y}(t) = f(\varphi)$ where function $f(\cdot)$ is a regression-type neural network, characterized by N_φ inputs, N_η non-linear hidden units, and a single effective linear/non-linear output (Hassoun, 1995; Hertz, Krog, & Palmer, 1991; Ljung, 1987; Ljung, Sjoberg, & Hjalmarsson, 1996).

We denote by $y_\Delta(x) = f_\Delta(\theta, \Delta, x)$ the mathematical description of the perturbed computation and by $\Delta \in D \subseteq \Re^p$ a generic p-dimensional perturbation vector, a component for each independent perturbation affecting the network weights of model $\hat{y}(t)$. The perturbation space D, is characterized in stochastic terms by providing the probability density function pdf_D.

To measure the discrepancy between $y_\Delta(x)$ and $y(t)$ or $\hat{y}(t)$ we consider a generic loss function $U(\Delta)$. A common example for U is the Mean Square Error –MSE- loss function:

$$U(\Delta) = \frac{1}{N_x} \sum_{i=1}^{Nx} (y(x_i) - \hat{y}(x_i, \Delta))^2 \qquad (1)$$

but a generic Lebesgue measurable loss function with respect to D can be taken into account (Jech, 1978). The formalization of the impact of perturbation on the performance function can be simply derived as:

Definition: Robustness Index

We say that a neural network is robust at level $\bar{\gamma}$ in D, when the robustness index $\bar{\gamma}$ is the minimum positive value for which:

$$U(\Delta) \le \bar{\gamma},\ \forall\ \Delta \in D,\ \forall \gamma \ge \bar{\gamma}. \qquad (2)$$

Immediately, from the definition of robustness index, we have that a generic neural network NN1 is more robust than another NN2 iff $\bar{\gamma}_1 < \bar{\gamma}_2$; the property holds independently from the topology of the two neural networks.

The main problem related to the determination of the robustness index $\bar{\gamma}$ is that we have to compute $U(\Delta)$, $\forall\ \Delta \in D$ if we wish a tight bound. The $\bar{\gamma}$-identification problem is therefore intractable from a computational point of view if we relax all assumptions made in the literature as we do.

To deal with the computational aspect we associate a dual probabilistic problem to (2):

Robustness Index: Dual Problem

We say that a neural network is robust at level $\bar{\gamma}$ in D with confidence η, when $\bar{\gamma}$ is the minimum positive value for which:

$$\Pr(U(\Delta) \le \bar{\gamma}) \ge \eta \ \text{ holds } \forall\ \Delta \in D,\ \forall \gamma \ge \bar{\gamma}. \qquad (3)$$

The probabilistic problem is weaker than the deterministic one since it tolerates the existence of a set of perturbations (whose measure according to Lebesgue is 1-η) for which $u(\Delta) > \bar{\gamma}$. In other words, not more than 100η% of perturbations $\Delta \in D$ will generate a loss in performance larger than $\bar{\gamma}$.

Probabilistic and deterministic problems are "close" to each other when we choose, as we do, η=1.

The non-linearity with respect to Δ and the lack of a priori assumptions regarding the neural network do not allow computing (2) in a closed form for the general perturbation case. The analysis, which would imply testing $U(\Delta)$ in correspondence with a continuous perturbation space, can be solved by resorting to probability according to the dual problem and by applying randomised algorithms (Alippi, 2002b; Bai, Tempo, & Fu, 1997; Tempo & Dabbene, 1999; Vidyasagar, 1996, 1998) to solve the robustness/sensitivity problem.

RANDOMIZED ALGORITHMS AND PERTURBATION ANALYSIS

In the following we denote by $p_\gamma = \Pr\{U(\Delta) \le \bar{\gamma}\}$ the probability that the loss in performance associated with perturbations in D is below a given—but arbitrary—value γ.

Probability p_γ is unknown, it cannot be computed in a form for a generic U function and neural network topology, and its evaluation requires exploration of the whole perturbation space D.

Anyway, the unknown probability p_γ can be estimated by sampling D with N independent and identically distributed samples Δ_i (intuitively a sufficiently large random sample explores the space); extraction must be carried out according

to the pdf of the perturbation. For each sample Δ_i we then generate the triplet:

$$\{\Delta_i, U(\Delta_i)\, I(\Delta_i)\},\ i = 1, N \text{ where } I(\Delta_i) \begin{cases} 1 \text{ if } U(\Delta_i) \le \gamma \\ 0 \text{ if } U(\Delta_i) > \gamma \end{cases}. \tag{4}$$

The true probability p_γ can now be simply estimated by means of the frequency as:

$$\hat{p}_N = \frac{1}{N}\sum_{i=1}^{N} I(\Delta_i). \tag{5}$$

Of course, when N tends to infinity, $\hat{p}_N$ somehow converges to p_γ. By introducing an accuracy degree ε on the difference $|p_\gamma - \hat{p}_N|$ and a confidence level 1– δ (which requests that the $|p_\gamma - \hat{p}_N| \le \varepsilon$ inequality is satisfied at least with probability 1– δ), our problem can be formalized by requiring that the inequality:

$$\Pr\{|p_\gamma - \hat{p}_N| \le \varepsilon\} \ge 1 - \delta \tag{6}$$

is satisfied for $\forall\gamma \ge 0$. In other words (6) states that the true probability and its estimates must be very close (they may differ not more than ε) and that the statement must be true with high probability. Of course, we wish to control the accuracy and the confidence degrees of (6) by allowing the user to choose the most appropriate values for the particular need. Finally, by extracting a number of samples from D according to the Chernoff bound (Chernoff, 1952):

$$N \ge \frac{\ln\frac{2}{\delta}}{2\varepsilon^2} \tag{7}$$

we have that $\Pr\{|p_\gamma - \hat{p}_N| \le \varepsilon\} \ge 1 - \delta$ holds for $\forall\gamma \ge 0$, $\forall\delta, \varepsilon \in [0,1]$.

As an example, by considering a 5% in accuracy and 99% in confidence we have to extract 1060 samples from D; with such choice we can approximate p_γ with $\hat{p}_N$ introducing the maximum error 0.05 ($\hat{p}_N - 0.05 \le p_\gamma \le \hat{p}_N + 0.05$) and the inequality holds at least with probability 0.99.

The Chernoff bound grants that the dual probabilistic problem related to the identification of the robustness index $\bar{\gamma}$ can be solved with a polynomial complexity algorithm in the accuracy and the confidence degrees independently from the number of weights of the neural model network. In fact, from (6) if accuracy ε and confidence δ are small enough we can confuse p_γ and $\hat{p}_N$ by committing a small error. As a consequence, the dual probabilistic problem requiring $p_\gamma \ge \eta$ becomes $\hat{p}_N \ge \eta$. We surely assume ε and δ to be small enough in subsequent derivations.

The final algorithm, which allows for testing the robustness degree $\bar{\gamma}$ of a generic neural network, can be summed up as:

1. Select ε and δ sufficiently small to have enough accuracy and confidence
2. Extract from D, according to its pdf, a number of perturbations N as suggested by (7)
3. Generate the indicator function $I(\Delta)$ and generate the estimate $\hat{p}_N = \hat{p}_N(\gamma)$ according to (5)
4. Select the minimum value γ_η from the $\hat{p}_N = \hat{p}_N(\gamma)$ function so that $\hat{p}_N(\gamma_\eta) = 1$ is satisfied $\forall\gamma \ge \gamma_\eta$. γ_η is the estimate of the robustness index $\bar{\gamma}$.

Note that with a simple algorithm we are able to estimate in polynomial time the robustness degree $\bar{\gamma}$ of a generic neural network. The accuracy in estimating $\bar{\gamma}$ can be made arbitrarily good by considering a larger number of samples as suggested by the Chernoff's bound.

SOME EXPERIMENTAL RESULTS

To shed light on how the methodology can be used to test the robustness of a neural network we focus the attention on a regression-type experiment. In particular, we consider the simple error-free function:

$$y = -x \cdot \sin(x^2) + \frac{e^{-0.23 \cdot x}}{1 + x^4}, \quad x \in [-3, 3]$$

and we approximate it with a regression-type neural network having 10 hidden units. Learning is then perfected with a Levenberg-Marquardt training algorithm (Hassoun, 1995).

We then wish to test the robustness of the trained neural network in the large (i.e., by not assuming the small perturbation hypothesis).

Perturbations affecting weights and biases of the neural network are defined in a perturbation space subject to a uniform distribution. In particular, a perturbation Δ_i affecting a generic weight w_i must be intended as a relative perturbation with respect to the weight magnitude according to the multiplicative perturbation model $w_{i,p} = (1 + \Delta_i)$, $\forall i = 1, n$ (e.g., see Edwards et al., 1998). As such, a t% perturbation implies that Δ_i is drawn from a symmetrical uniform distribution of extremes

$$\left[-\frac{t}{100}, \frac{t}{100}\right];$$

a 5% perturbation affecting weights and biases composing vector $\hat{\theta}$ hence requires that each weight/bias is affected by an independent perturbation extracted from the [– 0.05, 0.05] interval and applied to the nominal value according to the multiplicative perturbation model. By applying the algorithm suggested in the previous section we determine the $\hat{p}_\gamma = \hat{p}_\gamma(\gamma)$ functions corresponding to the 1%, 3%, 5%,

10%, 30% perturbations. Results are given in figure 1 where we considered $\varepsilon = 0.02$, $\delta = 0.01$ leading to N=6624.

From its definition, $\bar{\gamma}$ is the smallest value for which $\hat{p}_\gamma = 1$, $\gamma \geq \bar{\gamma}$; in the figure, if we consider the 5% perturbation case, $\bar{\gamma}$ assumes a value around 7. We observe that by increasing the strength of perturbation (i.e., by enlarging the extremes of the uniform distribution characterizing the pdf of D) $\bar{\gamma}$ increases. In fact, stronger perturbations have a worse impact on the performance loss function since the error-affected neural network diverges from the error-free one. Conversely, we see that small perturbations (e.g., the 1% one) induce a very small loss in performance since the robustness index γ_η is very small.

Another interesting use of the suggested methodology allows the neural network designer to test different neural models and identify, for instance, the most robust one.

To this end, we consider a set of performance-equivalent neural networks, each of which able to solve the application with a performance tolerable by the user. All neural networks are characterized by a different topological complexity (number of hidden units).

The $\hat{p}_\gamma = \hat{p}_\gamma(\gamma)$ curves parameterized in the number of hidden units are given in Figure 2 in the case of 1% perturbation. We see that by increasing the number of hidden units $\bar{\gamma}$ decreases. We immediately realize that neural networks with a reduced number of hidden units are, for this application, less robust than the ones possessing more degrees of freedom. For this application, large networks provide, in a way, a sort of spatial redundancy: information characterizing the knowledge space of the neural networks is distributed over more degrees of freedom.

It should be clear, anyway, that robustness is a property of the identified neural model and strongly depends on the neural network complexity, the envisaged training algorithm, the training starting point and the training data samples. Hence, for a generic application, it is not possible to assert that by increasing the network complexity (i.e., by increasing the number of hidden units) we improve the robustness of the obtained model.

We then tested the robustness of two trained neural networks in the large (i.e., by not assuming the small perturbation hypothesis) in case of a classification problem. We considered two equiprobable classes C0 and C1 with bidimensional Gaussian distributions:

$$(X,Y) \sim N\left(\begin{bmatrix} \mu_X & \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_X^2 & 0 \\ 0 & \sigma_Y^2 \end{bmatrix} \right)$$

$$\text{C0: } (X,Y) \sim N\left(\begin{bmatrix} 0 & 0 \end{bmatrix}, \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix} \right)$$

$$\text{C1: } (X,Y) \sim N\left(\begin{bmatrix} 10 & 10 \end{bmatrix}, \begin{bmatrix} 9 & 0 \\ 0 & 9 \end{bmatrix} \right).$$

We considered two neural networks: a feed-forward and a probabilistic neural network.

Similarly to what done for the regression-type experiment, we assumed perturbations affecting weights and biases to be defined in a perturbation space subject to the symmetrical uniform distribution of extremes

$$\left[-\frac{t}{100}, \frac{t}{100} \right],$$

Figure 1. $\hat{p}_\gamma$ as a function of γ for the 10 hidden units neural network

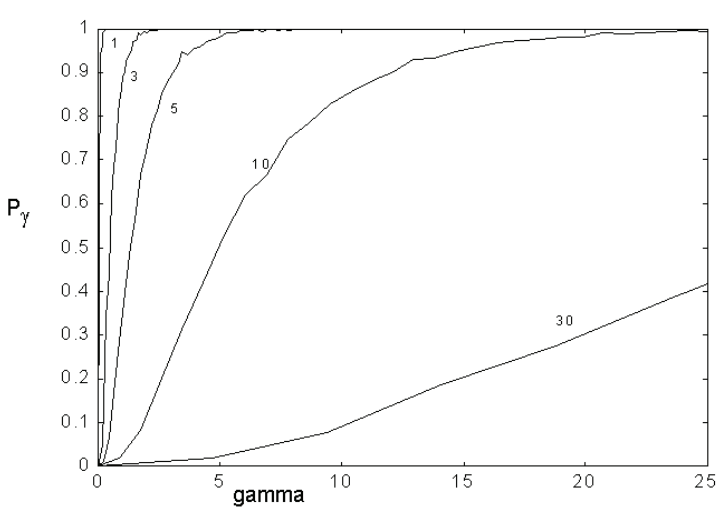
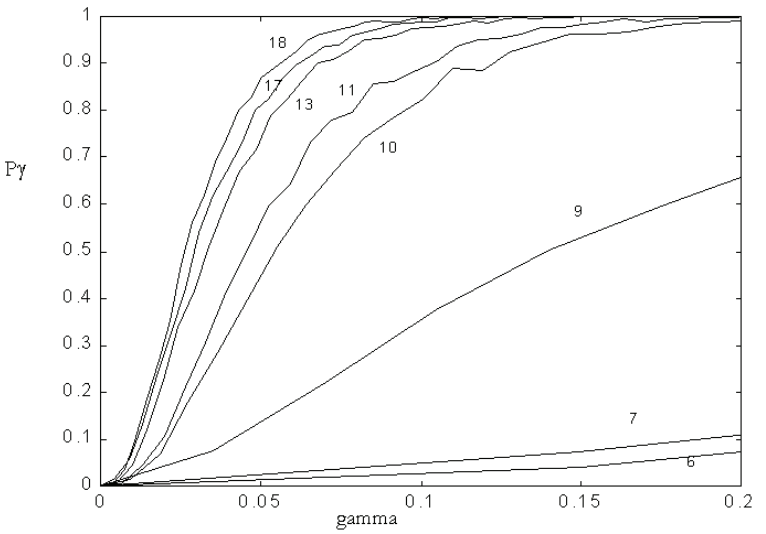


Figure 2. $\hat{p}_\gamma$ over γ and parameterized in the number of hidden units



with t equals to 1,3,5,10 or 30. Even in this case, the perturbation Δ_i affecting a generic weight w_i must be intended as a relative perturbation with respect to the weight magnitude according to the multiplicative perturbation model $w_{i,p} = (1 + \Delta_i)$, $\forall i = 1, n$.

By applying the algorithm suggested in the previous section we determined the $\hat{p}_\gamma = \hat{p}_\gamma(\gamma)$ functions corresponding to the 1%, 3%, 5%, 10%, 30% perturbations for the feed-forward neural network (10 hidden units neural network). Results are given in Figure 3 where we considered $\varepsilon = 0.02$, $\delta = 0.01$ (N=6624).

Even in this case it is possible to identify $\bar{\gamma}$ by definition; for instance, in the figure $\bar{\gamma}$ assumes a value around 0.4 for the 5% perturbation case. As expected by increasing the strength of perturbation (i.e., by enlarging the extremes of the uniform distribution characterizing the pdf of D) $\bar{\gamma}$ increases. It is worth noting that even small perturbations (e.g., the 1% one, induce a not neglectible loss in performance).

We then applied the suggested algorithm to the probabilistic neural network, whose spread value was set at 0.1. We evaluated the $\hat{p}_\gamma = \hat{p}_\gamma(\gamma)$ functions corresponding to the 1%, 3%, 5%, 10%, 30% perturbation cases: results are given in figure 4. Even in this case we considered $\varepsilon = 0.02$, $\delta = 0.01$.

As expected by increasing the strength of the perturbation, $\bar{\gamma}$ increases. Differently from the feed-forward neural network, small perturbations induce a very small loss in performance since the robustness index γ_η is very small.

By analyzing the last two figures, we realize that, in the considered classification problem, the probabilistic neural network is more robust than the feed-forward neural network with 10 hidden units. This is evident by considering the different values of $\bar{\gamma}$ provided by the probabilistic and the feed forward neural network in case of the considered perturbations. The values of $\bar{\gamma}$ of the probabilistic neural network are significantly lower than the ones provided by the feed-forward network.

However, it is important to recall that, as previously defined, robustness is a property of the identified neural model and strongly depends on the neural network complexity, the envisaged training algorithm, the training starting point and the training data samples. It is thus impossible to state that the probabilistic neural networks are generally more robust than the feed-forward ones in case of classification problems.

FUTURE TRENDS

As explained in the article, the robustness analysis in the large issues allows the designer to solve all limits posed by the small perturbation hypothesis. Anyway, further research needs to be done in this direction by enlarging the application domain (e.g., by considering recurrent neural networks and assessing with the suggested methodology stability issues) and studying the training modality. The latter aspect is of primary relevance to understand the intrinsic nature of the learning algorithms, how it works for an ensemble of neural models, and their efficacy in proximity of the minimum of the figure of merit associated with training and validation.

CONCLUSION

A robustness analysis in the large (i.e., by not assuming any hypothesis about the perturbation strength) can be ad-

Figure 3. $\hat{p}_\gamma$ *as a function of* γ *for the feed-forward neural network*

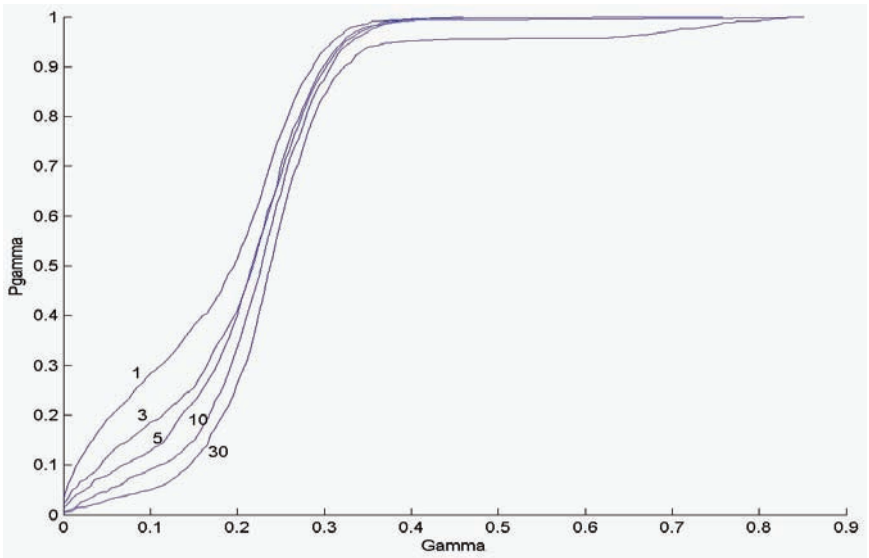
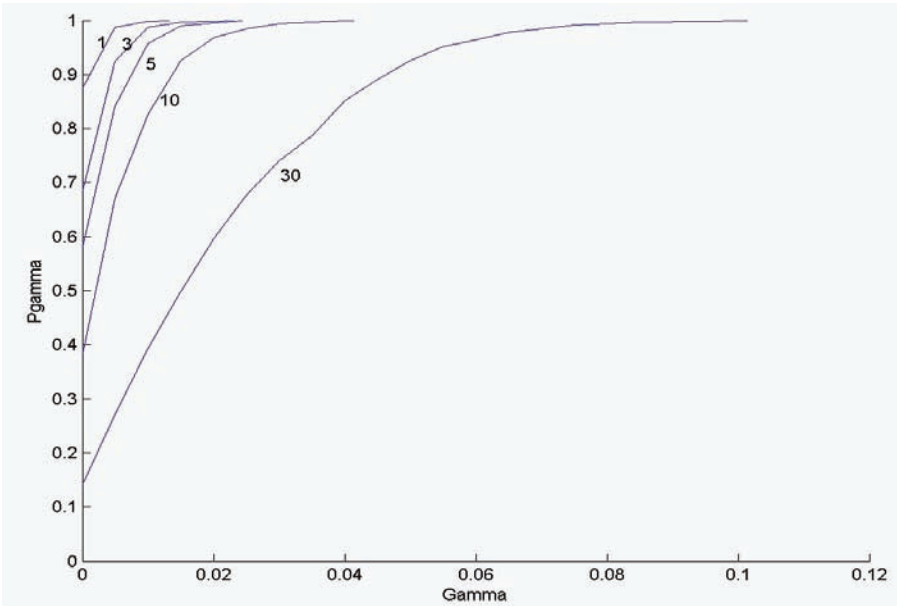


Figure 4. $\hat{p}_\gamma$ *as a function of* γ *for the probabilistic neural network*



dressed for a generic neural network model and accuracy loss figure of merit by resorting to randomization. In reality, the robustness analysis is even more general and can be considered for any Lebesgue-measurable computation (basically, all functions involved in signal/image processing are Lebesgue-measurable).

Hence, by considering trained neural networks, we can easily estimate the effects induced by perturbations affecting a generic neural network by considering a probabilistic approach. The robustness index, which can be used to investigate the relationships between knowledge space and neural network accuracy, can be computed with an effective poly-time algorithm, which spouses Monte Carlo and learning theories sampling methods.

REFERENCES

Alippi, C. (2002a). Randomized algorithms: A system level, poly-time analysis of robust computation. *IEEE Transactions on Computers, 51*(5).

Alippi, C. (2002b). Selecting accurate, robust, and minimal feedforward neural networks. *IEEE-Transactions on Circuits and Systems: Part I, Fundamental theory and applications*, *49*(12), 1799-1810.

Alippi, C. (1999, May 30-June 2). Feedforward neural networks with improved insensitivity abilities. *Proceedings of the IEEE-ISCAS99*, Orlando, Florida, USA, 1999.

Alippi, C., & Briozzo, L. (1998). Accuracy vs. precision in digital VLSI architectures for signal processing. *IEEE Transactions on Computers, 47*(4).

Alippi, C., Piuri, V., & Sami, M. (1995). Sensitivity to errors in artificial neural networks: A behavioural approach. *IEEE Transactions on Circuits and Systems: Part 1, 42*(6).

Alippi, C., Sana, D., & Scotti, F. (2004, July 25-29). A training-time analysis of robustness in feed-forward neural networks. In *Proceedings of the 2004 IEEE International Joint Conference on Neural Networks* (Vol. 4, pp. 2853-2858), 2004.

Bai, E., Tempo, R., & Fu, M. (1997). Worst-case properties of the uniform distribution and randomized algorithms for robustness analysis. In *Proceedings of the IEEE-American Control Conference* (pp. 861-865). Albuquerque.

Chernoff, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annuals of Mathematical Statistics*, *23*(4), 493-507.

Dundar, G., & Rose, K. (1995). The effects of quantization on multilayer neural networks. *IEEE-TNN, 6*(6), November.

Edwards, P. J., & Murray, A. F. (1998). Towards optimally distributed computation. *Neural Computation, 10*(4), 987-1005.

Hassoun, M. H. (1995). *Fundametals of artificial neural networks*. The MIT Press.

Hereford, J. M., & Kuyucu, T. (2005). Robust neural networks using motes. In *Proceedings of the Evolvable Hardware* (pp. 117-124), 2005. NASA/DoD.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation*. Addison-Wesley Publishing Co.

Holt, J., & Hwang, J. (1993). Finite precision error analysis of neural network hardware implementations. *IEEE- TC, 42*(3), March.

Jech, T. (1978). *Set theory* (series Pure and Applied Mathematics). New York: Academic Press.

Ljung, L. (1987). *System identification: Theory for the user*. Prentice-Hall.

Ljung, L., Sjoberg, J., & Hjalmarssoon, H. (1996). On neural networks model structures in system identification, in identification, adaptation, learning. NATO ASI series. *Series F: Computer and System Sciences*, *153*. Springer.

Manic, M., & Wilamowski, B. (2002, November 5-8). Towards the robustness in neural network training. In *Proceedings of the Industrial Electronics Society, IECON 02* (Vol. 3, pp. 1768-1771), 2002.

Ng, W. W. Y., Yeung, D. S., Xi-Zhao, W., & Cloete, I. (2004, August 26-29). A study of the difference between partial derivative and stochastic neural network sensitivity analysis for applications in supervised pattern classification problems. In *Proceedings of the Machine Learning and Cybernetics* (Vol. 7, pp. 4283-4288).

Nugent, A., Kenyon, G., & Porter, R. (2004, June 24-26). Unsupervised adaptation to improve fault tolerance of neural network classifiers. In *Proceedings of the Evolvable Hardware* (pp. 146-149), Proceedings of the NASA/DoD.

Piché, S. (1995). The selection of weights accuracies for Madalines. *IEEE Transactions on Neural Networks, 6*(2).

Qin, J., Wei, W., & Wang, P. (2004, June 15-19). Robust learning of neural networks ensemble for modeling. *Proceedings of the Intelligent Control and Automation* (Vol. 3, pp. 1927-1930). WCICA 2004.

Stevenson, M., Winter, R., & Widrow, B. (1990). Sensitivity of feedforward neural networks to weights errors. *IEEE Transactions on Neural Networks, 1*(1).

Tempo, R., & Dabbene, F. (1999). Probabilistic robustness analysis and design of uncertain systems. *Dynamical Systems, Control, Coding, Computer Vision—New Trends, Interfaces, and Interplay*, *25*, 263-282.

Vidyasagar, M. (1998). Statistical learning theory and randomized algorithms for control. *IEEE-Control Systems Magazine*, *18*, 69-85.

Vidyasagar, M. (1996). *A theory of learning and generalisation with applications to neural networks and control systems*. Berlin: Springer-Verlag.

KEY TERMS

Knowledge Space: The space defined by the neural network weights.

Neural Network Weights: The free parameters of the neural model.

Perturbation: A behavioral entity affecting the weights of a neural network.

Poly-Time Algorithm: An algorithm whose complexity evolves polynomially with respect to the envisaged complexity parameter.

Regression-Type Neural Network: A neural network with one hidden layer and a unique linear output neuron used to approximate a mathematical function.

Randomized Algorithms: A probabilistic sampling technique for exploring a space combining learning theories and Monte Carlo approaches.

Robustness: A property possessed by a system. A system is robust with respect to a perturbation when the perturbation effect on the system performance is tolerable according to a predefined threshold.

Small Perturbation Hypothesis: The strength of the envisaged perturbations is small enough to guarantee effectiveness for a linearized sensitivity analysis.

Wireless Sensor Network: A network of distributed sensors linked by wireless connections.

The Role of Business Case Development in the Diffusion of Innovations Theory for Enterprise Information Systems

Francisco Chia Cua
Otago Polytechnic, New Zealand

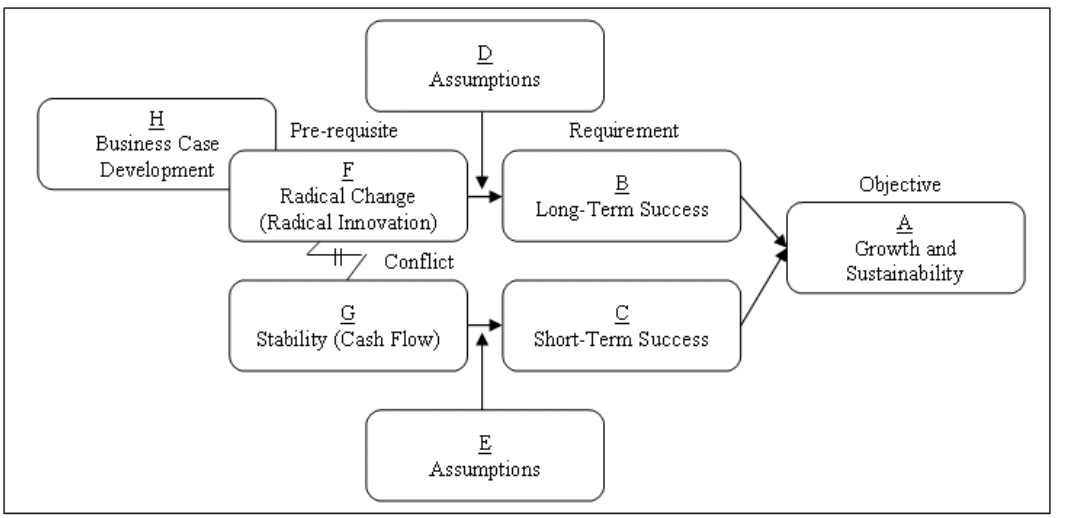
Tony C. Garrett
Korea University, Republic of Korea

INTRODUCTION

A successful organisation continually initiates and implements radical innovations. The innovation must not only be new. A radical innovation has a significant impact on how the organisation undertakes its business process. Impacting is different from affecting. The former has a more substantial effect on the organisation. This is precisely why new enterprise information systems represent a radical innovation. To be successful, the organisation undertakes an innovation-decision process to align itself, as much as possible, with the ever-changing external realities. The **innovation-decision process** dictates selling an idea (the business case) that the new enterprise information systems possess economic value to upper management.

This paper depicts a bird's-eye view of how innovation, in this case, the new enterprise information systems, diffuses (episteme) via business case development (techne) in the innovation-decision process. As shown in Figure 1, the adoption and implementation of new enterprise information systems constitute a radical change (prerequisite F). New enterprise information systems represent radical innovation. An **innovation-decision process** starts with an **initiation phase** through which the individuals or decision-making units move from identifying and knowing the new enterprise information systems, to the forming of an attitude toward the different competing software packages, and subsequently to deciding whether to adopt or reject the implementation and use of the new idea. A **business case** is a formally written document that argues about the adoption to a certain course

Figure 1. Conflict between radical change and stability



Interpreted from Burrell & Morgan (2005), Dettmer (2003), Trompenaars & Prud'homme (2004)

of action. It contains a point-by-point analysis to making a decision for a set of alternative courses of action to accomplish a specific goal. A **business case process** walks through the **initiation phase** of the innovation-decision process and talks about the project plans that concern the **implementation phase**, which follows the initiation phase. The **business case document** justifies, in detail, the innovation-decision process: what has transpired in the initiation phase and what will transpire in the implementation phase. It takes into account the innovation-decision process. In short, a business case process develops a detailed business case document of the innovation-decision process. Thus, a business case is both a means and an end.

An innovation-decision process is required to foster long-term success (Figure 1; Burrell & Morgan, 2005; Dettmer, 2003; Trompenaars & Prud'homme, 2004). Stability is another factor, that enhances short-term success. The crucial component of stability is the cash flow, the lifeblood of the organisation. Profitability is vital in generating the cash flow. Profitability is a crucial issue to stability, while uncertainty, risk management, and governance are crucial issues to radical innovation. The organisation cannot exploit the radical innovation effectively unless it manages uncertainty, mitigates the risk related to the uncertainty, and governs the innovation-decision process appropriately. Because the innovation-decision process disrupts stability, radical innovation and stability are in **conflict** with each other.

Inasmuch as a business case document must justify the radical innovation and its innovation-decision process, the business case must resolve the many issues and assumptions that affect the radical innovation, stability, and the conflict between them. There are two questions about radical innovation at the broadest level (Borge, 2001; Nadler & Tushman, 2004, pp 554-555; Trompenaars & Prud'homme, 2004). What innovation should the organisation prioritise? How should the organisation carry out the innovation-decision process? The first overriding question concerns the radical innovation (the object of innovation) and the change vision (the expected consequence). The second question assumes that the short-listed options in the business case provide the solution to achieve the change vision. The change vision represents an expected consequence, preceding the innovation-decision process. Another antecedent is the expected bad outcomes to avoid. The proximate trigger and the perceived attributes of the innovation likewise influence the executive sponsor to be in favour or against a certain package (brand) of the enterprise information systems.

The **diffusion of innovations (DOI)** (Rogers, 2003) is the theory of focus. DOI has three constructs: the antecedents, the innovation-decision process, and the consequences, especially the unexpected consequences.

- DOI underscores the antecedents of the innovation-decision process. The change vision, the proximate trigger, the expected "bad outcome" to avoid, and the perceived attributes of the innovation are important antecedents. There are other antecedents, such as organisational innovativeness, and the external and internal environments.
- DOI draws attention to the innovation-decision process (Baskerville & Pries-Heje, 2001; Bradford & Florin, 2003; Dechow & Mouritsen, 2004, 2005; Dillard, 2000; Dillard, Ruchala, & Yuthas, 2005; Van de Ven & Poole, 1990; Zaltman, Duncan, & Holbek, 1973). DOI consists of the initiation phase and the implementation phase (Rogers, 2003). There are stages in each phase. The initiation phase incorporates the agenda-setting stage, matching stage, and adoption decision stage. The implementation phase of new enterprise information systems includes the preproduction, production, postproduction, and confirmation stages.
- DOI highlights undesirable consequences. The executive sponsor, who is the champion of the initiative to implement the new enterprise systems, acknowledges their success when the information systems go live. Yet research suggests that organisations fail to achieve outstanding bottom-line improvements even after 3 to 5 years of going live (Baskerville & Pries-Heje, 2001). This is **successful failure**, the undesirable consequence of implementing new enterprise information systems. Three illustrations of undesirable consequences are given in Section 2.

This section has stated the prerequisites and requirements to achieve growth and sustainability (Figure 1). It introduces DOI theory, and highlights its three main constructs: the antecedents, the innovation-decision process, and the consequences. Section 2 makes clear and illustrates the undesirable consequences of innovation. The terminologies of DOI differ, but not their essence. The business case is a part of the matching stage in the initiation phase. Understanding the framework of the innovation-decision process simplifies the complexity of business case development. It is particularly useful toward developing a structured approach to writing a business case. Section 2, section 3, and section 4 clarify the concepts and framework of DOI and the innovation-decision process. Section 5 ends with the implications of the business case development (that is, the *episteme* and *techne* of DOI and the business case development).

UNDESIRIBLE CONSEQUENCES

The studies of the villagers of Los Molinas in Peru illustrate an unexpected consequence. The villagers in Los Molinas stubbornly drank water from a canal floating with dead monkeys instead of a nearby tap with clean drinking water. To combat the infections caused by the contaminated water,

the Peruvian government and health authorities instructed the villagers to boil the water. Even after 2 years of diffusing the new idea to the villagers (e.g., getting the clean water from the tap, boiling the water, and drinking the boiled water), a majority continued to drink from the filthy river. The diffusion process ignored the culture of Peruvian villagers (Rogers, 2003, pp 1-5): Healthy people do not drink boiled water; only sick people do.

An undesirable consequence is also seen in what happened to FoxMeyer. In December 1995, FoxMeyer installed SAP and subsequently, went into Chapter 11 bankruptcy proceedings. It claimed that its innovation-decision process of the enterprise information systems was one of the factors that eventually led to its bankruptcy (O'Leary, 2000; "SAP and Deloitte Sued by FoxMeyer," 27 Aug 1998). The undesirable consequence in the context of the new enterprise information systems is the successful failure explained in the previous section.

Uncertainty and risk are unavoidable with the innovation-decision process of new enterprise information systems. Wrong turns and missteps occur frequently (Van de Ven & Poole, 1999). One thing is certain. Of the many factors affecting the success or failure of implementing new enterprise information systems, people are a significant variable and therefore, organisational risk is the greatest risk (O'Leary, 2000; Van de Ven & Poole, 1999).

The "*episteme*" (knowledge) of the business case development in context of the new enterprise information systems (Allen Sr, 2005; Benco, 2004; Carson III, 2005; Mabert, Soni, & Venkataraman, 2000) makes the "*techne*" (practice) of business case development less complicated in the end. The "*episteme*" requires understanding the exact evidences at every stage of the innovation-decision process. Only the evidence at the agenda-setting stage is clear in diffusion studies (Rogers, 2003). The evidence is not clear in the matching stage and decision stage of the initiation phase. There is very limited evidence in the implementation phase. These research gaps reiterate the rigour necessary in developing the business case.

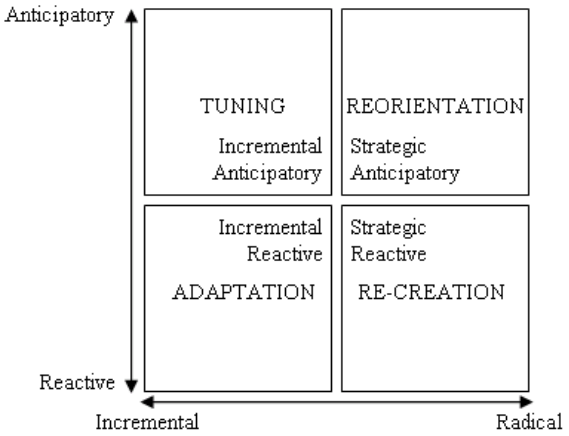
A cause of undesirable consequences relates to the quantity [and quality] of information, of knowledge, in the decision points (Ormerod, 2005, pp 23-25). The uncertainty or the probability of the consequences, of the new enterprise information systems, is in itself unknown. The concept of risk involves quantifying the probabilities. The business case document not only details the innovation-decision process, it quantifies the perceived risks inherent to the options in the decision stage of the initiation phase.

INNOVATION IN ORGANISATIONS

Organisational innovation generally involves a range of ideas that include reinvention (Van de Ven & Poole, 1999). The executive sponsor and the change agent learn and implement some ideas and discard others. The timeframe in radical innovation is generally long; so is the gestation during the innovation-decision process (Van de Ven & Poole, 1999). The innovation-decision process in organisations is like a mind map (Cua & Theivananthampillai, 2006). It diverges and converges, involving many people, multiple actions, and series of evaluations during implementation (Van de Ven & Poole, 1999). Not only is the innovation-decision process far more complex and unpredictable, with long gestation periods and multiple stops and starts, it is sensitive to [and constrained by] external influences (Van de Ven & Poole, 1999). With the new enterprise information systems, the radical innovation fosters new ways to undertaking the business process, and creates the environment that enhances the learning. Consequently, the new enterprise information systems create the environment, but are simultaneously constrained by the environment.

Innovation comes in a variety of dimensions (Christensen & Raynor, 2003; Cooper, 1998; Cua & Theivananthampillai, 2006; Nadler & Tushman, 1986; Tidd, Bessant, & Pavitt, 2001, p 7). Innovation may be sustaining or disruptive, anticipatory or reactive, and incremental or radical, process, service, or product, and technological or administrative. Organisations seek to build new-growth business in two ways. One is to build on the existing market from entrenched competitors with sustaining innovation by offering more functions and better features. The other way is to disrupt the competitors with innovation that creates new markets by targeting nonconsumers, or take root among the incumbent's worst customers (that is, the low end of an established market). The low-end product is cheaper with lesser or simpler functions. Christensen (1997) calls this type of innovation the disruptive innovation. Anticipatory innovation initiates innovation in anticipation of competitive advantage. Reactive innovation is an urgent, normally unplanned, response to an external event or change. Radical innovation redefines what the organisation is, and modifies the basic organisational framework composing of strategy, structure, people, processes, and core values. Incremental innovation does not modify the value, mode of organising, and general strategic framework. Furthermore, the innovation may be a business process, service, or product. The multiple dimensions allow upper and middle managers to investigate the innovation with their corresponding mechanisms and strategic advantages (Cooper, 1998). Regardless of the forms of innovation, exploiting the innovation must be understood beyond these dimensions. This is a necessary process that must happen. However, there must be diffusion before making this happens. The diffusion is the timely communication of the new idea through the business case. The business case document is a mechanism to facilitate the diffusion, which is the adoption decision.

Figure 2. Organisational innovation of Nadler and Tushman



Nadler and Tushman (1986) categorise innovation within an organisation as incremental rather than radical, and reactive as opposed to anticipatory. This results in two types of radical innovation (reorientation and recreation; Figure 2). As mentioned above, the change vision is a crucial antecedent to undertaking radical innovation. As an ongoing process, the strategic visioning must take place with enterprise redesign, value-stream reinvention, procedure redesign, and total quality management (TQM; Martin, 1995, pp 61, 384-385). Reorientation however requires greater strategic visioning than recreation. Recreation, however, demands greater care. Recreation is riskier than reorientation. It has a 90% mortality rate. The radical innovation, whether reactive or anticipatory, must balance with the stability (Figure 1) to mitigate the conflict between the radical change and stability. The business case development embodies the conflict.

PREREQUISITES FOR THE INNOVATION-DECISION PROCESS

The diffusion paradigm cuts across all social science disciplines (Rogers, 2003). The innovation-decision process, however, is articulated differently, depending on the discipline. One articulation is the business case development (project management, information technology, and information science). Other articulations are radical change process (change and innovation), strategic sourcing (logistics and supply chain), strategic investment decision (finance, accounting, and management), and capital investment process (finance and accounting).

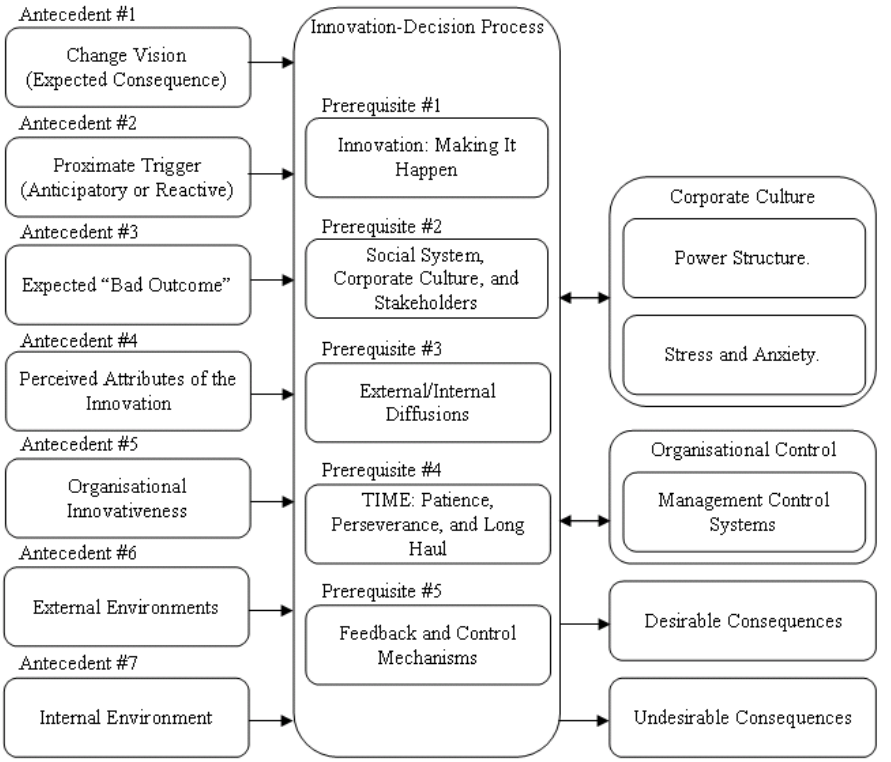
In finance and accounting parlance, organisations undertaking the innovation-decision process regards it as a "capital investment process" to maximise value and to enhance growth and sustainability (Figure 1) How the upper managers, the executive sponsor, and the project team members maximise shareholder value is an agency problem (Brealey & Myers, 2000). The shareholders are the ultimate principals. The upper management is their agent. The middle managers and the project team members are, in turn, the agents of the upper management. Thus, the middle managers are agents in relation to shareholders, and principals with regard to the rest of the organisation. The business case embodies the agency issue. Reflecting the agency issue, the exact evidence at each decision point in the initiation phase, the assumptions, the expected risk, and the performance measurement matter in the business case development.

However, shareholders are not the only concern. How the organisation is able to meet the needs of the other stakeholders defines the success of the innovation-decision process. The stakeholders analysis to maximise shareholder value in the longer-term (Morin & Jarrell, 2001; Rüegg-Stürm, 2005) is thus, an important prerequisite (#2 in Figure 3) in the business case development.

The five prerequisites of the innovation-decision process (Cua & Theivananthampillai, 2006; Nadler & Tushman, 1986, 2004; Rogers, 2003; Trompenaars & Prud'homme, 2004) are:

1. Innovation. An innovation is necessary for radical change (Trompenaars & Prud'homme, 2004). Foremost is its change vision (Slagmulder, 1997). The perceived newness, trialability, and observability (Rogers, 2003) ultimately affect the adoption (or rejection) decision. It involves analysing cost-benefit, mitigating uncertainty, and minimising risk necessary through information seeking and processing from RFI and RFP. Questions include (Rogers, 2003, p 14): What is the newness? How does the innovation work? Why does it work? What are the expected consequences? How cost-beneficial is it? What are the bad outcomes to avoid? With enterprise information systems, reinvention poses a difficult challenge.
2. Social system. Diffusion must occur in a social system (Rogers, 2003, p 24), which may comprise stakeholders (individuals, groups, organisations, or their subgroups). The success of diffusion lies in understanding the social system, the early adopters, their culture, norms, challenges, and value. How members in a social system interact or relate formally or informally matters. Because of their technical competence and accessibility, opinion leaders exert influence in their homogenous networks. Change agents, as will be explained in

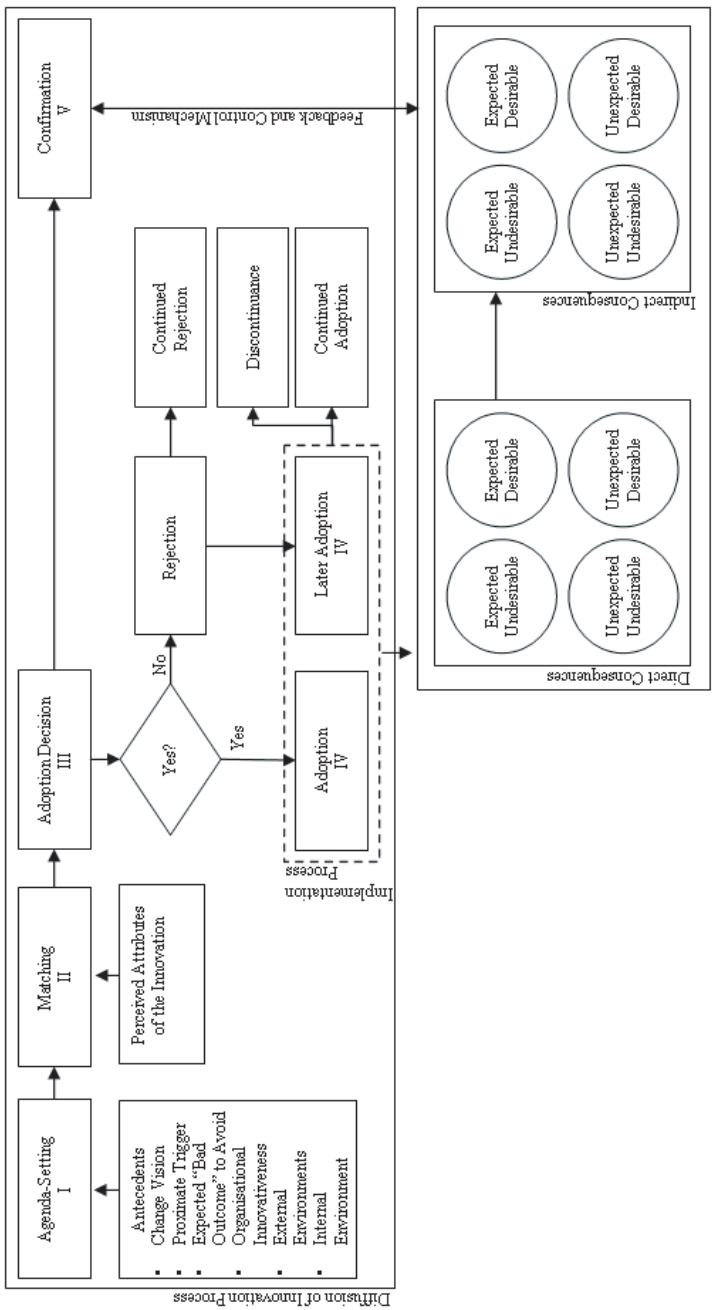
Figure 3. Conceptual framework of business case development Adopted from Burrell & Morgan (2005), Cua & Theivananthampillai (2006), Dettmer (2003), Nadler & Tushman (2004), Rogers (2003) , Trompenaars & Prud'homme (2004)



paragraph 5, are more effective with heterogeneous network, provided they employ the opinion leaders as their lieutenants. The middle managers also play important roles in the diffusion.

3. Communication channels. Studies show that most people do not analyse cost-benefit objectively (Cua & Theivananthampillai, 2006). Rather, they decide or respond to the innovation based on subjective evaluation from opinion leaders. In diffusion, interpersonal network is generally effective to foster a critical mass (Mahler & Rogers, 1999). Diffusion is more effective when people have similar culture or profile. Mass media channels are generally much faster and more efficient during the first stage of innovation-decision process. Interpersonal communication, however, is more effective in the next stage. In short, communication channels influence the speed of diffusion and the rate of adoption.
4. Time. The progress from one stage to another stage in the innovation-decision process discloses how crucial time functions in innovation. Patience is a virtue for adoption to advance toward reaching the critical mass. In context of time are the five broad categories of people who will adopt innovations, namely: the innovators, early adopters, early majority, late majority, and laggards.
5. Change agents. In a corporate setting undergoing radical change, there are individuals who influence an innovation-decision process and the people involved with the process (Rogers, 2003, p 366). They are the change agents. Their influences are usually in favour of the adoption of the new idea. Their priorities are different from those of the middle managers (Van de Ven & Poole, 1999). Change agents give emphasis to input criteria (antecedents) at the start of innovation process and later, to outcome criteria, whereas the middle managers are more concerned with the outcome criteria at the start and at the end, they look into the input criteria. The whole innovation-decision process embodies a range of information-seeking and informa-

Figure 4. The innovation-decision process



tion-processing activities pursued by decision makers to reduce uncertainty (Rogers, 2003). Inasmuch as radical change disrupts the status quo of organisation (Figure 1), organisational control (aka, management control systems) is vital. The conflicting priorities between the change agents and the middle managers compel timely and regular communications among the upper managers, middle managers, project team members, and other stakeholders.

Table 1. Important constructs of the business case development

Pre-requisites	1. Object of innovation (the new enterprise information systems)
	2. The social systems (the organisation, corporate culture, stakeholders, and extended social systems)
	3. External diffusion (external marketing) / internal diffusion (internal marketing)
	4. Time (patience, perseverance, long haul)
	5. Feedback and control mechanisms (management control systems)
Antecedents	1. Change vision (expected consequence)
	2. Proximate (anticipatory or reactive) trigger
	3. The expected "bad outcome" to avoid
	4. The perceived attributes of the innovation
	5. Organisational innovativeness
	6. External environments
	7. Internal environment
Process	1. Initiation phase: Agenda-setting stage, matching stage (RFI, RFP, and Business Case), adoption (or rejection) stage
	2. Implementation phase: Pre-production stage, production stage, post-production stage, confirmation stage
Consequences	1. Desirable consequences
	2. Undesirable consequences

In summary, Figure 3 maps out the five prerequisites in a change process. Diffusion of innovation shows innovation (#1 the new idea) as a process diffuses. Communication about innovation happens through mass media and personal interaction (#3 communication channels), and progresses through a structure of a social system (#2 social system) over a relatively long time frame (#4 time).The vital missing link in the process, where people are aware, make decisions, and manage uncertainty, is information. Timely information highlights the crucial management control systems (#5 organisational control) in a corporate setting undergoing change. Table 1 summarises the important constructs of the business case development.

THE COMPLEX INNOVATION DECISION PROCESS

The innovation-decision process of new enterprise information systems generally pass through the finance division (Rogers, 2003). It consists of five complex stages (Figure 4). The agenda-setting stage begins with learning about the innovation and understanding what it does and how it works. The matching stage happens when the executive sponsors and their team members incline toward or away from the innovation. The decision stage includes all activities leading to acceptance or rejection of the innovation. The implementation stage represents actual use of innovation, while the confirmation stage involves a search of evidence to support that the decision was right. Sometimes, it may lead to abandoning the innovation when the experiences or messages show conflict.

The innovation-decision process moves from something known to something unknown and disrupts status quo. The uncertainty exerts pressures on organisational power, anxiety, and control (Figure 3) and subsequently affects the innovation-decision process itself. The uncertainty and the complex interactions of the critical elements of the innovation-decision process threaten the desirable expected consequences. The risk of being unable to achieve the expected consequences and the risk of encountering unexpected consequences are inevitable.

CONCLUSIONS

Growth and sustainability! Organisations consider these big words the ultimate goal, and utilise radical innovation as the engine to achieve them. While exploiting the radical innovation, life has to go on. The stability of the current state makes it possible to maintain stability and generate cash flow amidst chaos and radical innovation. Both radical change and stability are necessary to ensure longer-term growth and sustainability. However, the essence of change opposes stability.

The business case development embodies the radical change and its conflict with the stability (Figure 1). The adoption and implementation of new enterprise information systems represent a radical business process innovation, which is complex and risky. A reactive radical innovation has greater risk, and demands greater care than an anticipatory radical change (Figure 2). In undertaking the innovation-decision process, the first question concerns thoroughly understanding the change vision, justifying why the organisation should prioritise the new enterprise information systems, and determining whether the radical innovation is anticipatory or reactive. The upper managers will be interested in getting detailed answers to the second set of questions. Why should the organisation prioritise the new enterprise information systems? How did the organisation carry out the initiation phase? How does the organisation undertake the adoption decision stage? How will the organisation implement the enterprise information systems? What assumptions underlie the innovation-decision process?

The diffusion of innovation theory, with its constructs listed in Table 1 and the model depicted in Figure 3, provides the insights of diffusing the innovation (*episteme*) via the business case development (*techne*). Table 2 suggests the area for future research. The DOI provides the big picture and a mind map necessary to succeed in undertaking the complex innovation-decision process and developing a good business case. The stakeholders of the radical innovation will have better opportunities to solve the many challenges if they possess the adequate *episteme* of the diffusion paradigm. Introducing a reinvention, adopting, and implementing new enterprise information systems, or extending the current management control systems to encompass linkages with suppliers and customers, are challenges that pose the diffusion problems. The knowledge of DOI certainly offers opportunities in these pursuits toward the development of the business case.

REFERENCES

Allen Sr, G. F. (2005). *The impact of enterprise resource planning on business processes in Allied Aerospace Corporation.* Unpublished PhD Dissertation, Walden University.

Baskerville, R. L., & Pries-Heje, J. (2001). A multiple-theory analysis of a diffusion of information technology case. *Information Systems Journal, 11*(3), 181-212.

Benco, D. C. (2004). *Empirical examination of the effect of enterprise resource planning investments.* Unpublished PhD Dissertation, The University of Texas at Arlington.

Table 2. Future research direction concerning the business case

1.	Understand the innovation-decision process at both the organisational level and the context level (that is, the new enterprise information systems).
2.	Develop and further refine a framework of innovation-decision process (aka, business case development) at both organisational and context levels.
3.	Identify and understand the decision points that permeate the whole innovation-decision process, especially those decision points in the matching stage (the RFI and RFP) and decision stage (the Business Case) in the initiation phase and pre-production stage, production stage, and post-production stage in the implementation phase.
4.	Understand the exact evidence necessary in making the decisions at the various decision points of the innovation-decision process .

Source: Cua & Garrett (2007), Rogers (2003).

Borge, D. (2001). *The book of risk*. New York: John Wiley & Sons, Inc.

Bradford, M., & Florin, J. (2003). Examining the role of innovation diffusion factors on the implementation success of enterprise resource planning systems. *International Journal of Accounting Information Systems, 4*(3), 205-225.

Brealey, R. A., & Myers, S. C. (2000). *Principles of corporate finance* (6th ed.). Boston: The McGraw-Hill Companies, Inc.

Burrell, G., & Morgan, G. (2005). *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life*. Ardershot: Ashgate Publishing Limited.

Carson III, W. A. (2005). *Successful implementation of enterprise resource planning software: A Delphi study.* Unpublished PhD Dissertation, Capella University.

Christensen, C. M. (1997). *The innovator's dilemma*. Boston, MA: Harvard Business School Publishing Corporation.

Christensen, C. M., & Raynor, M. E. (2003). *The innovator's solution: Creating and sustaining successful growth*. Boston, MA: Harvard Business School Press.

Cooper, J.R. (1998). A multidemensional approach to the adoption of innovation. *Management Decision*, *36*(8), 493-502.

Cua, F. C., & Garrett, T. C. (2007). *The decision-points of initiation-decision process of enterprise information systems innovation: An exploratory study of a large public sector asia-pacific organisation.* Paper presented at the Pan-Pacific Conference XXIV: Digital Convergence and e-Globalization, Dunedin-Queenstown, New Zealand.

Cua, F. C., & Theivananthampillai, P. (2006). *Diffusion of innovations in deployment of information technology.* Paper presented at the UK Academy of Information Systems (UKAIS) 2006 11th Annual Conference: Where Theory Meets Practice, The Park, University of Gloucestershire, England.

Dechow, N., & Mouritsen, J. (2004). ERP manuscripts of accounting and information systems. In K. V. Andersen & M. T. Vendelø (Eds.), *The past and future of information systems* (pp. 96-110). Amsterdam: Elsevier Butterworth-Heinemann.

Dechow, N., & Mouritsen, J. (2005). Enterprise resource planning systems, management control and the quest for integration. *Accounting, organizations and society, 30*, 691-733.

Dettmer, H. W. (2003). *Strategic navigation: A systems approach to business strategy*. Milwaukee, WI: ASQ Quality Press.

Dillard, J. F. (2000). Integrating the accountant and the information systems development process. *Accounting Forum, 24*(4), 407 (415p).

Dillard, J. F., Ruchala, L., & Yuthas, K. (2005). Enterprise resource planning systems: A physical manifestation of administrative evil. *International Journal of Accounting Information Systems, 6*, 107-127.

Mabert, V. A., Soni, A. K., & Venkataraman, M. A. (2000). Enterprise resource planning survey of US manufacturing firms. *Production and Inventory Management Journal, 41*(20), 52-58.

Mahler, A., & Rogers, E. M. (1999). The diffusion of interactive communication innovations and the critical mass: The adoption of telecommunications services by German banks. *Telecommunications Policy, 23*, 719-740.

Martin, J. (1995). *The great transition: Using the seven disciplines of enterprise engineering to align people, technology, snd Strategy*. New York: American Management Association.

Morin, R. A., & Jarrell, S. L. (2001). *Driving shareholder value*. New York: McGraw-Hill.

Nadler, D. A., & Tushman, M. L. (1986). *Managing strategic organizational change*. New York: Delta Consulting Group.

Nadler, D. A., & Tushman, M. L. (2004). Implementing new design: Managing organizational change. In M. L. Tushman & P. Andersen (Eds.), *Managing strategic innovation and change: A collection of readings* (2nd ed.). Oxford: Oxford University Press.

The New York Times. (27 Aug 1998). *SAP and Deloitte Sued by FoxMeyer*. Retrieved 17 Feb 2007, from http://query.nytimes.com/gst/fullpage.html?res=9A05E7D7123CF934A1575BC0A96E958260

O'Leary, D. E. (2000). *Enterprise resource planning systems: Systems, life cycle, electronic commerce, and risk*. New York: Cambridge University Press.

Ormerod, P. (2005). *Why most things fail: Evolution, extinction and economics*. London: Faber and Faber Limited.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Simon & Schuster, Inc.

Rüegg-Stürm, J. (2005). *The new St. Gallen management model: Basic categories of an integrated managment*. New York: Palgrave Macmillan.

Slagmulder, R. (1997). Using management control systems to achieve alignment between strategic investment decisions and strategy. *Management Accounting Research, 8*, 103-139.

Tidd, J., Bessant, J., & Pavitt, K. (2001). *Managing innovation: Integrating technological, market and organizational change* (2nd ed.). Chichester: John Wiley & Sons, Ltd.

Trompenaars, F., & Prud'homme, P. (2004). *Managing change across corporate cultures*. Chichester: Capstone.

Van de Ven, A. H., & Poole, M. S. (1990). Methods for studying innovation development in the Minnesota Innovation Research Program. *Organization Science, 1*(3), 313-335.

Zaltman, G., Duncan, R., & Holbek, J. (1973). *Innovations and organizations*. New York: John Wiley and Sons.

KEY TERMS

Business Case: Completed business case document. Business case process.

Business Case Process: Walks through the initiation phase of the innovation-decision process and talks about the project plans that concerned the implementation phase.

Completed Business Case Document: A formal written document that argues a course of action. It contains a point-by-point analysis to making a decision for a set of alternative courses of action to accomplish a specific goal.

Diffusion: Communicating the new idea within a social system, such as an organisation. It culminates in the adoption of the idea, which is the intention of diffusion.

Diffusion of Innovations: Theory concerns the how, why, and at what rate the new idea (commonly referred to as innovation) diffuses.

Implementation Phase: Proceeding after the initiation phase, the implementation phase of enterprise information systems consists of pre-production, production, and post-production (also known as upgrade and maintenance). Refer to innovation-decision process.

Initiation Phase: Consists of awareness stage, matching stage, and lastly, the decision stage. It is the first phase of the innovation-decision process. The second phase is the implementation phase. Refer to innovation-decision process.

Innovation-Decision Process: Starts with an **initiation phase** through which the individuals or decision-making units move from knowing (understanding/identifying) the new idea (the innovation), to forming of an attitude toward the innovation, and subsequently, to deciding whether to adopt or reject the implementation and use of the new idea. The awareness stage is the agenda setting stage. The attitude formation stage is the matching stage. In addition, the decision stage to adopt or reject the innovation terminates the initiation phase. An adoption decision continues the process toward the **implementation phase**, which consists of the pre-production, production, post-production, and confirmation stages.

The Role of E-Services in the Library Virtualization Process

Ada Scupola
Roskilde University, Denmark

INTRODUCTION

The networked ICT technologies (such as the Internet) are having a dramatic effect on how services and especially knowledge services are innovated, designed, produced and distributed. In addition ICT-networks such as the Internet have created the basis for the development of new types of services. E-services are defined here as services that are produced, provided and/or consumed through the use of ICT-networks such as for example Internet-based systems and mobile solutions. E-services can be used by both consumers and businesses, and can be accessed via a wide range of information appliances (Hoffman, 2003, p.53). E-services include also selling of physical goods on the Internet as for example an airline ticket that is purchased online, but delivered by surface mail to the buyers or government services offered on the Internet or e-government. There are three main characteristics of e-services:

- The service is accessible across the Internet or other electronic networks
- The service is consumed by a person across the Internet or other electronic networks
- There might be a fee that the consumer pays the provider for using the e-service, but that might not always be the case as for example in some e-services offered by the government.

Normally the production, provision or consumption of a service requires the interaction between the service provider and the user of the service. Traditionally this has been based on personal interactions, most often face-to-face interactions. In e-services, the production, consumption and/or provision of services takes place through the intermediation of an ICT-network such as Internet-based systems or mobile solutions. Examples of e-services are e-banking, e-library services, e-publishing, airline tickets, e-government, information and location services. The advent of e-commerce and e-services has raised a number of challenges for knowledge intensive service organizations such as consulting companies, libraries and publishers, as well as for companies selling physical goods.

The purpose of this study is to investigate the challenges that e-services are posing and will pose for research or academic libraries. The study has focused on the issues that Roskilde University Library (RUB) has had to deal with as a result of e-services adoption as well as the future challenges that e-services provide for RUB. The study is based on a number of interviews with RUB management, other secondary material provided by Roskilde University library and information provided on the Web page.

BACKGROUND

In order to understand how digitalization and e-services are changing the library and its activities it is important to understand what a library is, and what its major roles in learning are. Libraries can be defined as "an organized set of resources, which includes human services as well as the entire spectrum of media" (e.g. text, video, and hypermedia). Libraries have physical components, such as space, equipment, and storage media; intellectual components such as collection policies that determine what materials will be included and organizational schemes that determine how the collection is accessed; and people, who manage the physical and intellectual components and interact with users to solve information problems" (Marchionini & Maurer, 1995, p. 68). Marchionini and Maurer (1995) distinguish three major roles that academic and research libraries serve in learning. The first role is sharing expensive resources. These resources are physical resources such as books, periodicals, media, and human resources such as the librarians that provide a number of responsive and proactive services. The second role that libraries serve is a cultural role in preserving and organizing artefacts and ideas. Libraries have historically had the role of preserving material to make it accessible to future learners in addition to ensuring access to materials trough indexes, catalogues and other aids that allow users to find what they need. The third role of the library is that of serving as a physical knowledge space, where people meet to study and read and often to exchange ideas.

Roskilde Universtiy Library

Roskilde University Library (RUB) is a research library serving the students and staff at Roskilde University. Roskilde University is a smaller university located in Roskilde, a city about 35 km from Copenhagen, the capital City of Denmark.

The university counts circa 10,000 students. According to Roskilde University Statute (www.ruc.dk/library), Roskilde University Library has the following purposes:

1. To give teachers and students at Roskilde University access to information and materials containing information, that are necessary for research and teaching, as well as ensure information on and access to the university's teachers and students' research.
2. As a public research library to make available its collection to external users, among which regional research and teaching institutions, business, and citizens.
3. Participate to the national and international library collaboration.
4. To conduct research and development within the library subjects and functions, but also the surrounding community and businesses as well as anybody who would like to use the library because it is a public library.

Today the library counts circa 45 employees, and the number of employees has decreased due to the digitalization process and e-services adoption. The library acquires still 8,000-9,000 books in paper format per year. However they expect this number to go down, while the number of e-books to go up especially as the quality of e-books improves. In addition RUB counts today circa 18,000 e-journals, while the number of paper journals has gone down from circa 5,000 to 2,000. Today, in Denmark, libraries are the heaviest users of ICTs among the public sector institutions. The advent of the World Wide Web circa 10 years ago has completely revolutionized the way RUB operates and has made possible a number of e-services and self-services. The adoption and implementation of e-services and self-services has resulted into a number of organizational changes, changes in the organizational structure, the competencies of the librarians and relationships between the library and the publishers and the library and the users. In addition also the business model is changing as RUB is trying to sell the services to private businesses. RUB is moving towards a combination of physical and virtual library, as many services are getting transformed into e-services and self-services. Therefore Internet and e-services might change many aspects of the library and its relationships with users and publishers. However, RUB might preserve its historical role of knowledge space, even though after the implementation of library's online communities, such knowledge space can become also a virtual knowledge space.

E-Services Adoption at RUB

Over the last few years RUB has adopted a number of e-services and self-services that are changing many aspects of the way the library operates. Many of the services provided by RUB have been transformed into e-services after the advent of the World Wide Web. Nowadays RUB offers a number of e-services and Web based self-services. The main e-services offered at RUB are as follows:

1. Access to electronic journals
2. Access to electronic books
3. Digital repository of all the students projects
4. Chat with a librarian

Examples of self-services include:

1. Rucforsk: a self-service system for the online registration of research and other activities of the teachers.
2. Online reference search, online reservation of material not available in the library, etc.

The library is also working on developing a digital repository of the compendia used in the courses. These e-services and self-services are developed on the base of open source software, although the IT department at RUB modifies it to make the software fit to their needs. However they try to use the original open source software as much as possible since it is very expensive to modify it.

MAIN FOCUS OF THE CHAPTER

This session presents the main issues that RUB has encountered in e-services' adoption, the organizational transformations RUB had to go through as a consequence and the challenges that RUB is presently facing and expecting to face in the future.

Back Office

Back office processes have been completely automated as a result of e-service adoption and they have changed from manual to electronic. All library work is today done with the use of ICTs. Even when they get the physical magazine, they insert it into an integrated library system. Everyone working in the library is using ICTs to do their job.

Innovation

Innovation is very important at RUB. The entire e-services and self-services business model is based on one key word: innovation and especially IT-driven innovation. E-services related innovations at RUB are both user-driven and employees-driven. The sources of innovation are very different. A lot of projects are based on ideas coming from people employed at RUB such as librarians, management, the director, and the IT department. Also they provide courses to new enrolled students and faculty about how to use the e-services, and a

lot of ideas come from these teaching sessions. In addition they have a customer-complaint box and library users may send e-mails to the library. These e-mails get screened and RUB may use such suggestions for incremental innovations. DEFF, a major initiative of the Danish government (www.deff.dk), is also an important source of innovation especially regarding the technology aspect of e-services implementation. Through DEFF RUB can get ideas from and share experiences with other libraries. For example each library might be in charge of testing an IT solution, then they share experiences and finally they decide to choose and adopt a system. DEFF is also important in financing new ideas or innovation projects, as RUB might lack the financial resources to start all the projects they believe are worth pursuing.

The main driving forces of e-services adoption have been the government visions and policies for an "IT society for all", the technological development of Internet, World Wide Web and related IT solutions mainly in a technology push fashion, the pressure from cutting costs in the public sector coming either from the government or local university authorities, an IT innovation culture that has always existed in the Danish libraries, (as the director of reader services says "you want to be a little bit better then your neighbour library"), competition among the different libraries' top management and, even though to a less extent, the customer's wishes.

Organizational Change

The digitalization process has changed the structure of RUB's organization in several ways. First of all, a new organizational level, a management level has been introduced that can make the organization look more hierarchical then before, but it cannot really be compared with a classical hierarchical structure. In addition such a management level is mainly dealing with library development and with political issues. Most importantly the division of labor has changed. Especially the number of IT-related jobs has grown; for example 13 years ago RUB had 1 employee dealing with IT, while today they employ 6-7 people in the IT department. The IT department is expected to grow in the future in special fields. In addition almost everybody in the library has to be IT literate and librarians have to grow together with IT, as the trend goes fast. Each employee is participating into several projects, mostly dealing with e-services and e-services development. When Roskilde University started, RUB employed circa 70 people and was servicing circa 1/3 of the number of students and faculties that has today. Today, RUB employees 45 people and serve a number of students and faculties which is 3 times as large as the one that was servicing when the university was founded. The number of RUB's employees might decrease with in the next decade or so. This is because the adoption of e-services and self-services has decreased the need for competences such as the classical librarian, as more and more services that earlier were done by RUB employees are now done by the users of the library. There is a shift from the librarians to the users in the production-consumption of (e-) services. The use of e-services and self-services is increasing; circa 80-85% of the users of the library are using e-services and self-services. Earlier they needed 2-3 librarians at the reference desk, now one is enough. Due to all the organizational changes, this is causing resistance among the employees and users of e-services. As a matter of fact, even though most of RUB users (about 80-85%) are very satisfied with the digitalization trend and the introduction of e-services, there is still a small group that is missing the "old" library and is unsatisfied with e-services.

RUB Business Model

RUB's business model is changing as a result of e-services and self-service adoption and is going into different directions. Within Roskilde University RUB is getting more involved with Campus IT, which is presently developed by the IT department at Roskilde University. However, collaboration is sometimes difficult due to different priorities. RUB believes that they will play a central role in future e-learning projects at Roskilde University. In addition they are trying to collaborate with the teachers and instructors on how to best use the library for teaching and research, including a number of courses on how to use the e-services and self-services that the library offers. Outside Roskilde University, RUB is looking at offering consulting in the field of e-services for other libraries, including business libraries. They are also trying to open their market not only to the campus' students and faculties, but also to companies, especially small and medium enterprises. Participation to the DEFF project can influence the future RUB's business model as well. For example they presently provide an e-service called chat with a librarian, which they are running not only for RUB, but for all the other research libraries in Denmark as well.

Relationships with Customers/Users

Since the introduction of e-services and self-services the relationships with the users of the libraries have changed immensely. The number of users coming to the physical reference desk is decreasing quickly, while the number of inquiries at the virtual desk is increasing; yet the total number of inquiries is decreasing, and in addition the user behavior is changing as well. For example while paper books are still important for the readers, the total number of library loans is decreasing and the number of downloads of e-books is increasing. This trend is also observed for the journals. While RUB still has a substantial number of paper journals, more and more downloads of e-journals articles are taking place. They expect that the loans of physical books and journals will not be important in 5 years and that most of the material

will be provided in electronic form. The users that have a login to the library can access the e-services 24/hours per day, 7 days a week no matter where they are. So they will have everything they need on the computer. Some things are printed, but others are not. The relationships with the users are expecting to change even more in the future as a result of implementation of library blogs. In fact RUB is looking at blogs and how to use them or integrate them with e-services such as electronic journals or e-books. Blogs would have the objective of creating online communities around specific topics, specific books or journal articles. In addition RUB is negotiating with Google to have all its collection retrievable through Google search engines. Therefore e-services are leading to a digitalization of knowledge that was already codified in printed form. E-services are making it easier and quicker for users to find, store and analyze such knowledge. In addition e-services are making it easier for more users to get access to the same piece of knowledge or information. In fact if only one user at a time could get access to a specific journal in print form, in electronic form many users can get access to the same journal, article, or book chapter simultaneously. In addition e-services are pushing customer relationships towards a virtual form. This is the case both regarding the relationship user-librarian and the relationship among the library's users which, after the implementation of blogs, is expected both to become more virtualized and to increase in number due to the formation of online communities.

Relationships with Publishers (or Providers)

This relationship has also changed as a result of e-services. Many of the traditional transactions such as ordering, cataloguing, and so forth of journals has almost disappeared. The total number of transactions with the publishers has decreased. The e-journals are kept at the publishers' repository and RUB only buys the access or license to them. Initially the publishers offered a huge amount of e-journals at a smaller extra cost. As a result RUB cut the number of paper journals from circa 5,000 to circa 2,000 and instead has acquired access to circa 18,000 e-journals. However the publishers are now increasing prices on e-journals, therefore the total costs might increase as a result in the future. This kind of license agreement has contributed to the formation of a Danish library consortium, whose purpose is to get better prices to electronic journals and e-books from the publishers.

Relationships with Other Research Libraries

The trend towards adoption of e-services by the Danish libraries has changed the relationship between RUB and other research libraries in Denmark by increasing collaboration and partnerships among them. While earlier they were competing on services, number and type of journals and books offered, after adoption of e-services there is much more collaboration among Danish research libraries. Two key examples of this collaboration and partnerships which RUB is part of are Denmark Licensing Consortium and the DEFF initiative. Denmark Licensing Consortium is a consortium of libraries getting common licenses to publishers' e-journals and e-books. The major purpose is to put pressure on the publishers and decrease costs for the single library. Therefore the adoption of e-services is causing a convergence and standardization of the (e-) services offered by the different Danish libraries. Libraries were differentiating from each other much more before the adoption of e-services. Now all the research libraries members of the license consortium offer the same types of e-journals and e-books, and more or less the same type of e-services. Those few that are ahead get caught up within a six month period.

DEFF is a major initiative undertaken by the Danish government with the purpose of developing a network of electronic research libraries that make available their electronic and other information resources in a coherent and simple way. This is obtained partly through government funding and partly by joint license purchase (www.deff.dk). By participating to DEFF the libraries can achieve economies of scope and scale in the development of e-services.

FUTURE TRENDS

There are many challenges laying ahead for RUB. RUB will continue to exist and keep the role of library as an information centre, but the way the information and knowledge is provided will change. RUB will still face several organizational and technological challenges in the future.

From a technology point of view, the ICTs platforms used in delivering e-services become obsolete quite periodically and new e-services solutions have to be found. For example with the development of WEB 2.0 they will have to make new types of systems. Integration of RUB e-services into one simple system is also an important technical future challenge. Presently the e-services located on the Web page are connected to 6-7 different systems and a future challenge is to integrate all these different systems. Standardization is another technological challenge. Customers want fast response and RUB is working on this by looking at standardization issues and they have to keep doing so also in the future. Standards are very important for library's e-services. Finally, ensuring to get the best and same results for the same search is also a future technical challenge.

Copyrights and licenses are another important obstacle and challenge for the development of RUB's e-services. For

example they are running a project to convert the library's videos into files to be kept on the local servers. The problem is though that whenever a student wants to see a video, instead of seeing the file on the computer screen, they have to save the file on the tape, since the material that they loan out has to be in analog form due to copyrights restrictions. So copyrights of what can be digitized are a big barrier to further e-services development and especially use by the customers. Licenses on the other hand limit the use of the e-services for remote users that are not connected to the university and therefore do not have a log in to the library system. This implies that these users still have to walk into the library to be able to use the e-services, thus limiting to some extent their functionality.

Another future challenge comes from the library users. The users are becoming much more advanced and sophisticated in their online searches, young people have a lot of ideas about how to do things better. Here the challenge is to understand their needs and implement user-driven innovations in e-services. Budget problems are another challenge for RUB. In the last few years the budgets allocated to research libraries have been decreasing. This trend has been worsened by decentralizing the budgets concerning the research libraries from the government to the university the libraries are connected to. This creates the possibility for management at Roskilde University to cut the library's budget in favour of other activities.

Organizational challenges are also lying ahead. As the number of physical loans will decrease and the number of electronic downloads keeps on increasing, there is going to be less need for the reference desk and the number of positions in the library might decrease. The way of working in the library is changing, therefore the type of competences needed might change moving more towards IT specialists and going away from the classical librarians skills. Disagreement on e-services future development between the different groups in the library is also a major organizational and human resource challenge, even though most RUB's employees like e-services. This requires RUB to explore new functions and new directions to change their business model.

CONCLUSION

This chapter has contributed to the project "E-service—Knowledge services, entrepreneurship, and the consequences for business, customers and citizens" sponsored by the Danish Research Council by investigating a particular type of e-services: research library e-services. Specifically the study has investigated the implication of the advent of Internet and e-services for Roskilde University library (RUB) as well as the future challenges that e-services provide for RUB. The study has also investigated the consequences of e-services for Roskilde University library organization, its business model and relationships with customers, publishers (providers of information) and other research libraries. The picture that emerges is one of fast innovation, big transformations and change at organizational and business model level, as well as in the relationships with customers, publishers and other research libraries. In addition there are a number of challenges that RUB has to face in the future in response to e-services. Some are IT-related; others have to deal with copyrights, licenses, standardization, and user-driven innovation. The general trend is that RUB is becoming a combination of a virtual and physical library, moving more and more towards a virtual library by providing resources and knowledge mainly in digital form and by offering blogs and possibilities of online communities to discuss books and articles. On the other hand RUB is still keeping the traditional library function of a physical knowledge space. How will RUB look like 10 years down the road? The only certain answer according to RUB management is that it will still exist.

REFERENCES

Hoffman, K. D. (2003). Marketing+MIS=E-Service. *Communications of the ACM, 46*(6), 53-55.

Marchionini, G. & Maurer, H. (1995). The roles of digital libraries in teaching and learning. *Communications of the ACM, 38*(4), 67-75.

Sarkar, M. B., Butler, B., & Steinfield, C. (1995). Intermediaries and cybermediaries: A continuing role for mediating players in the electronic marketplace. *Journal of Computer Mediated Communication, 1*(3).

Scupola, A. (2002). The impact of electronic commerce on industry structure—The case of scientific, technical and medical publishing. *Journal of Information Science*, *28*(3).

http://www.bs.dk/publikationer/english/statistics/2004/index.htm

www.deff.dk

http://www.rub.ruc.dk

KEY TERMS

Library: Can be defined as an organized set of resources, which includes human services as well as the entire spectrum of media (e.g., text, video, and hypermedia). Libraries have physical components, such as space, equipment, and storage media; intellectual components such as collection policies that determine what materials will be included and organizational schemes that determine how the collection

is accessed; and people, who manage the physical and intellectual components and interact with users to solve information problems.

E-Services: Are defined as services that are produced, provided, and/or consumed through the use of ICT-networks such as for example Internet-based systems and mobile solutions.

Knowledge Intensive Service Organizations: Are service organizations whose core product is knowledge such as consulting companies.

Danish Research Library: Has the purpose to give teachers and students access to information and materials that are necessary for research and teaching, as well as ensure information on and access to the university's teachers and students' research.

Innovation: Is defined as a new idea, a new product, a new process or an organizational form. It is characterized by three stages: invention, innovation, and diffusion. An invention is a new idea or product, which becomes which becomes an innovation when it starts diffusing in the society or move into a usable form.

IT-Driven Innovation: It is defined as any innovation the creation of which is based on information technology.

Adoption: E-services adoption is here defined as the decision to make use of e-services in the daily operations of the library.

The Role of Human Factors in Web Personalization Environments

Panagiotis Germanakos
National & Kapodistrian University of Athens, Greece

Nikos Tsianos
National & Kapodistrian University of Athens, Greece

Zacharias Lekkas
National & Kapodistrian University of Athens, Greece

Constantinos Mourlas
National & Kapodistrian University of Athens, Greece

George Samaras
National & Kapodistrian University of Athens, Greece

INTRODUCTION

The explosive growth in the size and use of the World Wide Web as a communication medium as well as the new developments in ICT allowed service providers to meet these challenges, developing new ways of interactions through a variety of channels enabling users to become accustomed to new means of service consumption in an "anytime, anywhere and anyhow" manner. However, the nature of most information structures is static and complicated, and users often lose sight of the goal of their inquiry, look for stimulating rather than informative material, or even use the navigational features unwisely. Hence, researchers and practitioners studied adaptivity and personalization to address the comprehension and orientation difficulties presented in such systems, to alleviate such navigational difficulties and satisfy the heterogeneous needs of the users, allowing at the same time Web applications of this nature to survive.

There are many approaches to address these issues of personalization but usually, each one is focused upon a specific area, that is, whether this is profile creation, machine learning and pattern matching, data and Web mining or personalized navigation.

Some noteworthy, mostly commercial, applications in the area of Web personalization that collect information with various techniques and further adapts the services provided, are among others the Broadvision's One-To-One, Microsoft's Firefly Passport, the Macromedia's LikeMinds Preference Server, the Apple's WebObjects, and so forth. Other, more research-oriented systems, include ARCHIMIDES (Bogonikolos et al., 1999), Proteus (Anderson et al., 2001), WBI (Magglio & Barret, 2001), BASAR (Thomas & Fischer, 1997), and mPERSONA (Panayiotou & Samaras, 2004). Significant implementations have also been developed in the area of adaptive hypermedia, with regards to the provision of adapted educational content to students using various adaptive hypermedia techniques. Such systems are, among others, INSPIRE (Papanikolaou, Grigoriadou, Kornilakis, & Magoulas, 2003), ELM-ART (Weber & Specht, 1997), AHA! (De Bra & Calvi, 1998), Interbook (Brusilovsky, Eklund, & Schwartz, 1998), and so forth.

BACKGROUND

Once considering adaptation and personalization categories and technologies we refer to Adaptive Hypermedia and Web Personalization, respectively, due to the fact that they both make use of a user profile to achieve their goal, and consequently they can together offer the most optimized adapted content result to the user.

A Constructive Comparison of Adaptive Hypermedia and Web Personalization

In view of the aforementioned statement, it would be essential to highlight their similarities and differences and furthermore, to identify their convergence point which is their objective to develop techniques to adapt what is presented to the user, based on the specific user needs identified in the extracted user profiles.

Generally, Adaptive Hypermedia refers to the manipulation of the link or content structure of an application to

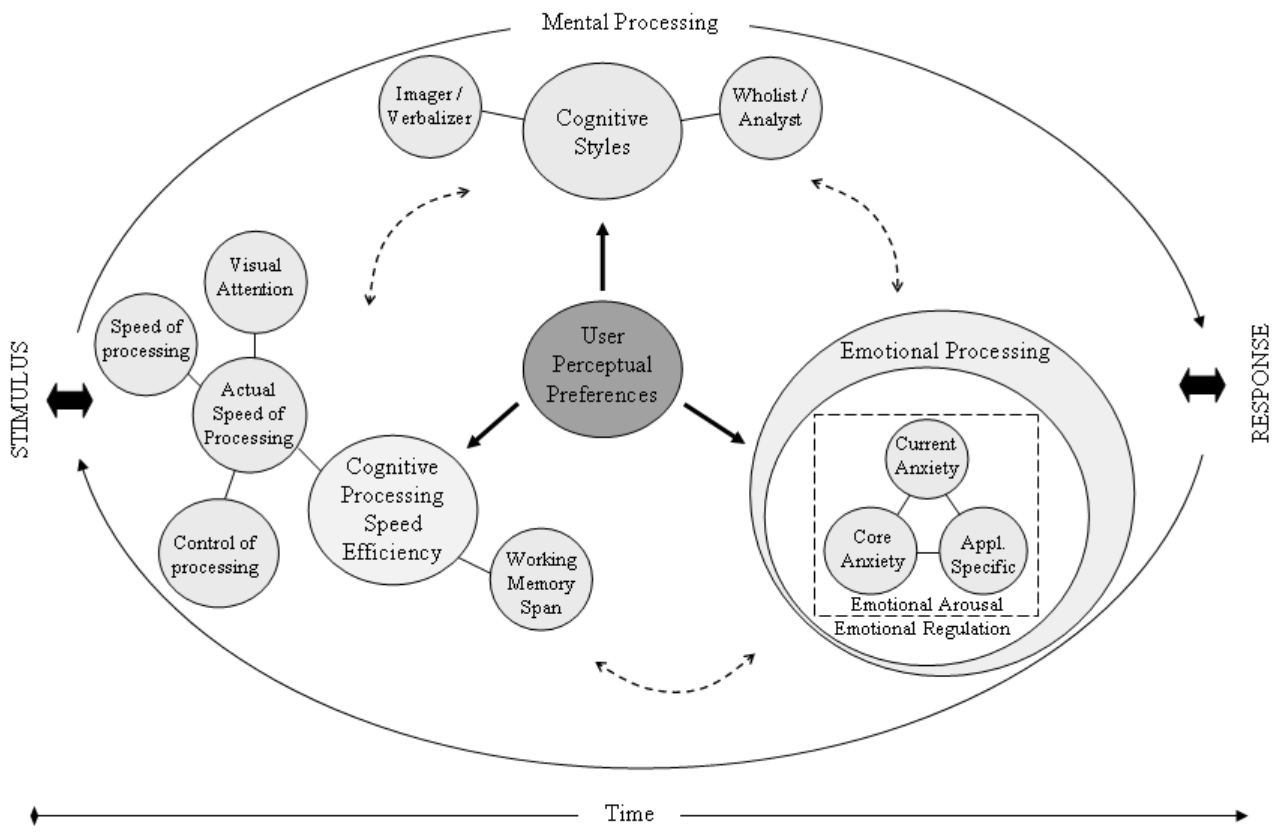
achieve adaptation and makes use of an explicit user model (Brusilovsky, 2001; Eklund & Sinclair, 2000). Adaptive Hypermedia is a relatively old and well established area of research counting three generations (Brusilovsky & Peylo, 2003). Educational hypermedia and online information systems are the most popular, accounting for about two thirds of the research efforts in adaptive hypermedia. Adaptation effects vary from one system to another. These effects are grouped into three major adaptation technologies: adaptive content selection (Brusilovsky & Nejdl, 2004), adaptive presentation (or content-level adaptation) and adaptive navigation support (or link-level adaptation) (Brusilovsky, 2001; Eklund & Sinclair, 2000).

On the other hand, Web personalization refers to the whole process of collecting, classifying and analyzing Web data, and determining based on these the actions that should be performed so that the user is presented with personalized information. Personalization levels have been classified into: Link Personalization, Content Personalization, Context Personalization, and Authorized Personalization (Lankhorst, Kranenburg, & Peddemors, 2002; Rossi, Schwade, & Guimaraes, 2001). The technologies that are employed in order to implement the processing phases mentioned above as well as the Web personalization categories are distinguished into Content-based Filtering, Rule-based Filtering, Collaborative Filtering, Web Usage Mining, Demographic-based Filtering, Agent Technologies, and Cluster Models (Pazzani, 2005; Mobasher, Dai, Luo, Nakagawa, & Wiltshire, 2002).

As inferred from its name, Web personalization refers to Web applications solely, and is a relatively new area of research. One could also argue that the areas of application of these two research areas are different, as Adaptive Hypermedia has found popular use in educational hypermedia and online information systems (Brusilovsky, 2001), whereas Web personalization has found popular use in e-business services delivery. From this, it could be implied that Web personalization has a more extended scope than Adaptive Hypermedia.

The most evident technical similarities of them are that they both make use of a user model to achieve their goal and they have in common two of the adaptation / personalization techniques: the adaptive-navigation support and the adaptive presentation. Last but not least, it is noteworthy to mention that they both make use of techniques from

Figure 1. User Perceptual Preference Characteristics – Three-Dimensional Approach



machine learning, information retrieval and filtering, databases, knowledge representation, data mining, text mining, statistics, and human-computer interaction (Mobasher, Anand, & Kobsa, 2007).

The User Profile Fundamentals

One of the key technical issues in developing personalization applications is the problem of how to construct accurate and comprehensive profiles of individual users and how these can be used to identify a user and describe the user behaviour, especially if they are moving (Panayiotou & Samaras, 2004). User profiling can either be *static*, when it contains information that rarely or never changes (e.g., demographic information), or *dynamic*, when the data change frequently. Such information is obtained either *explicitly*, using online registration forms and questionnaires resulting in static user profiles, or *implicitly*, by recording the navigational behavior or the preferences of each user (Germanakos, Tsianos, Lekkas, Mourlas, & Samaras, 2007a).

THE SIGNIFICANCE OF HUMAN FACTORS IN THE WEB PERSONALIZATION PROCESS

But, do the designers and developers attempt to build user-centric Web-based applications, taking into consideration the real users' preferences in order to provide them a really personalized Web-based content? Many times this is not the case. How can a user profile be considered complete, and the preferences derived optimized, if it does not contain parameters related to the user perceptual preference characteristics (see Figure 1)? *User Perceptual Preference Characteristics* could be defined, as all the critical factors that influence the visual, mental and emotional processes liable of manipulating the newly information received and building upon prior knowledge, that is, different for each user or user group.

These characteristics, which have been primarily discussed in Germanakos, Tsianos, Lekkas, Mourlas, Belk, and Samaras (2007a), and formulate a three-dimensional approach to the problem of building a user model that determines the visual attention, cognitive and emotional processing taking place throughout the whole process of accepting an object of perception (stimulus) until the comprehensive response to it (Germanakos, Tsianos, Mourlas, & Samaras, 2005).

The first dimension investigates users' *cognitive style*, the second their *visual and cognitive processing efficiency*, while the third captures their *emotional processing* during the interaction process with the information space.

Cognitive Style

Cognitive styles represent an individual's typical or habitual mode of problem solving, thinking, perceiving or remembering, and "are considered to be trait-like, relatively stable characteristics of individuals, whereas learning strategies are more state-driven…" (McKay, Fischler, & Dunn, 2003). Among the numerous proposed cognitive style typologies (Cassidy, 2004) has been selected Riding's Cognitive Style Analysis (Riding, 2001), because it is considered that its implications can be mapped on the information space more precisely, because it consists of two distinct scales that respond to different aspects of the Web. The imager/verbalizer axis affects the way information is presented, while the wholist/analyst dimension is relevant to the structure of the information and the navigational path of the user. Moreover, it is a very inclusive theory that is derived from a number of pre-existing theories that were recapitulated into these two axises.

Cognitive Processing Efficiency

The cognitive processing parameters (Demetriou & Kazi, 2001) that have been included in the model are:

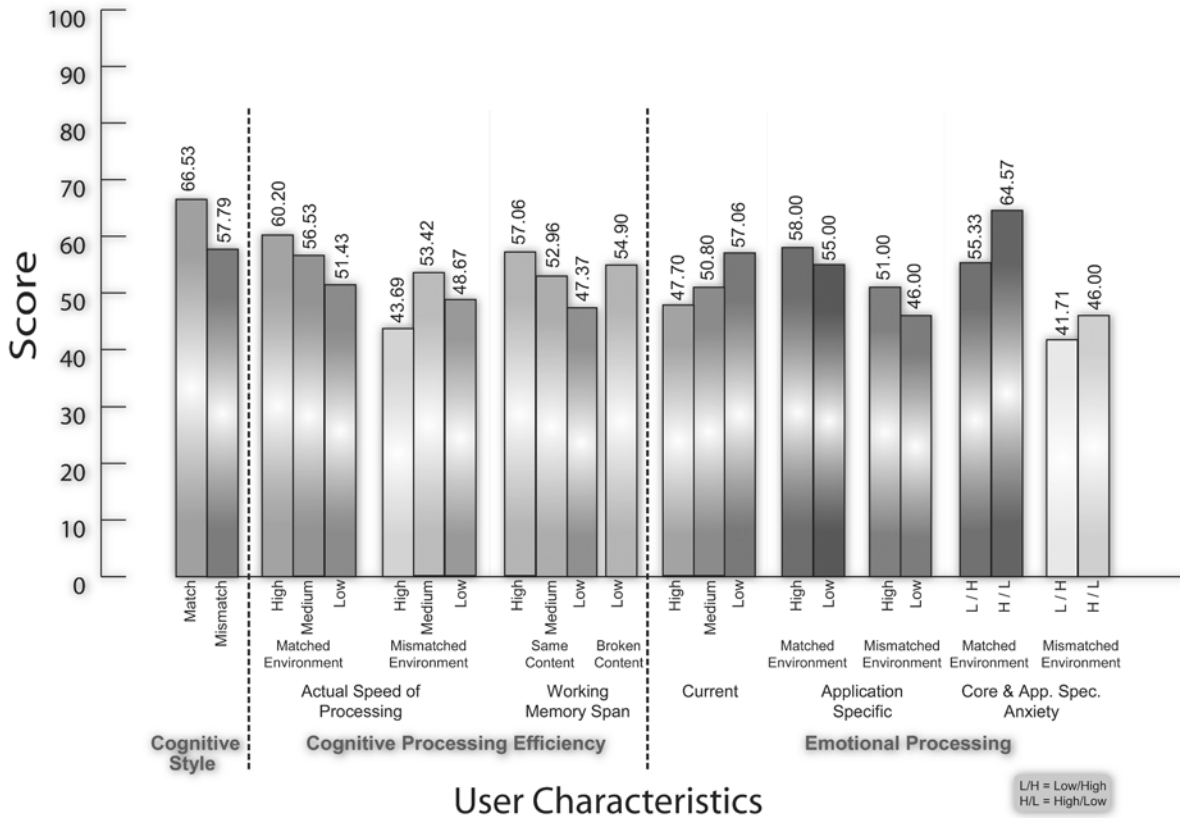
i. *control of processing* (refers to the processes that identify and register goal-relevant information and block out dominant or appealing but actually irrelevant information),
ii. *speed of processing* (refers to the maximum speed at which a given mental act may be efficiently executed),
iii. *working memory span* (refers to the processes that enable a person to hold information in an active state while integrating it with other information until the current problem is solved (Baddeley, 1992)), and
iv. *visual attention* (based on the empirically validated assumption that when a person is performing a cognitive task, while watching a display, the location of his/her gaze corresponds to the symbol currently being processed in working memory and, moreover, that the eye naturally focuses on areas that are most likely to be informative).

Emotional Processing

Emotional processing is a pluralistic construct which is comprised of two mechanisms:

- Emotional Arousal, which is the capacity of a human being to sense and experience specific emotional situations, and

Figure 2. Aggregated differences in matched/mismatch condition



- Emotion Regulation, which is the way that an individual perceives and controls his emotions.

The main focus has been placed on anxiety, as the main indicator of emotional arousal, because it is correlated with academic performance (Cassady & Johnson, 2004), as well as with performance in computer-mediated learning procedures (Smith & Caputi, 2007).

The construct of emotional regulation that has been used includes the concepts of emotional control (self-awareness, emotional management, self-motivation) (Goleman, 1995), self-efficacy (Bandura, 1994), emotional experience and emotional expression (Halberstadt, 2005). By combining the levels of anxiety with the moderating role of emotion regulation, it is possible to examine how affectional responses hamper or promote learning procedures (Lekkas, Tsianos, Germanakos, & Mourlas, 2007).

Evaluation: The Case of AdaptiveWeb System

Subsequently, an adaptive Web-based system has been built, the AdaptiveWeb[1] (Germanakos et al., 2007b), that takes into account users' cognitive and emotional parameters and provides them with information matched to their preferences. All the tests implemented so far to prove components efficiency have been based on a predetermined online content in the field of e-learning multimedia environment, due to the fact mainly that there is an increased interest on distant education via the Web. In this case, it has been feasible to control factors as previous knowledge and experience over distributed information. More specifically, the main research hypotheses drawn has been investigated:

- Are the cognitive and emotional parameters of the model significantly important in the context of an educational hypermedia application, and
- Does matching the presentation and structure of the information to users' perceptual preferences increase academic performance?

Sampling and Procedure

All participants were students from the Universities of Cyprus and Athens; phase I was conducted with a sample of 138 students, while phase II with 82 individuals. Thirty-

five percent of the participants were male and 65% were female, and their age varied from 17 to 22 with a mean age of 19. The environment in which the procedure took place was an e-learning course on algorithms. By controlling the factor of experience, the sample has been divided in two groups: almost half of the participants were provided with information *matched* to their Perceptual Preferences, while the other half were taught in a *mismatched* way. In order to evaluate the effect of matched and mismatched conditions, participants took an online assessment test on the subject they were taught as soon as the e-learning procedure ended, in order to control for long-term memory decay effects. The dependent variable that was used to assess the effect of adaptation to users' preferences was participants' score at the online exam.

Results

As expected, in both experiments the matched condition group outperformed those of the mismatched group (Tsianos, Germanakos, Lekkas, Lourlas, Mourlas, & Samaras, 2007). Figure 2 displays the aggregated differences in performance (the dependent variable of exam score), in matched and mismatched conditions.

Table 1 shows the differences of means (one way ANOVA) and their statistical significance for the parameters of Cognitive Style, Cognitive Efficiency Speed, and Emotional Processing.

Moreover, in many cases there is a high correlation between the dimensions of the various factors of the model, validating the psychometric tools that have been used. This fact also demonstrates the effectiveness of incorporating a variety of human factors in Web-based personalized environments.

Finally, emotional processing, and more specifically anxiety, turned out to be an equally important factor; medium levels of anxiety are supportive of increased performance, while aesthetics and extra navigation support helped significantly students that were highly (not extremely though) anxious, always in terms of performance.

FUTURE TRENDS

Future and emerging trends include the further investigation of constraints and challenges arising from the implementation of such issues on mobile devices and channels; study on the structure of the metadata coming from the providers' side, aiming to construct a Web-based personalization architecture that will serve as an automatic filter adapting the received content based on a comprehensive user profile; the incorporation of physiological measurements of emotions and anxiety in such a model, with the use of biometrical sensors; as well as the use of an eye-tracker device to clarify the role of visual attention in Web-based communication environments.

CONCLUSION

Adaptive hypermedia and Web personalization are two distinct, well-established areas of research, both investigating methods and techniques to move conventional static systems beyond traditional borders to more intelligent, adaptive and personalized implementations. They share a common goal: to alleviate navigational difficulties and satisfy the heterogeneous needs of the user population by adapting according to user specific characteristics. In order to do that, the user profile construction is considered necessary.

The basic objective of this article was to make an extensive reference of a combination of concepts and techniques coming from different research areas, adaptive hypermedia and Web personalization, all of which focus on the user. It has been attempted to approach the theoretical considerations and technological parameters that can provide the most comprehensive user profile, under a common filtering element

Table 1. Differences of means in the matched/mismatched condition for Cognitive Style and Cognitive Efficiency Speed

	Match Score	Match n	Mismatch Score	Mismatch n	F	Sig.
Cognitive Style	66.53% 5	3	57.79% 6	1	6.330	0.013
Cognitive Efficiency Speed	57.00% 4	1	48.93% 4	1	5.345	0.023
Emotional Processing	57.91% 2	3	48.45% 2	9	4.357	0.042

(User Perceptual Preference Characteristics), supporting the provision of the most apt and optimized user-centred Web-based result.

REFERENCES

Anderson, C., et.al. (2001). Personilizing Web sites for mobile users. In *Proceedings of the 10th Conference on the World Wide Web, 2001*.

Baddeley, A. (1992). Working memory. *Science, 255*, 556-559.

Bandura, A. (1994). Self-efficacy. In V. S. Ramachaudran (Ed.), *Encyclopedia of human behaviour* (Vol. 4, pp. 71-81). New York: Academic Press.

Bogonikolos, N., et al. (1999). ARCHIMIDES: An intelligent agent for adaptive-personalized navigation within a WEB server. In *Proceedings of the 32nd Annual Hawaii International Conference on System Science, HICSS-32,* (Vol. 5).

Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction, 11*, 87-110.

Brusilovsky, P., Eklund, J., & Schwarz, E. (1998, April 14-18). Web-based education for all: A tool for developing adaptive courseware. Computer networks and ISDN systems. In *Proceedings of the 7th International WWW Conference,* (Vol. 30, No.1-7, pp. 291-300).

Brusilovsky, P., & Nejdl, W. (2004). *Adaptive hypermedia and adaptive Web*. CSC Press.

Cassady, J.C. (2004). The influence of cognitive test anxiety across the learning-testing cycle. *Learning and Instruction, 14*(6), 569-592.

Cassidy, S. (2004). Learning styles: An overview of theories, models, and measures. *Educational Psychology, 24*(4), 419-444.

De Bra, P., & Calvi, L. (1998). AHA! An open adaptive hypermedia architecture. *The new review of hypermedia and multimedia* (Vol. 4, pp. 115-139). Taylor Graham.

Demetriou, A., & Kazi, S. (2001). *Unity and modularity in the mind and the self: Studies on the relationships between self-awareness, personality, and intellectual development from childhood to adolescence*. London: Routledge.

Eklund, J., & Sinclair, K. (2000). An empirical appraisal of the effectiveness of adaptive interfaces of instructional systems. *Educational Technology and Society, 3(*4), ISSN 1436-4522.

Germanakos, P., Tsianos, N., Lekkas, Z., Mourlas, C., Belk, M., & Samaras, G. (2007b, June 25-29). An adaptive Web system for integrating human factors in personalization of Web content. In *Proceedings of the 11th International Conference on User Modeling (UM 2007),* Corfu, Greece.

Germanakos, P., Tsianos, N., Lekkas, Z., Mourlas, C., & Samaras, G. (2007a). Capturing essential intrinsic user behaviour values for the design of comprehensive Web-based personalized environments. *Computers in Human Behavior Journal, Special Issue on Integration of Human Factors in Networked Computing*.

Germanakos, P., Tsianos, N., Mourlas, C., & Samaras, G. (2005, December 14-16). New fundamental profiling characteristics for designing adaptive Web-based educational systems. In *Proceedings of the IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA2005),* Porto, (pp. 10-17).

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York: Bantam Books.

Halberstadt, A.G. (2005). Emotional experience and expression: An issue overview. *Journal of Nonverbal Behavior, 17*(3), 139-143.

Lankhorst, M.M., Kranenburg, S.A., & Peddemors, A.J.H. (2002). Enabling technology for personalizing mobile services. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS-35'02).*

Lekkas, Z., Tsianos, N., Germanakos, P., & Mourlas, C. (2007, May 23-27). Integrating cognitive and emotional parameters into designing adaptive hypermedia environments. In *Proceedings of the Second European Cognitive Science Conference (EuroCogSci'07),* Delphi, Hellas.

Maglio, P., & Barret, R. (2000). Intermediaries personalize information streams. *Communications of the ACM, 43*(8), 96-101.

McKay, M.T., Fischler, I., & Dunn, B.R. (2003). Cognitive style and recall of text: An EEG analysis. *Learning and Individual Differences, 14*, 1-21.

Mobasher, B., Anand, S.S., & Kobsa, A. (2007). Intelligent techniques for Web personalization. In *Proceedings of the 5th Workshop ITWP 2007, held in conjunction with the 22nd National Conference in Artificial Intelligence (AAAI2007).*

Mobasher, B., Dai, H., Luo, T., Nakagawa, M., & Wiltshire, J. (2002). Discovery of aggregate usage profiles for Web personalization. *Data Mining and Knowledge Discovery, 6*(1), 61-82.

Panayiotou, C., & Samaras, G. (2004). mPersona: Personalized portals for the wireless user: An agent approach.

Journal of ACM/Baltzer Mobile Networking and Applications (MONET), Special Issue on Mobile and Pervasive Commerce, (6), 663-677.

Papanikolaou, K.A., Grigoriadou, M., Kornilakis, H., & Magoulas, G.D. (2003). Personalizing the interaction in a Web-based educational hypermedia system: The case of INSPIRE. *User-Modeling and User-Adapted Interaction, 13*(3), 213-267.

Pazzani J.M. (2005). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review, 13*(5-6), 393-408.

Riding, R.J. (2001). Cognitive style analysis—research administration. *Learning and Training Technology*. New Zealand.

Rossi, G., Schwade, D., & Guimaraes, M.R. (2001). *Designing personalized Web applications*. ACM 1-58113-348-0/01/0005.

Smith, B., & Caputi, P. (2007). Cognitive interference model of computer anxiety: Implications for computer-based assessment. *Computers in Human Behavior, 23*(3), 1481-1498.

Thomas, C., & Fischer, G. (1997). Using agents to personalize the Web. In *Proceedings of the ACM IUI'97*, Florida, (pp. 53-60).

Tsianos, N., Germanakos, P., Lekkas, Z., Mourlas, C., & Samaras, G. (2007, December 7-9). Evaluating the significance of cognitive and emotional parameters in e-learning adaptive environments. In *Proceedings of the IADIS International Conference on Cognition and Exploratory Learning in Digital Age (CELDA2007),* Algarve, Portugal.

Weber, G., & Specht, M. (1997). User modeling and adaptive navigation support in WWW-based tutoring systems. In *Proceedings of User Modeling '97*, (pp. 289-300).

KEY TERMS

Cognition: A human-like processing of information, applying knowledge and changing preferences. Cognition or cognitive processes can be natural and artificial, conscious and not conscious; therefore, they are analyzed from different perspectives and in different contexts, in anesthesia, neurology, psychology, philosophy, systemics and computer science.

Cognitive Styles: They are consistent individual differences in preferred ways of organizing and processing information and experience

Emotional Intelligence: It describes an ability, capacity, or skill to perceive, assess, and manage the emotions of one's self, of others, and of groups.

User Modeling: User modeling is a subarea of human-computer interaction, in which the researcher/designer develops cognitive models of human users, including modeling of their skills and declarative knowledge. User models can predict human error and learning time.

User Perceptual Preference Characteristics: User Perceptual Preference Characteristics are all the critical factors that influence the visual, mental and emotional processes liable of manipulating the newly information received and building upon prior knowledge, that is, different for each user or user group. These characteristics determine the visual attention, cognitive and emotional processing taking place throughout the whole process of accepting an object of perception (stimulus) until the comprehensive response to it.

Visual Processing: It is the sequence of steps that information takes as it flows from visual sensors to cognitive processing.

Web Personalization: It is the process of tailoring pages to individual users' characteristics or preferences. It is a means of meeting the user's needs more effectively and efficiently, making interactions faster and easier and, consequently, increasing user satisfaction and the likelihood of repeat visits.

The Role of Information in the Choice of IT as a Career

Elizabeth G. Creamer
Virginia Tech, USA

INTRODUCTION

Practitioners, researchers, and policy makers alike are puzzled by the continued intransigence to the integration of women to undergraduate and graduate majors, as well as occupation, in fields like engineering and information technology (IT). While strong advances in the direction of gender equity have been made in the last two decades in fields like biology and mathematics and in the professional fields of medicine and law, women only still represent about 20% of the undergraduate enrollments in engineering and computer science (NSF, 2000). This gender gap persists persist despite the near evaporation of evidence of gender differences in performance in these fields, such as in the dramatic narrowing of gender differences in the high school course taking patterns, including in advance placement courses (Clewell & Campbell, 2002). Gender differences in the enrollment in computer-related courses and out-of-class, informal programs in science and engineering persist, however (Volman & van Eck, 2001).

Academics have used several major groups of theories to try to understand the reasons for women's under-representation in IT and engineering. Social psychological theories are one of four major groupings of theoretical frameworks identified by Clewell and Campbell (2002). As compared to perspectives that seek biological or cognitive explanations for women's disinclination to pursue careers in some fields, social-psychological theorists consider environmental, social, and attitudinal influences. Factors such as teachers' and advisors' attitudes and beliefs, pedagogical practices in the way math and sciences courses are taught, and the influence of parents and the media are some of the factors considered by social-psychological theorists (Clewell & Campbell, 2002).

The research described in this entry belongs to the group of social-psychological theories that look to environmental, rather than individual, explanation for women's under-representation in certain fields in science and engineering, including information technology. It considers the role of parents and the role of interactions with teachers, counselors, and important others in interest in a career in information technology.

BACKGROUND

Career choice is often approached as if it were entirely a rational process whose outcome can be predicted by simply understanding an individual's abilities, attitudes, and interests. Researchers with a less individualistic perspective, however, point out that individual qualities are far less predictive of women's career choices than they are of men's (O'Brien & Fassinger, 1993) and that a number of social and cultural factors are required to understand the types of environments that promote women's interest in sex atypical careers, including IT (Blum, Frieze, Hazzon, & Dias, 2007). This fits with other research that documents that women continue to enter the IT field through nontraditional venues. Rather than taking a more direct route through an IT-related major in college, women often enter positions in IT as a result of propitious exposure or opportunity (Turner, Brent, & Percora, 2002). This is why some authors prefer to refer to "pathways" rather than to continue to utilize the "pipeline metaphor" as a way to capture the various career paths women follow prior to entering IT in a professional capacity (Leventmen, 2007).

A socio-cultural perspective considers career interest and choice to be the product of a complex interplay of factors that is primarily understood empirically through relatively sophisticated statistical procedures. This theoretical perspective frames research about gender and interest and success in fields in science, engineering, and technology (SET) to be the result of the interplay of personal and environmental factors. Key individual qualities related to interest in IT include positive attitudes about computers and computing fields (Dryburgh, 2000; Shashaani, 1997), and characteristics of parents, including their familiarity with IT and their views about the appropriateness of IT as a career choice (Meszaros, Laughlin, Creamer, Burger, & Lee, 2006). Attachment to parents is positively associated with career exploration among college women (Ketterson & Blustein, 1997).

Social qualities related to an interest in IT refer to dimensions of the environment, such as encouragement from educators, family and friends, role models, and peers. It also includes elements of the family, community, and educational environment that introduce opportunities to experience

creative and interactive applications of computing (Volman & Van Eck, 2001).

The stereotyping of many SET fields as inherently unfeminine is a key cultural dimensions that impacts women's' perceptions of the opportunity for success and advancement in the field (Blum et al., 2007).

CAREER INFORMATION AND INTEREST IN IT AS A CAREER

This entry summarizes key research findings about the relationship between sources of career information and information seeking behavior and interest in IT among high school and college men and women. Findings reported here are based on responses to multiple administrations of questionnaire *Career-Decision Making Survey* (Creamer, Lee, Meszaros, Laughlin, & Burger, 2006) and the response of 1,147 high school and college men and women in rural and urban locations in the US. The focus in this entry is on the "information processing" section of a larger, complex causal model that was confirmed through statistical analysis using path analysis. The term "information processing" refers to the impact of others on interest in IT. Details about the complete model are reported elsewhere (Creamer, Meszaros, & Lee, 2007; Creamer, Lee, Meszaros, Burger, & Laughlin, 2006).

Our findings indicate that for both young women and men, there is a startling gap between knowledge about job options in IT, the extent of career exploration, and interest in IT. The nature of the relationship is quite different, however, for men and women. Figure 1 summarizes key findings about career information processing related to interest in IT as a career field and how this varies by gender.

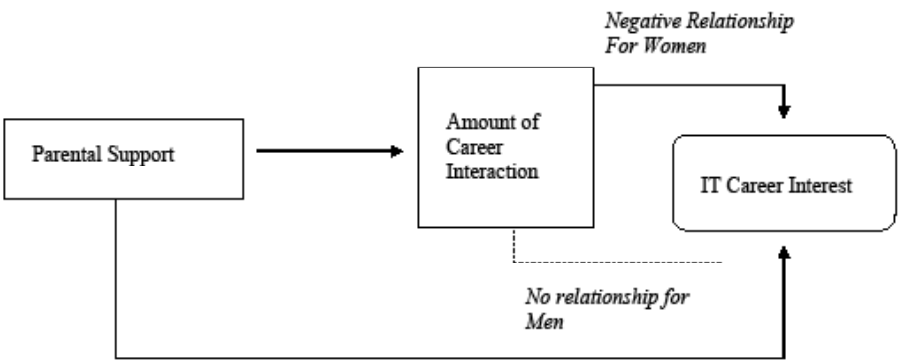
As shown in Figure 1, there is no significant or direct relationship for men between amount of career exploration, sources of career information, and interest in IT as a career option. Interactions with others about career options had no significant positive or negative effect on interest in IT. For men, computer use and positive attitudes about the attributes of IT workers were the only factors directly related to IT career interests.

The relationship between information and interest in IT as a career choice was even more surprising for women. Women were more significantly more likely than men to report talking to others about career options. For women, however, there is a significant, *negative* relationship between career information about IT and interest in IT as a career. The more interactions women had with others about IT as a career field, the less interest they reported in IT as a viable career choice. This is consistent with other research that suggests women are more influenced than men by the opinions of others when making career choices (Seymour & Hewitt, 1997). This is probably even more so when the choice involves a sex-atypical career, like IT. While it is not possible to determine from our data exactly what is occurring during these interactions, what is clear from our analysis is that many women are walking away from interactions with teachers, peers, counselors and others with less than positive views about IT as a personally viable career option.

The amount of interaction with others about career options increased for women with uncertainty about career choice and/or the ability to make a good career choice. Another possible explanation for why women may walk away from exchanges about career options with negative views about IT is that the more people they talk to, the more likely they are to encounter different viewpoints about appropriate career alternatives. Making a reasoned decision in the face of differing viewpoints is not something most high school or college students are developmentally equipped to handle.

For both men and women, parents play a significant role in the development of IT as a career interest. Perceptions of parents support for IT as an appropriate career choice directly impacted the amount of career exploration students

Figure 1. The impact of career information on IT career interests among high school and first- and second-year college students, by gender



reported, as well as whether they had positive views about the attributes of IT workers and the amount of computer use. The role of parents' expectations and attitudes, particularly mothers, had a stronger direct effect on women's than men's interest in IT.

During one-on-one interviews, female participants repeatedly highlighted the role of trusted and immediate others in the development of their career interests. These findings underscore the importance of involving parents activities designed to promote students' interest in IT as a career options.

FUTURE TRENDS

Some key findings from our research, such as about the role of trust in women's willingness to consider unfamiliar career options, only arose through the concurrent analysis of qualitative and quantitative data from the same participants. This points to one of the principal advantages of mixed methods research design and the value of intentionally engaging differences that appear from the analysis of qualitative and quantitative research (Creswell, Clark, Gutmann, & Hanson, 2003; Greene & Caracelli, 2003). Future researchers considering issues related to women and IT would gain immeasurably by pursuing research questions through both qualitative and quantitative data, particularly when opportunities are created to engage participants in discussions of differences in data they supply.

CONCLUSION

Findings from this research indicate that it is far too simplistic to assume that the key reason for the under-enrollment of women in IT majors is the lack of access to up-date information about the field. Although middle-class young women in urban and suburban settings have a clear advantage over their lower-income counterparts and those in rural settings because of the visibility of role models in the IT occupations, the Internet offers ready access to career information for those with the tools and willingness to invest the time to explore them. What is needed in the K-12 setting is a commitment of resources to provide up-to-date training to counselors about the day-to-day activities of people holding different types of professional IT jobs.

Engagement in out-of-class activities, often called informal education, has been found to be key to developing and sustaining women's interest in science, engineering, and technology fields (Campbell, Acerbo, & Hoey, 1999). Industry can contribute to the recruiting of women to the IT field by providing resources and expertise to the creation of informal activities outside of the classroom, such as intensive summer programs, that create the opportunity for students to engage in creative, interactive, and socially meaningful computer applications. These are the types of activities that can build on student's fascination with computers, create a sense of playfulness and experimentation, and help them to begin to see the link between enjoyment in using computers and IT as a viable career option. The development of a sense of community that is possible in intensive residential programs may promote the trust instrumental to widening the range of career options women consider beyond those modeled and/or supported by important people in their immediate environment.

While there is a good deal of overlap in what men and women find attractive about a career in IT, there are some important differences that should be considered in developing effective recruiting materials. Some of these are addressed in a professionally produced DVD developed by the WIT team, *The Power of Partners* (Meszaros, Laughlin, Burger, Creamer, & Lee, 2007). Highlighting IT as a demanding field with the potential for a competitive income appears, for example, to be persuasive to recruiting men, but not women to IT. The potential for flexibility in work schedules, and thus the potential to achieve some work-life balance, is an issue of growing importance in the career choices of women (Rosser, 2004) and one that human resource personnel in the IT field could leverage effectively in recruiting material.

ACKNOWLEDGMENT

This research was supported by funds from the National Science Foundation under grants HRD-0120458 and HRD-0522767.

REFERENCES

Blum, L., Frieze, C., Hazzan, O., & Dias, M. B. (2007). A cultural perspective on gender diversity and computing. In C. J. Burger, E. G. Creamer, & P. S. Meszaros (Eds.), *Reconfiguring the firewall: Recruiting women to IT across continents and cultures* (pp. 109-134). Wellesley, MA: AK Peters Publishing.

Campbell, P., Acerbo, K., & Hoey, L. (1999). *Educational equity concepts: Playtime is Science Plus. Final Evaluation Report*. Groton, MA: Campbell-Kibler Associates, Inc.

Clewell, B. C., & Campbell, P. B. (2002). Taking stock: Where we've been, where we are, where we're going. *Journal of Women and Minorities in Science and Engineering, 8*, 255-284.

Creamer, E. G., Lee, S., Meszaros, P., Burger, C. J., & Laughlin, A. (2006). Predicting women's interest and choice of

an IT career. In E. M. Trauth (Ed.), *Encyclopedia of gender and information technology* (pp. 1023-1028). Hershey, PA: Idea Group Reference.

Creamer, E. G., Lee, S., Meszaros, P. S., Laughlin, A., & Burger, C. J. (2006). Career decision making survey. Retrieved from http://www.wit.clahs.vt.edu

Creamer, E. G., Meszaros, P. S., & Lee, S. (2007). Predicting women's interest and choice of a career in information technology: A statistical model. In C. J. Burger, E. G. Creamer, & P. S. Meszaros (Eds.), *Reconfiguring the firewall: Recruiting women to IT across continents and cultures* (pp. 15-40). Wellesley, MA: AK Peters Publishing.

Creswell, J. W., Clark, V. P., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori, & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209-240). Thousand Oaks, CA: SAGE.

Dryburgh, H. (2000). Under-representation of girls and women in computer science: Classification of 1990's research. *Journal of Educational Computing Research, 23*(2), 181-202.

Greene, J. C., & Caracelli, V. J. (2003). Making paradigmatic sense of mixed methods practice. In A. Tashakkori, & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 91-110). Thousand Oaks, CA: SAGE.

Ketterson, T. U., & Blustein, D. L. (1997). Attachment relationships and the career exploration process. *Career Development Quarterly, 46*(2), 167-178.

Leventman, P. G. (2007). Multiple pathways toward gender equity in the U.S. information technology workforce. In C. J. Burger, E. G. Creamer, & P. S. Meszaros (Eds.), *Reconfiguring the firewall: Recruiting women to IT across continents and cultures*. Wellesley, MA: AK Peters Publishing.

Meszaros, P. S., Laughlin, A., Burger, C. J., Creamer, E. G., & Lee, S. (2007). *The power of partners: Helping females their way to high tech careers* [DVD]. Peggy Meszaros, Executive Producer. Retrieved from www.witvideo.org.vt.edu

Meszaros, P., Laughlin, A., Creamer, E. G., Burger, C. J., & Lee, S. (2006). Parental support for female IT career interest and choice. In E. M. Trauth (Ed.), *Encyclopedia of gender and information technology* (pp. 963-969). Hershey, PA: Idea Group Reference.

National Science Foundation. (2000). *Women, minorities, and persons with disabilities in science and engineering.* Arlington, VA.

O'Brien, K. M., & Fassinger, R. E. (1993). A causal model of the career orientation and career choice of adolescent women. *Journal of Counseling Psychology, 40*, 456-469.

Rosser, S. V. (2004). *The science glass ceiling*. NY: Routledge Press.

Seymour, E., & Hewitt, N. M. (1997). *Talking about leaving: Why undergraduates leave the sciences*. Oxford, UK: Westview Press.

Shashaani, L. (1997). Gender differences in computer attitudes and use among college students. *Journal of Educational Computing Research, 16*(1), 37-51.

Turner, S. V., Brent, P. W., & Pecora, N. (2002, April). *Why women choose information technology careers: Educational, social and familial influences.* Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.

Volman, M., & Van Eck, E. (2001). Gender equity and information technology in education: The second decade. *Review of Educational Research, 71,* 613-634.

KEY TERMS

Direct Effect: In a statistical model, a variable that impacts the dependent variable, or the variable being predicted in a direct and statistically significant way.

Information Processing: Relates to student's self-reports about (1) how much they know about job options within IT, (2) what information sources about IT they consider credible, and (3) how often they have interacted with others about career options.

Information Technology: Refers to a variety of jobs that involve the development, installation, and implementation of computer systems and applications. Careers in IT encompass occupations that require designing and developing software and hardware systems, providing technical support for computer and peripheral systems, and creating and managing network systems and databases.

IT Career Interest and Choice: The dependent variable in our statistical model that identifies the characteristics of respondents who express either an interest in a career in IT or who have already made a choice to pursue a career in IT.

Mixed Methods Research: Involves the analysis of both quantitative and qualitative data in a single study and the integration of findings at one or more stages in the process of research (Creswell et al., 2003).

Parental Support: In the statistical model presented in this entry, parental support was calculated from nine questionnaire items relating to the respondent's perceptions that her parents support the importance of a career and encourage career exploration, as well as agreement with the statement that parents have an idea about what would be an appropriate career choice.

Social-Psychological Theorists: Theories that consider environmental and social influences on career interests, including such factors as parents' attitudes, teachers' and advisors' attitudes and beliefs, and experiences and interactions in- and out-of-the-classroom.

Sources of Career Information: A set of questionnaire items in the *Career Decision Making Survey* where respondents indicated how often they had discussed career options and plans with ten groups of people: mother, father, teacher or professor, counselor or advisor, other family members, male friends, female friends, spouse or significant other, employer or boss, and family friends.

Satellite Network Security

Marlyn Kemper Littman
Nova Southeastern University, USA

INTRODUCTION

Satellite networks play a vital role in enabling essential critical infrastructure services that include public safety; environmental monitoring; maritime disaster recovery and reconnaissance; electronic surveillance; and intelligence operations for law enforcement, the military, and government agencies (Jamalipour & Tung, 2001). As demonstrated by the events following the terrorist attacks in the U.S. on the Pentagon in Washington, D.C. and the World Trade Center in New York City on September 11, 2001, satellite networks also provide redundant communications services when terrestrial networks are disrupted and/or unavailable. Despite their merits, satellite networks are nonetheless vulnerable to cyber attacks that pose threats to national security and the economy.

Satellite networks transport voice, video, images, and data through the air as electromagnetic signals, thereby making these transmissions susceptible to interception. Technical advances enable the interconnectivity of satellite systems to public and private wireless and terrestrial networks including the Internet. These advances, however, amplify the risk of cyber attacks that can compromise critical infrastructure functions dependent on satellite networks in sectors that include information technology (IT) and telecommunications; defense; government; banking and finance; utilities; agriculture; emergency services; public health; and transportation (U.S. Department of Homeland Security (DHS), 2003; U.S. Government Accounting Office (GAO), 2004). As a consequence, satellite networks employ an array of security tools and mechanisms for countering costly and widespread cyber incursions and, thereby, ensuring the continuity of critical infrastructure operations. Those cyber attacks that are politically motivated and specifically designed to disrupt essential services are generally attributed to *cyber terrorism*.

This chapter describes the technical fundamentals of satellite networks; examines security vulnerabilities; and explores initiatives for protecting the integrity of satellite network transmissions and operations from cyber incursions and physical attacks. Standards and protocols that safeguard satellite networks from unauthorized use and intentional disruptions and policies, and legislation that facilitate cyberspace asset protection are described. Capabilities of *encryption* in supporting secure satellite services and the distinctive attributes of the InterPlanetary Internet (IPN), also called the InterPlanetary Network**,** are explored.

BACKGROUND

Satellite Network Technical Fundamentals

Satellite networks consist of ground and space segments. The ground segment includes a ground or earth station that delivers communications services and monitors satellite operations by providing tracking, telemetry, and control (TT&C) functions. The space segment consists of the artificial satellite and its payload.

In contrast to a natural satellite or a celestial body that revolves around a larger sized planet, an artificial satellite is a wireless receiver/transmitter that orbits the earth and employs microwave technology in the super high and extremely high radio frequency (RF) bands of the electromagnetic spectrum to enable wide area interactive communications (Littman, 2002). The payload includes transceivers and antennas for RF signal reception, amplification, and retransmission.

The quality of the satellite signal reflects the quality of the uplink and downlink. An uplink describes signal transmissions from an earth station such as a gateway, teleport, hub, or very small aperture terminal (VSAT) to the satellite. A downlink refers to signal transmissions from the satellite to the designated reception site. Typically, satellite transmissions are asymmetrical with more information transported on the downlink than on the uplink (Littman, 2002). Generally classified in terms of the orbits in which they operate, satellite constellations are categorized as geosynchronous or geostationary earth orbit (GEO), medium earth orbit (MEO), and low earth orbit (LEO).

Satellite Network Vulnerabilities

Satellites' transmissions are subject to lengthy delays, low bandwidth, and high bit-error rates that adversely impact real-time, interactive applications such as videoconferences and lead to data corruption, performance degradation, and cyber incursions. Atmospheric and interstellar noise; cosmic radiation; interference from electronic devices; and precipitation and rain absorption in the spectral frequencies employed

by satellites impede network performance and information throughput and negatively affect provision of quality of service (QoS) guarantees (Littman, 2002).

Satellite network applications and services are also adversely impacted by geophysical events. In 1998, for example, tremendous explosions on the sun disrupted operations onboard PanAmSat's Galaxy IV Satellite. As a consequence of these solar flares, digital paging services, bank transactions, and cable television programs across the U.S. were disabled (U.S. GAO, 2002).

According to the U.S. GAO (2002), satellite network functions can be compromised by ground-based antisatellite weapons, high-altitude nuclear explosions, stealth micro satellites, space mines, space-to-space missiles, and directed energy space weapons. For instance, as a consequence of intentional jamming resulting from cyber attacks on a Telestar-12 commercial satellite in 2003, U.S. government-supported broadcasts promoting regime changes in Iran were blocked by the Iranian Ministry of Post, Telegraph, and Telephone (Waldrop, 2005). Satellite-based telephony services in Tehran were also disabled.

Satellite network operations are subject to denial of service (DoS) and distributed DoS (DDoS) attacks generated by automated tools that prevent authenticated users from accessing network services; the spread of viruses to mobile satellite-enabled appliances such as cellular phones; worms that self-propagate malicious data; and spy ware that enables intruders to gain unrestricted access to classified documents (U.S. GAO, 2005) as well. denial of information (DoI) attacks on satellite networks such as spam or unsolicited commercial e-mail and phishing or transmission of fraudulent e-messages are typically designed to deceive legitimate users into revealing confidential information to unauthorized sources (Conti & Ahamad, 2005; Wilson, 2005).

Satellite networks are also vulnerable to cyber terrorism or coordinated space-based and ground-based threats and attacks committed by unlawful and/or politically motivated terrorist groups who target critical communications systems such as satellite networks to cause data corruption, disruption of critical infrastructure services, economic damage, harm, and loss of life (Wilson, 2005). Satellite network attacks attributed to cyber terrorism can result in disruptions in financial markets and disclosure of government, law enforcement, medical, and/or military classified data (U.S. GAO, 2004). Intentional satellite system incursions motivated by cyber terrorism raise questions about the dependability, reliability, availability, and security of satellite network services and erode public confidence in the integrity of satellite-dependent, critical infrastructure applications (Bosch, 2002).

SATELLITE NETWORK SECURITY INITIATIVES

A multifaceted approach with multiple levels of security is required to protect satellite networks against cyber attacks that can culminate in malicious data corruption; system and service disruptions; unauthorized information disclosure; and physical destruction of satellite assets. Implementation of procedures for safeguarding satellite space and ground segments, TT&C functions, and satellite uplink and downlink transmissions; strategies to optimize satellite network performance; and satellite security protocols to provide authentication and authorization services must be based on a systematic assessment of satellite network risks and a comprehensive determination of satellite network security requirements (Roy-Chowdhury, Baras, Hadjitheodosiou, & Rentz, 2005). Tools, procedures, and measures that aid in safeguarding satellite operations include the enactment of public policies and legislation; the implementation of satellite security protocols and standards; and the utilization of security mechanisms and tools such as encryption.

Public Policies and Legislation

Presidential Decision Directives Nos. 49 (1996) and 63 (1998) define U.S. satellites' space activities as critical to national defense, economic security, and public health and safety and are essential in supporting critical infrastructure protection. U.S. National Security Telecommunications and Information Systems Security Policy (NSTISSP) No.12 establishes a foundation for a nationwide information assurance policy to guide the planning, design, implementation, and operations of secure U.S. space systems (NSTISSC, n.d.). NSTISSP No. 12 measures also mandate that U.S. space systems support information confidentiality, data integrity, user authentication, the availability of information services to authorized users, and service nonrepudiation. Empowered by the U.S. Homeland Security Act of 2002, the U.S. DHS supports comprehensive vulnerability assessments and coordinates nationwide response to threats and attacks classified as cyber terrorism in conjunction with entities that include the U.S. National Infrastructure Protection Center (U.S. DHS, 2003).

International cyber security agreements and public policies such as the Council of Europe's Convention on Cybercrime endorsed in 2001 by 38 countries including the U.S. promote development of international legislation to deter cyber terrorism activities (Wilson, 2005). In 2003, a joint declaration of Cooperation to Combat Terrorism supported by the European Union and the Association of South East Asian Nations (ASEAN, 2003) called for international cooperation in detecting and responding to threats of attacks on satellite assets.

SATELLITE STANDARDS AND PROTOCOLS

Space Communication Protocol Standards

Developed by the Consultative Committee for Space Data Systems (CCSDS), an international consortium that includes space agencies in countries such as Japan, Canada, and the U.S. among its membership, CCSDS establishes the Space Communication Protocol Standards (SCPS) to promote space system interoperability and security. The SCPS suite is based on the Transmission Control Protocol/Internet Protocol (TCP/IP) specifications that also serve as the foundation for the public or commodity Internet. Endorsed by the International Organization for Standardization (ISO) in 1999, the SCPS suite defines approaches for space data transmissions; secure communications to and from space and ground segments; and satellite information management (Hooke, 2001).

The CCSDS SCPS-File Protocol (SCPS-FP) describes processes for error recovery, access control, and privacy. Developed for the U.S. National Aeronautics Space Agency (NASA) Mercury, Surface, Space Environment, Geochemistry, and Ranging (MESSENGER) mission in 2004, the CCSDS-File Delivery Protocol (CCSDS-FDP) , also known as CFDP, supports dependable and secure file transfers across interplanetary distances and standardized file downlink operations between satellites and ground stations (Krupiarz et al., 2002). CFDP uses forward error correction coding to detect data loss and request retransmission.

The SCPS-Transmission Protocol (SCPS-TP) employs algorithms to control data loss resulting from congestion and signal corruption (CCSDS, 1999). Performance enhancing proxies (PEPs) such as TCP spoofing algorithms optimize SCPS-TP operations by enabling header compression, dynamic buffering, and reliable and fast transmissions.

Based on the Integrated-Network Layer Security Protocol (I-NLSP) and the Internet Protocol Security (IPsec) Encapsulation Security Header (ESH) and Authentication Header (AH) protocols, the SCPS-Security Protocol (SCPS-SP) supports data confidentiality; access controls for space operations with minimal overhead; and data protection, authentication, and authorization services (CCSDS, 1999). SCPS-SP encapsulates transport protocol data units (TPDUS) into secure protocol data units (SPDUS) to maintain integrity of space transmissions. It is important to note that SCPS-SP does not recommend the use of specific cryptographic algorithms or key management systems since these functions are handled by the Data-Link Layer or Layer 2 and the Physical Layer or Layer 1 of the seven-layer Open Systems Interconnection (OSI) Reference Model.

Satellite Internet Protocol Security (SatIPSec)

Endorsed by the Internet Engineering Task Force (IETF) and based on IPSec**,** SatIPSec facilitates secure IP unicast transmissions between a single sender and a single receiver and multicast transmissions between a single sender and a multiple group of receivers. In addition to working with IPv4 (IP version 4) and IPv6 (IP version 6), SatIPSec safeguards satellite network operations that are vulnerable to threats and incursions ranging from satellite terminal cloning to eavesdropping (Duquerroy, Josset, Alphand, Berthou, & Gayraud, 2004). SatIPSec maintains data integrity through the use of symmetric encryption that enables authorized multicast group members to verify the origin, identity, and source of multicast transmissions.

Satellite-Reliable Multicast Transport Protocol

Satellite-Reliable Multicast Transport Protocol (SAT-RMTP) was developed by the University of Aberdeen (n.d.) and endorsed by the IETF, SAT-RMTP enables reliable transport and delivery of multimedia files and video clips via GEO satellite constellations to terrestrial networks. Capabilities of SAT-RMTP were verified in tests supported by GEOCAST (Multicast over Geostationary Extremely High Frequency [EHF] Satellites), a European Commission Information Society Technologies (IST) initiative.

Security Tools and Mechanisms

Satellite network security operations are supported by an array of satellite tools and mechanisms ranging from antivirus software and stateful firewalls to attack-resistant or hardened satellite components and physical and logical access controls requiring the use of devices such as smart cards and biometric systems that employ retinal scans and fingerprints for authentication. Redundant security systems for surveillance and fire, flood, and windstorm protection safeguard satellite ground station operations from deliberate cyber attacks, unauthorized use, and natural and artificial disasters.

In the U.S., the Department of Defense Advanced Research Projects Agency (DARPA) is evaluating the security capabilities of pseudolites or pseudo satellites that support redundant communications services if ground station equipment is deliberately disabled. The U.S. Air Force employs antijamming units to safeguard ground station operations; outbound filters to prevent forged source addresses from infiltrating satellite systems; and space telescopes to monitor space activities classified as cyber terrorism. The U.S. Air Force also supports development of sophisticated high-power

space weapons equipped with laser beams to temporarily disable adversary satellites attempting to deny the U.S. utilization of its own space network (U.S. DHS, 2003).

A popular satellite network security tool—encryption—enables safe transmissions via insecure satellite uplinks and downlinks. Conventional encryption systems employ algorithms to scramble plaintext into ciphertext or a meaningless format prior to transmission (Littman, 2002). A key or a secret piece of information typically consisting of a string of random bits enables decryption and the restoration of the message to plaintext or its original format. Digital signatures and time stamps are used in conjunction with encryption to authenticate message integrity. In asymmetric cryptosystems, a public key that is shared by two or more individuals supports encryption, and a private key known only to the message recipient facilitates decryption. In symmetric cryptosystems, the same public key is used for encryption and decryption.

In the absence of strong encryption, cyber intruders can compromise satellite operations by eavesdropping and conducting brute force attacks of weakly encrypted data. In 2002, for instance, unauthenticated subscribers to satellite television programming in the European Union viewed unencrypted surveillance video of U.S. military bases in Bosnia when the cryptosystem was compromised. Satellite television providers generally use conventional encryption systems that employ mathematical algorithms for decryption to ensure that only subscribers with authorized receivers can decrypt television signals and receive delivery of television and pay-per-view entertainment programs. In accordance with NSTISSP No. 12 (NSTISSC, n.d.), U.S. space systems must employ robust encryption to safeguard command and control data and national security information transported between satellite space and ground segments.

The U.S. Army provides access to classified and unclassified information and support services to field units in Iraq via the Combat Service Support Satellite Communications VSAT network. This network employs Triple Digital Encryption Standard (3DES) and complies with the Federal Information Processing Standard (FIPS) 140-2 that specifies security requirements for cryptographic modules. Since 3DES and its forerunner DES are susceptible to hacker attacks, the U.S. National Institute of Standards and Technology (NIST) recommends their replacement by Advanced Encryption Standard (AES), a block cipher encryption algorithm that is unclassified, available without royalty charges worldwide, and complies with FIPS 197 (Burr, 2003). Also called the Rijndael block cipher after its developers Vincent Rijmen and Joan Daemen, AES uses 128-bit, 192-bit, and 256-bit encryption keys.

FUTURE TRENDS

Established by DARPA as a state-of-the-art, next-generation, Internet initiative, the IPN is a deep space backbone network that features a delay-tolerant network (DTN) architecture capable of operating in terrestrial and interplanetary environments with minimal bandwidth, limited power, high error rates, and latencies or delays in the length of time required for information transport from source to destination (Akyildiz, Akan, Chen, Fang, & Su, 2004). The IPN is designed to interlink terrestrial networks including the Internet with remotely located Internets on other planets or spacecraft in transit. Based on CFDP, the IPN bundle layer consists of a delay-tolerant protocol stack that complements DTN architecture. By supporting store-and-forward operations, the bundle layer relays voice, video, image, and data message fragments as bundles from heterogeneous networks via IPN nodes for secure transmission when forward links are established (Burleigh et al., 2003). The store-and-forward operations accommodate uncertain and intermittent interconnectivity between IPN nodes and lengthy propagation delays associated with space-based transmissions. Importantly, the bundling layer also supports the use of multiple data-protection mechanisms to ensure the security of IPN backbone operations and the integrity of information transmitted and exchanged across interplanetary distances in harsh space environments (Hooke, 2001).

Since deep space missions may not have direct line of sight between the earth and the final destination address, IPN bundles may also be transported to recipient locations via satellites that function as intermediate IPN nodes (Akyildiz et al., 2004). Approaches for ensuring secure and reliable IPN transmissions, protecting IPN infrastructure operations from cyber attacks, and utilizing access controls and authentication mechanisms to ensure bundle integrity and data privacy are in development.

CONCLUSION

Satellite network security is dependent on carefully designed and effectively implemented security tools, mechanisms, standards, and protocols; policies and legislation; and procedures that prevent unauthorized entities from gaining access to ground stations; eavesdropping on confidential satellite transmissions; altering satellite information in transit; falsifying command and control data; and destroying satellite assets (Roy-Chowdhury et al., 2005). Even when multiple security devices and countermeasures are in place, however, satellite networks remain vulnerable to ground-based and space-based attacks. The mounting incidents of satellite network cyber incursions attributed to cyber terrorism and the potentially

adverse impacts of these attacks on critical infrastructure operations underscore the importance of building secure satellite networks to protect the integrity, reliability, sustainability, and availability of critical infrastructure resources, services, and initiatives.

REFERENCES

Akyildiz, I., Akan, O., Chen, C., Fang, J., & Su, W. (2004). The state of the art in interplanetary Internet. *IEEE Communications Magazine, 42*(7), 108-118.

ASEAN (2003). *Joint declaration on cooperation to combat terrorism*. Retrieved June 27, 2006, from http://www.aseansec.org/14030.htm

Bosch, O. (2002). Cyberterrorism and private sector efforts for information infrastructure protection. *Creating Trust in Critical Networks. Workshop of the ITU Strategy and Policy Unit*. Retrieved November 15, 2005, from http://www.itu.int/osg/spu/ni/security/workshop/presentations/cnibosch paper.doc

Burleigh, S., Cerf, V., Durst, R., Fall, K., Hooke, A., Scott, K. et al., (2002, Oct. 10-19). The interplanetary Internet: A communications infrastructure for Mars exploration. *53rd International Astronautical Congress: The World Space Congress,* Houston, TX.

Burr, W. E. (2003). Selecting the advanced encryption standard. *IEEE Security & Privacy Magazine, 1*(2), 43-52.

Consultative Committee for Space Data Systems (CCSDS). (1999). *Report on the application of CCSDS protocols to secure systems*. Retrieved November 30, 2005, from http://telecom.esa.int/telecom/www/object/index.cfm?fobjectid=11703

Conti, G., & Ahamad, M. (2005). A framework for countering denial-of-information attacks. *IEEE Security & Privacy Magazine, 3*(6), 50-56.

Duquerroy, L., Josset, S., Alphand, O., Berthou, P., & Gayraud, T. (2004, May 9-12). Satellite Internet Protocol Security (SatIPSec): An optimized solution for securing multicast and unicasts satellite transmissions. *Twenty-second American Institute of Aeronautics and Astronautics (AIAA) International Communications Satellite Systems Conference (ICSSC) and Exhibit,* Monterey, CA.

Hooke, A. (2001). The interplanetary Internet. *Communications of the ACM, 44*(9), 38-40.

Jamalipour, A., & Tung, T. (2001). The role of satellites in global IT: Trends and implications. *IEEE Personal Communications, 8*(3), 5-11.

Krupiarz, C., Burleigh, S., Frangos, C., Heggestad, B., Holland, D., Lyons, K. et al. (2002). The use of the CCSDS file delivery protocol on MESSENGER. *NASA SpaceOps 2002 Conference Papers*. Retrieved October 8, 2005, from http://www.spaceops2002.org/papers/SpaceOps02-P-T5-35.pdf

Littman, M. K. (2002). *Building broadband metworks.* Boca Raton, FL: CRC Press.

National Security Telecommunications and Information Systems Security Committee (NSTISSC). (n.d.). Fact sheet. NSTISSP No. 12. National information assurance (IA) policy for U.S. space systems. Retrieved June 27, 2006, from www.cnss.gov/Assets/pdf/nstissp_12.pdf

Presidential Decision Directive 49. (1996, September 19). *Fact sheet—National space policy.* Retrieved June 1, 2006, from http://www.fas.org/irp/offdocs/pdd/pdd-63.htm

Presidential Decision Directive 63. (1998, May 22). *Fact sheet—Critical infrastructure protection.* Retrieved June 1, 2006, from http://www.fas.org/irp/offdocs/pdd/pdd-63.htm

Roy-Chowdhury, A., Baras, J., Hadjitheodosiou, M., & Rentz, N. (2005). *Hybrid networks with a space segment—Topology design and security issues*. (Tech. Rep. No. 2005-6.) College Park, MD: University of Maryland, Center for Satellite and Hybrid Communication Networks (CSHCN).

University of Aberdeen. (n.d.). *Satellite reliable multicast transport protocol—A network tool for multimedia file distribution*. Retrieved June 27, 2006, from http://geocast.netvizion.fr/download/sat-rmtp.pdf

U.S. Department of Homeland Security (DHS). (2003). *The national strategy to secure cyberspace.* Washington, DC: Author.

U.S. General Accounting Office (GAO). (2002). *Critical infrastructure protection: Commercial satellite security should be more fully addressed* (Pub. No. GAO-02-781). Washington, DC: Author.

U.S. General Accounting Office (GAO). (2004). *Critical infrastructure protection. Improving information sharing with infrastructure sectors* (Pub. No. GAO-04-780). Washington, DC: Author.

U.S. General Accounting Office (GAO). (2005). *Information security: Emerging cybersecurity issues threaten federal information systems* (Pub. No. GAO-05-231). Washington DC: Author.

Waldrop, E. (2005). Weaponization of outer space: U.S. national policy. *High Frontier: The Journal for Space and Missile Professionals, 1*(3), 35-45.

Wilson, C. (2005). *Computer attack and cyberterrorism: Vulnerabilities and policy issues for congress*. Washington, DC: Library of Congress.

KEY TERMS

Cyber Attack: A computer network attack that involves the use of wireline and/or wireless network connections to gain unauthorized access to computing resources in order to control network operations (Wilson, 2005).

Geostationary or Geosynchronous Earth Orbit (GEO): A satellite constellation with three to five satellites that orbit the earth at altitudes of 35,800 kilometers (km) such as military strategic and tactical relay satellite (MILSTAR) that provides jam-resistant communications services for the U.S. military.

Ground Segment: Terrestrial component in a satellite network that manages and controls satellite operations and processes data for storage and transmission.

IPv6: Developed by the IETF. IPv6 extends IP addresses from 32-bits to 128-bits, thereby overcoming IPv4 address shortages and ensuring continued Internet growth and expansion.

Low Earth Orbit (LEO): Satellite constellations that orbit the earth at altitudes ranging from 500 km to 900 km and support applications such as Internet connectivity.

Medium Earth Orbit (MEO): Mid-sized satellite constellations such as the U.S. Global Positioning System (GPS). GPS satellites maintain orbits at approximately 20,200 km above the earth and provide precise positioning services.

Open Systems Interconnection (OSI) Reference Model: Seven-layer architectural model developed by the International Organization for Standardization (ISO) to describe standardized network operations.

Space Segment: Refers to the artificial satellite and its payload in a satellite network. It enables diverse applications in sectors that include e-government, e-learning, and e-medicine.

Satellite-Based Mobile Multiservices Platform

Alexander Markhasin
Siberian State University of Telecommunications and Information Sciences, Russia

INTRODUCTION

The future fourth generation (4G) of the satellite-based wireless and mobile communications is particularly important for global providing of the mobile broadband global information technologies (IT) multi-services and mobile e-applications (m-applications) for geographically dispersed mass users in support of anytime, anywhere, and any required quality of service (QoS) capabilities in a low-cost way. The recent broadband satellite systems described in Ivancic et al. (1999), Evans et al. (2005), Skinnemoen, Vermesan, Iuoras, Adams, and Lobao (2005) are based mainly on *centralized low-meshed architecture* with very high traffic concentration. Such structure is not adequate in context of the traffic topology for *rural, remote, and difficult for access* (RRD) regions. Markhasin (2001) noted that the cost of centralized systems is unacceptably large for deployment of future mass broadband communications in RRD regions (North Siberia, Scandinavia, Greenland, Canada, Alaska, Central and South East Asia, South America, Australia, etc.).

As it was shown in Markhasin (2001, 2004), the future low-cost IT multi-service platforms for RRD regions can be built optimal on a mix of the terrestrial and satellite-based mobile and wireless communications with *radically distributed* (*neural-like) all-IP/ATM architecture* that requires breakthrough steps for search advanced satellite, mobile, and wireless 4G technologies. Markhasin (1996) and Frigon, Chan, and Leung (2001) noted that the improvement of medium access control (MAC) protocols has a dominant effect on ensuring the breakthrough features of future QoS-aware mobile and wireless technologies. The survey and analytical comparison of the fundamental principles of QoS-oriented MAC protocols were described in Markhasin, Olariu, and Todorova (2004, 2005). The radically novel *multi-functional MAC technology* (MFMAC) for long-delay space mediums with fully distributed dynamic control of QoS, traffic parameters, and bandwidth resources was proposed in Markhasin (2001, 2004). This article will be focused on future QoS-aware, satellite-based, fully distributed, mesh, and scalable mobile IT multi-service and m-Applications platform's networking technology 4G for RRD regions.

BACKGROUND

Fundamental Challenges

The fundamental challenges of the future 4G technologies provide answers to the following questions:

1. Can it be true everywhere the 4G declaration "Mobile broadband for all anytime, anywhere, any QoS, any bandwidth?"
- In fact: Now it can be ensured only for urban regions with big population density.
2. Evolutionary or revolutionary way?
- The technological restrictions and "rudiments" of outdated generations will be a burdensome "pay" for evolution way.
3. Multimodal (heterogeneity, the "Babel" of MAC protocols) or multi-functional (homogeneity, universal, scalable, adaptable, the "Esperanto" of MAC protocols)?
- The big-cost's of the multimodal solution prevails up to this time.
4. IP or ATM?
- It will effectively integrate the merits of both these perspective technologies on the base of the developing next generations of asynchronous transfer mode (ATM) and multi-protocol label switching (MPLS) technologies based on breakthrough QoS and space-aware MAC protocols.
5. Centralized, hierarchic or decentralized, distributed, peer-to-peer?
- The cost of centralized systems is unacceptably large for deployment of future mass broadband communications for RRD regions, which include many geographical distributed customers.
6. Ultrahigh bit rates and local areas (wireless technology, WiMAX) or middle bit rates and global areas (mobile technology, 4G)?
- The characteristics of the traditional known broadband MAC protocols depend strongly on wireless area distances, and degrade quickly, if this area is increasing (Markhasin, 2001; Markhasin et al., 2004).

This article will be focused on these 4G fundamental challenges.

Integration of IP and ATM Technologies

What is the most promising telecommunication technology of the future global wireless multimedia communication environment 4G: Internet Protocol (IP) technology or ATM? We suppose that an integration of next generations of these two leading technologies will provide the most promising basis for the near future. The *IP over ATM* integration technologies (IP/ATM, for short) have emerged as advanced concepts that are expected to provide broadband multi-service to the end users by making the best utilization of the advantages of IP (packet switching, adaptive routing, heterogeneous flexibility, scalability, etc.) and of QoS-oriented MAC-based ATM (soft QoS provisioning; fully distributed; dynamical traffic engineering and resource allocation; multimedia; high speed; guarantying; etc.) in a cost-efficient way.

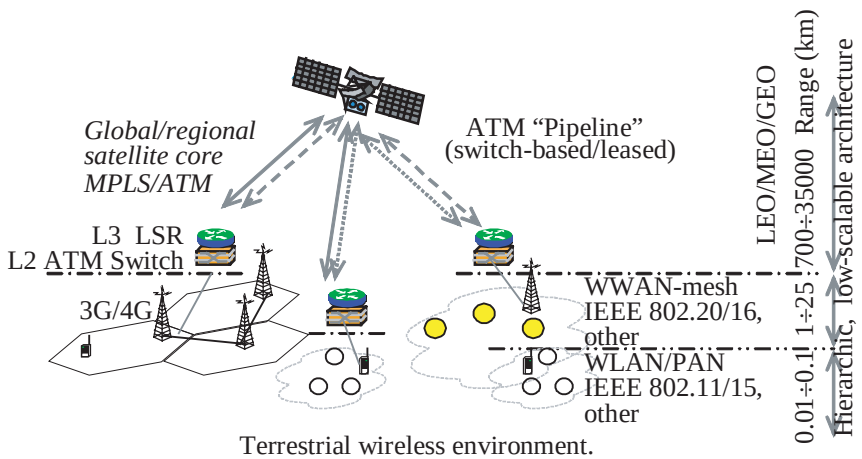
As it was described in Ivancic et al. (1999), Peyravy (1999), and many other authors, the widespread global wireless IP/ATM multimedia systems are based mainly on switch-based centralized architecture (Lawrence, 2001) with very high traffic concentration (see Figure 1). This global/regional satellite core networks domain will be built upon the big ATM switching/*label switch routing* (LSR) nodes, which are conned via satellite switch-based or leased "pipelines." Such structure is adequate to high urban areas especially. The wireless and mobile access networks domain may include the personal (WPAN), the local (WLAN), wide (WWAN), vehicular (VAN) area networks levels, and also cellular systems (B3G/4G, UMTS, S-UMTS). Salkintzis (2004) and many authors of the Special Issue "Migration Toward 4G Wireless Communications" (2004) define a key role of integration of the previously listed wireless and cellular technologies in the 4G of mobile data networks.

The another concept *IP-VAN over Digital Video Broadcast Satellite* of satellite-based broadband access VAN system (DVB-S) was proposed in Oh, Kim, Song, Jeon, and Lee (2005). Its satellite access channel consists of forward and return links having asymmetric star topology. The time division multiplexing (TDM) based forward link is transmitted over DVB-S channels where vehicle's IP packets are encapsulated into MPEG2-TS packets. The return links are composed of multi-frequency multiple access channels with time division (MF-TDMA) or code-division (MF-CDMA) based on Digital Video Broadcast-Return Channel Satellite (DVB-RCS) standard. Unfortunately, the DVB-S/RCS channels' opportunities limited the dynamical control of the QoS, symmetry, and other required characteristics of the VAN system.

MPLS over ATM

The MPLS over (MPLS/ATM, for short) were proposed as very promising extensions to the existing IP/ATM technique (Lawrence, 2001). Lawrence noted that the recent MPLS/ATM is based mainly on routing, switching, and also on further developing the IP capabilities. Switch-based MPLS/ATM's *label distribution protocol* (LDP) supports *hop-by-hop label routing* through the network. Switch-based (L3 LSR) up to date have been based on ATM switches. The ATM cell forwarding mechanism supports also *hop-by-hop ATM label switching*. In Markhasin (2001, 2004) it was

Figure 1. Switch-based MPLS on ATM centralized global satellite/mobile/wireless hierarchic architecture



Terrestrial wireless environment.
Mobile/wireless broadband access networks B3G/4G

shown that switch-based mechanisms support effectively the multimedia satellite networks with highly centralized architecture only (see Figure 1), because the number of such "hops" is strictly limited, and their main components—big ATM switches and high rate "pipeline"—need very high traffic concentration. Unfortunately, its cost is unacceptably high for the satellite multimedia personal networks in context of remote users with geographical distributed traffics and supporting mass market.

The strategic starting point for improvement of next generation of MPLS/ATM technology represents the thesis about dominant effect of MAC layer modernization to the ensuring of novel required abilities. The overview and analytical comparison of the QoS-oriented MAC protocols were considered in Markhasin et al. (2004, 2005). In fact, many MAC protocols proposed for multimedia wireless and satellite (Peyravy, 1999) are based on published per 1970-1980s' years (Rubin, 1979; Tobagi, 1980) classical free, controlled (reservation), and hybrid multi-access methods. As it was described in Markhasin (1996), Peyravy (1999), Rubin (1979), Tobagi (1980), and Wong, Zhy, and Leung (2001), the hybrid or reservation MAC protocols with the defined or fixed *superframe formats* (SFR) are often used in wireless and satellite long-delay mediums.

Requirements to Novel MAC Technology

In Markhasin (1996, 2004) it was shown that the specific conditions of the satellite-based wireless and mobile networks 4G (the long-delay space medium, the geographical distributed multimedia traffics, the soft QoS-oriented MAC, the completely distributed all-MPLS/ATM architecture, the mass market, and others) require the MAC protocol abilities to overcome three principal problems: so-called (1) *time barrier*, (2) *dynamical barrier*, and (3) *economic barrier*.

The time barrier appears due to the effect of *degradation* of long-delay MAC efficiency when the round trip time increases. The dynamical barrier occurs from such essential reason as dynamical instability of the well-known parallel processing reservation MAC schemes with the defined or fixed SFR (Markhasin, 1996). The economic barrier is due to unacceptable costs incurred by any well-known satellite centralized architecture, making a low-cost wireless broadband, ATM mass-market implementation difficult.

In summary, the following abilities must be achieved during creation of soft QoS-oriented long-delay-mediums multi-functional MAC, so-named MFMAC technology:

- high efficiency and throughput of access control to long-delay wireless and space mediums: the time barrier overcoming aspects;
- high controllability, differentiation, stability, and guarantee of dynamical control of QoS, traffic parameters (TP), and bandwidth resource (BR): soft QoS and the dynamical barrier overcoming aspects;
- multi-functional and universal abilities on the basis of the common dynamically controlled and adaptive ATM MAC protocol through the entire network hierarchy—core, backbone, and access networks: all-MPLS/ATM-MFMAC aspects; and
- low-cost, neural-like, fully distributed hyperbus architecture supporting of the MAC sublayers through all networks hierarchy: economic barrier overcoming aspects.

FULLY DISTRIBUTED MOBILE TECHNOLOY 4G

MFMAC Technology

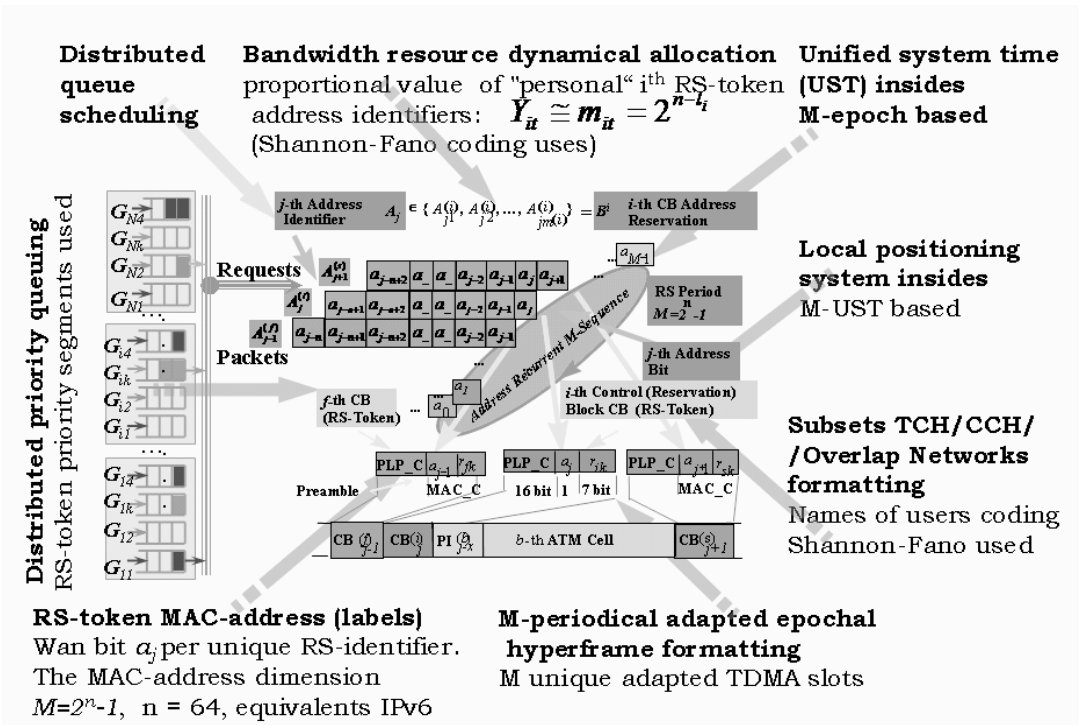
As it was shown in Markhasin et al. (2004), the best required QoS-oriented MFMAC technology capabilities ensure the MAC protocols with adaptive time frame access processes, controlled (reservation) access mechanisms, and parallel conveyer processing of MAC instruction. In one of them, such capabilities can be successfully realized on the basis of developing the *RS-token broadcast reservation* (RS-TBR) *MAC protocol* (Markhasin, 1996, 2001). This protocol uses the recurrent M-sequences (RS) MAC addressing opportunities in order to organize RS-token tools *all-by-one* for highly effective multiple access to long-delay space medium; soft QoS provision and distributed dynamical control of traffic parameters; and bandwidth resources (Figure 2).

Two types of ATM blocks are introduced: control mini-blocks (CB) and information blocks (IB). The control mini-blocks contain a 16 bit preamble PLP_C and an 8 bit access control field MAC_C, including an address bit and the request of the i^{th} station for a block transmission from some number of ATM cells of the k^{th} service class. The information blocks include a packet layer preamble (PLP_I), a field *logical linking control* (LLC), and a variable number of ATM cells. The RS-TBR described in Markhasin (1996, 2001) is used for completely distributed long-delay MAC. Each station adaptively divides the time axis into equidistant synchronous RS-labeled time slots using the common deterministic algorithm of parallel-conveyer requests processing, based on RS-token requests listened to in the broadcast channel. The M-subsequences $A_j = a_{j-(m-1)}, a_{j-(m-2)},...,a_j$ serve as RS-identifiers of the MAC addresses and other protocol subjects, including the labels. Some number m_{it} of "personal" identifiers

$$\{A_{jit}\} = \{A_{ji1},...A_{ji2},...,A_{jim_{it}}\} = B_{it} \quad (1)$$

for passing of requests r_{ik} in proportion to the required bandwidth resource value $[Y_{it}]$ is dynamically assigned to

Figure 2. RS-token "all-by-one" MAC control's mechanism



each i^{th} station on a decentralized basis by Shannon-Fano method. Each j^{th} RS-MAC address can be identified by a unique RS-token using one RS-bit a_j per one MAC address. The dynamical bandwidth assignments and requests scheduling mechanisms are based on this RS-token intensity (the number of RS-token per second) soft regulation in proportion to required resource. The bandwidth dynamical assignments policy can be defined on the basis of the game theory task (Markhasin, 2005). These tasks can create an analytical base for bandwidth brokers (BB). One can show that protocol's realized efficiency can be near to potential capacity (Markhasin, 2004). To support the required k^{th} service classes and queuing discipline, a mechanism of *priority segments* $[(j\text{-}l_{kt}),j)$, $M \geq l_{Kt} \geq ... \geq l_{Kt} \geq ... \geq l_{2t} \geq l_{1t} = 0$ is used (Markhasin, 1996).

At the same time, this protocol allows us to realize a universal RS-token MAC tools *all-by-one* for dynamical (up to real time) soft control of QoS, traffic parameters, and bandwidth resources. This tool implements effective RS-token MAC mechanisms for the optimal planning of control policies, the bandwidth resource scheduler, traffic shaping, dropping discipline, united system time, local positioning date, and, possibly, other added functions simultaneously.

MFMAC-Based Novel ATM Technology

The novel opportunities of the proposed MFMAC technology create a perspective base for developing the next generation of satellite, mobile, and wireless ATM technologies (ATM-MFMAC), which will be highly effective for RRD regions. The fundamental definition of the next generation satellite ATM-MFMAC technology presents the *Virtual Space Medium ATM Hyperbus*, which operates by MFMAC protocols (Brandt et al., 2001; Markhasin, 2001). In fact, it is a QoS-aware, fully distributed, dynamically controlled, and broadband ATM wireless hyperbus, which "was lifted up to the sky," that is, "Sky-Bus." The Space Medium ATM Hyperbus builds a virtual bus between all the users (which are in sight of a given satellite). The satellite in this case is merely a retranslator so those users in different geographical locations can share the same virtual bus.

The MFMAC's *all-by-one* mechanism implements an adaptive highly effective multiple access to long-delay space medium; soft QoS provision and distributed dynamical control of traffic and bandwidth resources; and other important MAC mechanisms. The dimension of this MAC-address space can be equivalent to IP-address capacity of

Table 1. Comparison of the recent ATM and the satellite next generation ATM-MFMAC

Features	The recent generation ATM	The proposed satellite next generation ATM-MFMAC
Physical channel topology	"Multipoint-to-multipoint," *not distributed ("pipe")*	"Multipoint-to-multipoint," *radically distributed ("hyperbus")*
Multiple access (MA) technology	*Single* station MA with statistical multiplexing, *fully centralized*	*Multi* station MA with statistical multiplexing, *radically distributed*
MAC technology	*Not QoS-aware*, *not adapted*, long-delay-aware only for *"pipe"* topology	*QoS-oriented*, multifunctional, *adapted "on-the-fly,"* long-delay-aware for fully distributed *"hyperbus"* topology
MAC address dimension	The dimension M = 256, used time resource—*eight* address bit per one address	$M=2^n-1$, n = 64÷128, equivalents IPv6, used time resource—*one* address bit per one address
QoS control	*Static* planning a set of fixed class of quality (CBR, ABR, etc.), *Hard QoS (not proportional)*	*Dynamical* ("on-the-fly") control of QoS, traffic parameters, and bandwidth resources, *Soft QoS (proportional)*
Cells format	53 octets	It is possible and expedient to increase
Used networks	Transport and corporative networks	Access and transport networks, integrated satellite/mobile/ wireless, and so forth.
Supports region conditions	*High urban*	*Rural, remote, difficult of access* (mountains, oceans, islands, tundra, etc.)
Supports services and applications	*Stationary* multimedia services and e-Applications	*Mobile* multimedia services and m-Applications
Cost	*Large-cost* per customer	*Low cost* per customer

IPv6. The ATM Hyperbus builds a virtual bus connection between all L2 nodes.

The proposed ATM Hyperbus architecture provides a wide range of services with different traffic streams and different QoS characteristics. QoS guarantees are essential for the support of real time services such as telephone and video. ATM technology allows temporal QoS requirements to be taken into account, flexibility for bursts traffic, and bandwidth allocation to isochronous applications with various adaptation layer (AAL) protocols.

The comparison of the recent ATM technology and the proposed next generation satellite ATM MFMAC-based technology is shown in Table 1.

MFMAC-Based Novel Generation Satellite MPLS/ATM

In Markhasin (2001) a novel MFMAC-based next generation MPLS over ATM-MFMAC integration technology (MPLS/ATM-MFMAC, for short) was proposed.

The proposed generation MPLS/ATM-MFMAC technology is suitable for completely distributed all-MPLS/ATM satellite networks architecture and soft QoS provisioning. This generation was based on the principle of using an ATM *selecting technique,* rather than ATM *switching technique*. In fact, the novel alternative solution combines a multi-functional, long-delay, broadband next generation L2 ATM-MFMAC technology (Markhasin, 2001) with fully distributed multi-functional L2 ATM Hyperbus architecture (Brandt et al., 2001; Markhasin, 2001), and the L3 MPLS over ATM routing technique (Lawrence, 2001).

The topology of virtual bus can be represented by a passive optic tree, wireless or satellite bus, cross-hexagonal (see Figure 3), ring-radial ("metro") topology, and so forth. The suggested MPLS/ATM structure can include a few of L3 LSR/MFMAC *routing nodes* (RN) for buses or edge interworking, and possibly many hundreds of L2 ATM/MFMAC scalable *transit nodes* (TN) for ATM data cells *label selecting*. The novel MFMAC-based MPLS/ATM does not require hop-by-hop switching mechanisms. The label distribution protocol uses a label selecting mechanism for TN and *label routing mechanism* for RN for setup in the selected path through the network. The ATM cell-forwarding protocol uses also a non-hop-by-hop label selecting mechanism for transit and routing.

CONCEPTUAL MODEL OF A SATELLITE-BASED FULLY DISTRIBUTED MOBILE PLATFORM'S TECHNOLOGY 4G

The conceptual look of the fully distributed (neural-like) all-MPLS/ATM-MFMAC architecture of future global

Figure 3. The next generation alternative MPLS over ATM-MFMAC (fiber optic analogy)

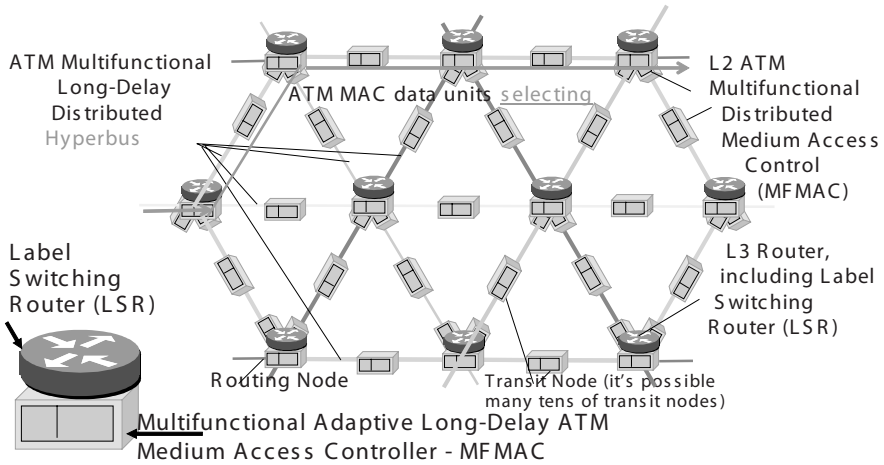
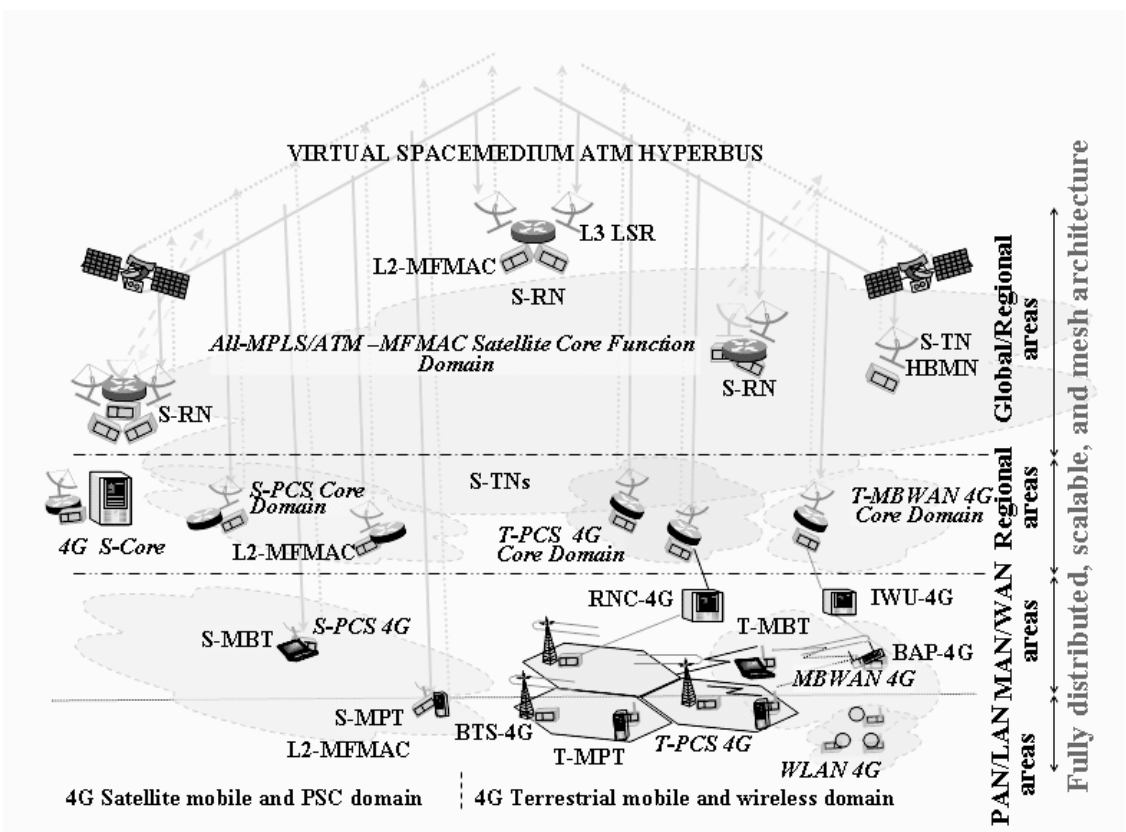


Figure 4. The conceptual model of global fully distributed scalable all-IP/ATM wireless mesh environment 4G: L2-MF-MAC—Universal MAC controller; L3 LSR—Label switch router; S-TN—Transit/selecting node; S-RN—Routing node; HBMN—Hyperbus managing node; MBWAN—Mobile broadband wireless access network; MBT—Mobile broadband terminal 4G; S-MBT—Satellite MBT terminal 4G; MPT—Mobile broadband PCS terminal 4G; S-MPT–Satellite PSC terminal 4G; BAP—Broadband access point; S-PCS—Satellite personal communication system 4G; T-PCS—Terrestrial PSC 4G.



satellite/mobile/wireless multimedia networks is explained in Figure 4.

The proposed architecture of the global wireless and mobile environment is based on previously described novel alternative MPLS over ATM-MFMAC integration, and advanced soft-QoS-oriented multi-functional MFMAC technology. Their structure is created on so-called distributed Virtual Spacemedium ATM Hyperbuses and uses the common universal dynamically adapted MFMAC protocol through the entire networks hierarchy—core, backbone, and access networks. The orbital groups geostationary (GEO), medium Earth (MEO), or low-Earth/high elliptical (LEO/HEO) orbits of satellite broadband digital retranslator ensure the global coverage.

The retranslator forms a multi-access up channel and broadcast down channels. The virtual hyperbusses L2 for various wireless and satellite MPLS/ATM systems, which differ according to the classes of the topology scales (W-MAN, W-WAN, S-UMTS, S-PSC), function hierarchy, technical characteristics, and kinds of medium, can be configured using up to three logically homogeneous basis components: (1) universal adaptive medium interface with medium type cartridge—L1-UAMI, (2) universal adaptive multifunctional distributed ATM-MFMAC medium access controller—L2-MFMAC, and (3) the hyperbus's bandwidth broker server (HBBS) of the distributed dynamical control of QoS, traffic parameters (TP), and bandwidth resource assignment (BR). Moreover, the same hyperbus can implement simultaneously (in parallel) a function of access network, and function of core network or/and backbone, that is, it is mult-ifunctional.

FUTURE TRENDS

The satellite-based domain will be having a common core of high layer standard protocols of the eMobility Technology Platform, and different (MFMAC-based) physical and channel layers protocols. In fact, it is necessary to develop the added MFMAC-based Universal Adaptive Interface with Mediums Type Cartridge (L1-UAMI) and Universal Adaptive Multifunctional Distributed ATM-MFMAC Medium Access Controller (L2-MFMAC) standards, and possibly some overhead extending of the high layers protocol, which realizes the soft QoS and other MFMAC novel opportunities (L2/L3 QoS Manager/Bandwidth Broker, etc.). For example, it is necessary to develop the novel 4G Air Interface (mobile MFMAC), IEEE 802.16x/20y (wireless mesh, soft QoS, long-delay, multi-functional, reconfigurable) IEEE 802.2Xz (satellite MFMAC), and so forth, standards.

CONCLUSION

What is our vision of the place in the future mobile and wireless world of the described MFMAC-based fully distributed neural-like All-IP/ATM integrated satellite, wireless and mobile networking technology 4G? It is very expedient and well grounded to consider this fully distributed networking technology 4G as satellite-based RRD-domain of the future mobile IT/ITC-Multiservices, m-Applications, and eMobility (Tafazolly, Correia, & Saarnio, 2005) Platform's Technology 4G, which will guarantee a low cost and friendly access to global/regional IT-Resources/Multiservices for mass subscriber by covering the RRD areas.

REFERENCES

Brandt, H., Todorova, P., Lorenz, P., Markhasin, A., Milne, P., & Ristol, S. (2001). Multifunctional distributed broadband ATM with dynamic control of QoS Hyperbus over satellite. *19 AIAA International. Communication Satellite System Conference,* Toulouse, France, (pp. 1-7).

Evans, B., Werner, M., Lutz, E., Bousquet, M., Copazza, G. E., Maral, et al. (2005), Integration of satellite and terrestrial systems in future multimedia communications. *IEEE Wireless Communications, 12*(5), 72-80.

Frigon, J.-F., Chan, H. C. B., & Leung, V. C. M. (2001). Dynamic reservation TDMA protocol for wireless ATM networks. *IEEE JSAC, 19*(2), 370-383.

Ivancic, W. D., Brooks, D., Frantz, B., Hoder, D., Shell, D., & Beering, D. (1999). NASA's broadband satellite networking research. *IEEE Communications Magazine, 37*(7), 40-47.

Lawrence, J. (2001). Design multiprotocol label switching networks. *IEEE Communications Magazine, 39*(7), 134-142.

Markhasin, A. (1996). Multi-access with dynamic control of the traffic and service quality in broadband ATM networks. *Optoelectronics, Instrumentation and Data Processing, 3,* 93-99.

Markhasin, A. (2001). Advanced cost-effective long-delay broadband ATM medium access control technology and multifunctional architecture. In *Proceedings of the IEEE International Communication Conference—ICC'2001,* Helsinki, Finland(pp. 1914-1918).

Markhasin, A. (2004). QoS-oriented medium access control fundamentals for future all-IP/ATM satellite multimedia personal communications 4G. *Proceedings of the IEEE International Communication Conference—ICC'2004*, Paris, France (pp. 3963-3968).

Markhasin, A. (2005, June 19-23). QoS-Oriented MAC technology for distributed All-MPLS/ATM satellite integrated platform for 3G+/4G and WLAN communications. *14th IST Mobile & Wireless Communications Summit, Dresden, Germany,* paper #215 (pp.1-5). Retrieved from, www.mobilesummit2005.org

Markhasin, A., Olariu, S., & Todorova, P. (2004). QoS-oriented medium access control for all-IP/ATM mobile commerce applications. In N. S. Shi (Ed.), *Mobile commerce applications* (pp. 303-331). Hershey, PA: Idea Group Publishing.

Markhasin, A., Olariu, S., & Todorova, P. (2005). QoS-oriented MAC protocols for future mobile applications. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (Vol. 1-4) (pp. 2373-2377). Hershey, PA: Idea Group Reference.

Migration toward 4G wireless communications (2004). *IEEE Wireless Communications, 11*(3), 6-42.

Oh, D. G., Kim, P., Song, Y. J., Jeon, I. J., & Lee, H.-J. (2005). Design considerations of satellite-based vehicular broadband networks. *IEEE Wireless Communications, 12*(5), 91-97.

Peyravy, H. (1999). Medium access control protocols performance in satellite communication. *IEEE Communications Magazine, 3*(37), 62-71.

Rubin, I. (1979). Access-control disciplines for multi-access communication channels: Reservation and TDMA schemes. *IEEE Transactions on Information Theory, IT-25*(5), 516-536.

Salkintzis, A. K. (2004). Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks. *IEEE Wireless Communications, 11*(3), 50-61.

Skinnemoen, H., Vermesan, A., Iuoras, A., Adams, G., & Lobao, H. (2005). VoIP over DVB-RCS with QoS and bandwidth on demand. *IEEE Wireless Communications, 12*(5), 46-53.

Tafazolly, R., Correia, L. M., & Saarnio, Y. (Eds.)(2005). Strategic research agenda, Version 4. eMobility and Wireless Communications Technology Platform (pp. 1-35). Retrieved November 23, 2005 from www.emobility.eu.org/research_agenda.html

Tobagi, F. A. (1980). Multi-access protocols in packet communications systems. *IEEE Transaction on Communications, 28*(4), 468-488.

Wong, W. K., Zhy, H., & Leung, V. C. M. (2003). Soft QoS provisioning using the token bank fair queuing scheduling algorithm. *IEEE Wireless Communications, 10*(3), 8-16.

KEY TERMS

All-IPArchitecture: A networks interworking's architecture based on the set of Internet protocols (IP) end-to-end.

Asynchronous Transfer Mode (ATM): A QoS-aware, broadband communications technology that supports multistreams transfer of traffic of any kind (multimedia, video, voice, date, etc.) into so named cells with fixed length of 53 bytes.

Medium Access Control (MAC): According to the Open System Interconnection (OSI) terminology, an interface between the Physical Layer (PhL) and Logical Link Control (LLC) layers.

Mobile Multi-Service: An integrate provisioning of the several kinds of mobile services (voice, video, data, mobile Internet, etc.).

Multi-Functional MAC (MFMAC): Universal multifunctional MAC technology which guarantees QoS-oriented, fully distributed, and dynamical control of the MA to long-delay mediums by any networks function—access, core, transport backbone, and so forth.

Multiple Access (MA): Procedures regulating the simultaneous use of a common transmission medium by multiple users.

Quality of Service (QoS): Collection of performance parameters for network service including bandwidth, average delay, jitter, packet loss probability, among many others.

Category: Global Information Technology

Sectoral Analysis of ICT Use in Nigeria

Isola Ajiferuke
University of Western Ontario, Canada

Wole Olatokun
University of Ibadan, Nigeria

INTRODUCTION

Information and communications technologies (ICTs) have become key tools and had a revolutionary impact of how we see the world and how we live (Dabesaki, 2005). They have the potential to be a major driving force behind the economic growth of any nation because of their potentially strong restructuring impact on existing economic activities and the ability to affect economic activities in a variety of ways. These include improving the quality of existing services, creating new services, raising labor and productivity, increasing capital intensity, enhancing economics of scale, and creating new economic structures. ICTs are also paving the way for greater ease of movement of technical and financial services, and are instrumental to development during the rapid globalization process. From the information technology revolution, a new kind of economy emerges. This is the information-based economy in which information along with capital and labor is a critical resource for creation of income and wealth for the enhancement of competitiveness. ICTs have also left their mark on the political and social dimensions of development, specifically by enhancing participation in decision-making processes at the corporate, local, and national levels.

It is an established fact that a few developing countries like China, India, and Brazil are successfully taking advantage of the opportunities information and communications technologies offer and have made significant improvement in their economic, and many more developing countries (including Nigeria) are beginning to derive some of the potential benefits. For most of the developing world, however, information and communications technologies remain just a promise, and it seems a distant one at that. There is little evidence from past experience of national and international development policies, strategies, and programs to suggest that much will change for large segments of the world's poorest people.

Nigeria, like most developing countries, is an "information-poor" country where the deployment and application of ICTs is still in its infancy. This article, which is an updated version of an earlier one (Ajiferuke & Olatokun, 2005), presents the current status of ICT in Nigeria, particularly its applications in some sectors of the nation's economy. It also identifies some inhibitions to the effective deployment and exploitation of ICT in Nigeria and concludes with a discussion of the policy issues, challenges, and prospects of ICT use in Nigeria.

BACKGROUND

ICT Initiatives, Policy Formulation, and Implementation

The Federal Government of Nigeria has accorded ICT a national priority. This is evident in the approval of the National Information Technology Policy (NITP) and the subsequent establishment of the National Information Technology Development Agency (NITDA) in 2001 to serve as a bureau for the implementation of the NITP. The policy recognized the private sector as the driving engine of the ICT sector. NITDA is to enter into strategic alliance, collaboration, and joint venture with the private sector for the actualization of the ICT vision, which is to make Nigeria an ICT-capable country using ICT as an engine for sustainable development and global competitiveness. It is also to be used for education, job creation, wealth creation, poverty eradication, and global competitiveness.

A sectoral application of ICT has been recognized in the formulation of the ICT policy, which involves the development of the following areas of the economy: Human Resource Development, Infrastructure, Governance, Research and Development, Health, Agriculture, Urban and Rural Development, Trade and Commerce, Arts, Culture and Tourism, National Security and Law Enforcement, Fiscal Measures, and so forth. According to Ajayi (2002), NITDA has embarked on a number of projects aimed at stimulating the growth of ICT in the country. The Public Service Network (PSNet) is one such project, aimed at addressing the major problem of ICT infrastructure, which will serve as a pipeline for ICT services. It consists of a Very Small Aperture Terminal (VSAT) sited in each state capital. This VSAT provides Internet access for that central location and all other locations connected to this center using broadband wireless access (BWA) technology. The various sites around the country are then connected to each other through a vir-

tual private network (VPN). Nine states have already been connected in the first phase of the project.

Human capacity building has been another focus of NITDA. Towards realizing this goal, NITDA has forged a thriving partnership with public and private organizations in what has become a public-private partnership (PPP). The Enterprise Technology Center (ETC) is one such PPP that is worthy of note. The ETC is a partnership between NITDA and two private companies to provide ICT training for civil servants. NITDA has also collaborated with several multinationals and international organizations to deliver specialized training in some train-the-trainer workshops. These institutions include UNESCO, the International Center for Theoretical Physics (ICTP), and Cisco Systems.

The Telephone System in Nigeria and the GSM Revolution

The telephone system in Nigeria has been challenged for years. A breakthrough in telephone infrastructure emerged in January 2001 when the sector was totally liberalized, leading to the Nigerian Communications Commission (NCC) issuing four wireless licenses to MTN Nigeria Communication, Econet Wireless Nigeria Limited (now Celtel), Communications Investment Limited (CIL), and state-owned NITEL. CIL, however, had its license withdrawn because of its inability to meet the deadline for payment of the license fee. The fourth GSM provider, Glomobile (Globalcom), though it won its multiple licenses in September 2002 for the provision of telecommunications services, did not commence provision of mobile phone services until August 2003. Since the GSM launch, mobile telephony has rapidly become the most popular method of voice communication in Nigeria. Indeed these developments have been truly explosive: today Nigeria has about five million mobile lines and about one million fixed lines, compared with just about 450,000 working lines from NITEL four years ago.

Internet Usage in Nigeria

The Internet has experienced relatively slow growth in Nigeria and other countries in the poorer regions of the world such as sub-Saharan Africa. Nigeria is among the nine countries in Africa that achieved full Internet access in their capital cities and some secondary towns by 1998 (Jensen, 1998). Available statistics show that Internet usage in Nigeria between 2000-2003 stood at 200,000, representing 0.1% of the total population. For Nigeria with a population of over 120 million, growth rate has been stagnant. This may have contributed to the slow pace in socio-economic development in the country. According to Internet Usage statistics for Africa, Internet usage in Nigeria remained stable until 2003, but increased to 5 million in 2006, about a 3.1% penetration rate (see Table 1). Many factors, including high cost of bandwidth and telephone lines, might have made the majority of the operators to charge within the range of N100 per hour, which many surfers might not be able to afford as a result of the prevailing economic realities.

MAIN THRUST OF THE ARTICLE

ICT Application in Various Sectors

ICT Usage in Manufacturing

ICT usage in manufacturing companies in Nigeria has improved production and made it easier to perform difficult tasks. ICT is used in product design, such as in computer-aided design and computer-aided manufacturing. ICT is also used in the control of processes, as well as the control of machinery. A survey of 22 manufacturing companies conducted in Lagos and Ibadan in March 2005 showed that ICT facilities exist at various levels and departments in the companies (Lasaki, 2005). ICT is also being used to capture information at all stages of design, manufacturing, and marketing. According to Lasaki (2005), "There is considerable evidence that the Nigerian manufacturing sub-sector is moving towards proper integration into the information society."

Use of ICT in Education

Education helps in the development of skills and the acquisition of knowledge. There is a need for continuous improvements in education. ICT is playing a tremendous role in educational developments in Nigeria. Online admission into Nigerian schools is becoming more and more popular. Many schools have institutional Web sites where information about

Table 1. Nigeria's Internet usage and population growth (ITU, 2003; Internet World Stats, n.d.)

Year	Users	Population	% Penetration	Usage Source
2000	200,000	142,895,600	0.1 %	ITU
2006	5,000,000	159,404,137	3.1 %	ITU

them is available to the public. There are students in Nigeria who will eventually obtain degrees via the Internet from institutions of higher learning that are outside the country. Within the country, admission registration processes have improved in schools and agencies that make good use of ICT facilities. ICT use has also improved distance learning, especially with the resuscitation of the national Open University of Nigeria (NOUN), which delivers its courses through a combination of Web-based modules, textual materials, audio and videotapes, as well as CD-ROMs.

Use of ICT in Banking

Use of ICT in banking has resulted in electronic banking (e-banking). E-banking in Nigeria has been in an upbeat mode. Some of its features include the use of automated teller machines (ATMs), pay-by-phone systems, personal computer banking, point-of-sale transfers, and the electronic check conversion. The financial marketplace has been actively promoting its online publications and functionalities. Nigerian banks are becoming very innovative, launching new services with the use of ICT facilities. For example, an automated check clearing system (NAACCS) was launched last year through the Central Bank of Nigeria.

Use of ICT in the Print Media

Print media are involved in publishing written materials such as newspapers and magazines for the public. However, the advent of the information age has brought about major changes in the Nigerian print media with the adoption of ICT in the industry. Some of the benefits of the use of ICT in the print media are enhanced performance (speed, skills, and quality), improvement in communication, improvement in picture resolution and general graphics, reduced cost of production (which results in increased profits), better page layout, good image, high rating, and realization of corporate goals (Ehikhamenor, 2003). Use of ICT therefore helps to gain competitive advantage, which is necessary for survival in the print media industry.

A survey of 19 print media houses to evaluate the level of utilization of ICTs in Nigerian print media industries found that all the print media houses surveyed have ICT facilities like computers, printers, scanners, telephone facilities, lithographic machines, FAX machines, and photocopiers (Erigha, 2006). Many of them now have online versions of their newspapers on the Internet.

Use of ICT in Medicine

Technological advancements have greatly improved healthcare delivery worldwide. Nigeria is not excluded in this development. Use of ICT in medicine has helped both in routine work and complicated medical procedures. ICT has especially helped in the area of public awareness and enlightenment about health issues, as there are over 10,000 health-related Web sites (Ashville Citizens Times, 1997, as cited in Ibeguam, 2004). In Nigeria for example, many people are aware of HIV/AIDS (means of infection, prevention, proper attitude towards those infected, and management of the infection) as a result of good use of ICT facilities. Medical research also depends a lot on the effective use of ICT. Data collection, analysis, interpretation, and report of findings are usually done by means of ICT facilities. For example, the University College Hospital (UCH), a center for medical research in Nigeria, has embarked on a study on effective diagnosis of cancer by electronic means.

ICT in Governance

Electronic government (e-government) involves the use of ICT facilities to deliver public services to citizens and businesses. It entails the transformation of public services to citizens using new organizational processes and new technologies (Gunter, 2006). E-government aims at making government services more accessible, more customer focused, and more relevant to citizens. It encourages citizens' involvement in governance, enhanced communication links, harmonized organizational practices, and partnerships between different layers of government. It focuses on improving public services and internal workings of government organizations.

E-governance has been partially adopted by the Federal Government of Nigeria and a few states. For example, the voters' registration for the April 2007 general elections was done electronically, while a few state governments are providing their citizens with online access to important forms. However, some states are still lagging behind in the adoption of ICT in governance. For example, a recent survey that sought to measure the preparedness of Oyo state government for e-governance shows that the level of computerization of government agencies in Oyo state is low, perception of e-governance is low, and information dissemination records and information management practices are poor (Ogunsola, 2006).

Issues and Challenges of ICT Usage In Nigeria

No doubt, there has been substantial improvement in access to telecommunications facilities and unprecedented growth in the telecommunications network in Nigeria. However, in view of Nigeria's size and requirements, the telecommunications infrastructure is still grossly inadequate. Among others, the following are the major challenges and barriers inhibiting ICT usage in Nigeria:

a. *Cost:* In Nigeria, high access cost is still a major barrier. While prices have definitely come down, the cost of access is still too high to have a transformatory impact. There are presently price competition battles going on involving private telecommunication operators (PTOs) and GSM providers which are steps in this direction. The provision of the Internet by PTOs is also bringing down the cost. But more needs to be done about bringing down call tariffs and rates, not just communications acquisition cost.
b. *Poor Energy Infrastructure:* Epileptic public power supply in Nigeria increases the cost of access. Supply of electricity needs to be optimal to enable businesses and banks to provide seamless online services through local areas networks, wide area networks, and the Internet. Inefficiency is the word to describe a situation where power generators are the only reliable suppliers of power. This constitutes a barrier to growth and sustainable development. The growth of real e-business cannot take place or be of any significance in an environment with unreliable public power supply.
c. *Quality of Service:* While availability has grown, this has not been matched by quality of service. It is not enough to have cheap lines and low-cost bandwidth. Efficiency and accessibility of telecommunications service should be paramount. Most operators have a lot of work to do in the area of quality of service, especially congestion control and support services. The National Communications Commission (NCC) may need to wield the big stick to ensure quality service by sanctioning poor performers.
d. *Appropriate Licensing Fees:* NCC has done a lot as a pacesetter. But NCC needs to review the appropriateness of its license fees. How realistic are such fees for healthy competition? Will such fees as they are stimulate telecommunications growth or increase the number of competent market players?

CONCLUSION AND FUTURE TRENDS

Nigeria's present status and indicators point to a possible growth in ICTs. For example, Africa's growth in the number of telephone subscribers from 14 million fixed lines and 1 million mobile lines in 1996 to 22 million fixed lines and 28 million mobile lines in 2001 and over 32 million fixed lines and 82 million mobile lines in 2005 (BBC News, 2005) is a strong indication that the growth of ICT in Africa and indeed Nigeria is real. It is noteworthy that Nigerian banks are beginning to embrace electronic banking (e-banking), and in spite of all the bottlenecks and assumed odds, e-banking will help produce a better banking industry, one that is accessible *anytime, anywhere.* Also, the adoption and deployment of ICTs in distance education is beginning to gain ground as presented in the preceding sections. So also is the usage of ICTs in governance becoming a reality in transacting business in some federal government ministries and agencies, as well as in some states.

Potential investors are attracted because of the country's size and population. But is the environment conducive for growth in the ICT sector? Is there ICT growth for the benefit of Nigerians? Considering Nigeria's size and potential, growth in ICT is definitely poor. The future of Nigeria depends not on oil, but on the quality of human capital. In the digital age, Nigeria needs quality manpower. We must get our priorities right by investing seriously in human capital. This means focusing on increased computer literacy and ICT professionalism.

In sum, for Nigeria to reap from the numerous opportunities offered by ICT and for ICT to be effectively deployed and applied in the various sectors, it is an imperative to adopt policies aimed along the following areas and directions:

- finding a holistic solution to the perennial energy problem;
- developing ICT infrastructure;
- increasing the telephone and mobile phone penetration rates;
- increasing general literacy of Nigerians;
- increasing ICT literacy through the introduction of computer education from the early stages of education, and organizing ICT training for civil servants in government ministries and agencies;
- further deregulating and liberalizing the mobile telephony sector;
- reducing tariffs on ICT equipment for it to become more affordable and widespread with a view to increasing accessibility;
- making the National Information Technology Development Agency (NITDA) more autonomous; and
- adopting strategies to bridge the digital divide (gender, income, literacy, etc.).

REFERENCES

Ajayi, G.O. (2002). African response to the information communication technology revolution: Case study of the ICT development in Nigeria. *ATPS Special Paper Series*, (8), 1.

Ajiferuke, I., & Olatokun, W. (2005). Information technology usage in Nigeria. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (vol. I-V, pp. 1508-1512). Hershey, PA: Idea Group.

Akpore, S.A. (2005). *E-learning takes root in Nigeria.* Retrieved December 6, 2006, from http://www.afrihub.com/pages/corporate/news/20050118002.htm

Awe, J. (2006). *Nigeria: Bridging the infrastructure divide.* Retrieved December 7, 2006, from http://www.jidaw.com/telecom/telecomm8.html

BBC News. (2005). *Mobile growth fastest in Africa.* Retrieved December 26, 2006, from http://news.bbc.co.uk/1/hi/business/4331863.stm

Brown, M.M. (2000). Commentary: The Internet and development. *Choices,* 9(2).

Dabeski, M. (2005). *Building e-governance capacity in African countries.* Retrieved December 17, 2006, from http://www.globelicsindia2006.org/II-3/Hassan%20Wunmi.doc

Ehikhamenor, F.A. (2003). The information society and the Nigerian print media. *African Journal of Library, Archival and Information Science, 13*(2), 187-199.

Erigha, A.O. (2006). *an evaluation of the level of utilization of ICTs in Nigerian print media*. Master's Degree Project, University of Ibadan, Nigeria.

Gunter, B (2006). Advances in e-democracy: Overview. *Aslib Proceedings, 58*(5), 361-370.

Ibegwam, A. (2004). Internet access and usage by students of the College of Medicine, University of Lagos. *The Information Technologist, 1*(1&2), 81-87.

Internet World Stats. (n.d.). Retrieved from http://www.internetworldstats.com/af/ng.htm

ITU (International Telecommunications Union). (2003). *Low Internet usage growth reported in Nigeria.* Retrieved October 6, 2006, from http://findarticles.com/p/articles/mi_qn4175/is_20030924/ai_n12930786

Jensen, M. (1998). Internet opens new markets for Africa. *Africa Recovery, 12*(3), 6-7.

Lasaki, O.M. (2005). *A survey of the adoption and use of information and communication technologies in the manufacturing industry.* Master's Degree Project, University of Ibadan, Nigeria.

Mac-Ikemenjima, D. (2003) The Integration of ICT into the school system: Our roles. *Proceedings of the Student Leaders IT Conference.*

Ogunsola, K. (2006). *E-government and e-governance in Nigeria: A case of preparedness of Oyo state government agencies*. Master's Degree Project, University of Ibadan, Nigeria.

Olorunda, O., & Oyelude, A.A. (2003). Professional women's information needs in developing countries: ICT as a catalyst. *Proceedings of the World Library and Information Congress: 69th IFLA General Conference and Council.* Retrieved December 7, 2006, from http://www.ifla.org/IV/ifla69/papers/600-Olorunda_Oyelude.pdf

Persaud, A. (2000). The perils of neglecting the Net: Unless developing countries act fast, the Internet will make them fall behind in the global economy. *The Financial Times, 17,* 15-17.

Petrazzini, B., & Kibati, M. (1999). The Internet in developing countries. *Communications of the ACM, 42*(6), 31-35.

Ugwoke, F. (2000). NCC laments Africa's poor Internet status. *Africa News Service,* (July 27).

KEY TERMS

Global System for Mobile Communication (GSM): A digital mobile telephone system that is widely used in Europe and other parts of the world. GSM uses a variation of TDMA and is the most widely used of the three digital wireless telephone technologies (TDMA, GSM, and CDMA). GSM digitizes and compresses data, then sends it down a channel with two other streams of user data, each in its own time slot.

Information and Communications Technology (ICT): An umbrella term that includes any communication device or application, encompassing radio, television, cellular phones, computer and network hardware and software, satellite systems, and so on, as well as the various services and applications associated with them.

Information Technology: Encompasses all forms of technology used in processing and disseminating information.

Internet: An interconnected system of networks that connects computers around the world via the TCP/IP protocol.

Policy: A plan of action to guide decisions and actions. The term may apply to government, private sector organizations and groups, and individuals.

Security and Privacy in Social Networks

Barbara Carminati
Università degli Studi dell'Insubria, Italy

Elena Ferrari
Università degli Studi dell'Insubria, Italy

Andrea Perego
Università degli Studi dell'Insubria, Italy

INTRODUCTION

Web-based social networks (WBSNs) are online communities that allow users to publish resources (e.g., personal data, annotations, blogs) and to establish relationships, possibly of a different type ("friend," "colleague," etc.) for purposes that may concern business, entertainment, religion, dating, and so forth. In the last few years, the usage and diffusion of WBSNs has been increasing, with about 300 Web sites collecting the information of more than 400 million registered users. As a result, the "net model" is today used more and more to communicate, share information, make decisions, and 'do business' by companies and organizations (Staab et al., 2005).

Regardless of the purpose of a WBSN, one of the main reasons for participating in social networking is to share and exchange information with other users. Recently, thanks to the adoption of Semantic Web technologies such as FOAF and other RDF-based vocabularies (Brickley & Miller, 2005; Davis & Vitiello, 2005; Golbeck, 2004), accessing and disseminating information over multiple WBSNs has been made simpler (Ding, Zhou, Finin, & Joshi, 2005). If this has been quite a relevant improvement towards an easier sharing of information, it makes more urgent that content owners have control over information access. In fact, making available possibly sensitive and private data and resources implies that they can be used by third parties for purposes different from the intended ones. As a matter of fact, users' personal data and resources are regularly exploited not only by companies for marketing purposes, but also by governments and institutions for tracking persons' behaviors and opinions, and in the worst case, by online predators (Barnes, 2006).

It is then a challenging issue to devise security mechanisms for social networks, able to protect private information and regulate access to shared resources. In this article, besides providing an overview of the characteristics of the WBSN environment and its protection requirements, we illustrate the current approaches and future trends to social network security, with particular attention paid to the emerging technologies related to the so-called Web 2.0.

BACKGROUND

Usually, a social network is defined as a *small-world network* (Watts, 2003), consisting of a set of individuals (persons, groups, organizations) connected by personal, work, or trust relationships. Social networking is then a quite broad and generic notion, which in the Web context might be applied to any kind of virtual community. For instance, users registered to a Web service, such as Web mail, online journals, or newspapers requiring a subscription, can be considered as a social network. In the following, we adopt the definition provided by Golbeck (2005), according to which an online community's Web site can be considered a Web-based social network only if it satisfies the following conditions:

- Relationships are explicitly specified by its members, and not inferred from existing interactions (e.g., a mailing list can be used to infer implicit relationships).
- Relationships are stored and managed by using technologies, such as database management systems, allowing relationship analysis and regulating access and retrieval of relationship data.
- Members are able to access relationship information, at least partially.

Born in the late 1990s, in the last few years WBSNs gained increasing interest and diffusion. Although the first and most successful ones, such as MySpace, Friendster, and Facebook, were formerly designed for entertainment and socialization purposes, they are currently establishing themselves as a business model, through which institutions and organizations can set up a collaborative environment for specific purposes, and where it is possible to share resources at an intra- and inter-organizational level. Due to the great amount of collected data, WBSNs are currently the subject of great interest for statistical analysis (Wasserman & Faust, 1994; Freeman, 2004), since they may provide useful information not only to social researchers, but also for marketing purposes.

WBSNs may provide different kinds of services, ranging from information and contact sharing, to collaborative

rating, collaborative work environments, and so on. However, independently from the specific purposes of a WBSN, members' relationships are the core information on which all the provided services are based. In fact, they can be used not only to create connections among people sharing similar interests, but also to customize WBSN services themselves. This is particularly true in WBSNs supporting collaborative rating: in such a context, ratings may be given different weights, depending on the relationships existing between WBSN members. For instance, it may be the case that a given WBSN member m_1 considers more relevant (or trustworthy) the opinions of member m_2 than, say, those of member m_3. For this purpose, some WBSNs allow their members not only to specify personal relationships (e.g., "friend of," "colleague of") but also to establish *trust* relationships, which express how much they trust the other members either with respect to a specific topic (*topical trust*) or in general (*absolute trust*). For a thorough discussion on trust relationships and how they can be used, we refer the reader to the work by Golbeck and Hendler (2006).

As far as security is concerned, current WBSNs enforce simple protection mechanisms, which only allow their members to label given information as public or private, or to make it available to WBSN members with whom there exists a direct relationship of a given type (friend, colleague, etc.). However, these solutions on one hand may dramatically reduce the possibility of sharing information, which is the basic function of a WBSN, and on the other hand, they do not necessarily grant the required protection to personal information. In fact, giving to WBSN members just the choice of stating whether a given resource is public or private may result in hiding a huge amount of information. Moreover, it may frequently happen that WBSN members make publicly available resources that are accessed by people different from the ones they intended—the most typical case is a student publishing photos or blogs in recreational WBSNs, without considering that they can be accessed by his or her teachers.

Additionally, personal information and relationships among WBSN members must be protected when WBSN data are analyzed by data mining tools, that is, tools capable of analyzing massive datasets of personal information with the purpose of extracting models of social and commercial interest.

SECURITY AND PRIVACY REQUIREMENTS IN SOCIAL NETWORKS

In this section we consider the security and privacy issues related to WBSNs from two different points of view. First, we discuss the privacy-preserving techniques adopted to allow statistical analysis on social network data without compromising WBSN members' privacy, and then we illustrate the current approaches aimed at enforcing privacy protection when performing access control.

Privacy-Preserving Social Network Analysis

Data collected by WBSNs are an important source for social and marketing analysis, which may provide useful information on the evolution of a social community, collaborative problem solving, information distribution, and so on. Additionally, they can also be used to optimize social network services and customize them with respect to users' preferences and interests. However, when analyzing WBSN data for statistical purposes, it is necessary to avoid as much as possible disclosing private information about WBSN members.

So far, this issue has been addressed by anonymizing the network graph according to two main strategies, namely, *node anonymization* and *edge perturbation.* The former strategy aims at hiding members' identities by labeling the corresponding network nodes with random identifiers (naïve anonymization). In case nodes are associated with attributes which can be used to identify the corresponding user, the possibility of using techniques based on k-anonymity (Sweeney, 2002) has been discussed−see, e.g., Zheleva and Getoor (2007). By contrast, edge perturbation performs a set of random edge deletions and insertions, which prevent an attacker from inferring the identity of network nodes based on the existing relationships but, at the same time, preserve the utility of the graph for network analysis.

It has been noticed that the proposed solutions to node anonymization do not grant total privacy protection. In particular, Backstrom, Dwork, and Kleinberg (2007) carried out an extensive analysis of the possible attacks, and argued that the most effective strategies for privacy protection are those based on *interactive* techniques. According to this approach, the anonymized network graph is not disclosed; rather it is analyzed by the social network management system itself upon submission of a query, and then the result is perturbed by adding noise to the real answer.

By contrast, edge perturbation, when combined with node anonymization, grants a greater degree of protection. Examples of how such techniques are applied are provided by Frikken and Golle (2006), Hay, Miklau, Jensen, Weis, and Srivastava (2007), and Zheleva and Getoor (2007). In particular, Hay et al. (2007) report experimental results which show that random edge deletions and insertions grant graph anonymity when the perturbation affects a percentage of graph edges ranging from 5% to 10%. By contrast, a perturbation rate greater than 10% dramatically increases information loss, thus making useless the results obtained by

analyzing the perturbed graph. Although Zheleva and Getoor (2007) do not provide experimental results, they enhance the edge perturbation strategy by considering the different possible methods according to which it can be performed, and by evaluating the obtained perturbed graph with respect to information loss and link re-identification attacks.

Note, however, that graph anonymization is based on the assumption that the only information that can be obtained by an attacker is the one publicly released by the social network service. By contrast, this strategy is useless when applied to social networks, as most WBSNs are, to which any Web user can register, and where each member has a total or partial view of the network graph. In such a case, attackers can infer the network structure and members' identity with more or less accuracy by using techniques like *node bribing*, that is, by obtaining access to the partial view of the WBSN graph of one or more of its members, as illustrated by Korolova, Motwani, Nabar, and Xu (2008). The authors argue that it is possible to reduce the effectiveness of such attacks by limiting the neighborhood visibility of a member (his or her *lookahead* ℓ) to his or her neighbors ($\ell = 0$), and to the neighbors of his or her neighbors ($\ell = 1$). By contrast, in case $\ell > 1$, the possibility of obtaining correct information on the WBSN graph exponentially increases.

In conclusion, available privacy-preserving techniques, both those based on graph anonymization and those limiting WBSN members' lookahead, have the goal of preserving users' privacy when network data are analyzed through data mining tools. An additional issue is to enable a WBSN user to state which information should be public or private, and which members are authorized to access it. In this respect, current WBSNs enforce very naïve default protection mechanisms which cannot be personalized by WBSN members. We elaborate more on this issue in the next section.

Privacy-Aware Access Control

WBSN resources have protection requirements that cannot be enforced by simple mechanisms, as those currently adopted by WBSNs. An access control model for WBSNs should therefore take into account the specific characteristics of the application domain, in order to devise the most suitable access control strategies. In the following, we first discuss the main requirements for an access control mechanism tailored to WBSNs. Then, we survey the solutions proposed so far.

According to the traditional approach, access control requirements are expressed by *authorizations*, which in their basic representation are tuples of the form $\langle s, p, o \rangle$, where s is the subject authorized to access object o under privilege p (Bertino & Sandhu, 2005). However, such an approach is not suitable for dynamic and distributed environments, as WBSNs are, since a member may be required to update the authorizations applying to his or her resources whenever he or she knows new members, or if relationships he or she participates in are revoked. In such a scenario, it is preferable to *intensionally* denote authorized members by specifying the *requirements* they must satisfy to access a given resource. According to this strategy, whenever any modification to the state of the WBSN structure occurs, the set of authorized members will dynamically change, without the need to modify the existing authorizations.

So far, a variety of access control models have been proposed, which denote authorized users in terms of their characteristics, and not only by their identities. The role-based model (Ferraiolo, Kuhn, & Chandramouli, 2003) is the most popular one; others are those based on credentials (e.g., Winslett, Ching, Jones, & Slepchin, 1997; Agarwal, Sprick, & Wortmann, 2004) or certificates (e.g., Thompson et al., 1999; Palomar, Estevez-Tapiador, Hernandez-Castro, & Ribagorda, 2006). An analogous approach can be applied to WBSNs. In fact, WBSN members usually publish resources having in mind a specific audience consisting of, for example, their friends or colleagues. Therefore, in a WBSN context, *relationships* can be used to intensionally denote authorized members.

The enforcement of relationship-based access control requires addressing two main issues. First, it must be possible to verify the authenticity and reliability of information about relationships, in order to avoid security attacks based on forging faked relationships. Second, relationship information may have privacy protection requirements, and thus mechanisms should be enforced to regulate their disclosure.

A further requirement is related to the support of content-based access control (Adam, Atluri, Bertino, & Ferrari, 2002). Actually, the practice of 'tagging' resources is currently diffused among WBSN members, and content analysis is another possible solution to enforce content-based access control. However, since resource rating is performed by each single member and content analysis gives only probabilistic results about the actual content of a resource, strategies should be devised in order to obtain accurate and unambiguous descriptions, usable for access control purposes.

Finally, access permissions should take into account the possible operations to be performed on WBSN resources. Besides the traditional 'read' privilege, in collaborative environments WBSN members may be authorized to modify/delete a resource or add content to it. In such a case, it may be useful to support different types of 'write' privileges, such as 'modify' (authorized members can modify existing content or add new content), 'delete', and 'append' (authorized members can only add content, but not modify existing content). Additionally, when supporting 'write' privileges, it is important that any modification performed on a resource can be associated with the member who performed it. This means that supporting different privilege types requires enforcing accountability in the WBSN framework.

A last issue to be addressed concerns the access control architecture to be adopted. According to the traditional ap-

proach, access control is enforced on the side of the content provider. However, this solution may not be suitable to WBSNs, which may have millions of registered members and, as a consequence, the WBSN management service would be a bottleneck to the whole system.

As far as we are aware, the only two proposals of an access control mechanism based on WBSN relationships are the ones by Carminati, Ferrari, and Perego (2006) and Hart, Johnson, and Stent (2007).

In the proposal by Carminati et al. (2006), access control requirements are expressed by *access conditions,* which denote authorized members not only in terms of relationship types (e.g., friend, colleague), but also with respect to the relationship *depth* and *trust level*. The depth of a relationship corresponds to the distance between two members, considering only the edges labeled with a given relationship type. Thanks to this, it is possible to specify authorizations stating that a given resource can be accessed only by the friends of Alice, or by the friends of Alice's friends. By contrast, the trust level denotes how much confidence a member has on the fact that another given member does not reveal protected information to unauthorized members.

As far as access control enforcement is concerned, Carminati et al. (2006) adopt the rule-based approach proposed by Weitzner, Hendler, Berners-Lee, and Connolly (2006). More precisely, access authorizations are expressed by Horn-like clauses (rules), and it is the requestor who is in charge of demonstrating to the content provider of being authorized to access a given resource, by providing a proof of the corresponding access rules. WBSN resources and the corresponding access rules are managed by the resource owner, whereas relationship certificates are stored in a central directory, stored and managed by the WBSN management system. Whenever an access control request is submitted, the resource owner sends back to the requestor the set of associated access rules. The requestor then contacts the WBSN management system, in order to retrieve the relationship certificates concerning the relationships denoted by the received access rules. Then, he or she computes a proof, if any, demonstrating that he or she satisfies the rules. The resource owner sends the resource to the requestor only in case the provided proof is valid.

Also the access control model proposed by Hart et al. (2007) in their position paper uses existing WBSN relationships to denote authorized members, but only the direct relationships they participate in are considered, and the notion of trust level is not used in access authorizations. In addition, differently from Carminati et al. (2006), resources are not denoted by their identity, but based on their content. Information about resources' content is derived based on users' tags and content analysis techniques. Hart et al. (2007) do not provide any information about access control enforcement.

Both the approaches by Carminati et al. (2006) and Hart et al. (2007) assume that relationships are public. Later, Carminati et al. (2007) have extended their earlier research (Carminati et al., 2006) by proposing a privacy-aware access control mechanism, where the existing relationships are protected by a set of rules, called *distribution rules.* Such rules are used to regulate the distribution of relationship certificates to authorized members. Carminati et al. (2007) also address the issue of protecting relationship information that may be inferred by access rules, when enforcing access control. In fact, if an access rule states that, in order to be able to access a given resource, the requestor must be a friend of Alice, it is possible to infer that Alice participates in at least one relationship of type *friendOf,* otherwise no member will be able to access that resource. In order to deal with this issue, WBSN members are equipped with a set of group keys (Rafaeli & Hutchinson, 2003), called *relationship keys,* used to encrypt access conditions. More precisely, each WBSN member *m* holds a key for each type of relationship he or she participates in. These keys are shared by all the WBSN members in his or her social network group, that is, all the WBSN members connected to *m* by paths labeled with those relationship types. Whenever *m* receives an access request to a resource he or she owns, he or she does not send the corresponding access rules in plaintext. Rather, each access condition in the access rule is encrypted with the corresponding relationship key. For instance, if an access condition puts a constraint on relationships of type *friendOf, m* will encrypt it with the key corresponding to that relationship type. As a consequence, the requestor will be able to read that access condition only when he or she belongs to the same group of type *friendOf m* participates in.

It is important to note that all the approaches we have described so far support 'read' privileges only. Of course, they can be extended with other types of access modes, but enforcing accountability would require a relevant extension to the access control mechanisms described above.

FUTURE TRENDS

WBSN security and privacy is quite a new and challenging research area, and as such, the proposals discussed in this article are just a starting point. It is then difficult, given the state of the art, to provide an exhaustive summary of all the possible future trends and research directions. However, some general considerations can be done on the main open issues with respect to the topics discussed in the previous sections.

First of all, it is clear that privacy-preserving social network analysis and privacy-aware access control address WBSN privacy from two different points of view: the former, from an *external* perspective—that is, the one of an

analyst carrying on social network analysis; the latter, from an *internal* perspective—that is, the point of view of WBSN members themselves. In privacy-preserving data mining, the goal is to provide, on average, an acceptable degree of privacy (e.g., by using anonymization techniques) to all the WBSN members to whom the data refer. By contrast, in privacy-aware access control, each WBSN member can explicitly state his or her privacy and/or access control requirements—for instance, some members may have stricter privacy requirements than others. This means that potential conflicts between social network analysis tools and WBSN privacy requirements may arise. Therefore, in the future it is desirable that these two research directions find some common points, in order to proceed towards the definition of a comprehensive framework, able to address all the privacy and security requirements of WBSNs. It must also be taken into account that social network analysis is carried out based on the assumption that the social network management system is able to release periodically, or upon demand, the network graph (or a perturbed version of it). This implies that the existing relationships must be stored in a central repository, accessible by the social network management system itself. However, this is not always the case. For instance, privacy protection mechanisms enforced in a WBSN might adopt approaches according to which relationship information is stored by WBSN members themselves, to avoid that the social network management system infers private information from the existing relationships. Therefore, privacy-preserving data mining tools must also take into account the different architectures according to which access control is enforced.

As far as privacy-aware access control is concerned, we argued in the previous sections that, when adopting a relationship-based approach to specify access control requirements, it is necessary at the same time that access to relationship information is regulated by proper protection mechanisms. The strategy proposed by Carminati et al. (2007) addresses this issue, but other solutions are also possible. For instance, instead of assuming that relationship information is directly distributed by the WBSN members involved in them, as in Carminati et al. (2007), an alternative is to support negotiations and privacy policies, similar to those provided by P3P (Cranor et al., 2006) and trust negotiation mechanisms. According to this approach, relationship information is held by WBSN members and released upon request after having verified whether the requestor satisfies given privacy protection policies, and/or whether he or she can be considered trustworthy about the use he or she will make of such information and the protection he or she can assure to it.

Content-based access authorizations are one of the other open issues. By using content-based access control, it is possible to simplify the task of policy specification as well as to express access control requirements related to the semantics of the protected objects. However, applying it to distributed environments such as WBSNs, where any member can use any vocabulary and any language (either standard or user defined) for describing resources, might make such strategy ineffective for access control purposes. Using content analysis tools has similar drawbacks, since, independently of the efficiency and effectiveness of the adopted tools, it may happen that a given resource is incorrectly described, thus granting unauthorized access to it or denying access to authorized members. Finally, the trade-off between accuracy and complexity in describing resources must be taken into account. Inaccurate and ambiguous descriptions are useless for access control purposes, but evaluating too complex descriptions may have computational costs that make unfeasible, in practice, the enforcement of content-based authorizations.

We think that a solution to this issue must satisfy two main requirements. First, resource descriptions should be encoded by using standard schemes, and the vocabularies used for describing resources must enforce semantic interoperability. Second, mechanisms should be devised that are able to confirm the actual validity of a description.

As far as the former issue is concerned, a possible solution might be provided by the outcome of the work currently carried on by the W3C working group named, "Protocol for Web Description Resources" (POWDER, 2007), which aims at defining a standard metadata format for describing the content/characteristics of a group of resources. In addition, POWDER aims at granting the accountability of such descriptions, referred to as *description resources* (DRs, for short), thus making any Web user able to verify their trustworthiness. Finally, DRs provide a simple mechanism for enforcing semantic interoperability. In fact, any Web user describing a resource can state that such description, independently of how it is specified, is equivalent to one or more given DRs released by other users.

However, POWDER DRs by themselves do not ensure the trustworthiness of resource descriptions. A possible solution is to use a content analyzer to validate the description provided by a given user. However, the results of a content analyzer are reliable when applied to resources all belonging to a given content domain, which is not the case of WBSNs. An alternative is to use social networking itself in order to validate resource descriptions, by exploiting *collaborative rating*. According to this strategy, WBSN members, on one side, can express their opinions on the trustworthiness of a description, and on the other side, can specify their personal descriptions of the same resource. The result is that, for the same resource, more descriptions are available, whereas a description is associated with ratings stating whether it is trustworthy. Given the huge population of WBSNs, it is possible to collect a data set having a size suitable to perform statistical analysis, which can provide a more accurate measure of how much the claims made by a given

description can be trusted. Such an approach is currently under development in the framework of the QUATRO Plus EU project (*http://www.quatro-project.org*).

Support for different types of access privileges is another of the issues not addressed by Carminati et al. (2006, 2007) and Hart et al. (2007). As we mentioned, a key issue is the support for accountability, in order to be able to identify who performed which access operation on which resource. This is extremely important for 'write' operations, especially in collaborative environments where the members of the working group should be able to identify, for instance, who inserted/modified/deleted given portions of a shared document. Finally, it is worth noting that future WBSNs may rely on architectures different from the current one, where the WBSN management service is in charge of running almost all the supported services. In fact, from the privacy protection and access control approaches proposed by Carminati et al. (2006, 2007), it comes out that a decentralized architecture grants a more accurate protection to WBSN data. In such a scenario, WBSN members themselves store and manage their personal data, relationships, and resources, and are in charge of carrying on most of the tasks concerning relationship establishment/revocation and the enforcement of privacy and access control policies. By contrast, the WBSN management system provides just basic services, such as user registration, and may be used as a common space from which it is possible to access all the information WBSN members wish to share publicly. Such decentralized architectures pose challenging research issues with respect to security and privacy protection as well as efficiency.

CONCLUSION

With the increasing diffusion and usage of online social networks, protecting personal data and resources of their members is becoming a fundamental issue. Contributions to this research area are currently very limited, and can be grouped into two main classes: on one side, anonymization techniques able to protect the privacy of social network members when performing social network statistical analysis, and on the other side, privacy-aware access control mechanisms, making social network members able to regulate access to their data, relationships, and resources by, at the same time, protecting the privacy of their relationships. The proposed solutions are far from addressing all the privacy and security requirements of social networks, and not all the potential approaches have been investigated. Although it is difficult to predict with enough precision the possible evolution of this new research area, it is very likely that the enforcement of security and privacy mechanisms for social networks, more sophisticated than the ones currently available, may have two main relevant results. First, it might lead to the development of new security paradigms able to address the distributed nature of social networks. Moreover, it might determine a dramatic modification of current online social networks into a decentralized architecture, where the management of social network information, and of the social network itself, will be carried out collectively by its members inside a collaborative framework.

REFERENCES

Adam, N.R., Atluri, V., Bertino, E., & Ferrari, E. (2002). A content-based authorization model for digital libraries. *IEEE Transactions on Knowledge and Data Engineering, 14*(2), 296-315.

Agarwal, S., Sprick, B., & Wortmann, S. (2004). Credential-based access control for Semantic Web services. *Proceedings of the AAAI Spring Symposium on Semantic Web services.* Retrieved from http://www.aifb.uni-karlsruhe.de/WBS/sag/papers/Agarwal_Sprick_Wortmann-CredentialBasedAccessControlForSemanticWebServices-AAAI_SS_SWS-04.pdf

Backstrom, L., Dwork, C., & Kleinberg, J. (2007). Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. *Proceedings of the 2007 World Wide Web Conference.*

Barnes, S.B. (2006, September). A privacy paradox: Social networking in the United States. *First Monday, 11*(9). Retrieved from *http://www.firstmonday.org/issues/issue11 9/barnes*

Bertino, E., & Sandhu, R. (2005). Database security—concepts, approaches, and challenges. *IEEE Transactions on Dependable and Secure Computing, 2*(1), 2-19.

Brickley, D., & Miller, L. (2005, July). *FOAF vocabulary specification* (RDF vocabulary specification). Retrieved from http://xmlns.com/foaf/0.1

Carminati, B., Ferrari, E., & Perego, A. (2006). Rule-based access control for social networks. *Proceedings of the OTM 2006 Workshops* (pp. 1734-1744). Berlin: Springer-Verlag.

Carminati, B., Ferrari, E., & Perego, A. (2007). Private relationships in social networks. *Proceedings of the ICDE 2007 Workshops* (pp. 163-171). IEEE CS Press.

Cranor, L., Dobbs, B., Egelman, S., Hogben, G., Humphrey, J., Langheinrich, M. et al. (2006, November). *The platform for privacy preferences 1.1 (P3P1.1) specification* (W3C working group note). *Proceedings of the World Wide Web Consortium.* Retrieved from http://www.w3.org/TR/P3P11

Davis, I., & Vitiello, E., Jr. (2005, August). *RELATIONSHIP: A vocabulary for describing relationships between people*

(RDF vocabulary specification). Retrieved from http://purl.org/vocab/relationship

Ding, L., Zhou, L., Finin, T., & Joshi, A. (2005). How the Semantic Web is being used: An analysis of FOAF documents. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences* (HICSS'05) (p. 113.3). IEEE CS Press.

Ferraiolo, D.F., Kuhn, D.R., & Chandramouli, R. (Eds.). (2003). *Role-based access control.* Norwood MA: Artech House.

Freeman, L.C. (2004). *The development of social network analysis: A study in the sociology of science.* BookSurge.

Frikken, K.B., & Golle, P. (2006). Private social network analysis: How to assemble pieces of a graph privately. *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society* (WPES 2006) (pp. 89-98).

Golbeck, J.A. (2004). *The Trust ontology* (OWL vocabulary). Retrieved from http://trust.mindswap.org/ont/trust.owl

Golbeck, J.A. (2005). *Computing and applying trust in Web-based social networks.* Unpublished Doctoral Dissertation, University of Maryland, USA. Retrieved from http://trust.mindswap.org/papers/GolbeckDissertation.pdf

Golbeck, J.A., & Hendler, J. (2006). Inferring binary trust relationships in Web-based social networks. *ACM Transactions on Internet Technology, 6*(4), 497-529.

Hart, M., Johnson, R., & Stent, A. (2007). More content—less control: Access control in the Web 2.0. *Proceedings of the Web 2.0 Security & Privacy 2007 Workshop.* Retrieved from http://seclab.cs.rice.edu/w2sp/2007/papers/paper-193-z6706.pdf

Hay, M., Miklau, G., Jensen, D., Weis, P., & Srivastava, S. (2007, March). *Anonymizing social networks.* Technical Report No. 07-19, University of Massachusetts Amherst, USA. Retrieved from http://www.cs.umass.edu/~mhay/papers/hay-et-al-tr0719.pdf

Korolova, A., Motwani, R., Nabar, S.U., & Xu, Y. (2008). Link privacy in social networks. *Proceedings of the 24th International Conference on Data Engineering* (ICDE'08).

Palomar, E., Estevez-Tapiador, J.M., Hernandez-Castro, J.C., & Ribagorda, A. (2006). Certificate-based access control in pure P2P networks. *Proceedings of the 6th IEEE International Conference on Peer-to-Peer Computing* (P2P'06) (pp. 177-184). IEEE CS Press.

POWDER. (2007). *Protocol for Web Description Resources working group.* Retrieved from http://www.w3.org/2007/powder

Rafaeli, S., & Hutchinson, D. (2003). A survey of key management for secure group communication. *ACM Computing Surveys, 35*(3), 309-329.

Staab, S., Domingos, P., Mika, P., Golbeck, J., Ding, L., Finin, T. W. et al. (2005). Social networks applied. *IEEE Intelligent Systems, 20*(1), 80-93.

Sweeney, L. (2002). *k*-anonymity: A model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), 557-570.

Thompson, M., Johnston, W., Mudumbai, S., Hoo, G., Jackson, K., & Essiari, A. (1999). Certificate-based access control for widely distributed resources. *Proceedings of the 8th USENIX Security Symposium.* Retrieved from http://dsd.lbl.gov/~mrt/papers/AkentiUsenixSec.pdf

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications* (vol. 8). Cambridge: Cambridge University Press.

Watts, D.J. (2003). *Small worlds: The dynamics of networks between order and randomness.* Princeton, NJ: Princeton University Press.

Weitzner, D.J., Hendler, J., Berners-Lee, T., & Connolly, D. (2006). Creating a policy-aware Web: Discretionary, rule-based access for the World Wide Web. In E. Ferrari & B. Thuraisingham (Eds.), *Web and information security* (pp. 1-31). Hershey, PA: Idea Group.

Winslett, M., Ching, N., Jones, V.E., & Slepchin, I. (1997). Using digital credentials on the World Wide Web. *Journal of Computer Security, 5*(3), 255-266.

Zheleva, E., & Getoor, L. (2007). Preserving the privacy of sensitive relationships in graph data. *Proceedings of the 1st ACM SIGKDD International Workshop on Privacy, Security, and Trust in KDD* (PinKDD 2007). Retrieved from http://www-kdd.isti.cnr.it/pinkdd07/Zheleva PinKDD07.pdf

KEY TERMS

Edge Perturbation: Graph anonymization technique aimed at hiding the actual social network relationships by performing a set of random edge deletions/insertions in the network graph.

Graph Anonymization: Technique aimed at hiding private information about social network members when performing social network analysis. Node anonymization and edge perturbation are the two main graph anonymization techniques currently used.

Node Anonymization: Graph anonymization technique aimed at hiding social network members' identities by labeling the corresponding nodes with random identifiers (naïve anonymization), or, in case nodes are associated with attributes which can be used to identify the corresponding user, by using techniques based on k-anonymity (Sweeney, 2002).

Privacy-Aware Access Control: In the context of social networks, denotes an access control paradigm where access control requirements of social network members are enforced without disclosing private information about the relationships they participate in.

Relationship Trust Level: In a social network, denotes the value associated with a trust relationship, providing a measure of how much a given member considers another member trustworthy. Depending on the purpose for which it is used, this notion may have different meanings. For instance, in a collaborative rating environment, it denotes how much a given member trusts the opinions of another member with respect to a specific topic (*topical trust*) or in general (*absolute trust*) (Golbeck & Hendler, 2006). By contrast, in an access control context, it has some similarities to the notion of *security level* used in mandatory access control models (Carminati et al., 2006, 2007).

Relationship-Based Access Control: An access control paradigm specifically tailored to social networks, according to which social network members authorized to access a given resource are denoted in terms of the relationships they must participate in to get the access.

Social Network: A *small-world network* (Watts, 2003) consisting of a set of individuals (persons, groups, organization) connected by personal, work, or trust relationships. Usually modeled as a graph, where nodes correspond to social network members, whereas edges denote the relationships existing between them.

Social Network Analysis: A discipline aimed at collecting statistical data from the analysis of social network topology (Wasserman & Faust, 1994; Freeman, 2004).

Social Network Relationship: A relationship concerning two members of a social network. In WBSNs, besides personal/work relationships (e.g., friend/colleague), also trust relationships may be supported which denote how much a one member trusts another. In the graph representation of a social network, relationships are usually denoted by edges, labeled with a relationship type and/or a relationship trust level.

Web-Based Social Network: A Web-based system that allows its registered members to establish relationships with other members and to share different types of information (e.g., personal data, contacts, multimedia resources). A precise, but not normative definition of Web-based social network has been provided by Golbeck (2005).

Security and Reliability of RFID Technology in Supply Chain Management

Vladimír Modrák
Technical University of Košice, Slovakia

Peter Knuth
Technical University of Košice, Slovakia

INTRODUCTION

RFID (radio frequency identification) technology can be expressed in the most universal manner as wireless identification technology, which does not need the line-of-sight to be read or written. It offers enhancement of identification technologies like barcode technology. Optical barcode technology was developed in 1948 by Silver and Woodland at Drexel Institute of Technology and first commercially used in 1966 (Adams, 2002). Barcode technology stores data in the widths and spacings of printed parallel lines, or in patterns of dots, concentric circles, and hidden within images. The most extended is UPC code which was invented in 1973 and since then became everyday part of our life. Other commonly used types of barcodes are Code 128, Code 93 (Groover, 1980) and DataMatrix 2D barcode. At this time, mostly the barcodes are keeping inventory and shipments moving. RFID and barcode technology complement each other because both of them are beneficial in different situations and can be used together in many applications.

RFID technology has several advantages for managing and collecting object's data or tracking it as it moves through the supply chain (SC). Two of them are related to the increased abilities of security and reliability of the identification systems. These two properties of identification technologies are equally important for their use in supply chain management (SCM).

The purpose of this chapter is to highlight selected areas of this technology that may be critical specific aspects of further RFID development and applications. We have also discussed about differences between RFID and barcode technologies especially in terms of their use in SCM and concluded this article with expectations of further development of this still progressive technology.

BACKGROUND

Security and reliability issues have their roots in history of RFID technology (RFID Journal, 2005). Principles of RFID technology are based on the fundamentals of electromagnetic energy, radio broadcast technology and radar technology. The first active identification friend or foe (IFF) system was developed by the British during World War II. It was radio frequency identification technology for identification of friendly aircrafts. Each plane was equipped with a transmitter, which began to broadcast signal back after receiving signals from radar stations on the ground (Landt, 2001). System was very simple and was not very secure. This is the point where not only reliability, but also security becomes significant issue. The reasons why security and reliability of RFID technology became important are at most actual and are implicitly involved in numerous polemics about security and reliability of RFID technology (Karygiannis, Eydt, Barber, et al., 2007; Rieback, Crispo, & Tannenbaum, 2006a; Thorton et al., 2006; Wyld, 2005; Rieback, Crispo, Tanenbaum, 2006b; Macaulay, Abeysinghe, 2004; Bono, 2005). Explicitly we can see the reasons in existing precedence about potential serious impacts that are mentioned later in this chapter. Another important milestone can be considered the patent on RFID that was granted to Mario Cardullo in 1973. In the same year Charles Walton invented access control system based on RFID (Rieback et al., 2006a). The first widespread RFID tag was 1-bit tag (the bit is either on or off) as a part of electronic article surveillance (EAS). EAS could only detect the presence or absence of the tag. In other words if someone does not pay, the tag remains on and the readers at the door detects the tag and the alarm sounds to alert un authorized removal. The U.S. government was also supporting research and development of RFID systems. Los Alamos National Laboratory in New Mexico soon became a leading center for R&D of this technology (Shepard, 2005). This laboratory developed automated toll payment systems and passive RFID tag to track cows. Both are still used all around the world. Further development lead to use higher frequencies

that allowed greater and faster data transfer rates. In 1999 Uniform Code Councile, EAN International, Procter and Gamble, and Gillette founded an Auto-ID Center project at Massachusetts Institute of Technology for development of RFID standards. The main result of this project was electronic product code (EPC). Other results are air interface protocols (Class 1 and Class 0) and network architecture scheme, which links objects to the Internet through the tag. After that Uniform Code Council and EAN International created joint venture EPCglobal Inc. to commercialize EPC technology due to high importance of EPC technology, since it could dramatically improve efficiencies within supply chain. Previous research responsibilities of Auto-ID center were delegate to EPC global. Recently, more attention is given to security of RFID technology. To corroborate it by facts the following events can be mentioned. In January, 2005, students at Johns Hopkins University broke encryption of SpeedPass electronic payment and RFID point of sale (POS) system. In February, 2006, Adi Shamir reported that he could monitor power levels in an RFID tag which can be used to compromise the secure hashing algorithm 1 (SHA-1) used in some RFID tags (Thorton et al., 2006). However, it is not the reason for a resignation, as a level of risks depends in generally, but also in the specific area, on preventive actions. The example supporting this statement offers the situation in privacy protection. In contrast to barcode technology, RFID technology has greater implications on individuals' privacy, because RFID tags used in personal identification cards can be read from an abundant distance without that person's knowledge or consent. This led to creation of groups like FoeBud or CASPIAN that are against this technology, because they fear, that they could be tracked by tags. In the meantime, blockers for passport RFID tags in a form of passport jackets containing physical barrier and other countermeasures as unique identifier numbers, encryption, and mutual authentification were developed to ensure greater security. On the other hand, the more sophisticated protections bring more opportunities for potential failures.

COMPARISON OF SECURITY AND RELIABILITY BETWEEN THE RFID AND BARCODE TECHNOLOGY

The main difference between barcode technology and RFID technology is that barcode technology is optical technology and RFID technology is radio technology. All other advantages, disadvantages and differences result from this fact. In the supply chain, the biggest advantage that RFID has over barcode is the ability to automatically read large groups of tags eliminating the labor needed to manually scan the large volumes involved in the supply chain. Improving visibility in the supply chain systems gives "management programs better visibility into the supply chain, which enables identification of bottlenecks, targeted recalls, and new forms of market research" (Karygiannis et al., 2007). "Both active and passive RFID tags have significant potential to provide low-cost, short-range, identification for many consumer goods and can help to identify objects" (Finkenzeller, 2003). Potential benefits of RFID implementation in the supply chain management are counterfeit and fraud reduction, improved efficiency, labor cost reduction, stock shrinkage reduction, stocking management improvements and return goods facilitation (improved customer satisfaction). RFID technology can be used in several levels in supply chain management (see Table 1).

When talking about reliability of automatic identification technology, an attention might be focused on the ability of readers to identify codes from tags at the first time. The potential interferences of barcodes make optical barriers such as objects placed between barcode and reader or dirt. Also, they are unreadable under extreme atmospheric conditions such as steam or when vertical damage occurs. Barcode readers are sensitive to dirt, dust, or other foreign object obstructing the lens. But 2D barcodes can be read even if part of the tag is destroyed. Passive RFID tags can interference with environments or fields and various materials such as liquids and metals that affect transmission of radio frequency. Active tags are less susceptible to interference. Despite this, they can be read under extremer weather conditions than barcodes. It is not clear whether tags, that could not be read, can be entered manually as barcodes. Reliability seems to be solved these days by knowing RFID physics (Schlosser, 2004). But there is no universal solution for implementation of RFID at all. It is always necessary to fit RFID system to meet company needs (if company really needs RFID) as far as 100 percent reliability. And because there are numerous types of RFID tags the selection of proper RFID tag system is essential. It is better to start with smaller project and obviously in detail defined problem rather than to fail (Sweeney, 2005).

Security of RFID technology in supply chain management can be seen from many aspects: health (radiation of devices), personnel or vehicle access control and tracking, inventory location, privacy issues, third-person attacks, software and hardware protection, encryption and tracking origin of goods

Table 1. Levels of RFID in supply chain application (Source: D'Hont, 2003, p. 13)

Level	Use	Application
Item	Consumer units	Products and individual items
Case or Carton	Traded units	Boxes (packaging) product carriers
Pallet	Distribution units	Pallets / Trucks

Table 2. RFID applications (Source: Wyld, 2005, p. 13)

Traditional RFID Applications						
Security/access control	Electronic article surveillance	Asset/fleet management	Mass transit	Library access	Toll collection	Animal identification
Emerging RFID Applications						
Warehouse management	Supply chain management	Reverse logistics	Shipment tracking	Asset tracking	Retail management	Document tracking	Anti-counterfeit
Advance access control	Mass transit —monthly and single trip	Airline baggage handling	Aircraft parts and tools	Healthcare applications	Regulatory compliance	Payments	

(very important for food safety). Data security of RFID technology depends on the class and generation of RFID tag. From the health aspect, there are no concerns about the risks. One can never be sure, but this is scientifically proved. For example frequencies 13.56 MHz, 915 MHz and 2.45 GHz have been used for many years without any known problems if levels are below 1 watt or 4 watt at frequency 13.56 MHz (Wyld, 2005).

The biggest security risks are related with privacy concerns, with ability of RFID system to identify, track and monitor each good uniquely and link it with the owner. It is possible to automatically track individuals or even to be a victim of robber just because you are carrying expensive product and the theft have read the tag of that product. U.S. based group CASPIAN worries about people-tracking with the help of tagged goods after they leave the store.

RFID security threads can be classified as sniffing, tracking, spoofing, denial of service or relay attacks. In the field of cryptography new low-power algorithms like stream ciphers (Finkenzeller, 2003), block ciphers (Feldhofer, Dominikus, & Wolkerstorfer, 2004), lightweight protocols for authentication (Vajda & Buttyán, 2003) and public key cryptographic primitives have been created. Other techniques like trusted RFID readers or access control mechanism that are located either on a tag like hash locks (Weis, Sarma, Rivest, & Engels, 2004) pseudonyms (Juels, 2004) or off the tag can prevent unauthorized threats and attacks too. Off the tag RFID access control mechanisms are RFID Guardian (Rieback, Gaydadjiev, Crispo, et al., 2006), RFID Enhancer Proxy (Juels, Syverson, & Bailey, 2005), The Blocker Tag (Juels, Rivest, & Szydlo, 2003) and FoeBud Data Privatizer. The easiest way still remains deactivating RFID tag permanently through "frying," "clipping"(Karjoth & Moskowitz, 2005) or "killing," or temporarily using sleep/wake modes (Spiekerman & Berthold, 2004) or Faraday cage.

FUTURE TRENDS

It is expected that modern technologies like proxies and firewalls will be adopted through central monitoring and managing of the communication environment. Also, development of privacy enhancing technologies is very important for future RFID implementation and acceptance of public. For example in the future mobile RFID technology will allow consumers to scan particular tag attached to an item using mobile phone and then connect to the manufacturer's EPCIS (electronic product code information service) accessing application in order to verify if product is genuine or fake (Konidala & Kim, 2006). RFID can also be used to trace the location of soldiers in a battle or to sense certain parameters as glucose level when implanted in human body. RFID tags with identification information implanted under the skin can soon replace credit cards and enable non-contact reading of bank account. The European Union is thinking to implement RFID tags in the Euro currency.

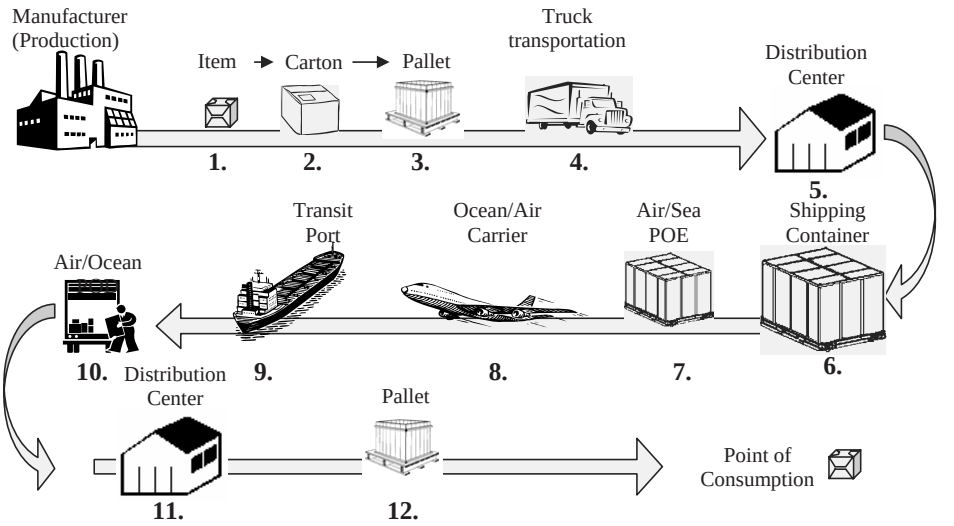
The RFID technology is here for a very long time, but with new research and inventions are coming into existence also new ideas and fields of use (see Table 2).

Likely we can expect that the barcode will be considerably replaced by the RFID. Due to unpredictable changes in corporate environment, the horizon of the milestone is question of the guesswork. In opposite case of predictable corporate environment, the business would not be so complicated, complex and fascinating.

As reliable indicator of the future trends advanced ideas and solutions that present challenge for new development programmes of companies can be considered. An example of such advanced idea is the supply chain model based only on the RFID technology connected to a global positioning system depicted on Figure 1.

Because entire supply chains present multimodal transport systems, number of transloadings in the whole supply chain is much greater than in unimodal transport. Then, the RFID plays important role in entire SC by eliminating human

Figure 1. Future supply chain model (Adopted from Wagner, 2006)



1.	*Passive tag labeling*	7.	*Tag reads at the gate; manifest for shipment*
2.	*Passive tag labeling*	8.	*Carrier track ship/airplane, provides visibility*
3.	*Active tag labeling*	9.	*Tag reads when discharged or reloaded*
4.	*Tag reads when truck departs*	10.	*Tag reads at discharge and at truck departure*
5.	*Automatic capture of transaction*	11.	*Tag reads and content reconfiguration*
6.	*Active tag with manifest data labeling*	12.	*Breaking pallets into individual items*

labor, which is needed when using barcodes. Furthermore, RFID systems reduce operator costs during discharge and multiply accelerates acquisition of item data into information system.

CONCLUSION

Questions about reliability and security issues of RFID technology are very important not only for the future development of RFID technology, but for all of us. It is obvious that RFID technology is not taken just with a great interest, but it is also facing rejection and concerns. Despite this, development of this technology will continue with the aim to reveal all potentials of RFID and not only in supply chain management. Over the next few years, it is theorized that decreasing price of tags, return on investment and better knowledge of this technology will allow wider spread of RFID technology in other business and civic areas.

REFERENCES

Adams, R. (2002). Barcode 1: Barcode History page. Retrieved from http://www.adams1.com/pub/russadam/history.html

Auto-ID Center (2003). 13.56 MHz ISM band class 1 radio frequency identification tag interface specification: recommended standard, Version 1.0.0, Technical Report.

Bono, S. (2005). Security Analysis of a Cryptographically-Enabled RFID Device. *Proceedings 14th USENIX Security Symposium, USENIX,* (pp. 1–15).

D'Hont, S. (2003). *The Cutting Edge of RFID Technology and Applications for Manufacturing and Distribution: A White Paper from Texas Instruments.*

Feldhofer, M., Dominikus, S., & Wolkerstorfer, J. (2004). Strong authentication for RFID systems using the AES algorithm. *Workshop on Cryptographic Hardware and Embedded Systems, LNCS,* Vol. 3156, (pp. 357–370).

Finkenzeller, K. (2003). RFID-handbook, fundamental and applications in contactless smart cards and identification, (2nd ed.). Swadlincote UK: Wiley & Sons Ltd.

Groover, M.P. (1980). *Automation, production, systems and computer-integrated manufacturing.* New Jersey: Prentice-Hall.

Juels, A. (2004). Minimalist cryptography for low-cost RFID tags. *In proceedings of the 4th International Conf. on Security in Communication Networks, LNCS*, Springer-Verlag.

Juels, A., Rivest, R., Szydlo, M. (2003). The blocker tag: Selective blocking of RFID tags for consumer privacy. *In Proceedings of the ACM Conference on Computer and Communications Security,* (pp. 103–111). New York: ACM Press.

Juels, A., Syverson, P., & Bailey, D. (2005). Highpower proxies for enhancing RFID privacy and utility. *Proceedings of the 5th Workshop on Privacy Enhancing Technologies*.

Karjoth, G., & Moskowitz, P. (2005). Disabling RFID tags with visible confirmation: Clipped tags are silenced. *Workshop on Privacy in the Electronic Society.*

Karygiannis, T., Eydt, B., Barber, G., Bunn, L., & Phillips, T. (2007). *Guidelines for securing radio frequency identification (RFID) Systems*. *Natl. Inst. Stand. Technol. Spec. Publ.* 800-98, 154 pages (April 2007)

Konidala, D.M., & Kim, K. (2006). *RFID tag-reader mutual authentication scheme utilizing tag's access password.* Auto-ID Labs White Paper

Landt, J. (2001). *Shrouds of Time: The history of RFID*. White Paper from AIM, Inc.

Macaulay, D., & Abeysinghe, G. (2004, April 6-7). Radio frequency identification: Putting price on user privacy. *In Proceedings IWWST*. London, UK.

Rieback M. R., Crispo, B., & Tanenbaum, A.S (2006a). The evolution of RFID security. Pervasive computing. *IEEE, Jauary.-March 2006, 5*(1), 62- 69.

Rieback, M.R., Crispo, B., & Tanenbaum, A. S. (2006b). Keep on Blockin' in the Free World: Personal Access Control for Low-Cost RFID Tags. *Proceedings 13th International Workshop Security Protocols*, Springer

Rieback, M.R., Gaydadjiev, G.N., Crispo, B., Hofman, R.F.H., & Tanenbaum, A.S. (2006). A Platform for RFID Security and Privacy Administration. *In proceedings of the 20th USENIX/SAGE Large Installation System Administration conference*, Washington DC.

Schlosser, C. (2004). Physics can solve your RFID puzzle. *RFID Journal.* Retrieved from http://www.rfidjournal.com/article/articleview/1118/1/82

Shepard, S. (2005). *RFID: Radio frequency identification.* New York, McGraw-Hill.

Spiekermann, S., & Berthold, O. (2004). Maintaining privacy in RFID enabled environments – proposal for a disable-model. *Workshop on Security and Privacy, Conf. on Pervasive Computing, April 2004.*

Stockman, H. (1948). Communication by Means of Reflected Power. *Proceedings of the IRE Institute of Radio Engineers*, (pp. 1196–1204).

Sweeney, P. J. (2005). *RFID for dummies.* Indianapolis, IN: Willey Publishing.

RFID Journal (2005).The history of TFID technology. RFID Journal. Retrieved from www. rfidjournal.com/article/articleview/1338/1/129

Thornton, F., Haines, B., Bhargava, H., Cambell, A., Kleinschmidt, J. (2006). RFID security. *Syngress Publishing*.

Vajda, I., & Buttyán, L. (2003). Lightweight authentication protocols for low-cost RFID tags. *2nd Workshop on Security in Ubiquitous Computing.*

Wagner, M.A. (2006). *Radio frequency identification (RFID)*. Industry White Paper.

Weis, S., Sarma, S., Rivest, R., & Engels, D. (2004) Security and privacy aspects of low-cost radio frequency identification systems. *Security in Pervasive Computing, LNCS, 2802*, 201–212.

Wyld, D. C. (October, 2005). RFID: The Right Frequency for Government, IBM Center of The Business of Government

KEY TERMS

Automatic identification (auto-ID): A broad term encompassing technologies used to help machines identify objects. A host of technologies fall under the automatic identification umbrella, including barcodes, biometrics, smart cards, voice recognition and RFID.

EAS (electronic article surveillance): Loss-prevention technology using passive RFID surveillance. This surveillance uses simple electronic tags that can be turned on or off. When an item is purchased at a store or checked out

EPC (electronic product code): A unique number, stored in the chip on an RFID tag, that identifies an item in the supply chain allowing for tracking of that item (EPC number). Also it is a protocol for data communication and data storage (EPC protocol).

Interrogation Zone: The area where RFID tag can be powered up and read, often between an array of antennas.

Middleware: Software for processing the streams of tag or sensor data from one or more readers and filters, aggregate and counts tag data. This process reduces amount of data before sending them to enterprise application.

Reader (also called an interrogator): A device that communicates with RFID tags. The reader has one or more antennas, which emit radio waves and receive signals back from the tag. Readers may have a digital display to relay information to the operator and may transmit data on to an organization's computer network infrastructure. Readers can be either fixed or portable, and today they are beginning to be integrated into other electronic devices, such as PDA (personal digital assistant) and cell phones, and even into objects such as pens.

Transponder: It receives and transmits radio signals at a prescribed frequency range. After receiving the signal a transponder will at the same time broadcast the signal at a different frequency.

UPC (universal product code): The barcode standard used in North America and administered by the Uniform Code Council (UCC).

Security for Electronic Commerce

Marc Pasquet
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Christophe Rosenberger
GREYC Laboratory (ENSICAEN – Université Caen Basse Normandie - CNRS), France

Félix Cuozzo
ENSICAEN, France

INTRODUCTION

E-commerce permits a dematerialized financial transaction between a customer and a merchant (Schafer, Konstan, & Riedl, 2001). It uses a complex architecture involving many aspects in computer science (security, database management) and in electronics (smartcards, tokens) (Tang, Waichee, & Veijalai, 2004). E-commerce is in a constant growth (Herrmann & Herrmann, 2004). To be used by the majority of individuals, electronic transactions must be secured to increase the confidence in the e-commerce. Security is necessary in commercial relationships for many reasons. First, the customer must be sure that the goods he/she is buying will be the expected ones, and will be well delivered at his/her address. Second, the merchant must be sure to be paid. If the customer uses banknotes or electronic payment, two or more partners are involved in that transaction: the customer's bank and the merchant's one. The two banks must be sure of the customer's identity and of the merchant's one in order to avoid banking frauds.

In the transaction process, many security systems are used to ensure the confidentiality, authentication, and integrity of exchanges. The security is guaranteed by using specific procedures and hardware. The objective of this chapter is to present how the classical security concepts are applied for an electronic payment and especially to limit the fraud.

The background section first gives a general idea of the problem generated by the electronic commerce. Second, we present briefly the public key infrastructure approach that is generally used for authentication within this context. The main thrust introduces two protocols that have been developed: SSL (secure sockets layer) and TLS (transport layer security), to create a secure channel where all transactions are encrypted by using specific architectures and algorithms. For the payment part of the transaction process, banks have been considered that SSL and TLS are not sufficiently secure. The main reason is that the cardholder is not authenticated by the issuer bank and the responsibility stays on the merchant side. Banks have so tried to implement different architectures to meet these requirements. These different methods, use of token with SET (secure electronic transaction) or a smartcard such as C-SET developed in the last fifteen years, began to converge to the 3D-secure (three domains security) protocol. These methods to secure the distant payment was adopted together by the card scheme Visa© and MasterCard©. The last, but not the least problem, concerns the distant authentication of the client by its bank, which is described in the future trends.

BACKGROUND

We first make a brief description of the e-commerce issues.

E-Commerce Description

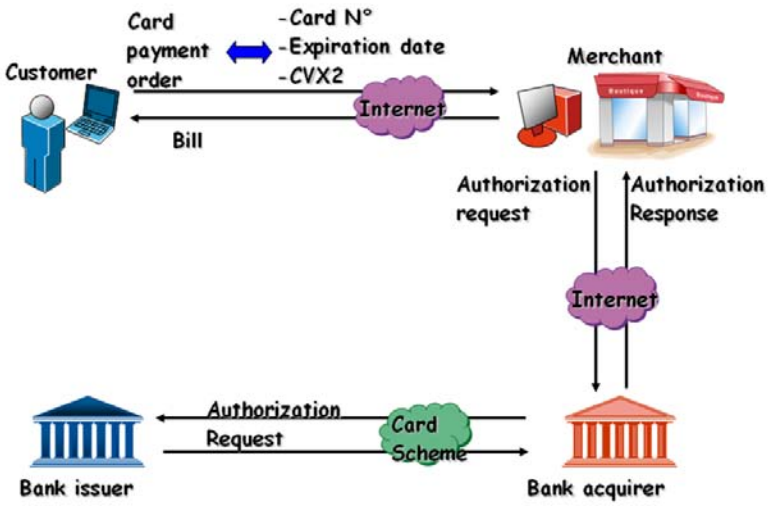
In order to better understand how the e-commerce works, Figure 1 shows the different partners and the different exchanges between them. A financial transaction between a customer and a merchant is, in fact, a transaction between the issuer and the acquirer banks. The payment is achieved through many authorization requests (customer authentication, bank transfer authorization) involving many security and cryptographic concepts.

In order to help the e-commerce development, some good practices are necessary to be applied:

- *The risk control*. The risk is partly taken by:
  - The merchant to not finally be paid;
  - The consumer to not receive the goods or the services;
  - The consumer bank in case of a systemic attack.

This risk, assumed by these different partners, must be as low as possible. The risk is as much loss of confidence in the system, as waste of money.

Figure 1. The different partners and flux in e-commerce payment



- *The facility of use for the consumer.* The reference model is the face-to-face commerce, and an ideal solution for the e-commerce must not create more constraints;
- *The use of international standards*. In one hand, Internet protocol is the base for e-commerce and in the other hand, the banking payment systems with chip or/and stripe cards, should also be used for e-payments;
- *The deployment of the different measures with a communication between banks and merchants.* The constraints and the added value must be studied with a great attention. If one of the four partners of the transaction (the customer and his/her bank and the merchant and his/her bank) is not interested in one architecture implementation, the system will have much more difficulties to be developed.

As conclusion, it is necessary to:

- Well balance the responsibilities between the four partners;
- Adapt the security level to the risk level;
- Integrate the legal constraints.

The Security Problematic

Additionally, in order to help the electronic commerce development, banks have to implement different solutions (Furnell & Karweni, 2000):

- *Visual cryptogram:* To improve the identification process, the EMV (Eurocard MasterCard Visa) cards include, on the back, a three figure code called CVX2, that the consumer must give to complete a payment transaction;
- *Authorization:* More than 7% of authorizations come from the e-commerce, and that part increases every year;
- *Individual supervision of frauds*: Coming from consumers or merchants;
- *PKI (public key infrastructure) for data protection*: To create the better possible protection for the different exchanges between all the transaction partners;
- *Authentication services*: To avoid the risk at the consumer level (CAP (chip authentication program ©MasterCard), SET, 3D-secure).

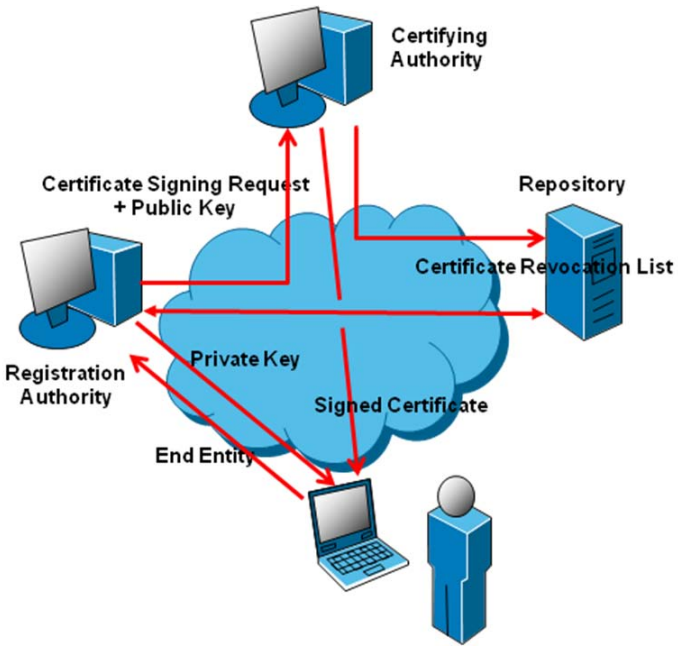
The Public Key Infrastructure

A public key infrastructure (PKI) includes a set of physical components (computers, cryptographic algorithms and equipments, smartcards), human procedures (verifications, validations), and software (systems and applications) to manage the life cycle of electronic keys or certificate. A certificate can be considered as proof of the existing relation between the identity of a customer or merchant and a public key. The major element is the certifying authority (CA) that signs the certificate, the registration authority (RA) that creates the pairs of keys, and the repository that stores the certificates.

Figure 2 shows the different transactions in a PKI infrastructure (Chanson & Cheung, 2002). There exist many certifying authorities (EuroPKI, E-certify Corporation, ID.Safe, Identrus, E-Commerce PKI CA, SwisSsign…).

A public key infrastructure delivers a set of services for its users (Critchlow & Zhang, 2004). The main services are:

Figure 2. The different parts of the PKI



- Users recording (or computers);
- Users identification and authentication;
- Pairs of keys generation (an encrypted message with a public key can be decrypted only with the corresponding private key);
- Public key certification;
- Certificate management (generation, renewal, revocation, publication, storage).

A PKI generates electronic certificates used for cryptographic operations, such as encryption and electronic signatures that offer some security guaranties for e-transactions such as:

- Confidentiality: Only the legitimate addressee of a message can read this message;
- Authentication: When a message is sent, the sender identity is perfectly known;
- Integrity: It is possible to know if a message has been deteriorated or falsified;
- Nonrepudiation: The message sender cannot deny he/she sent it.

MAIN TRUST

We focus in this part on existing solutions to secure e-payment.

SSL and TLS

The SSL security transaction process in HTTPS (hyper text transfer protocol secured) communication is detailed in Figure 3 (Guitart, Carrera, Beltran, Torres, & Ayguade, 2007; Nabi, 2005).

The pair of keys (public and private) has been received from one of the PKI registration authority. The security process continues by using the secret key AB (see Figure 4) all along the transaction. When the customer is ready to pay, he/she clicks on the corresponding option and one form asks him/her to give his/her card number, the expiration date, and the CVX2 writes on the back of the card. The merchant sends an authorization request to his/her bank and waits for an authorization response to conclude the transaction.

The SSL is limited for a payment use:

- Vulnerability to attack when keys less than 128 bits are used;
- The customer identification is not always done;
- SSL is not well protected in case of man in the middle attack. The attacker is able to receive the totality of presumably protected flow.

Very similar to the protocol SSL version 2, the TLS (transport layer security) protocol is promoted by Microsoft on its Windows browsers for the HTTPS communication. Only few differences can be pointed (Kwon, Cho, & Chae, 2001):

Figure 3. Security process. A= merchant site, B= customer PC

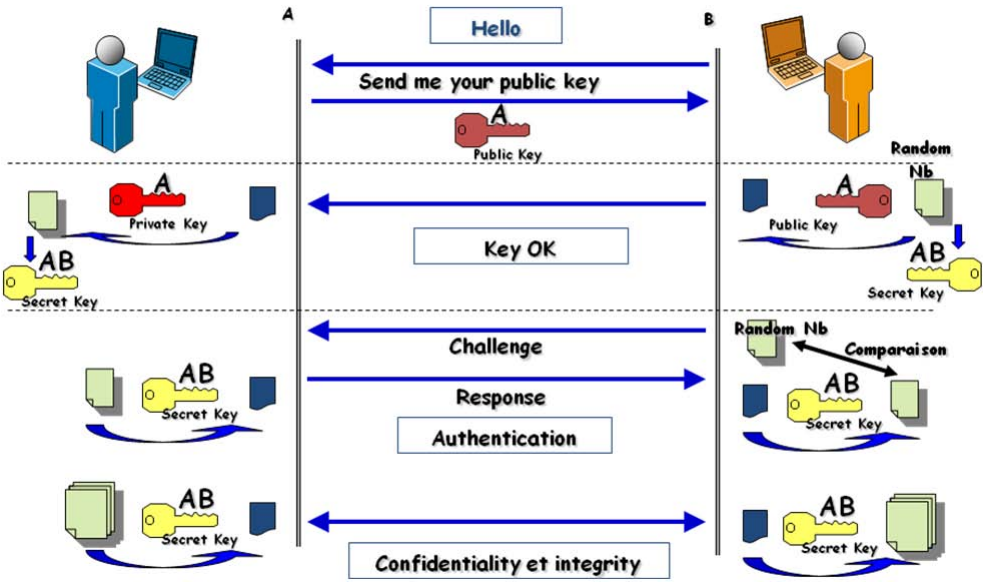
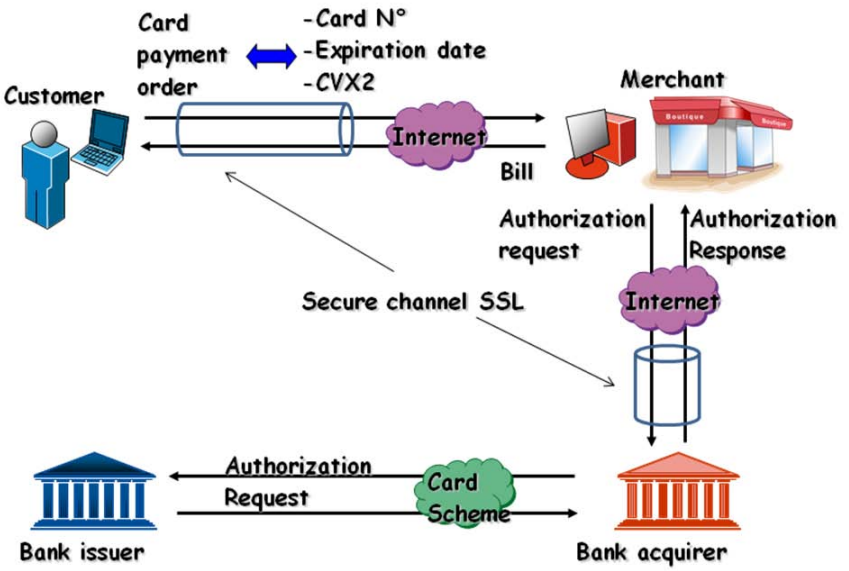


Figure 4. The different flux in SSL payment



- Encryption with the AES (advanced encryption standard) algorithm (256 bits key) instead of the DES (data encryption system) algorithm;
- More rigorous in the certificates use and less vulnerable to the man in the middle attack.

Trusted Partner and Electronic Purse

To protect the customer during the transaction, another solution is possible: use of a trusted partner that stores your banking information, debit your account, and pays for you without giving any information about your smartcard or your account to the merchant (Hawk, 2004). There are many methods, like Digicash or PayPal, that provide a digital cash implementation (see Figure 5). However, the limits of that type of payment are very quickly reached: You must be registered to the right trusted partner accepted by the merchant or having several trusted partners to pay freely on the Web.

The electronic purse is a similar solution:

Figure 5. The different flux in PayPal payment

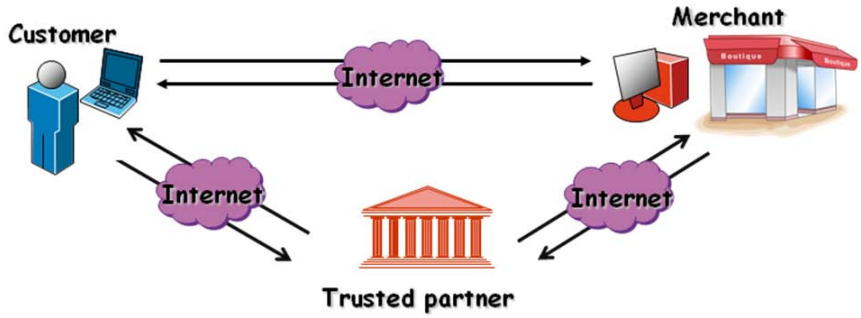
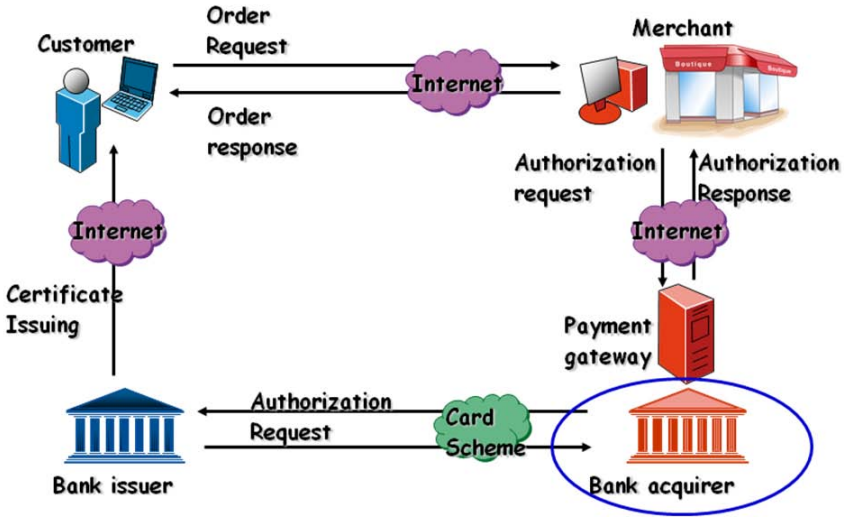


Figure 6. The different flux in SET payment



- It is a preloaded account evaluated in monetary units stored in the system of cashing of a nonbanking operator;
- The access to this electronic purse is done using software installed on the customer's PC to pay online.

SET

To limit the risk that the customer can repudiate his/her payment transaction, a set of companies (Visa, MasterCard, GTE, IBM, Microsoft, Netscape, SAIC, Terisa system, Verisign) have developed, in the eighties, one solution called SET (secure electronic transaction). The customer's bank sends him/her a certificate issued from one CA of a PKI that is stored on his/her computer. When he/she wants to make a payment on the Web, the customer must sign with the PKI keys as shown in Figure 6 (Rennhard, Rafaeli, Mathy, Plattner, & Hutchison, 2004).

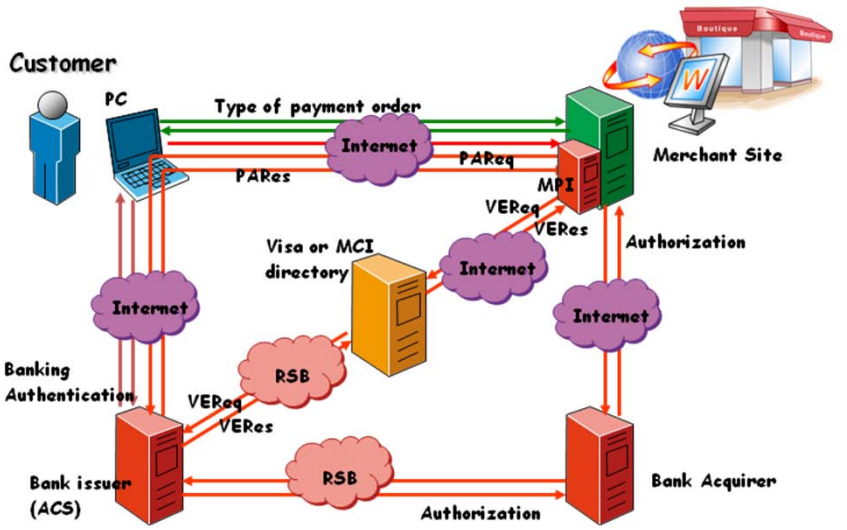
SET has not been deployed so much, but was at the origin of C-SET then to 3D secure (Visa) and SPA UCAF (MasterCard), and finally, to the 3D Secure generalization (Brlek, Hamadou, & Mullins, 2006). The idea developed by C-SET was to use banking smartcards and their certificates, through small card readers connected to the customer's computer, to secure the payment transactions. The price of that card reader working as a POS (point of sale) was a limit to the deployment of that solution.

3D-Secure

The current solution to solve the problem of electronic payments is 3D secure (3D–Secure Functional Specification, 2001), developed by VISA and used by MASTERCARD, which has gave up from SPA UCAF.

3D Secure is not only an authentication method; it is a payment architecture on Internet, launched by Visa in 2001. The commercial trademarks are « Secure Code » for MasterCard and « Verified by Visa » for Visa. The term 3D is the contraction of "Three Domains":

Figure 7. The different communications in 3D-secure payment



- Acquiring domain (bank acquirer and merchant);
- Issuer domain including the customer authentication;
- Interbank field, which makes it possible the two other fields to communicate on Internet.

3D Secure describes the different processing between the three domains to carry out a payment by bank smartcards and distributes the responsibilities in a balanced way between these domains:

- The customer's bank authenticates its client;
- The merchant's bank authenticates its merchant;
- The interbank domain makes it possible the merchant to start the customer's authentication using services of directories (MasterCard or Visa).

In the 3D-Secure authentication diagram, the first stage is the recording of the cardholder by his/her bank. Figure 7 presents the different communications for a transaction in the 3D-secure architecture. The recording procedure contains a series of questions, after which, the cardholder chooses a password, for example, that will ensure his/her authentication by its bank for each transaction.

In 3D-secure, the security of transactions lay on the banks and not on the merchant. The merchant benefits from the same level of payment guaranty as in the face-to-face trade in card payment. More than that, there is a responsibility transfer from the customer towards the issuer (or "Liability Shift"). During the payment phase, the bank issuer becomes responsible for the authentication of its cardholder as a preliminary step in the authorization request.

However, the programs "Secure Codes" and "Verified by Visa" leave the bank issuer free to choose the authentication method of their cardholder. Those complex exchanges are transparent for the merchant and the customer, and secure, very well, the e-commerce. The last problem is the customer authentication by his/her bank issuer. This will concern the future trends described in the next section of this chapter.

FUTURE TRENDS

We present, in this part, several issues concerning the perspectives of security in e-commerce.

Authentication

In fact, with 3D-secure, the authentication problem from the customer/merchant domain is replaced by the customer/issuing bank domain. The problem seems easier to solve because there is a constant relationship between the cardholder and his/her bank (Torres, Izquierdo, Ribagorda1, & Alcaide, 2005). Many solutions have been proposed to meet this need for safety, some based on biometrics (Jain & Pankanti, 2006), others on the use of the couple reader-smartcard, allowing a dynamic authentication of the cyber-consumer.

Whatever the chosen solution, the cost of the system of security is a key element, at the same time for the bank and the customer. For the bank, they are, concretely, the

technical and organizational costs of integration, deployment, management, and maintenance. On his/her side, the customer is interested by a simple tool of use at a moderate cost. It must represent a compromise between the security constraints and a convivial use of the tool of security of the bank-customer exchanges. Actually, bank payment chains use, mainly, two types of proof together to create a strong authentication:

- A smartcard to identify the cardholder and so, to authenticate the smartcard with the use of cryptographic keys and certificates ;
- A password to authenticate the cardholder.

The MasterCard initiative CAP integrates this type of strong authentication. We present, in the next two sections, two possible solutions that are explored in research works within this context.

Sopas Project

Connecting a simple terminal to the computer is realized in a project called SOPAS, in which we are involved. This project consists in developing a card reader with just a keyboard with 12 keys and a screen with two lines of 10 characters. The card reader is connected to the cardholder's computer by the USB port. A set of development software is installed on the computer. When the smartcard is introduced in the card reader, the screen indicates to type a PIN code. Then, the smartcard generates a token sent to the ACS, as shown in Figure 8.

To create a reader as simple as possible, it is necessary to use the new ISO 7816 specifications for cards that have an I^2C bus and a USB bus at one's disposal. All the certificates are calculated by the smartcard. The different protocols are indicated in Figure 9.

Dynamic Tokens

Another solution consists in using a smartcard reader unconnected, which generates a dynamic token that the customer has to enter on the computer keyboard. It is enough for the smartcard holder to insert its card, which can be a debit/credit card, in a pocket EMV reader in conformity with the CAP standards, such as ActivReader Solo (TM) of ActivCard, and to compose its code PIN there.

A dynamic onetime-password is generated with the card and can be used to check or sign a transaction. This password or signature is then subjected to the bank, through its Web site (or by telephone where it is confirmed by an operator),

Figure 8. Sopas secured solution for authentication

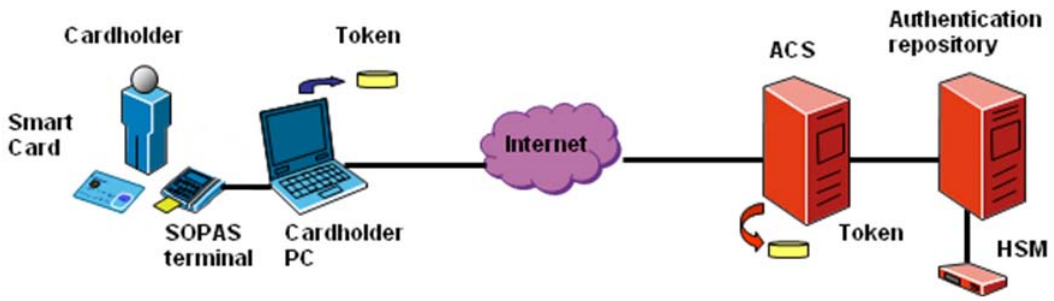
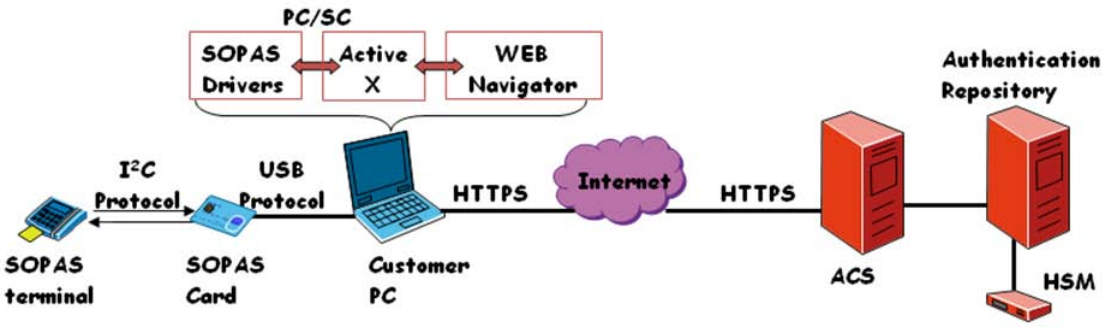


Figure 9. Protocols in SOPAS solution



for authentication, with an aim of checking the identity of the cardholder and the specific parameters of the transaction.

CONCLUSION

There are, currently, a great increase of the use of payment in e-commerce because many of their needs tend to be satisfied. Under no one circumstance, is it possible to create an absolute secured system, but new developments seem to be strong enough to protect, correctly, the e-commerce. The 3D-secure solution is now well implemented and will become a leader in the next few years for three main reasons:

- It is a method recommended by the card scheme Visa and MasterCard;
- There is a responsibility transfer from the customer towards the issuer ("Liability Shift");
- Banks are free to choose the authentication method, which is the visible part of the iceberg that is seen by the client of the bank, and can allow the banks to create a differentiation from their concurrent.

The authentication part is in progress (Walton, 2005), but there is, today, no one solution that can be considered as an emerging leader. The new authentication methods will be developed with three main principles: the solution must be secure, cheap, and easy to manipulate by a user. Today, many solutions are very secure but expensive, certain are complex for the cardholder, and certain are not secure enough. Many research and development laboratories are working today on that problem, and we can expect some good solutions in the near future.

REFERENCES

Brlek, S., Hamadou, S., & Mullins, J. (2006). A flaw in the electronic commerce protocol SET. *Information Processing Letters*, *97*, 104–108.

Chang, K. I., Bowyer, K. W., & Flynn, P. J. (2005). An evaluation of multimodal 2-D+3-D face biometrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(4), 619-624.

Chanson, S. T., & Cheung, T.-W. (2002). Design and implementation of a PKI-based end-to-end secure infrastructure for mobile e-commerce. In *World Wide Web archive,* vol. 4 (pp. 235 – 253). Hingham, MA: Kluwer Academic Publishers.

Critchlow, D., & Zhang, N. (2004). Security enhanced accountable anonymous PKI certificates for mobile e-commerce. *Computer Networks*, *45*, 483–503.

Furnell, S. M., & Karweni, T. (2000). Security implications of electronic commerce: A survey of consumers and businesses. *Internet Research*, *9*(5), 372-382.

Guitart, J., Carrera, D., Beltran, V., Torres, J., & Ayguade, E. (2007). Designing an overload control strategy for secure e-commerce applications. *Computer Networks*, *51*, 4492–4510.

Hawk, S. (2004). A comparison of B2C e-commerce in developing countries. *Electronic Commerce Research*, *4*, 181–199.

Herrmann, G., & Herrmann, P. (2004). Introduction: Security and trust in electronic commerce. *Electronic Commerce Research*, *4*, 5–7.

Jain, A.K., & Pankanti, S. (2006). A touch of money [biometric authentication systems]. *IEEE Spectrum magazine, 43*, 22-27.

Kwon, E.-K., Cho, Y.-G., & Chae, K.-J. (2001). Security enhancement on mobile commerce. *Lecture Notes In Computer Science,* vol. 2105, (pp. 164–176). Springer-Verlag.

Nabi, F. (2005). Secure business application logic for e-commerce systems. *Computers & Security*, *24*, 208-217.

Rennhard, M., Rafaeli, S., Mathy, L., Plattner, B., & Hutchison, D. (2004). Towards pseudonymous e-commerce. *Electronic Commerce Research*, *4*, 83-111.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, *5*, 115–153.

Tang, J. J., Waichee, F. A., & Veijalai, J. (2004). Supporting dispute handling in e-commerce transactions, A framework and related methodologies. *Electronic Commerce Research*, *4*, 393–413.

Torres, J., Izquierdo, A., Ribagorda1, A., & Alcaide, A. (2005). Secure electronic payments in heterogeneous networking: New authentication protocols approach. *Lecture Notes in Computer Science*, *3482*, 729-738.

Visa Corporation. (2001). *3D–secure functional specification*, Chip Card Specification v1.0.

Walton, R. (2005). Identity infrastructure: security considerations. In *Computer Fraud & Security*, (pp. 4-8).

KEY TERMS

CA: The certifying authority (CA) signs the certificates.

CAP: Chip authentication program (©MasterCard), CAP provides one line chip-based cardholder authentication within the SecureCode™ (3D-secure) program.

EMV: Eurocard, MasterCard and Visa specifications define the electronic payment transaction and its security.

PKI: Public-key infrastructure. The use of cryptography with public key on large scale, creates the need to manage large lists of public keys, for entity often repartee on the network. The public-key infrastructure manages that problem.

RA: The registration authority (RA) creates the pairs of keys

SET: Secure electronic transaction was a solution developed by a set of companies (Visa, MasterCard, GTE, IBM, Verisign...) to limit the risk that the customer can repudiate an e-commerce electronic payment transaction.

3D-SECURE: The current solution to solve the problem of e-commerce electronic payments, 3D-secure is used by VISA and by MASTERCARD.

Security Issues in Distributed Transaction Processing Systems

R. A. Haraty
Lebanese American University, Lebanon

INTRODUCTION

Transaction-processing systems (TPS) are becoming increasingly more available as commercial products. However, the approaches to the issues associated with using TPS in multilevel secure environments are still in the research stage. In this article, we address the issues of multilevel security in distributed transaction-processing systems. A distributed transaction-processing system (DTPS) is a collection of a finite number of centralized transaction-processing systems connected by a computer network. Each of these transaction-processing systems is controlled by a software layer and can be accessed both remotely and locally. Properties of a DTPS, such as data replication, may have a substantial effect on the security of the system. The security policies and integrity constraints adopted at each site may result in global security having inconsistent states. We address the issues of achieving a multilevel secure DTPS, and discuss the security constraints and data replication.

In this work, we address the issues of achieving a multilevel secure DTPSs system and discuss the security constraints and the replication of data items. The next section provides some background. Then, next, an overview of a distributed transaction-processing system is presented. In the fourth section, security-related issues are discussed. In the fifth section, a multilevel secure distributed transaction-processing system is presented. Then, in the next section, future trends are presented. The final section concludes the article.

BACKGROUND

Several commercial and military applications require a multilevel secure transaction-processing system (MLS/TPS). In an MLS/TPS, users are assigned classification levels that we denote by "clearances," and data items are assigned sensitivity levels. There are three interesting architectures that have been used to build MLS/TPSs from untrusted ones. These architectures are known as the integrity lock architecture, the kernelized architecture, and the data distribution architecture (Air Force Studies Board, 1983). While most of the techniques for TPS security are developed for traditional centralized TPSs, more TPS researchers are making substantial contributions to the development of a distributed TPS (Getta, 2003; Haraty, 1999; Haraty & Rahal, 2002; O'Connor & Gray, 1988).

A DTPS is a collection of a finite number of TPSs connected by a computer network (Ozsu & Valduriez, 1999). Each of these TPSs is controlled by a transaction management software layer and can be accessed both remotely and locally. A DTPS integrates information from the local TPS and presents remote users with transparent methods to use the total information in the system. An effective TPS system serves to maintain the ACIDity properties (i.e., atomicity, consistency, isolation, and durability) of transactions and must be superimposed on the preexisting local TPSs (Gray & Reuter, 1993).

One proposed architecture for MLS/TPS is the replicated architecture. This approach is being explored in several ongoing research efforts, including the Naval Research Laboratory Secure Information through replicated architecture (SINTRA) project (Thuraisingham, 1987). Data replication in DTPS has several implications for the security of the system. Replication allows data items in different local TPSs to be identified as logically belonging to the same entity. The security policies adopted by each site may result in global security having inconsistent states, because of the difference of local representation and management.

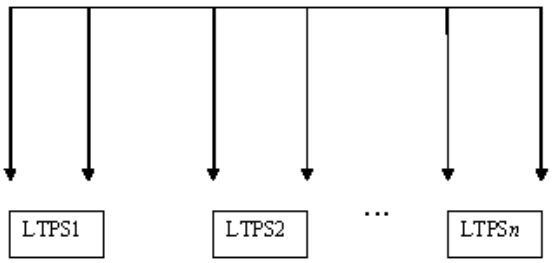
OVERVIEW OF DISTRIBUTED TRANSACTION-PROCESSING SYSTEMS

A DTPS consists of a set of preexisting local TPSs {LTPS*i* | 1 d•*i* d" *m*}, distributed among several interconnected sites. Each LTPS*i* is a software layer on a set of data items D*i*. Figure 1 depicts the architecture of a DTPS.

SECURITY ISSUES

Processes that execute on behalf of users are referred to as subjects. Objects, on the other hand, correspond to a data item. Objects can be files, records, or even fields. In this section, we present the notion of object classification with emphasis on the problem of conflicting security constraints due to replication.

Figure 1. Distributed transaction-processing system



A security classification is a function that associates each subject and each object with a given level of security. Many classifications, such as the security lattice, exist (Denning, 1976). However, a well-known classification is four-value function (DOD paradigm) that classifies objects into unclassified (U), confidential (C), secret (S), and adopt top secret (TS). A simple policy that can be established using a classification function SL is as follows:

Subject X can access (read) Object Y iff SL (Y) ≤ SL(X)

A security constraint consists of a data specification and a security value. The data specification defines any subset of the TPS. The security values can be given by a classification function. Specific values are unclassified, confidential, secret, and top-secret. Thuraisingham (1987) defined two types of security constraints—internal constraints and external constraints:

1. Internal constraints classify the entire TPS as well as relations, attributes, and tuples within a relation. These constraints can be applied to data, as they are actually stored in the TPS.
2. External constraints classify relationships between data and the results obtained by applying operations on the stored data, such as sum, average, and count. Among these constraints are the functional constraints and the dynamic constraints.

These security constraints are subject to inconsistency and conflicting local security constraints. A good global security approach should reject inconsistent security constraints and inconsistent clearance of users. Examples of the inconsistencies encountered include:

- **Conflicting security constraints:** Such constraints classify the same facts into different categories.
- **Overlapped security constraints:** These constraints cover overlapped data domains.
- **Inconsistent security level of replicated data:** Cases where different copies of replicated data may belong to different security cases.
- **Access privileges of users to replicated data:** Instances where a user may have different access rights on replicated data at different sites.

Several solutions have been proposed to solve these inconsistencies and define a global security policy that respects the local ones (Pfleeger, 1989; Thuraisingham, 1987).

There are several ways to combine local policies. The optimal combination should give a policy that defines all component policies and is still secure.

MULTILEVEL SECURE DISTRIBUTED TRANSACTION-PROCESSING SYSTEMS

There are two strategies for building MLS/DTPS from DTPS. These strategies include data replication and per-level-based distribution. The scope of this article does not include the issues associated with network security; but, it is particularly important to have the various local TPSs. Instead, we will assume that interconnection between the various local TPSs is secure and focus attention on security that has to be provided due to replication and other properties specific to the TPS.

The data distribution approach physically replicates low-level data at all higher-level TPSs. The advantage of the replicated architecture is that is fairly secure (McDermott & Sandhu, 1991). No performance overhead is associated with multilevel queries, because they are locally executed. On the other hand, because data is replicated, there is overhead associated with broadcasting updates of lower-level data to higher-level TPSs in a correct and secure manner. This broadcasting mechanism is known as "data synchronization" (Air Force Studies Board, 1983).

In the per-level-based approach, data are physically stored in separate local TPSs according to sensitivity level. Early examples of this approach were presented by Hinke and Schaefer (1975). The advantage of this approach is that updating transactions does not produce inconsistencies. Performance overhead associated with multilevel queries is a major disadvantage.

Global Commitment in Secure Environment

An important aspect of a correct TPS is atomic commitment (Bernstein et al., 1987). Unfortunately, the local TPS in a MLS/DTPS system cannot support atomic commitment, so the two-phase commit (2PC) protocol (Bernstein et al., 1987) cannot be implemented. 2PC is known to introduce covert channels. In order to establish a covert channel, there must be two cooperating agents/subjects in the system and an encod-

ing scheme. There are two main types of covert channels: covert storage channels and covert timing channels.

Covert storage channels disclose information from high to low subjects by manipulating a physical object that can or cannot be seen by the low subjects. For example, suppose there are two subjects of different security levels. Suppose also that these processes share a common resource—the available disk space. The secret subject creates a secret file that takes all of the available disk space to store the file. When the low subject attempts to create a file and store it onto the common disk, its request is denied. Through this denial, the high subject can signal information to the low subject. These signals are in terms of 0 and 1 bits that the low subject has to decode and turn into useful messages.

Covert timing channels can covertly send information by modulating observable delays of a common resource. This delay must be measured by low subjects cleanly; otherwise, the channel becomes noisy. For example, suppose we have two subjects again operating the low and high levels. The high subject can modulate the disk access time of the low subject by issuing numerous disk requests (thus transmitting a bit of 1) or zero disk requests (thus transmitting a zero). A system that is free from any type of covert channel is called covert channel secure.

Several distributed commitment protocols have been defined. A scheduler in MLS/DTPS that produces committable execution guarantees that a distributed transaction (a unit of work with execution sites: TPS1, TPS2, …, TPS*n*) becomes committed after it has been locally committed. The commitment of a distributed transaction means all of its subtransactions are committed. In this article, we follow the definition proposed by Bernstein et al. (1987):

If one subtransaction commits, then all other subtransactions will eventually commit.

We assume, in this article, that each subtransaction of a distributed transaction is designed to be executed in only one container. One can then say that a subtransaction has a security level.

FUTURE TRENDS

Future work will involve taking a closer look at MLS/DTPS and defining new and better ways of handling transaction management as well as query processing. Future work will also involve extending security issues to temporal and multimedia databases.

CONCLUSION

The security issues presented in this article highlight the intricacies required to architect a MLS/DTPS. We hope to address these issues further and to identify potential prototypes and engineering solutions that meet the requirements of MLS for DTPS.

REFERENCES

Air Force Studies Board, Committee on Multilevel Data Management. (1983). *Multilevel data management*. National Research Council.

Bernstein, P. A., Hadzilacos, V., & Goodman, N. (1987). *Concurrency control and recovery in database systems*. Reading, MA: Addison-Wesley.

Denning, D. (1976). Secure information flow in computer systems. Ph.D. dissertation. Purdue University.

Getta, J. R. (2003). Hybrid concurrency control in multilevel secure database systems. In *Proceedings of the IASTED International Conference—Applied Informatics*. Innsbruck, Austria.

Gray, J., & Reuter, A. (1993). Transaction processing: Concepts and techniques. San Francisco, CA: Morgan Kaufmann.

Haraty, R. A. (1999). A security policy manager for multilevel secure object oriented database management systems. In *Proceedings of the International Conference on Applied Modelling and Simulation*, Cairns, Queensland, Australia.

Haraty, R. A., & Rahal, I. (2002). A bit vectors algorithm for accelerating queries in multilevel secure databases. In *Proceedings of the International Conference on Computer Science, Software Engineering, Information Technology, e-Business, and Applications (CSITeA'02)*, Foz do Iguazu, Brazil.

Hinke, T., & Schaefer, M. (1975). Secure database management system, RADC-TR-75-266.

McDermott, J. P., & Sandhu, R. S. (1991). A single-level scheduler for the replicated architecture for multi-secure database. In *Proceedings of the Seventh Annual Computer Security Applications Conferences*.

O'Connor, J. P., & Gray, J. W. (1988). A distributed architecture for multilevel database security. In *Proceedings of the Security Conference*.

Ozsu, M. T., & Valduriez, P. (1999). *Principles of distributed database systems*. Upper Saddle River, NJ: Prentice Hall.

Pfleeger, C. P. (1989). *Security in computing*. Upper Saddle River, NJ: Prentice Hall.

Thuraisingham, M. B. (1987). Security of database systems. *Computer and Security, 6*(6).

KEY TERMS

Covert Channel: This is a channel that is not meant to route information, but nevertheless does.

Multilevel Secure Transaction-Processing System: This is a system whereby database users are assigned classification levels, and data items are assigned sensitivity levels.

Security Lattice: This is a partial (or total) order of security classes, where there is a least upper bound that dominates all the other security classes and a greatest lower bound that is dominated by all security classes.

Subject: This corresponds to a user or, more correctly, to a process that is running on behalf of a user.

Two-Phase Commit (2PC): This is an atomic commitment protocol that behaves as follows: The coordinator asks the participants to vote on commitment; if any votes No, the coordinator informs all participants to Abort; if all participants voted Yes, then the coordinator informs all participants to Commit.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2455-2458, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Security Issues in Mobile Code Paradigms

Simão Melo de Sousa
University of Beira Interior, Portugal

Mário M. Freire
University of Beira Interior, Portugal

Rui C. Cardoso
University of Beira Interior, Portugal

INTRODUCTION

Unlike mobile computing, in which hardware moves, *mobile code* moves from nodes to other nodes and can change the machines where it is executed. A paradigmatic example of such *mobile code* are Java applets that can be downloaded from a distant machine and executed by a virtual machine embedded in a browser. Multi-application smart cards (like Javacards) are an example of an execution environment that allows the loading and the execution of (mobile) programs into a card after its issuance. Code mobility allows the software reconfiguration without delivering a physical support, as done by Sun initially with Java to reprogram cable TV boxes, or nowadays, by Microsoft to promptly distribute software patches. PostScript files are another type of mobile programs which execute in printers to produce graphic images. Mobile code may also be used in distributed systems to adapt autonomously in order to balance loads or compensate for hardware failures (Brooks, 2004). Mobile code has received a great deal of interest as a promising solution to increase system flexibility, scalability, and reliability. However, to reach such objectives, some issues need to be matured, namely security issues. This article addresses security issues in *mobile code* paradigms.

BACKGROUND

Several *mobile code* paradigms have been reported (Brooks, 2004; Brooks & Orr, 2002; Fuggetta, Picco, & Vigna, 1998; Milojicic, Douglis, & Wheeler, 1999; Tennenhouse, Smith, Sincoskie, Wetherall, & Minden, 1997; Wu, Agrawal, & Abbadi, 1999). These paradigms differ on where code is executed and who determines when mobility occurs (Brooks & Orr, 2002; Brooks, 2004) and can be classified as follows:

- **Client-Server:** The user node invokes code resident on a distant node: the server or program node. This node fetches the required data from data nodes, executes the invoked program, and returns the result to the user node. Examples include the common object request broker architecture. CORBA integrates remote procedure calls (RPCs) with the object-oriented paradigm.
- **Remote Evaluation:** The user node requests the execution of code resident on a distant node. This node uploads the code to the node containing the data needed for its execution. The execution takes place in this node, and the result is then sent to the user node. Examples include CORBA, Simple Object Access Protocol (SOAP) and Web Services.
- **Code-On-Demand:** The user node requests the execution of code resident on a distant node. This code is downloaded on user node and locally executed. Examples include Java applets and Active X programs.
- **Process Migration:** The operating system dispatches processes from one node to others nodes in order to balance the load. Examples include Mosix and Sprite.
- **Mobile Agents:** The user node executes a program, called agent, which moves, along with its execution context, from node to node. The decision to move from one node to another node or to execute a specific set of operations on a particular node is made by the agent itself. The result of the execution is, at the end, transmitted within the program to the user node. There are several agent and multi-agent platforms.
- **Active Networks:** In this paradigm, the network configuration and infrastructure can be modified by the transmitted packets. Here, the packets act as mobile code. An example would be Capsules.

A mobile agent is a program that encapsulates code, data, and execution context. The mobile agent is sent by the client to another node. Unlike a procedure call, the agent does not have to return data to the client. The agent can migrate to other node, send information to the client, or come back to the client. However, the efficiency of each approach depends on network configuration and the size of programs and data files.

SECURITY ISSUES IN MOBILE CODE

One of the major challenges in the context of mobile code is the safety of the execution of untrusted code. This concern occurs naturally when we verify that mobile code to be executed comes from an eventually unknown source, or it was designed or compiled by unknown methods. In fact, the code may have been produced or changed by malicious sources. Thus, an execution environment for mobile code must be able to execute mobile code without allowing it to produce damages in the case of being a malicious code.

From a theoretical point of view, the problem of stating if a given program is inoffensive or malicious is not decidable in general. Thus, the quest of finding a universal filter that rejects every malicious code and accepts innocuous programs is an utopia. It is indeed very hard to universally and formally define what is a malicious program is. However, there exist several partial solutions which increase the safety of execution environments. They can be classified in these four approaches (Rubin & Geer, 1998; Zachary, 2003):

- Sandboxes, which limit or control the context in which code is executed;
- Code signing, which ensures that code comes from a trusted source and its integrity;
- Firewalls, which limits the accessibility; and
- Proof-carrying code (PCC), in which code carries explicit proof of its safety.

The first approach consists in the isolation of the code execution zone. Each mobile program is executed within a controlled context and isolated from the other processes (including memory). Control is assured by runtime monitoring of the performed operations. For instance, sensitive operations (whether operation on resources such as disks, memory, etc., or operations such as communications or data/files handling) may be forbidden or, at least, supervised. Enforcing security policies by confinement and runtime access control is relatively easy to implement (when compared with other approaches), easy to use, and provides a reasonable level of confidence. A successful example of such approach is the Java virtual machine and its security manager mechanism. However, runtime checking induces a penalty in terms of execution performance. In the same vein, access control policies limit the computational ability of mobile code (for instance, an innocuous applet could have access to the whole instruction set).

The next approach, the *code signing* approach, allows the execution of code which presents enough credentials. This mechanism is based on the extraction and the verification of a digital signature which is included in the code to be executed. This signature allows the identification of the code producer and the code integrity. If code comes from a source identified as secure and if the code has not been changed since it had to leave the source, then the execution environment may allow its execution. Such a mechanism takes place before the execution stage. Unfortunately, it does not provide information about the actions performed by the program and must be associated with other security mechanisms. Therefore, most popular mobile code execution/support systems such as Java and .NET integrate a combination of the two approaches, since this increases the flexibility of policy securities.

Another way to guarantee the security in a mobile context is based on the restriction and control of the mobility or the communication capability. These mechanisms rely on *firewalls* and other similar mechanisms. This approach allows precise control of the generated interactions by the executed program. However, since this mechanism acts in runtime, it leads to performance degradation of program execution and of the infrastructure that supports the execution. Another drawback is that the safety cannot be fulfilled exclusively in terms of safe interaction, but this approach can be used in conjunction to other security mechanisms.

Recent and emerging approaches try to minimize the need of runtime verification. Such approaches are known as proof-carrying code (Appel, 2001; Appel & Felty, 2001;Barthe, Grégoire, Kunz, & Rezk, 2006; Colby, Lee, Necula, Blau, Plesko, & Cline, 2000; Hamid, Shao, Trifonov, Monnier, & Ni, 2002) or static program analysis. These mechanisms operate on the code as soon as it is received and can get conclusions about the safety of the program without requiring its execution. From the code consumer point of view, the penalty is located in the loading time. The underlying principle is the following: the code to be executed is enriched in such a way that it contains enough information for the execution environment to verify the conformance of the program with respect to the security policies of the code consumer. If the program is approved, then it can be executed in a safe way and these policies do not need to be verified at runtime. The several approaches in these families of mechanisms differ in the quantity of the information required in the code to be executed. This information can vary from complete demonstrations (as the name "proof-carrying code" suggests) to simple type annotations. For instance, Java bytecode, the code executed by the Java virtual machine, is a typed low level language. This allows the Bytecode Verifier (BCV) of the Java platform to perform the static analysis of several safety policies. Because *proof-carrying code* is an emerging approach and a very promising technology (as witness recent initiatives like the European project MOBIUS IST 15905, the literature, or the emergence of certifying compilers (see the next section) for languages like JAVA), we will postpone its detailed description to the next section.

FUTURE TRENDS

The concept of *proof-carrying code* (PCC) was first defined by G. Necula and P. Lee in 1996 (Necula & Lee, 1996; Necula, 1997) and is based on two actors: the code consumer (the node which executes the code), the code producer (the node where the executable code is produced) and on the following two statements:

1. A proof is lot easier to check than to build (Aristotelian Principle); and
2. The code producer must be responsible for the behavior of the produced executable code.

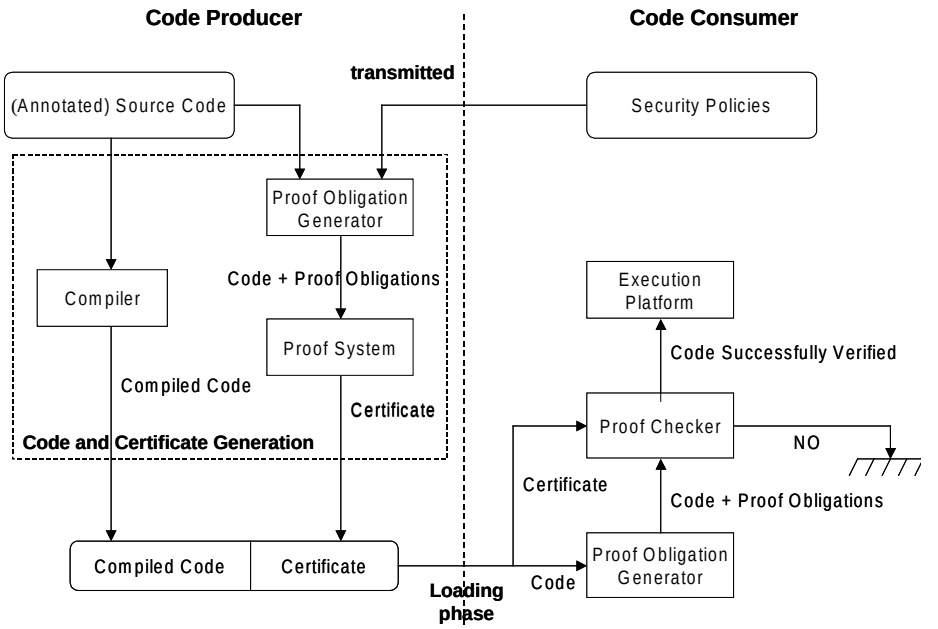
This gives rise to the following scenario:

- The code consumer defines a set of security policies and makes them available to the code producer.
- The code producer designs the program.
- The code producer annotates the program with pre-post-conditions, and invariants in order to provide enough information for the verification of the security policies.
- The code producer uses the compiler to produce executable code and verification tools to produce the formal proof that the executable code is compliant with the required policies. This formal proof is called certificate.
- The code consumer receives the executable code with its certificate. The code consumer verifies that the certificate is proof that the code is secure.
- In the positive case, the code can be safely executed without runtime checking.

Figure 1 summarizes the typical PCC architecture. The code and certificates generation stage (the emphasized rectangle in Figure 1) varies following the different implementations of PCC depending on whether the proof generation is done before, during, or after the compilation (or alternatively with the source code or with the compiled code). The overall infrastructure responsible for the generation of executable code and formal proofs (emphasized in Figure 1) receives source code and security policies and produces executable code and certificates. This is the reason why such a module is often referred in the literature by *certifying compiler*. Code annotations are used by both the producer and the consumer to produce proof obligations. These are the properties that the code must verify in order to be compliant with the security policies. The code producer must prove them and the code consumer must reproduce them to verify that the certificate contains exactly their proofs.

One definitive advantage of PCC is that the certificate only needs to be verified once during the loading time, and this process is automatic, efficient, and trustworthy. Moreover, the consumer does not have to trust the code producer, and a validated formal proof of a property is

Figure 1. General proof-carrying code architecture



a strong argument for its safety with respect to what the consumer assumes to be safe. Finally, PCC can be easily combined with other security mechanisms.

But, despite these interesting features, a lot of work around PCC remains to be done.

This is why PCC is an active and prolific research area essentially because the following challenges or open problems must be solved before engaging a real deployment of PCC architectures:

- Coupling certificates with executable code induces a network loading overhead. Certificates must be concise. This point is particularly difficult because this is not just a question of defining an adequate format for the certificate: For a property, there exist several proofs. Some are longer than others.
- On the other hand, the certificate must be checked as efficiently as possible.
- It is not reasonable to ask the code producer to carefully look at the compiled code. The producer side of the PCC architecture must provide tools that allow the programmer to focus its attention at the source level. The compiler is then a key element. Unfortunately, it is difficult to implement certifying features and mix them with code optimizers.
- Another point is the proof construction automation. It is possible to provide automatic tools for simple proofs. But proofs of security properties tend to be tricky and, so, hard to automate. PCC relieves the consumer from heavy verifications but must not weigh down the code producer effort.
- Designing a good security policy is an art, and the good performance of a PCC architecture strongly depends on it.

CONCLUSION

An overview of mobile code paradigms and its security has been presented. Safe execution of untrusted code, which is one of the major challenges in mobile code security, has been analyzed. The sandboxes, code signing and firewalls approaches have been described, and special attention has been taken in the description of the proof-carrying code approach. As an emerging trends, it has been shown that the Proof-carrying code technology is a powerful and very promising technique, but despite recent and outstanding advances these past few years, its challenges can still be resumed by: Is the PCC approach able to fill the gap between its alluring, underlying principles and a practical and industrial use?

REFERENCES

Appel, A. W. (2001). Foundational proof-carrying code. *IEEE Logic in Computer Science*.

Appel, A. W., & Felty, A. P. (2001). A semantic model of types and machine instructions for proof-carrying code. *POPL 2000: The 27th ACM SIG- PLAN SIGACT Symposium on Principles of Programming Languages*, Boston (pp. 243-253). ACM Press.

Barthe, G., Grégoire, B., Kunz, C., & Rezk, T. (2006). *Certificate translation for optimizing compilers.*

Brooks, R. R. (2004). Mobile code paradigms and security issues. *IEEE Internet Computing, 8*(3), 54-59.

Brooks, R.R., & Orr, N. (2002). A model for mobile code using interacting automata. *IEEE Transactions on Mobile Computing, 1*(4), 313-326.

Colby, C., Lee, P., Necula, G. C., Blau, F., Plesko, M., & Cline, K. (2000). *A certifying compiler for java. ACM SIGPLAN Notices, 35*(5), 95-107.

Fuggetta, A., Picco, G.P., & Vigna, G. (1998). Understanding code mobility. *IEEE Transactions of Software Engineering, 24*(5), 342-361.

Hamid, N., Shao, Z., Trifonov, V., Monnier, S., & Ni, Z. (2002). A syntactic approach to foundational proof-carrying code. *Journal of Automated Reasoning*.

Milojicic, D., Douglis, F., & Wheeler, R. (Eds.). (1999). *Mobility: Processes computers, and agents.* Reading, MA: Addison-Wesley.

Necula, G. (1997). Proof-carrying code. *Proceedings of the 24th Annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL '97)* (pp. 106-119).

Necula, G., & Lee, P. (1996). Safe kernel extensions without run-time checking. *Proceedings of 2nd Symposium on Operating Systems Design and Implementation (OSDI'96)* (pp. 229-243).

Rubin, A. D., & Geer, D.E. (1998). Mobile code security. *IEEE Internet Computing, 2*(6), 30-34.

Tennenhouse, D.L., Smith, J. M., Sincoskie, W. D., Wetherall, D. J., & Minden, G. J. (1997). A survey of active network research. *IEEE Communications Magazine, 35*(1), 80-86.

Wu, D., Agrawal, D., & Abbadi, A. (1999). StratOSphere: Unification of code, data, location, scope, and mobility. *Proceedings of International Symposium on Distributed Objects and Applications* (pp. 12-23).

Zachary, J.M. (2003). Protecting mobile code in the wild. *IEEE Internet Computing*, *7*(2), 78-82.

KEY TERMS

Active Networks Paradigm: In this mobile code paradigm, packets moving through the network reprogram the network infrastructure.

Client-Server Paradigm: In this mobile code paradigm, client invokes code resident on another node.

Code-On-Demand Paradigm: In this mobile code paradigm, local clients download and execute code as needed.

Mobile Agents Paradigm: In this mobile code paradigm, a program moves from site to site. It represents the current more advanced stage of mobile computing paradigms.

Process Migration Paradigm: In this mobile code paradigm, processes move from one node to another to balance the load.

Proof-Carrying Code: This approach for security in mobile code is based on two actors: the code consumer and the code producer, and on the following two statements: 1) a proof is lot easier to check than to build and 2) the code producer must be responsible for the behavior of the produced executable code.

Remote Evaluation Paradigm: In this mobile code paradigm, a remote node downloads code before executing it.

SOAP (Simple Object Access Protocol): Protocol developed by Microsoft, DevelopMentor, and Userland Software. It allows the communication between a program executed in a given operating system with a program in the same or in another operating system using HTTP and XML as a mechanism for message exchange.

Security-Based Knowledge Management

Shuyuan Mary Ho
Syracuse University, USA

Chingning Wang
Syracuse University, USA

INTRODUCTION

As knowledge is recognized as intellectual (or intangible) assets that can enhance an organization's competitive capability, how to effectively manage knowledge assets has become an important issue in the information age (Alavi, 2000). Literature in knowledge management (KM) emphasizes issues on knowledge creation, knowledge codification, knowledge sharing, and knowledge utilization; however, security perspectives on assuring knowledge confidentiality and knowledge integrity are left unaddressed.

This article takes an initial step to address different perspectives of security centric knowledge management. This article first presents the background of security-based knowledge management. It then discusses sources of security threats in knowledge-based organizations and identifies challenges in four aspects of knowledge management practices, which are culture-based, strategy-based, content-based (or standard-based), and technology-based, along with a discussion of 10 corresponding security domains. Real-world cases are intertwined with the challenges faced by knowledge-based organizations. This article ends with further envisioning the future trends of the security-based knowledge management.

BACKGROUND

While knowledge management enables collaboration within an organization to retain and share knowledge and experience, threats to knowledge confidentiality and integrity have raised corporate concern.

Threats against organizations are multi-faceted. Hackers and crackers have greatly threatened information[1] transmission over the virtual world such as the Internet, intranet and extranet. Many corporations have applied multi-layered security solutions to prevent threats of this kind. System-layered security solutions include system logs, host-based intrusion detection, file encryption, identity-based, or role-based access control. Network and infrastructure-layered security solutions include firewall, virtual private network, public key infrastructure, cryptography, network-based intrusion detection, and intrusion prevention. Physical separation of the networks is seen as the fundamental practice to protecting information assets. One of the best examples would be the practice of the de-militarized zone[2] (called DMZ). In the social context, personnel with a knowledge base of the corporate assets could pose potential threats to the organizations. Corporate regulations, best practice, and ethical codes could be security solutions to threats of this nature.

With the advance of information technology, threats to information security have come to corporations with increasing frequency and subtlety, and so information security should be an emphasis in the field of knowledge management for the new information era.

SOURCES OF INFORMATION THREATS

Threats against information security can be intentional or unintentional. These threats can be further differentiated into internal threats and externals threats. Sources of security threats are tabulated in Table 1.

External Threats

In a networked environment, intentional threats from outsiders include attacks from hacker/crackers (Harris, 2003).

Table 1. Sources of threats to information security

	Intentional	**Unintentional**
External	Malicious Hacker/Cracker; outsider ID theft	Natural disasters
Internal	Personnel fraudulence; unauthorized modification/leakage of knowledge/information[1];	System failure

These attacks include man-in-the-middle, Trojan Horse, denial-of-service (DOS), logic bombs, viruses, and so forth. These attacks could cause malfunctions of the systems and result in loss in terms of time and money. For instance, *The New York Times* suffered attacks in 1998, which resulted in its web servers being compromised and its Web front page replaced.

Identification (ID) theft is another type of external threat. The purpose of those malicious hackers in this scenario is to steal personal information such as social security numbers, credit card numbers, or to conduct banking transactions without authorized permission. Additionally, they could commit identification (ID) fraud by falsely presenting stolen identification in exchange of goods and services in this virtual e-commerce world. These misconducts could cause huge losses and creditability damage to innocent victims.

On the other hand, external threats from natural forces such as earthquakes, tornadoes, or tsunamis are generally unintentional and thus become difficult to prevent in advance. Damages caused by natural forces could be devastating and hard to restore. The tsunami that devastated Southern Asia in December 2004 was an example of this kind.

Internal Threats

Threats from insiders are subtle and complex (Hayden, 1999; Park & Ho, 2004). Like a double-edged sword, the knowledge from internal personnel could bring beneficial advantages, but also ironically, could bring potential security threats as well to an organization (Benkoil, 1998; Powell & Rosenberg, 1987). For example, in 1985, Jonathan Pollard, a U.S. Navy intelligence analyst, was arrested for passing classified U.S. intelligence information to Israel. The Israeli government was then able to analyze U.S. intelligence infrastructure such as the locations of facilities and identities of intelligence agents. As such the General Accounting Office (GAO) reported in 1993 that insiders' abusive use of the National Crime Information Center (NCIC) for personal reasons had threatened the safety of U.S. citizens (Benkoil, 1998; Powell & Rosenberg, 1987). Insider threats have been statistically increased since late 1980, and, more so, the cost of loss from the insider threats has exceeded the threats from outsiders (Hayden, 1999; Park & Ho, 2004). Normally insiders are not interested in sabotaging hardware systems or applications but in obtaining critical information and accessing the internal level of resources. Due to the insider threats, the paradigm has shifted toward a more sophisticated and security centric collaborative environment

On the other hand, infrastructure failures such as power outage and destruction of infrastructural devices such as routers and switches are seen as unintentional internal threats. These threats bring up issues on operational site redundancy, system maintenance, and information preservation.

ASPECTS OF KNOWLEDGE MANAGEMENT AND INFORMATION SECURITY

Knowledge management can be classified into four aspects: culture-based, strategy-based, content-based, and technology-based (Alavi & Leidner, 1999; Oostveen & van den Besselaar, 2004). These four aspects represent four major themes in managing knowledge. Each theme of knowledge management has related security concerns. International Information System Security Certification Consortium, Inc. (known as (ISC)2) has identified 10 domains of information security that would assure knowledge management practices in organizations. These 10 information security domains include law, investigation and ethics, risk assessment, operations security, business continuity planning, physical security, security architecture technology standards, access controls, telecommunications and network security, applications security, and cryptography. These aspects are not mutually exclusive but overlap one another. We map the 10 domains of information security (Hansche, Berti, & Hare, 2004; Harris, 2003) with four aspects of knowledge management identified in literature (Table 2).

Table 2. Conceptual map of knowledge management aspects and domains of information security

Aspects of Knowledge Management	Corresponding Domains of Information Security
Culture-based	Law, Investigation and Ethics
Strategy-based	Risk Assessment Physical Security Operations Security Business Continuity Planning
Standard-based	Security Architecture
Technology-based	Access Controls Telecommunications & Network Infrastructure Security Applications Security Cryptography

Culture-Based Aspect

This aspect of knowledge management focuses on cultivating both learning and sharing cultures in organizations (Alavi & Leidner, 1999; Oostveen & van den Besselaar, 2004). While creating a willing-to-share organization culture could enhance collaboration and productivity, this process is vulnerable to unethical information misuse (Alavi & Leidner, 1999; Oostveen & van den Besselaar, 2004). This brings up the legal aspect of information security in terms of laws, investigation, and ethics. As such, it is important to establish information policies, regulations and "boundary control" to govern information sharing and usage (Liddy, 2001). It is also important to establish an internal audit and monitoring system to supervise information activities. Proper background check and security clearance toward authorized insiders should be conducted to ensure organizational security.

Strategy-Based Aspect

Objective goals of an organization can be achieved with various corporate strategies which are applicable to knowledge management practice. An organization may face challenges from different levels of practices such as information collecting, processing, and storage (Green, Hurley, & Shaw, 2004; Nyaboga & Mwaura, 2004); it could enact different coping strategies to manage the challenges they experienced.

Strategy-based aspect of knowledge management relates to a wider scope of security issues that could be further differentiated into pre-stage security, in-stage security, and post-stage security. Risk assessment is a pre-stage security wherein it analyzes potential risk (or threats) and rates vulnerabilities so that effective controls can be implemented (Hall, 2004; Siponen, 2001). Pre-staged risk assessment informs operations security and physical security that are in-stage security. Operation security concerns the organization's ability to audit, monitor, and identify events when there are security breaches, and report to the right entities for incident response (Hall, 2004; Siponen, 2001). Physical security is a mandatory guard to protect knowledge-based organization from unethical or even brutal physical intrusion. Preventative, detective, and corrective systems are commonly used for security checks around the organization's buildings (Hall, 2004; Siponen, 2001). Although disaster management, disaster recovery, incidents response, and business continuity planning ought to be considered beforehand and planned for, they are indeed countermeasures that would strategize the post-stage security (Hall, 2004; Siponen, 2001).

Content-Based Aspect

This aspect of knowledge management defines how the content of information is collected or stored, which relates to security architecture (Alavi & Leidner, 1999). When applying security to ensuring corporate information and knowledge, it is important to adopt the standards that have been widely used. Standards in information security upon systems and applications always prevent the uncertain applets, codes, or backdoors programmed in the organizations critical knowledge-based systems. The standards that evaluate the security property and assurance level of the IT systems and products include Common Criteria and ITSEC[4]. They are transactional standards being recognized and practiced internationally. TCSEC[5], IETF[6] and IPSec[7], on the other hand, are Internet security standards. The challenges that most organizations would face remain in how to comply the security architecture of hardware, firmware, and software during system development cycle with the standards of the information technology security evaluation (Hall, 2004).

Technology-Based Aspect

Technology-based aspect of knowledge management relates to transforming the actionable and readily accessible information into knowledge with enabling technologies such as data mining, data warehouse, information retrieval, expert system, smart system, intranet/extranet, multimedia, and so forth security issues that are applied to ensuring technological aspect of knowledge management include access control, telecommunications and network infrastructure security, applications security, and cryptography.

Access Controls generally include mandatory access control, discretionary access control, and role-based access control (RBAC). All access control mechanisms guard the information assets from unauthorized access which specifies who has the permission to access what resource. For example, RBAC indicates the granted permission assigned to specific sets of roles in that organization. In other words, an authenticated personnel or insider with a specific set of role functions is only allowed to conduct authorized activities. The activity is defined as the work that an insider conducts over a permitted resource. RBAC's protection mechanism controls the gates that determine whether the authorized personnel are able to enter into an authorized area for authorized activities. In sum, non-legitimate activities are excluded and legitimate activities are guarded.

Telecommunications and Network Infrastructure Security emphasizes on securing the communication links. The technologies used are virtual private network with cryptography implementation, firewall to block out unspecified ports and services, and router to ensure the routed packets sent to the right destination address, and so forth. The challenge here is not only to adopt preventive, detective, and corrective measures at the infrastructural level, but also correlate security incidents from various infrastructural devices into meaningful threat analysis and knowledge. How to determine threat indicators, eliminate false positives and false negatives, and

generate early warnings of the intrusion would be the keys to secure intranet and extranet communications before the information is stolen or modified.

Applications Security Control should be adopted and implemented within systems and applications software in their early development stage. Information assurance at this domain refers to securing agents, applets, software, databases, data-warehouses, and knowledge-based systems from implanted backdoor programs or system development holes. Boundary control over software releases becomes essential in preventing software exploitation attack (Liddy, 2001). Understanding the system life-cycle and how to incorporate security features into a knowledge-based system would be key challenges.

Cryptography ensures the confidentiality and is used to encrypt texts, e-mails, messages, and documents that are transmitted over the link. It can further work with identification and authorization systems to specify designated recipients for the encrypted messages or documents. The challenges here remain in how to maintain confidentiality while the speed of transmission and encryption/decryption is not sacrificed. How to effectively integrate the security technologies such as public key infrastructure, digital signature, and key management into different user-centric applications becomes critical to a successful transformation of a knowledge-based organization.

FUTURE TRENDS

The challenges faced by today's knowledge-based organizations have been subtly shifted from sharing knowledge to protecting knowledge against all sorts of threats. The ethical and technical aspects of monitoring and detecting personnel's anomalous behavior based on their roles and functions within the organization have increasingly attracted public attentions. The ability to audit, monitor, encounter security threats and breaches would lead to successful implementations of the mechanism of security preventive control and incident response. The technological solutions for digital forensics and electronic crime investigation will change the way how a knowledge-based organization would plan and develop efficacious corporate security policy and procedure.

Although knowledge management is deemed a critical role to organization's competitive capabilities, the enabling aspect of information security toward a knowledge-based organization should be emphasized. Information security procedures and techniques not only protect intellectual assets from unexpected or unethical intrusions but also remedy the potential losses from the threats. In summation, information security ensures knowledge confidentiality from unauthorized disclosure, knowledge integrity from unauthorized modification, and knowledge availability and sharing in today's dynamic, but still vulnerable, digital world.

REFERENCES

Alavi, M. (2000). Managing organization knowledge. In R. W. Zmud (Ed.), *Framing the domains of IT management: Projecting the future through the past* (pp. 15-28). Cincinnati, OH: Pinnaflex Education Resources.

Alavi, M., & Leidner, D.E. (1999, February). Knowledge management systems: Issues, challenges, and benefits. *Communications of the Association for Information Systems, 1*(7).

Benkoil, D. (1998, October 25). *An unrepentant spy: Jonathan Pollard serving a life sentence*. Retrieved August 10, 2004, from ABCNEWS.com

Green, C. W., Hurley, T., & Shaw, P. (2004, May 23-26). Knowledge management in organizational settings: The effect of normative influence and technological support on knowledge creation and transfer. *Proceedings on 2004 Information Resources Management Association International Conference*, New Orleans, LA (pp. 444-446). Hershey, PA: Idea Group Publishing.

Hall, D. E. (2004, June 13-18). Requirements and policy challenges in highly secure environments. *SIGMOD 2004,* Paris.

Hansche, S., Berti, J., & Hare, C. (2004). *Official (ISC)2 guide to the CISSP exam*. Boca Raton, FL: CRC Press LLC.

Harris, S. (2003, June 17). *CISSP All-in-one exam guide* (2nd ed.). New York: McGraw-Hill Osborne Media.

Hayden, M. V. (1999, July). The insider threat to U.S. government information systems. *National Security Telecommunications and Information Systems Security Committee (NSTISSAM) INFOSEC 1-99*. Retrieved from http://www.nstissc.gov/Assets/pdf/NSTISSAM_ INFOSEC1-99.pdf

Liddy, E. D. (2001). Information security and sharing. *Information Today Online.* Retrieved March 3, 2005, from http://www.infotoday.com/online/OL2001/liddy5_01.html

Nyaboga, A. B., & Mwaura, M. F. (2004, May 23-26). A conceptual evaluation of the intranets to support knowledge management. *Proceedings on 2004 Information Resources Management Association International Conference*, New Orleans, LA (pp. 1203-1205). Hershey, PA: Idea Group Publishing.

Oostveen, A. M., & van den Besselaar, P. (2004). From small scale to large scale user participation: A case study of participatory design in e-government systems. *ACM Transactions on Computer Systems: Proceedings Participatory Design Conference 2004,* Toronto, Canada (pp. 173-182).

Park, J. S., & Ho, S. M. (2004, June 10-11). Composite role-based monitoring for countering insider threat. *Proceedings of Second Symposium on Intelligence and Security Informatics*, Tucson, AZ (pp. 201-213).

Powell, S., & Rosenberg, R. (1987). Spying between friends: Pollard case simmers on. *U.S. News 7 World Report*, March 16.

Siponen, M. T. (2001, June). Five dimensions of information security awareness. *Computers and Society*, 24-29.

Whitman, M. E. (2003, August). Enemy at the gate: Threats to information security. *Communications of the ACM, 46*(8).

KEY TERMS

Asymmetric Cryptography: A technique using different keys to encrypt and decrypt messages. Usually, a public key is used to encrypt a message, and a private key is used to decrypt it.

Buffer Overflow Attacks: A technique an attacker use to overwrite the data. Buffer overflow occurs when the program writes more information into the space than the buffer has in its memory. This allows the attackers to control the program and execute the code they wrote.

Cryptography: A technique that encrypts and decrypts a message for ensuring security and privacy during the exchange and communication process.

Denial-of-Service (DOS) Attack: A type of attack that intends to deprive legitimate users of access to services of a resource they would normally have.

Identification (ID) Fraud: Unauthorized use of other's personal information to commit crimes, usually in exchange of economic gains.

Identification (ID) Theft: Illegally obtaining other's personal information such as social security number, credit card number, and password without authorized access and permission. ID theft is related to ID fraud.

Information Security: Policy, procedure and techniques used to protect information from unauthorized use or other misconducts.

Knowledge Management: Refers to the managerial process of acquiring, storing and transferring knowledge intra- or inter-organizationally with the purposes of fulfilling tasks and enhancing competitiveness.

Symmetric Cryptography: A technique using the same secret key to encrypt and decrypt a message. Also called secret-key cryptography or conventional cryptography.

Trojan Horse Attacks: A maliciously security-breaking backdoor program that is disguised as benign so that dangerous program can be unleashed. It could result in damage to the disk, ID theft, a denial-of-service attack, and so on.

ENDNOTES

1 The difference between information and knowledge has been under debate for a long time. Due to the fact that its differentiation is outside the scope of this article, the terms of "information" and "knowledge" are used interchangeably.

2 Demilitarized zone (abbreviated as DMZ) is in a subnetwork that is located in between the intranet and Internet.

3 The difference between information and knowledge has been under debate for a long time. Due to the fact that its differentiation is outside the scope of this article, the terms of "information" and "knowledge" are used interchangeably.

4 ITSEC stands for Information Technology Security Evaluation Criteria.

5 TCSEC stands for Trusted Computer System Evaluation Criteria (commonly called the Orange Book). It sets the standards for trusted computer products.

6 IETF stands for Internet Engineering Task Force. It regards the development of the Internet architecture and the operations.

7 IPSEC stands for Internet Protocol Security, which defines a set of protocols developed by the IETF to support secure IP packet exchange over the network.

Self Organization Algorithms for Mobile Devices

M.A. Sánchez-Acevedo
CINVESTAV Unidad Guadalajara, Mexico

E. López-Mellado
CINVESTAV Unidad Guadalajara, Mexico

F. Ramos-Corchado
CINVESTAV Unidad Guadalajara, Mexico

INTRODUCTION

Self-organization is a phenomenon in nature which has been studied in several areas, namely biology, thermodynamics, cybernetics, computing modeling, and economics. Systems exhibiting self-organization have well defined characteristics such as robustness, adaptability, and scalability, which make self-organization an attractive field of study for two kinds of applications: a) maintaining the communication among mobile devices in wireless networks, and b) coordination of swarms of mobile robots.

In ad hoc networks, there is not necessarily an underlying infrastructure in which the nodes can maintain communicated with other nodes; so due to this feature, it is necessary to provide efficient self-organization algorithms for routing, managing, and reconfiguring the network.

Furthermore, self-organization in nature provide clear examples about how complex behaviors can arise from only local interaction between entities, namely the ants colony, feather formation, and flock of birds. Based on the above mentioned examples, several algorithms have been proposed to accomplish robot formations using only local interactions.

Due to resource constraints in mobile devices, self-organization requires simple algorithms for maintaining and adapting wireless networks. The use of resources for establishing robot formations can be reduced by improving simple rules to accomplish the formation. This article first presents a brief overview of several works developed in ad hoc networks; then, delves deeper into the key algorithms; and finally, challenges arising in this area are discussed.

Figure 1. Wireless mesh network

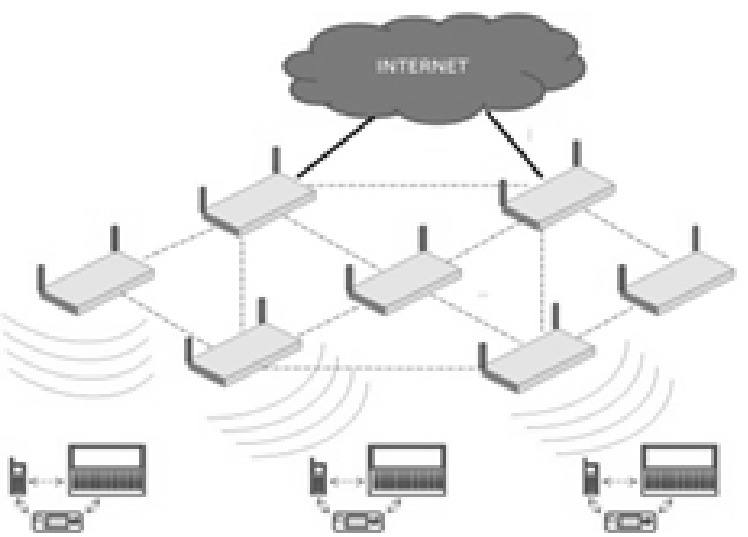


Figure 2. Wireless sensor network

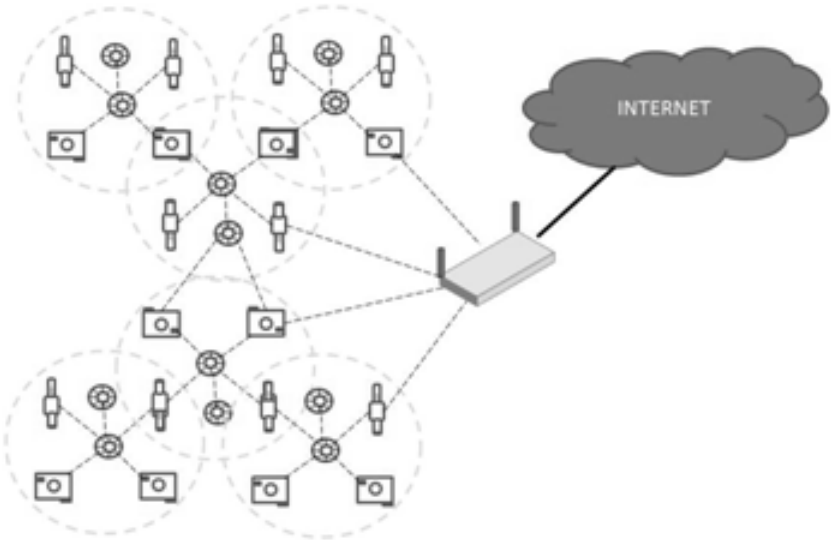
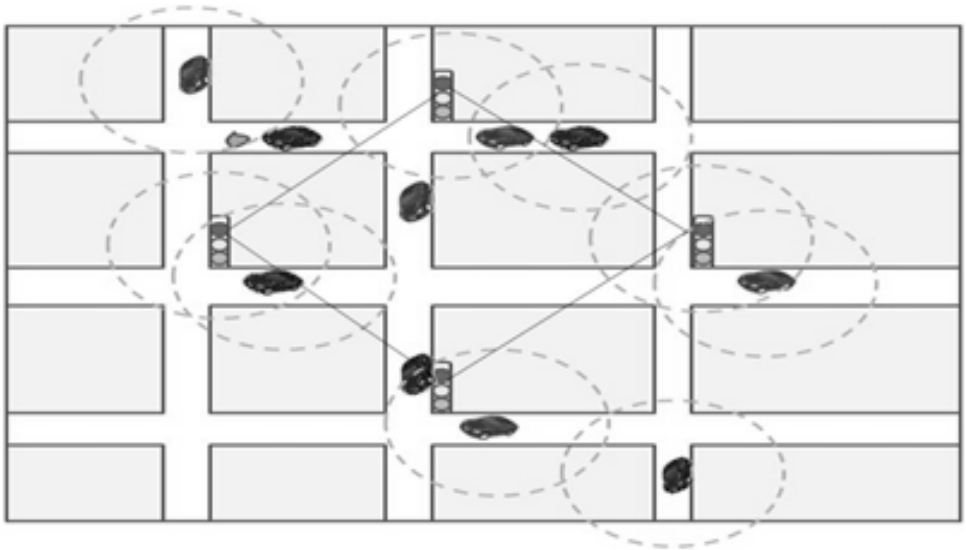


Figure 3. Vehicular ad hoc network



BACKGROUND

Self-organizing networks can be classified as follows:

- Mobile Ad hoc Networks (MANET). In this kind of network the nodes are mobile devices operating under energy consumption constraints.
- Wireless Mesh Networks (WMNs). The backbone of the network consists of mesh routers (which have reduced mobility) allowing the communications between mobile mesh clients. (Figure 1).
- Wireless Sensor Networks (WSNs). They are composed of a large number of sensor nodes widespread on a field; they are used for collecting information on the environment and transmitting such information to themselves or to a base station (Figure 2).
- Vehicular Ad Hoc Networks (VANETs). These networks use ad hoc communications for detecting obstacles on the road and emergency events by exchanging information obtained from the roadside or from other vehicles (Figure 3).

A common feature of these networks is the mobility of nodes; furthermore, there exists not always a relaying structure. These facts evidence the need for including within

the network management systems, strategies for maintaining communication between nodes despite the mobility.

In next section, several approaches for solving this problem are presented. Although there are some issues not addressed yet, this overview may be aware to the reader about the state of art in this area.

SELF-ORGANIZATION IN AD-HOC NETWORKS

An ad-hoc network may be structured in three ways: flat topology, hierarchical topology, and hybrid topology (Dressler, 2006; Tang & Tienfield, 2006). Hierarchical topology is the most widely used because it is scalable, allowing for efficient routing protocols for large networks; furthermore, it has well stated battery consumption schemas. The hybrid topology is suitable for combining ad hoc networks with existing network infrastructures.

Self-organization algorithms applied to networks can be classified as cluster-based, role-based, location information-aided, biological inspired, and economically inspired (Dressler, 2006; Tang & Tienfield, 2006). Cluster-based self-organization algorithms have received more attention, because they intend to provide efficient ways for creating and maintaining organization of nodes according to a hierarchical topology.

Cluster-based self-organizing algorithms

Several approaches have been proposed to self-organize nodes of an ad-hoc network in a cluster-based fashion. The work presented in Basagni (1999a) models the network as a graph G = (V, E) where V is the set of nodes and E is the set of links that exists between two communicating nodes. Every node is classed as either ordinary or clusterhead; it has an identifier and a weight w which depends on the characteristics of the node.

The key properties of the clusters formed by the algorithm are the following:

- Every ordinary node has at least a neighboring clusterhead (dominance property).
- Every ordinary node affiliates with the neighboring clusterhead which has the bigger weight.
- Two clusterheads cannot be neighbors (independence property).

The main idea of the algorithm is based on the ability of a given node for deciding its role that it has to play only when all its neighbors have decided their own roles. The algorithm is executed in every node with the sole knowledge of its ID, weight, and neighbors' ID. In order to accomplish the task, two procedures are considered: On Receiving_Clusterhead, which is executed when a node receives a Clusterhead message from other node, and the On Receive JOIN procedure, which is executed by the node that has received a JOIN message from a neighbor.

The disadvantage of the proposed approach is that the stability of the network is affected by a high mobility in the network, mainly in clusterheads. In order to avoid this problem a new approach is proposed (Basagni, 1999b) where the properties of the clusters are modified as follows:

- Every ordinary node affiliates with one and only one clusterhead.
- For every ordinary node, there is no clusterhead in the neighborhood of the node such that its weight is bigger than the clusterhead of the node, plus a threshold h.
- A clusterhead cannot have more than k neighboring clusterheads.

The modifications implemented in this work preserve the network topology, even though new clusterheads move into a cluster with a clusterhead assigned. When the new clusterhead weight is lower than the current clusterhead weight, plus a given threshold, it changes its role to ordinary node.

In Blazevic, Buttyan, Capkun, Giordano, Hubaux, and Le Boudec (2001), routing algorithms, positioning methods, mobility management, and algorithms including incentives for cooperation are proposed for designing a self-organized mobile ad-hoc network.

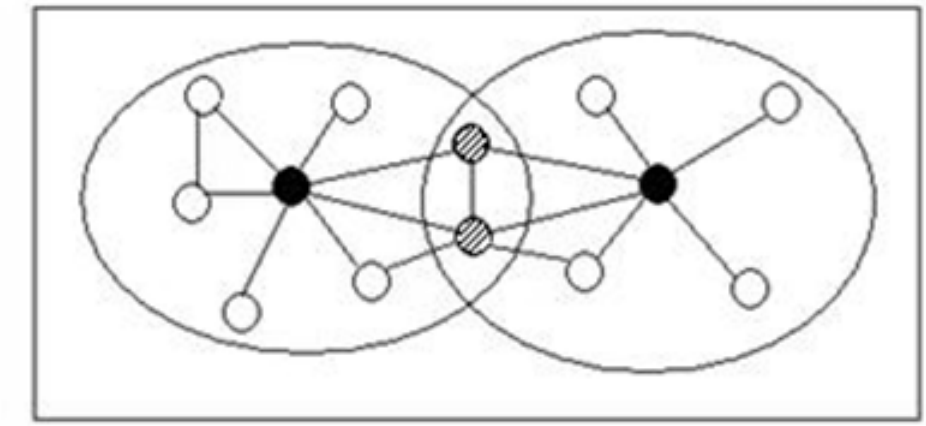
In order to accomplish the packet forwarding, two routing algorithms are proposed. Terminode Local Routing (TLR) allows reaching nodes located a limited number of hops away. Terminode Remote Routing (TRR) allows sending data to nonreachable nodes using the TLR algorithm. A Terminode attempts to maintain several paths for every destination. Several paths can be used by a terminode at the same time.

When the nodes are located in different geographical areas the packets are sent first to anchors, which are points specified by cartesian coordinates. The source nodes compute the anchors by using path discovery methods.

In this approach the mobility management is achieved by defining virtual home regions (VHR), where every node advertises its current position. The VHR has a fixed center and a variable radius, which adapts according to the density of the area, for maintaining a constant number of nodes.

However, in a network with resources constraints, every node will try to reduce the resource consumption; this behavior reduces the performance of the overall network. In order to avoid this situation an algorithm including incentive for cooperation has been proposed, where a nuglet is used as currency, then every time a node is used to forward a packet, it receives nuglets in return. When a node needs to send a packet it pays some nuglets for the service.

Figure 4. Black nodes represent the clusterheads, and dashed nodes represent the nodes acting as connection between clusters



Other works have proposed to organize the network nodes according to the preferences of the users, as in Herrmann and Geihs (2003), where a socio-aware approach is proposed. The goal of a socio-aware system is clustering users according to their common interest. The clusters are formed by nodes with the same preferences; different clusters are joined by nodes with shared preferences. An organized cluster is showed in Figure 4.

An important problem to solve in these networks is service allocation; it is necessary to find the nodes that have the highest number of interactions with other nodes, which could be good candidates for running services. Services should be placed in these nodes to facilitate the access to as many members as possible.

In the works mentioned above, the way to decide which node is a clusterhead depends on the weight of the node, which is obtained according to their characteristics; however, sometimes this is not enough for maintaining the topology in long time periods. In Mitton and Fleury (2003), a new metric called density is proposed. The k-density of a node is defined as the ratio of the number of edges between the node and its k-neighbors (the degree of the node), the number of edges between the k-neighbors of the node, and the number of nodes inside the k-neighborhood of the node.

To form the cluster and select the clusterhead, each node computes its k-density and broadcast it to all its k-neighbors. The node with the bigger value is the winner, and it is elected as a clusterhead; once the clusterhead is elected, its ID and its density is broadcasted by all nodes which have joined this cluster. The clusterheads will be distant at least three hops between them.

The cluster organization has to adapt to topology changes caused by the node's mobility. In order to accomplish the adaptation, the nodes have to check periodically their environment and their mobility. When a node becomes too mobile it will not join any cluster; but if the node is able to communicate, it joins to its neighbor that has the highest density.

The approach proposed above allows constructing a robust topology, which is not affected by nodes that become too mobile; however, the nodes defined as clusterheads will consume their energy faster than the ordinary nodes, because they have to forward packets coming from many nodes. For this reason it is necessary to establish a balance between energy consumption, and the number of nodes that belong to a cluster.

In Sivavakeesar, Pavlou, and Liotta (2004), the concept of virtual clusters is introduced, which facilitates the formation of stable clusters. A virtual cluster is a circular region that represents a geographical area in such a way that every node can determine the virtual region it is in. Every virtual cluster has a unique identifier based on the geographic location.

In order to maintain the topology of the network, it is necessary to construct models that allow adapting the topology based on the node mobility. In this approach an intelligent mobility prediction is proposed. It is accomplished by deriving probabilistic prediction of particular user mobility, by using its accumulated movement-history.

The movement-history is represented by a mobility tree constructed using the LZ78 algorithm (Bhattacharya & Das Sajal, 1999). The probability of a node arriving to a given virtual cluster is obtained by applying a second order Markov model. The algorithm proposed in this work predicts the availability of the nodes giving the possibility to provide QoS applications; however, it is necessary to consider energy constraints to avoid the waste of network resources in a disproportionate way.

The mobility management has been studied also in Bluetooth ad-hoc networks; in Song, Chaegwon, and Choi (2004), the scatternet topology management is dealt along with packet routing. In order to obtain a robust scatternet topology, is necessary to configure the structure for having

sufficient bridges to forward the packets. But if a bridge participates in many piconets, it could cause the decreasing of the network performance; so it is necessary to limit the degree of a bridge.

The steps followed for every node to form a scatternet are given below:

- Start the inquiry procedure.
- Establish a temporary connection.
- Examine the collected information.
- If there is a master, join the master's piconet as a slave through a master/slave role change.
- When there are several masters, become a slave/slave bridge that links the masters.
- If all nodes are slaves, make its own piconet by acting as a master.
- If it cannot find nearby nodes, repeat process.

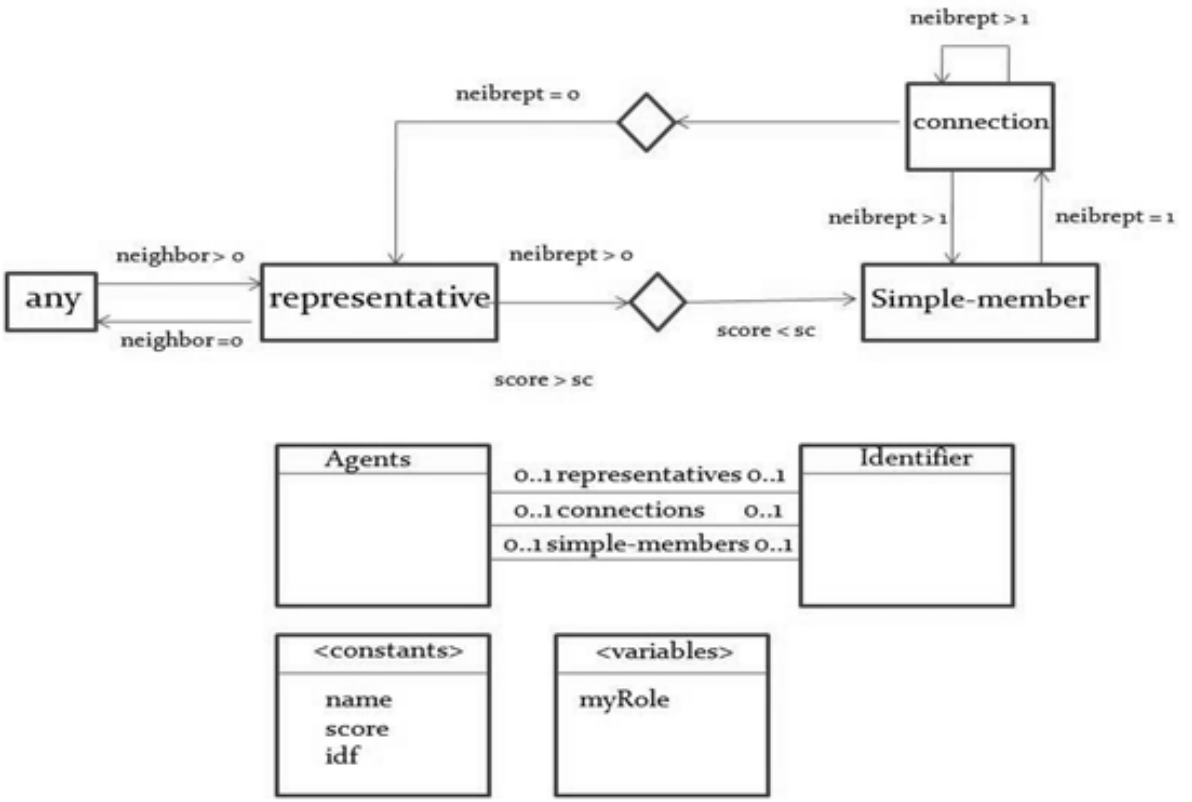
Every master in the network maintains three tables for routing the packets; when a bridge is created, this bridge node sends to each master an APT (Adjacent Piconet Table) update message. A master renews its PMT (Piconet Member Table) when it gets a PMT update message. When the PMT update message contains information about an unknown piconet, the master creates a new entry for this piconet in its PMT. The master also rebuilds its PRT (Piconet Routing Table), using the information contained in the updated PMT.

This approach forms a robust scatternet; however, in scenarios with high mobility, the network partition is increased.

According to the increase in mobile devices and wireless technology advances, new applications try to be implemented in wireless networks. As an example, in Wellnitz and Wolf (2006), an algorithm to support multiplayer games in wireless ad-hoc networks is proposed. The approach implements a server selection algorithm using three phases, discovery, determination, and marking.

In the discovery phase, every node determines its weight and degree, and broadcast them periodically. During the determination phase, the games servers are determined based on the highest weight among all neighboring nodes. Finally, in the marking phase, every node which has been determined as a game server tags itself as a server; the nodes who are neighbors of a server tag themselves as neighbors, and the nodes which are not servers or neighbors keep an empty tag.

Figure 5.Modeling self-organization algorithm with AUML



Nodes mobility and environment changes are dealt by a four phase process, where each node monitors changes during the game. When a node gets three or more hops away from a server, it initiates a new discovery phase. If there are no servers near, it starts a new server determination process.

The problem in game applications resides on the fact that the game state has to be stored by every server, so when a server is going to leave the network, it is necessary to move the information into the new server; if the network has high mobility or there is not a stable structure, the network performance is affected considerably.

Several works have tried to formalize and validate the self-organizing ad-hoc networks for proving properties' desirables in a network. In the next section we are going to present the main approaches.

Formalizing Self-organization in Ad-hoc Networks

In Fadil, Koning, Ramos, Jamont, and Occelo (2006), a graphical design technique for achieving cluster-based self-organization is proposed. In this approach, three roles are considered in the network structure. The representative node acts as a clusterhead, the node which can communicate with more than one representative play the role of connection, and the rest of nodes are simple members.

The assignment of a representative role is based on the weight of the node; the node with the bigger weight becomes a representative; if there is another node with the same weight, the identifier id is used to solve the conflict. This algorithm is modeled using first AUML, shown in Figure 5, and then translated into B language to prove safety and liveness properties. Two algorithms are proposed for detecting and resolve conflicts in role assignments for every node.

In Oleshchuk (2003), the network is modeled as a directed graph that may change dynamically; in the graph the set of vertices represent the nodes, and the set of edges represent the arcs between nodes which can communicate directly.

In this work two assumptions are made:

- The nodes are mobiles, not trusted, and have a unique id.
- The communication links are bidirectional and not trusted.

Every node is described as a process in Promela, and the properties that will be proven are described in Linear Temporal Logic (LTL). Such properties can be seen as part of requirement specification.

With this descriptions, the model checker SPIN (Holzmann,1997) can be used for proving the network properties described in LTL, and the network is modeled by processes in Promela. The LTL formulas and the processes are translated into a Büchi automaton by the model checker; then, an intersection of the two automata is obtained, and an acceptance cycle is searched.

The disadvantage of creating automatons is the state explosion; however, new techniques for reducing the space needed to store information have been implemented, increasing the possibility of solving more problems.

Another approach proposed in Johnen and Nguven (2006), give a self-stabilized version of the algorithms proposed in Basagni (1999a, 1999b), where the properties of the algorithms constitute attractors for the system, and it is proven that the self-stabilization is obtained in a finite time.

FUTURE TRENDS

There exist many issues which have to be addressed before the self-organized networks can be engineered, for example, how to ensure that the self-organized networks are secure, how to implement QoS policies in the network to be able of build multimedia networks, how to validate that the communication protocols are efficient, and how to prove that the self-organizing algorithms always converge to a stable state in an acceptable time. Bio-inspired self-organization algorithms are being used also in robot formation applications.

CONCLUSION

In this article, we reviewed the approaches proposed for accomplishing self-organization in wireless networks; we could see this is probably the main feature of future networks. The importance of self-organization due to computation cost, and for obtaining complex behaviors which are robust, scalable and adaptable, is clear. The approaches presented here give the line to continue addressing these issues.

Finally, we pointed out several issues to be addressed before self-organization can be engineered in wireless networks. In the next few years, the wireless networks in all the fashions presented in this article are going to be part of our daily life.

REFERENCES

Basagni, S. (1999a). Distributed clustering for ad hoc networks. In *Proceedings of the International Symposium on Parallel Architectures, Algorithms, and Networks*, (p. 130).

Basagni, S. (1999b). Distributed and mobility-adaptive clustering for multimedia support in multi-hop wireless networks. In *Proceedings of the IEEE Vehicular Technology Conference (VTC),* Amsterdam, The Netherlands, (Vol. 9, pp. 19-22).

Bhattacharya, A., & Sajal, K., Das. (1999, August). LeZi-Update: An information-theoretic approach to track mobile users in PCS networks. In *Proceedings of the Fifth ACM/ IEEE International Conference on Mobile Computing and Networking (Mobicom99),* Seattle, WA, (pp. 1-12).

Blazevic, L., Buttyan, L., Capkun, S., Giordano, S., Hubaux, J.P., & Le Boudec, J.-Y. (2001, June). Self-organization in mobile ad-hoc networks: The approach of terminodes. *IEEE Communications Magazine, 39*(6), 166174.

Dressler, F. (2006, March). *Self-organization in ad hoc networks: Overview and classification* (Tech. Rep. No. 02/06). University of Erlangen, Department of Computer Science 7.

Fadil, H., Koning, J.-L., Ramos, F., Jamont, J.-P., & Occelo, M. (2006). Graphically designing and formally checking self-organizations for wireless network systems. In *Proceedings of the International Conference on Self-Organization and Autonomic Systems in Computing and Communications*, (Vol. 2, No. 3, pp. 297-302).

Herrmann, K., & Geihs, K. (2003). Self-organization in mobile ad hoc networks based on the dynamics of interaction. In *Proceedings of the Spring Meeting of the GI-specialized Group of Operating Systems*.

Holzmann, G. J. (1997, May). The model checker SPIN. *IEEE Transactions on Software Engineering, 23*(5), 279-295.

Johnen, C., & Nguven, L. H. (2006). *Self-stabilizing weight-based clustering algorithm for ad-hoc sensor networks* (Vol. 4240, pp. 83-94). Lecture Notes in Computer Science: Springer-Verlag.

Mitton, N., & Fleury, E. (2003, December). *Self-organization in ad hoc networks* (INRIA research Rep. No. 5042).

Oleshchuk, V. (2003, September). Ad hoc sensor networks: Modeling, specification and verification. In *Proceedings of the IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*, (pp. 76-79).

Sivavakeesar, S., Pavlou, G., & Liotta, A. (2004, March). Stable clustering through mobility prediction for large-scale multihop intelligent ad hoc networks. In *Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC'04),* Georgia, USA, (Vol. 3, pp. 1488-1493).

Song, O., Chaegwon, L., & Choi, C.-H. (2004). Mobility management in Bluetooth ad hoc networks. In *Proceedings of the 14th Joint Conference on Communications and Information, JCCI 2004*.

Tang, H., & Tienfield, H. (2006). Self-organizing networks of communications and computing. *International Transactions on Systems Science and Applications, 1*.

Wellnitz, O., & Wolf, L. (2006). Assigning game server roles in mobile ad-hoc networks. In *Proceedings of the 16th ACM International Workshop on Network and Operating Systems Support for Digital Audio and Video (NOSSDAV'06).*

KEY TERMS

Mobile Devices: Computer-based communication devices that give us the possibility to be connected wherever we are at any time.

Self-Organization: Phenomenon in nature which allows the spontaneous appearing and maintaining of a functional structure (Tang & Tienfield, 2006).

Wireless Networks: Networks that do not need wires to allow the communication; instead they use wireless technologies as 802.11, WiFi, WiMax, Zigbi, and so forth.

Bio-Inspired Algorithms: Algorithms based on living beings' behaviors to accomplish a task efficiently.

Ad-Hoc Networks: Networks that do not need an infrastructure to allow the communication. The communication is established between devices using a wireless technology.

Cluster-Based Algorithms: These kinds of algorithms organize the nodes in clusters where a clusterhead is elected. The members of the cluster can communicate with nodes in other clusters through clusterhead.

QoS (Quality of Service): Set of technologies that provide a reliable data transmission ensuring data quality.

Self-Organization in Social Software for Learning

Jon Dron
Athabasca University, Canada

INTRODUCTION

The Internet has long been touted as an answer to the needs of adult learners, providing a wealth of resources and the means to communicate in many ways with many people. This promise has been rarely fulfilled and, when it is, often by mimicking traditional instructor-led processes of education.

As a large network, the Internet has characteristics that differentiate it from other learning environments, most notably due to its size: the sum of the value of a network increases as the square of the number of members (Kelly, 1998), even before aggregate effects are considered. Churchill (1943) said, "We shape our dwellings and afterwards our dwellings shape us." If this is true of buildings then it is even more so of the fluid and ever-changing virtual environments made possible by the Internet. Our dwellings are no longer fixed but may be molded by the people that inhabit them. This article discusses a range of approaches that make use of this affordance to provide environments that support groups of adult learners in their learning needs.

BACKGROUND

Darby (2003) identifies three generations of networked learning environments used in adult education. First-generation systems are direct analogues of traditional courses, simply translating existing structures and course materials. Like their traditionally delivered forebears, they are dependent on individual authors. Second-generation systems tend to be team-built and designed for the medium from pedagogical first principles, but still within a traditional course-based format. Third-generation systems break away from such course-led conventions and provide such things as just-in-time learning, guided paths through knowledge management systems, and personalized curricula. This article is concerned primarily with such third-generation environments.

Saba's interpretation of Moore's theory of transactional distance predicts that in an educational transaction, as structure increases, dialogue decreases and vice versa (Moore & Kearsley, 1996; Saba & Shearer, 1994). What is significant in differentiating learning experiences is not the *physical* distance between learners and teachers, but the *transactional* distance, measured by the degree of interaction between them. Highly structured educational activities have a high transactional distance, while those involving much discussion have a lower transactional distance.

In a traditional learning environment, the structure of the experience is provided by the teacher or the instructional designer. However, learners will not benefit equally from any given structure, as different learners learn differently. It would be better if learners could select appropriate approaches for their needs—to choose whether or not to choose, to control or to be controlled (Dron, 2007a). Without a teacher, help with this might be provided by the opinions of other learners. However, eliciting those opinions, assessing their reliability/relevance, actually finding the resources in the first place, and once found, fitting them into a structured learning experience is difficult. Several approaches to these problems are available, but first it is necessary to introduce a few concepts of self-organization.

SELF-ORGANIZING PROCESSES

Self-organization processes are emergent: the interactions of many autonomous agents lead to structure, not due to central control, but to the nature of the system itself. Such processes are very common in nature and in human social systems. Two in particular are of interest here, *evolution* and *stigmergy.*

Based primarily on work following that of Darwin (1872), evolution is one of the most powerful self-organizing principles, whereby a process of replication with variation combined with natural selection (survival of the fittest) leads to a finely balanced self-adjusting system. It is important to note that "fittest" does not mean "best" by any other measure than the ability to survive in a given environment.

Stigmergy, a form of indirect communication through signs left in the environment (Grassé, 1959), leads to self-organized behavior—examples range from ant trails and termite mounds to forest footpaths, money markets, and bank-runs. For example, ants wander randomly until they find food, after which they return to the nest, leaving a trail of pheromones. Other ants are more likely to wander where such pheromone trails mark the route. When they too find food,

they too leave a trail. The stronger the trail, the more other ants are drawn to it. This positive feedback loop continues until the food runs out, after which the trail evaporates.

A full discussion of the many factors that result in a successful self-organizing system is beyond the scope of this article. However, the following brief discussion should give a flavor of what is involved.

Self-organizing processes occur through local interactions. For systems to develop any sort of complexity, it is necessary for these interactions to occur at a number of scales. For instance, the interactions of bacteria in an ant's gut affect the ant, groups of ants can affect tree growth, tree growth can affect climate. Local interactions should form small clusters, which in turn interact with each other, leading to ever-increasing scales of self-organization. However, in general, the large and slow-moving affect the small and fast far more than vice versa, which is a common feature of self-organizing systems, from forests to cities (Brand, 1997). Parcellation is also an important feature of such systems (Calvin, 1997). As Darwin found in the Galapagos Islands, isolated populations tend to develop differently and more rapidly than their mainland counterparts. Any self-organizing system relies on interactions between more or less autonomous agents. The precise level of interactivity varies, but it is interesting to note that, for a system which teeters at the edge of chaos, neither too stable to change nor too changeable to develop, the average number of connections between interacting agents tends to stabilize around just over two (Kauffman, 1995). Systems must be capable of change, being in a more or less permanently unfinished state. Already perfect systems cannot evolve (Shirky, 1996). Equally, systems in perpetual flux can never achieve the stability to achieve self-organization.

SOCIAL SOFTWARE AND THE IMPORTANCE OF THE GROUP

Social software has been defined by Clay Shirky as that in which the group is a first-class object within the system (Allen, 2004). Early social systems such as discussion forums and mailing lists tended to provide a means of supporting individual interactions, with little consideration of the combinatorial effects of the behavior of the many. Typically, they scaled badly, suffering equally from too many as from too few users. In newer social software that underpins the hype-laden term 'Web 2.0', emergent patterns are capitalized upon and reified. For example, tag clouds provide a snapshot of aggregates of classifications by many individuals, social networking software provides structured webs that are generated from individual links between users, wikis gain their structure from individual decisions to link pages, and clusters of linked blogs give texture to the blogosphere. In all cases, the primary determinant of structure is the bottom-up, local behavior of the many. This means that (in general) social software gets better as more people use it.

Interactions within an e-learning environment have previously only considered agents such as the individual learner, the teacher, and the software with each other and with others of the same kind (Anderson, 2003). If the group is a distinct entity from the individuals of which it is composed, then there are more potential interactions to consider. In particular, the group may be seen as, in some ways, a potential teacher within the system (Dron, 2006a).

SOME EXAMPLES OF SELF-ORGANIZED LEARNING IN PRACTICE

For many knowledge-seekers, the starting point is often Google (*http://www.google.com*), perhaps the largest and most pre-eminent example of social software available today. Google's PageRank™ algorithm (Brin & Page, 2000) is based on the assumption that most Web pages provide links to other sites when those sites are considered in some way valuable. Implicitly, the more links that point to a given site, the higher its approval rating. Combined with a content-based search for keywords, documents returned therefore should have a high degree of relevance and reliability. This approach is self-organized, incorporating evolution as unlinked sites "die" and stigmergy as more-visited sites get more links pointing to them (Gregorio, 2003). It is social, not relying on a central controlling authority to provide decisions on resources' usefulness or give a structure to the content that is returned. However, limited parcellation, problems with term ambiguity, and the lack of a support for identifying relevant resources for specific learner needs beyond content-based searching make Google a relatively poor learning tool.

Social navigation, which explicitly capitalizes on stigmergy to enable the navigation or classification behavior of previous users to influence those who follow, is becoming almost ubiquitous on social sites. Tag clouds of the sort found on del.icio.us (*http://del.icio.us*), Flickr (*http://www.flickr.com*), or MySpace (*http://www.myspace.com*) emphasize popular tags by increasing the font size relative to those that are less popular and limiting the display to popular tags, providing a constantly changing map of a community's interests. Applied in an educational setting, such systems offer many benefits, but at the cost of many distractions, inappropriate content, and a breadth of focus that is as likely to discourage as to enthuse learners.

Wikis allow anyone, or sometimes a more closed community, to edit any page. The potential for chaos is enormous, and yet Wikipedia (*http://www.wikipedia.org*), an encyclopedia generated by thousands of volunteers with

little central authority, is hugely successful. At the time of writing, the English version of the site had well over a million articles. Soft-security (Cunningham, 2006) allows the wisdom of the crowd to override intentional or unintentional errors introduced by the individual: the fact that it is easier to undo changes than to make them, combined with a large community, leads to a highly reliable and comprehensive source of knowledge (Giles, 2005), which is widely used by learners as a means of tapping into collective expertise. Wikipedia's success is in part due to its strong structure and simple policies. Interestingly, it makes use of a meta-wikipedia where Wikipedians may discuss issues relating to articles to be posted. This scaled parcellation contributes greatly to the evolution of ideas.

Blogging communities form through links between blogs. These may occur inline, or through trackbacks (two-way hyperlinks between blog postings) or blogrolls (explicit lists of links to related sites), which together generate emergent webs of related blogs. As links between blogs grow, they start to form small, stigmergic clusters (Gregorio, 2003). As long as an appropriate cluster can be found, such networks provide a powerful means of finding structure in a subject. This may be facilitated through a recommender system or dedicated blog search tool such as Technorati (*http://www.technorati.com*), which itself uses social navigation in the form of tag clouds. Blogs usually offer interaction and can enable learners to discover and actively participate in relevant communities to help them to learn.

Collaborative filters are recommender systems that make use of implicit and/or explicit preferences to provide recommendations based on similarities between user models. Collaborative filters are very useful for matching users' preferences, but tend to be less successful when seeking learning resources, because to learn is to change. My previous preferences for particular resources will be a less reliable guide to my future needs than preferences for books, movies, or Web sites because the act of using a learning resource will (if successful) change me in ways that are likely to differ from how they have changed you. Nonetheless, systems such as Altered Vista (Recker, Walker, & Wiley, 2000) and RACOFI (Anderson et al., 2003) have achieved some success in this area.

Social networking sites, which are concerned with finding like-minded or otherwise related people, offer another means of structuring the environment from the bottom up. Sites such as MySpace, Orkut (*http://www.orkut.com*), Ecademy (*http://www.ecademy.com*), and FriendsReunited (*http://www.friendsreunited.co.uk*) are concerned with establishing webs of links between people. Again, no centralized controller determines the structure, which forms because of local links provided by individual users. When used to seek fellow learners or experts, such systems offer promising benefits for the learner.

EXPLICITLY EMERGENT LEARNING ENVIRONMENTS

Some social software explicitly exploits self-organizing principles for learning. The selection presented here is a relatively small subset that indicates how this area is developing.

An exception to the rule that collaborative filters cannot cope well with changing needs is CoFIND (Dron, Mitchell, & Boyne, 2003), which combines both evolutionary and stigmergic principles to provide both social navigation and collaborative filtering. Rather than simple good-bad ratings, it employs a multi-dimensional matrix of "qualities," which loosely translate into those things that learners find useful about a resource—for instance, if it is good for beginners, detailed, exciting, and so on. Because qualities are created by learners and used by other learners to provide explicit ratings, they provide a kind of footprint of the learning needs that led to a particular resource being recommended. This remains even after the learner has moved on. By basing its recommendations on an explicit metadata model, rather than a user model, it overcomes the problem of changing user needs that afflicts other collaborative filters. In keeping with evolutionary principles, not just the resources in the system but the metadata which describe them are in competition with each other. Combined with positive feedback loops generated by social navigation using stigmergy, this means that each CoFIND instance develops into a unique ecosystem composed of smaller, interacting ecosystems.

Elgg (*http://elgg.net*) is a powerful social networking system that enables learners to blog, podcast, share resources, and find other users with similar interests. Like many systems, it uses tags and stigmergic tag clouds to identify shared interests. With a focus on the educational sector, every resource, be it a file, a podcast, a blog entry, or a comment, offers fine-grained user control over the permissions to view or change it that are offered. It thus supports strong parcellation as communities can be as private or public as their members wish, while always providing weaker connections between clusters of interest. It has been used extensively around the world in educational settings (Anderson, 2006; Campbell, 2005; Dron, 2006b) as a means of breaking free of the strictures of more top-down learning management systems and of giving control to the end user.

Wiki-based systems for education are becoming increasingly sophisticated. For example, Sloep, van Rosmalen, Kester, Brouns, & Koper (2006) describe an application of latent semantic analysis to automatically generate wiki pages from a knowledge base in response to questions which may then be edited by humans to become more relevant, thus creating a hybrid combination of the strengths of automation with those of human experts. OurWeb (Miettinen, Kurhila, Nokelainen, & Tirri, 2005) overlays footprints and annotations on a wiki page, providing stigmergic social navigation

to structure a constantly evolving environment. Emergent patterns of use lend structure to temper the potential chaos of unconstrained wiki growth.

The Knowledge Sea II uses navigation behavior to identify resources of interest to a community of C programmers, using a combination of explicit and implicit ratings to generate a visualization resembling a sea, with greater depth of color representing greater levels of interest. Again, stigmergic self-organization develops through individual interactions with the system (Brusilovsky, Chavan, & Farzan, 2004).

EDUCO (Kurhila, Miettinen, Nokelainen, & Tirri, 2002) uses social navigation to provide not only visual indicators of the relative popularity of documents within the system, but also real-time indicators of who is currently viewing them. This is combined with a chat system that enables interactions between users, providing a powerful incentive to visit pages that are currently being viewed. A similar technique is employed in Dron's (2005, 2007b) Dwellings, which makes use of stigmergic self-organizing principles suggested by Jacobs (1961) in *The Death and Life of Great American Cities.*

The Comtella system uses peer-to-peer protocols to enable its users to share documents, and thence to find documents that others have found useful. Incorporating a ranking system partly inspired by self-organizing market mechanisms, Comtella is replete with emergent structure based on the aggregated behavior of individuals (Vassileva, 2004). Comtella is extensible and has spawned an interesting discussion forum that uses similar mechanisms to help find relevant people and postings (Webster & Vassileva, 2006).

FUTURE TRENDS

The tensions between bottom-up design and the top-down needs of educational institutions and organizations make it uncertain that, despite widespread use, self-organizing, social software will notably impact existing institutional education structures. The large and slow will always dominate the small and fast. Darby's third generation of learning environments already exists, but the majority of effort is still being expended on first- and (occasionally) second-generation systems. However, as the need for lifelong, just-in-time learning becomes ever more significant, it is certain that software combining the wisdom of crowds will have an important role to play in all walks of life.

With a trend towards meaningful metadata being appended to resources using tags or standards such as RDF (Resource Description Framework), the relevance of search results may be improved over the coming years. However, to turn such information into useful knowledge and learning, social software is needed that combines the knowledge of many people, effectively amplifying intelligence and operating in some senses as a kind of group mind. The massive growth in the use of social software and the technologies of Web 2.0 seems irrepressible, offering structure through dialogic processes, with many consequent benefits, enabling the learner to choose whether to be in control or to accept control by the emergent systems that arise (Dron, 2006a).

Interoperability is central: increasing use will be made of mashups, combinations of Web services, and RSS feeds that merge content from many sources (O'Reilly, 2005). Applications that are composed of constantly changing hosted services are quick to build and therefore easy to evolve as needs and communities change. This merging of parcellated populations offers many opportunities for cross-fertilization between systems and communities, as can be seen in sites such as Mappr (*http://www.mappr.com/*), The Daily Mashup (*http://dailymashup.com/*), or Doggdot (*http://doggdot.us/*). Evolutionary change in systems employing mashups occurs on a different scale from that of its individual components, potentially enabling richer and more complex structures to develop at more hierarchical layers. Exactly how this will be adopted in educational settings remains to be seen, but there are already some effective uses of the principle. Elgg, for example, integrates RSS feeds from other sites seamlessly into the local environment, allowing local, parcellated, evolutionary processes to transform the structure from another site into one that is more relevant to a given community's needs.

CONCLUSION

The World Wide Web is becoming ever-more dynamic. Because of social software, the generation of resources to learn from is moving away from rule-based machine- or human-governed information to a more symbiotic relationship, where the strengths of machines and the strengths of people are combined and, in the process, mutually enhanced. In the process, the high transactional distance of resource-based learning is reduced by glimpses of the footprints of others. In a world where roles are changing faster than the traditional course-based approaches to the delivery of learning can address, the resulting group mind can adapt itself more readily and effectively than any single person to the needs of individual learners.

REFERENCES

Allen, C. (2004). *Tracing the evolution of social software.* Retrieved December 29, 2005, from http://www.lifewithalacrity.com/2004/10/tracing_the_evo.html

Anderson, M., Ball, M., Boley, H., Greene, S., Howse, N., Lemire, D. et al. (2003). RACOFI: A rule-applying collab-

orative filtering system. *Proceedings of COLA'03,* Halifax, Canada.

Anderson, T. (2003). Modes of interaction in distance education: Recent developments and research questions. In M.G. Moore & W.G. Anderson (Eds.), *Handbook of distance education* (pp. 129-146). Hillsdale, NJ: Lawrence Erlbaum.

Anderson, T. (2006). Social software applications in formal online education. *Proceedings of the 6th IEEE International Conference on Advanced Learning Technologies,* Kerkrade, The Netherlands.

Brand, S. (1997). *How buildings learn.* London: Phoenix Illustrated.

Brin, S., & Page, L. (2000). *The anatomy of a large-scale hypertextual Web search engine.* Retrieved from http://www-db.stanford.edu/pub/papers/google.pdf

Brusilovsky, P., Chavan, G., & Farzan, R. (2004). Social adaptive navigation support for open corpus electronic textbooks. *Proceedings of AH 2004,* Eindhoven.

Calvin, W.H. (1997). The six essentials? Minimal requirements for the Darwinian bootstrapping of quality. *Journal of Memetics, 1.*

Campbell, A. (2005). *Weblog applications for EFL/ESL classroom blogging: A comparative review.* Retrieved November 30, 2006, from http://www-writing.berkeley.edu/TESL-EJ/ej35/m1.pdf

Churchill, W. (1943). *HC Deb 28 October 1943 c403.* Retrieved from

Cunningham, W. (2006). *Why wiki works.* Retrieved July 19, 2006, from http://c2.com/cgi/wiki?WhyWikiWorks

Darby, J. (2003). *UK eUniversities worldwide: Who we are and what we want from standards.* Retrieved December 14, 2003, from http://www.imsglobal.org/otf/IMS-Darby.pdf

Darwin, C. (1872). *The origin of species* (6th ed.).

Dron, J. (2005). A succession of eyes: Building an e-learning city. *Proceedings of E-Learn 2005,* Vancouver.

Dron, J. (2006a). Social software and the emergence of control. *Proceedings of ICALT 2006,* Kerkrade, The Netherlands.

Dron, J. (2006b). The pleasures and perils of social software. *Proceedings of the 7th Annual Conference of the ICS HE Academy,* Dublin, Ireland.

Dron, J. (2007a). *Control and constraint in e-learning: Choosing when to choose.* Hershey, PA: Idea Group.

Dron, J. (2007b). The safety of crowds. *Journal of Interactive Learning Research, 18*(1), 31-36.

Dron, J., Mitchell, R., & Boyne, C.W. (2003). Evolving learning in the stuff swamp. In N. Patel (Ed.), *Adaptive evolutionary information systems* (pp. 211-228). Hershey, PA: Idea Group.

Giles, J. (2005). *Internet encyclopaedias go head to head.* Retrieved April 12, 2006, from http://www.nature.com/news/2005/051212/full/438900a.html

Grassé, P.P. (1959). La reconstruction du nid et les coordinations inter-individuelles chez Bellicoitermes natalenis et Cubitermes sp. La theorie de la stigmergie: Essai d'interpretation des termites constructeurs. *Insect Societies, 6*, 41-83.

Gregorio, J. (2003). *Stigmergy and the World Wide Web.* Retrieved December 12, 2003, from http://bitworking.org/news/Stigmergy/

Jacobs, J. (1961). *The death and life of Great American cities.* London: Pimlico.

Kauffman, S. (1995). *At home in the universe: The search for laws of complexity.* London: OUP.

Kelly, K. (1998). *New rules for the new economy.* New York: Penguin Group.

Kurhila, J., Miettinen, M., Nokelainen, P., & Tirri, H. (2002). Use of social navigation features in collaborative e-learning. *Proceedings of E-Learn 2002,* Montreal, Canada.

Miettinen, M., Kurhila, J., Nokelainen, P., & Tirri, H. (2005). Our Web-transparent groupware for online communities. *Proceedings of the Conference on Web Based Communities 2005,* Algarve, Portugal.

Moore, M.G., & Kearsley, G. (1996). *Distance education: A systems view.* Belmont: Wadsworth.

O'Reilly, T. (2005, September 30). *What is Web 2.0: Design patterns and business models for the next generation of software.* Retrieved November 30, 2006, from http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

Recker, M.M., Walker, A., & Wiley, D.A. (2000). An interface for collaborative filtering of educational resources. *Proceedings of the 2000 International Conference on Artificial Intelligence,* Las Vegas, NV.

Saba, F., & Shearer, R.L. (1994). Verifying key theoretical concepts in a dynamic model of distance education. *American Journal of Distance Education, 8*(1), 36-59.

Shirky, C. (1996). In praise of evolvable systems. *ACM Net_Worker.*

Sloep, P.B., van Rosmalen, P., Kester, L., Brouns, F., & Koper, R. (2006). In search of an adequate yet affordable tutor in online learning networks. *Proceedings of the 6th International Conference on Advanced Learning Technologies,* Kerkrade, The Netherlands.

Vassileva, J. (2004). Harnessing P2P power in the classroom. *Proceedings of ITS 2004,* Maceio, Brazil.

Webster, A., & Vassileva, J. (2006). Visualizing personal relations in online communities. *Proceedings of AH 2006,* Dublin, Ireland.

KEY TERMS

Collaborative Filter: A form of recommender system that uses implicit or explicit recommendations of others to provide advice based on similarities between user models.

Emergent Behavior: Behavior that arises out of the interactions between parts of a system and which cannot easily be predicted or extrapolated from the behavior of those individual parts.

Latent Human Annotation (LHA): The unintentional communication of a recommendation or other information as a by-product of another process, for example the provision of hyperlinks in a Web page that are then used by search engines to provide rankings of the linked pages.

Recommender System: A computer program that recommends some sort of resource based on algorithms that rely on some sort of user model, some sort of content model, and some means of matching the two.

Social Navigation: The transformation of an interface (usually Web based) by using the actions of visitors.

Social Software: Software in which the group is a distinct entity within the system.

Stigmergy: A form of indirect communication whereby signs left in the environment influence the behavior of others who follow.

Transactional Distance: A measure of the relative amounts of dialogue and structure in an educational activity. Of necessity, as one increases, the other decreases and vice versa. More autonomous learners require less dialogue than more dependent learners.

Semantic Video Analysis and Understanding

Vasileios Mezaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Georgios Th. Papadopoulos
Aristotle University of Thessaloniki, Greece
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Alexia Briassouli
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Michael G. Strintzis
Aristotle University of Thessaloniki, Greece
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

INTRODUCTION

Access to video content, either amateur or professional, is nowadays a key element in business environments, as well as everyday practice for individuals all over the world. The widespread availability of inexpensive video capturing devices, the significant proliferation of broadband Internet connections and the development of innovative video sharing services over the World Wide Web have contributed the most to the establishment of digital video as a necessary part of our lives. However, these developments have also inevitably resulted in a tremendous increase in the amount of video material created every day. This presents new possibilities for businesses and individuals alike. Business opportunities in particular include the development of applications for semantics-based retrieval of video content from the Internet, video stock agencies or personal collections; semantics-aware delivery of video content in desktop and mobile devices; and semantics-based video coding and transmission. Evidently, the above opportunities also reflect to the video manipulation possibilities offered to individual users. Besides opportunities, though, the abundance of digital video content also presents new and important technological challenges, which are crucial for the further development of the aforementioned innovative services.

The cornerstone of the efficient manipulation of video material is the understanding of its underlying semantics, a goal that has long been identified as the "Holy grail of content-based media analysis research" (Chang, 2002). Efforts to understand the semantics of video content typically build on algorithms that operate at the signal level, such as temporal and spatiotemporal video segmentation algorithms that aim at partitioning a video stream into semantically meaningful parts. To support the goal of semantic analysis, these signal-level algorithms are augmented with a priori knowledge regarding the different semantic objects and events of interest that may appear in the video and their signal-level properties. The introduction of a priori knowledge serves the purpose of facilitating the detection and exploitation of the hidden associations between the signal and semantic levels, resulting in the generation of semantically meaningful metadata for the video content.

In this article, existing state-of-the-art semantic video analysis and understanding techniques are reviewed, including a hybrid approach to semantic video analysis that is outlined in some more detail, and the future trends in this research area are identified. The literature presentation starts in the following section with signal level algorithms for processing video content, a necessary prerequisite for the subsequent application of knowledge-based techniques.

BACKGROUND

Segmentation is in general the process of partitioning a piece of information into meaningful elementary parts termed segments. Considering video, the term segmentation is used to describe a range of different processes for partitioning the video into meaningful parts at different granularities (Salembier & Marques, 1999). Segmentation of video can thus be temporal, aiming to break down the video to scenes or shots, spatial, addressing the problem of independently

segmenting each video frame to arbitrarily shaped regions, or spatio-temporal, extending the previous case to the generation of temporal sequences of arbitrarily shaped spatial regions. The term segmentation is also frequently used to describe foreground/background separation in video, which can be seen as a special case of spatio-temporal segmentation. In any case, the application of any segmentation method is often preceded by a simplification step for discarding unnecessary information (e.g., low-pass filtering) and a feature extraction step for modifying or estimating features not readily available in the visual medium (e.g., texture, motion features , etc., but also color features in a different color space, etc.).

Temporal Video Segmentation

Temporal video segmentation aims to partition the video to elementary image sequences termed shots. A shot is defined as a set of consecutive frames taken without interruption by a single camera. A scene, on the other hand, is usually defined as the basic story-telling unit of the video, that is, as a temporal segment that is elementary in terms of semantic content and may consist of one or more shots.

Temporal segmentation to shots is performed by detecting the transition from one shot to the next. Transitions between shots, which are effects generated at the video editing stage, may be abrupt or gradual, the former being detectible by examining two consecutive frames, the latter spanning more than two frames and being usually more difficult to detect, depending among others on the actual transition type (e.g., fade, dissolve, wipe, etc.). Temporal segmentation to shots in uncompressed video is often performed by means of pair-wise pixel comparisons between successive or distant frames or by comparing the color histograms corresponding to different frames. Methods for histogram comparison include the comparison of absolute differences between corresponding bins and histogram intersection (Gargi, Kasturi, & Strayer, 2000). Other approaches to temporal segmentation include block-wise comparisons, where the statistics of corresponding blocks in different frames are compared and the number of "changed" blocks is evaluated by means of thresholding, edge-based and motion-based methods.

Other recent efforts on shot detection have focused on avoiding the prior decompression of the video stream, resulting to significant gains in terms of efficiency. Such methods consider mostly MPEG video, but also other compression schemes such as wavelet-based ones. These exploit compression-specific cues such as macroblock-type ratios to detect points in the 1D decision space where temporal redundancy, which is inherent in video and greatly exploited by compression schemes, is reduced. Regardless of whether the temporal segmentation is applied to raw or compressed video, it is often accompanied by a procedure for selecting one or more representative key-frames of the shot; this can be as simple as selecting by default the first or median frame of the shot or can be more elaborate, as for example in Liu and Fan (2005), where a combined key-frame extraction and object segmentation approach is proposed.

Spatial and Spatio-Temporal Segmentation

Several approaches have been proposed for spatial and spatio-temporal video segmentation (i.e., segmentation in a 2D and 3D decision space, respectively), both unsupervised and supervised. The latter require human interaction for defining the number of objects present in the sequence, for estimating an initial contour of the objects to be tracked or for grouping homogeneous regions to semantic objects, while the former require no such interaction. In both types of approaches, it is typically assumed that spatial or spatio-temporal segmentation is preceded by temporal segmentation to shots and possibly the extraction of one or more key-frames, as discussed in the previous section.

Segmentation methods for 2D images may be divided primarily into region-based and boundary-based methods. Region-based approaches rely on the homogeneity of spatially localized features such as intensity, texture, and position. They include among others the K-means algorithm and evolved variants of it, such as K-Means-with-Connectivity-Constraint (Mezaris, Kompatsiaris, & Strintzis, 2004a), the Expectation-Maximization (EM) algorithm (Carson, Belongie, Greenspan, & Malik, 2002), and Normalized Cut, which treats image segmentation as a graph partitioning problem (Shi & Malik, 2000). Boundary-based approaches, on the other hand, use primarily gradient information to locate object boundaries. They include methods such as anisotropic diffusion, which can be seen as a robust procedure for estimating a piecewise smooth image from a noisy input image (Perona & Malik, 1990). Additional techniques for spatial segmentation include mathematical morphology methods, in particular the watershed algorithm, and global energy minimization schemes, also known as snakes or active contour models.

Regarding spatio-temporal segmentation approaches, some of them rely on initially applying spatial segmentation to each frame independently. Spatio-temporal objects are subsequently formed by associating the spatial regions formed in successive frames using their low-level features (Deng & Manjunath, 2001). A different approach is to use motion information to perform motion projection, that is, to estimate the position of a region at a future frame, based on its current position and its estimated motion features. In this case, a spatial segmentation method need only be applied to the first frame of the sequence, whereas in subsequent frames only refinement of the motion projection result is required (Tsai, Lai, Hunga, & Shih, 2005). A similar approach is followed in Mezaris, Kompatsiaris, and Strintzis (2004b), where the need for motion projection is substituted by a Bayes-based approach to color-homogeneous region-tracking

using color information, and the resulting spatio-temporal regions are eventually clustered to different objects using their long-term motion trajectories.

Alternatively to the above techniques, one could restrict the problem of video segmentation to foreground/background separation. In Chien, Huang, Hsieh, Ma, and Chen (2004), a fast moving object segmentation algorithm is developed, based upon change detection and background registration techniques; this algorithm also incorporates a shadow cancellation technique for dealing with light changing and shadow effects.

Finally, as in temporal segmentation, the spatio-temporal segmentation of compressed video has recently attracted considerable attention. Algorithms of this category generally employ coarse motion and color information that can be extracted from the video stream without full decompression, such as macroblock motion vectors and DC coefficients of DCT-coded image blocks (Mezaris, Kompatsiaris, Boulgouris, & Strintzis, 2004).

SEMANTIC VIDEO ANALYSIS AND UNDERSTANDING TECHNIQUES

The result of pure segmentation techniques, though conveying some semantics, such as the complexity of the key-frame or video, measured by the number of generated regions, or the existence of moving objects in the shot, is still far from revealing the complete semantic content of the video. To alleviate this problem, the introduction of prior knowledge to the segmentation procedure, leading to the development of domain-specific knowledge-assisted analysis techniques, has been proposed.

Prior knowledge for a domain (e.g., F1 racing) typically includes the important objects that can be found in any given image or frame belonging to this domain (e.g., car, road, grass, sand, etc.), their characteristics (e.g., corresponding color models) and any relations between them. Given this knowledge, there exists the well-posed problem of deciding, for each pixel, whether it belongs to any of the defined objects (and if so, to which one) or to none of them.

Depending on the adopted knowledge acquisition and representation process, two types of approaches can be identified in the relevant literature: implicit, realized by machine learning methods, and explicit, realized by model-based approaches. The usage of machine learning techniques has proven to be a robust methodology for discovering complex relationships and interdependencies between numerical image data and the perceptually higher-level concepts. Moreover, these elegantly handle problems of high dimensionality. Among the most commonly adopted machine learning techniques are Neural Networks (NNs), Hidden Markov Models (HMMs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Genetic Algorithms (GAs) (Assfalg, Berlini, Del Bimbo, Nunziat, & Pala, 2005; Zhang, Lin, & Zhang, 2001). On the other hand, model-based video analysis approaches make use of prior knowledge in the form of explicitly defined facts, models and rules, that is, they provide a coherent semantic domain model to support "visual" inference in the specified context (Hollink, Little, & Hunter, 2005).

Knowledge Representation and Ontologies

Ontology, being a formal specification of a shared conceptualization (Gruber, 1993), provides by definition the formal framework required for exchanging interoperable knowledge components. By making semantics explicit to machines, ontologies enable automatic inference support, thus allowing users, agents, and applications to communicate and negotiate over the meaning of information. Typically, an ontology identifies classes of objects that are important for the examined subject area (domain) under a specific viewpoint and organizes these classes in a taxonomic (i.e., subclass/super-class) hierarchy. Each such class is characterized by properties that all elements (instances) in that class share. Important relations between classes or instances of the classes are also part of the ontology. Consequently, ontologies can be suitable for expressing multimedia content semantics so that automatic semantic analysis and further processing of the extracted semantic descriptions is allowed (Hollink et al., 2005). In (Dasiopoulou, Mezaris, Papastathis, Kompatsiaris, & Strintzis, 2005), an ontology was developed for representing the knowledge components that need to be explicitly defined for video analysis. More specifically, domain knowledge is combined with object low-level features and spatial descriptions realizing an ontology- aided video analysis framework. To accomplish this, F-logic rules (Angele & Lausen, 2004) are used to relate the extraction of the semantic concepts, the execution order of the necessary multimedia processing algorithms and the low-level features associated with each semantic concept, thus integrating knowledge and intelligence in the analysis process. The overall system thus consists of a knowledge base, an inference engine, the algorithm repository containing the necessary multimedia analysis tools and the system main processing module, which performs the analysis task, using the appropriate sets of tools and multimedia features, for the semantic multimedia description extraction.

Semantic Video Analysis Approaches

Few semantic video (or, visual content in general) analysis approaches have been presented so far in the literature. Starting with image analysis, a knowledge-guided segmentation and labeling approach for still images is presented in Zhang,

Hall and Goldgof (2002), where an unsupervised fuzzy C-means clustering algorithm along with basic image processing techniques are used under the guidance of a knowledge base. The latter is constructed by automatically processing a set of ground-truth images to extract cluster-labeling rules.

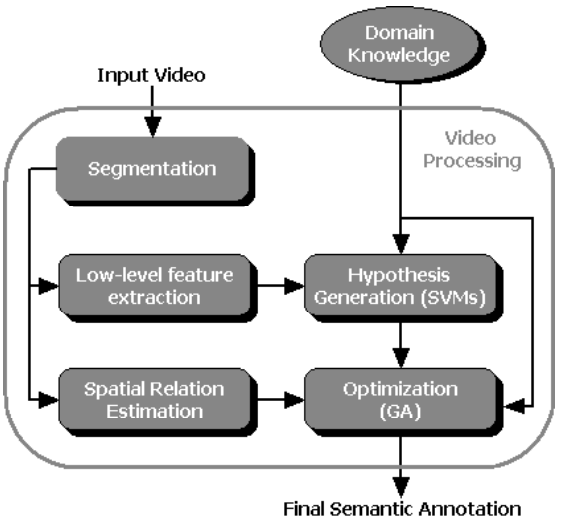
In Naphade, Kozintsev and Huang (2002), the understanding of the semantics of the video content for the purpose of indexing and in particular the association of low-level representations and high-level semantics is formulated as a probabilistic pattern recognition problem and is addressed with the introduction of a factor graph framework. Another approach to video semantic object detection is presented in Tsechpenakis, Akrivas, Andreou, Stamou and Kollias (2002), where semantic entities in the context of the MPEG-7 standard are defined; moving regions are extracted at the signal level by an active contour technique and then their low-level features are matched against those assigned to the previously defined semantic entities, resulting in the identification of associations between the latter and the video segments corresponding to moving objects. In Dasiopoulou et al. (2005), domain knowledge is combined with object low-level features and spatial descriptions realizing an ontology-aided video analysis framework, as described in the previous section. This approach is extended in Voisine et al. (2005), where a genetic algorithm is applied to the atom regions initially generated via simple segmentation in order to find the optimal scene interpretation according to the domain conceptualization. With respect to event detection, existing approaches include among others Sadlier and O'Connor (2005), where a framework for event detection in broadcast video of multiple different field sports is developed, based on robust detectors.

A Hybrid Approach to Semantic Video Analysis

In this section, an example of a semantic video analysis approach that combines two types of machine learning algorithms, namely SVMs and GAs, with explicitly defined domain-specific knowledge in the form of an ontology, is discussed Papadopoulos, Panagi, Dasiopoulou, Mezaris, and Kompatsiaris (2006). SVMs are used for acquiring the implicit knowledge that is required for the analysis process. Additionally, an ontology is developed for representing the knowledge components that need to be explicitly defined, that is, the semantic objects of interest for the selected domain, as well as their spatial relations. The latter appear in the form of fuzzy directional relations, which are used for denoting the relative positions of the depicted real-world objects. The GA is employed for exploiting the aforementioned spatial-related contextual information and deciding upon the final image semantic interpretation.

According to this approach, the video sequence is initially segmented into shots (temporal segmentation) and key-frames are extracted. Then, for every resulting key-frame, spatial segmentation is performed and, subsequently, low-level visual features and fuzzy directional relations are estimated at the region level. Region low-level visual features include the Scalable Color, Homogeneous Texture, Region Shape and Edge Histogram standardized MPEG-7 descriptors. Following the extraction of low-level visual features, SVMs employ them for performing an initial mapping between the image regions and the domain objects in the developed ontology (i.e., generating an initial hypothesis set for every image region). Finally, after the application of SVMs to each region independently, a GA is used to optimize the region-object mapping over the entire image, taking into account the region fuzzy directional relations and the corresponding spatial-related contextual knowledge that is stored in the ontology. This architecture is schematically presented in Figure 1. Application of the proposed approach to video frames of the specified domain results in the generation of fine granularity semantic annotations, that is, segmentation maps with semantic labels attached to each segment. A sample analysis outcome of this hybrid approach is illustrated in Figure 2.

Figure 1. Outline of hybrid semantic video analysis approach



FUTURE TRENDS

In the future, efforts will continue to concentrate on two different subfields of semantic video analysis and understanding: (a) the low-level analysis aiming at effectively decomposing the signal into semantically meaningful entities or parts thereof, without concentrating on understanding

Figure 2. Sample analysis outcome of a hybrid approach to semantic video analysis

Original Frame

Semantic Annotations

the actual semantics of them, and (b) the efficient use of knowledge, including the resolution of issues regarding the efficient acquisition, formulation and representation of it, for understanding the semantics of the signal parts formed after the application of a low-level analysis method. Of particular attention will most certainly be the closer interaction between the two aforementioned subfields: low level image analysis can benefit from the use of some form of knowledge about the content, though this may have to be represented in, for example, the form of a probability distribution (Freedman & Zhang, 2004), and thus may be different in nature from the knowledge used for understanding the semantics of already formed parts of the visual medium. Similarly, the latter task will most probably require the extension of existing knowledge representation formalisms, which were originally designed for tasks far different from the semantic analysis of video or multimedia content in general, so as to address the needs of analysis and provide support for improved reasoning based on the output of the latter.

CONCLUSION

In this article, semantic video analysis and understanding was discussed, starting with a literature review of the elementary task of video segmentation, which constitutes the first necessary step for the analysis and understanding of video content. Following that, the article focused on techniques that go beyond traditional segmentation to incorporate prior knowledge in the analysis procedure, so as to extract a high-level representation of the video content comprising semantic class memberships and recognized video objects. The dominant types of approaches presented in the literature for accommodating this task were identified and a hybrid approach combining machine learning algorithms with explicitly defined knowledge in the form of an ontology was presented in more detail. The future trends identified in the relevant section provide insights on how the algorithms outlined or presented in more detail in this article can be further evolved, so as to more efficiently address the problem of semantic video analysis and understanding and consequently pave the way for the development of innovative video applications.

REFERENCES

Angele, J., & Lausen, G. (2004). Ontologies in f-logic. *International handbooks on information systems.* Berlin, Germany.

Assfalg, J., Berlini, M., Del Bimbo, A., Nunziat, W., & Pala, P. (2005). Soccer highlights detection and recognition using HMMs. In *Proceedings of the IEEE International Conference on Multimedia & Expo (ICME),* (pp. 825-828).

Carson, C., Belongie, S., Greenspan, H., & Malik, J. (2002). Blobworld: Image segmentation using expectation-maximization and its application to image querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(8), 1026-1038.

Chang, S. F. (2002). The holy grail of content-based media analysis. *IEEE Multimedia*, *9*(2), 6-10.

Chien, S. Y., Huang, Y. W., Hsieh, B. Y., Ma, S. Y., & Chen, L. G. (2004). Fast video segmentation algorithm with shadow cancellation, global motion compensation, and adaptive threshold techniques. *IEEE Transactions on Multimedia, 6*(5), 732-748.

Dasiopoulou, S., Mezaris, V., Papastathis, V. K., Kompatsiaris, I., & Strintzis, M. G. (2005). Knowledge-assisted semantic video object detection. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(10), 1210-1224.

Deng, Y., & Manjunath, B.S. (2001). Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(8), 800-810.

Freedman, D., & Zhang, T. (2004). Active contours for tracking distributions. *IEEE Transactions on Image Processing, 13*(4), 518-526.

Gargi, U., Kasturi R., & Strayer, S. H. (2000). Performance characterization of video-shot-change detection methods. *IEEE Transaction on Circuits and Systems for Video Technology, 10*(1), 1-13.

Gruber, T. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Hollink, L., Little, S., & Hunter, J. (2005). Evaluating the application of semantic inferencing rules to image annotation. In Proceedings of the *3rd International Conference on Knowledge Capture (K-CAP05),* Banff, Canada.

Liu, L., & Fan, G. (2005). Combined key-frame extraction and object-based video segmentation. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(7), 869-884.

Mezaris, V., Kompatsiaris, I., Boulgouris, N. V., & Strintzis, M. G. (2004). Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(5), 606-621.

Mezaris, V., Kompatsiaris I., & Strintzis, M. G. (2004a). Still image segmentation tools for object-based multimedia applications. *International Journal of Pattern Recognition and Artificial Intelligence, 18*(4), 701-725.

Mezaris, V., Kompatsiaris, I., & Strintzis, M. G. (2004b). Video object segmentation using bayes-based temporal tracking and trajectory-based region merging. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(6), 782-795.

Naphade, M. R., Kozintsev, I.V., & Huang, T.S. (2002). A factor graph framework for semantic video indexing. *IEEE Transactions on Circuits and Systems for Video Technology, 12*(1), 40-52.

Papadopoulos, G. T., Panagi, P., Dasiopoulou, S., Mezaris, V., & Kompatsiaris, I. (2006, September). A learning approach to semantic image analysis. In *Proceedings of the 2nd International Mobile Multimedia Communications Conference (Mobimedia),* Alghero, Sardinia, Italy.

Perona, P., & Malik, J. (1990). Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(7), 629-639.

Sadlier, D. A., & O'Connor, N. E. (2005). Event detection in field sports video using audio-visual features and a support vector Machine. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(10), 1225-1233.

Salembier, P., & Marques, F. (1999). Region-based representations of image and video: Segmentation tools for multimedia services. *IEEE Transactions on Circuits and Systems for Video Technology, 9*(8), 1147-1169.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(8), 888-905.

Tsai, Y. P., Lai, C. C., Hunga, Y. P., & Shih, Z. C. (2005). A bayesian approach to video object segmentation via merging 3-D watershed volumes. *IEEE Transactions on Circuits and Systems for Video Technology, 15*(1), 175-180.

Tsechpenakis, G., Akrivas, G., Andreou, G., Stamou, G., & Kollias, S.D. (2002, September). Knowledge-assisted video analysis and object detection. In *Proceedings of the European Symposium on Intelligent Technologies, Hybrid Systems and Their Implementation on Smart Adaptive Systems (Eunite02)*, Algarve, Portugal.

Voisine, N., Dasiopoulou, S., Mezaris, V., Spyrou, E., Athanasiadis, T., Kompatsiaris, I., et al. (2005, April). Knowledge-assisted video analysis using a genetic algorithm. In *Proceedings of the Workshop on Image Analysis For Multimedia Interactive Services,* Montreux, Switzerland.

Zhang, M., Hall, L. O., & Goldgof, D. B. (2002). A generic knowledge-guided image segmentation and labeling system using fuzzy clustering algorithms. *IEEE Transactions on Systems, Man and Cybernetics, Part B, 32*(5), 571-582.

Zhang, L., Lin, F.Z., & Zhang, B. (2001, October). Support vector machine learning for image retrieval. In *Proceedings of the International Conference on Image Processing.*

KEY TERMS

Compressed Video Segmentation: Segmentation of video without its prior decompression.

Knowledge-Assisted Analysis: Analysis techniques making use of prior knowledge for the content being processed.

Machine Learning Techniques: Training-based techniques for discovering and representing implicit knowledge, such as complex relationships and interdependencies between numerical image data and perceptually higher-level concepts.

Ontology: Knowledge representation formalism, used for expressing explicit knowledge.

Semantic Video Analysis: Extraction of the semantics of the video, that is, detection and recognition of semantic objects and events.

Spatiotemporal Video Segmentation: Partition the video to elementary spatio-temporal objects, that is, sequences of temporally adjacent arbitrarily-shaped spatial regions.

Temporal Video Segmentation: Partition the video to elementary image sequences termed shots, defined as a set of consecutive frames taken without interruption by a single camera.

Semantic Web and E-Tourism

Danica Damljanović
University of Sheffield, UK

Vladan Devedžić
University of Belgrade, Serbia

INTRODUCTION

Offering tourist services on the Internet has become a great business over the past few years. Heung (2003) revealed that approximately 30% of travelers use the Internet for reservation or purchase of travel products or services.

Classic sites of tourist agencies enable users to view and search for certain destinations and book and pay for vacation packages. At a higher level of sophistication are tourism Web portals, which integrate the offers of many tourist agencies and enable searching from one point on the Web. Still, when using this kind of systems one is forced to spend a lot of time analyzing Web content with destinations that match his/her wishes. This problem is identified by Hepp, Siorpaes and Bachlechner (2006) as the "needle in the haystack" problem.

Applying artificial intelligence (AI) techniques in E-tourism could help resolve this problem by providing:

1. Data that are semantically enriched, structured, and thus represented in a machine readable form;
2. Easy integration of tourist sources from different applications;
3. Personalization of sites: the content can be created according to the user profile;
4. Improved system interactivity.

As an example of using AI in e-tourism, we present *Travel Guides*—a prototype system that offers tourists complete information about numerous destinations. They can search destinations by using several criteria (e.g., accommodation type, food service, budget, activities during vacation, and user interests: sports, shopping, clubbing, art, museum, monuments, etc.). He/She can also read about the weather forecast and events in the destination.

In a way, Travel Guides complements traditional information systems of tourist agencies. These systems require a lot of maintenance effort in order to keep the huge amount of data about tourist destinations up-to-date.

Travel Guides is created to minimize the user's input and his/her need to filter information. It shows how usage of semantically enriched data in a machine readable form can

- Increase interoperability in the area of tourism,
- Decrease maintenance efforts of tourist agents, and
- Offer tourists a better service.

Nowadays, there are just a few e-tourism systems that use AI techniques. We briefly discuss them in the next section. In this article, we explain why it would be good to use such techniques and how Travel Guides does it. Specifically, using Semantic Web technologies in the area of tourism can improve already existing systems (which are mostly available online) that do not use Semantic Web techniques yet. Likewise, the Semantic Web approach can help decrease the maintenance efforts required for existing e-tourism systems and ease the process of searching for vacation packages.

Travel Guides was initially developed as a large-scale expert system. Over time, it has evolved into a modern Semantic Web application.

BACKGROUND

According to Aichholzer, Spitzenberger and Winkler (2003), e-tourism comprises electronic services which include:

- Information services (e.g., destination, hotel information);
- Communication services (e.g., discussion forum);
- Transaction services (e.g., booking).

Transaction services are offered at many places on the Web, such as Expedia, Travelocity, and so forth. These Websites include some of the information services, but for complete details about certain destination (e.g., activities, climate, monuments, and events) one must search for other sources. Some Websites even help in planning the whole itinerary (e.g., HomeAndAbroad). Apparently, there is an "information gap" between these online services, and no interoperability. Semantic Web technologies can be used to overcome this problem and thus increase the quality of e-tourism.

Cunningham (2002) presents GATE (General Architecture for Text Engineering) as an infrastructure for developing and deploying software components that process human

language. It can annotate documents and recognize concepts such as: locations, persons, organizations and dates. GATE can annotate documents with respect to a particular ontology. Some of the recently built Knowledge Management Platforms, like KIM (Popov, Kiryakov, Ognyanoff, Manov, & Kirilov, 2004**),** use GATE for information extraction and retrieval.

Similar to other AI technologies, Semantic Web is not frequently used in real-time tourism applications. Integrating AI tools into mainstream applications can result in benefits to both sides (Djuric, Devedzic & Gasevic, 2007). Standard organizations like the Internet Engineering Task Force and the World Wide Web Consortium (W3C) are making major efforts at developing languages for sharing meaning (Shadbolt, Berners-Lee & Hall, 2006). Speaking at the WWW2006 conference in Edinburgh in May 2006, W3C director Tim Berners-Lee pointed out that Semantic Web has all the standards and technologies it needs to succeed and that it was time for Web developers and content producers to start using semantic languages in addition to HTML (Bennett, 2006).

Cardoso (2006a) addresses the lack of standards in the tourism domain: the prices for tourism activities are expressed in different currencies; the time units also do not follow the standards. He argues that use of Semantic Web and ontologies could overcome this problem. In Cardoso (2006b) he describes the ontology developed to achieve integration and interoperability through the use of a shared vocabulary and meanings for terms with respect to other terms in the area of tourism. His system creates vacation packages dynamically using previously annotated data in respect to the ontology. This is performed with a service that builds itinerary by combining user preferences with flights, car rentals, hotel, and activities in a single price. Similar to this, Jakkilinki, Georgievski and Sharda (2007) presents a tool for tour planning that is intelligent in the meaning that it generates travel plans by matching user preferences and available tourist offers from different travel agents in respect to the ontology which enables reasoning.

To use Semantic Web in e-tourism, two approaches could be applied. One is to make applications from scratch, based on the existing standards. The other one is to enrich already existing content with annotations based on ontology. The first approach is not cost-effective for tourist agencies. Although the second approach sounds more reasonable, it seems that there are not enough data in the domain of tourism on the Web. Hepp et al. (2006) made a research in this field using a sample of 100 accommodations in Austria. Their results showed that neither some of the hotels had their Web pages, nor the biggest Austrian portal for e-tourism (Tiscover) had any information about them. An additional problem they noticed was the incompleteness of the details such as the availability of the accommodation and the prices.

Many e-tourism portals store their data internally, and not on the Web. This means that even a perfect annotation of the Web content would not be sufficient enough; hence, it is limited to persistently published information (Hepp, 2006). To exceed this problem Stojanovic, Stojanovic and Volz (2002) developed a mapping mechanism for migrating relational database schemas into ontologies in order to form the conceptual backbone for metadata annotations which are automatically created from the database. A better approach would be to use Semantic Web services, for example, Web Service Modeling Ontology-WSMO (Roman et al., 2005) or OWL-based Web service ontology - OWL-S (Smith & Alesso, 2005).

Dell'erba, Fodor, Hopken, and Werthner (2005) present the Harmonize project that integrates Semantic Web technologies and merge tourist electronic markets using ontology as a mediator. Their ontology was taken over by the E-tourism Working Group (2004) at Digital Enterprise Research Institute (DERI). This group plans to develop an advanced e-tourism Semantic Web portal, which will connect the customers and virtual travel agents. This portal could be of importance to the travel industry in Austria, whereas for the rest of the world it could be an example of using Semantic Web technologies in a real business system.

In 2001, the industry tried to address the interoperability issue by forming a consortium called the Open Travel Alliance. OTA is producing XML specifications (schemas) for messages to be exchanged between the trading partners, for example, availability checking, booking, rental, reservation, query services, insurance, etc. The precondition for this improvement method to succeed is that each travel agent's application can produce and consume OTA-compliant messages.

Dogac et al. (2004) present the SATINE project as a peer-to-peer network that enables peers to deploy their semantically-enriched travel Web services and allows others to discover these services semantically.

In addition to attempts to semantically enrich tourism sources on the Internet, it has been very popular to develop Location Based Tourism Systems (LBTS). LBTS are computerized systems that depend on an automated location of a target which either deliver or collect information. Currently, LBTS applications are being used by mobile phones, iPods, and PDAs (Hawking et al., 2005). LBTS provide search for hotels and ATM machines near by the user's current location and additional information when the user visits a city for the first time. An example of such a system is "Mobility Agent" (Edwards, Blythe, Scott & Weihong-Guo, 2006). This system delivers Internet-based travel and tourism-related services through fixed and mobile devices. Intelligent agent technology (Devedzic, 2003) was used to provide European visitors the dynamic, mobile, personalized, location-based information and services, especially related to travel in complex urban environments. Kanellopoulos

and Kotsiantis (2006, p. 86) predict that "in a collaborative travel environment, agents will be essential in addressing the issues of security, negotiation, personalisation and Web Service procurement."

TRAVEL GUIDES

Travel Guides is a Semantic Web portal in the area of tourism. Designed with the idea to gather all vacation packages from different tourist agencies, this system is built to help searching for a "perfect" vacation package. It is also personalized, so that its content is adapted to the user regarding his/her interests and activities. It is an example of using Semantic Web to increase the interoperability, provide better interaction, and intelligent reasoning in the domain of tourism.

Requirements for an E-Tourism Intelligent Portal

To make a tourism portal capable of intelligent reasoning, it is necessary to:

- Build some initial knowledge in the system,
- Introduce user profiles, and
- Maintain the knowledge automatically during the user's interaction with the system.

That way, the knowledge collected in a machine-readable form reduces the user's input and improves search.

In addition, the portal should be able to gather useful Web content and extract information of potential interest to the user.

In order to produce knowledge in a machine-readable form, it is necessary to develop and use a set of *ontologies* to represent all important concepts and their relations. Ontologies also enable knowledge sharing and reuse between applications, and development of intelligent Web services by using Semantic Web technologies and tools.

In case of search-intensive applications like tourism portals, ontologies bring up an important feature of *semantic search*. Unlike keyword-based syntactic search, ontology-supported semantic search returns better-quality information, because of the possibility to find the pages that contain semantically similar albeit possibly syntactically different text. A related concept is that of *semantic interoperability*. It is best understood by contrasting it to syntactic interoperability: syntactic interoperability is about parsing the data, while semantic interoperability means using ontologies to define mappings between data and known concepts. For example, a tourist agency may use the term "day trip" on its Web site, whereas another one may use "1-day excursion." Ontology from the domain of tourism may be used to make the equivalence mapping between the two terms, thus enabling the portal to treat the related pieces of information equally.

User Profiles

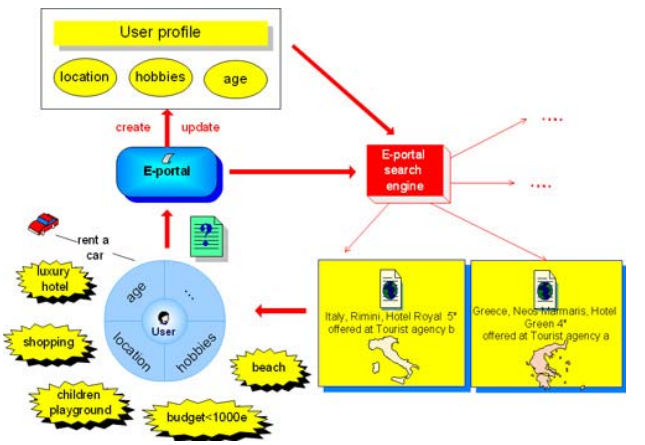
Personalization implies adapting content to a particular user and enhances the interactivity of the system. User profiles are based on the data that the system has about the user. These data are provided in two ways:

1. The user is willing to fill the forms where he/she inputs data about him/herself: gender, birth date, social data, address, profession, education, languages, interests and activities, budget, or visited destinations (cities, countries).
2. The system collects data about the user's interests and preferences while the user is visiting the portal and reading about or searching for vacation packages. Each time the user clicks on some of the vacation package details, the system stores his/her action in a database.

With the user profile, the system can adapt its look-and-feel and, more importantly, its content to the user. For example, the results of search for a destination (based on the requested criteria) can be additionally filtered according to the user profile, Figure 1. Each destination is identified by its latitude and longitude. When the system is aware of the user's location, the search results could be sorted by destinations that are physically closer to the user.

Whenever possible, the user profile extension and maintenance is performed dynamically by the system itself without the user's explicit intervention, using specific heuristics. There is a need to "observe" the user constantly and to use a built-in reasoner to infer the user's preferences and inten-

Figure 1. An adaptive portal using the user's profile



tions from the observations, that is, to create and maintain a valid user profile based on his/her behavior.

The Portal Architecture

Figure 2 depicts the architecture of the Travel Guides portal for tourism management. The *User interface* component collects data from the user and sends them to the *Controller*. *Controller* manages the requests from the interface and fires appropriate actions.

The *User Manager* takes care of the user data. It uses the *User DAO* (Data Access Object) to store or fetch the needed data from/to a database. These data are user details that are not subject to frequent change and are not important for determining the user profile: the username, password, first name, last name, address, birth date, phone and e-mail. The *User Profile Expert* is aware of the *User ontology* and also of the *User profile knowledge base* that contains instances of classes and relations from the *User ontology*.

The *Travel Manager* is responsible for fetching, storing and updating the data related to vacation packages. Storing and retrieving data from the database is performed by the *Vacation Package DAO*. The data stored in the database are those that are subject to frequent changes and are not important in the process of reasoning: start date, end date, prices (accommodation price, food service price, transport price), benefits, discounts, and documents that contain textual descriptions with details about the vacation packages. Some of these data are used in search queries as constraints. Actually, retrieving a "perfect" vacation package is performed in two steps:

1. Matching the user's wishes to certain destinations—the user profile is matched with certain types of destinations. This is performed by the *Travel Offer Expert* and the *World Expert*.
2. The list of destinations retrieved in the first step is filtered using the constraints the user may want to set (e.g., the start/end dates of the vacation). The filtering is done by the *Vacation Package DAO*.

The *Travel Manager* aggregates the *Tourist Offer Expert* and *World Expert* components, which include inference engines. These inference engines are aware of the *Travel* and *World ontology,* respectively. Their main role is to validate the content in the knowledge bases.

The *System Scheduler* is running daily at scheduled times. It takes data about the type of destinations that the user has visited and updates his/her user profile. Those data are fetched from the logged user's activities in the relational database, semantically processed and stored in a temporal knowledge base. After the temporal knowledge base is validated, it is merged with an already existing user-profile knowledge base. This validation is performed by an inference engine built inside the *User Profile Expert*.

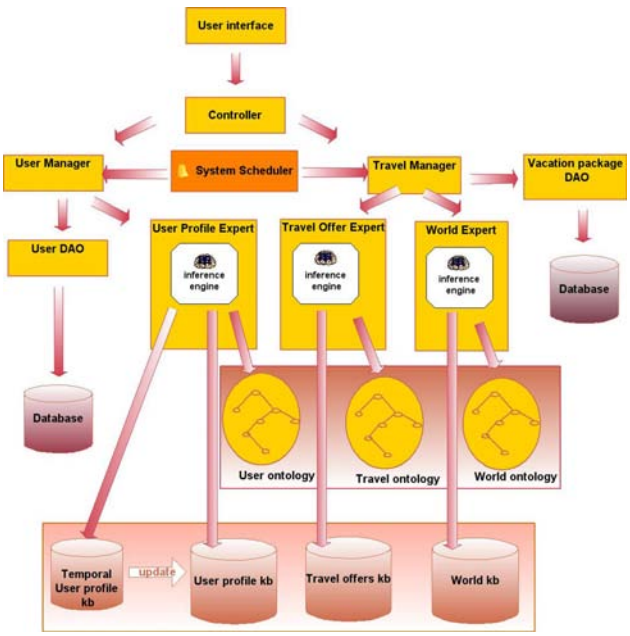
A similar process is performed to update tourist offers in the knowledge base. This happens when a travel agent updates data about new vacation packages, as well as when the portal administrator updates the *World knowledge base* (see Figure 2).

The relations and restrictions stored in Travel Guides ontologies represent the core of this system. Ontologies allow machine-supported travel-related data interpretation and integration (Kanellopoulos, Kotsiantis, & Pintelas, 2006).

The *World ontology* contains concepts and relations from the real world: geographical terms, locations with coordinates, land types, time and date, time zone, currency, languages, and all other terms that are expressing some concepts that are in a way related to tourism or tourists, but not to vacation packages that could be offered by tourist agencies. Because Travel Guides is intended to support semantic annotation, indexing, and retrieval of documents, this ontology is also meant to contain the general concepts necessary for expressing semantic annotation, indexing, and retrieval.

The *User ontology* is meant to contain data about the users–the travelers who visit the Travel Guides portal. This ontology describes user interests and activities, age groups, favorite travel companies, and other data related to user profiles. Each user can have one or more user profiles, depending on his/her behavior while visiting the portal, and also depending on the data he/she has provided to the system.

Figure 2. The architecture of the Travel Guides portal for tourism management



The *Travel (Tourism) ontology* includes all terms being specific to types of vacations, traveler types, and vacation packages offered in tourist agencies and being important to travelers, like the type of accommodation, food service type, transport service, type of room in a hotel, and the like. It is an ontology that makes an indirect connection between users and destinations.

The ontologies are implemented in OWL (Antoniou & Harmelen, 2004) and developed using the Protégé tool (Horridge, Knublauch, Rector, Stevens, & Wroe, 2004).

Travel Guides is designed so that all users contribute to the creation and updating of the knowledge base. Each group is contributing to the knowledge base in its own way:

- End users (tourists) feed the system with their personal data, which then get analyzed by the system in order to create/update user profiles. The system also uses logged data about each user's activities (mouse clicks) when updating the user's profile.
- Tourist agents are creating vacation packages and similar offers in tourist agencies. They feed and update the knowledge base with new information about destinations, arrangements, excursions, etc. To do this, they fill a form about a destination which includes fields like name, accommodation, hotel name, parking provided by the hotel, swimming pool inside the hotel, activities, transport service, etc.
- Portal administrators—they mediate the knowledge base updates with destinations not covered by the tourist agencies. The idea is that tourist agents can then use those updates as the basis for creating new vacation packages and other tourist offers. To alleviate the creation, extensions, and maintenance of this part of the knowledge base, Travel Guides exploits WordNet (Fellbaum, 1998), an open-source knowledge base that describes a number of concepts and terms. Some of these were copied to the Travel Guides knowledge base to avoid manual entrance of static (permanent) data about various destinations. A special tool is developed for copying instances of WordNet classes (concepts) and relations to instances of classes and relations of the Travel Guides World ontology.

FUTURE TRENDS

There is room for future improvements in several directions:

- The use of the Semantic Web services.
- The system is designed to possibly include information about all places in the world and all vacation packages. To achieve this, it is necessary to feed it and save a great amount of data. The system's prototype described here includes a limited collection of vacation packages, which must be extended and updated.
- Simplification of the document annotation process. Currently, it requires the knowledge and understanding of the GATE.
- Finally, Travel Guides supports only management of hotel accommodation. This should be expanded with hostels, campgrounds, and private apartments.

CONCLUSION

Artificial intelligence can play a significant role in improving e-tourism portals. The use of Semantic Web technologies can increase interoperability in the area of tourism, although the agreement of all involved parties about using the standards is essential. Without this agreement, the only useful thing that could be achieved with annotation process would be to analyze sites with content about countries, cities, beaches, and so forth. Most of the data regarding vacation packages are internally stored in the tourism portals, and hence only the usage of intelligent Web services would help their retrieval and representation in a machine readable form. Standardization of the way all tourist services are representing the data would speed up the process of their integration. This would ease searching for tourist deals from one place. The integration of geographical data would also decrease efforts of tourist agents who are responsible to feed the system and keep data up-to-date. If these data would be centralized in a repository available to tourist agents, this would significantly decrease the maintenance efforts.

The Travel Guides system is built to solve the problem of distributed tourist sources and to help users find a "perfect" vacation package quickly. The system includes intelligent components that perform reasoning and have some built-in heuristics. It is based on a number of ontologies, which support the process of adding semantics to data. The data are semantically enriched, which means that they can be understood by machines. The system is also personalized in order to adapt its content to each user. The more a user visits the portal, the more the system "knows" about him/her.

As this system is built using the latest Java technologies for building Web applications widely used in existing online tourism information systems nowadays, this prototype system is to show that such mainstream systems can benefit from integrating the Semantic Web components.

REFERENCES

Aichholzer, G., Spitzenberger, M., & Winkler, R. (2003, April). *E-tourism strategic guideline 6. PRISMA–providing innovative service models and assessment*. Vienna: Institute of Technology Assessment, Austrian Academy of Sciences.

Retrieved December 9, 2007, from: http://www.prisma-eu.net/deliverables/sg6tourism.pdf

Antoniou, G., & Harmelen, F. V. (2004). Web ontology language: OWL. In S. Staab & R. Studer (Eds.), *Handbook on ontologies (International Handbooks on Information Systems)* (pp. 67-92). Springer-Verlag.

Bennett, J. (2006, May 25). The Semantic Web is upon us, says Berners-Lee. *Silicon.com research panel: WebWatch*. Retrieved December 9, 2007, from http://networks.silicon.com/webwatch/0,39024876,39159122,00.htm

Cardoso, J. (2006a). Developing dynamic packaging systems using Semantic Web technologies. *Transactions on Information Science and Applications*, *3*(4), 729-736.

Cardoso, J. (2006b). Developing an owl ontology for e-tourism. In J. Cardoso & P. A. Sheth (Eds.), *Semantic Web services, processes and applications* (pp. 247-282). Springer-Verlag.

Cunningham, H. (2002). GATE, a general architecture for text engineering. *Computers and the Humanities*, *36*(2), 223-254.

Dell'erba, M., Fodor, O., Hopken, W., & Werthner, H. (2005). Exploiting Semantic Web technologies for harmonizing e-markets. *Information Technology & Tourism, 7*(3-4), 201-219.

Devedzic, V. (2003). *Intelligent information systems*. Digit, FON, Beograd (in Serbian).

Djuric, D., Devedzic, V., & Gasevic, D. (2007). Adopting software engineering trends in AI. *IEEE Intelligent Systems*, *22*(1), 59-66.

Dogac, A., Kabak , Y., Laleci, G., Sinir, S., Yildiz, A. & Tumer, A. (2004, October 27-29). *SATINE project: Exploiting Web services in the travel industry*. eChallenges 2004 (e-2004), Vienna, Austria.

Edwards, S. J., Blythe, P. T., Scott, S., & Weihong-Guo, A. (2006). Tourist information delivered through mobile devices: Findings from the image. *Information Technology & Tourism*, *8*(1), 31-46.

E-tourism Working Group. (2004, October). *Ontology collection in view of an e-tourism portal*. Retrieved December 9, 2007, from http://138.232.65.141/deri_at/research/projects/e-tourism/2004/d10/v0.2/20041005/

Fellbaum, C. (1998). *WordNet—an electronic lexical database*. The MIT Press.

Hawking, P., Stein, A., Zeleznikow, J., Pramod, S., Devon, N., Dawson, L., & Foster, S. (2005). Emerging issues in location based tourism systems. In *Proceedings of the International Conference on Mobile Business (ICMB'05),* (pp. 75- 81). IEEE Computer Society.

Hepp, M. (2006). Semantic Web and Semantic Web services: Father and son or indivisible twins? *Internet Computing, IEEE, 10*(2), 85- 88.

Hepp, M., Siorpaes, K., & Bachlechner, D. (2006, June 12-14). Towards the Semantic Web in e-tourism: Can annotation do the trick? In *Proceedings of the 14th European Conference on Information System (ECIS 2006),* Gothenburg, Sweden.

Heung, V.C.S. (2003). Internet usage by international travelers: Reasons and barriers. *International Journal of Contemporary Hospitality Management*, *15*(7), 370-378.

Horridge, M., Knublauch, H., Rector, A., Stevens, R., & Wroe, C. (2004, August). *A practical guide to building OWL ontologies using the Protege-OWL plugin and CO-ODE tools edition 1.0.* The University of Manchester. Retrieved December 9, 2007, from http://protege.stanford.edu/publications/ontology_development/ontology101.html

Jakkilinki, R., Georgievski, M., & Sharda, N. (2007). Connecting destinations with an ontology-based e-tourism planner. In M. Sigala, L. Mich, & J. Murphy (Eds.), Information and communication technologies in tourism 2007. In *Proceedings of the International Conference in Ljubljana, Slovenia, 2007,* (pp. 21-31). Vienna: Springer.

Kanellopoulos, D., & Kotsiantis, S. (2006). Towards intelligent wireless Web services for tourism. *IJCSNS International Journal of Computer Science and Network Security, 6*(7), 83-90.

Kanellopoulos, D.,Kotsiantis, S., & Pintelas, P. (2006). Intelligent knowledge management for the travel domain. *GESTS International Transactions on Computer Science and Engineering*, *30*(1), 95-106.

Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., & Kirilov, A. (2004). KIM—a Semantic platform for information extraction and retrieval. *Journal of Natural Language Engineering, 10*(3-4), 375-392. Cambridge University Press.

Roman, D., Keller, U., Lausen, H., Bruijn J. D., Lara, R., Stollberg, M., et al. (2005). Web service modeling ontology. *Applied Ontology*, *1*(1), 77-106.

Shadbolt, N., Berners-Lee T., & Hall, W. (2006). The Semantic Web revisited. *IEEE Intelligent Systems, 21*(3). 96-101.

Smith, C. F., & Alesso, H. P. (2005). *Developing Semantic Web services*. A K Peters.

Stojanovic, L.J., Stojanovic, N., & Volz, R. (2002). Migrating data-intensive Web sites into the Semantic Web. In *Proceed-*

ings of the 2002 ACM Symposium on Applied Computing, Madrid, Spain, (pp. 1100-1107). ACM Press.

KEY TERMS

Dynamic Packaging: The combination of different travel components, bundled and priced in real time, in response to the requests of the consumer or booking agent.

Intelligent Agents: Software elements which help the user find information of specific interest to him/her without their explicit assistance.

Intelligent Reasoning: The act of using reason to derive a conclusion from certain premises using a given methodology.

Location Based Services: Services that provide context-sensitive information based on the mobile user's location.

Ontology: A controlled vocabulary that describes objects and the relations between them in a formal way, and has a grammar for using the vocabulary terms to express something meaningful within a specified domain of interest.

OWL-Based Web Service Ontology (OWL-S): An ontology which supplies Web service providers with a core set of constructs for describing the properties and capabilities of their Web services in unambiguous, computer-interpretable form.

Semantic Web Services: Self-contained, self-describing, semantically marked-up software resources that can be published, discovered, composed and executed across the Web in a task-driven semiautomatic way.

Web Service Modeling Ontology (WSMO): A data model that provides the conceptual underpinning and a formal language for semantically describing all relevant aspects of Web services in order to facilitate the automation of discovering, combining and invoking electronic services over the Web.

Web Portal: A Web site or service that offers a broad array of resources and services, such as e-mail, forums, search engines, and online shopping malls.

Semantic Web in E-Government

Mamadou Tadiou Koné
Université Laval, Canada

William McIver Jr.
National Research Council and Institute for Information Technology, Canada

INTRODUCTION

Today, in many countries, looking for government information, filing taxes, renewing a driver's license, obtaining a certificate and notifying of a new address anytime, anywhere are becoming mundane online operations. For the satisfaction of their constituents, local governments are striving to deliver more effective and efficient online services through the use of innovative information and communications technologies.

E-government also known as "digital government" can be defined as the civil and political conduct of government using **information and communication technologies** (ICT) (McIver & Elmagarmid, 2002). The most accepted picture of e-government is that of a provider of online services to citizen (G2C), businesses (G2B) and the administration (G2G). The real value of an e-government rests on the effectiveness of its programs, the broad availability of its enhanced online services, the satisfaction of customers and the tangible savings in time, money and human resources (Koné, 2005).

E-government expansion and adoption by communities, citizens, businesses, and public administrations in most countries is generally seen as a four-step process: presence phase, interaction phase, transaction phase, and transformation phase. The goal of the last transformation phase is to integrate several internal services at the vertical and horizontal levels, into a one-stop, whole-of-government with innovative services operating seamlessly across departments, agencies and programs. To address the problems of **seamless integration** and **interoperability** (D'Auray, 2001), some actors in e-government are experimenting with the **semantic Web** promoted by **Tim Berners-Lee** (Berners-Lee et al., 1999, 2001), **Web service** technologies (McIlraith et al., 2001) as well as **service oriented architecture** (SOA) as a means for achieving integration and inter-operation in the service transformation phase.

Scope and Structure of the Article

This chapter aims at presenting the semantic Web technology applied to the transformation and advancement of e-government. After this introduction in the first section, we expose in the second section the nature of the semantic Web and e-government. Then, we explain in the fourth section, how semantic Web technologies can contribute to solving known issues in the transformation of e-government. Given this background, we are able to propose a simple illustration of our ideas: Web services and semantic Web-based architectures within the e-government project of Québec, Canada. We then give a glimpse of some future trends in the fourth section and the conclusion in the fifth and last section.

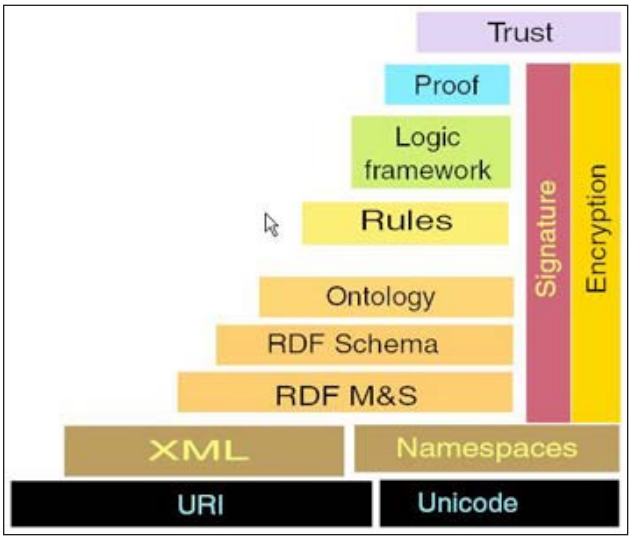
BACKGROUND

Semantic Web Technology

If the **Internet** is said to be "*the place where one can find anything,"* there are still concerns about really finding what one is looking for. In this context, Berners-Lee proposes (Berners-Lee et al., 1999, 2001) a whole new vision of the Web called the **semantic Web**. This semantic Web is "*an extension of the current Web, in which information is given well-defined meaning, better enabling computers and human to work in cooperation*" (Berners-Lee *et al.*, 2001). In line with this vision of the semantic Web, the **World Wide Web Consortium** (W3C) has developed a number of ontological languages for specific purposes. The following languages (Asuncion & Corcho, 2002) displayed in Figure 1 are called ontology languages because they can formally describe the meaning of terms and relations in Web documents.

- **The resource description framework** (RDF) is a flexible data model for resources described as objects and the relations among them. It provides a simple semantics for this data model, and these data models can be represented in XML syntax;
- **RDF schema** is a vocabulary for describing properties and classes of RDF resources, with semantics for hierarchies of such properties and classes;
- The **Darpa agent modeling language** (DAML) has been developed as an extension of XML and RDF. It

Figure 1. The Semantic Web 'layer cake' as proposed by Berners-Lee



is used to explicitly represent the meaning of terms in vocabularies and the relationships between these terms;

- The **ontology Web language** (OWL) is intended to provide a language suitable for describing the classes and relations inherent to Web documents. OWL has more facilities for expressing meaning than XML, RDF and RDF-S.

A more detailed and in-depth description of the semantic Web technology is given in "Semantic Web fundamentals" by Antoniou and Plexousakis in the Encyclopedia of Information Science and Technology, 2005.

E-Government

Considerable progress has been made in e-government over the past decade. A recent major study, in which 21 governments were surveyed, showed that e-government has created major changes along several dimensions: services, modes of operation, and organizational structures (Accenture, 2006). In particular, unified contact centers have been created to help government provide single entry points for citizen services.

Two broad classes of e-government technologies exist (Ashley, 1999). As seen in Table 1, one class comprises externalizing systems which provide interfaces to government entities through which citizens and other government entities can obtain services. The level of service of this range from one-way information delivery to complex transactional interactions, whereby legally-binding tasks, such as vehicle registration, can be completed. Another class comprises systems which provide:

Table 1. A summary of e-government characteristics

• Civil and political conduct of government using ICT; • Provision of online services to citizen (G2C), businesses (G2B) and the administration (G2G); • Two broad classes: externalizing and internalizing systems; • Evolution through presence, transaction and transformation phases.

1. Integrative communication functionality to improve intra-governmental workflows;
2. Domain-specific processing and knowledge management, such as data mining for public health or support for law enforcement investigations.

The needs and trends in e-government parallel those of the broader computing community with respect to semantic Web research. The current generation of research in e-government reflects an effort to make:

1. Services more widely accessible;
2. Services more integrated within organizations; and
3. Information more "intelligent" (Cencioni & Bertolo, 2006).

In a government context, the accessibility and integration of services is being addressed through Web services and business processes. Bringing intelligence to information has involved the injection of semantics into content as meta-data, largely XML-based, and corresponding processing techniques that allow those meta-data to be interpreted.

SEMANTIC WEB IN E-GOVERNMENT

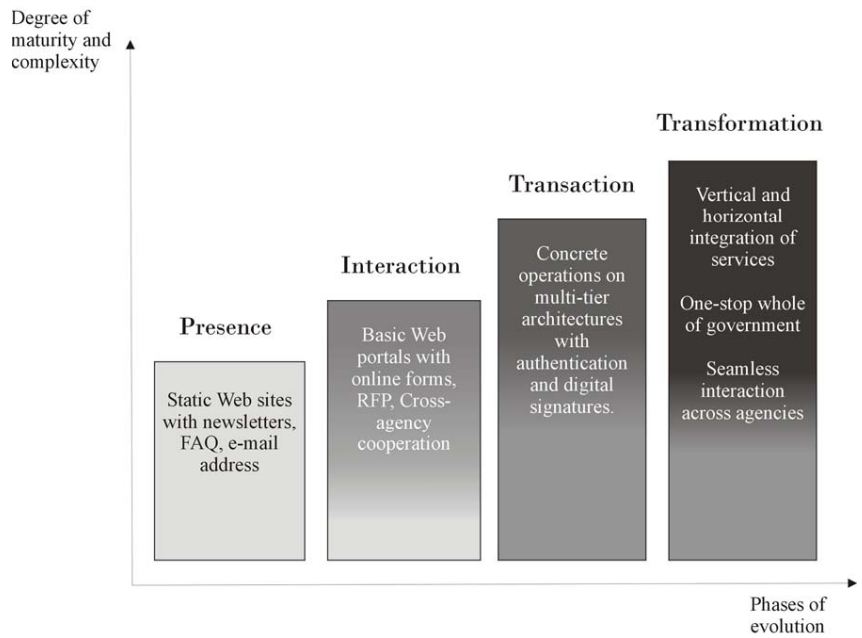
There is a need within e-government services to provide information whose format and methods of delivery are adapted to users and situations (Accenture, 2006). In its evolution, e-government is expected to format information from a given knowledge domain in different ways when presented to senior citizens, youth, or government officials.

E-Government Evolution

E-government expansion and adoption by communities, citizen, businesses and public administrations in most countries is generally seen (Government of Canada, 2003) as a four-step process: presence phase, interaction phase, transaction phase and transformation phase.

The initial **presence phase** is implemented through the publication on the Web of static information on government operations and services. Starting with few services, the initiative expands to a broad range of services with basic

Figure 2. E-government maturity model based on Gartner Research 2000



capabilities like official publications, newsletters, e-mail contact and a FAQ section.

The **interaction phase** appears through the building of basic Web portals containing online forms, requests for proposal and opinion surveys on critical issues of interest to citizen and local businesses. In addition, some kind of cross-agency cooperation appears: government agencies start to reach out to one another through links in their official Web sites.

The **transaction phase** offers online operations like a driver's license renewal, a car registration, a request for a new passport and requires payment of fees in a complete and secure online setting. The proper implementation of these online transactions draws much from the technical aspects of similar transactions in eCommerce. An e-government architecture, at this stage, uses complete multi-tier architectures with authentication and digital signatures.

The **transformation phase** aims at integrating several internal services at the vertical and horizontal levels, into a one-stop, whole-of-government with innovative services operating seamlessly across departments, agencies and programs. These services are therefore tailored to the needs of businesses, communities and citizens.

Current Challenges in E-Government

Significant challenges remain in e-government, for which S**emantic Web** technologies might offer solutions. The coming generation of research (Oreste, 2005) will focus on moving from intelligent information to information that is "actionable" (Bertolo, 2006). The Seventh Framework Programme of research of the European Commission has taken this view after extensive consultations with experts in this area. Semantic Web research directions that are relevant to e-government include:

1. Social networking;
2. Service composition and collaborative workflow;
3. Security and trust;
4. Automated collection and processing information; and
5. Adaptive information delivery.

As countries like United States, Denmark, Australia, Finland, United Kingdom, Germany, Ireland, and France are making remarkable progress in mature service delivery, questions arise about the appropriate technological platform for reaching the next level of service transformation. Until now, to design a platform for its service delivery, some government have quietly followed in the footsteps of the successful e-commerce where Web services technology with its Web services description language (WSDL), simple object access protocol (SOAP) and uniform description, discovery and integration (UDDI) protocols, appears to be the most popular. However, government providers not only have *different goals* in the design, organization, management and delivery of services than the private sector but face *massive challenges under different constraints*. According to D'Auray (2001), the most important of these challenges are safety, security, and integrity of online interactions

with government; privacy and confidentiality of personal and business information within government; information management with respect to accuracy and relevance when merging data across departments and agencies. Although security technology for Web services addresses the safety, integrity and confidentiality concerns through cryptography, digital certificates and trusted third-party authorities, it is neither convenient nor stable enough to inspire trust. In addition, current research has shown that it is possible to automatically discover, select and compose Web services on a syntactic, semantic and pragmatic (location, QoS, policy) levels. However, to date, no Web service infrastructure has the capability of dealing with laws (e.g., the *Privacy Act in Canada* mentioned in D'Auray, 2001) related to the use and sharing of personal information where the explicit written consent of a subject is required.

In Europe, there are a number of e-government projects using semantic Web technology. Access to e-government services employing semantic technologies (**Access-eGov**) is a project which aims at increasing the accessibility of public administration services for citizens and business users by supporting the interoperability among existing electronic and 'traditional' government services. Access-eGov is partially funded under the IST Programme of FP6 (e-government research). Its main objective according to the project designers is to create a server reference ontology covering basic domain knowledge and processes; rule-based editorial add-on component for Web sites and Web applications to insert semantic mark-up within public e-Gov applications.

SemanticGov (http://www.semantic-gov.org) is a project which aims at building the infrastructure (software, models, services) to support **semantic Web services** for interoperability across local or transborder public administrations (PA). It is based on the Service Oriented Architecture and Semantic Web Services technologies for enabling the discovery and execution of complex PA services.

e-govRTD2020 is a project co-funded by the European Commission under the 6th Framework Programme of IST. It aims at sketching e-government in 2020 through the identification of future strategic research fields in e-government. This project intends to use ontologies and well known knowledge management tools to provide information quality and economy.

The Online Address Change Service of Quebec

In Quebec, Canada, online services are organized according to visitors' profile (citizen or enterprise) and gathered according to specific topics such as finances, industry, business, education, employment, and legal matters. To have access to online government services, one must visit a constellation of Web sites whose architectures does not allow integrated service provision like a one-time change of address in several administrations. In addition, the personal data protection law requires each government organization in Québec to build and manage its own secure online database with no possibility of sharing personal information on citizen and businesses. Unfortunately, each of these organizations has its own concept of identification with rules and restrictions. The best example of this problem is the online address change Service of Quebec[1]. Here, the process of changing one's address has been made easy by a single online form which serves six ministries and agencies simultaneously. Then, each destination still bears the burden of validating the address changed. In the province of Québec, official records show that about 66 % of online address change requests through this traditional system is rejected during validation. Therefore, we proposed the design and implementation of an **e-government Web service** platform described in figure 3 (Ben Fadhel & Koné, 2005). This platform supports functionalities and modules made of a Web service-based portal which plays the role of online middle man; the services request folder, an emulation of an electronic commerce shopping cart system; the service search module as a channel to available government services and an online address change service.

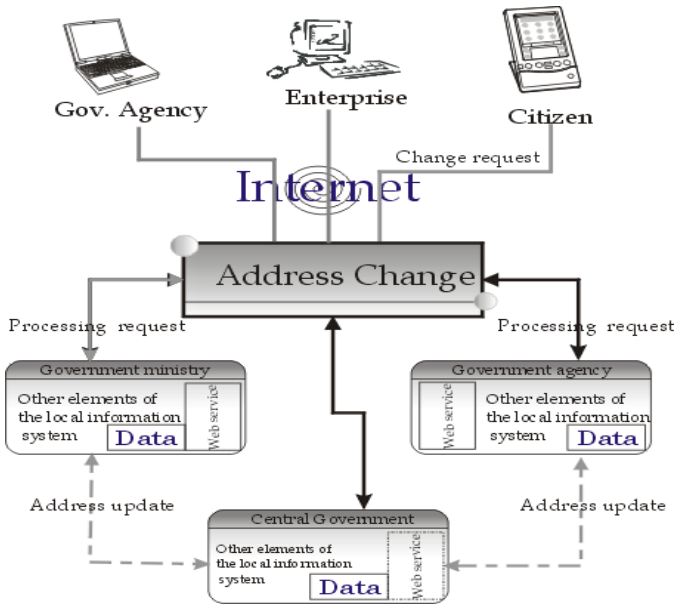
We designed an appropriate ontology for government services with the Protégé ontology editor to support these operations. With the local information at ministries and agencies being processed as Web services, our platform is able to efficiently collect and process data then display query results to consumers. Here, as required by the law, there is no need for interaction between the players.

A similar approach using semantic Web services technologies has been used by Medjahed, Bouguettaya, & Ouzzani (2003) for the automatic selection, interoperation and composition of e-government services in the context of the Family and Social Services Administration (FSSA) of their community.

FUTURE TRENDS

There are a number of semantic Web technologies which promises to bring significant innovations to the field of e-government. One remarkable example is the **service oriented architecture** or SOA based on semantic Web services (Erl, 2005). SOA is a reliable and relatively simple infrastructure which is flourishing and promises to spark greater data integration and interoperability between heterogeneous systems in administrations. It can articulate a process independent of any technology and allow the coordination and use of a collection of Web services. A number of e-government programmes (e.g., the state of Kentucky, in the United States, and Dubai Municipality) has already adopted this paradigm to support their ongoing effort in service transformation. Fortunately, SOA seems to help avoid the duplication of infrastructure and data across agencies and ministries. Therefore, coher-

Figure 3. Proposed structure of the online address change service



ence, accuracy of information, and trust among citizen are the first benefits in implementing this paradigm.

CONCLUSION

In this chapter, we first explained what e-government is and the state of its evolution. Then, we introduced the field of semantic Web applied to e-government to overcome the obstacles in achieving the last service transformation phase. As an illustration, we presented a couple of noteworthy ongoing projects in Europe called Access-eGov, SemanticGov and eGovRTD2020. Then, we presented an interesting online address change application that we suggested to the e-government of Québec in Canada to improve its architecture.

REFERENCES

Accenture. (2006). Leadership in Customer Service: Building the Trust, *Accenture's Seventh Global Report on Government Service Delivery.* Retrieved from http://www.accenture.com/countries/canada/research_and_insights/leadership-delivery.htm

Antoniou, G., & Plexousakis, D. (2005). Semantic web fundamentals. In *Encyclopedia of information science and technology*. Information Resources Management Association Press.

Ashley, K. (1999). Preserving the History of Government Computing: Social and Technological Change. Digital Resources for the Humanities 1999 (DRH'99). University of Edinburgh, Scotland. September 12-15, 1999.

Asunción, G., & Corcho, O. (2002). Ontology languages for the semantic web. *Intelligent IEEE Systems*, *17*(1), 54-60.

Ben Fadhel, D., & Koné, M.T. (2005). An e-government web services platform on the semantic Web. *Proceedings of EGOV05, International Conference on e-government 2005*, Copenhagen, Denmark, pp. 143-149.

Berners-Lee T., & Fischetti, M. (1999). *Weaving the web: The original design and ultimate destiny of the world wide web by its inventor*. San Francisco: Harper

Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34-43.

Cencioni, R., & Bertolo, S. (2006). From Intelligent Content to Actionable Knowledge: Research directions and opportunities under the EU's Framework Programme 7*, 2007-2013.* Talk at SAMT 2006, Athens, Greece.

D'Auray, M. (2001). Behind the portal: Revealing the challenges of Service integration. *ICA Information No. 74: General Issue*. 1-13.

Erl, T. (2005). *Service-oriented architecture (SOA): Concepts, technology, and design.* Prentice Hall.

Government of Canada. (2003). Connecting with canadians: Pursuing service Transformation. *Final Report of the Government On-Line Advisory Panel*. Retrieved from http://www.gol-ged.gc.ca/pnl-grp/reports

Kone, M.T., Ben Fadhel, J., & Msaid, A. (2006) A critical step in e-government evolution. In Proceedings of the AAAI Symposium Semantic Web meets e-government, (pp. 64-69).

McIlraith, S., et al. (2001). Semantic Web Services. *IEEE Intelligent Systems*, 46-53.

McIver, Jr., William J. and Elmagarmid, Ahmed K. (eds) (2002). *Advances in Digital Government: Technology, Human Factors, and Policy*. Boston: Kluwer.

Medjahed, B., Bouguettaya, & A. Ouzzani, M. (2003). *Semantic Web Enabled E-Government Services*. The dg.o 2003 NSF Conference for Digital Government Research, Boston, USA.

Signore, O., Chesi, F., & Palloti, M. (2005). E-government challenges and opportunities. In *Proceedings of the CMG Italy XIX annual conference*, Florence, Italy, Retrieved from http://www.w3c.it/papers/cmg2005Italy.pdf

KEY TERMS

E-Government Web Services: E-government applications deployed over the Web within a Web service infrastructure. This technology becomes really useful only when an ensemble of related distributed e-government services are composed in order to create a new one.

Interoperability: The ability of several software components based on different platforms to interact, exchange services and cooperate in solving complex tasks. ISO TC204 defines interoperability as "*the ability of systems to provide services to and accept services from other systems and to use the services so exchanged to enabled them to operate effectively together."*

Ontology: Originally used in philosophy to refer to the kind of things that exist, an ontology is interpreted as "*a specification of a conceptualization*" (Tom R. Gruber) in the context of artificial intelligence. In practical terms, an ontology is the set of terms of a vocabulary about a given domain and all the relationships between these terms. It can be written as an RDF document with classes and properties available for creating instances and making assertions.

Semantic Web: The best and most well known definition of the semantic Web is given by its inventor, Tim Berners-Lee in the May, 2001 issue of Scientific American as "*The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation.*" To achieve this goal, a data model called resource description framework (RDF), several data interchange formats like RDF/XML and N3, notations called RDF schema (RDFS) and the Web ontology language (OWL) have been developed and proposed by the World Wide Web Consortium (W3C) to give formal descriptions of concepts, terms, and relationships in a domain.

Semantic Web Services: When Web services and their related messages are semantically described (capabilities, interfaces) with appropriate ontologies, they are called semantic Web services.

Service Oriented Architecture (SOA): is literally an architecture which relies on service-orientation. It is a reliable and relatively simple infrastructure which allows greater data integration, interoperability and the coordination of a collection of heterogeneous systems. OASIS (the Organization for the Advancement of Structured Information Standards) defines SOA as: "A paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains. It provides a uniform means to offer, discover, interact with and use capabilities to produce desired effects consistent with measurable preconditions and expectations."

Web Services: Web services are a set of protocols named Web services description language (WSDL), uniform description, discovery and integration (UDDI) and simple object access protocol (SOAP) used to exchange data between applications regardless of their platform, language or object model. In this interaction, there are three actors:

1. A service provider defines with the WSDL language the format for request and response of services it generates;
2. A UDDI registry stores the services descriptions published by the service provider;
3. A service consumer in need can make a request and find a particular service description in the UDDI registry. It subsequently calls this service through the SOAP protocol and requires it to perform some action at the provider's location and send back the result.

Current examples Web services are weather information service, authentication service, Foreign exchange service and Knowledge base service.

World Wide Web: A system of interlinked multimedia documents distributed over the Internet created by Tim Berners-Lee around 1990.

ENDNOTE

[1] Service Québécois de Changement d'Adresse: SQCA (in French).

Semantic Web Uncertainty Management

Volker Haarslev
Concordia University, Canada

Hsueh-Ieng Pai
Concordia University, Canada

Nematollaah Shiri
Concordia University, Canada

INTRODUCTION

Since the introduction of the *Semantic Web* vision (Berners-Lee, Hendler, & Lassila, 2001), attempts have been made for making Web resources more machine interpretable by giving them a well-defined meaning through semantic markups. One way to encode such semantic markups is to use ontologies. An *ontology* is "an explicit specification of a conceptualization" (Gruber, 1993, p. 199). Informally, an ontology consists of a set of terms in a domain, relationships between the terms, and a set of constraints on the way in which those terms can be combined. By explicitly defining the relationships and constraints among the terms, the semantics of the terms can be better defined and understood.

Over the last few years, a number of ontology languages have been developed, most of which use *Description Logics* (DLs) (Baader, McGuinness, Nardi, & Schneider, 2003) as the foundation. The family of DLs is a subset of first-order logic (FOL) and is considered to be attractive as it keeps a good compromise between expressive power and computational tractability.

Uncertainty is a form of deficiency or imperfection in the information/data, where the truth of information is not established definitely. Uncertainty modeling and reasoning have been challenging issues for over two decades in many disciplines, such as database and artificial intelligence. Most of the information in the real world is uncertain or imprecise, for example, classifications of genes in bioinformatics, schema matching in information integration, finding best matches in a Web search, and so forth. Therefore, uncertainty management is essential for the success of many such applications and in particular DLs and the Semantic Web.

Despite its popularity, it has been realized that classical DLs are inadequate to model uncertainty. For example, in the medical domain, one might want to express that: "It is very likely that an obese person would have heart disease," where "obese" is a vague concept that may vary across regions and "likely" shows the uncertain nature of this information. Such an expression cannot be expressed using classical DLs.

The importance of incorporating uncertainty in DLs has been recognized by the knowledge representation community: "modeling primitives such as … fuzzy/probabilistic definitions" could be the next step for extension (Horrocks et al., 2000, p. 3). For this, a number of frameworks have been proposed to incorporate uncertainty in DLs. This paper provides a survey of these proposals.

The rest of this paper is organized as follows. We first provide the background on the classical DL framework. We then study representative extensions of DLs with uncertainty. This follows by some possible research directions for incorporating uncertainty in the Semantic Web. We conclude with a summary and some remarks.

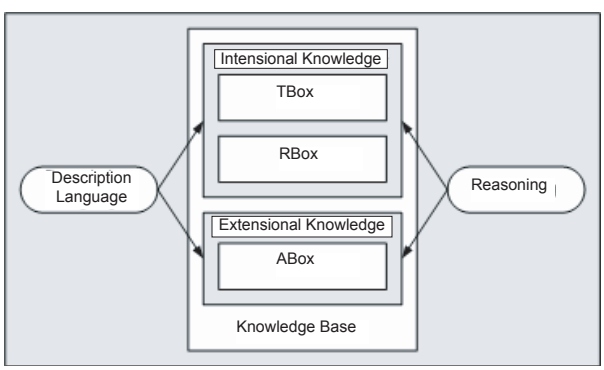
BACKGROUND

In this section, we review the basics of the classical DL framework, which provides facilities to represent knowledge bases and to reason about them.

As shown in Figure 1, the classical DL framework consists of three components:

1. **Description Language:** All description languages have elementary descriptions which include atomic

Figure 1. Classical DL framework



concepts (unary predicates) and atomic roles (binary predicates). Complex descriptions are built inductively from atomic ones using concept constructors. In this work, we focus on the description language $\mathcal{ALC}$ (Baader et al., 2003). Let C and D be concept descriptions. $\mathcal{ALC}$ includes atomic concepts A, atomic roles R, top/universal concept $\top$, bottom concept $\bot$, concept negation $\neg C$, concept conjunction $C \sqcap D$, concept disjunction $C \sqcup D$, role value restriction $\forall R.C$ (meaning $\forall y$: $R(x,y) \rightarrow C(y)$, for x in the domain), and role exists restriction $\exists R.C$ (meaning $\exists y$: $R(x,y) \wedge C(y)$, for x in the domain).

2. **Knowledge Base (KB):** The KB is composed of both intensional knowledge and extensional knowledge (see Figure 1). The former includes the Terminological Box (TBox or $\mathcal{T}$) consisting of a set of terminological axioms that could be concept subsumptions $C \sqsubseteq D$ and/or concept definitions $C \equiv D$ (where C and D are concepts), and the Role Box (RBox or $\mathcal{R}$) consisting of a set of role axioms that could be role subsumptions $R \sqsubseteq S$ and/or role definitions $R \equiv S$ (where R and S are roles). The extensional knowledge includes the Assertional Box (ABox or $\mathcal{A}$) consisting of a set of assertions/facts that could be concept assertions a: C (i.e., a is an instance of concept C) and/or role assertions (a,b):R (i.e., individuals a and b are related through relationship R).
3. **Reasoning Component:** A DL framework is equipped with reasoning services which allows that implicit knowledge be derived from explicit knowledge.

DESCRIPTION LOGICS WITH UNCERTAINTY

In this section, we study existing frameworks for DLs with uncertainty. We first provide a classification of the approaches of these frameworks. We then study representative extensions of DLs with uncertainty.

Approaches to Extend Description Logics with Uncertainty

On the basis of their mathematical foundation and the type of uncertainty modeled, we can classify existing proposals of DLs with uncertainty into three approaches: fuzzy, probabilistic, and possibilistic.

The fuzzy approach, based on fuzzy set theory (Zadeh, 1965), deals with vagueness in the knowledge, where a proposition is true only to some degree. For example, the statement: "Jason is obese with degree 0.4" indicates Jason is slightly obese. Here, the value 0.4 is the degree of membership that Jason belongs to the fuzzy concept obese.

The probabilistic approach, based on classical probability theory, deals with the uncertainty due to lack of knowledge, where a proposition is either true or false, but one does not know for sure which one is the case. Hence, the certainty value associated with the proposition refers to the probability that the proposition is true. For example, one could say: "The probability that Jason would have heart disease, given that he is obese, lies in the range [0.8,1]."

Finally, the possibilistic approach, based on possibility theory (Zadeh, 1978), allows both certainty (necessity measure) and possibility (possibility measure) to be handled in the same formalism. For example, by knowing that "Jason's weight is above 80 kg," the proposition "Jason's weight is 80 kg" is necessarily true with certainty 1, while "Jason's weight is 90 kg" is possibly true with certainty 0.5.

Description Logics with Uncertainty—Current State

To incorporate uncertainty into DLs, each component of the DL framework needs to be extended (Figure 2). In what follows, we survey how the description language, the KB, and the reasoning component have been extended with uncertainty using fuzzy, probabilistic, and possibilistic approaches.

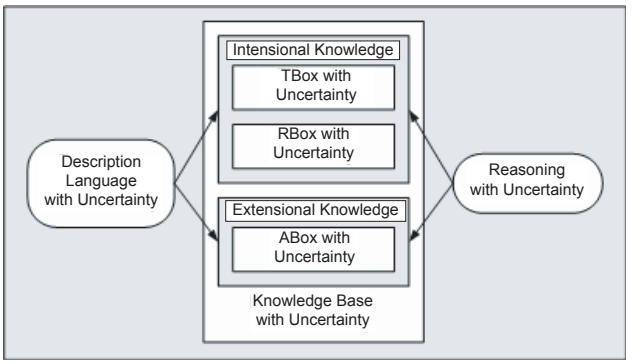
Description Languages with Uncertainty

The description languages contain a set of language constructors that serve as the building blocks of the description. In this section, we study how description languages have been extended using fuzzy and possibilistic approaches. To the best of our knowledge, no probabilistic extension of description languages has been proposed.

Fuzzy Description Languages

All existing proposals for fuzzy DL extend the semantics of the description language by fuzzifying their interpretation using fuzzy logic (Hölldobler, Khang, & Störr, 2002;

Figure 2. DL Framework with uncertainty



Sánchez & Tettamanzi, 2004; Straccia, 2001, 2004a, 2004b, 2005a, 2005b; Tresp & Molitor, 1998). In general, a fuzzy interpretation $\mathcal{I}$ is a pair $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where $\Delta^{\mathcal{I}}$ is the domain and $\cdot^{\mathcal{I}}$ is an interpretation function that maps language elements to some membership degree in [0,1]. For instance, the semantics of an atomic concept A is defined as $A^{\mathcal{I}}(a) \in [0,1]$ for all $a \in \Delta^{\mathcal{I}}$. That is, if individual a is an element of the domain, then the interpretation of the atomic concept A gives the membership degree that a belongs to A. The semantics of complex descriptions are defined in a straightforward way. For instance, the semantics of concept intersection is defined as $(C \sqcap D)^{\mathcal{I}}(a) = \min\{C^{\mathcal{I}}(a), D^{\mathcal{I}}(a)\}$, for all $a \in \Delta^{\mathcal{I}}$. That is, the certainty degree of concept C intersect D is the minimum of the certainty degrees of C and D.

In addition to the semantic extension, syntactic extensions to the description language are also proposed. The *manipulators/modifiers* (Hölldobler et al., 2002; Straccia, 2005a, 2005b; Tresp & Molitor, 1998) are unary operators that can modify the membership functions of the concepts they are applied to (e.g., "mostly," "very"), whereas *fuzzy quantifiers* (Sánchez & Tettamanzi, 2004) allow expressing vague quantities (e.g., "about 2") and quantity intervals (e.g., "roughly between 1 and 3").

Possibilistic Description Languages

Hollunder (1994) proposed a possibilistic extension to the description language. The idea is to keep the original description language syntax, while changing its interpretation using possibility theory. More specifically, the *possibility measure* Π for a concept (or event) C induced by a possibility distribution π on a set of interpretations $\Omega_{\perp}$ is defined as $\Pi(C) = sup\{\pi(\omega) \mid \omega \in \Omega_{\perp} \text{ and } \omega \vDash C\}$, characterizing the extent to which C is possible. On the other hand, the *necessity measure* N for a concept C is defined as $N(C) = inf\{1-\pi(\omega) \mid \omega \in \Omega_{\perp} \text{ and } \omega \nvDash C\}$, characterizing the extent to which this event is necessary or certain to occur. For example, the possibility that C and D are occurring at the same time is no more than the minimum of their possibilities, that is, $\Pi(C \sqcap D) \leq min\{\Pi(C), \Pi(D)\}$, but the necessity that both of them occur is $N(C \sqcap D) = min\{N(C), N(D)\}$.

Knowledge Base with Uncertainty

The fuzzy, probabilistic, and possibilistic extensions of KBs are defined in a similar way as the classical case, except that the interpretation $\mathcal{I}$ corresponds to the extension considered. An interpretation $\mathcal{I}$ *satisfies* (or is a *model* of) a KB Σ, denoted $\mathcal{I} \vDash \Sigma$, if it satisfies each element of Σ. Also, Σ is *satisfiable* if there exists an interpretation $\mathcal{I}$ that satisfies Σ. In this section, we study how each component of the KB can be extended with uncertainty.

Fuzzy Knowledge Base

Each component of the KB (i.e., the TBox, RBox, and ABox) has been extended with fuzzy logic.

For the TBox, two approaches are proposed. The first approach keeps the syntax the same as classical terminological axioms while extending only the semantics using fuzzy logic. Examples of this approach include Hölldobler et al. (2002), Sánchez & Tettamanzi (2004), and Straccia (2001, 2004a, 2004b, 2005b). A fuzzy interpretation $\mathcal{I}$ satisfies a fuzzy concept inclusion $C \sqsubseteq D$ if for all $a \in \Delta^{\mathcal{I}}$, we have that $C^{\mathcal{I}}(a) \leq D^{\mathcal{I}}(a)$. That is, the certainty of a subconcept C (e.g., *VeryTall*) is no more than the certainty value of a super-concept D (e.g., *Tall*). In the second approach, the TBox is extended both syntactically and semantically (Straccia, 2005a). Let C and D be concepts, $op \in \{\geq, \leq, >, <\}$, and $\alpha \in [0,1]$. A fuzzy terminological axiom is expressed as $\langle C \sqsubseteq D \text{ op } \alpha \rangle$. For example, $\langle C \sqsubseteq D \geq 0.8 \rangle$ indicates "the certainty that C is subsumed by D is at least 0.8."

Extension to RBox with fuzzy logic is proposed in Straccia (2005b). It is similar to the TBox counterparts, except that we have role axioms instead of terminological axioms.

In terms of the ABox, a fuzzy assertion can be represented as $\langle X \text{ op } \alpha \rangle$, where X is either a concept assertion $a{:}C$ or a role assertion $(a,b){:}R$, $op \in \{\geq, \leq, >, <, =\}$, and $\alpha \in [0,1]$ (Hölldobler et al., 2002; Sánchez & Tettamanzi, 2004; Straccia, 2001, 2004a, 2004b, 2005a, 2005b; Tresp & Molitor, 1998). For example, $\langle a{:}C \geq 0.5 \rangle$ means "the certainty that a is an instance of concept C is at least 0.5."

Probabilistic Knowledge Base

Several proposals extend TBox and ABox using probability theory as the mathematical basis.

A probabilistic TBox contains a set of classical terminological axioms and a set of probabilistic terminological axioms. The probabilistic information can either be embedded as part of the terminological axiom (Giugno & Lukasiewicz, 2002; Heinsohn, 1994; Jaeger, 1994) or be stored in Bayesian networks (Koller, Levy, & Pfeffer, 1997; Staker, 2002). For lack of space, we describe only the first approach here. A probabilistic terminological axiom is an expression of the form $P(C|D) \in [l,u]$, where C and D are concepts and $0 \leq l \leq u \leq 1$. This states that: "if an individual a is known to belong to concept D, then the probability that a belongs to concept C lies in $[l,u]$."

A probabilistic ABox (Giugno & Lukasiewicz, 2002; Jaeger, 1994) contains assertions of the form $P(a{:}C) \in [l,u]$, where C is a concept, a is an individual, and $l, u \in [0,1]$. Intuitively, this asserts that "the probability that an individual a belongs to concept C lies in $[l,u]$." The probabilistic assertions can also be applied to roles.

Possibilistic Knowledge Base

In Hollunder (1994), a possibilistic extension to TBox and ABox is proposed. Let X be an axiom or assertion, $\prod_\alpha$ be a possibility degree, and N_α be a necessity degree, where $\alpha \in (0,1]$. A possibilistic terminological axiom is either in the form $\langle X, \prod_\alpha \rangle$, meaning "$X$ is possibly true with degree at least α," or $\langle X, N_\alpha \rangle$, meaning "$X$ is necessarily true with degree at least α."

Reasoning with Uncertainty

The reasoning component provides inference services that enable implicit knowledge to become explicit. In this section, we study fuzzy, probabilistic, and possibilistic reasoning procedures.

Fuzzy Reasoning

For fuzzy reasoning, two main approaches are proposed. The first approach (Straccia, 2004a) transforms fuzzy DL into classical DL. The main advantage is that one can directly use existing reasoners developed for classical DL. The problem is that existing reasoners do not consider certainty values in fuzzy DL as something special, and hence no additional optimization techniques may be applied to inferences with certainty values involved.

The second approach extends existing reasoning procedure so that it takes into account the presence of certainty values during the reasoning process; hence, a reasoner has to be built from scratch. There are two variants of this approach. In the first one, certainty values are dealt with within the inference rules (Hölldobler et al., 2002; Straccia, 2001, 2004b). The basic idea is that, given a fuzzy ABox, a set of inference rules are applied to transform the ABox into simpler and satisfiability preserving equivalence, so that the implicit knowledge becomes explicit. For example, if we know that $\langle a{:}C \sqcap D \geq 0.5 \rangle$, then we could infer that $\langle a{:}C \geq 0.5 \rangle$ and $\langle a{:}D \geq 0.5 \rangle$. The inference rules are applied until either all branches in the extended ABox contain a contradiction/clash (meaning no model can be built), or there exists a clash-free completion of ABox (meaning the ABox is satisfiable). The second approach relies on integer programming and linear optimization to deal with certainty values (Straccia, 2005b). Here, a set of completion rules are applied to generate new assertions together with a set of constraints in the form of inequations over variables with values in [0,1]. The inference rules are applied until either the extended ABox contains a clash, or no rule can be further applied. If there is a clash, the ABox is unsatisfiable. Otherwise, an optimization technique is applied to solve the system of inequations to determine the satisfiability of the KB or the tightest bound such that an assertion is true.

Probabilistic Reasoning

The probabilistic reasoning procedure depends on how the probabilistic information is represented in the KB. In case Bayesian networks are used to express the probabilistic information, the inference procedure developed for Bayesian networks can be directly applied (Ding & Peng, 2004; Koller et al., 1997; Staker, 2002). On the other hand, if the probabilistic information is embedded in the KB (Baader et al., 2003, Giugno & Lukasiewicz, 2002), inference procedures similar to the fuzzy/linear optimization approach are applied to find the optimal bound for which, given a KB, a conditional probability $P(C|D)$ is satisfied.

Possibilistic Reasoning

The possibilistic reasoning is not well explored. No concrete calculus is provided in Hollunder (1994) for inferences on a given KB.

FUTURE TRENDS

In this section, we outline some possible directions for uncertainty management in the Semantic Web. First, it would be useful to have a generic framework that could handle various forms of uncertainty under a unifying umbrella. The first attempt has been made by Haarslev, Pai, and Shiri (2005), which is later extended with the reasoning component(2006). However, more work is required in this direction. For practical reasons, we also need to develop efficient tools that support DL with uncertainty. Finally, to handle real-life applications, more expressive fragments of DL (e.g., $\mathcal{SHOIN}$) should be extended with uncertainty.

CONCLUSION

We studied existing frameworks for DL with uncertainty which can form a basis for uncertainty management in the Semantic Web. Based on the mathematical foundations and the types of uncertainty modeled, we classified these frameworks into fuzzy, probabilistic, and possibilistic approaches. We also studied how these approaches extend the components of the DL framework, including description language, KB, and reasoning procedures. It is our hope that this survey would foster further research in incorporating uncertainty in the next generation of Web.

ACKNOWLEDGMENTS

This work is supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada, and

by Faculty of Engineering and Computer Science (ENCS), Concordia University. We also thank anonymous reviewers for their useful comments.

REFERENCES

Baader, F., McGuinness, D., Nardi, D., & Schneider, P. P. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge, UK: Cambridge University Press.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American, 284*(5), 34-43.

Ding, Z., & Peng, Y. (2004). A probabilistic extension to ontology language OWL. *The 37th Hawaii International Conference on System Sciences* (p. 40111a), Big Island, HI. IEEE Computer Society

Giugno, R., & Lukasiewicz, T. (2002). P-$\mathcal{SHOQ}$(D): A probabilistic extension of $\mathcal{SHOQ}$(D) for probabilistic ontologies in the semantic Web. Lecture notes in Computer Science. In *Proceedings of The 8th European Conference on Logics in Artificial Intelligence* (pp. 86-97). Cosenza, Italy. Springer Verlag.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Haarslev, V., Pai, H. I., & Shiri, N. (2005, November 7). A generic framework for description logics with uncertainty.In *Proceedings of the 2005 Workshop on Uncertainty Reasoning for the Semantic Web at the 4th International Semantic Web Conference* Galway, Ireland (pp. 77-86).

Haarslev, V., Pai, H. I., & Shiri, N. (2006, May 11-13). Uncertainty reasoning in description logics: A generic approach. In G. Sutcliffe & R. Goebel (Eds.) *Proceedings of the 19th International FLAIRS (Florida Artificial Intelligence Research Society) Conference,* Melbourne Beach, FL (pp. 818-823). AAAI Press.

Heinsohn, J. (1994, July 29-31). Probabilistic description logics. In R. López de Mántaras & D. Poole (Eds.) *Proceedings of The 10th Annual Conference on Uncertainty in Artificial Intelligence,* Seattle, WA (pp. 311-318). Morgan Kaufmann.

Hölldobler, S., Khang, T. D., & Störr, H.-P. (2002, December 3-5). A fuzzy description logic with hedges as concept modifiers. In *Proceedings of The 3rd International Conference on Intelligent Technologies,* Hanoi, Vietnam (pp. 25-34).

Hollunder, B. (1994, July 29-31). An alternative proof method for possibilistic logic and its application to terminological logics. In R. López de Mántaras & D. Poole (Eds.) *Proceedings of The 10th Annual Conference on Uncertainty in Artificial Intelligence* , Seattle, WA(pp. 327-335). Morgan Kaufmann.

Horrocks, I., Fensel, D., Broekstra, J., Decker, S., Erdmann, M., Goble, C., et al. (2000). *OIL: The ontology inference layer*. (Tech. Rep. No. IR-479). Vrije Universiteit Amsterdam, Faculty of Sciences, Department of Math and Computer Sciences.

Jaeger, M. (1994). Probabilistic reasoning in terminological logics. In J. Doyle, E. Sandewall, & P. Torasso (Eds.) P*roceedings of The 4th International Conference on Principles of Knowledge Representation and Reasoning,* Bonn, Germany (pp. 305-316). Morgan Kaufmann.

Koller, D., Levy, A. Y., & Pfeffer, A. (1997, July 27-31). P-CLASSIC: A tractable probabilistic description logic. In *Proceedings of The 14th National Conference on Artificial Intelligence*, Providence, RI (pp. 390-397). AAAI Press

Sánchez, D., & Tettamanzi, A. G. B. (2004,). Generalizing quantification in fuzzy description logics. In *Proceedings of The 8th Fuzzy Days in Dortmund*, Dortmund, Germany.

Staker, R. (2002). Reasoning in expressive description logics using belief networks. In *Proceedings of The International Conference on Information and Knowledge Engineering* . Las Vegas, NV(pp. 489-495). CSREA Press

Straccia, U. (2001). Reasoning within fuzzy description logics. *Journal of Artificial Intelligence Research, 14,* 137-166.

Straccia, U. (2004a, September 27-30). Transforming fuzzy description logics into classical description logics. In J. Alferes & J. Leite (Eds.), *Proceedings of The 9th European Conference on Logics in Artificial Intelligence* Lisbon, Portugal(pp. 385-399). Springer Verlag.

Straccia, U. (2004b, July 4-9). Uncertainty in description logics: A lattice-based approach. In *Proceedings of The 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems* Perugia, Italy (pp. 251-258).

Straccia, U. (2005a, May 29-June 1). Towards a fuzzy description logic for the semantic Web (preliminary report). Lecture notes in computer science In A. Gómez-Pérez & J. Euzenat (Eds.) *Proceedings of The 2nd European Semantic Web Conference* Heraklion, Greece(pp. 167-181). Springer Verlag.

Straccia, U. (2005b, July 26-28). Fuzzy $\mathcal{ALC}$ with concrete domains. *In Proceedings of The International Workshop on Description Logics* Edinburgh, UK (pp. 96-103).

Tresp, C., & Molitor, R. (1998, August 23-28). A description logic for vague knowledge. In H. Prade (Ed.) *Proceedings of The 13th European Conference on Artificial Intelligence* Brighton, UK (pp. 361-365). John Wiley and Sons.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control, 8*, 338-353.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems, 1*(1), 3-28.

KEY TERMS

Description Logics: A decidable subset of first order logic.

Inference: The process of deriving conclusions from a knowledge base.

Knowledge Base: A collection of axioms and assertions.

Knowledge Representation: A formalism used for expressing knowledge stored in a knowledge base.

Ontology: An explicit formal specification of conceptualization that consists of a set of terms in a domain and the relations among them.

Semantic Web: An extension of the current Web by giving well-defined meaning to Web resources.

Uncertainty: A form of deficiency/imperfection in the information where the truth of information is not established definitely.

Service Description Ontologies

Julia Kantorovitch
VTT Technical Research Centre of Finland, Finland

Eila Niemelä
VTT Technical Research Centre of Finland, Finland

INTRODUCTION

Services can be Internet-based e-commerce services, business services that abstract company-level interactions, or any other software services that are provided by surrounding devices that are mobile or embedded in nearly any type of physical environment (e.g., home, office, or cars). In brief, services are ubiquitous and executed in heterogeneous environments.

Surrounding the definitions and technologies that describe services, there are some important features that are in common. First, services always have some actions that are performed by an entity, possibly on behalf of another. Second, there always exists service interaction, including a service provider, service requestor, and service registry. Finally, services have inherent value that is transferred from the service provider to the service requestor as a result of the service's execution.

To invoke and operate a service in the most efficient way, the service is to be described via essential types of knowledge: a) what the service requires from the user/agent(s) and then provides for them; b) where and when the service is available; c) what quality level is to be guaranteed; d) how to access and interact with the service; and e) what access rights are granted over the service.

An accurate service description, including the specifications of functional and nonfunctional properties, benefits and facilitates several important activities, such as service discovery, service composition, and service administration, including the monitoring and controlling of the service's execution. However, due to the diversity of service contexts, service technologies shall be generic and adaptable to different domains and heterogeneous environments. Service description ontologies solve this problem by enabling a rich representation of services and a common understanding about their respective features. The use of ontologies enables computational entities and services to have a common set of concepts and properties for representing knowledge about a domain of interest. The deployment and customization of existing and emerging service systems can also be considerably facilitated by a common set of ontologies that is developed in order to describe service semantics.

BACKGROUND

The studies on the data schema in XML, RDFs, OWL, and service description languages all form the basis for defining the service semantics. The eXtensible Markup Language (XML), as defined by the World Wide Web Consortium (W3C), is a well known, and industry accepted, way for representing flexible information. It is used to create information objects consisting of elements encoded by tags and attributes. XML schemas express shared vocabularies and allow machines to carry out those rules that are established by people.

The resource description framework (RDF) and Web ontology language (OWL) are built on XML and facilitate greater machine interpretability of content by providing additional vocabulary along with formal semantics.

XML, SOAP, WSDL, and UDDI form the core of Web service standards. Simple object access protocol (SOAP) is a lightweight protocol for exchanging XML-based information in a distributed environment. Web service description language (WSDL) is an XML format for describing services as a set of endpoints operating on messages. The operations and messages are described abstractly, and then bound to a concrete protocol and message format in order to define an endpoint. UDDI (universal description, discovery, and integration) is concerned with the publishing and discovery of Web services.

Service Level Agreements (SLAs) related quality information is described by:

- Web service level agreement (WSLA) language (Ludwig et al., 2002), which is based on an XML schema and defines the SLAs in three parts; i) contractual parties, ii) the characteristics of the service and its observable parameters, and ii) obligations to various guarantees and constraints that may be imposed on the SLA parameters, or
- Web services offerings language (WSOL) (Tosic, Paguredk, Patel, 2003), which is a formal specification language for defining the functional and QoS constraints and access rights for Web services. WSOL is XML-based and compatible with WSDL.

SERVICE ONTOLOGIES

In the following, the existing semantic approaches for describing the services, including their functional capabilities, QoS, and context are compared in Tables 1-3. This comparison aims at identifying the main benefits and shortcomings and the missing aspects in the service description ontologies that serve as a basis for the conclusion and future research trends that are to be identified in the remainder of the paper.

Service Functionality Descriptions

Several service ontologies contribute to the service creation, provision, and execution, to a varying extent, by using different description languages. Table 1 compares the main properties of four existing service description ontologies from the viewpoint of their completeness to describe the service related aspects. Web ontology language for services (OWL-S), Web service modeling ontology (WSMO) (Roman, Keller, Lausen, Lara, Bruijn, Stollberg, et al., 2005) and Internet reasoning service (IRS) (Domingue, Cabral, Hakimpour, Sell, & Motta 2004; Motta, Domingue, Cabral, & Gaspari, 2003;) provide specific ontology building blocks for particular purposes of use. Conversely, METEOR-S (LSDIS Lab, 2005) is an approach that targets the extension and integration of the existent Web services and semantic Web technologies. OWL-S is the most widely used approach concerning service semantic modeling. OWL-S combines the expressiveness of description logics, as it builds on OWL. The WSMO is a relatively new effort and is based on the Web service modeling framework (WSMF) (Fensel, Bussler, Ding, & Omelayenko, 2002). All of the approaches provide for specific advantages that are missing from another approach. However, the approaches do not individually provide complete description support for service semantics (see Table 1, Difference row).

Service Quality Descriptions

Table 2 compares a set of QoS ontologies that address their benefits and shortcomings. The difference that is in focus results in different ontology layers. A lack of completeness is common for all the approaches; only one or a few qualities are considered, and the vocabulary or/and metrics are missing. Moreover, there is no support for making tradeoffs between quality attributes or managing QoS at run-time.

In Zhou et al. (Zhou, Chia, & Lee, 2004), the QoS ontology with the three layers covers the matchmaking, QoS property definition layer with domain and range constraints, and metrics with the measurement details. A drawback of this approach is that the proposed ontology is rather limited, while the QoS ontology vocabulary is absent. The framework presented in Maximilien and Singh (2004) is based on agents that enable dynamic Web services selection. On the other hand, work in Tosic et al. (Tosic, Esfandiari, Pagurek, & Patel, 2002) has focused on metrics, measurement units, and currencies to support QoS semantic management. An extended matchmaking mechanism with the concept of the service broker is addressed in Tian et al. (Tian, Gramm, Naumowicz, Ritter, & Schiller, 2003). It also classifies the QoS parameters into network-related and server/client-related parameters. The MOQ (mid-level ontologies for quality) framework (Kim, Sengupta, & Evermann, 2005) aims to minimize the ambiguities in QoS evaluations by defining the ontologies for the requirements, measurement, traceability, and quality management.

Service Context Descriptions

Concerning the contextual characteristics of services, several ontologies have been designed, some of which are more elaborate and others more succinct, depending on their scope. The most popular of these are context ontology language (CoOL) (Strang, Linnhoff-Popien, & Frank, 2003), context broker architecture (CoBrA) (Chen, Finin, & Joshi, 2003), service-oriented context-aware middleware (SOCAM) (Gu, Wang, Pung, & Zhang, 2004), COntext MAnagement oNTOlogy (COMANTO) (Strimpakou, Roussaki, Pils, & Anagnostou, 2006), and the standard ontology for ubiquitous and pervasive applications (SOUPA) (Chen, Finin, & Joshi, 2005). Their main characteristics are shown in Table 3.

Most of the approaches presented address the vocabulary ontology needs in the domain of pervasive computing. The context ontologies that are designed include a set of vocabularies for describing people, agents, and places, as well as a set of properties and relationships that are associated with these basic concepts. However, rather little emphasis is placed on services, including their functional properties and related aspects, such as user interfaces and devices on which these services are deployed, along with temporal contextual information. No attempts have been made to align service and context ontologies.

Tools for Describing Service Semantics

Numerous freeware and commercial tools to support the development and use of ontologies are currently available: SWOOP is a hypermedia based OWL ontology editor; Protégé is a free, open source ontology editor and knowledgebase framework; TopBraid Composer™ is an enterprise class platform. The advancement in these tools has greatly improved the ability to test and build ontologies from scratch or to reuse existing ontologies.

Application programming interfaces (APIs) for ontology languages provide programming language dependent means to load ontologies, manipulate the ontology classes and relations, perform reasoning, and provide persistent storage for

Table 1. Summary of service description ontologies

	OWL-S	***WSMO***	***IRS***	***METEOR-S***
Ontology building blocks	Service profile Service model Service grounding for invocation	Ontologies Goals Web Services Mediators Core nonfunctional properties	Domain models Task models Problem-solving methods Bridges between model components Utilizes WSMO ontologies	Utilizes OWL-S ontologies
Service middleware	Advertising Discovery Matching Invocation	Advertising Discovery Invocation (under specification)	Advertising task Brokering tasks Reasoning rules Mapping	Service annotation, discovery, and composition QoS mgt for business processes
Tool support	DAML-S virtual machine Mindswap/OWL-S API	WSMX, IRS-III, Semantic Web Fred, WSMO design studio	IRS-III framework (server, publisher, client)	Meteor-S framework (enhanced legacy service discovery, description, etc., tools)
Difference (advantages& drawbacks)	Declarative advertisement of service capabilities. Single modeling element for both views Request is expressed by the desired service description Does not address heterogeneity explicitly Constructs for service compositions and interactions. Grounding to WSDL More mature	Quality properties, for example, performance, reliability and security Distinguishes the requester and provider points of view Request is described in the form of goals that is, the results expected Consider the heterogeneity by mediators Orchestration of services is under specification Does not offer any grounding (however, in the future it might be grounding independent)	Publishing legacy as Web services User direct invoking Programmable framework	Semantic operations for existing services Dynamism and scalability

Table 2. Summary of QoS description ontologies

	Zhou et al.	***Maximilien et al.***	***Tosic et al.***	***Tian et al.***	***Kim et al.***
Ontology layers	Profile layer Property layer Metrics layer	Upper ontology Middle ontology	Metrics Measurement units Currencies	Network-related parameters Server-client related parameters	Requirements Measurement Traceability Quality mgmt
Focus	Defining QoS semantics	Web service agent proxy	QoS semantic management	Matchmaking mechanism	QoS development phases
Advantages	QoS properties and metrics	Language and vocabulary	Dependencies between metrics	Server status monitoring	Extensions to existing ontologies
Drawbacks	No vocabulary	No metrics concept	No integrated solution	Only XML Schema	No complete QoS ontology

Table 3. Summary of service context ontologies

	CoOL	***CoBrA***	***SOCAM***	***COMANTO***	***SOUPA***
Main concepts	Aspect-Scale-Context	People-Agents-Places	Person-Location-Activity-Computational entity	Person-Place- Preferences- Agenda-Activity- Subscribed service	Agent-Time-Space-Events-Users-Actions-Security
Domain	Separated customer, service provider and context provider domains	Intelligent spaces	Middleware for indoor applications	Requirements of large pervasive computing environments	Pervasive computing applications
Ontology layers	Metrics ontology	Vocabulary Low-layer ontology	Common and domain specific context ontology	Vocabulary	Core vocabularies Extension vocabularies
Advantages	Mappings between metrics	Inference engine	Knowledge sharing and reuse	Stakeholder support	Standardized context ontology

the model. Jena and OWLS API are the most popular Java frameworks for building semantic Web applications. These tools provide an application developer with programming language level support for working with ontologies.

There are several tools that provide reasoning capabilities for ontology applications offering a language dependent API or DIG interface (i.e., a standardized XML interface to description logics systems). The examples of such are FaCT++, Pellet, and RacerPro. These tools help in ontology testing and in the development of application level intelligence that is based on ontologies. Domain ontology specific editors such as OWLS Editor and WSMO design studio help in order to create error free semantic descriptions based on a specific ontology.

From the service modeling perspective, what is missing in those tools are the features that would make the contextual semantic information related to service descriptions easier to understand and to be used by an application developer. The visualization support that is provided by generic ontology editors is limited to mainly showing the abstract structural relations of the ontology classes and their instances. These tools are not well suited for understanding and modeling the semantic relations and services in a complex dynamic physical world related to application scenarios. This leaves the step of adopting semantic approaches in service-oriented application development excessively high for most programmers.

FUTURE TRENDS

Although ontologies are community-wide contracts about representations of specific domain knowledge, the existing ontologies mainly address the technical dimension of service semantics, in turn ignoring business and development dimensions in service engineering. Most of the ontologies discussed are specific to Web/Internet-based services (e.g., e-commerce), in which services provided by devices are little addressed. Therefore, the future research effort should be directed towards enhancing ontology engineering by a stakeholder-centric modeling approach and providing support for i) developing industry strength business ontologies, ii) generic ontologies for QoS metrics and QoS execution management, iii) application specific domain ontologies, and iv) integrated orchestration of the developed ontologies in service engineering practices. In the software architecture field, the use of viewpoints is a community-wide accepted approach to cluster stakeholder-related concerns into a single view. This principle can be lent to service semantics engineering. The use of multiple views is a necessity; the interests of stakeholders differ, application domains differ, and service functionality and quality differ according to the usage/execution contexts. Moreover, different application domains, for example, e-business and pervasive computing applications, require modeling languages that take into account the characteristics of a domain by providing a notation that can be enhanced and adapted by domain specific extensions. The approach used in software family engineering, namely the separation of commonality and variability, would be a viable approach that solves the problems in separation of common and domain specific semantics, and the integrated use of defined service ontologies. Last, but not least, developer tools are required to provide specific support for software engineers and system administrators through the development and management process including the engineering of semantic services, their deployment, and subsequent management.

CONCLUSION

The present study provided an overview of the existing service description ontologies and thereby, improves the understanding of the benefits and drawbacks of service ontologies in defining the functionality, quality, and contexts of services by the available tools.

From a technical point of view, service description ontologies are to provide knowledge for describing the required and provided properties of a service, ability and rights of achieving a service, and the quality guaranteed for a service. XML, RDF, and OWL schema provide a basis for service description languages and ontologies, such as OWL-S, WSMO, and IRS, which in turn provide building blocks for service semantics. However, existing service ontologies focus mainly on Web services, and none of them provides complete support for service descriptions as required in modern service centric systems. Furthermore, QoS description ontologies also have a specific focus, for example, one or few quality attribute(s) in defining, managing, or matching OoS. To guarantee QoS requires comprehensive support for defining and managing all the relevant quality attributes of services, at the design time and run-time. Context ontologies assist in adapting services to the execution environment and to the observed changes in environmental or usage contexts. SOUPA, as the only standardized context ontology, provides the most promising approach to which QoS ontologies and domain ontologies can be integrated as extension ontologies.

Building ontologies consumes resources, in which these resources are justified only if the effort needed to establish and keep alive consensual representations of domains is outweighed by business gain, either in terms of cost, added value, or strategic dimensions. Therefore, future research activities should address the business and process agility dimensions by enhancing ontology engineering by a stakeholder centric modeling and separation of commonality and variability in service semantics. These topics are studied in the software architecture field. Therefore, the fusion of the design principles and techniques applied to the quality and

model driven architecture design is the most promising approach to enhance the existing service ontologies in order to better fit the service engineering and service semantics that are required in service centric computing systems.

REFERENCES

Chen, H., Finin, T., & Joshi. (2003). An ontology for context-aware pervasive computing environments. *Knowledge and Engineering Review, Special Issue on Ontologies for Distributed Systems*, *18*(3), 197-207.

Chen, H., Finin, T. & Joshi, A.. (2005). The SOUPA ontology for pervasive computing. *Whitestein Series in Software Agent Technologies,* Springer.

Domingue, J., Cabral, L., Hakimpour, F., Sell, D., & Motta, E. (2004). *IRS-III: A platform and infrastructure for creating WSMO-based semantic Web services*. Paper presented at the Workshop on WSMO Implementations (WIW 2004), September 29-30, in Frankfurt, Germany.

Fensel, D., Bussler, C., Ding, Y., & Omelayenko, B. (2002). The Web service modelling framework WSMF. *Electronic Commerce: Research and Applications*, ,113-137.

Gu, T., Wang, X., Pung, H. K., & Zhang, H. K. (2004). *An ontology-based context model in intelligent environments*. Paper presented at the Communication Networks and Distributed Systems Modelling and Simulation Conference (CNDS 2004), January 18-21, in San Diego, California, USA.

Kim, H. M., Sengupta, A., & Evermann, J. (2005). *MOQ: Web services ontologies for QOS and general quality evaluations*. Paper presented at the European Conference on Information Systems (ECIS 2005), May 26-28, in Regensburg, Germany.

LSDIS Lab. (2005). *METEOR-S initiative of LSDIS Lab. 2005. Meteor-s: Semantic Web services and processes*. University of Georgia. Retrieved from http://lsdis.cs.uga.edu/projects/meteor-s/

Ludwig, H., Keller, A., Dan, A., King, R., & Franck, R. (2003). *Web service agreement (WSLA) language specification, v. 1.0*, *wsla-2003/01/28, IBM Technical Report*. IBM

Maximilien, E. M., & Singh, M. P. (2004). A framework and ontology for dynamic Web services selection. *IEEE Internet Computing*, *8*(5), 84-93.

Motta, E., Domingue, J., Cabral, L. & Gaspari, M. (2003). *IRS-II: A framework and infrastructure for semantic Web services.* Paper presented at the 2nd International Semantic Web Conference (ISWC 2003), October 20-23, in Sanibel Island, FL, USA.

Roman, D., Keller, U., Lausen, H., Lara, R., Bruijn, J., Stollberg, M., Polleres, A., Feier, C., Bussler, C., &Fensel, D. (2005). Web service modeling ontology. Applied Ontology, *ESWC 2005*, *1*(1),77-106.

Strang, T., Linnhoff-Popien, C., & Frank, K. (2003). *CoOL: A Context Ontology Language to enable Contextual Interoperability*. Paper presented at the 4th IFIP WG 6.1 Int. Conf. on Distributed Applications and Interoperable Systems (DAIS2003), November 19-21, in Paris, France

Strimpakou, M., Roussaki, I., Pils, C., & Anagnostou, M. (2006). COMPACT: Middleware for context representation and management in pervasive computing environments. *International Journal of Pervasive Computing and Communications (JPCC)*, *2*(3).

Tian, M., Gramm, A., Naumowicz, T., Ritter, H., & Schiller, J. (2003). *A concept for QoS integration in Web services*. Paper presented at the 1st Web Services Quality Workshop (WQW 2003), December 13, in Rome, Italy.

Tosic, V., Paguredk, B., & Patel, K. (2003). WSOL - A language for the formal specification of classes of service for Web services. In *The International Conference on Web Services* (pp. 375-381), Las Vegas, June 23-26, 2003. CSREA Press.

Tosic, V., Esfandiari, B., Pagurek, B., & Patel, K. (2002). *On requirements for ontologies in management of Web services.* Paper presented at the Int. Workshop on Web Services, E-Business, and the Semantic Web (CAiSE 2002), May 27-28, in Toronto, Canada.

Zhou, C., Chia, L., & Lee, B. (2004). DAML-QoS ontology for Web services. In *Proc. of the Int. Conference on Web Services 2004 (ICWS04),* San Diego, California, USA.

FaCT++, http://owl.man.ac.uk/factplusplus/

Jena, http://jena.sourceforge.net/

OWL, http://www.w3.org/TR/owl-features/

OWL-S, http://xml.coverpages.org/ni2004-01-08-a.html

OWLS Editor, http://hydromodel.com/dl.htm

Pellet, http://pellet.owldl.com/

Protége, http://protege.stanford.edu/

RacerPro, v. 1.9, http://www.racer-systems.com/

RDF-Schema, http://www.w3.org/TR/rdf-schema/

SOAP, http://www.w3.org/TR/soap/

SWOOP, http://www.mindswap.org/2004/SWOOP/

TopBraid Composer™, http://www.topbraidcomposer.com/

WSDL, http://www.w3.org/TR/wsdl

WSMO studio, http://www.wsmostudio.org/

XML, http://www.w3.org/TR/xml/

KEY TERMS

Ontology: Ontology is a shared knowledge standard or a knowledge model defining the primitive concepts, relations, rules, and their instances comprising a relevant knowledge topic. Ontology is used for capturing, structuring, and enlarging explicit and tacit topic knowledge across people, organizations, systems, and software services.

QoS (Quality of Service): Refers to the nonfunctional properties of services at different levels. QoS is the degree to which a service meets its quality requirements and end user needs. QoS quantifies the service fitness based the collective behavior of composite services.

Service: A stateless piece of software with the predefined functional and quality properties that is used for achieving a particular action.

Service Composite: A cluster of services that are combined together from a bottom-up fashion in order to achieve an enhanced behavior, which no service by itself can provide. Service composites are made proactively (design-time) or reactively (run-time).

Service Description: A definition of what a service provides and how it is accessed and used. A service description includes descriptions of the functional and nonfunctional properties of the service, service interfaces, and the legal and technical constraints or rules for its usage.

Service Modeling: Service modeling produces a service description by exploiting generic graphical modeling languages, such as unified modeling language (UML), and/or textual notations such as Web services description language (WSDL) and Web ontology language for services (OWL-S).

Stakeholder: An individual, team, or organization (or classes thereof) with interests in, or concerns relative to, service development, service deployment, or service operation.

SLA (Service Level Agreement): A contract between a service provider and a service consumer that is related to the quality of service guaranteed by the service provider.

Vocabulary: Vocabularies express equivalent or the semantically closest concepts or concept expressions related to the domain of interest.

Shortest Path Routing Algorithms in Multihop Networks

Sudip Misra
Cornell University, USA

INTRODUCTION

The topic of *shortest path algorithms* is very fundamental and important in information science and technology. Shortest path algorithms have evolved over many years and have found applications in different domains such as telecommunication networks, military, and transportation. There has been a lot of work undertaken on this topic in this area in the past. A lot of research is still being conducted. The topic is still poorly understood. This article should be helpful to readers, because it reviews some of the important works conducted in the area, out of the plenty of works available on this topic. The problem is so fundamental that whatever the interest areas of the readers may be, they will find the article useful.

The popular traditional shortest path algorithms date back to 1958/1959 and were proposed by Dijkstra (1959), and Bellman (1958). Their algorithms found wide applications in the abovementioned domains for many years. However, they were static. Thereafter, many other algorithms were proposed in the last few decades, all of which can be classified to be either dynamic or semi-dynamic.

BACKGROUND

Multihop networks, such as the Internet and *mobile ad hoc networks* (MANETs) contain several routers and mobile hosts. The Internet typically employs routing protocols such as the open shortest path protocol (OSPF) and the intermediate system—intermediate system protocol (IS-IS), and the MANETs employ protocols such as the fisheye state routing (FSR), the optimized link state routing (OLSR), and the ad hoc on-demand distance vector routing (AODV).

In many of these protocols, each router (or a routing device) computes and stores a shortest path tree (SPT) from one router to all other routers and hosts in a routing domain (Moy, 1997; Peterson & Davie, 2000; Schwartz & Stern, 1980). Such networks, which can be modeled as graphs (Misra & Oommen, 2005b; Ramalingam & Reps, 1996), typically contain several routers/switches (nodes) connected by links (edges) with constantly changing costs (weights), link-ups (edge-insertions), and link-downs (edge-deletions).

SINGLE-SOURCE SHORTEST PATH ROUTING: DYNAMIC VERSUS STATIC

The problem of computing and maintaining information about the shortest paths information in a graph (with a single source)—where the edges are inserted/deleted and edge-weights constantly increase/decrease—is referred to as the *dynamic single source shortest path problem* (DSSSP). Although this problem is important, it has received little attention in the literature. The importance of the problem lies in the fact that it is representative of many practical situations in daily life, where most environments are dynamically changing. In such environments, one needs to devise efficient solutions to maintain the shortest path even though there are updates on the structure of the graph by virtue of edge-insertion/deletion, or edge-weight increase/decrease, and hopefully this can be achieved without recomputing everything "from scratch" following each topology update. An example of a single-source shortest path graph, after the insertion of an edge is shown in Figure 1. The new edge C→F appears in the list of shortest paths, and the existing edge B→F, which was earlier in the list of shortest paths, ceases to be so.

Out of the four possible edge operations (insertion/deletion and increase/decrease), it has been shown that edge-insertion is equivalent to edge-weight decrease, and edge-deletion is equivalent to edge-weight increase. Increasing or decreasing an edge-weight can be performed by inserting a new edge (with the new weight) parallel to the edge under consideration, and then deleting the old edge (Ramalingam & Reps, 1996). If all edge operations are allowed, the problem is referred to as the *fully dynamic problem*. If only edge-insertion/weight-decrease (or edge-deletion/weight-increase) is allowed, the problem is referred to as the *semi-dynamic problem* (Frigioni, Marchetti-Spaccamela, & Nanni, 1996).

Typically, with many present-day routing protocols,[1] with a unit change in network topology (e.g., link-ups, link-downs, and link-cost changes), each router in the routing domain is intimated of the change. This change typically triggers recomputation of each router's SPT.

There are well-known *static solutions* to the traditional combinatorial *single source shortest path problem* (Bellman, 1958; Dijkstra, 1959) which are unacceptably inefficient in

such dynamic practical scenarios because using them would involve recomputing the shortest path tree from scratch each time a topology change occurs in the graph. Static algorithms are unarguably more effective in fixed-infrastructure networks because of their polynomial time complexity. But they are extremely inefficient for time-critical, rapidly changing infrastructures. For instance, if such an algorithm is used in a real-time, large network routing scenario, where there are fast link-state changes, such recomputations will delay the execution of important routing functions considerably.

Two of the earliest known works on the dynamic shortest path problem date back to the papers by Spira and Pan (1975) and McQuillan, Richer, and Rosen (1980). While the former is theoretically proven to be inefficient, the latter has neither been proven theoretically nor through simulations.

The most recent and well-known solutions to the DSSSP problem on general graphs with positive real-valued edge-weights were proposed by Ramalingam and Reps (1996), Franciosa, Frigioni, and Giaccio (1997), and Frigioni, Marchetti-Spaccamela, and Nanni (2000). However, the solution by Franciosa et al. (2000) is limited to the semi-dynamic problem only. Although these recent works are theoretical in nature, Ramalingam and Reps (1996)'s and Frigioni et al. (2000)'s results were recently experimentally evaluated through simulations (Demetrescu, Frigioni, Marchetti-Spaccamela, & Nanni, 2001; Frigioni, Ioffreda, & Nanni, 1998). While the former was found to be superior when it concerns running time, the latter was shown to be better suited when the worst-case time was the main concern, or when the number of edges to be updated had to be minimized.

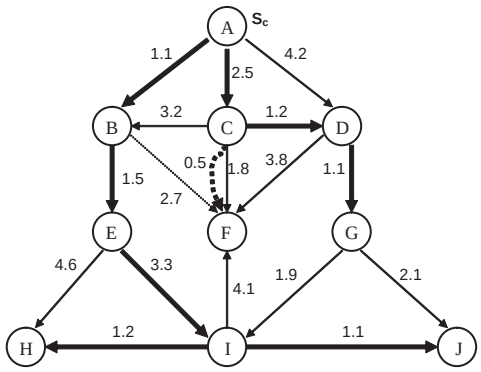
The well-known, fully dynamic algorithms (mentioned previously) are constrained by the following limitations:

1. The existing fully dynamic algorithms process unit changes to topology (i.e., edge-insertion/deletion or weight-increase/decrease) one change at a time, that is, sequentially. When there are several such operations occurring in the environment simultaneously, the algorithms are quite inefficient.
2. In environments where the edge-weights change stochastically and continuously, the existing algorithms (mentioned previously) would fail to converge to the actual underlying "average" solution.

The problems are worse in large topologies which have a large number of nodes and edges, and where a large number of topology changes can occur continuously at all times. In such cases the existing algorithms would fail to determine the shortest path information in a time-critical manner.

Since such scenarios are representative of the actual environments in which the dynamic shortest path algorithms are likely to operate, the existing solutions would be limitedly useful. Misra and Oommen (2005b) proposed a learning solution by taking the aforementioned aspects into consideration. To the best of my knowledge, other than their solution, there is no known solution to finding the shortest path in a real-weighted graph where multiple edges are changing stochastically at once, and at the same time, which is more efficient than calculating everything from scratch for every change. The work reported by them was inspired by the need for formulating an algorithm for finding the shortest path in such realistically occurring stochastic environments. Indeed, they sought to find the shortest path in the "average" graph (dictated by an "Oracle," also called the *environment*). Since, on query, the edge-weights supplied by the environment are assumed to follow an underlying unknown distribution, there exists a mean solution to the problem to which the algorithm would converge to after a sufficiently long time. Their intention was to find the "statistical" shortest path in the average graph that will be stable—regardless of the (possibly) continuously changing weights provided by the environment. Their solution generates superior results (when compared to the previous solutions). However, unfortunately, their scheme does not consider the insertion/deletion of edges. Thus, the problem they have considered assumes that there is one fixed structure graph with randomly changing edge weights, and that the distribution of these random variables is unknown.

Figure 1. Graph after the insertion of the edge C→F with weight 0.5



ALL-PAIRS SHORTEST PATH ROUTING: DYNAMIC VERSUS STATIC

Contrary to the previous discussions, in which the idea was of computing the shortest paths from one node to all the other nodes in a network, the problem of computing and maintaining all-pairs shortest paths information in a graph where the edges are inserted/deleted and edge-weights constantly increase/decrease is referred to as the *Dynamic*

all-pairs shortest paths problem (DAPSP) (Demetrescu & Italiano, 2003; Ramalingam & Reps, 1996). The DAPSP problem has found equal importance as the DSSSP problem, among researchers and practitioners alike. Both the problems aim to efficiently maintain shortest path solutions in environments that are representative of most practical situations in daily life, since most real-life environments are dynamically changing.

Similar to what was discussed for the DSSSP problem and for the DAPSP problem as well, it can be shown that edge-weight updates can be treated as edge-deletions and edge-insertions by setting the weights of edges to an infinitely large value (Demetrescu & Italiano, 2003). Likewise, an edge-update algorithm for the all-pairs shortest path problem is referred to as being fully dynamic if both weight-increase and weight-decrease operations are supported on the edges of the graph. A semi-dynamic algorithm handles only weight-increases or weight-decreases, but not both at the same time (Demetrescu & Italiano, 2003).

Like the static solutions of the single source shortest path problem, the well-known static solutions for the all-pairs shortest path problem, for example, the Floyd-Warshall's algorithm (Floyd, 1962), the all-pairs adaptations of Bellman-Ford's algorithm (Bellman, 1958), or the Dijkstra's algorithm (Dijkstra, 1959), which recompute the shortest paths from scratch each time a topology change occurs in the graph, are certainly inefficient in such dynamic practical scenarios.

Over the last few decades, there has been a lot of research done to solve the DAPSP problem. The earliest papers were written by Loubal (1967), Murchland (1967), and Rodinov (1968). Many other dynamic algorithms were proposed in the literature (e.g., Even & Gazit, 1985; Ramalingam & Reps, 1996; Rohnert, 1985); however, their worst-case running times were no better than recomputing the all-pairs shortest paths from scratch. Thereafter, a few solutions (Ausiello, Italiano, Marchetti-Spaccamela, & Nanni, 1991; Fakcharoemphol & Rao, 2001; Henzinger, Klein, Rao, & Subramanian, 1997; King, 1999) were proposed whose running times are better than a total recomputation. However, these solutions only work for integer weights. The algorithm proposed by Ausiello et al. (1991) is only applicable for the semi-dynamic case (decrease-only), and requires positive integer weights less than a constant "C." Their algorithm's amortized running time per insertion operation is O(Cnlogn) (Ausiello et al., 1991). Although Henzinger et al. (1997) provide a fully dynamic solution to the all-pairs shortest paths problem, their solution is only for planar graphs with integral values of edge-weights. The running time of their algorithm is $O(n^{4/3}\log(nC))$ per update operation. The first fully dynamic solution on general graphs was proposed by King (1999). Her solution too only works with positive integer weights less than C, and the running time of the algorithm is $O(n^{2.5}(C\log n)^{1/2})$. Later, Demetrescu and Italiano (2001) published a paper containing a fully dynamic algorithm that would perform edge-update operations on general graphs with real-valued edge-weights. If S represents the number of different real values, the amortized running time per update operation for their algorithm is $O(n^{2.5}(S\log^3 n))$. Finally, in 2003, Demetrescu and Italiano (2003) proposed a remarkable algorithm that solved the same problem for general digraphs with edge-weights that can assume positive real values but with a substantially improved running time per edge-update operation.

Misra and Oommen (2005a) proposed a learning algorithm that finds all-pairs shortest paths for the statistical average network topology, and the solution converges irrespective of whether there are new changes in edge-weights or not. They showed that their solution has better performance than the Demetrescu and Italiano (2003)'s solution. In their solution, not all the edges in a stochastic network topology are probed, and even if they are, they are not all probed equally often. Indeed, their algorithm attempts to almost always probe only those edges that will be included in the final list involving all pairs of nodes in the graph, while probing the other edges minimally. This increases the performance of the proposed algorithm. Their second contribution is the designing of a data structure, the elements of which represents the probability that a particular edge in the graph lies in the shortest path between a pair of nodes in the graph. Misra and Oommen (2005a)'s work is considered to represent the state of the art in this area.

APPLICATIONS

First of all, the application of shortest path algorithms in telecommunication is straightforward. They commonly find their use in routing equipments. In vast rapidly changing telecommunications wired or wireless networks, where links go up and down continuously and rapidly, and where there are simultaneous random updates in link costs, the existing algorithms are inefficient. Shortest paths need to be computed within a very short time (in the order of microseconds) with minimal number of nodes to be processed and edges to be scanned. Furthermore, in the *transportation domain*, there are complex road networks in urban areas. The costs of routing shipments from a warehouse to (all) other retail outlets constantly change. Changes occur because of a range of reasons such as road construction, accidents, traffic jams, office hours, and the presence of emergency vehicles. There are often several alternative routes that can be taken by a vehicle, whenever the predetermined route is unsuitable to deliver the shipments in time. Shipment vehicles may be equipped with suitable technology (like the global positioning system [GPS]) to guide them through alternative routes. The proposed learning algorithms can then be used to achieve such a vehicle routing. The same arguments

would be true if airplanes have to be redirected adaptively to take care of changes.

In the military, the proposed learning algorithms also have several potential applications. For example, in complex military networks, ground forces may have to be rapidly rerouted based on sudden changes in enemy information and strategies along existing routes, and self-guided missiles may need to change their trajectories (within microseconds) to account for changes in the path along which it travels.

FUTURE TRENDS

Interested researchers or practitioners could perform the following works to advance the state of the art in this area:

- Most of the aforementioned solutions (especially the recent ones) were reported on simulated networks of relatively small number of nodes. They have not been tested on massive sized networks. Testing the previous algorithms on large (on topologies with 10,000 to 100,000 nodes) simulated networks is required to understand their performance characteristics well.
- It is also required to test the performance of the algorithms in real-life networks, as most of the solutions mentioned previously have not been tried on such networks.
- Similarly, assessing the suitability of the aforementioned generic solutions in different application domains.

CONCLUSION

In this article we reviewed different shortest path solutions that have been proposed since 1958/1959 when Bellman and Ford, and Dijkstra proposed their solutions for a static network. Of all the algorithms, the solutions proposed by Misra and Oommen (2005a) and Misra and Oommen (2005a) represent the state of the art. The experimental evaluation of their algorithms shows the superiority of their proposed algorithms. The characteristic of their algorithm is that once their algorithms have converged, the average number of processed nodes, scanned links, and the time per update operation is superior to the previously proposed algorithms by a few orders of magnitude. The advantage of their solution is that in stochastic network environments, it possesses a statistical shortest paths list that should be actual shortest paths irrespective of whether there are new changes in link costs taking place continuously. In such cases, their solution converges to a shortest paths list, while the existing algorithms would recalculate affected shortest paths after every change in link cost.

REFERENCES

Ausiello, G., Italiano, G., Marchetti-Spaccamela, A., & Nanni, U. (1991). Incremental algorithms for minimal length paths. *Journal of Algorithms, 12*(4), 615-638.

Bellman, R. (1958). On a routing problem. *Quarterly of Applied Mathematics, 16*(1958), 87-90.

Demetrescu, C., Frigioni, D., Marchetti-Spaccamela, A., & Nanni, U. (2001). Maintaining shortest paths in digraphs with arbitrary arc weights: An experimental study, *Lecture Notes in Computer Science, 1982*, 218-229.

Demetrescu, C., & Italiano, G. (2001). Fully dynamic all-pairs shortest paths with real weights. *Proceedings of the 42nd IEEE Annual Symposium on Foundations of Computer Science (FOCS'01),* Las Vegas, NV (pp. 260-267).

Demetrescu, C., & Italiano, G. F. (2003). A new approach to dynamic all pairs shortest paths. *Proceedings of the 35th Annual ACM Symposium on the Theory of Computing* San Diego, CA (pp. 159-166).

Dijkstra, E. W. (1959). A note on two problems in connection with graphs. *Numerische Mathematik, 1,* 269-271.

Even, S., & Gazit, H. (1985). Updating distances in dynamic graphs. *Methods of Operations Research, 49,* 371-387.

Fakcharoemphol, J., & Rao, S. (2001). Planar graphs, negative weight edges, shortest paths, and near linear time. *Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science (FOCS'01)* Las Vegas, NV (pp. 232-241).

Floyd, R. W. (1962). Algorithm 97 (SHORTEST PATH). *Communications of the ACM, 5*(6), 345.

Franciosa, P. G., Frigioni, D., & Giaccio, R. (1997). Semi-dynamic shortest paths and breadth first search in digraphs. Symposium on Theoretical Aspects of Computer Science, (LNCS 1200, pp. 33-46).

Frigioni, D., Ioffreda, M., & Nanni, U. (1998). Experimental analysis of dynamic algorithms for the single source shortest path problem. *ACM Journal of Experimental Algorithmics, 3*, Article 5.

Frigioni, D., Marchetti-Spaccamela, A., & Nanni, U. (1996). Fully dynamic output bounded single source shortest path problem. *ACM-SIAM Symposium on Discrete Algorithms* (pp. 212-221).

Frigioni, D., Marchetti-Spaccamela, A., & Nanni, U. (2000). Fully dynamic algorithms for maintaining shortest paths trees. *Journal of Algorithms, 34,* 251-281.

Henzinger, M., Klein, P., Rao, S., & Subramanian, S. (1997). Faster shortest-path algorithms for planar graphs. *Journal of Computer and System Sciences, 55*(1), 3-23.

King, V. (1999). Fully dynamic algorithms for maintaining all-pairs shortest paths and transitive closure in digraphs. *Proceedings of the 40th IEEE Symposium on Foundations of Computer Science (FOCS'99)* (pp. 81-99).

Loubal, P. (1967). A network evaluation procedure. *Highway Research Record, 205,* 96-109.

McQuillan, J., Richer, I., & Rosen, E. (1980). The new routing algorithm for the ARPANET. *IEEE Transactions on Communications, 28*(5), 711-719.

Misra, S., & Oommen, B. J. (2005a, June 27-30). New algorithms for maintaining dynamic all-pairs shortest paths. *Proceedings of the 10th IEEE Symposium on Computers and Communications (IEEE ISCC 2005),* Cartagena, Spain.

Misra, S. & Oommen, B. J. (2005b, December). Dynamic algorithms for the shortest path routing problem: Learning automata-based solutions. *IEEE Transactions on Systems, Man, and Cybernetics, (Part B), 35*(6), (pp. 1179-1192).

Moy, J. (1997). OSPF Version 2, Internet Draft, RFC 2178. Retrieved from http://www.ietf.org/rfc/rfc2178.txt

Murchland, J. (1967). *The effect of increasing ordecreasing the length of a single arc on all shortest distances in a graph* (Tech. Rep., LBS-TNT-26). London, UK: London Business School, Transport Network Theory Unit.

Peterson, L., & Davie, B. (2000). *Computer networks:A systems approach*, (2nd ed.). San Francisco, CA: Morgan Kauffman Publishers.

Ramalingam, G., & Reps, T. (1996). On the computational complexity of dynamic graph problems. *Theoretical Computer Science, 158*(1), 233-277.

Rodinov, V. (1968). The parametric problem of shortest distances. *USSR Computational Mathematics and Mathematical Physics, 8*(5), 336-343.

Rohnert, H. (1985). A dynamization of the all-pairs least cost problem. *Proceedings of the 2nd Annual Symposium on Theoretical Aspects of Computer Science (STACS'85), LNCS 182* (pp. 279-286).

Schwartz, M., & Stern, T. (1980). Routing techniques used in computer communications networks. *IEEE Transactions on Communications, 28,* 539-552.

Spira, P., & Pan, A. (1975). On finding and updating spanning trees and shortest paths. *SIAM Journal of Computing, 4*(3), 375-380.

KEY TERMS

Algorithm: An algorithm is a set of clear steps that is used to define how a task or a set of tasks can be accomplished.

Dynamic Shortest Path Algorithm: An algorithm that is capable of finding a path that has the least distance (among all possible paths) between a pair of source and destination nodes in a network, when the status of nodes and links change with time.

Fully Dynamic All-Pairs Shortest Path Algorithm: If in a dynamic shortest-path algorithm, all edge operations (edge-insertion, weight-decrease, edge-deletion, weight-increase) are allowed, the algorithm is referred to as the fully dynamic algorithm.

Routing: The mechanism by which a path or a set of paths is selected to send information or commodities.

Semi-Dynamic Shortest Path Algorithm: If in a dynamic shortest-path algorithm, only edge-insertion/weight-decrease, or edge-deletion/weight-increase is allowed, the algorithm is referred to as the semi-dynamic algorithm.

Shortest Path Algorithm: An algorithm that is capable of finding a path that has the least distance (among all possible paths) between a pair of source and destination nodes in a network.

Static Shortest Path Algorithm: An algorithm that is capable of finding a path that has the least distance (among all possible paths) between a pair of source and destination nodes in a network, when all the nodes and links remain stationary.

ENDNOTE

[1] It should, however, be noted that some routing protocols in MANETs are based on the knowledge of a partial topology and provide the shortest path to any destination in the network. The advertised topology is a subset of the whole topology. Hence, ideally, only link-ups and link-downs concerning the advertised topology must be accounted for in the whole network.

Signal Processing Techniques for Audio and Speech Applications

Hector Perez-Meana
National Polytechnic Institute, Mexico

Mariko Nakano-Miyatake
National Polytechnic Institute, Mexico

INTRODUCTION

Since the apparition of the first standalone digital signal processor (DSP) in 1980, the development of very-large-scale integration (VLSI) technology has allowed an impressive improvement on the performance of signal processing devices. This fact has made it possible to implement more efficient systems for storage, transmission, enhancement, protection, and reproduction of speech and audio signals. Some of these successful applications, shown in Table 1, have contributed to improving the performance of communications, storage, and medical systems, as well as security and copyright protection.

BACKGROUND

Since the apparition of electronics technology, several devices have been introduced to improve received, produced, and recorded audio and speech signal quality—such as that of low pass and band pass analog filters, amplifiers, and so forth—that allowed suppression of some kinds of interfering signals. Next, with the development of solid state technology in the late of 1960s and the apparition of the op-amp, several analog signal processes were developed; although there were limitations, these allowed import improvements in audio and speech systems. The limitations of analog technology encouraged the development of digital technology. As a result in 1980 the Nippon Electric Corporation (NEC) and American Telegraph and Telephone (AT&T) released the first standalone complete digital signal processors, the PD7220 and the DSP1. Since then, digital signal processing technology has experienced impressive growth, allowing performance improvement of already available systems, as well as development of many other successful systems in several other fields, some described in this article.

Table 1. Main audio and speech signal processing applications

• Echo cancellation in telecommunication systems
• Acoustic echo cancellation
• Noise canceling
• Adaptive equalization
• Narrowband speech coding
• Broadband audio and speech coding
• Medical applications
• Watermarking

REVIEW OF MAIN SIGNAL PROCESSING APPLICATIONS IN SPEECH AND AUDIO FIELDS

As mentioned before, the signal processing applications in speech and audio fields have increased, contributing to the solution of many important problems, constituting an important part of many practical systems. To understand the importance of this technology and how it has contributed in the development of audio and speech fields, this section provides a review of some successful signal processing systems.

Echo Cancellation for Long-Distance Transmission

A very successful speech signal processing application is the adaptive echo cancellation used to reduce a common but undesirable phenomenon in most telecommunications systems, called echo. Here when mismatch impedance is present in any telecommunications system, a portion of the transmitted signal is reflected to the transmitter as an echo, which represents an impairment that degrades the system quality (Hansler & Schmidt, 2006; Manolakis, Ingle, & Kogon, 2005). In most telecommunications systems, such as a telephone circuit, the echo is generated when the long-distance portion consisting of two one-directional channels (four wires) is connected with a bidirectional channel (two wires) by means of a hybrid transformer. If the hybrid impedance is perfectly balanced, the two one-directional channels are uncoupled and no signal is returned to the transmitter side (Hansler & Schmidt, 2006; Manolakis et al., 2005). However in general the bridge is not perfectly balanced because the

required impedance to properly balance the hybrid depends on the overall impedance network. In this situation, part of the signal is reflected, producing an echo. To avoid this problem an adaptive filter is used to generate an echo replica, which is then subtracted from the signal to be transmitted. Subsequently, the adaptive filter coefficients are updated to minimize, usually, the mean square value of the residual echo (Madisetti & Williams, 1998; Perez-Meana, 2007). To obtain an appropriate operation, the echo canceller impulse response must be larger than the longer echo path to be estimated. Thus assuming a sampling frequency of 8kHz and an echo delay of about 60ms, an echo canceller with 256 or more taps is required (Manolakis et al., 2005; Perez-Meana, 2007). Besides the echo path estimation, another important problem is how to handle the double talk—that is, the simultaneous presence of the echo and the near speech signal (Hansler & Schmidt, 2006; Manolakis et al., 2005).

Acoustic Echo Cancellation

A critical problem affecting speech communication in a teleconferencing system is the acoustic echo. When a bi-directional line links two rooms, the acoustic coupling between the loudspeaker and microphones in each room causes an acoustic echo perceivable to the users in the other room. The best way to handle it appears to be the adaptive echo cancellation. An acoustic echo canceller generates an echo replica and subtracts it from the signal picked up by the microphones. The residual echo is then used to update the filter coefficients such that the mean square value of approximation error is kept to a minimum (Perez Meana, Nakano-Miyatake, & Nino-de-Rivera, 2002.; Huang & Banesty, 2005).

Adaptive Noise Cancellation

The adaptive noise canceller is a generalization of the echo canceller in which a signal corrupted with additive noise must be restored or enhanced. When a reference signal correlated with the noise signal but uncorrelated with the desired one is available, the noise cancellation can be achieved by using an adaptive filter to minimize the total power of the output of the difference between the corrupted signal and the estimated noise, such that the resulting signal becomes the best estimate, in the mean square sense, of the desired signal. This system works fairly well when the reference and the desired signal are uncorrelated among them. However, appropriate reference signals are not always available. To solve this problem several noise canceling algorithms have been proposed which are resistant to crosstalk situations. A different approach, developed by Dolby Laboratories, is used in the Dolby noise reduction systems in which the dynamic range of the sound is reduced during recording and expanded during the playback (Davis, 2002; Hansler & Schmidt, 2006). Several types of Dolby noise reduction systems have been developed including the A, B, C, and HXpro. Most widely used is the Dolby B, which allows acceptable playback even on devices without noise reduction. The Dolby B noise reduction system uses a pre-emphasis that allows masking the background hiss of a tape with a stronger audio signal, especially at higher frequencies. This effect is called psychoacoustic masking (Davis, 2002; Perez-Meana, 2007).

A related problem to noise cancellation is the active noise cancellation that intends to reduce the noise produced in closed places by several electrical and mechanical pieces of equipment such as home appliances, industrial equipment, air conditioning units, airplanes turbines, motors, and so forth. Active noise is achieved by introducing a canceling

Table 2. Digital speech coding standards

Rate Kb/s	Application	Type of Coder	Year
64	Public Switched Telephone Network	Pulse Code Modulation (PCM)	1972
2.4	U.S. Government Federal Standard	Linear Predictive Coding	1977
32	Public Switched Telephone Network	Adaptive Differential PCM	1984
9.6	Skyphone	Multi-Pulse Linear Predictive Coding (MPLPC)	1990
13	Pan-European Digital Mobile Radio (DMR) Cellular System (GSM)	Regular Pulse Excitation Linear Prediction Coding (RPE-LPC)	1991
4.8	U.S. Government Federal Standard	Codebook Excited Linear Prediction Coding (CELP)	1991
16	Public Switched Telephone Network	Low Delay CELP (LD-CELP)	1992
6.7	Japanese Digital Mobile Radio (DMR)	Vector Sum Excited Linear Prediction Coding (VSELP)	1977

anti-noise wave through an appropriate array of secondary sources, which are interconnected through an electronic system using adaptive systems with a particular cancellation configuration. Here, the adaptive filter generates an anti-noise that is acoustically subtracted from the incoming noise wave. The resulting wave is captured by an error microphone and used to update the noise canceller parameters, such that the total error power is minimized. This technology has been successfully applied in earphones, electronic mufflers, noise reduction systems in airplane cabins, and so forth (Davis, 2002; Kuo & Morgan, 1996).

Speech and Audio Coding

Besides interference cancellation, speech and audio signal coding are other very important signal processing applications. This is because low bit rate coding is required to minimize the transmission costs or provide cost-efficient storage. Here we can distinguish two different groups: the narrowband speech coders used in telephone and some video telephone systems in which the quality of telephone-bandwidth speech is acceptable, and the wideband coders used in audio applications that require a bandwidth of at least 20kHz for high fidelity (Madisetti & Williams, 1998; Spanias, Painter, & Atti, 2007; Huang & Benesty, 2005).

Narrowband Speech Coding

The most efficient speech coding systems for narrowband applications use an analysis-synthesis-based method in which the speech signal is analyzed during the coding process to estimate the main parameters of speech that allow its synthesis during the decoding process. Two sets of speech parameters are usually estimated: (1) the linear filter system parameters that model the vocal track, estimating use of the linear prediction method, and (2) the excitation sequence. Most speech coders estimate the linear filter in a similar way, although several methods have been proposed to estimate the excitation sequence that determines the synthesized speech quality and compression rates. Among these speech coding systems, we have the multi-pulse and regular pulse linear predictive coding and the codebook exited linear predictive coding (CELP), which achieve bit rates among 9.6 Kb/s and 2.4 kb/s, with reasonable good speech quality (Madisetti & Williams, 1998; Huang & Benesty, 2005; Spanias et al., 2007). Table 2 shows the main characteristics of some of the most successful speech coders.

Wideband Audio and Speech Coding

Bandwidths higher than that of telephone bandwidth result in major subjective improvements. Thus a bandwidth of 50 to 20 kHz not only improves the intelligibility and naturalness of audio and speech, but also adds a feeling of transparent communication and eases speaker recognition. However, this will result in the necessity to store and transmit a much larger amount of data, unless efficient wideband coding schemes are used. Wideband speech and audio coding intend to minimize the storage and transmission costs, while providing an audio and speech signal with not audible differences between the compressed and the actual signals with 20kHz or higher bandwidth and a dynamic range equal to or above 90 dB. Four key technology aspects play a very important role in achieving this goal: perceptual coding, frequency domain coding, window switching, and dynamic bit allocation. Using these features, the speech signal is divided into a set of non-uniform subbands to encode with more precision the components perceptually more significant and with fewer bits the perceptually less significant frequency components. The subband approach also allows the use of the masking effect in which the frequency components close to those with larger amplitude are masked and then they can be discharged without audible degradation. These features, together with a dynamic bit allocation, allow significant reduction of the total bits required for encoding the audio signal without perceptible degradation of the audio signal quality. Some of the most representative coders of this type are listed in Table 3 (Kondoz, 1994; Madisetti & Williams, 1998; Spanias et al., 2007; Huang & Benesty, 2005).

Medical Applications

Signal processing has been successfully used to improve the life quality of persons with hearing and speaking problems (Davis, 2002; Hansler & Schmidt, 2006). Among them we have the development of hearing aid devices, which attempt selectively to amplify the frequencies if the sound is not properly perceived. The enhancement of alaryngeal speech is another successful application in which signal processing and pattern recognition methods are used to improve the intelligibility and speech quality of persons whose larynx and vocal cords have been extracted by a surgical operation. Signal processing algorithms have also been developed to

Table 3. Some of the most used wideband speech and audio coders

Coder Bitrate Application
• CCITT G.722 64 kbits/s, 56 kbits/s, 48 kbits/s Speech
• Low Delay CELP 32 kbits/s Speech
• Compact Disc 1.41 Mbits/s Audio
• Perceptual Audio Coder 128 kbits/s Audio
• MP3 (MPEG-1 layer III) 96 kbits/s Audio
• Windows Media Audio 64 kbits/s Audio
• VQF 80 kbits/s Audio
• Mp3PRO 64 kbits/s Audio
• OGG Vorbis 96 kbits/s Audio
• WAV 10 MB/min Audio

modify the time scale of speech signal to improve the hearing capabilities of elderly people.

Watermarking

Digital watermarking is a technique used to embed a collection of bits into a signal in such way that it will be kept imperceptible to users and the resulting watermarked signal remains with almost the same quality as the original one. Watermarks can be embedded into audio, images, videos, and other formats of digital data in either the temporal or spectral domains (Cox, Miller, & Bloom, 2001; Perez-Meana, 2007; Spanias et al., 2007). Here the temporal watermarking algorithms embed watermarks into audio signals in their temporal domain, while the spectral watermarking algorithms embed watermarks in certain transform domains, such as the Fourier transform domain, wavelet domain, or cepstrum domain. Depending on their particular application, the watermarking algorithms can be classified as robust or fragile watermarks. Here the robust watermarking algorithms, which cannot be removed by common signal processing operations, are used for copyright protection, distribution monitoring, copy control, and so forth, while the fragile watermark, which will be changed if the host audio is modified, is used to verify the authenticity of an audio or speech signal.

FUTURE TRENDS

Audio and speech processing have achieved an important development during the last three decades, however several problems still remain that must be solved, such as how to develop more efficient echo canceller structures with improved double talk control systems. In adaptive noise canceling a very important issue that remains unsolved is the crosstalk problem. To get efficient active noise cancellation systems, it is necessary to cancel the anti-noise wave inside the reference microphone which distorts the reference signal to reduce the computational complexity of ANC systems and develop a more accurate secondary path estimation. Other important issues are to develop low distortion speech coders for bit rates below 4.8 kBits and increase the convergence speed of adaptive equalizers, to allow the tracking of fast time varying communication channels. The speech and audio processing systems will also contribute to improve the performance of medical equipment such as hearing aids and alaryngeal speech enhancement systems, as well as to security through the development of efficient and accurate speaker recognition and verification systems. Finally, in recent years digital watermarking algorithms have grown rapidly, however several issues remain such as developing an efficient algorithm taking into account the human auditory system (HAS), solving synchronization problems using multi-bit watermarks, and developing efficient watermarking algorithms for copy control.

Table 4. Other successful audio and speech applications

• Signal processing for hearing aids
• Virtual musical instruments synthesis
• Alaryngeal speech enhancement
• Cross-language voice conversion
• Speech and speaker recognition
• Adaptive equalization

CONCLUSION

Audio and speech signal processing have been fields of intensive research during the last three decades, becoming an essential component for interference cancellation and speech compression and enhancement in telephone and data communication systems, high-fidelity broadband coding in audio and digital TV systems, speech enhancement for speech and speaker recognition systems, and so forth. However despite the development that speech and audio systems have achieved, the research in those fields is increasing in order to: provide new and more efficient solutions in the above-mentioned fields and several others, such as the acoustic noise reduction to improve the environmental conditions of people working in airports, factories, and the like; improve the security of restricted places through speaker verification systems; and improve the speech quality of alaryngeal people through more efficient speech enhancement methods. Thus it can be predicted that speech and audio processing will contribute to more comfortable living conditions in future years.

REFERENCES

Bosi, M., & Goldberg, R. (2002). *Introduction to digital audio coding and standards.* Boston: Kluwer Academic.

Cox, I., Miller, M., & Bloom, J. (2001). *Digital watermark: Principle and practice.* New York: Morgan Kaufmann.

Davis, G. (2002). *Noise reduction in speech applications.* New York: CRC Press.

Gold, B., & Morgan, N. (2000). *Speech and audio signal processing.* New York: John Wiley & Sons.

Hansler, E., & Schmidt, G. (2006). *Topics in acoustic and noise control: Selected methods for cancellation of acoustic echoes, the reduction of background noise and speech processing.* Berlin: Springer-Verlag.

Haykin, S. (1991). *Adaptive filter theory.* Englewood Cliffs, NJ: Prentice Hall.

Huang, Y., & Benesty, J. (2005). *Audio for next generation multimedia communication systems.* Norwell, MA: Kluwer Academic.

Kondoz, A.M. (1994). *Digital speech.* Chichester, England: John Wiley & Sons.

Kuo, S., & Morgan, D. (1996). *Active noise control system: Algorithms and DSP implementations.* New York: John Wiley & Sons.

Madisetti, V., & Williams, D. (1998). *The digital signal processing handbook.* Boca Raton, FL; CRC Press.

Manolakis, D., Ingle, V., & Kogon, S. (2005). *Statistical and adaptive signal processing.* Norwood, MA: Artech House.

Perez-Meana, H., Nakano-Miyatake, M., & Nino-de-Rivera, L. (2002). *Speech and audio signal application. Multirate systems: Design and applications* (pp. 200-224). Hershey, PA: Idea Group.

Perez-Meana, H. (2007). *Advances in audio and speech signal processing: Technologies and applications.* Hershey, PA: Idea Group.

Spanias, A., Painter, T., & Atti, V. (2007). *Audio signal processing and coding.* New York: John Wiley & Sons.

KEY TERMS

Adaptive Filter: Linear system that modifies its parameters minimizing some given criterion of the difference between its output and a given reference signal. Widely used in echo and noise canceling, equalization of communication channels, antenna arrays, and so forth.

Adaptive Algorithm: Method used to modify the filter coefficients, online, in order to minimize the power of an adaptive filter output error.

Alaryngeal Speech: Speech produced by persons whose larynx and vocal cords have been extracted by a surgical operation.

Anti-Noise: Estimated replica of acoustic noise generated by an active noise canceller system, which is used to cancel an environmental noise.

Crosstalk: Interference present in a signal propagating through a communication produced by another signal present at an adjacent channel.

Digital Watermarking: A technique used to embed a collection of bits into a host signal in such a way that the watermark remains imperceptible to the users.

Double Talk: An interference produced when the speakers in both ends of a telephone line simultaneously speak. This phenomenon greatly disturbs the echo canceller performance.

Hybrid Transformer: Device used to connect two one-directional channels with a bi-directional channel, keeping uncoupled among them the two one-directional channels.

Narrowband Speech Signal: Speech signal with a frequency band equal to that of the telephone channel — that is, with a bandwidth of 300 to 3,300 kHz.

Signal Compression: Signal coding that allows a reduction of the total number of bits required to represent a given signal without distortion or with negligible distortion.

Speech Enhancement: Signal processing performed in a given speech signal to improve its intelligibility and signal-to-noise-ratio.

Speaker Verification: Signal processing required for verifying the speaker identity by using his or her speech features.

Wideband Signal: Signal with a bandwidth wider than that of the telephone channel, usually between 50 Hz and 20 kHz.. This fact results in major subjective improvements.

Simulation for Supporting Business Engineering of Service Networks

Marijn Janssen
Delft University of Technology, The Netherlands

INTRODUCTION

Today, the services industry provides the majority of all jobs in Western countries, and services tend to be delivered more and more using the Internet. The service economy is becoming increasingly dominant in developed economies, with knowledge assets playing a great role relative to physical and financial assets. (Rouse & Baba, 2006)

Services are often characterized as intangible, perishable, experience-based, difficult-to-standardize products needing many interactions between customers and services providers. Grönroos (2001) identified three basic characteristics of services:

1. Services are processed consisting of activities or a series of activities rather than things.
2. Services are at least to some extent produced and consumed simultaneously.
3. The customer participates in the service delivery process.

All kinds of information and communication technology (ICT) are applied to support the creation of service networks. *Service networks* are constellations of independent organizations that work together in various configurations in order to deliver services. The provisioning of services can be viewed as a series of activities leading to some observable behavior between service providers (or service brokers) and service requesters. They are delivered using the Internet, accessible from any place at any time and often involve no direct human involvement of the service provider. The term *e-services* is typically used to describe a variety of electronic interactions ranging from basic services, such as the delivery of news and taking out an insurance policy, to more complex services, such as the delivery of context-aware, personalized services.

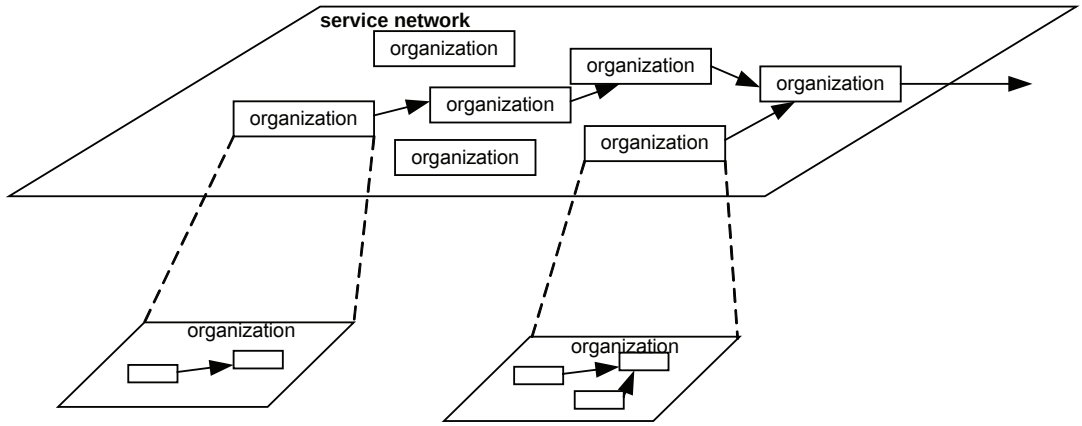
To understand a service network, both the network and the decision makers involved need to be understood. Stakeholder theory states that those who can effect change or be affected by it should be accounted for in the transformation process (Pfeffer, 1981). The diversity of key stakeholders and their interests makes evaluating the design and the efficiency and effectiveness of service networks very complex. Often stakeholders are characterized by opposing interests, having heterogeneous systems and being part of multiple service networks. The effective management of such services network is key to success, which requires understanding each other's interests, business processes, and information systems. Often organization network managers, a particular type of electronic intermediary (e-intermediary), specialize in coordinating such networks (Janssen & Verbraeck, 2005b). Design decisions are critical, as they determine the efficiency and effectiveness of the service networks. The development and growth of service networks requires the developer to carefully identify, evaluate, and understand the possible impact of the various design alternatives. A business engineering methodology can be of help in designing and developing service networks by providing insight into current network structure and potential structure, and by evaluating the implications of potential arrangements. Simulation can be used to compare the performance of the current and possible situations in a business engineering methodology. Simulation of service networks is much more difficult than physical networks, as the products often concerns intangibles. The *objective* of this article is to discuss research issues concerning the simulation of service networks to support business engineering.

BACKGROUND

Service networks can use a large variety of coordination mechanisms and structures to coordinate the activities of participants. Service networks consist of organizations cooperating together, and the management of the networks has a large impact on the total performance. Figure 1 shows a service network schematically.

The dynamic nature of service networks—that is, the changing number and/or types of partners, and the involvement in several networks—increases the difficulty and complexity to understand the dependencies in the network. The creation of flexible, temporary service networks results in the creation of business processes that are no longer self-contained within a single organization. The effectiveness of service networks depends more and more on the performance of external partners that are often unknown and viewed as black boxes (Tewoldeberhan & Janssen, 2007). Therefore,

Figure 1. Example of a service network



some organizations can be considered as white boxes and others as black boxes in the networks, as schematically depicted in Figure 1. Not all organizations must be involved in each service provisioning process. Organizations might be selected and dynamically assembled based on the services needed by the customers. A service network consists of multiple businesses having varying types of relationships.

The core of a service network is the *coordination* of the various interdependent activities performed by autonomous organizations. There are two opposing views on coordination. In a *coordination of tasks* approach, the design of processes is dependent on the coordination mechanisms that manage the dependencies between tasks (Malone & Crowston, 1994). The *coordination of commitments* approach emphasizes networks of commitments that organizations establish through intentional acts of speech (Winograd & Flores, 1987). This coordination approach emphasizes the fulfillment of human commitments and describes activities in terms of contracts and promises. A traditional approach to supply chains is the coordination of tasks view. In a service network, both views apply, as the activities performed by the independent organization needs to be coordinated in order to agree on and fulfill commitments.

The requirements of an organization are not easily elicited and can demand innovative mechanisms or deliberate trade-offs. The timely sharing of information among organizations is often a major issue (Christopher, 2003). Information sharing is necessary for efficient coordination of the service network and to optimize performance. The organizations making up the network often want to avoid that information is provided to other network members. Information might be used to negotiate lower prices or undermine competitive advantage, as competitors might learn from it. Another typical issue in the business engineering of service networks is the selection of coordination mechanisms, as members can have different and even opposing requirements. For

Table 1. A list of business engineering issues

• Which information architectures and structures are most beneficial in which situations?
• Aligning mechanism with service and markets characteristics?
• Integration of the information systems of network members?
• How to manage the service network?
• Should intermediaries be used to coordinate the service network?
• Level of coarse and fine-grained services
• Evaluation of implications of changes
• Pooling and sharing of services
• Conflicting interests of network members
• Incomplete information, ensuring information privacy
• Ensuring quality of product to buyer and payment to sellers
• Tracking and tracing
• Reducing transaction risk and increasing trust
• Use of software agents as assistance in search and evaluation
• Intelligent product and vendor matching mechanisms
• Product distribution/delivery
• Creating and disseminating product information
• Information processing and aggregation
• Open and closed networks
• Spot sourcing for dynamic networks vs. systematic sourcing for sustainable networks
• Type of management information and dissemination of management information

example one organization strategy might be to minimize trading time, while another might want to minimize costs. The most conspicuous opposing requirement is that buyers want to have the lowest price at the best possible trading conditions while sellers want to have the highest possible price to maximize revenue.

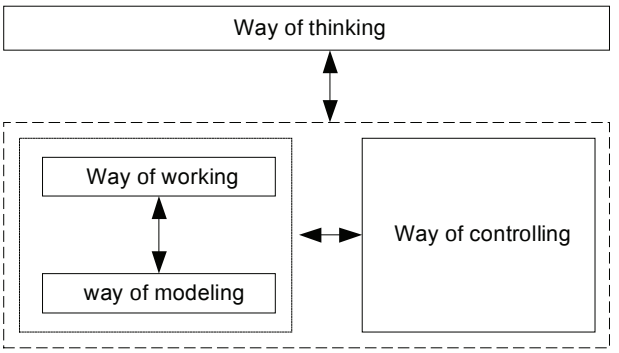
In short, a large number of other trade-offs and decisions need to be made before an efficient and effective service network can be established. A list of business engineering issues is shown in Table 1. Some limitations are coming from the state of the art of the technology and from market and/or service characteristics; others are coming from the opposing requirements and needs of the parties involved. Simulation for business engineering of service networks can help decision makers gain insight into these issues. This should support them in making deliberate choices without having to experiment in real life, which could be costly and even result in a loss of customers.

SIMULATION-SUPPORTING BUSINESS ENGINEERING

Service networks are by nature complex, and analytic methods can only be applied in a limited way. Although these approaches contribute to insight into and design of service networks, they do not help decision makers evaluate the impact and support their decision making in practice. Ideally, the implications of changes should be evaluated prior to implementation on criteria such as costs, utilization, trading time, delivery time, number of bids, matching chance, and so on. Analytical approaches also do not grasp the time-dependent dynamics resulting from the interplay between actors executing business processes.

A business engineering methodology can be seen as a continuum of approaches to process change (Kettinger, Teng, & Guha, 1997). Such a methodology is often seen as a way to tackle issues from an engineering perspective, as well as from a social perspective. The analytical framework of Sol (1990) provides a suitable way to describe business engineering methodologies. This framework classifies design methodologies by ways of thinking, working, modeling, and controlling. The *way of thinking* describes the philosophy on which the design methodology is based. It provides the basic assumptions of the business engineering approach and should contain, for example, whether it is aimed at radically changing the service network or at step-wise improvement, and will outside experts be used to analyze the situation or will the business engineering approach be based on stakeholders' participation? The *way of working* provides the subsequent steps that should be carried out to arrive at a new situation. An approach supporting relatively small steps and letting stakeholders participate so that they can gain the necessary knowledge to formulate their own incremental

Figure 2. Analytical framework for business engineering methodologies (based on Sol, 1990)



improvements or a radical approach can be taken. The *way of modeling* refers to the concepts that are used to abstract reality into models of the problem domain. Often models are used to support communication and evaluate the impact of changes. The *way of controlling* or management approach provides the components needed for the management of a design project. Often a project management approach is taken using milestones to guide the process.

The way of modeling provides the necessarily support for a business engineering approach. Simulation of business processes constitutes one of the most widely used applications of operational research, as it allows us to understand the essence of business systems, to identify opportunities for change, and to evaluate the effect of proposed changes on key performance indicators (Law & Kelton, 1991). The philosophy behind a business engineering approach is to develop a simulation model of the service network, experiment with this model, and experiment with alternative situations (Sol, 1982). An analysis of service networks needs to begin with an understanding of current processes and should investigate how conventional transaction methods are changed as a result of ICT adoption. One of the advantages of simulation is that what-if analysis can be carried out without changing reality at lower costs. These analyses often compare situations using time- and cost-based performance indicators such as delivery time, utilization of resources, cost of activities, and waiting time.

Animation is often a standard feature of simulation. An *animation* model is a graphical representation of a problem situation and includes visualization of the time-ordered dynamics of objects, a static background, an overview of performance indicators, and a user-interface (de Vreede & Verbraeck, 1996). The purposes of animation are to facilitate decision makers to acquire insight into the dynamic interactions of the modeled system, the performance of the 'as is' and 'to be' situation, and to facilitate communication between parties involved in a dynamic modeling study. Effecting enterprise-wide technology and business process change in the service networks is a massive and complex

undertaking, and animation might help to create a shared vision and understanding among participants.

FUTURE TRENDS

Business engineering methodology provides context to the simulation technique (Greasley, 2003). Business engineering approaches proposed by Streng (1994) and Giaglis, Paul, and Doukidis (1999) tackle the analysis of the added value of technology-enabled changes by making use of discrete-event simulation. Nikolaidou and Anagnostopoulos (2003) use a simulation approach for modeling distributed systems.

Dependent on the characteristics of the service network under study, certain issues shown in Table 1 dominate and should be simulated. A business engineering methodology should ensure the incorporation of the relevant business engineering issues, and simulation can be used to aid conscious decision making. A business engineering approach should help to focus on those issues most relevant to the particular situation and help to find solutions to those issues. A fruitful research direction seems to be which issues should be included in the business engineering for which types of situations.

The ability to create flexible alliances with partners to form supply chains or business networks becomes more and more important for businesses. Almost all of the current attention seems to be geared toward how static service networks can be configured, almost completely neglecting the dynamism of the environment in which a service network operates. Little attention is given to how these service networks evolve and emerge over time, or how environmental changes that require the service network to adapt can be dealt with. Various types of changes and their possible impact on the complete network and also on the individual organizations should be captured by further research.

In the early stages of service networks, management resembled like traditional business relationships, and the network participants where often modeled using black-box approaches. Most studies assume that all data was available, however in many service networks this might not be a valid premise. Many organizations will be reluctant to share data about their mission-critical business processes with others. Thus, the efficiency of service networks will be hampered by the fact that organizations often demand that their data and inner business processes remain hidden from the other organizations in the supply chain. Future research should also include the modeling of service networks for situations where the supply chain is not completely known and limited data is available.

Web services technologies are more and more used to support the creation of service networks and especially the use of Web service orchestration technology to coordinate the interactions between network participants. Fast adoption is often hampered by the need for experimentation to make efficient use of this technology (Tewoldeberhan & Janssen, 2007). As such, more insight in business engineering, simulation, and the effect of Web service technology is a research direction. Especially to evaluate the implications of new technologies, simulation of a service network should on *emulation*, which means that actual software is used and embedded in the simulation experiment. This combines a simulation and prototyping approach into one modeling approach.

A large number of independent organizations with their own strategies and sometimes even opposing aims carry out business with each other. The relations between organizations can change during the trading process, as organizations can enter or leave the playing field. The most powerful abstractions are the ones that minimize the semantic gap between the units of analysis that are intuitively used to conceptualize the problem and the constructs present in the modeling approach. Ideally, an organization should be simulated as distributed systems, where each system is represented by an autonomous entity. Software agents are autonomous entities that can be used for simulating organizations. The so-called agents are autonomous entities with their own interests and goals so they can decide to enter or leave a trading situation. *Agent-based simulation* has appeared for modeling organizations within electronic markets (Ramat & Preux, 2003; Janssen & Verbraeck, 2005a). Janssen and Verbraeck (2005a) developed an agent-based simulation technique for electronic markets. Agent-based simulation approaches view systems in terms of autonomous agents that engage in interactions to coordinate their activities. The coordination problem in a system consisting of a number of agents is analogous to the coordination problem of independent organizations trading with each other. This approach might also be suitable for a service network and seems to be a feasible research direction.

Communication over the Internet using a *Web-based* simulation is preferable for supporting communication between the researcher and persons involved in the design process. Another research issue is to use *distributed* simulation so multiple participants could interact with a simulation environment at the same time. In this way the participants can gain experiences with a hypothetical service network. This might lead to an increase in insight into the problem situation by participants and might help designers to make better decisions.

CONCLUSION

Organizations cooperate and compete more with each other in service networks, and the efficient and effective management of such networks determines success. Business engineering is aimed at creating change by simultaneously considering

organizational and technical aspects, and simulation can make change possible by supporting the business engineering approach. During the business engineering of service networks, numerous trade-offs and decisions must be made, influencing the performance and possible adoption. Business engineering using simulation should help decision makers focus on the most relevant issues and make appropriate decisions and trade-offs without having to experiment in real-life situations. It is still unclear what the most relevant issues are to focus on when engineering service networks.

Ideally, service networks should be modeled representing reality as close as possible. All types of electronic intermediaries can support the coordination of the dependencies among organizations in the service network. A fruitful direction seems to be the use of a distributed, agent-based simulation, making use of emulated mechanisms. In this way emergent behavior and changes can be modeled and their impact evaluated. Ideally, decision makers of various organizations should be able to view the animation of the simulation over the Internet, manipulate parameters, and view the consequences of their actions. The accomplishment of this ideal needs ample research attention in the domain of business engineering methodologies, distributed simulation, agent-based simulation, Web-based animation, and emulation of mechanism.

REFERENCES

Christopher, J. (2003). The effect of delays in information exchange in electronic markets. *Journal of Organizational Computing and Electronic Commerce, 12*(2), 121-131.

de Vreede, G.J., & Verbraeck, A. (1996). Animating organizational processes: Insight eases change. *Simulation Practice and Theory, 4*(4), 245-263.

Giaglis, G.M., Paul, R.J., & Doukidis, G.I. (1999). Dynamic modeling to assess the business value of electronic commerce. *International Journal of Electronic Commerce, 3*(3), 35-52.

Greasley, A. (2003). Using business-process simulation within a business-process reengineering approach. *Business Process Management Journal, 9*(4), 408-420.

Grönroos, C. (2001). *Service management and marketing. A customer relationship management approach.* Chichester: John Wiley & Sons.

Janssen, M., & Verbraeck, A. (2005a). An agent-based simulation testbed for evaluating Internet-based matching mechanisms. *Simulation Modeling Practice and Theory, 13*(5) 371-388.

Janssen, M., & Verbraeck, A. (2005b). Evaluating the information architecture of an electronic intermediary. *Journal of Organizational Computing and Electronic Commerce, 15*(1), 35-60.

Kettinger, W.J., Teng, J.T.C., & Guha, S. (1997). Business process change: A study of methodologies, techniques, and tools. *MIS Quarterly, 21*(1), 55-79.

Law, A.M., & Kelton, D.W. (1991). *Simulation modeling and analysis.* New York: McGraw-Hill.

Luo, Z., Jennings, N.R., Shadbolt, N., Leung, H., & Lee, J.H. (2003). A fuzzy constraint based model for bilateral, multi-issue negotiations in semi-competitive environments. *Artificial Intelligence, 148,* 53-102.

Malone, T.W., & Crowston, K. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys, 26,* 87-119.

Nikolaidou & Anagnostopoulos (2003). A distributed system simulation modeling approach. *Simulation Modeling Practice and Theory, 11,* 251-267.

Pfeffer, J. (1981). *Power in organizations.* Marshfield, MA: Pitman.

Ramat, E., & Preux, P. (2003). Virtual laboratory environment (VLE): A software environment oriented agent and object for modeling and simulation of complex systems. *Simulation Modeling Practice and Theory, 11,* 25-55.

Rouse, W.B., & Baba, M.L. (2006). Enterprise transformation. *Communications of the ACM, 49*(7), 67-72.

Sol, H.G. (1982). *Simulation in information systems development.* Doctoral Dissertation, University of Groningen, The Netherlands.

Sol, H.G. (1990). Information systems development: A problem solving approach. *Proceedings of the International Symposium on System Development Methodologies,* Atlanta, GA.

Streng, R.J. (1994). *Dynamic modeling to assess the value of electronic data interchange: A study in the Rotterdam port community.* Doctoral Dissertation, Delft University of Technology, The Netherlands.

Tewoldeberhan, T., & Janssen, M. (2007). Simulation-based experimentation for designing reliable and efficient Web service orchestrations in supply chains. *Electronic Commerce Research and Application, 6.*

Winograd, T., & Flores, F. (1987). *Understanding computers and cognition. A new foundation for design. Reading, MA:* Addison-Wesley.

KEY TERMS

Agent-Based Simulation: Simulates organizations as interacting autonomous entities with their own interests and goals.

Animation Model: A graphical representation of a problem situation which can consist of a visualization of the time-ordered dynamics of objects, a static background, an overview of performance indicators, and a user interface.

Business Engineering: The integral design of both organizational structures and information systems.

Coordination of Commitments: The actions by humans leading to the completion of work. Coordination is described in terms of contracts and promises consisting of recurring loops of requesting, making, and fulfilling commitments.

Coordination of Tasks: The management of dependencies between tasks.

Discrete-Event Simulation: Models a system by changing the systems state at discrete points in time.

Emulation: Actual software is written to execute something, instead of simulating it.

Service Network: Constellations of independent organizations that work together in various configurations in order to deliver services.

Category: Artificial Intelligence

Simulation Model of Ant Colony Optimization for the FJSSP

Li-Ning Xing
National University of Defense Technology, China

Ying-Wu Chen
National University of Defense Technology, China

Ke-Wei Yang
National University of Defense Technology, China

INTRODUCTION

The job shop scheduling problem (JSSP) is generally defined as decision-making problems with the aim of optimizing one or more scheduling criteria. Many different approaches, such as simulated annealing (Wu et al., 2005), tabu search (Pezzella & Merelli, 2000), genetic algorithm (Watanabe, Ida, & Gen, 2005), ant colony optimization (Huang & Liao, 2007), neural networks (Wang, Qiao, &Wang, 2001), evolutionary algorithm (Tanev, Uozumi, & Morotome, 2004) and other heuristic approach (Chen & Luh, 2003; Huang & Yin, 2004; Jansen, Mastrolilli, & Solis-Oba, 2005; Tarantilis & Kiranoudis, 2002), have been successfully applied to JSSP.

Flexible job shop scheduling problem (FJSSP) is an extension of the classical JSSP which allows an operation to be processed by any machine from a given set. It is more complex than JSSP because of the addition need to determine the assignment of operations to machines. Bruker and Schlie (1990) were among the first to address this problem. The flexible job shop scheduling problem may be formulated as follows.

1. There is a set of n jobs that plan to process on m machines;
2. The set machine is noted M, $M = \{M_1, M_2, ..., M_m\}$;
3. Each job j consists of a sequence of n_j operations $O_{j1}, O_{j2}, ..., O_{jn_j}$;
4. The execution of each operation i of a job j (noted O_{ji}) requires one machine out of a set of given machines called $M_{ji} \subseteq M$.

The problem is thus to both determine an assignment and a sequence of the operations on all machines that minimize following criteria.

1. Maximal completion time of machines;
2. Total workload of the machines;
3. Critical machine workload.

The weighted sum of the above three objective values is taken as the objective function.

$$F(c) = 0.5*F_1(c) + 0.2*F_2(c) + 0.3*F_3(c) \quad (1)$$

Where $F(c)$ denotes the total evaluation value of the schedule c; $F_1(c)$ denotes the maximal completion time of machines (makespan) of the schedule c; $F_2(c)$ denotes the total workload of the machines of the schedule c; $F_3(c)$denotes the critical machine workload of the schedule c.

BACKGROUND

For solving the realistic case with more than two jobs, two types of approaches have been used: hierarchical approaches and integrated approaches (Xia & Wu, 2005).

In hierarchical approaches, assignment of operations to machines and the sequencing of operations on the machines are treated separately. Kacem, Hammadi, and Borne (2002a; 2002b) proposed a genetic algorithm controlled by the assigned model for the FJSSP. Xia and Wu (2005) present an effective hybrid optimization approach, which makes use of particle swarm optimization to assign operations on machines and simulated annealing algorithm to schedule operations, for the multi-objective FJSSP.

Integrated approaches were used by considering assignment and scheduling at the same time. The integrated approach which had been presented by Dauzere-Peres and Paulli (1997) was defined a neighborhood structure for the FJSSP where there is no distinction between reassigning and resequencing an operation, and the tabu search procedure is proposed based on the neighborhood structure. Mastrolilli and Gambardella (2002) improved Dauzere-Peres' tabu search techniques and presented two neighborhood functions. Most researchers were interested in applying tabu search techniques and genetic algorithms to FJSSP in the past (Xia & Wu, 2005).

SIMULATION MODEL OF ANT COLONY OPTIMIZATION

Framework of the Simulation Model

The framework of our proposed simulation model is displayed as Figure 1.

Input Subsystem

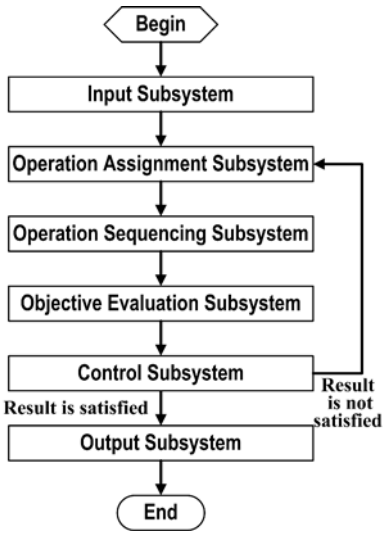
The leading function of input subsystem is inputting all necessary data for solving FJSSP. In our work, we apply file mode to implement the data inputting. Please note, it should have a verify function after data reading, for example, each input data should be a positive integer etc.

Operation Assignment Subsystem

The primary mission of operation assignment subsystem is achieving an excellent assignment of operations to machines. In this part, each assignment was evaluated by formula (2).

$$Fitness(\alpha) = 0.2*F_2(\alpha) + 0.8*F_3(\alpha) \quad (2)$$

Figure 1. The framework of our proposed simulation model



Also, the operation assignment machine knowledge (OAMK) is defined for operation assignment. OAMK is the accumulative knowledge of assigning the giving operation to a more appropriate machine. It was achieved from the near-optimal solution of FJSSP of each iterative. A knowledge matrix *OAMK* with size |*Oper*| × |*Mach*| is defined for the OAMK, where |*Mach*| denotes the number of machines, and |*Oper*| denotes the total account of all operations. For an arbitrary element *OAMK* [*i*][*j*], it means the probability of assigning the giving operation *i* to the current machine *j*.

For enhancing the assignment performance, we try to assign each operation to the machines which process the giving operation with a minimal processing time or the second minimal processing time. The implementing flow of operation assignment is listed as follows.

Step 1. Select an operation (i.e., O_{ij}) among all operations which need to be assigned.
Step 2. Search all machines (e.g., M_{ij}) which process the giving operation with a minimal processing time or the second minimal processing time.
Step 3. Choice a machine (i.e., M_{ijk}, $M_{ijk} \in M_{ij}$) among the achieved machine set M_{ij} randomly with a probability distribution, which was indicated by the OAMK, and assign operationO_{ij} to the selected machine M_{ijk}.
Step 4. Repeat *Step 1* to *Step 3* until all operations were assigned to the appropriate machines.

When obtaining the global optimal solution (the most excellent solution from the beginning of the trial), the OAMK will be updated by applying the rule of (3) according to the global optimal solution.

$$OAMK(i,m) = OAMK(i,m) + Q_G \quad (3)$$

Where *m* denotes each giving machine, *i* denotes each operation processed in machine *m*, Q_G denotes the incremental level in the knowledge updating phase.

Operation Sequencing Subsystem

In order to enhance the sequencing performance, we try to arrange each operation to the giving machine using ant colony optimization (ACO) algorithm. The computational flow of ACO algorithm is displayed as Figure 2.

1. Schedule knowledge initialization. In this part, operation assignment position knowledge (OAPK) is defined according to the traditional pheromone definition. OAPK is the accumulative knowledge of the more appropriate operation processing sequence at a giving machine. It is achieved from the near-optimal solution of FJSSP of each iterative.

In this work, a knowledge matrix *OAPK* with size |*Mach*| × |*Oper*| × |*Oper*| is defined for the OAPK, where |*Mach*| denotes the number of machines, and |*Oper*| denotes the total account of all operations. For an arbitrary element *OAPK*[*i*][*j*][*k*], it means the probability of processing sequence (from operation *j* to operation *k*) at machine *i*.

2. Schedule construction. In this phase, the ACO algorithm schedules operations on each machine and obtains the optimal sequencing of operations on the machines.
   a. Schedule construction mechanism. Based on the simulation advance mechanism, a new schedule construction mechanism was proposed and summarized as follows. If there are several idle machines, then the next operation will be selected according to the state transition rule (see next part) for these machines; if all the machines are busy, then the simulation process will be advanced until one or several machines are idle. The aforementioned process was executed till all the operations were finished.
   b. State transition rule. In the schedule construction phase, each artificial ant will choose the next operation from an allowed set for these idle machines. Let *allow*(*m*,*t*)denotes the allowed set (set of the allowed processing operation) for the machine *m* at the time *t*, and then it can be defined as follows (figure 3).

To the machine *m*, the following probability distribution (Formula 4) will be applied to select the next operation O_j when it finished the previous operation O_i. (See Box 1)

3. Local search algorithm. In ACO, the generated schedules by artificial ants may be so coarse that they should be improved by some complementary local search method. In this chapter, a proposed local search algorithm was applied to refine the feasible solution achieved by ACO. That is, the ACO algorithm, which has excellent exploration and information learning abilities, was applied to provide some appropriate initial schedules, and then these appropriate feasible solutions were refined by the proposed local search

Figure 2. The framework of the ant colony optimization algorithm

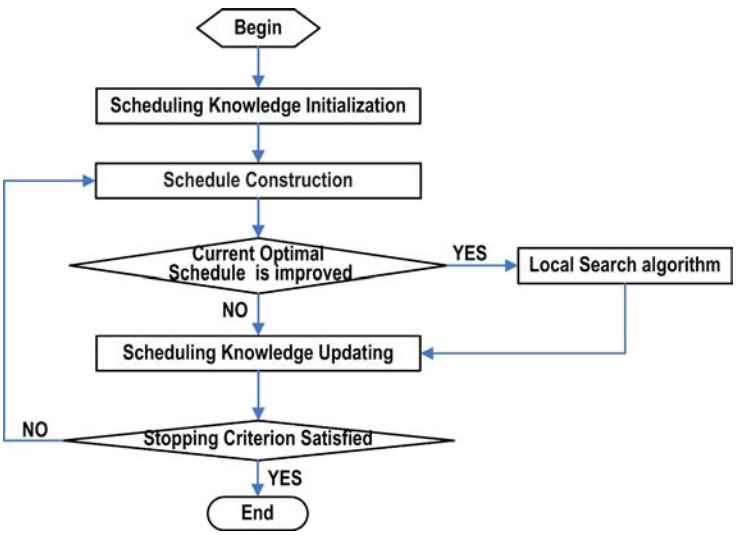


Figure 3. The definition for allowed processing operation set

```
for i = 1: OperNum    % OperNum denotes the account of the assigned operation to machine m;
    Oper = the i^th operation among the assigned operation to machine m;
    if (Oper is the first operation in its job)
        Oper ∈ allow(m,t);
    elseif (the predecessor of Oper was finished at the time t)
        Oper ∈ allow(m,t);
    end
end
```

algorithm. The essential points of this proposed local search algorithm is listed as follows.

a. Try to adjust each operation to the machine with the minimal processing time;
b. Try to adjust each operation to the machine with the second minimal processing time;
c Try to adjust each operation to the machine with the minimal total processing time.

4. Scheduling knowledge updating.
   a. **Local update rule:** The best ant, which constructed the best schedule at the current iterative, is allowed to deposit knowledge to its corresponding solution according to the local updating rule.

$$OAPK(m, i, j) = OAPK(m, i, j) + Q_L \quad (5)$$

Where, m denotes each giving machines; i and j denote two different operations processed at machine m respectively, and operation i was processed before operation j; Q_L denotes the incremental level in the local updating phase.

   b. **Global update rule:** The globally best ant, which constructed the most excellent solution from the beginning of the trial, is allowed to deposit knowledge to its corresponding solution according to the global updating rule.

$$OAPK(m, i, j) = OAPK(m, i, j) + Q_G \quad (6)$$

Where, Q_G denotes the incremental level in the global updating phase.

   c. **Knowledge evaporating rule:** The knowledge evaporating rule is performed after each iterative. In our applied ACO algorithm, knowledge trails on each solution component is limited to an interval $[\tau_{min}, \tau_{max}]$ to avoid stagnation state of possibility. (See Box 2)

5. In our work, the stopping criterion of ACO is defined through the iterative times. That is, the ACO algorithm is terminated when the maximal preset search iterative is exhausted.

Objective Evaluation Subsystem

The main function of objective evaluation subsystem is evaluating the objective value of these achieved schedules. In this part, the makespan, the total workload of the machines and the critical machine workload of the giving schedule were computed firstly, and the weighted sum of these above three objective values was computed as the objective function of the giving schedule according to the formula (1).

Control Subsystem

The elementary function of control subsystem is computation flow control, feasibility validation and other extensible function (module). Computation flow control includes whole flow design, termination condition setting, and so on. Feasibility validation is validating the feasibility of

Box 1.

$$\Pr(O_i, O_j, m, t) = \begin{cases} \dfrac{\left[\tau(m,i,j)\right]^a \times \left[\lambda(m,j)\right]^b}{\sum_{O_k \in allow(m,t)} \left(\left[\tau(m,i,k)\right]^a \times \left[\lambda(m,k)\right]^b\right)} & O_j \in allow(m,t) \\ 0 & O_j \notin allow(m,t) \end{cases} \quad (4)$$

Where, $\Pr(O_i, O_j, m, t)$ denotes the probability of selecting the next operation O_j when it finished the previous operation O_i on machine m at time t; $\tau(m, i, j)$ denotes the heuristic value achieved by OAPK, $\tau(m, i, j) = OAPK(m, i, j)$; $\lambda(m, j)$ denotes the heuristic value achieved by operation processing time, $\lambda(m, j) = 1/t(j, m)$; $t(j, m)$ denotes the time of processing operation j at machine m; a,b denote the weight of these two different heuristic information respectively. In our work, the pseudo-random-proportional rule (Dorigo & Stutzle, 2004) was used as state transition rule.

Box 2.

$$OAPK(m, i, j) = \max\{\tau_{min}, \min\{\tau_{max}, (1 - \rho)OAPK(m, i, j)\}\} \quad (7)$$

Where ρ is the knowledge evaporating parameter ($0 < \rho < 1$).

these interim solutions. If there has some errors, the control subsystem should respond these errors to the user.

Output Subsystem

The leading function of output subsystem is outputting some essential optimization result for the user. Data output modes and figure output modes were applied to the output subsystem.

SIMULATION EXPERIMENTS

To illustrate the effectiveness and performance of our proposed algorithm in this chapter, one representative instances based on practical data have been selected to compute. This instance has 15 jobs with 56 operations that plan to be processed on 10 machines with total flexibility. Larger scale instance is chosen to display the optimizing ability of our proposed approach. We have applied the proposed simulation system to them with following parameters (Table 1). The proposed approach was implemented on a Pentium IV 2.4 GHz personal computer with a single processor and 512M RAM. In order to eliminate the random fluctuation of the optimization process, each instance was run 20 times.

We present the comparison with the algorithms of Kacem et al. (2002a; 2002b) and Xia & Wu (2005) in Table 2. The column labeled 'AL+CGA' is one algorithm by Kacem et al. (2002a). The column labeled 'Hybridization of EA and FL' refers to Kacem et al. (2002b). 'PSO+SA' is proposed by Xia and Wu (2005). The experimental results of Table 2 suggest that our approach makes us reduce the makespan

Table 1. The parameter setting of this proposed algorithm

	Phase Name	Symbol	Value Setting	Material Signification
1	**Schedule Knowledge Initialization**	τ_0	5	The initialization of knowledge matrix $OAPK$;
2		η_0	$1/t(i,m)$	The initialization of knowledge matrix $OAMK$;
3		$t(i,m)$	---	the time of processing operation i in the machine m ;
4	**Operation Assignment**	Max_Iter1	10	The maximal iterative times in the implementation process of operation assignment subsystem;
5		Max_Iter2	5	The maximal iterative times in the local search phase;
6	**Operation Sequencing**	$AntSize$	50	The account of ant;
7		a	2	The weight of heuristic information of knowledge $OAPK$;
8		b	5	The weight of heuristic information of processing time;
9		Max_Iter3	3	The maximal iterative times in the local search phase;
10	**Scheduling Knowledge Updating**	τ_{max}	50	The maximal value of knowledge matrix $OAPK$ ($OAMK$);
11		τ_{min}	0.01	The minimal value of knowledge matrix $OAPK$;
12		ρ	0.02	The knowledge evaporating parameter;
13		Q_L	0.2	The incremental level in the local update knowledge phase;
14		Q_G	5	The incremental level in the global update knowledge phase;
15	**Stopping Criterion**	Max_Gen1	5	The maximal generation of ant colony optimization;
16		Max_Gen2	10-50 ①	The maximal generation of whole simulation system;

① It is preset as 10 when solving instance 1; it is preset as 20 when solving instance 2-4; it is preset as 50 when solving instance 5.

Table 2. Comparison of results on problem 15×10 *with 56 operations*

	AL+CGA		Hybridization of EA and FL			PSO+SA	Our Proposed Approach			
$F_1(c)$	23	24	23	23	24	12	12	11	11	11
$F_2(c)$	95	91	91	95	91	91	92	92	93	91
$F_3(c)$	11	11	10	11	11	11	11	11	10	11
$F(c)$	33.8	33.5	32.7	33.8	33.5	27.5	27.7	27.2	27.1	27.0

(12 instead of 11) and get the decline in terms of objective function value.

FUTURE TRENDS

About future research direction, we should pay more attention to the following two points. The first one is optimizing these parameters of our proposed approach for its quick constringency. The second one is reducing the sensitivity to the initial solution of our proposed approach.

CONCLUSION

The contribution of this chapter is summarized as follows. It presented a simulation model to work out flexible job shop scheduling problems with the multi-objective of minimizing makespan, the total workload of machines and the workload of the critical machine. The results obtained from the computational study have shown that the proposed approach is a feasible and effective approach for the multi-objective flexible job shop scheduling problem.

ACKNOWLEDGMENT

This research was supported in part by the National Natural Science Foundation of China (No.70272002).

REFERENCES

Bruker, P., & Schlie, R. (1990). Job-shop scheduling with multi-purpose machines. *Computing*, *45*, 369-375.

Chen, H., & Luh, P.B. (2003). An alternative framework to Lagrangian relaxation approach for job shop scheduling. *European Journal of Operational Research*, *149*, 499-512.

Dauzere-Peres, S., & Paulli, J. (1997). An integrated approach for modeling and solving the general multiprocessor job-shop scheduling problem using tabu search. *Annals of Operations Research*, *70*, 281-306.

Dorigo, M., & Stutzle, T. (2004). *Ant colony optimization*. Cambridge, MA: MIT Press.

Huang, K.L., & Liao, C.J. (2007). Ant colony optimization combined with tabu search for the job shop scheduling problem. *Computers & Operations Research*, article in Press.

Huang, W.Q., & Yin A.H. (2004). An improved shifting bottleneck procedure for the job shop scheduling problem. *Computers & Operations Research*, *31*, 2093-2110.

Jansen, K., Mastrolilli, M., & Solis-Oba, R. (2005). Approximation schemes for job shop scheduling problems with controllable processing times. *European Journal of Operational Research*, *167*, 297-319.

Kacem, I., Hammadi, S., & Borne, P. (2002a). Approach by localization and multi-objective evolutionary optimization for flexible job-shop scheduling problems. *IEEE Transactions on Systems, Man, and Cybernetics*, *Part C*, *32*(1), 1-13.

Kacem, I., Hammadi, S., & Borne, P. (2002b). Pareto-optimality approach for flexible job-shop scheduling problems: Hybridization of evolutionary algorithms and fuzzy logic. *Mathematics and Computers in Simulation*, *60*, 245-276.

Mastrolilli, M., & Gambardella, L.M. (2002). Effective neighborhood functions for the flexible job shop problem. *Journal of Scheduling*, *3*(1), 3-20.

Pezzella, F., & Merelli, E. (2000). A tabu search method guided by shifting bottleneck for the job shop scheduling problem. *European Journal of Operational Research*, *120*, 297-310.

Tanev, I.T., Uozumi, T., & Morotome, Y. (2004). Hybrid evolutionary algorithm-based real-world flexible job shop scheduling problem: application service provider approach. *Applied Soft Computing*, *5*, 87-100.

Tarantilis, C.D., & Kiranoudis, C.T. (2002). A list-based threshold accepting method for job shop scheduling problems. *International Journal of Production Economics*, *77*, 159-171.

Wang, S.F., & Zou, Y.R. (2003). Techniques for the job shop scheduling problem: A survey. *System Engineering Theory and Practice*, *23*(1), 49-55.

Wang, X.H., Qiao, Q.L., & Wang, Z.O. (2001). A method to solve job-shop schedule problems by neural network with transient chaos. *System Engineering*, *19*(3), 43-48.

Watanabe, M., Ida, K., & Gen, M. (2005). A genetic algorithm with modified crossover operator and search area adaptation for the job-shop scheduling problem. *Computers & Industrial Engineering*, *48*, 743-752.

Wu, D.W., Lu, T.D., Liu, X.B., & Meng Y.S. (2005). Parallel simulated annealing algorithm for solving job-shop scheduling problem. *Computer Integrated Manufacturing Systems*, *11*(6), 847-850.

Xia, W.J., & Wu, Z.M. (2005). An effective hybrid optimization approach for multi-objective flexible job-shop scheduling problems. *Computers & Industrial Engineering*, *48*, 409-425.

KEY TERMS

Ant Colony Optimization: ACO studies artificial systems that take inspiration from the behavior of real ant colonies and which are used to solve discrete optimization problems. ACO is a population-based approach to the solution of combinatorial optimization problems. The basic ACO idea is that a large number of simple artificial agents are able to build good solutions to hard combinatorial optimization problems via low-level based communications.

Combinatorial Optimization Problems: One can argue that combinatorial optimization and Integer programming are synonymous terms. This is because the majority (if not all) of the combinatorial optimization problems are integer programming problems, usually involving binary variables.

Job Shop Scheduling Problem: An instance of the job-shop scheduling problem consists of a set of *n* jobs and *m* machines. Each job consists of a sequence of *n* activities so there are *nm* activities in total. Each activity has a duration and requires a single machine for its entire duration. The activities within a single job all require a different machine. An activity must be scheduled before every activity following it in its job. Two activities cannot be scheduled at the same time if they both require the same machine. The objective is to find a schedule that minimizes the overall completion time of all the activities.

Flexible Job Shop Scheduling Problem: It is an extension of the classical job shop scheduling problem which allows an operation to be processed by any machine from a given set. The problem is to assign each operation to a machine and to order the operations on the machines, such that the maximal completion time (makespan) of all operations is minimized.

Multi-Objective Optimization: In the world around us it is rare for any problem to concern only a single value or objective. Generally, multiple objectives or parameters have to be met or optimized before any 'master' or 'holistic' solution is considered adequate. Most realistic optimization problems, particularly those in design, require the simultaneous optimization of more than one objective function.

Operation Assignment Machine Knowledge (OAMK): It is the accumulative knowledge of assigning the giving operation to a more appropriate machine. It was achieved from the near-optimal solution of FJSSP of each iterative.

Operation Assignment Position Knowledge (OAPK): It is the accumulative knowledge of the more appropriate operation processing sequence at a giving machine. It is achieved from the near-optimal solution of FJSSP of each iterative.

Simulation, Games, and Virtual Environments in IT Education

Norman Pendegraft
University of Idaho, USA

INTRODUCTION

The rapid change in information technology presents several problems to IS educators and trainers. In particular, the number of concepts that must be mastered is constantly increasing while the time available is not. This makes it essential to use class time efficiently as well as effectively. Simulations and games provide interesting and useful tools to help in this effort.

BACKGROUND

The idea that students learn better by doing goes back at least to Dewey (1938). The key idea underscoring this approach is that people learn better from experience than from reading or listening (Corbeil, Laveault, and Saint-Germain, 1989). This sort of experience can be gained in a simulation or game. By compressing time, the simulation allows the students to experience the consequences of their own actions or to see how a system operates.

Simulation, case studies, role playing, and gaming are related teaching methods based on experiential learning. They permit experience or experimentation with a situation modeling the real world (Senge, 1990). On a deeper level, simulation is claimed by some to be a fundamentally new way of studying the world (Pagels, 1998). Narayanasamy, Wong, Fung, and Rai (2006) distinguish between games, simulation games, and training simulators. Simulators are models used for systems analysis or policy formation. Simulators may use mathematical models and Monte Carlo or discrete event methodologies. They argue that training simulators offer real-world environments and challenges are focused on skill development rather than entertainment, and are not goal oriented. Klassen and Willoughby (2003) discuss the importance of assessment and present data supporting the notion that games help students learn more quickly than do lectures.

Case studies are a time honored approach of instruction in strategy courses (see, for example, Burgelman, Maidique, and Wheelwright, 2001). Barker (2002) suggests that they can also be very valuable for teaching technical skills such as software development. In some sense, a case study is a role play with the student acting the part of an analyst examining the case situation.

Role playing and simulation gaming are similar approaches in that they use simulated worlds, but instead of creating, or observing or analyzing that world, students are immersed in it. Role playing is a method in which students are presented a scenario simulating some real situation, and assigned roles in that scenario. The scenario can be based on real or simulated situations (Barker, 2003). Participants then assume the roles of relevant persons in the scenario and act out the situation to see what happens. Role playing is a commonly and successfully used tool in IS education (for example, Christozov, 2003).

According to Greenblat (1988), simulation gaming includes role playing as an element. Whereas role playing allows participants to play the roles as they please, simulation gaming emphasizes the interactions of the roles and constraints of various types on the players. In some sense, a simulation game strives to teach about a specific situation while a role play or game may have a more general lesson.

Greitzer, Kuchar and Huston. (2007) describe cognitive principles for learning and a process for designing and improving game based education. In particular, they argue that experience should be presented in realistic contexts. They identify features of games that attract extended play including levels, adaptability, clear goals, interactions, and shared experience with others. They describe the use of these principles in a security training game.

The use of modern information technology in developing these games and simulations, stimulated, no doubt, by the vibrant computer game industry, has led several authors to create "virtual" environments for training. This seems a good umbrella term to include all simulation and gaming approaches to training, as they do indeed create virtual realities in which the students operate. Summerfield (2004) suggests that role playing is superior for learning soft skills (like dealing with people) while technically based simulations are useful for learning hard skills.

Simulation and gaming have been used to enhance training in a variety of non IS areas including incident management (Jain and McLean, 2003), mass casualty medicine (Müller, Martens, Willen, and Müller, 2000), military technology

(Meeds, 2001), military tactics (Chatham, 2007), and immunology (Kelly, Howell, Glinert, Holding, Swain, Burrowbridge, and Roper, 2007). Mayo (2007) argues that video games can be used to teach science and engineering better than lectures.

After many years of using such exercises at all levels (undergraduate, graduate, and executive), it is the author's opinion that they are very useful and that major benefit accrues to the instructor in preparing the simulation as well as to the students when they play the game. Simulation and gaming are student centered learning, that is, the student is actively involved in the learning rather than passively observing the instructor (Greenblat, 1998). The student does the work, makes decisions and sees the impact of the decisions. Role playing and simulation gaming attempt to take advantage of this by creating a situation in which a student may "play a game" in which time is compressed and attention can be focused on a few key ideas. Finally, these kinds of exercises are fun. The class gets to move around, talk, and frequently laugh. Simulations and games epitomize the idea that learning should be fun.

SIMULATION AND GAMING IN IT

Simulation as a teaching tool suggests several approaches. Perhaps most obvious in an information systems curriculum is computer simulation. Using this technique, a computer program is written which exhibits behavior that models the behavior of the system under study. Butterfield and Pendegraft (1998) described a spread sheet simulation of a Fourier Series, adding sine wave to construct a square wave for a class demonstration of the impact of bandwidth limits on data rates. Campbell (1996) created a simulation of a computer and had his students write assembly language programs to execute on the simulation. Zant (2001) used a computer simulation of a CPU for in class demonstrations and for homework problems. Englander (2003) used Little Man Computer, a simple paper simulation of a CPU, as an example to explain basic CPU architecture, CPU operation, and machine language. In an extension of those ideas, Pendegraft and Stone (2003) had their students develop a Visual Basic simulation of a Little Man central processing unit on which they ran programs mandated by the instructor. In addition to having to execute simple programs written in Little Man's machine language, their simulation had to deal with other architectural issues like input and output.

Several authors have used simulations or games to teach networking. Voderhobli and Pattison (2005) developed a simulation system for network managers. Guo, Xiang and Wang (2007) developed a simulation network laboratory emphasizing the dynamics of network protocols rather than configuration. Leitner and Cane (2005) designed a virtual networking laboratory intended to address the need for a hands on experience in a distance education environment. Dennis (2002) and Pendegraft (2002, 2003) developed in class games to teach TCP/IP. These two games will be described in more detail in the next section. Leitner and Cane (2005) describe a virtual laboratory for distance education in IS. They argue that simulations do not provide direct experience. They distinguish virtual laboratories from simulations in that the former "is a true laboratory in which experiments are carried out under the control of remote users" (p.284).

There are many interesting efforts to improve the teaching of programming through the use of games. Several researchers have looked at the use of Robocode which uses a game to teach Java. Long (2007) evaluated Robocode as a vehicle for instruction and found that participants had fun and enjoyed the exercise. They mentioned that the exercise was intrinsically interesting because they learned something, but also that they had fun doing so. Bierre, Ventura, Phelps, and Egert (2006) had their students create simulated battle tanks. The assignment ended with a tournament in which the tanks fought against each other. They report that teams who used this environment learned more than those in traditional or Robocode environments. Alice is a visual environment for teaching programming by allowing them to easily create games or simulations (Alice.org). Students learn to think in terms of objects and their behaviors. Kelleher and Pausch (2007) used Alice to inspire middle school girls' interest in learning to program computers. Baker, Navarro, and van der Hoek (2003) describe a card game that they developed to teach software engineering. This approach is not limited on the software side to programming. Lawrence (2004) used a game to teach data structures.

Others have developed games to help students learn about IS management. For example, Jain and Boehm (2006) developed SimVBSE, a game to improve understanding of value-based software engineering. Students act as project managers to learn the fundamentals of software engineering. Sheng, Magnien, Kumaraguru, Acquisti, Cranor, Hong, and Nunge (2007) and Irvine, Thompson and Allen (2005) describe games used to teach about IS security. Curtin, Carpenter and Ritzo (2006) report on games developed to train help desk staff. They also offer a process for creating effective games.

EXAMPLE: USING A GAME TO TEACH TCP / IP

An example may help to clarify the mechanics of gaming and illustrate its utility. Consider two similar games, one designed by Dennis (2002) and one by Pendegraft (2002, 2003), to help teach how TCP/IP works. Both are published elsewhere and so will not be described in detail here. Both

games are designed to be run in one class session in a course on telecommunications and could be adapted to class sizes ranging from a dozen to more than 40. The games have similar structure. The class is divided into teams, each team representing a host. Each player represents one layer on that host.

In the play of the simulation, an application layer player writes a message to another application layer player on a paper form. The form is then handed to the TCP player of the sending host. The TCP player adds the TCP header and hands the packet to the IP player. The IP player adds the IP header and hands the packet to the DLL player. The message is then passed to another host where each player strips off the header for that layer and hands the message upward, or in the case of IP forwards it as necessary.

The games offer different points of view. Dennis' game allows many messages on the network at one time, while Pendegraft's game only one message is sent at a time and the entire class follows it along the way, discussing immediately problems that may occur such as a player incorrectly addressing a packet. In Dennis' game that sort of error is handled in discussion between the affected players.

Both games simplify TCP/IP ignoring some issues like handshaking or error detection. This is not to say that these issues are unimportant, but that these games focus attention on a limited set of issues of paramount importance.

FUTURE TRENDS

There is a trend in IT toward increased system complexity paralleled with efforts to hide that complexity from the user. Examples include third generation cell phone services, distributed databases, and web based applications in general. These systems raise questions about how to prepare students to deal with a world of increasing complexity. It may be that simulations and games offer a way to help students more quickly understand complex systems.

Holt (2000) identified several trends in modern training, especially as it related to IT. Among them are emphasis on individualized, self paced training that is realistic. Virtual environments seem ideal ways to meet these needs. The recent explosion of interest in computer based games and simulation for training has drawn from the entertainment industry for inspiration and techniques. As Malykhina (2005) and Narayansamy, Wong, Fung, and Rai (2006) note, simulation games and simulators are converging. As IS becomes more closely connected with other disciplines like the arts and entertainment, education needs will evolve. Hence, there is every reason to believe that this trend will continue and accelerate. It is interesting that most of the effort noted here has been directed at training as opposed to education.

CONCLUSION

Gaming and simulation have proven to be excellent ways to approach training in a variety of fields. In IS they have been used to teach about computer and network architecture, programming, and IS management. Increasingly these games are conducted as computer games making use of the very technology being studied. This use of simulation is consistent with the more conventional use of simulation as a vehicle for evaluating policies or designs prior to implementation. Those more traditional uses are in some sense about educating managers about the implications of their choices. More direct use in training seems a natural extension.

Understanding IT basics will remain an essential part of the education of IT professionals. As technology evolves it will continue to be a challenge to help students understand the basics and prepare themselves to keep learning. Simulation and gaming offer effective and fun tools to help students learn about new technology. They do so in a way that reinforces the need to learn how to learn and to continue learning.

There is an "ancient debate" over the difference between training and education (Fein, 1959). Education is understood to be theoretical or basic while training is practical and has immediate application. Both approaches address legitimate needs, and there is always tension between them. Most of the examples noted here seem to lean to training. In "Profession," one of the all time classic science fiction short stories, Asimov (1957) imagined a world in which young people were trained by being directly connected to a computer. In that world only a few were selected for education using books. One wonders if we a seeing a first step in that direction.

REFERENCES

Alice.org. www.alice.org. (2007, December 15).

Asimov, Isaac (1957). "Profession", *Astounding Science Fiction*, July, Street and Smith. http://www.abelard.org/asimov.htm (2007, December 15).

Baker, A., Navarro, E. O., and van der Hoek, A. (2003). Problems and Programmers: an educational software engineering card game. In *Proceedings of the 25th international Conference on Software Engineering* (Portland, Oregon, May 03 - 10, 2003). International Conference on Software Engineering. IEEE Computer Society, Washington, DC, 614-619.

Barker, S. (2002). Training Business Students to be End-Used [sic] Developers: Are Case Studies the Best Option?, *Proceedings of the IRMA International Conference*, Seattle.

Barker, S. (2003). Business Students as End-Use Developers: Simulating "Real-Life" Situation through Case Study

Approach, *Current Issues in IT Management*, 305-312, McGill, Tanya, Ed., IRM Press, Hershey PA.

Bierre, K., Ventura, P., Phelps, A., and Egert, C. (2006). Motivating OOP by blowing things up: an exercise in cooperation and competition in an introductory java-programming course. *Proceedings of the 37th SIGCSE technical symposium on computer science education*, 06, 38(1).

Burgelman, R.A., Maidique, M.A., and Wheelwright, S.C. (2001). *Strategic Management of Technology and Innovation*, 3rd ed. McGraw Hill, Boston.

Butterfield, J. and Pendegraft, N. (1998). Fourier Analysis: Creating a Virtual Laboratory Using Computer Simulation, *Informing Science*, 1#3.

Campbell, Robert A. (1996). Introducing Computer Concepts by Simulating a Simple Computer, *SIGCSE Bulletin*, Vol. 28#3.

Chatham, R. E. (2007). Games for training. *COMMUNICATIONS OF THE ACM* 50(7) 36-43. DOI= http://doi.acm.org/10.1145/1272516.1272537.

Christozov, D. (2003). Real Live cases in Training Management of Information Resources During the Transition to Market Economy, *Current Issues in IT Management*, 297-303, McGill, Tanya, Ed., IRM Press, Hershey PA.

Corbeil, P., Laveault, D.. and Saint-Germain, M. (1989). *Games and Simulation Activities: Tools for International Development Education*, Canadian International Development Agency, Quebec.

Curtin, M., Carpenter, N., and Ritzo, C. (2006). Adding fun and games to training programs. In *Proceedings of the 34th Annual ACM SIGUCCS Conference on User Services* (Edmonton, Alberta, Canada, November). SIGUCCS '06. ACM, New York, NY, 50-54. DOI= http://doi.acm.org/10.1145/1181216.1181228

Dennis, Alan (2002). *Networking in the Internet Age*, John Wiley and Sons, New York.

Dewey, J. (1938). *Experience in Education*. Collier, New York.

Englander, I. (2003). The *Architecture of Computer Hardware and Systems Software*, 3rd Ed., Wiley, New York.

Fein, L. 1959. The role of the University in computers, data processing, and related fields. *COMMUNICATIONS OF THE ACM* 2(9) 7-14. DOI=http://doi.acm.org/10.1145/368424.368427

Greenblat, C.S. (1988). Designing Games and Simulations: an Illustrated Handbook, Sage, Newbury Park, CA.

Greitzer, F., Kuchar, O., and Huston, K. (2007). Cognitive science implications for enhancing training effectiveness in a serious gaming context. *J. Educ. Resour. Comput*. 7, 3 (Nov. 2007), 2. DOI= http://doi.acm.org/10.1145/1281320.1281322

Guo, J., Xiang, W., Wang, S. (2007). Reinforce Networking Theory with OPNET Simulation, *Journal of Information Technology Education* 6, 215-226.

Holt, B. (2000). Technology Training: Trends for the 21st Century. *Linux Journal* 2000 71, 3.

Irvine, C., Thompson, M., Allen, K. (2005). CyberCIEGE: Gaming for Information Assurance, *IEEE Security and Privacy* 3(3) 61-64.

Jain, A, and Boehm, B. (2006). SimVBSE: Developing a Game for Value-Based Software Engineering, *19th Conference on Software Engineering Education & Training (CSEET'06)*, 103-114.

Jain, S. and McLean, C. (2003). Integrated Simulation and Gamin Architecture for Incident Management Training. *Proc of the Winter Simulation Conf*.

Kelleher, C. and Pausch, R. (2007). Using storytelling to motivate programming. *COMMUNICATIONS OF THE ACM* 50(7), 58-64. http://doi.acm.org/10.1145/1272516.1272540

Kelly, H., Howell, K., Glinert, E., Holding, L., Swain, C., Burrowbridge, A., and Roper, M. (2007). How to build serious games. *COMMUNICATIONS OF THE ACM 50 (*7) 44-49. DOI= http://doi.acm.org/10.1145/1272516.1272538

Klassen, K. and Willoughby, K. (2003). In Class simulation games: Assessing student Learning. *Journal of Information Technology Education* 2, 1-13.

Lawrence, R. (2004). "Teaching data structures using competitive games", *IEEE Transactions on Education*, 47(4) (459- 466).

Leitner, L., and Cane, J. (2005). A Virtual Laboratory Environment for Online IT Education, *ACM SIGITE'03.*

Long, J., (2007). Just for fun: Using Programming Games in Software Programming training and education- A field study of IBM Robocode community. *Journal of Information Technology Education 6*.

Malykhina, E. (2005). New School of Thought, *Information Week*, 14 Feb. http://www.informationweek.com/story/showarticle.jhtml?articleID=60400089 (December 12, 2007).

Mayo, M. J. (2007). Games for science and engineering education. *COMMUNICATIONS OF THE ACM* 50 (7) 30-35. DOI= http://doi.acm.org/10.1145/1272516.1272536

Meeds, H., (2001). University of Information Technology simulations: "learning by doing", http://www.gordon.army.mil/AC/winter/winter%2001/meeds.htm, retrieved Nov. 2007.

Müller, N., Martens, P., Willen, P., and Müller, H. (2000). Panic, a Computer Game for Training of Candidate Physician Confronted with Mass Casualty Incidents. *Proceedings of the IFIP Tc5/Wg5.7 Fourth international Workshop of the Special interest Group on integrated Production Management Systems and the European Group of University Teachers For industrial Management Ehtb: Games in Operations Management* (November 26 - 29, 1998). J. O. Riis, R. Smeds, and R. V. Landeghem, Eds. IFIP Conference Proceedings, vol. 170. Kluwer B.V., Deventer, The Netherlands, 125-136.

Narayanasamy, V., Wong, K. W., Fung, C. C., and Rai, S. (2006). Distinguishing games and simulation games from simulators. *Comput. Entertain*. 4 (2) 9. http://doi.acm.org/10.1145/1129006.1129021

Pagels, H. (1998). *Dreams of Reason*, Simon and Schuster, New York.

Pendegraft, N. (2002). The Internet Game,. *Proceedings of the International Conference of the Information Resources Management Association*, Seattle.

Pendegraft, N. (2003). The TCP / IP Game, *Current Issues in IT Management*, 117-124, McGill, Tanya, Ed., IRM Press, Hershey, PA.

Pendegraft, N. and Stone, R. (2003). "Using A Simulation Assignment to Teach CPU Operations", *International Conference on Informatics & Research*, Dec 12-14.

Senge, P. (1990). *The Fifth Discipline*, Doubleday, New York.

Sheng, S., Bryant Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L., Hong, J., and Nunge, E. (2007). Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. *Symposium On Usable Privacy and Security*, Pittsburg. http://cups.cs.cmu.edu/soups/2007/proceedings/p88_sheng.pdf.

Summerfield, B. (2004). Learning Simulations: Experiential Education, *Chief Learning Officer Magazine*, http://www.clomedia.com, November 15, 2007.

Voderhobli, K. and Pattison, C. (2005). Building Virtual Network Management Scenarios.

Zant, R. (2001). Using Simulation In IS Curriculum, ISECON, Chicago, http://isedj.org/isecon/2002/342d/ISECON.2002.Zant.pdf, 12 Dec 2007.

KEY TERMS

Case Study: An instruction tool containing a detailed description of a real world situation

Computer Simulation: A simulation built using a computer language

Experiential Learning: Learning based on experiences rather than listening or reading.

Game: A simulation in which people are part of the model and their decisions partially determine the outcome.

Role Playing: An element in gaming in which players assume the roles of other people.

Serious Games: Games used for serious training or analysis as opposed to entertainment.

Simulation: A model of a system focusing on selected behaviors or aspects of that system.

Simulator: A simulation model of a system used for systems analysis. May involve mathematical models, Monte Carlo or discrete event methods.

NOTE

Portions of this article were previously published in "Simulation and Gaming in IT Education" in the first edition of this encyclopedia.

Smart Assets Through Digital Capabilities

Jayantha P. Liyanage
University of Stavanger, Norway

Thore Langeland
Norwegian Oil Industry Association (OLF), Norway

INTRODUCTION

The history of oil and gas (O&G) production on the Norwegian Continental Shelf (NCS) dates back to the early 70s and began with the discovery and subsequent development of the great *Ekofisk asset* in 1969. Ever since, North Sea has been an attractive region, and Norway in particular has been a major supplier of O&G to the world energy market. The remaining production prospect on the Shelf is also said to be substantial and the more recent estimates indicate that it is equivalent to twice the already produced amount. However, a major part of the Norwegian O&G production portfolio will approach maturity by 2007/2008. Both *declining production* and *marginal discoveries* have given a clear indication that the Norwegian O&G industry will undergo a series of significant challenges during the next few years. These critical issues, in conjunction with volatile business environment, have brought a major turning point demanding immediate steps to enhance operational efficiency and to reduce operating risk in offshore exploration and production (E&P) activities on NCS.

BACKGROUND

Norwegian Oil Industry Association (OLF) constantly raised its concerns since the mid 90s about this emerging situation owing to its substantial commercial impact. By 2000-2002, major challenges for E&P activities on NCS became more visible and obvious. These included:

- declining investments and activity level on NCS and its immediate impact on the production profile and sustained growth;
- the need to enhance recovery efficiency to keep supply levels and to add more value;
- rising lifting costs and its direct impact on the cost of operatorship; and
- the volatile oil price and its direct implications on profit performance of offshore E&P activities.

Subsequently it appeared that the Norwegian O&G industry requires, a dedicated plan to re-engineer conventional practices in order to reduce commercial risk and to enhance value creation (see discussions by Lindgren & Bandhold, 2003). Thus, the O&G production scenario on the NCS stepped into its so-called *3rd efficiency leap* with the effect from 2003 (see Norwegian Oil Industry Association, 2003). This macro-scale national program is expected to induce a step change particularly related to smart use of advance information and communication technologies (ICT) and new data management techniques, and is completely dedicated to adapt to a fully integrated operational setting to smartly manage offshore-onshore activities by the year 2010. Issues related to this new development scenario were envisioned in the government white paper; *Storting white paper 38:2003–2004: On the petroleum activities* in 2003.

SMART ASSETS THROUGH *INTEGRATED E-OPERATIONS*

In more general terms, *smart assets* and *integrated e-operations* largely dedicate the Norwegian O&G industry toward:

- joint exploitation of advanced technologies, digital ICT capabilities, and real-time operational and technical data to optimize decisions, and
- tighter integration of work processes, decision loops with effective and efficient division of work to optimize activities.

This aims at enhancing connectivity and interactivity between offshore O&G assets and their onshore support systems (for further highlights see Barabasi, 2003). Advancement in information sciences and technologies, and long-term commercial benefits of their successful usage, have contributed much to the ongoing change process. Partner industries, particularly those related to electronic and communication technologies in general, play pivotal roles to establish the necessary stable and reliable *digital environment* around offshore assets and activities (see further discussions by

During, Oakey et al., 2002). The key enabling technologies that are already under implementation include:

- fiber-optic-based ICT-net laid on the sea-bed of the North Sea, and wireless communication capability ;
- smart sensors, intelligent transducers, and equipment with advanced functionalities;
- real-time visualization, 3D visualization, and simulation tools;
- online diagnostic and prognostic engineering capability;
- process automation and real-time data acquisition techniques; and
- online video monitoring and conferencing facilities.

Such a technological leap not only systematically builds strategic *digital capabilities*, but also provides necessary *digital environment* for active *knowledge and intelligent data* management. The onshore support system, with the support of such application technologies, systematically advances toward a highly connective and interactive *extended digital enterprise* through active strategic collaboration breaking the conventional inter-organizational gaps (see also Dyer, 2000; Faulkner & Rond, 2001; Lipnack & Stamps, 1997; Spekman, 2003; Tidd, 2001; Tonchia & Tramontano, 2004). This new organizational setting is capable of:

- joint online monitoring of offshore E&P activities at dispersed *onshore support centres*;
- real-time data acquisition, joint data analysis and data interpretation; and
- 24/7 network-based connectivity for collaborative decision making and work planning.

A diagramatic model of the new operating environment is illustrated in Figure 1.

The network-based and collaborative environment requires a highly robust ICT infrastructure to support decisions, core tasks, and activities (see also Barabasi, 2003; During, Oakey et al., 2002; Hosni & Khalil, 2004). *SOIL* (Secure Oil Information Link), introduced in 1998 to the NCS, is the result of the current industry demand to acquire necessary *digital infrastructure* using a large-scale information and communication network through a common data-hub, a large-scale network, and centralized information system, which is highly *reliable* (stable and dependable), *secure* (control of access and routing), and possesses a large *bandwidth* (high traffic capacity).

Today, SOIL is extended into the UK and functions as a collaboration arena, a secure interconnection point, and an industry network. This digital network between offshore facilities, major producers, and third-party organizations facilitates the connectivity through the use of fiber-optic cables, radio links, and satellite communications. A unique feature of SOIL application today is that it has moved the information sharing and communication capabilities from conventional setting to a *one-to-many* collaborative setting. SOIL allows consolidation of a traditional information and communication network into one single hub building 24/7 online and real-time information sharing and communication capabilities between O&G producers and their business partners. This is one of the most important application

Figure 1. New operational environment establishes an extended enterprise on the basis of advanced technological capability, digital infrastructure, and active collaboration

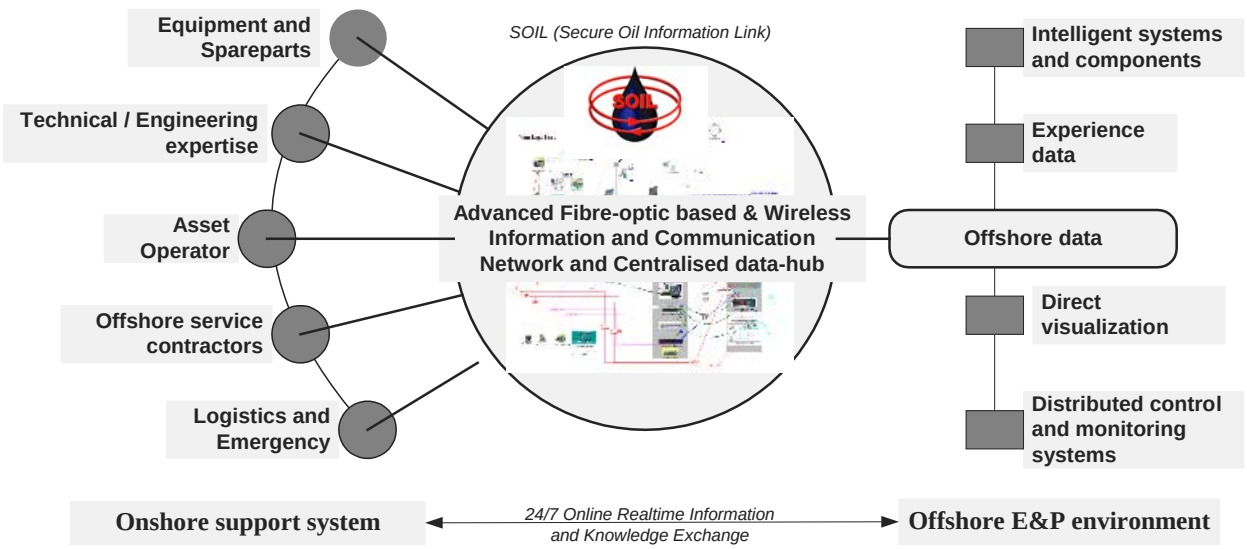
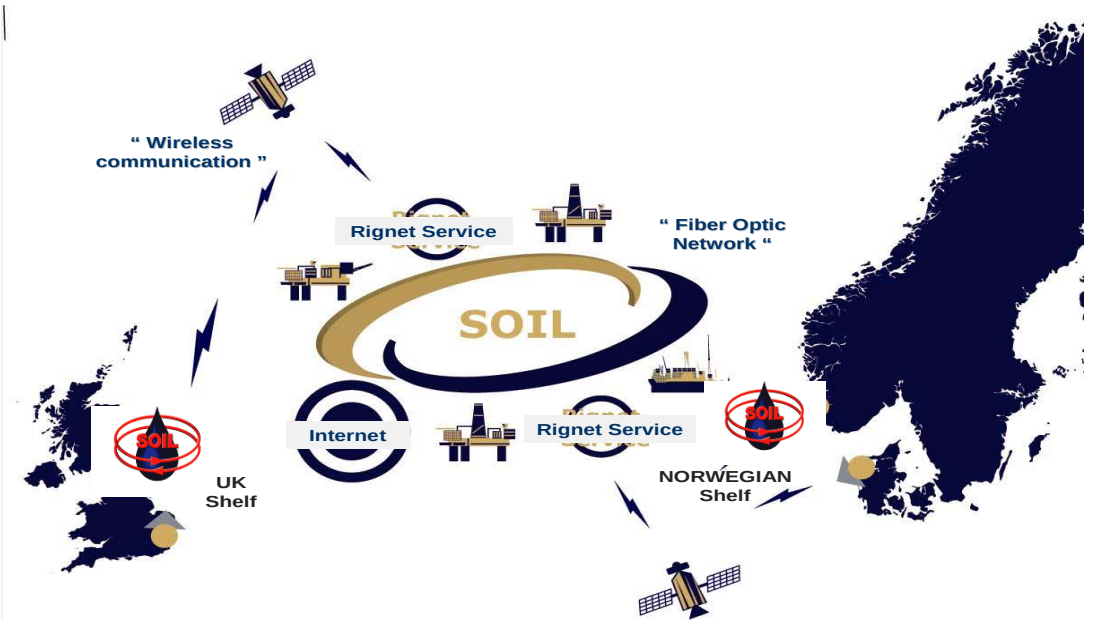


Figure 2. Digital infrastructure through SOIL is established using fibre-optic network and wireless communication capabilities enabling 24/7 online real-time communication between offshore assets and onshore support system



settings for onshore-based remote integrated operations or simply '*e-fields'* (Figure 2).

SOIL provides some key application services such as:

- *SOIL meeting* (that provides a virtual communication platform between member companies);
- *E2E monitoring* (that automatically monitor real-time network performance between operations on offshore rigs and onshore operation centers); and
- *Proex* (that is a Web-based solution to structure, define, execute, and follow-up tasks and activities in projects and/or work processes on a continuing basis).

The membership status of *SOIL* has grown from a total of 20 in 1998 to approximately 170 by 2005 with the recent extension to the UK sector.

The interactive nodes of the complex information and knowledge network are the *onshore online support centers*. Such centers are available at the premises of O&G producers (e.g., ConocoPhillips, BP, Hydro, Statoil) as well as service-support-supply organizations (e.g., BakerHughes INTEQ, RC-DEI). Technological capabilities built into these onshore centers, together with the licensed access via *SOIL*, allow real-time integration between offshore assets and onshore support system. For instance, the *Onshore Drilling Center (ODC)* of *ConocoPhillips of Norway (COPNO)* has the capability to actively integrate 3 functional arenas, namely:

- **Operational:** Where online monitoring of offshore drilling activities and equipment performance are enabled. This is equipped with table-top workstations and back-projected large VDUs.
- **Collaboration:** Where online communication between offshore facility and onshore support system takes place. This is equipped with video-conferencing facilities, CCTV, and other advanced technological capabilities and electronic gadgets (e.g. VisiWear, Smart boards) for joint decision-making.
- **Visualization:** Where complex data from reservoir and production/injection wells are processed to enhance visualization. This is equipped with advance technology to produce 3D images and to run simulations.

In general, *smart assets* and *intergated e-operations* exemplify a radical change in the offshore asset management practice in *North Sea* based on advances in information sciences and technologies.

OLF proclaims that the new step can directly contribute to, make offshore production on NCS more commercially attractive, improve recovery or production regularity, give more positive health, safety, and environmental results, and improve economical performance. The economical speculations, as of today, are in fact 10% increment in oil production from the continental shelf and, at the same time 30% reduction in overall operating costs.

FUTURE TRENDS

A unique feature of ongoing developments is that it induces a major transition in the industry infrastructure bridging the conventional gaps to jointly manage offshore E&P activities. It implies that new developments bring the O&G producers, service contractors, support and supply organizations, and expert services, closer together opening up greater opportunities through strategic partnerships (see further discussions in Dyer, 2000; Faulkner & Rond, 2001; Lipnack & Stamps, 1997; Spekman, 2003; Tonchia & Tramontano, 2004). The integration efforts across the industry will continue targeting full-scale integration by 2010. This will have to go through three major phases, namely:

- integration across different technical disciplines (e.g., data engineers, operation geologists, drilling engineers, reservoir engineers, safety risk analysts, etc.);
- integration across organizations, that is, O&G producers and external business partners (e.g., third party service-support-supply organizations); and
- integration across geographically diversified operational regions (with further expansions to the existing corporate network to other operating regions).

In general, the entire O&G industry remains optimistic about the future and the positive commercial impact of recent initiatives. However, ongoing developments and information technology implementation efforts have already brought a unique set of challenges that includes:

- data integration across disciplines and common standards for active data sharing;
- reliability of digital infrastructure and data security;
- standardisation of ICT platforms across different actors;
- information quality and data filtering techniques;
- work processes integration across disciplines and actors; and
- rapid integration of knowledge based industry.

While information technologies have brought their own challenges (Clarke, Coakes et al., 2003; Gunasekaran, Khalil et al., 2003; Hosni & Khalil, 2004), a major need is also to resolve some critical issues at human and organizational interface levels. A general concern both at socio-political and industrial levels is that ill-defined interfaces and increasing complexities of systems and data solutions can lead to unforeseen consequences, greater vulnerability, and greater risk (see for instance Duffey & Saull, 2003). This mainly demands further attention to a great extent on the critical interfaces between human, organizational, technical, and work process related aspects to realize full-scale benefits of reengineering efforts. For more information see Liyanage (2004a, 2004b, 2005a, 2005b) and Liyanage, Herbert, and Harestad (2006).

CONCLUSION

New asset management practice on NCS aims at reducing the risk exposure, enhancing the commercial value of O&G production assets on the continental shelf. The current setting and ongoing developments have given rise to an interesting *integrated e-operations* setting around offshore assets based on wireless & fiber-optic-based ICT system (i.e., SOIL) that allows 24/7 connectivity and online real-time interactivity among the partners leading the way towards *Smart assets*. Through the *digital capability* it brings together offshore assets, O&G producers, engineering, contractors, drilling companies, component and equipment producers, suppliers, consultants, service providers, and so forth, to a common 24/7 platform to share information, knowledge, and experience to optimize decisions and actions. The Norwegian industry has already experienced a number of success stories of the new environment. ConocoPhillips alone has claimed for substantial cost savings just over a period of nine months. The new environment created by the *smart assets* and *integrated e-operations* has resulted in major changes in human, technological, organizational, and operational aspects, and in particular within electronic information and knowledge sharing practice, organizational forms, work processes, decision loops, and knowledge industry integration. It has also resulted in a mass flow of R&D funds to stimulate innovative solutions to solve some unique challenges. These ongoing developments in general indicate a rapid reengineering process on North Sea assets and a completely new asset management practice based on digital capabilities.

REFERENCES

Barabasi, A. (2003). *Linked: How everything is connected to everything else and what it means for business, science, and everyday life.* Plume Books.

Clarke, S., Coakes, E., et al. (Ed.). (2003). *Socio-technical and human cognition elements of information systems.* Hershey, PA: Information Science Publishing.

Duffey, R. B., & Saull, J. W. (2003). *Know the risk - Learning from errors and accidents: Safety and risk in today's technology.* Butterworth Heinemann.

During, W., Oakey, R., et al. (Ed.). (2002). *New technology-based firms in the new millennium: Volume III.* Elsevier.

Dyer, J. H. (2000). *Collaborative advantage: Winning through extended enterprise supplier networks.* Oxford University Press.

Faulkner, D., & Rond, M. (Ed.). (2001). C*ooperative strategy: Economic, business, and organizational issues*. Oxford University Press.

Gunasekaran, A., Khalil, O., et al. (Ed.). (2003). *Knowledge and information technology management: Human and social perspectives.* Hershey, PA: Idea Group Publishing.

Hosni, Y. A., & Khalil, T. M. (Ed.). (2004). *Management of technology—Internet economy: Opportunities and challenges for developed and developing regions of the world.* Elsevier.

Lindgren, M., & Bandhold, H. (2003). *Scenario planning: The link between future and strategy.* Palgrave Macmillan.

Lipnack, J., & Stamps, J. (1997). *Virtual teams: Reaching across space, time, and organizations with technology.* John Wiley & Sons.

Liyanage, J. P. (2004a). Digital future of operations and maintenance in Norwegian oil and gas production environment: Issues, challenges and opportunities. *The 17th International Congress on Condition Monitoring and Diagnostic Engineering Management (COMADEM-2004),* University of Cambridge, UK (pp. 476-486).

Liyanage, J. P. (2004b). Smart integrated OMS: The 3rd leap to manage the integrity of high risk, capital intensive, and technologically complex offshore assets on NCS. *The 4th Asia Pacific Conference on Systems Integrity and Maintenance (ACSIM-2004),* New Delhi, India (pp. 409-414).

Liyanage, J. P. (2005a). Managing integrity of offshore assets through digital capability: Reducing risk and adding value through Integrated eOMS in North Sea. *The 4th International Conference on Quality & Reliability (ICQR-2005),* Beijing China (pp. 193-200).

Liyanage, J. P. (2005b). Reducing risk and adding value through Smart Integrated Assets: Managing complex industrial assets in complex environments. *The 18th International Congress on Condition Monitoring and Diagnostic Engineering Management (COMADEM-2005),* University of Cranfield, UK (pp. 123-130).

Liyanage, J. P, Herbert, M., & Harestad, J. (2006). Smart integrated e-operations for high-risk and technologically complex assets: Operational networks and collaborative partnership in the digital environment (accepted book chapter), Wang, Y. C. et.al. (Ed.), *Supply chain management: Issues in the new era of collaboration and competition.* Hershey, PA: Idea-Group.

Norwegian Oil Industry Association (2003). *eDrift for norsk sokkel: Det tredje effektiviseringsspranget (eOperations in the Norwegian continental shelf: The third efficiency leap), OLF* (in Norwegian).

Spekman, R. (2003). *Extended enterprise: Creating competitive advantage through collaborative supply chain*. Prentice Hall.

Tidd, J. (Ed.). (2001). *From knowledge management to strategic competence: measuring technological, market and organizational innovation.* Imperial College Press.

Tonchia, S., & Tramontano, A. (2004). *Process management for the extended enterprise: Organizational and ICT networks.* Springer.

KEY TERMS

Digital Capability: Abilities and strengths acquired through active integration of advanced application technologies (e.g., smart sensors, intelligent transducers, electronic gadgets, equipment with advanced functionalities, etc.) and digital infrastructure to optimize decisions and work processes through 24/7 online real-time connectivity.

Digital Infrastructure: Joint fiber-optic and wireless-based advanced information and communication technology platform with embedded multi-functional application services that facilitate 24/7 online real-time connectivity between nodes in the operational network to allow remote management of production assets.

E2E Monitoring: End-to-end monitoring of status of the operational network performed by the administrator of the digital infrastructure.

Extended Digital Enterprise: The organizational form resulted by the growth of conventional organizations into a virtual organizational setting where a number of organizations can simultaneously interact with each other, breaking the boundaries of traditional organizations, through digital infrastructure and digital capability for a common purpose.

Hybrid Techno-Organization: Highly technology dependent and team-based interactive organizational form that functions through a high-level synergy between human and technology.

Integrated Operations: The operational setting where both production assets and a technical support environment are tightly integrated across technical disciplines and organizations creating an active collaborative environment around production assets based on enhanced digital capabilities.

Integrated Work Processes: Work processes integrated across technical disciplines using large-scale information systems to streamline decisions and activities.

Intelligent Data: Data that are automatically acquired through advanced application technologies (smart sensors, intelligent transducers, electronic gadgets, equipment with advanced functionalities, etc.), systematically processed, and are presented in a meaningful form for active sharing, further assessments, and following interpretations.

One-to-Many Connectivity: The ability to connect from one active node to many active nodes in the digital infrastructure on top of the existing network infrastructures allowing consolidation of such conventional ICT networks into one-single common data-hub establishing 24/7 online real-time operational network.

Smart Assets: Those production assets that actively exploit digital capabilities and digital infrastructure smartly and strategically in conjunction with the extended data-knowledge-experience sharing enterprise setting, creating a highly interactive hybrid techno-organizational environment.

Category: IT Education

Smart Learning through Pervasive Computing Devices

S. R. Balasundaram
National Institute of Technology, Tiruchirappalli, India

Roshy M. John
National Institute of Technology, Tiruchirappalli, India

B. Ramadoss
National Institute of Technology, Tiruchirappalli, India

T. Balasubramanian
National Institute of Technology, Tiruchirappalli, India

INTRODUCTION

An increasing number of educators are calling for high standards and challenging learning activities for students. Learning blended with technology can especially provide all possible sources of education. The technologies are not only going to act as technical add-ons to the system but also they can try their best to improve the quality of education.

New technologies can provide meaningful learning experiences for all learners, especially those who are in the developing countries. Educational centers that capitalize on the technological and educational reforms will help students to develop higher order skills and to function effectively in the world beyond the classroom. Achieving such fundamental change, however, requires a transformation of not only the underlying pedagogy but also the kinds of technology applications typically used in classrooms serving at-risk students.

The vision of classrooms structured around student involvement in challenging, long-term projects and focused on meaningful, engaged learning is important for all students. Yet such a change in practice would be especially dramatic for those students who have been characterized as *economically disadvantaged* or *at risk*. Traditionally, schools have had lower expectations for such students. Teachers have emphasized the acquisition of basic skills for at-risk students, often in special pullout programs or in lower level tracks.

BACKGROUND

The impact of technology is seen everywhere—at work, at home, and, indeed, at educational institutions. Educators, policy makers, businesses, and other community groups are looking to technology as a tool for reshaping and improving education.

The educational sectors, whether academic or training divisions, have enjoyed the benefits of technologies in various ways. The technologies used for education range from the storage device technologies to the recent e-learning technologies. Earlier computers were used for storing the contents as well for better information presentation only. With the advent of e-learning, new dimensions are realized by learners, educators, and administrators.

The e-learning environment, where the use of electronic tools like computers and the Internet deliver content, has emerged as the fastest growing segment in the field of education/training and development. The "e" in e-learning focuses on the technology-enabling feature of the learning. The e-learning environment came to forefront for taking the traditional classroom training model and applying technology advancements to create new ways for teaching and learning (Thorpe, 2004).

According to the report produced by the National Committee of Enquiry into Higher Education (2001), the rapid growth in e-learning, has overcome many of the barriers of higher education, thereby providing universities and educational sectors with an opportunity to meet the changing demand for education. The advent of e-learning is inevitably linked to a number of challenges. The real challenge of e-learning centers most on the usage pattern. The biggest myth surrounding e-learning is, "Build it and they will use it."

Recently, e-learning is evolving from the initial technology-driven approach towards a more measured, sophisticated evaluation of its strengths and weaknesses. Recently, the "e" has shifted the focus of learning and training onto the choices in design and delivery of education. But soon, the focus of learning and education will be back where it should be—on learning; the act, process, or experience of gaining knowledge or skill.

The modern-day learners demand knowledge and content that are more sophisticated, dynamic, interactive, and more relevant. Both the learners as well the organizations that depend on them need to learn new skills quickly, and they should have the ability to apply them at the right time. The problem of how to get inside an expert's head and transfer the wealth of knowledge that resides there is being aided by e-learning. While this remains a real challenge, we are beginning to develop tools and approaches for better capturing of that knowledge and delivering it directly into the hands of those who need the same knowledge.

In cases where e-learning is appropriately deployed, educators can generally anticipate student academic performance that is at least equivalent to traditional classroom instruction (Cavanaugh, 2001).

MAIN FOCUS—MERGING PERVASIVE COMPUTING AND E-LEARNING

Generations of Computing

In this history of computing, we have seen several various generations of computing. Initially, the mainframe computing was introduced to enable several people to share one large computer. The problem was that individuals have to personally be present in the proximity of such systems—then came the personal computing era. Personal computers are general purpose devices, designed to do any task. Though it is advantageous, PCs can not be flexibly used for any individual task, that is, not human centric. Another problem related to them is that they are not highly secured. While focusing on stability, security, or transparency aspects of the system, the user's flexibility may be limited.

Network computing is the next big thing that happened. As distinct from stand-alone computing, this term first appeared informally in the late 1970s to denote computers working together over a network. It later came to have a specific technical meaning, denoting a graphical form of remote computing. This led to the introduction of Internet computing.

The growth of the Internet today has exploded into the latest craze. It is the newest wave of communication through e-mail, file transfer, telnet access, transaction applications, and more. This in turn with the invention of World Wide Web has revolutionized computing to a greater extent.

In the history of computing the move towards the next generation of computing—the fourth generation—has happened (see Table 1). Over time, the cost and size of computers has reduced significantly to allow more people to participate in the world of computing (Amor, 2001).

Pervasive Computing

Pervasive computing is the integration of computing power into almost anything, including household equipment, toys, housing, furniture, or even a coffee pot. The name pervasive computing tells only part of the story; a parallel revolution also exists in network-enabling these pervasive computing devices by providing transparent, ubiquitous access to e-business services (Satyanarayanan, 2001). Pervasive solutions enable anytime, anywhere information exchange and access to applications. Davis (2002) points out the analysis of the implications and consequences of pervasive solutions as "anytime/anyplace computing" (p. 3) for future knowledge work.

It is done by natural interaction and control of the ambient environment by people and by artifacts. Interfaces used in pervasive computing support natural communication such as speech and gestures taking into account the user's preferences, personality, and context of use, and enabling multisensory interaction.

Pervasive computing is about making technology and computers disappear. In fact, the more technology becomes transparent, the more business will prosper. The move is towards linking devices like mobile phones, hand-held digital devices, automobiles, refrigerators, and several other easy-to-use devices to the Internet, thereby allowing people to connect anytime, anywhere. It will be pervasive—global—and it will change forever the way we think about the Internet.

Pervasive Approach to e-Learning

Pervasive computing, sometimes called ubiquitous or nomadic computing, describes not only a class of computing device that does not fit the form factor of the traditional personal computer, but also a set of new business models supporting these devices. Ubiquitous computing, prior to pervasive computing, defines a world where computers and associated technologies become invisible and thus makes

Table 1. Generations of computing

Type	Features
Mainframe computing	Many people sharing one large computer
Personal computing	A person works with one computer
Internet computing	One person using several services through global network
Pervasive computing	Many devices serve several people in a personalized way on a global network

them indistinguishable from every day life (Weiser, 1991). With the maturity of computing technologies like wireless LANs, portable and wearable computers, and embedded sensors, a new area emerged from ubiquitous approach called pervasive computing (Satyanarayanan, 2001). Where a desktop computer uses a familiar keyboard, monitor, and mouse interaction model, pervasive computing devices interact in a variety of different ways.

They may use handwriting recognition, voice processing, or imagery. They are often portable and may or may not have a persistent network connection. A pervasive computing device is meant to integrate our lifestyle with the global network of computing, freeing the users from desk-bound application interaction. With the ability to take corporate and personal processes and information with us, no matter our destination, opportunities are plenty for improving and enhancing our personal and professional life. In general, pervasive computing is characterized by certain key areas like smart spaces, invisibility, localized scalability, and uneven conditioning (Satyanarayanan, 2001).

In this context, an approach towards using pervasive computing for e-learning is thought of. Especially a system where embedded devices can help the learners to inform the status and the way in which the teaching aids and laboratory equipment are handled.

Normally, learning about any teaching aid in the curriculum or handling any laboratory equipment in the practical classes needs specialized instructions to be followed by the learners. It is very difficult for all the learners to follow these instructions in the beginning. In the same way, the understanding of these instructions and handling patterns will vary from person to person. Whenever the instrument or equipment is handled abnormally or in the wrong way, instead of the instructor going to the individual student and informing him or her about the improper handling of the equipment, we suggest the ways in which the device itself can be self-acting as a warning or guiding tool. That is the pervasiveness embedded in the device or the context should help the learner to perform things in a better way. By this, the learner and the teacher can use the device *on-the-go*. They do not have to spend more time on reading instruction manuals or searching the *How-To* notes for getting the guidelines. The equipment itself will instruct the user with the help of indicators, that is, lights and sounds. All the user has to know is the meaning of the sounds and the lights. In such a scenario even the normal light context can be used—such as a red light for danger, green light if everything is fine, or a siren sound for an action done by the user if he or she uses the equipment in an improper way.

Case Study: The E-Pipette

The importance of pervasiveness in the e-learning context is highlighted based on a case study in the science laboratory.

We are illustrating the possibility of using pervasive computing in e-learning by adding computing capabilities to a normal conventional pipette used in the chemistry laboratory. The pipette is a laboratory instrument used to transport a measured quantity of liquid. These devices are neither accurate nor precise and cannot be calibrated. Since normal pipettes are made of glass, they are fragile and susceptible to damage. There is a mark on the neck of the pipette which shows the *limit*, and if the user is not watching it, the liquid will enter the mouth.

By adding a circuit board to the pipette, the pipette is made into an intelligent device. The circuit board contains a number of sensors and audiovisual indicators, which helps the user to use the apparatus with comfort and avoids going through any instruction manual. The e-pipette design is shown in Figure 1.

Hardware Design

The heart of the system consists of an ultra low power microcontroller from Texas Instruments powered up by CR2032 Button Cell battery. There is an ADXL202 Accelerometer to sense the tilt and vibrations occurring to the pipette. The liquid level sensor that is present in the system can warn the user if the liquid in the pipette is going above the marked region. The patterns of sounds and the different indicators on the board of the pipette warn the user about the different

Figure 1. E-pipette

conditions the pipette can be used for. The small IR transceiver associated with the system transmits the information regarding the usage patterns to a nearby computer. The details regarding the pattern of handling the device or the laboratory equipment are stored in a database. This database enables the course administrators to analyze the usage pattern and helps them to assess the learners. The block diagram of the e-pipette board is given in Figure 2.

Software Design

The microcontroller in the system has algorithms, which can condition the signals received from the different sensors onboard. The communication protocol enables the connectivity between the computer and the e-pipette. A program running in the personal computer interfaces it with the e-pipette and sends the data over the network. This data helps the invigilator or instructor to provide individual attention to the learners. The entire system block diagram is shown in Figure 3.

Observations from the Screen Shots

When the learner uses the equipment, the sensors capture the usage pattern and store these details in a database. Two graphs are produced for every learner, one depicting the usage pattern defined for a particular instance of time and the other, the real-time plotting for a particular subinterval. The variations in the graph plotting, as shown in Figure 4, are used by the device to inform the learners as well the instructors about the user's characteristics. The remarks provided by the context-aware device define a smart environment thereby providing the learners individual attention during their learning process. The observations from the graph can be used by the instructors to take appropriate actions as and when needed.

FUTURE TRENDS

This idea of linking a pipette with devices to enable better and smart learning can be extended to other learning scenarios also. Pervasive computing provides ample opportunities to define smartness in the academic environment, linking devices with the global network. Further improvements in the learning scenario could be to encourage not only individual learning but also collaborative learning. Through collaborative activities, learners as well as instructors are more closely associated with the system thereby improving the pedagogical aspects of the learning domain. With the promising features of pervasive computing, it becomes possible for both the academic and industrial sectors to devise innovative models for incorporating training related to any concept, at anytime, from any place, and through any device.

Figure 2. Hardware block diagram

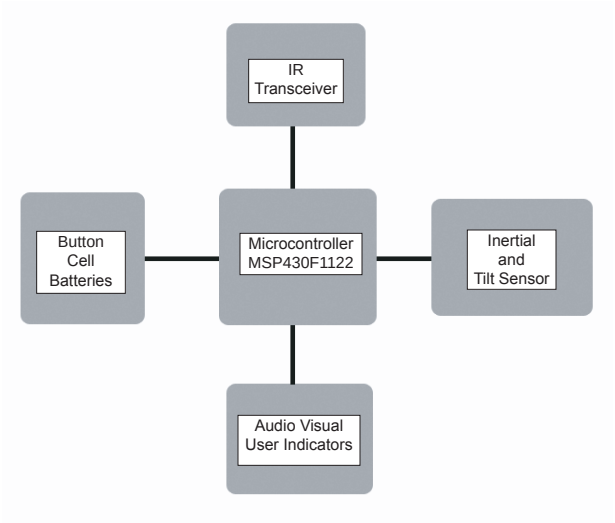


Figure 3. System block diagram

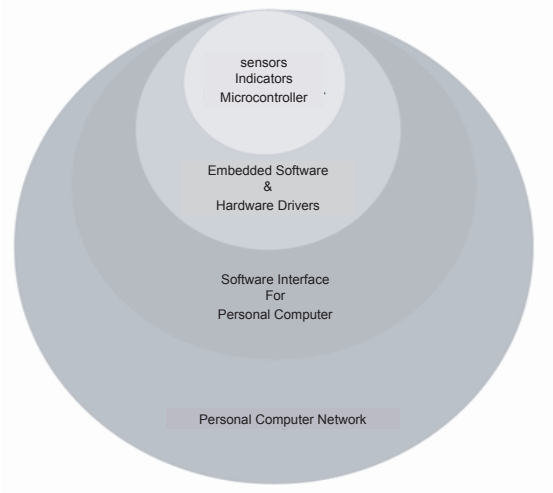
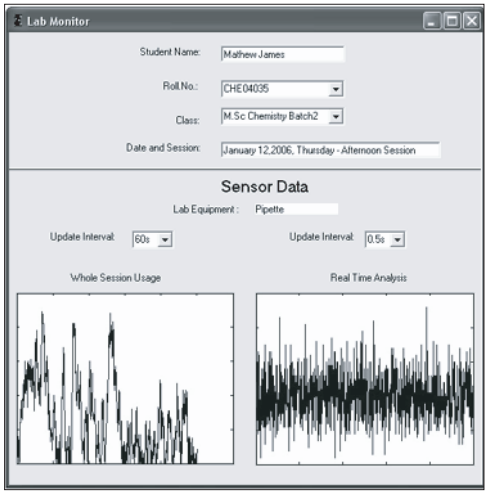


Figure 4. Sensor data analysis



CONCLUSION

The ever-growing nature of educational systems, say academic- or organizational-level projects, poses several challenges to learners and instructors. Due to advancements in technology, the e-learning scenario has brought potential benefits and features to the learning and teaching process. Even then, some of the technical- and context-level hurdles make e-learning to be less effective. Especially, a domain-like laboratory where individual attention is needed for the learners of varying size, or where learners have to be taught how to use the equipment/apparatus as per the rules; e-learning has to be made still more effective. In this context, the use of pervasive computing for the use of pipettes in the chemistry laboratory is discussed. Here, the e-pipette provides enough assistance as well as individual attention to the learners thereby promoting the academic skills to a greater extent. The opportunity and, more importantly, the technology for taking e-learning to the next level are available at our fingertips. The industry has yet to reach its full potential, but it will definitely start to emerge as a profitable business for companies.

REFERENCES

Amor, D. (2001). *Pervasive computing: The next chapter on the Internet*. IN: Prentice Hall.

Archived Information, Technology and Education Reform. (1995, August). Retrieved May 20, 2001, from http://www.ed.gov/pubs/SER/Technology/ch1.html

Beigl, M., Krohn, A., Zimmer, T., Decker, C., & Robinson, P. (2003, October 12-15) Aware-Con: Situation aware context communication. *Proceedings of Ubi-comp 2003*, Seattle, USA.

Beigl, M., Zimmer, T., & Decker, C. (2002). A location model for communicating and processing of context. *Personal and Ubiquitous Computing,* 6(5-6), 341-357.

Beigl, M., Zimmer, T., Krohn, A., Decker, C., & Robinson, P. (2003). *Smart-its—Communication and sensing technology for UbiComp environments*. (Tech. Rep. ISSN 1432-7864 2003/2). In *Proceedings UbiComp2003*, Seattle, WA, October 12-15. Retrieved from http://www.ubka.uni-karlsruhe.de/cgi-bin/psview?document=ira/2003/2

Cavanaugh, C. (2001). The effectiveness of interactive distance education technologies in K-12 Learning: Meta analysis. *International Journal of Educational Telecommunications, 7*(1), *73-88.* Retrieved June 1, 2002, from http: www.unf.edu/~ccavanau/CavanaughIJET01.pdf

Davis, G. B. (2002). Anytime/anyplace computing and the future of knowledge work. *Communication of the ACM, 45*(12), 67-73.

Mukherjee, S.D. (2003, March). Pervasive computing: A paradigm for the 21st century. *IEEE Computer, 36*(3), 25-31.

Dong, M. J., Yung, G., & Kaiser, W. J. (1999, August 18-20). Low power signal processing architectures for network microsensors. In B. Barton, M. Pedram, A. Chandrakasan & S. Kiaei (Eds.), *Proceedings of International Symposium on Low Power Electronics and Design,* Monterey, CA, (pp. 173-177). ACM.

Edwards, W. K., & Grinter, R. E. (2001). At home with ubiquitous computing: Seven challenges. In G. D. Abowd, B. Brumitt, & S. A. Shafer (Eds.) *Proceedings of the UbiComp 2001*, Atlanta, GA. Springer, LNCS.

Gellersen, H. W., Beigl, M., & Schmidt, A. (2000, June 7-9). *Sensor-based context-awareness for situated computing.* SEWPC00 (Wokshop on Software Engineering for Wearable Pervasive Computing), Limerick, Ireland.

Hands on technology transfer Inc. (2002, October 25). *E-learning myths and realities for the IT professional.* Retrieved November 27, 2005, from http://hosteddocs.ittoolbox.com/SM072402.pdf

Ljungstrand, P., Holmquist, & Lars, E. (1999). WebStickers: Using physical objects as WWW bookmarks. In *Extended Abstracts of CHI'99,* Pittsburg, PA (pp. 332-333). ACM Press.

National Committee of Enquiry into Higher Education. (2001). *National Report, Chapter 13; Communications and Information Technology*, Retrieved November 4, 2003, from www.leeds.ac.uk/eduol/ncihe/nr_202.htm

Satyanarayanan, M. (2001, August). Pervasive computing: Vision and challenges. *IEEE Personal Communications, 8*(4), 10-17.

Thorpe, T. W. (2004). Exploring e-Learning myths. Retrieved January 16, 2006 from, http://www.twthorpe.com/index.php/2004/05/29/exploring-e-learning-myths/

Weiser, M. (1991, September). The computer for the 21st century. *Scientific American, 265,* 94-104.

KEY TERMS

Accelerometer: An instrument that measures acceleration.

ADXL202E Accelerometer: A low cost, low power, complete 2-axis accelerometer with a measurement range of ±2 *g*. The ADXL202 can measure both dynamic acceleration (e.g., vibration) and static acceleration (e.g., gravity).

E-Learning: Learning that is facilitated and supported through the use of information and communication technology, e-learning can cover a spectrum of activities from supported learning, to blended learning (the combination of traditional and e-learning practices), to learning that is entirely online.

IR Transceiver: IR transceiver deals with infrared communications.

Microcontroller: A computer on a chip used to control electronic devices. It is a type of microprocessor emphasizing self-sufficiency and cost effectiveness, in contrast to a general purpose microprocessor, the kind used in a PC.

Pervasive Computing: Pervasive computing consists of inexpensive microprocessors embedded in everyday objects and environments. Characterized by being numerous, casually accessible, often invisible computing devices, frequently mobile or imbedded in the environment and connected to an increasingly ubiquitous network structure.

Transceiver: Communications device capable of both transmitting and receiving.

SMEs Amidst Global Technological Changes

Nabeel A. Y. Al-Qirim
United Arab Emirates University, UAE

BACKGROUND AND IMPLICATIONS

In small countries such as New Zealand, small to medium-sized enterprises (SMEs) are defined as enterprises employing 19 or fewer employees. Small enterprises are defined as those employing zero to five full-time employees (FTEs) (often called microbusinesses), and medium-sized enterprises as those employing six to nineteen FTEs. Other countries, such as the United States and European countries, define their SMEs as having a much larger number of employees (200–500 or fewer).

SMEs contribute significantly to the economies and to the employment levels of different countries in the world. For example, SMEs constitute around 95 percent of enterprises and account for 60–70 percent of employment within the countries of the Organisation for Economic Cooperation and Development (OECD, 1997) and other countries across the globe, including the United States. Not to forget that SMEs are usually the source of most of the profound inventions and innovations (Iacovou, Benbasat, & Dexter, 1995).

Historically, SMEs have been accused of being uncritical about the strategic importance of IT and its use in their businesses. This laggardness in adopting or using IT in business was attributed to various organisational, technological, and environmental deficiencies in SMEs. The recent emergence of the Internet, in general, and the Web, in particular, revolutionises business activities (Abell & Lim, 1996) and promises to provide unprecedented opportunities to SMEs to expand in scope and in market reach.

However, despite the apparent media hype (Premkumar & Roberts, 1999) and the enthusiasm among academicians (Adam & Deans, 2000; Abell & Lim, 1996; *Infotech Weekly*, 1997; Poon & Swatman, 1999a) and professionals (Deloitte, 2000; IDC, 1998; PWHC, 1999) about electronic commerce (EC), the published EC research portrayed a gloomy picture about EC uptake and use by SMEs. Thus, investigating reasons behind such laggardness in adopting and in using EC effectively is essential. This research attempts to highlight some of the important issues that could assist in bridging the existing divide between SMEs and EC. These issues could be of interest to SMEs and to other stakeholders interested in SMEs and EC.

ELECTRONIC COMMERCE SUCCESS IN SMES

In the SMEs scenario, different research emphasised the different EC advantages to SMEs (Abell & Black, 1997; Abell & Lim, 1996; Adam & Deans, 2000; Deloitte, 2000; Poon & Swatman, 1997, 1998, 1999a,b; PWHC, 1999):

1. The Internet is an efficient communication medium and a vast resource for information. The SMEs could use e-mail technology to communicate efficiently with their buyers and suppliers, reducing communication costs, including the buying of expensive equipment (e.g., fax/telex).
2. The Internet provides added-value services to customers/partners/suppliers by providing different primary/supplementary information about the organisation's industry, products, and services on their Web sites. This could result in increasing the loyalty and the stickiness of their customers (customer resource management; CRM). The preceding tangible and intangible tactics are of strategic importance in retaining and increasing customer bases by increasing switching costs.
3. The Internet would provide new opportunities to SMEs, otherwise not possible before the introduction of the Internet, such as the ability to reach global markets and the ability to mass-customise products and services to appeal to the different tastes of global consumers.
4. SMEs would adopt EC for image-enhancement purposes. Having an Internet account (URL, dot-com, Web page) and printing an e-mail address on business cards and letterhead were reported as major drivers as well. Whether the SMEs were able to elevate from such initial depiction to a more strategic posture in adopting more strategic EC initiatives, such as selling and buying online, is worth further investigation from the perspective of the different countries.

On the other hand, the EC research highlighted the following impediments:

1. Technological impediments: e.g., security (privacy concerns, viruses, e-payments), legalities (enforceability of contracts, confirmations of receipt, prosecutions), policies (lack of global or unified standards),

telecommunication services [bandwidth, convergence, reliability and quality of services (QoS)]
2. Organisational impediments: cost, busy nature, small size and limited resources, lack of knowledge/expertise about EC
3. Environmental impediments: Relating to the lack of regulatory frameworks pertinent to the above technological impediments highlighted in (1), above, either at the one-country level or even at the global level.

In the light of the above advantages and impediments, most of the existing EC research found most of the SMEs not witnessing real benefits (direct sales and tangible profits) in the short term due to difficulties in selling products over the Internet (Adam & Deans, 2000; Poon & Swatman, 1998, 1999a). Face-to-face interactions with customers and buyers proved to be more dominating than electronic interfaces (Ba et al., 1999; Poon & Swatman, 1997). They found the key motives for SMEs to adopt EC are the long-term indirect benefits, e.g., ongoing business transformation and new business initiatives (new opportunities), which could resemble a preparatory stage (infrastructure) for the long-run direct benefits stage (secure returning customers and form long-term business partnerships) (Poon & Swatman, 1998, 1999a). However, the biggest challenge for the SMEs here is to succeed in moving from such simple and preparatory EC initiatives (driven mostly by hype from the media, professionals, and researchers) to more sophisticated and strategic initiatives (e.g., efficiency à effectiveness à strategic advantage).

On the one hand, having EC requires an apparent investment in different areas: technological infrastructure upgrades or replacement, EC integration with existing IT systems, EC consultants, investments in bandwidth and applications (Web site, intranet, extranet, etc.). However, this considerable investment in the EC infrastructure is necessary to make it possible to process information efficiently, handle heavy traffic, and deliver satisfactory performance. SMEs would perceive this to be an expensive endeavour and, hence, represent a barrier to EC adoption (MOED, 2000; PWHC, 1999). It is worth mentioning here that unlike the investment in information science/information technology (IS/IT), which requires high initial investment and smaller ongoing maintenance and support costs, EC would require considerable continued investments in upgrading, overhauling, and replacing the whole EC system with an innovation or new designs, etc. Most probably, the investment in EC would materialise in the long term only as highlighted earlier (Poon & Swatman, 1998, 1999). However, this depends on different factors, such as the ability to develop economies of scale (Ba et al., 1999; Poon, 2000), e.g., having a well-established online customer base and ongoing business that enables the firm to sell massively and cheaply at the same time.

With the introduction of new EC technology like the intranet, Internet electronic data interchange (EDI), extranet, Web site, etc., there would be some fundamental changes in work processes and current practices (Alexander, 1999; Behrendorff & Rahman, 1999). EC is not only a new way of selling and marketing, but also a new way of thinking, which requires a change of mindset. Teo, Tan, and Buk (1998) pointed to the fact that organisations attempting to adopt the Internet should expect a possible change in communication and culture patterns. EC is changing the way business is conducted, even with individual customers. Firms that are able to streamline their products or processes or delivery agents on the Internet will be able to shift entirely to the pure EC arena (Choi et al., 1997). The success stories of small businesses using the Internet are apparent and are publicised and reported by the media. However, most of the businesses existing on the Internet are not necessarily transacting information-based products only, but rather complementing the sale and the delivery of a physical product with such things as publishing information about the usability of a physical product (e.g., user manuals), tracking the shipment, etc. (Teo et al., 1998).

Most of the IS literature on SMEs (Blili & Raymond, 1993; Cragg & King, 1992, 1993; Harrison et al., 1997; Jarvenpaa & Ives, 1991; Thong, 1999; Thong & Yap, 1995, 1996) and EC in SMEs (Poon & Swatman, 1998, 1997, 1999a, 1999b) emphasises the role and the characteristics of the manager (usually the owner) as a product champion. Poon and Swatman (1998, 1999a) pointed to the manager's role in their EC study, where they found direct management involvement was the norm in the different cases. Although the managers of small business lack formal IT qualifications and training, they were champions in adopting EC, specifically in microbusinesses, where the sole decision maker was the director of the business.

Due to the recent nature of EC, it is expected that the adoption decision for EC would include some sort of high-risk elements. Hence, the adoption decision for EC would require a risk-taking manager. Poon and Swatman (1997, 1998, 1999a) found that the entrepreneurial perspective differed between the different firms in their study. Managers/owners embraced EC technology and attempted to exploit it to the maximum. The managers who championed Internet adoption in their organisations demonstrated an innovative and risk-taking attitude toward EC, despite lacking formal IT training.

Adam and Deans (2000) and Poon and Swatman (1998) pointed to the market scope of small business, where SMEs transacting with international markets would perceive many advantages from the Internet, such as cost savings and market communication in comparison with other SMEs operating in local markets. In this scenario, EC is perceived to increase global competition and provide different opportunities to SMEs. Poon and Swatman (1999a) asserted that

if a small business retained a high percentage of customers and competitors online, this would increase the chances of adopting EC.

The field of EC is relatively new, and the actual functioning and utilisation of EC technologies are still unknown to most organisations (Teo et al., 1998), including SMEs. Therefore, it is expected that SMEs planning to adopt EC would seek assistance from consultants and vendors in the industry in different areas, such as planning and strategy, training, development, and implementation (Deloitte, 2000). Determining how efficient the technology vendors are in providing feasible and well-integrated EC products and services to SMEs is worth investigating in different countries.

CONCLUSION

SMEs contribute significantly to the national economies and to the employment levels of different countries and represent a viable source for inventions and innovations. The emergence of EC in the early 1990s could provide different opportunities to the small business sector to overcome its inadequacies. However, in review of the electronic commerce/business (EC) literature in organisations, in general, and in SMEs, specifically, it was observed that the available research portrayed a gloomy picture about EC uptake and use by SMEs. Therefore, this research attempted, by reviewing recent EC research, to depict an agenda for EC success in SMEs. By following the suggested guidelines in this research, SMEs could be in a better position to assess the viability of the new EC phenomenon to their survivability in the long term. Specifically, these points are addressed to the managers/owners of the SMEs in identifying the different perspectives surrounding the new innovations. These factors are of high importance to researchers, SMEs, professionals (including educational institutions), and policy makers in driving SMEs and EC forward.

REFERENCES

Abell, W., & Black, S. (1997). Business use of the Internet in New Zealand: A follow-up study. Retrieved August 8, 2000, from *http://www.scu.edu.au/ausweb96/business/abell/paper.htm*

Abell, W., & Lim, L. (1996). Business use of the Internet in New Zealand: An exploratory study. Retrieved August 8, 2000, from *http://www.scu.edu.au/ausweb96/business/abell/paper.htm*

Adam, S., & Deans, K. (2000). Online business in Australia and New Zealand: Crossing a chasm AusWeb2k—The Sixth Australian World Wide Web conference,

Alexander, A. (1999, December). Tuning small business for e-Commerce: Consultants say business consulting is essential, even in e-commerce. *Accounting Technology, 15*(11), 48–53.

Ba, S., Whinston, A., & Zhang, H. (1999, December). Small business in the electronic marketplace: A blue print for survival. *Texas Business Review.* University of Texas, Austin.

Behrendorff, G., & Rahman, S. (1999). Adoption of electronic commerce by small to medium enterprises in Australia. In F. Tan, P. Corbett, & Y. Wong (Eds.), *Information technology diffusion in the Asia Pacific: Perspective on policy, electronic commerce and education* (pp. 130–147). Hershey, PA; London: Idea Group Publishing.

Blili, S., & Raymond, L. (1993). Information technology: Threats and opportunities for small and medium-sized enterprises. *International Journal of Information Management, 13,* 439–448.

Choi, S., Stahl, D., & Whinston, A. (1997). *The economic of electronic commerce*. New York: Macmillan Technical Publishing.

Cragg, P., & King, M. (1992). Information systems sophistication and financial performance of small engineering firms. *European Journal of Information Systems, 1*(6), 417–426.

Cragg, P., & King, M. (1993). Small-firm computing: Motivators and inhibitors. *MIS Quarterly*, March.

Deloitte Touche Tohmatsu. (2000). Deloitte e-Business survey: Insights and issues facing New Zealand business. Retrieved August 8, 2000, from *http://www.deloitte.co.nz/images/acrobat/survey.pdf*

Iacovou, C., Benbasat, I., & Dexter, A. (1995, December). Electronic data interchange and small organisations: Adoption and impact of technology. *MIS Quarterly,* 465–485.

Infotech Weekly. (1997, April 1). New Zealand Internet use. Retrieved May 15, 2000, from *http://www.nua.net/surveys/index.cgi?f=VS&art_id=863080905&rel=true*

International Data Corporation (IDC). (1998). Ecommerce booming in New Zealand. Nua Internet Services: Retrieved April 30, 1998, from http://www.nua.ie/surveys/index.cgi?f=VS&art_id=905354498&rel=true. Retrieved May 15, 2000, from *http://www.nua.ie/surveys/index.cgi?f=VS&art_id=905354498&rel=true*

Jarvenpaa, L., & Ives, B. (1991, June). Executive involvement and participation in the management of information technology. *MIS Quarterly, 15*(2), 205–227.

OECD. (1997). Small business, job creation and growth: Facts, obstacles and best practices.

Poon, S. (2000). Business environment and Internet commerce benefits—A small business perspective. *European Journal of Information Systems, 9*, 72–81.

Poon, S., & Swatman, P. (1997). Internet-based small business communication. *International Journal of Electronic Commerce, 7*(2), 5–21.

Poon, S., & Swatman, P. (1998). A combined-method study of small business Internet commerce. *International Journal of Electronic Commerce*, *2*(3), 31–46.

Poon, S., & Swatman, P. (1999a). An exploratory study of small business Internet commerce issues. *Information & Management, 35,* 9–18.

Poon, S., & Swatman, P. (1999b). A longitudinal study of expectations in small business Internet commerce. *International Journal of Electronic Commerce, 3*(3), 21–33.

Premkumar, G., & Roberts, M. (1999). Adoption of new information technologies in rural small businesses. *The International Journal of Management Science (OMEGA*), *27*, 467–484.

PWHC (Pricewaterhousecoopers). (1999, September 24). SME electronic commerce study (TEL05/97T). Retrieved April 10, 2000, from *http://apec.pwcglobal.com/sme.html*

Rihga Colonial Club Resort, Cairns, June 12–17. Retrieved August 8, 2000, from *http://ausweb.scu.edu.au/aw2k/papers/adam/paper.html*

Teo, T., Tan, M., & Buk, W. (1998). A contingency model of Internet adoption in Singapore. *International Journal of Electronic Commerce*, *2*(2), 95–118.

Thong, J. (1999). An integrated model of information systems adoption in small business. *Journal of Management Information Systems, 15*(4), 187–214.

Thong, J., & Yap, C. (1995). CEO characteristics, organisational characteristics and information technology adoption in small business. *Omega, International Journal of Management Sciences*, *23*(4), 429–442.

Thong, J., & Yap, C. (1996). Information technology adoption by small business: An empirical study. In K. Kautz & J. Pries-Heje (Eds.), *Diffusion and adoption of information technology* (pp. 160–175). London: Chapman & Hall.

KEY TERMS

Economical Importance of SMEs: SMEs contribute significantly to the economies and to the employment level of different countries in the world. For example, SMEs constitute around 95% of enterprises and account for 60% to 70% of employment within the countries of the OECD (OECD, 1997) and other countries across the global including the U.S. Not to forget the SMEs are usually the source for most of the profound inventions and innovations (Iacovou, Benbasat, & Dexter, 1995).

IT/E-Commerce Adoption and Use in SMEs: Historically, SMEs have always been accused of being uncritical about the strategic importance of IT and its use in their businesses. This laggardness in adopting or using IT in business was attributed to various organizational, managerial, technological and environmental deficiencies in SMEs. The recent emergence of the Internet in general and the Web in particular revolutionizes business activities (Abell & Lim, 1996) and promises to provide unprecedented opportunities to SMEs to expand in scope and in market reach.

Small Business Internet Commerce: the use of Internet technology and applications to support business activities of a small firm (Poon, 1999).

Small- to Medium-Sized Enterprises (SMEs): In small countries such as New Zealand, SMEs are defined as enterprises employing 19 or fewer employees. Small enterprises are defined as those employing zero to five full-time employees (FTEs) (often called microbusinesses) and medium-sized enterprises as those employing six to nineteen FTEs. Other countries, such as the United States and European countries, define their SMEs as having a much larger number of employees (200–500 or fewer).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2519-2522, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Social and Legal Dimensions of Online Pornography

Yasmin Ibrahim
University of Brighton, UK

INTRODUCTION

The dialectics between private pleasures and public needs raise various dilemmas, especially in the domain of the erotic and aesthetics. These are relative and abstract terms that can vary from individual to individual. However, in the public spaces of the Internet, the need for community standards of decency, acceptability, and taste often drag many of the debates about the Internet into a legal space, despite its description as a virtual sphere and the libertarian endeavours to keep it free from government and organizational control. While the Internet is a global resource it is often ruled through the laws of its physical embeddedness, and the global nature of the Internet also means that it is consumed and assessed through the differing cultural practices and norms that prevail in various parts of the world. The Internet as a communication and information platform is then subject to varying codes of ethical and moral conduct by different communities whether online or off-line. While the realm of the erotic is often equated with individual pleasure and psyche, the proliferation of pornography on a public platform raises social, moral, and legal concerns for communities, states, and governments. One significant element in the development of the Internet as a market place has been the availability of explicit sexual material, and these electronic networks continue to feed the pornography boom and facilitate new methods for consumers to interact with sexual content as "porn" (Spencer, 1999). These networks highlight the "privatising" potential of technology, especially in relation to sexual matters, while illuminating new forms of formal and informal exchanges (Jacobs, 2004, p.72; Spencer, 1999). The Internet, from being a rather unregulated enterprise a few years ago, has recently become the focus of multiple ethical concerns and debates and in some cases, it has amounted to moral panic (Bkardjieva & Feenerg, 2000; Cavanagh, 1999).

BACKGROUND

The emergence of gaming culture and the simulation of reality through the design of gaming technology raises the age-old issues about image and representation; the effects it can have on our cognitive senses, and how these can, as a result, affect or mediate our ability to reason and engage with interactive technology. These questions become ever more salient with regard to online pornography or sexually explicit material. The distinctive element about online porn is its use of multimedia, its ubiquity, and consumer access to it. Due to the anonymity of the Internet and the difficulties in regulating this transnational and anonymous medium, transgressive forms of entertainment, including pornography, have flourished online. According to Spencer (1999), the Internet is structured at one level around the economics and politics of consumption, at another level around the politics of individuality, and at another around communitarian concerns (p. 242).

Online pornography has been acknowledged as a relatively new form of pornography. Authors Stack, Wasserman, and Kern (2004) point out that there were about 900 pornography sites on the Web in 1997 and just a year later, the figure had burgeoned to between 20,000 to 30,000 sites, with revenues reaching US$700 million by the late 1990s. Its growth has been attributed to the "triple a-engine" of accessibility, affordability, and anonymity (Cooper & Griffin-Shelly, 2002, p. 11). Fisher and Barak (2001) agree that "spectacular growth in availability of sexually explicit material on the Internet has created an unprecedented opportunity for individuals to have anonymous, cost-free, and unfettered access to an essentially unlimited range of sexually explicit texts, moving images and audio materials" (p. 312). This increased accessibility and convenience, as well as the exploiting of e-commerce by pornographers, means that the Internet makes it easier for individuals to come into contact with porn. Some suggest that this has enabled the normalization of practices that may have otherwise been stigmatized in traditional markets, leading to a mainstreaming of cyberporn through its visibility and presence (See Cronin & Davenport, 2001, p.35; O'Toole, 1998). In the last few years, undoubtedly, there has been increasing heterogeneity and decentralization on the Internet as a wider variety of producers and consumers participate in the making of globalized markets, and a contemporary notion of pornography should capture such networked sexual agency and politics (Jacobs, 2004).

MAIN FOCUS

Diane Russell (1998) defines sexually explicit material as that which "combines sex and/or the exposure of genitals with abuse or degradation in a manner that appears to endorse,

condone or encourage such behaviour" (p. 3). James Check (1985), on the other hand, terms pornography as "sexually explicit material" without further qualifying it. The Internet poses new questions about the reality, regulation, definition, and availability of pornography, as it has dramatically increased the accessibility of pornography, and of violent pornographic images in particular (Gosset & Byrne, 2002). The danger of pornography to adults is much more disputed, and often the arguments for pornography include freedom of speech and the expression of civil liberties, the right to choose, and the right to privacy (Kuipers, 2006).

Nevertheless, what constitutes pornography is often contested in societies. While in terms of ethics adult pornography is a contested terrain, child pornography, on the other, is almost universally prohibited. But in the online environment, a digital image can be manipulated and altered and consequentially it may be difficult to clearly define the distinction between adult and child images. Jenkins (2001) posits that child pornography can be accessed in various ways in the online environment, where it can be distributed via credit-card access Web sites, bulletin boards, and encrypted e-mails, as well as through peer-to-peer file sharing. These are constitutive of not only the new configurations between producers and users, but also of new forms of abuse (Oswell, 1999, 2006, p. 253).

A report by the Washington-based research and policy group, Third Way, highlights how this accessibility and presence can present new problems for Internet users, particularly children (cf. Whitehead, 2005). According to the report, only 3% of more than 450 million individual porn Web sites ask for proof of age. Additionally, a majority of these Web sites do not carry any warning of adult content, and nearly three-quarters display free teasers of pornography images on their homepages; it is therefore likely that children may accidentally come across a porn site while doing homework or surfing the Web. While child pornography is almost universally illegal, adult pornography is prevalent and easy to access on the Web. Whitehead (2005, p. 18) contends that unlike off-line pornography, which can be curbed through measures imposed by the community such as zoning laws and curfews, the politics of online pornography is very different, as online porn, through its technology, can be seen to be "everywhere and nowhere." This has meant the loss of power for parents to control what their children come into contact with.

The status of a photograph as a verifiable fact continues to linger with the Internet despite the radical impact of digital technology on photographic practice. This has been problematic as digital technology can manipulate and distort images, thus further compounding the relationship between reality and representation. David Oswell (2006) describes this as the ethics of virtual observation, where the referentiality of the image in representing the scene of abuse that is real, and our ethical response to it, is predicated by our perceptions of reality (p. 258).

In discussing the ethics of the virtual, Oswell (1999, 2006) observes that there are often epistemological inconsistencies and disjuncture between knowledge, law, and sociological perspectives. In the context of the US and UK, there is often an implicit understanding that child pornography is the record of actual child sexual abuse, and this has become widely used in legal discourses and public discourses of law enforcement charities and child protection agencies (Williams, 1991, p. 88).

According to Tate (1990), while child pornography has been a problem for decades, until relatively recently it has been a hidden crime (p. 1). In the UK, the principal legislation that addresses the indecent images of children is the Protection of Children Act 1978 (PoCA), which differentiates between different mediums representing the abuse. Photographic images are the subject of this specific legislation that focuses on child pornography, whereas all other mediums are treated as obscene articles and are subject to general obscenity legislation (Gillespie, 2005).

The ontological status of visual depiction has legal ramifications both in the context of the UK and the US. In the US, the Child Pornography Prevention Act (CPPA) of 1996 addresses the legal implications of visual depiction. In 2002, a ruling by the US Court of Appeals for the Ninth Circuit found that the CPPA ruled solely on the image without considering the set of contextual factors catered for in an earlier ruling in 1973, and in view of this, the CPPA's emphasis on the image was "overbroad" and "unconstitutional" (Oswell, 2006). It also reiterated that the proximity or distance of a photograph from the scene of the actual event is an important criterion in the legality of child pornographic images. The court also overruled the argument in the CPPA that virtual child pornography is "intrinsically related" to the sexual abuse of children as the link between the two is contingent and indirect. Harm here does not necessarily accrue from the image but is dependent on the potential for subsequent criminal abuse. The ruling showed the court's unease with the assumption that the image takes up the position of the "modest witness" whose account of the scene is "unadorned, factual and compelling" (Haraway, 1997, p. 26).

Virtual child pornography may have no link to crime or sexual abuse that has actually been committed, and in the same vein, the virtual image has no necessary link to future cases of abuse. As with child pornography, virtual child pornography cannot be prohibited on the basis of its possible harm to some children or the possibility some children may be exposed to it. In this sense, the CPPA defies the "principle that speech within the rights of adults to hear may not be silenced completely in an attempt to shield children from it" (Oswell, 2006, p. 251). In April 2002, the US Supreme Court found the Child Pornography Prevention Act (CPPA) unconstitutional. Though it remains illegal to make, show, or possess sexually explicit material of children, the court

found that there were not compelling reasons to prohibit the manufacture or exhibition of pictures that merely appear to be children. As a consequence, two categories of pornography that were prohibited under the act are now permitted in the US. These include sexually explicit pictures of actual models who appear to be younger than they are and computer-generated sexually explicit pictures of children (Gillespie 2005; Levy, 2002; Oswell, 2006).

In contrast, the UK treats indecent pseudo-photographs of children "as indecent photographs of children." In the UK, the term pseudo-photography was introduced by the Criminal Justice and Public Order Act of 1994. This act defines a "pseudo-photograph" as "an image, whether made by computer graphics or otherwise, which appears to be a photograph" (cf. Gillespie 2005, p. 435). With the police finding images on computers that could not be readily verified as those of a child or an adult, the 1994 amendment became the rationale for addressing this problem (Hansard, 1994, cf. Gillespie, 2005, p. 435).

In UK legislation, while the indecent photograph and the indecent pseudo-photograph are not identical, they are treated as identical. This means that the act of downloading child pornography constitutes a crime, regardless of whether these images are records or not of actual abuse. This legal response to pornography intrinsically associates an image and a crime and provides a legal platform for the authorities to act. In contrast, the actual images are illegal in the US but not the virtual. Oswell (2006, p. 252) points out that such an equivalence creates challenges in a court of law as reservations can be raised as to the evidential status of the image, i.e. as to whether the image is an image of the incident of sexual abuse at all. Oswell contends that the photograph becomes the measure of the real and its observation, and hence, the implicit prioritization in UK law of virtual child pornography means the crime of possession, making, or distribution of child pornography (whether virtual or real) is a crime not only against a particular child but against all children, invoking it as a universal crime against childhood. The debates on adult pornography often delineate between the vague boundaries of erotics and aesthetics but these distinctions can be subjective and may influenced by the context of the immediate society, making them arbitrary criteria.

Prior to the Internet, the debate about pornography in the US centred on the First Amendment, and the need to regulate pornography was stressed on the grounds that pornography violates community standards. Under the US Constitution, it is legal for adults to own and distribute most types of pornography. However, since the 2002 Supreme Court Ruling over COPA (Child Online Protection Act), the US government has made serious efforts to monitor and impose restrictions on Internet pornography traffic by arguing that juveniles or minors (18 years or less) are automatically exposed to and harmed by pornographic images (Jacobs, 2004). As Taylor and Quayle (2003, pp. 159-163) point out, one of the principal elements of Internet-facilitated child pornography is an exponential growth in the size of the individual collection. Here the imaging and archiving features of Internet technology cannot be overlooked. In the UK, for example, legislation does not distinguish between accessing images for personal use (including downloading) and the creation and distribution of images (Gillespie, 2005).

Often censorship and obscenity laws have provided the basis to tackle pornography on the Internet, and this has been the case in the US, Australia, the UK, and France. In Australia, the Western Australia Censorship Act (1996) is designed to protect the local and state territory from the influx of pornographic material over the Internet (Jacobs, 2004:, p.71). On the other hand, South Australia's Censorship Act criminalizes any content that is deemed "unsuitable for children online," even if the content is intended for adults. This leaves content open to police interpretation, as authorities can evaluate and arrest users who post information that is deemed offensive to children.

Akedeniz (2002) stresses that there is a difference between illegal and harmful content, as the former is criminalized by national law while the latter is merely deemed offensive or disgusting by some sections of society. In tandem with this, Jacob (2004) queries whether community standards of decency can be transmitted from one place to another. With the Internet being perceived as a global resource, the issue of community standards creates different cultural and legal approaches to solving the issues at hand.

While there is often a societal acknowledgement of child pornography as a heinous and universal crime, there is, nevertheless, a difficulty in defining what constitutes child pornography as different jurisdictions can define it differently, and equally, the issue of obscenity can also be culturally mediated in different environments. The consensus in terms of what constitutes child pornography can emerge within the context of supra-national agencies such as the Council of Europe, which defines child pornography as "any audiovisual material which uses children in a sexual context" (Oswell, 2006, p. 246).

Akdeniz (2001) points out that the legal regulation of this sort of Internet content may differ from country to country, and this is certainly the case within the European Union, where member countries have taken different approaches to sexually explicit material. Akdeniz streses that in terms of Internet content and young users, harm remains a criterion, and this is accepted within the jurisprudence of the European Court of Human Rights. Harm in societies again is culturally defined. In terms of illegal or harmful content, the UK adopts a multilayered approach with the involvement of both national and international levels. The government also favours a coregulatory approach in which there is a role to be played by the industry's own self-regulation.

Before the Internet, the US already had some antipornography legislation that has since been applied to the Internet.

In the US, pornographic sites are legally obliged to refuse access to minors (Kuipers, 2006). Until recently, further attempts to penalize or regulate Internet pornography failed due to the fact they conflicted with the First Amendment. Beyond legal restraints, countries can also encourage the use of technology to filter undesirable content deemed harmful to children. In June 2003, the Child Internet Protection Act (CIPA) was approved by the US Supreme Court to force libraries and schools to block pornographic sites. The CIPA requires public libraries to filter their computers if they want to retain federal funding, but such software is not completely reliable. Judith Levine (2002) has argued that it is important to promote media literacy and moral intelligence rather than to deal with the Internet through technology. This means that governments and societies should also invest in public awareness and education campaigns instead of phasing out controversial sexuality debates that can polarise the public.

FUTURE TRENDS

Online pornography in many countries has been regulated through existing censorship and decency laws. However, due to the complexity of the Internet environment and the availability of different technologies on one platform, there has been a need to enact and revisit what a photograph or "psuedo-photograph" can constitute in the digital environment and the consequences of admitting it as evidence in courts. The legal trajectories in the US and UK highlight the complexities of the digital image and its treatment in different legal contexts. In the future, new forms of technologies and the convergence of these on the Internet will pose more challenges for the legal domain. Societies would then have to enact new legislation to cope with legal systems that predate modern technologies and the challenges they present. It illuminates the sort of legal and moral dilemmas that have emerged from the Internet environment.

The criminalization of child pornography and the ubiquity of pornography on the Web raise legal issues for users as their private actions in their personal domain can have legal consequences. On an international level, in the domain of child pornography there has been cooperation at a global level between governments to share databases of perpetrators involved in child trafficking and pornography. The future as such alludes to both centralization and surveillance to curb criminal behaviour on the Web as well as to the rise of fragmented and diverse forms of pleasure-seeking on the Internet.

CONCLUSION

The issues of online pornography are entangled with private pleasures, community standards, increasing commercialisation of the Internet, and a proliferation of sites and images offering sexually explicit content. It raises the need to address numerous concerns and issues with regard to online pornography. Concepts of privacy, harm, offence, taste, decency, legality, and protection of children as a universal ideal as well as the evidential status of the digital image on the Internet compound the problems of online pornography. Online digital images, and their referential authenticity to the actual event or person, capture the complexities of the Internet as a medium where convergence of technologies and amalgamation of sound and images, as well as editing technologies, enable simulated realities and new forms of gaze and pleasures. These will continue to pose new moral panics and debates in societies and communities, particularly in terms of protecting the vulnerable and the young while catering to a well-established market that has capitalised on e-commerce and new forms of voyeurism.

REFERENCES

Akdeniz, A. (2002). UK government and the control of Internet content. *Computer Law and Security Report*, *17*(5), 303-318. Retrieved 20/08/2007, from http://www.cyber-rights.org/documents/clsr17_5_01.pdf

Bakardjieva, M., & Feenberg, A. (2000). Involving the virtual subject. *Ethics and Information Technology*, *2*, 233-240.

Cavanagh, A. (1999). *Behaviour in public: Ethics in online ethnography.* Retrieved 12/01/07, from http://www.socio.demon.co.uk/6/cavanagh.html

Check, J. (1985). *The effects of violent and non-violent pornography*. Ottawa: Department of Justice, Canada.

Cooper, A., & Griffin-Shelley, E. (2002). A Quick Tour of On-Line Sexuality: Part 1. *Annals of the American Psychotherapy Association*, 5, 11-13.

Cronin, B & Davenport, E. (2001). E-rogenous zones: Positing pornography in the digital economy. *The Information Society*, *17*, 33-48.

Fisher, W., & Barak, A. (2001). Internet pornography: A social psychological perspective on Internet sexuality. *Journal of Sex Research, 38*, 313-323.

Gillespie, A. A. (2005). Indecent images of children: The ever-changing law. *Child Abuse Review, 14*, 430-443.

Gossett, J. L., & Byrne, S. (2002). 'Click here' A content analysis of Internet rape aites. *Gender & Society, 16*(5),

689-709.

Haraway, D. (1997). *Modest_Witness@Second Millenium. FemaleMan_MeetsOnco-MouseTM: Feminism and Technoscience*. Routlege: London.

Jacobs, K. (2004). Pornography in small places and other spaces. *Cultural Studies*, *18*(1), 67-83.

Jenkins, P. (2001). *Beyond tolerance: Child pornography on the Internet*. New York: New York University Press.

Kuipiers, G. (2006). The social construction of digital danger: Debating, defusing and inflating the moral dangers of online humour, pornography in the Netherlands and the US. *New Media and Society, 13*(3), 379-400.

Levine, J. (2002). *Harmful to minors: The perils of protecting children from sex.* Minneapolis, MN: University of Minnesota Press.

Levy, N. (2002). Virtual child pornography: The eroticization of inequality. *Ethics and Information Technology, 4*, 319-323.

Oswell, D. (1999). The dark side of cyberspace: Internet content regulation and child protection. *Convergence, 5*(4), 42-62.

Oswell, D. (2006). When images matter; Internet child pornography, forms of observation and an ethics of the virtual. *Information, Communication and Society, 9*(2), 244-265.

O'Toole, L. (1998). *Pornocopia: Porn, sex. Technology and desire*. London: Serpent's Tail.

Russell, D. (1998). *Dangerous relationships: Pornography, misogyny, and rape*. Thousand Oaks, CA: Sage.

Spencer, J. (1999). Crime on the Internet: Its presentation and representation. *The Howard Journal, 38*(3), 241-251.

Stack, S., Wasserman, I., & Kern, R. (2004). Adult social bonds and use of Internet pornography. *Social Science Quarterly*, *85*(1), 75-89.

Tate, T. (1990). *Child pornography: An investigation*. London: Methuen.

Taylor, M., Holland, G., & Quayle, E. (2001). Typology of paedophile picture collections. *Police Journal, 74*(2), 97-107.

Whitehead, B. (2005). Online porn: How do we keep it from our kids? *Commonweal*, *132*, 18.

Williams, N. (1991). *False images: Telling the truth about pornography*. London: Kingsway Publications.

KEY TERMS

Cyber Porn: Sexually explicit material that is available on the Internet

Digital Image: A visual content constructed through pixels which can be altered or manipulated through technology.

Ethics: The code of conduct in a society or community that may be tacit or explicitly expounded

Pornography: Sexually explicit material that may be available in any medium

Regulations: Formal rules and legislation that are enacted to address a particular issue.

Social Learning Aspects of Knowledge Management

Irena Ali
Department of Defence, Australia

Leoni Warne
Department of Defence, Australia

Celina Pascoe
Department of Defence, Australia

INTRODUCTION

There are probably as many variations of knowledge management definitions as there are practitioners and researchers in the discipline. Complete consensus in such a group would be a surprising finding. This is because the two words are loaded with pre-existing meanings that do not always sit comfortably in juxtaposition, so what it means to "manage knowledge" is difficult to ascertain, and hence comes to mean different things to different people.

We do know however, that knowledge exists in the minds of individuals and is generated and shaped through interaction with others. In an organizational setting, knowledge management must, *at the very least*, be about how knowledge is acquired, constructed, transferred, and otherwise shared with other members of the organization, in a way that seeks to achieve the organization's objectives. Put another way, knowledge management seeks to harness the power of individuals by supporting them with information technologies and other tools, with the broad aim of enhancing the *learning capability* of individuals, groups, and in turn, organizations (Ali, Warne, Bopping, Hart, & Pascoe, 2004). Social learning, in this context, is defined as learning occurring in or by a cultural cluster or organizational group or team and includes procedures for transmitting knowledge and practices across different work situations, settings, and time. However, the application of technology must be guided by the needs of the organization and its workers. As Davenport (2005, p.162) states, "While I don't question the importance of technology in organizations today, it's only one source of knowledge and learning for knowledge workers."

BACKGROUND

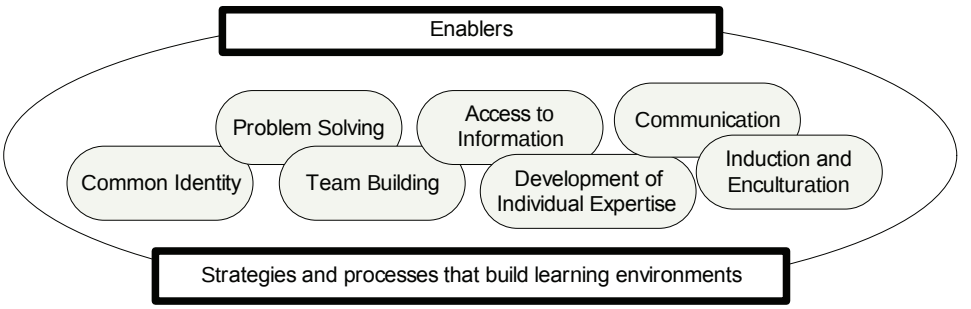
In this article, we examine both theoretical and practical socio-cultural aspects of knowledge management based on years of research by the authors in a large and diverse organization. The study involved numerous functional settings of the organization and the researchers used qualitative and quantitative methodology to gather data. Elements required to build an organizational culture that supports knowledge management are discussed. Unless otherwise specified, words in double quotes in the text are direct quotes from personnel in research settings.

ENABLERS OF SOCIAL LEARNING

The research team identified seven basic categories that constitute enabling processes and strategies to facilitate social learning: common identity, problem solving, team building, access to information, development of individual expertise, communication, and induction and enculturation (see Figure 1).

- **Common identity:** A common ground/understanding to which many people/groups can subscribe, and requires a shift from seeing oneself as separate to seeing oneself as connected to and part of an organization unit. Based on our research, motivators impacting on *common identity* are: goal alignment, cultural identity, gendered identity, language, morale, and workplace design (spatial and physical design).
  - Doney et al. (1998) discuss the relationship between goal alignment and group cohesiveness, claiming that the extent of group cohesiveness relies on the extent to which a team's goals are clear and accepted and also on the degree to which all members adopt team behaviors.
  - The term cultural identity refers to member's sense of self in relation to the specific "tribe" and "tradition" to which they belong and how this distinctiveness applies in their workplace. Cultural identity is another important motivator for social learning because, like common identity, it impacts on the extent to which staff feel that they are part of the system or alienated from it.

Figure 1. Constructs enabling social learning



- Gendered identity relates specifically to one's sense of self, which is imbued with the social, cultural and historical constructions surrounding femininity and masculinity. Gender identity, because of its relationship with common identity, was also seen to impact on social learning (Agostino, 1998).
- Language is another important factor fundamental to the overall social learning processes. By reflecting the social and political relationship between various members, language can impact on common identity. Language is also important in terms of creating a shared understanding among workers and their relationship to the wider organization. "Words are bullets. Never, never use imprecise language." Thus learning the specific work related language is of central importance to broader social learning development, and is an important outcome of the enculturation process.
- Morale has been a significant focus in the overall study because the research team found evidence of low morale being coupled with higher levels of alienation towards senior management. Such alienation has obvious implications for the broader understanding of a common identity and thus for social learning.
- Workplace design and proximity also threatens common identity when staff are not working in the same location. "[Building X] and us. We don't see them. There is not any spirit that we are belonging to one branch. I have more to do with [a specific area] than anything else and I've made some good contacts in there... who I sit around with."

- **Problem solving:** A core activity. It fosters social learning, because each problem represents an opportunity to generate knowledge. Motivators associated with this enabler are: networking, perceptions of the organization, systemic understanding, and time for inquiry and reflection.
  - An individual's personal and social networks are an important means of acquiring, propagating, and sharing knowledge. As Davenport and Prusack (1998) claim, when those who are in a position of "know-how" share their expertise, they contribute to problem solving. Personal networks were seen to function as channels supporting both "information pull" and "information push." Atkinson and Moffat (2005) state that sharing of information is based on trust developed through social interaction, shared values, and beliefs. A human is a node in such interactions and a link is a bond that people develop which is based on mutual trust. Therefore, a significant component of a person's information environment consists of the relationships he or she can tap into for various informational needs.
  - Individual and shared perceptions of the organization, and how they operate, provide an essential backdrop to problem solving within an organizational context. These perceptions may consist of deeply ingrained assumptions, generalizations, or even pictures or images that influence how people understand their organizational world and how they should act within it (Senge, 1992). The importance of these perceptions cannot be stressed enough, because they directly influence the construction of individuals' knowledge and understandings that they draw upon in their day-to-day-activities.
  - Effective problem solving often requires a systemic understanding of organizational and inter-organizational issues. Systemic understanding requires a holistic view of an organization and its inter-relationships, an understanding of the fabric of relationships and the likely effect of interrelated actions (Davenport, 2005; Senge, 1992).

- Inquiry and reflection together are a powerful means of enhancing social learning and knowledge creation. Inquiries, or questions, are triggered by problems that require solutions or explanation. Reflection allows time for examination, contemplation and, often, resolution of the inquiries. To use a common metaphor, it is perhaps the best means for distinguishing between the forests and the trees of everyday working life.

- **Team building:** Working together and understanding what each member is trying to do. Team building was seen to be essential for effective social learning and problem solving. As team-members got to know each other they become aware of each other's strengths and weaknesses, what they could or could not do, their expertise and experience. Motivators associated are: leadership, team-based morale, performance management, public recognition and reward systems, use of humour, and workplace design (Warne, Ali, & Pascoe, et al., 2003).
  - In general, the caliber of leadership within the settings studied was to be admired. The leaders and managers were innovative and they motivated and developed their staff, mainly by demonstrating that staff are highly valued and by acknowledging expertise and knowledge regardless of their pay or position. Another team building issue that emerged was that people were appreciative of informal "drop ins" by senior managers inquiring how they were doing. This "roving management" was said to contribute to better cohesion of teams, to promote system thinking, to help to focus on overall goals, and to facilitate communication and feedback.
  - "Team spirit" and "team cohesiveness" are both important values within the work culture and work ethic, nonetheless, there was nothing uniform about this in the settings studied. Some teams did not see the significance of their particular tasks to the overall goals of the organization. However, good examples of teamwork and team spirit were also evident. There were instances where teamwork was well integrated into daily work and where people worked collaboratively. Such teams were goal oriented and were not only teams in structure but in spirit and were led by a leader who saw his/her role as serving team members rather than just having the position of a leader (Warne et al., 2003).
  - For many employees, the performance cycle is annual and the outcome of a performance report often determines the prospects of one's career progression. Some felt somewhat uneasy as their performance evaluation was due relatively early into their posting cycle. A well planned performance appraisal system should help to make equitable and unbiased decisions regarding staff selection, placement, development, and training (Wood, 1989). Researchers were told that there was often a lack of clear communication about performance expectations. Also, an annual performance appraisal appears to be too long to wait for recognition of good work and too late to correct a performance problem. Morgan (1989) and Wood (1989) explain that to maximize positive results, the appraisal process should be two ways, it should facilitate and coach staff in doing their jobs effectively, and it should be frequent and informal.
  - It was observed that humor was used for smoothing discussions that were becoming heated and to stop the conflict from escalating whilst also enabling the conflicting subordinates to save face. At meetings, humor was used to assist in uniting people around common themes and to make criticism palatable.
  - One way of and increasing team and individual morale is to publicly acknowledge outstanding work. Making employees feel appreciated, and saying, "Thank you, we know that you are a good employee, we value you and your work," is a big factor in motivation (Mitchell, 2000). Key informants stated that public recognition of good work was scarce and that a written or verbal word of praise, a pat on the back often means more, for example, than a pay raise—"praise is better than money" and praise is needed at all levels.
  - Workplace design was seen to have impact on social learning. Staff located at small isolated outposts were at risk of feeling isolated and did not identify strongly with the parent organization. As stated earlier, outposted staff identified more with the workplace with which they were based than their Branch where they affiliated. This was further exacerbated by the fact that they often felt excluded by their colleagues.

- **Access to information:** The easy availability of corporate information in whatever format was observed to effect knowledge acquisition and generation of new knowledge and social learning. Motivators associated are: record keeping, networking, meetings, and information technology (IT) infrastructure.
  - The researchers observed that general familiarity with records keeping procedures was quite poor. Some people have developed their own

personal records keeping systems but there was little uniformity in these and no adherence to file naming conventions and standards. As some informants stated: "I believe that physical files in the … are no longer managed well because their management has been farmed out to outside bodies." or "I think we have problems with passing on information in the organization as a whole. We just don't do it very well." The issue of electronic records, particularly e-mail messages containing evidence of business transactions, posed problems not only in the setting studied but also in the ADO at large.

- Personal networks from previous postings as well as newly acquired contacts in the new environment play a vital role in knowledge construction and acquisition. New knowledge often begins with the individual and through conversations people discover what they know, what others know and in the process of sharing, new knowledge is created. Knowledge sharing depends on the quality of conversations, formal or informal, that people have. Sharing of information has a behavioural component and the emphasis is usually on one-to-one networking initiative and effort. It requires time and space (physical, cognitive and social) to develop the sense of safety and trust that is needed for information sharing. Webber (1993) aptly describes it "conversations—not rank, title, or the trappings of power—determine who is literally and figuratively 'in the loop' and who is not."
- Meetings are another means of accessing information and those that were observed varied significantly in format and the protocols in place. At the tactical headquarters, meetings that were mission related provided excellent opportunities for learning. Strict protocols were observed at these briefings (e.g., allowing participants to discuss errors or problems encountered during missions without assigning blame or shame to individuals). There were few equivalent meetings at the strategic headquarters, other than some induction sessions and briefings and it appeared that learning how to do one's job was not given quite the same priority.
- The researchers observed that information access due to failings in the IT infrastructure inhibited access to information within the strategic settings. Another issue that caused problems was the difficulty in finding information on the shared drive. Since there was no specific person responsible for maintaining the shared drive and for naming folders, it was left to the discretion of the document originator where information would be stored.

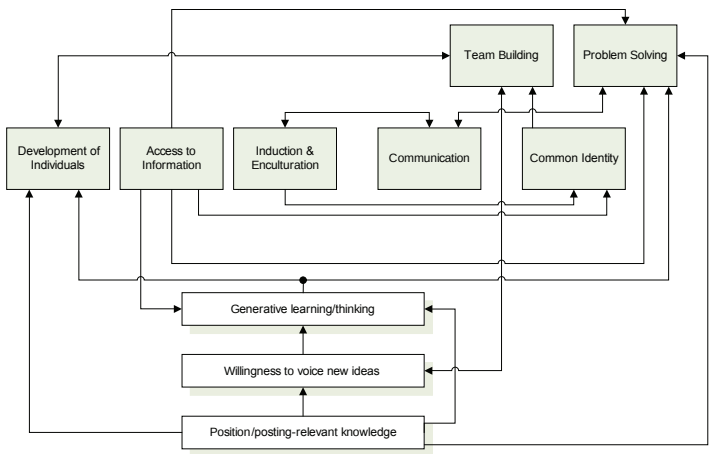
- **Development of individual expertise:** The acquisition and development of expertise was seen as an integral part of social learning. Motivators associated with this enabler are: career trajectories, professional currency, professional training, postings and promotion, and mentoring.
  - A career trajectory describes the positions, roles, and experience that individuals have accumulated, up to and including the position they currently hold. While not excluding personal experiences outside of a work or training context, a well designed career trajectory generally equips an individual with the skills, experience, maturity, and personal networks needed to successfully fill a particular posting.
  - The term professional currency has a somewhat different meaning within different environments. However, professional currency promotes social learning in the same way that appropriate career trajectories do so—by providing a foundation for the generation of new knowledge.
  - Appropriate professional training is a significant component of the development of individual expertise and, therefore, a fundamental for generating new knowledge. Training courses are important to furthering individuals' expertise, as well as for forming the personal networks that subsequently develop. However, in times of budgetary constraints, training money is often the first to go, with damaging consequences for the organization's ability to learn and manage their knowledge.
  - Mentoring is regarded as an effective method of assisting the development of individual expertise, especially for junior staff a degree of informal mentoring was seen to be built into elements of the training program in some of the settings studied. In terms of developing a career trajectory, the knowledge acquired through mentoring may also be important when individuals want to prepare themselves for specific roles in the future.
- **Communication:** Essential to effective learning within an organization and to effective social learning. Motivators associated with this enabler are: overall communication climate, formal and informal information flows, time for inquiry and reflection, use of humor, language, and workplace design.
  - Supportive communication climates are being positively linked to open and free exchange of information and constructive conflict management. Characteristics of a supportive commu-

nication climate include a culture of sharing knowledge, treating each other with respect, and generally behaving in a cooperative manner. Research has established the link between supportive organizational communication climates and generative learning (Bokeno, 2000; Ruppel, 2000) and with higher levels of organizational commitment (Guzley, 1992).

- An important element of generative learning is for organizational members to be able to engage in dialogue, which is open and is based on inquiry and reflection. A supportive communication climate is a prerequisite for such dialogue and it requires learning how to recognize defensive patterns of interaction that undermine learning (Senge, 1992).
- The issue of workplace design and its impact on teams, network building, and on accessing information arose repeatedly during the study. Physical location and proximity to each other had the potential to promote the transfer of pertinent knowledge. The point was made that in addition to more quickly obtaining answers to questions about particular tasks, an open plan workplace enabled one to tap into pertinent knowledge by overhearing others' conversations. Hutchins (1996) uses the term "horizon of observation" to describe the area of the task environment, which can be seen, and is therefore available as a context for learning by team members.

- **Induction and enculturation:** Facilitates social learning by providing a foundation upon which an individual can become fully productive. Issues associated with this enabler are: timeliness and comprehensiveness of the process, buddy/mentoring system, handovers and information packages, and training.
  - Good induction is more that just an introduction to new job and workmates; it is a way of helping people find their feet. Attitudes and expectations are shaped during the early days of new employment and work satisfaction is linked to well timed and conducted work orientation (Dunford, 1992; George & Cole, 1992). The interviews clearly indicated a relationship between meaningful and timely induction and subsequent job satisfaction. An interesting finding was that those who were not properly inducted or enculturated into the organization saw no need and responsibility to actually prepare any form of handover for anyone who may take over their position in the future.
  - Although highly desirable, it was not always feasible to conduct an induction program at the beginning of a new posting cycle. In the interim, a "buddy" or "mentoring" system could fill in the gap. A "buddy" would be an experienced workmate who could be available to answer questions and assist the orientation of new members during the initial few weeks. Some interviewees said that having a buddy when they started was invaluable to settling into a new job and to effective learning.
  - The researchers were repeatedly told that early training is an important part of effective induction and enculturation. It is an opportunity to learn the explicit knowledge that is taught as part of formal training. It is also an opportunity to be exposed to the attitude and cultural perceptions of colleagues and peers.

Figure 2. Enabling processes and their impact on social learning



These factors enabling social learning identified from our data are by no means exhaustive, however, based on the available data the research team could see a relationship between these enablers and social and generative learning. Figure 2 depicts these relationships and their impact on social learning.

FUTURE TRENDS

Whether by design or necessity, humans tend to collaborate to achieve set goals. In fact, this sharing of information and knowledge, and the willingness to cooperate, are key elements for learning, innovation and advancement in general. The progress and proliferation of information technology greatly facilitate this process. However, this widespread application of information technology and emphasis on sophisticated networks for information sharing and social learning leads to a false assumption that once all networks are in place the information will be shared and freely disseminated. The subtle difference between "the network" and "to network" is the key. "The network" is a noun, the information technology, and can only be the enabler. "To network" is the verb, the human behaviour, the action, and the main focus. Therefore, the future trends in the area of social and organisational learning must look beyond the acquisition of technical enablers to individual and organizational behaviour (e.g., organizational structure, processes, and tactics) in order to shift emphasis from a technology-centric approach to a people-centric capability, ensuring that people will get the systems they need and want.

CONCLUSION

Organizations seeking to improve information sharing and knowledge generation need to develop a greater awareness of the processes and strategies of organizational learning. Organizational knowledge is distributed across functional groups and its generation and continual existence is dependant on the overall communication climate which is embedded in the organizational culture. This study indicates that information sharing and subsequent knowledge generation would be successful when interactive environments are cultivated before other (e.g., technology-based solutions are implemented). Therefore, the communication strategy in any organization must take into account the role played by informal and personal networks and trust in information sharing to optimise the process of transferring critical data to facilitate speedier decision-making. Technology should only be designed and applied after a thorough investigation of the work practices and work preferences of the people and teams in the organization.

REFERENCES

Agostino, K. (1998). The making of warriors: Men, identity, and military culture. *JIGS: Australian Masculinities, 3*(2).

Ali, I., Pascoe, C., & Warne, L. (2002). Interactions of organizational culture and collaboration in working and learning. *Educational Technology and Society, 5*(2), 60-69.

Ali, I., Warne, L., Bopping, D., Hart, D., & Pascoe, C. (2004). Organisational paradigms and network centric organisations. *Journal of Issues in Informing Science and Science and Information Technology, 1,* 1089-1096.

Argyris, C. (1973). *On organisations of the future*. Beverly Hills, CA: Sage.

Atkinson, S. R., & Moffat, J. (2005). *The agile organization*. Washington, DC: CCRP Publication Series.

Bokeno, R. M. (2000). Dialogic mentoring. *Management Communication Quarterly, 14*, 237-270.

Cooke, R. (1998). Welcome aboard. *Credit Union Management, 21*(7), 46-47.

Davenport, T. H. (2005). *Thinking for a living: How to get better performance and results from knowledge workers.* Boston: Harvard Business School Press.

Davenport, T. H., & Prusack, L. (1998). *Working knowledge: How organisations manage what they know.* Harvard Business School Press.

Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the influence of national culture on the development of trust. *Academy of Management Review, 23*(3), 601-623.

Drucker, P. F. (1999). Beyond the information revolution. *The Atlantic Monthly, 284*(4), 47-57.

Dunford, R. W. (1992). *Organisational behaviour: An organisational analysis perspective*. Sydney: Addison Wesley.

Enneking, N. E. (1998). Managing email: Working toward an effective solution. *Records Management Quarterly, 32*(3), 24-43.

Ganzel, R. (1998). Elements of a great orientation. *Training, 35*(3), 56.

George, C. S., & Cole, K. (1992). *Supervision in action: The art of managing*. Sydney: Prentice Hall.

Guzley, R. M. (1992). Organizational climate and communication climate: Predictors of commitment to the organization. *Management Communication Quarterly, 5*(4), 379-402.

Hutchins, E. (1996). *Cognition in the wild*. Cambridge, MA: MIT Press.

Mitchell, S. (2000). *Be bold and discover the power of praise*. East Roseville: Simon & Schuster.

Morgan, T. (1989). *Performance management—The missing link in strategic management and planning.* In D. C. Corbett (Ed.), *Public sector policies for the 1990s* (pp. 243-250). Melbourne: Public Sector Management Institute, Faculty of Economics and Politics, Monash University.

Ruppel, P. C. (2000). The relationship of communication, ethical work climate, and trust to commitment and innovation. *Journal of Business Ethics, 25*, 313-328.

Senge, P. M. (1992). *The fifth discipline*: *The art & practice of the learning organisation.* Australia: Random House.

Warne, L., Agostino, K., Ali, I., Pascoe, C., Bopping, D. (2002). The knowledge edge: Knowledge management and social learning in military settings. In A. G. O. Khalil & S. M. Rahman (Eds), *Knowledge and information technology management: Human and social perspectives* (pp. 324-353). Hershey, PA: Idea Group Publishing.

Warne, L., Ali, I., & Pascoe, C. (2003). Team building as a foundation for knowledge management: findings from research into social learning in the Australian defense organisation. *Journal of Information & Knowledge Management, 2*(2), 93-170.

Warne, L., Hasan, H., & Ali, I. (2005). Transforming organizational culture to the ideal inquiring organization: Hopes and hurdles. In J. F. Courtney, J. D. Haynes, & D. B. Paradice (Eds.), *Inquiring organizations: Moving from knowledge management to wisdom* (pp. 316-336). Hershey, PA: Idea Group Publishing.

Webber, A. M. (1993). What's so new about the new economy? *Harvard Business Review,* 24-42, Jan-Feb.

Wood, R. (1989). Performance appraisal in the reform of public sector management practices. In D. C. Corbett (Ed.), *Public sector policies for the 1990's* (pp. 225-242). Melbourne: Public Sector Management Institute, Faculty of Economics and Politics, Monash University.

KEY TERMS

Career Trajectory: Describes the positions, roles, and experience that individuals have accumulated, up to and including the position they currently hold.

Common Identity: A common ground/understanding to which many people/groups can subscribe, and requires a shift from seeing oneself as separate to seeing oneself as connected to and part of an organizational unit.

Communication Climate: Extend to which there is an open and free exchange of information, transparency of decision-making, and how constructively conflict is managed.

Knowledge: An understanding gained through experience or learning: the sum, or a subset, of what has been perceived and discovered by an individual. Knowledge exists in the minds of individuals and is generated and shaped through interaction with others.

Knowledge Management: In an organizational setting, it must, *at the very least*, be about how knowledge is acquired, constructed, transferred, and otherwise shared with other members of the organization, in a way that seeks to achieve the organization's objectives.

Social Learning: Learning occurring in or by a cultural cluster and includes procedures for transmitting knowledge and practices across different work situations/settings and time.

Systemic Understanding: A holistic view of an organization and its inter-relationships, an understanding of the fabric of relationships and the likely effect of interrelated actions.

Socio-Cognitive Model of Trust

Rino Falcone
Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy

Cristiano Castelfranchi
Institute for Cognitive Sciences and Technology, National Research Council of Italy, Italy

INTRODUCTION

Humans have learned to cooperate in many ways and in many environments, on different tasks, and for achieving different and several goals. Collaboration and cooperation in their more general sense (and, in particular, negotiation, exchange, help, delegation, adoption, and so on) are important characteristics - or better, the most foundational aspects - of human societies (Tuomela, 1995).

In the evolution of cooperative models, a fundamental role has been played by diverse constructs of various kinds (purely interactional, technical-legal, organizational, socio-cognitive, etc.), opportunely introduced (or spontaneously emerged) to support decision making in collaborative situations.

The new scenarios we are destined to meet in the third millennium transfigure the old frame of reference, in that we have to consider new channels and infrastructures (i.e., the Internet), new artificial entities for cooperating with artificial or software agents, and new modalities of interaction (suggested/imposed by both the new channels and the new entities). In fact, it is changing the identification of the potential partners, the perception of the other agents, the space-temporal context in which interaction happen, the nature of the interaction traces, the kind and role of the authorities and guarantees, etc.

For coping with these scenarios, it will be necessary to update the traditional supporting decision-making constructs. This effort will be necessary especially to develop the new cyber-societies in such a way as not to miss some of the important cooperative characteristics that are so relevant in human societies.

BACKGROUND

Trust (Ganbetta, 1990; Luhmann, 1990; Dasgupta, 1990), in the general frame described above, might be considered as a socio-cognitive construct of main importance. In particular, trust building is always more recognized as a key factor for using and developing the new interactional paradigm.

Trust should not be made indistinct with security. The latter can be useful to protect – in the electronic domain - from the intrusiveness of an unknown agent, to guarantee an agent in the identification of its partner, to identify the sender of a message (for example, by verifying the origin of a received message; by verifying that a received message has not been modified in transit; by preventing that an agent who sent a message might be able to deny later that it sent the message [He, Sycara & Su, 2001]). With sophisticated cryptographic techniques, it is possible to give some solution to these security problems.

However, more complex is the issue of trust, that must give us tools for acting in a world that is in principle insecure (that cannot be considered 100% secure), where we have to make the decision to rely on someone in risky situations. (Consider the variety of cases in which it is necessary or useful to interact with agents whose identity, history or relationships are unknown, and/or it is only possible to make uncertain predictions on their future behaviors.)

Trust should not be made indistinct with reputation, too. In fact, communicated reputation (Conte & Paolucci, 2002) is one of the possible sources on which the trustier bases its decision to trust or not.

The more actual and important example of the usefulness of trust building is electronic commerce, but we must also consider other important domains of Multi Agent Systems and Agent Theory such as Agent Modeling, Human-Computer Interaction, Computer Supported Cooperative Work, Mixed Initiative and Adjustable Autonomy, Pervasive and Ubiquitous Computing. In fact, today many computer applications are open distributed systems (with many autonomous components that are spread throughout a network and interacting with each other). Given the impossibility to rule this kind of system by a centralized control regime (Marsh, 1994), it becomes essential to introduce local tools in order to choose the right partnership and at the same time reduce the uncertainty (deriving from the nature of an open distributed system) associated with that choice.

TRUST IN THE NEW TECHNOLOGICAL SCENARIOS

In fact, various different kinds of trust should be modeled, designed, and implemented:

- Trust in the environment and in the infrastructure (the socio-technical system)
- Trust in personal agents and in mediating agents
- Trust in potential partners
- Trust in sources
- Trust in warrantors and authorities.

Part of these different kinds of trust have a complementary relation with each other, that is, the final trust in a given system/process can be the result of various trust attributions to the different components. An exemplary case is one's trust in an agent that must achieve a task (and more specifically in its capabilities for realizing that task) as different from one's trust in the environment (hostile versus friendly) where that agent operates, or again as different from one's trust in a possible third party (arbitrator, mediator, normative systems, conventions, etc.) able to influence/constrain the trustee and representing a guaranty for the trustier (Castelfranchi & Falcone, 1998; Falcone & Castelfranchi, 2001).

Therefore, the "sufficient" trust value of one single component cannot be established before evaluating the value of the other components. In this regard, it is very interesting to characterize the relationships between trust and (partial) control (Castelfranchi & Falcone, 2000).

It is important to underline how trust is in general oriented towards not directly observable properties. It is, in fact, based on the ability to predict these properties and to rely or not to rely on them. Thus, it is quite complex to assess the real trustworthiness of an agent/system/process, not only because - as we have seen - there are many different components that contribute to this trustworthiness, but also because the latter is not directly observable (see [Bacharach & Gambetta, 2001] about signs of trust). The important thing is the perceived trustworthiness that is, in its turn, the result of different modalities of the trustier's reasoning about direct experience; categorization; inference, and communicated reputation.

SOCIO-COGNITIVE MODEL OF TRUST

The Socio-Cognitive model of trust is based on a portrait of the mental state of trust in cognitive terms (beliefs, goals). This is not a complete account of the psychological dimensions of trust. It represents the most explicit (reason-based) and conscious form. The model does not account for the more implicit forms of trust (for example, trust by default, not based upon explicit evaluations, beliefs, derived from previous experience or other sources) or for the affective dimensions of trust, based not on explicit evaluations but on emotional responses and an intuitive, unconscious appraisal (Thagard, 1998).

The word *trust* means different things, but they are systematically related with each other. In particular, three crucial concepts have been recognized and distinguished not only in natural language but also in the scientific literature. Trust is at the same time:

- A mere *mental attitude* (prediction and evaluation) toward another agent, a simple *disposition*;
- A *decision* to rely upon the other, i.e., an *intention* to delegate and to trust, which makes the trustier "vulnerable"(Mayer, Davis, & Schoorman, 1995);
- A *behavior*, i.e., the intentional *act* of trusting, and the consequent *relation* between the trustier and the trustee.

In each of the above concepts, different sets of cognitive ingredients are involved in the trustier's mind. The model is based on the BDI (Belief-Desire-Intention) approach for modeling mind that is inspired by Bratman's philosophical model (Bratman, 1987). First of all, in the trust model only an agent endowed with both goals and beliefs can "trust" another agent. Let us consider the trust of an agent *X* towards another agent *Y* about the (*Y*'s) behavior/action α relevant for the result (goal) *g* when:

- *X is the (relying) agent, who feels trust; it is a cognitive agent endowed with* internal explicit goals and beliefs (the *trustier*)
- *Y* is the agent or entity that is trusted (the *trustee*)
- *X* trusts *Y* about g/α and for g/α.

In the model *Y* is not necessarily a cognitive agent (for instance, an agent can - or cannot - trust a chair as far as to sustain his weight when he is seated on it). On the contrary, *X* must always be a cognitive agent: so, in the case of artificial agents we should be able to simulate these internal explicit goals and beliefs.

For all the three notions of trust defined above (*trust disposition*, *decision to trust*, and *trusting behavior*), we claim that someone trusts some other one only relatively to some goal (here the goal is intended as the general, basic teleonomic notion, any motivational representation in the agent: desires, motives, will, needs, objectives, duties, utopias, are kinds of goals). An unconcerned agent does not really "trust": he just has opinions and forecasts. Second, trust itself *consists of* beliefs.

Since *Y*'s action is useful to *X* (trust disposition), and *X* has decided to rely on it (decision to trust), this means that *X*

might delegate (act of trusting) some action/goal in his own plan to *Y*. This is the strict relation between trust disposition, decision to trust, and delegation.

The model includes two main basic beliefs (we are considering the trustee as a cognitive agent, too):

- *Competence Belief*: a *sufficient evaluation* of *Y*'s abilities is necessary. *X* should believe that *Y* is useful for this goal, that *Y* can produce/provide the expected result, and that *Y* can play such a role in *X*'s plan/action.
- *Willingness Belief*: *X* should think that *Y* not only is able and can do that action/task, but *Y* actually will do what *X* needs (under given circumstances). This belief makes the trustee's behavior predictable.

Another important basic belief for trust is:

- *Dependence Belief*: *X* believes -to trust *Y* and delegate to it- that either *X* needs it, *X* depends on it (*strong dependence*), or at least that it is better to *X* to rely on it, rather than not to rely on it (*weak dependence*). In other terms, when *X* trusts someone, *X* is in *a strategic situation*: *X* believes that there is interference and that his rewards, the results of his project, depend on the actions of another agent *Y*.

Obviously, the willingness belief hides a set of other beliefs on the trustee's reasons and motives for helping. In particular, *X* believes that *Y* has some motives for helping it (for adopting its goal), and that these motives will probably prevail -in case of conflict- on other motives, negative for it. Notice that motives inducing adoption are of several different kinds: from friendship to altruism, from morality to fear of sanctions, from exchange to common goal (cooperation), and so on. This is why, for example, it is important to have common culture, shared values, and the same acknowledged authorities between trustier and trustee.

Another important characteristic of the socio-cognitive model of trust is the distinction between trust "in" someone or something that has to act and produce a given performance thanks to its *internal characteristics*, and the global trust in the global event or process and its result, which is also affected by *external factors* like opportunities and interferences.

Trust in *Y* (for example, "social trust" in strict sense) seems to consists in the first two prototypical beliefs/evaluations identified as the basis for reliance: *ability/competence* (that with cognitive agents includes knowledge and self-confidence), and *disposition* (that with cognitive agents is based willingness, persistence, engagement, etc.). Evaluation about external opportunities is not really an evaluation about *Y* (at most the belief about its ability to recognize, exploit and create opportunities is part of our trust in *Y*). We should also add an evaluation about the probability and consistence of obstacles, adversities, and interferences.

Trust can be said to consist of or better to (either implicitly or explicitly) imply the *subjective probability* of the successful performance of a given behavior a, and it is on the basis of this subjective perception/evaluation of risk and opportunity that the agent decides to rely or not to rely on *Y*. However, the probability index is based on, and derives from those beliefs and evaluations. In other terms, the global, final probability of the realization of the goal *g*, i.e., of the successful performance of a, should be decomposed into the probability of *Y* performing the action well (*internal attribution*) and the probability of having the appropriate conditions (*external attribution*) for the performance and for its success, and of not having interferences and adversities (*external attribution*). This decomposition is important because:

a) The trustier's decision might be different with the same global probability or risk, depending on its composition (for example, for personality factors);
b) Trust composition (internal vs. external) produces completely different intervention strategies: to manipulate the external variables (circumstances, infrastructures) is completely different from manipulating internal parameters.

The idea that trust is gradable is usual (in common sense, in social sciences, in Artificial Intelligence). However, since no real definition and cognitive characterization of trust is given, the quantification of trust is quite *ad hoc* and arbitrary, and the introduction of this notion or predicate is semantically empty. On the contrary, in the socio-cognitive model of trust there is a strong coherence between the cognitive definition of trust, its mental ingredients, and, on the one side, its value, on the other side, its social functions. More precisely the latter are based on the former.

A degree of trust of *X* in *Y* is grounded on the cognitive components of *X*'s mental state of trust. More precisely *the degree of trust is a function of the subjective certainty of the pertinent beliefs*. The degree of trust is used to formalize a rational basis for the decision of relying and betting on *Y*. A "quantitative" aspect of another basic ingredient is relevant: the value or importance or utility of the goal (g). In sum, *the quantitative dimensions of trust are based on the quantitative dimensions of its cognitive constituents*.

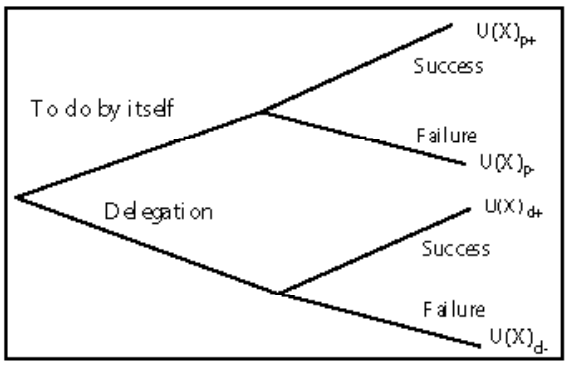
If we call $DoTXY\tau$ the degree of trust of an agent *X* about *Y* on the task $\tau=(\alpha,g)$ we have:

- $DoTXY\tau = DoCX[OppY(\alpha,g)] * DoCX[AbilityY(\alpha)] * DoCX[WillDoY(\alpha,g)]$

Where:

- $DoCX[OppY(\alpha,g)]$, is the degree of credibility of *X's* beliefs about *Y's* opportunity of performing α to realize *g;*

Figure 1.



- *DoCX[AbilityY(α)]*, the degree of credibility of *X's* beliefs about *Y's* ability/competence to perform α*;*
- *DoCX[WillDoY(α,g)]*, the degree of credibility of *X's* beliefs about *Y's* actual performance*;*
- *DoCX[WillDoY(α,g)] = DoCX[IntendY(α,g)] * DoCX[PersistY(α,g)]*
  (Given that *Y* is a cognitive agent)

In any circumstance, an agent *X* endowed with a given goal, has three main choices:

i) To try to achieve the goal by itself;
ii) To delegate the achievement of that goal to another agent, *Y*;
iii) To do nothing (relatively to this goal), renouncing.

Considering the simplified scenario in which only (i) and (ii) are the possible choices we have the Figure1.

Where if *U(X)* is the agent *X*'s utility function, more specifically:

- ***U(X)p+***, the utility of the *X*'s success performance;
- ***U(X)p-***, the utility of the *X*'s failure performance;
- ***U(X)d+***, the utility of a successful delegation (the utility due to the success of the delegated action);
- ***U(X)d-***, the utility of a failure delegation (the damage due to the failure of the delegated action).

In the previous scenario, in order to delegate we must have:

$$DoTXY\tau * U(X)d^{+} + (1 - DoTXY\tau)\, U(X)d^{-} > DoTXX\tau * U(X)p^{+} + (1 - DoTXX\tau)\, U(X)p^{-}$$

where $DoTXX\tau$ is the *selftrust* of *X* about τ.

More precisely, we have: $U(X)p^{+} = Value(g) + Cost[Performance(X)]$,

$U(X)p^{-} = Cost[Performance(X)] + Additional\ Damage\ for\ failure$

$U(X)d^{+} = Value(g) + Cost[Delegation(X\ Y)]$,

$U(X)d^{-} = Cost[Delegation(X\ Y)] + Additional\ Damage\ for\ failure$

Where it is supposed that it is possible to attribute a quantitative value (importance) to goals and where the costs of the actions (delegation and performance) are supposed to be negative.

FUTURE TRENDS

One of the main aspects that should be analyzed in the next few years is the dynamics of trust and the possibility of introducing all the dynamic aspects in the computational setting in which humans and machines will work together. Trust is a dynamic phenomenon in its intrinsic nature (Falcone & Castelfranchi, 2001). Trust changes with experience, with the modification of the different sources it is based on, with the emotional state of the trustier, with the modification of the environment in which the trustee is supposed to perform, and so on. But, trust is also influenced by trust, itself, in the same specific interaction: for example, how trust creates a reciprocal trust; how the fact that A trusts B can actually increase B's trustworthiness; and so on. In other words, in a computational model of trust relationships we have to consider all the dynamical aspects of the trust phenomenon.

CONCLUSION

The Socio-Cognitive model of trust analyzes the basic elements on which trust is founded in terms of the cognitive ingredients of the trustier. In fact, the richness of the referred model (trust is based on many different beliefs) allows us to distinguish between internal and external attributions (to the trustee) and for each of these two attributions it allows us to distinguish among several other sub-components such as: competence, disposition, un-harmfulness, and so on.

The model introduced a degree of trust instead of a simple probability factor since it permits us to evaluate trustfulness in a rational way.

In other words, if we understand what precisely the basic ingredients of trust are, we would be able to better model and build artificial systems in which this attitude should be present.

REFERENCES

Bacharach, M., & Gambetta, D. (2001). Trust as type detection. In C. Castelfranchi & Y. Tan (Eds.), *Trust and deception in virtual societies*. Dordrecht: Kluwer Academic Publishers.

Bratman, M.E. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Castelfranchi, C., & Falcone, R. (1998). Principles of trust for MAS: cognitive anatomy, social importance, and quantification. Proceedings of the International Conference on Multi-Agent Systems (ICMAS'98).

Castelfranchi, C. & Falcone, R. (2000). Trust and Control: A Dialectic Link. *Applied Artificial Intelligence Journal*, 14(8), 799-823.

Conte, R. & Paolucci, M. (2002). *Reputation in artificial societies. Social beliefs for social order*. Boston, MA: Kluwer Academic Publishers.

Dasgupta, P. (1990). Trust as a commodity. In D. Gambetta (Ed.), *Trust* (pp. 49-72). Oxford: Basil Blackwell.

Falcone, R., & Castelfranchi, C. (2001). The socio-cognitive dynamics of trust: does trust create trust? In R. Falcone, M. Singh, and Y. Tan (Eds.), *Trust in Cyber-societies: Integrating the human and artificial perspectives* (pp. 55-72). LNAI 2246 Springer.

Falcone, R., & Castelfranchi, C. (2001). Social Trust: A Cognitive Approach. In C. Castelfranchi & Y. Tan (Eds.), *Trust and deception in virtual societies.* Dordrecht: Kluwer Academic Publishers.

Gambetta, D. (Ed.) (1990). *Trust*. Oxford: Basil Blackwell.

He, Q., Sycara, K., & Su, Z. (2001). Security infrastructure for software agent society. In C. Castelfranchi & Y. Tan (Eds.), *Trust and deception in virtual societies*. Dordrecht: Kluwer Academic Publishers.

Luhmann, N. (1990). Familiarity, confidence, trust: Problems and alternatives. In D. Gambetta (Ed.), *Trust* (pp. 94-107). Oxford: Basil Blackwell.

Marsh, S. (1994). Formalising trust as a computational concept. Ph.D. thesis, Department of Computing Science, University of Stirling.

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3), 709-734.

Thagard, P. (1998). Emotional Coherence: Trust, Empathy, Nationalism, Weakness of Will, Beauty, Humor, and Cognitive Therapy, Technical Report, University of Waterloo.

Tuomela, R. (1995). The importance of us: a philosophical study of basic social notions. Stanford University Press.

KEY TERMS

Cyber-Societies: The set of natural, artificial, and virtual agents connected and interacting with each others through natural and artificial infrastructures within virtual institutions.

Reputation: The estimated trustworthiness in an agent as derived from the communicated opinions of other parts (directly or indirectly received); the resulting and emergent "common opinion" about the agent's trustworthiness.

Task: An action and/or a goal an agent has to realize as delegated by another agent; thus – in the opposite perspective - the couple action/goal that an agent intentionally delegates to another agent; where at least the delegating agent knows one between the action and the goal.

Trust: The attitude of an agent to delegate a part of its own plan/goal to another agent and rely upon it in a risky situation (possible failure) on the basis of its own beliefs about the other agent and on the environment in which it operates.

Trustee: The trusted agent in the trust relationship.

Trustier: The trusting agent in the trust relationship.

Ubiquitous Computing: The trend of the technological development to integrate into any kind of object information processing and communication capabilities.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2534-2538, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Sociological Insights in Structuring Australian Distance Education

Angela T. Ragusa
Charles Sturt University, Australia

INTRODUCTION

Sociology is well-known for analyzing institutions and social change (Holmes, Hughes, & Julian, 2007). Yet, a dearth of sociological research explores technology and distance education (DE) despite imperatives to include cultural issues (Jorgensen, 2002; Lum, 2006). Meta-analysis shows social studies scholars fail to prioritize technological research (Marri, 2007). Sociologists have examined Web-based instruction and anxiety levels (Gundy, Morton, Liu, & Kline, 2006), flaming (Lee, 2005) and the relationship between learning environment, pedagogy, social roles, relations (Jaffee, 2003) and unintended benefits of traditional classrooms using DE (Edwards, Cordray, & Dorbolo, 2000). This qualitative exploratory research looks at **asynchronous forum** (AF) and DE student experiences in Australia. Using social constructivism, learning is seen as *praxis*, or doing (Vygotsky, 1986) in contrast with ancient traditionalists' *tabula rasa*/"blank slate" understanding of learners waiting to be filled with knowledge (Palloff & Pratt, 2001). Case studies show how culture and learning environments affect **virtual communication** (VC) when *all* communication, student-teacher and student-student, is technologically mediated. Experiences from four 2005-2006 cohorts show social structure affects student perceptions' of learning, satisfaction and agency.

BACKGROUND

Knowledge is an interaction between learner and environment, subsequently reconfiguring both (Semple, 2000). What counts as knowledge is subjective and historically contingent. Advanced capitalistic societies are affected by information technologies (IT) in our "Information Age." In advanced capitalism, ownership and management of IT create global networks and change social interaction (Castells, 2000). This change affects education as technology increasingly facilitates dialogue across power structures and hierarchies (Sorenson, 2007). Virtual communities have emerged alongside, sometimes replacing, traditional communities. In e-**learning communities,** global citizens often use **virtual classrooms (VCM)**. "Globalization of the world's economies is leading to increased emphasis on internationalization of the curriculum" (Barjis, 2003, p. 1). AFs offer DE interaction opportunities that may be "an acceptable alternative to face-to-face [F2F] discussion" (Payne & Reinhart, 2008, p. 36). In VCM, **identity** is more complex than in F2F settings. Technology brings new cultural products and ways of thinking and acting. DE is a fragmented cultural product and pedagogic design and course management systems are contested as neutral (Payne & Reinhart, 2008; Sorensen, 2007).

The popularity of e-learning in post-2000 is growing. Technology has irrevocably altered business models and policies, including higher education worldwide (Stein, 2001). For example, the UK's OpenLearn project is "lead[ing] the learning revolution, experimenting with new models of content and technologies" as the introduction of tuition fees saw 15,000 less university entrances (NIACE, 2006, p. 4). E-learning is supplementing, and sometimes replacing, traditional classrooms as learners' age increases and universities add **flexible delivery**. In 2004, more than 130 countries were developing or offering DE courses, most using IT (Shields, Gil-Egui, & Stewart, 2004). By 2006, researchers claimed "Web-based distance learning environments is growing exponentially with no limits in sight" (Wijekumar & Spielvogel, p. 221). Adoption of IT for education exhibits great social change (Schifter, 2004) yet offers little consensus despite correspondence courses existing since the 1800s (Romeo, 2001). The global marketplace for e-learning varies widely among and within countries, courses offered and technologies available (Marcus, 2006) with DE shaped by cultural attitudes, communication, infrastructure and government policy (Bowles, 2004). Variation is compounded by multi-sector (education, corporate, government) involvement. As Ragusa (2007) cautions, excluding culture in the development, delivery and evaluation of education technologies poses undesirable learning, economic and communicative consequences.

In contrast with Webb, Jones, Barker, and van Schaik's (2004) quantitative analysis, much AF and DE research focuses on small numbers of graduate and professional experiences (Allan & Lewis, 2006; Beuchot & Bullen, 2005; Christopher, Thomas, & Tallent-Runnels, 2004; Marra, Moore, & Klimczak, 2004). Content analysis is common (Lee & Berter, 2007; Im & Lee, 2004; Marra et al, 2004; Marri, 2007; Zhu, 2006) and supplemented by surveys/interviews. Even when qualitative text analysis of

forum data is proclaimed among "the most valued analytic techniques" (Figaredo & Diaz, 2005, p. 4), much remains positivist. Quantitative analysis of student satisfaction in synchronous e-learning (Chen, Wu, & Yang, 2006) reveals social norms and socialization impact learning satisfaction more than technological systems and learning tasks. Structural change draws attention to the role of student agency in DE structures, an issue receiving little attention (van Aalst & Chan, 2007).

This research adds to proponents of case studies (Allan & Lewis, 2006; Hlapanis, Kordaki, & Dimitrakopoulou, 2006; Schrire, 2006) for analyzing VC to augment quantitative analyses (Beuchot & Bullen, 2005; Au-Yeung, Ha, & Au, 2004; Hawkey, 2004; Webb et al., 2004). Simultaneously, it addresses the common e-learning research limitation of inability to isolate "pure" e-learning, "learning that relies entirely on information and communication technologies" without supplementary F2F interaction which "is rare in Australia" (Bowles, 2004, pp. 25-26). In the U.S., Web-based technologies frequently supplement F2F classroom learning (Wijekumar & Spielvogel, 2006). However, research in Wales (Packham, Jones, Thomas, & Miller, 2006), and this study, demonstrates increasingly online university programs without F2F substantive learning. These structural changes foreground the timeliness and fruitfulness of contextualizing DE in organizational practices and procedures.

AF AND DE IN AUSTRALIA: PRACTICE-BASED EXPERIENCES

Descriptive surveys and qualitative secondary data from more than 800 2005-2007Australian DE undergraduates provide experiences, controversies and key issues on AF and VC. Anonymous student comments from two survey items (Q1 - *Aspects of this subject you found helpful to your learning* & Q2 - *Aspects of the subject you'd like to see changed*) a) reveal virtual realities are guided by communication norms/values and b) show identities are negotiated and recreated by computer-mediated communication set amid corporate policies and institutional cultures.

By comparing two DE environments, case studies show how social structure and culture impact perceptions, communication norms and identity formation. In Virtual Learning Environment 1 (VLE1) (2 cohorts: 2005, N=140 and 2006, N=15), students participated in instructor-driven AF with a peer-learning assessment item derived from their AF work. Virtual Learning Environment 2 (VLE2) (2 cohorts: 2005, N=280 and 2006, N=330) offered AF only as supplementary tool. This research argues VC *type* affects learner satisfaction, subject content and skills used. Findings are case-specific and nongeneralizable.

Main Findings

Research findings are presented in three general themes: 1) Structure and norms affect AF learning and dialogue; 2) AF require management of identities, cultural contests and unforeseen events; and 3) Variation in systemic practices affects AF success. Student perceptions and broader issues are presented by theme.

Structure and Norms Affect AF Learning and Dialogue

Variation in e-learning environments resulted in different learner practices (quantity and quality/type of forum postings, learning and teaching expectations and levels of professionalism). Examining responses from one 2006 third-year subject (N=15) in VLE1, 100% of respondents agreed: i) they enjoyed this form of online learning; ii) the subject forum was an appropriate way to support learning activities; and iii) their understanding of the subject improved because of the subject forum. This echoes experiences of 2005 first-year students (N=140). As one DE student and government employee wrote about her AF work, "this type of exercise mirrors how students would be asked to complete the work on campus and I think it's a really good learning tool" (2005, August 15). According to another, "it really makes a big difference and I am finding that what would normally be for me a very difficult subject [is] very stimulating" (2005, August 26). This adds to Webb et al.'s (2004) quantitative finding that participation in integral e-learning dialogue positively correlates with learning. This study lends qualitative support for the centrality of e-learning structure to student satisfaction and learner practices. Framed by Wenger's learning theory "as social participation in the process of active participation in communities of practice" (Sorenson, 2007, p. 165), students' experiences are part of a macrolevel participatory and reification process requiring adoption of microlevel competencies through online VC engagement.

Debate exists in the course management software (CMS) literature over social control and power in VCM, particularly classroom architectures as instructor/administrator-managed or learner-driven with integrated participation (Payne & Reinhart, 2008). The Australian experiences show dichotomizing learners and instructors/facilitators circumvents the complexity of "control" issues because variation in structural preference also exists between students. VLE1 and VLE2 were organized to specifically address issues of control and student ownership of learning. VLE1 evaluations show students supported highly structured AF assessment tasks. Comments such as "the assessments were extremely beneficial and enhanced learning in the subject" (Comment 1Q1, 2005) and "I found the student forum to be interactive

and reinforced learning of the subject" (2Q1, 2005) contrast with arguments for the centrality of dialogue to learning (Sorensen, 2007). Whereas VLE1 was structured to minimize impromptu student dialogue and student-driven discussion topics (in favor of instructor-determined objectives), VLE2 followed a social constructivist approach, enabling students to drive the nature/extent of virtual dialogue. Variation in computer literacy and AF preferences caused questioning of nonassessable forum discussions: "DE students are not all computer literate… I did not use forums…it never seemed to be the appropriate area to discuss a topic" (6Q2, 2006) with expectations differing by generation and younger generations exhibiting better computer knowledge.

In contrast to VLE1, VLE2 AF users produced little subject-related dialogue, primarily venting grievances with learning and life. In response to a student venting anger at the lecturer, another wrote "the student forum is not a chat room with people anonomously sending opinions and fantasy messages out to just see what happens or to vent without regard to others' feelings. The forum is in the real world and with real people: Would you have spoken to the person that way in front of the class of 170 people? You might say 'yes' but I would suggest you would instantly be very embarrassed if you did" (2005, June 30).

Student concern of supplemental AF usage continued in the 2006 VLE2 and yielded a number of negative comments. Noting the "appalling displays of bad manners on the forum" (25Q2, 2006) by other students one wrote, "I found the forums destructive to my learning. I frequently felt frustrated and annoyed by the attitude and online behavior of some of the other forum users. They did not seem to have a grasp of what was required of THEM at an academic level reducing the forum to a high school classroom. I frequently avoided using the forum for this reason" (2Q1, 2006). In VLE2, issues of quality management and structural arrangements led to debate over larger issues including freedom of speech, "from the 350 or so students enrolled in this [subject] it is unfair to the learning and teaching of this topic to have only ONE lecturer. At times it seemed that there was not enough control on the forums and when everything seemed to get out of hand the easiest way was to close the forum down. As a DE student this was seen as denying some students the freedom of speech" (3Q2, 2006).

While some interpreted AF moderation as censorship, others found moderation inadequate, asking for "stricter control over meaningless comments on the forums" (15Q2, 2006). Others said it was "a waste of time to read some of the students postings" noting "I am aware that it would be extremely helpful to talk about the topics…and this is the only way for DE students," but that "with such a huge workload, there is no time to communicate on the forum" (22Q2, 2006). Although netiquette guidelines were provided to VLE1 and VLE2, they did not meaningfully impact VC unless pedagogical guidelines and task-focused virtual learning were also supplied, as in VLE1, supporting Buelens, Deketelaere and Dierickx's (2007) AF research.

AF used for general communication with limited moderation in VLE2 fostered intolerance in the large, international and culturally diverse cohort. One student commented, "there were some really dumb people doing this subject. Maybe you could make people more patient?" (29Q2, 2006). Deciding whose perspective deserves legitimation is contentious. Virtual participants, equipped with multiple identities and knowledges (as professionals, in various locations with different social roles, etc.) used VC to act out real-world battles. Participants divided along political, ethnic, gender and class allegiances.

Perrotta (2006) argues social constructivism can help understand how culture affects identity production. Questioning how/why identity and tolerance develop, or flounder, via VC are issues these case examples raise, yet cannot solve. DE has potential to cultivate global citizenship, democratic skills and intercultural perspectives (Sorensen, 2007). In practice, learner variation stimulated "global awareness" but failed to promote tolerance, questioning whether building quality "online learning architectures" (Sorensen, 2007, p. 175) are sufficient for learners to work across cultural and geographical divides. Lifelong socialization fosters deeply-embedded norms and values, like nationalism, resilantly manifested in VC.

Despite negative VLE2 surveys, some positive evaluations reveal VC perceptions as highly individualistic. Positive comments stated "the forum was excellent, the lecturer made use of this communication very effectively" (7Q1, 2006), "the online forum is a must for DE" (10Q1, 2006), "great enthusiasm from lecturer and helpful hints regarding learning process" (1Q1, 2006) and "just reading the forum answered many queries" (18Q1, 2006). Student variation may explain how some could find the subject to have "a ridiculous amount in an introductory course" (9Q2, 2006) while others note "I have loved this course and the way in which it has been managed and taught… there has been feedback on the forums about the difficulty of some of the work. I sincerely hope that this kind of feedback will not lead to any kind of 'dumbing down'" (15Q2, 2006). Differences in social capital (Putnam, 2005) may explain such variation, offering insights for future sociological research.

In VLE2 62% of evaluating students agreed assessment tasks assisted their learning and 65% agreed supporting resources (i.e., online forums) facilitated learning. Eighty percent found DE provided adequate opportunities to communicate with the lecturer. Differences in AF quality and structure add to Jorgensen's (2002, p. 9) assertion "when students are actively involved in collaborative (group) learning online, the outcomes can be as good as or better than those for traditional classes, but when individuals are simply receiving posted material and sending back individual work, the results are poorer then in traditional classrooms."

Students in VLE2 and VLE1 raised competing life demands and perceived equity issues: "There was sometimes difficult[y] when other team members were shift workers, etc., and couldn't be online at the same time" (2Q2, 2005); "I would have really enjoyed the subject without the online forum weekly participation. I found that helpful but really hard to maintain with work and life commitments" (10Q2, 2005); "I would like to see the group forum a more equitable process" (5Q2, 2005). These contrast with arguments about nontraditional students being better suited to asynchronous technologies than F2F classrooms (Burgon & Williams, 2003). Still, VLE2 expressed broad agreement in DE technologies as beneficial with 88% agreeing assessment tasks assisted learning, 82% feeling online forums facilitated learning and 88% believing there were adequate communication opportunities with the lecturer.

AF Require Management of Identities, Cultural Contests and Unforeseen Events

The second theme is the importance of anticipating and managing unanticipated events. Unanticipated events may be technical (i.e., disrupted Internet service) or situational (i.e., outbreak of war). Regardless of event-type, AF can have a calm group of participants quickly turn into seething debaters (often outside business hours). Communication norms research (Palloff & Pratt, 2001) has used stage and systems theory to explain electronic group dynamics, including conflict and resolution. Management of unexpected dialogue often requires educators to quickly formulate solutions, put out "fires" or assume counseling roles. According to Lynch (2004, p. 109), "learning online supports dialogue and the collaborative development of understandings" while DE bridges individuals' life roles (Lynch, 2004). When engaging in AF, master statuses (gender, class, race, etc.) and social identities responsible for the enforcement of social norms are covert. VC's informal nature and transgression of traditional boundaries (geographical, social and economic) facilitate intimate dialogue among potentially isolated/segregated social groups creating potential for (politically incorrect) debate (Ragusa, 2007). The asynchronous and textual nature of VC dialogue poses unique challenges not experienced in F2F interactions:

- time to compile/access expertise from multiple resources for debate;
- lack of consensus regarding pre-established definitions (i.e., "indigenous" among more than 300 cross-cultural individuals);
- how to interpret meaning(s) of written text if intonation/context is lacking;
- uncertainty if anyone has heard/read one's contribution;
- lack of status symbols (i.e., uniform, job titles, credentials) to legitimate statements; and
- ability to affect/control the path of future dialogue.

Who has the right to speak is an epistemological and political issue (Roof & Wiegman, 1995)? Computers are more than IT tools; they are social artifacts mirroring self-images, encouraging reflection of self-presentation. How one defines situations is critical to the subjective meanings individuals attach to VC. According to Lynch (2004, p. 13), "when reading a typed message, there is a strong tendency to project—consciously or not—your own expectations, wishes, anxieties, and fears into what the person wrote. This may lead to further conflict and, because it is written, sometimes builds a feeling of resentment that lasts longer." VLE1 and VLE2 experiences showed VC encourages "self-talk," the thinking and rationalization process, to become public. Individuals often wrote "why" they advocated world views and reflected in writing for later review by all. Consequently, there is need for task clarity, focused collaboration and quick leadership intervention, if conflict arises for successful group functioning (McClure, 1998) and to diminish misinterpretion (Palloff & Pratt, 2001).

Variation in Systemic Practices Affects AF Success

The third theme is how systemic conditions impact delivery and quality of AF. Disgruntlment about educational design led individualist and capitalist comments, "I get nothing for my money... I feel completely ripped off" (11Q2, 2006). Statements like "I feel [the lecturer's] attempt to have the students take responsibility for their own learning may have been optimistic as it appeared many students were new to tertiary education and were overwhelmed by the need 'to think for themselves' (21Q1, 2006) articulate learners' comfort level with self-driven technologies and show to compare students' and instructors' perceptions (Mazzolini & Maddison, 2007). Students in both groups were unforgiving of IT failure (i.e., electronic assessment systems, incompatible files). Computer-generated reports led to outpours of computer-phobic sentiments and panic unique to VC. Structuring VLE1 with multiple educational technologies (CD readings, subforums, electronic lectures/student presentations, electronic exams, assignment submission) heightened potential for IT difficulty and elevated apprehension about IT inadequacies. Still, IT variation increased skill development and learning satisfaction, while practice virtual exercises reduced stress, and enhanced competence and willingness to use IT.

Student willingness to participate in AF is a key issue (Vredenburg, 2004). Vredenburg (2004) serendipitously found online minute papers enabled shy students to communicate with garrulous peers. Learners in VLE1 and VLE2

did not demonstrate similar behavior. "Students need to be encouraged to use the forum, would of [been] nice to hear from others instead of the same few" (7Q2, 2006). Making online tasks assessable made little difference to the garrulousness of students. Like F2F, AF had a few dominant voices interspersed by a quieter majority. This reveals resilience of social interaction norms. Research is needed to identify necessary conditions to enhance engagement. In contrast with Vredenburg (2004), all VLE1 and VLE2 learning was DE, yielding core and peripheral communicators. Deeper exploration of why/how individuals decide to participate, particularly where DE is the only learning mode, requires analysis. However, structuring learning environments so users can preselect when to contribute taught time management skills and empowered users to opt in/out of culturally sensitive issues (i.e., death, domestic violence).

FUTURE TRENDS

A number of trends emerge from this research. DE offers viable alternatives for higher education circumventing some infrastructure issues, such as lack of Australian rural universities. Yet, IT discrepancies among urbanites, rural and international students remain. Antiquated policies and systemic inconsistencies (i.e., postal return of assignments/grades coexisting with electronic exams) cause user frustration. In considering culture, one must transgress attitudes, behaviors and the social construction of sameness (Lum, 2006) to note how tacit knowledge, embedded in institutional culture, are acted out in policy. Infrastructural variation (i.e., slow mail delivery in Pipalatjara due to no rain in 2 years and by camel in the Sudan) are exacerbated by university policies.

Leaders in DE technologies must be aware of how socially created cleavages (culture, class, age, etc.) impact learning. Broader global adoption (Marcus, 2006) increases contact among individuals of divergent backgrounds. Diversity heightens culturally-contingency and interpretation in VC. As DE gains legitimacy and standing on par with F2F, new norms, customs and roles will develop from the unique group dynamics of virtual communities. Lacking the symbols and rituals of F2F (physical/verbal cues), mechanisms to manage unintended meanings, perceptions and consequences are needed because the frequency, depth and style of VC interactions differ. VC should not be left to ad hoc management, or postponed until convenient. VC requires active, consistent and flexible management, as learners are socialized to adopt communication norms and behaviors conducive to creating safe and supportive learning environments.

CONCLUSION

This research demonstrated culture is embedded in DE and fundamentally affects VC. Lack of sociological analysis was documented and arguments made for what socio-cultural analysis can offer DE organizations, participants, evaluators and planners. Experiences drawn from a large sample of Australian DE undergraduates in two differently-structured VLE highlighted key issues AF participants faced. The important role organization and management hold for satisfactory learning outcomes was revealed. The subjective nature of electronic communication was contrasted with the real-world impact of infrastructure limitations, both affecting social interaction norms. Potential and challenges of asynchronous communication was contextualized amid service delivery expectations in a consumerist world to argue changes necessitate re-evaluation of discursive practices and organizational policies. Findings reflect the complexity of virtual communication and how DE surpasses 18th-century correspondence courses. How VC structures affect social interaction in unique and patterned is argued to require further investigation.

REFERENCES

Allan, B., & Lewis, D. (2006). Virtual learning communities as a vehicle for workforce development. *The Journal of Workplace Learning, 18*(6), 367-383.

Au-Yeung, L.H., Ha, T., & Au, G. (2004). The experience of new WBI-adopters in Hong Kong. *Journal of Educational Technology Systems, 31,* 411-422.

Barjis, J. (2003). An overview of virtual university studies. In F. Albalooshi (Ed), *Virtual education* (pp. 1-20). Hershey, PA: IRM.

Beuchot, A., & Bullen, M. (2005). Interaction and interpersonality in online discussion forums. *Distance Education, 26*(1), 67-87.

Bowles, M.S. (2004). *Relearning to e-learn*. Victoria, Australia: Melbourne University.

Buelens, H., Totte, N., Deketelaere, A., & Dierickx, K. (2007). Electronic discussion forums in medical ethics education. *Medical Education, 41*(7), 711-717.

Burgon, H., & Williams, D.D. (2003). Bringing off-campus students on campus. *The Quarterly Review of Distance Education, 4*(3), 253-260.

Castells, M. (2000). *The information age* (Vols. 1-3). Oxford: Blackwell.

Chen, C.C., Wu, J., & Yang, S.C. (2006). The efficacy of online cooperative learning systems. *Campus-Wide Information Systems, 23*(3), 112-127.

Christopher, M.M., Thomas, J.A., Tallent-Runnels, M.K. (2004, Spring). Raising the bar. *Roeper Review, 26*(3), 1-13.

Edwards, M.E., Cordray, S., & Dorbolo, J. (2000). Unintended benefits of distance-education technology for traditional classroom teaching. *Teaching Sociology, 28*, 386-391.

Figaredo, D., & Diaz, L. (2005, February). Conference report: II online Congress for the observatory of the cyber society. *Qualitative Social Research, 6*(2), Art.2. Retrieved May 31, 2008, from http://www.qualitative-research.net/fqs-texte/2-05/05-2-2-e.htm

Hawkey, K. (2004). Assessing online discussions working "along the grain" of current technology & educational culture. *Education & Information Technologies, 9*(4), 377-386.

Hlapanis, G., Kordaki, M., & Dimitrakopoulou, A. (2006). Successful e-courses. *Campus-Wide Information Systems, 23*(3), 171-181.

Holmes, D., Hughes, K., & Julian, R. (2007). *Australian sociology*. Frenchs Forest: Pearson Education.

Im, Y., & Lee, O. (2004). Pedagogical implications of online discussion for preservice teacher training. *Journal of Research on Technology & Education, 36*(2), 155-170.

Jaffee, D. (2003). Virtual transformation. *Teaching Sociology, 31*, 227-236.

Jorgensen, D. (2002). The challenges and benefits of asynchronous learning networks. In H. Iver (Ed.), *Distance learning* (pp. 3-17). New York: Haworth Information.

Lee, H. (2005). Behavioral strategies for dealing with flaming in an online forum. *Sociological Quarterly, 46*(2), 385-403.

Lee, E.O., & Bertera, E. (2007). Teaching diversity by using instructional technology. *Multicultural Education & Technology Journal, 1*(2), 112-125.

Lum, L. (2006). Internationally-educated health professionals. *Education & Training, 48*(2/3), 112-126.

Lynch, M. (2004). *Learning online*. New York: Routledge Falmer.

Marcus, S. (2006). Measure by measure. Campus-Wide Information Systems, 23(2), 56-67.

Marra, R. M., Moore, J.L., & Klimczak, A.K. (2004). Content analysis of online discussion forums. *Educational Technology Research & Development, 52*(2), 23-40.

Marri, A.R. (2007). Working with blinders on. *Multicultural Education & Technology Journal, 1*(3), 144-161.

Mazzolini, M., & Maddison, S. (2007). When to jump in. *Computers & Education, 49*(2), 193-213.

McClure, B. (1998). *Putting a new spin on groups*. Hillsdale: Erlbaum.

National Institute of Adult Continuing Education. (2006, November). OU provides learning materials free online. *Adults Learning*, 4-5.

Packham, G., Jones, P., Thomas, B., & Miller, C. (2006). Student and tutor perspectives of on-line moderation. *Education & Training, 48*(4), 241-251.

Palloff, R.M., & Pratt, K. (2001). *Lessons from the cyberspace classroom*. San Francisco: Jossey-Bass.

Payne, C.R., & Reinhart, C.J. (2008). Can we talk? *On the Horizon, 16*(1), 34-43.

Perrotta, C. (2006). Learning to be a psychologist. *Journal of Computer-Assisted Learning, 22*(6), 456-466.

Putnam, R.D. (Ed.). (2005). *Democracies in flux*. Oxford: Oxford University.

Ragusa, A. (2007). The impact of socio-cultural factors in multi-cultural virtual communication environments. In K. St-Amant (Ed.), *Linguistic & cultural online communication issues in the global age* (pp. 306-327). Hershey, PA: IGI.

Romeo, L. (2001). Asynchronous environment for teaching & learning. *The Delta Kappa Gamma Bulletin, 6*(3), 24-28.

Roof, J., & Wiegman, R. (Eds.). (1995). *Who can speak*? Chicago: University of Illinois.

Schifter, C. (2004). Faculty participation in DE programs. In D. Monolescu, C.C. Schifter, & L. Greenwood (Eds.), *The distance education evolution* (pp. 1-21). Hershey, PA: ISP.

Schrire, S. (2006). Knowledge building in asynchronous discussion groups. *Computers & Education, 46*(1), 49-70.

Schopler, J., Abell, M., & Galinsky, M. (1998). Technology-based groups. *Social Work, 4*(3), 254-269.

Semple, A. (2000). Learning series & the influence on the development and use of educational technologies. *Australian Science Teachers Journal, 46*(3), 21-28.

Shields, S.F, Gil-Egui, G., & Stewart, C.M. (2004). Certain about uncertainty. In D. Monolescu, C.C. Schifter, &

L. Greenwood (Eds.), *The distance education evolution*. Hershey, PA: ISP.

Sorensen, E.K. (2007). Dialogic e-learning2learn. *Multicultural Education & Technology Journal, 1*(3), 162-177.

Stein, A. (2001). Preparation for e-learning. In F. Albalooshi (Ed.), *Virtual education* (pp. 140-155). Hershey, PA: IRM.

Van Aalst, J., & Chan, C.K.K. (2007). Student-directed assessment of knowledge building using electronic portfolios. *Journal of the Learning Sciences, 16*(2), 175-220.

Vygotsky, L. (1986). *Thought and language*. Cambridge, MA: MIT.

Webb, E., Jones, A., Barker, P., & van Schaik, P. (2004). Using e-learning dialogues in higher education. *Innovations in Education & Teaching International, 41*(1), 93-103.

Wijekumar, K.K., & Spielvogel, J. (2006). Intelligent discussion boards. *Campus-Wide Information Systems, 23*(3), 221-232.

Zhu, E. (2006). Interaction and cognitive engagement. *Instructional Science, 34*(6), 451-480.

KEY TERMS

Asynchronous Forum: Online communication billboard unconstrained by date/time. Users post statements for others' immediate or later review. Past postings stay viewable, depending on host server or project length.

E-Learning: E-learning uses technology to teach/learn and is a type of "distance education." University degrees via e-learning may use electronic assessments, virtual classrooms, and online resources.

Flexible Delivery: A mode of education more adaptable to time/geographical constraints than face-to-face classrooms. Often appeals to mature, rural/remote students or others with competing life demands (i.e., employment/child care).

Learning Communities: A social group sharing common experiences, cognitive boundaries, sense of belonging and geographical space (virtual or physical) who come together to learn.

Online Identity: Identity refers to a fluid social-psychological sense of "self" in relation to others. Online identities may be similar/different to "real world" identities.

Virtual Classrooms: Exist in contrast to face-to-face learning. Often used in distance education with asynchronous (consecutive) or synchronous ("real-time") communication.

Virtual Communication: Electronic information transfer between individuals/groups via the Internet. Can be text-based (e-mail, wikis, forums) or oral (podcasts). Meanings and communication norms differ from face-to-face communication.

Software Agents in E-Commerce Systems

Juergen Seitz
University of Cooperative Education Heidenheim, Germany

INTRODUCTION

The Internet introduces a new global marketplace for a large number of relatively unknown and not seldom small companies often offering substitutive or complimentary products and services. The merchants profit from reduced costs, reduced time, and unsold stocks. Customers are attracted by increasing convenience and fast fulfillment.

Merchants offering these products and services on this new marketplace need to acquire new customers and sustain ongoing relationships. Nowadays, most merchants' sites are passive catalogs of products and prices with mechanisms for orders (Dasgupta, Narasimhan, Moser, & Melliar-Smith, 1998). The pull strategy is also applied in auctions available over the Internet, where the seller waits passively for bids. The new push technologies for electronic commerce, like software agents, enable customers to compare a bewildering array of products efficiently, effectively, and automatically (Jennings, Sycara, & Wooldridge, 1998). Switching costs for customers and, thereby, their loyalty to previous suppliers in the marketplace decline (Phlips, 1989; Schwartz, 1999).

Using the Internet, the producers profit from reduced cost through direct, non-intermediated sales. The key elements to successful long-term relationships between merchants and customers will be the offering of personalized and value-added services, like one-to-one marketing services, discounts, guarantees, and savings coupons (Seitz, Stickel, & Woda, 2003).

In this article, we will analyze possible consequences of new push and pull technologies in electronic commerce for customer's loyalty, as well as the active technologies enabling customers to purchase efficiently and for the merchants to offer high personalized, value-added, and complimentary products and services. We will discuss some examples of such services and personalization techniques sustaining one-to-one relationships with customers and other actors involved.

BACKGROUND

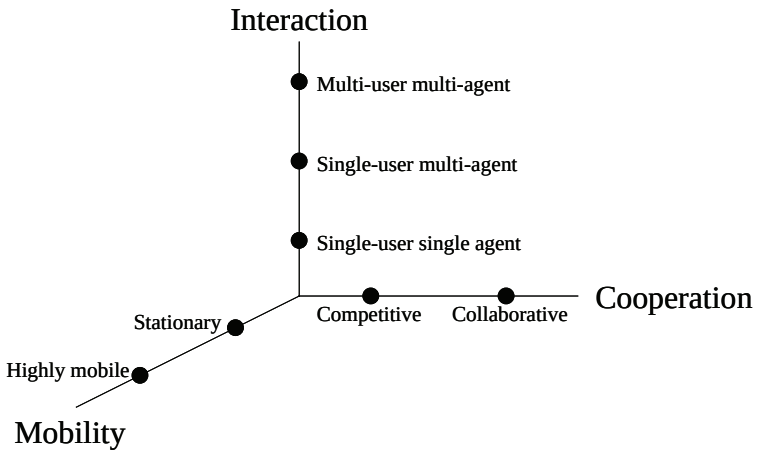
The World Wide Web provides a great opportunity to compare products and services. Customers as well as competitors may quickly gain detailed and up-to-date data. Especially, suppliers of digital goods are in fear of declining customer's loyalty. Customers compare catalogs of products of merchants and producers, and conduct transactions independently of their geographic localization. The crucial basic factors responsible for a limited loyalty of customers are convenience, time, and cost of fulfillment. So, an electronic commerce system should support the ability to embed intelligence to automate the decision process (Dasgupta et al., 1998). The system should not only compare products and prices, but also negotiate and finally purchase products (Teuteberg & Kurbel, 2002). Nowadays, most systems still involve a substantial human element, that is, from the consumer's perspective neither convenient nor efficient. The human involvement should be limited to transaction specification at the beginning of the process and to the buying or refusal decision at the end of the process (Chen, 2000). This means that an appropriate technology is necessary in the intermediate stages to coordinate between customers and suppliers (d'Inverno & Luck, 2003). Mobile software agents emerge as ideal mediators in electronic commerce and thereby as an appropriate technology for an automated procurement process. Customers may specify constraints on the features of products that enable mobile agents to select products from the merchants' catalogs and finally to determine the terms of the transaction. Otherwise, software agents may be used by suppliers as market surveyors to determine the current demand and an appropriate price for the good (Chernev, 2003; Fay, 2004; Hann & Terwiesch, 2003; Spann, Klein, Makhlouf, & Bernhardt, 2005; Spann, Skiera, & Schaefers, 2004; Spann & Tellis, 2006). Software agent technology also abolishes the problem of different technological standards, like hardware platforms and operating systems of remote computers. This means that geographical or technological barriers for customers are of no significant importance any more. The key factors are convenience, time, and cost of the procurement process.

AGENT-MEDIATED ELECTRONIC COMMERCE

Software agents are computer programs showing the following characteristics (Joshi & Ramesh, 1998):

- **Reactivity:** Agent perceives and reacts to environmental changes.
- **Autonomy:** Agent has its own program code, data, and execution state.

Figure 1. Classification of software agents (Joshi & Ramesh, 1998)



- **Proactivity:** Agent initiates changes to the environment.

The ability of an agent to travel around in networks enhances it to a mobile agent (Brenner, Zarnekow, & Wittig, 1998). Mobile software agents may be classified based on their attributes, like mobility, type of cooperation, and level of interactivity (see Figure 1) (Joshi & Ramesh, 1998). For further possible classifications see, for example, Nwana (1996), Sycara, Decker, Pannu, Williamson, and Zeng (1996), or Kurbel and Loutchko (2001).

Competitive agents, mostly single-agents, maximize the interests of their owners. Collaborative agents, on the contrary, share their knowledge and try to maximize benefits of the community on the whole (Joshi & Ramesh, 1998). Mobile agents differ also in terms of the ease of the mobility between remote computers. A continuously traveling nomadic agent, like mobile sales agents (containing information of the total quantity of the product to be sold, the initial price of the product, and the list of potential customers to visit), arrives at a customer's site and communicates with a stationary customer agent that determines the quantity to be purchased at a given price (Dasgupta et al., 1998). The customer agent uses market values and demand curves of the product for its decision. The sales agent has to adjust the price dynamically during negotiations in order to maximize the gross returns. The price for the product may not be settled too low (an agent sells all of his stock at a bargain price) or too high (a given quantity of the goods may be unsold). Such a supplier-driven electronic commerce system enables merchants to maximize their gross return, but also to identify quickly the customers' needs and finally to cultivate long-term relationships with them. The architecture of the supplier-driven system was presented by Dasgupta et al. (1998).

From a customer's perspective, software agents should be highly personalized, continuously running, and autonomous mediators that have to delegate some process management tasks (Guttman, Moukas, & Maes, 1998). A software agent should identify a customer's needs at first, then retrieve information about the product from the merchants' sites, compare the offers, and finally determine the terms of the transaction (Castro-Schez, Jennings, Luo, & Shadbolt, 2004). Nowadays, customer agents are mostly used for product and merchant brokerage and for negotiation (Guttman et al., 1998).

The price of a product may also be dynamically negotiated instead of being fixed (Phlips, 1989; Schwartz, 1999). For example, tête-à-tête agents cooperatively negotiate multiple terms of a transaction, like warranties, return policies, delivery times, and loan options (Guttman et al., 1998). The buyer agent in a tête-à-tête system negotiates toward a pareto-optimal deal with the sales agent (Fatima, Wooldridge, & Jennings, 2004). A system like this does not maximize gross returns to suppliers or price discounts for customers (Excelente-Toledo & Jennings, 2004; Rahwan et al., 2004). However, it takes into consideration the important value-added merchants' services (Weerakkody, Currie, & Ekanayaka, 2003).

Summarizing, software agents are helping customers to compare and to purchase goods in the Internet. Most of them are agents for a simple online product price comparison or for competitive negotiation over price without considering the value-added and post purchase services from merchants. Such agents decrease customers' loyalty to a merchant toward zero. However, additional services, like guarantees, return policies, loans, gifts, discounts, and insurance are of interest to customers. Therefore, they should rather use agents comparing or negotiation over multiple terms of a transaction (tête-à-tête). Otherwise, merchants may also send their

own sales agents to potential buyers in order to acquire new customers and remind the previous customers of new sales offerings (Dutta, Moreau, & Jennings, 2003).

FUTURE TRENDS

In general, software agents helping customers in the procurement process may minimize their loyalty to merchants. Suppliers who do not want to compete solely on the basis of price provide their customers with highly personalized and value-added services to sustain a long-term relationship.

Personalization and Privacy

Personalization is defined as the customization of a Web site to meet the particular needs of individual users (Chaffey, Mayer, & Johnston, 2000; Dean, 1998). The goals of personalization technology are to encourage repeated visits and to enhance user loyalty. The identification of customers' needs occurs through the observation of their behavior and the collection of data (filling out forms or following decision-tree sets of questions).

There exist some advanced techniques supporting personalization Web site contents, like rule-based matching and collaborative filtering. Using rule-base matching, users have to answer a set of yes/no or multiple choice questions to settle a set of user's criteria. Collaborative filtering methods combine the personal preferences of a user with preferences of like-minded people (Dean, 1998). In regard to personalization techniques, one-to-one-marketing should be noticed. This strategy enables targeting unique offers to specific customers (Chaffey et al., 2000). Institutions offering such individualized services have to dispose of accurate user profiles before.

A critical factor of personalization is the privacy issue. Filtering and customization techniques entail the collection and the use of personal data, like name, e-mail address, postal address, age, gender, income, Internet connection, and employment status, that must be protected from abuse (Dean, 1998). Furthermore, a lot of suppliers in the Internet deriving revenues mainly from advertising need to identify their users in order to better customize the content and to attract the advertisers being interested. Hence, the user should be informed by suppliers how they use the personal data and how they protect it. Nowadays, there are several initiatives and standardization projects for the privacy of personal data usage. Such initiatives increase user trust and confidence in electronic commerce. However, no organization or institution has the power to enforce it to the wide usage of suppliers.

Value-Added Services

Merchants who do not want to compete solely on the basis of prices often offer their customers value-added or complimentary services. Complimentary products imply higher benefits to the customer in the case he or she only buys the product he or she looks for (Seitz, Stickel, & Woda, 1999). Such products or services increase the value of the primary good to the customer. Examples of value-added services in electronic commerce are sales discounts, savings coupons, additional insurance and guarantees, gifts, and also free software to test. In general, value-added services enable customers to trade at favorable terms and with confidence. They increase the attractiveness of the merchant to present customers and attract new customers.

Reduced Financial Transaction Costs

Internet merchants might achieve additional reduction of transaction costs using electronic payment systems (Bakos, 1998). Electronic payment systems reduce cash handling costs for merchants and improve speed and convenience for customers. The aggregated cost of each payment consists of transformation costs, for example, the fees for conversion from assets to cash and vice versa, transport and storage costs, costs for safety measures, and search and time costs (Hakenberg, 1996).

CONCLUSION

This article discusses consequences of electronic commerce on customer's loyalty. Electronic commerce in the Internet offers the possibility to create a perfect marketplace. The intermediation in distribution will be reduced. This means lower costs for both suppliers and customers.

Software agents may be classified into different types with regard to their use on supporting electronic commerce. Most of the software agents only perform simple product price comparisons; some support the purchase of products. These software agents reduce customers' loyalty because the price is the only parameter. Quality and added values are not considered. Therefore, multi-agent systems allowing negotiation might be useful from a customer's perspective. Merchants may also send their own sales agents to potential buyers in order to remind previous customers of new sales offerings or to suppliers in order to maximize their gross return.

Personalization and customization of Web sites, value-added services, and the reduction of transaction costs are instruments for increasing customers' loyalty. Personalization techniques, like rule-based matching and collaborative filter-

ing, provide contents on Web sites that are appropriate to the customers' preferences or analyze past purchases and prior suggestions of other customers. One-to-one-marketing may be especially useful for sophisticated products demanding explanations or to enable cross selling of other products and services. User profiles allow merchants to make customer-oriented offers or build special offers including additional services. Value-added services attract the customer to trade at favorable terms. The usage of electronic payment systems may reduce transactions costs.

REFERENCES

Bakos, Y. (1998). The emerging role of electronic marketplaces on the Internet. *Communications of the ACM, 41*(8), 35-42.

Brenner, W., Zarnekow, R., & Wittig, H. (1998). *Intelligente Softwareagenten. Grundlagen und Anwendungen*. Berlin: Springer.

Castro-Schez, J. J., Jennings, N. R., Luo, X., & Shadbolt, N. (2004). Acquiring domain knowledge for negotiating agents: A case study. *International Journal of Human Computer Studies, 61*(1), 3-31.

Chaffey, D., Mayer, R., & Johnston, K. (2000). *Internet marketing*. London: Prentice Hall.

Chen, Z. (2000). Intelligent agents. In M. Zeleny (Ed.), *The IEBM handbook of information technology in business* (pp. 561-569). London: Thomson Learning.

Chernev, A. (2003). Reverse pricing and online price elicitation strategies in consumer choice. *Journal of Consumer Psychology, 13*(1/2), 51-62.

Dasgupta, P., Narasimhan, N., Moser, L. E., & Melliar-Smith, P. M. (1998). A supplier-driven electronic marketplace using mobile agents. In *Proceedings of the First International Conference on Telecommunications and E-Commerce* (pp. 42-50). Nashville, TN.

Dean, R. (1998, June 2). *Personalizing your Web site*. Retrieved August 29, 2003, from http://builder.cnet.com/webbuilding/pages/Business/Personal/

d'Inverno, M., & Luck, M. (2003). *Understanding agent systems*. Berlin: Springer.

Dutta, P. S., Moreau, L., & Jennings, N. R. (2003). Finding interaction partners using cognition-based decision strategies. In *Proceedings of the IJCAI workshop on Cognitive Modeling of Agents and Multi-Agent Interactions* (pp. 46-55). Acapulco, Mexico.

Excelente-Toledo, C. B., & Jennings, N. R. (2004). The dynamic selection of coordination mechanisms. *Journal of Autonomous Agents and Multi-Agent Systems, 9*(1-2), 55-85.

Fatima, S., Wooldridge, M., & Jennings, N. R. (2004). Bargaining with incomplete information. *Annals of Mathematics and Artificial Intelligence, 44*(3), 207-232.

Fay, S. (2004). Partial repeat bidding in the name-your-own-price channel. *Marketing Science, 23*(3), 407-418.

Guttman, R. H., Moukas, A. G., & Maes, P. (1998). Agents as mediators in electronic commerce. *EM-Electronic Markets, 8*(1), 22-27.

Hakenberg, T. (1996). Elektronische Zahlungssysteme im Wettstreit mit dem Bargeld. *Sparkasse, 59*(6), 271-274.

Hann, I.-H., & Terwiesch, C. (2003). Measuring the frictional costs of online transactions: The case of a name-your-own-price channel. *Management Science, 49*(1), 1563-1579.

Jennings, N. R., Sycara, K., & Wooldridge, M. (1998). A road map of agent research and development. *Autonomous Agents and Multi-Agent Systems, 1*(1), 7-38.

Joshi, N., & Ramesh, V. C. (1998). *On mobile agent architectures* (Tech. Rep.). Illinois Institute of Technology, ECE Department, Chicago, IL.

Loutchko, I., & Kurbel, K. (2001). A framework for multi-agent electronic marketplaces: Analysis and classification of existing systems. In *Proceedings of International ICSC Congress on Information Science Innovations (ISI 2001),* American University in Dubai, U. A. E. (pp. 335-339).

Nwana, H. S. (1996). Software agents: An overview. *The Knowledge Engineering Review, 11*(3), 205-244.

Phlips, L. (1989). *The economics of price discrimination.* Cambridge: Cambridge University Press.

Rahwan, I., Ramchurn, S. D., Jennings, N. R., McBurney, P., Parsons, S. & Sonenberg, L. (2003). Argumentation-based negotiation. *The Knowledge Engineering Review, 18*(4), 343-375.

Schwartz, E. I. (1999). *Digital Darwinism: 7 breakthrough business strategies for surviving in the cutthroat Web economy.* New York: Broadway.

Seitz, J., Stickel, E., & Woda, K. (1999). Electronic payment systems: A game-theoretic analysis. In M. Khosrow-Pour (Ed.), Managing information technology resources in organizations in the next millennium. *Proceedings of the 1999 Information Resources Management Association International Conference* (pp. 564-568). Hershey, PA: Idea Group Publishing.

Seitz, J., Stickel, E., & Woda, K. (2002). Impacts of software agents in e-commerce systems on customer's loyalty and on behavior of potential customers. In B. Fazlollahi (Ed.), *Strategies for e-commerce success* (pp. 208-223). Hershey, PA: IRM Press.

Spann, M., Klein, J., Makhlouf, K., & Bernhardt, M. (2005). Interaktive Preismassnahmen bei Low-Cost-Fluglinien. *Zeitschrift fuer Betriebswirtschaft (ZfB), 75*(EH1), 53-77.

Spann, M., Skiera, B., & Schaefers, B. (2004). Measuring individual frictional costs and willingness-to-pay via name-your-own-price mechanisms. *Journal of Interactive Marketing, 18*(4), 22-36.

Spann, M., & Tellis, G. (2006). Does the Internet promote better consumer decision? The case of name-your-own-price auctions. *Journal of Marketing, 70*(1), 65-78.

Sycara, K., Decker, K., Pannu, A., Williamson, M., & Zeng, D. (1996, December). Distributed intelligent agents. *IEEE Expert*. Retrieved July 11, 2006, from http://www.cs.cmu.edu/~softagents/papers/ieee-agents96.pdf

Teuteberg, F., & Kurbel, K. (2002). Anticipating agents' negotiation strategies in an e-marketplace using belief models. In W. Abramowicz (Ed.), *Proceedings of the Fifth International Conference on Business Information Systems (BIS 2002)* (pp. 91-100). Poznan, Poland: Akademia Ekonomiczna w Poznaniu, Katedra Informatyki Ekonomicznej.

Weerakkody, V., Currie, W. L., & Ekanayaka, Y. (2003). Reengineering business processes through application service providers—Challenges, issues and complexities. *Business Process Management Journal, 9*(6), 776-794.

KEY TERMS

Collaborative Filtering: Collaborative filtering methods combine personal preferences of a user with preferences of like-minded people to guide the user.

Customer Loyalty: Because there is no existing ownership to service products, suppliers have to make special efforts to get long-standing customers.

Customer Profiling: Usage of the Web site to get information about the specific interests and characteristics of a customer.

Customization: The adjustment of products or services to individual needs. Basic characteristics are implemented in the product or service and may be controlled by parameters.

Disintermediation: The elimination of agents, like wholesale dealers or brokers, who built the former relationship between producer and consumer. Disintermediation allows the direct supply of the consumer.

One-to-One-Marketing: The direct dialog between a producer and an individual consumer or a group of consumers with similar needs.S

Personalization: Web-based personalization means providing customized content to individual users using Web sites, e-mails, and push technologies.

Software Agents: Computer programs that are characterized by reactivity, autonomy, and proactivity. Therefore, the software agent interacts with its environment.

Software and Systems Engineering Integration

Rick Gibson
American University, USA

INTRODUCTION

With software an increasingly significant component of most products, it is vital that teams of software and systems engineers collaborate effectively to build cost effective, reliable products. This article will identify the key aspects of software engineering and systems engineering in an effort to highlight areas of consensus and conflict to support current efforts by practitioners and academics in the both disciplines in redefining and integrating their professions and bodies of knowledge.

In response to increasing concerns about software development failures, the Software Engineering Institute (SEI) pioneered a software process improvement model in 1988, with the fully developed version of their Capability Maturity Model for Software (SW- CMMâ) appearing in 1993. Since the early nineties, there have been comparable improvement models introduced in the systems engineering community as well, some of which have been published and widely accepted include: Systems Engineering Capability Maturity Model (SE-CMM), also known as the Electronic Industries Alliance Interim Standard (EIA/IS) 731, Systems Engineering Capability Model (SECM), and the Integrated Product Development Capability Maturity Model (IPD-CMM). The resulting avalanche of models and standards has been described by Sarah Sheard (Software Productivity Consortium) as a "Framework Quagmire". In December of 2000, the SEI initiated the Capability Maturity Model–Integrated (CMMI^{SM}) project, which combines best practices from the systems and software engineering

Table1. Software and system engineering similarities and differences

Similarities	Differences
Definition and analysis involves manipulation of symbols.	Software is not subject to physical wear or fatigue.
Highly complex aggregation of functions, requiring satisfying (though not optimizing) multiple criteria.	Copies of software are less subject to imperfections or variations.
Decisions driven by need to satisfy quality attributes such as reliability, safety, security, and maintainability.	Software is not constrained by the laws of physics.
Easy and dangerous to suboptimize solutions around individual subsystem functions or quality attributes.	Software interfaces are conceptual, rather than physical—making them more difficult to visualize.
Increasing levels of complexity and interdependency.	Relative to hardware, software testing involves a larger number of distinct logic paths and entities to check.
	Unlike hardware, software errors arrive without notice or a period of graceful degradation.
	Hardware repair restores a system to its previous condition; repair of a software fault generally does not.
	Hardware engineering involves tooling, manufacturing, and longer lead times, while software involves rapid prototyping and fewer repeatable processes.

disciplines. (Note: CMMâ and CMMI℠ are copyrights and service marks of the Software Engineering Institute.)

Recent studies (Carter et al., 2003; Goldenson & Gibson, 2003) have validated the SEI's assertion the each of the disciplines benefit from incorporation of principles from the other. Moreover, there appears to be no fundamental differences between the disciplines that would prevent their integration.

BACKGROUND

There is great hope that the SEI imitative will provide the impetus to overcome some long-standing discipline boundaries. The nature of the systems and software engineering work has led to terminology differences rooted in the very descriptions of the disciplines. One important problem with software is the difficulty in understanding its inherent level of quality.

Issues and concerns regarding such an integration were articulated by Barry Boehm and Fred Brooks as early as 1975. Boehm suggested that the adoption of systems engineering reliability techniques by software engineers was counterproductive. Moreover, Brooks' Law suggests that a common systems engineering solution to schedule slippage (add more people) will only make late software projects even later.

More recently, Boehm (1994) expressed concerns that, in spite of the central function of software in modern systems, the two engineering disciplines have not been well integrated. Boehm articulated similarities and differences as shown in Table 1.

Software engineering, as defined by the Institute of Electrical and Electronics Engineers (IEEE, 2001), is: (1) the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software; that is, that application of engineering to software; (2) The study of approaches as in (1)—and further identifies the body of knowledge for software engineering to be: software requirements, software design, software construction, software testing, software maintenance, software configuration management, software engineering management, software engineering process, software engineering tools and methods, and software quality.

A useful definition of systems engineering resides in an in-process body of knowledge document by the International Council on Systems Engineers (Leibrandt, 2001, p. 3), which defines systems engineering in terms of product and process: "…product oriented engineering discipline whose responsibility is to create and execute an interdisciplinary process to ensure that customer and stakeholder needs are satisfied in a high quality, trustworthy, cost effective and schedule compliant manner throughout a system's lifecycle". The process starts with customer needs, and consists of stating the problem, investigating alternatives, modeling, integrating, launching the system, and assessing performance. Moreover, the system engineer is responsible for pulling together all the disciplines to create a project team to meet customers' needs. The complete systems engineering process includes performance, testing, manufacturing, cost, schedule, training and support, and disposal. The body of knowledge recognizes that systems engineering processes often appear to overlap software and hardware development processes and project management. Thus, systems engineering is a discipline that focuses on processes; it develops structure, and efficient approaches to analysis and design to solve complex engineering problems. In response to concerns about integrated development of products, the system engineer plans and organizes technical projects and analyzes requirements, problems, alternatives, solutions and risks. Systems engineering processes are not specific to a particular discipline; they can be applied in any technical or engineering environment.

In short, software engineering is defined by IEEE Standard 610.12 as the application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software—that is, the application of engineering to software. Eisner (2002) adopts the International Council on Systems Engineering (INCOSE) definition of systems engineering as an interdisciplinary approach and means to enable the realization of successful systems.

When different process models are in place within developer groups, say for systems engineering and software engineering of an organization, the organizations will have communication problems, be unable to improve their processes, and if the combined performance of one advances beyond the other in capability, then the problems are even more profound (Johnson, 1998).

In 2002, the SEI released a single integrated capability model for systems engineering and software engineering, integrated product and process development and supplier sourcing. The new model, Capability Maturity Model Integrated (CMMI), is intended to improve organizations' development and maintenance of products. The CMMI will eventually replace the SEI's Software Capability Maturity

Process Management
Organizational Process Focus
Organizational Process Definition
Organizational Training
Organizational Process Performance
Organizational Innovation and Deployment

Engineering
Requirements Management
Requirements Development
Technical Solution
Product Integration
Verification
Validation
Validation

Project Management
Project Planning
Project Monitoring and Control
Supplier Agreement Management
Integrated Product Management
Risk Management
Qualitative Project Management
Integrated Teaming

Support
Configuration Management
Process and Product Quality Assurance
Measurement and Analysis
Causal Analysis and Resolution
Decision Analysis and Resolution
Organizational Environment for Integration
Organizational Environment for Integration

Model (Phillips, 2002). In the integrated model (SEI, 2002), CMMI, the categories and processes are:

One purpose of the CMMI was to evolve the software CMM while integrating the best features of the systems engineering capability models. The combination of the practices of the models into one single framework required more than just combining practices because of differences in interpretation, focus, and terminology. Compromises and intentional inefficiencies were required in order to integrate these models.

With the arrival of the CMMI, a wider continuum of the product life cycle has been targeted for possible enhancement, no longer limiting process improvement only to the development of software. This integrated approach provides a reduction in the redundancy and intricacy resulting from the use of multiple, separate process improvement models. For organizations that wish to assess their process improvement efforts against multiple disciplines, the CMMI provides some economies of scale in model training and assessment training. This one evaluation method can provide separate or combined results for the fields of software and systems engineering. Furthermore, software organizations can also focus on the amplifications for software engineering within the engineering shared process areas and take advantage of any systems engineering amplifications that are helpful. Although still subject to debate, a distinction is made between base and advanced engineering practices as model constructs. Adopting the continuous representation of CMMI not only forces software organizations to define business goals and choose process areas that should be implemented first to focus on these goals, but it also forces companies who are choosing a new subcontractor to do the same. One of the claimed benefits of a staged representation is that it facilitates comparisons among organizations (Shrum, 2000). While it may simplify comparisons, it does so at the loss of additional details. Using a continuous representation, the comparison can be done based on the process areas that are judged by the organization as important rather than simply comparing the organization's maturity score. When using a continuous representation, there is less likelihood that organizations will try to attain a specific level without reasons within their business to do so. It provides an incentive to address processes that would have the greatest impact on their business goals.

FUTURE TRENDS

Despite anticipated problems, bringing systems engineering best practices into the established software process improvement models is expected to be very beneficial. Boehm (1994) reminds us that an important reason to overcome or bridge these differences is to establish an adequate supply of people who can deal with complex systems problems. The Bureau of Labor Statistics (1997) estimates of anticipated growth in information technology jobs, shown in Table 2, provides further support for this concern.

Table 2. Anticipated employment growth 1996-2006

Type of Job	1996 Employment	2006 Employment	% Change
Database Administrators and Computer Support	212,000	461,000	118%
Computer Engineers	216,000	451,000	109%
Systems Analysts	506,000	1,025,000	103%
Data Processing Equipment Repair	80,000	121,000	52%
Engineering, Science, and Computer Systems *Managers*	343,000	498,000	45%

The final job type, managers, is a significant concern addressed by Jerry Weinberg in an interview (Layman, 2001). Weinberg explains that the software development problems are growing faster than individuals' levels of competence. Moreover, he asserts that the current state of practice is one where we need to apply a few fundamentals (e.g., requirements, reviews, configuration management); that is, things known to be useful, but not adopted in the sense of consistent application. Soloman (2002) provides guidelines for using the CMMI to improve earned value management.

It has been suggested that the systems engineering—hardware engineering interfaces have matured nicely over many years, but that the systems engineering—software engineering interface is not as mature as the various hardware engineering interfaces.

Meanwhile, the dependency on the systems engineering-software engineering interface has increased faster than it has matured. Specific concerns by discipline include:

The State of Systems Engineering

- Most successful projects rely on expertise established with similar systems.
- Lack of documented processes makes repeatability difficult.
- Development efforts for unprecedented or significantly different systems often encounter problems.

The State of Software Engineering

- The brief history of software development has been filled with problems of cost overruns, schedule slippage, and failure to achieve performance goals.
- Systems are increasingly dependent on software, yet hardware typically gets the most visibility.

CONCLUSION

Although Pierce (2000) suggests that the CMMI is more of a merged model than an integrated one, which may serve to prolong the separation of the disciplines, Rassa (2001) summarizes the benefits of the CMMI project as follows:

- Common, integrated vision of improvement for all organizational elements;
- Means of representing new discipline-specific information in a standard, proven process improvement context;
- Efficient, effective assessment and improvement across an organization's multiple process disciplines;
- Reduced training and assessment costs.

Although many software-only organizations remain adamant that they do not do systems engineering, all software must run on some computer system, and interface with others. This perceived separation of concerns exacerbates the difficulties associated with hardware/software/system tradeoff decisions, which are further complicated by terminology differences and disparate mental models. Interpretive guidance for CMMI implementation has been provided (e.g., Chrissis et al., 2003)

However, the integration potential of the CMMI[SM] can allow the system and software engineering communities to get the most out of their similarities. The CMMI[SM] allows organizations to tailor the model to mesh with their own mission and goal statements as well as their business objectives. Each individual project can use CMMI[SM] mod-

els for individual disciplines and discipline combinations because the architecture of the CMMI[SM] does not force the employment of every discipline for every organization implementing it. Before the CMMI[SM], the systems engineering models shared many of the same principles as the software version of CMM, but were written to address the needs and terminology of the systems engineering community. Because the CMMI[SM] includes the common and shared elements and best features of both software and system engineering together with discipline-specific elements, an organization can generate integrated capability maturity models or discipline-specific capability models. With CMMI[SM], an organization can still capitalize on these similarities and improve the efficiency of and the return on investment for process improvement. The resulting integrated capability models will adapt to an organization's business purposes.

The concept of an architecture continues to serve as a theoretical link for both the software/system tradeoffs and the integration of process improvement efforts. While respecting the legitimate differences in areas such as reliability testing, it is important to sustain the hope that overlapping or underlying theories will emerge regarding areas of common concern such as: requirements, security, safety, and performance.

In order to achieve true integration of software and system engineering practices into one process improvement model, the remaining differences of terminology and model construction have to be addressed. These two communities have well-developed disparate languages and methodologies that are reflected in their different origins, models and perspectives; differences that have become entrenched in their organizational cultures. With the adoption of an integrated process improvement model, an organization can assess both software and systems engineering functions, reduce conflict and increase consensus.

REFERENCES

Boehm, B. (1994, July-September). Integrating software engineering and system engineering. *The Journal of INCOSE, I*(I).

Carter, L., Graettinger, C., Patrick, M., Wemyss, G., & Zasadni, S. (2002). *The road to CMMI: Results of the first technology transition workshop.*

Chrissis, M., Wemyss, G., Goldenson, D., Konrad, M., Smith, K., & Svolou, A. (2003). *CMMI® interpretive guidance project: Preliminary report.*

Eisner, H. (2002). *Essentials of project and systems engineering management.* New York: John Wiley & Sons.

Goldenson, D., & Gibson, D. (2003). *Demonstrating the impact and benefits of CMMI®:* An update and preliminary results.

International Council on Systems Engineering (INCOSE). Retrieved February 2002, from *http://www.incose.org/*

Johnson, K.A., & Dindo, J. (1998, October). Expanding the focus of software process improvement to include systems engineering. *Crosstalk: The Journal of Defense Software Engineering,* 13-19.

Layman, B. (2001, September). An interview with Jerry Weinberg. *Software Quality Professional, 3*(4), 6-11.

Leibrandt, R. (2001, April 22). A guide to the systems engineering body of knowledge (SEBoK). Retrieved February 2002, from *www.incose.org/orlando/sebok/attach/sebok_text.doc*

Phillips, M. (2002, February). CMMI version 1.1: What has changed. *Crosstalk: The Journal of Defense Software Engineering,* 4-6.

Pierce, B. (2000, July). Is CMMI ready for prime time. *Crosstalk: The Journal of Defense Software Engineering.*

Rassa, B. (2001, March 13). Beyond CMMI-SE/SW v1.0. Software Engineering Institute. *http://www.sei.cmu.edu/cmmi/publications/sepg01.presentations/beyond.pdf*

Shrum, S. (1999, December). Choosing a CMMI model representation. SEI Interactive. Retrieved July 17, 2001, from *http://www.stsc.hill.af.mil/crosstalk/2000/jul/shrum.asp*

Software Engineering Institute. (2002, March). *Capability maturity model integrated (CMMI), version 1.1.*

Soloman, P. (2002). *Using CMMI® to improve earned value management.*

KEY TERMS

Capability Maturity Model (CMM): Contains the essential elements of effective processes for one or more disciplines. It also describes an evolutionary improvement path from ad hoc, immature processes to disciplined, mature processes with improved quality and effectiveness.

Process Architecture: Describes the ordering, interfaces, interdependencies, and other relationships among the process elements in a standard process. Process architecture also describes the interfaces, interdependencies, and other relationships between process elements and external processes (CMMI).

Process Improvement: A program of activities designed to improve the performance and maturity of the organization's processes, and the results of such a program.

Quality: The ability of a set of inherent characteristics of a product, product component, or process to fulfill requirements of customers.

Software Engineering: The software engineering discipline covers the development of software systems. Software engineers focus on applying systematic, disciplined, and quantifiable approaches to the development, operation, and maintenance of software.

Systems Engineering: The systems engineering discipline covers the development of total systems, which may or may not include software. Systems engineers focus on transforming customer needs, expectations, and constraints into product solutions and supporting those product solutions throughout the product life cycle.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2551-2556, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

The Software Industry in Egypt as a Potential Contributor to Economic Growth

Sherif Kamel
The American University in Cairo, Egypt

INTRODUCTION

During the 1960's computing was introduced to Egypt. Its use and applications was limited to the government and the public sector. During the 1980's the introduction and diffusion of computing was widespread following the global personal computer evolution. Personal computers effectively affected organizational development and growth due to the continuous developments in the information technology industry and caused by increasing hardware penetration, software innovations, and the build-up of the telecommunications and information infrastructures. This chapter assesses the recent developments in the software industry in Egypt, especially post to the establishment of the ministry of communications and information technology late in the 1990's as a major building block of the information technology industry and a possible active contributor to economic development at large through a strong, quality and much needed export-oriented software industry that could have concrete implications on the economy.

BACKGROUND

Although computing started in Egypt in the 1960's, it was only in 1985 that the active role played by the government caused a change in the way information technology was perceived as a vehicle for socioeconomic development and a tool to improve the decision making process (Kamel, 1999). This change was accelerated by the continuous development of new tools and techniques that had direct and concrete effects on socioeconomic development. Furthermore and after the establishment of a ministry for communications and information technology, Egypt's information society initiative (EISI) was launched in 2001 to provide a broad perspective on the strategic plan for information and communication technology diffusion in Egypt (Kamel, 2005). Therefore, it is perceived that the way developing countries will manage the computer driven process of change will influence whether its socioeconomic development goals will be promptly achieved. This will be bound to the continuous ability to invest in emerging technologies, the provision of skilled human resources and the completion of a state-of-the-art information and communication technology infrastructure. Many researchers have identified information technology as the combination of information, computing and communication technologies that through convergence could help the development process (The American Chamber of Commerce in Egypt, 2001).

Today, with the evolution and diffusion of the Internet, the integration of these technology elements is invaluable to societies around the world and strongly contributing to globalization. Moreover, newly evolving economies in the 21st century are mainly dependent on hardware to process information; communication that acquire and distribute information and software which helps manage the whole process. The importance of information technology has been greatly emphasized in most developing countries in a deliberate effort to ensure that they do not lag behind, with an emphasis on localization and adaptation to local community needs. In most developing nations, the government has played the most important role in the diffusion of information technology being the largest user of computers (Moussa & Schware, 1992) and through its policies, laws, and regulations it still exerts the largest influence on the diffusion of information technology throughout different organizations (Nidumolu & Goodman, 1993). Such concept has gradually started to change throughout the last decade through massive deregulation of the information and communication technology sector with a focus on telecommunications. For example, in the case of Egypt, the government used to be the primary user of information technology with an accumulated market share of 25% (Ministry of Communications and Information Technology, 2006). However, recently increasing use of information and communication technology has been greatly felt in the banking, health, employment, trade, and local administration and education sectors among others (CIT Egypt, 2006).

Since the establishment of Egypt's information society initiative in 2001 and its amendments in 2003 to cater for the changing local and global market needs, and through a public-private partnership (PPP), there has been a growing and effective role being played by the private sector through a win-win formula that is applied on a variety of information and communication technology projects that relates but not limited to (a) PC for every home, (b) free-subscription Internet model, (c) information technology clubs, (d) broadband diffusion, (e) mobile penetration, (f) electronic

government institutionalization, and (g) software incubation and development (Kamel & Ahmed, 2006).

Software as an integral element of the information and communication technology industry is attracting increasing attention of developing nations after many years where hardware was really the focal point. Moreover, it is important to note that the software industry is an excellent setting to understand the features of the knowledge-based economy (Seleim, Ashour & Bontis, 2004). It is also important for knowledge acquisition and transfer since the essence of software development is pure knowledge and the fact that 95% of software business is intangible capital (Grant, 2000; Hoch, Roeding, Purkert, Linder & Muller, 2000). India is a classical example with 6.5 billion US dollars industry and increasing steadily (Nasscom, 2000). In Egypt, the ICT market has been steadily increasing since the late 1990's (Economic News Bulletin, 2003). The software market according to the Egyptian Software Association (ESA) was estimated at 50 million US dollars in 1998 mainly locally developed with a focus on tailor-made applications and with an annual growth rate of 35% locally and expectation of 200% increase in exports annually (Harvard Consulting Group, 1999). During that time, the total market was estimated at being 140 million US dollars however with massive expectations of growth in the following years. It is also important to note that since the late 1980's, Egypt has been playing a leading role in software publishing in the Middle East and 80% of its software exports are regularly going to countries in the Gulf region and mainly Saudi Arabia (Arab Human Development Report, 2002). Table 1 demonstrates a number of elements that relates to the software industry in Egypt showing the different stakeholders, market outreach and the various applications whether customized or packaged.

IT FOR DEVELOPMENT

Egypt is the cradle of an ancient civilization dating back to 3000 BC. With a population of about 72 million, it is the most populous country in the region (Ministry of Communications and Information Technology, 2006). About 28% of its population is enrolled in education programs (schools and universities education), 58% are under the age of 25 and 19 million represent it workforce; around 6 million are working for the government sector (Information and Decision Support Center, 2003). Egypt is trying to expand its industrial base and modernize itself technologically with agriculture accounting for 17% of the gross domestic product, industry for 32% and a large service sector (51%) mainly built around tourism and transportation. A comprehensive economic reform program was implemented that enabled its current economic growth rate to stand at 6.1% annually with an inflation of 5.3% (Economic News Bulletin, 2003). Estimates show that unemployment is standing at 11% and the labor force is growing at around 2.7% annually (Information and Decision Support Center, 2003).

The government of Egypt is more determined than ever to build-up the national infrastructure and keeps pace with the IT evolution worldwide. Since 1999, the concerned ministry in collaboration with different stakeholders embarked on a master plan to build Egypt from an information and communication technology perspective that is based on the fact that as an emerging market, Egypt has already made considerable achievements in terms of economic development and is ready to move aggressively into the global market and the only vehicle to realize that objective is through a state-of-the-art information and communication technology infrastructure (Osman, 2000). There is no doubt that the information technology sector can act as a driving force behind a potential new economy for Egypt. With the growing size of the global

Table 1. Software industry elements

Industry Characteristics	
Stakeholders	Government Private sector Public sector Civil society organizations Individuals
Market Outreach	Europe Arab countries North America
Applications (customized and packaged)	Functional applications Educational applications Cultural applications Arabization of applications

software market, Egypt has promising potentials to compete in the middle market segment of companies because it cannot compete with the likes of India on the basis of prices. It can only compete based on the relatively low price of labor compared with the higher level of value added, business vision, and innovation (Rizk, 2002).

Developing the Software Industry

The software industry in Egypt has been gradually growing over the last few years. It is diverse and heterogeneous in nature with the presence of local vendors and multinationals like most mature markets. Most of the software development companies provide training services to support their products and clients. Overall, the number of information and communication technology companies grew since 1999 from 312 to 1,773 marking over 560% and employing over 35,000 compared to 6,000 in 2000 including managers, programmers, experts, consultants, and project managers mostly involved in the development and delivery of information systems to local and international markets (Ministry of Communications and Information Technology, 2006). The software industry is divided into four categories including (a) software tools, (b) packaged applications, (c) tailored applications and multimedia applications and (d) Arabization of applications. Software companies range in size between 1 to 5 staff member start-ups through to relatively mature firms with around 50 to 150 employees. The majority of firms are located in and around Cairo or Alexandria. However, recently, some of the new start-ups were located in the new industrial areas to benefit from the tax holidays they offer.

It is important to note that a growing number of firms since 2001 have been involved in a variety of business process outsourcing activities leading to expanding the outreach of these firms beyond the national borders and contributing to a healthy export-oriented software industry where playing a major part in such industry is companies like ITWorx, MNS, DMS, and Raya that are taking the lead among others. As an information technology-producing sector, the software industry can act as a driving force behind a potential new economy for Egypt. It is important to note that because of the cost structure of the industry that relies on information that is expensive to produce and inexpensive to reproduce; there are a variety of opportunities for economies of scale and increasing return on investment (Rizk, 2002). Respectively, given the important role of human capital in shaping an effective and rewarding software industry and Egypt's well-educated workforce it is important to highlight the importance of expanding investments in the information and communication technology sector and especially in the software industry.

Egypt's dream is to become a leading software exporter in the region with a global outreach which is relatively realistic considering the boom in the software industry in Egypt and the potentials in the regional marketplace with a growing need for Arabization in a market that totals over 300 million in population (International Telecommunication Union, 2006). The market for Arabized software is large in Egypt but there is also a great potential elsewhere throughout the Arabic speaking world that can be served with language-specific software produced in Egypt. In 2003, Egypt has put together a plan to lift its software exports to 2 billion US dollars within 5 years (AME Info, 2003). The plan presented by the Union of Egyptian Industries builds on the potential of the Egyptian software industry mainly in the domain of localization, customization and consulting although there were concerns for high taxation on IT imports as an obstacle to promoting a competitive software industry and a deterrent to investments in software research and development.

Capitalizing on over 160 software development houses and around 10,000 developers, Egypt has all the ingredients to become a regional leader in the domain of value-added services and localization (AME Info, 2003). The plan aimed at a cut on import taxes from 25% to 3% in order to make the industry more attractive and profitable. Lowering custom duties on imports would directly result in boosting the software industry by 300 million US dollars in the first year due to the decrease in cost of production. More importantly, the plan calls on the government of Egypt to include the software industry as a priority in trade agreements signed between Egypt and other countries.

While these numbers are not impressive if compared with more developed software industries, they do provide a foundation from which to start a serious development of the industry. Moreover, the ministry of communication and information technology has embarked since 2001 on a multiphases plan to train around 30,000 graduates to enter the labor market in the software industry and it the ministry has also allocated around 9 million US dollars for professionally qualifying them (Ministry of Communications and Information Technology, 2006). Egypt realizes that it must devote considerable resources to educating and training IT professionals to reach a reasonable critical mass that could act as agents of change and help create a healthy and strong software development platform that is an integral element of a high-tech industry and a contributing factor to the development of a knowledge-oriented and productive society. The plan aims to continuously increase and diversify the training of fresh graduates as well as IT professionals that could represent the core of the development of a high-tech industry (Osman, 2000). The expectation for growth in the domestic marketplace for IT products and services is expected to be in the range of 35% for services and products on annual basis for the many years to come.

The industry distribution channels in Egypt are still relatively underdeveloped with around 63% of software sales without intermediaries; 50% of tailored applications are sold

bundled with niche products and services. Moreover, software sales through system integrators are low because of limited subcontracting, technical cooperation, and interchange of skill and specialization's between local companies. Finally, function-oriented software is sold primarily through dealers (Seleim, Ashour & Khalil, 2005). With regard to software demand, the government purchases generate 25% of total software revenues, making it the largest demand segment with two major purchase determinants, which are quality and after-sales service for fear of system failure with cheaper systems (The American Chamber of Commerce in Egypt, 1998). However, only 6% of revenues are from sales to small office and homes, which is in part due to the widespread piracy rate of 86% that plagues this segment. This figure is gradually decreasing due to the newly introduced laws against violators of software piracy laws. Also, the number of software applications sold to households is increasing due to the boom in PC sales for household usage and the

Table 2. Software industry in Egypt SWOT analysis

Strengths	**Weaknesses**
▪ Well-educated graduates interested in ICT coupled with entrepreneurial skills ▪ Computer science schools in all national and private universities ▪ Basic and advanced training programs in IT management ▪ Growing technical skills in ICT ▪ Low and competitive labor costs ▪ Good command of English for dealing with overseas customers (other languages include French, German, and Spanish) ▪ Same time zone advantage with Europe and provides a second-shift for the United States (outsourcing services) ▪ Geographically well-located from most African and European cities and some Asian cities making it a hub for trading ▪ Encouragement of the government by facilitating procedures and logistics related to the software industry	▪ Technical skills are too broad and thin ▪ Lack of sufficient expertise in any one technology ▪ Software companies spend around 6 months to turn graduates into productive contributors ▪ Lack of project management, marketing, and managing start-ups skills ▪ Limited local demand for software ▪ Lack of management recognition to the value of using IT as a business vehicle ▪ Effective role of government still limited ▪ Infrastructure level and cost is high compared to the capacities of manufacturers and beneficiaries ▪ Government regulations needs to be firm and enforced ▪ Perception that software has little intrinsic value among commercial and government customers
Opportunities	**Threats**
▪ Creation of software business incubators such as the smart village model ▪ More proactive role played by educational institutions and training centers ▪ Internships and scholarships from software vendors both local and multinationals ▪ Promotional role played by software associations to activate the role of software development companies ▪ Government support role needs to be more at the macro and micro levels ▪ Changes in tax treatment, reduction on telephone tariffs and the introduction of new intellectual property laws ▪ Penetration of PC in homes and businesses is increasing	▪ Intellectual copyright violations ▪ Piracy rates are relatively high ▪ Lack of understanding for software products and their implications on business development ▪ Lack of market research in domestic and overseas markets that Egyptian companies could target ▪ Lack of financial support to the industry ▪ Limited distribution skills to serve the international software market

spread of Internet among younger generations due a few public-private partnership programs addressing the Internet diffusion and PC penetration in the society.

The competitive advantages of Egypt's domestic software production environment have attracted numerous international producers to subcontract programming of tailored applications. The industry was further boosted in 2004 with the establishment of the Information Technology Industry Development Authority (ITIDA) a governmental entity established through Law 15 and aiming at paving the way for the diffusion of the e-business services and supporting an export-oriented IT sector. This was coupled with a number of strong attributes that places Egypt as a potential center for offshore IT services such as (a) favorable technical staff and infrastructure costs, (b) language capabilities, (c) improving infrastructure, (d) pro-business governmental reforms, (e) strong government support for the IT sector and (f) a strategic geographic location.

With the presence of over 50 Internet service providers; there is an expected significant growth in services and software applications that are Internet-based contributing to the new economy with its driving forces and new rules. Additionally, there is an expected increase in the development of applications of a number of key sectors in the economy including the financial, petroleum, tourism, and health sectors (Loch, Straub & Kamel, 2000).

Software Industry SWOT Analysis

The software industry in Egypt is better analyzed through a SWOT analysis to be able to understand where it stands and where it is heading with an overview on its strengths and weaknesses and an identification of the opportunities available and the threats faced. Table 2 demonstrates the software industry in Egypt SWOT analysis.

FUTURE TRENDS

Based on the assessment of the software industry in Egypt, an action plan needs to be formulated in an attempt to capitalize on the opportunities available and overcome the challenges in the market. The plan can assist Egypt in improving its software industry performance and cope with the global trends and drivers especially in the domain of outsourcing. This could include (a) identifying and penetrating international target markets to realize growth, (b) investing in people to leverage the capacities of the key building block in the industry, (c) introducing incubator programs to link the industry to educational institutions and the government, (d) improving the infrastructure to ensure that firms have the vehicle to operate their business; and, (e) increasing government support role to demonstrate the value of software to Egypt's future (Kamel, 2003). Moreover, during 2005, the ICT sector, through public-private partnership programs have been focusing more on encouraging innovations and developments in ICT applications and in the development of the software industry with an emphasis on exports. Therefore, the government established four centers of excellence in data mining and computer modeling; wireless technologies, mobile and electronic services, and electronic design.

CONCLUSION

Egypt has an excellent opportunity to develop the ingredients of a small but effective software industry. The industry has high potential to be developed way beyond the current levels. All the efforts and support enabled by the government and backed by the industry, financial institutions and the educational system will ultimately determine the level of success of the industry making it profitable and active contributor to business and socioeconomic development. There is a wealth of opportunities for Egypt to improve all aspects of development for the software industry. Egypt could dramatically increase the level of revenue and growth in the software business with a relatively small investment and focus to be able to attain the levels of achievement realized in nations such as Ireland and India. Increasingly, Egyptian software companies are establishing partnerships with global software vendors offering complete and integrated solutions as well as customized applications designated for vertical markets. For example, since the establishment of the smart village, there has been a number of strategic alliances developed with the likes of IBM through the establishment of a software development center training 400+ skilled IT developers annually, Mentor Graphics established a design center training 170+ engineers, Intel established a regional platform definition center and a regional software development lab, HP established an imaging and printing technology center, Oracle established a global support center, Microsoft established a developer support center for software and a software innovation center, and finally Cisco established a core competency e-learning institute.

Egypt has all the resources to become a leading player in the regional industry and even to establish a global reputation as a software development center due to the growing presence of its manpower, skills, capacities and educational infrastructure. Such industry can benefit economic growth and work as a vehicle for knowledge acquisition and transfer which is an important element for societal development and building competitive advantages. The start could be the model of the smart village that could be replicated in different locations in Egypt. The smart village is a business park on 300 acres of land with a capacity of 67 companies' buildings, hosting 30,000 job openings with investment incentives like tax exemptions, high-speed network, host of a diversity of multinationals such as Microsoft, Alcatel,

Vodafone, and Ericsson among other as well as Egyptian companies and technology incubators; there are also areas available for SMEs and a plan to establish a financial district in which the Egyptian Stock Exchange and a number of financial institutions will be hosted.

RECOMMENDATIONS

Based on the development of the software industry worldwide and the potentials for the software industry in Egypt; there is a number of issues that need to be restructured to benefit from opportunities available both in Egypt and other countries in the region. These issues relate to (a) having a better link and coordination between the industry at large and academic establishments to cater for the growing needs in terms of human resource capacities as well as to enable an evaluation and feedback mechanism that relates both stakeholders; (b) enhancing the educational system to generate capacities ready to deliver the needs of local and international markets; (c) improving the salary scale in the industry to match other regional and relatively international markets to minimize the implications of brain-drain situations, (d) setting-up a regulatory authority to regulate the software industry, and (e) establishing a solid presence in the overseas markets through offshore business development and contact centers and benefit from a market for business development outsourcing that should reach 500 billion US dollars by 2008 (Ministry of Communications and Information Technology, 2006). The software firms in Egypt possess many elements that represents the platform for a growing, successful and competitive industry, however to have enable healthy and strong industry it is recommended to implement a number of steps that includes developing a strategy for the protection of intellectual property rights, providing a process for documentation of business practices and needs in the marketplace, and focusing on the needs of the regional markets for Arabic software and value-added applications such as e-government and Web-enabled government services.

REFERENCES

AME Info (2003). *Five-year plan to lift Egypt's software Exports to USD 2 billion.* Retrieved June 11, 2008, from www.ameinfo.com

Arab Human Development Report (2002). *United Nations development program.* Jordan: Regional Bureau for Arab States.

CIT Egypt (2006). Retrieved June 11, 2008, from www.citegypt.com

Economic News Bulletin (2003). Retrieved June 11, 2008, from www.economic.idsc.gov.eg

Grant, R. M. (2000). Shifts in the world economy: The drivers of knowledge management. In C. Despres & D. Chauval (Eds.), *Knowledge horizons: The present and promise of knowledge management.* Butter-worth-Heinemann

Harvard Consulting Group (1999). *Sector assessment of the Egyptian software industry*. Cairo: Egyptian Software Association.

Hoch, D. J., Roeding, C. R., Purkert, G., Linder, S. K., & Muller, W. R. (2000). *Secrets of software success: Management insights from 100 software firms around the world*. Boston: Harvard Business School Press.

Information and Decision Support Center (2003). Retrieved June 11, 2008, from www.idsc.gov.eg

International Telecommunication Union (2006). *The next India? Egypt's software dream*. Retrieved June 11, 2008, from www.itu.int

Kamel, S. (1999). Information technology transfer to Egypt. In *Proceedings of the Portland International Conference on Management of Engineering and Technology (PICMET), Technology and Innovation Management: Setting the Pace for the Third Millennium*, Portland, Oregon.

Kamel, S. (2003). The implications of the digital economy on a growing digital divide in developing nations. In *Proceedings of the 8th American University in Cairo Research Conference on Globalization Revisited: Challenges and Opportunities* (pp. 60-70). Cairo, Egypt.

Kamel, S. (2005). Assessing the impacts of establishing an internet cafe in the context of a developing nation. In *Proceedings of the 16th Information Resources Management Association International Conference on Managing Modern Organizations with Information Technology* (pp. 176-181), San Diego, California.

Kamel, S. & Ahmed, H. (2006). The impact of eReadiness on eGovernment in developing nations - Case study of Egypt. In *Proceedings the 17th Information Resources Management Association International Conference on Emerging Trends and Challenges in Information Technology Management*, Washington, DC.

Loch, K. D., Straub, D. W., & Kamel, S. (2000). Use of the internet: A study of individuals and organizations in the Arab world. In P*roceedings of the First Annual Global Information Technology Management World Conference* (p. 191). Memphis, Tennessee.

Ministry of Communications and Information Technology (2006). Retrieved June 11, 2008, from www.mcit.gov.eg

Moussa, A. & Schware, R. (1992). Informatics in Africa: Lessons from World Bank experience. *World Development, 20*(12).

Nasscom (2000). *India's national association of software and services companies*. Retrieved June 11, 2008, from www.nasscom.org

Nidumolu, S. R. & Goodman, S. (1993). Computing in India: An Asian elephant learning to dance. *Communications of the ACM, 236*(4).

Osman, H. (2000). Editorial, Business Today, February.

Rizk, N. (2002). Information technology and growth: Will the software industry lead Egypt into a new economy? In *Proceedings of the Middle East Economic Association, Topics in Middle Eastern and North African Economies Journal, 4*.

Seleim, A., Ashour, A., & Bontis, N. (2004). Intellectual capital in Egyptian software firms. *The Learning Organization, 11*(4/5), 332-346.

Seleim, A., Ashour, A., & Khalil, O. (2005). Knowledge acquisition and transfer in Egyptian software firms. *International Journal of Knowledge Management, 1*(4), 43-72.

The American Chamber of Commerce in Egypt (1998). Information technology in Egypt. Business Studies and Analysis Center, June

The American Chamber of Commerce in Egypt (2001). Annual general meeting agenda. Ministry of Communications and Information Technology, May

KEY TERMS

Arabization: The transformation of software applications into the Arabic language in terms of usage as well as interface to be able to cater for a community that stands in 2003 at around 300 million people.

Building Blocks: Reflects all the critical success factors of the information technology industry and that include: hardware, software, human resources "humanware", networking and information.

Business Process Outsourcing: Means the use of outsourcing mechanisms locally or offshore for the design and/or development of software applications.

Diffusion of Information Technology: Reflects the spreading of information technology concepts among the society of implementation whether that could be within an organization or within the community at large.

Government-Private Sector Partnership: The teaming of different entities in the government and the private sector to realize a change and a transformation in the development of information technology at large and in the software industry in specific.

Incubator Programs: A form of collaboration usually between the industry, corporations, and the business community and the educational sector aiming at identifying industry and market needs, catering for these needs and creating employment opportunities for the society especially, young graduates.

Informatics Projects: The projects that involve in any way possible the use, design, delivery, implementation, and management of information technology irrespective of the element involved including software, hardware, and so forth.

Information Technology Industry: The accumulation of all elements of information technology design, delivery, and management.

Smart Village: Model technology park represented by a landscaped development usually comprising of high specification office space and retail developments, designed to encourage localization of high technology companies such as information technology and software development thereby giving each the benefit of economies of scale; usually located outside the city areas as these are quite land intensive in nature.

Software Industry: Focuses on the needs of the software development industry in terms of infrastructure, know-how, capacities, and development.

Tailored-Applications: Applications based on industry or organizational needs to complement the off-shelf software applications available in the marketplace.

Software Reuse in Hypermedia Applications

Roberto Paiano
University of Lecce, Italy

INTRODUCTION

Hypermedia applications were, at the beginning, hand-coded pages with "ad-hoc" links. This production method was acceptable until a few pages had to be produced, but it became rapidly unmanageable when several hundreds of pages with complex interactive objects had to be considered. In particular, two interwoven problems rapidly became relevant: how to ensure the "usability" of modern large hypermedia-applications (Garzotto, Matera & Paolini, 1999), and how to improve the efficiency of its production/maintenance process.

In good hypermedia applications, in fact, the reader should be able to effectively exploit the information contained in the application: that is, he or she should be able to quickly locate the objects of interest, to understand the inner structure of the objects and to easily navigate from one object to another. Several factors concur to the achievement of usability: one of the most important is to have a good structuring of the information objects and a good structuring of the navigation patterns.

BACKGROUND

Several authors have recently proposed the adoption of design models (Garzotto, Mainetti & Paolini, 1995; Isakowitz, Stohr & Balasubramanian, 1995; Schwabe & Rossi, 1995)and design patterns (Rossi, Schwabe & Lyardet, 1999), in order to improve the quality of hypermedia applications, at least for those aspects concerning structure and navigation. Other authors (Conallen, 1999; Schwabe & Rossi, 2000) have proposed the use of object oriented paradigm to model this kind of application, but the navigation structures are more simple. Design models provide, in fact, the primitives that allow structuring the information objects and the corresponding navigation patterns along regular and systematic features, improving consistency, predictability (for the user), robustness of the design, and therefore improving usability. The ancestor of these models can be traced to HDM (Garzotto, Paolini & Schwabe, 1993) and its evolution: W2000 Model (Baresi, Garzotto & Paolini, 2000).

The adoption of W2000 to design the internal structure and the navigational features of hypermedia applications is desirable for three reasons:

- resulting applications are usable;
- the production process can be decomposed into subproblems easy to manage;
- the application model can be "executed" by a suitable "interpreter" to create the application pages in a way that is independent from the specific application.

Furthermore, in several real-life projects we encountered the problem of dealing with application families. An application family is a set of applications sharing (part of) the content and also (part of the) conceptual design. The problem of application families is the typical situation where the application owner, after a successful first application, needs a second one very similar to the first one. At first it seems a simple problem of "reuse" of content: use the same pictures, use the same texts, use the same data, and so forth (Garzotto, Mainetti & Paolini, 1996). After a while it becomes apparent that not only content, but also (pieces of the) conceptual structure must be "reused". Then comes a third "similar" application, and so on. So, the truth emerges: the designer should have started from the beginning having in mind a family of applications, knowing that several specific applications could have resulted from it. In other words the designer should have optimized the activity of "carving out," from a family, a specific application, for a specific need.

Therefore it became clear that the design process, the design model and the design support system should adopt the notion of family of applications. Such kind of activity is made easily using a structured model.

BRIEF DESCRIPTION OF W2000 METHODOLOGY

The methodology was developed by the UWA Consortium (UWA), and specifically by Polytechnic of Milan.

W2000 methodology assumes that it is essential to make a clear distinction between the different aspects of the application that need to be observed during the design phase, in order to make the design itself a structured and easily controllable process, and to obtain clear modeling, suitable for different users and delivery devices.

After the Requirements Analysis phase, guided by a goal-oriented approach, the methodology suggests a sequence of steps that may be briefly summarized as follows:

- Information Design: the goal is to describe the information that the application is going to deal with, giving it a structured organization from the user's point of view.
- Navigation Design: this reconsiders the information and its organization from the viewpoint of its fruition, defining the navigational paths the user can follow.
- Publishing Design: the results of the previous steps must be complemented with considerations on presentation and organized into "pages" and "fruition units".

According to the previous description, a database can store the application components described by the model, and then a run-time engine can extract those components from the database to display it to the reader. This kind of engine, named WAPS, is application independent, so it is really reusable and it may be defined as a W2000 methodology Interpreter. It is the last evolution of a family of navigation engines according to the evolution both of methodology (Bochicchio & Paiano, 2000; Paiano & Pandurino, 2003) and available technology (Bochicchio & Paolini, 1998; Bochicchio, Paiano & Paolini, 1999).

A REUSABLE INTERPRETER

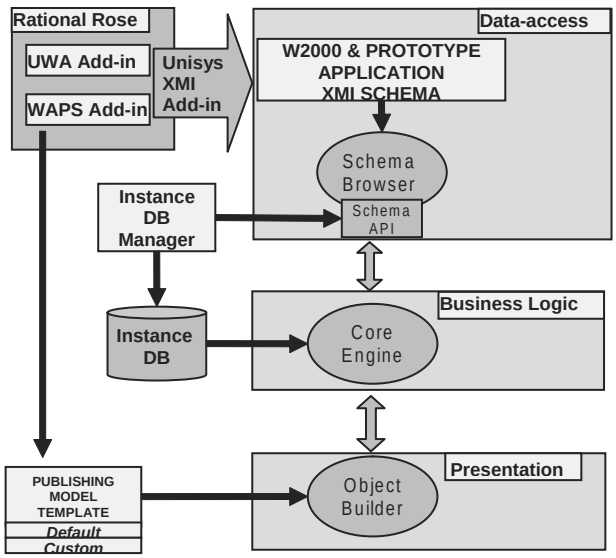
The run-time environment, the WAPS core, has the main task of creating a mock-up application starting from the W2000 model in XMI format.

In accordance with the modular structure of the W2000 methodology and the various aspects of WA, as shown in Figure 1, it is possible to identify a clear n-tier architecture for the WAPS run-time environment. This choice was highly suitable for the W2000 methodology: in order to manage the complexity, each architecture layer manages a single aspect and provides services to the other levels. All the data managed in the modules are in XML format; furthermore, all interaction between modules is in the same format according to the market trend and standard.

It allows the use of transformation parsing techniques like XSL in the visualization and processing phases, also allowing the following of the evolution of methodology.

- Rational Rose Add-in: Rose helps to design the WA in graphic format using standard UML notation, in accordance with W2000 methodology, in order to obtain a "machine readable" description. Another rational rose add-in: "Unisys Rose XML Tools," produced by Unisys, exports the UML diagram into a standard XMI output.
- Schema Browser: This module allows a unique entry point to the WA schema, hiding the complexity in order to manage the XMI in raw mode. The module provides a set of schema APIs (S-API) to navigate through the WA model via W2000 primitives, bypassing the UML MOF used by XMI.
- Core Engine: This module corresponds to the business level for a three-tier application. This module has the task of understanding the requests from the Object Builder, using the S-API of the schema browser to compose the reply schema that will contain the application data taken from the Instance DB. Since this module creates the reply schema, all design customizations take effect at this stage.

Figure 1. WAPS architecture



- Object Builder: This module is the door to WAPS systems: the user request comes in, the prototyped page goes out. The module moves the request to the Core Engine and receives the response in XML-like format. Its main task is to apply a template to make the page visible. WAPS uses the XSLT transformation to obtain HTML or WML pages.
- Instance DB: It is an E-R database that contains the data that will be shown to the user. The E-R schema is derived from the W2000 model; thus the schema is fixed and does not change with the domain of the WA being prototyped.
- Publishing Model Template: It is an E-R database that contains the references to the visualization template to create the page.

FUTURE TRENDS

We are now assisting the revival of MVC (Model View Controller) architecture applied to the Web. There are several frameworks oriented to the prototyping of Web applications, based on the MVC paradigm, but the Model site is not fully specified.

The next step of research in this area is to join the W2000 Model definitions with the Model site of these frameworks.

To achieve this goal, the SET Lab of University of Lecce is developing this kind of interface between W2000 model and Struts framework developed by Apache software Foundation, obtaining very interesting results.

CONCLUSION

The adoption of structured approaches and conceptual models, such as W2000, is an important step in the direction to improve the quality and the reusability of hypermedia applications, reducing at the same time the costs and the time required for their development.

The construction of a reusable engine for hypermedia application, based on the W2000 model, is the logical extension to the model-based approach. The reusable engine, in fact, is able to implement the needed hypermedia software in an application-independent way. The effectiveness of this approach is based on the ability of the model to describe a wide range of complex hypermedia applications.

REFERENCES

Baresi, F., Garzotto, F., & Paolini, P. (2000). *From Web sites to Web applications: New issues for conceptual modeling.* WWW Conceptual Modeling Conference.

Bochicchio, M.A., & Paolini, P. (1998). An HDM interpreter for on-line tutorials. In N. Magnenat-Thalmann & D. Thalman (Eds.), MultiMedia modeling (pp. 184-190). Los Alamitos: IEEE Computer Society.

Bochicchio, M.A., & Paiano, R. (2000). Prototyping Web applications. *ACM Symposium on Applied Computing, Como, (IT),* 978-983.

Bochicchio, M.A., Paiano, R., & Paolini, P. (1999). JWeb: An HDM environment for fast development of Web applications. *IEEE Conference on Multimedia Computing and Systems, Firenze (IT), 2,* 809-813.

Conallen, J. (1999). Modelling Web application achitectures with UML. *Communication of the ACM, 42,* 63-70.

Garzotto, F., Mainetti, L., & Paolini, P. (1995). Hypermedia application design: A structured approach. In J.W.Schuler, N.Hannemann & N. Streitz (Eds.), *Designing user interfaces for hypermedia.* Springer Verlag.

Garzotto, F., Mainetti, L., & Paolini, P. (1996). Information reuse in hypermedia applications. *ACM International Conference on Hypermedia.* Boston: ACM Press.

Garzotto, F., Matera, M., & Paolini, P. (1999). *Abstract tasks for hypermedia usability evaluation.* Technical Report No.03-99. Dept. of Electronics and Information, Polytechnic of Milan.

Garzotto, F., Paolini, P., & Schwabe, D. (1993). HDM: A model based approach to hypermedia application design. *ACM Transactions on Information Systems, 11*(1), 1-26.

Isakowitz, T., Stohr, E.A., & Balasubramanian, P. (1995). RMM: A methodology for structured hypermedia design. *Communications of the ACM, 38*(8), 33-44.

Paiano, R., & Pandurino, A. (2003). From the design to the development: A W2000 based framework, issues and guidelines. *IRMA International Conference,* Philadelphia (pp. 500-503).

Rossi, G., Schwabe, D., & Lyardet, F. (1999). Improving Web information systems with design patterns. *International WWW Conference.* Toronto: Elsevier Science.

Schwabe, D., & Rossi, G. (1995). The object-oriented hypermedia design model. *Communications of the ACM, 38*(8), 45-46.

Schwabe, D., & Rossi, G. (2000). An object oriented approach to Web-based application design. *http://www.telemidia.puc-rio.br/oohdm/oohdm.htm*

UWA Consortium. (2001). *General definition of the UWA framework.* Technical report EC IST UWA Project. *www.uwaproject.org*

KEY TERMS

HDM: Hypermedia Design Methodology, developed by Polytechnic of Milan (Italy). It is a methodology to design hypermedia applications.

Interpreter: The traditional definition in computer science is a program that translates and executes source language statements one line at a time. The meaning in this article is: a program that accesses the model to understand the requested navigation one user action at a time.

W2000: A methodology to conceptually design Web applications, developed by Polytechnic of Milan (Italy) in the UWA project.

WAPS: A reusable W2000 interpreter to prototype Web applications.

Web Application: An application that presents the characteristics and the issues of both hypermedia applications and traditional applications. In other words, this kind of application has the navigational issues of Web sites joint to the traditional operation issues.

XMI: XML Metadata Interchange is a widely used interchange format for sharing objects using XML. It is defined by OMG.

XML: EXtensible Markup Language is a mark-up language much like HTML designed to describe data using your own tags. It is a recommendation of W3C Consortium.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2567-2570, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Solutions for Wireless City Networks in Finland

Tommi Inkinen
University of Helsinki, Finland

Jussi S. Jauhiainen
University of Oulu, Finland

INTRODUCTION

Wireless urban networks can be approached from many perspectives. They are commonly studied on the basis of technology development (e.g., Chao, Uden, & Shih, 2005), business and service support (e.g., Friday, Davies, Wallbank, Catterall, & Pink, 2004; Jenisch, Orlamünder, Köstring, & Brügge, 2005), urban marketing and city image promotion (e.g., Dobers, 2004) or societal use of technology (e.g., Graham & Marvin, 1996; Rao & Parikh, 2003; Palm & Wihlborg, 2006).

This article gives a detailed outline of the provision and condition of public wireless local area networks (WLANs) in Finland. The cases of Oulu, Turku and Helsinki ("Arabianranta" residential area) are presented. These cities are relevant study locations because they have actively participated to the creation processes of public city WLANs. They are also using wireless networks as promotion tools in their image marketing. In addition, Finland has been regarded as one of the top "network ready" nations in the world and was ranked fourth in the latest Networked Readiness Index by WEF (2007) after Denmark, Sweden and Singapore.

BACKGROUND

Several relevant studies on wireless networks have been conducted recently (e.g., Harwit, 2005; Salkintzis, Pavlidou, Fitzek, & Varma, 2005; Zhuang, Gan, Loh, & Chua, 2003). For example, Tang and Baker (2002) analysed metropolitan area wireless networks in three locations in the U.S., including the San Francisco Bay Area, Washington D.C. and Seattle. They provided an extensive quantified analysis of network loading, activity and mobility patterns of data signals.

To generalise the work of Tang and Baker (2002), there are variations in 1) technologies, 2) organisational arrangements and 3) usage preferences related to WLANs. There are several "city networks" that could be studied here. First, several technical solutions exist for different purposes. Private radio access networks (RANs) and cellular wide area networks (WANs) are often deployed by public authorities (e.g., emergency units, police, maintenance) for the purpose of sharing and delivering information from their daily field operations. WLANs are the third typical solution structure for providing a higher bandwidth in outdoor conditions. WLANs are based on standardised industry technologies. Three main standards are used to define the communication protocol between the access point and the client. They are all variants of the IEEE 802.11 standard (802.11a, 802.11b and 802.11g). The (a) and (g) standards provide 54 Mbps rates and the (b) standard, 11 Mbps. However, in practice, the rates are considerably lower because of protocol overheads and distance decay between the client and the access device (Cisco, 2006).

A variety of interest groups are needed to realise a WLAN that provides services to the public, such as residents, tourists and business visitors. These include infrastructure providers, Internet service providers (ISPs), hardware providers and clearinghouse operators. In general, provision of a WLAN can be characterised either as a top-down approach in which network operators charge access fees, or as a bottom-up approach in which end-users are offered a free access on a noncommercial basis (Rao & Parikh, 2003). Thus, the main relationships are business-to-customer (B2C), business-to-business (B2B), business-to-government (B2G) and customer-to-government (C2G). Publicly provided noncommercial networks are often partly commercial, because their production and development usually requires outsourced services by collaborating partners. "Wireless city networks" are collections of networks under a same brand and they differ fundamentally from single actor networks (e.g., closed company network).

A second important issue of public and open access WLANs is information security and authentication. Commercial services always require an authentication procedure due to the need to charge for the service. From this viewpoint, noncommercial networks are more challenging because from the end user perspective, they can be either open access or ID-required networks. From the technical viewpoint, all connections have an "authentication" in the machine-to-machine interaction (requiring an IP address), but end user information is not included in open access networks.

Public organisations can provide two main elements for end-users: 1) the network access signal and 2) the access device. All other components and their provision are based on B2B or B2C arrangements. However, this is a generalised and spatially noncontextualised description of a WLAN provision. The issue becomes more complex when location-related specifics are included. For example, the physical attributes of the WLAN environment are important, because architecture (building heights and materials) and surface topology influence connection quality and speed.

The third, and perhaps the most problematic, field of WLAN analysis is human-to-computer interaction (see Davies, Cheverst, Friday, & Mitchell, 2002; Dawley & Anthony, 2003; Inkinen, 2006). This refers to user preferences, needs, contents and services of the Internet used via the WLAN connections. These include issues of P2P networking and distribution of copyrighted materials. Misuse of information networks is an identified problem of open access networks. Therefore, in most cases noncommercial freely accessible networks have an authentication system to prevent unwanted network behaviour. This implicates issues of network security and related protocols, for example, Wired Equivalency Privacy (WEP) that encrypts transmitted data. Also other security mechanisms, such as end-to-end encryption and virtual private networks (VPN) are tools that increase mobile workstation privacy and security.

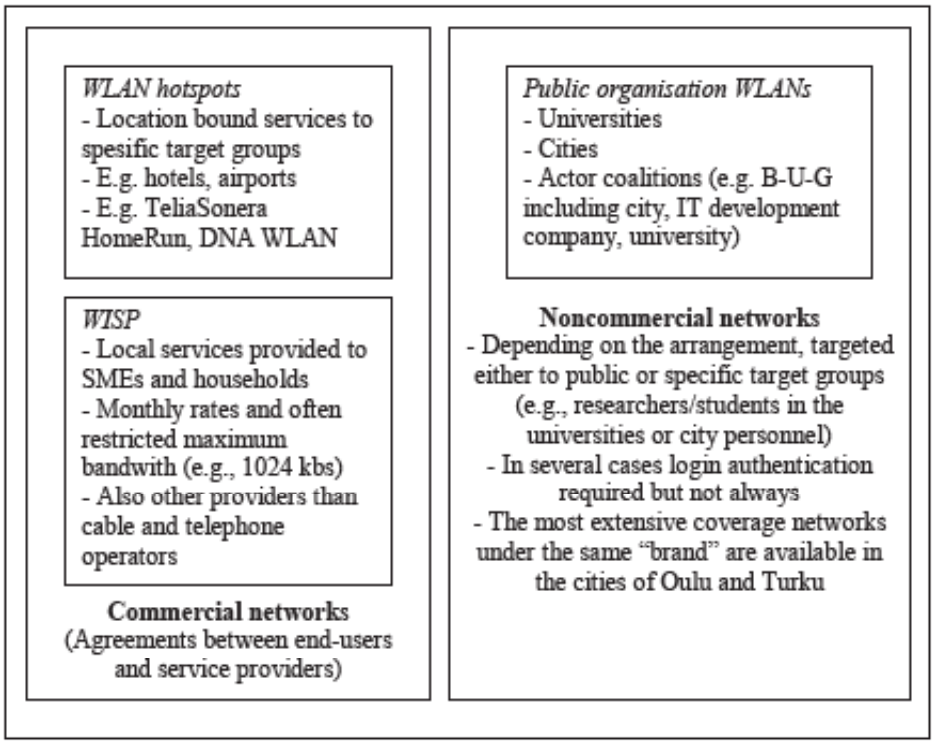
WIRELESS NETWORKS IN FINLAND

General Overview

Finnish public WLANs can be divided into three categories. First, there are commercial services provided by telephony operators, such as TeliaSonera (HomeRun) and Finnet Group (DNA WLAN). These services are provided on a "hot spot" principle, referring to spatially limited locations within a city space. Commercial hot spots are commonly located in hotels, airports, shopping malls, conference centres, restaurants and other gathering places in which the customer segment is expected to need wireless data transfer. These services have a charge, and depending on the tariff system, access can be obtained for minutes to weeks. Hot spot WLANs are targeted to professionals and they mainly serve the needs of business.

Second, there are WISP (Wireless Internet Service Provider) connections providing Internet access. These services

Figure 1. A characterisation of the public WLAN service provision in Finland



are provided mainly by local actors, in Finland commonly by electricity companies, telephone operators or cable-television operators. The target of these services is households and small companies. Compared to "hot spot" networks, WISP connections are less expensive for the end user.

Third, there are noncommercial WLANs providing Internet access in specific city locations. We use the term "city networks" mainly to refer to these noncommercial public WLANs provided or arranged by public sector. The extent of their network coverage varies, depending on the service provision arrangement. Commonly, noncommercial networks in Finland are developed and maintained by units of higher education, such as universities, or by public sector coalitions, including municipal organisations in cooperation with universities and local businesses. The triad cooperation between businesses, universities and government organisations is also called B-U-G cooperation (see terms and definitions). Generalised description of the available WLAN options is presented in the Figure 1.

In addition, it is possible to identify value-added WLAN services. They are private networks, but their provision is a complementary service (side product). Free-of-charge WLAN services in cafés, restaurants and hotels are examples of private noncommercial networks. In general, their provision is based on the consumption of another product (such as a beverage) in a location in which the WLAN service exists, and the role of the WLAN is to be a complimentary service for the customer. Moreover, these networks are commonly based on WISPs, so they can be regarded as realisations of the second category.

There are several solutions to building a cooperative network. Public noncommercial networks are implemented in the cities of Turku (Sparknet), Oulu (PanOulu) and Lappeenranta (WLPR). Sparknet and PanOulu cover the central areas of the respective cities, whereas WLPR is located at the university campus area outside the centre of Lappeenranta. Location-bound projects have also been implemented in Helsinki (e.g., Arabianranta WLAN) and Lahti (e.g., Wireless Lahti). In all cases, university institutions have a key role in the development and implementation of the networks. These networks are commonly segmented to certain user groups, requiring registration (authentication), which in most cases is based on the university network's user IDs. However, there are also exceptions in which authentication is not needed, as in the case of Oulu's RotuaariWLAN, which is a joint network within PanOulu.

A Closer Look: The Sparknet, PanOulu and Arabianranta networks

In the following, we focus on the cooperation arrangements of the noncommercial WLAN solutions in three case locations. PanOulu and Sparknet are actually "collections" of organisational networks and Arabianranta has a specifically built project solution. Thus, in several cases the noncommercial WLANs are themselves networks of networks consisting of various interest groups (see the background section and Figure 1). To combine these existing networks, cooperation is always required; there are business and social arrangements behind the pure technical interoperability. Outdoor WLAN provision in a city space is a multidimensional task that commonly requires efforts from several interest groups.

PanOulu is an open access network resulting from collaboration between the University of Oulu, the Oulu University of Applied Sciences, the City of Oulu and Oulu Telecom. The cooperation agreement was signed in October 2003. This shows that cooperation-based city WLANs are relatively new phenomena. Technically, PanOulu is a joint network including components from all the mentioned partner organisations. The available networks are "KampusWLAN," covering the university campus area and managed by the university; "OuluNet," which covers the area of the University of Applied Sciences; "OuKaWLAN," the network of the City of Oulu covering the city hall, the central library as well as cultural and science centres; and "RotuaariWLAN," an open-door network built cooperatively by university research teams and Oulu Telecom. The latter is an open access network without an authentication procedure (for details, see www.panoulu.net).

Sparknet is a collaboration network managed by ICT Turku Ltd, a local development company owned by the City of Turku in southwest Finland. Thus, the city's role is channelled through a nonprofit development company. The Spark WLAN was originally created by the University of Turku and MP-MasterPlanet Ltd. The City of Turku joined the consortium shortly after its launch in June, 2003. Thus, the city WLANs in Oulu and Turku were originated during the same time with different organisational structure. Currently, Sparknet is the largest and most widespread wireless network solution in Finland. Sparknet is used in both private and public sector organisations (for details, see www.sparknet.fi/en).

Sparknet is divided into two separate networks. The actual Sparknet brand is targeted to organisations, and "Openspark" is a community network mainly for individuals. Interestingly, Sparknet is based on a joint agreement to use the existing network infrastructures within the city. All Sparknet services require authentication. Business-targeted network services have a charge. Individuals are also able to register with Sparknet. Thus, for the "outsider," the Sparknet concept is similar to a commercial hot spot, but it allows substantially broader spatial coverage. The main operating principle is to provide extensive coverage for businesses with low investment costs.

The *Arabianranta network* is a spatially bound WLAN in the Arabianranta residential area (i.e., a local neighbourhood area network, NAN) in Helsinki, the capital of Finland. The Arabianranta WLAN project is mainly a publicly-funded

project operated by a private company. Thus, the Arabianranta network is a public-private partnership that supports attractiveness and economy in a particular location within the city space. The main target group comprises SMEs differentiating it from the resident-focused PanOulu.

Technically, two key elements are identified in the Arabianranta project: the WLAN and the "Helsinki Virtual Village" portal project. The project Web site states: "The fastest and most modern aerial network in Finland is located in Arabianranta. The speed of data transfer in the backbone network is as high as 1 Gbps" (Helsinki Virtual Village, 2006). The project has been implemented by Art and Design City Helsinki Ltd (ADC Ltd) in cooperation with the City of Helsinki and other local stakeholders. Helsinki Virtual Village portal is targeted to small and medium-size enterprises located on the area. The main user segment consists of various fields of business: design companies, media companies and support services for culture and travel. However, users must register themselves as individuals, not as organisations even though the individuals in the most cases represent organisations.

Discussion of WLAN Provision

There are similarities and differences in the urban WLAN solutions in Finland. First, the solutions vary in their provisional structure. All the described cases are joint actions, but the providers vary and actual service provision is mainly deployed by local businesses. Therefore, business-university-government cooperation and a combination of innovative actors within a city are essential in the creation of new services.

Because of their tight organisational relations, proximity and lower hierarchy, medium-sized cities normally have a better starting point for creating public noncommercial WLANs. Bureaucracy and hierarchical structures require unifying of administrative structures in the organisations of larger cities. The cases of Turku and Oulu show that both cities aim to provide an extensive "single service" WLAN in their central areas. Helsinki and its adjacent cities of Espoo and Vantaa have left WLAN provision to the markets and residential area driven development actions, such as Arabianranta.

Second, the target groups vary between cases. PanOulu is clearly targeted to inhabitants. It promotes the city's image of technological advance and openness, whereas the Arabianranta project is a business venture. Sparknet (including Openspark) has divided service provisions among the public and private segments through a separated arrangement. Marketing and "customer" identification are essential in public WLANs (see Davies et al., 2002). Also important are user and community involvement in the creation of the network. Creation of the information infrastructure is connected to the general definition of "information." Is it commodity priced or is it a public good, equally available to everyone?

Third, universities have an active role in WLAN provision in Finland. In the case of public noncommercial networks, universities have been the most important actors in the development of outdoor WLANs. All the universities in Finland are public nonprofit organisations, meaning that they have to gain something other than financial profit from WLAN cooperation. In addition, the absence of universities creates difficulties in providing low-cost WLANs in many cities. Voluntary-based nonprofit WLAN provisions may challenge the long-term robustness of the technology used (also Rao & Parikh, 2003).

One additional issue concerns the public-private operations and the use of public open space as a medium for information retrieval from the Internet. Creation of the information infrastructure, such as WLANs, is connected to the development of national, regional and local information societies (see Inkinen & Jauhiainen, 2006; Webster, 2002). Here, issues such as social division, digital divide and knowledge creation become relevant. They also include the legality of actions. For example, if public organisations take a strong position to promote open access WLANs, they must also consider security: if the infrastructure is misused due to poor information security, what are the responsibilities of the provider?

Last, but not least, is consideration of the spatial coverage of WLANs. This is becoming increasingly important because in the near future many people expect open access or payable WLANs to be available within cities. Various city organisations (tourist offices, business development, public relations) use a WLAN to attract "desired" individuals and businesses to the vicinity of the city. Areas of the city not covered by a WLAN may be seen as marginal or undeveloped urban space. Therefore, urban WLANs relate to urban segregation and differentiation resulting from economic and political relations and decision-making.

FUTURE TRENDS

Studied provision patterns in Finnish cities show that there are various ways to organise a noncommercial WLAN for wide user segments. Noncommercial WLAN service is commonly implemented as a cooperative effort by an organisation and the university and city authorities. The cases of Turku, Oulu and Arabianranta (Helsinki) show that public WLAN service provision includes the central issue of recognising the main service user groups. In addition, the expansion of WLAN coverage and service provision is increasing. This implies that technical proficiency has to be analysed jointly with business and customer segmenting; otherwise network expansion and long-term operability will suffer. Information security also continues to gain increasing attention. Database combining, user and consumer data and marketing have implications for privacy. There are unclear issues related to

composition and the pay-out logic of WLAN service provision. It will be interesting to see whether or not the business model, for example, the case of Sparknet, gains popularity elsewhere. Will one of the current models break through as the best practice for everyone?

There are several future challenges regarding the development and research of city WLANs. First, the development of technical solutions for new protocols and services continues to be an essential topic of engineering sciences. Especially, the emerging 4th generation mobile networks will challenge the short coverage of WLANs. Second, the economic questions of WLANs are gaining more attention by the research community. In particular, the practices and interconnections of cooperative networks need research on decision-making and management. Third, user-driven analysis of WLANs and the Internet is important. Individuals' and businesses' needs for services are relevant study topics for basic and applied research. Finally, the emergence of WLANs in public transportation systems such as trains and busses provides a field of further study of new business solutions taking advantage of wireless communication technologies.

CONCLUSION

This article provided an insight into local WLAN developments in Finland, technologically one of the most advanced countries. The following remarks concerning best practices, bottlenecks and challenges should be considered in similar efforts elsewhere:

- Recognition of the context: successful implementation of technology often requires knowledge of the target segment (consumers, organisations or locations and their particular technology solutions).
- Provision of free or low-charge WLAN solutions greatly depends on joint cooperation models between local organisations. The cases of Turku, Oulu and Arabianranta (Helsinki) are good examples of cooperative networking practices that combine the efforts of universities, city administrations and supportive private sector businesses.
- Selection of the technology employed should take into account the standardisation level of the technology. In several cases (e.g., public smart card traffic fare system in Helsinki), the selected technology does not comply with international standards, increasing the dependence of the whole structure on a single provider. This causes extensive losses if another provider must later reinstall large-scale systems.

Among the greatest challenges in the development of the information society and its technological solutions is interoperability. As discussed, it is not only about hardware-to-hardware, software-to-hardware or software-to-software interoperability, but also, and perhaps even more importantly, a question of human-to-human interaction and cooperative institutional arrangements. Target group identification is one practical example in the studied cases. Urban WLAN networks targeted to the masses (including all inhabitants) are likely to be more successful if their provision is based on wide cooperation between the stakeholders.

REFERENCES

Chao, H.-C., Uden, L., & Shih, F.Y. (Eds.). (2006). Special issue: Mobile IP. *Wireless Communications and Mobile Computing*, *6*(5), 543-739.

Cisco. (2006). *Overview of municipal wireless network applications and technology for city managers.* Retrieved December 12, 2007, from www.cisco.com/en/US/products/ps6548/prod_brochure0900aecd8056ea1c.pdf

Davies, N., Cheverst, K., Friday, A., & Mitchell, K. (2002). Future wireless applications for a networked city: Services for visitors and residents. *IEEE Wireless Communications, 9*(1), 8-16.

Dawley, D.D., & Anthony, W.P. (2003). User perceptions of E-mail at work. *Journal of Business and Technical Communication*, *17*(2), 170-200.

Dobers, P. (2004). Stockholm as a mobile valley. Empty spaces or illusionary images? *Journal of Urban Technology*, *11*(3), 87-108.

Friday, A., Davies, N., Wallbank, N., Catterall, E., & Pink, S. (2004). Supporting service discovery, querying and interaction in ubiquitous computing environments. *Wireless Networks*, *10*(6), 631-641.

Graham, S., & Marvin, S. (1996). *Telecommunications and the city*. London: Routledge.

Harwit, E. (2005). Telecommunications and the Internet in Shanghai: Political and economic factors shaping the network in a Chinese city. *Urban Studies*, *42*(10), 1837-1858.

Helsinki Virtual Village. (2006). *Made in Arabianranta*. Retrieved December 12, 2007, from http://www.helsinki-virtualvillage.fi/Resource.phx/adc/inenglish/index.htx

Inkinen, T. (2006). The social construction of the urban use of information technology: The case of Tampere, Finland. *Journal of Urban Technology, 14*(3), 49-75.

Inkinen, T., & Jauhiainen, J.S. (2006). Public authorities and the local information society. In A. Anttiroiko & M. Mälkiä (Eds.), *Encyclopedia of digital government* (Vol. 3, pp. 1370-1376). Hershey, PA: Idea Group.

Jenisch, M., Orlamünder, H., Köstring, N., & Brügge, T. (2005). My personal city information. *Wireless Personal Communications*, *33*(3), 271-279.

Palm, J., & Wihlborg, E. (2005). Governed technology? Urban management of broadband and 3G systems in Sweden. *Journal of Urban Technology*, *13*(2), 71-89.

Rao, B., & Parik, H. (2003). Wireless broadband drivers and their social implications. *Technology in Society*, *25*(4), 477-489.

Salkintzis, A.K., Pavlidou, F.N., Fitzek, F.H.P., & Varma, V.K. (Eds.). (2005). Special issue on advances on wireless LANs and PANs. *Wireless Personal Communications*, *34*(1), 1-108.

Tang, D., & Baker, M. (2002). Analysis of metropolitan-area wireless network. *Wireless Networks*, *8*(2-3), 107-120.

Webster, F. (2002). *Theories of the Information Society* (2nd ed.). London: Routledge.

WEF. (2007). *Global information technology report 2006-2007*. London: Palgrave MacMillan.

Zhuang, W., Gan, Y.-S., Loh, K.-J., & Chua, K.-C. (2003). Policy-based QoS management architecture in an integrated UMTS and WLAN environment. *IEEE Communications*, *41*(11), 118-125.

KEY TERMS

B-U-G Cooperation: An organisational cooperation model that includes partners from business, university and government. B-U-G co-operation is commonly locally based aiming to foster local development.

Public E-Service: A service available in digital form. Public e-services include all services provided by the public actor. These include Internet-based service solutions, other network-based solutions (also restricted), smart card solutions and other digitalised authentication methods targeted to users. Public e-services are provided by authorities of different spatial scales, including local actors (e.g., city organisations), regional actors (counties and districts), national and international actors (see Inkinen & Jauhiainen, 2006).

Public Noncommercial Wireless Network: A wireless network not based on market economy profit-making. Public organisations providing a wireless Internet connection without a charge or payment are noncommercial WLANs.

Public-Private Partnership: A joint agreement between public organisation and private business resulting into a product, service or development venture that is funded and operated through a cooperation of participating partners.

Wireless City Network: A location-bound WLAN provided by a public or a commercial organisation. Generally, a wireless city network can be used by anyone. The connection can be made with or without identification process by anyone located in the specified area. A city network can also be locally targeted inside the city to a certain location or it can be a full-coverage network.

Wireless Internet Service Provider (WISP): An actor providing wireless connection service to households and organisations. The WISP service customer (subscriber) gains access to network services within the vicinity of the server transmission range. These locations are called "hot spots" or "access points." WISPs provide basic service sets (BSS) and extended service sets (ESS) that are defined in the IEEE 802.11b specification.

Wireless Service End-User: A person or an organisation using the provided wireless Internet connection with a user interface that is commonly a lap-top computer, palm-top device such as a smart phone or other portable device.

Spatial Data Infrastructures

Clodoveu Augusto Davis, Jr.
Pontifical Catholic University of Minas Gerais, Brazil

INTRODUCTION

Spatial Data Infrastructures (SDI), also known as Spatial Information Infrastructures (SII), are a set of policies, technologies and standards that interconnect a community of spatial information users and related support activities for production and management of geographic information (Phillips, Williamson, & Ezigbalike, 1999). SDI reduces redundant effort and lowers production costs for new and existent datasets through interoperable information sharing, providing neutral means to access geographic data. Multiple information providers, commercial or public, may cover various interests and compete among themselves for clients.

SDIs present several challenges, at various levels of interaction. First, there is a societal and organizational level. Partners in a community should have convergent interests, agree on common rules, and be able to use information produced by others. Such agreements are not easy to achieve, and usually require long-term commitments. Within public organizations, it is usual to think in transnational terms, between national mapping agencies, but intranational relationships are also important.

Second, there are standardization issues. Guiding the technology standardization and defining the key elements for SDI, the Open Geospatial Consortium (OGC) has proposed a number of standards, through a framework called *OGC Reference Model* (Percivall, 2003).

Third, there are concerns on specific aspects of geographic information, such as scale (levels of detail, accuracy, uncertainty) and the need to integrate data from various sources. Geographic information from each source needs to be consolidated in order to be valuable to high-level decision-makers. In this case, SDI can be seen as a set of building blocks, in which hierarchies are built through the exchange and consolidation of information from corporate and local levels, to regional and global levels. In this hierarchy, lower levels (Davis & Alves, 2005) provide detailed information that helps to consolidate the upper, more general, levels (Rajabifard & Williamson, 2001). The integration problem also requires attention to semantics, because data produced by different organizations, for different needs, are not necessarily compatible, even if they refer to the same location or to the same real-world subject. In this particular issue, the development and use of ontologies may be required.

Finally, there is a technological level. The exchange of information can occur in several ways, but the most interesting one is the use of Web services, using a service-based architecture approach, thus achieving *loosely-coupled and distributed geographic information systems* (Bernard & Craglia, 2005; Davis & Alves, 2005). There are pending issues related to the compatibility between Web service standards defined by the OGC and by the World Wide Web Consortium (W3C), but there are already initiatives to bridge them (Bacharach, 2007; Kim, Kim, Lee, & Joo, 2005). There is also the need to define and propose higher-level services, so SDI can go beyond the simple discovery and download of geographic data, and provide solutions to location-related problems using multiple and distributed sources of information.

BACKGROUND

Creating geographic datasets is a complex and expensive undertaking. In the past, redundant efforts in dataset creation were commonplace: organizations with an interest in the same areas, therefore potential partners for sharing basic data, would not cooperate due to their diverse technological strategies, budgeting, and timing. Of course, such redundancy was undesirable, motivating the creation of cooperation efforts for data sharing.

An example of such an arrangement took place in Belo Horizonte, Brazil. A cooperation agreement, involving 29 different organizations, including government agencies, universities, and private-sector companies, has been active since 1994 (Davis & Fonseca, 2005). Even though political and organizational problems have been solved in this case, technologically there is still a lot to be done. Data for interchange reside in a FTP server, for download by authorized people in one of five different data formats. There is also a metadata sheet for each information class, presented as a simple and nonstandardized text file. Simple as it may seem, maintaining such a setting requires much effort by those who coordinate data gathering and distribution, because much work is performed manually or with little automation. Constant and efficient interpersonal communication is required, so that one organization's needs on some data can be informed to the organizations that generate that data. Thus, cooperative settings are a great improvement on the early approach to sharing, but their scalability potential is rather limited.

Data Transfer Standards

Large-scale off-line data sharing depends fundamentally on data translation, because each organization can use a different geographic information system (GIS). Many efforts in the past have tried to establish a neutral file format for exchange purposes, so that every GIS would only need translators to and from this common format (Lima, Câmara, & Monteiro, 2001).

In practice, commercial formats are used in most data transfer situations, reflecting the influence of the user community of a given GIS package. Regardless of using *de facto* or *de jure* standards, this approach addresses syntactic concerns only, avoiding semantic issues. Furthermore, data transfer formats are unsuitable for online access, maintaining the need for an export-import off-line cycle. Off-line sharing causes multiple copies of the same data to be distributed among interested parties at different times, causing serious synchronization problems.

Spatial Data Clearinghouses

From the establishment of a standard (or, at least, from some *de facto* standards), many national mapping agencies started to create *spatial data clearinghouses*, Internet-based settings that intend to facilitate access to spatial data. A centralized site, from which data from several sources can be found, is established, including services for searching, viewing, transferring, and ordering spatial data (Crompvoets, Bregt, Rajabifard, & Williamson, 2004). Clearinghouses allow data providers to make their offerings known by users, with descriptions (metadata) and instructions on how to access and use the data.

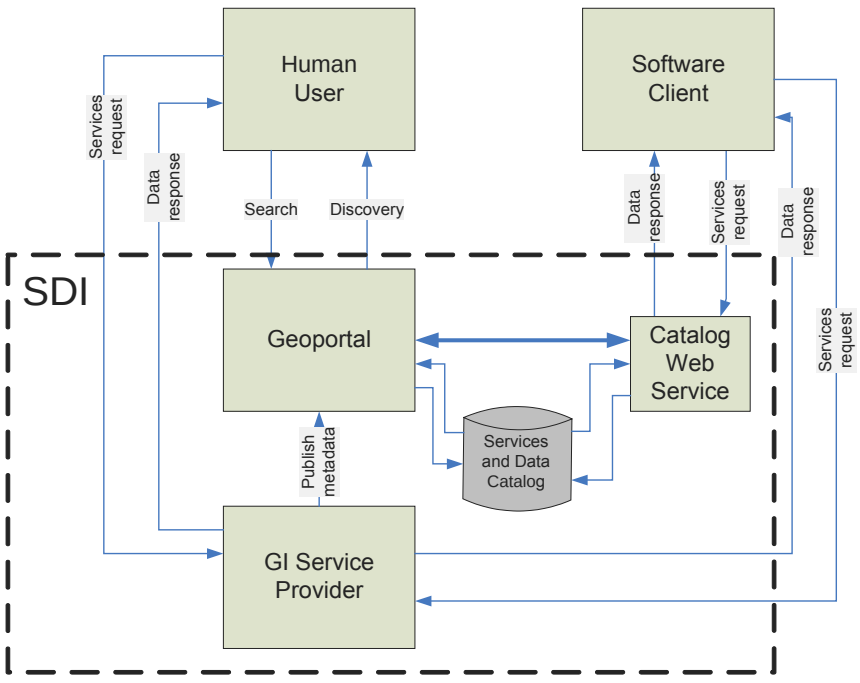
Clearinghouses have been more recently described as a kind of Web portal, that is, a site or a gateway through which commonly used services are offered (INSPIRE, 2002). The emphasis on services is recent, compared to previous implementations, which were mostly based on a combination of technical tools, institutional cooperation mechanisms, and commercial concerns, directed at "off-the-shelf" data dissemination (FGDC, 1997).

A recent study on clearinghouses (Crompvoets et al., 2004) showed that users are dissatisfied with their functionality, indicating that the focus should change from a data-oriented to a user- and application-oriented view. This can be achieved by using service-based architectures.

Early Spatial Data Infrastructures

The expression "spatial data infrastructure" was initially used to describe the provision of standardized access to geographic information (Maguire & Longley, 2005). Many clearinghouse initiatives evolved to what Masser (1999) calls "the first generation of national spatial data infrastructures," while observing that "infrastructure" implies the existence of some sort of coordination for policy formulation and implementa-

Figure 1. Geoportals and SDI



tion. This first generation of SDI focused on granting a broad thematic scope, which is consistent with the current analogy between SDI and other types of infrastructure: fostering economic development by granting access to publicly-available and multiple-use goods or services. By "publicly-available," we do not mean "government-supported." Even though SDIs are seen as drivers of economic development, some of the initiatives reviewed by Masser (1999) do not grant access to the private sector, or do so by charging an usage fee, as a means to recover some or all their costs.

Evolution from the first generation of SDI was made possible by the recent expansion of Web-based information systems. In the USA, the Geospatial One-Stop (GOS) Web portal was created to provide widespread access to geographic information[a], inaugurating the concept of *geoportals* (Maguire & Longley, 2005; Tait, 2005).

GeoPortals

The term "portal" has been widely used with the general meaning of an "entry point" for information and services available on the Web. Applying this concept to geoinformation, a geoportal is, therefore, a "Web site that presents an entry point to geographic content on the Web" (Tait, 2005). A geoportal includes the discovery of information sources and content, and online access to applications. Examples of geoportals include the previously mentioned Geospatial One-Stop from the USA, and the EU-Geoportal, a part of the INfrastructure for SPatial Information in Europe (INSPIRE) project (INSPIRE, 2002)[b].

It is important to distinguish between the concepts of SDI and geoportal. We consider that an SDI is formed by the confluence of (potentially) several geographic data providers, each of which grants data access through specific Web services. In order to select services to fulfill his needs, the user searches through a repository of metadata on available geographic data and services. In the case of a human user, searches are done interactively, through a geoportal, using search interfaces and other interactive tools; in the case of a software client, this can be done through a catalog Web service. Thus, we consider that a geoportal is a component of an SDI (Figure 1).

The use of Web services to grant direct access to data is the most important distinction between first- and second-generation SDIs. In fact, the numerous possibilities that arise from using services to encapsulate data from multiple sources, and thereby achieve interoperability, have led Bernard and Craglia (2005) to propose a new translation for the SDI acronym: *Service-driven Infrastructures*.

SPATIAL DATA INFRASTRUCTURES AND SERVICE-ORIENTED ARCHITECTURES

The most current view on spatial information infrastructures considers their insertion into the perspective of *service-based distributed system architectures*, which have been proposed as part of a strategy for developing complex information systems based on reusable components. One of the most interesting approaches in this field is the one of *service-oriented architectures* (SOA) (Papazoglou & Georgakopoulos, 2003).

Services, their descriptions and fundamental operations, such as *discovery*, *selection*, and *binding*, form the basis of SOA. SOA supports large applications with sharing of data and processing capacity, through network-based distributed allocation of applications and use of computational resources. In this architecture, services are self-contained, which means that information on the service's description, including its capabilities, interface, behavior, and quality, can be obtained from the service itself, through a standardized set of functions.

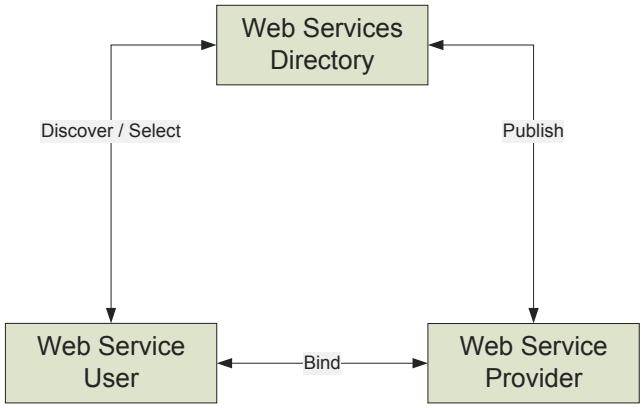
Service providers, service aggregators and service users are the actors that participate in this scenario. Providers implement and publish services, while aggregators design compositions of rules based on primary services. Available services should be listed in directories for user reference, or "discovery." Service users may be human or software clients, which need to access the services through the communications network.

Web Services

Web services are a particular class of services that use open Internet standards, such as connection and communication using the Hypertext Transfer Protocol (HTTP), identification using the Uniform Resource Identifier (URI), contents specification through the eXtensible Markup Language (XML), service descriptions expressed by the Web Services Definition Language (WSDL), and directory services using the Universal Description, Discovery and Integration (UDDI) protocol (W3C, 2002). Therefore, while services in general provide interoperability between different software components, Web services go a step further by facilitating cross-institutional interchange of data and services over the Internet, and by improving resources sharing among various data sources.

The main Web service operations are *publication*, *discovery*, *selection*, *binding*, and service *composition*.

Figure 2. Web service operations



Publication is performed by a service provider, and consists in creating a service description in WSDL and publishing it on Web-based discovery channels. These channels use the UDDI protocol to register the service, storing data that allow users to *discover* it. Users can then *select* services using UDDI to search through catalogs or directories, getting sets of URIs in response. Once the service descriptor is obtained (in WSDL) for the *binding* operation, client software initiates a direct communication with the service provider using HTTP, sending a request. The binding ends with the reception of the expected Web service response in XML (Figure 2). Finally, more complex services can be created by the *composition* (or *chaining*) of primary ones. In this case, service users access the composite services in the same way as simple services, but different results are possible. Thus, several alternatives in terms of service performance, costs and quality become feasible through the appropriate configuration of Web service chains.

OGC Web Services

The Open Geospatial Consortium (OGC) proposed the *OpenGIS Services Framework* (Percivall, 2003), an architecture for sharing of geographic data and functionality over the Internet, thus leading the standardization process on data formats, methods and interface specifications. The OpenGIS Services Framework does not necessarily employ the usual (W3C) Web services standards, such as the Simple Object Access Protocol (SOAP) and WSDL. Instead of using UDDI, the OGC proposes the use of catalog services for the implementation of the publication, discovery and selection operations. Moreover, OGC Web services have a particular interface for binding, which does not use service descriptors. This alternative poses difficulties for indexing and searching. OGC Web services use Geographic Markup Language (GML) to encode and transmit objects, while regular Web services use generic XML, but this is not actually a difference, because GML is based on XML. We observe that the main differences that exist between W3C and OGC Web services should be solved as soon as possible, in order to incentive the adoption of the proposed standards in a more universal way (Davis & Fonseca, 2005; Sonnet, 2004).

Some basic Web services were specified by OGC, as listed below:

- **Web Feature Service:** provides an interface for the insertion, selection, updating and removal of geographic features (objects).
- **Web Coverage Service:** provides access to geo-fields, much in the same manner of the Web Feature Service. Notice that this service does not return images of the geo-fields, but rather returns semantic details on them.
- **Web Gazetteer Service:** extends the Web Feature Service with resources for the implementation of interfaces to gazetteers.
- **Web Registry Service and OpenGIS Catalog Service (OCS):** implement an operational functionality similar to UDDI.
- **Web Coordinate Transformation Service:** provides an algorithm that converts coordinates for spatial objects between different spatial reference systems.
- **Web Map Service:** a service for the production of maps over the Web, or Web Maps. Maps, in this service, are renderings (presentations) of the reality, and do not include the actual geographic data.
- **Web Terrain Service:** similar to the Web Map Service, but geared toward three-dimensional renderings of surfaces.

OGC Web services may also be combined. A special OGC Web service, the Web Notification Service, can be used to send update notifications to registered clients that participate on a chain in which some Web service is to be altered. The

OpenGIS Services Framework shows the degree of commitment of the OGC with service-oriented architectures for interoperability purposes, which historically represent the core of OGC's purposes.

Such an open and flexible architecture will find its main uses in situations very similar to the ones presented in the discussion on SDI. Summing up our previous arguments, SDI must be distributed, must support multiple applications, multiple clients of several different types, multiple data sources, multiple data maintenance teams, under a heterogeneous computational environment. SDI must not force the adoption of specific products on their participants, but should instead provide an architectural view and determine a set of minimal standards. These standards should be as widely accepted as possible, and typical Internet standards as the ones we mentioned fill that description closely.

FUTURE TRENDS

It seems to be clear that the integration of the service approaches by the W3C and the OGC is desired by both groups and that could have a positive impact on future SDI initiatives. The differences only exist because the OGC, from its pioneering work on geoinformation integration, has conceived and implemented the concept of Web services even before the W3C, as the Web standards organization, reached that stage. This is a sign that information integration and interoperability problems that are being researched today have first arisen as significant concerns in the geoinformatics community.

Considering the tendency toward unification of the standards, SDI represents a new set of technologies that are central to a new generation of geographically distributed applications. SDIs are also a first step toward information availability through ubiquitous networking infrastructures.

CONCLUSION

A look at the OGC proposed Web services shows that most are concerned with data access, not with supplying problem-solving information. We consider that SDIs will benefit from the proposal of new services, in which actual information generation can be achieved. SDI, in the future, should deal with queries that are vague, if compared to what can be done using today's database query languages, and respond with information gathered from various distributed sources.

There is also potential for enabling the geographic information user as a "server" of various kinds of information. Information perceived directly by the user (Alves & Davis, 2007; Goodchild, 2007), along with previously collected data, perceived quality of information, ontologies about its interests, geographic position, and others, can be communicated to other users through services. However, additional studies concerning security and privacy in such applications are required.

ACKNOWLEDGMENT

The author wishes to acknowledge the support of his work receives from CNPq and FAPEMIG, Brazilian agencies in charge of fostering research and developement.

REFERENCES

Alves, L. L., & Davis, Jr, C. A. (2007). Evaluation of OGC Web services for local spatial data infrastructures and for the development of clients for geographic information systems. In C. A. Davis, Jr & A. M. V. Monteiro (Eds.), *Advances in geoinformatics* (pp. 217-234). Berlin: Springer-Verlag.

Bacharach, S. (2007, January 17). OGC joins W3C to help add geospatial to the Web. *OGC Press Release*.

Bernard, L., & Craglia, M. (2005). SDI—from spatial data infrastructure to service driven infrastructure. In *Proceedings of the Research Workshop on Cross-Learning Between Spatial Data Infrastructures and Information Infrastructures,* Enschede, The Netherlands.

Crompvoets, J., Bregt, A., Rajabifard, A., & Williamson, I. (2004). Assessing the worldwide developments of national spatial data clearinghouses. *International Journal of Geographic Information Science, 18*(7), 665-689.

Davis, Jr, C. A., & Alves, L. L. (2005). Local spatial data infrastructures based on a service-oriented architecture. In *Proceedings of the 8th Brazilian Symposium on GeoInformatics (GeoInfo 2005),* in CD-ROM, Campos do Jordão (SP).

Davis, Jr., C. A., & Fonseca, F. T. (2005). Considerations from the development of a local spatial data infrastructure. *Information Technology for Development, 12*(4), 273-290.

FGDC. (1997). *Metadata to clearinghouse hands—on tutorial*. Washington, DC: Federal Geographic Data Committee.

Goodchild, M. F. (2007). Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research, 2*, 24-32.

INSPIRE. (2002). *INSPIRE architecture and standards working wroup, INSPIRE architecture and standards position paper*. Brussels: Commission of the European Communities.

Kim, M., Kim, M., Lee, E., & Joo, I. (2005). Web services framework for geo-spatial services. In C. Claramunt, Y.-J. Kwon, & A. Boujou (Eds.), *Web and Wireless Geographical Information Systems: 4th International Workshop W2GIS 2004* (pp. 1-13). Goyang, Korea: Springer-Verlag.

Lima, P., Câmara, G., & Monteiro, A. M. V. (2001). Geographic data exchange: Models, formats, and converters (in Portuguese). In *Proceedings of the III Brazilian Workshop on GeoInformatics (GeoInfo 2001),* Rio de Janeiro, Brazil, (pp. 122-128).

Maguire, D. J., & Longley, P. A. (2005). The emergence of geoportals and their role in spatial data infrastructures. *Computers, Environment and Urban Systems, 29*(1), 3.

Masser, I. (1999). All shapes and sizes: The first generation of national spatial data infrastructures. *International Journal of Geographic Information Science, 13*(1), 67-84.

Papazoglou, M. P., & Georgakopoulos, D. (2003). Service-oriented computing. *Communications of the ACM, 46*(10), 25-28.

Percivall, G. (Ed.). (2003). *OpenGIS reference model*. Open Geospatial Consortium.

Phillips, A., Williamson, I., & Ezigbalike, C. (1999). Spatial data infrastructure concepts. *The Australian Surveyor, 44*(1), 20-28.

Rajabifard, A., & Williamson, I. (2001). Spatial data infrastructures: Concept, SDI hierarchy and future directions. In *Proceedings of Geomatics'80*, Tehran, Iran.

Sonnet, J. (2004). *OWS 2 common architecture: WSDL SOAP UDDI.* (Discussion Paper No. OGC 04-060r1, version 1.0.0). Open Geospatial Consortium.

Tait, M. G. (2005). Implementing geoportals: Applications of distributed GIS. *Computers, Environment and Urban Systems, 29*(1), 33-47.

W3C. (2002). *Web services architecture working group, Web services architecture requirements* (W3C working draft). World-Wide Web Consortium.

KEY TERMS

Geographic Information System (GIS): Information systems used to store, analyze, and manipulate geographic data, that is, data that represent objects or phenomena for which the geographic location is an important characteristic.

Geoportal: A Web site that presents an entry point to geographic content on the Web, used to discover and access geographic information and associated services on the Web.

GML: The Geography Markup Language is a XML grammar defined by the Open Geospatial Consortium (OGC) to adequately express and transfer, in a neutral way, the encoding of geographic features. Its purpose is to foster the integration of geographic data sources.

Service-Oriented Architectures (SOA): Information system architectures in which services encapsulate the exchange of data among modules, which can reside in different points throughout a computer network.

Spatial Data Clearinghouse: Internet-based components that intend to facilitate access to spatial data, by establishing a centralized site from which data from several sources can be found, and by providing complementary services, including searching, viewing, transferring, and ordering spatial data.

Web Services: Software applications from which interfaces and bindings are expressed in XML and that can be discovered using XML messages. In the W3C definition, Web services are "a software system designed to support interoperable machine-to-machine interaction over a network."

XML: The eXtensible Markup Language is a markup language developed to facilitate the sharing of structured data across different information systems.

ENDNOTES

[a] http://www.geo-one-stop.gov
[b] http://eu-geoportal.jrc.it

Category: Web Technologies

Spatial Search Engines

Cláudio Elízio Calazans Campelo
University of Campina Grande, Brazil

Cláudio de Souza Baptista
University of Campina Grande, Brazil

Ricardo Madeira Fernandes
University of Campina Grande, Brazil

INTRODUCTION

It is well known that documents available on the Web are extremely heterogeneous in several aspects, such as the use of various idioms, different formats to represent the contents, besides other external factors like source reputation, refresh frequency, and so forth (Page & Brin, 1998). Altogether, these factors increase the complexity of Web information retrieval systems.

Superficially, traditional search engines available on the Web nowadays consist of retrieving documents that contain keywords informed by users. Nevertheless, among the variety of search possibilities, it is evident that the user needs a process that involves more sophisticated analysis; for example, temporal or spatial contextualization might be considered. In these keyword-based search engines, for instance, a Web page containing the phrase "…due to the company arrival in London, a thousand java programming jobs will be open…" would not be found if the submitted search was "jobs programming England," unless the word "England" appeared in another phrase of the page. The explanation to this fact is that the term "London" is treated merely like another word, instead of regarding its geographical position. In a spatial search engine, the expected behavior would be to return the page described in the previous example, since the system shall have information indicating that the term "London" refers to a city located in a country referred to by the term "England." This result could only be feasible in a traditional search engine if the user repeatedly submitted searches for all possible England sub-regions (e.g., cities). In accordance with the example, it is reasonable that for several user searches, the most interesting results are those related to certain geographical regions. A variety of features extraction and automatic document classification techniques have been proposed, however, acquiring Web-page geographical features involves some peculiar complexities, such as ambiguity (e.g., many places with the same name, various names for a single place, things with place names, etc.). Moreover, a Web page can refer to a place that contains or is contained by the one informed in the user query, which implies knowing the different region topologies used by the system.

Many features related to geographical context can be added to the process of elaborating relevance ranking for returned documents. For example, a document can be more relevant than another one if its content refers to a place closer to the user location. Nonetheless, in spatial search engines, there are more complex issues to be considered because of the spatial dimension concerning on ranking elaboration. Jones, Alani, and Tudhope (2001) propose a combination of Euclidian distance between place centroids with hierarchical distances in order to generate a hybrid spatial distance that may be used in the relevance ranking elaboration of returned documents. Further important issues are the indexing mechanisms and query processing. In general, these solutions try to combine well-known textual indexing techniques (e.g., inverted files) with spatial indexing mechanisms. On the subject of user interface, spatial search engines are more complex, because users need to choose regions of interest, as well as possible spatial relationships, in addition to keywords. To visualize the results, it is pleasant to use digital map resources besides textual information.

BACKGROUND

Numerous contributions have been made in the information retrieval (IR) area since 1960's decade. Nevertheless, due to Web continuous growth, research in this field is still in infancy.

Baeza and Ribeiro (1999) say that IR brings some challenges, such as how to determine the real user needs, as well as supply their expectations through relevant document subsets. In IR systems, it is necessary to analyze both semantics and syntax of document contents, which may return imprecise results. According to Kowalsky (1997), the aim of an IR system is to minimize the overhead of finding the expected information. Classical IR models consider that a document is described by a set of indexed terms. Some of these models

also take into account different terms importance at the same document. This importance is called weight (w), and can be represented by a numeric value. The most well-known classical models are Boolean, probabilistic and vector. The vector one, proposed by Gerard Salton, has a greater acceptance between researchers and is the most utilized in current IR applications.

The Web brings new features and difficulties to the IR process, due to both the heterogeneity of the underlying documents and the approaches used to present them. Studies have demonstrated that a large amount of information disposed on the Internet has some kind of geographical context. For instance, the locale where the information was created, the referenced information locale, the place where most information consumers live, and so forth. However, traditional search engines do not consider this spatial context in their information organization and retrieval process.

The requirement for efficient information supported by the knowledge about a specific domain raised the concern on developing ontologies that model many associated concepts. Concerning geographical IR, the use of ontologies can be extremely important for geographical features representation of documents. Fu, Jones, and Abdelmoty (2005) suggest the existence of a primary ontological component, place ontology, that provides the terminology and geographical space structure modeling. Such ontology has a fundamental role, for instance, in user query interpretation, relevance ranking elaboration, and metadata extraction.

Surely, the relevance ranking elaboration is one of the main processes in a search engine, since it is directly related to user interest. In traditional systems, the ranking can be produced through a variety of techniques, for example, similarity measures between query and returned documents using the spatial-vector model (Baeza & Ribeiro, 1999). Currently, one of the most accepted methods is the PageRank (Page & Brin, 1999) that uses the Internet link structure to produce its ranking.

Research on spatial search engines is very incipient. It has addressed some information retrieval subareas aiming to develop efficient data structures and algorithms for space-textual indexing; to elaborate efficient approaches for relevance ranking using the geographic context; and to detect and model the geographic scope.

There are different approaches to extract geographic information from Web-crawled documents. Buyukkokten, Cho, Molina, Gravano, and Shivakumar (1999) associate domain name IP addresses to telephone code area, and by using postal code of the Web site, they enable to match place names to geographic coordinates. McCurley (2001) introduced the geocoding concept, which enables one to associate geographic coordinates to Web pages. McCurley also has used several terms that are useful to geocode a Web page, for instance, postal code, city names, and telephone numbers. Nonetheless, there is no discussion on techniques to extract and eliminate ambiguity. Gravano (Gravano, Hatzivassiloglou, & Lichtenstein, 2003) has implemented automatic geocoding.

More recently, some research projects have presented relevant results in this field, as, for example, the GeoTumba one (Chaves, Silva, & Martins, 2005). GeoTumba contains a repository based on a domain-independent metamodel to integrate geographical knowledge collected from several sources. Silva, Martins, Chaves, Afonso, and Cardoso (2006) focus on techniques for geographical features extraction from large collections of Web documents by using a method that involves the attribution of geographic scope through the GraphRank algorithm, which is inspired in the PageRank one (Page & Brin 1999). Silva et al. (2006) show three sets of heuristics used in the process of georeferencing Web pages.

Another important research project is the SPIRIT (*spatially-aware information retrieval on the Internet*), which focuses on geographic information retrieval and issues involving the semantic Web. This project proposes a multimodal interface that provides text and maps; spatially aware ontologies; query expansion and relevance ranking based on geographic ontologies; spatial indexes for the document collection; and a learning mechanism for extracting geographic context from Web documents that generates spatial metadata. An overview of this project can be found in Jones, Abdelmoty, Finch, Fu, and Vaid, 2004), which addresses the architecture, indexing mechanisms, and spatial ontologies.

Markowetz, Chen, Suel, Long, and Seeger (2005) use the geocoding concept divided into three steps: *geoextraction*, *geomatching and ge propagation*. Their work is inspired in Buyukkokten et al. (1999), and Ding, Gravano, and Shivakumar (2000). Markowetz, Brinkhoff, and Seeger (2004) propose a relevance ranking that may balance between text and spatial ranking.

TOWARD A SPATIAL SEARCH ENGINE

This section presents the GeoSEn, geographic search engine, project that has been developed in our laboratory, which may be accessed at http://www.lsi.dsc.ufcg.edu.br/geosen. GeoSEn is a Web spatial information retrieval system that uses geographical scope detection mechanisms (set of places that Web element contents can be associate) to better index Web documents.

Some studies have proposed a Web geographical information retrieval system. The major concern of these prototypes is to detect pages spatial features and represent them, providing documents retrieval according to a relevance ranking, elaborated considering the geographical location of the documents. However, the presented mechanisms for geographical scope detection and relevance ranking elaboration still have limitations and demand for innovative contributions.

Therefore, the main issue to be addressed in this research is to set the geographical scope of Web pages using more complex methods, and developing new techniques that are able to associate more flexible and complete contexts. The main contributions of the GeoSEn approach are:

- **Places associated by different factors:** besides administrative regions (e.g., cities, states), it will be possible to take, as valid places, regions formed by other factors such as weather (e.g., rainy or dried places), socioeconomic (e.g., countries with high level of violence), cultural (e.g., Paris Dakar rally places), and so forth.
- **Distinction between Web-page geographic scope and places of interest:** capability of distinguishing the places that are related to the page content and the places of interest of a particular page (Markowetz et al., 2004).
- **Multimodal interface:** the software interface should provide spatial, temporal, image, and textual interaction.
- **Multimedia information retrieval:** the system will be able to retrieve images and video based on their geographical scopes.
- **Innovative mechanism of Web page geographical scope attribution:** the proposed mechanism is capable of realizing analysis based on heuristics of document-referenced places distribution pattern, statistics about this distribution, many spatial relationships between places and weather, cultural and socioeconomic data. Moreover, well-known geographical references detection methods (Buyukkokten et al., 1999; Ding et al., 2000; Markowetz, Brinkhoff, & Seeger, 2005; Silva et al., 2006), as well as Web page link structure analyis techniques for geographical scope attribution (Buyukkokten et al., 1999; Markowetz et al., 2005) will be improved and used with the innovative methods developed.

After a particular page is captured by a Web crawler, it is parsed to find terms that can be mapped to some place name, for example, cities' names, zip codes, telephones, and so forth. Afterwards, these place names are associated with geographical coordinates, used internally by the system to manipulate the spatial information (e.g., gazetteer). Each identified place is associated to a certain level of trust. References to one or more places can be found in the same page, which allows the system to infer the document's geographical scope, thus, the region that the document is referenced. Hence, the system will take into account a variety of heuristics and specialized algorithms to combine features of discovered terms and different spatial relationship between analyzed regions. A major factor to be considered is the references distribution pattern. For example, suppose that a document has geographical references mapped to Brazilian states like Paraná, Santa Catarina, and Rio Grande do Sul. Thus, it is expected that the document is related to the Brazilian south region, once these three states are located at this region. Similarly, suppose a document that references many states at each of the five Brazilian Regions. It is possible that this document is associated to a national context.

Combined with this distribution pattern, GeoSEn utilizes several statistics to deduce geographical document scope. Still considering Brazil, suppose that two documents have references to 3 and 9 Brazilian swtates, respectively, from a total of 27. It is feasible that the second one is associated to a national context while the other is related to a restricted region. However, these many factors must be combined in order to obtain a more precise result. For instance, it is possible that all nine states referenced by the second document belong to the same region, whereas the first document references three states, where each one is from a different region, changing completely the initial hypothesis.

Regarding algorithms for document's geographical scope modeling, GeoSEn considers the various relationships between referenced places, for example, containment, distance, and adjacent properties. These relationships can be extremely useful, for instance, to detect regions that, even not belonging officially to a territorial division, are known by commercial trends, weather factors, and so forth.

- **URLs and WHOIS analyze:** besides checking the documents to search geographical references, the system analyzes URLs related to these documents and WHOIS (a set of public information that can be queried in a particular Internet domain).
- **External services interoperability:** in order to complement the geographical scope attribution and modeling necessary data involved in analysis, some external services are queried (e.g., gazetteers) via Web services.
- **Implementation issues:** after defining functional and nonfunctional requirements, 9 Web crawlers and 10 Web search engines, all of them open-source, were evaluated in order to choose the most appropriate to be extended with GeoSEn features. The final choice was the Nutch, from Apache Software Foundation. Details of the Nutch architecture can be seen at http://lucene.apache.org/nutch/. The main reasons for this choice were that it has shown great maturity when compared to the others; it is an active project that integrates, in one system, a Web crawler and a Web search engine; and it provides an easy extension mechanism due to its plug-in oriented architecture. Nutch is an extension of Lucene, a framework also from Apache that offers an API for textual indexing and the search system core. Hence, Nutch adds some functionality to Lucence, such as the capacity of retrieving Web pages (crawling) and

analyzing documents (parsing) from different formats. All mentioned software has been developed using the Java technology.

Users submit their queries and visualize the results through WebSearcher that delegates the search requests for GeoSEn and Nutch core. Both use their indices to find documents that satisfy the query, textually, in Nutch's case (Index) and geographically, in GeoSEn's case (GeoIndex). The Geodatabase stores a variety of geographical information essential to the spatial scope modeling and search process. In addition to Geodatabase information, GeoSEn accesses external services through Web services. Nutch offers several extension points, where some were chosen to access GeoSEn core functionalities, via API. Therefore, GeoSEn becomes independent from Nutch architecture, since its business layer is implemented in the core instead of directly in a plug-in. The GeoSEn plug-ins to be developed for Nutch are:

- **GeosenParser**, responsible for detecting geographical terms in the parsing process, the base for documents geographical scope analysis process;
- **GeosenOntology**, provides geographical semantics in the information retrieval process;
- **GeosenQueryFilter**, adds capacity of interpreting place descriptions in user queries;
- **GeoSenURLFilter**, offers geographical references and spatial relationships detection mechanisms in accessed URLs, during the crawling process.

The database that contains numerous place names and other information used by the parser (GeoDatabase) was implemented using the PostgreSQL database server, with its TSearch2 extension for indexing textual data. Currently, the parser capability of associating trust levels to place names functionality has been developed based on rules such as term frequency, abbreviation rate, and special terms existence in texts.

FUTURE TRENDS

The evolution of spatial search engines is directly related to the advances in spatiotemporal indexing, aiming scalability; more precise relevance ranking; and better extraction of the geographic scope in Web documents. It will be necessary for not only the advent of new techniques addressing those issues, but also the availability of spatially aware Web services. Hence, the service oriented architecture (SOA) will play an important role in the next generation of spatial search engines.

Another important issue to the geographical information system (GIR) area success is the definition of standards for place name ontologies and gazetteers, including a global scope and multilingual capabilities.

Other areas will continue to contribute to GIR, such as natural language processing, by promoting easy user interactions on spatial queries; and a strong and complete support for the temporal dimension, which will facilitate spatiotemporal interactions.

Lastly, the use of such spatial search engines in mobile devices will improve the quality of data manipulated by these devices; thus, an important research issue is how to integrate such spatial search engines with current location-based service architectures. Again, the use of SOA seems to be a good solution.

CONCLUSION

Search engines have become very popular nowadays and they are used ubiquitously by Web users. As most of the information available has a spatial footprint, the development of spatial search engines seems to be one of the next great waves in information retrieval techniques.

This chapter presents an overview of spatial search engines and discusses the GeoSEn prototype. We believe that by combining the ideas from traditional IR systems with those from geographical information systems will enhance the way of doing search on the Web.

REFERENCES

Baeza, R.Y., & Ribeiro, B.N. (1999). *Modern information retrieval.* New York: ACM Press Book.

Buyukkokten, O., Cho, J., Molina, G. H., Gravano, L., & Shivakumar, N. (1999). Exploiting geographical location information of Web pages. In *WebDB (Informal Proceedings)* (pp. 91-96).

Chaves, M. S., Silva, M. J., & Martins, B. (2005). A geographic knowledge base for semantic Web applications. In *Simpósio Brasileiro de Banco de Dados* (pp. 40-54), Uberlândia.

Ding, J., Gravano, L., & Shivakumar, N. (2000). Computing geographical scopes of Web resources. In *26th International Conference on Very Large Databases* (pp. 445-456), Cairo, Egypt.

Fu, G., Jones, C. B,. & Abdelmoty, A. I. (2005). Building a geographical ontology for intelligent spatial search on the Web. In *Proceedings of IASTED International Conference on Databases and Applications*(pp. 167-172), Innsbruck, Austria.

Gravano, L., Hatzivassiloglou, V., & Lichtenstein, R. (2003). Categorizing Web queries according to geographical locality. In *Proc. of the 12th ACM Conference on Information and Knowledge Management* (pp. 325-333).

Jones, C. B., Abdelmoty, A. I., Finch, D., Fu, G., & Vaid, S. (2004). The SPIRIT spatial search engine: Architecture, ontologies and spatial indexing. In Proceedings of Third International Conference on Geographic Information Science – GIScience, Maryland, USA. *Lecture Notes in Computer Science, 3234*, 125-139.

Jones, C. B., Alani, H., & Tudhope, D. (2001). Geographical information retrieval with ontologies of place. In Proceedings of Fifth Conference on Spatial Information Theory – COSIT, Morro Bay, CA, USA. *Lecture Notes in Computer Science, 2205,* 323-335.

Kowalsky, G. (1997). *Information retrieval systems: Theory and implementation.* Kluwer Academic Publishers.

Markowetz, A., Brinkhoff, T., & Seeger, B. (2004). Geographic information retrieval. In *3rd International Workshop on WebDynamics* (pp. 1-10). New York.

Markowetz, A., Brinkhoff, T., & Seeger, B. (2005). Exploiting the Internet as a geospatial database. In *International Workshop on Next Generation Geospatial Information* (pp. 5-14), Balkema Publishers.

Markowetz, A., Chen, Y. Y., Suel, T., Long, X., & Seeger, B. (2005). *Design and implementation of a geographic search engine*. Technical Report, CIS Department, Polytechnic University.

McCurley, K. S. (2001). Geospatial mapping and navigation of the Web. In *Tenth International World Wide Web Conference* (pp. 221-229).

Page, L., & Brin, S. (1998). The anatomy of a large-scale hypertextual Web search engine. *WWW7/Computer Networks, 30*(1-7), 107-117.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the Web*. Technical Report SIDL-WP-1999-0120, Stanford Digital Library.

Silva, M. J., Martins, B., Chaves, M. S., Afonso, A.P., & Cardoso, N. (2006). Adding geographic scopes to Web resources. *CEUS – Computers, Environment and Urban Systems, 30*(4), 378-399.

KEY TERMS

Crawler: Also known as robot or spider, it is a module of a search engine that is responsible for visiting Web sites and extracting their content to be further indexed by the search engine.

Gazetteer: A dictionary that translates a set of spatial coordinates to a place name and vice-versa.

Geocoding: A technique to provide spatial coordinates to places.

Geographical Information System: Software, for implementing geoprocessing, that is able to store and manipulate spatial data.

Indexing: A data structure technique used to speed up querying in large datasets.

Multimodal Interface: User-centered interface in which the computer may process more than one mode of communication.

Search Engine: Software that enables one to find documents on the Web according to user query.

Sponsorship in IT Project Management

David Bryde
Liverpool John Moores University, UK

David Petie
Petie Ltd., UK

INTRODUCTION

Since the 1970s academics and practitioners in the discipline of project management have sought answers to two inter-related questions: How is project success defined and measured? What are the influences on project success? To answer the first question people have studied project success criteria/key performance indicators. To answer the second, studies have focused on project critical success factors. Daniel (1961) introduced the concept of "success factors," stating that "in most industries there are usually three to six factors that determine success; these key jobs must be done exceedingly well for a company to be successful" (p.116). Approaches to the management of information have been established using Daniel's concept. For example, Rockart (1979) developed a Critical Success Factor (CSF) method for meeting the information needs of top executives. This method focused on understanding the objectives and goals of the company and the factors (CSFs) critical to their achievement, and establishing information systems to report on performance in these two areas. A key challenge has been to integrate the definitions and measures of success with CSFs, and in this respect work has been carried out to develop frameworks linking models of success criteria (the measures of success) with CSFs (see, for example, van Veen-Dirks & Wijn, 2002). The concept of CSFs has also been applied to project environments, with project CSFs being "those inputs to the management system that lead directly or indirectly to the success of the project" (Cooke-Davies, 2002, p. 185). Project management theory has also looked for a holistic answer to the questions of "How is project success defined and measured?" and "What are the influences on project success?", through the development of models linking project success criteria and project CSFs (Westerveld, 2002; Bryde, 2003).

BACKGROUND

Sponsorship as a Project Critical Success Factor

In respect of individual project CSFs, the importance of project sponsorship to achieving successful project outcomes has long been recognized. In a review of previous studies of CSFs, the sponsorship of projects by top management was highlighted as one of 8 major influences on success (Pinto & Slevin, 1987) and confirmed in a later study by the same authors as one of 10 influences (Pinto & Slevin, 1989). The importance of sponsorship is recognized through the distinction made between Macro CSFs, which involves activities in the realm of the sponsoring organisation and Micro CSFs, which are carried out in the domain of the project team (DeWitt, 1988). This crucial role of sponsorship has been identified in various manufacturing and service-related business environments, such as defence (Tishler et al., 1996), construction (Black et al., 2000), research & development (Pinto & Slevin, 1989) and management consultancy (Jang & Lee, 1998). Studies of project CSFs in IT environments have confirmed the pivotal influence of project sponsorship (see, for example, Bytheway, 1999; Fui-Hoon Nah et al., 2001; Procacinno et al., 2002).

ROLES OF PROJECT SPONSOR AND PROJECT MANAGER

A key step to delivering successful outcomes is gaining an understanding of the perspectives of all stakeholders to the project, including the sponsor and other stakeholders (Wright, 1998; Wateridge, 1995). In IT environments a lack of understanding has contributed to projects being unsuccessful. A failure by project managers to understand that users emphasized longer-term criteria relating to delivering workable systems, rather than short-term criteria linked to meeting time and cost objectives was a characteristic of IT projects perceived to be unsuccessful (Wateridge, 1998). A necessary step in achieving understanding is defining and delineating the roles and responsibilities of the project spon-

Table 1. Role of the project sponsor

- Define the business benefit/requirements
- Understand the risks to benefit realisation
- Agree the project definition, including project objectives
- Develop the project strategy, including priorities
- Help define the project success criteria
- Specify any constraints
- Determine the relative priorities of cost, time and quality
- Monitor the project's business environment
- If necessary, re-define or cancel the project
- Monitor project performance
- Take delivery at project completion
- Monitor benefit realization
- Champion the project, including making resources available
- Support the project manager in their role

Table 2. Role of the project manager

- Develop an effective working relationship with sponsor
- Deliver the project to time and cost, quality objectives
- Evaluate the risk profile and advise the sponsor
- Meet the defined project success criteria
- Manage the sponsor's and other stakeholders' expectations
- Define the project
- Build and lead the project team
- Monitor and control project progress
- Keep sponsor informed of progress and problems
- If necessary, recommend redefining or canceling of the project
- Hand over to the sponsor on completion

sor and project manager (Belassi & Tukel, 1996). *Table 1* summarizes the role of the project sponsor, drawing from the following extant literature: Snowdon (1976), Kliem & Ludin (1992, pp.163-169), Morris (1994, pp.188-189, 258-259), Briner et al. (1999, pp.65-67), Turner (1999, pp.50-53), and Hall et al. (2003).

The role of the project manager is summarized in *Table 2*. This table was constructed with reference to Gaddis (1959), Middleton (1967), Mantel et al. (2001, pp. 27-34), Anderson & Merna (2003), and Kendra & Taplin (2004).

FUTURE TRENDS

Although the importance of the sponsor to achieving project success is now well established (see earlier section "Background - Sponsorship as a Project Critical Success Factor"), and the roles of the sponsor and project manager are fairly well defined in theory, there are a number of critical issues that still need to be addressed in the future.

Awareness Among Project Sponsors of Their Role in Benefit Realization

Firstly, in some project organizations the role of the sponsor, especially in relation to their relationship with the project manager, is not clearly understood. This can lead to problems. For example, the responsibility for monitoring benefit realization (a sponsor role shown in *Table 1*) is often abdicated (without any corresponding authority) to project managers. However, it must be remembered that project managers only deliver products. A failure by the sponsor to fulfill their role in relation to benefit realization will lead to sub-optimal performance from the strategic perspective of the organisation.

Integration of Project Sponsor and Project Manager Perspectives

Secondly, there is the practical difficulty of integrating the perspectives of the sponsor and the project manager. The

sponsor should be focused on the benefits of an IT system, its impacts on the organisation and its contribution to the company vision. However the project manager should be focused on delivering products, IS/IT functionality and how it will be used operationally. The need for such a strategic and tactical integration underpins work on the development of new models for the delivery of successful IT projects (for example, Byers & Blume, 1994; Ward & Elvin, 1999). Project managers, by their very nature, are backward looking and tend to focus on what has been done, what has been spent and the problems they have delivering the product. Sponsors, on the other hand, must be forward looking, trying to ensure that the benefits will be delivered and that nothing gets in the way of this.

Distinction Between "Executive" Project Sponsorship and the Sponsorship of Individual Projects

Finally there is the question of what is meant by executive sponsorship? There is a difference between the sponsorship provided by one person to an individual project and the sponsorship given by the organisation to enable a strategic vision to be linked with effective delivery of an IT system. This second type of sponsorship, which is a potentially new paradigm, involves company directors, senior executives and middle managers understanding how to lead and manage change. It focuses on ensuring the organisation is ready for a project to be undertaken. For example, an exploratory study showed that the greatest influence on achieving successful outcomes from business process re-engineering initiatives was the innovative capacity of the organisation (Teng et al., 1998). In a similar fashion, in the new paradigm, the sponsors of projects must ensure that organizational competency issues linked to project management, in such areas as programme management, benefits management, sponsoring individual projects, long-term planning and governance, training and support, are properly addressed.

To ensure success, most organizations intuitively restrict projects to a particular department or area, with little or no cross-functional interaction. This is purely a competency issue, largely relating to executive sponsorship rather than project management competencies. The line of least resistance is to manage work packages or small projects with limited impact on other areas of the business. However, successful organizations are able to manage complex and inter-related projects through effective sponsorship. It is not hard to appreciate that decisions made regarding a particular initiative will have ramifications in other parts of the business. These elements need to be managed by individuals who have the interests of the organisation as a whole at heart, rather than a specific initiative. Effective executive sponsorship achieves this. Project managers cannot be expected to manage these aspects as well as deliver the projects on time, to the right quality and at the right cost.

This executive sponsorship equates to demonstrating effective leadership and being a good corporate citizen. A real test of this is when projects get cancelled because the benefits cannot be realized or the risk profile is too great.

CONCLUSION

Over the last 30 years there has been much work carried out in establishing the importance of project sponsorship to achieving successful outcomes and in defining the theoretical roles of the project sponsor and the project manager. This has resulted in a "traditional" view of the project sponsor and project manager roles, which in the case of the sponsor tends to focus on the sponsorship of individual projects. A further element, which was discussed earlier, is that of "executive" sponsorship, which focuses on organizational competency issues linked to project management. Further study is needed to establish the extent to which the two types of sponsorship, "traditional" and "executive," are undertaken in practice and to explore the relationship between different types of sponsorship and project success. This will require the taking of a holistic view of "project success," that is, incorporating project management-related criteria (meeting cost, time and quality objectives), satisfaction-related criteria (for the customer, project team and user) and benefits-related criteria (tangible and intangible). From this holistic perspective it will be possible to investigate how variations in the nature of project sponsorship impacts on the different success criteria, with the most effective sponsorship being that which meets all three.

Another focus for future work in this area needs to be on how project sponsors and project managers work together to ensure that IT project performance is optimized and benefits are realized. To achieve such optimization it may also be necessary to redefine the sponsorship paradigm to focus on addressing organizational competency in project management. For this to be useful in terms of project management practice there will need to be accompanying education, training and awareness-raising activities of the broader sponsorship role.

REFERENCES

Anderson, D.K., & Merna, T. (2003). Project management strategy: Project management represented as a process based set of management domains and the consequences for project management strategy. *International Journal of Project Management, 21* (6), 387-393.

Belassi, W., & Tukel, I. (1996). A new framework for deter-

mining critical success/failure factors in projects. *International Journal of Project Management*, *14* (3), 11-151.

Black, C., Akintoye, A., & Fitzgerald, E. (2000). An analysis of success factors and benefits of partnering in construction. *International Journal of Project Management*, *18*, 423-434.

Briner, W., Hastings, C., & Geddes, M. (1999). Project Leadership (2nd ed.). Aldershot: Gower.

Bryde, D.J. (2003). Modelling project management performance. *International Journal of Quality & Reliability Management*, *20* (2), 228-245.

Byers, C.R., & Blume, D. (1994). Tying critical success factors to systems development. *Information & Management*, *26* (1), 51-61.

Bytheway, A. J. (1999). Successful software projects and how to achieve them. *IEEE Software, 16* (3), 15-18.

Cooke-Davies, T. (2002). The "real" success factors on projects. *International Journal of Project Management*, *20* (3), 185-190.

Daniel, R.D. (1961). Management information crisis. *Harvard Business Review*, *39* (5), 111-121.

De Witt, A. (1988). Measuring project success. *International Journal of Project Management, 6* (3), 164-170.

Fui-Hoon Nah, F., Lee-Shang Lau, J., & Kuang, J. (2001) Critical factors for successful implementation of enterprise systems. *Business Process Management Journal*, *7* (3), 285-296.

Gaddis, P.O. (1959). The project manager. *Harvard Business Review*, *37* (3), 89-98.

Jang, Y., & Lee, J. (1998). Factors influencing the success of management consulting projects. *International Journal of Project Management, 16* (2), 67-72.

Kendra, K.A., & Taplin, L.J. (2004). Change agent competencies for information technology project managers. *Consulting Psychology Journal: Practice and Research*, *56* (1), 20-34.

Kliem, R.L., & Ludin, I.S. (1992). *The people side of project management*. Aldershot: Gower.

Hall, M., Holt, R., & Purchase, D. (2003). Project sponsors under new public management. *International Journal of Project Management*, *21* (7), 495-502.

Mantel, S.J., Meredith, J.R., Shafer, S.M., & Sutton, M.M. (2001). *Project management in practice*. New York: John Wiley & Sons.

Middleton, C.J. (1967). How to set up a project organization. *Harvard Business Review*, *45* (2), 73-83.

Morris, P.W.G. (1994). *The management of projects*. London: Thomas Telford.

Pinto, J.K., & Slevin, D.P. (1987). Critical factors in successful project implementation. *IEEE Transactions on Engineering Management*, *34* (1), 22-27.

Pinto, J.K., & Slevin, D.P. (1989). Critical success factors in R&D projects. *Research Technology Management*, pp. 31-35.

Procaccino, J.D., Verner, J.M., Overmyer, S.P., & Darter, M.E. (2002). Case study: Factors for early prediction of software development success. *Information and Software Technology, 44* (1), 53-62.

Rockart, J.F. (1979). Chief executives define their own data needs. *Harvard Business Review*, *57* (2), 81-93.

Snowdon, M. (1976). Project management in the early stages. *Engineering & Process Economics*, *1* (4), 257-264.

Teng, J.T.C., Fiedler, K.D., & Grover, V. (1998). An exploratory study of the influence of the IS function and organizational context on business process reengineering project initiatives. *Omega*, *26* (6), 679-698.

Tishler, A., Dvir, D., Shenhar, A., & Lipovetsky, S. (1996). Identifying critical success factors in defense development projects: A multivariate analysis. *Technological Forecasting and Social Change*, *51* (2), 151-171.

Turner, J.R. (1999). *The handbook of project-based management* (2nd Ed.). London: McGraw-Hill.

Van Veen-Dirks, P., & Wijn, M. (2002). Strategic control: Meshing critical success factors with the balanced scorecard. *Long Range Planning*, *35* (4), 407-427.

Ward, J.W., & Elvin, R. (1999). A new framework for managing IT-enabled business change. *Information Systems Journal*, *9*, 197-221.

Wateridge, J. (1995). IT projects: A basis for success. *International Journal of Project Management, 13* (3), 169-172.

Wateridge, J. (1998). How can IS/IT projects be measured for success? *International Journal of Project Management, 16* (1), 59-63.

Westerveld, E. (2003). The project excellence model: Linking success criteria and critical success factors. *International Journal of Project Management, 21* (6), 411-418.

White, D., & Fortune, J. (2002). Current practice in project management: An empirical study. *International Journal of Project Management*, *20* (1), 1-11.

Wright, J.N. (1998). Time and budget: The twin imperatives of a project sponsor. *International Journal of Project Management, 15* (3), 181-186.

KEY TERMS

Organisational Competency in Project Management: Ensuring that the organisation is in a ready state in order for projects to be able to deliver benefits.

Project Critical Success Factors: The influences on the success, or otherwise, of a project. A distinction can be made between the underlying factors, or causes of success or failure and the symptoms of an ineffective project management process. The lack of top management support is a typical project critical success factor, which can lead to a variety of symptoms, such as adequate resources not being made available to the project.

Project Manager: An individual with the responsibility of ensuring the project objectives are delivered.

Project Performance: The degree to which the project meets its overall objectives. This compares with *project management performance*, which is the degree to which the traditional objectives of cost, time and quality are met.

Project Sponsor: An individual or group with the responsibility and authority to ensure that the project benefits are realised.

Project Stakeholder: Any person or group that has an interest in the project. The interest could be in the project outcome, outputs or the project management process.

Project Success Criteria/Key Performance Indicators: The measures of success. The terms project success criteria and project key performance indicators are used interchangeably. Traditional measures are meeting cost, time and quality objectives. Other measures are linked to the attributes used by a stakeholder to judge whether their expectations have been met.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2597-2601, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Spreadsheet End User Development and Knowledge Management

Anders Avdic
Örebro University, Sweden

INTRODUCTION

In the early days of computers, expertise was needed in order to use computers. As IT tools have become more powerful and user friendly, more and more people have been able to use computers and programs as tools when carrying out working tasks. Nowadays, it is possible for people without special IT training to develop Information Systems (IS) that only IT specialists could have done some years ago.

In this paper End User Development (EUD) using a Spreadsheet Program (SP) is discussed from a knowledge management perspective. *EUD* can be a part of an organization's effort to take advantage of existing, often tacit, knowledge or creating new knowledge. An end user is a person who acts both as a user and a systems developer. A typical feature of an end user is that he has a good (often unique) knowledge of the business and the work related to the IS in question, which is called the User Developed Application. It is the combination of these two sorts of knowledge which is the key to EUD as knowledge management.

The aim of this of this chapter is to relate EUD to knowledge management and, specifically, to describe how tacit knowledge can be audited when end users develop spreadsheet systems for their own domain of expertise. The main source is a set of qualitative case studies carried out between 1995 and 2005. (Avdic, 1999; Westin, Avdic & Roberts, 2005)

BACKGROUND

There are many reasons for professionals to use personal IT tools such as spreadsheet programs in their daily work. One is to increase their knowledge and understanding of their professional domain in a changing world. The use of spreadsheet programs in order to develop systems for decision making is an alternative to more traditional systems development, where IT specialists assist in specifying information needs and other requirements as a basis for a systems development process where the IT specialist is the developer. This "traditional" procedure is often the only option since developing the system, especially its technical parts, is complicated and not possible to carry out by non-IT specialists. When the technical dimensions are uncomplicated, there are some interesting benefits for professionals to develop their own systems.

End User Development is herein defined as "…the use of the adoption and use of information technology by personnel outside the information systems department to develop software applications in support of organizational tasks." (Brancheau & Brown, 1993) In our case studies, we have studied end users developing systems using spreadsheet programs. The end users have worked as controllers, administrators, civil servants, production planners, and managers in private companies and local government organizations. Some studies have been longitudinal, the longest one ten years. One thing they have had in common is that they are all professional experts in their domains. All of them have had the possibility, to a large extent, chose their own working situation and how to carry out the way they work.

Knowledge Management "is the name given to the set of systematic and disciplined actions that an organization can take to obtain the greatest value from the knowledge available to it." (Marwick, 2001) Focus here is more on "obtain greatest value" than "systematic and disciplined." The forms of EUD that are discussed here are often related to groups that could be described as *Communities of Practice* (Wenger, 1998) and whose development activities, to a large extent, are about making *tacit knowledge* explicit. (Polanyi, 1967) Among the practitioners' working situation and domain of expertise, there has been organized and disciplined activities in order to explore the knowledge domain by developing spreadsheet programs in various forms of cooperation with colleagues and peers. This is, to a large extent, an expression of social learning, which is a fundamental concept regarding communities of practice. (Wenger, 1998)

The knowledge management approach discussed in this chapter is more inductive than deductive in the sense that development activities are not planned or organized by the management but by the end user from his practitioner perspective. This is also a typical way for communities of practice to exist. The process of creating and sharing knowledge is complex, abstract, and subtle and, to a certain degree, a tacit process in itself. In accordance with Walsham (2001) we believe that knowledge management processes of the kind we discuss benefit from putting the human before IT.

Organizations where EUD activities take place are decentralized organizations where skilled practitioners have

the mandate and resources to independently develop various sorts of applications. Decentralization of decision making is a necessary condition "…when tacit and detailed knowledge is involved in opposite to explicit aggregated knowledge." (Grant, 1997)

The distinction between *tacit* and explicit *knowledge* is originally presented by Polanyi (1967) and is frequently used ever since. The notion of tacit knowledge refers to knowledge, experiences, and abilities that are not able to represent or codify, while explicit knowledge is possible to represent in, for example, books and computer programs. Some claim that most knowledge (50% - 95%) is tacit. (Awad & Ghaziri, 2004) Since tacit knowledge is not explicit, it is not possible to question. When end users make tacit knowledge explicit, it becomes, at the same time, open to inspection. This is an important part of EUD and the focus of this chapter.

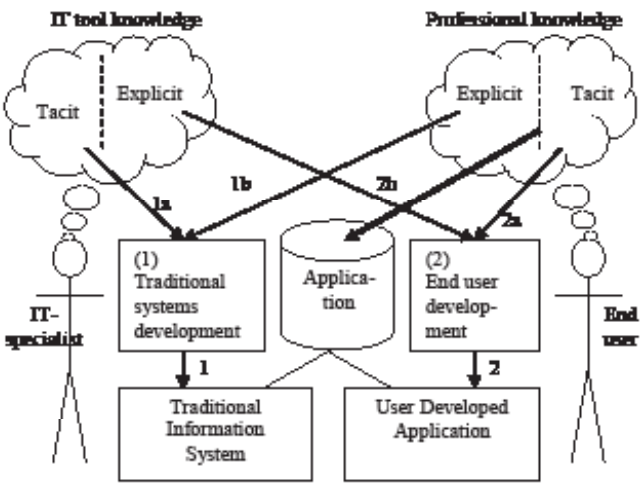
MANAGING KNOWLEDGE BY SPREADSHEET END USER DEVELOPMENT

Traditional and End User Systems Development

EUD is often compared to *systems development* carried out by IT specialists. This is not always relevant, but below we are applying the comparison in order to draw the attention to some central knowledge related properties inherent in EUD.

In Figure 1 the difference between traditional systems development (1) and EUD (2) is outlined. To the IT specialist, knowledge about systems development tools (e.g. methods, program languages) (1a) is in primary focus when developing an IS (1c). This is the core of his/her professional knowledge. Knowledge about business (1b) is, of course, essential but not primary. When the IT professional starts the next project, his/her basis is his/her IT specialist expertise but another business context is focused. To the end user, knowledge about business (2a) is of primary importance and knowledge about systems development tools (2b) is just a means to accomplish business-oriented tasks, eventually by developing user developed applications (2c). The IT specialist has access to knowledge about IS development tools that are hard to access for non-professionals, since they consist of tacit knowledge together with explicit knowledge. Some business knowledge is hard to access to the IT specialist, since this knowledge is not in the professional knowledge domain of the IT specialist. The end user, on the other hand, is the expert on business knowledge. His/her expertise depends on his/her knowledge about business. This professional knowledge also consists of tacit and explicit knowledge. No one can replace him/her in this matter. In order to perform EUD, the end user needs some knowledge about IS development tools. It is not possible, though, to have access to as much

Figure 1. The relation between knowledge and development



knowledge about IS development tools as the IT specialist has, since it is partly tacit.

To both the IT specialist and the end user, both kinds of knowledge are necessary. In order to make an IS though, the overall most important kind of knowledge is in general knowledge about business, since the IS is about the business and is supposed to be used in the business context. The bold arrow in Figure 1 demonstrates this circumstance.

In order to *develop information systems*, business specialists must share knowledge about business with IT specialists. This process is problematic since people have different frames of references. (Alter, 1996) The entire intention, closely related to professional ethics of the business specialist can, therefore, not be perceived by the IT specialist. The IT specialist can, therefore, not fulfill the requirements since he/she cannot completely understand the business specialist. Since the business specialist's knowledge about business to a certain extent is tacit, he/she cannot tell what he/she knows.

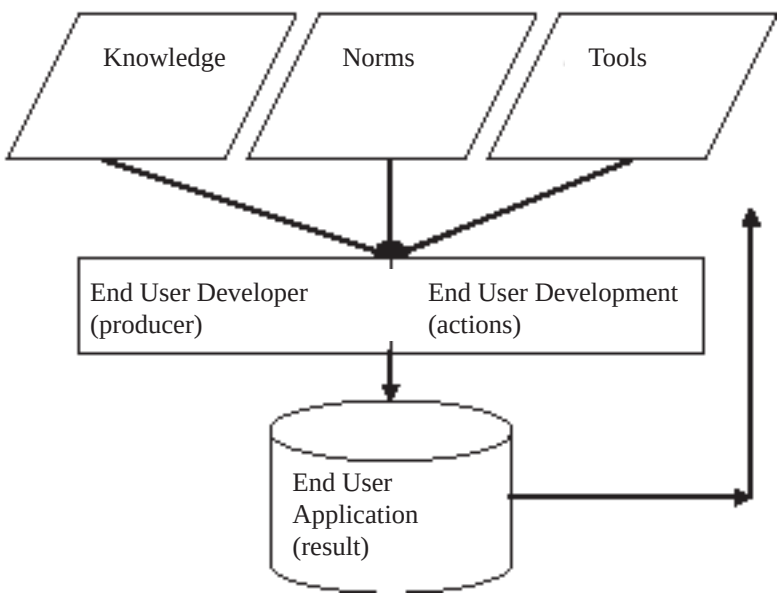
Even though the described problems of knowledge sharing are well known, complex systems development tasks still have to be performed the traditional way. The problem is addressed in various ways, for example, participative approaches or agile systems development methods. But as more powerful systems development tools are at hand, the possibilities to perform EUD are enhanced. Spreadsheet programs have properties that give the end user access to IS development features without the demands of being an IT specialist. The systems discussed here are often small and local, and thereby not suitable for traditional systems development projects.

The End User Development Practice

The practice of the end user is described below using the (modified) *model of generic practice* (the ToP model) (Goldkuhl, 2005) in order to systemize empirical findings and related theory. The model can be used to specify the conditions and result of a specific practice, for example, a controller practice or an IT specialist practice. The model consists of a set of conditional categories: *knowledge, norms, and tools.* The categories that express the specific practice are named *producers* and their *actions.* The last category is the *result* of the practice. When an end user develops an application he/she acts in at least two kinds of practices; the primary (e.g., controller) practice, and the secondary (developer's) practice. Each practice is related to a profession, for example, a controller and an IT specialist profession. The model makes it possible to separate the conditions of the different practices. It also makes it possible to discuss which parts of the developer's practice can improve the main practice without consulting an IT specialist. The nature of the categories is described in Figure 2 together with presentation of findings from the case studies.

A *user developed application* is an IS and an IS is a *result* of systems development. The difference between a

Figure 2. End user development as practice



traditional IS and a user developed application is mainly a question of how it is built. User developed applications are built by end users with a good knowledge of the business, while traditional information systems are built by IT specialists. (Avdic, 1999)

Traditional *Systems Development* can be characterized by the notion of the "Life Cycle", where tasks are specialized and activities are separated and systemized. *End User Development* (*actions*) and traditional development are profoundly different. EUD actions are seldom organized or planned. Specific work related tasks or problems make the end user aware of some information related requirements. EUD is by the end user perceived as work rather than systems development.

Risks are obvious when end users with limited IT skills start to develop systems that affect vital organizational decisions. Panko (2006) have shown that errors in spreadsheet systems are common and that there is a need for finding ways to prevent and find these errors.

An *end user* (*producer*) is a person with a deep knowledge of the business who develops user developed applications that support the end user in his/her work. The end user is primarily a professional (e.g., a controller) who integrates, to some extent, the role of one or more IT specialist, when performing EUD. The end user could have good knowledge about IS development tools. This does not disqualify him as an end user; it rather makes him even more efficient. When several end users work together with a common goal in order to understand and solve problems in their own domain, it can be described as a Community of Practice (Wenger, 1998).

When performing EUD, the *knowledge* necessary in the practice of the IT specialist practice to carry out the development, is divided between the end user and the tool (the spreadsheet program). IT-related knowledge that is not too complex, is formalized into the spreadsheet program (e.g., arguments of statistical functions and graph drawing) and can be used in the application. Other kinds of knowledge can be formalized by the end user into the application which is a knowledge condition of the business practice. Some kinds of knowledge (e.g., of critical evaluation of the relevance of formulas) cannot be formalized at all. Still, this kind of not-easily-formalized (sometimes tacit) knowledge can be taken into consideration when using the application, since the end user (with business knowledge) is the user of the system. We also claim that goals and norms, not easily formalized, can be taken into consideration when performing EUD.

Knowledge about tools can be used to deepen knowledge about business. End users make tacit knowledge explicit when developing EUD's, which in turn make it possible for others to evaluate and criticize the application and its output. One important aim of the end user is to articulate knowledge about business and EUD is one mean to do this.

Norms and knowledge are closely related and sometimes hard to keep apart. One set of norms that are central in a EUD context is professional ethics. Professional ethics are crucial to the end user, since the professionals' activities are monitored not by procedures but by professional and business ethics. Professional ethics as well as professional tacit knowledge cannot easily be transferred to IT specialists in systems development projects. Therefore, when EUD is performed by end users, professional ethics and tacit knowledge can be taken into consideration in a way not possible in traditional systems development. EUD can also change organizational norms. Ongoing questioning of business, using user developed applications, can implicitly or explicitly challenge existing models as well as their norms. (Avdic, 1999) This does not mean that revolutionary effects take place every time an end user develops an application. See also the section entitled "Processing the tacit".

EUD *tools* are closely related to norms and knowledge, since norms and knowledge are implemented in tools. The main tool when performing spreadsheet EUD is, of course, the *spreadsheet program*. The spreadsheet program integrates functions for input, output, storage, processing, and presentation. This integration results in interactive development and use. The open nature of the spreadsheet program can and does cause different kinds of errors. (Panko, 2006) Knowledge of business, tools, and design can prevent some of these errors. Another circumstance that makes spreadsheet programs suitable for EUD is the fact that they are very common. In Sweden, for example, almost all employees who might need it can have access to a spreadsheet program.

Because of the integrated nature of EUD, learning, using, and systems development take place at the same time. Learning applies to both the business and the tool. One conclusion of this is that training in the use of a tool can improve the quality of EUD, which in turn can improve business. One way for the management to support EUD, eventually in a Community of Practice, is to initiate and encourage end user-tailored training in the use of tools.

Processing the Tacit

Fahy & Murphy (1999) has shown that managers' ability to process their *tacit knowledge* by developing applications is rewarding in several ways. According to Fahy & Murphy (1999) managers are involved in "…activities that are characterized by high levels of tacit knowledge based on extensive experience." Development of systems is an important part of their efforts to understand and articulate their information requirements. "The process of developing systems provides managers with a mechanism for iterative analysis. This trial and error approach helps managers build up their understanding of the problem situation and allows them to be more explicit in describing and understanding the problem facing them." Göranzon (1990), on the other hand, shows evidence that when a professional domain (that of forest wardens) is formalized into a system that is

developed by someone else, there is a risk that professional knowledge disappears. He also states with examples from his study that the ability to calculate is closely related to the ability to assess and evaluate. This means that professional norms can be maintained by processing data.

According to Wulf & Jarke (2004), the initial costs for EUD exceeds those of traditional systems development. On the other hand, costs for "software adaptation in context of use" and, moreover, "costs for missed opportunities for appropriation" are larger for non-EUD applications. The reason for this is claimed to be the gap between domain experts and IT specialists.

FUTURE TRENDS

All over the world the use of computers, personal software, and Internet is increasing. More and more practitioners *develop end user applications*. According to Sutcliffe & Mehandijev, 2004) there were probably 55 million end users in USA 2005 in comparison to 2.75 million professional software developers. As knowledge of computer use is increasing, we can be sure that personal use of computers will increase. The growth of Internet use and maturity will promote an integration of Internet as an information source, which will open new possibilities for EUD.

A world that is changing faster and faster will put demands on managers and professional practitioners to explore how these changes affect their organization. To a certain extent this is possible with traditional systems, but due to the built-in inertia, these systems cannot provide means to explore aspects that are hard to define. Therefore, practitioners will have a need for personal IT tools like spreadsheet systems in the future as well. Actually, there are reasons to believe that organizations will more actively take advantage of communities of practice and support them (Wenger, 1998).

Finally, traditional systems development experts can benefit from EUD and integrate EUD in their development procedures in order to get access to tacit domain knowledge.

CONCLUSION

Organizational knowledge is to a large extent tacit. Due to ever ongoing changes in organizations and in the environment of organizations there is a need to understand current conditions in order to take decisions from what is as close as possible to the actual situation. Traditional *information systems* suffer from a certain inertia which sometimes makes them insufficient to produce current data. It is also rather obvious that existing systems cannot anticipate all future demands and there is, therefore, always a need for flexible and usable tools which make it possible for decision makers to collect, analyze, and present information in a not pre-specified way. Existing systems is a necessary condition but not always the most optimal way to serve as a tool for analyzing earlier, not explicitly known situations.

Increasing knowledge and decreasing costs of IT and IT tools have provided professionals with possibilities to process their explicit and tacit knowledge about their knowledge domain, their work, and their organization in order to create new knowledge and audit existing knowledge. EUD using *spreadsheet programs* provide opportunities to practitioners to explore and audit tacit knowledge.

REFERENCES

Alter, S. (1996). *Information Systems - A Management Perspective*. Benjamin/Cummings, Menlo Park, CA.

Avdic, A. (1999). *User and developer – End User Development Using a Spreadsheet Program*. Doctoral thesis, Linköping University, Sweden. [In Swedish]

Awad, E. M. & Ghaziri, H. M. (2004). *Knowledge Management*. Pearson, New Jersey.

Brancheau, J. C. & Brown, C.V. (1993). The Management of End User Computing: Status and Directions. *ACM Computing Surveys*, 25(4), 437-482.

Fahy, M. & Murphy, C. (1999). Managers, Information and Systems Development: Exploring the Tacit Dimension. In: *Information Systems at the Core: European Perspectives on Deploying and Managing Information Systems in Business*. Finnegan, P. and Murphy, C. (Eds.), Dublin: Blackhall, 188-204.

Goldkuhl, G. (2005). Workpractice Theory – What it is and Why we need it. *Proceedings of ALOIS 2005*, Limerick Ireland.

Göranzon, B. (1990). *Det praktiska intellektet* Carlssons, Stockholm. [In Swedish]

Grant, R. M. (1997). The knowledge-based view of the Firm: Implications for Management Practice. *Long Range Planning*, 30(3), 450-454.

Marwick, A.D. (2001). Knowledge Management Technology. *IBM Systems Journal*, 40(4), 814-830.

Panko, R. (2006). Spreadsheet Research, http://panko.shidler.hawaii.edu/SSR/index.htm. Accessed 2007, December 11.

Polanyi, M. (1967) *The Tacit Dimension*. Garden City, N.Y.: Doubleday.

Sutcliffe, A. & Mehandijev, N. (2004). End-User Development, *Communications of the ACM*, 47(9), 31-32.

Walsham, G. (2001). Knowledge Management: The Benefits and Limitations of Computer Systems. *European Management Journal*, 19(6), 599-608.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*, Cambridge: Cambridge University Press.

Westin, O. Avdic, A. & Roberts, H. (2005). Lookin' for a reason: local knowledge and the changing of control practices. Proceedings of ENROAC05, Antwerp.

Wulf, V. & Jarke, M. (2004). The Economics of End-User Development. *Communications of the ACM*, 47(9), 41-42.

KEY TERMS

Communities of Practice: Groups of people in organizations with common goals and a common repertoire of concepts and knowledge. They share what they know and learn from one each other regarding aspects of their work and provide a social context for that work.

End User Application: An information system developed by an end user developer in order to support him/her in his/her work.

End User Developer: A person with a deep knowledge of the business, who develops end user applications that supports him/her in his/her work.

End User Development: The adoption and use of information technology by personnel outside the information systems department to develop software applications in support of organizational tasks.

Explicit Knowledge: Codified knowledge.

Knowledge Management: A set of systematic and disciplined actions that an organization can take to obtain the greatest value from the knowledge available to it.

Tacit Knowledge: Knowledge, experiences, and abilities that are not articulated, represented or codified.

Category: IT Education

Standardization in Learning Technology

Maria Helena Lima Baptista Braz
DECIVIL/IST, Technical University of Lisbon, Portugal

Sean Wolfgand Matsui Siqueira
DIA/CCET, Federal University of the State of Rio de Janeiro (UNIRIO), Brazil

INTRODUCTION

The use of computers in education has been reported since the 1970s, but the Internet is fundamentally changing the way organizations operate, and these changes are spreading fast to educational organizations as they are eager to take advantage of the new possibilities.

In this context, new terms have been created to express new concepts related to the use of technology and following this trend, the term e-learning was coined. E-learning is an all-encompassing term generally used to refer to the use of technology in learning in a much broader sense than the computer-based training (CBT) or computer-aided instruction (CAI) of the 1980s. E-learning is extensively used and can include, just to name a few examples: educational Web sites; the use of hypermedia, discussion boards, e-mail, text chat, simulations and games in an educational context; computer-aided assessment; and learning management software. Although the term is not well defined and covers many possibilities, it has been mainly used when the Web is involved in the learning process.

The rapid growth in e-learning has led the community of designers, developers and users of learning resources to a point where they have an enormous variety of tools to support their work. However, if these tools use proprietary solutions, this would make the reuse of learning content outside the scope of the system where it was created difficult. It would also be hard to provide mechanisms for searching, accessing, reusing, and integrating such resources. One way to avoid these kinds of problems is the definition and use of open specifications and standards.

A standard is a set of technical definitions and guidelines for designers, manufacturers, and users, establishing the characteristics of a product, process, or service, such as dimensions, safety aspects, and performance requirements. Standards are written by experts with knowledge and expertise in a particular field (ASME, n.d). The Internet is a very good example of the importance of having standards to support the development and wide adoption of technology. It would have been impossible to connect so many different computers around the world if there were no standards to define the connections and communication protocols. This is also the case of e-learning technology, which needs standards in order to facilitate worldwide propagation.

Once e-learning standards are defined, accepted, and used, they will bring many advantages (Duval, 2004; MASSIE Centre, 2003):

- From the point of view of users, standards will prevent them from being locked-in to a particular vendor as it will be much easier to shift between tools and platforms and increase the reuse of existing resources.
- From the point of view of the tool vendor, they will not need to develop proprietary interfaces for other existing products lowering the cost and increasing the size of potential markets.
- From the point of view of content producers, they can use a standard format that will be understood by any delivery system conforming to the standards and increase the potential market of their products.

Above all, standards are a clear signal that a technology is mature and usually are seen as a first step towards a rapid growth phase and worldwide adoption.

BACKGROUND

It should be noted that there are different kinds of standards. *De facto* standards are developed by market—or technology-driven processes—and have come into use by general acceptance, custom, or convention, but have no formal recognition. When the standard is created by an accredited standard developing organization (SDO), it is called a *de jure* standard or an accredited standard and its scope can be national, regional, or international depending on the SDOs involved in its publication. Usually each country has a national standards body that is responsible for the adoption of standards within that country and that represents the country in international standards forums like ISO—International Standards Organization. ISO is a network of the national standards bodies of 157 countries, with one member per country and is a non-governmental organization with a Central Secretariat in Geneva, Switzerland (ISO, 2006).

When there is a product, which has been developed according to a standard, it is usually said to be conformant with the standard. If the product is tested by an external entity that verifies and attests the conformance (or conformity) of the

product to the standard, then it can be said to be certified.

The importance of a standard is usually measured by the extent of its actual acceptance and use. Nowadays, although accredited standards tend to have more credibility, *de facto* standards or even open specifications are commonly used as the need for standards is growing fast. This is the case in e-learning, where the most active standards groups are open consortia.

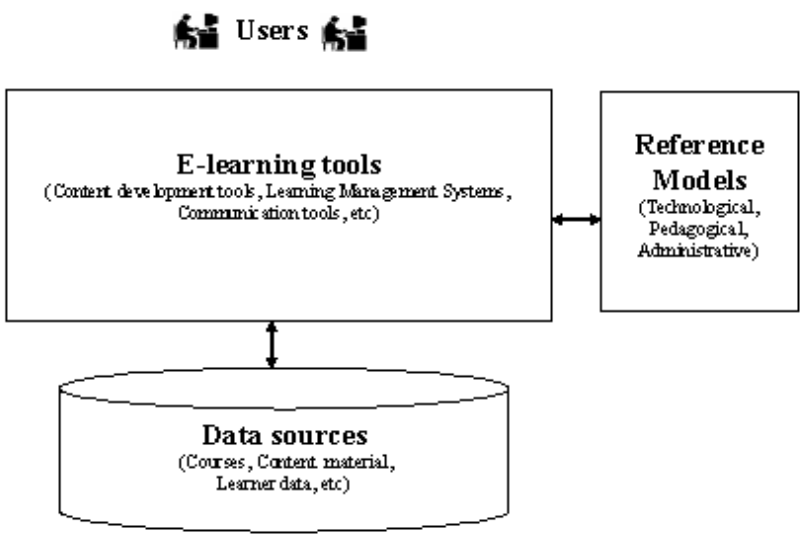
In e-learning, the development of standards usually follows four steps: specification, validation, standardization, and dissemination.

- In the specification phase, cooperating organizations develop initial specifications based on their analysis of the tasks accomplished during the learning process. All information exchanges and interfaces between components are specified, documented, and refined among all the participants.
- Based on the results gathered during the specification phase, pilot programs are developed to test the effectiveness of new products incorporating the initial specifications. This is the validation phase, where testbeds are established for validating conformance to the specifications and to reveal the existence of initial specifications that should be revised. Also, reference models are developed, showing how different specifications work together.
- The specifications that have been tested and proved to be valuable are then submitted for approval by accredited standards bodies that will refine, clarify, and follow the established procedures for reaching final accreditation of the specifications. In e-learning the bodies creating accredited standards are: IEEE Learning Technology Standards Committee (LTSC) (http://ieeeltsc.org/), ISO/IEC JTC 1/SC 36—Information Technology for Learning, Education, and Training (http://jtc1sc36.org/), and CEN/ISSS Learning Technologies Workshop (http://www.cen.eu). Within the bodies are working groups that address specific areas of e-learning and that coordinate all the internal process for reaching accreditation. Additional information about this process can be found in IEEE (2005).
- Finally, the last phase of the process is the dissemination of approved standards among all the stakeholders in e-learning, giving them support for understanding and correctly using the new standards.

CEN/ISSS Learning Technologies Workshop has created an observatory which reports the most important events in e-learning standardization available online at http://www.cen-ltso.net/.

To understand the relevant aspects that should be considered for standardization in e-learning, it is important to comprehend an e-learning environment. Figure 1 shows a simple model of the most important aspects of an e-learning environment. Users represent all the participants in the learning environment, including, for example, learners, teachers, content developers, and administrative staff. E-learning tools cover all the software and hardware systems that provide the necessary functionality for the environment and can include for instance such elements as a learning management system (LMS), content development tools, and communication tools. Data sources mean all the digital data that is needed to be able to fulfill the functionality of

Figure 1. A simplified framework of educational and training systems



the e-learning environment, including learner data such as personal data and grades, learning content materials, data about existing courses and conditions to be able to attend them, and so on. The reference models represent different administrative, pedagogical, and/or technological strategies, methods, and techniques adopted by the organization that can guide the learning processes. For instance, the pedagogical reference model could establish a learning approach based on an existing learning theory. Examples of the use of learning theories in online courses can be found in Modritscher (2006).

Considering this simplified framework of educational and training systems (Figure 1), it is possible to see how the different components play their part in an educational or training program. When starting the development of any educational or training program, the first step is to conduct a needs assessment. The developers can perform a needs assessment by accessing databases in order to find information revealing on which topics the students need more material, what professional or academic goals need further and/or deeper support and so on. Then, after choosing a new educational/training goal, knowledge about the audience must be acquired. This can be supported by a learner model and usually consists of accessing data from the learner's database. After this phase specific objectives are defined, and the system will support the design and development phases following the direction established by the course model and the pedagogical reference model.

If the needed content is not available, then its development will have to be considered. A widely accepted approach for authoring is related to grouping together pieces of information objects (Barritt & Lewis, 2001). This approach leads to the creation of what is generally called "learning objects" (LOs) and its main goal is to facilitate the development process and reuse of content. To fulfill this goal it is also essential to create metadata associated to the LOs. Metadata is simply defined as "data about data" and, in this case, it can be seen as structured, encoded data that describe characteristics of the LOs in order to allow faster search and easier access. It should be noted that authoring and composing tasks, in the LO approach, could be automated according to some special rules, allowing on-demand learning. In addition, there is the possibility that metadata could be automatically generated from the learning material. The learning content development process clearly includes the insertion of data and metadata into the content material database as well as possibly into the course database.

Another significant aspect connected to content development is the definition of learning activities. A learning activity may be loosely described as an instructional event or events related to content resources that will engage the student in executing a task that contributes to achieving learning objectives (Koper & Tattersall, 2005). The creation of learning activities is essential to achieve learning, and these activities should be tracked by e-learning tools and specified during educational or training program development.

Once an educational or training program is developed, there must be tools to allow its deployment. These tools are, of course, dependant on the teaching approach that was used during the design and development phases. In many educational and training systems the core functionality is to provide access to content material stored in the database on a previously specified sequence, following the visualization steps.

At this point, it is important to consider aspects such as user interface usability and information presentation, content hypermedia navigation, and content personalization and adaptation.

The goal of adaptation and personalization is to provide content according to the users' background and abilities which must be known by the tools and this capacity is now identified as an important factor in the improvement of pedagogical quality in e-learning (Brusilovsky & Peylo, 2003; Dagger, 2006; Paramythis & Loidl-Reisinger, 2004).

Another keyword is accessibility; this can be understood as the degree to which a system is usable by as many people as possible and now, in many countries, there are laws or regulations protecting users which specifically focus on people with disabilities and their access rights.

There are also systems based on constructivist principles. In this approach learning activities take a key role in the learning process, and there is a high degree of interaction among all the participants supported by communication and groupware tools. Other possible approaches include project-based learning, simulations, and educational games.

In addition, there are assessment tools that can be incorporated in the educational/training system in order to evaluate the students. Although assessment activities could be supported by groupware infrastructure or included as learning activities, it is interesting to consider specific assessment tools that could help teachers and students in the learning process.

Finally, security tasks and overall evaluation should be present on all the activities executed within the system. Course evaluation should lead to a revision on new iterations of the course while the evaluation of the educational and training system should lead to its development and improvement.

From this description it is easy to understand that data stored in the databases/data sources will be much more useful if standards exist. For instance, if the learner data has a standard format, all the tools that deal with this data can access that information regardless who originates it, who is maintaining it, or who uses it. In a more practical perspective, a learner having a standard learner data/profile can easily transfer this data from one environment to another conforming one. Therefore, the learner could have different personalized environments using and/or sharing the same learner profile.

From the point of view of the software tools, it is also possible to understand that it is essential that they are able to communicate and cooperate in order to provide all the services envisioned for the e-learning experience. This is equivalent of saying that interoperability is needed. Interoperability, according to ISO/IEC, is defined as follows: "The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units" (ISO, 2003).

E-LEARNING STANDARDS AND ORGANIZATIONS

The presented description, although not fully exhaustive, allows establishing the most significant aspects that standards for e-learning must address. First, general standards for interoperability of software and hardware are clearly needed as in any other information system. They cover such areas as data exchange formats, communication protocols, and services. These standards are obviously created independently—that is separately from the application area and, because of the role that the Internet has, nowadays many important standards are produced within W3C (World Wide Web Consortium) (http://www.w3.org). This consortium develops relevant technical standards that are referenced and used within learning technologies such as: HTTP (Hypertext Transfer Protocol), URL (Uniform Resource Locator), HTML (HyperText Markup Language), XML (eXtensible Markup Language), RDF (Resource Description Framework), OWL (Web Ontology Language) and SOAP (Simple Object Access Protocol), just to name a few.

Besides these standards and specifications, there are those that specifically address the area of e-learning. In this case, two crucial aspects for achieving a more flexible and interoperable e-learning environment can be noticed: learning content and learner data.

Considering the learning content, the main aspects to be standardized are:

- **Metadata:** To describe useful characteristics of the content such as author, language, and title.
- **Content Distribution Information:** To specify how that content can be accessed by existing tools.
- **Content Assembling:** To state how different units of content can be assembled and used by existing tools.
- **Learning Activities:** To allow e-learning tools to allocate activities and their associated resources to participants that play the various roles, and coordinate the runtime flow.

Considering learner data, the main aspects to be standardized are:

- **Personal Data:** To identify the learner, including information such as name and address.
- **Qualifications:** To show the courses or other studies that the learner has pursued and the corresponding grades.
- **Competencies:** To describe learner's skill.
- **Accessibility:** To cater for special needs of the students connected with personal preferences or possible disabilities.

With the development of the e-learning industry, many organizations understood the usefulness of having standards and, then, addressed this problem. IEEE LTSC (http://ieeeltsc.org/), IMS Global Learning Consortium (http://www.imsglobal.org/), Advanced Distributed Learning (ADL) Initiative (http://www.adlnet.gov/index.cfm), Aviation Industry CBT Committee (AICC) (http://www.aicc.org/), and Association of Remote Instructional Authoring and Distribution Networks for Europe (ARIADNE) (http://www.ariadne-eu.org) are important players within the standardization process and are now working in collaboration.

From the learning content perspective, two main results are broadly accepted and have been adopted by many e-learning developers: IEEE LOM (Learning Object Metadata) and SCORM (Sharable Content Object Reference Model).

IEEE LOM (Draft Standard, 2002) specifies a conceptual data schema that defines the structure of a metadata instance for a learning object. This metadata instance describes relevant characteristics of the learning object to which it applies. Such characteristics are organized in a hierarchy of elements that are grouped into nine classification categories: general, life-cycle, meta-metadata, educational, technical, rights, relation, annotation, and classification. The purpose of this standard is to facilitate search, evaluation, acquisition, and use of learning objects. IEEE LOM was approved as a standard by IEEE LTSC in 2002 and this work has been extended by other standards through the specification of bindings of LOM data model in XML and RDF (W3C standard languages/data models for Web documents).

SCORM is a standards reference model concerning the development, packaging and delivery of learning objects. There have been previous versions and the latest one is the SCORM 2004 3rd edition. This edition is a collection of specifications and standards that defines the interrelationship of content objects, data models, and protocols such that objects are sharable across systems that conform to the model.

It is interesting to note that SCORM is built on existing standards and specifications proposed by other organizations. For instance, SCORM 2004 metadata is defined using IEEE 1484.12.1-2002 LOM Standard and IEEE 1484.12.3

Standard for Extensible Markup Language (XML) Binding for LOM Data Model.

Besides these two results, many other important specifications are also related to learning content:

- IMS Content Packaging is a specification for sending learning resources (or learning objects) from one program to another, facilitating easier delivery, reuse, and sharing of materials.
- IMS Simple Sequencing is a specification used to describe navigation paths through a collection of learning activities.
- IMS Learning Design is a specification used to describe learning scenarios. It allows these scenarios to be presented to learners online, and enables them to be shared between systems.
- IMS Question and Test Interoperability is designed to make it easier to transfer information such as questions, tests and results between different software applications.

Considering learner data, there are: IMS Learner Information Package (LIP) ("IMS Learner", 2005), IMS Accessibility for LIP (ACCLIP) ("IMS AccessForAll", 2004), IEEE Public and Private Information (PAPI) for Learners (PAPI Learner) ("PAPI Learner", 2002).

IMS LIP is designed to allow information about learners (including their progress and awards received) to be shared between different applications. Within the 11 defined categories, there is one for describing hobbies and recreational activities, and also an accessibility category for describing physical issues (e.g., large print) and/or technical preferences (such as the computer platform).

IMS ACCLIP provides a means to describe how learners can interact with an online learning environment based on their preferences and needs. These preferences will have an influence on the user interface of learning delivery tools and how content is selected and presented.

The PAPI Learner Standard is a data interchange specification that describes learner information for communication among cooperating systems. It is a multi-part standard that specifies the semantics and syntax of learner information. It defines and/or references elements for recording descriptive information about: knowledge acquisition, skills, abilities, personal contact information, learner relationships, security parameters, learner preferences and styles, learner performance, learner-created portfolios, and similar types of information. This standard permits different views of the learner information (perspectives: learner, teacher, parent, school, employer, etc.) and substantially addresses issues of privacy and security.

Finally, it is important to mention two interesting and more recent initiatives:

- The E-Learning Framework (ELF) project (http://www.elframework.org/): ELF is endorsed by important institutions in the e-learning arena such as the Australian Department of Education (DEST), the United Kingdom's Joint Information Systems Committee (JICS), and the U.S. Advanced Learning Initiative (ADL). It aims at providing both a common vocabulary and a roadmap for the development of the component services in an e-learning infrastructure. Besides the identification of necessary services, ELF also lists if there are standards and specifications that already support, at least partially, each of these services, and promotes the development of open source implementation toolkits to assist developers in implementing instances of the services.
- The Content Object Repository Discovery and Registration/Resolution Architecture (CORDRA) project aims to contribute to the solution of the problems related to the interchange of educational content between different repositories. As it is stated in literature, CORDRA is an open, standards-based model for how to design and implement software systems for the purpose of discovery, sharing, and reuse of learning content through the establishment of interoperable federations of learning content repositories (Rehak, Dodds, & Lannom, 2005).

To conclude, the issue should be raised as to whether these standards are perceived as useful for the development of the area. Several authors, in fact, report doubts about this (Dagger, 2006; Marshall, 2004; Modritscher, Gutl, Barrios, & Maurer, 2004; Paramythis & Loidl-Reisinger, 2004), but almost all of them acknowledge at least some advantages in having them. Marshall (2004) states that, "For some aspects of e-learning, it is clear that standards are useful, even necessary, and the impact of their utility is clearly measurable" (p. 601). Also, there are now many systems developed according to the existing standards that reflect the relevance/usefulness of the work done (Duval, 2004).

FUTURE TRENDS

The e-learning industry and research have matured to the point where there is a structured vision of an overall architecture (e.g., ELF) establishing the services that are needed and how they relate to one another. This is a fertile environment for standardization to grow as it offers a roadmap for standards development.

In the future, the issue of the creation of conformance tests to allow the certification of e-learning products must be tackled, as this will bring more credibility to products. Without well-established certification procedures, the claim

for conformance could be regarded as a mere marketing strategy and not necessarily as an indicator of quality.

CONCLUSION

E-learning technology grew largely with no e-learning standards and those that existed were initially developed without being widely adopted. The past few years have seen wider adoption of e-learning standards by the community of e-learning tool developers, especially in corporate training through the adoption of SCORM.

It is possible to notice that an e-learning courseware industry for corporate training (e.g., Skillsoft, NETg, DigitalThink, etc.) has, through the adoption of SCORM, had a rapid development which might otherwise been different if it had not been for this unifying standard which promotes interoperability.

Although important results have been achieved, many products still claiming conformance to standards do not really interoperate without adjustments. So, more specifications to cover all aspects of the identified services are needed.

As e-learning standards bodies are now aware of the importance of collaboration, it is expected that instead of developing overlapping specifications, they will work towards the common aim of having complete, non-redundant, and more reliable standards.

REFERENCES

ASME (n.d.). *Introduction to ASME codes and standards*. Retrieved January 14, 2007, from http://files.asme.org/ASMEORG/Codes/About/Links/1028.pdf

Barritt, C., & Lewis, D. (2001). *Reusable learning object strategy – definition, creation process and guidelines for building – version 3.1*. Cisco Systems, Inc.

Brusilovsky, P., & Peylo, C. (2003) Adaptive and intelligent Web-based educational systems. In P. Brusilovsky, & C. Peylo (Eds.), *International Journal of Artificial Intelligence in Education, 13*(2-4), *Special Issue on Adaptive and Intelligent Web-based Educational Systems,* 159-172.

Dagger, D. (2006). Authoring standards based personalised elearning. In T. Reeves, & S. Yamashita (Eds.), *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2006* (pp. 2680-2685). Chesapeake, VA: AACE

Draft Standard for Learning Object Metadata. (2002). Retrieved January 14, 2007, from http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Duval, E. (2004).Learning technology standardization: Making sense of it all. *Computer Science and Information System,. 1*(1) 33-43.

IEEE. (2005). *Standards development at the IEEE Standards Association*. Retrieved January 14, 2007, from http://standards.ieee.org/announcements/bkgnd stdsprocess.html

IMS AccessForAll Meta-data Overview. (2004). Retrieved January 14, 2007, from http://www.imsglobal.org/accessibility/accmdv1p0/imsaccmd_oviewv1p0.html

IMS Learner Information Package Summary of Changes. (2005). Retrieved January 14, 2007, from http://www.imsglobal.org/profiles/lipv1p0p1/imslip_sumcv1p0p1.html

ISO. (2003). *Proposed draft technical report for: information technology—learning, education, and training—management and delivery—specification and use of extensions and profiles.* Retrieved January 14, 2007, from http://jtc1sc36.org/doc/36N0646.pdf

ISO. (2006). *Overview of the ISO system*. Retrieved January 14, 2007, from http://www.iso.org/iso/en/aboutiso/introduction/index.html#two

Koper, R., & Tattersall, C. (2005). *Learning design: A handbook on modelling and delivering networked education and training*. Springer-Verlag.

Marshall, S. (2004). E-learning standards: Open enablers of learning or compliance strait jackets? In R. Atkinson, C. McBeath, D. Jonas-Dwyer, & R. Phillips (Eds.), *Beyond the comfort zone: Proceedings of the 21st ASCILITE Conference* (pp. 596-605). Perth, Australia.

MASSIE Center. (2003).*Making sense of learning specifications & standards: A decision maker's guide to their adoption* (2nd ed.). S3 Working Group, Masie Centre e-Learning Consortium. Retrieved May 4, 2007, from http://www.masie.com/standards/s3_2nd_edition.pdf

Modritscher, F. (2006).e-Learning theories in practice: A comparison of the three methods. *Journal of Universal Science and Technology of Learning, 0*(0) 3-18.

Modritscher, F., Gutl, C., Barrios, V., & Maurer, H. (2004). Why do standards in the field of e-learning not fully support learner-centred aspects of adaptivity?. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications (EDMEDIA) 2004*, Lugano, Switzerland (pp. 2034-2039). Chesapeake, VA: AACE.

PAPI Learner, Draft 8 Specification. (2002). Retrieved January 14, 2007, from http://edutool.com/papi/

Paramythis, A., & Loidl-Reisinger, S. (2004). Adaptive learning environments and e-learning standards. *Electronic Journal of eLearning, 2*(1), 181-194.

Rehak, D., Dodds, P., & Lannom, L. (2005). A model and infrastructure for federated learning content repositories. Paper presented at the 14th International World Wide Web Conference—WWW 2005, Chiba, Japan.

KEY TERMS

Accessibility: The degree to which a system is usable by as many people as possible.

Conformance: Adherence of an implementation to the requirements of one or more specific specifications or standards.

E-Learning: Refers to the use of technology in learning or training that can be deployed either locally or globally.

Interoperability: The capability to communicate, execute programs, or transfer data among various functional units in a manner that requires the user to have little or no knowledge of the unique characteristics of those units.

Learning Objects: Any digital resource that can be reused to support learning.

Metadata: Structured, encoded data that describe characteristics of an object.

Standard: A standard is a set of technical definitions and guidelines for designers, manufacturers, and users establishing the characteristics of a product, process or service.

User Profile: A set of characteristics describing a user.

Staying Up to Date with Changes in IT

Tanya McGill
Murdoch University, Australia

Michael W. Dixon
Murdoch University, Australia

INTRODUCTION

Information and communications technology (ICT) has been changing rapidly over a long period and this rate of change is likely to continue or increase (Benamati & Lederer, 2001a; Lee & Xia, 2005). This rapid rate of change has produced many opportunities for organizations, but has also brought with it many challenges (Benamati & Lederer, 2001b). Among these challenges is the struggle for organizations to obtain personnel with the appropriate information technology (IT) knowledge and skills in order to meet their ICT needs (Byrd & Turner, 2001; Doke, 1999; Standbridge & Autrey, 2001). This is mirrored by the continual requirement for IT professionals to keep up to date with the skills required by organizations (Benamati et al., 2001a; Klobas & McGill, 1993; Moore, 2000).

Previous research has investigated the importance employers place on various skills and perceived deficiencies in these skills (e.g., Doke, 1999; Leitheiser, 1992; Nelson, 1991; Prabhakar, Litecky, & Arnett, 2005). While the call for improved communication and social skills has been consistent, the technical skills in demand have varied dramatically over time (Prabhakar et al., 2005; Van Slyke, Kittner, & Cheney, 1998). Less has been written about students' perceptions of the importance of various ICT skills, though this was addressed in a study that compared Australian and American students' perceptions of ICT job skills (von Hellens, Van Slyke, & Kittner, 2000). This article provides an overview of a project that investigated the channels of information that ICT students use to keep up to date with employers' needs.

BACKGROUND

Given that the skills required by IT professionals change over time, IT professionals need effective methods to keep up to date. The methods used by IT professionals to keep up to date were studied by Klobas et al. (1993). They identified the existence of a variety of information gathering strategies and noted that while IT professionals tended to be diligent in their efforts to keep up to date, a majority found it difficult to do so. In a more recent study, Benamati et al. (2001a) investigated the coping mechanisms adopted by IT professionals and noted that many mechanisms were not successful.

If it is difficult for experienced IT professionals to keep up to date, it is likely that it is even more difficult for ICT students to do so. New graduates require marketable IT skills in order to gain good employment, but the skills most in demand change regularly. Little is known about how ICT students keep informed of employers' requirements or about how they ensure that they can meet these requirements. Yet this knowledge would be of use to both educational institutions aiming to facilitate this process and to potential employers hoping to recruit students with the required skills.

Information about ICT skill requirements is available from a variety of sources in a variety of formats. Information sources include ICT suppliers, publishing companies, and universities. Formats include different types of publications, presentations, and personal contacts. The term "information channel" can be used to describe the various combinations of sources and formats of information.

HOW DO STUDENTS KEEP UP TO DATE?

Eighty-five information technology students at an Australian university were surveyed to investigate the channels of information that they use to keep up to date with employers' needs. Participants were recruited during class and completed a questionnaire on the spot.

The questionnaire listed information channels that may be used to keep up to date and asked participants firstly whether they had used each channel within the last three months, and also to rate the importance of each channel to them as a means of knowing what skills are in demand. Importance was measured on a 5 point scale ranging from (1) "Not important" to (5) "Vital." The initial list of channels of information was drawn from Klobas et al. (1993) report of the methods used by IT professionals to keep up to date with developments in ICT. Several additional channels were included after consultation with industry contacts. Table 1 lists the information channels included in the questionnaire.

Overall, the students appeared to be diligent in their efforts to keep up to date with employers' skill requirements. The average number of channels used by the students during the previous three months was 3.8 (and the most common number used was 5). Thirteen students (15.3%) had not made any attempt to keep up to date during this period and four (4.7%) had made use of all nine listed channels.

The information channels are ranked by frequency of use in Table 1. The most frequently consulted channels were newspaper employment and IT sections and Internet sources. University instructors had been consulted by about half of the participants during the previous three months. Other students had also been used as a source of information by quite a few students (40%). This high level of use of other students to provide information about employers' skill requirements is understandable given the easy accessibility of other students (Klobas et al., 1993). Work colleagues were ranked 7th overall, but as only around a third of the participants had ICT work experience this means that most of those with prior experience had consulted their colleagues (75% of those with prior ICT work experience had consulted their colleagues). The least used channels were books and vendor presentations. It is likely that students were conscious that information about employer skill requirements derived from books was not going to be sufficiently up to date to meet their needs.

Table 2 shows the importance rankings of the individual information channels. The most highly ranked information channel was Internet sources such as the Cisco and Lucent sites. As well as being frequently used, newspaper ICT sections and employment pages were also considered very important (ranked two and three). University instructors were ranked 4th in importance, which was consistent with their frequency of consultation by students. Although other students were consulted by many students they were not considered as an important channel of information (ranked 7th). This suggests that students recognize that although other students are an easily accessible source of information, they are not necessarily an accurate or reliable source. Both books and vendor presentations were considered of low importance. In future research, it would be interesting to determine how well student perceptions match those of employers.

In addition to the items about methods used to keep up to date, participants were also asked several questions that addressed whether they believed they were in fact obtaining the skills employers required. A majority of participants believed that their degree would provide the skills employers require (67.1% "yes," 5.9% "no," and 27.1% "not sure"). This high level of confidence suggests that although only around 50% of students had consulted their instructors about employer skill requirements during the previous three months (and instructors were only given a medium ranking of importance), students do implicitly accept that instructors know what skills students require. Industry certification was also seen as a very important means to ensure that students obtain the necessary skills (mean importance score was 4.18/5 for those students not yet working in the ICT industry). This is consistent with the results of a recent study on IT certification which found that students undertaking certification believe that the most important benefit of certification is that it provides "real world" experience (McGill & Dixon, 2005).

Are There Demographic Differences in Use and Importance?

Patterns of use and perceptions of importance were further examined to determine whether gender, level of study, or previous ICT work experience had an influence. Differences in use were explored using χ^2 tests and differences in importance were explored using independent sample t-tests. These factors had surprisingly little influence on patterns of use and perceived importance of information channels.

Table 1. Information channels ranked by frequency of use

Rank	Information channel	Number	Percentage
1	Newspaper employment pages	56	65.9
2	Newspaper ICT sections	52	61.2
3	Internet sources (e.g., Cisco, Lucent)	47	55.3
4	University instructors	43	50.6
5	Other students	34	40.0
6	ICT magazines (e.g., Packet Magazine)	29	34.1
7	Work colleagues	24	28.2
8	Books	20	23.5
9	Vendor presentations	17	20.0

Table 2. Information channels ranked by importance

Rank	Information channel	Mean	Standard deviation
1	Internet sources (e.g., Cisco, Lucent)	3.55	1.40
2	Newspaper IT sections	3.38	1.44
3	Newspaper employment pages	3.30	1.30
4	University instructors	2.88	1.42
5	ICT magazines	2.62	1.43
6	Work colleagues	2.54	1.43
7	Other students	2.41	1.13
8	Books	2.24	1.34
9	Vendor presentations	2.13	1.32

The first demographic factor considered was gender. No significant difference was found between the number of information channels used by male and female students. The only significant gender difference was for the levels of use and perceived importance of Internet sources. Male students used Internet sources more frequently and perceived them to be more important for keeping up to date with the skill requirements of employers.

The possible impact of previous ICT work experience was considered next. No significant difference was found between the number of information channels used by those with and those without previous ICT work experience. The only significant difference in usage of information channels was related to consultation with work colleagues and with other students. Those with previous work experience not surprisingly consulted with work colleagues more frequently, and appeared to consider work colleagues a more important channel of information. Presumably those with previous ICT experience would have received better quality information from their work colleagues, than would those without ICT work experience who would have been receiving information from a pool of people with perhaps limited direct ICT experience.

Those without ICT work experience consulted other students more frequently, but there was no difference in perceptions of the importance of other students between those with and those without previous ICT work experience. As previously mentioned, this suggests that other students are consulted because of their accessibility rather than their credibility as a source of information. Those with previous ICT experience have other accessible sources of more credible information and hence do not rely so heavily upon other students.

The differences between undergraduate and postgraduate students were similar to those between those with previous ICT work experience and those without. This is consistent with postgraduate students being more likely to have previous ICT work experience than are undergraduates (54.5% of postgraduates vs. 22% of undergraduates had previous ICT work experience). Undergraduate students consulted other students more frequently, but did not value their information more highly. Postgraduate students also consulted work colleagues more frequently, but they did not value their input more highly. This finding differs from the added importance given to work colleagues by those with previous ICT experience, but the means are in the same direction and the result may reflect the fact that 45.5% of the postgraduates did not have previous ICT work experience.

FUTURE TRENDS

The rapid rate of change in ICT is likely to continue (Benamati et al., 2001b; Lee et al., 2005), and in fact some authors believe that the rate of change is accelerating (Horn, 1999). This means that ICT students will continue to require access to up to date information about employers' ICT skill requirements. Given the increased role of electronic means of information dissemination (Bertot, 2003; Williams & Nicholas, 2001) it is likely that Internet sources of information will continue to be seen as the most important sources and that their frequency of use will increase rapidly so that Internet sources will soon be the most frequently used. Greater broadband access will enable delivery of richer content and greater interactivity. Convergence of information technologies such as notebooks, phones, and television and the development of pervasive computing will provided even greater flexibility to students who wish to keep up to date with employer skill requirements.

CONCLUSION

New graduates require marketable skills in order to gain good employment, but as the ICT industry is subject to rapid change, the skills most in demand change regularly. The study described in this article investigated the approaches that a group of ICT students used to keep up to date with employers' skill needs. Overall, they appeared to be diligent in their efforts to keep up to date with skill requirements. The most commonly used channels were newspaper employment and IT sections, and Internet sources. The same three channels were also rated most highly in terms of importance, with Internet sources being seen as most important.

Instructors were ranked relatively highly in terms of both frequency of consultation and importance and the results suggest an implicit confidence that the knowledge of instructors is up to date. Whilst students have a wide variety of information channels available to them and do indeed make use of them, instructors have a major role to play in providing up to date information about employers' needs. They need to be highly accessible and to ensure that their knowledge of employers' skill requirements remains current. Instructors should use studies of employers' requirements to assess their course offerings and to help guide their students.

REFERENCES

Benamati, J., & Lederer, A. L. (2001a). Coping with rapid changes in IT. *Communications of the ACM, 44*(8), 83-88.

Benamati, J., & Lederer, A. L. (2001b). How IT organanizations handle rapid IT change: 5 coping mechanisms. *Information Technology and Management, 21*(1), 95-112.

Bertot, J. C. (2003). World libraries on the information superhighway: Internet-based library services. *Library Trends, 52*(2), 209-227.

Byrd, T. A., & Turner, D. E. (2001). An exploratory analysis of the value of the skills of IT personnel: Their relationship to IS infrastructure and competitive advantage. *Decision Sciences, 32*(1), 21-54.

Doke, E. R. (1999). Knowledge and skill requirements for information systems professionals: An exploratory study. *Journal of IS Education, 10*(1), 10-18.

Horn, P. M. (1999). Information technology will change everything. *Research Technology Management, 42*(1), 42-47.

Klobas, J. E., & McGill, T. (1993). Computing professionals and information about developments in information technology. *The Australian Computer Journal, 25*(4), 149-158.

Lee, G., & Xia, W. D. (2005). The ability of information systems development project teams to respond to business and technology changes: A study of flexibility measures *European Journal of Information Systems, 14*(1), 75-92.

Leitheiser, R. (1992). MIS skills for the 1990s: A survey of MIS managers perceptions. *Journal of Management Information Systems, 9*(1), 69-91.

McGill, T., & Dixon, M. (2005). Information technology certification: A student perspective. *International Journal of Information and Communications Technology Education, 1*(1), 19-30.

Moore, J. E. (2000). One road to turnover: An examination of work exhaustion in technology professionals. *MIS Quarterly, 24*(1), 141-148.

Nelson, R. R. (1991). Educational needs as perceived by IS and end-user personnel: A survey of knowledge and skill requirements. *MIS Quarterly, 15*(4), 503-525.

Prabhakar, B., Litecky, C. R., & Arnett, K. (2005). IT skills in a tough job market. *Communications of the ACM, 48*(10), 91-94.

Standbridge, J., & Autrey, R. (2001). Rapid skill obsolescence in an IT company: A case study of Acxiom Corporation. *Journal of Organizational Excellence, 20*(3), 3-9.

Van Slyke, C., Kittner, M., & Cheney, P. (1998). Skill requirements for entry-level IS graduates: A report from industry. *Journal of Information Systems Education, 3*(9), 7-11.

von Hellens, L., Van Slyke, C., & Kittner, M. (2000). A comparison of Australian and American students' perceptions of IT job skills. In M. Khosrow-Pour (Ed.), *Challenges of information technology management in the 21st Century. 2000 IRMA International Conference* (pp. 915-916). Hershey, PA: Idea Group Publishing.

Williams, P., & Nicholas, D. (2001). *The Internet and the changing information environment*. London: Aslib-IMI.

KEY TERMS

Industry Certification: Certification involves passing a recognized standardized test (or set of tests) within particular subject areas. It intends to establish a standard of competency in defined areas. ICT industry certifications are designed to provide targeted skills that have immediate applicability in the workplace.

Information Channel: A term used to describe the various combinations of sources and formats of information.

Information Format: The arrangement and appearance of information. Format includes both the media used and the style of presentation.

Information Gathering Strategies: The approaches and processes used by information seekers. Information seeking behavior is influenced by previous experience, mental models, and preferences of information seekers.

Information Source: An organization or person from which information is obtained.

Information Technology Professionals: A term used to describe people for whom development and support of IT systems and related activities is their primary employment. The group includes people who design hardware, who develop and support information systems and who train end users. It does not include people who use ICT in the course of pursuing other professions.

Information Technology Skills: All IT professionals require some computer skills; these may include particular programming languages, database, or networking skills.

Category: Business Information Systems

Strategic Alignment Between Business and Information Technology

Fernando José Barbin Laurindo
University of São Paolo, Brazil

Marly Monteiro de Carvalho
University of São Paolo, Brazil

Tamio Shimizu
University of São Paolo, Brazil

INTRODUCTION

Information technology (IT) has assumed an important position in the strategic function of the leading companies in the competitive markets (Porter, 2001). Particularly, e-commerce and e-business have been highlighted among IT applications (Porter, 2001). Two basic points of view can be used for understanding IT's role: the acquisition of a competitive advantage at the value chain, and the creation and enhancement of core competencies (Porter & Millar, 1985; Duhan, Levy, & Powell, 2001).

Several problems have been discussed concerned with IT project results in effectiveness of their management. Effectiveness, in the context of this article, is the measurement of the capacity of the outputs of an information system or of an IT application to fulfill the requirements of the company and to achieve its goals, making this company more competitive (Shimizu, Carvalho, & Laurindo, 2006).

There is a general consensus about the difficulty of finding evidence of returns over the investments in IT (the "productivity paradox"), even though this problem can be satisfactorily explained (Farrell, 2003). Carr (2005) defends the idea that IT in itself has no more strategic value, since it is so widely disseminated that it could not be a source of strategic differentiation anymore.

In order to better use these investments, organizations should evaluate IT effectiveness, which allows the strategic alignment of objectives of implemented IT applications and their results with the company business vision (Shpilberg, Berez, Puryear, & Shah, 2007; Laurindo & Moraes, 2006). Besides, it must be highlighted that if IT applications are associated with changes in business processes, it is possible to notice greater impacts in business performance (Farrell, 2003).

According to Benko and McFarlan (2003), three aspects must be taken into account about IT strategic alignment: IT projects portfolio, business objectives, and the constantly changing situation of business environment.

Thus, the comparison and evaluation of business and IT strategies and between business and IT structures must be a continuous process, since the company situation is constantly changing to meet market realities and dynamics.

THEORETICAL BACKGROUND

Finding Strategic IT Applications

The discussion about the strategic impact of IT applications started in the 1970s, when technology began to provide more powerful alternatives not only for solving companies' problems but also for increasing their business competitiveness (Shimizu et al., 2006).

One of the first important proposals for studying the strategic role of IT was that of *critical success factors* (CSFs), which is still a widespread method used for linking IT applications to business goals, and for planning and prioritizing information systems projects. This method was proposed by Rockart (1979) and states that the information systems, especially executive and management information systems, are based on the current needs of the top executives. These information needs should focus on the CSFs.

Rockart defines CSFs as the areas where satisfactory results "ensure successful competitive performance for the organization." This author states that CSFs' prime sources are the structure of the industry, business (or competitive) strategy, industry position, geographic location, environment, and temporal factors.

Basically, the CSF method includes the analysis of the structure of the particular industry and the business strategy, and the goals of the organization and its competitors. This analysis is followed by two or three sessions of interviews with the executives, in order to identify the critical success factors related to business goals, define respective measures (quantitative or qualitative) for the CSFs, and define infor-

mation systems for controlling CSFs and their measures (Shimizu et al., 2006).

For Rockart, this process can be useful at each level of the company and should be repeated periodically, since CSFs can change through the time and also can differ from one individual executive to another.

The CSF method had an important impact on managerial and strategic planning practices, even though it was primarily conceived for information systems design, especially management and executive information systems.

Besides the utilization in information systems planning and information systems project management, it has been used in strategic planning and strategy implementation, management of change, and as a competitive analysis technique.

Furthermore, the continuous measurement of CSFs allows companies to identify strengths and weaknesses in their core areas, processes, and functions (Rockart, 1979).

More details of the process of implementation of the CSF method can be found in Rockart and Crescenzi (1984).

Understanding IT Strategic Role in Companies

McFarlan (1984) proposed the Strategic Grid that analyzes the impacts of IT-existent applications (present) and of an applications portfolio (future), defining four boxes, each one representing one possible role for IT in the enterprise: "Support," "Factory," "Turnaround," and "Strategic" (see Figure 1).

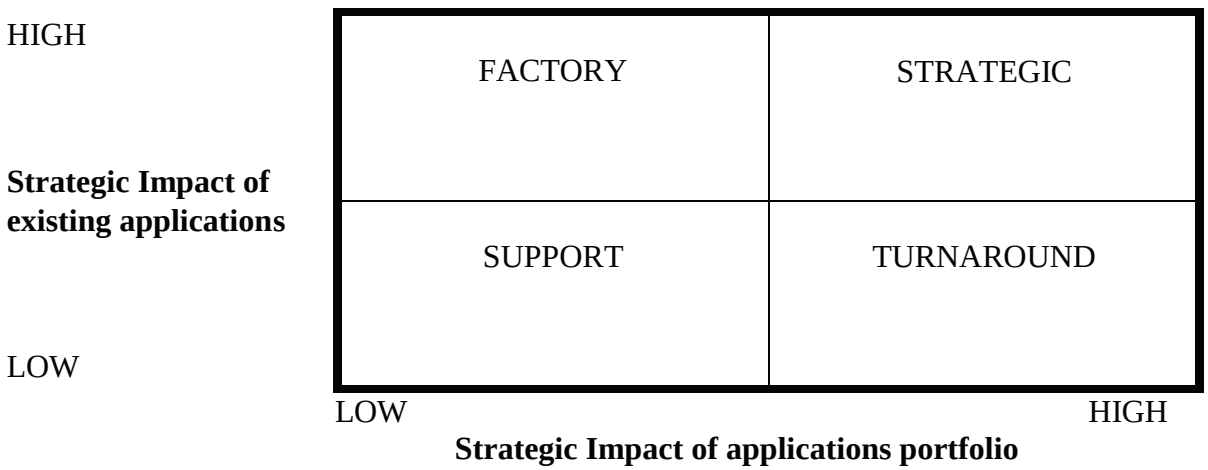
- *Support:* IT has little influence in present and future company strategies.
- *Factory:* Existent IT applications are important for the company's operations success, but there is no new strategic IT application planned for the future.
- *Turnaround:* IT is changing from one situation of little importance ("support" box) to a more important situation in business strategy.
- *Strategic:* IT is very important in business strategy in the present, and new planned applications will maintain this strategic importance of IT in the future.

In order to assess the strategic impact of IT, McFarlan proposed the analysis of five basic questions about IT applications, related to the competitive forces (Porter, 2008):

Can IT applications:

- build barriers to the entry of new competitors in the industry?
- build switching costs for suppliers?

Figure 1. Strategic grid of impacts of IT applications (McFarlan, 1984)



- change the basis of competition?
- change the balance of power in supplier relationships?
- create new products?

These questions should be answered considering both present and planned future situations.

Thus, IT may present a smaller or greater importance, according to the kind of company and industry operations. In a traditional manufacturing company, IT supports the operations, since the enterprise would keep on operating even when it could not count on its information systems. However, IT is strategic in a bank for business operations, since it is a source of competitive advantage and a bank cannot operate without its computerized IS.

Nolan and McFarlan (2005) have updated the Strategic Grid, changing the two "axes" for "Need for Reliable IT" (instead of "Present Impact") and "Need for New IT" (instead of "Future Impact"). These authors stated that companies in "Support" and "Factory" quadrants adopt a *defensive* approach regarding IT. On the other hand, companies classified in "Turnaround" and "Strategic" quadrants can be considered *offensive* in IT use. They also indicated the right policies in IT governance (Weil & Ross, 2005) for the board of directors' use in each of the four situations of the Strategic Grid.

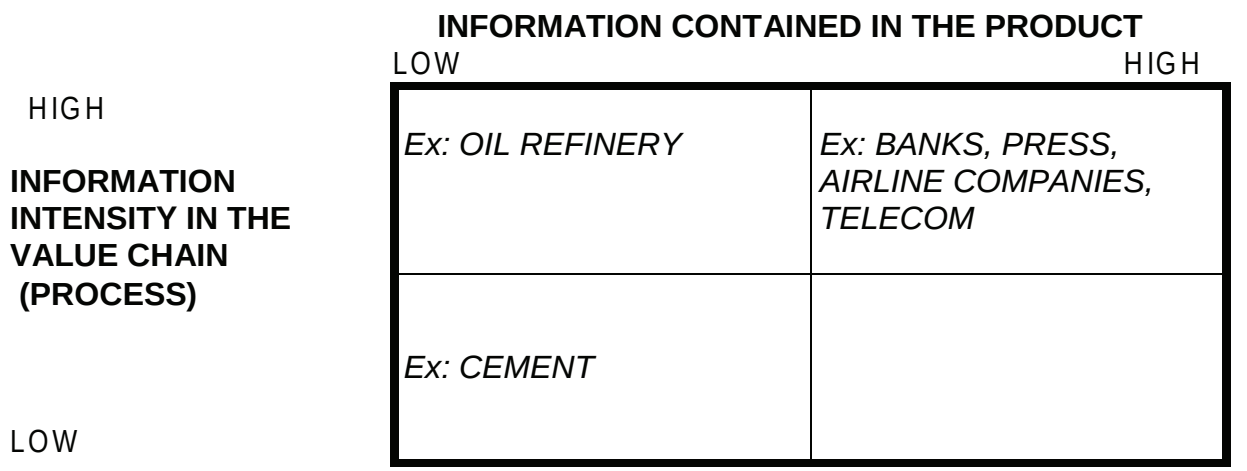
Porter and Miller (1985) highlight the concepts of the value chain (activities inside the company linked by connections and which have one physical component and another of information processing) and value systems (the set of value chains of an industry from the suppliers to the final consumer).

IT permeates the chains of value, changing the way of executing activities of value and also the nature of the connections among them and, therefore, IT can affect competition:

- by changing the structure of the sector since it has the ability to influence each of the five forces of competition (Porter, 2008);
- by creating new competitive advantages, reducing costs, increasing differentiation, and altering the scope of competition scope; and
- by generating completely new business.

The potential that IT has to make these changes varies according to the characteristics of the process (value chain) and of the product, regarding information needs. The "Information Intensity Matrix" considers the value chain and analyzes "how much" information is contained in the process and the product (see Figure 2). In companies whose products and processes contain a lot of information, information technology will be very important (Porter & Miller, 1985).

Figure 2. Information intensity matrix (adapted from Porter & Millar, 1985)



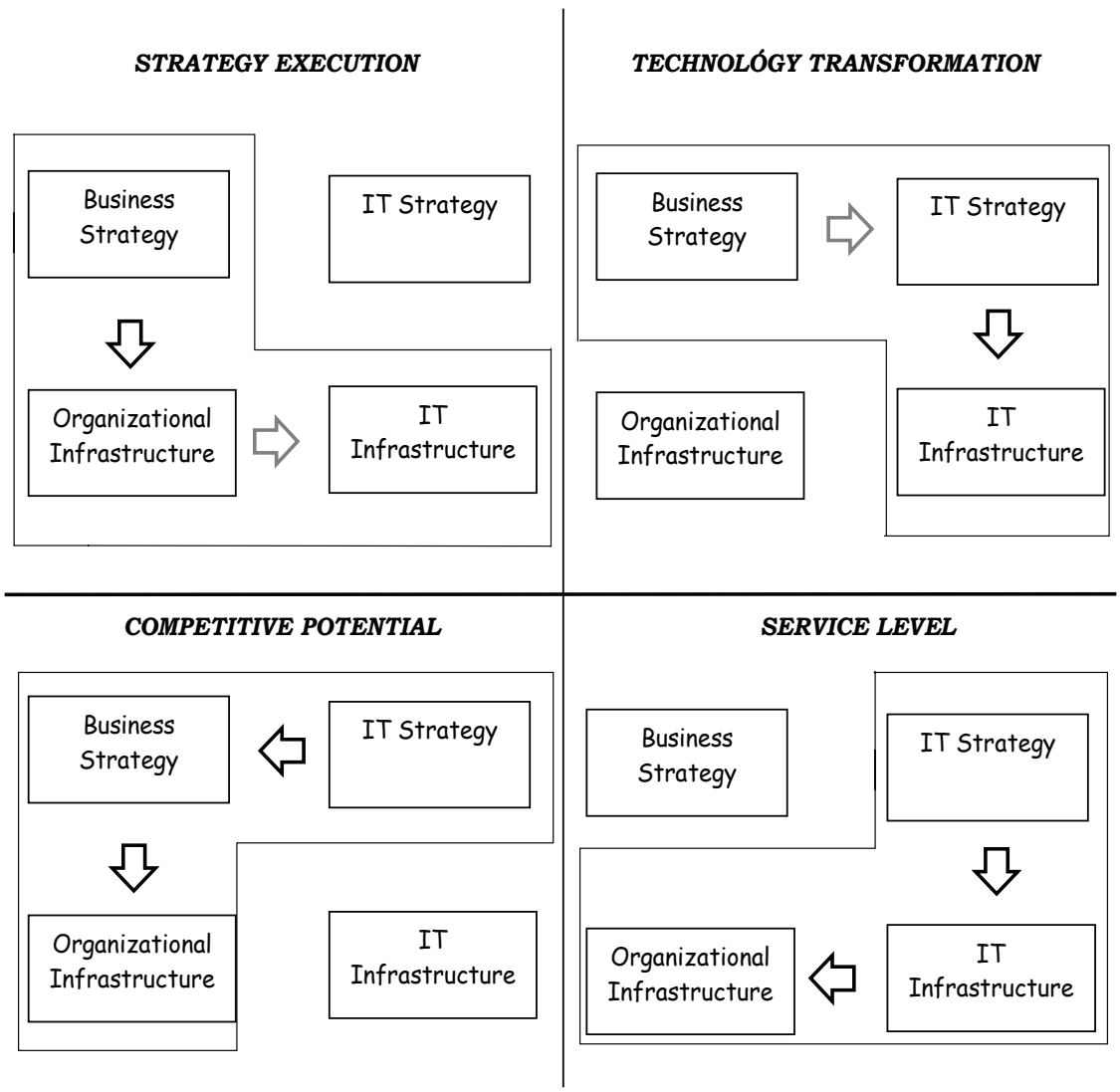
In their original article, Porter and Millar did not cite an example for "high information content in the product" or "low information intensity in the process" in the Information Intensity Matrix. However, for Duhan et al. (2001), this would be the case of educational and law firms, for consulting firms would also fit in this same quadrant.

Further according Duhan et al. (2001), an analysis of the value chain would be impaired in the case of knowledge-based companies (such as consulting firms) where it is hard to identify the value that is aggregated to each activity. In these situations, the authors propose that using the essential competencies would be more appropriate to plan the strategic use of information systems.

Henderson and Venkatraman (1993) proposed the "Strategic Alignment Model" that analyzes and emphasizes the strategic importance of IT in the enterprises. This model is based on both internal (company) and external (market) factors.

The authors emphasize that strategy should consider both internal and external domains of the company. Internal domain concerns administrative structure of the company; external domain concerns the market and the respective decisions of the company. Thus, according to this model, four factors (that the authors called domains) should be considered for planning IT:

Figure 3. Perspectives of strategic alignment (adapted from Henderson & Venkatraman, 1993)



1. business strategy,
2. IT strategy,
3. organizational infrastructure and processes, and
4. IS infrastructure and processes.

The Strategic Alignment Model brings the premise that the effective management of IT demands a balance among the decisions about those four domains above.

According to Henderson and Venkatraman, there are four main perspectives of Strategic Alignment, through the combination of the four factors, starting from business strategy or from IT strategy, as shown in Figure 3.

One important innovation of this model is that IT strategy could come first and change business strategy, instead of the usually general belief that business strategy comes before IT planning. This planning should be a continuous process, since external factors are in a permanent changing situation. If the company does not follow these changes, it will be in serious disadvantage in the fiercely competitive market. This is particularly true when a new technology is adopted by almost all companies in an industry, passing from a competitive advantage for those that have it to a disadvantage to those that do not use it. Thus, in this sense, the strategic alignment differs from the classic vision of the strategic plan, which does not present the same dynamic approach.

After the proposal of the four perspectives above, Luftman (1996) described four new perspectives that start in the infrastructure domains, instead of the strategies domains:

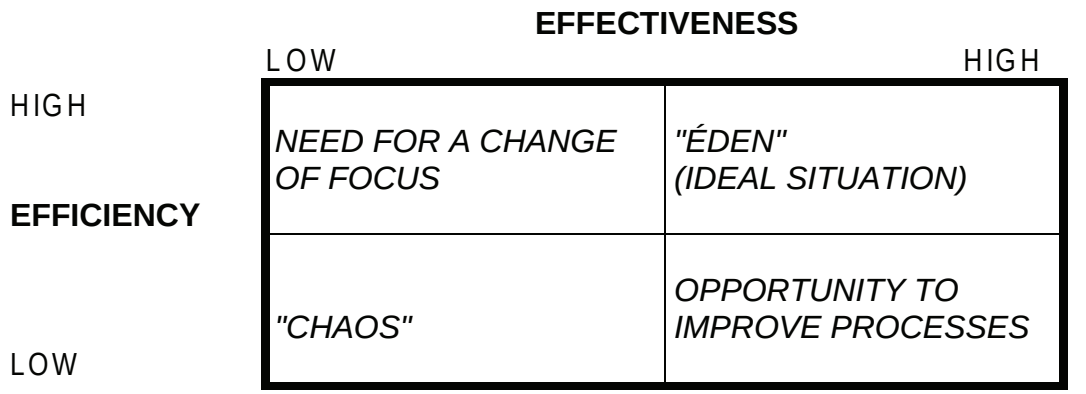
- *Organizational IT Infrastructure Perspective:* Organizational infrastructure → IT infrastructure → IT strategy
- *IT infrastructure Perspective:* IT infrastructure → IT strategy → Business strategy
- *IT Organizational Infrastructure Perspective:* IT infrastructure → Organizational infrastructure → Business strategy
- *Organizational Infrastructure Perspective:* Organizational infrastructure → Business strategy → IT strategy

Luftman (1996) also proposed that in some situations a fusion of two perspectives might occur. In these cases, two perspectives can be simultaneously assessed and impact the same domain: *IT Infrastructure Fusion, Organizational Infrastructure Fusion, Business Strategy Fusion, IT Strategy Fusion.*

Research has been developed in order to find the enablers of Strategic Alignment. Luftman (2001) listed five of them: senior executive support for IT; IT involved in strategy development; IT understands the business, business-IT partnership; well-prioritized IT projects; and IT demonstrates leadership. The absence or poor performance of these same factors are considered inhibitors of Strategic Alignment.

Some authors, like Ciborra (2004), state the strategic success of IT applications might be achieved through a tentative approach, rather than structured methods of strategic IT

Figure 4. Efficiency vs. effectiveness in IT applications (adapted from Shimizu et al., 2006)



planning. These authors argue that frequently the drivers of strategic IT applications are efficiency issues, instead of a result of a strategic IT plan. Some important and well-known successful information systems, with clear strategic impacts, do not present evidence of being previously planned, which seems to be in agreement with this kind of thinking (Eardley, Lewis, Avison, & Powell, 1996).

EFFICIENCY AND EFFECTIVENESS: DIAGNOSING THE ROLE OF IT IN COMPANIES

In this article the importance of focusing on the effectiveness of IT utilization has been emphasized, since frequently analysis is done only from the point of view of efficiency. However, this does not mean that being efficient is not positive; it means that one needs to be efficient in certain areas. In other words, once effectiveness is achieved, increased efficiency can result in important gains and there are many models that help to analyze and improve IT efficiency.

Figure 4 contains a proposed diagram for viewing the situations related to efficiency and effectiveness in the use of IT.

When companies demonstrate low efficiency and high effectiveness, they are in "Chaos"—in a critical situation. The first move to get out of this situation should be to aim at increased effectiveness, to align the IT strategy with the business strategy. If the company has low effectiveness, but high efficiency in the use of IT, it means that it should redirect its efforts, change the focus of its activities, in order to use its good capacity where it can add value to the company's competitiveness. In the case of a company with high effectiveness, but low efficiency in IT utilization, it is necessary to work to improve its processes, with a view to exploiting to the maximum the focus that is already on the right things, and which can contribute to the success of the company's strategy. Finally, a company that is efficient and effective in the use of IT will arrive in "Eden," the ideal situation, which should be the goal for all.

CONCLUSION

At present there are a series of applications that have captivated the attention of many and have opened up new possibilities. Both Knowledge Management and Customer Relationship Management have been closely associated to IT. In fact, without IT these concepts could hardly have been effectively used in companies. In this sense, one important example is the growing use of business intelligence applications.

Despite the failure of many virtual enterprises (the so-called "dot.coms"), e-business and e-commerce applications seems to have reached a new maturity level, especially B2B (business-to-business—the connection between companies via the Internet).

There are various success stories, and large companies are increasingly investing in this success. According to Porter (2001), the Internet is the IT tool that, up to the present, has shown the greatest potential of being a source of obtaining or stressing strategic advantages. Therefore, an appropriate analysis and evaluation of IT effectiveness can take on a fundamental role, enabling it to really become a powerful tool for competitiveness.

The concepts described above show the importance of a broad view for analyzing IT strategic alignment. Each of the described models (CSF, Strategic Grid, Information Intensity Matrix, and Strategic Alignment) focuses on specific aspect of this issue.

These widespread known models, in fact, have complementary characteristics, and concomitant use of them allows a better comprehension of the role of IT in an organization.

On the other hand, even the use of the three models does not solve the complexity of IT alignment in organizations. As highlighted by several authors, sometimes a tentative and evolutionary approach can be successfully adopted, in circumstances that structured methods do not work properly. By this continuous focus in the IT strategic alignment, the problems of the "productivity paradox" would be overcome.

Further studies would be necessary for a better and deeper understanding of the importance of IT effectiveness for the success of competitive companies. However, this chapter intended to help find a way for this understanding and to provide some tools.

REFERENCES

Benko, C., & McFarlan, F.W. (2003). *Connecting the dots.* Boston: Harvard Business School Press.

Carr, N.G. (2005). The end of corporate computing. *Sloan Management Review, 46*(3), 67-73.

Ciborra, C.U. (2004). *The labyrinths of information: Challenging the wisdom of systems.* Oxford: Oxford University Press.

Duhan, S., Levy, M., & Powell, P. (2001). Information systems strategies in knowledge-based SMEs: The role of core competencies. *European Journal of Information Systems, 10*(1), 25-40.

Eardley, A., Lewis, T., Avison, D., & Powell, P. (1996). The linkage between IT and business competitive systems: A reappraisal of some 'classic' cases using a competitive

analysis framework. *International Journal of Technology Management, 11*(3/4), 395-411.

Farrell, D. (2003). The real new economy. *Harvard Business Review, 81*(10), 104-112.

Henderson, J.C., & Venkatraman, N. (1993). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal, 32*(1), 4-16.

Laurindo, F.J.B., & Moraes, R.O. (2006). IT projects portfolio management: A Brazilian case study. *International Journal of Management and Decision Making, 7*(6), 586-603.

Luftman, J.N. (1996). Applying the Strategic Alignment Model. In J.N Luftman (Ed.), *Competing in the information age□strategic alignment in practice* (pp. 43-69). New York: Oxford University Press.

Luftman, J.N. (2001). Business-IT alignment maturity. In R. Papp (Ed.), *Strategic information technology: Opportunities for competitive advantage* (pp. 105-134). Hershey, PA: Idea Group.

McFarlan, W.E. (1984). Information technology changes the way you compete. *Harvard Business Review, 62*(3), 98-103.

Nolan, R.L., & McFarlan, W.E. (2005). Information technology and the board of directors. *Harvard Business Review, 83*(10), 96-106.

Porter, M.E., & Millar, V. (1985). How information gives you competitive advantage. *Harvard Business Review, 63*(4), 149-160.

Porter, M.E. (2008). The five competitive forces that shape strategy. *Harvard Business Review,* (1), 78-93.

Porter, M.E. (2001). Strategy and the Internet. *Harvard Business Review,* (March), 63-78.

Rockart, J., & Crescenzi, A.D. (1984). Engaging top management in information technology. *Sloan Management Review, 25*(4), 3-16.

Rockart, J.F. (1979). Chief executives define their own data needs. *Harvard Business Review, 57*(2), 81-92.

Shimizu, T., Carvalho, M.M., & Laurindo, F.J.B. (2006). *Strategic alignment process and decision support systems: Theory and case studies.* Hershey, PA: IRM Press.

Shpilberg, D., Berez, S., Puryear, R., & Shah, S. (2007). Avoiding the alignment trap in information technology. *MIT Sloan Management Review, 49*(1).

Weil, P., & Ross, J.W. (2005). A matrixed approach to IT governance. *MIT Sloan Management Review, 46*(2), 26-34.

KEY TERMS

Competitive Forces: According to Porter (2008), the state of the competition in a particular industry depends on five basic forces: new competitors, bargaining power of suppliers, bargaining power of customers, rivalry among current competitors, and substitute products or services.

Critical Success Factor (CSF): One of the areas where satisfactory results "ensure successful competitive performance for the organization," according to Rockart (1979).

Effectiveness: In the context of IT, the measurement of the capacity of the outputs of an information system or of an IT application to fulfill the requirements of the company and to achieve its goals, making this company more competitive. In other words, effectiveness can be understood as the ability of "do the right thing."

Productivity Paradox: The discussion about the lack of evidence about the return of investments on IT in the economy productivity indicators.

Strategic Alignment: The IT Strategic Alignment Model was proposed by Henderson and Venkatraman (1993) and consists of a framework for studying IT impacts on business and understanding how these impacts influence IT organization and strategy, as well as how it enables analysis of the market availabilities of new information technologies.

Strategic Grid: Nolan and McFarlan (2005) and McFarlan (1984) proposed the Strategic Grid, which allows the visualization of the relationship between IT strategy and business strategy and operations. This model analyzes the impacts of IT-existent applications (present) and of an applications portfolio (future), defining four boxes, each one representing one possible role for IT in the enterprise: "Support," "Factory," "Turnaround," and "Strategic."

Value Chain: According to Porter and Millar (1985), the set of technologically and economically distinct activities a company performs in order to do business.

Strategic IT Investment Decisions

Tzu-Chuan Chou
University of Bath, UK

Robert G. Dyson
University of Bath, UK

Philip L. Powell
University of Bath, UK
University of Groningen, UK

INTRODUCTION

As many as half the decisions taken in organizations result in failure (Nutt, 1999). As information technology (IT) assumes a greater prominence in firms' strategic portfolios, managers need to pay more attention to managing the technology. However, while IT can have a significant impact on organizational performance, it can also be a major inhibitor of change and can be a resource-hungry investment that often disappoints. Organizations can best influence the success of IT projects at the decision stage by rejecting poor ones and accepting beneficial ones. This may enable better implementation, as Nutt (1999) suggests most decision failures are due to implementation failure that tends to be under managers' control.

However, little is known about IT decision processes. Research demonstrates the importance of managing strategic IT investment decisions (SITIDs) effectively. SITIDs form part of the wider range of corporate strategic investment decisions (SIDs) that cover all aspects in which the organization might wish to invest. Strategic investment decisions will have different degrees of IT intensity that may impact outcome. IT investment intensity is the degree to which IT is present in an investment decision. That is, some decisions will be wholly about IT investments while others will have little or no IT—most, though, will be blended programs of IT and non-IT elements. Here, IT investment intensity is defined as the ratio of IT spending to total investment. The higher the IT investment intensity, the more important IT is to the whole investment. For example, Chou, Dyson, and Powell (1997) find IT investment intensity to be negatively associated with SID effectiveness. The concept of IT intensity is similar to, but also somewhat different from, the concept of information intensity. Information intensity is the degree to which information is present in the product or service (Porter & Millar, 1985).

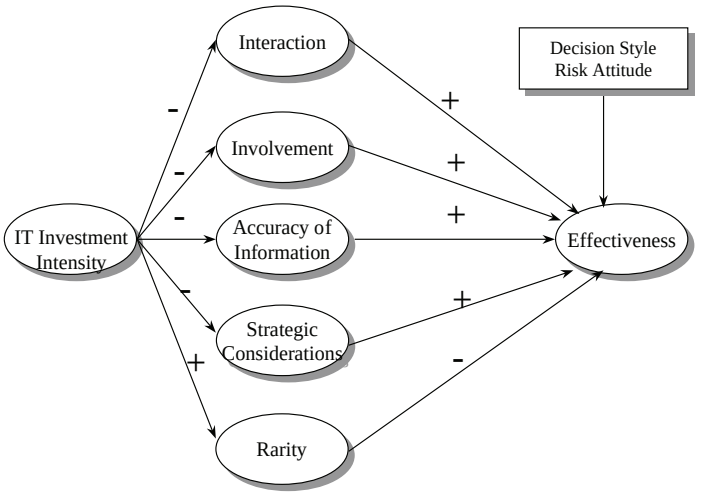
Management may use different processes in order to make different types of decisions (Dean & Sharfman, 1996). The link between decision process and outcome is so intimate that "the process is itself an outcome" (Mohr, 1982, p. 34). This may imply that the link between IT investment intensity and SID effectiveness is not direct but that the impact of IT investment intensity may be through the decision process. If different IT intensity in projects leads to different decision processes, leading to different outcomes, then it is important to know what factors act in this, in evaluating and managing SITIDs. This chapter presents an integrative framework for exploring the IT investment intensity-SID effectiveness relationship.

BACKGROUND

Studying decisions involves "contextualism" (Pettigrew, McKee, & Ferlie, 1988), which integrates process, content, and context, as all decisions need to be studied in context. Content refers to the decision itself, here exploring the nature and scope of SIDs. Process refers to actions, reactions, and interactions as managers allocate resources for the decision. The context includes the outer context of economic, political, and social actions, while the inner context involves ongoing strategy, structure, culture, management, and political processes.

Many researchers have investigated how strategic decisions are made, though most focus on the decision process rather than the implementation and the outcome. However, Hickson, Miller, and Wilson (2003) identify eight independent variables in decision implementation—familiarity, assessibility, specificity, resourcing, acceptability, structural facilitation, and priority—and uncover two distinct approaches—experience-based and readiness-based. The experience-based approach leads to acceptance of what is being done, while the readiness-based approach leads to implementation being given clear priority. Both approaches may be employed together. Hsu (2001) on the other hand introduces promethean rationality—the stealing back of order amid disorder. Sauer-Leroy (2004) argues that decision reality is much more complex than can ever be captured by

Figure 1. The theoretical model (adapted from Chou et al., 1997)



financial analysis and that projects often introduce inertia to the organization, as they are often irreversible. He stresses the role of subjective factors in strategic decisions.

In the IT investment intensity-SID effectiveness link, the roles of process, content, and context are unclear. Though unclear, it is likely that the links between variables are not direct; rather they are mediated or moderated by other variables or processes. Moderators and mediators are functions of third variables. A moderator "partitions a focal independent variable into subgroups that establish its domains of maximal effectiveness in regard to given dependent variables," while a mediator function "represents the generative mechanism through which the focal independent variable is able to influence the dependent variable of interest" (Baron & Kenny, 1986, p. 1174). Sambamurthy, Bharadwaj, and Grover (2003) employ the moderator concept in their work on reshaping agility through digital options. They, for example, see IT competence as an antecedent of firms' competitive actions, but the relationship is mediated by dynamic capabilities. Sambamurthy et al. (2003) demonstrate theoretically the IT investments and capabilities influence firm performance through organizational capabilities—agility, digital options, and entrepreneurial alertness, and through strategic processes involving capability building, entrepreneurial action, and co-evolutionary adaptation. This, they claim, is valid at the enterprise level, for business units and for processes. They finally call for empirical research that might validate their theoretical developments. This article takes a slightly different route as it focuses on IT at the project or at the decision level, and it is based on data that back up the model development.

Here, the proposal is that the impact of IT investment intensity on SID effectiveness is through decision processes. Accordingly, decision process constructs should have a mediating effect. Greater IT intensity in projects leads, inter alia, to a more technically orientated project that impacts SID effectiveness. The decision content has a mediating effect on the IT involvement-SID effectiveness link. The investment context impacts the outcome. Therefore, context constructs should act as covariances that impact SID effectiveness. Decision context, content, and process will involve many individual constructs, some unrelated to IT investment intensity. Two criteria can be employed in order to select the constructs of interest here. First, the decision construct is expected to vary with IT investment intensity. Second, it must impact at the decision level. Figure 1 outlines the basic model.

A hypothesized negative impact of IT investment intensity on several constructs suggests projects with high IT investment intensity are more challenging than those with low IT content. Effectiveness compares actual performance against planned target, outcomes, and policy objectives, measured by project success, correct choice, unexpected negative outcomes, learning, and satisfactory process (Butler, Davies, Pike, & Sharp, 1991).

Decision Context

The context of any investment is affected by many things, such as the firm's financial health, its market position, industry pressures, culture, and business strategy. SIDs often involve major change to the organization and environ-

ment. This suits managers with an innovative risk attitude. From a style perspective, decision quality is dependent on resources the leader is able to utilize. Consensus-driven management seems able to acquire more information than directive management, and leads to more effective decisions. Management's *attitude to risk* and *decision style* are predicted to relate to SID effectiveness, since other factors impact at an organizational level.

Decision Process

Strategic decision processes involve comprehensiveness, rational activity, participation, duration, and conflicts (Rajagopalan, Rasheed, & Datta, 1993). Comprehensiveness measures rationality and is the extent to which the organization attempts to be exhaustive in making and integrating strategic decisions. This includes formal meetings, assignment of primary responsibility, information-seeking and analytical activities, the systematic use of external sources, stakeholder involvement, use of consultants, reviews of historical data, functional expertise (Papadakis, 1995), and informal interaction. Hickson, Butler, Cray, Mallory, and Wilson (1986) define "politicality" as the degree to which influence is exerted on the outcome through a decision process. Strategic decision-making is not a matter of explicating alternatives and choosing on the basis of readily available criteria all participants perceive as appropriate (Fahey, 1981). It might be expected that interaction and involvement are related to IT investment intensity.

Interactions are contacts between people. Higher IT intensity reduces interaction and SID effectiveness. Decision-makers' IT knowledge, experience, and education are associated with alienated attitudes toward IT. Higher IT investment intensity leads to more technically oriented projects. Without IT knowledge, managers cannot discuss the project knowledgeably. It, therefore, reduces interaction and impacts decision quality. This article suggests that higher IT intensity reduces *involvement,* reducing SID effectiveness. Less involvement leads to less collective information and reduce decision effectiveness. This suggests that *IT investment intensity reduces interaction and adversely impacts decision effectiveness* and that *IT investment intensity reduces involvement and adversely impacts decision effectiveness.*

The evaluation process is important for investment decisions. An IT investment decision is problematic because the cost and benefits are hard to identify and quantify. Therefore, uncertainty of information used in evaluating IT investment is greater. The higher the information uncertainty, the lower the *information accuracy*. This article expects that lower accuracy of information reduces decision effectiveness, that is, *IT investment intensity reduces information accuracy and adversely impacts decision effectiveness.*

It can be argued that the IT evaluation problem is one of alignment. This article expects that management may fail to link IT's strategic purpose with organizational strategy, reducing decision effectiveness. This suggests that *IT investment intensity reduces strategic considerations and adversely impact decision effectiveness.*

Decision Content

A strategic decision is characterized by its novelty and complexity. Complexity relates to the number and variety of factors in the environment that impinge on decision-making behavior. SIDs evolve from the organizational context and have their own characteristics. Constructs that contribute to decision complexity include rarity and importance. Uncertainty is due to *rarity*. Rarity assesses the novelty of the decision to the firm. If a firm repeatedly makes similar decisions, then it will gain experience; conversely, a rare decision is likely to be more problematic. Importance is common to all SIDs irrespective of IT investment intensity, as they are all strategic. New technologies often require investments of a different nature because of high uncertainty, widespread organizational impact, and greater strategic importance. Even compared with other new technologies, the IT life cycle is short so that the IT component of projects is constantly changing, increasing rarity. Rarity inhibits effective feedback and learning. This article expects that the higher IT investment intensity, the higher decision rarity, reducing decision effectiveness; *IT investment intensity heightens decision rarity and adversely impacts decision effectiveness.*

In order to investigate these proposals, empirical work investigated a single strategic investment project in each if 80 Taiwanese manufacturers. IT investment intensity is measured by the ratio of IT spending to total investment in the project, while the measure of decision effectiveness is subjective. In order to capture the whole process from decision to outcome, only projects that were complete were researched. Thus each project had a decision process and an outcome to allow assessment of success.

DISCUSSION

A principal components factor analysis highlights five important factors in these decisions—information accuracy, strategic consideration, interaction, involvement, and rarity. However, only three proposed mediators, interaction, information accuracy, and strategic consideration are significant in predicting mediators.

The model as a whole is significant in predicting SID effectiveness. When contextual variables are added, IT investment intensity is still significant in predicting SID effectiveness. Interaction, information accuracy, and strategic consideration have a negative correlation with IT investment intensity, but a positive correlation with SID effectiveness. Hence, the impact of IT investment

intensity is transmitted to interaction, information accuracy, and strategic considerations and, through that, adversely impacts decision effectiveness. This suggests that *IT investment intensity does reduce interaction and adversely impacts decision effectiveness.* It is also the case that *IT investment intensity reduces information accuracy and adversely impacts decision effectiveness,* and that *IT investment intensity reduces strategic considerations and adversely impacts decision effectiveness*. However *IT investment intensity is not found to reduce involvement and adversely impacts decision effectiveness*, nor does *IT investment intensity heighten decision rarity and adversely impacts decision effectiveness*.

Three process-related constructs, interaction, strategic considerations, and information accuracy act as mediating factors in linking IT investment intensity and SID effectiveness, as they reduce the effects of IT investment intensity. Content-related constructs do not act as mediators. Although decision rarity is negatively associated with SID effectiveness, it is unrelated to IT intensity.

Interaction in the formulating process has a mediating effect on the linkage. Interaction is important in developing group behavior. IT investment intensity lowers interaction, reducing SID effectiveness.

Strategic considerations act as a mediating variable. The higher the IT intensity, the lower the strategic considerations, leading to reduced SID effectiveness. This demonstrates that the IT evaluation problem is really one of alignment.

Information accuracy acts as a mediating variable. The higher the IT investment intensity, the lower the information accuracy, reducing SID effectiveness. This supports Freeman and Hobbs (1991), who find managers ignoring reject signals given by capital budgeting techniques, and identify senior management's preference for qualitative information and IT investment as an "act of faith" (Powell, 1993). This suggests that high information uncertainty leads to a limited use of capital budgeting techniques.

IT investment intensity is still significant when interaction is tested as a mediator. This indirect transmission of influence from IT investment intensity to SID effectiveness via interaction shows that the effect of IT investment intensity on effectiveness is only partially mediated by interaction. The effect of IT investment intensity on SID effectiveness is completely mediated by strategic consideration and information accuracy—two evaluation-related constructs. This implies that, in seeking a better outcome of SITIDs, research that focuses on evaluation factors may be insufficient to capture the complexity of SITIDs.

FUTURE TRENDS

As demonstrated, the two evaluation-related constructs are highly correlated. From an IT investment perspective, alignment of IT and business strategy is problematic if there is a lack of accurate information for evaluation. However, evaluation of IT investments is problematic if there is a lack of alignment of IT and business strategy. To improve the effectiveness of IT investment, management needs to increase alignment of IT and business strategy and information accuracy for evaluation techniques simultaneously. This points to the issues to which management needs to attend in the future.

CONCLUSION

Much work on SITIDs ignores the continuous nature of decisions and the relationships between SITIDS and non-IT SIDs. This article proposed a model that explores mediators in the link between IT investment intensity and SID effectiveness. Survey data show that interaction, information accuracy, and strategic considerations are important factors that mediated the impact of IT investment intensity. Willcocks (1992) emphasizes that management faces a Catch-22 situation with IT investment. They know how important IT is, but they do not know how to evaluate IT projects. The implication is that managers need to pay special attention to the problematic nature of IT investment intensity in SIDs. They should focus on facilitating interaction, ensuring integration of IT strategy with corporate strategy, and improving information accuracy.

REFERENCES

Baron, R., & Kenny, D. (1986). Moderator-mediator variable distinction. *Journal of Personality and Social Psychology, 51*(6), 1173-1182.

Butler, R., Davies, L., Pike, R., & Sharp, J. (1991). Strategic investment decision-making. *Journal of Management Studies, 28*(4), 395-415.

Chou, T.-C., Dyson, R. G., & Powell, P. (1997). Managing stategic IT investment decisions: From involvement to effectiveness. In W. Baets (Ed.), *Proceedings of the Sixth European Conference on Information Systems*, Aix-en-Provence (pp. 938-952).

Dean, J., & Sharfman, M. (1996). Does decision process matter? A study of strategic decision-making effectiveness. *Academy of Management Journal, 39*(2), 368-396.

Fahey, L. (1981). On strategic management decision processes. *Strategic Management Journal, 2*, 43-60.

Freeman, M., & Hobbs, G. (1991). Costly information, informed investors, and the use of sophisticated capital budgeting techniques. In *Proceedings of the Accountants*

Association of Australia and New Zealand (AAANZ) Conference, Brisbane, Australia (pp. 68-74).

Hickson, D., Butler, R., Cray, D., Mallory, G., & Wilson, D. (1986). *Top decisions: Strategic decision-making in organizations*. San Francisco: Jossey-Bass.

Hickson, D., Miller, S., & Wilson, D. (2003). Planned or prioritized: Two options in managing the implementation of strategic decisions. *Journal of Management Studies, 40*(7), 1803-1836.

Hsu, F. (2001). Strategic decision making in a new millennium. *Creativity and Innovation Management, 10*(1), 40-48.

Mohr, L. (1982). *Explaining organizational behavior*. San Francisco: Jossey-Bass.

Nutt, P. (1999). Surprising but true: Half the decisions in organizations fail. *Academy of Management Executive, 13*(4), 75-90.

Papadakis, V. (1995). Contribution of formal planning systems to strategic investment decisions. *British Journal of Management, 16*, 15-28.

Pettigrew, A., McKee, L., & Ferlie, E. (1988). Understanding change in the NHS. *Public Administration, 66*, 297-317.

Porter, M., & Millar, V. (1985, July-August). How information gives you competitive advantage. *Harvard Business Review, 63*(4), 149-160.

Powell, P. (1993). Causality in the alignment of information technology and business strategy. *Journal of Strategic Information Systems, 2*(4), 320-334.

Rajagopalan, N., Rasheed, A., & Datta, D. (1993). Strategic decision processes. *Journal of Management, 19*(2), 349-384.

Sambamurthy, V., Bharadwaj, A., & Grover, V. (2003). Shaping agility through digital options. *MIS Quarterly, 27*(2), 237-263.

Sauer-Leroy, J-B. (2004). Managers and productive investment decisions. *Journal of Small Business Management, 42*(1), 1-18.

Willcocks, L. (1992). IT evaluation: Managing the catch 22. *European Management Journal, 10*(2), 220-229.

KEY TERMS

Contextualism: Integrates process, content, and context to study organizational decision-making.

Decision Content: Content refers to the particular decision under study. Content explores the basic nature and scope of decisions.

Decision Context: The context includes the outer context, which refers to the national economic, political, and social context for an organization, and the inner context, which is the ongoing strategy, structure, culture, management, and political process of the organization. Context helps to shape the process of decision-making.

Decision Process: The actions, reactions, and interactions of the various interested parties as they seek to make a commitment to allocate corporate resources. Process incorporates both the formulation and evaluation processes.

IT Investment Intensity: A concept similar to, but also somewhat different from, the concept of information intensity. Information intensity is the degree to which information is present in the product/service of a business. The degree to which IT is present in an investment decision reflects the IT level of intensity of that decision. Here, IT investment intensity is defined as the ratio of spending on IT to total investment.

Moderators and Mediators: In the social sciences, moderators and mediators have been identified as two functions of third variables. These are subgroups of independent variables that affect given dependent variables via a mediator function.

Strategic IT Investment Decisions: Significant, long-term decisions to invest in projects that have substantial information systems or information technology components. They form part of corporate strategic investment decisions.

Strategic Knowledge Management in Public Organizations

Ari-Veikko Anttiroiko
University of Tampere, Finland

INTRODUCTION

New public management and the more recent concept of new public governance have become the dominant management doctrines in the public sector. Public organizations have become increasingly network-like units with various governance relations with actors from the public, business, and voluntary sectors. Their organization is based more on networks than on traditional hierarchies, accompanied by a transition from the command-and-control type of management to initiate-and-coordinate type of governance.

Among the most critical factors in this transformation is knowledge, for most of what has happened has increased the overall demand to create and process knowledge, and to utilize it in the performance of governmental functions. The success of public organizations depends increasingly on how efficiently they utilize their knowledge assets and manage their knowledge processes in adjusting to local and contextual changes, as illustrated in Figure 1 (cf. Gupta, Sharma, & Hsu, 2004, p. 3; Skyrme, 1999, p. 34, Fletcher, 2003, pp. 82-83). This requires that special attention be paid to strategic knowledge management.

In the early organization theories of public administration, knowledge was predominantly conceptualized within the internal administrative processes, thus to be conceived of as bureaucratic procedures, rationalization of work processes, identification of administrative functions, and selected aspects of formal decision making. New perspectives emerged after World War II in the form of strategic planning and new management doctrines. The lesson learned from strategic thinking is that we need information on the external environment and changes therein in order to be able to adapt to and create new opportunities from these changes (see Ansoff, 1979; Bryson, 1995). As the complexity in societal life and related organizational interdependency has increased due to globalization and other trends, new challenges of managing organization-environment interaction also emerged (cf. Skyrme, 1999, p. 3).

Figure 1. The public organization as an institutional mediator (adopted from Anttiroiko, 2002, p. 272)

BACKGROUND

The branch of management doctrine that became known as knowledge management (KM) reflected actual changes and new ideas in the business world. Classic works that inspired later developments included Polanyi (1966) and Drucker (1969). During the 1980s knowledge became widely recognized as a source of competitiveness, and by the end of the 1990s, knowledge management had become a buzzword. Among the best-known thinkers who contributed to the rise of this field are Peter Senge (1990), Ikujiro Nonaka and Hirotaka Takeuchi (1995), Karl-Erik Sveiby (1997), and Thomas A. Stewart (1997). (For more on the evolution of knowledge management, see Barclay & Murray, 1997; Gupta et al., 2004, pp. 8-10.) It is becoming common understanding that in essence *knowledge management* is about governing the creation, dissemination, and utilization of knowledge in organizations (Gupta et al., 2004, p. 4; Lehaney, Clarke, Coakes, & Jack, 2004, p. 13).

Knowledge cannot be managed in the traditional sense of management. The processing and distribution of information can surely be managed, but it is only one part of the picture. The other concerns knowledge and especially managers' ability to create conditions which stimulate active and dynamic knowledge creation, learning, and knowledge sharing within the organization (e.g. Nonaka, Toyama, & Konno, 2000). To systematize this picture we may say that knowledge management includes four core areas (cf. Gupta et al., 2004; Lehaney et al., 2004):

- *Information Management:* Managing data and information, and designing information and knowledge systems
- *Intellectual Capital Management:* Creating and utilizing knowledge assets, innovations, and intellectual capital.
- *Knowledge Process Management:* Organizing, facilitating, and utilizing sense-making and other knowledge processes.
- *Organizational Learning:* Creating learning and knowledge sharing environments and practices.

Traditionally the most widely applied areas of knowledge management in public organizations used to be data and transaction processing systems, and management information systems serving mainly internal administrative functions. Yet, since the 1980s authorities started to facilitate the exchange of information by local area networks, followed by the Internet revolution of the 1990s. In the early 2000s the knowledge management agenda has focused increasingly on knowledge sharing and learning, and in inter-organizational network and partnership relations (e.g., Wright & Taylor, 2003). As reported by OECD (2003, p. 4), knowledge management ranks high on the management agenda of the great majority of central government organizations across OECD member countries, followed with some time lag by regional and local authorities. Many public organizations have even developed their own KM strategies. The leading countries in this respect include France, Sweden, Finland, and Canada (OECD, 2003, pp. 28-29).

As for more operational actions, there has been a wave of intranet projects at all levels of public administration since the late 1990s. The result is that some 90% of state agencies surveyed by OECD in the early 2000s had their intranets in place. Sectors that really stand out as being well above the OECD average include organizations in charge of finance and budget, of justice, and of trade and industry (OECD, 2003, pp. 20-30). Intranet projects in the public sector aim at creating an Internet-based computer network to securely share information or operations between politicians, administrators, and other employees. Extranet extends such a network outside the organization—that is, to users, partners, service providers, and other stakeholders. Many public organizations in different countries and at different institutional levels have set up such extranet and intranet projects. For example, New York City established in the early 2000s the Human Services Extranet (later renamed the Integrated Human Services Project) to link the city agencies with human service contractors. Similarly, in 2003 the Queensland Government, Australia, established a project aimed at enhancing the effectiveness of e-government service delivery, which was supported by a government-wide extranet. Such projects have been numerous in the public sector since the early 2000s, indicating a transition from information management towards genuine knowledge management. Yet, it is equally true that many public organizations have been slow to embrace knowledge management and knowledge technologies.

FOCUSING ON THE STRATEGIC ASPECT

Combining strategic thinking with knowledge management brings us to the very core of the life of organizations. *Strategic knowledge management* is a set of theories and guidelines that provides tools for managing an organization's knowledge assets and processes of strategic importance for the purpose of achieving organizational goals. It is in this sense about the development of an organization-wide knowledge management capability (Katsoulakos & Rutherford, 2005). The basic idea of strategic knowledge management in the public sector is to ensure that public organizations are capable of high performance by utilizing knowledge assets and knowledge processes when interacting with their environment.

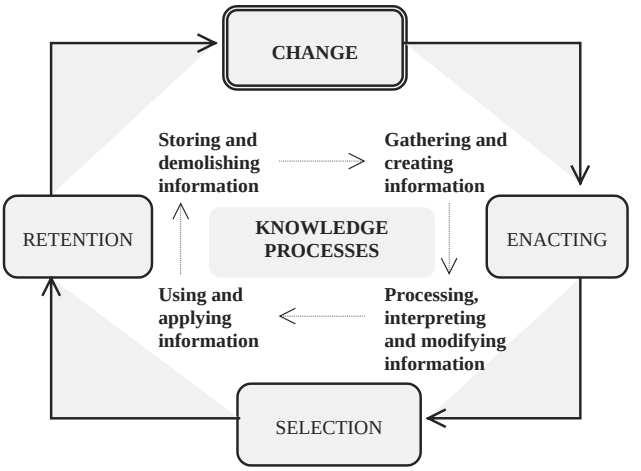
What is essential in strategic knowledge management is that it needs to be 'strategic' in the true sense of the word, as opposed to 'operational'. Public employees sometimes have a tendency to view their knowledge requirements from the point of view of their current work practices. At an organizational level, too, there is sometimes a temptation to map out the future on the basis of current strengths and well-defined short-term challenges. The strategic approach to knowledge aims to overcome such inertia and narrow perspectives by creative knowledge processes, which help to transform views from introspective to outward-looking, from resources to outcomes, and from formal duties to actual impacts and customer satisfaction.

In the knowledge management literature, knowledge has primarily been approached either as an object or a process (cf. Sveiby, 2001). The main focus of public organizations is on knowledge processes framed by certain institutional arrangements. Among the most important of these are the political dimension and democratic control and legally defined functions, competencies, and procedures within territorially defined jurisdictions (for more on KM in the public sector, see e.g. BSI, 2005). This theme will be discussed next.

FACILITATING STRATEGIC KNOWLEDGE PROCESSES

Public organizations possess and process a huge amount of information in their internal operations and external exchange relations. This is why the most important function of their knowledge management practice is to manage knowledge processes and to support knowledge-sharing practices.

Figure 2. Strategic sense-making and related knowledge process of the organization

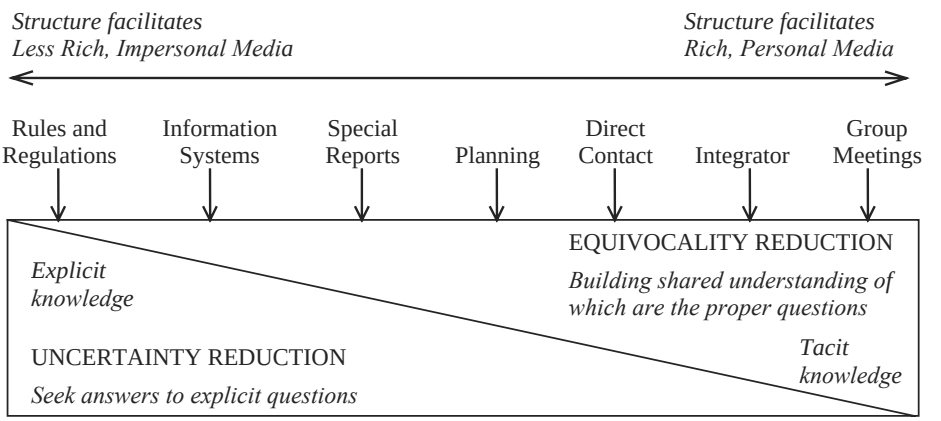


Nonaka (1994) considers an organization's ability to accomplish the task of acquiring, creating, exploiting, and accumulating new knowledge. This formulation takes us very close to how the knowledge process can be operationalized. The *knowledge process* can be defined as an organizational process in which information is gathered, created, processed, used, and dissolved in order to form an enriched orientation base for taking care of an organization's basic functions (cf. Gupta et al., 2004, p. 3; Mendes, Gomes, & Bátiz-Lazo, 2004, p. 153).

It is important to note that in the actual knowledge process, it is more or less meaningless to make a clear-cut distinction between knowledge and information, for both are processed in such a process. For example, knowledge is not simply extracted from information, for knowledge is possessed by human beings and serves as a background and built-in epistemic frame to deal with complexity, novelty, and the requirements of innovativeness (cf. Wiig, 2000). Thus, genuine aspects of the category of *knowledge* are in question when we deal with statements, assumptions, and understandings and such learning and communicative processes in which these knowledge assets can be shared, assessed, and enriched. Many theorists consider tacit knowledge in particular as the most challenging and important form of knowledge in organizations (Polanyi, 1966; Nonaka, 1994). It also needs to be stressed that it is not knowledge as something abstract, but a 'generative dance' or interplay between (a) *knowledge* we possess and (b) *knowing* as an epistemic aspect of the interaction with the world that generates organizational innovation and strategic understanding (Cook & Brown, 1999).

Strategic knowledge processes are those aspects of knowledge processes that have the most profound and far-reaching impact on an organization's adjustment to contextual changes and on its core competencies. A paradigmatic form of a strategic knowledge process is the *sense-making* or *strategy process* in which an organization devotes effort to analyzing its internal attributes and external conditions, and decides on that basis about the action lines in order to achieve its overall goals (cf. Weick, 1995). In such a strategic knowledge process, the organization seeks information on environmental changes and utilizes this in strategy formulation, in which such tools as SWOT analysis have traditionally been used. A basic model of the organizational knowledge-based adaptation process is presented in Figure 2 (Anttiroiko, 2002). This model serves as a heuristic tool to conceptualize knowledge processes. Yet, it is important to keep in mind that this is only a starting point. When taking this idea further, clear-cut sequential stages or phases of the KM lifecycle need to be 'recontextualized' as a set of continuous interdependent sub-processes (cf. Mendes et al., 2004, p. 165). Thus, context-related and situational aspects of knowledge need to be integrated with all essential connections to their environments into the key functions and operations of an organization in order to assess their meaning as a part of actual strategic adaptation and sense-making processes.

Figure 3. Continuum of knowledge facilitation mechanisms (Daft & Lengel, 1986)



Applying Daft and Lengel (1986), we may ask how organization structures and systems should be designed in order to meet the need to manage knowledge processes. Well-designed systems help to decrease the uncertainty and ambiguity faced by an organization by ordering the amount of relevant information and by enabling clarification of problems and challenges. Daft and Lengel (1986) propose seven structural mechanisms that can be used to deal with uncertainty and ambiguity in varying degrees, as illustrated in Figure 3. This model resembles the continuum of communication that has explicit knowledge and tacit knowledge as its extremities (Lehaney et al., 2004, p. 21).

The idea is that these mechanisms form a continuum starting from tools to be used to tackle well-defined problems and thus to reduce uncertainty, and proceeding towards more communicative mechanisms designed to facilitate sense-making processes that aim at reducing equivocality or ambiguity (Anttiroiko, 2002).

As stated, a paradigmatic case for strategic knowledge management is the strategy process of an organization (for more on strategy and information resources, see Fletcher, 2003, pp. 82-84). What is of utmost importance is that managers ensure that people feel involved in the strategy formulation process. The staff also needs to understand the meaning of strategy in their own situations. This would help to make strategy documents living guidance owned by all in the organization, as concluded by Wright and Taylor (2003, p. 198).

Another important premise relates to organization culture and work practices that often impede the development of knowledge management. For example, employees may resist a new knowledge management initiative if they perceive it only as extra work. Similarly, employees may be reluctant to share their knowledge if there are no rewards or tangible benefits for themselves or their organizations. In all, the 'human factor' is essential for improving KM practices, for most of the positive outcomes are the result of the commitment of all employees, successful structural changes in the organization, and the development of the organizational culture and climate (OECD, 2003, p. 4).

THE ROLE OF TECHNOLOGY

Information technology (IT) provides a range of tools that can be effectively used in knowledge management. Such applications are sometimes referred to as knowledge technologies. Relevant applications can support decision making, executive functions, planning, communication, and group work. Introduction of new knowledge technologies may be based on a stages of growth model for knowledge management technology, where organizations develop from the person-to-tools strategy, via the person-to-person strategy and the person-to-documents strategy, to the person-to-systems strategy (Gottschalk, 2005).

Tools and technologies available for knowledge management include generic communication tools (e.g., e-mail), blogs, computer-based information and decision support systems, document management systems, wikis, intranets and extranets, groupware, geographic information systems, help-desk technology, and a range of knowledge representation tools (Gupta et al., 2004, pp. 17-24; Grafton & Permaloff, 2003.) In general, the Internet may be suggested as the KM infrastructure due to its widespread availability, open architecture, and developed interfaces (Jennex, 2003, p., 138).

In real life, most of the tools applied in knowledge management are more or less conventional, such as training, seminars, meetings, and the like. Various KM-specific organizational arrangements had been adopted by about half of the organizations studied in the OECD survey on ministries, departments, and agencies of central government in the early 2000s. These measures include central coordination units for KM, quality groups, knowledge networks, and chief knowledge officers. Another important application area is the classification of information, referring to new filing mechanisms, e-archives, and new types of databases. In internal knowledge-sharing, intranet projects form the mainstream, combined with wide access to the Internet and having e-mail addresses for the staff. The external knowledge sharing goes largely hand in hand with the emergence of new practices of e-governance. These practices have increased the knowledge sharing in both local and wider governance processes (OECD, 2003, pp. 17-20; Anttiroiko, 2004).

FUTURE TRENDS

The future challenge for public organizations is to increase their responsiveness to stakeholders, especially to citizens. At the same time they need to be capable of strategic institutional mediation in the increasingly turbulent environment, thus bringing an element of continuity and stability to social life, and guaranteeing democratic and civic rights at different institutional levels.

Another trend that may change some premises of strategic knowledge management is the changing nature of media and information landscapes. One indication of this is Web 2.0, which refers to a second-generation of Internet-based services that allow people to collaborate and share information online in new ways (blogs, wikis, ubiquitous technologies, etc.). All this requires increasing capacity to manage knowledge of strategic importance and create innovations of knowledge management (see Montano, 2004).

CONCLUSION

Strategic knowledge management refers to the theory and practice of managing knowledge assets and processes of

strategic importance. Public organizations need to create favorable organization structures and environments for knowledge sharing, organizational learning, and other aspects of knowledge management in order to create all the knowledge they require in their adjustment and trend-setting processes.

A main return of strategic knowledge management is better capability to adjust to contextual changes. This is difficult to measure, even if such tools as Balanced Scorecard (BSC), the Intangible Assets Monitor (IAM), and Intellectual Capital Index (ICI) are available. This is because they provide only a partial picture of KM performance, as claimed by Chaudhry (2003, p. 63). What seems to be needed is more process-focused assessments that are able to analyze the steps of KM processes, thus highlighting the actual changes in organizational knowledge base, capacities, and processes. As usual, there is no measurement system that fits all organizations in all situations. Rather, measurement should be tailored to the actual needs of the organization.

REFERENCES

Ansoff, H.I. (1979). *Strategic management.* London: Macmillan.

Anttiroiko, A.-V. (2002). Strategic knowledge management in local government. In A. Grönlund (Ed.), *Electronic government: Design, applications & management* (pp. 268-298). Hershey, PA: Idea Group.

Anttiroiko, A.-V. (2004). Introduction to democratic e-governance. In M. Malkia, A.-V. Anttiroiko, & R. Savolainen (Eds.), *eTransformation in governance.* Hershey, PA: Idea Group.

Barclay, R.O., & Murray, P.C. (1997). *What is knowledge management?* Retrieved January 12, 2007, from http://www.media-access.com/whatis.html

Bryson, J.M. (1995). *Strategic planning for public and nonprofit organizations. A guide to strengthening and sustaining organizational achievement* (revised ed.). San Francisco: Jossey-Bass.

BSI. (2005). *Knowledge management in the public sector: A guide to good practice.* PD 7504:2005, British Standards Institution (BSI), UK.

Chaudhry, A.S. (2003). What difference does it make: Measuring returns of knowledge management. In E. Coakes (Ed.), *Knowledge management: Current issues and challenges.* Hershey, PA: IRM Press.

Cook, S.D.N., & Brown, J.S. (1999). Bridging epistemologies: The generative dance between organizational knowledge and organizational knowing. *Organization Science, 10*(4), 381-400.

Daft, R.L., & Lengel, R.H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

Drucker, P. (1969). *The age of discontinuity. Guidelines to our changing society.* London: Heinemann.

Fletcher, P.D. (2003). The realities of the Paperwork Reduction Act of 1995: A government-wide strategy for information resources management. In G. David Garson (Ed.), *Public information technology: Policy and management issues*. Hershey, PA: Idea Group.

Gottschalk. P. (2005). *Strategic knowledge management technology.* Hershey, PA: Idea Group.

Grafton, C., & Permaloff, A. (2003). Computer tools for better public sector management. In G. David Garson (Ed.), *Public information technology: Policy and management issues*. Hershey, PA: Idea Group.

Gupta, J.N.D., Sharma, S.K., & Hsu, J. (2004). An overview of knowledge management. In J.N.D. Gupta & S.K. Sharma (Eds.), *Creating knowledge-based organizations.* Hershey, PA: Idea Group.

Jennex, M.E. (2003). A survey of Internet support for knowledge management organizational memory systems. In E. Coakes (Ed.), *Knowledge management: Current issues and challenges.* Hershey, PA: IRM Press.

Katsoulakos, P., & Rutherford, A. (2005). *An introduction to knowledge oriented strategy KoS and strategic knowledge management capabilities.* Retrieved January 10, 2007, from http://www.inlecom.com/uploadfiles/An%20introduction%20to%20Knowledge%20oriented%20Strategy%20and%20Strategic%20Knowledge%20Management%20Capabilities.pdf

Lehaney, B., Clarke, S., Coakes, E., & Jack, G. (2004). *Beyond knowledge management.* Hershey, PA: Idea Group.

Mendez , M.M., Gomes, J.F.S., & Bátiz-Lazo, B. (2004). Management of knowledge in new product development in Portuguese higher education. In J.N.D. Gupta & S.K. Sharma (Eds.), *Creating knowledge-based organizations.* Hershey, PA: Idea Group.

Montano, B. (2004). *Innovations of knowledge management.* Hershey, PA: IRM Press.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(2), 14-37.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company. How Japanese companies create the dynamics of innovation.* New York: Oxford University Press.

Nonaka, I., Toyama, R., & Konno, N. (2000). SECI, Ba and leadership: A unified model of dynamic knowledge creation. *Long Range Planning, 33* , 5-34.

OECD. (2003, April 3-4). The learning government: Introduction and draft results of the survey of knowledge management practices in ministries/departments/agencies of central government. *Proceedings of the 27th Session of the Public Management Committee,* Organisation for Economic Cooperation and Development (OECD), Paris.

Polanyi, M. (1966). *The tacit dimension.* Garden City, NY: Doubleday.

Senge, P.M. (1990). *The fifth discipline. The art and practice of the learning organization.* New York: Doubleday.

Skyrme, D.J. (1999). *Knowledge networking. Creating the collaborative enterprise.* Oxford: Butterworth-Heinemann.

Stewart, T.A. (1997). *Intellectual capital: The new wealth of organizations.* New York: Currency/Doubleday.

Sveiby, K.E. (1997). *The new organizational wealth. Managing and measuring knowledge-based assets.* San Francisco: Berrett-Koehler.

Sveiby, K.-E. (2001, March). *What is knowledge management?* Retrieved January 10, 2007, from http://www.sveiby.com/Portals/0/articles/KnowledgeManagement.html

Weick, K. (1995). *Sense-making in organizations.* Thousand Oaks, CA: Sage.

Wiig, K. (2000). Knowledge management: An emerging discipline rooted in a long history. In C. Despres & D. Chauvel (Eds.), *Knowledge horizons. The present and the promise of knowledge management.* Boston: Butterworth-Heinemann.

Wright, G., & Taylor, A. (2003). Strategic knowledge sharing for improved public service delivery: Managing an innovative culture for effective partnerships. In E. Coakes (Ed.), *Knowledge management: Current issues and challenges.* Hershey, PA: IRM Press.

KEY TERMS

Intellectual Capital (IC): Knowledge and know-how possessed by an individual or an organization that can be converted into value in markets. Roughly the same as the concept of intangible assets.

Intellectual Capital Management (ICM): A management of value creation through intangible assets. Close to the concept of knowledge management.

Intellectual Property (IP): Any product of the human intellect that is unique and has some value in the marketplace. It may be an idea, composition, invention, method, formula, computer software, or something similar. In practice, special attention is paid to such intellectual property that can be protected by the law (e.g., patent and copyright).

Knowledge Assets (KAs): Statements, assumptions, abstract models, and other forms of knowledge regarding the organization itself and its environment (markets, customers, etc.) that an organization possesses. These assets provide economic or other value to an organization when interacting within it or with its environment.

Knowledge Management (KM): Management theory and practice on managing intellectual capital and knowledge assets, and also the processes that act upon them. In a practical sense KM is about governing the creation, dissemination, and utilization of knowledge in organizations.

Knowledge Management System (KMS): Set of tools and processes used by knowledge workers to identify and transmit knowledge to the knowledge base contained in the organizational memory.

Organizational Learning (OL): An organizational process in which the intentional and unintentional processing of knowledge within a variety of structural arrangements is used to create an enriched knowledge and orientation base, and a better organizational capacity for the purpose of improving organizational action.

A Structured Approach to Developing a Business Case for New Enterprise Information Systems

Francisco Chia Cua
Otago Polytechnic, New Zealand

Tony C. Garrett
Korea University, Republic of Korea

INTRODUCTION

The term business case is used to describe both a process and a document. A business case exploits an initiative. Exploiting the initiative from awareness to implementation encompasses a process, referred to in the diffusion of innovation parlance, as the innovation-decision process. The development of a business case concerns this innovation-decision process. The individuals or the decision-making units pass through the innovation-decision process, gaining knowledge of a new idea, forming an attitude toward it, and deciding whether to adopt or reject it (Rogers, 2003, p 20). Gaining the knowledge triggers the awareness or enforces it. Then, it leads to setting the agenda. After the agenda-setting stage is the examination of the available options. Attributes of competing options are matched together, enabling attitude formation in favour or against a particular option. This results in the creation of a shortlist of two or three options. A decision is generally reached at this point. The decision is, therefore, part of the matching stage. However, this is not always true in an organisational setting. There is a third stage after the matching stage. It is the decision (aka, business case) stage. Organisations generally demand rigour in making the decision. A business case document embodies the rigour in the business case development. Consequently, the decision stage culminates with a completed business case document and the decision that results from it: to adopt or reject the innovation. The three stages, agenda setting, matching, and decision stages, compose the initiation phase. If the decision favours adoption, then the implementation phase proceeds. In the context of implementing the new enterprise information systems, the stages in the implementation phase consists of pre-production, production, post-production (that is, maintenance), and confirmation stages. In summary, the business case development is a means, and its end is a business case document.

A complete business case is a formal written argument and a detailed "point by point" analysis (Cannon, 2006, p 4; Carruth, 2001). It purports to justify the adoption or rejection of investing and thereby, implementing the new enterprise information systems. The analysis takes into consideration the stakeholders (Ministry of Health, March 2005), especially the decision-makers and the end-users. Consequently, a business case document is formal, detailed, and complex.

Using the parlance of "diffusion of innovations" (DOI) theory (Rogers, 1962, 2003), a business case document is a communication tool used to diffuse the new enterprise information systems, and to justify their adoption and implementation. Diffusion refers to the process by which the executive sponsor, who owns the innovation-decision process of the new enterprise information systems, communicates to the upper managers to get their approval of the project and funding.

Diffusion via a business case document for a technological product, such as an enterprise information system, must be directed at a single target audience to be effective. The upper managers represent a chasm that needs to be bridged (Moore, 1991). A completed business case document, containing relevant information for the managers, can serve as that bridge.

In addition to its relevance, a business case must also be responsible and credible. Therefore, the business case must bring relevance, reputation, and responsibility (the 3 Rs) into a number of issues and challenges during its development.

This chapter proposes a business case structure, with the 3Rs underlying it. It continues from the big picture of business case development in the article, *The Role of Business Case Development in the Diffusion of Innovations Theory for EISs* (hereinafter referred to as *The Role of Business Case Development...*). The structure, suggested in Table 2, delineates the context to the new enterprise information systems. However, prior to that, certain issues must be addressed:

- What is the purpose of a business case?
- What should a business case document contain?
- How should the business case be structured? How should it be written?

THE PURPOSE OF A BUSINESS CASE

A good business case must have a purpose that is clear, specific, and relevant to the organisation and the upper managers. *The Role of Business Case Development...* mentions "growth and sustainability" as a strategic goal. That term is too broad to be useful in a business case. A similar ambiguous construct is sustainable competitive advantage (Hammer, 1996; Monczka, Carter, Petersen, & McDowell, 2006, p 213). A single detailed statement is far more relevant to the organization than several broad statements. An actual case study reveals one objective of selecting and implementing a proven, up-to-date enterprise financial system. The chosen systems must have the capacity to meet likely future financial-related requirements and growth. This vague objective can be made clearer by citing sustainable competitive advantages that are strategically valuable to the organisation, taking into consideration certain guiding principles (Table 1).

Enterprise information systems are the enabling technologies that foster sustainable competitive advantages under certain guiding principles. Table 1 helps to develop the business case backward from the purpose, and includes ALL the planning activities, resources, and metrics that are critical to the goal (Ministry of Health, 2005, p 9). Recent research indicates that most successful organisations have a crystal clear notion of the organisation's strategy, and how deploying information technologies can help actualise that strategy.

A Business Case Is a Two-Sided Coin

The purpose of a business case is to reflect the rigour of planning relative to the level of investment being undertaken (Ministry of Health, 2005, p 3). This is one side of the coin. The other side relates to justifying the innovation discussed. On one side, there are sustainable competitive advantages or other reasons to justify the implementation of the innovation. On the other side is the purpose of the business case. Justifying the implementation and matching the best fit pertain to the what and the why questions. Thinking about the innovation-decision process and substantiating the process with the necessary rigour concern the how (and why) questions. In the context of the new enterprise information systems, the specific sustainable competitive advantage represents the primary goals. The enterprise information systems enable seamless integration (Table 1) of information across the whole organisation and its extended social systems. The systems empower the internal people in the organisation to provide the best performance with shared information systems. The systems help to identify and develop centres of excellence with resources and expertise prior to outsourcing. The systems also maximise return on information systems investment across the organisation, maximise the exploitation of opportunities, and minimise the risk associated with the new implementation. All these technological advantages are about the innovation. However, that is not the purpose of the business case, which is a reflection of the rigour on the innovation-decision process.

WHAT SHOULD A BUSINESS CASE DOCUMENT CONTAIN?

The Scope and Content of Business Cases Vary

Not only must the business case contain the justification of the innovation (what and why) and the rigour of the process

Table 1. Guiding principles and sustainable competitive advantages

Guiding principles	Sustainable competitive advantages
• E-business supply chain • Economic value-added focus • Globalisation • Satisfaction of the needs of customers • Total value management • Value/supply chain integration, productivity, and collaboration (operational excellence and process redesign)	• Seamless integration • Best performance: Quality, price, delivery, technology, cycle time (velocity, responsiveness, service), safety • Enhanced EVA (increase revenue by broadening the offerings and improving customer value, reduce internal and external cost structure, reduce assets, and improve asset utilization) • Enhanced EBIT, ROI, cash flows • Perceived highest customer value • Revenue generation • Time-to-market/breakeven

Adopted from Carruth (2001); Hammer (1996); Ministry of Health (2005), Monczka et al. (2006)

Table 2. Suggested structure of a business case

	Section	Caption	Remarks
E		Title Page	
E		Table of Contents	
E		Executive Summary	Prepare a sharp and compelling summary.
E	1	INTRODUCTION	Brief, purpose, scope, limitations
E		Terms of Reference or Background	State the brief (that is, who asks to do what?). What do the readers need to know? If the background is crucial, then there can be a separate Background section.
E		Purpose of the Report	Briefly explain the purpose of the report.
		Gaps Analysis	Describe the expected consequences (the strategic vision in refer to Section 4). Describe the assessment of the present state (refer to Section 4). Describe the gaps between the present state and the desired future state.
E		The Proposed Solutions and Its Strategic Value	Highlight the strategic value, the selling point, of the solution. Describe the relevant issues and their action plans (related to Section 7, the proposed project). Some questions to ponder are: Is the solution reactive or anticipatory (Figure 3 in *The Role of Business Case Development...*)? How does the proposed solution FIT into the big picture? If the solution is reactive, describe the urgency of the radical change. If the solution is anticipatory, explain the critical strategic area (Table 1 and strategic vision in Section 4 and Section 6). Why does the solution matter?
E		Project Ownership (and Consultant)	Corporate power dictates the level of detail this section requires (assumed). Who is the executive sponsor? How supportive is the executive sponsor to the business case? Who is the expert the executive sponsor has consulted?
E		Scope and Delimitation	Delineate the scope of the problem.
O		Methodology	Briefly describe the methods or methodology of Section 3.
E		Structure of the Business Case Document	Briefly outline how the business case is presented.
	2	CONCLUSIONS AND RECOMMENDATIONS	
E	2.1	Conclusions	How does the executive sponsor evaluate the Business Case (eg, cost-benefit analysis)? How logical or intuitive is the evaluation? Based on the financial and nonfinancial analysis, summarise briefly how the business case will impact positively and negatively on the organisation? What are the critical success factors? The critical failure factors? How are the options considered and documented?
O	2.2	Recommendations	Which option has the executive sponsor chosen and why? The Recommendations section (or the Proposed Project section) may attempt to answer further questions such as: How does the executive sponsor propose to rally the support, involvement, and usage? How does the executive sponsor ensure success?
O	3	METHODS	What are the contingency plans? How does the executive sponsor come about the plan? How reliable is the evidence used to develop the Business Case?
O	4	FACTS AND ASSUMPTIONS	This section may form part of the Introduction section.

Legend: E = Essential, O = Optional

continued on following page

Table 2. continued

	Section	Caption	Remarks
O		Strategic Vision	Identify the strategic vision, the guiding principles, and sustainable competitive advantages (Table 1). Assess the key strengths and weaknesses. This may be incorporated under the future state of the Introduction section.
O		Needs Analysis	Describe briefly the needs and relevant issues. This section may be incorporated under the Gaps Analysis of the Introduction section.
O		Options Analysis	Describe briefly the options considered in this business case. This section may be incorporated under the Proposed Solutions of the Introduction section.
	5	FINANCIAL ANALYSIS	How does the EXECUTIVE SPONSOR explore the sensitivity of key assumptions in the analysis?
E		Total Cost of Ownership (TCO)	Combine the net present value, discounted cash flows, and total cost of ownership in the financial analysis. Determine which option enhances the value to the organisation and not which option has the lowest TCO. Make sure to scrutinise all "hidden costs."
E		Business Risks	How does the EXECUTIVE SPONSOR identify, assess, and resolve the risks involved?
E	6	NONFINANCIAL ANALYSIS	What external factors drive the Business Case? How will the competitive advantage, macro environments, and stakeholders (competitors, customers, and vendors) impact the organisation? With what of key assumptions in the analysis?
E		Sustainable Competitive Advantage(s)	Describe in detail the innovation, its scope, its goals (the sustainable competitive advantages in Table 1), and the expected consequences. How does the innovation fit into, for example, the seamless alignment or the organisation's strategy? How does the innovation enhance the competitive advantage? How does it create the differentiation of the organisation from its competitors? What is the "compelling reason" why the innovation must be put into action? What value does the innovation provide?
E		Macro environments	
E		Stakeholder Analysis	St Gallen Management Model suggests two approaches. One is the "strategic stakeholder value." The other is the "ethically critical stakeholder value." The former approach assumes that a balanced consideration of the long-term interests of all stakeholders is the best way to maximise shareholder value. The latter approach evaluates all potential stakeholders equally toward "ethically justifiable legitimacy." (Rüegg-Stürm, 2005)
E	7	THE PROPOSED PROJECT	
		Project Plan	How does the project plan look like? What are the key milestones? How abstract or detailed is the project plan? How does the project plan mitigate the risks of occurrence of undesirable consequences (refer to Appendix A section)?
		The innovation-decision process	How did the executive sponsor conduct the initiation phase? What is the change strategy? How will the executive sponsor implement the new enterprise information systems? Reiterate the change strategy, costs, and risks?

Legend: E = Essential, O = Optional

continued on following page

Table 2. continued

	Section	Caption	Remarks
		Project Participants	How detailed has the Business Case been documented with regards to the resourcing requirements (that is, the capacity and capability requirements)? Specifically, what are the skills, experience, and time commitment required of the project team? What are the skills, experience, and time commitment required within the business after the completion of the innovation-decision process? How does the executive sponsor propose to manage the risks related to the resourcing requirements? How does the Business Case take into account the additional resourcing required through and after the implementation-decision phase?
E	Appendix A	UNDESIRABLE CONSEQUENCES TO AVOID	Refer to Project Plan section above.
E	Appendix B	REFERENCE SITES	

Legend: E = Essential, O = Optional

(how and why), it must be relevant to the stage of the innovation-decision process at which it is written. A business case written at the awareness stage (Stage 1) or matching stage (Stage 2) differs in scope and content with a business case written at the decision stage (Stage 3). For example, the Ministry of Health (2005) of New Zealand has issued guidelines for the investment in information technology. Its business case development is composed of three stages equivalent to the three stages in the initiation phase: the strategic analysis, options analysis, and completed business case. Respectively, the outputs are the needs and options analysis, both of which are part of the completed business case document. Each output is a specifically differentiated business case. The needs analysis is a business case. So is the options analysis. The needs analysis and options analysis represent a section in the completed business case document. The completed business case inevitably contains sections that correspond to the three stages of the business case development.

Analysis, Depth, and Quality

The keys to a successful business case are in-depth analysis and quality. A Risk-adverse organisation with an analytical bent will usually spend considerable time and effort in fully understanding all the aspects and implications of any significant investment (Carruth, 2001, p 10). Too much analysis is not cost effective. The content of the business case must limit and highlight few but crucial aspects that impact the organisation. Implementing new information systems affect the operation and well-being of the organisation concerned. The question is: Which of these "effects" give the greatest impacts to the organisation? Thus, there is a need to limit and highlight the critical factors. The focus is, therefore, on the quality of the information and the depth of analysis. Cannon (2006, pp. 2-4, 194-252) suggests 29 analytical tools, such as balanced scorecards, cost-benefit analysis, critical success factor, life-time cost analysis, risk analysis, sensitivity analysis, and SWOT analysis. The amount of analysis carried out is irrelevant. In a good business case, the analysis must reflect the right depth of a critical factor that impacts the organisation as a consequence of strategically investing or not investing the chosen option (Weill & Ross, 2004).

Managing Risk and Defining Undesirable Consequences

Aside from the in-depth quality analysis, the business case must identify important assumptions (also referred in Figure 1 of *The Role of Business Case Development...*). Risk-taking attitude, sustainable competitive advantages, organisational innovativeness, politics, cash flows, and total cost of ownership are examples of relevant assumptions.

Risk is foremost to these assumptions. The implementation of new enterprise information systems involves risk. How does the organisation react to risk? The answer can be found in the organisations past actions.Risk-seeking, innovative organisations exploit innovative opportunities which have potential for big returns. Risk-averse organisations are very late adopters of technology, and are less innovative. To risk-averse organisations, the business case must mention similar implementations undertaken by competitors or other organisations (refer to Appendix A section in Table 2).

Managing risk is simply gaining more power over the uncertainty brought about by the innovation-decision process. Borge (2001) suggests certain techniques to managing risk:

- Being aware of the risks by defining, at the start, the possible undesirable outcomes;
- Knowing that taking deliberate action can increase the odds of desirable consequence and decrease the chance of unexpected or undesirable consequences;
- Weighing risks vs. benefits

Defining all possible undesirable consequences is easily overlooked or taken for granted because, as Borge (2001) puts it, there is no universal definition of a bad outcome. Like the "new idea" to exploit, the "undesirable consequence" to avoid depends on the perspective of the people involved. Therefore, the undesirable consequence must be explicit in the business case. As mandated by the Sarbanes-Oxley Act, the executive sponsor and the project team members should be able to see the threat early, and understand them before they become clear and present dangers (Green, 2004).

Here is a last word on the risky business of managing risk. The rigour of developing a business case is NOT minimising the risk. It is balancing risk with opportunity to create the best overall value (for details, refer to the concept of "value at risk" by Borge).

Other Important Assumptions

The attitude to risk is not the only relevant assumption. There are other assumptions that need to be addressed and identified. For example, how do new enterprise information systems foster seamless alignment in the organisation? How does the proposed implementation fit into the organisation's strategy? Table 1 has suggested certain sustainable competitive advantages. Therefore, how does the business case describe in detail the innovation, its scope, its goals (the sustainable competitive advantages), and its objectives (the expected consequences)?

Corporate culture, specifically the power structure mentioned in the conceptual framework of *The Role of Business Case Development...*, is another basic assumption. A business case may rely more heavily on the opinions of the executive sponsor or any other leaders with powerful opinions. The mention of the "power" person or people gives credibility (or reputation, the second R) and responsibility (the third R) to the business case. If the business case is to diffuse successfully, the premise of the corporate culture dictates a disclosure of ownership of the implementation. Furthermore, IT governance essentially involves stating who is the executive sponsor who is responsible to making the IT decision and who will be accountable for diffusing its usage in the organisation (Weill & Ross, 2004; refer to project ownership in the Introduction section of Table 2).

NPV, DCF, and TCO

The net present value is an old measure that most accountants and financial professionals rely heavily on valuation to business decision making.

One related valuation method is the discounted cash flows (DCF) method. It estimates the current market dollar value of the new enterprise information systems. Based on expected future cash flows and discount rates, the DCF value drives capital budgeting and therefore, the innovation-decision process. For organisations with positive cash flows, the cash outflows for future periods can be estimated with a certain degree of reliability.

Another important concept is the total cost of ownership, or TCO for short. The traditional notion of cost is the money paid in exchange for tangible and intangible projects or services, where costs can be either direct or indirect. Before making the adoption decision, a capital budgeting analysis will be required. It involves breaking down the amounts to be paid. For example, with the enterprise information systems, the acquisition costs include initial software, hardware, installations, training, configuration, supporting, and consultation costs. Other cash outflows are used for maintenance, further training requirements, modifications, supports, and upgrades (Cua & Theivananthampillai , 2006; Piedad, 2001). The initial acquisition cost represents a capital expenditure. The subsequent ongoing costs are operating expenditures. These so-called operating "expenses" accumulate over time, and equate to at least a third of the initial acquisition cost (Schweitzer, 2003). They must be included in the total cost of ownership, otherwise, they will become hidden costs.

Combining the net present value, discounted cash flows, and total cost of ownership, give to financial analysis the much needed context as to valuation, strategy, finance, and corporate governance (Morin & Jarrell, 2001). There is one caveat. TCO is like air. Its absence is fatal, but it must remain discreet and in the background when present. Although determining the lowest TCO is good, deciding on the option with the lowest TCO puts emphasis on minimising the cost through TCO. It is, therefore, not wise. Instead, the question to ask is not which option has the lowest TCO? Rather, which option enhances the value to the organisation (Cua & Theivananthampillai, 2006; Morin & Jarrell, 2001)?

Planning for the Past, Present, and Future

In DOI, the innovation-decision process develops along a normal course. Imagine the whole project as a life cycle consisting of the two major phases (the initiation and implementation phases) with several stages in each phase. The completed business case (the third type of the business case) occurs in the decision stage of the first phase, and

effectively links the initiation phase to the implementation phase. Thus, a completed business case must contain the story of the past (what was done in the awareness stage and matching stage), the present (the available options and critical factors to make the decision), and a preview of the project management (Nokes, Greenwood, Major, & Goodman, 2004) with regards to the implementation of the new enterprise information systems.

Nonfinancial Factors

Other than the sustainable competitive advantages, there are other nonfinancial factors, such as the macro environments and the stakeholders. These are two of the six categories under the St Gallen Management Model (Rüegg-Stürm, 2005). The macro environments are particularly useful in weighing the external circumstances that can impact on organisation's growth and sustainability. Organisations often define their success by the degree to which they are able to meet the needs of the various stakeholders. Some stakeholders are the conditions to success. Others affect the creation of value. The purpose of stakeholders analysis is to maximise shareholder value in the longer-term (Morin & Jarrell, 2001; Rüegg-Stürm, 2005).

STRUCTURING AND WRITING THE BUSINESS CASE

Using a business report format, the business case (Table 2) identifies the gaps between the expected future state and the present state (Section 1). The conclusions on the business case, together with the recommendations, follow in Section 2. Subsequent sections are the methods (Section 3), facts and assumptions (Section 4), financial analysis (Section 5), nonfinancial analysis (Section 6), and the implementation plan (Section 7).

As business cases are complex, the executive summary becomes the most important part of the business case. Pugh and Bacon (2004) emphasise the "executive treatment," and suggest several approaches of diffusing the innovation via the executive summary. The upper managers are busy people. They have time to read a summary but not an entire proposal. They also have a crucial role in the final decision and therefore, they will form their "attitude" towards the business case in three quick steps: the executive summary (first step), the introduction (the second step), and the conclusions and recommendations (the third step). By the third step, they usually have an idea whether or not the executive sponsor has undertaken the necessary rigour in developing the business case, and whether or not the business case is worth investing.

CONCLUSIONS

The key points in the business case are the 3Rs (relevance, responsiblity, and reputation) of the strategic value of the new enterprise information systems. What value impacts the organisation the most? With the relevance brought about by the strategic value come the in-depth analysis and the quality (not quantity) of information. The project ownership in Section 1 and the stakeholder analysis in Table 2 concern responsibility, crucial to Sarbanes-Oxley Act and corporate/IT governance. Project ownership sets accountability for the decision and diffusing the usage of the new enterprise information systems. Stakeholders' analysis sets the social responsibility. The rigour of the business case development foster a mindset that the organisation is able to fully consider risks, and that the organisation is able to mitigate the risk concerned with the initiation and implementation of the new enterprise information systems. The reputation of the organisation is at stake in this instance.

Lastly, the proposed structure in Table 2 is a blueprint, but should not be used as a boilerplate.

REFERENCES

Borge, D. (2001). *The book of risk*. New York: John Wiley & Sons, Inc.

Cannon, J. A. (2006). *Making the business case: How to create, write, and implement a successful business plan*. London: Chartered Institute of Personnel and Development.

Carruth, B. (2001). *Develop a successful business case for board approval*. Wellington, New Zealand: Institute of Chartered Accountants of New Zealand.

Cua, F. C., & Theivananthampillai, P. (2006). *Value management of sourcing decisions: The cost of ownership in performance management systems.* Paper presented at the Pacific Asian Consortium for International Business Education & Research (PACIBER) 2006, Cebu, Philippines.

Green, S. (2004). *Manager's guide to the Sarbanes-Oxley Act: Improving internal controls to prevent fraud*. New York: John Wiley & Sons.

Hammer, M. (1996). *Beyond reengineering: How the process-centered organization is changing our work and our lives*. New York: HarperCollins Publishers, Inc.

Ministry of Health. (2005). *Business case guidelines for investment in information technology*. Wellington, New Zealand.

Monczka, R. M., Carter, P. L., Petersen, K. J., & McDowell, C. P. (2006). Project 10X: The value proposition and

strategic impact to sourcing and supply effectiveness. In J. H. Cavinato, A. E. Flynn, & R. G. Kauffman (Eds.), *The supply management handbook* (7th ed.) (pp. 209-232). New York: McGraw-Hill.

Moore, G. A. (1991). *Crossing the chasm: Marketing and selling disruptive products to mainstream customers*. New York: HarperCollins Publishers.

Morin, R. A. & Jarrell, S. L. (2001). *Driving shareholder value*. New York: McGraw-Hill.

Nokes, S., Greenwood, A., Major, I., & Goodman, M. (2004). *The definitive guide to project management: Every executive's fast-track to delivering on time and on budget*. New York: Financial Times Prentice Hall.

Piedad, F. (2001). *Total cost of ownership: Principles and practical applications.* Retrieved 17 April 2006, from http://www.phptr.com/articles/printerfriendly.asp?p=24404

Pugh, D. G., & Bacon, T. R. (2004). *Powerful proposals: How to give your business the winning edge*. New York: American Management Association.

Rogers, E. M. (1962). *Diffusion of innovations*. New York: The Free Press of Glencoe.

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York: Simon & Schuster, Inc.

Rüegg-Stürm, J. (2005). *The new St. Gallen management model: Basic categories of an integrated managment*. New York: Palgrave Macmillan.

Schweitzer, D. (Sept 2003). Track the true TCO: Watch out for hidden costs over the long term. *Processor, 25*(39).

Weill, P. & Ross, J. W. (2004). *IT governance: How top performers manage IT decision rights for superior results*. Boston, MA: Harvard Business School Press.

KEY TERMS

Business Case: Completed business case document. Business case process.

B**usiness Case Development**: Walks through the initiation phase of the innovation-decision process and talks about the project plans that concern the implementation phase.

Completed Business Case Document: A formal written document that argues a course of action. It contains a point-by-point analysis to making a decision for a set of alternative courses of action to accomplish a specific goal.

Diffusion: Essentially communicating a new idea (aka, the innovation) within a social system (such as an organisation) with the intention that the audience of that communication adopts or use the innovation.

Diffusion of Innovations: Theory concerns the how, why, and at what rate the new idea (commonly referred to as innovation) diffuses.

Implementation Phase: Proceeding after the initiation phase, the implementation phase of enterprise information systems consists of pre-production, production, and post-production (also known as upgrade and maintenance). Refer to innovation-decision process.

Initiation Phase: Consists of awareness stage, matching stage, and lastly, the decision stage. It is the first phase of the innovation-decision process. The second phase is the implementation phase. Refer to innovation-decision process.

Innovation: Represents a product, a service, or an idea that is perceived, or should be perceived by the audience or the market in which this innovation is intended to be new and of value.

Innovation-Decision Process: Starts with an **initiation phase** through which the individuals or decision-making units move from knowing (understanding/identifying) the new idea (the innovation), to forming of an attitude toward the innovation, and subsequently, to deciding whether to adopt or reject the implementation and use of the new idea. The awareness stage is the agenda setting stage. The attitude formation stage is the matching stage. In addition, the decision stage to adopt or reject the innovation terminates the initiation phase. An adoption decision continues the process toward the **implementation phase**, which consists of the pre-production, production, post-production, and confirmation stages.

Risk: Connotes a possible negative impact to something of value. It symbolises the probability of a loss.

Total Cost of Ownership: Also known as TCO, is a rigorous and holistic methodology. It helps to estimate how much an investment will cost to operate over its lifetime. It takes into account all direct and indirect costs. The indirect costs are generally insignificant individually. However, they become very substantial when accumulated over time.

A Study of Image Engineering

Yu-Jin Zhang
Tsinghua University, Beijing, China

INTRODUCTION

Images are an important medium from which human beings observe the majority of the information they received from the real world. In its general sense, the word "image" could include all entities that can be visualized, such as a still image, video, animation, graphics, charts, drawings, even also text, and so forth. Nowadays, "image" rather than "picture" is used because computers store numerical images of a picture or scene. Image techniques, which are expanding over wider and wider application areas, have attracted more and more attention in recent years. Image engineering (IE), an integrated discipline/subject comprising the study of all the different branches of image techniques, is evolving quickly.

From 1969 to 2000, a well-known bibliography series had been developed to offer a convenient compendium of the research in picture processing until 1986, as well as in image processing and computer vision after 1986. This series has been ended in 2000 by the author after a total of 30 survey papers were published (Rosenfeld, 2000a). Some limitations of this series for the termination are (Zhang, 2002b):

1. No attempt is made to summarize the cited references for each year.
2. No attempt is made to analyze the distributions of the selected references from various sources.
3. No attempt is made to provide statistics about the classified references in each group.

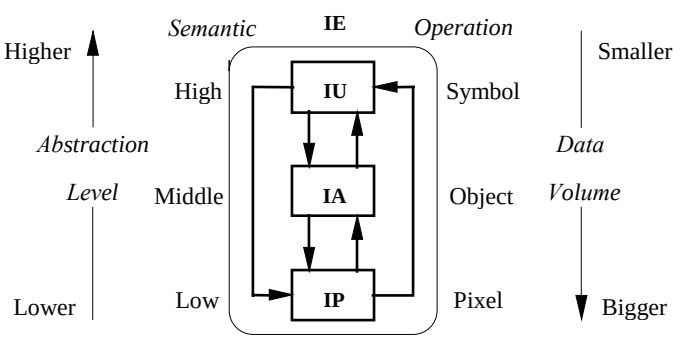
Another survey series, but on IE, has been started since 1996 (Zhang, 1996a, 1996b, 1996c, 1997, 1998, 1999, 2000a, 2001a, 2002a, 2003, 2004, 2005). The purpose of this survey work is mainly to capture the up-to-date development of IE, to make available a convenient means of literature searching facility for readers working in related areas, and to supply a useful reference for the editors of journals and potential authors of papers. This new series overcame the weakness of the earlier mentioned one by summarizing the cited references for each year, analyzing the distributions of the selected references from various sources, and providing various statistics about the classified references in each group. This new survey series has already made consecutively for ten years. This article will present an overview of this survey series by showing the idea behind and consideration on this work as well as the comprehensive statistics obtained from this work.

BACKGROUND

Image Engineering

IE, from a perspective more oriented to technique, could be referred to as the collection of three related and partially overlapped groups of image techniques, that is, image processing (IP), image analysis (IA), and image understanding (IU). In a structural sense, IP, IA, and IU build up three interconnected layers of IE as shown in Figure 1. Each of them operates on different elements (IP's operand is pixel, IA's operand is object, and IU's operand is symbol) and works with altered semantic levels (from low at IP to high at IU).

Figure 1. Three layers of image engineering



The three layers follow a progression of increasing abstractness and of decreasing compactness from IP to IU.

IP primarily includes the acquisition, representation, compression, enhancement, restoration, and reconstruction of images. While IP is concerned with the manipulation of an image to produce another (improved) image, IA is concerned with the extraction of information from an image. Compared to IP which takes an image as input and outputs also images, IA takes also an image as input but outputs data. Here, the extracted data can be the measurement results associated with specific image properties or the representative symbols of certain object attributes. Based on IA, IU refers to a body of knowledge used in transforming this extracted data into certain commonly understood descriptions and for making subsequent decisions and actions according to the interpretation of the images.

Related Subjects

IE is a broad subject encompassing studies of mathematics, physics, biology, physiology, psychology, electrical engineering, computer science, automation, and so forth. Its advances are closely related to the development of telecommunications, biomedical engineering, remote sensing, document processing, industrial applications, etc. (Zhang, 2002b).

According to different science politics/perspectives, various terms such as computer graphics (CG), pattern recognition (PR), computer vision (CV), scene analysis (SA) (just counted as another name of CV, see Rosenfeld, 2001) *etc*., are (partially) overlapped with IP, IA, and/or IU. A diagram describing the relationship among the earlier-mentioned subjects is given in Figure 2. Images are captured from the real world and processed to furnish the basis for IA or PR. The former produces data that can be visualized by CG techniques while the latter continually classifies them into one of several categories. Results produced by both of them can be further interpreted for human beings to understand the real world. The whole process aims to make computers capable of understanding environments from visual information, which is also the purpose of CV/SA.

THE CURRENT "PICTURE" OF IMAGE ENGINEERING

What is the current "picture" of IE? Answering this question is the foremost intention of the new survey series. For such a purpose, selection of reference source and classification of references according to contents are two important factors. Also for such a purpose, three statistics made by this survey are illustrated in the following.

Classification Scheme

The classification scheme used in the bibliography series should reflect the contents of references. A classification problem can be considered as a problem of partitioning a set into subsets. An appropriate classification of references into groups and/or sub-groups should satisfy the following four conditions:

1. Every reference must be in a group.
2. All groups together could include all references.
3. The references in the same group should have some common properties.
4. The references in different groups should have certain distinguishing properties.

Taking into consideration these conditions and the status of development in the field, a complete and compact classification of the theories and techniques of IE is proposed and listed in Table 1 (Zhang, 2002b). It is easy to verify that these conditions are fulfilled by this classification.

Figure 2. Image engineering and related subjects

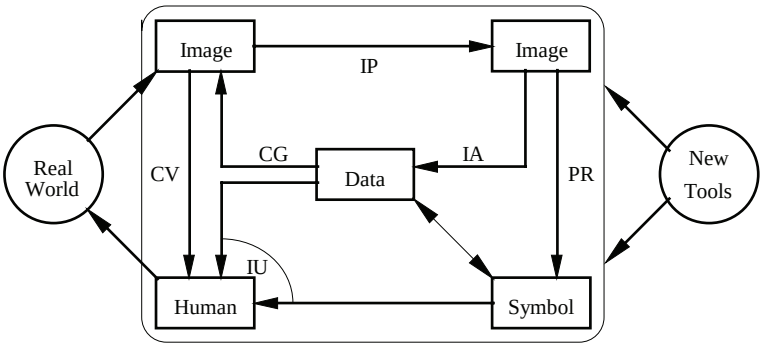


Table 1. Classification scheme of image engineering

Group	Sub-group
IP: Image Processing	P1: Image capturing and storage (including camera calibration) P2: Image reconstruction from projections P3: Filtering, transformation, enhancement, restoration P4: Image and/or video coding and standards P5: Image digital watermarking and image information hiding
IA: Image Analysis	A1: Edge detection, image segmentation A2: Representation, description, measurement (bi-level image) A3: Analysis of color, shape, texture, position, motion, etc. A4: (2-D) object recognition, extraction, tracking, classification A5: Human face and organ detection and location
IU: Image Understanding	U1: (Sequential, Volumetric) image registration and matching U2: 3-D modeling, representation and real world recovery U3: Image interpretation and reasoning (semantic, expert system) U4: Content-based image and video retrieval
TA: Technique Applications	T1: System and hardware (fast algorithm implementation) T2: Telecommunication, television T3: Documents (texts, digits, symbols) T4: Bio-medical imaging T5: Remote sensing, surveying and mapping T6: Others

Table 2. Selected journals and their abbreviations

#	Journal	Abbreviation.
1	Acta Automatica Sinica	AAS
2	Acta Electronica Sinica	AES
3	Acta Geodactica et Cartographica Sinica	AGCS
4	Chinese Journal of Biomedical Engineering	CJBE
5	Chinese Journal of Computers	CJC
6	Chinese Journal of Stereology and Image Analysis	CJSIA
7	Computerized Tomography Theory and Applications	CTTA
8	Journal of China Institute of Communications	JCIC
9	Journal of Data Acquisition and Processing	JDAP
10	Journal of Electronic Measurement and Instrument	JEMI
11	Journal of Electronics and Information	JEI
12	Journal of Image and Graphics	JIG
13	Journal of Remote Sensing	JRS
14	Pattern Recognition and Artificial Intelligence	PRAI
15	Signal Processing	SP

Source Selection

As with any other emerging discipline, a large number of references related to IE have been published worldwide. The continued growth of the literature has already made it impractical to cover all of them in one survey (Rosenfeld, 1999). Though references have been dispersed across many resources, the most popular ones are conference proceedings, journals, and books. Considering the fast publishing rate, the conference proceedings would be ranked first followed by journals and books. Considering the comprehensiveness, the books would be ranked first followed by journals and conference proceedings. Considering the quality and coverage, journal articles would be ranked higher than that of conference proceedings and books. Combining all these considerations, journals would be the best choice for such a survey series.

Based on a careful selection of literature for providing an appropriate coverage in this area, 15 important journals (in the sense defined by Lin & Zhang, 1996) with high standard articles that are published in Chinese have been selected to limit the volume of references to a manageable size. All of the papers in these journals have titles, abstracts, and keywords in English. The list of journals is given in Table 2.

Summary Over Years

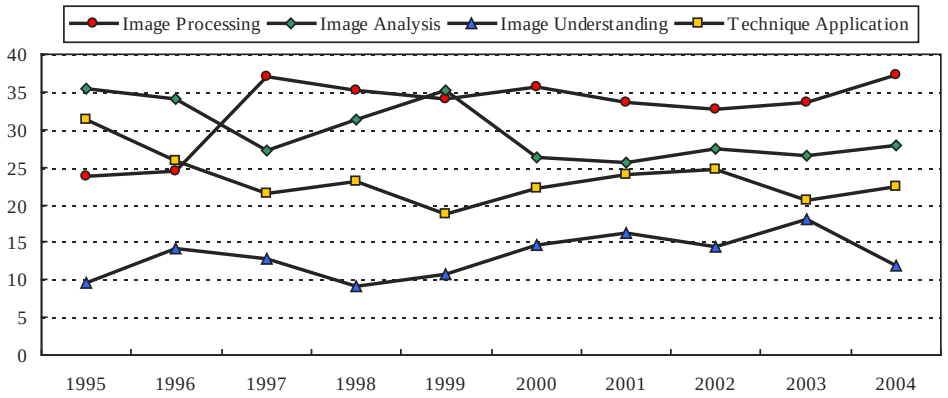
The first statistic made from this survey is a summary of the number of publications in the last ten years, as shown in Table 3. As in a survey of papers, the references have been classified into five groups: IP, IA, IU, TA and Survey. In Table 3, the total number of papers published in the selected journals (#T), the number of papers selected for survey as they are related to IE (#S), and the selection ratio (SR), which equals to #S/#T, for each year have been provided. In addition, the paper numbers for five groups (and their percentages in the year) are also listed.

Some interesting points can be noted from Table 3:

Table 3. Summary over the last 10 years

Year	#T	#S	SR	IP	IA	IU	TA	Survey
1995	997	147	14.74	35(23.8%)	52(35.4%)	14(9.52%)	46(31.3%)	
1996	1205	212	17.59	52(24.5%)	72(34.0%)	30(14.2%)	55(25.9%)	3(1.42%)
1997	1438	280	19.47	104(37.1%)	76(27.1%)	36(12.9%)	60(21.4%)	4(1.43%)
1998	1477	306	20.72	108(35.3%)	96(31.4%)	28(9.15%)	71(23.2%)	3(0.98%)
1999	2048	388	18.95	132(34.0%)	137(35.3%)	42(10.8%)	73(18.8%)	4(1.03%)
2000	2117	464	21.92	165(35.6%)	122(26.3%)	68(14.7%)	103(22.2%)	6(1.29%)
2001	2297	481	20.94	161(33.5%)	123(25.6%)	78(16.2%)	115(23.9%)	4(0.83%)
2002	2426	545	22.46	178(32.7%)	150(27.5%)	77(14.3%)	135(24.8%)	5(0.92%)
2003	2341	577	24.65	194(33.6%)	153(26.5%)	104(18.0%)	119(20.6%)	7(1.21%)
2004	2473	632	25.60	235(37.2%)	176(27.8%)	76(12.0%)	142(22.5%)	3(0.47%)
Total	18819	4032		1364(33.8%)	1157(28.7%)	553(13.7%)	919(22.8%)	39(9.67%)
Average	1882	403	21.44	136.4	115.7	55.3	91.9	3.9

Figure 3. Number variation of four groups in selected publications for last 10 years



1. This work has a quite large scale with nearly 19,000 papers involved and more than 4,000 papers selected and classified.
2. IE is an important topic for electronic engineering, computer science, and automation. The average SR is more than 1/5, which is remarkable considering the wide coverage of these journals.
3. IE publication evolves quite steadily. From Table 3, #S is increasing every year, and its value in 2004 is four times bigger than 10 years' ago. It is also noted that SRs in the recent three years are not only rising but also among the highest in ten years with SR > 1/4 for 2004.
4. The number of publications for IP and IA constitute 2/3 of the total number of IE publications. This shows the current research focus of IE. In contrast, research work on IU needs to be promoted.
5. The growing rates of publications for IP, IA, IU, and TA are comparable. To make it clear, Figure 3 shows the numbers of publications for these four groups graphically. The four curves in Figure 1 run quite smoothly and have not intercrossed in the last five years.

Distribution Analysis

The second statistic is the summary over the different journals (see Table 2), and the results are shown in Table 4. In Table 4, #I is the number of surveyed issues; #T and #S are now the total number of papers and the number of survey-selected papers, respectively. We also give the rank of the different journals according to their SR (selection ratio), and the rank of the different journals according to TR (total ratio, i.e., over all 15 journals). In Table 4, SR gives the relative frequency of IE publications in a journal. This relative frequency brings a measure of the probability of obtaining useful information from that journal. TR presents the relative contribution of each journal to IE publication and supplies a figure of importance of that journal among 15 journals. According to these rankings, readers could selectively scan the journal and judge the value of each journal.

From Table 4, the following observations can be made:

1. SR of a journal gives the probability of obtaining useful information from this journal. JIG has the highest SR among the 15 journals, and therefore, it should be checked frequently.
2. TR of a journal shows the contribution of this journal to IE publication. JIG has the highest TR among the 15 journals (and much higher than all competitors); therefore, it is evident that this journal offers a focused location for researchers in this field.
3. According to the scatter rule (Ding, 1993), most research papers of one discipline will be concentrated in a few number of journals, and other papers will be dispersed in a large number of journals. The leading five journals: JIG, AES, CJC, PRAI, and JEI, contained more than twice the number of IE papers compared to the other 10 journals.

Table 4. Summary over 15 journals

Journal	#I	#T	#S	SR (Rank)	TR (Rank)
AAS	60	1280	132	10.31% (14)	3.27% (10)
AES	88	3474	504	14.51% (11)	12.5% (2)
AGCS	40	592	95	16.04% (9)	2.36% (13)
CJBE	48	822	121	14.72% (10)	3.00% (11)
CJC	120	1946	319	16.39% (7)	7.91% (3)
CJSIA	36	488	111	22.75% (5)	2.75% (12)
CTTA	40	460	74	16.09% (8)	1.84% (14)
JCIC	106	2017	209	10.36% (13)	5.18% (8)
JDAP	40	971	211	21.73% (6)	5.23% (7)
JEI	84	2095	284	13.56% (12)	7.04% (5)
JEMI	40	565	57	10.09% (15)	1.41% (15)
JIG	102	1704	1223	71.77% (1)	30.3% (1)
JRS	48	639	155	24.26% (4)	3.84% (9)
PRAI	41	823	291	35.36% (2)	7.22% (4)
SP	48	943	246	26.09% (3)	6.10% (6)
Summary	941	18819	4032		

Detailed Classification Statistics

The third statistic is a detailed classification of IE publications in each sub-group and for each journal. The results are listed in Table 5.

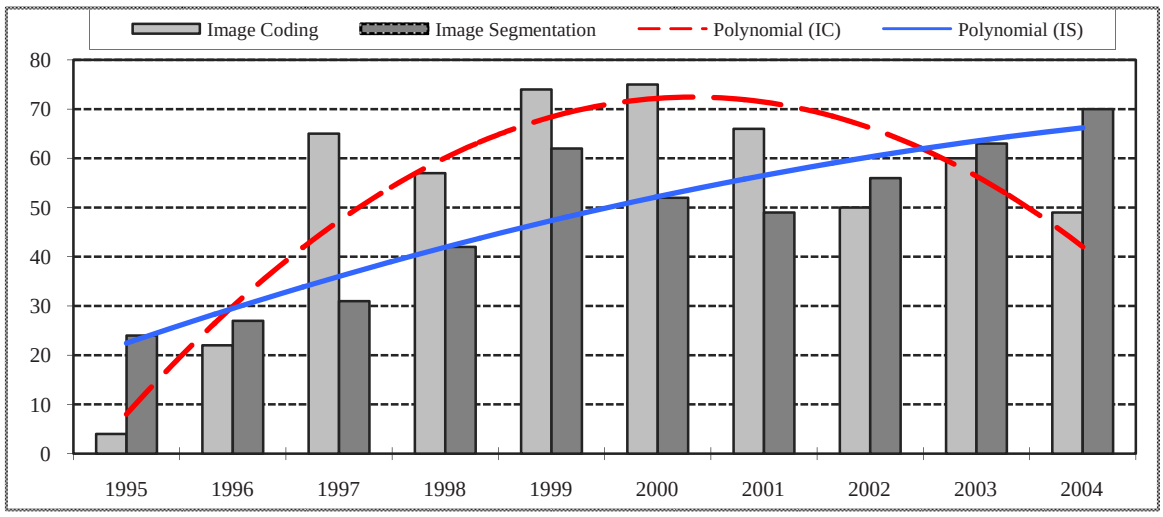
Many commentaries could be made on Table 5; however, we only point out three important observations:

1. From the number of publications in different sub-groups, it seems that image compression (P4), image segmentation (A1), and object extraction (A4) are the most important research topics in all these years. Note that classes A1 and A4 are closely related but different. A1 is concentrated for separating an image into meaningful parts while A4 is more related to direct detection with object model; the former is more like an unsupervised task and the latter is more supervised.
2. The detailed classification shows that different journals have their different emphasis; some of them cover different sub-groups of IE (for example, AES, CJC, JDIP, JEI, JIG, PRAI) evenly, while some of them are more specialized in certain sub-groups of IE (for example, CJBE for T4, CTTA for P2, JCIC for P4 and T2, and JRS for T5). That information would be useful for potential authors.
3. The top two sub-groups are P4 and A1, respectively. Both of them contain about 1/8 of the total publications and thus indicate two hot research areas in IE. How-

Table 5. Detailed classification of references

Journal	P1	P2	P3	P4	P5	A1	A2	A3	A4	A5	U1	U2	U3	U4	T1	T2	T3	T4	T5	T6	S1
AAS	16		7	7	5	17	5	1	15	11	10	17	3	3		1	7	1		6	
AES	22	16	40	97	41	69	16	14	24	21	36	9	3	17	25	22	15	6	7	4	
AGCS	10		6	5		4	6	9	7		14	3	3		1				20	6	1
CJBE	6	14	5	2		13	1	1		2	8	6			5			58			
CJC	22	4	17	32	23	40	21	9	16	23	21	21	7	17	5	6	23	3	2	6	1
CJSIA	4	4	7	5	3	16	12	4	5	1	6			2	6		1	20	2	11	2
CTTA	7	42	3			1	1				1	1			2			9		7	
JCIC	2	1	16	66	34	12	3	4	2	1	4			5	13	40	3		1	1	1
JDAP	11	2	21	31	5	22	5	9	15	8	12	1	1	3	19	12	10	7	6	10	1
JEI	27	3	26	43	24	33	7	8	18	9	15	6		7	13	10	3	3	19	9	1
JEMI	8		4	7		2	1	2	5		1				13	3	1	1		9	
JIG	30	17	117	185	37	152	50	48	86	31	94	53	12	51	31	24	27	41	43	64	30
JRS	18	2	19	2		8	1	7	8		12		2	1	2				67	6	
PRAI	7	3	8	11	5	51	22	11	40	15	25	12	2	10	1	1	50	5	2	8	2
SP	4	8	29	45	14	37	6	5	18	10	10	2		4	21	9	6	9	5	4	
Summary	194	116	325	538	191	477	157	132	259	132	269	131	33	120	157	128	146	163	174	151	39

Figure 4. Comparisons of publication quantities for image coding and image segmentation in last 10 years



ever, a detailed comparison for each year, as shown in Figure 4, illustrates that they have had quite different developing trends. Image coding had been progressed mostly from 1997 to 2001 and decreased since than. This can be seen clearly by the polynomial (IC) curve (which is the polynomial approximation of the image coding curve for the last 10 years) in Figure 4. On the other side, image segmentation is progressing steadily all these years.

FUTURE TRENDS

The field of IE has changed enormously in recent years. Many techniques have been developed, exploited, or applied only in the last decade. We now see techniques for IE being implemented and used on a scale few would have predicted a decade ago. It is also likely that these techniques will find many new applications in the future.

Viewing the prospective of IE, the work for survey on IE could also be pushed deeply, at least, in two ways. First is that since this survey provides an up-to-date picture regarding IE and its research advance, so further research could be advanced and promoted in appropriate directions. Second is that according to the principles and methods of bibliometrics, a systematic investigation of the factors of the articles indexed in the survey series could be made. This can include the number of authors, the author productivity, the number of collaborative publications, the average number of authors per paper, the active author group, and the author variation ratio, and so forth. Some preliminary works have been performed (Zhang & Li, 2000b, 2001b); an up-to-date and completed version is in preparation. Such a work would reveal the level, status, and alteration of researchers in IE, as well as provide useful information for summarizing the development, progress, trends, and application areas of IE.

CONCLUSION

This article shows an overview of a survey series on IE made in the last 10 years. The idea behind and consideration on this survey, as well as a thorough summary of obtained statistics are illustrated and discussed. All these provide much of useful information regarding the 10 years' tendency of fast progresses of IE in China and worldwide.

Such a work not only provides a convenient means for literature searching in IE but also presents a detailed picture of hot research topics in the field. Moreover, it may be useful for publishers who want to quickly capture the general trends of development in IE and for potential authors who wish to disseminate widely their research results in the associated communities.

ACKNOWLEDGMENT

This work has been supported by the National Natural Science Foundation under Grants NNSF-60573148 and the Ministry of Education under Grants SRFDP-20050003013.

REFERENCES

Ding, X.D. (1993). *Fundamentals of literature metrology.* Beijing University Publishers.

Lin, B.D., & Zhang, Q.S. (1996). *A guide to the core Journals of China.* Beijing University Publishers.

Rosenfeld, A. (1999). Image analysis and computer vision: 1998. *CVIU*, *74*(1), 36-95.

Rosenfeld, A. (2000a). Image analysis and computer vision: 1999. *CVIU*, *78*(2), 222-302.

Rosenfeld, A. (2000b). Classifying the literature related to computer vision and image analysis. *CVIU, 79*(2), 308-323.

Rosenfeld, A. (2001). From image analysis to computer vision: An annotated bibliography, 1955-1979. *CVIU*, *84*(2), 298-324.

Zhang, Y.J. (1996a). Image engineering in China: 1995. *Journal of Image and Graphics*, *1*(1), 78-83.

Zhang, Y.J. (1996b). Image engineering in China: 1995 (supplement). *Journal of Image and Graphics*, *1*(2), 170-174.

Zhang, Y.J. (1996c). Image engineering and bibliography in China. In *Technical Digest of ISIST'96* (pp. 158-160).

Zhang, Y.J. (1997). Image engineering in China: 1996. *Journal of Image and Graphics*, *2*(5), 336-344.

Zhang, Y.J. (1998). Image engineering in China: 1997. *Journal of Image and Graphics*, *3*(5), 404-414.

Zhang, Y.J. (1999). Image engineering in China: 1998. *Journal of Image and Graphics*, *4*(5), 427-438.

Zhang, Y.J. (2000a). Image engineering in China: 1999. *Journal of Image and Graphics*, *5A*(5), 359-373.

Zhang, Y.J. (2001a). Image engineering in China: 2000. *Journal of Image and Graphics*, *6A*(5), 409-424.

Zhang, Y.J. (2002a). Image engineering in China: 2001. *Journal of Image and Graphics*, *7A*(5), 417-433.

Zhang, Y.J. (2002b). Image engineering and related publications. *International Journal of Image and Graphics*, *2*(3), 441-452.

Zhang, Y.J. (2003). Image engineering in China: 2002. *Journal of Image and Graphics*, *8A*(5), 481-498.

Zhang, Y.J. (2004). Image engineering in China: 2003. *Journal of Image and Graphics*, *9*(5), 513-531.

Zhang, Y.J. (2005). Image engineering in China: 2004. *Journal of Image and Graphics*, *10*(5), 537-560.

Zhang, Y.J., & Li, R. (2000b). Statistical analysis on the articles and authors of "Journal of Image and Graphics". *Journal of Image and Graphics*, *5A*(1), 6-10.

Zhang, Y.J., & LI, R. (2001b). Statistical analysis on the authors of papers cited in the survey series "Image engineering in China". *Journal of Image and Graphics*, *6A*(1), 1-5.

KEY TERMS

Image: An entity that was captured by some visual systems in looking at the real world and that can be sensed to produce perception. It is a representation, likeness, or imitation of an object or thing, a vivid or graphic description, something introduced to represent something else.

Image Analysis: One of three layers of image engineering, which is concerned with the extraction of information (by meaningful measurements with descriptive parameters) from an image (especially from interesting objects).

Image Coding: A process for representing an image with some other representations in view of reducing data for storage and/or transmission of this image.

Image Engineering: An integrated discipline/subject comprising the study of all the different branches of image and video techniques.

Image Processing: One of three layers of image engineering, which encompasses processes whose inputs and outputs are both images, with the outputs being improved version of inputs.

Image Segmentation: A process consists of subdividing an image into its constituent parts and extracting these parts of interest (objects) from the image.

Image Understanding: One of three layers of image engineering, which transforms data extracted from images into certain commonly understood descriptions, and makes subsequent decisions and actions according to the interpretation of the images.

Supporting E-Commerce Strategy through Web Initiatives

Ron Craig
Wilfrid Laurier University, Canada

INTRODUCTION

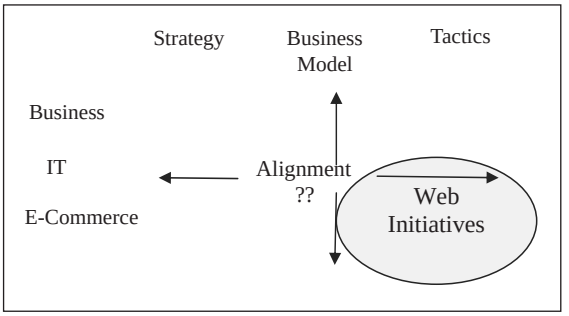
Our understanding of "the Web" and its e-commerce (EC) potential has grown rapidly during the past decade. While e-commerce has matured and is now mainstream, there continue to be opportunities to innovate as technology improves, the public is increasingly comfortable with and dependent up the e-approach, and new or enhanced applications appear. While historical roots of the Web go back several decades, it was only in the last two that business really started to embrace the Internet, and in the last one that commercial opportunities on the Web grew rapidly. Business use has gone from simple operational efficiencies (e-mail on the Internet, replacement of private EDI networks, etc.) to effectiveness (enhanced services, virtual products, and competitive advantage). Information and information products, available in digital form, and the ability to quickly transfer these from one party to another, have led to a paradigm shift in the way organizations operate. Many BPR (business process re-engineering) projects made use of the Web to streamline business processes and reduce or eliminate delays. Web self-service has emerged as a popular approach, with benefits for both customers and providers. Even governments have embraced the Web (e-government) for information and service delivery and interaction with citizens and businesses.

While the transition has followed the historical IT progression of automate, infomate, and transformate, the pace has been unprecedented. There have been successes and failures, with fortunes made and lost. After the dot-com boom/bust cycle, things settled down somewhat; yet the rapid pace of Web initiatives continues. At the forefront are innovators seeking competitive advantage. At the rear are laggards who can no longer ignore efficiencies provided by the Web and market requirements to be Web-enabled.

Paralleling the improvement in IT and the Internet has been a series of economic shifts including globalization, flattening of hierarchical organizations, outsourcing and off-shoring, increasing emphasis on knowledge work (contrasted with manual labor), plus growth in the service sector and information economy. IT has both hastened these economic shifts and provided a welcome means of addressing the accompanying pressures (often through EC or other Web initiatives).

To consider EC strategy and Web initiatives, one first needs to understand strategy and then extend this to the organization's business model and tactics. A firm's general business strategy includes, but is not limited to, its IT strategy (Figure 1). Similarly, EC strategy is a subset of IT strategy. Strategy should drive actions (tactics), through an appropriate business model. When strategy (business, IT, and EC) and tactics are closely aligned, and tactics are successfully executed, desirable results are obtained. Sometimes this normative view becomes reversed or otherwise changed. In the extreme, Web initiatives become the sole major focus (as was the case in the early days of the dot-com boom). However, without alignment between such tactics and the firm's strategy and business model, such an approach is either doomed to eventual failure or substantial modification.

Figure 1. Strategic alignment



In addition to commercial use of the Web, there are many non-commercial uses and non-commercial users (governments, educational institutions, medical organizations, etc.). The term e-business is often used to include both commercial and non-commercial activity on the Internet. In this article, the focus is on commercial activities (B2B and B2C). While e-government includes use of EC, governments are often driven by goals and responsibilities other than profit generation or cost reduction.

BACKGROUND: BUSINESS STRATEGY, IT STRATEGY, AND WEB INITIATIVES

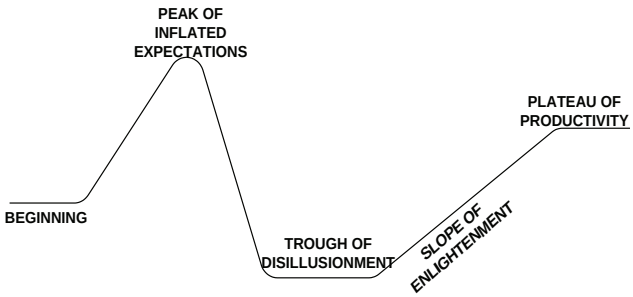
Business strategy and IT strategy have been extensively studied. The "strategic alignment model" of Henderson and Venkatraman (1993) identifies four domains of strategic

choice: business strategy, IT strategy, organizational infrastructure and processes, and IT infrastructure and processes. This model recognizes that a firm's IT operates within, and supports, a larger environment. As well, a firm's IT strategy can lead, lag, be independent of, or be aligned with a firm's business strategy. When alignment exists, there are significant payoffs (Tallon & Kraemer, 2003).

On the business strategy side, Porter provides several frameworks to guide firms in selecting their strategy and business model. His five-forces model, value chain network, and generic strategies (Porter, 1996) are useful frameworks when considering both business and IT strategies. In response to the question of whether or not the Internet renders established rules of strategy obsolete (as some had proposed), Porter answers that it makes strategy more vital than ever (Porter, 2001). He shows how the Internet has both positive and negative effects on industry structure, and identifies six principles of strategic positioning: (1) start with the right goal—superior long-term return on investment; (2) a firm's strategy enables it to deliver a value proposition, or set of benefits, that differentiates itself from competitors; (3) a firm's strategy is reflected in a distinctive value chain; (4) effective strategies require trade-offs; (5) strategy defines how all the elements of what a company does fit together; and (6) strategy involves continuity of direction. Porter (2001, p. 78) concludes, "In our quest to see how the Internet is different, we have failed to see how the Internet is the same." Today this conclusion seems almost self-evident, as our understanding of EC is much more comprehensive.

An extension to Porter's value chain is the virtual value chain (Rayport & Sviokla, 1995). Just as the physical value chain identifies the value-adding stages through which physical goods flow, the virtual value chain identifies the value-adding steps for information (gathering, organizing, selecting, synthesizing, and distributing). Firms can follow a three-stage development process: (1) visibility—improving ability to track operations more effectively, (2) mirroring—substituting virtual activities for physical, and (3) creating new customer relationships—using information to deliver value in new ways.

Figure 2. Technology hype cycle (Adapted from Gartner Group)



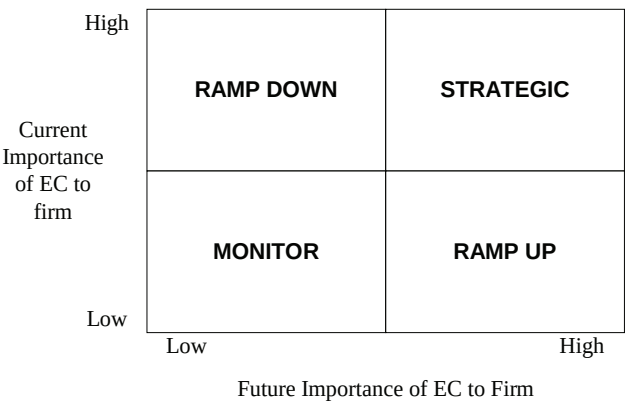
For virtual products and services, EC strategy and Web initiatives are especially important. EC usually takes advantage of both these value chains (the physical and the virtual). For example, supply chain management (SCM) initiatives have found that sharing information (virtual) about ultimate end-user demand with all members of the chain (physical) can result in significantly lower total chain costs along with improved delivery performance.

E-COMMERCE STRATEGY

During the rampant optimism of the mid to late 1990s, there seemed to be much more hype than reality concerning e-business. Statements were made that business was different now, that the Internet and Web changed everything, and that new e-business models were needed. The feeding frenzy among venture capitalists, eager to fund almost any start-up, allowed incomplete and ill-conceived concepts to be financed. It did not take long before reality took hold again, as the dot-com boom became the dot-com bust. The pendulum has now shifted from an overemphasis on "E" to a more balanced perspective on both "E" and "C." The Gartner Group Hype Cycle (Figure 2) provides a somewhat light-hearted, yet still realistic, view of this technology lifecycle. EC has gone through this cycle and emerged as an essential, productive process for most businesses.

Understanding an organization's strategic grid position (Figure 3) is critical for developing an appropriate IT and EC strategy and determining the requisite level of resources to commit. EC is not strategic to all firms, nor is all EC strategic. As Carr (2003) argues, much of IT today is a commodity-like service for many organizations, and can be managed as such (hence the popularity of IT outsourcing). Yet, EC and Web initiatives can be strategic. For firms transitioning from one quadrant to another within this grid

Figure 3. EC importance strategic grid



(i.e., changing their EC strategy) there are usually significant resource implications.

As stated earlier, an important component of corporate and e-commerce strategy is the business model used by a firm. As shown in the previous section, strategy is about making decisions and choices. For some firms, there will be much greater emphasis on the virtual side of their business; for others it will be the opposite. However, all firms need to consider the needs of their customers and the strategies (both business and IT) of their competitors, and be able to deliver required goods/services in a sustainable manner. Hence, the firm's business model must align with the strategy selected (be it EC strategy, IT strategy, or business strategy).

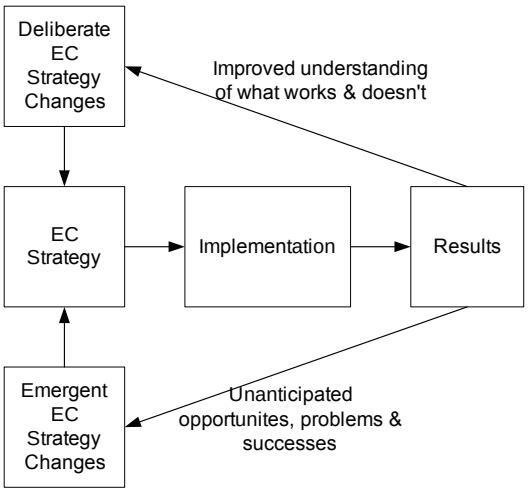
A business model is about much more than simply technology or the means of customer interaction. In the earlier days of the dot-com boom, there was an overemphasis on executing Web initiatives while ignoring the rest of the business model. Now, as the EC field matures, firms have a better understanding of the opportunities, the costs/benefits, risks, and the technology and applications to use. While technology is an important component, so are organizational characteristics (such as culture, interorganizational relationships, leadership, reward and control systems, staffing, and structure), resources (capabilities, financial, fixed assets, human, marketing, relationships, reputation and brand awareness, and technology), and managerial preferences (beliefs and values, personality and needs, job experience and context, leadership style, and political elements) (Crossan, Fry, & Killing, 2004).

An excellent discussion of business models is provided by Chesbrough and Rosenbloom (2002). They identify six functions:

- Articulates a customer value proposition
- Identifies a market segment (*who* will use the technology for *what* purpose; specifies the revenue generation process)
- Defines the venture's specific value chain structure
- Estimates the cost structure and profit potential
- Describes the venture's positioning within the value network linking suppliers and customers (includes identification of potential complementors and competitors)
- Formulates the venture's competitive strategy

Magretta (2002) identifies two tests for a powerful business model—the narrative test and the numbers test. The first test requires a logical, defendable explanation of who one's customers are, what they value, and how the firm will be profitable by providing that value. The second test requires evidence of ongoing financial viability, based on realistic assumptions and a financial analysis. Online auction giant eBay passed both these tests, as did Amazon. In contrast, most online grocery models failed because of false assumptions about what customers valued, and overly optimistic estimates of marketing, technology, and delivery cost. Another retailer, Lands' End, utilized Web initiatives and other IT to successfully pioneer mass customization for apparel (Ives & Piccoli, 2003).

Figure 4. Process by which EC strategy is defined & implemented (Adapted from Christensen & Raynor, 2003)



A business model can emerge quickly (e.g., for startups) or evolve slowly (as in a mature industry). The goal is to have a sustainable (profitable) business model. To develop this, components of the model (particularly business, IT, and EC strategy) may have to be changed. Christensen and Raynor's (2003) framework for strategy development can be applied to this process (Figure 4).

One question facing firms is when to use Web initiatives. Andal-Ancion, Cartwright, and Yip (2003) identify 10 drivers of new information technology (NIT), which also apply to Web initiatives. These different drivers determine the competitive advantages of deploying NIT, and fit under three general classifications: (1) Inherent characteristics of the product or service (electronic delivery, information intensity, customizability, and aggregation effects), (2) Interactions between a company and its customers (search costs, real-time interface, and contracting risk), and (3) Interactions between a company and its partners and competitors (network effects, standardization benefits, and missing competencies).

In their paper, they apply these drivers to various industries and firms, explaining the basis for disintermediation, remediation, or network-based mediation within an industry. These drivers should also be considered when a firm develops its EC strategy and considers Web initiatives.

A particular Internet benefit comes from its ability to integrate business processes across organizations, facilitated by the sharing of information between the various partners.

This is seen in B2B exchanges (whether private or public) and portals. After a rocky start with involvement in public exchanges (which tended to overemphasize price), many firms are now participating in, or hosting, private by-invitation-only exchanges (Hoffman, Keedy, & Roberts, 2002). Once again, it was lack of understanding of a sustainable B-model that resulted in so many failures. The original plan was to use the Internet's power and reach to form a more efficient marketplace. Most exchanges failed once their venture capital financing dried up. The strongest, surviving exchanges are those that added value for all participants (appropriate alignment, per Figure 1). Examples of successful industry exchanges include Covisint LLC (automotive, healthcare), Trade-Ranger Inc (energy/chemical), and Global Healthcare Exchange LLC (healthcare).

In the area of SCM, the sharing of information allows efficiency improvements all along the chain—firms can reduce inventory levels, increase the accuracy of their forecasting, and supply parts in a timely and cost-efficient manner. With the move from EDI to the Internet, even small firms can easily participate.

When developing an EC or Web initiative, there are many things to consider. A partial list includes the source and target (business, consumer, government), the focus (internal or external or both), whether the objective is efficiency or effectiveness, go it alone vs. partnership, proactive vs. reactive approach, targeting one-time vs. ongoing customers, physical vs. virtual goods, and single good/service vs. package.

From an ROI perspective, any investment in Web and e-commerce should provide a reasonable return. Yet, traditional investment analysis techniques may not be appropriate. Kohli, Sherer, and Baron (2003) discuss the unique challenges that e-business environments pose to the measurement of IT payoff. These include the productivity paradox, level of measurement, choice of metrics, and the measurement process. The authors identify four general areas for future research (appropriate metrics, the e-business environment, technology, and business process change).

FUTURE TRENDS

There is considerable uncertainty about predicting the future, particularly with IT and EC. Yet, there are current observable trends.

On the technology side, there is the gradual convergence of technical standards/protocols (although no single standard set has emerged). A benefit of collaborative consortiums is that a single protocol can be agreed upon by all participants. Certainly, the number of technologies and software capabilities will increase, and the move toward commoditization will continue, with the vendor community battling the Open Source community. For most firms, the problem will continue to be which technology to implement for what purpose. Overall, firms will continue to have more difficulty implementing technical initiatives because of non-technical issues. Usability of Web sites will continue to improve, as firms apply human factors analysis and other design procedures to implement better Web sites (see the Nielsen Norman group at www.useit.com). Research opportunities in this area will include empirical studies of how different types of organizations are using new technology, along with the study of implementation methods and challenges, and normative studies of how organizations should decide on which technologies to implement, when, and how.

Customer trends are also noticeable. Today's customers have heightened expectations. One proven approach to meeting these is through self-service applications (a "win/win" for both customers and the firm) for which Web initiatives are ideal. Self-service applications will proliferate.

The development and use of intra-organizational applications will continue to grow. Growing use of analytics will assist firms in segmenting their customers, more quickly becoming aware of demand shifts, and providing increased value at lower cost. Retail customers will have fewer Web site security concerns and continue to become more comfortable with Web purchasing of goods and services. Research opportunitities include developing better analytics and studying the behavior of early adopters, mainstreamers, and laggards.

At the firm level, several things are noticeable. As firms have gained experience with EC, many problems (such as channel conflict) have been identified and dealt with. SCM has led in consolidating firms into value chain networks. An increased understanding of the business levers for effective e-commerce strategies is leading to better decisions and standardization on "best practices" (which results in operational efficiency, but not necessarily in competitive advantage). In addition, there will be continued digitization of business processes; recent estimates were that only 20%-25% were currently digitized (Kalakota & Robinson, 2003). Smaller firms, facing resource scarcity, are likely to continue to lag in EC initiatives (Craig, 2002).

To date, m-commerce has not had the impact on EC that was expected. Yet the sheer number of mobile devices in the hands of consumers provides an attractive potential market. Okazaki (2005) provides an overview of m-commerce research, from both a past and future perspective. Xu and Gutierrez (2006) conducted a Delphi panel to identify potential "killer applications" for m-commerce. The panel results included SMS (short message service) and a killer portfolio (a package of applications designed to meet customer needs). In addition, four factors—convenience, ease of use, trust, and ubiquity—were identified as particularly important.

CONCLUSION

Business success comes from making difficult decisions about business, IT, and EC strategy, and then implementing these successfully, using appropriate resources. While spending on IT and EC will continue, it is possible to underspend, as well as overspend. Firms need to set priorities for investments, and ensure their business strategy aligns with their IT strategy (Farrell, Terwilliger, & Webb, 2003) and EC strategy. Sometimes this will result in being an IT and EC leader within an industry segment, and sometimes it will mean staying within the mainstream group or even lagging (Carr, 2003). This is an important choice for firms and should be made consciously, rather than by default.

Web initiatives, when undertaken as part of a firm's EC strategy, are only successfully implemented when adequate resources are provided for each project, and appropriate project management processes followed. For leading edge (bleeding edge?) projects, significant contingency allowances are important, and firms must be prepared for occasional failures (one could call these "learning experiences"). Risk management is necessary for all major IT projects, including Web initiatives.

Finally, firms should not overly focus on technology (Barua, Konana, Whinston, & Yin, 2001) when it comes to EC and Web initiatives. Good, appropriate technology (not necessarily the best), successfully implemented, will bring significant benefits at a reasonable cost.

REFERENCES

Andal-Ancion, A., Cartwright, P. A., & Yip, G. S. (2003). The digital transformation of traditional businesses. *MIT Sloan Management Review, 44*(4), 34-41.

Barua, A., Konana, P., Whinston, A. B., & Yin, F. (2001). Driving e-business excellence. *MIT Sloan Management Review, 43*(1), 36-44.

Carr, N. G. (2003). IT doesn't matter. *Harvard Business Review, 81*(5), 41-49.

Chesbrough, H., & Rosenbloom, R. (2002). The role of the business model in capturing value from innovation: Evidence from Xerox Corporation's technology spin-off companies. *Industrial and Corporate Change, 11*(3), 529-555.

Christensen, C. M., & Raynor, M. E. (2003). *The innovator's solution: Creating and sustaining successful growth*. Boston: Harvard Business School Press.

Craig, R. (2002). Web initiatives & e-commerce strategy: How do Canadian manufacturing SMEs compare? In S. Burgess (Ed.), *Managing information technology in small business* (pp. 193-208). Hershey, PA: Idea Group Publishing.

Crossan, M., Fry, J., & Killing, J. (2004). *Strategic analysis and action* (6th ed.). Toronto: Prentice Hall.

Farrell, D., Terwilliger, T., & Webb, A. P. (2003). Getting IT spending right this time. *The McKinsey Quarterly, 2003*(2), 118-129.

Henderson, J. C., & Venkatraman, N. (1993). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal, 32*(1), 4-16.

Hoffman, W., Keedy, J., & Roberts, K. (2002). The unexpected return of B2B. *The McKinsey Quarterly, 2002*(3), 97-105.

Ives, B., & Piccoli, G. (2003). Custom-made apparel at Lands' End. *Communications of the Association for Information Systems, 11*(3), 79-93.

Kalakota, R., & Robinson, M. (2003). *Services blueprint: Roadmap for execution*. Reading, MA: Addison-Wesley.

Kohli, R., Sherer, S. A., & Baron, A. (2003). Editorial—IT investment payoff in e-business environments: Research issues. *Information Systems Frontiers, 5*(3), 239-247.

Magretta, J. (2002). Why business models matter. *Harvard Business Review, 80*(5), 86-91.

Okazaki, S. (2005). New perspectives on m-commerce research. *Journal of Electronic Commerce Research, 6*(3), 160-164.

Porter, M. E. (1996). What is strategy? *Harvard Business Review, 74*(6), 61-78.

Porter, M. E. (2001). Strategy and the Internet. *Harvard Business Review, 79*(3), 62-78.

Rayport, J. F., & Sviokla, J. J. (1995). Exploiting the virtual value chain. *Harvard Business Review, 73*(6), 75-85.

Tallon, P. P., & Kraemer, K. L. (2003). Investigating the relationship between strategic alignment and IT business value: The discovery of a paradox. In N. Shin (Ed.), *Creating business value with information technology: Challenges and solutions* (pp. 1-22). Hershey, PA: Idea Group Publishing.

Xu, G., & Gutierrez, J. A. (2006). An exploratory study of killer applications and critical success factors in m-commerce. *Journal of Electronic Commerce in Organizations, 4*(3), 63-79.

KEY TERMS

Business Model: A specific arrangement of organizational strategies, goals, processes, resources (technologies,

finances, people, etc.), structures, products, and services that enable a firm to successfully compete in the market place. Many EC researchers have taken a narrower view, based on organizations involved (i.e., B2B, B2C, B2G, etc.), or specific framework used (i.e., hierarchy, hub, or intermediary for e-markets). While there is not yet a consensus about what makes up a business model, the trend is away from a narrower view.

E-Business Model: That subset of the general business model that supports e-business.

E-Commerce Strategy: A subset of general business and information technology strategy, focusing on Web-based commercial opportunities. It may dominate general strategy in some firms.

Electronic Commerce (E-Commerce): Commercial activities taking place over electronic networks (primarily the Internet); e-commerce is a subset of general commerce.

Mobile Commerce (M-Commerce): With wireless access to the Web.

Protocol: A set of rules and procedures that govern transmission between the components of an electronic network.

Strategy: The determination of the basic long term goals and objectives of an organization, and the adoption of courses of action and allocation of resources necessary for achieving these goals; major components of strategy include goals, product/market focus, business system focus, and competitive premise.

Web Initiative: Any use of the World Wide Web for a specific purpose.

Web Personalization: Customizing Web content, in real time, to a specific user.

Supporting Quality of Service for Internet Multimedia Applications

Yew-Hock Ang
Nanyang Technological University, Singapore

Zhonghua Yang
Nanyang Technological University, Singapore

INTRODUCTION

The Internet has gone from near-invisibility to near-ubiquity and penetrated into every aspect of society in the past decades (Department of Commerce, 1998). The application scenarios have also changed dramatically, and now demand a more sophisticated service model from the network. In the early 1990s, there was a large-scale experiment in sending digitized voice and video across the Internet through a packet-switched infrastructure (Braden, Clark, & Shenker, 1994). These highly-visible experiments have depended upon three enabling technologies: (1) Many modern workstations now come equipped with built-in multimedia hardware, (2) IP multicasting, which was not yet generally available in commercial routers, and (3) Highly-sophisticated digital audio and video applications have been developed. It became clear from these experiments that an important technical element of the Internet is still missing: multimedia, which dominate increasing proportion of today's data traffic, are not well supported on the Internet.

BACKGROUND

The Internet, as originally conceived, offers only best-effort service provided by the connectionless datagram service of IP (Internet Protocol). Such service does not promise whatsoever the end-to-end delay of packets nor about the variation of packet delay variation, and is not suitable for carrying multimedia traffic. To support multimedia services, the Internet must guarantee its transmission delay, delay variation, and loss according to the quality of service (QoS) required by multimedia applications and the service quality level expected by the Internet users. In other words, the Internet is to be QoS-aware of different service quality requirements for supporting multimedia applications and sensitive to the characteristics that are peculiar to multimedia data.

On the other hand, traditional networks are QoS-specific networks which were designed to support fixed classes of service; such as telephone networks for time-sensitive voice communication, cable networks for error-tolerant video streaming, and data networks for non real-time error-sensitive text messaging. The future Internet, however, is expected to be QoS-aware, and provide QoS guarantees to meet a wide range of QoS requirements for supporting multicast multimedia applications. In response to these expectations, the Internet Engineering Task Force (IETF) has specified three service models for supporting QoS on the Internet; namely, the Integrated Services (IntServ), Differentiated Services (DiffServ), and Multi-Protocol Label Switching (MPLS). This article discusses the design motivation, architecture and development of these service models for supporting QoS guarantees on the Internet.

PRINCIPLES FOR SUPPORTING MULTIMEDIA QOS

To support multimedia applications, the Internet service models must address the issue of quality of service that relates to the characteristics of multimedia data. These data characteristics are: (a) media type, (b) data synchrony, and (c) data persistency. The media type identifies the multimedia as one of image, video, audio, or text. Network services should be aware of different error susceptibilities of media types. For example, video data are handled by a QoS-aware network with a higher packet loss rate or treated with higher drop precedence than for error-sensitive audio data. Data synchrony defines the temporal synchronization of real-time data, such as interframe delay in video streams. A QoS-aware network must forward real-time data packets with higher priority than for non real-time data packets. Data persistency describes the transient nature of "live" data. Live data are not stored or buffered at its source; hence they are nonpersistent and are not retransmitted when received in error.

To be sensitive to QoS requirements of multimedia applications, a QoS-aware network must be designed according to the following four service principles and implemented with a corresponding set of traffic management/handling mechanisms:

1. *Service agreement*: a call admission mechanism to allow an application to negotiate for a service agreement

to share network resources or to operate at a required service level.

2. *Service specification*: a packet marking mechanism for the user or application to specify the application data type and a classification mechanism to allow routers to distinguish between different data specifications or classes of applications.
3. *Service conformance*: a traffic conditioning mechanism for flow policing and remarking of misbehaved packets so as to force application traffic to conform to service agreements.
4. *Service commitment*: a queue management and scheduling mechanism for QoS treatment of packets, while ensuring high resource utilization, to meet service agreements, and thereby fulfilling service commitments.

These service principles specify the manner a network is made aware of the service level requirements, specifications of traffic it is expected to handle, policing policies for service agreement enforcement, and service commitment to guarantee QoS of admitted flows. Accordingly, future Internet service models must all implement these principles for QoS-awareness.

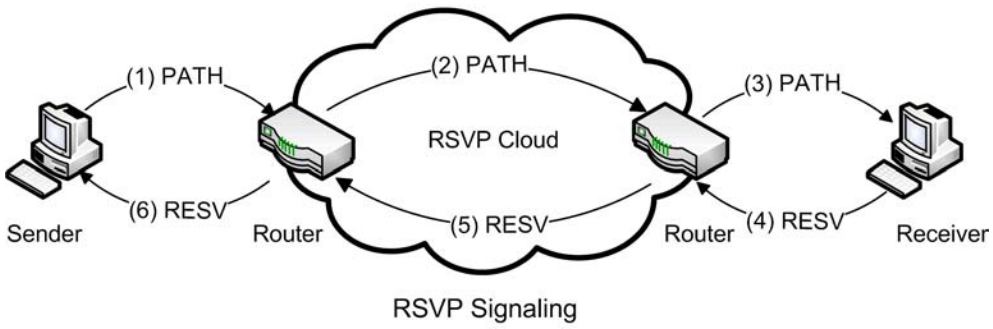
THE INTERNET INTEGRATED SERVICES (INTSERV) MODEL

The Integrated Services (IntServ) model extends the original connectionless point-to-point best-effort service of the Internet with a connection-oriented end-to-end QoS provisioned service (Braden et al., 1994). It achieves this by setting up a routed path between the endpoints with the necessary level of resources reserved along this path to guarantee end-to-end QoS on a per-flow basis (Figure 1).

Implementation of QoS support mechanisms for QoS guarantees in IntServ:

1. *Service agreement*: A call setup, using RSVP signaling (Braden, Zhang, Berson, Herzog, & Jamin, 1997) initiated by the application, requests for explicit resource reservation in routers to satisfy per-hop QoS requirements of an anticipated flow. The QoS requirements specified in a PATH message carries the *Sender_TSpec* and the *AdSpec* from the source toward the receiver. The *TSpec* contains the description of the traffic profile (traffic burst rate and average rate), while the *AdSpec* describes the properties of the data path and QoS requirements (packet delay and loss) of the sending application. The receiver responds to the service request with a RESV message for setting up the resources in the routers along the data path toward the sender.
2. *Service specification*: The RESV carries a *Flowspec* and a *Filterspec*. The *Flowspec* contains a *Receiver_TSpec* which describes the traffic profile, and *RSpec* (Reserved QoS spec) the QoS service level that each router along the routed path is expected to provide. The *Filterspec* provides the information required by the packet classifier in the router to identify packets belonging to a particular flow. Such flow-state information must be maintained in end-hosts and routers along the routed path for flow filtering and QoS packet treatment.
3. *Service conformance*: Packet classifiers in routers filter incoming flow packets based on the *Filterspec*. Packets identified as belonging to a provisioned flow are treated according to the reserved level of QoS. Flow packets not identified are treated on a best-effort basis.
4. *Service commitment*: Packet schedulers in routers handle a filtered flow according on its *TSpec* and service it such that its QoS commitments (*Rspec*) are met. The Token Bucket scheduler is typically used to service a flow carrying a *TSpec* of traffic burst rate, B bounded by the token bucket size, b; and its long-term traffic average rate, R bounded by the token arrival rate, r. The flow delay defined by b/R is guaranteed so long as the burst rate, B does not exceed the bucket size, b.

Figure 1. End-to-end QoS provisioning in Integrated Services Model



The key feature of the IntServ model is its ability to share available link capacity through explicit network resource reservation such that end-to-end QoS guarantee is maintained on a per-flow basis. However, large scale deployment of the IntServ model may be hampered in certain ways: (1) huge router memory required to maintain flow-state information which increases proportionally with the number of flows, (2) flow filtering and packet classification based on IP header multifield is computationally demanding, and (3) high implementation complexity is required as all QoS support mechanisms must be ubiquitously deployed in all routers and end hosts.

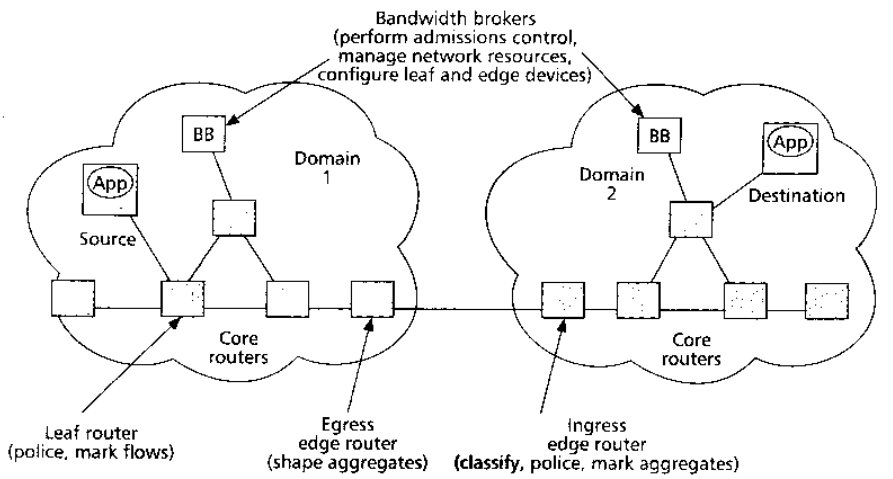
THE DIFFERENTIATED SERVICES (DIFFSERV) MODEL

The Differentiated Services (DiffServ) model provides per-class QoS service provisioning on a per-hop behavior (PHB) basis (Blake, Black, Carlson, Davies, Wang, & Weiss, 1998). Such per-class service avoided the scalability problem associated with per-flow QoS provisioning of IntServ. DiffServ allows flow packets to be marked differently by the application to receive different aggregate classes of service. It maintains edge-to-edge service commitments based on Service Level Agreement (SLA) and traffic conditioning at network edge (Figure 2).

Implementation of QoS support mechanisms for QoS guarantees in DiffServ:

1. *Service agreement*: An application initiates a SLA with the Bandwidth Broker (BB) for a service class and the traffic specifications it wants to be supported. The SLA includes a traffic conditioning agreement (TCA) which specifies the service level, traffic profile, and traffic conditioning policies. The service level is defined by a set of service performance parameters such as packet loss rate, and packet delay. The traffic profile describes the traffic peak rate and burst rate.
2. *Service specification*: Traffic entering the DiffServ domain are marked using the ToS field for differential service class treatment. ToS fields are mapped to behavioral aggregates (BAs) at leaf routers into Differentiated Service Code Points (DSCP). Packet classification into BAs using single DSCP value is far more efficient than multifield packet filtering in IntServ.
3. *Service conformance*: TCAs set up service classification and traffic conditioning policies in leaf and boundary routers for service conformance. BA flows exiting a DiffServ domain are shaped by egress routers to conform to contracted aggregate rate. Similarly, BA flows entering a DS domain are classified and conditioned to conform to TCA's traffic conditioning rules. Packets exceeding traffic profile are remarked as out-of-profile and could be potentially dropped.
4. *Service commitment*: Core routers maintain service commitments to BAs according to a set of service profiles on a per-hop behavior (PHB) basis. Service profiles define best-effort, assured forwarding, or expedited forwarding service classes and for each service class different packet drop precedence. BA reduces the number of service states needed to be maintained in the core routers; hence, a stateless core. Flow conditioning, policing, and conditioning are performed only at edge routers; hence, complexity at the network edge.

Figure 2. Per-hop behavior service in Differentiated Services Model



The DiffServ architectural framework has several advantageous features as compared to the IntServ model: (1) Stateless core allows high scalability of DiffServ domain, (2) Per-class aggregate SLA signaling is simpler than per-flow RSVP signaling, and (3) Efficient packet classification based on a single DSCP reference and forwarding based on a small number of states. However, DiffServ provides only PHB routing and no guarantee on end-to-end QoS. Traffic aggregation can potentially cause "hot spots" to form in the stateless core and hence disrupting end-to-end QoS guarantees; unless constraint based routing or traffic engineering has been used to even out "hot spots" and optimize resource utilization.

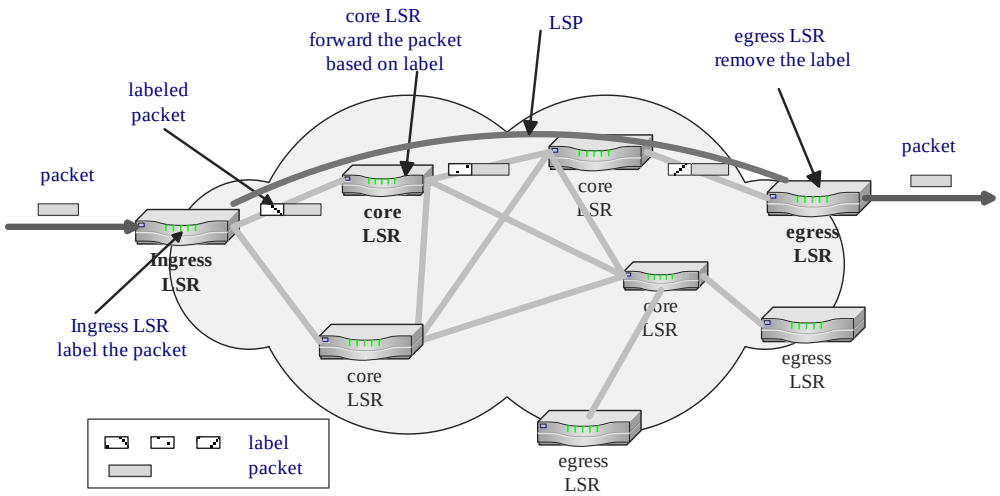
MULTI-PROTOCOL LABEL SWITCHING (MPLS)

Multi-Protocol Label Switching (MPLS) implements separate service control and forwarding functions (Rosen, Viswanathan, & Callon, 2001). Separate control allows the use of explicit routing or traffic engineering for QoS provisioning of label switched paths (LSPs) to achieve even distribution of traffic and optimized resource utilization. A LSP is end-to-end QoS provisioned and provides a forwarding equivalent class (FEC) service. Packets requiring the same QoS treatment are forwarded on a same FEC and switched along a LSP using label swapping (Figure 3).

Implementation of QoS support mechanisms for QoS guarantees in MPLS:

1. *Service agreement*: Explicit LSPs with labels are set up based on RSVP-TE (RSVP with Traffic Engineering extension) signaling (Awduche, Berger, Gan, Li, Srinivasan, & Swallow, 2001) or Label Distribution Protocol (LDP) (Andersson, Doolan, Feldman, Fredette, & Thomas, 2001). LSPs are QoS provisioned to support Forward Equivalent Classes (FECs) of service and are load balanced for high resource utilization. Label switched routers (LSRs) set up forwarding tables for packet forwarding using label swapping. An existing LSP of lower holding priority may be torn down for a new LSP with a higher setup priority to be established for a new FEC.
2. *Service specification*: An application flow is classified once at an ingress LSR into a particular FEC and assigned a label that binds the FEC to a QoS provisioned LSP. All packets in a FEC flow are treated in the same manner by intermediate LSRs along the routed LSP. Packet classification and forwarding uses a single-field MPLS label rather than IP header multifield filtering and longest-prefix address matching in IP routing. The experimental Class of Service (CoS) field in the MPLS header can also be used for microflow packet marking to provide PHB service integration with DiffServ. The CoS field may also be referred to as an EXP field and may be used for QoS bridging between MPLS and IEEE Layer-2 service (ANSI/IEEE Standard 802.1Q, 1988).
3. *Service conformance*: Ingress LSRs condition input flows to conform to FEC traffic specifications. Traffic conditioning in MPLS is done once at the ingress LSRs as compared to per-hop conditioning in IntServ.
4. *Service commitment*: QoS provisioned LSPs ensure end-to-end QoS guarantees of FEC and explicit routing of LSPs ensures load-balancing of FEC traffic for high resource utilization.

Figure 3. End-to-end packet forwarding in multiprotocol label switching



The MPLS architectural framework achieves several advantages over the IntServ and DiffServ service models: (1) Separation of service control and forwarding function allows efficient packet forwarding on explicit routed QoS provisioned LSPs, (2) Efficient flow filtering and packet forwarding using fixed-length label swapping, instead of multifield processing in IntServ, (3) Per-path QoS provisioning and end-to-end service guarantee; instead of per-hop service provisioning in DiffServ, and (4) Traffic engineered LSPs provides load-balancing of traffic and avoids network congestion.

THE FUTURE TRENDS

The IntServ, DiffServ, and MPLS, as described above, represent a generic class of service models for QoS support on the Internet. The intrinsic services they provide to different types of multimedia applications are characterized by the manner in which their service control and packet forwarding mechanisms are implemented (Table 1). These services provide QoS guarantees that match the QoS requirements of multimedia applications. For example:

- IntServ provides per-flow end-to-end service guarantee that matches QoS requirements of a dedicated flow of real-time audio stream.
- DiffServ provides per-class behavior aggregate service that matches different QoS requirements of intrastream class aggregates of I-, B-, and P-frames in a MPEG video stream. In such cases, I-frames are treated with the higher class of service and B-frames with the lowest class of service.
- MPLS provides per-path service guarantee that matches QoS requirements of a multiplexed audio/video stream. In this case, the EXP field in the MPLS header can be used to differentiate per-hop Class of Service (CoS) packet treatments that matches the distinct QoS requirements of audio and video flows in the multiplexed stream. Flow packets of the same EXP value are treated as a microflow with particular drop precedence.

Recently, research efforts have been directed at developing a framework for internetworking among the different service architectures to provide a truly integrated multiservice IP network. Such a framework includes the definitions of a common set of QoS parameters and a mechanism to map those QoS parameters across different service domains, which generally were composed of chains of multiple domains independently administered by different ISPs. A framework which provides end-to-end QoS in an IntServ-over-DiffServ architecture was proposed by extending the concept of PHB and treating each DiffServ domain as a Per-Domain Behavior (PDB) and a chains of DiffServ domains as a Per-Region Behavior (PRB), such that the QoS parameters of each of these behavioral domain/region can be externally defined and mapped between the user QoS requirements and the SLAs (Mammeri, 2005).

To facilitate the exchange and setting up of service agreements between the two domains, a set of Session Initiation Protocol (SIP) based description messages was proposed for establishing service agreements among end-to-end user terminals, policy servers, and edge routers. For IntServ-DiffServ interworking, the SIP descriptors declare the user traffic and QoS demands according to the SLA parameters and set the access control policies for traffic conditioning based on the status of resource reservation to guarantee seamless QoS connections (Cho, Shin, & Yoo, 2006). In fact, end-to-end QoS connections may also be guaranteed over any selected path that satisfies the user QoS constraint in the DiffServ domains by ensuring the residual resources in all servers along the selected path have sufficient capacity for the user request (Hsu, Tung, & Wu, 2007).

Interworking between per-hop DiffServ and per-path MPLS services can be implemented by mapping DSCP values onto specific QoS provisioned LSPs. The IETF, RFC

Table 1. Implementations of service control and packet forwarding functions in IntServ, DiffServ and MPLS service architectures

Service Model	Service Control	Packet Forwarding	Application Service/Type
IntServ	▪ Per-flow resource reservation ▪ End-to-end service	▪ flow filtering ▪ per-flow basis	▪ per-flow end-to-end QoS guarantee ▪ dedicated audio stream
DiffServ	▪ Per-class aggregates ▪ Per-hop behavior service	▪ packet classification ▪ per-class basis	▪ per-class QoS behavior aggregates ▪ MPEG video streams
MPLS	▪ Per-path provisioning ▪ Label swapping	▪ flow classification ▪ per-flow basis ▪ per-class basis	▪ per-path QoS guarantee ▪ per-hop CoS treatment ▪ Multiplexed streams

3270 (Wu, Davie, Davari, Vaananen, Krishnan, Cheval, & Heinanen, 2002) described two types of LSPs. An E-LSP which defines an EXP-Inferred per-hop scheduling class LSP, and a L-LSP which defines a normal Label-Inferred per-path scheduling class LSP and its PHB is determined from both the MPLS label and the EXP field. The E-LSP is useful for servicing microflows of I-, B-, and P-frames within a MPEG flow. For example, the DSCP values which define the per-hop aggregate class of services in DiffServ can be mapped to the EXP (or CoS) bits, and in turn mapped to the 3-bit user priority field in the Layer-2 MPLS (IEEE 802.1Q) frame to facilitate per-hop differential packet treatments of a FEC flow.

As QoS support on the Internet is becoming a reality there will be increasingly more demanding multimedia applications which require multicast services, such as in multimedia desktop collaboration, video conferencing, and network gaming. It was observed that further research work on multicast technologies were needed to develop multicast internetworking of IntServ, DiffServ, and MPLS into a practical solution (Agarwal, & Wang, 2003). In fact, practical developments to provide widespread interworking of service models have already begun at a Public Interoperability Event held at the MPLS World Congress 2006 on the interoperability of MPLS internetworked multicast services across many different MPLS vendors' equipments (Ganbar, Morin, Rossenhoevel, & Schrenk, 2006).

We believe that while cross domain QoS mappings are not yet widespread on the Internet, IntServ and DiffServ will continue to be the preferred service models for intranets and access networks implementations. However, as MPLS advances along with new protocols defined by the IETF, more devices will begin to include MPLS stacks. This will in turn push MPLS deployments closer to the edge and MPLS being adopted for access networks as well and as an integration model for widespread interworking of QoS-aware networks for supporting converged network services.

CONCLUSION

The Internet has evolved from a provider of the simple TCP/IP best effort service to an emerging QoS-aware network for supporting converged network services. The development of different service models, such as the IntServ, DiffServ, and MPLS forms the building blocks for a truly integrated multiservice IP network for supporting multicast multimedia applications over the Internet. These service models are characterized by a set of four common principles for supporting QoS guarantees and the implementation of their QoS support mechanisms led to unique characters of their service architectures for supporting different application services. With significant work and progress already completed over the last decade on the development of service architectures and network interoperability, it is anticipated that a truly integrated multiservice Internet may be realized sooner than later. Perhaps the MPLS architecture could well be the service model of choice because of its flexibility of service interoperability, end-to-end QoS guarantee, and high resource utilization.

REFERENCES

Agarwal, A., & Wang, K.B. (2003). Supporting quality of service in IP multicast networks. *Computer Communications, 26*, 1533-1540.

Andersson, L., Doolan, P., Feldman, N., Fredette, A., & Thomas, B. (2001, January). *LDP specification*. IETF, RFC 3036.

ANSI/IEEE Standard 802.1Q. (1998). *IEEE standards for local and metropolitan area networks: Virtual bridged local area networks*.

Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., & Swallow, G. (2001, December). *RSVP-TE: Extensions to RSVP for LSP tunnels*. IETF, RFC 3029.

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998, December). *An architecture for differentiated services*. IETF, RFC 2475.

Braden, R., Clark, D., & Shenker, S. (1994, June). *Integrated services in the Internet architecture: An overview*. IETF RFC 1633.

Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997, September). *Resource ReSerVation Protocol (RSVP), version 1 functional specification*. IETF, RFC 2205.

Cho, E-H., Shin, K-S., & Yoo, S-J. (2006). SIP-based QoS support architecture and session management in a combined IntServ and DiffServ networks. *Computer Communications, 29*, 2996-3009.

Department of Commerce. (1998, April). *The emerging digital economy*. United States.

Ganbar, J., Morin, J., Rossenhoevel, C., & Schrenk, G. (2006). Converged network services using MPLS. In *Proceedings of the MPLS World Congress 2006, Public Interoperability Event*, Paris, (pp. 1-11).

Hsu, W-H., Tung, M-C., & Wu, L-Y. (2007). An integrated end-to-end QoS anycast routing on DiffServ networks. *Computer Communications, 30*, 1406-1418.

Mammeri, Z. (2005). Framework for parameter mapping

to provide end-to-end QoS guarantees in IntServ/DiffServ architectures. *Computer Communications, 28*, 1074-1092.

Rosen, E., Viswanathan, A., & Callon, R. (2001, January). *Multiprotocol label switching architecture*. IETF, RFC 3031.

Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., & Heinanen, J. (2002, May). *Multiprotocol label switching (MPLS) support of differentiated services.* IETF, RFC 3270.

KEY TERMS

Behavior Aggregate (BA): A collection of packets with the same DS codepoint crossing a link in a particular direction.

DSCP: A specific value of the DiffServ codepoint portion of DS field (same as in IPv4 header TOS field or the IPv6 Traffic Class field) used to select a PHB.

Forwarding Equivalence Class (FEC): A group of IP packets which are forwarded in the same manner (e.g., over the same path, with the same forwarding treatment).

Label Switched Path (LSP): The path through one or more LSRs at one level of the hierarchy followed by a packets in a particular FEC.

MPLS Label: A label which is carried in a packet header, and which represents the packet's FEC.

Multifield: IP header multifield defines the 5-tuplet fields in the IP header, comprising of the source and destination addresses, source and destination port numbers, and the protocol type.

Per-Hop-Behavior (PHB): The externally observable forwarding behavior applied at a DS-compliant node to a DS behavior aggregate.

QoS: Quality of service refers to the suitability of a packet delivery service for the needs of a particular application, as defined by parameters such as achieved bandwidth, packet delay, and packet loss rates.

RSpec: A Service Request Specification is a specification of the quality of service a flow desires from a network element. The service request specification is highly specific to a particular service; for example, it might contain information about bandwidth allocated to the flow, maximum delays, or packet loss rate.

RSVP: The Internet standard protocol: Resource ReSerVation Protocol (RSVP).

Service Level Agreement (SLA): A service contract between a customer and a service provider that specifies the forwarding service a customer should receive.

Service Model: Consists of a set of service commitments, in response to a service request the network commits to deliver some service.

Token Bucket: A particular form of traffic specification consisting of a "token rate" r and a "bucket size" b. Essentially, the r parameter specifies the continually sustainable data rate, while the b parameter specifies the extent to which the data rate can exceed the sustainable level for short periods of time.

TSpec: Traffic Specification containing descriptions of traffic characteristic such as packet arrival peak rate and burst rate.

Traffic Conditioning Agreement (TCA): An agreement specifying classifier rules and any corresponding traffic profiles and metering, marking, discarding or shaping rules which are to apply to the traffic streams.

Supporting Real-Time Services in Mobile Ad-Hoc Networks

Carlos Tavares Calafate
Technical University of Valencia, Spain

Ingrid Juliana Niño
Technical University of Valencia, Spain

Juan-Carlos Cano
Technical University of Valencia, Spain

Pietro Manzoni
Technical University of Valencia, Spain

INTRODUCTION

Mobile ad-hoc networks (MANETs) are well known by their flexibility and usefulness, being an ideal technology to support ubiquitous computing environments. Such environments are expected to support a plethora of applications, including real-time video and voice communications.

In terms of applications, this technology can be used whenever there is a lack of infrastructure for support, which typically occurs in rescue missions, areas affected by natural disasters, remote areas, war scenarios, and also in the underground. The use of real-time voice and video communications could allow, for example, firemen rescue teams to communicate seamlessly and for the head officer to remotely supervise their activity using different video channels.

The deployment of real-time services over mobile ad-hoc networks requires QoS (quality of service) support at different network layers. QoS support is understood as the network ability to offer some guarantees about the traffic being delivered. Within the scope of QoS we often define performance in terms of availability (uptime), bandwidth (throughput), latency (delay), delay jitter, and error rate.

Offering QoS support in mobile ad-hoc network environments is, nevertheless, quite difficult due to the innate complexity of these networks. The problems that impact mobile ad-hoc networks can be split according to the network layer affected.

At the physical layer, frequent topology changes—in conjunction with channel contention and unstable radio links—make real-time services support in such networks very hard to achieve (Georgiadis, Jacquet, & Mans, 2004).

At the Medium Access Control (MAC) layer, channel access is typically distributed, provoking the well-known hidden and exposed node problems, which complicate bandwidth reservation.

At the network layer, routing protocols have to deal with frequent topology changes and simultaneously discriminate among the available paths to meet QoS requirements.

At the application layer, awareness of the type of networks and technologies being used allows applications to adapt themselves according to path conditions and so improve performance.

This article discusses the aforementioned issues related to QoS challenges and solutions in mobile ad-hoc networks. It first includes some background information on the history of QoS support in computer networks. It then refers to the problematic of QoS support in mobile ad-hoc networks by referring MAC and routing layer solutions, along with QoS architectures for ad-hoc networks. To conclude the article there is reference to future trends in terms of QoS support in ad-hoc networks.

BACKGROUND

The first attempts at providing significant QoS support improvements in computer networks took place on the Internet in the early 1990s. The main problem faced by engineers was that the Internet was initially created to handle best-effort traffic alone. This means that its infrastructure was not designed considering QoS-related functionality such as resource reservation, and so all users compete for bandwidth. For this reason the Internet protocol (IP) is connectionless, requiring no set-up "signaling" for admission control.

When enhancements in terms of available bandwidth and a terminal's capabilities brought up the need for supporting new services on the Internet, preliminary evaluation studies showed that the performance of these new services was very poor due to the best-effort policy. There was, therefore, a need to enhance the Internet infrastructure in order to allow

performing resource reservations in a similar fashion to telephony networks. The RSVP protocol (Braden, Zhang, Berson, Herzog, & Jamin, 1997) was created to fulfill this need as part of the Internet's integrated services (IntServ) architecture (Braden, Clark, & Shenker, 1994). RSVP follows a receiver-based model since it is the responsibility of each receiver to choose its own level of reserved resources, initiating the reservation and keeping it active. The actual QoS control, though, occurs at the sender's end. The sender will try to establish and maintain resource reservations over a distribution tree. If a particular reservation is unsuccessful, the correspondent source is notified.

The IntServ architecture proved to be complex and required too many resources, suffering from scalability problems. So, the differentiated services (DiffServ) architecture (Blake, 1998) emerged as a more efficient alternative. In the latter, service-level agreements (SLA) are achieved between different domains. One of the main virtues of the DiffServ architecture is that it drops the traditional concept of signaling, no longer requiring the reservation of resources in all the network elements involved. The strategy consists of performing admission control on domain boundaries, and then treating them in a differentiated manner inside the domain according to packet tagging on the domain borders, which is a much faster and lightweight process.

QoS SUPPORT IN MOBILE AD-HOC NETWORKS

MANET environments differ greatly from the wired environments the DiffServ and IntServ models were created for. The difference stems not only from the new problems encountered in MANETs (mobility, collisions, variable channel conditions, etc.), but also because MANETs do not follow the client/service provider paradigm inherent to both IntServ and DiffServ models. In MANETs the network is typically formed by users that cooperate and, except in situations where there is some centralized management entity (e.g., Army), it depends on the good behavior of users and limited resource sharing. So, new proposals were presented in order to achieve reliable QoS support in MANETs.

In this section we present an overview of the different proposals available in the literature offering QoS improvements to mobile ad-hoc networks. We first introduce QoS proposals at the MAC layer. Then we refer to QoS solutions at the routing layer. Finally we refer to complete QoS architectures for MANETs.

QoS Support at the MAC Layer

Most of the MAC layer protocols for ad hoc networks are characterized by being distributed (there is no central entity regulating channel access) and contention-based (channel access is not deterministic, being that stations compete to gain access to it). These characteristics, along with the well-known hidden (Kleinrock & Tobagi, 1975) and exposed (Shukla, Chandran-Wadia, & Iyer, 2003) node problems that are prone to occur in wireless multi-hop environments, complicate the process of offering QoS support. In fact, in such wireless environments, it is impossible to offer strict QoS guarantees to users, and so statistical QoS is achieved instead.

Despite this is a novel research area, we can already find products in the market offering QoS support at the MAC layer. The most relevant technology available due to its widespread adoption is IEEE 802.11e, which is a new MAC standard proposed by the IEEE 802.11 Working Group (2005) to enhance WiFi networks with QoS support.

The IEEE 802.11e standard relies on a hybrid coordination function, HCF, which defines two medium access mechanisms: the HCF controlled channel access (HCCA) and the enhanced distributed channel access (EDCA). From these two, only EDCA applies to ad-hoc network environments, the former being reserved for access point operation.

QoS support through EDCA requires introducing different access categories (ACs) and their associated backoff entities. Contrarily to the legacy IEEE 802.11 stations, where all the packets received by the MAC layer have the same priority and are assigned to a single backoff entity, IEEE 802.11e stations have four backoff entities—one for each AC—so that packets are sorted according to their priority. Each backoff entity has an independent packet queue assigned to it, as well as a different parameter set.

Table 1. IEEE 802.11e MAC parameters for an IEEE 802.11a/g radio

Access Category	AIFSN	CWmin	CWmax	TXOPLimit (ms)
Background	7	15	1023	0
Best effort	3	15	1023	0
Video	2	7	15	3.008
Voice	2	3	7	1.504

Table 1 presents the default MAC parameter values for the different ACs introduced by IEEE 802.11e. Notice that smaller values for the AIFSN, CWmin, and CWmax parameters result in a higher priority when accessing the channel; relative to the TXOPLimit, higher values result in larger shares of capacity and, therefore, higher priority.

Works such as Romdhani, Ni, and Turletti (2003) propose enhancements to the IEEE 802.11e technology to offer relative priorities by adjusting the size of the contention window (CW) of each traffic class, taking into account both application requirements and network conditions. In the literature we can also find other works, such as Sobrinho and Krishnakumar (1999) and Sivavakeesar and Pavlou (2004), which propose alternate QoS MAC schemes designed specifically for ad hoc network environments.

QoS Support at the Routing Layer

The issue of QoS support at the routing layer is quite complex in the MANET field since different authors use different QoS metrics. Besides generic QoS metrics that typically apply to all environments—minimum bandwidth, maximum delay, maximum delay jitter, and maximum packet loss ratio—in MANETs there is another set of metrics that is also considered important. Among these we have residual link/path capacity, node buffer space, energy consumed per packet, node/route lifetime, and so forth.

Hanzo and Tafazolli (2007) survey the most relevant QoS routing solutions developed to date. That survey categorizes QoS routing solutions according to their interaction with the MAC layer. So, the authors define three different categories for QoS routing protocols in MANETs:

(1) MAC layer dependent/contention-free
(2) MAC layer dependent/contended
(3) MAC layer independent

Concerning proposals in the first category, Lin and Liu (1999) propose CCBR (channel capacity-based routing), a QoS routing protocol that includes end-to-end bandwidth calculation along with bandwidth allocation schemes. Chen and Nahrstedt (1999) propose TBR (ticket-based routing), defining a distributed QoS routing scheme that selects a network path with sufficient resources to satisfy a certain delay (or bandwidth) requirement, though not both at the same time.

An example of a routing layer proposal for the second category is CAAODV (Lei & Heinzelman, 2005), a QoS-aware routing protocol that incorporates admission control and feedback schemes to meet the QoS requirements of real-time applications by offering an estimate of available bandwidth. Also in this category we have AQOR (Xue & Ganz, 2003), a resource reservation-based routing and signaling algorithm that provides end-to-end QoS support in terms of bandwidth and delay.

An example of a routing protocol for the third category is IAR (Gupta, Jia, Tung, & Walrand, 2005), which offers throughput-constrained QoS routing based on knowledge of the interference between links.

QoS Architectures for MANETs

In the literature we find proposals focusing essentially on the MAC and routing layers, being that only a few authors have proposed complete QoS architectures for MANETs. We will describe three QoS architectures for ad-hoc network environments that we consider especially relevant: INSIGNIA (Lee, Gahng-Seop, Zhang, & Campbell, 2000), SWAN (Ahn, Campbell, Veres, & Sun, 2002), and DACME (Calafate, Oliver, Cano, Malumbres, & Manzoni, 2007).

INSIGNIA

The INSIGNIA architecture (Lee et al., 2000) is an attempt to adapt the IntServ model proposed for the Internet to MANET environments. It consists of an in-band signaling system that supports fast reservation, restoration, and adaptation algorithms. With INSIGNIA all flows require admission control, resource reservation, and maintenance at all intermediate stations between source and destination to provide end-to-end quality of service support.

One of the characteristics of INSIGNIA is that it relies on in-band commands. These are put inside the IP option field and include service mode, payload type, bandwidth indicator, and bandwidth request fields. When a reservation is being established, each node along the path checks INSIGNIA's IP option field to see if it can offer the maximum QoS requested. The destination will become aware of QoS availability on the path by obtaining the value of this field.

The INSIGNIA signaling system is designed to be lightweight in terms of the amount of bandwidth consumed for network control and to be capable of reacting to fast network dynamics such as rapid host mobility, wireless link degradation, intermittent session connectivity, and end-to-end quality of service conditions.

SWAN

SWAN (Ahn et al., 2002) is an approach to integrated services support in MANETs through a flexible signaling system. SWAN relies on plain IEEE 802.11 plus rate-control for best effort traffic. Rate control is done at every node and relies on an AIMD (additive increase/multiplicative decrease) algorithm based on feedback from the MAC layer.

The acceptance of real-time traffic is dependent on local bandwidth estimates and admission control probes. When

a new flow must be admitted into the network, the source station sends a probing request packet to assess end-to-end bandwidth availability. Each node along the path will update the bandwidth value if its locally available bandwidth is lower than the one stated on the packet. The destination node then sends a probing response packet back to the source node with the bottleneck field copied from the probing request message it received. Based on that value the source decides whether to admit the flow or not.

Mobility may cause traffic to be re-routed through new paths where there are no resources available to accommodate the new traffic. In the case of congestion, SWAN uses ECN (explicit congestion notification) so that intermediate nodes are able to mark packets.

DACME

Calafate et al. (2007) propose a QoS architecture for MANETs whose core element is DACME (distributed admission control for MANET environments). It consists of a probe-based admission control mechanism that, when combined with a QoS-enabled MAC layer (such as IEEE 802.11e) and an appropriate routing protocol—usually a reactive one such as AODV (Perkins & Royer, 1999)—conforms a full QoS framework for MANET environments offering statistical QoS guarantees.

The main element of DACME is a QoS measurement module that assesses the QoS conditions of an end-to-end path. Path conditions are measured in terms of end-to-end bandwidth availability, delay, and jitter. All measurements rely on probes between source and destination, though the strategy varies depending on what is being measured.

Admission control is based on all the values collected during the probing process. The source makes accept/deny decisions based on statistical indexes obtained from the different probes generated.

One of the main advantages of DACME's framework is that it does not impose any demand on the lower protocol layers, nor does it rely on any sort of resource reservation, thus offering soft QoS. Moreover, it operates correctly independently of the MAC or routing protocol being used, offering a great flexibility.

As a summary, in Table 2 we point out the main differences between the three architectures described in this section.

FUTURE TRENDS

Despite efforts from a great number of researchers worldwide, QoS support in mobile ad-hoc networks remains an area with much room for improvement. One of the current trends is breaking the boundaries between layers to achieve a real cross-layer architecture that is able to improve performance by allowing network protocols at any layer to gain awareness of everything that is going on underneath it. This means that, for example, the statistics gathered on the MAC layer about available channel resources and link quality towards each neighbor node are made available to upper layers (such as routing, admission control, or transport) which can help at finding better paths, better estimate end-to-end resources, and so forth. Another trend is to base QoS decisions on statistical inference instead of using hard values. Such technique avoids strict reservations, offering a better adaptability to the network dynamics.

Both techniques are expected to greatly boost the QoS performance of MANETs, and in the future we shall see several proposals in these lines of work.

CONCLUSION

Supporting QoS demanding applications in ad-hoc networks is a very challenging goal that can only be accomplished by enhancing the different network layers involved to support QoS restrictions.

At the MAC layer the most important issue to solve is how to offer prioritized channel access using distributed, contention-based algorithms.

In terms of QoS routing protocols, there was a clear

Table 2. Main differences between INSIGNIA, SWAN, and DACME architectures

Characteristic	INSIGNIA	SWAN	DACME
Relationship with existing Internet QoS models	Integrated services	Differentiated services	None
Resource reservation	Yes	No	No
Initial bandwidth estimation	Per node	Per-node	End-to-end
Support for delay/jitter constraints	No	No	Yes

evolution in the last few years, especially in the assumptions made about the MAC layer. Though initial protocols required a TDMA-based MAC, current ones are able to offer QoS support on top of CSMA/CA-based MAC layers such as IEEE 802.11. At the routing layer the most complicated issue to solve has to do with multi-constraint routing, which remains a very complex problem.

In terms of QoS architectures for MANETs, there is a trend towards solutions that minimize restrictions on the terminals and technologies used. We consider that the most relevant architectures currently available are INSIGNIA, SWAN, and DACME, where we can observe a progression towards soft QoS guarantees.

Due to the incipient nature of this research area, we expect to see novel solutions in the near future that are expected to greatly improve the QoS performance of mobile ad-hoc networks.

REFERENCES

Ahn, G-S., Campbell, A.T., Veres, A., & Sun, L. (2002). Supporting service differentiation for real-time and best effort traffic in stateless wireless ad hoc networks (SWAN). *IEEE Transactions on Mobile Computing, 1*(3), 192-207.

Blake, S. (1998). *An architecture for differentiated services.* IETF RFC 2475.

Braden, R., Clark, D., & Shenker, S. (1994). *Integrated services in the Internet architecture—an overview.* IETF RFC 1633.

Braden, R., Zhang, L., Berson, S., Herzog, S., & Jamin, S. (1997). *Resource ReSerVation Protocol (RSVP)—version 1 functional specification.* IETF RFC 2205.

Calafate, C.T., Oliver, J., Cano, J.C., Malumbres, M.P., & Manzoni, P. (2007). A distributed admission control system for MANET Environments Supporting Multipath Routing Protocols. *Elsevier Microprocessors and Microsystems Journal, 31*, 236-251.

Chen, S., & Nahrstedt, K. (1999). Distributed quality-of-service routing in ad hoc networks. *IEEE Journal on Selected Areas in Communications, 17*(8), 1488-1505.

Georgiadis, L., Jacquet, P., & Mans, B. (2004). Bandwidth reservation in multihop wireless networks: Complexity and mechanisms. *Proceedings of the International Conference on Distributed Computing Systems Workshops* (ICDCSW'04), Hachioji-Tokyo, Japan.

Gupta, R., Jia, Z., Tung, T., & Walrand, J. (2005). Interference-aware QoS routing (IQRouting) for ad-hoc networks. *Proceedings of the Global Telecommunications Conference,* St. Louis, MO.

Hanzo, L. II, & Tafazolli, R. (2007). A survey of QoS routing solutions for mobile ad-hoc networks. *IEEE Communications Surveys & Tutorials, 9*(2), 50-70.

IEEE 802.11 Working Group. (2005). *802.11e: IEEE standard for information technology—telecommunications and information exchange between systems—local and metropolitan area networks—specific requirements part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Amendment 8: Medium access control (MAC) quality of service enhancements.*

Kleinrock, L., & Tobagi, F. (1975). Packet switching in radio channels part II: The hidden terminal problem in carrier sense multiple access modes and the busy-tone solution. *IEEE Transactions on Communications, 23*(12), 1417-1433.

Lee, S.B., Gahng-Seop, A., Zhang, X., & Campbell, A.T. (2000). INSIGNIA: An IP-based quality of service framework for mobile ad hoc networks. *Journal of Parallel and Distributed Computing, 60*(4), 374-406.

Lei, C., & Heinzelman, W.B. (2005). QoS-aware routing based on bandwidth estimation for mobile ad hoc networks. *IEEE Journal on Selected Areas in Communications, 23*(3), 561-572.

Lin, C.R., & Liu, J.-S. (1999). QoS routing in ad hoc wireless networks. *IEEE Journal on Selected Areas in Communications, 17*(8), 1426-1438.

Perkins, C.E., & Royer, E.M. (1999). Ad hoc on-demand distance vector routing. *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications,* New Orleans, LA.

Romdhani, L., Ni, Q., & Turletti, T. (2003). Adaptive EDCF: Enhanced service differentiation for IEEE 802.11 wireless ad-hoc networks. *Proceedings of the Wireless Communications and Networking Conference* (vol. 2), New Orleans, LA.

Shukla, D., Chandran-Wadia, L., & Iyer, S. (2003). Mitigating the exposed node problem in IEEE 802.11 ad hoc networks. *Proceedings of the 12th International Conference on Computer Communications and Networks,* Dallas, TX.

Sivavakeesar, S., & Pavlou, G. (2004). Quality of service aware MAC based on IEEE 802.11 for multihop ad-hoc networks. *Proceedings of the IEEE Wireless Communications and Networking Conference,* Atlanta, GA.

Sobrinho, J.L., & Krishnakumar, A.S. (1999). Quality-of-service in ad hoc carrier sense multiple access networks. *IEEE Journal on Selected Areas in Communications, 17*(8), 1353-1368.

Xue, Q., & Ganz, A. (2003). Ad hoc QoS on-demand routing (AQOR) in mobile ad hoc networks. *Journal of Parallel and Distributed Computing, 63*(2), 154-165.

KEY TERMS

CSMA/CA: Stands for Carrier Sense Multiple Access with Collision Avoidance. The distributed channel access mechanism adopted by the IEEE 802.11 standard for operation.

DiffServ: Abbreviation that refers to Internet's Differentiated Services architecture.

IntServ: Abbreviation that refers to Internet's Integrated Services architecture.

Media Access Control (MAC): A communications protocol sub-layer that provides addressing and controls channel access to allow several terminals to communicate over a shared medium.

Mobile Ad-hoc NETwork (MANET): An autonomous type of network where terminals participate both as clients and routers, eliminating the need for any type of support infrastructure.

Resource ReSerVation Protocol (RSVP): A transport layer protocol designed to perform resource reservation on the Internet; one of the key elements in the integrated services architecture proposed for the Internet.

Service-Level Agreement (SLA): A formally negotiated agreement between customers and their service provider, or between service providers, which defines the expected standards in terms of services, priorities, responsibilities, guarantees, and so forth, which is usually referred to as the "level of service."

Supporting the Evaluation of Intelligent Sources

Dirk Vriens
Radboud University of Nijmegen, The Netherlands

INTRODUCTION

To survive, organizations need to produce and process information about their environment, for instance, about customers, competitors, suppliers, governments, or all kinds of socioeconomic and technological trends. The process of obtaining this information is often called competitive intelligence (cf Fleisher & Blenkhorn, 2001; Kahaner, 1997; Vriens, 2004). An important stage in the competitive intelligence process is the collection stage. In this stage, one has to determine relevant sources, access them, and retrieve data from them (cf Bernhardt, 1994; Kahaner). For each data class, many possible sources are available, and determining the right ones is often difficult. Moreover, accessing sources and retrieving data may require a lot of effort and may be problematic (cf Cook & Cook, 2000; Fuld, 1995; Kahaner, 1997). In this chapter, we present a tool for supporting the effective and efficient use of sources: the source map. In essence, a source map links data classes to sources and contains information about these links. This information indicates the adequacy of sources in terms of ease of access, ease of retrieval, and usefulness of the retrieved data. A source map can support the selection of appropriate sources and it can support the assessment of the overall adequacy of available sources.

BACKGROUND

The process of competitive intelligence is often described as a cycle of four stages (the intelligence cycle; see Kahaner, 1997; Vriens, 2004). This cycle comprises (a) the direction stage (in which the organization determines about what aspects in the environment data should be collected), (b) the collection stage (where sources are determined and data are collected), (c) the analysis stage (in which the data are analyzed to assess whether they are useful for strategic purposes), and (d) the dissemination stage (where the data are forwarded to decision makers; Bernhardt, 1994; Gilad & Gilad, 1988; Herring, 1999; Kahaner, 1997; Sammon, 1986). The collection stage is considered to be the most time-consuming stage (e.g., Chen, Chau, & Zeng, 2002) and if it is not performed carefully, many difficulties arise (e.g., too much time spent on search, collection stage leads to irrelevant data, information overload; see, for example, Cook & Cook, 2000; Chen et al.; Teo & Choo, 2001; Vriens & Philips, 1999). For successfully carrying out collection activities, knowledge about what sources contain what kind of data and knowledge about how to approach these sources (metaknowledge regarding the collection of data) would be very helpful. This chapter presents a tool to structure and deal with this metadata: the source map.

To collect data about the environment one has to

1. identify possible sources,
2. judge the value of the source (in terms of different criteria; e.g., does it contain relevant data? What are the costs of employing this source? Is it reliable?), and
3. use value judgments to select the appropriate sources.

Many authors discuss Step 1 by pointing to a variety of available sources (cf Fuld, 1995; Kahaner, 1997; Sammon, 1986). Typical sources include the Internet, online databases, sales representatives, internal or external experts, CEOs, journals, tradeshows, conferences, embassies, and so forth.

The literature treats the valuation step more implicitly. It discusses distinctions regarding sources, such as open versus closed sources, internal versus external sources, or primary versus secondary sources (Fleisher & Blenkhorn, 2001; Kahaner, 1997). These distinctions implicitly refer to criteria used in the valuation of sources. The distinction of open versus closed sources implicitly refers to, for instance, criteria such as ease in collection or relevance. The distinction of primary versus secondary sources implicitly refers to the criterion of the reliability of the data. In our view, it is possible to value sources more precisely when the valuation criteria are stated explicitly and not implicitly in the form of these distinctions.

The selection step is even more elusive in literature (and practice). This step integrates value judgments to select appropriate sources for collecting the required data. Few methods seem to be designed for source selection.

In this article, we propose a tool to structure and support the valuation and selection of sources: the source map. This tool builds on Fuld's (1995) intelligence maps and knowledge maps (e.g., Davenport & Prusak, 1998). The purpose of the source map is to help pin down the appropriate sources quickly and detect weaknesses in the available sources.

THE SOURCE MAP AS A TOOL FOR ASSESSING SOURCES

What is a Source Map?

A source map links data (or classes of data) to sources in such a way that the (most) appropriate sources can be selected for the collection of the requested data. If viewed as a matrix, the column entries may refer to data classes (e.g., products under development by competitor X) and the row entries to possible sources. Each column then indicates what sources may be used to gather the requested data (e.g., a patent database, economic journals, or the Internet site of competitor X). To determine what sources are (most) appropriate, the source map needs to contain information about criteria for appropriateness and their valuation. The cells in the source map (connecting the data classes to sources) should contain this information. To get this information, it should be clear (a) what the relevant criteria are, (b) how they can be given a value, and (c) how to integrate them into an overall judgment of the appropriateness of the sources. The next two sections deal with these issues.

Note that we treat the source map as a tool for supporting and structuring collection activities *given* the data classes. We assume that the data (classes) are already defined in the direction phase (the first phase of the intelligence cycle).

Criteria and Scores for Judging Sources

The criteria for assessing the appropriateness of sources link up with the three activities required to deal with sources. These activities are the following.

1. Accessing the source. Accessing means determining the exact location and approaching the source to prepare retrieval.
2. Retrieving (in interaction with the source) the data from the source.
3. Using the retrieved data in further processing (i.e., for the production of intelligence).

Referring to these activities, the appropriateness of sources depends on four dimensions: (a) ease of access, (b) ease of retrieval, (c) usefulness of the content of the retrieved data and processing ease, and (d) cost effectiveness. Below, we discuss criteria in these dimensions.

Criteria for Access and Retrieval

To assess the appropriateness of sources regarding access and retrieval, barriers in employing a source can function as criteria (cf Fuld, 1995; Davenport & Prusak, 1998). Examples of these barriers are as follows.

- A language barrier.
- A cultural barrier (i.e., a difference in culture between collector and source).
- An institutional barrier. In some (bureaucratic) organizations, it may be very hard to locate and approach certain people and documents.
- A personal barrier. Personal characteristics can make it difficult to approach and interact with someone.
- A geographical barrier. Some sources need to be dealt with on location.
- A technological barrier. Accessing some sources and retrieving data from them may sometimes be possible only by means of specific information and communications technology, requiring specific knowledge or skills.
- A fee barrier. For accessing some sources and/or retrieving data, a fee may be charged.
- A time barrier. For some sources, the response time may be very slow.
- A clarity barrier. This barrier refers to the effort one has to give to make sense of the data from the source. Factors that increase this barrier are the use of specific jargon and the lack of (requested) structure in the data.
- A stability barrier. This barrier refers to the stability of access to the source (some sources may cease to exist, some are not available at the expected moment, others may decide to stop providing their services, etc.).

In our view, these criteria can also be used to express the costs associated with using a particular source. We therefore prefer to deal with the above criteria, instead of using cost estimates that may be derived from them, because (a) it is difficult to translate the criteria into costs and (b) if only cost estimates are used, one loses information about the appropriateness of sources.

Using a barrier as a criterion to assess appropriateness, it can be scored on a five-point Likert scale where 1 means *very problematic* and 5 means *nonexistent*.

Criteria for the Use of Data

There are four criteria for assessing the appropriateness of sources regarding the use of the data for the production of intelligence. One of them is a processing criterion and three of them are content criteria.

The processing criterion refers to the ease of processing. This can be determined by the format in which the data are delivered; that is, does the source deliver the data in a format that can be used directly for the purposes of the collector or does it need reformatting? One may score this criterion on a five-point scale ranging from 1, *much reformatting needed*, to 5, *right format*.

The content criteria are completeness, reliability, and timeliness (cf O'Brien, 1998, for a summary of these criteria). When applied to the value of sources, these criteria mean the following.

- Completeness: The source can deliver all the data required to gain insight into the data class for which the source is used. This can, for instance, be measured in terms of the number of times the source was unable to deliver the requested data and/or the number of aspects of a data class for which the source could not provide data.
- Reliability: This refers to the reliability of the data from the source. It can be measured, for instance, in terms of the number of times the data from the source proved to be incorrect.
- Timeliness: The data from the source is up to date. It can be measured in terms of the number of times that the source delivered obsolete data.

In literature, one often finds relevance as an additional criterion to assess the content of data. Relevance then refers to the suitability of the data in gaining insight into the data class for which the source was used. However, relevance can be adequately expressed in terms of completeness, reliability, and timeliness. Completeness links the data provided by a source to the required data defined by the data class. Given the completeness, the data should further be correct and up to date to be relevant. Relevance, therefore, can be treated as an overarching concept, referring to the other three content criteria.

The content criteria can, again, be scored on a five-point scale, where 1 means *very incomplete*, *very unreliable*, and *very obsolete*, respectively, and 5 means *very reliable, very complete*, and *very timely*.

Content of Source Map Cells

The criteria for the appropriateness of sources and their scores should be put in the source map. To this end, each cell in a source map contains the following information (see also Figure 1).

1. General information about the source, consisting of the name of the source, the data-carrier (human, data or electronic) and (if known) the exact or default location.
2. Scores on the criteria for access, retrieval, content and processing of the (data from the) source.
3. Information about what data could not be delivered if the source was incomplete. This is useful for analyzing the appropriateness of the sources (see next section).
4. Remarks concerning one of the above aspects.

Using the Source Map

A source map allows for two different uses. First, it is used to find appropriate sources for a particular data class. Second, it is used to assess the overall adequacy of the sources. For both types of use, it is necessary to compare the sources. In this section, we discuss how to compare sources and how to use this method for comparison for the two different uses.

Comparing Sources

Sources can be compared using a single criterion (e.g., which source scores highest on completeness?). It is also possible to integrate the values of (several) individual criteria and

Figure 1. Content of cells in a source map (the shaded areas are not applicable)

Name:
Carrier:
Location:

	Language barrier	Cultural barrier	Institut. barrier	Personal barrier	Geogr. barrier	Techno. barrier	Time barrier	Fee barrier	Clarity barrier	Stability barrier
Access	1…5									
Retrieval										

Content:
Completeness: 1…5 If incomplete: What data could not be delivered?
Timeliness: 1…5
Reliability: 1…5

Process/format: 1…5

Remarks: …

compare these integrated scores. To integrate these values into an overall score, we propose the following procedure.

1. Define two classes of criteria: efficiency criteria and effectiveness criteria. The class of efficiency criteria consists of the access criteria, the retrieval criteria, and the ease-of-processing criterion. The class of effectiveness criteria consists of the criteria completeness, reliability, and timeliness.
2. Estimate weights for the criteria in the two classes. A possible way of determining the weights of the individual criteria is to have CI professionals produce a rank order of the criteria (in each of the two classes) expressing their ideas about the relative relevance of the criteria. Next, one could discuss the results and produce one rank order for each class. (This procedure could be supported by groupware, such as Group Systems; cf Nunamaker, Dennis, Valacich, Vogel, & George, 1991).
3. Compute, for each source, the overall scores for the two classes. For both classes, we suggest taking the weighted average score for the given criteria. To compute these scores, the scores on the individual criteria should be available. These scores might be obtained initially by asking CI professionals. From that moment on, they should be evaluated every time a source is used and updated when necessary.

Finding Appropriate Sources

The most straightforward use of the map is to find out what sources are available for a particular data class. A step beyond merely enumerating available sources is to give a judgment about the appropriateness of these sources in terms of the criteria presented in the previous section. To this end, we use the efficiency and effectiveness scores of the sources. For a particular data class, all the available sources can be plotted regarding these two scores (see Figure 2).

The figure states that Source 5 scores best on effectiveness, Source 4 best on efficiency, and Source 1 scores lowest on both classes of criteria.

Figures like the above can help in analyzing the appropriateness of a source for a particular data class. As a general heuristic for ranking the sources, we suggest that sources in the upper right quadrant should be preferred to those in the lower right quadrant, and these should be preferred to the ones in the upper left quadrant. Sources in the lower left quadrant should probably be discarded.

Sources that come up as appropriate should be checked for completeness. If they are complete, they can be added to the list with preferred sources. If they are incomplete, it is necessary to find out if there are sources that can compensate for this lack. To this end, information about what data the source is unable to deliver can be used. This information directs the search for an appropriate compensating source.

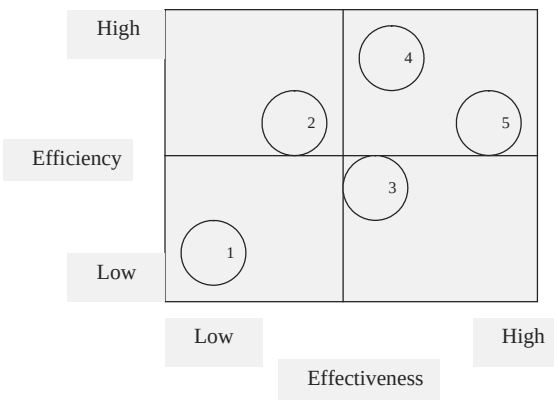
For sources that score high on effectiveness but low on efficiency (lower right quadrant), it should be examined (a) whether the relevance of the data class makes the effort (and costs) worthwhile and/or (b) whether measures can be taken to make the use of the source more efficient, for example, the efficiency of scale in gathering data (a subscription to an often-used online database).

For sources that score in the two left quadrants, it can be established what exactly causes the score. Dependent on the outcome of this investigation, it may be decided to stop using the source.

Assessing the Adequacy of Sources

To judge the overall adequacy of the sources, the map may help in answering the following questions.

Figure 2. Scores of five sources regarding their appropriateness for a particular data class (see text)



1. Do the sources cover all data classes?
2. Do we have adequate sources for the required data classes? If some data classes only have sources that have scores in the lower left quadrant of Figure 2, problems may arise. If a rank order of the data classes regarding their relevance exists, one can also establish whether the most relevant data classes are covered by appropriate sources.
3. Do we have enough different sources for the (most important) data classes? This question refers to the flexibility in collecting data. If a source is suddenly unavailable, one needs to have adequate alternatives. It is also useful to have different sources for the purpose of validating the data.

Answers to these questions help intelligence officers to identify weaknesses in the available sources and direct their efforts to repair them.

Implementing a Source Map

To build, maintain, and use a source map does not require exceptional resources. IT applications for implementing the map range from sophisticated applications to simple solutions. An example of a simple solution is an implementation of the map by means of Microsoft Excel sheets. However, it is also possible to use more sophisticated application, for instance, Web-based applications of the map. Making the map available via an intranet, for instance, can enhance its use and maintenance. In addition to these technological issues, it is important to define and allocate tasks and responsibilities regarding maintenance and use of the map. Finally, data collectors should be motivated to use the map to define their search strategies. In our experience, data collectors see the benefits of a good map and will be inclined to use and maintain it.

FUTURE TRENDS

To aid intelligence officers in their task to evaluate sources, the source map was introduced. Given the increasing need for organizations to collect data about their environment, it can be expected that the need for tools to evaluate sources (like the source map) will also increase. In order to deal with this, information technology tools may be tailored to support the implementation of source maps and the process of keeping them up to date (see, for instance, Philips, 2004).

CONCLUSION

To produce actionable intelligence, the efficient and effective use of sources is imperative. However, up until now, little attention has been paid to supporting the selection of sources. In this paper, we deal with this omission by presenting the source map as a support tool. Properly implemented source maps can be valuable instruments in the support of collection activities. In our view, they can aid in both the everyday use of sources and in the assessment of the overall adequacy of available sources.

REFERENCES

Bernhardt, D. C. (1994). I want it fast, factual, actionable: Tailoring competitive intelligence to executive needs. *Long Range Planning, 27*(1), 12-24.

Chen, H., Chau, M., & Zeng, D. (2002). CI-spider: A tool for competitive intelligence on the Web. *Decision Support Systems, 34*, 1-17.

Cook, M., & Cook, C. (2000). *Competitive intelligence.* London: Kogan Page.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge.* Boston: Harvard Business School Press.

Fleisher, C. G., & Blenkhorn, D. L. (2001). *Managing frontiers in competitive intelligence.* Westport: Quorum, CT.

Fuld, L. M. (1995). *The new competitor intelligence.* New York: Wiley.

Gilad, B., & Gilad, T. (1988). *The business intelligence system.* New York: Amacon.

Herring, J. P. (1999). Key intelligence topics: A process to identify and define intelligence needs. *Competitive Intelligence Review, 10*(2), 4-14.

Kahaner, L. (1997). *Competitive intelligence.* New York: Touchstone.

O'Brien, J. (1998). *Introduction to information systems* (2nd ed.). New York: McGraw-Hill.

Nunamaker, J. F., Dennis, A. R., Valacich, J. S., Vogel, D. R., & George, J. F. (1991). Electronic meetings to support group work. *Communications of the ACM, 34*(7), 40-61.

Philips, E. A. (2004). Building a competitive intelligence system: An infrastructural approach. In D. Vriens (Ed.), *Information and communication technology for competitive intelligence* (pp. 227-247). Hershey, PA: IRM Press.

Sammon, W. L. (1986). Assessing the competition: Business intelligence for strategic management. In J. R. Gardner, R. Rachlin, & H. W. Sweeny (Eds.), *Handbook of strategic planning* (pp. 4.12-4.19). New York: Wiley.

Teo, T. S. H., & Choo, W. Y. (2001). Assessing the impact of using the Internet for competitive intelligence. *Information & Management, 39*, 67-83.

Vriens, D. (2004). *Information and communication technology for competitive intelligence.* Hershey, PA: IRM Press.

Vriens, D., & Philips, E. A. (1999). Business intelligence als informatievoorziening voor de strategievorming. In E. A. Philips & D. Vriens (Eds.), *Business intelligence.* Deventer, Netherlands: Kluwer, 11-44.

KEY TERMS

Collection Stage: Stage of the intelligence cycle. In this stage, sources regarding the required environmental data are located and accessed, and the data are retrieved from them.

Competitive Intelligence: In the literature, two definitions are used: a product definition and a process definition. In the product definition, competitive intelligence is defined as information about the environment, relevant for strategic purposes. The process definition highlights producing and processing this environmental information. Process definitions often refer to the intelligence cycle.

Intelligence Cycle: Cycle of four stages (collections of intelligence activities). The stages are direction (also referred to as planning, in which the strategic information requirements are determined), collection (determining sources and retrieving data), analysis (assessing the strategic relevance of data), and dissemination (of the intelligence to strategic decision makers).

Source: Something or someone containing data and from which the data can be retrieved. Many distinctions regarding sources are given in the competitive intelligence literature, for instance, open versus closed sources, primary versus secondary sources, internal versus external sources, and a distinction referring to the carrier of the data (human, electronic, or paper).

Source Evaluation: The process of assessing the efficiency and effectiveness of a source or several sources, given certain criteria. The result of this process can be (a) a judgment about the usefulness of a particular source for collecting data and/or (b) an insight into the relative usefulness of all available sources. See also "Source map."

Source Identification: Identifying suitable sources (i.e., efficient and containing the relevant data) given a certain data need. See also "Source map."

Source Map: A source map is a matrix linking data classes to sources. In the cells of the matrix, the sources are valued according to different criteria (e.g., accessibility, costs, timeliness of the data, etc.).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2690-2695,

Supporting the Mentoring Process

Karen Neville
University College Cork, Ireland

Ciara Heavin
University College Cork, Ireland

INTRODUCTION

While the concept of knowledge management (KM) is not new, the focus on knowledge management as a strategy has heightened in recent times as organizations realize the importance of knowledge as an intangible asset contributing to the enhancement of competitive advantage (Bolloju & Khalifa, 2000). In the 21st century, it is believed that successful companies are those that effectively acquire, create, retain, deploy, and leverage knowledge (Cecez-Kecmanovic, 2000). Knowledge work is the ability to create an understanding of nature, organizations, and processes, and to apply this understanding as a means of generating wealth in the organization. Evidently, the focus on knowledge management as a strategy has become central to organizations (Davenport & Prusak, 1998). Ichijo, Von Krogh, and Nonaka (1998) view knowledge as a resource that is unique and imperfectly imitable, allowing firms to sustain a competitive advantage. Additionally, many approaches to managing knowledge are marred by obstacles of sustainability (Kulkarni, Ravindran, & Freeze, 2006). As a direct result organizations fail to realize the expected returns on investment from knowledge management implementations or strategies (Zyngier, 2007). However, if knowledge management as a formalized organizational strategy is supported, it can be sustained. Therefore in an economic environment where organizations have been forced to take a step back and reevaluate their core competencies and ability to innovate, organizational knowledge has come to the forefront as a valuable strategic asset (Haghirian, 2003). It is the objective of this article to provide an example of knowledge workers and experts collaborating to implement successful training and learning programs to support knowledge management activities in their organization. The authors hope that the case discussed will inform researchers of an appropriate model in designing an interactive learning environment which enables a positive knowledge sharing environment and in turn contributes to the growth of an organization's memory.

BACKGROUND

The intensity of competition in the business market, advances in technology, and a strong shift towards a knowledge-based economy have each contributed to the demand for Web-based mentoring systems. "There is no knowledge that is not power," according to Emerson (1843), and the organization (public or private) that can utilize its knowledge resources more effectively than its competitor will persevere. An effective mentoring system between knowledge workers and experts can provide an organization with a strategic advantage in the market. Mentoring environments can help create and maintain skills, and therefore the corporate knowledge base. They both alleviate the strain on corporate resources and facilitate employees' changing training needs through knowledge sharing. Therefore the majority of organizations face the enormous challenge of supporting their employees' thirst for expanding their skill base and effectively their corporate assets, as "knowledge implies a knower; *the rest is just information.*" Some companies exploit the capabilities of Web technology to facilitate knowledge sharing at workgroup and company levels (Davenport, 1996). Recent evidence points to the deployment of organizational systems with the primary objectives of improving customer services, increasing revenue, containing costs, and improving internal processes—in other words, creating competitive advantage. In the case under consideration, the organization implemented a successful mentoring system in order to develop employee skills and knowledge in both IT and managerial issues such as knowledge management. This article is focused on the development of a Web-based mentoring system (WBMS) to mentor (Neville, Adam, & McCormack, 2002) workers and enhance learning. The case study indicates a strong requirement for the utilization of such an environment to both increase support for and collaboration between the knowledge workers.

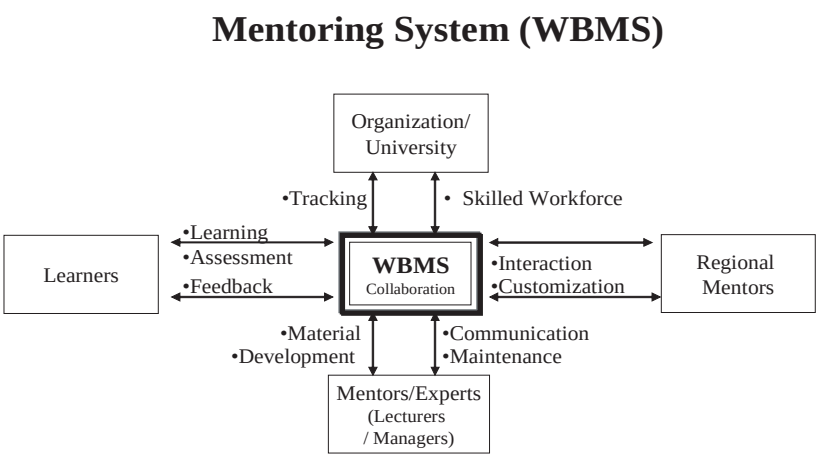
MAIN FOCUS OF THE ARTICLE

Mentoring is a traditional method of teaching that strengthens the concept and objectives of learning/training (Neville et

al., 2002; Heavin & Neville, 2006). The Oxford dictionary defines the word mentor as a "wise counselor, who tutors the learner in intellectual subjects." When this model is applied to a learning network, the student is called a teleapprentice who studies using appropriate methods. The teleapprentice reads messages, answers questions, participates in discussions, and conducts research online to master his or her subject. Mentorship is a method of teaching that has been used for hundreds of years; this design is incorporated into learning/knowledge networks to develop more effective learning and collaborating practices, and provide additional support and mediation to the learners/workers. 'Access to experts' is one of the many advantages provided through learning networks (Harasim, Hiltz, Teles, & Turoff, 1995; Hansen, Nohria, & Tierney, 1999). Networks are, in fact, modeled on this method (Harasim et al., 1995). Therefore, Web-based mentoring systems allow students/workers to communicate with experts in a field and collaborate with their peers. WBMSs can be described as learning delivery environments in which the WWW is its medium of delivery (Neville et al., 2002). The possibilities of WBMSs are limited only by constraints imposed by the university or organization in question, such as technological or managerial support (Neville, 2000). Innovative companies and universities are using this implementation for a number of reasons, specifically to keep employees or students abreast of emerging technologies in their fields, and to provide effective training to both staff and customers on new products and skills. Designing a WBMS requires a thorough investigation into the use of the Web as a medium for delivery (McCormack & Jones, 1997). The designer must be aware of the attributes of the WWW and the principles of instructional design to create a meaningful support environment (Gagne, Briggs, & Wagner, 1988). The Web-based training room is viewed as an innovative approach to teaching (Relan & Gillani, 1997). The virtual training room, like the traditional method, requires careful planning to be both effective and beneficial (Dick & Reiser, 1989). As stated by McCormack and Jones (1997), a Web-based classroom must do more than just distribute information; it should include resources such as discussion forums to support collaboration between learners and ultimately it should also support the needs of both the novice and advanced learner. A WBMS is composed of a number of components that are integral to the effective operation of the environment, for example the development of content, and the use of multimedia, Internet tools, hardware, and software (Reeves, 1993a). A developer must understand the capabilities of these components (search engines, feedback pages, and movie clips), as their use will determine the success or the failure of the learning environment. In this article we provide an example of a WBMS to help illustrate the main elements, issues, components, and problems encountered through the implementation of learning systems to enhance knowledge management in organizations.

The WBMS (see Figure 1) was constructed to support and develop knowledge sharing for personnel who seek to acquire and develop their knowledge management skills. Training material is available online, but in addition, a discussion forum enables both learners and experts to exchange ideas and add to the environment. This allows learners to provide feedback (anonymously, if desired) to the experts. It also enables them to pose queries, which other participants or the experts can answer. All participants are able to see the initial queries and the discussion stream of answers from

Figure 1. The Web-based mentoring system (WBMS)



other participants and the instructors. This further extends the reach of the training material, as employees can log on to the WBMS at home or at work and pose questions for which answers are available when they next log on. The facility also allows the learners to voice their satisfaction regarding the different elements of the environment. This provides the participants with the opportunity to take part in the ongoing design of the WBMS, and therefore increases user acceptance.

Figure 1 illustrates the opportunities available to the participants of the case. The system provides professional training to a range of employees, such as full-time staff at all levels including senior management. Its core aim is to develop further the knowledge and abilities of personnel so that they are increasingly aware of IT and management issues within the organization, with a particular focus on capturing, storing, disseminating, and creating knowledge. The course is designed on a distance-learning basis and is supported by a tutorial system. The main purposes of the tutorials are to facilitate the learning process, assist in the completion of interactive assignments, and encourage team playing within the group. Learners are presented with written modules, which act as 'the lecture', and the expert plays the role of the facilitator, enabling the students to combine the written materials with their own experiences. Feedback from the students identified the need to provide additional learning support through an online environment. The WBMS has provided an improved learning process and has enabled enhanced collaboration among employees. This article focuses on the development of these requirements through an interactive learning environment for employees, and the WBMS is designed or customized for the requirements of the individual learner. This approach accounts for the varying learning abilities of students and overcomes the limitations of traditional training environments, which are restricted to rules in order to adequately facilitate the group. The educator/expert instructs a class, but the level of both collaboration and the development of problem-solving skills can be directly correlated to class size. The greater the size of the group, the less attention individual learners gain or the more intimidated a student is to participate in discussions, thus reducing collaboration. The WBMS, when adequately designed, can reduce the limitations of the classroom and allow the learner to work at his or her own pace with structured support from both the educators and the other learners.

DESIGN CONSIDERATIONS

Harasim et al. (1995) describe knowledge networking as "the use of electronic linkages among different teaching and learning communities to facilitate information acquisition and knowledge building" (p. 10). Prime examples of knowledge networking are the thousands of discussion groups that exist on the Internet that exchange information and discuss a wide range of various different topics representing an "active forum of informal learning and information exchange: a knowledge network." "Knowledge networking is based on self-directed learning and growth through the pursuit of information, skills, and knowledge" (Harasim et al., 1995, p. 11), and learning occurs when experts/students/mentors/people interact with each other (for example by discussing a certain topic of interest). One of the main disadvantages to traditional distance education (characterized by "correspondence, paper-based assessment and tutor-student communication") is the "lack of immediacy and adequate communication" between mentors and knowledge workers or students (Louvieris & Lockwood, 2001). However, Louvieris and Lockwood (2001) argue that the use of ICT can enhance the mentor process, and that this is evident in the objectives of some higher education institutions who now "offer access to learning materials to students who are located geographically locally or remotely, facilitate support from their tutors/lecturers and also encourage participation in online activities while [taking] a course" (p. 1100).

Organizations are increasingly using and investing in Web technology as it can support all aspects of organizational work, including mentoring (Isakowitz, Bieber, & Vitali, 1998; Carstensen & Vogelsang, 2001), resulting in the fact that "higher education institutions are aiming to leverage both information and communication technologies, particularly the Internet, to redefine and/or extend their business scope in increasingly competitive global/international learner markets" (Louvieris & Lockwood, 2001, p. 1099).

Turoff and Hiltz (1998) argue that Web technology "is fundamentally a new medium of human communication" (p. 116). As students and mentors work together collaboratively, they form what Harasim et al. (1995) define as a 'learning community' that "can be both personally and educationally enriching," as the Web-based classroom offers the students more opportunities to voice their opinions or comments over the traditional classroom environment, so students feel that "online environments enable them to communicate with their colleagues more than in face-to-face classes." Learning through the Web is regarded as the 'silver bullet' solution to learning issues faced by both universities and organizations, despite little quantitative evidence to support claims of its effectiveness. Therefore, it is essential to define the characteristics of interactive education/mentoring that can be achieved through the WWW to promote learning. The identification of these characteristics is necessary to implement such a concept. Thus, this section reviews additional dimensions proposed by Reeves and Reeves (1993) for interactive training and collaboration including: (1) educational philosophy, (2) learning theory, (3) goal orientation, (4) task orientation, (5) source of motivation, (6) role of the teacher, (7) metacognitive support, (8) collaborative learning, (9) cultural sensitivity, and (10) structural flexibility.

The dimensions are proposed to describe the characteristics of a learning environment—the characteristics of a WBMS. Each of the dimensions identified are outlined in the following paragraphs:

1. *Educational philosophy* emphasizes the belief that learners build their cognitive strategies on previous knowledge and on the learning environment. Therefore a rich and stimulating environment is required to train the different adult learners. Thus, direct instruction is also replaced with challenging tasks.
2. The design of the environment should be based on researched *learning theories*. The two dominant theories identified in the design of training environments are behavioral and cognitive psychology. Behaviorists believe that the most important factors that should be taken into consideration are the arrangement of stimuli, responses, feedback, and reinforcement to shape the desirable behavior of the learners. By contrast, cognitive psychologists place more emphasis on internal mental states rather than on behavior. As a result, the WBMS design, using cognitive theory, will be based on direct instruction and practice exercises.
3. The *goals* for a WBMS can vary from sharply focused, where a specific environment is required, to a more general approach.
4. The orientation of *tasks* can range from academic to authentic. By contrast, an authentic design for adult education would require the learners to tackle job-related exercises or cases (tacit knowledge). The design orientation of a WBMS should support the transfer of skills to the learners.
5. *Motivation* is the main factor for the success of any learning environment. The source of motivation ranges from two extremes, from extrinsic (outside the learning environment) to intrinsic (a part of the learning environment).
6. Lecturers and tutors fulfill different *roles* from the traditional role of instructor (didactic) to the facilitative role.
7. *Metacognition* is described as the learner's ability to identify objectives, and plan and understand learning strategies. Thus, a WBMS can be designed to challenge the learner to solve course-related problems.
8. The *collaborative* learning dimension for a WBMS can also range from lack of support to the inclusion of facilities to support it.
9. Reeves et al. (1993) argue that all training environments have *cultural implications*. However, the development of a WBMS cannot be designed to adjust to every rule. Therefore, WBL should be designed to be as culturally aware as possible.
10. *Structural flexibility* describes a WBMS as either asynchronous or synchronous. Open or asynchronous environments refer to the use of such an environment at any particular time or from any location. However, synchronous refers to fixed environments that can only be used in the training room of an organization. The WWW provides educators and students alike with the opportunity to avail of resources from more open environments through which students are supported or mentored in the acquisition of both tacit and explicit knowledge.

The dimensions were used as an aid in the production of the generic WBMS (see Figure 1). Both the study of the different dimensions and the factors necessary for the collaboration and structure of learning provide valuable information and steps for the analysis and therefore the development of the solutions.

FUTURE TRENDS

Knowledge workers have praised the hands-on approach provided through this expert-driven system. As knowledge sharing has increasingly become a key organizational objective, this type of environment provides an extensive communication channel leveraging the technology to support a wide variety of knowledge-sharing activities. It also enables the experts and learners to collaborate, therefore providing 24-hour online support. This case is a prime example of a successful KM support tool that can and will continue to avail of technological advances to ensure ongoing success. Further research exploring the various pedagogy and technology mixes to produce a set of options, which would identify the integration of a particular pedagogy with an appropriate technology, would prove beneficial if WBM is to meet its full potential. Additionally, the WBMS illustrated in this chapter is primarily concerned with the downstream development process that incorporates key design and development considerations. Therefore, further research exploring the upstream development process would be worthwhile. This would involve exploring some of the development options that were identified in the development of off-the-shelf packages, and exploring the open source development option for a WBMS. These further studies may yield interesting results and therefore increase the level of understanding of the development of effective WBMSs.

CONCLUSION

Creating a knowledge or learning community is one of the major challenges for organizations and universities alike. Poe and Stassen (2002, p. 29) state:

"A sense of community—where students are able to work cooperatively with peers on course material, have the opportunity for positive interaction with the instructor, and where the learning environment is respectful and motivates students to do their best—is key to a positive and successful learning experience."

After an in-depth analysis it was apparent that the learners lacked an efficient online support system, which would complement alternative communication channels such as face-to-face encounters and traditional training classes as a means of knowledge sharing. An effective KM training support system can provide an organization with a strategic advantage in the market. Learning environments can help create and maintain skills and therefore increase the corporate knowledge base. They both alleviate the strain on corporate resources and facilitate employees' changing training needs. This article focuses on the design of a suitable environment to support knowledge workers and encourage collaboration. The research outlines the factors necessary for the successful implementation and use of the system. It also highlights the potential of the system to overcome the physical barriers of traditional knowledge sharing and learning channels. Interactive learning environments can, when properly mediated and structured, facilitate cooperation and enhanced learning practices, reduce conflict, and avail all of the benefits that technology can provide.

REFERENCES

Bolloju, N., & Khalifa, M. (2000, July 9-11). A framework for integrating decision support and knowledge management in enterprise-wide decision making environments. *Proceedings of the IFIP TC8/WG8.3 International Conference on Decision Support through Knowledge Management,* Stockholm, Sweden.

Carstensen, P.H., & Vogelsang, L. (2001, June 27-29). Design of Web-based information systems—new challenges for systems development? *Proceedings of the 9th European Conference on Information Systems* (ECIS 2001) (vol. 1, pp. 536-547), Bled, Slovenia.

Cecez-Kecmanovic, D. (2000, July 9-11). Understanding knowledge sharing in organizational decision making supported by CMC. *Proceedings of the IFIP TC8/WG8.3 International Conference on Decision Support Through Knowledge Management,* Stockholm, Sweden.

Davenport, T.H. (1996). Some principles of knowledge management. *Strategy and Business,* (Winter), reprint no. 96105.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know.* Boston: Harvard Business School Press.

Gagne, R.M., Briggs, L.J., & Wagner, W.W. (1988). *Principles of instructional design* (3rd ed.). New York: Holt Reinbank Winston.

Haghirian, P. (2003). Does culture really matter? Cultural influences on the knowledge transfer process within multinational corporations. *European Journal of Information Systems.*

Hansen, M., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? *Harvard Business Review,* (March-April), 106-116.

Harasim, L., Hiltz, S.R., Teles, L., & Turoff, M. (1995). *Learning networks: A field guide to teaching and learning online.* Cambridge, MA: The MIT Press.

Heavin, C., & Neville, K. (2006). Mentoring knowledge workers. In D. Schwartz (Ed.), *Encyclopedia of knowledge management* (pp. 621-626). Hershey, PA: Idea Group.

Ichijo, K., Von Krogh, G., & Nonaka, I. (1998). *Knowledge enablers. Knowing firms: Understanding, managing and measuring knowledge.* London: Sage.

Isakowitz, T., Bieber, M., & Vitali, F. (1998). Web information systems. *Communications of the ACM, 41*(7), 78-80.

Kulkarni, U.R., Ravindran, S., & Freeze, R. (2006) A knowledge management success model: Theoretical development and empirical validation. *Journal of Management Information Systems, 23,* 309-347.

Louvieris, P., & Lockwood, A. (2001, June 27-29). An analysis of the UNICAFE experience and its implications for IT induced business transformation in higher education. *Proceedings of the 9th European Conference on Information Systems* (ECIS 2001) (pp. 1098-1109), Bled, Slovenia.

McCormack, C., & Jones, D. (1997). *Building a Web-based education system.* New York: John Wiley & Sons.

Neville, K. (2000). A Web-based training (WBT) system development framework: A case study. *Proceedings of the 10th Annual Conference on Business Information Technology Management,* Manchester, UK.

Neville, K., Adam, F., & McCormack, C. (2002, June). Mentoring distance learners: An action research study. *Proceedings of the 10th European Conference on Information Systems,* Gdañsk, Poland.

Poe, M., & Stassen, M.L.A. (2002). *Teaching and learning online: Communication, community, and assessment—a*

handbook for UMass faculty. Office of Academic Planning and Assessment, University of Massachusetts Amherst, USA.

Reeves, T.C. (1993a). Research support for interactive multimedia: Existing foundations and future directions. In C. Latchem, J. Williamson, & L. Henderson-Lancett (Eds.), *Interactive multimedia: Practice and promise* (pp. 79-96). London: Kogan Page.

Relan, A., & Gillani, B. (1997). Web-based information and the traditional classroom. In C.M. Reigeluth & R.J. Garfinkle (Eds.), *Systemic change in education.* Englewood Cliffs, NJ: Educational Technology.

Turoff, M., & Hiltz, S.R. (1998). Superconnectivity. *Communications of the ACM, 41*(7), 116.

Zyngier, S. (2007a). Knowledge management governance: A framework for knowledge management benefits realization. *Proceedings of the 8th International Research Conference on Quality, Innovation and Knowledge Management,* New Delhi, India.

KEY TERMS

Explicit Knowledge: Information that has specific meaning and can be easily and clearly understood.

Knowledge Web: The use of electronic linkages among different teaching and learning communities to facilitate information acquisition and knowledge building.

Knowledge Work: The ability to create an understanding of nature, organizations, and processes, and to apply this understanding as a means of generating wealth in the organization.

Mentoring: A method of teaching that has been used for hundreds of years; this design is incorporated into learning networks to develop more effective learning practices and provide additional support to the learner.

Tacit Knowledge: Knowledge gained through an individual's own experiences.

System Dynamics Based Technology for Decision Support

Hassan Qudrat-Ullah
York University, Canada

INTRODUCTION

Managers face problems that are increasingly complex and dynamic. Decision support systems (DSS) are designed to assist them make better decisions. However, the empirical evidence concerning the impact of DSS on improved decision making and learning in dynamic tasks is equivocal at best (Klabbers, 2003; Sharda, Steve, Barr, & McDonnell, 1988; Sterman, 2000; Todd & Benbasat, 1999). Over four decades of dynamic decision making; studies have resulted in a general conclusion on why people perform poorly in dynamic tasks. In dynamic tasks, where a number of decisions are required rather than a single decision, decisions are interdependent, and the decision-making environment changes as a result of the decisions or autonomously or both (Edwards, 1962), most often the poor performance is attributed to subjects' misperceptions of feedback. That is, people perform poorly because they ignore time delays between their "actions and the consequences" (Sterman, 2000) and are insensitive to the feedback structure of the task system (Diehl & Sterman, 1995). Decision maker's mental models about the task are often inadequate and flawed (Kerstholt & Raaijmakers, 1997; Romme, 2004). In this paper we argue that system dynamics based interactive learning environments (ILEs) could provide effective decision support for dynamic tasks by reducing the misperceptions of feedback.

BACKGROUND

Dynamic Decision Making

Dynamic decision-making situations differ from those traditionally studied in static decision theory in at least three ways: (1) a number of decisions are required rather than a single decision, (2) decisions are interdependent, and (3) the environment changes, either as a result of decisions made or independently of them or both (Edwards, 1962). Recent research in system dynamics has characterized such tasks by feedback processes, time delays, and nonlinearities in the relationships between decision task variables (Romme, 2004). Driving a car, managing a firm, and controlling money supply are all dynamic tasks (Diehl & Sterman, 1995) In these tasks, contrary to static tasks such as lottery-type gambling, locating a park on a city map, and counting money, multiple and interactive decisions are made over several periods, whereby these decisions change the environment, giving rise to new information and leading to new decisions (Forrester, 1961; Sterman, 2000).

ILE

We use *ILEs* as a term sufficiently general to include microworlds, management flight simulators, DSS, learning laboratories, and any other computer simulation-based environment—the domain of these terms is all forms of action whose general goal is the facilitation of dynamic decision making. Based on the on-going work in the system dynamics discipline (Moxnes, 2004; Otto & Struben, 2004; Qudrat-Ullah, in press; Sterman, 2002), this conception of ILE embodies learning as the main purpose of an ILE. Under this definition of ILE, learning goals are made explicit to the decision makers. A computer simulation model is built to represent adequately the domain or issue under study with which the decision makers can experience and induce real world-like responses (Qudrat-Ullah, 2005). Human intervention refers to active keying in of the decisions by the decision makers into the computer simulation model via the interface of an ILE.

Performance in Dynamic Tasks

How well do people perform in dynamic tasks? The empirical evidence (Diehl & Sterman, 2000; Klabbers, 2003; Moxnes, 2004; Sterman, 2000) suggests almost a categorical answer: "very poorly." Very often the poor performance in dynamic tasks is attributed to subjects' misperceptions of feedback (Moxnes, 2004; Sterman, 2000). The misperception of feedback (MOF) perspective concludes that subjects perform poorly because they ignore time delays and are insensitive to feedback structure of the task system. The paramount question remains, are people inherently incapable of controlling system with time lags, nonlinearities, and feedback loops? Contrary to Sterman's MOF hypothesis, an objective scan of real-world decisions would suggest that experts can deal efficiently with highly complex dynamic systems in real life, such as, for example, maneuvering a ship through restricted waterways. The expertise of river

pilots, for example, seems to consist more of using specific knowledge (e.g., pile moorings, buoys, leading lines) that they have acquired over time than in being able to predict accurately a ship's movements (Schraagen, 1994). This example suggests that people are not inherently incapable of better performance in dynamic tasks. Instead, decision makers need to acquire the requisite expertise.

SUPPORTING DYNAMIC DECISION MAKING THROUGH ILES

There exists some fundamental barriers to developing expertise in dynamic tasks: (1) *dynamic complexity*: our limited ability to understand the impact of time delays between our actions and their consequences coupled with the interactions between feedback loops that are multiple and nonlinear in character and are ever present in the task systems we face in the real world, (2) *information availability limitations*: information we estimate, receive, and communicate is often oversimplified, distorted, delayed, biased, and ambiguous, (3) *information processing limitations*: when it comes to decision making people generally adopt an event-based, open-loop view of causality, ignore feedback processes; fail to appreciate time delays and are insensitive to nonlinearities present in the feedback loop structures of the task system; perceive flawed cognitive maps of the causal structure of the systems; make erroneous inferences even about the simplest possible feedback systems; and fall prey to judgmental errors and biases, defensive routines, and implementation failure (Sterman, 2000). The effective DSS, therefore, should allow the users to overcome such impediments to decision making and learning in dynamic tasks.

ILEs meet this challenge through the provisions of (1) a representative simulation model of the task system, (2) powerful interface, and (3) human tutor support—the three fundamental components of any ILE.

Decision Support Through the Simulation Model

The greatest strength and appeal of an ILE in supporting decision making and learning in dynamic tasks lies in its underlying simulation model. In an ILE, the simulation model is built on system dynamics methodology (Forrester, 1961). The fundamental premise of system dynamics methodology is that "the structure of the system drives its behavior." That structure consists of feedback loops; stocks and flows; and nonlinearities arising from the interaction of these basic structures (Oliva, 2003; Sterman, 2000). A typical system dynamics model allows that

- the interaction and feedback between the systems variables, over time, in and across various sectors (e.g., demand, supply, production, finances, etc.) of the task system be explicitly represented and the structural assumptions are made explicit and open;
- the disequilibrium framework for modeling be established, where the adjustments, say in the need for variable "A" in response to the changes in variable "B" to new equilibria typically create imbalances and transient behavior;
- delays and other distortions in perceiving the true value of the variables be explicitly modeled;
- desired and actual variable magnitudes be explicitly distinguished from real magnitudes in the model; and
- nonlinear responses to actions be explicitly represented.

The significance of the modeling capabilities of system dynamics methodology is its contribution to our understanding of the structure and behavior of complex, dynamic systems. An understanding of the relationship between the structure(s) and behavior(s) leads to the formulation of a better mental model of the task system (Sterman, 2002) and improved decision making (Brekke & Moxnes, 2003; Romme, 2004).

Decision Support Through the Interface Design

Dörner (1980) asserts that decisions makers in dynamic tasks must acquire some reasonably precise notions of relationships among key task variables and develop an understanding of the most influential delays and feedback loops in the task system. System dynamics methodology provides powerful tools to represent qualitatively the connections between structure and behavior of the task system through (1) causal loop diagrams and (2) stock and flow structures. Utilizing these tools together with advances in modern IT, powerful interface, whereby references to the underlying simulation model are facilitated interactively in an ILE, can be constructed (for an excellent illustration please see, Romme, 2004). In this way, ILEs aid decision making by allowing the learners to examine the structure-behavior relationship as and when needed in an ILE session.

Decision Support Through Tutor Support

Decisional aid in the form of human tutor support constitutes the distinguishing and fundamental component of an ILE model. In an ILE session, decisional aids can be provided at three levels: (1) pre-, (2) in-, and (3) post-task levels. Pre-task level decisional aids can be conceptualized as information

provided by the human tutor to a decision maker about the model of the task prior to performing the task (Corner, Buchanan, & Henig, 2001; Davidsen & Spector, 1997). In-task decisional aids attempt to improve the individuals' decision-making performance by (1) making the task goals explicit at early stages of learning, (2) helping them keep track of goals during the task, and (3) providing them with "diagnostic information" (Cox, 1992). Post-task level decisional aids aim at improving performance by providing the decision makers an opportunity to reflect on their experiences with task (Cox, 1992; Davidsen & Spector, 1997).

Thus, an ILE could support the user's understanding of dynamic tasks by offering the opportunity to, experimentally, design, test, and evaluate their decision strategies.

FUTURE TRENDS

Turning to the future, the most fundamental research question for dynamic decision-making research seems to be: How to acquire expertise in dynamic tasks? A solution to this question would effectively provide a way to improve dynamic decision making. We present a potential solution as shown in Figure 1. Our conceptual model presents an integrated approach to improve dynamic decision making.

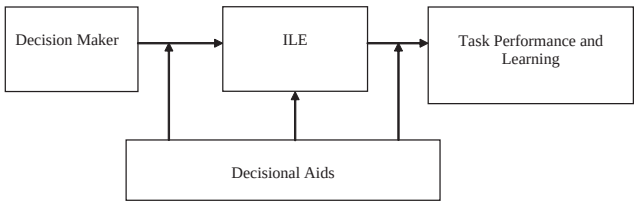
In future studies, by utilizing our model, we could then compare results, build a cumulative knowledge base of effective decisional aids, and study the trade-offs among various kinds of decisional aids to dynamic decision making. Pursued vigorously and systematically, research on decisional aids such as ILEs should be beneficial to both dynamic decision-making researchers and practitioners.

CONCLUSION

Dynamic decision-making research is highly relevant to the managerial practice (Diehl & Sterman, 1995; Kerstholt & Raaijmakers, 1997). We need effective DSS to help the managers cope with the ever present dynamic tasks. We presented ILE as a viable decision support for dynamic tasks. Investigations regarding the overall effectiveness of ILEs, we believe, will advance our insights into the design conditions for an effective DSS to promote decision making and learning in dynamic tasks.

REFERENCES

Figure 1. The conceptual model: How to improve dynamic decision making?



Brekke, K. A., & Moxnes, A. (2003). Do numerical simulation and optimization results improve management? Experimental evidence. *Journal of Economic Behavior and Organization, 50*(1), 117-131.

Corner, J., Buchanan, J., & Henig, M. (2001). Dynamic decision problem structuring. *Journal of Multicriteri Decision Analysis, 10*(3), 129-143.

Cox, R. J. (1992). Exploratory learning from computer-based systems. In S. Dijkstra, H. P. M. Krammer, & J. J. G. Van Merrienboer (Eds.), *Instructional models in computer-based learning environments* (pp. 405-419). Berlin, Heidelberg, Germany: Springer-Verlag.

Davidsen, P. I., & Spector, J. M. (1997). Cognitive complexity in system dynamics based learning environments. *International System Dynamics Conference.* Istanbul, Turkey: Bogacizi University Printing Office.

Diehl, E., & Sterman, J. D. (1995). Effects of feedback complexity on dynamic decision making. *Organizational Behavior and Human Decision Processes, 62*(2), 198-215.

Dörner, D. (1980). On the difficulties people have in dealing with complexity. *Simulations and Games, 11,* 8-106.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors, 4,* 59-73.

Forrester, J. W. (1961). *Industrial dynamics.* Cambridge, MA: Productivity Press.

Kerstholt, J. H., & Raaijmakers, J. G. W. (1997). Decision making in dynamic task environments. In R. Ranyard, R. W. Crozier, & O. Svenson (Eds.), *Decision making: Cognitive models and explanations* (pp. 205-217). New York: Routledge.

Klabbers, J. H. G. (2003). Gaming and simulation: Principles of a science of design. *Simulation & Gaming, 34*(4), 569-591.

Moxnes, E. (2004). Misperceptions of basic dynamics: The vase of renewable resource management. *System Dynamics Review, 20,* 139-162.

Oliva, R. (2003). Model calibration as a testing strategy for system dynamics models. *European Journal of Operational Research, 51,* 552-568.

Otto, P., & Struben, J. (2004). Gloucester fishery: Insights from a group modeling intervention. *System Dynamics Review, 20*(4), 287-312.

Qudrat-Ullah, H. (2005). MDESRAP: A model for understanding the dynamics of electricity supply, resources, and pollution. *International Journal of Global Energy Issues, 23*(1), 1-14.

Qudrat-Ullah, H. (in press). Behavior validity of a simulation model for sustainable development. *International Journal of Management and Decision Making.*

Romme, A. G. (2004). Perceptions of the value of microworld simulation: Research note. *Simulation & Gaming, 35,* 427-436.

Schraagen, J. M. C. (1994). *What information do river pilots use?* (Rep. No. TNO TM 1994 C-10). Soesterberg, The Netherlands: Human Factor Research Institute.

Sharda, R., Steve, H., Barr, J., & McDonnell, C. (1988). Decision support system effectiveness. *Management Science, 34*(2), 139-159.

Sterman, J. D. (2000). *Business dynamics.* New York: McGraw-Hill.

Sterman, J. D. (2002). All models are wrong: Reflections on becoming a system scientist. *System Dynamics Review, 18*(4), 501-531.

Todd, P., & Benbasat, I. (1999). *Information Systems Research, 10,* 356-381.

KEY TERMS

Behavior of System: The patterns of performance of the variable(s) of the system over time.

Diagnostic Information: Information that helps to understand why a particular consequence happened or did not happen.

Feedback: It is a process whereby an input variable is fed back by the output variable. For example, an increased (or decreased) customer base leads to an increase (or decrease) in sales from word of mouth which then is fed back to the customer base, increasingly (or decreasingly).

Mental Model: A mental model is the collection of concepts and relationships about the image of real world things we carry in our heads. For example, one does not have a house or a city or a gadget in his/her head but a mental model about these items.

Nonlinearity: A nonlinearity exists between a cause (decision) and effect (consequence), if effect is not proportional to cause.

Simulation Model: A simplified, computer, simulation-based construction (model) of some real world phenomenon (or the problem task).

Time Delays: Often the decisions and their consequences are not closely related in time. For instance, the response of gasoline sales to the changes in price involves time delays. If prices go up only after a while may sales drop.

Systems Thinking and the Internet from Independence to Interdependence

Kambiz E. Maani
The University of Queensland, Australia

INTRODUCTION

Despite our most impressive advances in sciences and technology, our prevailing worldview and the way we work and relate is deeply rooted in the thinking that emerged during the Renaissance of the 17th century! This thinking was influenced by the sciences of that era and in particular by Newtonian physics. Newton viewed the world as a *machine* that was created to serve its master–God, (Ackoff, 1993). The machine metaphor and the associated mechanistic (positivist) worldview, which was later extended to the economy, society, and the organization, has persisted until today and is evident in our thinking and vocabulary. The mechanistic view of the enterprise became less tenable in the 20th century partly due to the emergence of the corporation and the increasing prominence of human relation issues in the workplace. Today, this way of thinking has reached its useful life – The futurist, Alvin Toffler declared in 1991 "the Age of the Machine is screeching to a halt".

For well over a century, the western world has subscribed to a way of thinking known as analysis (Ackoff, 1995). In analysis, in order to understand things—a concept, a product, a law, an organization, human body—we break it into pieces and study the pieces *separately*. This approach tends to overlook the interdependencies and connections between the constituent parts, which are responsible for dynamic change in systems, say aging in human body.

On the one hand, this "divide and conquer" approach has served us well in the past. It has enabled efficient mass production of goods and services, which has brought a new social and economic order creating unprecedented wealth and standards of living in the industrialized world. On the other hand, this thinking has resulted in over-fragmentation and has created complexity and cross-purposes within organizations.

In the early part of the 20th century, a new breed of scientists, in particular quantum physicists such as Werner Heisenberg (Uncertainty Principle) and Norbert Weiner (Cybernetics) began to challenge the Newtonian precepts (Zohar & Marshal, 1994). In 1968, Austrian biologist Von Bertalanffy (1968) published "General Systems Theory"—a major departure from conventional fragmentation in science. Similarly, Jay Forrester of MIT introduced and demonstrated the applications of feedback theory in organizations (Forrester, 1958). Forrester's seminal work marks the birth of a new discipline known as System Dynamics. System Dynamics is concerned with applications of systems theory and computer modeling in complex problems in business, economics, and the environment. System Dynamics is the forerunner and the scientific foundation of Systems Thinking.

Today, biologist and physicists as well as social and cognitive scientists are working on new fields such as complexity and network theory, and Gaia theory. These emerging fields come under the broader umbrella of "systems theory" or "living systems" and "they are working in the systems sciences and are contributing to advancing the integrated, systemic understanding of life" (Capra, 2007).

BACKGROUND

The major intellectual and philosophical precepts that form the bedrock of our modern society, such as free-market economics, mass production, division of labor, and scientific management embed the following machine age characteristics (Zohar et al., 1994):

- The hierarchy
- Need for certainty, stability, and the absolute
- Treating organizations and the society as consisting of isolated, separate and interchangeable parts
- Relationships based on conflict and confrontation (rationality and self-interest)
- Desire for control and bureaucratic methods
- Persistence of "single points of view" leading to friction and polarisation
- Over-emphasis on specialist expertise, leading to fragmentation and loss of relevance

Machine-age thinking, still prevailing today, is based on the following notions, that:

- Complete understanding of the universe is possible
- All relationships can be described through simple (linear) cause-and-effect
- The world could be understood through analysis (breaking the wholes into pieces)

SYSTEMS THINKING

Systems Thinking (ST) is a discipline for understanding the dynamics of change and complexity underlying business, economic, scientific and social systems. Systems Thinking has three distinct but related dimensions: paradigm, language, and methodology. These dimensions are outlined next (Maani & Cavana, 2007):

- *Paradigm*: Systems Thinking is a way of thinking about the world and relationships. This paradigm relates to the dynamic relationships that influence the behaviour of complex systems. A number of expressions that we use in daily language reflect the Systems paradigm—vicious/virtuous cycle, ripple effect, snowballing, spiral effect, domino effect and chronic behaviour are among these.
- *Language*: As a language, Systems Thinking provides a tool for understanding complexity and group decision-making. The Systems Thinking language is known as Causal Loop Diagrams.
- *Methodology*: Systems Thinking provides a sophisticated computer modeling technology and associated learning environments for group interactions and learning.

Systems Thinking and the Internet

For centuries, knowledge was the preserve of the aristocrats and the clergy who controlled it to dominate and manipulate the masses. In the past century, the "knowledge" privilege extended to the teacher, the manager, and the boss who assumed this as part of their role and superiority. This knowledge divide, for its part, has strengthened the hierarchy and to some extent has widened the gap between the haves and have-nots.

In the past two decades, two movements have had a profound influence on the way we learn, think, communicate and do business—the Internet and Systems Thinking. Both are grounded in science and technology and complement each other in principle and practice. While one has become a daily necessity, the other is coming out of obscurity. The Internet was developed in military and academic quarters in the late 1960's. In the nineties, the Internet emerged in the public domain and rapidly became a mass movement. Today, the Internet is the engine driving the economy, globalization and convergence of various markets, services and industries (Query & Jin, 2003).

Systems Thinking also originated in scientific centers in the 1950's and is now growing rapidly in appeal and applications. It offers a way of thinking based on the primacy of the "whole" and relationships. Systems Thinking deals with hidden complexity, ambiguity, and mental models. It provides tools and techniques to leverage change and to create lasting interventions (Maani, 2001).

Although they may be regarded as purely technical advances, both Systems Thinking and the Internet challenge the age-old paradigms and the ways information and knowledge are disseminated. At a more fundamental level, they challenge the hierarchy and authority, power and leadership. In essence, the Internet has ushered in a new culture, social movements, and "new politics" around the globe (Webster, 2001). Through its unimpeded access and reach, the Internet has in effect brought down the boundaries that define business, trade, and even nationhood. For example, today, Facebook, an Internet portal, has over 100 million members—as a "nation" it would be the eighth largest "country" in the world (Bessant, 2007).

Likewise, Systems Thinking, through its unifying and compelling scientific principles, breaks down the superficial dichotomies of the whole vs the part, the individual vs. the collective, integration vs. autonomy, growth vs. sustainability, and nature vs. progress. Together, the Internet and Systems Thinking can provide powerful synergies blending new concepts, tools, and technologies.

Over the past 20 years, new management concepts and models have emerged that have dramatically challenged the prevailing assumptions and practices in business and organizations. Among these—the Just-in-Time philosophy and

Table 1. Why we need Systems Thinking (Maani & Cavana, 2007)

• Increasing complexity in our lives
• Growing interdependence of the world
• Revolutions in management theories and practice
• Increasing global consciousness and yet "local" decision-making
• Need for multistakeholder decision making and consensus building
• Increasing recognition of learning as a key organizational capability

techniques, total quality management, and, more recently, Supply Chain Management and Enterprise Systems. These paradigms have progressively removed the conventional boundaries between the organization, the customer, the supplier, and to some extent the competition.

Supply chain management (SCM) is an example of the Internet—Systems Thinking synergy. The conceptual underpinning of SCM is systemic in nature in that the "business" or organizational boundaries span to cover the entire chain of supply. In this model, the supply chain participants and stakeholders regard themselves as partners and collaborators in a "whole" where their good depends on the good of the whole. This notion stands in sharp contrast to the business models preceding it, which are characterized by competition, control, and optimization of the parts in isolation. For example in SCM, Internet companies extensively use the practice of drop-shipping, where the wholesaler stocks and owns the inventory and ships products directly to customers at retailers' request (Netessine & Rudi, 2006).

Concomitantly, the reach and speed of the Internet has enabled such cross boundary and integrative business models as supply chain management and e-business, e-commerce to rapidly grow and multiply. With the daily spread of globalization and e-commence there is a dire need for reliable and efficient ways to manage, monitor, and coordinate global large-scale initiatives which contain thousands of members and hundreds of organizations located at different sites. The terms "information exchange", "compatibility", and "interoperability" have become ubiquitous in this environment" (Badir, Founou, Stricker & Bourquin, 2003). To survive, companies will have to constantly develop innovative ways for selecting and managing information. These can be accomplished efficiently and effectively with Web-based models that automate the business processes and extend them beyond the organization.

The astronomical growth of the Internet is epitomized in China where, as of December 2004, 94 million people had gone online, making China the second largest Internet-user market in the world, behind only the US (Zhu & Wang, 2005). This, for example, has enabled over 50% of domestic Chinese insurance companies to offer policies on line (Query & Jin, 2003). However, the impact of the Internet goes well beyond business and economics and into cultural and political realms. This has sparked some serious questions in recent literature, such as, "Can the Internet open up a new public sphere that will foster democracy in China? To what extent will Internet communication erode the social and cultural fabrics and affect Chinese society negatively?" (Kang, 2004).

In summary, the Internet has steadily pushed our thinking and practice closer to a systemic model of the world and the enterprise. At a conceptual level, the accelerated interdependence and interconnectedness of the global economy, trade, and governance calls for a systemic (holistic) view of the global community, the environment, and nationhood. To this end, the Internet has substantially shifted the distribution of economic activity around the world through its impact on communication, information, uniform pricing, and marketing strategies that target individuals rather than regions, hence creating in essence a new "economic geography" (Anderson, Chatterjee & Lakshmanan, 2003) and "Invisible Continent" (Ohmae, 2004). Thus, Systems Thinking and the Internet can play a synergistic role in the transformation of social, cultural and political paradigms toward a unified global socioeconomic order. Given recent trends, this appears not only plausible but also inevitable.

FUTURE TRENDS

Agent based modeling (ABM) is an emerging technology, which draws its theories and techniques from complexity science (Rothfeder, 2003). While Systems Thinking/System Dynamics (SD) and Agent-Based Modeling (ABM) use different modeling philosophies and approaches, they can be used complementarily and synergistically.

System Dynamics focuses on modeling known structures (i.e., relationships, policies, strategies) that underlie behavior of systems. This may be viewed as a weakness of system dynamics approach in that behavior is assumed to be solely a function of structure (model relationships defined a priori). In contrast, in ABM, organizations are modelled as a system of semiautonomous decision-making elements—purposeful individuals called agents. In ABM, each agent individually assesses its situation and makes decisions based upon value hierarchies representing goals, preferences, and standards for behavior. Thus, macrobehavior is not modeled separately but emerges from the microdecisions of individual agents. In other words, in agent based modeling, "emergent" behavior unfolds as a result of agents' interactions. This is a key difference between the two approaches.

Recent advances in video game technology allow the development of multiagent, artificial "society" simulators with capabilities for modeling physiology, stress and emotion in decision-making (Silverman, 2002). While system dynamics acknowledges the critical role of individual and organizational mental models (e.g., motivations, values, norms, biases, etc.) it does not explicitly model them. System Dynamics utilizes factual data or "cold knowledge" and does not take into account decision makers "mood". In contrast, ABM attempts to capture warm knowledge, representing emotional and human context of decision-making.

At the simplest level, an agent-based model consists of a system of agents and their relationships. This new approach enables superior understanding of the complexity in organizations and their relevant business environments. This in turn provides an opportunity for new sophistications in learning environments that enhances decision-making.

Experience with agent-based modeling shows that even a simple agent-based model can exhibit complex behavior patterns and provide valuable information about the dynamics of the real world system that emulates them. "Agent-based complex systems are dynamic networks of many interacting agents; examples include ecosystems, financial markets, and cities. The search for general principles underlying the internal organization of such systems often uses bottom-up simulation models such as cellular automata and agent-based models" (**Grimm,** Revilla, Berger, Jeltsch, Mooij & Railsback**, 2005)**.

Despite their differences, SD and ABM can be used in a complementary fashion. Both ABM and SD are powerful tools for transforming information into knowledge and understanding leading to individual and group learning. Nonetheless, the transition from knowledge to understanding may not be immediate or transparent. This requires a deep shift in mental models through experimentation and group learning which is enabled by both SD and ABM.

CONCLUSION

The contemporary era is characterized by interdependence, change, complexity, speed, and a rapid breakdown of social and political norms. In order to succeed (or even survive) collectively, we need new paradigms and tools. According to Ackoff (1995), there is a progression from information to knowledge, to understanding and ultimately to wisdom. The Internet has facilitated part of this progression, namely, instant and free access to information and an un-impeding portal for global communication transcending conventional barriers of the past.

Systems Thinking, on the other hand, addresses a compelling need for a deeper understanding of the ever-increasing complexity and interdependence of the world and the human society. Systems Thinking and System Dynamics tool unravel forces shaping this interdependence. In social, organizational context, Systems Thinking makes it possible to:

- Examine and foresee the consequences of policy and strategic decisions;
- Implement fundamental solutions to chronic problems;
- Avoid mistakenly interpreting symptoms as causes;
- Test assumptions, hypotheses, and scenarios;
- Reconcile the dilemma of short vs. long term interventions
- Harmonize divergent mental models
- Create win-win solutions for complex multistakeholder problems
- Implement change management without adverse side effects.

The most urgent and fundamental dilemmas that we face today in the organization and in society alike require a collective will that comes from a deeper understanding of our place in history and our collective destiny towards an organic unity. Systems theory provides a scientific way of thinking about the world and relationships. It offers the art and the science of seeing the tree *and* the forest – it reconciles the superficial dichotomy between the individual and the collective and focuses attention on the common good. The principles of systems theory, collectively, provide a paradigm and a language for deeper understanding of the chronic issues besetting our times. Used appropriately and with care, they can transform our age-old views and assumptions and move us towards a shared understanding. In this respect, the challenge for the Internet is to transcend from an information repository and a business tool and become a channel to enhance understanding and goodwill. The evidence to date is promising—the Internet continues to break down national, political, and social barriers and boundaries thus bringing us closer to a holistic and systemic understanding of our global society.

REFERENCES

Ackoff, R. A. (1995). In *Proceedings of the Systems Thinking in Action Conference*, Boston

Ackoff, R. A. (1999) Re-creating the corporation – A design of organizations for the 21st century. Oxford University Press.

Anderson, V. & Johnson, L. (1997). *Systems thinking basics*. Pegasus Communications, Inc

Anderson, W. P, L. Chatterjee, T. R. Lakshmanan, (2003). E-commerce, transportation, and economic geography. *Growth and Change, 34*(4), 415.

Argris, C. (1990). *Overcoming organizational defences - Facilitating organizational learning*. Boston: Allyn and Bacon.

Badir, Y. F, Founou, R., Stricker, C., & Bourquin, V. (2003). Management of global large-scale projects through a federation of multiple web-based workflow management systems. *Project Management Journal*, *34*(3), 40.

Bessant, J. (2007). Meeting the innovation challenge. In *Proceedings of the Keynote Plenary, 21st ANZAM Conference 2007*, Sydney.

Checkland, P. (1981). *Systems thinking, systems practice*. John Wiley.

Capra, F. (2007). Complexly and life. *Systems Research and Behavioral Science on Complexity, Democracy and Sustainability, 24(*5), 475-479 [Special Issues].

De Geus, A. (1995). In *Proceedings of the Systems Thinking in Action Conference*, Boston

Forrester, J. (1958). Industrial dynamics—A major breakthrough for decision makers. *Harvard Business Review, 36*(4).

Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W. M., Railsback, S. F., Thulke, H. H., Weiner, J., Wiegand, T., DeAngelis, D. L. (2005). Pattern-oriented modeling of agent-based complex systems: Lessons from ecology. *Science, 310*(5750), 987 – 991.

Kang, L. (2004). The internet in China: Emergent cultural formations and contradictions. Globalization and the Humanities

Kauffman, D. L., Jr. (1980). *Systems one, An introduction to systems thinking*. The Innovative Learning Series, Future Systems, Inc

Maani, K. (2001). Systems thinking and the internet. *Internet management issues*. Hershey, PA: Idea Group Publishing.

Maani, K. & Cavana, R. (2007). *Systems thinking and modeling – Managing change and complexity* (2nd ed.). Prentice Hall

Maani, K., Pourdehnad, J., Sedehi, H. (2003). Integrating system dynamics and intelligent agent-based modelling – Theory and case study. In *Proceedings of the Euro INFORMS*, Istanbul.

Netessine, S. & Rudi, N. (2006). Supply chain choice on the internet. *Management Science, 52*(6), 844-864.

Ohmae, K. (2004). *The invisible continent – Four strategic imperatives of new economy*. HarperCollins/Nicholas Brealey Publishing

Query, J. T. & Jin, Z. (2003). Walking tiger. *Best's Review, 104*(4), 119.

Rothfeder, J. (2003). Expert voices: Icosystem's Eric Bonabeau. *CIO Insights*

Senge, P. (1990). *The fifth discipline – The art and practice of the learning organization*. Doubleday/Currency

Senge, P. (1992). Building learning organizations. *Journal for Quality and Participation*.

Shiba, S., Walden, D., & Graham, A. (1994). *A new American TQM*. Productivity Press

Silverman, B. G. et al. (2002). Using human models to improve the realism of synthetic agents. *Cognitive Science Quarterly, 3*

Toffler, A. (1991). *The third wave*. Bantam Books

Von Bertalanffy, L. (1968). *General system theory: Foundations, development, applications*. New York: George Braziller, Inc.

Webster, F. (2001). *Culture and politics in the information age: A new politics?* New York: Routledge.

Zhu, J. H. & Wang, E. (2005). Special issue: Transforming China. *Communications of the ACM, 48*(4), 49 – 53.

Zohar, D. & Marshal, I. (1994). *The quantum society*. Morrow Press

KEY TERMS

Balancing (Counteracting) Feedback: A systemic pattern that is responsible for stability, balance and control. It represents adjusting, correcting and counteracting processes that resist, slow down or impede change and growth.

Causal Loop Diagram (CLD): A tool that captures the causal interrelationships amongst a set of variables. CLDs reveal systemic patterns underlying complex relationships and highlight hidden causes and unintended consequences.

Flow: or rate is the amount of change in a variable over time. Flow represents the change in the status of a variable over a specified time unit.

Leverage: Leverage knows which actions may yield long lasting outcomes. It knows where and when to intervene/influence a system to gain long lasting desired change using minimal effort and energy.

Reinforcing (Positive) Feedback: A systemic pattern that represents growth and self-feeding processes. It explains the dynamics underlying vicious and virtuous cycles with downward and upward spiral effects.

Stock: is the accumulation of a variable such as asset, debt, energy, morale, and reputation. Stock represents the status of a variable at a given point in time, that is, a snapshot of reality.

System Dynamics: Is a scientific tool which embodies principles from biology, ecology, psychology, mathematics, and computer science to model complex and dynamic systems.

System: Is a purposeful entity whose parts interact with each other to function as a whole. Thus, a system is not the sum of its parts—it is the product of their interactions (Ackoff, 1993). A system can be part of a larger system.

Systems Archetype: A generic pattern of relationships that occurs in a wide range of systems and circumstances, natural, biological, political, social, and economic—a powerful tool for seeing high-level dynamics.

Systems Delay: Is the time lapse between action and response. Delays often destabilize the system and slow a system down from reaching its goal. Systems delays often mask anticipated outcomes as well as unintended consequences of actions as the intervening time lapse is often longer than expected.

Systems Thinking: Is thinking holistically and conscientiously about the world by focusing on the interaction of the parts and their influence within and over the system.

Tacit Knowledge and Discourse Analysis

Michele Zappavigna-Lee
University of Sydney, Australia

Jon Patrick
University of Sydney, Australia

INTRODUCTION

Much of human experience is *below-view*, unattended to as we operate in the world, but integral to our performance as social creatures. The tacit knowledge involved in our practice allows us the experiential agility to be at once efficient and creative, to assimilate the novel and the familiar: in essence, to develop expertise. The possessors of skilful practice, the artisan, the witchdoctor or the physician, have occupied a position of both importance and mystery in most cultures since ancient times. Our interest over the ages in such hidden knowledge has caused us to mythologise expertise, placing it beyond the common by constructing it as unspeakable. Thus, in contemporary times it is not surprising that the dominant research perspective on tacit knowledge maintains that it is ineffable, that is, tacit knowledge cannot be understood by looking at what and how people communicate verbally. Indeed the word tacit has its origins in the Latin, *tacitus*, meaning silent.

As information technologies have begun to alter the way in which we think about our own processes while looking for ways to automate and retain our practices, we have been compelled to consider how the experience of the artisan mentioned above can engage with the constraints of the computational world. Capturing and sharing tacit knowledge has thus been a consistent problem in information systems and knowledge management research (Boisot, 1995; Nonaka & Takeuchi, 1995; Tsoukas, 2002; Wenger, 1998). Polanyi's Theory of Tacit Knowing (TTK) is the dominant theoretical perspective in this research. In this theory, Polanyi (1958) suggests that tacit knowledge is inherently personal, underlying our ability to perform tasks we find difficult to explain, such as facial recognition. Concepts in TTK have been made available to the information systems (IS) community largely through the work of Nonaka & Takeuchi (1995) who reinterpreted the theory, precipitating research directions in information systems that are misaligned with Polanyi's theses. A notable example is the movement in IS research to differentiate tacit and explicit knowledge (Johnson & Lundvall, 2001). In contradistinction, Polanyi asserts that explicit knowledge cannot be adequately separated from its tacit coefficient:

> *Now we see tacit knowledge opposed to explicit knowledge; but these two are not sharply divided. While tacit knowledge can be possessed by itself, explicit knowledge must rely on being tacitly understood and applied. Hence all knowledge is either tacit or rooted in tacit knowledge. A wholly explicit knowledge is unthinkable. (Polanyi, 1969, p. 144)*

The emphasis in information systems research is typically on converting tacit knowledge into explicit knowledge (Hershel, Nemati, & Steiger, 2001). Attention is also given to setting up a dichotomy of tacit and explicit knowledge in terms of articulation (can it be carried in language?), codification (can it be turned into an artifact?) or judgment (is it objective or subjective?).

This article is structured to critique the dominant position in information systems research that tacit knowledge is ineffable. The background section provides an introduction to the extensive interdisciplinary literature on tacit knowledge, providing context for the subsequent section that deconstructs the assumptions that this literature makes about what it means to, in Polanyi's (1966, p. 4) terms, "know more than we can tell." To conclude, the role of linguistic and semiotic analysis in realising the growing trend toward theorising "community knowing," rather than knowledge as an artifact, is suggested in the final sections.

BACKGROUND: UNDERSTANDING TACIT KNOWLEDGE, AN INTERDISCIPLINARY PURSUIT

Both prior and adjacent to the popularisation of Polanyi's TTK, there has been extensive interdisciplinary research in tacit knowledge and a very large body of research looking at the implicit, situated nature of practice. These disciplines include philosophy, psychology, linguistics, semiotics, sociology, history, philosophy of science, and, most recently, knowledge management. A theme running through all these domains is uncovering the below-view. For example, Freud's (1970) theory of psychoanalysis and his corresponding concept of the unconscious locates a significant aspect of human experience below the view of awareness. Similarly, a movement in psychology investigates implicit learning

(Reber, 1993), that is, learning that is below-view in the sense that it occurs without the attention of the subject on the process of learning. Recently, there have been attempts by psychologists to create psychological metrics for tacit knowledge. Wagner & Sternberg (2000) develop a method for measuring tacit knowledge in psychology and business management. Their "Tacit Knowledge in Management" (TKIM) inventory seeks to "identify individuals whose 'street smarts' indicate the potential for excellent performance in managerial and executive careers" (Wagner & Sternberg, 1991, p. 1). Conceiving of tacit knowledge as a facet of *practical intelligence*, they construct a triadic view of tacit knowledge as "managing self," "managing tasks" and "managing others," and test subjects in each area through multiple-choice responses to scenarios.

In the social sciences, the focus has been on identifying the implicit subject positions which have been naturalised by culture (Bernstein, 1971; Bourdieu, 1990). This naturalisation means that these positions remain below scrutiny. This is part of a research program that suggests that our experience is constructed through implicit social processes and parallels a body of research in psychology that claims that social behaviours are encoded automatically and without intention (Bargh, 1999). Bourdieu (1990) suggests that the individual internalises the cultural habitat in which they reside, their *habitus*. This means that they form dispositions to behave and construe their experience in certain ways. The acquisition of these structural constraints is a process of acculturation into specific socially-established groups or classes (Bourdieu, 1990, p. 130) and is akin to Bernstein's (1971) notion of *code orientation*. This is a different conceptualisation to the structuralist position, which argues that we follow unconscious rules in enacting our practice. Instead, it is a view of socially constructed dispositions that are inscribed in the habitus but which may shift with changes in context: "Agents to some extent fall into the practice that is theirs rather than freely choosing it or being impelled into it by mechanical constraints" (Bourdieu, 1990, p. 90).

Research in the history and philosophy of science examines science as a discipline that operates to efface the tacit component of its practice in order to maintain an ideology of objectivity (Collins, 1974, 2001a; Kuhn, 1962; Ravetz, 1971; Turner, 2001). In their investigation of science as a social practice, theorists in this domain have sought to reveal the tacit nature of the methods scientists employ. In fact, Polanyi's model of tacit knowledge itself was directed at critiquing the scientific method and to providing "a stable alternative to…[science's] ideal of objectivity" (Polanyi, 1966, p. 25). It provided a conceptual basis for understanding scientific activity as the practice of a craft (Ravetz, 1971). This is part of a program which deconstructs what Schuster (1984) refers to as the mythic construction of method. Method discourses in science typically try to generate the idea that they are systematic, explicit and objective (Collins, 2001a, 2001b; Ravetz, 1971). As such, they present themselves as incommensurable with other practices such as astrology, which they claim to be pseudoscience. Here we have an example of an institutional discourse, science, operating to efface the role of tacit knowing in the production of knowledge.

IS TACIT KNOWLEDGE INEFFABLE?

The attribute most consistently ascribed to tacit knowledge across the disciplines is ineffability (Baumard, 1999; Collins, 2001a; Nonaka & Takeuchi, 1995; Reber, 1993). The strong position is that tacit knowledge cannot be articulated in any linguistic form, while the weak position holds that it is difficult to articulate. Polanyi's (1966:4) widely cited suggestion that "we know more than we can tell" asserts the epistemological significance of tacit knowing in terms of its ineffability. In assessing this proposal, it is important to consider what it means "to tell." If telling means making explicit, codified artifacts that are directly transferred to the mind of the listener, then this kind of telling is not a possible means of exposing tacit knowledge. However, if we allow that telling involves processes of which the speaker is not necessarily aware and which are, in turn, subject to both unconscious and conscious interpretation by the listener, linguistic structure is reinstated as relevant to understanding tacit knowledge. These below-view processes are akin to Peirce's notion of the interpretant in semiosis, introduced in the previous section.

Thus, it appears that Polanyi's statement needs to be refined. We know more than we can tell only if we think about telling as making explicit knowledge. Such an assumption utilises an impoverished model of communication. This model, referred to as the mathematical model of communication (Shannon & Weaver, 1949), presupposes that meaning in communication is absolute and, as such, may be seamlessly transferred from the mind of the speaker to that of the listener. It applies what Reddy (1979) terms the *conduit metaphor*, that is, the notion that words are boxes with meanings inside that are unpacked by the person to which they are directed. Reddy (1979:287) argues that the metalingual resources of English privilege this kind of view, as the following examples suggest:

Whenever you have a good idea, practice capturing it in words

You have to put each concept into words very carefully

Just as in uttering the sentences above we are unlikely to focus on the presuppositions about communication that they presume, when we speak, that which we utter cannot be viewed as an overt object. We may well articulate what we know implicitly through patterns and features of language

to which we do not directly attend. This is an argument that articulation does not produce a form that by definition is explicit, or in alternative terms, that articulation is not the equivalent of codification. However, many studies in information systems research equate these two modes of meaning-making.

There is a substantial tradition within psychotherapy that has approached language as a way of understanding a person's unconscious experience (Ferrara, 1988, 1994; Freud, 1960; Labov & Fanshel, 1977; Lentine, 1988; Parker, 1995; Pittenger, Hockett, & Danehy, 1960). This notion is further specified in linguistics by Halliday & Webster (2002, p. 303) who assert the significance of the relationship between grammar and the unconscious. Meaning-making with grammar is, according to this view, implicit meaning-making:

Conscious language achieves its creative force mainly by lexical means; and lexical items are semantically close to experience. Unconscious language depends much more for its creative force on grammar – and grammatical categories are far removed from experience. (Halliday & Webster, 2002, p. 303)

This appears in accord with the argument that description of the grammatical features in a subject's discourse will give us insight into the nature of their unconscious experience. It follows from this, that if the features of a subject's grammar involving meanings that are effaced are explicated, then the knowledge to which they point may be elicited. In looking at implicit meaning, Hasan et al. (1996) introduce the notion of implicit style in discourse. They give the example of the clause, "they will," which they argue is an example of maximal implicitness as it does not contain a string that is not implicit. The clause raises the questions: whom "will" and what "will" they? Hasan et al. (1996, p. 194) argue that we may distinguish between implicit and explicit ways of saying and that, when an implicit style is adopted by a speaker, "precise meanings become available only if certain additional conditions are met; the average working knowledge of the language is necessary but not sufficient." It is at this point that we require the services of a linguist.

Zappavigna-Lee & Patrick (2004) present a method for eliciting tacit knowledge from the language of interviewees through a process of directed interviews based on a linguistic model of tacit knowledge. This process centers upon the interviewer identifying semantic and grammatical features in the interviewee's language that suggest knowledge that the participant possesses, which remains "below-view." The knowledge is below-view in the sense that the linguistic-choice that the interviewee has made indicates *under-representation*. Under-representation occurs when components of knowledge are effaced in discourse, as they have been automatised by the individual. For example, the agent in a clause may be omitted. In addition to simply being left out of discourse, the knowledge may be effaced through generalisation that construes it as unavailable for deconstruction. For example, a verb may be nominalised meaning that something that was a process with component steps is rendered as a static object This means that there is less potential for these steps to be analysed. Tacit knowledge is subsidiary in the sense that we do not attend to such obfuscation. The directed interview method entails:

1. Identifying the semantic feature that suggests knowledge that is under-represented in the interviewee's discourse and important for their current knowledge management task.
2. Asking a question that elicits a more delicate response from the interviewee and which prompts them to elaborate on this feature.

In applying this interview method in an on-going case study in an Australian broadcasting organisation, Zappavigna-Lee & Patrick (2004) demonstrate that identifying such features will contribute in eliciting a more delicate description of the interviewees' meaning than a strategy based solely on eliciting content. The description is more delicate not only in the sense of being more specific lexically, but also in the sense of being increasingly precise lexicogrammatically.

In a directed interview with a senior manager in the digital media division of the organisation the interviewer noted that the manager possessed tacit knowledge that was embedded in nominalizations. For example, the manager described his division as a "service area" that "provide[s] services including IT services." "Service" is the nominalisation of a range of processes involving understanding, communicating and delivering feedback to clients. An underlying component of these processes is negotiating shared cultural experience. Through elaborating this nominalisation in the directed phase of the interview the interviewer uncovered that the manager believed that the greater the shared cultural experience and active cultural processes that he was able to foster, the greater the shared knowledge and cohesion of his employees. This information was not elicited merely by asking the manager to be more specific, as this would simply have elicited a more detailed rendering of his explicit style. This may have merely produced a taxonomy of the IT services in the organisation. Instead, embedded phenomena about the manager's beliefs and practices were uncovered by analysing his implicit style. This phenomenon is more "specific" in a particular way: it is the elaboration of parts of the interviewee's language of which they were not aware. This is an exercise below the surface of the text, and below the content plane on which most interviews are conducted. As such it involves a richer elicitation of the interviewee's experience.

FUTURE TRENDS: TACIT KNOWING OR TACIT KNOWLEDGE?

There is a growing movement in the IS literature that argues that it is not possible to capture tacit knowledge but that we should instead manage processes which facilitate its social transfer (Boland & Tenkasi, 1995; Stenmark, 2001; Wenger, 1998). This is part of a movement which acknowledges that that knowledge is human and social rather than an artefact that can be abstracted in a database (Weick, 1995; Wenger, 1998). It suggests that tacit knowledge is dynamic and carried in communities of practice (Huysman, 2004; Wenger, 1998). Rather than taxonomising tacit knowledge as if it were an object, this movement adopts a community-oriented model of knowledge management in which IT services are aimed at connecting people with relevant experts rather than attempting to externalize and codify this expertise (Swan, Newell, Scarborough, & Hislop, 1999).

The community-oriented model is aligned with the post-critical epistemological orientation of Polanyi's thesis, which deals with "knowing" rather than "knowledge;" with a process rather than an object. This conceptual position is in accord with the movement in semiotics and other disciplines concerned with theorising knowledge, such as philosophy and linguistics, away from reification: that is, away from a constituency-based view of knowledge as an object, toward a view of knowledge as dynamically produced by human subjects. Future research adopting this kind of framework should seek to employ existing tools such as linguistic analysis to understand how humans "do knowing" as opposed to "construct knowledge."

CONCLUSION

Polanyi's maxim that "we know more than we can tell" remains the dominant perspective on tacit knowledge in information systems research. However, few studies consider what it means to "tell" and the implications this has on asserting the ineffability of tacit knowledge. As this research begins to shift to considering tacit knowledge a process rather than an object and draws upon the large body of existing scholarship in other disciplines, attention should be given to what kind of meaning-making tacit knowledge involves. Tacit knowledge is potentially not as taciturn as we have assumed.

REFERENCES

Bargh, J.A. (1999). The unbearable automaticity of being. *American Psychologist, 54* (7), 462-479.

Baumard, P. (1999). *Tacit knowledge in organizations*. London: Sage.

Bernstein, B.B. (1971). *Class, codes and control*. London: Routledge and K. Paul.

Boisot, M. (1995). *Information space. A framework of learning in organizations, institutions and culture*. London: Routledge.

Boland, R. J., & Tenkasi, R.V. (1995). Perspective making and perspective taking in communities of knowing. *Organization Science, 4* (4), 350-372.

Bourdieu, P. (1990). *In other words: Essays towards a reflexive sociology*. Stanford, CA: Stanford University Press.

Collins, H.M. (1974). The TEA set: Tacit knowledge and scientific networks. *Science Studies, 4*, 165-186.

Collins, H.M. (2001a). Tacit knowledge, trust and the Q of Sapphire. *Social Studies of Science, 31* (1), 71-85.

Collins, H.M. (2001b). What is tacit knowledge? In T.R. Schatzki, K. Knorr-Cetina & E.V. Savigny (eds.), *The practice turn in contemporary theory* (pp. 107-119). London, New York: Routledge.

Ferrara, K. (1988). Variation in narration: Retellings in therapeutic discourse. In J. Baugh (ed.), *Linguistic Change and Contact*. Austin, TX: University of Texas.

Ferrara, K. (1994). *Therapeutic ways with words*. New York: Oxford University Press.

Freud, S. (1960). *Jokes and their relation to the unconscious* (J. Strachey, Trans.). New York: W.W. Norton.

Freud, S., & Strachey, J. (1970). *An outline of psycho-analysis*. New York: W.W. Norton.

Halliday, M.A.K., & Webster, J. (2002). *On grammar*. London: Continuum.

Hasan, R., Williams, G., Butt, D., & Cloran, C. (1996). *Ways of saying, ways of meaning: Selected papers of Ruqaiya Hasan*. London, New York: Cassell.

Hershel, R.T., Nemati, H., & Steiger, D. (2001). Tacit to explicit knowledge conversion: Knowledge exchange protocols. *Journal of Knowledge Management, 5* (1).

Huysman, M. (2004). Communities of practice: Facilitating social learning while frustrating organizational learning. In H. Tsoukas & N. Mylonopoulos (eds.), *Organizations as Knowledge Systems: Knowledge, Learning, and Dynamic Capabilities* (pp. 67-85). New York: Palgrave Macmillan.

Johnson, B., & Lundvall, B. (2001). Why all this fuss about codified and tacit knowledge? *Industrial and Corporate Change, 11* (2), 245-262.

Kuhn, T.S. (1962). *The structure of scientific revolutions*. Chicago, London: University of Chicago Press.

Labov, W., & Fanshel, D. (1977). *Therapeutic discourse: Psychotherapy as conversation*. New York: Academic Press.

Lentine, G. (1988). Metaphor as cooperation in therapeutic discourse. In *16th Annual Conference on New Ways of Analyzing Variation: Linguistic Change and Contact* (Vol. 30). Austin, TX: University of Texas, Dept of Linguistics.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. New York: Oxford University Press.

Parker, I. (1995). Everyday behavior(ism) and therapeutic discourse: Deconstructing the ego as verbal nucleus in Skinner and Lacan. In J. Siegfried (ed.), *Therapeutic and Everyday Discourse as Behavior Change: Towards a Micro-Analysis in Psychotherapy Process Research* (pp. 447-467). Norwood, NJ: Ablex Publishing Corporation.

Pittenger, R.E., Hockett, C.F., & Danehy, J.J. (1960). *The first five minutes: A sample of microscopic interview analysis*. Ithaca, NY: P. Martineau.

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. Chicago: U.P.

Polanyi, M. (1966). *The tacit dimension*. London: Routledge & K. Paul.

Polanyi, M. (1969). *Knowing and being: Essays*. Chicago: University of Chicago Press.

Ravetz, J.R. (1971). *Scientific knowledge and its social problems*. Oxford: Clarendon Press.

Reber, A.S. (1993). *Implicit learning and tacit knowledge: An essay on the cognitive unconscious*. New York, Oxford: Oxford University Press & Clarendon Press.

Reddy, M. (1979). The conduit metaphor. In A. Ortony (ed.), *Metaphor and Thought*. Cambridge: Cambridge University Press.

Schuster, J. (1984). Methodologies as mythic structures: A preface to future historiography of method. *Metascience: Annual review of the Australasian Association for the History, Philosophy and Social Studies of Science, 1* (2), 17.

Shannon, C.E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Stenmark, D. (2001). Leveraging tacit organizational knowledge. *Journal of Management Information Systems, 17* (3), 9-24.

Swan, J., Newell, S., Scarborough, H., & Hislop, D. (1999). Knowledge management and innovation: Networks and networking. *Journal of Knowledge Management, 3* (4), 262-275.

Tsoukas, H. (2002). Do we really understand tacit knowledge? In M. A. Lyles (ed.), *Handbook of Organizational Learning and Knowledge* . Blackwell.

Turner, S. (2001). Throwing out the tacit rule book: Learning and practices. In T. R. Schatzki, K. Knorr-Cetina & E. V. Savigny (eds.), *The practice turn in contemporary theory* (pp. ix, 239). London, New York: Routledge.

Wagner, R.K., & Sternberg, J. (1991). *Tacit knowledge inventory for managers: User manual*. The Psychological Corporation.

Wagner, R.K., & Sternberg, J. (2000). Tacit knowledge and management in the everyday world. In R. J. Sternberg (ed.), *Practical Intelligence in Everyday Life* (pp. xiv, 288). Cambridge, New York: Cambridge University Press.

Weick, K.E. (1995). *Sensemaking in organizations*. Thousand Oaks: Sage Publications.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, New York: Cambridge University Press.

Zappavigna-Lee, M., & Patrick, J. (2004). *Literacy, tacit knowledge and organisational learning.* Paper presented at the The 16th Euro-International Systemic Functional Linguistics Workshop, Madrid.

KEY TERMS

Below-View: Elements of experience which are not available to direct inspection without applying some semiotic tool such as linguistic analysis.

Conduit Metaphor: A metaphor about communication, which suggests that an addresser's ideas are objects contained in packages, known as words, that are directly sent to the addressee.

Focal Awareness: Polanyi's term for conscious perception that the individual can directly access. Contrast with *subsidiary awareness*.

Habitus: Bourdieu's term for the cultural context in which an individual resides and which influences their practice.

Subsidiary Awareness: Polanyi's term for perception to which an individual does not have direct access, as it is not part of their focal awareness. Contrast with *focal awareness*.

Tacit: From the Latin *tacitus*, meaning silent.

Tacit Knowing, Tacit Integration: Polanyi's concept of the process of implicit integration of subsidiary and focal elements by an individual. The individual attends "from" the element in their subsidiary awareness "to" the element in their focal awareness.

Tacit Knowledge: Implicit understanding of which the individual is not directly aware and which is involved in their skilful practice.

Under-Representation: Zappavigna-Lee & Patrick's term for a set of specific linguistic features that indicates the presence of tacit knowledge in an individual's talk.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2724-2729

Taxonomy of C2C E-Commerce Venues

Kiku Jones
The University of Tulsa, USA

Lori N. K. Leonard
The University of Tulsa, USA

INTRODUCTION

Commerce can be conducted face-to-face or electronically. Electronic commerce (e-commerce) refers to buyers and sellers transacting online. As Figure 1 illustrates, businesses and individual consumers participate in various forms of this commerce, for example, business-to-business (B2B); business-to-consumer (B2C); and consumer-to-consumer (C2C). B2B e-commerce is the conducting of online transactions between businesses. It is the largest form of e-commerce being practiced today. B2C e-commerce is the conducting of online transactions between a business and a consumer. It is the largest form of e-commerce being researched and the second largest form of e-commerce being practiced. C2C e-commerce is the conducting of online transactions between consumers. It has not been researched as much as B2B or B2C e-commerce, but it is steadily catching up to B2C e-commerce, in practice. For example, online auction use is expected to reach $54 billion in 2007, which is a growth rate of 33% compounded annually since 2002 (Johnson, 2002). Even though more transactions are occurring, research in C2C e-commerce has not kept up in this growing field. Perhaps the reason for this lack of synchronization is that researchers are unaware of the many venues in which C2C e-commerce can and is being conducted.

The most cited and researched example of C2C e-commerce is through online auctions such as eBay. However, there are many other venues for conducting C2C e-commerce that should be explored that may not necessarily be apparent. As demonstrated by existing research in C2C e-commerce, researchers have focused their studies on venues in which C2C e-commerce is the specified purpose of a site. In addition to these venues, C2C e-commerce can be facilitated in places such as online communities, Web-based discussion forums, consumer blogs, and chat rooms. Each of these venues needs to be explored before a complete representation of C2C e-commerce can be made. Differences in how a consumer views each of the C2C e-commerce venues may be expected to alter the factors affecting his/her determination to participate in C2C e-commerce in the various venues.

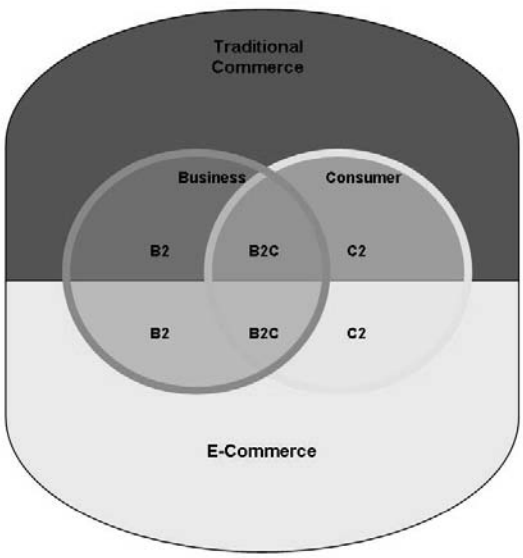
To help researchers in structuring the potential venues and classifying factors affecting participation, this article presents a taxonomy of the C2C e-commerce venues and a description of the types of venues found in the categories. The next section provides a background discussion of each venue discussed in the taxonomy. Following that will be the presentation of the taxonomy. Future trends and conclusions are provided at the end of the article.

BACKGROUND

C2C e-commerce has been examined in terms of trust, reputation systems, and value in communities. However, the venues for conducting C2C e-commerce have not been explored, nor have the multitude of venues available to conduct C2C e-commerce been recognized. This section will present potential C2C e-commerce venues and the relevant research to date for each venue.

Online auctions have been heavily researched in the literature. An online auction is designed to allow consumers to buy and sell from one another in a structured environment.

Figure 1. Commerce channels



In this venue, payment and product exchange mechanisms are established. One area of concern in online auctions is the reputation system. Many researchers have examined the impact of a seller's reputation (given that he/she may be anonymous) on the willingness of buyers to bid on and purchase items in an online auction (Lin, Li, Janamanchi, & Huang, 2006; Melnik & Alm, 2002). Online auctions have also been researched in many other ways, such as price setting (Bapna, Goes, & Gupta, 2001), bidding strategies (Ward & Clark, 2002), and trust (Klein & O'Keefe, 1999).

Third party listing services allow consumers to post items for sale as one would traditionally post in the classifieds. Third party listing services are also well structured and appear to be an established way to conduct C2C e-commerce; however, they have not been explored in the literature with regard to selling and purchasing. Currently, only the recognition of a third party in the buying/selling process has been studied as to its impact on consumer trust (Schneiderman, 2000).

Online communities offer consumers a venue to post comments regarding a topic of interest to the community. The feeling of being part of a "community" opens the online community up to more than just topical discussions. They offer the opportunity to post items for sale based on the community interest. The "community" feeling may make the member feel more comfortable with buying a product from another community member. Online communities have been greatly researched. Areas examined the most regarding online communities are: value (Armstrong & Hagel, 1996), design (Andrews, 2002; Lutters & Ackerman, 2003), success (Cothrel, 2000), and use in health care (Leimeister, Ebner, & Krcmar, 2005). However, research has not been conducted regarding the commerce that exists in online communities.

Web-based discussion forums are an online venue that allows individuals to post information related to a particular topic. Discussion forums are similar to online communities; however, individuals can be a part of an online forum without being a member of the online community. Web-based discussion forums have been researched in relation to rules of communication (Fayard, DeSanctis, & Roach, 2004), influence on consumer purchase decisions (Dellarocas, 2006), and learning (DeSanctis, Fayard, Roach, & Jiang, 2003).

Consumer blogs are online journal-like Web sites created by users to display their personal thoughts and ideas, much like a diary, and individuals that read the blog can leave messages which make this venue highly interactive. Therefore, blogs can be used for more than displaying journal entries; they can be utilized to conduct transactions. Research related to consumer blogging has currently been completed in relation to why people blog (Nardi, Schiano, Gumbrecht, & Swartz, 2004) and protecting bloggers (Robben, 2006), but blogging has yet to be studied in relation to C2C e-commerce.

Chat rooms offer a locale for individuals to meet and "chat" as often as they would like in real time. Chat rooms also offer the opportunity for consumers to meet and establish how a sale will take place; however, this aspect has yet to be explored in the literature. Chat rooms have been researched regarding information exchange (Shoham, 2004) and promotional chat where word-of-mouth and advertising is used (Mayzlin, 2006).

While research concerning commerce in these venues is lacking, anecdotal evidence suggests that all of these venues can be utilized to conduct C2C e-commerce. In order for researchers to fully study the venues, a taxonomy must first be established for guidance. The next section develops a taxonomy for C2C e-commerce venues.

C2C E-COMMERCE VENUES TAXONOMY

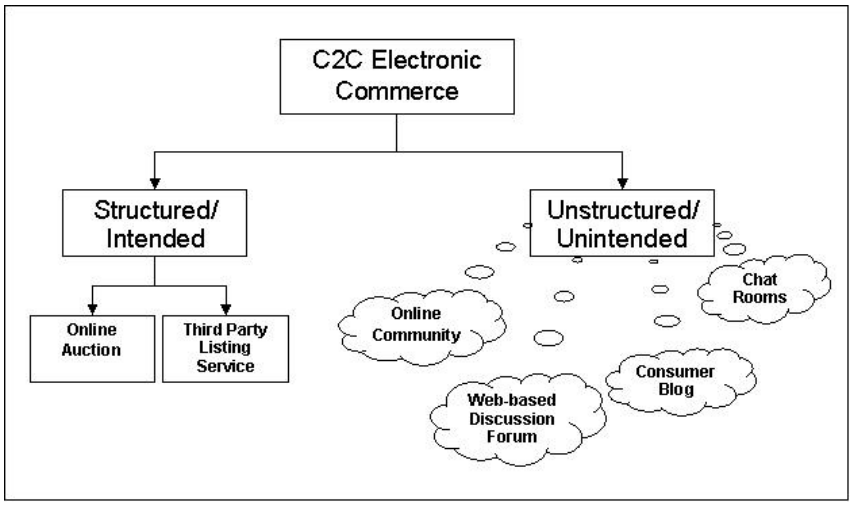
Taxonomies can help to better detail a particular phenomenon. When an area of research begins to blossom, it can be difficult to understand how research areas fit together. A taxonomy can help to structure current research, to identify holes in the existing literature for future research, and to build a roadmap for a given area of research. No such taxonomy exists for any aspect of C2C e-commerce. Without this valuable tool, researchers may find it difficult to see the connections among the current C2C e-commerce research and to develop future C2C e-commerce projects. This article develops a taxonomy of the venues in which C2C e-commerce can take place. Defining the characteristics of these various venues will help researchers to classify aspects of their C2C e-commerce research and begin to provide prescriptions to consumers participating in these various venues.

C2C e-commerce venues can broadly be broken into two main categories: Structured/Intended and Unstructured/Unintended (see Figure 2). Each of these two main categories can contain numerous venues where C2C e-commerce can be conducted. Below is a description of each of these categories.

Structured/Intended

The structured/intended category is made up of C2C e-commerce which is performed in a venue set up specifically for C2C e-commerce. For example, online auctions (e.g., eBay) and third party listing services (e.g., Half.com) are intended to facilitate C2C e-commerce. These venues may have other features available on the sites, but the main purpose for the site itself is to facilitate the exchange of goods and services. In order to participate on these sites consumers are required to adhere to various standards set forth by the third party hosting the site. The third party enforces its right to restrict consumers from participating if they do not adhere to the standards. These venues may or may not require payment

Figure 2. C2C electronic commerce venues taxonomy



from the seller in order to list on their site. The company governing the site provides consumers with information regarding what is and is not allowed to be exchanged on the site. The third party continually monitors the site to ensure that proper activities are being conducted. For example, the third party ensures that the buyer and seller of a transaction complete the given transaction under the terms agreed upon by the two parties. In many cases, the third party provides a guarantee or insurance to the buyer/seller.

These venues are very structured with formalized methods for the exchange of goods and services to and from consumers. Consumers are able to search the sites for specific products they wish to purchase or simply browse merchandise by category. They are then able to see the prices (or, in cases of auctions, the bids) of each item, description, method of payment accepted, and in most cases, pictures of the product. After a transaction takes place, consumers are able to rate one another regarding the transaction. The buyer's/seller's ratings are grouped together to provide a consumer with information regarding the number of positive and negative ratings received. This provides additional structured information for consumers to review and compare to other buyers/sellers.

Many factors can affect a consumer's willingness to participate in C2C e-commerce. For example, a consumer may feel that making purchases from consumers rather than businesses puts them at more risk of receiving defective products. C2C e-commerce conducted in the structured/intended venues may help to alleviate some of this feeling of risk by providing a third party to govern the transactions. While the consumer is still ultimately purchasing from another consumer, the structured and formal process of the venues in this category may provide more of a feeling of purchasing through a business entity.

Unstructured/Unintended

The unstructured/unintended category is made up of C2C e-commerce performed in a venue which was not created for the purpose of facilitating C2C e-commerce. For example, chat rooms are created for various purposes, such as exchanging information regarding a particular topic (more examples of venues in this category are discussed in the Background section and represented in Figure 2). These venues, while unintended, can still facilitate C2C e-commerce. To demonstrate, consider a participant in a "horror movies" themed chat room who entered the chat room to simply discuss the overall theme. While this participant is chatting in the room, the participant determines he/she doesn't have a copy of the current movie being discussed. In turn, he/she makes a request to the other participants to purchase a copy of the movie. Another participant indeed has an extra copy and informs the first participant of the price and acceptable method of payment (e.g., money order). Additional information such as name and address are then submitted by each participant. The payment and movie are then exchanged via postal mail. While the delivery of the payment and movie are not done online, the transaction was discussed, decided, and completed (in terms of the exchange of address information) through the chat room. So, while the intention of the chat room was not to facilitate C2C e-commerce, it was certainly a viable venue to conduct such a transaction.

Transactions performed under the unstructured/unintended category are done in an informal manner. While these venues can be hosted by a third party, there are no governing bodies to ensure that the transaction is completed as agreed. If one party submits payment and then does not receive the product/service, there are no guarantees that the individual will receive restitution. There are also no formalized payment methods set up on the venue itself. This may limit the ways in which consumers can send/receive payment. Compared

to the structured/intended venues, consumers may find it more difficult to search for items they wish to purchase. There are also no formalized rating systems for consumers to reference when determining if they want to transact with another consumer. Since the transactions in these venues are unintended, it may take longer for a consumer to receive information (i.e., description, price, comparison prices of other similar items, shipping cost, acceptable payments, etc.) regarding a particular product in order for the seller to gather and submit this information.

With this lack of structure and formalization may come a sense of distrust, which needs to be considered before a consumer feels comfortable in transacting with other consumers. Researchers will want to determine if there are differences not only between structured/intended and unstructured/unintended categories, but also between the various venues in the unstructured/unintended category regarding trust. For example, a consumer may feel more comfortable transacting with someone from their online community rather than someone who posted a product for sale on his blog.

FUTURE TRENDS

As C2C e-commerce continues to increase, so will the venues in which the phenomenon is occurring. Researchers will need to continue to study each aspect of these venues in order to determine the factors that will affect consumers in these areas. This article mainly discussed third party hosted sites with regard to both structured/intended and unstructured/unintended categories (with the exception of consumer blogs). One future trend might be that consumers begin holding online auctions on their own sites. In these cases, only the consumer owning the site would list items for sale. While the site would certainly contain aspects of the structured/intended (i.e., formality in auction, listing of payment methods, etc.), it would also have aspects of the unstructured/unintended category as well (i.e., any guarantees are only given by the consumer himself, and are not accountable for any standards set by a governing agency). In addition, consumers may wish to hold reverse auctions on their own sites. This, too, would fall somewhere between structured/intended and unstructured/unintended. In these cases, researchers will need to determine how that influences the factors affecting a consumer's decision to participate. Also, social networking sites, such as MySpace.com and FaceBook.com, will be used more for C2C e-commerce, if not already. Many businesses (B2C e-commerce) are currently advertising in this manner.

There is much to be done in the research area of C2C e-commerce. One factor of particular interest is trust. Many research studies have been conducted regarding trust on the Web, trust in B2C e-commerce, and related areas; however, little has been done in the area of C2C e-commerce trust beyond looking at online auctions. Additionally, satisfaction with C2C e-commerce should be explored, as well as satisfaction with C2C e-commerce venues, and how these venues transform from an area of discussion to an area for conducting commerce. The emotional draw of one consumer to another consumer in online commerce should be fully explored, examining aspects such as age, gender, and education.

CONCLUSION

The C2C e-commerce venues taxonomy provides a way for research to be organized in this valuable area. It also provides a roadmap to aid researchers in the areas not yet explored. The taxonomy includes structured/intended (online auctions and third party listing services) and unstructured/unintended (online communities, Web-based discussion forums, consumer blogs, and chat rooms) categories. C2C e-commerce venues offer consumers a multitude of ways to buy/sell from other consumers, beyond just online auctions. Researchers have yet to explore commerce in the venues classified as unstructured/unintended; however, commerce is indeed taking place in these venues. Additional research is needed in each venue to understand the factors that affect a consumer's decision to participate in C2C e-commerce.

REFERENCES

Andrews, D. C. (2002). Audience-specific online community design. *Communications of the ACM, 45*(4), 64-68.

Armstrong, A., & Hagel, J. (1996). The real value of online communities. *Harvard Business Review, 74*(3), 134-141.

Bapna, R., Goes, P., & Gupta, A. (2001). Comparative analysis of multi-item online auctions: Evidence from the laboratory. *Decision Support Systems, 32*(2), 135-153.

Cothrel, J. P. (2000). Measuring the success of an online community. *Strategy & Leadership, 28*(2), 17-21.

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science, 52*(10), 1577-1593.

DeSanctis, G., Fayard, A. L., Roach, M., & Jiang, L. (2003). Learning in online forums. *European Management Journal, 21*(5), 565-577.

Fayard, A. L., DeSanctis, G., & Roach, M. (2004). Language games in online forums. *Academy of Management Proceedings*, D1-D6.

Johnson, C. A. (2002). *The boom in online auctions*. Retrieved December 9, 2007, from news.com.com/2009-1069-962530.html

Klein, S., & O'Keefe, R. M. (1999). The impact of the Web on auctions: Some empirical evidence and theoretical considerations. *International Journal of Electronic Commerce, 3*(3), 7-20.

Leimeister, J. M., Ebner, W., & Krcmar, H. (2005). Design, implementation, and evaluation of trust-supporting components in virtual communities for patients. *Journal of Management Information Systems, 21*(4), 101-135.

Lin, Z., Li, D., Janamanchi, B., & Huang, W. (2006). Reputation distribution and consumer-to-consumer online auction market structure: An exploratory study. *Decision Support Systems, 41*(2), 435-448.

Lutters, W. G., & Ackerman, M. S. (2003). Joining the backstage: Locality and centrality in an online community. *Information Technology & People, 16*(2), 157-182.

Mayzlin, D. (2006). Promotional chat on the Internet. *Marketing Science, 25*(2), 155-163.

Melnik, M. I., & Alm, J. (2002). Does a seller's ecommerce reputation matter? Evidence from Ebay auctions. *The Journal of Industrial Economics, 50*(3), 337-349.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., & Swartz, L. (2004). Why we blog. *Communications of the ACM, 47*(12), 41-46.

Robben, P. D. (2006). Protecting the anonymity of bloggers and blog sources: Evolving case law applies old principles to new technology. *Journal of Internet Law, 10*(6), 17-21.

Schneiderman, B. (2000). Designing trust into online experiences. *Communications of the ACM, 43*(12), 57-59.

Shoham, A. (2004). Flow experiences and image making: An online chat-room ethnography. *Psychology & Marketing, 21*(10), 855-882.

Ward, S. G., & Clark, J. M. (2002). Bidding behavior in online auctions: An examination of the eBay Pokeman card market. *International Journal of Electronic Commerce, 6*(4), 139-155.

KEY TERMS

C2C E-Commerce: Consumer-to-consumer electronic commerce, that is, the buying and/or selling of goods and services from one consumer to another consumer online.

C2C E-Commerce Venue: The place or format for which C2C e-commerce can be conducted.

Consumer Blog: A Web log that is created by a user to display journal-like entries.

Online Auction: An established online venue designed to allow consumers to buy and sell from one another.

Online Community: Otherwise known as a virtual community; an online venue that allows individuals with the same interests to interact, exchange ideas, and potentially determine the exchange of products.

Third Party Listing Service: An established online venue that is similar to using the classifieds; it allows consumers to post items online for sale and for other consumers to receive seller contact and product information so that a formal exchange can take place.

Web-Based Discussion Forum: An online venue that allows for the posting of and discussion of information related to a particular topic.

Category: Social Computing

Technical Communication in an Information Society

John DiMarco
St. John's University, USA

INTRODUCTION

Historical analysis of technical communication elucidates an evolution of tools, techniques, and roles that are connected to the establishment and growth of the information society. The emergence of technical communication as an identifiable force in the matriculation of Western and Eastern information economies and societies has been quite evident. Historical literature provides accounts that point towards graphic communication taking 30,000 years to evolve (Meggs, 1998). From cave paintings during 15,000-10,000 BC, to the invention of writing with pictographs in Mesopotamia; through the evolution of illustrated manuscripts, into the invention of paper and Chinese relief printing; transitioning to the rise of late medieval illuminated manuscripts and into the breakthrough of movable type in Europe, the role of the technical communicator and the function of technical communication was born. Investigation of technical communication today, and during the last century, reveals patterns of technological, economic, occupational, spatial, and cultural developments that can be attributed to the creation of an information-driven economy and information society which relies on technical communication for stability and growth.

Gutenberg developed movable type and revolutionized communication. O'Hara (2001) makes identification that "from the fourteenth century on, the social system of science has depended on technical communication to describe, disseminate, criticize, use, and improve innovations and advances in science, medicine, and technology" (p.1). O'Hara's reference provides a clear pathway to further discussion and interpretation on the rapidly changing tools, techniques, and roles that have caused the permutation of technical communication from an original tool of science and medicine in the 1400s to an academic discipline and a universally desired societal skill set for all who engage the information society.

The purpose of this research is to identify the stature of technical communication in societies which engage heavily in information design, social technological product consumption, and publishing. This chapter addresses the past, present, and future issues, controversies, and roles that technical communication has had and will have on the information society.

BACKGROUND

On the broadest level, technical communication techniques can be defined as technical writing, research, information management, digital document design, Web design, and foremost, persuasive, action-based communication (Sheehan, 2005). A simple definition of the information society is a society that relies on information products and serves to thrive and prosper (Webster, 2002). Investigation of the history of technical communication and the birth of the information society, revealed some interesting research questions. The connection between technical communication and the information society provides a pathway to gaining a deeper understanding of the role of technical communication within modern day society. To clarify the connections, I explored and answered two research questions. First, how do the tools, techniques, and roles of technical communication enable an information society to exist? Second, are there metaphors that may be translated into predictable analogies that can be uncovered that connect technical communication as a driving force in the information society? Answering these questions revealed commonality, pattern, and evolution within the tools, techniques, and roles of technical communication as they relate to the information society.

Historical Connections

The proliferation of technical communication into disciplinary maturity has occurred over the past sixty years and has yielded academic programs and a body of innovative research (Staples, 1999). Pringle and Williams (2005, p. 362) explain that "evidence exists, in fact, that traces technical and scientific writing back to ancient times where anonymous technical writers wrote on tablets in Babylon". The rise of moveable type and the English renaissance enabled technical writing to emerge as a "by-product of print technology and literacy" (Pringle and Williams, 2005, p. 362). This historical plateau in technical communication gave rise to the 1800s and the rise of technical writing in England. Pringle and Williams (2005) explain that the first works were "books that provided instruction on performing work in a broad range of fields such as medicine, agriculture, navigation, and military science" (*Ibid.* 362). These pre-industrial revolution and pre-war events helped to structure and enable the

foundations of technical communication, technical writing, and technology.

Technical Writing Connection

"Technical" occupations in the early 1890s through the 1940s, before the title and designation of technical communication practitioner ever existed, meant serving the tools, techniques, and roles of technical writing. The evolution of technical writing as a staple skill and a staple role within the context of technical communication has bred innovation and fostered change in the information society on economic, occupational, and cultural levels. Connors (1982) describes the early years as being established by the need for engineering education soon after the passage of the two Morrill Acts in 1862 and 1877. These laws provided land grant opportunities for agricultural and mechanical colleges (A&M) which made college education possible for large numbers of people in the later nineteenth century. This economic and cultural development initiated the birth of colleges which had engineering students going through freshman composition within an English department. In the period from 1880-1905 or so, Connors (1982) explains that this education in writing was thought to be adequate for the engineer of the day. However, this was not the case and in response, many engineering schools opened separate English Departments to cater to teaching not only freshman level English courses, but also advanced upper level composition courses which addressed the needs of student engineers. The metamorphosis of Engineering and English collaboration into technical writing brought the role of technical writing into both the arenas of science and humanities, and provided new occupational purposes for people to learn and communicate. These purposes were driven by technological needs of the day such as world war and post world war development of the defense industry. O'Hara (2001) points out that specifically during the war, technical writers were used to write "standardized procedure documents, definitions, descriptions, instructions, and training". These skills were transferred into writing proposals and other military procurement documentation when war was over and the national defense sector evolved into a prominent player within the US government.

Technology Connection

After the postwar development of the transistor by Shockley, Bardeen, and Brattain (O'Hara, 2001) the United States, the most prominent ubiquitous application that has evolved from printed circuit board technology, is the personal computer (O'Hara, 2001). Having a computer on your own desktop, a personal computer, changed the nature of technical communication by complementing the technical writing skill set and knowledge base of technical communicators with the unbridled creative power of digital tools. This caused a paradigm shift in the tools, techniques, and roles of technical communication in the late 1950s through the new millennium. Hardware and software pioneered by Microsoft, IBM, Apple, Aldus (now Adobe), Macromind (formerly Macromedia, now Adobe) helped usher in technical communication as a profession and an academic discipline that went beyond technical writing and digital tools towards research, design, development, publishing, and presentation.

The tools and technologies of today are not only specific to technical communication practitioners, students, and academics, but to all people who need to live, work, and survive in the information society. Steiner (1999, p. 389), looks to Heidegger's definition of communication and argues against it by calling for innovation and individuality. Steiner (1999) explains that Heidegger finds no place for the technical specialist because he or she is part of an objective world in which there is no humanity to share. Therefore, Heidegger believes that "scientist, engineers, and technologists have no humanity to share" (*Ibid.* 389). Technical communication for the everyday person, professional, and academic will become cross functional as a commodity in the form of unlimited data content, information development based on needs and wants, and knowledge development based on problems and processes. Using e-mail, PDA's, iPods, kiosks, interactive television, video games, and personal computers have become the technical communications tools, techniques, and roles of everyday life. These everyday technical communication activities drive the information society in the form of ubiquitous technology, occupational shifts, economic realities, spatial communication methods, and cultural identity changes.

TECHNICAL COMMUNICATION WITHIN THE INFORMATION SOCIETY

Literature review of technical communication theories and histories, graphic design histories, the various schools of thought behind the information society, and professional technical communication practices were interpreted to decipher connections between past, present, and future events. These connections are presented in a series of definitions and an interpretive narrative.

Communication and Technical Communication

To begin the investigation, quantification of the term technical communication needs to be established. Severin and Tankard (1979, p. 5) reinforce the idea that the definition of communication has been extended by a plethora of scholarly articles and numerous schools of thought. To establish meanings for this communication research project, several notable

definitions on the terms, communication, and technical that stress sharing and influence will be discussed. Commonality and sharing are two active features of communication noted by scholars. These features play a role in the life of the technical communicator through creating a common thread of information for all to follow and sharing data with others so that it evolves into information or knowledge within each receiver's individual experiences. Severin and Tankard (1979) quote Alexander Gode's definition of communication "It [communication] is a process that makes common to two or several what was the monopoly of one or some" (p. 6). Schramm (1971) describes communication in terms of "sharing informational signs" (p. 13), which is quite evident in everyday technical communication activities that people share with their new media devices, such as e-mailing each other images and text, or downloading and burning mp3 files, to an Apple iPod that may be redistributed to others later.

The root of technical communication activity is purely communication driven. This includes persuasion, which has been essential to the definition of communication by many scholars. Berlo (1960, p. 16) provided a definition of communication that stated: "All communication behavior has as its purpose the eliciting of a specific response from a specific person (or group of persons)." This idea that persuasion is central to communication is consistent with the goals of technical communication being a facilitator to understanding and action.

Technical meanings serve as an adjunct to context and content, as well as to tools and techniques. Severin and Tankard (1979, p. 6) discuss the meaning of communication through looking at the etymology of the word, or the words in other languages from which it was derived. In the case of communication, the Latin word is *communicare* which means "to make common" (p. 6). Common experience of information is something that technical communicators facilitate. Technical communicators who develop detailed animations depicting natural and mechanical disasters provide visual elements that make understanding complex occurrences "common". On such example is the heavily shown 911 animation which explained to the American people and the world how the steel infrastructure of the Twin Towers in New York City failed from the massive heat brought on by raging, jet fuel driven fire. Although disturbing, the animation helped people come to a common understanding of what happened, even if they did not have a background in structural engineering.

Examination of the word "technical" takes us to the realm of several important distinctions. DiMarco (2004) explained that the modern day word "technical" comes from the Greek word *technikos* which means "of art" and "of or relating to a technique". The term also carries the meaning of "having special knowledge of a mechanical or scientific subject" (DiMarco, 2004, p. vii).

Establishing a definition of technical communication, through both academic and professional lenses, begins with looking at historical documentation from the Society of Technical Communication. The STC was born in 1971 with the merger of technical writing and technical editing organizations. STC "helps professionals design effective communication for a technical world". Understanding technological ubiquity, STC encourages the development of better-educated professionals whose jobs are to make complicated information usable by many (http://www.stc.org). In its early years, "STC was primarily made up of engineers who wrote instructions and descriptions of how electrical and mechanical products worked" (http://www.stc.org/about/history01.asp). The STC and the field of technical communication adapted as "profound change took place as the pervasiveness of technology and the need to understand it became an integral part of our everyday lives" (http://www.stc.org/about/history01.asp). The emergence of the Internet and online communication within the information society has caused the need for all people to be technical communication savvy. This socio-technological need stems from the desire and societal lifeline enabled by creation through navigation and understanding of ubiquitous and professional technologies and applications.

Richard Johnson-Sheehan's introductory textbook: *Technical Communication Today* (2005) defines technical communication based on the role of technical communication as a critical information management activity. Sheehan states, (p. 6) "Technical communication is a process of managing technical information in ways that allow people to take action".

A synthesis of the terms "technical" and "communication", based on prominent scholarly viewpoints, produces a comprehensive definition of technical communication as *techniques involving special knowledge of a mechanical or scientific subject for the purpose of sharing informational signs to make them common, to evoke shared experiences, and to persuade people to act*. A real life example of this proposed definition can be seen in assembly product instructions used when we buy a product that "assembly is required" and we are directed to "read this first". Taking furniture assembly and translating it into a series of steps supported by graphics and language which take the form of informational signs (the parts of the furniture) and make them common to all (everyone who buys the furniture), and then persuades the reader to (act) begin to put the furniture together.

We also see the common person, or someone who is assumed to not be a technical communication practitioner, use *"everyday technical communication skills & activities"* to share informational signs to evoke shared experiences. This is clear in video games. When a young teen (or someone 35 years old) plays a true to life video game such as Tiger Woods Golf, the player has the option to choose the distinct

characteristics of their virtual golfer. When the player does this, he or she is using everyday technical communication skills including menu navigation, scrolling, and visual communication to create a virtual player. That virtual player will become part of a gaming experience which was enabled by technical communication and creativity on the part of the game developers and the game players. The teenager uses the created golfer to "act out" the experience of virtually playing professional golf against modern day athletes who exhibit their own unique traits within game play.

INFORMATION SOCIETY CONNECTIONS

Many scholars cite Webster's debate that questions the notion of an information society. Webster explains that information society theorists contend that "technological innovation produces social change" (Webster 2002, p. 264). On the other side of the debate, of which Webster is a staunch proponent, scholars charge that no information society exists and that information and technology are simply following a path of continuity with historical change. More importantly, Webster (p. 6) makes the distinction that many scholars occupy various points along the continuum of both constructs. Webster explains that there exist five criteria (technological, economic, occupational, spatial, and cultural) for definitions of an information society. The criteria for these definitions are driven by the thought that quantitative changes in information are evoking qualitative changes in society, thus contributing the notion of an information society (Webster 2002, p. 9). By repurposing Webster's five criteria for defining the information society and translating it into the context of technical communication, the connection between technical communication and an information society become more evident.

Technological

It seems obvious from its name that technical communication has technological foundations attached to it. With technological innovation comes societal change. To clarify this, Webster (2002, p. 10) highlights the 1980 suggestions of Alvin Toffler and the three waves of technological innovations. The three waves Toffler refers to are the agricultural revolution, the industrial revolution, and the information revolution (Webster, 2002). These waves are seen by scholars as periods of change both in commercial, professional, and academic circles, but also in societal measures. Webster points to several pieces of evidence that technology is an integral part of the social systems that we occupy. He exemplifies that many studies show that R & D departments have engaged society's needs and wants by providing technologically savvy products that focus on helping establish identity and social comfort for the user. Such products include the Apple IPod and downloadable ring tones, both customizable technologies that provide a sense of identity.

Another such example is creating a Web portfolio. DiMarco (2005) explained that a Web portfolio is a collection of artifacts that is organized and presented in a Web site. A Web site, once considered a high end professional project, requires technical communication skills including Web design, digital imaging, and information organization abilities which have become more common in technologically savvy societies.

Economic

The work of Machulup (1962) has focused on establishing the idea that information business is central to the economies of a specific nation and that each sector of an economy can be categorized into informational and non-informational elements and counted to assess the weight information based activities hold in an economy. By assessing informational and non-informational activities within an economy, a distinction can be made that an economy heavily weighted in information services and products can be seen as an information economy (Webster, 2002). An information society is one in which information economy drives public and private sectors. How is technical communication connected? Technical communication activities are based on managing, manipulating, and publishing information. Therefore, as economic dependence on information based elements grows, so does the economic stake of the technical communications practitioners and academics. The value of technical communication in an information economy is built around the need for dissemination of information that stimulates the economy. Dissemination occurs through many public and private media vehicles. These vehicles and activities drive society to experience, learn, understand, and act.

Occupational

Discussing the work of post industrial society theorist Daniel Bell, Webster (2002) references the evidence that "in western Europe, Japan and North America over 70 percent of the workforce is now found in the service sector of the economy, and the white collar jobs are now in the majority" (p. 14). Webster points out that Bell evidence provides evidence that insists that the information society must exist because "predominant group [of occupations] consists of information workers" (p. 14). Technical communication activities contribute to this trend. As the information society grows, so will the unyielding need for technical communication skills in all occupations that involve service. Whether it is using a PDA for recording package deliveries or designing e-learning for aeronautics training, technical communica-

tion skill is present in all levels of occupation within the information society.

Spatial

Castells (1996) identified spatial boundaries in the information society by explaining that with the network society, the boundaries of the clock have disappeared. Technical communication on everyday and professional levels uses information networks to relinquish boundaries of time and space. One such example is that of e-mail. It is an everyday act executed by many people and is also relied upon heavily in the professional and business world for dynamic communication.

Cultural

In organizations, there are many people who prefer to create their own presentations using PowerPoint. This is positive from an empowerment standpoint but can be devastating from a communication standpoint. As seen in Edward Tufte's 2003 manifesto on the evils of bad presentation design, "The Cognitive Style of PowerPoint", Tufte exalts that the tools are not the content. The content is the content and those who create content must take care of how it is presented in the form of quantity, style, pace, and, most importantly, message. Technical communication specialists who are formally trained in design, writing, interactive multimedia, and presentation technology can decipher between weak and strong communication and utilize tools for achieving a goal other than simply to use a tool. Myspace.com is cultural example of technical communication. A myspace.com Web account allows you to upload text and graphics to a Web page that appears under your specific profile. Creating Web pages and uploading them to myspace.com and the Internet through templates is a technical communication activity. It involves special knowledge to evoke shared experiences and persuade the visitor to act by providing feedback via chat or e-mail.

FUTURE TRENDS

As the future serves more information and more information manipulation devices, building the context of the information society into the definition of technical communication will be critical to keeping a true measurement of the impact of the discipline. Investigation of the connection between technical communication and the information society is suggested as a topic for further historical, ethnographic, and action research.

Reliance on information based skills, products, and services will continue as society grows more attached to socio-technological ways of life. Using tools and processes in technical communication in everyday life will require, as well as foster, information literacy across populations. The growth of information literacy and skill building will be a necessity for prosperity in information based economies worldwide. Further research in the value of technical communication in the information society is suggested to enable continued growth of technical communication as a social science that reveals ubiquitous communication challenges to evaluating, understanding, designing, publishing, and presenting information for professionals, academics, and everyday people.

CONCLUSIONS

A interweaved definition technical communication has been established in this paper as techniques (technical writing, research, information management, digital document design, Web design) involving special knowledge of a mechanical or scientific subject for the purpose of sharing informational signs to make them common, to evoke shared experiences, and to persuade people to act as developer, producer, receiver, or user.

Finally, this paper identifies technical communication as evolving as a critical societal skill set for professional or ubiquitous contexts and all who live in an information society.

REFERENCES

Berlo, D. (1960). *The Process of Communication: An introduction to Theory and Practice.* San Francisco: Rinehart Press. p.16.

Castells, M. (1996). *The Internet Galaxy*: Reflections on the Internet, Business, and Society. Oxford: Oxford University Press.

Connors, R. (1982). The Rise of Technical Writing Instruction in America. *Journal of Technical Writing and Communication.* 12.4, 329-352.

DiMarco, J. (2004). *Computer Graphics and Multimedia: Applications, Problems, and Solutions.* Hershey: Idea Group. p. vii.

DiMarco, J. (2005). *Web Portfolio Design and Applications.* Hershey: Idea Group.

Machulup, F. (1962). *The Production and Distribution of Knowledge in the United States.* Princeton: Prince University Press.

Meggs, P. (1998). *A History of Graphic Design.* New York: Wiley & Sons.

O'Hara, F. (2001). *A Brief History of Technical Communication*. Retrieved from http://www.stc.org/confproceed/2001/PDFs/STC48-000052.pdf; 2001 STC Proceedings.

Pringle, K., and Williams, S. (2005). The Future is the Past: Has Technical Communication Arrived as a Profession? [Electronic version].*Technical Communication,* 52(3), 361-370.

Severin, W., and Tankard, J. (1979). *Communication Theories: Origins, Methods, Uses.* New York: Hastings House. p. 5-7.

Sheehan, R. (2005). *Technical Communication Today.* New York: Pearson. p. 5-9.

Society of Technical Communication Website. *STC History.* Retrieved from http://www.stc.org/about/history01.asp

Staples, K. (1999). Technical Communication from 1950-1998: Where are we now? [Electronic version]. *Technical Communication Quarterly, 8 (2), 153-164.*

Steiner, C. J. (1999). Getting Personal: Individuality, Innovation, and Technical Communication [Electronic version]. *Technical Writing and Communications, 29*(4), 383-399.

Tufte, E. (2003). *The Cognitive Style of PowerPoint.* Cheshire: Graphics Press, LLC.

Webster, F. (2002). *Theories of the Information Society.* New York: Routledge. p. 6-9.

KEY TERMS

Communication: A transactional process that involves the exchange of information by sharing and evaluating informational signs which include verbal (auditory), non-verbal (visual), written, and mass (radio, television, print, and Web medias) message types.

Everyday Technical Communication Activities: Activities that require technical communication skills. (playing video games, operating cell phone, using a digital camera).

Everyday Technical Communication Skills: Technical communication skills (menu navigation, scrolling, visual communication) used to share informational signs to evoke shared experiences in the execution of everyday technological activities.

Information Society: A society that relies on information products and serves to thrive and prosper.

New Media: Media characterized by digital content, development, and delivery. New media includes digital video, computer animation, 3d Modeling, video games, motion graphics, kiosks, PDA, iPod, and Web sites.

Technical Communication: (1)Techniques (technical writing, research, information management, digital document design, Web design) involving special knowledge of a mechanical or scientific subject for the purpose of sharing informational signs to make them common, to evoke shared experiences, and to persuade people to act. Technical communication can occur within the contexts of developer, producer, receiver, and user because of the dynamic, ubiquitous nature of technical communication tools, techniques, and roles. (2) An academic discipline focusing on technical writing, research, information management, digital document design, and Web design.

Technical Writing: Writing medical or technology based materials which include: standardized procedure documents, definitions, descriptions, instructions, and training.

Ubiquitous Technology/Ubiquitous Computing: Microprocessor based devices that people to use in everyday personal life. Examples of ubiquitous computing are IPods, cell phones, and automobile navigation systems.

Technological and Social Issues of E-Collaboration Support Systems

Nikos Karacapilidis
University of Patras, Greece

INTRODUCTION

Removal of communication impediments and provision for techniques that systematically direct the pattern, timing, and content of cooperative processes are two key prerequisites in the contemporary organization. Their establishment has been proven to facilitate the solution of ill-structured problems by a set of individuals working together as a team, through the interactive sharing of information between them. *E-collaboration* involves a variety of both communication and cooperation issues, in that it leverages the connective powers of a computer network to coordinate the efforts of a group of people. By using e-collaborative capabilities in an organization, people can operate as a single business entity, thus making joint decisions of added value.

At the same time, the representation and visualization of social structures and interaction in a collaborative environment is also of major importance. This is associated to the perception and modeling of actors, groups and organizations in the diversity of collaborative contexts. A problem to be addressed is to provide the means to represent and manage user and group profiles, as well as social relationships in a collaborative context. Neither relationships nor contexts are static; they are emerging and change over time, which necessitates the development of adaptive services. Furthermore, social relationships are diverse and of different intensity. What is required is development and utilization of appropriate mechanisms that perceive given structures in order to extract implicit information.

Issues to be addressed in the establishment of an e-collaboration environment should have a strong *organizational focus*. These include work structuring in order to improve coordination, use of communication technology to make collaboration more efficient and effective, enforcing of rules and procedures for achieving consistency, exploitation of social structures and interaction, and automation of data processing in data intensive situations (Angehrn and Jelassi, 1994).

BACKGROUND

The environment in which a collaborative process takes place sets a series of important requirements. Issues to be taken into account in the design and implementation of an e-collaboration system include:

- The *spatial distance* between team members. This refers to whether full face-to-face communication among them is possible. Depending on the group size and the proximity of members during a decision making procedure, various settings have been identified (DeSanctis and Gallupe, 1987).
- The *temporal distance* among the activities performed by the individual group members. This refers to whether collaboration is taking place through meetings at a particular time, such as in conventional meeting or teleconferencing environments, or whether participants submit their input at different points in time, based on electronic mail, bulletin boards, newsgroups, and computerized conferencing concepts.
- The *type of participants' goals* distinguishes between an environment in which a group wants to solve its common problem cooperatively, and another, in which bargaining takes place. Issues arisen in the first case concern knowledge sharing, preference aggregation, and negotiation support.
- The type of *control* over the collaborative process. There may be cases where the participants follow a democratic process in order to reach a solution, and cases where the system is supported by a human group leader or *mediator*.
- *Separating people from the problem.* The system designer has to evaluate the individual and group characteristics of the participants, as well as their motivations, disagreements, and conflicts, in order to reduce (if not avoid) the negative impact that misunderstandings, emotions and bad communication may have.
- *User modeling.* The term user modeling refers to the process of acquiring knowledge about a user in order to provide adapted services or information to his/her specific needs (McTear, 1993). By having the characteristics of users explicitly represented within a system, it can be used as a resource in various types of computations and services in order to bring user-tailored services. User modeling is in general motivated by the observations that different users have different

needs. The user model is an essential component when considering personalized interaction and adaptive filtering in information systems. It provides the means to control and confront important problems such as cognitive overhead.

- *Social networking.* Social structures, relationships and interaction should be represented and visualized in a way that makes it possible to reflect on them in their context. In order to provide this, appropriate structure representations and visualizations must be provided. Specialized applications for representing social structure and relationships are usually known as *social network applications* (Atzenbeck and Tzagarakis, 2007).
- The *type of communication* between the participants. Collaborative environments can be based either on *point-to-point communications*, or on *broadcasting* of messages.

Furthermore, approaches for the development of a framework for e-collaboration have to address both behavioral and technical aspects (Zigurs, Poole, and DeSanctis, 1988). Behavioral issues reported concern the diffusion of responsibility, pressures toward group consensus and problems of coordination.

COMPUTER SUPPORTED COOPERATIVE WORK

Computer-supported cooperative work (CSCW) has been defined as computer-assisted coordinated activity, such as communication and problem solving, carried out by a group of collaborating individuals (Greenberg, 1991). The multi-user software supporting CSCW is known as *groupware* (Ellis, Gibbs and Rein, 1991). CSCW may also be viewed as the emerging scientific discipline that guides the thoughtful and appropriate design and development of groupware (Greenberg, 1991). Key issues of CSCW are group awareness, multi-user interfaces, concurrency control, communication and coordination within the group, shared information space, and the support of a heterogeneous, open environment which integrates existing single-user applications.

A principal aim for the designer of an e-collaboration framework is to apply state-of-the-art telematics and groupware technology to provide advanced support for the users over wide area networks, in particular the Internet. Generally speaking, CSCW tools can harness the complexity of the social and knowledge processes involved, thus providing benefits in terms of speed and accuracy, and facilitating the development of business policies. Such tools can be used to support the group reasoning processes, that is, to facilitate the evaluation of proposed solutions and their support, to structure the decision-making process through the implementation of specific methodologies, and to help group members in reaching a shared understanding of the issue by supporting knowledge elicitation, knowledge sharing and knowledge construction. Moreover, by exploiting intranet or Internet technologies, they can connect participants with similar interests, encouraging dialogue and stimulating the exchange of knowledge.

A plethora of systems that support capturing of decision rationale and argumentation for different types of user groups and application areas has been already developed. For instance, *QuestMap* (Conklin and Begeman, 1987) can capture the key issues and ideas during meetings and attempts to create a shared understanding by placing all messages, documents and reference material for a project on a "whiteboard", while *Sibyl* (Lee, 1990) is a system that provides services for the management of dependency, uncertainty, viewpoints and precedents. Generally speaking, this category of systems meets the collaboration requirements concerning the type of control, conflict resolution, and behavioral issues, by providing a cognitive argumentation environment that stimulates reflection and discussion among participants. However, issues related to temporal and spatial distances are not fully addressed. These systems do not exploit any network infrastructure, thus users can work in an asynchronous way only through a human mediator who receives their contributions and appropriately deploys them to the system. Most important, this category of systems does not integrate any reasoning mechanisms to (semi)automate the underlying decision making and negotiation processes.

Increasing interest has been also developed in implementing Web-based conferencing systems, such as *AltaVista Forum Center*, *Open Meeting* and *NetForum*. Such systems exploit the platform-independent communication framework of the Web, as well as its associated facilities for data representation, transmission and access. They usually provide means for discussion structuring and user administration tools, while the more sophisticated ones allow for sharing of documents, on-line calendars, and embedded e-mail and chat tools. Discussion is structured via a variety of links, such as simple responses or different comment types to a previous message. This category of systems meets fully the requirements that are related to the spatial and temporal distances between members of a team. However, these systems merely provide threaded discussion forums, where messages are linked passively. This usually leads to an unsorted collection of vaguely associated comments. As pointed out by the developers of *Open Meeting*, there is a lack of consensus seeking abilities and decision-making methods (Hurwitz and Mallery, 1995). Moreover, as in the previous category of systems, issues related to the appropriate storage of knowledge in order to be exploited in future collaboration settings are not addressed.

This last category of systems belongs to the family of *social media*. Social media have increased in popularity

during the last few years, due to the development of capable web technologies and the growth of a wider and more mature user community discovering the potential of user-created content and many-to-many communications. Hence, large virtual communities are today formed around a multitude of subjects, collaborate through wikis, express and comment through blogs, share material through resource sharing sites (such as *delicio.us* for bookmarks, *Flickr* for photos, and *YouTube* for videos), and use forums for structured discussions. The usefulness of these systems with respect to large virtual communities is in many cases evident, as they are successful and have a large user base. However, with respect to capturing collaborative discourses, there is a range of weaknesses as far as the inherent needs of communities are concerned.

In parallel to the growth of social media, tools and methodologies for supporting collaborative discourses have been matured. One example is the *Compendium* tool (http://www.compendiuminstitute.org/), which supports various concept mapping paradigms and dialogue mapping. The approach followed originates from IBIS (Issue Based Information System), introduced back in 1970 (Kunz and Rittel, 1970). Also building on IBIS concepts, the *Hermes* system (Karacapilidis and Papadias, 2001) supports collaborative decision making through formal argumentative discourse acts. Another example is *Conzilla* (http://www.conzilla.org/), a tool that works with "context-maps", which are collaborative crafted contexts or presentational layers on top of semantic information. Finally, *CoPe_it!* (http://copeit.cti.gr/) is an innovative Web-based system that attempts to assist and augment collaboration being held among members of Communities of Practice by facilitating the creation, leveraging and utilization of the relevant knowledge (Karacapilidis and Tzagarakis, 2007).

SOCIAL NETWORK APPLICATIONS

Social network applications support users in jointly keeping track of their connections to others. Examples can be found for various domains, such as business (e.g., LinkedIn.com, Xing.com), students (e.g., studiVZ.net), or entertainment (e.g., MySpace.com). Many research projects deal with representing social structures. Examples include studies on large distributed organizations (Hinds and McGrath, 2006) or surveys on various special groups, such as researchers (Spence, Reddy and Hall, 2005) or software developers (Elliott and Scacchi, 2003). A recent survey on selected social network or contact management software has shown that their structures are based on simple syntax and semantics (Atzenbeck and Tzagarakis, 2007). They mostly represent social relationships apart from their contexts, for example, their connections to assets or the intensity of relations. This makes structure representations highly abstract.

Dealing with human and intellectual capital can be also approached by social network analysis, because organizations function by way of a social network of employees influencing, giving, hoarding, or accumulating data. From this network sprout the innovations that will produce the next revenue generating or cost saving product or service. Although no organization can survive without such a network, some organizations are beginning to realize that they can profit by analyzing these invisible - sometimes described as tacit - communication links. Real working knowledge lies in the relationships between employees. Many companies confuse hierarchical structure with their social network, but a hierarchical tool such as an organizational chart reflects procedural, not social knowledge.

Communities' activities concerning collaborative asset production is a special form of technologically embedded social networks. Various research from the social sciences has to be acknowledged in order to respect special roles, hierarchies, and positions within an organization. In network organizations (Castells, 2004), new forms of links between the individual level and the organizational level, namely networked links in project teams, do have several implications for software design, especially concerning the management of knowledge (Elmholdt, 2004). Of special relevance for the use of asset management applications are power differences in collaborative teams. Research shows that team members with insufficient formal power have difficulties with such work (Cabrera and Cabrera, 2002). Moreover, computer-mediated communication over long distances in worldwide working teams has been proven to often lead to new adjusted roles, hierarchies and individual timetables (Licoppe and Smoreda, 2005).

Finally, attempts to address collaborative behavior modeling have been confined in terms of generality and flexibility. In the proposed approaches, a participant's intentions and plans are identified by the system based on the behavior and intentions regarding a particular issue in hand (Introne and Alterman, 2006).

FUTURE TRENDS

We argue that services to be provided in a contemporary e-collaboration framework can be classified in four levels:

- The *information services* should deal with the interoperability of proprietary systems, providing efficient and cost-effective access to multimedia data in heterogeneous, distributed databases over wide-area networks. In particular, services should be included for finding relevant data and converting proprietary data to standard formats for data interchange. Additionally, these services should include ways of controlling remote servers from within compound documents

and general-purpose electronic mail, conferencing systems, and hypermedia systems, such as the World Wide Web. Another major issue here concerns the provision for customized solutions, which adapt to a team member's profile according to his/her preferences, abilities, experience, and collaboration mode, as well as aspects related to technical specifications of his/her platform, software available, and network connection. In order to be effective, such solutions have to remove barriers imposed by non-interoperable collaboration tools, inadequate infrastructure, undefined data sharing policies and standards, and differing priorities for presentation formats. What is often required is generation of customized content through approaches such as document transformation, dynamic documents generation, adaptive hypermedia, and provision for personalized collaboration tools, based on adaptive learning techniques, that track a team member's activity and interactions with the system, analyze the feedback, and accordingly identify his/her needs or interests.

- The *documentation services* should provide a "shared workspace" for storing and retrieving the documents and messages of the participants, using appropriate document formats, such as XML. As argued in (Prahalad and Hamel, 1990), an organization's only advantage in today's business environment is its ability to leverage and utilize its knowledge. While a firm comprises individuals and a set of objectified resources, its most strategically important feature is its body of collective knowledge (Spender, 1996). Such knowledge resides in an evolving set of assets including the employees, structure, culture, and processes of the organization. Of these, employee knowledge, and, particularly, tacit knowledge is identified as the dominant one, which is decisive at all mental levels and has to be fully exploited (Nonaka, 1994). Security and privacy issues should be also addressed here. Moreover, controlled experimentation by simulation may augment the quality of a collaborative process by providing insight into the dynamic interactions and feedback loops formed by the problem elements (Sterman, 2000). A simulation model can map organizational knowledge onto appropriate graphs quantifying the problem under consideration, thus providing a clearer understanding of which alternative solution seems to be more prominent at the moment. Finally, databases containing project documents may also become part of the *collective memory* of a community, facilitating the design and re-use of plans.
- The *social network services* should satisfy requirements related to the user and group modeling, as well as to the associated social structures, interactions and relationships. First of all, user (group) modeling requires an explicit representation of the notion of user (group), usually called the *profile*. This takes usually the form of a predefined attribute hierarchy that characterizes the user (group) in a particular system. The associated attributes are usually specific to the domain of the application. They can be categorized, depending on how they are populated and who may modify them, as explicit (their values are provided by users themselves and include personal data such as name, address, birth date, preferences, competencies, skills etc.) or implicit (their values are not provided by users explicitly, but implicitly, by observing their behavior within the system). User (group) modeling is also associated with *mechanisms for the acquisition of user (group) related information.* The role of these mechanisms is to acquire the implicit information of users (groups) that has been mentioned above. They observe and log the operations and discourse moves of users within the system and record them in the user's profile. Finally, user (group) modeling services need to offer *inference engines.* The role of these engines is to analyze all data present in the profile and either extract new information about a user (group) or take appropriate actions towards adapting the software according to the user's (group's) current state. One important aspect that most *user modeling* approaches have in common is that they focus on individual users and can provide their advanced services only on the user level; thus, they do not deal at all – or deal marginally - with the community-related aspects (i.e. relationships between individual users and relationships between users and knowledge artifacts). While this may be acceptable in situations where the individual is of prime importance, such as in e-commerce and learning systems, it is not adequate in collaboration environments, where the emphasis is on the community and not on the individual. In such situations, adaptation and role management at the community – rather than at the individual – level is required.

  Management of social structures, interactions and relationships is also critical in a contemporary e-collaboration framework. As mentioned in the previous section, many applications and projects dealing with social relationships mainly support explicit and abstract structures. However, social structures may gain from the expertise of structure domain research, including various structure abstractions or ways for implicit structuring. Another issue to be addressed concerns the elaboration of social relationships in their contexts, that is, how they relate to assets, locations, or change over time. Social network analysis has to be extensively used to find who is depending on whom in a network. Such an analysis will also help to detect hidden hierarchy of social networks. Finally, another service of this category concerns the (semi)automatic

role-specific cognitive mapping for each participant, based on his/her overall behavior (with respect to the knowledge content of every issue of concern and his/her issue-specific collaborative behavior), as well as development of (cognitive map related) complexity, centrality, and collaboration metrics, which will be associated to the context-sensitive and collaborative use of assets.

- The *mediation services* should regulate the group's activities and facilitate the underlying decision making processes. Commercial workflow systems can be used to support well-defined, formal administrative procedures within organizations. Decisions should be considered as pieces of descriptive or procedural knowledge referring to an action commitment. In such a way, the decision making process is able to produce new knowledge, such as evidence justifying or challenging an alternative or practices to be followed or avoided after the evaluation of a decision, thus providing a refined understanding of the problem. On the other hand, in a decision making context the knowledge base of facts and routines alters, since it has to reflect the ever-changing external environment and internal structures of the organization (Bhatt and Zaveri, 2002). Knowledge management activities such as knowledge elicitation, representation and distribution influence the creation of the decision models to be adopted, thus enhancing the decision making process (Bolloju, Khalifa and Turban, 2002).

CONCLUSION

We have summarized a series of technological and social issues to be considered in the development of systems supporting e-collaboration in the contemporary organization. Services to be provided by such systems have been classified in four levels, namely, information, documentation, social network, and mediation services. We argue that more research and applied work needs to be carried out on issues concerning the synergy of knowledge management and decision making, while this should be further enhanced by providing advanced argumentation, visualization and experimentation features.

REFERENCES

Angehrn, A. and Jelassi, T. (1994). DSS Research and Practice in Perspective. *Decision Support Systems*. 12, 267-275.

Atzenbeck, C. and Tzagarakis, M. (2007). Criteria for social applications. In V. Dimitrova, M. Tzagarakis, J. Vassileva (eds.). *SociUM: Adaptation and personalization in social systems: groups, teams, communities - Workshop Proceedings in conjunction with the 11th International Conference on User Modeling*. 45–49.

Bhatt, G. and Zaveri, J. (2002). The Enabling Role of Decision Support Systems in Organizational Learning. *Decision Support Systems*. 32(3), 297-309.

Bolloju, N., Khalifa, M. and Turban, E. (2002). Integrating Knowledge Management into Enterprise Environments for the Next Generation Decision Support. *Decision Support Systems*. 33, 163-176.

Cabrera, A. and Cabrera, E. (2002). Knowledge-sharing dilemmas. *Organization Studies*. 23, 687–710.

Castells, M. (2004). *The Network Society: A Cross-cultural Perspective*. Edward Elgar.

Conklin, E.J. and Begeman, M.L. (1987). gIBIS: A Hypertext Tool for Team Design Deliberation. In *Proceedings of the Hypertext'89 Conference*. ACM Press, New York, 247-252.

DeSanctis, G. and Gallupe, R.B. (1987). A Foundation for the Study of Group Decision Support Systems. *Management Science*. 33(5), 589-609.

Ellis, C.A., Gibbs, S.J., and Rein, G.L. (1991). Groupware: Some issues and experiences. *Communications of the ACM*. 34(1), 39-58.

Elliott, M.S. and Scacchi, W. (2003). Free software developers as an occupational community: resolving conflicts and fostering collaboration. In *Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work*. ACM Press, 21–30.

Elmholdt, C. (2004). Knowledge management and the practice of knowledge sharing and learning at work: A case study. *Studies in Continuing Education*. 26, 327–339.

Greenberg, S. (1991). *Computer-Supported Cooperative Work and Groupware*. Academic Press, London, UK.

Hinds, P. and McGrath, C. (2006). Structures that work: social structure, work structure and coordination ease in geographically distributed teams. In *Proceedings of the 20th Anniversary Conference on Computer Supported Cooperative Work*. ACM Press, 343–352.

Hurwitz, R. and Mallery, J.C. (1995). The Open Meeting: A Web-Based System for Conferencing and Collaboration. In *Proceedings of the 4th International World Wide Web Conference*, Boston, MA (available at: http://www.ai.mit.edu/projects/iiip/doc/open-meeting/paper.html, last accessed: Dec 12, 2007).

Introne, J. and Alterman, R. (2006). Using Shared Representations to Improve Coordination and Intent Inference. *User Modeling and User-Adapted Interaction*. 16, 249-280.

Karacapilidis, N. and Papadias, D. (2001). Computer Supported Argumentation and Collaborative Decision Making: The HERMES system. *Information Systems*. 26(4), 259-277.

Karacapilidis, N. and Tzagarakis, M. (2007). Supporting Incremental Formalization in Collaborative Learning Environments. In *E. Duval, R. Klamma and M. Wolpers (Eds.): Proceedings of the EC-TEL 2007 Conference*, Crete, Greece, September 17-20. Springer-Verlag, Berlin, LNCS 4753, 127–142.

Kunz, W. and Rittel, H. (1970). *Issues as Elements of Information Systems*. Technical Report, 0131. Universität Stuttgart, Institut für Grundlagen der Planning.

Lee, J. (1990). SIBYL: A Tool for Managing Group Decision Rationale. In *Proceedings of the CSCW'90 Conference*. ACM Press, New York, 79-92.

Licoppe, C. and Smoreda, Z. (2005). Are social networks technologically embedded? How networks are changing today with changes in communication technology. *Social Networks*. 27, 317–335.

McTear, M. (1993). User modelling for adaptive computer systems: a survey of recent developments. *Artificial intelligence review*. 7, 157-184.

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*. 5(1), 14-37.

Prahalad, C.K., and Hamel, G. (1990). The Core Competence of the Corporation. *Harvard Business Review*. 68(3), 79-91.

Spence, P.R., Reddy, M.C., and Hall, R. (2005). A survey of collaborative information seeking practices of academic researchers. In *Proceedings of the 2005 International ACM SIGGROUP Conference on Supporting Group Work*. ACM Press, 85–88.

Spender, J. (1996). Organizational Knowledge, Learning and Memory: Three Concepts in Search of a Theory. *Journal of Organizational Change Management*. 9(1), 63-78.

Sterman, J.D. (2000). *Business Dynamics: Systems Thinking and Modeling for a Complex World*. McGraw Hill, New York.

Zigurs, I., Poole, M.S., and DeSanctis, G.L. (1988). A Study of Influence in Computer-Mediated Group Decision Making. *MIS Quarterly*. December 1988 Issue, 625-644.

KEY TERMS

Computer-Supported Cooperative Work (CSCW): A computer-assisted coordinated activity, such as communication and problem solving, carried out by a group of collaborating individuals.

E-Collaboration: The process in which a set of individuals communicate through an intranet or Internet to coordinate their efforts towards the solution of a problem.

Group Decision Support System: An interactive, computer-based system that aids a set of decision makers working together as a group in solving ill-structured problems. It enables decision makers to analyze problem situations and perform group decision-making tasks.

Groupware: The multi-user software supporting CSCW. Sometimes this term is broadened to incorporate the styles and practices that are essential for any collaborative activity to succeed, whether or not it is supported by computer.

Knowledge Management: The active management of the expertise in an organization involving collection, categorization, and dissemination of knowledge; the activity of representing and processing knowledge.

Social Network Applications: Specialized applications for representing social structure and relationships.

User Modeling: The process of acquiring knowledge about a user in order to provide adapted services or information to his/her specific needs.

Technologies for Information Access and Knowledge Management

Thomas Mandl
University of Hildesheim, Germany

INTRODUCTION

Internet search engines established themselves as an everyday technology for many users. Search engines provide the main access to information on the Internet. According to estimates, some 500 million queries are sent to search engines every day (http://searchenginewatch.com/reports/article.php/2156461) in order to find the most relevant pages among the billions of pages on the Internet. The underlying technology for search engines is provided by information retrieval (IR), which is a key technology in the knowledge society.

IR deals with the search for information and the representation, storage, and organisation of knowledge. Information retrieval is concerned with search processes in which a user needs to identify a subset of information that is relevant for his or her information need within a large amount of knowledge.

For many years, information retrieval systems were mainly used by professional database hosts that provide online access to bibliographic references. Most knowledge objects were manually (or intellectually) indexed. This means that the representation in the form of index terms is selected by human information specialists usually out of a controlled vocabulary. The dominant retrieval model was the Boolean model that advocates an exact match between query and documents, which is inspired by set theory.

In the 1960s, automatic indexing methods for texts were developed. They had already implemented the "bag-of-words" approach, which still prevails. Although automatic indexing is widely used today, many information providers and even Internet services still rely on human information work.

In the 1970s, research shifted its interest to partial-match retrieval models and proved their superiority over Boolean retrieval models. Vector-space and later probabilistic retrieval models were developed. However, it took until the 1990s for partial-match models to succeed in the market. The Internet played a great role in this success. All Web search engines were based on partial-match models and provided ranked lists as results rather than unordered sets of documents. Consumers got used to this kind of search systems, and all big search engines included partial-match functionality. However, there are many niches in which Boolean methods still dominate, for example, patent retrieval.

The basis for information retrieval systems may be pictures, graphics, videos, music objects, structured documents, or combinations thereof. This article is mainly concerned with information retrieval for text documents.

BACKGROUND

The user is in the center of the information retrieval process. Nevertheless, most research tends either to be more user oriented or more algorithm and system oriented. User-oriented research tries to pursue a holistic view of the process whereas system-oriented research is concerned with measuring the effect of system components and tries to resolve efficiency issues.

The information retrieval process is inherently vague. In most systems, documents and queries traditionally contain natural language. The content of these documents needs to be analyzed, which is a hard task for computers. Robust semantic analysis for large text collections or even multimedia objects has yet to be developed. Therefore, text documents are represented by natural-language terms mostly without syntactic or semantic context. This is often referred to as the bag-of-words approach. These keywords or terms can only imperfectly represent an object because their context and relations to other terms are lost.

As information retrieval needs to deal with vague knowledge, exact processing methods are not appropriate. Vague retrieval models like the probabilistic model are more suitable. As a consequence, the performance of a retrieval system cannot be predicted but must be determined in evaluations. Evaluation plays a key role in information retrieval. Evaluation needs to investigate how well a system supports the user in solving his or her knowledge problem (Baeza-Yates & Ribeiro-Neto, 1999).

Web search engines take the information retrieval process to the Internet. They need to contain the following modules (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001).

- A crawler collects pages on the Web by starting from known pages, following the links encountered in these seed pages, and iteratively following all links found in further pages (Baeza-Yates & Castillo, 2002).

- An indexer builds a representation of the pages passed on by the indexer. Well-known information retrieval technology is used for this process including linguistic preprocessing and weighting schemes dealing with several occurrences of the same term.
- The user interface (usually a Web client) allows the user to enter queries, presents the results, and should support user strategies like iterative retrieval.
- The query processor analyzes the queries and compares them to the pages represented in the index. Based on the similarity between page and query, a ranking is produced.

Except for the crawler, the other modules are necessary for any information retrieval system and will be introduced in the following sections.

REPRESENTATION AND RETRIEVAL OF TEXT DOCUMENTS

Information retrieval deals with the storage and representation of knowledge and the retrieval of information relevant for a specific user information need. The information seeker formulates a query trying to describe the information need. The query is compared to document representations that were extracted during the indexing phase. The representations of documents and queries are typically matched by a similarity function such as the cosine. The most similar documents are presented to the users, who can evaluate the relevance with respect to their problem.

Representation

Indexing is a process during which words describing the content of a document are chosen as content representation of this document. During automatic indexing, algorithms assign keywords to documents. The indexing process for natural-language documents typically consists of the following steps.

- Word segmentation
- Elimination of stop words
- Stemming
- Compound analysis (for some languages)

Linguistic preprocessing is limited to the level lexemes and morphology. Syntax and semantics are not analyzed because current technology is not able to achieve satisfying quality for mass data. Consequently, the word provides the core for the content representation. The meaning of a text is seen as a set of basic word forms of words that occur in the text. Each word contains its meaning; however, the specific meaning within the text or a sentence can be reconstructed based on the information in the index.

Segmentation is defining the boundaries between the individual words. In European languages, most boundaries can be found by considering blanks. However, other characters need to be considered additionally. In Asian languages, where words are not segmented by blanks, segmentation is a difficult task.

Subsequently, many words that occur frequently are eliminated. These are called stop words and comprise usually articles, prepositions, pronouns, and conjunctions. These words obviously do have meaning; however, because the content is represented following the bag-of-words method, the words are isolated and taken out of their context. Stop words cannot be used for representation in such an approach. In addition, few users would post queries containing stop words. The elimination of stop words also reduces the corpus size typically by 30% and thus leads to higher efficiency (Savoy, 1999).

The most important operation during linguistic preprocessing is stemming. It maps conjugated word forms to their basic form or their stem (e.g., runs -> run, walking -> walk). Morphological variations of words fulfill their function only within their grammatical context. In a bag-of-words approach, all variations can be reduced to their basic form. Stemming improves efficiency also. Three main methods are used for stemming.

- Rule based
- Lexicon based
- Similarity based

The most important algorithms are rule based. The rules describe which steps are necessary in order to obtain the stem from a word form. The number of rules necessary is still under debate (Savoy, 2006).

All index terms are stored in an inverted list from which the documents belonging to a term can be easily accessed.

Weighting

Weighting determines the importance of a term for a document. A term weight measures how well the term represents a document. These weights mirror different levels of relevance. First, the frequency of each term is counted. Weighting assumes that words occurring more often are better representatives for a document. The relationship is not modeled as linearly increasing but is governed by a logarithmic function. The second important parameter of weighting systems is the frequency of a term in the whole collection. Very frequent words contain little discriminative power. Rare words better

distinguish between documents. This assumption is formalized in the inverse document frequency (IDF) formula for weighting (Sparck Jones, 2004). Weighting also needs to perform length normalization; otherwise, longer documents would be more likely to be encountered than short ones. Currently, advanced weighting schemes take the average and maximum length of all documents into account. Other formulas like OKAPI consider even the query terms. OKAPI has led to excellent results (Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1995).

Similarity Calculation

Once the representation has been created, users can post queries to a system. First, the words in the query are preprocessed as the words in the documents. The comparison of the representation of the query to the representation of the documents is the central part of the matching process. The documents most similar to the query will be returned. The system calculates a retrieval status value (RSV) or system relevance, which is a measure for the similarity between document and query.

The Boolean model allows only the similarity values 1 and 0. A document either belongs to the relevant set or not. The ranking or partial-match systems allow different degrees of similarity and order the result documents according to the similarity or relevance. An example for a partial-match model is the vector-space model, which interprets the retrieval process using a spatial metaphor. All documents and the query are points in a high-dimensional space. Closeness is seen as similarity. The retrieval status value can be calculated as a measure of the distance between the documents, which can be determined, for example, by the Euclidian distance. The axes in the model are defined by the terms, and the weights define the exact position of a document.

Other typical similarity functions like the cosine similarity exploit the direction of the vector, which each document defines. The angle between document and query denotes the degree of relevance. When both vectors point toward the same direction, the angle is small. The cosine function returns a high value for such an angle and a small similarity value for a large vector.

In Web retrieval, systems often take noncontent factors into account. A popular way to add additional terms to the representation is the analysis of link labels. How do other pages refer to a certain page? This text is added to the model. An additional way to calculate the similarity is the consideration of authority by link analysis. Algorithms for link analysis take the number of links that point to a page as a measure for their quality or authority. This absolute value of a page is used during the similarity calculation (Thelwall, 2004).

EVALUATION

Information retrieval systems can be implemented in many ways by selecting a model and specific language processing tools. They interact in a complex system and their performance for a specific data collection cannot be predicted. As a consequence, the empirical evaluation of the performance is a central concern in information retrieval research.

A holistic evaluation of a system is difficult and needs to set the satisfaction of the user as the yardstick. The user is satisfied when the initial information need is stilled. A holistic evaluation would need to consider the user interface, the speed of the system, as well as adaptivity. In order to achieve such an evaluation, individual and subjective factors as well as the context need to be considered. However, such results may be of little value for situations different from the one studied.

As a consequence, information retrieval research has adopted an evaluation scheme that tries to ignore subjective differences between users in order to be able to compare systems and algorithms. The user is replaced by a prototypical and constant user. Relevance judgments are carried out by domain experts who evaluate the relevance of a document independent of subjective influences. This approach is called the Cranfield paradigm after one early evaluation study (Robertson, 1982). Current research shows that the Cranfield paradigm leads to reliable results for comparing systems (Buckley & Voorhees, 2005).

The most important measures are recall und precision. Recall shows how good a system is in finding relevant documents whereas precision measures how good a system is in finding only relevant documents without ballast.

Important experiments are carried out within evaluation initiatives. The three major evaluation initiatives are historically connected. TREC (http://trec.nist.gov) was the first large effort, which started in 1992 (Voorhees & Buckland, 2005). Subsequently, CLEF (http://www.clef-campaign.org; Peters et al., 2006) and NTCIR (Kando, 2005) adopted the TREC methodology and developed specific tasks for multilingual and crosslinguistic searches. TREC achieved a high level of comparability of system evaluations for the first time in information science. The test data and collections have stimulated research and are still a valuable resource for development. The initial TREC collections for ad hoc retrieval were newspaper and newswire articles. In the first few years, the effectiveness of the systems approximately doubled. In order to cope with the new requirements and the changing necessities of different domains and information needs, new tasks were continuously established (Mandl, in press). Evaluation initiatives provide collections of documents and topics as descriptions of information needs, and after the experiments of the participating research groups,

they organise the intellectual relevance assessments and publish comparative results.

OPTIMIZATION APPROACHES

The most important technique to improve the quality of a retrieval system for a specific query is relevance feedback and subsequent iterative retrieval. Relevance feedback requires activity by the user, who needs to specify which of the documents presented in the initial result list are relevant. The system can exploit this information as an additional knowledge source and can modify the query to better suit the information need. Frequent terms in the documents judged as relevant are added to the query. Terms of documents judged as irrelevant may be omitted or weighted lower.

Although relevance feedback has proved to lead to very good results, few systems offer it because many users are unwilling to provide relevance judgments. The technique is mimicked by so-called blind relevance feedback, where the system assumes that the top documents are relevant and initiates another retrieval step. The results are presented to the user after only one iteration.

There were two main directions of research on relevance feedback: first, to reweight terms automatically depending on their distribution over relevant and irrelevant documents, and second, to provide term expansion within the initial query (Harman, 1992). Later on, the community paid attention to more detailed aspects, namely, the relation between the modification of weights and query expansion, the selection of further query terms, the effectiveness of the number of iterations, and so forth.

Many evaluation studies have shown that the results of similarly well-performing information retrieval systems often differ. This means that while the systems find the same percentage of relevant documents, the overlap between their results is sometimes low. Therefore, fusion seems to be a promising approach and has been applied to retrieval (McCabe, Chowdhury, Grossman, & Frieder, 1999). Fusion methods combine different perspectives of representation and query-document matching. They delegate a task to different systems and integrate the results into one final result. Fusion requires the integration of different probabilities for the relevance of a document. Such approaches are also referred to as polyrepresentation (Ingwersen, 1994).

FUTURE TRENDS

The integration of information resources is an important trend. Integration tries to combine several data sources under one virtual system so that users need to interact only with one interface for querying all collections. This often leads to semantic heterogeneity, which can lead to a degradation of the quality from the perspective of the user. Consequences may be that the user's terminology does not match the terminology of the system, or the cognitive landscape he or she has formed of a domain does not match the structures presented within an interface. Heterogeneous ontologies occur in many areas. The most typical attempt to resolve this problem is standardization and concentration on one ontology. However, this may not always be possible, and aspects of and perspectives on the domain may get lost. As a consequence, heterogeneity treatment is necessary to overcome the incompatibilities between different ordering systems. A traditional but expensive approach is the creation of concordances. Currently, automatic systems based on machine learning are being developed (Hellweg et al., 2001).

More and more collections contain documents written in different natural languages. Multilingual or crosslinguistic information retrieval (CLIR) is concerned with the retrieval of documents in other languages than the query language. This is of importance if the user has only passive knowledge of a language that enables him to read a document. However, in many cases, users lack the active knowledge to formulate a good query. CLIR needs to bridge language borders and applies several methods from natural-language processing. Often, machine translation is used to translate the query. In these cases, the translation quality is not the most important factor. Sometimes, a bad translation may still lead to good retrieval results. Other approaches include statistic transfer methods without an explicit translation step (Oard, 1997).

Apart from text documents, other media types are collected in large amounts as well. Image, music, and video collections require content-based retrieval also. So far, the content of a picture cannot be extracted fully automatically. The identification of objects or faces in pictures and the classification of similar images is an active research area. Current systems can retrieve documents based on color, texture, forms, objects, and spatial relationships between objects. Still, much research is needed to achieve systems in which the user can naturally query for images (Bimbo, 1999).

Information retrieval systems often offer value-added tools in order to support the user. These tools may provide personalization, geographic constraints on queries, support during query formulation, or the display of relationships between documents.

CONCLUSION

Information retrieval is a technology that guarantees access to large collection of unstructured text. It is the basic technology behind Web search engines and, as such, is an everyday technology for many Web users.

Information retrieval deals with the storage and representation of knowledge and the retrieval of information relevant for a specific user problem. Information retrieval systems

need to process queries that typically contain a few words. The query is compared to document representations that were extracted during indexing. The most similar documents are presented to the users.

REFERENCES

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2-43.

Baeza-Yates, R., & Castillo, C. (2002). Balancing volume, quality and freshness in Web crawling. In A. Abraham, J. Ruiz-del-solar, & M. Köppen (Eds.), *Soft-computing systems: Design, management and applications. Frontiers in artificial intelligence and applications* (Vol. 97, pp. 565-572). IOS Press.

Baeza-Yates, R., & Ribeiro-Neto, B. (Eds.). (1999). *Modern information retrieval.* Harlow, United Kingdom: Addison-Wesley.

Bimbo, A. d. (1999). *Visual information retrieval.* Morgan Kaufman.

Buckley, C., & Voorhees, E. (2005). Retrieval system evaluation. In *TREC: Experiment and evaluation in information retrieval* (pp. 53-75). Cambridge, MA: MIT Press.

Fuhr, N. (2005). *Scriptum information retrieval.* Universität Duisburg-Essen. Retrieved from http://www.is.informatik.uni-duisburg.de/courses/ir_ss06/folien/irskall.pdf

Harman, D. (1992). Relevance feedback and other query modification techniques. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures & algorithms* (pp. 241-263). Englewood Cliffs, NJ.

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M., Mutschke, P., et al. (2001). *Treatment of semantic heterogeneity in information retrieval* (Tech. Rep. No. 23). IZ Bonn. Retrieved from http://www.gesis.org/Publikationen/Berichte/IZ_Arbeitsberichte/index.htm#ab23

Ingwersen, P. (1994). *Interactive information retrieval.* Retrieved from http://www.db.dk/pi/iri/

Kando, N. (Ed.). (2005). *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access.* National Institute of Informatics (NII). Retrieved from http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings5/toc.html

Lew, M. (Ed.). (2000). *Principles of visual information retrieval table of contents.* London: Springer.

Mandl, T. (in press). *Recent developments in the evaluation of information retrieval systems: Moving toward diversity and practical applications.*

McCabe, M.C., Chowdhury, A., Grossmann, D., & Frieder, O. (1999). A unified framework for fusion of information retrieval approaches. In *Eighth ACM Conference on Information and Knowledge Management (CIKM)* (pp. 330-334). New York: ACM Press.

Oard, D. (1997, December). Serving users in many languages: Cross-language information retrieval for digital libraries. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/december97/oard/12oard.html

Peters, C., Gey, F., Gonzalo, J., Jones, G., Kluck, M., Magnini, B., et al. (Eds.). (2006). *Accessing Multilingual Information Repositories: Sixth Workshop of the Cross-Language Evaluation Forum, CLEF 2005* (LNCS 4022). Berlin, Germany: Springer.

Rijsbergen, K. v. (1979). *Information retrieval.* London: Butterworths. Retrieved from http://www.dcs.gla.ac.uk/Keith/Preface.html

Robertson, S. E. (1981). The methodology of information retrieval experiment. In K. Sparck Jones (Ed.), *Information retrieval experiment.* Butterworths.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1995). Okapi at TREC-3. In D. Harman (Ed.), *Proceedings of the Third Text Retrieval Conference (TREC 2005)*, Gaithersburg, MD. National Institute of Standards and Technology.Retrieved from http://trec.nist.gov/pubs/trec3/papers/city.ps.gz

Savoy, J. (1999). A stemming procedure and stopword list for general French corpora. *Journal of the American Society for Information Science, 50*(10), 944-952.

Savoy, J. (2006, April 23-27). Light stemming approaches for the French, Portuguese, German and Hungarian languages. In *Proceedings of 2006 ACM SAC Symposium on Applied Computing (SAC)*, Dijon, France (pp. 1031-1035).

Sparck Jones, K. (2004). IDF term weighting and IR research lessons. *Journal of Documentation, 60*, 521-523.

Thelwall, M. (2004). *Link analysis: An information science approach.* Elsevier Academic Press.

Voorhees, E., & Buckland, L. (Eds.). (2005). *The 14th Text Retrieval Conference (TREC 2005) Proceedings.* National Institute of Standards and Technology. Retrieved from http://trec.nist.gov/pubs/trec14/t14_proceedings.html

KEY TERMS

Indexing: Indexing is the assignment of terms (words) that represent a document. Indexing can be carried out manually or automatically. Automatic indexing requires the elimination of stop words and stemming.

Information Retrieval: Information retrieval is concerned with the representation of knowledge and subsequent search for relevant information within these knowledge sources. Information retrieval provides the technology behind search engines.

Inverse Document Frequency (IDF): IDF is a traditional weighting scheme for terms. It can be calculated as the logarithm of the term frequency in the document divided by the frequency of the term in the whole collection.

Link Analysis: The links between pages on the Web are a large knowledge source that is exploited by link analysis algorithms for many ends. Many algorithms similar to PageRank determine a quality or authority score based on the number of incoming links of a page. Furthermore, link analysis is applied to identify thematically similar pages, Web communities, and other social structures.

Precision: Precision is a quality measure for information retrieval evaluation. It gives the percentage of relevant documents within the document set. Precision can be calculated by dividing the number of relevant documents that were found by the number of documents found.

Recall: Recall is a quality measure for information retrieval evaluation. It can be calculated by dividing the number of relevant documents that were found by the number of relevant documents in the collection. The second figure can often only be estimated.

Stemming: Stemming refers to the mapping of word forms to stems or basic word forms. Word forms may differ from stems due to morphological changes necessary for grammatical reasons. The plural versions of English nouns, for example, are mostly constructed by adding an *s* to the basic noun. In most European languages, stemming needs to strip suffixes from word forms.

Term Weighting: Weighting determines the importance of a term for a document. Weights are calculated by many different formulas that consider the frequency of each term in a document and in the collection, as well as the length of the document and the average or maximum length of any document in the collection.

Technologies in Support of Knowledge Management Systems

Murray E. Jennex
San Diego State University, USA

INTRODUCTION

Knowledge management systems (KMSs) support the various knowledge management (KM) functions of knowledge capture, storage, search, retrieval, and use. To do this, KMSs utilize a variety of technologies and enterprise systems. This chapter surveys the various technologies and enterprise systems. Specific attention is placed on enterprise systems that integrate KM into organizational business processes, and technologies that enhance the effectiveness of these implementations. The chapter is based primarily on research summarized in *Case Studies in Knowledge Management* (Jennex, 2005a) and articles published by the Knowledge Management Track at the Hawaii International Conference on System Sciences (HICSS).

BACKGROUND

Knowledge

Davenport and Prusak (1998) view knowledge as an evolving mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. They found that in organizations, knowledge often becomes embedded in artifacts such as documents, video, audio, or repositories and in organizational routines, processes, practices, and norms. They also say that for knowledge to have value, it must include the human additions of context, culture, experience, and interpretation. Nonaka (1994) expands this view by stating that knowledge is about meaning in the sense that it is context specific. This implies that users of knowledge must understand and have experience with the context, or surrounding conditions and influences in which the knowledge is generated and used for it to have meaning to them. This also implies that for a knowledge repository to be useful, it must also store the context in which the knowledge was generated. That knowledge is context specific argues against the idea that knowledge can be applied universally, however it does not argue against the concept of organizational knowledge. Organizational knowledge is considered to be an integral component of what organizational members remember and use, meaning that knowledge is actionable.

Polanyi (1967) and Nonaka and Takeuchi (1995) describe two types of knowledge, tacit and explicit. Tacit knowledge is that which is understood within a knower's mind, and which cannot be directly expressed by data or knowledge representations and is commonly understood as unstructured knowledge. Explicit knowledge on the other hand is that knowledge which can be directly expressed by knowledge representations and is commonly known as structured knowledge. Current thought has knowledge existing as neither purely tacit nor purely explicit. Rather, knowledge is a mix of tacit and explicit, with the amount of explicitness (only one dimension needs to be measured) varying with each user. This is the knowledge continuum where purely tacit and purely explicit form the end points, with knowledge existing somewhere on the continuum between the two end points. Smolnik, Kremer, and Kolbe (2005) have an individual position of knowledge on the continuum through context explication, where context explication reflects the experience and background of the individual. Nissen and Jennex (2005) expand knowledge into a multidimensional view by adding the dimensions of reach (social aggregation), lifecycle (stage of the knowledge lifecycle), and flow time (timeliness) to explicitness. Research is continuing to refine the concept of knowledge and its dimensions.

Knowledge Management

Jennex (2005c) utilized an expert panel, the editorial review board of the *International Journal of Knowledge Management*, to generate a definition of KM as the practice of selectively applying knowledge from previous experiences of decision making to current and future decision-making activities, with the express purpose of improving the organization's effectiveness. Another key definition of KM includes Holsapple and Joshi (2004) who consider KM as an entity's systematic and deliberate efforts to expand, cultivate, and apply available knowledge in ways that add value to the entity, in the sense of positive results in accomplishing its objectives or fulfilling its purpose. Finally, Alavi and Leidner (2001) concluded that KM involves distinct but interdependent processes of knowledge creation, knowledge storage and retrieval, knowledge transfer, and knowledge application. Taken in context, these definitions of KM focus on the key elements of KM: a focus on using knowledge for

decision making and selective knowledge capture. This is important as the selective focus on knowledge capture separates KM from library science, which attempts to organize all knowledge and information, and the decision-making focus emphasizes that KM is an action discipline focused on moving knowledge to where it can be applied. Ultimately, KM may best be described by the phrase, "getting the right knowledge to the right people at the right time," and can be viewed as a knowledge cycle of acquisition, storing, evaluating, dissemination, and application.

Knowledge Management Systems

Jennex (2005c) views a KM system as that system created to facilitate the capture, storage, retrieval, transfer, and reuse of knowledge. The perception of KM and KMSs is that they holistically combine organizational and technical solutions to achieve the goals of knowledge retention and reuse to ultimately improve organizational and individual decision making. This is a Churchman (1979) view of KM that allows KMSs to take whatever form necessary to accomplish these goals. Alavi and Leidner (2001, p. 114) defined KMSs as "IT-based systems developed to support and enhance the organizational processes of knowledge creation, storage/retrieval, transfer, and application." They observed that not all KM initiatives will implement an IT solution, but they support IT as an enabler of KM. Maier (2002) expanded on the IT concept for the KMS by calling it an information and communication technology (ICT) system that supported the functions of knowledge creation, construction, identification, capturing, acquisition, selection, valuation, organization, linking, structuring, formalization, visualization, distribution, retention, maintenance, refinement, evolution, accessing, search, and application. Stein and Zwass (1995) define an organizational memory information system (OMS) as the processes and IT components necessary to capture, store, and apply knowledge created in the past on decisions currently being made. Jennex and Olfman (2006) expanded this definition by incorporating the OMS into the KMS, and adding strategy and service components to the KMS.

INTERNET KMS

Discussion

One of the most commonly cited KMS success factors (Jennex & Olfman, 2005) is having an integrated technical infrastructure including networks, databases/repositories, computers, software, and KMS experts. KM designers are using the Internet to obtain this integrated network and are using browsers as common software. Various approaches are being utilized by KMS designers to achieve common databases and repositories. Common taxonomies and ontologies are being used to organize storage of unstructured knowledge files and to facilitate knowledge retrieval, while other Internet-based KMSs serve as interfaces to large enterprise databases or data warehouses. Some Internet KMSs are being used to facilitate communication and knowledge transfer between groups. Knowledge portals are being used by organizations to push knowledge to workers and be communities of practice (CoPs) to facilitate communication and share knowledge between community members. The following section describes some examples of Internet-based KMSs.

Internet networks can be scaled to fit any size KMS. Browsers can be tailored to fit processes as desired. Taxonomies can be created that support unstructured knowledge sharing for any size KMS. The following examples illustrate this flexibility as the examples include a project KMS, an industry-wide project KMS, and an enterprise KMS. Knowledge portals can be scaled to fit either form of KMS but are more commonly used for enterprise KMS. A community of practice KMS is a variation of process/task KMSs.

Examples of Internet-Based KMSs

Project-Based KMS for a Single Organization

Jennex (2000) discussed an intranet-based KMS used to manage knowledge for a virtual Y2K project team. This KMS used two different site designs over the life of the project. The purpose of the initial site was to facilitate project formation by generating awareness and providing basic information on issues the project was designed to solve. The design of this site was based on Jennex and Olfman (2002), who suggested a structure providing linkages to expertise, and lessons learned were the knowledge needed by knowledge workers. This was accomplished by providing hot links to sites that contained Y2K knowledge, a project team roster that indicated the areas of expertise for each of the project team members and additional entries for individuals with expertise important to the project, and some basic answers to frequently asked questions. This site was accessed from the corporate intranet site through the special projects section of the IT division page. This made the site hard to find for those who did not know where to look, forcing the project team leadership to provide direction to the site through e-mail directions. The site did not contain guidelines and accumulated knowledge as reflected in test plans, test results, inventories of assets referenced to the division who owned them, and general project knowledge such as project performance data, meeting minutes and decisions, presentations, and other project documentation. This information had not

been generated at the time the site was implemented. Once generated, this information was stored on network servers with shared access to acknowledged project team members. This was done due to a lack of resources allocated to the initial site. No dedicated personnel or special technologies were allocated for the design or maintenance of the site. This site was in effect from early 1998 through mid-1998.

As the project team formed and began to perform its tasks, the requirements for the intranet site changed from generating awareness to supporting knowledge sharing. The site was redesigned and expanded to include detailed frequently asked questions (FAQs), example documents, templates, meeting minutes, an asset database, guidelines for specific functions that included lessons learned, and so forth. The knowledge content of the site was distributed into the other components of the site, and persons were identified as being responsible for the information and knowledge content of their responsible areas. Additionally, access to the site was enhanced by the addition of a hot link to the Y2K site placed and prominently displayed on the corporate intranet homepage. The basic layout of the site provided for access to seven specific sub-sites: Major Initiatives, Contacts, Documents, What's New, Hot Links, Issues and Questions, and Y2K MIS.

Access to this site was granted to all employees; however, several of the sub-sites were password protected for restricted use. Most of the knowledge contained on the site was contained in these protected sub-sites. The knowledge from the initial site was rolled over into the Hot Links and Contacts sub-sites. Additionally, information that had been previously stored on network servers was left on those sites, but access was provided through the intranet site. The network structure was expanded to include more sub-structures for storing more documents, information, and knowledge.

The effectiveness of the two sites was considered good. The first site was successful in generating interest and starting the project. The second site succeeded in taking a project that was performing in the bottom third of projects to being a leading project within six months after its release. Effectiveness of the sites was established using the model in Figure 2 and by ensuring the information quality was high and the system quality, especially the search, retrieval, and infrastructure, was good. Use of both sites was established by ensuring the sites met the needs of the project team and the company.

Project-Based KMS for an Industry

This example is the extranet site used by the utility industry for Y2K (Jennex, 2000). Its purpose was to facilitate information/knowledge sharing between industry members. It initially provided documents, procedures, and guidelines for getting projects started. It also provided an electronic forum for questions and answers. As projects progressed, more test data became available and this information was posted. Finally, this site provided links to other important sites and sources of information.

The effectiveness of the site was limited. A great deal of knowledge was stored on the site, but searching was difficult and time consuming, reducing system quality. The consensus of the Nuclear and Non-Nuclear Generation Y2K project personnel was that the site provided little benefit as many companies did not post test results, thus reducing information quality. The Substation Controls Y2K project personnel also found it limited, except they did use the knowledge to put together a statistically valid test sample as requested by the North American Electric Reliability Council (NERC). Industry consensus was that the site had limited knowledge value. A redesign of the site with more emphasis on knowledge search and retrieval was not available until after most projects were complete. It was anticipated the new site would be available for the expected onslaught of lawsuits following the rollover to 2000, which of course did not happen. A further inhibitor to effectiveness was that the member companies did not categorize equipment and system information in the same format. This lack of a shared ontology contributed to the search and retrieval difficulties and made understanding the posted information and knowledge more difficult to users from other companies.

KMS as a Knowledge Portal

This example, from Cross and Baird (2000), is an intranet site built by Andersen Consulting. Consulting firms have had a long tradition of brokering their knowledge into business. In the early 1990s, Andersen Consulting began to produce global best practices CDs for distribution to project personnel. This evolved into the development of a intranet site called KnowledgeSpace, which provided consultants with various forms of knowledge including methodologies and tools, best practices, reports from previous like engagements, and marketing presentations. Support was also provided for online communications for online communities of practice and virtual project teams. The site was effective for personnel with access to the Internet and adequate bandwidth. It should be noted that current modem technology and improved dial-in access, as well as the proliferation of broadband connections, have made sites such as this much more effective for field or remote personnel.

The second example, from Bartczak (2005), describes the system used by the United States Air Force to support the Material Command called the AFKM Hub. The AFKM Hub is the primary Web site for the AF Lessons Learned utility. Although the Web site has evolved, Lessons Learned are still the centerpiece of the Hub. Lessons Learned have been captured and categorized by subject area and provide valu-

able knowledge about past processes and events. The AFKM Hub also acts as a portal for all other AFKM components and, as such, also serves as the default AFKM homepage. The AFKM Hub provides a conduit to select relevant information and knowledge resources, and provides an avenue for creating a knowledge-sharing organization. The AFMC Help Center of the AFKM Hub allows AFMC customers to perform a natural language or keyword search of over 130 AFMC Web sites and selected databases. It connects AFMC customers throughout the Air Force and Department of Defense with the appropriate AFMC information source or point of contact. The search engine used dynamically creates a unique results page separated into four categories: ranked list of related Web documents and links, top priority Major Command issues, bulletin board discussion entries, and contact information for the AFMC command liaisons and topic area points of contact The CoP Workspace supports the growing number of Air Force communities of practice. A community of practice is a network of people who share a common goal. CoP workspaces are virtual environments where members of these CoPs can exchange information to complete work tasks and solve problems. Each CoP serves a specific customer set. The AFKM Hub provides workspaces for a variety of CoPs and supports more than 1,300 active CoPs. The effectiveness of the AFKM Hub has also been mixed. Air Force leadership sees the value in KM, and many examples of successful uses of knowledge have been recorded. However, articulating a knowledge and KM strategy has been difficult and has allowed for wasted effort in supporting Air Force KM needs.

KMS as a Topic Map

The last examples come from Eppler (2001). There are five types of knowledge maps: source, asset, structure, application, and development. A multimedia company intranet site is used to illustrate a knowledge source map. This site provides graphical buttons representing individuals with specific expertise color-coded to indicate the expert's office location. The Knowledge Asset map provides a visual balance sheet of an organization's capabilities of a skills directory or core competency tree. Colors are used to indicate knowledge domains, while the size of symbols indicates level of expertise. Knowledge Structure maps divide knowledge domains into logical blocks that are then broken into specific knowledge areas. The Knowledge Application map breaks an organization's value chain into its components parts and then indicates what knowledge, tools, or techniques are needed to implement the component part.

The last example is a Knowledge Development map. This map is used to plot the activities needed to acquire the indicated knowledge competence. Clicking on the displayed competence displays the steps needed to develop the competence. Effectiveness of these maps has only been determined for the Knowledge Asset map. This map, developed for a telecommunications consultant firm, was found to be very useful for the planning of training activities and for identifying experts quickly when required during an emergency. It should be noted that knowledge maps enhance the linkage aspects of information quality.

ENTERPRISE SYSTEM SUPPORT FOR KMS

Discussion

As organizations strive to improve their competitive position/advantage, they are implementing enterprise-wide systems. These systems integrate processes and data/information/knowledge across the enterprise, and in many cases with suppliers and customers to improve efficiency and effectiveness (Koch, 2002). This usually results in lowered operating costs and improved response times, economies of scale, and user satisfaction. Typical of these systems are enterprise resource planning (ERP), customer relationship management (CRM), supply chain management (SCM), and data warehouse implementations. As these systems are refined and improved, organizations are finding that incorporating knowledge and KM into them improves system performance. Additionally, KMS designers are finding that enterprise systems satisfy several KMS success factors including: a knowledge strategy (meeting the needs of the enterprise system processes); a common enterprise-wide knowledge structure (inherent in the enterprise systems); a clear goal and purpose for the KMS (supporting the enterprise system); search, retrieval, and visualization functions of the KMS support easy knowledge use (they are built into the enterprise system); and work processes are designed that incorporate knowledge capture and use (also inherent in the enterprise system work processes). Unfortunately, there are also several issues involved in successfully using enterprise systems to support KM; chief among these are organizational culture issues. Many enterprises suffer from fragmentation, meaning that many organizations within the enterprise own and use their own data and systems. Enterprise systems seek to integrate these systems, but to be successful they must overcome issues of ownership and a reluctance to share data, information, and knowledge. This issue is usually characterized by the presence of "silos" in the enterprise. Corral, Griffen, and Jennex (2005) discuss this issue with respect to integrating data warehouses and KM. This is considered the principal issue affecting successful integration of enterprise systems and KM. The following examples describe how enterprise systems and KMSs are being fused together.

Enterprise systems support primarily infrastructure/generic KMSs, as the primary reason for using these systems is to integrate data, information, and knowledge and to create a common infrastructure. However, data warehouse systems can be used to support process/task KMSs through the use of data marts, and CRM and ERP can be used as transitional technologies. Transitional technologies preserve the process/task KMS while converting their stored knowledge into generic knowledge through back-office operations. The issue in doing this is in getting the process/task users to agree to knowledge capture process changes that incorporate capturing supporting context and culture data and information.

Examples of Enterprise System Support for KMS

ERP and KMS

Li, Yezhuang, and Ping (2005) describe an ERP implementation in a Chinese paper manufacturing company. The ERP was implemented to help the company respond to market and customer changes more rapidly by integrating enterprise data information and knowledge and centralizing process control. Unfortunately, China lacked experience with ERP implementation (although it has experience with MRP) and is not overly familiar with western concepts of centralized data, information, and knowledge management. This was an issue in getting the ERP implemented and utilized. Once this was accomplished, decision making was greatly enhanced through improved knowledge transfer provided by the ERP's integration of organizational data, information, and knowledge into a single accessible location. Other issues faced in implementing the ERP were management support of the various sub organizations being integrated into the ERP and creating a culture that used data, information, and knowledge in the expected way.

White and Croasdell (2005) describe ERP implementations in Nestle, Colgate-Palmolive, and Chevron-Texaco. Each of these implementations was performed to improve data, information, and knowledge integration, with an expectation of improved decision making and transfer of key knowledge such as lessons learned and process improvements. All three implementations were ultimately successful after initial difficulties, including cost overruns due to unrealistic project estimates of schedule and cost, and overcoming employee resistance to changing to new processes and merging data, information, and knowledge ownership. These examples also incorporated the KMS success factor of metrics for measuring success and illustrate the importance of measuring KMS performance.

CRM and KMS

Al-Shammari (2005) describes a knowledge-enabled customer relationship management (KCRM) system in a large middle-eastern telecommunications company. The KCRM was composed of three major parts: enterprise data warehouse (EDW), operational customer relationship management (CRM), and analytical CRM. The KCRM initiative was designed to automate and streamline business processes across sales, service, and fulfillment channels. The KCRM initiative was targeted at achieving an integrated view of customers, maintaining long-term customer relationships, and enabling a more customer-centric and efficient go-to-market strategy. The driver for the initiative was that the company faced deregulation after many years of monopoly. The company initiated a customer-centric knowledge management program, and pursued understanding customers' needs and forming relationships with customers, instead of only pushing products and services to the market. Unfortunately, the KCRM program ended as an ICT project. The company did not succeed in implementing KCRM as a business strategy, but did succeed in implementing the KCRM as a transactional processing system. Several challenges and problems were faced during and after the implementation phase. Notable among these is that the CRM project complexity and responsibilities were underestimated, and as a result, the operational CRM solution was not mature enough to effectively and efficiently automate CRM processes. Changing organizational culture was also a tremendous effort in terms of moving towards customer-centric strategy, policy, and procedures, as well as sharing of knowledge in a big organization with lots of business 'silos'. Employee resistance to change posed a great challenge to the project. Ultimately, this project failed to achieve expectations.

Data Warehouse, Enterprise Databases, and KMS

White and Croasdell (2005) describe Xerox's use of an enterprise database to facilitate the sharing of experience knowledge across the company. Xerox had difficulty in fostering best practice among its group of printer maintenance employees. The problem centered on an inability to circulate employee expertise using existing organizational infrastructure. To help the maintenance technicians share their experience and expertise, Xerox created a database to hold top repair ideas in order to share those ideas with other technicians in all areas. This strategy called for only the most favored ideas to be kept within the database, as it often occurred that what one person thought useful, others found absurd or redundant. Xerox also realized that many

databases had been created by managers who filled the databases with information they thought would be useful for their employees. However, most of those databases were rarely used by the employees. When Xerox created the Eureka database, it also formed a process for entering and updating the ideas within the database. The process is based on a peer review system. Within this practice the representatives, not the organization, supply and evaluate tips. In this way a local expert would work with the representative to refine the tip. Representatives and engineers evaluate the tips, calling in experts where appropriate. As of July of 2000 the Eureka database held nearly 30,000 ideas and was being utilized by 15,000 Xerox technicians who answered a quarter-million repair calls per year. The shared knowledge in Eureka saved Xerox about $11 million in 2000, and customers also saved money in terms of the reduction in downtime.

Eureka later extended the role of the Eureka database to collect, share, and reuse solutions to software and network problems as well as those involving hardware. Additionally, Xerox Web-enabled, or made available over the Web, the Eureka database system, allowing technicians to gain access to the system from anywhere in the world though the Internet. The system added features including a search function, called "Search Light," and a wizard that aids in searching for tips, including those waiting to be validated. The technicians trust the Eureka system and constantly use the system because it helps them get any problem fixed quickly. In the old process many technicians would have to call a specialist to find solutions to problems they could not solve themselves.

As mentioned above, Al-Shammari (2005) describes the use of a data warehouse with its KCRM implementation. The data warehouse became the primary repository of data, information, and knowledge for the KCRM implementation and was designed using knowledge of customer needs and questions. Data mining tools were provided to support KCRM users in discovering knowledge contained in the data warehouse. Also as discussed above, key issues in implementing a data warehouse with KM were organizational culture as reflected in a "silo" mentality, and overall data, information, and knowledge ownership.

ADVANCED TECHNOLOGIES

Discussion

Although there is strong support for using the Internet as a knowledge infrastructure, there are concerns. Chief among these concerns is the difficulty in organizing, searching, and retrieving unstructured knowledge artifacts. Ezingeard, Leigh, and Chandler-Wilde (2000) point out that Ernst & Young UK in the beginning of 2000 had in excess of one million documents in its KMS. Another concern is the tendency to not use the system. Cross and Baird (2000) discuss this tendency but come to the conclusion that repositories are essential. Jennex (2007) found that use and importance for knowledge do not correlate, suggesting that use is not a true measure of the value of a KMS. Jennex and Olfman (2002) found that voluntary use is enhanced if the system provides near and long-term job benefits, is not too complex, and the organization's culture supports sharing and using knowledge and the system. Stenmark (2002) found that if the Internet is visualized as a system for increasing awareness of knowledge and the KMS, a system for retaining and sharing knowledge, and as a system for enhancing communication and collaboration between teams and knowledge experts and users, then it should be successful as a KMS. In all cases, researchers are experimenting with technologies that improve the handling of unstructured knowledge. These are discussed in the following paragraphs.

Technologies

Newman and Conrad (2000) propose a framework for characterizing KM methods, practices, and technologies. This framework looks at how tools can impact the flow of knowledge within an organization, its role in manipulating knowledge artifacts, and the organizational behavior most likely to be affected. The framework also looks at the part of the KM process the tool works in. The Activity phase looks at the utilization, transfer, retention, and creation of knowledge. This framework can be used to show that Internet and browser-based KMS tools are effective.

Gandon, Dieng, Corby, and Giboin (2000) propose using XML to encode memory and knowledge, and suggest using a multi-agent system that can exploit this technology. The proposed system would have improved search capabilities and would improve the disorganization and poor search capability normally associated with Internet systems. Chamberlin et al. (2001) and Robie, Lapp, and Schach (1998) discuss using XML query language to search and retrieve XML-encoded documents.

Dunlop (2000) proposes using clustering techniques to group people around critical knowledge links. As individual links go dead due to people leaving the organization, the clustered links will provide a linkage to people who are familiar with the knowledge of the departed employee. This technique would improve the reliability of the links to knowledge called for in Figure 2. Lindgren (2002) proposes the use of Competence Visualizer to track skills and competencies of teams and organizations.

Te'eni and Feldman (2001) propose using task-adapted Web sites to facilitate searches. This approach requires the site be used specifically for a KMS. Research has shown that some tailored sites, such as the ones dedicated to products or

communities, have been highly effective. This approach is incorporated in the examples in this article with the exception of the use of dynamic adaptation.

Eppler (2001), Smolnik, and Nastansky (2002) and Abramowicz, Kowalkiewicz, and Zawadzki (2002) discuss the use of knowledge maps to graphically display knowledge architecture. This technique uses an intranet hypertext clickable map to visually display the architecture of a knowledge domain. Knowledge maps are also known as *topic maps* and *skill maps.* Knowledge maps are useful as they create an easy-to-use standard graphical interface for intranet users and an easily understandable directory to the knowledge.

The use of ontologies and taxonomies to classify and organize knowledge domains is growing. Zhou, Booker, and Zhang (2002) propose the use of Rapid Ontology Development (ROD) as a means of developing an ontology for an undeveloped knowledge domain.

Making sense of seemingly unrelated structured data, information, and knowledge can also be difficult. Data mining is being used as a method for identifying patterns in this data, information, and knowledge that can then be assessed for meaning. Zaima and Kashner (2003) describe data mining as an iterative process that uses algorithms to find statistically significant patterns in structured data, information, and knowledge. These patterns are then analyzed by business process experts to determine if they actually have meaning in the business process context. CRM tends to use this technology the most, as illustrated by the example from Al-Shammari (2005).

Organizing and visualizing data and information into usable knowledge is a challenge that digital dashboard technologies are seeking to solve. Few (2005) describes dashboards as providing single-screen summaries of critical data and information. The key to developing effective dashboards is the use of KM to identify critical knowledge for key decision making and then linking it to the appropriate context data and information that indicates the status of the key knowledge. Dashboards can be used with a Internet browser or any other KMS infrastructure.

FUTURE TRENDS IN KNOWLEDGE MANAGEMENT

KM is a young and evolving discipline. The last section of this chapter looks at some of the emerging future trends affecting the discipline. Many of these trends have been discussed in relative detail in previous sections so they will just be summarized. The first trend is towards more formal measurement of KM success and performance. This is an indicator of discipline maturity, organizations are moving from "it's a good idea to do KM" to "KM has this impact on our performance."

A newer trend is a set of technologies used to empower knowledge groups such as communities of practice. KM was initially viewed as a formal process for managing organizational knowledge; this is still true, but it is also being recognized that KM can enable groups and teams to self-manage their knowledge. CoPs are groups that may or may not be in the same organization. What makes them a CoP is a shared common interest in a knowledge domain. CoPs have a shared context of understanding and may or may not share culture. CoPs self-identify critical knowledge and transfer it to those in the CoP that need it. Finally, CoPs need technology that facilitates CoP communication and collaboration. Wikis and other open source tools as well as knowledge portals are examples of technology used by CoPs. Murphy and Jennex (2006) illustrate how these tools in the hands of a CoP can lead to leaderless development in times of emergency and the quick creation of KM-enabled systems.

Finally, the Internet itself is changing. Efforts to improve the Internet's ability to facilitate collaboration and handle unstructured data, information, and knowledge, both by users and by software agents, is leading to the development of the Web 2.0 and Web 3.0/semantic Web. Web 2.0 refers to improving the Web to handle social networking sites and collaborative technologies such as Wikis, blogs, and other community-based sites (such as Craigslist, Skype, etc.). Web 3.0/semantic Web is predicted to be an environment where data, information, and knowledge can be understood and used by any user and software agent. Technologies used include Resource Description Framework (RDF), Web Ontology Language (OWL), and the Extensible Markup Language (XML). Both of these developments will improve the ability of the Web to store, search, and retrieve knowledge and to facilitate knowledge transfer.

CONCLUSION

The conclusion is that the Internet, data warehouse, CRM, and ERP are effective infrastructures for a KMS. However, there are issues associated with using these technologies that KMS designers need to be aware of. Chief among these are knowledge representation and search. Several tools such as Knowledge Maps, XML, adaptive Web sites, clustering, and dashboards are being developed to address these issues. However, as knowledge bases grow, designers need to be aware of increasing search times as well as a variety of knowledge artifacts. This is perhaps the most important area for future research. Developing ontologies and taxonomies to aid in classifying and structuring knowledge domains is critical.

REFERENCES

Abramowicz, W., Kowalkiewicz, M., & Zawadzki, P. (2002). Tell me what you know or I'll tell you what you know: Skill map ontology for information technology courseware. *Proceedings of the 2002 Information Resources Management Association International Conference.*

Al-Shammari, M. (2005). Implementing knowledge-enabled CRM strategy in a large company: A case study from a developing country. In M.E. Jennex (Ed.), *Case studies in knowledge management* (pp. 249-278). Hershey, PA: Idea Group.

Alavi, M., & Leidner, D.E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Bartczak, S.E., & England, E.C. (2005). Challenges in developing a knowledge management strategy for the Air Force Material Command. In M.E. Jennex (Ed.), *Case studies in knowledge management* (pp. 104-128). Hershey, PA: Idea Group.

Chamberlin, D., Clark, J., Florescu, D., Simon, J., Robie, J., & Stofancscu, M. (2001). *Xquery 1.0: An XML query language* (WSC working draft 2001). Retrieved from http://www.w3.org/TR/xquery/

Churchman, C.W. (1979). *The systems approach* (revised and updated ed.). New York: Dell.

Corral, K., Griffin, J., & Jennex, M.E. (2005). Expert's perspective: The potential of knowledge management in data warehousing. *Business Intelligence Journal, 10*(1), 36-40.

Cross, R., & Baird, L. (2000). Technology is not enough: Improving performance by building organizational memory. *Sloan Management Review, 41*(3), 41-54.

Davenport, T.H., & Prusak, L. (1998) *Working knowledge.* Boston: Harvard Business School Press.

Dunlop, M.D. (2000). Development and evaluation of clustering techniques for finding people. *Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management* (PAKM2000).

Eppler, M.J. (2001). Making knowledge visible through intranet knowledge maps: Concepts, elements, cases. *Proceedings of the 34th Hawaii International Conference on System Sciences.*

Ezingeard, J.-N., Leigh, S., & Chandler-Wilde, R. (2000). *Knowledge management at Ernst & Young UK: Getting value through knowledge flows* (teaching case, pp. 807-822).

Few, S., (2005). Dashboard design: Beyond meters, gauges, and traffic lights. *Business Intelligence Journal,* 18-24.

Gandon, F., Dieng, R., Corby, O., & Giboin, A. (2000). A multi-agent system to support exploiting an XML-based corporate memory. *Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management* (PAKM2000).

Holsapple, C.W., & Joshi, K. (2004). A formal knowledge management ontology: Conduct, activities, resources, and influences. *Journal of the American Society for Information Science and Technology, 55*(7), 593-612.

Jennex, M.E. (2000). *Using an intranet to manage knowledge for a virtual project team. Internet-based organizational memory and knowledge management.* Hershey, PA: Idea Group.

Jennex, M.E. (2005a). *Case studies in knowledge management.* Hershey, PA: Idea Group.

Jennex, M.E. (2005b). The issue of system use in knowledge management systems. *Proceedings of the 38th Hawaii International Conference on System Sciences.*

Jennex, M.E. (2005c). What is KM? *International Journal of Knowledge Management, 1*(4), i-iv.

Jennex, M.E. (2007). Exploring system use as a measure of knowledge management success. *Journal of Organizational and End User Computing.*

Jennex, M.E., & Olfman, L. (2002). Organizational memory/knowledge effects on productivity, a longitudinal study. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.*

Jennex, M.E., & Olfman, L. (2005). Assessing knowledge management success. *International Journal of Knowledge Management, 1*(2), 33-49.

Jennex, M.E., & Olfman, L. (2006). A model of knowledge management success. *International Journal of Knowledge Management, 2*(3), 51-68.

Koch, C. (2002, February 7). *The ABC's of ERP.* Retrieved from http://www.cio.com/research/erp/edit/erpbasics.html

Li, Z., Yezhuang, T., & Ping, L. (2005). Organizational knowledge sharing based on the ERP implementation of Yongxin Paper Company, Limited. In M.E. Jennex (Ed.), *Case studies in knowledge management* (pp. 156-164). Hershey, PA: Idea Group.

Lindgren, R. (2002). Competence Visualizer: Generating competence patterns of organizational groups. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.*

Maier, R. (2002). *Knowledge management systems: Information and communication technologies for knowledge management.* Berlin: Springer-Verlag.

Murphy, T., & Jennex, M.E. (2006). Knowledge management and Hurricane Katrina response. *International Journal of Knowledge Management, 2*(4), 52-66.

Newman, B., & Conrad, K. (2000). A framework for characterizing knowledge management methods, practices, and technologies. *Proceedings of the 3rd International Conference on Practical Aspects of Knowledge Management* (PAKM2000).

Nissen, M., & Jennex, M.E. (2005). Knowledge as a multidimensional concept: A call for action. *International Journal of Knowledge Management, 1*(3), i-v.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company—how Japanese companies create the dynamics of innovation.* Oxford: Oxford University Press.

Polanyi, M. (1967). *The tacit dimension.* London: Routledge and Kegan Paul.

Robie, J., Lapp, J., & Schach, D. (1998). XML Query Language (XQL). *Proceedings of the Query Language Workshop.*

Smolnik, S., Kremer, S., & Kolbe, L. (2005). Continuum of context explication: Knowledge discovery through process-oriented portals. *International Journal of Knowledge Management, 1*(1), 27-46.

Smolnik, S., & Nastansky, L. (2002). K-discovery: Using topic maps to identify distributed knowledge structures in groupware-based organizational memories. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.*

Stein, E.W., & Zwass, V. (1995). Actualizing organizational memory with information systems. *Information Systems Research, 6*(2), 85-117.

Stenmark, D. (2002). Information vs. knowledge: The role of intranets in knowledge management. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.*

Te'eni, D., & Feldman, R. (2001). Performance and satisfaction in adaptive Websites: An experiment on searches within a task-adapted Website. *Journal of the Association for Information Systems, 2*(3).

White, C., & Croasdell, D. (2005). A comparative case study of knowledge resource utilization to model organizational learning. In M.E. Jennex (Ed.), *Case studies in knowledge management* (pp. 235-248). Hershey, PA: Idea Group.

Zaima, A., & Kashner, J. (2003). A data mining primer for the data warehouse professional. *Business Intelligence Journal,* 44-54.

Zhou, L., Booker, Q.E., & Zhang, D. (2002). ROD—toward rapid ontology development for underdeveloped domains. *Proceedings of the 35th Annual Hawaii International Conference on System Sciences.*

KEY TERMS

Knowledge: An evolving mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information (Davenport & Prusak, 1998).

Knowledge Management: The process established to capture and use specific knowledge in an organization for the purpose of improving organizational performance.

Knowledge Management System: The system created for users to interact with the organizational memory system.

Knowledge Map: An intranet hypertext-clickable map to visually display the architecture of a knowledge domain. Knowledge maps are also known as topic maps and skill maps.

Knowledge Ontology: Common definitions established for captured knowledge, similar to keywords, used to capture and express a common context for search, retrieval, and use of knowledge.

Knowledge Taxonomy: The hierarchical organization of knowledge categories within a knowledge management system.

Technology and Transformation in Government

Vincent Homburg
Erasmus University Rotterdam, The Netherlands

INTRODUCTION

Information technology and public administration are an odd couple. Students of information technology have long neglected arduous issues of public sector reform and public policymaking (Borins, Kernaghan, Brown, Bontis, & Thompson, 2007; Orlikowkski & Barley, 2001). Likewise, public administration scholars have rarely paid attention to information technology beyond treating it pragmatically (Gruening, 2001), at the periphery of governments' core activities of policy making and policy implementation. This situation of disciplinary negligence, however, has changed since the advent of the admittedly voguish term electronic government ("e-government"). E-government refers to a practice in which governments throughout the world embrace information and communication technologies in order to transform the machinery of governance (Bekkers & Homburg, 2007; Borins et al., 2007; Chadwick & May, 2003; Dunleavy, Margetts, Bastow, & Tinkler, 2006; Heeks, 2006).

The relation between technology and transformation is not as straightforward as might appear at first sight (Williams & Edge, 1996), for at least two reasons. First, the clamor for transformation and reform was first heard in the beginning of the 1990s (Osborne & Gaebler, 1992) without technology playing a role. Rather, the focus was on organizational and managerial changes, in particular focusing on establishing customer orientation and use of market-type mechanisms (Guy Peters, 1996; Hood, 1991; Pollitt, van Thiel, & Homburg, 2007), that later blended with the emergence of new technologies. Second, e-government practices throughout the world display a huge variety of forms, shapes and effects that are not easily attributed to technology alone. In the national policies of the United Kingdom and the United States, for instance, the focus is on achieving one-stop service shops that enable transactions with citizens on the basis of clearly defined "service themes" (Chadwick & May, 2003). At municipal levels in Sweden, on the other hand, e-government takes the form of electronic interactions between municipal commissioners and citizens, in such a way that citizens can watch video broadcasts of city council meetings, and can submit questions to commissioners during the half-way break (Grönlund, 2003).

The above discussion makes clear that the use of ICTs in government has moved from being a peripheral concern, to a topic that concerns the core activities of government, policy making and policy implementation, and that e-government is intrinsically linked to transformation and reform of governments. It does not, however, make clear how to circumscribe and define "e-government," and what obstacles and dilemmas can be witnessed in practice. The remainder of this chapter addresses these issues.

BACKGROUND

Electronic government (or e-government) has emerged as a powerful catchphrase to indicate situations in which ICTs are associated with bureaucratic renewal and institutional innovation in general (Homburg & Bekkers, 2005). The term New Public Management appeared in the 1980s in Anglo-American discussions about how to reform rather traditional bureaucratic structures and practices. One of the dominant observations related to bureaucratic renewal and New Public Management was that it truly was management ideology: In talk, writing and discussions, there was a powerful and almost compelling rhetoric of administrative reform, yet in practice the clamor for reform suffered from a lack of useful and practical instruments with which actual change could be accomplished. Since the advent of Web technology, many reform adepts have embraced information and communication technology, and have used the concept of e-government as a "tool" to actually implement changes in and around governments. In *The Economist* of June 24, 2000, it is stated that the once fashionable idea of reinventing government, is now finally being made possible by the Internet (Symonds, 2000).

Central to the reform ideas at the corner stones of New Public Management and the emergence of communication technologies is the focus on client (or citizen) orientation. Not surprisingly, many definitions of e-government emphasize electronic service delivery as a main objective for e-government (for a review, see Yildiz, 2007), thus portraying e-government as "e-commerce for governments" (Wimmer, Traunmüller, & Lenk, 2001). There are, however, various arguments for declaring such a definition too narrow in focus (Bekkers & Homburg, 2005b).

First, e-commerce concerns itself with transactions between suppliers and customers. If we extrapolate that to ICTs

in relation to government, we see that the notion of "customer" is far more problematic. Citizens can be customers, in the sense that they are beneficiaries of public services, but at the same time they are co-creators of the policies (in the case of Bollnäs mentioned above), and, more importantly, they are sometimes involuntarily involved in transactions with governments (e.g., in the case of electronic tax services and electronically administered fines for speeding).

Second, the objectives of e-government applications address, in many cases, more values than efficiency of service delivery and customer orientation alone. E-government implementations can also serve other purposes like increasing transparency of the government apparatus (Homburg, 2008; LaPorte, de Jong, & Demchak, 2000), bridging the gap between citizens and administration (Bekkers & Homburg, 2005a), or addressing (and preferably decreasing) the democratic deficit.

Third, many public electronic one-shop facilities necessitate data sharing and standardization of practices among multiple, relative autonomous agencies in order to provide integrated services. From a technological point of view, it is understood that data sharing is severely hampered by lack of consistency of data and, in general, a lack of data standardization. In the information systems literature, various Strategic Information Systems Planning (SISP) methodologies have been proposed that can be put to use to alleviate this situation. In specific e-government initiatives, however, data sharing is not so much hampered by more or less operational inconsistencies, but rather by checks-and-balances (e.g., between executive and judicial branches in penal law enforcement) and disagreement over professional values (of social workers and medical professionals in cases of child protection services).

Fourth, it may be tempting to assume that e-government is a more or less direct translation of a global, unequivocal and consistent wave of administrative reform, New Public Management. A closer look at the phenomenon New Public Management reveals, on the other hand, that the trajectories of reform are different in various institutional contexts (Pollitt et al., 2007). New Public Management takes many forms and shapes in Singapore as opposed to Denmark, Spain, or Guatemala, to name a few institutional contexts, and so does e-government. This issue is furthermore addressed in the subsequent section.

In recognition of the arguments set out above, e-government is defined not as e-commerce for government, but rather as a redesign of information relations of a public agency with stakeholders in its environment (Bekkers & Homburg, 2005b; Homburg, 2008). Redesign, in this definition, can apply to front offices, that is, to relations between governments and citizens (in either of the roles of customer, voter, "citoyen" and subordinate of policy) but also to back offices, indicating a redesign of information relations between various agencies, or even branches of government. The various issues and obstacles of redesign are presented and discussed in the following section.

REDESIGN OF INFORMATION RELATIONSHIPS: RESULTS AND ISSUES

A first issue concerns the type of front office services (at the government-society interface) that are offered in the actual practice of e-government: information services, contact services, transaction services and participation services. Chadwick and May have convincingly argued that explaining the kinds of front office services offered (i.e., whether authorities offer information services, or electronic participation services, or several of these kinds of services at the same time), is not so much a technological issue of "maturity" but is the result of an underlying normative frame of reference (Chadwick & May, 2003). Hence, understanding e-government is not a question of understanding technology, but rather grasping the concept of democracy (Barber, 1997).

In practice of national e-government initiatives at the federal level in the United States (Chadwick & May, 2003) and national levels of policy making in the United Kingdom (Bekkers & Homburg, 2007; Chadwick & May, 2003), the Netherlands, Denmark and Australia (Bekkers & Homburg, 2007), e-government policies focus on information services and especially on transaction services, especially at the expense of participation services.

A second issue with the redesign of information relationships is if and how the *appearance* of service counters (i.e., the *front office*) relates to the back office organizational (departmental or interorganizational) structures. Alfred Tat-Kei Ho (2002) analyzed Web portals of 55 of the most populous cities in the United States. He concluded that most cities had, over time, transformed their Web presence from an administrative-oriented portal design (reflecting bureaucratic logic of a variety of functionally differentiated departments) to user-oriented portals. Furthermore, responses by city Web masters indicated that many city officials had abandoned a departmental mentality in Web management. Donald Norris, on the other hand, noticed that most municipal Web sites offer information services, but few transaction services. Moreover, Norris concluded on the basis of survey data of American local authorities that services that horizontally or vertically span various authorities, are notably lacking (Norris, 2005). Obviously, e-government is being used to break down departmental barriers, but collapsing interorganizational boundaries still results in many problems.

A third issue, which is related to the issue of collapsing interorganizational boundaries mentioned above, is the re-organization of information relations in the back office. Sharing information across organizational boundaries is far

from an operational, neutral issue, neither in the private sector (Webster, 1995) nor in the public sector (Homburg, 2000). The issue of back office interorganizational information exchange is addressed in the public management literature by using the term joined-up government (JUG). Joined-up government refers to the aspiration to achieve horizontal and vertical coordination ("cross-cutting approaches") in and among organizations (Ling, 2002; Pollitt, 2003). Both the information systems literature as well as the public management literature report formidable difficulties in actually achieving interorganizational information integration and joined-up government. It may be tempting to propose all-encompassing and unequivocal ontologies (which in fact takes place at numerous e-government conferences). A recent European study, however, has shown that there is in practice a large divergence in ways in which joined-up government and integration among agencies is actually achieved (Millard, Iversen, Kubicek, Westholm, & Cimander, 2004), and that JUG and integration are far more contextually bound than is apparent at first sight. The authors compared JUG implementations in several service clusters: income tax, car registration, citizen certificates, family allowances, student grants, social benefits, building permissions, enrolment in higher education, citizen portals, social contributions for employers, corporation taxes, customs declarations, business registration, environmental-related permits and business portals in Austria, the Benelux countries, Finland, France, the UK, Greece, Iceland, Ireland, Italy, Norway, Portugal, Spain and Sweden, and found a large variety of mechanisms with which interoperability and joined-up government was actually enforced (Millard et al., 2004, pp. 27-44):

- by means of digitization of organizationally unchanged back-offices;
- by means of fundamental redesign of back-offices;
- by means of centralization of back-offices;
- by means of simultaneous centralization and back-offices and decentralization of front-offices;
- by means of modularizing common back-office components over broad areas, while retaining flexibility and possibilities to adapt to specific requirements; and
- by means of clearing houses that enable the exchange of information from various sources without the need to necessarily integrate data sources.

The authors conclude that "One of the clearest conclusions emerging from the present study is that state structures, and institutional, legal, regulatory and cultural factors, can be extremely important in determining the nature, cost and success of eGovernment. (…) Different countries across Europe need to develop their own paths as each has unique identities, cultures, legal systems and institutional structures" (Millard et al., 2004, p. 61).

The importance of institutional context is exemplified in a study on integration and joined-up government in the back office of social security in the Netherlands and Belgium (Homburg, 2007). Although the Netherlands and Belgium are, seen from a distance, rather comparable nation states (in terms of size, geographic location, etc.) and the specific sector of social insurance has been confronted with comparable issues (fiscal crises since the 1970s, reforms of welfare state regimes, etc.), and hence, one would expect a certain amount of resemblance of technological solutions to integrate these sectors. In practice rather different solutions were implemented to allow for joined-up government. In Belgium, a relatively centralized initiative was taken which resulted in changes in legislation, the introduction of a new, powerful player in the form of the Cross Point Bank (a centralized clearing house), and large-scale standardization throughout organizations in the sector, enabled by missionary activities of the Cross Point Bank. In the Netherlands, on the other hand, executive agencies themselves initiated a decentralized clearing house, governed by representatives of executive agencies, thereby fostering decentralization and strengthening the power base of executive organizations (vis-à-vis policy making bodies like the Ministry of Social Affairs and Employment). The diversity of organizational form and organizational practice can be explained in terms of varying power bases of the executive in social security in Belgium and the Netherlands, but also in terms of technological competencies in policy making and policy implementation.

FUTURE TRENDS

Setting the difficulties of understanding e-government and explaining the diversity of form and impact aside, it is possible to discern a number of future trends.

A first trend—observable in the government-society interface—that can be observed in the realm of public service delivery is a trend toward further personalization. Increasingly, service delivery by a variety of authorities is organized around a limited number of *life events* and is presented to individual citizens and companies in a personalized manner: Citizens increasingly do not have to know anymore which agency exactly is responsible for specific services, and services are tailored as much as possible toward individual circumstances. In the Netherlands, for example, municipal electronic service counters increasingly become the *de facto* integrated service outlets for various governmental agencies.

A second trend—observable in the back-office of government—is the formation of information networks, implying a transformation of traditional bureaucracies into virtual public bureaucracies (Bekkers, 1998), called infocracies by Zuurmond (1998). Characteristic for the latter types of

organization is the disappearance of functionally defined organizational boundaries, and changes in ways in which accountability takes place. Whereas political, vertically-oriented accountability is a central value in traditional public bureaucracies, this type of accountability is blended with horizontal and public forms of accountability. Horizontal accountability is accountability targeted at peers (comparable organizations). Public accountability is focused at increasing legitimacy in the eyes of the general public.

CONCLUSION

E-government has been defined as redesign of information relations between government agencies and stakeholders in their environments. In this chapter, the focus was on understanding variety in forms and shapes that e-government initiatives take, both in the front office (the appearance of service counters and other manifestations of government-society interfaces) as well as in the back office (cooperative relations and information exchanges between governmental agencies). Although it is tempting to assume "maturity" models of growth here, understanding front office and back office developments in e-government initiatives requires in-depth understanding of values, norms, normative frames of reference and taken for granted assumptions (in short: institutions) of localized worlds of government and public administration. If we accept this view and its consequences, we see that developing e-government applications is by far a neutral exercise, and that design requirements are constituted in settings in which normative frames of reference may collide. Developing and implementing e-government initiatives, therefore, is a complex endeavor in which politicians, policymakers, public managers and developers face formidable challenges. Nevertheless, e-government progresses and has shown to be able to actually instrument government reform and desired institutional change.

REFERENCES

Barber, B. (1997). The new telecommunications technology: Endless frontier or the end of democracy. *Constellations*, (4), 208-228.

Bekkers, V. J. J. M. (1998). Wiring public organizations and changing organizational jurisdictions. In I. T. M. Snellen & W. B. H. J. v. d. Donk (Eds.), *Public administration in an information age* (pp. 57-77). Amsterdam: IOS Press.

Bekkers, V. J. J. M., & Homburg, V. M. F. (2005a). The information ecology of e-government: Background & concepts. In V. J. J. M. Bekkers & V. M. F. Homburg (Eds.), *The information ecology of e-government: E-government as institutional and technological innovation in the public dector* (pp. 1-20). Amsterdam: IOS Press.

Bekkers, V. J. J. M., & Homburg, V. M. F. (Eds.). (2005b). *The information ecology of e-government*. Amsterdam: IOS Press.

Bekkers, V. J. J. M., & Homburg, V. M. F. (2007). The myths of e-government: Looking beyond the assumptions of a new and better government. *The Information Society, 23*(5), 373-382.

Borins, S., Kernaghan, K., Brown, D., Bontis, N. P., & Thompson, F. (2007). *Digital state at the leading edge*. Canada: Toronto University Press.

Chadwick, A., & May, C. (2003). Interaction between states and citizens in the age of the Internet: "E-government" in the United States, Britain, and the European Union. *Governance, 16*(2), 271-300.

Dunleavy, P., Margetts, H., Bastow, S., & Tinkler, J. (2006). *Digital era governance (IT corporations, the state and e-government)*. Oxford: Oxford University Press.

Grönlund, A. (2003). Emerging electronic infrastructures (Exploring democratic components). *Social Science Computer Review, 21*(1), 55-72.

Gruening, G. (2001). Origin and theoretical basis of New Public Management. *International Public Management Journal, 4*(1), 1-25.

Guy Peters, B. (1996). *The future of governing: Four emerging models*. Kansas University Press.

Heeks, R. (2006). *Implementing and managing e-government (An international text)*. London: SAGE.

Homburg, V. M. F. (2000). Politics and property rights in information exchange. *Knowledge, Policy and Technology, 13*(3), 13-22.

Homburg, V. M. F. (2007). A comparative account of joined-up government initiatives in Dutch and Belgian social security. *International Journal of Cases on Electronic Commerce, 3*(2), 1-12.

Homburg, V. M. F. (2008). *Information systems and public administration: Understanding e-government*. London: Routledge.

Homburg, V. M. F., & Bekkers, V. J. J. M. (2005). E-government and NPM: A perfect marriage? In V. J. J. M. Bekkers & V. M. F. Homburg (Eds.), *The information ecology of e-government: E-government as institutional and technological innovation in public administration* (pp. 155-170). Amsterdam, Berlin, Oxford, Tokyo, Washington, DC: IOS Press.

Hood, C. (1991). A public management for all seasons? *Public Administration, 69*(1), 3-19.

LaPorte, T. M., de Jong, M., & Demchak, C. C. (2000). Public organizations on the World Wide Web: Empirical correlates of administrative openness. *Administration & Society, 34*(4), 411-446.

Ling, T. (2002). Delivering joined-up government in the UK: Dimensions, issues and problems. *Public Administration, 80*(4), 615-642.

Millard, J., Iversen, J. S., Kubicek, H., Westholm, H., & Cimander, R. (2004). *Reorganisation of government back-offices for better electronic public services--European good practices (back-office reorganisation)*. Brussels: EU DG Information Society.

Norris, D. F. (2005). Advancing e-government at the grassroots: Tortoise or hare. *Public Management Review, 65*(1), 64-75.

Orlikowkski, W. J., & Barley, S. R. (2001). Technology and institutions: What can research on information technology and research on organizations learn from each other? *MIS Quarterly, 25*(2), 145-165.

Osborne, D., & Gaebler, T. (1992). *Reinventing government: How the entrepreneurial spirit is transforming the public sector*. Reading, MA: Addison-Wesley.

Pollitt, C. P. (2003). Joined-up government: A survey. *Political Studies Review, 1*(1), 34-49.

Pollitt, C. P., van Thiel, S., & Homburg, V. M. F. (Eds.). (2007). *New public management in Europe: Adaptation and alternatives*. Basingstoke: Palgrave McMillan.

Symonds, M. (2000). The next revolution. *The Economist, 355*(8176).

Tat-Kei Ho, A. (2002). Reinventing local governments and the e-government initiative. *Public Administration Review, 62*(4), 434-444.

Webster, J. (1995). Networks of collaboration of conflict? Electronic data interchange and power in the supply chain. *Journal of Strategic Information Systems, 5*(1), 31-42.

Williams, R., & Edge, D. (1996). The social shaping of technology. *Research Policy, 25*, 865-899.

Wimmer, M., Traunmüller, R., & Lenk, K. (2001). Electronic business invading the public sector: Considerations on change and design. In *Proceedings of the 34th Hawaii International Conference on Information Systems*.

Yildiz, M. (2007). E-government research: Reviewing the literature, limitations, and ways forward. *Government Information Quarterly, 24*(6), 646-665.

Zuurmond, A. (1998). From bureaucracy to infocracy. In I. T. M. Snellen & W. B. J. H. van de Donk (Eds.), *Public administration in an information age* (pp. 259-272). Amsterdam: IOS Press.

KEY TERMS

E-Government: Redesign of information relations of a public agency with stakeholders in its environment.

NPM (New Public Management): Management ideology with which private-sector business management techniques (performance management systems, benchmarking, autonomization) are introduced in the public sector.

JUG (Joined Up Government): Aspiration to aspiration to achieve horizontal and vertical coordination ("cross-cutting approaches") in and among public sector organizations.

Institutions: Values, norms, normative frames of reference, taken for granted assumptions and practices.

Front Office: Part of the organization that is specialized in interaction with society (citizens, companies) and that is responsible for, among other things, managing the government-society interface.

Back Office: Part of the organization that is specialized in meeting information requirements of front office processes and is responsible for, among other things, registering and exchanging information between public, private and hybrid organizations.

Technology Discourses in Globalization Debates

Yasmin Ibrahim
University of Brighton, UK

INTRODUCTION

Globalization, a key concept in our modern and postmodern discourse, is a highly contentious term that continues to generate endless debates about its form and consequences on our societies. Anthony Giddens (1999) professes that while the term is "not particularly attractive or an elegant one, absolutely no one who wants to understand our prospects and possibilities can ignore it." While many agree that it denotes the occurrence of social change, there is, however, less agreement what these changes may be and whether they, in effect, represent the transition of one form of society to another (i.e., the industrial to the postindustrial or information society). Nevertheless, the increase in the volume of discourses surrounding the term is significant in illuminating that the increased interdependence of the world can lead to new forms of challenges, concerns, empowerment, and resistance with the symbolic and material exchanges of ideas, products, and services, as well as the formation of social networks (Castells, 1998). Castells (1996, 2000, 2001), in his numerous reflections on the network society, asserts that since the 1980s, a new economy has emerged that is global, information-based, and interconnected. This new form of economy remains capitalist in form but is situated on an informational rather than an industrial form of development; at the core of the informational mode of development are networks contributing to a network society.

The term globalization then captures a complex set of processes that involve political, social, economic, cultural, and technological factors, and these intersect with each other in crucial but unpredictable and uneven ways (Stammers & Eschle, 2005, p. 57). Crucial to the debates on globalization is the role of information and communication technologies (ICTs) in making the world a connected entity. ICTs are often seen as providing a new platform to enable the flows, exchanges, and networks in analysing globalization as a multidimensional process. According to Hamid Mowlana (1997), a useful way of viewing the "postindustrial age" or the "information age" is to look at it as a tangible or material infrastructure being built into contemporary societies (p. 176). In this sense, modern communication and information technologies, in the form of satellites, computers, and radar, comprise the new infrastructure. Characteristics of new media technologies, like the Internet and mobile phones, include digitization, convergence, and networking, which enables new ways to distribute, store, consume, and interact with information that facilitates a networked distribution on a global scale (Flew, 2007, p. 22).

BACKGROUND

While globalization is a relatively a new concept that has gained much currency and controversy in the last two to three decades, the idea of high-speed transport and communications in some ways altering our social reality in terms of the construction of the temporal or social space is not in itself a new idea. There is an abundance of references in literary texts and historical annals to human interaction and concepts of geography being negotiated through technological innovations. In literary fiction, specifically, the future is often imaged through the advances in technology that prophesy degrees of technological determinism on human society. There have been references to a global economy through industrialization and technological innovations dating back from 1870 to 1914 (See Kobrin). Marshall McLuhan's (1964) concept of the "global village" constructed the notion of a world community brought about by communication technologies and, consequently, technical biases, he argues, were intrinsic to our cognitive constructions of reality.

Similar ideas were propounded by Martin Heidegger (1971), who prophesied the "abolition of distance" as a distinctive feature of our modern condition where all "distances in time and space are shrinking" (p. 165). Addressing the ability of modern technology to premise on simultaneity and instantaneousness, Heidegger perceived technology as a levelling experience for individuals and in the process it produces a loss of meaning in defining what is far or near. Equally, Robertson (1992) alludes to the compression of time and space and an intensification of a consciousness of the world. Benedict Anderson (1991), in *Imagined Communities*, locates cultural artefacts and technological innovations, including books and press, as crucial in shaping our social imagination, political engagements, and our notions of community. Much of the rhetoric on globalization, as such, has revolved around the notion of "deterritorialization," where the material space is replaced with social activities that emphasise the connectedness of the world. However,

globalization, in our contemporary context, is deemed to be more expansive in terms of the nations involved and deeper in terms of the intensity of interactions and interdependence (Haleja, 2005, p. 7). In later debates, globalization has entailed the reconfiguration of material spaces where social space is no longer tightly defined by territoriality and, as such, the world becomes borderless. In a nutshell, globalization as advanced by ICTs is seen as both empowering as well as causing conflict and divisiveness within and between countries (Wilson, 1998, p.2).

The Implication of ICTs

Globalization has been discussed as a philosophy of ideas and as empirically (i.e., quantitatively and qualitatively) discernible processes that are occurring throughout the world. The process of change associated with globalization has been defined and viewed from various perspectives spanning the economic, social, political, legal, technological, and cultural, and has involved a range of viewpoints from the cynics to the ardent supporters of an open, borderless world of trade, commerce, and symbolic exchanges. Kaldor et al. (cf. Zanfei, 2005, p. 7) refer to the different perspectives of economic globalization ranging from the "supporters" who have championed the value of free international integration of economies since the early 1990s, to the "regressives" who have endorsed globalization when only deemed beneficial to an indigenous nation, and "rejectionists" who have clamoured for the greater protection of national economies.

These debates often implicate information and communication technologies (ICTs) as driving the change. ICTs encompass the full range of the "production, distribution and consumption of messages, across all media from radio and television, to satellite to Internet, and tangentially the 'information revolution' denotes the rapid advances in the power and speed of computers, the digitalization of information, and the convergence of once-separate industries to a new amalgam of production, distribution and consumption activities" (Wilson, 1998, pp. 6-7). According to Wilson (1998), this conveys both the "cross-border flow of information content as well as hardware used nationally and locally to produce, distribute and consume information" (pp. 6-7). Globalization has also been viewed as being primarily driven by the Internet and ICTs, which enable peoples, ideas, investments, goods, and services to come together through interconnected economies (Lang, 2001). Manuel Castells (2000) situates ICTs as fundamental in enabling the expansion of social and organizational networks in the information age, where information generation and processing are intrinsically entwined in the transformation of societies.

From this perspective, "globalization and technological advances in ICTs signify a fundamental transformation of the economy" (Alecke & Untiedt, 2000). Inevitably, the ICT revolution has been viewed as the primary catalyst for the process of globalization. It provides the tools for the postindustrial age and the foundations for a knowledge economy facilitating the rapid transfer and acquisition of knowledge (Ajayi, 2000; Morales-Gomez & Melesse, 1998). The wiring of the planet through ICTs has been seen as the "death of distance" (Caincross, 1997), where there is a shrinking of physical spaces and temporal distances. Anthony Giddens (1990, p. 64) similarly stresses the reframing of social geography due to globalization where there is "an intensification of worldwide social relations which link distant localities in such a way that local happenings are shaped by events occurring many miles away and vice versa." While the death of distance through the speed of communication presents opportunities for economies and polities, the increasing modernization and industrialization, as well as the attendant interconnectedness of the world, is also seen as giving rise to new forms of risk (Beck, 1992). Ulrich Beck (1992) argues that "a universalization of hazards accompanies industrial production, independent of the place where they are produced; food chains connect practically everyone on earth to everyone else" (p. 39). For example, terrorism in the age of globalization is seen as a global threat that can have consequences for both the global economy as well as individual governments, posing new forms of security threat which require both local and global governance.

In fact, reductionist models of globalization have limited their analyses to primarily economic and/or technological perspectives. These perspectives stress the "growing integration and liberalisation of worldwide markets, the development of communications and transport technologies and the rapid growth of global governance institutions above and beyond the state" (Stammers & Eschle, 2005, p. 55). Marxist and neo-Marxists view technological and institutional developments as indicative of a shift in the more fundamental structures of capitalism while being uncertain of whether these have constrained or empowered local spaces (Stammers & Eschle, 2005, p. 55).

Giddens (1990, 1999) defines globalization as not a singular process, but a complex set of processes that operate in a contradictory and dialectical manner. While nations lose a degree of economic power through these processes, Giddens reiterates that globalization has also been the reason for cultural revival of local identities in different parts of the world. In tandem with the resonant theme of deterritorialization, the discourses of globalization in its initial phases also mooted the withering of nation states and the inability of politicians to influence events. Hardt and Negri (2000) propound that in the age of globalization, a system of imperialism is created through a network of global and supranational entities that have forced nation states in some ways to compromise their jurisdiction and sovereignty. While postmodernity has not witnessed the demise of the nation state, it has nevertheless witnessed the emergence of global governance through intermediary and international institu-

tions such as the World Trade Organization, International Telecommunication Union, and the World Intellectual Property Organization (WIPO). The nation state is vulnerable to market pressures to liberalize and attract investments and equally, to preserve its sovereignty and jurisdictions despite these external and global pressures.

As such, globalization has also been equated with internationalization and liberalization denoting the increase in interdependence between nation states and the "removal of government-imposed restrictions on movements between countries in order to create an open and borderless world economy" (Scholte, 2000, p. 6). In tandem with the discourse of globalization and proliferation of ICTs is the use of the term "Information Society," which stresses the "rapid adoption and diffusion of new ICTs which has also precipitated the call for the liberalization of the telecommunications industry to reduce the barriers to the adoption of the new technologies and to ensure equal access to the global infrastructure" (Wolfe, 2000).

On the other hand, globalization has been associated with hegemonic discourses of cultural imperialism, where there is a flow of ideas and tangible and intangible cultural goods from the West to the East. Thus, globalization has been discussed as synonymous with modernization or westernization where the social structures of modernity, such as industrialization and rationalization, are spread over the world. Information and entertainment themselves have emerged as major industries and are presently the fastest growing segments of the global economy, characterised by the highest degrees of concentration and centralization compared to other sectors (Ghosh, 2004, p. 101). As Ghosh (p. 101) asserts, the "multimedia boom has spawned large multimedia companies that constitute the largest Multinational Corporations (MNCs). Cultural imperialism is then manifested in the bulk of the content, forms of expressions and structures of ownership and management reflect the domination of the core capitalist countries, especially the US." Nicholas Negroponte argues that in the information age, the mass media is both small and big, with global players, such as CNN, that reach a global audience. In comparison, new types of technology have also enabled narrowcasting, which caters to small demographic groups.

As Stammers and Eschle (2005, p. 57) point out, one main argument that has dominated discussions about globalization is the "tension between homogenising and fragmenting tendencies within globalizing processes and potential emergence of diverse innovative and hybrid cultural, political and even economic forms." In line with this, local responses to global processes become a major issue of scrutiny in globalization debates, where localization denotes both the tensions between the local spaces and global processes as well as cultural and social appropriation, negotiations, and hybridization that can happen in local spaces as a response to these global processes and influences. Robertson (1992), though acknowledging the time-space compression enabled by ICTs, contends that it also accentuates the tensions between global, societal, and communal attitudes and in this sense, while certain imported themes can be indigenized, others may be resisted or adapted in particular ways. In a similar vein, Appadurai (1997) negates the homogenization thesis through a recognition that the appropriation of ideas and practices in indigenous societies can be mediated through their cultural landscape (i.e., geographies, histories, or languages).

Thoman Friedman (2005) proposes that we are entering an accelerated phase, which he terms as Globalization 3.0, that captures the global changes facilitated by the semantic Web, where there is a fusion of people and lifestyles through a complex layering of economies and cultures. For Friedman, this entails ubiquitous computing, fibre optics, better bandwidth capacity, and software that connects nations, corporations, and people in an unrelenting manner. Friedman argues that the semantic Web, with English as its *lingua franca*, will homogenize through language and will accentuate the flattening of cultures. The recent discourses about the semantic Web argue for the need to pursue a "universal" imperative, viewing it more than a linguistic concept. Cho and Giustini propound that the semantic Web is actually a way to translate and merge languages, including computer languages, into something universal. This entails promoting standards for databases through the incoporation of artificial intelligence so that similar concepts are not confused within the Internet, enabling users to retrieve information without confusion between similar terms or concepts. While Web 2.0 is seen as a "disjoined space where everything is miscellaneous and governed by the dichotomy of the global-local dynamic" (Cho & Giustini), the semantic Web reinstates the need for a universal logic or language beyond the linguistic notions of language. While Web 2.0 conveyed the increased intractivity, customization, and social aspects of the Internet along with its viability as an advertising medium, the semantic Web emphasises the integration of data where there will be cross-referencing between countless data bases. This will affect the social and economic ways in which people, advertisers, and industries engage with technology and information in a global society.

From an economic perspective, globalization denotes a global economy in which capital markets are interconnected, and where investments and savings in indigenous economies can be affected by the behaviour of global financial markets (Castells, 1998). Undeniably, foreign direct investments and the activities of transnational companies are also seen as important, as well as contentious, components of the globalization process in the last few decades (Zanfei, 2005, p. 9). According to Greg Buckman (2004, p. 35), there are two groups of institutions that are relentless in their promotion and expansion of economic globalization. These include the world's transnational corporations (TNCs), which control most of the investment, trade, and employment decisions of

economic globalization. The other is the public international financial institutions created to oversee the management of economic globalization, which include organizations such as the World Bank and the International Monetary Fund (IMF).

In stressing the role of the MNCs, Bailey et al. (cf. Castells, 1998) point out that while in the early 1990s MNCs employed directly only about 70 million workers, producing one-third of the world's total private output, these, at times, outweighed the total value of world trade. Consequentially, MNCs in manufacturing, services, and finance constitute the core of the world economy, which creates dependence between local firms and global marketplaces, thus impacting the fate of workers and their life outcomes at local levels (Castells, 1998). This highlights the fact that the mode of production has been reconfigured in the information society, where knowledge-based activities have come to play a central role in the production process and the rise in the proportion of the labour force that deals with the production, distribution, and processing of information and knowledge, in contrast to the proportion which handles tangible goods (Wolfe, 2000, p. 2). Additionally, the globalization of technology is also linked to the increasing significance of research and design (R&D) in the new paradigm, while ICTs have been the main catalyst in stimulating new technical and social networks that can enable strategic planning for MNCs (Wolfe, 2000). This can include technology transfer and outsourcing, which can transform the structure of the global economy (Nelson, 2005). The transfer of technology across borders to different cultural environments is one way in which ICTs are involved in cross-cultural interaction (Walsham, 2001). Thus, in the context of rapid development and dissemination of new knowledge, innovation has become a critical element of competitiveness, which has led to global innovation networks enabling sharing and reciprocity of knowledge, as well as the exploitation and the reinstatement of national identity and ownership through patents and copyrights (Fruchterman, 2007; Rycroft, 2002).

The concept of globalization, Wolfe (2000) argues, implies that "individual economies are becoming more transnationalized or integrated into the international economy and this entails losing an important degree of national sovereignty and autonomy" (p.2). Globalization, Wolfe contends, can encompass the "growing integration of markets and production strategies which facilitates the design and production of goods for global, rather than simply national markets as well as the sourcing of components on a global basis." The reduction in transportation and communication costs, along with the digitalization of information, has resulted in the physical distintegration of production, as different components of a final product are now manufactured in several different countries (Fruchterman 2007). The growth, demand, and migration of highly-skilled labour is also indicative of a world economy. Castells (1998) reiterates that this global economy is a relatively new entity in terms of its history as the infrastructure (i.e., telecommunications, information systems, microelectronics-based manufacturing and processes, information-based air transportation, container cargo transport, etc.) required to facilitate it has only been available in the last two decades.

Castells (1998, p. 3) contends that since the 1990s, the entire planet has become organized around telecommunicated networks of computers at the heart of information systems and telecommunication networks, and categorises the availability of ICTs as a prerequisite for economic and social development in our world. Technological advances in microprocessors, fibre optics, memory chips, along with other complementary technologies, have dramatically increased the speed, storage, and processing capacity of computers and telecommunication networks. Kahin and Neeson (cf. Wilson, 1998, p. 8) are of the view that the present 'information revolution' will have a far greater and qualitatively different impact compared to any previous phenomenon. The role of ICTs has also been bound with development studies literature where the use and dissemination of media is seen crucial to both economic and political development (See Lerner 1958). In contrast others have argued that such media technologies are amenable to abuse by governments for political propaganda.

On the other hand, the adoption and proliferation of ICTs are seen as an enabling force not just for business and trade, but also for governments where innovations in information technology and policy, including data warehousing, civic networking, and the Internet, can provide new means to create public agencies that emphasise democratic participation and citizen access to information (Sunarno 2001, p. 64). ICTs have been framed as both enhancing and debilitating democracy. Theorists (See Talero, 1997; Toffler, 1980) have suggested that the information revolution will enable citizens to be connected to one another and to sources of power that will bring transparency to internal governmental processes and help liberate political debates and, in the process, deliver new ways of expressing and aggregating individual and communal expressions.

Globalization from this perspective has also been viewed as "bottom up," where globalization of ICTs can lower communication and coordination costs for NGOs and social movements to distribute their messages, mobilize support, and influence public discourse (Wilson, 1998, p. 34). As pointed out by Benjamin Barber (1996, p. 17), the information revolution unleashes both centrifugal and centripetal forces where homogenization tendencies will be counter-balanced by balkanization in terms of political and social expression. According to Bill Robinson (2004, p.184), the rise of a global justice movement is the clearest example that popular and transitional forces had begun to transnationalise in the 1990s,

moving to create alliances, networks, and organizations that transcend national and even regional borders.

Many nations around the world have recognised the role that ICTs can play in socioeconomic development (Ogunsola, 2005), whether they be developing or developed nations. Manuel Castells (1998) argues that though technology on its own does not resolve social problems, the availability and use of information and communication technologies are prerequisites for economic and social development in our world. For Castells (1998), ICTs can be a double-edged sword for, on the one hand, they have the ability to boost economies when complemented with the right policies of augmenting technical literacy and education in a country. On the other hand, countries that do not adapt to the changing technical landscape can be left behind, thus imposing an uneven spread of economic and social benefits that can then exacerbate the digital divide. Alan Freeman (2004, p. 47) supports the hypothesis that globalization has doubled the inequality between the advanced countries and the rest of the world. Freeman argues (2004) that in 20 years (assuming that the starting point of globalization is 1980), globalization has reasserted and sharpened the division of the world's nations into two fundamentally unequal blocs (p.47). As Buckman (2004, p. 68) points out, since the start of the Industrial Revolution rich countries have benefited more from economic globalization than poorer countries, and in recent decades some countries have become even poorer as a result of it.

FUTURE TRENDS

The processes of change that globalization have unleashed in developed and developing economies will inevitably be contradictory and uneven, warranting closer cooperation between countries and governments. In the process, they will increase the growing interdependence of countries, but also cement existing power structures in which the more economically powerful countries will increase their power to make decisions on many aspects of global governance, including poverty, environmental concerns, disease control, the digital divide, and global migration patterns, among others. The recent rhetoric on globalization has focused on issues of inequality, resistance to globalization, and the new forms of dependence that less developed countries are subjected to in a globalized world. The rise of global civil society organizations, and their use of ICTs to express and harness support locally and globally against international organisations such as the WTO and the IMF, exemplifies the present forms of concern that occupy globalization debates. In addition, since the events of 9/11, increased globalization is perceived to result in increased risks. Such risks transcend "national borders and bring into being supranational and non-class specific global hazards with a new type of social and political dynamism" (Beck 1992, p. 13). Due to an overwhelming recognition in the world community of environmental issues, exploring resources, services, and solutions within local contexts is also being put on the political agenda in different societies, and will continue to be a major concern in the coming years.

CONCLUSION

Globalization as a phenomenon has been analysed from various perspectives. The optimists, pessimists, rejectionists, and the anti-globalisation advocates have discussed issues ranging from the economic, political, social, and technological to the cultural, illuminating the diverse and dialectical processes that embody the term.

The interconnectedness of the world through ICTs has provided nations with opportunities and problems. It has allowed richer nations to profit from the ability to make strategic alliances, and to produce and source globally while creating global markets for their products and services. For less developed nations, it has created a siege mentality and a fear that they may be left far behind if they do not adapt to the changing nature of the world economy. This, in tandem, has produced varying degrees of media literacy, as well as digital and economic divides in various parts of the world. Globalization has also been discussed as a form of cultural imperialism, with the emergence of global media players and the rise in Western cultural imports to the rest of the world. Equally, local responses to global cultural flows and indigenous forms of production and consumption have also been argued as forms of resistance, as have localization and hybridization of global and local products and practices. Globalization and ICTs have also been bound with both political democratization and its opposite, where technologies present opportunities for abuse by governments through centralization of information, but also empowerment of citizens through information dissemination and aggregation of public opinion.

Despite the new forms of engagement and connectivity presented by ICTs, and the emergence of global governance as well as social networks, age-old problems, such as poverty and disease, still plague the globe. In addition, the new connectivities facilitated through communication and transport networks will constitute new forms of risks that will be borderless, global, and instantaneous, such as the threat of terrorism and disease (Beck 1993). Zygmunt Bauman (2006, p. 98) captures the zeitgeist of this postmodern condition by eloquently arguing that in a "planet tightly wrapped in the web of human interdependence, of nothing which the others do or can do we may be sure that it won't affect our prospects, chances and dreams."

REFERENCES

Ajayi, G. O. (2000). Challenges to Nigeria of globalization and the information age. Keynote Address at Workshop on National Information Communication (NIC) Infrastructure Policy, Plans and Strategies for Implementation. National University Commission (NUC) Auditorium Ironsi Street, Maitama, Abuja, March 28-30.

Alecke & Untiedt, 2000

Anderson, B. (1991). *Imagined communities: Reflections on the origin and spread of nationalism*. London: Verso.

Appadurai, A. (1997). *Modernity at large: Cultural dimensions of globalization*. New Delhi, India; Oxford University Press.

Bailey, P. et al. (Eds.). (1993). *Multinationals and employment: The global economy in the 1990s*. Geneva: ILO.

Barber, B. (1996). *Jihad vs. McWorld. How globalism and tribalism are reshaping the world.* USA: Ballantine.

Bauman. Z. (2006). *Liquid fear*. London: Polity Press.

Beck, U. (1992). *Risk society: Towards new modernity*. London: Sage.

Buckman, G. (2004). *Globalization: Tame it or scrap it?* London: Zed Books.

Caincross, F. (1997). *The death of distance: How the communication revolution will change our lives.* Cambridge, MA: Harvard Business School.

Castells, M. (1996). *The rise of network society. The Information Age: Economy, society and culture*, vol. 1. Oxford: Blackwell.

Castells, M. (1998). *Information, technology, globalization and social development.* Paper Presented at the UNRISD Conference on Information, Technological and Social Development, Palais des Nations, Geneva, 22-24 June 1998. Retrieved 27/10/2007, from http://www.unrisd.org/unrisd/website/document.nsf/ab82a6805797760f80256b4f005da1ab/f270e0c066f3de7780256b67005b728c/$FILE/dp114.pdf

Castells, M (2000). End of a millennium. *The Information Age: Economy, society, culture*, vol. 3. Oxford: Blackwell.

Castells, M. (2001). *The Internet galaxy: Reflections on the Internet, business and society*. Oxford: Oxford University Press.

Cho, A. & Giustini, D. *The semantic Web as a large, searchable catalogue: A librarian's perspective.* Retrieved 14/03/2008, from http://www.semantic.com/index2.php?option=com_content&task=view&id=52

Flew, T. (2007). *Understanding global media*. New York: Palgrave Macmillan.

Freeman, A. (2004). The inequality of nations. In A. Freeman & B. Kagarlitsky (Eds.), *The politics of empire*. London: Zed Books.

Friedman, T. (2005). *The world is flat: A brief history of the twenty-first century*. New York: Farrar, Straus and Giroux.

Fruchterman, (2007). *Developing information technology to meet social needs. Innovations World Economic Special Forum*. Retrieved 14/03/2008, from http://benetech.blogspot.com/2007_04_01_archive.html

Ghosh, J. (2004). Imperialist globalisation and the political economy of South Asia. In A. Freeman & B. Kagarlitsky (Eds.), *The politics of empire*. London: Zed Books.

Giddens, A. (1990). *The consequences of modernity*. Cambridge: Polity Press.

Giddens, A. (1999). Globalization: An irresistible force. In *Daily Yomiuri*, 7 June 1999. Retrieved 27/10/2007, from http://globalpolicy.igc.org/globaliz/define/irresfrc.htm

Haleja, S. (2005). *Role of information and communication technologies in managing globalization at the national and regional levels*. Paper prepared for the International Conference on Strengthening Regional Cooperation For Managing Globalization, Moscow, 28-30 September 2005.

Hardt, M. & Negri, A. (2000). *Empire*. Cambridge: Harvard University Press.

Heidegger, M. (1971). *Poetry, language, thought*. New York: Harper and Row.

Kobrin, S. J. *Economic governance in an electronically networked global economy*. The Wharton School, University of Pennsylvania.

Lang, J. C. (2001). Managing in knowledge-based competition. *Journal of Organizational Change Management, 14*, 539-53.

McLuhan, M. (1964). *Understanding media*. New York: Mentor

Morales-Gomez, D., & Melesse M. (1998). Utilizing information and communication technologies for development: The social dimensions. *Information Technology for Development, 8*(1), 3-14.

Mowlana, H. (1997). *Global information and world communications*. London: Sage.

Nelson, D. (2005.) *Outsourcing and the political economy of globalization: A discussion note*. Workshop on The Political Economy of Globalization: How Firms, Workers and Policymakers are Responding to Global Economic Integration, Centre for Globalization and Governance at Princeton University, 28-30 April 2005.

Ogunsola, L. A. (2005). Information and communication technologies and the effects of globalization: Twenty-first century 'digital slavery' for developing countries – Myth or reality? *Electronic Journal of Academic and Special Librarianship, 6*(1-2). Retrieved 27/10/2007, from http://sounthenlibrarianship.icaap.org/content/v06n01/ogunsola_101.htm

Robertson, R. (1992). *Globalization: Social theory and global culture*. Sage: London.

Robinson: B. (2004). The crisis of global capitalism. In A. Freeman & B. Kagarlitsky (Eds.), *The politics of empire*. London: Pluto Press.

Rycroft, R. (2002). Technology-based globalization indicators: The centrality of innovation network data. *The Centre for the Study of Globalization, Occasional Paper Series*, Oct 7 2002.

Scholte, 2000

Stammers, N., & Eschle, C. (2005). Social movements and global activism. In W. de Jong, M. Shaw, & N. Stammers (Eds.), *Global activism global media*. London: Pluto Press.

Sunarno, S. (2001). Globalization and information technology: Forging new partnerships in public administration. *Asian Review of Public Administration, 8*(2), 63–76.

Toffler, A. (1980). *The third wave*. New York: William Morrow and Co.

Tolero, E. (1997). National information infrastructure in developing economies. In B. Kahin & E. J. Wilson III (Eds.), *National information infrastructure initiatives: Visions and policy design*. Cambridge, MA: MIT Press.

Walsham, G. (2001). Globalization and ICTs: Working across cultures. *Research Papers in Management Studies*, working paper 8.

Wilson. E. J. (1998). *Globalization, information technology and conflict in the second and third worlds: A critical review of the literature*. New York: Rockerfeller Brothers Fund, Inc. Retrieved: 13/10/2007, from http://www.rbf.org/usr_doc/Globalization,_Information_Technology,_and_Conflict.pdf

Wolfe, D. A. (2000). Globalization, information and communication technologies and local and regional systems of innovation. In K. Rubenson & H. Schuetze (Eds.), *Transition to the knowledge society: Conference proceedings*. Vancouver: University of British Columbia Press. Retrieved 07/10/07, from http://www.utoronto.ca/progris/pdf_files/Ictreginnov.pdf

Zanfei, A. (2005). Globalization at bay? Multinational growth and technology spillover. *Critical Perspectives On International Business, 1*(1), 7-19.

KEY TERMS

Digital Divide: The separation between the information rich, or haves, and information poor, or have-nots.

Globalization: The integration of the world through a complex set of social, economic, political, technical and cultural processeses

ICTs: Information and communication technologies, seen as the driving force of globalization and new forms of connectivity.

Information Society: The transition from the modern and industrial age in which modes of production, exchange, and social capital are increasingly defined through information

MNCs or Multinational Corporations: In the age of globalization, MNCs have an increasing role in global economies, flow of capital, and in the formation of strategic economic and global alliances.

Network Society: The rise of information society will see the emergence of a network society in which information and technology will enable the formation of networks and strategic planning.

Postindustrial Society: The transition from the industrial society will lead to a postindustrial society characterised by increasing emphasis on global connectivity and capitalisation of information.

Technology Leapfrogging for Developing Countries

Michelle W. L. Fong
Victoria University, Australia

INTRODUCTION

Information and communication technologies (ICTs) have been acknowledged, in research works on developed and industrialized countries, for their potential in opening up development opportunities. At firm level, ICTs can facilitate communications and coordination of processes within a firm or between firms in a supply chain such as through e-collaboration (Fong, 2005; Hammant, 1995; Jin, 2006; Porter, 2001; Porter & Millar, 1985). These technologies can also improve management decision-making process through better and faster marshalling of information. Gains from these applications may be in the form of scale economies, cost-savings, increased productivity, and improved competitiveness (Bourlakis & Bourlakis, 2006; Farrell, 2003; Hammant, 1995; Howgego, 2002; Pilat, 2004; Porter & Millar, 1985). At industry level, ICTs can improve the functioning or governance of market (James, 2000; Malone, Yates, & Benjamin, 1989; Matsuda, 1994; OECD, 2005). From the social-economic perspective, ICTs can improve quality of life of communities, provide greater access to health and education services, and create economic opportunities for any underprivileged population groups (Mercer, 2001; Oberski, 2004; Reisman, Roger, & Edge, 2001; The World Bank, 2001; UNDP, 2001a; United Nations, 2006). All these improvements in efficiency and access are likely to be aggregated at the national level in the form of economic growth or sustainability, and welfare gains (Madden & Savage, 2000; OECD, 2005).

Developing countries are generally latecomers to the ICT revolution, but if they can emulate industrialized countries in their adoption of ICTs, they will be afforded the same technological opportunities. Successful exploitation of such opportunities by developing countries can significantly narrow the economic gap between them and developed countries as they catch up in economic development.

In ICT's advancement trajectory, the opportunities offered by a newly emerged ICT tend to be superior to those of prior versions of technology. If a developing country leapfrogged to a newly emerged ICT, it would then be exposed to unprecedented potential in alleviating poverty and securing economic growth, as well as the possibility of surpassing developed and industrialized countries in economic development. Thus, technology leapfrogging is an attractive notion to developing countries, but is it a realistic goal?

BACKGROUND

"Technology leapfrogging" refers to the adoption of advanced or state-of-the-art technology in an application area where immediate prior technology has not been adopted. Discussions of ICT leapfrogging have largely focused on developing countries, which generally lag behind on technology adoption, and unlike the developed countries, are not inhibited by entrenched intermediate technology. New and advanced technology provides developing countries with the opportunity to accelerate economic development (Hanna, Guy, & Arnold, 1995; Prayag, 2001; OECD, 2005; UNDP, 2001b). In addition, the advancement of ICTs has reduced costs and imposed lesser demands on the skill of the users due to user-friendly features (Ensley, 2005). The possibility of achieving significant economic growth through advanced and less costly technology thus seems exceptionally attractive to developing countries. It has also been suggested that developing countries do not have any alternative in technology adoption, except to leapfrog to new and advanced technologies (Choucri, 1998; Mansell & Wehn, 1998; Davison, Vogel, Harris, & Jones, 2000).

The concept of technology leapfrogging for ICTs first emerged in the early 1990s (Antonelli, 1991; Lamberton, 1994; Mody & Dahlman, 1992). However, research works in this area were still limited, hindered by a lack of clear empirical data from developing countries, measurement difficulties, the relatively short-time span of newly emerged and advanced ICTs, and the long time span involved in gathering reliable data to understand technology leapfrogging in developing countries (Alzouma, 2005; Ausubel, 1991; Prakash 2005; Sharif, 1989; UNESCO, 1996).

As a result, technology leapfrogging remains a controversial concept. For example, in the case of Africa, Chisenga (2000) believes that the implementation of this concept would facilitate global integration of businesses, and provide a better learning environment for African children, all to the benefit of the economy. Ochieng (2000), on the other hand, believes that investments in ICT compete with the provision of basic necessities for the poor. One of the general arguments against leapfrogging has been that it might turn out to be an expensive trajectory in the short run for developing countries, which tend to bear a high burden of debts (Chen, Farinelli, & Johansson, 2004). In addition, investment in a

new technology is likely to involve a long payback period for developing countries because of the nascent conditions of their market. It has been further argued that the worst possible outcome from this exercise would be when the new investment was displaced by a major breakthrough before these developing countries geared up in their capabilities to sufficiently harness their technology investment. Some commentators believe that the developmental effects of ICT applications have been greatly exaggerated and caution that advanced technology makes little difference and can even inflict harmful effects, such as creating a digital divide, in the lives of people in developing countries (Mansell, 1999; Sussman, 1997; United Nations Commission on Science for Technology and Development, 1997; Van Dijk, 1999; Wang, 1991).

Kojima (2003, p. 1), however, highlighted that the concept of technology leapfrogging requires selective and discerning application. She noted that technology leapfrogging is applicable to certain technologies such as ICTs but not to emission control technology, a field where developing countries should wait for developed countries to test out the emerging technology before adopting it.

LEAPFROG TO WIRELESS ICTs

It has been claimed that leapfrog technologies are largely those that do not rely on tangible grid. These include mobile phones, satellite communications, and decentralized power sources like solar power (Article 13, 2005). Wireless ICTs, such as mobile phones, have emerged as a leading leapfrog technology (BBC News, 2002; Nkwae, 2002; Cascio, 2004). Mobile phone communication technology has even substituted the traditional fixed networks in developing countries such as China and many African countries.

Leapfrog wireless communications technologies (ICTs) have often generated significant benefits for communities. For example, villages in Robib, Cambodia, were reported to have leapfrogged from an agricultural to an information economy through wireless network (oneworld radio, 2006). The villagers were able to access medical and health services, and a global marketplace for their cottage industry through wireless communications technology. In Africa, information available through mobile phones enabled farmers in Senegal to double the prices of their crops and herders in Angola to locate their cattle through GPS (global positioning system) technology (oneworld radio, 2006). In another study by Williams on "The relationship between mobile telecommunications infrastructure and FDI in Africa" (as cited in Vodafone, 2005), at least 50% of small businesses surveyed in South Africa and Egypt attributed profit increases to mobile phone usage. The same study also found that more than 75% of respondents in Tanzania and South Africa experienced improvement in contact and relationships with close ones because of mobile phone communications technology (Vodafone, 2005). In all these cases, however, the actual benefits from leapfrogging were yet to be ascertained at the aggregate level.

Technology by itself does not solve problems, but the availability and use of ICTs are a prerequisite for economic and social development in developing countries. In other words, ICTs are not a standalone solution to development problems. Developing countries characteristically lack many of the conditions needed to harness and sustain leapfrog-type development offered by new and advanced technology. Because the conditions in developing countries are usually weak or inadequate, cultivating, building, and deepening these conditions to support the mastery, applications, diffusion, and innovation of leapfrog technologies are unlikely to prove easy or straightforward. It is therefore anticipated that developing countries would not be able to fully utilize or exploit the potential of advanced technology at the early stage of leapfrogging due to their limited infrastructure. Despite this, there is no point for developing countries to go through the fiber optic building process for their telecommunication infrastructure when they could start out with wireless telecommunications. Leapfrogging to wireless communications technology is a valid strategy based on the promising technology's potential for economic advancement, and the lower costs and resources involved in setting up a telecommunication infrastructure. It has been estimated that wireless technologies cost about 20% of traditional wired installations (United Nations General Assembly and Economic & Social Council, 2000). Developing countries should develop and strengthen their infrastructure within the framework of advanced technology rather than time- and resources-consuming intermediate technology that is likely to prolong their subservient position in development. In capabilities development and learning processes for example, developing countries should be exposed to and educated in the skills required by higher-level technologies.

However, technology leapfrogging may pose very high risks to developing countries if immature or unproven technologies are involved. Despite this, the risks can be managed and minimized by careful planning and evaluation, to capitalize on the opportunities of leapfrog technology. Otherwise, they are likely to end up as expensive failures. In the planning and implementation of any technology leapfrogging process, developing countries need to take into consideration a number of factors (shown in Figure 1) which are likely to impact on the time span involved in the mastery, applications, diffusion and innovation of the technologies involved.

Figure 1 includes certain conditions such as literacy and education that cannot be immediately emulated or leapfrogged despite their undoubted importance in reaping the greatest advantages from advanced ICTs. The nurturing of human capabilities requires substantial investment in

Table 1. Factors relevant to conditions for technology leapfrogging

Factors	Example of issues	Source
Market condition: • Market demand • Market competition	• Market competition for rational pricing of ICT & access, • Development of locally relevant content and languages to promote advanced technology uptake, • Foreign participation through investment to break down monopoly structure.	Adzadi, 2001; Alzouma, 2005; Choucri, 1998; Davison et al., 2000; Ensley, 2005; Garcia & Gorenflo, 1999; Grace et al., 2001; Haddad & MacLeod, 1999; International Telecommunication Union, 2004; Mansell, 1999; Mbambo, 1996; Nkwae, B. (2002); OECD, 2005, 2006; Praskash, 2005; Pringle & David, 2002; Raji et al., 2006; Sehrt, 2003; Sinha, 2005; UNDP, 2001a, 2001b; United Nations General Assembly and Economic & Social Council, 2000; Vodafone, 2005; Wijkman & Afifi, 2002.
Institutional capacity	• Support for intellectual capital development, • Development of stable learning and attractive investment environment, • Ensuring security and stability in the environment, • Establishment of an enabling regulatory and legislative framework, • Economic, social, and political stability.	
Social	• Ensuring equity in digital access, • Narrowing or erasure of digital divide.	
Human capabilities	• Improvement of literacy and computer literacy levels, • Nurturing of requisite skills and expertise. • Continuous investment.	
Government	• Definitive guiding policies, • Strategic deployment of ICT, • Coordination and linkages among actors in the system.	
Stakeholders	• Interaction and strategic links among actors in the system, • Regional and international cooperation and collaboration, • Identification of e-champions and e-leaders to spearhead technology leapfrogging projects.	
Utility infrastructure	• Electricity, transportation networks, etc.	

education and skills transfer, and also government intervention to expedite the development and accumulation of human capabilities. Governments, besides being providers of national education, can use incentives to encourage private initiatives in the provision of training resources.

In regard to market conditions, telecommunication markets that are operated by a monopoly which charges high access fees represent a serious obstacle to advanced technology uptake and investment. The ITU (2004) observed that mobile markets of competitive structure have significantly higher rates of mobile penetration than monopoly markets, even where per-capita incomes are the same. Competition is important for making access cost affordable to users, and for developing products and services that meet users' requirements. This in turn would help to generate substantive market demand and confidence in future technology investment.

In addition to achieving competitive market conditions, strong legislative frameworks must be introduced to provide environmental stability and security, such that businesses are confident to invest and consumers are confident to uptake new technology and trial its sophisticated potentials. Governments in developing countries are likely to face the challenge of designing an appropriate regulatory environment to support and enable effective operation of a sophisticated telecommunication infrastructure. However, an appropriate regulatory environment could be achieved through goodwill assistance from experienced international aid agencies and developed countries. Such assistance should not be limited to regulatory or legislative environment; developing countries in fact require partnerships and long-term support at the local, national and international levels for their capabilities and other institutional capacity building, research and development, and innovation of leapfrog technologies.

It must also be noted that a new and advanced technology, which is capable of offering development opportunities to developing countries, may also be capable of creating a digital divide within these economies. Therefore, advanced technology adoption and diffusion must be user-focused rather than technology-focused, as social issues can turn out to be a formidable barrier to technology leapfrogging. To prevent the emergence of digital divide, government-donor-community-enterprise partnerships may be initiated to support ongoing projects of expanding wireless communication access to rural areas. To advance the goals of a country's leapfrogging strategy, commitments from a strong network of governmental and nongovernment participants including e-leaders and e-champions, are vital for its success.

FUTURE TRENDS

In communications technology adoption, wireless and mobile technologies offer a relatively quicker and less costly way to leapfrog the more expensive and time-consuming task of building fixed-line telephone networks. In addition, these technologies are built on an easily deployable infrastructure. As a result, wireless and mobile communication technologies will become a dominant medium in developing and transitional countries in the coming years.

The wireless mobile phone industry is in the early phase of deploying its third generation (3G) technologies, which are capable of sophisticated communication features. These technologies have the potential to integrate data communications and computing capabilities into handsets such as mobile Internet. The full deployment of 3G technologies will change industry value networks through their wide range of nonvoice services.

CONCLUSION

Technology leapfrogging is likely to be difficult and challenging for developing countries. Rather than start out with traditional fixed-line solutions, developing countries should leapfrog to wireless technologies which offer a relatively quicker and less costly way of building a telecommunication infrastructure. However, economic development benefits may not quickly accrue to developing countries in their leapfrogging efforts. These countries are likely to take a considerable period of time in building their capabilities to absorb, master, use, and innovate leapfrog technologies. Nevertheless, the process may be expedited by the deliberate policies and guidance of governments in these countries, with support from the international arena. At the same time, governments in these countries must constantly reassess the impact of policies and align them with the social objectives so as to remain user-focused in the need to harness the technology quickly.

REFERENCES

Adzadi, G. K. (2001). *Critical analysis of the impact of socio-economic and political factors on information technology disparity between developed and developed countries*. Retrieved December 12, 2007, from http://cyber.law.harvard.edu/archive/oe/msg00044.html

Alzouma, G. (2005). Myths of digital technology in Africa: Leapfrogging development? *Global Media and Communication, 1*(3), 339-356.

Antonelli, C. (1991). *The diffusion of advanced telecommunications in developing countries*. Paris: OECD.

Article 13. (2005, September). *Leapfrogging: A different route to development*. Retrieved December 12, 2007, from http://www.article13.com/A13_PrintablePages.asp?strAction=GetPublication&PNID=1192

Ausubel, J. H. (1991). Rate-race dynamics and crazy companies: The diffusion of technologies and social behaviour. *Technological Forecasting and Social Change, 39*(1-2), 11-22.

BBC News. (2002). *Mobiles to leapfrog into the future*. Retrieved December 12, 2007, from http://news.bbc.co.uk/2/hi/technology/2287913.stm

Bourlakis, M., & Bourlakis, C. (2006). Integrating logistics and information technology strategies for sustainable competitive advantage. *Journal of Enterprise Information Management, 19*(4), 389-402.

Cascio, J. (2004, December 15). *Leapfrog 101*. Retrieved December 12, 2007, from http://www.worldchagning.com/archives/001743.html

Chen, Y., Farinelli, U., & Johansson, T. B. (2004, June). Technological leapfrogging–a strategic pathway to modernization of the Chinese iron and steel industry. *Energy for Sustainable Development, VIII*(2), 18-26.

Chisenga, J. (2000). Global information and libraries in Subsaharan Africa. *Library Management, 21*(4), 178-187.

Choucri, N. (1998). Knowledge networking for technology leapfrogging. *Cooperation South Journal, 2*, 40-52. Retrieved December 12, 2007, from http://tcdc.undp.org/CoopSouth/1998_2/cop9827.pdf

Davison, R. M., Vogel, D. R., Harris, R. W., & Jones, N. (2000). Technology leapfrogging in developing countries–an

inevitable luxury? *The Electronic Journal on Information Systems in Developing Countries, 1*(5), 1-10.

Ensley, L. (2005). *Information and communications technological leapfrogging in developing countries of the world*. Retrieved December 12, 2007, from http://www.ischool.utexas.edu/~i385q/archive/ensley_l-Technological%20Leapfrogging.doc

Farrell, D. (2003, October). The real new economy. *Harvard Business Review, 81*(10), 104-112.

Fong, M. (Ed.). (2005). *E-collaborations and virtual organizations*. Hershey, PA: Idea Group.

Garcia, D. L., & Gorenflo, N. R. (1999, April). *Rural networking cooperatives: Lessons for international development and aid strategies.* Retrieved December 12, 2007, from http://www.fao.org/WAICENT/FAOINFO/SUSTDEV/CDdirect/CDre0033.htm

Grace, J., Kenny, C., Qiang, C., Liu, J., & Reynolds, T. (2001, February 14). *ICTs and broad-based development; A partial review of the evidence*. Retrieved December 12, 2007, from http://www.cag.lcs.mit.edu/ict4dev/papers/grace02.pdf

Haddad, H., & MacLeod, S. (1999). Access to medical and health information in the developing world: An essential tool for change in medical education. *Canadian Medical Association Journal, 160(*1), 63-5.

Hammant, J. (1995). Information technology trends in logistics. *Logistics Information Management, 8*(6), 32-37.

Hanna, N., Guy, K., & Arnold, E. (1995). *The diffusion of information technology: Experience of industrial countries and lessons for developing countries*. Washington, DC: The World Bank.

Howgego, C. (2002). Maximising competitiveness through the supply chain. *International Journal of Retail and Distribution Management, 30*(12), 603-605.

International Telecommunication Union. (2004). *Africa: The world's fastest growing mobile market. Does mobile technology hold the key to widening access to ICTs in Africa*? Retrieved December 12, 2007, from http://www.itu.int/newsarchive/press_releases/2004/04.html

James, G. (2000). Empowering bureaucrats. *MC Technology Marketing Intelligence, 20*(12), 62.

Jin, B. (2006). Performance implications of information technology implementation in an apparel supply chain. *Supply Chain Management: An International Journal, 11*(4), 309-316.

Kojima, M. (2003, February). *Public policy for the private sector: Leapfrogging technology*. The World Bank Group: Private Sector and Infrastructure Network, Note Number 254. Retrieved December 12, 2007, from http://lnweb18.worldbank.org/SAR/sa.nsf/Attachments/Leapfrog/$File/Leapfrogging+technology.pdf

Lamberton, D. (1994, November). The information revolution in the Asian-Pacific region. *Asian-Pacific Economic Literature, 8*(2), 31-57.

Madden, G., & Savage, S. J. (2000, July). Telecommunications and economic growth. *International Journal of Social Economics, 27*(7-10), 893-906.

Malone, T. W., Yates, J., & Benjamin, R. I. (1989, May-June). The logic of electronic markets. *Harvard Business Review, 67*(3), 166-170.

Mansell, R. (1999). Information and communication technologies for development: Assessing the potential and risks. *Telecommunications Policy, 23*, 35-50.

Mansell, R., & Wehn, U. (1998). *Knowledge societies: Information technology for development*. Oxford University Press.

Matsuda, T. (1994). The use of information technology to achieve accurate pricing in agricultural commodity markets in Japan. *Information Technology and People, 7*(3), 37-49.

Mbambo, B. (1996). Virtual libraries in Africa: A knight in shining armour? *IFLA Journal, 22*(3), 229-32.

Mercer, K. (2001). Examining the impact of health information networks on health system integration in Canada. *Leadership in Health Services, 14*(3), 1-30.

Mody, A., & Dahlman, C. (1992). Performance and potential of information technology: An international perspective. *World Development, 20*(12), 1703-1719.

Nkwae, B. (2002). *Information and communications technologies: Can Africa leapfrog the digital divide*? Retrieved from December 12, 2007, from http://www.svt.ntnu.no/geo/Prodec/ict2002/Pdf/Nkwae.pdf

Oberski, I. (2004). University continuing education: The role of communications and information technology. *Journal of European Industrial Training, 28(*5), 414-428.

Ochieng, R. O. (2000). Global information flows. *Library Management Journal, 21*(4), 215-216.

oneworld radio. (2006, December 18). *How radio, cell phones, wireless Web are empowering developing nations*. Retrieved December 12, 2007, from http://radio.oneworld.net/article/view/78640/1/

Organisation for Economic Cooperation and Development. (2005). *Good practice paper on ICTs economic growth and*

poverty reduction. Retrieved December 12, 2007, from http://www.oecd.org/dataoecd/2/46/35284979.pdf

Organisation for Economic Cooperation and Development. (2006). *Information technology outlook highlights*. Retrieved December 12, 2007, from http://www.oecd.org/dataoecd/27/59/37487604.pdf

Pilat, D. (2004). The ICT productivity paradox: Insights from micro data. *OECD Economic Studies, 2004/1*(38), 37-65.

Porter, M. E. (2001, March). Strategy and the Internet. *Harvard Business Review, 79*(3), 62-78.

Porter, M. E., & Millar, V. E. (1985, July-August). How information gives you competitive advantage. *Harvard Business Review, 63*(4), 149-160.

Prakash, G. (2005, September). Leapfrogging into the knowledge era: Use of ICT for development. *IIMB Management Review*, 47-56.

Prayag, A. (2001, September 19). *Leapfrog and catch up, says ILO*. Retrieved December 12, 2007, from http://www.hinduonnet.com/businessline/2001/09/20/stories/01202001.htm

Pringle, I., & David, M. J. R. (2002). Rural communities ICT applications: The Kothmale model. *The Electronic Journal on Information Systems in Developing Countries, 8*(4), 1-14.

Raji, M. O., Ayoade, O. B., & Usoro, A. (2006). The prospects and problems of adopting ICT for poverty eradication in Nigeria. *The Electronic Journal of Information Systems in Developing Countries, 28*(8), 1-9.

Reisman, S., Roger, G., & Edge, D. (2001). Evolution of Web-based distance learning strategies. *International Journal of Educational Management, 15*(5), 245-251.

Sharif, M. N. (1989). Technological leapfrogging: Implications for developing countries. *Technological Forecasting and Social Change, 36*(1-2), 201-208.

Sehrt, M. (2003). *E-learning in the developing countries: Digital divide into digital opportunities.* Retrieved December 12, 2007, from http://www.un.org/Pubs/chronicle/2003/issue4/0403p45.asp

Sinha, C. (2005, October 31). Effect of mobile telephony on empowering rural communities in developing countries. In *Proceedings of the International Research Foundation for Development (IRFD), Conference on Digital Divide, Global Divide, Global Development and the Information Society*, November 14-16.

Sussman, G. (1997). *Communication, technology, and politics in the information age*. London: Sage.

The World Bank (2001). *World development report*. Oxford University Press.

United Nations. (2006, November 3). *Special Panel on e-government for participation and inclusion.* Retrieved April 11, 2008, from http://panl.un.org/intrade/groups/public/documents/UN/unpan024444.pdf'

United Nations Commission on Science and Technology Development. (1997, May 12). *Report of the working group on information and communication technologies for development*. Prepared for the 3rd session (Item No. E/CN.16/1997.4). Geneva, March 7.

United Nations Development Programme. (2001a). *Human development report*. New York: Oxford University Press.

United Nations Development Programme. (2001b). *Role of UNDP in information and communication technology for development*. Retrieved December 12, 2007, from http://www.undp.org/execbrd/pdf/DP2001CRP8.PDF

United Nations Educational, Scientific and Cultural Organization. (1996). *Information and communications technologies in development: A UNESCO perspective: Submission to the UNCSTD working group on information technology for development and the ITU development study group 1*. Retrieved December 12, 2007, from http://www.itu.int/acc/rtc/unesco.htm

United Nations General Assembly and Economic & Social Council. (2000, May 22). *Report of the high-level panel of experts on information and communication technology.* Retrieved December 12, 2007, from http://www.un.org/documents/ecosoc/docs/2000/e2000-55.pdf

Van Dijk, J. (1999). *The network society*. London: Sage.

Vodafone. (2005). *Many factors affect the spread of mobile phones (Key findings of SIM research)***.** Retrieved December 12, 2007, from http://www.vodafone.com/article/0,8118,CATEGORY_ID%253D3040302%2526LANGUAGE_ID%253D0%2526CONTENT_ID%253D266250,00.html

Wang, I. K. (1991). Indigenous and new technologies for technology development. *Asian Economies, 79*, 5-19.

Wijkman, A., & Afifi, M. (2002). *Technology leapfrogging and the digital divide*. Retrieved December 12, 2007, from http://www.cs.berkeley.edu/~mattkam/tased/wijkman2002.pdf

KEY TERMS

Digital Divide: Refers to the disparity between two or more groups of people in their access to digital technology. Digital divide can occur at national level (between different groups within the economy) and/or global level (between different countries or regions).

E-Collaboration: A process by which internal and external individuals and/or groups work together on a practical endeavour through integrated electronic networks enabled by ICTs or coordination technologies.

GPS: A global positioning system that uses satellites, computers, and receivers. It can be used for navigation and tracking purposes, based on computer calculation of time difference between signals emitted from satellites and received by receivers.

ICT: Encompasses all the technology that facilitates the processing, transfer and exchange of information and communication services.

Productivity: In economic terms, is the value of output produced using one unit of input. Workplace productivity generally means output per worker. For a considerable period of time, economists failed to determine the relationship between investments in ICT and productivity. This phenomenon was known as "productivity paradox". Although evidence of this relationship has emerged from studies at the firm level in recent times, measures of it at the aggregate or industry level remain nebulous (Pilat, 2005).

Category: IT Education

Technology-Enhanced Progressive Inquiry in Higher Education

Hanni Muukkonen
University of Helsinki, Finland

Minna Lakkala
University of Helsinki, Finland

Kai Hakkarainen
University of Helsinki, Finland

INTRODUCTION

In higher education, students are often asked to demonstrate critical thinking, academic literacy (Geisler, 1994), expert-like use of knowledge, and creation of knowledge artifacts without ever having been guided or scaffolded in learning the relevant skills. Too frequently, universities teach the content, and it is assumed that the metaskills of taking part in expert-like activities are somehow acquired along the way. Several researchers have proposed that in order to facilitate higher-level processes of inquiry in education, cultures of education and schooling should more closely correspond to cultures of scientific inquiry (e.g., Carey & Smith, 1995; Perkins, Crismond, Simmons & Under, 1995). Points of correspondence include contributing to collaborative processes of asking questions, producing theories and explanations, and using information sources critically to deepen one's own conceptual understanding. In this way, students can adopt scientific ways of thinking and practices of producing new knowledge, not just exploit and assimilate given knowledge.

BACKGROUND

The best practices in the computer-supported collaborative learning (CSCL) paradigm have several features in common: consideration in an interrelated manner of the development of technological applications, use of timely pedagogical models, and attention to the social and cognitive aspects of learning. Emphasis is placed on creating a collaborative community that shares goals, tools, and practices for taking part in an inquiry process.

Synthesizing these demands, Hakkarainen and his colleagues at the University of Helsinki have developed a model of *progressive inquiry* as a pedagogical and epistemological framework. It is designed to facilitate expert-like working with knowledge in the context of computer-supported collaborative learning. It is primarily based on Scardamalia and Bereiter's (1994) theory of knowledge building, on the interrogative model of scientific inquiry (Hintikka, 1999; Hakkarainen & Sintonen, 2002), and on the idea of distributed expertise in a community of learners (Brown & Campione, 1994). The model has also been implemented and studied in various educational settings from elementary to higher education (see, e.g., Hakkarainen, Järvelä, Lipponen, & Lehtinen, 1998; Lipponen, 2000; Veermans & Järvelä, 2004; Muukkonen, Lakkala, & Hakkarainen, 2005; Lakkala, Lallimo, & Hakkarainen, 2005; Lakkala, Ilomäki, & Palonen, 2007).

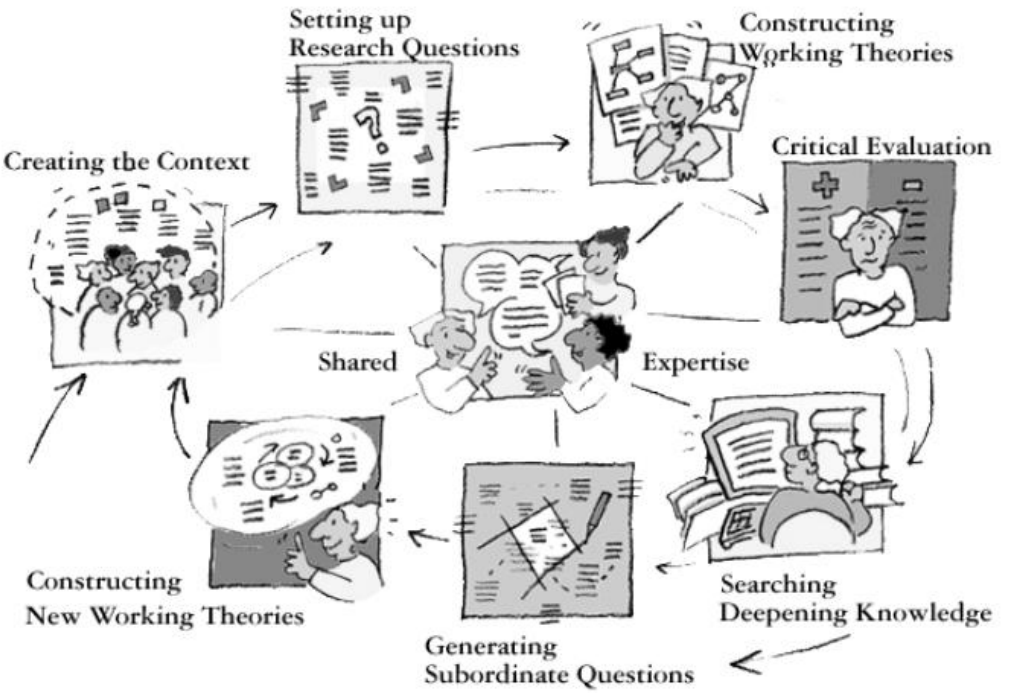
The Progressive Inquiry Model

In progressive inquiry, students' own, genuine questions and their previous knowledge of the phenomena in question are a starting point for the process, and attention is drawn to the main concepts and deep principles of the domain. From a cognitive point of view, inquiry can be characterized as a question-driven process of understanding; without research questions, there cannot be a genuine process of inquiry, although in education, information is frequently conveyed or compiled without any guiding questions. The aim is to explain the phenomena in a deepening question-explanation process, in which students and teachers share their expertise and build new knowledge collaboratively with the support of information sources and technology.

The progressive inquiry model specifies certain epistemologically essential processes that a learning community needs to go through, although the relative importance of these elements, their order, and actual contents may involve a great deal of variation from one setting to another. As depicted in Figure 1, the following elements have been placed in a cyclic, but not step-wise succession to describe the progressive inquiry process (Hakkarainen, 2003; Muukkonen, Hakkarainen, & Lakkala, 1999, 2004):

a. *Distributed expertise* is a central concept in the model. Progressive inquiry intends to engage the community in a shared process of knowledge advancement, and to convey, simultaneously, the cognitive goals for collaboration. Diversity in expertise among participants, and interaction with expert cultures promotes knowledge advancement (Brown et al., 1993; Dunbar, 1995). Acting as a member in the community includes sharing cognitive responsibility for the success of its inquiry. This responsibility essentially involves not only completing tasks or delivering productions on time, but also learners taking responsibility for discovering what needs to be known, goal setting, planning, and monitoring the inquiry process (Scardamalia, 2002). There should be development of students' (and experts') social metacognition (Salomon & Perkins, 1998): students learning to understand the cognitive value of social collaboration and gaining the capacity to utilize socially distributed cognitive resources.
b. The process begins by *creating the context* to anchor the inquiry to central conceptual principles of the domain or complex real-world problems. The learning community is established by joint planning and setting up common goals. It is important to create a social culture that supports collaborative sharing of knowledge and ideas that are in the process of being formulated and improved.
c. An essential element of progressive inquiry is *setting up research questions* generated by students themselves to direct the inquiry. Explanation-seeking questions (Why? How? What?) are especially valuable. The learning community should be encouraged to focus on questions that are knowledge driven and based on results of students' own cognitive efforts and the need to understand (Bereiter, 2002; Scardamalia & Bereiter, 1994). It is crucial that students come to treat studying as a problem-solving process that includes addressing problems in understanding the theoretical constructs, methods, and practices of scientific culture.
d. It is also important that students explain phenomena under study with their own existing background knowledge by *constructing working theories* before using information sources. This serves a number of goals: First is to make visible the prior (intuitive) conceptions of the issues at hand. Second, in trying to explain to others, students effectively test the coherence of their own understanding, and make the gaps and contradictions in their own knowledge more apparent (e.g., Hatano & Inakagi, 1992; Perkins et al., 1995). Third, it serves to create a culture in which knowledge is treated as essentially evolving objects and artifacts (Bereiter, 2002). Thoughts and ideas presented are not final and unchangeable, but rather utterances in an ongoing discourse (Wells, 1999).
e. *Critical evaluation* addresses the need to assess strengths and weaknesses of theories and explanations that are produced, in order to direct and regulate the community's joint cognitive efforts. In part, it focuses on the inquiry process itself, placing the process as the center of evaluation and not only the end result. Rather than focusing on individual students' productions, it is more fruitful to evaluate the community's productions and efforts, and give the student participants a main role in this evaluation process. Critical evaluation is a way of helping the community to rise above its earlier achievements, creating a higher-level synthesis of the results of inquiry processes.
f. Students are also guided to engage in *searching deepening knowledge* in order to find answers to their questions. Looking for and working with explanatory scientific knowledge is necessary for deepening one's understanding (Chi, Bassok, Lewis, Reiman, & Glaser, 1989). A comparison between intuitive working theories produced and well-established scientific theories tends to show the weaknesses and limitations of the community's conceptions (Scardamalia & Bereiter, 1994). The teacher of a course must decide how many of the materials should be offered to the students and how much they should actually search out for themselves. Questions stemming from true wonderment on the part of the students can easily extend the scope of materials beyond what a teacher can foresee or provide suggestions for. Furthermore, searching for relevant materials provides an excellent opportunity for self-directed inquiry and hands-on practice in struggling to grasp the differences between various concepts and theories.
g. *Generating subordinate questions* is part of the process of advancing inquiry; learners transform the initial big and unspecified questions into subordinate and more specific questions, based on their evaluation of produced new knowledge. This transformation helps to refocus the inquiry (Hakkarainen & Sintonen, 2002; Hintikka 1999). Directing students to return to previously stated problems, to make more subordinate questions and answer them, are ways to scaffold the inquiry.
h. *Developing new working theories* arises out of the fresh questions and scientific knowledge that the participants attain. The process includes publication of the summaries and conclusions of the community's inquiry. If all productions to the shared database in a collaborative environment have been meaningfully organized, participants should have an easy access to prior productions and theories, making the development of conceptions and artifacts a visible process.

Figure 1. Elements of progressive inquiry (reprinted by permission, Muukkonen et al., 2004)



Cases of Progressive Inquiry in Higher Education

Progressive Inquiry in a Cognitive Psychology Course

In a study reported by Muukkonen, Lakkala, and Hakkarainen (2001), the progressive inquiry model was implemented in a cognitive psychology course with the use of the Future Learning Environment (FLE). The FLE is an open-source collaborative tool that has the progressive inquiry model embedded in its design and functionality (http://fle3.uiah.fi/; Muukkonen et al., 1999). All the students in the course were guided, during the first two lectures, to formulate research problems. In the beginning, they individually produced these formulations. They continued by discussing their research problems with a peer and, finally, within a small group, selected the most interesting questions to pursue. These questions were then presented to all the participants in the lecture. After this initial problem setting, the technology-mediated groups (three groups of 4-7 volunteers) were instructed to continue their inquiry processes between the weekly lectures in the FLE. The tutor-facilitators took part in the FLE, whereas the teacher conducted the weekly lectures without participating in the database discourse. The rest of the students also formed groups based on their questions, but continued their inquiry process by writing learning logs and commenting on the logs produced by other members of their group without collaborative technology.

A comparative analysis of the knowledge produced by the students in the two conditions provided evidence that the technology-mediated groups were more engaged in problem-setting and redefining practices. Further, they reflected on the process they had undertaken, in respect of the collaboration and their individual efforts. In the productions of the groups who had not used collaboration tools, the social and communal aspects of inquiry and knowledge building were not evident at all in their learning logs, although they were engaged in collaboration during the lectures. The type of comments they provided to two of the learning logs written by other members of their group were very general, and they concentrated mainly on evaluating the level of writing, not on advancement of ideas. However, many of the learning logs were conceptually well developed and integrated. Discourse interaction within the FLE was different in respect of

the participants sometimes engaging in extensive dialogues with ideas presented by the fellow students.

Progressive Inquiry in a Design Course

Two studies carried out by Seitamaa-Hakkarainen and her colleagues (Seitamaa-Hakkarainen, Lahti, Muukkonen, & Hakkarainen, 2000; Seitamaa-Hakkarainen, Raunio, Raami, Muukkonen, & Hakkarainen, 2001) analyzed a collaborative design process as it occurred in a complex and authentic design task of designing clothing for premature babies. The framework of the studies was based on evidence from cognitive research on expertise, which indicated that novices in design tend to generate problem solutions without engaging in extensive problem structuring; experts, by contrast, focus on structuring and restructuring the problem space before proposing solutions (Glaser & Chi, 1988). The studies described in this case were designed to examine whether an expert-like engagement in design process would be supported in the FLE-environment. Features of the environment were used to encourage the users to engage in expert-like designing, and to enable graphic presentation of the knowledge artifacts in the form of importing students' drafts and prototypes into the collaborative environment and developing multiple versions of the designs.

During the collaborative design course, the students were first guided to find out information about the constraints of their design task, such as the size of the babies, special needs for the usability of the clothing, and about the materials. Then they were asked to produce their own sketches and work in small groups to share design ideas and develop their designs. Following this development, each group produced a prototype, which was tested by actual end users in a hospital. Feedback and suggestions were then used to develop advanced design ideas.

In these studies of designing with the support of a networked collaborative environment, Seitamaa-Hakkarainen et al. (2001) found that a key aspect of these environments is the provision of tools for progressive discourse interaction between the designers and users of the future products. Further, the environments offer shared spaces and tools to elaborate conceptual knowledge related to the design problem. The collaborative technology made design thinking more explicit and accessible to the fellow designers, and enabled participants to share their ideas and construct a joint understanding of design problems and solutions.

Supporting and Guiding Students' Progressive Inquiry Processes

A special question in implementing progressive inquiry and knowledge-building practices in higher education is the teacher's or tutor's role in supporting and guiding students' collaborative inquiry. In progressive inquiry, the traditional role of a teacher as an expert, who delivers the essential information by lecturing, is radically changed. The important roles of the teacher and the facilitators of collaboration are:

1. to create the context for the collaborative inquiry practices by organizing the community's activities and establishing the underlying conditions of the educational setting (Lakkala et al., 2005), and;
2. to supervise and scaffold the process, keep it active and in focus during the progression of the course, and to help students gradually take on themselves the responsibility for the higher-level cognitive processes (Scardamalia, 2002).

Our recent studies indicate that, in order to successfully engage students' in progressive inquiry, the whole educational setting should be carefully planned and constructed to support the eligible activities, because expert-like inquiry practices do not emerge spontaneously. However, it does not mean pre-planning detailed tasks, but *indirectly* designing conditions for the community's activities (Jones, Dirckinck-Holmfeld, & Lindström, 2006). Building on previous studies (e.g., Bielaczyc, 2001; Guribye, 2005; Lakkala et al., 2005; Paavola, Lipponen, & Hakkarainen, 2002), we have started to use the notion *pedagogical infrastructures* to illustrate how the pedagogical design of collaborative inquiry practices resembles the construction of basic physical infrastructure to support smooth and effective functioning of people's daily activities. We have developed the *framework of pedagogical infrastructures* (Muukkonen, Lakkala, & Paavola, in press; Lakkala, Muukkonen, Paavola, & Hakkarainen 2008), consisting of *technical*, *social*, *epistemological*, and *cognitive infrastructure*, to be used to classify, design, and evaluate the elements of educational settings based on technology-enhanced collaborative inquiry. The separate infrastructures exist in parallel, and in a successful educational setting, all aspects are taken care of and designed to foster collaborative knowledge creation: appropriate technology is easily accessible and it is used meaningfully (technical infrastructure); there is deliberate collaboration built into the tasks (social infrastructure); students' activity is organized to be creative working with knowledge, not just internalizing certain content (epistemological infrastructure); and students' autonomy and the development of skills are supported by explicit cognitive modeling of expert-like practices (cognitive infrastructure).

In addition to appropriate and supportive overall design of the educational setting for progressive inquiry, timely guidance and tutor's participation throughout the process is important for further shaping and directing students' engagement in inquiry. In one study (Muukkonen et al., 2005) we compared students' technology-enhanced inquiry practices in

two conditions, either with or without tutor participation. As evidence for the tutors' role, the process analysis recovered a more focused problem-setting tendency in tutored groups: instead of opening up new lines of questions, the tutored groups were more likely to present subordinate questions to previous ones. The results of another study, comparing three tutors' scaffolding practices (Lakkala et al., 2007b), indicated that the kind of advice and expert model that the tutor offered had an effect on the style of the inquiry discourse in the group: either it concentrated more on theory reviewing, focusing of the inquiry, or generating of divergent ideas. The scaffolding that appears to have promoted deepening discourse may be characterized as explicitly built on the students' preceding discourse and, accordingly, as providing a content-specific and well-timed expert recommendation to refocus the inquiry.

FUTURE TRENDS

Productive changes in educational systems towards establishing inquiry-based approaches in studying and teaching call for an alignment of pedagogical and institutional goals and actions (Muukkonen et al., 2004). Learning technologies also need to be critically viewed for their role in fostering expert-like skills in advancing knowledge. For instance, availability of scaffolding, support for multiple forms of collaboration, and shared development of knowledge objects are challenges for designing learning technologies.

CONCLUSION

The progressive inquiry model may be utilized in a variety of educational settings to provide a heuristic framework for the key activities and epistemic goals of a knowledge-building community. The community may provide multiple levels of expertise and, equally importantly, social support for engaging in a strenuous quest for learning and advancing knowledge. Especially in higher education, a progressive inquiry approach may support the development of academic literacy, scientific thinking, and epistemic agency, particularly when integrated with the use of appropriate collaborative technology and supportive arrangements in curriculum design.

REFERENCES

Bereiter, C. (2002). *Education and mind in the knowledge age.* Hillsdale, NJ: Lawrence Erlbaum.

Bielaczyc, K. (2001). Designing social infrastructure: The challenge of building computer-supported learning communities. In P. Dillenbourg, A. Eurelings, & K. Hakkarainen (Eds.), *European perspectives on computer-supported collaborative learning.* Maastricht: Maastricht McLuhan Institute.

Brown, A.L., Ash, D., Rutherford, M., Nakagawa, K., Gordon, A., & Campione, J. (1993). Distributed expertise in the classroom. In G. Salomon (Ed.) *Distributed cognitions: Psychological and educational considerations* (pp. 188-228). Cambridge: Cambridge University Press.

Brown, A.L., & Campione, J.C. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory & classroom practice* (pp. 229-287). Cambridge, MA: MIT Press.

Carey, S., & Smith, C. (1995). On understanding scientific knowledge. In D.N. Perkins, J.L. Schwartz, M.M. West, & M.S. Wiske (Eds.), *Software goes to school* (pp. 39-55). Oxford, UK: Oxford University Press.

Chi, M.T.H., Bassok, M., Lewis, M.W., Reiman, P., & Glaser, R (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145-182.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real world laboratories. In R.J. Sternberg & J. Davidson (Eds.), *Mechanisms of insight* (pp. 365-395). Cambridge, MA: MIT Press.

Geisler, C. (1994). *Academic literacy and the nature of expertise.* Hillsdale, NJ: Lawrence Erlbaum.

Glaser, R., & Chi, H.T.M. (1988). Overview. In H.T.M. Chi, R. Glaser, & M. Farr (Eds.), *The nature of expertise* (pp. xv-xxviii). Hillsdale, NJ: Lawrence Erlbaum.

Guribye, F. (2005). *Infrastructures for learning: Ethnographic inquiries into the social and technical conditions of education and training.* Doctoral Dissertation, University of Bergen, Norway. Retrieved September 16, 2007, from http://hdl.handle.net/1956/859

Hakkarainen, K. (2003). Emergence of progressive inquiry culture in computer-supported collaborative learning. *Learning Environments Research, 6,* 199-220.

Hakkarainen, K., Järvelä, S., Lipponen, L., & Lehtinen, E. (1998). Culture of collaboration in computer-supported learning: Finnish perspectives. *Journal of Interactive Learning Research, 9,* 271-287.

Hakkarainen, K., & Sintonen, M. (2002). Interrogative model of inquiry and computer-supported collaborative learning. *Science & Education, 11,* 25-43.

Hatano, G., & Inagaki, K. (1992). Desituating cognition through the construction of conceptual knowledge. In P. Light & G. Butterworth (Eds.), *Context and cognition:*

Ways of knowing and learning (pp. 115-133). New York: Harvester.

Hintikka, J. (1999). Inquiry as inquiry: A logic of scientific discovery. *Selected papers of Jaakko Hintikka* (vol. 5). Dordrecht, The Netherlands: Kluwer.

Jones, C., Dirckinck-Holmfeld, L., & Lindström, B. (2006). A relational, indirect, meso-level approach to CSCL design in the next decade. *International Journal of Computer-Supported Collaborative Learning, 1*(1), 35-56.

Lakkala, M., Ilomäki, L., & Palonen, T. (2007). Implementing virtual, collaborative inquiry practices in a middle school context. *Behaviour & Information Technology, 26*(1), 37-53.

Lakkala, M., Lallimo, J., & Hakkarainen, K. (2005). Teachers' pedagogical designs for technology-supported collective inquiry: A national case study. *Computers & Education, 45*(3), 337-356.

Lakkala, M., Muukkonen, H., Paavola, S., & Hakkarainen, K. (2008). Designing pedagogical infrastructures in university courses for technology-enhanced collaborative inquiry. *Research and Practice in Technology Enhanced Learning, 3*(1), 33-64.

Lipponen, L. (2000). Towards knowledge building discourse: From facts to explanations in primary students' computer mediated discourse. *Learning Environments Research, 3*, 179-199.

Muukkonen, H., Hakkarainen, K., & Lakkala, M. (1999). Collaborative technology for facilitating progressive inquiry: Future learning environment tools. In C. Hoadley & J. Roschelle (Eds.), *Proceedings of the Computer Support for Collaborative Learning (CSCL) 1999 Conference* (pp. 406-415). Mahwah, NJ: Lawrence Erlbaum.

Muukkonen, H., Hakkarainen, K., & Lakkala, M. (2004). Computer-mediated progressive inquiry in higher education. In T.S. Roberts (Ed.), *Online collaborative learning: Theory and practice* (pp. 28-53). Hershey, PA: Information Science.

Muukkonen, H., Lakkala, M., & Hakkarainen, K. (2001). Characteristics of university students' inquiry in individual and computer-supported collaborative study process. In P. Dillenbourg, A. Eurelings, & K. Hakkarainen (Eds.), *Proceedings of the 1st European Conference on Computer-Supported Collaborative Learning* (pp. 462-469). Maastricht, The Netherlands: Maastricht McLuhan Institute.

Muukkonen, H., Lakkala, M., & Hakkarainen, K. (2005). Technology-mediation and tutoring: How do they shape progressive inquiry discourse? *Journal of the Learning Sciences, 14*(4), 527-565.

Muukkonen, H., Lakkala, M., & Paavola, S. (in press). Promoting knowledge creation and object-oriented inquiry in university courses. In S. Ludvigsen, A. Lund, & R. Säljö, R (Eds.), *Learning in social practices. ICT and new artifacts transformation of social and cultural practices.* Oxford, UK: Pergamon (EARLI Series: Advances in Learning).

Paavola, S., Lipponen, L., & Hakkarainen, K. (2002). Epistemological foundations for CSCL: A comparison of three models of innovative knowledge communities. In G. Stahl (Ed.), *Computer support for collaborative learning: Foundations for a CSCL community* (pp. 24-32). Hillsdale, NJ: Lawrence Erlbaum.

Perkins, D.A., Crismond, D., Simmons, R., & Under, C. (1995). Inside understanding. In D.N. Perkins, J.L. Schwartz, M.M. West, & M.S. Wiske (Eds.), *Software goes to school* (pp. 70-87). Oxford, UK: Oxford University Press.

Salomon, G., & Perkins D.N. (1998). Individual and social aspects of learning. *Review of Research of Education, 23*, 1-24.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67-98). Chicago: Open Court.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *Journal of the Learning Sciences, 3*, 265-283.

Seitamaa-Hakkarainen, P., Lahti, H., Muukkonen, H., & Hakkarainen, K. (2000). Collaborative designing in a networked learning environment. In S.A.R. Scrivener, L.J. Ball, & A. Woodcock (Eds.), *Collaborative design: The proceedings of CoDesigning 2000* (pp. 411-420). London: Springer.

Seitamaa-Hakkarainen, P., Raunio, A.M., Raami, A., Muukkonen, H., & Hakkarainen, K. (2001). Computer-support for collaborative designing. *International Journal of Technology and Design Education, 11*, 181-202.

Veermans, M., & Järvelä, S. (2004). Generalized achievement goals and situational coping in inquiry learning. *Instructional Science, 32*(4), 269-291.

Wells, G. (1999). *Dialogic inquiry: Towards a sociocultural practice and theory of education.* Cambridge, UK: Cambridge University Press.

KEY TERMS

Distributed Expertise: Cognition and knowing are distributed over individuals, their tools, environments, and networks.

Epistemic Agency: Taking responsibility over one's own learning efforts and advancement of understanding.

Knowledge Building*:* A framework for collective knowledge advancement and development of knowledge artifacts.

Learning Community/Community of Learners: All participants in a learning process (students, teachers, tutors, and experts) have valuable expertise and skills, which benefit collective efforts.

Metaskills: Skills involved in academic literacy, as well as metacognitive skills related to planning, monitoring, and regulating comprehension-related activities.

Progressive Inquiry: A pedagogical model for structuring and supporting a group of learners in a deepening question-explanation process.

Pedagogical Infrastructures: In the present context, the concept is intended to refer to the elementary preconditions that should be designed to shape and support collaborative inquiry practices in educational settings.

Scaffolding: Providing support, which enables a learner to carry out a task that would not be possible without that support, and enabling the learner gradually to master that task without support.

Teens and Information and Communication Technologies

Leanne Bowler
McGill University, USA

INTRODUCTION

In the late 20th century, the digital revolution in information and communication technology (ICT) moved into the homes and private lives of ordinary people. Unsurprisingly, the early adopters of domesticated ICT have been youth, young people between the ages of 12 and 19, whose lives have become increasingly shaped and mediated by information and communication technologies. Called the "Net Generation" (Tabscott, 1998, p. 3), these young people are leading the charge toward what the United Nations has called "an unprecedented and global media culture" (United Nations, 2003. p. 311).

The focus of this article is on how young people, ages 12–19, in the early 21st century use information and communications technologies. The wide and diverse nature of the landscape, composed of multiple platforms and applications in continuous change, necessitates a broad approach. Information technologies are now bundled with communications capabilities and vice versa, making a focus on one and not the other virtually impossible. Furthermore, one of the difficulties in studying ICT use among children and teenagers is that statistics and studies are still limited, even within digitally privileged countries. Ironically, while research in this area has focused on the educational use of ICT, young people overwhelmingly use it for personal reasons. This article, therefore, looks at ICT through a wide angle and offers a snapshot of the role of ICT in the lives of young people in the early days of the 21st century, suggesting in broad terms where the emerging issues and trends may lie.

BACKGROUND

Youth have traditionally been the early adopters of digital ICT, forging new patterns of information and communication behavior. The next generation of ICT users, those born after 2000, will move with even greater ease among the emerging information and communications technologies. This generation will enter their teen years never having known a world without personal computers, the Internet, cellular telephones (more commonly called "cell phones" in North America and "mobile phones" in the UK), and personal digital assistants. While the Net Generation's first experiences on the Web, and with ICT in general, were typically asynchronous and tied to a physical location, namely the home or classroom, young people who are now entering their teens increasingly find that information and communication technologies are accessible anywhere, anytime, and anyplace. Cell phones are quickly becoming personal digital assistants, providing a broad range of information services beyond basic voice capabilities. Portable hardware such as MP3 players and the "podcasts" used to deliver content from the Internet to the device have helped move the Internet beyond the desktop and into the street. The onset of Web 2.0 — the social Web — has further enhanced the immediacy of the experience.

For many young people living in digitally privileged societies, ICT represents a world of entertainment, the most popular activities being communicating with friends, online gaming, and downloading music (United Nations, 2003). ICT now rivals home and school as a "space" for socialization and identity development. While opportunities await technology-savvy educators and marketers — reaching young people "where they live" and in a language they understand — these same opportunities can turn to manipulation and threat in a technology-rich, media-saturated world that is sometimes disconnected from the worlds of parents and other adults significant in the lives of teens. Whether young people will be at risk in this world, or will adapt to and even shape it, is a question to consider.

ACCESS TO ICT

How pervasive is ICT in the lives of youth? In the United States, 9 out of 10 teens are Internet users. The vast majority (84%) report owning one personal media device — a computer, a cell phone, or a personal digital assistant — and half of American families with teens have broadband connections to the Internet. Eighty-seven percent of American youth between 12 and 17 years old have used the Internet, and of that number, half (51%) report going online at least once a day (Lenhart, Madden, & Hitlin, 2005). Across the border in Canada, the situation is similar: 94% of young people in grades 4 to 11 (ages 9 to 17 years) report going online from home. Sixty-one percent of Canadian online youth have high-speed access and 23% have their own cell phone, 44% of which have Internet capability (Environics Research Group, 2005). Australian youth are among the world's leading users of computers, with 94% of Australian students reporting that they have access to a home computer for schoolwork and 100% reporting that they have access to a computer at school

(OECD, 2003). In the United Kingdom, 75% of youth between the ages 9–19 have accessed the Internet from a computer at home, and school access is nearly universal (92%). Young people in the UK use as diverse a range of platforms as those in the United States and Canada, with 71% living in a home with a computer and 38% owning a cell phone (Livingstone & Bober, 2005). Access to computers is almost universal for 15-year-olds living in countries of the Organization for Economic Cooperation and Development (OECD): 98% or more in 21 of the 25 OECD countries that participated in the 2003 PISA study have experience with computers, and the vast majority of these young people report confidence performing basic ICT skills, such as opening, deleting, and saving files, and using the Internet (OECD, 2003). Internet access is high, if not universal, in schools throughout much of Europe and large areas of Asia, specifically Australia, Singapore, Korea, Hong Kong, Taiwan, and New Zealand (Kirkman, Cornelius, Sachs, & Schwab, 2002).

Despite the seeming pervasiveness of ICT in the lives of youth, inequity of access exists. The OECD (2001) defines this as the digital divide — "the gap between individuals, households, businesses and geographic areas at different socio-economic levels with regard to both their opportunities to access information and communication technology (ICTs) and to their use of the Internet for a wide variety of activities."

Internationally, the lines between those with access and those without are clearly drawn between developed and developing nations. In developed countries, virtually every child has access to a telephone, a television, and a computer, either at home, school, or a public library. The same cannot be said for youth in developing countries (United Nations, 2003). Access to computers in developing countries is typically through the school, but not all children go to school because many countries do not have a universal, free education system. The Internet reaches little more than 10% of the world's population; while 331 per 1,000 people in Europe use the Internet, only 37 per 1,000 in the Middle East and Africa, 92 per 1,000 in Latin America and the Caribbean, and 15 per 1,000 in South Asia and sub-Saharan Africa use the Internet (United Nations, 2005). While these figures include all Internet users, and not just youth, young people are often the first adopters of ICT. Therefore these figures suggest a profound gap between the ICT experiences of youth in developed countries and developing countries.

Even within developed nations, the distribution of ICT access can be uneven. Physical access to a computer, a network connection, and increasingly a cell phone or personal digital assistant, is the starting point. Income, social class, and proximity to urban areas play a key role in determining the accessibility of such tools. While many schools provide computers for student use, the student-computer ratio may be so high, the quality of hardware and software so poor, and the network connection so slow, that access to technology is more theoretical than practical. Beyond the basic physical access, users of ICT must have intellectual access in the form of multiple literacies: the basic literacy skills of reading and writing, the language skills to understand and contribute to the discourse, and the critical thinking skills required to decode media messages and sift through a myriad of information sources.

HOW ARE YOUTH USING ICT?

This section looks at what young people are doing with ICT and studies this question within the framework of "purpose" rather than format, application, or specific technology. So, how are young people using ICT in their lives? Regardless of the mode of delivery, young people use ICT for three principle reasons: as a tool for learning, a channel for human interaction, and a form of entertainment. At times, these purposes are deeply intertwined, as in the case of networked learning environments or virtual reality games that teach.

ICT and Learning

ICT has become an essential educational tool in the 21st century, and for purposes of learning, young people most commonly use it to find information resources on the Web (OECD, 2003). Fifty-five percent of students in OECD countries report searching the Internet for information about people, things, or ideas, with the highest use of the Internet as a source for information resources in Canada (75%), the United States (74%), and Australia (74%) (OECD, 2003). For Canadian youth, searching the Internet for

information is as popular as playing games online, and they willingly choose the Internet over other information sources (Environics Research Group, 2004). American youth look for information about current events, politics, religion, careers and colleges, and increasingly, health. While its not clear whether American teens seek information on these topics for educational or personal reasons, the Internet is now a key source for information, especially for those who have access to broadband connections (Lenhart et al., 2005).

The shear volume of information on the Internet can be more frustrating than useful when it threatens to overwhelm young people. Young adults report feelings of being lost in a sea of information and say they have problems filtering the "good" information from the "bad" (Environics Research Group, 2004; McMillan & Morrison, 2006). A meta-analysis of research related to youth information-seeking behavior revealed that young people are not experiencing the richness of the Internet because of "poorly developed information-seeking skills or a propensity to take the easiest path possible" (Dresang, 2005, p. 181). For those teens who do happen to stumble upon information they feel is useful, the operative phrase "use with caution" still remains. In an environment

where the traditional filters of editor and publisher have been cast aside and anyone with Web publishing software can launch their own Web page, it is difficult to determine the reliability of information. The problem is more acute in the world of Web 2.0 — the second generation of the Internet, the social Web — where the Internet it is the facilitator of conversations and users *are* the content. In this new online world of wikis, blogs, and socially mediated knowledge bases such as *YahooAnswers*, the lines between author, publisher, and user have grown ever more fuzzy. Now, typing and clicking a mouse is all that is required to add content to the Internet. Given this profusion of information choices/sources and the corresponding decrease in their reliability, the ability to effectively find, choose, and use information has become a key educational outcome for schools — generally falling under the rubric of information literacy, the set of problem-solving skills needed to negotiate a complex world of information.

Information seeking on the Internet is the principle online learning activity of teens (OECD, 2003). Increasingly, young people have access to networked educational software, such as *Blackboard* or *WebCT*. As its name suggests, networked educational software is an ICT tool designed specifically for the purposes of teaching and learning in an online environment. It is actually a suite of applications and it operates in a closed, password-protected environment where teachers can moderate online discussions, share course materials, assess student learning, and communicate with students, parents, and other teachers. On a smaller scale, educational content can now be delivered directly into the hands of teens via their cell phones; class reminders can be e-mailed to mobile devices, Web-enabled browsers for small screens provide access to digital information resources, and podcasts — audio files distributed by subscriptions over the Internet — can be sent to MP3-enabled cell phones.

While the Internet plays an increasingly important role in education, its overall use by young people for school-related reasons remains surprisingly low in comparison to other functions such as communication with friends, downloading music, and online gaming (United Nations, 2003). However, as the size and cost of digital devices decreases and the number of broadband and wireless connections increases, it seems safe to say that the digital classroom, as yet more theory than practice, will move closer to the reality.

ICT and Human Interaction

One of the principle tasks of adolescence is to establish a personal identity within the social framework of family, friends, school, and community. Friendship and a sense of belonging are critical to the development of self- and social identity. It is no surprise, then, that young people increasingly take advantage of ICT tools that enable social networking. The communications possibilities offered through ICT — such as e-mail, chat and instant messaging, text messaging, multimedia messaging, blogs, and Web sites that leverage social networking capabilities — offer young people a forum for the expression of identity and a way to connect with others. Teens who live in a "wired world" can choose to communicate anytime, anywhere, and sometimes more worryingly, with anyone.

Social networking on the Internet is an essential component of Web 2.0, the next generation of the Web. The digital equivalent to "hanging out at the mall," social networking Web sites provide a popular platform for teen communication and personal expression. Web sites such as MySpace and Facebook allow users to post a profile and build a personal network, all within the context of the free and open Web. Teens with something more to say can begin a blog (a "Web-log"), which is essentially an online journal akin to the handwritten diaries and the personal homepages of earlier days. Blogs, however, are interactive and allow readers of the journal to post comments and contribute to the discourse.

More than half (55%) of all online American youths ages 12–17 use online social networking sites (Lenhart & Madden, 2007). In the UK, Bebo attracted more than 22 million members in its first 13 months, most of them school and college students (Ward, 2006). Social networking sites are about friendship: girls use them to reinforce pre-existing friendships, while boys use them for flirting and making new friends (Lenhart & Madden, 2007). Young people use Web sites' social networking sites to build community, to construct a social world that is distinct from their parents. Social networking is also about identity. Users in such environments can create false personas — an interesting experiment for a teenager exploring self-identity, but one that is more worrisome when created by an online predator. How to find a balance between the exploratory, identity-building behavior that is crucial for adolescent development and their online safety is a critical question for parents, educators, policymakers, and ICT developers in future decades.

E-mail and instant messaging (IM) are popular channels of communication via ICT, but a clear division between the two has arisen: E-mail is for communicating with adults and IM is for friends. E-mail is task-oriented; IM is for fun (Schiano et al., 2002; Lenhart et al., 2005). Young people in the early 21st century are indeed the "IM Generation," instant messaging being the most popular of Internet communication modalities (Greenfield, Subrahmanyam, Suzuki, & Tynes, 2006, p. 198). IM allows young people to have private real-time conversations with a friend while "hanging out" with a group of friends — two important functions built into one application. While there are concerns regarding who teens let into their "buddy" group when communicating via IM, they tend not to stray beyond their in-person network of friends, most typically other teens they have met at school or who live nearby (Boneva, Quinn, Kraut, Kiesler, & Shklovski, 2006). But interestingly, IM can be

used to maintain ties with family; in the United States, almost one in three American teens (29%) say they use IM to communicate with their parents (Lenhart et al., 2005). As a space for personal expression, IM offers users many options beyond text-based communication. IM-using teens also share photos, music and video files, and links to Web sites or articles (Lenhart et al., 2005). The medium of IM opens up possibilities for information seeking on topics that teens may be too embarrassed to ask in person. That information is, of course, only as reliable as the person who answers the question. In a move to fill this "reliability gap," libraries have begun to offer synchronous, online reference services using IM technology. Called "virtual reference," these services are aimed squarely at the young adult clientele.

As IM spreads to new, smaller platforms, such as cell phones and handheld digital assistants, teens are taking their text-based conversations to the street, with the subsequent blurring between the distinction of "home" and "not home" (Minoura, 2001). While this means that teens are no longer "land locked" by their desktop computers, they are also "placeless" and often "faceless," a development that threatens to break the social bonds that are built through face-to-face communication. But teens still prefer face-to-face time spent with friends and seem to use IM as a supplement to, and not a replacement for, in-person communication (Lenhart et al., 2005; Boneva et al., 2006).

The cell phone has become a part of the everyday lives of youth living in affluent countries. Portable, easy to use, and inconspicuous, cell phones are a basic tool for young people growing up in the information age. For example, in North America at least 45% of American teens and 46% of Canadian youth in grade 11 own cell phones (Lenhart et al., 2005; Media Awareness Network, 2005). Approximately 80% of young people in the European Union use a cell phone at least once a week (United Nations, 2005). The growth of cell phone use among young people has been phenomenal and continues to expand. In 1999, only 15% of Finnish 15-year-olds owned a cell phone. Three years later, in 2001, that number had climbed to 66% (United Nations, 2003).

The functionality of cell phones has moved beyond the traditional voice capabilities to include text messaging, multimedia messaging, wireless e-mail, MP3 players, Internet browsers, and digital cameras, and young people are taking full advantage of this array of options. For many young people, cell phones are the e-mail terminal of choice, some even accessing a scaled-down Web built for small screens, slower speeds, and keypunching. "Texting" (or SMS) — short text messages sent between cell phones — is growing in use, with at least 33% of American youth reporting that they text, rather than talk by cell phone. Interestingly, almost one in three (29%) of teens who use text messaging or IM use it to communicate with their parents, indicating the emergence of a new channel of communication between parent and child (Lenhart et al., 2005).

ICT and Entertainment

Teens and technology go hand in hand when it comes to leisure and recreation, and for many young people, ICT is synonymous with entertainment (United Nations, 2003, p. 321). Most large-scale studies that have looked at teens and ICT have focused on its use in educational settings and as a tool for communication. Reports from the PEW/American Life series and the Kaiser Family Foundation, which both looked at youth in the United States, provide a snapshot of how online teens use ICT to interact with media in the service of fun and entertainment (Lenhart et al., 2005; Lenhart & Madden, 2007; Rideout, Roberts, & Foehr, 2005). While these reports do not represent the global picture, they illustrate how the separate worlds of teens, entertainment, and ICT are converging in the 21st century. Calling today's young people "Generation M" — for media — the Kaiser Foundation reports that youth, ages 8–18, in the United States spend an average of up to six-and-a-half hours each day with media (Rideoute et al., 2005). While this statistic includes both analog and digital media, such as television, movies, magazines, music, and video games, it nevertheless represents an enormous chunk of time in the lives of youth; and given that the "amount of media a person used to consume in a month can be downloaded in minutes and carried in a device the size of a lipstick tube" (Rideout et al., 2005), media is now rooted in all aspects of the lives of teens who have access to ICTs. The following section discusses briefly two uses of ICT for entertainment — music downloading and online gaming.

Music is a vital part of many teens' lives: on a typical day, 68% of young people in the United States listen to music, either by CD, tape, or MP3 player (Rideout et al., 2005), so it is no surprise that the Internet, embedded as it is in the lives of many, has been appropriated by young people as a tool for retrieving and sharing music. Using peer-to-peer networks, online music services such as iTunes or BuyMusic.com, instant messaging, or e-mail, music can be downloaded and then listened to using a computer or an MP3 player. In the United States, just over half (51%) of online teens in the United States say they have downloaded music, and nearly one-third (31%) have downloaded video files (Lenhart & Madden, 2005). While these teens are at ease downloading content, many have little concern for the actual source of music, with the majority (55%) of downloading teens saying they do not care whether the music is copyrighted or not (Lenhart & Madden, 2005, p. 14). But over half of online teens are creating content too, for their own amusement and to entertain other teens (Lenhart & Madden, 2005).

Eight in ten teens in the United States (81%) who use the Internet play online games (Lenhart & Madden, 2007), interacting with other players in virtual worlds via multiplayer online games. Some games are limited to a few players; others have worldwide subscriptions in the millions. One example

is the game "Second Life," a 3D virtual world inhabited by 1.3 million people/players. A teen version, called "Teen Second Life" exists as well. Players in "Second Life" occupy a digital continent, where they meet new people, buy property and build homes, learn new skills, and create businesses. In this virtual world, entertainment, communication, and education have merged, and educators have taken note. Already one public library in the United States, aware that gaming, multimedia, and ICTs play a huge role in the lives of many of their teen patrons, has launched a pilot project exploring the creation of teen library services in a virtual world (PLCMC, 2006). McMaster University in Canada has opened a library facility in "Second Life." Students can visit it and even speak to a librarian, represented of course by an avatar (McMaster University, 2006). The use of online gaming as a platform for information seeking and learning is still in its infancy, but opportunities for such experiences can be expected to expand in the 21rst century.

FUTURE TRENDS

There is little doubt that teens will increasingly live in a wired world. Computers and cell phones will in time converge into mobile personal digital assistants — the digital "Swiss Army knives" of the 21st century (Rainie, 2006) — and the Internet will continue to spread worldwide. And yet little is known about teen patterns of behavior in relation to ICT and, more importantly, its impact on their lives. Many profound questions regarding the effect of ICT on the lives of teens remain, suggesting fruitful areas for research:

- How do we bring ICT to those who as yet have no access to it?
- For those who do have access to ICT, how are they using it? As ICT becomes more ever-present in the lives of teens worldwide, we might begin asking if the patterns of use are universal or if local culture makes a difference in the way young people use ICT.
- Given the ubiquitous nature of ICT in the lives of many teens, it seems critical to ask whether it has an effect on the intellectual, emotional, and physical development of young people.
- Do teens have the intellectual skills needed to safely and effectively use ICT? If not, what should be the focus of intervention?
- Should ICT-related applications and interfaces be designed specifically for teens, and if so, how? To what extent do teens adapt to existing technology? To what extent do they shape it?
- We know that teens use ICT as a channel for socializing and interacting with their peers. What, then, is the nature of teen online culture? Does it affect family relationships? Does participation in social groups mediated by ICT differ from traditional participation? Are traditional modes of socializing and participating falling to the wayside, or does ICT simply supplement them?
- The Internet is a primary source of information for online teens and yet we are only just beginning to understand how they use it to look for information. How do teens use ICTs to fulfill their information needs?
- For teens who use ICTs, the lines between learning, socializing, and entertainment are merging. How close can the lines be drawn? Are there aspects of these activities that should remain distinct? If not, what is the best way to leverage one, without sacrificing the other two?
- Computers and cell phones will in time converge into mobile personal digital assistants. How are teens using these portable tools, and what is the impact on their lives? A natural extension of this question should be asked by educators: Can personal digital assistants be used in service to teaching and learning, and if so, how?
- And finally, does ICT function within a policy framework that meets the needs of young people?

CONCLUSION

Despite its phenomenal growth, information and communication technology remains a tool for the privileged, and access to it is not the norm throughout the world. It is still the case that on a global scale, very few teens actually have access to ICTs. This digital divide is an inequity so fundamental that one could compare it to the effects of being illiterate in a world of people who read. Even so, experience has shown that when they do have access to ICT, it is young people who lead the charge in adopting it into their everyday lives.

REFERENCES

Boneva, B., Quinn, A., Kraut, R., Kiesler, S. & Shklovski, I, (2006). Teenage communication in the instant messaging era. In R. Kraut, M. Robert, & S. Kiesler (Eds.), *Computers, phones, and the Internet* (pp. 201-218). New York: Oxford University Press.

Dresang, E.T. (2005). The information-seeking behavior of youth in the digital environment. *Library Trends, 54*(2), 178-196.

Environics Research Group. (2005). *Young Canadians in a wired world.* Retrieved November 15, 2006, from http://www.media-awareness.ca/english/research/YCWW/index

Greenfield, M., Gross, E., Subrahmanyam, K., Suzuki, L., & Tynes, B. (2006). Teens on the Internet: Interpersonal connection, identity, and information. In R. Kraut, M. Brynin, & S. Kiesler (Eds.), *Computers, phones, and the Internet* (pp. 185-200). New York: Oxford University Press.

Kirkman, G., Cornelius, P., Sachs, J., & Schwab, K. (2002). *The global information technology report: Readiness for the networked world.* New York: World Economic Forum.

Lankes, R.D., Silverstein, J., & Nicholson, S. (2007). *Participatory networks: Libraries as conversations.* Retrieved February 10, 2007, from http://iis.syr.edu/projects/PNOpen/ParticiaptoryNetworks.pdf

Lenhart, A., & Madden, M. (2005). *Teen content creators and consumers.* Retrieved January 15, 2007, from http://www.pewinternet.org/pdfs/PIP_Teens_Content_Creation.pdf

Lenhart, A., & Madden, M. (2007). *Social networking Web sites and teens: An overview.* Retrieved February 8, 2007, from http://www.pewinternet.org/pdfs/PIP_SNS_Data_Memo_Jan_2007.pdf

Lenhart, A., Madden, M., & Hitlin, P. (2005). *Teens and technology.* Retrieved January 15, 2007, from http://www.kff.org/entmedia/upload/Executive-Summary-Generation-M-Media-in-the-Lives-of-8-18-Year-olds.pdf

Livingstone, S., & Bober, M. (2005). *UK children go online: Final report of key project findings.* Retrieved January 15, 2007, from http://personal.lse.ac.uk/bober/UKCGOfinal-Report.pdf

McMaster University. (2006). *Getting into virtual worlds.* Retrieved February 10, 2007, from http://ulatmac.wordpress.com/2006/12/06/getting-into-virtual-worlds/

McMillan, S., & Morrison, M. (2006). Coming of age with the Internet: A qualitative exploration of how the Internet has become an integral part of young people's lives. *New Media and Society, 8*(1), 73-95.

Minoura, Y. (2001). Children and media. In N. Kobayashi (Ed.), *The bright and dark sides of the information revolution. A cultural ecological perspective* (pp. 87-103). Tokyo: Hoso-Bunka Foundation.

OECD (Organization for Economic Cooperation and Development). (2003). *Are students ready for a technology-rich world? What PISA tells us.* Retrieved January 11, 2007, from *http://www.oecd.org/dataoecd/28/4/35995145.pdf*

PLCMC (Public Library of Charlotte and Mecklenburg County). (2006). *Teen library in Second Life.* Retrieved June 1, 2007, from http://plcmc.org/teens/secondlife.asp

Rainie, L. (2006, October 27). *Digital natives: How today's youth are different from their "digital immigrant" elders and what that means for libraries.* Retrieved January 10, 2007, from http://www.pewinternet.org/PPF/r/71/presentation_display.asp

Rideout, V., Roberts, D., & Foehr, U. (2005). *Generation M: Media in the lives of 8–18 year olds. Report.* Retrieved January 15, 2007, from http://www.kff.org/entmedia/upload/Generation-M-Media-in-the-Lives-of-8-18-Year-olds-Report.pdf

Schiano, D.J., Chen, C., Ginsberg,, J., Gretarsdottir, U., Huddleston, M., & Isaacs, E. (2002, April). Teen use of messaging media. *Proceedings of the ACM CHI 2002 Conference on Human Factors in Computing Systems,* Minneapolis, MN. Retrieved January 15, 2007, from http://home.comcast.net/~diane.schiano/CHI2002.short.ttalk.pdf

Tapscott, D. (1998). *Growing up digital: The rise of the Net Generation.* New York: McGraw-Hill.

United Nations. (2003). *World youth report 2003: The global situation of young people.* Retrieved January 10, 2007, from http://www.un.org/esa/socdev/unyin/documents/ch12.pdf

United Nations. (2005). *World Youth Report 2005: Young People Today, and in 2015* (pp. 76-79). New York: Author.

Ward, M. (2006, March 23). *Teen craze over networking sites.* Retrieved February 10, 2007, from http://news.bbc.co.uk/1/hi/technology/4826218.stm

KEY TERMS

Information and Communication Technology (ICT): Technology that enables the handling of information and facilitates different forms of communication.

Instant Messaging (IM): A text-based method for communicating one to one or in groups in real time over the Internet using standard IP protocol.

OECD Countries: The member countries of the Organization for Economic Cooperation and Development (OEDC) include: Australia, Austria, Belgium, Canada, the Czech Republic, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, The Netherlands, New Zealand, Norway, Poland, Portugal, the Slovak Republic, Spain, Sweden, Switzerland, Turkey, the United Kingdom, and the United States.

Personal Digital Assistant (PDA): A handheld digital device that combines the functionality of the telephone with computing and networking. It can operate as a cell phone, a Web browser, an e-mail terminal, digital camera, and personal assistant.

Podcast: A method of publishing audio files to the Internet. The term is a combination of the word "iPod" and "broadcasting." Podcasts are often distributed through RSS feeds.

Social Networking Web Sites: Web sites that facilitate the development of online social networks and collaborative knowledge building through the use of social software.

Social Software: Software tools for computer-mediated communication. Includes instant messaging, text chat, blogs, wikis, and Internet forums. From these have arisen new areas of collaborative knowledge building such as folksonomies, social bookmarking, social citations, and knowledge bases.

Text Messaging: Short text messages received by and sent to a mobile, handheld communication device such as a cellular phone, a personal digital assistant, or a pager. Text messages can be also sent from the Web, either through the Web page of the cellular service provider or through some Web sites that offer to send text messages free of charge. Also called *texting, short message service,* or *SMS.*

Virtual Reference: An online information service that uses computer-mediated communication to answer questions. The service can be asynchronous (e-mail) or synchronous (instant messaging). Also called *digital reference.*

Web 2.0: The so-called second generation of the Web. A suite of Web-based services where users control the content by contributing, collaborating, and sharing. Sometimes called the *social Web,* Web 2.0 architecture is dependant on the participation of its users.

Telemedicine Applications and Challenges

Lakshmi S. Iyer
The University of North Carolina at Greensboro, USA

INTRODUCTION

Telemedicine is a function of information and communication technologies (ICT) that facilitates exchange of medical data to assist the health care industry in providing services to the society more competently. Its applications range from diagnosis, treatment, and prevention of disease, to continuing education of medical professionals, research and evaluation. Telemedicine is not a process aimed to replace traditional practices of medicine. It simply acts as a partner of the industry to reduce inadequacies in time and resources. ICT should not be viewed only as a competitive advantage of health care organizations, but rather a fundamental commodity intrinsic to the delivery of global health care (Iyer & Dey, 2005; Nash & Gremmillion, 2004).

Pedigo (1997) illustrates the essence of telemedicine with the follow example: In April 1995, a student at Peking University sent an e-mail requesting medical assistance for a fellow student, Zhu Ling. Zhu Ling was experiencing rapid hair loss and paralysis. An extensive online network of physicians, toxicologists and other experts collaborated with Ling's physician in Beijing to respond to the SOS e-mail. With the assistance and suggestions from over 2,000 responses, the Beijing physician was able to treat Ling in the best possible way and prevent death. The Zhu Ling case was the first recorded use of the Internet to seek diagnosis and patient care from a distance.

Telemedicine has the potential to help bridge the time and distance gaps that can mean life or death for some patients. It can provide live video conferencing between local, rural doctors and clinics to the necessary specialists at a major hospital or research center. These conferences can provide quick and accurate diagnosis and save both the patient and the doctor time and money.

This article presents a background on telemedicine including components, applications and benefits of telemedicine, challenges and trends in telemedicine, and conclusion with some direction for future research in telemedicine.

BACKGROUND

Telemedicine removes geographic barriers and is anticipated to save money by treating patients on-site rather than in an expensive hospital setting, improve patient care by giving health care providers access to teaching medicine resources, and target services to populations that have been hard to reach (remote rural areas), expensive to serve (prisons, mental institutions), and historically neglected (urban poor). The most important benefit of telemedicine is its ability to access patient data from any remote location (Demiris, 2004). It is impossible to have specialists in all areas available at all times to any given hospital or emergency care service. There are people worldwide that live in rural and remote areas who are not able to receive the type of care they need due to their distance from the nearest facility that specializes in their illness. Moreover, in most developing countries, there is a severe scarcity of medical specialists. Lack of capital, facilities, and systems are some of the common problems faced by developing countries. Telemedicine coupled with telecommunications can provide a solution to some of the above problems.

The U.S. Department of Defense has been using telemedicine technologies to support their operations in Saudi Arabia, Kuwait, Somalia, Haiti, Cuba, Panama, Croatia, and Macedonia (Garshnek, Logan, & Hassell, 1997). The telemedicine project in the Persian Gulf in 1993 had computerized tomography (CT) scanners installed in transportable modular military hospital units and deployed in the Saudi desert just south of the Iraqi and Kuwaiti borders. During Operation Restore Hope, physicians in Somalia were able to communication and share medical data with specialists in Washington DC.

Telemedicine has always played an important role in astro medicine as well. From the 1960s, astronauts have been monitored by groups of medical specialists through telemetry during the space operations. Currently, NASA is making efforts to hold conferences in the micro-gravity environment between astronauts on the orbiting space-crafts and the medical specialists on earth (Garshnek et al., 1997). These one-way video and two-way audio conferences would make a phenomenal difference in the safety and security of the astronauts on board.

Treatment of inmates in the prison (Cooper, 1997) is another application of telemedicine. It helps to maintain a secure prison system by minimizing movement of the prisoners in case of a medical problem. The state of Iowa has implemented a telemedicine project via which medical staff of the prison can consult with doctors at the University of Iowa through a two-way video conference. This system transmits captured images letting physicians located at a remote place view a patient's ears, throat, or skin. It also

enables sharing of x-rays and other information to help with diagnosis and follow-up care.

Telemedicine and telehealth also eliminates travel cost as well as travel delay (Jossi, 2005). Moreover, immediate real-time access to patient data gets rid of time lag and accelerates early detection of diseases that can improve overall performance of the health care industry (Jossi, 2005).

Medical information shared over a network can support research collaboration by allowing researchers to exchange findings over the networks at no additional cost. Informational networks online also provide a means to establish official and unofficial educational programs over a wide area across the globe.

Components of Telemedicine

The success of telemedicine depends on how effectively the capabilities of technology have been exploited to benefit the health care industry. Health care industry requirements should be analyzed carefully before considering technology as a solution. Telemedicine systems may be developed using two key dimensions: internal and external integrations (Raghupathi & Tany, 2002). Internal integration refers to technologies that are applied to integrate systems with one another within an organization. External integration refers to systems and technologies interfacing with outside organizations and agency computer systems.

The fundamental telemedicine integration should be planned to allow a scope for future expansion if necessary. Scalability should be used as a valuable measuring rod for every telemedicine project. The basic components of a telemedicine project infrastructure are discussed in the following sections.

Telecommunication

The first step is to ensure a network connecting all remote facilities in order to communicate with each as desired. This could vary from a basic telephone service to broadband Internet. Considering complex operations requiring huge amounts of data being interchanged across the globe in seconds between systems, telemedicine networks often require a high bandwidth. Asynchronous transfer mode (ATM) coupled with resilient synchronous optical network (SONET) has been one of the most popular configurations from the early 2000s. It offers high-quality and low-delay conditions. These systems are supported by fiber optic cables that allow data to be transferred up to 40 gigabytes per second.

Mobile communication systems are also critical to telemedicine industry. This includes cordless, cellular, satellite, paging, and private mobile radio systems (Ackerman, et al., 2002). Wireless technology is the next big step for telemedicine. Wireless end users within a physician's office, hospital building, or even medical campus can be connected with a wireless local area network (WLAN).

Interoperable Systems

Interoperability adds value to the system by ensuring flexibility and cost-effectiveness (Ackerman et al., 2002). The system design should allow stations developed by independent vendors to interact with each other. Medical devices and other peripherals connected to one vendor's station should be able to interact with that of another station created by another vendor. Systems should be further designed to allow creation of individual stations in a plug-and-play

Table 1. Telemedicine interactions and technical requirements (Adapted from Garshnek et al., 1997)

Applications	Interaction Processes	Data Transferred	Min. Bandwidth Reqd.
* Telepsychiatry * Remote Surgery * Interactive Exams	Real time, one-way or two-way interactive motion video	Voice, sound, motion video, images, text	Moderate to High
* Dermatology * Cardiology * Otolaryngology * Orthopedics	Still images or video clips with real-time telephone voice interaction, 'store and forward' with data acquired and sent for later review	Voice, sound, still video images, text	Low to Moderate
* Distance Education * Training	One-way or two-way real-time or delayed video	Voice, sound, motion video, images, text	Full Spectrum: Low to High
* Health Info. Networks * Medical Records	Transfer of electronic text, image, or other data	Text, images, documents, related data	Low to High

fashion from components developed by multiple vendors. Middleware is a possible solution to ensure interoperability with systems.

Computing Processes

Store-and-Forward: It is usually used for sending digital data between hosts in a telemedicine network. Images taken on digital cameras or still videos are sent as simple e-mail attachments. Depending on the data range, computing power and speed requirements may vary from two desktops connected to the Internet to a whole grid-enabled network.

Real-Time Video Linking: Typically higher bandwidth communication channels (ISDN lines) are required to enable real-time processing (Biomedical Informatics Ltd., 2003). It is the basis for "face-to-face" consultation between patients and specialists located at two different parts of the world. Specialized video conferencing equipments at both locations are used to facilitate "real-time" consultation.

Computing power and speed are determined based on the combination of processes involved in the particular telemedicine application. Table 1 illustrates the differences in the technical requirements of telemedicine in respect to its application.

Telemedicine Equipment

There are a huge range of devices that are used in telemedicine for acquisition, presentation, storage and delivery of medical data (Mirza 2004). In this article, a few typical equipments are discussed as follows:

- **Electric Phone Stethoscope** is used to pass high quality auscultation sounds over low bit rate with switches at receiving units picking diaphragm frequency sounds. An e-steth produces phonocardiogram by digitalizing heart and lung sounds using a sound card. The relevant data is then attached to an email for transmission. Physicians opening the email attachments receive the pertinent diaphragm frequency with the phonocardiogram and sound playing in the background automatically.

- **Telemedicine Video Imaging System** assists diagnosis of patient data. Special cameras with specific features to power zoom, auto focus, frame capture, and electronic image polarization together are used for video imaging.
- **Vital Signs Devices** is typically used for homebound patients. It helps in the constant monitoring of heart rate, respiratory rate, blood pressure, and temperature of patients located in remote areas. Patient data can be transmitted to the hospital through this machine while the patient remains at their home or any other remote location.

APPLICATIONS OF TELEMEDICINE

Telemedicine enables health care providers to deliver efficient and cost effective quality care to persons at some distance from the provider. Organizations such as the European Space Agency (ESA) are providing funding for projects to support the provision of health care services in rural and remote areas (see Figure 1).

The most common applications of telemedicine are telecardiology, teledentistry, teledermatology, telepathology, telepsychiatry, and telesurgery.

Telecardiology

Telecardiology, along with ECG interpretation service, aids physicians by providing them with instant access to cardiac assessment. Besides direct telephone access to cardiologists, general practitioners are prepared with hand-held, automatic standard 12-lead electrocardiogram ECG transmitters. These 12- ECG transmitters allow for online cardiac consultations and ECG interpretation. A full medical report including ECG signals are sent out by the general practitioners to the cardiologists. In turn, the respective cardiologists respond to the report with their consultation to the general practitioners.

Teledentistry

Teledentistry benefits patients in remote locations by allowing them to get specialized dental consultation over a network. Their dental information is electronically sent and reviewed

Figure 1. Visiting patients in their homes through telemedicine (Source: ESA, http://www.esa.int/esaCP/SEMMT0M26WD_index_0.html)

by dental specialists. Teledentistry encompasses real time and offline dental care which includes diagnosis, treatment planning, consulting and follow-up.

Teledermatology

Teledermatology is the process of providing patients situated at remote locations with dermatology consultations using information technology mechanisms. Telemedicine is exceptionally valuable to teledermatology since it is visual in nature, and health practitioners other than dermatologists are poor at diagnosing skin diseases. Studies have shown that 20% of general practitioners are not able to diagnose twenty of the most common dermatological problems.

Telepathology

Telepathology is the transmission of digitalized histological or macroscopic images between remote locations. It is used for diagnostic, prognostic, quality control, research, and educational purposes. An example of a typical telepathological structure would consist of a CCD or digital camera connected to a microscope with a computer having a good graphics card and software to control the images over a network. Telepathology is an important contribution to the health care industry as it allows for faster diagnosis and consultations by pathologists located at remote places.

Telepsychiatry

Telepsychiatry provides specialized psychiatry care or support from remote settings. Patients, physicians, and specialists communicate with each other by phone, fax, e-mail, the Internet, still imaging, and live interactive two-way audio-video conferences. Video conferencing is primarily used for clinical consultative sessions. Telepsychiatry services include assessments, diagnosis, treatment, psychological testing, medico legal assessment, case conferencing and management, education, supervision, support, administration, and research. However, patient's privacy and confidentiality of communication are vital concerns in telepsychiatry. Telepsychiatry sessions are making sincere efforts to maintain the same standards as those followed by face-to-face consultations to protect patient information.

Telesurgery

Telesurgery is the most interesting application of telemedicine. Surgeons perform micro-surgery by manipulating the hands of a robot (Angood, 2001). Over 2,000 brain surgeries performed by telesurgery have been successful. Currently, the main problem is that the robotic tools do not let surgeons feel patients' tissues. Researchers at the Biorobotics Laboratory at Harvard are conducting further research to enhance this issue. Sensors enabled to send three-dimensional information to tiny pins on the surgeon's fingertips are being designed to let the doctors feel changes in texture or the strength of his or her grip. This technology is being developed to be used as a medium to detect lung tumors or to insert needles into delicate tissues.

Telecardiology, teledentistry, teledermatology, telepathology, telepsychiatry, and telesurgery are only a few of the extensive list of applications of telemedicine. However, there are several challenges of telemedicine that it needs to overcome before being utilized to its full potential.

TELEMEDICINE CHALLENGES

There are many issues involved with telemedicine that must be addressed before it can be utilized or applied to its full potential. Some of these issues include: licensure of those that provide the service over state/country lines, insurance payment issues, privacy and security, cost and accessibility, and industry-wide standards, especially relating to safety and liability (Kantor, 1997). Organizations such as the Sandia National Laboratories (www.sandia.gov) are developing secure online telemedicine techniques. Figure 2 shows Sandia's Linda Gallagher checking her blood oxygenation and pressure with sensors connected to a state-of-the-art unit from TelAssist Corp.

In general the challenges of telemedicine can be broadly categorized into the two following sections.

Figure 2. Use of online techniques at Sandia National Lab for checking blood oxygenation and pressure (Source: http://www.sandia.gov/media/NewsRel/NR1999/telemed.htm)

Telecommunications Challenges

In rural and underdeveloped areas, where telemedicine is most needed, telecommunication technology is not up-to-date. Standard telephone lines do not provide the bandwidth necessary for many telemedicine projects.

In addition, setting up telecommunication channels is an expensive mission. Proponents of telemedicine are still not sure whether they can afford such a huge investment solely because cost of telemedicine projects is not yet accurately justified (Huston & Huston, 2000). Even though a large number of prison-based teleconsultations cases have been able to show cost savings, this data could not be used for reimbursement purposes considering its exclusive security and transportation expenses.

Socioeconomic Challenges

Legal issues regarding physician licensing, liability, and patient confidentiality exist. As physicians are licensed by states, there is a legal problem when a physician consults across state lines. It is necessary for states to engage in interstate provisions of service in order to fully benefit from telemedicine. Currently, interstate agreements vary greatly. Several states maintain that physicians must be licensed in both the sending and receiving states. Other states have entered reciprocity agreements with neighbors.

Liability is an obstacle in providing telemedicine. There is a debate related to which physician would be liable for a poor patient outcome: the primary care or the consulting physician. In the case of a poor outcome, it is not clear if the patient should file suit in the residing state or in the state the practitioner is located.

Medicaid covers telemedicine consultation in only 10 states. Medicare will reimburse for telemedicine services provided in rural counties that are designated as health professional shortage areas. This Medicare provision, authored by North Dakota Senator Kent Conrad, was part of the Balanced Budget Act of 1997 (Moreno 1998). Most commercial payers do not cover routine telemedicine consultation. Physician reluctance and patient apprehension are also obstacles. Some rural physicians fear the loss of patients to urban facilities.

The public and physicians worry about the impersonality of telemedicine as well. Differences in resources available, language, and literacy together with cultural differences in acceptability of telemedicine are other serious obstacles telemedicine needs to overcome.

TELEMEDICINE TRENDS

The full scope of telemedicine is currently being defined, but its future will be driven primarily by the new economic imperatives which are dramatically changing health care delivery. Telemedicine will not only allow physicians and other health care workers to take better care of patients, but will provide patients with tools to allow them to take a much more active and effective role in taking care of themselves. There have been several measures taken to solve the issues discussed. In the United States, the Telecommunications Act of 1996 established supplements to the cost of providing the necessary structure to support the technology in some rural areas. In addition, telecommunication companies are working on projects to connect the government with education, health care and business (Brown, 2002). There are scientific advancements such as "intelligent clothing," which monitors the condition of a patient's health then relays that information to medical specialist. Nevertheless, these measures are still at their development stages. The full potential of telemedicine can only be understood when all or most of the barriers of telemedicine are eliminated.

CONCLUSION

Developing a reliable delivery system has been slow, which contributed to the cautious pace of telemedicine in the early 90s. Reliability is an issue for some aspects partly due to a lack of industry standards. Although the technology appears as simple and as familiar as turning on a TV set, in fact multiple technical elements are involved, and users must be trained to operate the equipment. Telemedicine projects require broad-based planning, installation, and operational support. Legal and jurisdiction issues are also a concern for some types of telemedicine. If the diagnosis is incorrect, who is liable: the consulting doctors or the one that is present with the patient? Which state or country is responsible? Who gets paid by the insurance company and how much? These are questions that are still being considered for the future of telemedicine.

REFERENCES

Ackerman, M., Craft, R., Ferrante, F., Kratz, M., Mandil, S., & Sapci, H. (2002). Chapter 6: Telemedicine technology. *Telemedicine Journal and eHealth, 8*(1), 71-78.

Angood, P. (2001). Telemedicine, the Internet, and World Wide Web: Overview, current status, and relevance to surgeons. *World Journal of Surgery, 25*(11), 1449-1657.

Biomedical Informatics Ltd. (2003, December). *Telemedicine*. Retrieved April 15, 2004, from http://www.biohealthmatics.com/healthinformatics/telemedicine/telemed.aspx

Brown, N. (2002, May 03). About telemedicine: What is telemedicine? *Telemedicine Research Center*. Retrieved

April 15, 2004, from http://trc.telemed.org/telemedicine/primer.asp

Cooper, R. M. D. (1997, Spring). The University of Iowa Launches Telemedicine Project with Department of Corrections. *Health Connections*, 4. Retrieved April 15, 2004, from http://telemed.medicine.uiowa.edu/TRCDocs/Pubs/4HC/4hc11.html

Demiris, G. (2004). Electronic home health care: Concepts and challenges, *International Journal of Electronic Healthcare, 1*(1), 4-16.

Garshnek, V., Logan, J. S., & Hassell, L. H. (1997). The telemedicine frontier: Going the extra mile. *Space Policy, 3*(1), 37-46.

Huston, T. L., & Huston, J. L. (2000). Is Telemedicine a Practical Reality? *Communications of the ACM, 43*(6), 91-95.

Iyer, L. S., & Dey, D. (2005). Global healthcare applications: Telemedicine. *Fortune Journal of International Management, 1*(1), 57-71.

Jossi, F. (2005, February). *Telehealth.* Retrieved July 22, 2005, from http://www.healthcare-informatics.com/issues/2005/02_05/cover.htm#telehealth

Kantor, M. (1997, January 31). Telemedicine report to Congress. Retrieved April 15, 2005, from http://www.ntia.doc.gov/reports/telemed/index.htm

Mirza, K. (n.d.). LastByte – Health goes digital: Telemedicine. *Bytes for All. A Voluntary Online Initiative from South Asia*. Retrieved on April 15, 2004, from http://www.bytesforall.org/7th/lastbyte.htm

Moreno, D. (1998, April).What is Telemedicine? *The UND Center for Rural Health: A collaboration between the North Dakota State Data Center and the Center for Rural Health*. Retrieved April 25, 2005, from http://www.med.und.nodak.edu/depts/rural/pdf/whatistele.pdf

Nash, M. G., & Gremmillion, C. (2004). Globalization impacts the healthcare organization of the 21st century, Demanding new ways to market product lines successfully. *Nursing Administration Quarterly*, 28(2), 86-91.

Pedigo, T. L., & Sr, M. D. (1997, October). *Musings on telemedicine.* Retrieved February 22, 2004, from http://medicalcomputingtoday.com/0opmusetelemed.html

Raghupathi, W., & Tan, J. (2002). Strategic IT applications in health care. *Communications of the ACM, 45*(12), 56-61.

KEY TERMS

ICT (Information and Communication Technologies): Includes computers, software, peripherals, and connections that are intended to fulfill information processing and communications functions.

Internet: A computer network consisting of a worldwide network of computer networks that use the network protocols to facilitate data transmission and exchange.

ISDN (Integrated Services Digital Network): A system of digital phone connections that allows voice and data to be transmitted simultaneously across the world using end-to-end digital connectivity.

SONET (Synchronous Optical NETwork): The standard for synchronous data transmission on optical media.

Telehealth: The use of ICTs to deliver health and health care services and information over large and small distances.

Telemedicine: Derived from the Greek 'tele' meaning "at a distance" and the present word "medicine" which itself derives from the Latin "mederi" meaning"healing".

Teleradiology: A means of transmitting radiographic patient images and consultative text from one location to another with the use of ICTs.

Telesurgery: The use of medical technology such as robotics, sensory devices and imaging video that allows a surgeon to operate long distance. This technology provides doctors the full sensory experience of hands-on surgery.

Telescopic Ads on Interactive Digital Television

Verolien Cauberghe
University of Antwerp, Belgium

Patrick De Pelsmacker
University of Antwerp, Belgium

INTRODUCTION

The adoption of interactive digital television (IDTV) and related technologies such as the personal video recorder (PVR) make IDTV attractive for advertisers in terms of reach. Although IDTV leads to an increase in advertising avoidance behaviour, it also offers new advertising opportunities (Cauberghe & De Pelsmacker, 2006). One of them is the telescopic advertisement. This format consists of a "30-second TV ad with a call-to-action button with clickable content or micro sites featuring individual still screens providing additional product information" (Bellman & Varan, 2004, p. 2). When the viewer clicks on the call-to-action button, he or she leaves the linear broadcast stream to enter a dedicated advertising location (DAL). There, the viewer can navigate through the additional information, which can be structured in different layers. The purpose of this study is to investigate the impact of two aspects of the complexity of a telescopic ad by experimentally manipulating the amount of information and the level of interactivity in the DAL. Additionally, the role of time spent in the DAL is explored.

BACKGROUND

Advertising Complexity and Information Load

To keep the consumer's interest focused on a persuasive message, an appropriate level of complexity is recommended (Putrevu, Tan, & Lord, 2004). Complexity has been identified as one of the major dimensions of information load. An increase in information load generates a positive effect on information processing until a certain threshold is reached, at which the consumer will be overloaded with information. At this point the consumer will no longer consider additional information and will become confused, and it is harder to recall previous information (Lang, 2000). The effects of complexity and information load follow an inverted *U* shape, by which a moderate level of complexity leads to the most optimal advertising results (e.g., Geissler, Zinkhan, & Watson, 2006; Martin, Sherrard, & Wentzel, 2005; Wang, Chou, Su, & Tsai, 2007).

Amount of Information

The effect of the amount of information follows an inverted *U* shape (e.g., Meyer, 1998), leading to negative effects in decision quality when too much information is provided. Although an increase in the quality of the additional information decreases message complexity, increases the credibility of product information, and leads to a positive effect on decision making (Keller & Staelin, 1987), increasing the quantity of information leads to an increase in complexity and can, due to limited cognitive capacity and information overload, lead to confusion and negative evaluative effects after a certain threshold is reached (Lang, 2000). In the present study, the quantity of information in the DAL is manipulated, keeping the quality constant. Therefore, the following can be expected.

H1: A high amount of information in the DAL leads to lower brand recall and a less positive brand attitude than a low amount of information.

Level of Interactivity

Interactivity has the capability to develop feelings of flow, "an intrinsically motivated optimal enjoyable mental state" (Csikszentmihalyi & Lefevre, 1989). This mental state increases the cognitive involvement with the interactive content due to focused attention and the possibility for consumers to take control over the time, structure, and order in which they want to be exposed to the information (Liu & Shrum, 2002); it also increases the processing of the information presented in the interactive context (e.g., Chung & Zhao, 2004; Macias, 2003; Sicilia, Ruiz, & Munera, 2005). The intrinsically motivated joy evoked by flow may be transferred to the persuasive message and brand, leading to a positive effect on the attitude toward the ad, brand attitude, and pur-

chase intention (Chung & Zhao; Ko, Cho, & Roberts, 2005; Macias, 2003). Therefore, we expect the following.

H2: A high level of interactivity in the DAL leads to higher brand recall and a more positive brand attitude than a low level of interactivity.

Mediating Role of Time Spent in the DAL

Longer ads have more opportunities to provide extra product arguments and to repeat key points of the message compared to their shorter equivalents. The longer a consumer is exposed to an advertisement, the more opportunity the consumer has to process it, and the more he or she will remember of it. Longer commercials lead to more positive brand attitudes (e.g., Danaher & Mullarkey, 2003). Since both the amount of information and interactivity increase the level of complexity, we can expect that more complexity will lead to more time spent in the DAL and that, next to a direct effect of the amount of information and the level of interactivity on brand responses, part of this effect is mediated by the time spent in the DAL.

H3: Time spent in the DAL mediates the effects of the amount of information and the level of interactivity on brand recall and brand attitude.

EMPIRICAL STUDY

Research Method

The hypotheses were tested using a 2x2 (information x interactivity) between-participant factorial design. A telescopic ad was developed, consisting of a 30" television ad, a call-to-action button ("click on the red button for more information"), and an interactive DAL. To avoid confounding effects, a traditional 30" advertisement for a travel agency originating in The Netherlands was used, unknown to Dutch-speaking Belgians. The attitude toward the 30" advertisement was controlled to be positive to avoid negative affective reactions. The ad contained a feel-good conversation between two men on an airplane.

In the DAL, the amount of information was manipulated at two levels. Information about different kinds of holiday formulas and hotels was provided for different countries. The information in the high-level condition was more of the same compared to the low information level (e.g., information about 29 hotels vs. 106 hotels). Interactivity was manipulated at two levels by means of the amount of links in the DAL (12 vs. 92), the availability of a navigation bar (yes or no), and the possibility of two-way communication (e.g., "search an address," yes or no). The layout of the DAL was kept stable over conditions.

Out of a database of a Belgian market research agency, a gross sample of 521 individuals was randomly selected based on age, gender, and education quota. A net sample of 282 participants cooperated in the study. The average age of respondents in the sample was 38 (range 21-56); 61.8% were males, and 55.3% finished higher education. The respondents were randomly assigned to one of the four experimental conditions. The respondents were individually invited to an experimental living-room setting. After the briefing, they watched a 6-minute neutral-mood excerpt of a TV programme followed by the advertisement. At the end of the 30" television ad, a call-to-action button appeared on the screen in combination with a voice-over that invited the respondent to press the button for more product information. After viewing the DAL, the respondents entered a computer-assisted questionnaire. The experiment lasted 40 minutes in total. Each respondent received €25.

The questionnaire contained a brand attitude (Ab) measurement (which did not contain the brand name; seven-point four-item scale, α =.87) and a measure of the attitude toward the 30" ad (used as a covariate in the analysis; seven-point four-item scale, α =.95). After questions about the age, gender, and education level of the participants, unaided brand recall was measured using an open-ended question.

Results

Manipulation Check

The perceived amount of information (two-item five-point semantic differential scale, α =.93) and the perceived level of interactivity were measured by three-item five-point Likert scales (α =.93 and .74). The t-tests indicate that both manipulations were successful ($M_{low\ information}$=3.281 vs. $M_{high\ information}$=4.022, t=5.885, p<.001; $M_{low\ interactivity}$= 3.286 vs. $M_{high\ interactivity}$=3.644, t=6.024, p<.001).

Main Effect of Amount of Information

The amount of information in the DAL had no effect on brand recall (low amount of information=52.1% vs. high amount of information=47.9%, chi^2=.845, p=.358). To measure the effect of the amount of information on brand attitude, an ANCOVA (analysis of covariance) was conducted in which the attitude toward the 30" television ad was used as a covariate. The attitude toward the 30" television ad had a significant positive effect on Ab (F (1, 240)=65.128, p<.001). However, there was no significant main effect of the amount of information on Ab (F (1, 240)=.039, p=.843; $M_{low\ information}$=4.28 vs. $M_{high\ information}$=4.37). H1 is not supported.

Main Effect of Level of Interactivity

The level of interactivity had a positive effect on brand recall and attitude. As compared to a low level of interactivity, a high level of interactivity increased brand recall from 43.6% to 56.4% (chi^2=5.371, p=.020). This positive effect of interactivity was also found for Ab (F (1, 240)=9.502, p=.002; $M_{low\ interactivity}$=4.19 vs. $M_{high\ interactivity}$= 4.46). H2 is supported.

Mediating Role of Time Spent in the DAL

We chose not to constrain the time in our study to keep the external validity high. In real life, the viewer also has the freedom to choose how much time he or she spends to process the additional information. To investigate the mediating effect of time spent in the DAL on the relation between the amount of information, the level of interactivity, and the attitude toward the brand, we used the three-step procedure of Baron and Kenny (1986).

Brand Recall

The amount of information has no significant direct effect on brand recall (see above main effect of the amount of information; chi^2=.845, p=.358). However, to detect a mediation effect, significance in this first step is not required. The effect of the amount of information on the time spent in the DAL is positive and significant (Step 2; low amount of information: M_{time}=4.768; high amount of information: M_{time}=5.745, t=2.514, p=.013). In Step 3, both the amount of information (dichotomous variable) and the time spent in the DAL were inserted as independent variables in a logistic regression with brand recall as a dichotomous dependent variable. The results show that the time spent (p=.007) fully mediates the effect of the amount of information (p=.224) on brand recall. The main effect of the level of interactivity on brand recall is positively significant (Step 1; see above main effect of the level of interactivity; chi^2=5.371, p=.020). The level of interactivity also significantly increases the time spent in the DAL ($M_{low\ interactivity}$=4.373 vs. $M_{high\ interactivity}$=6.148, t=4.710, p<.001). The logistic regression indicates that the effect of the level of interactivity on brand recall is only partially mediated by the time spent in the DAL (p=.047). Interactivity is still (marginally) significant (p=.08).

Brand Attitude

The amount of information has no significant effect on brand attitude (Step 1; see effect of the amount of information, $M_{low\ information}$=4.279 vs. $M_{high\ information}$=4.366, t=.698, p=.486). The amount of information has a positively significant effect on time spent in the DAL (Step 2; $M_{low\ information}$=4.767 vs. $M_{high\ information}$=5.745, t=2.514, p=.013). ANCOVA was used to investigate the combined effect of the amount of information (dichotomous variable) and the time spent in the DAL (Step 3). The effect of the amount of information on brand attitude stays insignificant (F(1, 2.13), p=.645), and time spent had a positive, significant effect (F(1, 4.612), p=.033). These results indicate that the effect of the amount of information on brand attitude is fully mediated by the time spent in the DAL. The level of interactivity has a positive, significant effect on brand attitude (Step 1; see above main effect of the amount of information, $M_{low\ interactivity}$= 4.187 vs. $M_{high\ interactivity}$=4.457, t=2.188, p=.030). Next, the level of interactivity significantly increases the time spent in the DAL ($M_{low\ information}$=4.373 vs. $M_{high\ information}$=6.148, t=4.710, p<.001). The ANCOVA (Step 3) illustrates that the effect of interactivity on brand attitude is no longer significant (F(1, 2.093), p=.149) when integrating the time spent in the DAL (F(1, 2.769), p=.097) as a covariate. However, the significance levels of the effects of both variables (interactivity and time) on brand attitude are rather weak. This indicates that time spent in the DAL fully but weakly mediates the effect of interactivity on brand attitude.

H3 is supported. Both the effects of the level of interactivity and the amount of information on brand recall and brand attitude are at least partly mediated by the time spent in the DAL.

FUTURE TRENDS

The main conclusion of this study is that the positive effect of interactivity in the DAL is remarkably noticeable on both brand recall and attitude. Since the amount of information did not have a negative effect in any way, advertisers may be advised to provide their DAL with enough information. High amounts of information will increase the time spent in the DAL, and will therefore indirectly have a positive effect on brand attitude and recall.

A telescopic ad fits into a number of new advertising and marketing trends (see also Rappaport, 2007). A telescopic ad can be used on demand given the technical possibility to store the DAL on the television. This allows the viewer to ask for product information whenever convenient for him or her and fits into the trend of consumer-controlled two-way communication. A telescopic ad can also be the starting point to increase engagement with the brand. All consumer data (clicking behaviour, time spent, e-mail, etc.) can be stored and used to develop long-lasting relationships between the customer and the company. In this way, it fits into the trend of shifting the focus from transaction-oriented marketing (push) to relationship marketing. The telescopic ad can also be used as an extra service for the consumer. This is particularly relevant for product categories that are at least moderately involving, such as banking, insurances,

and durables. In mature and saturated markets, brands will have to be full-scale solution providers rather than merely offering functional product benefits. Telescopic ads can also tap into the experiential marketing trend: Brands should be experienced, not merely communicated. Telescopic ads can offer the extra experiential context that a modern brand needs.

Our study provides some directions for research that would provide guidance with respect to what successful telescopic ads should look like in the future. First, our results are only pertinent when the respondent reacts to the call for action and uses the interactive possibilities. A real-life study is necessary to explore the actual click-through behaviour more validly. Further research could investigate under which circumstances a viewer will react to the call to action that appears in the 30" ad. Further research could examine the role of the product message or offer in the ad, the creative style of the ad, the impact of the congruency between the ad message and the information provided in the DAL (in terms of expectancy, relevance, humour, etc.), and so on. Only the amount of information and the level of interactivity were manipulated in this study. Other aspects of the DAL, such as the quality of the information, the vividness (color, animation, graphics), or the occurrence of a simultaneous media context, could also be investigated.

CONCLUSION

The cognitive involvement and the intrinsic enjoyable experience evoked by the interactivity may explain these results, which are in line with earlier Internet studies (e.g., Sicilia et al., 2005). A higher level of interactivity presumably makes the DAL moderately complex without demanding too much cognitive resources to process it, and therefore leads to positive results. The amount of information had no negative effect on brand recall or on brand attitude. The level of interactivity had a positive effect on both brand recall and attitude. Therefore, this study could not confirm the expectations based on the complexity theory and the cognitive-load theory, which states that too much information causes the respondents to feel lost or overwhelmed, leading them to lose focus and interest rapidly.

There are several possible explanations for not finding the information overload effect. The effect of the amount of information appears to be fully mediated by the time spent in the DAL. Also, the effect of the level of interactivity is partially mediated by the time spent in the DAL. Consequently, if the amount of information and the level of interactivity increase, the respondent will also spend more time processing that interactive information. This may offer an explanation as to why the overload phenomenon did not occur. A second explanation can be offered based on the information processing insight that a consumer can experience information overload in decision making processing, but will not necessarily do so. Maybe in this study the information overload, and thus the negative advertising results, did not occur because the individuals did not allow it to happen, only processing the amount of information that they desired. A third possible explanation is based on the premise that when individuals get familiar and experienced with interactivity, the cognitive load induced by it may be reduced. As such, experienced Web users have less difficulty with processing interactive content than less experienced ones. The participants in our study, like many people nowadays, were probably used to frequently engaging in interactive information processing (e.g., on the Internet, third-generation mobile phones, ATMs, etc.), and therefore did not experience overload.

In conclusion, the study illustrates that the negative impact on brand responses of information overload induced by a complex advertising stimulus (manipulated by the amount of information and level of interactivity) is not likely to occur in a real-life interactive context. The interactivity does increase user control, and the user can decide on the sequence and how much time he or she spends processing the information, therefore managing the information load him- or herself. Our results show that interactivity can increase advertising effectiveness regardless of the amount of information provided. Moreover, the longer the viewer stays in the DAL (as a result of interaction possibilities and/or extra information), the more positive the advertising effects become.

REFERENCES

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bellman, S., & Varan, D. (2004). *The impact of adding additional information to television advertising on elaboration, recall, persuasion.* Paper presented at the ANZMAC Conference, Wellington, New Zealand.

Cauberghe, V., & De Pelsmacker, P. (2006). Opportunities and thresholds for advertising on interactive, digital TV: A view from advertising professionals. *Journal of Interactive Advertising, 7*(1). Retrieved from http://www.jiad.org/vol7/no1/cauberghe/index.htm

Chung, H., & Zhao, X. (2004). Effects of perceived interactivity on Web site preference and memory: Role of personal motivation. *Journal of Computer-Mediated Communication, 10*(1). Retrieved from http://jcmc.indiana.edu

Csikszentmihalyi, M., & Lefevre, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology, 56*(6), 815-822.

Danaher, P., & Mullarkey, G. (2003). Factors affecting online advertising recall: A study of students. *Journal of Advertising Research*, pp. 252-267.

Geissler, G. L., Zinkhan, G. M., & Watson, R. T. (2006). The influence of home page complexity on consumer attention, attitudes and purchase intent. *Journal of Advertising, 35*, 69-80.

Keller, D. K., & Staelin, R. (1987). Effects of quality and quantity of information on decision effectiveness. *Journal of Consumer Research, 14*, 200.

Ko, H., Cho, C. H., & Roberts, M. S. (2005). Internet uses and gratifications: A structural equation model of interactive advertising. *Journal of Advertising, 34*, 57-70.

Lang, A. (2000). The limited capacity model of mediated message processing. *Journal of Communication, 50*(1), 46-70.

Liu, Y., & Shrum, L. J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person and situation for the influence of interactivity on advertising effectiveness. *Journal of Advertising, 21*(4), 53-64.

Macias, W. (2003). A beginning look at the effects of interactivity, product involvement and Web experience on comprehension: Brand Websites as interactive advertising. *Journal of Current Issues and Research in Advertising, 25*(2), 31-44.

Martin, B. A., Sherrard, M. J., & Wentzel, D. (2005). The role of sensation seeking and need for cognition on Web-site evaluations: A resource-matching perspective. *Psychology & Marketing, 22*(2), 109-126.

Meyer, J. (1998). Information overload in marketing management. *Marketing Intelligence & Planning, 16*, 200-209.

Putrevu, S., Tan, J., & Lord, K. R. (2004). Consumer responses to complex advertisements: The moderating role of need for cognition, knowledge and gender. *Journal of Current Issues and Research in Advertising, 26*(1), 9-24.

Rappaport, S. D. (2007). Lessons from online practice: New advertising models. *Journal of Advertising Research, 47*, 135-141.

Sicilia, M., Ruiz, S., & Munera, J. L. (2005). Effects of interactivity in a Website. *Journal of Advertising, 34*(3), 31-45.

Wang, K. C., Chou, S. H., Su, S. J., & Tsai, H. Y. (2007). More information, stronger effectiveness? Different group package tour advertising component on Web page. *Journal of Business Research, 60*, 382-387.

KEY TERMS

Cognitive Overload: Cognitive overload occurs when the volume of information supply exceeds the information processing capacity of the individual.

Dedicated Advertising Location (DAL): A DAL contains clickable content featuring individual still screens providing additional product information.

Flow: Flow is an intrinsically motivated, optimal enjoyable mental state leading to increased attention, cognitive involvement, and a feeling of fun and escapism.

Interactive Digital Television (IDTV): IDTV is the merging of the Internet and television.

Interactivity: It is human-to-human interaction that places the emphasis on two-way communication, mutual discourse, feedback, and so forth, and message-to-human interaction, which is related to aspects of content such as user choice, user (information) control, structure, and so on.

Stimulus Complexity: This refers to the amount of variety and diversity in a stimulus pattern.

Telescopic Ad: It is a 30-second TV ad with a call-to-action button and clickable content or microsites featuring individual still screens providing additional product information.

Testing Graphical User Interfaces

Jaymie Strecker
University of Maryland, USA

Atif M Memon
University of Maryland, USA

INTRODUCTION

In recent years, an emerging trend in software products has been toward the use of graphical user interfaces (GUIs). More user-friendly than traditional, text-based interfaces, GUIs serve as the front-end for a large portion of today's software applications. Technologies like Ajax are helping to spread familiar GUI interaction styles to Web applications. With the rise of ubiquitous computing, users are interacting with GUIs in a widening range of situations—not just with their PCs, but with their dishwashers and cars. Critical applications, such as banking systems, are moving to GUIs as well. Thus, quality assurance for GUI-based software is growing more important every day.

With GUIs, users enjoy many degrees of freedom in the way they interact with the software. While this benefits users, it challenges testers. Because users may interact with a GUI in a variety of unexpected ways, it is difficult to insure that the software meets its functional requirements (correctness) and non-functional requirements (e.g., usability) for all possible interactions. The difficulties are compounded by the frequent intersection of GUIs with other emerging technologies, including component-based and service-oriented architectures. New trends in software development, such as rapid development cycles, globally distributed developers, and open-source projects, make the quality assurance process ever more challenging.

This chapter describes the state of the art in testing GUI-based software. Traditionally, GUI testing has been performed manually or semimanually, with the aid of capture-replay tools. Since this process may be too slow and ineffective to meet the demands of today's developers and users, recent research in GUI testing has pushed toward automation. Model-based approaches are being used to generate and execute test cases, implement test oracles, and perform regression testing of GUIs automatically. This chapter shows how research to date has addressed the difficulties of testing GUIs in today's rapidly evolving technological world, and it points to the many challenges that lie ahead.

BACKGROUND

A GUI provides a visual front-end through which a user can interact with a software application. Although there are various models for GUI design, the most commonly used in practice and in software-testing research—and hence the model assumed in this chapter—is the WIMP model with *w*indows, *i*cons, *m*enus, and *p*ointing devices (Nielsen, 1993). The GUI is made up of *widgets*—such as buttons, text boxes, and labels—that the user can manipulate to send input to the underlying software and the software can, in turn, manipulate to send output to the user. Each widget has a set of *properties*—for example, "font", "width", "enabled"—each of which has some *value*—for example, "Helvetica", "100", "true" (Yuan & Memon, 2007).

Widgets are contained in *windows*, which may either be *modal* or *modeless*. A modal window blocks the user's interaction with other windows while it is active, whereas a modeless window imposes no such restrictions. A *window's state* at any particular time is the set of all triples (*w*,*p*,*v*) such that *w* is a widget in the window, *p* is a property of *w*, and *v* is the value of *p*. The *GUI state* then consists of the state of all windows in the GUI (Yuan & Memon, 2007).

As the user interacts with the GUI, the state of both the GUI and the underlying software can change. When the user performs an *event* on the GUI—such as clicking a button or typing in a text box—a piece of application code called an *event handler* is executed. The event is the basic unit of interaction with a GUI. To accomplish a task, a user typically must perform multiple events in sequence. Hence, a *GUI test case* consists of a sequence of events (Yuan & Memon, 2007).

Several tools and techniques are available to aid testing of GUI-based applications, varying greatly in the level of automation they provide. Ignoring the GUI altogether, test harnesses like JUnit can interact directly with the underlying software much like the GUI would. However, this may require major changes to the GUI's architecture, and, at any rate, it leaves an important part of the end-user software untested.

JUnit has been extended in tools such as JFCUnit, Pounder, and Jemmy Module to interact with the application under test through its GUI. With these tools, test cases must be written manually. Alternatively, a tester can generate test cases by recording sequences of events, which the tester manually performs on the GUI, using a capture-replay tool. Some capture-replay tools—for example, CAPBAK and TestWorks—record events in terms of mouse coordinates, while others—for example, WinRunner, Abbot, and Rational Robot—record the GUI widgets associated with events. The latter are more robust in the face of superficial changes to the GUI layout (Memon & Xie, 2005).

All of the tools and techniques mentioned so far automate the execution of test cases but still require substantial effort on the part of the tester to generate test cases, define the test oracle, and modify the test suite as the application under test evolves. Tools like the visual test-development environment created by Ostrand, Anodide, Foster, and Goradia (1998) streamline the testing process but do not depart from the conceptualization of GUI testing as a fundamentally manual process. Similarly, while Kasik and George (1996) have shown how genetic algorithms can be used to augment a test suite, they leave much work to the tester. Fortunately, new techniques based on various types of models of the GUI are shifting much of the burden of the testing process from humans to machines.

The most popular type of GUI model, the state-machine model, makes it possible to generate test cases—or perform model-checking, a related activity—automatically (Belli, 2001; Berstel, Reghizzi, Roussel, & Pietro, 2005; Dwyer, Carr, & Hines, 1997; Holzmann & Smith, 1999; Shehady & Siewiorek, 1997; White & Almezen, 2000). But techniques based on state-machine models have serious drawbacks. These techniques require that the model be created manually, that a formal specification be written, or that the source code be annotated—in any case, a potentially laborious task susceptible to human error. Further, since the effectiveness of the test cases generated from the state-machine model depends on the model creator's definition of "state", two testers testing the same application may get quite different results (Yuan & Memon, 2007). Techniques for generating test cases from UML diagrams suffer from similar weaknesses (Vieira, Leduc, Hasling, Subramanyan, & Kazmeier, 2006).

Rather than modeling a GUI in terms of states, others have modeled it in terms of events. Memon, Pollack, and Soffa (2001) have used automated planning to generate test cases that consist of sequences of events chosen to accomplish tasks specified by the tester. In this approach, model creation requires substantial human effort: although the events in the model are identified automatically, their preconditions and effects must be defined manually. More recently, techniques have used event-based models to further reduce the amount of effort required in the testing process while improving its effectiveness. These are described in the next section.

GUI TESTING WITH EVENT-FLOW MODELS

Events are central to the dynamic structure of a GUI-based application. A user accomplishes tasks via the GUI by performing sequences of events. Thus, the execution of the application occurs as the execution of a sequence of event handlers, each of which may depend on and may also affect the state of the application. Users may interact with the application in unexpected ways, so the event handlers may be executed in unexpected orderings. In these respects, GUI-based applications differ from traditional, batch-style software (e.g., compilers), which receives some input, processes it, produces some output, and terminates. Traditional testing techniques like code-based coverage criteria that were designed for such software may not work as well for much differently-structured GUI-based applications, so new techniques have been developed to address GUIs' event-driven nature (Memon, 2002).

The previous section showed how GUI-testing tools and techniques have evolved to be faster and more effective. Notable advances have been achieved through model-based testing, using various types of models. In recent years, one type of model has proved particularly successful: the *event-flow graph*.

Event-Flow Graph

In an event-flow graph, a GUI is represented by a graph whose vertices represent events and whose edges represent the *follows* relationship. Event e_1 is said to *follow* event e_2 if e_1 can be executed immediately after e_2, with no events intervening. Test cases can be generated rapidly and automatically by traversing the EFG, and coverage criteria can be defined in terms of the EFG. Variations of the EFG have been used to further improve the cost-effectiveness of GUI testing. Each of these topics will be elaborated upon after the process of creating an EFG is explained (Xie & Memon, 2006).

An EFG can be reverse engineered semi-automatically from a GUI in a process called *GUI ripping*. A single GUI window is ripped by identifying and recording properties of all of the widgets it contains, then executing any events available in the window that open new windows. This can be accomplished by running the GUI with reflection to access the currently open windows and inspect their widgets. Widgets likely to open new windows can be identified based on conventions in GUI design: clicking on a widget whose caption ends in "..." typically opens a window. As

new windows are discovered, each is ripped until no more new windows are found. Since the GUI may not be ripped perfectly—indeed, the GUI itself may contain defects that show up in the resulting GUI model—the tester must examine and possibly edit the model. The GUI model provides the information necessary to determine the *follows* relationships for all of the events and, hence, to construct the EFG (Memon, Banerjee & Nagarajan, 2003).

Any path through the EFG starting at an event that is available when the application under test is launched can serve as a test case. However, the number of possible test cases grows exponentially with the length of the path. Short test cases can find some faults, but many faults can only be detected with longer event sequences (Yuan & Memon, 2007). Thus, one challenge of GUI testing is to identify which of the longer sequences are most likely to add value to testing. Several variations of the EFG have been created to address this challenge.

Variations of the Event-Flow Graph

An important observation about the EFG is that it contains many events that are unlikely to be defective. The handler for an event that opens a menu, for example, is almost always located in library code and does not interact with the application code in any way. This observation leads to a variation of the EFG called the *event interaction graph (EIG)* that achieves marked size reduction without sacrificing fault-detection potential by omitting events that need not be tested as rigorously as the rest. The vertices in the EIG represent *system-interaction events*—events that either close windows or perform actions without opening or closing any windows or menus. Examples of system-interaction events include clicking the "OK" button in a preferences window and using the "copy" event to copy objects to the clipboard. Edges in the EIG represent the *interacts-with* relationship. A system-interaction event e_1 is said to *interact with* a system-interaction event e_2 if there is a path from e_1 to e_2 in the EFG that contains no system-interaction events other than e_1 and e_2. Test cases can be generated by traversing the EIG to get sequences of system-interaction events. To make the test cases executable on the GUI, they must be mapped onto the EFG to fill in the necessary nonsystem-interaction events before and between system-interaction events (Memon & Xie, 2005).

Even with its substantial size reduction compared to the EFG, the EIG can be unwieldy for large applications. Testing all length-*n* event sequences in the EIG is only feasible for the smallest values of *n*, often just 2. But many of the longer event sequences can safely be skipped during testing—namely, sequences in which the event handlers do not affect each other. If the handlers of events e_1 and e_2 do not interact, then executing e_1 and e_2 in sequence will not reveal any faults that could not be revealed by executing each of e_1 and e_2 by itself. The *event semantic interaction graph (ESIG)* is a subgraph of the EIG that omits edges between unrelated events. In the ESIG, an edge from event e_1 to e_2 means that there is an *event semantic interaction* relationship between e_1 and e_2, or, in other words, executing e_1 followed by e_2 results in a GUI state that is in some sense different from the state that would have resulted if e_1 and e_2 had been executed in isolation. An example of an event semantic interaction in a word processor occurs between the events "select all" and "delete": performing delete just after select all modifies the text on the screen differently than executing either select all or delete does. But select all would not be expected to semantically interact with, say, "change page orientation". The event semantic interaction relationship is defined in terms of dynamic GUI state, rather than static source-code properties such as variables shared by event handlers, because pervasive design elements in GUI software—such as multiple languages, callbacks for event handlers, multi-threading, and the use of libraries—limit the applicability of static analysis. The ESIG is constructed automatically by running an initial test suite that covers all edges in the EIG and analyzing the states into which each test case drives the GUI. As with the EIG, test cases can be generated from the ESIG by traversing the graph and filling in any necessary nonsystem-interaction events (Yuan & Memon, 2007).

Another variation of the EFG, the *probabilistic event flow graph (PEFG)*, has been used to focus testing on the event sequences that users are most likely to perform. The PEFG consists of an EFG in which paths are annotated with probabilities indicating the likelihood that a user would follow that path. These probabilities come from *usage profiles* (also called *operational profiles* or *session data*), which are event sequences captured automatically as users interact with the program. Although the usage profiles could simply be replayed on the application (as in capture-replay tools), test cases generated by traversing high-probability paths in the PEFG offer some advantages. First, because the PEFG represents a composite of different usage patterns, some high-probability paths, while not executed by any individual user during usage-profile collection, are likely to be executed by future users. Testing these paths, then, can reveal faults that future users would be likely to encounter. Second, generating test cases from the PEFG is more flexible in the face of changes to the GUI than replaying users' interactions verbatim (Brooks & Memon, 2007).

Running Test Cases

Test cases generated from any of these models—the EFG, the EIG, the ESIG, or the PEFG—can be executed automatically. Much like in GUI ripping, the application under test is run under reflection, enabling the widgets specified in the test cases to be identified and manipulated to perform

events. The resulting output can be recorded and checked (run through the oracle procedure) automatically (Xie & Memon, 2007).

As mentioned previously, the way GUIs communicate information to users departs from the traditional batch style. GUI output is complex: rather than a number or text string, it consists of the property values of all of the widgets the user can see, including their structure and position relative to one another. Each event the user performs can drive the GUI into a new state, producing new output in the form of changes to the GUI. Even if the final GUI state resulting from an event sequence is correct, those at intermediate steps may not be. Recording and checking the entire GUI state after each intermediate event, while effective at detecting faults, costs time and space. The cost of applying the test oracle can be reduced by collecting less *oracle information* or relaxing the *oracle procedure*, at the price of reduced fault detection. The amount of oracle information—property values of widgets in the GUI—that must be collected and checked each time the GUI state is captured can be reduced by omitting widgets outside the currently active window or even outside the most recently manipulated widget. The oracle procedure—which compares the actual oracle information to the expected behavior—may be called less frequently—for example, only after the last event in the test case. Although less stringent oracles may be cheaper, thorough oracles can be more cost-effective. Moreover, stricter oracles can make up for a shortage of longer test cases (Xie & Memon, 2007).

Another factor that affects the cost-effectiveness of GUI testing is the coverage criterion (or test-adequacy criterion) that defines when the GUI has been tested enough. Coverage criteria have been defined in terms of the event-flow models described above—for example, "all events", "all EIG edges", or "all length-*n* event sequences in the EFG". Event-based coverage criteria are more closely tied than code-based coverage criteria to the way the application is used and the way its components interact. Model-based approaches to GUI testing not only provide a way to define coverage criteria, but, by enabling many testing tasks to be automated, also make it feasible to satisfy those criteria (Memon, 2002; Memon, Soffa & Pollack, 2001).

Regression Testing

Of course, a testing technique does not just need to be successful the first time around—it must be cost-effective throughout the life-cycle of the software, across many iterations of regression testing. This is especially critical given the rise of rapid development cycles and multiple, geographically-distributed developers and maintainers. Enabling testers to rapidly construct a GUI model and generate test cases from it, event-flow models support the creation of a new, disposable test suite for each revision of the software. In addition, testers may want to create a more permanent set of test cases for regression testing. Although minor changes to the GUI tend to break test cases generated from the EFG, test cases generated from the EIG have proved to be more robust, withstanding changes like moving events from one window to another and changing the structure of menus (Memon & Xie, 2005; Xie & Memon, 2006).

FUTURE TRENDS

As GUI testing is integrated into development processes, we expect that different styles of GUI testing will be adapted to different parts of the processes. Xie and Memon (2006) have proposed three concentric loops of testing that vary in speed and thoroughness. For rapid, fully automated testing, which may be performed each time a code revision is committed, a few EIG edges are covered and the oracle procedure simply checks for crashes; this technique is called *crash testing*. Over multiple iterations of crash testing, different EIG edges are covered, so that eventually all are tested. In *smoke testing*—a slower, more comprehensive form of testing that may be performed in conjunction with a nightly build—test cases covering all EFG vertices or edges are generated and used to reference test the current version of the software against an earlier version. The most labor-intensive forms of GUI validation—such as manually-created test cases and state-model-based techniques—may be reserved for release testing.

GUIs belong to the larger class of event-driven (or reactive) software—a class of software that challenges testers in many ways. The GUI-testing techniques described here may in the future be generalized to other kinds of event-driven software. This may include component-based applications, object- and service-oriented applications, network protocols, and Web applications, as well as nonWIMP GUIs. The event-flow models of GUI usage will have to be extended to handle complications like event timing, multiple users or input streams, and relaxed assumptions about turn-taking between users and computers (Memon, 2004; Nielsen, 1993).

CONCLUSION

As GUIs become more pervasive and more is demanded of them, the importance of their quality assurance is growing. The need to incorporate GUI testing into testing processes throughout the software life-cycle is becoming apparent. Over the past several years, advances in model-based GUI testing have made this more cost-effective by providing test-adequacy criteria and by automating test-case generation, test execution, test oracles, and regression testing. In the future, GUI testing may commonly be woven into the testing process, with different levels of testing designed to

meet goals at different time scales. Lessons learned from GUI testing will likely be applied to the testing of other kinds of event-driven software.

ACKNOWLEDGMENT

This work was partially supported by the US National Science Foundation under NSF grant CCF-0447864 and the Office of Naval Research grant N00014-05-1-0421.

REFERENCES

Belli, F. (2001). Finite state testing and analysis of graphical user interfaces. In *Proceedings of the 12th International Symposium on Software Reliability* (pp. 34-43).

Berstel, J., Reghizzi, S. C., Roussel, G., & Pietro, P. S. (2005). A scalable formal method for design and automatic checking of user interfaces. *ACM Transactions on Software Engineering and Methodology, 14*(2), 124-167.

Brooks, P. & Memon, A. M. (2007). Automated GUI testing guided by usage profiles. In *Proceedings of the 22nd IEEE International Conference on Automated Software Engineering*.

Dwyer, M. B., Carr, V., & Hines, L. (1997). Model checking graphical user interfaces using abstractions. In *Proceedings of the 6th European Software Engineering Conferences held jointly with the 5th ACM SIGSOFT International Symposium on Foundations of Software Engineering* (pp. 244-261).

Holzmann, G. J. & Smith, M. H. (1999). A practical method for verifying event-driven software. In *Proceedings of the 21st International Conference on Software Engineering* (pp. 597-607).

Kasik, D. J. & George, H. G. (1996). Toward automatic generation of novice user test scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground* (pp. 244-251).

Memon, A. M. (2002). GUI testing: Pitfalls and process. *Computer, 35*(8), 87-88.

Memon, A. M. (2004). Developing testing techniques for event-driven pervasive computing applications. In *Proceedings of the OOPSLA 2004 Workshop on Building Software for Pervasive Computing*.

Memon, A. M., Banerjee, I., & Nagarajan, A. (2003). GUI ripping: Reverse engineering of graphical user interfaces for testing. In *Proceedings of the 10th Working Conference on Reverse Engineering* (pp. 260-269).

Memon, A. M., Pollack, M. E., & Soffa, M. L. (2001). Hierarchical GUI test case generation using automated planning. *IEEE Transactions on Software Engineering, 27*(2), 144-155.

Memon, A. M., Soffa, M. L., & Pollack, M. E. (2001). Coverage criteria for GUI testing. In *Proceedings of the 8th European Software Engineering Conference held jointly with the 9th ACM SIGSOFT International Symposium on Foundations of Software Engineering* (pp. 256-267).

Memon, A. M. & Xie. Q. (2005). Studying the fault-detection effectiveness of GUI test cases for rapidly evolving software. *IEEE Transactions on Software Engineering, 31*(10), 884-896.

Nielsen, J. (1993). Noncommand user interfaces. *Communications of the ACM, 36*(4), 83-99.

Ostrand, T., Anodide, A., Foster, H., & Goradia, T. (1998). A visual test development environment for GUI systems. In *Proceedings of the 1998 ACM SIGSOFT International Symposium on Software Testing and Analysis* (pp. 82-92).

Shehady, R. K. & Siewiorek, D. P. (1997). A method to automate user interface testing using variable finite state machines. In *Proceedings of the 27th International Symposium on Fault-Tolerant Computing* (pp. 80-88).

Vieira, M., Leduc, J., Hasling, B., Subramanyan, R., & Kazmeier, J. (2006). Automation of GUI testing using a model-driven approach. In *Proceedings of the 2006 International Workshop on Automation of Software Test* (pp. 9-14).

White, L. & Almezen, H. (2000). Generating test cases for GUI responsibilities using complete interaction sequences. In *Proceedings of the 11th International Symposium on Software Reliability Engineering* (pp. 110-121).

Xie, Q. & Memon, A. M. (2006). Model-based testing of community-driven open-source GUI applications. In *Proceedings of the 22nd IEEE International Conference on Software Maintenance* (pp. 145-154).

Xie, Q. & Memon, A. M. (2007). Designing and comparing automated test oracles for GUI-based software applications. *ACM Transactions on Software Engineering and Methodology, 16*(1).

Yuan, X. & Memon, A. M. (2007). Using GUI run-time state as feedback to generate test cases. In *Proceedings of the 29th International Conference on Software Engineering* (pp. 396-405).

KEY TERMS

Graphical User Interface (GUI): A visual front-end through which a user can interact with a software application.

GUI State: The collection of states of all windows in the GUI, where a window's state is the set of all triples (w,p,v) such that w is a widget in the window, p is a property of w, and v is the value of p.

Event: The basic unit of input to a GUI, triggered by such user actions as clicking a button or typing in a text box.

GUI Test Case: A sequence of events to be performed on the GUI.

Event Handler: A piece of application code that executes in response to an event.

System-Interaction Event: An event that either closes a window or performs some action without opening or closing any windows or menus.

Event-Flow Graph (EFG): A graph representation of a GUI in which vertices represent events and an edge from event e_1 to event e_2 signifies that e_2 can be performed immediately after e_1.

Event-Interaction Graph (EIG): A graph representation of a GUI in which vertices represent system-interaction events and an edge from event e_1 to event e_2 signifies that there is a path from e_1 to e_2 in the EFG that contains no system-interaction events other than e_1 and e_2.

Event Semantic Interaction Graph (ESIG): A graph representation of a GUI in which vertices represent system-interaction events and an edge from event e_1 to event e_2 signifies that performing e_1 followed by e_2 results in a GUI state that is qualitatively different from the state that would have resulted had e_1 and e_2 been performed in isolation.

Probabilistic Event-Flow Graph (PEFG): A graph representation of a GUI that consists of an EFG whose paths are annotated with probabilities of traversal by users.

Third Places in the Blackosphere

C. Frank Igwe
The Pennsylvania State University, USA

INTRODUCTION

Although times change, there are certain human elements that survive through the ages. These elements include the need for expression, companionship, involvement, connection, and information. The avenues by which humans engage in these social practices have evolved, and with the dawn of the Information Age we are seeing the emergence of new forms of computer mediated communication (CMC), with Weblogs (or blogs) being a manifestation of this transformation. This chapter deals with these Information and Communicative Technologies (ICT), and more specifically, how blogs are being used by African Americans on the positive side of the digital divide to create virtual "third places", to rebuild aspects of community dialogue that have been lost in the physical "real-world". These "third places" arise out of a need for individuals to find a dependable, neutral place of refuge to gather and interact, away from first places (home) and second places (work), often conferring or dealing with issues that may considered too taboo for public discussion by the community at large.

With this in mind the researcher identified an issue within the African American community that was of consequence, and yet was not being addressed due to individual or social pressures. The problem that presented itself was the lack of discussion and social support pertaining to the Human Immunodeficiency Virus (HIV), and Acquired Immune Deficiency Syndrome (AIDS).

BACKGROUND

African American HIV/AIDS Statistics

HIV/AIDS statistics paint a particularly disturbing picture for African American females, due to the fact that they account for a disproportionate number of infections relative to other social groups (Phillips 2005), and 75% of new HIV/AIDS cases within the larger African American population. The Center for Disease Control (CDC), states that HIV/AIDS is among the top 4 causes of death for African American women aged 25–54 years, and the number 1 cause of death for African American women aged 25–34 years (CDC 2006). In 2001, HIV/AIDS was among the top three causes of death for African American men 25-54 years of age, and of persons diagnosed with AIDS since 1995, a smaller percentage of African Americans (60%) were alive after 9 years compared with Whites (70%) (due in part to late diagnosis) (National Institute of Health 2007). Despite these figures, there is still a deafening silence associated with the discussion of the disease, because contraction of the HIV virus is seen as a consequence of behaviors that are stigmatized within the largely religious and conservative African American community (i.e., promiscuity, homosexuality, or drug use), framed within the context of sin and immorality (Baker 1999).

Communities and Expression

An interesting element of any functional community is that it is self-sustaining. In order to be self-sustaining a community has to possess the ability to address issues that affect members' wellbeing, in either a direct or indirect fashion, to ensure that what members are getting out of the association exceeds the cost. Every healthy community discusses issues that threaten its survival. However, the number of African Americans infected and dying from HIV/AIDS is staggering, and the silence associated with the epidemic is akin to having "an elephant in the room" that nobody wants to talk about. This conflict, and the fear of violating group discussion norms, has created a prevalent silence on the subject, and degraded certain aspects of community, namely: emotional safety, sense of belonging, and positive reinforcement when members of the community engage the problem through dialogue.

It is believed that in an effort to "heal" itself of this silence, and restore aforementioned communicative elements of community that have been diminished, African Americans have resorted to finding other outlets to discuss the epidemic; one outlet is online third places. Blogs were chosen because they represent a single place, outside of large social gatherings, where people can engage in real time conversations on a grand scale, and unlike their physical counterparts, users are empowered by the relative cloak of anonymity afforded by the Internet. Never before has a medium such as ICT existed that can connect and enable conversations from members representing all classes of the community, with potentially everyone able to contribute to the discussion and be heard. By technology being an enabler for rebuilding aspects of community, it adds impetus to the drive towards eliminating the "digital-divide" through tangible benefits, such as improved health outcomes through preventative, rather than reactive, practices.

VIRTUAL COMMUNITIES

With the advent of the Internet, we are seeing a movement away from the traditional depiction of communities built around geographic lines, and are seeing the emergence of "communities of interest", or self-organizing virtual communities[a] that are born of individuals who share similar interests on a topic, or topics, that is independent of their geographic location. As stated by Weinberger, "what holds the Web together isn't a carpet of rocks [i.e., the physical Earth], but the worlds collective passion" (Weinberger 2002). With this statement in mind, Milne (2004) provided a germane and useful working "technological" definition of community that suits the needs of this paper. According to Milne, "Community is a social technology for bonding people together through shared characteristics that leads to a sense of belonging". Milne goes on to say that "community" also encompasses the people who are so bonded [technologically], and forming a community is a way to foster a sense of belonging, which serves a wide range of human needs and is the basic survival strategy for individuals and groups (Milne 2004).

Weblogs and Community

Weblogs are defined as "frequently modified Web pages in which dated entries are listed in reverse chronological sequence" (Herring, Scheidt et al. 2005). These Weblogs have also been characterized as having a "community-like" nature to them, due to the inherent interactiveness of the posts, which allow readers to respond to individual entries, which fosters "conversational" exchanges on the blog site itself. Marlow (2004) states that "The weblog medium, while fundamentally an innovation in personal publishing, has also come to engender a new form of social interaction on the Web: a massively distributed but completely connected conversation covering every imaginable topic of interest" (Herring, Kouper et al., 2005; Marlow 2004). Graham (1999) describes blogs as "a community, of sorts, a small town sharing gossip and news, recreation and sport, laughter and tears, all for the commonwealth." The power of real social bonds of virtual communities was captured by one blog participant when he stated that "Weblogs are the first example I've encountered where people are meeting each other in masses, and forming real social bonds, the type of relation you'd call your friend…There's a real sense of solidarity in the relationships we're forming."[b] (Bausch, Haughey et al., 2002). These virtual relationships may prove handy when individuals are faced with a myriad of pressures that may go unabated, with no source of social support systems upon which to lean, as evidenced by the HIV/AIDS epidemic, and its minimal discussion by African American's in "the real world".

THE DIGITAL DIVIDE

Past research has shown that there exists a chasm between individuals who have access to information via computers and the Internet, and those that do not. This chasm has resulted in a corresponding gap between the information rich and the information poor, the haves and the have-nots. This phenomenon has been described in literature as the "digital divide". The digital divide was first observed as distinct group clusters most likely to use the Internet, namely: white, men, residents of urban areas, greater access to education, income and other resources necessary to get ahead (Kvasny & Keil, forthcoming; Mossberger, Tolbert et al., 2003; Norris 2000).

Noting the effects of the digital divide, and the increased likelihood that poorer African Americans will not have access to computers, and thus access to virtual third place sites, one may be tempted to believe that this research is germane only to middle and upper-middle class African Americans. However, this assumption would be false, because research has shown that interactions between African Americans spans class, with upper and middle income members of the community in frequent contact with poorer friends and family. The socioeconomic Web of African Americans transcends the nuclear family that they themselves create, and reaches back to the families to which they were born, their siblings, extended family and friends (Pattillo, 2006). Therefore, the ensuing conversations on the site serve as a means to enrich personal networks via the accessing of new information by means of weak ties (or casual acquaintances), and sharing this newfound information with those with strong ties (or family and friends).

Oldenburg's "Third Places"

The silence associated with HIV/AIDS by traditional African American institutions created a need to form social bonds in other arenas. This need provided fertile conditions for the rise of an alternate place to gather and discuss transcendent issues; such places are typically born out of a need, and can be labeled as "third places". Every stable community is comprised of a first place (home), a second place (work), and a third place (informal gathering location). If one of these components is missing, it affects the stability of the community (Baker-Eveleth, Eveleth et al. 2005). The third place (bars, cafes, barbershops, etc.) provides a context for sociability, spontaneity, community building and emotional expressiveness" (Oldenburg & Brissett, 1982). Within the framework of computer mediated communication there has been a realization that cyberspace (such as blogs, chatrooms, etc.) resemble traditional types of physical social settings described by Oldenburg, providing an informal place where individuals gather to rebuild communicative aspects of community that may be lost (Soukup, 2006).

Oldenburg (1999) describes and elaborates the essential characteristics of third places as:

1. Being on neutral ground
2. Being levelers
3. Conversation being the main activity, with the mood being playful
4. Accessible
5. Are a home away from home, and have "regulars"

NEUTRAL GROUND

Neutrality is not used in a political or position-taking sense, rather it is used to denote that within third places nobody is burdened with the role of being a host or a visitor. This freedom gives individuals in these settings the ability to come and go as they please, easing the task of association, essential to community life. Neutral ground allows for more informal and intimate relationships among people, which cannot necessarily be found in the home.

Levelers

Places that are considered "levelers" are inclusive places that do not have any formal rules for membership and exclusion, and are accessible to the public at large. In locations that are levelers, worldly status claims do not carry as much weight, and there is a realization that there is more to a person than status may indicate. The nature of online environments make physical attributes or material wealth moot in regards to the ability to interact with other members of the community (Baker-Eveleth, Eveleth et al., 2005).

Conversation/Fun

Third places provide a fun atmosphere, due in part to the fact that friends gathered in numbers create a festive mood, and the burden to contribute to the conversation is spread out over many people, making individual interaction easy. This "fun" is created by the members of the community themselves, with the sustaining activity being conversation, which covers the entire range of being passionate and lighthearted, serious and witty, informative, and silly. Everyone who is part of these communities is expected to understand that it is in good fun, and also expected to give-and-take with civility and humor, with the style of conversation emphasized over vocabulary.

Accessible

Third places are places where individuals can go to at almost anytime of the day or night and be reasonably sure that acquaintances will be there to relieve loneliness, boredom or frustrations of the day. Accessibility should be qualified more specifically as "easy accessibility" in order for the third place to survive and serve. The nature of online conversations make it available to anybody who has access to a Web browser and an Internet connection, making it accessible 24 hours a day, seven days a week (Baker-Eveleth, Eveleth et al., 2005).

The Regulars

Third places are often considered more homelike than home, with the definition of a "home" conforming to the notion that it provides a "congenial environment". However, third places will remain simply a "space" without the right people who can transform it into a "place" (Harrison & Dourish, 1996). The people who make this transformation possible are the "regulars", who assure that on most visits a member of the core "gang" is there. It is the approval of the regulars that is critical for welcome and acceptance. Regulars are formed by reappearances and fairly decent game play, with the person able to give and take according to the group norms.

African American Virtual Third Places

Research has found that there is indeed a vibrant community of Black bloggers (Poole, 2005) that is known as the "Blackosphere", as described by Francis Hollander (2007):

These blogs are by and principally for Black people, focusing not only upon Black people but upon people and issues deemed relevant to the Black people who write these blogs and post comments. At Black blogs, we comment on the issues of the day raised in white newspapers and blogs, but we also highlight issues that whites mostly ignore, such as the unfair criminal prosecution of individual humble and unknown Blacks. Our commentary and the relative importance that we give news are informed by our unique historical perspective on and position in America. From our vantage point, we share with each other a distinct perspective and critique that white people, including white progressives, cannot have and generally do not want.

Empirical findings from research of this "Blackosphere" reveals that there are indeed a significant number of individuals within the African American community that utilize blog sites to discuss the crisis of HIV/AIDS, as opposed to the lack of discussion or communal support found in traditional African American "real world" communities (Igwe, forthcoming). This use of technology, specifically blogs, to discuss HIV/AIDS served as a means to circumvent and release frustration directed towards the inactivity of established institutions. This medium was also used to try to make sense of the HIV/AIDS dilemma while discussing

practical protective measures. The blog was a place where rich conversations took place, which helped increase community cohesiveness and emotional wellbeing as participants realized that they were not alone in dealing with the ramifications of the disease running rampant in the community (Igwe, forthcoming).

FUTURE TRENDS

Although information does not guarantee better health practices, it is an important step in health behavior change (Freimuth, Stein et al., 1989). Future research should investigate how conversations in virtual third places manifest themselves in real world action. For example, how do virtual third place HIV/AIDS conversations correspond to health information seeking from physical institutions? It should be said that in order for individuals to engage in health information seeking, uncertainty must be experienced, whether through physical symptoms, or external sources such as television, posters, friends, or information technology (Freimuth, Stein et al., 1989). Further research should explore whether technology (via virtual third places), stimulates this uncertainty, and is this uncertainty enough to stimulate further health information seeking among online African Americans?

CONCLUSION

Findings from researching the "Blackosphere" reveal that they do indeed exhibit characteristics of Oldenburg's third place communities, in virtual world environments (Igwe, forthcoming). The first level of inquiry finds that there is support for Oldenburg's characterization of conversations being the major activity found in third places, and these conversations tend to have a witty or playful nature and are on neutral ground. The second level of inquiry shows that there is support for Oldenburg's characterization of third places being accessible and having regulars. A third level of inquiry showed that there was support for Oldenburg's characterization of leveling found on the Website. The Websites found in the Blackosphere were places where rich conversations took place, and the anonymity afforded by the Internet allowed individuals to establish themselves based more on the content of their textual discourse, rather than real world physical characteristics.

The chapter makes a significant and novel contribution to the existing body of literature and knowledge within the Information Science domain through the utilization of an existing framework to build a strong foundation for future studies of blogs, virtual third places, and HIV/AIDS. This is a foundation that the author hopes will be built upon. This foundation further serves to aid in the understanding of factors that contribute to the utilization of technology by African Americans to rebuild communicative aspects of community. Finally, it aids in the understanding of how blogs serve as virtual third place for those in need of a place to discuss, vent, or cope emotionally with the HIV/AIDS epidemic, and circumvent the silence and inaction from traditional institutional forces within the community.

REFERENCES

Baker, S. (1999). HIV/AIDS, nurses, and the black church: A case study. Journal of the Association of Nurses in AIDS Care, 10(5), 71-79.

Baker-Eveleth, L., Eveleth, D. M. et al. (2005). An emergining on-line "Third place" for information systems (IS) students: Some preliminary observations. Journal of Information Systems Education, 16(4).

Bausch, P., Haughey, M. et al. (2002). We blog: Publishing online with weblogs. New York: John Wiley and Sons.

Department of Health and Human Services: Centers for Disease Control and Prevention CDC (2006). Fact sheet.

Freimuth, V., Stein, J. et al. (1989). The cancer information service model. Philadelphia, University of Pennsylvania Press.

Graham, B. L. (1999). Why I weblog: A rumination on where the hell I'm going with this website. We've got blog: How weblogs are changing our culture. Cambridge: Perseus Book Group.

Harrison, S. & Dourish, P. (1996). Re-place-ing space: The roles of place and space in collaborative systems. In Proceedings of the CSCW '96, ACM.

Herring, S. C., Kouper, I. et al. (2005). Conversations in the blogosphere: An analysis "from the bottom up". In Proceedings of the 38th Hawai'i International Conference on Systems Sciences (HICSS-38).

Herring, S. C., Scheidt, L. A. et al. (2005). Weblogs as a bridging genre. Information Technology & People 18(2), 142.

Igwe, C. F. (forthcoming). Beyond the digital divide into computer-mediated communications: A content analysis of the role of community weblogs in building Oldenburg's virtual third places in black america.

Kvasny, L. & Keil, M. (forthcoming). The challenges of redressing the digital divide: A tale of two US cities. Information Systems Journal.

Marlow, C. (2004). Audience, structure and authority in the weblog community. In Proceedings of the International

Communication Association Conference, May 27-June 1 2004, New Orleans, LA.

Milne, J. M. (2004). Weblogs and the technology lifecycle: Context, geek-chic and personal community. Department of Anthropology, University of South Florida.

Mossberger, K., Tolbert, C. J. et al. (2003). Virtual inequality: Beyond the digital divide. Washington D.C.: Georgetown University Press.

National Institute of Health (2007). Health disparities in HIV/AIDS: Focus on African Americans (R03).

Norris, P. (2000). The worldwide digital divide: Information poverty, the internet and development. Cambridge: Cambridge University Press.

Oldenburg, R. & D. Brissett (1982). The third Place. Qualitative Sociology, 5(4).

Pattillo, M. (2006). Poverty in the family: Race, siblings, and socioeconomic heterogeneity. Social Science Research, 35, 804-822.

Phillips, L. (2005). Deconstructing "down low" discourse: The politics of sexuality, gender, race, AIDS, and anxiety. Journal of African American Studies, 9(2), 3-15.

Poole, A. (2005). Black bloggers and the blogosphere. In Proceedings of the Second International Conference on Technology, Knowledge and Society, Hyderabad, India.

Soukup, C. (2006). Computer-mediated communication as a virtual third place: Building Oldenburg's great good places on the world wide web. New Media & Society, 8(3).

Weinberger, D. (2002). Small pieces loosely joined (a unified theory of the web). Cambridge: Perseus Publishing.

KEY TERMS

Virtual Communities of Interest: Self-organizing virtual communities that are born of individuals who share similar interests on a topic, or topics, that is independent of their geographic location.

Information Age: The period beginning in the last quarter of the 20th century marked by the increased production, transmission, consumption of, and reliance on information.

Third Places: a location that has a role intermediate between the home and the workplace and that allows people to be around other people without being in a structured setting. Examples include coffee shops, pubs, libraries, and public plazas. The essential characteristics of third places are: (1) being on neutral ground, (2) being levelers, (3) conversation being the main activity, with the mood being playful, (4) accessible, (5) are a home away from home, and have "regulars"

Weblogs: Frequently modified Web pages in which dated entries are listed in reverse chronological sequence.

Digital Divide: Chasm between individuals who have access to information via computers and the Internet, and those that do not.

Blackosphere: Blogs that are by, and principally for, Black people, focusing not only upon Black people but upon people and issues deemed relevant to the Black people who write these blogs and post comments.

Health Information Seeking: The search and retrieval of messages that help to reduce uncertainty regarding health status and construct a social and personal (cognitive) sense of health.

ENDNOTES

[a] Also referred to as "Online", "Nonplace based", "Electronic", "Computer mediated communication" (CMC), or "Chosen" communities.

[b] Comments made by Cameron Marlow (Milne; Bausch)

Category: Artificial Intelligence

3-D Digitization Methodologies for Cultural Artifacts

K. Lee
The University of Auckland, New Zealand

X.W. Xu
The University of Auckland, New Zealand

INTRODUCTION

Historic cultural artifacts are objects of high importance and value. They give evidence of a civilization's culture, heritage, and development over time, and often date back thousands of years and are irreplaceable. Thus, the preservation and protection of artifacts against damage or theft is an issue of importance to museums and other conservation organizations.

Developments made in the fields of computer vision and technology have allowed information about artifacts to be archived digitally. These developments have facilitated the use of electronic models and replicas, and have led to numerous organizations, worldwide, increasing research into methods of artifact digitization . (Surendran, Xu, & Stead, 2007)

The three main methods of digitization can be broadly defined as contact digitization, image-based digitization (photogrammetry), and geometry-based digitization (laser scanning). With the development of the latter two digitization methods, and advanced rendering technologies, virtual displays and museums can now be used widely. (Hung, 2007) Furthermore, recent developments in interactive 3-D computer graphics technology have seen an increased interest in, and use of, 3-D digitization for cultural heritage objects. (Muller-Wittig, Zhu, & Voss, 2007) Technologies for reconstructing or remodeling physical components in 3-D formats are not new in the engineering field, in particular within manufacturing engineering. However, 3-D digitization used for the preservation and archiving of cultural artifacts is relatively recent.

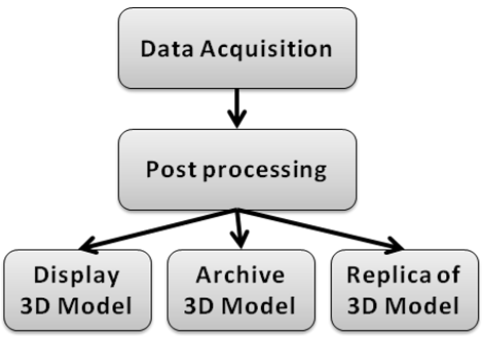
BACKGROUND

Digitization of artifacts is the process of converting spatial and color information into digital formats. 3-D digitization refers specifically to creating a digital representation of an object in three spatial dimensions, that is, Cartesian x, y, and z coordinates. During 3-D digitization, depth, size, proportion, and textural information about the artifact are recorded and stored in electronic form.

There are a wide range of techniques available in the field of 3-D digitization. The specific approach to digitization differs depending on the artifact and the final intended application of the data. The overall process of 3-D digitization involves three broad steps, as shown in Figure 1. Data are acquired using a method of determining and recording the spatial details of the artifact. The raw data are then processed to form complete rendered 3-D models, and then the data are applied to its intended purpose.

3-D digitization of artifacts is an effective method of archiving historic information. It is currently used within museums and other organizations for documentation and security purposes. Furthermore, electronic models give museums the ability to display digital models on the Internet, increasing public awareness and accessibility to the cultural artifacts. The models also facilitate the use of digital interactive displays within museums, allowing viewers to explore objects without risk of damaging the original artifact. Digital information can also be supplied to computer numerical control (CNC) machines to manufacture accurate replicas or

Figure 1. Overall digitization process



support pieces for artifacts. A further application of digitized data is in analysis of artifacts for historical restoration.

CURRENT DIGITIZATION PRACTICES

A number of organizations, worldwide, are promoting research into the digitization of cultural artifacts. These include the Canadian Heritage Information Network (CHIN), the Virtual Heritage Acquisition and Presentation (ViHAP3D) project in Europe, and the Salzburg Research Institute (SRI) in Austria. Some museums are working in collaboration with universities to further research in digitization; in 2004, approximately 35% of museums, worldwide, had initiated developments in some form of 3-D digitization of objects (White, Mourkoussis, Darcy, Petridis, Liarokapis, Lister, et al., 2004). These include the Museum of New York, the Royal Ontario Museum, the Museum of Science Boston, and the American Museum of Natural History. Efforts to establish entire "virtual museums" include The Canadian Museum of Civilization and the National Research Council of Canada collaborating on the production of the Inuit3D Virtual museum, launched in April 2001, and the Computer Science Department of Zhejiang University developing a 3-D Dunhuang cultural relic exhibition system in 2004. (Zhang, Pan, Ren, & Wang, 2007)

The field of 3-D artifact digitization also extends to independent projects, and in several cases, organizations have been assigned specifically to digitize iconic monuments. A project led by Gabriele Guidi, in 2005, involved digitizing the "Plastico di Roma antica," a model of ancient Rome created in the last century. A modulated light scanner was used to provide the accuracy needed to capture the detail of the model's features. The scanner was supplemented by a triangulation scanner to capture the more intricate parts of the model (Guidi, Micoli, Russo, Frischer, De Simone, Spinetti, & Carosso, 2005).

In 2004, the spiral motif at England's Castlerigg stone circle in Cumbria was digitized using the noncontact techniques of laser scanning (using a Minolta 910 scanner) and ground-based remote sensing. No motif was identified through the digitization process, despite the fact that in previous years, the motif image had been observed. This indicated that the spiral was probably painted or had faded due to natural events, and was a novel application of the highly objective methods of 3-D digitization to record the presence of an artifact feature (Diaz-Andreu, Brooke, Rainsbury, & Rosser, 2006).

In 2003, Subodh Kumar, and a team of students from the Johns Hopkins University in Baltimore, undertook the 3-D scanning of ancient cuneiform tablets. Cuneiform documents exhibit writing on three-dimensional surfaces. The team aimed to provide accurate, high-resolution 3-D models of these tablets for scholars' use in their research and for digital preservation of the unique historical artifacts. A laser triangulation scanner was used, using a regular grid pattern at a resolution of 0.025 mm. It was found that conclusive scanning was a challenge using current technologies (Kumar, Snyder, Duncan, Cohen, & Cooper, 2003).

In 2002, David Luebke, and a team from the University of Virginia's Computer Science Department, scanned Thomas Jefferson's Virginia home using a commercial time-of-flight laser scanner, the DeltaSphere 3000. The resultant data from the process was later combined with color data from digital photographs to create the Virtual Monticello, and the Jefferson's Cabinet exhibits, displayed in the New Orleans Art Museum in 2003 (Wang & Luebke, 2003).

The examples described are just a selection of projects that indicate the diversity of applications of 3-D digitization to cultural heritage artifacts.

DATA ACQUISITION TECHNIQUES

The past 15 years have seen the development of different 3-D data acquisition technologies. These include acoustic position trackers, close range photogrammetry, coordinate measurement machines (CMM), holography, laser scanners, magnetic position trackers, and touch probes. The methods of digitization available can be broadly classified as contact or noncontact (mechanical or optical) techniques. (Surendran et al., 2007)

Contact Digitization

Contact digitizing systems work by manually recording points, on the surface of the object, to be digitized. To acquire each data point a probe is activated and runs over the target object. Contact digitizers are commonly used in reverse engineering and manufacturing applications, and are inexpensive and efficient methods for obtaining digital models of objects with low geometric detail. Though these methods could be extended to artifact digitization, in some cases, the process is slower and does not capture as much geometric information as optical techniques. Furthermore, the process of physically touching the object has the potential to damage the artifact itself. For this reason, the two techniques that are predominantly used for artifact digitization are noncontact techniques of laser scanning and photogrammetry (Granero, Sánchez, Micó, Esteve, Hervás, Simón, & Perez, 2007).

Laser Scanning

Laser scanning is a method whereby surface information is captured using laser technology. A 3-D laser scanner is an active system that uses laser light to explore the surface of an object with a process called triangulation, as shown in

Figure 2. Laser triangulation diagram

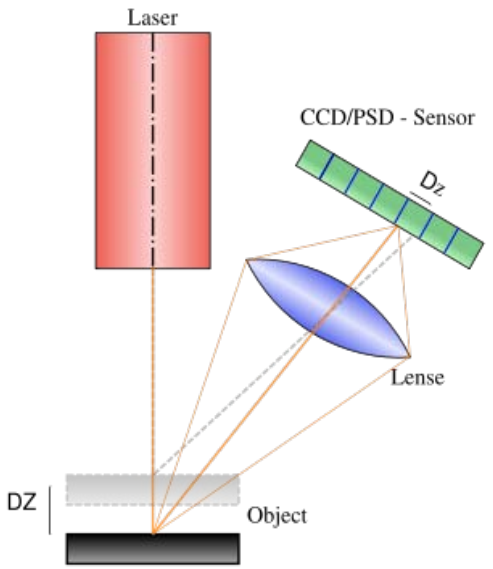


Figure 3. White Light digitisation process

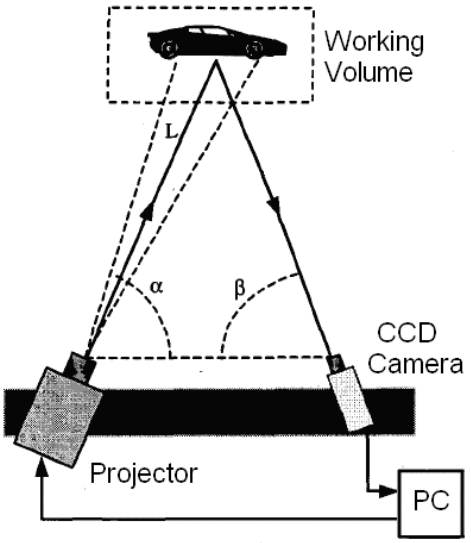


Figure 2 (Wiora, 2006). During the triangulation process, the scanner projects a laser point to the surface, and an angled camera is used to locate the position of the point in its frame of reference. The information from the location of the laser in the camera's field of view is used to determine how far away the object is. The process is called triangulation because the laser emitter, laser point, and camera form a triangle. The distance between the camera and laser emitter, and the angle of the laser emitter, are known. The angle of the camera corner can thus be determined by looking at the location of the laser point in the camera's field of view. This information is sufficient to determine the size and shape of the triangle giving the location of the laser point (Beraldin, 2004). In most scanners, a laser beam, rather than a point, is used to aid faster data acquisition.

Photogrammetry

Photogrammetry, also known as optical digitization, is a common technique used for the acquisition of 3-D data from cultural heritage artifacts. It is a remote sensing technique that uses photographic images to form geometrical information about objects. Photogrammetry systems can be roughly divided into passive or active digitization systems (Petrov, Talapov, Robertson, Lebedev, Zhilyaev, & Polonskiy, 1998).

Passive systems construct a 3-D model of an object's surface using stereoscopic images. Most stereoscopic digitisers are not truly passive, as they interact with the subject by projecting a grid onto its surface to aid image correlation (Petrov et al., 1998). An example of a stereoscopic system is the stereoSCAN produced by Breuckmann (Figure 3). This system comprises two 1.4mega pixel CCD digital cameras and a central halogen projector for projecting a grid onto the object. The camera angle can be adjusted to cater to different artifacts. Different lenses can also be used to alter the range of measurement (Breuckmann, 2006).

Active photogrammetry systems operate using a triangulation technique similar to that used in laser scanning. The difference is that white light, rather than laser light, is projected and reflected into a CCD camera to collect the geometric information (Petrov et al., 1998).

For this reason, they are also known as white light digitisers. White light digitisers are able to project light over a larger angle than laser scanners. Because of this, systems are able to project light from the same location as the sensors, and the size of the system can be reduced. This technique is able to capture color data, which makes it even more suitable for cultural artifact applications. A disadvantage is that white light digitisers are generally less accurate than laser digitisers (Petrov, et al., 1998).

POSTPROCESSING OF DATA

The data from scanning systems is in the form of a dense set of surface points, and is often referred to as "point cloud data." Each point has a distinct position in space dictated by Cartesian coordinates. In the case of color scanners, each vertex will have color information attached to its position data. (Chambard & Chalvidan, 2007)

Postprocessing is the stage responsible for the conversion of this point cloud data into complete and realistic 3-D models. There are several steps that can be implemented to form a final surface model. These include cleaning of irregularities, and the joining of points to form surfaces, filling holes and correcting data, and applying textural or color information.

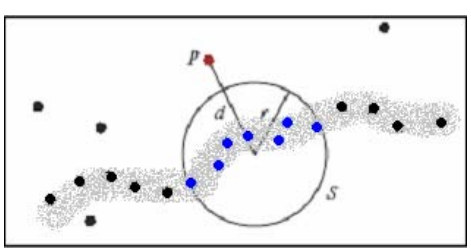
Postprocessing tasks are generally accomplished by using specifically designed software packages. These software packages not only render scan data into 3-D models, but can also provide an efficient means of reverse engineering. The models created through such software can be further used to generate highly accurate engineering drawings. Two of the most popular postprocessing software packages used in industry are Geomagic Studio and RapidForm.

Point Sampling

Point sampling is the first step to be implemented in postprocessing. In this step, data is selectively removed, which also helps to reduce the size of the data files that are being processed. Noisy data is the data that is captured in the process of scanning, but which is not part of the artifact itself. Noisy data can be removed by sampling large sections of data manually, or by allowing the software package to identify noisy points.

Outliers are erroneous data points that do not fall into the threshold of describing the object surface. This concept is illustrated in Figure 4 (Weyrich, Pauly, Keiser, Heinzle, & Gross, 2004). The thick grey line shows the form of the raw point cloud data, while the threshold P dictates the distance from the bulk of the data at which the point will be recognized as noise.

Figure 4. Outliers in scanned point cloud



Often the distribution of points within a scan is inconsistent. Some areas may have densely populated points while other areas may have sparse point data. Point relaxation is the step that aims to create a point cloud with evenly distributed data. By sampling areas according to a uniform density, redundant data can be removed, and the size of the file can be reduced.

Surface Creation

Once the points have been sampled they are able to be formed into surface data. The surface creation step joins adjacent points, in a process called triangulation, to form triangular faces. Each small face plane is built using three neighboring points. Every data point in the scan data is used to model the surface, and so, depending on the scanner accuracy, surface texture detail can be captured. There is a trade-off between model surface accuracy and data size, as a model generated with a larger number of triangular faces will be more representative of the real object, but requires more data.

Missing data or holes can be filled using software tools within postprocessing packages. Software packages consider surrounding face location to generate a best estimate fit for missing data. This step should be considered carefully in the field of artifact archiving, to preserve the integrity of digital models. Improvement of data accuracy to density ratios (higher accuracy for lower file size) and improved methods of interpolating missing data are likely to continue developing, and present a major research challenge in 3-D digitization.

Color and Texture Mapping

Once a surface model has been created, color data can be applied to the model to create realistic color and texture effects. If the scan data included color information, this data may have remained with the spatial data throughout the cleaning and surface generation stages, in which case, color mapping does not need to take place.

In the case where scan data does not include color information, this can be added through application of photographic images to the surface. Distinct geometric features on the model should be matched with their corresponding location on a photograph to position the image. This step can be difficult to accomplish accurately if surfaces have no distinctive features.

Model Validation

Once a model is produced, there must be a method to determine its accuracy. The simplest method is to compare the

original artifact, or images of the original artifact, with the generated model for consistency of features. Dimensions of the final model can be measured on screen and compared to actual dimensions.

A comparison of initial scan data to subsequent surface models can be made within postprocessing software, to form a quantitative test of the deviation of the surface model from its original data. Another method to ensure accuracy is if the textural details of the model surface and color information align. These methods are useful in ensuring that methods used within postprocessing were successful, but do not indicate whether the scan itself contained error. Research is ongoing to improve both the accuracy of scan data itself, as well as model validation methods.

MODEL VISUALIZATION

On completion, models can be displayed in a variety of formats. Exhibiting digitized models online can provide an effective avenue for increasing public awareness and knowledge surrounding historic artifacts.

The original data needs to be exported into a Web-compatible format to enable user viewing. Postprocessing software packages often come with the ability to export model files into publishable formats, for example, RapidForm's ICF format. Other more generic formats can be used for publication, for example, extensible 3-D (X3D), the ISO standard for real-time computer graphics from Web 3-D Consortium (Web 3-D, 2007) and Adobe Flash. Generic formats can be more easily accessible to users than proprietary display formats.

"Virtual museum" environments can be created to display a collection of artifacts models. Virtual digital museums carry out exhibition, preservation, education, and research through a variety of multimedia avenues, such as Web pages, animation, and video clips. These digital museums play an important role in the fields of information management, sharing, preservation, education, and scientific research (Zhu, Zhou, Sean, Tian, & Yan, 2007).

Security and Cultural Issues

Models that are used for Web site display should be protected against theft and plagiarism. Access to high quality 3-D data online presents the risk that information will be misused, for example, in creating unregistered replicas. Further research needs to be conducted to establish clear formats for protecting electronic information in these formats.

The publication of historic and cultural information online should be handled sensitively. Certain artifacts are culturally sensitive, and the parties concerned may not wish to have 3-D models of the artifacts available online. Consideration of these factors should be made in all digitization and display projects.

FUTURE TRENDS

3-D digitization of cultural artifacts is increasing worldwide. Currently, a large proportion of the projects undertaken use specialized equipment and processes developed for each specific project. As methods become standardized and futher research occurs, optimal digitization strategies may be developed and applied across a wider range of projects.

There are a number of research challenges including improving accuracy of data collection, developing data validation techniques, improving accuracy vs. economic viability of digitization, as well as providing secure Internet data structures for models.

As the standardization of digitization process occurs, methods of 3-D digitization for preservation of cultural artifacts will become more commonplace and accessible to museums. The increased development of "virtual museum" gallery environments will facilitate greater user interaction with artifacts online and across other digital media.

CONCLUSIONS

This chapter discussed the technical aspects of the digitization process, as well as applications in archiving, artifact preservation, and display and replica creation.

The digitization of artifacts offers an attractive alternative solution to preserving the historical information embodied in the spatial and textural qualities of artifacts. An accurate digital form of this information can be stored in multiple locations and safeguarded for future use. 3-D digitization also allows the application of 3-D model data to other purposes outside of digital archiving, for example, the 3-D geometric models give museums the ability to display artifacts on the Internet, increasing public awareness and accessibility to the cultural artifacts. The models may also facilitate the use of interactive digital displays within museums, allowing viewers to explore objects without risk of damaging the original artifact. 3-D geometric models can also be used to produce accurate replicas or support pieces for artifacts. A further application of digitized data is in analysis of artifacts and historical restoration.

It is believed that the field of 3-D digitization will continue to grow in scope as methods of digitization become progressively standardized and accessible. It provides an effective tool in the field of cultural heritage artifact preservation.

REFERENCES

Beraldin, J. A. (2004). Integration of laser scanning and close-range photogrammetry – The last decade and beyond. In *Proceedings of the XXth ISPRS congress, Istanbul, Turkey*,

Commission VII (pp. 972-983).

Breuckmann. (2006). StereoSCAN 3D- *The measuring system for highest demands*. Retrieved October 20, 2007, from http://www.breuckmann.com/index.php?id=stereoscan&L=2

Chambard, J., & Chalvidan, V. (2007). Digitization of art pieces based on 3-D, colour and texture parameters. In *Proceedings of SPIE - The International Society for Optical Engineering* , 6618, 66180C.

Diaz-Andreu, M., Brooke, C., Rainsbury, M., & Rosser, N. (2006) The spiral that vanished: The application of noncontact recording techniques to an elusive rock art motif at Castlerigg stone circle in Cumbria. *Journal of Archaeological Science*, *33*(11), 1580-1587.

Granero, L., Sánchez, J., Micó, V., Esteve, J.J., Hervás, J., Simón, S., & Pérez, E. (2007). 3-D digitizing using structured illumination. Application to mould redesign. In *Proceedings of SPIE - The International Society for Optical Engineering*, *6166*, 66164B.

Guidi, G., Micoli, L., Russo, M., Frischer, B., De Simone, M., Spinetti, A., & Carosso, L. (2005). 3-D digitization of a large model of imperial Rome. In *Fifth International Conference on 3-D Digital Imaging and Modeling* (pp. 565 – 572), 13-16 June 2005.

Hung, Y. (2007). An image-based approach to interactive 3-D virtual exhibition . *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4872*, 1.

Kumar, S., Snyder, D., Duncan, D., Cohen, J., & Cooper, J., (2003). Digital preservation of ancient cuneiform tablets using 3-D-scanning. In *Proc. of Fourth International Conference on 3-D Imaging and Modeling* (pp. 326-333).

Muller-Wittig, W., Zhu, C., & Voss, G. (2007). Cultural heritage as digital experience: A Singaporean perspective . *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *4563*, 680-688.

Petrov, M., Talapov, A., Robertson, T., Lebedev, A., Zhilyaev, A., & Polonskiy, L. (1998). Optical 3-D digitizers: Bringing life to the virtual world. *Computer Graphics and Applications, IEEE*, *18*(3), 28-37.

Surendran, N., Xu, X., & Stead, O. (2007). Contemporary technologies for 3-D digitization of Maori and Pacific Island artifacts. *International Journal of Imaging Systems and Technology.*

Wang, R., & Luebke, D. (2003). Efficient reconstruction of indoor scenes with color. In *Proc. of Fourth International Conference on 3-D Imaging and Modeling*, 402-409.

Web 3-D Consortium. (2007). *What is X3D? Web 3-D Consortium*. Retrieved October 9, 2007, from http://www.web3d.org/

Weyrich, T., Pauly, M., Keiser, R., Heinzle, S., & Gross, M., (2004). Postprocessing of scanned 3-D surface data. *Eurographics Symposium on Point Based Graphics*.

Wiora, G. (2006). *Laserprofilometer*. Retrieved March 13, 2008, from http://en.wikipedia.org/wiki/Image:Laserprofilometer_EN.svg

White, M., Mourkoussis, N., Darcy, J., Petridis, P., Liarokapis, F., Lister, P., Walczak, K., Wojciechowski, K., Cellary, W., Chmielewski, J., Stawniak, M., Wiza, W., Patel, M., Stevenson, J., Manley, J., Giorgini, F., Sayd, P., & Gaspard, F. (2004). ARCO—An architecture for digitization, management and presentation of virtual exhibitions. In *Proceedings of the Computer Graphics International*, *19*(19), 622–625.

Zhang, M., Pan, Z., Ren, L., & Wang, P. (2007). Image-based virtual exhibit and its extension to 3-D. *International Journal of Automation and Computing*, *04*(1), 18-24.

Zhu, T., Zhou, Y., Sean, H.S., Tian, F., & Yan, X. (2007). Plant modeling and its application in digital agriculture museum. *Lecture Notes in Computer Science, 4563.* Springer Berlin/Heidelberg.

KEY TERMS

CHIN: Canadian Heritage Information Network – a network of professionals and volunteers whose objective is to promote the development, the presentation, and preservation of Canada's digital heritage content for current and future generations.

ICF: INUS compression format file for publication

Point cloud: The collection of points in 3-D space resulting from scanning an object

RapidForm: A comprehensive suite of software designed to convert real-world data from 3-D scanning devices into high quality, accurate, and useful data for a variety of applications.

SRI (Salzburg Research Institute): An institute fully owned by the Province of Salzburg, specializing in digital media

Stereoscopic: Two 2-D images that, when combined, give depth perception.

Triangulation: Within postprocessing of digitized spatial information, the creation of faces or surfaces by

joining vertices according to spatial distribution to create triangular planes.

ViHAP3D: Virtual heritage high quality 3-D acquisition and presentation – a project founded by the European Union deigned to increase public awareness of Europe's most precious artefacts and documents.

3D Graphics Standardization in MPEG-4

Marius Preda
Institut Telecom/Telecom & Management Sudparis, France

Françoise Preteux
Institut Telecom/Telecom & Management Sudparis, France

INTRODUCTION

Computer graphics is currently the spine of several industries of information technology. From computer aided design of almost all objects surrounding us, through virtualization of 3D environments such as Google Earth, to video games, 3D graphics becomes a valuable media with a similar impact as the one of video, image and audio. Such digital assets should be exchanged between producers, service providers, network providers and device manufacturers. Supporting large-scale interoperability needs the development and deployment of open standards. ISO[a], as a major international standardization organization, anticipated this issue and proposed in the last decade several standards addressing 3D graphics assets exchange. VRML (ISO, 1997) (Virtual Reality Modeling Language), published in 1997, provides basic geometry primitives, appearance models and animation mechanisms for representing 3D objects and scenes. Built on top of VRML, the first version of MPEG-4 (ISO, 1998) supports tools for the compression and streaming of graphics assets. Since then, MPEG improved the 3D graphics compression technologies and published the MPEG-4 Part 16 (ISO, 2004) in order to address theses issues within a unified and generic framework.

This chapter is dedicated to professionals of 3D graphics industry, solution providers for on-line systems involving 3D content (games, persistent universes, virtual spaces with graphical representation), students and professors in digital sciences.

The first section aims at presenting the background of the compression for multimedia signals. The second section presents the latest developments in MPEG-4 with respect to the compression and streaming of 3D graphics objects and scenes. The MPEG-4 tools, categorized into geometry, appearance and animation, are introduced. While MPEG standards specify only the bit-stream syntax and the architecture of the decoder, scheme for encoders' implementation are presented in this section.

In the third section we describe a recently adopted MPEG model for 3D graphics consisting in opening the representation of graphics primitives to any XML-based format and completing it with binarization and compression layers. This shift in the manner of using MPEG-4 for 3D graphics is concretized in Part 25 of MPEG-4 (Preda, Jovanova, Arsov& Prêteux, 2007).

BACKGROUND

A generic compression schema aims at eliminating redundancy in the data representation. Additionally when dealing with lossy compression, it makes also possible to identify

Figure 1. Generic signal compression schema

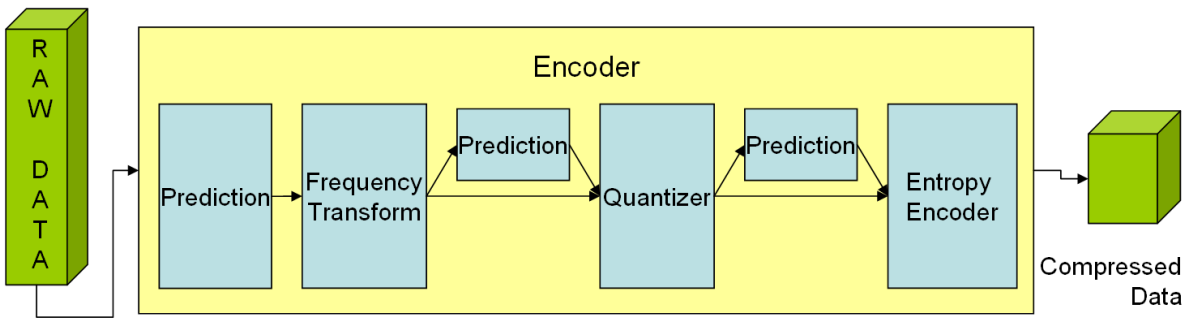


Table 1. MPEG-4 tools for 3D graphics compression

Compression Tool	Type
3D Mesh Compression	Geometry
Wavelet Subdivision Surface	Geometry
Coordinate, Orientation and Position Interpolator	Animation
Bone-based Animation	Animation
Frame-based Animated Mesh Compression	Animation
Octree Compression for Depth Image-based Representation	Appearance
Point Texture	Appearance

the attributes for which a less precise reconstruction results in an acceptable signal distortion for human observers. As example, image and video encoders exploits the way that humans perceive the colors. In general, a compression schema is based on prediction, frequency transform, quantization, and entropy encoder as illustrated in Figure 1.

Compression is necessary when data should be transmitted over the networks or to optimize the data storage. Nowadays, it is used in many applications of the real life: photo cameras, digital television, DVDs, video servers on the Internet (like Youtube) and so forth. Forecasting the importance of the compression for the development of multimedia applications, the MPEG consortium initiated in 1998 an international standardization project for specifying bit-stream syntax for audio and visual information. The first product, called MPEG-1 was designed for compressing video and audio at low bit-rates. A world-wide recognized part of MPEG-1 is the MP3format, used for compressing music data. The second product, MPEG-2 was designed for compressing video and audio in high-quality. The technical performances of MPEG-2 explain the success of applications such as DVD[b] (used by the movie industry) and DVB[c] (used in television). MPEG-4 completes the compression solutions provided by the MPEG consortium and is the main focus of the current chapter.

MAIN FOCUS OF THE CHAPTER: MPEG-4 TOOLS FOR 3D GRAPHICS COMPRESSION

A major advancement of MPEG-4 over its predecessors, MPEG-1 and MPEG-2, is the extension of data types to rich media, offering the possibility to handle, in a unique format, pixel-based image representation, 2D scalable vector graphics and 3D graphics. On top of these representations, the MPEG consortium developed generic and data-specific compression tools. In order to handle the composition and presentation of various media elements in the scene as well as the user interactivity, MPEG-4 introduced the concept of scene graph by adopting the VRML specifications and adapting it to specific requirements of streaming. A first tool responding to such adaptation is BIFS[d] (ISO, 2005), a binary encoded version of an extended set of VRML. Designed as a generic tool, BIFS attempts to balance the compression performances with the extensibility, ease of parsing and simple bit-stream syntax. It includes the traditional modules such as prediction, quantization and entropy coding, without pushing them to extreme complexity. With respect to the textual description of the same data, BIFS may compress by a factor up to 15:1, depending on quantization step. However, for 3D graphics primitives and for animation, BIFS does not fully exploit the spatial and respectively, temporal correlation of (animated) 3D objects. To overcome this limitation, MPEG defined specific tools in Part 16 of the MPEG-4 standard.

For compressing 3D graphics assets, MPEG-4 offers a rich set of tools classified with respect to the data type, namely geometry, appearance and animation (Table 1).

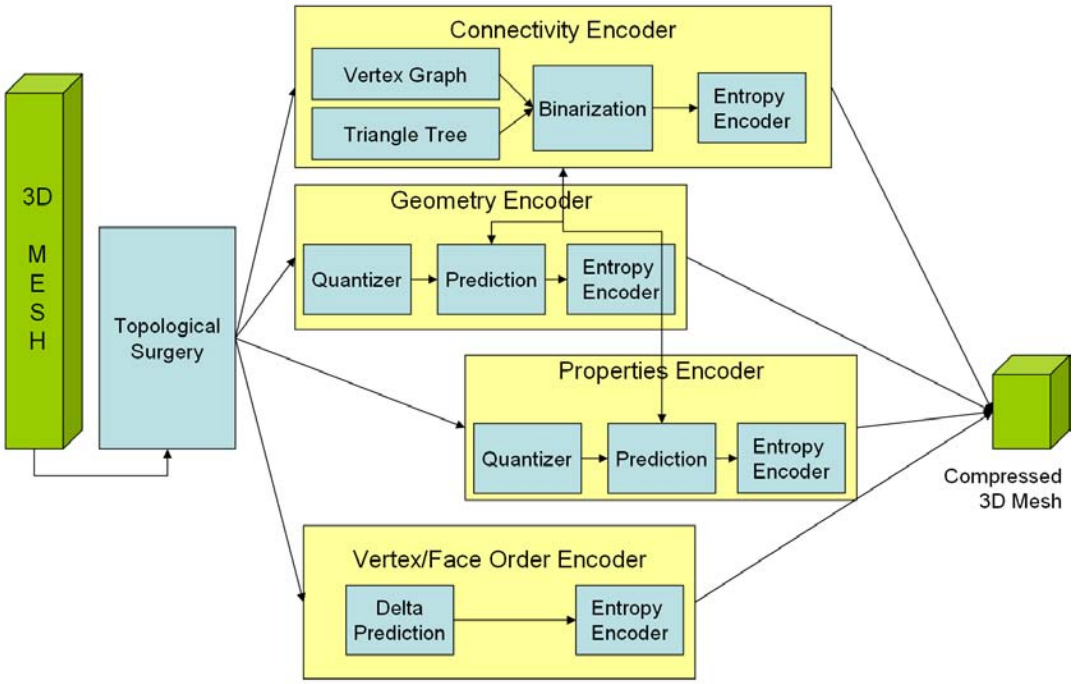
The following sections describe the key elements for geometry and animation compression tools.

3D Mesh Compression (3DMC)

3DMC, initially published in 1999 (ISO, 1999) and extended in 2007 (ISO, 2007) is based on the Topological Surgery (TS) representation (Taubin & Rossignac, 1998). It applies to 3D meshes defined as an indexed list of polygons and consists of geometry, topology and properties (e.g., color, normal, texture coordinate, and other attributes).

The connectivity information is encoded loss-less, whereas the other information could be quantized before compression. In order to maintain the congruence of the system, geometry and properties information are encoded in a similar fashion. 3DMC supports three modes of compressing the mesh: single resolution, when the entire mesh is encoded as indivisible data, incremental representation, when data is interleaved such as each triangle may be rendered immediately after decoding, and the hierarchical mode, when an initial approximation of the mesh is improved by additional decoding of the details. 3DMC supports error resilience and computational graceful degradation. The extension published

Figure 2. General block diagram of an MPEG-4 3D mesh encoder



in 2006 allows preserving vertex and/or triangle order and improves compression of the texture mapping information. When using 3DMC or its extension, the compression gain is up to 40:1 with respect to textual representation. Figure 2 shows the encoder scheme highlighting the main components.

Wavelet Subdivision Surface Streams (WSS)

Wavelet methods for geometry encoding are a superset of multi-resolution analysis which has proven to be very efficient in terms of compression and adaptive transmission of three-dimensional (3D) data. The decorrelating power and space/scale localization of wavelets enable efficient compression of arbitrary meshes as well as progressive and local reconstruction. Especially for signal transmission purposes, subband decomposition is a very important concept. Not only does it permit to send a coarse version of the signal first and progressively refine it afterwards, but it also enables a more compact coding of the information carried by signals whose energy is mostly concentrated in their low-frequency part.

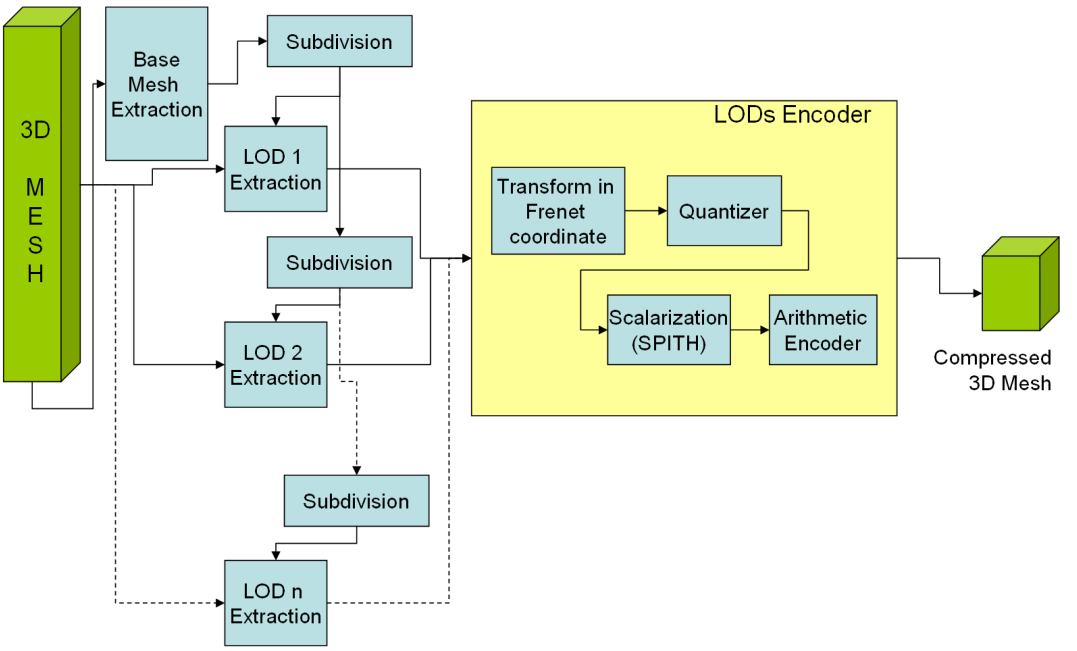
WSS of MPEG-4 is a hierarchical compression tool that uses a list of indexed triangles as a base mesh and encodes the vertex positions at different resolution levels based on subdivision surface predictors. The WSS contains only corrections of vertex position prediction encoded by using SPIHT (Set Partitioning In Hierarchical Trees) technique (Said & Pearlman, 1996); for all the other attributes defined per vertex (normals, colors, texture coordinates, etc.) linear interpolation schemes are used. WSS is well suited for applications such as terrain navigation. Optimal systems using WSS (Gioia, Aubault & Bouville, 2004) combines algorithms for local updates, cache management and server/client dialog and reports compression gain of up to 40:1. The schema of a WSS encoder is illustrated in Figure 3.

Coordinate Interpolator, Orientation Interpolator and Position Interpolator Streams (CI, OI and PI)

Interpolator representation in key-frame animation is currently the most popular method for computer animation. The interpolator data consist of key and key value pairs, where a key is a time stamp and a key value is the corresponding value to the key. Depending on the data type (coordinate, orientation, or position) the key value may have different dimension. The key value for coordinate and position is the set of X, Y and Z (dimension 3), while for the orientation it is represented as an axis and an angle (dimension 4).

The structure of the Interpolator Compression (IC) is illustrated in Figure 4. First, the original interpolator data may be reduced through an analyzer with the role of removing the redundant or less meaningful keys from the original set. A redundant key is defined as a key that can obtained from its neighbors by interpolation and a less meaningful

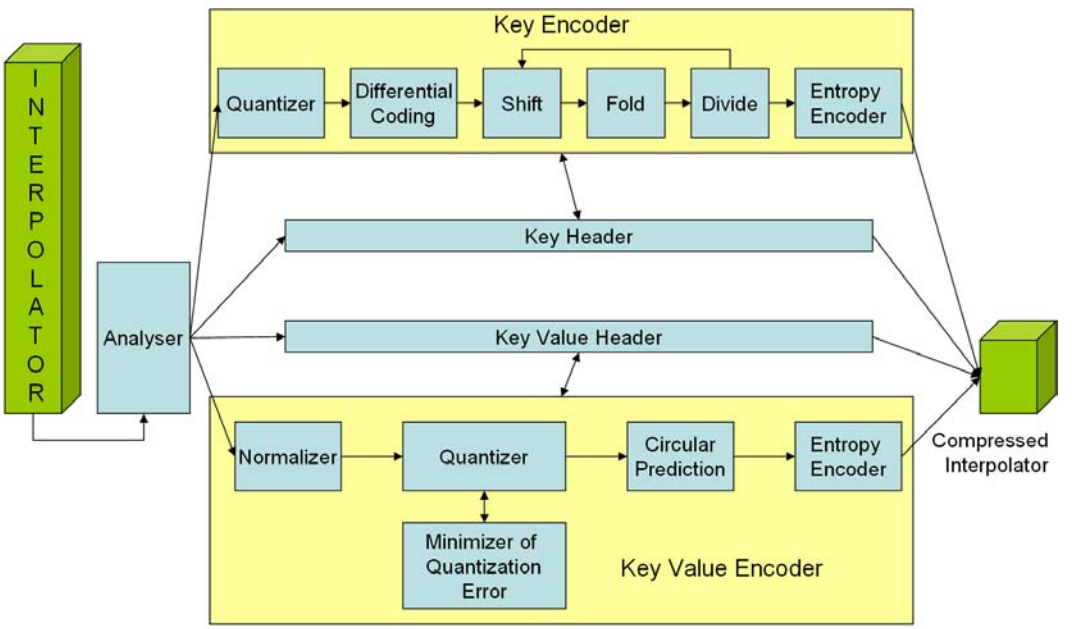
Figure 3. General block diagram of an MPEG-4 WSS Encoder. (LOD means Level of Details)



key is one for which the distortion between the original signal and the one reconstructed by interpolation is below a given threshold.

Once the significant keys are selected, each component of a pair (key, key value) is processed by a dedicated encoder. The key data, an array with monotonically non-decreasing values and usually unbounded, is first quantized and a delta prediction is applied. The result is further processed by a set of shift, fold, and divide operations with the goal of reducing the signal range (Jang, 2004).

Figure 4. Interpolators encoder



Data such as number of keys and quantization parameters are encoded by a header encoder using dictionaries. For the key values, the data is first normalized within a bounding box and then uniformly quantized. The quantized values are predicted from their one or two already transmitted key-values and the prediction errors are arithmetically encoded.

By using this method for representing interpolators, the compression performances are up to 30:1 with respect to textual representation of the same data.

Bone-Based Animation Stream (BBA)

In order to represent compactly the animation data (varying different vertex attributes, mainly spatial coordinates, but also normals or texture coordinates), some kind of redundancy in the animation is exploited: either temporal or spatial. In the first case, linear or higher order interpolation is used to compute the value of an attribute based on key values. In the second, vertices are clustered and a unique value or geometric transform is assigned to each cluster. For avatar animation, MPEG published BBA (Preda, Salomie, Prêteux & Lafruit, 2004) which is a compression tool for geometric transforms of bones (used in skinning-based animation) and weights (used in morphing animation). These elements are defined in the scene graph and should be uniquely identified. The BBA stream refers to these identifiers and contains, at each frame, the new transforms of bones (expressed as Euler angles or quaternion) and the new weights of the morph targets.

A key point for ensuring a compact representation of the BBA animation parameters consists in decomposing the geometric transformations into elementary motions. For example, when using only the rotation component of the bone geometric transformation, a binary mask indicates that the other components are not involved. The compactness of the animation stream can still be improved when dealing with rotations by expressing them in local coordinates system. The rotation can be represented either as a quaternion (4 fields to be encoded) or as Euler angles (3 fields to be encoded).

BBA follows traditional signal compression schema by including two encoding methods as illustrated in Figure 5. Optimized implementation (Preda et al., 2007) of the BBA encoder obtains up to 70:1 compression factor with respect to a textual representation.

Frame-Based Animated Mesh Compression Stream (FAMC)

FAMC is a tool to compress an animated mesh by encoding on a time basis the attributes (position, normals, etc.) of vertices composing a mesh. FAMC is independent on the manner how animation is obtained (deformation or rigid motion). The data in a FAMC stream is structured into segments of several frames that can be decoded individually. Within a segment, a temporal prediction model, used for motion compensation, is represented.

Each decoded animation frame updates the geometry and possibly the attributes of the 3D graphic object that FAMC is referred to. Once the mesh is segmented with respect to motion of each cluster, three kinds of data are obtained and encoded as illustrated in Figure 6. First, the mesh partitioning

Figure 5. BBA Encoder. (DC means Discreet Coefficient, AC means Alternative Coefficients).

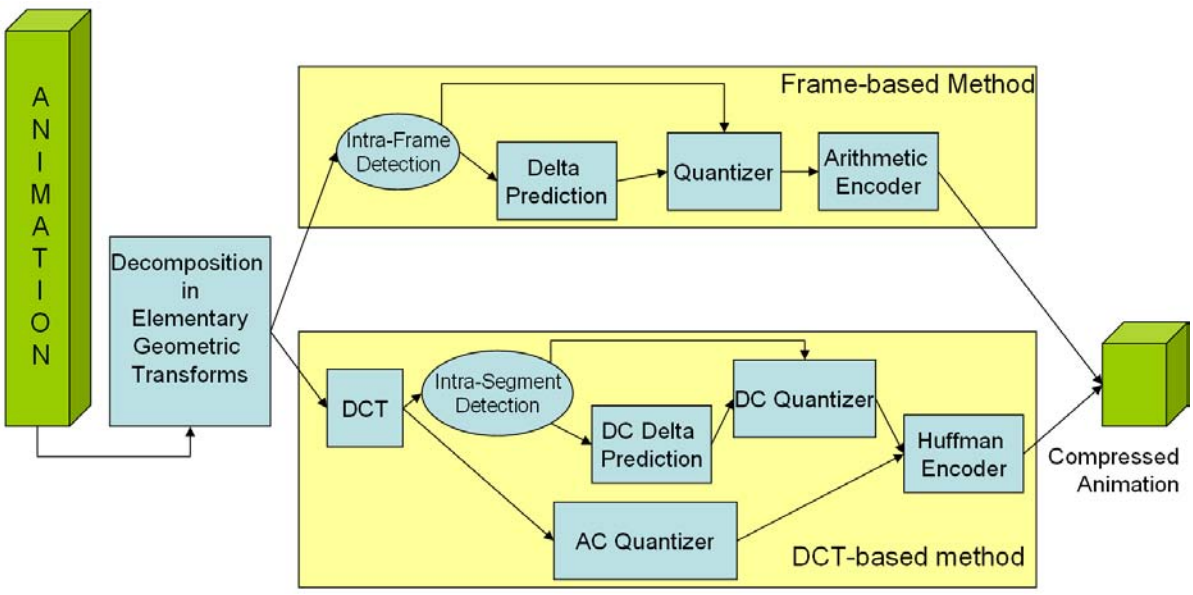
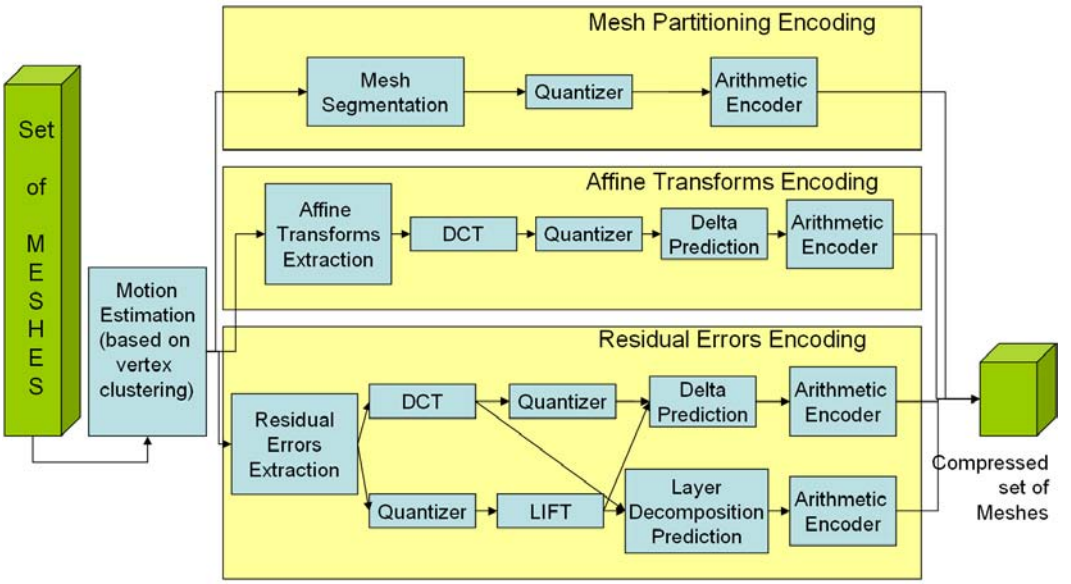


Figure 6. FAMC encoder



indicates for each vertex the attachment to one or several clusters. Secondly, for each cluster an affine transform is encoded by following a traditional approach (frequency transform, quantization, prediction of a subset of the spectral coefficients, and arithmetic encoder). Last, for each vertex a residual error is encoded. The frequency transform may be chosen between DCT and Wavelet (LIFT) and the prediction may be delta prediction or one based on multi-layer decomposition of the mesh. Current implementation (Mamou, 2007) of the FAMC encoders reports up to 45:1 compression factor with respect to a textual representation.

FUTURE TRENDS: MPEG-4 PART25, A NEW MANNER TO ACCESS MPEG COMPRESSION

Initially, all the 3D graphics compression tools developed by MPEG were designed to be applied on top of graphics primitives specified by BIFS. Each compression tool corresponds to one or several nodes in the scene graph, being able to efficiently encode the information of the node (as in the case of geometry tools) or to update some fields of the node (as in the case of animation tools).

Recent developments in the space of 3D graphics formats show a large diversity for definition of scene graph and graphics primitives, several standards being available today. The most known are COLLADA by Khronos Group[e], X3D by Web3D Consortium and XMT by MPEG. Other proprietary formats are defined, in general each authoring tool having a proprietary format. Some of them include rich sets of primitives; others are specialized for specific data (e.g. avatars for H-Anim). The majority of them use XML (XML, 2006). However, beside MPEG-4, none of them provide tools for compression, in general an entropy encoder (usually gzip) satisfying this need.

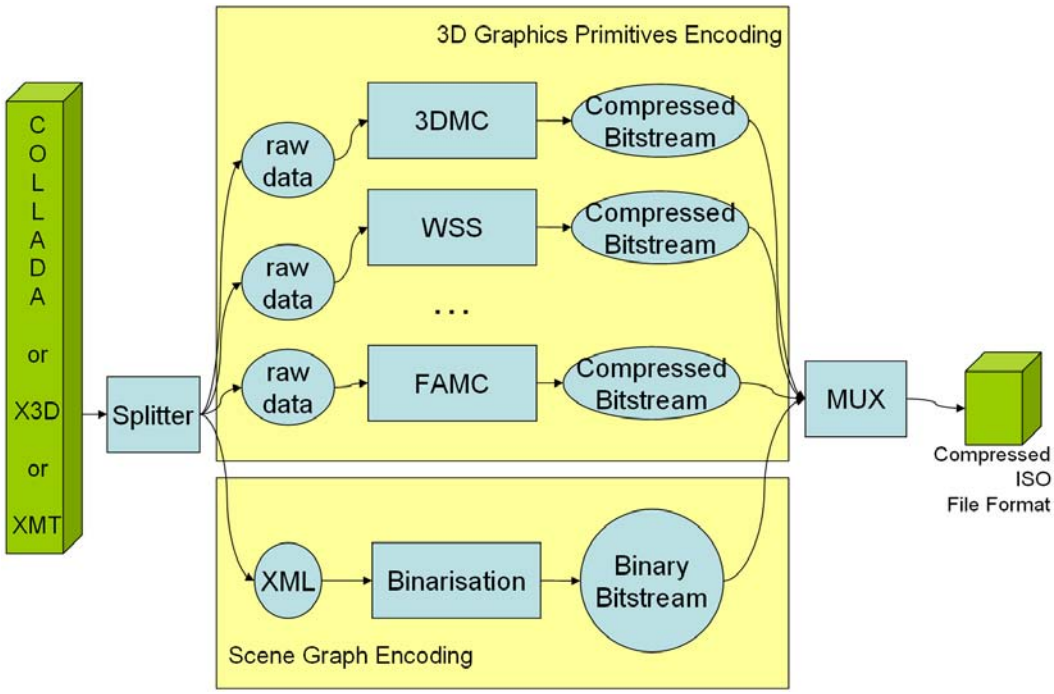
The goal of P25 of the MPEG-4 standard, also called 3D Graphics Compression Model, is to specify an architectural model able to accommodate third-party XML based description of scene graph and graphics primitives with binarisation tools and with MPEG-4 3D Graphics compression tools.

The advantages of such approach are on one side the use of powerful MPEG-4 compression tools for graphics and, one another side, the support of a large set of graphics primitives formats. Hence the compression tools described in previous section would not be applied only to the scene graph defined by MPEG but to any scene graph definition. The bit-streams obtained when using the P25 model are MP4 formatted and contain XML (or binarized XML) data for scene graph and binary elementary streams for graphics primitives (geometry, texture and animation).

Architecture Model

The architectural model has three layers: Textual Data Representation, Binarisation and Compression. In the Textual Data Representation layer, the model can accommodate any scene graph and graphics primitives' representation formalism. The only requirement on this representation is that it should be expressed in XML. Any XML Schema (specified by MPEG or by external bodies) may be used. Currently P25 supports the following XML Schemas: XMT, COLLADA, X3D, for each one the standard indicating the connection between the XML elements and the compression tool.

Figure 7. Encoding path in Part 25



The Binarisation layer provides a generic binarisation tool of the XML Schema. The XML data is represented in a "meta" atom of an MP4 file and can be textual or binary (gzip).

Finally, the compression layer includes the elementary streams listed in Table 1 and encoded as specified in the previous sections. A usual implementation of an MPEG-4 encoder generating bit-streams compliant with P25 is indicated in Figure 7. The MUX is used for multiplexing the elementary streams and formats the MP4 file.

CONCLUSION

This overview is purposed to the standardization of compression for different graphics primitives as specified in MPEG-4. This standard addresses a complete set of such primitives, ensuring efficient representation and streaming capabilities and allowing for reduction of the data size up to 70 times. Initially designed as a complete solution for 3D graphics (by specifying a formalism for textual representation usable for production purposes together with the compression tools), MPEG-4 recently opened the door to third-party solutions for graphics primitive formalisms, building a generic architectural model, called P25. The model is able to accommodate the MPEG compression tools on top of XML representation. Therefore, with P25, MPEG-4 for 3D graphics becomes a transparent layer for storage and transmission, without imposing a format for content production or/and consumption.

REFERENCES

International Standards and Recommendations

ISO/IEC JTC1/SC24, (1997). Standard ISO/IEC 14772, a.k.a. The Virtual Reality Modeling Language, ISO, 1997.

ISO/IEC JTC1/SC29/WG11 (1999). Standard 14496-2, a.k.a. MPEG-4 Part 2: Visual, ISO, 1999.

ISO/IEC JTC1/SC29/WG11 (2004). Standard 14496-16, a.k.a. MPEG-4 Part 16: Animation Framework eXtension (AFX), ISO, 2004.

ISO/IEC JTC1/SC29/WG11 (2005) Standard 14496-11, a.k.a. MPEG-4 Part 11**: Scene description and application engine, ISO, 2005.**

ISO/IEC JTC1/SC29/WG11 (2007) Standard 14496-16 AMD1, a.k.a. MPEG-4 Part 16 AMD1**: Geometry and shadow, ISO, 2007.**

Preda, M., Callow, M., Hwan, J. A. (2007). Committee draft of ISO/IEC JTC1/SC29/WG11 Part 25, 3D graphics

compression model. *In Proceedings of the Shenzhen MPEG Meeting,* October 2007.

XML (eXtensible Markup Language) Core Working Group(2006). *XML 1.0* (4th ed.), W3C (World Wide Web Consortium). Retrieved June 18, 2008, from http://www.w3.org/TR/2006/REC-xml-20060816/.

Books, Book Chapters and Journal/Conference Papers

Gioia, P., Aubault, O., & Bouville, C. (2004). Real-time reconstruction of wavelet-encoded meshes for view-dependent transmission and visualization. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(7), 1009-1020.

Jang, E. S., Kim, J. D. K., Seok Yoon Jung, Mahn-Jin Han, Sang Oak Woo, & Shin-Jun Lee (2004). Interpolator data compression for MPEG-4 animation. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(7), 989- 1008.

Mamou, K., Zaharia, T., & Preteux, F. (2007). A skinning approach for dynamic3d mesh compression. *Computer Animation VirtualWorlds, 17*(3–4), 337–346.

Preda, M., Salomie, A., Prêteux, F., & Lafruit, G. (2004). Virtual character definition and animation within the MPEG-4 standard. In M. Strintzis, N. Sarris (Ed.), *3D modeling and animation: Synthesis and analysis techniques for the human body* (pp. 27-69). Hershey, PA: IRM Press.

Preda, M., Jovanova, B., Arsov, I., & Prêteux, F. (2007) Optimized MPEG-4 animation encoder for motion capture data. In *Proceedings of the 12th International Conference on 3D Web Technology (Web3D'2007)*, Perugia, Italy.

Said, A. & Pearlman, A. (1996). A new, fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Circuits and Systems for Video Technology, 6*(3), 243-250.

Taubin, G. & Rossignac, J. (1998). Geometric compression through topological surgery. *ACM Transactions on Graphics (TOG), 17*(2), 84-115.

KEY TERMS

MPEG: "Motion Picture Expert Group", marketing name of the "ISOIEC SC29 WG11" standardization committee, affiliated to ISO (International Standardization Office) and creator of the multimedia standards: MPEG-1, MPEG-2, MPEG-4, MPEG-7 and MPEG-21.

3DGC: "3D Graphics Compression" – an MPEG working group dealing with specifications of 3D Graphics tools and integration of synthetic and natural media in hybrid scenes.

3DMC: "3D Mesh Compression" – a part of MPEG-4 specifications dealing with the specification of the bit-stream syntax for compressed meshes.

BIFS: "Binary Format for Scene" – a binary formalism defined by MPEG in the standard ISO/IEC 14496-11 for compressing the scene graph.

XMT: "eXtensible MPEG 4 Textual Format" – a XML formalism defined by MPEG in the standard ISO/IEC 14496-11 for representing the scene graph.

WSS: "Wavelet Subdivision Surfaces" – a part of MPEG-4 specifications dealing with the specification of the bit-stream syntax for compressed meshes described in a hierarchical manner.

IC: "Interpolator Compression" – a part of MPEG-4 specifications dealing with the specification of the bit-stream syntax for compressed animation interpolators.

BBA: "Bone-based Animation" – a part of MPEG-4 specifications dealing with the definition and the animation at very low bit-rate of a generic articulated model based on a seamless representation of the skin and a hierarchical structure of bones and muscles.

FAMC: "Frame-based Animation Compression" – a part of MPEG-4 specifications dealing with the specification of the bit-stream syntax for a set of meshes with consistent connectivity and temporal updates of vertex attributes.

COLLADA: An XML based formalism standardized by the Khronos consortium for representing 3D graphics assets.

ENDNOTES

[a] International Standards Organisation, www.iso.ch
[b] DVD – Digital Versatile Disk
[c] DVB – Digital Video Broadcast
[d] BIFS – BInary Format for Scene
[e] www.khronos.org

T-Learning Technologies

Stefanos Vrochidis
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Francesco Bellotti
ELIOS Lab, University of Genoa, Italy

Giancarlo Bo
Giunti Labs S.r.l., Italy

Linda Napoletano
O.R.T. France, France

Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

INTRODUCTION

Television (TV) is a ubiquitous consumer electronics device representing the traditional information and entertainment medium for the majority of the people.

Following the rapidly growing technology, TV started to switch-off from the analogue world to the modern digital technologies of broadcasting. Digital technology has the potential to offer the audience a variety of services apart from the common audiovisual stream. Many of the new services are inherited from the Personal Computers (PC) world, including on-demand features, games, transactions, and other interactive options.

Television has had a long history of performing an educational function for the mass audience, typically by broadcasting culturally-relevant movies, documentaries and news, as well as educational programmes. The idea of Distance Learning through a TV blossomed extensively in particular as a complementary educational option besides PC-based e-learning and traditional analogue TV educational programs. In particular, TV-based interactive education promises a huge potential due to its ability to support interactivity while compensating for the low penetration of Internet-enabled computers in comparison with the penetration of a TV in a household.

"T-learning" was the new term, which prevailed for the definition of TV-based interactive learning (Aarreniemi-Jokipelto, 2005).

BACKGROUND

The first forms of learning with interactive Digital TV (iDTV) have been little more than modified or enhanced videoconferencing. Today, iDTV platforms for learning provide a big amount of audiovisual and educational contents to the viewer through interactive and content personalization. iDTV is considered as the convergence of television and computer technologies by encompassing three important features typical of computer-based technologies (Lytras Lougos, Chozos, & Pouloudi, 2002):

- **Interactivity:** The control of the whole activity and of the elements of a single activity can be placed into the hands of the potential consumer (Watheieu & Zoglio, 2002);
- **Personalization:** Use of technology and viewer information, to tailor interactive content to each individual viewer profile (Lekakos & Giaglis, 2001); and
- **Digitization:** Technological advancements that allow better quality of sound and picture (Kenyon, Miles, & Rose, 2000).

In particular, considering the use of the media by its audience, TV has some features that make it different from PCs. First of all, TV is usually watched by more than one person (co-viewing), and usually triggers social interactions that are very useful for a more effective experience and interiorization of the contents. Secondly, the logic of broadcasting to a wide population enables social mass mechanisms that typically enhance the impact of the broadcast program.

Nowadays, there are signs that the TV providers are moving to interactive education by broadcasting educational programs that exploit the interactivity of iDTV. A characteristic example is the BBC channel, which offers a learning portal (BBC learning) that provides interactive learning services and covers all the most widespread media, such as radio, TV, iDTV, Web and broadband. Some of the Web interactive services of BBC are also available in BBCi Interactive TV as the ones devoted to preschool children (BBC CBeebies) and support "Learning through play"[1]. Although t-learning as a rather new concept has not been applied so widely in interactive TV, there is a number of projects that support and investigate the future penetration of t-learning as the Enhanced Learning Unlimited (ELU) project which is currently dealing with the iDTV technologies for the design and the implementation of an integrated t-learning system[2].

The article is organized as follows: the pedagogical aspect of t-learning is presented in the next section while the part "Technologies Involved in T-learning" is dedicated to a description of the available technologies and standards of iDTV which are exploited in an education-effective way by t-learning.

PEDAGOGICAL ASPECTS OF T-LEARNING

In defining a t-learning pedagogy it is crucial to deal with an active learning model, the constraints imposed by the actual development of the technology and the nature of the allowed interactions. Related research acknowledges active learning as an exceptionally effective teaching technique (Clark, Nguyen, & Sweller, 2006). More specifically, active learning strongly relies on the learners' interactions with their environment that lead to mental actions through which they construct ideas about what they are encountering.

In this context, the challenge is to exploit the added value of providing an interactive learning environment and the potential of allowing people to access learning activities and contents directly in their house, at distance, through media easy to access and simply to use.

This reflection produces a twofold vision that aims at balancing learning and teaching strategies:

1. Leave the control to the learner.
2. Guide the learner.

Thus, to draw a pedagogy for t-learning experiences, two dimensions have to be explored and taken into account as the drivers of the design process:

- The context where learning happens and the behaviour of learners in this environment;
- The specific features of the medium.

The interactivity, audio/video-based experiences, narrative learning environment and informal learning/edutainment are the key points that emerged from this exploration.

TECHNOLOGIES INVOLVED IN T-LEARNING

T-learning exploits in an educational manner the available technologies and standards for iDTV such as the broadcast technology, the supported middleware for applications and the variety of related tools.

Interactive Digital TV

iDTV has been pushed into the marketplace by the broadcast industry and the network operators in the last decade, introducing two major features, which will be presented in this section: the digitization of the broadcasting and the availability of interactive programs (Baker, Pulles, & Sasno, 2004).

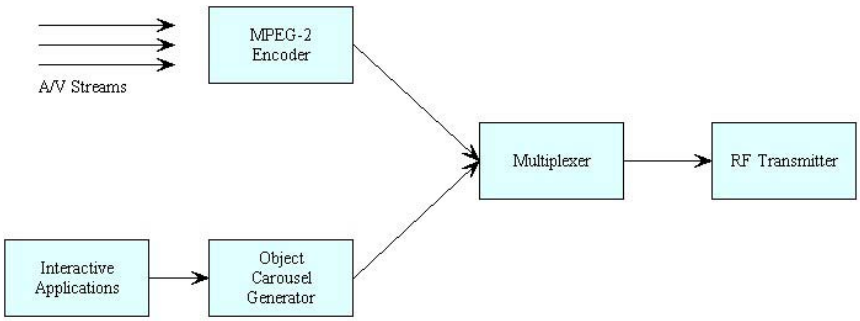
Audio Visual and Data Broadcast Technology

Digital television mostly relies on the Digital Video Broadcasting (DVB) standard, characterized as DVB-T for terrestrial, DVB-S for Satellite and DVB-C for Cable transmissions. DVB has been defined by a consortium of public and private organizations in the iDTV sector[3].

In the DVB schema, the digital TV signal is transmitted as a stream of MPEG-2 data known as a transport stream. This stream consists of a set of substreams (elementary streams), where each substream can contain MPEG-2 encoded audio, MPEG-2 encoded video or data encapsulated in MPEG-2 stream. The elementary stream which carries the application data is constructed using a Digital Storage Media-command and Control (DSM-CC) Object Carousel. Subsequently, the transport stream is passed to the multiplexer and then to a Radio Frequency (RF) transmitter in order to be broadcast. The overall broadcasting system for digital TV is illustrated in Figure 1.

The received signal is demodulated and afterward it has to be decoded appropriately. The common TV sets are manufactured to deal with analogue signals. Hence, a device called Set Top Box (STB) is used to transform the digital signal. Moreover, it also provides a middleware, based on an embedded Operating System (OS), which is an execution environment for running the interactive applications that are broadcast in a channel together with the main audiovisual stream. Execution environments are standard and the most common are: the European Multimedia Home Platform

Figure 1. Schema of the broadcast system



(MHP), the American Open Cable Application Platform (OCAP) and DTV Application Software Environment (DASE), the Japanese STD-B23/STD-B24. Because MHP is the standard in Europe and a subset of it, the Globally Executable MHP (GEM), is becoming the common reference worldwide, in this section we focus on MHP.

MHP

MHP is the middleware system for interactive TV development designed by the DVB Project[4]. The first draft of MHP was released in August 1999 and the first version of MHP 1.0 was approved by DVB in February 2000. MHP offers a standard platform for application developers. Applications are written in Java and HTML, so they don't depend on any single hardware platform or operating system. Due to the iDTV's special context, MHP-Java applications are slightly different from normal Java applications. However, due to the similarities with Java applets, MHP-Java applications are called Xlets.

On the one hand, MHP Java limitations are mainly related to the constraints given by the STB's hardware and OS in terms of computational power, memory size, storage, communication facilities, screen resolution, font and colour availability and their size is severely constrained by the limited bandwidth available. On the other hand, MHP provides support for those special features which are essential in the digital TV world such as low-level access to the transport stream, service information access, and support for the specialized graphics model of the digital TV. MHP can be extensively exploited by t-learning as it offers the proper middleware for learning interactive applications.

Tools for T-Learning

Based on the aforementioned technologies, a number of tools are appearing to support t-learning, such as authoring tools, games, personalization techniques and virtual tutor avatar.

Authoring Tools for T-Learning Courses

Authoring tools are software environments that support content providers in the creation of applications. There is a large number of authoring tools for the editing of e-learning courses in the market and in the last years, authoring tools for the creation of MHP applications appropriate for iDTV were developed.

On the e-learning side, authoring tools provide an environment that allows the insertion of learning resources (video, text, games, etc.) for creating the course. The structure of the course is organized usually with the definition of a sequence that relates these resources and allows different paths based on rules set by the author. The output of such kind of tools is appropriate to be used for learning purposes on a computer, but it is cannot be used for iDTV.

On the other hand, MHP authoring tools typically allow the creation of MHP-Java applications, suitable for iDTV but they are not learning-oriented. Up to now there are not any authoring tools dedicated specifically to the creation of t-learning courses. The tools that are used nowadays for t-learning purposes are the MHP authoring tools, which have to support the implementation of learning strategies suitable to TV-based education through the realization of MHP Java Xlets .

Games

Games in iDTV could play an important role to t-learning, although today they are used mostly for entertainment. Nowadays, there is a number of games for iDTV, covering various categories, such as arcade, adventure, puzzle and educational games. Quizzes, multiple-choice and memory games could increase the interest of the viewer-learner supporting the concept of relaxed-learning that seems suited to TV.

iDTV games are typically broadcast as applications resided inside the object carousel of a transport stream. Most of current games are stand-alone: they are not related with the A/V stream which is broadcast in parallel. This is because it is difficult to design and develop meaningful applications with a live, simultaneous interaction of games synchronized with the A/V stream. Even more important, synchronization drastically limits the use-timeframe of a game that can be played only at a given moment in time.

T-learning has the ambition of creating educational games for a wide range of users, in particular those with limited attitude to computers. The games are considered as an integral part of a t-learning course as they could support the learning procedure involving a wide audience through challenges and engaging activities that are anyway able to meet the typical user need for relax and sympathy.

Personalization

In general, the final goal of personalized learning is to provide a learning path that is matched to the learner's needs and abilities, resulting in a more efficient and high quality learning process. In order to obtain this matching of learner's profile and objectives, current learning context and available pedagogical resources, a well-defined description of each component involved in the process is needed, with specific focus on the user model. An additional interesting aspect of the personalization process is that, once the user model has been identified, the accuracy of the personalization can be iteratively improved with time, as more dynamic data are collected and stored regarding the ongoing interactions of the user with the system and the continuous monitoring and re-assessment of the user's satisfaction. This also allows for a classification and "clustering" of learners (Blanco-Fernández et al., 2004).

Personalization in terms of t-learning implies that a potential iDTV learner can easily be offered on his/her TV equipment a selection of available pedagogical contents and services according to his/her interests, skills and preferences.

From the conceptual and technological perspective, supporting personalization implies designing and developing suitable *services* to be integrated in the final t-learning application and able to provide contents and learning processes adapted to the user profile. The minimum set of such services includes:

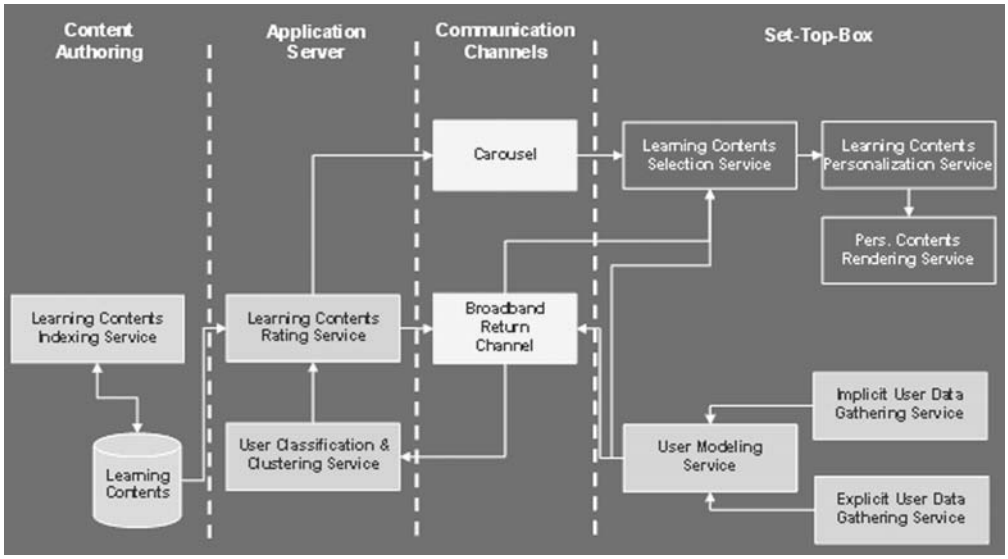
- **Implicit User Data Gathering Service:** Transparent acquisition of learner-related information through the automatic analysis of his/her behaviour while selecting/navigating learning contents.
- **Explicit User Data Gathering Service:** This service should offer to the user the possibility of providing personal information through an easy-to-use interactive interface.
- **User Modeling Service:** The information acquired by the two previous services need to be processed by the user modeling module, in order to produce a user profile that can be given as an input to other services in charge of automatically select and adapt available learning contents.
- **Users Classification and Clustering Service:** This service receives as input a set of user profiles, which are then classified and clustered. This would allow for the definition of "group" or "category profiles."
- **Learning Contents Rating Service:** The content rating engine is in charge of matching the existing available contents with the current user or category profile. The output is a list of relevant contents, among which the user is allowed to select interactively.
- **Learning Content Selection Service:** It allows for a selection of a subset of relevant contents coming from the carousel. This can be used to offer a limited degree of personalization on the Set-To-Box.
- **Learning Content Personalization Service:** When a specific learning object has been selected, a further personalization step is performed at the single content level, again according to the user or category profile.
- **Personalized Learning Content Rendering Service:** This service would be in charge of properly rendering the personalized content, for a better fruition by the user.

In Figure 2, a complex technological framework suitable to integrate, personalization services for iDTV-based knowledge management and t-learning/t-training applications is presented.

Besides the constraints coming from computational power and local storage features on the Set-Top-Box, the offered degree of interactivity in terms of available return channel has a strong influence on the personalization level. Having this in mind, three main profiles can be identified:

- STB without return channel (*basic profile*): includes local user profile processing, selection of relevant contents from the carousel and local processing of the selected content;

Figure 2. The personalization framework in t-learning



- STB with narrowband return channel (*advanced profile*): it offers the same features as the basic profile plus server side user profiles processing, dynamic adaptation of broadcast contents, server side synchronization of user behaviour, limited on-demand access to contents; and
- STB with broadband return channel (*full interaction profile*): same as advanced profile plus full on-demand access to contents through the return channel.

Virtual Teacher

A success factor in the distance courses is the simulation of an interaction between teacher and student. This can be achieved by using the figure of a tutor. It is generally accepted that the presence of the tutor is important for better understanding and motivating students. In particular, it was identified that a virtual teacher avatar could be a good means to achieve better audiovisual communication with the learner. An avatar is an intelligent agent with graphical representation, often humanoid and with speech capabilities (Ortiz, Aizpurua, Oyarzun, Arizkuren, & Posada, 2003).

Nowadays, avatars are used in the iDTV world mainly in online games in order to represent the viewer. However, they can also play the role of a virtual teacher (VT) in a t-learning course. The implementation of a virtual teacher for t-learning courses has to carefully react to the user behaviour and performance. Events can be generated from the AV stream and from other applications, and the VT may react to them on the basis of specific rules defined by the educational designer. In such a case, an events manager is needed by the system architecture to manage the incoming events and dispatch them to the VT.

User Interface and Synchronization Issues in T-Learning

During a t-learning course, the viewer-learner would be able to watch rich multimedia content accompanied by MHP applications. These applications can be either synchronized or not with the broadcast video. The screen can be divided in parts where every part could visualize different content as the video, the application or a virtual teacher including control instructions and help. Additionally, an application could overlay the A/V stream, also exploiting transparencies and semitransparencies. These display modalities can be supported in both synchronized and not synchronized courses. Synchronization of the video with the MHP application can be achieved with the insertion of triggers in the A/V stream that launch events to the listening application. In the case of a synchronized application, the viewer would have a specific amount of time to interact with the application. In a nonsynchronized application, the course would allow a more relaxed learning, but it would not benefit from synergies with the underlying A/V stream.

The user can provide inputs to an application through the remote control buttons. The arrows play an important role in the choice of preferences or the navigation, while the "OK" button would be usually used for confirmation purposes. The color buttons are typically used to provide access to

options. Text insertion is possible, in particular by using virtual keyboards on the screen, but its use is not encouraged as it is rather time-consuming and awkward because it is not directly supported by the remote control.

Architecture of an Integrated T-Learning System

The proposed architecture which is illustrated in Figure 3 is split in two parts: the production side where the content is prepared and the receiver side where the course is presented to the viewer through the appropriate terminal.

The production side is the area where the content and the applications are prepared. The educative A/V stream for the t-learning course is built by a TV producer, while the applications are developed in the authoring tool by the educators. The content including games, images, VT characters and text is used for the development of personalized courses and can be retrieved from a server where learning resources are stored. Eventual A/V-application synchronization is achieved with the aid of an authoring tool as well, where the content created is matched on specific time stamps inserted in the A/V stream. Subsequently, the A/V stream is fed into the MPEG2 encoder, while the content and the Xlet produced by the authoring tool are inserted into the object carousel. In this way the substreams are constructed and then multiplexed the final transport stream, which is broadcast.

The signal is received at the receiver side and processed by the STB where the A/V stream and the applications are restored from the transport stream. The Xlet that contains the t-learning course runs on the STB MHP middleware presenting the content of the course. The existence of an Internet IP return channel on the STB allows the use of on-demand features. Through this return channel it is possible to send requests regarding the retrieval of additional learning resources as well as information about the viewer in order to support more advanced personalization features.

FUTURE TRENDS

T-learning is based on modern technologies employed in Digital TV. Despite the fact that the world policies are supporting the switch-off from analogue TV to digital, there are still many analogue TVs in many homes and the people are not yet familiar with the new technologies and the interactive services introduced by iDTV. However, it is likely that the evolution of digital TV in the future years, accompanied by an increasing familiarity of users with its interactive services, will augment the penetration of t-learning into the households.

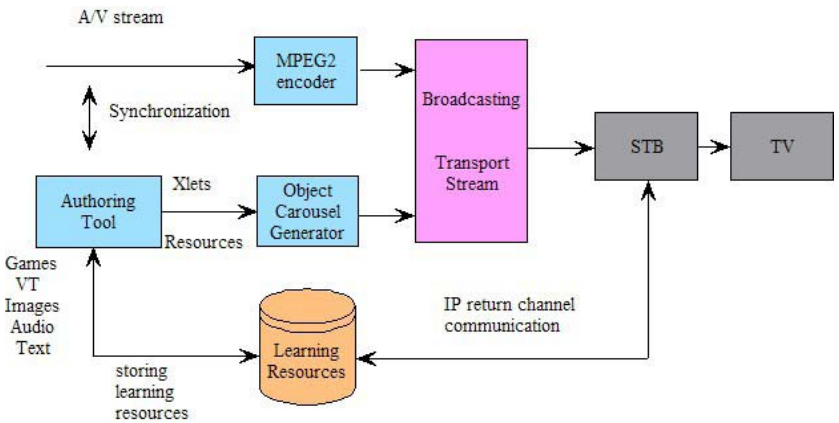
The spreading of the new technologies will further speed up the evolution of t-learning, as they will increase the recording, playing and computational capabilities of the TV-sets, increasing the scope and the quality of the offered services.

CONCLUSION

The most important technologies involved in t-learning were discussed and described in this article. These technologies are still evolving in order to support more efficiently t-learning and other interactive services for TV.

T-learning is a relatively new concept. It is a challenge for the world of interactive TV and distance learning with a potential to concretely complement e-learning, in particular given its potential to reach a much wider audience, and in

Figure 3. T-learning architecture system



different contexts (e.g., in relaxed situations). The ambition of more opportunities for learning in home may be fulfilled by t-learning as TV is still the medium that is present in every household. T-learning exploits the available technologies of iDTV and it can also spur them by introducing new needs and requirements.

The real potential of t-learning is promising, but the best ways to offer it to a wide audience and different target groups has still to be carefully investigated, in particular through extensive field experiments.

REFERENCES

Aarreniemi-Jokipelto P. (2005). T-learning model for learning via digital TV. In *Proceedings of the 16th EAEEIE Annual Conference on Innovation in Education for Electrical and Information Engineering (EIE),* Lappeenranta, Finland.

Baker, K., Pulles, R. & Sasno, P. (2004). Terminal architecture issues for interactive and reactive TV media using MHP-Java and MPEG4. In *Proceedings of the IEEE International Symposium on Consumer Electonics*, University of Reading, UK.

Blanco-Fernández, Y., Pazos-Arias, J. J., Gil-Solla, A., Ramos-Cabrer, M., Barragáns-Martínez, B., & López-Nores, M. (2004). A multi-agent open architecture for a TV recommender system: A case study using a Bayesian strategy. In *Proceedings of the Sixth IEEE International Symposium on Multimedia Software Engineering*, Miami, FL, USA.

Clark, R., Nguyen, F., & Sweller, J. (2006). Efficiency in learning: Evidence-based guidelines to manage cognitive load. John Wiley & Sons.

Kenyon B., Miles, A., & Rose J. (2000). Unscrambling digital TV. *The McKinsey Quarterly.*

Lekakos, G., & Giaglis, G. (2002). Delivering personalized advertisements in digital television: A methodology and empirical evaluation. In *Proceedings of the AH 2002 Workshop on Personalization in Future TV*, Malaga, Spain.

Lytras, M., Lougos, C., Chozos, P., & Pouloudi A. (2002). Interactive television and e-learning convergence: Examining the potential of t-learning. In *Proceedings of the ECEL2002: The European Conference on E-learning*, Brunel University, UK.

Ortiz, A., Aizpurua, I., Oyarzun, D., Arizkuren, I., & Posada J. (2003). ASEDUC–e-learning course delivery using 3D conversational avatars. *Computer Graphik topics, (4).*

Watheieu L., & Zoglio M. (2002). *TIVO: Case study*. Harvard Business School Review, Boston, USA.

KEY TERMS

Authoring Tool: Environment for applications creation without the need of programming and technical skills.

E-Learning: Distance learning with the aid of a personal computer.

iDTV: Interactive Digital TV is the evolution of the traditional TV set based on digital transmission and has the capability of running interactive applications.

MHP: Multimedia Home Platform is the common middleware for running applications for iDTV.

Personalization: The customization and categorization procedure of a viewer-learner.

Return Channel: Port that allows IP connectivity of the STB to support on demand features.

STB: Set Top Box is the device that decodes and processes the digital received signal.

Synchronization: This term is used to specify the time-matching between the t-learning content and the respective video.

T-Learning: Term that defines the TV-based interactive learning.

Virtual Teacher: Avatar that plays the role of a teacher during a t-learning course by providing instructions and help.

ENDNOTES

1 BBC, http://www.bbc.co.uk/learning/

2 Enhanced Learning Unlimited (ELU), http://www.elu-project.com

3 Digital Video Broadcasting (DVB), http://www.dvb.org

4 Interactive TV Web, http://www.interactivetvweb.org

Category: IT Education

Toward a Framework of Programming Pedagogy

Wilfred W. F. Lau
The University of Hong Kong, Hong Kong

Allan H. K. Yuen
The University of Hong Kong, Hong Kong

INTRODUCTION

As a major topic in information technology education, computer programming has been taught to both major and non-major students in universities. While there has been ongoing debate on whether students should be taught programming (Soloway, 1993), the literature shows that learning to program poses a lot of difficulties to novices (Bonar & Soloway, 1989). Dijkstra (1989) describes programming as "radical novelty" in which our usual strategy of metaphors and analogies simply does not apply. Pea (1986) identifies three types of conceptual bugs which are rooted in a superbug where "there is a hidden mind somewhere in the programming language that has intelligent interpretive powers".

Why is learning to program so difficult? One difficulty is that learning to program needs the acquisition of a multitude of inter-related skills. Jenkins (2002) argues that programming is a complicated task, which requires the mastery of a number of skills such as problem solving, abstraction, mathematical logic and testing, debugging and so forth. A novice programmer simply lacks these skills. More importantly, success in learning to program demands knowledge of computer itself. Ben-Ari (1998) points out that students lack a viable mental model to learn programming. On the other hand, undue emphasis is placed on the learning of programming syntax (Deek, 1999). In this article, we will focus on approaches of teaching computer programing. Winslow (1996) introduced the term "programming pedagogy" in his paper. Although programming pedagogy is not explicitly defined in the paper, the term here refers to any instructional methods and strategies which are used to teach students introductory programming. Due to these reasons, programming pedagogy calls for special attention.

BACKGROUND

Over the years, pedagogical innovations have been proposed to cope with these difficulties. These include a variety of programming tools (Smith & Webb, 2000). These tools help novice programmers to develop programs through program visualisation and algorithm animation. Deek and McHugh (1998) evaluate programming tools used to teach programming. One common problem among these tools is that they fail to integrate into the curriculum. This suggests that there is a need to investigate what programming pedagogy should be adopted together with tools to bring about innovations in programming instruction. This article intends to review on programming pedagogy reported in the literature. A theoretical framework on programming pedagogy grounded on literature review is proposed which attempts to conceptualise pedagogy in terms of the cognitive and technological dimensions. For researchers, this review not only provides a summary of pedagogy adopted to date, but also theoretical underpinning for future research on programming pedagogy. For practitioners, the proposed framework can help them to evaluate and reflect on their own pedagogy with an aim to improve the quality of teaching and learning computer programming.

APPROACHES OF TEACHING PROGRAMMING

We identified seven pedagogical approaches of teaching computer programming arising from the review of appropriate literature. The following sections provide a brief description of each of these approaches.

Structured Programming Approach

In the 1970s, one catchword in programming is structured programming. It is an approach, which intends to "support the production of correct, understandable programs which are easy to modify and maintain" (Freiburghouse & Liskov, 1973). The approach allows control structures of sequence, selection, and repetition only. The GOTO statement is considered detrimental to structured programming (Dijkstra, 1968). To facilitate the development of a structured program, a top-down design, which decomposes a large program into a manageable smaller program, is used. Programs are improved successively through stepwise refinement. In this

manner, it is hoped that quality program can be produced. Although it was advocated in the 1970s, it is still one of the prevalent programming approaches today.

Problem Solving Approach

Barnes, Fincher, and Thompson (1997) describe a programming methodology consisting of four steps, namely, Understanding, Designing, Writing and Reviewing. Following the same line of thinking, Thompson (1997) proposes a problem solving approach in teaching functional programming. He claims that using the approach, "a novice can make substantial progress in completing a programming task before beginning to write any program code." Gries (1974) emphasizes the importance of problem solving in programming. He argues that usual assumption that students should have learned programming after giving tools and examples is not pedagogically sound. To address this problem, he suggests the four-phase process of problem solving by Polya (1957).

Software Development Approach

It is equally important that students should know how to translate algorithms into syntactically and semantically correct solution of the problem that form the program. In this regard, Deek (1999) develops a methodology which incorporates both the problem solving skills and the programming skills into a single process that provides a framework for beginning students. As noted by Deek (1999), there are three kinds of difficulties faced by students when learning to program: (1) deficiencies in problem solving strategies and tactical knowledge; (2) ineffective pedagogy of programming instruction; and (3) misconceptions about syntax, semantics, and pragmatics. The new approach, which incorporates both the problem solving skills and the programming skills, can help address all the three kinds of difficulties.

Small Programming Approach

If programming creates a large cognitive load on novices, it is reasonable to reduce such load by programming in a "small" scale. In this sense, we distinguish between the terms Programming-in-the-small and Programming Language-in-the-small.

Glaser, Hartel, and Garratt (2000) introduce the idea Programming by Number in teaching ML and Java. Programming by Number is intended to get students started in writing program by providing a step-by-step guidance to students while allowing flexibility in the design of the solution to a problem. When writing functions, they suggest the following steps: (1) name the function; (2) write down its type; (3) enumerate all cases; (4) deal with any simple case(s); (5) list the ingredients in preparation for the complex case(s); (6) deal with the complex case(s), where some inspiration is required; and (7) think about the result.

Brusilovsky, Kouchnirenko, Miller, and Tomek (1994) review on three approaches of teaching introductory programming, namely, the incremental approach, the mini-language approach, and the sub-language approach. In the incremental approach, new language subsets, which introduce new programming language constructs while retaining all the constructs of preceding subsets, are introduced successively to novices. In the mini-language approach, a small and simple language is used to support the first steps in learning to program. In most cases, a student learns how to program by controlling an actor, which can be a turtle, a robot, or any other active entity, in a microworld. The sub-language approach uses a special starting subset of the full language which contains easily visualizable operations to introduce programming to novices. In short, these three approaches provide a simple and small language subsets and a visually appealing metaphor embedded in a context-rich environment to help novices start programming.

Language Teaching Approach

Robertson and Lee (1995) give a research manifesto for the application of a second natural language acquisition pedagogy to the teaching of programming languages. They argue that programming has traditionally taught with little reference to natural language pedagogy. To conclude, they provide some areas for further research such as the value of reading programs before writing, the use of authentic programs, the study of the cultural milieu of programs, and so forth. Baldwin and Macredie (1999) argue that research in the learner strategies in second language pedagogy may provide insight into programming pegadogy. Based on the call for a more learner-centred environment, they believe that learner strategies are one of the issues in teaching programming that can help address difficulties in learning to program. Deek and Friedman (2001) describe their ideas of how programming and writing are learned in parallel. They argue that the common element that exists in both domains, problem solving and program development, provides "new ways for students to transfer skills between domains".

Learning Theory Approach

Lister and Leaney (2003) argue that traditional norm-referencing approach to grading tends to target at average students. As a result, weaker students cannot program well and stronger students are not challenged. They suggest a criterion-referencing approach to grading so that explicit and clear criteria are set for each grade. In deciding the criteria, reference is made to the Bloom's Taxonomy of Educational Objectives (Bloom, 1956). Macfarlane and Mynatt (1988) examine the effectiveness of advance organizer in teaching

the syntax of arrays. The tasks of the subjects were to enter syntactically and semantically correct Pascal statements to manipulate arrays in a training program. The study used a metaphorical advance organizer–a one-story apartment building. Results revealed that the groups did not differ on syntactic knowledge (near transfer). Yet the advance organizer group outperformed the other two groups in semantic knowledge (far transfer).

Other Approaches

There are other innovative attempts in programming pedagogy. Astrachan and Reed (1995) propose an apprenticeship model of learning in which students begin by reading, studying, and extending programs written by expert programmers. Through this approach, students learn from real-world examples, practice code, and concept reuse. Jenkins (1998) attempts to get students actively participate in learning to program. His participative approach aims to make the learning of some abstract programming concepts such as procedures, parameters, pointers, and linked lists more accessible to students. Proulx (2000) introduces the use of programming patterns in learning to program. The idea is to organize the common used methods like reading data in programming as patterns for reuse.

DISCUSSION

In this section, pedagogical remarks are made on these approaches. Then, it is followed by the discussion of a proposed theoretical framework which attempts to conceptualise programming pedagogy in terms of two dimensions.

Empirical Evaluation

It is common that many approaches lack classroom evaluation of their effectiveness. In many cases, they are simply ideas based on the experience of the practioners in teaching. While some studies show empirical evaluations of their approaches (Glaser et al., 2000; Macfarlane & Mynatt, 1988), many simply do not provide sufficient evidence to support their claims. For instance, Proulx (2000) reports that, "Overall, we felt the course was a success. Students performed better on the midterm exam and seemed more confident than in the past". Yet, no information on the students' background or even results of the mid-term examination are provided. Nevertheless, it is of paramount importance that any claims on the effectiveness of an approach should be grounded on empirically tested evidence (Deek & McHugh, 1998) in pedagogical practices.

Support from Learning Theory

If approaches are not grounded on empirical data, an established theory may serve as the foundation for the approaches. However, not many approaches have explicitly mentioned the use of any learning theory in its design. The risk of not basing the approach on established theory can be detrimental since the approach remains ad hoc and may not be generalized to other settings. As pointed out by Berglund (2002), the problem is, "Although valuable as a mean of sharing experiences between computer science educators, the results are…often hard to generalise, since they are not based on pedagogically sound theories of learning or carried out with sound methodological principles".

Individualization

One of the foci of our curricula in the school education today is learner-centerness. It is our responsibility to provide such an environment for the needs of individuals. In terms of pedagogy, the usual strategy adopted is "one size fits all". This results in weaker students learn well while stronger students are unchallenged (Lister & Leaney, 2003). It is argued that learning styles provide a vehicle through which individual differences are addressed. The literature has shown that learning styles do affect programming performance (Ross, Drysdale, & Schulz, 2001). It is thus possible to make use of learning style data to individualize pedagogy to bring about innovation in teaching.

Towards a Conceptual Framework

Learning to program is recognized as a cognitively demanding process in the literature. It involves expressing solutions to problems in terms of programming language syntax as well as understanding of how programs work in the computer. In the former case, Linn and Dalbey (1985) propose a chain of cognitive accomplishments for computer programming instruction consisting of three main stages: (1) single language features, (2) design skills, and (3) general problem-solving skills. Besides, Bayman and Mayer (1988) propose three interrelated types of programming knowledge necessary to understand the underlying complex processes involved in programming: syntactic, conceptual, and strategic. Such discussion indicates a developmental change from learning programming syntax to learning programming concepts in terms of programming knowledge. Recently, technology advancement has made computer programming more visually than before. This is achieved in one way by visual programming (Green, 1995). It points to a dimension in programming pedagogy in terms of programming representation.

In this article, an initial framework on programming pedagogy is suggested which attempts to conceptualise

pedagogy along the two dimensions, namely, programming knowledge and programming representation. For the dimension of programming knowledge, the two poles are programming syntax and programming concept, which correspond to the developmental change in learning programming. For the dimension of programming representation, the two poles are textual representation and visual representation, which reflects the advancement of technology in programming presentation. Figure 1 shows a classification of aforementioned programming approaches under this framework. Structured programming is traditionally learned with focus on programming syntax in textual representation. As a result, it is considered as a more syntax-oriented and textual-based (i.e., syntax-textual pedagogy). Problem Solving, Software Development, Language Learning, and Learning Theory approaches are regarded as more concept-oriented and textual-based (i.e., concept-textual pedagogy) since they emphasise the mastery of problem solving skills. Mini-language and Sub-language approaches usually make use of visual metaphor to help beginners to start programming. In this sense, they focus more on concepts building through a visual programming representation (i.e., concept-visual pedagogy). However, we could not find any approach which is under syntax-visual pedagogy. With a trend towards user-friendliness and technology development, it is anticipated that there will be a gradual shift from the concept-textual pedagogy to concept-visual pedagogy in the future.

FUTURE TRENDS

Guzdial and Soloway (2002) advocate the use of authentic multimedia construction to engage students to learn programming. Jimenez-Peris, Pareja-Flores, Patino-Martinez and Velazquez-Iturbide (2000) foretell a day in a 2020 university, saying that, "current lab programming environments are truly educational, and they are highly visual and intuitive. They detect and accurately diagnose most errors…" From these observations, in the future, learning to program will be probably dominated by a visual and media-rich environment, and this is in agreement with our guess.

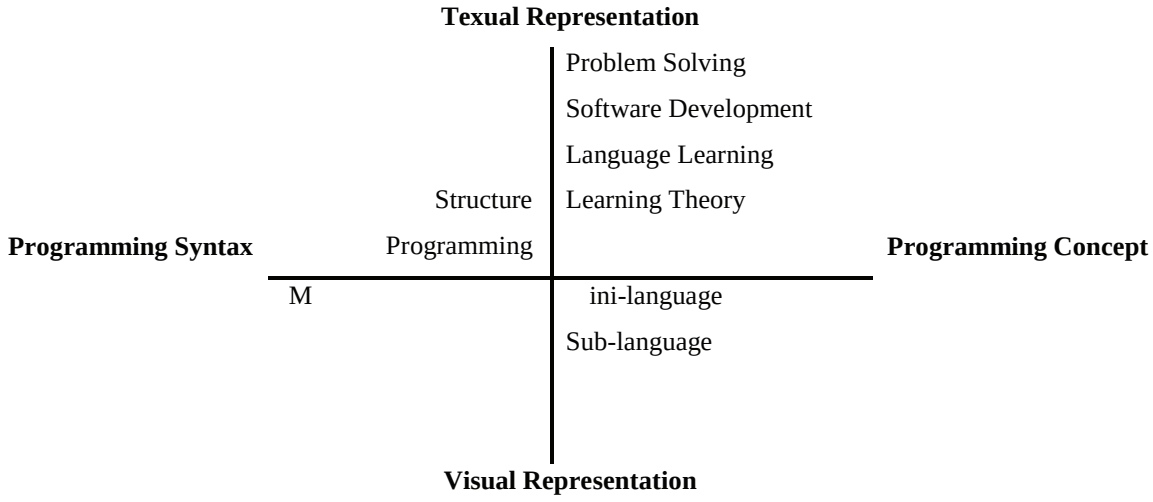
CONCLUSION

The fact that programming is difficult to novices is reported enormously in the literature. Pedagogical innovations have been suggested to address the issue. In this article, programming pedagogy is reviewed. Approaches of teaching programming are discussed and grouped under seven categories. Problems associated with these approaches are discussed. To conceptualise pedagogy, a theoretical framework consisting of the programming knowledge and programming representation dimensions is proposed. An attempt has been made to classify the approaches under this framework. While the approaches discussed are not meant to be exhaustive, it is hoped that this article may shed light on the discussion of programming pedagogy and arouse further exploration and research in programming education.

REFERENCES

Astrachan, O., & Reed, D. (1995). *AAA and CS 1: The applied apprenticeship approach to CS 1*. Paper presented at the Twenty-Sixth SIGCSE Technical Symposium on Computer Science Education, Nashville, Tennessee.

Figure 1. Classification of programming pedagogy



Baldwin, L. P., & Macredie, R. D. (1999). Beginners and programming: Insights from second language learning and teaching. *Education and Information Technologies, 4*(2), 167-179.

Barnes, D. J., Fincher, S., & Thompson, S. (1997). Introductory problem solving in computer science. In G. Daughton & P. Magee (Eds.), *5th Annual Conference on the Teaching of Computing* (pp. 36-39). Ireland: Dublin City University.

Bayman, P., & Mayer, R. (1988). Using conceptual models to teach BASIC computer programming. *Journal of Educational Psychology, 80*(3), 291-298.

Ben-Ari, M. (1998). *Constructivism in computer science education*. Paper presented at the Twenty-Ninth SIGCSE Technical Symposium on Computer Science Education, Atlanta, GA.

Berglund, A. (2002). *On the understanding of computer networks*. Unpublished Licentiate thesis, Uppsala University.

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives, Handbook I: Cognitive domain*. New York: Longmans.

Bonar, J., & Soloway, E. (1989). Uncovering principles of novice programming. *Communications of the ACM*, 10-13.

Brusilovsky, P., Kouchnirenko, A., Miller, P., & Tomek, I. (1994). *Teaching programming to novices: A review of approaches and tools*. Paper presented at the ED-MEDIA 94 - World Conference on Educational Multimedia and Hypermedia, Vancouver, BC, Canada.

Deek, F. P. (1999). The software process: A parallel approach through problem solving and program development. *Computer Science Education, 9*(1), 43-70.

Deek, F. P., & Friedman, R. (2001). Computing and composition: Common skills, common process. *Journal of Computer Science Education - ISTE SIGCS, 1*, 8-14.

Deek, F. P., & McHugh, J. (1998). A survey and critical analysis of tools for learning programming. *Computer Science Education, 8*(2), 130-178.

Dijkstra, E. W. (1968). Letters to the editor: Go to statement considered harmful. *Communications of the ACM, 11*(3), 147-148.

Dijkstra, E. W. (1989). On the cruelty of really teaching computer science. *Communications of the ACM, 32*, 1398-1404.

Freiburghouse, R., & Liskov, B. (1973). *Report of session on structured programming*. Paper presented at the ACM SIGPLAN - SIGOPS Interface Meeting on Programming Languages - Operating Systems, New York.

Glaser, H., Hartel, P. H., & Garratt, P. W. (2000). Programming by numbers: A programming method for complete novices. *The Computer Journal, 43*(4), 252-265.

Green, T. R. G. (1995). Noddy's Guide to Visual Programming. Interfaces (Newsletter of the British Computer Society Human-Computer Interaction Group).

Gries, D. (1974). *What should we teach in an introductory programming course?* Paper presented at the Fourth SIGCSE Technical Symposium on Computer Science Education, New York.

Guzdial, M., & Soloway, E. (2002). Teaching the Nintendo Generation to Program. *Communications of the ACM, 45*(4), 17-21.

Jenkins, T. (1998). *A participative approach to teaching programming*. Paper presented at the 6th Annual Conference on the Teaching of Computing and the 3rd Annual Conference on Integrating Technology into Computer Science Education, Dublin City University, Ireland.

Jenkins, T. (2002). *On the difficulty of learning to program*. Paper presented at the 3rd Annual LTSN-ICS Conference.

Jimenez-Peris, R., Pareja-Flores, C., Patino-Martinez, M., & Velazquez-Iturbide, J. A. (2000). New technologies in computer science education. In T. Greening (Ed.), *Computer science education in the 21st century* (pp. 113-136). New York: Springer-Verlag.

Linn, M. C., & Dalbey, J. (1985). Cognitive consequences of programming instruction: instruction, access, and ability. *Educational Psychologist, 20*, 191-206.

Lister, R., & Leaney, J. (2003). *Introductory programming, criterion-referencing, and bloom*. Paper presented at the 34th SIGCSE Technical Symposium on Computer Science Education, Reno, NV.

Macfarlane, K. N., & Mynatt, B. T. (1988). *A study of an advance organizer as a technique for teaching computer programming concepts*. Paper presented at the Nineteenth SIGCSE Technical Symposium on Computer Science Education, Atlanta, GA.

Pea, R. D. (1986). Language-independent conceptual bugs in novice programming. *Journal of Educational Computing Research, 2*(1), 25-36.

Polya, G. (1957). *How to solve it* (2nd ed.). NJ: Princeton University Press.

Proulx, V. K. (2000). *Programming patterns and design patterns in the introductory computer science course*. Paper

presented at the Thirty-First SIGCSE Technical Symposium on Computer Science Education, Austin, TX.

Robertson, S. A., & Lee, M. P. (1995). The application of second natural language acquisition pedagogy to the teaching of programming languages: A research agenda. *SIGCSE Bulletin, 27*(4), 9-12.

Ross, J. L., Drysdale, M. T. B., & Schulz, R. A. (2001). Cognitive Learning Styles and Academic Performance in Two Postsecondary Computer Application Courses. *Journal of Research on Computing in Education, 33*(4), 400-412.

Smith, P., & Webb, G. (2000). The efficacy of a low-level program visualization tool for teaching programming concepts to novice C programmers. *Journal of Educational Computing Research, 22*(2), 187-215.

Soloway, E. (1993). Should we teach students to program. *Communications of the ACM, 36*(10), 21-24.

Thompson, S. (1997). Where do I begin? A problem solving approach to teaching functional programming. In K. Apt, P. Hartel, & P. Klint (Eds.), *First International Conference on Declarative Programming Languages in Education*. Springer-Verlag.

Winslow, L. E. (1996). Programming pedagogy: A psychological overview. *SIGCSE Bulletin, 28*(3), 17-22.

KEY TERMS

Computer Programming: Mainly consists of designing the underlying algorithm and representing that algorithm as a program.

Curriculum: Consists of four interacting components, namely, teaching objective, teaching material, teaching method, and assessment.

Novice Programmer: A computer programmer who is not experienced at programming.

Problem Solving: To solve, problem is to find a way where no way is known off-hand, to find a way out of a difficulty, to find a way around an obstacle, to attain a desired end that is not immediately attainable by appropriate means. The emphasis of problem solving in programming instruction focuses on the expression of solutions in terms of mathematical functions.

Programming Pedagogy: The study and theory of the methods and principles of teaching and learning computer programming. It refers to any instructional methods and strategies which are used to teach students introductory programming.

Textual Programming: The representation of programming processes is based on written texts.

Visual programming: The representation of programming processes is based on graphics and images. It allows the manipulation of visual information and provides support for visual interaction.

Toward Societal Acceptance of Artificial Beings

Daniel I. Thomas
Technology One Corp,, Australia

Ljubo B. Vlacic
Griffith University, Australia

INTRODUCTION

Modern organizations are faced with many challenges with the trend toward distribution of their workforce across the planet. With this situation becoming more common, it is important for organizations to find ways of encouraging effective leadership and strong teamwork. Training and evaluation of the effectiveness of those employed can be an expensive exercise due to geographic separation of the parties involved. To this end, we propose a collaborative play scenario, using humans and artificial beings as *fully equal partners* (FEPs), to facilitate training and evaluation of a dispersed workforce.

While this scenario is a simple example of collaboration among human and artificial entities, moving this concept forward in other application areas creates questions about how artificial entities influence outcomes in the context of group decision making. The idea of social influence and acceptance of artificial beings as equal decision makers is explored, and how they may integrate into larger societies.

In this article, we present a simple training exercise designed to test a candidate's leadership ability to negotiate with other members of the organization, using their influence to achieve (partially or fully) their goals. While in practice, the play scenario would consist of combinations of human and artificial beings, the training scenario presented shall consist solely of artificial FEPs in order to demonstrate how influence can affect a result in a collaborative process.

BACKGROUND

Artificial Beings as Fully Equal Partners

When the average person is confronted with the term "Artificial Intelligence" it is more likely to conjure images of science fiction than science fact (Khan, 1998). Yet throughout our daily lives, we experience various degrees of artificial intelligence in such mundane devices as washing machines and refrigerators. Beyond this, organisations have been using intelligent systems in a myriad of endeavours. Beyond today however, "We may hope that machines will eventually compete with men in all purely intellectual fields" (Turing, 1950, p 460). Turing's remarks may not be fully realized today; however, the integration of artificial beings into human organizations and society evoke powerful images of both positive and negative possibility.

One possibility is artificial beings emerging as partners rather than tools in various collaborative situations. Unlike past revolutions of mechanical automation, the presence of artificial beings should not imply a redundancy for human partners, but rather a complimentary relationship. Group decision making, including both humans and artificial beings as equals, increases the diversity of the knowledge pool (Dunbar 1995), improving the likelihood of positive outcomes.

In order for artificial beings to be realized as collaborative partners, as opposed to an intelligent tool, they must be able to articulate their perspectives and opinions, while taking onboard the knowledge and opinions of others. For this to occur, artificial beings require a degree of social influence. For this influence to occur, the artificial being needs to become acceptable within the social system: Society, organization or group (Kelman, Fiske, Kazdin, & Schacter, 2006). In making the transition to societal acceptance of artificial beings, there are great challenges, both technical and social. To better study artificial beings as collaborative partners, it is possible to focus on a smaller, group social setting, with an assumption of social acceptance (and therefore the capability to influence) collaborative group decision making. For this reason, computer games provide an excellent environment for understanding how humans and artificial beings can positively influence outcomes in a collaborative group situation.

Much of our work into collaboration has been influenced by the use of intelligent autonomous agents in computer games. Jennings and Wooldrige (1995) describe an intelligent agent as one that enjoys the attributes of autonomy, situatedness, social ability, reactivity and proactiveness.

Basing intelligent entities around this core concept of agency has led researchers such as Laird (2001) and Kaminka et al. (2002) to create intelligent opponents for human players.

Taking this a step further, we see future applications for intelligent artificial beings as more than just opponents or nonplayer characters (called NPCs) in computer games, but rather we see artificial beings being utilised as *fully equal partners*.

Extending these concepts of humans and artificial entities interacting collaboratively in computer games, it is necessary to define a type of entity that:

- Does not treat human and artificial players differently during interaction;
- Can work cooperatively with other fully equal partners (including humans);
- "Plays" the game as a human would;
- Does not work to a defined script or take direction from an agent "director" such as those described by Magerko et al. (2004) and Riedl, Saretto, and Young (2003) and;
- Is not necessarily aware of the nature of other FEP beings (human or artificial in nature).

Simply, a *Fully Equal Partner* (or *FEP*) is an intelligent entity that performs tasks cooperatively with other FEPs (human or artificial), but is also capable of being replaced one with another. These beings are not necessarily aware of the nature of their fellow partners.

A Collaborative Architecture

In order to facilitate the collaboration among fully equal partners, the involved computer games must support a number of key features (Thomas & Vlacic, 2003), including:

i) A clean and well-defined interface or separation between the beings and the game (Vincent et al., 1999);
ii) A concept of time and causality; and
iii) Support for experimentation (Cohen, Hanks, & Pollack, 1993).

To create a computer game that enjoys many of these features, Thomas and Vlacic (2005) developed a layered architectural approach to collaborative games. Collaborative computer games that involve FEPs have three architectural layers: A communications, a physical and a cognitive layer.

The communications layer is essentially the protocols and low-level software that facilitate interaction and communication. The physical layer describes items and entities within the game world and how they may be manipulated. The cognitive layer describes the processes required to facilitate intelligent collaboration.

Collaborative process

If considered at a high level, the collaborative process *c* involves taking a set of fully equal partners *P* with a set of goals *G* and producing a set of outcomes *O*. An outcome may not necessarily satisfy the set of goals (e.g., a failure outcome).

$O = c(P,G)$

In order to obtain these outcomes to the collaborative process, FEPs engage in conversations. The result of these conversations are pieces of group collective knowledge *K*; that is, knowledge that is known to the group. Outcomes of the collaborative group are a result of the collaborative process between the group of FEPs and the goals of the collaborative process.

$$O = c(P,G)$$
$$O = c(P,G)$$
$$O = \{o_1,..,o_n\}$$
$$o_n = s(P,n(G,K^P))$$

where *s* is a function of all partners *P* applied to an interpretation function *n* of the set of goals *G*, the set of group collective knowledge across the entire set of partners K^P, resulting in an outcome o_n.

INFLUENCE IN THE COLLABORATIVE PROCESS

In human to human interactions, we see many forces at play that influence one person to agree or take the side of another in a discussion. These influences need to be taken into account when collaborative work is undertaken. Even the size of a group (Fay, Garrod, & Carletta, 2000) can change the way in which partners are influenced, and by whom.

Collaborative FEPs may create an affinity with one or more entities and are more likely to accept their position during negotiation. Possible methods for obtaining an affinity with one or more FEPs include:

1. The degree to which one FEP's responses convey a perception/opinion that matches that of another FEP. The more that one partner's position matches that of another partner, it becomes more likely that the partner will "trust" the statements of that partner.
2. Some arbitrary/authoritative influence factor that has the partner tending toward the position of one or more

other partners. An example in a business sense might be seniority or position within an organization.

3. Trade: changing a given position in order to influence another partner's position on another item in the collaborative process.
4. A pre-existing relationship (e.g., a friendship) that exists beyond the scope of the collaborative process.

The scope of the play scenario discussed shall focus on two forms of influence, that of arbitrary/authoritative influence and common interest.

During the collaborative process, any partner p_l, where $l \neq k$, may ask a question q_m of any other partner p_k in order to receive a response r_m, where $m = j+i$

$$r_m = f(p_k, q_m)$$

The set of Group Collective Knowledge K obtained by the group through the collaborative process is a subset of the responses obtained during the collaborative process.

$$K \subseteq R$$
$$K \subseteq \{r_1,..,r_m\}$$
$$\{k_1,..k_q\} \subseteq \{r_1,..,r_m\}$$

To influence group collective knowledge, resulting responses r_m that contribute to K must be changed in some way. Assuming that all responses contribute to collective knowledge:

$$K = R$$
$$\{k_1,..,k_q\} = \{r_1,..,r_m\}$$
i.e., $q = m$

During the negotiation phase, an influence function changes the response for a given partner's initial decision based on the degree of influence the other partners have with the first partner. The influence function that is used here is simply the sum of the proportion difference between one partner's response (obtained during the conversation process) and that of another partner:

$$i(r_{pk}) = \sum_{1-n,n\neq i} p_{nf} * (r_{pn} - r_{pk})$$

Where $i(r_{pk})$ is the influenced response which is the sum of all influence factors p_{nf} multiplied by the response difference partner p_n and the partner under influence p_k. FEPs using this influence function cannot influence themselves.

PLAY SCENARIO: PROJECT PLANNING

In order to investigate influence in the collaborative process, a play scenario was devised based upon a documented real-world project planning process from industry. The scenario is used as a training exercise for management candidates to investigate their negotiation skills in representing their business group's interests in a collaborative situation.

The scenario involves a particular enterprise software developer that utilises a common infrastructure upon which all its enterprise solutions are based. When a new version of the infrastructure is to be developed, the managers and architects of this business unit formulate a list of potential projects that can be pursued within the next version's development timeframe. Each project has an estimated development time budget.

Unfortunately, the list of potential projects greatly exceeds the number of development days available within the new version development period.

In order to satisfy the needs of the enterprise product business groups, the manager responsible for infrastructure therefore contacts the managers of the various enterprise products to obtain their feedback on what projects are most desirable in their respective areas.

Because there are timeframe limitations, and the enterprise products address different business needs, a significant amount of time is involved in negotiating a "best fit" for all parties. As such, each of the business group managers wants to influence the decisions of the others in order to influence the inclusion in the development plan of their highest priority requirements.

Table 1 documents the business units involved.

The Experiment

The play scenario is calibrated against the original business data that was collected prior to the meetings. Each business group was requested to review the list of potential projects (88 in total) and place a numbered priority next to each item until the development days from the items totaled approximately 400 days.

When a candidate undertakes the training scenario, they must represent their business group's interests, maximising the number of inclusions within the project.

Unlike a regular play scenario, the following exercises have been undertaken with artificial FEPs only, in order to investigate their effectiveness in influencing the final outcomes. The following influence methods were undertaken:

1. No Influence (baseline);
2. Arbitrary Influence; and
3. Common Interest.

The participating FEPs utilize a collaborative fuzzy logic process (Thomas & Vlacic, 2007) to determine the responses given during the collaborative process.

Determining Influence

In order to determine influence between the FEPs involved, the following methods were used.

An arbitrary percentage influence from 0% to 10% was allocated to each business unit representative. As the collaborative process progressed, the responses by each representative were influenced by the other 10 member's influence. The breakdown of the allocated influence was:

Common interest influence was determined by informing each FEP of the other member's "Top 5" projects. If a project matched one of the other's in the top five, an influence percentage was awarded between 1% and 5%. This influence was allowed to compound if more than one project was in common. A 5% influence was awarded when there was a total match (i.e., Partner *A*'s first project choice is the same as Partner *B*), reducing to 0% the further away a project was in terms of order (i.e., Partner *A*'s project is ranked 3, and partner *B*'s project is ranked 5; therefore, the influence is 3%). Table 2 documents the common influence factors across the business units.

Results

After completing each run of the play scenario, the results were compared to the actual results collected during the industry process.

Across the 88 projects identified, approximately 70% of projects were common to all results. It can also be seen from the captured results that influence does impact the final makeup of the projects to be included within the allowable project budget timeframe.

While the percentage of match with the industry results is fairly high, of more interest are the reasons of difference between the industry and the play scenarios. As stated earlier, factors such as trade and pre-existing relationships were beyond the scope of this experiment. In addition, given the business unit group size of 11, they may also be factors of influence such as the group size (Fay et al., 2000). Subsequent play scenarios will require the incorporation of these additional factors. Closing this "collaborative gap" between a human social group and an artificial social group with the same goals and objectives is a compelling area of research, and a large step along the road to achieving social acceptance of artificial beings as FEPs.

Actual Results

As noted earlier, the data collected for this play scenario is sourced from real-world business data. As such, the collected data does not have the rigour true experimental collection. For example, the business groups involved in the real process interpreted the ranking instructions sent out by the infrastructure team in different ways. While the scenario itself was a compelling real-world case, further work on industry scenarios will require more explicit information to ensure the resulting data collection is of experimental quality.

FUTURE TRENDS

As artificial entities become more widely used in collaborative cognitive applications, negotiation skills become more important. There are many applications where the use of collaborative human and artificial FEPs may be utilised. Broadly speaking, to apply these concepts can be applied to any collaborative decision making scenario, as long as the following criteria are well-defined: The scenario's objectives and goals must be established; the communication, physical and cognitive layers are clearly defined; and the collaborative process is applied as a framework to facilitate the collection and application of group collective knowledge in order to achieve the defined objectives/goals.

In the enterprise, collaborative FEPs can be applied in training scenarios such as the project planning exercise. Beyond this, FEPs can also play an effective role in collaborative business intelligence and strategic planning. Stepping out of the enterprise, the nature of FEPs being human or artificial, allows artificial FEPs to find application in automated transport systems that can "collaborate" with other human and artificial road users. In addition, there are compelling reasons for artificial FEPs to be used in pure entertainment applications, by allowing computer game developers to create richer, more interactive experiences. Interestingly, collaborative FEPs could facilitate the study of human behaviour and interaction, through the creation of safe environments for behavioural studies and clinical psychology (we strongly believe that it may be used for the purpose of assisting patients in overcoming difficulties of expression and engaging in collaborative play scenarios).

For artificial beings emerging as equal partners in society, the future becomes more speculative. As stated earlier, artificial beings need to move from the position of an intelligent tool to some form of acceptance of equality, in order to become "collaboratively equal" to their human counterparts. For this kind of acceptance to occur, FEP applications such as business analysis, transportation and entertainment need to form the groundwork for basic acceptance.

Beyond overcoming the technical challenges of integrating artificial beings into collaborative partnerships

with humans, there are also the challenges of acceptance. Preconceived notions of science fiction (Khan, 1998) and cultural perceptions between different cultures (Kaplan, 2004) are some of these challenges. The positive benefits of collaborative artificial FEPs also faces challenges of negative perceptions within different social groups and societies. Addressing these issues effectively shall ensure that collaboration among human and artificial beings is an effective method of group decision making.

CONCLUSION

Human and artificial beings as Fully Equal Partners (FEPs) offer compelling applications in industry. There are many challenges faced in the integration and acceptance of artificial FEPs into the social groups of humans. One challenge described was the capability of artificial FEPs to influence group decision making outcomes. FEPs and the collaborative process were described, showing how these are applied in a simple training play scenario. The negotiation processes of the artificial FEPs were presented and how different forms of influence may be used to affect a collaborative outcome. The outcome of the play scenarios indicates a "collaborative gap" between humans and artificial beings. Further work on reducing the size of this gap will contribute to the acceptance of artificial beings as collaborative entities within the group decision making process.

Beyond this play scenario and its application in training, FEPs offer a wide variety of applications across varied business sectors including training, business analysis, transportation, entertainment and behavioural studies. By understanding the scenario objectives, architectural layers and collaborative process framework, it is possible to apply this concept to any collaborative process. As humans work more closely with artificial entities in the future, collaborative FEPs offer a mechanism to ensure that work is mutually productive to all entities involved, regardless of their physical nature.

REFERENCES

Cohen, P. R., Hanks, S., & Pollack, M. E. (1993, Winter). Benchmarks, testbeds, controlled experimentation, and the design of agent architectures. *AI Magazine*, *14*(4), 17-42.

Dunbar, K. (1995). How scientists really reason: Scientific reasoning in real-world laboratories. In R. J. Sternberg & J. Davidson (Eds.), *Mechanisms of insight* (pp. 365-395). Cambridge MA: MIT press.

Fay, N., Garrod, S., & Carletta, J. (2000). Group discussion as interactive dialogue or as serial monologue. *The Influence of Group Size Psychological Science*, *11*(6), 481-486.

Horling B., Lesser V., & Vincent R. (1999). Experiences in simulating multi-agent systems using TMS. In *Proceedings of the Fourth International Conference on Multi-agent Systems (ICMAS 2000)*, Boston, July, 2000. AAAI. (Also submitted as UMASS Tech. Rep. No. 1999-75).

Jennings, N. R., & Wooldridge, M. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review*, 10.

Kaminka, G. A., Veloso, M., Schaffer, S., Sollitto, C., Adobbati, R., Marshal, A. N., et al. (2002, January). GameBots: The ever-challenging multi-agent research test-bed. *Communications of the ACM*.

Kaplan, F. (2004). Who is afraid of the humanoid? Investigating cultural differences in the acceptance of robots. *International Journal of Humanoid Robotics, 1*(3), 465-480.

Kelman, H., Fiske, S. T., Kazdin, A. E., & Schacter, D. (Eds.). (2006). Interests, relationships, identities: Three central issues for individuals and groups in negotiating their social environment. *Annual Review of Psychology, 57*, 1-26.

Khan, Z. (1998). *Attitudes towards intelligent service robots, TRITA-NA-P9821, IPLab-154*.

Laird, J.E. (2001). Using a computer game to develop advanced AI. *IEEE Computer, 34*, 70-75.

Magerko, B., Laird, J. E., Assanie, M., Kerfoot, A., & Stokes, D. (2004). AI characters and directors for interactive computer games. In *Proceedings of the 2004 Innovative Applications of Artificial Intelligence Conference.*

Nakanishi, H., Nakazawa, S., Ishida, T., Takanashi, K., & Isbister, K. (2003). Can software agents influence human relations? In *Proceedings of Balance Theory in Agent-mediated Communities, AAMAS2003*, (pp. 717-724).

Riedl, M., Saretto, C.J., & Young, M.R. (2003). Managing interaction between users and agents in a multi-agent storytelling environment. In *Proceedings of the Second International Conference on Autonomous Agents and Multi-agent Systems*.

Thomas, D., & Vlacic, L. (2005). TeamMATE: Computer game environment for collaborative and social interaction. In Proceedings of the *IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO'05)*.

Thomas, D., & Vlacic, L. (2007). Collaborative decision making amongst human and artificial beings. In G. Phillips-Wren & L. C. Jain (Eds.), *Intelligent decision making: An AI based approach*. Springer-Verlag.

Thomas, D., Vlacic, L., Hyungsuck Cho, J. K., & Lee, J. (Eds.). (2003). *Selecting an environment for cooperative*

autonomous robot research in intelligent robots: Vision, learning and interaction (pp. 187-198). ***KAIST Press*.**

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, LIX*, 433-460.

KEY TERMS

Fully Equal Partner (FEP): An intelligent entity that performs tasks cooperatively with other FEPs (biological or artificial), but is also capable of being replaced one with another. These beings are not necessarily aware of the nature of their fellow partners.

Group Collective Knowledge: Information that is presented to a group of FEPs collaboration.

Collaborative Process: Involves the interaction of FEPs to achieve defined outcomes. The process involves questions, responses, actions and negotiation.

Influence: Among FEPs is the ability of a FEP, during collaborative negotiation, to align another FEPs response with their own.

Collaborative Gap: The difference between the outcomes of groups of human and artificial entities in a collaborative decision-making scenario, where both sets of entities have the same beliefs and objectives.

Intelligent Tools: Intelligent artificial entities that are utilised by humans but do not impact any collaborative decision-making process via social influences.

Social Acceptance: Social acceptance of artificial FEPs is the ability of human FEPs to accept Artificial FEPs into the collaborative process and influence, and be influenced, by these entities.

Transforming Recursion to Iteration in Programming

Athanasios Tsadiras
Technological Educational Institute of Thessaloniki, Greece

INTRODUCTION

The main advantage of a Recursive Algorithm (an algorithm defined in terms of itself) is that it can be easily described and easily implemented in a programming language (van Breughel, 1997). On the other hand, the efficiency of such an algorithm is relatively low because for every recursive call not yet terminated, a number of data should be maintained in a stack, causing time delays and requiring higher memory space (Rohl, 1984). Solving the same problem iteratively instead of recursively can improve time and space efficiency. For example, to solve a problem that involves N recursive procedure calls, it will require stack space linear to N. On the contrary, using iteration, the program will need a constant amount of space, independent of the number of iterations. There are **programming languages**, such as **Prolog**, that do not possess built-in iterative structures and so recursion should be used instead. Nevertheless, there are ways to write recursive programs that have similar behaviour with that of the corresponding iterative programs.

BACKGROUND

The transformation of **recursion** to **iteration** is not a new problem and it has been studied by various scientists, for example, De Moor and Sittampalam (2001) and Clinger (1998). This competition between recursion and iteration is interesting because they represent two different schools of thought in Computer Science. Various programming languages use **recursion** extensively, especially high-level programming languages that cope with Artificial Intelligence problems (Luger, 2002) such as Prolog (Bratko, 2000; Ramachandran, 1986), LISP (Lamkins, 2004; Seibel, 2005) or Scheme (Dybvig, 2004; Watson, 1996). The Prolog language will be used to exhibit the transformation of recursion to iteration (Clocksin & Mellish, 2003; Shoham, 1994). **Prolog** is a **Logic Programming Language** (Bramer, 2005) that uses heavily recursion. This happens because Prolog lacks iterative structures such as "for," "while-do" or "repeat-until" (Holmes, 2001; Langfield, 2003). In Prolog, a clause can be iterative even if it contains a recursive call. That is, a Prolog clause is iterative if it has zero or more calls to Prolog system predicates before the recursive call (Sterling & Shapiro, 1986). Furthermore, a Prolog procedure is iterative if it contains only unit clauses (facts) and iterative clauses. Having these in mind, we will try to transform **recursion** to **iteration** in Prolog, but that can apply to all other languages that use recursion.

TRANSFORMING RECURSION TO ITERATION

The fact is that there is no easy or general way to transform a recursive algorithm to an iterative algorithm. What a programmer can do to increase the efficiency of a recursive algorithm is to implement the recursive algorithm in an iterative manner. This can be done by using special variables called accumulators that will be used to keep intermediate results and facilitate acceleration.

To demonstrate the technique, the calculation of the sum of all integers from 1 to N will be used. The recursive relation that can be used to calculate the sum S of all positive integers from 1 to N is S(N)=S(N-1)+N. A Prolog predicate sum(N,S) that evaluates sum S of all integer numbers from 1 to N, is the following:

```
sum(1,1).
sum(N,S):-N1 is N-1,
          sum(N1,S1),
          S is S1+N.
```

This is a recursive implementation because the second clause involves the call of the same predicate and after it, there is a Prolog system call (S is S1+N). To illustrate the computational effort of the above implementation, the trace of the call to find the sum of all integers from 1 to 4 is shown below (Figure 1).

In Figure 1, we see that after a series of recursive calls the call sum(1,S3) stops the recursion and initiates a series of value returns, until the initial call sum(4,S) is reached. The common process of recursion is now apparent where first the problem is getting smaller and smaller, until it meets the limit case. After that, values of intermediate call are returned until the initial call is answered. If we call the same predicate for an integer that is above a certain limit, the computer will

Figure 1. The trace of the recursive implementation

```
?-sum(4,S).
   → sum(3,S1)
     → sum(2,S2)
       → sum(1,S3)
       ← sum(1,1)
     ← sum(2,3)
   ←sum(3,6)
 sum(4,10)
S=10
```

Figure 2. Trace of the iterative implementation, involving two additional parameters, one accumulator and one counter

```
?-sum_iterative1(4,S).
  → sum1(1,4,1,S)
    → sum1(2,4,3,S)
      → sum1(3,4,6,S)
        → sum1(4,4,10,S)
S=10
```

reach stack overflow because it cannot afford to provide memory space for all the recursive calls.

We will try to implement the same algorithm in an iterative manner using predicate sum_iterative1(N,S). This will require the introduction of two new parameters. The first one will play the role of the **counter** of the iteration and the second will play the role of the **accumulator**. That is, the predicate sum_iterative1(N,S) will call an augmented predicate sum1(I,N,T,S) that has the two additional parameters. Counter "I" represents the I-th iteration and accumulator "T" the temporal sum until iteration I. Both the counter and the accumulator should be initiated with value 1. For this reason, the call of sum_iterative1(N,S) lead to the call of predicate sum1(1,N,1,S) that has the first and the third argument equal to 1. The complete code is given below.

```
sum_iterative1(N,S):-
sum1(1,N,1,S).

sum1(I,N,T,S):-
      I<N,
      I1 is I+1,
      T1 is T+I1,
      sum1(I1,N,T1,S).
sum1(N,N,S,S).
```

Predicate sum1 is iterative because, as it is mentioned in the "Background" section, after the pseudo-"recursive" call at sum1(I1,N,T1,S), there is no other call. This kind of **iteration** is also called **Tail Recursion**. The introduction of the 2 parameters leads to the evaluation of the partial result of the sum at each step of the iteration and cause the termination of the iteration when counter I becomes equal to N. It is apparent that the recursion is transformed into an iteration of N steps. The iterative nature of the above implementation is illustrated in Figure 2, which shows the trace of the call to find the sum of all integers from 1 to 4.

Comparing this trace with that of Figure 1, we see that now the solution is found as soon as the call reaches the limit case at sum1(4,4,10,S). This means that the second stage of returning values and making calculations that is present in the trace of Figure 1 is now avoided, shortening the whole process.

The same problem can be solved iteratively even without having the counter I described above. In this case only one additional parameter, that of the accumulator, will be needed. The accumulator will store the partial result up to the current step of iteration. The accumulator should be initiated with value 0; this is why the second argument of the call of predicate sum2(N,0,S)is zero. The code is given below:

```
sum_iterative2(N,S):-
sum2(N,0,S).

sum2(N,T,S):-
      N>0,
      T1 is T+N,
      N1 is N-1,
      sum2(N1,T1,S).
sum2(0,S,S).
```

In this second implementation variable N also plays the role of the counter, because at every iteration step, its value decreases by one, terminating when its value becomes zero. Once again, the recursion is transformed into an iteration of N steps. This can be shown with the following trace of the call sum_iterative2(4,S).

The trace of Figure 3 is similar of that of Figure 2, that is, the second stage of returning values found in Figure 1, is omitted, and additionally, needs less memory space than the implementation traced at Figure 2, because only one additional parameter is added instead of two.

Figure 3. Trace of the second iterative implementation, involving only one additional parameter, that of an accumulator

```
?-sum_iterative2(4,S).
  → sum2(4,0,S)
   →sum2(3,4,S)
    → sum2(2,7,S)
      → sum2(1,9,S)
        → sum2(0,10,S)
S=10
```

Figure 4. Trace of the recursive implementation

```
?-sumlist([3,5,7,8],S).
   → sumlist([5,7,8],S1)
    → sumlist([7,8],S2)
     → sumlist([8],S3)
        → sumlist([],S4)
        ← sumlist([],0)
     ← sumlist([8],8)
    ← sumlist([7,8],15)
  ←sumlist([5,7,8],20)
 sumlist([3,5,7,8],23)
S=23
```

We conclude that both above implementations demonstrate the technique of the transformation of the recursion to iteration. This transformation leads to more efficient use of computer resources, something especially useful in cases of time or space critical applications.

The same technique can be applied also to programs that involve lists. The data structure of list is recursive in nature, because it can be defined as a structure that has a single element at its head and all other elements at its tail (Clocksin, 2003). Because the tail is also a list, we conclude that a list is an element and a list. This recursive nature is the reason why most of the problems concerning lists are solved in a recursive way. For example, the following predicate calculates recursively in Prolog, the sum of a list of integers.

```
sumlist([],0).
sumlist([H|Tail],Sum):-
        sumlist(Tail,Sum1),
        Sum is Sum1+H.
```

Because the second clause has a recursive call to the same predicate and this call is followed by a Prolog system call (Sum is Sum1+H), this is a recursive implementation. To illustrate this recursion, the trace of a Prolog call that asks for the sum of the numbers in the list [3,5,7,8], is shown below.

As it is shown in Figure 4, after a series of recursive calls that gradually shorten the length of the list of numbers, the call sumlist([],S4) is reached. According to the first clause of the predicate sumlist([],0), S4 should equal zero. This result is returned to the parent recursive call. This process of returning the result to the parent recursive call is continued until the final result S=23 is returned.

To increase the efficiency of the program above, an iterative version of the same predicate must be implemented. Once again, an additional argument should be introduced, playing the role of the accumulator. The accumulator will keep the partial sum of numbers that have been added until that point. It should be initiated with value 0, which is why the second argument of the call of predicate sumlist1 is zero.

```
sumlist_interative(L,S):-
        sumlist1(L,0,S).
sumlist1([],S,S).
sumlist1([H|Tail],T,Sum):-
        T1 is T+H,
        sumlist1(Tail,T1,Sum).
```

We see that predicate sumlist1 is iterative because, as it is mentioned in the "Background" section, after the pseudo-"recursive" call at sumlist1(Tail,T1,Sum), there is no other call. The second argument T plays that role of an accumulator. This accumulator has the complete solution as soon as the whole list is processed.

To illustrate this version of sumlist, the trace of the same Prolog call with that of Figure 4 is shown below.

It is clear that the sum of the numbers in the list [3,5,7,8] is gradually calculated at the accumulator, that is the second argument of sumlist1. When the list reaches its limit state, that is, when it becomes empty, the value of the accumulator is the solution. In Figure 5, the iterative nature of the implementation is apparent, especially if we compare it with that of Figure 4. In Figure 4, where recursion is used, there is a

Figure 5. Trace of the iterative implementation, involving the use of an accumulator

```
?-sumlist_iterative([3,5,7,8],S).
  → sumlist1([3,5,7,8],0,S)
   → sumlist1([5,7,8],3,S)
    → sumlist1([7,8],8,S)
      → sumlist1([8],15,S)
        → sumlist1([],23,S)
S=23
```

second phase in which the return of values is happening. In Figure 5, this second phase is completely omitted.

FUTURE TRENDS

The "competition" between the use of recursion or iteration is not new. Both techniques have advantages and disadvantages. The programmer should be aware of these advantages and disadvantages and use the right technique to the right problem.

As soon as Artificial Intelligence Applications and other applications start to incorporate time critical subprograms, the recursion should to be avoided because it can cause unnecessary time delays. In such cases, the transformation of recursion to iteration is a valid solution. The same applies when memory space is limited and recursion is a luxury that we cannot afford.

Furthermore, the technique that transforms recursion to iteration can be useful from an educational point of view. Students that learn how to program both recursively and iteratively should know about the technique presented above, in order to understand how recursion and iteration are achieved internally in a computer and choose the more suitable implementation. Because time and space efficiency is a requirement in modern programming, the transformation of recursion to iteration as presented above will become more and more popular.

CONCLUSION

Transforming recursion to iteration/tail recursion can provide many advantages, especially from a time or a memory space point of view. The current article has presented a technique to achieve such a transformation. This technique incorporates the introduction of one or two internal parameters that play the role of the accumulator and the counter. These two additional parameters are used in order:

a. the accumulator to keep partial results that are evaluated during the execution of the program and
b. the counter to keep the number of iterations that are made until that point of execution.

REFERENCES

Bramer, M. (2005). *Logic programming with Prolog*. Springer-Verlag.

Bratko, I. (2000). *Prolog–programming for artificial intelligence* (3rd ed.). Addison-Wesley.

Breughel van, F. (1997). *Comparative metric semantics of programming languages, nondeterminism and recursion*. Springer-Verlag.

Clinger, W.D. (1998). Proper tail recursion and space efficiency. In *Proceedings of the ACM SIGPLAN 1998 Conference*.

Clocksin, W.F., & Mellish, C.S. (2003). *Programming in Prolog: Using the ISO standard* (5th ed.). Springer-Verlag.

De Moor, O., & Sittampalam, G. (2001). *Higher-order matching for program transformation.* Theoretical Computer Science.

Dybvig, K. (2003). *The scheme programming language* (3rd ed.). MIT Press.

Holmes, B.J. (2001). *Pascal programming*. Thomson Learning.

Lamkins, D. (2004). *Successful lisp: How to understand and use common lisp*. bookfix.

Langfield, S. (2003). *Learning to program in Pascal and Delphi*. Payne-Gallway.

Luger, G. (2002). *Artificial intelligence. Structures & strategies for complex problem solving* (4th ed.). Addison-Wesley.

Pettorossi, A., Proietti, M., (1996). Rules and strategies for transforming functional and logic programs. *ACM Computing Surveys*, *26*(2), 360-414.

Ramachandran, B. (1986). *Introduction to PROLOG*. TAB Books.

Rohl, J.S. (1984). *Recursion via Pascal*. Cambridge University Press

Seibel, P. (2005). *Practical common lisp*. Apress.

Shoham, Y. (1994). *Artificial intelligence techniques in PROLOG*. Morgan Kaufmann.

Sterling, L., & Shapiro, E. (1986). *The art of Prolog*. MIT Press.

Watson, M. (1996). *Programming in Scheme: Learn Scheme through artificial intelligence programs*. Springer-Verlag.

KEY TERMS

Accumulator: A variable used for the accumulation of arithmetic sums, usually evaluated in the form Sum:=Sum+X. Usually, it is initialized with the value zero.

Counter: A variable that stores a single natural number and usually represents the number of iterations that have to be done so far or the number of iterations that are needed from that point on. In the first case, its value is usually evaluated by the form N:=N+1 and its initial value is zero. In the second case, its value is usually evaluated by the form N:=N-1 and its final value is zero.

Iteration: The repeated execution of the same process a given number of times or until a specified result is obtained.

Predicate: A predicate is an operator or function which returns a Boolean value (e.g., true or false).

Recursion: The definition of a program in terms of itself. This gives the program the ability to call itself.

Stack Overflow: The condition that occurs when a computer program makes too many calls to subprograms, leading the memory stack to run out of space.

Trace: A report showing the detailed step-by-step execution of a computer program

Transmission of Scalable Video in Computer Networks

Jânio M. Monteiro
University of Algarve and IST/INESC-ID, Portugal

Carlos T. Calafate
Technical University of Valencia, Spain

Mário S. Nunes
IST/INESC-ID, Portugal

INTRODUCTION

The Internet as a video distribution medium has seen a tremendous growth in recent years with the advent of new broadband access networks and the widespread adoption of media terminals supporting video reception and storage. This growth of Internet video transmission resulted from the advances in video encoding solutions and the increase in the bandwidth of terminals. However, it has also placed new challenges in the developments of video standards, due to the heterogeneous characteristics of current terminals and of the wired and wireless distribution networks.

As the terminal capabilities increase in terms of display definition, processing power, and bandwidth, users tend to expect higher qualities from the received video streams. Additionally, as different types of terminals will likely coexist in the same network, it will be necessary to adapt the content transmitted according to the receiving terminal. Instead of re-encoding (or transcoding) the bitstream, which requires a high computational power on intermediate nodes, rate adaptation would preferably be done by extracting parts of the original bitstream.

In terms of network scalability, encoding and transmitting the same video sequence in a large-scale live video distribution system is a challenge, which may only rely on point-to-multipoint transmissions like IP multicast or broadcast. The traditional solution for point-to-multipoint video transmission in heterogeneous networks and with terminals with very different capabilities relies on a process usually called simulcast or replicated streams transmission. In this process, a discrete number of independent video streams are encoded and distributed through the multicast or broadcast path. Terminals request and decode the video stream that better fits their characteristics and communication rate, switching between video streams according to bandwidth variations. However, the major drawback of *simulcast* is that much of the information carried in one stream is also carried in adjacent streams, and therefore the total rate required for a video transmission is much higher than the rate of a single stream. In these scenarios it would be preferable to encode different levels of quality—one base layer quality and one or more enhancement layers—which could be used to increase the quality of the base layer. Accordingly, terminals with lower bandwidth or computational power could request the reception of lower layers, and terminals with higher capabilities could request additional enhancement layers.

In this article we analyze scalable video transmission, from the perspective of video coding standards and the necessary developments in protocols that support media distribution in current and future network architectures. In the next section we start by describing the first contributions to this topic and following developments in related video coding standards. We then describe the structure of a scalable video bitstream, taking the novel H.264/SVC standard as reference, and we further proceed with an analysis of the protocols that can be used for the description, signaling, and transport of scalable video. We describe different network scenarios and examples where scalable video offers significant advantages, before moving on to some remarks on future trends in this area, discussing those mechanisms that must be associated with SVC techniques to achieve an efficient and robust transmission system, and concluding the article.

BACKGROUND

The use of layered video transmission in IP multicast was originally proposed by Deering (1993), who suggested the transport of different video layers in different multicast groups. With this solution the encoder would produce a set of interdependent layers (one base layer and one or more enhancement layers), and the receiver, starting with a base layer, could adapt his quality by joining and leaving multicast trees, each one carrying a different quality layer.

Deering's proposal was followed by several protocols like: receiver driven layered multicast (RLM) protocol (Mc-

Canne, Jacobson, & Vetterli, 1996), layered video multicast with retransmission (LVMR) protocol (Li, Paul, Pancha, & Ammar, 1997; Li, Paul, & Ammar, 1998), and ThinStreams (Wu, Sharma, & Smith, 1997).

The advantages of layered video multicast were confirmed by Kim and Ammar (2001) for scenarios where receivers are distributed in the same domain, with multiple streams sharing the same bottleneck link, as occurs in many video distribution scenarios.

In terms of video coding, layered video transmission requires a layered encoding of video, a process usually referred to as scalable video coding (SVC). Video coding standards like *ITU-T Recommendation H.263* from the International Telecommunication Union-Telecommunication (ITU-T, 2000) and *MPEG-2 Video* from the ITU-T and the International Organization for Standardization/International Electrotechnical Commission (ITU-T & ISO/IEC, 1994) included several tools that supported the most important scalability options. However, none of these scalable extensions was broadly implemented since they imply a loss in coding efficiency and also a significant increase in terms of decoder complexity.

In January 2005, the Joint Video Team (JVT) from ISO/IEC Moving Picture Experts Group (MPEG) and the ITU-T Video Coding Experts Group (VCEG) started developing a scalable video coding extension for the H.264 Advanced Video Coding standard (ITU-T & ISO/IEC, 2003), known as H.264 Scalable Video Coding (ITU-T & ISO/IEC, 2007). The H.264 SVC augments the original encoder's functionality to generate several layers of quality. Enhancement layers may enhance the content represented by lower layers in terms of temporal resolution (i.e., the frame rate), spatial resolution (i.e., image size), and the quality—specified as signal-to-noise ratio resolution (i.e., SNR).

By using the H.264 SVC, different levels of quality could be transmitted efficiently over both wired and wireless networks, allowing seamless adaptation to available bandwidth and to the characteristics of the terminal. However, the transport of SVC presents many challenges which must be considered in order to take full advantage of its potential.

CODING AND TRANSMISSION OF SCALABLE VIDEO

The most adequate technique for efficiently transmitting scalable video is highly dependent on the video encoding technology itself. Hence, for this article we have used the scalable video extensions to the H.264 standard as reference since these represent the most advanced technology currently available in this area.

The H.264 Advanced Video Coding (AVC) standard is currently emerging as the preferred solution for video services in third-generation (3G) mobile networks, which include packet-switched streaming services, messaging services, conversational services, and multimedia broadcast/multicast services (MBMS) (3GPP TS 26.346, 2005). It will also be used for mobile TV distribution to handheld devices (DVB-H) (ETSI TR 102 377, 2005).

SVC Bitstream Structure

The scalable extension of H.264/AVC includes several layers of quality. Relative to the base layer of an SVC bitstream, and for compatibility purposes, the JVT decided to make it compatible with the H.264/AVC profile.

The SVC bitstream may be composed of multiple spatial, temporal, and SNR layers of combined scalability. Temporal scalability is a technique that allows supporting multiple frame rates. In SVC, temporal scalability is usually implemented by using hierarchical B-pictures.

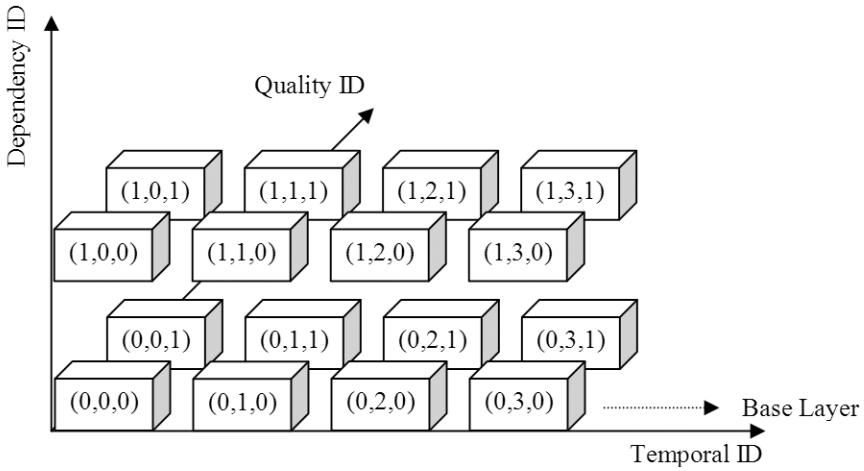
Quality (or SNR) scalability relies on both *coarse-grain quality scalable* (CGS) and *medium-grain quality scalable* (MGS) coding. While CGS encodes the transform coefficients in a nonscalable way, in MGS, which is a variation of CGS, fragments of transform coefficients are split into several network adaptation layer (NAL) units, enabling a more graceful degradation of quality when these units are discarded for rate adaptation purposes. The JVT also considered the possibility of including another form of scalability named *fine-grain scalability* (FGS), which was proposed in MPEG-4 Visual. FGS arranges the transform coefficients as an embedded bitstream, enabling truncation of these NAL units at any arbitrary point. However, in the first specification of SVC (Phase I) (ITU-T & ISO/IEC, 2007), FGS layers were not supported.

Spatial scalability provides support for several display resolutions (e.g., 4CIF, CIF, or QCIF) and is implemented by decomposing the original video into a spatial pyramid.

These spatial, temporal, and SNR layers are identified using a triple identification (ID), consisting of the *dependency ID* identifying the spatial definition, the *temporal ID*, and the *quality* (i.e., SNR) *ID*, which is referred as tuple (D,T,Q). For instance, a base layer NAL unit of the lowest temporal resolution and SNR scalability should be identified as (D,T,Q)=(0,0,0). Accordingly the network adaptation layer structure of H.264/AVC has been extended to include these three IDs. These three layers may be represented using a three-dimensional graph, such as the one in Figure 1.

SVC layers can be highly interdependent from each other, which means that the loss of an NAL unit of a certain layer may cause a severe reduction of quality or even prevent the correct decoding of other layers. This implies that lower layers should be protected from bit errors or packet losses.

Figure 1. Example of a coded video sequence with four temporal layers, two image definitions, and two quality layers per image definition



Protocols for Scalable Video Description, Signaling, and Transport

Scalable transmission of H.264 SVC is based on the Internet Engineering Task Force (IETF)-defined architecture and suit of protocols. SVC NAL units are encapsulated in real-time transport protocol (RTP) (Schulzrinne, Casner, Frederick, & Jacobson, 2003) packets, which are usually carried over user datagram protocol (UDP) (Postel, 1980) and Internet protocol (IP) (Postel, 1981) datagrams.

Concerning the RTP payload format definition, the IETF is currently working on such a draft (Wenger, Wang, & Schierl, 2007). Special attention is being taken in order to make the SVC payload format an extension of the original H.264 payload format defined in RFC 3984 (Wenger, Hannuksela, Stockhammer, Westerlund, & Singer, 2005), so that SVC servers can distribute data to legacy receiver equipment. The main mechanisms for fragmentation and aggregation defined in RFC 3984 are being maintained in Wenger et al. (2007).

In terms of session signaling, two main protocols are used to initiate and control streaming sessions, namely the session initiation protocol (SIP) (Rosenberg et al., 2002) and the real-time streaming protocol (RTSP) (Schulzrinne, Rao, & Lanphier, 1998). Both of these protocols use the session description protocol (SDP) (Handley, Jacobson, & Perkins, 2006) to describe terminals' capabilities. Thus SVC information must be conveyed using SDP, and this protocol must be updated to include the description of interlayer dependencies.

The communication topology defined for SVC transmission considers the possibility of including one or more intermediate nodes, named Media Aware Network Element (MANE) (Wenger et al., 2005), which could adapt unicast and multicast transmission modes, aggregate several streams, or perform rate-adaptation tasks adapting a bitstream according to the capabilities of a particular receiver. These elements also constitute session signaling endpoints between servers and clients.

Scenarios Where Scalable Video Offers Significant Advantages

Layered video transmission was initially intended for IP multicast networks. However, IP multicast is not being largely used on the Internet. Instead, it is currently considered as an important solution for edge IP networks, like those of service providers. Nevertheless, even in these networks, IP multicast constitutes a good solution for scalable transmission of video with the advantage that it is easier to associate it with quality of service (QoS) solutions that guarantee a higher priority to video or other sensitive data.

Two basic video distribution scenarios for SVC can be considered. In a first scenario IP multicast is used between a server and clients. As previously explained, this may be the case for a large private IP network of a service provider, as

for example MBMS (3GPP TS 26.346, 2005) in mobile 3G networks. In this scenario receivers request the SVC layers according to their capacity, dynamically joining or leaving multicast groups. The signaling for session establishment is also performed on an end-to-end basis. In this architecture, it is important to specify a protocol or mechanism that guarantees equality between receivers, since a greedy or naive user could request a rate far beyond the network capacity, thus leading the network to a persistent congested state.

One of the problems pointed out to this solution is the number of firewall ports opened in receivers, which may constitute a security problem.

In a second scenario, one or more MANEs could be placed between a server and clients. MANEs constitute a signaling and RTP endpoint. They can perform rate adaptation tasks, removing packets from an incoming RTP stream and rewriting RTP headers, or even aggregate several video layers in one unique RTP flow of video in case an endpoint does not have the processing power or display size to decode all layers. Using the example of a mobile 3G network, the MANE should be placed close to a base station, with access to both signaling and media traffic. This element can also be associated with other modules or functions which, for instance, perform adaptation between unicast and multicast transmission modes, QoS broking, adaptation between QoS-capable and best-effort networks, or even admission control tasks. This scenario could constitute part of an overlay network that manages the distribution of media at the application layer.

FUTURE TRENDS

The integration of scalable video transmission techniques in current and emerging network architectures still presents some challenges that must be considered prior to being able of taking the most of it.

QoS solutions should be considered for giving higher priority to video data when compared with less sensitive data. QoS solutions for wireless networks, like for instance the IEEE Standard 802.11e (2005), usually include several mechanisms like queue management and automatic repeat request (ARQ) solutions. Link layer ARQ solutions can be associated with QoS provisioning, quickly recovering from packet losses in more important layers.

Concerning large-scale multicast and broadcast networks, the retransmission of lost packets could constitute a scalability problem, and therefore the protection of lower layers may be performed by applying forward error correcting (FEC) codes. In such scenarios the use of raptor codes (Shokrollahi, 2006) is gaining popularity due to their high flexibility and effectiveness. As an example, the use of raptor codes was proposed for H.264/AVC video transmission in mobile MBMS networks (3GPP TS 26.346, 2005).

Finally, in terms of coding, current H.264/SVC encoding and decoding processes present additional complexity when compared with the AVC standard. In particular, decoding time must be minimized in order to encourage the widespread adoption of this novel technology.

CONCLUSION

In this article we describe the main issues associated the transmission of scalable video coding. Although the basic mechanisms of scalable video transmission were defined more than a decade ago, there is still a lack of encoding and transmission mechanisms to support it. In terms of video coding, the flexibility offered by the scalable extension of the H.264/AVC provides considerable opportunities to extend current video distribution networks without the penalty of having to change or upgrade H.264/AVC terminals: its base layer and signaling mechanisms are designed to guarantee the compatibility with legacy equipments.

In terms of network support, several challenges remain that must be considered. However, the advent of mobile and wireless networks with very different terminal capabilities, combined with wired access networks capable of delivering high-quality videos, will require a flexible technology that supports easier rate-adaptation mechanisms with graceful degradation when facing bandwidth reduction.

REFERENCES

3GPP TS 26.346. (2005). *Multimedia broadcast/multicast service (MBMS); Protocols and codecs—version 6.1.0.* Third Generation Partnership Project.

Deering, S.E. (1993, October). Internet multicast routing: State of the art and open research issues. *Proceedings of the Multimedia Integrated Conferencing for Europe (MICE) Seminar at the Swedish Institute of Computer Science.*

ETSI TR 102 377. (2005). *Digital video broadcasting (DVB)□DVB-H implementation guidelines.* European Telecommunications Standardization Institute.

Handley, M., Jacobson, V., & Perkins C. (2006). *SDP: Session description protocol (RFC 4566).* Internet Engineering Task Force.

IEEE Standard 802.11e. (2005). *Institute of Electrical and Electronics Engineers standard for information technology – telecommunications and information exchange between systems – local and metropolitan area networks – specific requirements part 11: Wireless LAN medium access control (MAC) and physical layer (PHY) specifications: Amend-*

ment 8: Medium access control (MAC) quality of service enhancements. IEEE.

ISO/IEC International Standard 14496. (2001). *Information technology—coding of audio-visual objects.* ISO/IEC.

ITU-T. (2000). *ITU-T recommendation H.263: Video coding for low bit rate communication* (version 1: November 1995; version 2: January 1998; version 3: November 2000). Author.

ITU-T & ISO/IEC. (1994). *ITU-T recommendation H.262 and ISO/IEC 13818-2: Generic coding of moving pictures and associated audio information—part 2: Video (MPEG-2 video).* Author.

ITU-T & ISO/IEC. (2003, May). *ITU-T recommendation H.264 and ISO/IEC 14496-10: Advanced video coding for generic audiovisual services (MPEG-4 AVC).* Author.

ITU-T & ISO/IEC. (2007, April). *ITU-T recommendation H.264 and ISO/IEC 14496-10: Advanced video coding for generic audiovisual services (MPEG-4 AVC).* Author.

Kim, T., & Ammar, M.H. (2001, June). A comparison of layering and stream replication video multicast schemes. *Proceedings of the IEEE 11th International Workshop on Network and Operating System Support for Digital Audio and Video* (NOSDAV'01) (pp. 63-72), New York.

Li, X., Paul, S., Pancha, P., & Ammar M. (1997, May). Layered video multicast with retransmission (LVMR): Evaluation of error control schemes. *Proceedings of the IEEE 7th International Workshop on Network and Operating System Support for Digital Audio and Video* (NOSDAV'97) (pp. 161-172), St. Louis, MO.

Li, X., Paul, S., & Ammar, M. (1998, March). Layered video multicast with retransmission (LVMR): Evaluation of hierarchical rate control. *Proceedings of the 17th Annual Joint Conference of the IEEE Computer and Communications Societies* (INFOCOM'98) (vol. 3, pp. 1062-1072), San Francisco.

McCanne, S., Jacobson, V., & Vetterli M. (1996, August). Receiver-driven layered multicast. *Proceedings of the Special Interest Group on Data Communications of the Association for Computing Machinery* (ACM SIGCOMM) (pp. 117-130).

Postel, J. (1980). *User datagram protocol (STD 0006, RFC 768).* Internet Engineering Task Force.

Postel, J. (1981). *Internet protocol DARPA Internet program protocol specification (STD 0005, RFC 791).* Internet Engineering Task Force.

Ramos, N., & Dey, S. (2007, September). A device and network-aware scaling framework for efficient delivery of scalable video over wireless networks. *Proceedings of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications* (pp. 1-5).

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., & Schooler, E. (2002). *SIP: Session initiation protocol (RFC 3261).* Internet Engineering Task Force.

Schulzrinne, H., Casner, S., Frederick, R., & Jacobson, V. (2003). *RTP: A transport protocol for real-time applications (STD 0064, RFC 3550).* Internet Engineering Task Force.

Schulzrinne, H., Rao, A., & Lanphier, R. (1998). *Real time streaming protocol (RTSP) (RFC 2326).* Internet Engineering Task Force.

Shokrollahi, A. (2006). Raptor codes. *Institute of Electrical and Electronics Engineers Transactions on Information Theory, 52*(6), 2551-2567.

Schierl, T., Stockhammer, T., & Wiegand, T. (2007, September). Mobile video transmission using scalable video coding. *Institute of Electrical and Electronics Engineers Transactions on Circuits and Systems for Video Technology, 17*(9), 1204-1217.

Wenger, S., Wang, Y.-K., & Schierl, T. (2007, November 19). *RTP payload format for SVC video (Internet draft: draft-ietf-avt-rtp-svc-03).* Internet Engineering Task Force.

Wenger, S., Hannuksela, M.M., Stockhammer T., Westerlund, M., & Singer, D. (2005). *RTP payload format for H.264 video (RFC 3984).* Internet Engineering Task Force.

Wu, L., Sharma, R., & Smith, B. (1997, May). Thin streams: An architecture for multicast layered video. *Proceedings of 7th International Workshop on Network and Operating Systems Support for Digital Audio and Video* (NOSDAV'97) (pp. 173-182), St. Louis, MO.

KEY TERMS

Automatic Repeat Request (ARQ): A method for error control in packets or frames that uses redundant bits, acknowledgment packets, and timeouts to detect and retransmit data affected by errors.

Broadcast: Process of transmitting a flow of data from a source to all receivers of the broadcast domain or of the network where the terminal operates.

Forward Error Correction (FEC): A method for error correction that uses redundant bits to detect and correct errors in data without the need for retransmission.

IP Multicast: The transmission of Internet protocol (IP) datagrams from a source to all the receivers that belong to a certain IP multicast group. In IP multicast, terminals that wish to receive data must register with that multicast group. IP multicast routing protocols are used to create distribution trees.

Network Adaptation Layer (NAL): An H.264 syntax structure that represents a frame or part of it, integrating information about its encoding parameters. NAL structures are appropriate for the transport of H.264 over several types of networks.

Quality of Service (QoS): Process of providing different priorities to different flows of data or to guarantee a certain level of quality to a data flow. Examples of quality parameters are transmission rate, delay, delay variation (jitter), and packet loss.

Simulcast: Also referred to as replicated streaming, the process of simultaneously encoding and transmitting a same video sequence through a discrete number of quality video streams to several receivers.

Transcoding: The process of partially or totally decoding and re-encoding a certain video stream in order to change its characteristics in terms of definition, frame rate, or transmission rate.

Unicast: Process of transmitting a flow of data from a source to a particular receiver.

The Trends and Problems of Virtual Schools

Glenn Russell
Monash University, Australia

INTRODUCTION: THE EMERGENCE OF THE VIRTUAL SCHOOL

Until recent times, schools have been characterized by the physical presence of teachers and students together. Usually, a building is used for instruction, and teaching materials such as books or blackboards are often in evidence. In the 20th century, alternatives to what may be called "bricks-and-mortar" schools emerged. These were forms of distance education, where children could learn without attending classes on a regular basis. The technologies used included mail, for correspondence schools, and the 20th century technologies of radio and television.

Virtual schools can be seen as a variant of distance education. Russell (2004) argued that they emerged in the closing years of the 20th century and can be understood as a form of schooling that uses online computers to provide some or all of a student's education. Typically, spatial and temporal distancing is employed, and this results in students being able to use their computers at convenient times in their homes or elsewhere, rather than being subject to meeting at an agreed upon time in a school building.

The concept of a virtual school is agreed upon only in broad terms, as there are a number of variants. Some virtual schools insist on an agreed upon minimum of face-to-face contact, while others are so organized that a student might never set foot in a classroom. It is possible for a virtual school to have no physical presence for students to visit, and an office building in one state or country can be used to deliver virtual school services to interstate or international students.

One way of categorizing virtual schools is by imagining where they might be placed on a scale of face-to-face contact between students and teachers. At the conservative end of this scale, there would be conventional schools, where students use online computers in classrooms or labs for some of their lessons. A trained teacher in the same subject area might be available to help students, or other teachers, volunteers, or parents could supervise them.

Toward the middle of such a scale would be mixed-mode examples, where some subjects are offered in virtual mode, but students are asked to visit the school on a regular basis to monitor their progress or to participate in other face-to-face subjects, such as sport, drama, or art. At the other end of the scale are virtual schools, where the student and teacher never meet, and there is no requirement for the student to enter a school building for the duration of the course. One example of such a virtual school is Florida High School, where there is no Florida High School building, and students and teachers can be anywhere in the world (Florida High School Evaluation, 2000, 2002).

FACTORS PROMOTING THE RISE OF VIRTUAL SCHOOLS

Russell (2005) has argued that the principal factors that account for the growth of virtual schools include globalization, technological change, availability of information technology (IT), economic rationalism, the model provided by higher education, perceptions about traditional schools, and the vested interests of those involved in them.

The first of these factors, globalization, refers to a process in which traditional geographic boundaries are bypassed by international businesses that use IT for globally oriented companies. It is now possible for curriculum to be delivered remotely from across state and national borders. Educational administrators can purchase online units of work for their school, and parents in developed countries can sometimes choose between a traditional school and its virtual counterpart.

As IT continues to develop, there is a correspondingly increased capacity to deliver relevant curricula online. As broadband connections become more common, students will be less likely to encounter prolonged delays while Web pages load or other information is downloaded. Advances in computers and software design have led to developments such as full-motion video clips, animations, desktop videoconferencing, and online music. Collectively, what is referred to as the Internet is already very different from the simple slow-loading Web pages of the early 1990s.

Economic rationalism also drives the spread of virtual schools, because the application of economic rationalism is associated with productivity. For education, as Rutherford (1993) suggested, the collective or government provision of goods and services is a disincentive to private provision, and that deregulation and commercialization should be encouraged. Consistent with this understanding is the idea that schools, as we know them, are inefficient and should be radically changed. Perelman (1992) argued that schools are remnants of an earlier industrial era that ought to be replaced with technology.

The ways in which higher education has adopted online teaching provide an example of how online education can be accepted as an alternative. The online courses provided by universities in recent years have proliferated (Russell & Russell, 2001). As increasing numbers of parents complete an online tertiary course, there is a corresponding growth in the conceptual understanding that virtual schooling may also be a viable alternative.

Those convinced that existing schools are unsatisfactory can see virtual schools as one alternative. Criticism of schools for not adequately meeting student needs, for providing inadequate skills required for employment, or not preparing students for examinations and entrance tests, are continuing themes that can be identified in a number of educational systems. Discussions related to school reform can include funding, resourcing, teacher supply, curriculum change, and pedagogy, but they can also include more radical alternatives, such as virtual schooling.

PROBLEMS OF VIRTUAL SCHOOLS AND THEIR SOLUTIONS

Virtual schools face a number of challenges related to the way that teaching and learning are implemented in online environments. While similar problems can also be identified in conventional schools, the different natures of virtual schools serve to highlight these concerns. These problems include authenticity, interactivity, socialization, experiential learning, responsibility and accountability, teacher training, certification, class sizes, accreditation, student suitability, and equity.

The first of these problems, authenticity, relates to the verification of the student as the person who has completed the corresponding assignments and tests from a virtual school. Virtual schools may assign students a secure password to use over the Internet, but this procedure would not preclude students from giving their passwords to a parent or tutor who completed the work on their behalf. A possible solution that may have to be considered is to test students independently to confirm that they have the understanding, knowledge, and skills suggested by their submitted work. Some virtual schools, such as Louisiana Virtual School (2006), ask students and parents to sign an honesty policy in an effort to maintain the authenticity of students' work.

Interactivity describes the relationship between the learner and the educational environment. For virtual school students, there is an interactive relationship involving the multimedia, the online materials used, and the teacher. Students would typically access materials on the World Wide Web, respond to them, and send completed work electronically to their teachers. The preferred way for students to become involved in online learning is to have an active engagement involving a response. If a student is directed to a static Web page containing a teacher's lecture notes, learning may be less effective, unless other teaching methods are used to supplement it. The solution to this problem will be found in both the increased capability of students' online computers to operate in a rich multimedia environment, and the recognition by course designers that virtual schools should take advantage of advances in learning theory and technological capability. In the U.S., the National Education Association's *Guide to Online High School Courses* (NEA, 2006) has maintained that online courses should reflect current research on learning theory and recognize the opportunities provided by online learning environments.

Socialization continues to be a problem with virtual schools, because there is an expectation in conventional schooling that students will learn how to work cooperatively with others and will internalize the norms and values necessary for living in a civilized community. Moll (1998) is concerned with disruption to the tradition of public education as the primary vehicle for the transference of national narratives and humanistic and democratic values. Clearly, socialization will still occur if students use online learning supplemented by some contact with teachers and opportunities for organized sports. However, students' ability to relate to others in society is likely to change. Despite this concern, a type of virtual school that routinely insists on organized face-to-face learning and social situations, with peers, teachers, and other adults, will reduce the problems that otherwise are likely to arise.

A related concern to that of socialization is the belief that Web culture is inherently isolating, and that by encouraging students to pursue their education with a virtual school, an existing trend toward loss of community may be exacerbated. Kraut et al. (1998) originally suggested that Internet use could be associated with declines in participants' communication with family members in the household, declines in the size of their social circles, and increases in depression and loneliness. However, more recent research (Kraut et al., 2002) found that negative effects had largely dissipated.

There are some teaching activities in conventional schools referred to as experiential. These usually involve some form of hands-on activity or physical interaction with others. Typically, a teacher will provide a demonstration, explanation, or modeling of what is to be learned, and activities that follow provide opportunity to correct errors. While virtual schools commonly offer subjects such as mathematics and social studies, the study of physical education, drama, art, and the laboratory component of science is more problematic. Sometimes the problem does not arise, because students will enroll only for subjects that they missed or that they need for credit toward a qualification.

A common solution to these problems is for the virtual school to provide online or print-based teaching materials, as with other subjects in the range to be offered. Students complete the activities and send evidence of the completed

work to the school. The Open School (2002) in British Columbia, Canada, offers art in both elementary and secondary school levels. At the Fraser Valley Distance Education Centre (2006), students are invited to participate in a science fair by sending in digital pictures and a digital video clip of their project to the supervising teacher.

Changing notions of responsibility, accountability, and student discipline are also likely to arise in virtual school environments (Russell, 2002). In a traditional school, teachers accept responsibility for the students in their charge, including the prevention of physical injury, and accountability for using appropriate teaching techniques. When there is a spatial and temporal distance between teacher and student, teachers are unable to exercise some of their accustomed responsibilities. While there is still a requirement to act ethically, and to ensure that appropriate teaching materials and methods are used, much of the responsibility shifts to parents, students, and to the suppliers of the online materials.

Teacher training is also emerging as an area of concern. Virtual teachers will find that some new skills are required, while others are less important. Class management skills in a face-to-face environment will differ from their online equivalents, as will many of the teaching practices. Salmon (2000) identified a number of skills that will be required by online teachers in the future. It is clear that there will be an ongoing need to use technological skills and to apply these skills to an appropriate educational context. However, it is unlikely that many teachers' colleges and other providers of trained teachers have modified their courses to reflect these changes, as mainstream teacher education is still focused on conventional school education. There are, nevertheless, some hopeful signs. The California Virtual School Report (2000) provided evidence of the use of online modules for teachers at Durham Virtual High School, in Canada, and a 15-week teacher-training program in Fairfax County School District.

Parents would normally expect that the virtual teacher working with their child would be a competent online teacher and be certified or registered with the corresponding school system. Where a student is working from home, and the principal contact with the teacher is by e-mail, the anonymity of the communication mode could conceivably cover the use of unqualified teachers. The necessity for demonstrating that a high-quality educational experience is being supplied is, however, likely to reduce this possibility. Florida Virtual High School uses only certified classroom teachers (Schnitz & Young, 2002). As the online environment becomes more competitive, it is likely that virtual schools will provide evidence of their teachers' certifications.

With conventional schools, the issue of class sizes is a perennial problem. The diversity of virtual schools means that it is not easy to determine corresponding workloads. The evaluation of Virtual High School (VHS) (Kozma et al., 2000) revealed that some of the teachers involved in the case study had to complete their VHS work at home in addition to their normal teaching load during the day. When teachers are asked to take responsibility for large groups of students, the time available for individual attention is likely to be reduced, and the quality of the educational service provided may be less satisfactory. There are indications that some virtual schools have recognized this problem. Louisiana Virtual School (2002), for example, is limited to 20 students per course.

Accreditation of courses across geographic regions will also become an increasing problem. Palloff and Pratt (2001) noted concerns with the quality of online high school programs as early as 2001. Varying standards can mean that a course in one area is not recognized in another. Students will increasingly be able to choose programs across state and even national borders and complete their schoolwork by sitting at home with their computers.

An important item relating to the quality of a student's educational experience in a virtual school is the recognition that not all students are suited to online learning. Already, some virtual schools try to determine whether the prospective student is suited to online learning by using questionnaires. Typically, these questionnaires ask students about their independent learning skills, motivation, time management abilities, and comfort with technology.

If virtual schools are perceived to be advantageous for those enrolled in them, there are also concerns as to when the access to them is seen as inequitable. Bikson and Paris (1999) found that there were "highly significant differences in household computer access based on income" (p. 9), in the United States. It is reasonable to assume that households with children will have less access to computers to use in a virtual school if they are part of a disadvantaged group. Unless there is careful planning, the use of technology-mediated education is likely, in the short term, to further entrench those inequalities that exist in society.

FUTURE TRENDS IN VIRTUAL SCHOOLS

Three broad trends can be identified in the growth of virtual schools. These are the continued expansion in the number of virtual schools, the trend from virtual high schools to virtual K-12 schools in North America, and the growth of virtual schools in other parts of the world. Research by Clark (2001) indicated that more virtual schools began their operations in the United States during the period 2000-2001 (43%) than in the previous four years combined. Fifty-one percent of virtual schools surveyed offered junior high and middle school courses as well as high school courses, and about one in four schools offered courses across the whole K-12 spectrum (Clark, 2001). In Canada, there is also evidence of growing demand for virtual schools. The two-year

cumulative growth rate for Alberta virtual schools was 125% (SAEE, 2002).

Collectively, the implication of the trends from North America toward an increased number of virtual schools and the extension of virtual schooling to K-12 is that there will be increased attention devoted to solving problems associated with virtual schooling. When virtual schools made their first appearance, it would have been possible for some educators to dismiss them because they were experimental, or ignore their existence because they catered only to a niche market of high school students. In some cases, this suggestion may still be valid, but support for virtual schooling is increasing, rather than decreasing, and the nature of what is offered is becoming more comprehensive and established.

Virtual schools have also continued to develop in other parts of the world. Russell (2006) has also observed a new trend in the evolution of virtual schools in Europe. The *conventional school* is likely to be the most common way in which online learning is used in European schools. In this schooling mode, students physically meet with their teachers in a dedicated building or classroom where online facilities are available, although some additional work may be completed at home. *Resources* are often available online from a provider, vendor, or umbrella organization. In Europe, one of the key providers is *European Schoolnet*, an international association of ministries of education from Europe and elsewhere. European Schoolnet provides a portal for members that includes a number of innovative projects for schools, in addition to policy information and online services. The European Schoolnet (EUN) describe their operations as follows:

The European Schoolnet is a unique international partnership of 26 Ministries of Education developing learning for schools, teachers and pupils across Europe and beyond. We provide insight into the use of ICT (information and communications technology) in Europe for policy-makers and education professionals. This goal is achieved through communication and information exchange at all levels of school education using innovative technologies, and by acting as a gateway to national and regional school networks. (European Schoolnet, 2005)

In the UK, a type of virtual school is being trialed that aims to re-engage school-age students into learning who have previously been out of more traditional educational systems. The project (Notschool, 2006) aims to establish a virtual community and develop students' self-esteem using new technology and community support.

In China, Hung, Chen, and Wong (2006) cite evidence that there are now more than 600,000 students enrolled in virtual schools, with more than 30 online virtual schools having been established in Beijing alone. Typically, many of these virtual schools are sponsored by enterprises. An example of one of these virtual schools is the 101 Distance Learning Center (2006) in Beijing.

In Australia, virtual schools have been established in Qld (VSS, 2006) and Tasmania (Distance Education Tasmania, 2006) by the relevant educational authorities. These virtual schools use conventional schools. Schools elect to receive or provide instruction in designated subject areas, but there are important differences in the way that these approaches to virtual schooling have evolved.

CONCLUSION

Virtual schools continue the tradition whereby students learn at a distance from their teachers. The availability of online courses through the Internet has simultaneously reduced the emphasis given to older forms of distance education and increased the opportunities for students to explore alternatives to traditional school education. It is likely that there will be an increase in the number of virtual schools and that they will continue to attract students. The expected increase in the number and type of virtual schools is likely to provide both exciting possibilities and daunting challenges.

REFERENCES

101 Distance Learning Center. (n.d.). Retrieved June 20, 2006, from http://www.chinaedu.com/english/index.html

Bikson, T. K., & Paris, C. W. A. (1999). *Citizens, computers and connectivity: A review of trends*. Retrieved January 20, 2004, from http://www.rand.org/publications/MR/MR1109/mr1109.pdf

California Virtual School Report. (2000). *The California Virtual High School report: A national survey of virtual education practice and policy with recommendations for the state of California.* Retrieved January 20, 2004, from http://www.uccp.org/docs/VHS_Report_lowres.pdf

Clark, T. (2001). *Virtual schools: Trends and issues—A study of virtual schools in the United States*. Retrieved January 10, 2006, from http://www.WestEd.org/online_pubs/virtualschools.pdf

Distance Education Tasmania. (2006). *Virtual schooling service in Tasmania, Australia*. Retrieved June 20, 2006, from http://www.distance.tased.edu.au/onlinecampus/background.htm

European Schoolnet. (2005). Retrieved August 8, 2005, from http://www.eun.org/eun.org2/eun/en/About_eschoolnet/entry_page.cfm?id_area=101

Florida High School Evaluation. (2000). *The Florida High School evaluation: 1999-2000 year-end report for the Orange County School Board*. Tallahassee, FL: Center for the Study of Teaching and Learning, Florida State University. Retrieved March 20, 2006, from http://www.flvs.net/educators/documents/pdf/archived_evals/FLVS%20Annual%20Evaluations/99-2000/99-2000%20Year%20End%20Evaluation.pdf

Florida High School Evaluation. (2002). *The Florida High School evaluation: 1999-2000 year-end report for the Orange County School Board*. Tallahassee, FL: Center for the Study of Teaching and Learning, Florida State University. Retrieved December 12, 2003, from http://www.flvs.net/_about_us/pdf_au/fhseval_99-00.pdf

Fraser Valley Distance Education School. (2006). Retrieved March 20, 2006, from http://www.fvdes.com/grade7/sciencefair.html

Hung, D., Chen, D.-T., & Wong, A. F. L. (2006). An overview of virtual learning environments in the Asia-Pacific: Provisos, issues and tensions. In J. Weiss, J. Nolan, J. Hunsinger, & P. Trifonas (Eds.), *The international handbook of virtual learning environments* (Vol. 1, pp. 699-721). Dordrecht: Springer.

Kozma, R., Zucker, A., Espinoza, C., McGee, R., Yarnell, L., Zalles, D., et al. (2000). *The online course experience: Evaluation of the Virtual High School's third year of implementation, 1999-2000*. Retrieved January 9, 2004, from http://www.sri.com/policy/ctl/html/vhs.html

Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V., & Crawford, C. (2002). Internet paradox revisited. *Journal of Social Issues, 58*(1), 49-74.

Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukopadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist, 53*(9), 1017-1031.

Louisiana Virtual School. (2002). Retrieved January 10, 2004, from http://www.icet.doc.state.la.us/distance

Louisiana Virtual School. (2006). Retrieved March 21, 2006, from http://www.louisianavirtualschool.net/

Moll, M. (1998). No more teachers, no more schools: Information technology and the "deschooled" society. *Technology in Society, 20*(3), 357-369.

NEA. (2006). *Guide to online high school courses*. Washington, DC: National Education Association. Retrieved March 23, 2006, from http://www.nea.org/technology/onlinecourseguide.html

Notschool. (2006). Retrieved July 19, 2006, from http://www.notschool.net/ns/template.php?id=about

Open School. (2002). *Open School in British Columbia, Canada.* Retrieved December 2, 2003, from http://openschool.bc.ca

Palloff, R. M., & Pratt, K. (2001). *Lessons from the cyberspace classroom: The realities of online teaching*. San Francisco: Jossey-Bass.

Perelman, L. (1992). *School's out: Hyperlearning, the new technology and the end of education*. New York: William Morrow and Company.

Russell, G. (2002). *Responsibility for school education in an online globalised world*. Focus paper presented to Technology Colleges Trust Vision 2020 Online Conference, UK.

Russell, G. (2004). Virtual schools: A critical view. In C. Cavanaugh (Ed.), *Development and management of virtual schools: Issues and trends* (pp. 1-25). Hershey, PA: Information Science Publishing.

Russell, G. (2005, September). *The knowledge economy and virtual schooling*. Paper presented at the European Conference on Educational Research (ECER 2005), Dublin, Ireland. Retrieved January 6, 2005, from http://www.leeds.ac.uk/educol/documents/142870.htm

Russell, G. (2006). Online and virtual schools in Europe. *European Journal of Open, Distance and E-Learning*. Retrieved June 18, 2006, from http://www.eurodl.org/materials/contrib/2006/Glenn_Russell.htm

Russell, G., & Russell, N. (2001). Virtualisation and the late age of schools. *Melbourne Studies in Education, 42*(1), 25-44.

Rutherford, T. (1993). Democracy, markets and Australian schools. In C. James, C. Jones, & A. Norton (Eds.), *A defence of economic rationalism* (pp. 151-159). St. Leonards, Australia: Allen and Unwin.

SAEE. (2002). *Executive summary of e-learning: Studying Canada's virtual secondary schools*. Retrieved December 3, 2003, from http://www.saee.bc.ca/vschoolsum.html

Salmon, G. (2000). *E-moderating: The key to teaching and learning online*. London: Kogan Page.

Schnitz, J., & Young, J. E. (2002). *Models of virtual schooling*. Retrieved January 10, 2004, from http://www.can.ibm.com/k12/pdf/Virtualschool.pdf

VSS. (2006). *Virtual Schooling Service in Qld, Australia.* Retrieved June 20, 2006, from http://www.learningplace.com.au/default_community.asp?orgid=64&suborgid=368

KEY TERMS

Bricks-and-Mortar Schools: Traditional schools, where students attend at a physical school building.

Distance Education: A generic term referring to education where teachers and students are geographically separate. Modes employed include print and non-print technologies.

Experiential Learning: Learning based on direct and unmediated instruction or on physical interaction with people and materials.

Globalization: The bypassing of traditional geographic borders using information technology to enable global orientation of business and remote curriculum delivery.

Interactivity: The relationship between the learner and the educational environment.

Socialization: The process by which students internalize the norms and values necessary for living in a civilized community.

Virtual School: A form of schooling that uses online computers for part or all of a student's education.

Trends in Information Technology Governance

Ryan R. Peterson
Information Management Research Center, Spain

INTRODUCTION

Information technology (IT) *governance* has been a perennial item on the corporate agenda of many organizations. Ever since IT proved to be more than an administrative tool, researchers and practitioners have pondered its governance. Defined as the locus of IT decision-making authority (Brown & Magill, 1994; Sambamurthy & Zmud, 1999), discussions concerning IT governance have flourished for more than four decades across research communities and boardrooms. Posed as a question of *centralization* during the 70s, IT governance drifted towards *decentralization* in the 80s, and the recentralization of IT decision-making was a 90s trend.

Today, IT governance is experiencing yet another transformation, and persists as a complex and evolving phenomenon (Grembergen, 2003). As business environments continuously change and new technologies evolve rapidly, how to govern IT effectively remains an enduring and challenging question. This chapter discusses past developments and the present status quo of IT governance, and outlines several critical questions, which are pending future investigation.

BACKGROUND

Traditionally, three IT governance models have been distinguished (Brown & Magill, 1998; Sambamurthy & Zmud, 1999). In each model, stakeholder constituencies take different lead roles and responsibilities for IT decision-making across the *IT portfolio*. In the centralized model, corporate IT management has decision-making authority concerning *IT infrastructure* and *IT applications*. In the decentralized model, division IT management and business management have authority for IT infrastructure and IT applications. In the *federal* model, corporate IT has authority over IT infrastructure, and (either or both) division IT and business-units have authority over IT applications.

In general, it is argued that centralization provides greater efficiency, control, and standardization, while decentralization improves business ownership, flexibility, and responsiveness (Brown, 1997; Rockart, Earl, & Ross, 1996). Literature suggests that the federal model provides the benefits of both centralization and decentralization (see Table 1). Research indicates that organizations adopt a federal model when pursuing multiple competing objectives involving a simultaneous focus on cost-efficiency and business-flexibility

Table 1. Drivers and design of IT governance (Adapted from Hodgkinson, 1996; Peterson, O'Callaghan, & Ribbers, 2000; Sambamurthy & Zmud, 1999)

Model / Drivers	Centralized IT Governance	Decentralized IT Governance	Federal IT Governance
Synergy	+	-	+
Standardization	+	-	+
Specialization	+	-	+
Customer responsiveness	-	+	+
Business ownership	-	+	+
Flexibility	-	+	+

1.

(Peterson, O'Callaghan, & Ribbers, 2000; Sambamurthy & Zmud, 1999).

MAIN THRUST

While the federal model seems to be the dominant configuration in contemporary firms (Peterson, O'Callaghan, & Ribbers, 2000; Sambamurthy & Zmud, 1999), empirical studies regarding the complexity of this configuration are sparse. Specifically, *allocation of IT decision-making authority does not resolve the need for effective coordination between corporate IT, division IT and business-unit management*. Continuous differentiation leads to fragmentation, unless a corresponding process of integration complements it. The problems reported in practice and research regarding the lack of, for example, IT prioritization, top management IT commitment, IT management business understanding, business management IT responsibility, and IT value generation, are symptomatic of this fragmentation and are typically encountered in the federal IT governance model (Peterson, 2001; Weill & Broadbent, 1998).

In order to provide direction and achieve organizational effectiveness, differentiation begets integration (Daft, 1998; Galbraith, 1994; Lawrence & Lorsch, 1967). Designing effective IT governance is dependent on both the *differentiation* and *integration* of decision-making for IT across the portfolio of business IT investments and processes (see Figure 1).

Whereas differentiation focuses on the distribution of IT decision-making rights and responsibilities among different stakeholders in the organization (i.e., the locus of IT decision-making), integration focuses on the coordination of IT decision-making/-monitoring processes and structures across stakeholder constituencies. Organizations thus need to consider and implement integration mechanisms for the effective governance of IT.

FUTURE TRENDS

Integration mechanisms for IT governance can be classified according to two dimensions (Peterson, 2003). Vertically, integration mechanisms focus either on integration structures or integration processes; whereas horizontally, a division is made between formal positions and processes, and relational networks and capabilities. Collectively, this provides four types of generic integration mechanisms for IT governance (see Figure 2).

Formal integration structures involve appointing IT executives (e.g., CIO) and IT functions (e.g., client-account and user relationship managers), and institutionalizing special and standing IT committees and councils. Committees and/or

Figure 1. Differentiation and integration IT decision-making (Adapted from Weill & Broadbent, 1998)

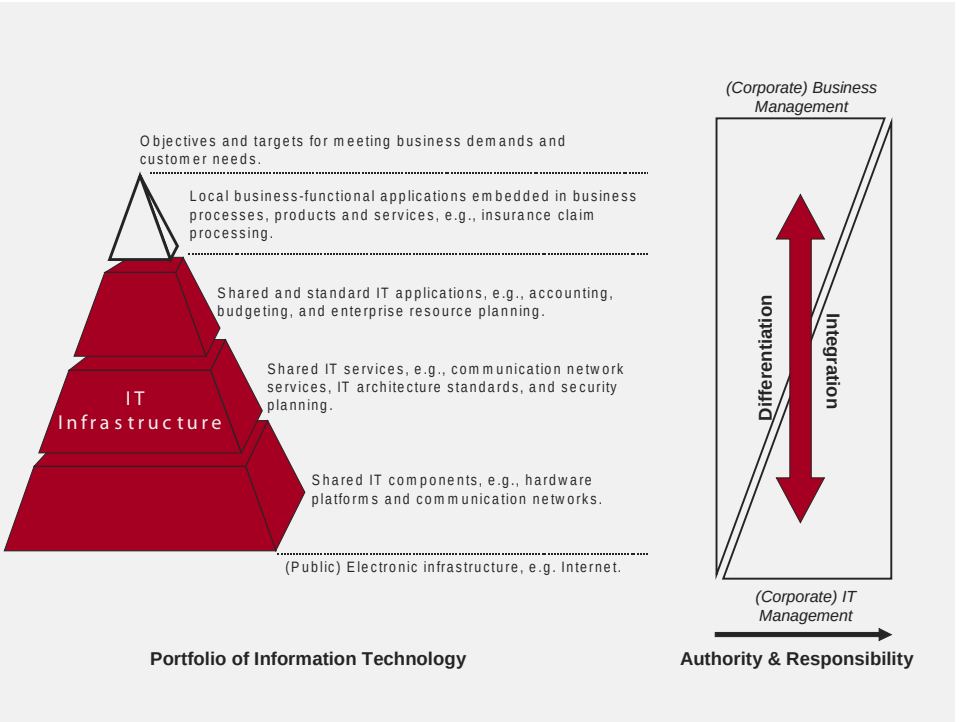
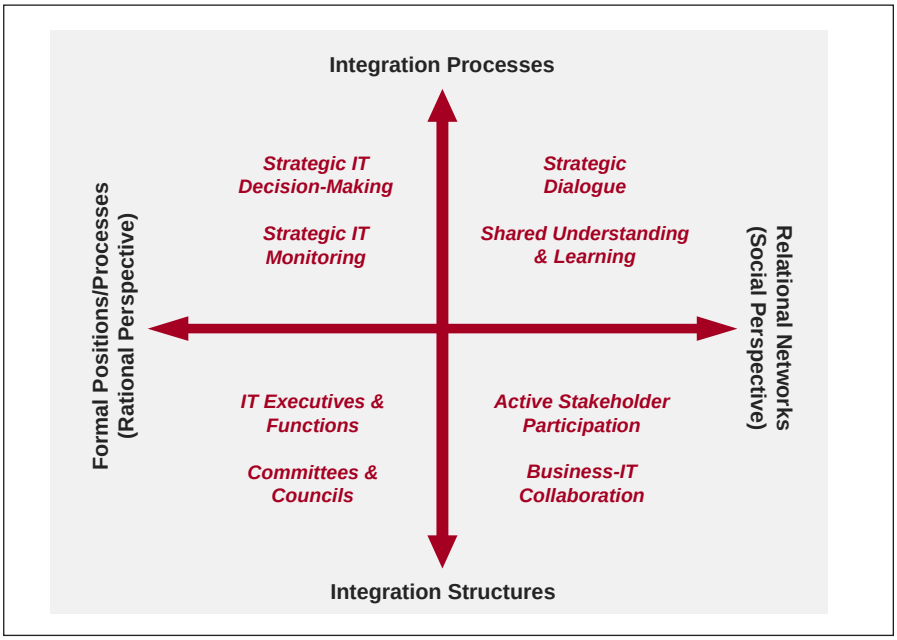


Figure 2. Generic integration mechanisms for IT governance (Adapted from Peterson, 2003)



executive teams can take the form of temporary task forces (e.g., project steering committees), or can alternatively be institutionalized as an overlay structure in the organization in the form executive and/or IT management councils. Formal integration processes describe the formalization and institutionalization of IT decision-making-/monitoring procedures and performance. Formal integration processes vary with levels of IT governance *comprehensiveness*, *formalization*, and *integration*.

Whereas the foregoing formal integration mechanisms tend to be mandatory and tangible, relational integration mechanisms are "voluntary" and "tacit" actions, which cannot be programmed and/or formalized. While formal integration mechanisms are necessary, they are insufficient for designing effective IT governance in competitive environments (Peterson, O'Callaghan, & Ribbers, 2000).

Relational integration structures involve the active *participation* and *collaboration* between corporate executives, IT management, and business management. Central to relational integration is the participative behavior of different stakeholders to clarify differences and solve problems in order to find integrative solutions. The ability to integrate relationally allows an organization to find broader solutions, and unleashes the creativity involved in joint exploration of solutions that transcend functional boundaries and define future possibilities. Relational integration processes describe strategic dialogue and *shared learning* between principle business and IT stakeholders. Strategic IT dialogue incorporates a wide range of initially unstructured business perspectives and IT views, and involves rich conversation to resolve diverging perspectives and stakeholder conflicts. Shared learning describes the co-creation of mutual understanding by members of organizational sub-units of each other's goals and objectives.

In summation, formal and relational structures and processes collectively constitute and determine the integration capability of IT governance, and underscore the importance of flexible management systems in complex and uncertain environments. The organizing logic is characterized by a collaborative network, where communication is more likely to be lateral, task definitions are more fluid and flexible (i.e., related to competencies and skills, rather than being a function of position in the organization), and where influencing of business-IT decisions is based on expertise, rather than an individual's (or group's) position (Peterson, 2003).

CONCLUSION

Amidst the challenges and changes of the 21st century, involving hyper-competitive market spaces, electronically-enabled global network businesses, and corporate governance reform, IT governance has become a fundamental business imperative. IT governance is a top management priority, and

Table 2. Future research

Suggested Research Questions	Suggested Research Design and Methodology
What types of integration mechanisms are used in conjunction with different IT governance models? Do certain IT governance *gestalts* exist that combine IT governance differentiation and integration mechanisms?	A quantitative approach (large scale survey) to identify and validate patterns of IT governance differentiation and integration. Multiple case studies can also be used to explore and discover new patterns.
What are the most effective integration mechanisms under different IT governance models?	A quantitative approach (large scale survey) to identify effective integration mechanisms for IT governance. Multiple case studies can be of value when considering the business and industry context.
How do integration mechanisms moderate the relationship between IT governance and IT business value?	A quantitative approach (large scale survey) to measure and validate the moderating impact of integration mechanisms on IT business value realization
How do organizations design and implement integration mechanisms for IT governance?	A longitudinal research design involving field-research and multiple case studies. A single in-depth case study can also provide rich (theoretical) insights.
What are the (inter-) dependencies between integration mechanisms? Do companies follow a specific path when developing and implementing integration mechanisms (e.g., do they start with positions, the processes and finally relationships?)	A longitudinal or retrospective field-study on the development and evolution of integration mechanisms in certain companies
Are certain drivers (e.g., standardization, responsiveness, flexibility) related to the use (or non-use) of specific integration mechanisms? How do the competing drivers of efficiency and flexibility impact the adoption and usage of integration mechanisms?	Both surveys and/or case studies can be used. A (quantitative) survey would identify empirically significant relationships amongst certain drivers and integration mechanisms. Multiple case studies would help explain why and how these (competing) drivers lead to the adoption and usage of integration mechanisms

rightfully so, because it is the single most important determinant of IT value realization (Mata, Fuerst, & Barney, 1995; Peterson, 2001; Rockart, Earl, & Ross, 1996; Sambamurthy & Zmud, 1999; Weill & Broadbent, 1998).

More than simply assigning and allocating IT decision-making authority, IT governance is the system by which an organization's IT portfolio is directed, and describes the distribution of IT decision-making rights and responsibilities among different stakeholders in the organization, and the rules and procedures for making and monitoring decisions on strategic IT concerns. These rules and procedures address the integration mechanisms, which are fundamental to effective IT governance.

Nevertheless, research has only recently focused on the use and effectiveness of integration mechanisms for IT governance. More research is definitely and urgently required in this area (Table 2). It is only through empirical investigation of these and other related research questions that we will be able to advance the current body of knowledge and understanding in the area of IT governance.

REFERENCES

Brown, C.V. (1997). Examining the emergence of hybrid IS governance solutions: Evidence from a single case site. *Information Systems Research, 8*(1), 69-94.

Brown, C.V., & Magill, S.L. (1994, December). Alignment of the IS functions with the enterprise: Toward a model of antecedents. *MIS Quarterly*, 371-403.

Brown, C.V., & Magill, S.L. (1998). Reconceptualizing the context-design issue for the information systems function. *Organization Science, 9*(2), 177-195.

Daft, R.L. (1998). *Organization theory and design* (6th ed.). South Western College Publishing.

Galbraith, J.R. (1994). *Competing with flexible lateral organisation*. MA: Addison-Wesley.

Grembergen, W. van (2003). Strategies for information technology governance. Hershey, PA: Idea Group Publishing.

Hodgkinson, S.T. (1996). The role of the corporate IT function in the federal IT organisation. In M.J. Earl (Ed.), *Information management: The organisational dimension*. Oxford University Press.

Lawrence, P.R., & Lorsch, J.W. (1967). *Organisation and environment. Managing differentiation and integration.* Harvard University Press.

Mata, F.J., Fuerst, W.L., & Barney, J.B. (1995). Information technology and sustained competitive advantage: A resource-based analysis. *MIS Quarterly*, *19*(4), 487-505.

Peterson, R.R. (2001). Information governance: An empirical investigation into the differentiation and integration of strategic decision-making for IT. Tilburg University, The Netherlands.

Peterson, R.R. (2003). Integration strategies and tactics for information technology governance, In W. van Grembergen (Ed.), *Strategies for information technology governance*, Hershey, PA: Idea Group Publishing.

Peterson, R.R., O'Callaghan, R., & Ribbers, P.M.A. (2000). Information technology governance by design. *Conference Proceedings of the Twenty-First International Conference on Information Systems*, Brisbane, Australia.

Rockart, J.F., Earl, M.J., & Ross, J.W. (1996, Fall). Eight imperatives for the new IT organization. *Sloan Management Review*, 43-55.

Sambamurthy, V., & Zmud, R.W. (1999). Arrangements for information technology governance: A theory of multiple contingencies. *MIS Quarterly, 23*(2), 261-290.

Weill, P., & Broadbent, M. (1998). *Leveraging the new infrastructure: how market leaders capitalise on information technology.* Boston, MA: Harvard Business School Press.

KEY TERMS

Centralized Model: The concentration of decision-making in a single point in the organization, in which a single decision applies.

Collaboration: A close, functionally interdependent relationship, in which organizational units strive to create mutually beneficial outcomes. Collaboration involves mutual trust, the sharing of information and knowledge at multiple levels, and includes a process of sharing benefits and risks. Effective collaboration cannot be mandated.

Decentralized Model: The dispersion of decision-making, in which different independent decisions are made simultaneously.

Differentiation: The state of segmentation or division of an organizational system into subsystems, each of which tends to develop particular attributes in relation to the requirements posed by the relevant environment. This includes both the formal division, as well as, behavioral attributes of the members of organizational subsystems.

Federal Model: A hybrid configuration of centralization and decentralization, in which decision-making is differentiated across divisional and corporate units.

Integration: 1) The process of achieving unity of effort among various subsystems in the accomplishment of the organizational task (*process focus*). 2) The quality of the state of collaboration that exists among departments, which is required to achieve unity of effort by the demands of the environment (*outcome focus*).

IT Applications: Local business-functional applications embedded in business processes, activities, products and/or services.

IT Governance: 1) Locus of IT decision-making authority *(narrow definition).* 2) The distribution of IT decision-making rights and responsibilities among different stakeholders in the organization, and the rules and procedures for making and monitoring decisions on strategic IT concerns *(comprehensive definition).*

IT Governance Comprehensiveness: Degree to which IT decision-making/-monitoring activities are systematically and exhaustively addressed.

IT Governance Formalization: Degree to which IT decision-making/-monitoring follows specified rules and standard procedures.

IT Infrastructure: The base foundation of the IT portfolio, delivered as reliable shared services throughout the organization, and centrally directed, usually by corporate IT management.

IT Governance Integration: Degree to which business and IT decisions are integrated administratively, sequentially, or reciprocally.

IT Portfolio: Portfolio of investments and activities regarding IT operations and IT developments spanning IT infrastructure (technical and organizational components) and IT applications.

Participation: Process in which influence is exercised and shared among stakeholders, regardless of their formal position or hierarchical level in the organization.

Shared Learning: The co-creation of mutual understanding by members of organizational sub-units of each other's goals and objectives.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2865-2870, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Trends in the Higher Education E-Learning Markets

John J. Regazzi
Long Island University, USA

Nicole Caliguiri
Long Island University, USA

INTRODUCTION

This article describes research undertaken at the Scholarly Communications Lab of the College of Information and Computer Science at Long Island University in the area of higher education e-learning market in the United States. It is organized around three topics: a definition of e-learning and distance education; a description of the size, growth, and future outlook for this market; and the identification of some of the key growth drivers both historically and for the future.

The distant education market is now a mature market and has been around for a long while, with its antecedents established decades ago. The market has grown substantially in the last 10 years, and will continue to grow significantly, with an estimated market size of over $17 billion in 2010, representing a penetration rate of over 30% of the total higher education market in the United States. Many institutions of higher education have embraced some type of online programs, with 96% of universities with enrollments of 15,000 students having some e-learning programs. Major new entrants from the for-profit sector are now active in this marketplace. With the combination of new entrants and new technology advancements, the opportunities for value creation have increased while the nature of competition has intensified substantially.

BACKGROUND

According to Western Governors University (USA), distance learning and e-learning is simply:

Education that takes place when the instructor and student are separated by space and/or time. The gap between the two can be bridged through the use of technology—audio tapes, videoconferencing, satellite broadcasts and online technology, just to name a few—and/or more traditional delivery methods, such as the postal service. (WGU, 2004)

As communications and computer technology evolve, the definition of distance learning continues to develop. Asynchronous or time-delayed computer conferencing has given institutions the capability to network groups of learners over a period of time, allowing students in distance learning programs to be taught in groups rather than as individuals (Gunawardena & McIsaac, 2003).

Distance learning, before it evolved into primarily e-learning, has been around for a long period of time. It began in the second half of the 19th century with the exchange of print materials, assignments, and feedback by mail. Over the course of the 20th century, the development of radio and television made the delivery of additional materials (lectures and demonstrations) by electronic means possible. The 1950s saw the growth of a number of video projects that sought to identify expert science, math, and language teachers who could spread their expertise to students across a region or across the whole country. In 1989, Congress enacted the Star Schools legislation, intended to deliver quality instruction to largely rural or underserved areas. Among the Star School projects were three courses designed for adult learners, two of which used a studio teacher providing regular classes on topics ranging from job-seeking skills to skill-building needed to qualify for the GED. (www.ed.gov/prog_info/StarSchools/).

CURRENTLY CHANGING DEVELOPMENTS

More recent technologies have expanded the number of communication channels available to distant educators. E-mail and computer conferencing began in the early 1970s as part of the government sponsored ARPANET (Advanced Research Projects Agency Network) (Harasim, Hiltz, Teles, & Turoff, 1995). Although scientific work groups quickly adopted these communicative tools to advance collaborative activity at a distance, they were not available to educators and off-campus students for another decade. In education, these tools could permit learners to exchange and debate

ideas. But only in recent years have educators recognized the potential of these tools to support a different model of distance education, a model built on more constructivist principles of learning. Over the 20th century, the technological possibilities have changed, although the pedagogical model has not (Tolmie & Boyle, 2000). Most distance courses that use the modern information handling technologies are still built on a transmission model in which instructors create material to be consumed by learners, and learners are given exercises and tests that they submit to the instructor demonstrating their mastery of the material; that they understand it, remember it, and can apply this knowledge in testing situations (Askov, 2003).

In the 1990s, new tools became available to the scientific community: the Internet and the Web. By the mid 1990s, these were made available to the broader public. Educators recognized the potential of these technologies immediately, and a few distance educators began to recommend a new model of education that emphasized the qualitative improvements in learning itself, if learners had ready access to a variety of electronic materials and were supported in examining and discussing these materials with other learners. These educators sought to distinguish this form of distance learning from more traditional forms by using new terms: distributed or flexible learning. (Bates, 2000) It is estimated that by the year 2012, schools and colleges will routinely use "computerized teaching programs and interactive television lectures and seminars, as well as traditional methods" ("Emerging," 2003, p. 8). Videoconferencing and other technologies will also help enrich media and provide many benefits of face to face instruction (Wonacott, 2002).

In 2003, the first in a series of annual reports by The Sloan Consortium on the state of online learning in U.S. higher education, "Sizing the Opportunity: The Quality and Extent of Online Education in the United States, 2002 and 2003" was released. The initiation of this annual study emerged from a search for an authoritative answer to the question: "How many students are learning online?"

Market Size estimates. The answer determined by that first study was that for the fall 2002 term, slightly more than 1.6 million students took at least one online course at U.S. degree granting institutions. This same study asked institutions to predict the rate of growth (or decline) in their online enrollments for the following year, and respondents projected an average annual growth rate of 19.8%. This number was substantially above the annual rate of increase in the overall population of higher education students, whose annual growth has been estimated as between 0.8 and 1.3% (Allen & Seaman, 2006). It is evident that higher education administrators have been both optimistic and aggressive in their views of the growth of online learning in the U.S. This view of the marketplace has been validated, and significant investments into this market for e-learning courses have followed consistently. For example, the second annual study, "Entering the Mainstream, The Quality and Extent of Online Education in the United States, 2003 and 2004," found that the overall growth in the number of online learners actually exceeded the optimistic projections of the previous year, increasing at a 22.9% rate, to reach 1.9 million online students for fall 2003 (Allen & Seaman, 2006).

The strength of this market in terms on new students, new courses, and new entrants continue unabated. In 2003, the yearly increase of about 360,000 new online learning students was matched by the results of the 2005 study, Growing by Degrees, Online Education in the United States, 2005, with more than 2.3 million students taking at least one online course during the 2004 fall term. Despite a lower percentage increase reported in 2005 of 18.2%, as there is an increasing larger base population, there are more and more students taking online courses and further there are increasingly more students taking online courses for the first time. This trend was further illustrated in the results the fourth annual Sloan Consortium study showing there has been no leveling in the growth rate and strength of this market. During the 2005 term higher education institutions taught nearly 3.2 million online students, an increase of about 850,000 students and a growth rate of 35%. 2005 marks both the largest increase in the number of online students and the largest percentage increase since these tracking surveys have begun. In 2005,

Table 1. Students taking at least one online course—Fall 2005 (Source: Allen & Seaman, 2006)

Undergraduate	2,621,713
First Professional	39,350
Graduate	443,827
Other for-credit	75,159
Total	3,180,050

the overall size of the higher education student population is sized at approximately 17 million with the number of students taking online courses at approximately 3 million or now representing nearly 17% of all higher education students (Allen & Seaman, 2006) (Table 1).

Using our current estimates of growth, the current 2007 market size is estimated between 4.5 to 5.0 million students enrolled in e-learning courses or 25% to 30% of the overall higher education market. The demand for this form of education is continuing steadily and strongly.

In meeting this strong and increasing demand for online education, it appears that the large universities have responded first and most strongly. The size of an institution has a clear impact on the average number of online students at institutions. The largest institutions (defined as institutions with enrollments of 15,000 or more), for example, are each teaching an average of more than 3,200 online students at the undergraduate level alone. This compares to about half that amount (1,500 online students) for the next smaller-sized institution type (those with overall enrollments between 7,500 and 14,999). It is estimated that currently the average number of online students enrolled is proportional to the size of the institution (Kariya, 2003). In 2005, for each institutional size type, the average number of students is around 20% of the lower range of each size type, except for the smallest of the institutions (those with less than 1,500 total students) which is below this level, while the largest institutions are slightly above this average. However it is estimated, the proportion of the student population that is taking at least one online course has begun to reach a level that institutions of all sizes must address this as an important educational issue (Table 2).

The Doctoral/Research and Master's institutions have the largest average online total enrollments since they are more likely on average to be the largest schools. Associates institutions offering 2 year only Associate Degree programs also have a sizable average online enrollment (nearly 800 undergraduate students per institution), but the large number of Associates institutions is what accounts for the large number of online students at such schools (Allen & Seaman, 2006) (Table 3).

Course offerings and institutional profiles. Evidence has shown in the past a very uneven distribution of online course and program offerings by type of institution. Public institutions and the largest institutions of all types have consistently been at the forefront of online offerings. Those that are the least likely to offer online courses, and typically have the most concern about online education in general, have been the small, private, 4-year institutions. It must be taken into account that not all schools offer online courses, and not all schools that have online courses offer fully online programs (Hickman, 2003). Examining the pattern of online offerings

Table 2. Mean undergraduate online enrollment by size of institution—Fall 2005 (Source: Allen & Seaman, 2006)

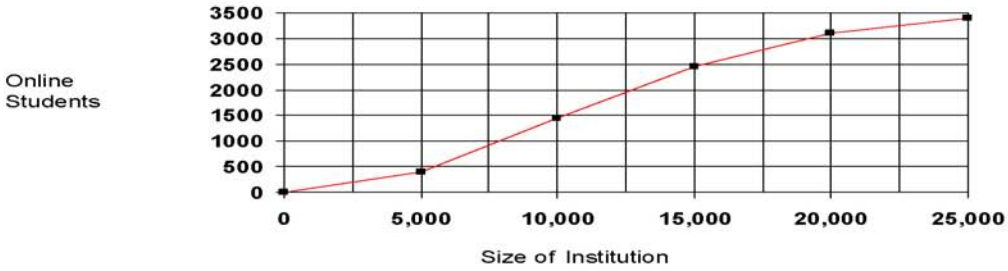


Table 3. Mean number of online students per institution—Fall 2005 (Source: Allen & Seaman, 2006)

	Doctoral/Research	Masters	Baccalaureate	Associates	Specialized
Undergraduate	1017.1	988.3	148.7	797.5	84.2
Professional	21.7	1.1	0.8	2.6	10.3
Graduate	520.6	365.5	14.1	0.3	47.1
Other for credit	16.9	9.7	0.6	23.3	1.8

does show some interesting patterns when the results are compared to the distribution of online students. For example, Doctoral/Research institutions, which enroll 13% of all online students, have the greatest penetration of offering online programs as well as the highest overall rate (more than 80%) of having some form of online offering (either courses or full programs). Although Associates schools have by far the largest contingent of online students, they trail both Doctoral/Research and Master institutions in the proportion with online programs or any type of online offering.

More than 96% of the very largest institutions have some online offerings, which is more than double the rate observed for the smallest institutions. The proportion of institutions with fully online programs rises steadily as institutional size increases, and about two-thirds of the very largest institutions have fully online programs, compared to only about one-sixth of the smallest institutions. Doctoral/Research institutions have the greatest penetration of offering online programs as well as the highest overall rate (more than 80%) of having some form of online offering. The proportion of institutions with fully online programs rises steadily as institutional size increases, and about two-thirds of the very largest institutions have fully online programs, compared to only about one-sixth of the smallest institutions. Additionally, Doctoral/Research institutions are far more likely to have online programs than are Baccalaureate institutions (Table 4).

Market competitive issues. With markets no longer defined by geography and as the quality and quantity of online courses increases, the trend toward electronic learning is perhaps entering a new era of growth and competition in higher education.

According to a 2004 report by Eduventures, while enrollment in fully online distance education programs is growing at 30% (Gallagher, 2004), the average tuition per student is approximately $5,500 and represents a total market of about $5 billion in tuition revenue for 2004. Our estimates suggest that the fully online distance education will represent a market of $17 billion by 2010. The overall revenue market size for fully online education is expected to grow approximately 38% in 2004. The estimated growth rates for the number of students enrolled in online education programs will remain above 20% for at least the next 3 to 5 years, slowing as the market matures further (Table 5).

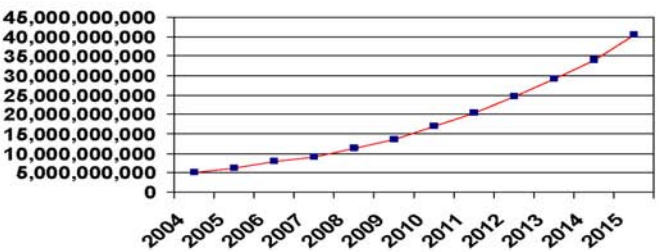
It is estimated that undergraduate student enrollment programs will drive approximately 73% of revenues in the online education market in 2004. The majority of students in online education programs are undergraduate students. However, the share of the online market regarding graduate level education is greater than its share of campus-based education, mostly due to the disproportionate number of working adults in graduate programs. It is expected that the undergraduate market will continue to grow, while the graduate-level programs will experience a surge in the coming years due to the current undergraduates continuing their education after obtaining their undergraduate degree.

For-profit institutions have taken a significant market share of the online education market, clearly disproportionate to their share of higher education overall market. For-profit institutions were early entrants and aggressive investors in this market and these factors coupled with their programs typically priced higher than nonprofit institutions, have permitted the for-profits a command of nearly 44% of the

Table 4. Online offerings—Fall 2005 (Source: Allen & Seaman, 2006)

	Doctoral/Research	Masters	Baccalaureate	Associates	Specialized	Total
Online Programs	55.7%	43.6%	17.2%	31.2%	26.0%	31.4%
Courses Only	24.9%	33.9%	24.0%	39.8%	22.7%	31.5%
No Online	19.4%	22.5%	58.8%	29.0%	51.3%	37.0%

Table 5. Fully online distance education enrollment (Source: Gallagher, 2004)



online market revenues in 2004, up from 30% in 2000. In fact, online education revenues in the for-profit section of the market have grown by more than 50% per year for several years and have grown at an estimated 25 to 30% per year through 2006. The for-profit sector represents significant and sustained competition in the online higher education market. As the overall higher education market moves to a more online offering, for-profit institutions will become a more significant player in the overall market.

FUTURE TRENDS

In reviewing the development of the last 20 years of the online higher education market, we see the growth of this market in three distinct phases. Each phase is characterized by a different competitive advantage. These are: Phase1. Technology Superiority; Phase 2. Breath of Content Offerings; and Phase 3. Brand equity.

1. Technology Superiority 1980-1997

The 1980s and early1990s represented the pioneering days of online distance education. With the emergence and availability of the World Wide Web, this set the stage for colleges and universities with a history of distance programs to then expand and tap into a broader market. During this phase, student enrollment grew exponentially as leading colleges launched and marketed new and innovative programs.

- University of Phoenix launches online division via Compuserve
- Institutions with histories of distance education and outreach development Web-based courses
- Early course management system software emerges

This phase of development is characterized by a competitive landscape where the leaders made significant investments in the technology platforms and software systems that allow for the production and distribution of online education courses. In this initial phase, those institutions that held the technology to develop these courses held a significant competitive advantage. The cost of technology was a significant barrier to entry, and those organizations that had access to capital and a taste for risk represented the early leaders. With this profile it is not surprising that the for-profit institution gained early and significant market share from the onset.

2. Building the Inventory of Course Offerings 1998-2003

The promise of online education as a market was enhanced by the Internet "Gold Rush" of the late 1990s. Private investors began to invest aggressively in for-profit universities and software providers in the higher education market with more than $1.2 billion in capital in 1999 and 2000 alone, the majority of which was devoted to e-learning. This capital provided not only marketing dollars and infrastructure to enable distance education to take off, but during this phase most importantly much of this capital was used to develop a strong and robust portfolio of course offerings. These course offerings were developed by for-profits both on their own and in partnership with other higher education institutions. Nonprofit educational institutions built this inventory of courses by offering incentives to their faculty for the production of courses, while developing training programs for faculty to learn this new form of distance education.

With technologies and a robust inventory of courses in place, these years also triggered the initial boom in consumer awareness of online education. Marketing and organizational development became the pivotal factor in defining competitive winners and losers in this phase. For example,

- University of Phoenix spins out an its online subsidiary in an initial public offering (IPO), the first of its kind in this area
- Nonprofit universities such as NYU, Columbia, and the University of Maryland University College create for-profit spin offs to tap Internet course opportunities and have the access to capital available in this sector
- In order to leverage their investments in technology and content, collective advertising and marketing expenditures for online education top $200 million at for-profit institutions
- The number of colleges offering online education programs balloons into the thousands
- U.S. News and World Report launches an annual e-learning section evaluating online education programs.

Course development still continues today. Online courses now can be made available, but the faculty training and incentives needed to develop these programs are significant, particularly for those institutions that have not embraced online programs. We estimate that this development of content is still a barrier of entry for institutions without substantial programs now in place and will represent increasingly substantial investments for these institutions to "catch up" and try to level the competitive landscape for themselves.

3. Market Development and Brand Equity 2004 and Forward

Colleges and universities at all levels have utilized online education to tap new markets of working adults for whom participation in traditional postsecondary education was inconvenient or impossible. Institutions recognizing the potential of online education have made important investments not only in infrastructure, but also in faculty training, student services, and marketing to participate in the growth of the online education market.

Recognizing the growth opportunity, institutions are investing more heavily than they have in the past. Effective marketing is becoming more targeted and segmented as distinct value propositions, areas of focus and brands emerge. Additionally, many of the students of the early boom years are graduating, creating a powerful new word of mouth and viral marketing channel. For instance,

- Majority of degree-granting institutions, particularly large institutions over 15,000 students (at 96% for this group), are offering online education programs
- Capella University was selected by Wal-Mart as a preferred education provider
- University of Phoenix Online enrollments approach 100,000, nearly 50% of all enrollment at the nation's largest private university
- Nonprofit institutions partner with for-profit education experts such as UNext, Capella and Collegis in order to launch best-of-breed online offerings.

Many of these trends in higher education will influence the future of distance learning. Student enrollments are growing and with this learner profiles are changing, and students are becoming more demanding in the quality, breath and convenience of offerings, and as a result are shopping for educational programs that can be easily tailored to their needs (Brodsky, 2003). The Internet and other information technology devices are becoming more ubiquitous while technological fluency is becoming a common expectation. As these expectations are common and as the technology is broadly available, we see the competitive landscape moving beyond technology and content, and residing much more around the reputations of the online programs themselves. Though innovation will always be a factor, as will the breath of content offering, the key emerging competitive differentiate is likely to be the perceived quality of the offering and the support services around those programs. Customers will expect, demand, and evaluate the value of these online offerings based upon their complete experience, including the market positioning, the program support, and administrative functions, as well as the technology and quality of the faculty and content (Bates, 2003).

CONCLUSION

Technological advances and increased fluency will continue to open opportunities for distance education. Although higher education institutions are changing to favor distance education, the complexities of major transformations will require substantial investment and planning. As Bates (2000) suggests, perhaps "the biggest challenge [in distance education] is the lack of vision and the failure to use technology strategically" (p. 7). Though this challenge is understandable, given the complexity of the issues involved, institutions which now wish to enter this market will find the barriers to be substantial. Universities without e-learning programs may easily find the distribution technologies that will launch their programs, but their challenges will be in building an inventory of courses, faculty competent in this new educational media, the creation of adequate administrative and support services, and most importantly the need to market and build brand equity in their new e-learning programs. Further, institutions currently active will strengthen their distance-learning strategic plans by identifying and understanding distance-education trends for new student enrollments, faculty support, and larger academic, technological and economic issues (Howell, Williams, & Lindsay, 2003). The gap will increase between the current players and new aspiring entrants, who will need to be willing to make significant financial investments with a clear development plan, if they hope to catch up.

REFERENCES

Allen, E., & Seaman, J. (2006). *Making the grade. Online education in the United States.* Retrieved December 14, 2007, from http://www.sloan-c.org/publications/survey/pdf/making_the_grade.pdf

Askov, E. (2003). Expanding access to adult literacy with online distance education. *National Center for the Study of Adult Learning and Literacy.* Retrieved December14, 2007, from http://www.ncsall.net/fileadmin/resources/research/op_askov.pdf

Bates, T. (2000). *Distance education in dual mode higher education institutions: Challenges and changes*. Retrieved December 14, 2007, from http://bates.cstudies.ubc.ca/papers/challengesandchanges.html

Bates, T. (2003). *Higher education and e-learning: Integration or change?* Presentation, University of British Columbia. Retrieved December 14, 2007, from http://bates.cstudies.ubc.ca/

Brodsky, M. W. (2003). E-learning trends today and beyond. *Learning and Training Innovations*. Retrieved December

14, 2007 from http://www.elearningmag.com/ltimagazine/article/articleDetail.jsp?id=56219

Emerging technologies and ground-floor investment opportunities. Special Report: Forecasts for the next 25 years. (2003). The World Future Society: Bethesda, Maryland.

Gallagher, S. (2004). *Online distance education market update: A nascent market begins to mature*. Boston MA: Eduventures. Retrieved December 14, 2007, from www.eduventures.com

Gunawardena, C.N., & McIsaac, M.S. (2003). Distance education. In D.H. Jonassen (Ed.), *Handbook of research on educational communications and technology* (2nd ed., pp. 113-142). Mahwah, NJ: Lawrence Erlbaum.

Harasim, L., Hiltz, S. R., Teles, L., & Turoff, M. (1995). *Learning networks: A field guide to teaching and learning online*. Cambridge MA: MIT Press.

Hickman, C. J. (2003). *Results of survey regarding distance education offerings*. University Continuing Education Association (UCEA) Distance Learning Community of Practice, Research committee report.

Howell, S, Williams, P., & Lindsay, N. (2003). *Thirty-two trends affecting distance education: An informed foundation for strategic planning. Online Journal of Distance Learning Administration, 6*(3), 10-19. State University of West Georgia, Distance Education Center.

Kariya, S. (2003). Online education expands and evolves. *IEEE Spectrum*, *40*(5), 49-51.

Tolmie, A., & Boyle, J. (2000). Factors influencing the success of computer mediated communication (CMC) environments in university teaching: A review and case study. *Computers and Education, 34*(2), 119-140.

U.S. Department of Education. Retrieved December 14, 2007, from http://www.ed.gov/prog_info/StarSchools/index.html

Western Governors University. (2004). Retrieved December 14, 2007 from http://www.wgu.edu/wgu/index.html

Wonacott, M. E. (2002). *Blending face-to-face and distance learning methods in adult and career-technical education* (Practice Application Brief No. 23). Columbus, OH: Clearinghouse on Adult, Career, and Vocational Education. (ERIC Document).

KEY TERMS

Collaborative Learning Online: Technologies that link together people in several locations so that they can interact with one another.

Computer-Based Learning: Refers to the use of computers as a key component of the educational environment. Broadly refers to a structured environment in which computers are used for teaching purposes.

Distance Education: The process of extending learning, or delivering instructional resource-sharing opportunities, to locations away from a classroom, building or site, to another classroom, building or site by using video, audio, computer, multimedia communications, or some combination of these with other traditional delivery methods.

Distance Learning: Courses in home: education for students working at home, with little or no face-to-face contact with teachers and with material provided remotely, for example, by e-mail, television, or correspondence.

E-Learning: Learning using electronic means: the acquisition of knowledge and skill using electronic technologies such as computer- and Internet-based courseware and local and wide area networks.

Higher Education: Education provided by universities, vocational universities and other collegial institutions that award academic degrees.

Online Learning: Learning via educational material that is presented on a computer via an intranet or the Internet.

Web-Based Training: Also referred to as "online courses" or "Web-based instruction," a form of learning in which the training material is contained on Web pages on the Internet or an intranet.

Triangular Strategic Analysis for Hybrid E-Retailers

In Lee
Western Illinois University, USA

INTRODUCTION

For traditional retailers, the success of an e-channel lies largely in formulating and implementing a sound e-channel strategy that leverages their resource base. Numerous evidences show that poorly developed e-channels have added little value to retailers (Huang, 2003; Prencipe & McCarthy, 2002). Some of the poorly developed e-channels have had a negative impact on business performance due to an excessive investment, disappointing sales, and low margin (Nataraj & Lee, 2002). The poor performance arose from focusing on the Internet as a separate channel not affected by the activities in other existing channels (Kannan, 2001). For many traditional retailers, their costly and frequent e-channel reorganizations could have been avoided if they had adequately analyzed the strategic fit between their external environment and e-channel organization.

There has been no universally applicable business strategy for e-channels. For many retailers, the right mix of the traditional channels and e-channel is critical to their business success (Gulati & Garino, 2000). Despite the strategic value of the e-channel, there have been only a paucity of frameworks that help managers initiate and formulate the e-commerce strategy (Allen & Fjermestad, 2001; Lee, 2001). A number of strategic analysis models such as SWOT analysis, five forces model, resource-based view, and critical success factors have been applied to the e-commerce strategy development. These models attempted to formulate a business strategy from different perspectives of a business organization, but have not been fully integrated with each other. Formulating a business strategy based on the analysis and integration of multiple perspectives will result in a more competitive strategy than those of a single perspective.

Based on a number of e-channel case studies and strategic management theories, this short article presents (1) an overview of a triangular strategic analysis and (2) an example application of the triangular strategic analysis with an Office Depot case study. Data on the Office Depot's e-channel strategy and implementation were collected through secondary sources such as trade journals and Office Depot's official publications. The triangular strategic analysis consists of (1) competitive forces analysis, (2) resource base analysis, and (3) critical success factor analysis

BACKGROUND

To capture the ever-increasing B2C population, retailers have experimented with a variety of B2C business models (Gulati & Garino, 2000). Some of the widely used e-commerce models include auction models (e.g., eBay.com), reverse auction models (e.g., Priceline.com), portal models (e.g., Yahoo.com), stand-alone e-retailer models (e.g., Amazon.com), and hybrid e-retailer models (e.g., Walmart.com). While B2B e-commerce applications such as e-procurement systems and the Internet-based supply chain management have brought significant benefits to business organizations, many B2C business models have failed to generate sustainable long-term profits.

In the late 1990s, most stand-alone e-retailers of commodity type products suffered the hardest hits due to low margin, rising customer acquisition cost, and the lack of financial support of investors (Stockport, Kunnath, & Sedick, 2001). Numerous stand-alone e-retailers such as Garden.com, Boo.com, and Petopia.com were consolidated with traditional retailers or liquidated (Kujubu & Martin, 2001). These failures were attributed to the poor business plan, weak complementary resources in distribution network and customer services, lack of brand name recognition, and low entry barriers.

Evidence shows that a misdirected e-channel development leads to costly and frequent revisions of e-commerce strategies. Kmart and Wal-Mart experienced a costly revision of their e-channel strategies. Kmart initially created a spin-off entity, BlueLight.com, in December 1999 as a joint venture between Kmart and Softbank Venture Capital. After Kmart withdrew from a planned initial public offering (IPO) for BlueLight.com in 2000, it acquired all of the interests of BlueLight.com in 2001. Walmart.com is another example of the costly revision of an e-channel strategy. Walmart.com was established in January 2000 as an independent company operating as a joint venture between Wal-Mart and Accel Partners. In 2001, Wal-Mart acquired all the minority interest in Walmart.com in order to establish the tight integration between its e-channel and physical stores.

Since each organization is uniquely positioned in a market with a different set of competitive forces, critical success factors, and capabilities, no single e-channel strategy would be suitable for all organizations. The successful deployment

of an e-channel requires a thorough review and analysis of all major business activities, including business strategies, processes, functions, and vendor/customer relationships. For traditional retailers, poorly deployed e-channels without cross-channel coordination and integration mechanisms in place cannot create competitive advantages. These e-channels may also have a negative impact on other channels by losing customers who value a seamless cross-channel experience. The triangular strategic analysis will provides managers with a unified view of a business organization by combining and presenting multiple organizational perspectives.

TRIANGULAR STRATEGIC ANALYSIS FOR HYBRID E-RETAILERS

The purposes of strategic analysis are to examine the current and future business environments, to identify new business opportunities and threats, and to develop strategies to counter competition and achieve strategic goals. A number of strategic analysis models have been developed with the emphasis on different perspectives of a business strategy development. Table 1 summarizes major strategic theories/models, their purposes, advantages, and disadvantages. Based on the complementarities of these models, we utilize three analysis models in an e-channel development framework. The triangular strategic analysis consists of (1) competitive forces analysis, (2) resource base analysis; and (3) critical success factor analysis. While each of these strategic analysis methods was developed in isolation of the others, they contribute unique yet complementary perspectives on the development of the business strategy. The integration of three analyses implicitly subsumes the essence of the core competence analysis. The balanced scorecard model can be used in parallel with the triangular strategic analysis or in the subsequent planning stage, and later in evaluating the organizational performance.

Table 1. Summary of major strategic management theories/models

Major Theories/Models	Proponents	Characteristics
Five Forces Model of Industry Competition	Porter (1980)	His basic theory was that dynamics of five competitive forces determine the nature of competitiveness in an industry and influence the strategies available to firms in the industry. The competitive forces are: (1) threat of new entry into an industry; (2) intensity of rivalry among existing competitors; (3) pressure from substitute products; (4) bargaining power of buyers; and (5) bargaining power of suppliers.
Resource Based View (RBV)	Wernerfelt (1984)	RBV suggests that firms compete not just in terms of final products, but more fundamentally in terms of the underlying "resources" which make production and product diversification possible. From a resource-based view every firm has a unique set of resources that the firm can leverage to exploit opportunities and counter threats.
Core Competence	Prahalad and Hamel (1990)	Core competencies are the collective learning in the organization that gives the company a unique advantage over its competitors. Core competence can manifest itself in many ways. Core competence is communication, involvement, and a deep commitment to working across organizational boundaries. It is the skills of individuals who can blend their expertise with that of others in new and interesting ways.
Balanced Scorecard (BSC)	Kaplan and Norton (1996)	The Balanced Scorecard is a method for turning a company's vision and strategy into a coherent set of performance measures distributed among four perspectives: Financial, Customer, Internal Business Processes, and Learning and Growth. The framework provides a balance between short- and long-term objectives, financial and nonfinancial measures, and external and internal performance indicators.
Critical Success Factor (CSF) Analysis	Rockart (1979)	CSF analysis is a method developed to guide businesses in creating and measuring success. CSFs are key areas where satisfactory performance is required for the organization to achieve its goals. Rockart provided the following as an example of the CSFs: new product development, good distribution, and effective advertising - factors that remain relevant today for many firms.

Table 2. Impacts of e-commerce on competitive forces' threats and firm's opportunities from retailers' perspective

Competitive Forces	Threats	Opportunities
Suppliers	Disintermediation Sell-Side Forward Auction	E-Procurement Group Purchasing Reverse Auction Internet-based EDI
Traditional Competitors	E-Channel E-Services Intranet Extranet E-Procurement	E-Channel Strategic Alliances Third-Party E-Marketplaces
New Market Entrants	E-Commerce Strategic Alliances	E-Channel Strategic Alliances, Merger/Acquisition
Customers (Corporate)	E-Procurement Third-Party Order Aggregation Price Comparison	E-Channel One-to-One Marketing Internet-based EDI Extranet
Customers (Consumers)	Price Comparison Third-Party Order Aggregation	E-Channel Web Personalization E-Services User Profiling
Substitute Products/Services	Digitized Products/Services On-Demand Delivery Services Portals	E-Channel Digitized Products/Services On-Demand Delivery Services New Services

Competitive Forces Analysis

Porter's Five Forces Model (1980) has been widely used in analyzing competitive forces that would shift an organization's strategic position in the industry. The components of the model include: (1) the traditional competitors, (2) the bargaining power of the customers, (3) the bargaining power of the suppliers, (4) the potential threat of new entrants, and (5) the threat of substitute products/services. The competitive forces analysis presents a picture of where the company's current strategy stands against competitive forces and a road map of where it should go for the success in the competitive environment. The emergence of e-commerce has changed the ways all competitive forces perform businesses in the industry. E-commerce provides customers with opportunities to comparison-shop and thereby raises the bargaining power of customers. E-commerce also provides suppliers with opportunities to disintermediate retailers. Traditional competitors establish e-commerce Web sites to enter into the e-marketplaces.

The competitive forces analysis provides managers with vital information on the opportunities and threats arising in the e-commerce market. Table 2 summarizes the opportunities and threats introduced by the e-commerce from the retailers' perspective. While the competitive forces analysis is effective in understanding external environment, some researchers have pointed out its limitations (Luffman, Lea, Sanderson, & Kenny, 1996). The analysis has not explicitly addressed what the critical factors are and what resources are needed for a successful execution of a business strategy. The following resource base analysis utilizes the results obtained from the competitive forces analysis and determines what resources are needed and how they are used in the execution of an e-channel strategy.

Resource Base Analysis

Many researchers have indicated that the resource-based view of a firm is the most important theory of sustainable competitive advantage (Conner, 1991). The fundamental logic of the resource-based view is that the desirable outcome of a business strategy is a sustainable competitive advantage. Since not all resources are equal in creating sustainable competitive advantage, many researchers focused on identifying advantage-creating resources. Barney (1986, 1991) suggested that the advantage-creating resources must be firm-specific, rare, and difficult to imitate. The resource-based view of a firm was applied to understand how the superior IT resources of organizations render the cost and value of IT innovations different from competitors (Bharadwaj,

2002). In a changing environment, firms must continuously invent and upgrade their resources and capabilities if they are to maintain a competitive advantage and growth. The sequential development of resources and capabilities can make a firm's advantage inimitable (Barney, 1991; Lado, Boyd, & Hanlon, 1997).

Each retailer has a different set of firm-specific resources to utilize, and an addition of an e-channel can be viewed as a unique resource utilization and development process in achieving sustainable growth. The mapping between the competitive forces and the resource base provides answers to two important questions: (1) what kinds of existing resources a retail organization can leverage and (2) what resources it needs to differentiate itself from the competitors. While strong resources are leveraged to survive, weak or non-existent resources need to be critically examined for the future resource development and sustainable competitive advantage. Traditional retailers' resources include physical stores, distribution centers, patent, trademark, brand, reputation, enterprise-wide database, integrated

Table 3. E-commerce analysis matrix of competitive forces and resource base: Office Depot's case

Competitive Forces	Suppliers		Traditional Competitors		New Market Entrants		Customers		Substitute Products/ Services	
Resource Base	Threats	Opportunities	Threats	Opportunities	Threats	Opportunities	Threats	Opportunities	Threats	Opportunities
Physical Stores							●			
Distribution Centers							●			
Database				●				●		
Integrated Information Systems	●							●		
Skills and Knowledge of Employees			●					●	●	
Organizational Culture and Trust								●		

Table 4. E-commerce analysis matrix of critical success factors and resource base: Office Depot's case

Competitive Forces	Suppliers		Traditional Competitors		New Market Entrants		Customers		Substitute Products/ Services	
Resource Base	Threats	Opportunities	Threats	Opportunities	Threats	Opportunities	Threats	Opportunities	Threats	Opportunities
Physical Stores							●			
Distribution Centers							●			
Database				●				●		
Integrated Information Systems	●							●		
Skills and Knowledge of Employees			●					●	●	
Organizational Culture and Trust								●		

information systems, skills and knowledge of employees, and organizational culture and trust. Some of these resources are more difficult to imitate and more valuable as a source of competitive advantage than others. In general, intangible resources would be more difficult to transfer and imitate than tangible resources.

Table 3 shows the e-commerce strategy matrix of the competitive forces and the resource bases of Office Depot. The rows list two types of resources: strong resources (sustaining resources) and weak/non-existent resources (developmental resources). Certain weights can be assigned to the competitive forces based on their importance to prioritize the resource allocation and development activities. The analysis of Office Depot suggests that the existing integrated information systems can be leveraged to counter the threat of the suppliers' e-commerce and to benefit from the growth of the customers' online purchases.

The matrix needs filling of e-commerce strategies and projects in relation to resources, threats and opportunities. Threats posed by new market entrants may force a retailer to choose the e-channel introduction as a strategic option. If a retailer has a strong resource base in support of an e-channel, then an immediate introduction of an e-channel may be feasible with little developmental resources. Overall, the e-commerce strategy matrix of the competitive forces and the resource base provides a conceptually grounded framework for assessing sustainable and developmental resources, and enables these resources to be examined in terms of the opportunities and threats for establishing sustainable competitive advantages.

Critical Success Factors Analysis

Once the previous analyses lead to an e-commerce strategy, critical success factors (CSFs) analysis decides what the most important determinants are in achieving strategic goals (Rockart, 1979). The successful execution of CSFs requires resources. Identifying a match between the firm's resources and the critical success factors in the industry is a demanding task, and the success of the match is a function of the accuracy of managerial expectations about the value of the strategy (Barney, 1986).

The e-commerce success factors for retailers include efficient business process, integrated distribution systems, multi-channel coordination and integration, customers' trust, strong brand recognition, superior customer service operations, financial stability, effective online technologies, and strategic alliances. Table 4 shows a matrix that develops matches between the Office Depot's resources and the critical success factors. For example, the business process redesign requires the integrated information systems, skills and knowledge of employees, and organizational culture and trust as critical resource bases.

The matrix helps managers determine what existing resources can be leveraged or what new resources are needed to achieve the CSFs and ultimately the business strategies. To successfully achieve e-channel strategic goals and sustain competitive advantages, retailers need to monitor the performance of these CSFs with a measurable objectives and metrics. Once CSFs and metrics are defined, a detailed e-channel plan should be developed to achieve these critical success factors and business strategies.

FUTURE TRENDS AND CONCLUSION

The rapid penetration of the World Wide Web and the explosion of e-commerce startups have changed the dynamics of the competitive forces across all industries. Due to the explosion in e-commerce competition, the e-channel has become a critical factor in the strategy development by the hybrid e-retailers. While managing the e-channel is one of the most important tasks for marketing managers, many managers are still unclear about e-channel strategies and lack core e-channel knowledge needed to analyze business environments, to develop strategies, and to evaluate alternative e-channel solutions. Empirical evidence shows that experimenting with different types of e-channels is very costly. Some retailers such as Kmart, Wal-Mart, CVS, and Staples experimented with spin-offs. These retailers later struggled to retrofit the spin-offs into their parent companies.

A number of strategic analysis models such as SWOT analysis, five forces model, resource-based view, and critical success factors have been applied to e-commerce strategy development. However, there has been little effort to integrate these models for e-channel strategy development. These models attempted to formulate a business strategy from different perspectives of a business organization, but have not been fully integrated with each other. The analysis and integration of multiple dimensions will result in a more comprehensive strategy than those of a single dimension. To develop a comprehensive e-channel strategy, the triangular strategic analysis attempted to integrate three well-known analysis models: competitive forces analysis, resource base analysis, and critical success factor analysis.

The triangular strategic analysis provides managers with a unified view of a business organization by combining and presenting multiple perspectives of an organization. The unified view will enhance managers' ability to effectively identify their present strategic position, internal resources, and critical success factors, and decide upon the most appropriate e-channel strategy. As important, the unified view will allow managers to utilize the emerging opportunities of e-commerce and to prevent threats that would be posed by carelessly developed e-channel strategies.

The triangular strategic analysis was applied to analyze Office Depot's e-channel development strategy. The evalua-

tion of Office Depot's e-channel development strategy suggests a number of its success factors that may be useful for other retailers' e-channel development: (1) Office Depot was an early adopter of an e-commerce technology, and continued to explore different e-commerce business models over time; (2) its e-channel was not only another distribution channel but also a service channel; (3) it also leveraged its own e-channel expertise and resources in office supplies in expanding to other markets; and (4) it treated an e-channel not as an independent entity but as an internal business unit in an integrated business organization; (5) its senior management supported an e-commerce project from a strategic point of view; and (6) it pursued a resource-based e-channel IT development.

REFERENCES

Allen, E., & Fjermestad, J. (2001). E-commerce marketing strategies: An integrated framework and case analysis. *Logistics Information Management, 14*(1/2), 14-23.

Barney, J.B. (1986). Strategic factor markets: Expectations, luck, and business strategy. *Management Science, 32*(10), 1231-1241.

Barney, J.B. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99-120.

Bharadwaj, A.S. (2000). A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly, 24*(1), 169-196.

Conner, K.R. (1991). A historical comparison of resource-based theory and five schools of thought within industrial organisation economics: Do we have a new theory of the firm. *Journal of Management, 17*(1), 121-154.

Gulati, R., & Garino, J. (2000). Get the right mix of bricks & clicks. *Harvard Business Review, 78*(3), 107-114.

Huang, X. (2003). Research on Australian e-tailers: Strategic issues, success factors, and challenges. *International Journal of Services Technology & Management, 4*(4-6), 563-573.

Kannan, P.K. (2001). Introduction to the special issue: Marketing in the e-channel. *International Journal of Electronic Commerce, 5*(3), 3-6.

Kaplan, R.S., & Norton, D. (1996). Using the balanced scorecard as a strategic management system. *Harvard Business Review, 74*(1), 75-85.

Kujubu, L., & Martin, A. (2001), Opportunity in failure. *InfoWorld, 23*(15), 36-37.

Lado, A.A., Boyd, N.G., & Hanlon S.C. (1997). Competition, cooperation, and the search for economic rents: A syncretic model. *Academy of Management Review, 22*(1), 110-141.

Lee, C-S. (2001). An analytical framework for evaluating e-commerce business models and strategies. *Internet Research, 11*(4), 349-359.

Luffman, G., Lea, E., Sanderson, S., & Kenny, B. (1996). *Strategic management*. Blackwell Publishers Inc, Oxford.

Nataraj, S., & Lee, J. (2002). Dot-com companies: Are they all hype? *S.A.M. Advanced Management Journal, 67*(3), 10-14.

Porter, M.E. (1980). *Competitive strategy*. New York: Free Press.

Prahalad, C., & Hamel, G. (1990). The core competence of the corporation. *Harvard Business Review, 68*(3), 79-91.

Prencipe, L.W., & McCarthy, J. (2002). Battle of the shopping carts. *InfoWorld, 24*(40), 46.

Rockart, J.F. (1979). Chief executives define their own data needs. *Harvard Business Review, 57*(2), 81-93.

Stockport, G.J., Kunnath, G., & Sedick, R. (2001). Boo.com - The path to failure. *Journal of Interactive Marketing, 15*(4), 56-70.

Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal, 5*(2), 170-180.

KEY TERMS

Channel Conflict: Situation in which an e-channel creates a conflict with existing channels because of real or perceived damage from inter-channel competition.

Complementarities: Products or services that provide more value together than individually. For example, hybrid e-retailers can leverage complementarities by providing offline services to online shoppers.

E-Channel: An online marketing channel where companies and customers conduct business, no matter where they are. Since the e-commerce revolution, many brick-and-mortar businesses have expanded their marketing channel to include e-channel.

Hybrid E-Retailer: A click and mortar company which conducts retailing through e-channel as well as physical stores and other distribution channels. Compared to its pure e-commerce competitors, a hybrid e-retailer can leverage existing physical stores, brand recognition, distribution network, existing customer base, and so forth.

On-Demand Delivery Services: Express delivery of products made with highly efficient transportation systems after an online order is received.

Order Aggregation: A group purchase designed to achieve a volume discount by aggregating orders placed by individual buyers.

Reverse Auction: A fixed-duration auction hosted by a single buyer in which multiple sellers compete for business.

Sell-Side Forward Auction: An auction where a seller announces the items for quick sale and buyers bid on them.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2871-2877, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Triune Continuum Paradigm

Andrey Naumenko
Triune Continuum Enterprise, Switzerland

INTRODUCTION

This article reviews the Triune Continuum Paradigm—a logically rigorous theoretical base for organization of conceptual frameworks that are used for system modeling in different contexts (e.g., in software development, in enterprise architecture, in the architecture of financial services, in jurisprudence, etc.). This paradigm is an important contribution to the system modeling domain, because currently none of the prevailing system modeling frameworks has a satisfactory formal theoretical foundation.

The absence of a theoretical foundation for modeling frameworks leads to the practical application experiences where modelers are constrained to be guided by chance and not by a founded reason. This often leads to the inadequate choices of modeling frameworks, that is, to the situations where a chosen modeling framework is not designed to deal with the targeted modeling problems. Possible consequences of such choices include incorrect (e.g., inadequate with regard to the requirements) information systems specifications, contradictory data architectures, incomplete service specifications, and so forth—all of these being the decisive contributions to failures of many projects. The paradigm, which we review in this article, fixes this problem providing missing theoretical foundations for frameworks positioned in the domain of general system modeling.

Many of the existing system modeling frameworks appeared as an integration of the best modeling practices. The reviewed paradigm does not repudiate the practical experience that was gathered by these different frameworks, but fixes its inconsistencies and complements it supporting with logically rigorous theoretical foundations. Therefore the paradigm brings a significant constructive potential to the evolution of modern system modeling frameworks. This potential could be realized if people responsible for the design of modeling frameworks and tools would heed the proposed paradigm.

BACKGROUND

The Cambridge Dictionary of Philosophy (Audi, 1999, p. 641) provides the following definition of the term "paradigm": "Paradigm, as used by Thomas Kuhn (The Structure of Scientific Revolutions, 1962), a set of scientific and metaphysical beliefs that make up a theoretical framework within which scientific theories can be tested, evaluated and if necessary revised."

In practice, a paradigm is usually defined for a collection of sciences. In this context a paradigm introduces and justifies a set of basic assumptions and principles on which any of sciences from the collection can rely as on their foundations. Then, starting from the principles provided by a paradigm, different sciences build their specific frameworks of knowledge. And if some sciences share the same paradigm, then they can bind and synchronize their specific frameworks of knowledge. By doing so they can mutually enrich each other with the knowledge obtained from the different (but consistent with regard to the basic principles) points of view.

The Triune Continuum Paradigm (Naumenko, 2002) is a paradigm for general system modeling. Thus the Triune Continuum Paradigm serves the sciences that have diverse interests in system modeling. As any paradigm, it introduces and justifies a set of principles that provide the sciences with the necessary starting points for building their diverse conceptual frameworks of scientific knowledge, in our case the principles that are necessary for building modeling frameworks.

THREE PRINCIPLES OF THE TRIUNE CONTINUUM PARADIGM

The Triune Continuum Paradigm is composed of three principles.

The first principle is the result of application of Tarski's Theory of Truth (Tarski, 1956) for the case of general system modeling. This application allows defining coherent semantics for the concepts of a modeling framework. This is done by constructing formal descriptions for the relations between the subjects that are interesting to be modeled on one side, and the concepts that have to represent these subjects in the models on the other side. This principle is necessary to assure the *coherency* and *unambiguity* within modeling interpretations performed using a single system modeling framework.

An application of the first principle provided by the Triune Continuum Paradigm results in a system modeling framework that features modeling terms with a coherently defined semantics in the form of Tarski's declarative semantics. The justifications of importance of this principle for the information systems modeling were presented and analyzed

in details (Naumenko, Wegmann, & Atkinson, 2003). In particular, it was demonstrated that Tarski's declarative semantics are:

- formally *sufficient* for the definition of the application scope of a modeling language;
- formally *sufficient* for unambiguity in coherency of interpretations within modeling representations; and
- formally *necessary and sufficient* for unambiguity in adequateness of modeling representations.

The second principle of the Triune Continuum Paradigm is the result of application of Russell's theory of types (Russell, 1908) for the case of general system modeling. This application defines the way to categorize concepts of a modeling framework so that in applications of this framework the concepts make up *internally consistent* structures of propositions. Thus this principle is necessary to assure the consistency of descriptions and specifications, which are constructed with the aid of the modeling frameworks.

The importance of this principle is justified by the fact that Russell's theory of types was formulated to resolve Russell's paradox, "the most famous of the logical or set-theoretical paradoxes" (Irvine, 2003). Thus with an application of the second principle of the Triune Continuum Paradigm, the resulting modeling framework in its own applications will produce internally consistent system specifications (i.e., system specifications that are devoid of self-contradictions).

The name of Triune Continuum Paradigm originates from the third theory that was employed for the paradigm definition, from the Theory of Triune Continuum. This theory was defined by Naumenko (2002). This theory allows for the introduction of the abstract ontologies that are formally *necessary and sufficient* to cover the modeling scope of different modeling contexts on the most abstract level.

In particular, the Theory of Triune Continuum was applied in the context of general system modeling (Naumenko, 2002), and this application contributed to the definition of the Triune Continuum Paradigm. The application is the third paradigm principle that allowed introducing and justifying a minimal set of modeling concepts that are necessary and sufficient to cover the representation scope of the general system modeling domain on the most abstract level. This principle is necessary for different system modeling frameworks to justify the existence of their basic modeling concepts.

The Theory of Triune Continuum introduces three continuums that represent in models the scope of general system modeling. The first two continuums are:

- **Spatiotemporal Continuum:** Where subjective space-time metrics are defined to be used in the subjective representations.
- **Constitution Continuum:** Where subjective constitutional metrics are defined to be used in the subjective representations, for example, objects defined in relation with their environments.

These two continuums are introduced in relation with each other as complements within the universal general system modeling scope. In other words, everything in the scope that is not space-time is constitution; and everything in the scope that is not constitution is space-time.

The third continuum is:

- **Information Continuum:** Which emerges from the mutual relations of the first two continuums and contains the corresponding information about these relations, for example, information about objects and their environments being related to the spatiotemporal intervals or to the points in space-time).

Thus the three continuums are *triune*: none of them exist without the others; either the three exist altogether, or they do not exist at all. Indeed, as soon as the first (spatiotemporal) continuum is introduced, everything in the universal scope that does not belong to the first continuum immediately shapes the second (constitution) continuum; and the third (information) continuum immediately emerges as the information about the mutual relations of the first two continuums (e.g., as spatiotemporal information about the constitution).

The third principle of Triune Continuum Paradigm is important for various system modeling frameworks, which are used in diversified domains of human activity (e.g., the frameworks used to analyze, design, and develop coherent structures providing useful functionalities in domains spread from jurisprudence and health care to software engineering and machine-building industries). Using the notion of Triune Continuum it is possible to introduce and justify minimal sets of modeling concepts that are necessary and sufficient for those diversified frameworks to cover their respective representation scopes.

APPLICATIONS OF THE TRIUNE CONTINUUM PARADIGM

The Triune Continuum Paradigm can be applied in practice either to improve an existing system modeling framework or to design a new system modeling framework for a given purpose. Let us mention here three of the existing applications of the paradigm:

- case of the Unified Modeling Language (UML);
- case of the reference model of open distributed processing (RM-ODP); and
- case of the systemic enterprise architecture methodology (SEAM).

The first two of the three cases illustrate the paradigm applications targeting improvements of the existing system modeling frameworks. The third case illustrates the paradigm application contributing to the design of a new system modeling framework.

Case 1: Triune Continuum Paradigm Application for UML

"The Unified Modeling Language (UML) is a language for specifying, visualizing, constructing, and documenting the artifacts of software systems, as well as for business modeling and other non-software systems" (OMG, 2003, section 1.1). UML is a proposition of the Object Management Group (OMG) that emerged from the integration of different industrial practical experiences and became an influential phenomenon in the system modeling. As a matter of fact, due to the multiple efforts of different interested parties, UML has gained a relative domination over the other modeling techniques in the current industrial practices. This is why it was interesting to apply the Triune Continuum Paradigm for the case of UML conceptual framework. Results of this application were presented to the UML research community (Naumenko & Wegmann, 2002). With the aid of the Triune Continuum Paradigm it was shown that the metamodel of UML features a number of undesirable properties, in particular:

- absence of an explicit structural organization defined for the UML metamodel;
- absence of Tarski's declarative semantics in the UML metamodel; and
- absence of theoretical justifications for the UML metamodel to represent the modeling scope that is targeted by UML.

The paradigm-based solutions were presented for each of the three identified problems (Naumenko & Wegmann, 2002) providing designers of UML with the benefits of the paradigm's logical rigor, of its formal presentation, and of its solid theoretical foundations.

Case 2: Triune Continuum Paradigm Application for RM-ODP

The RM-ODP is an ISO and ITU standard for system modeling, designed to model ODP-systems (ISO & ITU, 1998). The result of Triune Continuum Paradigm application for the RM-ODP case is especially interesting because it allowed accomplishing a single consistent formalization of the RM-ODP conceptual framework, providing the denotational semantics for the basic modeling and specification concepts of RM-ODP. Such formalization was officially declared as a goal of the ISO and ITU activities in the scope of RM-ODP standardization (ISO & ITU, 1998). But this goal was not achieved by the standard; and so far the paradigm-based formalization remains the only solution achieving the defined objective.

The formalization was expressed in a computer interpretable form using Alloy formal description technique (Jackson, 2002). Alloy was chosen because of the public availability of the corresponding software tool, "Alloy Constraint Analyzer," that allows simulating the instances of conceptual structures formalized with Alloy and representing these instances in a graphical form. However, due to the nature of denotational semantics (Naumenko et al., 2003), any choice of the formal description technique does not change semantic interrelations within the formalization. So, another formal description technique could also be used to express the paradigm-based formalization of the RM-ODP conceptual framework in a computer interpretable form.

The paradigm-based formalization of RM-ODP presents a concrete example of formal ontology for general system modeling. Thanks to the Triune Continuum Paradigm, the metamodel that is realized by the formal ontology is internally consistent, introduces logical coherency of interpretation of a subject of modeling, defines formal semantics for the modeling concepts, and its models are verifiable with the aid of computer tools. These results were presented to the RM-ODP research community (Naumenko & Wegmann, 2001, 2005; Naumenko, Wegmann, Genilloud, & Frank, 2001), and they attracted interest of the ISO/ITU committee that is responsible for the RM-ODP standardization. This provides the Triune Continuum Paradigm with a chance to influence future evolution of the ISO/ITU standard.

Case 3: Triune Continuum Paradigm Application for SEAM

The systemic enterprise architecture methodology (SEAM) is a methodology proposed by LAMS-EPFL (Wegmann, 2003) for system modeling in the domain of enterprise architecture, which is the domain that considers integration of IT systems and business systems in the context of an enterprise.

Applying the Triune Continuum Paradigm, a logically rigorous framework of concepts covering the representation scope of SEAM was designed and implemented as a specialization of the RM-ODP standard conceptual framework (ISO & ITU, 1998). Thus in this case the paradigm application provided a formal ontology for SEAM. The corresponding research results were reported to the Enterprise Architecture community (Wegmann & Naumenko, 2001) and provided the necessary basis for ongoing evolution of SEAM.

FUTURE TRENDS

The Triune Continuum Paradigm provides a set of theoretical foundations for different frameworks of knowledge belonging to the general system modeling domain. In most of the cases currently existing industrial and academic frameworks for system modeling do not feature such theoretical foundations, because these frameworks are developed using the so-called "best practices" approach (when results of different practical experiences of system modeling within a given domain are integrated to build a modeling framework for the domain). And if the "best practices" approach is not accompanied by theoretical foundations, then it is impossible to justify a number of important properties for the resulting system modeling frameworks (e.g., to guarantee the necessity and sufficiency of a framework for its domain representation, to assure internal consistency within different pieces of practical experience integrated in a single framework, etc.).

So, the Triune Continuum Paradigm provides an indispensable contribution to the general system modeling. And the future trends should assure practical realization of the significant constructive potential that the paradigm features for those numerous system modeling frameworks that currently do not have satisfactory theoretical foundations. This potential will be realized through the paradigm applications to these concrete system modeling frameworks. Some of the examples of such applications were presented in this article.

CONCLUSION

The Triune Continuum Paradigm provides system modelers (in particular, IS modelers) with a set of principles that are essential to build adequate system modeling frameworks. These principles are based on the solid theoretical foundations discovered in the last century: Tarski's Theory of Truth was presented in 1935, while Russell's Theory of Types was formulated in 1908. The authors of these two theories, Alfred Tarski and Bertrand Russell, are recognized to be among the greatest logicians throughout the history of humankind. Thus the Triune Continuum Paradigm, through its applications used in the computer-aided environment, promotes the use of fundamental logical theories to the practices of regular modelers, information system designers, and architects. The paradigm-based theoretically founded approaches to the information systems development make a constructive difference in the IS development projects where the usual "best practices methodologies" do not perform well enough due to a number of reasons (e.g., lack of flexibility, lack of representation possibilities, lack of internal consistency, etc.).

REFERENCES

Audi, R. (Ed.). (1999). *The Cambridge dictionary of philosophy* (2nd ed.). Cambridge, UK: Cambridge University Press.

Irvine, A. D. (2003). Russell's Paradox. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (Summer 2003 ed.).* Retrieved October 31, 2006, from http://plato.stanford.edu/

ISO & ITU. (1998). Open distributed processing—Reference model. *ISO/IEC 10746-1, 2, 3, 4 | ITU-T Recommendation X.901, X.902, X.903, X.904.* 1995-98.

Jackson, D. (2002). Alloy: A lightweight object modeling notation. *ACM Transactions on Software Engineering and Methodology, 11*(2), 256-290.

Kuhn, T. S. (1962). *The structure of scientific revolutions* (3rd ed.). Chicago: University of Chicago Press.

Naumenko, A. (2002). *Triune continuum paradigm: A paradigm for general system modeling and its applications for UML and RM-ODP* (-No. 2581). Lausanne, Switzerland: Swiss Federal Institute of Technology—Lausanne (EPFL).

Naumenko, A., & Wegmann, A. (2001). *A formal foundation of the RM-ODP conceptual framework* (Tech. Rep. No. DSC/2001/040). Lausanne: Swiss Federal Institute of Technology—Lausanne (EPFL).

Naumenko, A., & Wegmann, A. (2002). A metamodel for the Unified Modeling Language. In J-M. Jézéquel, H. Hussmann, & S. Cook (Eds.), *LNCS 2460: Proceedings of UML 2002* (pp. 2-17). Dresden, Germany: Springer.

Naumenko, A., & Wegmann, A. (2005). Formalization of the RM-ODP foundations based on the Triune Continuum Paradigm. *Computer Standards & Interfaces*. Elsevier B.V. Retrieved October 31, 2006, from http://dx.doi.org/10.1016/j.csi.2005.10.001

Naumenko, A., Wegmann, A., & Atkinson, C. (2003). *The role of Tarski's declarative semantics in the design of modeling languages* (Tech. Rep. No. IC/2003/43). Lausanne: Swiss Federal Institute of Technology—Lausanne (EPFL).

Naumenko, A., Wegmann, A., Genilloud, G., & Frank, W. F. (2001). Proposal for a formal foundation of RM-ODP concepts. In J. A. M. Cordeiro & H. Kilov (Ed.), *Proceedings of ICEIS 2001, Workshop on Open Distributed Processing—WOODPECKER 2001* (pp. 81-97). Setúbal, Portugal: ICEIS Press.

OMG. (2003). *Unified Modeling Language Specification*. Version 1.5. Retrieved October 31, 2006, from http://www.omg.org/uml

Russell, B. (1908). Mathematical logic as based on the theory of types. *American Journal of Mathematics, 30*, 222-262.

Tarski, A. (1956). *Logic, semantics, meta-mathematics*. Oxford, UK: Oxford University Press.

Wegmann, A. (2003). On the Systemic Enterprise Architecture Methodology (SEAM). In *Proceedings of ICEIS 2003* (Vol. 3, pp. 483-490). Anger, France: ICEIS Press.

Wegmann, A., & Naumenko, A. (2001). Conceptual modeling of complex systems using an RM-ODP based ontology. In *Proceedings of the 5th IEEE Conference—EDOC 2001* (pp. 200-211). Seattle, WA: IEEE Computer Society.

KEY TERMS

Reference Model of Open Distributed Processing (RM-ODP): An ISO and ITU standard for system modeling designed to model open distributed systems.

Russell's Theory of Types: A theory proposed by British logician Bertrand Russell to resolve Russell's paradox, which appears when the set of all sets that are not members of themselves is considered in naive set theory. The paradox is that such a set appears to be a member of itself if and only if it is not a member of itself.

Systemic Enterprise Architecture Methodology (SEAM): A methodology proposed by LAMS-EPFL for system modeling in the domain of enterprise architecture (the domain that considers integration of IT systems and business systems in the context of an enterprise).

Tarski's Theory of Truth: A theory proposed by Polish logician Alfred Tarski. The theory defines the criteria for a formal definition of a true sentence; the theory allows deriving the notion of Tarski's declarative semantics for a modeling language where the modeling language terms are put in the unambiguous correspondence with the subjects of modeling interest that they represent in applications of the language.

Theory of Triune Continuum: A modeling theory proposed by Andrey Naumenko. The theory introduces three continuums (spatiotemporal, constitution, and information continuums) to justify a minimal set of modeling concepts that are formally necessary and sufficient to cover the representation scope of different modeling contexts on the most abstract level.

Triune Continuum Paradigm: A paradigm for general system modeling. The paradigm introduces and justifies a set of principles that provide designers and architects of system modeling methodologies with the necessary theoretical support for building their modeling frameworks. The principles are derived from the Tarski's Theory of Truth, from the Russell's theory of types, and from the Theory of Triune Continuum.

Unified Modeling Language (UML): Proposed by the Object Management Group (OMG) for system modeling in the domains of software systems, of business systems, and others.

Trust in B2C E-Commerce Interface

Ye Diana Wang
University of Maryland, Baltimore County, USA

THE NATURE OF TRUST

Electronic commerce (e-commerce) is changing the way people make business transactions, especially in the business-to-consumer (B2C) area, and it is becoming a significant global economic force. Since Internet technologies and infrastructures to support e-commerce are now in place, attention is turning to psychological factors that affect e-commerce acceptance by online users and their perceptions of online transactions. One such factor is trust, seen to be key to the proliferation of e-commerce.

Trust has existed as long as the history of humans and human social interactions, and it has been studied long before the emergence of the Internet or e-commerce. With respect to consumer behavior, studies have mainly focused on trust and trust relationships in the off-line world and have emerged from numerous disciplinary fields since the 1950s (Corritore, Kracher, & Wiedenbeck, 2001). These disciplines, including philosophy, sociology, psychology, management, marketing, ergonomics, human-computer interaction (HCI), and industrial psychology (Corritore, Kracher, & Wiedenbeck, 2003), have together contributed an extensive body of literature on trust in general, and therefore, they are important grounding points for the examination of trust in the online world. However, "trust is an extraordinarily rich concept, covering a variety of relationships, conjoining a variety of objects," as Nissenbaum (2001, p. 104) has pointed out. Due to the complex and abstract nature of trust, each discipline has its own understanding of the concept and different ways to conceptualize it according to the features of a particular context.

Even with the diverse trust research, researchers from every discipline do acknowledge the value of trust and generally observe and accept four characteristics of trust. First, there must exist two specific parties in any trusting relationship: a trusting party (trustor) and a party to be trusted (trustee). The two parties, comprised of persons, organizations, and/or products, constantly evaluate each other's behaviors. Second, trust involves vulnerability. Trust is only needed, and actually flourishes, in an environment that is uncertain and risky. Third, trust decreases complexity in a complex world and leads people to take actions, mostly risk-taking behaviors. "Without trust people would be confronted with the incomprehensible complexity of considering every possible eventuality before deciding what to do" (Grabner-Krauter & Kaluscha, 2003, p. 787). And fourth, trust is a subjective matter. It is directly related to and affected by individual differences and situational factors.

The previously mentioned characteristics of trust make it especially needed in e-commerce because people perceive economic transactions in a virtual environment as posing a higher degree of uncertainty than in traditional settings. Most e-commerce transactions are not only separated in time and space, but are also conducted via limited communication channels and impersonal interfaces, making trust a crucial facilitator for people to overcome fear, risks, and complexity. Therefore, online consumers need trust as a mental shortcut to reduce the complexity of conducting business transactions with online vendors (Luhmann, 1989). Such trust occurring in cyberspace is commonly termed "online trust," and we limit the scope to the online trust that is pertinent to B2C e-commerce, namely, the trust that occurs for an individual Internet user toward a specific e-commerce Web site or the online vendor that the Web site represents. Derived from the general definition for trust (Rousseau, Sitkin, Burt, & Camerer, 1998), online trust can be defined as follows: *an Internet user's psychological state of risk acceptance based upon the positive expectations of the intentions or behaviors of an online vendor*.

There are almost certainly many potential sources of influence that promote or hinder online trust. However, the current article focuses on the HCI or interface design perspective in inducing online trust, that is, to use what consumers can see on an e-commerce interface to affect their feelings of trust toward the online merchant that the e-commerce interface represents.

ONLINE TRUST IN THE HCI LITERATURE

Online trust is a relatively new research topic and has recently drawn great interest from researchers in HCI and human factors. There are several main themes that the majority of the existing studies can be divided into. First, some studies attempt to understand the online consumer's mind by investigating the underlying elements, antecedents, or determinants that are pertinent to the formation of online trust. For example, Gefen (2002) examined trust from a multi-dimensional perspective. According to the researcher, the specific beliefs of integrity, ability, and benevolence were seen as antecedents to overall trust. Other research-

ers, such as Corritore et al. (2003), also proposed that the consumer could perceive trust before, during, or after the online transaction, and they further concluded that online trust was characterized by its stage of development.

The second stream of studies focuses on conceptualizing trust into theoretical models or frameworks and dividing trust elements into various dimensions. For example, the Model of Trust for Electronic Commerce (MoTEC), is proposed by Egger (2001). The model consists of four components: the pre-interactional filters taking place before any online interaction, the interface properties of the Web site, the information content of the Web site, and relationship management. The Cheskin/Sapient Report (1999) focused on Web site interface cues and presented a model of six building blocks of online trust. These six building blocks were seals of approval, brand, navigation, fulfillment, presentation, and technology. The building blocks could be further divided into a total of 28 components to establish perceived trustworthiness. Such studies provide a theoretical account for exploring and enhancing trust in an online context and often take the effects of customer relationship management into consideration.

The third stream of studies aims to validate those conceptual frameworks or trust scales, often by analyzing data acquired directly from the consumers (e.g., Ba & Pavlov, 2002; Bhattacherjee, 2002). The main objective of these studies is to theoretically derive and empirically validate a scale that can be used to measure either individual online trust or the trustworthiness of an e-commerce Web site. In developing such an instrument, as for developing any other kind of scale, the researchers need to stress establishing its reliability, content validity, and construct validity. Factor analysis, structural equation modeling, and multiple linear regression analysis are some of the most commonly used statistical analysis methods in those efforts.

And finally, the rest of the studies suggest Web design guidelines that are intended to enhance consumer online experience and induce the feeling of trust from the consumers (Karvonen & Parkkinen, 2001; Kim & Moon, 1998; Neilsen, 2000). In other words, the main goal for the researchers of these studies is to explore Web interface design implications to maximize consumer trust or, more precisely, trust perception. A representative study of this kind is the Nielsen Norman Group Report (2000), in which explicit trust-inducing guidelines — including graphic design, surface cue, and Web usability features — are provided based on a large number of user testing observations carried out by experts.

These preceding studies provide important insights into trust in an online context. However, the research field of online trust is still far from maturity and expected to be significantly substantiated and enhanced. For example, the terms *element*, *antecedent*, *dimension*, *determinant*, and *principle* are sometimes used interchangeably due to the lack of agreement on a clear definition for each term among researchers in the field. Nevertheless, this is the current body of work from which any potential implementation is to be derived.

BUILD ONLINE TRUST BY WEB DESIGN

To initiate and build a consumer's online trust is inevitably a challenging task. Due to the nature of the Internet, people nowadays browse different e-commerce Web sites as fast as they switch TV channels. Consequently, to succeed in e-commerce, online vendors must be able to convey their trustworthiness to first-time visitors and effectively and efficiently build trust in the eyes of consumers. This requires online vendors to implement optimal electronic storefronts that can attract potential consumers and induce their trust. According to Ang & Lee (2000), "if the web site does not lead the consumer to believe that the merchant is trustworthy, no purchase decision will result" (p. 3). In other words, applying trust-inducing features to the Web sites of online vendors is the most effective method of enhancing online trust, given the current state of knowledge.

Efforts have been taken to establish a framework that classifies various trust-inducing Web design features into three broad dimensions: visual design, content design, and social-cue design (Wang & Emurian, in press). The framework is not exhaustive in the sense that it does not attempt to capture every possible trust-inducing feature that web designers can apply. It is focused on articulating the most prominent set of trust-inducing features and presenting them as an integrated entity that can be empirically evaluated and appropriately implemented in Web design. *Table 1* illustrates the framework in detail, including the explanations and design feature examples.

All the trust-inducing interface design factors that are identified in the framework have been illustrated on a synthetic e-commerce interface and evaluated by 181 survey respondents (Wang & Emurian, 2004). Along with identifying the three dimensions, the factors were found to significantly contribute to online trust ratings. This has confirmed what most HCI researchers believe — as Kim & Moon (1998) pointed out — that informative emotions such as trust can be triggered by the customer interfaces and further aid decision making while using e-commerce systems. It may also be concluded that the three dimensions of the framework can act together to promote online trust and reflect the different aspects of Web interface design.

The first two dimensions, which are visual design and content design, are seemingly straightforward, and they have been traditionally the focus of the research that aims to promote online trust by Web design. The last dimension, the social-cue design dimension, is a relatively new design strategy being suggested by numerous researchers (Riegels-

Table 1. Framework of trust-inducing features

Dimensions	Explanations	Design Feature Examples
Visual Design	Defines the graphical design aspect and the structural organization of displayed information on the Web site.	• Use of three-dimensional and half-screen size clipart • Symmetric use of moderate pastel color of low brightness and cool tone • Use of well-chosen, good-shot photographs • Implementation of easy-to-use navigation (simplicity, consistency) • Use of accessible information (e.g., no broken links and missing pictures) • Use of navigation reinforcement (e.g., guides, tutorials, instructions, etc.) • Application of page design techniques (e.g., white space and margin, strict grouping, visual density, etc.)
Content Design	Refers to the informational components that can be included on the Web site, be they textual, graphical, etc.	• Display of brand-promoting information (e.g., prominent company logo or slogan, main selling point) • Up-front disclosure of all aspects of the customer relationship (company competence, security, privacy, financial, and legal concerns) • Display of seals of approval or third-party certificates • Use of comprehensive, correct, and current product information • Use of a relevant domain name
Social-Cue Design	Relates to embedding social and interpersonal cues, such as social presence and face-to-face interaction, into the Web interface via different communication media.	• Inclusion of a representative photograph or video clip • Use of synchronous communication media (e.g., instant messaging, chat lines, video telephony, etc.)

berger, Sasse, & McCarthy, 2003; Steinbruck, Schaumburg, Duda, & Kruger, 2002; Wang, in review). This approach is aimed to remedy the prominent problem of e-commerce known as "lack of human touch" that eliminates online shopping for a considerable number of people.

At least two reasons, or disadvantageous characteristics of e-commerce, contribute to such a problem. First, e-commerce transactions are mostly separated in space and time. Second, a Web site is the only primary and direct "contact point" that online vendors can rely on to interact and communicate with their customers. While the face-to-face interaction can help to establish and stabilize consumer trust in off-line situations, the business transaction in e-commerce is deficient in the personal communication dimension. Therefore, there is need to bring e-commerce interactions closer to off-line shopping experiences by implementing social and interpersonal cues that moderate the disadvantages of an impersonal e-commerce interface and induce online trust. It is such an initiative that compels the social-cue design dimension of the framework.

The social and interpersonal cues, being investigated and embedded into e-commerce Web sites, refer to voice, gestures, appearance, and other communication cues that have been found to have a strong impact on triggering people's trust in face-to-face encounters. Using richer communication media has been seen as a valid means for facilitating the conveyance of these interpersonal cues and providing more opportunities for personal contacts between consumers and online vendors. With the advancement of technology and the increase of bandwidth, a huge collection of communication media is presently available. However, to choose suitable communication media for adding social or interpersonal cues in e-commerce Web sites, designers need to be aware of the different features of each medium, such as channel availability, synchrony, and channel symmetry (Greenspan, Goldberg, Wimer, & Basso, 2000). When implementing social cues, special care should also be taken, as advised by Riegelsberger et al. (2003), to prevent online shoppers from being disappointed by elements lacking functionality other than giving cues of social interaction.

The existing research on interpersonal cues and online trust still remains preliminary. Most research only focuses on examining the trust-inducing capacity of photography, which is the simplest form of communication media to be employed in e-commerce, and the outcomes are found to be somehow contradictory (e.g., Riegelsberger et al., 2003; Steinbruck et al., 2002). Therefore, empirical evidence and valid methodologies are in great need in this research area to address a number of intriguing research questions.

CONCLUSION AND FUTURE DIRECTIONS

As e-commerce gains widespread attention and rapidly emerges as a competitive business form, online merchants are facing an urgent challenge of building and sustaining consumer trust on the Internet. While the issue has initiated numerous investigations for valid research methods and effective solutions by researchers from multiple disciplines, this article shows the merit of an HCI approach in confronting the challenge. Adding trust-inducing interface features and interpersonal cues in Web design has been proposed as a fruitful strategy for building online trust.

Based upon the present overview, five potential areas of suggested research include (1) the effects of culture on online trust; (2) the effects of domain (e.g., .com, .edu, .org) on online trust; (3) the reasons for losing online trust and the ways to repair it; (4) the importance of civil remedies for consumers in case of violations of privacy laws; and (5) the transferability of online trust from the Internet to other activities. In addition, there is obvious need for further investigation on the effects of social and interpersonal cues on online trust, including both methodology development and experimental testing. Finally, it should be pointed out that while well-crafted Web interfaces can induce trust in those who intend to purchase online, online vendors should also pay attention to other methods, such as customer relationship management (CRM) and off-line marketing strategies, to obtain consumer trust and nurture strong business relationships (Tan, Yen, & Fang, 2002).

REFERENCES

Ang, L., & Lee, B.C. (2000). Influencing perceptions of trustworthiness in Internet commerce: A rational choice framework. In *Proceedings of Fifth CollECTer Conference on Electronic Commerce* (pp. 1-12). Brisbane.

Ba, S., & Pavlov, P.A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS Quarterly*, *26* (3), 243-268.

Bhattacherjee, A. (2002). Individual trust in online firms: Scale development and initial test. *Journal of Management Information Systems, 19* (1), 211-241.

Cheskin Research and Studio Archetype/Sapient. (1999). *Ecommerce Trust Study*. Retrieved from *www.cheskin.com/p/ar.asp?mlid=7&arid=40&art=0.*

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2001). Trust in the online environment. In M.J. Smith, G. Salvendy, D. Harris, & R.J. Koubek (eds.), *Usability Evaluation and Interface Design: Cognitive Engineering, Intelligent Agents and Virtual Reality* (pp. 1548-1552). Mahway, NJ: Erlbaum.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, and a model. *International Journal of Human-Computer Studies, 58*, 737-758.

Egger, F.N. (2001). Affective design of e-commerce user interface: How to maximize perceived trustworthiness. In *Proceedings of the International Conference on Affective Human Factors Design.* London: Asean Academic Press.

Gefen, D. (2002). Reflections on the dimensions of trust and trustworthiness among online consumers. *The DATA BASE for Advances in Information Systems, 33* (3), 38-53.

Grabner-Krauter, S., & Kaluscha, E. A. (2003). Empirical research in on-line trust: A review and critical assessment. *International Journal of Human-Computer Studies, 58*, 783-812.

Greenspan, S., Goldberg, D., Wimer, D., & Basso, A. (2000). Interpersonal trust and common ground in electronically mediated communication. In *Proceedings of the ACM2000 Conference on Computer Supported Cooperative Work*, Philadelphia, PA.

Karvonen, K., & Parkkinen, J. (2001). Signs of trust. In *Proceedings of the 9th International Conference on HCI.* New Orleans, LA.

Kim, J., & Moon, J. Y. (1998). Designing towards emotional usability in customer interfaces: Trustworthiness of cyber-banking system interfaces. *Interacting with Computers, 10*, 1-29.

Luhmann, N. (1989). *Vertrauen. Ein Mechanismus der Reduktion sozialer Komplexitat (3rd edition).* Enke, Stuttgart.

Nielsen Norman Group. (2000). Trust: Design guidelines for e-commerce user experience. In *E-commerce User Experience.* Retrieved from *www.nngoup.com/reports/ecommerce.*

Neilsen, J. (2000). *Designing Web usability: The practice of simplicity*. Indianapolis: New Riders Publishing.

Nissenbaum, H. (2001). Securing trust online: Wisdom or oxymoron? *Boston University Law Review, 81*, 101-131.

Riegelsberger, J., Sasse, M. A., & McCarthy, J. D. (2003). Shiny happy people building trust? Photos on e-commerce Websites and consumer trust. In *Proceedings of CHI2003*, Ft. Lauderdale, FL.

Rousseau, D.M., Sitkin, S.B., Burt, R.S., & Camerer, C. (1998). Not so different after all: A cross disciplinary view of trust. *Academy of Management Review, 23* (3), 393-404.

Steinbruck, U., Schaumburg, H., Duda, S., & Kruger, T. (2002). A picture says more than a thousand words: Photographs as trust builders in e-commerce Websites. In *Conference Extended Abstracts on Human Factors in Computer Systems*, Minneapolis, MN.

Tan, X., Yen, D.C., & Fang, X. (2002, Spring). Internet integrated customer relationship management: A key success factor for companies in the e-commerce arena. *Journal of Computer Information Systems*, 77-86.

Wang, Y.D., & Emurian, H. H. (in press). An overview of online trust: Concepts, elements, and implications. *Computers in Human Behavior.*

Wang, Y.D. & Emurian, H.H. (2004). Inducing consumer trust online: An empirical approach to testing e-commerce interface design features. In *Proceedings of the 15th International Conference,* New Orleans, LA.

KEY TERMS

Channel Availability: A feature of any communication medium. A communication medium's channel can be contextual, audio, visual, or any combination of the three. For example, telephone is an audio-only communication medium, while videoconferencing is an audio-visual communication medium.

Channel Symmetry: A feature of any communication medium. A communication medium affords symmetry if the recipient of a message can respond with the same type of message. For example, telephone and e-mail tools are symmetric (two-way) communication media, while television and Web sites are asymmetric (one-way) communication media.

Communication Media: The methods or tools in which information can be exchanged and communication can be facilitated. Examples include telephone, television, e-mail, Web sites, video conferencing, and instant messaging, to name a few.

Customer Relationship Management (CRM): An approach that recognizes that customers are the core of the business and that a company's success depends on effectively managing its relationship with them. CRM is about locating and attracting customers and thereby building long-term and sustainable relationships with them.

Electronic Commerce (E-Commerce): An emerging business form in which the process of buying, selling, or exchanging products, services, and information is undertaken via computer networks, including the Internet.

Online Trust: An Internet user's psychological state of risk acceptance based upon the positive expectations of the intentions or behaviors of an online vendor.

Synchrony: A feature of any communication medium. A communication medium is synchronous if the recipient of a message can respond immediately. For example, telephone and instant messaging are synchronous communication media; while e-mail and voice mail are asynchronous communication media.

Trust-inducing Design: The application of empirically verified features of a Web site to enhance a consumer's perception that the online vendor is trustworthy.

Trustworthy Web site: A Web site that reduces a consumer's perception of risk and that increases confidence in the online vendor's integrity.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2887-2891,

Trust Management in Virtual Product Development Networks

Eric T.T. Wong
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Business-to-business partnerships are gaining rising attention in management and academic research. Increasingly, companies are advised to pursue their collaborative advantage (Dyer, 2000) in order to co-create world-class products, attract the most valuable customers and generate exceptional profits. Today, there is significant global overcapacity in most industries. In this environment of scarce demand, customers are becoming more demanding of customised and innovative products or services. With the advance of information and communication technology (ICT) and the resulting globalisation of markets and manufacturing, innovative product designs are generating new opportunities. In such a change-driven environment, a single manufacturer rarely provides everything on its own anymore. Rather, the most attractive offerings involve buyers and suppliers, allies and business partners in various combinations. Consequently, manufacturers or suppliers do not really compete with one another anymore. Rather, it is offerings that compete for the time and money of customers. The networked business can take different shapes ranging from integrated product development through a key player, to virtual production networks, strategic alliances, virtual organizations, extended enterprises, and so forth.

A review of business publications indicates that companies are extensively using ICT in their new product development activities. Based on an analysis of numerous industrial, high-tech, and business-to-business applications, it appears that ICT can facilitate new product development in a number of areas. These areas can include: speed, productivity, collaboration, communication and coordination, versatility, knowledge management, decision quality, and product quality (Ozer, 2000).

The number of strategic alliances between large, established firms and small, new ventures is on the rise, especially in industries affected by technological change. Theoretically, the combination of a smaller firm's innovative design capabilities with a larger firm's production system and financial prowess promises synergies that can contribute to both firms' competitive advantage, for example, Parts Manufacturers Approval (FAA, 2006) parts as an alternative to Original Equipment Manufacturer (OEM) parts in the aviation industry. Yet, not many of these partnerships result in successful collaboration.

New product development is inherently risky, particularly when new technology or emerging markets are involved. Although collaborative product development has been promoted as a means for reducing or at least sharing risk, such partnerships have their own limitations. Collaboration can also accentuate many of the risks inherent in product development projects. In the case of virtual production networks (VPN), this challenge is even greater because the new product development team spans geographical as well as organizational boundaries.

The basic hypothesis forwarded in this chapter is that a major cause for VPN failure is managerial, and therefore controllable and potentially avoidable. Although today's managers are well-trained in competitive behavior, cooperative processes in VPN require special trust management(TM) skills, skills that a majority of managers do not possess. As a result, cooperation often appears to be managed reactively, rather than being based on a deliberate, proactive cooperation strategy. For a VPN to be competitive and successful in a dynamic environment characterized by constantly changing customer demands and technological innovations, it must be capable of rapid adjustment in order to reduce the time and cost needed to deliver to the customer quality products. The main objective of this chapter is to propose essential guidelines for developing and maintaining partnership trust in Virtual Product Development Networks (VPDN) such that these networks can be managed in a proactive manner. In the following sections, the background of VPDN collaboration will be described, followed by an analysis of the key factors likely contributing to successful VPN collaboration. Based on findings reported in the trust and product development literature, basic requirements for developing and maintaining effective partnership trust in VPDNs have been proposed. Barriers likely to occur in practice are also outlined.

BACKGROUND

The advantages of VPDN collaboration can be significant. The pooling of resources and capabilities can generate synergistic growth between virtual organizations, either in terms of

developing a current product or service offering, or through the creation of an entirely new venture. Increased competitive power can help firms to leapfrog jointly over larger competitors, and generation of higher returns on investment levels can provide the means for further expansion into new market areas at relatively little cost. In an increasingly unpredictable and complex international arena, the flexibility and speed of entry associated with collaboration is opening up new opportunities and possibilities which outright investment through merger or acquisition cannot offer. Indeed, sharing the risks and costs of new product development through collaboration has been advocated by various authors. Securing access to new processes or technologies or gaining information for product development is another frequently mentioned benefit of collaboration. The apparently increasing complexity of technological and product development and convergence of industries provides a strong motive for such collaborative product development relationships.

Marketing considerations may also play an important role in collaborating for product development, especially in the face of increasing globalization of industries. The rapid rate of product obsolescence does, according to some, focus attention on securing rapid access to markets so that new products can be marketed virtually simultaneously in several regions. Collaborative product development relationships may also be seen as a means of overcoming various barriers to entry to foreign markets.

Collaboration can, however, have its shortcomings. History is strewn with the wrecks of failed partnerships left as a warning to the unwary. Many more struggle on without realizing their full potential, frequently to be ultimately bought out by one partner

or the other. With almost 50% deemed as failures, collaborative partnerships are proving to be complex relationships which demand a particular level of expertise and trust management skill in order to navigate the relationship through the hazards associated with this form of virtual network.

Such failures might be due to various causes. For example, there can be a leakage to collaborating partners of a firm's design and analysis skills, experience, and product knowledge that may form a significant part of the basis of its competitiveness. There is a danger that its partners not only acquire the competencies that the design partner brings to the product development, but also gain access to the knowledge and skills that the firm uses in other business areas. A VPDN partner may also fire the opportunism of its collaborators by providing information and insights into possible markets and future possibilities that otherwise may have been its exclusive domain.

Although collaboration is frequently suggested as a means of reducing the cost and duration of the product development process, one would need to consider the additional financial and time costs incurred in managing the collaboration, including the time involved in harmonizing what are likely to be fundamentally different management styles and budgeting processes of the collaborating parties. Furthermore, there can be significant potential opportunity costs because undue effort and resources are directed toward the collaborative product development project, such that the maintenance of the VPDN collaboration itself becomes the prime objective, at the expense of the specific product development. Given the small but growing number of studies reporting dissatisfaction with the outcomes of collaborative product development by one or more of the parties involved, it is understandable that attention should be directed toward key factors affecting the chance of success.

Defining success in product development has been the subject of much research attention and has been shown to be less than straightforward. Defining success in collaborative product development is similarly problematic, given that the perspectives of two or more organizations are involved. The most straightforward measure of the success of a collaborative product development project is likely to relate to whether or not the product was developed as planned and to cost and time allocations. The termination of an agreement cannot inevitably mean the collaboration has been unsuccessful, because the original objectives may have been met. Moreover, the objectives might change as product development progresses. It also has to be recognized that "success" in collaborative product development, as in any product development project, can be multifaceted. There can, for instance, be unintended advantageous side effects, whereas even a prematurely terminated collaborative product development project might yield beneficial experience and knowledge and assist in developing future products.

MAIN FOCUS OF THE CHAPTER

The main focus of this chapter is to examine factors affecting VPDN collaboration with the objective of proposing essential guidelines for developing and maintaining effective partnership trust in VPDNs.

Factors affecting VPDN Collaboration

There has been considerable research into the factors affecting both the success of product development and the outcome of collaborative projects (Lam & Chin, 2005; McDonough, 2000). A number of factors that appear to have some bearing on the success of collaborative ventures have been identified and these will be briefly reviewed here. It is recognized that some of these factors might also have an impact on product development per se, whether collaborative or not, but other factors referred to here are clearly of importance specifically to collaborative product development through the VPN.

A major factor relates to the choice of partner. A particular issue here is the compatibility of the respective cultures of the cooperating organizations. It was suggested that the partnering organizations must be able to communicate with each other, having a language that they all understand. They must have a working style which is complementary, in the way they go about reaching decisions, their problem solving style and so forth. Above all, their management styles must be compatible. There is also evidence suggesting that collaborations that are related to the existing activities of the production partners are more likely to be seen as successful, while others emphasize the value of general experience of collaborations as a factor that enhances the probability of future collaboration success.

Some researchers have stressed the importance of clearly establishing the ground rules for collaboration, such as ensuring that there are clearly defined goals, objectives, and responsibilities for the collaboration that are fully understood by all parties involved. Some of the literature stresses the necessity of preparing detailed and binding initial collaboration agreements in order that future ambiguity is avoided. Such advice corresponds with the recognition of the importance of early and upfront investment in any product development project. It also needs to be recognized, of course, that circumstances change and this alone suggests that there may be need for, first, frequent appraisal of the collaboration and, second, the scope for adaptability. The importance of establishing the limits to the collaboration has also been noted to avoid the transfer of general knowledge and experience during the process of joint product development. Some researchers advise collaborators to impose restrictions and exclusivity clauses in order to limit the transfer of core technologies.

There does, though, need to be a balance between protecting the proprietary interest of the firm while establishing trust and openness with its partners. This factor is regarded by many as a critical ingredient in the continuation and effectiveness of interorganizational relationships. The task for those involved in the management of collaborative product development is to balance these potentially conflicting issues as the project evolves. Related to the establishment of clear ground rules for collaboration is the corresponding need for the monitoring of progress such as through the establishment of milestones, significant points at which progress can be assessed. However, it is obvious, too, that at the outset it is difficult to plan for all the possibilities that might emerge as product development proceeds and this again highlights the need for frequent reappraisal and for a degree of flexibility.

The importance of allocating sufficient financial resources to a collaborative product development project is frequently emphasized, as has been the case for product development more generally. Of course, it is often the allocation of management time and effort that can have a disproportionate influence. The perceived mutuality of contribution and benefits from the various parties involved in a joint product development project has also been highlighted as important. A well managed collaboration, however, will not necessarily result in a profitable outcome. The broader context within which product development takes place is also likely to have a significant bearing. Changes in the partners' markets, in their competitive fields, in the range of technologies available, in the wider economic environment, or in the policies of government agencies can have a critical effect on the project, as can a redefinition of the collaborators' own missions and objectives. Maintaining the necessary external focus may, however, be awarded subsidiary importance given the administrative demands of maintaining the collaboration and the often overriding desire to ensure the collaboration per se is perceived by the partners as proceeding successfully.

Conflict, which affects NPD performance, is inevitable in collaborative NPD. Practicing effective conflict management improves NPD performance as well as helps maintain a long-lasting collaborative relationship. By incorporating the judgments of clients and suppliers using Analytic Hierarchy Process, Lam and Chin (2005) identified and prioritized four categories of success factors for conflict management. The results, based on the synthesized judgments, indicate that of all the factors, the most critical is communication management, followed by trust and commitment to the collaboration.

Existing theories give inadequate attention to differences among VPN members in recognition of these misalignments, interpretation of their origin, proposed corrective actions, and reconciliation of differences. It was found that lack of trust and increased diversity among team members exacerbate such differences (Susman, Gray, Perry, & Blair, 2003).

An effective interface between production and marketing is considered to be vital for the successful development and commercialization of new products. Studies in the U.S., Japan and the UK have, however, identified that conflict between engineers and marketers can act as a barrier to effective cooperation (Shaw, Shaw, & Enke, 2003). It was found that German engineers recognize the importance of trust, good understanding, common knowledge, integration and teamwork in building a good relationship with marketers. The main sources of conflict between German engineers and their marketing colleagues are differences in education and training and different goals and priorities.

To enhance the performance of collaborative product development a VPN must overcome such barriers as resistance to sharing proprietary information, and the not-invented-here syndrome. It may be seen from above that most of the success factors noted suggest that overcoming such barriers depends to a large extent on formal trust development processes. Within the ICT literature, only a handful of studies have examined the recent introduction of ICT applications aimed at helping virtual enterprises electronically collaborate (Mezgar, 2003, 2005; Ratcheva, 2006; Wong, 2005, 2006).

Of these studies, Mezgar (2003, 2006) and Wong (2005, 2006) have helped to advance theoretical understanding of how trust connects with security services and mechanisms and how it affects virtual enterprise operation, respectively. According to Wong (2005), because many organizations will become increasingly more reliant on geographically dispersed NPD teams in the future, companies will need to understand the essential conditions for successful trust building and maintenance in virtual enterprises. In a Pricewaterhouse-Cooper's study on corporate innovation in companies listed on the Financial Times 100, trust was ranked the number one differentiator between the top 20% of companies surveyed and the bottom 20% (Schaub & Altimier, 2006). The top performers' trust empowered individuals to turn strategic aims into reality. People are more innovative in a climate of trust. It is therefore expected that the development of a trust management framework would provide an opportunity for VPN partners to overcome the problems mentioned and eventually improve collaborative product development.

Trust Management Guidelines

Partnerships are distinct from ordinary relationships, as they require at least the restraint of partners from abusing power, a high level of trust and a cultivation of common norms. Trust has been regarded as the foundation of the digital economy (Keen, 2000; Mezgar, 2005; Wong, 2005). A virtual enterprise network is characterized by the impersonal nature of the online environment:

a) the extensive use of information and communication technology (ICT) as opposed to face-to-face transactions,
b) the implicit uncertainty of using an open technological infrastructure for transactions, and
c) the newness of the transaction medium.

Given these attributes, trust development in VPNs presents significant challenges because it is difficult to evaluate partners' trustworthiness without ever having met them (McDonough, Kahn. & Barczak, 2001). Moreover, as the life of many virtual teams is relatively limited, trust must quickly develop (Jarvenpaa & Leidner, 1999).

According to traditional studies, trust builds incrementally and accumulates over time. VPN business relationships, however, are characterized by project-oriented relationships that may entail no past history, nor any plan for future association. In these temporary relationships, time is a vital but often elusive component in the trust building process. This does not mean, however, that trust cannot be apparent in temporary groups. On the contrary, McKnigh et al. (1998) have shown that trust in initial relationships can often be high. Further, Jarvenpaa and Leidner (1999) argue that trust is maximally important in new and temporary organizations, because it acts as a substitute for the traditional mechanisms of control and coordination.

Creating a VPN takes more than just information technology. A study on issues of information technology and management concluded that there is no evidence that IT provides options with long-term sustainable competitive advantage. The real benefits of IT derive from the constructive combination of IT with organisation culture, supporting the trend toward new, more flexible forms of organization. Information technology's power is not in how it changes the organisation, but the potential it provides for allowing people to change themselves. Creating these changes, however, presents a new set of human issues. Among the biggest of these challenges is the management of trust between partner organisations in the VPN (Wong, 2005; Wong & Lau, 2002). Based on findings reported in the literature, a framework for trust management in a VPN can be developed through a serious consideration the following guidelines.

Common Business Understanding

In order to choose the appropriate virtual partners for a collaborative product development project, Fuehrer and Ashkanasy (2001) note that an important element in any business cooperation is the establishment of common business understanding. An earlier work suggests that there are three specifications necessary for the establishment of a common business understanding in the virtual context. The first is a clear product specification: the design, quality, and functionality of the product. The second is specification of the level of cooperation, which requires agreement about deadlines, liability, prices, profit allocation, and staff and resource input. The third is formal specification of agreements between the virtual partners. In a virtual organization, these specifications need to be communicated clearly between the partners to achieve a common business understanding. There is always varying uncertainty between members, however. Therefore, there is a need to guard against opportunistic behavior between the partners. This depends on the risk that the member is prepared to sustain as a potential loss, and also the partner's fear of opportunistic exploitation and the uncertainty of their behavior.

The three specifications (production, cooperation, and agreements between partners) can be achieved by negotiating relational contracts that guide the formation, operation, and dissolution of the virtual organization, thereby facilitating an increase in the level of collaboration-enabling trust. VPNs, like other organizations, create fiscal and legal issues that must be clarified, but they lack a formalized legal framework. Therefore, it is incumbent on the VPN's members to develop their own guidelines for the operation of the enterprise. Such agreements may include clarification of members' tasks and responsibilities, agreement on contracts, allocation of funds, potential liability, and how members will contribute their

expertise. In this sense, clear guidelines, spelled out in an early stage of the partnership, serve to reduce misperceptions and to foster the establishment of trust.

Other mechanisms to establish a common business understanding in VPNs include development of an organization handbook, design of a mutual Internet site, chat room technology, or the use of team addresses for e-mail. A specific example is Livelink, a software selected by Siemens to enable creation of a common business understanding through a standard computer interface.

The concept of common business understanding therefore shares similarities with Organizational Identity, which may be described as a set of distinctive and enduring traits that members associate with their organization. Scott and Lane (2000) have proposed that identity is determined in part by the nature of stakeholder networks. Common business understanding, however, is more akin to Barney's (see Barney et al., 1998, p. 103) broader concept of identity: "the theory organizational members have about who they are." In this respect, the author agreed with Gioia, Schultz and Corley (2000) that Organizational Identity is not necessarily a stable phenomenon, but mutates to suit the prevailing environment. In the virtual context, therefore, a common business understanding may be defined as a transient understanding between network partners as to what they stand for, about the nature of the business transactions that they engage in, and about the outcomes that they expect; their "vision."

Scott and Lane (2000) emphasize that a common business understanding requires the creation of a shared vision, together with communication of mutual aims through clear definition of the roles and expectations within the team, especially in the early stages of the partnership. In this respect, the process is typically initiated by agreement on a symbolic logo or design for a product or service. This is because understanding each member's role, together with group identification, determines critical behaviors such as willingness to cooperate with others, and willingness to engage in mutual goal setting. The VPN partners thus need rapidly to establish group identity and an awareness of mutual needs and expectations, along with the clarification of tasks and responsibilities. In traditional partnerships, awareness and identity are in part shaped by the legal framework that regulates organizational relationships, as well as by networks, artifacts, and the organization chart. In the case of the VPNs, however, mechanisms outside of the domain of traditional organizations need to be put in place to establish a common business understanding, which constitutes an important precursor of trust formation (Jarvenpaa & Leidner, 1999).

Ploetner and Ehret (2006) suggest that successful collaboration rests on a system-wide identification of benefits and that metrics and incentive systems constitute decisive barriers as soon as they no longer apply to system-wide benefits. Hence, the common vision for future benefits, that is, the development of new design capabilities, new products or new technologies, serves as a prime driver for VPDN collaboration.

These examples illustrate how the creation of a sense of shared meaning, member identification, and mission identity, especially in an early stage of the partnership, facilitates collaboration at an individual level and the operation and productivity of the VPN as a whole. As such, a common business understanding provides an essential condition for the development of trust within the organization. In effect, a common business understanding provides the virtual organization's members with an opportunity to share their perceptions of the organization's defined features, and creates a feeling of ownership and trust.

High Ethical Standards

Three factors uniquely characterize the virtual organization's position in regard to business ethics. Firstly, VPNs are rarely guided by pre-existing codified laws, where values and standards are written into legal systems enforceable in court. Because the organization's partners are not usually legally bound to the organization, any negative outcomes or perceptions attributed to poor business ethics could result in the organization's reputation suffering. Second, because VPNs are intrinsically temporary, corporate ethics are difficult to develop because members will typically be finishing one virtual collaboration and entering into another in a short period. Thirdly, VPNs are intrinsically boundary spanning in nature, so that they must incorporate a diversity of culturally-based values and morals.

Researchers focused on the notion of advances in ICT and the related effects on social behavior agree that unethical behavior in the virtual context is predominantly caused by technological changes and by the inside keepers of the information systems. They also agree that social behavior needs more than new laws and modified edicts, and that ethical issues will become increasingly important to enable business transactions to be carried out safely and securely. Although technology has been largely secured by advancing software and technology for virus detection, as well as en/decryption of information to ensure the security of business processes, Johnson (1997) notes that technology can never be sufficient to control all aspects of social behavior. Consequently, online behavior is predicated on an awareness and acceptance of ethical norms and behaviors. This can best be achieved through specification and clarification of the members' tasks, responsibilities and agreed sanctions for proscribed behavior.

Johnson (1997) posits further that the "only hope to control online behavior is for individuals to internalize norms of behavior," and suggests three rules for online ethics: (1) know and follow the rules of the forums participated in; (2) respect the privacy and property rights of others and, if there is any doubt, assume the user's desire for privacy

and ownership; and (3) respect interacting partners by not deceiving, defaming, or harassing them. Not surprisingly, these rules for online behavior are essentially identical to rules for off-line behavior. Indeed, there is no reason why the same ethical guidelines that apply to regular behavior should not be employed in respect to online behavior.

Pearson, Crosby, and Shim (1997) reported on ethical standards for the IS profession proposed by three major professional associations in this field. These associations share an agreed set of behavioral obligations to society, to colleagues, and to professional organizations. The standards aim to promote the principle that individuals within the professions act in an ethical and responsible manner in order to influence the success of their organizations (Pearson et al., 1997). Clearly, similar standards can be developed for the operation of individual VPNs specifying, for instance, the obligation to virtual organization members and clients.

Other possible mechanisms to promote ethical behavior in VPNs include formal codes of ethics, which comprise statements of prescribed and proscribed values or behaviors, and thus provide a strategic tool within organizations to inculcate and to demonstrate ethical standards. Ethical standards also fulfill a strategic external role through recognition by government agencies and insurance companies. Recent surveys show that in the case of VPNs, informal rules known as "netiquette" are usually in place, but a lack of a formal legal infrastructure means that a code of ethics is simultaneously both imperative and difficult to achieve. This is further compounded by different ethical standards and regulations between countries. Nevertheless, trust in interorganizational VPDNs clearly cannot be established until all members recognize that ethical standards are in place and are made aware of what the standards are.

Mutual Forbearance Between VPN Partners

Some researchers approach the issue of trust by defining cooperation as coordination effected through mutual forbearance. Forbearance is refraining from cheating. Cheating may take a weak form (failing to perform a beneficial act for the other party), or a strong form (committing a damaging act). The incentives for forbearance arise from the possibility of reciprocity, leading to mutual forbearance. Parties that are observed to forbear may gain a reputation for this behaviour, which makes them potentially attractive partners for others. The parties to a successful agreement may develop a commitment to mutual forbearance, which cements the partnership, and, in this way, mutual trust is created, which alters the preferences of the parties toward a mutually cooperative mode. Thus, short-term, self-interested behaviour becomes converted to cooperative trusting behaviour.

Demonstrated Capability of VPN Partners

In a VPDN participants will be more willing to share knowledge when they trust in others' ability. It is only natural that they would want to converse with others who have the knowledge and skills regarding the product development project at hand because VPDNs almost always center around a common theme.

Effective Communication and Interaction between VPN Partners

Through communicating with people, we calibrate them, we get a better sense of them and we understand their priorities. Members of VPDN can therefore increase the trust they are giving and the amount they will trust others, by actively seeking opportunities to communicate with other members.

Conflict Recognition and Reconciliation

It is widely acknowledged that effective integration of marketing, product design and manufacturing is vital for the successful development and commercialization of new products and services. The literature suggests, however, that there is much conflict between marketing and engineering personnel that can have a detrimental impact on integration and thus successful new product development (Shaw et al., 2003). A main reason for the existence of such kinds of conflicts appear to be the polarization of functions, with marketing and product designers wanting customized products, whereas the production department wants to manufacture standardized products (Susman et al., 2003). This phenomenon clearly needs to be addressed in order to improve the design-manufacturing interface. One way is to provide education and training for all VPDN partners. This will help different functions in the VPDN to become more sensitive to each other's needs.

Flexible Coordination of Design Activities

As product development involves processes mainly executed by humans, rigid forms of procedural control would create unnecessary conflicts between VPDN partners because people like to keep their freedom regarding the way they work. Product design, similar to other creative processes, evolves according to a kind of anarchic flow of activities. It is therefore necessary to support loosely constrained sets of business processes. Additionally, temporal interdependencies among activities would need to be considered. For example, in the case of product design and process planning, although both processes can proceed with some degree of concurrency

(e.g., process planning can start once the first draft of the engineering design is available), process planning cannot finish before product design finishes. Usually, some details of the process plan depend on the final commitments on the product model. One way of achieving coordination flexibility is through the use of a multi-agent approach (Camarinha-Matos & Afsarmanesh, 2003).

Realistic Expectations of VPN partners

Prior research has indicated that trust creation may be a history-dependent process in which trust accumulates and builds incrementally. Based on this concept, interview and questionnaire data obtained by Adobor (2005) on strategic partnerships from chief executive officers and senior management in the biotechnology, pharmaceutical and medical equipment manufacturing companies in North America showed that trust building in partnerships may be a sort of self-fulfilling prophecy in which initial expectations positively impact behavior and trust building. The results also show that there may be some optimal level of expectations. Both too low and too high an expectation was counterproductive to trust building. It is suggested that VPDN coordinators should develop reasonable product development targets, such that each partner will be able to form realistic initial expectations about each other's design output.

Assuming that most of the above-mentioned conditions have obtained top management support, it is expected that the mutual trust created would enhance the openness within a VPDN (i.e., freedom from censorship and willingness to express innovative ideas) and VPDN partners' ability to reconcile their differences and reach agreements on most of their product development projects.

FUTURE TRENDS

Current research on the management of trust in VPN has limitations. Most work was based on a limited number of case studies. These case studies cannot be considered representative of all VPDN because of their industrial and cultural biases. As most of these models are longitudinal, for generalization purposes it is necessary to test them against the behavior of VPDN over time, a difficult, costly and time-consuming exercise. A possible next step is to confront the above models with a richer, more widely dispersed set of cases, with more cultural and structural variety in the VPDN analyzed in order to investigate its degree of robustness.

While some studies (e.g., Pavlou 2002) posit positive relationships between trust and its consequences, these relationships are nonlinear. Given a minimum threshold for trust to become effective, it is important to recognize this nonlinearity.

Currently, a popular experimental paradigm employed by Human-Computer Interaction (HCI) researchers to assess trust between people interacting via computer-mediated communication covers social dilemma games based on the Prisoner's Dilemma (PD). HCI researchers employing this experimental paradigm currently interpret the rate of cooperation, measured in the form of collective pay-off, as the level of trust the technology allows its users to develop. Some researchers argue that this interpretation is problematic, because the game's synchronous nature models only very specific trust situations (Reigelsberger, Sasse, & McCarthy, 2003). Furthermore, experiments that are based on PD games cannot model the complexity of how trust is formed in the real world, because they neglect factors such as ability and benevolence.

It is noted from the literature that little theoretical explanation exists in order to understand the impact of trust in various forms of VPN relationships. It has been found that firms in horizontal alliances would display a lower level of organizational trust and a weaker relationship between interfirm cooperation compared to firms in vertical integration of alliances and that trust is unrelated to cooperation in horizontal alliances (Rindfleisch & Moorman, 2001). It was suggested that this different impact of trust could be due to higher opportunism, lower interdependency, and stronger institutional linkages among horizontal collaborators compared to their vertical counterparts. If this finding is substantiated by future empirical research in the VPN domain, researchers may need to reconsider the popular notion that trust is an essential component of all types of relationship exchanges in the VPN.

In view of the positive association between expectations and trust, finding out how VPDN partners form expectations about each other will yield important insights. Perhaps they are influenced by factors external to the VPDN relationship, such as the amount of technical investments, rate of technological progress or the reputation of partners.

CONCLUSION

Collaborative teamwork offers greater chance for VPDN to be successful in nowadays' agile competition. Trust models based on traditional familiarity would not meet the special needs of the VPDN, which is having a much shorter life cycle and involves a large number of partners who have never met before. Consequently, the VPDN partners must realize the need to effectuate this paradigm shift. This chapter identifies eight essential guidelines needed for effective trust management in a virtual environment: a common business understanding, high ethical standards, mutual forbearance between partners, capability of partners, effective communication and interaction within the VPDN, conflict rec-

ognition and reconciliation, flexible coordination of design activities, and reasonable expectations about VPDN partners. VPDNs with high levels of trust among their members can effectively utilize interactions and communication processes at their interfaces so members can learn together, and can develop shared mental models of reliability and a shared culture of safety. It is anticipated that the likely impact of trust management would imply continuous product innovation because trust plays an important synthesis role. VPDN with its flexible organizational structures can leverage the partners' ability and willingness to learn, thereby enhancing new product developments.

REFERENCES

Adobor, H. (2005). Trust as sensemaking: The microdynamics of trust in interfirm alliances. *Journal of Business Research, 58*, 330-337.

Barney, J.B., Bunderson, J.S., Foreman, P., Gustafson, L.T., Huff, A.S., Martins, L.L., et al. (1998). A strategy conversation on the topic of organizational identity. In D.A. Whetton & P.C. Godfrey (Eds.), *Identity in organizations: Building theory through conversations* (pp. 99-168). Thousand Oaks, CA: Sage.

Camarinha-Matos, L.M., & Afsarmanesh, H. (2003). Elements of a base VE infrastructure. *Computers in Industry, 51*, 139-163.

Dyer, J.F. (2000). *Collaborative advantage: Winning through extended enterprise supplier networks*. Oxford University Press.

Federal Aviation Administration. (2006). *Parts manufacturer approval: Regulations & policies*. Retrieved May 27, 2008, from http://www.faa.gov/aircraft/air_cert/design_approvals/pma/pma_regs/

Fuehrer, E.C., & Ashkanasy, N.M. (2001). Communicating trustworthiness and building trust in interorganizational virtual organizations. *Journal of Management, 27*(3), 235-254.

Gioia, D.A., Schultz, M., & Corley, K.G. (2000). Organizational identity, image, and adaptable instability. *Academy of Management Review, 25*, 63-81.

Jarvenpaa, S.L., & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organizational Science, 10*(6), 791-815.

Johnson, D. (1997). Ethics online. *Communications of the ACM, 40*(1), 60-65.

Keen, P.G.W. (2000). Ensuring e-trust. *Computerworld, 34*(11), 46.

Lam, P.K., & Chin, K.S. (2005). Identifying and prioritizing critical success factors for conflict management in collaborative new product development. *Industrial Marketing Management, 34*(8), 761-772.

McDonough, E.G. (2000). Investigation of factors contributing to the success of cross-functional teams. *Journal of Product Innovation Management, 17*(3), 221-235.

McDonough, E., Kahn, K., & Barczak, G. (2001). An investigation of the use of global, virtual, and collocated new product development teams. *The Journal of Product Innovation Management, 18*(2), 110-120.

Mezgar, I. (2003). Role of trust in networked production systems. *Annual Reviews in Control, 27*(2), 247-254.

Mezgar, I. (2005). Building trust in virtual communities, In S. Dasgupta (Ed.), *Encyclopedia of virtual communities and technologies* (pp. 4-9). Hershey, PA: Idea Group Reference.

Ozer, M. (2000). Information technology and new product development: Opportunities and pitfalls. *Industrial Marketing Management, 29*(5), 387-396.

Pavlou, P.A. (2002). Trustworthiness as a source of competitive advantage in online auction markets. In *Best Paper Proceedings of the Academy of Management Conference*, Denver, CO, (pp. 9-14).

Pearson, J.M., Crosby, L., & Shim, J.P. (1997). Measuring the importance of ethical behavior criteria. *Communications of the ACM, 40*(9), 94-100.

Ploetner, O., & Ehret, M. (2006). From relationships to partnerships—new forms of cooperation between buyer and seller. *Industrial Marketing Management, 35*, 4-9.

Riegelsberger, J., Sasse, M.A., & McCarthy, J. (2003, April 20-25). Shiny happy people building trust? Photos on e-commerce Web sites and consumer trust. In *Proceedings of CHI 2003*, Ft. Lauderdale, FL, (pp. 121-128).

Rindfleisch, A., & Moorman, C. (2001). The acquisition and utilization of information in new product alliances: A strength-of-ties perspective. *Journal of Marketing, 65*, 1-18.

Schaub, A., & Altimier, L. (2006). Tenants of trust: Building collaborative work relationships. *Newborn and Infant Nursing Reviews, 6*(1), 19-21.

Scott, S.C., & Lane, V.R. (2000). A stakeholder approach to organizational identity. *Academy of Management Review, 25*, 43-62.

Shaw, V., Shaw, C.T., & Enke, M. (2003). Conflict between engineers and marketers: The experience of German engineers. *Industrial Marketing Management, 32*(6), 489-499.

Susman, G.L., Gray, B.L., Perry J., & Blair, C.E. (2003). Recognition and reconciliation of differences in interpretation of misalignments when collaborative technologies are introduced into new product development teams. *Journal of Engineering and Technology Management, 20*(1-2), 141-159.

Wong, T.T. (2005). Trust in virtual enterprises. In M. Khosrow-Pour (Ed.), *Encyclopedia of Information Science and Technology, 5*, 2902-2909. Idea Group.

Wong, T.T. (2006). Neural data mining system for trust-based evaluation in smart organizations. In I. Mezgar (Ed.), *Integration of ICT in smart organizations* (pp. 159-185). Idea Group.

Wong, T.T., & Lau, H.C.W. (2002). The impact of trust in virtual enterprises. In A. Gunasekaran (Ed.), *Knowledge and information technology management in the 21st century organizations: Human and social perspectives* (pp. 153-168). Idea Group.

KEY TERMS

Virtual Enterprise: A temporary business organization set up between trading partners operating from geographically dispersed sites, for the duration of a common project. The design and manufacture of new products or services frequently requires the talents of many specialists. When many corporations combine their specialties to create a product or service, the result can be called a virtual enterprise. A virtual enterprise must be able to form quickly in response to new opportunities and dissolve just as quickly when the need ceases.

Virtual Private Network (VPN): A private communication network often used within a company, or by several different companies or organizations, to communicate confidentially over a publicly accessible network.

Agile: Being agile means being proficient at change, and allows an organization to do anything it wants to do whenever it wants. As virtual enterprises do not own significant capital resources of their own, this helps to make them agile, as they can be formed and changed very rapidly.

Ethical: Conforming to standards of professional or social behavior agreed by all members of a virtual enterprise.

Self-Fulfilling Prophecy: A predetermined idea or expectation one has toward oneself that is acted out, thus "proving" itself. For example, in the stock market, if it is widely believed that a crash is imminent, investors may lose confidence, sell most of their stock, and actually cause the crash.

Traditional Familiarity: Traditional familiarity combines an assumption of continuity with the past experience of a partner. Traditional trust therefore relies on the fact that VPN partners who could be observed as trustworthy in the past will display the same kind of behavior in the future.

U.S. Disabilities Legislation Affecting Electronic and Information Technology

Deborah Bursa
Georgia Institute of Technology, USA

Lorraine Justice
Georgia Institute of Technology, USA

Mimi Kessler
Georgia Institute of Technology, USA

INTRODUCTION

The Americans with Disabilities Act (ADA) is the cornerstone legislation to address the civil rights of people with disabilities, including making products, services, and physical environments accessible to them. Almost everyone in the U.S. is familiar with the ADA, but designers of technology products and services need to be aware of accessibility standards that go beyond the ADA: specifically, Section 508 of the Rehabilitation Act and Section 255 of the Communications Act. These laws define accessibility standards and guidelines that impact the design of electronic, information, and telecommunication technologies, and they are intended to promote products and services that are as accessible to persons with disabilities as those without (Section 508, 1998). Furthermore, with the aging of the U.S. and world populations (Forrester, 2004), the number of people who want to use technology but cannot, because of disabilities, is on the rise. Designers of technology need to understand how modifications in traditional design will make products more marketable and usable by a wider range of customers. This article reviews important aspects of Sections 508 and 255, assistive technology and accessible design, and additional sources of information and training.

BACKGROUND

Key Implications of the Legislation

Section 508 of the Rehabilitation Act states that "each Federal department or agency, including the United States Postal Service…when developing, procuring, maintaining, or using electronic and information technology (EIT)…shall ensure…that individuals with disabilities…have comparable access" (Section 508, 1998). Section 508 required the Architectural and Transportation Barriers Compliance Board (also known as the Access Board) to develop accessibility standards for EIT, and it stipulated that the law applied to procurements after June 25, 2001 (Federal Acquisition Regulations, 2001).

The technology addressed in the Section 508 technical standards cover: "software applications; operating systems; Web-based intranet and Internet information and applications; telecommunications products; video and multimedia products; self-contained, closed products; and desktop and portable computers" (Federal Acquisition Regulations, 2001). The standards also address information and documentation including "product support in alternative formats, descriptions of accessibility and compatibility features in alternative formats, and product support services in alternative communications modes" (Access Board, 2001). Federal departments and agencies can be exempt from compliance only if they can show that compliance is an "undue burden" as a result of "significant difficulty or expense." Prior to the Section 508 amendments, Section 501 (federal employment) and Section 504 (federally funded programs and activities) of the Rehabilitation Act addressed accommodation of individuals. The amendments to Section 508 addressed the technology itself, that is, making it accessible to everyone right "out of the box" (although this does not necessarily eliminate the need for individual accommodation). It was thought that by making accessibility requirements a part of the federal procurement process, there would be financial incentives for companies to design products that meet these standards.

The Section 508 standards apply specifically to the United States government when purchasing, developing, maintaining, or using electronic and information technology. In addition, an increasing number of states purchase electronic products and services that conform to the Section 508 standards (or similar state-developed regulations or stan-

dards). Due to the complexity of the regulations and range of requirements, creators and vendors of technology products need information and training to ensure their products and services adhere to the 508 standards.

The Telecommunications Act of 1996, which was the first major overhaul of American telecommunications policy in nearly 62 years, added *Section 255* to the Communications Act of 1934. Section 255 requires telecommunications manufacturers and providers of telecommunications services to make their products and services accessible to and usable by people with disabilities if "readily achievable." The Federal Communications Commission (FCC) makes readily achievable determinations on a case-by-case basis, but generally, companies with more resources need to do more to make their products and services accessible to people with disabilities. When it is not possible to provide direct access, Section 255 requires manufacturers and providers to make their devices and services compatible with peripheral devices and specialized customer premises equipment (CPE) that are commonly used by people with disabilities, if readily achievable. Examples of specialized CPE include teletypewriters (TTYs) and assistive listening devices.

Section 255 also requires companies that develop telecommunications products and services to include the following activities as business practices:

- When the company conducts market research, product design, testing, pilot demonstrations, and product trials, it should include individuals with disabilities in target populations of such activities.
- Companies should work cooperatively with disability-related organizations.
- Companies should undertake reasonable efforts to test access solutions with people with disabilities. (Federal Communications Commission, 2002)

Unlike Section 508, Section 255 is not restricted to just the federal marketplace; it applies to telecommunications products and services purchased by anyone in the U.S. Section 508 covers a wide variety of disabilities: people who are deaf or hard of hearing, who have mobility or dexterity impairments, who have speech impairments, and those who have low vision or who are blind. Section 255 includes all of these impairments as well as cognitive disabilities. While the standards, guidelines, and directives associated with Section 508 and Section 255 may appear to complicate the product design process, often these challenges bring about innovation and new product ideas. In most cases, these innovations lead to product features that are desired by a customer market that is much larger than the disability community.

Disabilities and Applications

People with disabilities need different product and service features so they can access information and communicate with others at a level that is equal to those without disabilities. Often an assistive technology is needed by someone with a disability to overcome access barriers. For example, a text-to-speech software program, commonly called a "screen reader," is an assistive technology that allows someone who is blind to access electronic information, such as Web pages on the Internet. Designers of Web sites and software applications need to understand accessibility requirements so they can make their content accessible to users of screen readers. For example, Web sites and other software programs need to have text "tags" that describe every non-trivial image used in their applications. The screen reader reads these descriptions aloud so the user with low or no vision can understand the information being conveyed. These tags benefit sighted people too because the tag's text appears whenever they "mouse over" the image, which can help the users identify the function of an icon.

People who are deaf need visual assistance to access information that is typically delivered aurally. As a result of the Television Decoder Circuitry Act of 1990, televisions must now be manufactured with the circuitry necessary to show captioning. In this case, the viewer requires no assistive technology, and these caption-ready TVs are examples of making technology accessible "out of the box." Continuing the example related to Web site design, when sound is used to communicate information, there should be text to notify the user of the presence of sound and to describe the sound itself. Often the solution is as simple as a caption that says "music."

Another assistive technology is voice-to-text software, commonly called "voice recognition software," which assists those with motor or dexterity limitations to transmit information through speech. While this assistive technology was initially developed for people with disabilities, voice recognition is a popular feature for many technology products and is in high demand by the mass market.

Training for Accessible Design

Accessibility training is available to designers in several places around the country, and there is a wealth of information on the Internet. Web sites listed in the resources section of this article provide contacts for online training, courses, and conferences. Many of these courses are "hands on" so designers and information specialists can see the assistive technology software packages in action. Online training sessions are available through the national centers listed at the end of this article. Several centers have training materials

that can be downloaded or requested through the U.S. Mail. Professional societies, such as the Human-Computer Interaction (HCI) group of the Association for Computing Machinery (ACM), American Institute of Graphic Arts (AIGA), and the Industrial Design Society of America (IDSA) include peer-reviewed papers and research on related topics.

FUTURE TRENDS

It is important for information providers and designers to know the requirements and guidelines of Sections 508 and 255 before they begin a project. If they include people with disabilities early in their user-centered design processes, they can avoid the cost of retrofitting later. In addition, increased software and hardware capabilities will ease compliance through software that easily converts information to text, or sound to image.

CONCLUSION

Section 508 of the Rehabilitation Act and Section 255 of the Communications Act are the two primary federal laws that directly affect the design of accessible electronic, information, and telecommunication technology in the U.S. These laws go beyond the Americans with Disabilities Act by defining specific standards and guidelines that address the accessibility needs of people with disabilities. Designers of technology, therefore, should understand and apply these principles when designing electronic and information technology that will be sold to the federal government (Section 508), or when developing telecommunications products and services that will be sold to the public (Section 255).

It would be a mistake, however, to assume that federal legislation is the only factor driving the need for more accessible technology. The U.S. population is aging, and with the years come age-related disabilities such as visual, hearing, and mobility impairments. The appeal of accessible technology, therefore, extends to a larger portion of the overall market, and accessibility requirements should become an essential consideration in the technology design process.

REFERENCES

Access Board. (2000). *36 CFR, SubPart 1194 text.*

Federal Acquisition Regulations (FAR). (2001). *Federal Register*, (April 25). Retrieved from *www.section508.gov/index.cfm?FuseAction=Content&ID=13*

Forrester. (2004). Study commissioned by Microsoft. Retrieved from *www.microsoft.com/presspass/press/2004/feb04/02-02AdultUserBenefitsPR.asp*

Section 508. (1998). *Section 508 of the Rehabilitation Act (29 U.S.C. 794d), as amended by the Workforce Investment Act of 1998* (P.L. 105-220), August 7, 1998.

Vanderheiden, G.C. & Tobias, J. (2000). *Universal design of consumer products: Current industry practice*. Madison, WI: Trace Research and Development Center.

RESOURCES

Architectural and Transportation Barriers Compliance Board

- Also known as the Access Board: www.access-board.gov
- ADA Accessibility Guidelines for Buildings and Facilities: www.access-board.gov/adaag/html/adaag.htm
- Section 508 standards: www.section508.gov/index.cfm?FuseAction=Content&ID=12
- Section 255 guidelines: www.access-board.gov/telecomm/html/telfinl2.htm

Section 508 of the Rehabilitation Act

- Also known as Section 508: www.Section508.gov

Federal Communications Commission

- Also known as the FCC: www.FCC.gov
- Section 255 information page: www.fcc.gov/cgb/dro/section255.html

Department of Justice

- Also known as the DOJ: www.usdoj.gov
- Guide to Disability Rights Laws: www.usdoj.gov/crt/ada/cguide.htm
- ADA Title III Technical Assistance Manual: www.usdoj.gov/crt/ada/taman3.html

Information Technology Technical Assistance and Training Center

- Also known as ITTATC: www.ittatc.org
- Accessibility in the user-centered design process: www.ittatc.org/technical/access-ucd/

- Web accessibility online course: www.ittatc.org/training/Webcourse/
- Product accessibility evaluation Webcast: www.tvworldwide.com/event_020314_ittatc.cfm
- Other Webcasts about accessibility and Sections 508 and 255: www.ittatc.org/training/Web_training.cfm
- Speak Out! about inaccessible information and telecommunication technology: www.ittatc.org/technical/speakout/index.cfm

The TRACE Research Center at the University of Wisconsin–Madison

- Also known as the TRACE Center: www.trace.wisc.edu
- Collation of Access Board's 508 Final Rule and Guides: trace.wisc.edu/docs/508-collation/index.shtml?style=default
- Product design ideas browser: trace.wisc.edu/docs/browser/index.html

National Institute on Disability and Rehabilitation Research

- Also known as NIDRR: www.ed.gov/about/offices/list/osers/nidrr/index.html?src=mr
- List of NIDRR-funded programs and projects: www.ed.gov/rschstat/research/pubs/programs.html

Center for Assistive Technology and Environmental Access

- Also known as CATEA: www.catea.org
- List of CATEA projects: www.catea.org/projects.html

KEY TERMS

Accessible Technology: "[T]echnology that can be used by people with a wide range of abilities and disabilities. It incorporates the principles of universal design. Each user is able to interact with the technology in ways that work best for him or her. Accessible technology is either directly accessible—in other words, it is usable without assistive technology—or it is compatible with standard assistive technology" (Knowledgebase entry from University of Washington: www.washington.edu/accessit/articles?110).

Assistive Technology: Any item, piece of equipment, or system that is commonly used to increase, maintain, or improve functional capabilities of individuals with disabilities. Examples include screen readers, teletypewriters (TTYs), and Braille keyboards.

Disability: Under the ADA, an individual with a disability is a person who: (1) has a physical or mental impairment that substantially limits one or more major life activities; (2) has a record of such an impairment; or (3) is regarded as having such an impairment.

Rehabilitation Act: A federal law that was created to empower individuals with disabilities to maximize employment, economic self-sufficiency, independence, and inclusion and integration into society. It also was enacted to ensure that the federal government plays a leadership role in promoting the employment of individuals with disabilities. You can read the text of the act at www.ed.gov/policy/speced/leg/rehabact.doc.

Section 255: A part of the Communications Act. Section 255 requires telecommunications manufacturers and providers of telecommunications services to make their products and services accessible to and usable by individuals with disabilities, if readily achievable. For more information, see www.access-board.gov/telecomm/html/telfinal.htm.

Section 508: A part of the Rehabilitation Act. Under Section 508, agencies must give employees with disabilities and members of the public access to information that is comparable to the access available to others. The law applies to all federal agencies when they develop, procure, maintain, or use electronic and information technology. For more information, see www.section508.gov.

Telecommunications Act: The first major overhaul of American telecommunications policy in nearly 62 years. This Act added Section 255 to the Communications Act of 1934. See www.fcc.gov/telecom.html for more information.

Universal Design: "A process of creating products (devices, environments, systems, and processes) which are usable by people with the widest possible range of abilities, operating within the widest possible range of situations (environments, conditions, and circumstances), as is commercially practical" (Vanderheiden & Tobias, 2000).

ENDNOTES

1 Forrester, 2004. Study commissioned by Microsoft: http://www.microsoft.com/presspass/press/2004/feb04/02-02AdultUserBenefitsPR.asp

2 Section 508, 1998. Section 508 of the Rehabilitation Act (29 U.S.C. 794d), as amended by the Workforce Investment Act of 1998 (P.L. 105-220), August 7, 1998 Full text of Section 508

3 Access Board, 2000 36 CFR, SubPart 1194 text

[4] Federal Acquisition Regulations (FAR), As published in the Federal Register April 25, 2001 http://www.section508.gov/index.cfm?FuseAction=Content&ID=13

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2916-2920, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

U.S. Information Security Law and Regulation

Michael J. Chapple
University of Notre Dame, USA

Charles R. Crowell
University of Notre Dame, USA

INTRODUCTION

The American legal system, along with many of its counterparts around the globe, is only beginning to grapple with the legal challenges of the knowledge age. The past decade has witnessed a multitude of new laws and regulations seeking to address these challenges and provide a common framework for the legal and technical professions. Those charged with information security responsibilities face a myriad of complex and often vague requirements. In this article, we establish a four-level taxonomy for information security laws and explore the major components of each level.

BACKGROUND

Mohamed Chawki, in a study of computer crime law, points out that the traditional definition of a computer crime as any crime that involves "the knowledge of computer technology for its perpetration, investigation, or prosecution" is far too broad for practical application (Chawki, 2005, p. 7). Virtually every crime involves computer technology at some point in the investigative process. For example, a common burglary should not be considered a computer crime merely because the booking officer entered data on the crime into a department information system. Similarly, the fact that the criminal looked up driving directions on the Internet should not make a bank robbery a computer crime.

We seek to clarify these issues by creating a general taxonomy of information security laws. Our taxonomy includes the following four levels:

- **Intellectual property laws** protect the rights of authors, inventors and creators of other intellectual works.
- **Computer-focused crime laws** define transgressions and applicable punishments for offenses where the use of a computer is intrinsic to the crime.
- **Computer-related crime laws** are those laws that involve the use of a computer but where the criminal activity is not defined by the use of a computer. This category also includes those laws that require the use of computers to assist in the investigation of a crime.
- **Industry-specific laws** do not apply to society as a whole but, rather, govern particular industries and are typically focused on protecting the confidentiality, integrity and/or availability of personal information.

It is also important to note that many information security crimes are prosecuted under traditional laws, rather than the specific laws presented in this taxonomy. Smith (2005) points out two examples of this: the charging of an individual with a felony offense for accessing an unprotected wireless network and a school district's charge of criminal trespass against 13 students who accessed laptops issued to them with an administrative password that was taped to the bottom of the machines.

In the remainder of this chapter, we seek to explore this taxonomy in further detail. While the taxonomy may be applied to any body of law, due to space constraints, this article limits the discussion to federal laws in the United States. A myriad of state and local laws, as well as the laws of other nations, may also be classified under this taxonomy.

INTELLECTUAL PROPERTY LAW

The legal principles protecting the rights of owners of creative works date back several centuries. As our society shifts from an industrial economy to a knowledge economy, these laws become increasingly important, as they protect the very essence of our economic engine. These intellectual property laws are critical to any information security program, as they provide the legal basis for protecting the intellectual property rights of individuals and organizations.

Copyrights

Copyrights protect any original work of authorship from unauthorized duplication or distribution. The Copyright Act defines eight categories that constitute covered works (Copyright Act, 1976). One of these categories, literary works, is broadly interpreted to include almost any written work. This category has traditionally been used to include computer software, web content and a variety of other works of direct interest to information security professionals.

Copyright protection is automatic upon the creation of a work. For works created after 1978, copyright protection lasts for 70 years after the death of the last surviving author.

Trademarks

Trademark law protects words, phrases and designs that identify the products or services of a firm. The essential characteristic of a trademark is that it must uniquely distinguish the trademark holder's goods or services from those of other providers. Therefore, trademarks may not be simply descriptive of the product or service but must contain the element of uniqueness. For example, it would not be possible to gain trademark protection on the term "blue automobile," while it may be possible to gain protection for the term "Blue Streak Automobiles".

Trademark protection is afforded by the Lanham Act (1946). The U.S. Patent and Trademark Office grants registrations with an initial duration of 10 years and the option to renew.

Patents

Patents protect inventions, processes, and designs. They grant the inventor substantial protection in the form of exclusive rights to the patented concept. To protect against the abuse of this privilege, the U.S. Patent and Trademark Office strictly governs the issuance of patents. The three requirements for patent protection are that the invention must be novel, useful, and nonobvious. Patents granted for inventions or processes are valid for 17 years while design patents are valid for 14 years (Patent Act, 1952).

Trade Secrets

The Economic Espionage Act of 1996 makes it illegal to steal, misappropriate, duplicate or knowingly receive or possess a trade secret without appropriate permission. Trade secrets include any information that "derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable by proper means by the public" and are the subject of "reasonable measures to keep such information secret" (Economic Espionage Act, 1996).

When designing an information security program, it is essential to recognize that trade secrets are defined by the confidentiality protection afforded them. If an organization fails to take reasonable efforts to maintain the confidentiality of a trade secret, this protection is lost. This is a major departure from patent protection, which requires public disclosure of the invention. Public disclosure of a trade secret nullifies the protection afforded to that secret and effectively releases it into the public domain. Unlike patents, however, trade secrets enjoy indefinite protection under the law.

Digital Millennium Copyright Act

The Digital Millennium Copyright Act (1998) instituted a number of significant changes in U.S. copyright law. In addition to procedural changes required to implement World Intellectual Property Organization (WIPO) treaties, DMCA makes several modifications to the law designed to accommodate the changing digital environment of the Internet. For example, DMCA offers a safe harbor provision for Internet service providers, absolving them of liability for the infringing acts of customers, provided that they have policies to terminate the accounts of repeat copyright offenders and do not interfere with the technical measures used by copyright holders to protect their works. If providers meet these requirements, they are protected from liability caused by transitory communications, system caching, information residing on systems or networks at the direction of users and information location tools (such as search engines).

COMPUTER-FOCUSED CRIME LAW

Computer-focused crime laws center upon the transgressions and associated punishments when the use of a computer is intrinsic to the crime. When drafting computer-focused crime laws, legislators have the specific intent of outlawing the use of a computer to commit a crime. This category is distinct from the next category, computer-related crime laws, crimes in which the perpetrator may utilize a computer as a support tool. For example, a law prohibiting the use of a computer to eavesdrop on the electronic mail of an individual is a computer-focused crime law. It is the act of using the computer to eavesdrop that is the essential nature of the crime.

Computer Fraud and Abuse Act

Congress originally passed the Computer Fraud and Abuse Act of 2001 in 1986 and later amended it in 1994, 1996, and 2001 to reflect the rapidly changing digital environment. Originally intended to protect data contained on the computers of government agencies and financial institutions, later amendments expanded the scope to include any system involved in interstate commerce (Burke, 2001). Offenses under the Computer Fraud and Abuse Act include gaining unauthorized access to a computer.

Electronic Communications Privacy Act

The Electronic Communications Privacy Act (ECPA) of 1986 protects the rights of individuals who become the subject of electronic surveillance by government agencies or other third parties. It includes two separate components: the Wiretap Act and the Stored Communications Act (SCA).

The Wiretap Act makes it illegal to intercept (or attempt to intercept) any wire, oral or electronic communication outside of several specific circumstances identified in the law (such as when approved by a court order or when conducted as part of a quality assurance monitoring effort). The SCA protects communications stored on a computer against unauthorized access or alteration.

COMPUTER-RELATED CRIME LAW

The third category in our taxonomy, computer-related crime laws, includes laws that govern crimes which commonly involve the use of a computer but do not meet the criteria of a computer-focused crime.

Child Pornography Laws

Society has long held the tenet that any molestation or exploitation of children via the creation or distribution of pornography is objectionable and has codified this abhorrence in the law. Unfortunately, the growth of the Internet has led to an increased ability of child pornography traffickers to market their wares with a greater degree of anonymity. This technological shift required a corresponding shift in the law.

The majority of child pornography prosecutions take place under Title 18, Section 2252 of the United States Code (Waters & Harrell,1997). This law bans the interstate or foreign transportation of sexually explicit materials that involve minors and was amended specifically to include the transmission of such materials through the use of a computer.

The Child Protection and Obscenity Enforcement Act of 1988 requires that the producers of sexually explicit materials maintain documented records of the ages of all actors and models used in their productions.

Identity Theft and Assumption Deterrence Act

The Identity Theft and Assumption Deterrence Act of 1998 amended federal law to address computer-related elements encompassed by the crime of identity theft. Specifically, Congress outlawed the possession or use of electronic devices, computer hardware and computer software designed primarily for the production of false identity documents. This Act also modified the definition of "means of identification" under the law to include biometric data, electronic identification numbers, addresses or routing codes, and telecommunication identifying information or access codes.

USA PATRIOT Act

The Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism (USA PATRIOT) Act enhances the authority granted to the federal government when conducting counterterrorism operations and places additional requirements on service providers. In a legal summary of the act, Iuliano (2003) notes that the critical changes that impact information security programs are that the Act:

- Exempts voicemail from wiretap requirements, allowing law enforcement officials access through a search warrant
- Provides law enforcement officials with authority to track and monitor Internet traffic
- Increases penalties for computer-focused crimes

Communications Assistance for Law Enforcement Act

Law enforcement agencies have long employed the use of court-ordered wiretaps in investigations to obtain evidence of criminal activity. Up until the past two decades, agents could implement these wiretaps simply by attaching electronic eavesdropping devices to an analog telephone network. The emergence of digital and mobile communications devices increased the technical difficult of implementing wiretaps and caused Congress to pass the Communications Assistance for Law Enforcement Act (CALEA) of 1994. CALEA requires that communications providers cooperate with law enforcement efforts to obtain wiretaps and to do so in a manner that cannot be detected by the communicating parties.

For ten years, both the government and telecommunications providers interpreted CALEA to apply to voice communications over telephone networks. In a 2005 notice of proposed rulemaking, the Federal Communications Commission stated that the government intends to apply CALEA to Internet service providers (Federal Communications Commission, 2004). This new interpretation of CALEA raises a number of critical issues as it requires service providers to make substantial equipment investments in order to comply. For example, a coalition of higher education argued to the FCC that the proposed interpretation would impose an unjustified cost burden upon academia (Higher Education Coalition, 2005).

INDUSTRY-SPECIFIC LAW

In addition to the broad laws identified in the previous sections of this taxonomy, there are a number of laws that apply

to specific industries, due to their unique access to sensitive data. These include regulations on healthcare providers, financial institutions, public corporations, and others.

Child Online Privacy Protection Act

The Child Online Privacy Protection Act (COPPA) of 1998 regulates the conduct of business with minors using the Internet. It requires that businesses obtain parental consent before knowingly collecting personal information from children under the age of 13. It also requires that these online services provide parents with any information collected from their children, offers them the opportunity to revoke consent, and demands the removal of such information at any time upon parental request. As pointed out by Isenberg (2000), the cost of compliance with this Act may be steep. He illustrates this point through the case of SurfMonkey, a site which reportedly spent over $50,000 on COPPA compliance and instituted a 4,673 word privacy policy.

Health Insurance Portability and Accountability Act

In 1996, Congress enacted the Health Insurance Portability and Accountability Act (HIPAA). Among its many provisions, HIPAA implemented privacy and security requirements for healthcare providers, health insurance plans and health information clearinghouses. The Privacy Rule creates a new classification of data: protected health information (PHI) which includes several categories of data related to an individual's health.

The Privacy Rule requires that covered organizations only disclose PHI to authorized individuals and organizations. The HIPAA Security Rule provides five categories of safeguards that must be applied to electronic PHI (ePHI):

- Administrative Safeguards
- Physical Safeguards
- Technical Safeguards
- Organizational Requirements
- Policies and Procedures

HIPAA is one of the most comprehensive industry-specific laws and it provides a full blueprint for the protection of healthcare data. It also provides specific consequences for knowing and willful violations of the law. Anyone who obtains or discloses PHI in violation of HIPAA may be fined up to $250,000 and imprisoned for up to 10 years.

Gramm-Leach-Bliley Act

Just as HIPAA requires that healthcare fs protect the privacy and security of patient records, the Gramm-Leach-Bliley Act (GLBA) of 1999 requires that financial institutions protect the privacy and security of individual financial records. GLBA's Financial Privacy Rule requires that financial institutions provide customers with a copy of the institution's privacy policy. GLBA's Safeguards Rule requires financial institutions to design and implement safeguards to protect customer information.

Sarbanes Oxley Act

The Sarbanes Oxley Act (SOX) of 2002 instituted sweeping reforms in the way public corporations conduct business and report financial results. While the majority of SOX requirements pertain to financial reporting and traditional accounting controls, the law does specifically impact information security. The law requires that institutions implement sufficient information security controls to ensure the validity and reliability of financial reports. The broad applicability of this law to all publicly traded corporations makes it one of the highest impact information security laws of the past decade.

Payment Card Industry Data Security Standard

The Payment Card Industry Data Security Standard (PCI DSS) offers an interesting case of industry self-regulation becoming embodied in state law. In 2004, Visa, MasterCard, American Express, and Discover aligned their previously disparate merchant data security requirements into a uniform standard (PCI DSS) that applies to all businesses involved in the storage, processing or transmission of cardholder data. PCI DSS imposes twelve specific requirements (Payment Card Industry, 2005):

- "Install and maintain a firewall configuration to protect cardholder data
- Do not use vendor-supplied defaults for system passwords and other security parameters
- Protect stored cardholder data
- Encrypt transmission of cardholder data across open, public networks
- Use and regularly update anti-virus software
- Develop and maintain security systems and applications
- Restrict access to cardholder data by business need-to-know
- Assign a unique ID to each person with computer access
- Restrict physical access to cardholder data
- Track and monitor all access to network resources and cardholder data
- Regularly test security systems and processes

- Maintain a policy that addresses information security"

While PCI DSS does not carry the force of law in its own right, merchants who agree to accept payment cards are under a contractual obligation to comply with the standard's requirements and face severe fines of up to $250,000 in the event of a compromise.

There is also a recent trend among state legislatures to incorporate PCI DSS requirements into state law. The Minnesota State Legislature recently passed the Plastic Card Security Act making it illegal to store certain elements of cardholder data subsequent to transaction authorization (Plastic Card Security Act, 2007). At the time of this writing, both Texas (HB 3222, 2007) and California (AB 779) had pending legislation that would incorporate additional PCI DSS requirements into that state's body of law.

FUTURE TRENDS

As we have demonstrated, the landscape of federal information security law is quite complex and includes a number of overlapping regulations that apply to different industries, technologies, and constituencies. These laws have evolved significantly over the past three decades and matured in their understanding of the unique technological issues posed by the ubiquity of broadband Internet access. We expect that these laws will continue to evolve as technology does and that the United States will eventually adopt a broad-reaching information security and privacy law, similar to the European Union's Data Privacy Directive of 1995.

We also expect to see an increase in the already heightened public awareness of security and privacy issues. As consumers become more familiar with the present and future laws and regulations protecting the privacy and security of their personal data, we will see an increase in judicial activity on these issues. The courts will be used to test the existing enforcement provisions of information security laws in actions brought by the government and we will also likely see civil liability cases brought by private individuals, individually or as members of a class, seeking damages for negligent activity. This heightened awareness will serve a greater purpose – it will provide the stimulus necessary to bring information security reform to the forefront of industry.

CONCLUSION

It is important to reiterate that this article presents a snapshot in time of a rapidly changing landscape of information security laws. In addition to the laws presented in this section, a variety of other federal laws impact information security decisions. Further, there are a myriad of state, local and international laws that govern information security controls. In addition, private contractual relationships, such as PCI DSS, often impose specific information security requirements on individual business relationships or broad industries.

Information security is a rapidly developing field and the body of law regulating related activities is evolving with each development. The four-tier taxonomy presented in this article serves as a framework for understanding the context of existing and future information security laws.

REFERENCES

AB 779 (2007). *California Assembly Bill 779*, 2007 Legislative Session.

Burke, E. (2001). *The expanding importance of the Computer Fraud and Abuse Act*. Retrieved June 18, 2008, from http://www.gigalaw.com/articles/2001-all/burke-2001-01-all.html

Chawki, M. (2005). A critical look at the regulation of cybercrime. *ICFAI Journal of Cyber Law, 3*(4), 1-55.

Child Online Privacy Protection Act. (1998). 15 USC 6501-6506.

Child Protection and Obscenity Enforcement Act. (1988). 18 USC 2257.

Communications Assistance for Law Enforcement Act. (1994). Public Law 103-414, 108 Stat. 4279.

Computer Fraud and Abuse Act. (1984), 18 USC 1030, as amended 1994, 1996 and 2001.

Copyright Act. (1976), 17 USC.

Digital Millennium Copyright Act. (1998), Public Law 105-304, 112 Stat. 2860.

Economic Espionage Act. (1996), 18 USC 90.

Electronic Communications Privacy Act. (1986), 18 USC 2510-2522 and 18 USC 2701-2711, as amended.

Federal Communications Commission (2004). *In the matter of communications Assistance for Law Enforcement Act and Broadband Access and Services*. Notice of Proposed Rulemaking RM-10865, ET Docket No. 04-295.

Gramm-Leach-Bliley Act. (1999). 15 USC 6801-6809.

Isenberg, D. (2000). *The problems with online privacy laws*. Retrieved June 18, 2008 from http://www.gigalaw.com/articles/2000-all/isenberg-2000-07a-all.html

HB 3222 (2007), *Texas House Bill 3222*, 2007 Legislative Session

Health Insurance Portability and Accountability Act. (1996). Public Law 104-191, 110 Stat. 1936.

Higher Education Coalition (2005). *Comments before the Federal Communications Commission in the matter of Communications Assistance for Law Enforcement Act and Broadband Access and Services.* Retrieved June 18, 2008 from http://www.educause.edu/ir/library/pdf/EPO0536.pdf

Iuliano, R. W. (2003). *Summary of the USA PATRIOT Act and related legislation.* Retrieved June 18, 2008, from http://www.security.harvard.edu/usa_patriot.php

Identity Theft and Assumption Deterrence Act. (1998), Public Law 105-318, 112 Stat. 3007.

Lanham Act. (1946), 15 USC.

Mota, S. A. (2002). The U.S. Supreme Court addresses the Child Pornography Prevention Act and Child Online Protection Act in Ashcroft v. Free Speech Coalition and Ashcroft v. American Civil Liberties Union. *Federal Communications Law Journal, 55*(1), 85-98.

Payment Card Industry (2005). *Payment card industry data security standard*. Retrieved June 18, 2008, from http://www.visa.com/cisp

Patent Act. (1952). 35 USC.

Plastic Card Security Act. (2007). Minnesota Session Laws 2007, Chapter 108, H.F. No. 1758.

Sarbanes Oxley Act. (2002). Public Law 107-204, 116 Stat. 745.

Smith, S. W. (2005). Pretending that systems are secure. *IEEE Security & Privacy, 3*(6), 73-76.

USA PATRIOT Act. (2001). Public Law 107-56, 115 Stat. 272.

Waters, M. & Harrell, J. (1997). *Child pornography on the internet*. Retrieved on June 18, 2008, from http://gsulaw.gsu.edu/lawand/papers/sp97/f

KEY TERMS

Computer-Focused Crime Laws: Involve criminal acts where the use of a computer is intrinsic to the crime.

Computer-Related Crime Laws: Involve the use of a computer but where the criminal activity is not defined by the use of a computer.

Copyright: Protects any original work of authorship from unauthorized duplication or distribution.

Industry-Specific Laws: Protect the confidentiality, integrity and/or availability of personal information maintained by specific industries.

Intellectual Property Laws: Protect the rights of authors, inventors and creators of other intellectual works.

Patents: Grant exclusive use rights to the creators of inventions, processes and designs.

Trademark Law: Protects words, phrases and designs that identify the products or services of a firm.

Trade Secrets: Include any information that derives independent economic value, actual or potential, from not being generally known to, and not being readily ascertainable by proper means by the public and is the subject of reasonable measures to keep such information secret.

Ubiquitous Computing and Communication for Product Monitoring

Rinaldo C. Michelini
University of Genova, Italy

Roberto P. Razzoli
University of Genova, Italy

INTRODUCTION

The present discussion summarises the benefits provided with resort to IT instruments, by dealing with the delivery of *extended* artefacts, under the responsibility of *extended* enterprises. In view to establish the IT environment, one needs to address the market paradigm-shift, from earlier commodity- to mainly utility-based economics, having supply chains concerned by *products-services*, where the latter delivering often outruns the former. Outstripping the pertinent material flows, effectiveness quickly turns on the information flows, supported by networked organisations and cooperative set-ups, with mainly, a twofold outcome: (1) value added in totally new provisions, enhancing the supply effectiveness; (2) value added to the joined information and related transparency of overall environmental impact.

Ambient intelligence is technology-driven opportunity based on the user friendly exploitation of ubiquitous computing and communication (Riva, Vatalaro, Davide, & Alcañiz, 2005; Stephanidis, 2001; VanLoenen, 2003). Turning ambient intelligence toward collaboration activities and eco-compatibility certifying duties could be the winning option to support enterprise competitiveness, privacy protection and eco-system safeguard through cooperative organisations. The involved IT aids basically will move from the existing World Wide Web capabilities, enhancing the *extended* enterprise, with the qualifying functions of service engineering, and fitting out the on-duty incumbents by users' adaptive interfaces.

BACKGROUND

The IT options grant new prospects, as for manufacturing and market organisation, leading to new traded items, *products-services* or *extended* artefacts, by means of new industrial set-ups, the *networked* facilities. Indeed, the recalled concepts lead to address the ambient intelligence and the supporting IT processing Web options for enterprise cooperation and business deployment according to an innovative scenario to grant competitive advantage of richer or enhanced delivery with lifecycle transparent eco-conservativeness. This scenario corresponds to a shared vision aiming at development sustainability based on key aspects (Figure 1) where product on-duty properties and enterprise point-of-service responsibility are transparently reported, assuring the eco-impact data management under third-party certification.

The information set-up consistent with the sketched scenario faces two conditioning lines:

- technical feasibility incumbents, which can be dealt with by suitably implemented IC innovative aids;

Figure 1. Key aspects of the information frame for sustainability

• ***extended enterprise co-operative environment***, the net concerns operate with unified responsibilities, under headquarters ruling the traded provisions on their lifelong span;
• ***product-service unified data-frame***, the delivery of ***extended*** artifacts is primary achievement, and on-duty visibility is basic knowledge, to keep conformance-to-specification levels;
• ***total connectedness***, all authorized stakeholders are linked by communication infrastructures that deliver the right data at the right time, according to their permit and priority labels;
• ***supply-chain transparent reporting***, business productivity gives account, out of finance and lobar factors, also technology and natural resources (e.g., by the ***KILT*** model);
• ***eco-impact data management***, the entropy trend is monitored on the ***extended*** artefacts span, and assessed by acknowledged standards (e.g., the ***TYPUS*** metrics);
• ***third-party certification data-vaulting***, accredited bodies oversee downgrading on the ***extended*** artefacts, lifecycle, and charge consumers for the net natural capital depletion.

- politico-legal and socio-economical hindrances, which will evolve with sustainability consciousness.

Along the first line, the technical literature (Abowd, 2004; Ailisto, Kotila, & Strömmer, 2003; Ameri &Dutta,2004; Dekker & Scarf, 1998; Garetti, 2004; Michelini &Kovacs, 2005), deals with the networked infrastructure technology, namely, the IT aids that need be added to the supply chain for lifecycle management. These are enabling support of product data visibility and eco-consistency assessment, once politico-legal and socio-economical pertinent rules are established. The coherent description of the product impact, over the operation horizon, including reverse logistics, needs address the consumable decay, explicitly making account of involved natural resources.

The approach leads, for instance, to the *KILT* model (Michelini, Acaccia, Callegari, Molfino, & Razzoli, 1999; Michelini & Razzoli, 2004a, 2004b) linking the delivery, ***Q***, of the manufacture activity to the four inputs corresponding to all contributed capitals, say:

- ***K***, *know-how* or *technical capital*, the knowledge and technology exploited in manufacturing,
- ***I***, *invested financial capital*, traditional driving input of earlier industrial economical set-ups,
- ***L***, *directly engaged labour*, conventional work-force counterpart of industrial organisations,
- ***T***, *natural capital*, actually consumed tangibles to support the whole actual supply chain.

Return on invested capitals is built on all factors, and fair competition requires *equal opportunity* players, compelled:

- to bring out the dependence on tangible resources consumption ***T***, when pricing items;
- to equitably remunerate the direct labour ***L***, along the product-service supply-chain, dismissal included;
- to repay the fixed assets ***I***, for the share- and stake-holders profits;
- to exploit the underlying knowledge ***K***, both enterprise solid practice or non-proprietary technologies.

COOPERATIVE ORGANISATION FOR LIFE-LONG SERVICE

The earlier outlined scenario is consistent with new supply chains, delivering *products-services* supported by cooperating organisations. The changes open new business paradigms, based on product lifecycle management (Ameri & Dutta, 2004) and service engineering, embedding high-intensity information flow with intangibles value added and enhanced transparency of natural capital exploitation. These paradigms (Figure 2) encompass three layers: the business *networked concern*, the *extended* artefacts delivery layout, and the *certified visibility* set-up; the IT tools differ, as the horizons broaden, to include clients and controllers.

The manufacturer business concern has to deal with all layers, especially today, as the supplier responsibility expands to cover lifecycle conformance-to-specification prerequisites and dismissal requirements out of the *point-of-sale*. The *point-of-service* tasks address items operation properties on two facts: on-duty reliability for users' satisfaction and eco-impact control for environmental protection. This leads to widely expanding the domain of intervention of existing corporations with new tasks out of traditional workshops based on competencies up until new not dealt with, and mostly covered by providers timely taken in by users after the *point-of-sale*. The social interest of third parties, in environment protection and natural capital preservation, is new fact, entitling governmental regulations explicitly involving who conceives and brings out the traded goods. Then, the efficient answer brings to *extended* enterprises with the new business paradigms of *product-service* delivering.

Thereafter, the *service engineering* (SE) will appear as challenging duty, linked to the design steps by the *product-lifecycle-management* (PLM) for accessing the technical sheets for *point-of-service* and *point-of-dismissal* tasks.

Figure 2. Collaborative networks for lifecycle visibility

- The networked organization, granting the information service for customers, is required:
  - # to provide collaborative forms and behaviors for product life-cycle management;
  - # to rule conformance assessment and restoration within networked responsible bodies.
- The *extended* enterprise profits of a networked organization to expand buyers satisfaction:
  - co-operative design and shared knowledge make multi-technology *extended* artifacts possible;
  - lifecycle data pricing becomes relevant and *new* competition feature between companies.
- The sustainability assures *fair* trade conditions, provided that networks are available, in order:
  - to employ objective, world wide referenced, metrological standards (e.g., *TYPUS* metrics);
  - to record the artifact lifecycle behavior, controlled by independent certifying bodies.

Figure 3. Crafty techniques to improve the eco-conservativeness

Expansion of the artifacts on-duty availability, by *monitoring* maintenance options, through:
- *reactive mode*: at the failure diagnosis, the process self-ruling is started, enabling the up-keeping and repairing schedules;
- *proactive mode*: at the deviation from nominal conditions, the procedures for process regulation and duty fixing are accomplished.

Protraction of the artifacts operation life, after renovation or revision deeds:
- to re-vamp: the recovery of consistent functional state, by proper renovation and updating of every critical modules, according to enacted bylaws;
- to re-integrate: the restoration to the original state, by full revision and (possible) replacement of every defective modules.

SE distinguishes from PLM when the service providers are independent partners, and the supply refers to shared responsibilities; the two practices bear similar database, conception, planning, and implementation with emphasis on applied maintenance and restoring options (Figure 3) to achieve function reliability while improving eco-conservativeness. The present discussion is specifically turned on the IT aids, as these are enabling technologies of the prospected business paradigms shif, by means of innovative options such as ambient intelligence.

The ambient intelligence (AmI) shows how IT tools can be incorporated as invisible witness of people life, so that social interaction and product functionality will move foreground for further use. The AmI concept leads to cooperative settings, where entities communicate to each other. These entities can be humans, artificial agents, Web/grid services, virtual-objects representing real things (not only human beings), descriptions of human knowledge (knowledge-based systems), and so forth; the whole layout (Figure 4) leverages the full potentiality of net-centric facilities, for creativity improvement boosting innovative duties. The options support people to experience interaction with human and artificial agents in their working environments; they are basic help to grant lifecycle transparency by SE tasks routines for eco-conservativeness duties. By *ubiquitous computing*, the effective, pleasant, and unobtrusive presence of computing devices everywhere is supplied; by *ubiquitous communication*, the everywhere access to network and computing facilities is provided; by *intelligent user adaptive interfaces*, the perception of remaining in a natural world is preserved as the facility automatically adapts to the human preferences. The options are important when trying to understand the implications that AmI has on the world we live in, notably due to the ability of giving visibility of our current behavior.

The AmI potentials are in the invisible support of flexible and natural communication with other users or computers providing input and perceiving feedback by utilising indifferently all senses and communication channels. The *ubiquitous computing* aims at invisibly and unobtrusively means to free people from tedious routine jobs. In its final form, the concept leads to a computing device, which, when moving with the user, incrementally builds dynamic models of the changing environment and configures its services accordingly;

Figure 4. Ambient intelligence for real life monitoring

Ambient intelligence, AmI: convergence of ubiquitous computing and ubiquitous communication, with duty-driven interfaces, adapting to the users. Three facts deserve explanations:
ambient - the concept assumes the ability of *existing* or *be present on all sides*;
ubiquitous - it assumes that something exists *everywhere*, *at the same time*, *on a steady level*;
natural perception - it involves self-adaptation, to match *real-life conditions*.

The *ambient intelligence*, AmI, incorporates properties of:
• mobile (nomadic) computing, by multiple-function devices and remote interaction capabilities;
• distributed interactivity, to enable communication by invisible processing computer resources;
• self-adaptive multi-mode configuration, to deal with the current human behavioral habits.

Ubiquitous computing is roughly the opposite of virtual reality. Where virtual reality puts people inside a computer-generated world, ubiquitous computing forces the equipment to operate in the real world, automatically adapting with the people preferences.

it is able either to remember past patterns or proactively build up new ones. The *ubiquitous communication* is major change, promoting data transfer and allowing to integrate new modules like sensors or diagnosis modules by natural procedures. Distributed agents and wireless networking are enabling technologies by sensing-driven modules grouped into proper categories, say: visual recognition and output (e.g. face 3D gesture/location); sound-recognition and -output (e.g., speech, melody); scent recognition and output; tactile-recognition and output; other sensor technologies. By *profiling*, the ambient has the ability to personalise and automatically adapt itself, to any particular user behavioral patterns. Major importance is given to *natural-feeling* human and to multi-mode interfaces. Humans speak, gesture, touch, sense, and write in their interactions with other humans and with the physical world. The idea is that these *natural* actions are used as explicit or implicit input to AmI systems. The AmI services split on two-fold incumbents:

- to support restoring and maintenance duties for clients' satisfaction of the delivered *product-service*;
- to help data collection and vaulting for tangibles consumption assessment and certification.

All technical sides of the collaborative networked organisations are met by ambient intelligence. The links binding the three-parties net-concern where the eco-protection is dealt with (Figure 5) help explaining in which frames the ambient intelligence will represent the enabling innovation. Indeed, AmI provides the effective work-environment to maintenance technicians (Arnaiz, Arana, Maurtua, & Susperregi, 2004; Dekker & Scarf, 1998; Gross & Fleisch, 2004), giving access to ubiquitous and up-to-date views of the ongoing operations wherever the equipment or the operator is (enabling remote maintenance and lifecycle management) by user friendly and intelligent interfaces (running context-aware data acquisition).

When looking at *product-service* monitoring on the life-long span, one realises that implementing the AmI concepts will drastically change the SE environments. The *ubiquitous* property integrates into a unified scheme the different processes combining (human and virtual) entities leading to a product-oriented dynamic infrastructure. Benefits are: concentrating information on delivery; creating service-oriented bent for all steps in the lifecycle; performing supply chain information updates in real-time and in intelligent ways; promoting information sharing with easier access and management of operation data; enabling products to carry and to process information which affect their destiny; and the likes. The resort to full AmI vision will require developing large, sophisticated, heterogeneous, distributed systems built on flexible platforms capable of providing seamless networking to support the provision of value-added services to industry, individuals, and administrations. The resulting layouts covering several interacting embedded components will need to be ubiquitous, self-configuring, self-healing, self-protective, and self-managed; it will lead to appearance of new type of services: the AmI services.

Up until now, the *extended enterprise* concept has mostly been explored to foster a business architecture, and the inclusion of the *extended artefact* lifecycle service is considered as wily advantage in front of competitors. If manufacturers' responsibility would expand at the *point-of-service*, and, further, at the *point-of-dismissal*, the new scenario establishes requiring totally different organisations: the goods conformance assessment, out-of-client satisfaction, becomes legal prerequisite with connected prescriptions for the environment protection. The involvement of third party certifying bodies (Figure 6) is an obvious issue requiring proper upgrading of the networked infrastructure to grant the different, specialised accesses. Of course, the eco-conservativeness incumbents are even more important. To that purpose, the networked infrastructure shall assure varying topology link

Figure 5. The collaborative networks for sustainability

The collaborative infrastructure requires the interlinked participation of:

- **purveyors**, covering the entire *supply-chain*: materials provision, items manufacture, life-cycle upkeeping, backward recovery; the ecological responsibility is dealt with by clustering several firms within a factural alliance of co-operating multi-sectional interests businesses;
- **users**, purchasing *extended* artifacts (*products-services*), to profit of the delivered functions with reliability figure close to one; the payments shall include conformance certification at the point of service, after tax collection agaisnt tangibles depletion;
- **supervisors**, assuring *third party* incumbents for (today and tomorrow) environment and society protection; the certifying bodies report to govenmental authorities and use objective standards,

having access to the *extended* artifacts life-cycle databases.

The transparency of the environment impact is achieved by continuous monitoring and recording the actual running conditions, both, of the *forward* and the *backward* cycle.

The governmental agencies collect the charges for the net consumed tanbibles, following assessments of the authorized certifying bodies.

Figure 6. The third party certifying service

The third party certifying body operates:
• interacting with the collaborative network support, established by the extended enterprise to manage the extended artifact administration, with full visibility on the lifecycle data and proper security restrictions;
• verifying the monitoring and control network integrity, and providing conformance assessment records by progression statements, as partial estimates for T factor refunds, based on the TYPUS metrics;
• overseeing the extended enterprise accomplishments in terms of eco-charges payments, accounting progress balances, based on unified responsibility established by the two consumers parties indentation;
• guaranteeing data vault and privacy protection, with proper specifications in the case the certifying duty is transferred to a different body and/or the joint consumers parties modify the binding contracts;
• operating within an accreditation scheme, notified to national (government) authorities and international organizations, having worldwide acknowledgment.

The involvement of third party certifying bodies needs, of course, proper regulations, enacted by the national authorities, but suitably harmonized to assure worldwide equivalence.

between users and certifying bodies by friendly layouts not requiring on-process experts. Now, AmI provides all useful prerequisites: intelligent resource management; distributed knowledge with multi-layer computing; hierarchical access with priority assignment; automatic *profiling* and interface adaptation; no communication/computation programming requests, and so forth. Thus, we need to address:

- the management of the *extended* artefacts assuring effective servicing by *ambient intelligence*; and
- the full enabling of the *extended* enterprise with integration of collaborative communication.

The two facts need to be analysed on their technological feasibility and on their eco-benefits. The potentialities of any innovation are not fully exploited unless the enabling surrounding appears as the driving requirement supported by IT tools. In these conditions, the technology-driven options, such as AmI, become standard requisites for service-engineering, assuring effectiveness to *extended* artefacts provision, and powerful support for the supply chain monitoring under unified responsibilities with the specific charges on assessing the net resources consumption and environment impact.

FUTURE TRENDS

On longer terms, new means to create tangible resources to need be envisaged joining to better exploitation of material processes (the reverse logistics of *backward* cycle factories), possibly novel (***K***-driven) *technical* capital. In fact, the *knowledge* society by itself does not modify the entropy laws of the physical world; it simply widens the value-chain patterns adding high-intensity information. On earth, bio-processes only seem to violate this decay, and such technologies could promote regenerative trends with restoration of original assets and remediation of downgraded sites (by bio-mimetic industries). The scenario is taken with caution, as bio-manipulation is associated to Frankenstein's myth and the ubiquitous knowledge society with full monitoring of on-progress developments is a reasonable option to relax anxieties.

CONCLUSION

The turning of *ambient intelligence* toward collaboration duties and eco-compatibility certifying activities is explored as basic option, to join enterprise competitiveness, privacy protection, and eco-system safeguard through ubiquitous computing and communication. The facilities supported by collaborative infrastructures are analysed with focus on the transparency of consumable decay (by appropriate metrics) over the supply chain lifecycle horizon, dismissal, and reverse-logistics included (Dekker, Fleischmann, Inderfurth, & Van Wassenhove, 2004). The discussed IT options are consistent with the scenario describing the manufacturing activity by the *KILT* model; they grant novel paradigms as for business deployment and enterprise organisation leading to new traded items, *products-services* (or *extended* artefacts), by means of new industrial facilities, the *networked* companies; they assure a very effective surroundings to enable quantitative assessments aimed at monitoring and recording tangibles net yield along any material supply chains.

The presentation equivalently addresses *point-of-service* and *point-of-dismissal* incumbents, as both are considered for their falls-off in the material flows, when environment impact is dealt with. The IT networked options are developed

to fully exploit enterprise cooperation and to foster business deployment by means of the competitive advantage of *products-services* delivery with lifecycle transparent eco-consistency. When *point-of-service* incumbents are dealt with, useful practices are found by enhancing reliability by condition monitoring maintenance (Michelini, Crenna, & Rossi, 2001) and by expanding on-duty life by re-vamping and re-integration; when *point-of-dismissal* incumbents are considered, the very intriguing aspects of reverse logistics appear, which require deep understanding of the overall processes to weigh the ecological-return conservativeness, even when the economical-return is granted due to the enacted by-laws. In such context, the pervasive visibility on the supply chain progression provided by *ambient intelligence* is a powerful instrument to help selecting the legal setting for worldwide effectiveness.

ACKNOWLEDGMENT

The article has, mainly, been based on the chapter contributed by the two co-authors (Putnik & Cunha, 2005):

Michelini, R.C., & Razzoli, R.P. (2005). Collaborative networked organisations for eco-consistent supply-chains. In G.D.Putnik & M.M.Cunha (Eds.), *Virtual enterprise integration: Technological and organisational perspectives* (pp. 45-75). Hershey, PA: IRM Press.

REFERENCES

Abowd, G.D. (2004). *Investigating research issues in ubiquitous computing: The capture, integration, and access problem.* Retrieved from http://www.cc.gatech.edu/fce/c2000/pubs/nsf97/summary.html

Ailisto, H., Kotila, A., & Strömmer, E. (2003). *Ubicom applications and technologies*. Presentation at ITEA 2003, Oulu, Finland (pp. 1-21). Retrieved from http://www.vtt.fi/ict/publications/ailisto_et_al_030821.pdf

Ameri, F., & Dutta, D. (2004). Product lifecycle management needs: Concepts and components. Technical Report (pp.1-3). Retrieved from http://plm.engin.umich.edu/PLMDC-TR3-2004.pdf

Arnaiz, A., Arana, R., Maurtua, I., & Susperregi, L. (2004, May 17-19). Maintenance: Future technologies. *Intl. IMS Forum 2004: Global Challenges in Manufacturing*, Villa-Erba Cernobbio, Italy (Vol. 1, pp. 300-307).

Dekker, R., Fleischmann, M., Inderfurth, K., & vanWassenhove, L.N. (2004). *Reverse logistics: Quantitative models for closed-loop chains* (pp. viii - 436). Berlin: Springer Verlag.

Dekker, R., & Scarf, P. (1998). On the impact of optimising models in maintenance decision making: A state of the art. *Reliability Engineering and System Safety,* (60), 111-119.

Garetti, M. (2004, May 17-19). PLM: A new business model to foster product innovation. *Intl. IMS Forum: Global Challenges in Manufacturing,* Villa-Erba Cernobbio, Italy (Vol. 2, pp. 917-924).

Gross, S., & Fleisch, E. (2004, April 5-7). Maintenance improvement by unique product information enabled by ubiquitous computing. *The 11th Intl. IFAC Symp. Information Control Problems in Manufacturing*, INCOM 2004, Salvador, Brazil (pp. 65-70).

Michelini, R.C., Acaccia, G.M., Callegari, M., Molfino, R.M., & Razzoli, R.P. (1999). Artefact integration by concurrent enterprises and productive break-up. In G.Jacucci, G.J.Olling, K.Preiss, & M.Wozny (Eds.), *Globalisation of manufacturing in the digital communication era* (pp. 221-234). Boston: Kluwer Academic.

Michelini, R.C., Crenna, F., & Rossi, G.B. (2001). Diagnostics for monitoring maintenance and quality manufacturing. In C.T. Leondes (Ed.), *Computer aided design and manufacturing: Techniques and applications* (Vol. I, pp. 5.01-5.56). Boca Raton, FL: CRC Press.

Michelini, R.C., & Razzoli, R.P. (2004a). Product-service eco-design: knowledge-based infrastructures. *Intl. J. Cleaner Production, 12*(4), 415-428.

Michelini, R.C., & Razzoli, R.P. (2004b). Product-service for environmental safeguard: A metric to sustainability. *Intl. J. Resources, Conservation and Recycling, 42*(1), 83-98.

Michelini, R.C., & Kovacs, G.L. (2005). Information infrastructures and sustainability. In L. Camarinha Matos (Ed.), *Emerging solutions for future manufacturing systems* (pp.347-356). Springer.

Putnik, G.D., & Cunha, M.M. (Eds.). (2005). *Virtual enterprise integration: Technological and organisational perspectives* (pp. 1-300). Hershey, PA: IRM Press.

Riva, G., Vatalaro, F., Davide, F., & Alcañiz, M. (2005). Ambient intelligence: the evolution of technology. *Communication and cognition towards the future of human-computer interaction* (pp. 1-320). IOS Press.

Stephanidis, C. (2001). Ambient intelligence in the context of universal access. *ERCIM News,* (47), 10-11.

VanLoenen, E.J. (2003). *Ambient intelligence: Philips' vision.* Presentation at ITEA 2003, Oulu, Finland. Retrieved from http://www.vtt.fi/ele/new/ambience/evert_van_loenen.ppt

KEY TERMS

Ambient Intelligence (AmI): is the convergence of ubiquitous computing, ubiquitous communication, and interfaces adapting to the user; it relies on provisioning *ubiquitous* computing (i.e., useful and unobtrusive presence of computing devices everywhere).

Condition Monitoring Maintenance (CMM): The maintenance intervention applies once the product state is recognised; it needs support of special diagnostics and prognosis tools and sophisticated knowledge-based systems.

Extended Artefact or Product-Service: Any supply joining manufactured commodities and enabling utilities. *Artefact*: any object made by man, especially with a view of subsequent use; something made with skill.

Interoperability: The ability of two or more systems, subsystems, products, or applications to work together and/or share information or inputs and outputs.

Lifecycle: The collective set of phases a product or system may go through during its lifetime (e.g., concept definition, development, production, operation and support, dismissal, decommissioning, and disposal).

Natural Language: Ordinary human language; unlike precisely defined computer languages, it is often ambiguous and is thus interpreted differently by different hearers.

Net Concern or Extended Enterprise: A group of companies that work together and act as a single business entity to satisfy a particular set of customer needs.

Product Lifecycle Management, PLM: The process of, or a system for, managing all data about a product as it moves through the all lifecycle, from materials provision, to on-duty requirements and dismissal.

Service Engineering, SE: The activity that deals with improving the design process of the service, supplied with the extended artefacts, developing and implementing the duties, which assure maintenance, restoring and conformance assessment on the lifecycle.

Underwater Wireless Networking Techniques

Manuel Perez Malumbres
Miguel Hernandez University, Spain

Pedro Pablo Garrido
Miguel Hernandez University, Spain

Carlos Tavares Calafate
Technical University of Valencia, Spain

Jose Oliver Gil
Technical University of Valencia, Spain

INTRODUCTION

Underwater sound has probably been used by marine specimens for millions of years as a communication capability among the members of a same species. It is said that in 1490, Leonardo Da Vinci wrote the following sentence: "If you cause your ship to stop and place the head of a long tube in the water and place the outer extremity to your ear, you will hear ships at a great distance from you" (Urick, 1983); being perhaps the first recorded experiments about hearing underwater sounds.

In 1826 on Lake Geneva, Switzerland, the physicist Jean-Daniel Colladon, and his mathematician friend Charles-Francois Sturm, made the first recorded attempt to determine the speed of sound in water. In their experiment, the underwater bell was struck simultaneously with ignition of gunpowder on the first boat. The sound of the bell and flash from the gunpowder were observed 10-miles away on the second boat. The time between the gunpowder flash and the sound reaching the second boat was used to calculate the speed of sound in water. Colladon and Sturm were able to determine the speed of sound in water fairly accurately with this method. (Colladon, 1893).

This experiment on sound propagation through water laid the foundation for underwater acoustic technology, which paved the way for the development of this technology up to our days. In 1906, Lewis Nixon invented the very first sonar-type listening device, increasing the demand of this technology during World War I to detect submarines. In 1915, the physicist Paul Langévin and the engineer Constantine Chilowski, invented the first sonar-type device for detecting submarines, called an "echo location to detect submarines," using the piezoelectric properties of the quartz. He was too late to offer any help to the war effort; however, Langévin's work heavily influenced future sonar designs.

After using underwater sound technology for measuring the proximity to the shore and other ships, researchers soon realized that, if the sound device was pointed down at the seafloor, the depth could be accurately determined. So, new applications of sonar devices were discovered, like active depth measuring (bathymetry), seafloor shape registering, search for geological resources (i.e., oil, gas, etc.), detecting and tracking fish banks, submarine archaeology, and so forth.

Although the underwater acoustic applications were mainly focused in ranging applications, exploration of seafloor and fishery by means of sonar devices, the interest in underwater multipoint communications was stressed in the 1990's, where synoptic, spatially sampled oceanographic surveillance has provided an impetus to the transfer of networked communication technology to the underwater environment. One of the former deployments was the autonomous oceanographic surveillance network (AOSN), supported by the US Office of Naval Research (ONR) (Curtin, Bellingham, Catipovic, & Webb, 1993). It calls for a system of moorings, surface buoys, underwater sensor nodes, and autonomous underwater vehicles (AUVs) to coordinate their sampling via an acoustic telemetry network.

BACKGROUND

Wireless networking technologies have experienced a considerable development in the last 15 years, not only in the standardization areas, but also in the market deployment of a bunch of devices, services, and applications. Among this plethora of wireless products, wireless sensor networks are exhibiting an incredible boom, being one of the technological areas with greater scientific and industrial development pace (Akyildiz, Sankarasubramaniam, & Cayirci, 2002). The interest and opportunity in working on wireless sensor network technologies is endorsed by (a) technological indicators like the ones published by MIT (Massachusetts Institute of Technology) in 2003 (van der Werff, 2003),

where wireless sensor network technology was defined as one of the 10 technologies that will change the world, and (b) economic and market forecasts published by different economic magazines like (Rosenbush, Crockett, & Yang, 2004), where investment in wireless sensor network (WSN) ZigBee technology was estimated over 3.500 nillion dollars during 2007.

Recently, wireless sensor networks have been proposed for their deployment in underwater environments, where a lot of applications like aquiculture, pollution monitoring, offshore exploration, and so forth, would benefit from this technology (Cui, Kong, Gerla, & Zhou, 2006).

Despite having a very similar functionality, underwater wireless sensor networks (UWSNs) exhibit several architectural differences, with respect to the terrestrial ones, that are mainly due to the transmission medium characteristics (sea water) and the signal employed to transmit data (acoustic ultrasound signals) (Akyildiz, Pompili, & Melodia, 2006). Then, the design of appropriate network architecture for UWSNs is seriously hardened by the conditions of the communication system and, as a consequence, what is valid for terrestrial WSNs is perhaps not valid for UWSNs. So, a general review of the overall network architecture is required in order to supply an appropriate network service for the demanding applications in such an unfriendly submarine communication environment.

Major challenges in the design of underwater acoustic networks are:

- Battery power is limited and usually batteries can not be recharged because solar energy cannot be exploited;
- The available bandwidth is severely limited;
- The channel suffers from long and variable propagation delays, multipath and fading problems;
- Bit error rates are typically very high;
- Underwater sensors are prone to frequent failures because of fouling, corrosion, and so forth.

In the next section, we discuss the main issues in the design of efficient underwater wireless sensor networks. Following a bottom-to-top approach, we will review the network architecture, highlighting some critical design parameters at each of the different network layers, and how to overcome the limitations and problems introduced by UWSN environments.

UNDERWATER WIRELESS NETWORKING TECHNOLOGIES

Basically, a UWSN is formed by the cooperation among several nodes that establish and maintain a network through the use of bidirectional acoustic links. Every node is able to send/receive messages from/to other nodes in the network, and also to forward messages to remote destinations in case of multihop networks. Every node may have one or several sensors that are actively recording environmental data that should be forwarded to special sink nodes, typically platforms or buoys at the surface. Sink nodes have communication channels to forward and/or locally store the collected data to the remote control station in the coast, typically through a radio frequency (RF) link.

So, the UWSN allows an interactive environment where scientists can extract real-time data from multiple distant underwater sensor instruments. After evaluating the obtained data, control messages can be sent to individual network nodes so the overall network can be adapted to changing situations.

Topology

In Partan, Kurose, and Levine (2006), taxonomy of UWSN regimes is proposed. They classify different UWSNs in terms of both spatial coverage and node density. For every kind of network topology, different architectural approaches have to be considered in order to improve the network performance (throughput, delay, power consumption, packet loss, etc.). So, it is important to design the network architecture taking into account the intended network topology.

Physical Layer: Acoustic Link

The most common way to send data in underwater environments is by means of acoustic signals, just like dolphins and whales use to do for communicating between them. Radio frequency signals have serious problems to propagate in sea water, as shown in Schill, Zimmer, and Trumpf (2004), being operative for radio-frequency only at very short ranges (up to 10 meters) and with low-bandwidth modems (tens of Kbps).When using optical signals, the light is strongly scattered and absorbed underwater, so only in very clear water conditions (often very deep) does the range go up to 100 meters with high bandwidth modems (several Mbps) and blue-green wavelengths.

Since acoustic signals are mainly used in UWSNs, it is necessary to take into account the main aspects involved in the propagation of acoustic signals in underwater environments, including (1) the propagation speed of sound underwater is around 1,500 m/s (5 orders of magnitude slower than the speed of light), and so the communication links will suffer from large and variable propagation delays and relatively large motion-induced Doppler effects; (2) phase and magnitude fluctuations lead to higher bit error rates compared with radio channels' behaviour, being mandatory the use of forward error correction codes (FEC); (3) as frequency increases, the attenuation observed in the acoustic channel

also increases, this being a serious bandwidth constraint; (4) multipath interference in underwater acoustic communications is severe due mainly to the surface waves or vessel activity, being a serious problem to attain good bandwidth efficiency. Several approaches were taken to combat multipath effects, being the use of multiple-input multiple-output (MIMO) transducer arrays (Freitag, Stojanovic, Singh, & Johnson, 2001), a good solution for reducing multipath effects and therefore increasing the link throughput.

Several works in the literature propose models for an acoustic underwater link, taking into account environmental parameters, such as salinity degree, temperature, depth, environmental interference, and so forth. In Harris and Zorzi (2007), you will find a clear description of the different issues that take part in the development of an acoustic channel model.

Medium Access Control (MAC) Layer

The main task of MAC protocols is to provide efficient and reliable access to the shared physical medium in terms of throughput, delay, error rates, and energy consumption. However, due to the different nature of the underwater environment, there are several drawbacks with respect to the suitability of the existing terrestrial MAC solutions for the underwater environment. In fact, channel access control in UWSNs poses additional challenges, due to the aforementioned peculiarities of underwater channels.

As shown in Akyildiz et al. (2006), existing MAC solutions are mainly focused on carrier sense multiple access (CSMA) or code division multiple access (CDMA). However, frequency division multiple access (FDMA) is not suitable for UWSNs due to the narrow bandwidth available in underwater acoustic channels, and the vulnerability of limited band systems to fading and multipath effects. Moreover, time division multiple access (TDMA) shows limited bandwidth efficiency because of the long time guards required in the underwater acoustic channel. Furthermore, the variable delay makes it very challenging to achieve a precise synchronization through a common timing reference.

CSMA Based

In general, CSMA-based protocols are vulnerable to both hidden and exposed terminal problems (Karl & Willig, 2005). In order to reduce the effects of hidden terminals, MAC proposals should include techniques similar to the ones used in terrestrial networks like MACA (Karn, 1990), which uses RTS/CTS/DATA packets to reduce the hidden terminal problem, and MACAW (Bharghavan, Demers, Shenker, & Zhang, 1994), which adds to the previous one an ACK packet at the link-layer, which can be helpful in an unreliable underwater channel. FAMA (Fullmer & Luna-Acebes, 1995) extends the duration of RTS and CTS packets in order to avoid data packet collisions, and so, contention is managed at both sender and receiver sides before sending data packets. The efficiency of these protocols is heavily impacted by propagation delays due to their multiple handshakes.

A number of adaptations have been proposed to adopt MACA, MACAW, and FAMA for underwater networks. In Molins & Stojanovic (2006), there was proposed the slotted FAMA approach, adding timeslots to the FAMA protocol to limit the impact of propagation delays. Kebkal, Kebkal, and Komar (2005) proposed a means to reduce the impact of propagation delay on FAMA- and MACAW-based protocols, with ACK and DATA packets simultaneously in flight. They also suggest an extension to FAMA, using CDMA for the RTS packets, to develop a collision-free FAMA protocol.

CDMA Based

CDMA is a contention-free multiple-access method that is promising for future underwater networks. In fact, CDMA is robust to frequency selective fading caused by multipath, since it is able to distinguish among signals simultaneously transmitted by multiple devices through codes that spread the user signal over the entire available band.

In Freitag et al. (2001), two code-division spread-spectrum physical-layer techniques for underwater communications in shallow water are compared, namely direct sequence spread spectrum (DSSS) and frequency hopping spread spectrum (FHSS). In the case of DSSS, limitations in temporal coherence of the channel affect the maximum spreading factor, leading to situations that may be better suited to FHSS signals. Conversely, the multipath resolving properties of DSSS minimize the effects of frequency-selective fading that degrade the performance of FSK modulation in FHSS systems.

More recently, Stojanovic and Freigtag (2006) reported very promising CDMA experimental results. An important caveat for this work, however, is that the received power for each of the users should be similar. If the received power for all users is not roughly similar, signals from distant users cannot be received successfully. This is known as the near-far problem, and requires the transmit power of each user to be controlled when switching the channel.

In Pompili, Melodia, and Akyildiz (2007) the authors propose a distributed medium access control (MAC) protocol called UW-MAC. It defines a transmitter-based CDMA scheme that incorporates a novel closed-loop distributed algorithm to set the optimal transmit power and code length to minimize the near-far effect. It compensates for the effect of multipath by exploiting the time diversity in the underwater channel, thus, achieving high channel reuse and a low number of packet retransmissions.

Network Layer

This layer is mainly responsible of routing packets to the proper destinations. So, a routing protocol is required when a packet must go through several hops to reach its destination. It is responsible for finding a route for the packet and making sure it is forwarded through the appropriate path. The way paths are selected for every source-destination pair will have a direct impact on the overall network performance.

Most of the routing proposals for UWSN are based on the ones developed for terrestrial ad hoc and wireless sensor networks. An overview of the different approaches proposed in the literature can be found in Abolhasan, Wysocky, and Dutkiewicz (2004).

Nevertheless, as stated by Akyildiz et al. (2006), proactive and reactive routing protocols are not suitable for UWSNs. The former introduce large-signalling packet exchanges every time the network changes, so that each network node knows the path to the rest of nodes. Concerning the latter, reactive routing protocols, these require a source-initiated flooding of control packets to establish the path(s), and so the latency to establish the path is usually high, being further amplified in the underwater environment due to the slow propagation of acoustic signals.

Therefore, it seems that geographical routing protocols are the most promising approaches for their use in UWSNs. However, the GPS (global positioning system) radio receivers (around 1.5 GHz band) do not work properly in underwater environments, as explained before. There are some works related to underwater localization, also required for AUV navigation systems, and the need of mapping sensor data with spatial localization becomes evident, representing an open problem that requires further research. So, in the following, we will briefly describe several examples of routing protocols especially devoted to underwater sensor network scenarios. Most of the proposals take into account the energy consumption constraint.

In Xie and Gibson (2001), a routing protocol for UWSNs is proposed, being able to initialize the network topology and work in a centralized manner at the surface station (typically the sink node). The paths to the sink are established by the central manager avoiding congestion, introducing quality of service (QoS) support, and managing the overall network energy consumption; it operates in a similar fashion to the point coordination function (PCF) of IEEE 802.11 networks.

Another proposal can be found in Xie, Cui, and Lao (2006), where a routing protocol called vector-based forwarding (VBF) is described. With VBF, each packet carries the positions of the sender, the destination, and the forwarder. Packets are forwarded along redundant and interleaved paths (routing pipes) from a source to a destination node, being robust against packet loss and node failure. Jointly with the routing strategy, a localized and distributed self-adaptation algorithm is proposed to enhance the performance of VBF. This algorithm allows the nodes to weigh the benefit of forwarding packets, reducing energy consumption by discarding low-benefit packets.

The solution proposed in Pompili et al. (Pompili, Melodia, & Akyildiz, 2006) relies on the use of virtual circuits that are established a priori between each source and sink. So, every packet associated with a particular connection follows the same path. This requires centralized coordination from a sink node (station usually located at the surface), and leads to a less flexible routing architecture. However, as other centralized proposals, it is able to exploit optimization tools in order to achieve optimal network layer performance with minimum signalling overhead. In order to increase the reliability of network due to potential node failures, the algorithm finds two paths, primary and backup virtual circuits, between every source-destination pair of nodes.

APPLICATIONS

As mentioned in the introduction, underwater sensor networks can enable a broad range of applications, following the same path as the terrestrial sensor networks. Therefore, we can use UWSN technology for

- *Ocean dampling networks*. Networks of sensors and AUVs with the ability to perform synoptic, cooperative adaptive sampling of the 3-D coastal ocean environment in order to build geology information databases.
- *Environmental monitoring*. UWSNs can perform pollution monitoring (chemical, biological, and nuclear), monitoring of ocean currents and winds, improved weather forecast, detecting climate changes, understanding and predicting the effect of human activities on marine ecosystems, and biological monitoring, such as tracking of marine biology activity or aquaculture industry.
- *Undersea explorations*. Underwater sensor networks can help at detecting underwater oilfields or reservoirs, determining routes for laying undersea cables, and assisting in the exploration for valuable minerals. Also, it can be used to find out wrecks, or as an invaluable tool for submarine archaeology.
- *Disaster prevention*. Sensor networks that measure seismic activity from remote locations can provide tsunami warnings to coastal areas, or study the effects of submarine earthquakes.
- *Assisted navigation*. Sensors can be used to identify hazards on the seabed, locate dangerous rocks or shoals in shallow waters, mooring positions, locating

submerged wrecks, and performing bathymetry profiling.

- *Distributed tactical surveillance.* AUVs and fixed underwater sensors can collaboratively monitor areas for surveillance, reconnaissance, targeting, and intrusion detection.

FUTURE TRENDS

Underwater sensor networks represent an emerging technology that needs a lot of research effort in the following years. The benefits that this technology would offer to maritime industry are invaluable. However, there are a lot of open issues that would require further research in the future.

More effort is required at the physical layer in order to develop efficient, low-power acoustic modems that are able to maximize the available bandwidth and minimize the delivery error rates by using proper FEC coders.

Currently, there are a lot of works related to MAC layer proposals, since this is one of the more sensible parts of the UWSN architecture. It seems that distributed CDMA-based schemes are the candidates for underwater environments, but it depends on many factors, such as the application and network topology. Also, MAC protocols should be designed taking energy consumption into account as a main design parameter.

With respect to routing protocols, they heavily depend on other design factors, such as network topology (even application requirements), node mobility patterns, and energy consumption. Up to now, geographically based routing algorithms seem to be the most adequate for UWSNs, despite requiring the use of localization schemes. Most of the research efforts should be applied to algorithms and protocols that detect and deal with disconnections due to failures, unforeseen mobility of nodes, or battery depletion. Due the underwater nature, cross-layer interaction between all layers should also be required, in order to make a better use of the available resources, and to be able to perform fast adaptations in such a continuously changing environment.

CONCLUSION

Underwater sensor networks are a very recent technology that tries to follow the same steps as terrestrial wireless networks in a very different and challenging network environment. There is an increasing interest in USWN technologies and their potential applications. However, there are several open issues to solve in order to provide an efficient and reliable data transport to the applications.

In years to come, it is expected that UWSNs technology is widely adopted by the industry, resulting in the deployment of new commercial products and solutions that will represent very important revenues to the maritime technology market.

REFERENCES

Abolhasan, M., Wysocky, T., & Dutkiewicz, E. (2004). A review of routing protocols for mobile ad hoc networks. *Ad Hoc Networks Journal, 2*, 1-22.

Akyildiz, F., Pompili, D., & Melodia, T. (2006). State of the art in protocol research for underwater acoustic sensor network. In *Proceedings of the ACM International Workshop on UnderWater Networks (WUWNet)* (pp. 7-16), Los Angeles CA, September 25th.

Akyildiz, F., Su, W., Sankarasubramaniam, Y., & Cayirci, E. (2002). Wireless sensor networks: A survey. *Computer Networks*, *38*(4), 393–422.

Bharghavan, V., Demers, A., Shenker, S., & Zhang, L. (1994). MACAW: A media access protocol for wireless LAN's. In *Proceedings of the ACM SIGCOMM Conference* (pp. 212-225), London, UK, August 31- September 2.

Colladon, J. D. (1893). *Souvenirs et Memoires*. Geneva, CH: Albert-Schuchardt.

Cui, J-H., Kong, J., Gerla, M., & Zhou, S. (2006). Challenges: Building scalable mobile underwater wireless sensor networks for aquatic applications. *IEEE Network, 3*, 12-18.

Curtin, T. B., Bellingham, J. G., Catipovic, J., & Webb, D. (1993). Autonomous oceanographic sampling networks. *Oceanography, 6*, 86–94.

Freitag, L., Stojanovic, M., Kifoyle, D., & Preisig, J. (2004). High-rate phase-coherent acoustic communication: A review of a decade of research and a perspective on future challenges. In *Proceedings of the 7th European Conf. on Underwater Acoustics*, Delft, HL, 5-8 July.

Freitag, L., Stojanovic, M., Singh, S., & Johnson, M. (2001). Analysis of channel effects on direct-sequence and frequency-hopped spread-spectrum acoustic communication. *IEEE Journal of Oceanic Engineering, 26*(4), 586–593.

Fullmer, C. L., & García-Luna-Acebes, J. J. (1995). Floor acquisition multiple access (FAMA) for packet-radio networks. *Computer Communication Review, 25*(4), 262-273.

Harris, A. F., & Zorzi, M. (2007). Modeling the underwater acoustic channel in NS2. In *Proceedings of the ACM International Workshop on Network Simulation Tools (NSTools)*, Nantes, FR, October 22.

Karl, H., & Willig, A. (2005). *Protocols and architectures for wireless sensor networks*. Sussex, UK: John Wiley & Sons Ltd.

Karn, P. (1990). MACA - A new channel access method for packet radio. In *Proceedings of the ARRL 9th Computer Networking Conference* (pp. 134-140), London, Ontario (CA), September 22.

Kebkal, A., Kebkal, K., & Komar, M. (2005). Data-link protocol for underwater acoustic networks. In *Proceedings of the IEEE Oceans Europe* (pp. 1174–1180), Brest, FR, 20-23 June.

Molins, M., & Stojanovic, M. (2006). Slotted FAMA: A MAC protocol for underwater acoustic networks. In *Proceedings of the IEEE OCEANS 2006* (pp. 1-7), Singapore, 16-19 May.

Partan, J., Kurose, J., & Neil Levine, B. (2006). A survey of practical issues in underwater networks. In *Proceedings of the ACM International Workshop on UnderWater Networks (WUWNet)* (pp. 17-24), Los Angeles CA, September 25th.

Pompili, D., Melodia, T., & Akyildiz, I. F. (2006). A resilient routing algorithm for long-term applications in underwater sensor networks. In *Proceedings of Mediterranean Ad-hoc Networking Workshop (Med-Hoc-Net)*, Lipari, Italy, 14-17 June.

Pompili, D., Melodia, T., & Akyildiz, I. F. (2007). A distributed CDMA medium access control for underwater acoustic sensor networks. In *Proceedings of the Mediterranean Ad-hoc Networking Workshop (Med-Hoc-Net)*, Corfu, Greece, 13-15 June.

Rosenbush, S., Crockett, R. O., & Yang, C. (2004). Sin cables, sin normas y a bajo precio. *Dinero: Inteligencia empresarial, 932*, 70-73.

Schill, F., Zimmer, U. R, & Trumpf, J. (2004). Visible spectrum optical communication and distance sensing for underwater applications. In *Proceedings of the Australasian Conference on Robotics and Automation*. Canberra, Australia, 6-8 December.

Stojanovic, M., & Freitag, L. (2006). Multichannel detection for wideband underwater acoustic CDMA communications. *IEEE Journal of Oceanic Engineering 31*(3), 685-695.

Urick, R. J. (1983). *Principles of underwater sound* (3rd ed.). NewYork, USA: McGraw-Hill.

van der Werff, T. J. (2003). Ten emerging technologies that will change the world. *Technology Review (MIT*). Retrieved December 5, 2007, from http://www.technologyreview.com/read_article.aspx?id=13060&ch=infotech

Xie, G., & Gibson, J. (2001). A network layer protocol for UANs to address propagation delay-induced performance limitations. In *Proceedings of the MTS/IEEE OCEANS Conference, 4*, 2087–2094, Honolulu, HI (USA), 5-8 November.

Xie, P., Cui, J.-H. & Lao, L. (2006). VBF: Vector-based forwarding protocol for underwater sensor networks. *Lecture Notes In computer Science, 3976*, 1216–1221.

KEY TERMS

AUVs (autonomous underwater vehicles): Those underwater vehicles able to navigate autonomously to collect sensor data in a specific control area. The most common use of these vehicles is related with the oil and gas industry, allowing one to obtain maps of the seafloor before starting to build the subsea infrastructure.

Direct Sequence Spread Spectrum (DSSS): A modulation technique where the data signal is multiplied by a pseudorandom binary sequence at a frequency much higher than that of the original signal, thereby spreading the energy of the original signal into a much wider band.

Frequency Hopping Spread Spectrum (FHSS): A modulation technique that can be described as a frequency modulation that is repeatedly changing the frequency of the carrier signal in the full available spectrum for transmission.

Hydrophone: A microphone designed to be used underwater for recording or listening to underwater sound. Most hydrophones are based on a piezoelectric transducer that generates electricity when subject to a pressure change.

MANETs (mobile ad hoc networks): Refer to those wireless networks composed of mobile nodes that can communicate between them without the need of any kind of infrastructure (base stations).

Point Coordination Function (PCF): A MAC protocol, used in wireless local area networks (WLANS), that relays a central coordinator, usually known as an access point (AP). The access to the medium is governed by APs unit allowing an ordered and collision-free access to the network.

Proactive and Reactive Routing Protocols: Represent two different styles of routing packets. The former ones maintain updated the routing info by periodically distributing routing info throughout the network. In contrast, a reactive routing protocol finds a route on demand by flooding the network with route request packets.

SONAR (sound navigation and ranging): A device that uses the properties of underwater sound propagation to communicate, navigate, or detect other vessels. It sends pulses of sound to probe the sea, and the echoes are then processed to extract information (shape, distance, composition, etc.) about the sea, its boundaries, and submerged objects.

WSNs (wireless sensor network): Can be defined as a particular case of MANETs in the sense that each node is required to record and wirelessly distribute environmental data obtained through a set of sensors attached to it. They are typically small and low-power consuming devices, and use to be deployed in high node density networks.

Underwriting Automobile Insurance Using Artificial Neural Networks

Fred Kitchens
Ball State University, USA

INTRODUCTION

As the heart of the insurance business, the underwriting function has remained basically unchanged for the past 400 years, since Lloyd's of London was a place where ship owners would seek out financial supporters. The two would contractually agree to share the financial risk in the unlucky event that the ship would be lost at sea (Gibb, 1972; Golding and King-Page, 1952).

In the modern insurance market, insurance underwriters perform a similar financial function on behalf of their respective insurance companies. Underwriters gather pertinent information and analyze their potential clients to determine whether or not they should underwrite the risk; and if so, what premium they would require for the insurance policy. Insurance companies employ actuaries to assist the underwriter in this process by studying past insurance losses and making predictive models for future risks. Using traditional statistical methods, insurance actuaries look for loss-contributing characteristics within the risk (Webb, Harrison, et al., 1992). When the actuaries find positive relationships between the policy characteristics and subsequent losses, they create "underwriting guidelines" for the underwriters to follow when analyzing potential clients and setting premiums (Malecki and Underwriters, 1986).

For hundreds of years, actuaries used pencil and paper to perform their statistical analysis It was a long time before they had the help of a mechanical adding machine. Only recently have they had the benefit of computers. As recently as 1981, computers were not considered important to the process of insurance underwriting. Leading experts in insurance underwriting believed that the judgment factor involved in the underwriting process was too complex for any computer to handle as effectively as a human underwriter (Holtom, 1981).

Recent research in the application of technology to the underwriting process has shown that Holtom's statement may no longer hold true (Gaunt, 1972; Kitchens, 2000; Rose, 1986). The time for computers to take on an important role in the insurance underwriting process may be upon us. The author intends to illustrate the applicability of artificial neural networks to the insurance underwriting process.

BACKGROUND

The American Institute for Chartered Property Casualty Underwriters (CPCU) reports that the most common considerations found in automobile underwriting guidelines are:

- age of operators;
- age and type of automobile;
- use of the automobile;
- operator's driving record;
- territory;
- gender;
- marital status;
- operator's occupation;
- operator's personal characteristics; and
- physical condition of the vehicle.

Traditionally, these comprise the core variables used in determining the acceptability, classifying, and rating of private passenger automobile insurance policies (Malecki and Underwriters, 1986).

Private passenger automobile insurance is well-suited for artificial intelligence applications applied to the underwriting function. There are three primary reasons for this:

- there is a fixed set of finite data used to make the underwriting decision;
- policies are highly standardized; and
- deviations from the standard insurance contract are rare.

In recent years, researchers have considered the application of computers to the process of automobile insurance underwriting. Two studies attempted to predict the acceptability of a given policy from a broad underwriting standpoint (Gaunt 1972; Rose 1986). Two other studies considered the possibility of predicting a loss on an individual-policy basis (Lemaire, 1985; Retzlaff-Roberts and Puelz, 1966). Another study focused the relationship between premium and customer retention from year-to-year. One study was designed to predict losses on individual policies using artificial neural networks (Kitchens, 2000).

The recent use of artificial neural networks represents what may result in the most accurate application of computers to the underwriting process. Originally developed in the 1940s, artificial neural networks were designed to replicate and study the thought process of the human brain (Cowan and Sharp, 1988). Early research showed that all processes that can be described with a finite number of symbolic expressions could be represented with a finite number of interconnected neurons (Wilson, Starkweather, et al., 1990). Thus, artificial neural networks also provide a means of economic problem solving.

The author believes that for a number of reasons discussed in the following section, artificial neural networks can be successfully applied to the insurance underwriting process in order to reduce the ratio of insurance losses to insurance premiums.

NEURAL NETWORKS FOR INSURANCE UNDERWRITING

Artificial neural networks were first developed in the 1940s as a mathematical model used to study the human thought process (Cowan and Sharp, 1988). In 1943, McCulloch and Pitts proved that all processes which can be described with a finite number of symbolic expressions can be represented in a network of interconnected neurons (Wilson, Starkweather, et al., 1990). This makes the artificial neural network a mathematical modeling tool in addition to a representation of the human brain.

Using a data set consisting of dependent and independent variables, an artificial neural network can be trained until it converges on an optimal solution for the dependent variable(s). If properly developed, the resulting model will be at least as accurate as traditional statistical models (White, 1989).

The insurance business, as practiced in the United States, has certain characteristics that produce less than optimal financial results. There are five basic reasons that the unique abilities of artificial neural networks can improve the underwriting process:

First, an artificial neural network model will be successful because the inequity of the current rate classification system will allow neural networks the opportunity to more accurately assess the risk level of each and every individual policyholder, rather than a class of policyholders (Wood, Lilly, et al., 1984).

Second, an artificial neural network model will produce improved results because current actuarial methods of study will benefit from the broad range of available tools, such as more recent developments in the field of artificial intelligence (Cummins and Derrig, 1993; Kitchens, 2000).

Third, an artificial neural network model will improve the current state of actuarial research. Traditionally, the primary method of research in this field has been to predict the *pure premium* (the amount of premium required to pay all of the losses in a given class of insured accounts, a.k.a. "relative rates"). In comparison, *actual premiums* include the pure premium along with other important factors such as profit margin and operating expenses. The traditionally used pure premium models follow an actuarial approach, but not necessarily an underwriting approach. While it is intended to reduce corporate loss ratios, current actuarial research does not take an underwriting approach to the process. A fresh perspective on the problem could produce improved results (Kitchens, Booker, et al., 2002).

Fourth, an artificial neural network will produce improved results because historically, statistical models used in predicting insurance losses have been able to produce only marginal incremental improvements. Given the current state of technology, the time has come for new insurance actuarial models to take advantage of the available speed and flexibility of artificial neural networks to solve what is clearly a complex problem, which will require extensive training and is likely to involve a complex architecture (Kitchens, 2000, 2004; Kitchens, Johnson, et al., 2001).

Fifth, even if the actuarial models are "perfect" (which the author contends they are not), the neural network should be capable of at least matching the current statistical results, if not improving upon them (Kitchens, Booker, et al., 2002). This is because artificial neural networks comprise a class of nonlinear statistical models whose processing methods are designed to simulate the functioning of the human brain (Hawley, Johnson, et al., 1990). The advantage of neural network models over other modeling methods grows with the complexity of the relationship between input and output variables; however, greater complexity of the underlying relationships between variables requires a more complex design (Lee, White, et al., 1993). Provided the appropriate network architecture, a neural network output function can accurately approximate any mathematical function (White, 1989). Further, a model can achieve any degree of desired accuracy if the neural network is properly designed (Funahashi, 1989).

NEURAL NETWORK MODELS: DESIGN ISSUES

Automobile accidents occur with a certain degree of randomness, and it is expected that they will be very difficult to predict on an individual-policy basis. Previous research has shown that an underwriter's ability to predict the actual value of a paid claim is exceedingly difficult, if possible at

all (Kitchens, Johnson, et al., 2001). However, a successful system needs only to predict the incident (occurrence) of a loss, not the dollar value. In addition, a successful model would not have to predict each and every accident, as long as the predictions that the model makes are accurate. In fact, a new model needs only to out-perform any current models in order to prove itself worthwhile. As an industry rule-of-thumb, the average loss-to-gross-premium ratio is approximately 60 percent. The rest of the collected premium is used to pay operating expenses and a small profit of approximately 3 percent (Kitchens, Jr., 1999). Thus, if a new model could reduce losses by 1 percent, it would represent a 33 percent increase in operating profit. If a corresponding decrease in operating expenses such as loss-adjustment expenses is incurred, the operating profit could be increased by as much as 53 percent. This in itself is not a justification for using artificial neural networks, but it is enough incentive to try nontraditional techniques.

As Data Mining and Knowledge Management grow in their importance to business, trade secrets grow as well. Description of new software such as *Risk Manager* by Valen Technologies has an uncanny resemblance to an artificial neural network. The company will not reveal the details behind their technology except to call it, "a predictive insurance underwriting software system" with "computational learning technology" (Stodder, 2005). While it is difficult, if not impossible, to say how much underwriting is currently performed by artificial neural networks, it is certainly a growing field. One prediction indicates that by 2009 as much as fifty percent of all insurance underwriting decisions will be automated using data mining technology (Betts, 2004). While it may be theoretically possible for a computer program to handle the underwriting function, the larger hurdles may be obtaining regulatory permission and gaining user acceptance of a non-human process. Aside from the highly regulated nature of the insurance industry, the human side must also be considered. In the 1970s and 1980s managers and regulators were reluctant to accept technology that made decisions for them. In the 1990s social acceptance started to turn. Two important features have lead the cause for acceptance. New applications do not require human intervention; the user is not required to "do" anything. And, new applications are configured to translate decisions into action (Davenport and Harris, 2005). A computer-based model that aids the underwriter in the decision-making process by suggesting an appropriate course of action is a step in the right direction. An artificial neural network as an underwriter's tool could be used in several ways: to help train new underwriters, to provide a suggested course of action, or to handle routine policies, allowing the underwriter to spend more time on more complex policies. As an underwriter's tool, a model should be useful, reliable, and convenient while providing information of value.

In the development of an artificial neural network model as an underwriter's tool, several things must be determined: the output required, the input variables, the type of artificial neural network, the architecture of the artificial neural network, and the interpretability of the output.

The underwriter's decision-making process boils down to two basic decisions. First, a decision must be made whether to accept or reject the risk. Second, if accepted, a decision as to the premium that will be charged.

Depending on the purpose of the model, the required output may place stringent requirements on the required input. One reason all previous models have had limitations on their applicability has been due to the lack of quantity or quality in the available data sets used to generate the model.

For purposes of insurance loss-prediction modeling, the Genetic Adaptive Neural Network Training (GANNT) algorithm is an appropriate choice. The GANNT algorithm is designed to overcome difficulties associated with the popular gradient and back propagation techniques (Dorsey and Mayer, 1994).

The genetic algorithm was first proposed in 1975 (Nygard, Ficek, et al., 1992). It was shown that the biological evolutionary process could be applied to an artificial mathematical modeling system (Konza, 1992). The concept is based on the theory that an optimization problem can be encoded as a list of concatenated parameters (nodal weights), which are used in the artificial neural network (Whitley, Starkweather, et al., 1990). The genetic algorithm works through a process of modeling founded on the biological process by which DNA replicates, reproduces, crosses over, and mutates (Crane, 1950). These procedures are then modeled in a computer-based algorithm to solve complex problems (Nygard, Ficek, et al., 1992). The actual operations of the genetic GANNT algorithm are explained in detail by Dorsey, Johnson, and Mayer (1991).

RESULTS

Recent research has shown that the automobile insurance underwriting process practiced in the United States is lacking in precision. Underwriters are not efficiently utilizing all of the information available to them. The current human-based underwriting process uses only a portion of the available information. In some cases, so much valuable information is overlooked that the remaining unutilized information can be used to make a more accurate accept/reject decision than the initial underwriter made (Kitchens, 2000,,2004). With the benefit of the flexibility and adaptability of an artificial neural network, the unutilized information may be used in the future to make more accurate and more precise underwriting decisions.

FUTURE TRENDS

Future research should be focused on two primary areas. First, to be properly trained, an artificial neural network requires data from both the "accepted" and the "unaccepted" policies. Thus, some effort needs to be focused on obtaining information about the policies that are currently being rejected by the underwriter. This is difficult information to obtain because insurance companies have no reason to track losses on policies they previously rejected. But this data will be valuable in the development of a more complete underwriting model.

Second, future research should go beyond the accept-or-reject decision and investigate the premium-setting decision. A model capable of second-guessing an underwriter's accept-or-reject decision might be capable of reducing an insurance company's losses. But a model that can both accept or reject and set the premium, might be capable of reducing the cost of underwriting, streamline the business process, and produce policies that are more appropriately priced. These are the loftier long-term goals (Davenport and Harris, 2005).

CONCLUSION

Since the process of insurance underwriting began 400 years ago, until recently the biological neural network (human brain) has been the fastest, most efficient, and most readily available information processing tool available. It has naturally been the tool of choice for underwriters when making accept-or-reject and pricing decisions on insurance policies.

In 1981, it was a common belief that computers could not be used to replace insurance underwriters (Holtom, 1981). In the past 20 years or so, computers and technology have made tremendous advancements. During the same time period, sophisticated mathematical models and the algorithms used to generate them, including the Genetic Adaptive Neural Network Training algorithm, have taken advantage of increased computing power and availability. Artificial neural networks and the technology used to run them have been shown to outperform the traditional human-based mental decision making practices, in both speed and accuracy, if only for limited domains and applications, such as insurance underwriting.

REFERENCES

Betts, M. (2004). *Predictions For BI's Future*. Computer-World: Business Intelligence: 4.

Cowan, J. D. and Sharp, D. H. (1988). *Neural Nets*. Quarterly Reviews of Biophysics 21: 305-427.

Crane, H. R. (1950). *Principles and Problems of Biological Growth*. The Scientific Monthly LXX(6): 376-386.

Cummins, J. D. and Derrig, R. A. (1993). *Fuzzy Trends in Property-Liability Insurance Claim Costs*. Journal of Risk and Insurance, 60(3): 429-466.

Davenport, T. H. and Harris, J. G. (2005). *Automated Decision Making Comes of Age*. MIT Sloan Management Review, 46(4): 83-89.

Dorsey, R. E., Johnson, J. D. ,, et al. (1991). *The Genetic Adaptive Neural Network Training (GANNT) Algorithm for Genetic Feedforward Artificial Neural Systems*. Working Paper, The University of Mississippi.

Dorsey, R. E. and Mayer, K. J. (1994). *Optimizing Using Genetic Algorithms*. Greenwich, CT: JAI Press Inc.

Funahashi, K. (1989). *On the Approximate Realization of Continuous mappings by Neural Networks*. Neural Networks 2: 183-192.

Gaunt, L. D. (1972). *Decision-Making in Underwriting: Policyholder Selection in Private Passenger Automobile Insurance*, Georgia State University - College of Business Administration: 00310.

Gibb, D. E. W. (1972). *Lloyd's of London: a study in individualism*. London, Lloyd's.

Golding, C. E. and King-Page, D. (1952). *Lloyd's*. New York: McGraw-Hill.

Hawley, D. D., Johnson, J. D. , et al. (1990). *Artificial Neural Systems: A New Tool for Financial Decision-Making*. Financial Analysts Journal 46(November/December): 63-72.

Holtom, R. B. (1981). *Underwriting: Principles and Practices*. Cincinnati, Ohio: The National Underwriter Company.

Kitchens (2000). *Using Artificial Neural Networks to Predict Losses in Automibile Insurance. Graduate School*. Oxford, The University of Mississippi: 150.

Kitchens, F. L. (2004). *Financial implications of neural networks in automobile insurance underwriting*. International Conference on Fuzzy Sets and Soft Computing in Economics and Finance, St. Petersburg, Russia.

Kitchens, F., Johnson, J. D. , et al. (2001). *Predicting severity in automobile insurance losses using artificial neural networks*. Production and Operations Management Society International Conference, Sao Paulo, Brazil.

Kitchens, F. L., Booker, Q. E. , et al. (2002). *An Application of Neural Networks to Insurance Underwriting*. 33rd Annual COnference of the Decision Sciences Institute, Southwest Region, St. Louis Missouri.

Kitchens, Jr., F. L. (1999). Past President, Cherokee Insurance Co.; Past President Coastal Plains Insurance Associates.

Konza, J. R. (1992). *Genetic programming: On the Programming of Computers by Means of Natural Selection*. Cambridge, Massachusetts: The MIT Press.

Lee, T. H., White, H., et al. (1993). *Testing for neglected nonlinearity in time series model: a comparison of neural network methods and alternative tests*. Journal of Econometrics, 56(3): 269-290.

Lemaire, J. (1985). *Automobile Insurance : Actuarial Models*. Boston, MA: U.S.A., Kluwer-Nijhoff, Distributors for North America Kluwer Academic Publishers.

Malecki, D. S. and A. I. F. P. A. L. Underwriters (1986). *Commercial liability risk management and insurance*. Malvern, Pa., American Institute for Property and Liability Underwriters.

Nygard, K. E., Ficek, R. K., et al. (1992). *Genetic Algorithms: Biologically Inspired Search Method Borrows Mechanisms of Inheritance to Find Solutions*. OR/MS Today (August): 28-34.

Retzlaff-Roberts, C. and Puelz, R. (1966). *Classification in Automobile Insurance Using a DEA and Discriminant Analysis Hybrid*." Journal of Productivity Analysis 7(4): 417-27.

Rose, J. C. (1986). *An Expert System Model of Commercial Automobile Insurance Underwriting (Knowledge Base)*, The Ohio State University: 00280.

Stodder, D. (2005). *A Prediction: Data Integration Will Improve Sefety*. Intelligent Enterprise.

Webb, B. L., Harrison C. M., et al. (1992). *Insurance operations*. Malvern, PA:, American Institute for Chartered Property Casualty Underwriters.

White, H. (1989). *Neural Networks and Statistics*. AI Expert 49(December).

Whitley, D., Starkweather, T. et al. (1990). *Genetic Algorithms and Neural Networks: Optimizing Connections and Connectivity*. Parallel Computing 14: 347-361.

Wilson, D., Starkweather, T., et al. (1990). *Genetic Algorithms and Neural Networks: Optimizaing Connections and Connectivity*. Parallel Computing 14: 347-361.

Wood, G. L., Lilly, C. C. , et al. (1984). *Personal Risk Management and Insurance*. U.S.A., American Institute for Property and Liability Underwriters.

KEY TERMS

Actuary: A statistician who practices the collection and interpretation of numerical data, especially someone who uses statistics to calculate insurance premiums.

Artificial Neural Network: (commonly referred to as "neural network" or "neural net"**)** A computer architecture, implemented in either hardware or software, modeled after biological neural networks. Nodes are connected in a manner suggestive of connections between the biological neurons they represent. The resulting network "learns" through directed trial and error. Most neural networks have some sort of "training" algorithm to adjust the weights of connections between nodes on the basis of patterns found in sample or historical data.

Back Propagation: A learning algorithm for modifying a feed-forward neural network which minimizes a continuous "error function" or "objective function." Back propagation is a "gradient descent" method of training in that it uses gradient information to modify the network weights to decrease the value of the error function on subsequent tests of the inputs. Other gradient-based methods from numerical analysis can be used to train networks more efficiently.

Biological Neural Network: A network of neurons that function together to perform some function in the body such as thought, decision making, reflex, sensation, reaction, interpretation, behavior, etc.

Genetic Algorithm (GA): A class of algorithms commonly used for training neural networks. The process is modeled after the methods by which biological DNA are combined or mutated to breed new individuals. The crossover technique, whereby DNA reproduces itself by joining portions of each parent's DNA, is used to simulate a form of genetic-like breeding of alternative solutions. Representing the biological chromosomes found in DNA, genetic algorithms use arrays of data, representing various model solutions. Genetic algorithms are useful for multi-dimensional optimization problems in which the chromosome can encode the values for connections found in the artificial neural network.

Insurance: Protection against future loss. In exchange for a dollar value (premium), insurance is a promise of reimbursement in the case of loss. Contractual arrangement of insurance may be voluntarily or by government mandate

(such as minimum requirements for automobile insurance for licensed drivers).

Nodal Connections: Connections between nodes in an artificial neural network. They are communication channels that carry numeric data. They simulate the axons and dendrites used to carry electrical impulses between neurons in a biological neural network.

Node: A mathematical representation of a biological neuron. Multiple layers of nodes are used in artificial neural networks to form models of biological neural networks.

Risk: An individual or organization's exposure to a chance of loss or damage.

Underwriter: (Insurance Underwriter) An employee of an insurance company whose job duties include analyzing an application for insurance and making a decision whether to accept or reject the application. If accepted, the underwriter further determines the premium to be charged. Underwriters also review existing insurance contracts for renewal.

Unified Modeling Language 2.0

Peter Fettke
Institute for Information Systems (IWi) at the DFKI, Germany

INTRODUCTION

Mature engineering disciplines are generally characterized by accepted methodical standards for describing all relevant artifacts of their subject matter. Such standards not only enable practitioners to collaborate, but they also contribute to the development of the whole discipline. In 1994, Grady Booch, Jim Rumbaugh, and Ivar Jacobson joined together to unify the plethora of existing object-oriented systems engineering approaches at semantic and notation level (Booch, 2002; Fowler, 2004; Rumbaugh, Jacobson & Booch, 1998). Their effort leads to the unified modeling language (UML), a well-known, general-purpose, tool-supported, process-independent, and industry-standardized modeling language for visualizing, describing, specifying, and documenting systems artifacts.

UML is applicable to software and non-software domains, including software architecture (Medvidovic, Rosenblum, Redmiles, & Robbins, 2002), real-time and embedded systems (Douglass, 2004), business applications (Eriksson & Penker, 2000), manufacturing systems (Bruccoleri, Dieaga, & Perrone, 2003), electronic commerce systems (Saleh, 2002), data warehousing (Dolk, 2000), bioinformatics (Bornberg-Bauer & Paton, 2002) and others. The language uses multiple views to specify system's structure and behavior. Modeling tools supporting the development of UML diagrams are available from a number of commercial vendors and the open source community (OMG, 2006b; Robbins & Redmiles, 2000).

Table 1 depicts the origin and descent of UML. The recent version UML 2.0 supports thirteen different diagram types. Table 2 overviews the main concepts of each diagram, a more detailed description is given below. For a full description of all semantics see (OMG, 2005a, 2005b, 2006a, 2006c) respectively the available secondary literature (Fowler, 2004; Rumbaugh et al., 1998).

UML version 2, first planned for 2001 (Kobryn, 1999, p. 30), was finally completed in 2006. This major revision mainly focuses on language extensibility, language specification, language precision and expressiveness. Although the complete language specification was almost fully rewritten, this revision is primary an internal reorganization with just minor consequences for the end user. For example, the new diagrams mainly clarify and resemble existing diagram types.

The description of UML 2.0 consists of four separate documents (Kobryn, 2002):

- **Infrastructure:** This document is concerned with core language features. It specifies the base classes that provide the foundation for UML modeling concepts.
- **Superstructure:** Advanced topics such as component and activity modeling are defined by this specification. It describes the constructs that developers use to build UML models.
- **Object constraint language (OCL):** This specification describes the language used for invariants, operation specifications etc.

Table 1. History of UML (Fowler, 2004, pp. 151-159; Kobryn, 1999, p. 30)

Year	Version	Comments
1995	0.8	Origin of UML, so-called "Unified Method"
1996	0.9	Refined proposal
1997	1.0	Initial submission to OMG
1997	1.1	Final submission to OMG
1998	1.2	Editorial revision with no significant technical changes
1999	1.3	New use case relationships, revised activity diagram semantics
2001	1.4	Minor revisions, addition of profiles
2003	1.5	Adding action semantics
2005	1.4.2	Standardized by the International Organization for Standardization (ISO/IEC 19501:2005)
2005/6	2.0	Deep changes to meta-model, new diagram types, improved expressiveness

Table 2. UML diagram types

Focus	Diagram	Purpose	Main Concepts	Supported Since
Structure diagrams	Class	Object structure	Class, features, relationships	UML 1
	Object	Example configuration of instances	Object, link	UML 1 (unofficially)
	Component	Structure and connections of components	Component, interface, dependency	UML 1
	Composite structure	Decomposition of a class during runtime	Part, interface, connector, port	UML 2
	Package	Interrelationships between packages	Package, dependency	UML 1 (unofficially)
	Deployment	Deployment of components to nodes	Node, component, dependency	UML 1
Behavior diagrams	Use case	User interaction with system	Use case, actor	UML 1
	Activity	Procedural and parallel behavior	State, activity, completion, transition, fork, join	UML 1
	State machine	Change of events during object's lifetime	State, transition, event, action	UML 1 statechart diagram
Interaction diagrams	Sequence	Interaction between objects emphasizing sequences	Interaction, message	UML 1
	Communication	Interaction between objects emphasizing collaborations	Collaboration, interaction, message	UML 1 collaboration diagram
	Timing	Interaction between objects emphasizing timings	Object, timing constraint, state, event	UML 2
	Interaction	Interplay between activities and sequence interactions	Combination of sequence and activity diagram (see there)	UML 2

- **Diagram interchange:** The storage and exchange of model including the layout of UML models is covered by this specification.

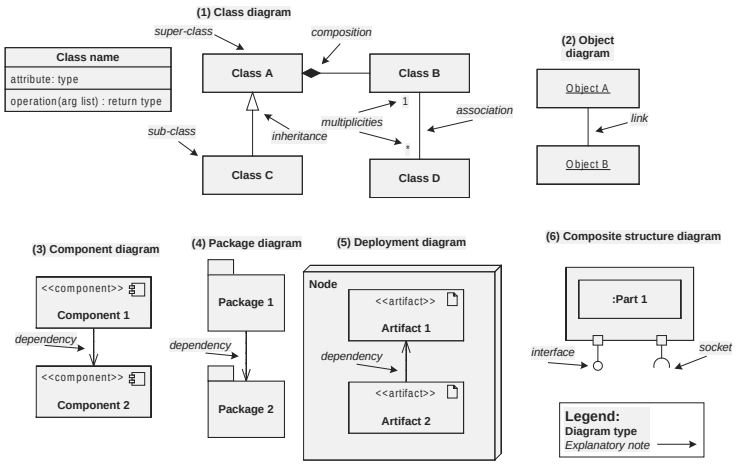
The specification of the UML is publicly available and maintained by the Object Management Group (OMG). OMG's standardization process is formalized and consists of several proposal, revision, and final implementation activities (Kobryn, 1999, p. 31f.). Note, UML 1.4.2 is adopted by the International Organization for Standardization, too.

BACKGROUND

There is a great deal of terminological confusion in the modeling literature. A modeling language or grammar provides a set of constructs and rules that specify how to combine the constructs to model a system (Wand & Weber, 2002, p. 364). It can be distinguished between an abstract syntax and a concrete syntax or notation of a language. While the abstract syntax specifies conceptual relationships between the constructs of the language, the concrete notation defines symbols representing the abstract constructs. In contrast, a modeling method provides procedures by which a language can be used. A consistent and suited set of modeling methods is called a methodology. A model is a description of a domain using a particular modeling language.

The UML specification provides an abstract syntax and a concrete notation for all UML diagrams as well as an informal description of the constructs' semantics. The UML's language specification is independent of but strongly related to other OMG standards such as Common Data Warehouse Model, XML Metadata Interchange or Meta Object Facility. A modeling method or a modeling methodology is not defined by the UML standard. Hence, the language is process-neutral and can be used with different software development processes.

Figure 1. UML structure diagram examples



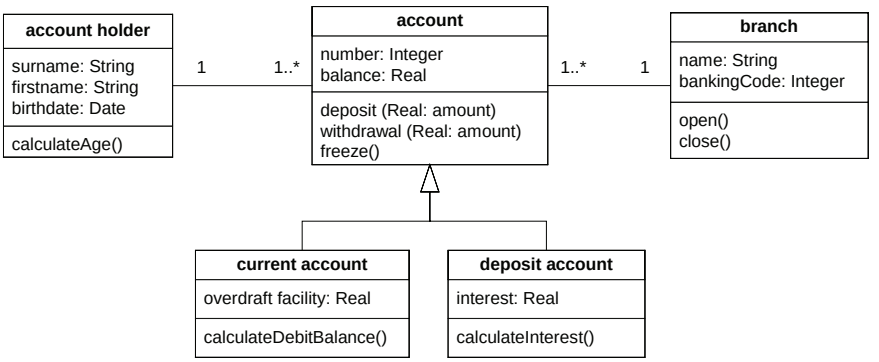
Conceptual modeling has a long history. Other modeling approaches that are to a certain degree accepted in practice, for instance the Entity-Relationship Model or flow charts, have a much more limited scope than UML. These approaches address just some aspects of systems' specification, namely data and process view. In contrast, UML supports the specification of static as well as dynamic aspects. Other approaches with a similar scope, e.g. Open Modeling Language (Firesmith, Henderson-Sellers & Graham, 1998), are not widely accepted in practice.

STRUCTURE DIAGRAMS

Structure or static diagrams describe the objects of a system in terms of classes, attributes, operations, relationships, interfaces and connectors (see Figure 1).

1. **Class diagram:** A class diagram can be viewed as a graph of several elements connected by static relationships. The main element is a class. Classes represent concepts within the system being modeled and are descriptors for a set of objects with similar structure, behavior, and relationships. An object represents a particular instance of a class. Each class has a unique name among other classes within a specific scope (usually a UML package). A class can hold several attributes and operations. Attributes have names and belong to particular types that can be simple data types such as Integer, String, Boolean as well as complex types (e.g., other classes). Operations are services offered by an instance of the class and may be requested by other objects during run-time. Different relationships between classes can be defined. Figure 2 depicts a class diagram for banking systems. An account is described by the attributes 'number' and 'balance'. The operations 'deposit', 'withdrawal', and 'freeze' are offered by an account. Each account is kept by a 'branch' and is assigned to a 'holder'. The classes 'deposit account' and 'current account' reuse the structure and behavior of the class 'account' (inheritance relationship). In addition, the specialized account classes define further feature, e.g. an object of the class 'current account' is described by the property 'overdraft facility' and offers an operation calculating the current debit balance.
2. **Object diagram:** An object diagram is an instance of a class diagram and depicts the state of the system at a point in time (e.g., a particular configuration of several objects). It contains objects including their actual values of attributes and links describing object references.
3. **Component diagram:** Component diagrams capture the physical structure of a software system during build-time. Components in UML are physical elements such as source, binary or executable modules and files respectively. Simple components can be aggregated to complex components to specify physical containment relations. Directed relationships between components specify that one component relies or refines the other. Such relationships are called dependencies.
4. **Package diagram:** Packages are grouping constructs that allow to group elements together into high-level units. Typically, packages are used to group classes. It is possible that one package is also be member of another package, so packages build a hierarchic structure. A package diagram visualizes the packages of a

Figure 2. Class diagram for banking systems



system. The sub-package relationship is shown by a nested package, dependencies between packages are depicted by dashed arrows.

5. **Deployment diagram:** While component diagrams primary show build-time dependencies of components, deployment diagrams show a run-time configuration of the system's components. In addition, a deployment diagram uses nodes representing a processing resource, for instance a server or workstation, that can execute system operations during run-time.
6. **Composite structure diagram:** This diagram shows how a class is hierarchically decomposed into several parts. The decomposed parts are typically represented by interfaces. Each part can support or require a particular interface. It is possible that a part implements an interface by using a delegating connector. Additionally, ports are used to group the required and provided interfaces into logical chunks that connect a component to other comments.

BEHAVIOR DIAGRAMS

Behavior diagrams describe the dynamics between objects of a system in terms of interactions, collaborations, and state histories (see Figure 3).

7. **Use case diagram:** A use case specifies a complete set of events within a system to fulfill tasks or transactions in an application from a user's point of view. In a use case diagram, a set of use cases, actors, and relationships between these elements are depicted. Several use cases may optionally be enclosed by a rectangle that represents the boundary of the containing system. An actor describes a particular role of a human or non-human user of the system being modeled.
8. **Activity diagram:** While state chart diagrams are used to specify the behavior of a single object, activity diagrams can describe behavior that crosses object boundaries. They are analogous to traditional flowcharts and are often used to document (business) processes or the dynamics inside a use case. So-called fork bars are used to describe activities that can be executed in parallel. Parallel activities get synchronized by so-called join bars. Guards are used to specify conditional forks that are only executed if particular conditions hold.
9. **State machine diagram:** Object behavior is represented by state chart diagrams that can specify the behavior of an entire object or a single method. A state describes a condition during the lifetime of an object. Transitions are relationships between two states describing that an object's state can change from the first to the second state. The change of a state is triggered by an event that occurs in the modeled system. There are two special types of states: An initial state identifies the point at which behavior starts when an object is created, a final state identifies the point at which behavior ends (end of object's lifetime).

INTERACTION DIAGRAMS

Interaction diagrams are all derived from the more general behavior diagrams and are used for meeting particular modeling requirements (see Figure 4).

10. **Sequence diagram:** Sequence diagrams describe interactions between different objects. An interaction consists of a partially ordered set of messages that are exchanged by the participants of that interaction. Sequence diagrams have two dimensions: The horizontal dimension represents the participants of the

Figure 3. UML behavior diagram examples

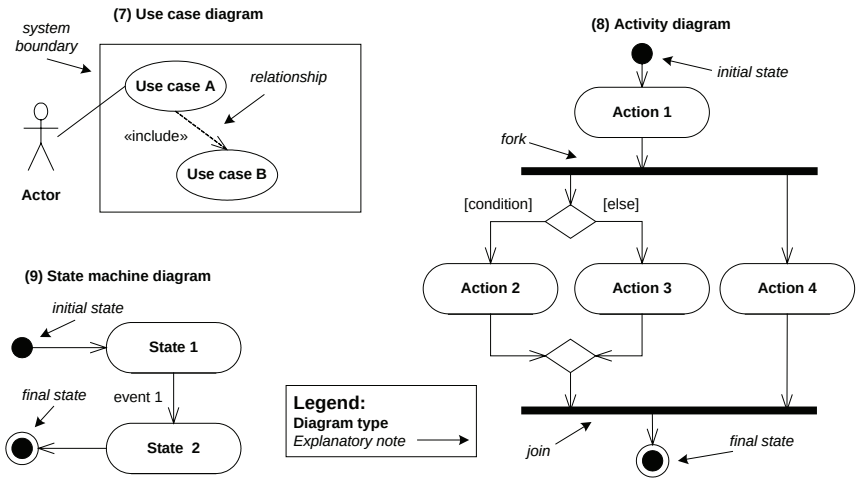
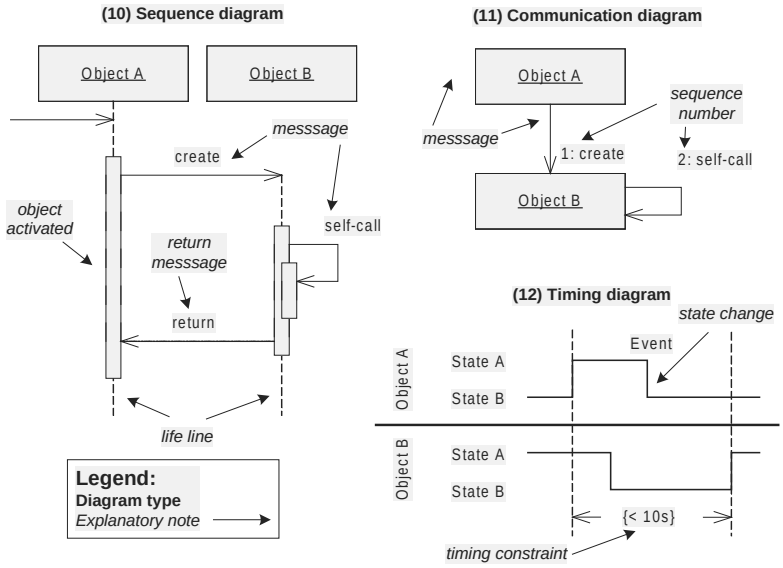


Figure 4. UML interaction diagram examples



interaction; the vertical dimension represents the flow of time (usually time proceeds from up to down).

11. **Communication diagram:** Communication and sequence diagrams use the same underlying information and can easily be transformed into each other. While sequence diagrams emphasize the sequence of communication between objects, communication diagrams show the roles of the participants of an interaction and their relationships. A sequence number specifies the flow of messages in an interaction, so no time dimension is needed in this diagram. Simple communication patterns can be depicted by communication diagrams; sequence diagram can better specify complex message exchanges or requirements for real-time systems.
12. **Timing diagram:** This diagram focuses on timing constraints for a single object or multiple objects. Additionally to the information gathered by a state machine diagram, minimal or maximal timing constrains between the occurrence of events can be specified. For example, a specification might say that at least ten seconds must pass between two events.
13. **Interaction diagram:** An interaction diagram is a mix of activity diagrams and sequence diagrams (hence it is not explicitly shown in the Figure 4). One way to use this kind of diagram is to detail the activities of an activity diagram by a particular sequence diagram. These sequence diagrams show the flow of messages between objects during execution of the corresponding activity.

ADVANCED TOPICS

- **Object constraint language (OCL):** There is often a need to capture unambiguously system's semantics in a precise and rigorous way. OCL is used for that purpose. It is a formal textual language inspired by the 'Design by Contract' concept and provides concepts for the definition of constraints such as invariants, pre- and post-conditions (Cengarle & Knapp, 2004; Warmer & Kleppe, 2003).
- **Language specification and meta-model:** The UML itself is specified using textual descriptions and a four-layered meta-modeling approach (Atkinson & Kühne, 2002, pp. 291-296). In this approach, the semantic constructs at each layer are recursively refined. The top layer, the meta-meta-model (M3), provides a so-called Meta Object Facility (MOF) to specify meta-models on the next lower layer. The MOF is used on the meta-model (M2) layer to specify the concepts of UML diagrams, e.g. class diagram etc. The model (M1) and object (M0) layer are user-defined. The M1 layer specifies concrete UML models, the M0 layer instances of the former.
- **Extension mechanisms:** So-called heavyweight extensions are supported by MOF and carried out on the meta-model (M2) layer. Such extensions have great impact on the language and are not performed by a particular modeler. User extensions are usually lightweight extensions that are built-in mechanisms of the UML. Lightweight extensions comprise constraints (OCL expressions), tagged values (attached additional information to model elements), and stereotypes (most powerful lightweight mechanism ranging from concrete syntax modifications to semantics redefinitions (Berner, Glinz, & Joos, 1999)).

FUTURE TRENDS

There has been always a strong discussion about how UML should and should not evolve in the future (Engels, Heckel, & Sauer, 2001; Henderson-Sellers, 2005; Kobryn, 2004; Miller, 2002). These trends include:

- **Model driven architecture (MDA):** MDA promotes modeling through the whole system's life-cycle (Frankel, 2003). Its objective is to fully automate the system's development process.
- **Executable UML:** Executable UML enriches modeling concepts with execution semantics (Mellor & Balcer, 2002). This opens the possibility of software development without "classical" programming.
- **Model libraries:** UML is used to standardize domain specific models fostering the usage of reference models (Fettke & Loos, 2003a; Fettke, Loos, & Zwicker,Loos, 2007). Known reference models, e.g. OMG's Business Enterprise Integration or Finance Domain Task Forces, support model reuse.
- **Ontological analysis and semantics:** This research line evaluates UML from an ontological point of view and incorporates real-world semantics into UML constructs (Opdahl & Henderson-Sellers, 2002). The aim of an ontological evaluation is to examine whether all constructs of an ontology can be mapped onto the constructs of UML and vice versa.
- **Component-based development:** UML is primary an object-oriented language. To fully support component-based development, some enhancements are needed (Dahanayake, 2003; Fettke & Loos, 2003b; Kobryn, 2000). Particularly, component descriptions must include dependencies on other components, quality specifications for needed and offered services, and domain-specific semantics.
- **Tailoring UML:** The adaption of UML to particular needs of a development project is of high importance. There are already some ideas for tailoring UML (France, Ghosh, Dinh-Trong & Solberg, 2006, p. 61).

- **Empirical foundations:** Until now, the research on UML is mainly conceptual without any strong empirical foundation (see for counter-examples (Dobing & Parsons, 2005; Glezer, Last, Nachmany, & Shoval, 2005; Grossman, Aronson, & McCarthy, 2005)). Future research has to adesss important questions such as: How is UML used in practice? Are some constructs of UML not relevant for modeling practice? Does the use of UML improve information system quality and developers productivity?

CONCLUSION

Although almost everyone acknowledges the practical benefits of a standardized modeling language (e.g., protection of investments in technology, easier model exchange and reuse, better professional training (Frank, 1997, p. 13)), there are important opportunities that have to be challenged. UML's size (UML 2 has approximately 1000+ pages) and complexity is compared with other languages overwhelming (Erickson & Siau, 2004; Siau & Cao, 2001). Therefore users have difficulties in writing and reading diagrams (Agarwal & Sinha, 2003; Laitenberger, Atkinson, Schlich, & Emam, 2000) and tool venders have problems to fully support the UML standard. Furthermore, the maintenance of the standard is very expensive and error-prone, (e.g., Fuentes, Quintana, Llorens, Génova, & Prieto-Díaz, 2003 identified several hundred errors in UML's meta-model). Other authors criticize UML for its semantic inconsistency, construct ambiguity, notation inadequacy, and cognitive misdirection (Champeaux, 2003; France et al., 2006; Frank, 1998; Henderson-Sellers, 2002, 2005; McLeod, Halpin, Kangassalo, & Siau, 2001; Shen & Siau, 2003; Thomas, 2002; Wang, 2001).

On the other hand, UML is the de-facto standard for object-oriented modeling and an important milestone in software engineering. Modeling of software systems increases the degree of abstraction during system development tremendously. This change is similar to the replacement of assembly languages by high-level languages in the 1960s and 1970s. Today, high-level languages are not used in all but most domains. We predict that, in the future, UML has an analogous position as high-level languages have today. Hence, UML continues to play a major role in systems development.

REFERENCES

Agarwal, R., & Sinha, A. P. (2003). Object-oriented modeling with UML: A study of developers' perceptions. *Communications of the ACM, 46*(9), 248-256.

Atkinson, C., & Kühne, T. (2002). Rearchitecting the UML infrastructure. *ACM Transactions on Modeling and Computer Simulation, 12*(4), 290-321.

Berner, S., Glinz, M., & Joos, S. (1999). A classification of stereotypes for object-oriented modeling languages. In R. France & B. Rumpe (Eds.), *UML '99—The unified modeling language—Beyond the standard. Second International Conference, Fort Collins, CO, October 28-30, 1999* (Vol. 1723, pp. 249-264). Berlin: Springer.

Booch, G. (2002). Growing the UML. *Software and Systems Modeling, 1*, 157-160.

Bornberg-Bauer, E., & Paton, N. W. (2002). Conceptual data modelling for bioinformatics. *Briefings in Bioinformatics, 3*(2), 165-180.

Bruccoleri, M., Dieaga, S. N. L., & Perrone, G. (2003). An object-oriented approach for flexible manufacturing control systems analysis and design using the unified modeling language. *The International Journal of Flexible Manufacturing Systems, 15*, 195-216.

Cengarle, M. V., & Knapp, A. (2004). OCL 1.4/5 vs. 2.0 Expressions—Formal semantics and expressiveness. *Software and Systems Modeling, 3*, 9-30.

Champeaux, D. d. (2003). Extending and shrinking UML. *Communications of the ACM, 46*(3), 11-12.

Dahanayake, A. (2003). Methodology evaluation framework for component-based system development. *Journal of Database Management, 14*(1), 1-26.

Dobing, B., & Parsons, J. (2005). How the UML is used. *Communications of the ACM*, to appear.

Dolk, D. R. (2000). Integrated model management in the data warehouse era. *European Journal of Operational Research, 122*, 199-218.

Douglass, B. P. (2004). *Real-time UML: Developing efficient objects for embedded systems* (3rd ed.). Reading, MA: Addison-Wesley.

Engels, G., Heckel, R., & Sauer, S. (2001). UML - A Universal Modeling Language? In M. Nielsen & D. Simpson (Eds.), *Application and theory of Petri Nets 2000: 21st International Conference, ICATPN 2000 , June 2000, Aarhus, Denmark* (pp. 24-38). Berlin: Springer.

Erickson, J., & Siau, K. (2004). Theoretical and practical complexity of unified modeling language: A Delphie study and metrics analyses. *International Conference on Information Systems 2004* (pp. 183-204). Washington, DC.

Eriksson, H.-E., & Penker, M. (2000). *Business modeling with UML—Business patterns at work*. New York: John Wiley & Sons.

Fettke, P., & Loos, P. (2003a). Classification of reference models—A methodology and its application. *Information Systems and e-Business Management, 1*(1), 35-53.

Fettke, P., & Loos, P. (2003b). Specification of Business Components. In R. Unland (Ed.), *Objects, components, architectures, services, and applications for a networked world—International Conference NetObjectDays, NODe 2002, Erfurt, Germany, October 7-10, 2002, Revised Papers* (Vol. 2591, pp. 62-75). Berlin: Springer.

Fettke, P., Loos, P., & Zwicker, J. (2007). Using UML for reference modeling. In P. Rittgen (Ed.), *Enterprise modeling and computing with UML* (pp. 174-205). Hershey, PA: Idea Group.

Firesmith, D., Henderson-Sellers, B., & Graham, I. (1998). *OPEN modeling language (OML)—Reference manual*. Cambridge, UK: Cambridge University Press.

Fowler, M. (2004). *UML distilled—A brief guide to the standard object modeling language* (3rd ed.). Boston: Addison-Wesley.

France, R. B., Ghosh, S., Dinh-Trong, T., & Solberg, A. (2006). Model-driven development using UML 2.0: Promises and pitfalls. *Computer, 39*(2), 59-66.

Frank, U. (1997). *Towards a standardization of object-oriented modelling languages?* Working Paper No. 3. Koblenz, Germany: Institut für Wirtschaftsinformatik der Universität Koblenz Landau.

Frank, U. (1998). Object-oriented modelling languages: State of the art and open research questions. In M. Schader & A. Korthaus (Eds.), *The unified modeling language: technical aspects and applications* (pp. 14-31). Heidelberg: Physica.

Frankel, D. S. (2003). *Model driven architecture—Applying MDA to enterprise computing*. Indianapolis, IN: Wiley.

Fuentes, J. M., Quintana, V., Llorens, J., Génova, G., & Prieto-Díaz, R. (2003). Errors in the UML metamodel? *ACM SIGSOFT Software Engineering Notes, 28*(6), 1-13.

Glezer, C., Last, M., Nachmany, E., & Shoval, P. (2005). Quality and comprehension of UML interaction diagrams—An experimental comparison. *Information and Software Technology, 47*(10), 675-692.

Grossman, M., Aronson, J. E., & McCarthy, R. V. (2005). Does UML make the grade? Insights from the software development community. *Information and Software Technology, 47*(6), 383-397.

Henderson-Sellers, B. (2002). The use of subtypes and stereotypes in the UML model. *Journal of Database Management, 13*(2), 43-50.

Henderson-Sellers, B. (2005). UML—The good, the bad or the ugly?—Perspectives from a panel of experts. *Software and Systems Modeling, 4*, 4-13.

Kobryn, C. (1999). UML 2001: A standardization odyssey. *Communications of the ACM, 42*(10), 29-37.

Kobryn, C. (2000). Modeling components and frameworks with UML. *Communications of the ACM, 43*(10), 31-38.

Kobryn, C. (2002). Will UML 2.0 be agile or awkward? *Communications of the ACM, 45*(1), 107-110.

Kobryn, C. (2004). UML 3.0 and the future of modeling. *Software and Systems Modeling, 3*, 4-8.

Laitenberger, O., Atkinson, C., Schlich, M., & Emam, K. E. (2000). An experimental comparison of reading techniques for defect detection in UML design documents. *Journal of Systems and Software, 53*(2), 183-204.

McLeod, G., Halpin, T., Kangassalo, H., & Siau, K. (2001). UML: A critical evaluation and suggested future. *34th Hawaii International Conference on System Sciences*, Hawaii.

Medvidovic, N., Rosenblum, D. S., Redmiles, D. F., & Robbins, J. E. (2002). Modeling software architectures in the unified modeling language. *ACM Transactions on Software Engineering and Methodology, 11*(1), 2-57.

Mellor, S. J., & Balcer, M. J. (2002). *Executable UML: A foundation for model-driven architecture*. Boston: Addison Wesley.

Miller, J. (2002). What UML should be. *Communications of the ACM, 45*(11), 67-69.

OMG. (2005a). *Unified modeling language: Diagram Interchange, Version 2.0, ptc/05-06-04*. Needham.

OMG. (2005b). *Unified modeling language: Superstructure, Version 2.0, formal/05-07-04*. Needham.

OMG. (2006a). *Object constraint language, Version 2.0, formal/06-05-01*. Needham.

OMG. (2006b). *UML tools*. Retrieved July 7, 2006, from http://www.omg.org/uml

OMG. (2006c). *Unified modeling language: Infrastructure, Version 2.0, formal/05-07-05*. Needham.

Opdahl, A. L., & Henderson-Sellers, B. (2002). Ontological evaluation of the UML using the Bunge-Wand-Weber model. *Software and Systems Modeling, 1*(1), 43-67.

Robbins, J. E., & Redmiles, D. F. (2000). Cognitive support, UML adherence, and XMI interchange in Argo/UML. *Information and Software Technology, 42*, 79-89.

Rumbaugh, J., Jacobson, I., & Booch, G. (1998). *The unified modeling language reference manual*: Addison-Wesley.

Saleh, K. (2002). Documenting electronic commerce systems and software using the unified modeling language. *Information and Software Technology, 44*(5), 303-311.

Shen, Z., & Siau, K. (2003). An empirical evaluation of uml notational elements using a concept mapping approach. *International Conference on Information Systems (ICIS)*, Seattle, Washington, USA, 194-206.

Siau, K., & Cao, Q. (2001). Unified modeling language (UML)—A complexity analysis. *Journal of Database Management*, *12*(1), 26-34.

Thomas, D. (2002). UML—Unified or universal modeling language? *Journal of Object Technology, 2*(1), 7-12.

Wand, Y., & Weber, R. (2002). Research commentary: Information systems and conceptual modeling—A research agenda. *Information Systems Research*, *13*(4), 363-377.

Wang, S. (2001). Experiences with the unified modeling language (UML). *Seventh Americas Conference on Information Systems (AMCIS) 2001* (pp. 1289-1293).

Warmer, J. v., & Kleppe, A. (2003). *The object constraint language—Getting your models ready for MDA* (2nd ed.). Boston: Addison-Wesley.

KEY TERMS

Conceptual Modeling: an action describing a domain with the help of a particular artificial or formalized language.

Meta-Model: A meta-model is a model of model.

Methodology: A consistent and suited set of modeling methods providing procedures to apply the constructs of a modeling language.

Model: A model is a particular product of conceptual modeling. It is a description of a domain using a particular language.

Object-Oriented Analysis and Design (OOA & OOD): Software engineering approach to construct software systems by building object-oriented models that abstract key aspects of the target system.

Object-Oriented Programming (OOP): Object-oriented programming emphasizes the hiding or encapsulation of the inner state of objects and the specification of these objects by an interface. OOP languages support objects, classes and inheritance.

Reference Model: A reference model is a model representing a class of domains, e.g. a reference model for production planning and control systems. It is a conceptual framework or blueprint for system's development.

A University/Community Partnership to Bridge the Digital Divide

David Ruppel
The University of Toledo, USA

Cynthia Ruppel
The University of Alabama in Huntsville, USA

INTRODUCTION

Companies want employees with core values who ascribe to corporate values. Emotional intelligence (EQ) is used by companies in recruitment (Foote, 2001), and guides managers in dealing with team performance problems. Similarly, leadership requires refocusing on core values, which over time builds character (Badaracco, 1998). Thus, educational institutions should devote considerable attention to character building (Foote, 2001).

Service-learning is designed to help. Jacoby (1996a, p. 5) has defined service-learning as "…a form of experiential education in which students engage in activities that address human and community needs together with structured opportunities intentionally designed to promote student learning and development".

Service-learning is important in information technology where students need technical skills and experience, and a strong ethical foundation. Legal aspects of technology have not kept pace with technology; often IT people are confronted with complex ethical decisions. It has been argued that service-learning represents a "unique pedagogy…that enhances the ability of private sector managers to be effective stewards of society's moral authority" (Godfred, p. 364). Service-learning in colleges is tightly linked with K-12 education (Jacoby, 1996B) due to the growing number of at-risk children, a vested interest for colleges to improve the future students, and because students will view service-learning as an appropriate college activity if they benefited from it prior to college (Jacoby, 1996b).

A policy concern in the information age is the "digital divide," a gap between those who have easy access to technology and those who do not. References are made to information "haves" and "have-nots" in an age where information is equivalent to wealth (Holloway, 2000). The "have-nots" are in danger of exclusion from the new economy and marginalization into low-wage jobs (Dunham, 1999). In 2000, the President of the United States asked the IT community to help close this digital divide for moral reasons and to ensure that the economy flourishes with the availability of skilled workers (Shewmake, 2000).

This overview summarizes a five-phase service-learning project accomplished through a partnership between the University of Toledo and a local K-8 parochial/non-profit school. The students were primarily enrolled in a Systems Analysis, Design and Implementation course (SAD). This longitudinal project was undertaken to plan, design, and wire a network for the school and to assess and implement continuing and future computer needs. It allowed students to gain "real-life" experience while contributing to the growth of IT among children in a non-profit setting.

BACKGROUND

The school is a parochial school enrolling approximately 200-250 students. All grades have a dedicated classroom; a computer lab and library are also provided.

Existing computers consisted of a classroom set of older Macintosh computers in the 8th grade room. Each classroom had an older Macintosh computer for the teacher, all with unused LAN capability. The computer lab contained older Apple computers used in the primary grades for computer literacy and keyboarding skills.

Phase 1

The school had accumulated technology funds and hired a teacher with a Master's degree in Educational Technology. The teacher and principal agreed to participate in the project since an estimate from a local company exceeded the funds accumulated. While the teacher had pedagogic knowledge of computers, he did not possess the expertise to evaluate the quotation or analyze the technical aspects of the network. The school indicated that it hoped to apply for a grant, but needed technical information.

Students self-selected into the project: *The goal of the project was to educate themselves as to alternatives, costs and provide background information concerning networking to prepare a grant application.* They had the opportunity to examine the existing environment and interview stakeholders.

The instructor and students toured the building, including attic and closet locations where existing asbestos could not be disturbed. The stakeholders were asked to determine the number of "drops/connections" required in each room based on immediate and future use. Two drops were requested in each classroom — one for the teacher's computer and another for a classroom network. The group submitted a plan to the school including alternatives, costs, and technical information for the design of the campus network to be used for grant preparation.

Phase 2

Phase 2 included completing a grant proposal and the physical networking of the building. The wiring project was popular among students and required instructor selection to participate. Two students had experience, while others were chosen based on enthusiasm and desire to learn the "hands-on" aspects of networking.

Using the plan, a grant proposal was submitted providing evidence of the school's commitment and a plan for the educational use of the network. The university students' involvement was documented, and the authors were listed as consultants.

The grant writing was divided among the authors, the teacher, and the principal. Guidelines required a technology plan and a specific grant request. The accumulated funds were sufficient to wire the building without grant funding. Subsequently the maximum grant was awarded for continuation of the project.

The wiring plan included an Ethernet LAN in the 8th grade room and a campus LAN with connections in all classrooms and offices. Microsoft Windows NT Server 4.0 with Services for Macintosh (SFM) was chosen as the network operating system (NOS) based on the requirements of the heterogeneous network. With SFM, the Windows server appears as an Apple server to Macintosh clients.

The NOS was installed on a computer with a dial-up Internet connection, a temporary arrangement until high-speed access was available. Proxy server and content-filtering services were installed, providing low-bandwidth Internet access to all clients in the 8th grade classroom. Secure storage for network users was provided.

In wiring the building, a storage closet was used as the wiring closet where hubs were installed and all building wiring runs terminated. Since the computers were regularly used, work was partitioned into elements that could be done over a weekend to maximize availability for students and teachers. A file server for administrative applications and intranet e-mail was installed in the wiring closet with a tape drive to provide network backup.

Phase 3

The next step was to install high-speed Internet access for which the school receives state funding. The authors recommended the installation of a T-1 line (1.544 Mbps).

In 2001 the network consisted of three servers, each running Microsoft NT Server 4.0. A laser printer was available to over 50 network clients running various operating systems.

Phase 4

We now needed to recommend replacements for outdated equipment and provide information for equipment grant applications. Students self-selected into this project, and met with the stakeholders. The main issue was whether the replacement equipment should be Apple or Windows-based. Pertinent factors were educational needs, existing equipment, and available expertise. The group analyzed these factors together with their survey of surrounding schools to determine local norms. Their recommendation was Apple equipment.

After network completion, it was obvious that the computers were outdated. An early goal was to minimize expenses by leveraging existing equipment. An obsolete Web browser was installed since the machines' capabilities prohibited current versions. While slow, the clients provided Internet access where none existed before. In today's world, fast access is a necessity. Revenaugh (2000) suggests that access does not equal equity; just connecting a wire to a school does not mean it will be used well. Also, the browser security certificates had expired, so secure Web sites could not be accessed.

Phase 5

Using this recommendation, a second grant proposal was prepared requesting the maximum award to purchase new computers for the teachers. The replaced computers will be placed in the classrooms in small networks for student use.

Another group of students worked on designing a Web site for the school. A preliminary school Web site, created 2 years earlier, was out-of-date and the students were asked to produce an up-to-date, easily maintainable Web site.

CONCLUSIONS AND LESSONS LEARNED

There must be a close working relationship between the community organization and the university faculty. Trust

was necessary to order materials, remain in the building after-hours, and obtain quotations on behalf of the school.

Since this is a functioning organization, work cannot be left unfinished. Work must be analyzed and divided into tasks that can be completed within the available time. Student self-selection should be combined with faculty approval since the faculty member(s) have the ultimate responsibility for project completion.

Students appreciate being part of a project that benefits others. Food was provided on Saturdays, motivating the students and keeping them on site. The students had time to relax and discuss the project among themselves and with the instructors informally as recommended by the service-learning literature (Godfrey, 1999; Kenworthy-U'ren, 1999). During this time it was interesting to note their perceptions of how elementary education has changed since "their day". They even became attached to the classroom pets – feeding them, and so forth.

The school purchased materials on time and provided significant access to the building. They valued the students' work, since they had received a commercial bid. We recommend this procedure where applicable to eliminate the perception that free means valueless.

Choose the students both to ensure project completion and to make it a learning experience. From the volunteers, students were selected to ensure some students had necessary skills and also included inexperienced students. This facilitated a sharing of knowledge and maximized learning, giving the experienced students leadership opportunities.

Expect unanticipated problems. Since the building had been built in stages (to our surprise), we needed to wire through old exterior walls. Also, the 8th grade class rabbit ate through the cables and had to be sent to a new home. We recommend exploring insurance coverage for students involved in labor with the potential for injury.

It is important to seek knowledge from a variety of sources. Experience with Apple computers was limited among the students and the instructors. A student, not part of this project, was familiar with Apples and made relevant suggestions. Another former student, who works with Apple computers in his employment, supplied information, agreed to troubleshoot computers, and created a CD containing the software set-up for the computers. He showed the students how to install the set-up and was so impressed with one of the students that he arranged a job interview for the student, resulting in a job offer.

In the service-learning literature, one caveat is the time required of the instructor(s), particularly relative to the current faculty reward systems (Godfrey, 1999; Kenworthy-U'ren, 1999; Kolenko, Porter, Wheatley & Colby, 1996). The amount of instructor time and nature of instructor tasks varied according to the type of student projects undertaken. This is consistent with the findings of Bush-Bacelis (1998), who suggests that service-learning projects do not require more work, but a different type of work both for the students and the instructor. The wiring project required a significant amount of dedicated instructor time, as did the grant writing. However, students indicated they found the wiring project particularly fulfilling, especially since the feedback was a working network and a tangible skill set was learned.

Student Assessment

In group projects, students are asked to assess each member's contribution to the group. To assess the quality grade for those completing the physical wiring, the instructors were present in the building. For the more traditional projects, the instructor assessed quality and adherence to SAD principles after obtaining "perceived value and usefulness" feedback from stakeholders.

Benefits to the School

The age of the existing computers limited their usability, and the school did not have the funds to replace them. By placing them on a network, their useful life was extended. Due to limited funds and lack of technical knowledge the school was prevented from completing these projects.

Benefits to the Students

Kenworthy-U'ren (1999) suggests that service-learning projects should be grounded in experiential learning, not merely a service-related activity. These projects allowed the students' involvement in projects related to their career objectives. The students appeared determined to provide a quality product because it was being given to a "real" organization for decision making and to the next group who would use their work and attempt to "better it". These projects, if carefully planned, are win-win situations for the students and the community partner.

FUTURE TRENDS

The necessity of networking schools is becoming increasingly important and at lower levels such as primary school also. To continue to eliminate the digital divide, students, even grade school students, must be able to access and evaluate information to learn to make informed decisions. This project was designed to both give students "hands-on" experience and to provide access to young students so they can become familiar and comfortable with this access.

The technology in providing this access will continue to change. Many devices today are wireless devices (as schools can afford to purchase them), which would eliminate the need to wire the building and would use more current technology.

However, in a school setting, due to cost constraints, the available technology appears to change more slowly than a business environment would. The issues involved in setting up a wireless network may seem less involved since no physical wiring must take place, but different complex issues arise, such as providing coverage to all areas and methods of extending the signal on large campuses. So the exercise for the students becomes no less important.

REFERENCES

Badaracco, J.L. (1998). The discipline of building character. *Harvard Business Review, 76*(2), 114-124.

Dunham, R.S. (1999, August 2). Across America, a troubling 'digital divide.' *Business Week, n3640,* 40.

Foote, D. (2001, February 12). What's your 'emotional intelligence'? *Computerworld*.

Godfrey, P.C. (1999). Service-learning and management education: A call to action. *Journal of Management Inquiry, 8*(4), 363-378.

Holloway, J.H. (2000). The digital divide. *Educational Leadership, 58*(2), 90-91.

Jacoby, B. (1996a). Service-learning in today's higher education. In B. Jacoby (Ed.), *Service learning in higher education: Concepts and practices* (pp. 3-25). San Francisco: Jossey-Bass Publishers.

Jacoby, B. (1996b). Securing the future of service-learning in higher education: A mandate for action. In B. Jacoby (Ed.), *Service learning in higher education: Concepts and practices* (pp. 317-335). San Francisco: Jossey-Bass Publishers.

Kenworthy-U'ren, A.L. (1999). Management students as consultants: An alternative perspective on the service-learning "call to action." *Journal of Management Inquiry, 8*(4), 379-387.

Kolenko, T.A., Porter, G., Wheatley, W., & Colby, M. (1996). A critique of service learning projects in management education: Pedagogical foundations, barriers, and guidelines. *Journal of Business Ethics, 15,* 133-142.

Shewmake, B. (2000). Clinton to IT execs: Help close digital divide. *InfoWorld, 22*(7), 12.

KEY TERMS

Bandwidth: A measure of the data transmission capacity of a communications link.

Digital Divide: The term digital divide describes the fact that the world can be divided into people who do and people who do not have access to - and the capability to use - modern information technology, such as the telephone, television, or the Internet.

Ethernet: A communications standard (IEEE 802.3) for Local Area Networks (LAN). When a device wishes to transmit, it waits until the link is empty and then transmits. In the event that two or more devices transmit simultaneously (collision), all devices stop transmitting and wait a random time period before attempting to retransmit.

Heterogeneous Network: A network where network clients run a variety of operating systems. An example would be a network of machines running Windows, Unix, and Mac OS X.

Intranet: Computer network contained entirely within an organization.

LAN: Local Area Network. Group of computers and other devices sharing a single communications link. Typically, the longest distance between any two connections is a few kilometers. Usually owned by a single organization.

Network Operating System: An operating system for a computer designated as a server in a LAN. Manages network resources available to network clients such as printers, files and databases. Responsible for maintaining a database of names and addresses of clients attached to the network.

Service-Learning: A form of experiential education in which students engage in activities that address human and community needs together with structured opportunities intentionally designed to promote student learning and development.

Systems Analysis and Design Course (SAD): A course in which the student learns to understand how an information system can support organizational needs, how to design the system, build it and deliver it to users.

T-1 Line: A communications link that can be used for digital data transmission. Provides a data rate (bandwidth) of 1.544 million bits per second (Mbps).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 29-32, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: Software & Systems Design

Updated Architectures for the Integration of Decision Making Support Functionalities

Guisseppi Forgionne
University of Maryland, Baltimore County, USA

INTRODUCTION

Information systems research continues to examine ways to improve support for decision making. The evolution from simple data access and reporting to complex analytical, creative, and artificially intelligent support for decision making persists (Holsapple & Whinston, 1996). In the evolution, existing information systems still, and new intelligent systems have been created to, provide the desired decision making support.

By studying the existing, and new, systems' characteristics, advantages, and disadvantages, researchers and practitioners can better design, develop, and implement robust decision making support systems (Kumar, 1999). The original article facilitated such study by presenting and illustrating the underlying information system architectures for robust decision making support (Forgionne, 2005).

This article updates the original by offering additional contributions to the subject. New literature on intelligent decision making support is examined, and the relevant findings are discussed. The title has been modified slightly to reflect the updates.

BACKGROUND

Several frameworks have been developed to describe the human decision making process. The most popular is Simon's three-phase paradigm of intelligence, design, and choice (Simon, 1960). This paradigm seems to be the most general, implying virtually all other proposed frameworks, and the Simon paradigm appears to have best withstood empirical testing (Martinsons, Davison, & Tse, 1999). Such scrutiny, however, has suggested the expansion of the basic formulation to conclude with an implementation phase.

During the intelligence phase, the decision-maker observes reality, gains a fundamental understanding of existing problems or new opportunities, and acquires the general quantitative and qualitative information needed to address the problems or opportunities. In the design phase, the decision-maker develops a specific and precise model that can be used to systematically examine the discovered problem or opportunity. This model will consist of decision alternatives, uncontrollable events, criteria, and the symbolic or numerical relationships between these variables. Using the explicit models to logically evaluate the specified alternatives and to generate recommended actions constitute the ensuing choice phase. During the subsequent implementation phase, the decision maker ponders the analyses and recommendations, weighs the consequences, gains sufficient confidence in the decision, develops an implementation plan, secures needed financial, human, and material resources, and puts the plan into action.

A variety of individual information systems have been offered to support the decision-making phases and steps. Much can be learned about this support by examining the individual systems' components, architectures, and operations.

RENDERING EFFECTIVE DECISION MAKING SUPPORT

Issues, Controversies, and Problems

Decision making support has evolved over time and across disciplines (Mirchandani & Pakath, 1999). Initial support was offered by a decision support system (DSS). In the typical DSS, the decision maker utilizes computer technology to: (a) organize the data into problem parameters, (b) attach the parameters to a model, (c) use the model to simulate (experiment with) alternatives and events, and/or (d) find the best solution to the problem. Results are reported as parameter conditions (status reports), experimental forecasts, and/or recommended actions. Feedback from the user-controlled processing guides the decision maker to a problem solution, and created information and knowledge are stored as additional inputs for future or further processing.

The DSS concept presumes that the problem pertinent data and models have been created and stored in the system prior to use (Hooghiemstra, Kroon, Odijk, Salomon, & Zwaneveld, 1999). This concept also assumes that the user can utilize the computer technology to perform the technical processing operations and computations required by the system (Lawrence & Sim, 1999). In fact, DSS users rarely have the technical skill to recognize, capture, and process pertinent data and models or to interpret the results of the models' processing within the problem context (Raghunathan, 1999). In short, the DSS concept offers little direct support

for the intelligence, early design, and implementation phases of decision making.

To be useful, problem pertinent data must be identified, located, captured, stored, accessed, and interpreted (Seely & Targett, 1999). Data warehousing can facilitate access and reporting, while data mining can help with the interpretation function. An executive information system (EIS) can deliver these data access, reporting, and interpretation functions to the decision maker in an intuitive and appealing manner.

In a typical EIS, the decision maker utilizes computer technology to: (a) organize the data into specified broad categories, (b) view (slice and dice) the data from interesting perspectives, (c) generate "warnings" for the decision maker by scanning current trends, and (d) mine the data for less obvious relationships. Results are reported as category summaries (status reports), sliced and diced details (drill down reports), and/or suggested problem parameters (events). Feedback from the user-controlled processing guides the decision maker to a general problem understanding, and the created parameters are stored as additional inputs for further processing.

The user should exit EIS processing with a general understanding of the problem or opportunity and with relevant problem information (such as general objectives, range of decision alternatives, and range of pertinent events). Since additional decision analysis will be required to explicitly formulate the problem and complete the decision making process, an EIS directly supports only the intelligence phase of decision making.

Technical and domain expertise will be needed to recognize, formulate, and solve most complex and significant decision problems or opportunities. Although such expertise will be available within, and outside, an organization, the expertise may be difficult, costly, and time-consuming to locate, access, and utilize. Often, the corresponding knowledge can be acquired, embedded within a knowledge-based system (KBS), and the system can be used to capture, store, and deliver the expertise to the decision maker (Ayyub, 2001). A typical KBS captures and stores as inputs problem pertinent knowledge, either from experts, cases, or other sources, and the models (inference engine or reasoning mechanisms) needed to draw problem solution inferences from the knowledge. In the process, a KBS directly facilitates problem structuring (thereby supporting part of the design phase), the selection of alternatives, and the evaluation of the selections (hence supporting the choice phase of decision making).

Since decision making is a sequential and continuous process, learning will be essential to the successful completion of the process. Users will learn from their interactions with a KBS (or other individual decision making support system) and, in the process, gain skills that can be applied to further decision making tasks. Applying learning to the solution of the current problem, however, often will require system support (Bolloju, Khalifa, & Turban, 2002). Machine-learning systems (MLS) can provide such support by mimicking the learning processes of physical systems. In a typical MLS, the decision maker utilizes computer technology to: (a) organize the problem data, (b) structure (operationalize) the learning model, and (c) simulate learning. Results are reported as problem conditions (status reports), forecasted problem outcomes, and/or and an explanation of the learning logic.

Besides learning, creativity often is needed to successfully complete the decision making process (Keys, 2000). While the previous systems free decision makers to concentrate on the creative aspects of decision making, they do not provide direct support for the creative process (Savransky, 2001). Since decision makers may not be inherently creative, support for creativity can considerably enhance their decision making. A creativity enhancing system (CES) offers such support (Forgionne, Clements, & Newman, 1995). In a typical CES, the decision maker utilizes computer technology to: (a) organize (chiefly categorize and classify) problem ideas and concepts, (b) structure ideas and concepts into problem elements and relationships, and (c) simulate conceptual problem solutions. Results are reported as problem elements (status reports), the problem's conceptual structure (criteria, alternatives, events, and relationships), and/or forecasted outcomes from the conceptual analyses.

The major individual systems, and their primary and direct support, are summarized in Table 1. An examination of this table shows that none of the individual systems offers complete and integrated support for all phases and steps of the decision making process.

Table 1. Individual decision making support systems

System	Type	Support
Decision Support System	Individual	Specifying relationships between criteria, alternatives, and events; choice
Executive Information System	Individual	Intelligence; developing decision criteria; identifying relevant uncontrollable events
Knowledge-Based System	Individual	Develop decision alternatives; choice
Machine-Learning System	Individual	Logically evaluate decision alternatives
Creativity Enhancing System	Individual	Design; develop an implementation plan; put implementation plan into action

Table 2. Integrated decision making support systems

System	Type	Support
Intelligent Decision Support System (IDSS)	Integrates the functions of DSS and KBS (and/or MLS)	Developing decision alternatives; specifying relationships between criteria, alternatives, and events; choice
Executive Support System (ESS)	Integrates the functions of DSS and EIS	Intelligence; developing decision criteria; identifying relevant uncontrollable events; specifying relationships between criteria, alternatives, and events; choice
Whole-Brained Decision Support System (WDSS) and Group Decision Support System (GDSS)	Integrates the functions of DSS and CES	Gain problem/opportunity understanding; design; choice
Management Support System (MSS)	Integrates the functions of DSS, EIS, and KBS (and/or MLS)	Intelligence; design; choice

SOLUTIONS AND RECOMMENDATIONS

The need for complete and integrated decision making support has encouraged researchers to seek the synergistic effects that can be achieved by combining the functionalities of the individual systems. The result has been the development of the various integrated systems for decision making support summarized in Table 2.

Of particular note is the intelligent decision support system concept. In a KBS, knowledge is captured and stored in a well structured manner prior to system use. The same approach is used to capture and store the learning process within a MLS. This static approach works well when the decision involves a one-time problem, where a current action will be unaffected by previous choices, and have no influence on subsequent selections. When the problem is instead sequential, the decision maker must make a series of interrelated decisions over time or across decision stages. Under such circumstances, the system's knowledge and models must be restructured dynamically to reflect the new or changing reality and in real time to offer just in time decision making support (Gupta, Forgionne, & Mora, 2006).

To provide this dynamic intelligence, a new IDSS, called the intelligent just in time decision support system (IJDSS), has been offered (Conteh & Forgionne, 2005). In the IJDSS, a model, formed from technical and domain expertise, captures and stores problem pertinent knowledge and the explicit problem structure needed to evaluate decision alternatives. As the decision environment changes, the model is updated dynamically and in real time to reflect the new reality. The dynamically evolving model is used to simulate (experiment with) alternatives and events or find the best solution to the problem, with or without guidance from the system. Results are reported as parameter conditions (status reports), experimental forecasts, recommended actions, and explanations for the recommendations.

As Table 2 indicates, each integrated system, such as an IDSS or MSS, integrates the functionality of particular individual systems to provide decision making support. Even the IJDSS form of an IDSS has this characteristic. While the integrated functionality has created more complete and unified decision making support, the suggested synergies still leave significant gaps in decision making support. For example, an IDSS (or IJDSS) still leaves gaps in design support, while an MSS does not provide creativity support. With even more system choices available than previously, the decision maker and/or staff are forced to match the relevant functionality with his/her/their decision making support needs. Decision makers, and/or staff, may be ill equipped to make these selections and design, build, and implement the desired system.

An alternative strategy is to create one decision making support system that synthesizes the main features and functions for the various decision making support systems. The decision technology system (DTS), which is presented in Figure 1, has been proposed to support this alternative strategy. As Figure 1 illustrates, a DTS has a database, knowledge base, and model base. The database contains the data directly relevant to the decision problem, including the values for the uncontrollable events, decision alternatives, and decision criteria. The knowledge base holds problem knowledge, such as formulas for converting available data into the problem's parameters, guidance for selecting decision alternatives and problem relationships, or advice in interpreting possible outcomes. The model base is a repository for the formal (tabular, graphic, conceptual, or mathematical) models of the decision problem and the methodology for developing results (simulations and solutions) from the formal models.

Decision-makers utilize computer technology (hardware and software) to process the inputs into problem-relevant outputs. Processing will involve:

a. **Organizing Problem Parameters:** Accessing the data base, extracting the decision data, and organizing the

information in the form needed by the solution model and methodology;

b. **Structuring the Decision Problem:** Accessing the model base, retrieving the appropriate decision model, and operationalizing (attaching organized parameters to) the decision model;
c. **Simulating Policies and Events:** Using the operationalized decision model to perform the computations needed to simulate outcomes from user-specified alternatives and then identifying the alternative (or alternatives) that best meets the decision criterion (or criteria) among those tested;
d. **Finding the Best Problem Solution:** Accessing the model base, retrieving the appropriate solution method, and using the retrieved method to systematically determine the alternative (or alternatives), among all possible alternatives, that best meets the decision criterion (or criteria).

The DTS can use problem ideas, concepts, and knowledge drawn from the knowledge base to assist users in performing these processing tasks.

Processing will generate status reports, forecasts, recommendations, and explanations. The status reports will identify relevant uncontrollable events, decision alternatives, and decision criteria and show the current values for these problem elements. Forecasts will report the events and alternatives specified in the simulations and the resulting projected values of the decision criteria. The recommendations will suggest the values for the decision alternatives that best meet the decision criteria, and the corresponding criteria values, under current and forecasted values for the uncontrollable events. Explanations will justify the recommendations and offer advice on further processing. Such advice may include suggestions on interpreting the output and guidance for examining additional scenarios.

Input feedback from the processing provides additional data, knowledge, and models that may be useful for future decision making. This feedback is provided dynamically to update the models and parameters in real time within external intervention. Output feedback (which can include outcomes, cognitive information, task models, and what-if, goal-seeking, and other types of sensitivity analyses) is used to extend or modify the original analyses and evaluations. Often, computer-based system agents will be used to provide the dynamic and real time input and output feedback for the user.

Figure 1's general DTS architecture can support all phases of the decision making process in a complete, integrated, and continuous manner. Critical problem data can be captured in a DTS database, and the system can be used to organize this captured information, generate timely focused reports, and project trends. Such processing helps the decision maker to quickly monitor the decision environment, set objectives, and evaluate the processed information for opportunities or problems, thereby supporting the intelligence phase of decision making.

The DTS, augmented by the managers' (or perhaps staff) insights and judgments, can be used to process captured constructs and frameworks into criteria, events, and alternatives needed to formulate a model of the decision problem. Additional processing with the captured statistical methodologies can estimate the parameters required to operationalize the formulated decision problem model, thereby supporting the design phase of decision making. The formulated models, again augmented by the managers' insights and judgments, are used to evaluate alternatives in a systematic and analytic fashion and to recommend alternatives, thereby supporting the choice phase of decision making.

Decision technology systems can provide the analyses in vivid detail with tables, graphs, and other supporting material. Such supporting material will increase the decision maker's confidence in the recommendations, improve the decision maker's perception of support system effectiveness, and enable the decision maker to better explain, justify, and communicate the decisions during implementation, thereby supporting the implementation phase of decision making.

Along with the original analyses and evaluations, DTS feedback loops increase the users' confidence in the recommendations and enable the decision-maker to better explain, justify, and communicate the decisions during implementation. The loops also support decision making in a continuous and dynamic manner (Forgionne, 1999, 2000).

FUTURE TRENDS

Realizing the DTS promise presents significant technical and management challenges. Problem pertinent data, models, and knowledge must be identified, located, retrieved, and captured (Balasubramanian, Nochur,, Henderson, & Kwan, 1999). Intelligent data warehousing and mining can support the data retrieval tasks, and it may be possible to adapt these methodologies for model and knowledge retrieval support. Differences in data, knowledge, and model structures, however, may necessitate the development of new methodologies for knowledge and model retrieval tasks.

Also, it will be challenging to collect and deliver the tools and to manage the design, development, and implementation effort. Agents and object-oriented methods can be used to capture the tools and make them available for system operation in the dynamic and real time manner suggested by the IJDSS concept (Gupta et al., 2006; Siskos & Spyridakos, 1999). The resulting system, however, will profoundly challenge the nature of the decision maker's work as well as altering the structure of the organization. By providing complete

and integrated decision making support, a DTS will enable the decision maker to perform technical tasks previously outsourced to specialists. The result may be an organization with fewer hierarchical levels and smaller staffs.

CONCLUSION

Over the years, support for decision making has taken a variety of forms. As the forms have evolved, decision making support has become more comprehensive and integrated. Today, there are many system choices available, and matching the appropriate system to the particular problem or opportunity has created a new task for management.

The evolution has illustrated the synergistic value that can be achieved through higher levels of functional integration. This article has presented a concept, the DTS, that can offer a mechanism to consolidate the advances and promote a revolution in management. The proposed system has also created significant research opportunities—determining the best integration strategy, identifying the best design and development tools to achieve the strategy, and examining the impact of integrated decision support on management, decision making, and organizational structure, among others. It also clarifies the needs to: (a) have effective and efficient information reporting and communication systems in place and (b) to integrate the decision making support systems with information reporting and communication systems.

REFERENCES

Ayyub, B. M. (2001). *Elicitation of expert opinions for uncertainty and risks*. Andover, UK: CRC Press.

Balasubramanian, P., Nochur, K., Henderson, J. C., & Kwan, M. M. (1999). Managing process knowledge for decision support. *Decision Support Systems, 27*(1-2), 145-162.

Bolloju, N., Khalifa, M., & Turban, E. (2002). Integrating knowledge management into enterprise environments for the next generation decision support. *Decision Support Systems, 33*(2), 163-176.

Conteh, N., & Forgionne, G. A. (2005). A just in time approach to intelligent decision support systems. *Journal of Decision Systems, 14*(1-2), 39-54.

Forgionne, G. A. (1999). An AHP model of DSS effectiveness. *European Journal of Information Systems, 8*, 95-106.

Forgionne, G. A. (2000). Decision-making support system effectiveness: The process to outcome link. *Information Knowledge Systems Management, 2*(2), 169-188.

Forgionne, G. A. (2005). Functional integration of decision making support. In M. Khosrow-Pour (Ed.), The encyclopedia of information science and technology (pp. 1236-1242). Hershey, PA: Idea Group Reference.

Forgionne, G. A., Clements, J. P., & Newman, J. (1995). Qualitative thinking support systems (QTSS). *Journal of Decision Systems, 4*(2), 103-137.

Gupta, J. N. D., Forgionne, G. A., & Mora, M. (2006). *Intelligent decision-making support systems (I-DMSS): Foundations, applications, and challenges*. London: Springer.

Holsapple, C. W., & Whinston, A. B. (1996). *Decision support systems: A knowledge-based approach*. New York: ITP.

Hooghiemstra, J. S., Kroon, L. G., Odijk, M. A., Salomon, M., & Zwaneveld, P. J. (1999). Decision support systems support the search for win-win solutions in railway network design. *Interfaces, 29*(2), 15-32.

Keys, P. (2000). Creativity, design and style in MS/OR. *Omega, 28*(3), 303-312.

Kumar, R. L. (1999). Understanding DSS value: An options perspective. *Omega, 27*(3), 295-304.

Lawrence, M., & Sim, W. (1999). Prototyping a financial DSS. *Omega, 27*(4), 445-450.

Martinsons, M., Davison, R., & Tse, D. (1999). The balanced scorecard: A foundation for the strategic management of information systems. *Decision Support Systems, 25*(1), 71-88.

Mirchandani, D., & Pakath, R. (1999). Four models for a decision support system. *Information & Management, 35*(1), 31-42.

Raghunathan, S. (1999). Impact of information quality and decision-maker quality on decision quality: A theoretical model and simulation analysis. *Decision Support Systems, 26*(4), 275-286.

Savransky, S. D. (2001). *Engineering of creativity: Introduction to TRIZ methodology of inventive problem solving*. Andover, UK: CRC Press.

Seely, M., & Targett, D. (1999). Patterns of senior executives' personal use of computers. *Information & Management, 35*(6), 315-330.

Simon, H. (1960). *The new science of management decision.* New York: Harper and Row.

Siskos, Y., & Spyridakos, A. (1999). Intelligent multiple criteria decision support: Overview and perspectives. *European Journal of Operational Research, 113*(2), 236-246.

KEY TERMS

Artificially Intelligent Systems: Information systems that help users manage data and models by delivering virtual expertise and other forms of artificial intelligence in support of these tasks.

Creativity Enhancing Systems: Information systems that are designed to offer creative tools that help users formulate problems and perform other creative tasks in decision making.

Decision Making Process: The process of developing a general problem understanding, formulating the problem explicitly, evaluating alternatives systematically, and implementing the choice.

Decision Support Systems: Information systems that interactively support the user's ability to evaluate decision alternatives and develop a recommended decision.

Decision Technology Systems: Information systems that are designed to support all phases of the decision making process in a complete and integrated manner.

Executive Information Systems: Information systems that access, report, and help users interpret problem pertinent information.

Integrated Decision Making Support Systems: Information systems that integrate the functions of one or more individual (stand-alone) decision making support systems.

Intelligent Just-in-Time Decision Support Systems: Information systems that provide expertise just in time to support the decision making process.

Knowledge-Based Systems: Information systems that capture and deliver problem pertinent knowledge to users.

Machine-Learning Systems: Information systems that mimic the human learning process and deliver the knowledge to users.

Usability Engineering of User-Centered Web Sites

Theresa A. O'Connell
National Institute of Standards and Technology, USA

Elizabeth D. Murphy
U.S. Census Bureau, USA

INTRODUCTION

For Web sites to succeed, they must be user-centered. A user-centered focus throughout Web site development life cycles promotes Web site usability. This is accomplished through usability engineering carried out within the context of software engineering.

BACKGROUND

Starting with an understanding of users, and proceeding in concert with software engineering, usability engineering promotes user success and satisfaction. In this section, we introduce principal concepts of Web site usability engineering.

Users

In the context of this chapter, users are people who interact with Web sites. The term *user* is restricted to the intended users of a Web site. It excludes usability engineers (UEs), the site's providers, and others who have any stake in the Web site. Users of Web sites differ across many dimensions, for example, age, gender, technology experience, intellectual or aesthetic preferences, interaction styles, and abilities.

Usability

A definition of usability, from the International Organization for Standardization (ISO), underlies UEs' focus on users' needs and their goal of meeting users' needs through usability engineering. ISO defines usability in specific contexts of use: efficiency, effectiveness, and user satisfaction (ISO, 1998). Efficiency and effectiveness are components of user success. Satisfaction is an equal factor for usability: Usability = user success + user satisfaction.

The ISO definition implies that usable software must be accessible to users with special needs. Accessibility is a subdomain of usability in which users have physical and/or cognitive disabilities. Accessibility enables people with disabilities to experience success and satisfaction with software to a degree comparable to that enjoyed by people without disabilities (W3C, 2005).

User Participation in Web Site Development

Usability engineering relies on close interaction with users at strategic points where their input is crucial. Techniques span the software development life cycle. Examples include focus groups, interviews, surveys, design discussions, and observing as users interact with prototype Web sites. Collection of human performance data is key to evaluating users' success. Typical measures of human performance include accuracy and speed of task completion. These measures of user success complement, and often contrast with, self-reported ratings of user satisfaction.

Mental Models

From experiencing computers and Web sites, users build mental models, that is, psychological representations of the ways in which computers and Web sites work (Carroll, 1990; Johnson-Laird, 1983). Highly experienced users have mental models of different categories of Web sites, for example, entertainment and informational sites. Novice users, however, may not have differentiated their mental model of Web sites into unique categories. UEs help designers make user-interface design consistent with frequent users' expectations from prior experience with other Web sites. This consistency helps novices develop expectations as a basis for forming mental models that will apply across other Web sites. This is not to imply, however, that all user groups will have congruent mental models. Realistically, the mental models of different users and user groups will have some commonalties, yet, at the same, time exhibit many individual differences.

USER INTERFACE

A user interface (UI) is software that people use to interact with technology. The UI encompasses more than what users see or hear. In the broadest sense, the UI is the virtual place where the user's mental model meets the designers' system model (Bolt, 1984.) Aligning these models is a goal of usability engineering (Norman & Draper, 1986).

Usability Engineering

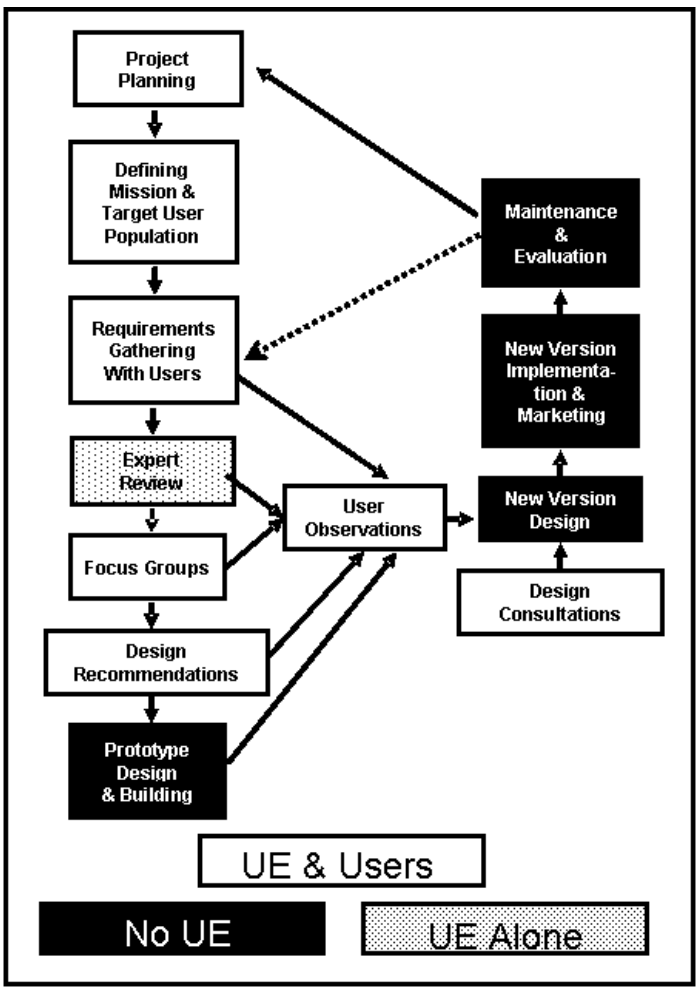
Usability engineering is a set of defined, user-centered processes grounded in research-based principles. Its purpose is to raise the potential for user success and satisfaction and, thereby, to support Web site providers' goals. UEs must understand a complex set of user variables to promote user success and satisfaction.

UEs' work is user-oriented, but usability engineering goes beyond user advocacy. UEs must also understand organizational and project goals. They must work within the structure of software development life cycles.

Usability Principles

With roots in human factors, usability engineering also draws on disciplines such as software engineering, linguistics,

Figure 1. A software development life cycle showing UEs interacting with users. Cited resources demonstrate how some usability engineering processes can be integrated into a software development life cycle (Addelston & O'Connell, 2005, Hix & O'Connell, 2005, Mayhew, 1992, O'Connell & Murphy, 2007). Next, we focus on usability engineering activities in four life-cycle critical iterative process areas: project planning, requirements definition, design, and evaluation/testing.



biology, cognitive psychology, information technology, and graphic design. These fields' diverse contributions converge in research-based usability principles and widely accepted guidelines for achieving usability (e.g., Koyani, Bailey, Nall, Allison, Mulligan, Bailey, & Tolson, 2003; Mayhew, 1992; Shneiderman & Plaisant, 2004). Usability principles are constantly updated in response to new research and technologies (e.g., O'Connell & Choong, 2008).

Usability principles trace to human capabilities and limitations. For example, to accommodate users with color vision deficiencies, one usability principle tells us never to rely solely on color to convey meaning. Another usability principle warns designers not to overload human working memory. Two goals of user-interface design are to capitalize on human capabilities and to compensate for human limitations.

Software Engineering

Software engineering is the "application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software" (IEEE, 1990, p. 70). On large Web site projects, software engineers typically oversee both the technical and management aspects of development by following life cycles that stipulate, for example, a project's phases, methods, activities, inputs, outputs, milestones, documentation, and risk-mitigation strategies.

A shared understanding of users' attributes and needs underpins the collaboration necessary to incorporate user-centered processes into software engineering processes. Usability engineering processes are compatible with software engineering processes. If software engineering is to develop useful tools, it must incorporate usability engineering. Books could be written about multimillion and multibillion dollar software engineering projects that failed because of inattention to usability. It is entirely possible for software to fulfill functional requirements and not be usable.

INTEGRATING USABILITY ENGINEERING INTO SOFTWARE-ENGINEERING LIFE CYCLES

Life cycles are structured frameworks for software development activities. For usability engineering to be fully integrated into Web site development, its practices must be integrated into software development life cycles from the start. Engineering usability up front reduces the need for changes after programming because user-centered design (UCD) is more likely to capture users' needs. Results include increased productivity, flatter learning curves, longer and repeated visits, increased profits, and decreased costs (e.g., Bias & Mayhew, 2005; Kalin, 1999; Mayhew 1999). Software engineering and usability engineering life cycle processes occur in parallel with compatible activities and outputs. They share the goal of producing usable Web sites.

In a Web development life cycle, each activity has goals, inputs, processes, and products. Each activity passes output into subsequent activities. Usability engineering has corresponding activities for most software engineering activities. Within each activity, usability engineering and software engineering processes occur simultaneously. Activity sequence and frequency can vary. For Web sites, when all the activities have been completed, it is not unusual for the life cycle to start again, resulting in a frequently updated Web site that adapts to changing user needs. Next, we focus on usability engineering activities in four life-cycle critical iterative process areas: project planning, requirements definition, design, and evaluation/testing.

Project Planning

The project plan is a blueprint for all Web site development activities. If usability engineering and user involvement are not addressed from the start in the plan's schedule and budget, they have little chance of later inclusion.

As a result of usability engineering's unique contributions during planning, the Web site's design will consider the user; the life cycle will be hospitable to user-centered processes; and the project will have a framework not only for funding and scheduling, but also for building usability into the product (Murphy, Nichols, Anderson, Harley, & Pressley, 2001).

Requirements Definition

In software engineering, requirements definition is a set of processes that identify and document a Web site's goals in terms of how the site will fulfill its providers' vision by delivering information and functionality. Usability engineering adds the users' perspective to verify that users' needs and expectations are met.

Requirements definition brings UEs face-to-face with targeted users during interviews, focus groups, and observations. Close interaction with users is key because existing documentation often omits users' real-world practices. The Web requires designing for a wider range of users than a project can usually involve. Therefore, UEs rely on usability principles and knowledge of human capabilities and limitations within the context of Web use.

Usability engineering processes during requirements definition start by considering users and their needs within the context of the Web site's intended offerings. As other team members set performance and system-function requirements, UEs investigate whether proposed content will meet both providers' goals and users' needs. For an existing site,

UEs address what worked and what did not work for its users not only in terms of content, but also in terms of layout and navigation.

Requirements definition produces artifacts such as user profiles, user classes, and user task descriptions and, most importantly, usability requirements that become a check list for everything that must be accomplished to promote successful and satisfactory user experiences. Critically, the underlying information architecture of the site must be logical from the user's perspective. This may mean that more than one path should lead to the same content because different user groups conceptualize the content differently.

User Profiles and Classes

Requirements definition starts by creating user profiles, documenting user attributes such as computer literacy; subject matter expertise, and experience with site functionality; physical and cognitive capabilities and limitations; education; mental models; interaction styles; goals; and tasks–all factors that impact human computer interaction. In user class analysis, UEs allocate user profiles into groups according to shared attributes.

UEs track the findings of these processes to the site's intended content to assure that its presentation empowers users to achieve their goals at the site. For example, during user class definition, UEs specify groups of people with disabilities, associating needed assistive devices with the user group. Assistive devices may include screen readers, screen magnifiers, and alternative means of entering data.

User Task Descriptions

In behavioral and cognitive task analysis, UEs describe users' goals and tasks. UEs study existing documents and observe users to learn the steps and sequences they take to accomplish goals. Behavioral task analysis documents observable tasks such as receiving and inputting information. Cognitive task analysis documents users' mental transformations and decisions.

Task analyses specify steps and work flows, describing the user experience, beginning-to-end. For example, when users must fill out an online form, UEs identify necessary initial steps, such as receiving an information package in the mail, locating the site's URL, and locating the user's login number, all long before the user navigates to the form.

Use Cases

Use cases are an output of requirements definition. They are formal descriptions of ways a product can be used. For each module of a system, use cases describe common processes and their prerequisites; steps users and the system will take; and resultant changes. Use cases help to ensure that the system supports frequent processes, that processes are relatively straightforward, and that the system architecture reflects the process structure.

Use cases do not accommodate the fact that users use software in ways that developers do not expect, based on cues from the UI in the context of the moment. Use cases do not stress cognitive tasks; therefore, they do not always represent users' natural interaction behaviors. UEs add value to use cases by making user-centered recommendations that would be missed otherwise. For example, the UE will recommend avoiding having each screen accommodate only one use case. UEs assure that screens accommodate users' own work flows, not someone else's model of required interactions.

Usability Requirements

Web sites have functional, system-performance and usability requirements. Functional requirements define what a Web site is supposed to do. System performance is a measure of how well the Web site does what it is supposed to do. Usability requirements address user interaction with the Web site. Usability requirements stipulate, for example, that users can purchase a product, making no more than one error, and that the site must receive a satisfaction rating of seven out of a possible nine. Some usability requirements are yes/no, for example, the Web site will provide anchor links on every page at its top two levels. Others are quantified, for example, users will be able to navigate, with only one click, from a product description page to an order placement page. Measurable usability requirements become goals for assessing the design against users' needs during usability evaluation (Whiteside, Bennett, & Holtzblatt, 1990).

In setting usability requirements, UEs view the site's goals through the lens of users' needs. Consider a case where the site aims to enhance the organization's image by showing its Chief Technology Officer talking about a new technical approach. For users with hearing impairments, the UE introduces a requirement for captions.

Some requirements are standards driven. Standards can be defined by the site provider, for example, an in-house standard for displaying the company logo. Standards can come from groups such as ISO. Sometimes requirements are legally mandated. For example, the UIs of some US federal government sites must comply with Section 508 of the Rehabilitation Act of 1973, as amended in 1998 (US Government, 1998), which gives accessibility requirements.

A Web site can fulfill all functional and system-performance requirements, yet still not be usable. In the end, it is the user's experience of a Web site that determines whether the site has achieved the product vision; so, usability engineering promotes a perspective that incorporates the product

concept, but expands the understanding of targeted users and their needs.

Design

UCD focuses on accommodating user attributes that impact how users will interact with the Web site. Its goal is to achieve user success and satisfaction by incorporating users' perspectives into design. Throughout the life cycle, UEs perform design consultations on iterative products, informing design with knowledge of users and usability principles.

UCD processes, called interaction design, consider ways that users attempt to accomplish goals at a Web site. UEs base interaction design on all that they have learned about the users. Usability principles inform UCD decisions. Consider a site intended for senior citizens who expect a prominent link to articles about leisure activities for seniors. UEs apply usability principles on legibility for older users with decreased visual acuity and recommend a large font and a strong contrast between the font and background colors (e.g., Czaja & Lee, 2003).

Incorporating user input from requirements definition, UEs participate in designing the site's information architecture. An information architecture sets out paths that users follow within and between pages in a Web site. The UEs' role is to assure that the information architecture facilitates navigation and makes finding information natural for users.

UEs know how to manage the impact of features such as animations on users with disabilities. For example, it is simple, but detrimental to usability, to give an information site a bright colorful background with flashing bold titles. UEs recommend treatments that accommodate the biological impacts of such an approach, thereby reducing the potential for the eyes to become fatigued because they are unable to focus on flashing elements (Travis, 1991).

UEs bring to UCD an understanding of semiotics, the science of signs and symbols. When incorporating icons into a UI design, for example, it is important to use standard icons and to test designs for user comprehension. If an icon is used to mean something different than the user expects, it will likely cause confusion. With effort, users can learn arbitrary meanings for icons, but they easily forget arbitrary meanings. Such icons need text labels to clearly indicate the actions that will occur when they are activated. (Horton, 1994).

Participatory design is a UCD process in which users offer opinions on mock-ups or prototypes. A typical participatory design process is card sorting (Usability.gov, 2008). Users sort terms to be used in the Web site into groups that they name. UEs combine results from all participants through a statistical technique. Applying usability principles, UEs then derive a meaningful user-centered organization of Web site topics, that is, the site's information architecture.

We distinguish between *user-centered* design and inappropriate *user-driven* design (UDD) where user input translates directly into design directives. In turning user requests into design decisions without looking at them in light of usability principles, UDD runs a high risk of producing Web sites that do not meet users' needs. Another UDD pitfall is requirements creep that extends schedules and strains budgets as users add inappropriate features and functions that can have a negative impact on their experience at the site (Andre & Wickens, 1995).

Evaluation/Testing

In software engineering, verification is iterative testing against requirements. Validation is end-of-life-cycle testing against requirements. Usability evaluation is a set of verification and validation (V&V) processes that are applied in conjunction with other V&V activities. In addition to checking for conformance to usability requirements, iterative usability evaluation has the goal of assessing a wide range of user experiences at the site. When performed early in the life cycle, evaluation keeps the door open for new requirements based on the way real users interact with the site.

Usability evaluation is not a matter of simply asking users what they like or dislike about a site. Key user-centered usability evaluation processes entail observing users interacting with a Web site. Activities for user observation include writing a test plan; identifying participant users and their task expectations; working with site providers to set usability goals for each user group and task; defining tasks; writing scenarios (statements of goals that never tell users how to achieve those goals); writing user satisfaction surveys; preparing ancillary materials such as consent forms; carrying out the observations; collecting success and satisfaction data from participants; analyzing data and reporting findings (Lazar, Murphy, & O'Connell, 2004).

Nothing speaks louder about the quality of a Web site than the experiences of its users. During usability evaluation, UEs often employ click-capture software; record numbers and types of user errors; document latency periods when users pause for a significant period of time, pondering what to do next; and record critical incidents when users must stop work. Observing users interacting with the site, developers and stakeholders see the need for changes.

Using a method called *think aloud*, UEs encourage users to talk about their expectations and reactions while they work with a Web site (Boren & Ramey, 2000; Van Waes, 2000). The output is metric data on user success accompanied by anecdotal data, the users' own comments on what they were doing and why, and what they think of the site. UEs are specially trained to put users at ease during observations and to facilitate the users' evaluation experience.

After usability observations, users often complete satisfaction surveys on site components that they have experienced. These surveys collect ratings on a numerical scale to produce metric data. They also offer opportunities

for users to elaborate on their experiences. UEs never rely solely on satisfaction data, but use it to inform analysis of data collected during user interactions with the Web site. UEs interview users about their experience at the end of the session. Interviews also give users opportunities to bring up points that no one else has anticipated.

It is unusual to hold formal usability observations at every iteration. Indeed, some projects find user observations cost-prohibitive. However, other simple, less expensive processes incorporate the user perspective. In an expert review, one or more UEs independently assess an interface against their understanding of users; usability principles; and applicable laws and standards. When more than one UE has performed the expert review, they discuss and prioritize their findings. Another kind of expert review employs automated accessibility tools to inspect the Web site's UI code for conformance with accessibility regulations, and to recommend remedies.

Life cycles typically include several V&V iterations. Usability evaluation produces recommendations to improve the potential for user success and satisfaction with the UI. The principal benefit is an understanding of how users interact with the site (Dumas & Redish, 1999). A unique benefit of usability engineering is the coordination of these recommendations with other organizational and project goals.

FUTURE TRENDS

Usability engineering is increasingly recognized as an integral part of Web site development. The field will continue to refine its usability principles and life cycle activities to address evolving Web technologies and user-interaction behaviors.

CONCLUSION

Usability engineering is rigorous, process-based, and addresses needs of site providers as well as users. Software engineering and usability engineering processes include simultaneous, complementary activities whose products can benefit later activities without impacting schedules. Usability engineering is the means to providing successful and satisfactory experiences for Web site users while fulfilling the goals of the site's providers.

This paper reports the results of research and analysis undertaken by U. S. Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research, and to encourage discussion of work in progress.

REFERENCES

Addelston, J. D., & O'Connell, T. A. (2005). Integrating accessibility into the spiral model of the software development life cycle. In *Proceedings of the 11th International Conference on Human-Computer Interaction*, vol. 8, Universal access in HCI: Exploring new dimensions of diversity. Las Vegas, NV: Mira Digital.

Andre, A. D., & Wickens, C. D. (1995). When users want what's NOT best for them. *Ergonomics in Design, 4*, 10-14.

Bias, R. G., & D. J. Mayhew. (2005). *Cost justifying usability: An update for the Internet age* (2nd ed.). San Francisco, CA: Elsevier.

Bolt, R. A. (1984). *The human interface: Where people and computers meet*. Toronto: Wadsworth.

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*, 261-278.

Carroll, J. M. (1990). Mental models in human-computer interaction. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 45-65). NY: Elsevier.

Czaja, S. J., & Lee, C. C. (2003). Designing computer systems for older adults. In J. A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (pp. 413-427). Mahwah, NJ: Erlbaum.

Dumas, J. S., & Redish, J. C. (1999). *A practical guide to usability testing* (Rev. ed.). Portland, OR: Intellect.

Hix, D., & O'Connell, T. A. (2005). Usability engineering as a critical process in designing for accessibility. In *Proceedings of the 11th International Conference on Human-Computer Interaction*, vol. 8, Universal access in HCI: Exploring new dimensions of diversity. Las Vegas, NV. Mira Digital.

Horton, W. (1994). *The icon book.* NY: Wiley.

IEEE. (1990). *IEEE Standard glossary of software engineering terminology* (IEEE Standard 610.12-1990). New York: Institute of Electrical and Electronics Engineers.

International Organization for Standardization (ISO). (1998). *Guidance on usability* (ISO 9241-11).

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge UK: Cambridge University Press.

Kalin, S. (1999). Usability mazed and confused. *CIO Web Business Magazine*. Retrieved July 16, 2005, from http://www.cio.com/archive/webbusiness/040199_use.html#price

Koyani, S., Bailey, R. W., Nall J. R., Allison, S., Mulligan, C., Bailey, K., & Tolson, M. (2003). *Research-based Web design & usability guidelines*. Washington, DC: U. S. Department of Health and Human Services, Office of Communications, U.S. National Cancer Institute.

Lazar, J., Murphy, E. D., &. O'Connell, T. A. (2004). Building university-government collaborations: A model for students to experience usability issues in the federal workplace. *Journal of Informatics Education Research, 6*(3), 57-78.

Mayhew, D. J. (1992). *Principles and guidelines in software user interface design*. Englewood Cliffs, NJ: Prentice-Hall.

Mayhew, D. J. (1999). *The usability engineering life cycle: A practitioner's handbook for user interface design*. San Francisco, CA: Morgan Kaufmann.

Murphy, E. D., Nichols, E. M., Anderson, A. E., Harley, M. D., & Pressley, K. D. (2001). Building usability into electronic data-collection forms for economic censuses and surveys. In *Proceedings of Federal Committee on Statistical Methodology Research Conference* (Statistical Policy Working Paper 34, Part 4 of 5, pp. 113-122). Retrieved April 24, 2008, from http://www.fcsm.gov

Norman, D. A., & Draper, S. W. (Eds.). (1986). *User centered system design*. Hillsdale, NJ: Erlbaum.

O'Connell, T., & Choong, Y-Y. (2008). Metrics for measuring human interaction with interactive visualizations for information analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1493-1496). Florence, Italy. New York: Association for Computing Machinery.

O'Connell, T. A., & Murphy, E. D. (2007). The usability engineering behind user-centered processes for Web site development life cycles. In P. Zaphiris & S. Kurnianwan (Eds.), *Human computer interactive research in Web design and evaluation* (pp. 1-21). Hersey, PA: Idea Group Publishing.

Shneiderman, B. & Plaisant, C. (2004). *Designing the user interface: Strategies for effective human-computer interaction* (4th ed.). Reading, MA: Pearson Addison Wesley.

Travis, D. (1991). *Effective color displays: Theory and practice*. NY: Academic Press.

US Government. (1998). *Rehabilitation act of 1973* (29 USC 794d), as amended by the workforce investment act of 1998 (PL 105-220).

Usability.gov. (2008). *Perform card sorting*. Retrieved April 24, 2008, from http://www.usability.gov/design/cardsort.html

Van Waes, L. (2000). Thinking aloud as a method for testing the usability of Web sites: The influence of task variation on the evaluation of hypertext. *IEEE Transactions on Professional Communication, 43*, 279-291.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1990). Usability engineering: Our experience and evolution. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 791-817). NY: Elsevier.

World Wide Web Consortium (W3C). (2005). *Web content accessibility guidelines 2.0*. Working draft 30 June 2005. Retrieved July 13, 2005, from http://www.w3.org/TR/WCAG20/#robust

KEY TERMS

Accessibility: A subdomain of usability; enables people with disabilities to experience success and satisfaction with software to a degree comparable to that experienced by people without disabilities.

Evaluation: The process of assessing user-interface (UI) software for its ability to support users' success and satisfaction in meeting their goals; not simply a matter of asking what the user likes and dislikes, but a process of collecting and analyzing human performance data (e.g., accuracy, task-completion times) and ratings of user satisfaction with the "look and feel" of the user interface; results of evaluation inform development of recommendations to improve the UI design.

Life Cycles: Structured frameworks for software development activities.

Mental Models: Users' psychological representations of the components and behavior of, for example, processes, products, services, and relationships.

Software Engineering: Application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software (IEEE, 1990, p. 70).

Usability: The potential for user efficiency, effectiveness and satisfaction, or simply, user success and satisfaction, when interacting with technology.

Usability Engineering: A set of defined, user-centered processes, grounded in research-based principles.

Usability Evaluation of E-Learning Systems

Shirish C. Srivastava
National University of Singapore, Singapare

Shalini Chandra
Nanyang Technological University, Singapare

Hwee Ming Lam
Nanyang Technological University, Singapare

INTRODUCTION

The traditional approach for designing user interface for information systems[1] has focused on the capabilities of the technology. This "technology-centered approach" has frequently neglected the actual user requirements. The focus of such a design philosophy is to create interface systems that are based on opportunities presented by the capabilities of technology. In contrast, the user-centered design starts with the requirements of the end users and exploits the capabilities of technology to address users' needs, preferences and abilities. There is no doubt about the fact that such a shift from "technology-centered approach" to "user-centered approach" has increased the *usability* of the designed systems. However, in the context of e-learning systems, user-centered design is, in itself, not adequate to meet all the learner's needs. In addition to being user-centric, in terms of convenience, e-learning systems must also achieve the desired "learning outcomes". Thus, usability of e-learning systems has wider connotations compared to other information systems. The design for such systems, which aims at satisfying the learning needs more closely, is often referred to as learner-centered design (LCD) and goes beyond the usual user-centered design.

Usability evaluation which refers to a series of activities that are designed to measure the effectiveness of a system as a whole, is an important step for determining the acceptance of system by the users. Usability evaluation is becoming important since both user groups, as well as tasks, are increasing in size and diversity. Users are increasingly becoming more informed and, consequently, have higher expectations from the systems. Moreover "system interface" has become a commodity and, hence, user acceptance plays a major role in the success of the system. Currently, there are various usability evaluation methods in vogue, like cognitive walkthrough, think aloud, claims analysis, heuristic evaluation, and so forth. However, for this study we have chosen *heuristic evaluation* because it is relatively inexpensive, logistically uncomplicated, and is often used as a discount usability-engineering tool (Nielsen, 1994). Heuristic evaluation is a method for finding usability problems in a user interface design by having a small set of evaluators examine an interface and judge its compliance with recognized usability principles.

The rest of the chapter is organized as follows: we first look at the definition of e-learning, followed by concepts of usability, LCD, and heuristics. Subsequently, we introduce a methodology for heuristic usability evaluation (Reeves, Benson, Elliot, Grant, Holschuh, Kim, Kim, Lauber, & Loh, 2002), and then use these heuristics for evaluating an existing e-learning system, GETn[2]. We offer our recommendations for the system and end with a discussion on the contributions of our chapter.

BACKGROUND

E-Learning

According to MSN Encarta, electronic learning (e-learning) is "the acquisition of knowledge and skills using electronic technologies such as computer and Internet-based courseware and local and wide area networks". E-learning applications can be broadly classified into two categories: *offline learning,* where learning is imparted through the use of digital media devices like CD-ROMs, DVDs, and so forth, and *online learning*, where learning is imparted through computer networks using Web-based tools like, virtual classrooms, digital collaboration (discussion forum, chat, electronic bulletin boards, listserv, etc.). Web-based learning operates in a computer-networked environment and many of these systems make use of the Internet. In this chapter, we restrict our discussion to the usability of Web-based learning management systems (LMS). Such LMS not only offer online courseware, but also track participants' progress in learning.

Usability of E-Learning Systems

The ISO 9241 (1998) defines usability as "the extent to which a product can be used by specified users to achieve specified

goals with effectiveness, efficiency, and satisfaction in a specified context of use." Usability is the quality attribute that assesses the *ease of using* the application by users to accomplish their specified goals effectively, efficiently, and with a high level of satisfaction. In addition to ease of use, a usable e-learning system should be *useful* for the learners in accomplishing their learning task (Venkatesh, Morris, Davis, & Davis, 2003). Usability analysis helps increase the likelihood of a system being classified as not only easy to use but also useful from the learners' perspective.

Apart from the two basic objectives highlighted above, researchers have suggested several additions to the usability model. Constantine & Lockwood (1999) highlighted that a usable e-learning system must achieve: learnability, rememberability, efficiency in use, reliability in use, and user satisfaction. An effective e-learning system should be interactive and provide feedback, have specific goals, motivate users, communicate a continuous sensation of challenge, provide suitable tools, and help avoid distractions interrupting the learning stream (Costabile, De Marsico, Lanzilotti, Plantamura, & Roselli, 2005). Incorporating feedback, curiosity, comprehensiveness, and challenges in the e-learning system can help achieve the motivational aspect of usability (Shilwant & Haggarty, 2005). The *usability* definition for e-learning is incomplete without incorporating *learnability* aspect into the systems. Hence, in addition to being user-centric, e-learning systems should be designed with a learner-centric approach.

Learner-Centered Design

Learnability is defined as the ease and speed with which users can figure out the way to use a product (Soloway, Guzdial, & Hay, 1994). To incorporate learnability into the usability framework, learner-centered design (LCD) was proposed by Soloway and Pryor (1996). The key to developing effective e-learning systems is to adopt the LCD methodology, which targets to help learners acquire knowledge efficiently and effectively even through an e-learning tool, which is new for them. Since learners do not undergo extensive training before using the e-learning system, system design should be such that the learners can focus on the "actual learning" and not on "learning to use the e-learning system".

In LCD, the design process considers a variety of learner categories due to differences in personal learning strategies, experiences in the learning domain, and motivations in affording the learning task (Ardito, De Marsico, Lanzilotti, Levialdi, Roselli, Rossano, & Tersigni et al., 2004). The designer must also consider prior knowledge and self efficacy (computer skills) of the learner and focus on learners' needs and goals in a simple way. For example, the use of multimedia files as instructional mode should not overwhelm the users and disrupt their learning process. The e-learning interface design should also integrate various pedagogical methodologies and traditional teaching strategies. The aim should be to help learners achieve their learning objectives in a more productive, efficient and effective way.

USABILITY EVALUATION IN E-LEARNING

Usability evaluations in the case of e-learning systems are difficult to employ due to the time, budget, and knowledge constraints. Consequently, there is a lack of e-learning usability studies, which has adversely influenced e-learning design and development in practice.

Feldstein (2002) pointed out that producers and consumers of e-learning applications have no standardized means to evaluate the extent to which any e-learning application is usable. The norm generally followed for assessing usability is the *satisfaction level* of the user. This may not be a fair indicator of the actual usability because it does not take into account the *learning objectives*. A user may be satisfied with attributes of the system, other than those achieving learning outcomes. Hence there is a need for understanding e-learning usability in greater detail.

Heuristic Usability Testing

Feldstein (2002) suggested the use of "heuristics" for testing the usability of e-learning systems. Heuristic usability testing techniques offer simple and cheap means for assessing e-learning systems. Nielsen (2000, 1994) offered a set of simple usability heuristics, which focus on giving timely and useful feedback. Notess (2001) pointed out that heuristic evaluations hold promise for online learning but the challenge for most types of online learning is that established sets of heuristics do not exist. Many researchers are of the view that Web design heuristics that have developed around e-commerce can be suitably modified and used for the evaluation of e-learning systems. Squires & Preece (1996) established the ineffectiveness of simple heuristic usability testing and highlighted that Web design heuristics have to incorporate various learning theories and pedagogical guidelines to be effective for e-learning. They added socio-constructivist tenets to Nielsen's (1994) heuristics and proposed a set of heuristics for "learning with software" (Squires & Preece, 1996). In due course, taking into account usability as well instructional design features, Reeves, et al. (2002) expanded and customized Nielsen's (1994) ten heuristics developed for software in general, to enunciate fifteen heuristics for evaluating e-learning systems (Table 1), which we use in our current study.

Table 1. Reeves fifteen heuristics for evaluating e-learning programs

No	Heuristic	Description
Protocol for heuristic evaluation of e-learning programs		
H1	Visibility of system status	The e-learning program keeps the learner informed about what is happening, through appropriate feedback within reasonable time.
H2	Match between system and the real world	The e-learning program's interface employs words, phrases and concepts familiar to the learner or appropriate to the content, as opposed to system-oriented terms. Wherever possible, the e-learning program utilizes real world conventions that make information appear in a natural and logical order.
H3	Error recovery and exiting	The e-learning program allows the learner to recover from input mistakes and provides a clearly marked "exit" to leave the program without requiring the user to go through an extended dialogue.
H4	Consistency and standards	When appropriate to the content and target audience, the e-learning program adheres to general software conventions and is consistent in its use of different words, situations, or actions.
H5	Error prevention	The e-learning program is designed to prevent common problems from occurring in the first place.
H6	Navigation support	The e-learning program makes objects, actions, and options visible so that the user does not have to remember information when navigating from one part of the program to another. Instructions for use of the program are always visible or easily retrievable.
H7	Aesthetics	Screen displays do not contain information that is irrelevant and "bells and whistles" are not gratuitously added to the e-learning program.
H8	Help and documentation	The e-learning program provides help and documentation that is readily accessible to the user when necessary. The help provides specific concrete steps for the user to follow. All documentation is written clearly and succinctly.
Usability and Instructional Design Heuristics for evaluation of e-learning Programs		
H9	Interactivity	The e-learning program provides content-related interactions and tasks that support meaningful learning.
H10	Message design	The e-learning program presents information in accord with sound information-processing principles.
H11	Learning design	The interactions in the e-learning program have been designed in accord with sound principles of learning theory.
H12	Media Integration	The inclusion of media in the e-learning program serves clear pedagogical and/or motivational purposes.
H13	Instructional assessment	The e-learning program provides assessment opportunities that are aligned with the program objectives and content.
H14	Resources	The e-learning program provides access to all the resources necessary to support effective learning.
H15	Feedback	The e-learning program provides feedback that is contextual and relevant to the problem or task in which the learner is engaged.

Heuristic Evaluation of GETn System: An Illustrative Application

In this section we demonstrate the applicability of Reeves (2002) fifteen heuristics for evaluating an existing e-learning system GETn (acronym for Get Knowledge!). GETn is a widely known, well-established and leading global enterprise, which has been in this industry for the last thirty-five years and has partnered with a pool of leading technology and content partners. GETn is a web-based interactive and self-paced tutorial in a simulated environment. It offers a wide variety of courses: self-improvement courses such as improving communication skills, listening skills, relationship skills and IT courses for the use of various softwares like adobe photoshop, and so forth. The progress status of the learner for a particular topic is marked by symbols, which denote "not started", "started", "completed" and "mastery". It is highly interactive and gives an immediate feedback after each response from the user.

Method

Heuristic evaluation involves having a system examined independently by not more than two to five evaluators who understand the system's goals and have good knowledge of established usability guidelines. These evaluators develop a list of items that they must address, creating a structured format for the evaluation. Their reports typically turn up some major and some minor usability flaws, without ranking them by importance.

Table 2. Heuristic evaluation and recommendations for GETn

Heuristics	Violation of Heuristics	Recommendations
Visibility of system status	**No** display of percentage of course completion (progress status). **No** alert message to inform the learner that the system has been idling. This erroneously inflates the training hours.	**The** progress status bar should be incorporated to motivate learners. **Pop-up** messages should periodically alert learners about the inactivity of the lesson and ask them if they intend to continue.
Match between system and the real world	**No** logical classification of training courses into categories based. They are just arranged in alphabetical order. **Learner** does not know how the current score is computed. **No** glossary/index for quick references is available.	**The** courses should be classified into learning categories to facilitate quick access and choice by the learner. **The** methodology for computing scores should be made clear to the learners. **Glossary** and indexes should be made available online to facilitate quick cross-referencing.
Error recovery and exiting	**No** provision for error recovery when learner accidentally clicks on other tasks or exits from the topic while he/she is in the midst of learning. Learner has to restart that particular topic all over again.	**System** should have an alert pop-up message to confirm exit and should also allow learners to return to the last screen attempted.
Consistency and standards	Nil	Nil
Error prevention	**No** instructions for setting up GETn are displayed on the login screen to connect users to the training plan. **System** error often reported on course completion due to un-attempted items. The course does not display "completed" status.	**Instructions** for setting up should be shown on the login page to inform the users. **System** should avoid "uncompleted" status by pointing out the specific uncompleted or overlooked tasks, in a timely fashion so that learners are not frustrated doing the whole course again.
Navigation support	**No** option to "expand" contents to view all the sub-topics simultaneously. **No** search facility provided. **No** print option for course material. **Bookmark** option not available.	**Contents** menu should include "expand" and "collapse" options for quick viewing. **Should** have options for searching topics. **Print** option should be provided. **Bookmark** option should be provided for subsequent referencing later.
Aesthetics	**Viewing** window is small and has a fixed size. Users cannot adjust the screen size.	**Should** allow learners to adjust the screen size depending on their preferences.
Help and documentation	**No** provision for online or e-mail help. **Instructions** for downloading content are not clear.	System should provide online help, like live chat, and also provide help e-mail addresses. **Downloading** options should be clearly highlighted to enable offline learning also.
Interactivity	**No** synchronous and asynchronous tools for collaborative learning such as chat, discussion forums, resulting in no opportunity to learn from others.	**Discussion** forums, live chat, and other collaborative features should be incorporated so that learners can effectively gain knowledge through social learning.
Message design	**Important** concepts and topics are neither highlighted in the title nor in the body	**Important** concepts and key words should be highlighted. This is similar to teachers emphasizing the key concepts.
Learning design	Nil	Nil
Media Integration	**Use** of video is minimal.	**Videos** should be incorporated so as to make the learning processes more interesting and captivating.
Instructional assessment	Nil	Nil
Resources	**No** links to other resources such as good Web sites.	**Useful** and good resources such as recommended readings and useful Web sites should be listed for learners to pursue the course further.

continued on following page

Table 2. continued

Heuristics	Violation of Heuristics	Recommendations
Feedback	**Feedback** is rather brief.	**Elaborate** feedback should be given after each test or exercise.
	No option for sending feedbacks to instructors and experts by e-mails or any other means of communication.	**Provision** for sending feedbacks to relevant persons should be incorporated to have their expert comments for better learning.

For our GETn e-learning system, Reeves (2002) heuristics were chosen, as they are formulated specifically for e-learning systems. Two of the co-authors of this chapter did separate usability evaluations, after which the results were reconciled. The reported violations of heuristics (Table 2) are the ones which were either recorded or agreed to by both evaluators. This gave some objectivity to our subjective evaluations. We present our results and also recommendations in Table 2.

Discussion

From the heuristics evaluation results (Table 2) we see that even for well-established e-learning systems (like GETn) the usability standards may not be up to the desired level. Hence, there is an imperative need for firms manufacturing e-learning systems to have a systematic usability evaluation in order to have a better impact on the learners.

Heuristic usability evaluation can be employed extensively without any concerns about the cost aspect, since it is an inexpensive, quick, and reliable means of evaluation. However, as suggested by Reeves, et al. (2002) heuristic evaluation is not sufficient to determine whether an e-learning system is "learner-centered" or not, since heuristic approach is based only on the views of the experts. The evaluation techniques of e-learning systems should include learners' perceptions as well. Heuristic evaluation is not a replacement for the actual usability testing with the users of the systems. It is only an initial assessment done by the experts, developers or designers.

FUTURE TRENDS

As the issue of integrating usability and learning is still in its infancy, more research needs to be conducted in order to develop and standardize the usability evaluation methods for e-learning. The major challenge is to provide a usability evaluation technique, which incorporates learners' perceptions, integrates usability and instructional design, and at the same time is simple, cheap and reliable.

Extending the usability evaluation techniques a Systematic Usability Evaluation (SUE) technique has been proposed recently (Ardito, et al., 2004; Costabile, et al., 2005). SUE methodology aims to provide reliable evaluation by systematically combining inspection with user-based evaluation. Future designers of e-learning systems can also bring in the concept of *personas* (Shilwant & Haggarty, 2005)**,** to help them focus on the needs of end users throughout the design process. These developed personas serve as proxy users, when actual users are not available for testing. Newer techniques of usability like Microsoft Usability Guidelines (MUG), which are being used for Web sites and wireless Web sites, can also be adapted to the context of e-learning systems (Venkatesh & Ramesh, 2006; Agarwal & Venkatesh, 2002).

CONCLUSION

To design a good and usable e-learning system that is effective, efficient, learnable, and has high user satisfaction, designers have to put users and their needs in the forefront. Pedagogical methodology, instructional strategies, and also the learning context should be incorporated in the design process. Although developing an attractive and easy to use, system interface is essential, the focus of e-learning systems should be on "learning outcomes" rather than satisfying the user only with a good interface. Designers or developers should pay special attention to the learning goals and needs of the users, instructional strategies, and quality of learning. This chapter discusses the applicability and usefulness of a simple and inexpensive usability evaluation technique, namely, heuristic evaluation. E-learning software designers are looking for ways and means of providing better products to their customers, and this chapter helps them understand

one such way without a heavy cost burden. As already discussed, actual user studies are more realistic, but due to cost constraints many firms do not actually conduct them for their products. Moreover, user studies can be done only when the product is actually ready for use. On the other hand, heuristic evaluation is a useful inexpensive evaluation technique, which can be employed even before the final product is launched in the market.

REFERENCES

Agarwal, R., & Venkatesh, V. (2002). Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability. *Information Systems Research*, 13(2), 168-186.

Ardito, C., De Marsico, M., Lanzilotti, R., Levialdi, S., Roselli, T., Rossano, V., & Tersigni, M. (2004). Improving Interaction: Usability of E-learning Tools. *Proceedings of the Working Conference on Advanced Visual Interfaces-2004*, ACM Press: New York, pp. 80-84.

Constantine, L., & Lockwood, L. (1999). *Software for Use: A Practical Guide to the Essential Models and Methods of Usage-Centered Design*. Addison-Wesley: Reading, MA.

Costabile, M.F., De Marsico, M., Lanzilotti, R., Plantamura, V.L., & Roselli, T. (2005). On the Usability Evaluation of E-Learning Applications. *Proceedings of the 38th Hawaii International Conference on System Sciences – 2005*, ACM Press: New York, pp. 1-10.

Feldstein, M., (2002). What is 'Usable' E-learning? *eLearn Magazine*. Retrieved February 15, 2006, from http://www.elearnmag.org/subpage.cfm?section=tutorials&article=6-1

ISO 9241. (1998). Ergonomics Requirements for Office Work with Visual Display Terminal (VDT). Retrieved February 16, 2006, from http://www.usabilitynet.org/management/b_standards.htm

Nielsen, J. (1994). *Heuristic Evaluation, in Usability Inspection Methods, Nielsen, J. and Mack, R.L.* (Eds.). John Wiley & Sons: New York.

Nielsen, J. (2000). Designing Web Usability. *New Riders*. Indianapolis, IN.

Notess, M. (2001). Tutorial: Usability, User Experience, and Learner Experience. *eLearn Magazine*, 2001(8).

Reeves, T. C., Benson, L., Elliot, D., Grant, M., Holschuh, D., Kim, B., Kim, H., Lauber, E., & Loh, S. (2002). Usability and Instructional Design Heuristics for E-Learning Evaluation. *Proceedings of the World Conference on Educational Multimedia, Hypermedia & Telecommunications*, June 24-29: Denver.

Shilwant, S., & Haggarty, A. (2005). Usability Testing for E-Learning. Retrieved February 16, 2006, from http://www.clomedia.com/content/templates/clo_article.asp?articleid=1049

Soloway, E., Guzdial, M., & Hay, K. E. (1994). Learner-Centered Design: The Challenge for HCI in the 21st Century. *Interactions*, 1(2), 36-48.

Soloway, E. & Pryor, A. (1996). The Next Generation in Human-Computer Interaction. *Communications of the ACM*, 39(4), 16-18.

Squires, D. (1999). Usability and Educational Software Design: Special Issue of Interacting with Computers. *Interacting with Computers*, 11 (5), 463-466.

Squires, D., & Preece, J. (1996). Usability and Learning: Evaluating the potential of educational software. *Computer & Education*, 27(1), 15-22.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward A Unified View. *MIS Quarterly*, 27(3), 425-478.

Venkatesh, V., & Ramesh, V. (2006). Web and Wireless Site Usability: Understanding Differences and Modeling Use. *MIS Quarterly*, 30(1), 181-206.

KEY TERMS

E-Learning: According to MSN Encarta, e-learning (or electronic learning) is "the acquisition of knowledge and skill using electronic technologies such as computer and Internet-based courseware, and local and wide area networks." In other words, it is learning through electronic means where knowledge and skills are transferred either through the computer networks or through digital media like CD-ROM, DVDs, and so forth. In the Web environment, users may use virtual classrooms, digital collaboration, discussion forums, chat rooms, and so forth, to obtain information and facilitate learning.

Heuristic Evaluation: Heuristic evaluation is a method for quick, cheap, and easy evaluation of a user interface design by comparing the system to a set of identified heuristics.

Learnability: Learnability is defined as the ease and speed with which the users get familiar with the use of a new product. With high learnability, users can intuitively learn to use a product without training or manuals. However, in the context of e-learning, the definition of learnability includes

the ability of users to effectively learn and retain the skills and knowledge.

Learner-Centered Design: Learner-centered designs focus on developing a learner's understanding, rather than on improving usability, of the designed system. In the context of e-learning systems, the focus is more on the achieving the "learning outcomes."

Personas: Personas are archetypal users of an intranet or Web site that represent the needs of larger groups of users in terms of their goals and personal characteristics. They represent the real users and help guide the designers in making decisions about the system designs and functionalities. Personas represent behavior patterns by identifying user motivations, expectations, and goals from the system. Although personas are fictitious, they are based on knowledge and profiles of real users.

Usability: The ISO 9241 defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."

User-Centered Design: It is a design approach in which the emphasis is on the users' requirements and not on the designers' capabilities. User-centered design helps achieve a high level of usability.

ENDNOTES

1 'Systems' can be hardware or software, however, in our paper, we use it for software systems.

2 This is a fictitious name to hide the real identity of the e-learning software.

Usable M-Commerce Systems

John Krogstie
IDI, NTNU, SINTEF, Norway

INTRODUCTION

Today, the PC is only one of many ways to access information resources. Traditional computing technology has become more mobile and ubiquitous and more and more computing tasks are possible to do using new types of mobile devices.

According to Siau, Lim, and Shen (2001), the essence of m-commerce (also termed "mobile information systems") is to reach customers, suppliers, and employees regardless of where they are located and to deliver the right information to the right person(s) at the right time. The ability to develop and evolve usable m-commerce systems is becoming increasingly critical for enterprises.

BACKGROUND

M-commerce systems differ from more traditional information systems along several dimensions (Krogstie et al., 2004). We have grouped the differences into three areas:

1. User-orientation and personalization
2. Technological aspects including convergence and need for multi-channel support
3. Methodology for development, evolution, and operations to ensure organizational returns

User-Orientation and Personalization

M-commerce systems often address a wide user-group, which means that user-interface aspects should feature prominently and early in the design process and often need to be very simple. A number of examples exist indicating that complex services often do not get adopted (Blechar, Constantiou, & Damsgaard, 2005; Steinert & Teufel, 2005). Input and output facilities of the end-user device may be severely restricted (no keyboard, small screen-size, etc.) or based on new modalities (speech-recognition and -synthesis). This means that personalization of mobile information systems becomes increasingly important, both at the individual level where user-interface details are tailored to personal preferences and hardware, and at the work-group level where functions are tailored to fit the team's preferred way of working.

Personalization means information systems that both automatically adapt themselves to the preferences and context of the user, and that can be explicitly tailored by users. The main goal is to achieve usability of the applications on all possible interfaces, based on adaptation to the different physical devices. This calls for intelligent, adaptive, and self-configuring services that enable automatic context-sensitivity and user profiling (Hella & Krogstie, 2006).

Technological Aspects Including Convergence and Multi-Channel Support

Mobile devices still have limited processing, memory, and communication capacities compared to what one are familiar with on traditional PCs. Performance considerations therefore is still important, as is bandwidth analysis. Analytically-based predictive methods are necessary in order to assess a large number of alternatives during the design (Gruhn & Köhler, 2006). M-commerce systems also pose new challenges to achieving information systems dependability. The new mobile devices provide integration and convergence of technologies into a wide range of innovative mobile and multi-modal applications. Mobile and other new technologies provide many different ways to offer the same or similar services. Thus, there is a need for novel approaches for the development and evolution of applications on and across different mobile and traditional platforms.

Methodology for Development and Operations to Ensure Organizational Return

M-commerce systems are often radically different from the existing systems. They therefore reward an increased focus on idea generation early in the requirements and design process. Understanding the mobile users requirements for new services is thus of large importance. One needs both to be able to develop these systems and to address the major hurdles for the deployment of applications and services (Amberg, Wehrman, & Zimmer, 2004). Another effect of the radically new approaches is that the introduction of m-commerce systems often spawns several other initiatives for changing other information systems and processes within an organization. It is important to focus on the interoperability of services and seamless access to corporate and government resources from the mobile devices (Pernici, 2006). A model-driven approach appears to be a promising approach to address many of these methodological challenges.

STATUS FOR MODEL-DRIVEN DEVELOPMENT AND EVOLUTION OF USABLE M-COMMERCE SYSTEMS

When speaking about model-driven system development, we refer to models developed in languages that have the following characteristics:

- The languages are diagrammatic with a limited vocabulary (states, classes, processes etc).
- The languages utilize powerful abstraction mechanisms.
- The languages have a formal syntax and semantics. The formal semantics is either operational enabling, for example, generation of other models including executable programs or mathematical enabling advanced analyses.
- The languages are meant to have general applicability across problem domains.

Although most software developers are aware of model-driven methodologies, they are traditionally mostly used in initial development stages. Newer approaches such as model-driven architecture (MDA) and especially service-oriented architecture (SOA) on the other hand appear to be changing this (Pernici, 2006).

In general, a model-driven approach to information systems development can be argued to have the following advantages (Krogstie & Sølvberg, 2003):

- Explicit representation of goals, organizations and roles, people and skills, processes and systems
- An efficient vehicle for communication and analysis
- Basis for design and implementation
- Readily available documentation as a basis for extensions and personalization

One striking aspect in connection to contemporary information systems development and evolution is that there is an increasing demand for shorter development time for new products and services (Pries-Heie & Baskerville, 2001). This is specifically important for m-commerce systems, where the convergence of different platforms continuously creates opportunities for new functionality. Some would argue that this highly dynamic situation would make model-based approaches impractical. To the contrary, we claim that this means that systems must be developed for change, which make model-based techniques specifically useful.

We can identify the following areas for potentially increased utility of techniques developed as part of model-driven development:

- **User-orientation and personalization:** Traditionally, support for workers performing nomadic processes has not been provided. Recent approaches to workflow modeling is starting to take the specific aspects of m-commerce systems into account (Modafferi, Bentallah, Casati, & Pernici, 2005; Sørensen, Wang, & Conradi, 2005). Functions of the mobile information system should be tailored to fit the user's preferred work processes, which typically involve other persons. To support teamwork, raising awareness of the status of knowledge resources is increasingly important in a mobile setting. To enhance social mobility, organizations and industries need to develop "social ontologies," which define the significance of social roles, associated behaviors, and context (Lyytinen & Yoo, 2002). Given that knowledge resources include both individuals and technology that can be mobile, one should look into interactive systems to improve group performance. Wegner's interaction framework (Wegner, 1997) was inspired by the realization that machines that must interact with users in the problem solving process can solve a larger class of problems than algorithmic systems computing in isolation. The main characteristic of an interaction machine is that it can pose questions to human actors (users) during its computation. The problem solving process is no longer just a user providing input to the machine, which then processes the request and provides an answer (output), but rather is a multi-step conversation between the user and the machine, each being able to take initiative. A major research question in this area is how to specify and utilize interaction machines on a multi-channel platform. Process support technologies are a natural choice for enabling interaction machines. Such technologies are typically based on process models, which need to be available in some form for people to alter them to support their emerging goals. Thus, interactive models should be supported (Krogstie & Jørgensen, 2004). The outset for this thinking is that models can be useful tools in a usage situation, even if the models are changing and incomplete. The user is included as an interpreter and changer of the models.
- **Technological aspects including convergence and multi-channel support:** There is a multitude of competing technologies available for providing the underlying infrastructure and access devices for distributed and mobile applications. A central element when addressing this is the development of model based specification techniques that are powerful enough to be used as a basis for the development of systems on a large number of technical platforms, but still general enough to represent the commonalties at one place only. One interesting approach is the use of architectural patterns (Risi & Rossi, 2004). A major initiative within

object management group (OMG) is on model-driven architectures (MDA) where both platform independent and platform specific modeling notations, including refinement techniques and mappings between platform independent and platform specific notations, are specified. Meta-modeling techniques and domain-specific modeling (DSM) have found a special application for the design of mobile phone software for some time already (Kelly & Pohjohnen, 2003). It could be argued that mobile information systems are a particularly good area for using domain-specific modeling:

- The software (on the client side) is partly embedded and thus needs a higher reliability than traditional software and can be supported by restricting choices through the application of modeling rules and code-generation.
- You need many very similar variants of the same application.
- There are a number of standards to adhere to, and the technology and standards change rapidly.

Model-based development can be used to support dependability analyses (i.e., the use of methods, techniques, and tools for improving and estimating dependability) such as risk analyses, probabilistic safety assessment, testing, formal walkthrough, simulation, animation, exhaustive exploration, and formal verification.

The new separation between content and medium found in m-commerce systems serves as a major challenge. Generally, the context should be explicitly modeled to maintain an overview of, analyze, and simulate the multitude of possibilities open for adapting to the context and the use of context traces (Henricksen, Wishart, McFadden, & Indulska, 2005). For the more general use of mobile applications, it is also important to be able to adapt these systems to the user at hand, thus making a case for simple user-models to guide the adaptation. Banavar and Bernstein (2002) highlight the importance of semantic modeling in this respect.

Over the recent years, work within user interface modeling has focused increasingly on mobile and multi-channel user interfaces (Vanderdonckt, 2005). This is often done to facilitate some level of common models for both the mobile and more traditional user interfaces. A central element in this is the development of model-based approaches that are powerful enough to form the basis for the development of user-interfaces on the multitude of platforms needed, but still general enough to represent the commonalties in a single place. One approach is to define user-interface patterns with general usability principles as powerful building blocks (Nilsson, 2002).

- **Methodology for development to ensure organizational return:** Siau et al. (2001) highlights as an important application-oriented research area the development of m-commerce business models. In order for m-commerce to succeed, it is vital to ensure that all the related applications and services can be accessed with ease and at little cost. Thus, in addition to externalizing the business models in the manner of independent computing, it is important to integrate these models with the internal enterprise models and enterprise architecture in order to pinpoint the links to, for example, internal systems for the efficient billing of services provided.

M-commerce systems are often radical and therefore reward an increased focus on idea generation early on in the development phase. This also means that services or situations in which to anchor problem analysis efforts such as using as-is analysis as a starting point for to-be design do not always exist. Technology in the field still develops rapidly, which means that idea generation should not be limited by currently available technologies. One needs to enhance the techniques for modeling of scenarios to take this into account (Skattør, 2005). Applications of the new technology call for highly distributed systems that comprise new user-interface systems on the client side, new and existing back-end systems, as well as new bridging systems (which port information between other systems). The new technologies therefore highlight the need for principled, long-term IS-architecture management and for integrating architecture management with software development methodologies. Often there is a need to interface with existing enterprise systems and architectures to enable the new workflow. Another aspect is how to integrate the user-interface models discussed above with other parts of the requirements and design model, for instance the entity model, process model and goal model. On both the process and the user-interface side, the challenges can be attacked by extending existing approaches to modeling, although research is needed to investigate the techniques that should be extended and how they could be best adapted to the new problem areas.

FUTURE TRENDS

The large-scale application of m-commerce is just getting started and increasingly work is done to provide usable m-commerce systems or on model-based development and evolution of usable m-commerce systems. As 3G- (UMTS-) and WLAN-infrastructures now are available in many locations providing higher bandwidth and constant connection to the network from virtually everywhere, the number of m-commerce applications is predicted to explode as will the number of users of complex applications on a mobile

platform. Thus, the need for evolving best practice within systems development will be increasingly important.

CONCLUSION

We have highlighted the need for adapting and extending modeling-based approaches to ensure usable m-commerce systems in this chapter. We are not starting this work from scratch, it is possible to build on existing work within the usability, user interface, and modeling fields, specifically on techniques for modeling of functional and non-functional requirements, data modeling, process modeling, user-centered design, model-driven architectures, service-oriented architecture, requirements specifications of web-applications, domain-specific modeling and dependability analysis.

REFERENCES

Amberg, M., Wehrman, J., & Zimmer, R. (2004). Factors influencing the design of mobile services. In E. Lawrence, B. Pernici, & J. Krogstie (Eds.), *Proceedings of IFIP TC8 Working Conference on Mobile Information Systems (MOBIS)*. Oslo, Norway: Kluwer.

Banavar, G., & Bernstein, A. (2002). Software infrastructure and design challenges for ubiquitous computing. *Communications of the ACM, 45*(12), 92-96.

Blechar, J., Constantiou, I., & Damsgaard, J. (2005). Seeking answers to the advanced mobile services paradox: Minimal acceptance and use despite accessibility. In J. Krogstie, D. Allen, & K. Kautz (Eds.), *Proceedings of the 2nd IFIP TC8 Working Conference on Mobile Information Systems (MOBIS)*. Leeds, UK: Kluwer.

Gruhn, V., & Köhler, A. (2006). Modeling communication behaviour of mobile applications. In *Proceeding from EMMSAD'06 included in CAiSE workshop Proceedings*, Presses Universitaires de Namur.

Hella, L., & Krogstie, J. (2006). *Semantic Web as enabling technology for m-commerce personalisation: Personalisation scenarios*. In Semantic Web Personalization Workshop at the 3rd European Semantic Web Conference, Budva, Montenegro, June 12.

Henricksen, K., Wishart, R., McFadden, T., & Indulska, J. (2005, March 8-12). Extending context models for privacy in pervasive computing environments. In *Proceedings of the 3rd International Conference on Pervasive Computing and Communications Workshops (PreCom 2005 Workshops)* (pp. 2024). Kanuai Islands, Hawaii, USA.

Kelly, S., & Pohjohnen, R. (2003). *Domain-specific modelling for cross-platform product families*. In Advanced Conceptual Modeling Techniques: ER 2002 Workshops, Springer LNCS 2784.

Krogstie, J., & Jørgensen. H. (2004). Interactive models for supporting networked organisations. The *16th Conference on Advanced Information Systems Engineering*. Riga, Latvia: Springer Verlag.

Krogstie, J., & Sølvberg, A. (2003). *Information systems engineering—Conceptual modeling in a quality perspective*. Trondheim, Norway: Kompendiumforlaget

Krogstie, J., Lyytinen, K., Opdahl, A., Pernici, B., Siau, K., & Smolander, K. (2004). Research areas and challenges for mobile information systems. In J. Krogstie, K. Lyytinen, K. Siau, & B. Lin (Eds.), *Modeling mobile information systems: Conceptual and methodological issues*. International Journal of Mobile Communication.

Lyytinen, K., & Yoo, Y. (2002). The next wave of nomadic computing: A research agenda for information systems research. *Information Systems Research*, April.

Modafferi, S., Bentallah, B., Casati, F., & Pernici, B. (2005). A methodology for designing and managing context-aware workflows. In J. Krogstie, D. Allen, & K. Kautz (Eds.), *Proceedings of the 2nd IFIP TC8 Working Conference on Mobile Information Systems (MOBIS)*. Leeds, UK: Kluwer.

Nilsson, E. G. (2002, June 12-14). Combining compound conceptual user interface components with modeling patterns—A promising direction for model-based cross-platform user interface development. In *Proceedings of the 9th International Workshop on the Design, Specification, and Verification of Interactive Systems*. Rostock, Germany.

Pernici, B. (2006). *Mobile information systems: Infrastructure and design for adaptivity and flexibility*. Springer Verlag.

Pries-Heie, H., & Baskerville, R. (2001, June 22-23). eMethodology. In *Proceedings of the IFIP TC 8 Conference on Developing a Dynamic, Integrative, Multi-Disciplinary Research Agenda in E-Commerce/E-Business*, Salzburg.

Risi, W. A., & Rossi, G. (2004). An architectural pattern catalogue for mobile Web information system. In J. Krogstie, K. Lyytinen, K. Siau, & B. Lin (Eds.), *Modeling Mobile Information Systems: Conceptual and Methodological Issues*, International Journal of Mobile Communication.

Siau, K., Lim, E. P., & Shen, Z. (2001). Mobile commerce: Promises, challenges, and research agenda. *Journal of Database Management, 12*(3).

Skattør, B. (2005). Creating and performing scenarios for mobile services supporting mobile work in exposed physical environments. In J. Krogstie, D. Allen, & K. Kautz (Eds.), *Proceedings of the 2nd IFIP TC8 Working Conference on Mobile Information Systems (MOBIS)*. Leeds, UK: Kluwer.

Steinert, M., & Teufel, S. (2005). The European mobile data service dilemma: An empirical analysis on the barriers of implementing mobile data services. In J. Krogstie, D. Allen, & K. Kautz (Eds.), *Proceedings of the 2nd IFIP TC8 Working Conference on Mobile Information Systems (MOBIS)*. Leeds, UK: Kluwer.

Sørensen, C. F., Wang, A. I., & Conradi, R. (2005). Support of smart work processes in context rich environments. In J. Krogstie, D. Allen, & K. Kautz (Eds.), *Proceedings of the 2nd IFIP TC8 Working Conference on Mobile Information Systems (MOBIS*). 2005. Leeds, UK: Kluwer.

Vanderdonckt, J. (2005, June 13-17). A MDA-compliant environment for developing user interfaces of information systems. In *Proceedings of the 17th Conference on Advanced Information Systems Engineering*. Porto, Portugal: Springer Verlag.

Wegner, P. (1997). Why interaction is more powerful than algorithms. *Communications of the ACM, 40*(5).

KEY TERMS

M-Commerce: A technological approach to reach customers, suppliers, and employees regardless of where they are located and to deliver the right information to the right person(s) at the right time.

Mobile Computing: The capability to physically move computing services with us.

Mobile Information Systems: Information systems that include end-user terminals that are easily movable in space, are operable independent of location and have wireless access to information resources and services.

Multi-Channel Information Systems: Information systems that are to be used by different types of end-user equipment such as traditional PC, PDA, and a mobile phone in an integrated manner.

Nomadic Computing: The use of computers while on the move.

Pervasive Computing: An environment where computers have the capability to obtain information from the environment in which it is embedded and utilize it dynamically

Ubiquitous Computing: An environment where computers are embedded in our natural movements and interactions with our environments. Combines mobile and pervasive computing.

Use Cases in the UML

Brian Dobing
University of Lethbridge, Canada

Jeffrey Parsons
Memorial University of Newfoundland, Canada

INTRODUCTION

The unified modeling language (UML) emerged in the mid-1990s through the combination of previously competing object-oriented systems analysis and design methods, including Booch (1994), Jacobson, Christerson, Jonsson, and Overgaard (1992), Rumbaugh, Blaha, Premerlani, Eddy, and Lorensen (1991) and others. Control over its formal evolution was placed in the hands of the object management group (www.omg.org), which recently oversaw a major revision to UML 2.0 (OMG, 2005). The UML has rapidly emerged as a standard language and notation for object-oriented modeling in systems development, while the accompanying unified software development process (Jacobson, Booch, & Rumbaugh, 1999) has been developed to provide methodological support for applying the UML in software development.

Use cases play an important role in the unified process, which is frequently described as "use case driven" (e.g., Booch et al., 1999, p. 33). The term "use case" was introduced by Jacobson (1987) to refer to a text document that outlines "a complete course of events in the system, seen from a user's perspective" (Jacobson et al., 1992, p. 157). The concept resembles others being introduced around the same time. Rumbaugh et al. (1991), Wirfs-Brock, Wilkerson, and Wiener (1990), and Rubin and Goldberg (1992) use the terms "scenario" or "script" in a similar way. While use cases were initially proposed for use in object-oriented analysis and are now part of the UML, they are not inherently object-oriented and can be used with other methodologies.

The official UML 2.0 documentation (OMG, 2005) includes some examples of use case diagrams, which provide an overview that shows which "actors" are involved in each use case. However, the only indication of the "text document" format is that "use cases are typically specified in various idiosyncratic formats such as natural language, tables, trees, etc." (UML, 2005, p. 574). However, virtually every book on the UML offers some format suggestions for use cases (sometimes termed "use case narratives" or "use case descriptions" to clearly distinguish them from diagrams). Together, the use case diagram and narrative are referred to as the "use case model." There are now several books focusing on use cases including Adolph and Bramble (2003), Armour and Miller (2001), Bittner and Spence (2003), Cockburn (2001), Denny (2005), and Övergaard and Palmkvist (2005) along with a few Web sites, notably Cockburn's (http://www.usecases.org). Thus, use cases seem to be well established within the UML despite the lack of any officially endorsed format from the OMG.

BACKGROUND

A use case "describes the system's behavior under various conditions as the system responds to a request from one of the stakeholders" (Cockburn, 2001). A use case should have a clear goal and describe what should happen (but not how it should happen) as users interact with the system. Common examples would include a customer renting a video, purchasing an item, withdrawing funds from a bank account, etc. The use case also identifies the main "actors" involved which, in the previous examples, could include the customer, employees (e.g., rental clerk), a device (bank machine), time (clock), etc. The use case must provide something of value to one or more actors; otherwise there would be no need for it. While the main use case narrative would describe a successful rental, purchase, or withdrawal, alternative outcomes would handle problems such as rejected credit cards, insufficient funds, etc.

The use case differs from typical structured requirements analysis tools that preceded it in two important ways. First, the use case is largely text-based (with the use case diagrams playing a minor role). Structured analysis emphasized the importance of graphical tools, such as work flow and data flow diagrams. The UML has not abandoned diagrams; thirteen are now included with UML 2. The class, activity, communication (previously collaboration), sequence, state machine (previously statechart), and use case diagrams have always played important roles. But use case narratives are text-based so that "users and customers no longer have to learn complex notation" (Jacobson et al., 1999, p. 38).

Second, use cases focus on complete transactions, from initiation to achievement of the defined goal, from the user's perspective. In particular, a use case has a goal, which comes from the goals of those who will be using the system (Cockburn, 2001). This keeps the focus on the key

requirements and helps facilitate communication with the system's intended users. In UML terminology, a use case is initiated by an actor, usually a person in a particular role (e.g., cashier) but actors can also be external systems or devices. A single use case can involve many actors.

Consistent with an object-oriented approach, use cases can also have generalizations and include and extend relationships. Generalizations allow a child use case to override the behavior of its parent use case in certain situations, but are "not widely used" according to Arlow and Neustadt (2004). An "include" relationship is generally used when the same steps are required by several use cases (e.g., logging into a system), in order to avoid repetition. An included use case is dependent on base use cases and "never stands alone" (Booch et al., 1999, p. 221). An "extend" relationship exists when a base use case incorporates another use case depending on certain conditions, such as exceptional situations where including the additional detail in the base use case adds too much complexity.

Writing use cases may seem simple enough because they are text-based. However, as discussed in the next section, the content and format of use cases vary somewhat among published books and articles. Those new to use cases would be well advised to read at least a couple of the books devoted to use cases (referenced previously) before incorporating them into a system development project.

ISSUES

Use cases have been all but universally embraced in object-oriented systems analysis and development books written since Jacobson et al. (1992). Despite this strong endorsement, there are many variations on Jacobson's original theme. First, there is a difference in content. Use cases, at least during the analysis phase, were intended to be a conceptual tool. The use case should emphasize "what" and not "how" (Jacobson et al., 1994, p. 146). This principle was not strictly followed by much of the early literature, including Jacobson et al. (1992, p. 162) who referred to a display "panel," "receipt button," and "printer" in one of their previous examples. Constantine and Lockwood (2000) distinguish "essential" use cases containing few if any references to technology and user interface implementation, from "concrete" use cases that specify the actual interactions. Others make a similar distinction using the terms "business use cases" and "system use cases." While this provides flexibility, developers need to be careful about what type of content is appropriate at any given time.

Second, there are several variations proposed for use case formats. While the first use cases in Jacobson et al. (1992) were written as a paragraph of text, most others have adopted numbered steps. Soon after, Jacobson et al. (1994, p. 109) did so as well. There also seems to be more acceptance of including exception and error steps, which were less common in earlier books.

Third, the granularity of use cases varies from coarse (few use cases) to fine (many). In principle, use cases should offer "measurable value to an individual actor" (Jacobson et al., 1994, p. 105) and "the collected use cases specify the complete functionality of the system" (White 1994, p. 7). But how to determine the number of use cases this requires is not easily articulated. While Dewitz (1996) uses 11 use cases in her video store example, the IBM object-oriented technology center (1997) has 24. Kulak and Guiney (2000, p. 37) suggest that "most systems would have perhaps 20 to 50 use cases and some small systems even fewer." But, as they later point out (p. 88), "there are no metrics established to determine correct granularity." Övergaard et al. (2005, p.45) suggest the same range (20-50) for "a normal medium-sized system." Armour et al. (2001, p. 244) claim that large systems may have hundreds of use cases.

Fourth, the level of detail within each use case also varies. For example, both Kulak et al. (2000, p. 125) and Armour et al. (2001, p. 125) recommend limiting the length of the flow of events to two pages of text, but the latter also note that some practitioners prefer a few longer use cases to many short ones. Bittner et al. (2003) suggest they are typically 5 to 15 pages, but with 60 to 80% of the content handling exception and error conditions. Jacobson et al. (1999) advocate an iterative development approach in which both the number of use cases and their level of detail increase as the work progresses. They suggest that only the most critical use cases (less than 10%) be detailed in the first (inception) phase. As analysis progresses and requirements become firmer, additional use cases can be added and each can be expanded to include considerably more detail. For example, Kulak et al. (2000) have identified four levels. However, knowing what should be a use case, how much detail is appropriate at each phase, and when to stop are important issues that are difficult to resolve precisely.

To further complicate the issue, some of those who favor fewer or less detailed use cases supplement them with "scenarios." Rumbaugh et al. (2005, p.579) say that "a scenario may be used to illustrate an interaction or the execution of a use case instance." "Add a customer" is a use case. Adding a specified customer with a particular name, address, etc. is a scenario. Others use scenarios to provide further detail on exception handling and other special cases (e.g., customers with missing, improbable, or unusual data) (Bennett, Skelton, & Lunn, 2001) rather than alternative paths in the use case. How many scenarios, alternate paths, and exception paths should be developed, and what their role should be in developing class diagrams, is not clear. A minimalist approach to use cases combined with extensive scenarios and paths may still result in a large and very detailed set of specifications.

While the general consensus seems to be in favor of a smaller set with relatively brief descriptions, "use case modeling concepts may be applied very informally and loosely or very rigorously and formally" (Armour et al., 2001, p. 70). The difficulty is determining when each is appropriate. Stories about organizations mired in hundreds or even thousands of use cases suggest that some limits need to be applied. Users will not read, or at least not properly understand, long and complex use cases. But a smaller set may be insufficient to fully capture the requirements. Thus, the two key roles of use cases, to gather requirements and to support development of the Class and other diagrams, may conflict somewhat and the ideal set of use cases for each role are perhaps different.

Fifth, several surveys have found that use cases are not always used with the UML. Grossman, Aronson, and McCarthy (2005) reported 93% of developers used use case narratives, Zeichick (2002) reported 89% used use case diagrams, and Dobing and Parsons (2006) found 93% for diagrams and 85% for narratives. However, these figures are based on developers, not projects. Only 44% of developers reported that use case narratives are used on two-thirds or more of their projects (Dobing et al., 2006). Thus, some projects are clearly not "use case driven" as recommended by Booch et al. (1999, p. 33) and most of the UML literature. However, this survey also found that clients are much more involved with other UML diagrams than the UML literature would suggest. How the absence of use cases affects project outcomes is unknown.

In a related issue, use case narratives have been primarily designed for communication between system analysts and clients, while other diagrams (particularly the class diagram) are used by analysts to communicate with programmers. While it might appear that anyone should be able to understand a use case narrative, Arlow et al. (2004, p. 92) point out that programmers may lack the necessary domain expertise and vocabulary. But because of their technical nature, they argue that clients have difficulty comprehending most of the other key UML diagrams. Thus, there is a potential communication disconnect that can occur when clients and developers are relying on different UML artifacts to understand the proposed system.

FUTURE TRENDS

Use cases, and the use case driven approach when using the UML, seem to be overwhelmingly accepted in the UML literature but, as previously noted, not by all practitioners. Better tools may encourage greater use, as well as help to standardize the ways in which use cases are written and used. Many developers are still relying largely on word processors to develop use cases. A closer integration with other UML diagrams would help ensure consistent specifications across all of them and greatly facilitate changing the specifications when modifications are required.

There remains considerable opportunity for researchers to examine issues in use case writing, particularly format, granularity, level of detail and usage. Perhaps the most fundamental issues deal with the effectiveness of use cases in their two key roles, communicating with users and supporting development of UML diagrams (particularly the class diagram). How effective are use cases in helping users find incomplete or incorrect specifications? More importantly, how effective are they at helping find better ways to do things? Do use cases support creative problem-solving?

Table 1 summarizes a research framework for studying the need for, and effectiveness of, use cases in the UML.

CONCLUSION

In summary, review of the literature shows some variation in how use cases are defined and used. This is not surprising, given their (and the UML's) relatively short history. But these differences can also be attributed to the lack of a theoretical foundation. The UML began with as a "best practices" approach (Booch et al., 1999, p. 449) and has continued to

Table 1. A framework for empirical research on use cases

Research Question	**Primary Independent Variable**	**Primary Dependent Variable**	**Methodology**
Do design/implementation details in use cases impede process redesign efforts?	Use case structure	Process innovation	Experiment; Case study
Can class diagrams be effectively extracted from use cases?	Use cases	Class diagram completeness	Case study; Experiment; developer surveys
Do use cases facilitate communication between developers and users?	Communication medium (use cases or class diagrams)	User understanding domain coverage	Experiment; user surveys

develop that way. The authors of books on the UML and use cases generally offer little more than their experience as evidence that their approach works. While a valuable contribution, research is needed to support these "best practices" claims (which are sometimes inconsistent).

REFERENCES

Adolph, S., & Bramble, P. (2003). *Patterns for effective use cases.* Boston: Addison-Wesley.

Arlow, J., & Neustadt, I. (2004). *Enterprise patterns and MDA: Building Better software with archetype patterns and UML.* Boston: Addison-Wesley.

Armour, F., & Miller, G. (2001). *Advanced use case modeling.* Boston: Addison-Wesley.

Bennett, S., Skelton, J., & Lunn, K. (2001). *Schaum's outline of UML.* New York: McGraw-Hill.

Bittner, K., & Spence, I. (2003). *Use case modeling.* Boston: Addison-Wesley.

Booch, G. (1994). *Object-oriented analysis and design with applications* (2nd ed.). Redwood City, CA: Benjamin/Cummings.

Booch, G., Jacobson, I., & Rumbaugh, J. (1999). *The unified modeling language user guide*. Reading, MA: Addison-Wesley.

Cockburn, A. (2001). *Writing effective use cases*. Boston: Addison-Wesley.

Constantine, L. L., & Lockwood, L. A. D. (2000). Structure and style in use cases for user interface design. In M. Van Harmelen & S. Wilson (Eds.), *Object modeling user interface design*. Reading, MA: Addison-Wesley.

Denny, R. (2005). *Succeeding with use cases: Working smart to deliver quality*. Upper Saddle River, NJ: Addison-Wesley.

Dewitz, S. (1996). *Systems analysis and design and the transition to objects.* New York: McGraw-Hill.

Dobing, B., & Parsons, J. (2006). How UML is used. *Communications of the ACM, 49*(5), 109-113.

Grossman, M., Aronson, J., & McCarthy, R. (2005). Does UML make the grade? Insights from the software development community. *Information and Software Technology, 47*(6), 383-397.

IBM Object-Oriented Technology Center. (1997). *Developing object-oriented software.* Upper Saddle River, NJ: Prentice Hall.

Jacobson, I. (1987). Object-oriented development in an industrial environment. *OOPSLA'87 Conference Proceedings, SIGPLAN Notices, 22*(12), 183-191.

Jacobson, I., Booch, G., & Rumbaugh, J. (1999). *The unified software development process*. Reading, MA: Addison-Wesley.

Jacobson, I., Christerson, M., Jonsson, P., & Overgaard G. (1992). *Object-oriented software engineering: A use case driven approach.* Reading, MA: Addison-Wesley.

Jacobson, I., Ericsson, M., & Jacobson, A. (1994). *The object advantage: Business process reengineering with object technology.* Reading, MA: Addison-Wesley.

Kulak, D., & Guiney, E. (2000). *Use cases: Requirements in context*. New York: ACM Press.

OMG. (2005). *Unified modeling language: Superstructure, Version 2.0*, formal/05-07-04. Retrieved from http://www.omg.org/technology/documents/formal/uml.htm

Övergaard, G., & Palmkvist, K. (2005). *Use cases: Patterns and blueprints.* Indianapolis, IN: Addison-Wesley.

Rubin, K., & Goldberg, A. (1992). Object behavior analysis. *Communications of the ACM, 35*(9), 48.

Rumbaugh, J., Blaha, M., Premerlani, W., Eddy, F., & Lorensen, W. (1991). *Object-oriented modeling and design.* Englewood Cliffs, NJ: Prentice Hall.

Rumbaugh, J., Jacobson, I., & Booch, G. (2005). *The Unified modeling language reference manual* (2nd ed.). Boston, Addison-Wesley.

White, I. (1994). *Rational rose essentials: Using the Booch method.* Redwood City, CA: Benjamin/Cummings.

Wirfs-Brock, R., Wilkerson, B., & Wiener, L. (1990). *Designing object-oriented software*. Englewood Cliffs, NJ: Prentice Hall.

Zeichick, A. (2002). *Modeling usage low; developers confused about UML 2.0, MDA. SD* Times. Retrieved July 15, from http://www.sdtimes.com/news/058/story3.htm

KEY TERMS

Actor: An actor plays one or more roles in relation to a set of use cases. An actor could correspond to a job title (e.g., purchasing agent, sales clerk) or can be non-human (e.g., another system, device, or database). Each actor in a use case must be directly involved at some point and is not merely a stakeholder (someone or something that is affected by the success or failure of a particular transaction).

Use Case: A use case describes a way in which a system can be used to provide value to one or more actors. They take a client perspective, focusing on how users will interact with the system.

Use Case Description: See use case narrative.

Use Case Diagram: The use case diagram shows one or more use cases (by title) and the actors involved in them. This provides an overview of the use case structure and also shows how each actor is involved in a system.

Use Case Model: Together, the use case diagrams and use case narratives form the use case model. This identifies all the actors and describes the functionality of the system.

Use Case Narrative: A use case narrative is a largely text-based description of a use case that could be supplemented with decision trees or other easily understood notations. The description should be written in the user's language, and thus provides an important communication tool between developers of systems and the intended users. Narratives follow a structured format, typically using a numbered sequence of steps for the main activity accompanied by preconditions, post-conditions, alternative or exception paths, etc.

Use Case Scenario: A scenario, as specified in the UML, is an instance of a use case that can help illustrate its use. For example, a single use case (rent video) might have different scenarios for renting a video to a child, a new customer, an existing customer with overdue videos, etc. However, the term is also used in other ways outside the UML.

Category: Electronic Business

The Use of Electronic Banking and New Technologies in Cash Management

Leire San Jose Ruiz de Aguirre
University of Basque Country, Spain

INTRODUCTION

The use of new information and communication technologies (ICT) as a business tool has increased rapidly for the past 10 years (Bonsón, Coffin, & Watson, 2000; Claessens, Glaessner, & Klingebiel, 2000; Vasarhelyi & Greenstein, 2003). More specifically, financial software, e-banking, and the Internet, as core aspects of the various technologies used, have become driving forces behind the expansion of firms and the development of cash management. New technologies are considered as one of the most attractive ways for businesses to increase revenue and achieve economies of scale that can reduce unit costs (Ballantine & Stray, 1998; Barajas & Villanueva, 2001; Daniel, 1999; Daniel & Storey, 1997; Deyoung, 2001; Downes & Muy, 1998; Faulder, 2001; Jayawardhena & Foley, 2000).

There are different studies about the use of ICT in the management of the enterprise that explain the obtaining of enterprise performance. Brynjolfsson and Hitt (2000) and Nájera (2005) have done a review of these works and a classification of these types of researches. Unfortunately, there are not specific works or empirical researches about the use of e-banking in cash management; consequently, this work is focused in this.

The rest of the chapter is structured as follows. The theoretical foundation on which the study is based is explained in Section 2. Section 3 presents the data and the analysis procedure used to conduct the empirical study. The main results of the investigation are shown in Section 4, and Section 5 presents conclusions. The chapter ends with a list of bibliographical references.

THEORETICAL FOUNDATION: E-BANKING IN FINANCIAL PRACTICES

Three different periods can be distinguished in the development of ICT in cash management (Williams, Chen, & Russell, 1997). In period one, prior to the 1970s, treasurers engaged in accounting and in managing the cash-flow of their companies, and did not use IT tools in their work. Period two, from the 1970s to the 1990s, is characterised by a vision based on corporate relations and integrated systems. Since the 1990s we have moved into period three, the era of networking, in which the responsibilities of cash managers have come to include the use of electronic banking and new technologies to obtain the efficiency in their financial decisions because the great advantages for business management entailed by the development of technology. Internet is a space that can be shared freely at zero expense. However, the introduction of new technologies needs to be analysed thoroughly if business management efficiency is to be maximised (Levinsohn, 2001).

In this context, specifically, electronic banking management becomes an essential function in which information can be obtained electronically on market conditions, financial products, trends, and financial services. Financing and investment of treasury deficit and surpluses is optimised by comparing the terms of the different financial products on the market, and then contracting products online (Mooney & Pittman, 1996; Vasarhely & Greenstein, 2003; Welch, 1999).

In short, financial services based on new technologies use the e-banking as a single communication standard and thus, obtain economies of scale (Barajas & Villanueva, 2001; Eije & Westerman, 2002; Mishkin & Strahan, 1999) and positive synergies at treasury departments that were formerly difficult to achieve.

METHOD AND SAMPLE

To draw up the explanatory model of e-banking use in cash management, we used an exploratory factorial analysis of variables with Version 14.0 of the SPSS program.

In the following table, we have described the sample that is considered representative of the population of Spanish firms. This study was conducted on Spanish firms with more than 10 employees. The sample was chosen by proportional allocation according to criteria of company size (defined by the number of employees) and sector of activity. The total number of firms used was 501, and the error is smaller than 5%, necessary in this type of study.

Table 1. Acknowledgements: The sample

SAMPLE	THE CLASIFICATION OF THE SAMPLE	RANDOM ERROR	INFORMATION COLLECTION TECHNIQUE. TIME
501 valid questionnaires. The interviewed person was the finance manager or cash manager.	Criteria: company size (defined by the number of employees) and sector activity.	± 3, 52% with a confidence level of 95,5%, p=q=0.5,	Telephone's interview. June of 2005.

RESULTS

The results indicate that the new technologies more utilized by firms to financial practices are financial software, Internet and electronic banking. Furthermore, these results have permitted us to develop an explanatory model of the use of electronic banking to treasury management.

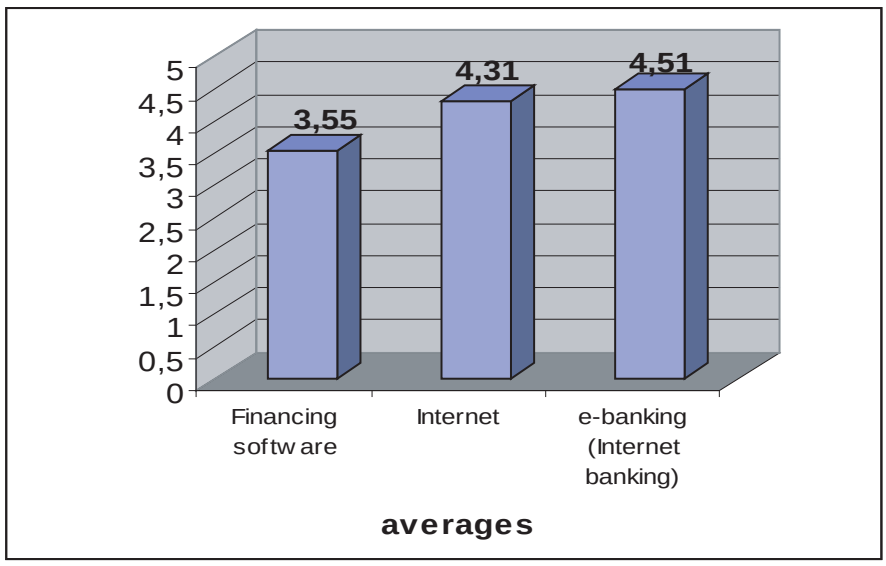
Preliminary Results

The Use of ICT in Cash Management

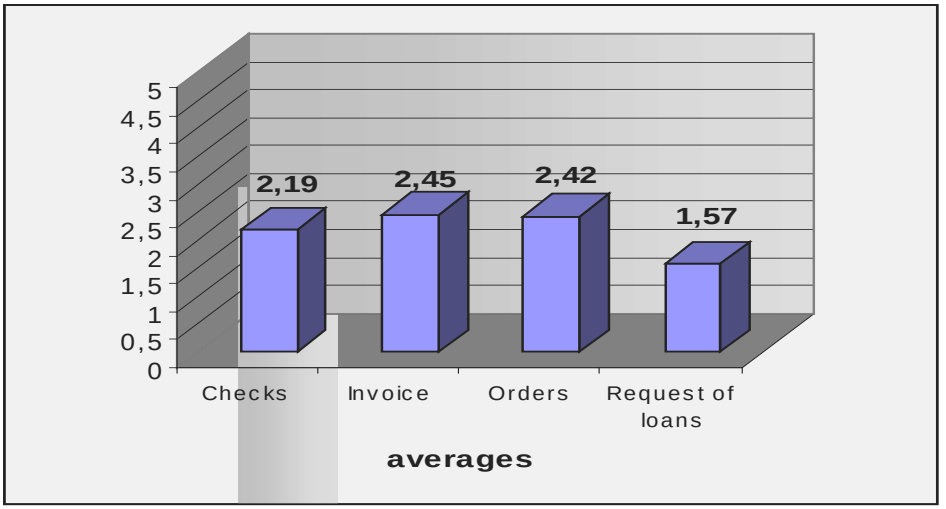
The ICT's most widely used in financial operations and more specifically, in treasury management are financial software, the Internet, and e-banking, though it is the introduction of the Internet into all areas of corporate life that has been the major revolution of the past 10 years. All these technologies entail benefits for financial management, so the next step is to analyse their average levels of use and determine which ICT's are most widely used in this area (see Graph 1).

This analysis shows that e-banking (Internet banking) is the most widely used tool in treasury operations, with treasury managers awarding it an average score of 4.512 out of 5. The second highest score is that of the Internet, with 4.312, followed by financial software with an average of less than 4. Specifically, e-banking is used habitually by 73.1% of the firms analysed, the Internet by 65.5% and financial software by 44.6%.

Graph 1. The level of utilization of ICT in cash management: Averages



Graph 2: Electronic financial instruments: Averages



Electronic Financial Instruments

In this sense, we analyzed electronic financial instruments, the financial instruments via e-banking mostly used by the companies. Concretely, four financial instruments have been analyzed that are emitted via electronic bank; the check, the invoices, the orders, and the request of loans (see Graph 2). The scale has been used starts in 1, which represents the smallest use, and finishes in 5, the most used.

The invoice (2,45) is, on the average, the financial instrument emitted by e-banking that is used more, followed by the orders (2,42) and checks (2,19). The request of loans (1,57) via electronics is considerably inferior. These data, in general, denote that the financial instruments via e-banking are not used habitually in the companies, although the electronic invoices and orders are used.

Explanatory Model: E-banking in Cash Management

Exploratory Factorial Analysis

The basic assumptions underlying factorial analysis, linearity, normality, and homoscedasticity, are conceptual rather than statistical. Therefore, from a statistical point of view, these assumptions can be obviated in the awareness that their fulfilment causes a drop in the correlations observed (Hair, Anderson, Tatham, & Black, 1999). However, these correlations are still sufficient if it is determined that factorial analysis is appropriate. This can be done by analysing the Kaiser-Meyer and Olkin (KMO) measurement and examining the whole correlation matrix, contrasting it with Bartlett's sphericity test.

The results shown in Table 2 are satisfactory for both tests, so an exploratory factorial analysis can be performed for e-banking.

These results (Table 3) show that the eight variables for the use of e-banking in treasury management can be grouped into two components with minimal information loss. The first component explains 41,728% of the variance, the second 28,725%. In all, this grouping into two factors explains 70,453%[1] of the overall variability of the sample.

Saturations lower than 0.6 in absolute value have been eliminated.

An analysis of the sensitivities in Table 4 shows that for the first component, *negotiation with financing institutions, management of the financing of treasury deficit, management of the placement of treasury surpluses and interest-rate, and exchange-rate risk management* have high, positive values. Considering the significance of these variables, this component seems to be reflecting aspects concerned with ***advances use of e-banking in cash management.***

Use of e-banking in *collects and payments management, day-to-day control of banking positions, short-term treasury forecasts, and monitoring of banking positions at the value*

Table 2. Determining factor of the correlation matrix, KMO, and Bartlett´s test

Kaiser-Meyer-Olkin simple suitability measure		,806
Bartlett´s sphericity test	Chi-square	1004,000
	fd	28
	p-value	,000

Table 3. Principal component analysis. Final statistics with three components of rotate variables

	Communality	Comp.	Eigen-value	% of Var.	%Var. Accum.
Collects and Payments management	,638	1	3,343	41,728	41,728
Day-to-day control of banking positions	,648	2	1,494	28,725	70,453
Short-term treasury forecasts	,587				
Monitoring of banking positions at the value data	,474				
Negotiation with financing institutions	,585				
Management of the financing of treasury deficit Management of the placement of treasury surpluses Interest-rate and exchange-rate risks management	,681				
Management of the placement of treasury surpluses	,622				
Interest-rate and exchange-rate risks management	,602				

Table 4. Rotated component matrix; Varimax normalization with Kaiser

	COMP. 1	COMP. 2
Collects and Payments management		,798
Day-to-day control of banking positions		,792
Short-term treasury forecasts		,709
Monitoring of banking positions at the value data		,663
Negotiation with financing institutions	,746	
Management of the financing of treasury deficit Management of the placement of treasury surpluses Interest-rate and exchange-rate risks management	,813	
Management of the placement of treasury surpluses	,772	
Interest-rate and exchange-rate risks management	,764	

data can be grouped around the second factor as ***basic use of e-banking in cash management.***

The Explanatory Model of Use E-Banking in Cash Management

The use of ICT in cash management is expanded in the last decade. The electronic cash management is focused in the use of new technologies and particularly, e-banking in treasurer's financial practices that permit they obtain more information to select the best financial decision that is conduced to obtain economic scales in the enterprise and reduction of transaction cost with positive synergies. There are two levels to use e-banking, the first level, basic e-banking is only used in the repetitive actions of the cash managers, the second level, we have denominated advance e-banking, and it is used in strategic financial practices (Graph 3). Is important to use e-banking in two levels, but nowadays some enterprises, small and medium principally, have not used advance e-banking to cash management.

CONCLUSIONS

The ICT more utilized by firms to financial practices is electronic banking. The principal electronic financial instrument that enterprises use to cash management is the invoice. Furthermore, these embrace not only the most repetitive treasury functions denominated as basic e-banking referred to use of e-banking to collect and payment management, but also they are used in treasury management functions that depend largely on corporate decisions and are strategic rather than operational denominated as advance e-banking.

REFERENCES

Ballantine, J. A., & Stray, S. J. (1998). Financial appraisal and the IS/IT investment decision making process. *Journal of Information Technology*, *13*, 3-14.

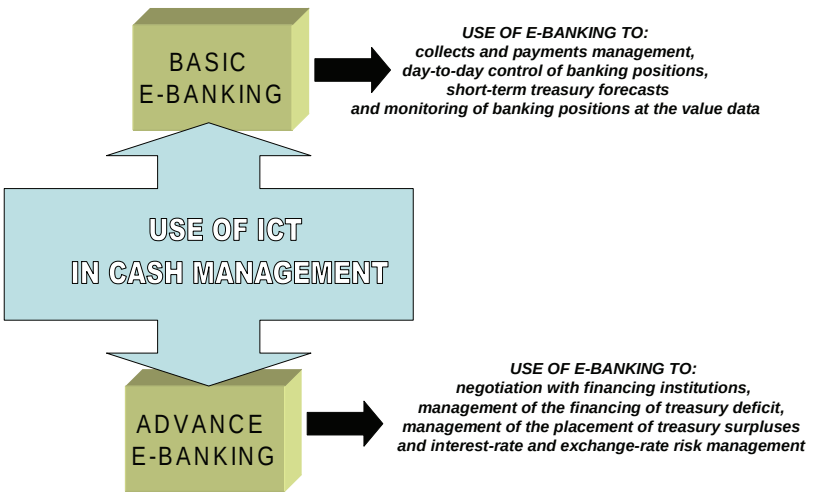
Barajas, A., & Villanueva, M. (2001). Escenario de la banca en Internet. *Banca y Finanzas*, *66*, 29-32.

Bonsón, E., Coffin, Z., & Watson, L. (2000). Un lenguaje para el reporting digital. *Partida Doble*, *17*, 16-22.

Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspective*, *14*(4), 23-48.

Claessens, S., Glaessner, T., & Klingebiel, D. (2000). *Electronic finance: reshaping the financial landscape around the world.* Working Paper, Nº 4, 1-26. Retrieved from http://www.ssrn.com.

Graph 3. The different use of e-banking in cash management



Daniel, E. (1999). Provision of electronic banking in the UK and the Republic of Ireland. *International Journal of Bank Marketing*, *17*(2), 72-82.

Daniel, E., & Storey, C. (1997). Online banking: Strategic and management challenges. *Long Range Planning*, *30*(6), 890-898.

Deyoung, R. (2001). The financial performance of pure play Internet banks. *Economic Perspectives, Federal Reserve Bank of Chicago*, *25*(1), 60-76. Retrieved from http://www.chicagofed.org

Downes, L., & Muy, C. (1998). *Killer app-digital strategic for market dominance*. Boston: Harvard Business School Press.

Eije, H., & Westerman, W. (2002). Multinational cash management and conglomerate discounts in the euro zone. *International Business Review*, *April*, 1-25.

Faulder, G. (2001). Foster's choose treasury. *Corporate Finance*, *198*, 25-26.

Hair, J. F., Anderson, R. E., Tatham, R. I., & Black, W. C. (1999). *Análisis Multivariante* (5ª ed,). Madrid: Prentice-Hall.

Jayawardhena, C., & Foley, P. (2000). Changes in the banking sector-The case of Internet banking in the UK. *Internet Research: Electronic Networking Applications and Policy*, *10*(1), 19-30.

Levinsohn, A. (2001). The wild, wired world of e-finance. *Strategic Finance*, *82*, 27-32.

Mishkin, F., & Strahan, P. E. (1999). *What will technology do to financial structure?* Working Paper, Nº 6892, National Bureau of Economic Research. Retrieved from http://www.nber.org

Mooney, J. L., & Pittman, W. D. (1996). A guide to electronic commerce. *Management Accounting*, *78*(3), 43-47.

Nájera, J. J. (2005). El estudio del impacto de la tecnología de la información sobre los resultados empresariales: una revisión de la literatura. *XV Congreso Nacional Acede*, La Laguna.

Vasarhelyi, M., & Greenstein, M. (2003). Underlying principles of the electronization of business: A research agenda. *International Journal of Accounting Information Systems*, *4*, 1-25.

Welch, B. (1999). *Electronic banking and treasury security*, (2nd ed.). Blackwell, Oxford.

Williams, B. C., Chen, J. C., & Russell, P. O. (1997). Understanding changes in systems, accounting and auditing: the impact of EDI. *Managerial Auditing Journal*, *12*(6), 298-304.

U

KEY TERMS

Cash Management: The cash management corresponds to the obtaining of available the necessary one, at the suitable moment, to the smaller possible cost for which the treasury is planned, is decided what short-term financing and investment to make, analyze the relations with the financial organizations and the risks are managed. In addition, the pursuit and the analysis of the management of the circuit of collections and payments are essential, along with the enterprise culture.

E-Banking: The electronic bank consists of the use of electronic channels by means of which the financial institutions can send products or offer the banking services. Between the services and products, they are possible to be included, deposits, financial management of accounts, warnings, payments of electronic accounts, and provision of other products of electronic payments.

Electronic Cash Management: It is possible to be defined as the set of procedures and practices of integrated management of treasury with the developments in the technologies of the information.

Electronic Financial Instruments: They consist of financial products that are contracted, emitted, and paid without the use in paper of financial documents.

ICT in Cash Management: It consists of the use of different techniques, such as Internet, Intranet, software, electronic bank, and the bank by Internet, with object to essentially obtain the efficiency of the management of treasury by means of the reduction of costs.

Information and Communication Technologies (ICT): The ICT has defined as the grouping of the technologies of information, that they are characterized by the technologies of registries of contents (computer science, communications, Telematics), and the technologies of the communication, that essentially group the radio, the television and the telephony.

Internet Banking: From the global concept of electronic bank, we considered that the bank by Internet consists of the use of the Internet channel like communication channel, banking product distribution, and hiring on the part of the financial organizations.

ENDNOTE

[1] The number of factors extracted was determined by prioritising the "percentage of variance" criterion over the "latent root" criterion so that with commonalities above the set minimum of 0.5 it was decided to select the number of factors necessary to explain at least 60% of the variance.

The Use of ICTs in Small Business

Stephen Burgess
Victoria Universtiy, Australia

INTRODUCTION

This article examines the main drivers and barriers facing small business owner/managers in the manner in which they use information and communications technologies (ICTs) within their businesses. The early part of the article examines the notion of what is meant by small business. The discussion then moves onto describing some of the drivers and barriers to the use of ICTs in small business and the implications of these to small businesses.

BACKGROUND

What is Small?

When studying the use of ICTs in small business, the range of definitions used to describe small business ranges from micro businesses to small and medium sized enterprises. This range can make it extremely difficult for researchers to match up different small business studies. A 2003 study by members of the information resources management association special research cluster on small business and information technology (Burgess, 2003) found that:

- Definitions of small business ranged from less than 20 (Australasia), 50 (Europe), and 100 (North America) employees (with some definitions including annual turnover and asset levels)
- Definitions of micro business ranged from less than 5 to less than 10 employees
- Definitions of medium business ranged up to 200, 250 and 500 employees

A common acronym used to represent small and medium sized businesses is SME. There is some argument as to whether the term is of any use at all given the vast differences between small and medium sized businesses. Still, there continue to be studies that examine the use of ICTs in SMEs.

For my purposes, small businesses are those businesses with 20 employees or less. However, most small businesses are micro businesses (Fillis, Johansson, & Wagner, 2004a), made up of less than five employees.

Size is Important!

Why is it so important to consider the size of the business? A number of studies suggest that there is a relationship between the size of a business and its level of adoption of ICTs (McDonagh & Prothero, 2000). There is also a relationship between the size of a business and the different characteristics it will have that can lead to the successful use of ICTs (Igbaria, Zinatelli, Cragg, & Cavaye, 1997; Pollard & Hayne, 1998). As such, research findings based upon traditional information systems in larger businesses are not necessarily directly applicable to small businesses. This will be touched upon again later in the article.

DRIVERS AND BARRIERS TO THE USE OF ICT IN SMALL BUSINESS

The literature around the area of small business and information technology is rife with what is now a fairly accepted list of barriers to the successful implementation of ICTs in small businesses. These barriers typically include (Igbaria et al., 1997; Management Services, 1997; McDonagh et al., 2000; Pollard et al., 1998):

- The cost of ICTs—this is perhaps not so much of an issue these days (well, in developed countries anyway)
- Lack of time to devote to the implementation and maintenance of ICTs
- A lack of ICTs knowledge combined with difficulty in finding useful, impartial advice
- Lack of use of external consultants and vendors
- Short-range management perspectives
- A lack of understanding of the benefits that ICTs can provide, and how to measure those benefits
- A lack of formal planning or control procedures

What about small businesses and e-commerce? Small businesses do face a series of barriers that they need to overcome before moving into the digital economy. These are (Taylor & Murphy 2004b):

- Many SMEs are unaware of the potential of e-commerce to enhance their business.

- Some SMEs occupy clearly defined (and small) niche markets that they are satisfied with. They do not need the extended connectivity provided by the Internet. In addition, small businesses may be happy with their existing activities because they are adequate enough to enable to maintain a particular "lifestyle" (Fillis et al., 2004a).
- There are still perceptions of unresolved security and privacy issues related to use of the Internet.
- Many SMEs lack the necessary skill base to engage in the digital economy.
- The perceived high initial and ongoing costs associated with ICTs and e-business can be seen as a barrier. This is also supported by Fillis et al. (2004a).
- Many SMEs cannot experiment with ICT investments like larger businesses. They need to be sure that there will be returns for the e-commerce investments they make.

Having identified some of the barriers to the successful use of ICTs, there is also a fairly common list of drivers that are listed in the literature that appear to indicate a greater chance of successful implementation of ICTs in small businesses.

Some of these factors are (Naylor & Williams, 1994; Yap & Thong, 1997; Swartz & Walsh, 1996; Zinatelli, Cragg, & Cavaye, 1996):

- The involvement of owner/managers in the implementation of ICTs
- The involvement of users (employees) in development and installation
- The training of users
- The use of disciplined planning methodologies in setting up applications
- The number of analytical/strategic (versus transactional) applications being run
- The level of ICTs expertise within the organisation
- The role of the external environment (especially consultants and vendors)

In relation to e-commerce, Mehrtens, Cragg, and Mills (2001) developed a model of Internet adoption using an innovation theory approach, suggesting that the decision to adopt was a based upon:

- **Identifying perceived benefits:** These can take the form of:
  - Efficiency benefits from the *relative advantage* that the Internet can provide over traditional methods
  - Employees gathering information in a more effective manner
  - A tool to build the image (or "brand") of the business
- Organisational readiness:
  - The level of Internet knowledge in the business amongst non-IT professionals, for instance the owner/manager; an Internet aware owner might support or initiate Internet adoption.
  - The suitability of systems within the business to access and use the Internet. Historically, smaller business sizes have been identified as a factor in lower ICT adoption rates and Internet technologies (Martin & Matlay, 2001). One important aspect here is if the infrastructure is available for businesses to take advantage of ICTs. This can especially be important in rural and remote areas and developing countries.
- External pressure:
  - From existing users, particularly customers but perhaps external business partners (such as suppliers).

We can now examine some of the factors mentioned in these lists more closely.

FACTORS IN THE ADOPTION OF ICT BY SMALL BUSINESSES

Role of Owner/Manager

One of the key factors leading to successful use of ICTs in small businesses identified in the previous section was the involvement of small business owner/ managers in the ICTs implementation.

There is some evidence to indicate that managers in small businesses are less likely to know how to use ICTs effectively or to keep up with the latest trends in ICTs than their counterparts in larger businesses (Pollard et al., 1988). Igbaria et al. (1997) cite a number of references to support the view that management support can promote the acceptance of ICTs. They found that the support of management positively affected the perceived ease of use and the perceived usefulness of ICTs within the small business.

The motivation of owner/managers is a key factor in shaping the e-business development of a small business. Their perceptions can range from highly positive entrepreneurial viewpoints to very conservative stances. The orientation of the business and the owner/manager can affect the level of Internet connectivity of the business. Business that exhibit entrepreneurial characteristics are more likely to adopt e-business (and will do so at a faster rate) (Fillis et al., 2004a). An important finding from Lee's (2004) study is that owner/managers with higher computer self-efficacy were more likely to adopt Internet applications. Indeed, Martin et al. (2001) suggest that the managerial knowledge,

skills, and experience of owner/managers can make a crucial difference in identifying Internet opportunities.

In a study of 59 Irish SMEs, Barry and Milner (2002) determined that the owner/manager was the principal driving force behind the decision to adopt e-commerce. Decision making within small businesses is often based upon owner/manager intuition rather than on the basis of market data gathered by the business (Schlenker & Crocker 2003).

One of the barriers identified earlier that hinder the effective use of ICTs in small businesses is a lack of formal planning and control methodologies. This relates to a lack of knowledge of how to plan effectively, lack of time and money to seek this knowledge, lack of time to apply ICTs even if they have the knowledge and a lack of understanding that they even need the knowledge! Small businesses are, however, concerned with issues relating to how they can operate more effectively and efficiently and/or how they can grow (El Louadi, 1998). One of the problems is that management practice in small businesses is often based on the short term and is informal and ad hoc. Much of the time is spent "surviving," so that little time can be devoted to examine ICTs projects (Pollard et al., 1988).

Location

Why is location important when considering the use of ICTs in small business? The major answer to this is a combination of *resources* and *distance***.** The further you are away from resources, the longer it takes and the more it costs to get them. This can particularly be the case with hardware and software purchases, training, and support.

Another reason for examining location is *culture***.** Some countries, and even different regions within countries, have their own traditions and their own established ways of doing things. This can influence the behaviour of small businesses and the manner in which they use ICTs.

In addition, distance from customers provides problems in relation to logistics, the regular requirement for a personal touch in closing sales or providing service and problems in relation to differing language and culture (Schlenker et al., 2003).

Developing Countries

Small businesses make up a major portion of businesses in developing countries (in some countries the percentage is higher than in developed countries). One of the major barriers faced by small businesses in developing countries is access to information, especially information used in decision-making. Another problem is the lack of data sources from which to obtain the type of information required. Problems with the technological infrastructure of developing countries only exacerbate this (Sawyerr, Edbrahimi, & Thibodeaux, 2000).

Taylor et al. (2004b) point to significant differences between e-business adoption in developed and developing countries. Many business people in developing countries are disillusioned with substandard telecommunications infrastructures, as well as a lack of expert advice and support systems in general (Schlenker et al., 2003). In relation to connecting to the Internet, the infrastructure in many countries is either not available or costs too much to connect to. Thus, new initiatives tend to occur in urban areas where affordable connections are available (Warren, 2004).

Rural Small Businesses

Some of the problems facing rural small businesses are similar to those facing small businesses in developing countries. One of the benefits that the Internet may provide is remote access to many desired ICTs resources such as training (Gallagher, 1999). A year 2000 survey of Australian small businesses revealed that 29% of metropolitan small businesses had a Web site, compared to 20% of rural small businesses. The main reason given by rural small businesses for not having a Web site was that they did not have access to the skills needed to design, build, and maintain a Web site (Telstra Corporation and NOIE, 2000). By 2005, the divide had increased, with 54% of metropolitan small businesses having Web sites, compared to 37% of regional small businesses (Sensis, 2005).

Martin et al. (2001) report that businesses located in rural regions of the UK had less access to ICT advice. SMEs in London and the Southeast region of the UK are more likely to have a Website than in other regions (Taylor et al., 2004b).

Industry

There is some relationship between the industry that small businesses are involved in and the types of ICTs that they use.

The level of e-business activity will vary across small businesses in different industry sectors (Fillis, Johansson, & Wagner, 2004b). In a study of the usage of ICTs by 24 businesses in Northern Ireland spanning the retail, construction, distribution, and wholesale industry areas Shiels, McIvor, & O'Reilly, (2003) discovered differences in the usage of ICTs in separate industry groupings. In some instances, they attributed these to the nature of their business operations and the type of customer they target. Martin et al. (2001) suggest that micro businesses specialising in business services are more likely to adopt ICTs than similar sized manufacturing businesses. Fillis et al., (2004b) expect that small businesses in the ICT industry would have a higher degree of uptake and usage of ICTs than non-ICT businesses. Small farms in the UK have rather low Internet connectivity rates. One of the reasons is a lack of suitable on-farm hardware. As at 2003,

a significant proportion of farms still did not have a PC, and for those that did many were old and slow (Warren, 2004). The hotel industry is an information intensive industry. As such, ICTs are a key enabler in distributing information and facilitating transactions (O'Connor & Frew 2004).

In a study of 129 SMEs in New Zealand, Al-Qirim and Corbitt (2002) found that a higher proportion of businesses in the ICT and business services industries had adopted applications using Internet technologies.

Working Out the Benefits of ICTs

Another barrier to the successful use of ICTs in small businesses is a lack of understanding of the benefits that ICTs can provide and how to measure those benefits. The most common way used to determine the level of ICTs success is to measure small business user satisfaction with information technology. Such measures of user satisfaction have one major problem—they are linked with user expectations (Naylor et al., 1994). For instance, an owner/manager understanding the strategic benefits that ICTs can provide may be less satisfied with a simple transactional system than an owner/manager who is unaware of these strategic benefits. This is despite the possibility that they may be reviewing systems that perform in a similar manner. Again, the problem falls back to a lack of proper knowledge about the advantages that ICTs can provide.

Success can vary in relation to industry and culture. The measurement of success can be measures such as quality of life, company culture, and maximising profits or generating revenues. The idea of quality of life as a goal is favoured by many smaller businesses and public organisations (Schlenker et al., 2003). In fact, Schlenker et al. (2003) suggest that 70% of businesses in the SME sector are more concerned with "quality of life" than the value of their stock. Their primary purpose is to generate an income for their owners rather than maximise revenues.

The value of technology to a business may therefore be in how it can assist management and employees to define, operationalise and measure success (Schlenker et al., 2003).

It is important to distinguish between the evaluation of a potential ICT investment in relation to determining whether or not to invest in it and the ongoing evaluation of an ICT project after it has been implemented. Similar issues affect both types of activities as the difficulties involved in assessing the financial value of ICT systems to a business are well documented. These systems often generate indirect, qualitative and contingent impacts that are difficult to measure in monetary terms. The projects often cross organisational boundaries and are related with investments in other investment projects, making the ICT component difficult to isolate (O'Connor et al., 2004).

Employee Skills and Training

Factors relating to a lack of knowledge of ICTs or lack of understanding of the benefits of ICTs have been mentioned a number of times already in this article. Human capital in small businesses is seen as being an important factor in recognising ICT opportunities and adopting them. Businesses where employees have a lower understanding and knowledge of ICTs have difficulties in understanding the usefulness of ICTs (Martin et al., 2001). Barry et al. (2002) suggest that human resource issues are perhaps the most important factor to consider when implementing new technologies in an SME. Businesses with a mix of technological and business skills are generally better placed to exploit e-business opportunities (Fillis et al., 2004a).

In relation to development of e-business systems, Taylor et al. (2004a) observe that Web-based systems development is different to other types of ICT development and requires a wide range of skills. Taylor et al. identified the skills sets required for e-commerce projects in the SME sector, incorporating technical, business, and overlapping skills and knowledge.

Developing Skills and Knowledge in Small Businesses

Barry et al. (2002) have identified a number of predominant features of SME training:

- Very small businesses are the least interested in providing training.
- Training is often considered to be too general and not specific enough to the needs of SMEs.
- SMEs face disincentives to train employees. They have less potential to offer the higher pay and benefits that accompany increased expertise.
- SMEs are often not aware of their training needs and are not capable matching needs to suitable training options.
- Training in SMEs is more likely to be informal than formal—with the more formal approaches providing too many constraints in relation to the consumption of time and other resources.

Barry et al. (2002) found that the owner/manager was a catalyst for organisational e-commerce training in SMEs and that some 85% of the 59 Irish SMEs in their study did provide training, but most of it was delivered internally and was informal in nature.

In a case study of three small businesses, Taylor et al. (2004a) encountered the following techniques for developing e-business skills:

- Hands-on experience
- Short courses
- Higher education courses
- Technical manuals
- Viewing the Web sites of other businesses; this allowed a benchmarking exercise to be performed against rival e-business Web sites.
- Consulting experienced staff—mainly external ICT consultants or external software house staff.

Simpson and Docherty (2004) discuss the role of personal business advisors (small business consultants in the UK)—who typically had marketing or finance backgrounds but could not equal the knowledge of the owner/manager in their particular industry. Also, their fee structures were not quite suited to small businesses.

Another option for small businesses, especially ones that may feel isolated, it to set up a mentoring system where mentors are able to provide external and relevant experience and impartial advice (Simpson et al., 2004).

Building Capacity through External Sources

Small business managers perceive more uncertainty in their environment than their counterparts in larger businesses. Effective management of external information can help them to reduce the level of uncertainty that they feel (El Louadi, 1998).

In many instances, small businesses have to rely on the ICT expertise of vendors and/or consultants because of a lack of internal ICT expertise. Igbaria et al. (1997) have found that good external support provided by vendors and/or consultants, such as technical support, training and a harmonious working relationship can reduce the risk of ICTs failure in small businesses.

There is a view, however, that vendors and consultants do not understand the small and medium business market and that the level of support provided by them is only adequate or less than adequate (Management Services, 2000). Careful selection of vendors and/or consultants is vital.

Governments worldwide are beginning to realise the importance of the small business community. The role of government in developing countries has already been touched upon in this article. ICTs are one of the areas that are the subject of increased government resources, through improved information programs, increased training opportunities and technology support grants and awards.

As time goes on, governments are increasingly becoming aware of the importance of small businesses to their economies and are providing increased resources to their support. This has resulted in various support programs for small businesses that have directly or indirectly resulted in improvements in the use of ICTs by small businesses. It is anticipated that this trend will continue. There is also an increasing awareness of the importance of the efficient and effective use of ICTs by small businesses by the research community. There continues to be increased amounts of research being carried out in the field, which can only lead to greater understanding.

CONCLUSION

This article has introduced a number of issues related to the use of ICTs in small businesses. When looking at small business research, it is important to determine what the researcher's view of "small" actually is. This is so that proper comparisons can be made across studies.

Drivers and barriers in relation to the use of ICTs were identified for small businesses as being common areas covered in the literature. The importance of owner/managers, the location of the business, differences between industries, and employee skill levels and training were discussed.

REFERENCES

Al-Qirim, N., & Corbitt, B. (2002). An empirical investigation of an e-commerce adoption model in small to medium-sized enterprises in New Zealand. In *Proceedings of the 6th Pacific Asia Conference on Information Systems (PACIS 2002): The Next e-what? for Business and Communities* (pp. 2-4).

Barry, H., & Milner, B. (2002). SMEs and electronic commerce: A departure from the traditional prioritisation of training? *Journal of European Industrial Training, 26*(7), 316-326.

Burgess, S. (2003). *A definition for small business?* IRMA Special Research Cluster - Small Business and IT, Melbourne, Australia.

Burgess, S. (1997). *Information technology and small business: a categorised study of the use of it in small business.* Detailed Survey Report, Small Business Victoria, Melbourne.

El Louadi, M. (1998). The relationship among organization structure, information technology, and information processing in small Canadian firms. *Canadian Journal of Administrative Sciences, 15*(2), 180-199, June.

Fillis, I., Johansson, U., & Wagner, B. (2004a). A qualitative investigation of smaller firm e-business development. *Journal of Small Business and Enterprise Development, 11*(3).

Fillis, I., Johannson, U., & Wagner, B. (2004b). Factors impacting on e-business adoption and development in the smaller firm. *International Journal of Entrepreneurial Behaviour and Research, 10*(3), 178-191.

Gallagher, P. (1999). E-commerce trends. *International Trade Forum*, *35*(2), 16-18.

Igbaria, M., Zinatelli, N., Cragg, P., & Cavaye, A. L. M. (1997). Personal computing acceptance factors in small firms: A structural equation model. *MIS Quarterly*, *21*(3), 279-305.

Lee, J. (2004). Discriminant analysis of technology adoption behavior: A case of Internet technologies in small businesses. *Journal of Computer Information Systems*, *44*(4), 57-66.

Martin, L. M., & Matlay, H. (2001). "Blanket" approaches to promoting ICT in small firms: Some lessons from the DTI ladder adoption model in the UK. *Internet Research*, *11*(5), 399-410.

Management Services. (1997). Computers fail to click with small businesses. *Enfield*, *41*(9), 4

Management Services. (2000). Nearly half of SMEs believe that the Internet and IT has no impact on them. *Enfield*, *44*(10), 6.

McDonagh, P., & Prothero, A. (2000). Euroclicking and the Irish SME: Prepared for e-commerce and the single currency? *Irish Marketing Review*, *13*(1), 21-33.

Mehrtens, J., Cragg, P. B., & Mills, A. M. (2001). A model of Internet adoption by SMEs. *Information and Management*, *39*(3), 165.

Naylor, J. B., & Williams, J. (1994). The successful use of IT in SMEs on Merseyside. *European Journal of Information Systems*, *3*(1), 48-56.

O'Connor, P., & Frew, A. J. (2004). An evaluation methodology for hotel electronic channels of distribution. *Hospitality Management*, *23*(2), 179-199.

Pollard, C. E., & Hayne, S. C. (1998). The changing faces of information systems issues in small firms. *International Small Business Journal*, *16*(3), 70-87.

Sawyerr, O. O., Edbrahimi, B. P., & Thibodeaux, M. S. (2000). Executive environmental scanning, information source utilisation, and firm performance: The case of Nigeria. *Journal of Applied Management Studies*, *9*(1), 95-115.

Schlenker, L., & Crocker, N. (2003). Building an e-business scenario for small business: The IBM SME Gateway project. *Qualitative Market Research: An International Journal*, *6*(1), 7-17.

Sensis. (2005). *Sensis E:Business Report: The Online Experience of Small and Medium Enterprises*, Sensis, Australia.

Shiels, H., McIvor, R., & O'Reilly, D. (2003). Understanding the implications of ICT adoption: Insights from SMEs. *Logistics Information Management*, *16*(5).

Simpson, M., Docherty, A. J. (2004). E-commerce adoption support and advice for UK SMEs. *Journal of Small Business and Enterprise Development*, *11*(3).

Swartz, E., & Walsh, V. (1996). Understanding the process of information management in small firms: Implications for government policy. The *19th ISBA National Conference Proceedings* (pp. 387-399). Birmingham.

Taylor, M. J., McWilliam, J., England, D., & Akomode, J. (2004). Skills required in developing electronic commerce for small and medium enterprises: Case based generalization approach. *Electronic Commerce Research and Applications*, *3*(3), 253-265.

Taylor, M., & Murphy, A. (2004). SMEs and e-business. *Journal of Small Business and Enterprise Development*, *11*(3).

Telstra Corporation and NOIE (The National Office for the Information Economy). (2000). *Small business index: Survey of computer technology and e-commerce in Australian small and medium businesses.* Pacific Access Pty Ltd, Melbourne, Australia.

Warren, M. (2004). Farmers online: Drivers and impediments in adoption of Internet in UK agricultural businesses. *Journal of Small Business and Enterprise Development*, *11*(3), 371-381.

Yap, C., & Thong, J. Y. L. (1997). Programme evaluation of a government information technology programme for small businesses. *Journal of Information Technology*, *12*, 107-120.

Zinatelli, N., Cragg, P. B., & Cavaye, A. L. M. (1996). End user computing sophistication and success in small firms. *European Journal of Information Systems*, *5*(3), 172-181.

KEY TERMS

Medium Business: This term is used to describe businesses that are too large to be considered as being small and too small to be considered as being large. This somewhat vague description is matched by the varying definitions of "medium" sized business there are around. In relation to use of IT, medium sized businesses usually exhibit more of the characteristics of larger businesses than smaller ones.

Micro Business: This term is used to describe very small businesses. Many of these are operate as family businesses. They form the majority of businesses.

Owner/Manager: In most small businesses, the owner/manger is the driving force behind the business, and as such can be the catalyst for change in the business that occurs through or around IT. In some small businesses, the positions of owner and manager are separated.

Small Business: "Small business" can be measured by number of employees, annual turnover and/or assets. It usually represents those businesses with up to 20, 50, or 100 employees (depending upon the region being investigated). This term encompasses micro businesses.

SME: This acronym is used to refer to small and medium sized enterprises as a collective group.

User Modeling and Personalization of Advanced Information Systems

Liana Razmerita
University of Galati, Romania

INTRODUCTION

Enterprise information systems are among the key enablers of the leadership agility on competitive market places and therefore the design of advanced information systems (IS) is a continuous challenge for modern organizations. IS can include knowledge management systems (KMS), customer relationship management solutions, business-to-business applications, e-commerce, e-government or e-learning systems. Advanced IS featuring intelligence have recently implemented as complex applications with modular architecture relying on Web services, Semantic Web technology integrating user modeling, machine learning approaches, and/or agent-based technology. Personalization and more recently contextualization have emerged as key issues for achieving intelligent features in advanced IS.

In general, the goal of personalization is to improve the efficiency of interaction with the users, to simplify the interaction, and to make complex systems more usable. Blom (2000) distinguishes between two main roles of personalization: (1) to facilitate work and (2) to accommodate social requirements. In the first category he includes enabling access to information content, accommodating work goals, and accommodating individual differences, while the second category contains eliciting an emotional response and expressing identity.

This chapter presents a set of personalization techniques for IS. The second section provides background information related to personalization of IS, and it proposes a set of personalization mechanisms. In the third section, these personalization mechanisms are exemplified in the context of KMS. The fourth section outlines future work related to personalization of IS. Finally, the fifth section summarizes the main ideas presented in this article.

BACKGROUND

Personalization techniques enable IS to adapt their structure and content to match the needs and preferences of users based on a user model, which is stored, inferred, or updated dynamically. A simplified form of a user modeling system including personalization mechanisms is represented in Figure 1.

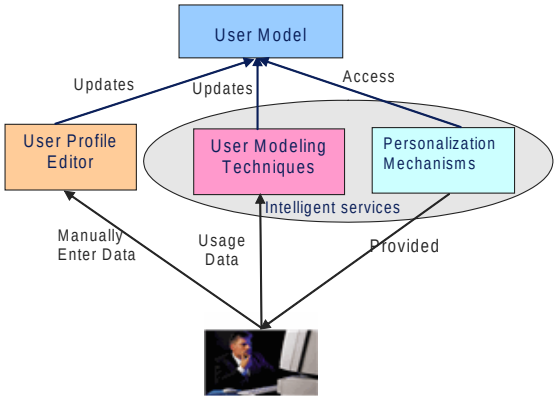
The user modeling server acquires and maintains the user's data through a user profile editor (explicitly) and through different user modeling techniques (implicitly). Among the most used techniques for implicitly constructing user models, acquiring user data, and deriving new facts are: logic-based techniques, stereotype-based reasoning, machine learning techniques, and reasoning with uncertainty (using either Bayesian networks or fuzzy logic techniques). User modeling techniques for personalization mechanisms are extensively overviewed in Razmerita (2003).

The following definition is proposed for personalization of IS: "Personalization of Information Systems is the process that enables interface customization, adaptation of the functionality, structure, content or/and presentation and modality in order to increase its relevance for its individual users." ISs can include customization and/or adaptive personalization features.

- **Customization is usually initiated by the user.** The user decides to select or exclude certain options from the interface. It is usually associated with interface customization. Many of the actual IS include customization features based on the user's preferences.
- **Adaptive personalization.** These features are triggered by the system, based on the user's interaction with the system, or based on the user's data available in the system. User's data are usually addressed as user profiles or user models. User models can be created by the users, who enter their data explicitly, or they can be inferred by the system. In the later case, the system tracks the user's activity with the system and infers characteristics of the user interacting with the system. These characteristics (e.g., domains of interest, goal) are further used for providing personalized interaction.

Agent-based systems can be used beyond adaptive personalization with different objectives. Agents can support users to perform different tasks, or they can perform tasks delegated by the users (e.g., intelligent information agents, personal assistants); they can search and guide users to find different knowledge assets (e.g., information filtering agents), or they can enhance learning processes (e.g., pedagogical agents) (Brna, Cooper, & Razmerita, 2001; Greer et al., 2001; Maes, 1998).

Figure 1. User modeling and personalization mechanisms



Important application areas of personalization include customer relationship management (Alpert, Karat, Karat, Brodie, & Vergo, 2003; Ardissono & Goy, 2002; Kobsa, Koenemann, & Pohl, 2001; Schafer, Konstan, & Riedl, 2001), educational software (Brusilovsky, 2001; Clark & Mayer, 2003), and information search and retrieval (Ardissono, Goy, Petrone, & Segnan, 2003; Kurki, Jokela, Sulonen, & Turpeinen, 1999; Tanudjaja & Mui, 2002; Waern, 2004,). Personalization has already proven its utility in e-commerce and e-learning. Fink and Kobsa (2001) provide data from communication reports showing that personalization based on purchased data and personal data has a considerable payoff in customer relationship management. Techniques for selection of relevant items according to the user's profiles in e-commerce are described by Ardissono and Goy (2000). In the following, a set of adaptive features that can be integrated in IS are identified and classified. Adaptation techniques can be classified in three categories: (1) adaptation of structure, (2) adaptation of content, and (3) adaptation of modality and presentation (Kobsa et al., 2001). These personalization techniques enable users to spend less time to search and retrieve relevant knowledge.

Adaptation of Structure

Adaptation of structure refers to the way in which the hypermedia space is structured and presented to the different groups of users. Fischer (2001) provides some insights in the design of human-centered systems supported by user modeling techniques. He emphasizes that high functionality applications must address three problems: (1) the unused functionality must not get in the way; (2) unknown existing functionality must be accessible or delivered at times when it is needed; and (3) commonly used functionality should not be too difficult to be learned, used, and remembered. Taking into account these principles, apart of a global view or a default view of an IS, several types of personalized views and layouts can be designed and integrated into the system. "Personalized views are a way to organize an electronic workplace for the users who need an access to a reasonably small part of a hyperspace for their everyday work" (Brusilovsky, 1998, p.).

Adaptation of Content

The users give different relevance to information/knowledge assets according to their goals, interests, background, or hobbies. Adaptation of content refers to the process of dynamic tailoring the information that is presented to the different users according to their specific profiles (e.g., needs, interests, level of expertise, etc). It enables the user to filter, retrieve, or rank relevant documents according to the user's characteristics. Adaptation of content relies on techniques for information filtering, information retrieval, information visualization, and adaptive hypermedia. Adaptation of content can include: filtering of content; personalized recommendations; personalized hints or automatic summarization; and optional detailed information. These techniques are further exemplified in the context of KMS.

Adaptation of Presentation and Modality

Adaptation of presentation empowers the users to choose between different presentation styles such as different layouts, skins, or fonts. Other preferences can include the presence or absence of anthropomorphic interface agents, the preferred languages, and so forth. Different types of sorting, bookmarks, or shortcuts can also be included in a highly functional system. Adaptation of presentation traditionally overlaps

with interface customization, but recently new modalities for information visualization have emerged.

Information portals such as Kartoo, Brain, and ClusterMap incorporate a graph-based view of the user interface. These new forms of information visualization go beyond a tree-based view of documents or knowledge assets using two or three dimension spatial representations. Documents can have associated visual cues to present information about documents, or they can be related with other documents, people, or institutions through relationships. These new forms of visualization have a higher expressive power of information visualization and an associated higher level of interactivity. Techniques for adaptive information visualization have been proposed by the Lighthouse system (Leusky & Allan, 2004).

Adaptation of modality enables changes from text to other types of media to present the information to the user (image, video, animations, or audio)—if they are available in the system. In modern adaptive hypermedia different types of media can present the same content. Adaptation of modality enables the automatic selection of the media type based on various criteria such as: cognitive style, learning style, preferences, physical abilities or disabilities, and so forth.

PERSONALIZATION TECHNIQUES APPLIED IN THE CONTEXT OF KMS

Adaptation of structure can be designed in the form of personalized views or layouts. These personalized views can be designed as domain-oriented subsystems of the KMS. Fischer (2001) argues that high functionality applications have often migrated to a collection of domain-oriented subsystems containing their own forms or templates. For instance the system can offer a personalized view of the corporate knowledge based on the interest areas and the knowledge of the users or based on the role and competencies of the users. For a KMS two main types of personalized views are proposed:

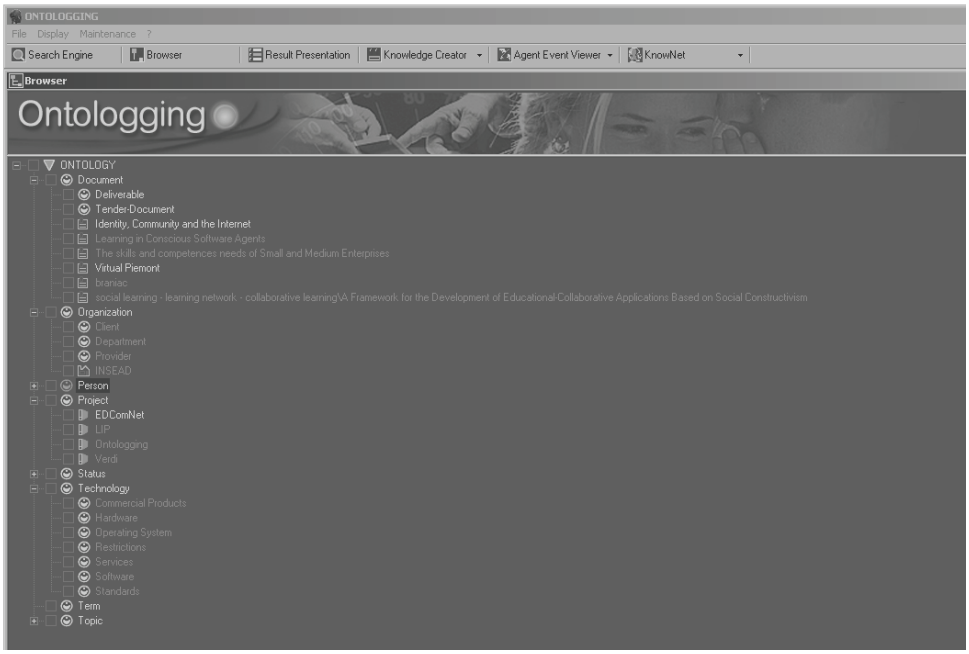
- **Personalized views based on the job title.** Based on the role of the users and the associated work tasks, different layouts can be designed. Taking into account the job titles and the most important associated work tasks, several stereotyped views can be designed. A set of richer features and complex functionality can be associated with certain jobs. Similar to a system administrator, a knowledge manager or a knowledge engineer has access to a richer functionality than a normal end user. Simplified views with limited functionality can be offered to other categories of users.
- **Personalized views based on the interest areas and the knowledge of the users.** This implies the possibility of selecting different views or subdomains of the domain ontology/taxonomy. In an ontology-based KMS the domain ontology or the corporate memory can be made up of different subontologies. For example, financial experts could have the possibility to select the finance ontology, ignoring other parts of the domain ontology. The distributed user interface of Ontologging system is designed using XML-driven components. The interface includes a "template mechanism" to customize a personalized view. The distributed user interface includes two main layers: (1) the user interface layer and (2) the component layer. The user interface layer consists of: search interface; browse interface; property edition and upload; result presentation interface; document download; and agent management interface. Thus through this template mechanism the different work units can personalize their view of the organizational memory or select parts of it, which are related to their expertise, interests, and so forth. Figure 2 represents the "tendering process" ontology of the Ontologging[1] system.

In a more sophisticated approach, the system is in charge with the selection of relevant knowledge assets for the user (the most suited items). The system evaluates how closed the knowledge assets match to the user's background and interests, and it presents the best ones. In an ontology-based KMS the system selects the concepts which match with the user's background and interests. A subontology of the domain ontology is further displayed to the user.

In a KMS, recommender systems, information filtering agents, and collaborative filtering techniques can be employed for adaptation of content. More appropriate for the use within KMSs are the following techniques:

- **Filtering of content.** Techniques for filtering the content help the user to select and retrieve information. Different systems integrate various filtering mechanisms. Among the well-known engines enabling filtering and automatic classification of content are: Verity ("Verity," 2003) and Autonomy (Autonomy, 2003). For example Autonomy integrates techniques such as: "active matching and content matching." Active matching enables users to enter the task and to extract a list of relevant documents for the task at hand. Content matching extracts conceptually related documents.
- **Personalized recommendations.** These recommendations inform users about available relevant information in the system. Recommendations can be provided via human (collaborative filtering) or artificial agents. In the category of recommender systems based on collaborative filtering Xerox research developed systems such as Knowledge Pump (KPump) and CWall (Snowdon &

Figure 2. A view of the domain ontology in the Ontologging system



Grasso, 2002). KPump allows users to submit recommendations of URLs, local files (via upload), or text. A recommendation consists of a rating and, preferably, a comment, along with the user's classification of the item into one or more communities. In turn, the KPump calculates a personalized set of recommendations for a user for each community to which he/she belongs. Communities are built based on domains of interest: "a community is a set of domains of interest plus the people in the organization with that set of interests."

- **Personalized hints for marking presumed interests.** Different types of agents or services can inform the users about new knowledge assets available in the system related to their interests and expertise (e.g., notification agents, etc.). Different relevant events can also be delivered to the users based on their interests and hobbies.

Let us have a look at some specific examples. All users interested in a certain domain (e.g., knowledge management) can be notified by the agents about various events related to their domain of interests. The domain of interest corresponds to the interest area in the user ontology. For instance, let us consider the following event: Mr. Popescu is in charge of a new project in the area of knowledge management. The notification agent system tells all users interested in knowledge management (via pop-up or e-mail depending on their preferences) that: "George Popescu from Competence Center team has started a new project in the area of Knowledge Management."

Another example: when Mr. Ionescu submits a new document into the system all users interested in knowledge management are notified that: "A new document related to knowledge management authored/submitted by Adrian Ionescu is available in the system."

- **Optional detailed information and automatic summarization.** Techniques of automatic summarization enable the automatic generation of summary from a certain piece of content. Autonomy is an example of a KMS embedding techniques of automatic summarization (Autonomy, 2003). This summary can be displayed by default to the users based on his/her preferences. Filtering certain parts of the documents could be provided based on semantic annotations associated with the document. For example, paragraphs of documents can be annotated in order to adapt the content of the displayed pages according to the user's interests and preferences. Techniques for tailoring Web pages according to the user's characteristics relate to research in the domain of adaptive hypermedia.

FUTURE TRENDS

Personalization is required by the end users of IS, and it is a step forward to overcome the "one size fits all" design of most of the actual IS. A key issue for integrating personalization is the access to the user's data, associated user model-

ing techniques, or collaborative filtering techniques. As the process of user modeling is limitative and sensitive, due to privacy issues, personalization techniques of successful systems such as Amazon or e-Bay rely on relatively simple collaborative filtering techniques.

Research on personalization of IS has opened many research issues. Future work will explore the possibility of reusing a user's profile for personalization and contextualization across different types of IS, new user modeling mechanisms combined with more complex personalization strategies, personalization of Web data extraction, and new collaborative approaches for personalization. The design of advanced IS and associated personalization techniques has recently converged with Semantic Web research. "Semantic Web enabled information systems will extract relevant information from the Web, process and combine different pieces of distributed information in such a way that the content selection and presentation fits to the individual needs of the user" (Baumgartner et al., 2005, p.)

CONCLUSION

Personalization techniques rely on the user's characteristics; captured and named user models; or user profiles and relate to specific objectives of various types of IS. In an e-commerce system personalization support for a potential client to retrieve relevant items corresponding to his/her goal, needs, and preferences are essential. KMSs aim to motivate people to create knowledge and submit new knowledge assets in the system; to stimulate collaboration and knowledge sharing between knowledge workers; to alleviate information overload; to simplify business processes and work tasks; and so forth.

The paper has proposed and exemplified a set of personalization techniques that can be applied for various types of IS. Personalized services for IS have been classified in: adaptation of structure; adaptation of content; and/or adaptation of presentation and modality. The use of personalization techniques has been exemplified in the context of KMS.

I argue that personalization has a utility function and a conviviality function (Razmerita, 2005). From the utility perspective, personalization helps fitting the functionality of the system to the user's needs, and personalization reduces the information overflow by providing users with the most relevant information.

From the conviviality perspective, personalization helps to bridge the gap between the designer's view of the system and the end user's view of the system and to take into account the user's preferences.

However, personalization techniques rely on the access of user's data (such as preferences, goal) available in the system, or on associated user modeling techniques that enable inferring dynamically the user's characteristics. The evaluation of Ontologging project has emphasized that even end users perceive personalization as an important feature to be integrated in IS; they are concerned with privacy issues and users want to be in control of their user profiles.

REFERENCES

Autonomy. (2003). *Autonomy white paper.* Retrieved September 27, 2003, from http://www.autonomy.com

Alpert, S., R., Karat, J., Karat, C.-M., Brodie, C., & Vergo, J. G. (2003). User attitude regarding a user-adaptive e-commerce Web site. *User Modeling and User-Adapted Interaction* 13, 373-396.

Ardissono, L., & Goy, A., (2000). Tailoring the interaction with users in Web stores. *User odeling and ser-dapted nteraction, 10*(4), 251-303.

Ardissono, L., Goy, A., Petrone, G., & Segnan, M. (2003). A multi-agent infrastructure for developing personalized Web-based systems. *ACM Transactions on Internet Technology.*

Baumgartner, R., Enzi, C., Henze, N., Herrlich, M., Herzog, M., Kriesell, M., et al. (2005). Semantic Web enabled information systems: Personalized views on Web data. *International Ubiquitous Web Systems and Intelligence Workshop (UWSI 2005),* Co-located with ICCSA 2005, Suntec Singapore.

Blom, J. (2000). Personalization—A taxonomy. *CHI'00 Conference on Human Factors in Computing Systems,* The Hague, Netherlands

Brna, P., Cooper, B., & Razmerita, L. (2001). Marching to the wrong distant drum: Pedagogic agents, emotion and student modeling. *Proceedings of the Workshop on Attitude, Personality and Emotions in User-Adapted Interaction in conjunction with User Modeling 2001,* Sonthofen, Germany.

Brusilovsky, P. (1998). Adaptive educational systems on the World-Wide-Web: A review of available technologies. *In 4th International Conference in Intelligent Tutoring Systems,* San Antonio, TX.

Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction,* 87-110.

Clark, R. C., & Mayer, R. E. (2003). *e-Learning and the science of instruction: Proven guidelines for consumers and designers of multimedia learning.* Wiley & Sons.

Fink, J., & Kobsa, A. (2000). A review and analysis of commercial user modeling servers for personalization on the World Wide Web [Special issue]. *User Modeling and User Adapted Interaction, 10,* 204-209.

Fischer, G. (2001). User modeling in human-computer interaction. *User odeling and User Adaptive Interaction*, 69-85.

Greer, J., McCalla, G., Vassileva, J., Deters, R., Bull, S., & Kettel, L. (2001). Lessons learned in deploying a multi-agent learning support system: The I-Help experience. *Proceedings of AIED'2001*, San Antonio, 410-421.

Kay, J. (2001, July). Scrutability for personalised interfaces. [Special Issue]. *ERCIM NEWS, 46*, 49-50.

Kobsa, A., Koenemann, J., & Pohl, W. (2001). Personalized hypermedia presentation techniques for improving online customer relationships. *The Knowledge Engineering Review, 16*, 111-155.

Kurki, T., Jokela, S., Sulonen, R., & Turpeinen, M. (1999). Agents in delivering personalized content based on semantic metadata. In *Proceedings 1999 AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace* (pp. 84-93). Stanford, USA.

Leusky, A., & Allan, J. (2004). Interactive information retrieval using clustering and spatial proximity. In *User Modeling and User Adapted Interaction, 14.*

Maes, P. (1998). Modeling adaptive autonomous agents. *Artificial Life Journal, 1*(1&2), 135-162.

Razmerita, L., (2003). *User model and user modeling in knowledge management systems: An ontology-based approach.* Unpublished doctoral thesis, University of Toulouse, France.

Razmerita, L. (2005). User modeling and personalization of the knowledge management systems. In S. Chen & G. Magoulas (Eds.), *Adaptable and adaptive hypermedia*. Hershey, PA: Idea Group.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*.

Snowdon, D., & Grasso, A. (2002). Diffusing information in organizational settings: Learning from experience. *Conf. on Human Factors and Computing Systems*, Minnesota (pp. 331-338).

Tanudjaja, F., & Mui, L. (2002). Persona: A contextualized and personalized Web search. In *Proceedings of the 35th Hawaii International Conference on System Science, (HICSS).*

Verity. (2003). Retrieved October 10, 2003, from http://www.verity.com/

KEY TERMS

Adaptive Hypermedia: Adaptive hypermedia is the dynamic generation of hypermedia spaces tailored to the characteristics and preferences of the different users.

Knowledge Management System (KMS): KMS are IS dedicated to manage organizational knowledge.

Information System (IS): IS is defined as "a system consisting of the network of all communication channels used within an organization" (Wordnet).

Ontology: Ontology is a specification mechanism used for knowledge representation based on a shared conceptualization.

Personalization: The personalization of IS is the process that enables interface customization, adaptations of functionality, structure, content, and modality in order to increase its relevance for its individual users.

User Model: A user model is an explicit representation of the system of a particular user's characteristics that may be relevant for personalized interaction.

ENDNOTE

[1] http://www.ontologging.com/

User Profile Modeling and Learning

Evangelia Nidelkou
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Vasileios Papastathis
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Maria Papadogiorgaki
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Ben Bratu
Motorola Ltd, France

Myriam Ribiere
Motorola Ltd, France

Simon Waddington
Motorola Ltd, UK

INTRODUCTION

A major theme of Information Science and Technology research is the study of personalization. The key issue of personalization is the problem of understanding human behaviour and its simulation by machines, in a sense that machines can treat users as individuals with respect to their distinct personalities, preferences, goals and so forth. The general fields of research in personalization are user modeling and adaptive systems, which can be traced back to the late 70s, with the use of models of agents by Perrault, Allen, and Cohen (1978) and the introduction of stereotypes by Rich (1979). With the wide progress in hardware and telecommunications technologies that has led to a vast increase in the services, volume and multimodality (text and multimedia) of content, in the last decade, the need for personalization systems is critical, in order to enable both consumers to manage the volume and complexity of available information and vendors to be competitive in the market.

BACKGROUND

The goal of personalization is to endow software systems with the capability to change (adapt) aspects of their functionality, appearance or both at runtime to the particularities of users to better suit their needs. The recent rapid advances in storage and communication technologies stress the need for personalization. This need is more evident in consumer-oriented fields, like news content personalization systems, recommendation systems, user interfaces, and applications like home audiovisual material collection and organization, search engines in multimedia browsing and retrieval systems, providing services for personalized presentation of interactive video content. Among these applications, some are Web-based, but there are also versions for PDAs and mobile devices (Tuoriniemi & Parkkinen, 2007) and mobile devices.

In this article, current approaches of user modeling and user profile representation are discussed, and then the focus is on methods for automatic learning of user models and profiles. The presented learning approaches cover a wide range of machine learning (vector-based or probabilistic) methods and also extend to support the most recent advances in personalization systems such as collaborative filtering, ontology-based user modeling and user social context.

OVERVIEW OF LEARNING AND ADAPTATION METHODS IN PERSONALIZATION SYSTEMS

User Modeling–User Profile Representation

User modeling describes the process of creating a set of system assumptions about all aspects of the user, which are relevant to the adaptation of the current user interactions. This can include user goals, interests, level of expertise, abilities and preferences. The most reliable method of user modeling is by explicit entry of information by the user. In most practical systems, this is too time-consuming and complex for the user. Hence implicit user modeling, based on analysis of past and current user interactions, is critical. The user profile is a machine-processable description of the user model.

The information included in user profiles can be divided into a number of categories such as user demographic information, semantic interests, context and location information, and privacy and user interface preferences (Heckmann & Krueger, 2003). Semantic preferences reflect user preferences for particular content topics. User interests and semantic user preferences are the most important source of information widely used in the personalization systems. More specifically, user interests are distinguished between *short-term* that are determined by a particular user interaction or current context, and *long-term* interests which are determined by the user behaviour and preferences over a longer period of time. User interests can also be classified into *gradual* (as a result of user experience), *abrupt* (as a result of an external stimulus) or *repetitive*. Loeb (1992) mentions two types of repetitive changes, *repetitive but predictable* (according to time of day) and *repetitive but unpredictable* (according to user mood).

There is a variety of structures and paradigms that have been used in the academic literature and in commercial personalization systems for the representation of the knowledge and information concerning the user, including the ones listed below. *Attribute-value pairs* are a fundamental data representation in many computing systems and applications. The advantage of such a structure is that it is an open-ended data structure, thus allowing for future extension without any need for modification. In such situations, all or part of the data model may be expressed as a collection of tuples (attribute name, value), where each element is an attribute-value pair. Several attempts have been put forward to standardize this type of user information structure, such as the IEEE Personal and Private Information (PAPI) (PAPI, 2002) and IMS Learner Information Package (LIP) (IMS, 2001).

The *vector space model* (VSM) is an algebraic model used for information filtering, information retrieval, indexing and relevancy rankings. It resembles the attribute-value pairs, but it has a more mathematical structure, in the sense that each element (term, or generally attribute) has a corresponding value or weight representing it and the vector has length and direction, both used, for example, in a similarity metric. The space of all vectors is often called vector domain or domain model. It has been extensively used in documents retrieval and indexing (Salton, Wong, & Yang, 1975). This representation approach has also been followed in a variety of personalization systems (Billsus & Pazzani, 2000; Lawrence, Almasi, Kotlyar, Viveros, & Duri, 2001; Ricci, Arslan, Mirzadeh, & Venturini, 2002).

One of the earlier representation approaches in user modeling has been the use of *stereotypes*. Stereotyping consists of creating a set of prototypical user profiles that represent the features of classes of similar users (Rich, 1979). Instead of keeping an individual model for each user, users are classified into the stereotypical description that best matches their individual characteristics, from which they inherit additional properties and rules.

The need to automatically learn user profiles has given rise to the use of more complicated representation methods such as the *classifier-based models*. These are based on decision trees, neural networks, inducted rules and Bayesian networks. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance and each branch corresponds to one of the possible values for this attribute (Cho, Kim, & Kim, 2002). In contrast to the limited decision trees representation range, artificial neural networks can represent real-valued, discrete-valued and vector-valued functions. The classifier-based models often take as input the *usage history and ratings*. The usage history is a log of the user transaction or interaction with the personalization system, which can be seen as a form of implicit user profile. It is a very practical model used in learning and adapting the user profile (Kang, Lim, & Kim, 2005).

Finally, the recent emerge of the Semantic Web technologies has led to *ontology-based representation* in user profiling. The Semantic Web vision of a next generation Web provides the mechanisms to identify those resources that better satisfy the requests not only on the basis of descriptive keywords but also on the basis of knowledge. The most common ways of representing semantic user profiles are the ontology-based and description logic based representations (Baldoni, Baroglio, & Henze, 2005). In recent work, semantic Web languages, such as Resource Description Framework (RDF), Ontology Web Language (OWL) are used to represent users and their semantic preferences. Gauch, Chaffee and Pretschner (2003) exploit hierarchical structures in ontologies to imply generalizations of user preferences upward in topic hierarchies (e.g., interest in football implies interest in sports).

Automatic Acquisition and Adaptation of User Models–User Preferences

The different representation approaches lead to a variety of methods used for the automatic acquisition and adaptation of user models, which are being presented in this section. Acquisition and adaptation models of user interests is a research area of steadily increasing importance, as it allows intelligent computer systems to adapt to users' information needs in a personalized way. Several machine learning approaches exist to build user profiles, such as Bayesian classifier, nearest neighbor, decision trees, neural networks and genetic algorithms (Pohl & Nick, 1999).

The classifier-based models are related to the use of *Classifier-Based (Statistical) Learning Methods*. The method most widely used in user profile learning is the Bayesian learning, which provides a probabilistic approach to inference. It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data. Bayesian learning methods are among the most practical approaches. The Naïve Bayes algorithm is the simplest form of probabilistic model for learning and classifying. It can easily be estimated by training data and in some cases it outperforms other learning methods (Billsus & Pazzani, 2000, pp. 147-180).

Decision tree learning is a method for approximating discrete-values target functions, in which the learned function is represented by a decision tree (Cho, Kim, & Kim, 2002).

In the case of the stereotypical user models, there are different methods of *Learning Stereotypical Sequences of User Interactions* depending on the purposes of the recommender systems: *Collaborative recommendations systems* are based on demographic, geographic and semantic information and they have manually predefined user-stereotypes. Other systems first build a data base of user profiles using the usage log (Azman & Ounis, 2004). In *marketing-based recommendations systems*, the demographic and usage log data can be used to discover rules that capture the truly personal behaviour of a user by means of data mining algorithms. Grouping similar profiles and creating a cluster representative set of rules firstly avoids the privacy problems and secondly reduces the computation during the recommendation process (Wei, Moreau, & Jennings, 2005). Finally, the *social matching recommender systems* match people to each other instead of recommending items to people. The system first creates different set of similar users and then builds a model (i.e., stereotype profile) for each set of users (Terveen & McDonald, 2005).

The user profiles represented with vector space models are related to learning methods based on the user's *relevance feedback*. The term feedback is normally used to describe the mechanism by which a system can improve its performance on a task by taking account of past performance. Adaptive systems using relevance feedback have to choose how relevance feedback should be represented, acquired and used. There are three different methods for representing the relevance feedback. *Boolean relevance* describes whether a document is relevant or not relevant (Objective feedback) vs. useful or not useful (Subjective feedback) to the user. *Multi valued relevance* has been proposed by Bookstein (1983) where the possible relevance classes might be: Very Relevant, Relevant, Indifferent, Irrelevant, Very Irrelevant. In

Quasi-Ordered Relevance, the Document Preference Relation (Wong & Yao, 1990) method relies on a quasi-order of documents. For each pair of documents, the user can either prefer one to the other or have no opinion.

Once the relevance feedback is acquired, there exist multiple formulas that propose to reweight the terms used in the initial query (query reformulation). In vector processing methods, the most commonly used is the Rocchio formula presented in Salton and Buckley (1990), where the new query vector is the vector sum of the old query vector plus the vectors of the relevant and non relevant documents. An extension of this formula proposed by Salton and Buckley (1990), called Ide, eliminates the normalization of the number of relevant and nonrelevant documents and allows limited negative feedback from only the top-ranked nonrelevant document.

In the case of the ontology-based user models, the learning process needs to exploit the deeper ontological knowledge about the underlying domain, thus allowing the personalization systems to handle heterogeneous and complex objects based on their properties and relationships and to automatically explain or reason about the user models or user recommendations. During the learning process, the concepts represented in the ontology are rated according to user-specified preferences such as semantic relevance, syntactic relevance, and categorical match (Kerschberg, Kim, & Scime, 2001).

The learning methods in the ontology-based user models are often enhanced with one of the most expressive and human readable representations for learned hypotheses, which is to use sets of if-then rules. One important special case involves learning sets of rules containing variables. First order rules are more expressive than propositional rules. In general, in many cases it is useful to learn the target function represented as a set of if-then rules that jointly define the function. One way to learn sets of rules is to first learn a decision tree, then translate the decision tree into an equivalent set of rules; one rule for each leaf node in the tree (Mobasher, Dai, Luo, & Nakagawa, 2001). Another method is to use a genetic algorithm that encodes each rule set as a bit string and uses genetic search operators to explore this hypothesis space. There are also a variety of algorithms that directly learn rule sets (Mitchell, 1997).

In all the above-mentioned learning methods, content filtering agents attempt to alleviate information overload by identifying which items a user will find worthwhile. Content filtering focuses on the analysis of item content and the development of a personal user interest profile.

However, to overcome the problem of handling end users solely as units and missing possible information and trends beyond those within the scope of user's history, *collaborative filtering* learning methods are introduced. Collaborative approaches find and recommend information sources for an individual user that have been rated highly by other users who have a pattern of ratings similar to that of the user (Pazzani, 1999). Each technique has advantages and limitations. Current methods for collaborative filtering can be divided into two categories: memory-based, which use all of the available data when making recommendations, and model-based methods which, at some point, learn a statistical model from data and use that model for predicting user interests. Collaborative filtering has a number of advantages over Content-based filtering methods. The quality of memory-based collaborative filtering algorithms typically increases with the size of the user population, and Collaborative filtering recommendations benefit from improved diversity when compared to Content-based filtering recommendations (Claypool et al., 1999). For a start, memory based algorithms are not suitable for recommending new items or one-off content items because these techniques can only recommend items already rated by other users. On the other hand, the model based methods can use smoothing methods to give prior probabilities to items without any ratings. Collaborative filtering matches people with similar interests and then make recommendations on this basis. Collaborative filtering is based on a statistical analysis of patterns and analogies of ratings obtained explicitly or implicitly from user system usage. Typically, for each user a set of nearest neighbors is defined using the correlation between past ratings. Collaborative filtering techniques can be classified to two categories according to the source of information; the user-based and item-based collaborative filtering. The main deficiency of user-based collaborative filtering systems is that they usually make recommendations from very thinly scattered data. In item-based filtering, there are techniques for computing item-item similarities and for obtaining recommendations from them. Linden, Smith, and York (2003, p. 76) follow this approach in Amazon's recommendation system. The main advantage of the item-based approach over the user-based one is its scalability. A combination of the two approaches can be seen in the work of Renda and Straccia (2005).

A further extension of the collaborative filtering approaches is to exploit the *social user context*, which is mainly composed of the user's relationships with other users. Social Information filtering exploits similarities between the tastes of different users to recommend (or advise against) items. It relies on the fact that people's tastes are not randomly distributed: there are general trends and patterns within the taste of a person as well as between groups of people. The basic idea is that the system maintains a user profile, a record of the user's interests (positive or negative) in specific items. It compares this profile to the profiles of other users, and weighs each profile for its degree of similarity with the user's profile. Mika (2005) presents the advances in exploiting the opportunity of semantically-enriched network data.

PERSONALIZATION IN COMMERCIAL APPLICATIONS

Personalization, besides its value in the research field, has also been deemed as an important part of many commercial applications due to the innovation in the services it provides. An example of automatic personalization in commercial systems is Amazon.com's personalized recommendations (Linden, Smith, & York, 2003). Google Inc. has also filed two U.S. patents on personalization technologies for Web search (Badros & Lawrence, 2005; Zamir, Korn, Fikes, & Lawrence, 2005). The Leiki concept aims at combining personalized user interfaces, communities and content targeting (Pennanen & Alatalo, 2001). The Leiki platform is applied as a personalized news service. MovieLens, http://movielens.umn.edu/login, is a free Web-based movie recommendation service provided by the GroupLens research team from the University of Minnesota. It works by matching together users with similar opinions about movies using a collaborative filtering algorithm. TiVo is a television show collaborative recommendation system (Ali & van Stam, 2004, pp. 394-401). The success and innovation of TiVo relies in their personalised television-viewing service, which recommends or automatically records programmes based on user preferences.

FUTURE TRENDS

Mobile ad-hoc networks, wireless broadcasting and open mobile applications are three prominent examples in which computation and communication intermingle with the real world changing the role of context information. Context includes user activities, goals, abilities, preferences, and surroundings.

Current personalization systems do not fully support such flexible and self-adapting models based on context. Thus, future research opportunities within the field of automatic personalization systems include the study of context-aware systems as well as seamless mobility, which is the key future trend in distributed mobile environments. These areas involve research in privacy and sharing of context informa-

tion and also in the synchronization of user profile between different devices.

CONCLUSION

In this article, the state of the art on the user modeling and user profile representation was presented. More specifically, the standardization and categorization of user profiles was introduced, along with the information included in user profiles and user profile structures. Then, the emphasis is given to the automatic learning of user profiles, where different approaches are being discussed. Automatic learning of user profiles is the current trend in the academic literature and also the key requirement of the current and future commercial personalization systems.

REFERENCES

Ali, K., & van Stam, W. (2004). TiVo: Making show recommendations using a distributed collaborative filtering architecture. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Seattle, WA, USA, (pp. 394-401).

Azman, A., & Ounis, I. (2004). Discovery of aggregate usage profiles based on clustering information needs. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '04*. New York, (pp. 470-471). ACM Press.

Badros, G. J., & Lawrence, S. R. (2005, June). *Methods and systems for personalized network searching* (U.S. Patent Application 20050131866).

Baldoni, M., Baroglio, C., & Henze, N. (2005). Personalization for the Semantic Web. *Reasoning Web. LNCS tutorial* (Vol. 3564, pp. 173-212). Springer-Verlag.

Billsus, D., & Pazzani, M. (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction, 10*(2/3), 147-180.

Bookstein, A. (1983). Outline of a general probabilistic retrieval model. *Journal of Documentation, 39*(2), 63-72.

Cho, Y.H., Kim, J.K., & Kim, S.H. (2002). A personalized recommender system based on Web usage mining and decision tree induction. *Expert Systems with Applications, 23*(3), 329-342.

Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., & Sartin, M. (1999). Combining content-based and collaborative filters in an online newspaper. In *Proceedings of the SIGIR '99 Workshop on Recommender Systems: Algorithms and Evaluation*, Berkeley, CA.

Gauch, S., Chaffee, J., & Pretschner, A. (2003). Ontology-based personalised search and browsing. *Web Intelligence and Agent System*, *1*, 219-234.

Heckmann, D., & Krüger, A. (2003). A user modeling markup language (UserML) for ubiquitous computing. In *Proceedings of the 9th International Conference on User Modeling (UM'2003)*. Johnstown, PA, USA, (Vol. 2702, pp. 393-397). Springer-Verlag.

IMS Global Learning Consortium, Inc. (2001). *IMS learner information packaging model specification*. Retrieved December 13, 2007, from http://www.imsglobal.org/profiles/lipinfo01.html

Kang, S., Lim, J., & Kim, M. (2005). Statistical inference method of user preference on broadcasting content. *LNCS*. (Vol. 3514, pp. 971-978).

Kerschberg, L., Kim, W., & Scime, A. (2001). A semantic taxonomy-based personalizable meta-search agent. In *Proceedings of the Second International Conference on Web Information Systems Engineering* (WISE'01), (Vol. 1, pp. 41).

Lawrence, R.D., Almasi, G.S., Kotlyar, V., Viveros, M.S., & Duri, S.S. (2001). Personalization of supermarket product recommendations. *Data Mining and Knowledge Discovery, 5*(1-2), 11-32.

Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing, 7*(1), 76-80.

Loeb, S. (1992). Architecting personalized delivery of multimedia information. *Communications of the ACM, 35*(12), 39-48.

Mika, P. (2005). Social networks and the Semantic Web: The next challenge. *IEEE Intelligent Systems, 20*(1), 80-93.

Mitchell, T.M. (1997). *Machine learning*. McGraw-Hill.

Mobasher, B., Dai, H., Luo, T., & Nakagawa, M. (2001). Effective personalization based on association rule discovery from Web usage data. In *Proceedings of the 3rd ACM Workshop on Web Information and Data Management (WIDM01)*, Atlanta, Georgia.

PAPI, IEEE Computer Society. (2002). *Public and private information for learners (PAPI Learners)*. Retrieved December 13, 2007, from http://edutool.com/papi/

Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review, 13*(5/6), 393-408.

Pennanen, P., & Alatalo, T. (2001). *Leiki—a platform for personalized content targeting*. A demo presentation at the ACM Hypertext 2001.

Perrault C. R., Allen, J. F., & Cohen, P. R. (1978). Speech acts as a basis for understanding dialogue coherence. In *Proceedings of the Theoretical Issues in Natural Language Processing-2,* Urbana-Campaign, IL, USA, (pp. 125-132).

Pohl, W., & Nick, A. (1999). Machine learning and knowledge representation in the LaboUr approach to user modeling. In *Proceedings of the 7th International Conference on User Modeling*, Banff, Canada, (pp. 179-188).

Renda, M.E., & Straccia, U. (2005). A personalized collaborative digital library environment: A model and an application. *Information Processing and Management: An International Journal*, *41*(1), 5-21.

Ricci, F., Arslan, B., Mirzadeh, N., & Venturini, A. (2002). ITR: A case-based travel advisory system. In S. Craw & A. Preece (Eds.), *6th European Conference on Case Based Reasoning, ECCBR 2002,* (pp. 613-627). Berlin: Springer-Verlag.

Rich, E. (1979). User modeling via stereotypes. *Cognitive Science*, *3*, 329-354.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, *41*(4), 288-297.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620.

Terveen, L., & McDonald, D. W. (2005). Social matching: A framework and research agenda. *ACM Transactions on Computer-Human Interaction*, *12*(3), 401-434.

Tuoriniemi, S., & Parkkinen, J. (2007, March). *Voucher driven on-device content personalization* (U.S. Patent Application 7191343B2).

Wei, Y. Z., Moreau, L., & Jennings, N. R. (2005). A market-based approach to recommender systems. *ACM Transactions on Information Systems*, *23*(3), 227-266.

Wong, S. K. M., & Yao, Y. Y. (1990). Query formulation in linear retrieval models. *Journal of the American Society for Information Retrieval*, *41*(5), 334-341.

Zamir, O. E., Korn, J. L., Fikes, A. B., & Lawrence, S. R. (2005, October). Personalization of placed content ordering in search results (U.S. Patent Application 20050240580).

KEY TERMS

Acquisition and Adaptation of User Profiles: The automatic creation and adaptation of a user profile by monitoring the user interaction with the system and employing efficient machine learning algorithms.

Collaborative Filtering: The method of making automatic predictions (filtering) about the interests of a user by collecting information from other similar users.

Content-Based Filtering: The method of making recommendations to a user by matching user profile entries to content attributes.

Machine Learning: The method for processing a training input and offering support for decision based on this input.

Personalization: Delivery of content according to the individual user's needs, characteristics and preferences.

User Modeling: The process of creating a set of system assumptions about all aspects of the user, which are relevant to the adaptation of the current user interactions.

User Profile: A machine-processable description of the user model.

Using an Architecture Approach to Manage Business Processes

Shuk Ying Ho
The University of Melbourne, Australia

INTRODUCTION

Business process management has long been a topic of great interest in operations management research. Early research on business process management focuses on workflow analysis and process optimization. These types of research evaluate and analyze a predefined set of procedures from a process perspective. That said, with a list of activities, constraints and criteria, the procedural workflow are specified and examined. Then, process analysts come up with suggestions to optimize the process and speed up the workflow. Research findings are widely applied in production and logistics; however, some works are criticized as being too rigid and only suitable for a stable business environment (Burns, 1993).

With the recent advances of information technologies (IT), research topics, including enterprise resource planning (Hong & Kim, 2002), computer integrated manufacturing (Burns, 1993) and total integrated management (Azhashemi & Ho, 1996), have become popular. Technologies are now playing a more significant role in organizations than before. They help organizations achieve higher competitive advantages, and facilitate the operations in all functional areas, such as marketing and sales, cash receipts, purchasing, cash disbursement, production and logistics and human resources (Valiris & Glykas, 2004). Transactions and financial data are gathered and stored, and technologies make data available for operational units and management to make decisions. Operations are sped up and an enduring dialogue in the intra- and inter-organizational contexts is built.

Technologies, on one hand, create substantial values to organizations. On the other hand, technologies are moving too fast, resulting in a rapidly-changing business environment. With the rapidly-changing environment, organizations face challenges. From a technological point of view, new technologies emerge and organizations conduct business in a more dynamic environment (Neiderman, Brancheau, & Wetherbe, 1991). For instance, although the Internet provides rich opportunities for organizations to exchange information, it greatly reduces the switching cost of users. This results in fierce competitions among organizations. Organizations have to reduce their costs in the hope of remaining competitive. Moreover, communities, such as W3C and IEEE, constantly propose new technology standards. Among various standards, Web services are the most dominant. Sets of Web services standards, such as extensible markup language, were put forward in the early 2000s. An organization can use these standards to easily integrate multiple systems across platforms. The standards also allow various organizations to share data and applications (Coetzee & Eloff, 2007). Much research (e.g., Moitra & Ganesh, 2005) explores how Web services increase the flexibility of business processes. At present, many technology leaders, such as Microsoft, IBM, Google and Amazon, have already adopted Web services. With high external pressure, other organizations are likely to follow the technology trend and quickly adopt Web services to maintain their competitive advantage.

New technologies not only shape the operation and business environments, but also influence government regulations that pose new requirements and constraints on business processes. With transactions migrating to a computer platform, IT frauds become a concern for process analysts, management and auditors. Thus, the federal government proposes new regulations. For instance, Section 404 of the Sarbanes-Oxley Act (SOA) expands the significance of internal controls of processes. It explicitly states that managers and auditors are responsible for enacting and enforcing proper internal controls throughout their organizations. Technologies become a means to achieve improved quality of operational controls, and the ultimate objective of using technologies in business processes is to achieve high effectiveness, high efficiency and high security of organizations.

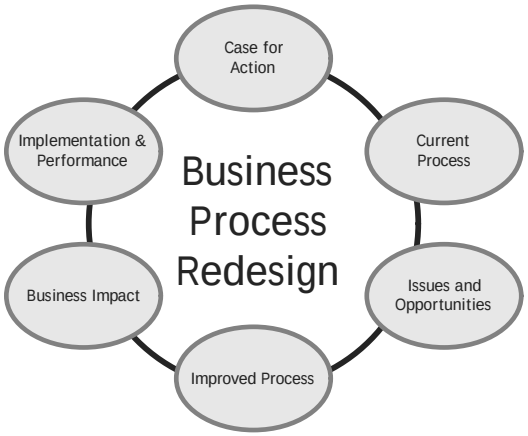
The article describes an architecture approach for business process management, and is organized as follows: first, we review the literature on architecture. Next, we outline a de facto standard for the architecture approach, and highlight the strength of using an architecture approach. Finally, we describe future trends, and conclude the article.

BACKGROUND

Business Processes and Their Importance

A business process is defined as a set of related, structured activities, or a chain of events, that produces a specific out-

Figure 1. General steps in business process redesign



come for a particular user or users. This set of activities is in pursuit of a common goal. Typical business goals include receiving sales orders, marketing services, selling products, delivering services, distributing products, invoicing for services, accounts receivable, purchasing and accounts payable. Business processes directly support organization strategies, and hence the processes are valuable corporate assets.

To tackle the challenges by the rapid changes in technologies, presumably, organizations are required to evaluate business processes on a regular basis. As a result, organizations are paying more attention to supporting business process management and redesigning the processes to adapt to the new environments (Davenport & Stoddard, 1994). Figure 1 depicts the general steps of business process redesign.

During business process redesign, organizations have a specific case for examination. They identify the processes of interests and evaluate the opportunities and constraints. They improve the processes by integrating them into new technology platforms, and processes are redesigned in the hope of cutting costs, improving controls, speeding up the operations and achieving higher efficiency in the interactions between the internal processes and interactions of the processes with the environment (Arlbjorn, Wong, & Seerup, 2006; Hofmann & Reiner, 2006). Process redesign can help to achieve strategic advantages. Thus, it is not surprising to see the increasing amount of resources that organizations invest in redesigning their business process (Tinnilä, 1995).

Prior research shows that most organizations are not familiar with business process redesign and change management. Also, they do not have a structured methodology for process redesign. However, even if a few organizations adopt a structured methodology, some researchers criticize that their methodology is prescientific and ad hoc (Valiris & Glykas, 2004), and most approaches cannot provide a holistic view from management and operations (Avison & Wood-Harper, 1990). In the following, we will provide a high-level architecture approach, which is influenced by organizational theories, IS development and existing work in business process redesign. Architecture views are representations of the overall system that are meaningful to all stakeholders, such as management board, chief information officer, users, designers and analysts, in the enterprise. It also provides a holistic approach to help stakeholders gather relevant and sufficient information for business process redesign.

An Architecture Approach

The word, architecture, originates from Latin. It is the art and science of building structures. In recent years, the word, architecture, has been used in application coordination (King, 1995), software management (Greefhorst, Koning, & Vliet, 2006) and enterprise management (Johnson, Lagerstrom, Narman, & Simonsson, 2007). It is the abstraction used to deal with complexity. By extension, the term "architecture" has come to denote the art and discipline of creating an actual, or inferring an implied or apparent plan of any complex object or system. Architecture is analogous to a blueprint for the object or the system. It details how the design is to be divided into individual functional components and the way in which these components are to interact to provide the overall functionality. According to IEEE Standard 1471-2000, architecture is "the fundamental organization of a system embodied in its components, their relationships to each other, and to the environment, and the principle guiding its design and evolution."

In the context of an organization undergoing business process redesign, architecture is a blueprint which specifies data architecture, business architecture, technical architecture and process architecture. An architecture approach is usually taken by large enterprises. As an organization grows in complexity and size, several factors hinder its abilities to solve the problems that it faces. The point is rapidly reached where the factors that come into play in structuring and conducting the business of the enterprise become too numerous and complex to manage. When dealing with such complex systems, a complicated problem is usually broken into subsets or domains, which is less complex and more manageable. This helps to make sure the orchestration of the interaction among those subsets or domains. Thereby, an architecture approach is typically adopted by large enterprises. We refer to it as *enterprise architecture*. Enterprise architecture is a specific type of architecture and can be considered to be the description and documentation of the current and desired information and technology environment, and its relationship to processes and business strategy. It coordinates various facets, including process models, diagrams or textual documents (Puschmann & Alt, 2005).

ARCHITECTING AN ENTERPRISE

Major Issues when Architecting an Enterprise

Enterprise architecture is a means for describing processes that connect business structures. Its objective is to let the stakeholders of the organization have a clear vision of the desired future state of the entire system, including such dimensions as business strategies, process design and technology infrastructure. To achieve this vision, all stakeholders must have a common context both for diagnosing the needs for changes and for managing the process of changes. Hence, the common context acts as an integrating force for the multitude of apparently disparate changes to be made. It is where the concept, architecture, comes in.

There are four dimensions in enterprise architecture. They are data architecture, business architecture, technical architecture and process architecture. These dimensions provide a holistic view of an enterprise (see Figure 2). *Data architecture* defines how data is stored, managed and used in an enterprise. It establishes common guidelines for data operations and thus it is possible to predict, model, gauge, and control the flow of data in the enterprise. It provides the basis for business architecture, technical architecture and process architecture. *Business architecture* structures the accountability over business strategies and activities prior to any further effort to structure individual aspects. *Technical architecture* refers to hardware and software architecture. It is the structured process of designing and building technology hardware at enterprise level. Issues including the legacy systems, middleware and network communications are all within the scope of hardware architecture. Technical architecture also deals with software architecture, which focuses on functions of the software programs and data transfer between programs. *Process architecture* is the structural design of general process systems and covers process design, logistics, policy and procedures. The overall inputs, outputs and functionality of the process architecture are also specified.

The enterprise architecture approach facilitates the business process management by providing frameworks and methodologies to support: (1) a shared understanding of business functionality and data usage across the virtual enterprise; (2) a baseline from which applications can be deployed or integrated in a coordinated fashion; (3) a context for performing business-driven process and information modeling; (4) an understanding of how business models drive the creation of functional IT and data models; (5) the capability for planners and business owners to communicate and exchange requirements with IT; and (6) the ability for chief information officers to understand and respond to dynamic business impacts on the IT environment (Johnson et al., 2007).

When architecting the enterprise, stakeholders generally have concerns about cost, value, risks, time, control and so

Figure 2. An overview of enterprise architecture

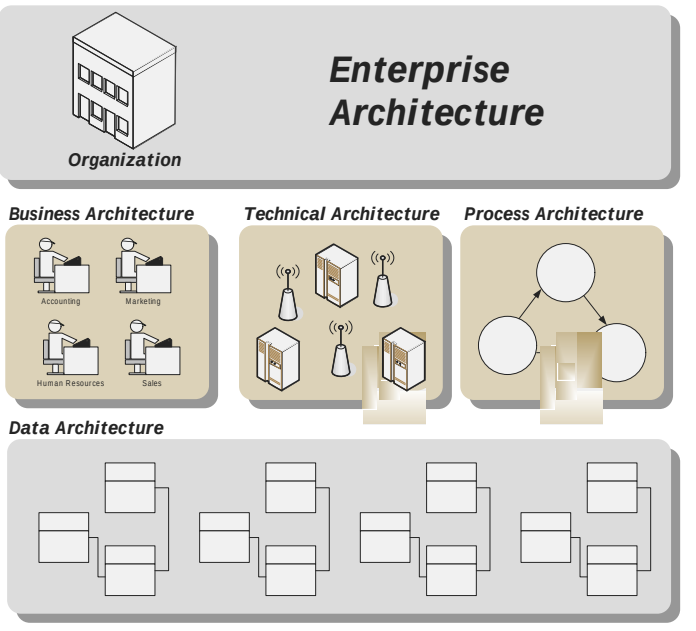


Table 1. Examples of typical questions raised by architects and top management

Process issues	*Performance* ● Is IT making the business more agile? ● Is significant improvement of information technology actually possible? ● Is full advantage being taken of the progression in hardware technology and software techniques? ● Is there an ongoing program for rationalizing systems? *Expertise and Resources* ● Can we identify the solutions specialists in our organization? *Risk and Security* ● What are the possible risks? What are their levels of damages? ● Are security and availability risks fully managed?
Strategy issues	*Cost and Return on Investment* ● Is the value of IT investments being measured? If so, by how? ● Are the results of these investments as good as possible? ● How can the finite resources be most effectively used? ● Is the asset value of IT understood and acceptable? *Impacts on Business Strategies* ● Is information technology focused on business outcomes? ● Is the business strategy being realized?
Policy issues	*Governance* ● Are our key development partners independent? ● How does the organization deal with increasing regulatory requirements?

forth. Table 1 shows a list of typical questions raised by the stakeholders.

Technologies Related to the Architecture Approach

Tools supporting architecture are commonly available. One of the most widely applied tools is Zachman Framework for Information Systems Architecture by John Zachman (Zachman, 1987). The Zachman Framework is an enterprise-class framework which provides a formal and highly structured way of defining an enterprise. It has a series of views describing how business planners and owners envision IT requirements, and how designers and developers view a logical or physical implementation of those same requirements.

The Zachman Framework is made up of two dimensions. The first dimension contains six aspects by using the basic communication interrogatives (What, How, Where, Who, When, and Why). The second dimension consists of six viewpoints by six groups of stakeholders (Planner, Owner, Designer, Builder, Subcontractor and Worker). Table 2 presents the six aspects and the six viewpoints.

Figure 3 shows a picture of the Zachman Framework. There are 36 (=6×6) cells in the framework model. The intersecting cells depict the related documentation modeling techniques or technology to address the questions by the stakeholders. The framework draws stakeholders' focus to relevant domains and stimulates them to ask relevant questions. It also organizes descriptive artifacts of an enterprise, such as models, pictures, diagrams, or textual documents, and classifies them into one of the 36 cells.

Apart from the Zachman Framework, other enterprise-class frameworks are available. They include the Information Framework (Evernden, 1996); the Open Group Architecture Framework (Open Group, 2006); and the Methodology for Architecture Description (Meinema, 1999). These are all de facto standards. Architects decide to use which framework based on criteria, such as their expertise and organizational practice.

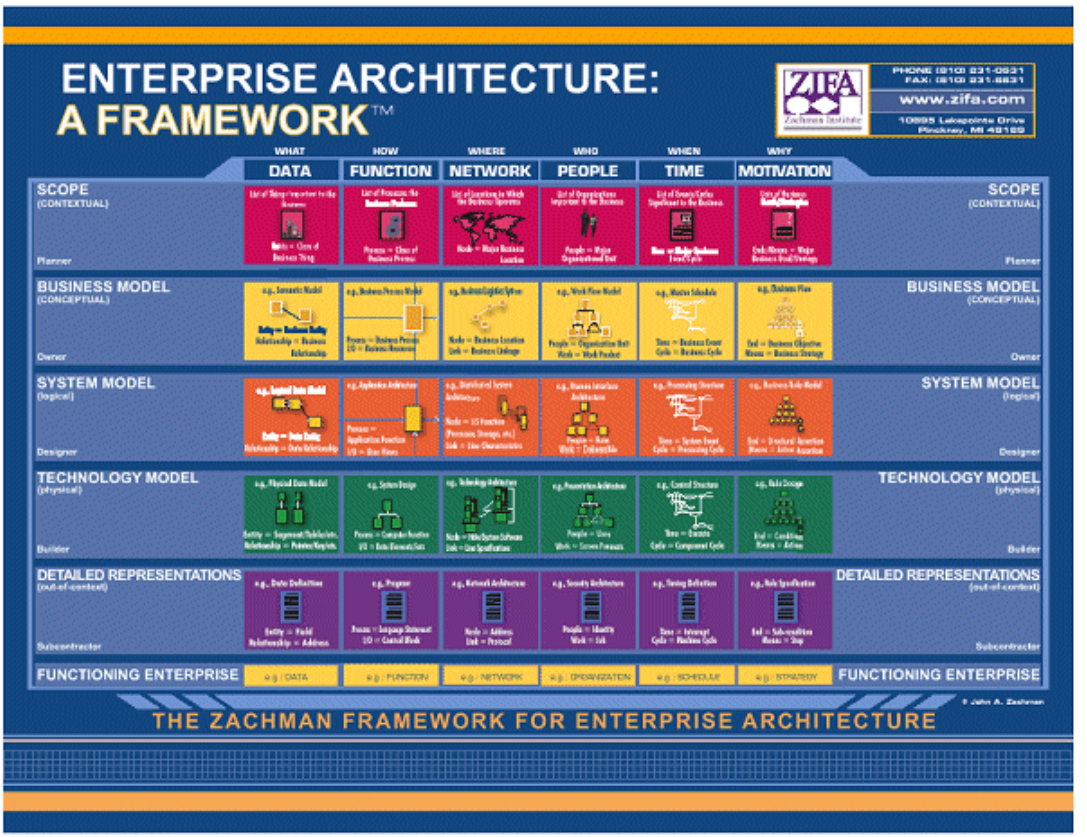
Reasons to Adopt an Architecture Approach

To deal with global competition, organizations seek ways to regularly reconsider their IT environments and optimize

Table 2. The six aspects and the six viewpoints of the Zachman Framework

The Six Aspects	1. The Data aspect - What? 2. The Function aspect - How? 3. The Network aspect - Where? 4. The People aspect - Who? 5. The Time aspect - When? 6. The Motivation aspect - Why?
The Six Viewpoints	7. The Scope (Contextual) viewpoint aimed at the planner 8. The Business Model (Conceptual) viewpoint aimed at the owner 9. The System (Logical) viewpoint aimed at the designer 10. The Technology (Physical) viewpoint aimed at the builder 11. The Detailed Representations (Out-of-Context) viewpoint aimed at the subcontractor 12. The Functioning Enterprise viewpoint aimed at the worker

Figure 3. The Zachman Framework (Reference: http://www.opengroup.org/architecture/togaf8-doc/arch/chap39.html)



their business processes. Organizations redesign their information systems and business processes, in the hope of reducing the cost and helping themselves respond to a changing environment rapidly.

Enterprises taking an architecture approach to redesign their process can achieve the following benefits: Firstly, enterprise architecture allows the business and IT to work together as harmoniously as the business and finance do. It

provides a set of tools, the common language for which is needed to allow the stakeholders to become partners with the same mission (Zachman, 1987). With an enterprise architecture approach, organizations describe and document the current and desired information and technology environment, and its relationship to business strategy and processes, thus providing a platform for better quality and quicker decision making.

Secondly, it provides a means for the business and IT to communicate effectively, ensuring that IT is supporting the business. Effective communication leads to a clear understanding of the current IT situation and how well it fits the business. It also facilitates IT planning so that all impacts on the business are clearly understood.

Thirdly, an architecture approach allows a business to be more flexible and agile, thus enabling a quick and cost effective response to changes in the environment. The approach also helps to manage the business risk relating to the use of IT, and reduce the risk of problems with technology and how it is applied.

FUTURE TRENDS

Managing business processes, in particular for an enterprise, is a challenging task because business processes interact with data, technology infrastructure and business strategies. While careful planning typically goes into its design, architecture actually emerges as a result of implementing business processes with the consideration of aspects from various stakeholders. It is a de facto standard and provides the capabilities for executing business strategies, and understanding this emergent architecture is of paramount importance.

Further research can be conducted to improve the architecture framework. One of the limitations of the framework is its complexity. Organizations have to invest time and resources to fill in each cell of the framework, and sometimes, not all information in all cells are important. Attempts by researchers and practitioners can be made to simplify the framework. They can prioritize the dimensions of the framework and customize the framework to different industrial domains. We believe the simplified version will be very useful to medium-sized firms.

CONCLUSION

The fierce competition of business leads to a strong need of improving and adapting business processes to the new environment. This chapter introduces an architecture approach of business improvement. We particularly focus on the enterprise architecture approach and its related technologies. We also provide explanations to address the importance of an architecture approach and how it helps to improve business process management. Currently, enterprise architecture is still considered to be a costly and highly sophisticated technology, which requires input from the management board, chief information officer, users, designers and analysts. With the high complexity of business activities and continuous awareness of situations, we believe an architecture approach will gain significant attention in the near future.

REFERENCES

Arlbjorn, J.S., Wong, C.Y., & Seerup, S. (2006). Achieving competitiveness through supply chain integration. *International Journal of Integrated Supply Management, 3*(1), 4-24.

Avison, D.E., & Wood-Harper, A.T. (1990). *Multiview: An exploration in information systems development*. Oxford: Blackwell Scientific.

Azhashemi, M.A., & Ho, S.K. (1996). Business process redesign and total integrated management. *The TQM Magazine, 8*(6), 42-47.

Burns, M. (1993). *Automated fabrication: Improving productivity in manufacturing*. Upper Saddle River, NJ: Prentice Hall.

Coetzee, M., & Eloff, J.H.P. (2007). Web services access control architecture incorporating trust. *Internet Research, 17*(3), 291-305.

Davenport, T.H., & Stoddard, D.B. (1994). Reengineering: Business change of mythic proportions? *MIS Quarterly, 18*(2), 121-127.

Evernden, R. (1996). The information framework. *IBM Systems Journal, 35*(1), 37-68.

Greefhorst, D., Koning, H., & Vliet, H.V. (2006). The many faces of architectural descriptions. *Information Systems Frontiers, 8*(2), 103-113.

Hofmann, P., & Reiner, G. (2006). Drivers for improving supply chain performance: An empirical study. *International Journal of Integrated Supply Management, 2*(3), 214-230.

Hong, K.K., & Kim, Y.G. (2002). The critical success factors for ERP implementation: An organizational fit perspective. *Information and Management, 40*(1), 25-40.

Johnson, P., Lagerström, R., Närman, P., & Simonsson, M. (2007). Enterprise architecture analysis with extended influence diagrams. *Information Systems Frontiers, 9*(2/3), 163-180.

King, W.R. (1995). Creating a strategic capabilities architecture. *Information Systems Management, 12*(1), 67-69.

Meinema, J.L. (1999). *Corporate architecture: A conceptual approach*. University of Twente.

Moitra, D., & Ganesh, J. (2005). *Web services and flexible business processes: Towards the adaptive enterprise. Information and Management, 42*(7), 921-933.

Neiderman, F., Brancheau, J.C., & Wetherbe, J.C. (1991). Information systems management issues for the 1990s. *MIS Quarterly, 15*(4), 475-502.

Open Group. (2006). *The open group architecture framework version 8.1.1, Enterprise Edition*. San Francisco.

Puschmann, T., & Alt, R. (2005). Developing an integration architecture for process portals. *European Journal of Information Systems, 14*(2), 121-134.

Tinnilä, M. (1995). Strategic perspective to business process redesign. *Business Process Management Journal, 1*(1), 44-59.

Valiris, G., & Glykas, M. (2004). Business analysis metrics for business process redesign. *Business Process Management Journal, 10*(4), 445-480.

Zachman, A.J. (1987). A framework for information systems architecture. *IBM Systems Journal, 26*(3), 276-292.

KEY TERMS

Blueprint: It is a document to detail how the design is to be divided into individual functional components and the way in which these components are to interact to provide the overall functionality.

Business Process: A business process is defined as a set of related, structured activities, or a chain of events, that produces a specific outcome for a particular user or users. This set of activities is in pursuit of a common goal.

Business Architecture: It structures the accountability over business strategies and activities prior to any further effort to structure individual aspects.

Enterprise Architecture: It is a specific type of architecture and can be considered to be the description and documentation of the current and desired information and technology environment, and its relationship to processes and business strategy.

Data architecture: It defines how data is stored, managed and used in an enterprise, and establishes common guidelines for data operations.

Technical architecture: It refers to hardware and software architecture.

Process architecture: It is the structural design of general process systems and covers process design, logistics, policy and procedures. The overall inputs, outputs and functionality of the process architecture are also specified.

Zachman Framework: It is a tool supporting enterprise architecture, and was developed by John Zachman.

Using Audience Response Systems in the Classroom

David A. Banks
University of South Australia, Australia

INTRODUCTION

Audience response systems (ARS) are increasingly being introduced into educational settings, having previously proved their value in business. These systems make use of hand-held numeric input devices to allow students to enter data in response to questions or statements displayed on a public screen. The captured data is displayed on a public screen and enables both academics and students to immediately see how the whole group has responded. The anonymity afforded by an ARS encourages individuals to fully participate without fear of ridicule or loss of face.

The low cost ARS technology is simple to use by both students and academics, can be used with large (up to several thousands) or small groups and has applications in all topics of study and at all levels of study. ARS are highly portable, require little set-up time and are easy to use by anyone who has had some experience with software such as PowerPoint.

AUDIENCE RESPONSE SYSTEMS

ARS comprise hand-held input devices that transmit data to a receiving device connected to a personal computer. Software processes the datum and presents it in a variety of formats to the participants for discussion. Key components of the system are:

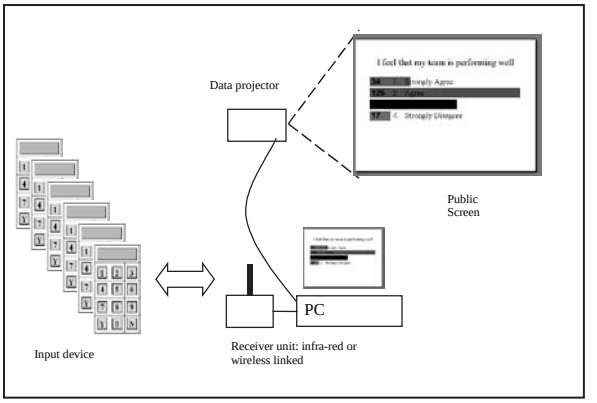
- **Hand-held input devices:** A variety of sizes and designs exist, from the credit-card size keypad with basic numeric input (Figure 1) through to systems that include a display screen to provide feedback to the user.
- **Receiver:** Utilizes infrared or other wireless communication media to collect data from the keypads.
- **Software:** Manages collection and processing of data and supports display of the data in a variety of presentational formats. The software may be embedded in other container software such as PowerPoint. The output from the system is usually displayed on a public screen via a data projector (Figure 2).

ARS IN HIGHER EDUCATION

Draper and Brown (2004, p. 20) suggest that "The dream of personal teaching is really about adaptive teaching; where what is done depends on the learner's current state of understanding." ARS can provide timely feedback to support this adaptive teaching goal, but Draper and Brown make the point that this can only be achieved through ap-

Figure 1. Credit-card size kepdad (Source: KEEpad Pty Ltd)

Figure 2. ARS components



propriate pedagogic design and action and not through the technology alone. In one-to-one or small group settings the learning facilitator may have a sense of the current state of the learner if the learner feels sufficiently comfortable in revealing it. With large groups in more formal settings the availability of cues to the learning facilitator can be more limited. The immediate feedback that an ARS offers can publicly identify differences or similarities of opinion within groups and provide a trigger for further discussion or analysis of data and re-adjustment of pacing or content. Audience Response Systems can be used with both large (hundreds of participants) and small groups (Banks, 2006) to support lectures, workshops, seminars, and to explore a wide range of subjects. They can be used at undergraduate and postgraduate levels, and within traditional and post-modern paradigms. Subject areas that value discussion, debate, multiple interpretations and direct challenges to accepted wisdom can benefit from this technology, but equally an ARS can be used in subject areas where demonstration of understanding of a fixed body of knowledge is vital. ARS can be used for formative and summative assessment, in the gauging of preliminary level and subsequent stages of understanding of a subject and in the exploration of the concepts that underpin critical issues.

Mitchell (2001) suggests that ARS can be used for mechanistic purposes such as monitoring class attendance via individual handsets, providing instant marking and feedback and for gathering data that can be used to support research activities related to classroom processes. McCabe, Heal and White (2001) used an ARS to support computer-assisted assessment (CAA) approaches with mathematics students and consider that it not only reinforced existing CAA activities but also served as a valuable tool for motivating higher levels of student learning. Hunt, Irving, Read and Knight (2003) used an ARS in a first-year information systems unit, in a decision-making subject in a third-year psychology course and also with second-year BS Pharmacy students. In the pharmacy course questions were posed via the ARS and the resulting answers were displayed and discussed by the whole group. A key issue here is that what is being sought is not necessarily a 'correct' answer but instead an examination and exploration of all possible answers and the reasons that individuals give for selecting a specific answer. The students expressed enthusiasm for the system, particularly in its ease of use, the ability to discuss answers immediately after making their choice and in the way it helped students identify where further reading was required. Importantly they also found it to be both easy and fun to use.

Post graduate HRM and MBA students using case-based approaches supported by an ARS indicated that the level of participation and number of ideas generated and explored was greater than usual and that the influence of individual personalities was greatly reduced. (Jones, Gear, Connolly, & Read, 2001). The students also observed that the technology was simple to use and to some extent became 'invisible' to the users. Williams (2003) notes that students on an MBA course using an ARS were strongly in favor of the technology and had negative views about passive learning approaches that simply involved reading or listening. Uhari, Renko and Soini (2003) reported that 80% of students studying a pediatrics course felt that an electronic voting system helped improve their learning and enhanced questioning during lectures. Witt (2003) found that 87% of students studying statistics for a psychology course saw more benefits than disadvantages in the use of keypads. Judson and Sawada (2002) reported that students consistently indicated that they are more likely to attend class, are challenged to think more deeply and feel that staff using such technologies learn more about them as individuals.

The anonymity afforded by ARS provides an opportunity for engaging students in sensitive subject areas where they may normally be reluctant to raise their hands. For example,

Wired News (2005) reported that Cheit used an ARS in an Ethics and Public Policy class at Brown University (USA) to explore the question 'Are you morally obliged to report cheating if you know about it?' The ARS quickly captured around 150 student responses indicating 64% in agreement with the statement and 35% in disagreement. The actual data captured here is of less significance, in this context, than the opportunity that is raised for discussion and for students to appreciate their own position in the light of others. The data captured during sessions such as this may prove useful for longitudinal studies of changing student attitudes or the responses of a variety of groups to such questions.

Although ARS are typically associated with large groups they can also offer useful support for small groups (Birdsall, 2002; Ward, Reeves, & Heath, 2003). Banks (2003) discusses the use of an ARS to support a diagnostic peer-review process with groups of around five students. This process offers an opportunity for students to become aware of the perceptions of their peers of their performance against a number of agreed group performance indicators. The use of this diagnostic session early in the course allows problems to be surfaced and identified, solutions discussed and action plans developed before any major tensions negatively affect the groups. One interesting effect that needs further investigation is the observation that overseas students who would normally be unwilling to engage in open discussion about the way they were perceived by others, or how they perceived others, would be willing to 'talk through the screen'. The ARS-based process seems to have produced a de-contextualization that allowed them to talk about themselves without embarrassment. ARS can also be used to support scenario-based activities (Banks & Bateman, 2004) that encourage groups to reflect on the ways in which individuals and groups perform and to act as a trigger for discussion of issues of communication, trust, negotiation, and teamwork in general.

ISSUES IN THE USE OF ARS IN HIGHER EDUCATION

Benefits of Anonymity

In groups where a number of different cultures are represented there is a danger that in traditional face-to-face settings some students will not offer an answer for fear of 'loss of face'. Some shy students may also feel inhibited in large groups and thus not feel able to contribute to the discussion. The anonymity afforded by these systems provides an opportunity for these potentially disadvantaged students to fully participate in the learning process. Chaitman (2005), however, cites the view of Bill Lewis, a computer science lecturer at Columbia University, that "Students will become more involved in their class and their own lives if they're not anonymous," More research is needed in this area.

More than Just the Numbers

Using ARS to collect numeric data from the individual keypads for testing or feedback purposes is an obvious asset. However, in more discursive learning environments the 'shape' of the collected data patterns is more useful in triggering critical discussion. Differences in displayed score patterns may indicate differing worldviews or interpretations of the data and this can provide an opportunity for reflection, sharing of views, critical thinking and deep learning. 'Flat' data may suggest that an ambiguous question has been asked whereas responses that locate at extreme ends of the scale, or are bipolar, suggest strong or opposing views that are worthy of further exploration.

Process vs. Content

The greater emphasis on process and engagement may lead to a feeling that some content has to be abandoned (Slain, Abate, Hodges, Stomatakis, & Wolak, 2004). This may be a useful outcome as it leads academics to consider what material needs to be included in any given part or mode of a course, and about what needs to be explored face-to-face and what can be supported by other mechanisms outside contact time. Slain et al. comment that their ARS based approach did suggest that there was a greater need for students to attend sessions having read appropriate material. Consideration of the balance of face-to-face, external and blended modes may offer improvements to the learning environment that go beyond the technology itself.

Question, then Question Again

By asking a question a number of times and critically evaluating the distribution of the ARS responses it becomes possible to explore the reasons for the differing or changing responses. Gauging the responses of students through the use of an ARS allows for quick and multiple loops around the material if learning appears to be problematic. D'Inverno, Davis & White (2003) report that the use of an ARS suggests that typically around 40% fail to identify the correct answer to simple questions, and that if the same question is asked again around 20% still provide the wrong answer. (They do, however, suggest that there may be some deliberate entry of incorrect answers as not all students feel that the technology offers them benefit).

Accessibility/Equity for All Students

Even though the cost of keypads is already low, and falling, the provision of one keypad per student clearly represents a large investment and technology management issue for an educational institution. Ownership of keypads by the institution also raises the problem of issuing and collecting

the keypads before and after learning sessions, maintenance and so on. One way to overcome this problem is to make the students responsible for the keypads. This can be achieved through the currently developing approach of providing students with shrink-wrapped packages of textbooks and keypads, the cost of the keypads being built in to the package purchase price. However, Stone (2004) notes that "Adding the clicker as a required supplement bundled with the new textbook not only helps sell new texts, it stifles sales of used copies that take away from publisher revenues … Further, clickers provided by most publishers won't work without activation, so the used clicker has little market value."

Standards

The number of systems available has grown dramatically in recent years. If different departments in an institution adopt an ad hoc approach to the adoption of these systems, the resulting collection of different brands will lead to incompatibility (Branen, 2005; Stone, 2004). Appropriate policies will need to be implemented to provide a managed environment.

- **Reliability:** Ideally the technologies can be taken into a room and quickly set up for use. In practice a number of problems have been reported in both set-up and operation. Delays impact upon the view that students gain of such systems, Branen (2005) noting that students in a Marketing class felt the systems to be 'an inefficient use of time' due to the wastage of fifteen minutes every day getting them to work. It is also noted that Infra Red keypads were not always detected by the system, leading to problems when they were used as a way of checking attendance or for assessment exercises.
- **Multiple choice:** ARS encourage the use of multiple choice questions and the apparent simplicity of implementation of such questions needs to be balanced against the need for careful design of the questions themselves. (Barrow & Blake, 2004). The development of approaches that allow publishers test banks linked to some books to be quickly converted into an ARS presentation may also raise concerns despite the potential benefits.

Current Developments

A number of systems that utilize personal digital assistant (PDA) technology and mobile phone technology are currently being developed to support text and graphical input, further increasing the versatility and power of these systems (Pelto, & Pelton, 2006; Jones, Marsden, & Gruijters, 2006; Dominick & Bishop, 2006). Many existing keypad vendors are developing their products so that the basic software can be used with virtual keypads implemented on networked PCs. For example one vendor has a virtual keypad that can be used with PDAs, laptops and networked PCs to provide all of the functions of the normal keypad plus text messaging, moment to moment polling and fastest response slides (KEEpad.com). As ARS continue to develop they will also start to link into Web-enabled learning systems to provide support for fully blended learning environments.

CONCLUSION

Considerable time and effort is being invested in distance learning via the Web, but it is equally important that the benefits of technology are applied to support and enhance more traditional face-to-face learning environments.

ARS technology provides educators with an opportunity to supplement their existing teaching and learning strategies in a way that provides them with improved, immediate and dynamic insight to the progress of learners. Students have an opportunity to engage with learning in an active way that helps them see how they, and their peers, are performing on a moment-to-moment basis. The opportunities for learning to be a shared experience are improved and there is the potential for engagement with subject material at a deeper level.

There are many research questions that need to be explored as this technology is introduced. These will include the potential for the use of the technology to promote deep learning, the long-term reactions of students to the technology versus the initial novelty factor, their effectiveness at different levels of study and in different subjects, and many other areas of concern will offer many opportunities for investigation. The use of frequent in-course evaluations rather than a single exit evaluation will allow student concerns to be addressed in a more timely way and will also allow discussion to take place between teacher and learner about both the evaluation instrument and the meaning of the captured data.

REFERENCES

Banks, D. A. (2003). Using Keypad-based Group Process Support Systems to Facilitate Student Reflection. In *Proceedings of the 20th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE)*, Adelaide (pp. 37-46).

Banks, D. A., & Bateman, S. (2004.) Using an audience response system to support a 'lost in the desert' learning scenario. In *Proceedings of the International Conference on Computers in Education (ICCE 2004)*, Melbourne.

Banks, D. A. (2006). Reflections on the use of ARS with small groups. In *Audience Response Systems in Higher Education: Applications and Cases*.

Barrow, G., & Blake, R. (2004). *Guessing or Assessing: Multiple Choice and the False Pass Problem*, GR Business Process Solutions. Retrieved from http://www.grbps.com/bkmini.pdf

Birdsall, S. (2002.). Assessment and Student Response Systems, *The Teaching Exchange*. Retrieved on February 6, 2006, from http://www.brown.edu/Administration/Sheridan_Center/pubs/teachingExchange/sept2002/assessment.shtml

Branen, J. (2005). In *The Spectator*, Student Newspaper of the University of Wisconsin, 16th May 2005, Campus News section. Retrieved on February 6, 2006, from http://www.qwizdom.com/articles/articles4web/article13.htm

Chaitman, , M. (2005). Is new ARS system the end of hand-raising? *The Tufts Daily*. Retrieved on February 8, 2006, from http://www.tuftsdaily.com/media/storage/paper856/news/2005/04/06/Features/Is.New.Ars.System.The.End.Of.HandRaising1490741.shtml?norewrite200608222030&sourcedomain=www.tuftsdaily.com

d'Inverno, R. A. Davis, H. C., & White, S.A. (2003). Student feedback: A lesson for the teacher. *Teaching Mathematics and Its Applications, 22,* 163-169.

Dominick, J., & Bishop, A. (2006). Instructor Mobile Audience Response System. In D. A. Banks (Ed.), *Audience response systems in higher education: Applications and cases*, Hershey, PA: Idea Group Inc.

Draper, S.W., & Brown, M. I. (2004). Increasing interactivity in lectures using an electronic voting system. *Journal of Computer Assisted Learning, 20,* 81-94.

Draper, S., Cargill, J., & Cutts, Q. (2001).Electronically Enhanced Classroom Interaction. In *Proceedings of the 18th Annual Conference of the Australasian Society for Computers in Tertiary Education* (pp. 161-167).

Hunt, A., Irving, A., Read, M., & Knight S. (2003). *Supporting learning with a group decision support system (GDSS)*. Retrieved on May 16, 2003, from http://www.pbs.port.ac.uk/~readm/GDSS/GDSS.htm

Jones, C., Gear, A., Connolly, M., & Read, M. (2001. Developing the professional skills of postgraduate students through Interactive Technology. In *Proceedings of the 2001 Information Resources Management Association Conference*, Toronto, Canada (pp. 759-761).

Jones, M., Marsden, G., & Gruitjers, D. (2006). Using Mobile Phones and PDAs in Ad Hoc Audience Response Systems. In D. A. Banks (Ed.), *Audience response systems in higher education: Applications and cases*. Hershey, PA: Idea Group Inc.

Judson, E., & Sawada. (2002). Learning from past and present: Electronic response systems in college lecture halls. *Journal of Computers in Mathematics and Science Teaching, 21*(2), 167-81. Retrieved on May 18, 2005, from http://www.aace.org/dl/files/JCMST/JCMST212167.pdf

KEEpad. Retrieved on May 14, 2003, from http://www.keepad.com

Keen, P.G.W., & Scott Morton, M. S. (1978). *Decision support systems: An organizational perspective*, Reading, MA: Addison-Wesley.

Littauer, R. (1972). Instructional implication of a low-cost electronic student response system. *Educational Technology, 12*(10), 69-71.

Mallach, E. G. (1994). *Understanding decision support systems and expert systems*. Sydney, Australia: Irwin.

McCabe, M., Heal, A., & White, A. (2001). Computer assisted assessment (CAA) of proof = proof of CAA: New approaches to computer assessment for higher level learning. In *Proceedings of the 5th International Conference on Technology in Mathematics Teaching.*

Menon, A. S., Moffett, S., Enriquez, M., Martinez, M. M., & Grappone, P. D. T. (2004). Audience response made easy: Using personal digital assistants as a classroom polling tool. *Journal of American Medical Information Association, 11*, 217-220.

Mitchell, C. (2003). *PRS Support System—A Summer Project, 2001*. Retrieved on May 16, 2003, from http://grumps.dcs.gla.ac.uk/papers/PRS_Support_System3.doc

Pelton, T., & Pelton, L. F. (2006). Creating a constructed response system to support active learning. In D. A. Banks (Ed.), *Audience response systems in higher education: Applications and cases*. Hershey, PA: Idea Group Inc.

Slain, D., Abate, M., Hodges, B. M., Stomatakis, M. K., & Wolak, S. (2004). An interactive response system to promote learning in the doctor of pharmacy curriculum. *American Journal of Pharmaceutical Education, 68*(5), Article 117. Retrieved on May 18, 2006, from http://www.ajpe.org/aj6805/aj6805117.pdf

Stone, T. (2005). Beware Publishing Reps Bearing, *Syllabus News*, July 25, 2004. Retrieved on June 20, 2006, from http://216.239.59.104/search?q=cache:0aO12F2mgpYJ:oregonstate.edu/itcc/wgs/ARS/OhioStateARS.doc

Uhari, M., Renko, M., & Soini, H. (2004). Experiences of using an interactive audience response system in lectures, *BMC Medical Education, 3*(12). Retrieved on February 23, 2004, from http://www.biomedical.com/1472-

Ward, C. R., Reeves, J. H., & Heath, B. P. (2003). Encouraging active student participation in chemistry classes with a Web-based, instant feedback, student response system.

CONFCHEM: Conferences on Chemistry, 2003. Retrieved on May 18, 2006 from http://aa.uncw.edu/chemed/papers/srs/confchem/confchem_srs.htm

Williams, J. B. (2003). Learning by remote control: Exploring the use of an audience response system as a vehicle for content delivery. In *Proceedings of the 20th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE)*, Adelaide (pp. 739-742).

Wired News (2006). Associated Press, May 15th 2005, *No wrong answer: Click it*. Retrieved on June 19, 2006, from http://www.wired.com/news/culture/1,67530-0.html

Witt, E. (2003). Who wants to be … The use of a personal response system in statistics teaching. *MSOR Connections*, *3*(2).

KEY TERMS

Anonymity: A feature of an ARS that can protect the identity of a participant. The default state is 'anonymity', but it is possible to collect unique keypad identifiers if the response of a specific individual or group is required (e.g. for attendance or testing).

Audience Response System (ARS): An electronic system designed to support and enhance face-to-face group interaction by means of individual hand-held communication devices. May also be referred to by a variety of names, including Classroom Response Systems, Student Response Systems, Interactive Response Systems, Interactive Student Systems and Electronic Voting Systems.

Group Decision Support System (GDSS): A collection of hardware and software used to support decision-makers.

Keypad: A hand-held device that allows a participant to communicate data to an Audience Response System. Also known as clickers or zappers.

Receiver or Base Station: A device that provides communication between the keypads and the ARS software. Increasingly these are permanently mounted in large lecture theatres to reduce set-up time.

Using Ontology and User Profile for Web Services Query

Jong Woo Kim
Georgia State University, USA

Balasubramaniam Ramesh
Georgia State University, USA

INTRODUCTION

Web services have received much attention because of their potential for realizing service oriented architecture (SOA). As the number of Web services increase exponentially, discovering Web services that satisfy user's specific needs has become a difficult task. In fact, the ability to find relevant Web services has been recognized as a key challenge in the realization of the potential of service oriented architectures.

The first step in the creation of system using SOA is to find relevant Web services which can be integrated or composed to provide the desired functionality (Kim & Jain, 2005). For example, developing a travel reservation application would require developers to search several Web services such as airline booking, hotel reservation, car rental, and weather forecasts. The commonly used UDDI-based Web services discovery approach is based on keyword search. Developers have to search for Web services using keywords such as "airline," "hotel," and "hotel" separately. This approach does not provide convenient search environment and deliver satisfactory query results to users because it does not take into account the context of the search (Zhou, Chia, & Lee, 2005). As the number of Web services grows explosively, it is almost impractical, if not impossible, for the user to efficiently analyze and combine the results with a keyword-based search approach (Matskin & Rao, 2002).

Ontology is a conceptualization of a domain into a human-understandable and machine-readable format. It consists of terms, their definitions, and axioms relating them (Gruber, 1993). Since ontology is the foundation for the semantic Web, which has been proposed as a mechanism to manage semantics and context of the World Wide Web, it can be used to improve the discovery of Web services by modifying or expanding query terms (McIlraith & Martin, 2003). However, little research has been done on the creation and use of ontology for Web services discovery. Our research seeks to address this void.

Furthermore, individual users may have unique preferences and needs even though they use seemingly similar queries to retrieve relevant Web services. User profiles can be used to capture these preferences and needs such that the queries return results that are of specific relevance to them. We propose the development of a query system that can incorporate user profiles so that queries accurately represent user's preferences and needs.

This chapter is organized as follows: In the next section, we discuss related research and various challenges and issues in Web services discovery. Then, we present several types of ontology and user profiles that may be used in the discovery process. Finally, we introduce a system architecture and discuss future trends and conclusions.

BACKGROUND

In this section, we provide background information about ontology and user profiles which can be used to address some of the challenges involved in Web services discovery.

Web Services

A Web service is an interface that describes a collection of operations that are network accessible through standardized XML messaging specifications such as SOAP, WSDL and UDDI. It provides open XML-based mechanisms for application interoperability, service description, and service discovery (Kim & Jain, 2005). A large number of Web services are already available on the internet, making Web services discovery a major task in service-oriented business application development. A widely used approach for discovering Web services is based on UDDI (Bin, Yan, Po, & Juanzi, 2005). UDDI uses a keyword based discovery feature which may not provide satisfactory query results because it does not take into account the context of the query. Recent research has proposed the use of ontology based query to improve the accuracy and relevance of search results (Bin et al., 2005; Maximilien & Singh, 2004; Zhou et al., 2005).

Ontology

The term ontology originally defines a philosophical discipline. As a branch of the philosophy, ontology deals with

the nature and organization of reality. Today, ontology is not only created by philosophers but also by computer scientists. In computer science, the term ontology is defined as the explicit specification of a conceptualization (Gruber, 1993). In other words, ontology represents the knowledge related to one or more domains in a way that may be interpreted by both for humans and computer programs. Ontology can support a variety of applications including the development of common understanding, enabling the reuse of domain knowledge, information extraction and concept tagging, knowledge management and intelligent systems (Noy & McGuinness, 2001).

User Profile

In the search for relevant Web services, which takes into account the context of the search, the profile of the user can play an important role. Integrated with the user's background and needs, a search can provide more personalized results (Storey, Sugumaran, & Burton-Jones, 2004). Such a search can exclude a large portion of irrelevant Web services by taking into account the user's particular interests.

When a query is contextualized, it produces results that account for (1) the meaning of query terms in the context in which they are used and (2) the user's preferences. In the development of a contextualized query, ontology can minimize the use of wrong concepts in the query whereas use profiles can help constrain the concepts requested (Storey et al., 2004) to those of interest to the user.

USING ONTOLOGY TO IMPROVE WEB SERVICE DISCOVERY

In this section we introduce several types of ontologies and propose that these ontologies may be used in a systematic manner to improve Web services query. We introduce several methodologies to create domain ontology which is a key ontology for Web services discovery.

Systematic Use of Ontology

Web services search can be improved by expanding users' original query with ontology (Storey et al., 2004). Different types of ontologies that can help improve queries are shown in Table 1 with examples. ResearchCyc is a general ontology that can handle terms which are independent of specific domains. In a similar fashion, existing upper-level ontologies such as SUMO or the Cyc Top-Level Vocabulary can be used for handling universal concepts such as time and space (Niles & Pease, 2001).

Linguistic ontology can handle a synonym in a query. Users may employ different terms to describe the same concept. In Web services discovery, a query system has to consider synonyms of terms used in a query as well as in the description of the Web services specified in the WSDL. For example, WordNet can be used as a thesaurus to cover extensive number of synonyms. Similarly, application ontologies can play a major role of improving users' query. Finally, search results can be improved with the use of domain ontology (Bin et al., 2005).

When there are a huge number of Web services, a query may return a large number of Web services which provide basically the same functions. Then, these results need to be organized by some criterion such as the quality of service (QoS). QoS ontology can help a query system rank order Web services according to the QoS level specified by user (Zhou et al., 2005). Although QoS typically represents non-functional requirements such as reliability and scalability, a QoS ontology can help formulate a more comprehensive assessment of quality (Maximilien & Singh, 2004; Zhou et al., 2005). Maximilien and Singh (2004) distinguish three ontologies for QoS: upper, middle, and lower. The QoS upper ontology describes general quality concepts such as quality measurement and relationships, whereas the QoS middle ontology captures several domain independent quality concepts such as availability, interoperability, and security. Domain-specific quality requirements are specified in the lower QoS ontology.

In summary, ontologies identified in Table 1 may be used to modify a user's query in Web services discovery. The user of such an ontology enhanced discovery process may provide feedback on whether the modifications to the query are appropriate. Through such an interactive process, a more comprehensive and appropriate query may be created.

Ontology Development

Existing ontologies shown in Table 1 are an excellent starting point for the development of an ontology based query system. Creating and using the first three ontologies in Table 1 is relatively easy compared with the rest of ontologies. Domain ontologies such as DAML ontology are also being developed as a part of large common-sense ontology. In fact, several ontologies which capture the same domain knowledge with different perspectives may be developed. Domain ontology needs to capture and represent concepts and their relationships shared among users. Therefore, the creation of domain ontologies is considered a time-consuming and difficult task (Cristani & Cuel, 2005; Noy & McGuinness, 2001). To reduce time and effort required to develop ontology, several methodologies have been proposed.

Cristani and Cuel (2005) classify ontology creation methodologies (such as DOLCE, OTK, TOVE, etc.) as top-down and bottom-up. Top-down methods start with an abstract view of domain and expand it with detailed specifications. Bottom-up methodologies start from the specification of a

Table 1. Types of ontology

Type	Task	Ontology
General ontologies	Represent general knowledge, independent of specific domain	ResearchCyc
Linguistic ontologies	Represent linguistic knowledge: different meanings of a given word, synonymy and antonym relationships between terms.	WordNet
Upper-level ontologies	Related to universal concepts like time and space	SUMO or Cyc Top-Level Vocabulary
Domain ontologies	Reflect a specific domain, independent of tasks	
Application ontologies	Imbedded in application, driven by tasks	The logical schema of a database
QoS ontology	Represent the level of non-functional requirements	DAML-QoS Ontology

certain task and obtain generalization. The choice of a methodology is dependent on the specific context and the needs of the application (Cristani & Cuel, 2005). For Web services discovery, the choice of methodology is influenced by the whether the Web services is in public UDDI or private UDDI. With a public UDDI, top-down methodology for ontology development is appropriate because public UDDI has to cover a broad range of topics. On the other hand, a bottom-up methodology is appropriate for a private UDDI because an organization running a private UDDI dedicated to specific domains may have very specific and detailed requirements from users. The bottom-up approach has been successfully used to develop ontology from text documents (Maedche & Volz, 2001; Shamsfard & Barforoush, 2004). Applying this methodology to Web services, we can create domain ontology from a collection of documents such as WSDL specifications. However, a challenge to this approach is the lack of adequate text documents. The amount of information available in WSDL specifications of Web services might be limited because developers tend to express the features of Web services succinctly. The other challenge is the time and effort required for the creation of a domain ontology.

Given these challenges involved in domain ontology creation for Web services query improvement, we propose two methodologies. Following Storey and Kim (2006), we propose a semi-automatic methodology. As this methodology uses World Wide Web as its resources, it can be easily applied to create a domain ontology for Web services. The proposed methodology shown in table 2 includes six steps: *(1) Identification of category, (2) specification of target domain keywords, (3) crawling and scanning relevant Web pages, (4) extraction of concepts, (5) clustering extracted concepts, and 6) construction of domain ontology*.

The six-step methodology has been used to create domain ontologies with less time and effort compared with manual development proposed by Storey and Kim (2006).

The second involves the pruning of a large common-sense ontology such as Cyc. Using keywords used in the domain, we can gather information related to those terms from a common-sense ontology, thereby creating relatively large domain ontology. Because this domain ontology is usually large and may contain unnecessary concepts, pruning methodologies proposed by Kim, Caralt, and Hilliard (2007) may be used to create a smaller domain ontology by removing irrelevant concepts. Then, developers can modify the domain ontology to suit their needs.

Tools for Creating and Managing Ontology

For ontology development and management, several tools and standards are available. Various tools for ontology editing and management include Protégé-2000, SWOOP, KAON, WSMX, OWL-S Editor, and OntoManager (Cristani & Cuel, 2005). The domain-independent languages used to specify ontologies in these tools include RDF(S), DAML+OIL, OWL, and KIF. Among these, OWL is considered a de facto standard for the semantic Web. OWL enables users to richly express ontologies and reason with this knowledge because it provides more comprehensive vocabulary for expressing meaning and semantics when compared to other standards such as RDF. For example, to more richly describe classes and properties, relations between classes and characteristics of properties are included in OWL. Ontology development tools can be used with the two approaches for domain ontology creation proposed in the previous section.

Using User Profiles to Improve Web Services Discovery

The role of user profile in improving the quality of search has long been investigated (Korfhage, 1984). The interactive development of a query can be made more effective by taking into account user's different preferences that are specified in a user profile.

Several types of user profiles and methods to create them have been proposed in the literature. Two types of

Table 2. Six-step methodology for ontology creation tools for creating and managing ontology

Steps	Activity
1. Identification of category	Identify category of target Web site using DMOZ or Clusty.
2. Specification of target domain keywords	Specify target domain (or Web sites) with initial domain keywords.
3. Crawling and scanning relevant Web pages	Collect Web pages identified relevant from step 2.
4. Extraction of concepts	Retrieve and organize words according to criteria such as frequency and TFIDF.
5. *Clustering extracted concepts*	Analyze the relation among concepts and cluster them.
6. *Construction of a domain ontology*	Create a domain ontology.

categorization of user profiles have been commonly used in prior research. One divides user profiles as knowledge-based or behavior based (Kuflik, Shapira, & Shoval, 2003). User's knowledge in the form of semantics is reflected in knowledge-based profiles whereas records of user's actions are used to create behavior-based user profiles (Middleton, Shadbolt, & De Roure, 2004). Knowledge-based profiling employs static models of users and dynamically matches each user to the closest model. To obtain the knowledge required for such matching, questionnaires and interviews are often used. Behavior-based profiling captures the user's behavior as a model. Machine learning techniques are employed to discover meaningful patterns in user's behavior (Kobsa, 1993).

User profiles may also be classified according to the level at which preferences are aggregated: (1) personal, (2) stereotype (i.e., held by a class of individuals), or (3) community (i.e., held by an entire community) (Middleton et al., 2004).

As community level profile is less likely to be useful for a specific user and context, we recommend the use of individual and stereotype level profile for query improvement. Further, depending on the expertise level of user, a query system can decide which level of group profile is appropriate (Storey et al., 2004). Queries of domain experts tend to be very stable and accurate. Therefore the query system should use only query terms provided by domain experts. In contrast, initial queries created by novices tend to be obscure and unstable. Therefore, the query system should use the expert stereotype profile first and then use the personal stereotype later. User profiles can be collected by both directly and indirectly depending on the user's expertise in a domain (Storey et al., 2004). When user is an expert in a certain field, direct method can provide accurate user profile information. However when user has little expertise in a certain field, it is difficult to collect profile information. To cope with this challenge, a novice profile can be created indirectly by using practice queries in a given topic.

A METHODOLOGY FOR THE CREATION AND USE OF USER PROFILES

We introduce a five-step methodology for the creation and use of user profiles that takes into account the recommendations made above. This methodology has been adapted from related research on the use of user profiles to improve search on the World Wide Web (Storey et al., 2004). Table 3 shows the five steps and relevant activities.

The five step methodology described above can be readily used in conjunction with the approach for the creation and use of domain ontology described in the previous section. Together, they synergistically help improve the retrieval of appropriate Web services.

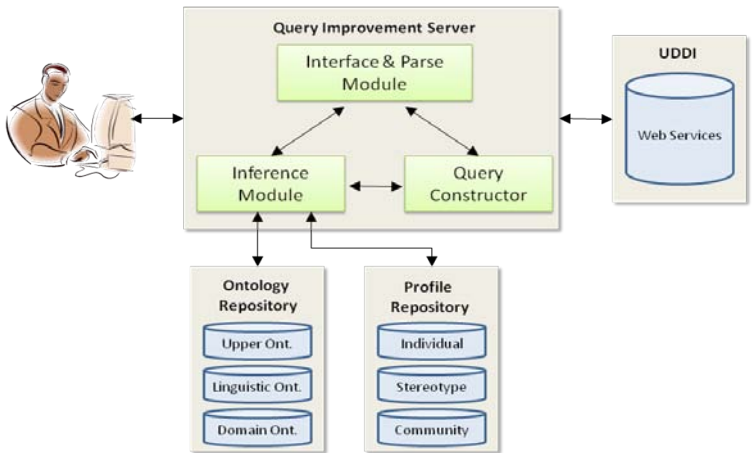
System Architecture and Scenario

Figure 1 shows the architecture of our proposed system for ontology and user profile driven Web services query. It consists of three modules and two repositories: (a) interface and query parser module, (b) inference module, (c) query constructor module, (d) ontology repository, and (e) user profile repository. The interface and query parser module interacts with a user to receive user's original query and deliver the result to the user. In addition, this module parses the user's original query. The inference module receives the terms from the interface and parse module. Interacting with ontology repository, the inference module tries to find synonyms and relationships between terms. Then, it uses the user profile information to further expand the query. The expanded query is used to search the Web services repository

Table 3. Five-step methodology for query development using profile

#	Steps
1	**Identify approach:** UDDI determines user's domain knowledge level from either interaction with user or previous records. Depending on user's level of domain knowledge, UDDI may suggest a delegated or user-driven query development.
2	**Identify appropriate type of user profile:** In case of a delegated query, UDDI determines a type of user profile (e.g., personal, stereotype, community). In case of a user-driven query, UDDI determines a type of user profile if a user is a novice.
3	**Identify terms from profile to increase query:** Unless user is an expert and chooses a delegated query, UDDI provides semantically relevant terms so that user can expand his/her query to increase accuracy of query.
4	**Execute query:** UDDI executes the query developed in step 3.
5	**Obtain feedback:** UDDI receives a feedback from user. When user is not satisfied with the query result, a new query creation procedure will be started.

Figure 1. System architecture



located in UDDI. Finally, the interface and parse module presents the selected Web services to the user.

We illustrate the use of the system with the following scenario in which ontology and user profile information are used to develop intelligent query. Suppose the user is interested in renting a GPS equipped car and creates a query which simply includes the terms "renting a GPS." When this incomplete query is used to search travel related Web services, if the UDDI does not have a specific Web services for renting a GPS, the query system will return no useful results. In contrast, our system establishes that (1) it does not find any Web service for renting a GPS, (2) it finds that GPS is part of a car from the domain ontology, and (3) it find that there is a Web service for renting cars. It will, therefore, return Web services that are relevant for renting a car.

The system selects the car reservation Web service by using the heuristic "if a specific Web service that describes an action A (reserve) for a concept C (GPS) does not exist, then such an action tends to be specified in the Web service that describes the same action A (reserve) over a concept C' (car) that includes C (GPS).

User profile information can be used to further improve query results. If the user profile information shows that the user lives in Atlanta and prefers discounted services, then the system presents only Web services from companies which are located in Atlanta and which offer a discount.

FUTURE TRENDS

For better support in Web services discovery, Web services need to be annotated in a semantically accurate manner. WSDL which is commonly used for this purpose is not descriptive enough to richly describe Web services. New standards such

as Web Services Agent Framework (WSAF) (Maximilien & Singh, 2004) and Web Services Modeling Framework (WSMF) (Fensel & Bussler, 2002) offer much promise.

Ontology can be used in Web services matchmaking to support composition (Zhou et al., 2005). For matchmaking, QoS ontology such as those focused on availability, interoperability, and security are promising.

A query may potentially be expanded with several domain ontologies. Research on ontology pruning and refactoring (Conesa & Olive, 2004) can support ontology expansion while creating accurate queries.

Further research on the appropriate ways to categorize user profiles is needed. Also, the relative effectiveness of knowledge-based or behavior-based user profile in improving Web services discovery needs to be empirically examined. In addition we are currently investigating the effectiveness of different levels of aggregation of user (e.g., individual, stereotype, and community) in the context of search for Web services.

CONCLUSION

Ontology and user profile can be used to improve Web services search by expanding query with more relevant terms which accurately represent context and user's preferences and needs. We identify several types of ontologies which capture different level of contextual information. The use of these ontologies in a systematic manner can help represent the context of a search in a query. While domain ontology may be very effective in improving queries, their creation requires a lot of time and effort. We introduce two methodologies to help create domain ontology with minimal effort. In addition, user profiles can further improve query results by capturing user's preferences and needs. We introduce a five-step methodology for the effective use of user profiles in the search for appropriate Web services. Finally, we propose an architecture for a system that incorporates our approach using ontology and user profile. The approach and system described here are aimed at achieving effective and efficient search for Web services.

REFERENCES

Bin, X., Yan, W., Po, Z., & Juanzi, L. (2005). *Web services searching based on domain ontology.* Paper presented at the Service-Oriented System Engineering, 2005. SOSE 2005. IEEE International Workshop.

Conesa, J., & Olive, A. (2004). Pruning Ontologies in the Development of Conceptual Schemas of Information Systems. *LECTURE NOTES IN COMPUTER SCIENCE,* 122-135.

Cristani, M., & Cuel, R. (2005). A survey on ontology creation methodologies. *Int'l Journal on Semantic Wen & Information Systems, 1*(2), 49-69.

Fensel, D., & Bussler, C. (2002). The web service modeling framework WSMF. *Electronic Commerce Research and Applications, 1*(2), 113-137.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199-220.

Kim, J. W., Caralt, J. C., & Hilliard, J. K. (2007). *Pruning Bio-Ontologies*. Paper presented at the 40th Annual Hawaii International Conference on System Sciences (HICSS'07).

Kim, J. W., & Jain, R. (2005, 03-06 Jan). *Web Services Composition with Traceability Centered on Dependency.* Paper presented at the HICSS '05. Proceedings of the 38th Annual Hawaii International Conference.

Kobsa, A. (1993). User modeling: Recent work, prospects and hazards. *Adaptive User Interfaces: Principles and Practice*, 111-128.

Korfhage, R. R. (1984). Query enhancement by user profiles. *Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*, (pp. 111-121).

Kuflik, T., Shapira, B., & Shoval, B. (2003). Stereotype-based versus personal-based filtering rules in information filtering systems. *Journal of the American Society for Information Science and Technology, 54*(3), 243-250.

Maedche, A., & Volz, R. (2001). The ontology extraction and maintenance framework text-to-onto. *Proceedings of the ICDM?1 Workshop on Integrating Data Mining and Knowledge Management*.

Matskin, M., & Rao, J. (2002). Value-Added Web Services Composition Using Automatic Program Synthesis. *Web Services, E-Business, and the Semantic Web, CAiSE 2002 International Workshop, WES*.

Maximilien, E. M., & Singh, M. P. (2004). A framework and ontology for dynamic web services selection. *IEEE Computer Society, 8*(5), 84-93.

McIlraith, S. A., & Martin, D. L. (2003). Bringing semantics to Web services. *IEEE Intelligent Systems, 18*(1), 90-93.

Middleton, S. E., Shadbolt, N. R., & De Roure, D. C. (2004). Ontological user profiling in recommender systems. *ACM Transactions on Information Systems, 22*(1), 54-88.

Niles, I., & Pease, A. (2001). Towards a standard upper ontology. *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001*, (pp. 2-9).

Noy, N. F., & McGuinness, D. L. (2001). *Ontology development 101: A guide to creating your first ontology* (No. KSL-01-05). Stanford: Stanford Knowledge Systems Laboratory.

Shamsfard, M., & Barforoush, A. A. (2004). Learning ontologies from natural language texts. *International Journal of Human-Computer Studies, 60*(1), 17-63.

Storey, V. C., & Kim, J. W. (2006). *Developing Domain Ontologies from Web Pages.* Paper presented at the Proceedings of the 5th AIS SIGSAND Symposium on Research in System Analysis and Design, Vancouver, Canada.

Storey, V. C., Sugumaran, V., & Burton-Jones, A. (2004). The role of user profiles in context-aware query processing for the semantic web. In F. M. A. E. Métais (Ed.), *Proceedings of the NLDB 2004* (Vol. 3126, pp. 51-63): LNCS.

Zhou, C., Chia, L.-T., & Lee, B.-S. (2005). Web services discovery with DAML-QoS ontology. *International Journal of Web Services Research, 2*(2), 43-66.

KEY TERMS

Ontology: A conceptualization of a domain. It is consisted of entities, attributes, relationships and axioms in a format which is human understandable and machine-readable.

Ontology Pruning: A method used to delete the elements of an ontology that are irrelevant.

SOA: An architectural style whose goal is to achieve loose coupling among interacting software agents. A service is a unit of work done by a service provider to achieve desired end results for a service consumer.

SOAP: An XML-based protocol to access objects from Web services.

TFIDF: A value to evaluate how important a selected term is within documents.

UDDI: An XML-based registry for businesses to publish and search Web services.

User Profile: User information representing his or her preferences and needs.

WSDL: An XML-based language used to describe Web services

Using Prolog for Developing Real World Artificial Intelligence Applications

Athanasios Tsadiras
Technological Educational Institute of Thessaloniki, Greece

INTRODUCTION

Artificial Intelligence Applications are becoming crucial for enterprises that want to be successful by having the advantage of using high information technology. The development of such applications is assisted by the use of high level computer programming languages that are closer to the programmer than to the computer. Such a programming language is Prolog.

Prolog is a logic programming language (Clocksin & Mellish 2003) that was invented by Alain Colmerauer and Phillipe Roussel at the University of Aix-Marseille in 1971. The name *Prolog* comes from *programmation en logique* (i.e., "programming in logic" in French). Together with LISP, they are the most popular Artificial Intelligence programming languages. Prolog was generated by an attempt to develop a programming language that extensively uses expressions of logic instead of developing a program by providing a specific sequence of instructions to the computer. Theoretically, it is based on a subset of first-order predicate calculus that allows only Horn clauses (Bratko, 2000). The control of the program execution is based on Prolog's built-in search mechanism that in fact is an application of theorem proving by first-order resolution.

BACKGROUND

Prolog has a clear contribution in solving a series of **Artificial Intelligence** (**AI**) problems (Sterling & Shapiro, 1986). There are many features that make **Prolog** suitable as a programming language for developing AI applications (Luger, 2002). Some of them are:

- **Declarative nature:** Programming in Logic using Prolog, remove the imperative (serial order) nature of other languages and allow programmer to solve a problem by describing the problem itself rather than defining a solution. The programmer writes programs by declaring the facts and the rules that apply to the problem in hand and then makes queries in order for Prolog to return valid solutions. This high level of abstraction makes **Prolog** suitable for AI applications where the programmers should give emphasis on the problem itself rather than on the computer idiosyncratic commands that they should impose to the computer system. **Prolog**'s built-in search mechanism takes the control of the program execution, leaving the programmer to concentrate on the problem itself.
- **Backtracking:** Even when a search path ends at a dead end, the backtracking mechanism of **Prolog** retreats back down the search path to try another path. This feature makes Prolog exceptionally suitable for a number of search problems that AI faces. It also provides the additional advantage of finding more than one solution of the problem, in the case that backtracking is forced, after finding a first solution. Because a great number of AI problems can be represented as a problem of finding the right path in the search space, the built-in depth first mechanism of Prolog, accompanied with backtracking, make **Prolog** suitable for such applications.
- **Don't care and don't know nondeterminism:** In the execution of a Prolog program, the nondeterminism feature is really apparent. Although the rule that will execute is the first one that matches the goal, we can ask for more than one solution. Using the backtracking mechanism, alternative rules will apply and other valid solutions will be found. This introduces: a) "don't know nondeterminism" implying that all possible ways to find the solutions will be followed because the execution does not know how to find the solution, or b) "don't care nondeterminism" meaning that we just need one solution and we do not care which solution is that among the many that exist. Both of these types of nondeterminsm are considered useful in AI applications, especially for those dealing with logic.
- **Recursion**, instead of **Iteration:** Because iteration constructs are not provided in **Prolog**, recursion should be used instead. The simplicity of solving problems recursively makes Prolog programs smaller and understandable even when coping with large, real life **AI** problems. Additionally, many AI problems are recursive in nature, increasing the suitability of Prolog.
- **List handling mechanism:** The data structure of List is very important for handling AI problems (e.g., LISP which is also used for solving AI problems, stands for

"LISt Processing"). Lists are built-in in Prolog while in most other languages they are not, making faster and easier the writing of Prolog programs that require list handling. List's recursive nature allows the extensive use of recursion in problem solving, providing an additional advantage for solving AI problems with Prolog.

- **Pattern matching** and **unification:** The use of unification to find the most general common instance of two formulas or patterns makes pattern matching a build-in feature of **Prolog**. This intelligent feature can assist AI problem solving where in many cases the decisions that are made are based on situation matching. This Prolog ability can be found really helpful in specific areas of AI, such as natural language processing, computer vision and intelligent database search.

PROLOG AND ARTIFICIAL INTELLIGENCE APPLICATIONS

The features described above make Prolog suitable for developing applications that solve **AI** Problems. Such an area is that of **Decision Support Systems**. Rules that support decisions can be expressed as Prolog rules, declaratively in pure logic, making development and maintenance of these systems mush easier. An example of a successful decision support system is the "Options Trading Analysis System" (OTAS, Lassez, McAloon, & Yap, 1987), used for the analysis of stock options and investment strategies. OTAS automatically generates and analyzes investment strategies based on standard vertical option combinations. Its main elements are: a portfolio maintenance module that creates and updates portfolios and provides expert recommendations, a numeric database containing stock market data, a symbolic database containing rules describing standard options combinations and a linear algebra module for the analysis of options combinations.

Except from financial decisions, medical decisions can be supported by similar systems written in Prolog. The OSM medical decision support system for general practitioners (Fox, Glowinski, Gordon, Hajnal, & O'Neill, 1990) is such an example. OSM supports a number of knowledge and information retrieval functions, providing the user with rapid access to textual information from text sources, knowledge bases or patient database. The system incorporates a version of the Oxford Textbook of Medicine, a 300-author general medical reference work, and uses its indexes to retrieve text.

Decision support systems can also take the form of a computer based advisor. For example, the RoadWeather Pro (Reiter, 1991) is used as an Expert Weather Advisor written in Prolog that permits mouse point-and-click manipulation of weather "objects," thereby allowing forecast upgrades based upon recent observational data received from sensors or human observers. This decision support system estimates weather-related effects on highway maintenance operations, as well as on airports, transportation, recreational activities, agribusiness and so forth. Its purpose is to be used as a user-interactive 24-hour weather prediction system for snow and ice control.

Another major **AI** area that Prolog contributes is that of **Natural Language Processing.** Pattern matching capabilities and the declarative nature of grammar definitions, make Prolog a handy and powerful tool for processing natural language. For example, CAT2 (Sharp, 1991) is a unification-based natural language processing system, designed for analysis, generation and translation of natural language sentences. CAT2 is used for multilingual machine translation and for automatic translation of informative texts although the emphasis has been on European Commission texts, as well as general purpose texts. It embodies a particular formalism for natural language processing, as well as a grammar development environment. Grammars have been written for English, German, French, Spanish, with experimental versions for Russian, Greek, and Japanese.

The Logic-based Machine Translation system, LMT (McCord, 1982, 1986) is another such application. This is a machine translation system for translating English to German. The system is based on a grammatical formalism called Modular Grammars based on slot filling techniques, which includes some automatic semantic translation and handling of metagrammatical rules. The principle aim of the system is to translate computer manuals from English into German.

Among the various **AI** applications using Prolog, many **Knowledge-based Systems** can be found. This is because, in most of the cases, knowledge in such systems is expressed in the form of rules. These rules can be easily expressed in **Prolog** syntax. After that, by using unification and Prolog search mechanism, inference can be done. AGATHA Electronic Diagnosis Knowledge Based System (Allred, Lichtenstein, Preist, Bennett, & Gupta, 1991) is such a system. Its aim is to test and diagnose complex printed circuit boards. Agatha uses a suite of smaller subsystems, each one of them customized to diagnose a particular kind of test, something that is necessary due to the diversity and complexity of the various tests. Agatha could reason about the test results as well as suggest further tests to run.

Moreover, Gunga Clerk (Woodin, 1989) is an example of a Legal Knowledge-based System written in Prolog. This system is a substantive legal knowledge-based advisory system in New York State Criminal Law, advising on sentencing, pleas, lesser included offences and elements. Gunga Clerk is designed to accept key facts of a criminal case and provide guidance to attorneys and judges as to statutory rules affecting sentence parameters, regulation of plea bargaining, identification of lesser included offences,

and offences chargeable based on designated conduct. Explanations include citations to legal authority and display of chains of legal inferences.

Another example of a Knowledge-based System written in Prolog is the ISCN Expert (Cooper & Friedman, 1990). This Health Knowledge-based System is able to interpret chromosomal abnormalities and allows geneticists to better reference and interpret chromosomal abnormalities such as those which result in Down Syndrome, mental retardation or physical disabilities. It interprets the International Human Cytogenetic Nomenclature, which is the standard notation used to represent human chromosomal abnormalities. These notations, each representing a person's genetic layout, are maintained in a computerized registry for reference and comparison against each other.

Prolog can also contribute to the area of **Intelligent Search in Large Databases.** This is because the only way to run a Prolog program is by making a query. This built-in feature can be used in order to make complex and efficient queries in large databases. An example of a database system coping with Prolog is ADAM (Gray, Kulkarni, & Paton, 1992; Paton & Diaz, 1991). ADAM is a general purpose object-oriented database, with emphasis on extensibility with new modeling constructs by using metaclasses. It adds the ability to structure Prolog programs and data using the object-oriented paradigm.

It is a fact that every **AI** problem that can be represented using graphs can be handled by Prolog search and backtracking mechanism. Having such an advantage, various Prolog systems have been developed to solve **Graph Theory Problems**. Conceptual Graph Tools (CGT, Wermelinger & Lopes, 1991), for example, is a Prolog-based knowledge representation system that has a partial implementation of Sowa's Conceptual Structures. As Sowa (1984) states, Conceptual Structures "are a system of logic with a graph based formalism that aims for a very wide expressive power. Its primary purpose is to serve as an intermediate language between natural language and other formalisms including database query languages [...] and predicate calculus." CGT includes predicates to implement the most important operations on conceptual graphs, like the canonical formation rules and the propositional inference rules. CGT reads and writes conceptual graphs using their linear notation. It also provides facilities to manipulate graph databases.

Another **AI** area where Prolog is used is that of **Scheduling and Planning**. The "Generate & Test" technique and the pattern matching mechanism of **Prolog**, make the test of candidate generated solutions a much simpler task. CAS/FPS, Computer-aided Synthesis of Flexible Production Scheduling (Csukas, Kozar, & Arva, 1989), for example, is a Prolog-based system that is used for the Production Planning and Scheduling of Multiproduct (Batch) Plants. Using multicriteria design, the system provides a computer-aided synthesis of the production plans and schedules from the possible building elements. In Prolog represented structural models, the various solutions are synthesized from the "free" active and passive elements of the structural model.

The OMAR, Operative Management of Aircraft Routing system (Torquati, Paltrinieri, & Momigliano, 1990) is also written in Prolog. It is an interactive Aircraft Scheduling system designed for the predictive scheduling of the Alitalia Fleet. The salving strategy that it uses combines network consistency and tree search techniques.

The unification and pattern matching mechanism of Prolog can also be found useful for **AI** problems of **Computer Vision**. Such a system written in Prolog is GEONS (Fairwood, 1991). The purpose of this system is to recognize the class of a 3-D volumetric primitive object in an image description which consists of curve properties and relations. The two-dimensional images of 3-D volumetric primitives are "input" in the form of facts about curves, lines and their properties and relationships (e.g., curved/straight, connectivity). This program models, in a qualitative way, (a) the 3-D objects, (b) the model-scene projection relationships, and (c) the image structure. These declarative models constitute a "parser" for the input curve data which is analyzed by the program to recognise the appropriate category of geometric primitive.

Prolog can also be useful for developing AI applications of **Game Playing.** The search space generated by games can be searched by Prolog's search mechanism, augmented by the backtracking feature. An example of such a Prolog system is the PYTHON (Sterling & Nygate, 1990). This intelligent system is used for recognizing and performing squeeze plays, an advanced strategy in the game of bridge. It performs, in its limited domain, at a truly expert standard, comparable to players of national ranking. PYTHON's core recognizes when a simple squeeze exists according to well-established theory. The core was extended to handle more complicated squeezes, also described by theory, making PYTHON's performance truly expert. Furthermore, methods were added for recognizing and executing squeezes not covered by existing theory by analogy with the other methods.

It should be also mentioned that Prolog's inference mechanism is suitable for developing applications coping with **Reasoning**, especially qualitative reasoning. GARP-General Architecture for Reasoning about Physics (Bredeweg, 1989), for example, is an integrated approach to qualitative prediction of behaviour. Given the description of a system (usually a physical system) GARP predicts the states of behaviour that the system will go through, in qualitative terms.

FUTURE TRENDS

Prolog is a successful logic programming language for developing AI applications. This will not change in the following

years, especially with the integration of **Prolog** with Internet applications and other computer languages such as Java. For example, the JIProlog (Java Internet Prolog, Chirico 2006), a Java platform Prolog interpreter which integrates Prolog and Java languages can add much of the functionality of Java to Prolog making Prolog programs easily delivered through Internet, all over the world. Furthermore, Prolog is a high level computer language that is closer to human than the computer machine. This type of languages is more probable to be used in the future because of the computer science trend of making computers friendlier to the users.

CONCLUSION

The logic programming features of **Prolog** make it suitable for developing programs coping with a wide variety of AI problems. The declarative nature of **Prolog**, accompanied by aspects such as backtracking, unification and pattern matching, recursion, the list handling mechanism and the built-in inference mechanism, provides the programmer with important and useful tools in order to confront AI problems.

The use of Prolog in the Internet and its integration with languages such as Java is a new and promising challenge for Prolog programmers that will give Prolog new potentials.

REFERENCES

Allred, D., Lichtenstein, Y., Preist, C., Bennett, M., & Gupta, A. (1991). AGATHA: An integrated expert system to test and diagnose complex personal computer boards. *Innovative Applications of Artificial Intelligence, 3*, 87-103. AAAI Press.

Bratko, I. (2000). *Prolog–programming for artificial intelligence* (3rd ed.). Addison-Wesley.

Bredeweg, B. (1989). Introducing meta-levels to qualitative reasoning. *Applied Artificial Intelligence, 3-2*, 85-100. New York.

Chirico, U. (2006). *JIProlog: Java Internet Prolog*. Retrieved December 14, 2007, from http://www.ugosweb.com/ jiprolog/

Clocksin, W.F., & Mellish, C.S. (2003). *Programming in Prolog: Using the ISO standard* (5th ed.). Springer-Verlag.

Cooper, G., & Friedman, J.M. (1990). Interpreting chromosomal abnormalities using Prolog. *Computers and Biomedical Research, 23*, 153-164.

Csukas, B., Kozar, Z., & Arva, P. (1989). Multicriteria evaluated PROLOG. *Synthesizing Algorithms, Computers Chemical Engineering, 13*, 595-602.

Fairwood, R.C. (1991). Recognition of generic components using logic-program relations of image contours. *Image & Vision Computing*, *9*(2), 113-122.

Fox, J., Glowinski, A., Gordon, C., Hajnal, S., & O'Neill, M. (1990). Logic engineering for knowledge engineering: Design and implementation of the Oxford system of medicine. *Artificial Intelligence in Medicine, 2*, 323-339.

Gray, P.M.D., Kulkarni, K.G., & Paton. N.W. (1992). *Object-oriented databases: A semantic data model approach*. Prentice Hall.

Lassez, C., McAloon, K., & Yap, R. (1998). Constraint logic programming and options trading. *IEEE Expert, 2*(3).

Luger, G. (2002). Artificial intelligence. *Structures & strategies for complex problem solving* (4th ed.) Addison-Wesley.

McCord, M.C. (1982). Using slots and modifiers in logic grammars for natural language. *Artificial Intelligence, 18*, 327-367.

McCord, M. (1986). Design of a Prolog-based machine translation system. In *Proceedings of the 3rd International Conference in Logic Programming* (pp. 350-374). Berlin: Springer-Verlag.

Paton, N.W., & Diaz, O. (1991). Metaclasses in object-oriented databases. In W. Meersman et al. (Eds.), *Object-oriented databases: Analysis, design and construction (DS-4)* (pp. 331-348). North-Holland.

Reiter, E.R. (1991). Hybrid modeling in meteorological applications. Part 1: Concepts and approaches. *Meteorology and Atmospheric Physics*, *46*, 77-99.

Sharp, R. (1991). CAT2: An experimental Eurotra alternative. *Machine Translation*, *6*, 215-228.

Sowa, J.F. (1984). *Conceptual structures: Information processing in mind and machine*. Addison-Wesley.

Sterling, L., Nygate, Y. (1990). PYTHON: An expert squeezer. *Journal of Logic Programming*, *8*, 21-39.

Sterling, L., & Shapiro, E. (1986). *The art of Prolog*. MIT Press.

Torquati, F., Paltrinieri, M., & Momigliano, A. (1990). A constraint satisfaction approach to operative management of aircraft routing. In *Proceedings of the Third International Conference of Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, Charlotte, NC, (pp. 1140-1146). ACM Press.

Wermelinger, M.L., & Lopes, G. P. (1991). A tool for knowledge acquisition and representation based on con-

ceptual graphs. In *Proceedings of the Eighth Brazilian AI Symposium*.

Woodin, D.E. (1989). Design and implementation of Gunga clerk: A substantive system in New York criminal law. *The Defender, Journal of the New York State Defenders Association, 35*.

KEY TERMS

Declarative Programming: Programming in the form of passive data structures such as facts and rules which can afterwards be used by active inference mechanisms. On declarative programming, the programmer states what is to be computed, and not how this is to be computed.

Backtracking: Backtracking occurs when in a search tree, the end of a search path is reached without a solution being found, and then the algorithm retreats back to a different path.

Don't Care Nondetermination: Solving a problem having in mind that just any one solution is enough. We do not care if there are other solutions or which one of the solutions we find.

Don't Know Nondetermination: Solving a problem having in mind that the execution of the program does not know how to find the solution, so all possible ways to find the solutions can be followed.

Generate and Test: A trial and error method of problem solving where a possible solution is generated and tested to check if it is successful. If it is not successful, then another possible solution is generated and tested and this goes on until a satisfactory solution is found. The solution found may not be optimal and not all possible scenarios are tested.

Iteration: The repeated execution of the same process a given number of times or until a specified result is obtained.

Recursion: The definition of a program in terms of itself. This gives the program the ability to call itself.

Unification: The process of finding the most general common instance of two expressions. This is the pattern matching technique used by Prolog computer language to match goals and subgoals in a program.

Video Content-Based Retrieval

Waleed E. Farag
Indiana University of Pennsylvania, USA

INTRODUCTION

Recently, multimedia applications have undergone explosive growth due to the monotonic increase in the available processing power and bandwidth. This incurs the generation of large amounts of media data that need to be effectively and efficiently organized and stored. While these applications generate and use vast amounts of multimedia data, the technologies for organizing and searching them are still immature. These data are usually stored in multimedia archives utilizing search engines to enable users to retrieve the required information.

Searching a repository of data is a well-known important task whose effectiveness determines, in general, the success or failure in obtaining the required information. A valuable experience that has been gained by the explosion of the Web is that the usefulness of vast repositories of digital information is limited by the effectiveness of the access methods. In a nutshell, the above statement emphasizes the great importance of providing effective search techniques. For alphanumeric databases, many portals (Acuna, Marcos, Gomez, & Bussler, 2005) have become widely accessible via the Web. These portals use search engines that adopt keyword-based search models in order to access the stored information, but the inaccurate search results of these search engines is a known issue.

For multimedia data, describing unstructured information (such as video) using textual terms is not an effective solution because they cannot be uniquely described by a number of statements. That is mainly due to the fact that human opinions vary from one person to another (Tešic & Smith, 2006), so that two persons may describe a single image by totally different statements. Therefore, the highly unstructured nature of multimedia data renders keyword-based search techniques inadequate. Video streams are considered the most complex form of multimedia data because they contain almost all other forms such as images and audio in addition to their inherent temporal dimension.

One promising solution that enables searching multimedia data in general, and video data in particular, is the concept of content-based search and retrieval (Deb, 2005). The basic idea is to access video data by their contents—for example, using one of the visual content features. Realizing the importance of content-based searching, researchers have started investigating the issue and proposing creative solutions. Most of the proposed video indexing and retrieval prototypes have the following two major phases (Marques & Furht, 2002):

1. The **database population phase** consists of the following steps:
   - *Shot Boundary Detection:* The purpose of this step is to partition a video stream into a set of meaningful and manageable segments (Hanjalic, 2002), which then serve as the basic units for indexing.
   - *Key Frames Selection:* This step attempts to summarize the information in each shot by selecting representative frames that capture the salient characteristics of that shot.
   - *Extracting Low-Level Features from Key Frames:* During this step, some of the low-level spatial features (color, texture, etc.) are extracted in order to be used as indexes to key frames and hence to shots. Temporal and other features (e.g., object motion) are used also.
2. In the **retrieval phase,** a query is presented to the system that in turns performs similarity matching operations and returns similar data (if found) back to the user.

It is worth mentioning that a growing trend in current content-based retrieval systems is the application of contextual constraints to enrich those systems with additional metadata (Davis, King, Good, & Sarvas, 2004). The use of context makes video retrieval systems both content-based and context-based systems at the same time. Besides, context-based techniques try to improve the retrieval performance by using associate contextual information, other than those derived from the media content (Hori & Aizawa, 2003).

In this article, each of the above stages will be reviewed and expounded. Background, current research directions, and outstanding problems will also be discussed.

VIDEO SHOT BOUNDARY DETECTION

The first step in indexing video databases (to facilitate efficient access) is to analyze the stored video streams. Video analysis can be classified into two stages (Farag & Abdel-Wahab, 2002b), *shot boundary detection* and *key frames extraction.* The purpose of the first stage is to partition a

video stream into a set of meaningful and manageable segments, whereas the second stage aims to abstract each shot using one or more representative frames.

In general, successive frames (still pictures) in motion pictures bear great similarity among themselves, but this generalization is not true at boundaries of shots. A shot is a series of frames taken by using one camera. A frame at a boundary point of a shot differs in background and content from its successive frame that belongs to the next shot (except in the case of gradual transitions). In a nutshell, two frames at a boundary point will differ significantly as a result of switching from one camera to another, and this is the basic principle that most automatic algorithms for detecting scene changes depend upon.

Due to the huge amount of data contained in video streams, almost all of them are transmitted and stored in compressed format. While there are many algorithms for compressing and representing digital video data, the MPEG family (Watkinson, 2004) is the most famous one and the current international standard. In MPEG, spatial compression is achieved through the use of a Discrete Cosine Transform (DCT)-based algorithm similar to the one used in the JPEG standard. In this algorithm, each frame is divided into a number of blocks (8x8 pixel), then the DCT transformation is applied to these blocks. The produced coefficients are then quantized and entropy encoded, a technique that achieves the actual compression of the data. On the other side, temporal compression is accomplished using a motion compensation technique that depends on the similarity between successive frames on video streams. Basically, this technique codes the first picture of a video stream (I frame) without reference to neighboring frames, while successive pictures (P or B frames) are generally coded as differences to those reference frames. Considering the large amount of processing power required in the manipulation of raw digital video, it becomes a real advantage to work directly upon compressed data and avoid the need to decompress video streams before manipulating them.

Several research techniques were proposed to perform the shot detection task for both cuts and gradual transitions. For instance, template matching and histogram comparison are commonly used. Statistical models are also proposed as in Hanjalic (2002). Farag and Abdel-Wahab (2002b) proposed the use of supervised learning systems in order to detect shout boundaries. Moreover, other techniques such as finite state machine and support vector machines are used to identify various types of transitions (Liu, Gibbon, Zavesky, Shahraray, & Haffner, 2007).

KEY FRAMES SELECTION

The second stage in most video analysis systems is the process of *key frames* (KFs) selection (Deb, 2005), which aims to abstract every shot using one frame or more. Ideally, we need to select the minimal set of KFs that can faithfully represent each shot. KFs are the most important frames in a shot, hence they may be used to represent the shot in the browsing system as well as be used as access points. Moreover, one advantage of representing each shot by a set of frames is the reduction in the computation burden required by any content analysis system to perform similarity matching on a frame-by-frame basis, as will be discussed later. KFs selection is an active area of research in visual information retrieval, and a quick review of some proposed approaches follows.

Clustering algorithms are proposed to divide a shot into M clusters, then choose the frame that is closest to the cluster centroid as a KF. Cooper and Foote (2005) introduced the use of linear discriminant analysis to select representative frames. The VCR system (Farag & Abdel-Wahab, 2002a) employs two algorithms to select KFs, *accumulated frames summation* (AFS) and *absolute luminance differences* (ALD). The AFS is a dynamically adapted algorithm that uses two levels of threshold adaptation, one based on the input dimension while the second level relies on a shot activity criterion to further improve the performance and reliability of the selection. AFS employs the accumulated summation of luminance differences of corresponding *direct current* (DC) coefficients in successive frames. The second algorithm, ALD, uses absolute luminance difference between the summation of all DC terms in a frame and the same summation of the next frame. It utilizes a statistical criterion for the shot-by-shot adaptation level. A comprehensive review and classifications of video abstraction techniques introduced by various researchers in the field is presented in Truong and Venkatesh (2007). That work reviewed different methodologies that use still images (key frames) and moving pictures (video skims) to abstract video data and provide fast overviews of the video content.

FEATURE EXTRACTION

To facilitate access to large video databases, the stored data need to be organized; a straightforward way to do such organization is through the use of index structures. In case of video databases, we even need multi-dimension index structures to account for the multiple features used in indexing. Moreover, we are in need of tools to automatically or semi-automatically extract these indexes for proper annotation of video content. Bearing in mind that each type of video has its own characteristics, we also need to use multiple descriptive criteria in order to capture all of these characteristics.

The task of the feature extraction stage is to derive descriptive indexes from video data content such as from selected key frames in order to represent the original data. These indexes are then used as metadata, and any further similarity

matching operations will be performed over these indexes and not over the original video data. Ideally, content-based retrieval (CBR) of video should be accomplished based on automatic extraction of content semantics, which is usually a difficult task. Thus, most of the current techniques only check the presence of semantic primitives or calculate low-level visual features. There are mainly two major trends in the research community to extract indexes for proper video indexing and annotation. The first one tries to automatically extract these indexes, while the second trend performs iconic annotation of video by manually (with human help) associating icons to parts of the video stream. One example of the latter trend uses a multi-layered representation to perform the annotation task where each layer represents a different view of video content. On the other hand, works on the first trend, automatic extraction of content indexes, can be divided into three categories:

- *Deriving indexes for visual elements using image-indexing techniques*—For example, using the color and texture as low-level indexes.
- *Extracting indexes for camera motion (panning, zooming, etc.)*—Generally, optical flow is used in such techniques.
- *Deriving indexes for region/object motion*—One system detects major objects/regions within the frames using optical flow techniques.

Color and texture are commonly used indexing features in many indexing systems. Moreover, they are among the proposed descriptors in the MPEG-7 standard (Manjunath, Salembier, & Sikora, 2002). Color feature extraction can work directly on the original decoded video frame or on its DC form. One technique converts the color space of DC video frames (YCbCr in the case of MPEG) to the traditional RGB color space, then derives color histograms from the RGB space. Deriving the histogram can be done in many ways. An efficient technique uses some of the most significant bits of each color component to form a codeword. For instance, the most significant two bits of each color component can be selected and concatenated to form a 6-bit codeword. This codeword forms a 64-bin color histogram that is used as the color feature vector. This histogram is a good compromise between computational efficiency and representation accuracy.

Many techniques are proposed in the literature to perform texture feature extraction. Some of them use auto-regression and stochastic models, while others use power spectrum and wavelet transform (Bimbo, 1999). The main disadvantage of these techniques is they are computationally expensive.

THE RETRIEVAL SYSTEM

The basic objective of any automated video indexing system is to provide the user with easy-to-use and effective mechanisms to access the required information. For that reason, the success of a content-based video access system is mainly measured by the effectiveness of its retrieval phase. The general query model adopted by almost all multimedia retrieval systems is Query By Example (QBE; Marchionini, 2006). In this model, the user submits a query in the form of an image or a video clip (in the case of a video retrieval system) and asks the system to retrieve similar data. QBE is considered to be a promising technique since it provides the user with an intuitive way of query presentation. In addition, the form of expressing a query condition is close to that of the data to be evaluated.

Upon the reception of the submitted query, the retrieval stage analyzes it to extract a set of features, then performs the task of similarity matching. In that task, the query-extracted features are compared to the features stored in the metadata, then matches are sorted and displayed back to the user based on how close a hit is to the input query. A central issue here is how the similarity matching operations are performed, and based on what criteria (Farag & Abdel-Wahab, 2003). This central theme has a crucial impact on the effectiveness and applicability of the retrieval system.

Many techniques have been proposed by various researchers in order to improve the quality, efficiency, and robustness of the retrieval system. Some of these techniques are briefly reviewed below:

- *Relevance Feedback:* In this technique the user can associate a score to each of the returned clips, and this score is used to direct the following search phase and improve its results (Oerlemans, Rijsdam, & Lew, 2007).
- *Clustering of Stored Data:* Media data are grouped into a number of clusters in order to improve the performance of the similarity matching.
- *Use of Linear Constraints:* To come up with better formal definitions of multimedia data similarity, some researchers proposed the use of linear constraints that are based on the instant-based-point-formalism.
- *Improve Browsing Capabilities:* In this technique KFs and mosaic pictures are used to allow easier and more effective browsing.
- *Use of Time Alignment Constraints:* The application of this technique can reduce the task of measuring video similarity to be equivalent to finding the path with minimum cost in a lattice. The latter task can be accomplished using dynamic programming techniques.

- *Optimize Similarity Measure:* Defining optimized formulas to measure video similarity instead of using the exhaustive similarity technique in which every frame in the query is compared with all the frames in the database (a computationally prohibitive technique).
- *Human-Based Similarity Criteria:* This technique tries to implement some factors that most humans probably use to assess the similarity of video data (Farag & Abdel-Wahab, 2003).

Many of the above techniques end up calculating the similarity between two frames. The equation below is one example of how similarity between the colors of two frames (*I* and *M*), represented by their color histograms, can be calculated using the normalized histogram intersection (Farag & Abdel-Wahab, 2003). Each histogram is scaled before applying the equation in order to account for variations in the dimension of video frames. If Sim approaches 0, the frames are dissimilar, and if it tends toward 1, the frames are similar in color.

$$Sim(I, M) = \frac{\sum_{i=1}^{n} \min(I_i, M_i)}{\sum_{j=1}^{n} M_j}$$

where:

n = the number of bins in the color histogram, and
min(Ii, Mi) = the intersection between the ith bin values in both frames.

FUTURE TRENDS

Some of the important issues under investigations are surveyed briefly in this section. There is still a need for efficient algorithms to parse video streams containing gradual transition effects. Detecting semantic objects inside video frames and successfully correlating them to other extracted low-level features are open challenges that need to be addressed. Bridging what is called the semantic gap is one of the most promising areas of research in visual information retrieval. Moreover, multi-dimension indexing structure is one area that requires further examination. Research in effective and efficient techniques for assessing the similarity of video data, in particular those that capture human notions of multimedia similarity, is very crucial to the advancement in the field. Investigation of methodologies for performance evaluation of multimedia retrieval systems and the introduction of benchmarks such as the TRECVID effort are two other areas that need more research. Also, the application of human-computer interface techniques and information visualization strategies in order to include the user intelligence requires further investigation. Finally, the recent introduction of various standards for representing digital video data, such as the MPEG-7 standard, calls for in-depth evaluation of the effectiveness of these standards in improving the performance of video retrieval systems. Moreover, research on how to optimize the use of tools and techniques provided by these standards to enhance current systems and how to properly apply these standards in various domains is a promising trend.

CONCLUSION

In this article, we briefly reviewed the need and significance of video content/context-based retrieval systems and explained their four basic building blocks. Current research topics and future work directions were covered as well. The first stage, the Shot Boundary Detection, divides video streams into their basic shots. Each shot is then represented by one or more key frame(s) in a process known as key frames selection. The feature extraction stage, the third one, derives descriptive indexes such as color, texture, shapes, object motion, and so forth from video content and stores these feature vectors as metadata. Finally, the retrieval system accepts a user query, compares indexes derived from the submitted query with those stored in the metadata, then returns search results sorted according to the degree of similarity to the query. In the end, we need to emphasize the importance of content/context-based video retrieval systems and assert that there is still a considerably large number of open issues that require further research to achieve more efficient and robust indexing and retrieval systems.

REFERENCES

Acuna, C., Marcos, E., Gomez, J., & Bussler, C. (2005). Toward Web portals integration through semantic Web services. *Proceedings of the IEEE International Conference on Next Generation Web Services Practices* (pp. 223-228).

Bimbo, A. (1999). *Visual information retrieval.* San Francisco: Morgan Kaufmann.

Cooper, M., & Foote, J. (2005). Discriminative techniques for key frame selection. *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 502-505).

Davis, M., King, S., Good, N., & Sarvas, R. (2004). From context to content: Leveraging context to infer media metadata. *Proceedings of the ACM International Conference on Multimedia* (pp. 188-195).

Deb, S. (2005). *Video data management and information retrieval.* Hershey, PA: Idea Group.

Farag, W., & Abdel-Wahab, H. (2002a). Adaptive key frames selection algorithms for summarizing video data. *Proceedings of the 6th Joint Conference on Information Sciences* (pp. 1017-1020).

Farag, W., & Abdel-Wahab, H. (2002b). A new paradigm for analysis of MPEG compressed videos. *Journal of Network and Computer Applications, 25*(2), 109-127.

Farag, W., & Abdel-Wahab, H. (2003). A human-based technique for measuring video data similarity. *Proceedings of the 8th IEEE International Symposium on Computers and Communications* (ISCC'2003) (pp. 769-774).

Hanjalic, A. (2002). Shot-boundary detection: Unraveled and resolved? *IEEE Transactions on Circuits and Systems for Video Technology, 12*(2), 90-105.

Hori, T., & Aizawa, K. (2003). Context-based video retrieval system for the life-log applications. *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval* (pp. 31-38).

Liu, Z., Gibbon, D., Zavesky, E., Shahraray, B., & Haffner., P. (2007) A fast, comprehensive shot boundary determination system. *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 1487-1490).

Manjunath, B., Salembier, P., & Sikora, T. (2002). *Introduction to MPEG 7: Multimedia content description language.* New York: John Wiley & Sons.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM, 49*(4), 41-46.

Marques, O., & Furht, B. (2002). *Content-based image and video retrieval.* Boston: Kluwer Academic.

Oerlemans, O., Rijsdam, J., & Lew, M. (2007). Real-time object tracking with relevance feedback. *Proceedings of the 6th ACM International Conference on Image and Video Retrieval* (pp. 101-104).

Tešic, J., & Smith, J. (2006). Semantic labeling of multimedia content clusters. *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 1493-1496).

Truong, B., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications, 3*(1), 1-37.

Watkinson, J. (2004). *The MPEG handbook.* St. Louis, MO: Focal Press.

V

KEY TERMS

Content-Based Access: A technique that enables searching multimedia databases based on the content of the medium itself and not based on keywords description.

Context-Based Access: A technique that tries to improve the retrieval performance by using associate contextual information, other than those derived from the media content.

Key Frames Selection: The selection of a set of representative frames to abstract video shots. Key frames (KFs) are the most important frames in a shot, hence they can represent the shot in both browsing and similarity matching operations, as well as being used as access points.

Query By Example: A technique to query multimedia databases where the user submits a sample query such as an image or a video clip and asks the system to retrieve similar items.

Retrieval Stage: The last stage in a content-based retrieval system where content features are extracted from the submitted query then compared with those stored in the metadata. Finally, matched ones are returned to the user.

Shot Boundary Detection: A process with the objective of partitioning a video stream into a set of meaningful and manageable units.

Video Indexing: The selection of indexes derived from the content of the video to help organize video data and metadata that represents the original video stream.

Video Shot: A sequence of contiguous video frames taken using the same camera.

Videoconferencing for Schools in the Digital Age

Marie Martin
Carlow University, Pittsburgh, USA

INTRODUCTION

Wallis and Steptoe (2006) tell of a "dark little joke" that is bandied about among certain educators. It recounts the tale of Rip Van Winkle, who on reawakening in 2006 after his hundred years' sleep, experiences utter bewilderment until at last he finds solace in the familiar environment of a classroom, where teaching is going on as it did back in 1906.

The story is amusing. The message is blunt. In the middle of the first decade of the 21st century, despite ongoing technology-driven societal transformation, schools are still functioning largely in the easily recognizable traditional model of the industrial era (Steinkuehler, 2006; Veletsianos, 2007). The rush to computerize the classroom has generally not brought about a corresponding change of mindset on the part of educators (Cuban, 2006; Spector, 2000; Shaffer, Squire, Halverson, & Gee, 2005; Thornburg, 2003). Schools are failing to address the needs of the Net generation of learners (Barnes, Marateo, & Ferris, 2007; van 't Hooft, 2007). These **digital learners** who have grown up in a technology-saturated world that has defined and shaped their way of learning find school irrelevant and boring (McCombs, 2000).

By drawing on the literature and on case studies from within the experience of the author and other educators in Northern Ireland (NI), this article seeks to demonstrate that **videoconferencing,** alone as well as alongside other technologies, and used with appropriate **pedagogy,** can help transform the traditional classroom and make it a place hospitable to the learning needs of the Net generation.

BACKGROUND

Higher education led the way in the use of **videoconferencing** in education by introducing it into distance learning courses (Anderson & Rourke, 2005; Hayden, 1999). Although the technology has been available to schools since at least the early 1990s (Cole, Ray, & Zanetis, 2004), it has generally remained underutilized and undervalued (Anderson & Rourke; 2005; Comber, Lawson, Gage, Cullum-Hanshaw, & Allen, 2004). In addition, there is a dearth of specific research on the use of videoconferencing in the K-12 classroom (Anderson & Rourke; 2005; Heath & Holznagel, 2002). However, much of the literature on videoconferencing in higher education is relevant to teaching and learning in the pre-tertiary sector. Coventry (1995) writes that the effectiveness of videoconferencing for learning lies in exploiting its capacity to facilitate effective learning through enabling dialogue. She defines learning as "a social process involving the active construction of new knowledge and understanding through individual learning and group and peer interaction" (Part Two: 1). This resonates closely with how **digital learners** actually learn (Barnes et al., 2007).

Rowan (2000) identifies a significant barrier to the uptake of videoconference technology by classroom teachers,viz a lack of information concerning the ways in which it can operate within an educational program. She lists three crucial points for teachers wishing to use videoconferencing which, again, would help to engage digital learners: the need for **interaction** to keep students engaged; the need for videoconferencing to be part of a **multimode delivery**; and the need for teaching with videoconferencing to be integrated with new teaching methodologies and with appropriate professional development.

A recurring warning in both higher education and specific K-12 literature is that the acquisition of technology does not guarantee acceptance and effective use (Greenberg, 2004), nor does sophisticated technology necessarily equate with an effective learning experience (Coventry, 1995). Generally, research findings focus on the primacy of **pedagogy**, and highlight three main points: the need to set the learning goals first; the need to become familiar with the unique capabilities of videoconferencing, enabling real time interactive visual and verbal communication between two or more sites; and the need to be aware that **videoconferencing** challenges the traditional teaching and learning paradigm, requiring a **pedagogy** that will exploit its capabilities and achieve the learning goals. The key to this is **interaction** which must be designed into the distant lesson plan and fostered throughout the videoconference (Amirian, 2003; Anderson & Rourke, 2005; Comber et al., 2004; Greenberg, 2004).

Another recurrent theme is the potential of **videoconferencing**, particularly when used with other technologies, to address the needs of learners with different **learning styles** (Heath & Holznagel, 2002). Yet another theme is the need for teacher training, both technical and pedagogical, in the use of **videoconferencing** (Coventry 1995; Greenberg, 2004;

Rowan, 2000) and for **student scaffolding** (Amirian, 2003; Anderson & Rourke, 2005).

In terms of achievement, there is little research on quantifiable outcomes (Comber et al., 2004). Cavanaugh (2001) reports that videoconferencing yields better learning outcomes, when used as a supplement to other technologies and other approaches to enable more reality-based learning. An equally constant theme is the enhancement of "soft" skills or "portable" skills by good **videoconferencing**–mediated learning experiences. One study from the UK on a major *Videoconferencing in the Classroom* project (Comber et al., 2004) makes the significant point that these **"life skills"** are perceived by teachers as feeding into performance gains. Another study from Ireland has found that special needs students in particular gain enhanced self-esteem, motivation, and improved communication and social skills, as well as greater attention to task, from regular experience of **videoconferencing**, and that this is by far their preferred technology (Abbott, Austin, Mulkeen, & Metcalfe, 2004).

GOOD PRACTICE FROM NORTHERN IRELAND

Videoconferencing has remained in the **"early adopter"** phase for the last 15 years (Greenberg, 2004). The barrier identified by Rowan (2000) still exists and teachers remain unsure how to use videoconferencing effectively. There is, however, much isolated good practice in many parts of the world, but it is generally not widely disseminated. To encourage more fruitful and creative take-up, this good practice needs to be made more easily available. Teachers like to hear it from teachers. In this spirit, I offer a few—out of many (Martin, 2002, 2005)—case studies from Northern Ireland. They cover the period from the mid-1990s to the present day. They are selected to give an idea of how **videoconferencing** can be used in the classroom to meet the needs of 21st century learners. Taken together, they exemplify many of the good pedagogical strategies identified by the literature.

Global Awareness and Curriculum Enrichment through Collaboration at the Elementary Level

This case study involves a small rural school with partners in Denmark and Italy. Learning goals were set first by the teachers and the project was developed as an integral part of the schools' history and geography curricula. The schools met regularly by multipoint **videoconference** to allow the grade 4 students to socialize, and to engage in interactive learning events. The teachers used the **videoconference** as a **virtual staffroom** where they met to plan the content of the next phase and to devise strategies for **interaction**, for variety of pace, and for different learning styles. They were enthusiastic about exploiting the capabilities of the technology to enrich and extend the learning experiences of their students. They also began to explore the implications for teachers of this new learner-centered approach, seeing the traditional "sage on the stage" role as being ill-suited to the new learning environment. The students were motivated by the "hard fun" aspect of preparing work for a real audience. They deepened their understanding of their history and geography syllabus, and developed cultural awareness, and communication and presentation skills. The principal of the NI school pointed out how greatly the videoconferencing learning environment had widened the horizons of his students, most of whom had never been out of Northern Ireland and were not likely to be in the near future.

Extending the Learning Environment through Collaboration between Students with Special Needs

A recent project, involving two schools for students with severe learning difficulties (SLD) in different parts of Northern Ireland, set out to explore the potential of **videoconferencing**, together with text conferencing for the more able, to support effective distant collaborative learning in the SLD environment. The project was on the theme of "My town." The aim was to give students, whose experience of the outside world was necessarily limited, a sense of their own and other places. The students communicated once a week by videoconferencing. The teachers used the technology as a **virtual staffroom** where they met regularly for discussion. Teacher enthusiasm for this technology, their willingness to change their **pedagogy** to exploit its potential, and their commitment to **student scaffolding** both before and during the virtual lessons were the key success factors.

The outcomes were very encouraging. Videoconferencing was experienced as being inclusive of all the children, including those with little or no oral skills who used body language to communicate, or who had their more articulate peers interpret for them. It increased their motivation and improved their concentration. Students with serious behavioral difficulties remained totally attentive and engaged throughout the 30-minute sessions. Both classes perceived videoconferencing as fun. It brought a new way of learning into the classroom. It awakened their curiosity about their distant peers and the distant town. Videoconferencing also whetted their appetite for a physical encounter with their virtual friends. This happened subsequently on two very happy occasions, with each school hosting the visit of the other.

Accessing and Sharing Experts

The capability of **videoconferencing** to overcome the barriers of distance, time, and expense, and to engage students enthusiastically in authentic learning was demonstrated in a multipoint linkup between 70 A-Level (grade 12) politics students from schools in Northern Ireland, their peers in New Jersey, and the House of Representatives in Washington, DC. From here, a senior U.S. Congressman held a "town meeting" with both groups of students, answering their questions about the U.S. constitution and the workings of the House, while an Irish expert on the British and Irish situation responded from the Irish site to the questions of the U.S. students. To accommodate the "intimate" nature of communicating by videoconference (Coventry, 1995), a panel of six students at each site interacted with the experts. Very thorough preparation of content and student scaffolding in advance of the conference, including raising awareness of the unique capabilities of videoconferencing, and development of presentation skills, led to a completely student-centered learning experience. During the preparatory phase, NI school district personnel liaised between the schools and the experts. Questions were sent in advance and the interactive nature of the conference was emphasized. Understanding was reached that information was to be delivered in "nuggets" rather than as a lecture. Effective chairing during the videoconference ensured a high degree of dialogue and **interaction** between students and experts and between students and their distant peers. Following the panel discussion, a roving radio microphone enabled the "town meeting" to be opened to the large group of observers at both sites. The two sets of students then had the opportunity to discuss their new learning and to build on previous knowledge. The whole "meeting" was a superb example of exploiting the potential of videoconferencing to meet the needs of **digital learners** by facilitating learning as a social process, enabling "hard fun," active construction of meaning through dialogue in a social process of group and peer interaction as described by Coventry (1995) and Hayden (1999).

Virtual School Day

A Virtual Day experiment was undertaken in 2007 by two NI schools. A number of students from both schools accessed learning from home, using **videoconferencing** and Learning NI—the Northern Ireland online managed learning environment—on which teachers had posted the lessons for collaborative learning that day. Videoconferencing enabled students to take part in classes being held at each school and to have brief tutorials with teachers. It was also used for collaborative work between the two schools. This experiment demonstrated that the use of **videoconferencing** as part of what Rowan (2000) describes as a **"multimode" delivery** had allowed good learning to take place. The informal evaluation indicated that students particularly enjoyed the independent and collaborative learning that the technologies enabled. Teachers saw great potential in this media-rich approach and realized how it would challenge the tradtional teaching and learning paradigm.

Teachers-as-Learners

In addition to being used to facilitate distance professional development courses within Northern Ireland (Martin, 2000), **videoconferencing** has occasionally provided an invaluable global dimension to the formal development of teachers-as-learners. An example of good practice in this domain was the delivery by a panel of three math specialists from a U.S. university, of part of an in-service course for primary school numeracy (math) coordinators in a school district. The focus was parental involvement in numeracy. Again, the strategy of a panel followed by open discussion was used. On each side, the panel comprised faculty, practising teachers, a parent and a student. This enabled the issues to be viewed from a variety of perspectives and fostered a high degree of **interaction.** Teachers subsequently reported that the experience had promoted their professional development, raised their self-esteem, widened their horizons, and helped them appreciate the value of videoconference as a technology for learning.

FUTURE TRENDS

In terms of technology, the future of **videoconferencing** is bright. Earlier criticisms of unreliability, complexity and cost are no longer valid (Arnold, Cayley, & Griffith, 2002) and with the advent of Internet Protocol (IP), calls are free. Technology convergence and mobile technologies will hasten anytime-anywhere conferencing, while the combination of videoconferencing with video streaming technologies is already enabling archiving of conferences as well as facilitating observation by a wider audience who can provide feedback through other media (Arnold et al., 2004). In addition, microwave radio and video links are beginning to enable videoconferencing from open space locations (Martin, 2008).

The future also looks brighter in terms of the adoption of videoconferencing in schools. Greenberg (2006) estimates that about 25% of U.S. schools have been equipped with videoconferencing. In Northern Ireland, where the technology infrastructure is supplied free, a basic videoconferencing system is in the process of being rolled out to all schools. Some schools are also purchasing high specification dedicated videoconferencing units.

Ultimately, what matters, however, is what mindset we will bring to **videoconferencing**. The main barrier to its effective use will be what Thornburg (2003) calls the "hu-

man challenge," getting educators to change their thinking about learning and the role of technology, so that schools will come to embody the spirit, not of the past, but of the future (Papert & Caperton, 1999).

There are signs that this is beginning to happen. The New Commission on the Skills of the American Workforce has released a "blueprint for rethinking American education" to better prepare students for the challenges of the global economy (Wallis & Steptoe, 2006). In Northern Ireland, the new thinking has led to the revised curriculum with its reduction of prescribed content and its emphasis on providing students with the skills necessary for life and work in the 21st century (Department of Education, 2005). Here, videoconferencing is increasingly being seen as a technology that can help meet these 21st century learning needs.

CONCLUSION

The literature and the case studies make it clear that videoconferencing can address the new needs of **digital learners** and help prepare them for the new market place, as well as being inclusive of the needs of learners with special needs and of teachers-as-learners. The case studies in particular give us grounds for hoping that schools, like Rip Van Winkle, may at last be awakening from a hundred years' sleep and becoming more aware of the need for that change of mindset and **pedagogy** that will exploit the unique capabilities of videoconferencing and allow it to help transform learning in the 21st century classroom.

REFERENCES

Abbott, L., Austin, R., Mulkeen, A., & Metcalfe, N. (2004). The global classroom: Advancing cultural awareness in special schools through collaborative work using ICT. *European Journal of Special Needs Education, 19*(2).

Amirian, S. (2003, October 31). Pedagogy & videoconferencing: A review of recent literature. In *Paper presented at the First NJEDge.NET Conference*, Plainsboro, NJ. Retrieved May 29, 2008, from http://www.lle.mdx.ac.uk/lle/alt/amirian_megacon.pdf

Anderson, T., & Rourke, L. (2005). *Videoconferencing in kindergarten-to-grade 12 settings: A Review of the literature*. Retrieved May 29, 2008, from the Canadian Association of Distance Education Research and Athabasca University, Center for Distance Education: http://www.vcalberta.ca/community/litreview.pdf

Arnold, T., Cayley, S., & Griffith, M. (2002). (rev. ed. 2004). *Video conferencing in the classroom: Communications technology across the curriculum.* Devon: Devon County Council.

Barnes, K., Marateo, R. C., & Ferris, S. P. (2007). Teaching and learning with the Net generation. *Innovate 3*(4).

Cavanaugh, C. S. (2001). The effectiveness of interactive distance learning technologies in K-12 learning: A meta-analysis [Electronic version]. *International Journal of Educational Telecommunications, 7*(1), 73-88. Retrieved May 29, 2008, from http://www.unf.edu/~ccavanau/CavanaughIJET01.pdf

Cole, C., Ray, K., & Zanetis, J. (2004). *Videoconferencing for K-12 classrooms.* Eugene, OR: International Society for Technology in Education.

Comber, C., Lawson, Gage, J., Cullum-Hanshaw, A., & Allen, T. (2004). *Report for schools of the DfES video conferencing in the classroom project*. University of Leicester School of Education, University of Cambridge. Retrieved May 29, 2008, from http://www.Becta.org.uk/page_documents/research/video_conferencing_report_may04.pdf

Coventry, L. (1995). *Video conferencing in higher education.* Retrieved May 29, 2008, from http://www.agocg.ac.uk/reports/mmedia/video3/video3.pdf

Cuban, L. (2006, October 31). *1:1 Laptops transforming classrooms: Yeah, sure* (ID No. 12818). Retrieved May 29, 2008, from http://www.tcrecord.org/Content.asp?ContentID=12818

Department of Education, Northern Ireland. (2005). *Questions and answers on the revised curriculum*. Retrieved May 29, 2008, from http://www.deni.gov.uk/index/22-postprimaryarrangements-new-arrangements_pg/22-ppa-questions_and answers_pg/22-ppa-faq-curr_pg.htm

Greenberg, A. (2004*). Navigating the sea of research on videoconferencing-based distance education: A platform for understanding research into the technology's effectiveness and value.* Retrieved May 29, 2008, from http://www.wainhouse.com/files/papers/wr-navseadistedu.pdf#search=%22Alan%20Greenberg%202004%20Navigating%20Sea%20Research%20Video%20conferencing%22

Greenberg, A. (2006). *Taking the wraps off video conferencing in the U.S. classroom: A state-by-state analysis*. Retrieved May 29, 2008, from http://www.wrplatinum.com/Downloads/5912.aspx?Relo=1

Hayden, K. L. (1999). *Videoconferencing in K-12 education: A Delphi study of characteristics and critical strategies to support constructivist learning experiences*. Doctoral dissertation, Pepperdine University, CA (UMI No. 9934596).

Heath, M. J., & Holznagel, D. (2002, October). Interactive videoconferencing: A literature review. In *Paper presented at the K-12 National Symposium for Interactive Videoconferencing*, Dallas, TX. Retrieved May 29, 2008, from http://neirtec.terc.edu/K12vc/resources/litpolicy.pdf

Martin, M. (2000). Videoconferencing in teaching & learning: Case studies and guidelines. Western Education & Library Board, Omagh, Northern Ireland. Retrieved May 29, 2008, from http://www.welb-cass.org/site/projects%5Finitiatives/video%5Fconferencing/downloads/VCT&L.pdf

Martin, M. (2002). Unlocking the potential of videoconferencing. Information Management, 15(3/4), 22-26.

Martin, M. (2005). Seeing is believing: The role of videoconferencing in distance learning. *British Journal of Educational Technology, 36*, 397-405.

Martin, M. (2008). Integrating videoconferencing into the classroom: A perspective from Northern Ireland. In D. L. Newman, L. Falco, & S. Silverman (Eds.), *Videoconferencing technology in K-12 instruction: Best practices and trends*. Hershey, PA: Idea Group.

McCombs, B. L. (2000, September). Assessing the role of educational technology in the teaching and learning process: A learner-centered perspective. In *Paper presented at Secretary's Conference on Educational Technology 2000*, Washington, D.C. Retrieved May 30, 2006, from http://www.ed.gov/rschstat/eval/tech/techconf00/mccombs_paper.html

Papert, S., & Caperton, G. (1999). *Vision for education: The Caperton-Papert platform*. Retrieved May 29, 2008 from http://www.papert.org/articles/Vision_for_education.html

Rowan, L. (2000). *Surfing electronic waves: The application of videoconference technology in tertiary teaching.* Retrieved May 29, 2008, from http://ultibase.rmit.edu.au/Articles/online/rowan1.htm

Shaffer, D. W., Squire, K. R., Halverson, R., & Gee, J. P. (2005, June). *Video games and the future of learning* (WCER Working Paper No.2005-4). Madison: University of Wisconsin-Madison, Wisconsin Center for Education. Retrieved May 29, 2008, from http://www.academiccolab.org/resources/gappspaper1.pdf

Spector, J. M. (2000). *Trends and issues in educational technology: How far we have not come*. Retrieved May 29, 2008, from http://suedweb.syr.edu/faculty/spector/publications/trends-tech-educ-eric.pdf

Steinkuehler, C. (2006, November 17). *Virtual worlds, learning, & the new pop cosmopolitanism* (ID No. 12843). Retrieved May 29, 2008, from http://www.tcrecord.org/Content.asp?ContentId=12843

Thornburg, D. (2003). *Instructional technology research online*. Retrieved May 29, 2008, from http://www2.gsu.edu/~wwwitr/interviews/thornburg.htm

Valetsianos, G. (2007, June 22). *Book review: Web-based instruction: A practical guide for online courses*. Teachers College Record (ID No. 14529). Retrieved February 6, 2008 from http://www.tcrecord.org/PrintContent.asp?ContentID=14529

van 't Hooft, M. (2007). Schools, children, and digital technology: Building better relationships for a better tomorrow. *Innovate 3*(4).

Wallis, C., & Steptoe, S. (2006, December 18). How to bring our schools out of the 20th century. *Time,* 51-56.

KEY TERMS

Digital Learners: Also known as the Net generation, technology-savvy students who have grown up immersed in technology and whose way of learning is shaped by this.

Early Adopters: Teachers who embrace new technologies at an early stage and experiment with its use in the classroom.

Pedagogy: The science or theory of teaching young people.

Talking Head: A teacher who uses videoconferencing in lecture-style format without interacting with learners at the remote site.

Twenty-First Century Skills: Skills required by the global market place, such as communication, ability to work in teams often across cultures, problem solving.

Videoconferencing: Audiovisual communication, either point-to-point (between two sites) or multipoint (between more than two sites), that enables interaction and application sharing between people in real time.

Virtual Staffroom: Meeting place for geographically separated teachers, where interaction is made possible by videoconferencing or computer technology.

Viewing Text-Based Group Support Systems

V

Esther E. Klein
Hofstra University, USA

Paul J. Herskovitz
College of Staten Island, CUNY, USA

INTRODUCTION

"[T]he word is not necessarily what it seems…." (Bialik, Revealment and Concealment, 2000, p. 17)

With interdisciplinary approaches leading to new and enriched perspectives, we argue that an encounter between information technology (IT) and sociology will result in a heightened understanding of the problem of textual ambiguity in text-based computer-mediated communication (CMC)[1] in general and in group support systems (GSS) in particular. Such approaches where IT meets sociology have already been taking place in other areas of group research (e.g., see Ahuja & Carley, 1998). "[W]ith the global and technological transformations of the workplace" (Aakhus, 2001, p. 341), as IT and the Internet gain wide acceptance throughout society as well as the global economy (e.g., see Friedman, 2000) and as CMC and computer-supported cooperative work (CSCW) become commonplace, both information systems (IS) scholars and sociologists have increasingly studied the patterns of human behavior in virtual groups. This article[2] is an attempt to advance that effort. Specifically, the purpose of this article is to apply the insights of Georg Simmel—an early and oft-neglected German theorist of sociology working in the late nineteenth and early twentieth centuries—on written communication to text-based GSS, which are interactive computer-based information systems that support and structure group interaction and intellectual teamwork (see also Ackermann & Eden, 1994; Fjermestad, J., 2004; Klein, 2000; Klein & Dologite, 2000; Nunamaker, 1997; Poole & DeSanctis, 1990; Zigurs & Buckland, 1998), "promot[ing] communication, collaboration and coordination among teams of people" (Ahalt, 2000, p. 1159).

BACKGROUND: COMPUTER-MEDIATED COMMUNICATION AND TEXTUAL AMBIGUITY

CMC research has pointed out that, unlike face-to-face (FTF) communication, CMC is distinguished by the absence of contextual (also referred to variously as situational or emotional) cues, which contributes to miscommunication, misunderstanding, misinterpretation, and distortion of the text message (e.g., see Sproull & Kiesler, 1986). In particular, text-based CMC media generate a written message unaccompanied by nonverbal communication. Nonverbal communication refers to "the exchange of information and meaning through facial expressions, gestures, and movements of the body" (Giddens & Duneier, 2000, p. 96; see also Schaefer, 2001, pp. 71-72), which are known as nonverbal cues. Nonverbal communication also includes verbal cues, or voice patterns, such as loudness, pitch, rate, and tone.

According to Easterbrook (1995, p. 6), contextual cues "are used [in FTF communication] for constant feedback and as a signalling mechanism …, indicat[ing] whether the listener is hearing, and understanding." Text-based CMC does not convey the context and emotional nuances that are necessary for an accurate understanding of the text message. For example, by e-mail and other text-based CMC, the pitch and tone of voice, hand motions, facial expressions, and eye movements are absent, often leading to textual ambiguity, which undermines the accuracy of the message.

Research from a variety of disciplines has recognized the significant role that facial expressions, body language, and voice patterns play in giving context to words. By way of illustration, Gottman (1994), investigating facial expression of emotion in married couples, reported that eye rolling by one spouse following comments by the other spouse is an important indicator of a troubled marriage and a strong predictor of divorce. In contrast to FTF interactions, text-based CMC forecloses the facial and vocal expression of emotion.

GROUP SUPPORT SYSTEMS AS LEANER MEDIA

With e-collaboration having assumed a pivotal role in organizations, a widely-used type of text-based CMC media is GSS, which have been used as e-collaboration tools to assist in intellectual teamwork in such activities as problem solving, decision making, idea generation, strategic planning, conflict resolution, and negotiations. GSS have been defined

as "networked, computer-based systems designed to facilitate structured, interactive discussion in a group of people communicating face-to-face or remotely, synchronously or asynchronously" (Davison & Vogel, 2000, p. 3; see also Aiken & Carlisle, 1992; Anson, Fellers, Kelly, & Bostrom, 1996). GSS, which permit "a group of users to collaborate electronically, sharing and updating a common database while allowing for intergroup communications" (Ullrick, 2000, p. 11; see also Hopkins, 1998, p. 96, note 5) are text-based in that "[g]roup member type their contributions into the system, which immediately makes each contribution available to all other participants" (Davison & Vogel, 2000, p. 3). Thus, what participants in a GSS-supported see is the written text without the benefit of contextual cues to serve as an *explication de texte* of sorts.

As GSS is a text-based e-collaboration tool, we suggest that it is worthwhile to analyze GSS in terms of media richness theory, which asserts that different media differ in their ability to transmit information, convey meaning, and change understanding (Daft & Lengel, 1984, 1986; see also Bastress & Harbaugh, 2003; Kahai & Cooper, 2003; Martz & Reddy, 2005; Simon & Peppas, 2004; Stenmark, 2002; Ware, 2000). Media with multiple contextual cues (e.g., body posture, gazes, voice) and rapid feedback are denoted as "rich," while media with few or no contextual cues and without quick feedback are classified as "lean."

Under many, but not all, circumstances (see Kock, 1998), media richness is an important determinant of the effectiveness of groups engaged in intellectual teamwork and collaboration. According to Majchrzak, Rice, King, Malhotra, and Ba (2000, p. 45):

Because of the kind of information they can transmit (nonverbal cues, etc.), some channels (face-to-face, videoconferencing, etc.) are particularly suited for tasks that are unanalyzable, non-routine, equivocal and involve manageable amounts of information. Unanalyzable tasks that teams might perform include strategic direction-setting, brainstorming, and conflict resolution.

FTF communication, from the perspective of media richness theory, is rich because of the multiplicity of nonverbal and verbal cues, which can be used to clarify and interpret the spoken message. By contrast, text-based GSS are generally leaner media to the extent that there is an absence of these multiple contextual/emotional cues. However, it should be noted, that multiple cues need not be absent in all GSS configurations. Kock (1999, p. 14) characterized the difficulty of enriching GSS through added features by noting "the persistent attempts of developers of commercial group support software, through adding features to their products, to achieve the elusive communication richness of face-to-face interaction."

In discussing the characteristics of electronic media, Yamauchi, Yokozawa, Shinohara, and Ishida (2000, p. 330) commented:

[E]lectronic media make it difficult to transmit equivocal messages, whose ambiguity in meaning permits multiple interpretations, because of the limited amount of communicative cues and sluggish interaction …. Face-to-face informal conversations are the richest medium and thus easily accommodate equivocal messages while written messages are more rigid and convey less information.

TEXT-BASED GSS THROUGH THE PRISM OF SIMMELIAN SOCIOLOGY

We now consider GSS, a staple of twenty-first century IT, from the vantage point of early twentieth century Simmelian sociology. Simmel has been "generally considered the most neglected of the founders of modern sociology" (Marshall, 1998, p. 601), although over the past two decades he has received increasing recognition in the law review literature (e.g., see Froomkin, 1996; Nagan & Hammer, 2004; Patterson, 1988; Reisman, 1983; Seul, 2004; Warner, 2005; West-Newman, 2005). We argue that Simmel anticipated media richness theory and the studies on the absence of contextual cues in CMC media, in his analysis on written communication. Contrasting the letter with FTF communication, Simmel (1908/1964, p. 353) asserted:

Individuals in physical proximity give each other more than the mere content of their words. Inasmuch as each of them sees *[emphsasis on original] the other, is immersed in the unverbalizable sphere of his mood, feels a thousand nuances in the tone and rhythm of his utterances, the logical or the intended content of his words gains an enrichment and modification for which the letter offers only very poor analogies. And even these, on the whole, grow only from the memories of direct personal contact between the correspondents.*

This insight represents a basis with which to view and analyze written communication, which, for Simmel, fails to convey the information, predominantly of an emotional nature, embedded in verbal and nonverbal cues. (For a recent paper on Simmel and cyberspace communication, see Bogard, 2000.) The absence or paucity of these contextual cues make letters prone to ambiguity, which, in Simmel's view, is a "sociological categor[y] of first rank" (p. 354). In terms of media richness theory, written communication is a lean communication medium as it is without the interpretive glosses provided by multiple contextual cues and rapid feedback. Simmel, thus, was an advocate of media richness theory *avant la lettre*.

Writing within the legal tradition, law scholar Jeffrey Rosen (2000), in investigating privacy in the workplace, applied Simmel's thought on letters to e-mail, where verbal and nonverbal cues are not available to supply emotional background to the text message, thus making the text message highly susceptible to being "wrenched out of context" (p. 76). According to Rosen, "e-mail shares the informality of a conversation, but like a letter, lacks the contextual accompaniments that provide clues to meaning in face-to-face encounters" (p. 76). Employing a Simmelian analysis, Rosen (p. 75) argued:

Messages sent by e-mail are often far more impetuous than face-to-face conversations, where "situational cues," such as body language and facial expressions, from the person with whom we are conversing temper what we say. Because e-mail messages are often dashed off quickly and sent immediately, without the opportunity for second thoughts that ordinary mail provides, they may, when wrenched out of context, provide an inaccurate window on someone's emotions at any particular moment.

Taking the works of Simmel and Rosen as our starting point, we extend the Simmelian analysis to text-based GSS, which we argue is, in many respects, similar to e-mail. Synchronous (same time) GSS messages are generally "dashed off quickly and sent immediately" without much deliberation. Even asynchronous (different times), GSS messages are often composed quickly without careful crafting. Moreover, the text messages in GSS, both synchronous and asynchronous, are without "contextual accompaniments" (Rosen, 2000, p. 75) that convey the emotional nuance of the text and clarify its meaning. As in letters, "a lack of all those accompaniments — sound of voice, tone, gesture, facial expression" (Simmel, 1908/1964, p. 354) are a source of ambiguity in text-based GSS messages because by doing without these accompaniments, GSS messages are without "[t]hese powerful denoters of information" (McNeil, Robin, & Miller, 2000, p. 701).

FUTURE TRENDS

Recently, an increasing amount of scholarly attention has been directed to GSS-supported and online focus groups. Hughes and Lang (2004) noted the following:

[I]n the few years that Internet communications have become widely acceptable, substitute cues have been developed which are already to some extent standardized and familiar to experienced online users. We have found that emoticons [also known as smileys or graphic accents] …, typographical cues …, standard acronyms …, all-uppercase text …, and interjections … are spontaneously introduced by group members to richen the text-only experience. (p. 98)

It is suggested that as individuals gain more experience with e-mail and other forms of e-communication, they will integrate emoticons seamlessly within their text messages thereby injecting some emotional context in an otherwise affectively neutral medium.

CONCLUSION

A Simmelian approach to written communication can contribute to a better understanding of the limitations of text-based GSS. The challenge to scholars and practitioners of IS is to develop pragmatic solutions to prevent textual ambiguity in GSS messages. These solutions should include fostering increased use of emoticons, which convey emotion and feeling and thus contribute to reducing misunderstandings and distortions of the computer-mediated text (e.g., see Crystal, 2001).

It is worth noting that the emoticons can only convey emotional cues that the sender of the message intentionally chooses to reveal. Emoticons cannot convey emotions that would be unintentionally reflected in facial expressions and body movements.

REFERENCES

Aakhus, M. (2001). Technocratic and design stances toward communication expertise: How GDSS facilitators understand their work. *Journal of Applied Communication Research, 29*, 341-371.

Ackermann, F., & Eden, C. (1994). Issues in computer and non-computer supported GDSSs. *Decision Support Systems, 12,* 381-390.

Ahalt, A. M. M. (2000). Remaking the courts and law firms of the nation: Industrial age to the information age. *Texas Tech Law Review, 31*, 1151-1165.

Ahuja, M. K., & Carley, K. M. (1998). Network structure in virtual organizations. *Journal of Computer-Mediated Communication, 3*(4). Retrieved August 4, 2002, from http://jcmc.huji.ac.il/vol3/issue4/ahuja.html

Aiken, M., & Carlisle, J. (1992). An automated idea consolidation tool for computer supported cooperative work. *Information & Management, 23,* 373-382.

Anson, R., Fellers, J., Kelly, G. G., & Bostrom, R. P. (1996). Facilitating research with group support systems. *Small Group Research, 27,* 179-214.

Bastress, R. M., & Harbaugh, J. D. (2003). Computer-mediated interviewing, counseling, and negotiating. *Clinical Law Review, 10,* 115-156.

Bialik, H. N. (2000). Revealment and concealment in language. In H. N. Bialik (Ed.), *Revealment and concealment: Five essays* (pp. 11-26). Jerusalem: Ibis Editions.

Bogard, W. (2000). Simmel in cyberspace: Strangeness and distance in postmodern communications. *Space and Culture, 4/5,* 23-46.

Crystal, D. (2001). *Language and the Internet.* Cambridge, UK: Cambridge University Press.

Daft, R. L., & Lengel, R. H. (1984). Information richness: A new approach to manager information processing and organization design. In B. Staw & L. L. Cummings (Eds.), *Research in organizational behavior* (pp. 191-233). Greenwich, CT: JAI Press.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural determinants. *Management Science, 32,* 554-571.

Davison, R., & Vogel, D. (2000). Group support systems in Hong Kong: An action research project. *Information Systems Journal, 10,* 3-20.

Easterbrook, S. (1995). *Coordination breakdowns: Why groupware is so difficult to design.* Retrieved July 29, 2002, from http://www.cs.toronto.edu/~sme/papers/1995/csrp343.pdf

Fjermestad, J. (2004). An analysis of communication mode in group support systems research. *Decision Support Systems, 37,* 239-263.

Friedman, T. L. (2000). *The Lexus and the olive tree*. New York: Anchor Books.

Froomkin, A. M. (1996). Regulation and computing and information technology: Flood control on the information ocean: Living with anonymity, digital cash, and distributed databases. *Journal of Law and Commerce, 15,* 395-507.

Giddens, A., & Duneier, M. (2000). *Introduction to sociology* (3rd ed.). New York: Norton.

Gottman, J. (1994). *What predicts divorce?: The relationship between marital processes and marital outcomes*. Hillsdale, NJ: Erlbaum.

Hopkins, K. (1998). Law firms, technology, and the double-billing dilemma. *Georgetown Journal of Legal Ethics, 12,* 95-106.

Hughes, J., & Lang, K. R. (2004). Issues in online focus groups: Lessons learned from an empirical study of peer-to-peer filesharing system users. *Electronic Journal of Business Research Methods, 2,* 95-110.

Kahai, S. S., & Cooper, R. B. (2003). Exploring the core concepts of media richness theory: The impact of cue multiplicity and feedback immediacy on decision quality. *Journal of Management Information Systems, 20,* 263-299.

Klein, E. E. (2000). The impact of information technology on leadership opportunities for women: The leveling of the playing field. *Journal of Leadership Studies, 7* (3), 88-98.

Klein, E. E., & Dologite, D. G. (2000). The role of computer support tools and gender composition in innovative information system idea generation by small groups. *Computers in Human Behavior, 16,* 111-139.

Kock, N. (1998). Can a leaner medium foster better group outcomes? A study of computer-supported process improvement groups. In M. Khosrowpour (Ed.), *Effective utilization and management of emerging information technologies* (pp. 22-31). Hershey, PA: Idea Group Publishing.

Kock, N. (1999). Can the adoption of a leaner medium increase group outcome quality? *Journal of Information Technology Impact, 1*(1), 13-20.

Majchrzak, A., Rice, R. E., King, N., Malhotra, A., & Ba, S. (2000). Computer-mediated inter-organizational knowledge-sharing: Insights from a virtual team innovating using a collaborative tool. *Information Resources Management Journal, 13*(1), 44-53.

Marshall, G. (Ed.). (1998). *A dictionary of sociology* (2nd ed.). Oxford, UK: Oxford University Press.

Martz, W. B., Jr., & Reddy, V. K. (2005). Looking for indicators of media richness theory in distance education. *Proceedings of the Thirty-Eighth Annual Hawaii International Conference on System Sciences.* Retrieved September 18, 2005, from http://csdl.computer.org/comp/proceedings/hicss/2005/2268/01/22680005c.pdf

McNeil, S. G., Robin, B. R., & Miller, R. M. (2000). Facilitating interaction, communication and collaboration in online courses. *Computers & Geosciences, 26,* 699-708.

Nagan, W. P., & Hammer, C. (2004). The changing character of sovereignty in international law and international relations. *Columbia Journal of Transnational Law, 43,* 141-187.

Nunamaker, J. F., Jr. (1997). Future research in group support systems: Needs, some questions and possible directions. *International Journal of Human-Computer Studies, 47,* 357-385.

Patterson, D. M. (1988). Wittgenstein and the code: A theory of good faith performance and enforcement under article nine. *University of Pennsylvania Law Review, 137,* 335-429.

Poole, M. S., & DeSanctis, G. (1990). Understanding the use of group decision support systems: The theory of adaptive structuration. In J. Fulk & C. W. Steinfeld (Eds.), *Organizations and communications technology* (pp. 173-193). Newbury Park, CA: Sage.

Reisman, W. M. (1983). The tormented conscience: Applying and appraising unauthorized coercion. *Emory Law Journal, 32,* 499-544.

Rosen, J. (2000). *The unwanted gaze: The destruction of privacy in America.* New York: Random House.

Schaefer, R. T. (2001). *Sociology* (7th ed.). Boston: McGraw-Hill.

Seul, J. R. (2004). Settling significant cases. *Washington Law Review, 79,* 881-968.

Simmel, G. (1964). Written communication. In K. H. Wolff (Ed. & Trans.), *The sociology of Georg Simmel* (pp. 352-356). New York: Free Press. (Original work published in 1908).

Simon, S. J., & Peppas, S. C. (2004). An examination of media richness theory in product Web site design: An empirical study. *Info, 6,* 270-281.

Sproull, L., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32,* 1492-1512.

Stenmark, D. (2002). Group cohesiveness and extrinsic motivation in virtual groups: Lessons from an action case study of electronic brainstorming. *Proceedings of the Thirty-Fifth Annual Hawaii International Conference on System Sciences.* Retrieved June 17, 2002, from http://dlib2.computer.org/conferen/hicss/1435/pdf/14350016b.pdf

Ullrick, B. A. (2000). The alternative billing diner: Serving up a new billing scheme for the technological age. *Journal of Technology Law & Policy, 5,* 1-50.

Ware, S. (2000, October 16). *Communication theory and the design of live online reference services.* Paper presented at the Second Annual Digital Reference Conference. Retrieved August 2, 2002, from http://www.vrd.org/conferences/VRD2000/proceedings/ware-intro.shtml

Warner, R. (2005). Surveillance and the self: Privacy, identity, and technology. *DePaul Law Review, 54,* 847-871.

West-Newman, C. L. (2005). Feeling for justice? Rights, laws, and cultural contexts. *Law and Social Inquiry, 30,* 305-335.

Yamauchi, Y., Yokozawa, M., Shinohara, T., & Ishida, T. (2000). Collaboration with lean media: How open-source software succeeds. *Proceedings of the ACM 2000 Conference on Computer Supported Cooperative Work (CSCW 2000)* (pp. 329-338).

Zigurs, I., & Buckland, B. K. (1998). A theory of task/technology fit and group support systems effectiveness. *Management Information Systems Quarterly, 22,* 313-334.

KEY TERMS

Computer-Mediated Communication (CMC): A communication system that involves or is assisted by computers. Computer-mediated communication includes group support systems, e-mail, videoconferencing, chat rooms, and instant messaging.

E-Collaboration: The collaboration of two or more individuals performing a specific task over a computer network. With the growth of GSS scholarship and its openness to interdisciplinary approaches, it is anticipated that future research will treat the textual ambiguity problem with the attention that it deserves and will find new approaches to place the GSS text within its context.

Face-to-Face Communication: Real-time communication between two or more individuals in physical proximity to each other.

Group Support Systems (GSS): Interactive computer-based information systems that support and structure group interaction and facilitate group meetings.

Lean Media: A vehicle of communication having few or no verbal cues, thereby precluding the clarification and interpretation of the spoken message.

Nonverbal Communication: The conveying of information, usually of an emotional nature, by means of body movement, facial expressions, hand gestures, and voice patterns.

Rich Media: A vehicle of communication having a multiplicity of nonverbal and verbal cues, which can be used to clarify and interpret the spoken message.

ENDNOTES

1 Computer-mediated communication (CMC) is a communication system that involves or is assisted by computers. Computer-mediated communication includes group support systems, e-mail, videoconferencing, chat rooms, and instant messaging.

2 This article is an expanded and revised version of an IRMA 2003 conference paper, which was published

in its proceedings under the title "Text-Based Group Support Systems: A Simmelian Perspective on E-Collaboration."

Virtual Communities of Practice

Chris Kimble
University of York, UK

Paul Hildreth
K-Now International Ltd., UK

INTRODUCTION

When knowledge management (KM) began to emerge in the 1990s it was seen as an innovative solution to the problems of managing knowledge in a competitive and increasingly internationalised business environment. However, in practice it was often little more than information management re-badged (Wilson, 2002). More recently, there has been recognition of the importance of more subtle, softer types of knowledge that need to be shared. This raises the question as to how this sort of knowledge might be managed. Communities of practice (CoPs) have been identified as means by which this type of knowledge can be nurtured, shared and sustained (Hildreth & Kimble, 2002). Do CoPs offer a means of managing the softer aspects of knowledge and, if they do, are they applicable to today's increasingly "virtual" world?

BACKGROUND TO COMMUNITIES OF PRACTICE

The term communities of practice (CoPs) was coined in 1991 when Jean Lave and Etienne Wenger used it in their exploration of situated learning (Lave & Wenger, 1991). Although the examples they used (non-drinking alcoholics, Goa tailors, quartermasters, butchers and Yucatan midwives) were all based on what might be broadly termed an apprenticeship model, the concept of a CoP is not restricted to this form of learning.

Lave and Wenger (1991) saw the acquisition of knowledge as a social process in which people participated in communal learning at different levels depending on their authority or seniority in the group, that is, whether they were a newcomer to the group or had been an active member for some time. The process by which a newcomer learns by being situated in the group was central to their notion of a CoP; they termed this process legitimate peripheral participation (LPP).

LPP is both complex and composite; legitimation, peripherality and participation are each indispensable in defining the other. Legitimation is concerned with power and authority relations in the community but is not necessarily formalised. Peripherality is not a physical concept or a measure of acquired knowledge, but concerned with the degree of engagement with the community. Participation is engagement in an activity where the participants have a shared understanding of what it means in their lives.

For Lave and Wenger (1991), the community and participation in it were inseparable from the practice. Being a member of a CoP implied participation in an activity where participants have a common understanding about what was being done and what it meant for their lives and their community. Thus, it would appear that CoPs with their concentration on situated learning and the exchange of understanding might be well suited to the management of the softer aspects of knowledge: but can this idea be applied to the business world?

EXTENSIONS TO THE COMMUNITY OF PRACTICE CONCEPT

Interest in CoPs continued to grow throughout the 1990s and several attempts were made to re-define Lave and Wenger's (1991) original model to encompass new areas such as communities of circumstance, communities of interest and communities of purpose. In particular, several attempts were made to re-define CoPs in a way that was more relevant to the commercial environment (e.g., Seely Brown & Duguid 1991, 1996; Stewart 1996). One of the most popular work related definitions of a CoP was offered by John Seely Brown and Estee Solomon Gray in their 1995 article called "The People Are the Company":

"At the simplest level, they are a small group of people … who've worked together over a period of time. Not a team not a task force not necessarily an authorised or identified group … they are peers in the execution of "real work". What holds them together is a common sense of purpose and a real need to know what each other knows" (Brown & Gray, 1995).

In 1998, Wenger (1998) published the results of an ethnographic study of a claims processing unit in a large

insurance company that described how employees exchanged knowledge during meetings and by the passing of handwritten notes. He proposed a view of the company not as a single community, but as a constellation of interrelated CoPs. CoPs arise out of the need to accomplish particular tasks and can provide learning avenues that exist within, between and outside organisations. CoPs are formed through mutual engagement in a joint enterprise and will share a repertoire of common resources (e.g., routines, procedures, artefacts, vocabulary) that members develop over time.

Thus, according to Wenger (1998) a CoP becomes defined in terms of:

- What it is about:

The particular area of activity/body of knowledge around which it has organized itself. It is a joint enterprise in as much as it is understood and continually renegotiated by its members.

- How it functions:

People become members of a CoP through shared practices; they are linked to each other through their involvement in certain common activities. It is this mutual engagement that binds its members together in a single social entity.

- What it produces:

The members of a CoP build up a "shared repertoire" of communal resources over time. Written files are a more explicit aspect of this, although less tangible aspects such as procedures, policies, rituals and idioms can also included.

Wenger (1998) also identified two key processes at work in CoPs: participation and reification. He described participation as:

"... the social experience of living in the world in terms of membership in social communities and active involvement in social enterprises" (Wenger, 1998, p. 55)

and reification as:

"... the process of giving form to our experience by producing objects that congeal this experience into thingness" (Wenger, 1998, p. 58)

Wenger emphasises that like LPP, participation and reification are analytically separable, but are inseparable in reality. Participation is the process through which people become active participants in the practice of a community and reification gives concrete form to the community's experience by producing artefacts. One is meaningless without the other and vice versa. In day-to-day work, people both negotiate meaning through participation in shared activities and project that meaning onto the external world through the production of artefacts.

Wenger's (1998) work with CoPs shows that the concept can be applied in a business setting. Since then, several other authors have identified the business benefits of CoPs (e.g., Fontaine & Millen, 2004: Lesser & Storck, 2001). However, almost all of the previous work on CoPs has described co-located communities. With the increasing globalisation of business and the heavy reliance on information and communication technology (ICT), the next question is "Can CoPs continue to operate in a modern business environment?"; that is, "Can a CoP be virtual?"

FUTURE TRENDS

Concerning the future of CoPs, and virtual CoPs in particular, two main issues must be considered. The first concerns the relationship between a CoP and its wider (electronic) environment; the second concerns the nature of the "work" that CoPs do; that is, do processes in a virtual CoP differ from one that is co-located?

CoPs in an Electronic Environment

Internet-based networking technologies, which can provide a single platform for groups or networks of groups to form within larger organisations, have led to the development of various forms of virtual groups and communities. Seely Brown and Duguid (2000) coined the phrase "networks of practice" (NoPs) to describe one type of virtual group. NoPs are composed of people who are geographically separate and may never even get to know each other, but who share similar work or interests. Thus, NoPs are organised more at the individual level and based on personal social networks than CoPs with their notions of mutuality and the collective social will of the community.

In a study of job seeking activity, Granovetter (1973) introduced the notion of strong and weak social ties. In terms of the previous description, CoPs are characterised by strong social ties, whereas NoPs are characterised by weak social ties. Within a wider network consisting of weak ties, an individual may act as a "local bridge" or broker that enables the network to react more quickly and provide a coordinated response. Nevertheless, within a network there is also a need for strong ties to encourage local cohesion and avoid fragmentation that would make knowledge sharing and the adoption of innovation more difficult.

CoPs can be seen in the role of hub for the wider network, providing a more tightly knit sub-network that serves as knowledge generating centres for the larger NoPs. CoPs can act as bridges drawing together different groups and

combining knowledge in new ways. They can also provide the access points for individuals to engage with the wider network and to establish a local identity within the larger organisation. Previous research has shown that the most common distributed form of a "virtual" CoP has a co-located active core (Hildreth, Kimble & Wright, 1998), which tends to support this view of distributed working.

A more recent example was provided by Lundkvist's (2004) study of customer networks as sources of innovation. This case study was generated from a long-term study of the Cisco Systems newsgroup, which identified user networks as peripheral and yet vital sites of innovation. In this case, the co-located core of the network was provided by a group of university technicians.

Work in Virtual CoPs

How might the balance between reification and participation be maintained in virtual working? This issue was addressed in an earlier paper (Kimble, Hildreth & Wright, 2000) where we described how a geographically distributed CoP managed both hard (reified) and soft (social) knowledge. In this case, the CoP was made up of four members co-located in the UK, a group of five members in the USA and one member in Japan.

In this situation, it might have been expected that sustaining participation would be more difficult and therefore reification would play a greater role. However, the findings of the case study showed that this was not necessarily the case. Shared artefacts, such as a planning document, did play an important role but the importance of social relationships remained paramount. While the group was able to sustain itself using e-media, it was still dependent on the development of relationships in the physical environment through face-to-face meetings.

It is interesting to observe how artefacts such as a planning document (reification) were used not only as ways of projecting knowledge from within the CoP but were also instrumental in the process of creating it (participation). The document stimulated discussion, problem solving, innovation and further participation. It was used both to drive meetings and as the focus of meetings. During discussions based on the document, new and innovative ideas would be triggered that could form the basis for new projects. Thus, as well as acting as a stimulus for innovation, the document acted as a catalyst leading to further participation.

A similar account can be found in Bradshaw, Powell and Terrell (2004) that describes how a team of remote workers developed into a CoP. They describe not only how the group deploys a variety of technologies to maintain contact but also the efforts that went into building commitment, ownership, engagement and focus in the group. In this case, the members of the group were all engaged in collaborative research. Writing about their work and presenting papers for peer- review was seen as a key factor in maintaining cohesion and developing the community's shared understanding of goals, development of knowledge and sense of belonging.

CONCLUSION

Reporting on a recent case study of how CoPs translate to a geographically distributed international environment, Hildreth (2003) throws further light on a number of these issues. The study examines the work of an internationally distributed CoP that spans three continents. In particular, it highlights the role that shared artefacts play in the process of creating, sharing and sustaining both types of knowledge and highlights the role that the creation of artefacts plays in enabling and sustaining participation in CoPs. Hildreth (2003) observes that the process of creating the artefact and regular (although not necessarily frequent) face-to-face contact are instrumental in maintaining the relationships that allow a CoP to function successfully in a virtual environment. Thus, paradoxically, it appears that one of the keys to a successful virtual CoP is an occasional, non-virtual, face-to-face meeting.

However, the changes that are sweeping the corporate infrastructure mean that increasingly workers find themselves forced into one or another form of virtual working. Instead of inhabiting a world of fixed roles with easy access to co-located resources, today's workers are increasingly based in a world of weak ties where resources are only obtained through personal and individual relationships. Rather than being embraced by a collective CoP, workers often find themselves functioning as individuals and building up networks, one contact at a time. Again, paradoxically, as social networks such as NoPs become more important, the fundamental unit for many examples of virtual working is not the group but the individual. This is not to say that collective groups such as CoPs and teams have ceased to exist but simply that the difficulty of building and maintaining the strong social ties needed to build a sense of community in a virtual environment should not be underestimated.

REFERENCES

Bradshaw, P., Powell, S., & Terrell, I. (2004). Building a community of practice: Technological and social implications for a distributed team. In P. Hildreth & C. Kimble (Eds), *Knowledge networks: Innovation through communities of practice* (pp. 184-201). Hershey, PA: Idea Group Publishing.

Fontaine, M.A., & Millen, D.R. (2004). Understanding the benefits and impact of communities of practice. In P. Hildreth & C. Kimble (Eds), *Knowledge networks: Innovation through communities of practice* (pp. 1-13). Hershey, PA:

Idea Group Publishing.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology, 78*(6), 1360-1380.

Hildreth, P. (2003). *Going virtual: Distributed communities of practice.* Hershey, PA: Idea Group Publishing.

Hildreth, P., & Kimble, C. (2002). The duality of knowledge. *Information Research, 8*(1). Paper no. 142. *http://InformationR.net/ir/8-1/paper142.html*

Hildreth, P., Kimble, C., & Wright, P. (1998, March). Computer mediated communications and communities of practice. *Proceedings of Ethicomp'98,* Erasmus University, the Netherlands (pp. 275 – 286).

Kimble, C., Hildreth, P., & Wright, P. (2000). Communities of practice: Going virtual. In Y. Malhotra (Ed.), *Knowledge management and business model innovation* (pp. 220-234). Hershey, PA: Idea Group Publishing.

Lave, J., & Wenger, E. (1991). *Situated learning. Legitimate peripheral participation.* Cambridge: Cambridge University Press.

Lesser, E.L., & Storck, J. (2001). Communities of practice and organizational performance. *IBM Systems Journal, 40*(4), 831 – 841. Retrieved from *http://researchweb.watson.ibm.com/journal/sj/404/lesser.pdf*

Lundkvist, A. (2004). User networks as sources of innovation. In P. Hildreth & C. Kimble (Eds.), *Knowledge networks: Innovation through communities of practice* (pp. 96-105). Hershey, PA: Idea Group Publishing.

Seely Brown, J., & Duguid, P. (1991). Organizational learning and communities of practice. *Organization Science, 2*(1), 40-57.

Seely Brown, J., & Duguid, P. (1996). Universities in the digital age. *Change*, 11-19.

Seely Brown, J., & Duguid, P. (2000). *The social life of information.* Boston, MA: Harvard Business School Press.

Seely Brown, J., & Solomon Gray, E. (1995). The people are the company. *Fast Company.* Retrieved February 10, 2004, from *http://www.fastcompany.com/online/01/people.html*

Stewart, T.A. (1996). The invisible key to success. *Fortune Online*. Retrieved October 4, 1996, from *http://pathfnder.com/@@V3AagAUAZyqOEYKS/fortune/magazine/1996.960805/edg.html*

Wenger, E. (1998). *Communities of practice. Learning, meaning and identity.* CUP.

Wilson, T.D. (2002). The nonsense of 'knowledge management'. *Information Research, 8*(1). Paper no. 144. Retrieved from *http://InformationR.net/ir/8-1/paper144.html*

KEY TERMS

Artefact: An artefact in the context of CoPs indicates objects, articles, and "things" which have been created by the CoP to assist the members in their work and which may have some of the community's knowledge embedded in them. Artefacts do not have to be concrete – a process or procedure may be an artefact.

Communities of Circumstance: Communities of circumstance are driven by position, circumstance or life experiences. Communities of circumstance are distinguished from CoPs in that they tend to be personally focused and are often built around "life stages," such as teenagehood, university, marriage or parenthood.

Communities of Interest: Communities of interest are groups of people who share a common interest. Members exchange ideas and thoughts about the given interest, but may know little about each other outside of this area. Participation in a community of interest can be compelling and entertaining but is not focussed on learning in the same way as a CoP.

Communities of Practice: Communities of practice are groups of people who have a common goal and who are internally motivated to reach the goal. The members have some form of common background and shared language.

Communities of Purpose: Communities of purpose form around people who are to achieve a similar objective. Such communities only serve a functional purpose. Members of the community can assist each other by sharing experiences, suggesting strategies and exchanging information on the process in hand.

Hard Knowledge: Hard knowledge is unambiguous and unequivocal, can be clearly and fully expressed, can be formalised and structured, can be "owned" without being used and is both abstract and static: it is about, but not in, the world.

Knowledge Management: Knowledge management is the means whereby an organisation "manages" and leverages its knowledge resources. This can include reports, databases and patents; it also includes people – identifying experts, sharing knowledge, and helping people learn.

Legitimate Peripheral Participation: LPP is the process by which a newcomer gradually works his/her way towards full participation in the community. Lave and Wenger's (1991)

examples were based on the apprenticeship model, where a newcomer (the apprentice) was allowed to undertake basic tasks. As they became more experienced, they were given more complicated tasks until they could fully participate in the practice of the community and became old-timers.

Network of Practice: People who are not directly connected to each other but still engage in similar kinds of activities are said to belong to a network of practice (NoP). NoPs link local communities whose members have similar interests and give a minimal coherence to the network.

Soft Knowledge: Soft knowledge is implicit and unstructured, cannot be articulated, can be understood without being openly expressed, is associated with action and cannot be possessed, is about what we do and is acquired through experience.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 2991-2995, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Virtual Communities of Practice for Health Care Professionals

Elizabeth Hanlis
Ehanlis Inc., Canada

Jill Curley
Dalhousie University, Canada

Paul Abbass
Merck Frosst Canada Limited, Canada

INTRODUCTION

Wenger is typically credited with the development of the metaphor of communities of practice where "learning requires an atmosphere of openness and the key is to build an atmosphere of collective inquiry" (Wenger, 1998). However, the focus of creating a sense of belonging as well as the formulation of knowledge as a social process is not as new. Rather, it can be found in the form of a learning community. Senge (1990) introduced this concept of the learning organization to explain strategies to enhance the capacity of members to consistently collaborate on mutual goals.

With the increased use of the Internet over the past decade, health professionals are examining how to effectively use this medium to support collaboration and learning, while improving patient care (Casebeer, Bennett, Kristofco, Carillo, and Center (2002).

It is therefore understandable that Continuing Medical Education (CME) on the Internet has grown exponentially over the last several years. Curran and Fleet (2005) describe that in order for physicians to be able to adapt to the demands and changes of an ever-evolving technical world, they must think about "the new dimension and innovative opportunities that the Internet affords for doctors to access CME in the 21st century" (Curran & Fleet, 2005). Online professional learning opportunities offer more flexibility than traditional face-to-face CME and are able to overcome barriers to learning like travel and irregular work hours. As a result, the Communities of Practice concept was taken to the Web and the term Virtual Communities of Practice (VCoP) appeared in the medical literature (Dube, Bourhis, & Jacob, 2006).

The first virtual communities in medicine involved patients, and focused on providing a space for mutual support, along with news on innovative treatments and helpful resources (Demiris, 2006; Nagy et al., 2006).

Virtual communities for professionals, or virtual communities of practice (VCoP), were recently generated based on similar needs (Nagy et al., 2006). According to Bates and Robert (2002), VCoP are vitally important for health care professionals and organizations, as they spread best practices and change practice (as cited in Sandars & Heller, 2006).

Unfortunately, there is limited literature that examines VCoP for health care professionals (Moule, 2006). This article will review existing literature to determine the requirements for establishing and maintaining an effective VCoP within the health care context, in support of continuing professional development. Specifically, the article will focus on the benefits of a VCoPs, the characteristics of successful VCoPs and examples of existing VCoPs with a focus on health professionals.

BACKGROUND

Communities of Practice (CoP) are groups "of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an on-going basis" (Wenger, McDermott, & Snyder, 2002, p. 4). Generally, such communities seem to be an innovative way to share and manage knowledge and sustain innovation (Wenger et al., 2002).

Virtual Communities of Practice (VCoP), without excluding face-to-face meetings, rely primarily on information and communication technologies (ICT) to connect their members. A VCoP may use a large array of traditional media (phone teleconference, fax, etc.) and more or less sophisticated technological tools, such as e-mail, videoconference, newsgroups, online meeting space, or a Website Intranet to establish a common virtual collaborative space (Demiris, 2006; Dube et al., 2006).

Virtual communities in health care refer to the group of people (and the social structure they create), who communicate via ICT for the purpose of collectively conducting activities related to health care and education. Such activities may include: discussions around problems, cases, best practices, management of diseases, or treatments, collaboration around patient care or research projects, sharing of docu-

ments and resources on topics of interest, consulting with experts, or generating new ideas and innovation (Demiris, 2006; Endslay, Kirkegaar, & Linares, 2005).

Dimensions of a Community of Practice

The following three dimensions are essential to a community of practice: 1) mutual engagement, 2) joint enterprise, and 3) shared repertoire (Wenger et al., 2002).

- **Mutual engagement** involves regular interaction among participants within the community, including both informal communication (i.e., e-mail), or more formal structured communication (i.e., monthly Web meetings) (Wenger et al., 2002).
- **Joint enterprise** refers to the process that maintains the community. This includes negotiating the endeavors of the community (Wenger et al., 2002).
- **Shared repertoire** includes the ways, routines, and even language developed by the community (Wenger et al., 2002). Shared repertoire implies longevity, as such successful communities cannot flourish in a few months (Moule, 2006).

Benefits of a VCoP for Health Care Professionals

"Practitioners reflect on and learn from their practice in ways that incorporate both tacit (implicit) and explicit (codified) knowledge" (Doak & Assimakopoulos, 2007; Rynes & Bartunek, 2001 (as cited in Bartunek, Trullen, Bonet, & Sauquet, 2003)). While explicit knowledge can be found in books, journal articles, or other formal learning events, tacit (implicit) knowledge comprises a range of conceptual and sensory information and images that are difficult to articulate in words, but rather can be demonstrated or imitated (Polanyi, 1967). Tacit knowledge may therefore include intuition, perspectives, beliefs, values, and culture. As such, tacit knowledge can only be gained through individual experience and by collective participation in communities of practice (Bartunek et al., 2003). Some research indicates that tacit knowledge is better diffused within an organization whose structure and work environment promote face-to-face interaction and employee sharing at close physical proximity (Busch, 2006). Nevertheless, tacit knowledge can also be shared via a VCoP through metaphors, analogies, and stories of practice, a form of knowledge transmission that builds on contextual cues (Bartunek et al., 2003).

One of the major benefits of a VCoP is the ability of a diverse group of health care professionals to communicate and collaborate quickly across institutions and geographical locations (Demiris, 2006; Endslay et al., 2005; Robinson & Cottrell, 2005).

This exchange of knowledge within a VCoP leads to the creation of new knowledge and change in practice (Robinson & Cottrell, 2005). Nonmedical literature further supports that communities of practice can result in increased productivity and innovation (Sandars & Heller, 2006).

Characteristics of a Successful Virtual Community for Health Professionals

The literature around virtual Communities of Practice for Health Professionals was examined to determine common traits and characteristics of successful communities. As such, the following list of characteristics was developed.

Community Coordinator or Moderator

According to the literature, a high degree of structured management is critical for the success of a virtual community of practice. One way to achieve this is through online moderating of the group (Salmon, 2000) (as cited in Sandars & Heller, 2006).

The community coordinator helps the community identify important issues, focus on relevant topics, develop and maintain relationships, and develop its practice, including lessons learned and best practices. They may not be leading experts in their field as their role is not to "give all the answers" but to link people and guide members to appropriate resources (Endslay et al., 2005; Wenger et al., 2002)

Active Participants and Lurkers

Regardless of the size of the community, in order for it to be successful it needs to have both active participants and "lurkers." Active participants will ensure that there is regular interaction among members, making the community vibrant and energetic. These members provide intellectual and social leadership (Dube et al., 2006). Lurkers are members of the community who do not contribute regularly. However, lurkers may constitute up to two thirds of their community and their knowledge and resources are still important (Endslay et al., 2005; Wenger et al., 2002).

Trust and Opportunity for Socialization

Virtual communities of practice can be as effective as groups that have face-to-face meetings, provided there is the development of trust (Hildreth, Kimble, & Wright, 2000 (as cited in Sandars & Heller, 2006, p. 343)).

However, when participants cross boundaries and are from different professions and organizations, it is difficult to develop a level of trust and to buy into the idea of knowledge sharing (Wenger et al., 2002). In such cases, more effort needs to be made to break organizational silos

and promote sharing and collaboration, which takes time (Dube et al., 2006).

Building trust is also difficult when there is a wide geographical dispersion across a country, as physical distance encourages psychological distance in many cases (Dube et al., 2006; Preece, 2000). One effective strategy for building trust within virtual communities is to provide opportunities for socialization (either face-to-face or online) when launching the community (Sandars & Heller, 2006).

Clear Purpose or Goal of the Community

A VCoP can be used to support a variety of goals and endeavors. These goals and the process used to support them need to be negotiated and agreed upon by its members (Moule, 2006).

Communities of practice for health care professionals may serve one or more of the following purposes: actual delivery of health care services, staff education discussing health- and treatment-related issues and problems, sharing of documents, sustaining relationships beyond face-to-face meetings and conferences, and supporting virtual care delivery teams or research teams (Demiris, 2006).

Community Activity and Events

Regular activity and events organized for the virtual community of practice assist in "anchoring" the community at the onset, and maintaining the energy and enthusiasm of the group in the long term (Nagy et al., 2006; Wenger et al., 2002). Both structured and less structured activities are encouraged. This could range from regularly scheduled meetings, via Web conferences, teleconferences, or videoconferences, to recommending members to attend medical conferences and share lessons learned within the VCoP (Endslay et al., 2005).

Inclusion of Face-to-Face Interaction

While a VCoP relies primarily on the use of ICT, the degree of reliance on it may vary greatly. For one VCoP there may be one face-to-face meeting a year at a conference, and the remaining communication takes place via ICT. For another VCoP there may be monthly face-to-face meetings, while the remaining communication takes place online. It is widely accepted that ICT will never be a perfect substitute for face-to-face interaction and most VCoPs need some face-to-face time to be most effective (Deloitte Research, 2005).

Credibility

A VCoP needs to prove its credibility to its members. This may be accomplished by ensuring that the site remains unbiased and transparent. Some level of expert oversight also increases credibility (Nagy et al., 2006).

Ownership of Site

Participants need to be encouraged to own elements of the site. This helps to increase a feeling of ownership and belonging of participants (Nagy et al., 2006).

Educational Content

In order for a VCoP to remain vital, organizers must provide online access to a variety of educational resources and content (Nagy et al., 2006). Such resources could be interesting or problematic patient cases, Websites, or articles that are freely available on the Internet. All members should also have the opportunity to share, upload, and review resources on the site in order for a just-in-time knowledge-base to be developed (Nagy et al., 2006).

Technology and Usability

There are many communication and learning technologies that could be used to support a VCoP. Possible technologies include: e-mail, asynchronous discussion forums (or Bulletin Board), Listservs, text chat, Web conferencing, teleconferencing, wikis, blogs, and course/learning management systems.

It is important that the technology supports the essential requirements of the community, including online facilitation, and storage of tacit knowledge into explicit knowledge (Sandars & Heller, 2006). Usually a combination of tools and approaches are appropriate to meet the needs of the learner.

Regardless of the approach used, the technology should be evaluated on its ability to enhance the natural information-seeking behaviors of individual practitioners and its usability (Parboosingh, 2002). According to Preece (2000), a virtual community of practice that is usable allows members to communicate with each other, find information, and navigate the community site with ease (Demiris, 2006).

Leadership

A strong community of practice requires strong leadership. The health care organizations and its individual members need encouragement, guidance and influence to help them accept the value of communities of practice in implementing best practices and motivating change of practice (Sandars & Heller, 2006).

It is advisable to create a formal leadership structure when launching a CoP, rather then letting leadership roles and authority structures emerge on their own over time (Dube et al., 2006).

Table 1. Examples of existing VCoP for health care professionals

VCoP – Name and URL	Brief Description
Braintalk.org (*www.braintalk.org*)	An online patient support group with almost 50,000 members. ("Designing virtual communities for medical professionals",2006).
ClubPACS (http://www.clubpacs.com/)	A virtual community for picture archiving and communication system (PACS) administrators with 2,500 members (Nagy et al., 2006).
CoPosis™ (www.coposis.ca)	CoPosis™ is a Canadian online Community of Practice (CoP) for health care professionals interested in advancing discussions and understanding of issues surrounding the identification and management of fungal infections.
Emergency Care Community of Practice (http://www.nhmrc.gov.au/nics/asp/index.asp?cid=5263&gid=207&page=programs/programs_article)	The National Institute of Clinical Studies in Australia has established an emergency care CoP for health care professionals.
Health Infomatics NHS Community (http://www.informatics.nhs.uk/index.html)	Developed for UK health care professionals to support the delivery of health care and promote health (Sandars & Heller, 2006).
Healthboards.com (*www.healthboards.com)*	A collection of more than 140 discussion boards pertaining to virtually every major health topic ("Designing virtual communities for medical professionals", 2006).
Pain-Talk (*www.pain-talk.co.uk/*)	A vibrant, national community for UK health care professionals with an interest in acute, chronic, or palliative pain management ("Designing virtual communities for medical professionals", 2006).

Institutional Commitment and Recognition

Ensuring that a VCoP is recognized and supported by an institution is critical to its success (Dube et al., 2006). Institutions need to recognize participation in the VCoP as continuing professional development and allow time for members to participate (Sandars & Heller, 2006). Such recognition provides credibility to the community, which will encourage participation (Dube et al., 2006).

FUTURE TRENDS

There seems to be an abundance of literature around communities of practice in general, since the inception of the term in 1998. While VCoP share many characteristics with CoP, they have a set of distinct characteristics (Dube et al., 2006). Nevertheless, there seems to be limited literature around VCoP and even fewer studies examining the effectiveness of VCoP for health care professionals (Moule, 2006).

In order to develop evidence-based guidelines for the design and maintenance of VCoP for health care professionals, it is necessary to conduct research beyond exploratory pilot studies to clinical trials and interventions that follow an experimental design (Demiris, 2006).

To increase the broad application of findings, researchers need to document the social context, recruitment strategies, the implied and stated rules that govern the community, its utilization rate, its rate of growth, leadership roles, funding strategy, and strategies used to encourage participation (Demiris, 2006).

Most ICTs that specifically focus on workplace learning activities, such as knowledge sharing, knowledge creation and

facilitation of collaboration are recent, and studies on their effectiveness to enhance learning have not been done yet. A research agenda must be planned to assess the effectiveness of information and communication technologies that support and enhance learning in practice (Parboosingh, 2002). This research agenda needs to include a pre-post design to determine the impact of the VCoP on knowledge and skills, practice, and patient care (Demiris, 2006).

CONCLUSION

With limited time and increased costs of travel, interacting face-to-face is becoming increasingly difficult. These limitations along with the increased use of technology and the need to interact and share implicit knowledge, has led to the creation of Virtual Communities of Practice for Health Professionals.

When "a specialized profession reaches a critical mass, a virtual community of shared knowledge can become a self-sustaining and beneficial resource that grows and develops along with the needs of its participants. Such a community can be highly beneficial for the discipline (Nagy et al., 2006, p. 5).

While this article is a first attempt at determining the characteristics of successful VCoP for health care professionals, more research is needed to ensure that we are using the right technologies to support VCoP and appropriate processes to design, develop, and maintain such communities in a manner that positively influence patient outcomes and patient care.

REFERENCES

Bartunek, J., Trullen, J., Bonet, E., & Sauquet, A. (2003). Sharing and expanding academic and practitioner knowledge in health care. *Journal of Health Services & Research Policy, 8*(2), 62-68.

Bates, S. P., & Robert, G. (2002). Knowledge management and communities of practice in the private sector: Lessons for modernizing the national health service in England and Wales. *Public Adminisration, 80*, 642-663.

Busch, P. (2006). Organisation design and tacit knowledge transferal: An examination of three IT firms. *Journal of Knowledge Management Practice, 7*(2).

Casebeer, L., Bennett, N., Kristofco, R., Carillo, A., & Centor, R. (2002). Physician Internet medical information seeking and online continuing education use patterns. *The Journal of Continuing Education in the Health Professions, 22*, 33-42.

Curran, V. R., & Fleet, L. (2005). A review of evaluation outcomes of Web-based continuing medical education. *Medical Education, 39*, 561-567.

Deloitte Research. (2001). *Collaborative knowledge networks: Driving workforce performance through Web-enabled communities*. Retrieved May 31, 2008, from http://www.affinitiz.net/enterprise/deloitte_knowledge_networks.pdf

Demiris, G. (2006). The diffusion of virtual communities in health care: Concepts and challenges. *Patient Education & Counseling, 62*(2), 178-188.

Designing virtual communities for medical professionals. (2006). *Journal of Visual Communication in Medicine, 29*(3), 130-131.

Doak, S., & Assimakopoulos, D. (2007). How do forensic scientists learn to become competent in casework reporting in practice: A theoretical and empirical approach. *Forensic Science International, 167*(2-3), 201-206.

Dube, L., Bourhis, A., & Jacob, R. (2006). Towards a typology of virtual communities of practice. *Interdisciplinary Journal of Information, Knowledge and Management, 1, 69-93.*

Endslay, S., Kirkegaar, M., & Linares, A. (2005). Working together: Communities of practice in family medicine. *Family Practice Management, 12*(1), 28-32.

Hildreth, P. M., Kimble, C., & Wright, P. (2000). Communities of practice in the distributed international environment. *Journal of Knowledge Management, 4*, 27-10.

Moule, P. (2006). E-learning for health care students: Developing the communities of practice framework. *Journal of advanced nursing, 54*(3), 370-380.

Nagy, P., Kahn, C.E., Jr., Boonn, W., Siddiqui, K., Meenan, C., Knight, N., et al. (2006). Building virtual communities of practice. *Journal of the American College of Radiology, 3*(9), 716-720.

Parboosingh, J.T. (2002). Physician communities of practice: Where learning and practice are inseparable. *The Journal of Continuing Education in the Health Professions, 22, 230-236.*

Preece, J. (2000). *Online communities: Designing usability and supporting sociability*. New York: John Wiley & Sons.

Polanyi, M. (1967). *The tacit dimension.* New York: Anchor Books.

Robinson, M., & Cottrell, D. (2005). Health professionals in multi-disciplinary and multi-agency teams: Changing professional practice. *Journal of Interprofessional Care, 19*(6), 547-560.

Rynes, S.L., & Bartunek, J.M. (2001). Across the great divide: Knowledge creation and transfer between practitioners and academics. *Academy of Management Journal, 44*, 340-356.

Salmon, G. (2000). *E-moderating the key to teaching and learning on-line*. London: Kogan Page.

Sandars, J., & Heller, R. (2006). Improving the implementation of evidence-based practice: A knowledge management perspective. *Journal of evaluation in clinical practice, 12*(3), 341-346.

Senge, P., Ross, R., Smith, B., et al. (1994). The fifth discipline fieldbook: Strategies and tools for building a learning organization. New York: DoubleDay.

Stuckey, B., Hedberg, J., & Lockyer, L. (2001). *Building on-line community for professional development.*

Wenger, E. (1998). *Communities of practice*. Cambridge: Cambridge University Press.

Wenger, E., McDermott, R.M., & Snyder, W.M. (2002). *Cultivating communities of practice*.

KEY TERMS

Active Participant: An actively engaged in a virtual community. Such participants display interactive behavior which includes collaborative or positive interactive behavior or even hostile behavior (Demiris, 2006).

Asynchronous Discussion Forums (or Bulletin Board): Web-based conversations that allow users to post, view, and reply to messages (Kevin, 2006). The benefit of such a tool is that the information posted by the CoP is archived for later viewing and in most systems is key-word searchable (Nagy et al., 2006).

Communities of Practice (CoP): Groups "of people who share a concern, a set of problems, or a passion about a topic, and who deepen their knowledge and expertise in this area by interacting on an ongoing basis" (Wenger et al., 2002, p. 4).

Explicit Knowledge (or codified knowledge): A transmittable in formal systematic language expressed in symbols, words, or numbers (Rynes & Bartunek, 2001). This is the type of knowledge or information that we can easily document and obtain from books and journals.

Lurkers: Members of the community who do not contribute regularly (Endslay et al., 2005). Lurkers limit their participation to passively observing, rather than actively participating in community discussions (Demiris, 2006).

Tacit Knowledge: Personal, context-specific knowledge, rooted in action and experience that is difficult to formalize and communicate (Rynes & Bartunek, 2001). It is knowledge that is shared via stories, anecdotes, metaphors, personal reflections, and communication (Sandars & Heller, 2006).

"Virtual Communities of Practice (VCoP): Rely primarily on ICT to connect members of a CoP and may use a large array of traditional media (phone teleconference, fax, etc.) and more or less sophisticated technological tools to establish a common virtual collaborative space (Demiris, 2006; Dube et al., 2006).

Category: Electronic Business

Virtual Corporations

Sixto Jesús Arjonilla-Domínguez
Freescale Semiconductor, Inc., Spain

José Aurelio Medina-Garrido
Cadiz University, Spain

INTRODUCTION

At the end of the 20th century, many authors tried to predict what new structures companies would be likely to adopt in the 21st century. Now, in the 21st century a clear tendency is emerging: the virtual organization (Agrawal & Hurriyet, 2004; Alsop, 2003; Bekkers, 2003; Camarinha-Matos & Afsarmanesh, 2005; Heneman & Greenberger, 2002; Lee, Cheung, Lau, & Choy, 2003; Talukder, 2003; Vakola & Wilson, 2004). This type of organization offers the most promising response to an increasingly complex business reality. In this respect, current organization theory is beginning to change its focus to new, flexible, and virtual organizational forms.

This article is organized as follows: The background section defines different concepts of virtual organization. The first model equates the virtual corporation to a temporary network of firms that quickly comes together to exploit temporary market opportunities. The second model focuses on the manufacture of virtual products by means of stable and trusting relationships with suppliers and customers. The third model of virtual corporation tries to turn the fixed workforce costs into variable costs. The third section points out the shared characteristics of this type of organization and the role of the manufacturing function, information and information technology, the network structure, and a new type of worker. The final sections discuss future trends and our conclusions.

BACKGROUND

The term *virtual corporation* was coined by Jan Hopland, an executive at Digital Equipment Corporation at the end of the 1980s, to describe firms that can marshal more resources than they actually possess by means of both internal and external collaborations (Fitzpatrick & Burke, 2001; Weisenfeld, Fisscher, Pearson, & Brockhoff, 2001). The term can be traced to the computing concept of *virtual memory*, which describes how a computer can behave as if it had much more processing power than it really has.

The expression virtual corporation has been used in the literature to refer to concepts ranging from simply using teleworking and outsourcing intensively (Buhman, 2003; Coates, 2005; Matthews, 2004) to the wholesale restructuring of the firm. However, three approaches predominate in the literature: (1) virtual corporation as a temporary network of firms that rapidly forms to exploit temporary opportunities appearing in the market (Alsop, 2003; Beckett, 2003; Chalmeta & Grangel, 2003); (2) virtual corporation as a firm that produces virtual products, and which develops strong and stable links with its suppliers and customers (Biondi, Bonfatti, Monari, Giannini, & Monti, 2003; Lee et al., 2003; Mo & Zhou, 2003); and (3) a final model which considers that the virtual firm is an organization whose costs are essentially variable, only being generated when the firm is sure that it will recover them by selling the product or service (Matthews, 2004; Talukder, 2003).

The virtual corporation can also be defined by what it is not. The virtual corporation is not a takeover or merger between firms, nor is it a temporary employment agency, nor a "hollow" firm seeking to cut costs by closing down factories in one country and opening them up again in another one with lower labor costs.

CHARACTERISTICS OF VIRTUAL CORPORATIONS

As we mentioned in the previous section, there are three different perspectives of the concept of virtual corporation. Among the characteristics shared by these three models of virtual corporation, we would stress: excellence, technology, trust, opportunism, and absence of borders:

- **Excellence:** Since each member contributes its core competencies (Mo & Zhou, 2003), it is possible to create an organization with the best of each of them, so that every process of the virtual corporation can be the best in its class, something that no one firm could achieve on its own.
- **Technology:** Information technology—and more specifically, communications networks—will facilitate the transfer of knowledge and technologies between firms and will allow firms and workers to work together (Helling, Blim, & O'Regan, 2005; Heneman

& Greenberger, 2002; Im, Yates, & Orlikowski, 2005; Kovacs & Paganelli, 2003).

- **Trust:** This type of relationship makes firms more dependent on others, and hence requires a much greater trust than would normally be the case between firms that are just partners (Camarinha-Matos & Afsarmanesh, 2003; Clases, Bachmann, & Wehner, 2003; Gallie & Guichard, 2005).
- **Opportunism:** In the first and third model, firms will come together to exploit a very specific market opportunity, disbanding as soon as the opportunity disappears. In the second model, the opportunity to form a virtual corporation is brief. Once a firm adopts a virtual corporation structure, other firms have less chance of doing the same.
- **Absence of Organizational Borders:** The close cooperative ties established between competitors, suppliers, and customers will make it difficult to see where one firm ends and another begins (Lee et al., 2003). This organizational structure shares the knowledge and resources needed to carry out the work to be done, regardless of which firm owns or manages them.

In virtual corporations the role of the manufacturing function also changes (Martinez, Fouletier, Park, & Favrel, 2001). This type of organizational structure seeks to generate high-quality products and services rapidly in response to the demand (Offodile & Abdel-Malek, 2002; Weisenfeld et al., 2001). Generating virtual products, also known as mass-customized products (Biondi et al., 2003), requires both the customers' participation in their conception and design and the firms' implementation of time-based strategies. This implies combining the customers and suppliers in a type of highly efficient network (Lee et al., 2003), which will require information systems that support relationships at all levels (Hsieh, Lin, & Chiu 2002). Companies will focus on one, maybe two, or three core competencies, all else will be outsourced (Buhman, 2003; Erickson, 2004; Mo & Zhou, 2003; Porter, 2000). Time is regarded as a component that, like any other, can be improved by means of intelligent planning and the use of technology (Hao, Shen, & Wang, 2005; Lee et al., 2003).

Mass customization of products combines the effects of lean manufacturing processes (Guisinger & Ghorashi, 2004), zero-inventory production methods, or just-in-time, and total quality management processes. These management processes make it possible to produce the great variety of products that could once only be made by craft manufacturing, but often at a lower cost than mass production, and with excellent quality.

Moreover, incorporating the new information technology into the manufacturing processes has given rise to *flexible manufacturing systems* (FMS) and *computer-integrated manufacturing* (CIM) (Presley, Sarkis, Barnett, & Liles, 2001). Such systems help manufacturers to achieve many of the objectives of the virtual corporation, such as shorter production cycles, a smaller and more qualified labor force, smaller batches, flexibility, a better short-term response, and long-term adaptability.

In virtual corporations, the role of information and information technology is also important (Aerts, Szirbik, & Goossenaerts, 2002; Alsop, 2003; Gallie & Guichard, 2005; Heneman & Greenberger, 2002; Joukhadar & Binstock, 2000; Khalil & Wang, 2002; Kovacs & Paganelli, 2003; Stowell, 2005; Xu, Wei, & Fan, 2002). This role becomes clear in the three models described previously. The virtual corporation understood as a network of firms is also an information system where there is an exchange of the data generated by the activities or processes (internal and external) that it carries out. The success of the virtual corporation will depend on its capacity to acquire, distribute, store, analyze, and integrate this massive flow of information through its organizational elements, supported by information technology.

The model of virtual corporation defined as a generator of virtual products is characterized by the customers' intense participation in the creation of the products, and by the high information component of the latter. This growing information component of products is evident in the mass customization of products that the virtual corporation is applying.

On the other hand, the virtual corporation is organized as a *network organization structure*. Firms have gone from competing against each other to cooperating. This cooperation was initially based on more or less temporary alliances, and subsequently on connections of variable geometry. This new organizational structure of variable geometry, or dynamic network structure, considers that the main components of firms can be assembled and reassembled again and again to adapt to the changing environmental conditions (Aerts et al., 2002; Huang, Gou, Liu, Li, & Xie, 2002). This network structure is accepted by all of the authors (Camarinha-Matos & Afsarmanesh, 2003, 2005; Weisenfeld et al., 2001), although they refer to it using different analogies and names. Where there is consensus is in the idea that the dynamic network is the most flexible organizational form known. Its main characteristics are: (1) disaggregation of the firm's functions (design and development, manufacturing, marketing, and distribution), now carried out by independent organizations in a network; (2) existence of internal and external agents, responsible for linking the business functions of different firms (Aerts et al., 2002; Zarour, Boufaida, Seinturier, & Estraillier, 2005); (3) existence of market mechanisms for coordinating the functions, rather than plans and controls (Biggs, 2000); and (4) existence of freely accessible information systems for verifying each member's contribution (Zarour et al., 2005), rather than the slow processes of mutual trust building.

The virtual corporation is founded on a *new type of worker*. This organizational structure is based on the management of processes, demanding a flexible number of versatile and autonomous human resources. Thus, a radical change in corporate culture is required for the concept of virtual corporation to work. This culture will be reflected throughout the entire organization, requiring new management skills (Heneman & Greenberger, 2002; Khalil & Wang, 2002)—focusing on how to lead "persons" and manage "relationships" (Beranek & Martz, 2005)—and a new type of worker.

With regards to human resource management, the emphasis is on retaining highly qualified workers in this type of organization (Erickson, 2004), and there is significant rotation of workers between the different functions and divisions of the firm. These workers turn the firm into a learning organization.

Agile Web Inc. is an example of virtual corporation. It is made up of 20 successful small and medium-sized manufacturing enterprises in north-eastern Pennsylvania. The aim of this project was that small firms could become more competitive, more productive, and provide a greater value-added service to their customers by working together in new ways to provide totally integrated solutions. This virtual corporation has the ability to knit companies together in the right configuration to achieve each task with speed and efficiency. Agile Web Inc. acts as a single point of contact that directs all the aspects of multi-phase design and production projects, ever changing to meet the customer's needs and anticipating them. The member companies are able to communicate by e-mail and electronic data interchange (EDI) to ensure a quick response. Agile Web Inc. offers a totally integrated chain of suppliers with high flexibility by a sharing of needed information rather than within a traditional hierarchical structure. The company's mission is to provide customers with solutions at the lowest possible price, with the fastest delivery time and highest technical performance. This example of virtual corporation illustrates the above mentioned roles of the manufacturing function, information, and information technology, and the network structure.

FUTURE TRENDS

It will be many years before we see the first real virtual corporation. There is no empirical evidence that fully supports this theoretical model, but there is growing interest in the model in the literature, as well as among firms. This model never gained widespread acceptance because the technology needed to integrate companies with their suppliers and customers did not exist. Now, the technology exists and companies are beginning to revisit the virtual corporation concept (Joukhadar & Binstock, 2000). Thus, for example Hector Ruiz (personal communication, January 24, 1997), after being named executive president of Motorola's semiconductor sector, wrote in an internal memo that organizations like his were becoming increasingly virtual. This trend is recognized by the Electronics Manufacturing Service Industry (EMSI), which detects a shift towards the virtual corporation in companies as important as Apple, Cisco, Bay Networks, and HP.

From the academic perspective, important lines of research are opening up. On the one hand, researchers could investigate which of the three models of virtual corporation described previously is most likely to be adopted by firms in the future. Another line of research would be to design an approach integrating the three models. Finally, it would be interesting to analyze the virtual corporation model from the perspective of the most consolidated theories of the firm (transaction cost economics, institutional theory, agency theory, options theory, population ecology theory, resource-based view of the firm, network theory, sociological theory, etc.). Analyzing the virtual corporation phenomenon with each of these theories would allow us to increase understanding of the phenomenon, its impact on firms and society, and future trends.

CONCLUSION

The new knowledge-based economy requires flexibility, new organizational structures, and managerial processes—based more on information than on hierarchy. Firms need to encourage self-learning, adapt to and exploit rapid technological change, and use their core competencies to differentiate themselves from their competitors.

The characteristics described previously are reflected in the different models of virtual corporation discussed in this article. The first model equates the virtual corporation to a temporary network of firms that quickly comes together to exploit temporary market opportunities, using the best resources and capabilities of each firm. The second model focuses on the manufacture of virtual products—which vary according to customer needs—by means of stable and trusting relationships with suppliers and customers. The third model of virtual corporation tries to turn the fixed workforce costs into variable costs, which are only incurred when the market demands the product or service.

This article outlines both the basic characteristics that define the virtual corporation—excellence, technology, trust, opportunism, and absence of borders—and other elements that, while they are not exclusive to virtual corporations, must always be present to make them possible. This is the case of the set of technologies making up the so-called lean manufacturing, of information and information technology, of the new type of worker, and of the organizational structures of variable geometry or dynamic networks.

Nevertheless, all authors coincide in noting that we have not yet seen the pure virtual corporation defined in the

theoretical literature. At present it remains just an important trend towards which many firms are gradually orienting their management and organizational structure.

REFERENCES

Aerts, A. T. M., Szirbik, N. B., & Goossenaerts, J. B. M. (2002). A flexible, agent-based ICT architecture for virtual enterprises. *Computers in Industry, 49*(3), 311-327.

Agrawal, R. K., & Hurriyet, H. (2004). The advent of manufacturing technology and its implications for the development of the value chain. *International Journal of Physical Distribution & Logistics Management, 34*(3/4), 319-336.

Alsop, S. (2003). I've seen the real future of tech and it is virtual. *Fortune, 147*(7), 390.

Beckett, R. C. (2003). Determining the anatomy of business systems for a virtual enterprise. *Computers in Industry, 51*(2), 127.

Bekkers, V. (2003). E-government and the emergence of virtual organizations in the public sector. *Information Polity, 8*(3/4), 89-101.

Beranek, P. M., & Martz, B. (2005). Making virtual teams more effective: improving relational links. *Team Performance Management, 11*(5/6), 200-213.

Biggs, M. (2000, September). Tomorrow's workforce. *InfoWorld: CTO FirstMover*, S59-S61.

Biondi, D., Bonfatti, F., Monari, P. D., Giannini, F., & Monti, M. (2003). A product manager supporting a new co-design methodology for SMEs. *International Journal of Computer Applications in Technology, 18*(1/4), 174-188.

Buhman, C. H. (2003). Oncoming wave of collaboration. *Industrial Engineer, 35*(8), 43.

Camarinha-Matos, L. M., & Afsarmanesh, H. (2003). Elements of a base VE infrastructure. *Computers in Industry, 51*(2), 139.

Camarinha-Matos, L. M., & Afsarmanesh, H. (2005). Collaborative networks: A new scientific discipline. *Journal of Intelligent Manufacturing, 16*(4-5), 439-452.

Chalmeta, R., & Grangel, R. (2003). ARDIN extension for virtual enterprise integration. *The Journal of Systems and Software, 67*(3), 141.

Clases, C., Bachmann, R., & Wehner, T. (2003). Studying trust in virtual organizations. *International Studies of Management & Organization, 33*(3), 7-27.

Coates, J. (2005). At 14 technology trends. *Research Technology Management, 48*(5), 7-9.

Erickson, K. (2004). The tangible presence of virtual agribusiness. *Agri Marketing, 42*(8), 20-22.

Fitzpatrick, W. M., & Burke, D. R. (2001). Virtual venturing and entry barriers: Redefining the strategic landscape. *S.A.M. Advanced Management Journal, 66*(4), 22-30.

Gallie, E. P., & Guichard, R. (2005). Do collaboratories mean the end of face-to-face interactions? An evidence from the ISEE project. *Economics of Innovation and New Technology, 14*(6), 517-532.

Guisinger, A., & Ghorashi, B. (2004). Agile manufacturing practices in the specialty chemical industry: An overview of the trends and results of a specific case study. *International Journal of Operations & Production Management, 24*(5/6), 625-635.

Hao, Q., Shen, W., & Wang, L. (2005). Towards a cooperative distributed manufacturing management framework. *Computers in Industry, 56*(1), 71-84.

Helling, K., Blim, M., & O'Regan, B. (2005). An appraisal of virtual networks in the environmental sector. *Management of Environmental Quality, 16*(4), 327-337.

Heneman, R. L., & Greenberger, D. B. (2002). *Human Resource Management in Virtual Organizations*. Greenwich, CT: Information Age.

Hsieh, Y., Lin, N., & Chiu, H. (2002). Virtual factory and relationship marketing—A case study of a Taiwan semiconductor manufacturing company. *International Journal of Information Management, 22*(2), 109-126.

Huang, B., Gou, H., Liu, W., Li, Y., & Xie, M. (2002). A framework for virtual enterprise control with the holonic manufacturing paradigm. *Computers in Industry, 49*(3), 299-310.

Im, H. G., Yates, J., &. Orlikowski, W. (2005). Temporal coordination through communication: Using genres in a virtual start-up organization. *Information Technology & People, 18*(2), 89-119.

Joukhadar, K., & Binstock, A. (2000). Virtual enterprise comes of age. *Information Week, 811*, 141.

Khalil, O., & Wang, S. (2002). Information technology enabled meta-management for virtual organizations. *International Journal of Production Economics, 7*5(1-2), 127-134.

Kovacs, G. L., & Paganelli, P. (2003). A planning and management infrastructure for large, complex, distributed projects—Beyond ERP and SCM. *Computers in Industry, 51*(2), 165.

Lee, W. B., Cheung, C. F., Lau, H. C. W., & Choy, K. L. (2003). Development of a Web-based enterprise collaborative platform for networked enterprises. *Business Process Management Journal, 9*(1), 46-59.

Martinez, M. T., Fouletier, P., Park, K. H., & Favrel, J. (2001). Virtual enterprise—Organisation, evolution and control. *International Journal of Production Economics, 74*(1-3), 225-238.

Matthews, J. (2004). The rise of the virtual company. *Supply Management, 9*(15), 32-33.

Mo, J. P. T., & Zhou, M. (2003). Tools and methods for managing intangible assets of virtual enterprise. *Computers in Industry, 51*(2), 197.

Offodile, O. F., & Abdel-Malek, L. L. (2002). The virtual manufacturing paradigm: The impact of IT/IS outsourcing on manufacturing strategy. *International Journal of Production Economics, 75*(1-2), 147-159.

Porter, A. M. (2000). The virtual corporation: Where is it? *Purchasing, 128*(4), 40-48.

Presley, A., Sarkis, J., Barnett, W., & Liles, D. (2001). Engineering the virtual enterprise: An architecture-driven modeling approach. *International Journal of Flexible Manufacturing Systems, 13*(2), 145.

Stowell, C. (2005). Real-time collaboration with flair. *Communications News, 42*(3), 40-42.

Talukder, M. I. (2003). The perception of professionals and management personnel on the virtual organization. *The Journal of Computer Information Systems, 43*(3), 92-99.

Vakola, M., & Wilson, I. E. (2004). The challenge of virtual organisation: Critical success factors in dealing with constant change. *Team Performance Management, 10*(5/6), 112-120.

Weisenfeld, U., Fisscher, O., Pearson, A., & Brockhoff, K. (2001). Managing technology as a virtual enterprise. *R&D Management, 31*(3), 323-334.

Xu, W., Wei, Y., & Fan, Y. (2002). Virtual enterprise and its intelligence management. *Computers & Industrial Engineering, 42*(2-4), 199-205.

Zarour, N., Boufaida, M., Seinturier, L., & Estraillier, P. (2005). Supporting virtual enterprise systems using agent coordination. *Knowledge and Information Systems, 8*(3), 330.

KEY TERMS

Information Component of Products: All that the buyer needs to know to obtain and use the product, and hence achieve the desired result (information about product characteristics, instructions for use, and maintenance).

Knowledge-Based Economy: An economy characterized by the recognition of knowledge as a source of competitiveness; the increasing importance of science, research, technology, and innovation in knowledge creation; and the use of computers and the Internet to generate, share, and apply knowledge.

Lean Manufacturing: "A philosophy of production that emphasizes the minimization of the amount of all the resources (including time) used in the various activities of the enterprise" (APICS Dictionary).

Strategic Network: Agreements by which firms establish a web of close, stable relationships to provide products and services in a coordinated and flexible way.

Teleworking: Professional activity that takes place in a firm and that is independent of the firm's physical location. The person who does this work at a distance—the teleworker—keeps in contact with the firm using information technology.

Variable Cost Virtual Corporation: These firms turn fixed personnel costs into variable costs. They make use of freelance teleworkers, who can be called in or sent home at the employer's convenience, and outsourcing (Coates, 2005).

Variable Geometry Structure: Dynamic network of firms whose main components can be assembled and reassembled again and again to adapt to the complex and changing environmental conditions.

Virtual Corporation as Temporary Network: Set of firms that come together quickly to exploit temporary opportunities appearing in the market, using each firm's best resources and capabilities.

Virtual Corporation that Manufactures Virtual Products: Stable and trusting relationships with suppliers and customers to manufacture products that adapt to each customer's changing needs.

Virtual Product: A product that adapts to the customer's changing needs.

Virtual Organization

James J. Lee
Seattle University, USA

Ben B. Kim
Seattle University, USA

INTRODUCTION

Virtual organization has been well documented as both a tool for organizations to seek further profitability through the removal of traditional barriers, as well as a method to extend the provision of services to clientele in a manner previously achievable by only large, multinational corporations (Markus 2000). The widespread implementation of information technology and its many applications in modern business has moved the act of management towards a virtual focus, where managers are able to complete tasks through the use of teams in varying physical locations, with members that may or may not be employees of that firm, sharing a wide variety of data and information. With so many companies now employing virtual organization techniques, referring to a company as "virtual" or to its components as possessing "virtuality" lacks the clarity and specificity needed when using these firms as examples for others. The variety of methods through which a firm can achieve virtuality represents a span nearly as wide as the business community itself.

BACKGROUND

The earliest definitions of a virtual organization appeared when the concept of *virtuality* was applied to studies of management, before information technology existed in a refined state to support the theory. Giuliano (1982) saw that with the addition of telecommunications and networking technology, there was little need for work teams to assemble at the same time or even at a contiguous location. A structured concept of virtual organization was formed by Mowshowitz (1994, 2002), who defined virtual organization in previous work as a group of employees communicating and being managed electronically through *metamanagement.* This concept defines the way in which a *virtual task* is managed and further categorizes a virtual organization as a series of virtual tasks, completed by virtual teams in strategic global locations. As each team has a certain commitment to the parent organization, the similarity in purpose and communication style allows for clear distribution of work across multiple groups of organizational members.

As with Net-enabled organizations, the concept of virtual organizations has gained prominence among researchers and practitioners. As shown by the recent work of Schultze and Orlikowski (2001), virtuality can be understood through the perception of time and space. This article extends the scope of the virtual organization in terms of 'virtual space', a metaphor used in *time* and *space* (beyond the constraints of the actual location we belong to) dimensions (Allcorn, 1997). As opposed to the virtual organization, time and space dimensions are constrained in traditional or 'real' organizations. Time constraints occur in real organizations due to the operational time dimension of such organizations, while space dimension occurs due to constraints of location.

It is true that a virtual organization inherits the attributes of virtual dimensions—a newly defined concept of time and space. In other words, a virtual organization does not exist in our time and space, but rather exists only in virtual space (the perceptual world), which is only a metaphor of our consciousness and not reality. A virtual organization, in this sense, is the metaphor of our designed and structured consciousnesses that exists in virtual space to perform the intended actions of interest. However, the most important thing in a virtual organization is to identify the role of human actors who get involved in both the physical and the perceptual world (Orlikowski, 2002). I attempt to explain the relationships between the human actors, the real and virtual organizations, and our perceptions of these concepts.

DUALITY OF HUMAN MINDS

Metaphors play a very powerful role in structuring virtual organizations, because terms like 'virtual' and 'virtuality' originate from symbolic languages (Faucheux, 1997). These metaphors provide the meaning of existence, thus we can treat the organization like a real organization in virtual space. Continuous analogical processes between virtual and real organizations explain the existence of virtual organizations because there exist similarities and discrepancies in them (Ahuja & Carley, 1999). A virtual organization, operating within virtual space imagery, exists in our consciousness, while an actual organization physically exists in various

forms (more tangible or definable manner) such as culture, politics, resources, and so forth (Morgan, 1986). Although a virtual organization exists in our consciousness, it is associated with its physical counterpart in the 'real' world such as a parallel virtual organization and bureaucratic hierarchical organization counterpart (Allcorn, 1997). However, there is a possibility that a 'real' organization will exist only when its virtual counterpart exists in the human mind. Mowshowitz (1994, 2002) described this as 'a dominant paradigm' of virtual organization due to its unique advantages in the efficiency, cost, and effectiveness of goal-oriented activity. Surprisingly, human minds streamline these two opposing ideas of real and virtual worlds—thus, it becomes obvious that humans possess duality of existence in both the real and the virtual world.

This article discloses the social aspects of a virtual organization and identifies the role of human actors in a virtual organization (or 'consciousness' in Faucheux, 1997). This consciousness exists in the perceptual world that we create beyond the limits of time and space (Allcorn, 1997). However, its counterparts exist in various forms (entities) in the real world. To bridge the gaps between the consciousnesses and the entities, there exists a need for human interveners who possess dual identities in both virtual and real worlds. This research provides the meaning of virtual organization, and proceeds to explain the relationship between the consciousnesses (virtual organizations) and entities (real organizations) with human intervention (human actors).

Schultze and Orlikowski (2001) examine rhetorical oppositions between real organizations and virtual organizations, and in doing so apply metaphors to the discourse. The visions or views of two opposing elements are not divergent or dichotomous; rather, they offer substitutes for the opposition through a process referred to as dualism. As Orlikowski (1991) proposed in her earlier paper, "The Duality of Technology," this dualism is not mutually exclusive. The dualism originated from the work by Giddens (1984) in *The Constitution of Society.* Giddens's (1984) structuration theory integrated two main streams of sociology—objectivism and subjectivism. It appears that the structuration theory adopts the notion of phenomenology, as it seeks to make explicit the implicit structure and meaning in human experiences (Sanders, 1982). Phenomenology searches for the essence of what an experience *essentially is* and is the intentional analysis between objective appearance and subjective apprehension. Structuration theory (the process of structuration of an organization) seeks a complementary essence in the structure of organization science and in the process of struggles between objectivism and subjectivism. Interestingly, the conflict of objectivism and subjectivism was reflected in metaphors, as Lakoff and Johnson (1980, pp. 189) stated:

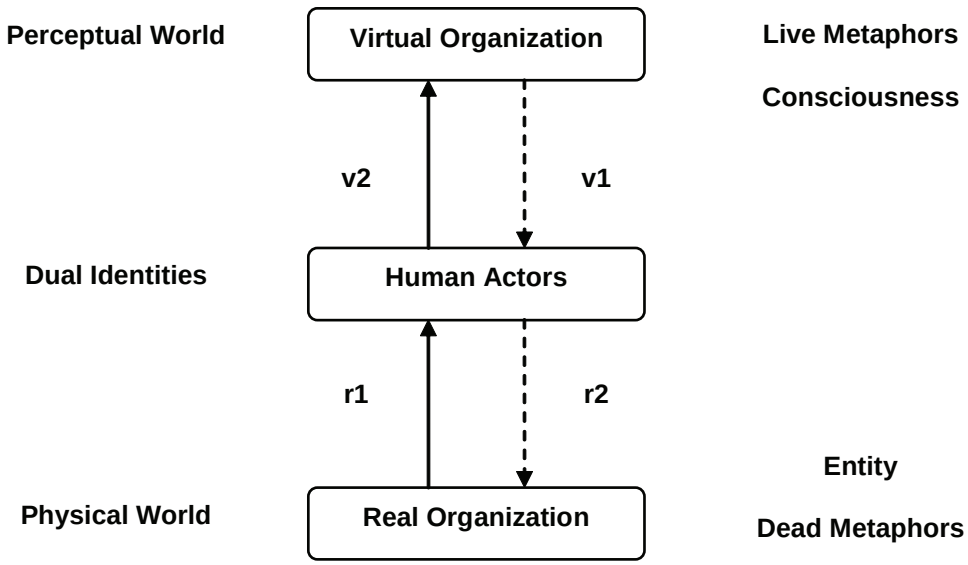
"Objectivism and subjectivism need each other in order to exist. Each defines itself in opposition to the other and sees the other as the enemy. Objectivism takes as its allies: scientific truth, rationality, precision, fairness, and impartiality. Subjectivism takes as its allies: the emotions, intuitive insight, imagination, humaneness, art, and a 'higher' truth…They coexist, but in separate domains. Each of us has a realm in his life where it is appropriate to be objective and a realm where it is appropriate to be subjective."

Human actors have very important roles in both phenomenology and metaphors due to their valuable experience. The key differentiator between objectivism and subjectivism is always human experience. Another important fact (usually overlooked by researchers) is that the use of metaphors appears in both the physical world and in the perceptual world (Harrington, 1991) because the terminology 'organization' itself results from *dead* metaphors. Tsoukas (1991) describes the process in which metaphors "have become so familiar and so habitual that we have ceased to be aware of their metaphorical nature and use them as literal terms." It implies that the metaphors of virtual organizations are *live* metaphors (Tsoukas, 1991), "knowing that these words are substitutes for literal utterances" that use dead metaphors (organization per se). Therefore, live metaphors are used to describe virtual organizations in another dimension where we can do things that are not possible in the real world because the virtual world operates without the constraints of time and space, unlike the real world.

The process of structuration involves the reciprocal interaction of human actors and institutional properties of organizations (Giddens, 1984); as Orlikowski (1991) pointed out, "The theory of structuration recognizes that human actions are enabled and constrained by structures, yet these structures are the result of previous actions." Because we live in both real and virtual worlds, we have both objective and subjective understandings of each world—dual identities. Figure 1 shows the relationship between real organizations and virtual organizations in the presence of human interveners. Both real and virtual organizations consist of rule resource sets that are implicated in their institutional articulation, thus these rule resource sets act as structures of the organizations (both virtual and real), where a structure is the medium and the outcome of organizational conduct. The structural properties do not exist outside of the actions of the organizational actors. Therefore, structural properties, related to space and time, are implicated in the production and reproduction of the organizations. In other words, both real and virtual organizations undergo structuration across the different sets of dimensions of time and space based on the perspectives of each human player.

The model in Figure 1, which is adopted from the duality of technology of Orlikowski (1991), depicts four processes that

Figure 1. Dual identities of human actors in both real and virtual organizations



operate continuously and simultaneously in the interaction between human actors and both real and virtual organizations. These processes include: (i) institutional properties, represented by arrow r1 (objective appearance of the real organization) and arrow v1 (objective appearance of the virtual organization), which are the *medium* of human actors; (ii) structures, represented by arrow r2 (subjective construction of the real organization) and arrow v2 (subjective construction of the real organization), which are the *product* of human actors; (iii) the interaction of human actors in both worlds, and the resultant influences on the social contexts of the real organization within which it is built and used (the direction of arrow r1 and v2); and (iv) how the virtual organization is built and used within particular social contexts in a real organization (the direction of arrow v1 and r2).

In Figure 1, there are two structurations from human actors, one for the real organization and the other for the virtual organization. The realms of virtual organization and real organization are objective, while the consciousness of the human player is subjective. The process of structuration (Figure 1) involves the reciprocal interaction of human actors and institutional properties of organizations (Giddens, 1984); as Orlikowski (1991) pointed out, "The theory of structuration recognizes that human actions are enabled and constrained by structures, yet these structures are the result of previous actions." We have both objective and subjective understandings of each world (dual identities) because we live in both real and virtual worlds.

The maturity phase of the real organization (with its established tradition) and the fledgling state of the virtual organization (with its newly emerging phenomena) indicate that the objective appearance of the real organization (arrow r1) and subjective construction of the real organization (arrow v2) dominate the structuration process in modern organizations. Many observations show that the knowledge and experiences accumulated in real organization enforce the formulation of the virtual organization in the abstraction process of efficient and effective goal-oriented activity (Mowshowitz, 1994). It is partly true that the creation of the virtual organization is only for the representation of the real organization in virtual space. A considerable amount of explanation arises from rethinking the basic assumptions of time and space. The real organization, whether tangible or intangible, is bound in time and space, while the virtual organization, an imaginative concept established in computer hardware and software, is free from the constraints of time and space.

Barley and Tolbert (1997) defined an institution as "shared rules and typifications that identify categories of social actors and their appropriate activities or relationships." As they explained their recursive model (institutions and actions), institutionalization involves the behavior of revision or replication of organizational abstracts (work procedures), and entails objectification and externalization of behaviors. In this sense, the successful functioning of a virtual organization is reaching institutionalization in virtual space. Through this process, the virtual organization becomes stable and helps serve as the constitution where human actors can follow their activities. Upon further inference, institutions from business processes of real organizations constrain human

actors (constitutive nature, r1), who in turn construct institutions of virtual organization (constituted role, v2) and/or vice versa (from v1 to r2).

The above arguments provide complementary insights to the social process explained by the structuration theory (Giddens 1984). In this theory, actions and institutions continuously interact, thereby determining the structure. The structuration theory lacks the explanation of how these interactions (revising and reproducing an institution or structure) are processed, although this is arguable as Giddens explains the role of reflection, interaction, and so forth. However, Barley and Tolbert (1997) clearly stated that their work, 'the aim of institutional theory', is "to develop the implications of structuration theory for the interplay between actions and institutions and to address the practical problem of how to study institutional maintenance and change in organizations."

The results of this study are compatible with the belief of Barley and Tolbert (1997) that "the institutional perspective must come to grips with institutionalization as a process if it is to fulfill its promise in organization studies." The focus of this study is the explanation of: "What is going on at a virtual organization?" The result revealed by this research is a rich description of theoretical induction. A limitation of this process is that it only reflects one part of the recursive model of institutional theory (Barley & Tolbert, 1997).

FUTURE TRENDS

Organizations today are usually faced with a turbulent environment that requires flexible and dynamic responses to changing business needs. Many organizations have responded by adopting decentralized, team-based, and distributed structures. Advances in communication technologies have enabled organizations to acquire and retain such distributed structures by supporting coordination among people working from different locations (Ahuja & Carley, 1998). Thus, virtuality comes from important roles to achieve above goals of current business environments. Conceptually, virtual organization can be seen as "emptying of organization," where emptying of information and knowledge has occurred in current communication technology (Giddens, 1984, 1990). IT (or information systems) generates and stores information that helps in the emptying (separation) of information from organizations. Knowledge, a supposedly higher format of information, is managed by knowledge management systems (KMSs), further evidence of the emptying of knowledge from organizations. Because data, information, and knowledge of organizations are emptying from their organizations (i.e., can be stored and manipulated in information systems), the separation of the organization from its four-dimensional entity **(**time factor and three location factors—latitude, longitude, altitude**)** can be achieved in the form of virtual organization (Giddens, 1984, 1990) if the rest of the components of the organization are implemented in network.

The e-business model has been widely accepted by modern business organizations. Some companies initiated this model hoping to reduce transaction costs; others are running it for its future value. Many cases show that e-business is now a necessity because Internet technology revolutionized communications. Only a few years ago many researchers and practitioners perceived e-business as an extension of conventional business practices for doing business at a lower cost. However, Internet technology changed our lifestyles and affected the way of doing business (Chudoba, Wynn, & Watson-Manheim, 2005). The e-business model is no longer a short-term strategy; it needs a thorough and systematic investigation with the foundation of virtual organization concept as discussed in Kaplan and Norton (2006).

Virtual organization incorporates all the revolutionary practices in the global network (e.g., e-business, e-commerce, virtual community, virtual company, etc.) and provides a unique opportunity for society. The foremost contribution of this study is to provide a new vision of the future organization. IT is not simply a technological product of communication networks, rather it is a vehicle whereby 'virtualization' occurs in ways to reconstruct what has been eliminated from the traditional socialization process. Virtual organization is established mainly to achieve the shared goal of participants or to take advantage of the availability of advanced IT and communication networks. When social mechanisms are developed in virtual organization, it flourishes as an entity with its own norms, culture, and SOP. As IT is integrated further into virtual organization to manage conflict and encourage cooperation, it grows into a complete organization (McKinney & Whiteside, 2006). IT becomes extended linkages among members with which they simulate frank communication and open consideration of different alternatives and opinions. The proactive members seek a productive discussion and creative thinking.

CONCLUSION

The approach of this article is that organizational transformation is "the ongoing practices of organizational actors, and emerges out of their (tacit and not so tacit) accommodations to and experiments with the everyday contingencies, breakdowns, exceptions, opportunities, and unintended consequences that they encounter" (Orlikowski, 1996). The above statement is identical to the theoretical framework of this article in that users of the system continuously interact with the system through producing, reproducing, and transforming work practices (Giddens, 1984).

The goal-oriented depiction of a virtual organization (Mowshowitz, 1997) is limited on a computer with a communication tool or computer network that increases the

efficiency and effectiveness of organization performance (Mowshowitz, 1994). This article complements the virtual organization as a social system giving the new meanings of time and space (Maznevski & Chudoba, 2000; Orlikowski & Yates, 2002). This study rethinks the philosophy of virtual organization, providing insight into the concept of duality of human identity. It is not only a lens for understanding virtual organizations, but also a socio-technical understanding of virtual organizations through structuration.

REFERENCES

Ahuja, M., & Carley, K. (1999). Network structure in virtual organization. *Organization Science, 10*(6), 741-757.

Allcorn, S. (1997). Parallel virtual organizations: Managing and working in the virtual workplace. *Administration & Society, 29*(4), 412-439.

Barley, S.R., & Tolbert, P.S. (1997). Institutionalization and structuration: Studying the links between action and institution. *Organization Studies, 18*(1), 93-117.

Chudoba, K.M., Wynn, E., Lu, M., & Watson-Manheim, M.B. (2005). How virtual are we? Measuring virtuality and understanding its impact in a global organization. *Information Systems Journal, 15*, 279-306.

Faucheux, C. (1997). How virtual organizing is transforming management science. *Communications of the ACM, 40*(9), 50-55.

Giddens, A. (1984). *The constitution of society.* Berkeley, CA: University of California Press.

Giddens, A. (1990). *The consequences of modernity.* Stanford, CA: Stanford University Press.

Harrington, J. (1991). *Organizational structure and information technology.* Hertfordshire, UK: Prentice Hall International.

Kaplan, R.S., & Norton, D.P. (2006). How to implement a new strategy without disrupting your organization. *Harvard Business Review, 84*(3), 100-109.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Markus, L.M. (2000). What makes a virtual organization work? *MIT Sloan Management Review, 42*(1), 13-26.

Maznevski, M.L., & Chudoba, K.M. (2000). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science, 11*(5), 473-492.

McKinney, V.R., & Whiteside, M.M. (2006). Maintaining distributed relationships. *Communications of the ACM, 49*(3), 82-86.

Morgan, G. (1986). *Images of organization.* Beverly Hills, CA: Sage.

Mowshowitz, A. (1994). Virtual organization: A vision of management in the information age. *The Information Society, 10*, 267-288.

Mowshowitz, A. (2002). *Virtual organization: Toward a theory of societal transformation stimulated by information technology.* CT: Quorum Books.

Mowshowitz, A. (1997). Virtual organization. *Communications of the ACM, 40*(9), 30-37.

Orlikowski, W.J. (1991). The duality of technology: Rethinking the concept of technology in organizations. *Organization Science, 3*(3), 398-427.

Orlikowski, W. (1996). Improvising organizational transformation over time: A situated change perspective. *Information Systems Research, 7*(1), 63-92.

Orlikowski, W. (2002). Knowing in practice: Enacting a collective capability in distributed organizing. *Organization Science, 13*(3), 249-273.

Orlikowski, W., & Yates, J. (2002). It's about time: Temporal structuring in organizations. *Organization Science, 13*(6), 684-700.

Sanders, P. (1982). Phenomenology: A new way of viewing organizational research. *Academy of Management Review, 7*(3), 353-360.

Schultze, U., & Orlikowski, W.J. (2001). Metaphors of virtuality: Shaping an emergent reality. *Information and Organization, 11*(1), 45-77.

Tsoukas, H. (1991). The missing link: A transformational view of metaphors in organizational science. *Academy of Management Review, 16*(3), 566-585.

KEY TERMS

Collaborative Culture: By their nature, virtual organizations foster camaraderie between members even in the absence of face-to-face communications. Since the built-in communications tools are so easy to access and use, relationships form between members who have not even met. A corporate culture forms out of friendship that produces a highly collaborative nature, unlike traditional organizations where such extensive communicating is not required.

Complementary Core Competencies/Pooling of Resources: The ease with which two members of a virtual organization can communicate allows them to pool their resources, even with members not directly involved in a specific project. Separate entities can quickly be called upon to provide secondary service or consult on a project via virtual channels.

Customer-Based/Customized Products: A virtual organization provides the unique opportunity to provide their customers with highly specialized products as per their specific needs. This can be accomplished through outsourcing work to a separate organization or through the use of a virtually connected interorganizational node located closer to the customer. Either way, it becomes simple to add a function based on the customer's request and seamlessly integrate that function into the existing framework.

Electronic Communication: A vital concept to the virtual organization is the ability to communicate through purely electronic means, eliminating the need for physical contact and allowing the geographical dispersion of organization members. Online collaboration via e-mail, discussion boards, chat, and other methods, as well as telephone and facsimile communications, are primary contributors to the removal of time and space in this new organizational concept.

Explicit Goals: Similar to meta-management, each member of the organization is charged with an explicit task to complete as it relates to the overall function of the organization. Often times, after this single goal is completed, the link between the organization and the entity is dissolved until a further need for it is realized. At this point, the link is reestablished.

Flexibility: Virtual organizations are, by their nature, flexible. Traditional organizational structures are rooted in the physical world and rely on structures, unalterable networks, and specific locations to function properly. Because of this, when it becomes necessary to introduce change into a specific organization, a barrier is reached where further alteration requires physical, costly modifications. A virtual organization is unhindered by these problems. These structures are designed so that they can operate regardless of time or place, independent of existing physical realities.

Functional or Cultural Diversity: The nature of global diversity and the ability to locate organizational functions across the globe creates a diverse environment for the entire organization. Since members are all in different locations and charged with different tasks, diversity exists that is only found in the very largest multinational corporations.

Geographical Dispersion: The combination of virtual organization with IT allows groups of employees to make progress on one project while working in tandem with another group in a distant physical location. Because information can be shared and meetings can be held with the use of high-speed networks and computers, tasks can be carried out in the location that is most appropriate and germane to that function.

Information Technology (IT): The crucial component of a modern virtual organization. Without advances in technology, many of the realities of today's virtual companies would be merely science fiction. These components include the Internet, LAN and WAN networks for business, e-mail and online chat/bulletin boards, and real-time videoconferencing. These technologies allow smaller workgroups as part of a larger company to operate independently of each other, across a room or the globe.

Open Communication: The foundation of a virtual organization is its set of communications components that exist in absence of face-to-face exchanges. A virtual organization can only survive if its members communicate freely through the provided channels between them, be they based on the Internet or more traditional telephone technologies. The organization cannot continue to function unless it is aware of what all its members are currently completing, and often times, when communication is more closed, work that is being completed in tandem by more than one member can be hindered or brought to a halt.

Participant Equality: Each individual facet of a virtual organization is expected to contribute an equal amount of work towards a given goal, if appropriate. While the equality may not be measured best in quantity, it can be restated as effort and the successful completion of all tasks assigned to it, be they large or small. Since every task is considered to be essential as a part of the project, the equality comes in the addition of that piece to a larger puzzle.

Sharing of Knowledge: Members of a virtual organization collaborate to share their knowledge gained from individual activities performed. Since collaboration is facilitated through the communications channels that are afforded through the virtual organization, it is common to find "knowledge bases" or other database systems that contain information and documents pertaining to past experience.

Switching: The switching principle is a fundamental advantage that a virtual organization has over a traditional one. Because the links between organizational functions are largely electronic and non-physical, it is easy to replace a weak component with a stronger one. Where this activity could be considerably expensive if the item in question was a physical supply chain, it may only be a change of suppliers for the virtual organization and can be made with a phone call and database edit, as opposed to a building project.

Temporary: Virtual organizations are often formed to fill temporary needs, only extending to the end of the specific project that is charged to them. In a manufacturing project, a virtual organization may be formed between the engineers who design the project, the suppliers who provide the raw materials, and the factory who processes those into finished goods. At the end of that particular project, those alliances are dissolved, as they are no longer necessary to benefit the three independent groups.

Trust: The lack of physical interaction places a higher regard on the trust that exists between each entity involved in the organization. Since fewer "checks and balances" can be placed on appropriate departments, management and other entities trust that they will complete the appropriate work on time or be straightforward about delays or problems. If two entities working on a project together separated by thousands of miles are unwilling to trust each other, the work slows and suffers to a critical point.

Vague/Fluid/Permeable Boundaries: As a continuation of flexibility, the virtual organization is characterized by vague boundaries as to the extent of its use and purpose. Since small tweaks can easily and largely affect the overall organization, it is quite possible to extend the boundaries of an organization so that they encompass new purpose, people, or control.

A Virtual Reality System for Learning Science in a Science Center

Sharlene Anthony
Singapore Science Centre, Singapore

Leo Tan Wee Hin
Nanyang Technological University, Singapore

R. Subramaniam
Nanyang Technological University, Singapore

INTRODUCTION

Current trends in informal science learning tend to place more emphasis on science centers as tools to bridge the technological gap for their visitors (Salmi, 2003; Sandifer, 2003). In line with compelling evidence in the multimedia literature, which shows that technology-based environments do provide good instructional support for meeting learning needs (Kim, 2006; Lim, Nonis, & Hedberg, 2006), it would be useful to investigate the potential of technology-based exhibits at science centers to create new multisensory experiences for learning science topics in a way that is different from traditional methods of teaching. This can provide pointers for schools to see how such attractions can be used to assist or complement the formal science learning in schools.

The principal objective of this research is to investigate the effectiveness of technology-based exhibits in promoting affective learning outcomes among students of mixed ability visiting a science centre. The chosen exhibit is the CAVE (cave automated virtual environment), a supercomputer-based multimedia system.

BACKGROUND

The CAVE is basically a virtual reality system. Its genesis can be traced to the need to develop compact virtual reality systems that can overcome the inconvenience of using head-mounted display sets and the limitations of single-user interaction at a time, both of which have plagued earlier versions. The ideas of Thomas DeFanti and Don Sandin of the Electronic Visualization Laboratory at the University of Chicago, in 1991, provided the basis for the development of the first working model of the CAVE by Carolina Cruz-Neir in 1992 (Cruz-Neir, Sandin, DeFanti, Kenyon, & Hart, 1992; Defanti, Sandin, & Cruz-Neira, 1993).

Relying on the use of computer-generated graphics and multisensory digital data, the CAVE provided, for the first time, the relishing of virtual reality as immersive and interactive experiences for a group of people in a specialized setting. Soon, the scope for using the CAVE as a platform to simulate complex scientific phenomena, as well as for generating walkthroughs in a range of virtual environments, was recognized. Over the years, several applications have been modelled to exploit the unique features of the CAVE. Some of these include

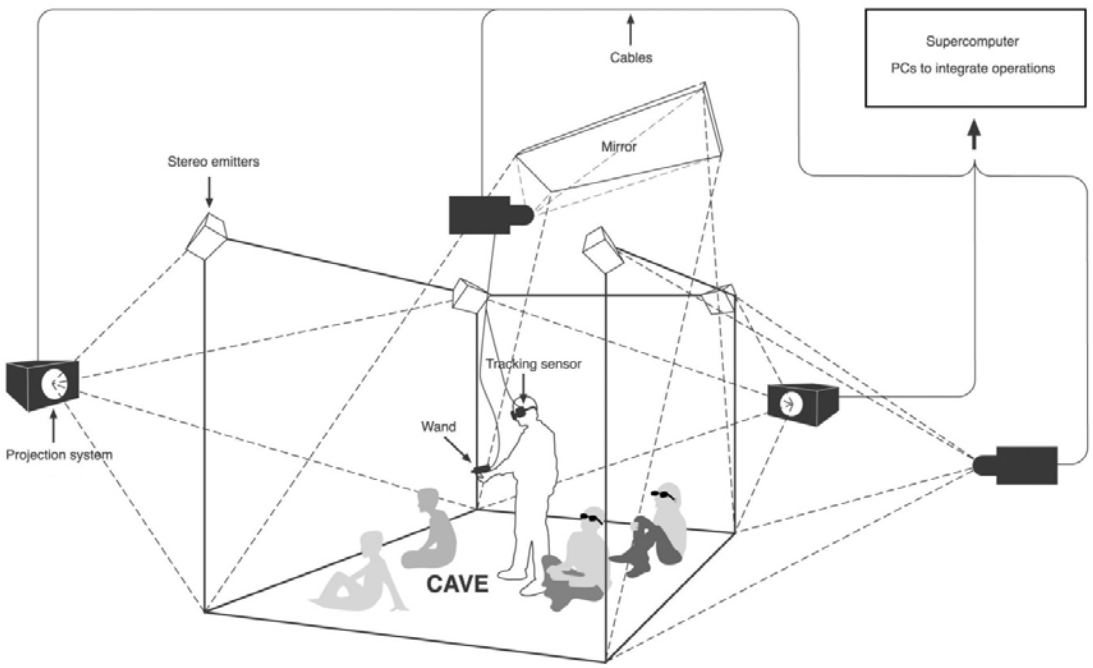
a. Exploration of a sprawling virtual plains inhabited by diverse vegetation (Moher, Johnsoon, Yongjoo, & Ya-Ju, 1999)
b. Collaborative construction, cultivation, and tending of a healthy virtual garden by young children (Roussos, Johnson, Leigh, Vaslakis, Barnes, & Moher, 1997)
c. In-depth probing of an ant, the inside of the Earth, an iceberg, a volcano, the solar system, and the human heart (Johnson, Moher, Ohlsson, & Gillingham, 1999)

The CAVE

This study focuses on the CAVE (Tan, Subramaniam, & Anthony, 2005) at the Singapore Science Centre. It is of interest to note that the only other CAVE to be set up outside a research laboratory in the world is the one at the National Museum of Emerging Science and Innovation in Miraikan, Japan.

Modelled as a cube of area 27 m^3, the CAVE comprises display screens mounted at right angles to the plane of image projection, acoustic speakers placed at the upper vertices to produce sonic effects, stereo emitters situated at the edges to ensure proper mapping between the frame rate and the configuration of the stereo glasses used by participants, tracking sensors on the stereo glasses used by the lead user to ensure that what the lead user sees is also what the participants wearing stereo glasses see, and a supercomputer (Silicon Graphics Onyx 2 Reality engine) to coordinate the overall operations. Navigating through the virtual environment is facilitated by the joystick on the wand held by the lead user,

Figure 1. Architectural elements of the CAVE



while the three buttons on the wand can be used to set off various acts of interactivity The stereo glasses used by the participants produce the 3-D effect, and this allows them to be immersed interactively in the virtual environment.

Figure 1 shows the principal elements of the CAVE.

CAVE Program on Water

The program that is the focus of this study is on the molecular structure of water. Appreciation of the following 3-D scenarios is made possible in the CAVE with this program:

- Coupling of two hydrogen atoms and one oxygen atom via covalent bonds to form a molecule of water; this exists as a 3-D structure in the space of the CAVE and can be "touched"!
- Motion of electrons going around the nucleus of the hydrogen and oxygen atoms; it is possible for participants to walk through into the interior of these atomic configurations and get a nuanced view from any perspective!
- Excursion into the interior of the crystal structure of ice.
- Clustering of H_2O molecules in various packing densities to form the three states of matter: ice, water, and vapour.

To enhance the reality of the immersive and interactive experience, suitable sound effects are produced when various modes of interactivity are triggered, and participants are also able to "touch" the 3-D images in the CAVE.

Samples for this Study

In the present study, all participants were from the Primary five level (Grade 5). One group was a class of 35 students (16 males and 19 females) from the EM1 stream. The second group comprised a class of 33 students (16 males and 17 females) from the EM2 stream, while the last group comprised 34 students (16 males and 18 females) from the EM3 stream. All three groups were from different schools,

and they were exposed to the CAVE experience on separate days. In terms of academic ability, EM1 > EM2 > EM3. The CAVE can accomodate about 10 persons at a time

EVALUATION INSTRUMENTATION

In order to test the efficacy of the CAVE programme offered on water, an evaluation instrument was developed, validated, and piloted. Based on a five-point Likert-type scale, ranging from Strongly Agree (SA) to Strongly Disagree (SD), the final version of this instrument was used to measure the affective outcomes of the students' experience. The instrument took about 5 minutes for completion by the students.

RESULTS

Internal consistency of the evaluation instrument was obtained by extracting the Cronbach Alpha coefficient; the value of 0.91 is well above the recommended norm of 0.70, and thus indicates good reliability of the instrument developed. The Cronbach Alpha coefficients for the subscales are as follows: 0.86 for Educational Potential sub-scale, 0.77 for the Effectiveness of Learning subscale, and 0.70 for Learning Climate sub-scale. The overall means for the individual statements are presented in Table 1.

The highest overall mean score for the entire survey was found among the EM3 males ($\underline{M}$ = 67.87, $\underline{SD}$ = 4.69), while the lowest mean score was found among the EM1 males ($\underline{M}$ = 59.75, $\underline{SD}$ = 14.67). The relevant descriptive statistics are shown in Table 2.

A one-way between groups analysis of variance was conducted to explore the impact of stream on the affective survey instrument. Equal variance was assumed (Levene's test $p>.05$). There was no significant difference found between the streams at the $p<.05$ level [$F(2, 99) = 1.57$, $p=.21$].

An independent-samples t-test was conducted to compare the affective survey scores for males and females. Equal variance was assumed (Levene's test $p>.05$). There was no significant difference in scores for males ($\underline{M}$ = 63.71, $\underline{SD}$ = 10.19) and females [($\underline{M}$ = 62.72, $\underline{SD}$ = 11.65); $t(100)=0.45$, $p=0.65$]. The magnitude of the differences in the means was very small (eta squared = 0.002).

The mean scores for each individual statement were compared using an independent-samples t-test to investigate the effect of gender. Equal variance was again assumed (Levene's test $p>.05$) for all statements except for Statement 10 (*The guide who led the CAVE presentation facilitated effectively my learning*). There was no statistically significant difference found between the mean scores of the individual statements in the affective survey for males and females.

DISCUSSION

This study builds on the notion that technology, used appropriately, can enhance learning and be a vital element for the education system (Cohen, 2001). Specifically, the use of technology-based exhibits in science centers can provide instructional support to enhance students' visual-spatial thinking, a key component that is much neglected in boosting scientific creativity and communication in science education (Mathewson, 1996), especially in the formal school environment. It further builds on growing evidence that science centers, by their very nature, can provide an environment conducive for inculcating an interest in science and providing the formal learning system with a much needed avenue to broaden and spark interest in science among their students through the use of technology-rich environments.

The CAVE uses high quality visualizations, immersive experiences, interactivity, and stereoscopic imagery to provide a unique virtual environment that is conducive for learning. This contrasts sharply from the first few attempts at virtual reality systems, which were fragile and required the use of head-mounted display sets to showcase the visualization of learning scenarios (Teitel, 1990) and that allowed only single users at any one time. The CAVE afforded, for the first time, a generational advance in immersive and participatory experiences for a group of people at the same time. The more rugged and robust CAVE provides a quantum leap in virtual reality technology (Cruz-Neira, et al., 1992; DeFanti, et al., 1993).

Virtual reality technology has been noted to provide a plethora of benefits. Pantelidis (1997) (as cited in Bakas & Mikropoulos, 2003) highlights that virtual reality can be motivating, has the capacity to depict various processes more accurately, allows objects to be appreciated via multidimensional perspectives spanning the micro to the macro domains, provides instructional support for topics that are rather risky or time-consuming to teach, and promotes active participation.

However there remain many obstacles in the implementation of such advanced technologies in schools. Johnson et al. (Johnson, Moher, Cho, Lin, Hass, & Kim, 2002), list a few of such factors: "insufficient resources to support teaching practices constrained by conventional school organization; lack of alignment between technology-based materials and school curriculum/performance goals; lack of authentically motivated teacher preparation and training offered; lack of time for teachers to pursue additional training; insufficient technical support both for maintaining and upgrading hardware and software systems and for providing assistance on the operation and capabilities of applications software; and failure to provide specific mechanisms for assessing the pace

Table 1. Overall means and standard deviations for the statements in the evaluation instrument

S/n	Statement	Overall Mean (SD)
	Learning Climate	
1	The CAVE makes learning fun and interesting	4.42 (0.88)
2	The design of the CAVE is appropriate for learning	4.08 (1.13)
3	The amount of material used in the multimedia presentation in the CAVE was just right	4.60 (0.80)
4	The sequence of material shown in the multimedia presentation in the CAVE was logical and systematic	4.44 (1.0)
5	I liked the immersive environment in which the topic was taught	3.93 (1.16)
	Effectiveness of Learning	
6	The use of multimedia technology in the CAVE helped to illustrate concepts in a way that facilitated my understanding	4.13 (1.19)
7	I learnt more about the structure of water from the CAVE than from my text book	3.87 (1.20)
8	The CAVE builds on my knowledge of water learnt from text books and in the classroom	4.44 (0.79)
9	The interactive environment in which the topic of water was explored in the CAVE contributed to greater learning	3.98 (1.22)
10	The guide who led the CAVE presentation facilitated effectively my learning	4.05 (1.22)
	Educational Potential	4.41 (1.12)
11	The CAVE is a good teaching tool to learn science	
12	I would like to learn other science topics through the CAVE	4.29 (1.22)
13	The CAVE is an exciting media for learning	4.11 (1.27)
14	The use of technology in the CAVE increased my motivation to learn	4.10 (1.17)
15	Because of the way information is presented in the CAVE, I understand the topic of water better	4.32 (0.83)

Table 2. Overall means and standard deviations of the evaluation instrument by stream and gender

Groups	Total Mean (SD)
EM1 Males (N=16)	59.75 (14.67)
EM1 Females (N=19)	61.53 (11.39)
All EM1 (N=35)	60.71 (12.82)
EM2 Males (N=16)	63.50 (7.27)
EM2 Females (N=17)	63.76 (8.36)
All EM2 (N=33)	63.64 (7.73)
EM3 Males (N=16)	67.87 (4.69)
EM3 Females (N=18)	63.00 (14.74)
All EM3 (N=34)	65.29 (11.31)
All Males (N=48)	63.71 (10.19)
All Females (N=54	62.72 (11.65)
Overall (N=102)	63.19 (10.95)

and effectiveness of technology integration with respect to student learning and school climate."

The siting of the CAVE in a public access setting, such as a science centre, can help to realize the educational benefits of virtual reality technology by reaching out to a large student population through field trips that complement the school science syllabus and thus, providing a novel and exciting medium for students to learn..

Gender

Although evidence exists in support for and against the existence of the so-called gender gap in science between males and females, the difference seems to be mainly in attitudes towards science (Martin, Mullis, Gonzalez, Gregory, Smith, et al., 1999; Weinburgh, 1995), and not so apparent in achievement in science (Kotte, 1992). This trend follows through in Singapore, as reported in the Third International Mathematics and Science Study (TMSS) report. For Singapore, the results indicated that gender differences are less prominent in terms of achievement in science, but are more apparent in the attitude realm. In the 1995 and 1999 tests, Singapore was ranked third and second respectively among 14 countries (that teach science as a single subject). A very apparent point from the TMSS results was the existence of a wide gender gap in attitudes towards science, with males showing more positive attitudes towards science than females (Martin et al., 1999, p. 152).

Analysis of the effect of the CAVE experience, with respect to gender, showed no statistical difference in scores between males and females. However, the mean scores for males ($\underline{M}$ = 63.71, $\underline{SD}$ = 10.19), although not significant, were higher compared to the females ($\underline{M}$ = 62.72, $\underline{SD}$ = 11.65). There is a strong expression of interest by both groups of students in wanting to learn about other science topics through the CAVE; mean for Statement 12 in respect of this is 4.29. The CAVE is thus an exhibit that provides favourable affective outcomes for both males and females, and puts them on equal learning terms. Overall, the feedback was very positive on the use of the CAVE as a teaching tool, regardless of gender.

Academic Ability

A review of the literature has shown that academic ability (as determined by intelligence quotient, achievement test scores, academic performance, and teacher feedback) correlates significantly with science achievement (Neathery, 1997). In Singapore, a similar effect was observed among Secondary Three (Grade 9) students, with the above-average students outperforming the lower academic ability students in science achievement tests (Lim, 1986). The same study also found that academic ability was linked with attitudes towards science. Here, the higher ability students reported greater enjoyment in science, and placed the subject in a favourable position in society, as compared to those in the EM3 stream (Lim, 1986). According to Weinburgh (1995), there exists an interaction effect between academic ability and gender. The low to moderate ability males showed higher positive attitudes towards science than their female counterparts, while the high ability girls had more positive

attitudes than high-ability boys. On the other hand, females of both high and low abilities reported higher correlation between attitude and achievement than the males.

In the present study, there was no significant difference found between the various academic streams at the $p<.05$ level [$F(2, 99) = 1.57$, $p=.21$]. Although not statistically significant, the results actually showed that the lowest ability students EM3 ($\underline{M} = 65.29$, $\underline{SD} = 11.31$) had the highest total mean score in the affective survey, followed by the EM2 ($M = 63.64$, $SD = 7.73$), and finally the EM1 ($M = 60.72$, SD, 12.82) students. Perhaps this could be due to the higher ability students' (in this case, the EM1 students) preference for receiving information in a more traditional manner. No interaction effect between stream and gender was observed.

Overall, the experience was well received by the students. The use of high-resolution graphics, wide field view contributing to an enveloping experience, 3-D imagery, and special sound effects have proved to be a powerful combination for promoting learning. Whilst not their first experience with virtual reality, since virtual reality arcades and home PC-based virtual reality games are common in Singapore, the feeling of immersion is powerful, as is also the stereoscopic reality of the imagery. The virtual reality environment managed to generate a good deal of interest and excitement. The guided experience and ability to pause and take questions as they arise also puts the CAVE in a positive light.

All these provide the scaffolding necessary to entrench the concepts in the cognitive psyche of students to a reasonable extent. More importantly, the shared learning experience in the communal setting of the CAVE is a factor that has found support among students coming for such experiences. Teachers have also commented favourably on the realistic portrayal of the various processes and structures. Coupled with the visible interest and enthusiasm generated by their students during the CAVE session, the quantitative data collected on the CAVE, in this study, provides the necessary indication of its potential as an educational tool, and makes it easier to convince teachers of the direct and indirect benefits of an out-of-school trip to experience technology-based exhibits at a science center The role of the teacher in exposing students to new learning environments in order to make learning enjoyable is thus, of paramount significance (Barry & King, 1993; Lucas, 2000). In Singapore, it is the teacher who makes the decision on what out-of-school learning experiences their students are exposed to.

Clearly, the learning climate, the effectiveness of learning, and the educational potential of the CAVE are rated positively by the students in this study. This is not surprising, as the CAVE program on water was specifically developed as an application to complement this topic in the science curricula in schools in Singapore.

FUTURE TRENDS

The advent of the CAVE has made virtual reality applications in 3-D to be enjoyed by a good number of people at the same time, unlike the days of head mounted displays, when only one person can use it at a time. Since the CAVE uses a supercomputer, it is likely that it will be used mainly for high-end applications in universities, research institutes, and public access settings, such as museums and science centres. The wide field presentation permits scaling up of processes and phenomena, as well as navigation through these environments, to an extent hitherto not thought possible.

CONCLUSION

Science centers provide resources in the form of tools and facilities for teachers to make use of to engage their students. Teachers who turn to participatory science museums as a resource to supplement classroom learning will, no doubt, find that some learning does take place on the school trip (Falk, Koran, Jr., & Dierking, 1986; Wellington, 1990). In the case of attractions like the CAVE, which provide a novel learning environment in the form of 3-D, teachers can be assured that besides the fun and entertaining elements that are usually associated with such visits (Hofstein & Rosenfeld, 1996), some content learning is also transferred. The experience of the CAVE, if integrated with the classroom learning, can provide a holistic learning experience for students.

The results of this study provide further support for the successful use of virtual reality in science centres (Bell & Rabkin, 2002), and indicate that schools can capitalize on such technology-based attractions to promote learning for their students, since it permits realistic visualizations of complex phenomena on a variety of topics.

On the downside, often these attractions are expensive and beyond the reach of schools to own. In addition, the equipment requires regular maintenance and a dedicated technician to operate and troubleshoot.

As can be seen from this study, teachers can take advantage of the informal learning environment of a science center, where they are presented with more opportunities to build on their mutual relationship with students. Students, on the other hand, have an opportunity to gather information from sources other than their teachers, and in an environment that is novel. Science centres, on the other hand, would have played their part in promoting science, while stimulating interest in learning science concepts in a manner that is exciting.

REFERENCES

Bakas, C., & Mikropoulos, T. A. (2003). Design of virtual environments for the comprehension of planetary phenomena based on students' ideas. *International Journal of Science Education*, *25*(8), 949-967.

Barry, K., & King, L. (1993). *Beginning teaching.* Sydney: Social Science Press.

Bell, L., & Rabkin, D. (2002). A new model of technology education for science centres. *The Technology Teacher*, 26-28.

Cohen, V. L. (2001). Learning styles and technology in a ninth-grade high school population. *Journal of Research on Computing in Education*, *33*(4), 355-366.

Cruz-Neira, C., Sandin, D., DeFanti, T., Kenyon, R., & Hart, J. (1992). The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM*, 64-72.

DeFanti, T., Sandin, J., & Cruz-Neira, C. (1993). A room with a view. *IEEE Spectrum, October*, 30-33.

Falk, J. H., Koran, Jr. J. J., & Dierking, L. D. (1986). The things of science: Assessing the learning potential of science museums. *Science Education, 70*(5), 503-508.

Hofstein, A., & Rosenfeld, S. (1996). Bridging the gap between formal and informal science learning. *Studies in Science Education*, *28*, 87-112.

Johnson, A., Moher, T., Cho, Y. J., Lin, Y. J., Hass, D., & Kim, J. (2002). Augmenting elementary school education with VR. *IEEE Computer Graphics and Application*s, *March/April*, 6-9.

Johnson, A., Moher, T., Ohlsson, S., & Gillingham, M. (1999). The round earth project: Collaborative VR for conceptual learning. *IEEE Computer Graphics and Applications*, *19*(6), 60-69.

Kim, P. (2006). Effects of 3-D virtual reality of plate tectonics on fifth grade students' achievement and attitude towards science. *Interactive Learning Environments*, *14*(1), 25-34.

Kotte, D. (1992). *Gender differences in science achievement in 10 countries*. Frankfurt: Peter Lang.

Lim, C. P., Nonis, D., & Hedberg, J. (2006). Gaming in a 3D multiuser virtual environment: Engaging students in science lessons. *British Journal of Education Technology*, *37*(2), 211-231.

Lim, C. T. (1986). *Attitude and learning behaviour correlates of science performance of secondary three students.* Unpublished Master's thesis, National University of Singapore.

Lucas, K. B. (2000). One teacher's agenda for a class visit to an interactive science center. *Science Education*, *84*, 524-544.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Gregory, K. D., Smith, T.A., et al., (1999). *TIMSS 1999 international science report: Findings from IEA's repeat of the Third International Science and Science Study (TIMSS) at the eighth grade* (Executive summary). Chestnut Hill, MA: Ceter for the Study of Testing, Evaluation, and educational Policy Boston College. Retrieved from http://isc.bc.edu/timss1999b/pdf/TB99_Sci_4.pdf

Mathewson, J. H. (1996). Visual-spatial thinking: An aspect of science overlooked by educators. *Science Education, 83*, 33-54.

Moher, T., Johnsoon, A., Yongjoo, C., & Ya-Ju, L. (1999). Observation-based ambient environments. In B. Fisherman & S. O'Connor–Divekbiss (Eds.), *Proceedings of the Fourth International Conference of the Learning Sciences* (pp. 238-145).

Neathery, M. F. (1997). Elementary and secondary students' perception towards science: Correlations with gender, ethnicity, ability, grade and achievement. *Electronic Journal of Science Education*, *2*(1). Retrieved 14 April, 2005, from http://unr.edu/homepage/jcannon/ejse/neathery.html

Roussos, M., Johnson, A., Leigh, J., Vaslakis, C., Barnes, C., & Moher, T. (1997). NICE: Combining constructivism, narrative and collaboration in a virtual environment. *Computer Graphics*, *31*(3), 62-63.

Salmi, H. (2003). Science centres as learning laboratories: Experiences of Heureka, the Finnish Science Centre. *International Journal of Technology Management*, *25*, 460-476.

Sandifer, C. (2003). Technological novelty and open-endedness: Two characteristics of interactive exhibits that contribute to the holding of visitor attention in a science museum. *Journal of Research in Science Teaching*, *40*(2), 121-137.

Tan, W. H. L., Subramaniam, R., & Anthony, S. (2005). Cave automated virtual environment; A supercomputer-based multimedia system for learning science in a science centre. In S. Sharma & S. Mishra (Eds.), *Interactive multimedia in education and training* (pp 327-349). Hershey, PA: Idea Group Publishing.

Teitel, M. (1990). The eyephone: A head-mounted stereo display. In *Proceedings of the SPIE Conference on Stereoscopic Displays and Applications*, *1256*, 168-171.

Weinburgh, M. (1995). Gender differences in student attitude towards science: A meta-analysis of the literature from 1970 to 1991. *Journal of Research in Science Teaching*, *32*(4), 387-398.

Wellington, J. (1990). Formal and informal learning in science: The role of the interactive science centres. *Physics Education, 25*, 247-252.

KEY TERMS

3-D: An abbreviation for three dimensions.

CAVE: An abbreviation for **c**ave **a**utomated **v**irtual **e**nvironment.

Edutainment: A term formed by the fusion of education and entertainment, and used to denote experiences that are both educational and entertaining.

Multimedia: A term used to denote the combination of image, sound, and graphics.

Science Center: An institution for the popularization of science and technology.

Stereo Glasses: A kind of goggles that is necessary to see images on a screen in 3-D.

Supercomputer: A very powerful computer that can process large amounts of data at enormous speeds

Virtual Reality: A technology for creating interactive and immersive experiences in cyberspace.

Virtual Teams

Robert M. Verburg
Delft University of Technology, The Netherlands

INTRODUCTION

Global market developments and the large-scale use of diverse applications in the area of information and communication technology have been key factors in the emergence of distributed teams. Such teams are often referred to as virtual teams. Virtual teams enable collaboration between people across traditional boundaries and offer tremendous opportunities for various achievements. Businesses are no longer tied to a single time zone and are, for example, able to develop software around the 24-hour clock. The Internet as the almost universal medium for interaction across boundaries has created an infrastructure that enables many organizations to launch virtual teams. Hardly any technical obstacle for communication and collaboration across geographic boundaries remain as these processes are supported by high tech collaboration solutions, such as groupware and other collaborative applications (e.g., videoconferencing, electronic blackboards). Virtual teams have a number of opportunities that are not found with colocated teams, such as involving rare expertise.

For example, a group of eight scientists from different organizations rapidly developed a revolutionary rocket engine design by working under geographically dispersed conditions and without prior work relationships (Majchrzak, Rice, Malhotra, King & Ba, 2000). The complex and innovative design could not have been developed without the expertise of the eight highly specialized scientists. However, the design was not only a result from a careful combination of expertise but required a number of interdependent iterative "virtual" brainstorming sessions among the team of rocket scientists. All these activities were performed through a collaboration tool called "the Internet notebook" whereby the specialists spend no more than 15% of their time on the project.

As the example illustrates, virtual teams have the advantage of bringing people together without the obvious constraints with regard to travel time, workspace, and socialization. Virtual teams perform a variety of tasks and are also defined in various ways. Martins, Gilson, and Maynard (2004) have defined virtual teams as teams whose members use technology to varying degrees in working across locational, temporal, and relational boundaries to accomplish an interdependent task. Earlier definitions were focused more on making a distinction between virtual teams and conventional colocated teams, mostly based on geographic distribution and mediated communication. Virtual team research is focusing increasingly on real world virtual teams, which often have some virtualness characteristics, but only seldom resemble "pure forms". Therefore virtualness is now widely accepted as being dimensional in nature. More attention is also given to the fact that virtual teams are first and foremost teams, who are carrying out interdependent tasks under difficult circumstances.

BACKGROUND

Being virtual is a matter of degree and refers, according to various authors, to dimensions such as spatial distance, time, cultural diversity, temporality, organizational contract, and mode of interaction (DeSanctis, Staudenmayer & Wong, 1999; Jarvenpaa & Leidner, 1998; Mowshowitz, 1997). Mediated communication is an important dimension. Some teams meet regularly face-to-face, but may have also some e-mail-based interaction, while other teams interact intensively and almost exclusively via various media and sophisticated groupware tools. Geographic distance and different timeframes may obviously be important reasons for groups to communicate electronically.

"Virtuality" refers to the extent to which a group is geographically distributed (Bell & Kozlowski, 2002), and to the extent that team members rely on ICT mediated communication (Dubé & Paré, 2004). Proposed indicators or measures of virtuality are therefore the relation of face-to-face to non face-to-face communication, the average distance between the members, but also the number of working sites represented in the team together with the number of members at each site (see also Kirkman, Rosen, Tesluk & Gibson, 2004; O'Leary & Cummings, 2002). Teams that span large geographic distances between members, will likely encounter additional complicating factors such as cultural diversity, different organizational affiliation, and distribution of members over different time zones. Apart from the above factors, virtual teams are also often associated with shorter life cycles and low member stability.

A useful definition of a team (or work group) is a collection of individuals who see themselves and who are seen by others as a social entity, who are interdependent because of the tasks they perform as members of a group, who are embedded in one or more larger social systems (e.g., community, organization) and who perform tasks that affect others (Guzzo & Dickson, 1996). Although often not defined,

a number of implicit characteristics of conventional teams seem to include that members are often permanent employees of one organization, are often colocated and the main form of interaction consists of face-to-face contact.

Virtual teams may not seem to be crucially different from colocated teams. There are comparable levels of responsibility for adequately performing basic processes of groups, such as information sharing, cooperation, coordination and team building. Virtual teams do also have to mobilize the necessary resources, and need to develop a cohesive team with clear goals. However, virtual teams have to care for these processes under conditions of geographic distribution, which has been found to be significantly and negatively related to work processes and team effectiveness (Cramton, 2005). Inadequate ICT tools or infrastructures and the incompatibility of technology will also result in barriers for cooperation. But with sufficient attention to team building and adequate ICT tools these problems may be overcome. The process of team building can be difficult in the virtual context, specifically when the "life cycle" of a team is short, the stability of membership is limited and face-to-face meetings are scarce. *Global* virtual teams have to deal with the additional issues of communicating across different time zones, languages, and cultures (Montoya-Weiss, 2001).

Other problems may include missing nonverbal cues in communication and a lack of unplanned social encounters, resulting in problems with awareness of availability and state of others, of progress of the work or of the setting in which others work (see e.g., Steinfield, 2002). These barriers may result in a lack of trust and cohesion, which often may lead to lower performance levels. Jarvenpaa and Leidner (1998) confirmed that global virtual teams might start with a form of swift trust (Meyerson, Weick & Kramer, 1996), but that such trust appears to be fragile and temporal. Cramton (1997) illustrates, for instance, the multiple interpretations members of virtual teams may give to the meaning of silence of their distant team members. Additionally, virtual team membership can be highly fluid, demanding for continuous adaptation processes between the existing team and new members, who bring their own beliefs and frame of reference. It is this system of views and beliefs people hold that is often considered very important for team functioning. This system is often referred to as a mental model, which can reflect knowledge and belief systems about members in the team, the teams' task, team interaction processes, and the technology used in the team (Cannon-Bowers, 1993). A high degree of sharedness of mental models has been suggested to lead to more effective teams. However, the distributed nature of, and ICT mediated communication in virtual teams hamper efficient development of shared mental models. Member diversity in organizational affiliation, professional background, and national cultures can complicate matters further.

TEAM PERFORMANCE

A crucial difference between colocated and virtual teams is the fact that virtual teams have the opportunity to combine and integrate both colocated and distributed interaction. Virtual teams may combine the better of two worlds and may therefore have an advantage over conventional teams. Virtual teams require certain tools in the area of information and communication technology (ICT) to support interaction. Some modern tools have sophisticated functionalities that provide such teams with opportunities that conventional teams do not have. One of the major effects of the introduction of collaboration technology has been that certain types of meetings can now be held with a large number of participants. Moreover, some tools allow for easy storage and retrieval of information and for collaborative editing of documents. Research results on the performance of virtual teams, relative to face-to-face teams have been mixed. Virtual teams have been generally found to take more time to complete tasks. However, virtual teams have been found to outperform face-to-face teams on idea generation tasks. In general research results in the area of team performance, and quality of work have been mixed and often contradictory (Martins et al., 2004).

So far, the development of virtual teams has mostly been technology-driven, almost neglecting other aspects of work, such as knowledge sharing, combining expertise, and dividing tasks.

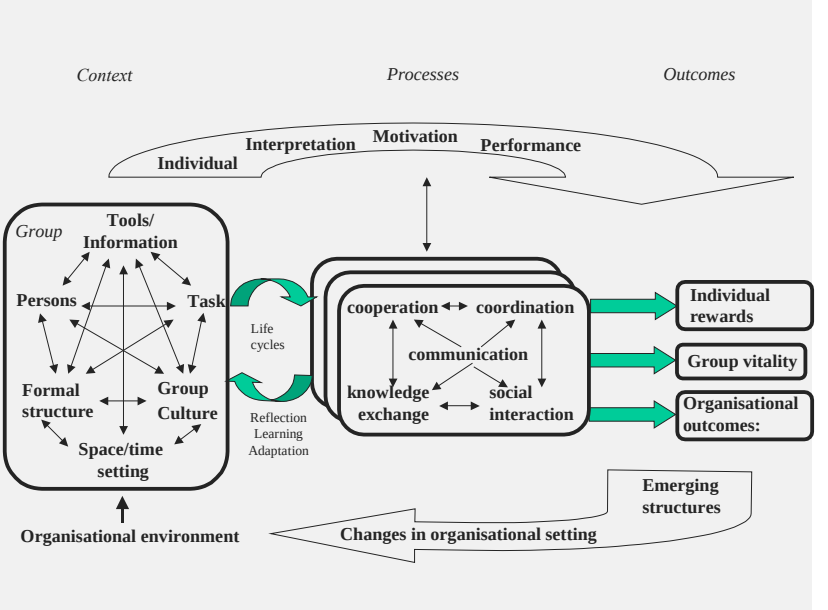
In order to reach an optimal level of functioning, these new types of collaboration require new ways of organizing and managing. Major challenges for both managers and employees are the consequences of dealing with virtual teams. Systematic insight in the design and performance of effective (global) virtual teams is therefore an important prerequisite. It is clear that virtual teams may face substantial barriers for effective cooperation and that the probability of failure is ever present. The next section presents a model for analyzing the reasons for failure and can support the design of virtual groups.

Analyzing Virtual Teams: A Model

The model is based on a general model of group functioning, called the Dynamic Group Interaction model (DGIn-model), which is applied in several case studies (Andriessen, 2002; Andriessen & Verburg, 2004). The purpose of this model is not to limit the analysis of collaborative activities to specific aspects, but to structure the analysis by providing ideas and insights that have proven their value in other contexts.

In this model, elements of several theories are brought together. Three levels of behavior are taken into account, that

Figure 1. Adapted from the Dynamic Group Interaction Model (DGIn -model) (Andriessen, 2002)



is, individual goal directed behavior and cognitive processes (based on Action Theory, Activity Theory, Media Richness Theory), interpersonal and group processes (Activity Theory, Adaptive Structuration Theory, Social Information theory, Coordination theory) and a macrosocial perspective (Structuration Theory). The various notions are brought together in a heuristic model concerning group processes, related to traditional 'input-process-output' schemas (see e.g., Hackman, 1987; Kraemer & Pinsonneault, 1990; McGrath, 1984; McGrath & Hollingshead, 1994). However, they are enriched with interpretative and structurational notions and feedback cycles (see Figure 1).

The DGIn-model has the following four basic principles that can be applied to virtual teams.

Effectiveness

Some virtual groups cooperate only once, so in those cases vitality and continuity of the group as outcomes may not be that interesting for the members. In case of real (project) teams, however, it is not enough to come up with clear results, and with rewards for their members. They also need to cater for trust, group vitality, and continuity in order to be effective. Virtual teams do not differ from conventional teams in this respect. However, developing vitality is more difficult than in colocated groups. High performing teams have been found to be able to maintain high levels of trust throughout the team's life (Kanawattanachai & Yoo, 2002).

The Quality of Group Processes

Six basic group processes were distinguished: communication, and the five other processes that can only exist on the basis of communication (cooperation, coordination, learning, reflection, and team building). These processes need to be aligned.

The type of *communication*–mediated to a smaller or larger extent–constitutes the core dimension for the concept of virtuality. In case of virtual groups, the model implies that collaboration, coordination, knowledge exchange, social interaction and reflection need to be adjusted to the degree of mediation of communication. This is reflected, amongst other things, in the fact that both remote *cooperation* and *social interaction* in mediated meetings need to be much more explicitly structured than face-to-face meetings in order to be effective. The already mentioned lack of non-verbal cues in communication, resulting in problems with awareness of availability and state of others, makes it difficult to interact. Overall, face-to-face meetings allow for more flexibility during meetings and do not need to be as structured as mediated meetings. It is important to provide minutes of virtual meetings as these help to assure that all members understand the same conclusions. In case of virtual student teams, Cramton (1997) showed that team members have difficulty in extracting information about the context in which their distant partners operate, while members themselves often fail to provide important information about their own context.

Globally distributed teams should give sufficient time and attention to group members who are less assertive than most members from a number of Western countries. *Leadership and coordination* of virtual teams play therefore a critical role in facilitating the work and organization of virtual teams (Bell & Kozlowski, 2002). In general the activities of virtual teams appear to need much more preparation and explicit coordination than colocated teams. The role of coordinators is, therefore, vital for the performance of virtual teams (Connaughton & Daly, 2004).

The Quality and Match of the "Context" Characteristics

The quality of group processes depends on characteristics of the "context". Six groups of characteristics are distinguished: the task of the team, tools, member characteristics (knowledge, skills, attitudes), team structure (such as role division and meeting type), culture (norms, trust, cohesion, cognitive distance), and time-space setting (e.g., geographical distribution). The context characteristics need to match each other in order to optimally support the group processes.

ICT support. The technical tools and their usage should be adjusted to the virtuality of the group. The following suggestions can be made:

- Virtual groups require information storage and exchange tools.
- Virtual groups may benefit from a database with information on background and expertise of the group members (*yellow pages*).
- Virtual groups with intensive and nonroutine interaction may benefit from tools for synchronous communication: chat features, where possible video links.
- Virtual groups with complex and time sensitive tasks require workflow management tools for providing information regarding the progress of the project and activities of group members.
- The tools have to be easy to use and equally accessible to all members.
- Group members should be sufficiently trained in remote interaction and in using the technology.
- Global virtual teams should be careful in choosing and using the right tools. Research suggests that people from some cultures prefer direct expression of ideas whereas others may be more sensitive to nonverbal cues and group relations (see for instance Trompenaars, 1993, for examples). The first group of people would probably prefer a collaboration tool that enables synchronous communication, such as telephone, video and chat. People from the other group would be happy with an asynchronous communication tool enabling them to express themselves more carefully. The choice for a suitable collaboration tool to facilitate both groups is, therefore, complicated.

Storage of information. Special attention should be given to information (document) exchange and storage. Effective virtual teams rely heavily on information exchange. Systems and procedures that allow for swift information exchange are therefore a prerequisite. Such systems need to be usable and accepted by all members of the team. In multi-cultural teams such systems are not always easy to obtain. Differences in preferred practices of communication and storing information will limit the choice of an equally useable tool.

Cultural diversity may be large in virtual teams. In order to avoid conflicts and facilitate a smooth work process, group members should be trained to understand the potentially disturbing effect of diversity in national, organizational and professional cultures (Dubé & Paré, 2004). The next step is to learn about each other's background so that differences in solving problems and ways of working will not form a source of major misunderstandings. As soon as members respect and trust distant team members, virtual teams will be able to benefit from the diversity of their members.

Development and Adaptation: Team Building

Groups develop and tools are adopted and adapted, through interpretation, interaction processes and feedback. One of the processes through which this development and adaptation can be explicitly structured is team building. Team building proves to be a critical aspect of team performance and acts as the foundation for the development of necessary levels of trust, cohesion and cognitive closeness among team members. In many cases, team building in virtual teams can benefit strongly from a face-to-face kick off meeting (see Maznevski & Chudoba, 2001, for an overview). Coordinators should be alert to organize such meetings whenever needed or possible.

FUTURE TRENDS

As more and more organizations will explore the opportunities of working across boundaries, the number of virtual teams will increase in the coming years. These teams will experience the complexity of coordination of cross-border work activities. Future research in the area of virtual team analysis will highlight the relationship between virtuality and team effectiveness more closely. Especially, research with regard to the coordination and management of virtual teams will get more attention. So far, research into coordination of virtual teams has primarily focused on the role of leaders of virtual teams (e.g., Cascio & Shurygailo, 2003). Other

possible ways of coordination in virtual contexts, such as the successful use of groupware tools and the role of substitutes for leadership, did not receive much attention yet but will be trends for research in the coming years.

CONCLUSION

Virtual teams offer great opportunities for collaboration across boundaries, which have encouraged many companies to form such teams. However, virtual teams also face challenges, particularly in the areas of communication and coordination. We have presented the DGIn model for team analysis. On the basis of our analysis we recommend that virtual teams should more explicitly pay attention to issues of team building, awareness, preparation, and information storage in order to work and collaborate effectively. Virtual teams should also benefit from the use of specialized groupware tools if applied properly.

ACKNOWLEDGMENT

This enhanced chapter is based on an article authored by Robert Verburg, Erik Andriessen, and Joris De Rooij previously published in the *Encyclopedia of Information Science and Technology*.

REFERENCES

Andriessen, J. H. E. (2002). *Group work and groupware: Understanding and evaluating computer-supported interaction*. London: Springer Verlag.

Andriessen, J. H. E. & Verburg, R. M. (2004). A model for the analysis of virtual teams. In S. Godar & S. P. Ferris (Eds.), *Virtual and collaborative teams: Process, technologies and practice*. Hershey, PA: Idea Group.

Bell, B. S. & Kozlowski, S. W. J. (2002). A typology of virtual teams: Implications for effective leadership. *Group & Organization Management, 27*(1), 14-49.

Cannon-Bowers, J. A., Salas, E., Converse, S. (1993). Shared mental models in expert teams decision making. In N. J. Castellan (Ed.), *Individual and group decision making* (pp. 221-246). Hillsdale, NJ: Lawrence Erlbaum.

Cascio, W. F. & Shurygailo, S. (2003). E-leadership and virtual teams. *Organizational Dynamics, 31*(4), 362-376.

Connaughton, S. L. & Daly, J. A. (2004). Leading from afar: Strategies for effectively leading virtual teams. In S. H. Godar & S. P. Ferris (Eds.), *Virtual and collaborative teams*. Hershey, PA: Idea Group Publishing.

Cramton, C. D. (1997). Information problems in dispersed teams. *Academy of Management Best Paper Proceedings* 1997: 298-302.

Cramton, C. D. & Webber, S. S. (2005). Relationships among geographic dispersion, team processes, and effectiveness in software development work teams. *Journal of Business Research, 58*, 758-765.

DeSanctis, G., Staudenmayer, N., & Wong, S-S. (1999). Interdependence in virtual organizations. In C. Cooper & D. Rousseau (Eds.), *Trends in organizational behavior*. New York: John Wiley.

Dubé, L. & Paré, G. (2004). The multifaceted nature of virtual teams. In D. J. Pauleen (Ed.), *Virtual teams: Projects, protocols and processes* (pp. 1-40). Hershey, PA: Idea Group Publishing.

Guzzo, R. A. & Dickson, M. W. (1996). Teams in organizations: Recent research on performance effectiveness. *Annual Review of Psychology, 47*, 341-370.

Hackman, J. R. (1987). The design of work teams. In J. W. Lorsch (Ed.), *Handbook of organizational behavior* (pp. 315-342). Englewood Cliffs, NJ: Prentice Hall.

Hertel, G., Geister, S., & Konradt, U. (2005). Managing virtual teams: A review of current empirical research. *Human Resources Management Review, 15*, 69-95.

Hutchinson, C. (1999). Virtual teams. In R. Stewart (Ed.), *Handbook of team working*. Aldershot, Hampshire, UK: Gower.

Jarvenpaa, S. L. & Leidner, D. E. (1998). Communication and trust in global virtual teams. *Journal of Computer-Mediated Communication, 3*(4).

Jarvenpaa, S., Knoll, K., & Leidner, D. (1998). Is anybody out there? Antecedents of trust in global virtual teams. *Journal of Management Information Systems, 14*(4), 29-64.

Kanawattanachai, P. & Yoo, Y. (2002). Dynamic nature of trust in virtual teams. *Journal of Strategic Information Systems, 11*, 187-213.

Kirkman, B. L., Rosen, B., Tesluk, P. E., & Gibson, C. B. (2004). The impact of team empowerment on virtual team performance: The moderating role of face-to-face interaction. *Academy of Management Journal, 47*, 175–192.

Kraemer, K. L. & Pinsonneault, A. (1990). Technology and groups: Assessment of the empirical research. In J. Galegher, R. E. Kraut, & C. Egido (Eds.), *Intellectual teamwork:*

social and technological foundations of cooperative work. Hillsdale, NJ: Lawrence Erlbaum.

Majchrzak, A., Rice, R. E., Malhotra, A., King, N., & Ba, S. (2000). Technology adaptation: The case of a computer-supported inter-organizational virtual team. *MIS Quarterly, 24*(4), 569-600.

Maznevski, M. L. & Chudoba, K. M. (2001). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science, 11*(5), 473-492.

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, NJ: Prentice Hall.

McGrath, J. E. & Hollingshead, A. B. (1994). *Groups interacting with technology: Ideas, evidence, issues and an agenda.* London: Sage.

Meyerson, D., Weick, K. E., & Kramer, R. M. (1996). Swift trust and temporary groups. In R. M. Kramer & T. R. Tyler (Eds.), *Trust in organizations: Frontiers of theory and research* (pp. 166-195). Thousand Oaks, CA: Sage Publications.

Montoya-Weiss, M. M. (2001). Getting it together: Temporal coordination and conflict management in global virtual teams. *Academy of Management Journal, 44*(6),1251 – 1263.

Mowshowitz, A. (1997). Virtual organization. *Communications of the ACM, 40*(9), 30-37.

O'Leary, M., Orlikowski, W., & Yates, J. (2002). Distributed work over the centuries: Trust and control in the Hudson's Bay Company, 1670–1826. In P. Hinds & S. Kiesler (Eds.), *Distributed work* (pp. 27–55). Cambridge, MA: MIT Press.

Steinfield, C. (2002) Realizing the benefits of virtual teams. *IEEE Computer, 35*(3), 104-106.

Townsend, A., DeMarie, S., & Hendrickson, A. (1998). Virtual teams: Technology and the workplace of the future. *Academy of Management Executive, 12*(3), 17-29.

Trompenaars, F. (1993). *Riding the waves of culture: Understanding cultural diversity in business*. London: Nicholas Brealey.

KEY TERMS

V

Action Theory: Perspective on action facilitation that makes a distinction between acts, actions, and operations in performing a task. A basic principle of the theory is that the tools used should provide sufficient feedback to allow for adaptation of task execution.

Group Dynamics: Field of inquiry dedicated to advancing knowledge about the nature of groups.

Groupware: ICT applications that support communication, coordination, cooperation, learning and/or social encounters through facilities such as information exchange, shared repositories, discussion forums, and messaging**.**

Media Richness Theory: Theory on mediated communication that highlights the extent to which a medium is capable of sending rich information (i.e., text, smell, pictures, noise, etc.) as well as the proposition that media use is most adequate if the medium is matched with the complexity of the task at hand.

Dynamic Group Interaction (DGIn) Model: In this model elements of several theories with regard to group performance are brought together. Three levels of behavior are taken into account, that is, individual goal directed behavior, group processes and a macrosocial perspective. The various notions are brought together in a heuristic model concerning group processes. They are related to traditional input-process-output schemas.

Structuration Theory: A theory of societal processes on a high abstraction level. Adaptive Structuration Theory (AST) focuses on the analysis of the way existing technologies are taken up by groups and evolve in their role during the appropriation process (i.e., the process of adaptation to new technical tools, which changes the original situation).

Team: A collection of individuals who see themselves and who are seen by others as a social entity, who are interdependent because of the tasks they perform as members of a group, who are embedded in one or more larger social systems (e.g., community, organization) and who perform tasks that affect others.

Virtuality: The extent to which a group is geographically distributed, is organizationally and culturally divers, has different time frames for work, communicates electronically and whose members are freelance or have fixed contracts with an organization.

Virtual Work Research Agenda

France Bélanger
Virginia Polytechnic Institute and State University, USA

INTRODUCTION

The paper by Bélanger, Watson-Manheim, and Jordan (2002) addresses the gap between research conducted and practitioner concerns in virtual work. One of the key difficulties in conducting research in this area is the overlap between terms used (McCloskey & Igbaria, 1998; Pinsonneault & Boisvert, 2001). While there are other distributed work arrangements such as hotelling, neighborhood work centers and flextime, most of the previous literature has focused on telecommuting (telework) and virtual teams/organizations. In this article, the term virtual work represents work environments where individuals spend some time working in a non-face-to-face (FTF) mode, using information and communication technologies to perform work activities.

Virtual work environments are increasingly employed by organizations. While there is increased complexity and potential for problems, virtual work strategies allow organizations a great deal of flexibility to compete in a rapidly changing business environment. While existing research provides insights into such environments, it does not clearly deal with major concerns faced by managers (referred to as the "gap" between research and practice). One of the potential reasons for this gap is that practicing managers are concerned with current challenges in their own work setting while academics are concerned with developing more generalizable rules and understanding.

This article addresses these issues, with three particular objectives:

1. examine the gap between research and practice in virtual work;
2. investigate factors leading to the gap; and,
3. identify a research agenda that addresses emerging issues and concerns relevant to practice in virtual work.

BACKGROUND

To explore the gap between virtual work research and practice, the authors first review previous literature, which they then compare to concerns raised by practitioners in two organizations. To identify relevant academic research, the authors searched for articles in mainstream IS journals. They then used the "snowball" technique, mining citations in articles for further references. They did not include the large number of conference papers and studies of homeworkers, entrepreneurs, or supplemental work at home. They focused on empirical and/or theoretically grounded studies.

Literature Review

Their review of recent literature (1998 to 2001) revealed six literature reviews and 35 empirical studies. In the original paper, a table including methodology details and key concepts was provided but is not included here for brevity, although sample references are provided. Overall, literature addresses the following questions:

- Who are the virtual workers? There are two types of studies that discuss who virtual workers are. The first type is descriptive, usually presenting demographics and other characteristics of virtual workers based on general surveys or public records (e.g., Johnson, 2001). The second type investigates characteristics of telecommuters (e.g., Bélanger, 1999b).
- How is communication influenced by virtual work? This area has been the most researched in recent years. The published work comprises studies of communication patterns (e.g., Bélanger, 1999a) and studies of choices of communication modes (e.g., Wiesenfeld, Raghuram, & Garud, 1999).
- What technologies are used and how do they influence virtual work outcomes? There were few studies prior to 1998 focusing on technologies in virtual work.

Recent studies look at computer-mediated communication systems, the Web, the role of technologies in affecting productivity of teleworkers (e.g., Bélanger, Collins, & Cheney, 2001), and usage patterns of technology in virtual teams (e.g., Majchrzak, Rice, Malhotra, King, & Ba, 2000).

- What is the nature of the work-family conflict in virtual work? There are some recent studies looking at stress but most studies were published prior to 1998 (e.g., Duxbury, Higgins, & Mills, 1998), or in non-IS mainstream journals.
- What are the outcomes of virtual work environments? Most hypothesis-driven studies used outcomes of virtual work, such as productivity, satisfaction, or level of communication, as dependent measures (e.g., McCloskey, 2001). Potential outcomes were also discussed extensively in pre-1998 literature.
- What happens in virtual group work? Studies investigate trust, development processes, and performance in virtual teams, and perceptions in virtual groups (e.g., Maznevski & Chudoba, 2000).
- What are the key issues in managing remote group workers? Studies typically look at issues with managing teleworkers (e.g., Staples, 2001).

Overall, the review showed that a number of barriers, enablers, and outcomes of virtual work have been studied. The samples have often been limited, for example, one organization, which can limit the generalizability of the findings. However, sample size has increased in recent years. In general, given the complexity of organizations, the current research still seems to be narrowly focused.

Case Narratives

To investigate the gap, interviews were conducted in two organizations with distributed workers but with quite different levels of worker distribution. The first organization is Booz Allen Hamilton. It is a global management and technology consulting firm. For their IT practice of the Worldwide Technology Business (WTB), they rely on distributed teams comprised of geographically dispersed employees. The teams are classified as functional, delivery, development, and external teams. Their flexible matrix allows members to participate on multiple teams.

The organization has had success using technology support for team communication. The technologies available include project management, collaboration, and knowledge management tools. Management challenges, however, do occur with practical issues such as the needs for breadth of multi-disciplinary domains, collaboration tools, and training. In managing the breadth of multi-disciplinary domain, identifying the right "mix" of team members with requisite skill sets is a challenge. While electronic collaboration tools are available, managers are not sure whether and how using particular tools makes positive outcomes more likely. Computer and collaboration technology training is left up to each consultant. Management wonders whether project managers are effective because of interpersonal qualities or because of automated tools usage.

The second case is a Fortune 100 telecommunications company headquartered in Southeast USA. It services residential and business telephone customers in nine states. The narrative focused on the management of 700 network service technicians responsible for installation and repair of telephone services within one district. Technicians complete four to five work orders per day assigned by a centralized provisioning center. They are evaluated based on efficiency in completing orders and quality of the work performed. Some teams are staffed in shifts 24 hours/day, seven days/week, while others have eight-hour days with frequent overtime.

Technicians and supervisors use information and communication technologies extensively, including cell phones, pagers, and wireless laptops. The first work order is loaded on the technician's laptop before the work day begins. Technicians update the work order, and completed orders are updated in the system as soon as a worker establishes a connection. A new work order is then assigned. Such a system allows dynamic assignment of work orders based on changing priorities throughout the day.

Managers are responsible for overseeing eight to 15 technicians, including visiting and inspecting the site where work was completed. Supervisory duties also include providing training for technicians (formal and informal). They must respond to individual questions which arise due to unique field conditions or changes in technology. In addition, they must conduct performance evaluations, counsel technicians on career development, and mentor technicians new to the job.

Such responsibilities pose challenges for managers. Management challenges include work activity coordination, measurement tools, training, and information sharing for team building. Coordinating activities of distributed field workers who face process changes, often due to field conditions, is challenging. Supervisors are not able to

Table 1. Gaps and overlap in virtual work research and practice

Practitioner Issues	Examples of questions of interest	Overlap*	Examples of questions/areas researched	Research in Virtual Work
Team building	What factors are critical for building and maintaining a solid team culture and effective communications process in distributed environments?		What is the effect of virtual work on communication patterns and structures in teams and work groups?	How is communication influenced by virtual work?
	What communication structures and mechanisms can be used to distribute information and coordinate tasks among team members in timely manner?		What effects does telework have on group communication structures?	How is communication influenced by virtual work?
Organizational/ management structure	Does management role change in distributed environments?		What factors lead to more successful management of remote workers?	What are key issues in managing remote workers?
	What role does distance play in determining organizational structure and job design?			
	How are resources best allocated in distributed environments?			
Information sharing & distribution	How can information sharing be facilitated in distributed environments?		How is communication and coordination performed in virtual work?	How is communication influenced by virtual work?
	How is time-sensitive information best distributed?			
	What are effective interpersonal networking techniques used by successful distributed workers?		How is communication in teams affected by virtual work?	How is communication influenced by virtual work?
Employee assessment & development	What are the best methods and metrics for employee assessment in virtual work?		Are demographics of teleworkers linked to success or other outcomes?	Who are the virtual workers**?**
	What is the role and appropriateness of employee monitoring in this environment?			
	How is the critical management function of employee development best performed in distributed work?			
	In distributed work teams that do not meet daily, are there differences between the factors considered in team members' evaluations and those of co-located teams?		How can trust be developed between distributed team members?	What happens in virtual group work?
	Should some performance evaluation factors be weighted more heavily than others in virtual work? Which?			
Work process training	What skills need to be developed for different types of virtual work environments?		What are the characteristics of individuals performing virtual work, including skills?	Who are the virtual workers?
	How can we train distributed workers for optimum job knowledge sharing and work in virtual teams?			
IT training and readiness	What IT training is needed and how is it most effectively employed in distributed work?			
	At what stage are employees (and org.) in their acceptance of and readiness to use information and communication tools?			
Communication tools & technology choice	How does management assess the appropriateness of available communication technologies and applications?		Which communication tools lead to greater success in telework?	What technologies are used and how do they influence virtual work outcomes?
	What tools are available, and which are best to support collaborative distributed work?		Which communications tools are available and used in virtual work?	What technologies are used and how do they influence outcomes?
	How does management assess the effectiveness of collaborative tools, where they are successful, and under what conditions?			
				What is the nature of work-family conflicts in virtual work?
				What are outcomes of virtual work?

* Filled cells indicate overlap. White cells indicate limited or no overlap.

observe the work performed, which causes difficulty in providing feedback to technicians on the quality of their work. Training in a timely and consistent manner is difficult with staggered schedules of field workers. In addition, it is difficult for distributed workers to develop relationships with team members. Thus, inability to share information with team members, especially for newer people, poses a challenge for managers.

MAIN TRUST OF THE ARTICLE

A number of managerial themes emerged from the interviews. The themes are summarized in Table 1, and only a few sample questions are provided in the following text for each theme. Please refer to Table 1 for further research questions.

Team building is an issue due to the complexity of lateral communication among team members in virtual work environments. The lack of physical interaction causes difficulty in creating relationships. An example of a question practitioners need answered is: What factors are critical for building and maintaining a solid team culture and effective communication process in distributed work?

Hierarchical communication between employees and managers also becomes more complex in virtual work. Coordination of unpredictable tasks is difficult for supervisors. In relation to organizational management structure, practitioners need answers to issues like: How are resources best allocated in the distributed environment?

Communicating organizational information in a timely manner is a challenge. Information sharing among team members may be difficult since work teams are physically distributed. Timely updates are difficult without methods for sharing information such as posting system messages and alerts for fellow workers. A sample question of interest to practitioners could be: What are effective interpersonal

networking techniques used by the most successful teleworkers?

Measuring and monitoring employees' work is complex in distributed work. Better measurement criteria on work patterns or processes would enable management to more fairly assess and provide effective feedback to employees. This leads to questions like: What are the best methods and metrics for employee assessment in virtual work?

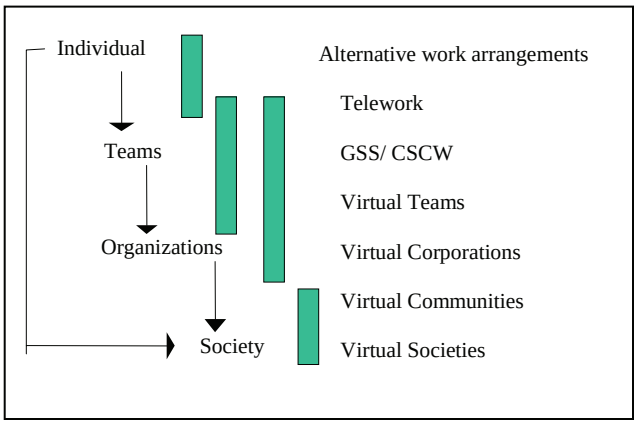
Development and training of employees for future assignments can be difficult, particularly for new employees in distributed settings. A sample question that needs further investigation is: How are the critical management functions of employee development best performed in the distributed environment?

Ensuring that employees are performing work activities most effectively is difficult in distributed work. Tension occurs between organizational consistency and employee independence. This area has just started to be researched by academics (Cooper & Kurland, 2002; Kurland & Cooper, 2002). A sample question needing further investigation in this area is: How can we train distributed workers for optimum job knowledge sharing and work in virtual teams?

The use of information and communication technologies is critical to effective performance in distributed settings. However, the complexity and rapid change of technologies adds to the challenge of providing training at a distance. An example of a question of interest is: How is necessary IT training most effectively conducted in distributed settings?

Choosing the best communication technology is critical. Management in both organizations interviewed, however, had little guidance in how to make the choice. A typical question of interest could be: What tools are available, and which are best to support collaborative work in distributed settings?

Figure 1. Proposed virtual work units of analysis



FUTURE TRENDS

A summary of the issues raised in the case narratives as compared to the literature is presented in Table 1. It provides a high level view of potential gaps between virtual work research and practice. In general, the literature does not always adequately capture the complexity of virtual work environments, which creates a gap between managerial concerns and academic research. In addition, topics addressed by research are not always addressed in as much depth as what is needed by practitioners. For example, researchers look at coordination and communication in general while practitioners are interested in how communication within virtual work can be better used for information sharing, information distribution, performance feedback and/or relationship development. Some gaps, however, may be justifiable. For example, work-family conflict issues should be studied by researchers, but do not seem to be major concerns of practitioners. The need to address longer-term issues about societal effects of virtual work may justify research in this area. One overlap is team building where issues of trust and communication in virtual teams are researched and are seen as practical concerns by managers. Another overlap is evaluation of tools and technologies for virtual work. Research has focused on e-mail and the Web, also important to managers. In addition, managers are interested in groupware and knowledge management tools.

While exploring the gap, several factors that might cause a disparity between research on virtual work and concerns of practitioners became apparent. Possible reasons for the gap include multidisciplinary nature of managerial concerns, time-intensive requirements for research methodologies, and lack of proper definition of the unit of analysis.

Virtual work research is fragmented by areas while problems faced by managers are multi-disciplinary in nature. Business organizations require a more systemic and holistic approach to studying virtual work. For example, issues of technology and organizational communication cannot be separated in virtual work. Understanding interpersonal relations is critical to understand how relationships are formed and maintained in an environment where cooperating individuals are working in different contexts with different technologies.

Longitudinal case studies or multiple case studies are the most appropriate research methodologies to study virtual work. The type of research needed requires substantial

time investments from the researchers' perspective. It is, therefore, more difficult to accomplish this research and get the appropriate rewards of timely publications.

The unit of analysis needs to be appropriate to the research question, and the research question should be relevant to the "unit" being studied. A proposed view of the appropriate unit of analysis in virtual work research adapted from Agres, Edberg, and Igbaria (1998) is shown in Figure 1. For example, the organizational level unit of analysis could include studies of telecommuting, GSS/CSCW, virtual teams or virtual corporations.

CONCLUSION

Through an in-depth review of virtual work literature and insights from two organizations, a gap between research and practice was identified since practitioners are faced with issues and challenges in virtual work environments for which research does not always capture the complexity. Although possible reasons provided for a gap between research and practice is not exhaustive, it is apparent that a need exists to have better communication between practitioners and researchers on issues of importance and on how each can benefit from one another's work. As academics, we should consider these as opportunities to perform research of importance to both the academe and practitioners.

REFERENCES

Agres, C., Edberg, D., & Igbaria, M. (1998). Transformation to virtual societies: Forces and issues. *Inform Soc, 14*, 71-82.

Bélanger, F. (1999a). Communication patterns in distributed work groups: A network analysis. *IEEE Transactions on Professional Communication, 42*(4), 261-275.

Bélanger, F. (1999b). Workers' propensity to telecommute: An empirical study. *Information and Management, 35*(3), 139-153.

Bélanger, F., Collins, R., & Cheney, P.H. (2001). Technology requirements and work group communication for telecommuters. *Information Systems Research*.

Bélanger, F., Watson-Manheim, M.B., & Jordan, D.H. (2002). Aligning IS research and practice: A research agenda for virtual work. *Information Resources Management Journal, 15*(3), 48-70.

Cooper, C.D., & Kurland, N.B. (2002). Telecommuting, professional isolation, and employee development in public and private organizations. *Journal of Organizational Behavior, 23*, 511-532.

Duxbury, L.E., Higgins, C.A., & Mills, S. (1998). After-hours telecommuting and work family conflict: A comparative analysis. *Information Systems Research, 3*(2), 173-190.

Johnson, N.J. (2001). Case study of the St.Paul Companies' virtual office for the risk control division. In N.J. Johnson (Ed.), *Telecommuting and virtual offices: Issues and opportunities* (pp. 148-161). Hershey, PA: Idea Group Publishing.

Kurland, N., & Cooper, C. (2002). Manager control and employee isolation in telecommuting environments. *Journal of High Technology Management Research, 13*, 107-126.

Majchrzak, A., Rice, R.E., Malhotra, A., King, N., & Ba, S. (2000). Technology adaptation: The case of computer-supported inter-organizational virtual team. *MIS Quarterly, 24*(4), 569-600.

Maznevski, M.L., & Chudoba, K.M. (2000). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science, 11*(5), 473-492.

McCloskey, D.W. (2001). Telecommuting experiences and outcomes myths and realities. In N.J. Johnson (Ed.), *Telecommuting and virtual offices: Issues and opportunities* (pp. 231-246). Hershey, PA: Idea Group Publishing.

McCloskey, D.W., & Igbaria, M. (1998). A review of the empirical research on telecommuting and directions for future research. In M. Igbaria & M. Tan (Eds.), *The virtual workplace* (pp. 338-358). Hershey, PA: Idea Group Publishing.

Pinsonneault, A., & Boisvert, M. (2001). The impacts of telecommuting on organizations and individuals: A review of the literature. In N.J. Johnson (Ed.), *Telecommuting and virtual offices: Issues and opportunities* (pp. 163-185). Hershey, PA: Idea Group Publishing.

Staples, D.A. (2001). Making remote workers effective. In N.J. Johnson (Ed.), *Telecommuting and virtual offices: Issues and opportunities* (pp. 186-212). Hershey, PA: Idea Group Publishing.

Wiesenfeld, B.M., Raghuram, S., & Garud, R. (1999). Communications patterns as determinants of organizational identification in a virtual organization. *Organization Science, 10*(6), 777-790.

KEY TERMS

Computer Supported Collaborative Work (CSCW): Research area that focuses on investigations and development of technologies that can be used for collaborative work in distributed settings.

Group Support Systems (GSS): A set of technologies used to help groups in their decision making processes.

Hotelling (Neighborhood Work Center): Organizational facility for employees to work at but where they do not have a permanently assigned desk. They must "check-in" every time they come to work there.

Longitudinal Case Studies: Research method that involves looking at particular cases over a longer period of time, with repeated measures to observe a phenomenon as it evolves.

Telecommuting (Telework): Work arrangement that allows employees to work at home during regular work hours.

Virtual Teams/Organizations: Teams and/or organizations where some or all of the members work from different physical locations.

Virtual Work: Work environments where individuals spend some time working in a non-face-to-face (FTF) mode, using information and communication technologies to perform work activities.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 3013-3017, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Virtual Work, Trust and Rationality

Peter Murphy
Monash University, Australia

INTRODUCTION

Since the development of the Internet—and the emergence of computer networking as a mass medium in the mid-1990s—many organizations and institutions have experimented with Internet protocol (IP)-based communications to coordinate work and activities across geographical distance. This has been in response to growing needs to coordinate business and projects between different offices, firms, regions, and states. Rather than organizations flying people to meet face-to-face, network technology presents opportunities for persons located apart to work together. It offers the potential for cheap and efficient collaborations across distance. Yet, while economic pragmatics drive organizations to adopt virtual work methods, virtual working is difficult to implement. This is because it strains many conventional assumptions about work behaviour and the cognitive and emotional foundations of collaboration.

BACKGROUND

Since the 1970s, there has been a general trend worldwide for organizations to move from being closed systems to open systems. This has involved growing pressures on organizations to interact with their environment rather than trying to internalize their environment. The most visible consequences of this have been the escalating tendency of organizations to contract out functions, to relocate parts of their operations across the world, and to grow the number of strategic collaborations with other organizations. The result is more and more organizational actors working with persons—often persons they do not know—in other locations. Working with people at a distance means working virtually (Duarte & Snyder, 1999; Franke, 2002; Igbaria & Tan, 1998; Jackson, 1999; Kisielnicki, 2002; Lipnack & Stamps, 2000; Mowshowitz, 2002; O'Hara-Devereaux & Eccles, 1994). Virtual collaborators (teams and partners) have no shared physical presence. Collaborators may see one another only rarely if at all.

The technologies of virtual collaboration are relatively straightforward: e-mail, ftp, collaborative groupware, and audio-video conferencing. Groupware and IP-based conferencing is still relatively under-utilized. Third-party hosted groupware offers solutions to high-level collaboration across firewalls. IP-based conferencing provides opportunities to enrich interactions with sound and visuals. Groupware to date, however, does little more than make conventional file storage and threaded discussion available to persons working in multiple locations across organizational boundaries. Conferencing software is only beginning to be able to deliver quality audio across low bandwidth connections. Typically, high-quality video and the sharing of complex software applications still require high network bandwidth, and are often unavailable from roaming and non-institutional locations.

While technology shapes the possibilities of virtual interactions, psychology is a more powerful factor in determining the viability of such interactions. A basic condition of virtual collaboration is the ability to work with others without seeing them, knowing them, or meeting them in person. While technology can enable such work, to effectively leverage these technological possibilities, organizations have to adapt themselves to different ways of working, and in some cases they have to re-invent themselves. Working virtually at the micro-level of teams, groups, and pairs is only effective where the larger organizational environment lends itself to virtual interaction.

There are three basic types of organization: social, procedural, and the virtual or self-organizing (Miller, 2002). Social organizations are the most common type. These are based on face-to-face interactions and on character norms such as loyalty, dedicated service, and "keeping your word". Procedural organizations are built on impersonal roles and rules. Virtual organizations are structured around more abstract patterns and forms. The family firm and the relationship-driven Japanese corporation are examples of the social organization (Fukuyama, 1995). The Fordist-type American corporation typifies the procedural kind (Chandler, 1977). In contrast, production and distribution reliant on intangible or intellectual capital, such as licensing, patents, or correspondence, encourages forms of virtual collaboration based on high degrees on self-organization (Barley, Freeman & Hybels, 1992).

In order to be effective, any organized human activity must be rational. Rationality is another word for continuity, identity, and stability of expectation. Organizational behaviours deteriorate or collapse if the members of an organization cannot see that these behaviours are for the most part rational. The emotional correlate of rationality is trust. What is felt to be reliable, and worthy of trust, is

also that which is recognized to be rational. Any organization where people trust one other is more effective than an organization where persons are suspicious of each other (Kramer & Tyler, 1996).

In social organizations, people "with character" are generally recognized as rational actors. These might be persons who are dependable, loyal, and unwavering in their treatment of each other. Through demonstrating that they are good at following social norms, such agents generate trust (Fukuyama, 1995; Handy, 1995). With the development of equity corporations and modern management in the late nineteenth century, social organizations in many places were replaced at least in part by procedural or bureaucratic organizations (Chandler, 1977; Yates, 1989). These developed around rules, roles defined by rules, procedures, work demarcations, impersonal written communication, and file management. Knowledge of rules rather than of people provided organizational continuity, identity, and stability. Thus persons who were consistent at following and applying rules acquired reputations for trustworthiness. Predictability in decision-making and task execution became the primary source of trust in bureaucratic organizations—complementing and often superseding the loyalty and patronage work cultures of social organizations.

Virtual work does not follow the logics of either social or procedural organizations. Without face-to-face interaction, character norms cannot be the basis of organized action. At the same time, procedural rules are difficult to agree on, to follow, or to enforce because virtual collaborators do not share the same office, organization, or manager. Virtual actors have to deal with multiple rule sets across diverse institutions, geographies, and cultures. Under these conditions, rules become ambiguous, conflicted, and uncertain. One party's rationality becomes another's irrationality. Such conflicting expectations breed distrust. Thus, under virtual conditions, rationality and trust have to be generated by other means (Murphy, 2003).

CRITICAL ISSUES

Because there is not the same history of working virtually as there is of working socially or working procedurally, identification of the means by which virtual partners and teams generate rationality and trust is less developed. If virtual collaborators cannot rely on personal moral character or on impersonal "rules and roles" to facilitate their interaction, then what can they rely on? The simplest answer is that, in the absence of social cues or clear-cut procedural direction, persons working have to be self-organizing. The key to successful self-organization is the sense of pattern or designing intelligence. Where self-directed activity (Ray & Bronstein, 1995) dominates cooperative and peer interaction, design intelligence and pattern rationality function as the coordinating medium of organized activity and group behaviour. If not, collective cohesion readily collapses.

Human beings have a strong design sense. They pick up exceptionally quickly on design characteristics such as rhythm, harmony, and proportion. Pattern recognition is central to brain processing (Davies, 1992). For instance, we use our pattern sense to make judgments about regular sentences, trustworthy buildings, and reliable machines (Alexander, 1977; Fodor & Pylyshyn, 1988; Gelernter, 1998). Such pattern rationality is also conducive to building trust. Patterns generate feelings of surety, satisfaction, and reliability. This applies as much to work environments as to cities, machines, or sentences. To create patterns, organizations employ tacit forms of aesthetic cognition (Calas & Smircich, 1996). Aesthetic cognition uses beauty, elegance and economy rather than rules or roles to achieve its ends.

Successful virtual work is conducted like a design process (Murphy, 2003). It relies less on the passing around of overt messages, and more on the ability of collaborators to understand through the exercise of imagination where their part "fits" into the overall design of the workflow. "Fit" is achieved by thinking in aesthetic terms of proportionality, rhythm, and harmony rather than in terms of rules or roles. The rationality of a virtual organization is not the rationality of character or procedure but of design. Much of this "acting by design" is intuitive or unspoken. It rests on imaginative cognition. Persons who work virtually by necessity cannot talk a lot or interact a lot with each other—so they need to imagine a lot. They need to be good at projective or anticipatory thinking. This projective thinking is not the same as the anticipatory thinking involved in either relationship empathy or in Gantt chart style project management. Rather, it is much more figurative in nature. The virtual collaborator who uses imagination is good at "seeing the shape of things" in lieu of dense social relationships or strong procedural guidance.

Virtual team or partnership work relies heavily on imaginative visualization and intuition. This is a kind of tacit knowledge. It is tacit in the sense that it involves picture thinking and pattern cognition rather than verbalization. It requires the cognitive-psychological capacity to "figure" things out (Mintzberg & Westley, 2001). Such cognitive figurative methods are closer in kind to processes of creative design than they are to processes of social recognition. In this context tacit does not mean the implicit understanding we derive from the warm handshake or the disapproving stare of another person. The tacit nature of figurative work methods thus are different in nature from the tacit knowledge that we draw from the bodily presence of collocated work partners. In the case of the imagination, tacit refers to high levels of picture-like abstraction. At the same time, however, because many aspects of this design intelligence operate non-discursively, the imaginative abstraction that is required in virtual working is quite unlike the explicit rules

of procedural organizations or the rule-driven inferential reasoning typical of procedural rationality.

Virtual work elides socio-emotive contents and makes conventional discursive (mile-stone) office planning difficult to implement. For virtual work to be successful, even on a micro-scale, it must draw heavily on the faculty of imaginative and figurative thinking. To be proficient at cooperation and interaction, virtual workers must be able to picture what is absent and organize that picture "aesthetically".

In virtual collaborations, designing intelligence generates shared integrative schemas—such as asynchronous rhythms of interaction or proportionate distributions of task load. Where virtual teams and collaborators "find their rhythm," they will—more likely than not—produce good work. Correspondingly, it is the act of producing good work that builds trust amongst virtual collaborators. Trust arises where collaborators know that each will "do their part". "Doing their part" does not mean sending lots of social messages, nor does it mean showing procedural fluency. Virtual trust is generated not through relationships or procedures but rather through the aura of reliability that arises from the visible effects of the "invisible" qualities of beauty, elegance and economy that virtual actors employ when they cooperate at a distance to produce things in imaginative ways. This means producing well-designed objects—be they physical goods, processes, systems, or learning objects—in a "choreographed" manner. The choreography of the virtual team rests not on social gestures or on rules but on a good sense of rhythmic or harmonic "fit". The sense of satisfaction and surety derived from such "fit" provides the emotional trust that is the counterpart of aesthetic rationality.

FUTURE TRENDS/CONCLUSION

As business and government are pressed to operate over increasingly large scales and long distances, and as inter-organizational, inter-agency, and inter-state activity becomes more common, the need for distance communications and virtual teams is growing and a class of virtual workers is gradually emerging.

All the while, established paradigms of work remain entrenched, meaning that there is a latent propensity to try and build virtual teams and partnerships around social communication and procedural norms. This creates conflict between the intrinsic nature of virtual cooperation and the extrinsic goal extending the reach of traditional organizational structures. As organizations at the micro-level of team and peer relations continue to expand their geographical scope, and as the boundaries of organizations become increasingly fluid as a result, a major challenge for the future will be increasing the understanding of the role of aesthetic rationality and designing trust in the formation of productive relations between actors who are separated by distance and time.

REFERENCES

Alexander, C. (1977). *A pattern language.* Oxford: Oxford University.

Barley, S.R., Freeman, J., & Hybels, R.L. (1992). Strategic alliances in commercial biotechnology. In N. Nohria & R.G. Eccles (Eds.), *Networks and organizations: Structure, form, and action.* Boston: Harvard Business School Press.

Calas, M., & Smircich, M. (1996). Essays on aesthetics and organization. *Organization, 3*(2).

Chandler, A.D. (1977). *The visible hand: The managerial revolution in American business.* Cambridge, MA: Harvard University Press.

Davies, P. (1992). *The mind of God: The scientific basis for a rational world.* New York: Simon & Schuster.

Duarte, D., & Snyder, N. (1999). *Mastering virtual teams.* San Francisco: Jossey-Bass.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28.* Lausanne: Elsevier.

Franke, U. (2002). *Managing virtual Web organizations in the 21st century.* Hershey, PA: Idea Group Publishing.

Fukuyama, F. (1995). *Trust: The social virtues and the creation of prosperity.* New York: Free Press.

Gelernter, D. (1998). *Machine beauty: Elegance and the heart of technology.* New York: Basic Books.

Handy, C. (1995). Trust and the virtual organization. *Harvard Business Review, 73,* 3.

Igbaria, M., & Tan, M. (1998). *The virtual workplace.* Hershey, PA: Idea Group Publishing.

Jackson, P. (1999). *Virtual working.* New York: Routledge.

Kisielnicki, J. (2002). *Modern organizations in virtual communities.* Hershey, PA: IRM Press.

Kramer, R.M., & Tyler, T.R. (1996). *Trust in organizations.* Thousand Oaks, CA: Sage.

Lipnack, J., & Stamps, J. (2000). *Virtual teams: People working across boundaries with technology.* New York: John Wiley.

Miller, K. (2002). *Organizational communication.* Belmont, CA: Wadsworth.

Mintzberg, H., & Westley, F. (2001). Decision making: It's not what you think. *MIT Sloan Management Review, 42,* 3.

Mowshowitz, A. (2002). *Virtual organization.* Westport, CN: Quorum.

Murphy, P. (2003). Trust, rationality and the virtual team. In D. Pauleen (Ed.), *Virtual teams: Projects, protocols and processes.* Hershey, PA: Idea Group Publishing.

O'Hara-Devereaux, M., & Johansen, R. (1994). *Global work: Bridging distance, culture, and time.* San Francisco, CA: Jossey-Bass.

Ray, D., & Bronstein, H. (1995). *Teaming up: Making the transition to a self-directed, team-based organization.* New York: McGraw-Hill.

Yates, J. (1989). *Control through communication: The rise of system in American management.* Baltimore: Johns Hopkins University Press.

KEY TERMS

Design: The structured composition of an object, process or activity.

Distance Communication: Communication under conditions of geographic separation that minimize the possibility of face-to-face and synchronous interactions.

IP-Based: Network technologies based on Internet protocols.

Open System: Any system without strong boundaries, where information and other goods flow to and from the system's environment.

Organization: The deliberate integration of persons in order to achieve a goal or outcome.

Procedure: A rule that governs a formal organizational process.

Rationality: The ability to infer with relative certainty from existing or past behaviour and statements future behaviour and statements.

Trust: Confidence that an object, process, institution, or another person's actions can be relied upon to produce some good.

Virtual Interaction: Computer-mediated interaction between persons who do not occupy the same physical space.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 3018-3021, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Virtualization and Its Role in Business

Jerzy A. Kisielnicki
Warsaw University, Poland

INTRODUCTION

A new management trend of the global information technology (IT) application—virtualization—has appeared in the contemporary management. Virtualization is a process of enterprise transformation (using IT) that allows breaking through various limitations of organizational constraints. Virtualization changes dramatically the image of business, especially of small and medium enterprises (SMEs); by adopting the concept of virtualization, they can become fully competitive and may effectively operate in the global market. Barriers of the scale between SMEs and large organizations disappear. This new type of organizations is often called in literature *modern organization* or *virtual organization*. Organizations of this type have an effective decision-making process, and function based on economic criteria. Consequently, their opportunities to grow and to compete in the global market are greater than for traditional SMEs. Hence the thesis that virtualization allows individual organizations to enter strategic co-operative alliances with other similar businesses. Such of virtual organizations have a competitive position in the global market.

In the literature, there are many terms used to define virtual organization: "network organizations" (Drucker, 1988, p. 9), "organizations after re-engineering" (Hammer & Champy, 1993, pp. 77-79), "crazy organization," "crazy time for crazy organization" (Peters, 1994, pp. 5-7), and "intelligent enterprise" (Quinn, 1992, p. 3).

BACKGROUND

Virtualization, defined as a process of continuous transformation, is a herald of a new direction in the science of organization management. In the context of this analysis, this process may assume such form that will allow them to become competitive in the global market. The process of transformation consists of quick adjustments of the enterprise to new requirements (Hendberg, Dahlgren, Hansson, & Olive, 2000). This is done through changes in the organizational structure as well as in the portfolio of products and services. These changes are possible due to development in the IT sector, particularly Internet applications (Kenny & Marshall, 2000).

From the theoretical perspective, we can separate the following forms of virtualization:

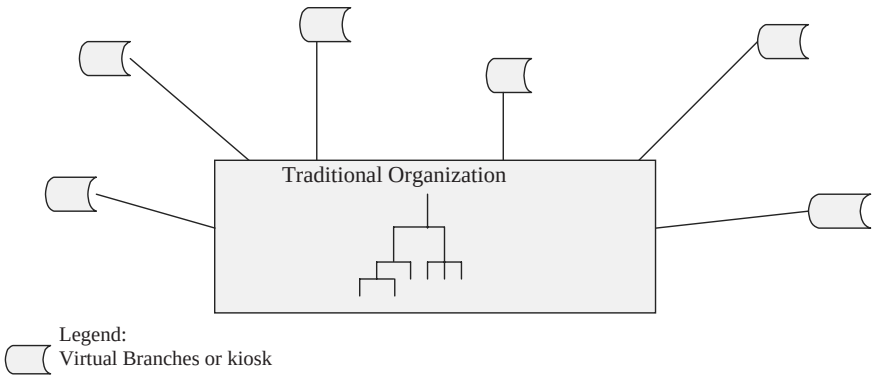
1. Functional extension (i.e., a vertical development). This occurs when the enterprise either wishes be closer to the customer and it does not have adequate resources or when the establishment of a traditional branch is not profitable. The enterprise creates for this purpose virtual branches or *kiosks*. Sometimes it enables their customers to use its services via computer or mobile phone. Examples are Internet banks, bookshops (best known is amazon.com), department stores, and travel agencies. Large companies also commonly extend their scope through such vertical development. It ensures increased competitiveness with a simultaneous control over the whole organization. SMEs apply such a strategy to a limited extent, most often for the purpose of marketing their presence in the Internet.
2. Creation of the virtual organization, or the horizontal development. Such a development occurs through a virtual incorporation of other organizations. The literature lacks a unanimous definition of this concept (Hendberg et al., 2000; Kisielnicki, 1998; Quinn, 1992; Scholzch, 1996).
3. Specialist structures being created in order to collaborate. In physical terms this is a computer or a network of computers equipped with specialist software. This form of virtualization is used for raising qualifications of the personnel by both SMEs and large enterprises (Fong, 2005).

For the purpose of this analysis, we assume that:

Virtual organization is created when its members voluntarily enter in relations of various types to achieve their common goal. Every member who creates this organization defines duration of the relation. The member who first admits that the existence of that relation is unfavorable, makes the decision on its liquidation, and withdraws. The virtual organization operates in the *cyberspace* and requires existence of the Internet and global IT infrastructure.

LSEs use IT to strengthen their competitive position in relation to other enterprises. As Hammer and Stanton (1999) rightfully notice, IT becomes—for a certain class of organizations—a "wall" that divides them from other enterprises. Large enterprises are described as "castles" (Hammer & Stanton, 1999). LSEs build these "castles" to protect themselves from competition. SMEs do not have such a sheath. They are more flexible than LSEs at a price: they are more prone to infiltration. In general, however, the

Figure 1. Virtual organization based off traditional structure



more experienced and knowledgeable the SMEs are, the more attractive they are to other enterprises seeking their share in the virtual enterprise.

MAIN THRUST OF THE ARTICLE

Virtualization in Traditional Organizations

Virtualization allows for organizational development at much lower cost than through the traditional process. The following diagram presents a virtual organization based off a traditional structure and connected with its virtual elements. These elements are flexible, allowing the organization quick adjustment to changing environments.

The virtualization may be developed as follows:

- The organization creates virtual kiosks or shops. For example, an organization that sells furniture using the Internet may present its products on a computer screen. It may also receive orders from customers located outside their traditional market. Other organizations (tourist, real estate, bookshops, stock exchange, etc.) may operate in the same manner.
- An organization places information about its activity on the Internet (Porter, 2001). It is available to its clients 24 hours a day, seven days a week, rather than being limited to office hours. Internet banking services are a good example. In addition, the expenses connected with the services of an organization are shifted to the client (the client pays for the terminal, access to the Internet, etc.). The organization covers the cost of the development and maintenance of an application.
- The organization creates a possibility of working from home (called Tele-work). It is a good way of increasing professional activity in those regions where it is difficult to find "traditional" employment. Also, through a Tele-work, a local organization may become a global one.

Virtualization allows traditional organizations to have a wider range of influence. Society is better informed of the organization's activities both by the organization itself and by its clients. The restrictions on this development are varied. The most common include available financial resources for IT infrastructure, language, and a necessity to have global, reliable computer networks. It should also be stressed that, unfortunately, virtualization enables organizations that are socially unaccepted (pornography, terrorism) to operate freely.

Based on the analysis of organizations that use virtualization, it is estimated that their operations require five times lower investment outlays. Generally, minimum savings obtained as a result of virtualization exceeded 60%. Only in a few cases, this proportion was less favorable.

In the organizations using Tele-work, the proportions are difficult to calculate. The analysis of organizations using Tele-work for outsourcing their services to developing countries confirms the effectiveness of virtualization. A good example is software development. The companies from highly developed countries (U.S., Great Britain, etc.) employ programmers from India, China, or Pakistan. This situation is beneficial for both the company and the countries providing resources.

A different situation occurs when Tele-work is connected with professional activation of the disabled or unemployed. Direct costs are higher as we deal with the poorer part of the society. Thus additional costs have to be incurred for training, hardware, and software. Unfortunately, there is no data available to make a precise estimate. In many countries, the cost of training and equipment is covered by special social programs. It is also difficult to estimate advantages. It may be said that social effect (i.e., decreased unemployment)

and—in case of the disabled—the ability to live a normal life is the most important one. This is possible only through virtualization.

Generally, Tele-work reduces the operating costs (parking space, office space) and provides timesaving (i.e., no travelling to work) to employees.

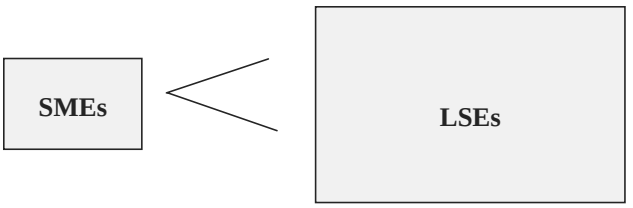
VIRTUAL ORGANIZATION AS A CHANCE FOR SMEs

Virtualization allows more and more SMEs to leave local markets and become global enterprises. Thus, the global market, for many years available only for LSE, opens up to a wide range of enterprises. As Yip (2004, p. 27) states, "the skill of defining and implementation a global strategy is a true test of leadership and management skills." The contemporary global market most often exists as e-market as well as e-business (Turban, King, Lee, Warkenting, & Chung, 2002). The enterprise, irrespective of its size, that wants to make its existence in the global market, should meet a number of conditions, such as:

1. Possessing a well-known and reputable brand
2. Built-up distribution and service network
3. Product or services that are unique and in demand
4. Management team able to support the global enterprise

To meet these conditions, the enterprise must have at its disposal adequate funds as well as material and human resources. Comparing SMEs with LSEs, the first are in a very difficult situation. The enterprise, to make its appearance in the global market, must incur definite outlay (a break-even point) to enter the global market. This break-even point is determined by:

Figure 2. Comparison of individual SMEs and LSEs (LSEs have a competitive advantage over individual SMEs)

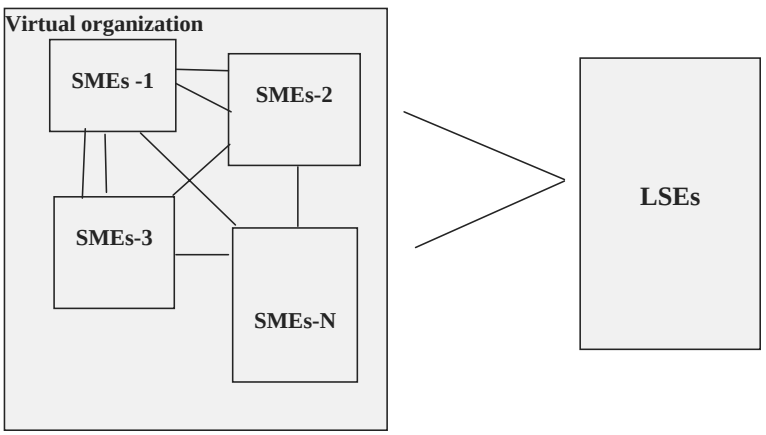


1. Advertising and promotional campaign
2. Volume of production or services, allowing for selling price below the current market price
3. Quality of product or service meeting international standards, such as ISO
4. Access to the distribution channels
5. Formal barriers (legal, duties, taxation, etc.)

Operating in the global market is much more difficult than in local markets, as it demands application of advanced IT solutions, access to the international data warehouses, and so forth. Only with the use of IT and access to information, a global company can reduce the risk of failure.

Operational costs of the enterprise are still larger if it operates in the global e-market (Elstrom, 2001). Surveys (Reichheld, 2001; Reichheld & Schefter, 2000) have shown that the cost of gaining a customer in this market is significantly higher than in the traditional market. For example, in the fashion industry, the cost of gaining a customer in the e-market is up to 40% higher. The profits following years, however, grow much faster. In the near future, organizations that cannot succeed in this market will not be able to succeed at all.

Figure 3. Comparison of a virtual organization and LSEs (A virtual organization as a set of SMEs is more competitive than LSEs)



Identification of the break-even point requires empirical surveys. It depends on the industry and the degree of globalization of the enterprise (there are not many enterprises that have been operating on all continents). To identify the break-even point, we may apply the methods of the strategic analysis, and especially the analysis of Porter's five forces. For example, the break-even point may be identified by the investment required to realize 2% of revenue from the global market. Sometimes, we can consider as an organization of such a type the one that has more than 50% of revenue from the global market. Because single SMEs do not have such resources to surpass the break-even point, they create strategic alliances in the form of virtual enterprise.

Theoretically, development of the virtual organization may be significant. Practically, a degree of its development is connected with technical barriers and with the necessity to achieve the goals set up for the virtual SMEs. Modification of the goal causes a fast modification of the enterprise.

Modern companies use such development strategies, which allow them to have a competitive position on both home and world markets. In different organizations, nowadays, it is connected with possibilities that are provided by virtualization and the Internet in particular. Among others, this fact is pointed out by Champy (2002) who presented foundations of the X re-engineering.

FUTURE TRENDS

My research suggests that a significant number of students feel the need to learn business administration (Kisielnicki, 2001). At the same time, learning should be as close to the reality as possible. Virtualization is the discipline of studies, which may cause the distance between the theory and practice to diminish. One may even risk a statement that it is the exact direction of the virtualization, which, in the nearest future, shall have the most significant impact on the society. Virtualization increases the effectiveness of the teaching process in the widest possible sense by supporting traditional teaching methods. Virtualization may be applied in such areas as lowering the cost and decreasing time of training jet pilots and also in improvement of military command, operating on the stock exchange, or cognitive analysis of genetic processes. In science, there are many examples of big discoveries, which were first tried on computer-simulated models. It is virtualization that allows for business simulation of both the decision-making process and the analysis of complex technical or sociological processes. In virtualization of teaching, there are two basic directions:

The first one is the direction of common education where everybody can, using the IT, gain a given knowledge. A classic example is a virtual stock exchange. In many countries, a lot of people want to learn how to operate on the stock exchange before they actually start using it to make money. They can get the necessary experience using appropriate software. They can acquire necessary skills in the virtual world. There are also numerous games available through the Internet. These games not only provide pleasant time spent on playing, but also teach foreign languages or how to drive a car.

Another direction is dedicated teaching. There are the following activities where virtualization can help:

- **Self-Evaluation:** Using special software to assess the level of language knowledge.
- **Learning Assistance:** Includes, amongst others, enterprise laboratories, business games, and special simulators that teach how to use a specific technical equipment, for example, flying simulators.
- **Distant Learning:** A student who has a proper terminal and software may participate in classes from a distance. This direction of virtualization is similar to the previously presented distance work.

The researches that I have conducted until now have shown that strategies connected with using only the virtual organization do not always provide positive results and meet people's expectations. It is due to a number of factors, concerning ethics and trust toward the virtualization issues (these matters were considered by Sutherland, 2004, and Kisielnicki, 2002). Therefore, obtaining a better position on a competitive market requires applying more of the convergence strategy. The convergence strategy combines traditional strategy with the use of virtualization (Kisielnicki, 2006; Thomenendal, 2006).

CONCLUSION

Virtual organization operates as a *transparent organization* in a sense that all its initiatives are known to the competition. This enforces responsibility. The image of a virtual organization is influenced by activities of its individual participants. Virtual organizations do not have common administration, offices, and real estate. In addition, virtual organizations do not have a joint executive team, middle management, or coordinators. Since the influence of each individual virtual organization is limited, the key success factor is a mutual trust.

Virtual organizations give a new insight into business management. They may trigger increased entrepreneurship and competitiveness. They also introduce new management methods different than traditional ones. Virtual organizations are an interesting alternative for current organizations. They are especially attractive for developing countries that want to operate in the international environment. It is fascinating how virtual organizations—without a formal reporting structure and control—can achieve high operational per-

formance and thus have a competitive advantage over their traditional counterparts.

Regardless of speculation on future solutions, we can now safely define the areas where virtual organizations can easily out-perform traditional organizations. They are trade, tourism, and services. In these areas, the benefits of virtual organizations are as follows:

1. Operational flexibility is much higher, especially when a quick reaction to the emerging market niche is required.
2. A transaction lifecycle is much shorter (especially the closure).
3. Use of cyberspace to close some transactions (despite legal and organizational barriers).
4. Lower operational costs.
5. Lower start-up costs.
6. Ability to cooperate with organizations that LSEs cannot accept (for political, geographical, racial, religious reasons).

Depending on each individual opportunity, virtual organizations can be successful in other industries as well. Very rarely, traditional organizations can outperform the virtual ones.

The economic metrics supporting the benefits of virtual organizations are generally available for individual cases. However, for illustration purposes, I include the results of research recently carried out in Poland:

1. Virtual travel agencies achieved 20% higher profit per transaction than traditional organizations.
2. Transaction life cycle was 100% faster in the Internet bank than in a brick-and-mortar bank.
3. The average transaction cost on the Internet bank was $0.10-0.15 compared to $1.10 in the brick-and-mortar bank.

We can put forward a hypothesis that progress in IT will create a snowball of virtual organizations. Tapscott (1998), introducing the 12 rules of the new economic deal, writes that they lead toward "virtual reality." Future strategy of business wills strategy convergence strategy.

REFERENCES

Byrne, J. A., & Brandt, R. (1993, August 2). The virtual corporation. *Business Week*, p. 5.

Champy, J. A. (2002). *X-engineering the corporation. Reinventing your business in the digital age*. New York: Warner Books Inc.

Drucker, P. (1988). The coming of the new organisation. *Harvard Business Review,* 66(1-2), 9.

Elstrom, E. (2001, July). E-money. *Business Week*, p. 63.

Fong, M. W. L. (Ed.). (2005). *E-collaborations and virtual organizations*. Hershey, PA: IRM Press.

Hammer, M., & Champy, J. (1993). Reengineering the corporation. *HarperBusiness.* New York.

Hammer, M., & Stanton, S. (1999, November-December). How process enterprises really work. *Harvard Business Review, 77*(6), 108.

Hendberg, B., Dahlgren, G., Hansson, J., & Olive, N. (2000). Virtual organizations and beyond; discovering imaginary systems.

Kenny, D., & Marshall, J. F. (2000, November-December). The real business of the Internet. *Harvard Business Review, 78*(6) 119.

Kisielnicki, J. (1998). Virtual organization as a product of information society. *Informatica, 22*, 3.

Kisielnicki, J. (2001). Virtual organization as a chance for enterprise development. In M. Khosrow-Pour (Ed.), *Managing information technology in a global economy* (p. 349). Idea Group.

Kisielnicki, J. (Ed.). (2002). *Virtual organization in modern society*. Idea Group.

Kisielnicki, J. (2006). The convergence strategy in small and medium sized companies exemplified by e-business and t-business organisations. In M. Khosrow-Pour (Ed.), *Emerging trends and challenges* (pp. 78-81). Idea Group.

Peters, T. (1994). Crazy times call for crazy organisations. The Ton Peters Seminar.

Porter, M. E. (2001, March). Strategy and the Internet. *Harvard Business Review, 79*(3), 62.

Quinn, J. B. (1992). *The intelligent enterprise*. New York: The Free Press.

Reichheld, F. F. (2001, July-August). Lead for the loyalty. *Harvard Business Review, 79*(4), 76.

Reichheld, F. F., & Schefter, P. (2000, July-August). E-loyalty, your secret weapon on the Web. *Harvard Business Review, 78*(4), 105.

Scholzch, C. (1996). Virtuelle Unternehmen—Organisatorische Revolution mit Strategischer Implikation. *Management & Computer, 2*, 16.

Sutherland, P., Tan, B. F. (2004). *The nature of consumer trust in B2C electronic commerce: A multi-dimensional conceptualism.* IRMA.

Tapscott, D. (1998). *Digital economy*. Warsaw: Business Press.

Thomenendal, M. (2006). Complex dynamics in the virtual corporation. *Master of Business Adminstration, 80*(3), 25.

Turban, E., King, D., Lee, J., Warkenting, M., & Chung, M. (2002). Electronic commerce: Managerial perspective.

Yip, G. S. (2004). *Global strategy*. Warszawa: PWE.

KEY TERMS

Cyberspace: Defined by the Miriam Webster Online dictionary as an "online world of computer networks." This definition can be augmented by the following characteristics: the network consists of various, globally distributed computers that can send and receive information using common protocols. In addition, this network does not have physically defined measured boundaries. The examples of the network types are electronic mail (e-mail), World Wide Web (WWW), electronic data interchange (EDI), business to business (B2B) applications, business to customer (B2C) applications, and peer-to-peer (P2P) applications.

Virtual Organization: (a) Virtual organization (Kisielnicki, 2002, p. 102) is created voluntarily; its members create this organization to accomplish a certain goal. Virtual organization is created anytime a manger realizes that he or she needs the cooperation of other organizations to accomplish its goal. The duration of a virtual organization is defined by each of its members; virtual organization operates in the cyberspace, which means that its duration can be very short, so short in fact, that it would be impossible to cooperate with other organizations using the traditional methods. The decision to restructure or fold the organization can be made anytime by any of its members. Virtual organization does not have one president or general manager. The virtualization process is a process of transformation from traditional organization into virtual using informational technology. This process can result in two forms of virtual organization: traditional organization with virtual divisions, or an association of many organizations, as depicted in Figure 3. (b) Artificial organizational structure where individual organizations provide base competencies. The integration and coordination of these base competencies allows for effective execution of chain process to deliver a product and satisfy a customer. This integration and coordination does not incur additional costs and maintains the customer focus (Scholzch, 1996). (c) Temporary network of independent enterprises—suppliers, customers, even former competitors—linked through the IT and working together to share their knowledge and costs to enter a new market (Byrne & Brandt, 1993).

Visual Medical Information Analysis

Maria Papadogiorgaki
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Vasileios Mezaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

Yiannis Chatzizisis
Aristotle University of Thessaloniki, Greece

George D. Giannoglou
Aristotle University of Thessaloniki, Greece

Ioannis Kompatsiaris
Informatics and Telematics Institute, Centre for Research and Technology Hellas, Greece

INTRODUCTION

Images have constituted an essential data source in medicine in the last decades. Medical images derived from diagnostic technologies (e.g., X-ray, ultrasound, computed tomography, magnetic resonance, nuclear imaging) are used to improve the existing diagnostic systems for clinical purposes, but also to facilitate medical research. Hence, medical image processing techniques are constantly investigated and evolved.

Medical image segmentation is the primary stage to the visualization and clinical analysis of human tissues. It refers to the segmentation of known anatomic structures from medical images. Structures of interest include organs or parts thereof, such as cardiac ventricles or kidneys, abnormalities such as tumors and cysts, as well as other structures such as bones, vessels, brain structures and so forth. The overall objective of such methods is referred to as computer-aided diagnosis; in other words, they are used for assisting doctors in evaluating medical imagery or in recognizing abnormal findings in a medical image.

In contrast to generic segmentation methods, techniques used for medical image segmentation are often application-specific; as such, they can make use of prior knowledge for the particular objects of interest and other expected or possible structures in the image. This has led to the development of a wide range of segmentation methods addressing specific problems in medical applications. In the sequel of this article, the analysis of medical visual information generated by three different medical imaging processes will be discussed in detail: Magnetic Resonance Imaging (MRI), Mammography, and Intravascular Ultrasound (IVUS). Clearly, in addition to the aforementioned imaging processes and the techniques for their analysis that are discussed in the sequel, numerous other algorithms for applications of segmentation to specialized medical imagery interpretation exist.

BACKGROUND

Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is an important diagnostic imaging technique attending to the early detection of the abnormal conditions in tissues and organs because it is able to reliably identify anatomical areas of interest. In particular for brain imaging, several techniques which perform segmentation of the brain structures from MRIs are applied to the study of many disorders, such as multiple sclerosis, schizophrenia, epilepsy, Parkinson's disease, Alzheimer's disease, and so forth. MRI is particularly suitable for brain studies because it is virtually noninvasive, and it achieves a high spatial resolution and high contrast of soft tissues. To achieve the 3D reconstruction of the brain morphology, several of the existing approaches perform segmentation on sequential MR images. The overall process usually includes noise filtering of the images and edge detection for the identification of the brain contour. Following, perceptual grouping of the edge points is applied in order to recover the noncontinuous edges. In many cases, the next step is the recognition of the various connective components among the set of edge points, rejection of the components that consist of the smallest number of points, and use of the finally acquired points for reconstructing the 3D silhouette of the brain, as will be discussed in more detail in the sequel.

Mammography

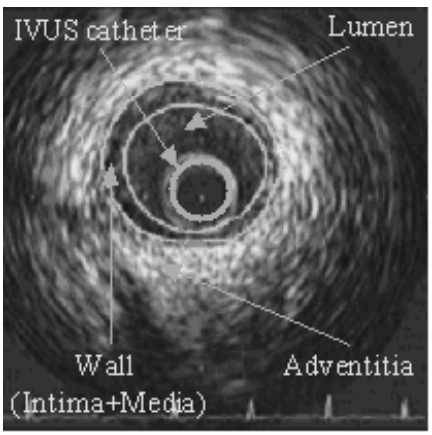
Mammography is considered to be the most effective diagnostic technique for detecting abnormal tissue conditions on women's breast. Being used both for prevention and for diagnostic purposes, it is a very commonly used technique that produces mammographic images by administering a low-dose of x-ray radiation to the tissue under examination. The analysis of the resulting images aims at the detection of any abnormal structures and the quantification of their characteristics, such as size and shape, often after detecting the pectoral muscle and excluding it from the further processing. Methods for the analysis of mammographic images are presented in the sequel.

Intravascular Ultrasound

IVUS is a catheter-based technique that renders two-dimensional images of coronary arteries and therefore provides valuable information concerning luminal and wall area, plaque morphology and wall composition. An example IVUS image, with tags explaining the most important parts of the vessel structure depicted on it, is shown in Figure 1. However, due to their tomographic nature, isolated IVUS images provide limited information regarding the burden of atherosclerosis. This limitation can be overcome through 3D reconstruction techniques in order to stack the sequential 2D images in space, using single-plane or biplane angiography for recovering the vessel curvature (Giannoglou et al., 2006; Sherknies, Meunier, Mongrain, & Tardif, 2005; Wahle, Prause, DeJong, & Sonka, 1999).

The analysis of IVUS images constitutes an essential step toward the accurate morphometric analysis of coronary plaque. To this end, the processing of IVUS images is necessary so that the regions of interest can be detected. The coronary artery wall mainly consists of three layers: intima, media and adventitia, while three regions are supposed to be visualized as distinguished fields in an IVUS image, namely the lumen, the vessel wall (made of the intima and the media layers) and the adventitia plus surroundings. The above regions are separated by two closed contours: the inner border, which corresponds to the lumen-wall interface, and the outer border representing the boundary between media and adventitia. A reliable and quick detection of these two borders in sequential IVUS images constitutes the basic step towards plaque morphometric analysis and 3D reconstruction of the coronary arteries.

Figure 1. Example IVUS image with tags explaining the most important parts of the vessel structure depicted on it



VISUAL MEDICAL INFORMATION ANALYSIS TECHNIQUES

Magnetic Resonance Imaging Analysis

Several techniques have been proposed for the analysis of MR images. In Grau, Mewes, Alcaniz, Kikinis, and Warfield (2004), a modification of a generic segmentation technique, the watershed transform, is proposed for knee cartilage and gray matter/white matter segmentation in MR images. This introduces prior information in the watershed method via the use of a previous probability calculation for the classes present in the image and via the combination of the watershed transform with atlas registration for the automatic generation of markers. As opposed to Grau et al. (2004), other methods are more application specific; in Woolrich, Behrens, Beckmann and Smith (2005), for example, segmentation tools are developed for the study of the function of the brain, that is, for the classification of brain areas as activating, deactivating, or not activating, using functional magnetic resonance imaging (FMRI) data. This method performs segmentation based on intensity histogram information, augmented with adaptive spatial regularization using Markov random fields. The latter contributes to improved segmentation as compared to nonspatial mixture models, while not requiring the heuristic fine-tuning that is necessary for nonadaptive spatial regularization previously proposed.

Because MR images contain a significant amount of noise caused by operator performance, equipment, or even the environment, the segmentation on them can lead to several inaccuracies. In order to overcome the effects of noise, Shen, Sandham, Granat and Sterr (2005) propose a segmentation technique based on an extension to the traditional fuzzy c-means (FCM) clustering algorithm. The segmentation performance is improved using neighborhood attraction, which depends on the relative location and features of neighboring pixels. The degree of attraction is optimized by applying a neural network model. Greenspan, Ruf and

Goldberger (2006) have also developed an algorithm for the automated brain tissue segmentation on noisy, low-contrast (MR) images. Under their approach, the brain image is represented by a model that is composed of a large number of Gaussians. For the algorithm's initialization an atlas or parameter learning are not required. Finally, segmentation of the brain image is achieved by affiliating each voxel to the component of the model that maximizes an a posteriori probability. In Valdes-Cristerna, Medina-Banuelos and Yanez-Suarez (2004) a hybrid model for the segmentation of brain MRI has been investigated. The model includes a radial basis network and an active contour model. The radial basis network algorithm generates an initial contour, which is following used by the active contour model to achieve the final segmentation of the brain.

Mammography Image Analysis

Several applications have been proposed, which process the mammographic images in order to assist the clinicians in their diagnostic procedure. In Székely, Toth and Pataki (2006) the mammographic images are analyzed using segmentation in order to identify regions of interest. The applied segmentation technique includes texture features, decision trees, and a Markov random field model. The extracted features which refer to the object's shape and texture parameters are linearly combined to lead to the final decision. Because the pectoral muscle should be excluded from processing on a mammogram intended for the breast tissue, its identification is important. Kwok, Chandrasekhar, Attikiouzel and Rickard (2004) have developed an adaptive segmentation technique for the extraction of the pectoral muscle on digitized mammograms. The method uses knowledge about the position and shape of the pectoral muscle. Other approaches, such as Cascio, Fauci, Magro, Raso, Bellotti, De Carlo et al. (2006) use supervised neural networks for detecting pathological masses in mammograms. A segmentation process provides features of geometrical information, or shape parameters which constitute the input to the neural network that computes the probability of the lesion existence.

IVUS Image and Image Sequence Analysis

Traditionally, the segmentation of IVUS images is performed manually, which is a time consuming procedure with results affected by the high inter- and intra-user's variability. To overcome these limitations, several approaches for semi-automated segmentation have been proposed in the literature. In Herrington, Johnson, Santago and Snyder (1992) after the manual indication of the general location of the boundary of interest by the user, an edge detection filter is applied to find potential edge points within the pointed neighborhood. The extracted image data are used for the estimation of the closed smooth final contour. Sonka, Zhang, Siebes, Bissing, DeJong, Collins et al. (1995) implemented a knowledge-based graph searching method incorporating a priori knowledge on coronary artery anatomy and a selected region of interest prior to the automatic border detection. Quite a few variations of active contour model have been investigated, including the approach of Parissi et al. (2006), where a user interaction is required, by drawing an initial contour as close as possible to its final position. Thus, the active contour is initialized and tends to approximate the final desired border.

The active contour or deformable models principles have been used to allow the extraction of the luminal and medial-adventitial borders in three dimensions after setting an initial contour in Kovalski, Beyar, Shofti and Azhari (2000). However, the contour detection fails for low contrast interface regions such as the luminal border where the blood-wall interface in most images corresponds to weak pixel intensity variation. In order to improve the included active surface segmentation algorithm for plaque characterization, Klingensmith, Nair, Kuban and Vince (2004) use the frequency information after acquiring the radiofrequency (RF) IVUS data through an electrocardiogram scheme. Radio frequency data are also used in Perrey et al. (2004), after in vivo acquisition for the segmentation of the luminal boundary in IVUS images. According to this approach, tissue describing parameters are directly estimated from RF data. Subsequently, a neuro-fuzzy inference system trained to several parameters is used to distinguish blood from tissue regions.

For clinical practice the most attractive approaches are the fully automatic ones. A limited number of them has been developed so far, such as the segmentation based on edge contrast (Zhu, Liang, & Friedman, 2002); the latter is shown to be an efficient feature for IVUS image analysis, in combination with the gray level distribution. Specific automated approaches which utilize the deformable models principles in combination with other various techniques and features reported in the related literature have been investigated. Brusseau, de Korte, Mastik, Schaar and van der Steen (2004) exploited an automatic method for detecting the endoluminal border based on an active contour that evolves until it optimally separates regions with different statistical properties without using a preselected region of interest or initialization of the contour close to its final position. Another automated approach based on deformable models has been reported by Plissiti, Fotiadis, Michalis and Bozios (2004), who employed a Hopfield neural network for the modification and minimization of an energy function as well as a priori vessel geometry knowledge. An automated approach for segmentation of IVUS images based on a variation of an active contour model is presented in Giannoglou et al. (2007). This technique is in vivo evaluated in images originated from human coronary arteries. The initialization of

the contours in each IVUS frame is automatically performed using an algorithm, which is based on the intensity features of the image. The initially extracted boundaries constitute the input to the active contour model, which then deforms the contours appropriately, identifying their correct location on the IVUS frame.

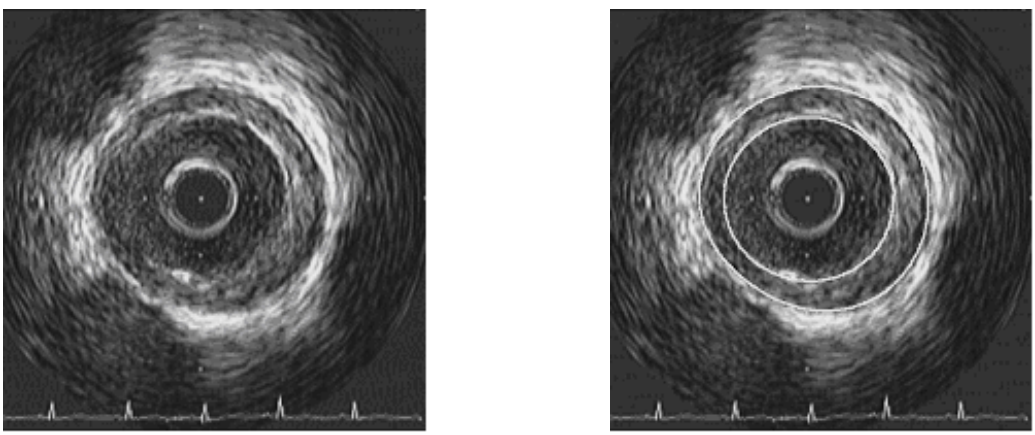
A fuzzy clustering algorithm for adaptive segmentation in IVUS images is investigated by Filho, Yoshizawa, Tanaka, Saijo and Iwamoto (2005). Cardinal et al. (2006) present a 3D IVUS segmentation applying Rayleigh probability density functions (PDFs) for modeling the pixel gray value distribution of the vessel wall structures. Other approaches are based on the calculation of the image's energy. In Luo, Wang and Wang (2003) the lumen area of the coronary artery is estimated using the internal energy, which describes the smoothness of the arterial wall and the external energy which represents the grayscale variation of images that constitute an IVUS video. The minimal energy which defines the contour is obtained using circular dynamic programming. Other methods include statistical analysis, such as Gil, Hernandez, Rodriguez, Mauri and Radeva's (2006), where the presented approach uses statistical classification techniques for the IVUS border detection.

In Papadogiorgaki, Mezaris, Chatzizisis, Kompatsiaris and Giannoglou (2006), a fully automated method for the segmentation of IVUS images and specifically for the detection of luminal and medial-adventitial boundaries is presented. This technique is based on the use of the results of texture analysis, performed by means of a multilevel Discrete Wavelet Frames decomposition. Following image preprocessing, to remove catheter-induced artifacts, a two-step process is employed for the detection of the boundaries of interest. Objective of the first step, termed contour initialization, is the detection of pixels that are likely to belong to the lumen and media-adventitia boundaries, taking into consideration the previously extracted texture features. As a result of this step, initial contours are generated; however, these are not smooth and are characterized by discontinuities, as opposed to the true lumen and media-adventitia boundaries. Thus, at the second step, a filtering or approximation procedure is applied to the initial contour functions, so as to result in the generation of smooth contours that are in good agreement with the true lumen and media-adventitia boundaries. This approach does not require manual initialization of the contours and demonstrates the importance of texture features in IVUS image analysis. A sample IVUS image and the corresponding analysis result of this approach are illustrated in Figure 2.

FUTURE TRENDS

With the number of medical imaging techniques used in everyday practice constantly rising and the quality of the results of such imaging techniques constantly increasing, to the benefit of the patients, it is clear that the need for accurate and automated to the widest possible extent analysis of medical images is a key element in supporting the diagnosis and treatment process for a wide range of medical conditions. To this end, future research will continue to concentrate on the development of analysis methods that are automated, robust and reliable. In addition to that, particular emphasis is expected to be put on the coupling of the results of automated analysis with techniques for the formal representation of them.

Figure 2. Sample IVUS image (left) and the corresponding analysis result of a texture-based approach to the detection of luminal and medial-adventitial boundaries (right)



Examples of early systems for the formal representation of knowledge extracted from medical images, though not in an automated manner, include those discussed in Dasmahapatra et al. (2006), Hu, Dasmahapatra, Lewis and Shadbolt (2003). These are concerned with the annotation of medical images used for the diagnosis and management of breast cancer, such as those generated by mammography and MRI, by expressing all the extracted features and regions of interest using domain knowledge and assigning them to specific concepts of a knowledge structure. In an analogous approach, in Gedzelman, Simonet, Bernhard, Diallo and Palmer (2005) a knowledge structure of cardiovascular diseases is constructed in order to be used for the representation of the findings of the relevant imaging techniques, so as to support concept-based information retrieval.

Combining automated analysis results with techniques such as those briefly discussed above for the formal representation of them will empower new possibilities in the areas of retrieval in extensive medical databases and reasoning over the results of analysis, consequently providing the physicians not only with the analysis results themselves but also with hints on their meaning, minimizing the risk of misinterpretation.

CONCLUSION

In this article, visual medical information analysis was discussed, starting with an introduction on the current use of medical imaging and the needs for its analysis. The current article then focused on three important medical imaging techniques, namely Magnetic Resonance Imaging (MRI), Mammography, and Intravascular Ultrasound (IVUS), for which a detailed presentation of the goals of analysis and the methods presented in the literature for reaching these goals was given. The future trends identified in the relevant section provide insights on how the algorithms outlined in this article can be further evolved, so as to more efficiently address the problem of medical image analysis and consequently pave the way for the development of innovative doctor decision support applications that will make the most out of the available image data.

REFERENCES

Brusseau, E., de Korte, C.L., Mastik, F., Schaar, J., & van der Steen, A. F. W. (2004). Fully automatic luminal contour segmentation in intracoronary ultrasound imaging—a statistical approach. *IEEE Transactions on Medical Imaging, 23*(5), 554-566.

Cardinal, M.-H.R., Meunier, J., Soulez, G., Maurice, R.L., Therasse, E., & Cloutier, G. (2006). Intravascular ultrasound image segmentation: A three-dimensional fast-marching method based on gray level distributions. *IEEE Transactions on Medical Imaging, 25*(5), 590-601.

Cascio, D., Fauci, F., Magro, R., Raso, G., Bellotti, R., De Carlo, F., et al. (2006). Mammogram segmentation by contour searching and mass lesions classification with neural network. *IEEE Transactions on Nuclear Science, 53*(5), 2827-2833.

Dasmahapatra, S., Dupplaw, D., Hu, B., Lewis, H., Lewis, P., & Shadbolt, N. (2006). Facilitating multi-disciplinary knowledge-based support for breast cancer screening. *International Journal of Healthcare Technology and Management, 7*(5), 403-420.

Dos S., Filho, E., Yoshizawa, M., Tanaka, A., Saijo, Y., & Iwamoto, T. (2005, September). Detection of luminal contour using fuzzy clustering and mathematical morphology in intravascular ultrasound images. In *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE-EMBS)*, China.

Gedzelman, S., Simonet, M., Bernhard, D., Diallo, G., & Palmer, P. (2005, September). Building an ontology of cardiovascular diseases for concept-based information retrieval. *Computers in Cardiology*, Lyon, France.

Giannoglou, G.D., Chatzizisis, Y.S., Koutkias, V., Kompatsiaris, I., Papadogiorgaki, M., Mezaris, V., et al. (2007). A novel active contour model for fully automated segmentation of intravascular ultrasound images: In-vivo validation in human coronary arteries. *Computers in Biology and Medicine, accepted for publication*.

Giannoglou, G.D., Chatzizisis, Y.S., Sianos, G., Tsikaderis, D., Matakos, A., Koutkias, V., et al. (2006). In-vivo validation of spatially correct three-dimensional reconstruction of human coronary arteries by integrating intravascular ultrasound and biplane angiography. *Coronary Artery Disease, 17*(6), 533-543.

Gil, D., Hernandez, A., Rodriguez, O., Mauri, J., & Radeva, P. (2006). Statistical strategy for anisotropic adventitia Modelling in IVUS. *IEEE Transactions on Medical Imaging, 25*(6), 768-778.

Grau, V., Mewes, A.U.J., Alcaniz, M., Kikinis R., & Warfield, S.K. (2004). Improved watershed transform for medical image segmentation using prior information. *IEEE Transactions on Medical Imaging, 23*(4), 447-458.

Greenspan, H., Ruf, A., & Goldberger, J. (2006). Constrained gaussian mixture model framework for automatic segmentation of MR brain images. *IEEE Transactions on Medical Imaging, 25*(9), 1233-1245.

Herrington, D.M., Johnson, T., Santago, P., & Snyder, W.E. (1992, October). Semi-automated boundary detection for intravascular ultrasound. In *Proceedings of Computers in cardiology* (pp. 103-106). Durham, NC, USA.

Hu, B., Dasmahapatra, S., Lewis, P., & Shadbolt, N. (2003, November). Ontology-based medical image annotation with description logics. In *Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, Sacramento, CA, USA.

Klingensmith, J.D., Nair, A., Kuban, B.D., & Vince, D.G. (2004, August). Segmentation of three-dimensional intravascular ultrasound images using spectral analysis and a dual active surface model. In *Proceedings of the IEEE Ultrasonics Symposium*, Montreal, Canada.

Kovalski, G., Beyar, R., Shofti, R., & Azhari, H. (2000). Three-dimensional automatic quantitative analysis of intravascular ultrasound images. *Ultrasound in Medicine & Biology, 26*(4), 527-537.

Kwok, S.M., Chandrasekhar, R., Attikiouzel, Y., & Rickard, M.T. (2004). Automatic pectoral muscle segmentation on mediolateral oblique view mammograms. *IEEE Transactions on Medical Imaging, 23*(9), 1129-1140.

Luo, Z., Wang, Y., & Wang, W. (2003). Estimating coronary artery lumen area with optimization-based contour detection. *IEEE Transactions on Medical Imaging, 22*(4), 564-566.

Perrey, C., Scheipers, U., Bojara, W., Lindstaedt, M., Holt, S., & Ermert, H. (2004, August). Computerized segmentation of blood and luminal borders in intravascular ultrasound. In *Proceedings of the IEEE Ultrasonics Symposium*, Montreal, Canada.

Papadogiorgaki, M., Mezaris, V., Chatzizisis, Y.S., Kompatsiaris I., & Giannoglou, G.D. (2006, September). A fully automated texture-based approach for the segmentation of sequential IVUS images. In *Proceedings of the 13th International Conference on Systems, Signals & Image Processing (IWSSIP)*, Budapest, Hungary.

Parissi, E., Kompatsiaris, Y., Chatzizisis, Y.S., Koutkias, V., Maglaveras, N., Strintzis, M.G., et al. (2006, December). An automated model for rapid and reliable segmentation of intravascular ultrasound images. In *Proceedings of the 7th International Symposium on Biological and Medical Data Analysis (ISBMDA)*, Thessaloniki, Greece.

Plissiti, M.E., Fotiadis, D.I., Michalis, L.K., & Bozios, G.E. (2004). An automated method for lumen and media–adventitia border detection in a sequence of IVUS frames. *IEEE Transactions on Information Technology in Biomedicine, 8*(2), 131-141.

Shen, S., Sandham, W., Granat, M., & Sterr, A. (2005). MRI fuzzy segmentation of brain tissue using neighborhood attraction with neural network optimization. *IEEE Transactions on Information Technology in Biomedicine, 9*(3), 459-467.

Sherknies, D., Meunier, J., Mongrain, R., & Tardif, J.-C. (2005). Three-dimensional trajectory assessment of an IVUS transducer from single-plane cineangiograms: A phantom study. *IEEE Transactions on Biomedical Engineering, 52*(3), 543-549.

Sonka, M., Zhang, X., Siebes, M., Bissing, M.S., DeJong, S.C., Collins, S.M., et al. (1995). Segmentation of intravascular ultrasound images: A knowledge-based approach. *IEEE Transactions on Medical Imaging, 14*(4), 719-732.

Székely, N., Toth, N., & Pataki, B. (2006). A hybrid system for detecting masses in mammographic images. *IEEE Transactions on Instrumentation and Measurement, 55*(3), 944-952.

Valdes-Cristerna, R., Medina-Banuelos, V., & Yanez-Suarez, O. (2004). Coupling of radial-basis network and active contour model for multispectral brain MRI segmentation. *IEEE Transactions on Biomedical Engineering, 51*(3), 459-470.

Wahle, A., Prause, G.P.M., DeJong, S.C., & Sonka, M. (1999). Geometrically correct 3-D reconstruction of intravascular ultrasound images by fusion with biplane angiography—methods and validation. *IEEE Transactions on Medical Imaging, 18*(8), 686-699.

Woolrich, M.W., Behrens, T.E.J., Beckmann, C.F., & Smith, S.M. (2005). Mixture models with adaptive spatial regularization for segmentation with an application to FMRI data. *IEEE Transactions on Medical Imaging, 24*(1), 1-11.

Zhu, H., Liang, Y., & Friedman, M.H. (2002). IVUS image segmentation based on contrast. In *Proceedings of SPIE*, Durham, NC, USA, (Vol. 4684, pp. 1727-1733).

KEY TERMS

Active Contour Model: Energy-minimizing parametric curve that is the basis of several medical image analysis techniques.

Computer-Aided Diagnosis: The process of using computer-generated analysis results for assisting doctors in evaluating medical data.

Coronary Angiography: X-ray diagnostic process for obtaining an image of the coronary arteries.

Intravascular Ultrasound (IVUS): Diagnostic catheter-based technique that renders two-dimensional images of coronary arteries.

Magnetic Resonance Imaging (MRI): Imaging technique that uses a magnetic field to provide two-dimensional images of internal body structures.

Mammography: Diagnostic X-ray technique which produces breast images and is used to detect breast tissue abnormalities.

Medical Image Segmentation: The localization of known anatomic structures in medical images.

Web Access by Older Adult Users

Shirley Ann Becker
Florida Institute of Technology, USA

INTRODUCTION

The older adult population in the U.S. continues to increase at a rapid pace due to aging baby boomers and increased life expectancy. Older Americans, 60 years and older, will comprise about 20% of the total population by 2030, which is more than twice the number of aging adults than in 2000 (Administration on Aging, 2002).

The Web offers an unprecedented opportunity for older adults to access a wealth of online resources. Increasingly, older adults are using the Web for information on self-diagnosis, treatment, and prevention of health problems (Preidt, 2003). They are taking advantage of electronic government resources to vote, file taxes, obtain social services, voice their opinions to officials, and search historical records. Older adults are also using the Web to stay socially active in terms of communicating with other users via lists and chat rooms (Czaja, Guerrier, Nair & Landauer, 1993; Kerschner & Hart, 1984).

Older adults are getting online by an estimated growth rate of 15% per year (Coulson, 2000). They log over eight hours of online time per week and visit more Web sites than persons in younger age groups when they are online (Morrell, Dailey, Feldman, Holt, Mayhorn & Echt, 2002). Their use of the Internet is predicted to increase as much as 358%, from 3.7 million users in 2001 to 17.3 million in 2005 (Scanlon, 2001).

Unfortunately, older adults may have trouble accessing a Web site because of design issues that impede its use. Barriers may be encountered due to a site's color scheme, font size and type, navigation, vertical screen length, image mouseovers, and sentence complexity, among others. In this information-rich society, many older adults will remain "information have-nots" unless these barriers are removed.

BACKGROUND

In the U.S., the Internet has emerged as a major communications medium with the potential to disseminate information to all citizens including older adults. Yet, there is an ongoing concern that the opportunities associated with Internet access may not be readily available to all citizens. This concern has been expressed in a recent Pew Foundation study (Lenhart, Horrigan, Rainie, Allen, Boyce, Madden & O'Grady, 2003, p. 6):

"Internet non-users will have less power as consumers and fewer economic opportunities, less access to high-quality health information, fewer options for dealing with government agencies, no chance to learn about their world from the millions of organizations and learning centers that have posted their material on the Web, and less opportunity to interact with others through email and instant messaging."

Older adults in particular may encounter Web accessibility barriers due to vision, cognition, and physical changes that are associated with the normal aging process. Reading complexity may also become a barrier when literacy skills of older adults are not taken into account in the design of Web content.

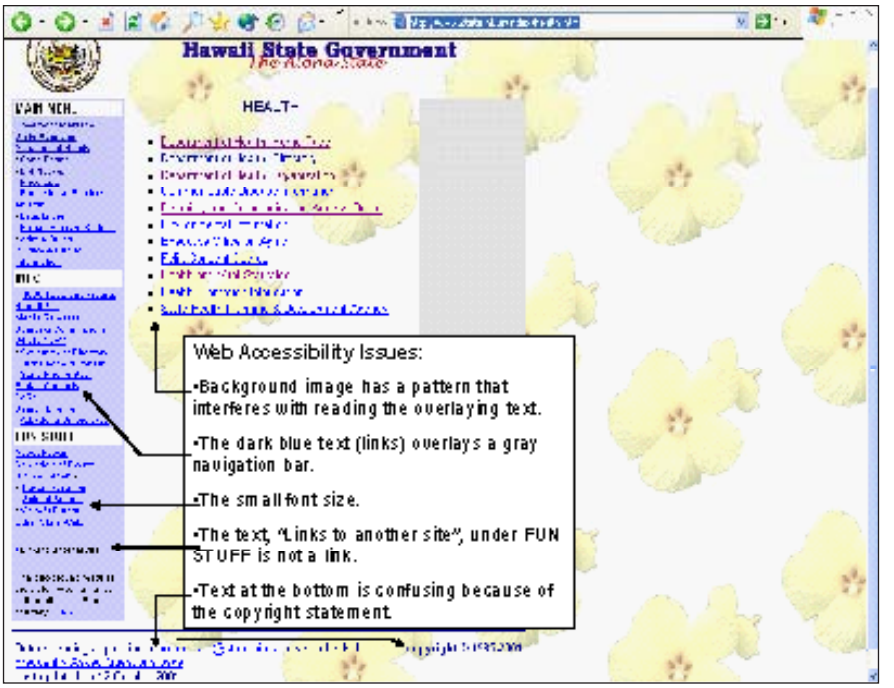
Vision

The aging eye has a reduced ability to focus on close objects due to a reduction in the elasticity in the lens. Other vision changes due to the normal aging process include: a decline in visual acuity, impacting the ability to see objects clearly, yellowing and thickening of the lens, impacting color perception, decreased light sensitivity, impacting adaptation to changes in light levels, increased sensitivity to glare from light reflecting or shining into the eye, and reduced depth perception, making it more difficult to judge the distance of an object (American Foundation for the Blind, 1999). These vision changes impact the use of the Web in terms of the legibility of written content on the page. They also impact searches, navigation, and reading speed and comprehension (Echt, 2002).

Figure 1 illustrates readability issues associated with a state government Web page when taking into account aging vision. The patterned background image may negatively impact readability, especially given the small size of the foreground text[1]. The format of the text at the bottom of the page also impacts readability given that the sentence "Before sending a question to <webmaster>, please check the Frequently Asked Questions page" breaks to accommodate the copyright statement appearing on the right.

The use of color can also impact the readability of information content on a Web page due to aging vision (refer to Becker, 2004a for a discussion on color and Web accessibility for older adults). For many older adults, foreground and background color combinations may render a Web page visually inaccessible. In Figure 1, the contrast between the text and background colors in the navigation bar may be

Figure 1. Illustration of Web barriers and aging vision (http://www.state.hi.us/index/health.htm)



insufficient for older adult readers. Figure 2 shows the New Mexico state government homepage with saturated colors for both the foreground and background. The edges of the text tend to blur when bright or neon colors are used in combination with saturated colors (e.g., bright yellow text displayed on a red background), thus reducing legibility of the text for many older adult users.

Cognition

Studies show that an older adult's working and spatial memory task performance declines with age (Holt & Morrell, 2002). As a result, an older adult may not be able to discern details in the presence of distracting information. In addition, complex navigational schemes, nonintuitive searches and cluttered pages may impede use of a Web site because of declines in working and spatial memory.

Of 40 U.S. state government sites assessed, over 60% required traversing three or more screen pages to navigate directly to resources for older adults (Becker, 2004b). Less than 8% of these sites had a descriptive link on the homepage linking to senior resources. Those Web sites having nondescript links required trial and error searches for senior resources. This type of navigational complexity impedes the use of electronic government by many older adult users.

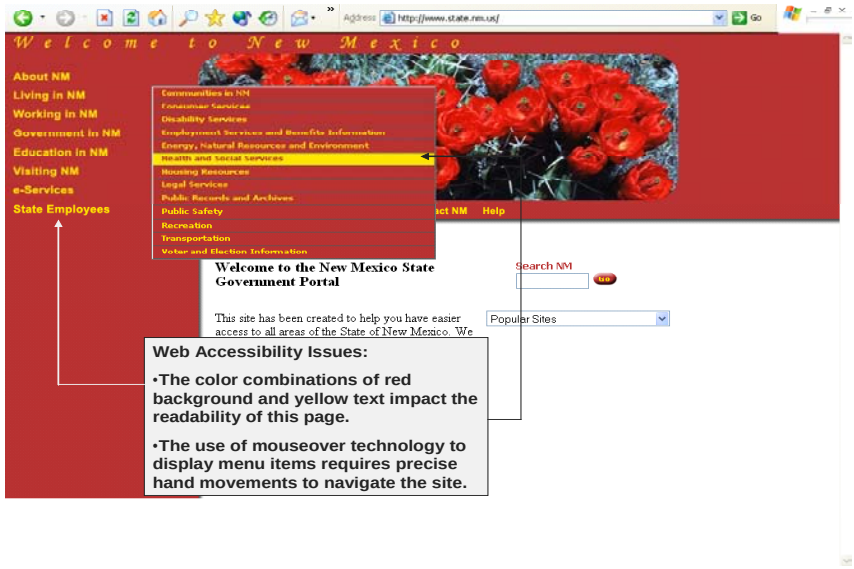
Physical Impairments

Older adults experience a decrease in motor coordination, and as such, may have difficulty with cursor positioning, precise mouse movement, and clicking on links (Chaparro, Bohan, Fernandez, Choi & Kattel, 1999; Ellis & Kurniawan, 2000; Hawthorne, 2000). Figure 2 shows a Web page that requires the precise use of mouseover technology in order to navigate the Web site. This site becomes virtually inaccessible to those users who cannot precisely click on a link and move the mouse to the pop-up menu.

Literacy

The Joint Committee on National Health Education Standards (1995) defines health literacy as the capacity of an individual to obtain, interpret, and understand basic health information and services and the competence to use such information and services in ways that are health-enhancing. This includes the ability to understand instructions on prescription drug bottles, appointment slips, medical education brochures, doctor's directions and consent forms, and the ability to negotiate complex health care systems. In today's wired world, health literacy is expanded to include the comprehension of online health information.

Figure 2. Illustration of Web barriers due to color and mouseovers (http://www.state.nm.us)



The Web is transforming health care by providing online health information to patients, clinicians, caregivers, and family members. There are hundreds of health-related Web sites offering access to unprecedented amounts of health information. Because of their heavy use of health services, older adults could be among the major beneficiaries of Web-accessible health information. But, when the literacy requirements of these pages are misaligned with the literacy skills of older adults, the results could be devastating.

The Web offers an extraordinary opportunity to disseminate timely and much needed health care information to older adults. This opportunity is expanding further with technological advances in wireless communication and increasing availability of community computing resources through institutions such as nursing homes, schools, and libraries.

Older adults in general have a lower education level than younger adults according to the U.S. Census (www.census.gov). The National Adult Literacy Survey found that 71% of older adults performed in the lowest two levels of literacy defined in the survey (Kirsch, Yamamato, Norris, Rock, Jungeblut & O'Reilly, 2001). Approximately 67% of older adults appeared to have difficulty with finding and processing quantitative information when compared to younger adults (Brown, Prisuta, Jacobs & Campbell, 1996).

The reading complexity of content targeting older adults is illustrated by the following sentence appearing on the Alabama state Web site (http://www.adss.state.al.us/seniorrx.htm):

"In case you're not familiar with the program, SenioRx is a partnership of state agencies and community organizations designed to assist senior citizens (ages 60 and older) with chronic medical conditions who have no prescription insurance coverage and limited financial means (living below 200% of the poverty level) with applying for drug assistance programs provided by pharmaceutical manufacturers."

The reading grade level associated with this sentence is far beyond a 12th grade level because it is composed of 58 words. In addition, about 33% of the sentence is composed of three or more syllable words, adding to its complexity. Though there is only one sentence in this sample, it illustrates the reading comprehension barriers that older adults may encounter while surfing the Web.

WEB ACCESSIBILITY INITIATIVES

The National Institute on Aging in conjunction with the National Library of Medicine has developed Web accessibility guidelines for making sites senior-friendly (NIA & NLM, 2001). These guidelines are based on scientific findings from research in aging and cognition and human factors (Morrell et al., 2002). They provide information on how to improve the design, navigation, and information content of Web sites to remove accessibility barriers for older adult users. Table 1 lists several of these guidelines for promoting the development of accessible Web sites. Note that several

Table 1. NIA/NLM guidelines for making senior-friendly Web sites

Guideline	**Description**
Sans serif typeface	Sans serif font types should be used to display information content because they are not condensed.
12 point or greater font size	The use of a large font size improves legibility of information content such that text body, buttons, links, images, and other textual objects are readily seen by an older adult.
Mixed case letters in text body	The text body should be in mixed case text to improve readability. Upper case text should be reserved for headlines on a page.
Left justification	Text should be left justified because spacing between letters is consistently the same.
Plain background images	Patterned background images should be removed from a Web page because they reduce the legibility of text overlaying them.
Text effects only in headlines	Text effects including underlining, italics, bold, or strikethrough should not be used in the body of the text.

of these guidelines, font size, and patterned background images were not followed in the design of the government Web page previously shown in Figure 1.

Another initiative is the National Cancer Society's usability.gov Web site, which provides information about making Web content more usable, accessible, and useful (*www.usability.gov*). It provides research and practitioner-based guidelines on design layout, navigation, information content, and other usability aspects of Web design targeting the general population of computer users. Though it does not specifically target the usability needs of older adults, many of these resources will improve Web site usability for this user group.

Nonprofit groups, including the SPRY foundation (*www.spry.org*), Seniornet (*www.seniornet.org*), and AARP (*www.aarp.org*), provide online resources and support research in Web accessibility for older adults, health literacy, and related areas. Much of the research conducted by these groups provides feedback on the digital divide, health literacy issues, community resources, and the status of older adults getting and staying online.

From a broader user perspective, there have been initiatives in promoting Web accessibility for those with disabilities. Microsoft, IBM, Apple, Watchfire and other corporations have provided resources, tools, and techniques for improved accessibility of Web designs. The Trace Research and Development Center at the University of Wisconsin (*www.trace-center.org*) is one of several centers that focus on universal usability in order to make technology accessible to anyone, anytime, and anyplace. The ACM SIGCAPH (Special Interest Group on Computers and the Physically Handicapped) Web site provides a full listing of Web accessibility resources (*http://www.hcibib.org/accessibility/#ORGANIZATIONS*).

Section 508, an amendment to the 1973 Rehabilitation Act, was enacted by the U.S. government in order to eliminate information barriers for those persons with disabilities. It requires that individuals with or without disabilities have equal access to information provided by federal agencies (www.Section508.gov). Web content guidelines have been put forth by the World Wide Web Consortium (www.w3c.org/WAI/) in order to eliminate accessibility barriers. Though Section 508 does not specifically address the barriers to Web use due to the normal aging process, many guidelines on design layout, information content, navigation, and design consistency improve usability from an older adult perspective.

CONCLUSION

Though significant strides have been made to promote Web accessibility for older adults, there are still barriers to overcome. Web accessibility and online literacy research requires further study in determining an optimal presentation of content to meet the literacy needs of older adult users. The NIA/NLM guidelines recommend active versus passive voice sentence structures, short sentences, and appropriate reading grade levels. Because these guidelines are rather vague, it may be difficult to enforce them during the design of Web content. For example, what is a sufficiently "short" sentence in terms of word count and syllable complexity? Cultural diversity and English proficiency of the older adult population also require further study in terms of potential barriers to Web use.

REFERENCES

Administration on Aging. (2002). A profile of older Americans: 2002. Administration on Aging, U.S. Department of

Health and Human Services, http://www.aoa.gov/ prof/Statistics/profile/2002profile.pdf

American Foundation for the Blind. (1999). Normal changes in the aging eye fact sheet. Retrieved June 20, 2003, from http://www.afb.org/info_document_view.asp?documentid=203

Becker, S.A. (2004a). E-government visual accessibility for older adult users. *Social Science Computer Review, 22,* 1.

Becker, S.A. (2004b). Architectural accessibility and reading complexity of U.S. state e-government for older adult users. Forthcoming in *Electronic Government.*

Brown, H. Prisuta, R. Jacobs, B., & Campbell, A. (1996). *Literacy of older adults in America: Results from the national adult literacy survey.* U.S. Department of Education, National Center for Education Statistics, NCES 97-576, Washington DC.

Chaparro, A., Bohan, M., Fernandez, J.E., Choi, S.D., & Kattel, B. (1999). The impact of age on computer input device use: Psychophysical and physiological measures. *International Journal of Industrial Ergonomics, 24,* 503-513.

Coulson, I. (2000). Introduction: Technological challenges for gerontologists in the 21st century. *Educational Gerontology, 26,* 307 - 315.

Czaja, S.J., Guerrier, J.H., Nair, S.N., & Landauer, T.K. (1993). Computer communications as an aid to independence for older adults. *Behaviour and Information Technology, 12,* 197 - 207.

Echt, K.V. (2002). Designing Web-based health information for older adults: Visual considerations and design directives. In R.W. Morrell (Ed.), *Older adults, health information, and the World Wide Web* (pp. 61 - 88). Mahwah, NJ: Lawrence Erlbaum Associates.

Ellis, R.D., & Kurnaiwan, S.H. (2000). Increasing the usability of on-line information for older users. A case study in participatory design. *International Journal of Human-Computer Interaction, 12,* 263-276.

Hawthorne, D. (2000). Possible implications of aging for interface designers. *Interacting with Computers, 12,* 507-528.

Holt, B.J., & Morrell, R.W. (2002). Guidelines for Web site design for older adults: The ultimate influence of cognitive factors. In R.W. Morrell (Ed.), *Older adults, health information, and the World Wide Web* (pp. 109-129). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Joint Committee on National Health Education Standards. (1995). *National health education standards.* Available from the American School Health Association, P.O. Box 708, 7263 State Route 43, Kent, OH 44240.

Kerschner, P.A., & Chelsvig Hart, K.C. (1984). The aged user and technology. In R.E. Dunkle, M.R. Haug & M. Rosenberg (Eds.), *Communications technology and the elderly: Issues and forecasts* (pp. 135-144). New York: Springer.

Kirsch, I., Yamamato, K., Norris, N., Rock, D., Jungeblut, A., & O'Reilly, P. (2001). *Technical report and data file users manual for the 1992 national adult literacy survey.* Washington DC: National Center for Education Statistics, U.S. Department of Education, NCES 2001-457.

Lenhart, A., Horrigan, J., Rainie, L., Allen, K., Boyce, A., Madden, M., & O'Grady, E. (2003). *The ever-shifting Internet population.* Washington DC: The Pew Internet and American Life Project.

Morrell, R.W., Dailey, S.R., Feldman, C., Mayhorn, C.B., & Echt, K.V. (2002). *Older adults and information technology: A compendium of scientific research and Web site accessibility guidelines.* Bethesda, MD: National Institute on Aging.

NIA & NLM. (2001). *Making your Web site senior friendly: A checklist.* National Institute on Aging and National Library of Medicine. Retrieved July 30, 2003, from *http://www.nlm.nih.gov/pubs/checklist.pdf*

Preidt, R. (2003). Seniors turning to Internet for health help. Retrieved September 23, 2003, from *http://www.healthscout.com/template.asp?page=newsdetail& ap=1&id=511832*

Scanlon, B. (2001, October). The future of the net: Surf's up for seniors. Retrieved August 21, 2003, from *http://www.eweek.com/article2/0%2C3959%2C950787%2C00. asp*

KEY TERMS

Digital Divide: The digital divide is a term used to describe the disparity between persons who have access to information and computing technology and those who do not. Often, it is used to describe the lack of Internet accessibility to those living in rural or remote areas or who lack computing knowledge and skills.

Electronic Government: Electronic government (e-government) refers to the use of information and computing technologies by government agencies to deliver services, information, and resources to citizens, businesses, and other organizations.

Health Literacy: Health literacy is the capacity of an individual to obtain, interpret, and understand basic health information and services and the competence to use such

information and services in ways that are health-enhancing (refer to text box 1).

Older Adults: An older adult is defined as a person who is 60 years or older in the National Institute on Aging's guidelines on making senior-friendly sites.

Visual Aging: Visual aging takes into account age-related changes in vision that have consequences on daily activities. In this article, the consequences are related to using the Web.

Web Accessibility: Web accessibility means that any person, regardless of disabilities, is able to use Web technology without encountering any barriers.

Web Usability: Web usability refers to the user satisfaction associated with a Web site. It typically includes the effectiveness of the site in meeting the needs of the user. It also includes the site's performance, reliability, and overall efficiency in supporting specified user goals.

ENDNOTES

[1] The impact on readability is minimized when displaying in black and white the patterned background image and its overlaying text and foreground and background color combinations.

The National Science Foundation under Grant No. 0203409 supports this work. Any opinions, findings and conclusions or recommendations expressed in this content are those of the authors and do not necessarily reflect the views of the National Science Foundation. Web accessibility for older adult research efforts can be found at: http://www.cba.nau.edu/facstaff/becker-a/Accessibility/main.html

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 3036-3041, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Web Accessibility and Compliance Issues

Shirley Ann Becker
Florida Institute of Technology, USA

INTRODUCTION

The impetus for accessible electronic and information technology was driven by federal initiatives with the objective of "bridging the digital divide" (U. S. Department of Commerce, National Telecommunications and Information Administration, 2000). This initiative focused on improving quality and longevity of life, addressing social disparities, promoting small businesses, and providing educational opportunities, among others. As an outgrowth of this initiative, the concept of building an "information society for all" was promoted in the form of universal usability of all electronic and information technology. The long-term goal was to ensure that no one was left behind in terms of inaccessible electronic and information technology.

In 1998, congress amended the Rehabilitation Act of 1973[1] with section 508 to require federal agencies to make electronic and information technology accessible to people with disabilities. Section 508 was enacted to eliminate barriers in electronic and information technology, make available new opportunities for people with disabilities, and encourage the development of technologies that will help achieve these goals (www.section508.gov). The law applies to all federal agencies when they develop, procure, maintain, or use electronic and information technology. Under section 508, agencies must give disabled federal government employees and citizens access to information that is comparable to the access available to others without disabilities.

The Internet and supporting technological advances opened doors for government employees and citizens to access electronic information that in the past were not readily available. However, those with disabilities or normal aging considerations found it difficult if not impossible to use basic technology that nondisabled individuals could use freely (McLawhorn, 2001). This disparity of access to electronic data and information was addressed by congress when it amended the Rehabilitation Act (1973) with section 508. Congress recognized that the federal government is the largest technology consumer in the U.S.; and as such, it can influence the design and manufacture of accessible technologies and supporting products.

BACKGROUND

Section 508, which went into effect in June 2001, requires all federal agencies to comply with accessibility standards administered by the Architectural and Transportation Barriers Compliance Board (referred to as the Access Board).[2] These standards ensure that electronic and information technology is accessible to disabled persons to the extent it does not pose an undue burden on an agency. When section 508 went into effect, federal agencies could no longer procure noncompliant electronic and information technology (Charles, 2001). This meant that vendors, who supply hardware, software, Web, telecommunications, and other information technologies, must ensure compliance with section 508 accessibility in order to obtain government contracts.

The Access Board put together the Electronic and Information Technology Access Advisory Committee (EITAAC) in order to develop section 508 standards. The EITAAC is comprised of industry, government, academic, and disability advocacy organizations. The EITAAC (1999) developed generic standards that were organized into three areas including: (1) accessibility of operation and information, (2) compatibility with peripheral devices, and (3) documentation and services associated with electronic and information technology. The committee made recommendations for implementation of section 508, formalized a definition of electronic and information technology for interpreting the statute, and developed recommendations for procurement processes.

The Access Board defines electronic and information technology as, "information technology and any equipment or interconnected system or subsystem of equipment used in the creation, conversion, or duplication of data or information" (U.S. Access Board, 1999). This definition encapsulates telecommunications, information kiosks, transaction machines, Web sites, copiers, faxes, and other multimedia office equipment. It does not include embedded information technology; back office equipment used only by service personnel for maintenance, repair, or similar purposes; or computer hardware and software, equipment, services, and other resources that automatically manipulate, acquire, store, manage, move, control, display, switch, interchange, or transmit data or information (McLawhorn, 2001).

Table 1. Section 508 technology standards (Federal Register, 2000)

Technical Categories	Subsection
Software applications and operating systems. The criteria primarily focuses on software specifications related to vision disabilities including navigation, animated displays, color and contrast settings, flash technology, electronic forms and others.	1194.21
Web-based information or applications. The criteria used is based on the access guidelines put forth by the World Wide Web Consortium (W3C). The focus of these standards is on accessible, federal government Web sites and private sector Web sites under contract to a federal agency.	1194.22
Telecommunications products. The criteria is designed primarily to ensure access for deaf or hard of hearing persons. The standards require technology that is compatible with hearing aids, implants, and devices for communicating over a telephone, among others.	1194.23
Video or multimedia products. The criteria focuses on video programs, computer generated presentations, and other products that use more than one media. The standards require alternate presentations (user-selectable) for training and informational multimedia productions developed or procured by federal agencies.	1194.24
Self-contained, closed products. The criteria focus primarily on products that have embedded software that cannot readily support assistive technology such as kiosks, copiers, printers, and faxes. As such, assistive technology must be built into the product with requiring assistive technology devices.	1194.25
Personal computers and portable computers. The criteria focus on keyboards and other mechanically operated controls, touch screens, ports, and connectors, among others.	1194.26

Table 2. Statutes and laws related to section 508 of the Rehabilitation Act (1973)

Statute or Law	Description
American with Disabilities Act (ADA) of 1990	American with Disabilities Act (1990) prohibits discrimination and ensures equal opportunity for persons with disabilities in employment, state and local government services, public accommodations, commercial facilities, and transportation (http://www.usdoj.gov/crt/ada/pubs/ada.txt).
Assistive Technology Act of 1998	The Assistive Technology Act (1998) establishes a grant program, administered by the U.S. Department of Education, to provide federal funds to support state programs addressing assistive technology needs of individuals with disabilities (http://www.section508.gov/docs/AT1998.html).
Section 501 of the Rehabilitation Act	Section 501 prohibits discrimination on the basis of disability in federal employment and requires federal agencies to establish affirmative action plans for the hiring, placement, and advancement of people with disabilities in federal employment (www.section508.gov/index.cfm?FuseAction=Content&ID=17).
Section 504 of the Rehabilitation Act	Section 504 prohibits discrimination based on disability in federally funded and federally conducted programs or activities in the U.S. including employment programs (www.section508.gov/index.cfm?FuseAction=Content&ID=15).
Section 505 of the Rehabilitation Act	Section 505 establishes enforcement procedures for title V of the Rehabilitation Act. Section 505 (a) (1) specifies that procedures and rights set forth in section 717 of the Civil Rights Act of 1964 shall be available with respect to any complaint under section 501. Section 505 (a) (2) specifies that remedies, rights and procedures set forth in title VI of the Civil Rights Act of 1964 shall be available to any person alleging a violation of section 504. Section 508 is also enforced through the procedures established in section 505 (a)(2) (http://www.section508.gov/index.cfm?FuseAction=Content&ID=18).
Section 255 of the Telecom-munications Act of 1996	Section 255 of the Telecommunications Act (1996) requires manufacturers of telecommunications equipment and providers of telecommunications services to ensure that equipment and services are accessible to persons with disabilities, if readily achievable (www.fcc.gov/telecom.html).

Table 1 summarizes each section of the technology standards put forth by the board and as recommended by the EITAAC. The technical categories in the table focus on the functional capabilities covered under section 508. These standards focus on assistive technologies (e.g., screen reader devices) and alternative technologies (e.g., keyboard navigation instead of mouse navigation) that allow access to those with disabilities. Though not specifically described in the table, each standard outlines the technical and information dissemination requirements for the use of electronic and information technologies.

Though section 508's enforcement mechanisms apply only to procurement, both sections 501 and 504 of the Rehabilitation Act (1973) require accommodations for individuals with disabilities. (Table 2 provides definitions of sections 501 and 504 along with other related statutes and laws.) As such, federal agencies cannot use section 508 to avoid accessibility requirements and must provide alternative means of access to information for federal employees and individuals with disabilities. If an undue burden claim prevents procurement of accessible electronic and information technology, then a federal agency must provide alternative means of access (www.section508.gov)

SECTION 508 AND WEB ACCESSIBILITY

The W3C, as the standards setting body for the Web, has provided support to the Web Accessibility Initiative (WAI) in the development of Web Content Accessibility Guidelines (WCAG) (Chisolm, Vanderheiden, & Jacobs, 2001). These guidelines enforce the section 508 Web accessibility standard (refer to www.w3.org/TR/WCAG10/ for more information). They continue to play an integral role in the implementation of section 508 by federal government and contracting vendors as mandated by law. They are also used by commercial, education, nonprofit, and other government sectors in voluntarily promoting Web accessibility for all users.

Web Accessibility Guidelines

The WAI guidelines for Web content accessibility focus on making online information accessible to those with disabilities. This includes text, images, links, audio, and other elements that compose a Web page or application. These guidelines are briefly described.

1194.22 (a) A text equivalent for every non-text element shall be provided. Pictures, graphs, and other elements on a Web page not in electronically readable form are supplemented with a text description to be read by screen reader software and Braille displays.

1194.22 (b) Equivalent alternatives for any multimedia presentation shall be synchronized with the presentation. Captioning provided for audio information as well as audio descriptions of visual information are examples of equivalent alternatives that must be synchronized for accessible Web content.

1194.22 (c) Web pages shall be designed so that all information conveyed with color is also available without color. Color alone cannot be used to convey information as it may render the Web site inaccessible to a user with color deficient vision.

1194.22 (d) Documents shall be organized so they are readable without requiring an associated style sheet. A style sheet is a file that contains instructions on how information is to be presented to the user. It is important to design style sheets such that specialized software can provide access to information by a disabled user.

1194.22 (e) Redundant text links shall be provided for each active region of a server-side image map. A server-side image map sends coordinates of the image map to a computer server that then identifies the URL (link) being selected. For those who cannot accurately click on a specific region, it would be important to provide redundant links on the map.

1194.22 (f) Client-side image maps shall be provided instead of server-side image maps except where the regions cannot be defined with an available geometric shape. A client-side image map associates a link with a particular image map region. It would be important to use client-side instead of server-side image maps whenever possible for improved accessibility.

1194.22 (g) Row and column headers shall be identified for data tables.

1194.22(h) Markup shall be used to associate data cells and header cells for data tables that have two or more logical levels of row or column headers. Descriptive identifiers for each row and column header allow a screen reader device to read table data in a meaningful format.

1194.22 (i) Frames shall be titled with text that facilitates frame identification and navigation. For those sites using frames, it is important to label each frame with descriptive information for accessible navigation.

1194.22 (j) Pages shall be designed to avoid causing the screen to flicker with a frequency greater than 2 Hz and lower than 55 Hz. Objects that flicker on a Web page may be a problem for users with photosensitive epilepsy, as there is the potential for triggering a seizure when flickering in the 4 to 59 flashes per second (Hertz) range.

1194.22 (k) A text-only page, with equivalent information or functionality, shall be provided to make a Web site comply with the provisions of this part, when compliance cannot be accomplished in any other way. The content of the text-only page shall be updated whenever the primary page changes. For those Web pages that cannot be designed to be compliant with the standards put forth by section 508, a text-only page can be provided.

1924.22 (l) When pages utilize scripting languages to display content, or to create interface elements, the information provided by the script shall be identified with functional text that can be read by assistive technology. Some Web sites use scripting language to customize a page with the display of dynamic content. It is important that text is associated with these elements to ensure that a visually impaired user is provided meaningful information.

1924.22 (m) When a web page requires that an applet, plug-in or other application be present on the client system to interpret page content, the page must provide a link to a plug-in or applet that complies with §1194.21(a) through (l). For example, PDF files require the downloading of Adobe Acrobat Reader software. As such, a download link would be provided adjacent to the PDF file.

1194.22 (n) When electronic forms are designed to be completed online, the form shall allow people using assistive technology to access the information, field

elements, and functionality required for completion and submission of the form, including all directions and cues. A Web form requires a label (e.g., "Last Name") adjacent to each element in order for the screen reader to present meaningful information to the user.

1194.22 (o) A method shall be provided that permits users to skip repetitive navigation links. Most Web sites have repetitive links on the top, side navigation bars, and bottom of a page. This design method allows repetitive links to be ignored by a screen reader device.

1924.22 (p) When a timed response is required, the user shall be alerted and given sufficient time to indicate more time is required. Web pages may time out when a response is not received within a specified time period rending the site inaccessible. An alert feature would warn a user of the impending time out.

Web Accessibility Compliance

According to Reed (2003), Matthews (2002), and others, federal agencies were initially slow to adopt compliance practices aggressively. This is reflected in early studies in section 508 Web accessibility compliance. Stowers (2002) conducted a study on 148 federal Web sites and found that only 13.5% of the sites met the accessibility standards of section 508. Additional studies were carried out by the World Markets Research Center and Brown University (2001) and West (2002) in assessing accessibility features of government sites. These studies also uncovered the lack of full compliance with the section 508 Web accessibility standard.

Since these early studies, most federal sites have made great strides in complying with section 508 through the use of automated support, such as provided by Watchfire Bobby™ 5.0, Microsoft FrontPage®, and Macromedia's Dreamweaver®. These tools provide feedback on noncompliance errors associated with the Web content accessibility guidelines. Unfortunately, there are still accessibility barriers even when automated tools report no compliance errors. A study by Becker (2005), for example, found that 60% of sampled federal Web sites had nondescript labels on navigational links. Though these would not be flagged as noncompliance errors by automated tools, they pose accessibility barriers when using assistive technology given that they provide insufficient information for navigating the site.

An illustration of this type of accessibility barrier is the text equivalent message "Americans with Disabilities" associated with an icon on the U.S. Small Business Administration Web site (http://www.sbaonline.sba.gov/). Though this message is in compliance with *1194.22 (a) A text equivalent for every non-text element shall be provided*, a user relying on a screen reader device would be unaware that the icon links to the Americans with Disabilities Act Web page. In contrast, the U.S. Treasury Department Web site (www.ustreas.gov) has a meaningful text equivalent message associated with its graphic. The message, "Social Security Banner that links to http://www.strengthening socialsecurity.gov," when read aloud by assistive technology relays information about the graphic as well as the navigational link associated with it.

FUTURE TRENDS

Government, nonprofit organizations, academic institutions, and commercial vendors continue to provide information, resources, and tools to assist in both mandatory and voluntary compliance with section 508. The WAI guidelines continue to provide important information about each accessibility barrier and potential solutions to it. The University of Wisconsin's Trace Center (2004) has recently collated the Access Board's 508 final rule and guidelines. John Hopkins University offers an updated list of Web accessibility resources and tools (Reinzi, 2005).

However, more needs to be done in promoting Web accessibility beyond a strict interpretation and adherence to standards. Though compliant with section 508, the lack of a meaningful text equivalent description for a graphic or image is an accessibility issue that needs to be addressed. For those with unsteady hands, arthritis, or other motor skill disabilities, mouseover technology may pose an impediment to navigating a site. For users with vision impairments, insufficient contrast between foreground and background colors may impede readability of Web contents. A Spanish version of a federal Web site that has nontranslated components becomes inaccessible when read aloud by assistive technology to a user with low or no English proficiency.

Jaeger (2003) recommends a rating system for no, partial, or full compliance; as well as, an identification component for whom the site is accessible. Future research is needed to explore the potential for this type of rating that could expand voluntary compliance with section 508 by nongovernmental entities.

CONCLUSION

Section 508 has made a significant impact on bridging the digital divide especially for those with disabilities. It has provided the impetus for building e-government sites that are universally accessible to all citizens. However, universal Web accessibility cannot be achieved through the sole use of automated technologies or strict adherence to section 508 guidelines (Brewer, 2004). Automated technologies can only guide the developer through an accessibility assessment and may not uncover barriers related to meaningful information content. Full compliance with section 508 will require inspection of Web elements including the good use of color,

meaningful labels on graphics and links, and full translation of content, among others.

REFERENCES

Becker, S. A. (2005). Technical opinion: E-government usability for older adults. *Communications of the ACM, 48*(2), 102-104.

Brewer, J. (2004). Web accessibility highlights and trends. *Proceedings of the 2004 International Cross-disciplinary Workshop on Web Accessibility (W4A)* (pp. 51-55). New York: ACM Press.

Charles, S. (2001). Section 508 compliance—Not a simple yes or no. Immix Group. Retrieved November 14, 2005, from http://www.immixgroup.com/pslibrary/eu_display.cfm?ID=91

Chisolm, W., Vanderheiden, G., & Jacobs, I. (2001). Web content accessibility guidelines. *Interactions of the ACM, 8*(4), 35-54.

Electronic and Information Access Advisory Committee (EITAAC). (1999). *Electronic and information access advisory committee recommendations for accessibility standards: Electronic and information technology.* Retrieved October 9, 2005, from http://www.access-board.gov/sec508/commrept/eitaac.pdf

Federal Register (2000). Electronic and information technology accessibility standards. Architectural and Transportation Barriers Compliance Board, Part II. *Federal Register, 65*(246), December 21. Retrieved June 24, 2006 from http://www.access-board.gov/sec508/508standards.pdf

Jaeger, P. T. (2003). The importance of accurately measuring the accessibility of the federal electronic government: Developing the research agenda. *Information and Technology and Disabilities E-Journal, 9*(2). Retrieved March 1, 2006, from http://www.rit.edu/~easi/itd/itdv09n1/jaeger.htm

Matthews, W. (2002). One year and counting: Section 508. *Federal Computer Week.* Retrieved June 24, 2006, from http://www.fcw.com/article77009-06-24-02-Print

McLawhorn, L. (2001). Recent development: Leveling the accessibility playing field Section 508 of the Rehabilitation Act. *North Carolina Journal of Law & Technology, 3*(1), 63-100.

Reed, M. A. T. (2003, August 11). Agencies still on the learning curve. *Federal Computer Week*. Retrieved October 22, 2005, from http://www.fcw.com/article80520-08-11-03-Print

Reinzi, G. (2005, May 9). JHU focuses on online accessibility. *The Gazette, 34*(33). Retrieved November 21, 2005, from http://www.jhu.edu/~gazette/2005/09may05/09access.html

Stowers, G. N. L. (2002). *The state of Federal Websites: The pursuit of excellence.* PricewaterhouseCooper Endowment for the Business of Government report. Retrieved June 24, 2006, from http://www.businessofgovernment.org/pdfs/StowersReport0802.pdf

University of Wisconsin, Trace Center. (2004). *Trace center collation of Access Board's final rule and guides. University of Wisconsin-Madison.* Retrieved November 21, 2005, from http://trace.wisc.edu/docs/508-collation/06092004v1.1.pdf

U.S. Access Board (1999). *Questions & Answers about Section 508 of the Rehabilitation Act Amendments of 1998*. Retrieved June 24, 2006 from, http://www.access-board.gov/sec508/FAQ.htm#14

U.S. Department of Commerce, National Telecommunications and Information Administration. (2000). *Falling through the net: Toward digital inclusion. A report on Americans' access to technology tools.* Retrieved November 15, 2005, from http://www.ntia.doc.gov/ntiahome/fttn00/contents00.html

West, D. M. (2002). State and federal e-government in the United States. Center for Public Policy, Brown University, Providence, Rhode Island. Retrieved March 18, 2006, from http://www.insidepolitics.org/Egovt02us.html

World Markets Research Center and Brown University. (2001). *Global e-government survey.* Providence, RI: World Markets Research Center.

KEY TERMS

Assistive Technology: Equipment, device, or other product that assists a disabled user in performing tasks that otherwise would be difficult or not possible to accomplish.

Color Deficiency: Color deficient vision results in an inability to distinguish certain colors and shades when compared to normal vision.

Digital Divide: The digital divide is the gap between those who have access to electronic and information technology and those who do not.

Electronic and Information Technology: Information technology and any equipment or interconnected system or subsystem of equipment used in the creation, conversion, or duplication of data or information (www.access-board.gov).

Motor Disabilities: Physical impairments that can impede movement, coordination, or sensation. They can include weakness and lack of muscle control.

Screen Reader: Speech synthesis software used by a vision-impaired person to read aloud what is displayed on a computer screen.

Section 508: Amendment to the 1973 Rehabilitation Act requiring federal agencies to make electronic and information technology accessible to people with disabilities.

Web Accessibility: Web accessibility means that a person, regardless of disabilities, is able to use Web technology without encountering any barriers.

ENDNOTES

1 Rehabilitation Act of 1973, *as amended by* § 508, 29 U.S.C. § 794(d) (1998).

2 Architectural and Transportation Barriers Compliance Board is an independent federal agency, established by Section 502 of the 1973 Rehabilitation Act. Its primary function is to promote accessibility for individuals with disabilities.

Web-Based GIS

Anselmo Cardoso de Paiva
University of Maranhão, Brazil

Cláudio de Souza Baptista
University of Campina Grande, Brazil

INTRODUCTION

Currently, the Internet is becoming the main vehicle for publishing geographical information, which enables data interchange, analysis, and geographical data visualization. The rapid evolution of Web technology has lead to an improvement of the geographical information utilization and availability.

The rise of the Internet has created an infrastructure ideally suited to the widespread distribution and dissemination of geographical information. By using Internet GIS applications, users may view, query, analyze, and download spatial information from anywhere at anytime. While this improvement provides new opportunities for public domain as well as commercial use of spatial datasets, new problems arise. One of them is the problem of transferring data efficiently from server to client. Geographical datasets are generally very large, and this process may demand too much time.

The Internet has created an interesting environment for geospatial data sharing, in which data providers make their databases available through the Web, and users may transfer, visualize, manipulate, and interact with them (Bertolotto & Egenhofer, 1999). This environment introduces new problems that must be addressed to make possible an efficient and effective use of these datasets. One such problem is related to the availability of huge geospatial data amounts in repositories with limited connection bandwidth (Flewelling & Egenhofer, 1999).

One of the main concerns on Web-based geographical information systems (Web GIS) is related to performance issues, as these datasets need to migrate, as fast as possible, from server to client tiers (Peng & Tsou, 2003). This is the problem of generation and transmission of the digital maps, which are suitable for user needs. Another important issue is to enable fast and easy Internet GIS application deployment.

In this chapter, we discuss solutions proposed for Web GIS based on the vector format, particularly the iGIS framework (Baptista, Leite Jr., Silva, & Paiva, 2004). The iGIS is a Web GIS that renders maps using multiresolution techniques, and enables user interaction them, using the W3C SVG specification. The chapter also addresses the issues that affect Web GIS performance, presenting an overview of the techniques that increase Web GIS performance and the results obtained by using the iGIS framework.

BACKGROUND

Web GIS applications evolved from the delivery of static maps conveyed in raster formats (e.g., JPEG, GIF, PNG) to an actual stage in which users may choose the window, the scale, and possibly also the data layers to be displayed, and the map is generated dynamically from a database and transmitted in the vector format. Web mapping deals at least with two basic problems: map generation, and map transmission.

Regarding map generation, there are, basically, two approaches. The first one is based on the use of a database that stores geometry information at the most precise scale, and each visualization at less precise scales are automatically derived from that, mainly through cartographic generalization (Müller, Lagrange, Weibel, & Salgé, 1995; Weibel & Dutton, 1999). This implies the execution of map generalization procedures on the fly (Harrie, Sarjakoski, & Lehto, 2002; Lehto & Kilpelainem, 2001). The second approach utilizes a multiresolution database in which spatial objects may be associated to a variety of geometric representations that are scale dependant (Zhou, Prasher, & Kitsuregawa, 2001). In this case, it is necessary to process an off-line computation of a multiscale database containing several independent levels of detail.

Associated with the map generation problem, there is also the problem of vector map transmission. A possible solution is the use of progressive data transmission that is well known and successfully applied to raster images, in which coarser versions of the data are displayed before the complete image is downloaded.

One of the first works on progressive vector data transmission on the Internet has been proposed by Buttenfield (1999, 2002). His work focuses on the hierarchical subdivision of vector data by using tree structures based on the Douglas-Peucker generalization algorithm (Douglas & Peucker, 1973) similarly to the BLG tree by (Oosterom & Schenkelaars, 1995).

Bertolloto and Engenhofer (2001) did a work based on progressive transmission of vector data in conjunction with on-the-fly mapping over the Internet. They proposed the creation of multiple dataset representations corresponding to different levels of details. These levels of details are then sequentially transmitted and added to the currently displayed representation. Although the proposal is very interesting, it remains an open issue; the integration of two different levels of details dataset on the client side in a way that maintains the visualization completely and correctly at each step.

WEB-BASED GIS: DEVELOPMENT TOOLS AND PERFORMANCE ANALYSIS

The iGIS framework makes it possible and easy to build a GIS application in few minutes. By configuring an XML file using the deployment tool, a user can put any application online. Firstly, it is necessary to populate the database. Then, the user sets data location by providing the parameters: login, password, and database URL; the layers to be displayed; the colors that are going to be used; and other parameters.

The iGIS architecture is based on three tiers: presentation, application, and database. In the presentation layer, Java Server Pages, JSP, are used for implementing dynamic pages. Scalar vector graphics (SVG) format and JavaScript are used for map drawings and graphical tools. SVG was chosen due to many reasons including the fact of being a W3C recommendation; the ability of using vector maps with client manipulation operations such as zooming and panning; and map data compression.

The application layer is responsible for the business logic, and it is composed of the following modules:

- Data loader, which is responsible for loading data from different data sources, configured in XML. To add a new data source, it is necessary to extend some classes from this package;
- Application data model, which uses the OpenGeo-Spatial Simple Features standard for implementing the system logic classes;
- Data formatter, which formats data according to the rendering type chosen. This module is extensible and is configured using XML. Currently, the iGIS uses SVG for rendering; and
- Static maps, which loads and stores SVG maps, improving system performance.

Finally, the data layer contains different data sources. Currently, there are drivers for Oracle, IBM DB2, and Postgresql database servers, and spatial data in ESRI shapefile format. Again, this module can be extended to support other data source types.

iGIS copes with spatial and nonspatial data in an integrated way: both data types are stored in the database server and manipulated via data manipulation language and data definition language statements. For example, information about a hospital would include name, address, number of rooms, physicians, equipment, and its latitude and longitude coordinates. Fortunately, there are several database servers with support to spatial dimension, such as Oracle, IBM DB2, Postgresql, MySQL, and, more recently, the Microsoft SQL Server. Therefore, by using SQL statements, users may submit spatial queries together with nonspatial ones.

iGIS works with different measure units that facilitate internationalization. Moreover, iGIS implements multiresolution by enabling users to navigate through several levels of detail in a map. For example, an application initially presents information at country level of detail. Then, after some operations of zooming in, it can present information at state level of detail, then at city level, and so on. In order to achieve that, iGIS deals with static and dynamic map generation.

On the other side, it is also important to address the aspects that affect the Web GIS performance once it is required to a Web-based GIS an efficient data transmission for the publication of spatial information. Furthermore, the provision of adapted data to heterogeneous clients, which may range from mobile to desktop ones, is another important requirement. Thus, the use of optimization techniques in order to obtain both data reduction and just-in-time delivery mechanism are also relevant to reach high performance. Some of these techniques, as pointed out by Baptista, Nunes, Sousa, Silva, Leite Jr., and Paiva (2005), are data simplification, relative coordinates, static maps, multiresolution, compression, on-demand loading, and progressive transmission.

Data simplification is one of the most used techniques for Web GIS optimization. Depending on user needs, a more detailed view may be required in map rendering. However, for some users, it is enough to have a broader view of map contents and some details in few zoom levels. Hence, map rendering can be done according to the details required by users. Moreover, maps are rendered for clients using their specific resolution, which may also introduce information loss.

Data simplification is a process that identifies, a priori, the map information that can be omitted, and removes it before map transmission.The amount of transmitted data without loss in visual quality is reduced. Map coordinates are generally set with high precision; consequently, the number of points is too large. The simplification process will reduce the number of transmitted points.

In general, vector graphics languages accept the representation of polygonal in two ways: using either absolute

or relative coordinates. Absolute coordinates specify each point position based on a fixed coordinate system. Relative coordinates express each point position in either a polyline or a polygon based on the previous point position.

A point expressed using relative coordinates represents the translation of the absolute coordinate point relative to the previous point in either polyline or polygon description. The initial point is always expressed using absolute coordinates.

A Web GIS usually stores its maps in a spatial database system. Hence, each user session needs both to establish a connection to the database system and to load the data. This loading process is time consuming, as a spatial query is executed and the resulting set must be transformed according to the GIS output format, such as SVG, Flash, WebCGM, and so on. As most Web GIS applications are read-only, it is worthy to load these maps statically as soon as they are needed. Thus, the first access to the required map will take more time to be rendered then next access, as the spatial query will be executed, and the transformation for the specific output format will be done. Nonetheless, next accesses to this map will be fast, as it is already formatted in an appropriate transmission format.

Let us consider a map that contains country, city, street, and quarter layers. A better performance can be achieved if maps are presented in different levels of detail. This is the purpose of multiresolution procedure.

Multiresolution divides the maps into cells, forming a hierarchy. Hence, detailed maps are presented progressively (such as streets) as the user zooms in. Each multiresolution level in the hierarchy has a number of zoom levels, and when the last one is reached, the next zooming operation points to another level. The first level is divided into nine subregions, called cells; each of these is represented in a file. The overlapping of cells is necessary to enable the implementation on the client-side of the pan visualization operation.

Each cell adopts the same procedure, that is, it has a number of zoom levels, and then it is subdivided into nine cells, and so on. This subdivision is done recursively, until the last level of multiresolution is reached. A cell is sent to a client only when it reaches the map representing it. This procedure tends to reduce data transmission. Paiva, Silva, Leite Jr., and Baptista (2004) explain this procedure with more detail.

On-demand map loading is another alternative to increase the availability of maps to Web GIS. It has been defined as the creation of a map upon user request, with the following features: appropriated scale; addressing the chosen area of interest; and focus on the required purpose.

Thus, the map geometries are transmitted to the client on the fly, through either a zooming or panning operation. This approach is based on two main concepts: layer visibility and map window.

The layer visibility concept defines that each map layer has a range of scales in which it is visible. This is a common organization task in map generation to avoid the map pollution confounding the user. Thus, at a given scale, some layers may be invisible, and this contributes to the map understanding.

The map window concept defines a region of interest from all geographical databases in which the user is interested; this defines a rectangle area that must be used to clip all the features that do not need to be integrated into the map because they will not be visualized. When a user requests a map, the client sends to the map server a window, and this window is used to locate all the geographic features in the database that intercept it.

As the user requests a change in the interest area (e.g., a pan operation) or in the current map scale (e.g., a zoom operation), the map server will generate a new map appropriated to user request, and will send it to the client side. Apparently, this process may seem slow, but for some class of maps with a large amount of geographical data, this process may improve map visualization performance.

Finally, another important aspect that must be used to improve the performance of Web GIS is the progressive transmission. In the progressive transmission, the map server divides the map into a low-resolution version and a set of incremental versions that, when incorporated to a certain map version, will generate a more detailed map version. The client is responsible for receiving a map detail increment at some level n, and for integrating it into the current map version, which generates a map version at level $n+1$.

Other approaches to this problem explore device restrictions, like visualization resolution. This can determine the level of detail that can be visualized in a specific device. In general, the device can represent less detailed information than the amount that is sent by the map server. The level of detail that cannot be visualized increases the transmission cost, and does not improve the quality of the map visualized by the client. Thus, the elimination of such information reduces response time at the client side without compromising map quality.

Baptista et al. (2005) present an evaluation of the impact of some of these techniques, and we may see that with a combination of these techniques, we may obtain about 63% in the improvement of performance. This was achieved using GZIP filter, simplification with precision value three, and relative coordinate techniques.

Additionally, Costa, Paiva, Teixeira, Baptista, and Silva (2006) present a progressive transmission scheme for the iGIS Web mapping framework. The proposed scheme enables the Internet GIS developer to plan a scale discretization for multiple levels of detail in offline maps. This approach provides, to the user application, a more efficient environment to visualize and interact with geographic data, especially when considering low bandwidth networks.

The proposed scheme is based on the visualization device resolution. The map features are simplified and a compact map representation is generated, aiming to improve response time. To evaluate the impact of our proposed scheme in the visualization task, several experiments were performed. The experimental results have demonstrated that the amount of information transmitted is significantly less in our proposed scheme, being, therefore, very effective for Web-mapping systems.

FUTURE TRENDS

We also expect that soon, we will make the transition from the historical 2-D planimetric maps to the 3-D view of the terrain, completely integrated with a set of multimedia information.

Finally, a major challenge is the integration of current technologies to generate the 4-D GIS (XYZ and time). This will open the way for integration with predictive modeling, to introduce a look into the future, or more traditionally, study the dynamics of various phenomena.

CONCLUSION

This chapter focused on the theme of Web-based GIS. Web GIS is becoming very popular, especially due to the high variety of open source solutions, such as Degree, GeoTools, GeoServer, Mapserver, and others, which makes this complex technological solution more affordable.

We conclude that, most of the time, the proposed optimal solutions, which maximize performance and accuracy of spatial data, may be employed. These solutions are very useful when there are constraints on both network bandwidth and device memory and screen sizes, as it is the case of mobile phones.

REFERENCES

Baptista, C. S., Leite Jr., F. L., Silva, E. R., & Paiva, A. C. (2004). Using open source GIS in e-government applications. In *Electronic Government* (pp. 428-421). Zaragoza, Spain: Springer Berlin/Heidelberg.

Baptista, C. S., Nunes, C. P., Sousa, A. G., Silva, E. R., Leite Jr., F. L., & Paiva, A. C. (2005). On performance evaluation of Web GIS applications. In *16th International Conference on Database and Expert Systems Applications. DEXA Workshops* (pp. 497-501). Copenhagen, Denmark: IEEE Computer Society.

Bertolotto, M., & Egenhofer, M. (1999). Progressive vector transmission. In C. B. Medeiros (Ed.), *7th ACM international symposium on Advances in geographic information systems*. (pp.152-157). Kansas City, MO: ACM.

Bertolotto, M., & Egenhofer, M. (2001). Progressive transmission of vector map data over the World Wide Web. *GeoInformatica*, *5*(4), 345-373.

Buttenfield, B. (1999). Progressive transmission of vector data on the Internet: A cartographic solution. In *19th ICA/ACI Conference* (pp. 581-590). Ottawa.

Buttenfield, B. (2002). Transmitting vector geospatial data across the Internet. In M. J. Egenhofer & D. M. Mark (Eds.), *Second International Conference: Vol. 2478/2002. Geographic Information Science* (pp. 51-64). Berlin: Springer Berlin/Heidelberg.

Costa, D. C., Paiva, A. C., Teixeira, M. M., Baptista, C. S., & Silva, E. R. (2006). A progressive transmission scheme for vector maps in low-bandwidth environments based on device rendering. In *Advances in Conceptual Modeling - Theory and Practice* (pp. 150-159). Tucson, AZ, USA: Springer Berlin/Heidelberg.

Douglas, D., & Peucker, T. (1973). Algorithms for the reduction of the number of points required to represent a digitalized line or its caricature. *The Canadian Cartographer*, *10*(2), 112-122.

Flewelling, D., & Egenhofer, M. (2002). Using digital spatial archives effectively. *International Journal of Geographical Information Science*, *13*(8), 1-8.

Harrie, L., Sarjakoski, T., & Lehto, L. (2002). A variable-scale map for small display cartography. In J*oint International Symposium on Geoespatial Theory*. Ottawa.

Lehto, L., & Kilpelainem, T. (2001). Generalizing XML-encoded spatial data on the Web. In *20th ICA/ACI Conference: Vol. 4* (pp. 2390-2398). Beijing, China.

Müller, J. C., Lagrange, J. P., Weibel, R., & Salgé, F. (1995). Generalization: State of the art and issues. In J. C. Müller, J.P. Lagrange, & R. Weibel (Eds.), *GIS and generalization: Methodology and practice* (pp. 3-17). London, U.K: Taylor & Francis.

Oosterom, P., & V. Schenkelaars, V. (1995). The development of an interactive multiscale GIS. In *International Journal of Geographic Information Systems*, *9*(5), 489-507.

Paiva, A. C., Silva, E. R., Leite Jr., F. L., & Baptista, C. S. (2004). A multiresolution approach for Internet GIS applications. In *15th International Workshop on Database*

and Expert Systems Applications. *DEXA Workshops* (pp. 809-813). Zaragoza, Spain: IEEE Press.

Peng, Z,. & Tsou, M. (2003). *Internet GIS: Distributed geographic information services for the Internet and wireless networks*. John Wiley & Sons.

Weibel, R., & Dutton, G. (1999). Generalizing spatial data and dealing with multiple representations. In P. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), *Geographical information systems: Principles, techniques, management and applications*, vol. 1, (2nd ed,) (pp. 125-155). Cambridge: Geoinformation International.

Zhou, X., Prasher, S., & Kitsuregawa, M. (2002). Database support for spatial generalisation for WWW and mobile applications. In *3rd International Conference on Web Information Systems Engineering*, WISE (pp. 239-246). Singapore: IEEE Computer Society.

KEY TERMS

Framework: A set of software components that provides a foundation structure for an application. Frameworks maximize developer productivity and produce more reliable code, as there is reuse of already tested code, which enables not writing an application from scratch.

GIS: Geographical information system is a system of hardware and software used for storage, retrieval, mapping, and analysis of geographic data.

Map Simplification: A set of algorithms to find simpler representations for each map object.

Multiresolution: A strategy to construct different representations of vector maps in different scales based on the construction of the map objects simplification.

On-Demand Map Loading: A scheme for map transmission that sends, to the client, just the map portion that is being visualized.

Progressive Transmission: Map transmission scheme in which the map server divides the map into a low-resolution version and a set of incremental versions that, once incorporated to a certain map version, will generate a more detailed map version.

Spatial Operators: Set of SQL operators to deal with the spatial dimension of the data, and making possible the definition of spatial queries.

SVG: Scalar vector graphics, is an XML specification of the W3C to deal with vectors graphics in the web.

Web Caching

Antonios Danalis
University of Delaware, USA

INTRODUCTION

The popularity of the World Wide Web has led to an exponential increase of the traffic generated by its users for over a decade. Such a growth, over such a long period of time, would have saturated both the content providers and the network links had Web caching not been efficiently deployed. Web caching can improve the overall performance of the World Wide Web in several ways, depending on the decisions made regarding the deployment of the corresponding caches. By placing caches in strategic positions, the core network traffic can be reduced, the load of a content provider can be scaled down, and the quality of service, as the users perceive it, can be improved. In this article we present an overview of the major design and implementation challenges in Web caching, as well as their solutions.

BACKGROUND

A Web cache can be placed at different positions on the Web and yield different benefits. Namely, it can be placed next to a content provider, in strategic positions inside the network, at the boundaries of local area networks, or inside the host that runs the requesting browser.

Most users tend to access particular pages quite often. An example could be the "home page" that has been set on their browser, or Web sites that contain documentation they are interested in. Furthermore, during browsing, many users visit the same pages multiple times in short periods of time, by using the history of their browsers. To exploit these access patterns, most browsers keep local copies of the frequently or recently accessed pages. This way, the user is served in near zero time, and the network traffic is reduced. Although this technique can have a positive effect on the browsing experience of a single user, it has a minimal effect on the network traffic. This is mostly due to the small disk space used by this type of caching and the lack of sharing between different user caches.

To yield significant results, dedicated computer systems, called proxies, are installed at the edges of local or wide area networks. These systems can achieve significant reduction in the network traffic and improvement in the user perceived quality of service, by filtering and serving the Web requests generated inside the entire network they serve. If a user has defined in the browser's settings a particular proxy to be used, every time he or she requests a Web page the browser will send this request to the proxy. If the proxy happens to have the page, the user will be served promptly without the original content provider being contacted. If the proxy cannot serve the request, it will fetch the appropriate Web objects (such as the text documents, images, applets) from the original server, and possibly keep a copy for later use. Transparent proxies are a special kind of proxy that end users do not explicitly specify in their browser's settings. Rather, the gateway of the network identifies Web requests and forwards them to the proxy. This model is more efficient, and easier to administrate, since all the users of the local network will be using the proxy, and they need not know about possible changes concerning the proxy.

To improve the performance of a content provider, a cache can be deployed on the server side. These caches, also called server accelerators, or reverse proxies, are usually deployed in front of clusters of Web servers (Challenger, Iyengar & Dantzig, 1999) and cache the most popular documents in their main memory. Since a few highly popular documents are in most cases responsible for a large percentage of requests, by using a server-side cache, a Web server can decrease its disk I/O load (Abrams et al., 1995; Markatos, 1996; Tatarinov, Rousskov & Soloviev, 1997) and significantly improve the request serving quality of service. Nevertheless, a reverse proxy benefits only a particular content provider and does not reduce the network traffic.

In addition to client and server-side caching, there exists the approach of adaptive caching (Michel et al., 1998). In this scheme caches can exist at different points inside the network, and be configured to cooperate with one another, in order to improve the overall performance and balance the load. In this model, different caches, potentially belonging to different organizations, are organized into dynamic groups, which each one can join or leave depending on content demand. The communication needed for a group to form and for documents to be located is done through the Cache Group Management Protocol (CGMP) and the Content Routing Protocol (CRP), respectively. Adaptive caching can deal very well with request surges, occurring when some documents become highly popular in a short period of time (usually articles in online newspapers), but assumes that cooperation across administrative boundaries is not an issue.

CRITICAL ISSUES OF WEB CACHING

To improve their effectiveness, caches can be combined in groups to serve requests cooperatively. Such groups may be organized in meshes, as in the case of adaptive caching, or in hierarchical formations, as in the case of the Harvest Cache project (Chankhunthod et al., 1996). In hierarchical designs, caches are usually organized in tree-like structures, and are configured such that child nodes can query their parents and their siblings, but never their children. Although grouping several caches can improve the scalability of the caching system, control messages between a large number of nodes could saturate the parent nodes and the network links (Baentsch, 1997). To make the communication between the caches efficient, several special purpose protocols have been introduced.

The most popular inter-cache communication protocols are ICP, CRP, CARP, cache digest, and WCCP (Melve, 1999). The Internet Cache Protocol (ICP) was developed for the communication of the caches in the Harvest project and was refined within Squid (SQUID). It is a lightweight message format, used by the caches for issuing queries and replies, to exchange information among one another regarding the optimal location from where a certain object can be retrieved. The large number of messages exchanged in ICP, when the number of caches is high, has been shown to impede scalability (Baentsch, 1997; Fan et al., 2000). To avoid similar effects, the Content Routing Protocol (CRP) uses multicast to query cache meshes. CRP exploits the overlapping that exists in cache group formations to propagate queries or popular objects between the groups. To further improve on the performance of the inter-cache communication protocol, cache digests can be used (Fan et al., 2000; Rousskov & Wessels, 1998). Cache digests compact (using a lossy algorithm) the information about the contents of a cache, and make it available to its neighbors. By checking the digests of its neighbors, a cache can identify, with some uncertainty, which neighbors are likely to have a given document. In addition to these protocols, there exist two proprietary ones, CARP and WCCP, designed by Microsoft and Cisco respectively. Unlike the protocols mentioned before, the Cache Array Routing Protocol (CARP) implements routing in a deterministic way. In particular, it uses hashing to identify which member of the proxy array has the requested document. This way it avoids the scalability problems that appear in ICP due to the large number of control messages. The Web Cache Communication Protocol (WCCP) is designed to support transparent caching. The idea is that a router (such as Cisco Cache Engine) that can recognize Web traffic will intercept user requests and redirect them to a cache.

The protocol used for the communication between the Web clients, proxies, and servers is the Hyper-Text Transfer Protocol (HTTP). HTTP runs on top of TCP/IP, which is the most common protocol used in the Internet for reliable transfers, flow control and congestion avoidance. To initialize a connection, TCP sends a special packet (SYN) from the client to the server, and the server responds with an acknowledgment. After initializing the connection, in order to request a page, the client has to send an additional message to the server, and wait for the reply. In the first version of HTTP, this four-step procedure had to take place for all individual objects of a page, such as images, or applets, separately. To reduce the number of round-trips needed, persistent connections, that is, connections that remain open after the retrieval of a file, were proposed (Caceres et al., 1998; Heidemann, Obraczka & Touch, 1997; Mogul, 1995). Persistent connections allow a client to fetch all the components of a Web page by facing the connection startup latency only once. In addition to the multiple steps handshake, slow-start is another feature of TCP that has negative effects in the context of Web transfers. As the name implies, slow starts demand that newly created TCP connections transmit data very slowly and they increase the throughput as they transmit more and more. In the general case, where flow control and congestion avoidance is necessary, this technique is very efficient. In the case of Web transfers though, where the majority is short-lived since most documents are a few kilobytes long (Arlitt & Williamson, 1996; Cunha, Bestavros & Crovella, 1995; Shriver et al., 2001), slow start translates to slow transfer. In order to deal with these issues, HTTP/1.1 (Fielding et al., 1997) was introduced. This new version, among other improvements, supports persistent connections and has been shown to dramatically outperform its predecessor (Nielsen et al., 1997).

In heavily loaded Web caching systems, disk I/O can become a significant bottleneck (Markatos et al., 1999; Mogul, 1999; Rousskov & Soloviev, 1998), because disks are significantly slower than the main memory, and traditional file systems are not optimized to handle Web traffic. To reduce this overhead, Gabber and Shriver (2000) proposed the use of a special purpose file system and Markatos et al. (1999) and Maltzahn, Richardson and Grunwald (1999) proposed the use of special purpose storing policies that reduce the overhead of file creation, deletion, and access. Such policies take into account Web access patterns when they make decisions regarding file placement. For example, in most cases, the HTML text of a Web page and the embedded objects (such as images) of the page will be requested one after the other. By placing these objects in files located close to each other on the disk (Maltzahn, Richardson & Grunwald, 1999), or in neighboring locations within the same file (Markatos et al., 1999), both reading and writing will access almost sequential disk blocks, yielding a significant performance improvement. Additionally, a specialized file system (Gabber & Shriver, 2000) could avoid keeping metadata to improve performance, or could use in-memory data structures to quickly identify cached objects without performing any disk I/O (Tomlinson, Major & Lee, 1999).

Regardless of the method used to access the disk, and no matter how large the storage space is, it is not possible to store the whole Internet. Therefore, there must exist some techniques to decide which documents to keep in the cache and which to replace when the cache fills up. Several algorithms have been proposed, each trying to keep the most valuable documents in the cache. To assign a value to a document, most algorithms take into account one or more parameters such as recency of access, frequency of access, document size, and fetching latency, and try to evict the documents with the lowest value. The effectiveness of a replacement policy is measured by Hit Rate (HR) and Byte Hit Rate (BHR). HR is the percentage of requests that are served from the proxy, and BHR is the percentage of bytes served from the cache. The Least Recently Used (*LRU*) algorithm is a replacement policy already known and used in virtual memory paging systems. Based on the fact that the possibility of a file to be requested again drops dramatically as the time since the previous request increases (Cao & Irano, 1997; Rizzo & Vicisano, 2000), it tries to replace the least recently used document, when space is needed for a new one. The *SIZE* algorithm (Rizzo & Vicisano, 2000) considers the large files to be the least valuable, and thus removes the largest document to make room for several smaller. Due to the high popularity of small files (Shriver et al., 2001), *SIZE* is among the best replacement algorithms. Capitalizing on frequency of use, the Least Frequently Used (LFU) algorithm evicts the documents that have been used the least amount of times. Although this algorithm can perform similarly to LRU for many access patterns, its main disadvantage is that it does not evict documents that were popular once, but not any more (such as old newspaper articles). To deal with this issue, Arlitt et al. (1999) presented LFU-DA, which is LFU with Dynamic Aging. Their variation takes into account the age of a document and thus manages to perform better than most existing algorithms in terms of both HR and BHR. To achieve even higher efficiency, researchers have introduced algorithms that combine the different dimensions. Lowest Relative Value (LRV) (Rizzo & Vicisano, 2000), for example, assigns a value to each document, calculated based on its age, popularity and size, and replaces the one with the lowest value. Hybrid (Wooster & Abrams, 1997) takes into account even more parameters, such as bandwidth to the content provider, and time to connect, for calculating the value of a document, and has been shown to outperform most traditional algorithms in terms of both HR and latency reduction. Finally, Cao and Irani (1997) proposed GreedyDual-Size, which consists of a family of replacement policies, all working under the same basic idea, but optimized for different goals. Therefore, depending on the goals a cache wants to achieve, GreedyDual-Size can be configured to maximize the HR, or minimize the average download latency, or reduce the network traffic.

Besides evicting documents when space needs to be made for new ones, a proxy needs to delete documents that are not valid any more, to preserve the consistency of its cache. The most common cache consistency mechanisms are time-to-live (TTL) fields, client polling, and invalidation protocols (Gwertzman & Seltzer, 1996). TTL is actually an estimation of how long the current copy of a document will be valid for. Although this approach can be very useful in cases where this information is known, as in the case of periodically changing news articles, it can lead to waste of bandwidth if the estimation is too conservative, or cache inconsistency if the value of TTL is too large. Adaptive TTL tries to find a better estimation by taking into account the document's age (Gwertzman & Seltzer, 1996). In client polling, every time a document is requested, the cache sends an if-modified-since (IMS) request to the original server. According to the result, the file is either proven valid and served by the cache, or the new version is fetched. This approach generates a large number of control messages, since an IMS is sent to the server for every document requested. To gain the benefits of both techniques, squid (SQUID) uses an adaptive TTL policy to decide when to send an IMS validation request to the server. Invalidation callbacks are an approach where the original server keeps track of the proxies that have a copy of its documents, and contacts them when the documents change. Callbacks provide strong consistency without saturating the network with control messages (Cao & Liu, 1998), but raise privacy and security concerns.

Dynamic pages, either created according to the preferences of each individual user, or updated dynamically to meet real-time changing data, such as the stock market or sports scores, raise important issues in regard to cache consistency. If a dynamic document is cached and served by a traditional proxy, it is very likely to be invalid even before the very next request. On the other hand, if a proxy decides not to cache dynamic objects, it will lose an important percentage of Web traffic (Caceres et al., 1998; Wolman et al., 1999). Markatos (2001) suggests caching of search engine results for hours or days, based on the fact that query results are already several days old. Several approaches deal with dynamic pages by caching static, sub-document objects (Douglis, Haro & Rabinovich, 1997; Meira et al., 1999; Shi et al., 2003).Furthermore, Mohapatra and Chen (2002) and Challenger et al. (1999) construct a graph of the objects that constitute a Web page and upon a change, graph traversal reveals the objects that need to be updated for the page to be reconstructed. The transfer of "deltas", that is, the differences between the cached and the original document, is another technique shown to work well (Mogul et al., 1997; Savant & Suel, 2003). Finally Cao, Zhang and Beach (1998) suggested an approach where proxies can fetch from the server a special purpose program (a CacheApplet) attached to the data, to deal with reconstructing or refetching the page the way the original content provider wishes.

CONCLUSION AND FUTURE TRENDS

Web caching has been extensively deployed during the last years to support the needs of the exponentially growing World Wide Web. Depending on their placement, Web caches can accelerate the performance of a content provider, improve content availability, reduce network traffic and download latency, and improve the overall user experience of the Web. In this article we have presented the most important design and implementation issues concerning Web caches, as well as their solutions. Some issues, such as dynamic document handling, remain open and new ideas are still being proposed. In particular, the issues of caching dynamic documents (Rhea, Liang & Brewer, 2003; Yuan, Chen & Zhang, 2003; Yuan, Hua & Zhang, 2003) and ensuring cache consistency (Mikhailov & Wills, 2003; Pandey, Ramamritham & Chakrabarti, 2003) in the presence of changing, non-static Web pages, are expected to attract considerable scientific interest in the near future. Regardless of the final decisions to prevail though, it is most likely that in the following years, caching will keep being a key solution for the performance of the World Wide Web.

REFERENCES

Abrams, M., Standbridge, C.R., Abdula, Williams, S., & Fox, E.A. (1995, December). *Caching proxies: Limitations and potentials.* WWW-4, Boston Conference.

Arlitt, M.F., & Williamson, C.L. (1996, May). Web server workload characterization: The search for invariants. *Proceedings of the ACM SIGMETRICS* (pp. 126-137).

Arlitt, M.F., Cherkasova, L., Dilley, J., Friedrich, R., & Jin, T. (1999, May). Evaluating content management techniques for Web proxy caches. *Proceedings of the Second Workshop on Internet Server Performance (WISP '99).*

Baentsch, M., Baum, L., Molter, G., Rothkugel, S., & Sturm, P. (1997, June). World-Wide Web caching – the application level view of the Internet. *IEEE Communications Magazine, 35*(6).

Caceres, R., Douglis, F., Feldmann, A., Glass, G., & Rabinovich, M. (1998, June). Web proxy caching: The devil is in the details. *ACM SIGMETRICS Workshop on Internet Server Performance.*

Cao, P., & Irani, S. (1997, December). Cost-aware WWW proxy caching algorithms. *Proceedings of the 1997 USENIX Symposium on Internet Technologies and Systems (USITS99)* (pp. 193-206).

Cao, P., & Liu, C. (1998). Maintaining strong cache consistency in the World Wide Web. *Proceedings of 3rd International Conference on Web Caching.*

Cao, P., Zhang, J., & Beach, K. (1998, September). Active cache: Caching dynamic contents (objects) on the Web. *Proceedings of the IFIP International Conference on Distributed Systems Platforms and Open Distributed Processing (Middleware '98).*

Challenger, J., Iyengar, A., & Dantzig, P. (1999, March). A scalable system for consistently caching dynamic Web data. *Proceedings of the IEEE Infocom '99 Conference.*

Chankhunthod, A., Danzig, P., Neerdaels, C., Schwartz, M., & Worrell, K. (1996). A hierarchical Internet object cache. *Proceedings of the USENIX Technical Conference.*

Cunha, C., Bestavros, A., & Crovella, M. (1995, April). *Characteristics of WWW client-based traces.* Technical Report 95-010. Boston University.

Douglis, F., Haro, A., & Rabinovich, M. (1997, December). HPP: HTML macropre -processing to support dynamic document caching. *Proceedings of the 1st USENIX Symposium on Internet Technologies and Systems (USITS97).*

Fan, L., Cao, P., Almeida, J., & Broder, A. (2000). Summary cache: A scalable wide-area Web cache sharing protocol. *IEEE/ACM Transactions on Networking, 8*(3), 281-293.

Fielding, R., Gettys, J., Mogul, J., Frystyk, H., & Berners-Lee, T. (1997, January). *Hypertext transfer protocol – HTTP/1.1. RFC2068.*

Gabber, E., & Shriver, E. (2000, September.). Let's put NetApp and CacheFlow out of business. *Proceedings of the SIGOPS European Workshop.*

Gwertzman, J., & Seltzer, M. (1996, January.). World-Wide Web cache consistency. *Proceedings of 1996 USENIX Technical Conference* (pp. 141-151).

Heidemann, J., Obraczka, K., & Touch, J. (1997). Modeling the performance of HTTP over several transport protocols. *IEEE/ACM Transactions on Networking, 5*(5), 616-630.

Maltzahn, C., Richardson, K.J., & Grunwald, D. (1999, June). Reducing the disk I/O of Web proxy server caches. *Proceedings of the 1999 USENIX Annual Technical Conference.*

Markatos, E.P. (1996, May). Main memory caching of Web documents. *Fifth International WWW Conference.*

Markatos, E.P. (2001, February). On caching search engine query results. *Computer Communications, 24*(2).

Markatos, E.P, Katevenis, M.G., Pnevmatikatos, D., & Flouris, M. (1999). Secondary storage management for Web proxies. *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems (USITS99).*

Meira, W., Cesario, M., Fonseca, R., & Ziv, N. (1999). Integrating WWW caches and search engines. *Proceedings of the IEEE 1999 Global Telecommunications Internet Mini-Conference.*

Melve, I. (1999). *Inter-cache communication protocols.* IETF WREC Working group draft.

Michel, S., Nguyen, K., Rosenstein, A., Zhang, L., Floyd, S., & Jacobson, V. (1998). Adaptive Web caching: Towards a new global caching architecture. *Computer Networks & ISDN Systems, 30*(22-23), 2169-2177.

Mikhailov, M., & Wills, C. (2003, May). Evaluating a new approach to strong Web cache consistency with snapshots of collected content. *12th Int'l World Wide Web Conference (WWW 2003),* Hungary.

Mogul, J.C. (1995, May). *The case for persistent-connection HTTP.* Western Research Laboratory. 95.4-Research Report 95/4.

Mogul, J.C. (1999). Speedier squid: A case study of an Internet server performance problem. *Login: The USENIX Association Magazine, 24*(1), 50-58.

Mogul, J.C., Douglis, F., Feldmann, A., & Krishnamurthy, B. (1997). Potential benefits of delta-encoding and data compression for HTTP. *Proceedings of the ACM SIGCOMM 97.*

Mohapatra, P., & Chen, H. (2002) WebGraph: A framework for managing and improving performance of dynamic Web content. *Special Issue of Proxy Servers in the IEEE Journal of Selected Areas in Communications.*

Nielsen, H.F., Gettys, J., Baird-Smith, A., Prud'hommeaux, E., Lie, H.W., & Lilley, C. (1997, September). Network performance effects of HTTP/1.1, CSS1 and PNG. *Proceedings of ACM SIGCOMM '97.*

Pandey, S., Ramamritham, K., & Chakrabarti, S. (2003, May). Monitoring the dynamic Web to respond to continuous queries. *12th Int'l World Wide Web Conference (WWW 2003),* Hungary.

Rhea, S., Liang, K., & Brewer, E. (200, May). Value-based Web caching. *12th Int'l World Wide Web Conference (WWW 2003),* Hungary.

Rizzo, L., & Vicisano, L. (2000). Replacement policies for a proxy cache. *IEEE/ACM Transactions on Networking, 8*(2), 158-170.

Rousskov, A., & Soloviev, V. (1998, June). On performance of caching proxies. *Proceedings of the 1998 ACM SIGMETRICS Conference.*

Rousskov, A., & Wessels, D. (1998). Cache digests. *Computer Networks & ISDN Systems, 30*(22-23), 2155-2168.

Savant, A., & Suel, T. (2003, September). Server-friendly delta compression for efficient Web access. *Proceedings of the 8th International Workshop on Web Content Caching and Distribution.*

Shi, W., Collins, E., & Karamcheti, V. (2003, September). Modeling object characteristics of dynamic Web content. *Journal of Parallel and Distributed Computing (JPDC), special issue on Scalable Internet Services and Architecture.*

Shriver, E., Gabber, E., Huang, L., & Stein, C. (2001, June). Storage management for Web proxies. *Proceedings of the 2001 USENIX Annual Technical Conference.*

SQUID proxy. *http://www.squid-cache.org*

Tatarinov, I., Rousskov, A., & Soloviev, V. (1997, September.). Static caching in Web servers. *Proceedings of the 6th IEEE Conference on Computer Communications and Networks.*

Tomlinson, G., Major, D., & Lee, R. (1999). High-capacity Internet middleware: Internet caching system architectural overview. *Second Workshop on Internet Server Performance.*

Wolman, A., Voelker, G., Sharma, N., Cardwell, N., Brown, M., Landray, T., Pinnel, D., Karlin, A., & Levy, H. (1999). Organization-based analysis of Web-object sharing and caching. *Proceedings of the 2nd USENIX Symposium on Internet Technologies and Systems (USITS99).*

Wooster, R., & Abrams, M. (1997, April). Proxy caching that estimates page load delays. *6th International World Wide Web Conference.*

Yuan, C., Chen, Y., & Zhang, Z. (2003, May). Evaluation of edge caching/offloading for dynamic content delivery. *12th Int'l World Wide Web Conference (WWW 2003),* Hungary.

Yuan, C., Hua, Z., & Zhang, Z. (2003, September). Proxy+: Simple proxy augmentation for dynamic content processing. *8th International Web Caching Workshop and Content Delivery Workshop (WCW'03).*

KEY TERMS

Byte Hit Rate: The ratio of bytes served by the cache over the total number of bytes requested by the clients. BHR can be significantly different from HR in a case where only few, but large files are being served by the cache.

Hit Rate: The ratio of requests served by the cache (hits) over the total number of requests made by the clients.

HTTP: Hyper-Text Transfer Protocol. The protocol used for most of the communication on the Web, between the clients, the proxies, and the servers.

Layer 4 Switch: A switch that can retrieve from the network packets information about the port number they are using, and thus the application that generated them.

Proxy Cache: A machine dedicated to serving client requests for Web documents. Usually it is installed at the edge of local or wide area networks to provide fast responses to the users and reduce the traffic exiting the network.

Server Accelerator: A machine dedicated to improving the performance of a Web server by caching popular Web objects. Usually it is installed in front of a single or a farm of Web servers in order to improve their performance and load the balance among them.

Transparent Proxy: A proxy that closely cooperates with either a router or a Layer 4 switch, to intercept Web requests while invisible to the users. Other than being invisible it works as a regular proxy cache.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 3048-3053, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Web Portal Research Issues

Arthur Tatnall
Victoria University, Australia

INTRODUCTION

In general terms, a portal can be seen as "a door, gate or entrance" (Macquarie Library, 1981), and in its simplest form the word just means a gateway; however, it is often a gateway to somewhere other than just to the next room or street. *The Oxford Reference Dictionary* defines a portal as "a doorway or gate etc, especially a large and elaborate one" (Pearsall & Trumble, 1996). In the context of this article, a Web portal is considered to be a special Internet (or intranet) site designed to act as a gateway to give access to other specific sites.

A Web portal can be said to aggregate information from multiple sources and make this information available to various users (Tatnall, 2005c). It consists of a Web site that can be used to find and gain access to other sites, but also to provide the services of a guide that can help to protect the user from the chaos of the Internet and direct him or her toward a specific goal. More generally, however, a portal should be seen as providing a gateway not just to sites on the Web, but to all network-accessible resources, whether involving intranets, extranets, or the Internet. In other words, a portal offers centralised access to all relevant content and applications.

BACKGROUND

The Web-portal concept developed from search-engine sites such as Yahoo, Excite, and Lycos, which offered access to a large amount of general information and acted as general jumping-off points to the contents of large parts of the Web. These general portals then began offering extra services in addition to search capabilities (Rao, 2001) as the first step in their evolution. In an attempt to describe the early stages of this evolution of the general portal, Eckerson (1999) outlines four generations of portals whose focus, in each case, was on the generic, personalised, application, and role (Tatnall & Davey, 2007).

An early classification of portals had them being either horizontal or vertical (Lynch, 1998). The original portal sites mentioned above would have been considered horizontal portals because they were used by a broad base of users, whereas vertical portals were focused toward a particular audience. Davison, Burgess, and Tatnall (2004) offer the following list of portal types.

- General portals provide links to all sorts of different sites of the user's choosing. Many of these general portals have developed from being simple search tools (such as Yahoo), Internet service providers (such as AOL), and e-mail services (like Hotmail).
- Vertical industry portals are usually based around specific industries. They aggregate information relevant to particular groups or online trade communities of closely related industries to facilitate the exchange of goods and services in a particular market as part of a value chain. They often specialise in business commodities and materials.
- Horizontal industry portals are portals utilised by a broad base of users across a horizontal market. Horizontal industry portals are typically based around a group of industries or a local area.
- Community portals are often set up by community groups, or based around special group interests. They attempt to foster the concept of a virtual community where all users share a common location or interest, and provide many different services depending on their orientation.
- Enterprise information (or corporate) portal is the term being applied to the gateways to corporate intranets that are used to manage the knowledge within an organisation. These are designed primarily for business-to-employee processes and offer employees the means to access and share data and information within the enterprise.
- E-marketplace portals often offer access to a company's extranet services and are useful for business-to-business processes such as ordering, tendering, and supplying goods.
- Personal or mobile portals are increasingly being embedded into mobile phones, wireless PDAs (personal digital assistants), and the like. Some appliances are also being equipped with personal portals aimed at allowing them to communicate with other appliances, or to be used more easily from a distance.
- Information portals can be classified into one of the other categories; however, they can also be viewed as a category in their own right as portals whose prime aim is to provide a specific type of information.
- Specialised or niche portals are designed to satisfy specific niche markets. In many cases, these can also be classified as information portals.

A major problem with any classification, however, is that new types and categories of portals are appearing all the time, portal types are reclassified, and most classification schemes include overlapping categories. Even given the difficulty in classifying portals or attempting to count the numbers of each type, it has become clear that specific rather than general portals are very much the topic of research interest around the world (Tatnall, 2005a, 2005b).

RESEARCH IN A WIDE RANGE OF WEB PORTAL APPLICATIONS

Web portals are now quite ubiquitous, and researching their use in organisations and by individuals is an important aspect of information systems research (Tatnall, 2007a). To illustrate the range of quite specific applications now being filled by portals, the following list of topic categories (Tatnall, 2005b) comes from articles by many academics from around the world who contributed to the *Encyclopaedia of Portal Technology and Applications* (Tatnall, 2007b). Anyone with an interest in conducting a research project needs to see what else has been done in this area, and this list of topics is provided to give such an indication. It does not show the level of usage of the different types of portal, but should be seen as much more than just a list. Rather, it provides an indication of the current research interest in portals and portal technology, and gives an idea of possible future research directions in this area.

Education Portals

To begin, there is considerable interest in education portals, and the topics covered include academic management portals; large-scale, integrated academic portals; mobile education portals; artificial intelligence and education portals; high school portals; primary school portals; corporate e-learning portals; Weblogs; knowledge portals in education; and subject-teaching portals.

Health and Medical Portals

This is another popular area with topics such as empowerment and health portals, bioinformatics portals, biotechnology portals, nursing knowledge portals, network-centric health care and the entry point into the network, and genomic and epidemiologic medical data portals.

Community, Personal, and Mobile Portals

Topics researched in this area included how to promote community portals, a community geographic domain names portal, designing a portal and community-to-community generator, local community Web portals and small businesses, the paradox of social portals, accessible personalised portals, mobile portal technologies and business models, mobile portals as innovations, mobile portals for knowledge management, the MP3 player as a mobile digital music-collection portal, widgets as personalised mini portals, wireless local communities in mobile commerce, and portals supporting a mobile learning environment.

Government and National Portals

In a related area, there was much research interest in government portals around the world: portals in the public sector, e-government portals, e-value creation in a government Web portal in South Africa, government portals as a gateway for enhancing electronic governance in Singapore, interoperability in integrating e-government portals, modeling public administration portals, service quality in the case of e-government portals, and state portals as a framework to standardise e-government services.

There is also research into portals relating to national issues: African Web portals, business module differentiation and a study of the top three Chinese portals, cross-cultural dimensions of national Web portals, the growth of e-portals in Dubai, how portals help Chinese enterprises operate successfully in global markets, impacts and revenues models from Brazilian portals, Web museums, and a case study of the French population.

Knowledge Management, Libraries, and Professional Societies

Knowledge management, especially relating to libraries and professional societies, is another area that attracts a number of researchers. They are interested in topics such as designing portals for knowledge work, mobile portals for knowledge management, knowledge servers, the portal as information broker, portal strategy for managing organisational knowledge, a prototype portal for use as a knowledge management tool to identify knowledge assets in an organisation, library portals and an evolving information legacy, open access to scholarly publications and Web portals, the IFIP portal, and portal features of major digital libraries.

Portal Concepts, Design, and Technology

As one might expect, portal concepts are also an area of particular interest with topics such as defining the portal, benefits and limitations of portals, comparing portals and Web pages, the evolution of portals, factors affecting the adoption of portals using activity theory, information visualisation, the ubiquitous portal, and portals of the mind.

Research on portal design and technology also features prominently with topics such as collaborative real-time information services via portals, digital interactive channel systems and portals, designing spatiotemporal portals to continuously changing network nodes, dynamic taxonomies and intelligent user-centric access to complex portal information, factors affecting portal design, developing semantic portals, an evolutionary approach to developing online learning portals in low-bandwidth communities, the role of ontologies in portal design, the evaluation of Web portals, portal quality issues, Java portals and the Java portlet specification API, the large-scale ASP replication of database-driven portals, WSRP specification and alignment with the JSR 168 portlet specification, and user-centric knowledge representation for the personalisation of Web portals.

Portal Uses and Applications

The largest area of research interest is in how portals are applied and used, and most of this research refers to quite specific applications, such as the Bizewest portal, the Bluegem portal, the European quality observatory portal, the future of portals in e-science, hosting portals on an e-marketplace, how corporate portals support innovation, how the Internet is modifying the news industry, industry portals for small businesses, portals for business intelligence, strategic planning portals, study of a wine industry portal, supplier portals in the automotive industry, supply chain management and portal technology, portal economics and business models, portals for integrated competence management, cultivating memories through the Beijing Olympics (2008) "advertainment" portal, portals for workflow and business process management, economical aspects when deploying enterprise portals, project management Web portals, the provision of product support through enterprise portals, e-management portals and organisational behaviour, employee self-service portals, a generic model of an enterprise portal, portal technologies and executive information systems implementation, user acceptance affecting the adoption of enterprise portals, the role of portals in consumer search behaviour and product customisation, guided product selection and comparison of e-commerce portals, enabling technology and functionalities of shopping portals, business challenges of online banking portals, Web museums as the last endeavour, Web portals as an exemplar for tourist destinations, and a Web portal for the remote monitoring of nuclear power plants.

APPROACHES TO RESEARCHING THE ADOPTION OF PORTALS

Before any technology can be used, however, it must first be adopted. Unfortunately, the adoption of any technology is not a straightforward process as it involves many factors, both human and technological.

The adoption of a portal, by an organisation or an individual, is particularly complex as by its very nature such an investigation must involve both humans and technology and so be treated as a sociotechnical study. I would argue that innovation translation, informed by actor-network theory (ANT), has many advantages as an explanatory framework over both the more commonly applied approaches of innovation diffusion and the technology acceptance model (TAM) in sociotechnical studies like this.

Both innovation diffusion (Rogers, 1995) and TAM (Davis, 1986) suggest that adoption decisions are made primarily on the basis of perceptions of the characteristics of the technology concerned (Tatnall & Davey, 2007). Using an innovation-diffusion approach, a researcher would probably begin by looking for characteristics of the specific portal technology to be adopted, and the advantages and problems associated with its use. He or she would think in terms of the advantages offered by portals in offering a user the possibility of finding information, but would do so in a fairly mechanistic way that does not allow for an individual to adopt the portal in a way other than that intended by its proponent: It does not really allow for any form of translation. If using TAM, this researcher would similarly have looked at characteristics of the technology to see whether the potential user might perceive it to be useful or easy to use.

One of the problems of using approaches based on innovation diffusion or TAM is that of essentialism: the idea that the technology has some characteristics that determine whether and how it will be adopted and used. This means that fixed and unproblematic properties or essences can then be assigned to the technology and used in any explanation of change. Whilst initially appealing, the problem with an approach like this is that it can lead to a simplistic explanation that fails to reveal the true complexities of the situation under investigation (Tatnall, 2003).

A researcher using an innovation-translation approach (informed by ANT) to study innovation, on the other hand, would concentrate on issues of network formation, investigating the human and nonhuman actors, and the alliances and networks they build up. The researcher would attempt to identify the actors and then follow them (Latour, 1996) in identifying their involvement with the innovation and how they affect the involvement of others. The researcher would then investigate how the strength of these alliances may have enticed the individual or organisation to adopt the portal or, on the other hand, to have deterred them from doing so (Tatnall, 2002; Tatnall & Burgess, 2006; Tatnall & Gilding, 1999). This approach allows the innovation to be translated into a different form by each potential adopter rather than just simply being accepted or rejected. Especially in investigations involving the interaction of humans and nonhumans (technical artefacts), such an approach has much value.

FUTURE TRENDS

The demise of the portal has been predicted for a long time (Tatnall, 2006). White (2003) noted some years ago that portals were undergoing a metamorphosis in which they were merging with technologies such as content management, collaboration, and business intelligence. Predictions that the portal would disappear into application servers have also proved untrue (Plumtree Software, 2003).

The first Web portals were designed by companies like Yahoo, Excite, and Lycos to act as general jumping-off points to the contents of large parts of the Web. The goal of these general portals was to become the user's home page, from which the user could select from a wide variety of categories of Web sites and return to the portal when they had finished looking at each site. Useful research has been done on portals of this type (Burgess & Tatnall, 2007; Sieber & Valor-Sabatier, 2002, 2005). Current research, however, is much more into specialised portals designed to provide a gateway to Web content in a specific area, and in enterprise portals that act as gateways to a corporate intranet (Tatnall, 2005b).

The marketing value of portals is considerable, and Schneider and Perry (2001) note that Web managers have discovered that increased sales and advertising income can result from the portal's ability to attract more people and retain them longer. They point out that Web portal companies have added "sticky" features like chat rooms, e-mail, and calendar functions in order to retain visitors longer at their sites. The success of the portal industry is closely linked to marketing, but Michael (2005) points out that advertisers and marketers have yet to understand the full potential of the Internet. Michael argues that a key function of marketing is to match buyers and sellers and to facilitate transactions, but to do this a proper institutional infrastructure is required. He points out that marketers need to be aware of new demographic segments that are being attracted to the Internet for searching and shopping purposes, and that portals should be regarded as strategic tools in the marketing process (Tatnall, 2006).

To give a rough idea of the growing need for researching the Web portal, a Google search of the Web in December 2003 showed 36 million entries. In late 2007, this had increased to 1,500 million entries relating to portals. Now, of course, not all these entries are concerned with research, and some will be duplicates of other entries, but these figures do provide a crude measure of the increasing importance of this topic. Future research of all types is required, both into portal applications and technology.

CONCLUSION

Research interest in the applications and technology of Web portals is high. The portal has now become quite ubiquitous, and researching its applications and design has become an important aspect of information systems and computer science research. While research into portal technology and design follows scientific principals, research into the applications and uses of portals requires a sociotechnical perspective. There is a considerable amount of research into the impact and use of Web portals, but not much into explaining the uptake of this technology. I have argued that for research of this type, it is useful to consider the portal in terms of technological innovation and to make use of an approach based on innovation theory. Actor-network theory provides a perspective that can resolve the dilemma of how to handle both the human and nonhuman contributions to technological innovation, and provides a useful explanatory system for doing so.

REFERENCES

Burgess, S., & Tatnall, A. (2007). A business-revenue model for horizontal portals. *Business Process Management Journal*, *13*(5), 662-676.

Davis, F. D. (1986). *A technology acceptance model for empirically testing new end-user information systems: Theory and results.* Boston: MIT.

Davison, A., Burgess, S., & Tatnall, A. (2004). *Internet technologies and business* (2nd ed.). Melbourne, Australia: Data Publishing.

Eckerson, W. (1999). *Plumtree blossoms: New version fulfils enterprise portal requirements.* Retrieved March 2003 from http://www.e-global.es/017/017_eckerson_plumtree.pdf

Latour, B. (1996). *Aramis or the love of technology.* Cambridge, MA: Harvard University Press.

Lynch, J. (1998, November 13). Web portals. *PC Magazine*.

Macquarie Library. (1981). *The Macquarie dictionary.* Sydney, Australia: Author.

Michael, I. (2005). Portals: Gateways for marketing. In A. Tatnall (Ed.), *Web portals: The new gateways to Internet information and services* (pp. 61-73). Hershey, PA: Idea Group Publishing.

Pearsall, J., & Trumble, B. (Eds.). (1996). *The Oxford English reference dictionary* (2nd ed.). Oxford, United Kingdom: Oxford University Press.

Plumtree Software. (2003). *The corporate portal market in 2003: Empty portals, the enterprise Web, composite applications.* San Francisco: Author.

Rao, S. S. (2001). Portal proliferation: An Indian scenario. *New Library World, 102*(9), 325-331.

Rogers, E. M. (1995). *Diffusion of innovations* (4th ed.). New York: The Free Press.

Schneider, G. P., & Perry, J. T. (2001). *Electronic commerce* (2nd ed.). Boston: Course Technology.

Sieber, S., & Valor-Sabatier, J. (2002). *Horizontal portal strategies: Winners, losers and survivors.* Paper presented at the 15th Bled Electronic Commerce Conference: eReality. Constructing the eEconomy, Bled, Slovenia.

Sieber, S., & Valor-Sabatier, J. (2005). Competitive dynamics of general portals. In A. Tatnall (Ed.), *Web portals: The new gateways to Internet information and services* (pp. 64-79). Hershey, PA: Idea Group.

Tatnall, A. (2002). Modelling technological change in small business: Two approaches to theorising innovation. In S. Burgess (Ed.), *Managing information technology in small business: Challenges and solutions* (pp. 83-97). Hershey, PA: Idea Group Publishing.

Tatnall, A. (2003). Actor-network theory as a socio-technical approach to information systems research. In S. Clarke, E. Coakes, M. G. Hunter, & A. Wenn (Eds.), *Socio-technical and human cognition elements of information systems* (pp. 266-283). Hershey, PA: Information Science Publishing.

Tatnall, A. (2005a). Portals, portals everywhere... In A. Tatnall (Ed.), *Web portals: The new gateways to Internet information and services* (pp. 1-14). Hershey, PA: Idea Group Publishing.

Tatnall, A. (2005b, November). *Web portals: From the general to the specific.* Paper presented at the Sixth International Working for E-Business (We-B) Conference, Melbourne, Australia.

Tatnall, A. (Ed.). (2005c). *Web portals: The new gateways to Internet information and services.* Hershey, PA: Idea Group Publishing.

Tatnall, A. (2006). Web portal gateways. In M. Khosrow-Pour (Ed.), *Encyclopedia of e-commerce, e-government and mobile commerce* (pp. 1217-1221). Hershey, PA: Idea Group Publishing.

Tatnall, A. (2007a). The ubiquitous portal. In A. Tatnall (Ed.), *Encyclopaedia of portal technology and applications* (Vol. 2, pp. 1040-1044). Hershey, PA: Information Science Reference.

Tatnall, A. (Ed.). (2007b). *Encyclopaedia of portal technology and applications.* Hershey, PA: Information Science Reference.

Tatnall, A., & Burgess, S. (2006). Innovation translation and e-commerce in SMEs. In M. Khosrow-Pour (Ed.), *Encyclopedia of e-commerce, e-government and mobile commerce* (pp. 631-635). Hershey, PA: Idea Group Reference.

Tatnall, A., & Davey, B. (2007, May 19-23). *Researching the portal.* Paper presented at IRMA: Managing Worldwide Operations and Communications with Information Technology, Vancouver, Canada.

Tatnall, A., & Gilding, A. (1999). *Actor-network theory and information systems research.* Paper presented at the 10th Australasian Conference on Information Systems (ACIS), Wellington, New Zealand.

White, C. (2003, July). Is the portal dead? *DM Review*.

KEY TERMS

Community Portals: Often set up by community groups or based around special group interests, they attempt to foster the concept of a virtual community where all users share a common location or interest, and provide many different services.

E-Marketplace Portals: They are extended enterprise portals that offer access to a company's extranet services.

Enterprise Information Portals: These are the gateways to corporate intranets that are used to manage knowledge within an organisation. They are designed primarily for business-to-employee processes and offer employees the means to access and share data and information within the enterprise.

General (or Mega) Portals: General portals provide links to all sorts of different sites of the user's choosing, often from a menu of options.

Horizontal Industry Portals: They are portals utilised by a broad base of users across a horizontal market.

Information Portals: Information portals can also be viewed as a category in their own right as portals whose prime aim is to provide a specific type of information.

Personal/Mobile Portals: These portals are embedded into mobile phones, wireless PDAs, appliances, and the like.

Specialised/Niche Portals: They are portals designed to satisfy specific niche markets. Sometimes they provide detailed industry information, often available only for a fee.

Vertical Industry Portals: Usually based around specific industries, they aim to aggregate information relevant

to these groups of closely related industries to facilitate the exchange of goods and services in a particular market as part of a value chain.

Web Portal: It is a special Internet (or intranet) site designed primarily as a gateway to provide access to other sites.

Web Services Coordination for Business Transactions

Honglei Zhang
Cleveland State University, USA

Wenbing Zhao
Cleveland State University, USA

INTRODUCTION

Many e-commerce companies such as Amazon.com, Yahoo.com, and eBay.com started to offer Web services to their partners and customers. Through such Web services, new value-added services could be provided and hence higher revenues would be generated. Essentially, the Web services technology is transforming the World Wide Web from a predominantly publishing platform to a programmable platform, which undoubtedly will make it easier to conduct business online, and enable automated business-to-business communications (Papazoglou, 2003). The Web services technology is particularly useful for Application Service Providers that offer various on-demand services and software-as-a-service (SAAS) to their customers (Chakrabarty, 2007). Such service-oriented computing is attractive to many businesses because they can save valuable resources and money by avoiding installing and maintaining sophisticated enterprise software on-site. Furthermore, most of business interactions are transactional, which require well-defined coordination support. To meet this requirement, a number of specifications have been proposed, and OASIS has recently rectified the Web Services Transactions specifications (Feingold & Jeyaraman, 2007; Freund & Little, 2007; Little & Wilkinson, 2007).

In this chapter, we provide an overview of the Web services technology, together with the set of standard specifications for Web services transactions. The core components of the Web services technology include eXtensible Markup Language (XML), HyperText Transfer Protocol (HTTP), Simple Object Access Protocol (SOAP), and Web Services Description Language (WSDL). Both SOAP and WSDL are based on XML. All these protocols and languages have the characteristic of strong extensibility, which lays a solid foundation for the success of the Web services technology. Due to the extensibility design, the protocols specified in the Web services transactions standards can be plugged into the Web services core seamlessly to provide the additional coordination needed for business transactions. Furthermore, we point out the need to protect the business transactions from the hardware failures and malicious faults, and for more robust coordination for Web services transactions.

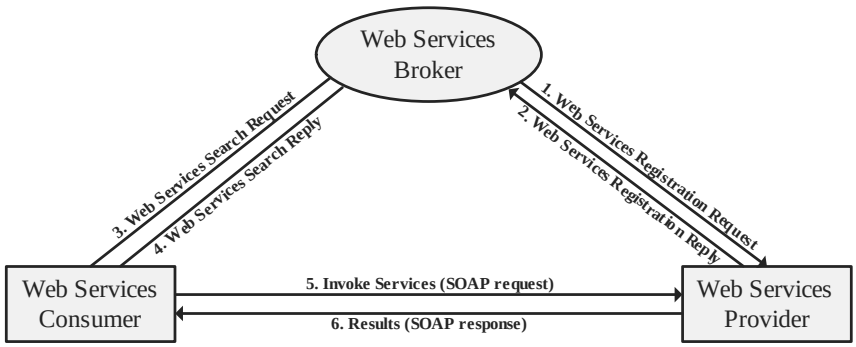
BACKGROUND

In this section, we introduce the Web services concept and the basic building blocks of the Web services platform. There is no universal definition of the term Web services and its interpretation varies drastically. Web services can be loosely defined as any services offered over the World Wide Web. On the other hand, only the services enabled by the Web services technology are referred to as Web services by many researchers and practitioners. In this chapter, we use the latter interpretation. The Web services technology refers to the set of standards that enable automated machine-to-machine interactions over the Web. The corner stone of the Web services technology consists of eXtensible Markup Language (XML) (Bray, Paoli, Sperberg-McQueen, Maler, Yergeau & Cowan, 2006), HyperText Transfer Protocol (HTTP), Simple Object Access Protocol (SOAP) (Gudgin, Hadley, Mendelsohn, Moreau, Nielsen, Karmarkar et al., 2007), Web Services Description Language (WSDL) (Christensen, Curbera, Meredith & Weerawarana, 2001), and the Universal Description, Discovery and Integration (UDDI) service (Clement, Hately, Riegen & Rogers, 2004). From an architecture point of view, the Web services platform consists of Web services providers, Web services consumers, and UDDI registries that broker the providers and the consumers, as shown in Figure 1. If a Web services consumer wants to request a service, it can search for available service providers via UDDI. Based on the returned information, the consumer can invoke the service by sending the request directly to the service provider.

EXTENSIBLE MARKUP LANGUAGE

XML is designed to facilitate self-contained, structured data representation and transfer over the Internet. It allows users to define their own tags, which is why it is easily extensible. XML messages enable different applications to communicate with each other over the network using a variety of transport-level protocols such as HTTP and SMTP. To invoke a Web service, a user only needs to send an XML request message to the Web services provider. The provider will then send

Figure 1. The architecture of Web service



back an XML reply message containing the results the user wanted. Typically, the XML messages must conform to the SOAP standard.

Simple Object Access Protocol

SOAP, a communication protocol for message exchanges over the Internet, provides a standard modular packaging model, a data encoding method and a way to perform remote procedure calls (RPCs). SOAP is easy to use and it can be easily extended due to its use of XML as the messaging format. Like many public-domain application-level protocols, such as SMTP, a SOAP message contains a SOAP Envelop and a SOAP Body. A SOAP message often contains an optional SOAP Header element and a Fault element if an error is encountered by the sender of the SOAP message.

Web Service Description Language

WSDL provides a structured way to describe a Web service based on an abstract model. For each Web service, the corresponding WSDL document specifies the available operations, the messages involved with the operations, and a set of endpoints to reach the Web service. Due to the use of XML, WSDL is also extensible. In particular, it allows the binding of multiple different communication protocols and message formats.

Universal Discovery Description and Integration

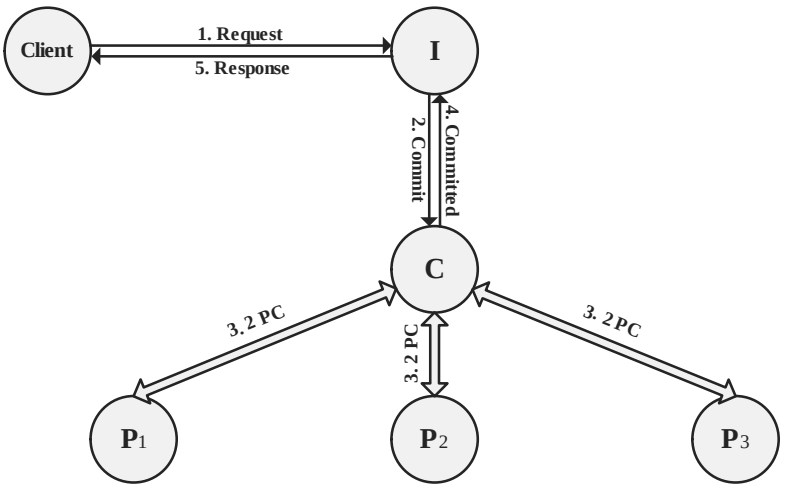
A UDDI registry service acts like yellow pages for business providers and consumers. Business owners publish their Web services to the UDDI registry, and their partners and consumers can locate the Web services they needed and obtain detailed information regarding the services by searching the registry. There are three main components in UDDI, often referred to as White Pages, Yellow Pages, and Green Pages. The White Pages provide the Web service provider's information, such as name, address, contact information and identifiers. The Yellow Pages describe industrial categories based on standard taxonomies. The Green Pages present technical information in detail regarding the Web services. The UDDI also support several ways to carry out the search, for example, one can search by service provider's location, or by specified service types.

Web Services Coordination of Business Transactions

Web services interactions are becoming more and more complex in structure and relationships. More complex means we need longer time to execute them, because of business latencies and user interactions. The Web Services Coordination specification (WS-Coordination) (Feingold & Jeyaraman, 2005) describes an extensible framework for plugging in protocols that coordinate the actions of distributed applications. Such coordination protocols can be used to support a variety of business applications, including those that require strict consistency and those that require agreement of a proper subset of the participants. The framework enables a Web service to create a context needed to propagate an activity to other Web services and to register for a particular coordination protocol.

There are two types of business transactions. One follows the traditional atomic transaction semantics, and the other is referred to as business activities, which implies that the atomicity property may be relaxed. The former is suitable for short transactions that require strong atomicity, such as a fund transfer transaction. The latter is more suitable long running transactions, such as those used in supply chain management. Based on WS-Coordination, two specifications, namely Web Service Atomic Transaction (WS-AtomicTransaction) (Little & Wilkinson, 2007) and Web Service Business Activity (WS-BusinessActivity) (Freund & Little, 2007), have been

Figure 2. The architecture of WS-AT



recently standardized by OASIS to address the coordination needs for common types of business transactions.

Web Services Atomic Transactions

The Web Services Atomic Transactions specification defines a coordination framework for Web services atomic transactions. In a distributed atomic transaction, all participants must reach the same final agreement as to whether the transaction has succeeded or not. This is ensured by the coordination mechanisms specified in WS-AtomicTransaction. In WS-AtomicTransaction, there are three actors for each transaction: Completion Initiator, Coordinator and Participants. Each provides a different set of services for the atomic transaction, and they interact with each other via two protocols, the Completion Protocol and the Two-Phase Commit Protocol (Gray & Reuter, 1993) (Tanenbaum & Steen, 2002). The architecture of WS-AtomicTransaction is shown in Figure 2.

The completion initiator is responsible to start and terminate a transaction. It also provides the Completion Initiator Service so that the coordinator can inform it the final outcome of the transaction, as part of the completion protocol. The coordinator provides the following services:

- Activation Service: At the beginning of a transaction, the initiator invokes the Activation Service for creating a coordinator object, which will generate a new coordination context for the transaction and return it to the initiator. The coordination context contains a unique transaction identifier and an endpoint reference for the Registration Service. This coordination context will be included in every request messages within the transaction boundary.
- Registration Service: The participants and the completion initiator use this service to register their endpoint references for other associated participant-side services. Later these endpoint references will be used by the coordinator to contact the participants during the two-phase commit.
- Coordinator Service: When a participant gets a Prepare request from the coordinator, it places its vote by invoking the coordinator service. The participants also use this service to notify the coordinator their acknowledgments to the commit/abort request. The participants obtain the endpoint reference of the Coordinator Service during the registration step.
- Completion Service: The initiator invokes this service to notify the coordinator to start a distributed commit. The Completion service, together with the Completion Initiator service on the participant side, implements the WS-AtomicTransaction completion protocol. The endpoint reference of the Completion Service is returned to the initiator during the registration step.

The participant provides the Participant Service, which allows the coordinator to solicit votes from, and to send the transaction outcome to the participants according to the two-phase commit protocol.

The Completion Protocol is used by the completion initiator to initiate the atomic termination of a transaction. When the initiator decides that it is time to commit the transaction, it sends a Commit request to the coordinator. The coordinator will then launch an instance of the Two-Phase Commit (2PC) protocol to carry out the coordination for atomic commitment of the transaction. When the 2PC completes, the coordinator notifies the initiator the outcome of the transaction (i.e., committed or aborted). If the request

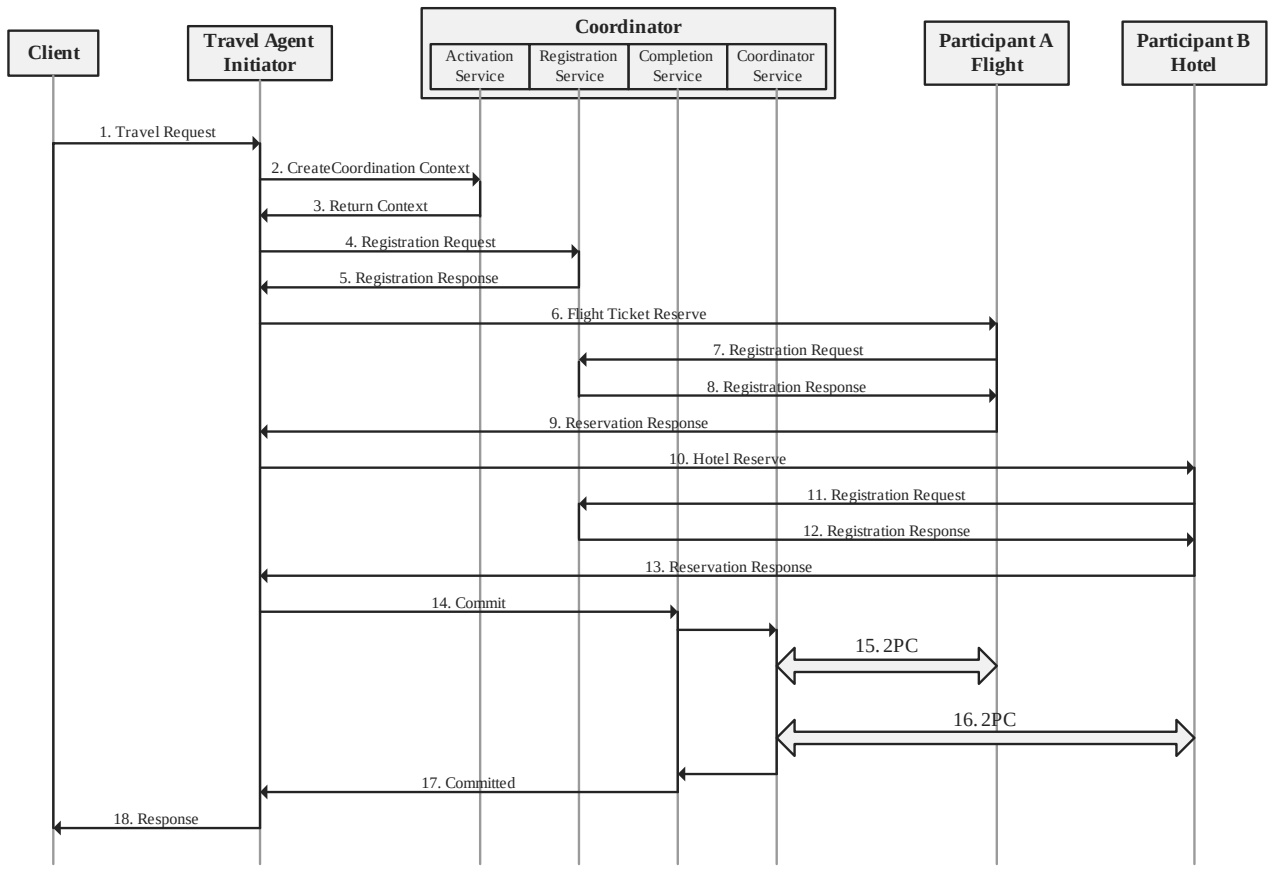
from the initiator is Rollback instead, the coordinator will abort the transaction directly.

The 2PC Protocol is used by the coordinator and participants to guarantee atomic commitment of a transaction, and it executes in two phases. During the first phase, that is, the prepare phase, the coordinator sends a Commit request to all registered participants soliciting their votes. When the coordinator receives votes from all participants, or a timeout has occurred, it starts the second phase, that is, the commit phase, to notify the participants the outcome of the transaction.

2PC has two variants used for different resources, Volatile 2PC and Durable 2PC. Volatile 2PC is used for volatile resources such as caches and Durable 2PC focuses on durable resources like a database. Participants must register in an appropriate protocol before the termination of the transaction. A participant can register in more than one protocol. Upon receiving a Commit request from the initiator in the completion protocol, the coordinator begins the prepare phase first for every participant who has registered in the Volatile 2PC protocol by sending a Prepare request to them before it begins the prepare phase for Durable 2PC. The participant that gets the request must respond with a Prepared, Aborted or ReadOnly message based on its own decision. During this period, other participants can continuously register with the coordinator for Durable 2PC, but the registration progress has to be done by the coordinator before the start of the first phase for durable resources. After the prepare phase for Volatile 2PC, the coordinator begins to take care of the prepare phase for Durable 2PC participants. Same as the prepare phase of Volatile 2PC, all participants in the Durable 2PC protocol have to respond appropriately before the protocol advances to the second phase, in which the coordinator will issue the Commit requests to all participants for both Volatile and Durable 2PC protocols if all participants have provided positive feedbacks. If there are any negative votes, even if only one, the coordinator has to abort the transaction. After the participants get a Commit notification, they will commit the transaction and send the Committed acknowledgement back to the coordinator.

Figure 2 shows an example of a Web services atomic transaction for a travel reservation coordinated by WS-AtomicTransaction. In this example, a client contacts a Travel Agent to make travel arrangement. The Agent, which

Figure 3. A travel reservation example using WS-AtomicTransaction



acts as the completion initiator, is responsible to make flight and hotel reservations in the context of an atomic distributed transaction, on behalf of the client. We assume that the Agent relies on a Flight reservation Web service and a Hotel reservation Web service, for booking a plane ticket and a hotel room for the traveler.

To begin this transaction, the client sends a request to the initiator (step 1). The initiator invokes the Activation Service on the Coordinator which will create a unique coordination context for the transaction (step 2). This context contains the Endpoint Reference for the Registration Service. After the Activation step, the initiator gets the reply back with the transaction context (step 3). In next step, the initiator registers the Completion Initiator Service with the coordinator so that the coordinator could inform the initiator after it obtains the outcome of the transaction (steps 4 and 5). Now the booking service offered by the initiator carries out the reservations for the plane ticket and a hotel room (steps 6 and 10). The flight and hotel booking Web services must register their participant endpoint references with the coordinator (steps 7 and 8; 11 and 12). After the registration step, these services will send their responses for the reservation requests (steps 9 and 13). Subsequently, the initiator asks the Completion Service to commit the transaction (steps 15 – 17). Finally, the Travel Agent sends the result back to the client (step 18).

Web Services Business Activity

WS-BusinessActivity is different from WS-AtomicTransaction because it focuses on the coordination of long running business activities where the atomic transaction model is not appropriate. WS-BusinessActivity is also built on top of the WS-Coordination framework. The WS-BusinessActivity specification describes two coordination types, Atomic-Outcome and Mixed-Outcome, and two coordination protocols, Business-Agreement-with-Participant-Completion and Business-Agreement-with-Coordinator-Completion. Either protocol can be used with either coordination type. If the Atomic-Outcome Coordination type is used, all participants must reach an agreement about the activity outcome (i.e., either to close or to compensate). If the Mixed-Outcome coordination type is used, some participants may be directed to close while others to compensate. All WS-BusinessActivity frameworks must implement the Atomic-Outcome coordination type.

In a WS-BusinessActivity framework, a participant registers either one of the two protocols, which are managed by the coordinator of the business activity. The only difference between the two protocols is that the Business-Agreement-with-Participant-Completion protocol assumes that the participants know when the coordinator has completed all the processing related to a business activity and the Business-Agreement-with-Coordinator-Completion protocol notifies the participants when the coordinator has received all requests.

A participant who has registered the Business-Agreement-with-Participant-Completion protocol informs its coordinator by sending a Completed notice when it has done its entire works for a business activity. The coordinator should reply with either Close or Compensate message depends on the circumstance. The participant receives a Close instruction if the activity has completed successfully. If it gets a Compensate instruction instead, the coordinator will undo the completed work and will have to restore the data recorded from the initial condition. The participant may encounter a problem or fail during the processing of the activity, in which case, it must signal the coordinator with an error message. If it gets the Fail message, the coordinator will acknowledge the participant by a Failed notification.

Upon receiving a CannotComplete notification, the coordinator learns that the participant cannot successfully finish its work. On sending out this message, the participant discards all its pending work and cancels all related executions, and exits the current business activity. On receiving such a message, the coordinator is required to notify the participant with a NotCompleted message. In the active state, the coordinator could cancel any transaction by using the Cancel notification, and the participant will respond with a Canceled message if it receives the message.

In the Business-Agreement-with-Coordinator-Completion protocol, the completion decision comes from the coordinator. The coordinator sends a Complete message to the participants informing them that they won't receive any new requests within the current business activity and it is time to complete the processing. The participant then replies with a Completed message if it could successfully finish its work.

FUTURE TRENDS

We identify two future research directions regarding Web services coordination for business transactions. First, the coordinator plays a critical role in both the WS-AtomicTransaction and WS-BusinessActivity frameworks. The integrity of the coordinator services directly impact the expected outcome among the participants involved in an atomic transaction or a business activity. For example, if the coordinator is compromised by an adversary, it could easily cause non-atomic commitment of transactions, which would cause serious trouble for transaction participants. For business activities, a compromised coordinator could also render an inconsistent outcome among the activity participants. Due to the untrusted Internet operating environment, such threat cannot be ignored, which calls for enhancement of the WS-AtomicTransaction and the WS-BusinessActiv-

ity frameworks by intrusion tolerance design (Zhao, 2007). The Byzantine fault tolerance technology (Castro & Liskov, 2002; Lamport, Shostak & Pease, 1982) might be a good fit to achieve the goal. Byzantine fault tolerance ensures that as long as the number of compromised coordinator replicas is below some threshold, the replicated coordinator cannot disseminate conflicting information (i.e., the transaction outcome) to different transaction participants, which is how it protects the coordinator's integrity. Byzantine fault tolerance also protects the system against hardware failures, which has the benefit of high availability.

Second, even though WS-BusinessActivity recognized the need for flexible transaction outcomes for business activities, its reliance on compensation transactions might limits its use for some applications. The reservation-based extended transaction protocol (Zhao, Moser & Melliar-Smith, 2005) seems to be an excellent candidate to augment the compensation-based approach. The basic idea of the reservation-based transaction protocol is that any business activity is carried in two steps. In the first step, a reservation is placed on a set of resources. Depending on the outcome of the reservation step, the coordinator could choose to confirm some reservations while cancel the remaining ones. The use of the extra reservation step eliminates the need for compensation transactions, which could be very expensive and error prone in practice.

CONCLUSION

In this chapter, we introduced the Web Service technology and the related technical standards such as XML, SOAP, WSDL and UDDI. We focused on the issues about the Web services business transaction coordination and presented three related Web services specifications, WS-Coordination, WS-AtomicTransaction and WS-BusinessActivity. We note that the Web services technology is still rapidly evolving. We have seen more standards being rectified and some old specifications are rendered obsolete. The Web services technology is also of great interest to the research community. In particular, there has been a tremendous effort to enhance the security and dependability of Web services.

REFERENCES

Bray, T., Paoli, J., Sperberg-McQueen, C. M., Maler, E., Yergeau, F., & Cowan, J. (Eds.). (2006). XML 1.1 (2nd Edition): World Wide Web Consortium.

Castro, M. & Liskov, B. (2002). Practical Byzantine fault tolerance and proactive recovery. ACM Transactions on Computer Systems, 20(4), 398–461.

Chakrabarty, S. (2007). Strategies for business process outsourcing: An analysis of alternatives, opportunities and risks. E-Business Process Management: Technologies and Solutions (1st ed.) (pp. 204-229). Hershey, PA: IGI Publishing.

Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001). Web services description language (WSDL) 1.1: World Wide Web Consortium.

Clement, L., Hately, A., Riegen, C., & Rogers, T. (Eds.). (2004). UDDI Version 3.0.2. OASIS Standard.

Feingold, M. & Jeyaraman, R. (Eds.) (2007). Web services coordination (WS-Coordination) Version 1.1: OASIS Standard.

Freund, T. & Little, M., (2007). Web services business activity (WS-BusinessActivity) Version 1.1: OASIS Standard.

Gray, J. & Reuter, A. (1993). Transaction processing: Concepts and techniques. Morgan Kaufmann Publishers.

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J.-J., Nielsen, H. F., Karmarkar, A., et al. (Eds.) (2007). SOAP Version 1.2: World Wide Web Consortium.

Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine generals problem. ACM Transactions on Programming Languages and Systems, 4(3), 382–401.

Little, M. & Wilkinson, A. (Eds.) (2007). Web services atomic transaction (WS-AtomicTransactionomicTransaction) Version 1.1: OASIS Standard.

Papazoglou, M. P. (2003). Web services and business transactions. World Wide Web: Internet and Web Information Systems, 6(1), 49-91.

Tanenbaum, A. S. & Steen, M. (2002). Distributed systems: Principles and paradigms. 393-398.

Zhao, W., Moser, L. E., & Melliar-Smith, P. M. (2005). A reservation-based coordination protocol for web services. In Proceedings of the IEEE International Conference on Web Services (pp. 49-56). Orlando, FL.

Zhao, W. (2007). A Byzantine fault tolerant coordination for web services atomic transactions. In Proceedings of the 3rd IEEE International Symposium on Dependable, Autonomic and Secure Computing (pp. 37-44). Columbia, MD.

KEY TERMS

Atomic Transaction: An atomic transaction in the context of Web services refers to a distributed transaction to be executed atomically. It should exhibit the atomicity,

consistency, isolation and durability properties, just like a local transaction.

Distributed System: A distributed system is a computer network system, shown to end users as a single machine but actually work with a set of independent computers connected.

Endpoints Reference (EPR): A collection of information used in web service to describe a resource's address, such as a URI (uniform resource identifier). This information is contained in the header of a SOAP message.

Remote Procedure Calls (RPC): A technical way that allows a call to a subroutine or a procedure which may run on another machine. To the caller, the call appears to be no different from a local call.

SOAP: Stands for Simple Object Access Protocol. It was originally designed to conduct remote procedure calls over the Web. It has evolved to become the main communication protocol to exchange XML documents. A SOAP message contains a SOAP Envelop and a SOAP Body. A SOAP message often contains an optional SOAP Header element, and a Fault element if an error is encountered by the sender of the SOAP message.

UDDI: Stands for Universal Description, Discovery and Integration. UDDI provides the standard way for Web services providers to describe their services, and the consumers to search and discover the available services.

Web Services: Electronic services enabled by the Web services technology. The Web services technology refers to the set of standards that enable automated machine-to-machine interactions over the Web. The core standards include XML, HTTP, SOAP, WSDL and UDDI.

WSDL: WSDL stands for Web Services Description Language. It is an XML-based language used to describe Web services. For each Web service, the corresponding WSDL document specifies the available operations, the messages involved with the operations, and a set of endpoints to reach the Web service. Due to the use of XML, WSDL is also extensible. In particular, it allows the binding of multiple different communication protocols and message formats.

XML: XML stands for eXtensible Markup Language. It is designed to facilitate self-contained, structured data representation and transfer over the Internet. It is extensible because it allows users to define their own tags.

Web Usability

Shirley Ann Becker
Florida Institute of Technology, USA

INTRODUCTION

The study of computing technology and user interfaces was initiated during the 1970s when industrial research laboratories began to focus on human-computer interaction (HCI) (Badre, 2002). In the 1980s, the personal computer was introduced, thus expanding the need for designing effective user interfaces. HCI became a discipline during this time, and the Association for Computing Machinery (ACM) established the Special Interest Group in Computer Human Interaction. One of the first textbooks on HCI, *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (Schneiderman, 19891), was published. Shortly thereafter, HCI became part of the ACM curriculum promoting the development of effective user interfaces. Software tools were developed in order to assist in designing usable interfaces while employing usability engineering methods. Many of these methods focused on usability from the perspective of ease of use, ease of learning, user satisfaction, and zero defects (Nielsen, 1993).

The World Wide Web (Web) became an integral part of HCI research in the 1990s, as organizations rushed to deploy a corporate Web site. Many of these Web sites took advantage of cutting-edge technology, including graphics and animation, with little regard for the impact on the user. As a result, users became disgruntled by lengthy download times, complex navigation schemes, nonintuitive search mechanisms, and disorganized content.

While others were predicting a "Y2K meltdown," Jakob Nielsen (1999a) correctly predicted a "Web meltdown" due to the number of poorly designed Web sites that cluttered the Internet. Numerous studies showed that users were frustrated with glitzy Web sites that had too many usability barriers. A Forrester report estimated a 50% loss of potential online sales due to users not finding a product or service on the Web site (Manning, McCarthy & Souza, 1998). As importantly, 40% of users did not return to a site when their initial visit was a negative one.

Shortly after 2000, electronic commerce sites (dot coms) began to fail at an increasing rate. A Deloitte and Touche report found that many retailers had developed online sites to "test the waters" for consumer demand with no clearly articulated strategy for success (Speigel, 2000). The demise of many dot coms has been attributed to unfriendly user interfaces that negatively impacted the online experience.

BACKGROUND

Many researchers and practitioners alike have studied usability in order to develop Web sites that are navigable, consistent, appealing, clear, simple, and forgiving of user mistakes (Murray & Costanza, 1999). Existing user interface design recommendations were extended to include user interfaces for the Web (Lynch & Horton, 1999; Schneiderman, 1998). Those experienced in designing user interfaces provided heuristics and guidelines for designing Web pages, often by identifying design layout, navigation, and performance issues associated with particular Web sites (Flanders & Willis, 1998; Hurst, 1999; Spool, Scanlon, Schroeder, Snyder & DeAngelo, 1999). Jakob Nielsen, a well-known usability expert, provided much needed guidance on Web usability through featured online articles (www.useit.com/alertbox) and published guidelines (Nielsen, 1999b; Nielsen & Tahir, 2002).

Web usability has been defined as the measure of the quality of the user's online experience. There are several factors that are commonly used as a means of measuring this experience. These factors include (www.usability.gov):

- Learnability – A measure of the user's learning time for accomplishing basic tasks given that the user interface has not previously been used (or used infrequently).
- Efficiency – A measure of the user's time and error rate for task completion.
- Effectiveness – A measure of user productivity in performing a task.
- Satisfaction – A measure of the attitude, perceptions, and feelings about the site.
- Memorability – A measure of user recall such that a previously visited site can be used effectively with no new learning curve.

It is commonly accepted that the usability of a Web site is impacted by the user's online goal, the user's profile, and his or her computing environment. A user, for example, would have some tolerance for lengthy download times when searching for medical information with graphic illustrations. This tolerance level is greatly reduced when searching for information on the cost of an airline ticket. The user profile, including age, gender, income, education, computer skills, and other factors, influences the online experience. Web

content written at a high reading grade level, for example, may be difficult to comprehend for users with low English proficiency. The use of color to convey meaning on a Web site may impede its use by those who have color-deficient sight. Small font size, patterned background images, and pastel colors may become Web barriers to older adults experiencing vision degradation due to aging (Morrell, 2002). The user's computing environment also has an impact on Web usability. Environmental factors, such as hardware, software, browsers, connectivity, and bandwidth, impede the use of a Web site when it is cluttered with graphics, animation, and other objects adding little value to the online experience.

Since 1998, much has been accomplished in promoting Web usability for persons with disabilities. Section 508 of the 1973 Rehabilitation Act was enacted to eliminate information technology barriers in order to provide those with disabilities equal access. The law applies to all federal agencies when they develop, procure, maintain, or use electronic and information technology (http://www.Section508.gov). As a result of this initiative, significant strides have been made to electronic government access by enforcing the Web content guidelines put forth by the World Wide Web Consortium. Though not mandated by law, many commercial and nonprofit Web sites have implemented Section 508 in order to provide access to a broad user base.

WEB USABILITY ASSESSMENT METHODS

There are several popular methods that have been employed to effectively study Web usability. The inquiry approach makes use of field observation, interviews, self-reporting logs and online sessions. The inspection approach utilizes heuristic evaluations, walkthroughs, and checklists. Usability testing may also be used in conjunction with the other approaches to gather feedback during and after Web site design (Hom, 1998).

- Field Observation – The user is observed while surfing a Web site in order to gather usability data in a real-world setting.
- Interviews, Surveys, and Questionnaires – The objective of these methods is typically to gather feedback about the user's perspective of usability. In terms of data gathering, the interview is a formal, structured process, whereas the survey is an informal, interactive process. Interviews and surveys may involve one or more users in a focus group setting. The questionnaire provides the means to obtain written responses regarding a user's online experience.
- Session and Self-Reporting Logs – The user records his or her actions and makes observations during an online session. Software is often used during a session to automatically record data about the user's online experience. The self-reporting log requires the user to manually record data while surfing the Web.
- Heuristic Evaluation – A usability expert (or group of experts) assesses a user interface to determine whether the Web design follows established usability practices (heuristics).
- Walkthrough – A usability expert (or group of experts) evaluates online experiences by constructing scenarios of Web use and then role-playing the targeted user.
- Usability Inspection – A usability expert (or group of experts) conducts usability inspections of a user interface in order to uncover usability problems in the design.
- Checklists – A usability expert (or group of experts) uses a checklist often in conjunction with an inspec-

Table 1. Web usability online resources

Resource	Description
http://www.usability.gov	National Cancer Institute summarizes research activities on Web usability. It also provides links to usability resources.
http://www.itl.nist.gov/iad/vvrg/index.html	National Institute of Standards and Technology provides resources and tools for usability testing.
http://www.useit.com	Jakob Nielsen and colleagues provide alert box articles, summaries of usability studies, and other usability resources.
http://www.acm.org/sigchi/	ACM Special Interest Group on Computer-Human Interaction provides a bibliography of usability research.
http://www.w3.org/WAI/	The World Wide Web consortium (W3C) Web initiative provides resources on making sites accessible to those with disabilities.
http://www.usabilitynews.org	The *Software Usability Research Laboratory (SURL)* specializes in software and Web site user interface design research, human-computer interaction research, and usability testing and research.

tion to ensure that established usability practices are evaluated.
- Usability Testing – Experiments are conducted regarding usability aspects of a Web design. The objective is to gather data about an online experience in order to draw conclusions about the usability of the site. Though usability testing can involve sophisticated technology including usability labs, videotaping, and eye-tracking, this does not have to be the case (Murray & Costanzo, 1999). Often, usability test cases can be generated from an existing Web design without the use of sophisticated technology.

WEB RESOURCES

There are valuable online resources promoting usable Web designs. Many of these sites offer links for usability design and testing, good practices and lessons learned in the field. Some of the more popular sources for Web usability guidance are listed in Table 1. Though too numerous to list, there are many educational sites that offer resources on research activities and course materials. The Trace Research and Development Center at the University of Wisconsin-Madision, in particular, offers usability resources and links promoting universal usability (*http://trace.wisc.edu/world/web/*).

FUTURE TRENDS

Murray and Costanzo (1999) point out that from a HCI perspective, Web development is significantly different from software development. As such, there are challenges facing developers who are pursuing usable Web designs. These challenges include the following:

- The demographic diversity of online customers makes it difficult to develop user-friendly interfaces to meet all needs. For example, older adult users (60 years and older) may have trouble seeing Web content based on the use of color, font size, font type, and patterned background images (Becker, 2004; Morrell, Dailey, Feldman, Mayhorn & Echt, 2002). These design elements may have no usability impact on a younger adult for whom aging vision changes have not yet occurred. Web site images or textual references to religious holidays (e.g., Valentine's Day), as another example, may be offensive in certain global regions due to local religious or cultural beliefs (Becker, 2002).
- There is significant diversity among hardware, software, and network components being used to surf the Web. The usability of mobile technology, for example, must take into account the tiny screen in which Web content is displayed (Russell & Chaparro, 2002).
- Slower network access speeds impact usability due to performance degradation for a Web page with graphics and animation. Usability is also impacted by the browser version being used to surf the Web. Web content may display differently in older browser versions of Netscape© and Internet Explorer© than newer versions.
- The internationalization of many Web sites must account for culture, religion, and language in designing localized, user-friendly interfaces. Too often, organizations develop localized versions that do not meet the needs of regional customers (Marcus & Gould, 2000). The localized site may still have design aspects of the country of origin such as: English content, clichés, acronyms, and abbreviations (Becker & Mottay, 2001). Graphics may become a usability issue when cultural and religious beliefs are not taken into account during Web design (e.g., scantily clad figure on a homepage).
- Unlike software, users do not have a vested interest in a particular site. Often times, a user has purchased software and therefore is willing to accept usability barriers associated with it. Since there is no personal investment in a Web site, a user is more likely to leave and not return to a Web site that is perceived as unusable.

CONCLUSION

Web usability remains an important consideration in the design of effective user interfaces. There has been significant research on Web usability in terms of design layout, performance, navigation, and searches, among other areas. This initial work has been broadened to include usable Web designs for all users regardless of age, skills, education, culture, language, or religion. The Web accessibility initiative has promoted Web designs that take into account users with vision, physical, and cognitive disabilities. The internationalization of Web sites has promoted Web designs that meet the needs of a particular locale taking into account the customs, culture, religion, education, and other factors. Though much has been accomplished in developing usable Web sites, there is still work to be done. New technologies and expanded marketplaces pose unique challenges for universally usable Web designs.

REFERENCES

Badre. (2002). Shaping Web usability: Interaction design in context. *Ubiquity.* Retrieved October 1, 2003, from *http://www.acm.org/ubiquity/book/a_badre_1.html*

Becker, S.A. (2002). An exploratory study on Web usability and the internationalization of U.S. electronic businesses. *The Journal of Electronic Commerce Research, 3*(4), 265-278.

Becker, S.A. (2004). E-government visual accessibility for older adult users. Forthcoming in *Social Science Computer Review.*

Becker, S.A., & Mottay, F. (2001). A global perspective of Web usability for online business applications. *IEEE Software, 18*(1), 54-61.

Flanders, V., & Willis, M. (1998). *Web pages that suck.* San Francisco, CA: SYBEX.

Hom, J. (1998, June). The usability methods toolbox. Retrieved October 1, 2003, from *http://jthom.best.vwh.net/usability/usable.htm*

Hurst, M. (1999, September). Holiday '99 e-commerce. *Research Report.* Creative Good, Inc. Retrieved October 1, 2003, from *http://www.creativegood.com*

Lynch, P.L., & Horton, S. (1999). *Web style guide: Basic design principles for creating Web stes.* New Haven, CT: Yale University Press.

Manning, H., McCarthy, J.C., & Souza, R.K. (1998). *Why most Web sites fail*. *Interactive Technology Series, 3*(7). Forrester Research.

Marcus, A., & Gould, W.E. (2000). Crosscurrents: Cultural dimensions and global Web user-interface design. *ACM Interactions, 7*(4), 32-46.

Morrell, R.W. (Ed.). (2002). *Older adults, health information, and the World Wide Web.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Morrell, R.W., Dailey, S.R., Feldman, C., Mayhorn, C.B., & Echt, K.V. (2002). *Older adults and information technology: A compendium of scientific research and Web site accessibility guidelines.* Bethesda, MD: National Institute on Aging.

Murray, G., & Costanzo, T. (1999). Usability and the Web: An overview. *Network Notes, 61.* Information Technology Services, National Library of Canada. Retrieved October 1, 2003, from *http://www.nlc-bnc.ca/9/1/p1-260-e.html*

Nielsen, J. (1993). *Usability engineering.* Cambridge, MA: Academic Press.

Nielsen, J. (1999a). User interface directions for the Web. *Communications of the ACM, 42*(1), 65-72.

Nielsen, J. (1999b). *Designing Web usability: The art of simplicity.* Indianapolis, IN: New Riders Publishing.

Nielsen, J., & Tahir, M. (2002). *Homepage usability 50 Websites deconstructed.* Indianapolis, IN: New Riders Publishing.

Russell, M.C., & Chaparro, B.S. (2002). Reading from a Palm Pilot™ using RSVP. *Proceedings of the Human Factors and Ergonomic Society 46th Annual Meeting* (pp. 685-689).

Schneiderman, B. (1989). *Designing the user interface: Strategies for effective human-computer interaction.* Boston, MA: Addison-Wesley.

Schneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd ed.). Boston, MA: Addison-Wesley.

Schneiderman, B. (2000). Universal usability. *Communications of the ACM, 43*(5), 85-91.

Schneiderman, B., & Plaisant, C. (2004). *Designing the user interface: Strategies for effective human-computer interaction* (4th ed.). Boston, MA: Addison-Wesley.

Spiegel, R. (2000, January). Report: 70 percent of retailers lack e-commerce strategy. *Ecommerce Times.* Retrieved October 1, 2003, from *http://www.ecommercetimes.com/news/articles2000/000126-1.shtml*

Spool, J., Scanlon, T., Schroeder, W., Snyder, C., & DeAngelo, T. (1999). *Web site usability: A designer's guide.* San Francisco, CA: Morgan Kaufman.

KEY TERMS

Dot Com: A Web site that is intended for business use, though the term is commonly used to represent any kind of Web site. The term evolved from the "com" part of a Web site's address, which represents commercial sites. It came to be associated with Internet companies that failed during the mid 2000s (*www.searchWebservices.com*).

Internationalization: It is the process of making a Web site interoperable in a specific market or locale. In general, interoperability means that the functionality of the site is not dependent on a specific language or culture and is readily adaptable to others.

Localization: It is the process of adapting an internationalized Web site to meet language, culture, religion, and other requirements of a specific market or locale.

Universal Usability: Universal usability can be defined as having more than 90% of all households as successful users of information and communications services at least once a week (Schneiderman, 2000, p. 85).

Usability: The ISO 9241-11 standard states that usability is the "effectiveness, efficiency and satisfaction with which a specified set of users can achieve a specified set of tasks in a particular environment".

Usability Engineering: It is a systematic approach to making software (Web designs) easy to use, thus meet the needs of the targeted users.

Web Accessibility: Web accessibility means that any person, regardless of disabilities, is able to use Web technology without encountering any barriers.

ENDNOTES

[1] The textbook is now in its 4th edition taking into account human factors associated with interactive systems (Schneiderman & Plaisant, 2004).

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 3074-3078, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Web Usage Mining

Stu Westin
University of Rhode Island, USA

INTRODUCTION

Research studies concerning the use of the World Wide Web (WWW) have become quite common in the MIS, education, marketing, and e-commerce literature. Increasingly, the research methodology employed in these studies involves some form of Web usage mining. This research technique seeks to uncover the Web user's access and navigation behaviors through analysis of real-time data artifacts of Web usage. These data artifacts are sometimes referred to as *mouse droppings* since each datum results from a specific user action involving the mouse. The so-called *click streams*, the sequence of URLs visited by a Web user, are often the focus of Web usage mining. These data can be supplemented with timestamp information to reveal page viewing time. Mouse click coordinates (i.e., X, Y location) can also be of interest, depending on the research question.

Studies that rely on Web usage mining can be experimental or observational in nature. The focus of such studies is quite varied and may involve such topics as predicting online purchase intentions (Hooker & Finkelman, 2004; Moe, 2003; Montgomery, Li, Srinivsan, & Liechty, 2004), designing recommender systems for e-commerce products and sites (Cho & Kim, 2004; Kim & Cho, 2003), understanding navigation and search behavior (Chiang, Dholakia, & Westin, 2004; Gery & Haddad, 2003; Johnson, Moe, Fader, Bellman, & Lohse, 2004; Li & Zaiane, 2004), or a myriad of other subjects. Regardless of the issue being studied, data collection for Web usage mining studies often proves to be a vexing problem, and ideal research designs are frequently sacrificed in the interest of finding a reasonable data capture or collection mechanism. Despite the difficulties involved, the research community has recognized the value of Web-based experimental research (Saeed, Hwang, & Yi, 2003; Zinkhan, 2005), and has, in fact, called on investigators to exploit "non-intrusive means of collecting usage and exploration data" (Gao, 2003, p. 31) in future Web studies.

In this article we discuss some of the methodological complexities that arise when conducting studies that involve Web usage mining. We then describe an innovative, software-based methodology that addresses many of these problems. The methods described here are most applicable to experimental studies, but they can be applied in ex-post observational research settings, as well.

BACKGROUND

Approaches to Web usage mining can be server-centric or client-centric. In the former case the data are harvested from a server machine. In some instances this approach requires no special software mechanisms since server logs are maintained routinely by server software. Client-centric approaches always require special data collection mechanisms because standard browsers do not document user actions. An example of this is the *PCMeter* usage mining software application. This software runs in the background on the client machine recording click-stream data as the research subject interacts with a Web browser (see Johnson et al., 2004, and Montgomery et al., 2004, for usage examples).

Server logs provide the most frequent data source for usage mining studies. This is because the data are readily available in a standard, machine-readable format, and preexisting Web sites can be used as long as the server log data can be procured for analysis. However, the literature is rife with criticism and complaints about the shortcomings of this data source (e.g., Bracke, 2004; Fenstermacher & Ginsburg, 2003; Huysmans, Baesens, & Vanthienen, 2004; Montgomery et al., 2004; Spiliopoulou, Mobasher, Berendt, & Nakagawa, 2003). The problems arise from such confounding elements as multiple server types (e.g., proxy servers, image servers, and application servers), server farms and load balancing procedures, caching activities, stateless nature of sessions, and so forth. In the words of Shahabi, Banaei-Kashani, and Faruque (2001, p. 1) "... usage data acquisition via server logs is neither reliable, nor efficient. It is unreliable due to the side effects of the network ... [it is] inefficient because of usage data requiring extensive preprocessing before it can be utilized."

Other server-centric data collection approaches based on server-side scripting (e.g., ASP, ASP.NET, etc.) can prove useful in some circumstances. Consider, for example, the situation where one wants to investigate the impact of download time as a factor affecting user satisfaction or Web site success. Using server-side scripting, a delay mechanism can be easily built into a Web page so that the server will delay serving the requested page to the client until some precise, predetermined time has passed. Different experimental treatment levels are accomplished by merely manipulating the delay time that is scripted into the Web page. Here, the

experimental subject, using an ordinary browser, will have the perception that the page is slow to download because of the delay between when the page is requested (e.g., by clicking a hyperlink) and when the page is available in the browser.

As another scenario, consider the situation where the researcher wants to study the end user's Web search strategy by analyzing the click-streams (e.g., Chiang et al., 2004). Here again, server-side scripts in the Web pages could provide a simple data collection mechanism by logging each page request (page ID, server timestamp) in a server database. The advantages of this approach over relying on server logs are that the server-side scripts can be designed to capture the precise data objects in the particular format that is desired, and many of the aforementioned confounding items can be circumvented.

In considering these scripting approaches, it is obvious that client-side data collection mechanisms can be constructed just as easily. In most cases, Java applets, Java scripts, or VB scripts can be embedded into Web pages to handle the required tasks. The only difference in this client-side approach is that the data collection is being handled by the client rather than by the server machine. Neither approach provides any obvious benefits over the other, although in the client-side approach the Web pages for an experiment could be stored locally and thus WWW, or even network access, is not required. In all of the previous research settings, including those that harvest data from server logs, standard Web browser software such as Internet Explorer (IE) can be used in the research study.

One flaw in all of these research approaches (except, perhaps, the *PCMeter* tactic) lies in the fact that experimental access must be restricted to either (1) a limited set of Web pages that have been appropriately scripted for data collection, or (2) a specific set of servers from which log data can be procured ex-post. If the experimental subject is allowed to "wander" beyond this limited set of pages or sites (an activity that is quite fundamental to the nature of using the Web), then these actions will be unrecorded or inaccessible, and the validity of the research will be nullified. In the script-based approaches, a related complexity stems from the fact that all Web pages used in the experiment must be developed and maintained by the investigator—a task that can be quite labor intensive if a large number of pages are to be made available. Obviously, the experimental pages should usually be large in number and professional in appearance if external validity is to be preserved.

In some situations the research data can be collected without the use of client- or server-side scripting, or server logs. Click-stream data, for example, can often be gleaned through the use of standard network management software, or through *network sniffers* that can be configured to monitor Internet requests and/or page downloads. In this case the experimental subject can be allowed to roam beyond a predefined set of pages, and, again, standard browser software can be used on the client side. The problem here can be in the precision or in the format of the data, as the software was not designed for this purpose. Pages containing multiple frames, for example, may be logged as individual (frame) downloads in some circumstances and as a single page download in others. Client requests that are satisfied through the local cache may not be logged at all. Indeed, this approach suffers from many of the problems that plague server logs.

A problem with all of the data collection methodologies discussed thus far is that they suffer from a lack of experimental control. This lack of control comes from the fact that the instrument with which the experimental subject is interacting (a standard Web browser such as IE) was not designed to be used as a research tool.

Consider the situation in which we wish to study WWW use behavior through analyzing click-stream data. There are numerous ways of gathering data on page requests or page downloads, as noted previously. However, there are no means, short of direct visual observation, of recording *how* a particular page was requested. The page request could have come in the form of a click on a hyperlink, but the request could just as likely have been generated automatically through a dynamic action on the page (e.g., *meta refresh*), or through the *Back* or *Forward* buttons in the browser interface. Normal click-stream data will not distinguish between these circumstances, so the precise behavior or intentions of the experimental subject cannot be determined. This specific problem is noted by Montgomery et al. (2004, p. 580) as a shortcoming of the aforementioned *PCMeter* usage mining software: "However, the meter does not distinguish how the user navigates between pages (e.g., whether the user selects a hyperlink, a bookmark, or directly types in the URL to navigate to a page)."

Another problem has to do with the occurrence of multiple windows. Many hyperlinks open in new browser windows, and the user often has the option of requesting a new window at his or her discretion. The problem here is that the data collected cannot reflect which of the open windows is active when actions occur, or even that there are multiple windows in use (the opening and closing of windows are not logged). Again, the data cannot capture, or misrepresent the behavior in question; true *streams* cannot be traced.

A CLIENT-CENTRIC ALTERNATIVE

As noted earlier, the methodological problems, for the most part, stem from a lack of experimental control. Logic and research experience suggest that, for maximum experimental control, any experimental manipulations (treatments) and the data collection mechanisms should be as close to the experimental subject as possible. That is, they should be

Table 1. Sample WebBrowser members

Event/Method	Behavior/Details	Possible Use
BeforeNavigate2 event	Triggered after page request, but before navigation begins. Provides target URL and frame information. Allows cancellation of navigation.	Can analyze URL request and covertly cancel or redirect inappropriate requests. Software can contain list of irrelevant URLs or heuristic rules (e.g., requests containing forbidden protocols or Web domains can be halted) (e.g., no returns to prior pages).
Navigate2 method	Forces browser to navigate to a new location (URL).	Can be used to covertly modify target for (cancelled) navigation request. Can be used to covertly effect programmatic delay between user's request and download activity.
DocumentComplete event	Triggered when a page has been successfully downloaded and is available for browsing. Provides full URL details.	Can be used to capture click-stream data (record URL). Can be used to timestamp click-stream data and to determine and record page-viewing time.
NewWindow3 event	Precedes creation of a new browser window. Window creation can be canceled and target redirected to main browser window.	Can force all URL targets to a single window. Can be used to block popup windows that might enter noise into experimental setting.
GoForward/*GoBack* methods	Forces navigation to subsequent (or previous) page in system-maintained URL history list.	Can emulate full behavior of IE Forward (and Back) buttons, but this behavior can be modified if needed. Note that button action can be included in click-stream data record to indicate source of request (see example later).
TitleChange/ *StatusTextChange* events	Fires whenever title (status information) of an IE session would change. Provides new text.	Can emulate behavior of an IE Title Bar (Status Bar) as needed.
GoHome, *GoSearch*, *Refresh*, *Stop* methods	Forces appropriate action of browser control.	Can easily emulate remaining features of IE interface, but all actions can be filtered through programmed decisions.

embedded in the browser itself. This leads to the development of a custom IE-look-a-like browser for use in Web-based research. The creation of such a software application is quite feasible with currently available programming tools and software techniques. The numerous benefits of this approach certainly outweigh the software development costs. The benefits are greatest when research designs are complex and when precision is of prime importance. This particular methodology has been employed in several research studies including Chiang et al. (2004) and Norberg (2003).

With custom browser software there is no need to depend on scripts or applets in experiment-specific Web pages to administer experimental treatments or to record user actions, nor is there the need to gain access to server logs. Consequently, there is no need to restrict the experimental domain to a limited set of custom Web pages, or even to a specific set of servers. With this approach, the experimental domain can include the entire Web. The custom browser software can be built with the ability to precisely record user activity and to preempt or modify actions that could be harmful or inappropriate in the experimental context. Experimental control and experimental manipulation can be integrated into the browser itself.

The software that we know as Internet Explorer (iexplore.exe) is essentially a software interface to a dynamic link library (DLL) that provides the requisite Internet processing functionality. This software component is named shdocvw.dll, but it is commonly referred to as the *WebBrowser Control* (see Microsoft Corporation, 2006a). Object-oriented programs developed for a Windows platform (e.g., C++, C#, or VB applications created in the *Visual Studio.NET* software development suite) can host this *WebBrowser* object to add Web browsing functionality to software applications. The *WebBrowser* object works with the standard event-based model of Windows computing.

With the *WebBrowser* object, event handlers are provided for all of the major occurrences in an Internet session such as *request to navigate to a page*, *page download complete*, or *request for a new window*. Key data such as URL, Target Frame, and Page Title are available with the events. In some cases, actions can be preempted through a *Cancel* argument in the event handler. One important example of this is the *BeforeNavigate* event handler. This routine is triggered *after* a navigation has been requested by the client, but *before* the request is fulfilled. This allows the hosting software application to inspect and evaluate the situation, and to possibly modify or cancel the request before it is allowed to proceed.

Properties and methods of the *WebBrowser* object can be used to dynamically emulate all of the features of the IE interface such as the status bar, the browser window caption, and the standard buttons (Back, Forward, Stop, Refresh, Home, etc.). In short, an emulation of IE can be built with the inclusion of as few or as many features of the IE interface as are needed in the research context.

Table 2. Data record structure

	Field	Details
1	Subject_ID	Auto generated by application.
2	URL_Sequence#	Auto generated by application.
3	Target_Frame	If any.
4	URL_Start_Time	Since session began (sec./1000).
5	URL_Duration	Sec./1000.
6	User_Action	Click, Back-Button, Forward-Button, Home-Button, Automatic.
7	Full_URL	Can be parsed if needed.

Table 1 describes selected events and methods of the *WebBrowser* object that are pertinent to this discussion. See Microsoft Corporation (2006b) for a full list and description of the members of this object.

Table 2 provides a sample structure of click-stream data that has been generated with this methodology. This represents one of several system-generated data tables analyzed by Chiang et al. (2004). This data table holds one record per subject per URL visited (i.e., one item in a stream). Fields 1 and 2 were generated automatically by the custom software, and provide primary and secondary sort fields for click-stream analysis. Field 3 provides data for frame-based pages and is derived from the *BeforeNavigate2* event. Field 4 (captured in *BeforeNavigate2* event), records the time at which the page was initially requested (e.g., user click). Field 5 records the duration for which the page was available to the user (*DocumentComplete* event for *current* page → *BeforeNavigate2* event for *next* page). Field 6 indicates how this page was requested ("Automatic" indicates a redirect, as by a script action or a meta refresh). Field 7 was captured by the *DocumentComplete* event handler. By writing the data record within the *DocumentComplete* event routine, unsuccessful or aborted downloads were culled from the stream. Sequence gaps in Field 2 indicate such.

In this particular experimental study, randomized assignment of subjects to treatment cells, as well as experimental session duration, was managed completely by the browser software application (after a fixed duration, the browser was programmatically disabled and the user was shown a "thank you" page). A second data table holding Subject_ID and experimental treatment information was later merged with Table 2 for analysis.

FUTURE TRENDS

As WWW becomes more and more woven into the fabric of our personal and professional lives, software developers, marketers, e-tailers, and social scientists will become increasingly interested in understanding Web access and search patterns. Most likely, this will result in an even greater proliferation of research studies involving Web usage mining. There is no reason to presume that harvesting from server logs will become significantly more efficient or reliable. Consequently, we should expect that the use of real-time, software-based approaches such as scripted data collection, the *PCMeter*, and the custom browser approach described previously will become more commonplace, especially in experimental research settings. Improvements in object-oriented software development environments such as .NET should facilitate this trend.

CONCLUSION

By developing software to host a custom browser research instrument, many of the problems that afflict other Web usage mining approaches can be circumvented. With this tactic, the investigator is free to include (covertly) all of the requisite mechanisms of experimental control and data capture into the software itself; no external scripting, log access, or network monitoring is needed. Timers to control the duration of the experiment or the occurrence of experimental treatments can be embedded into the application.

Experimental treatment randomization can also be built in. User activity down to the keystroke or mouse-click level can be monitored and recorded with millisecond accuracy if needed. Certain events can also be blocked or modified if necessary. For example, an attempt to open a page in a new window can be intercepted and the page redirected to the initial window. No special (e.g., scripted) Web pages are needed, but attempts to "wander" to irrelevant sites or inapposite protocols (e.g., *mailto*, *ftp*) can easily be halted if desired. The local cache can be controlled programmatically, thus avoiding systematic bias. Once the basic system is developed, modifications and new features are a fairly simple to put into effect.

The previous discussion and summary tables touch on the basic advantages and capabilities of this custom software approach. Many of the complexities, such as the need to override the standard context menus and some key sequences, are ignored here. Also beyond the scope of this discussion are many of the advanced capabilities such as how to gain access to an additional rich source of experimental data through use of the document object model (DOM). A more thorough presentation is presented by Westin (2003).

REFERENCES

Bracke, P. J. (2004). Web usage mining at an academic health sciences library: An exploratory study. *Journal of the Medical Library Association, 92*(4), 421-428.

Chiang, K., Dholakia, R. R., & Westin, S. (2004). Needle in a cyberstack: Consumer search for information in the Web-based environment. In B. Kahn & M. Luce (Eds.), *Advances in consumer research* (pp. 88-89). Provo, UT: Association of Consumer Research.

Cho, Y. H., & Kim, J. K. (2004). Application of Web usage mining and product taxonomy t collaborative recommendations in e-commerce. *Expert Systems with Applications, 26*(2), 233-246.

Fenstermacher, K. D., & Ginsburg, M. (2003). Client-side monitoring for Web mining. *Journal of the American Society for Information Science and Technology, 54*(7), 625-637.

Gao, Y. (2003). Web site interactivity and amusement: Techniques and effects. In S. Gordon (Ed.), *Computing information technology: The human side* (pp. 22-34). Hershey, PA: IRM Press.

Gery, M., & Haddad, H. (2003). Evaluation of Web usage mining approaches for user's next request prediction. In R. Chiang, H. Laender, & E. Lim (Eds.), *Proceedings of the Fifth ACM International Workshop on Web Information and Data Management* (pp. 74-81). New York: ACM Press.

Hooker, G., & Finkelman, M. (2004). Sequential analysis for learning modes of browsing. In B. Mobasher, B. Liu, & O. Nasraoui (Eds.), *Proceedings of the Sixth WEBKDD Workshop: Webmining and Web Usage Analysis* (pp. 1-12). Retrieved November 7, 2006, from http://maya.cs.depaul.edu/webkdd04-proceedings.pdf

Huysmans, J., Baesens, B., & Vanthienen, J. (2004). The influence of caching on Web usage mining. In A. Zanasi, N. Ebecken, & C. Brebbia (Eds.), *Proceedings of Data Mining 2004: The Fifth International Conference on Data Mining, Text Mining, and their Business Applications* (pp. 15-17). Southampton, UK: WIT Press.

Johnson, E. J., Moe, W. W., Fader, P. S., Bellman, S., & Lohse, J. L. (2004). On the depth and dynamics of online search behavior. *Management Science, 50*(3), 299-308.

Kim, J. K., & Cho, Y. H. (2003). Using Web usage mining and SVD to improve e-commerce recommendation quality. (LNCS 2891, pp. 86-97). Berlin; Heidelberg: Springer.

Li, J., & Zaiane, O. R. (2004). Combining usage, content, and structure data to improve Web site recommendation. (LNCS 3182, pp. 305-315). Berlin; Heidelberg: Springer.

Microsoft Corporation. (2006a). *WebBrowser control overviews and tutorials*. Retrieved November 11, 2006, from http://msdn.microsoft.com/library/default.asp?url=/workshop/browser/webbrowser/browser_control_ovw_entry.asp

Microsoft Corporation. (2006b). *WebBrowser control, reference for visual basic developers*. Retrieved November 11, 2006, from http://msdn.microsoft.com/library/default.asp?url=/workshop/browser/webbrowser/reflist_vb.asp

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology, 13*(1&2), 29-39.

Montgomery, A. L., Li, S., Srinivsan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science, 32*(4), 579-595.

Norberg, P. (2003). *Managed profiles: The value of personal information in commercial exchange*. Unpublished doctoral dissertation, University of Rhode Island, RI, Kingston, RI.

Saeed, K. A., Hwang, Y., & Yi, M. Y. (2003). Toward and integrative framework for online consumer behavior research: A meta-analysis approach. *Journal of End User Computing, 15*(4), 1-26.

Shahabi, C., Banaei-Kashani, F., & Faruque, J. (2001). *A reliable, efficient, and scalable system for Web usage data acquisition.* Retrieved November 11, 2006, from http://stanford.edu/~ronnyk/WEBKDD2001/ShahabiWebKDD01.pdf

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M. (2003). A framework for the evaluation of session reconstruction heuristics in Web usage analysis. *Informs Journal on Computing, 15*(2), 171-190.

Westin, S. (2003). Building a custom client-side research tool for online Web-based experiments. In S. Gordon (Ed.), *Computing information technology: The human side* (pp. 253-266). Hershey, PA: IRM Press.

Zinkhan, G. M. (2005). The marketplace, emerging technology, and marketing theory. *Marketing Theory, 5*(1), 105-115.

KEY TERMS

ASP (Active Server Page) Scripting: A simple server-side scripting approach where script code (usually VBScript or Jscript) is mixed with HTML code on a Web page. The script code is processed by a script engine before the page is rendered by the server. This can be used to create dynamic Web pages and to share data within or between Web sessions. This is a predecessor of ASP.NET technology and is sometimes called *Classic ASP*. ASP pages are identified by an ".asp" file extension. (See **ASP.NET**.)

ASP.NET: The new generation of ASP provided by the Microsoft .NET environment. ASP.NET supports a number of advanced features including server-side controls, dynamic data binding, Web services, and Web forms. All .NET programming languages (e.g., C++, C#, VB) are fully supported, so the developer is no longer restricted to using simple scripting languages. ASP.NET components are compiled thereby providing major security and performance enhancements. ASP.NET pages are identified by an ".aspx" file extension. (See **ASP Scripting**.)

Click-Stream: In Web research, the click-stream is the sequence of Web pages that is visited by the experimental subject. A click-stream data record can be as simple as URL and sequence number, or a timestamp can also be added. This latter approach allows for analysis of page viewing time.

Client-Side/Server-Side Scripting: In a Web environment, this term relates to the fact that scripted tasks can be handled by the browser software (client side) or by the Web server software (server side). A single Web page may contain both client side and server side scripts. The script host is determined by the *RUNAT* attribute of the *SCRIPT* tag.

Event Handler: A procedure (subroutine) that executes in response to an event. The event may represent a specific user action (e.g., a mouse click), or may be a manifestation of a system process (e.g., page has finished loading). Details surrounding the event are provided as arguments of the procedure.

Meta Refresh: A special HTML tag that automatically redirects the visitor to a new page. The result is that the final destination of the navigation is different from the initial target.

Method: In Object Oriented Programming, methods are the actions or behaviors that an object can perform. At the coding level, a method is created by including a procedure (function or sub) within the class.

Network Sniffer: A hardware and/or software mechanism that monitors, and possibly records, data traffic on a network.

Web Usage Mining: Harvesting and processing data for the purpose of uncovering usage and navigation patterns of Web users. This is usually recognized as a sub-category of *Web Mining*. The additional elements of this broader term are *Web Content Mining* and *Web Structure Mining*. These focus, respectively, on the information content of Web documents, and on the hyperlink structure among Web documents.

Web-Based 3D Real Time Experimentation

C. C. Ko
National University of Singapore, Singapore

Ben M. Chen
National University of Singapore, Singapore

C. D. Cheng
NDI Automation Pte Ltd, Singapore

INTRODUCTION

Spurred by development in computer science and network technology, the use of the Internet has been expanding exponentially. It is now extensively used as a connectivity and reference tool for numerous commercial, personal, and educational purposes. In education, the Internet opens a variety of new avenues and methodologies for enhancing the experience of learning as well as expanding educational opportunities for a larger pool of students. Specifically, distance education and non-traditional classrooms have the capability to reach more students using specialized instruction and self-paced learning.

In the area of distance education, many **Web-based real time experimentation** systems have been reported in the literature (Ando, Graziani, & Pitrone, 2003; Daponte, Grimaldi, & Marinov, 2002; Ko, Chen, Chen et al., 2000; Ko et al., 2001; Kumar, Sridharan, & Srinivasan, 2002; Yeung & Huang, 2003). These Internet-based remote laboratories allow users or students to carry out physical experimental work at their own pace anytime anywhere. They generally require very little physical space and minimal manpower to maintain, and are ideal for the sharing of expensive equipment. However, all these experimental systems can only provide 2D operation panels. Due to this limitation, the actual shapes of 3D instruments and equipment, some of which may have controls or display components on different sides, may not be possible to be reflected on the remote user's client display window.

BACKGROUND

Although many 3D visualization schemes on the client side have been presented (Geroimenko & Geroimenko, 2000; Hobona, James, & Fairbairn, 2006; Nakano, Sato, Matsuo, & Ishimasa, 2000; Oellien, Ihlenfeldt, & Gasteiger, 2005; Osawa, Asai, Takase, & Saito, 2001; Ueda, 2006; Vormoor, 2001) and some additional collaborative functions have been proposed for communication amongst multiple remote users or between client and server (Bender, Klein, Disch, & Ebert, 2000; Engel, Hastreiter, Tomandl, et al., 2000; Nielsen, 2006; Zhuang, Chen, & Venter, 2000), applications and issues such as Web-based real time control and 3D-based monitoring have not been addressed. We present in this article the development of **Web-based 3D real time experimentation** using Java 3D visualization tools.

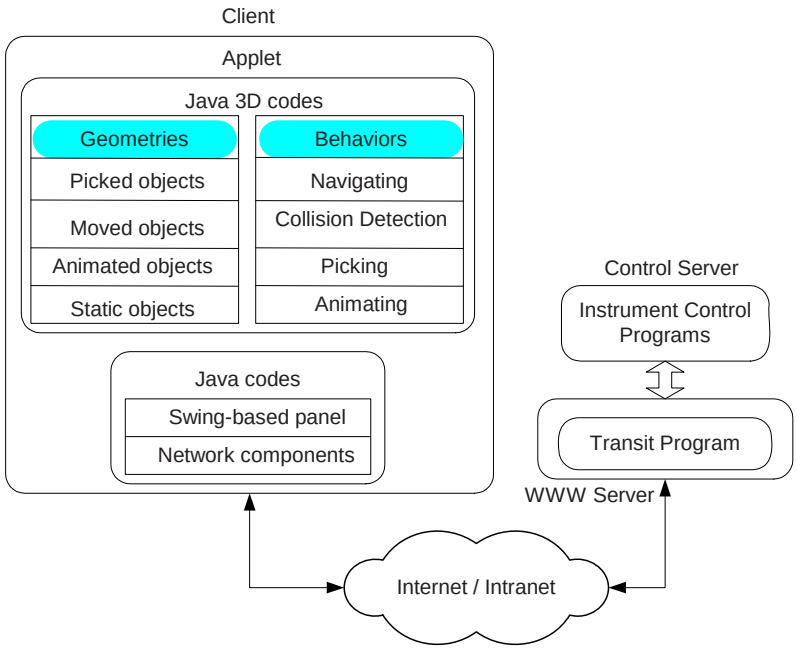
Among the various tools available, **Java 3D** is ideal from certain perspectives. Specifically, **Java 3D** is an efficient tool that provides a very flexible platform for building a wide range of Web-based three-dimensional graphics applications, and is becoming one of the most attractive tools for creating 3D user interfaces, **3D visualizations** and virtual environments. It provides not only strong **3D programming** but also excellent integration with previous version of Java components.

In comparison with other 3D virtual experimental systems, this chapter attempts to address all the important issues with an emphasis to provide a complete solution. Specifically, issues on connecting actual experimental instruments, real time data transmission, three-dimensional virtual scene and three-dimensional behaviors are addressed. These ensure that the user will get a more realistic feeling when operating and controlling three-dimensioanl experimental instruments as well as monitoring actual experimental results without any significant delay.

PROPOSED SYSTEM REFERENCE MODEL

Figure 1 shows our reference model for the creation of **Web-based, 3D, real-time experimentation** using Java 3D visualization tool. Note that on the client side, the combination of Java 3D API and Java realizes 3D visualization and network connection. Usually, 3D visualization consists of geometry and **behavior objects**. The former includes the picked, moved, animated and static objects, and the lat-

Figure 1. System reference model for creating Web-based, real-time experimentation



ter consists of navigating, collision detection, picking and animating behaviors.

The picked objects cover all controls such as buttons, knobs, sliders and connectors of the virtual equipment and experiment, while the moved objects include curve, text and screen displays on the **virtual instruments**. The animated objects cover all active and periodically moved objects. The other visual objects, such as walls, windows, tables and certain components of some **virtual instruments**, are taken to be static ones. These **geometry objects** designed using Java 3D helps to promote the rendering efficiency of the 3D virtual scene. In the relevant behavior objects, only the animating behaviors provided by Java 3D API are used without modifications.

PROPOSED HARDWARE ARCHITECTURE

The system reference model of Figure 1 can be supported by the double-server-client distributed hardware architecture of Figure 2. The whole system includes a user's computer on the client side, the Internet and/or an intranet to transmit command and data, a Web server to host the Web site of the remote experimentation, and a control server with control cards attached to programmable instruments together with some circuit boards.

In particular, programmable instruments have to be connected to the control server through control cards and cables in a 3D remote experimentation system. For example, two separate TCP/IP interface modules are used for real time control and retrieval. The commands coming in through the TCP/IP control interface are converted into the format required before being sent to the programmable instrument to be controlled. Experimental data for the generation of real time curve or text for the user is transmitted to the client through the TCP/IP retrieval module.

3D INSTRUMENTS AND SCENE

To be as realistic as possible, and to overcome certain limitation posed by 2-D operation panels, using which the actual shapes of **3D instruments** and equipment cannot be shown, the use of 3D visualization tools in real time Web-based experimentation may be considered. Figure 3 shows a typical example GUI realization on the client computer developed based on Java 3D.

Ideally, anyone conducting an experiment through the Internet should be able to do it in the same manner as in a real laboratory. This can be accomplished in a 3D environment through three behavior modules on navigating, collision detection and picking in the GUI interface.

Figure 2. Hardware architecture for creating web-based real time experimentation

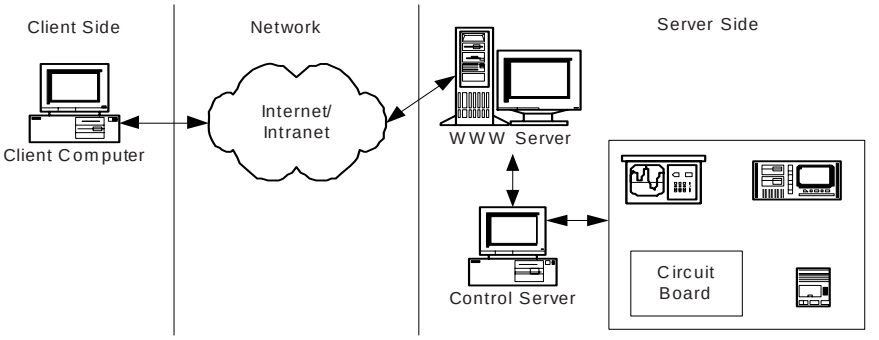
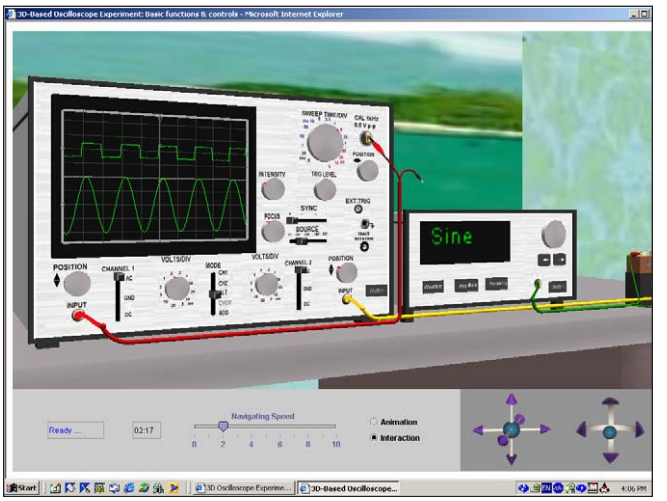


Figure 3. 3D view on an experiment with a battery



The module on **navigating behavior** controls how the user walks around in the **virtual laboratory**. Also, as the user attempts to get a better view, it controls indirectly the positions and angles of the view platform. The **collision detection** module ensures that the user does not traverse any solid objects such as walls, tables and instruments. Through the picking behavior module, the user will be able to adjust the controls of available experimental apparatus precisely.

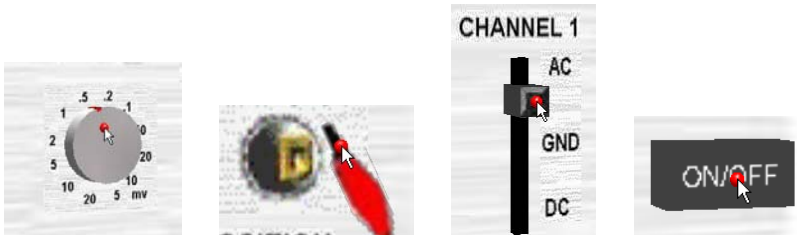
The experiment is performed when the user enters a virtual laboratory as shown in Figure 4. Apart from the **virtual instruments** and circuit board placed on a table, the virtual laboratory also includes the floor, the ceiling, a few walls, two windows and one door. To enter the virtual laboratory, a user will need to "walk" to the door and press the door open button. Upon opening, the user can walk through the door, move around and carry out the experiment by adjusting the instruments and circuit board on the table.

MAIN FUNCTIONS

The main functions and features in such a **Web-based 3D remote laboratory** system are summarized below:

1. A 3D remote laboratory controlling actual instruments and displaying real signals is implemented through a dynamic **virtual scene** via the Internet. In the example in Figure 3, the laboratory has an oscilloscope, a signal generator, a battery, a few cables and some other visual objects.
2. A navigation tool for walking around the virtual laboratory is provided. For example, the tool on the bottom right hand corner in Figure 3 allows the user to move around the virtual laboratory and view instruments from different positions and directions through the mouse.

Figure 5. Picking controls for instruments



(5a) **(5b)** **(5c)** **(5d)**

3. A collision detection mechanism is implemented. This guarantees that the viewing platform will not traverse any solid objects such as walls, doors, windows, tables and virtual instruments.
4. Through the appropriate picking function, the user can adjust individual controls on the instruments in the 3D environment and connect circuits in the same way as he or she operates an actual instrument in the real laboratory. As shown in Figure 5, the operations of turning a knob, adjusting a slider, pressing a button, and making a connection to a terminal can be performed by simply dragging the mouse to move the relevant control when the control is in "focus." To make it as user friendly as possible, a red point is displayed when the mouse is over a control that has received focus.
5. The adjusted controls are converted into the relevant commands and sent to a control server to control real instruments in the actual physical laboratory. The result of the experiment is sent back by the server to the client to be displayed in the 3D virtual laboratory in real time.

FUTURE TRENDS

Despite a number of efforts, there are still some limitations on the widespread application of **Web-based, 3D experimentation system**. Specifically, such systems generally require more complicated software and hardware configurations and support such as high-end display card and DirectX/OpenGL plug-in to run DirectX or OpenGL graphic libraries. Also, the Internet has certain inherent constraints for the transmission of large experimental data streams. Lastly, due partly to the variety of non-standard development systems that need to be used and the complexity of integration hardware apparatus and software applications, the number of sites that supports such systems is still rather small.

Nevertheless, as powerful computer hardware becomes more readily available and new development and software tools are introduced, these problems may be overcome to a certain extent in the next five to ten years. Web-based 3D experimentation will perhaps provide an evolutionary influence on 3D network applications, and as the Internet becomes faster and more universities and research institutes become interested in this area, more experimental sites may become available.

CONCLUSION

A 3D Internet oscilloscope experiment has been developed, and has been used to support the teaching of undergraduate courses in the Department of Electrical and Computer Engineering, National University of Singapore. It has also received good rating from experts in the Java 3D interest group on aspects such as user interface, ease of use and usefulness.

REFERENCES

Ando, B., Graziani, S., & Pitrone, N. (2004). Stand-alone laboratory sessions in sensors and signal processing. *IEEE Transactions on Education, 47*(1), 4-9.

Bender, M., Klein, R., Disch. A., & Ebert, A. (2000). A functional framework for Web-based information visualization systems. *IEEE Transactions on Visualization and Computer Graphics, 6*(1), 8-23.

Daponte, P., Grimaldi, D., & Marinov, M. (2002). Real-time measurement and control of an industrial system over a standard network: implementation of a prototype for educational purposes. *IEEE Transactions on Instrumentation and Measurement, 51*(5), 962- 969.

Engel, K., Hastreiter, P., Tomandl, B., Eberhardt, K., & Ertl, T. (2000). Combining local and remote visualization techniquues for interactive volume rendering in medical applications. *IEEE Conference on Visualization*, 449-452.

Geroimenko, V., & Geroimenko, L. (2000). Visualizing Human Consciousness Content Using Java 3D/X3D and Psychological Techniques. *Proceedings Information Visualization Conference*, (pp. 529-532).

Hobona, K., James, P., & Fairbairn, D. (2006). Web-based visualization of 3D geospatial data using Java3D. *IEEE Computer Graphics and Applications, 26*, 28-33.

Ko, C. C., Chen, B. M., Chen, S. H., Ramakrishnan, V., Chen, R., Hu, S.Y. et al. (2000). A large scale Web-based virtual oscilloscope laboratory experiment. *IEE Engineering Science and Education Journal, 9*(2), 69-76.

Ko, C. C., Chen, B. M., Hu, S. Y., Ramakrishnan V., Cheng, C. D, Zhuang, Y. et al. (2001). A Web-based virtual laboratory on a frequency modulation experiment. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 31*(3), 295-303.

Ko, C. C., Chen, B. M., Chen, J., Zhuang, Y., & Tan, K. C. (2001). Development of a web-based laboratory for control experiments on a coupled tank apparatus. *IEEE Transactions on Education, 44*(1), 76-86.

Kumar, B. R., Sridharan, K., & Srinivasan, K. (2002). The design and development of a Web-based data acquisition system. *IEEE Transactions on Instrumentation and Measurement, 51*(3), 427-432.

Nakano, H., Sato, Y., Matsuo, S., & Ishimasa, T. (2000). Development of 3D Visualization system for the study of physical properties of quasicrystals. *Materials Science and Engineering*, 294-296.

Nielsen, J. F. (2006). A modular framework for development and interlaboratory sharing and validation of diffusion tensor tractography algorithms, *Journal of Digital Imaging, 19*, 112-117.

Oellien, F., Ihlenfeldt, W., & Gasteiger, J. (2005). InfVis—Platform-independent visual data mining of multidimensional chemical data sets, *Journal of Chemical Information and Modeling, 45*, 1456-1467.

Osawa, N., Asai, K., Takase, N., & Saito, F. (2001). An immersive system for editing and playing music on network-connected computers. *Proceedings 5th International Conference on Information Visualization*, 630–635.

Ueda, M. (2006). Making of the simplest interactive 3D digital globe as a tool for the world environmental problems, *WSEAS Transactions on Environment and Development, 2*, 973-979.

Vormoor, O. (2001). Quick and easy interactive molecular dynamics using Java 3D. *Computing in Science & Engineering*, 98-104.

Yeung, K., & Huang, J. (2003). Development of a remote-access laboratory: A dc motor control experiment. *Computers in Industry, 52*(3), 305-311.

Zhuang, Y., Chen, L., & Venter, R. (2000). CyberEye: An internet-enabled environment to support collaborative design. *Concurrent Engineering: Research and Applications, 8*(3), 213-229.

KEY TERMS

Internet Remote Experimentation: The use of the Internet to carry out physical experimental work at a remote location.

Online Experiment: An experiment that is running and controlled by a computer terminal.

Virtual Laboratory: A computer accessible laboratory which may be simulated by running a software package or which may involve real remote experimentation.

Web-Based Control: The control of instruments or apparatus through the Internet.

Web-Based Laboratory: A laboratory that typically involves physical experiments and that can be accessed remotely through the use of the Internet.

Web-Based 3D Navigation: The use of the mouse or keyboard to navigate in a Web-based three-dimensional virtual scene.

Web-Based 3D Picking: The use of the mouse to manipulate the controls of Web-based three-dimensional virtual or real instruments.

Web-Based 3D Visualization: A 3-dimensional scene display that can be accessed through normal Internet explorer such as IE or Netscape.

Web-Based Algorithm and Program Visualization for Education

Cristóbal Pareja-Flores
Universidad Complutense de Madrid, Spain

Jaime Urquiza-Fuentes
Universidad Rey Juan Carlos, Spain

J. Ángel Velázquez Iturbide
Universidad Rey Juan Carlos, Spain

INTRODUCTION

Programming is a central activity in the computing profession. It is facilitated by different tools (editors, compilers, debuggers, etc.), which are often integrated into programming environments. Programming also plays a central role in computer science education. For this purpose, a number of complementary tools were developed during the last decade: algorithm animators, program visualizers, problem generators, assignment graders, and so forth.

After the Web explosion, teachers of programming rapidly turned their attention to the Web. Although the Web has speed and power limitations, it also has several advantages that make it invaluable for educational purposes. Mainly, it provides universal accessibility and platform independence, and solves the distribution problem by always making available the last version of any tool.

Probably, the most common use of the Web for programming courses is as a communication medium, facilitating submission and administration of assignments and grades (Burd, 2000). Another common use of the Web for programming education is as a public repository of high quality problems, such as the *Lab Repository* (Knox, 2006) and the *ACM International Collegiate Programming Contest* (Skiena & Revilla, 2003). Web sites may also host other resources, such as slides and audio lectures (Skiena & Revilla, 2003), algorithm animations (Brummond, 2001), or programming tools (English, 2001). These collections have no structure or, at best, are lineally or hierarchically structured, but more advanced repositories are possible. In this case, a management system must be delivered that, using (semi)structured mark-up languages, allows retrieving, maintaining, and publishing. A good representative is the eXercita system (Gregorio-Rodríguez et al., 2000, 2002). Finally, programming tools have been ported to be executed on the Web (Pareja-Flores & Velázquez-Iturbide, 2002).

This article describes a different class of Web-based tools for programming education, namely tools for algorithm and program visualization. After the Background section, we describe the evolution of these systems, educational uses, and lessons learned. Finally, we outline future trends in the use of the Web for programming education and our personal conclusions.

BACKGROUND

Algorithm animation is a research field that is now 20 years old and still evolving. There is a consensus with respect to the videotape *Sorting out Sorting* presented in 1981 by Baecker (1998), which is considered a landmark on animation. It included animations of nine sorting algorithms. Afterward, some works established the main techniques for specifying and implementing algorithm animation: BALSA (Brown, 1988), Tango (Stasko, 1990), and Pavane (Roman, Cox, Wilcox, & Plun, 1992). Systematic categorizations of software visualizations have been proposed since then (Price, Baecker, & Small, 1998).

In a general setting, software visualization studies the visual representation of software entities (Stasko, Domingue, Brown, & Price, 1998). Visualization requires an effort to abstract the target entity to visualize and to make a good graphical design that may yield many different representations: text versus graphics, level of abstraction, displaying control versus data structures, static versus dynamic visualizations, one or multiple views, behavior versus performance, and so forth. This general aim can be achieved in different ways. Program visualization is aimed at the visualization of a piece of software so that the representation has a close relationship to the source code. Algorithm animation is aimed at the dynamic visualization of a piece of software illustrating the main ideas or steps (i.e., its algorithmic behavior), but without a close relationship to the source code. The graphical nature of software visualization in general and algorithm animation in particular makes them very conductive to the hypermedia features of the Web.

WEB-BASED ALGORITHM AND PROGRAM VISUALIZATION SYSTEMS

In the mid 1990s, many of the algorithm animation systems were ported to the Web, and many additional systems were specifically designed for the Web. A representative work from these years is that of Naps (1996), in which he carried out a study of the technical alternatives that could be used to make animations produced by previous systems available on the Web. Other pioneer systems are Mocha (Baker, Cruz, Liotta, & Tamassia, 1996), JCAT (Brown & Raisamo, 1997), and Jeliot (Haajanen et al., 1997).

Some animation systems are general purpose, so they can be used to generate visualizations of any entity, including non-programming entities. They typically provide a scripting language that makes animation creation a relatively easy task. Good representatives are the ANIMAL (Roessling & Freisleben, 2002), JAWAA (Pierson & Rodger, 1998), and Samba (Stasko, 1997) systems. The JHAVÉ system (Naps, Eagan, & Norton, 2000) is not an algorithm visualization system itself, but rather a support environment to render visualizations from a variety of visualization systems. Currently, it supports algorithm visualizations in three scripting languages: ANIMAL, GAIGS (Naps, 1990), and Samba. With JHAVÉ, students can explore algorithms, controlling movement and responding to pop-up questions. Leonardo Web (Bonifaci, Demetrescu, Finocchi, & Laura, 2003) is a collection of tools for creating animated representations via the use of a visual editor and a Java library. The presentations can be viewed by a simple Java player, so being possible to integrate the animations in Web pages, or to browse them off-line.

Another category of systems automatically produce program animations, enhanced with graphical representations. Some of these systems are computer-supported learning systems that only provide visualization support for a given domain and with a specific didactic purpose. For instance, the KIEL system (Berghammer & Milanese, 2001) gives support to learning the ML functional language. The system allows students to input a program and an expression to evaluate and then graphically displays the evaluation of the expression. The student may navigate both forward and backward through the evaluation process. A successful experience comes from the "problets" by Kumar (2003). They are tutors aimed at specific programming topics, which randomly generate problems and ask questions to the student. Visualizations are used as a visual aid for problem solving. Problets have been developed for a number of topics: expression evaluation, identifier scope, and so forth. A similar approach has been addressed elsewhere for data structures (Baker, Boilen, Goodrich, Tamassia, & Stibel, 1999).

A different category of systems providing automatic visualizations are based on the availability of complete language processors. The WinHIPE system (Velázquez-Iturbide, Pareja-Flores, & Urquiza-Fuentes, 2006) is an integrated development environment (IDE) for functional programming that allows (semi)automatically generating animations of the evaluation of a functional expression. Then, the animation can be played within the IDE or exported to the Web (Urquiza-Fuentes & Velázquez-Iturbide, 2005), also giving support to manage collections of animations.

Other automatic visualization systems use the Web as a user interface for programming, and allow loading and executing programs. For instance, the ISVL is designed to learn Prolog programming online (Domingue & Mulholland, 1998). ISVL allows students to generate animations based on AND-OR trees that can be communicated to the course tutor and even stored as films. The tutoring may insert annotations on the student animations to respond to their questions or to explain where a programming error was made.

The Jeliot family (Ben-Ari, Myller, Sutinen, & Tarhio, 2002) is a collection of program and algorithm visualization tools designed for novices learning programming, algorithms, and data structures with Java. The last member of the family is Jeliot 3 (Moreno, Myller, Sutinen, & Ben-Ari, 2004), specially focused to animate object oriented features, such as inheritance.

The last class of animation systems is based on simulation. MatrixPro (Karavirta, Korhonen, Malmi, & Stalnacke, 2004) enables instructors to create algorithm animations by direct manipulation of a data structures library. Teachers can demonstrate the visual simulation of an algorithm by providing on the fly different input sets. "What-if" questions can also be invoked in lectures. A related tool, TRAKLA2 (Korhonen, Malmi, & Silvasti, 2003), is designed for building and solving interactive algorithm simulation exercises, made available as applets, for learning data structures and algorithms. In these exercises, students directly manipulate conceptual visualizations of data structures in order to simulate the working of algorithms.

EDUCATIONAL USES OF WEB-BASED ALGORITHM AND PROGRAM VISUALIZATION

From the outset, the main use of algorithm animation was educational, rather that industrial (for instance, as a debugging tool). Algorithm animation systems have been used in several ways: as a complement to lectures on algorithms, for self-study, or within laboratories. A more demanding use of animation systems consists in requiring students to build their own animations.

The best documented experience ran for about 20 years at the Computer Science Department of Brown University (Bazik, Tamassia, Reiss, & van Dam, 1998). All the experiences reported in the available literature agree that students

are highly motivated by algorithm animations. Hundhausen (Hundhausen, Douglas, & Stasko, 2002) provides a comprehensive review of experiences in using animations for education.

Design of Effective Visualizations

It is difficult to provide canned recipes for the design of visualizations. Several authors (Gloor, 1998; Khuri, 2001; Saraiya, Shaffer, McCrickard, & North, 2004) summarize some general recommendations for their design. Some commonly accepted suggestions for educational use follow:

- Make the meaning of the different graphical representations explicit, explained either by embedding them in the system or reinforced by allocating time to the subject during the course.
- Adapt to the knowledge level of the user, either novice or expert. In the first case, the system must be easy of use, whereas in the second case, it may be more comprehensive.
- Complement visualizations with explanations.
- Be interactive, allowing flexible control of animations (including movement in a backward direction), customization, input data, or unfolding unexpected short questions.
- Provide multiple views, which must be consistently coordinated.
- Include performance information, for example, data collected during the execution of an algorithm, or animating several algorithms simultaneously.
- Include execution history.
- Include canned examples.

Issues for Educational Use

Unfortunately, visualization systems have failed to find widespread use in computer science education. Two main problems can be identified (Naps, Roessling, Almstrum et al., 2003). The first problem is the lack of evidence on the educational effectiveness of visualization systems, that is, whether students learn more using them. In last years many evaluations have been performed on the educational effectiveness of visualization systems. The results vary and even contradict. Researchers in the field have finally come to the conclusion that these systems are effective only if students are actively engaged (Hundhausen et al., 2002).

Based on this conclusion, a taxonomy consisting in several levels of engagement has been proposed (Naps, Roessling, Almstrum et al., 2003). The lowest level of engagement corresponds to no-visualization, the next one to passive viewing, and the following levels successively force the student to engage in answering questions, changing the visualization, constructing a new visualization, and finally presenting a new visualization to an audience.

A second problem is on the instructor's side. Using a visualization system often poses a heavy workload on the instructor (Naps, Roessling, Almstrum et al., 2003; Naps, Roessling, Anderson et al., 2003). The many issues the instructor must face have been identified, and suggestions to constructors of these systems have been given (Naps, Roessling, Anderson et al., 2003). Many of these suggestions are solved or alleviated by means of the Web. For instance, the availability of an official and comprehensive Web site is a good way to announce a system and give support to potential users.

The main bottleneck is the huge amount of time necessary to build new visualizations. The term "effortlessness" has been coined (Ihantola, Karavirta, Korhonen, & Nikander, 2005) to refer to the feature of an algorithm or program visualization system that supports the creation or use of visualizations in a course without much effort from the instructor's perspective. There is no definitive definition of a system being effortless, but is seems to be related to how wide its scope is, how easy its adaptation to a given course is, and how easy different educational uses are supported.

FUTURE TRENDS

Algorithm and program animation achieved technical maturity in the early 1990s. In the years thereafter, most efforts have focused on porting to the Web and on enhancing their educational adequacy. We believe that we will see more efforts aimed at designing and evaluating the elements of effective and effortless animations.

Currently, there are several research efforts aimed at designing and developing dynamic electronic books, with dynamic contents and virtual laboratories. We believe that these virtual laboratories for programming will incorporate interfacing with programming tools such as program compiling and execution, and will include automatic visualization and animation. Some inspiration can be obtained from the hypertext books by Rockford Ross (Ross & Grinder, 2002) on automata theory.

CONCLUSION

We have provided a panoramic view of the educational use of algorithm and program visualization on the Web. We have offered an overview of the main technical achievements and the educational uses of them. We have also explained issues to consider in an educational setting and some of the lessons that have been learned from experience. In the near future, we expect that there will be increased interest in Web-based tools to support these activities.

REFERENCES

Baecker, R. (1998). Sorting out sorting: A case study of software visualization for teaching computer science. In J. T. Stasko, J. Domingue, M. H. Brown, & B. A. Price (Eds.), *Software visualization* (pp. 369-381). Cambridge, MA: MIT Press.

Baker, J. E., Cruz, I. F., Liotta, G., & Tamassia, R. (1996). Algorithm animation over the World Wide Web. In T. Catarci, M. F. Costabile, S. Levialdi, & G. Santucci (Eds.), *Proceedings of the International Workshop on Advanced Visual Interfaces (AVI '96)* (pp. 203-212). New York: ACM Press.

Baker, R. S., Boilen, M., Goodrich, M. T., Tamassia, R., & Stibel, B. (1999). Testers and visualizers for teaching data structures. *ACM SIGCSE Bulletin, 31*(1), 261-265.

Bazik, J., Tamassia, R., Reiss, S. P., & van Dam, A. (1998). Software visualization in teaching at Brown University. In J. T. Stasko, J. Domingue, M. H. Brown, & B. A. Price (Eds.), *Software visualization* (pp. 383-398). Cambridge, MA: MIT Press.

Ben-Ari, M., Myller, N., Sutinen, E., & Tarhio, J. (2002). Perspectives in program animation with Jeliot. In S. Diehl (Ed.), *Software visualization* (pp. 31-45). Berlin-Heidelberg, Germany: Springer-Verlag.

Berghammer, R., & Milanese, U. (2001). KIEL—A computer system for visualizing the execution of functional programs. In M. Hanus (Ed.), *Proceedings of the International Workshop on Functional and (Constraint) Logic Programming, (WFLP 2001)* (pp. 365-368). Kiel, Germnay: Christian-Albrechts-Universitt zu Kiel.

Bonifaci, V., Demetrescu, C., Finocchi, I., & Laura, L. (2003). A Java-based system for building animated presentations over the Web. *Science of Computer Programming, 53*(1), 37-49.

Brown, M. H. (1988). *Algorithm animation*. Cambridge, MA: MIT Press.

Brown, M. H., & Raisamo, R. (1997). JCAT: Collaborative active textbooks using Java. *Computer Networks and ISDN Systems*, *29*, 1577-1586.

Brummond, P. (2001). *The complete collection of algorithm animations*. Retrieved July 21, 2006, from http://www.cs.hope.edu/~alganim/ccaa

Burd, D. D. (2000). Web based support of a programming class. In A. Aggarwal (Ed.), *Web-based learning and teaching technologies: Opportunities and challenges* (pp. 175-197). Hershey, PA: Idea-Group Publishing.

Domingue, J., & Mulholland, P. (1998). An effective Web based software visualization learning environment. *Journal of Visual Languages and Computing, 9*(5), 485-508.

English, J. (2001). *BURKS 6 online*. Retrieved July 21, 2006, from http://burks.bton.ac.uk/

Gloor, P. A. (1998). User interface issues for algorithm animation. In J. T. Stasko, J. Domingue, M. H. Brown, & B. A. Price (Eds.), *Software visualization* (pp. 145-152). Cambridge, MA: MIT Press.

Gregorio-Rodríguez, C., Llana-Díaz, L., Palao-Gostanza, P., Pareja-Flores, C., Martínez-Unanue, R., & Velázquez-Iturbide, J. Á. (2000). *EXercita*. Retrieved June 2, 2001, from http://aljibe.sip.ucm.es

Gregorio-Rodríguez, C., Llana-Díaz, L., Palao-Gostanza, P., Pareja-Flores, C., Martínez-Unanue, R., & Velázquez-Iturbide, J. Á. (2002). A system to generate electronic books on programming exercises. *The Electronic Library, 20*(4), 314-321.

Haajanen, J., Pesonius, M., Sutinen, E., Tarhio, J., Teräsvirta, T., & Vanninen, P. (1997). Animation of user algorithms on the Web. *vl 00*, 360-367.

Hundhausen, C. D., Douglas, S. A., & Stasko, J. T. (2002). A meta-study of algorithm visualization effectiveness. *Journal of Visual Languages and Computing, 13*(3), 259-290.

Ihantola, P., Karavirta, V., Korhonen, A., & Nikander, J. (2005). Taxonomy of effortless creation of algorithm visualizations. In *Proceedings of the International Computing Education Research Workshop (ICSE 2005)* (pp. 123-133). New York: ACM Press.

Karavirta, V., Korhonen, A., Malmi, L., & Stalnacke, K. (2004). MatrixPro—A tool for on-the-fly demonstration of data structures and algorithms. In A. Korhonen (Ed.), *Proceedings of the Third Program Visualization Workshop* (pp. 26-33). Coventry, UK: University of Warwick.

Khuri, S. (2001). Designing effective algorithm visualizations. In E. Sutinen (Ed.), *Proceedings of the First Program Visualization Workshop* (pp. 1-12). Joensuu, Finland: University of Joensuu.

Knox, D. L. (2006). *The computer science teaching center*. Retrieved July 21, 2006, from http://www.cstc.org/

Korhonen, A., Malmi, L, & Silvasti, P. (2003). TRAKLA2: A framework for automatically assessed visual algorithm simulation exercises. In *Proceedings of Koli Calling—Third Annual Baltic Conference on Computer Science Education* (pp. 48-56). Helsinki, Finland: Helsinki University Printing House.

Kumar, A. N. (2003). Learning programming by solving problems. In L. Cassel & R. A. Reis (Eds.), *Informatics curricula and teaching methods* (pp. 29-39). Norwell, MA: Kluwer Academic Publishers.

Moreno, A., Myller, N., Sutinen, E., & Ben-Ari, M. (2004). Visualizing programs with Jeliot 3. In *Proceedings of the Working Conference on Advanced Visual interfaces (AVI'04)* (pp. 373-376). New York: ACM Press.

Naps, T. L. (1990). Algorithm visualization in computer science laboratories. *ACM SIGCSE Bulletin, 22*(1), 105-110.

Naps, T. L. (1996). Algorithm visualization served on the World Wide Web: Why and how. *ACM SIGCSE Bulletin, 28*(SI), 59-61.

Naps, T. L., Eagan, J. R., & Norton, L. L. (2000). JHAVÉ: An environment to actively engage students in Web-based algorithm visualizations. *ACM SIGCSE Bulletin, 32*(1), 109-113.

Naps, T. L., Roessling, G., Almstrum, V., Dann, W., Fleischer, R., Hundhausen, C., et al. (2003). Exploring the role of visualization and engagement in computer science education. *ACM SIGCSE Bulletin, 35*(3), 131-152.

Naps, T. L., Roessling, G., Anderson, J., Cooper, S., Dann, W., Fleischer, R., et al. (2003). Evaluating the educational impact of visualization. *ACM SIGCSE Bulletin, 35*(4), 124-136.

Pareja-Flores, C., & Velázquez-Iturbide, J. Á. (2002). Program execution and visualization on the Web. In A. Aggarwal (Ed.), *Web-based learning and teaching technologies: Opportunities and challenges* (pp. 236-259). Hershey, PA: Idea-Group Publishing.

Pierson, W. C., & Rodger, S. H. (1998). Web-based animations of data structures using JAWAA. *ACM SIGCSE Bulletin, 35*(1), 267-271.

Price, B., Baecker, R., & Small, I. (1998). An introduction to software visualization. In J. T. Stasko, J. Domingue, M. H. Brown, & B. A. Price (Eds.), *Software visualization* (pp. 3-27). Cambridge, MA: MIT Press.

Roessling, G., & Freisleben, B. (2002). ANIMAL: A system for supporting multiple roles in algorithm animation. *Journal of Visual Languages and Computing, 13*(3), 341-354.

Roman, C.-G., Cox, K., Wilcox, C., & Plun, J. (1992). Pavane: A system for declarative visualization of concurrent computations. *Journal of Visual Languages and Systems, 3*, 161-193.

Ross, R. J., & Grinder, M. T. (2002). Hypertextbooks: Animated, active learning, comprehensive teaching and learning resources for the Web. In S. Diehl (Ed.), *Software visualization* (pp. 269-283). Berlin-Heidelberg, Germany: Springer-Verlag.

Saraiya, P., Shaffer, C., McCrickard, D. S., & North, C. (2004). Effective features of algorithm visualizations. *ACM SIGCSE Bulletin, 36*(1), 382-386.

Skiena, S. S., & Revilla, M. A. (2003). *Programming challenges: The programming contest training manual*. Berlin-Heidelberg, Germany: Springer-Verlag.

Stasko, J. T. (1990). Tango: A framework and system for algorithm animation. *Computer, 23*(9), 27-39.

Stasko, J. (1997). Using student-built algorithm animations as learning aids. *ACM SIGCSE Bulletin, 29*(1), 25-29.

Stasko, J. T., Domingue, J., Brown, M. H., & Price, B. A. (Eds.). (1998). *Software visualization*. Cambridge, MA: MIT Press.

Urquiza-Fuentes, J., & Velázquez-Iturbide, J. Á. (2005). Effortless construction and management of program animations on the Web. In R. W. H. Lau, Q. Li, R. Cheung, & W. Liu (Eds.), *Advances in Web-based learning—ICWL 2005* (pp. 163-173). Berlin-Heidelberg, Germany: Springer-Verlag.

Velázquez-Iturbide, J. Á., Pareja-Flores, C., & Urquiza-Fuentes, J. (2006). An approach to effortless construction of program animations. *Computers & Education*.

KEY TERMS

Algorithm Animation: Visualization of a piece of software illustrating the main ideas or steps (i.e., its algorithmic behavior), but without a close relationship to the source code.

Effortlessness: Feature of an algorithm or program visualization system that refers to its support to create or use visualizations in a course without much effort from the instructor's perspective.

Engagement Levels: Taxonomy that communicates the degree of involvement of the learners in an educational situation that includes visualization.

Information Visualization: Visualization of phenomena by means of appropriate representations. It is a field different from scientific visualization since information visualization emphasizes delivering visualizations that improve comprehension, whereas scientific visualization emphasizes delivering realistic visualizations.

Multiple Views: Visualization technique consisting in showing several aspects or views of a program, where any view typically shows few aspects and must be comprehen-

sible by itself. Typical views are code vs. data, several levels of abstraction, logic vs. performance, history, and several simultaneous algorithms.

Program Visualization: Visualization of a piece of software so that the representation has a close relationship to the source code.

Visualization: To make something visible by means of some representation. Visualization consists of two elements: a mental process and a graphical language. Note that "to visualize" is different from "to see."

Web-Based Customer Loyalty Efforts and Effects on E-Business Strategy

Guisseppi Forgionne
University of Maryland, Baltimore County, USA

Supawadee Ingsriswang
Information Systems Laboratory, BIOTEC Central Research Unit, Thailand
National Center for Genetic Engineering and Biotechnology (BIOTEC), Thailand
National of Science and Technology Development Agency (NSTDA), Thailand

INTRODUCTION

Despite continued market growth, a number of Web sites have been unprofitable. Some notable failures were eToys.com, boo.com, bluefly.com, buy.com, and valueamerica.com. An examination of the companies' IPO filings suggests that the collapses were caused by cut-price strategies, over-investment, incorrect expectations, and non-profitability. Surviving in the digital market has become a critical challenge for Web managers.

To face the business challenge, Web managers and marketers demand information about Web site design and investment effectiveness (Ghosh, 1998). As the rate and diversity of product/service innovation declines and competition intensifies, Web managers need better research on Internet-related investment decisions (Donath, 1999; Hoffman, 2000). The original article examined the role of customer retention actions in these decisions (Forgionne & Ingsriswang, 2005). This article updates the original analysis by incorporating new research on comprehensive e-business strategy models.

BACKGROUND

Since online consumers can switch to other Web sites or competitive URLs in seconds with minimal financial costs, most Web sites invest heavily in programs to attract and retain customers. The Web site's ability to capture consumers' attention is known widely as "stickiness."

From one perspective (Rubric, 1999, p. 5), stickiness is the ability of a Web site to capture and keep a visitor's attention for a period of time. Alternatively, stickiness can be described as the ability of the site in attracting longer and more frequent repeat visits or the ability of the site to retain customers (Anders, 1999; Davenport, 2000; Hanson, 2000; Murphy, 1999; O'Brien, 1999; Pappas, 1999). On the other hand, Demers and Lev (2000) represent stickiness by the average time spent at the site per visit.

These views suggest that stickiness is similar to, if not the same as, the marketing concept of customer loyalty (Morgan & Hunt, 1994). The idea is to develop and maintain long-term relationships with customers by creating superior customer value and satisfaction (Reichheld, 1996). Enhanced customer satisfaction results in customer loyalty and increased profit (Anderson, Fornell, & Lehmann, 1994; Reichheld & Sasser, 1990). Loyal customers, who return again and again over a period of time, also are valuable assets of the Web site. The ability to create customer loyalty has been a major driver of success in e-commerce (Reichheld, Markey, & Hopton, 2000; Reichheld & Schefter, 2000) since enhanced customer loyalty results in increased long-term profitability.

On the basis of marketing theory, then, stickiness can be viewed as the ability of a Web site to create both customer attraction and customer retention for the purpose of maximizing revenue and profit. Customer attraction is the ability to attract customers at the Web site both frequently and for long durations, while customer retention is the ability to maintain customer loyalty.

MODELING E-COMMERCE LOYALTY

E-commerce customer loyalty, or stickiness, results from goodwill created by the organization's marketing efforts (Reichheld, 1996; Reichheld & Sasser, 1990), or

$$\text{Stickiness} = f(\text{Goodwill}) \quad (1)$$

and

$$\text{Goodwill} = f(\text{Marketing Mix}). \quad (2)$$

By encouraging current and return visits, stickiness will influence the organization's sales volume. Marketing theory also suggests that the mix of price (including switching costs to consumers), product/service (including site characteristics), and promotion (including banner and other

Web site ads), as well as other factors (including consumer characteristics), will influence this volume (Page, Pitt, & Berthon, 1996; Storbacka, Strandvik, & Gronroos, 1994). Conceptually, then,

Sales Volume = f (Stickiness, Promotion, Product, Price, Other Factors). (3)

Such volume will determine revenue, cost, and thereby profit for the Web site.

According to standard accounting practice and economic theory, profit is defined as the excess of revenue over cost, or

Profit = Revenue - Cost, (4)

while revenue will equal sales volume multiplied by price, or

Revenue = Sales Volume × Price. (5)

The same standard accounting and economic theory indicates that an organization's costs will have fixed and variable components, or

Total Cost = Fixed Cost + Variable Cost, (6)

and variable cost will equal sales volume multiplied by unit cost, or

Variable Cost = Sales Volume × Unit Cost (7)

with sales volume as defined in equation (3).

These marketing-economic-accounting-based relationships are illustrated in Figure 1. This figure and equations (1) through (7) provide a conceptual model that specifies the manner in which the stickiness investment contributes to an e-business's success by generating sales volume, revenue, and profit. As such, the model identifies the information that must be collected to effectively evaluate investments in an e-commerce customer loyalty plan. These equations also provide a framework to objectively evaluate these plans and their impact on organizational operations and activities.

In practice, the model can be applied in a variety of ways. The general relationships can be used for strategy formulation at the macro-organizational level. In addition, the equations can be decomposed into detailed micro level blocks with many variables and interrelationships. At this micro level, tactical policies to implement the macro strategies can be specified and evaluated.

Empirical Testing

Much of the required model information is financial in nature. Such data are largely proprietary and therefore not readily available.

Summarized financial information is available from some Internet companies' quarterly (10-Q) and annual reports and other public sources. In particular, data were obtained from the 10-Q and annual (10-K) reports for 20 Internet companies over the period of 1999-2000. These pooled cross-section, time series data provided values for the revenue and marketing mix variables found in equations (1) through (7). Customer characteristics were proxied through demographic variables, and data for these variables were obtained from the U.S. Census Bureau, Statistical Abstract of the United States for the period of 1999-2000. The quarterly and annual report and Census data provided 120 observations to operationalize the model embodied in equations (1) through (7).

Goodwill, as defined in the economic and marketing literature (repeat business based on happy customers) is not available from the annual and quarterly reports. However, an accounting measure of goodwill, amortization of goodwill and other intangible assets, is available from the reports. Although not strictly the same concept as marketing goodwill, the accounting measure is likely to be correlated with economic goodwill and is thereby used as a proxy for

Figure 1 Stickiness conceptual model

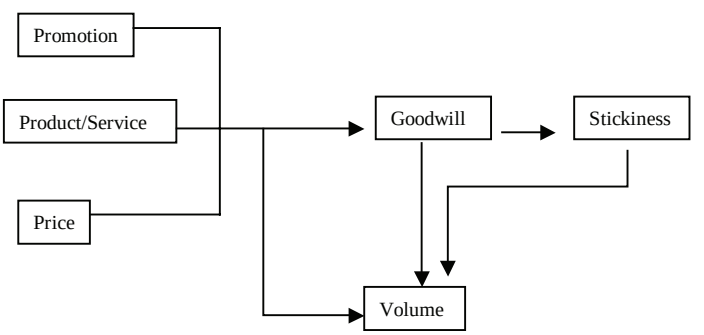


Table 1. Variables with data

VARIABLE	DEFINITION
Volume **V**	total sales volume of the Internet company
Goodwill **GW**	amortization of goodwill and other intangible assets
Promotion **A**	sales and marketing expenditures
Product Quality **Q**	product development costs, which includes research and development and other product development costs
Average Income **AI**	average consumer income
Stickiness **S**	number of unique visitors at the firm's Web site

economic goodwill (Chauvin & Hirschey, 1994; Jennings, Robinson, Thompson, & Duvall, 1996; McCarthy and Schneider, 1995).

No other data were provided by the available sources on a consistent and reliable basis. These data limitations reduced the variable list that could be used to operationalize the stickiness model specified in equations (1) through (7) to the list summarized in Table 1. To test the theory, then, it was necessary to use a truncated form of Figure 1's stickiness model.

Since, equations (4) through (7) in Figure 1's stickiness model are identities, only equations (1) through (3) must be estimated statistically from the available data. Prior research has shown that marketing investments in one period can continue to affect volume in subsequent periods (Hanson, 2000). Marketers call this phenomenon the carryover, lagged, or holdover effect. Stickiness investments and the marketing mix can be expected to create such an effect from new customers who remain with the Web site for many subsequent periods. Because of the data limitations, this important lagged effect had to be measured quarterly, and the carryover had to be restricted to one period. From the relationships in Figure 1, it seems reasonable to assume that the holdover will be expressed through the goodwill variable. Namely, goodwill in the previous quarter is assumed to affect revenue in the current quarter.

At time period *t*, the inputs may have joint, rather than independent, effects on sales volume. Moreover, elasticities (the percentage change in an output generated by a 1% change in an input) can provide useful information for resource allocation analysis. To account for the non-linearities, to facilitate the computation of elasticities, and to account for the carryover effect, equations (1) through (3) can be specified with the popular Cobb-Douglas production function or in log-linear form as follows:

$$\log (V_t) = \log (a_0) + a_1 \log(S_t) + a_2 \log(A_t) + a_3 \log (AI_t) + a_4 \log(GW_{t-1}) + z_1 \quad (8)$$

$$\log (S_t) = \log (b_0) + b_1 \log (GW_t) + z_2 \quad (9)$$

$$\log (GW_t) = \log (c_0) + c_1 \log (Q_t) + z_3 \quad (10)$$

with *t* being the considered time period (quarter), the a, b, c, and z labels denoting parameters to be estimated, and the other variables as defined previously. Price is excluded from the volume equation because data for the price variable are not available from the annual and quarterly reports or from any other published sources.

Equations (8) through (10) form a simultaneous system of equations. Hence, simultaneous equation estimation techniques should be used to separate the effects of the marketing instruments and stickiness on sales volume. Otherwise, there will be a statistical identification problem.

There are several simultaneous estimation techniques that can be used to estimate the operational stickiness model (equations [8] through [10]). Using coefficients of determination, theoretical correctness of the estimated parameters, and forecasting errors as guidelines, the best results were generated by seemingly unrelated regressors (SURs). These results are summarized in Table 2 and Table 3.

Table 2 shows that the volume equation (8) has an $R^2 = 0.74$ and thereby accounts for about 74% of the variation in this variable. More importantly, Table 3 shows that the estimated coefficient of the stickiness variable in equation (9) is significant at the $\alpha = .05$ level, suggesting that sales volume is significantly influenced by the stickiness of the Web site.

Table 2 shows that the model accounts for a relatively small percentage of the variation in stickiness and goodwill. Nevertheless, Table 3 indicates that goodwill has positive and significant effects on the stickiness of the Web site, while

Table 2. Results of the SUR estimation

Equation	MSE	R-Square	Adj. R-Square
Volume (equation [8])	0.3894	0.7408	0.7299
Stickiness (equation [9])	0.2190	0.1944	0.1862
Goodwill (equation [10])	3.5312	0.1296	0.1208

Table 3. SUR-estimated log-linear result

Variable	Volume (V_t)		Stickiness (S_t)		Goodwill (GW_t)	
	Est.	Prob>|T|	Est.	Prob>|T|	Est.	Prob>|T|
Constant	-5.09	.8282	8.74	<.0001	1.19	.0037
Stickiness	0.58	.0002				
Promotion	0.95	<.0001				
Income	0.09	.9885				
Lagged Goodwill	0.05	.1739				
Goodwill			0.17	<.0001		
Product Quality					0.77	<.0001

marketing instruments, such as product quality, also create positive effects on goodwill.

In short, the operational stickiness model provides reasonable statistical results. These results validate the theory that stickiness is an investment that, along with the marketing mix, will influence organizational performance. The results also indicate that sales volume will have inelastic responses to stickiness and promotion investments, that is, a 1% increase in such investments will lead to a less than 1% increase in sales volume. There are similar inelastic responses of stickiness to goodwill and goodwill to product quality.

E-Business Strategy Models

Recently, comprehensive e-business strategy models have been developed and tested that build on and incorporate the original stickiness model. In one model (Ha & Forgionne, 2006), current, rather than lagged, goodwill is treated as a determinant of quantity demanded (sales volume). In addition, there is a competition equation that includes supply chain, rivalry, and other customer loyalty measures that replaces the stickiness (9) and goodwill (10) equations in the operational stickiness model. The first e-business strategy model also offers equations for quantity supplied and employee efficiency. Empirical results, not yet reported in the literature, offer support for the e-business strategy model.

In another model (Wang & Forgionne, 2006), a balanced scorecard approach is used to provide a comprehensive view of quantity demanded and supplied and their roles in determining revenue, cost, and profit. This model explicitly links goodwill, customer retention, customer loyalty, and customer satisfaction. In this view, goodwill and customer satisfaction are determinants of quantity demanded, and customer satisfaction is determined by the product and supply characteristics. The second e-business strategy model also involves more relationships between the various factors than presented in the first e-business strategy or operational stickiness model. Empirical results, not yet reported in the literature, offer support for the second e-business strategy model.

While the two recent strategy models offer different perspectives, each indicates that e-business management is a complex problem involving many interrelated factors. Stickiness, or customer retention and loyalty as expressed in the recent models, is only a component in a comprehensive view of e-business management. Developing a strategy that focuses on stickiness while downplaying the other factors and relationships, then, is likely to result in less than optimal e-business performance. Moreover, the incomplete focus may create problems in managing the other aspects of the e-business.

FUTURE TRENDS

The stickiness analysis suggests future research directions. Neither of the recent e-business strategy models considers dynamic effects. It is useful to determine whether there are significant carryover effects from goodwill. If significant carryover effects are discovered, research should examine how long the investment in stickiness (or customer retention and loyalty) would contribute to the firm's short-term or long-term financial success. The recent e-business strategy models considered competition, and such research should continue. In particular, research should evaluate what role competition plays in stickiness (or customer retention and loyalty). For example, additional research may examine how competitors' stickiness (or customer retention) investments impact the firm.

The recent e-business strategy models demonstrated that empirical e-commerce evaluations will require the collection, capture, and retrieval of pertinent and consistent operational financial and other data for evaluation purposes. Appropriate statistical methodologies will be needed to estimate the proffered model's parameters from the collected data. The recent e-business strategy research also indicated that information systems will be required to focus the data, assist managers in the estimation process, and help such users to evaluate experimental customer loyalty plans. These tasks and activities offer new, and potentially very productive, areas for future research.

CONCLUSION

Figure 1's stickiness model shows how to measure the impact of online stickiness on revenue and profit, and the model also shows how stickiness and goodwill are related. While this operational model can forecast the long term revenue, profit, and return on investment from a Web site's products/services, the new e-business strategy models provide a more comprehensive view of the underlying relationships. In addition, the operational stickiness model, and its e-business strategy successors: (a) provide a mechanism to evaluate Web redesign strategies, (b) help Web managers evaluate market changes on custom loyalty plans, and (c) provide a framework to determine the "best" stickiness and other marketing policies.

In sum, in the e-commerce and e-business world, marketing still provides leadership in identifying consumer needs, the market to be served, and the strategy to be launched. Web managers must realize that desirable financial outcomes depend on their marketing effectiveness. With Figure 1's conceptual stickiness model and the recent e-business strategy models, Web managers and marketers have tools to evaluate their investment decisions.

REFERENCES

Anders, G. (1999). The race for sticky Web sites—Behind the deal frenzy, a quest to hang onto restless clickers. *Wall Street Journal*.

Anderson, E. W., Claes, F., & Lehmann, D. R. (1994). Customer satisfaction, marketshare and profitability: Findings from Sweden. *Journal of Marketing, 58*(July), 53-66.

Chauvin, K. W., & Hirschey, M. (1994). Goodwill, profitability, and the market value of the firm. *Journal of Accounting and Public Policy, 13*, 159-180.

Davenport, T. H. (2000). Sticky business. *CIO Magazine* (February). Retrieved from www.cio.com/forums/ec

Demers, E., & Lev, B. (2000). *A rude awakening: Internet value-drivers in 2000*. New York University. Retrieved from http://www.stern.nyu.edu/~blev/newnew.html

Donath, B. (1999). *The quest for e-business frameworks* (White Paper). The Pennsylvania State University, E-Business Center.

Forgionne, G. A., & Ingsriswang, S. (2005). Stickiness and Web-based customer loyalty. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and technology* (pp. 2610-2615). Hershey, PA: Idea Group Reference.

Ghosh, S. (1998). Making business sense of the Internet. *Harvard Business Review, 76*(2), 126-135.

Ha, L., & Forgionne, G. A. (2006). Econometric simulation for e-business strategy evaluation. *International Journal of E-Business Research, 2*(2), 38-53.

Hanson, W. (2000). *Principles of Internet marketing*. South-Western College Publishing.

Hoffman, D. L. (2000). The revolution will not be televised: Introduction to the special issue on marketing science and the Internet. *Marketing Science, 19*(1), 1-3.

Jennings, R., Robinson, J., Thompson, R. B., & Duvall, L. (1996). The relationship between accounting goodwill numbers and equity values. *Journal of Business Finance and Accounting, 23*(4), 513-533.

McCarthy, M. G., & Schneider, D. K. (1995). Market perception of goodwill: Some empirical evidence. *Accounting and Business Research, 26*(1), 69-81.

Morgan, R., & Hunt, S. (1994). The commitment trust theory of relationship marketing. *Journal of Marketing*, (July), 20-38.

Murphy, K. (1999). Stickiness is the new gotta-have. *Internet World, 5*(12), 1-2.

O'Brien, J. (1999). Sticky shopping sites. *Computer Shopper, 19*(7), 121.

Page, M., Pitt, L., & Berthon, P. (1996). Analysing and reducing customer defection. *Long Range Planning, 29*(6), 821-824.

Pappas, C. (1999). Let's get sticky. *Home Office Computing, 17*(1), 90-91.

Reichheld, F. E. (1996). Learning from customer defections. *Harvard Business Review,* (March-April), 56-69.

Reichheld, F. E., & Sasser, W. E. (1990). Zero defections: Quality comes to services. *Harvard Business Review, 14*(March), 495-507.

Reichheld, F. F., Markey, R., & Hopton, C. (2000). The loyalty effect—The relationship between loyalty and profits. *European Business Journal, 12*(3), 134-139.

Reichheld, F. F., & Schefter, P. (2000). E-loyalty: Your secret weapon on the Web. *Harvard Business Review, 78*(4), 105-113.

Rubric, Inc. (1999). Evaluating the sticky factor of e-commerce sites. Retrieved from http://www.rubricsoft.com

Storbacka, K., Strandvik, T., & Gronroos, C. (1994). Managing customer relationships for profit: The dynamics of relationship quality. *International Journal of Service Industry Management, 5*(5), 21-38.

Wang, F., & Forgionne, G. A. (2006). EBBSC: A balanced scorecard based framework for strategic e-business management. International Journal of E-Business Research, forthcoming.

KEY TERMS

Customer Attraction: The ability to attract customers at the Web site.

Customer Loyalty: The ability to develop and maintain long-term relationships with customers by creating superior customer value and satisfaction.

Customer Loyalty Plan: A strategy for improving financial performance through activities that increase stickiness.

Customer Retention: The ability to retain customers and their allegiance to the Web site.

Goodwill: The amount of repeat business resulting from happy and loyal customers.

Stickiness: The ability of a Web site to create both customer attraction and customer retention for the purpose of maximizing revenue or profit.

Stickiness Model: A series of interdependent equations linking stickiness to an organization's financial performance.

Web-Based Expert Systems

Yanqing Duan
University of Luton, UK

INTRODUCTION

Convergence of technologies in the Internet and the field of expert systems (ES) offer new ways of sharing and distributing knowledge (Sedbrook, 1998). Power (2000) argues that rapid advances in Internet technologies have opened new opportunities for enhancing traditional decision support systems and expert systems. Internet technology can change the way that an expert system is developed and distributed. For the first time, knowledge on any subject can directly be delivered to users anywhere and anytime through a Web-based ES. Because the main function of an ES is to mimic expertise and distribute expert knowledge to nonexperts, these benefits can be greatly enhanced with the emergence of the Internet.

The current focus on networked and Internet-based applications demands new architectures for "intelligent" systems as well as creating new possibilities for research and development in this field (Caldwell, Breton, & Holburn, 2003). This article provides an overview of Web-based expert systems with examples. Benefits and challenges are discussed by comparing Web-based ES with traditional standalone ES from both the development and the application perspectives using Turban and Aronson's knowledge engineering framework.

BACKGROUND

An expert system is "*a system that uses human knowledge captured in a computer to solve problems that ordinarily require human expertise*" (Turban & Aronson, 1998, p. 440). Durkin (1996) reports that many organisations have leveraged ES to increase productivity and profits through better business decisions. Although there have been reports of ES failures (Wong, 1996), research (Yoon, Guimaraes, & O'Neal, 1995) shows that many companies have remained enthusiastic proponents of the technology and continue to develop important ES applications.

The early applications of ES were standalone applications based on mainframe, Artificial Intelligence (AI) workstation or PC platforms. Later came local area network (LAN)-based distributed applications. Grove (2000) identified several problems associated with traditional ES applications, including knowledge bottleneck, performance brittleness, availability of the system, problems with individual software installation and upgrading, and a lack of common protocols for knowledge transfer.

Web-Based ES Application

The Internet offers an ever-expanding set of capabilities and Web-based ES is capable of offering much more than traditional ES. However, the literature appears to offer contradictory pictures on the current status of Web-based ES in practice. Grove (2000) provides some examples of Web-based ES in industry, medicine, science and government and claims that "there are now a large number of ES available on the Internet" (p. 130). Grove (2000) argues that there are several factors that make the Internet, by contrast to stand-alone platforms, an ideal base for Knowledge Based System (KBS) delivery.

Grove (2000) also identifies several problems that are associated with the development of Web-based KBS, such as to keep up with rapid technological change, including servers, interface components, inference engines, and various protocols; and the potential delivery bottleneck caused by communication loads and limited infrastructure. Adams (2001) points out in line with Grove that there are numerous examples of expert systems on the Web, but many of these systems are small, noncritical systems.

Perhaps the most successful example is Web-based legal ES, reported by Bodine (2001), which enable law firms to collect hundreds of thousands of dollars in subscription fees from clients who use their advisory services based on the Web. PT Consulting Partners in the USA (2005) reported that they have helped its clients build a number of successful Web-based ES, which have brought significant benefits to the client company. Grupe (2002) reported a Web-based ES called Student Advisement, which is an online ES helping students select an academic major.

The question remains: are there really not many ES on the Web or is it the case that most Web-based ES are not being reported in academic literature? It is evident that the situation on Web-based ES in practice is not very clear and further formal investigation needs to be carried out to offer better insight into the current situation.

As well as the limited number of reports on applications of Web-based ES, there also appears to be a lack of a general methodology for the development of Web-based ES. While there is a significant promise in the idea of developing Web-

based ES, there are also some challenges that have not yet been fully explored.

Knowledge Engineering for Expert System Development

To better understand the challenges and benefits of Web-based ES, the traditional knowledge engineering process for ES development is revisited. According to Turban and Aronson (2001), Knowledge Engineering (KE) is a process unique to ES development. It deals with knowledge acquisition, validation, representation, inferencing, explanation, and maintenance.

In order to cover all issues related to Web-based system development and application, the KE process in this article is extended to include evaluation, implementation and maintenance, which is depicted in Figure 1. This extended KE process is used as a framework that informs the analysis and discussion of ES benefits and challenges.

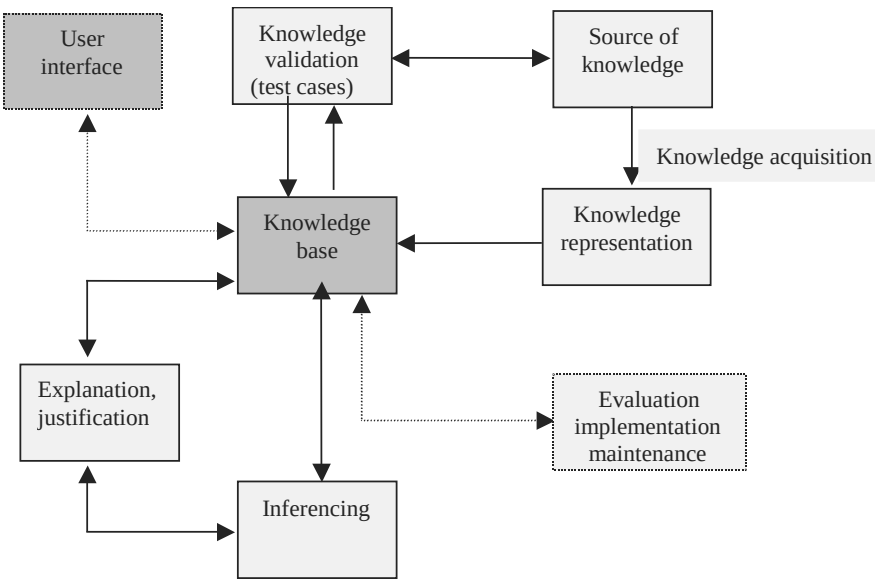
EXAMPLES OF WEB-BASED EXPERT SYSTEMS

Example 1: WITS

WITS is a Web-based Intelligent Training and Support system. It is developed for providing training and intelligent support for Small and Medium sized Enterprises (SMEs) on the use of Information and Communication Technologies (ICT). The research was inspired by the evidence from the literature that lack of adequate skills and knowledge is one of major barriers for SMEs in successfully adopting and running e-commerce and e-business. As a result of this deficiency, there is an emerging need for better education and decision support to SME managers who are eager to embrace the technology and afraid of being left behind.

WITS advisor has three subsystems, which are designed to facilitate SME managers' decision making process in e-commerce and e-business adoption. Compared with traditional ES, Web-based ES makes the evaluation and implementation of WITS much easier. There is no need to install the system in advance. It is easy to collect feedback from online forms. By using Web site analysis software, visitors can be easily traced and analyzed. By collecting visitors' information, it is possible to profile the users, and determine the usefulness of the system. The use of Web design software makes the user interface design easier. HTML-based user interfaces allow the incorporation of rich media elements. Hyperlinks in HTML provide an extra facility in enhancing ES explanation and help functions as users can access the relevant Web site easily. This is normally not possible with stand alone ES. Also, WWW has been proven to be a useful knowledge source for knowledge acquisition in constructing the WITS knowledge base. With a Web-based knowledge base, any knowledge updating and maintenance can be handled centrally, and no reinstallation needs to be carried out. Useful links are incorporated in the system which can help the user to understand and interpret the expert system's recommendations. E-mails, feedback forms and other Internet communication functions allow users to question and comment on the system, thus making an expert system more interactive and personal.

Figure 1. Extended process of knowledge engineering (Adapted from Turban and Aronson, 1998)



As with all ES, WITS also has certain limitations. It does not have an easy to use knowledge updating facility. Any attempt to update and modify domain knowledge has to be done by the system developers using the original programs. As e-commerce and e-business is a fast moving area, keeping the system's knowledge base updated is a serious challenge.

Example 2: Fish-Expert

Fish disease diagnosis is a rather complicated process in aquaculture production activities. Fish-Expert is a Web-based expert system for fish disease diagnosis in China. This Web-based expert system can mimic human fish disease expertise and diagnose a number of fish diseases with a user-friendly interface. A fish disease diagnosis expert system contains a large amount of fish disease data and images, which are used to conduct online disease diagnosis. More information about Fish-Expert can be found in Li, Duan and Fu (2002). The system has been tested and is in pilot use in certain regions of North China.

A number of points related to the benefits and challenges of Web-based ES emerged with the development and use of Fish-Expert, and these include:

- Online knowledge acquisition is welcomed by domain experts, but a knowledge engineer is still needed to check and transfer the knowledge into the knowledge base.
- The online user feedback form and online evaluation of ES seem to be effective and popular.
- Internet access speed is seen as a bottleneck for Web-based ES applications, especially in developing countries.
- The multimedia interface in Fish-Expert is effective in helping the user to query the system, but it slows down the access speed to it.
- ES are known for their inability to deal with exceptions or complex problems due to the inflexibility and limits of the knowledge base. The interactive nature of Internet communication provides an opportunity to reduce these limitations of ES with a Web-based telediagnostic system. In this case, users can talk to a human expert via telediagnostic equipment in either a synchronous or asynchronous manner.

Example 3: Online Legal Expert Systems

People may argue that the variety of skills and breadth of knowledge required of a lawyer mean that a computer could not feasibly be built to replace a lawyer (Booth, 2003), but one of the most successful online ES applications is online legal expert systems. Bodine (2001) reports that law firms find new paths to profit on the Web. "They are collecting hundreds of thousands of dollars in subscription fees from clients who use their question-and-answer advisory services based on the Web" (Bodine, 2001, p. 1). These online "expert systems" are working from local legal aid societies all the way up to international mega law firms. It is argued that rather than acting as a static Intranet repository of policies, the expert system responds to the user's input and generates tailored output, highlighting key issues for further consideration (Booth, 2003).

According to Bodine's (2001) report, Web-based legal expert systems are working like a question-and-answer session. Users go to a Web site, where the expert system prompts them to explain their question. The questions vary according to the answers and facts that the users give in response. The session concludes when the user reaches the Web page with the answer; a copy is typically e-mailed as a report to a lawyer in the firm.

Although it may be expensive to set up an online ES, the long term benefits and potential profit are also prominent. The advantages of using online legal expert systems include:

- **Internal:** The system operates on an Intranet to provide expertise and answer routine legal questions among a firm's employees (Bodine, 2001).
- **Private for clients:** The system operates on an extranet. Clients pay a fee and get a username and password to log on to the system (Bodine, 2001).
- **Public on a Web site:** The public can visit a Web site to get answers about their legal rights or complete an intake form (Bodine, 2001).
- **24/7 and worldwide access** to some legal advice and **documentary evidence** of compliance (Booth, 2003).

The profits are found in the private client expert systems. For example, a large international law firm in Australia operates its Advertising Copy Compliance Advisor to give legal guidance to brand managers at a client corporation. The lawyers were not always available immediately, so the law firm created the system on the Web that is running all day, every day. A London law firm runs an online service, which gives practical legal advice in many jurisdictions on setting up an e-commerce venture. The firm charges $3,000 per topic per jurisdiction per year for unlimited users. One big accounting firm operates an online business advisor. Subscribers pay $5,000 per year to log onto the system for answers to questions about tax, technology and business matters.

"The question is not whether computers can replace lawyers," said Kevin Mulcahy, Director of Jnana Technologies in New York. "The question is how will lawyers use the Web to optimize the way they sell their knowledge and expertise" (Bodine, 2001, p. 3). Challenges to the implementation of

Web-based legal expert systems identified by Booth (2003) include setting up cost, identifying suitable processes or areas of law, cultural and behavioral factors, relationships and business improvements.

BENEFITS AND CHALLENGES OF WEB-BASED EXPERT SYSTEMS

Challenges related to Web-based ES can be examined from different perspectives: technological, methodological, economic, and social. The aforementioned three examples have demonstrated some of the benefits and also the technological and methodological challenges of developing and using Web-based ES. The following paragraphs provide some discussion on the benefits and challenges using Turban and Aronson's knowledge engineering process presented in Figure 1.

Knowledge acquisition–The impact of the Internet on knowledge acquisition can be profound. Firstly, it provides another valuable knowledge source for knowledge acquisition. Secondly, it makes knowledge elicitation from the domain expert possible at a distance, thus eliminating the time and space barriers for experts to contribute to the knowledge base. Thirdly, as Basden (2000) argued, not only can knowledge be made available, but also the users can be closely involved in its selection and generation. For example, online communities can be a potential knowledge source, as one of the aims people have in joining such a community is to share knowledge in their domain of common interest. However, these benefits bring with them associated problems and challenges. These include dealing with information overload with effective knowledge mining techniques, locating and verifying online "experts," filtering knowledge collected via the Web, managing conflict when multiple online experts are involved, and security and reliability considerations.

Knowledge representation and inferencing-Traditional development methodology, tools and techniques which work effectively in a standalone environment may not work well in a Web environment. It appears that there is a need to investigate if the development methodology for a traditional expert system needs to be revised to suit the nature of the Web-based environment. Some Web-based ES have been developed following an ad hoc approach. There is a lack of systematic methodology or process suitable for Web-based ES development and implementation (Dokas, 2005). There is a lack of skilled knowledge engineers and an absence of standard procedures and development practices for online ES (Booth, 2003).

Knowledge validation–The knowledge validation, verification and testing process is likely to be one of the most affected processes in ES development and applications. Users can directly submit their test cases or provide feedback to system developers via the Internet. Alternatively, the knowledge base can be uploaded for validation and be accessed directly by users with their test cases. However, this user-centred approach needs a centrally managed validation process. Generic online debugging tools, such as rule base debugging in the cases discussed here, would be very welcome to developers.

Explanation and justification–One of the distinguishing features of an ES is its capability for explanation and justification. Compared with stand-alone ES, this function can be greatly enhanced with Internet technology. Explanation and justification can be improved with online hyperlinks to external resources, such as other relevant sites, and also with extensive use of graphics which do not have to be stored locally. It is also possible to receive explanation and justification from a human expert via Internet facilities such as video conferencing, chat room, message board or e-mail.

System evaluation, implementation and maintenance. As with the validation process, great benefits can be obtained by making an expert system available on the Web. From the users' point of view, systems can be easily accessed globally. Users will not be disadvantaged by their location. From the technical point of view, no installation is needed at the users' location. Any updating and maintenance can be carried out centrally at the server side. ES can provide users the means to submit their thoughts experiences and knowledge. Users' feedback on the overall system performance can be collected via online feedback forms and automatically saved in the database for easy analysis afterwards. Web-site analysis tools can be installed to trace the number of visitors and their visiting behaviour, which is not normally possible with traditional ES. The challenge is how to turn this valuable information directly into knowledge for system improvement and enhancement.

Web-based ES development tools-Traditional ES were developed for stand-alone computers and a number of development tools have long been available for their development. However, many traditional ES shells do not support the openness and interoperability required for deploying ES over a wide area network (Sedbrook, 1998). With the rapid development of the Internet, more Web-based ES are beginning to emerge. Unfortunately, the Web was originally conceived simply as a document distribution infrastructure (Riva, Bellazzi, & Montani, 1998) and any attempt to use it for distributing ES must cope with certain difficulties (Huntington, 2000; Riva et al., 1998).

Culture and behavioural challenges associated with the wide spread of online ES should also not be underestimated. The take up of Web-based ES requires a culture shift for staff to move from traditional face-to-face communication to an online inquires. This lack of human interaction will stir up the ongoing debate around the dehumanisation of society through increased use of machines, the role of ICT and its impact on interpersonal communication and relationships.

Moreover, ICT support staff needs to be well trained on how to deal with a large number of ES users if an online ES is to be widely deployed.

FUTURE TRENDS

It is envisaged that Web-based ES will become more sophisticated, complex, and capable (Grove, 2000) and fulfil their great promise in enhancing traditional ES. However, this relies on further research to address the challenges posed by developing and implementing Web-based ES, as discussed in this article. Web-based ES could bring new life to the somewhat "out of fashion" field of ES and generate a new era for expert system applications. A number of Web-based ES have been reported in the literature, but this number is expected to grow (Dokas, 2005). It is a technology that has yet to realise its full potential (Booth, 2003). Eventually, online ES will be an essential integral component of any online system by working along with online database and information processing systems and performing critical intelligent functions.

CONCLUSION

This article aims to provide an overview of Web-based ES and its associated benefits and challenges. The integration of Internet technology with the field of ES offers new ways of sharing and distributing knowledge and has changed the way that ES can be developed and distributed. The development and application of Web-based ES have benefits in many ways. The essence of an expert system is to mimic expertise and distribute expert knowledge into nonexperts' hands. This ultimate benefit can be enhanced greatly with the Internet technologies. However, the development of Web-based ES also brings a number of challenges to its development and applications.

REFERENCES

Adams, J.A. (2001, October). The feasibility of distributed Web-based expert systems. In *Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics*, Tucson, AZ. Retrieved April 14, 2008, from www.cs.yit.edu/~jaa/papers/SMC01.pdf

Basden, A. (2000). Some technical and non-technical issues in implementing a knowledge server. *Software–Practice and Experience, 30*(10), 1127-1164.

Bodine, L. (2001, June 12). Delivering legal services via Web-based expert systems: Law firms find new paths to profit on the Web. *The Sugarcrest Report* (No. 4). Retrieved December 11, 2007, from http://www.sugarcrest.com/newsletters/v1_issue4.htm

Booh, K. (2003). A solution in search of a problem? Challenges to implementing Web-based legal expert system technology. *Computerisation of Law Resources, 22*. Retrieved December 11, 2007, from http://bar.austlii.edu.au/au/other/CompLRes/2003/22.html

Caldwell, N. H. M., Breton, B. C., & Holburn, D. M. (2003). Web-based expert systems: Information clients vs. knowledge servers. *Intelligent exploration of the Web (Studies in fuzziness and soft computing)* (pp. 402-417). Heidelberg, Germany: Physica Verlag GmbH.

Dokas, I. M. (2005, September). Developing Web sites for Web-based expert systems: A Web engineering approach. In *Proceedings of the Information Technologies in Environmental Engineering (ITEE'2005),* Otto-von-Guericke-Universität Magdeburg, Germany, (pp. 202-207).

Durkin, J. D. (1996). Expert systems: A view of the field. *IEEE Expert, 11*(2), 56-63.

Grove, R. F. (2000). Internet-based expert systems. *Expert Systems, 17*(3), 129-136.

Grupe, F. H. (2002, December). Student advisement: Applying a Web-based expert system to the selection of an academic major. *College Student Journal*. Retrieved December 11, 2007, from http://www.findarticles.com/p/articles/mi_m0FCR/is_4_36/ai_96619963

Huntington, D. (2000). Web-based expert systems are on the way: Java-based Web delivery. PC *AI Intelligent Solutions for Desktop Computers, 14*(6), 34-6.

Li, D., Fu, Z., & Duan, Y. (2002). Fish-expert: A Web-based expert system for fish disease diagnosis. *Expert Systems with Applications: An International Journal, 23*(3), 311-320.

Power, D. J. (2000, August 10-13). Web-based and model-driven decision support systems: Concepts and issues. In *Proceedings of the Americas Conference on Information Systems (AMCIS 2000),* Long Beach, CA.

PT Consulting Partners. (2005). *Web enabled expert systems*. Retrieved December 11, 2007, from http://ptcpartners.com/Expert.htm

Riva, A., Bellazzi, R., & Montani, S. (1998). A knowledge-based Web server as a development environment for Web-based knowledge servers. IEE Colloquium on Web-based Knowledge Servers (Digest No.1998/307). London.

Sedbrook, T. A. (1998). A collaborative fuzzy expert system for the Web. *Data Base for Advances in Information Systems, 29*(3), 19-30.

Turban, E., & Aronson, J.E. (1998). *Decision support systems and intelligent systems* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Wong, B. K. (1996). The role of top management in the development of expert systems. *Journal of Systems Management, 47*(4), 36-40.

Yoon, Y., Guimaraes, T., & O'Neal, Q. (1995). Exploring the factors associated with expert systems success. *MIS Quarterly, 19*(1), 83-106.

KEY TERMS

Artificial Intelligence (AI): The study on human intelligence and on how to simulate human intelligence via machines, such as computers.

Decision Support Systems (DSS): An interactive computer-based information system to support decision-making activities by utilising databases and model bases.

Expert Systems (ES): A computer-based system that uses human knowledge captured in a computer to solve problems that ordinarily require human expertise.

Inference Engine: A computer program that tries to derive answers from a knowledge base.

Knowledge-Based Systems (KBS): A computer system that is programmed to imitate human problem-solving by means of artificial intelligence. Its core components are the knowledge base and the inference mechanisms. An expert system can be considered as one type of KBS.

Knowledge Engineering: An entire process of developing and maintaining artificial intelligent systems. The major activities in the process include knowledge acquisition, validation, representation, inferencing, explanation, and maintenance.

Web-Based Expert Systems: An expert system developed and distributed using Internet technologies.

Web-Based Personal Digital Library

Sheng-Uei Guan
National University of Singapore, Singapore

INTRODUCTION

Memex became an influential ideal and was hailed as the inspiration for hypertext and other new ways for information retrieval and organization. To deal with the explosion of scientific information, Bush's proposal for Memex focused on the problems of "locating relevant information in the published record and recording how that information was intellectually connected" (Bush, 1945). The expansion of information will be more and more serious as the Internet grows and most future computers are network-based. Thus, researchers have to consider the problems of "locating relevant published records in the published machines and recording how those records are intellectually connected".

This article proposes a mechanism that provides a new service paradigm for a network-based personal computer to browse, search, retrieve, organise, share, and publish information on the Web. Information in a machine should be made shareable with other machines, and other machines could be part of a machine's "memory extender".

As collections of relevant information either for public use or for personal use in a Web-based personal computer look like a digital library to its owner, this mechanism is called a Personal Digital Library (PDL). PDL will realise the selections of books, PDLs, and related persons through intellectual associations. As a server, PDL can store information simply and efficiently for easy retrieval and search, and provide intelligent supports for users (clients) to browse and find information. On the other hand, as a client, PDL can concurrently browse and automatically retrieve information from different PDLs (servers), and find a related person to communicate with during browsing. In addition, other PDLs can serve as memory extensions to overcome storage limitation in the local computer and avoid information duplication on the Web. PDL has potential uses in many areas such as personal use, education, commerce, finance, and entertainment. Personal users can employ any PDL with an Internet connection to manipulate their distributed information from anywhere around the world.

This article introduces PDL design and prototyping. The prototype implemented shows that the PDL concept is feasible under existing technologies. Although PDL is designed to manage personal information collections in a network-based personal computer, it will be fruitful if the design or ideas presented in this article can stimulate further development in the Digital Library research.

BACKGROUND

Memex

Bush described Memex as "a sort of mechanised private file and library" and as "a device in which an individual stores his books, records, and communications, and which is mechanised so that it may be consulted with exceeding speed and flexibility". Memex would store information on microfilms, which would be kept on a user's desk. Memex would have a scanner to enable the user to input new material, and it would also allow him to create hand-written marginal notes and comments.

In addition to the establishment of individual links, Bush also wanted Memex to support the building of trails through the material in the form of a set of links that would combine information of relevance for a specific topic. Bush emphasised (1967), that the mechanisation of "selection by association" would bring about a successful personal machine that would allow a human being to "think creatively and wisely, unencumbered by unworthy tasks," and that would allow people to "face an increasingly complex existence with hope".

The proposed PDL is a network-based Memex, which will be embedded in a personal computer. Information including bookmarks in a PDL can be made shareable with other PDLs. A Web-based "memory extender" is formed by linking PDLs together. Another essential feature of PDL is to allow people to locate and browse a certain or similar resource quickly and easily through intelligent associations.

Digital Library

Although PDL is not a particular digital library (DL) project, it has very similar goals as most DLs. In terms of service and resource provision in a DL, a number of service levels should be considered (Ormes & McClure, 1997): no services or resources provided, resource provision, self-assisted services, interactive services, video-on-demand services (Lin & Guan, 1996), and knowledge-based services.

PDL still needs to focus on the Internet, resource sharing and providing access to more information. However, there are some reasons that PDL can reach the top level. First, it is fully self-assisted because any PDL uses the same way to organise its collections, and has the same user interface,

therefore, users do not need to be assisted. Second, as to interactive service, a PDL user might discuss the use of a particular book or PDL with other users, authors, or PDL owners. Third, PDL has knowledge-based services. Examples include regularly checking changes of the favourite books or PDLs and then automatically informing the owner; providing access details and statistics of each book; and automatically forwarding a query to other PDLs. Due to such similarities with digital libraries, this proposed mechanism is called the Personal Digital Library.

In the following, we introduce two DL research projects (Cousins, 1996; Fox, 2003; Frew, Freeston, Freiras, Hill, Janee, Lovette, Nideffer, Smith, & Zheng, 2000; Guan, Yu, & Yang, 1998; Hassan & Paepcke, 1997; Kamiya, Röscheisen, & Winograd, 1996; Liu, Maly, Zubair, & Nelson, 2001; Paepcke, Cousins, Hector, Hassan, Ketchpel, Röscheisen, & Winograd, 1996; Wilensky, 1995, 1996) and some recent work on personalized bookmark services (Chen, Chen, & Sun, 2002; Kanawati & Malek, 2000, 2002; Li, Vu, Agrawal, Hara, & Takano, 1999; Yamada & Nagino, 1999).

The U.C. Berkeley Digital Library Project (Wilensky, 1995)

The goal of this project is to develop technologies that support "work-centred" digital library services, oriented to address the mission of work groups. Research areas include automated indexing, intelligent retrieval and search processes; database technology to support digital library applications; new approaches to document analysis; and data compression and communication tools for remote browsing.

PDL is a work-centred digital information system because the services provided by PDL also concentrate on the user needs. PDL improves the way to organise information for easy retrieval and search. PDL users can browse and even use different types of collections (books) not only from the local PDL but also from external PDLs. In addition, PDL provides users with a universal, customisable interface to perform information organising, browsing, retrieving, searching, and publishing.

The Stanford Digital Library Project (Stanford Group, 1995)

This project focuses on integration and interoperation (Hassan & Paepck, 1997; Paepcke et al., 1996). Its research areas include information sharing and communication models (Kamiya et al., 1996), client information interfaces (Cousins, 1996), and information finding services (Guan, Yu, & Yang, 1998).

The Integrated Digital Library will create a shared environment that links everything from personal information collections, to collections found today in conventional libraries, to large data collections shared by scientists. PDL is an integrated information system designed for network-based personal computers. Information including bookmarks in a PDL can be made shareable with other PDLs. Since bookmarks become shared resources, the information collections from other PDLs can be linked together with the local collections.

PowerBookmarks (Li et al., 1999)

PowerBookmarks (Li et al., 1999) provides personalized organization and management of bookmarks by combining the database with Web technologies. It can achieve advanced query, classification, and navigation functions and classifies the bookmarks of all users.

Personal Web Map (PWM) (Yamada & Nagino, 1999)

SeiJi Yamada and Norikatsu Nagino (1999) proposed a database named Personal Web Map (PWM) and developed the Anytime-Control algorithm to let users control their own Web map construction. PWM can help users to gather relevant information in the WWW to a small database for convenient retrieval. It will be interesting to see how their work can be extended for group work.

Semantic Web Personal Agents

Subhash, Kunjithapatham, Sheshagiri, Finin, Joshi, Peng, and Cost (2002) described the semantic Web as a vision to simplify and improve knowledge reuse on the Web. It uses software agents to collect, process, and exchange information. The PDL differs from the semantic Web personal agent in the sense that the PDL is a Memex-based approach as opposed to an agent based approach.

DESCRIPTION OF PERSONAL DIGITAL LIBRARY

Architecture

Figure 1 shows a typical PDL architecture and information flow. A PDL consists of the following components:

- **Library Explorer:** An integrated tool for information organising, browsing, retrieving, publishing, and representing. It is also a control panel to manipulate other mechanisms and application tools, and can give assistance to visitors during navigation through a series of views and facilities. Thus, the Library Explorer is a service centre in a PDL.

- **Reading Area:** Books (information collections) collected for reading (using) are placed in the Reading Area together with the corresponding reading tools (application packages). The Book Browser is the tool for reading hypermedia books and has traditional browser features.
- **Communication Area:** This area contains communication and collaboration tools such as e-mail, online chat and net conferencing. PDL provides a new type of one way collaboration with a Remote Bookshelf pane.
- **Interoperability and Protocol:** PDL can be embedded in either a personal computer or a mainframe account, which can be network-based and can run a Web server. PDL is therefore platform independent.
- **Local Storage:** A repository in which information collections are made shareable and organised in a topic-based hierarchy.

Features

PDL offers a number of important features in seven key areas: information organising, information browsing, information publishing, information searching, communication and collaboration, intelligent support, and security protection. Once these features are implemented, PDL will give Internet users a completely new browsing experience on the Web. We will detail some of them in the following sections.

Information Organising

In PDL, a book is a collection of relevant information for a specific purpose and are classified by subjects, and placed into related folders. As a folder can have sub-folders, PDL is structured as a hierarchical tree. Figure 2 shows how books and folders are hierarchically organised.

In Figure 2, it can be seen that bookmarks are located in a hierarchical tree. This is an important innovation, compared to traditional bookmark management. The hierarchical tree is used to organize shared bookmarks and facilitate searching.

- **Book:** A collection of relevant information (hypermedia, application software) and is a leaf node in the PDL hierarchical tree.
- **Folder:** A container of related books and sub-folders in a PDL. Since a folder might have children, those properties, except for *Title*, *Description*, *Type*, *Address* and *Access Times*, depend on the corresponding values of its children.
- **Bookmark:** PDLs, books, folders and bookmarks can be bookmarked. In this way, the information that summarises the item can be kept. Bookmarks can also work as a reference for an item to be used in the local PDL.

Figure 1. PDL architecture

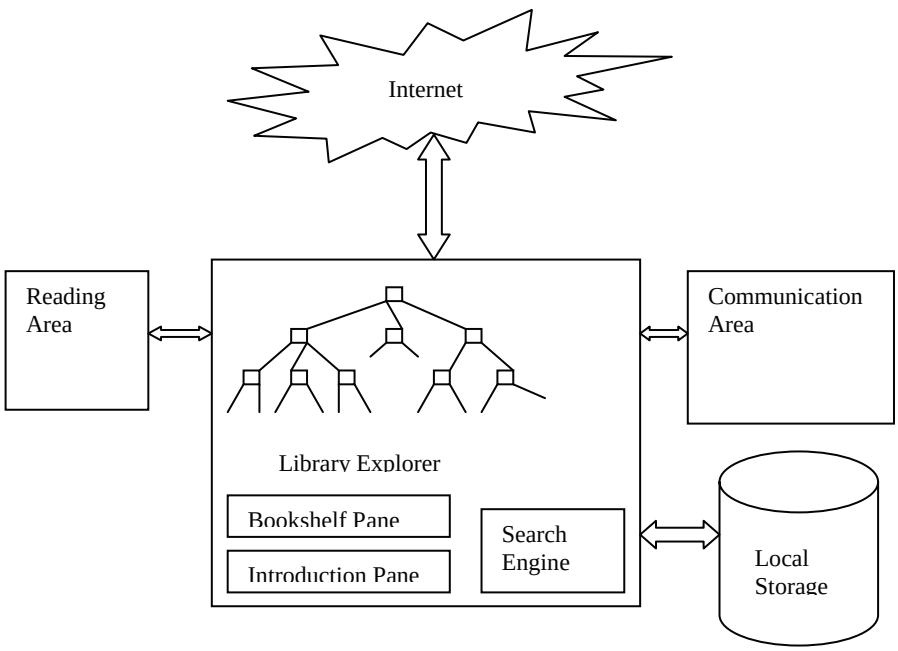
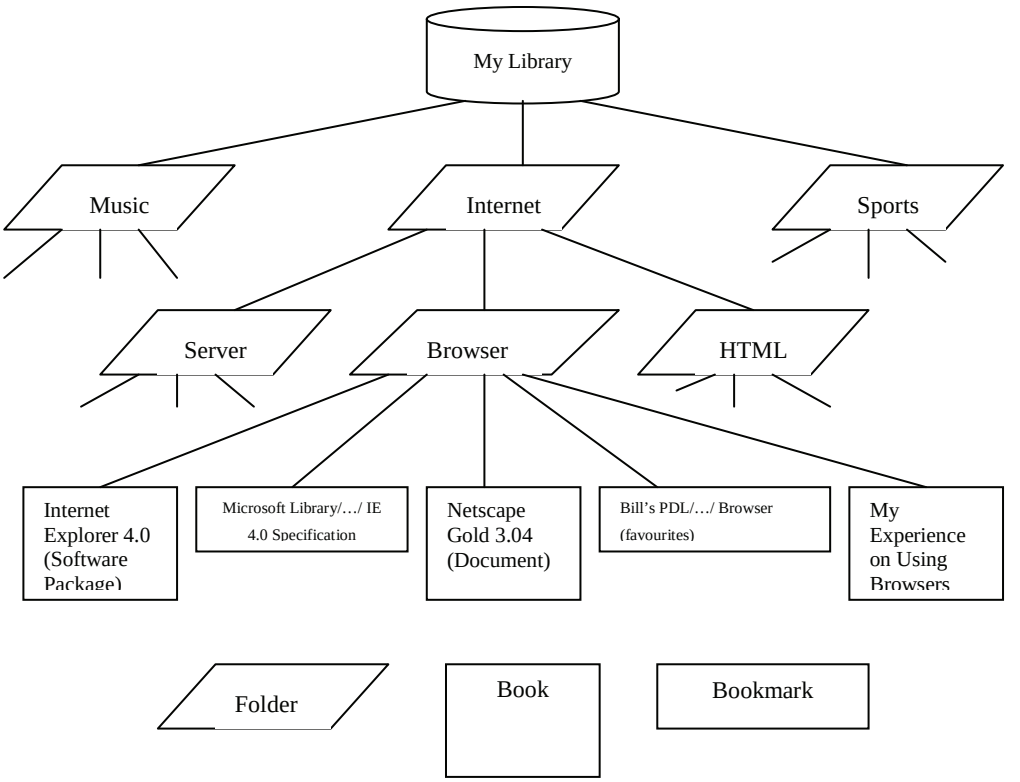


Figure 2. PDL hierarchical tree



In fact, together with the shareable, hierarchical information structure, this data structure will also bring benefits to other areas, such as information searching and intelligent support.

Information Browsing

Figure 3 shows the GUI design of the Library Explorer and the view of Figure 2. The Library Explorer provides four display panes to show information in a PDL. The Library Pane shows the currently connected PDLs and the local PDL with a tree-view structure. The Main Pane shows folders, books, and bookmarks in the currently opened folder with a list-view structure. The books that are more frequently accessed are displayed in the Introduction Pane together with the description of a selected item. The Bookshelf Pane contains the books selected from different PDLs (including the local PDL) by the PDL owner. In the Main Pane, a flag or colour is used to identify those top accessed items in the opened folder. This pane is also designed for showing a list of search results, a guest list of a selected item, and the content files of a selected book.

PDL realises true offline browsing. The PDL owner can decide whether to build a mirrored PDL for the connected PDL in the local storage during navigation or to create a cache folder for information retrieved as is done in current browsers. PDL does not assign a fixed place for cache contents. When a folder in the PDL is first opened, its corresponding cache folder is created using the same name under the "parent" cache folder in the local storage. The mirrored PDL is removed from the local storage after disconnection. This is implemented by deleting the "root" folder.

Information Searching

PDL offers title and keyword searches as in a traditional Web search engine, but both the *Title* and the *Description* fields of each related item will be searched in a keyword search process with more matched results. Figure 4 shows this unique interface linking the PDL search mechanism. The first three boxes decide what type of item is to be found. The others tell the search engine how to search.

Figure 5 illustrates how PDL handles a query either from the external PDL or the local PDL.

Intelligent Support

PDL offers a series of intelligent support during navigation. PDL provides an area (i.e., Introduction Pane) to inform

Figure 3. GUI design of the Library Explorer

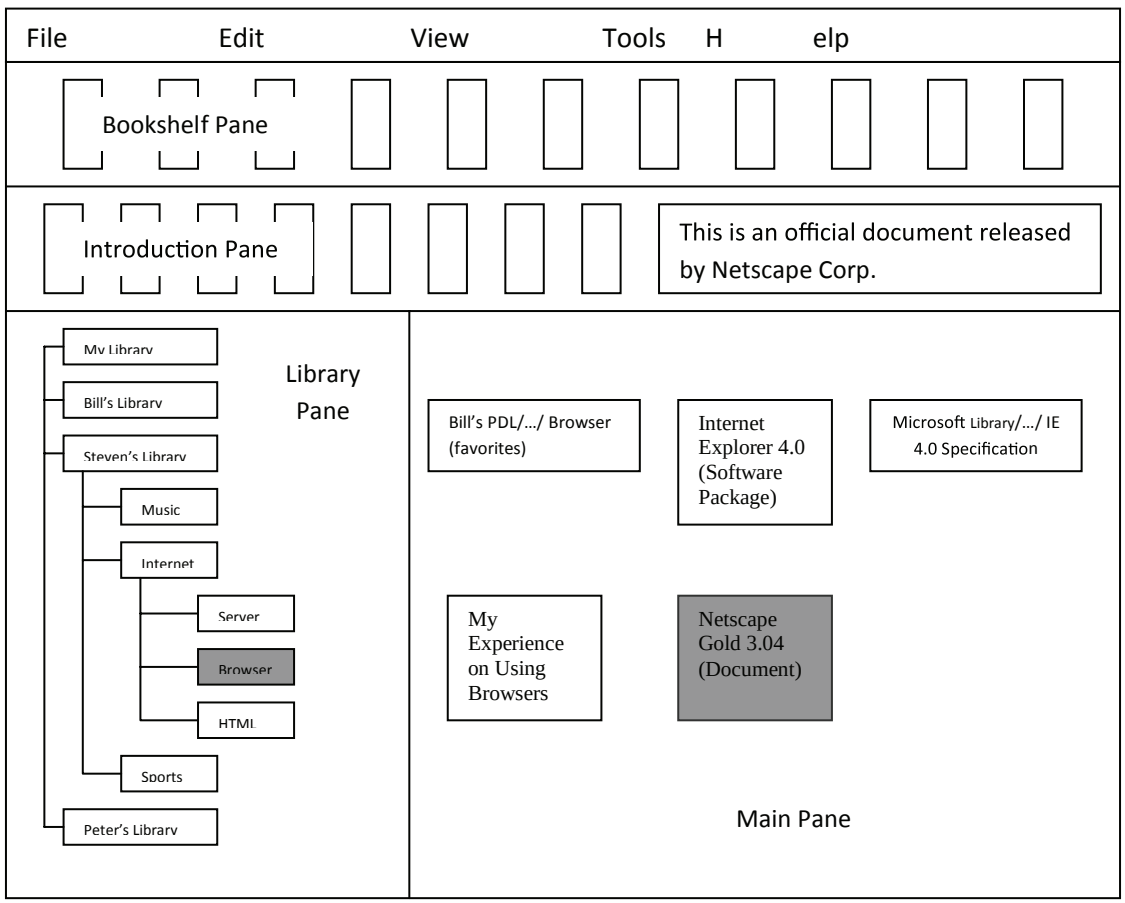


Figure 4. PDL search engine GUI

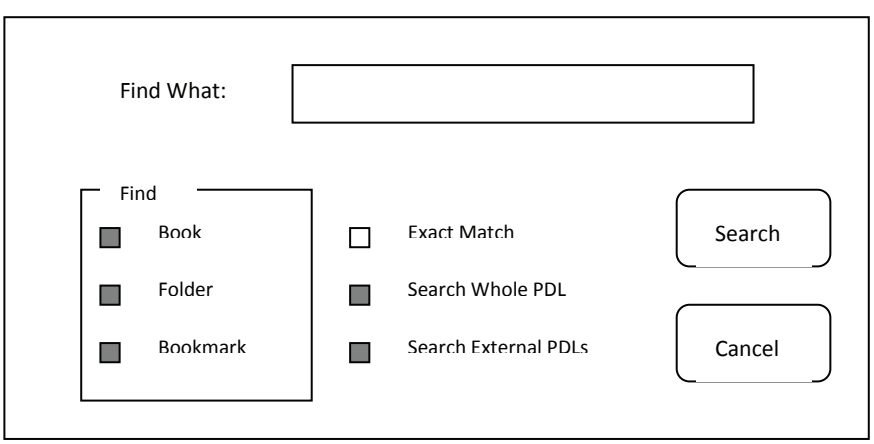
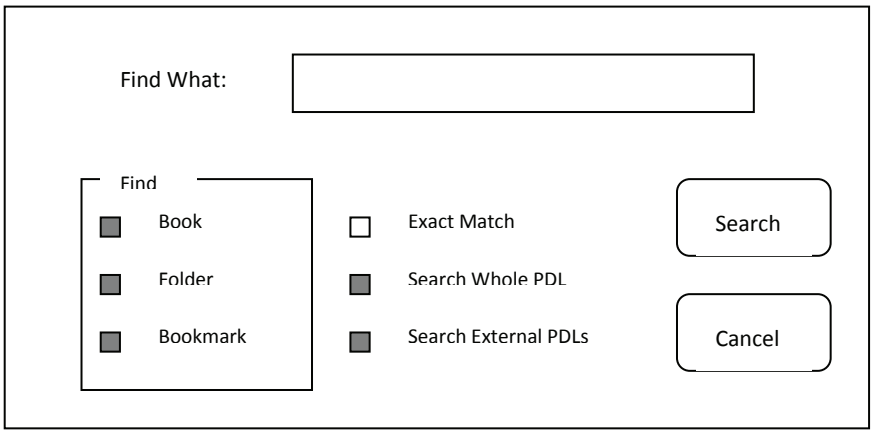


Figure 5. PDL processing query



visitors which books are more frequently visited by other people.

PDL is able to inform the owner automatically and accurately when those bookmarked items have been changed. In addition, PDL provides two options for keeping bookmarks updated and consistent: it background checks regularly and checks when there is little or no workload.

PDL also provides efficient subscription for bookmarked hypermedia books. Figure 6 illustrates the process to subscribe to a book regularly. PDL starts this subscription process regularly in the background for each bookmark marked by the PDL owner and places the subscribed books in a special folder. The PDL owner can go to this folder and read those books when desired.

Implementation

A runtime PDL prototype has been developed under Windows 95 using Microsoft Visual Basic 5.0 (Microsoft Corporation, 1998). To demonstrate that PDL can be available across platforms, a PDL is set up manually in Sun OS (Release 4.1.1) Unix machine running a Web server. By using this PDL as a server role, the communication between two PDLs can be tested. Any Web server can be used at this stage. The current selection is Apache 1.3.

FUTURE TRENDS

Research on PDL is still in a primary stage. Several further research and implementation issues include remote application use, remote management, the development of a search protocol, security, error recovery, and a 3D user interface.

Figure 6. Book subscription process

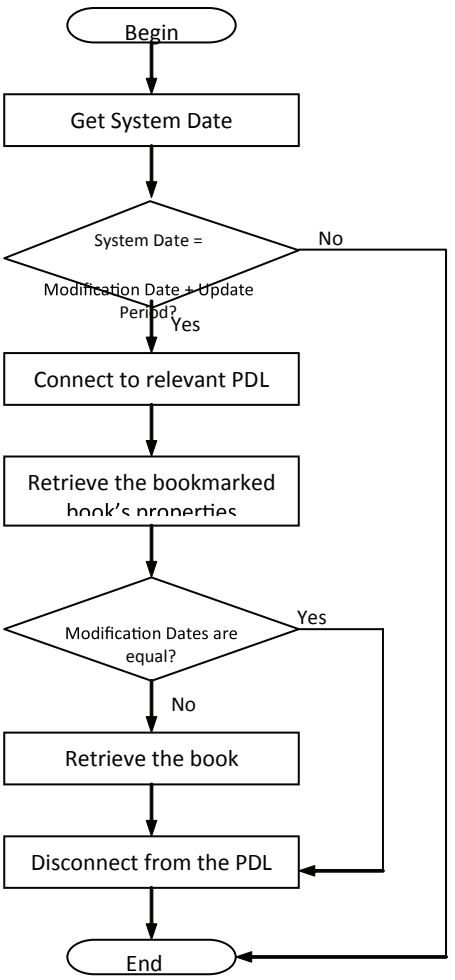
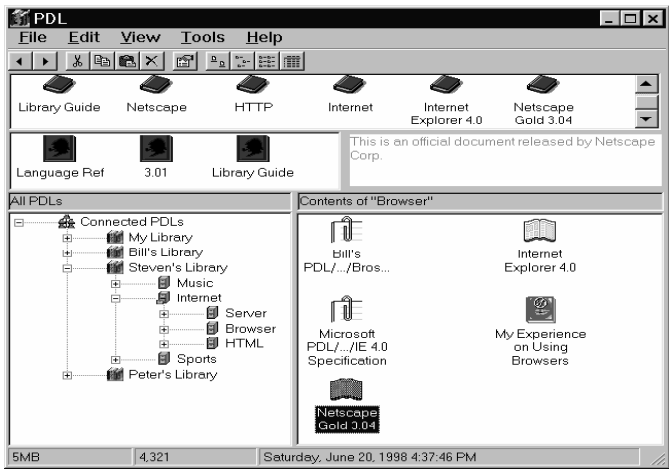


Figure 7. A Library Explorer for navigation



CONCLUSION

As a mechanism for providing a new service paradigm on the Web, the Personal Digital Library goes further to address the "trails" problems by hierarchically organising information and also making it shareable. By linking PDLs together, a Web-based "memory extender" is formed.

Although the PDL architecture still inherits hierarchical information organisation from a traditional file system, innovations are made to facilitate information sharing. One important feature is to organise and classify bookmarks along with other shareable information rather than place them into a separate folder or a file. In addition, two facilities are designed for recording "trails". These are the Library Pane, which shows the currently connected PDLs and the current navigating "trails" in each PDL, and the Bookshelf Pane, which shows a book list for the reading "trail".

On the other hand, PDL owners can quickly do a search not only in the local PDL but also in external PDLs. As the same search engine is embedded in every PDL, the search interface, the query, and the result format will be uniform. In addition, a query can be forwarded to other PDLs by the searched PDL if no results are available.

In conclusion, PDL is a Memex-like entity enhanced with the ability to connect to other PDLs and interact with their users. A fully implemented PDL can fulfil Bush's vision in more aspects than Memex. Therefore, this Memex-like entity will become a successful personal machine that allows a human being to "think creatively and wisely, unencumbered by unworthy tasks," and that allows people to "face an increasingly complex existence with hope" on the Internet.

REFERENCES

Bush, V. (1945). As we may think. *Atlantic Monthly, 176*(1), 101-108.

Bush, V. (1967). *Memex revisited, science is not enough*. William Morrow and Co.

Chen, C. C., Chen, M. C., & Sun, Y. P.V. A. (2002). A self adaptive personal view agent. *Journal of Intelligent Information Systems, 18*(2-3), 173-194.

Cousins, S. B. (1996). A task-oriented interface to a digital library. *CHI 96 Conference Companion*, 103-104.

Fox, E. A. (2003). Case studies in the U.S. National Digital Library: DL in a box, CTIDEL, and OCKHAM. *Lecture Notes in Computer Science*, 7-25.

Frew, J., Freeston, M., Freiras, N., Hill, L., Janee, G., Lovette, K., et al. (2000). The Alexandra Digital Library Architecture. *International Journal on Digital Libraries, 2*(4), 259-268.

Gravano, L., Hector, G. M., & Tomasic, A. (1994). The effectiveness of GlOSS for the text database discovery problem. In *Proceedings of the 1994 ACM SIGMOD Conference* (pp. 126-137).

Guan, S. U., Yu, H. Y., & Yang, J. S. (1998). A prioritized petri net model and its application in distributed multimedia systems. *IEEE Transactions on Computers, 47*(4), 477-481.

Hassan, S. W., & Paepcke, A. (1997). *Stanford Digital Library Interoperability Protocol.* Technical Report SIDL-WP-1997-

0054, Stanford University. Retrieved from http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1997-0054

Kamiya, K., Röscheisen, M., & Winograd, T. (1996). Grassroots, a system providing a uniform framework for communicating, structuring, sharing information, and organising people. In *Proceedings of the Fifth International WWW Conference* (pp. 1157-74).

Kanawati, R., & Malek, M. (2000). Informing the design of shared bookmark systems. *RIAO 2000* (pp. 170-180).

Kanawati, R., & Malek, M. (2002). A multiagent system for collaborative bookmarking. *International Joint Conference on Autonomous Agents and Multiagent Systems* (pp. 1137-1138).

Li, W. S., Vu, Q., Agrawal, D., Hara, Y., & Takano, H. (1999). PowerBookmarks: A system for personalizable Web information organization, sharing and management. *Computer Networks, 31*(11-16), 1375-1389.

Lin, C. L., & Guan, S. U. (1996). The design and architecture of a video library system. *IEEE Communications (Featured Topic Issue on Enterprise Networking), 34*(1), 86-119.

Liu, X., Maly, K., Zubair, M., & Nelson, M. L. (2001). Arc - An OAI service provider for Digital Library Federation. *D-Lib Magazine, 7*(4).

Microsoft Corporation. (1998). *Microsoft Visual Basic 5.0.* Retrieved from http://msdn.microsoft.com/vbasic/

NSF/DARPA/NASA Digital Libraries Initiative Projects. Retrieved from http://www.cise.nsf.gov/iis/dli_home.html

Ormes, S., & McClure, C. R. (1997). A comparison of public library Internet connectivity in the USA and UK. In S. Ormes & L. Dempsey (Eds.), *The Internet, networking and the public library* (pp. 24-40). Library Association Publishing.

Paepcke, A., Cousins, S. B., Hector, G. M., Hassan, S. W., Ketchpel, S. P., Röscheisen, M., et al. (1996). Using distributed objects for digital library interoperability. *IEEE Computer, 29*(5), 61-68.

Stanford Digital Libraries Group. (1995). The Stanford Digital Library Project. *Communications of the ACM, 38*(4), 59-60.

Subhash, K., Kunjithapatham, A., Sheshagiri, M., Finin, T., Joshi, A., Peng, Y., et al. (2002). *A personal agent application for the Semantic Web.* AAAI Fall Symposium Series.

Wilensky, R. (1995). The U.C. Berkeley Digital Library Project. *Communications of the ACM, 38*(4), 60.

Wilensky, R. (1996). Toward work-centred digital information services. *IEEE Computer, 29*(5), 37-44.

Yamada, S., & Nagino, N. (1999). Constructing a personal Web map with anytime-control of Web robots. *CoopIS'99 Proceedings. IFCIS Int Conf on Cooperative Information Systems* (pp. 140-147).

KEY TERMS

Database Technology: Organization of a collection of data in digital format.

Digital Library: Like a traditional library, a collection of books and reference materials.

Distributed Information: Information distributed across and cross accessible from several machines.

Hypertext: User interface paradigm allowing display of documents which branch or perform on request.

Information Retrieval (IR): The art and science of searching for information in documents for text, sound, images, or data.

Integrated Digital Library: Enterprise resource planning for a digital library system.

Internet: The Internet, or simply the Net, is the publicly available worldwide system of interconnected computer networks that transmit data by packet switching using a standardized Internet Protocol (IP) and many other protocols.

Knowledge-Based Systems: Systems based on the methods and techniques of *artificial intelligence*. Their core components are the knowledge base and the inference mechanisms.

Memex: The "Memex" was a theoretical analog computer described by the scientist and engineer Vannevar Bush. It is proposed as a device linked to a library, allowing automatic cross referencing.

Personal Digital Library: A personal database of documents where a significant portion of resources are in machine readable format.

Personalized Bookmark: Extended bookmarks for the PDL with scope for sharing and distribution.

Search Engine: A program designed to help find information stored on a computer system such as the World Wide Web or a personal computer.

World Wide Web: The World Wide Web ("WWW", or simply "Web") is an information space in which the items of interest, referred to as resources, are identified by global identifiers called Uniform Resource Identifiers (URI).

A Web-Enabled Course Partnership

Ned Kock
Texas A&M International University, USA

Gangshu Cai
Texas A&M International University, USA

INTRODUCTION

Notwithstanding fluctuations in enrollment, virtually every university in the U.S. and overseas has seen a significant increase in demand for information technology (IT) courses and programs in the last 15 years (Greenspan, 1999; Monaghan, 1998; Ross, 1998). At the source of this demand is an ever-growing need for qualified IT professionals in most companies, whether the companies are in technology industries or not (Alexander, 1999; Andel, 1999; Kock, 2005; Lee, 2006; Trunk, 2000; Wilde, 1999).

Given the previous practical motivation, one would expect university IT courses to be closely aligned with the industry's basic needs. Nevertheless, the gap between industry and academia in the field of IT (King, 1998; Kock, 2006; Kock, Auspitz, & King, 2002; Richter, 1999) seems to be widening rather than contracting, which is evidenced by some symptoms: (a) students complaining about their lack of "real world" IT experience when they graduate; (b) industry representatives pointing out that universities do not prepare students for the challenges and complexity of corporate IT management; and (c) faculty teaching topics that are related to their research yet far removed from the daily reality faced by IT professionals.

One way of addressing the problematic situation just mentioned is to establish industry-university partnerships. Such partnerships, particularly those involving research universities, have been commonplace for quite some time, and are arguably on the rise (Burnham, 1997; Wheaton, 1998). Irrespective of economic sector or industry, the vast majority of industry-university partnerships are of the *research partnership* type, which predominantly involves applied firm-specific research. In this type of partnership, funding from the industry partner is received in exchange for "intellectual horsepower" in the form of research services and technology transfer (Hollingsworth, 1998; Meyer-Krahmer, 1998).

A much less common type of industry-university partnership is what we refer here to as a *course partnership*, which gravitates around a regular university course (or set of courses) rather than a research project or program. In these types of partnerships, the industry partner agrees to sponsor one or more courses in which the students are expected to apply concepts and theory learned in class to the solution of some of the industry partner's key problems. Students benefit from the direct contact with the industry they are likely to join after they graduate as well as professional relationships they are able to establish during the course.

This article discusses a *course partnership* involving a large engineering and professional services company, and a public university, both headquartered in Philadelphia,. An action research study of the course partnership is used as a basis.

Like typical action research studies (Checkland, 1991; Lau, 1997; Peters & Robinson, 1984; Winter, 1989; Wood-Harper, 1985), ours aimed at providing a service to the research clients (Jonsonn, 1991; Rapoport, 1970; Sommer, 1994), while at the same time performing an exploratory investigation of the effect of Web-based collaboration technologies on course partnerships. The research clients in question were the students and the industry partner. Also, in line with a subclass of action research, namely participatory action research (Greenwood, Whyte, & Harkavy, 1993; Elden & Chisholm, 1993; McTaggart, 1991; Whyte, 1991), one of the research clients, the industry partner, participated actively in the compilation and analysis of the exploratory research data, as well as in the interpretation of the findings.

BACKGROUND

Our study was centered on a different and arguably promising approach to implementing course partnerships that was recently proposed to address the problems outlined previously (Kock, Auspitz, & King, 2000, 2002, 2003). The approach involves conducting certain courses, particularly senior undergraduate and graduate courses, in close partnership with companies. Such courses are designed so that the concepts and theory discussed in class are applied in team course projects geared at solving immediate problems at the company partner. Other fundamental characteristics of these course partnerships are:

- **All Team Projects are Conducted in One Single Organization:** Letting student teams identify organizations they would want to work with, based on criteria defined by the instructor, usually leads to dif-

ferent student teams conducting projects in different organizations, and thus to significant discrepancies in project complexity, project scope, and organizational support across different student teams. These problems can have a negative impact on learning and are considerably reduced when all team projects are conducted in one single organization.

- **Potential Projects are Identified in Advance:** The identification of a potential project by student teams can take up to five weeks of a 14-week course. One may argue that this is acceptable, as long as concepts and theory are covered in the classroom during those initial five weeks. However, in addition to identifying a project, a student team also needs to learn about the organizational culture, key people, and specific business processes they will be dealing with. This can easily take up another five weeks, leaving little time for other key project activities (e.g., business process redesign and IT implementation). The solution to this problem is to identify potential projects in advance, prior to the formal start of the course, and distribute them among student teams in the first week of the course.
- **Top Management Personally Sponsors the Course Partnership:** Often, when students are asked to come up with their own company-sponsored course projects, the individuals who sponsor the projects are not senior managers. As a result, a project sponsor may be reluctant or lack the authority to approve organizational changes or purchases of hardware and software necessary for a project to be effectively completed. These difficulties are mitigated when top management directly sponsors team projects.

It is important to note that course partnerships with these characteristics require a considerable amount of extra time and effort from the students and instructor, well beyond what is usually expected in traditional courses. In addition to applying the concepts and theory learned in class, students would also have to learn "on-the-fly" how to effectively deal with issues that are found in real-world projects (e.g., organizational culture and politics). The instructor, on the other hand, has to also take on project management, industry-university liaison, and inter-organizational team facilitation responsibilities, in addition to traditional course delivery and student mentoring duties.

IMPLEMENTING A COURSE PARTNERSHIP: TEMPLE UNIVERSITY AND DAY & ZIMMERMANN, INC.

The course partnership idea discussed previously has been implemented through a collaboration between Temple University, a large public university located approximately two miles from downtown Philadelphia, and Day & Zimmermann, Inc., a U.S. $1.5 billion engineering and professional services company headquartered in downtown Philadelphia. The course was a pilot version of CIS650—Process Design and Information Technology, a newly developed course in Temple's Computer and Information Science Department dealing with process analysis and redesign issues.

The course instructor (the author of this article) initiated the course partnership by sending a letter to one of the senior executives at Day & Zimmermann. In the letter, the course instructor inquired if the company would be interested in partnering with Temple University, providing details about the partnership. The partnership was approved after an initial meeting involving the course instructor and senior managers at the company.

The course project required students to analyze and redesign five of Day & Zimmermann's business processes using the concepts, theory, and techniques taught in class. The course partnership and related project had direct support from Day & Zimmermann's chief information officer (CIO) from the outset. A senior manager at Day & Zimmermann, who reported directly to the CIO, was assigned the responsibility of managing the project together with the course instructor. The project involved, directly and indirectly, over 30 Day & Zimmermann employees and 26 Temple students.

The students were split into five process redesign teams, which periodically met with key Day & Zimmermann employees at the company's headquarters in downtown Philadelphia. Each team analyzed and redesigned one process, generated three reports, and delivered an oral presentation to Day & Zimmermann management at the end of the course. The first report generated by each team contained a detailed description of the process targeted; the second a detailed description of the redesigned process and the rationale behind the redesign decisions; and the third a detailed analysis of IT solutions to enable the new (redesigned) process.

WEB SITE REMOVES OBSTACLES TO PARTICIPATION

Before the course partnership was started, two main obstacles had to be dealt with. First, Day & Zimmermann employees were expected to actively participate in the process redesign efforts. In order to do so, they had to understand the concepts and theory used by the students. Yet, most of the Day & Zimmermann employees likely to be involved in this project could not come to Temple to audit the course together with the students. Also, given that Temple students and Day & Zimmermann employees were not co-located, a great deal of their interaction would have to occur by means other than face-to-face meetings. The solution to overcome these two obstacles was the development of a password-protected

Table 1. Course evaluation scores

Question/statement	Score
The objectives and requirements of the course were made clear.	3.82
The instructor clearly communicated the subject.	3.88
The instructor was open to questions and comments.	3.82
The instructor was accessible outside the classroom.	3.70
The instructor was impartial in evaluating my performance.	3.47
The instructor expected academic excellence from students.	3.47
Overall, the instructor did an excellent job teaching.	3.82
Overall, I have learned a great deal from this course.	3.41
Overall, this is one of the best courses I have had at Temple.	3.53

Note: Score = average score for the class regarding the question; Range: 0-4)

Web site, which allowed Day & Zimmermann employees to access all the course material online. The Web site also supported interaction between them and Temple students through shared document areas, multimedia components, and discussion boards.

WAS THE COURSE PARTNERSHIP SUCCESSFUL?

The partnership was considered a success by Day & Zimmermann management and employees, as well as by Temple students. Managers emphasized the anytime/anyplace collaboration between Day & Zimmermann employees and Temple students enabled by the Web site as one of the key elements that made the course partnership a very successful collaborative effort.

Temple students emphasized the real-world experience as one of the most positive aspects of the course. Following is a representative comment by a student extracted from one of the anonymous course evaluation forms completed at the end of the course: "The learning experience was very rich. The group project gave us hands on experience in applying the redesign techniques we learned in the course. It was a great experience to work with upper level IT management!"

Table 1 shows the average scores for several question/statements asked from the students in connection with the course. The question/statements were part of a standard course and faculty evaluation, and were answered anonymously.

Several students pointed out that the course required considerably more time and effort from them than most traditional university courses they had taken before. In spite of that, their anonymous evaluations of the course were very positive, as it can be seen in Table 1. The average answer to the question/statement "Overall, this is one of the best courses I have had at Temple" was 3.53, on a 0-to-4 scale. The average answer to the question/statement "Overall, I have learned a great deal from this course" was 3.41, also on a 0-to-4 scale.

An added benefit for Day & Zimmermann was the ability to identify young talent based on observation of business-relevant action (as opposed to the traditional analysis of résumés). Day & Zimmermann was able to observe a group of 26 students in action over a two-month period and identify several students whom they would like to consider hiring. This is not as easy to accomplish with other approaches for identifying new graduates for hiring, of which internships are perhaps most popular. There are two key reasons for this. First, the number of interns that could be hired for a two-month period by an organization would typically be considerably smaller, thus significantly reducing the number of students that Day & Zimmermann managers would be able to observe in action during that period of time. Second, the tasks that the interns would be assigned to would not usually be nearly as complex and strategically relevant as those carried out in this course would.

CONCLUSION AND LESSONS FOR FUTURE PARTNERSHIPS

The general perception at the end of the course partnership was that it had been an exciting and rewarding experience for all those involved. Students saw the course as a valuable experience that provided them with a unique view of IT management and which complemented the concepts, theory, and techniques learned in the course and throughout their university program. Day & Zimmermann managers perceived the input provided by the students as very valuable and likely to lead to concrete business process improvements.

Also, a few lessons have been learned along the way that can be useful for universities and companies planning to implement similar course partnerships in the future. These lessons are summarized next.

- **The Course Partnership Should have Two Main Co-pProject Managers, One From Academia and One fromIndustry:** As with most inter-organizational initiatives, the scope of management authority does not usually extend beyond organizational boundaries. Students will respond more quickly to requests by the instructor than to requests by a manager of the partner organization. Similarly, employees of the partner organization will respond more quickly to a request by someone who has formal authority within the organization than to a request by the instructor. Therefore, the instructor should share the responsibility of managing the project with a member of the partner organization who has enough formal authority to oversee all the team projects.
- **The Course Partnership Should Include at Least One Purely Social Activity:** Social activities allow for important information exchanges that would not normally occur in formal meetings. For example, prior to the final oral presentation by student teams at Day & Zimmermann, a pizza party (paid for by the company) was held in downtown Philadelphia. After the party, several people remarked that the personal knowledge they learned from informal conversations during the party was invaluable to them. The party was also seen as a "thank you" gesture by the partner organization to the students. Not only did this boost morale, but it also helped the students relax for the presentation the next day, as they got to know the people they would be presenting to at a more personal level.
- **The Business Problems Addressed through the Course Partnership Should be "Real" and of High Relevance to the Partner Organization:** Because students are involved, not professional consultants, the partner organization may be tempted to create "toy" problems to be solved through the course partnership, rather than address real and relevant business problems. The motivation for this may be the perceived risks linked to not accomplishing the goals of the project (e.g., wasted time, internal conflicts, and reluctance to engage in future organizational change efforts). The problem with this approach is that it is likely to relieve the students from any real responsibility and, at the same time, decrease the motivation for employees to get involved. A better alternative to reduce risk is to involve experienced consultants at critical points in the course partnership (the costs are relatively low since few consultant-hours are likely to be used).
- **Partner Organization Employees Should be Asked to Report on and Make an Oral Presentation of their Projects, too:** Course partnerships such as the one described here are, as the name implies, collaborative endeavors in which members of the two organizations involved should contribute evenly. Therefore, key members of the partner organization should also be asked to make presentations about their projects. This is particularly important because, in some cases, despite full dedication from the students, a project may fail to accomplish its initial goals. And, a closer examination may indicate that the failure was not the students' fault, but that it was caused by lack of interest or commitment from the part of one or more employees. Employee reporting and presentations are useful in assessing whether this is the case, which in turn is important for appropriately grading the students' coursework. Moreover, the requirement to report and present their projects communicates to the partner organization employees that they are equally responsible for the outcomes of the project, which is likely to increase their level of commitment and dedication to the project.

The course partnership discussed in this article went well beyond the idea of having students conduct their course projects in real organizations, which is an approach adopted by many project-based courses around the world. It seems that the close cooperation between Temple University and Day & Zimmermann that characterized the course partnership presented here was the key reason for its success. This type of cooperation requires extra time and effort from everyone involved—students, instructor, company management, and employees. Yet, the positive (tangible and intangible) outcomes of the partnership seem to easily outweigh its costs.

REFERENCES

Alexander, S. (1999). High demand for hot skills. *Computerworld, 33*(39), 4-6.

Andel, T. (1999). IT is your business. *Material Handling Engineering, 54*(7), 18.

Burnham, J. B. (1997). Evaluating industry-university research linkages. *Research Technology Management, 40*(1), 52-56.

Carnes, K. C., & Gierlasinski, N. J. (1999). Have you considered a faculty intern? *The National Public Accountant, 44*(3), 31-32.

Checkland, P. (1991). From framework through experience to learning: The essential nature of action research. In H. Nissen, H. K. Klein, & R. Hirschheim (Eds.), *Information systems research: Contemporary approaches and emergent traditions* (pp. 397-403). New York: North-Holland.

Elden, M., & Chisholm, R. F. (1993). Emerging varieties of action research. *Human Relations, 46*(2), 121-141.

Greenspan, A. (1999). The interaction of education and economic change. *The Region, 13*(1), 6-11.

Greenwood, D. J., Whyte, W. F., & Harkavy, I. (1993). Participatory action research as a process and as a goal. *Human Relations, 46*(2), 175-191.

Hollingsworth, P. (1998). Economic reality drives industry-university alliances. *Food Technology, 52*(7), 58-62.

Jonsonn, S. (1991). Action research. In H. Nissen, H. K. Klein, & R. Hirschheim (Eds.), *Information systems research: Contemporary approaches and emergent traditions* (pp. 371-396). New York: North-Holland.

King, J. (1998). Labor confab issues call for training. *Computerworld, 32*(2), 1, 16.

Kock, N. (2005). *Business process improvement through e-collaboration: Knowledge sharing through the use of virtual groups*. Hershey, PA: Idea Group Publishing.

Kock, N. (2006). *Systems analysis and design fundamentals: A business process redesign approach*. Thousand Oaks, CA: Sage Publications.

Kock, N., Auspitz, C., & King, B. (2000). Using the Web to enable industry-university collaboration: An action research study of a course partnership. *Informing Science (Special issue on Organizational Learning), 3*(3), 157-167.

Kock, N., Auspitz, C., & King, B. (2002). Bridging the industry-university gap: An action research study of a Web-enabled course partnership. In E. Cohen (Ed.), *Challenges of information technology education in the 21st century* (pp. 166-186). Hershey, PA: Idea Group Publishing.

Kock, N., Auspitz, C., & King, B. (2003). Web-supported course partnerships: Bringing industry and academia together. *Communications of the ACM, 46*(9), 179-183.

Kock, N., Gray, P., Hoving, R., Klein, H., Myers, M., & Rockart, J. (2002). IS research relevance revisited: Subtle accomplishment, unfulfilled promise, or serial hypocrisy? *Communications of the AIS, 8*(23), 330-346.

Lau, F. (1997). A review on the use of action research in information systems studies. In A. S. Lee, J. Liebenau, & J. I. DeGross (Eds.), *Information systems and qualitative research* (pp. 31-68). London: Chapman & Hall.

Lee, K. S. (2006). The IT hiring forecast for 2006. *Certification Magazine, 8*(2), 16.

McTaggart, R. (1991). Principles for participatory action research. *Adult Education Quarterly, 41*(3), 168-187.

Meyer-Krahmer, F. (1998). Science-based technologies: University-industry interactions in four fields. *Research Policy, 27*(8), 835-852.

Monaghan, P. (1998). Growing demand for computer animators spurs a new program at U. of Washington. *The Chronicle of Higher Education, 44*(48), 23-35.

Peters, M., & Robinson, V. (1984). The origins and status of action research. *The Journal of Applied Behavioral Science, 20*(2), 113-124.

Rapoport, R. N. (1970). Three dilemmas in action research. *Human Relations, 23*(6), 499-513.

Richter, A. (1999, November 7). Silicon island, wired but underpopulated. *New York Times*, Sec. 14LI, 1.

Ross, P. E. (1998). Enjoy it while it lasts. *Forbes, 162*(2), 206.

Sommer, R. (1994). Serving two masters. *The Journal of Consumer Affairs, 28*(1), 170-187.

Trunk, C. (2000). Information technology in logistics: Material flow at moen. *Material Handling Management, 55*(1), 8-10.

Wheaton, Q. (1998). Government-University-Industry Cooperation: Does it Work?, *Quality*, 37(5), 20-24.

Whyte, W. F. (Ed.). (1991). *Participatory action research*. Newbury Park, CA: Sage.

Wilde, C. (1999). Hiring in triplicate. *Computerworld, 33*(28), 77.

Winter, R. (1989). *Learning from experience: Principles and practice in action-research*. New York: The Falmer Press.

Wood-Harper, A. T. (1985). Research methods in information systems: Using action research. In E. Mumford, R. Hirschheim, G. Fitzgerald, & A. T. Wood-Harper (Eds.), *Research methods in information systems* (pp. 169-191). Amsterdam, The Netherlands: North-Holland.

KEY TERMS

Action Research: Type of research approach in which the researcher attempts to improve the research client, which can be an organization, while at the same time generating relevant academic knowledge.

Co-Partnership Managers: The course instructor, on the university side, and a senior manager of the client organization, who jointly manage a course partnership project.

Course Partnership: Course-based industry-university partnerships, where a course is designed so that the concepts and theory discussed in class are applied in team course projects geared at solving immediate problems at the company partner.

Industry-University Gap: Disconnect between the knowledge and skill needs of industry practitioners and the knowledge and skills imparted on students by universities.

Process: Set of interrelated activities through which an organization transforms inputs into value-added outputs. Inputs and outputs can be tangible (e.g., materials, parts) or intangible (e.g., services, information) items.

Process Redesign: Approach to organizational improvement through transformation of business processes. The term refers to business process change approaches emphasizing gradual change (e.g., total quality management) or radical change (e.g., business process reengineering).

Research Partnership: Industry-university partnership involving collaboration in applied firm-specific research. In this type of partnership, funding from the industry partner is received in exchange for "intellectual horsepower" in the form of research services and technology transfer.

Web-Based Course Partnership: A course partnership in which a Web site is developed to serve as a central communication hub and document repository for the partnership.

A Web–Geographical Information System to Support Territorial Data Integration

V. De Antonellis
Universitá di Brescia, Italy

G. Pozzi
Politecnico di Milano, Italy

F.A. Schreiber
Politecnico di Milano, Italy

L. Tanca
Politecnico di Milano, Italy

L. Tosi
Comune di Milano, Italy

INTRODUCTION

The design of a Web-geographical information system, Web-GIS (Worboys & Duckham, 2004; Zhong Ren & Ming Hsiang, 2003), strongly requires methodological and operational tools for dealing with information distributed in multiple, autonomous and heterogeneous data sources, and a uniform data publishing methodology and policy over Internet Web sites. In this article we describe the experience of the Politecnico di Milano group in the activities of requirement analysis and conceptual design of the DEAFIN Web-GIS (Schreiber et al., 2003), whose objective is to provide a common environment for comparison of information about available vacant industrial sites coming from different regional data sources. Heterogeneity and Web availability requirements have been taken into account in the system architecture design; the system is thus conceived as a federated Web-based information system, apt to manage and provide access to all the regional relevant information in an integrated and complete fashion. Furthermore, since the data available by a given region partner can be both spatial and alphanumeric, a Web-GIS is defined for each regional component system.

BACKGROUND

The DEAFIN (development agencies and their impact on foreign direct investments) project has been launched with the purpose of allowing companies and investors to get a comprehensive information framework about areas located in European regions suited for potential investments. The aim is to make the regional data about possible investment areas homogenous and comparable, and internationally accessible. Potential investors need both a survey and a detailed view of vacant sites in different locations in order to compare different opportunities and decide their convenience. Quite naturally, such requirements call for a federated information system (FIS), which grants local sites a great deal of autonomy while enabling interoperation by means of a global integrated conceptual schema, that is, the federated data schema. Furthermore, owing to the capillarity of the end-user locations and to the need of a simple and widely known interface, Web-based access is mandatory. To define the functional specification of the system, the following activities have been carried out:

- *analysis of the requirements* of a distributed Web-based information system relying on a common conceptual database schema of the regional information that was initially (almost completely) available on paper support;
- *conceptual design* of the DEAFIN FIS, centered on the conceptual design of the federated conceptual database schema. The regional databases must be built according to the federated schema and then made accessible via the Web. The availability of data on the Web allows potential investors to navigate in the DEAFIN site according to various and customizable

criteria, based on a benchmarking model developed within the project.

INFORMATION REQUIREMENTS

Three regional administrations from Germany, Italy, and Sweden were involved. The project started with a data-gathering phase, aimed at collecting requirements about data and processes managed at the partner Public Administrations. A questionnaire-based tool was circulated to collect common information to be stored in the FIS.

The basis of the questionnaire is a set of the data categories managed in Public Administration information systems. The relevant data categories concern land use plans (master and regional or specific), territorial services, industrial vacant sites, mobility data, statistical and social-economic data, base cartography data, and information on cadastral units data. Information on vacant industrial sites is the main focus of the investigation. For each category, the questionnaire collected the data characteristics reported in Table 1.

In general, the data collected at the sites show uniformity with respect to the attention paid to cartographic availability, regulations and laws about reuse of vacant areas, and centralization of resources. In particular, the need exists at each regional site to introduce tools able to treat heterogeneous data, since these data more and more intensively are to come from various data sources, to be mapped into the federated schema. Also, the degree of automation is similar, since cartographic systems and basic data management tools are available at the three partners' sites.

Several ongoing local projects concern the digital acquisition of land use plans, the automation of document management, and the development of various thematic databases and Web sites. What is required is a unified common schema for the regional databases. Moreover, the need of a uniform data publishing methodology and policy over Internet Web sites emerges clearly from the participants and involves almost all data categories.

Table 1. Summary of data characteristics collected in the data-gathering phase

DATA CHARACTERISTICS
Availability of the data category
Location of the data source
Support technology and name of products/tools
Data format
Data owner
User roles involved in data access
Restrictions applying to data access
Performances
Maintenance policies
Availability on the Web

USAGE MODES

The data-gathering phase has also detailed a set of user profiles, which specify how different access demands can be supported by the system towards a variety of user groups. The design and implementation of profiles have obviously a deep impact on the usability of the system. Hence, a careful analysis of user typologies and profiles has been performed during the specification phase, while an enrichment of profiles and access modes has been planned in a post-implementation follow-up. The purpose is to have the system start with a set of pre-configured access typologies, and then tune the access modes and user profiles against the most typical uses observed for a fixed period on the DEAFIN pilot implementation. The first broad difference is between the profile of public administrations and that of private users, due to different data needs. Moreover, two basic interaction modes must be provided: browsing (using thematic areas and other refined search parameters) and querying (using simple and user-friendly interfaces). The system can be regarded as: a passive tool, when responding to user questions; an active tool, when used as a decision support system or when used as a standard application, allowing new European partners to join the DEAFIN consortium.

As a passive tool, the contents of the federated database can show the advantages and disadvantages of an area: the information provided can be related to the specific search needs of a specific user. These searches are different according to user types, but apart from the function of locating (or re-locating) business activities, the motivation probably exists to retrieve general information about a region.

SYSTEM ARCHITECTURE

In the literature, several approaches and tools for handling heterogeneous data sources have been developed, and standards for distributed information systems have been defined (Mylopoulos & Papazoglou, 1997; Wiederhold, 1992). For these systems, the use of multiple layer, mediator-based architectures, and of a common data model have been

employed (Garcia Molina et al., 1997). Wrapper/extractor and mediator tools (see Key Terms section) are proposed to obtain a uniform data representation (abstracting from the formats of the original data sources) and to facilitate federated access. Following this direction, in order to handle heterogeneity at each regional site, an architecture has been defined where extractors translate data coming from local heterogeneous data sources to a common reference format defined at the mediator level (Schreiber et al., 2003).

The aspect of data publishing over Internet Web sites has emerged clearly as a need from the partner regions. The system is thus conceived as a Web-based federated information system, apt to manage and provide access to all the regional relevant information in an integrated and complete fashion, thus satisfying the Web availability need.

Generally, data made available by a given partner region can be both spatial and alphanumeric, requiring each regional component system to be a GIS. Each regional site is in charge of data publishing and certification over the system according to the common federated schema. Each regional Web site contains a specific interface and stores regional data in a database by means of a Web-GIS technology, and operates as a central server that contains the common interface and some aggregate data. The global system is composed of Web-interconnected sites: the central Web site is the reference site for aggregate data and for requirements about new regional sites to be added in the federation or new applications to be developed in the component systems.

FEDERATED DATABASE CONCEPTUAL DESIGN

The federated database conceptual schema provides an integrated high-level representation of all data to be handled by the FIS. In order to achieve the highest readability and the widest possible diffusion, a common set of terms is needed. We adopted the dictionary of terms as defined by the EuroDicAutom (EuroDicAutom, 2001) of the EU; the international system of measurements, and commonly agreed upon data categories were adopted.

The database conceptual design produced a set of conceptual views, represented in terms of ER diagrams (see Key Terms section), that provide schematic requirement representations, each related to a specific information set. The following views were defined: vacant site description view, land use plan view, administration view, eco/reclamation view, transportation view, service view, document view, procedure and task view; the global conceptual data schema is a merge of these different views.

The design identified 36 entities (including entities with spatial representation that describes their shape, extension and location on the earth surface, marked as georeferenced), with a total number of about 200 attributes. The XML (eXtensible Markup Language) language was chosen for logical specification, providing a common reference format to facilitate data customization, translation and mediation.

The FIS should enable the user to dynamically query the databases and in particular the geographical data they contain; these requirements, as we discussed previously, are fulfilled by a Web-GIS system able to perform advanced operations guided by the client side.

THE WEB-GIS

The interface of the central Web site allows the users to access data in two search modes: direct search, which leads directly to the vacant area of interest, including also information about the surroundings; and navigation, which presents a list of vacant sites to be filtered progressively, according to various criteria. Indeed, a static benchmarking table presents a list of site features, along with the corresponding values for each partner: from this table the user can choose a partner and obtain pages that describe the partner's area, highlighting its most attractive features. Some aggregate data about the vacant sites of each region are also displayed for subsequent browsing operations, where the user can execute queries in order to filter the vacant sites on the basis of preferences or specific criteria, such as: the type of usage, area size, costs, surrounding services, and accessibility. Finally, links to the regional Web sites are provided.

The Web interfaces of regional sites support more specific functions: the exploration of the partner area by choosing a specific theme such as land planning, business and market activities, mobility system, services, demographic and macroeconomic data and development strategies; the access to business information such as contact offices and particular business opportunities that are currently available for some vacant sites of the area; the access to aggregated data about vacant sites of the area and to a benchmarking table where the user can apply some selection conditions based on the available comparison parameters; the search

of vacant sites based on forms that guide the user in the specification of the query, including geography related conditions (SQL with geographic extensions) like "search the vacant sites with a railway station within a radius of <parameter value> km"; the selection of specific features of the chosen vacant site, for example, buildings, eco quality, themes presenting the characteristics technological networks, land planning, mobility system, territorial services, procedures and tasks and documents.

FUTURE TRENDS

Web-GIS applications are becoming more and more important to a growing number of activities. Geography, geology, environmental studies, business marketing, and other disciplines have gained benefits from GIS tools and methods. Continual improvements in GIS hardware and software will lead to a much wider application of this technology throughout government, business, and industry. In particular, integrated geodata services based on data format standardization will increasingly facilitate the exchange of information among users of different systems by allowing data sharing and improving cooperation and communication among the organizations involved in environmental protection, planning, and resource management.

CONCLUSION

The development of the DEAFIN Federated Information System started from an analysis of the requirements and of the locally available information, performed through a data-gathering phase. The federated database was modeled in the form of a set of conceptual views, both as ER diagrams and XML specifications. The architecture of the overall system is in the form of a federated Web-GIS: at the main Web site a high-level view over the entire federated system is allowed, while local specializations are allowed at the level of every regional Web site adhering to the recommendations for the entire database.

The proposed system offers a marketing and information platform in a European context based on advanced Web functions, integrating geographic information systems (GIS), decision support tools, and database federation features. Analogous projects in the European area have been developed, mainly in the field of tourism (Pühretmair & Wöß, 2001), where integration of data sources and access to GIS in a graphical context are needed.

Further details about the project can be found in Schreiber et al. (2003).

REFERENCES

EuroDicAutom. (2001). Automatic translation of terms, European community. *http://eurodic.ip.lu:8086/cgi-bin/edicbin/EuroDicWWW.pl?page= expert*

Garcia Molina, H., Papakonstantinou, Y., Quass, D., Rajaraman, A., Sagiv, Y., Ullman, J., Vassolos, V., & Widom, J. (1997). The TSIMMIS approach to mediation: Data models and languages. *Journal of Intelligent Information Systems, 8*, 117-132.

Mylopoulos, J., & Papazoglou, M. (Eds.). (1997). Cooperative information system *IEEE Expert* [Special issue], *12*(5).

OpenGIS Consortium. (2001, February 20). *Geography Markup Language (GML).* OGC Document Number: 01-029.

OpenGIS Consortium. (2001, June 21). *Web map service implementation specification.* OGC Document Number: 01-047r2.

Pühretmair, F., & Wöß, W. (2001). XML-based integration of GIS and heterogeneous tourism information. In K. Dittrich, A. Geppert & M. Norrie (Eds.), *Advanced information systems engineering* (pp. 346-358). Berlin: LNCS Springer Verlag.

Schreiber, F.A., Belussi, A., De Antonellis, V., Fugini, M.G., Pozzi, G., Tanca, L., & Tosi, L. (2003). The design of the DEAFIN Web-geographical information system: An experience in the integration of territorial reclamation support services. In A. Dahanayake & W. Gerhardt (Eds.), *Web-enabled systems integration: Practice and challenges* (pp. 142-168). Hershey, PA: Idea Group Publishing.

Wiederhold, G. (1992). Mediators in the architecture of future information systems. *IEEE Computer, 25*, 38-49.

Worboys, M.F., & Duckham, M. (2004). *GIS: A computing perspective* (2nd ed.). Boca Raton: CRC Press

The World-Wide Web Consortium. *http://www.w3.org*

Zhong Ren, P., & Ming Hsiang, T. (2003). *Internet GIS.* New York: John Wiley and Sons Inc.

KEY TERMS

Conceptual Schema of a Database: A semi-formal high-level description of the database, independent of its implementation.

ER (Entity-Relationship) Diagrams: The most widely used model to express the database conceptual schema.

eXtensible Markup Language (XML): Markup language proposed by the World Wide Web Consortium [W3C] for data and documents interchange.

Federated Information System (FIS): An information system is named federated when it supports interoperation among several autonomous and possibly heterogeneous information systems, by means of a shared global data schema.

Geographical Information System (GIS): Information system storing geographical data along with alphanumeric and spatial components. GIS systems also provide the data structures and algorithms to represent and efficiently query a collection of geographical data.

Mediator: A software component providing a uniform integrated interface to process and execute queries over data stored in multiple, heterogeneous data sources.

Web-GIS: A GIS system empowered with a Web-based interface.

Wrapper: A software tool to extract content from data sources and perform data format translation.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 33-37, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Wireless Ad Hoc Networking

Fazli Erbas
University of Hanover, Germany

INTRODUCTION

Mobile ad hoc networks represent a new form of communication consisting of mobile wireless terminals (e.g., handset, PDAs, notebooks). These type of networks are wireless multi-hop packet networks without any fixed infrastructure and centralized administration, in contrast to today's wireless communications, which is based on a fixed, pre-established infrastructure. The design of wireless ad hoc networks faces many unique challenges. In this article, mobile ad hoc networks and their characteristics are described, and the design issues, applications and future trends of such networks will be discussed.

BACKGROUND

In recent years, widespread availability of wireless communication and handheld devices stimulated the research and development on self-organizing networks that do not require a pre-established infrastructure and any centralized architecture. Those spontaneous networks provide mobile users with ubiquitous communication capability and information access regardless of their location. This type of networking is called mobile ad hoc networks.

The idea of mobile ad hoc networks has been under development from the 1970s and 1980s in the framework of Mobile Packet Radio Technology (PRNET-1973) (Jubin & Tornow, 1987) and Survivable Adaptive Networks (SURAN-1983) (Schacham & Westcott, 1987). These projects supported research on the development of automatic call set up and maintenance in packet radio networks with moderate mobility. However, interest in this area grew rapidly due to the popularity of a large number of portable digital devices, such as laptop and palmtop computers, and the common availability of wireless communication devices.

In the middle of the 1990s, with the definition of standards, commercial radio technologies have begun to appear and the wireless research community identified in ad hoc networks a challenging evolution of wireless networks. The success of a network technology is associated with the development of networking products that can provide wireless network access at a competitive price. A major factor in achieving this goal is the availability of appropriate networking standards. Today's emerging standards and technologies for constructing a mobile ad hoc network are IEEE 802.11, Bluetooth and ETSI Hiperlan/2. The deployment of mobile ad hoc networks opens a wide-range of potential utilisation from military to miscellaneous commercial, private and industrial scenarios (Perkins, 2001).

MOBILE AD HOC NETWORKS AND THEIR CHARACTERISTICS

A mobile ad hoc network (MANET) consists of a collection of mobile wireless and autonomous hosts –in this sense simply referred to as "*nodes*"– which spontaneously form a temporary network. The devices may be of various types (e.g., notebook computers, PDAs, cell phones, etc.) and various capacities (e.g., computing power, memory, disk, etc.).

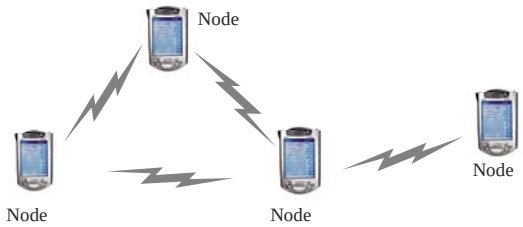
The most important characteristic of such a network is its independence of any fixed infrastructure (e.g., base station or access point) or centralized administration. All networking functions, such as determining the network topology, multiple access, and routing of data over the most appropriate paths, must be performed in a distributed way. These tasks are particularly challenging due to the limited communication bandwidth available in the wireless channel.

Actually, the idea of ad hoc networking is sometimes also called infrastructureless networking (Frodigh, Johansson & Larsson, 2000). An ad hoc network is able to operate autonomously and is completely self-organized and self-configured. Therefore, it can be easily and rapidly installed. In an ad hoc environment people and vehicles can be interworked in areas without a pre-existing communication infrastructure, or when the use of such infrastructure requires wireless extension.

Autonomous nodes may move arbitrarily so that the topology changes frequently without any prior notice. The wireless transmission range of the nodes is also limited; therefore the connection (e.g., wireless link) between the neighboring nodes may break as soon as they move out of range. Consequently, topology of the network and the interconnection patterns among nodes may change dynamically so that links between nodes become unusable. Because of the dynamic nature of ad hoc networks, new routes must be considered and maintained using routing protocols.

Another important property of ad hoc networks is the multi-hop capability. It is given that cellular networks — also called single-hop networks — rely on a fixed wired infrastructure to achieve the task of routing and maintain the

Figure 1. A mobile ad hoc network



connection end-to-end. On the contrary, a mobile node in an ad hoc network that cannot reach the destination directly, because it does not lie within its radio transmission range, will need to relay its information flow through other nodes. This implies the mobile hosts to incorporate routing functionality so that they can act both as routers and hosts.

Other important characteristics of MANET include (Perkins, 2001):

- **Dynamic topologies:** Nodes are free to move arbitrarily; thus, the network topology may change randomly and rapidly at unpredictable times.
- **Bandwith-constrained links:** Caused by the limits of the air interface. Furthermore, multiple access, multipath fading, noise and signal interference decrease the limited capacity available at the allocated frequency rate.
- **Energy-constrained operation:** MANETs inherently imply an underlying reliance on portable, finite power sources.
- **Limited security:** Mobile networks are in general more vulnerable to eavesdropping, spoofing and denial-of-service attacks than fixed-cable networks.

DESIGNING ISSUES

Physical Layer and MAC Layer

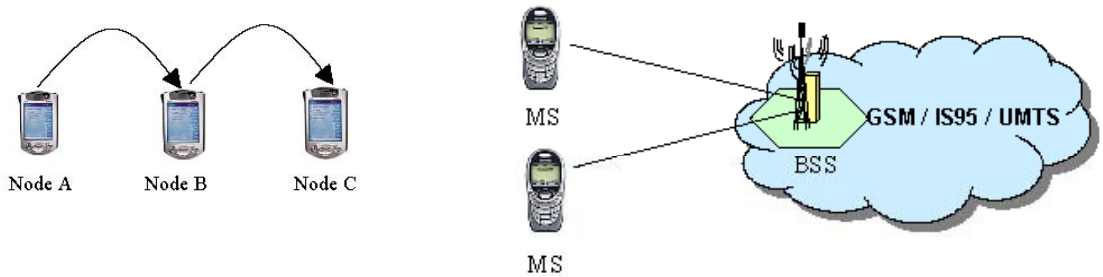
A well-designed architecture for mobile ad hoc networks involves all networking layers, ranging from the physical layer to the application layer. Information as node distribution, network density, link failures, and etcetera, must be shared among layers, and the MAC (medium access control) layer and the network layer need to collaborate in order to have a better view of the network topology and to optimise the number of messages in the network (Bruno, Conti, & Gregori, 2001; Kurose, Schwartz, & Yemini, 2000).

The main aspects of designing the physical transmission system are dependent on several characteristics of the radio propagation channel such as path loss, interference and fading. In addition, since mobile terminals usually have limited power resources, the transceiver must be power efficient. These aspects are taken into account while designing the modulation, coding, and power control features in the radio equipment. In principle, the radio equipment in the nodes forming a mobile ad hoc network can use any technology as long as it provides reliable links between neighboring mobile terminals on a common channel. Candidate physical layers that have gained prominence are infrared and spread spectrum radio techniques.

The MAC (medium access control) layer plays the key role in determining the channel usage efficiency by resolving contention amongst a number of unsupervised terminals sharing the common channel. An efficient MAC protocol allows coordinated access to the limited resources. The main goal of a MAC protocol is therefore maximizing the probability of successful transmissions and maintaining fairness amongst all users.

Though research on medium access schemes for wired local area networks (LANs) have been done for many years, the same concepts cannot be directly applied to wireless

Figure 2. Mobile ad hoc networks (multi-hop networks) in comparison to today's cellular (single-hop) networks



(Node C is reached from node A via node B in multihop way)

LANs. In a wired medium, the transmitted signals are almost received with the same signal strength at all terminals connected to the same shared medium. Hence a terminal in a LAN can avoid contention by sensing the presence of a carrier to determine if any other terminal is using the channel before it starts a transmission.

However, designing MAC protocols for wireless networks faces a different set of challenges. Propagation path losses in the wireless channel cause the signal power to decline with distance.

Since the strength of the received signal depends on the distance from the transmitter, the same signal is not heard equally well by all terminals. Hence carrier sensing is not very effective in wireless. Typical problems of using carrier sensing to determine the availability of the wireless channel are *hidden node problem* and *exposed node problem* (Nasipuri, 2002). Both the hidden node and the exposed node problems arise due to the fact that carrier sensing is only performed at the transmitter, whereas its effect is determined by the interference power at the receiver, which are usually different due to propagation path loss characteristics.

In order to address the hidden terminal problem, IEEE 802.11 has the option of adding the mechanism of an exchange of *REQUEST TO SEND (RTS)* and *CLEAR TO SEND (CTS)* control packets between a transmitting and receiving nodes before initiating the transmission of a data packet.

Several concerns with the IEEE 802.11 MAC has motivated researchers to explore newer techniques to improve the channel utilization and throughput in mobile ad hoc networks. The basic access method of the IEEE 802.11 MAC protocol is susceptible to inefficiencies due to the hidden and exposed terminal problems. The RTS/CTS option reduces the hidden terminal problem but not the inefficiency caused by the exposed terminal problem.

A wireless transceiver cannot transmit and receive at the same time as the transmitted signal will always be far stronger than any received signal. Hence, a wireless terminal cannot detect if its transmission has been successful. To inform the transmitting node about the successful packet transmission, the receiver sends an ACK (acknowledgement) packet back to the transmitter after it receives a data packet. If the transmitter does not receive an ACK within a fixed period of time, it assumes that the transmitted packet has been lost. Many different schemes have been designed for reducing these problems in wireless channel access.

ROUTING

Movements of nodes in a mobile ad hoc network cause the nodes to move in and out of range from one another. Consequently, topology of the network and the link and connection patterns between nodes may change dynamically so that links between nodes become unusable. As depicted, in contrast to conventional wireless networks, mobile ad hoc networks have no fixed network infrastructure or centralized administrative support for their operations. Because of the dynamic nature of ad hoc networks, new routes must be considered and maintained using routing protocols. Since the network relies on multi-hop transmissions for communication, this imposes major challenges for the network layer to determine the multi-hop route over which data packets can be transmitted between a given pair of source and destination nodes.

Because of this time-varying nature of the topology of mobile ad hoc networks, traditional routing techniques, such as the shortest-path and link-state protocols that are used in fixed networks, cannot be directly applied to ad hoc networks. A fundamental quality of routing protocols for ad hoc networks is that they must dynamically adapt to variations of the network topology. This is implemented by devising techniques for efficiently tracking changes in the network topology and rediscovering new routes when older ones are broken. Since an ad hoc network is infrastructureless, these operations are to be performed in a distributed fashion with the collective cooperation of all nodes in the network (Royer & Toh, 1999; Perkins, 2001). Some of the desirable qualities of dynamic routing protocols for ad hoc networks are:

- **Routing overhead:** Tracking changes of the network topology requires exchange of control packets amongst the mobile nodes. These control packets must carry various types of information, such as node identities, neighbor lists, distance metrics, and etcetera, which consume additional bandwidth for transmission. Since wireless channel bandwidth is at a premium, it is desirable that the routing protocol minimizes the number and size of control packets for tracking the variations of the network.
- **Path optimality:** With constraints on the routing overhead, routing protocols for mobile ad hoc networks are more concerned with avoiding interruptions of communication between source and destination nodes rather than the optimality of the routes. Hence, in order to avoid excess transmission of control packets, the network may be allowed to operate with suboptimal (which are not necessarily the shortest) routes until they break. However, a good routing protocol should minimize overhead as well as the path lengths. Otherwise, it will lead to excessive transmission delays and wastage of power.
- **Loop freedom:** Since the routes are maintained in a distributed fashion, the possibility of loops within a route is a serious concern. The routing protocol must incorporate special features so that the routes remain free of loops.

- **Complexity:** Another problem of distributed routing architectures is the amount of storage space utilized for routing. Ad hoc networks may be applied to small portable devices, such as PDA and handhelds, which are memory and hardware scarce. Hence, it is desirable that the routing protocol be designed to require low storage complexity.
- **Scalability:** Routing protocols should be able to function efficiently even if the size of the network becomes large. This is not very easy to achieve, as determining an unknown route between a pair of mobile nodes becomes more costly in terms of the required time, number of operations, and expended bandwidth when the number of nodes increases.

Because of its many challenges, routing has been a primary focus of researchers in mobile ad hoc networks. The MANET working group within the IETF (IETF, 2004) studies solutions for routing framework and develops IP-based routing protocols for mobile ad hoc networks. (Macker & Corson, 1998; Corson & Macker, 1999). Consequently, a large number of dynamic routing protocols applicable to mobile ad hoc networks have been developed.

APPLICATIONS

The term ad hoc means "as needed for a specific case," "done or set up solely in response to a particular situation or problem without considering wider issues." Both definitions stress out the urgent, specific and short-term need character. Indeed, an ad hoc network will be deployed in an area where a support for mobile communication is not available due to low expected usage and high costs. The network can disappear after its function has been served.

One of the original motivations for the development of this technology lay in satisfying military needs like battlefield survivability. Soldiers must be able to move about freely without any restrictions imposed by wired communication devices. Moreover, the military cannot rely on access to a fixed preplaced communication infrastructure especially during manoeuvres in enemy territories (Perkins, 2001).

Other potential practical utilities of mobile ad hoc networks could include:

- Commercial and industrial scenarios: Associates sharing information during a meeting, participants in a conference exchanging documents or presentations
- Educational scenarios: Students using laptop computers to participate in an interactive lecture
- Emergency coordination in disaster areas, where a hurricane or earthquake have destroyed the communication infrastructure
- Extension of the coverage area of cellular networks
- Communication between smart household appliances in the home environment
- Communication between "wearable" computing devices
- Inter-vehicle communication like within a convoy or in highway
- Sensor Networks: represent a special kind of ad hoc networks. They typically consist of nodes equipped with sensing, processing and communication abilities. Sensor networks are used to monitor remote or inhospitable physical environments

FUTURE TRENDS

Need for wireless ad hoc networking will arises in the context of shared desktop meetings (e.g., cooperate learning, workshops, conferences), disaster recovery, wireless inter-vehicle communication, proprietary local networks for private, commercial, business and governmental, use (universities, hospitals, home environments, office environments) or in various military applications.

New technological standards, regulations (e.g., IEEE 802.11, IEEE 802.15), emerging products (e.g., PDAs, notebooks) and new business models will enable fast deployment of such networking concepts.

CONCLUSION

Wireless ad hoc networking is a new form of communication offering new emerging technologies and applications. Several issues have to be considered which are arising due to characteristics of this type of networks. The critical issues discussed in this article offer many implications and challenges to technology, business and the user community as well.

The successful deployment and a broad acceptance of such networks will depend on the availability of appropriate standards and regulations. Further, the development of new products and business models at a competitive price will play key role.

REFERENCES

Bruno, R., Conti, M., & Gregori, E. (2001). WLAN technologies for mobile ad hoc networks. In *Proceedings of the 34th Hawaii International Conference on System Sciences*.

Corson, S., & Macker, J. (1999, January). *Mobile Ad Hoc Networking* (MANET). IETF RFC 2501. Retrieved from *www.ietf.org/rfc/rfc2501.txt.*

Frodigh, M., Johansson, P., & Larsson, P. (2000). *Wireless ad hoc networking: The art of networking without a network*. Ericsson Review No. 4. Retrieved from *www.ericsson.com/about/publications/review/2000_04/files/2000046.pdf*

Giardano, S. (2000). Mobile ad-hoc networks. In I. Stojmenovic (ed.), *Handbook of Wireless Networks and Mobile Computing*. New York: Wiley (Imprint) Inc.

IETF MANET Working Group. (2004). Retrieved from *http://www.ietf. org/html.charters/manet-charter.html*

Jubin, J., & Tornow, J.D. (1987). The DARPA packet radio network protocols. *Proceedings of the IEEE*, *75* (1), 21–32.

Kurose, J.F., Schwartz, M., & Yemini, Y. (2000). Multiple access protocols and time constraint communications. *ACM Computing Surveys*, *16*, 43-70.

Macker, J., & Corson, S. (1998). Mobile ad hoc networking and the IETF. *ACM Mobile Computing and Communications Review*, *2* (1).

Asis, N. (2004). Mobile ad hoc networks. In Farid Dowla (ed.), *Handbook of RF and Wireless*. Retrieved from *http://www.ece.uncc.edu/~anasipur*

Perkins, C. E. (2001). *Ad hoc networking*. Addison Wesley

Royer, E. M., & Toh, C. K. (1999). A review of current routing protocols for ad hoc mobile wireless networks. *IEEE Personal Communications*, pp. 46–55.

Schacham, N., & Westcott, J. (1987). Future directions in packet radio architectures and protocols. *Proceedings of the IEEE*, *75* (1), 83–99.

Schiller, Jr., J. (2003). *Mobile Ccommunications*. (2nd ed.). Pearson Education.

KEY TERMS

Carrier Sensing: Determination that the medium is not being used by a neighboring transmitter before accessing the channel.

Cellular Networks: A network consisting of several cells served by fixed, pre-established infrastructure to cover a geographic area, for example, GSM, IS-95, UMTS.

Dynamic Topology: Due to the node mobility the network topology of mobile multi-hop ad hoc networks are changing continuously in time.

Exposed Node: This is the reverse problem, where a transmitting or "exposed" node is within range of a sender, but is out of range of the intended destination.

Hidden Node: A node may be hidden or out of range from a sender but within range of its intended receiver.

MANET (Mobile Ad Hoc Networks): A network of wireless mobile nodes formed dynamically, self-organizing and without any central administration.

Multi-Hop: The mobile nodes are cooperating to forward data on behalf of one another node to reach distant stations that would otherwise have been out of range of sending node.

Node: Mobile terminal.

RTS/CTS (Request To Send/Clear To Send): Control packets between the transmitting and receiving nodes before initiating the transmission of a data packet.

Single-Hop: Direct communication between two nodes or entities without any intermediate station or network entity.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 3090-3094, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Wireless Networks for Vehicular Support

Pietro Manzoni
Technical University of Valencia, Spain

Carlos T. Calafate
Technical University of Valencia, Spain

Juan-Carlos Cano
Technical University of Valencia, Spain

Antonio Skarmeta
University of Murcia, Spain

Vittoria Gianuzzi
University of Genova, Italy

INTRODUCTION

Vehicular Ad hoc NETworks (VANETs) is an area under intensive research that promises to improve security on the road by developing an intelligent transport system (ITS). The main purpose is to create an inter-communication network among vehicles, as well as between vehicles and the supporting infrastructure. The system pretends to offer drivers data concerning other nearby vehicles, especially those within sight.

The problem of information sharing among vehicles and between the vehicle and the infrastructure is another critical aspect. A general communication infrastructure is required for the notification, storage, management, and provision of context-aware information about user travel. Ideally an integrated vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication paradigm enriched with an information management system would solve the problem. The infrastructure should manage all the collected safety events garnered from vehicles and the interesting information to be provided to the user, which is adapted to the car context and driver preferences.

Finally, security issues should be considered. Since the information conveyed over a vehicular network may affect critical decisions, fail-safe security is a necessity. The first directive for any V2V communication scheme is, therefore, that every safety message must be authenticated. Because of the high speed and therefore short duration within which communication between two cars is possible, communication must be non-interactive, and message overhead must be very low. The urgency of safety messages implies that authentication must be instantaneous without additional communication.

Moreover, providing strong security in vehicular networks raises important privacy concerns that must also be considered. Safety messages include data that is dangerous to the personal privacy of vehicle owners. Most relevant is the danger of tracking a vehicle through positional information. A set of security basics to address these challenges should be proposed that can be used as the building blocks of secure applications.

In this article we will focus on the aforementioned technologies and engineering issues related to vehicular ad-hoc networks, emphasizing the challenges that must be overcome to accomplish the desired vehicular safety infrastructure.

BACKGROUND

Ubiquitous computing is nowadays an emerging research field in mobile communications, due to more and more integration of heterogeneous services over different operation environments. The capacity of customizing services to the client, and the adaptation of its behavior according to the context, will offer the user value-added features in the new age of computer communications. Taking into account this premise, in this article our aim is to create a feasible environment for providing integrated services in the vehicle field.

Wireless communications in the vehicle field through ad hoc networks (or Vehicular Ad hoc NETworks–VANETs) are currently being used as a novel and promising technology to improve driving safety. Mainstream research usually considers these communication patterns to offer intelligent transportation systems (ITSs), where one of the most important aims is the creation of communication networks among vehicles, in vehicle-to-vehicle transmissions (V2V),

but without forgetting communications between vehicle and infrastructure (V2I). The usefulness of these developments is focused on providing every vehicle with information about the surrounding vehicles, and especially the ones not located in the field of vision.

Due to the continuous improvements of communication technologies, a great number of considerations must be taken into account when a network system is elected. Although VANET developments have predominated in V2V communications, it is necessary to study whether the facilities offered by the network design cover the requirements collected from the future deployable services. Collision avoidance applications have been the main safety service implemented for the vehicle field. However, a great number of non-safety services are appearing. When the amount of services for the vehicle side grows, more consideration is needed. It is mandatory to research in technological solutions which deal with the requirements of a generic and flexible architecture for service provisioning and usage.

For a correct design of such systems, it is necessary to take into account the vehicular environment. Here, high-speed mobility and special movement patterns can be found, where the creation and breakage of links between nodes appear continuously. The presence of an excessive or null rate of equipped vehicles is another important factor which must be considered. This fact and the need for reliable end-to-end communication make the design of a vehicular communication system a complex task.

The main objective of any project in this area is to offer an integrated solution for the deployment of an ITS using the VANET communication technology overall. The schema considered can be seen in Figure 1. The system should apply techniques inside the field of ubiquitous computing in the vehicle field, where numerous programmatic devices can interact with the user in a transparent way. This way, the system should have enough intelligence to analyze the context and efficiently detect hazardous and emergency situations, generate driver warnings in critical cases, and interact with the infrastructure when a global knowledge can be useful.

A CROSS-LAYER APPROACH TO VANET DESIGN

The design of a complete VANET solution must cover different technological areas, and several disciplines are involved in the development of such a system. In the following we will outline all the interrelated layers that must work together to provide a comprehensive architecture.

Modeling, Evaluation, and Simulation

In classical MANETs, researchers often use a typical set of simulation parameters. These parameters are inadequate for

Figure 1. Overall scheme of the proposed architecture

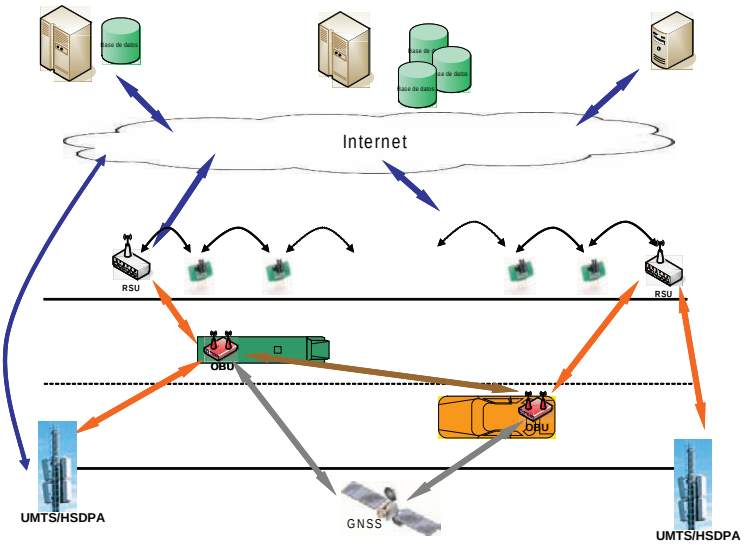
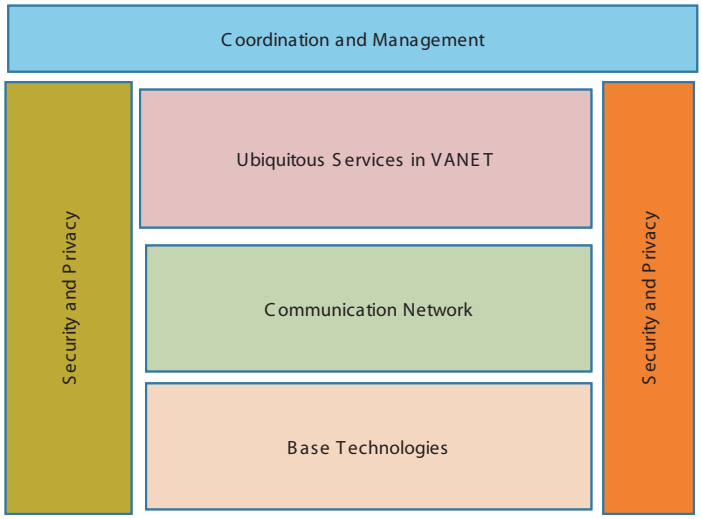


Figure 2. A VANET system vertical architecture



many MANETs, and particularly for VANETs. The MANET research community is aware of resulting limitation because of the simplification of some hypotheses (Kotz et al., 2004). In the VANET context, different researching groups are working to improve models in order to perform experiments close to reality. For example, Wu, Fujimoto, Guensler, and Hunter (2004) use their own simulator, called CORSIM, for creating mobility traces for other simulators. In the same way, Choffnes and Bustamante (2005) use a new mobility model called STRAW for modeling real traffic patterns. This model has an easy car persecution model and the chance of controlling the traffic to create congestion situations.

On-Board and Roadside Equipment

Current studies about technologies for on-board equipment are directed to the hardware that is needed to incorporate into the vehicle in order to offer one or several services with certain functionality (Skarmeta et al., 2002; Massaki, Xuchu, Hideki, Mami, & Masaki, 2004). An important issue will be the implementation of a positioning system with a high degree of reliability. Also we will pay special attention to the computer considered in the on-board unit. At the start there was a tendency to employ specific-purpose computers. Nevertheless, the development of several services advises the employment of general purpose computers (Simonds, 2003; Santa, 2007). This computer must be fitted with the communication interfaces needed to connect to the communication system that will be proposed in the project. In the car environment, although having used many communication technologies (Nolte, Hansson, & Lo Bello, 2005), the most interesting ones are the 802.11 and cellular networks. 802.11 networks have usually been considered in VANET networks and allow the deployment of cooperative services among vehicles. Nevertheless, new solutions were developed for communicating with the infrastructure (Okabe, Shizuno, & Kitamura 2005). On the other hand, UMTS is the most appropriate cellular technology for the vehicular environment. Usually, connections through the cellular network are being directed to infrastructure communications. However, in the project development, a V2V and V2I communication architecture based on cellular networks will be proposed.

A V2I communication infrastructure will need roadside support. Although in cellular networks operators offer hardware deployment, the research in new V2I communication technologies must take into account several technological considerations of the road infrastructure.

Horizontal and Vertical Handoffs Between Communication Systems

The interest in offering Internet access in heterogeneous scenarios and on mobile platforms like trains, buses, or airplanes has caused the creation of the NEtwork MObility

Working Group (NEMO) of the IETF (Internet Engineering Task Force). This group is in charge of defining mechanisms for managing network mobility as a whole, allowing network handoffs between different physical systems, but maintaining the same logical connection over IP (Internet Protocol). NEMO objectives are becoming real due to the improvement of cellular networks and the integration of WLAN technologies like WAVE (IEEE 802.11p, or Wireless Access in the Vehicular Environment) or CALM (Continuous Air interface for Long and Medium distance).

An important objective resides in offering mobile nodes an efficient and scalable mechanism for changing their network attachments, keeping their IP address, and keeping alive all the high-level connections (Yabusaki, Okagawa, & Imai 2005). In this way there are some interesting works, like MIPv6 and Hierarchical Mobile IPv6 (HMIPv6) (Yan & Atwood, 2006), with the aim of improving mobile communications in certain circumstances, making the handoff a secure and efficient process.

Security in Data Networks

Being able to protect members of possible security attacks is essential in this kind of network, without eliminating the possibility of giving context services. Attacks can be classified into passive attacks and active attacks. Passive attacks consist of "listening" to the information that is transmitted in the network, without directly affecting to the network operation (Yanchao, Wei, & Wenjing, 2005). Generally, hackers enter with the aim of examining the general state and the network traffic. This analysis can be made with the objective of damaging the network availability or traffic. The privacy about position is a special type of information confidentiality, which is defined as the possibility of stopping third parties from knowing the current or a previous vehicle position.

Active attacks (Liu, Fu, Xiao, & Lu, 2007) imply a bigger attacker interaction with the network, by means of devices manipulation, "man-in-the-middle" techniques, or denial-of-service strategies. To solve these problems it is necessary to use cryptography-based techniques, which allow authenticating both the transmitted information and the sender, for example by means of digital signatures and hash functions. The aim is providing VANET vehicles with classic services of authentication, non-repudiation of origin, and information integrity, as well as specific for VANET networks. Due to the high speed of movement, techniques for detection of so-called "false alarms," which are produced by information delays or alarms repetition, must been researched. In monitoring systems where an information management station for vehicles is situated at the infrastructure, security issues for information transmissions are also an important fact to be considered (Hoh, Gruteser, Xiong, & Alrabady, 2006).

Vehicle to Infrastructure Communications and Alternatives to VANET

Though VANET networks are the main research point in vehicular networks, there are situations where current developments can limit the use of some services. When the distance between two transmitting vehicles increases considerably, a communication based on VANET can present poor performance. On the other hand an interconnection with a fixed network requires efforts that can not be scalable in cost. With this meaning, the project will direct an effort in finding other alternatives that complement the VANET communication system.

Some authors have noticed VANET cadences and the usefulness of P2P (peer-to-peer) networks to create an architecture that allows cooperative work between vehicles (Rybicki et al., 2007) or a connection with the infrastructure (Baresi et al., 2005), this way avoiding VANET network restrictions. By means of a P2P network, nodes can use in a transparent mode an underlying network infrastructure and work cooperatively using a logical interconnection mesh. Vehicles can take advance of this system and communicate with its environment through several services in a V2V or V2I pattern. However, an underlying technology is necessary to "physically" connect the vehicle to the network.

Cellular networks were especially important in connecting with the infrastructure. Previous connection rates, of approximately 300kps, have been overcome by HSPA technology, which already offers significant improvements in delay and bandwidth to start considering UMTS as an appropriated technology for vehicular networks. For this reason, the project will cover issues regarding the use of cellular networks in vehicles and will consider an alternative network system for V2V and V2I communications based on P2P over UMTS.

FUTURE TRENDS

Vehicular networks introduce a new complex environment for communication engineers. The communication channel varies from a simple point-to-point microwave link for cars in open areas, to rich Rayleigh fading within the cities. Also the channel varies considerably every few seconds and the blockage of line of sight can occur frequently. Adaptive and efficient channel estimation algorithms must be found, diversity techniques to overcome fading effects should be examined, and Doppler effects should be carefully considered. The link layer is expected to provide a varying delay and different QoS classes to satisfy the different requirements of the applications.

Techniques of service provision which make use of vehicle position are the so-called location-based services (LBSs). The traffic environment and the information that vehicles receive require a more complete modeling about the destination of the user and his or her preferences in order to reach the so-called context-based services. The subscription strategy is of special interest, talking about a generic service provision system. Typical Internet services, like Web navigation, transactions management, bookings, and so forth, use a client/server strategy where the server replies to a previous client request, and operation requisites of vehicular services respond to other, more suitable communications strategies, known as publish/subscribe. Using this methodology, the user does not have to send an explicit request to the system to receive certain information, but neither would receive continually non-desired information, as happens with the vehicular RDS (radio data system). In a publish/subscribe scheme, the user manifests his or her interest about receiving certain information, and the system is in charge of sending pertinent notifications (Cugola & Jacobsen 2002). This way, a publish/subscribe system perfectly operates not only with a context-aware service provision, but also with users' requirements modeling.

The Ontology Web Language (OWL) and the adaptation of Web content by means of captive portals will be combined in this project with the service subscription strategy in order to present a ubiquitous environment of services for the driver. The usefulness of OWL has been proven in different environments, including aiding systems for museum visits (Chou, Hsieh, Gandon, & Sadeh 2005). These applications model the museum environment and adapt the information to the user according to his or her movement throughout the facilities. The adaptation of this idea to a vehicular environment is considered of special interest (Santa, 2007).

CONCLUSION

In this article, we presented the VANET research area, a very active area that is expected to truly enhance car drivers' lives. The main focus is security, but also infotainment (the combination of information and entertainment) is of interest, especially from the economic point of view.

We described the areas where most of the research activity is focusing and evidenced that a transversal solution is crucial for this type of network.

Fast and reliable dissemination of alerts on accidents or other road hazards is a critical application of vehicular networks. A protocol that supports the dissemination of emergency messages should ensure the timely delivery of such messages to all the vehicles passing through the potentially affected region during the lifetime of an emergency. This protocol might exploit a hybrid, cluster-based vehicular network architecture, where the bulk of communication between vehicles takes place over ad hoc networks for improved scalability and efficiency. Cellular communication can also be used to improve reliability when network partitions hinder the message delivery.

REFERENCES

Baresi, L., Ghezzi, C., Miele, A., Miraz, M., Naggi, A., & Pacifici, F. (2005). Hybrid service-oriented architectures: A case-study in the automotive domain. *Proceedings of the 5th International Workshop on Software Engineering and Middleware* (SEM'05) (pp. 62-68).

Choffnes, D.R., & Bustamante, F.E. (2005). An integrated mobility and traffic model for vehicular wireless networks. *Proceedings of ACM VANET.*

Chou, S., Hsieh, W., Gandon, F.L., & Sadeh, N.M. (2005). Semantic Web technologies for context-aware museum tour guide applications. *Proceedings of the 19th International Conference on Advanced Information Networking and Applications,* Washington.

Cugola, G., & Jacobsen, H. (2002). Using publish/subscribe middleware for mobile systems. ACM *SIGMOBILE Mobile Computing and Communications Review, 6*(4), 25-33.

Hoh, B., Gruteser, M., Xiong, H., & Alrabady, A. (2006). Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing, 5*(4), 38-46.

Kotz, D., Newport, C., Gray, R.S., Liu, J., Yuan, Y., & Adid Elliott, C. (2004). Experimental evaluation of wireless simulation assumptions. *Proceedings of ACM MSWIM.*

Liu, J., Fu, F., Xiao, J., & Lu, Y. (2007, July 30-August 1). Secure routing for mobile ad hoc networks. *Proceedings of the 8th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing* (SNPD 2007) (vol. 3, pp. 314-318).

Massaki, W., Xuchu, M., Hideki, H., Mami, M., & Masaki, S. (2004). iCAN: Pursuing technology for near-future ITS. *IEEE Intelligent Systems,* (January/February), 18-23.

Nolte, T., Hansson, H., & Lo Bello, L. (2005, July). Wireless automotive communications. *Proceedings of the Euromicro Conference on Real-Time Systems* (ECTRS'05), Palma de Mayorca, Spain.

Okabe, T., Shizuno, T., & Kitamura, T. (2005, June). Wireless LAN access network system for moving vehicles. *Proceedings of the 10th IEEE Symposium on Computers and Communications* (ISCC05), La Manga del Mar Menor, Spain.

Rybicki, J., Scheuermann, B., Kiess, W., Lochert, C., Fallahi, P., & Mauve, M. (2007, September 9-14). Challenge: Peers on wheels—a road to new traffic information systems. *Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking* (MobiCom '07) (pp. 215-221), Montréal, Canada.

Santa, J., Muñoz, A., & Skarmeta, A.F.G. (2007). A context-aware solution for personalized en-route information through a P2P agent-based architecture. *Proceedings of the 2007 International Conference on Computational Science and Its Applications* (ICCSA 2007) (pp. 710-723). Berlin: Springer-Verlag (LNCS 4707).

Santa, J., Ubeda, B., & Skarmeta, A.F.G. (2007, January). A multiplatform OSGi based architecture for developing road vehicle services. *Proceedings of the Consumer Communications & Networking Conference 2007* (CCNC 2007), Las Vegas, NV.

Simonds, C. (2003). Software for the next-generation automobile. *IT Professional*, (November/December), 7-11.

Skarmeta, A.F.G., Barbera, H.M., Izquierdo, M.Z., Minaro, B.U., de Leon, F.C.G., & Balibrea, L.M.T. (2002). MIMICS: Exploiting satellite technology for an autonomous convoy. *IEEE Intelligent Systems, 17*(IV/V), 85-89.

Wu, H., Fujimoto, R., Guensler, R., & Hunter, M. (2004). MDDV: A mobility centric data dissemination algorithm for vehicular networks. *Proceedings of ACM VANET.*

Yabusaki, M., Okagawa, T., & Imai, K. (2005). Mobility management in All-IP mobile network: End-to-end intelligence or network intelligence? *IEEE Communications Magazine, 43*(12).

Yan, C., & Atwood, J.W. (2006). Towards minimizing service degradation during MIPv6 handovers. *Proceedings of the 31st IEEE Conference on Local Computer Networks* (pp. 549-553).

Yanchao, Z., Wei, L., & Wenjing, L. (2005). Anonymous communications in mobile ad hoc networks. *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies* (vol. 3, pp. 1940-1951).

KEY TERMS

Global Navigation Satellite System (GNSS): A satellite system that is used to pinpoint the geographic location of a user's receiver anywhere in the world.

Mobile Ad-hoc NETwork (MANET): A type of wireless network, made of a self-configuring network of mobile devices. These devices are connected by wireless links. The union of all these devices forms an arbitrary topology.

On-Board Unit (OBU): Computing device located in a vehicle that establishes the connection with the RSU and the other OBUs on nearby vehicles.

Roadside Unit (RSU): Computing device located on the roadside that provides connectivity support to passing vehicles.

Universal Mobile Telecommunications Service (UMYS): A third-generation (3G) broadband, packet-based transmission of text, digitized voice, video, and multimedia at data rates up to 2 Mbps.

Vehicle to Infrastructure (V2I): Involves all the communication aspects between vehicles and the supporting infrastructure.

Vehicle to Vehicle (V2V): Relates all the data communication that takes place between various vehicles on the road.

Vehicular Ad hoc NETwork (VANET): A mobile network whose node are vehicles (i.e., car, trucks, etc.); considered an extension of a MANET (Mobile Ad hoc Networks) because of the possibility of working without a fixed infrastructure.

Wireless Technologies to Enable Electronic Business

Richi Nayak
Queensland University of Technology, Australia

INTRODUCTION

Research and practices in electronic business (e-business) have witnessed an exponential growth in the last few years (Liautand & Hammond, 2001). Wireless technology has also evolved from simple analog products designed for business use to emerging radioactive, signal-based wireless communications (Shafi, 2001). The tremendous potential of mobile computing and e-business has created a new concept of mobile e-business or e-business over wireless devices (m-business).

BACKGROUND

M-business can be defined as the use of mobile technology in exchange of goods, services, information, and knowledge. M-commerce is the execution of transactions done on mobile equipment via mobile networks, which may be wireless or switched public networks. M-business includes the range of online business activities, business-to-business and business-to-consumer, for products and services through wireless devices such as mobile phones with display screens, personal digital assistance (PDA), two-way pagers, and low-end or reduced-size laptops.

Example applications are mobile ticketing and receipting, banking, gaming, e-mail communication, weather forecast, sport scores access, movie database access, stock exchange information, ordering of books, and other daily needs such as food and groceries. With new emerging mobile applications, users only receive selective and real-time information personalized to their interests (Ratsimor, Korolev, Joshi & Finin, 2001). For example by using a positioning system, the advertising information of local services and entertainment can be sent whenever a user is passing by a shopping mall. Another application is "inventory management" that tracks the location of goods, services, and people to determine delivery times. Multiple trucks carry large amounts of inventory that companies could access for just-in-time delivery (Varshney & Vetter, 2002).

Significant benefits of m-business to consumers are convenience, portability, safety, integrating existing mobile phones with mobile computing technology, verifiable receipts, and transaction records that can be made available instantly and permanently on a smartcard. Significant advantages of m-business to service providers and content providers include driving additional revenue and decreasing consumer attrition by offering new m-business services to specific groups of customers.

WIRELESS TECHNOLOGIES TO ENABLE M-BUSINESS

Many wireless technologies exist to enables m-business services (Tsalgatidou, Veijalainen, Markkula, Katasonov & Hadjiefthymiades, 2003). All technologies try to achieve benefits such as being powerful, simple, economical, and secure. Some examples of these techniques follow.

Wireless Application Protocol technology links wireless devices to the Internet by optimizing Internet information so it can be displayed on the small screen of a portable device.[1] Web pages accessed by WAP-enabled mobile portals during m-business transactions must be written in WML.[2] It is not sure how well WAP will be able to proliferate (Tsalgatidou et al., 2000). Developments such as third-generation (3G) mobile communications and XYPOINT WebWirelessNow applications (Wen, 2001) already allow mobile phone consumers to experience the Web services without WAP.

Wireless Internet connecting technologies that offer textual interface such as WAP significantly suffer from the constraints of wireless communication such as having a small display screen. An alternative solution is providing voice access to users. Advances in speech recognition and text-to-speech technologies have made voice-based communication possible between computers and users over the phone.

VoxML[3] technology, based on the W3C XML standard, enables the application interface to be in the form of dialogues. However, there is an extra overhead for content providers to offer the same Web service through different channels, for example, providing a voice-enabled browser for their wireless customers along with the HTML/XML/WML browser. Another overhead is the processing power that speech recognition requires. Also this type of data transfer mode is not appropriate for applications with confidential data where one could be overheard. Overall, the success of this technology depends on public acceptance of mobile

phones as data-delivering tools and the type of applications best suited to their use.

The *Bluetooth* technology further enhances the sphere of mobility by conducting m-business without a heavy network infrastructure unlike WAP and VoxML technologies.[4] The Bluetooth technology is designed to allow low-cost, short-range data (asynchronous) and voice (synchronous) radio link (2.4 GHz, 1 Mb/sec) to facilitate protected connections for stationary (homes, buildings, shopping centers, restaurants, cars, etc.) and mobile (phones, PDAs) computing environments. A simple example of a Bluetooth application is to automatically update mobile phone contents such as phone list, e-mails, and memos without any user involvement when the phone comes within the range of the home/office PC. Currently, the Bluetooth networks providing m-business services are limited to 10 meters only. Also, it has too many flaws in terms of security for the services to be trusted. A promising future of Bluetooth technology is its integration with WAP or VoxML.

Based on infrared technology, the *IrDA (Infrared Data Association)* easy-to-use technology provides low-cost, short-range, point-to-point connectivity between devices, interoperable/cross-platform at a wide range of speeds (115.2kb/s to 4Mb/s) with a physical range of 1 meter. IrDA technology is embedded into 40 million new devices each year such as personal computers, laptops, mobile phones, PDAs, digital cameras, pagers, and so forth.[5] The keyword of IrDA advantages is simplicity for ad-hoc, point-to-point exchange. However, the requirement of direct line of sight for devices to communicate is surely a disadvantage for conducting m-business.

IEEE802.11 (Wi-Fi) technology provides a high data rate over different ranges (54Mbps using the 2.4 and 5 GHz ISM band, 11Mbps using the 2.4 GHz ISM band).[6] The single Media Access Control protocol helps to keep the cost down, but interoperability is a problem. The data transmission rate has to be defined before the transmission between devices can start. In terms of the transmission itself, it is based on the well-known TCP/IP protocol. The availability of unlicensed spectrum is a significant enabler for broad acceptance of this technology. However, the technology has some security flaws. Because of the large physical range (100+ meters) and "always-on" connection model, this technology consumes a lot of power, limiting its use in PDAs, phones, and other lightweight mobile devices. The greatest advantage of this technology for conducting m-commerce is its speed.

HiperLAN, a specification substandard of IEEE802.11, is a short-range technology (from 10 to 100 meters) adapted to 3G networks with low power requirements.[7] HiperLAN provides flexible services such as mobility management and quality of service at low cost. The technology has a potential for conducting m-commerce in terms of supporting both ad hoc and client/server networks.

Ultra Wideband technology is a recent RF technology with advantages like large bandwidth, high data transfer rates, and immunity to interference.[8] Still, the technology is in its early stage of development, and there are not many products using this technology yet. However, in the future this network technology may be a very good alternative for conducting m-commerce.

Mobile Agent technology offers a new computing paradigm in which a program, in the form of software agents, is initiated at the host, can suspend its execution on a host computer, launch itself to another agent-enabled host on the network, resume execution on the new host, and return back to its host with the result (Hayzelden & Bigham, 1999). This type of paradigm advocates the client/server model where the client is a mobile portal and the server is a fixed network. The mobile agent performs various optimizations on the server in lieu of its mobile portal to reduce the problems such as C-autonomy, limited bandwidth, and limited computational power. The fixed network offers its services to the agent such as access to local resources and applications, the local exchange of information between agents via message passing, basic security services, creation of new agents, and so froth. Many research papers emphasize that one of the most promising approaches for developing e-business applications is mobile agent technology (Dikaiakos & Samaras, 2001; Tsalgatidou et al., 2000).

Figure 1. A typical platform enabling m-business services

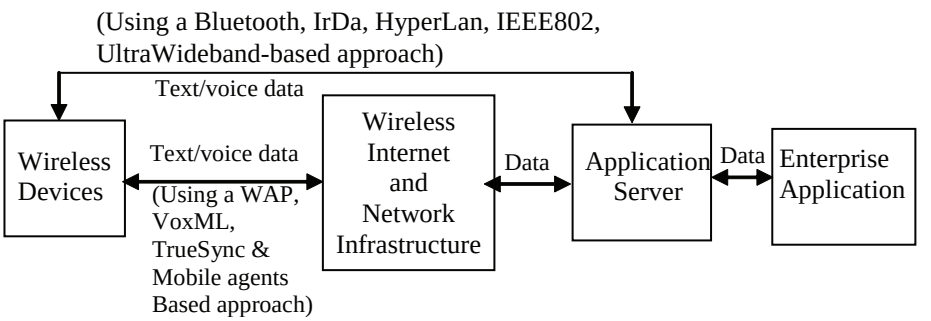


Figure 2. A mobile commerce cycle (Varshey & Vetter, 2002)

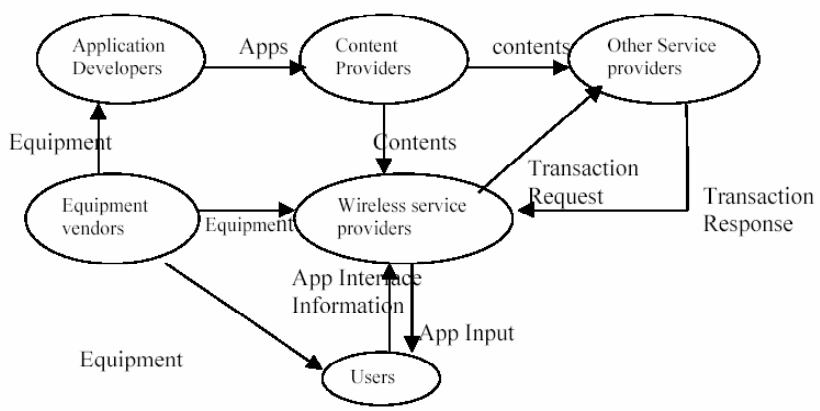


Figure 1 illustrates a typical platform to enable an m-commerce application. According to the wireless technology, a request from the wireless device is passed on to the application server, where the request is processed and output is returned.

Varshey and Vetter (2002) proposed a mobile commerce cycle that shows the possible interaction between various entities (see Figure 2).

CHALLENGES IN M-BUSINESS

As m-business bears its roots in e-business, many technical, legal, and business perspectives are extrapolated from e-business. But m-business embraces a number of technologies to enable its activities, such as wireless devices, software, and protocols, as well as eliciting a number of issues of its own.

Limited Environment of Mobile Devices

These devices usually have limited display size, input capabilities, data transfer rate, computation power, storage, power usage, and so forth; some may be overcome with the technological developments. For example, low bandwidth problem can be handled by the use of software compression techniques or by only sending the relevant information with the use of positioning systems to reduce the traffic. Such mobile applications should be developed so that a large footprint is not required (one innovation is NTT DoCoMo).[9] However, a low-power, inexpensive, and high-resolution color and larger size display will positively affect the m-business.

Security

Consumers should feel the same sense of security when they shop using a mobile phone as when they shop in the physical world. Connection to a wireless link does not require physical access to the network or devices in an "always-on" mode when connected, as in 3G devices. This makes wireless devices more susceptible to attack. In addition, mobile devices are prone to be lost or stolen. Various encryption and integrity mechanism are required to protect digitally encoded speech and control information in m-business services (Chari, Kermani, Smith & Tassiulas, 2001).

Transactional Issues

Usually transactional models with ACID (Atomicity, Consistency, Isolation, Durability) properties assist an application developer in providing powerful abstraction and semantics to concurrent executions and recovery in transaction management. M-business transactions require these properties to be redefined due to additional requirements such as: (1) the need to transfer money and goods along with data transfer, and (2) increased risk of incomplete transactions as mobile terminals can easily lose network connections.[10] Approaches such as the asymmetric cryptographic algorithms (also called Public Key algorithms) with certification authorities are utilized to fulfill ACID properties of m-business transactions (Veijalainen, 1999; Tsalgatidou et al., 2000).

Interoperability

The m-business market is still in its infancy with many standards, and many standards are competing with each other. This creates doubt for the success of global roaming and for the infrastructure. There is interference to one technology with the use of another technology in terms of broadband

spectrum. There have been some improvements in convergence of functions between phones and PDAs. But still, each PDA device comes with its own operating system.

Slow Growth

M-business has not lived up to the promises of the past. Users are not keen to upgrade their mobile phones to take full benefit of m-business applications. Users feel that existing m-business applications do not drive the force to adopt new communications such as 3G. Accordingly, there is delay in implementing system infrastructure such as 3G mobile networks.

FUTURE TRENDS AND CONCLUSION

M-business applications improve the value of the service for the providers, and give an easy and natural interfacing and interacting to the users with mobile portals. Considering it is difficult to provide full convenience due to the limited nature of the wireless devices, the application seems to be able to offer ease of navigation and provide real-time information to users using mobile devices, anytime anywhere.

Many optimists see m-business as a technology that is just one step from everyday use. Many pessimists see many unsolved problems and predict that m-business will not break through in the next few years. As usual, the truth lies somewhere in the middle. Basic techniques are already available. Millions of people are already using mobile portals. Businesses are making profit by moving on to e-business solutions. The potential of m-business is enormous.

So why not integrate them all? Major mobile service providers are taking initiatives (such as MeT,[11] GMCF,[12] WDF,[13] NTT DoCoMo) to envision this technology to flourish. The remaining tasks are rigorous testing and refining of protocols especially suited for m-business applications, resolving related technical and business issues, thus winning the trust of consumers to use m-business services.

REFERENCES

Chari, S., Kermani, P., Smith, S. & Tassiulas, L. (2001). Security issues in m-commerce: A usage-based taxonomy. In J. Liu & Y. Te (Eds.), *E-commerce agents* (pp. 264-282). Berlin: Springer-Verlag (LNAI 2033).

Dikaiakos, M.D. & Samaras, G. (2001). Performance evaluation of mobile agents: Issues and approaches. In R. Dumke et al. (Eds.), *Performance engineering* (pp. 148-166). Berlin: Springer-Verlag (LNCS 2047).

Hayzelden, A. & Bigham, J. (1999). *Software agents for future communication systems*. Berlin, New York: Springer-Verlag.

Liautaud, B. & Hammond, M. (2001). *E-business intelligence: Turning information into knowledge into profit*. New York, London: McGraw-Hill.

Ratsimor, O., Korolev, V., Joshi, A. & Finin, T. (2001). Agents2Go: An infrastructure for location-dependent service discovery in the mobile electronic commerce environment. *ACM Mobile Commerce Workshop*. Retrieved October 5, 2001, from *citeseer.nj.nec.com/455113.html.*

Shafi, M. (2001). *Wireless communication in the 21st century*. John Wiley & Sons/IEEE Press.

Tsalgatidou, A. & Veijalainen, J. (2000). Mobile electronic commerce: Emerging issues. *Proceedings of 1st International Conference on E-Commerce and Web Technologies* (EC-Web), London-Greenwich, UK.

Tsalgatidou, A., Veijalainen, J., Markkula, J., Katasonov, A. & Hadjiefthymiades, S. (2003). Mobile e-commerce and location-based services: Technology and requirements. *ScanGIS 2003,* 1-14.

Varshney, U. & Vetter, R.J. (2002). Mobile commerce: Framework, applications and networking support. *ACM Mobile Networks and Applications, 7*, 185-198.

Veijalainen, J. (1999). Transactions in mobile electronic commerce. In G. Saake, K. Schwarz & C. Trker (Eds.), *Transactions and database dynamics.* Berlin: Springer-Verlag (LNCS 1773).

Wen, H. (2001). Doing the Web without WAP: A discussion with XYPoint's Patrick Carey. *The Wireless Communication Channel*. Retrieved October 1, 2001, from *www.wireless-devnet.com/channels/lbs/features/xypoint.html*

KEY TERMS

Electronic Business (E-Business): Any type of business transaction or interaction in which the participants operate or transact business or conduct their trade electronically.

Mobile Electronic Business (M-Business): A range of online business activities, business-to-business and business-to-consumer, for products and services through wireless devices such as mobile phones with display screens, personal digital assistance (PDA), two-way pagers, and low-end or reduced-size laptops.

Voice Markup Language (VoxML): Based on the W3C XML standard; designed to support interactive dialogues.

VoxML masks the technology behind the voice-to-voice communications by using XML data-tagging structures to link the text-to-speech that generates audio with the speech-recognition software that interprets a user's command.

Wireless Application Protocol (WAP): An approach to link wireless devices to the Internet by optimizing Internet information so it can be displayed on the small screen of a portable device.

ENDNOTES

1 Edited and retrieved September 28, 2001, from *www.wapforum.org*

2 Edited and retrieved September 25, 2001, from *www.oasis-open.org/over/wap-wml.html*

3 *voxml.mot.com*

4 Edited from "Technology Overview"; retrieved October 1, 2001, from *www.bluetooth.com/v2/document*

5 *www.irda.org*

6 *grouper.ieee.org/groups/802/dots.html*

7 *www.hyperlan.com*

8 *www.uwb.org*

9 *www.nttdocomo.com*

10 Interested readers can refer to Veijalainen (1999) for a detailed study of transaction issues in m-business.

11 MeT, *www.mobiletransaction.org*, targets to establish a framework for secure mobile transactions, ensuring a consistent user experience independent of device, service, and network.

12 The Global Mobile Commerce Forum (GMCF) was established in 1997 by a diverse group of companies from around the world to promote the development of mobile commerce services for the benefit of consumers and the companies involved.

13 The Wireless Data Forum (WDF) was established in 1999 to help the wireless industry to develop new e-business products and services, and to use the Internet to sell products and services.

This work was previously published in Encyclopedia of Information Science and Technology, edited by M. Khosrow-Pour, pp. 3101-3105, copyright 2005 by Information Science Reference, formerly known as Idea Group Reference (an imprint of IGI Global).

Category: IT Education

The World Wide Web and Cross-Cultural Teaching in Online Education

Tatjana Takševa Chorney
Saint Mary's University, Canada

INTRODUCTION

The increasing number of virtual universities and online training with a global reach indicates that the opportunities and demands for successful cross-cultural communication expand exponentially, and that instructional paradigms are shifting. Online and distance education is increasingly becoming part of traditional universities as well (Irele, 2005). In 1997, over 60% of all public institutions of higher learning in the U.S. offered distance education courses; by 2001, that number rose to 90% (IES, 1997; 2001). In Canada that number is currently estimated to be 85%[a]. An online teaching environment "goes beyond the replication of learning events that have traditionally occurred in the classroom and are now made available through the Internet"; it provides for different and new approaches to learning, and calls for "flexible teaching….that incorporates a variety of access opportunities as well as a variety of learning modes" (CATL, p. 1). Online teaching here refers to teaching that takes place in programs and courses that incorporate an online component such as WebCT, those that rely completely on WebCt and other similar applications to deliver course or program content, as well as courses offered internationally as part of institutions' distance education degree programs. As online teaching is gaining prominence, educators are compelled to interact meaningfully with individuals from different cultures daily. These interactions demonstrate that teaching and learning are culturally-based processes and that instructional content and how it is experienced reflects the values and practices of a particular cultural group.

The new realities place new demands on educators' knowledge and skills. The cross-cultural context of instruction poses a number of challenges associated with cross-cultural communication in general, such as different communication and decision-making styles, different approaches to task-completion, knowledge, disclosure, and different attitudes toward the learning situation in general. These challenges can lead to misinterpreting the intentions behind certain actions and behavior. In addition, teaching in an environment where many students possess knowledge that they do not, educators have to become collaborative designers, instructional planners, mentors and facilitators of learning, rather than transmitters of authoritative knowledge in a traditional sense. They need to acquire greater familiarity with different learning styles, as well as understand that many of the components determining the nature of learning styles and attitudes toward learning are culture-based (Chorney, 2007; Hao, 2004; Kim, 2001).

Computer-mediated communication (CMC) and the properties of the online environment in general are inherently suited to help educators reconceptualize their role and engage in constructive cross-cultural communication. This is due to the new technologies' potential to enable collaborative teaching in an environment of diverse users and to support multiple learning styles. At the same time, the presence of collaborative technology itself does not guarantee that successful cross-cultural communication and learning will take place. The disembodied nature of online communication can sometimes add to the inherent challenges that accompany face-to-face cross-cultural communication.

Instructors who teach in cross-cultural contexts online will need to engage with the new technologies in a more purposeful way and apply that engagement to program design and teaching practice. They will need to devote some time to designing for interaction and collaboration in order to overcome common challenges in cross-cultural communication.

A more systematic study of the open-ended and interaction-enabling properties of the World Wide Web would help those who design for diversity in online educational environment. The open-ended and interactive nature of the World Wide Web, as the main platform for online cross-cultural teaching, can serve as a conceptual model to help teachers overcome common challenges in cross-cultural communication.

BACKGROUND

As e-learning is gaining prominence, and as distance education turns our world into a "global village", compelling educators to interact meaningfully with individuals from different cultures daily, it is becoming clear that both learning and teaching and culturally-based processes, and that instructional design is not culturally neutral (Campbell, 2004; Chorney, 2007). Instructional content, and the way that content is experienced, reflects the values and practices of a particular cultural group—most commonly, English speaking western cultures. Unless greater care is taken, this situation can alienate a number of students.

Since all education is based on interaction and communication, and cultural differences are often at the root of communication challenges, educators' ability to deal with those differences will determine largely how successful they are in practice. In cross-cultural contexts, teachers acknowledge that learners bring prior knowledge and experience to the learning environment. In these contexts, teachers can no longer see themselves as exclusive sources of knowledge. Rather, they need to see themselves as guides who facilitate the learners' navigating through networks of existing meanings to create new ones. They need to encourage learners to make connections between previous and new knowledge, to integrate previous knowledge with new knowledge, and transfer it from one context to another. In the new paradigm, teachers teach "for transfer", and embrace collaborative teaching.

Collaborative teaching rests on the assumption that learning is "more of a process than a product, in which internal meaning is made through the building and reshaping of personal knowledge through interaction with the world" (Campbell, 2004, p. 152). Collaborative teaching means engaging learners in the learning process and encouraging them through various activities to construct knowledge in a way that is meaningful to them. Teaching collaboratively means being willing to recognize and practice explicitly the reality that there is always more than one way to solve a problem, and more than one point of view in interpretation. Collaborative teaching is reflective, as it implies that instructors will be willing to reflect on their teaching, words, claims, and so on, on an ongoing basis, and be prepared to change their perspective at any given point if change is needed. The instructor who is committed to teaching collaboratively will be teaching students the nature and value of successful communication and collaboration by example.

Like the organization of materials on the WWW, this teaching model is inherently nonlinear, as it encourages the making of connections and identifying of differences among a multiplicity of perspectives on the same issue in no particular linear order. The collaborative model is based on flexible thinking, and is best achieved through the practice of so-called "transformative communication". There are a number of indicators that transformative communication is happening. Among them are the following:

1. The student teaches the instructor something that he or she did not know before, either about the technology or about content.
2. More emphasis is placed upon finding support or backing for a position than on conforming to an authority.
3. Students participate in setting the agenda for the class by helping choose content, learning methods, or both.
4. Students are calling the instructor's attention to valuable learning resources.
5. While the instructor helps establish expectations and articulates a clear assessment standard, the students collaboratively guide much of their own learning.
6. The instructor finds him or herself saving student work—not merely as examples of student work, but as content resources for future reference (cf. Sherry & Wilson, 1997).

This flexibility of approach relying on collaboration and learning as a process becomes crucial in the context of cross-cultural instruction. Individuals process information and approach learning in different ways, which results in different learning or "mind" styles. There are a number of different classifications concerning learning styles (cf. Gardner, 1983, 1993, 1999; Keefe, 1979). One such classification accounts for cognitive differences among learners according to two criteria: the way learners acquire information—though concrete experience or abstract conceptualizations; and according to how they internalize or process information—through active experimentation based on the method of scientific, deductive reasoning, or reflective observation (Kolb, 1984, 1985). Some learners prefer and appropriate knowledge and information offered through text, others through images and graphic representation. While most students have become proficient in interpreting text or print, only a portion of those students is actually composed of so-called "verbal learners," those who prefer to learn from texts and lectures (Campbell, 2004, p. 178).

Individual responses to the learning situation will be influenced by the learners' prior knowledge and the way they think of the individual's past experience, and this, in turn, will depend in some definite measure upon each person's background, including the individual's culture. While models of cognition are not entirely predetermined, they are also shaped through social interaction (Helwig, 2005; Nations Johnson, 1993; Oishi, Hahn, Schimmack, Radhakrishan, Dzokoto & Ahadi, 2005; Smetana, 2002). Since individual development is mediated by social interaction in a culture-specific, historical setting, and since culture influences one's cognitive processes, including the attitudes governing the assimilation of information, there is common ground on which culture and its impact on cognition can be studied (Abi-Nader, 1999; Neff & Helwig, 2002). The relationships between learning styles and cultural backgrounds are therefore strong and complex (Hao, 2004; Kim, 2001).

The nature of this relationship and its implication for cross-cultural education can be understood in the context of the often cited differences among cultures and the attitude those differences shape. There are a number of ways according to which cultural differences have been conceptualized (cf. McCutcheon, 1993). One recent model interprets the differences in terms of a "global learning style," associated with Japanese learners, vs. an "analytical learning style," associated with learners from Europe, North

America, Australia, and New Zealand (Ito, 2002). Japanese students, who are seen as exemplifying the "global learning style" are generally image-oriented, cooperative, learn by experience, depend on insight and intuition, prefer indirect expressions, value the subjective, and avoid standing out. On the other hand, learners associated with the "analytical learning style" generally learn by reasoning; they compete, assert themselves, value the objective, are text oriented, and prefer direct expressions (Ibid.).

What is clear is that attitudes toward learning are built into the education system, and that as difficult as they are to systematize, they will influence the learners' own approaches to learning and knowing. In turn, this is likely to result in different communication styles, different attitudes toward task completion, disclosure, assertiveness, and so forth. However, it is very important that these classifications, like any other attempt at systematizing human thought and behavior, are understood in a broad, nonrestrictive sense. Individual characteristics will always vary and overlap and no single model can completely define any one person.

Although we can and should continuously strive to learn as much as we can about the values and assumptions of different cultures, we can never hope to learn everything there is to learn about all cultures. The issue, here, therefore, is not so much about learning cultural content, as much as it is about learning a method of approaching and dealing with cultural differences in general. Since we tend to design "from our own experience and based on our own needs and values" (Ito, 2005), educators who teach online and in a cross-cultural context need to reflect on the diversity among the students they teach, and the implications of diversity for the course and program design. The activity in which knowledge is developed and used is neither separable not ancillary to what is learned, but an integral part of it, as both learning and cognition are fundamentally situated and embedded in the context in which they take part (Seely Brown, Collins, & Duguid, 1998). Teaching and learning, in both content and methodology, are interdependent processes, just as activity, concept, culture, and affect and cognition are interdependent and inextricably linked (cf. Seely Brown et al., 1989).

Due to these realities, educators need to reflect on their own cultural and pedagogical assumptions. Seeing that our own culture provides the lens though which we see ourselves and others in the world, it is of paramount importance to reflect purposefully and specifically on the values and experiences that have shaped our own pedagogical approach. This kind of reflection with the purpose of attaining greater understanding of ourselves as members of a particular culture will aid in instructional planning and design in a cross-cultural context, and lead to the practice of collaborative teaching. This approach is of crucial relevance in cross-cultural communication where the existence of multiple perspectives will be the beginning point of interaction.

Although it may seem that the instructor who teaches collaboratively teaches only method and not content, this is not the case. This practice, on the one hand, demonstrates the teacher's engagement with his or her discipline or subject, which is an active process and the basis of good scholarship. On the other hand, it enables students to develop a composite understanding of the subject matter and understand the connections that exist between different approaches. It shows students the "the legitimacy of their implicit knowledge"; it stresses that "heuristics are not absolute but assessed with respect to particular tasks" and thus always situated. It also helps students to "generate their own solution paths…making them conscious, creative members" in the process of knowledge creation (Seely Brown et al., 1989, p. 40).

As a conceptual model and a platform where most online education takes place, the World Wide Web offers unprecedented opportunities for instructors to accommodate preferences relating to different learning or "mind" styles, different models of acquiring and processing information, and different ways of creating knowledge.

THE WORLD WIDE WEB AND CROSS-CULTURAL COLLABORATIVE TEACHING

The nature of the online environment can support the features of active and collaborative teaching and the reconceptualization of the teacher's role. The World Wide Web is a "universe of network-accessible information" (W3C, 2001). Through the use of hypertext, multimedia techniques and Web 2.0 applications, "the web is easy for anyone to roam, browse and contribute." Examples of Web 2.0 applications, such as Wiki projects, refer to collaborative computer software used to create collaborative websites that now provide an infrastructure for an even more dynamic user participation, social interaction and collaboration, and demonstrate the direction of future knowledge creation and dissemination. As a concept, the WWW thus represents a "seamless world in which all information, from any source can be accessed" and connected though links (Ibid.).

As a conceptual model, the WWW demonstrates the coexistence of and interrelationships between multiple and apparently contradictory perspectives on a single issue. It is a good example of a "space" allowing for various "narratives", and "knowledges" to circulate, and to be added to existing collections and systems of meaning (cf. Seely Brown et al.,1989, p. 39). The WWW exemplifies that there is not only one kind of valid knowledge that can be viewed as absolute and universally applicable. It shows instead that there are various kinds of knowledges, situated within various communities of knowing, each operating according to

particular cultural dynamics, yet sharing in universal issues that affect us all. When applied to the educational context, these properties also point toward a collaborative model of instruction, suitable for a culturally diverse student body.

Collaborative teaching practices are in many ways exemplified by the open-ended structure and nonlinearity of the Web, which has great potential to support communication across and among diverse communities of knowers (Campbell, 2004). The diversity of materials found on the Web and their coalescence is similar to the diversity of perspectives and knowledges that coalesce within a cross-cultural community of learners. Like collaborative teaching and learning, the loose organization of the World Wide Web demonstrates the coexistence of multiple and apparently contradictory perspectives on a single issue. One Google keyword search will retrieve hundreds of documents linked by one single term, but applied in a variety of contexts. This characteristic of the Web environment, just like collaborative and cooperative teaching, emphasizes the importance of individual contexts to which a single concept can be applied productively. In the fluid online environment, supported by teaching that is social and interactive, learners no longer rely on a situation where the instructor provides a single learning context and suggests the desired connections among the offered concepts. They are encouraged to generate connections between problems and solutions, and apply the findings in a transformative way to a variety of contexts meaningful to them.

The Web, as a platform in which collaborative teaching can happen, is open, flexible, multimodal, and networked, and as such it provides "the richest opportunity to date to bring the elements of active learning together" (Campbell, 2004, p. 154). It resists systematization in a traditional sense, it allows for multimodal teaching and learning, and it accommodates multiple 'literacies' and learning styles. The non-linear and interactive multimedia capacity inherent in the Web and forms of CMC can present knowledge and information in ways that combine orality, literacy and "videocy" (Ulmer, 1989, p. vii). In this way, CMC and the properties of the World Wide Web can support the complex range of learning needs, characteristics and preferences. Web-based communication tools such as email, relay-chat, forums, and synchronous (i.e., realtime) conferencing are seen as potentially enabling dialogue. In turn, dialogue encourages critical thinking and cooperative learning, and enhances opportunities for "generative learning, wider diversity of ideas, most reflective thinking, and increased creative responses" (Oliver, Omari & Herrington, 1998). The hypermedia most effectively support tasks and forms of interaction requiring high-level reasoning, problem-solving skills, and critical thinking (Nunes & Fowell, 1996; Ryser, Beeler & McKenzie, 1995).

Forms of online teaching and learning relying on CMC, in general, have been termed "collaborative technologies" since they have an inherent affinity with definitions of learning emphasizing social, interpersonal, and collaborative interaction. Even simple, text-based CMC "equalizes the participants to the extent that everyone, regardless of gender, race, authority, age, etc., is limited to exchanging texts" (Markham, 1998, p. 155). There are factors that need to be taken into consideration, such as economic status or individual facility with manipulating texts in CMC, as they may affect the potential equality of participants (Mason, 2002). It is clear, however, that new models of communication enabled by the WWW lead to breaking down, or at least diminishing the distinction between "private and public writing" (Bolter, 1991, p. 102), and thus potentially create a space for constructive communicative intimacy. This trait is valuable in cross-cultural CMC where the onus is on the instructor to create a sense of trust and open-mindedness in an inclusive and culturally diverse learning community. The potential of the Web to support collaborative teaching, multiple learning styles, and to foster critical thinking is especially important for cross-cultural teaching where the diversity of learning styles is matched and often conditioned by diversity in cultural perspectives.

The nonlinearity of the presentation of materials and ides on the World Wide Web is a useful conceptual and philosophical model for collaborative, cooperative teaching and learning. The nature of the presentation of material and ideas on the WWW encourages active, intentional involvement on the part of learners, as there is no longer one "solution" or a single "interpretation," but a variety, all situated within their own context and knowledge. Due to its multimedia capacity, online communication and the WWW facilitate the use of tools promoting connectivity, transfer, the seeking of interrelationships among texts and/or audio-visual materials. They enable the creation of various interpretive contexts in instruction, relevant given the cultural differences in learning and communication styles. The nonlinearity of the Web models scaffolding as a way of grouping materials; it promotes the idea that conceptual knowledge cannot be separated from the contexts in which it is represented, as well as that learning works best when it is situated in the individual learner's implicit knowledge.

There are a number of practical issues that instructors who teach collaboratively in cross-cultural contexts in online education could consider in order to utilize the conceptual potential of the WWW. What follows are some issues arising from the ideas presented:

1. **Success in communication depends on the context of reception:** The outcome and success of communication in general is largely dependent on the context of reception. In face-to-face interaction, body language allows interlocutors to interpret meaning that may not be present in verbal content. This context is mostly absent in online communication. With cross-cultural communication in particular, educators should be aware

that "cultures vary in what they consider humor and taboo, which may give rise to misinterpretation and resentment, and that speed of delivery…as well as turn-taking should be respected, as much as the rules for entering conversations in progress" (Zeinstejer, 2002). In addition to using emoticons to reveal the intention in which a statement is made and the hoped-for response, a good way of dealing with potential ambiguity would be to acknowledge openly at the beginning of the learning unit that language, as a communicative tool, is embedded in a number of culture-specific uses, and as such, it can be ambiguous[b].

2. **Design for interaction and collaboration by gaining a clear sense of "audience":** In order to avoid some of the common challenges that accompany cross-cultural communication, instructors could ask participants to fill out a "pre-course" questionnaire. In designing the questionnaire, the instructor may consider how to obtain answers to questions such as: Who are the learners? What is their attitude toward interaction? What do they need or want to learn? In what contexts will the learning be applied? What do they already know?
3. **Organize material in a nonlinear, open-ended format and provide opportunities for learning in context:** Part of the interaction and the structuring of materials should be done in the form of scaffolding, which is a nonlinear, open-ended way of organizing data and enabling learners to adopt the suggested material in a way that allows them to apply it to their own needs, values and contexts. Scaffolding also enables learning in context in that it encourages "continuous sorting and sifting as part of a 'puzzling' process—the combining of new information with previous understanding to construct new ones" (McKenzie, 1999). Open-endedness of approach can be also encouraged by involving students at all level of instruction in "the choice of content, method, medium, reward, assistance, feedback, quantity, pacing, sequencing, or difficulty of instruction" (Sutton, 2004, p. 34), and by creating opportunities for group-learning.
4. **Set clear and explicit expectations for online behavior and communication:** Because of different cultural attitudes toward communication and participation in a group, it will be the instructor's role at the beginning of the learning unit to express general expectations for online behavior and communication. These expectations should be explicit about removing language that appears to stereotype learners and reducing the violations of cultural rules during discussions (Zeinstejer, 2002). The instructor may model continuously the idea that many of the beliefs we take for granted are in fact culturally determined, and that successful communication is based on an honest, benevolent and non-judgmental approach to cross-cultural differences.
5. **Provide opportunities for different kinds of online interaction:** Collaborative interaction occurs when learners have the opportunity to engage in activities enabling individual learning, learning done in pairs or through email (one-to-one), through the use of a bulletin board (one-to-many), and the use of computer conferencing techniques (many-to-many) (Hao, 2004, p. 25). The variety of interactions will successfully address the variety of different learning styles, as well as the students' culturally-determined approaches to communication and to knowing.
6. **Asynchronous communication is preferable for online learning:** Asynchronous communication—enabled by tools such as discussion and bulletin boards, blogs, messaging, surveys and polls—provides opportunities for active input from all members of the learning community with flexibility in time and place, so learners have greater control over the learning environment (Carr, 1998; Graham, Scarborough & Goodwin, 1999; Hao, 2005), as well as an opportunity for "vicarious interaction". Vicarious interaction takes place when a student actively observes and processes both sides of interaction between another student and instructor. This type of interaction is of special value in cross-cultural education as it promotes indirectly, but through conversational situations, awareness and understanding of the issues involved in cross-cultural communication (Chorney, 2007). Synchronous or "realtime" interaction enabled by tools such as Web, audio or video conferencing, chatting, instant messaging and white boarding, is a good supplement to the asynchronous delivery medium (Hao, 2004).
7. **The evaluation scheme should be varied and flexible:** When knowledge is understood as an active, flexible process of "meaning negotiation" in which multiple ways of arriving at different "knowledges" exist, are validated and lead to the same learning goals, the nature and design of student evaluation shifts. In a cross-cultural educational context, content itself may be redefined from emphasizing the gathering and arrangement of factual information in traditional academic formats, to emphasizing various aspects of the research process resulting in a variety of nontraditional formats that can be evaluated with equal academic rigor. Content can be information, as well interpretation of information by experts, novices or students. It can be in the form of research reports generated individually or with a partner/group, arguments, journalistic accounts, and essays represented through text, graphics or any other multimedia format. Similarly, in addition to the instructor, peers can provide feedback.

These practical approaches can help instructors create online learning environments that are inclusive, and that demonstrate the dynamic of successful cross-cultural communication in an educational context using the potential offered by the WWW and CMC.

FUTURE TRENDS

The long-term value of the collaborative approach in cross-cultural teaching contexts can be seen in the correlation between dominant media and cultural practices, including education. Whereas traditional "mass education tended to see life in a linear fashion based on print models and developed pedagogies which broke experience into discrete moments and behavioral bits," new critical pedagogies enabled through the online medium could produce "skills that enable individuals to better navigate the multiple realms and challenges of contemporary life" (Kellner, n.d., p. 9).

A more systematic study of the ways diverse information can be linked and presented on the WWW can take us one step further toward revisioning the goals of education in the 21st century. Most important for the future, further study of collaborative online education in relation to cross-cultural contexts has the potential to provide concrete answers to the call for meaningful reform of higher education (cf. Association of American Colleges and Universities, 2002). Technology and international distance education are already an undeniable presence and a growing trend in higher education. The student body increasingly reflects an extraordinarily diverse array of cultural backgrounds. These facts place an increasing need on institutions and instructors to develop new approaches to educational quality in a way that would serve meaningfully the needs of contemporary students who live in complex, technologized and interconnected world.

Modeling collaborative teaching practices on the WWW may promote the forms of learning needed for the 21st century, as they have been identified by the "Greater Expectations National Panel Report" (Association of American Colleges and Universities, 2002). Taking into consideration that the intellectual and practical skills today's students need are extensive, sophisticated and expanding with the explosion of new technologies, and the changes they have brought to education, cross-cultural collaborative teaching in the online environment will help students become "intentional learners". Intentional learners are those who can "adapt to new environments," who can integrate different kinds of knowledge from a variety of sources, who can "demonstrate intellectual agility and the ability to manage change," and who have the skills to engage in a meaningful dialogue and deal with the interrelations within and among global and cross-cultural communities (adapted from the recommendations of the National Panel Report, pp. xi-xii). Future research may investigate whether systematization of training in cross-cultural communication and collaboration in a variety of contexts, including online education, would help alleviate many of the crises in cross-cultural communication we face today.

CONCLUSION

As a concept, the WWW can illustrate many of the philosophical aspects that should accompany successful cross-cultural communication. Many of the practices associated with collaborative teaching and not new and are being used by many instructors in any classroom. However, the increase in online educational programs compels us to thoroughly examine pedagogical practices and the goals of teaching in an increasingly interconnected global world. These new educational trends encourage us to explore and apply systematically certain teaching practices that can improve the quality of cross-cultural education, especially when delivered by monocultural teachers.

Intentional assessment of the nature the World Wide Web in its potential to enable successful cross-cultural communication in online education can result in the development of pedagogical strategies suited to contemporary realities and the needs of contemporary students (cf. Association of American Colleges and Universities, 2002). The changes will ensure a decrease of instances where many cultural groups feel excluded from e-learning opportunities because the instructional content and the nature of the interaction is not culturally inclusive (cf. Campbell, 2004).

The online environment, usually synonymous with the World Wide Web, can be used as a tool and a model to encourage cross-cultural interactivity and to support forms of open-ended, collaborative teaching techniques. The structural flexibility that is the hallmark of this kind of instruction is similar to the structural flexibility and conceptual "open-endedness" of the Web. The multiplicity of perspectives represented on the Web and their coexistence could remind us that being exposed to and learning to see the world from another's point of view is a process of

cognitive and social growth that can deepen our understanding of ourselves and others.

END NOTES

a. The author gratefully acknowledges the support provided by the Social Sciences and Humanities Research Council of Canada

b. This is the author's estimation based on statistical data available through the Canadian Association of University Teachers and the searchable database of

distance education courses and programs currently offered at various Canadian universities created by the Canadian Association for University Continuing Education (CAUCE) and the Ontario Council for University Lifelong Learning (OCULL).

REFERENCES

Abi-Nader, J. (1999). Meeting the needs of multicultural classrooms: Family values and the motivation of minority students. In M. J. O'Hair & S. Odell (Eds.), *Diversity and teaching: Teacher education yearbook I* (pp. 212-228). Fort Worth, TX: Harcourt Brace Jovanovich.

Association of American Colleges and Universities (2002). *Greater expectations: A new vision for learning as a nation goes to college.* Retrieved June 16, 2008, from http://www.greaterexpectations.org

Bolter, J. D. (1991). *Writing space: The computer, hypertext and the history of writing.* Hillside, NJ: Lawrence Erlbaum.

Campbell, K. (2004). *E-ffective writing for e-learning environments.* Hershey, PA: Information Science Publishing.

CATL: Centre for the Advancement of Teaching and Learning (2005). *Toward a definition of online learning at UWA.* The University of Western Australia. Retrieved June 17, 2008, from http://www.catl.uwa.edu.au/elearning/online/definition

Chariot (2002). *Cross-cultural communication online: Perspectives from around the globe.* Retrieved June 17, 2008, from http://users.chariot.net.au/~michaelc/ccc/pres.htm

Chorney, T. (2007). Teaching, learning, negotiating: The WWW as a model for successful cross-cultural communication. In K. St. Amant (Ed.), *Linguistic and cultural online communication issues in the global age* (pp. 253-275). Hershey, PA: Information Science Reference.

Crawford, K. (1996). Vygotskian approaches to human development in the information era. *Educational Studies in Mathematics,* (31), 43-62.

Gardner, H. (1983, 1993). *Frames of mind: Theories of multiple intelligences.* New York: Basic Books.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century.* New York: Basic Books.

Graham, M., Scarborough, H., & Goodwin, C. (1999). Implementing computer mediated communication in an undergraduate course—A practical experience. *Journal of Asynchronous Learning Networks, 3*(1), 32-45. Retrieved June 17, 2008, from http://www.sloan-c.org/publications/jaln/v3n1_graham.asp

Hao, Y-W. (2005). *Students attitudes toward interaction in online learning: Exploring the relationship between attitudes, learning styles, and course satisfaction.* Unpublished doctoral dissertation, University of Texas at Austin.

IES: Institute of Education Sciences (2001). Distance education at degree-granting post-secondary institutions, 2000-2001. *National Center for Education Statistics.* Retrieved June 17, 2008, from http://nces.ed.gov/surveys/peqis/publications/2003017/

IES: Institute of Education Sciences (1997). Distance education course offerings. *National Center for Education Statistics.* Retrieved June 17, 2008, from http://nces.ed.gov/surveys/peqis/publications/98062/index.asp?sectionID=3

Irele, M. E. (2005). Can distance education be mainstreamed? *Online Journal of Distance Learning Administration, 3. 2.* Retrieved June 17, 2008, from http://www.westga.edu/~distance/ojdla/summer82/irele82.htm

Ito, S. (2002). Cultural inclinations in learning styles. *Cross cultural communication online: Perspectives from around the globe.* Presented by the Webheads Community of Networking. Retrieved June 17, 2008, from, http://users.chariot.net.au/~michaelc/ccc/pres.htm

Jonassen, D. (1999). Designing constructivist learning environments. In C.M. Reigeluth (Ed.), *Instructional-design theories and models: A new paradigm of instructional theory.* (Vol. 2, pp. 215-237). Hillside, NJ: Lawrence Erlbaum.

Kellner, D. (n.d.) *Technological transformation, multiple literacies, and the revisioning of education.* Retrieved June 17, 2008, from http://www.gseis.ucla.edu/faculty/kellner

Keefe, J. W. (1979). Learning style: An overview. In *NASSP, Student learning styles: Diagnosing and prescribing programs* (pp. 1-17). Reston, VA: National Association of Secondary School Principals.

Kim, K-S. (2001). Implications of user characteristics in information seeking on the World Wide Web. *International Journal of Human-Computer Interaction, 13*(3), 323-340.

Kolb, D. (1984).

Kolb, D. (1985). *Learning style inventory.* Boston: McBeer and Company.

Markham, A. (1998). *Life online.* Walnut Creek, CA: Altamira Press.

Mason, J. (2002). *From Gutenberg's galaxy to cyberspace: The transforming power of electronic hypertext.* CITD Press.

Retrieved June 17, 2008, from https://tspace.library.utoronto.ca/citd/JeanMason/orientation.html

McCutcheon, G. (1993). Curriculum: Overview and framework. In M. J. O'Hair & S. Odell (Eds.), *Diversity and teaching: Tteacher education yearbook I* (pp. 237-268). Fort Worth, TX: Harcourt Brace Jovanovich.

McKenzie, J. (1999). Scaffolding for success. *From now on: The educational technology Journal, 9*(4). Retrieved June 17, 2008, from http://www.fno.org/dec99/scaffold.html

Nations Johnson, L. (1993). On becoming a responsive teacher: A self-observational process analysis. In M.J. O'Hair & S. Odell (Eds.), *Diversity and teaching: teacher education yearbook I* (pp. 138-151). Fort Worth, TX: Harcourt Brace Jovanovich.

Neff, K. D. & Helwig, C. C. (2002) A constructivist approach to understanding the development of reasoning about rights and authority within cultural contexts. *Cognitive Development, 17,* 1429-1450.

Nunes, J. M. B. & Fowell, S. P. (1996). Hypermedia as an experiential tool: A theoretical model. *Information Research, 2*(1). Retrieved June 17, 2008, from http://informationr.net/ir/2-1/paperl2.html

Oishi, S., Hahn, J., Schimmack, U., Radhakrishan, P., Dzokoto, V., & Ahadi, S. (2005). The measurement of values across cultures: A pairwise comparison approach. *Journal of Research in Personality, 39*, 299-305.

Oliver, R., Omari, A., & Herrington, J. (1998). Developing converged learning environments for on and off-campus students using the WWW. In R. Corderoy (Ed), In *Conference Proceedings ASCILITE '98* (pp. 529-538). Wollongong, Australia: The University of Wollongong.

Relan, A. & Gillani, B. B. (1997). Web-based instruction and the traditional classroom: Similarities and differences. In B. H. Khan (Ed.), *Web-based instruction* (pp. 41-46). Englewood Cliffs, NJ: Educational Technology Publications.

Ryser, G. R., Beeler, J. E., & McKenzie, C. M. (1995). Effects of a computer-supported intentional learning environment on student's self-concepts, self-regulatory behaviour, and critical thinking ability. *Journal of Education Computing Research, 13*(4), 375-385.

Seely Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Sherry, L. & Wilson, B. (1997). Transformative communication as a stimulus to web innovations. In B. H.Khan (Ed.), *Web-based instruction* (pp. 67-73). Englewood Cliffs, NJ: Educational Technology Publications.

Smetana, J. G. (2002). Culture, autonomy, and personal jurisdiction in adolescent-parent relationships. In H. W. Reese & R. Kail (Eds.), *Advances in child development and behaviour* (Vol. 29, pp. 51-87). New York: Academic.

Sutton, L. A. (2000). The principle of vicarious interaction in computer-mediated communication. *International Journal of Education Telecommunications, 7*(3), 223-242.

Tripathi, A. K. (2006). *Coping with innovative technology: Albert Borgman on how does technology change learning and teaching in formal and informal education.* Retrieved June 17, 2008, from http://www.acm.org/ubiquity/views/pf/v7i23_coping.pdf.

Ulmer, G. (1989). *Teletheory: Grammatology in the age of video.* New York: Routledge.

W3C (2001). About the World Wide Web. Retrieved June 17, 2008, from http://www.w3.org/WWW

Zeinstejer, R. (2002). Teachers learning to achieve successful cross-cultural communication. *Presented by the Webheads community of networking*. Retrieved June 17, 2008, from http://users.chariot.net.au/~michaelc/ccc/pres.htm

KEY TERMS

Analytical Learning Style: (In contrast to global learning style); According to one classification, a style of learning associated with students from Europe, North America, Australia and New Zealand, and with being text -oriented and competitive, asserting oneself, learning by reasoning, preferring direct expressions, and valuing the rational and objective.

Collaborative Teaching: "Open-ended" teaching practice according to which learning is a process achieved through social and interpersonal interaction. It demonstrates the coexistence of multiple and often contradictory perspectives on the same issue; it encourages the discovery of connections among those perspectives, and emphasizes the importance of individual contexts to which new concepts can be applied productively.

Collaborative Technologies: Technologies enabling computer-mediated communication (CMC) have been termed "collaborative" because of their inherent affinity with definitions of learning emphasizing social, interpersonal, and collaborative interaction.

Cross-Cultural Communication: Communication between members of different cultures through which each member's values and patterns of thinking, communication and behavior are often revealed as contrasting the values,

patterns of thinking, communication, and behavior of the other.

Culture: Sets of social relationships, values, patterns of thinking, communicating and behaving that reflect ideas and actions established and accepted by one group of people as habitual, appropriate, or traditional.

Global Learning Style: (In contrast to analytical learning style); According to one classification, a style of learning associated with Japanese students, and with being image-oriented and cooperative, avoiding standing out, depending on insight and intuition, learning by experience, preferring indirect expressions, and valuing the subjective.

Intentional Learners: Model learners of the 21st century, as identified in the Association of American Colleges and Universities Report (2002). They are those who can adapt to new environments, engage in meaningful dialogue, integrate different kinds of knowledge from different sources, demonstrate intellectual agility, and the ability to deal successfully with interrelations within and among global and cross-cultural communities.

Learning Styles: a broad, nonrestrictive combination of cognitive, affective and physiological factors influencing how a learner perceives, interacts and responds to the learning environment.

Transformative Communication: The model of communication between students and instructor through which collaborative teaching happens. It emphasizes the instructor's willingness to learn from students while helping to establish expectations and clear assessment standards.

Vicarious Interaction: Indirect kind of interaction that takes place when a student actively observes and processes both sides of interaction between two other students or between another student and instructor.

.

Index

B

C

D

E

F

G

H

I

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z